Visual Data Analysis for EU Public Sector Data usingPython app MyDataApp

Georgi Petrovski[1]Snezana Savoska[2], Blagoj Ristevski[3], Andrijana Bocevska[4], Ilija Jolevski[5] and Natasha Blazheska-Tabakovska[6],

[1,2,3,4,5,6]*University "St Kliment Ohridski"-Bitola, Faculty of Information and Communication Technologies – Bitola, ul. Studentska bb 7000 Bitola, R. North Macedonia*

georgi.petrovski@uklo.edu.mk; snezana.savoska,@uklo.edu.mk;blagoj.ristevski@uklo.edu.mk; andrijana.bocevska@uklo.edu.mk;ilija.jolevski@uklo.edu.mk;natasa.tabakovska@uklo.edu.mk

Abstract:

The growing number of data collected in the public sector should contribute to the analysis and detection of situations and anomalies and provide appropriate measures to deal with them. This data analysis process generally requires a lot of time and resources because usually the data being analyzed is of the big data type. According to the new trends, it is necessary to analyze it according to the methods of big data analysis, to prepare the data in advance depending on the intentions of those responsible for decision making and to visualize it to achieve the greatest eloquence of the data for analysts and decision-makers. Due to all these findings, this paper creates an application solution for the analysis of data for the public sector to facilitate the process of loading data sources, using them, preparing them for visualization and performing visual data analysis of the data of interest with the purposeof providing decision-makers withthe information they need. For this aim, a user-friendly application form for analysis was created and used for the visualization of data from the public sector in the EU. The obtained results are analyzed to highlight the pros and cons of the software solutions.

Keywords:

Big data, visual data analysis, public sector data, visualization

1. Introduction

Data in the public sector for each country and especially for the EU are collected in many ways, placed in different databases and those needed for statistics are placed in the Eurostat database and are publicly available. All collected data in the public databases of the countries and EU contain data on many entities, with the necessary attributes and values for the period for which they are collected, analytically or synthetically. Because of this, they have characteristics of big data and their analysis is therefore quite demanding and requires the use of methods and techniques for big data visualization [1]. However, the process of visual data analysis itself is not simple and requires the use of various big data analysis tools[2], data access tools, their preparation for visualization, sometimes normalization and some kind of categorization and then visualization tools [3, 4]. Therefore, knowledge of all these parts of the analysis process is required.

The visual analysis of big data from the public sector with the support of algorithms and advanced analytics helps decision-makers perceive the situation, identify solutions, create forecasts for future events and improve the lives of citizens in general[5]. One way to explore this is to use large data from the statistics of the EU members (Eurostat) on a certain social topic, which are usually entered by all institutions of a public nature on the territory of Europe.

Firstly, the project task should be set in order to know what data to look for and what to analyze, which is the job of decision makers and analysts. It starts from the existence of a problem or need of a community. After that, it is necessary to make assumptions, to ask questions that provide information about what knowledge the data carries and what can be learned from it, what problem occurred, why that problem occurred, and predict what is to be expected in the future if (no) changes are performed [6].

The next stage is data processing. Inthe beginning, the processing consists of the analysis of the data placed in tables that are usually organized in a spreadsheet. However, sometimes the data comes in chaotic and messy and is hard to understand [7]. Such data structures are mostly in .csv (Comma Separated Values) and JSON format. They can easily become readable with today's software tools at our disposal. They are compared with the data we need for one of the set goals.

After the analysis of the data sources, the next step is the use of various data analysis techniques and methodologies (application of visualization techniques and methods, various graphs), frameworks and libraries that are specially designed for visual data analysis or data visualization [8].

It is considered that the support of big data visual data analysis is great today, especially with the development of many interactive tools for big data visual data analysis [2]. By using the Python programming language as a specialized language for data analysis, software for data processing and visual data analysis can be developed [9]. Therefore, in this paper, Python was used to create a custom tool for visual data analysis (VDA).

The paper is structured as follows. The first section after the introduction is devoted to the preparation of data for VDA, while the next section describes the objectives of VDA for the public sector. Next, the created software solution MyDataApp is describedalong with its advantages and disadvantages. Subsequently, the process of visualization of big data from the data for the public sector is shown with the proposed application, followed by an analysis of the eloquence of the given visualizations. Finally, a conclusion with insights for further research is given.

