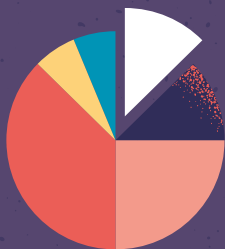


# OPEN DATA TOOLKIT





Institute for Development  
of Freedom of Information



The toolkit was prepared by the Institute for Development of Freedom of Information (IDFI) in frames of the project - *Empowering Watchdog Community for Government Accountability*. The project is co-financed by the Governments of Czechia, Hungary, Poland and Slovakia through Visegrad Grants from *International Visegrad Fund*. The mission of the fund is to advance ideas for sustainable regional cooperation in Central Europe. The responsibility of the content of the document lies with the Institute for Development of Freedom of Information (IDFI). It does not necessarily reflect the opinions of International Visegrad Fund.

**Author:** Teona Turashvili

© Open Data Toolkit, 2020

All rights reserved. The toolkit can be copied with consent from IDFI.

# OPEN DATA TOOLKIT

TBILISI  
2020

# CONTENTS



# INTRODUCTION

Availability of open data in easily processable formats has given journalists a unique opportunity to conduct comprehensive studies of various issues. Combination of various data sets and programmatic analysis has revealed hidden business connections and interests, information about property and income, illegal dealings, and misappropriation of public funds by public figures and high-ranking officials.

However, these new opportunities bring new challenges and a necessity of technical skills. While until recently, statistical data analysis did not concern journalists, today such analysis is the foundational component of journalistic investigations. Such journalism is known as data or investigative journalism. Data journalism encompasses not only description or review of facts but study and determination of their causes and in most cases proposal of solutions to the identified problems.

Data journalism is often characterized as an inverted pyramid, consisting of the following **data processing stages**:



### COMPILING

A journalist gathers and compiles information from various sources.



### CLEANING

This stage comprises of sorting the data, specification, correcting any data inaccuracies, transformation of the data in the appropriate format for identification of correlation and for comparison.



### CONTEXTUALIZATION

During data processing, a journalist must consider who is the primary source of the data, who compiled and published it, when and for what purpose. This enables the journalist to prevent any issues related to the validity or trustworthiness of the data.



### COMBINATION

Often most interesting conclusions can be drawn when combining various data sets. Therefore, data journalists often process various data sets connected to the issue they are working on, finding interesting connections and trends, as well as, validating the already-acquired data in other sets and finding a correlation between the two.



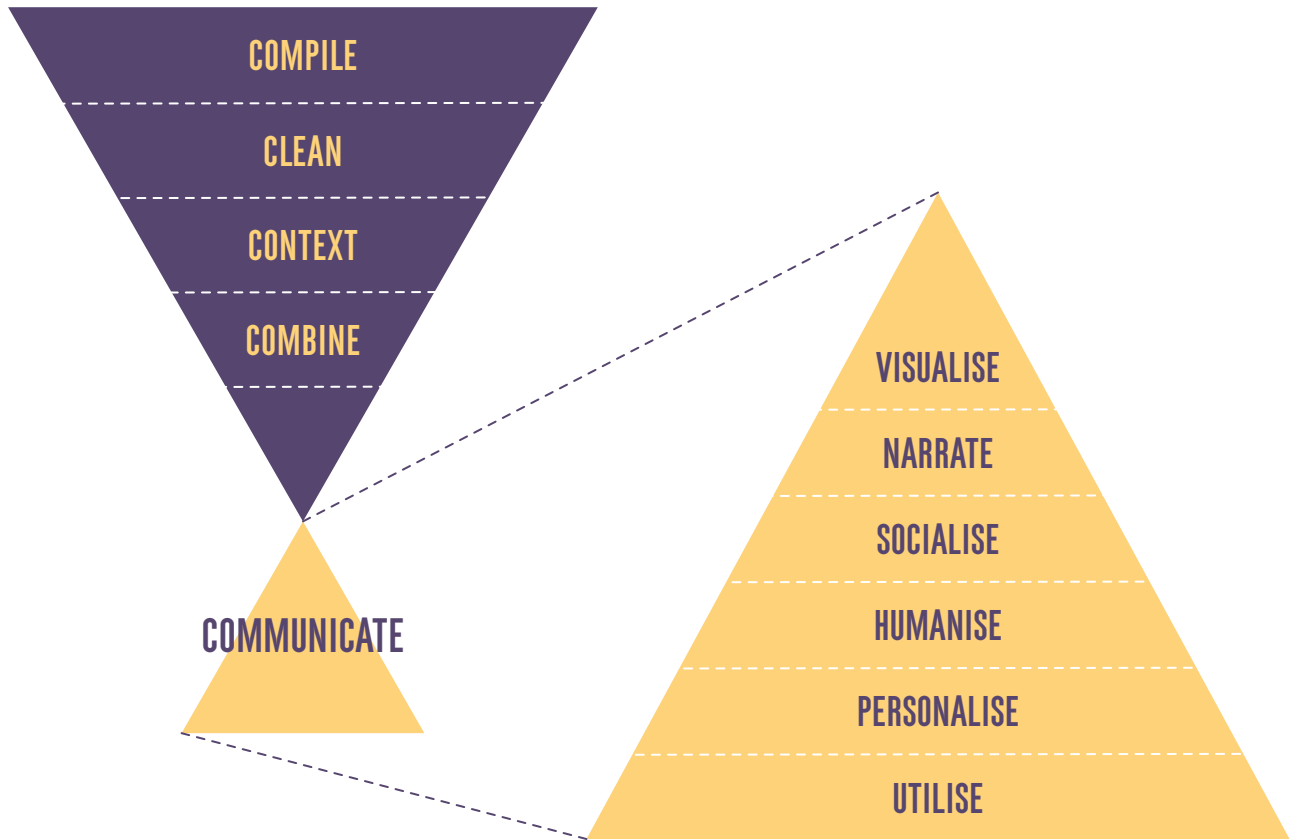
### COMMUNICATION

Data journalists vigorously utilize visualization for communication, as such communication forms have particular significance considering the fast-paced development of new technologies.

---

# THE INVERTED PYRAMID OF DATA JOURNALISM (COMPLETE)

PAUL BRADSHAW, [ONLINEJOURNALISMBLOG.COM](http://ONLINEJOURNALISMBLOG.COM)



To competently study and process data, journalists and activists must comprehend the definitions, fundamental principles, and particularities of working with open data.

## DEFINITION OF OPEN DATA

[Open Knowledge Foundation](#)<sup>1</sup> defines open data as follows: Open data means anyone can freely access, use, modify, and share it for any purpose.

Several important characteristics of open data are included in the definition:



### I. AVAILABILITY

The data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet.



### II. REUSE AND REDISTRIBUTION

The data must be provided under terms that permit reuse and redistribution including the intermixing with other datasets.



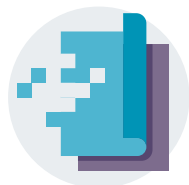
### III. UNIVERSAL PARTICIPATION

Everyone must be able to use, reuse and redistribute — there should be no discrimination.

<sup>1</sup> Open Knowledge Foundation is a global network that supports the distribution of information, including open data, in an open format without any fees. It was founded in 2004 in Cambridge, UK. For more information, visit: [www.okfn.org](http://www.okfn.org)

# OPEN DATA PRINCIPLES

Organizations working in this area emphasize several open data principles, including:<sup>2</sup>



## 1. COMPLETE

Public data should be complete and cover the topic to the fullest extent possible.



## 2. PRIMARY

Open data is provided by the primary source, ensuring its trustworthiness. Alternatively, a user must be able to check whether the data is correctly aggregated. This requires metadata (general description of data), which, along with other parameters, defines how the data was created and processed.



## 3. TIMELY

Published data must be available to the public in a timely manner. It's important that compiled and processed official data is published immediately and that published data is still valuable at the time of publication.

<sup>2</sup> For example: Open Knowledge Foundation. The Open Data Handbook. Available at: <https://okfn.org/opendata/>

Larry Lessig on Open Government Data principles. Available at: <https://opengovdata.org/>

Vivek Kundra's 10 Principles for Improving Federal Transparency. Available at: <https://bit.ly/37b8Y4o>

United Kingdom's Public Data Principles. Available at: <http://data.gov.uk/library/public-data-principles>



#### 4. ACCESSIBLE

Public data must be available to the widest extent possible. Access should be simple, both in electronic and physical forms. A user should not have a need to visit a public institution to receive public data. For increasing accessibility of electronic data, it's recommended that such data be compiled on a unified, central platform/online location. The data should be available for downloading.



#### 5. MACHINE PROCESSABLE

Data must be structured in a way that allows its automated processing for various purposes. This principle is also called machine-readable format.



#### 6. NON-EXCLUSIVE

The data must be available in a format that doesn't give any one party exclusive rights for its distribution. It is recommended that data be available in as many different formats as possible. This will prevent anyone from creating limitations on usage and distribution of data.



#### 7. NON-DISCRIMINATORY

The data must be available to everyone. No registration, requiring identification of a person, should be necessary.



#### 8. NON-PROPRIETARY

The data must not be covered by any intellectual property or copyright regulations. Only limitations related to protection and security of private data are acceptable.

---



## 9. PERMANENT

Access to open data should not be limited in time. It is important that published data remain available on specific web address for as long as possible, allowing utilization of the data by a user with a link to the address.

After covering the definitions and main principles of open data, this toolkit will review compiling, processing, theoretical and practical issues related to analysis and processing of data in the following chapters, reviewing modern techniques and methods of working with data. This toolkit is based on various books, guides and articles related to this topic.<sup>3</sup>

<sup>3</sup> Following sources have been used for this guide:

*Data Journalism*. MaryJo Webster's training materials. Available at: <http://mjwebster.github.io/DataJ>

Kuang Keng & Kuek Ser. *Best Practices for Data Journalism*. Available at: <https://bit.ly/2ESVlej>

*Data Journalism Manual*. EDECA. Available at: <http://www.odcanet.org/data-journalism-manual>

*Data Journalism*. Google News Initiative. Available at: <https://bit.ly/2Zs8m7L>

Paul Bradshaw. *Finding Stories in Spreadsheets*. Available at: <https://leanpub.com/spreadsheetsstories>

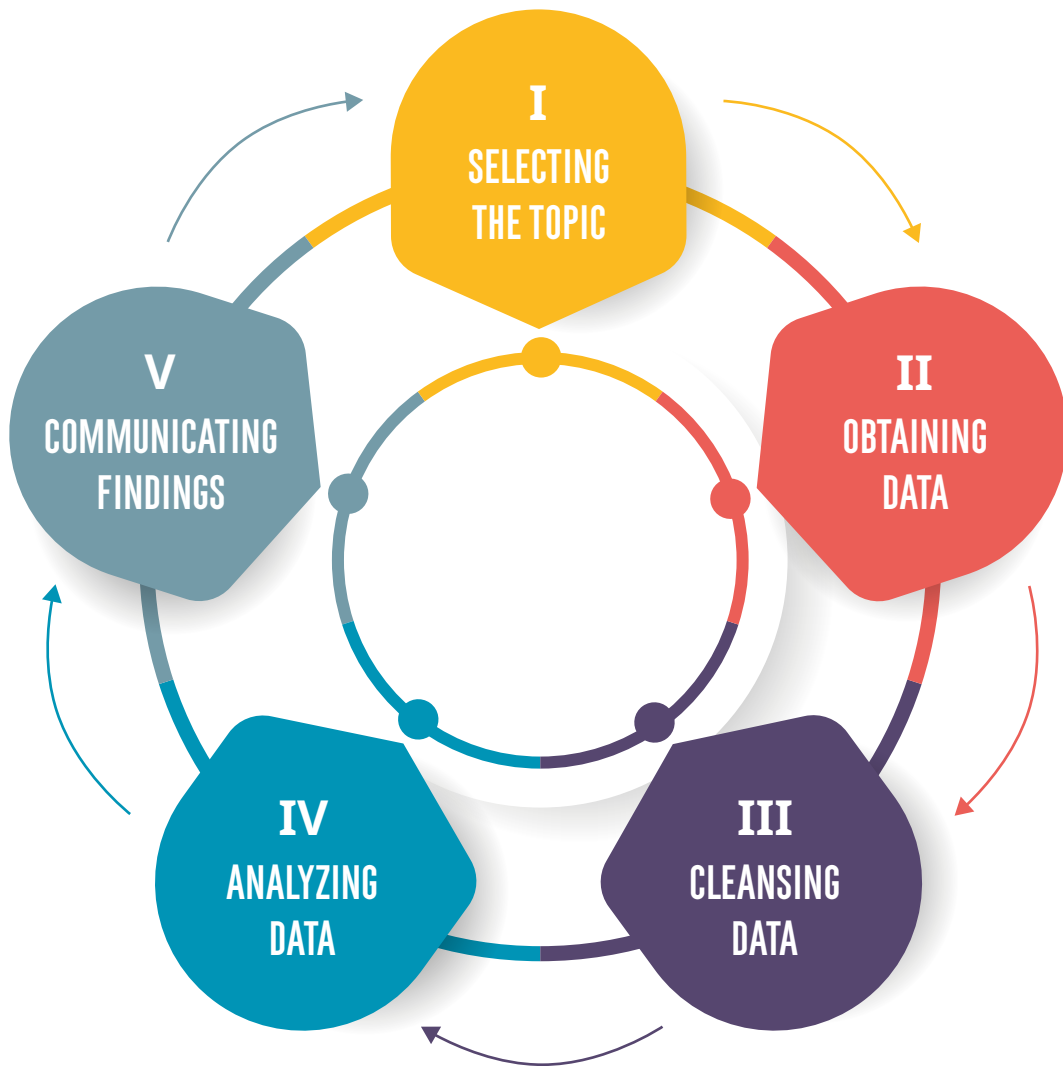
Lawrence Marzouk & Crina Boros. *Getting Started in Data Journalism*. Available at: <https://bit.ly/2QkPN1l>

Jānis Gulbis. *Data Visualization – How to Pick the Right Chart Type?* Available at: [https://eazybi.com/blog/data\\_visualization\\_and\\_chart\\_types/](https://eazybi.com/blog/data_visualization_and_chart_types/)

Anna Vital. *How To Think Visually Using Visual Analogies – Infographic*. Available at: <https://bit.ly/2SrjuAz>

Jami Oetting. *Data Visualization 101: How to Choose the Right Chart or Graph for Your Data*. Available at: <https://bit.ly/2EVEInC>

Writing an analytical article using data processing consists of five stages:



This toolkit reviews recommendations and main methods used on each stage.



# SELECTING THE TOPIC



At the initial preparation stage of an analytical article, it is crucial to understand how relevant this topic is for the audience and whether you will be able to access necessary information. However, the most important aspect is **creativity and innovative approach to the subject**. At the initial stage, you might not have a lot of information, however, as you study the issue, you might be able to make connections between issues and data that didn't seem relevant to each other before. This preparation stage of an analytical and investigative article might take longer than you think.

## WHEN CHOOSING A SUBJECT OR A TOPIC, CONSIDER THE FOLLOWING RECOMMENDATIONS:



### INVESTIGATIVE ARTICLES








If you want to find new circumstances, you might have to process various datasets, connect them, and determine a cause.



### IDENTIFYING CAUSES OF A PROBLEM

The article might concern a publicly articulated problem, however, if you plan to study the issue from a different angle and identify causes of the problem, this will allow your reader to see the topic in a new light. This requires fundamental study of the issue.

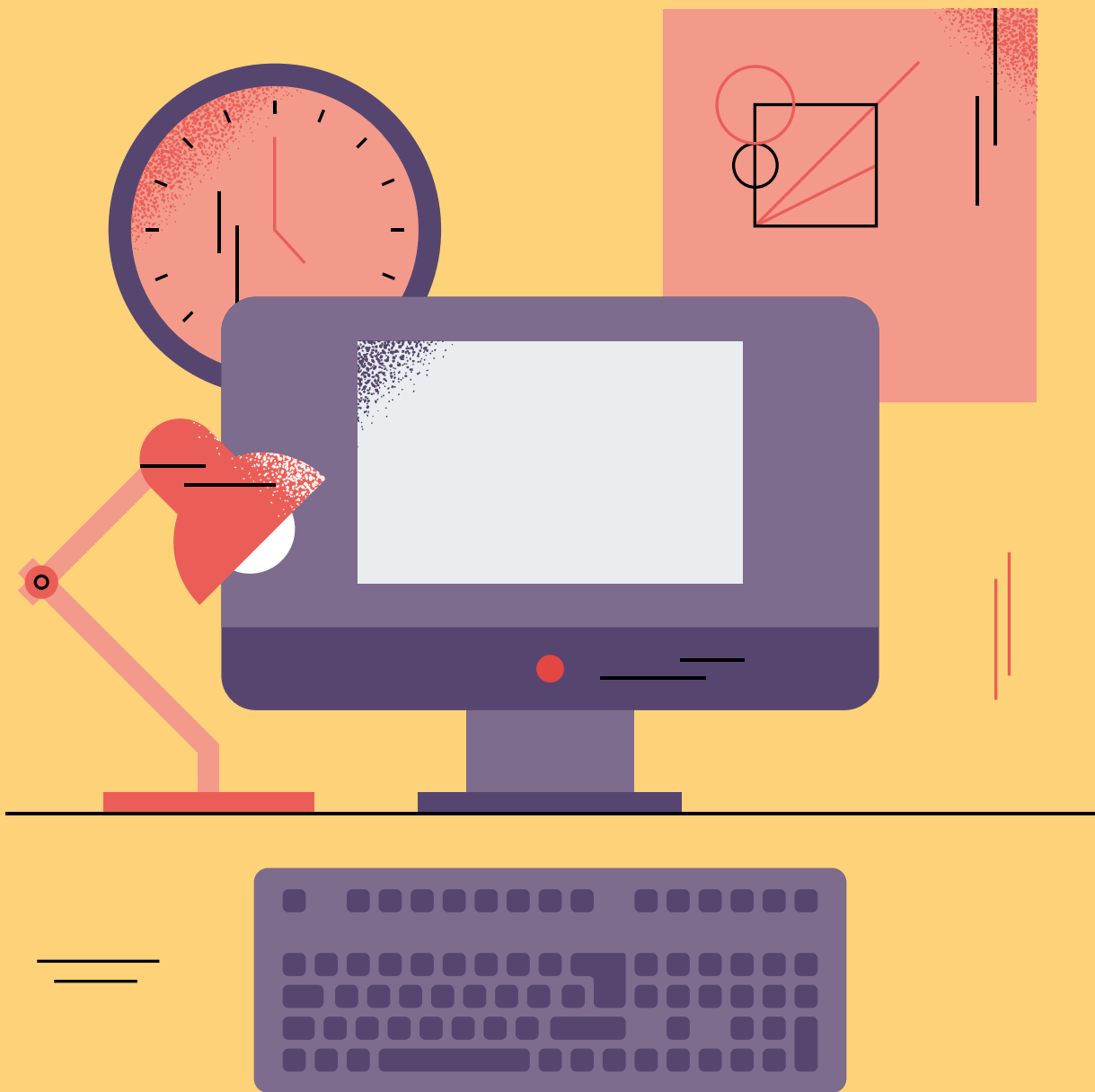
A topic of an analytical article is identified based on individual circumstances, however, several general sources of inspiration can be mentioned:

-  An interesting issue or information provided to you by a **primary source**;
-  An interesting trend identified in a **study and an analysis of official data**;
-  **Personal interest** in the issue;
-  **Follow-up on an already-written article or return to the topic**;
-  **Response** to a report or study published by another person or organization;
-  **Suspicious circumstances** or violations circulating or published on various mediums (e.g. social network, media, or personal conversation), that require additional in-depth study;
-  **Evergreen topics** – are topics that can be reviewed periodically or based on political context might become particularly relevant. For example, election campaigns and financing of political parties, crime statistics, state budget, key economic indicators of the country, healthcare topics..





When selecting a topic, it's important to have a hypothesis/thesis or a main question around which you will build the article. After you formulate your hypothesis/research question, you should create a research plan, that includes methods of information gathering, ways of reaching out sources, criteria for confirmation of hypothesis/answering the question and detailed timeline. Additionally, if you are working on an investigative article, you should **map main actors** of your research topic and write down **methods** of acquiring necessary information.

When preparing the article, you should constantly critically approach the assumptions of your hypothesis and the extent to which the analysis of the compiled information for the study of the topic will confirm the hypothesis.

# COMPILING DATA



Official data can be compiled **several ways**, specifically:

-  Downloading from websites and platforms of the public entity;
-  Requesting public information;
-  Retrieving it from various online sources;
-  Conducting an interview.

When compiling information, consider the fact that the information source might be primary or secondary.

### PRIMARY SOURCE:



a document or a person that is a direct evidence of the fact. The primary source might be an official statement, a journalistic report from the location, autobiographical work, statistical data processed by a public institution. It is a primary, original source of information on an issue.

### SECONDARY SOURCE:



any other type of published information, including reports and studies. Secondary sources are particularly significant as they allow for contextualization and communication of the issue, as well as directing us towards important primary sources. The facts presented in secondary sources must be checked for validity and accuracy.

Below we review each method of data collection and provide recommendations:

# GOVERNMENT AND CIVIC DATA PLATFORMS IN GEORGIA

In recent years, in the process of public administration reforms in Georgia several important official platforms were created. Each branch of the government publishes important data about their work on these platforms. Below you will find a list of these platforms, along with appropriate links and descriptions of data on these platforms.<sup>4</sup>

PLATFORM NAME AND LINK	RESPONSIBLE ENTITY/ ORGANIZATION	DESCRIPTION OF THE PUBLISHED DATA
<b>Unified Electronic System of Public Procurements</b> <sup>5</sup>	State Procurement Agency	Data about electronic procurements is available for every stage from its announcement to its completion – details of announced tender, documentation, tender applications and results, including subsequent contracts.

<sup>4</sup> For detailed recommendations on data published on these platforms, see IDFI Guide for Data Journalists: <https://datalab.ge/ge/toolkittext/toolkit/3/>

<sup>5</sup> <https://tenders.procurement.gov.ge/>

<b>Business Registry<sup>6</sup></b>	National Agency of Public Registry	Official data on commercial and non-commercial legal entities registered in the public registry: identification code, name, legal form, registration date, status, updated company data, including leadership, executive body, partners and their identification numbers, their names and shares in the company, other information related to property.
<b>Company Info<sup>7</sup></b>	Transparency International - Georgia	A platform created based on the Business Registry of the National Agency of Public Registry with simplified search function using a company name or identification number of company leaders. The search can be performed using company name or the name of an individual.
<b>Political Party Donations<sup>8</sup></b>	State Audit Service	The section of political party donations includes donations made to political parties by legal entities and physical persons since 2012. The data includes: name and last name, name of the political party, identification code or identification number, amount donated, date, legal form (physical, legal or both), type of donation.

<sup>6</sup> <https://napr.gov.ge/dziebakomp>

<sup>7</sup> <https://www.companyinfo.ge/>

<sup>8</sup> <https://monitoring.sao.ge/>

<b>Political Party Donations<sup>9</sup></b> (connected to other platforms)	Transparency International - Georgia	The platform allows a user to find information about business interests of donors. The platform connects information from various entities: political donations made since 2012, state procurement platform and business registry.
<b>Tax Lien/ Mortgage Register<sup>10</sup></b>	National Agency of Public Registry	The register displays whether tax lien, mortgage or any other financial sanction has been imposed on an individual or a company.
<b>Debtors Register<sup>11</sup></b>	National Agency of Public Registry	This is a register of companies and individuals who owe a debt to the state. Includes their name, identification code/personal number, address, and the registration number of the register.
<b>Voter List<sup>12</sup></b>	Central Election Commission	A list of voters, data on voters registered at specific addresses. This source is useful for identification of (1) family members, (2) date of birth and (3) address of a person.