2. Big data and its preparation for visual data analysis

Because big data has 6V characteristics (Volume, Variability, Velocity, Veracity, Value, Volatility), its analysis is complex. Usually, the data can be stored in different formats, such as tables, databases or streaming data [1]. In all known situations, the preparation of data for visualization may be different. Usually, the metadata is read from the table headers or displayed accordingly. In the preparation, it is essentialto detect the necessary columns - metadata and discard the others, if needed create new tables with parts of the data or tailor them according to the needs of the analysis [7].It is also important to provide the missing data, usually employing machine learning methods [10]. Sometimes shorteningof field lengths or transposing rows and columns should be applied. The cardinality of some data should usually also be reduced or normalized and some additional actions such as filtering, sampling or binned aggregation have to be provided [7]. Because the research is part of the bigger project for VDA of public data, for now we do not focus on the metadata description in this paper.

The methods used for big data analysis include data mining, Prediction analysis, Text analysis, Voice analysis, Statistical analysis and others. Generally, big data is stored in data warehouses or cloud storages [10]. The process of visual data analysis itself can be done in many ways. The best way is to have a big picture of big dataand then provide visual data analysis of the data on demand[11]. For analysis of the intended data and obtainingvital statistics [12, 13] from EU data,the MyDataApp application was created using Python and its Pandas and Plotly libraries, i.e. Plotly express [9]. For this purpose, both libraries are installed and called by the application in the big data VDA process. Pandas is a library that allows viewing the data in the application itself, from the prepared table, without the need of using another visualization program. With this feature, the user will be able to view the data from the tables more easily.Plotly express in collaboration with Pandas displays the data in readable and understandable graphs of different characters with automatically generated selection options for each selected graph [9]. Charts are fully interactive and provide many options including zooming, full-screen display, selecting specific data for easier viewing and downloading the chart as animage in different formats. In such a way, those charts can be used for presentations, meeting reports, etc.

The intention of creating MyDataApp tool was to provide a desktop VDA for selected data from public EU databases. But the usage of this tool is not limited only for this – in fact, it also provides usage for VDA on data on some URL location (for example, online VDA can be provided as in the given link: https://raw.githubusercontent.com/Lexie88rus/bank-marketing-analysis/master/bank.csv).

3. VDA Objectives for Public Sector

The public sector is becoming aware of the value of big data and its role in society. Governments collect large amounts of data with their daily activities, such as wages, contributions, taxes, data on the performance of state and health systems, traffic data and many others. Of course, it takes into account the overall socio-economic situation, technological trends and the increase in the need for medical and social services, especially in this time of a pandemic. The potential benefits that governments can get from big data are trend detection, data transparency, citizen sentiment analysis, citizen segmentation and personalization, financial and tax analysis, smart cities, information security data, and others [3, 5].

In order to collect all this data and achieve the benefits, the collected data "goes" through a cycle called "The government data value cycle"which consists of data generation and collection, preservation, security and processing, sharing, selecting and publishing and their reuse as shown inFigure 1, [14]. The purpose of these analyzes is broad and certainly necessary to obtain valid information for taking future actions and adopting development strategies.

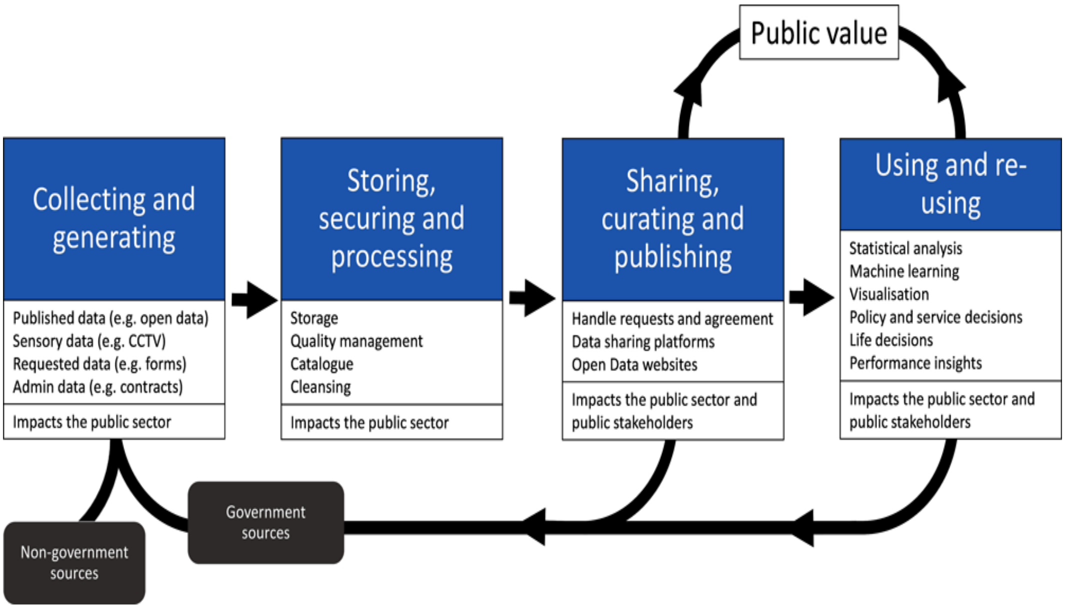

Public value
Collecting and generating
Published data (e.g. open data)
Sensory data (e.g. CCTV)
Requested data (e.g. forms)
Admin data (e.g. contracts)
Impacts the public sector
Storing, securing and processing
Storage
Quality management
Catalogue
Cleansing
Impacts the public sector
Sharing, curating and publishing
Handle requests and agreement
Data sharing platforms
Open Data websites
Impacts the public sector and public stakeholders
Using and re-using
Statistical analysis
Machine learning
Visualisation
Policy and service decisions
Life decisions
Performance insights
Impacts the public sector and public stakeholders
Government sources
Non-government sources

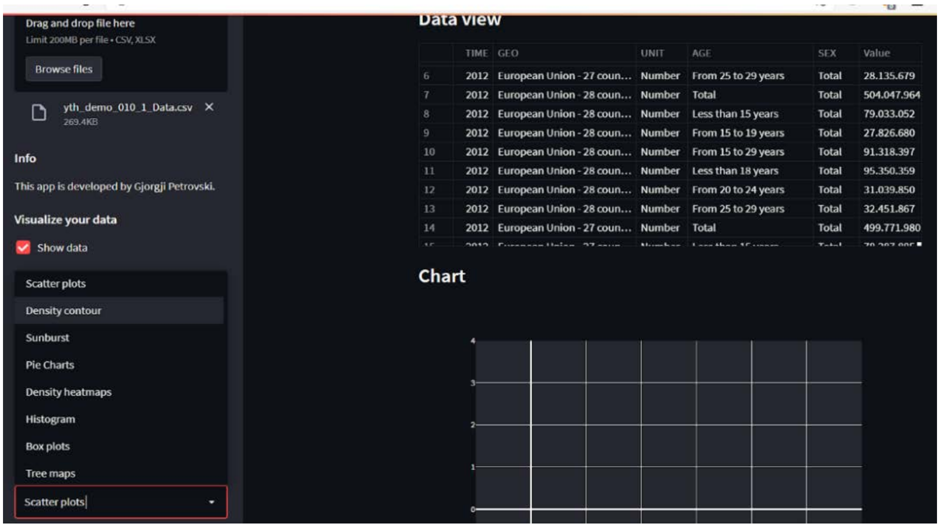
Figure 1: – Cycle of government data valorization [14].

4. Creating VDA of public sector data with the MyDataApp tool

For VDA purposes, the MyDataApp application was created with the Python programming language for EU public sector data. It is a web application designed for the review, analysis and visualization of data for all kinds of data from different topics. This application allows an easy way, with a few clicks, to display the data from the source visually with different graphs, as shown in Figure 2.

On the left side, the user can select a data source in .csv or .xlsx format by clicking on "Browse files". The maximum document size limit that the application can display is 200MB which means that larger files need to be adjusted to accommodate the VDA data in this memory frame.Once the data source is selected, the data is loaded and other options automatically appear that allow the user to view the data as a table, easily see which columns to use and input into the VDA, and choose which type of chart to use. By selecting the data range and the visual display type, the application reads all the columns and displays them with checkbox and dropdown options.