<sup>9</sup> <https://www.transparency.ge/politicaldonations/>

<sup>10</sup> <https://naprImr.reestri.gov.ge/#/>

<sup>11</sup> <https://debt.reestri.gov.ge/main.php?s=1>

<sup>12</sup> <https://voters.cec.gov.ge/>



---

<b>Property</b>	Civil Service	Includes the following information on public officials and their family members: name, last name and address, real-estate property exceeding the value of GEL 10,000; bonds; bank accounts and deposits; money in cash (exceeding GEL 4000); business activities (company ownership); income from work; active contracts (exceeding GEL 3000), accepted gifts (exceeding GEL 500); any income or expenditure (each exceeding GEL 1500 and is not mentioned in any other section of the declaration).
<b>Declarations</b> <sup>13</sup>	Bureau	

---

<b>Budget Section</b> <sup>14</sup>	Ministry of	Includes data related to the state budget from the recent years: analytical indicators, BDD – key indicators and directions, state budget and its implementation reports, budgets of autonomous republics and municipalities, budget calendar, public debt.
of the Website	Finance of	
of the Ministry of Finance	Georgia	

---

<sup>13</sup> <https://declaration.gov.ge/>

<sup>14</sup> <https://mof.ge/4536>

<b>Budget Monitor</b> <sup>15</sup>	State Audit Service	The platform provides information on state and municipality budgets and reports on the accounts audited by the State Audit Service. The data is visualized, it can be downloaded in various formats, includes explanatory notes and data sources.
<b>Statistics Section of the National Bank</b> <sup>16</sup>	National Bank	Includes important financial data on the following: GDP and national income, prices, monetary and financial statistics, financial markets, exchange rates, payment cards market, export sector, etc.
<b>Gambling Business License Register</b> <sup>17</sup>	Revenue Service	Following information is provided on each company: identification code, name, license type, license number, validity period (start and end dates), address.
<b>Mountainous Region Business Register</b> <sup>18</sup>	Revenue Service	The register includes the following information on businesses of mountainous regions: identification code, tax payer id/name, the number of the government decree on assignment of the status and the date of the order.

<sup>15</sup> <https://budgetmonitor.ge/ka>

<sup>16</sup> <https://www.nbg.gov.ge/index.php?m=304>

<sup>17</sup> [https://www.rs.ge/Default.aspx?sec\\_id=6181&lang=1](https://www.rs.ge/Default.aspx?sec_id=6181&lang=1)

<sup>18</sup> [https://www.rs.ge/Default.aspx?sec\\_id=6342&lang=1](https://www.rs.ge/Default.aspx?sec_id=6342&lang=1)

<b>Charity Organization Register<sup>19</sup></b>	Revenue Service	Includes identification number, name of the taxpayer, the date of status assignment and/or status termination.
<b>Unified Electronic Register of Tax-Exempt Persons<sup>20</sup></b>	Revenue Service	Includes projects and organizations, including state bodies and non-commercial organizations that are tax-exempt based on various international treaties.
<b>Intellectual Property Register<sup>21</sup></b>	Revenue Service	Includes a list of registered intellectual property: name of the intellectual property object, the owner of the property or their representative, product that includes this intellectual property.
<b>Social and Healthcare Data<sup>22</sup></b>	Social Service Agency	Includes a wide range of statistical and financial data on the following topics and state programs: pension, livelihood subsidy, targeted social assistance, livelihood assistance, program for improvement of demographic conditions, assistance for permanent residents of mountainous regions, state healthcare program, etc.

<sup>19</sup> [https://www.rs.ge/Default.aspx?sec\\_id=4761&lang=1](https://www.rs.ge/Default.aspx?sec_id=4761&lang=1)

<sup>20</sup> <https://www.rs.ge/5440>

<sup>21</sup> [https://www.rs.ge/Default.aspx?sec\\_id=4867&lang=1](https://www.rs.ge/Default.aspx?sec_id=4867&lang=1)

<sup>22</sup> [http://ssa.gov.ge/index.php?lang\\_id=GEO&sec\\_id=610](http://ssa.gov.ge/index.php?lang_id=GEO&sec_id=610)

---

**Legislative Herald of Georgia<sup>23</sup>**

Legislative Herald of Georgia




Includes Georgian laws and normative acts. Including normative acts of all bodies, international treaties, constitutional legal decisions, and local government acts. A user can see codified acts and their history, including revisions at every stage.

---

**Election Results<sup>24</sup>**

Central Election Commission

The section of elections provides detailed information on all elections conducted since 2010:

-  The number of votes received by each election candidate;
-  Data according to election districts;
-  Data according to election precincts with appropriate protocols (scanned).

---

**Voting Results<sup>25</sup>**

Parliament of Georgia

Includes voting data for each legislative bill. MPs that have voted for or against a legislative bill can be found by searching their name or the name of the bill. The data includes results since 2008 and is periodically updated.

---

<sup>23</sup> <https://matsne.gov.ge/ka>

<sup>24</sup> <http://results.cec.gov.ge/>

<sup>25</sup> <https://votes.parliament.ge/ka>

<b>Parliament Monitoring Website</b> <sup>26</sup>	Transparency International Georgia	Includes a) data on legislative activity, income, and biography of MPs; b) current legislative bills and their review stages
<b>Draft Legislation</b> <sup>27</sup>	Parliament of Georgia	A user can access draft legislation and other documents (laws, resolutions, statements) with appropriate details. Along with general information, each draft legislation and a law includes explanatory notes, review stages to be cleared and any other documents at each stage of the review.
<b>General Statistical Data</b> <sup>28</sup>	National Service of Statistics	Official statistics from various public spheres are published: GDP and national revenue, prices and inflation, external trade, employment, salaries, legal statistics, healthcare and social welfare, education and culture, agriculture, environment and food safety, statistics related to state finances, gender, tourism, etc.

<sup>26</sup> <https://www.chemiparlamenti.ge/>

<sup>27</sup> <https://info.parliament.ge/#law-drafting>

<sup>28</sup> <https://www.geostat.ge/ka>

---

**Open Data Lab<sup>29</sup>**

Institute for the  
Development  
of Freedom of  
Information (IDFI)

Includes data from hundreds of central and local public entities acquired by IDFI since 2009. The databases concern such issues as: public governance and administrative expenditure, local self-governance, economy, finances, healthcare, crime statistics, social policies, education, environmental protection, transportation, and society. The platform provides access in machine-readable formats (Excel and CSV) and allows for simple visualization.

---

**Tourism Data<sup>30</sup>**

National Tourism  
Administration

Includes data on international visitors according to time periods and countries, as well as, revenue from tourism.

---

<sup>29</sup> <https://datalab.ge/>

<sup>30</sup> <https://bit.ly/2oHhSoR>
















# PROACTIVE PUBLICATION OF PUBLIC INFORMATION

August 26, 2013 [Resolution](#),<sup>31</sup> of the Georgian Government established a standard for publication of public data for government entities. The resolution defined a list of information to be published proactively by the government entities on their websites with certain periodicity. The list includes important information related to public institutions, such as:





<sup>31</sup> 26 August Resolution №219 on “Requesting Public Information through Electronic Means and Proactive Publication,” Available: <https://matsne.gov.ge/ka/document/view/2001875?publication=0>

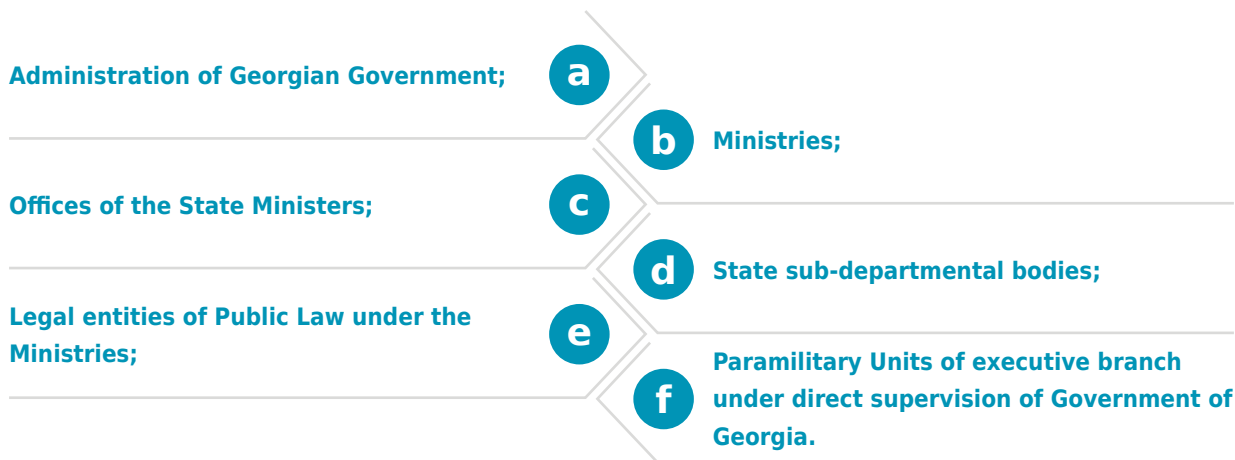
Related to work of the institution and appropriation of public funds, following categories of data can be underlined in this list:

-  An annual report about the work performed by the administrative body;
-  Information about the leadership of the body, their deputies, leaders of the structural units and territorial units (in case of a Legal Entity of Public Law – their heads and deputy heads): their names, last names, photos, biographies;
-  Information about public procurements and annual procurement plan;
-  Information regarding transfer of property;
-  Expenditure on advertisement;
-  Approved and revised budgets;
-  Information about implementation of the budget (including surplus);
-  Administrative expenditure:
  -  Salary, premium, and bonus issued to public officials (in total) and to other employees (total);
  -  Expenditure on official and business trips of officials (total) and other employees (total) (separately inside and outside of the country);
  -  Automotive vehicles on the balance with their respective model denominators;
  -  Fuel expenditure;
  -  Expenditure on technical service of vehicles (total);
  -  Real Estate property on balance;
  -  Communication expenditure (local and international calls) (total).



-  Financial aid by foreign governments, international organizations, and other state entities (grants, credits);
-  Grants issued by the institution (grantee, aim, size of the grant and transferred amount).

This resolution concerns the following public institutions:



The information defined in the resolution **must be published by public entities on their website in the special section titled Public Information.** Visit the website of the public entity of your interest. You may find interesting trends in the already published information, providing a future research topic.

## EXAMPLE:

If you find that there has been a sharp increase in bonuses, representative or other administrative expenses, appropriateness of these expenses might be of interest. After reviewing the already-published public information, you might request public information on detailed budgets. You may ask for granulated information for each year, position, country, event. You can also request information on expenses or signed contracts.

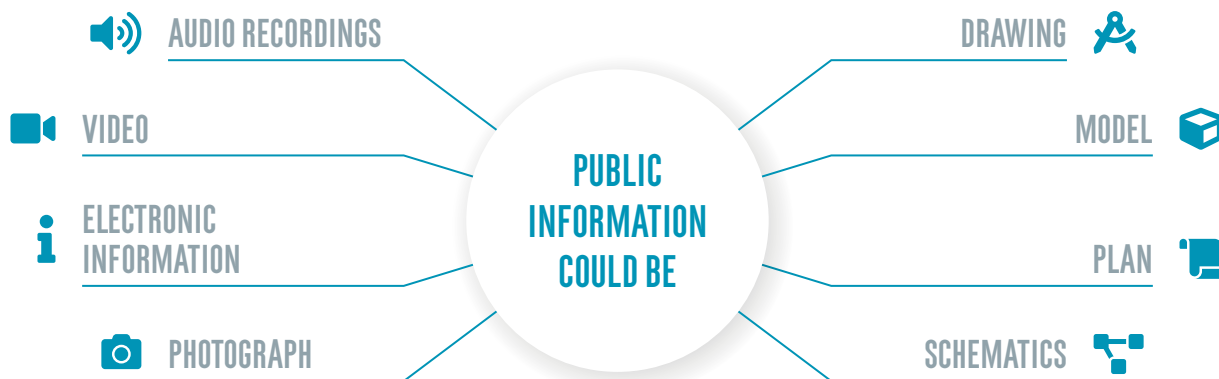
When requesting public information, you must describe your request and expected data format in as many details and as specifically as possible. If you want data to be granulated according to various categories (e.g. year, region, city, article, sex, age group, etc.), you must write so in your request! Otherwise, based on your request formulation, the public entity might not disclose detailed information. In the following section, we offer practical recommendations based on IDFI experience.

# REQUESTING PUBLIC DATABASES

At the initial stage of analytical and investigative resource discovery, you should research public official data available online. However, for the in-depth analysis you might need to request public data, and particularly public databases from various public entities. This section reviews definitions of public information and database provided by the law, including their forms, rules for requesting such information and practical recommendations that will enable you to find official data about the subject of your interest.

## WHAT IS PUBLIC INFORMATION?

Public information is an **official document**, thus is stored at a public institution, is generated by the public entity or an employee of such entity during the performance of their duties, is processed, created or sent information or information proactively published by the public entity.<sup>32</sup>



<sup>32</sup> See Chapter III on Freedom of Information of the General Administrative Code of Georgia, Available at: <https://matsne.gov.ge/document/view/16270?publication=28>

According to the Georgian legislation, every public entity is required to enter public information available to them into the public information register.<sup>33</sup> This provision designates a requirement to public entities to maintain a public register, which reflects public information and databases stored at the entity, including the name of the information, and dates of receipt, creation, processing, and publication.

Since June 2011, Georgia has a law on the **Unified State Registry of Information**.<sup>34</sup> It establishes requirements regarding registration of significant changes, expansion, combination, annulment, archiving and transfer of a register, data base or informational system. Specifically, the law provides that a public entity is required to notify LEPL Data Exchange Agency no later than 30 days after the creation or 30 days before the destruction of a database or a register.

Therefore, you can write a request to the public entity that you are researching and solicit databases maintained by them. Additionally, you can request from the Data Exchange Agency a list of databases registered in the Unified State Register of Information, select the interesting ones and send a request for additional information.

## WHO CAN REQUEST PUBLIC INFORMATION?



Anyone



You can send a request in your own name or in the name of your organization

<sup>33</sup> General Administrative Code of Georgia, Article 35

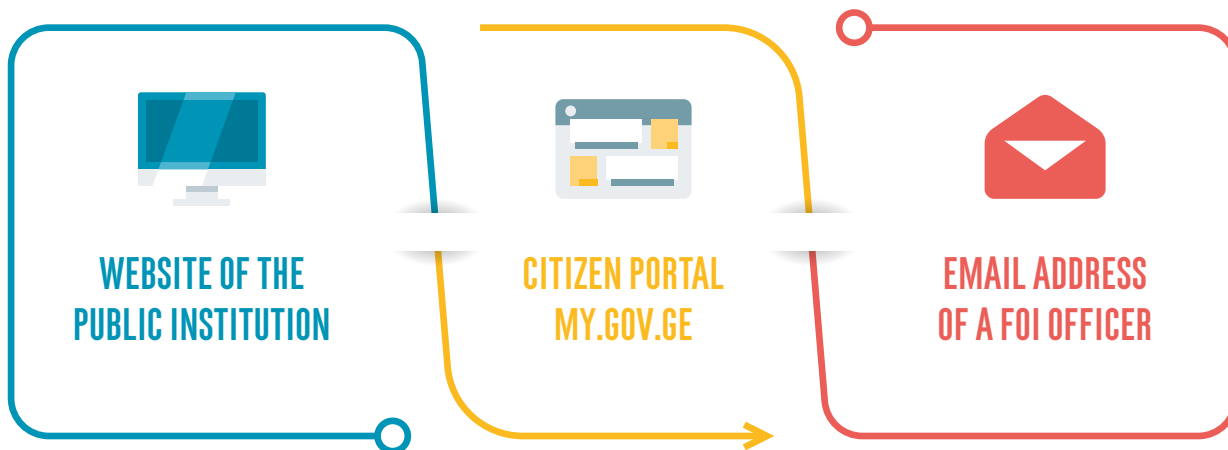
<sup>34</sup> Law of Georgia on Unified State Registry of Information, Available at:  
[http://ror.dea.gov.ge/Docs/ertiani\\_geo.pdf](http://ror.dea.gov.ge/Docs/ertiani_geo.pdf)

## NOTE:

- ✓ It is not required to mention a reason for your request in the letter.
- ✗ You can choose the form in which you would like to receive public information (electronic, a copy of a document, study it on the location), you can specify this in your request.
- 📄 When asking for a database or a register, it is recommended to choose an electronic form.

# ELECTRONIC REQUEST OF PUBLIC INFORMATION

You can request public information electronically using one of three methods:



## WEBSITE OF THE PUBLIC INSTITUTION

On the website of a public institution, you can find a section **Request Public Information**, enter your details (name, last name, personal number, email address, telephone number and address), fill out the request text, confirm it, and the request letter will be sent electronically to the appropriate public institution. However, not all public entities have implemented this model, and public information cannot be requested using their websites.

## CITIZEN PORTAL MY.GOV.GE

MY.GOV.GE is another way to request public information. For this, you need to register on the portal using the ID card reader. If you do not have an ID card reader, visit your local Public Service Hall, where you will be provided with an appropriate password. Using the password, you can register on the portal. You can then select the public entity and send them your request. You will receive a confirmation at your email address.

## EMAIL ADDRESS OF A FOI OFFICER

Each public entity is required to have a person responsible for public information (so called FOI Officer). Their contact information (name, last name, and email address) must be published on the official website of the public entity, in the section of public information.

The contact information is also available in the IDFI database of persons responsible for access and proactive publication of public information, which is periodically updated by IDFI.<sup>35</sup>

After determining the contact information, send a request for public information to the email address of the public official. **Always request confirmation of receipt of the letter!**



When you request information electronically, **you will receive requested information on your email address.**

<sup>35</sup> FOI officer database: [https://idfi.ge/ge/database\\_of\\_foi\\_officers\\_in\\_georgia](https://idfi.ge/ge/database_of_foi_officers_in_georgia)

# FINDING INFORMATION USING SEARCH ENGINES



[www.google.com/advanced\\_search](http://www.google.com/advanced_search)






The search engine has several sections that will allow you to conduct an effective search for the information you need. The sections are reviewed bellow:

all these words	Similar to standard Google search function, you can enter your search query here.
this exact word or phrase	Putting a word in ("---") will tell the search engine to search for the exact order of those words.
any of these words	Type OR between all the words you want and the engine will display results that include any of those words.



none of these words	Allows you to remove from search results any results that include this word. Put a minus symbol (-) in front of the word you want to exclude (e.g. -budget).
numbers ranging from	If you are searching for data according to years, amount, or any other numerical value, you can put two consecutive dots between the values. Search engine will treat this as a range. You can also specify the unit of measurement (e.g. kg).

You can further narrow down your search using the following parameters:

-  **language** - select the language for your search result scope;
-  **region/country** - **select a region or country** where the information was published;
-  Specify when the link was last updated;
-  Select a domain ending where the information was published (e.g. gov, ge);
-  **format** - select format in which you wish to receive information (e.g., .pdf, .xls, .ppt, .doc, .csv).

# CHECKING THE VALIDITY OF DATA



Often a database includes complex data and before starting analysis, we should ensure that we have complete information for reading and understanding the information. Specifically, we should know what the data represents, what is the unit of measurement, to what extent do we know the context and whether we need to know any specific abbreviations, etc.

A lot of databases include a special codebook or dictionary to provide explanations of each symbol and unit.

Additionally, when evaluating a raw database, you should critically approach the primary source, data categories and measurement units. Use the following questions for your guidance:

DATA SOURCE	UNDERSTANDING INDICATORS	MEASUREMENT UNITS
<b>Questions related to data</b>		
<ul style="list-style-type: none"> <li>➔ Which organization/entity collected the data?</li> <li>➔ To what extent was the primary source used in the compilation of the data?</li> <li>➔ Are definitions/explanations related to data available?</li> <li>➔ Is the data available in open/machine-readable format (Excel or another)?</li> <li>➔ When was the data created?</li> </ul>	<ul style="list-style-type: none"> <li>➔ Which indicators are included and what do they signify?</li> <li>➔ Will you be able to find definitions for indicators that are not known to you?</li> <li>➔ Which indicators are missing that could help you better understand the context?</li> </ul>	<ul style="list-style-type: none"> <li>➔ What do numbers signify?</li> <li>➔ What is the measurement unit?</li> <li>➔ Does the data use indicators from other sources?</li> </ul>

DATA SOURCE	UNDERSTANDING INDICATORS	MEASUREMENT UNITS
<b>Questions related to context</b>		
➔ Is the source of the data reliable?	➔ What would society like to know based on this data?	➔ Does the data indicator reflect the context?
➔ Is the data relevant and updated?	➔ Do indicators answer my questions?	➔ Does the indicator communicate well what I want to say?
➔ Can I find more information about the data source?	➔ Which other information would explain the data comprehensively?	➔ What type of text would communicate the indicators effectively to readers?

Additionally, the questions below will help you clarify how reliable the data source is. Specifically, consider the following circumstances when evaluating each database:



### Where does the data come from?

- ➔ Which organization published the data?
- ➔ Does the organization have experience in data collection?
- ➔ Is a report available on the website?



### Who collected data?

- ➔ Did the organization collect the data themselves or through a hired contractor?
- ➔ Are the employees of the organization qualified?
- ➔ Does any other source collect the same data and how similar are they?



### How?

- ➔ Is the data collected from the primary source or from a report?
- ➔ Was the data collected based on a survey or a census?
- ➔ Did the method of data collection change over the years?



### For what purpose was the data collected?

- ➔ For what purpose was the data collected?
- ➔ Does the publishing organization have any interest towards the issue of the data?
- ➔ Was the data collected by an independent actor?



### How complete is the data?

- ➔ Is information about data collection method available?
- ➔ Are the data limitations defined?
- ➔ Which demographic groups are included in the dataset and which are excluded?
- ➔ Does the data include rural and urban areas? Men and Women? Which social groups are included?
- ➔ When was the data collected and to what extent does it reflect the current situation?



### How exact is the data?

- ➔ Based on the dataset, how possible is generalization?
- ➔ How compatible is the data with other sources?

# DATA PROCESSING



# MAIN DATA FORMATS

You may receive data in various forms and formats. There are two main groups of data formats:

1



**Machine-readable, structured** - This type of data is generated by a computer and is organized in columns and rows. Examples include CSV (Comma-separated values), TSV (Tab-separated-values), Excel (.xls).

In case of Excel (XLS), data is organized in tables that can be read using Microsoft Excel.

In CSV (comma-separated values) and TSV (tab-separated values) data is given as a direct text separated by commas and tabs. This format presents data in encoded table where:

- ➔ Each line is a row;
- ➔ In each line values separated by commas and tabs represent columns.

2



**Non-structured** - Sometimes data is generated by a computer but it is not organized in tables. For example, PDF (Portable Document Format), Word, and bitmap photos (GIF, JPEG, PNG, BMP).

You should take **three issues** into account when working with PDF documents:



### Is it a scanned photo or not?

If a document was generated by a computer and is stored in PDF format, retrieving data from it can be easier. However, if the data was first printed and then scanned, it is more difficult to work on it.



### Is it structured?

Is the data in the document provided in tables?



### Is the document searchable?

If a document is generated by a computer, it will be searchable using keywords.

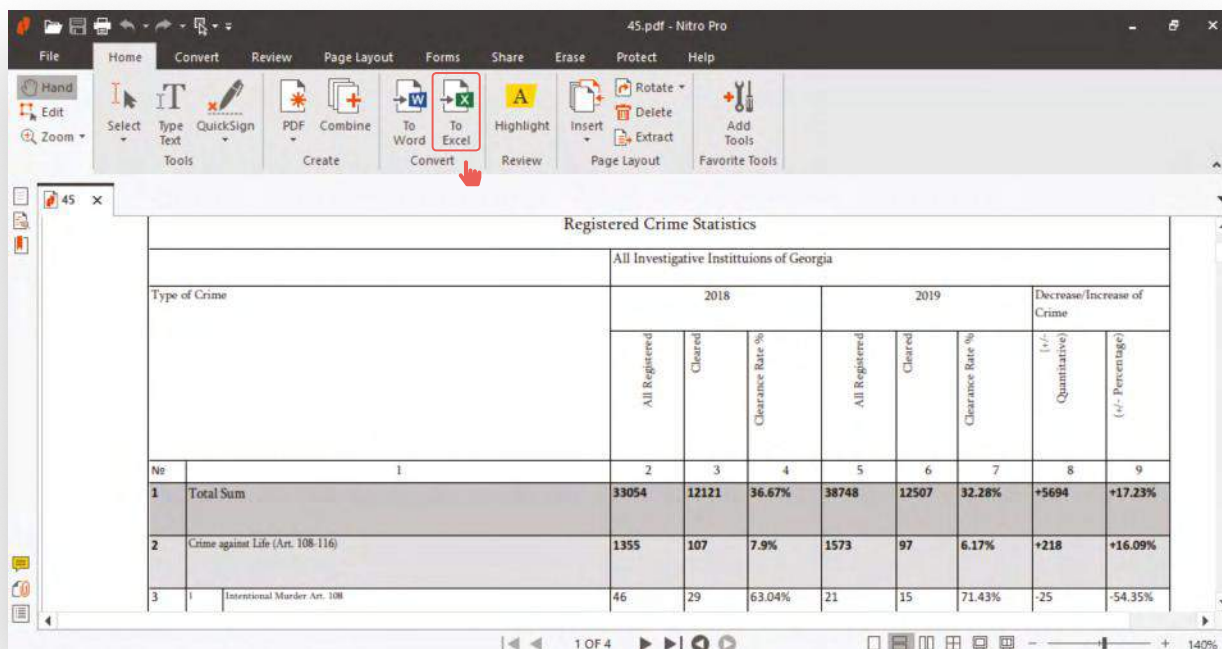


# CONVERTING PDF DOCUMENT IN EXCEL

There are several ways of converting a PDF document into Excel. Sometimes you might have to use several methods to extract maximum amount of data. Below we describe characteristics of working with each method.