MyDataApp is a web application entirely programmed in the Python programming language. It is built with Streamlit which is a Python framework installed and called a library in the code.The application, in order to enable the analysis and overview in the form of a table of data from different formats such as.csv, JSON, SQL databases of tables or queries and Microsoft Excel, uses the Pandas library which is a package written in Python, which is installed and called as a library in the code. The visualization of data using graphs is made possible by the Plotly library for Python and Plotly Express, which is a high-level module built into the library. Plotly allows visualization of data and graphical objects. Plotly and Plotly Express are installed separately and only Plotly Express is called in the code as px. In the application with Plotly Expressseveral types of graphical displays such as Scatter plots, Histogram, Box plots, Sunburst, Tree maps, Pie Charts, Density contour, Density heatmaps and Violin plots can be used, Figure 2 and Figure 3.



yth_demo_010_1_Data.csv
269.4KB

Info

This app is developed by Gjorgji Petrovski.

Visualize your data

☑ Show data

Scatter plots

Data View

	TIME	GEO	UNIT	AGE	SEX	Value
6	2012	European Union - 27 coun...	Number	From 25 to 29 years	Total	28.135.679
7	2012	European Union - 28 coun...	Number	Total	Total	504.047.964
8	2012	European Union - 28 coun...	Number	Less than 15 years	Total	79.033.052
9	2012	European Union - 28 coun...	Number	From 15 to 19 years	Total	27.826.680
10	2012	European Union - 28 coun...	Number	From 15 to 29 years	Total	91.318.397
11	2012	European Union - 28 coun...	Number	Less than 18 years	Total	95.350.359
12	2012	European Union - 28 coun...	Number	From 20 to 24 years	Total	31.039.850
13	2012	European Union - 28 coun...	Number	From 25 to 29 years	Total	32.451.867
14	2012	European Union - 27 coun...	Number	Total	Total	499.771.980

Figure 2: Display of the data from the source visually with different graphs in MyDataApp.

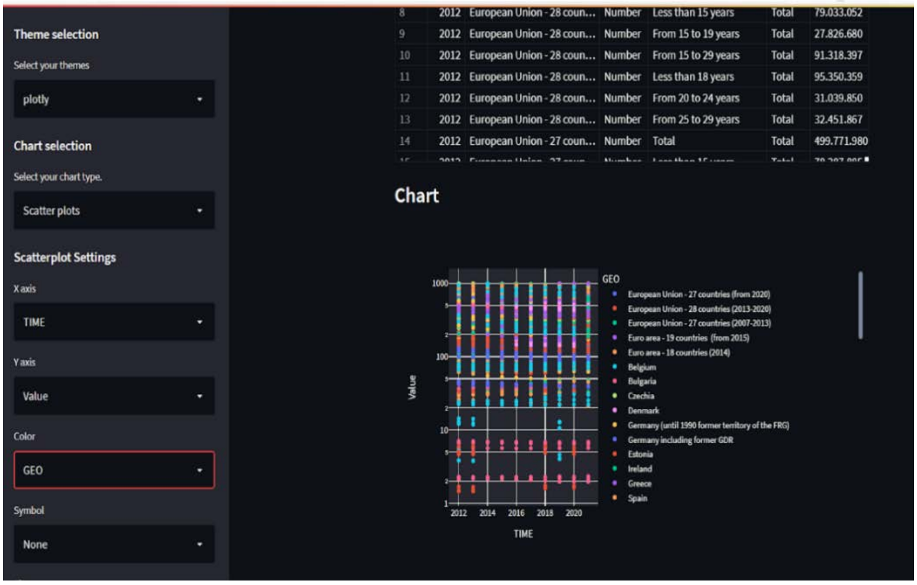


Chart
GEO
European Union - 27 countries (from 2020)
European Union - 28 countries (2013-2020)
European Union - 27 countries (2007-2013)
Euro area - 19 countries (from 2015)
Euro area - 18 countries (2014)
Belgium
Bulgaria
Czechia
Denmark
Germany (until 1990 former territory of the FRG)
Germany including former GDR
Estonia
Ireland
Greece
Spain
Value
1000
100
10
1
TIME
2012
2014
2016
2018
2020

Theme selection

Select your themes: plotly

Chart selection

Select your chart type.: Scatter plots

Scatterplot Settings

X axis: TIME

Y axis: Value

Color: GEO

Symbol: None

8	2012	European Union - 28 coun...	Number	Less than 15 years	Total	79.033.052
9	2012	European Union - 28 coun...	Number	From 15 to 19 years	Total	27.826.680
10	2012	European Union - 28 coun...	Number	From 15 to 29 years	Total	91.318.397
11	2012	European Union - 28 coun...	Number	Less than 18 years	Total	95.350.359
12	2012	European Union - 28 coun...	Number	From 20 to 24 years	Total	31.039.850
13	2012	European Union - 28 coun...	Number	From 25 to 29 years	Total	32.451.867
14	2012	European Union - 27 coun...	Number	Total	Total	499.771.980

Figure 3: Selection of the type of graphicaldisplaysfor data visualization.

MyDataApp is a simple, easy-to-use web application with a very intuitive user interface that does not require the user to have advanced computer skills. The main advantage is that when the data source is selected, it is not necessary to consider in which format it is given, but with the selection, the formats that the application supports are automatically displayed. By selecting, the data is displayed in a table and you can select the number of shown rows.

The advantages of MyDataApp can be listed asan easy and simple user interface, available user manual for the web application, easy data search procedure, fast reading of data and their display in a table for easier orientation, easy selection of graph type, ability to manipulate charts, ability to ignore empty columns or columns with NULL values, listing all options for each specific chart, automatic display of all columns from the selected chart, enabling communication with individuals and companies via email directly from the application. Also, MyDataAppis ofopen-source nature, allows possibilities for online work, provides desktop support and is also available as a mobile application.

The disadvantages of MyDataApp can be listed as the need forconnection to the Internet to use the application and the graphs, limited memory capacity for displaying a larger amount of data when MyDataApp is used for desktop VDA, inability to display multiple data structures, inability to download large data via the Internet due to a lack of memory in local computer and API, lack of manipulating and transforming data in tables and ability to display only one chart at a time. All these shortcomings are for this version of the application and some of them are planned to be overcome in newer versions.

Figure 4shows a visualization with a scatterplot matrix graph created with MyDataApp with a color representation of demographic balance and growth rate in Europe based on data obtained from Eurostat. The x-axis represents the years, while the y-axis the values. The legend shows the states and types of civil national level. If the cursor is placed on any of the points, a tooltip with additional information will be displayed.