First, attempt to convert the PDF document using PDF reader application on the computer. [Nitro PDF](#) is recommended for this purpose.

Open the document in this application and click the button in the top left corner, marked **to Excel**.



The screenshot shows the Nitro Pro application window with the 'Forms' tab selected. The 'To Excel' button is highlighted with a red box and a red arrow. Below the toolbar, a table titled 'Registered Crime Statistics' is visible, showing data for 'All Investigative Institutions of Georgia' for the years 2018 and 2019.

Registered Crime Statistics									
All Investigative Institutions of Georgia									
Type of Crime		2018			2019			Decrease/Increase of Crime	
		All Registered	Cleared	Clearance Rate %	All Registered	Cleared	Clearance Rate %	(+/- Quantitative)	(+/- Percentage)
Nr		2	3	4	5	6	7	8	9
1	Total Sum	33054	12121	36.67%	38748	12507	32.28%	+5694	+17.23%
2	Crime against Life (Art. 108-116)	1355	107	7.9%	1573	97	6.17%	+218	+16.09%
3	1 Intentional Murder Art. 108	46	29	63.04%	21	15	71.43%	-25	-54.35%

If you do not have Nitro installed on your computer, you can convert documents to different formats using their website:

1. Visit: [www.pdfstoexcelonline.com](https://www.pdfstoexcelonline.com)
2. Select from which format to which you wish to convert your document (in this case, from PDF to Excel)
3. Upload the document
4. Enter your email address where you wish to receive the document
5. Click on the link you receive in your email inbox and download the document.

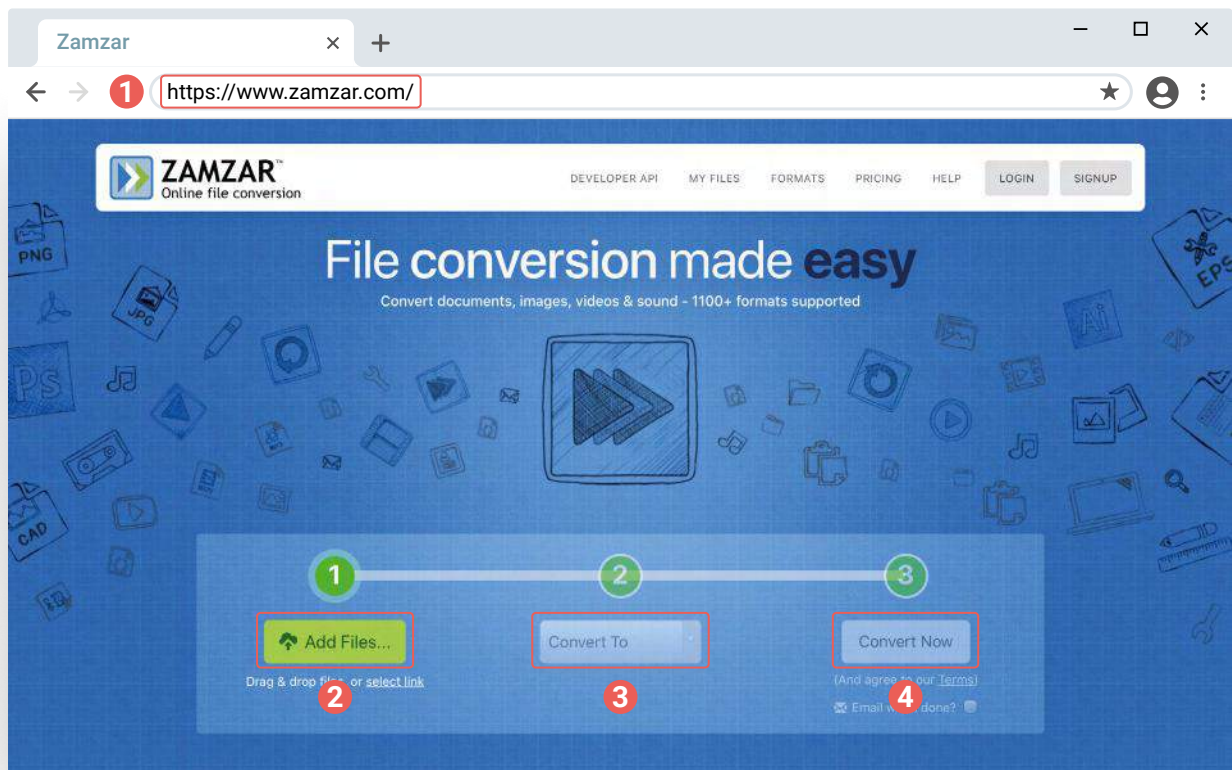
The screenshot shows a web browser window with the address bar displaying <https://www.pdfstoexcelonline.com>. The website has a header with the Nitro logo and navigation links. The main content area is titled 'PDF to Excel Converter' and contains a form with five numbered steps:

1. Select your PDF file to convert
2. Email converted file to:
3. Convert my PDF to Excel
4. Convert Now
5. (Implied step: Click on the link you receive in your email inbox and download the document.)

The form includes a dropdown menu for 'PDF' and 'Excel', a 'Select your file' button, an email input field with the placeholder 'your-email@example.com', and a 'Convert Now' button. A 'Try Free' button is also visible. The right side of the page features a 'Do More with Nitro' section with a list of features and a 'Quarterly Business Review' document preview.

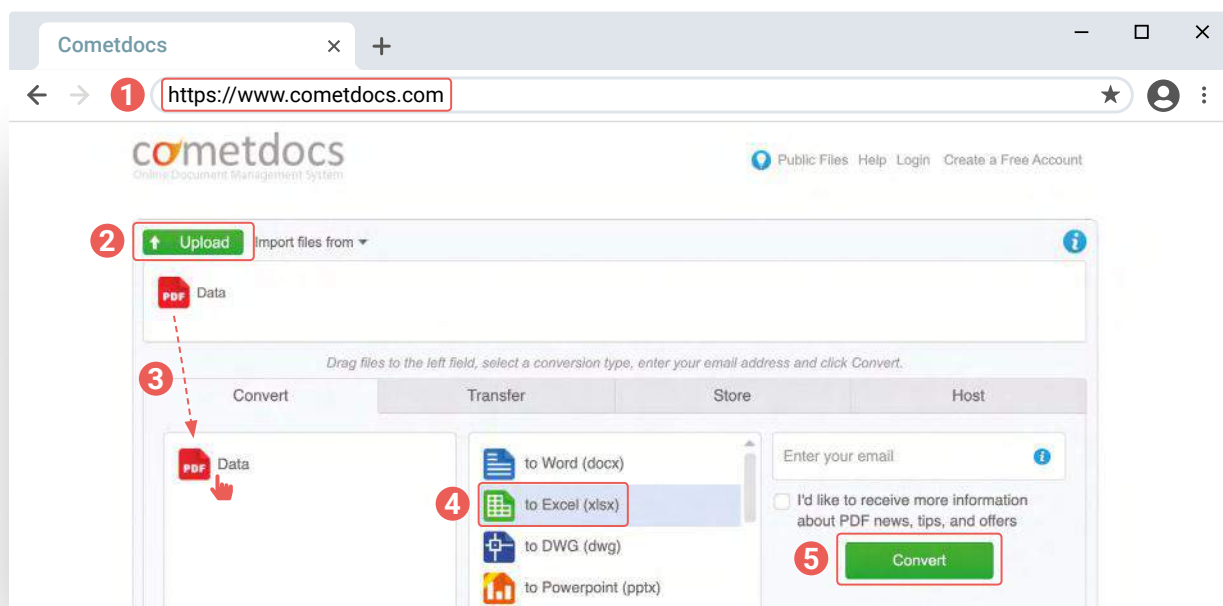
Follow these steps to use **ZAMZAR**:

1. Visit: [www.zamzar.com](https://www.zamzar.com/)
2. You will see three stages for document conversion. Click on **Add Files**
3. Select desired format (Excel)
4. Finally click **Convert Now**
5. After converting the document click **Download**.



**Cometdocs** - is particularly useful for conversion of colored tables. Follow these steps:

1. Visit: [www.cometdocs.com](https://www.cometdocs.com), create a profile and activate your account with a link you receive in your email inbox
2. Visit the website again and sign in, upload the document that you want to convert
3. Click **Convert**, and drag the file to the below, empty box
4. You will see the available formats, select one (in this case, Excel)
5. Click **Convert** and download the ready file



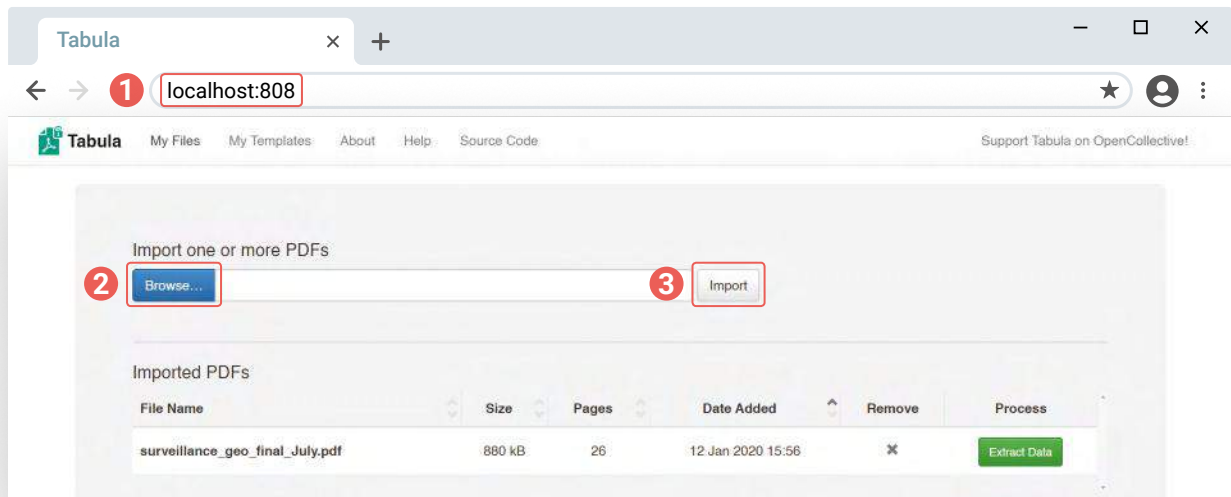
# CONVERTING PDF INTO EXCEL USING TABULA

You can use Tabula to convert tables and data in large documents into open formats. For this, you need two programs:

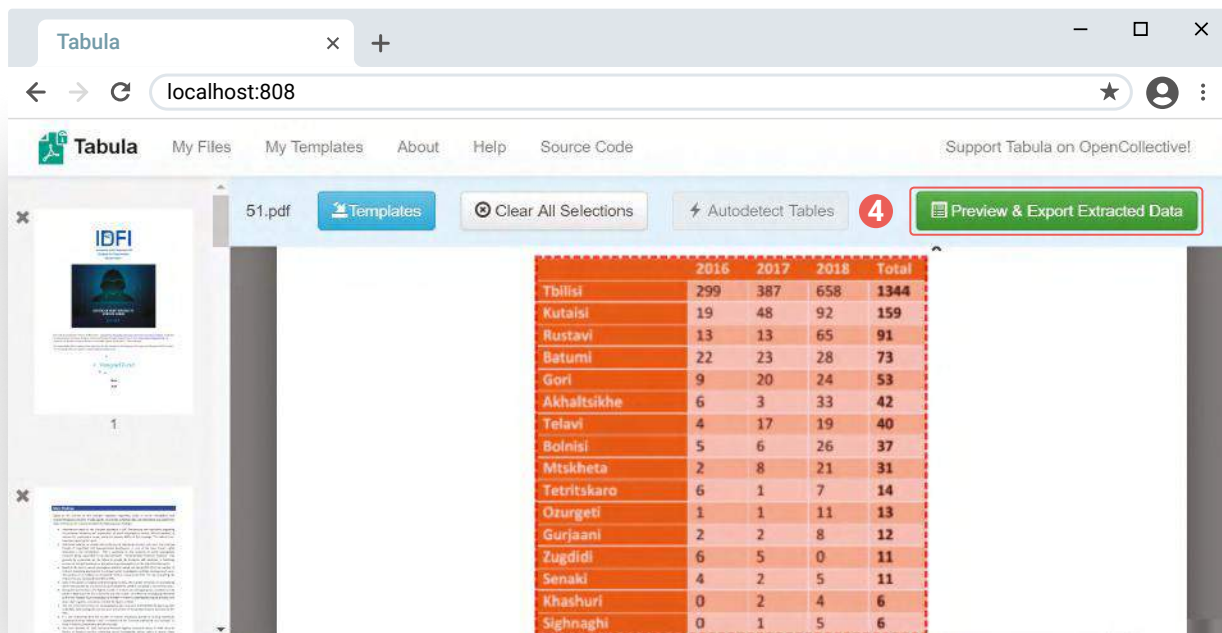
1. Java - available: [www.java.com/en/download](http://www.java.com/en/download)
2. Tabula - a) [tabula.technology/](http://tabula.technology/) - download the appropriate version of Tabula
  - ð) After downloading the zip files, a folder named “tabula” will appear.
  - ç) Enter the folder and launch **tabula.exe** (on Windows).
  - ø) After launching the program, Tabula will open in the web-browser. If the browser doesn't launch itself, open it and type in: <http://localhost:808> Tabula will open!

## How to use Tabula?

1. Open Tabula in your web browser
2. Upload your document
3. Click **Import**



4. After uploading the document, select the tables and data and click **Preview & Export Extracted Data**. You can also use **Autodetect Tables**, that will allow you to find and convert tables automatically.



The screenshot shows the Tabula web application interface. The browser address bar shows 'localhost:808'. The application has a navigation bar with 'Tabula', 'My Files', 'My Templates', 'About', 'Help', and 'Source Code'. A 'Support Tabula on OpenCollective!' link is also present. The main interface features a document preview on the left, a toolbar with '51.pdf', 'Templates', 'Clear All Selections', 'Autodetect Tables' (with a red circle containing the number 4), and 'Preview & Export Extracted Data'. The table of data is displayed on the right.

	2016	2017	2018	Total
Tbilisi	299	387	658	1344
Kutaisi	19	48	92	159
Rustavi	13	13	65	91
Batumi	22	23	28	73
Gori	9	20	24	53
Akhaltikhe	6	3	33	42
Telavi	4	17	19	40
Bolnisi	5	6	26	37
Mtskheta	2	8	21	31
Tetritskaro	6	1	7	14
Ozurgeti	1	1	11	13
Gurjaani	2	2	8	12
Zugdidi	6	5	0	11
Senaki	4	2	5	11
Khashuri	0	2	4	6
Sighnaghi	0	1	5	6

5. After finding and selecting appropriate format, click **export** and download the document. We recommend you download the table in CSV format so it can be read in Excel.



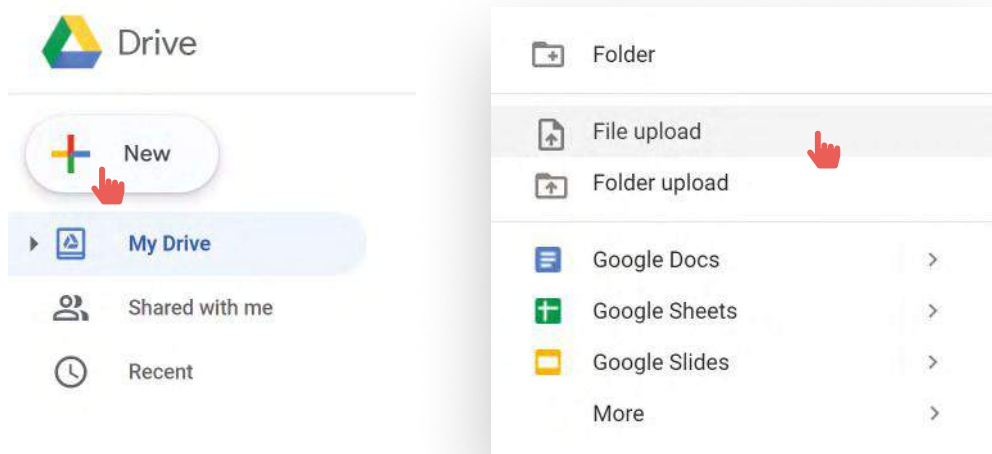
If the Georgian alphabet is not properly displayed in Excel, upload the document to Google Drive and open it as spreadsheet. After this you will be able to use and process it.

# EXTRACTING DATA FROM A TABLE IN A PHOTO

## USING GOOGLE DOCS

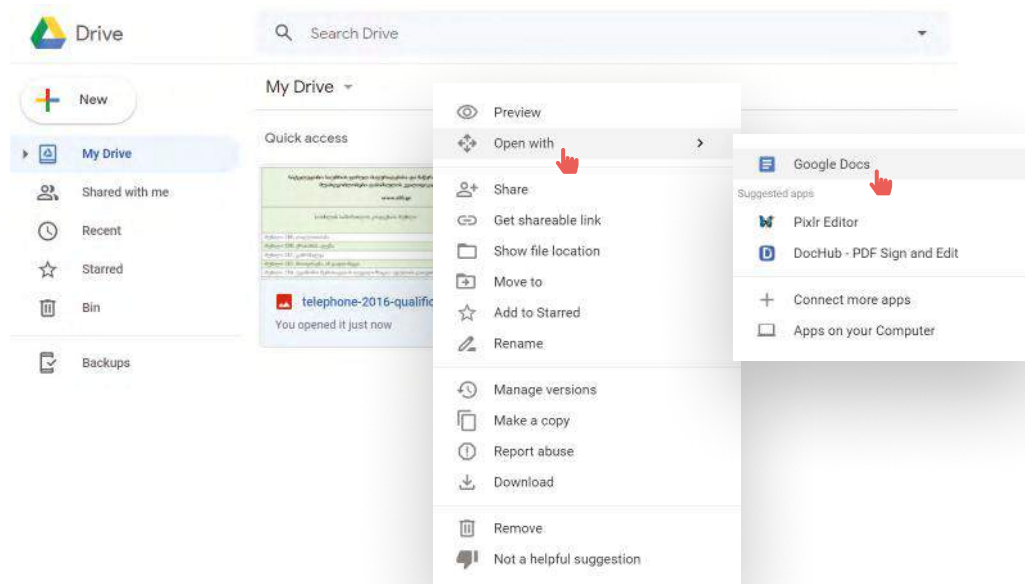
If the table that contains your data is available only as an image and you wish to extract the data, you can use Google Docs:

1. Download/save the table as a photo (for example, you can download a table from the following article: <https://bit.ly/2kcg3jY>)
2. After saving the photo, open [Google Drive](#).
3. In the top left corner of Google Drive click **New** and **File Upload** and upload the photo file.





4. After uploading the photo, find it in the Google Drive file list.
5. Open the photo with Google Docs, right click the file and select **Open with** ➔ **Google Docs**



6. A new Google Docs will open, along with the table it will include the text from the table
7. Ensure that the text is identical to the data in the table.

## Alternatives

There are better ways to extract tables from photos, however they are not free:

-  Adobe Acrobat Professional
-  Optional Character Recognition (OCR)



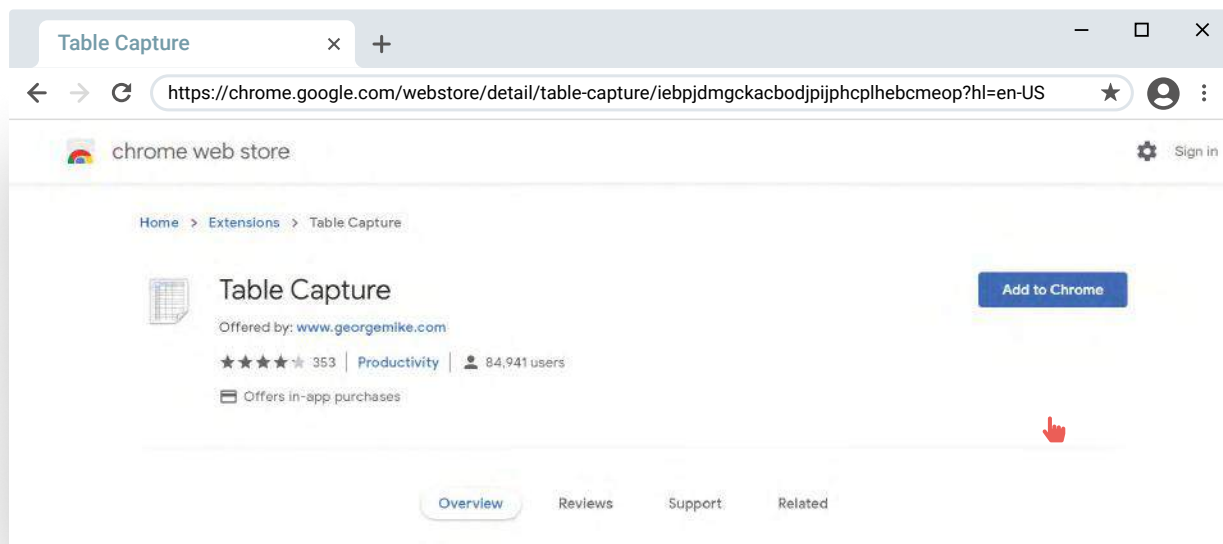
# DOWNLOADING HTML TABLES AS EXCEL

There are several web browser extensions that will enable you to download tables from a website. Let's review one:

## TABLE CAPTURE

To install the extension in Chrome:

1. Click on the link: <https://chrome.google.com/webstore/detail/table-capture/iebpjdmgckacbodjpjph-cplhebcmeop?hl=en-US>
2. Click ➞ **Add to Chrome**



3. Open the website from which you wish to download the table, e.g.: <https://bit.ly/2VMuO9Z>
4. After opening the link, click on the extension symbol next to the web-address section
5. All tables on the page will be displayed (in this case, it displays 6 tables)
6. Find the right table and click on the Google Spreadsheet symbol next to it.

2012 Georgian

https://en.wikipedia.org/wiki/2012\_Georgian\_parliamentary\_election

Region	Georgian Dream	United National Movement
Kakheti	48.06%	47.06%
Guria	58.79%	37.33%
Imereti	57.87%	37.47%
Mtskheta-Mtianeti	62.84%	32.64%
Ajara	57.53%	37.01%
Shida Kartli	51.48%	42.92%
Kvemo Kartli	38.72%	57.05%
Samegrelo-Zemo Svaneti	38.61%	55.23%
Racha-Lechkhumi and Kvemo Svaneti	46.45%	48.63%
Samtskhe-Javakheti	29.44%	67.03%
Tbilisi	65.27%	27.15%

Source: Election Portal

Reactions | edit

One day after the elections, President Saakashvili conceded that his United National Movement had been part of a coalition.<sup>[23]</sup> Georgian Dream leader Ivanishvili called on the president to resign to avoid a "sort of dual government" and to form a working group to consult with the outgoing executives over a smooth shift of power.<sup>[24]</sup> On 4 October, the Georgian Dream activists gathered in front of some District Election Commissions in constituencies where the opposition had won. Zurab Kharatishvili, complained that electoral commissioners had been intimidated. From the joint opposition list challenged the official figures and asserted that Georgian Dream had prompted his supporters to halt their protests in front of District Election Commissions.<sup>[25]</sup>

International | edit

This section may lend undue weight to certain ideas, incidents, or controversies. Please help to create a more balanced presentation. Discuss and resolve this issue before removing this message. (January 2015)

Upgrade to Pro refresh workshop display inline options · support · activity · get help now

Tables (6)

Select All Tables Selected To Google Sheets Copy Selected

(19 x 3) infobox event CSV

(10 x 3) CSV

(13 x 1) vertical-navbox... CSV

(22 x 9) wikitables CSV

(13 x 3) wikitables sort... CSV

(6 x 2) nowraplinks hli... CSV

Preview

Headers: Region,Georgian Dream,United National Movement

First row: Kakheti,48.05%,47.06%

7. A new Google Spreadsheet will open. Click **CTRL+V (or Command+V)** and the table will be pasted into the file. Now you can download the file.
8. If you buy a license for Table Capture, you will be able to download files in other formats like Excel and CSV.