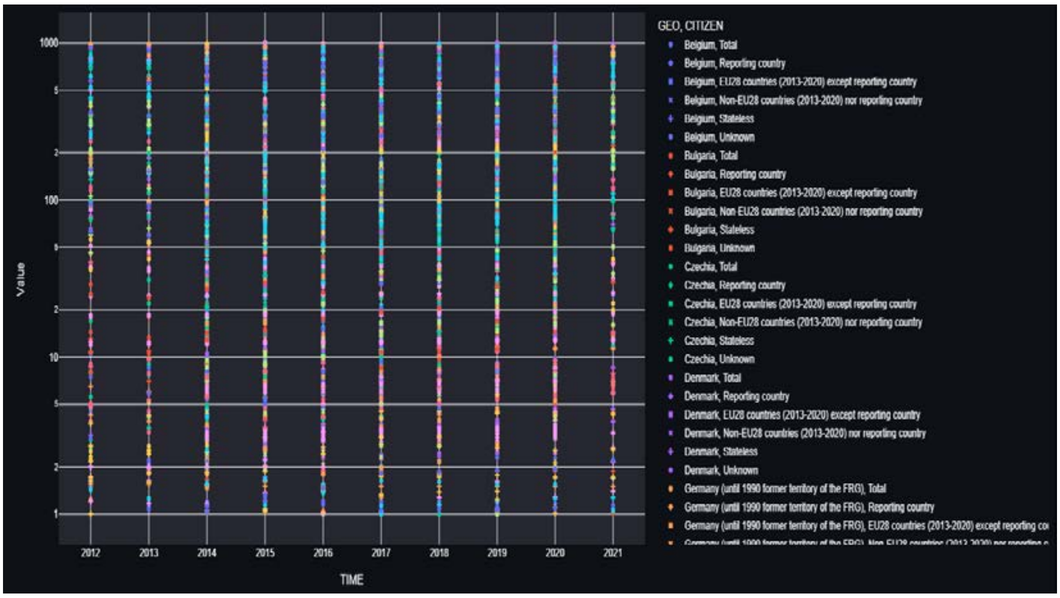


Figure 4: Visualization with a scatterplot matrix graph created with MyDataApp

Analysis of data for daily cigarette consumption by gender, country and education was performed. In Figure 5, the states are shown on a legend in purple, orange and green. The size of each bubble depends on the size of the value. The legend is interactive and therefore allows selecting a country and filtering the data only for that country. It has a tooltip property that is displayed by placing the cursor (crosshair) on the point in order to obtain additional information.

The available methods and techniques can help to visualize very complex data in an inventive way [8]. Figure 6shows a visualization with a density heatmap made with MyDataApp based on data by Eurostat. Data visualization of the frequency of smoking is in regard to the country, gender, age, years and level of education of the citizens. The values are displayed on a heat scale where each value has a specific color according to the magnitude of the value. The highest value is shown in light yellow and the lowest value is shown in blue. In order to get additional information, it has the tooltip property, which displays all the data for certain metadata that were previously selected when creating the graph.

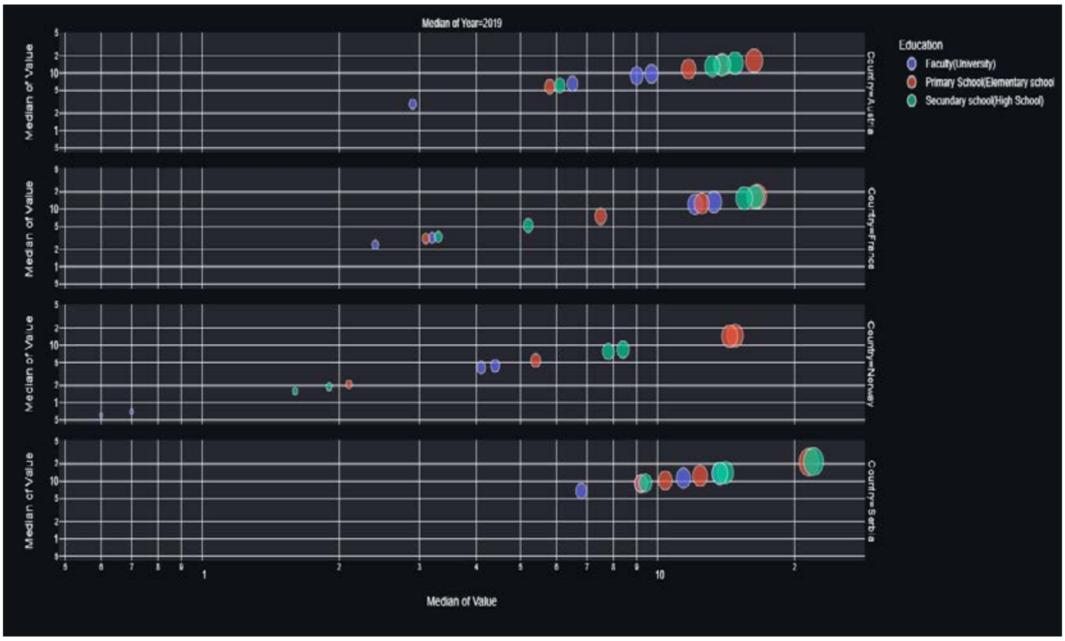


Figure 5: Data visualization with scatterplot graph and bubble property created with MyDataApp.