Another extension: ➔ [Scraper has similar functionality.](#)

More complex software:

➔ [OpenRefine](#)

➔ [Import.io](#)

➔ [Regular Expressions](#)

➔ [Data Miner](#)

➔ [Outwit Hub](#)

➔ [ScraperWiki](#)

## CONVERTING CSV INTO EXCEL

- ➔ Download the document in CSV and open a new Excel document
- ➔ Select **Data** tab and click **From Text/CSV**
- ➔ A window for uploading CSV document will appear, find the appropriate file and click **Import**
- ➔ A window will appear displaying option of how imported data will look. Select **Delimiter** and click **Next**.
- ➔ Select **Comma** and confirm. The data will be converted into Excel format.

Text Import Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.  
If this is correct, choose Next, or choose the data type that best describes your data.

Original data type:

Choose the file type that best describes your data:

☒ Delimited - Characters such as commas or tabs separate each field.

☐ Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: 1 File origin: 866 - Cyrillic (DOS)

☐ My data has headers.

Preview of file C:\Users\Teona.Turashvili\Desktop\Demographic\_Statistics\_By\_Zip\_Code.csv

JURISDICTION NAME	COUNT PARTICIPANTS	COUNT FEMALE	PERCENT FEMALE	COUNT MALE	PERCENT MALE
00001	44	22	0.5	22	0.5
10002	35	15	0.43	20	0.57
20003	1	1	1.0	0	0.0
30004	0	0	0.0	0	0.0
40005	2	1	0.5	1	0.5
50006	6	2	0.33	4	0.67

Cancel **3** Next > Finish

Text Import Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters:

☐ Tab

☐ Semicolon

☒ Comma

☐ Space

☐ Other:

☐ Treat consecutive delimiters as one

Text qualifier:

Data preview

JURISDICTION NAME	COUNT PARTICIPANTS	COUNT FEMALE	PERCENT FEMALE	COUNT MALE	PERCENT MALE
00001	44	22	0.5	22	0.5
10002	35	15	0.43	20	0.57
20003	1	1	1.0	0	0.0
30004	0	0	0.0	0	0.0
40005	2	1	0.5	1	0.5
50006	6	2	0.33	4	0.67

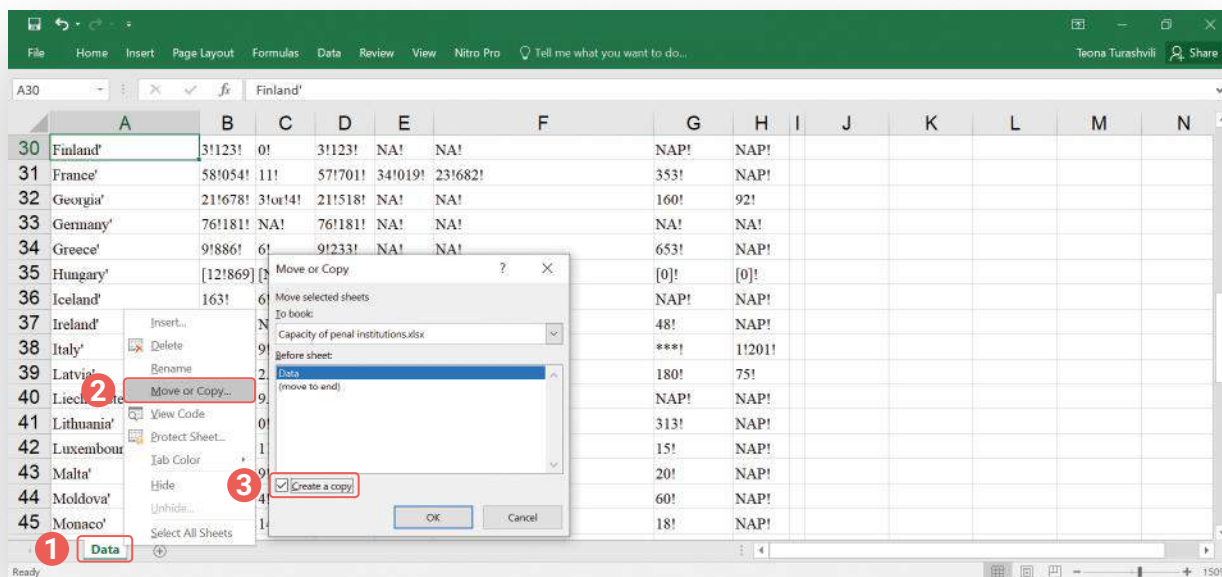
Cancel **5** Next > Finish

# DATA CLEANSING

Sometimes after a file or data has been converted into the Excel format, it requires cleaning and standardization. Excel has several useful functions for this.

Firstly, it is important **to save the original file** in Excel, so you can view and review the original source and any changes made during the cleansing. For this:

- ➔ Right click in Excel
- ➔ Select **Move or Copy** in the window that will appear
- ➔ In the new window, click **Create a copy**
- ➔ Choose a name for the new Excel page, that is the copy of the original and continue cleansing of data in this section.



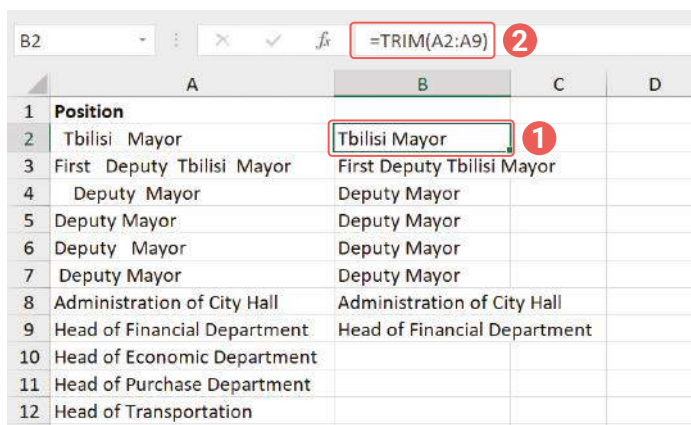
Pay attention to the following:

- ➔ Ensure that the data categories are correctly displayed as columns
- ➔ Check whether the numerical data is perceived as numbers or text
- ➔ Check if there are any mistakes
- ➔ Check if any data is missing
- ➔ Check whether during the conversion process an extra symbol has appeared with the numbers, for example a symbol instead of a space, etc.

Below we review several functions that might help you in the data cleansing process.<sup>36</sup>

## 1. REMOVING EXTRA SPACE

If the table or database includes an extra space between the words (more than one space), there is a TRIM Function. In an empty column write **=TRIM(cell/A1)** and drag the mark sign over the whole column.



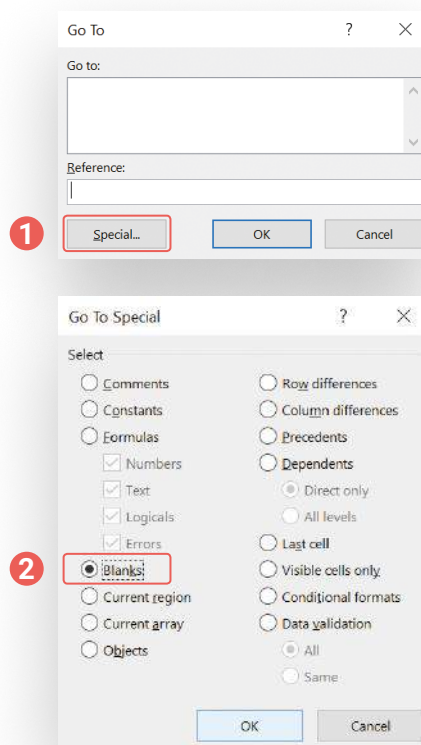
	A	B	C	D
1	Position			
2	Tbilisi Mayor	Tbilisi Mayor		
3	First Deputy Tbilisi Mayor	First Deputy Tbilisi Mayor		
4	Deputy Mayor	Deputy Mayor		
5	Deputy Mayor	Deputy Mayor		
6	Deputy Mayor	Deputy Mayor		
7	Deputy Mayor	Deputy Mayor		
8	Administration of City Hall	Administration of City Hall		
9	Head of Financial Department	Head of Financial Department		
10	Head of Economic Department			
11	Head of Purchase Department			
12	Head of Transportation			

<sup>36</sup> For more detailed information see: <https://bit.ly/2MQ0jgC>

## 2. FORMATTING AN EMPTY CELL

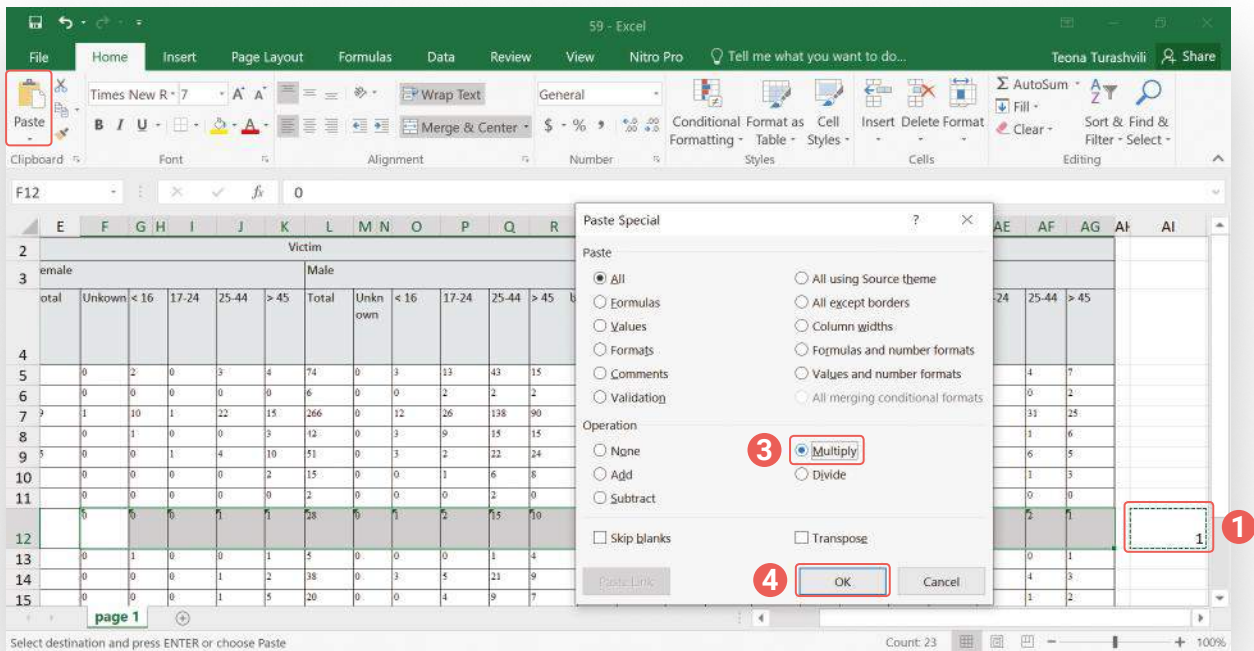
If the database includes empty cells and you want to format them the same way (for example, to write 0 or highlight with a color or write N/A), then:

- ➔ Select the data
- ➔ Click **F5 (on Windows)**, a **Go To** window will appear
- ➔ Select **Special**
- ➔ in the new window select **Blanks**
- ➔ Click **OK** and all blank cells will be selected
- ➔ Type the text or select the function you wish to implement in all empty cells and click **Control + Enter**. All blank cells will be formatted.



## 3. CONVERTING TEXT NUMBERS INTO NUMERICAL VALUES

- a** Often when converting data from PDF document into Excel, the numbers will be converted into text value. They need to be transformed into numerical values.
- ➔ Write in any cell 1, select the cell and push **Control + C**
  - ➔ Select all numbers that you want to convert into numerical values
  - ➔ In the top left corner, select **Paste** ➔ **Paste Special** (or push **Alt + E + S**)
  - ➔ In the newly-opened window (Paste Special Dialogue) select **Multiply**
  - ➔ Click **OK** and all numbers will be converted into numerical values.

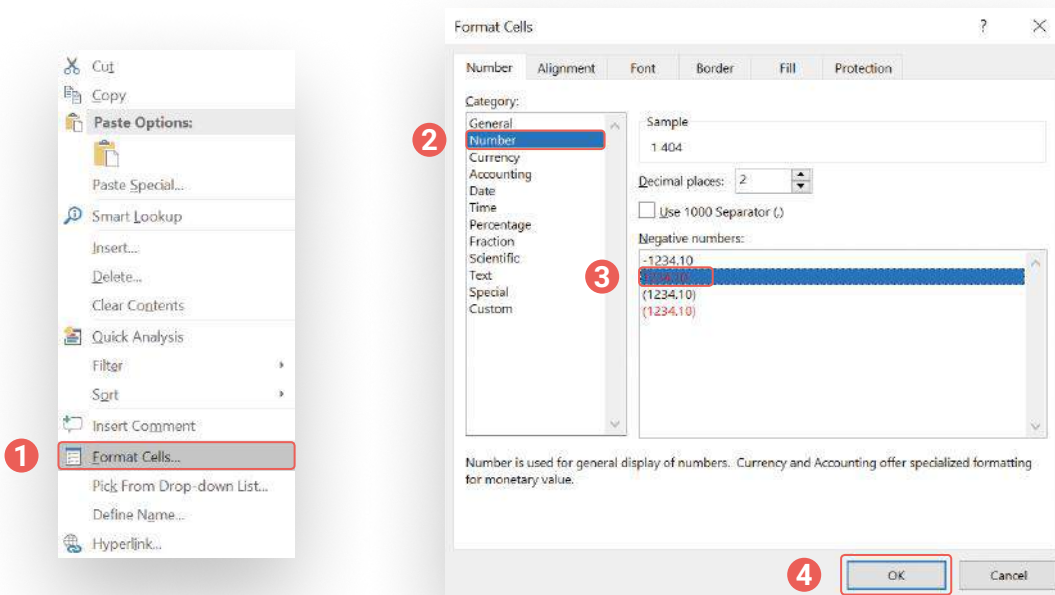


This function won't work, if one of the cells is empty, as this cell will be filled with 1.

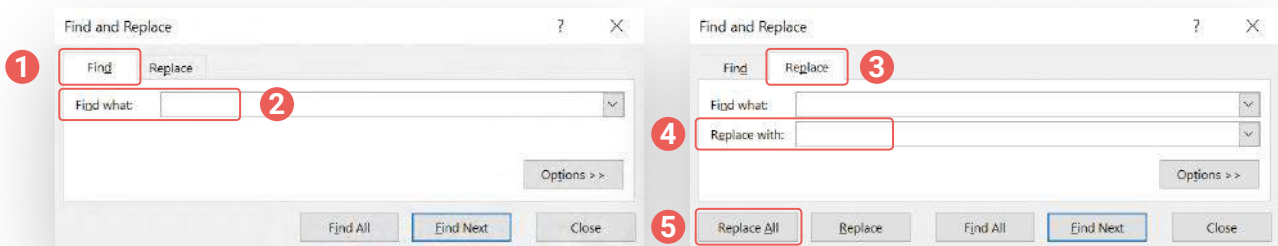
## b There is another way to convert text numbers into numerical values:

- Select all numbers, right-click and in the window select **Format Cells**
- A new window will appear, select Number and choose how they should be formatted. Click **OK**.





- ➔ After this, once again select the numbers and enter **Control + F** (or in the **Home** tab, **Find & Replace** select **Find**)
- ➔ In the Find window push **Enter**
- ➔ Leave **Replace with** section blank and click **Replace All**
- ➔ If you perform these actions correctly, the numbers will be converted into numerical values, the space between decimals will be removed and the numbers will be aligned to the right of cells.



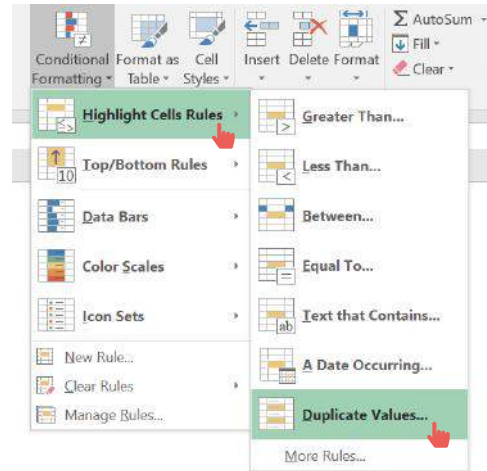


## 4. REMOVING DUPLICATES

If the data includes duplicates, you can select and/or delete them.

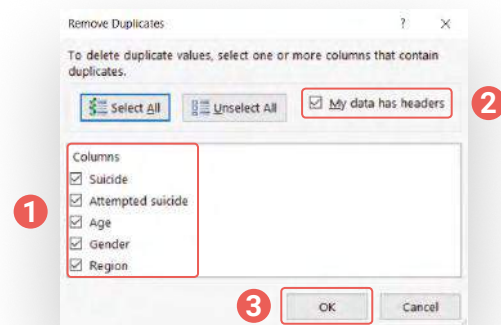
### a Selecting duplicates

- ➔ Select data
- ➔ In the Home tab select **Conditional Formatting** ➔ **Highlight Cells Rules** ➔ **Duplicate Values**.
- ➔ Select how you want numbers/cells to be displayed



### b Removing duplicates

- ➔ Select numbers and in the **Data** tab select **Remove Duplicates**
- ➔ In the new window, select the columns where you wish to remove duplicates
- ➔ If your data has appropriate names, in the window select the appropriate function (**My data has headers**) and click **OK**.

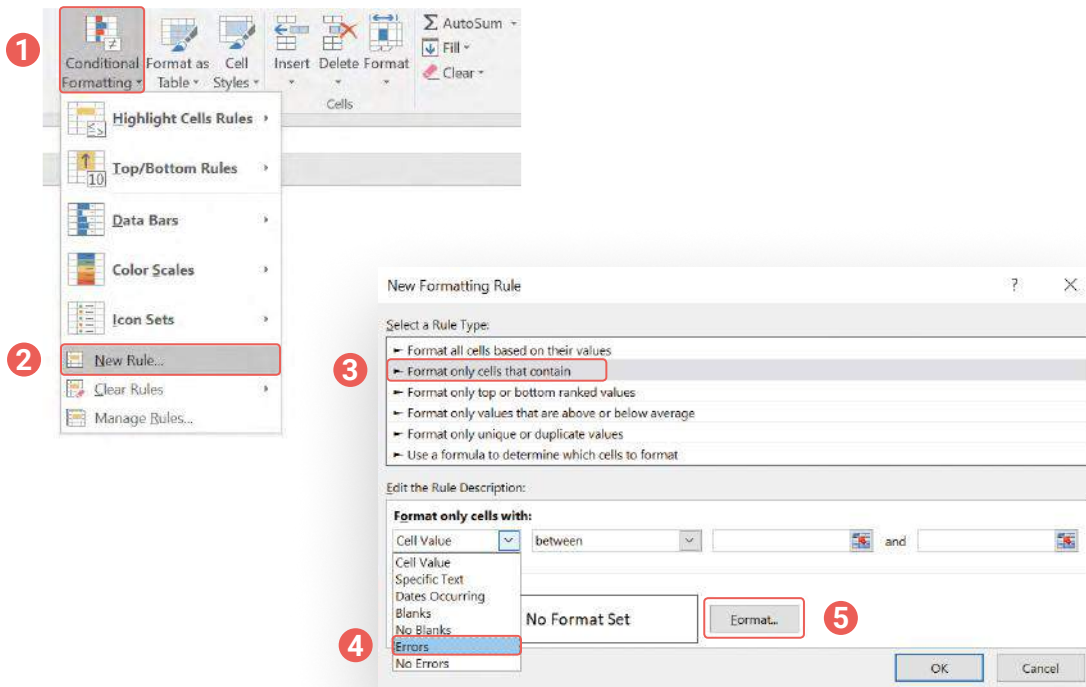


## 5. IDENTIFYING MISTAKES/FLAWS

There are two methods:

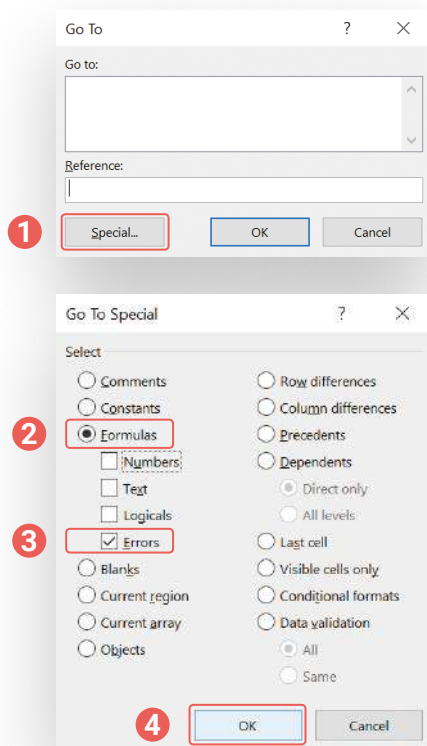
### a Using Conditional Formatting

- ➔ Select the whole database.
- ➔ In the Home tab, find **Conditional Formatting** ➔ **New Rule**
- ➔ In the window select **Format Only Cells that Contain**
- ➔ In the second section of the window from the drop-down menu select **Errors**, specifying that you want to format cells that have errors.
- ➔ Finally, by clicking Format, you can select how to format the errors.
- ➔ Click **OK** and the errors will be formatted.



## b Using Go To Special

- ➔ Select the whole database, push **F5 (on Windows)** button. **Go To** window will appear.
- ➔ Click **Special** button.
- ➔ Select **Formulas** and leave only **Errors** selected.
- ➔ Click **OK** and errors will be selected.

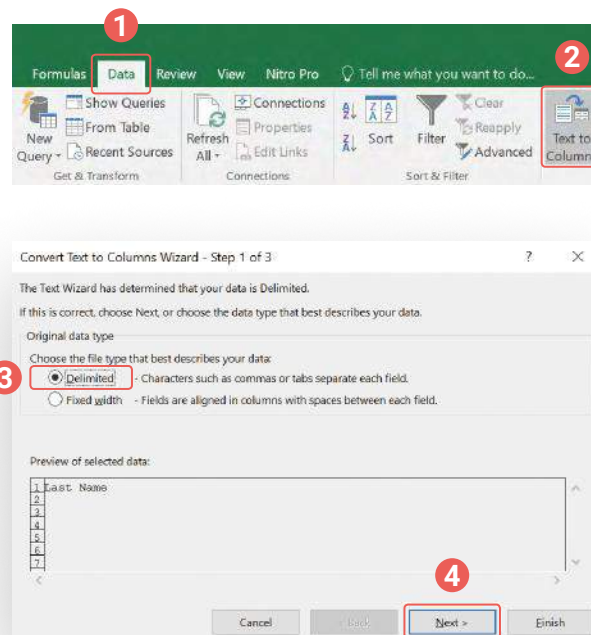


## 6. SEPARATING DATA IN COLUMNS AND MERGING COLUMNS

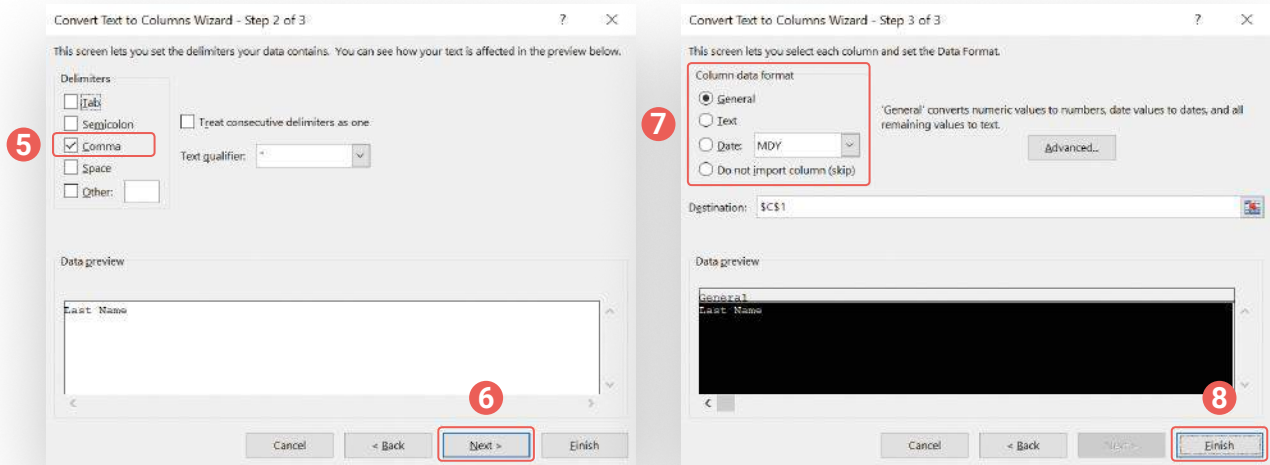
### a Separating data in columns

Sometimes data is converting from various documents into excel, and the data in one column needs to be divided into several. For grouping data into appropriate columns, follow these steps:

- ➔ ● In **Data** tab select **Text to Column** and select **Delimited**
- ➔ Click **Next** and a new window will appear



- ➔ In the new window select how the data that needs to be separated into columns is currently delimited (Comma, tab, space, or other)
- ➔ Click **Next** and a new window will appear
- ➔ Select the format of the data (date, text or general). The lower section of the window will display a preview of how the newly-formatted data will look
- ➔ Click **Finish** and check if the data is correctly separated into columns

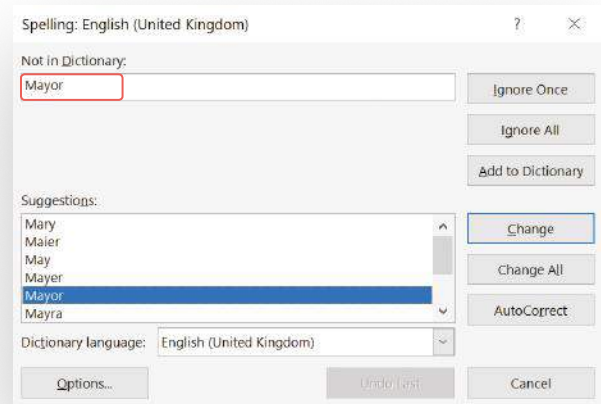


## b Merging columns

Sometimes data spread over several columns needs to be merged into a single column. For example, merging of name and last name. To implement this function, write `=CONCATEBATE (cell," "cell)` in one of the cells and drag the selection over the entire columns.

## 7. SPELLCHECK

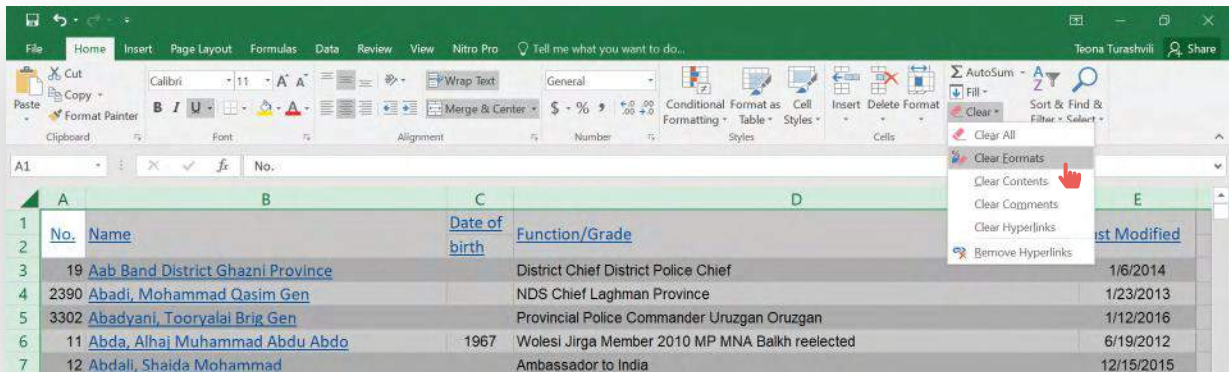
If you would like to spellcheck your database, select the data and push F7 (on Windows). Misspelled words will be identified and can be corrected.



## 8. REMOVING FORMATTING

If data converted into Excel includes formatting, e.g. words with links, and you need to remove the formatting, proceed with the following steps:

- ➔ In the Home tab, click **Clear** ➔ **Clear Formats**
- ➔ In the dropdown menu select formats, comments, links, etc. that you wish to remove.



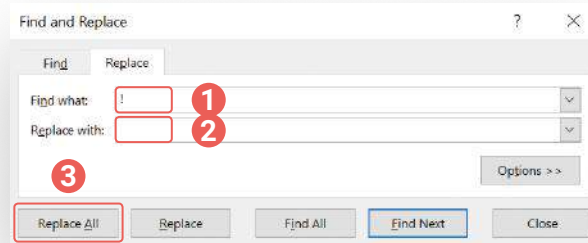
## 9. FIND AND REPLACE

### Removing extra symbols in dataset

If the numbers converted into Excel come with extra symbols, you can use the Find & Replace function to remove them.

- ➔ In the Home tab click on **Find & Select** and then **Replace** (or type **Control + F**)
- ➔ In **Find** type the symbols that you want to remove, leave **Replace** empty
- ➔ Click **Replace All** and the extra symbol will be removed.

The same function can be used to remove or replace individual (repeated) data or words

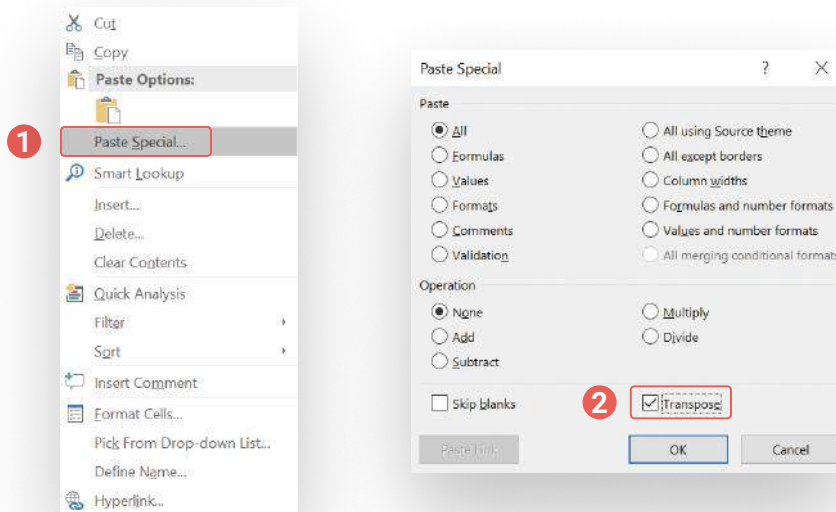


## 10. TRANSFORMING ROWS INTO COLUMNS AND VICE VERSA

If you want to transform columns into rows or vice versa, you can use one of the two ways.

### Method 1 (Static):

- ➔ Select data and make a copy using **Control + C**
- ➔ Right Click on the cell along which you wish to paste data
- ➔ In the window, select **Paste Special**
- ➔ In the new window, select **Transpose** and the data will be pasted.



## Method 2 (dynamic):

- ➔ Select a location where you want to copy your data. Try to determine the exact number of rows and cells.
- ➔ Type in formula **=TRANSPOSE**, while keeping the location selected.
- ➔ Type in the range of cells that you wish to copy or (or simply select the data and the formula will automatically be applied).
- ➔ Type in **Control + Shift + Enter** (not just enter **Enter**)
- ➔ If you haven't selected all cells, simply drag the right corner of the last cell onto the next one. If the data is not properly displayed, go back to the formula and push **Control + Shift + Enter** one more time.



The screenshot shows an Excel spreadsheet with a table of salaries. The formula bar at the top displays `=TRANSPOSE(A1:D12)`. The table data is as follows:

Position	Salary	Supplement	Total
Tbilisi Mayor	38160	0	38160
First Deputy Tbilisi Mayor	75600	0	75600
Deputy Mayor	73800	0	73800
Deputy Mayor	73800	0	73800
Deputy Mayor	73800	0	73800
Deputy Mayor	73800	0	73800
Administration of City Hall	73800	0	73800
Head of Financial Department	73800	0	73800
Head of Economic Department	73800	0	73800
Head of Purchase Department	73800	0	73800
Head of Transportation	73800	0	73800

Below the table, in cell A14, the formula `=TRANSPOSE(A1:D12)` is entered, which will dynamically transform the table data into a single row.

As this transformation is dynamic, any changes to the initial data will be automatically reflected in the transformed set.

## 11. FORMATTING TEXT/WORDS

If the text is using Latin alphabet and you want to transform the first letters to uppercase, use Lower/Upper/Proper functions. In the empty column type **=LOWER** (cell to be transformed/A1) and drag it across the whole column. For example, the functions will transform the words White House in the following manner:

➔ LOWER - white house ➔ UPPER - WHITE HOUSE ➔ Proper - White House

When processing data, it is important to know what kind of expectations we can have towards the information. The next chapter reviews what role data plays in preparing analytical articles.



# HOW TO FORM A STORY FROM DATA?



Before starting to work on the analysis of the material, it is important how we choose the focus or the main topic of the article. The data can be very helpful in this. Reading the data and identifying trends is particularly important for this. Following **functions of data** could help us better define their purpose for our article:

---



### DATA ALLOWS US TO CREATE/TEST A HYPOTHESIS

The article/story has a hypothesis/discovery that is confirmed or reinforced by data. Data allows us to find what caused specific events.

---



### DATA DEMONSTRATES TRENDS AND DEVIATIONS

Interesting narratives/facts are often found when we compare data. You can compare data based on different values – years, cities, age groups, regions, public institutions, and other characteristics. It is also interesting to compare subjects with the highest and lowest data points.

---



## DATA SHOWS OUTLIERS

A subject with characteristics significantly differentiating from others in a large dataset is particularly interesting – in a positive or a negative sense. If you identify such an outlier, **it's important that you study in depth why it deviates so much from others and what could be the reasons. You might make an interesting discovery.**



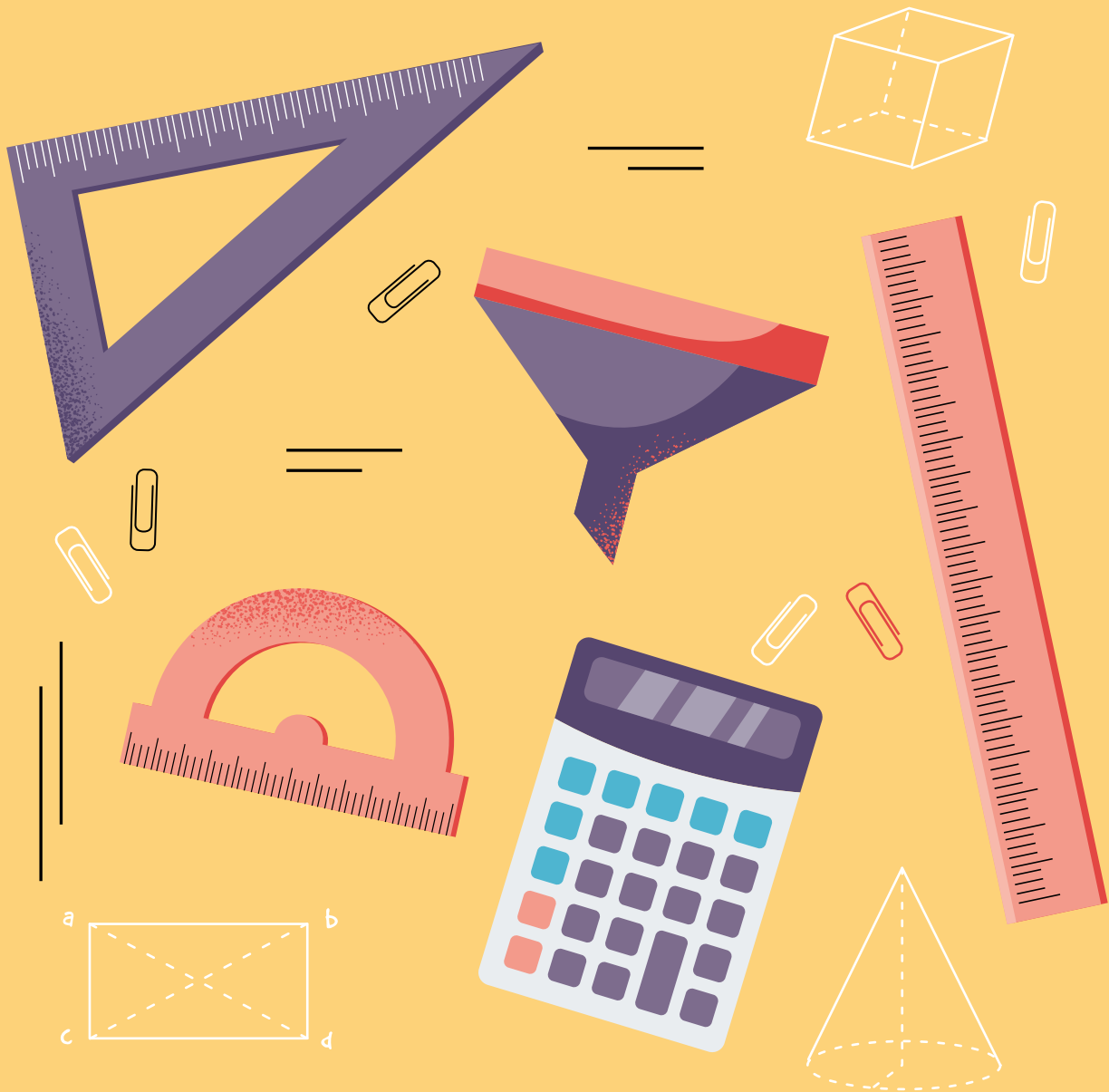
## DATA SHOWS HIDDEN CONNECTIONS

Cross-checking data in various databases and validating them, often uncover hidden circumstances. For example, undeclared property or income of politicians, public procurement tenders won by companies of their relatives, etc.

---

Using data to its fullest potential requires understanding of particularities of working with data and its analysis. The following chapter concerns this topic.

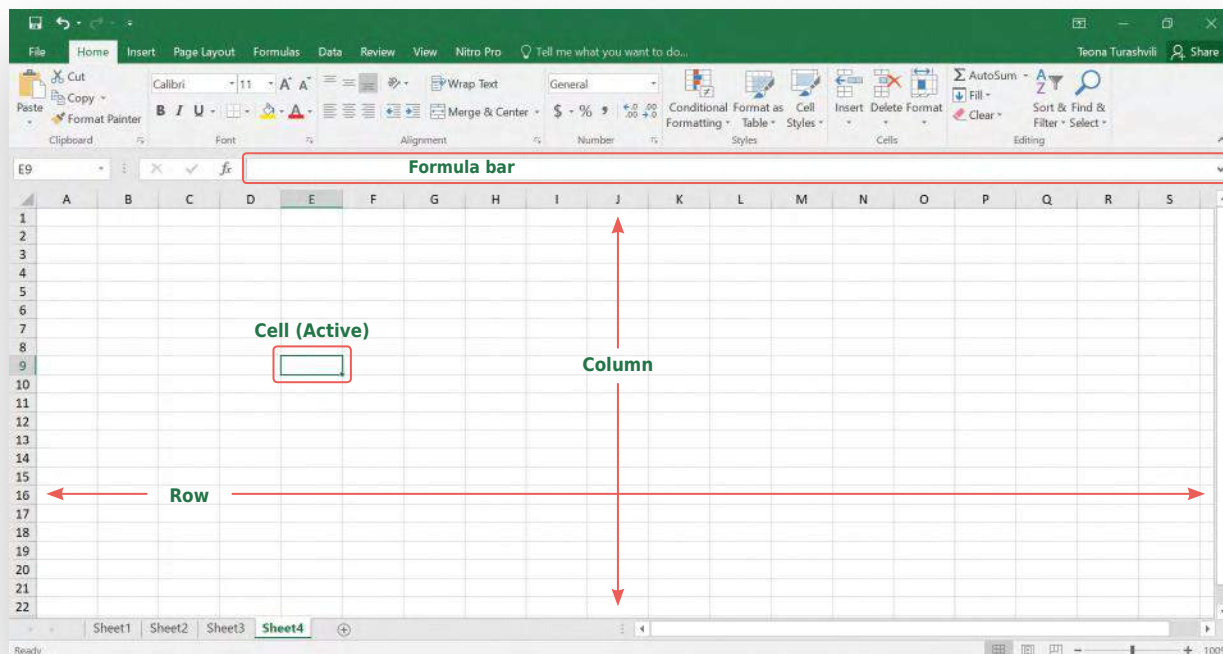
# DATA ANALYSIS



After grouping, cleansing, and standardization of data comes an important step of **data analysis**. Excel functions are very helpful in this process. Here we review several such functions and give you some recommendations on how to best utilize them.

## GLOSSARY:

- ➔ **Column** is a vertical set and often represents different categories of data.
- ➔ **Row** is a horizontal set and represents individual entries/numbers for each category (year, months, etc.).
- ➔ **Cell** is a location of an individual entry and is represented with coordinates of column and row (e.g. B5)
- ➔ **Formula bar** is located in the menu line, you can view and change the formula for each cell here.



In Excel documents, movements of the pointer (mouse) are represented graphically differently for different functions:



Is used to select a cell

---



Denotes a possibility to enter information/number in a specific cell

---



Is used for formula duplication/copying from one cell into another or filling in date or day set

---



Is used to adjust the width of a cell, increasing or decreasing it. A rotated version of the symbol allows for adjusting row height

---



Is used to select an entire column or row

---

## BASIC EXCEL FUNCTIONS:

## PERCENTAGE CHANGE

Subtract initial number from the most recent one and divide it by the initial value (e.g.,  $= (C2-B2)/B2$ ). To transform the subsequent value into the proper format, right click on the cell and select **Format Cell**, in the window the **Number** section allows you to select a format (in this case percent).

	A	B	C	D	E	F	G	H	I
1	Public Institution	2011	2012	Percentage Change					
2	Ministry of Defense	7,572	11,976	=C2-B2/B2					
3	Ministry of Foreign Affairs	24,428	28,776						
4	Ministry of Finance	1,979	1,695						

After calculating the percent, you can copy the formula to the rest of the column. To do this, select the cell, in the right corner a plus (+) sign will appear, click on it and drag across the column.

## SUM

If you have a long column or a row with a lot of values, it is more efficient to use the **SUM** function (=SUM(B2:B21))

	A	B	C	D	E	F	G
19	Sport Federation 18	491643.68					
20	Sport Federation 19	441575.94					
21	Sport Federation 20	422281					
22	<b>Total</b>	<b>=sum(B2:B21)</b>					

However, if you want to add only several values or the values from different rows or columns, you can use addition formula by selecting the cells and adding a plus (+) sign between the cell coordinates (=B3+B7+B13).

## MINIMUM AND MAXIMUM VALUES

If you want to find the highest and lowest values in a large dataset, you can use the following formulas:

- ➡ =MAX(B2:B21) - will display the highest value in column B, between rows 2 and 21;
- ➡ =MIN(B2:B21) - will display the lowest value in column B, between rows 2 and 21.

B24		=max(B2:B21)								
	A	B	C	D	E	F	G	H	I	
16	Sport Federation 15	595423.17								
17	Sport Federation 16	584980.4								
18	Sport Federation 17	524491.09								
19	Sport Federation 18	491643.68								
20	Sport Federation 19	441575.94								
21	Sport Federation 20	422281								
22	Total	44288784.93								
23										
24	Maximum	=max(B2:B21)								

PERCENT OF TOTAL

This function allows to calculate what percentage of the total value does an individual entry represent. For the calculation you need a sum, divide the individual value by the total and format the result as a percent. For example, =B21/B22 and then select percent.

	A	B	C	D	E	F	G
16	Sport Federation 15	595423.17					
17	Sport Federation 16	584980.4					
18	Sport Federation 17	524491.09					
19	Sport Federation 18	491643.68					
20	Sport Federation 19	441575.94					
21	Sport Federation 20	422281	=B21/B22				
22	<b>Total</b>	44288784.93					



## SUBTRACTING

To subtract, select the two cells with appropriate values and put a minus (-) sign between them (=C2-B2)

B2

✕

✓

*fx*

=C2-B2

	A	B	C	D	E	F	G
1	Public Institution	2011	2012	Difference			
2	Ministry of Difense	7,572	11,976	=C2-B2			
3	Ministry of Foreign Affairs	24,428	28,776				
4	Ministry of Finance	1,979	1,695				

## AVERAGE AND MEDIAN

Excel has inbuilt functions for calculating average and median values.

**Average** - is a number representing central tendency and is calculated as arithmetic mean, by summing up all numbers in the set and dividing it by the number of elements in the set. Formula: =AVERAGE(b2:b21)

SUM							✕		✓		fx		=average(B2:B21)	
	A				B		C	D	E	F				
16	Sport Federation 15				595423.17									
17	Sport Federation 16				584980.4									
18	Sport Federation 17				524491.09									
19	Sport Federation 18				491643.68									
20	Sport Federation 19				441575.94									
21	Sport Federation 20				422281									
22	Total				44288784.93									
23														
24	Average				=average(B2:B21)									

**Median** is a number that is in the middle of a set after the elements in the set have been arranged in the order of increasing magnitude. Formula: =MEDIAN(B2:B21). It is used to calculate a statistical average of such values where there is a very big difference between the lowest and highest values of the set.

B24										
	A	B	C	D	E	F	G	H	I	
15	Sport Federation 14	718172.51								
16	Sport Federation 15	595423.17								
17	Sport Federation 16	584980.4								
18	Sport Federation 17	524491.09								
19	Sport Federation 18	491643.68								
20	Sport Federation 19	441575.94								
21	Sport Federation 20	422281								
22	Total	44288784.93								
23										
24	Median	=median(B2:B21)								

It is recommended that you calculate both values and if there is a big difference between the two, **use median**.

## MODE

Mode is a value that is repeated most often in the set. Formula: =MODE(B2:B21).

B24		=mode(B2:B21)								
	A	B	C	D	E	F	G	H	I	
15	Sport Federation 14	718172.51								
16	Sport Federation 15	595423.17								
17	Sport Federation 16	584980.4								
18	Sport Federation 17	524491.09								
19	Sport Federation 18	491643.68								
20	Sport Federation 19	441575.94								
21	Sport Federation 20	422281								
22	Total	44288784.93								
23										
24	Mode	=mode(B2:B21)								

## COUNTA FUNCTION

If some cells are blank in a column and you want to calculate how many of the cells is populated with a value, you can use CountA function.

This function only calculates the number of the cells and doesn't take into account what type of data is stored in the cells (number, text, symbol, etc.).

**Formula:** =COUNTA(A3:A593) (=COUNTA(First cell:Final cell)) and drag the formula across the entire column.

SUM	X	✓	fx	=COUNTA(A7:A629)						
	A	B	C	D	E	F	G	H	I	J
1	Information about Suicide and Attempted Suicide as of 2017									
2										
3										
4	Suicide	Attempted suicide	Age	Gender	Region					
5										
620	Suicide		67	Male	Imereti					
621	Suicide		46	Male	Imereti					
622	Suicide		62	Male	Imereti					
623	Suicide		22	Male	Imereti					
624	Suicide		37	Male	Imereti					
625		Attempt	54	Female	Imereti					
626		Attempt	27	Female	Imereti					
627		Attempt	61	Female	Imereti					
628		Attempt	23	Female	Imereti					
629	Suicide		28	Male	Imereti					
630										
631	=COUNTA(A7:A629)									

## COUNTBLANK

If you want to calculate the number of blank cells, you can use the formula: =COUNTBLANK(A3:A593) (=COUNTBLANK(first cell:final cell)) and drag the formula across the entire column.

## COUNTIF FUNCTION

If you want to calculate the values from a large dataset based on individual parameters, you can use CountIf function. For example, if you have statistics according to regions and you want to calculate the number of instances in Tbilisi, write =COUNTIF(E3:E626,"\*Tbilisi\*") - Cells, "the parameter to be used") and drag it across the entire column.

	A	B	C	D	E	F	G	H
1	Information about Suicide and Attempted Suicide as of 2017							
2								
3								
4	Suicide	Attempted suicide	Age	Gender	Region			
5								
620	Suicide		67	Male	Imereti			
621	Suicide		46	Male	Imereti			
622	Suicide		62	Male	Imereti			
623	Suicide		22	Male	Imereti			
624	Suicide		37	Male	Imereti			
625		Attempt	54	Female	Imereti			
626		Attempt	27	Female	Imereti			
627		Attempt	61	Female	Imereti			
628		Attempt	23	Female	Imereti			
629	Suicide		28	Male	Imereti			
630								
631	Tbilisi	=COUNTIF(E6:E629,"*Tbilisi*")						

## COUNTIFS FUNCTION

If you want to group data elements according to some criteria into categories or groups (e.g. age group, from some value to another, dates, etc.), you can use Countifs function. It's important to enter the categorization principle correctly in the formula. You can use the following symbols for categorization: >, <, = (for numbers) and "<\*, ?" (for text).

For example, if the dataset includes ages of individuals and you want to group them into age groups, you can use the formula this way:

- ➔ Under 18 age group =COUNTIFS(C6:C871,"<18")
- ➔ 18-25 age group =COUNTIFS(C6:C871,">=18", C6:C871,"<=25")
- ➔ 25-45 age group =COUNTIFS(C6:C871,">25", C6:C871, "<=45")
- ➔ 45-60 age group =COUNTIFS(C6:C871,">45", C6:C871, "<=60")
- ➔ 60-75 age group =COUNTIFS(C6:C871,">60", C6:C871, "<=75")
- ➔ +75 age group =COUNTIFS(C6:C871,">75")

SUM				=COUNTIFS(C6:C871,"<18")				
	A	B	C	D	E	F	G	H
1	Suicide and Attempted Suicide Statistics of 2018							
2	Suicide	Attempted suicide	Age	Gender	Region			
864	Suicide		48	Male	Adjara			
865		Attempt	48	Male	Adjara			
866	Suicide		41	Male	Adjara			
867	Suicide		27	Male	Adjara			
868	Suicide		75	Male	Adjara			
869		Attempt	17	Male	Adjara			
870		Attempt	40	Female	Adjara			
871		Attempt	20	Male	Adjara			
872		<18	=COUNTIFS(C6:C871,"<18")					
873		18-25	118					
874		26-45	321					
875		46-60	188					
876		61-75	110					
877		>75	66					

## COMBINING TWO OR MORE DATABASES BASED ON COMMON VALUES

For example, if you have data according to different months in different Excel sheets and you want to calculate an annual value, you can create a new sheet and use a sum formula. You can do so in the following steps:

- ➔ If the data is categorized according to various parameters (e.g. region, country, etc.), make sure that all Excel sheets are using the same order of the parameters.
- ➔ Copy these parameter categories in the same order as in all other sheets to the newly created sheet.
- ➔ In one of the parameter categories write a formula summing all cells from the same categories from all other sheets by putting a plus (+) sign between each value. E.g.

=(January!B90+February!B90+March!B90+April!B90+May!B90+June!B90+August!B90+August!B90+September!B90+October!B90+November!B90+December!B90)

Drag the same formula across the whole column and Excel will combine data from all other sheets into this one.

B2

fx

=January!B4+February!B4+March!B4+April!B4+May!B4+June!B4+July!B4+August!B4+September!B4+October!B4+November!B4+December!B4

	A	B	C	D	E
	Region/District	PwD aged 0-18	A person with a clearly expressed disability	A person with a significant degree of disability	
1					
2	Gidani Nadzaladevi	11,075	24,859	60,945	
3	Didube Chughureti	4,148	7,570	21,792	
4	Isani Samgori	11,259	22,823	57,848	
5	Vake Saburtalo	7,950	15,501	35,809	
6	Old Tbilisi	3,544	8,975	21,197	
7	Tbilisi	37,976	79,728	197,591	
8	Lanchkhuti	1,350	3,269	15,525	
9	Ozurgeti	482	1,745	2,964	
10	Ozurgeti district	1,708	4,667	11,964	
11	Chokhatauri	735	1,901	5,431	
12	Guria	4,275	11,582	35,884	

## SORTING

Sometimes we need to sort values in the data analysis process according to various criteria: ascending or descending, alphabetizing, etc. To perform this function, click on the Data tab in the menu, and click **Sort**. A new window will appear where you can select the criteria for sorting (ascending, descending, etc.).

The screenshot shows the Microsoft Excel interface with the 'Data' tab selected in the ribbon. A red circle with the number '1' highlights the 'Data' tab. Another red circle with the number '2' highlights the 'Sort' button in the 'Sort & Filter' group. A third red circle with the number '3' highlights the 'Sort' dialog box, which is open and shows the 'Sort by' dropdown set to 'Values' and the 'Order' dropdown set to 'A to Z'. The dialog box also has a checkbox for 'My data has headers' which is checked. The background shows a table with 16 rows of data for 'Sport Federation' and their corresponding 'Amount'.

Name	Amount
Sport Federation 1	13450326.26
Sport Federation 2	6997844.19
Sport Federation 3	3945683.05
Sport Federation 4	2962800.78
Sport Federation 5	2623516.87
Sport Federation 6	2344388.47
Sport Federation 7	2318942.21
Sport Federation 8	1505000
Sport Federation 9	1152892.8
Sport Federation 10	828891.5
Sport Federation 11	797269.28
Sport Federation 12	792778.53
Sport Federation 13	789883.2
Sport Federation 14	718172.51
Sport Federation 15	595423.17
Sport Federation 16	584980.4

## OTHER FUNCTIONS

### CALCULATING A VALUE PER CAPITA

If you are working on such issues as crime, it's important to utilize rate function to better demonstrate the issue. Along with providing data on the total number of crime instances in an area, it's important to calculate crime per capita. To get this value, you need to divide the total number of crime instances by the population. To round up the number, multiply it by 1000. The formula:  $\text{=(crime/population) *1000}$ .

For example, the table below represents [information](#) on the distribution of economic activity across the population, according to the National Service of Statistics of Georgia. Data is presented per 1000 persons. This allows for easier perception. Additionally, along with employment and unemployment rates, their percentage relating to the total population is presented.

**Distribution of population aged 15 and older by economic status, 1998-2019\***

Thousand persons

	2012	2013	2014	2015	2016	2017	2018	2019
Total 15 + population	3057.3	3036.9	3031.6	3019.1	3009.4	3012.3	3034.3	3037.1
Active population (labour force), total	2004.5	1978.6	1984.6	2018.0	1996.2	1983.1	1939.9	1911.2
Employed	1659.4	1643.4	1694.4	1733.8	1717.3	1706.6	1694.2	1690.2
Hired	716.2	693.7	743.5	798.3	801.5	824.2	860.2	849.3
Self-employed	935.7	940.4	944.4	928.0	909.5	881.6	833.4	840.4
Not-identified worker	7.6	9.2	6.5	7.5	6.3	0.8	0.6	0.5
Unemployed	345.1	335.2	290.2	284.2	278.9	276.4	245.7	221.0
Population outside labour force	1052.8	1058.3	1047.0	1001.1	1013.2	1029.2	1094.3	1125.9
<b>Unemployment rate (percentage)</b>	<b>17.2</b>	<b>16.9</b>	<b>14.6</b>	<b>14.1</b>	<b>14.0</b>	<b>13.9</b>	<b>12.7</b>	<b>11.6</b>
<b>Economic activity rate (percentage)</b>	<b>65.6</b>	<b>65.2</b>	<b>65.5</b>	<b>66.8</b>	<b>66.3</b>	<b>65.8</b>	<b>63.9</b>	<b>62.9</b>
<b>Employment rate (percentage)</b>	<b>54.3</b>	<b>54.1</b>	<b>55.9</b>	<b>57.4</b>	<b>57.1</b>	<b>56.7</b>	<b>55.8</b>	<b>55.7</b>



## FREEZING ROW/COLUMN

When working with a large database, it is recommended to use pane freeze function. The frozen column or row will remain on display even when you navigate away from it. To utilize this function in Excel, click **View**, select **Freeze Panes**, three options will be displayed:



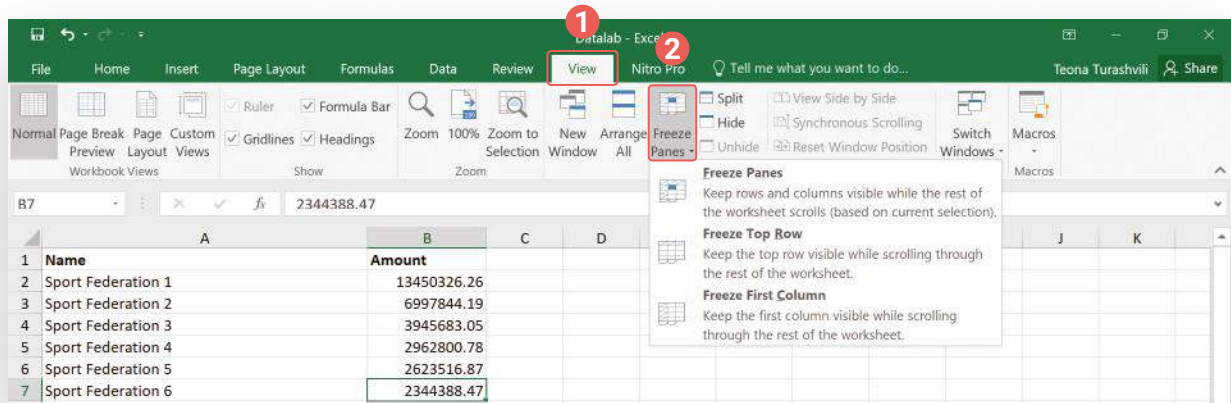
**Freeze Panes** - Choose the range of a column or row that you wish to freeze and click freeze.



**Freeze Top Row** - If you want to freeze only the top row, use this function.



**Freeze First Column** - If you want to freeze only the first column, use this function.



## HIDING COLUMNS

If you want to temporarily hide individual columns, select it, right click and in the window select **Hide**.

If you want the column to reappear, right click left and right columns of the hidden one and select **Unhide Columns**.

## ADDING A NEW COLUMN/ROW

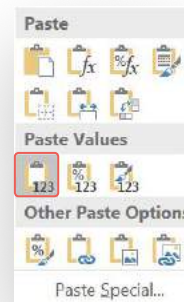
If you want to add a new column, for example between B and C, select the column C, right click and choose **insert**. A new column will appear. The same principle can be used for adding a new row.

## GOING TO THE TOP

If you are working in a large database and want to go to the top, click **Control+Home**.

## COPYING DATA WITHOUT FORMULAS

If you want to copy data without the accompanying formulas, copy the data with **Control+C**, in the new sheet select where you want to paste, right click and in the window **Paste Special** select **Values** - this will copy the values without the formulas.



## EXCEL PIVOT TABLES

Pivot tables are recommended to process raw data and group them appropriately.

When using pivot tables, ensure:

- ➔ That the database doesn't contain an empty column or a row;
- ➔ That every column has a name. Try to use short names;
- ➔ That the database has raw data and doesn't contain sums or values resulting from a formula;
- ➔ That the database doesn't include some values.

Select the entire database, click **Control+T** and the database will be transformed as displayed below:

	A	B	C	D	E	F
1	Suicide	Attempted suicide	Age	Gender	Region	
2		Attempt	42	Male	Tbilisi	
3	Suicide		68	Female	Tbilisi	
4		Attempt	19	Male	Tbilisi	
5	Suicide		61	Male	Tbilisi	
6	Suicide		20	Male	Tbilisi	
7	Suicide		50	Male	Tbilisi	

Select **Insert** tab and a pivot table function will appear. The window will offer you recommended pivot tables that can be automatically generated by Excel.

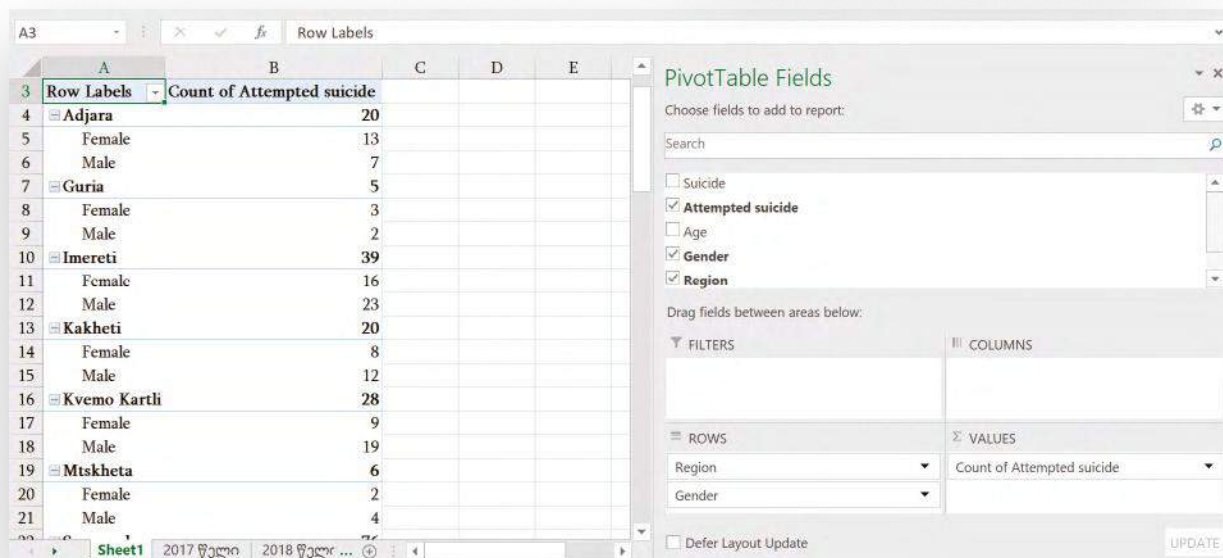
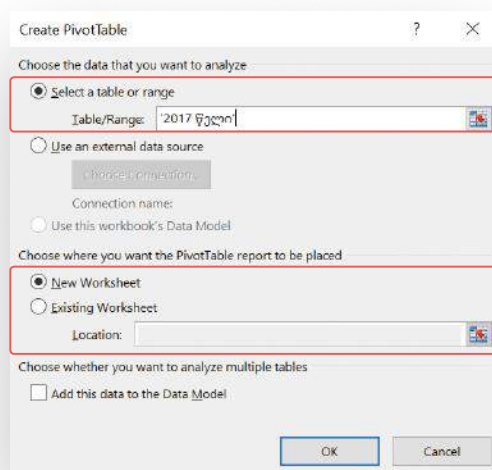
The screenshot shows the Excel interface with the 'Insert' tab selected. The 'Recommended PivotTables' task pane is open, displaying several suggested pivot tables. The first suggestion, 'Count of Age by Region', is selected. The pane also shows other suggestions like 'Count of Age by Suicide' and 'Count of Age by Gender'.

**Count of Age by Region**

Row Labels	Count of Age
Adjara	45
Guria	21
Imereti	93
Kakheti	48
Kvemo Kartli	63
Mtskheta	18
Samegrelo	118
Samtskhe	20
Shida Kartli	54
Tbilisi	143
<b>Grand Total</b>	<b>623</b>

To build a new table, click **Pivot Table**, in the new window select data location you want to analyze and the location where you want the pivot table to appear (new worksheet or the same, it is recommended to select the first option).

After the confirmation, a new page will appear, where in the right corner you can select categories based on which you want to build the table. For example, if you want to categorize suicides according to regions, in the row section select regions and in the value section number of suicides. If you want to further categorize the data based on sex, you can add sex to the row and data for each region will be groups for each sex.



If your data relates, for example, to salaries for various positions and you want an average salary after grouping, after moving salaries to the Value section, click on the arrow next to the name and in the dropdown menu select **Value Field Settings**. In the new window, you can select how the data should be calculated (sum, average, minimum, maximum, etc.). You can also choose the format of the final values. Click **Number format** and select the most suitable one.

The screenshot illustrates the steps to configure a PivotTable field. The main window shows a PivotTable with 'Row Labels' containing region and gender data, and a single value field 'Count of Attempted suicide'. The 'PivotTable Fields' task pane on the right shows the field list with 'Attempted suicide', 'Gender', and 'Region' checked. The 'COUNT of Attempted suicide' is in the Values area. The 'Value Field Settings' dialog box is open, showing the 'Summarize value field by' dropdown set to 'Count'. The 'Number Format' button is highlighted. Red numbers 1 through 4 indicate the sequence of steps: 1. Click the dropdown arrow next to 'Count of Attempted suicide' in the Values area. 2. Click 'Value Field Settings...'. 3. Select 'Count' in the 'Summarize value field by' list. 4. Click 'Number Format'.

If you want to view one of the grouped data categories in detail, based on all categories of the initial raw data, double-click on the value and you will see a dropdown version with all included values. In this example, there were 5 cases in Guria. By double-clicking on it we can view all five cases.

	A	B	C	D	E	F	G	H
1	Suicide	Attempted suicide	Age	Gender	Region			
2	Suicide		36	Female	Guria			
3	Suicide		39	Female	Guria			
4	Suicide		60	Female	Guria			
5	Suicide		45	Male	Guria			
6	Suicide		55	Male	Guria			

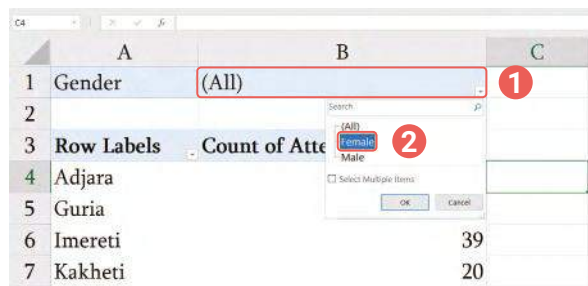
## DATA FILTERING

If you want to view only certain categories of data, you can use the filter function. For example, if you want to view data only from certain region, click on the dropdown menu and select only the regions you want to view. After the confirmation, only data from those regions will be displayed.

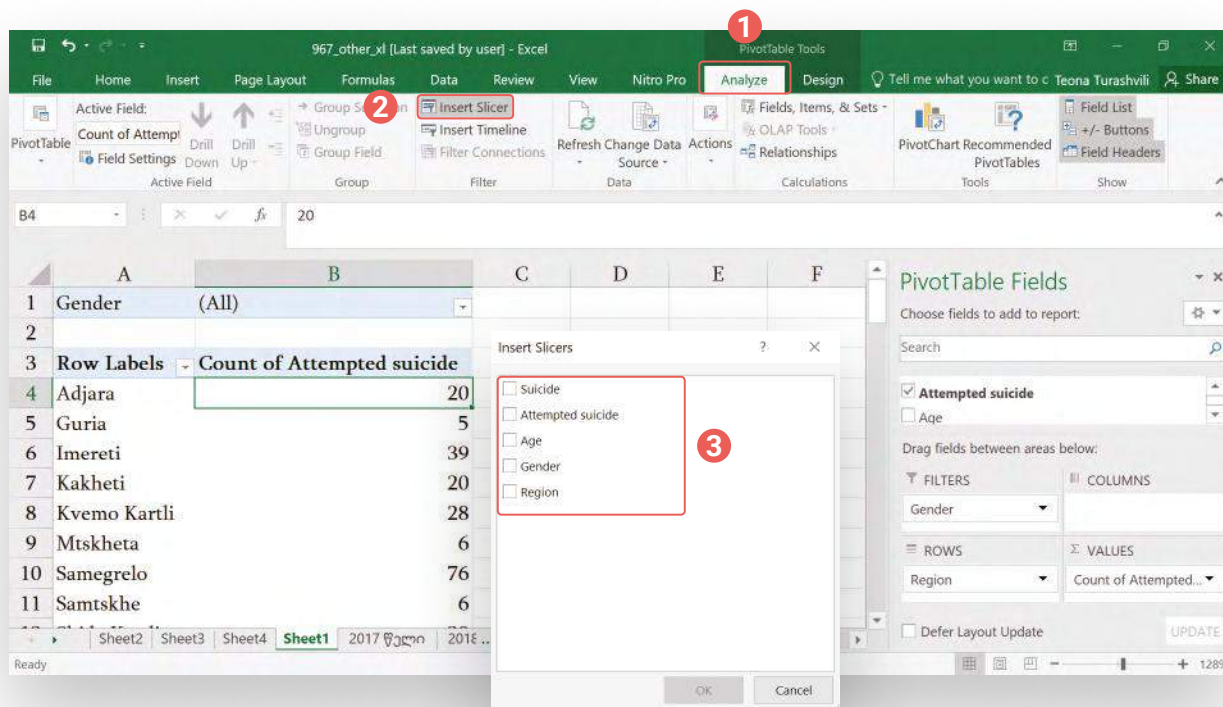
	A	B	C	D	E
2					
3	Row Labels	Count of Attempted suicide			
4	Adjara		20		
5	Female		13		
6	Male		7		
7	Guria		5		
8	Female		3		
9	Male		2		
10	Imereti		39		
11	Female		16		
12	Male		23		
13	Kakheti		20		
14	Female		8		
15	Male		12		



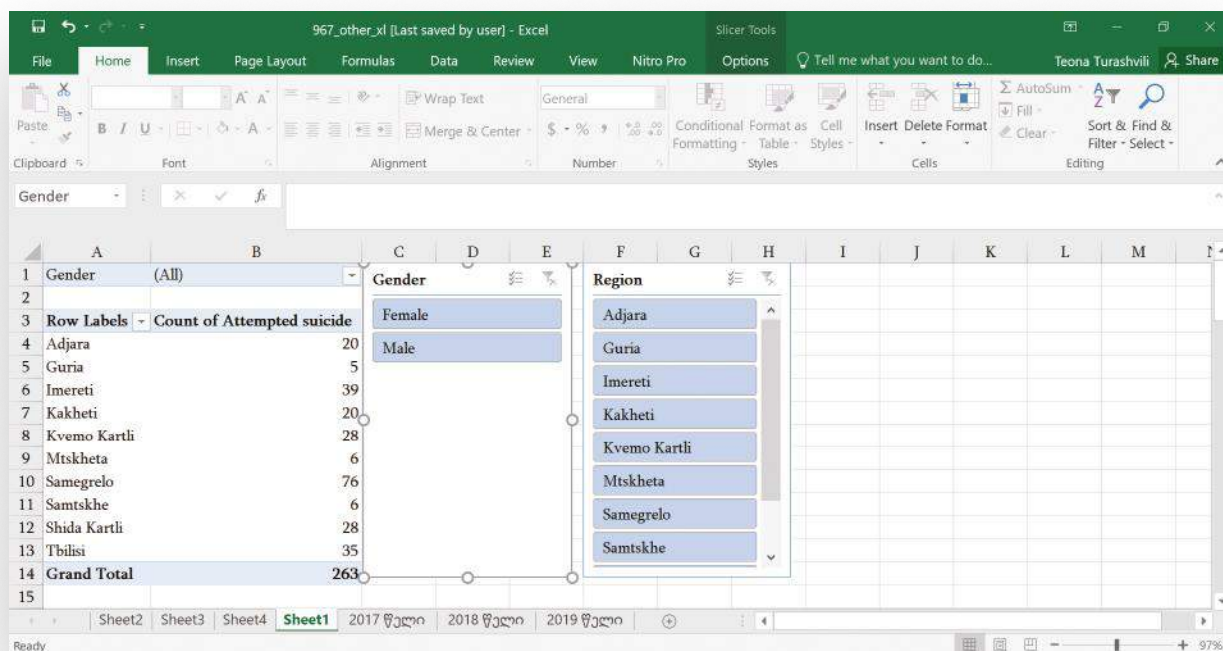
If you want to filter data based on a category that is not included in the processed data, e.g. sex, drag this category to the **Report Filter** section. After this, this category will appear on the top of the data, click on the dropdown filter and select the category, confirm and only data from this category will appear.



You can also use **Slicer** function for filtering. If you filter data frequently, it is recommended to use a slicer. Go to **Analyze**, select **Insert Slicer**, a new window will appear and you can select appropriate filters.

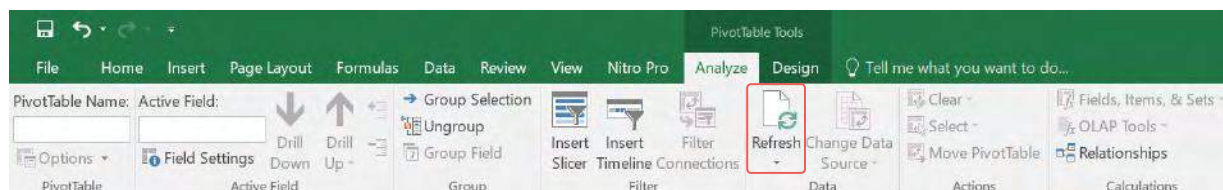


After the confirmation, options for the selected filter/category will appear and you can filter data based on sub-categories, as well.



## UPDATING DATA

If you make any changes to the raw/initial data, before starting working in pivot tables, go to the **Analyze** tab and select **Refresh**.





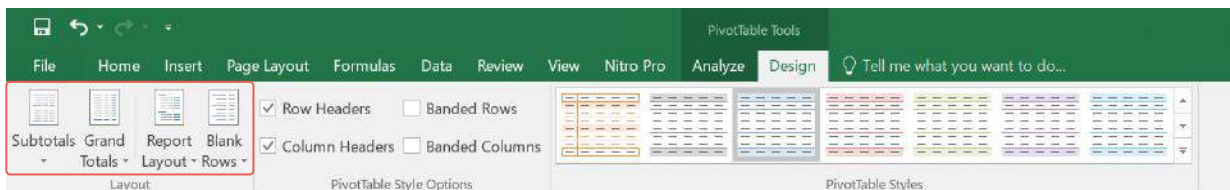
## TIMELINE

If your data includes dates and you want to sort them chronologically, make sure that all dates are in the same format. Then go to the **Analyze** tab and click **Insert Timeline**.



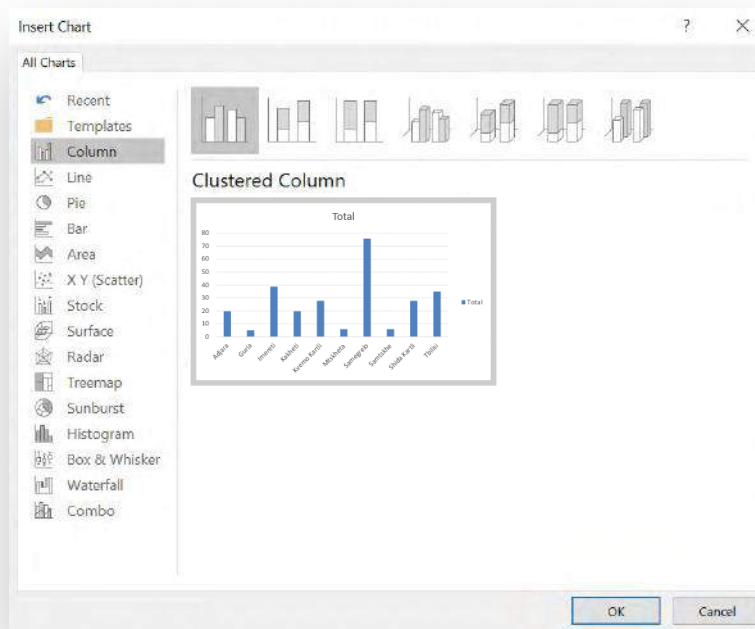
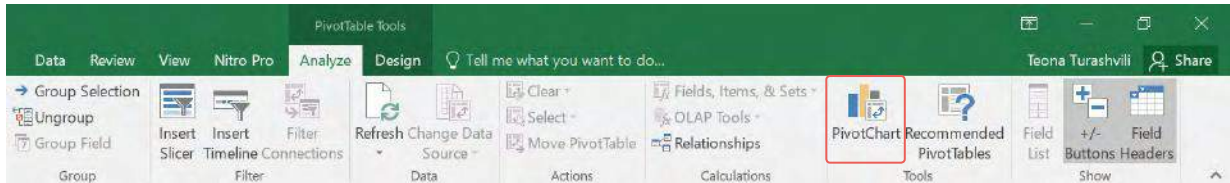
## TABLE DESIGN

After building your tables, you can select a design for them. Go to the **Design** tab, where you can add rows and columns, subtotals and grand totals of columns and subcategories. You can also select layout design (Report Layout, Blank Rows).



## CHARTS

Along with data processing, pivot tables allow you to create charts and graphs based on the filtered and categorized data. Go to the **Analyze** tab, select **PivotChart**, options for various types of charts will appear and you can select one.



# FINDING TRENDS AND INSIGHTS IN DATA

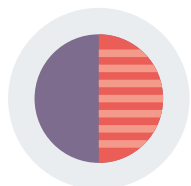


Data processing and analysis will allow you to find important insights and identify trends:



## TRENDS

How does a value change according to years and various groups.



## CONTRAST

Differences among the values.



## DIVERGENCE FROM STANDARD

Individual values diverge from the rest. Validate the correctness of the entry to prevent any mistakes.

---

Remember, finding insights in data is a constant process and you might find yourself analyzing, reading, processing, and visualizing the same data from various perspectives. Here are some general recommendations:

## REGROUPING, ADDING, AND COMPARING

If the data represents the issue through various categories (age group, city, etc.), it's important that your findings are also for different categories. When studying the issue in-depth, it's important to consider the circumstances. For example, if you want to compare crime in various regions of the country, you should consider that the size of the population of each region varies, to ensure that the conclusions

of the comparison are just and adequate. It might also be interesting to find the causes of crime based on particularities of each region.

## CAUSE-EFFECT AND CORRELATION





After processing and analyzing several datasets, you might find a correlation between various values/trends. However, be cautious when drawing conclusions on cause-effect relationship. A correlation between two values (e.g. number of homeless people and crime rate) doesn't imply a cause-effect relationship (e.g. homeless people commit crimes). You might find that there are other, independent reasons causing increase or decrease in crime.

## VISUALIZATION OF DATA FOR EFFECTIVE PRESENTATION OF TRENDS

Graphical representation of data will allow you to see relationships, trends, and insights more easily. This is particularly true in cases of geolocational data. You can also represent this data using various charts. Engaging in this type of experiments will allow to see the same data from various viewpoints and you might be able to identify new significant insights.




Along with data analysis, when preparing an analytical article for publication online, it is important to consider how you will display the data. It's well-known that readers do not like large texts, they first read headline, sub-titles, specially-marked text, graphics, and numbers.

# WHEN USING NUMBERS, CONSIDER THE FOLLOWING RECOMMENDATIONS:

-  Use numbers to express small numerical values rather than words (30 and not thirty);
-  Numbers up to one million are better expressed in number symbols rather than words (5,000 and not five thousand);
-  Sometimes it's better to use a combination of both numerical symbols and words, especially for larger numbers, e.g. 24 million and not 24,000,000;
-  Use commas to denominate thousands - 1,500, 25,000.

# SIMPLIFYING PERCENT:

When viewing information in percent, a reader might not be able to fully grasp the gravity of the issue. Therefore, it is often more effective if you present the percent as the percentage share of the whole population. For example, if you want to make 21%, 35% or 74% more easily perceivable for readers, you can use the following alternatives:

PERCENT	SHARE	POPULATION RATE	
21%	One fifth	Every fifth person, one in five people	
35%	Approximately one third	Approximately every third person, one in three people	
74%	Three quarters	Three is very four people	

## EXAMPLE:

If the government spends 21% of the budget on social programs, instead of percent, you can say that one fifth of the budget is spent on social programs.

If 35% of the population is unsatisfied with the services provided by the local government, you can say that every third person is unsatisfied with the services.

## NUMBER COMPARISON

If you want to compare two percent values, you can present the difference in numbers.

## EXAMPLE:

If you have the following numbers:

- ➔ 85% of the population of the capital has access to the Internet;
- ➔ 49% of the rural population has access to the Internet.

Instead of percent you can use numbers:

- ➔  $85\% = 85/100 = 17/20$
- ➔  $49\% = 49/100 = \frac{1}{2}$

Therefore, you can rephrase the initial sentences:

➔ 17 people out of 20 in the capital have access to the Internet;

➔ In regions 10 out of 20 people have access to the Internet.

or

In regions every second person has access to the Internet.

## ROUNDING UP NUMBERS

Large numbers are difficult to perceive for readers. If you don't need an exact number, it is recommended to round them up.

### EXAMPLE:

**If you have the following sentence:** In 2018 Gel 8,949,777 was spent from the reserve funds of 60 municipalities.

**We can simplify it:** In 2018, approximately, GEL 9 million was spent from the contingency funds of 60 municipalities.

**In visualizations it is recommended to use rounded numbers.**



## CALCULATING RATE

With rate calculation you will be able to identify the share of population that was impacted by a certain factor. You can calculate rate in proportion to the population.

### EXAMPLE:

If it was identified, that 13% of diseased population died in a natural disaster, you can turn this into a rate. To do so, divide the entire population (100%) by 13.  $100/13=7.6923$  round up to 8)

Based on this calculation you can say that 1 in every 8 deceased died in a natural disaster.

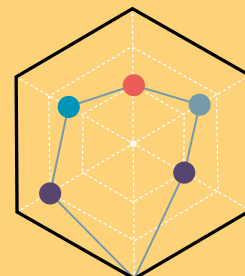
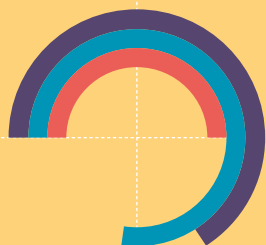
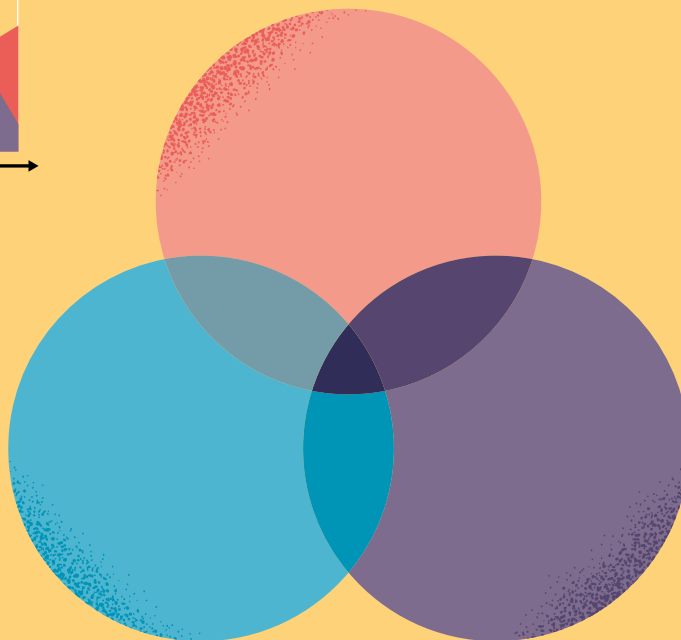
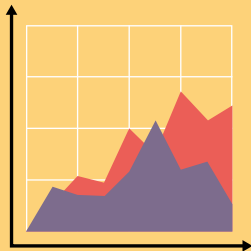
The number will be more accurate if you multiply it by 10. Thus, we get: 10 out of every 77 deceased died in a natural disaster.

To demonstrate rate and percent more effectively use the following divisions:

- |                                                            |                                              |
|------------------------------------------------------------|----------------------------------------------|
| ➔ 0.05 - 5% - one in twenty / every twentieth              | ➔ 0.5 - 50% - one in two / every second      |
| ➔ 0.1 - 10% - one in ten /every tenth                      | ➔ 0.6 - 60% - three in five                  |
| ➔ 0.2 - 20% - one in five / every fifth                    | ➔ 0.66 - 66% - two in three                  |
| ➔ 0.25 - 25% - one in four or $\frac{1}{4}$ / every fourth | ➔ 0.75 - 75% - $\frac{3}{4}$ , three in four |
| ➔ 0.33 - 33% - one in three / every third                  | ➔ 0.8 - 80% - four in five                   |
| ➔ 0.4 - 40% - two in five                                  | ➔ 0.9 - 90% - nine out of ten                |
|                                                            | ➔ 0.95 - 95% - nineteen out of twenty        |

If the percent/rate is different, use phrases like “approximately”, “more than” and round up the percent/rates.

# DATA VISUALIZATION



Visualization and graphical representation are the most effective ways to deliver an interesting story with data to you readers.

### Why you should use visualization?

- ➔ Brain remembers visual information better;
- ➔ It is easier to perceive and compare visualized data;
- ➔ Visualization attracts attention from readers;
- ➔ Along with attracting the attention of the readers, they are able to better perceive trends;
- ➔ Visualization communicates trends and insights effectively.

To use these opportunities to their fullest potential, it is important to understand which charts demonstrate which data better. Otherwise, the chart will not properly communicate your insights.

Before creating the visualization, ask the following **questions**:

- ➔ Who is your target audience?
- ➔ Which information does your target audience need to remember?
- ➔ What is the main message behind your visualization and how can you connect the reader emotionally to the story?
- ➔ What do you want to show?
- ➔ What actions does your target audience need to take after seeing your visualization/studying your data?

Answers to these questions will help you determine **which chart** to use.

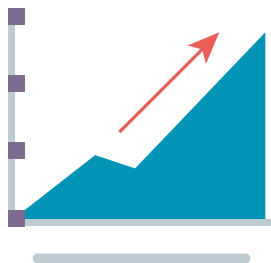
Visualization is commonly used for **four purposes**:

- ➔ Data comparison;
- ➔ Structure demonstration;
- ➔ Distribution demonstration;
- ➔ Data correlation.

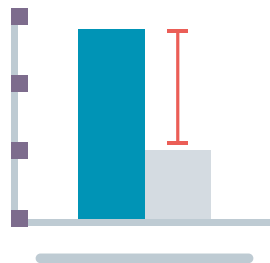
To select the proper chart for one of these purposes, you should consider the following questions:

- ➔ How many values do you want in each chart?
- ➔ How many numbers for each value do you want to present?
- ➔ How do you want to organize the data: time, various groups, or categories?

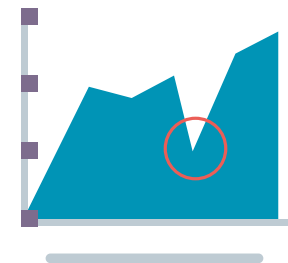
## WHAT TO LOOK FOR IN DATA?



TRENDS

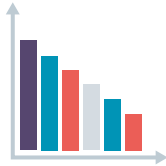


CONTRAST



OUTLIERS

## ONE DIMENSION



Bar



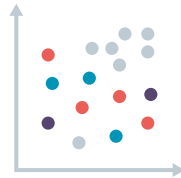
Pie



Picto

## COMPARISON

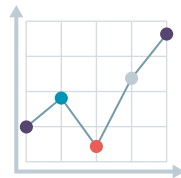
## TWO DIMENSIONS



Scatterplot



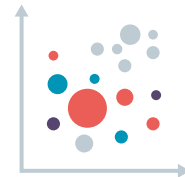
Area



Line

## TENDENCY/RELATIONS

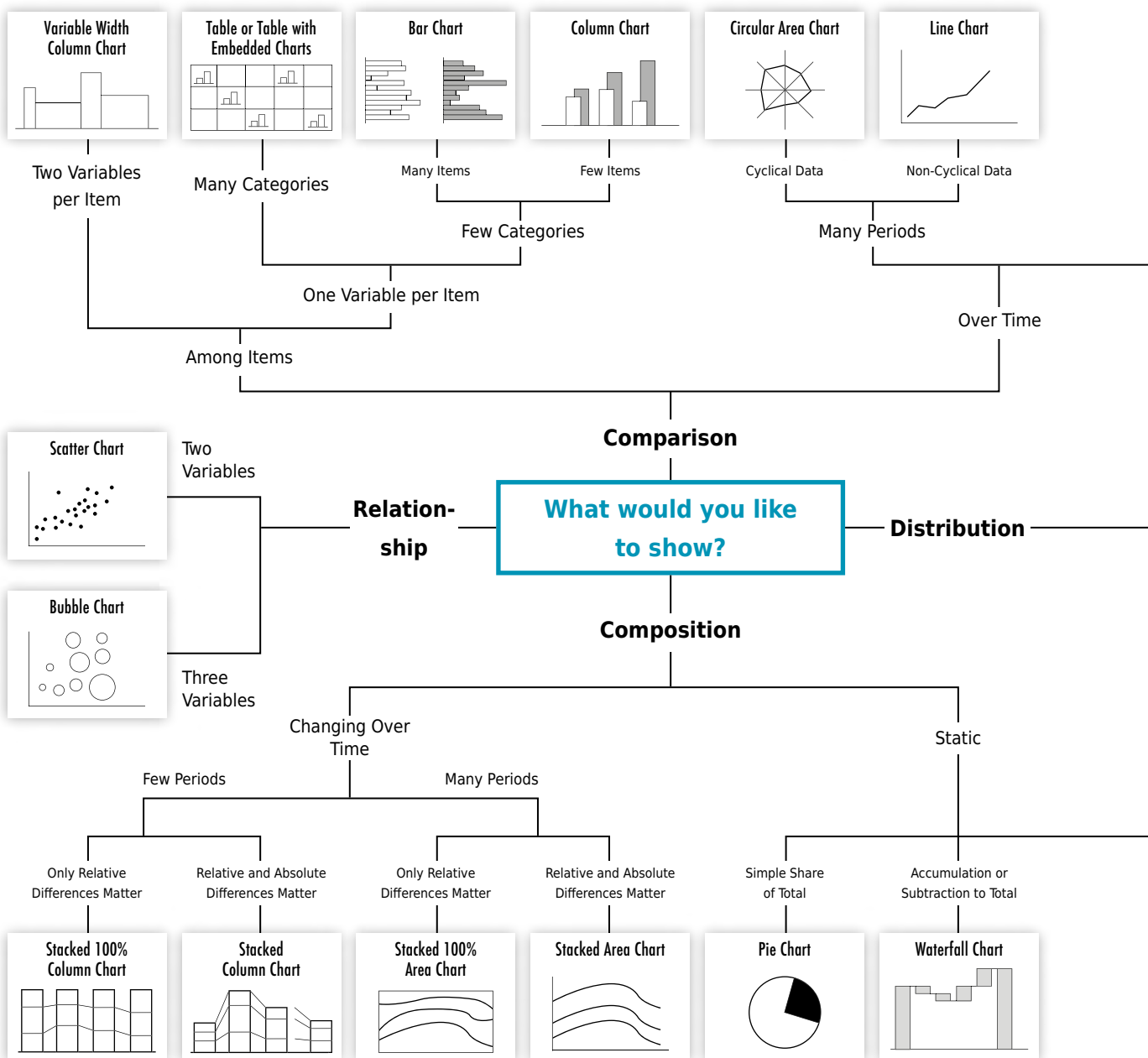
## THREE+ DIMENSIONS

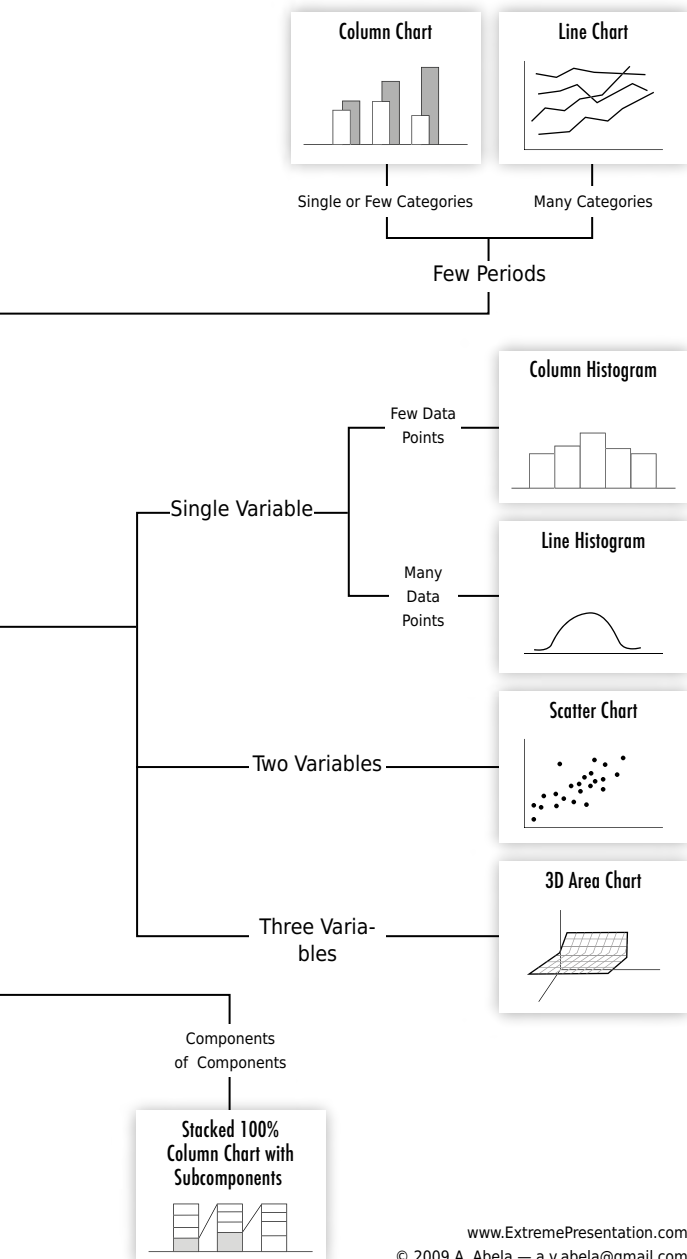


Bubble

## COMPARISON/ RELATIONSHIPS

# CHART SUGGESTIONS - A THOUGHT-STARTER



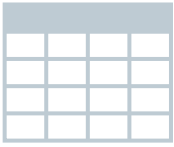


## Simple Rules:

- ➔ **Bar chart** is used for comparison of different categories.
- ➔ **Line graph** is used to display trends over time, as well as correlation between various values.
- ➔ **Scatter plot** is best used to demonstrate distribution and correlation.
- ➔ > **Pie chart** is used to only show the structure. **Never use a pie chart for comparison!** It represents one entity or a sum.
- ➔ **Maps** are used for comparison of geographically distributed data.
- ➔ **Pictogram** is used as a human/object representation for data comparison.

Here are some particularities of several charts:

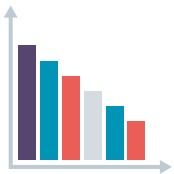
---



## TABLES

Tables is the main source for any visualization. Values populating a table are the building blocks for any chart. However, sometimes table itself communicates data most effectively. Generally, tables are used for showing **comparison, composition, or connection**. Tables are used when:

- ➔ It's important to use/demonstrate every value;
- ➔ You need exact numbers;
- ➔ There are different types/categories of data which are irrelevant to each other, however it is important to display them in one chart;
- ➔ Displayed numbers do not demonstrate a trend and contain only quantitative information.



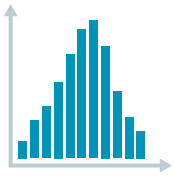
## COLUMN CHARTS

Column charts are most frequently used charts. They are perfect for comparison, especially, when display of numerical values is important. A reader can compare columns representing the numbers to the same value in different categories, or observe its change over time. It's recommended to use column chart:

- ➔ For comparisons, if the number of categories is not too large, up to five categories is acceptable, however no more than seven.
- ➔ If one category of data is time – year, quarter, month, week, day, or hour. You should allocate this data on the horizontal axis.
- ➔ In the diagram, the time passage should be denoted from left to right and from top to bottom.

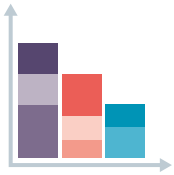


- ➔ The numerical axis should start at 0. Otherwise, relative proportions of various numbers might be misconstrued by a reader.



## HISTOGRAM

is used for demonstration of distribution and share of a value across different categories. A good example of a histogram is a value distribution across age groups. For example, crime rate across age groups.



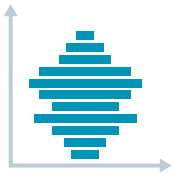
## STACKED BAR CHART

is used to show the composition. It is not recommended to use a stacked bar chart if the subject consists of more than four components. It is also recommended that the objects that are being compared do not differ very significantly. Otherwise, it will be difficult to effectively perceive data on the chart.



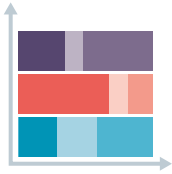
## BAR CHARTS

is a horizontal column chart. If you have a lot of data categories, bar charts are most effective. However, more than 15 categories cannot be effectively displayed.



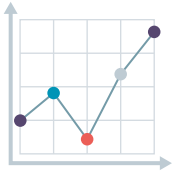
## BAR HISTOGRAM

is most frequently used to demonstrate distribution of population according to various categories (age, sex).



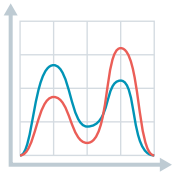
## HORIZONTALLY STACKED BAR CHART

is recommended when you have a small amount of values and want to show their composition (and not comparison). This chart is not very useful for comparison or correlation analysis.



## LINE CHARTS

is one of the most frequently used charts. It represents well trends and change in values over time. It is most useful when you have a large number of values, including over 20.



## TIMELINE CHARTS

is a variety of a line chart. The main difference is that timeline chart allows for a more detailed view of the trends as it reflects change over days, hours, or even over shorter periods. A user can view the change over the time frequency they find most appropriate.

This type of chart is most often used on stock market to show price change, number of website visitors or changes in sales volume.

---



## AREA CHART

is another variety of line chart and displays trends and comparisons well. In this chart area below the line is colored, and thus better displays change in volume or number over time (e.g. number of employed persons, savings, etc.). Do not use this chart for stock market or to display price change.



## STACKED AREA CHART

demonstrates change in composition over time. For example, share of users among the main players on the internet market over several years. The chart uses colors, however, be careful when selecting them as the chart needs to be easily perceivable.



## PIE CHART

is probably most used chart and often incorrectly. Usually, a pie chart demonstrates division and composition of one entity (e.g. budget, population, etc.) and percentage/share of each component. This chart is best used for demonstrating composition and connection.

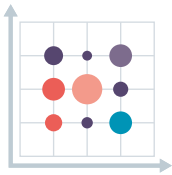
The chart is ineffective if it has too many components (no more than six is recommended) or if the difference between the shares is not very significant.

Do not use a pie chart if you want to compare individual categories. Column chart will be more useful. Avoid using 3D or other effects, as this might distort the accuracy and proportions of the chart.



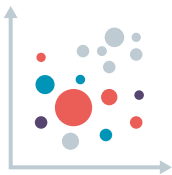
## SCATTER CHARTS

is mostly used for correlation and distribution analysis. Using the chart, you can demonstrate how two values interact and how much impact they have on each other. It also shows how data is distributed and grouped, allowing us to underline outliers.



## BUBBLE CHARTS

will be most useful, if you want to add one more dimension to the scatter charts. Scatter chart compares two dimensions/values, while bubble chart allows you to add a third one represented with the bubble size. If the bubble sizes do not vary significantly, use categories/legend. This chart is not useful if you want to compare exact values. They are often used to approximately compare budgets or populations of various countries.



You can add one more dimension to the bubble chart by varying the bubble colors or by representing bubbles as pie chart or by allocating different values to X and Y axis. In such cases, the location of the bubble on the chart also represents a value, if only Y axis is used, closer to the tip of the axis the bubble is, the bigger the value it represents, for example level of education of a country's population. If you use both axes, then closer to the tip of the Y axis the bubble is, and further right on the X axis it is, the higher the education level of the country's population and bigger their life expectancy. However, do not overpopulate the chart.

Bubble diagrams are most often used to display correlation between market expenditure, revenue, and profit. The standard chart can show positive correlation between expenditure and revenue while scattered bubble chart can show how market expenditure impacts revenue.

Use scatter and bubble charts:

- ➔ To demonstrate correlation between two (scatter) and three (bubble) values;
- ➔ To demonstrate trends (linear and nonlinear trends, correlation, cluster, or outlier) in large data-sets;
- ➔ To display large values without the timeline. More data you enter in the scatter chart, the better it will display the comparison.
- ➔ Not for comparison of exact numbers but for general comparison and correlation.



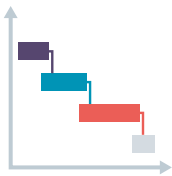
## MAP CHARTS

are most effective to display geographic distribution/context of data. It allows to show which regions perform well and which don't, what are the trends in each region/country. You can build such charts if you have the following location data: coordinates, country, region, address, etc.

Map charts are not very effective for comparison of exact values. You can use additional bubbles or numbers nested on the map. If you need exact comparison and general categories, you can use different color combinations for values in different categories.

When are map charts used?

- ➔ To demonstrate numbers on a map;
- ➔ To demonstrate geographic trends, distributions, and correlations;
- ➔ When data has geographic context/significance;
- ➔ If data is standardized and during the data grouping process geographic areas have the same scale.



## GANTT CHARTS

Gantt chart is used for project management. It represents a project map, demonstrating what needs to be done, in which order and at what time. The diagram allows you to demonstrate how long the project will run, what resources are needed, and what is the order of task execution and their relation to each other. You need start and end dates for the chart. To create more complex a Gannt chart, you will need task completion percentages and/or their relationship to other activities.

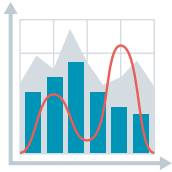


## GAUGE CHART

Gauge chart is used to show indicator achievement. Generally, it displays one value that is represented with a color combination. This chart is useful:

- ➔ To show progress stage as it relates to the final goal;
- ➔ To show the percentage of progress;
- ➔ To display one value exactly.

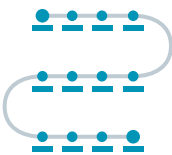
One disadvantage of this chart is that it displays only one value and occupies a big space. If you want to demonstrate achievement of several indicators and their comparison, column charts are recommended. They are more complex and compare various values more effectively.



## MULTIPLE AXIS CHART

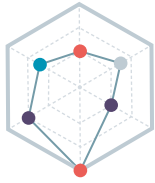
In some cases one chart is unable to fully communicate the story, as you want to demonstrate various types of connections and comparisons. Multiple axis chart is useful in such cases. Using the chart, you can display data along two or more Y axis and one common X axis. This chart is great for demonstration of connection, trends, and correlation among various types of data.

While this chart can communicate complex relationships, they might be difficult for readers to read or perceive. Such charts also often do not allow us to compare exact values, displaying only general trends.



## TIMELINE

is used to demonstrate an order of events over a timeframe. The timeline starts in the top left corner and continues down a spiral. This chart is used for biographies, histories, and instruction demonstrations.



---

## RADAR CHART

is used to display/analyze components, characteristics, and ingredients of one entity/indicator. Each value represents one component/characteristic of the entity and displays its volume/amount. Similar to the column chart, radar chart compares values to each other, however, it does so in the context of the main entity and its components.



---

## POLAR GRID

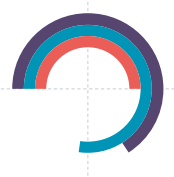
similar to radar chart, displays components of one entity. However, it allows us to add more circles as you are not limited to add categories in corners.



---

## TREE MAP

is used, for example, for budget visualizations, when we need to demonstrate components of the budget and compare their volumes. Bigger the share of the component, larger the rectangle will be. It's important that the names of the small rectangles are legible for the readers.



---

## CIRCULAR CHART

is used for comparison of the length/duration of several values. Each circle represents 100%. This chart is used for agendas, life expectancies, and program implementation period comparisons.

---





## VENN DIAGRAM






represents an overlap between two or more subjects. However, it doesn't display the scale of the overlap.

---




The main principles of **data visualization**:

- ➔ **Simplicity** - do not use more than six colors. Use color combinations that are complementary. Use a single-color combination. Do not use 3D effects.
- ➔ **Hierarchy/order** - every story needs the main theme/idea, around which the analysis is built. In column or bar charts populate data in descending order. When using a time dimension, allocate the timeline on the horizontal axis, from left to right.
- ➔ **Conciseness** - use short sentences. It's important that you know what you want to say and what your main message is before creating a visualization.
- ➔ **Creativity** - In the visualization use a creative/humorous design related to the context of the topic.
- ➔ **Clarity** - Use a legend or directly write categories next to their graphical representation to make chart easier to read. If you have only one data category, you don't need a legend. It's essential that the target audience is able to read and perceive you chart easily.

## (FREE) ONLINE PLATFORMS FOR DATA VISUALIZATION:

## PLATFORMS FOR PUTTING DATA ON MAPS:

	Google My Maps
	Google Earth
	Google Maps Platform

## ADDITIONAL RESOURCES:

---

### **Timeline JS** <sup>37</sup>

A simple, free way to display information in a timeline

---

### **StoryMapJS** <sup>38</sup>

A free resource to display data on a map

---

### **Gephi** <sup>39</sup> and **NodeXL** <sup>40</sup>

Platforms for correlation and radar charts

---

### **Color Brewer 2.0** <sup>41</sup>

A resource for selecting a color palette

---

### **Color Hunt** <sup>42</sup>

A color scheme library for designers

---

### **Data Visualization** **Catalogue** <sup>43</sup>

Around 60 chart types are demonstrated. Charts can be selected based on their type and function.

---

<sup>37</sup> <https://timeline.knightlab.com>

<sup>38</sup> <https://storymap.knightlab.com>

<sup>39</sup> <https://gephi.org>

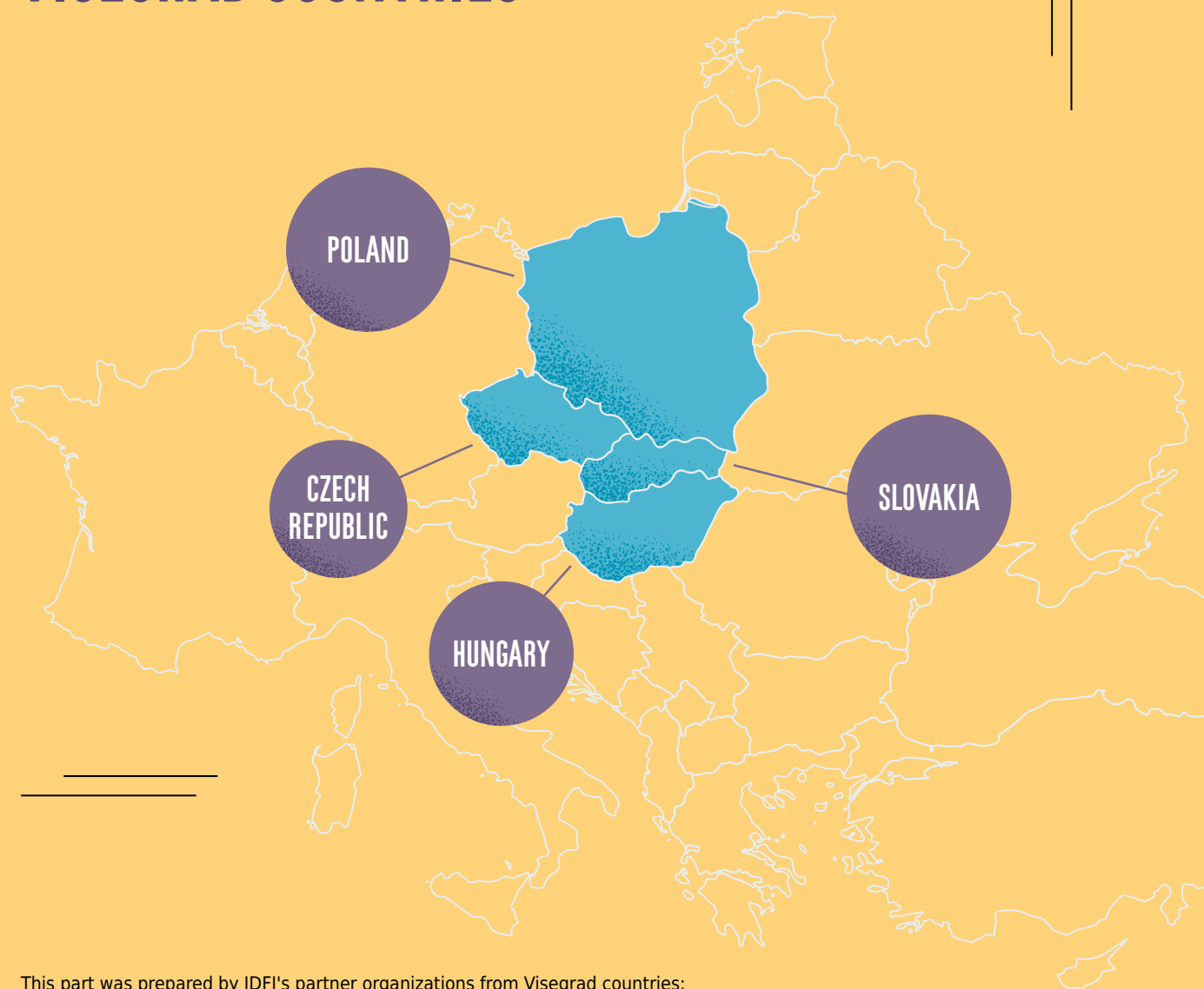
<sup>40</sup> <https://www.smrfoundation.org/nodexl>

<sup>41</sup> <http://colorbrewer2.org>

<sup>42</sup> <https://colorhunt.co>

<sup>43</sup> <https://datavizcatalogue.com/index.html>

# OPEN DATA PLATFORMS FROM VISEGRAD COUNTRIES



This part was prepared by IDFI's partner organizations from Visegrad countries:

[KohoVolit.eu](https://kohovolit.eu) (Czech Republic and Slovakia), [K-Monitor Public Benefit Association](https://k-monitor.hu) (Hungary), [ePaństwo Foundation](https://e-panstwo.org) (Poland)

# HUNGARY

---

## **Name of the platform:**

Glass Pocket Tracker - <https://uvegzseb.hu/>

---

## **Name of the initiators:**

Investigative portal Atlatszo.hu with the support of Google's Digital News Initiative

---



## **Major aim of the initiative:**

Glass Pocket Tracker (Üvegsebfigyelő in Hungarian) helps citizens to track public data published by government institutions, state-owned enterprises, municipalities and state agencies. It was launched by the investigative portal atlatszo.hu with the aim to make government publications searchable in one single platform and be informed about updates of databases or changes of datasets, contracts, etc.

## **Major details about the initiative:**

The service works with crawlers that are checking the websites of the above-mentioned government bodies and companies, and as soon as the crawler finds a change in the datasets on their sites or finds a new dataset uploaded, it will immediately connect that dataset to the Glass Pocket Tracker. From that moment, users are only one click away from the document – be it a pdf, an excel document or an html page. At the time of launching, users had access to 73 thousand documents published by 240 different government agencies, state-owned enterprises and municipalities. This number is constantly increasing.

The Glass Pocket Tracker also makes it extremely easy to search these documents that include contracts, documentation of public tenders, business reports and data about the organizations themselves, as there is no need to browse hundreds of different entities and check subpages and disclosure lists.

The Glass Pocket Tracker also rates all data owners. Its algorithm records how often datasets are updated and also how complex and accessible they are. A rating between 1 to 5 is assigned to each institution.

Users can subscribe to updates of the databases they are interested in.

---

**Name of the platform:**

Red Flags - [redflags.eu](http://redflags.eu)



---

**Name of the initiators:**

K-Monitor Association, TI Hungary with the support of the European Commission

---

**Major aim of the initiative:**

The Red Flags tool was launched in 2015 with the purpose of bringing more transparency to the Hungarian public procurement system and support the prevention of problematic procurements by creating an IT system capable of generating automatic alerts whenever a contract notice or contract award notice seems suspicious or may contain corruption risks. The main target group of the tool are procurement experts and journalists.

### **Major details about the initiative:**

The Red Flag tool processes tender notices and contract award notices published in the Tenders Electronic Daily (TED) to flag risky procurements in Hungary through algorithms. The tool operates with a set of 40 risk indicators, from which 32 are used to check contract notices and 9 are used to check award notices.

Indicators are built around data on technical capacity and economic and financial ability requirements as well as the use of specific procedures considered favorable ground for corruption. For contract award notices, indicators can include procedures without prior publication, the ratio of the total final value and estimated value, or unsuccessful procedures for risky reasons or without statement of reason.

### **Open Data component of the initiative:**

The tool is fully open source both regarding the database and the source code. The documentation of the tool is available under [docs.redflags.eu](https://docs.redflags.eu).

### **Innovation of the initiative:**

It is one of the most developed open source tool of this kind developed by civil society actors alone. The application of algorithms allows to easily check risky procurements without having to read through hundreds of procurement notices.

### **Some statistics:**

The platform has over 700 registered users. Around 20 people use the site daily.

# POLAND

---

## Name of the initiative:

[Rejestr.io](https://rejestr.io)

---

## Name of the initiator:

ePaństwo Foundation

---



## Major aim of the initiative:

Rejestr.io is a tool created by ePaństwo Foundation and it is the biggest and most popular open data portal in the CEE region. It provides an easy and free access to a vast amount of public data including business data, information on public tenders and more. It provides citizens and journalists with updated and users friendly data on public officials and business representative. It was established in 2012 under the name [mojepanstwo.pl](https://mojepanstwo.pl) and later evolve to [rejestr.io](https://rejestr.io) more concentrated on business registries and public procurement databases.

## Major details about the platform:

Two main components of the platform are:

**I. Data from National Business Registry** including information on companies, non-governmental organizations and other. Each organization from the registry has its profile page on which users can browse basic facts about the organization, as well as explore other data relevant to the organization, such as:

- ➔ people in organization's boards;
- ➔ connections with other organizations;
- ➔ financial documents and reports.



**II. Application for analyzing data on Polish public tenders.** On the one hand, it enables entrepreneurs to search for offers. On the other hand, it makes possible for activists to analyze who received largest contracts. Because of an integration with data from National Business Registry, it allows to analyze additional data and business connections of companies, which won tender offers.

### **Open Data component of the platform:**

It transforms pdf versions of excerpts from the official National Business Registry into open data enabling for interoperability with public procurement data or to establish and visualize connection between politicians and business representatives.

### **Innovation of the initiative:**

ePaństwo developed a special, unplanned feature of the Rejestr.io portal that turned out to be very important for the project and its users. It is called “Analysis”. It allows users to analyze content of documents in order to find information regarding public figures’ professional activities, connections, scientific achievements, stages of career, etc. Users can upload their documents (in the PDF format), which are then automatically processed (also using Optical Characters Recognition technology). At the end, users are able to see their documents with all mentions of public figures highlighted. They can then click highlighted names to get knowledge about specific persons from our databases.

### **Impact of the initiative:**

The tool is widely used by journalists and reporters. At least, two investigative books on public figures were mainly based on data from the portal. It is also the first source for looking for names of politicians and business representatives in case some important corruption or transparency topics arise in the public debate.

### **Some statistics:**

Data from the portal are presented around 5 times average weekly in mainstream media and nearly 40 in social media. The tool has more than 600k unique users per month.

---

**Name of the platform:**

[Sejmometr.pl](http://Sejmometr.pl)

---

**Name of the initiators:**

ePaństwo Foundation

---

**Major aim of the initiative:**

Application provides information about the work of the Parliament and MPs. It was established in 2009 and is constantly improving. The aim of the website was to support public with the user friendly information on the activities of the Polish Parliament.

**Major details of the platform:**

Application provides information about the work of the Parliament and MPs. It includes data about meetings of the Sejm, the latest bills and resolutions, MPs speeches, voting results, written questions and benefits registries. The application makes it easy to find MPs responsible for a user's constituency, based on a postal code. Every member of the Parliament has its profile page, with a feed of the newest data, which users can subscribe. Portal users can also send letters to MPs directly from their profile pages.

**Open Data component of the platform:**

The following information is available on the platform:

- ➔ Speeches at the Sejm (the lower chamber of the Polish parliament)
- ➔ Speeches at the Senate (the higher chamber of the Polish parliament)
- ➔ Voting results of the Sejm and Senate
- ➔ MPs' interventions and questions
- ➔ Bills and other documents presented to the Sejm and Senate

Thanks to transforming data from the official website to open data it allows for better search results and interoperability.

# CZECH REPUBLIC

---

## Name of the platform:

Guardian of the State - Hlídač státu

---

**Website:** <https://www.hlidacstatu.cz/>

---



## Initiator:

Michal Bláha (former successful IT businessman)

---

## Major details about the platform:

Guardian of the State consists of several services:

- ➔ Guardian of Contracts ("Hlídač smluv", [hlidacsmluv.cz](https://www.hlidacsmluv.cz/)) - the original and the most important service - it republishes the information from the Register of Contracts in a more user-friendly way (better search, etc.).
- ➔ Guardian of People aggregates public information about politicians and people connected with politics.
- ➔ Guardian of Public Contracts republishes the information about public contracts in a more user-friendly way.
- ➔ Guardian of [political] Sponsors aggregates information about sponsors of political parties and political campaigns.
- ➔ Guardian of Political Finances aggregates information from (compulsory) transparent accounts of political parties.
- ➔ Guardian of [public] Websites is regularly checking availability of the governmental websites.

The Guardian of the State started in 2016 as Guardian of Contracts. It reacted to the new law that required many public institutions to publish information about their contracts (above the level of about € 2000). However, the official website with these information was not very user-friendly, which opened the opportunity to provide better service.

It quickly became one of important sources for journalists reporting about many issues of public institutions. While the media is the main target group of the portal, it also serves wider group of civic activists, interested people in general - or also politicians and the public institutions themselves.

### Statistics:

The project attracts tens of thousands users every month.

---

### Name of the platform:

Election calculator - Inventory of votes

---

### Website:

Czech version: <https://volebnikalkulacka.cz/>

English version: <https://electioncalculator.org/>



---

### Initiator:

KohoVolit.eu (NGO)

---

### Major details about the platform

Inventory of votes is a voting advice application based on the real votes from the parliament (using the real voting records) - the Czech parliament or European parliament (or other parliaments).

The users answer a set of questions (usually 20-40) and at the end, they get easy comparison between their opinions and the opinions of the members of the parliament, listing from the parliamentarians with highest match to the lowest match.

The applications are usually prepared before elections, when there is most public interest in the parliaments. Other period of the year, the platform prepares annual performance report of MPs.

The idea behind these applications is to show what is really done / voted in parliaments, which MPs vote with the consideration of users' viewpoints and generally provide more information about what is really happening in parliaments.

The Inventory of votes was first prepared in Czechia before the national parliamentary elections in 2006. Since then, more than 10 different Inventories of votes were prepared - for Czech parliament or for European parliament and even for some municipal assemblies. All of them were done by KohoVolit.eu, sometimes with a cooperation with other NGOs or experts.

### **Open Data component of the platform:**

This kind of voting advice applications rely heavily on accessibility of the data - the roll-call voting results. These are available for the Czech parliament as an open data (database dumps): <http://www.psp.cz/sqw/hp.sqw?k=1300>

The application was also prepared in other countries using the same open-source software - in Slovakia, Hungary, Poland and Serbia.

The source code of the application: <https://github.com/KohoVolit/electioncalculator.org>

# SLOVAKIA

---

## Name of the platform:

Open Public Contracts in Slovakia

---

## Platform Website:

<https://tender.sme.sk/en/>

---

 Open public contracts in Slovakia

## Initiators:

Transparency International Slovensko (NGO), Datlab (software / data company), SME (media, web provider)

---

## Major details about the platforms:

The Open Public Contracts started in 2010. As the official data was not provided in a user-friendly way, this project aimed to fill this gap. The project was one of the first projects in the world that published similar data in such big amounts and in such level of user-friendliness.

It allows to control both side of the public contracts - buyers (the public institutions) and suppliers. It is also an important source of information for journalists.

The project uses data from the official website that publishes the information about public contracts:

<https://www.uvo.gov.sk/vestnik-590.html> The official website does not provide open data, it needs to be scrapped. However, the project Open Public Contracts publishes the information as basic open data (database dumps) for others to use.

---

**Name of the platform:**

Price of the State

---

**Platform website:**

<http://www.priceofthestate.org/>

---

 PRICE OF THE STATE

**Initiators:**

Institute of Economic and Social Studies (NGO), SME (media)

---

**Major details about the platforms:**

The project started in 2007 visualizing the national budget for the first time, similarly as in other countries. This information about budget is usually scattered across ministry websites and hidden in various government documents. It is hard to find the information regarding public finance from the official documents.

The Price of the State concentrates detailed data on revenues and expenditures for public administration. It also show how the values are changing over time. Another important goal of the project is to inform the general public of the dimensions of Slovak public finance. Therefore, it uses both the real “big” numbers and recalculates them “per capita”, so the users get better understanding of the numbers.

The information sources are official data published on the websites of ministries and public institutions.

# BIBLIOGRAPHY AND ADDITIONAL REFERENCES

1. Open Knowledge Foundation. The Open Data Handbook. Available at: <https://okfn.org/opendata/>
2. Larry Lessig on Open Government Data principles. Available at: <https://opengovdata.org/>
3. Vivek Kundra's 10 Principles for Improving Federal Transparency. Available at: <https://bit.ly/37b8Y4o>
4. United Kingdom's Public Data Principles. Available at: <http://data.gov.uk/library/public-data-principles>
5. Data Journalism. MaryJo Webster's training materials. Available at: <http://mjwebster.github.io/DataJ/>
6. Kuang Keng & Kuek Ser. Best Practices for Data Journalism. Available at: <https://bit.ly/2ESVlej>
7. *Data Journalism Manual*. EDECA. Available at: <http://www.odcanet.org/data-journalism-manual/>
8. *Data Journalism*. Google News Initiative. Available at: <https://bit.ly/2Zs8m7L>
9. Paul Bradshaw. *Finding Stories in Spreadsheets*. Available at: <https://leanpub.com/spreadsheetsstories>
10. Lawrence Marzouk & Crina Boros. *Getting Started in Data Journalism*. Available at: <https://bit.ly/2QkPN1I>
11. Jānis Gulbis. *Data Visualization - How to Pick the Right Chart Type?* Available at: [https://eazybi.com/blog/data\\_visualization\\_and\\_chart\\_types/](https://eazybi.com/blog/data_visualization_and_chart_types/)
12. Anna Vital. *How To Think Visually Using Visual Analogies - Infographic*. Available at: <https://bit.ly/2SrjuAz>
13. Jami Oetting. *Data Visualization 101: How to Choose the Right Chart or Graph for Your Data*. Available at: <https://bit.ly/2EVEInC>
14. Practical Toolkit of Data Journalism: <https://datalab.ge/ge/toolkittext/toolkit/3/>
15. The list of additional toolkits:
  - I. [https://www.journaliststoolbox.org/2019/11/21/online\\_journalism/](https://www.journaliststoolbox.org/2019/11/21/online_journalism/)
  - II. <https://ksj.mit.edu/data-journalism-tools/>





**INSTITUTE FOR DEVELOPMENT OF FREEDOM OF INFORMATION (IDFI)**

20, T. Shevchenko Street 0108. Tbilisi Georgia  
Phone: + 995 32 2 92 15 14 • Email: [info@idfi.ge](mailto:info@idfi.ge)

**[WWW.IDFI.GE](http://WWW.IDFI.GE) • [WWW.DATALAB.GE](http://WWW.DATALAB.GE)**