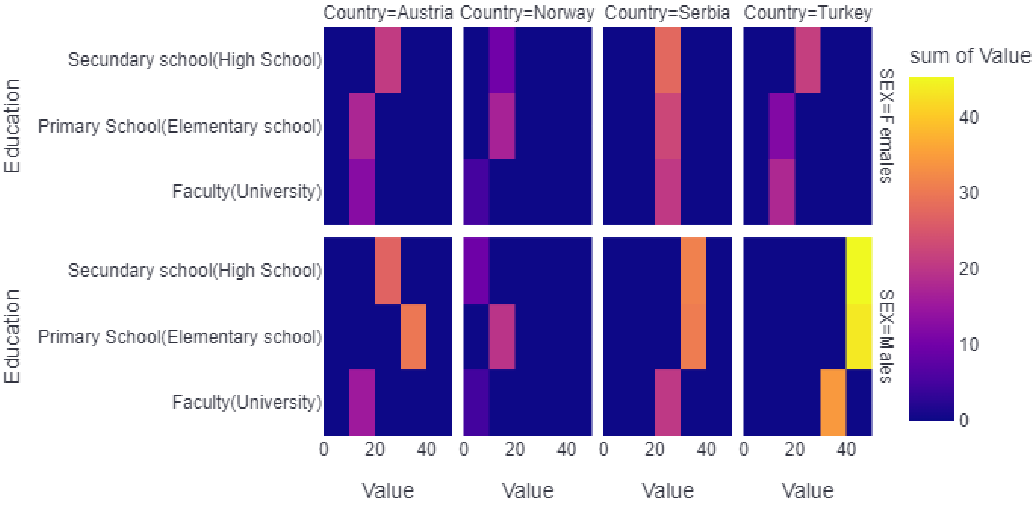


Figure 6: Data visualization of frequency of smoking with density heatmap technique.

Figure 7shows a visualization with a matrix of the density contour technique made in MyDataApp, where the four states are shown in columns.For each state, values are represented in a column expressed in percentages (%) on the x-axis, while the y-axis corresponds to the degrees of education. States are shown in a legend with a specific color for each state in blue, orange, red, green and purple. This chart is not interactive and therefore the data is displayed by hovering over the tooltip property, while the legend is interactive.

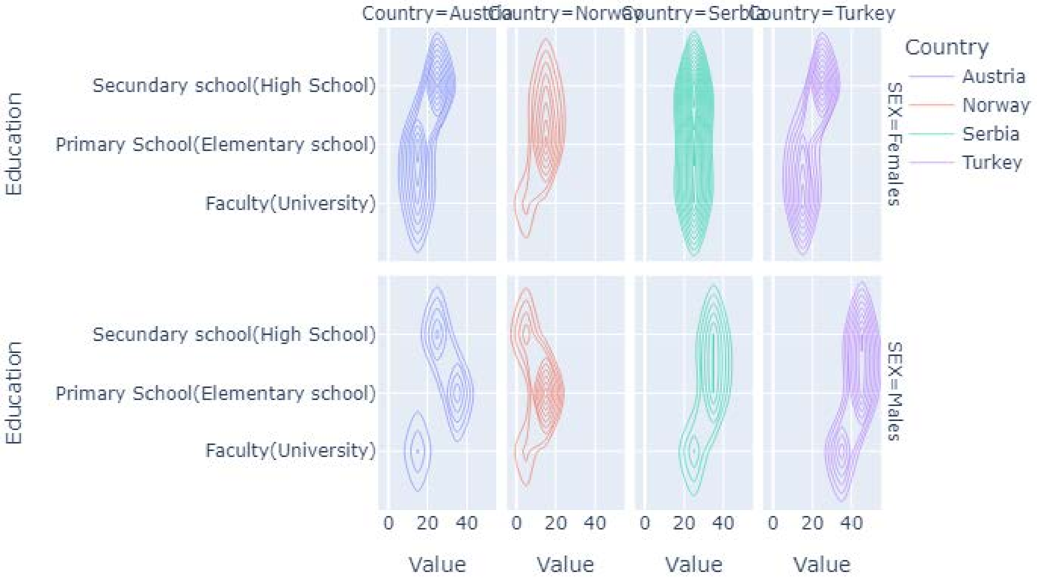


Figure 7: Data visualization of frequency of smoking with a matrix of the density contour technique

5. Conclusions

This paper presents one way of VDA intended for the analysis and processing of large amounts of data for the needs of the public sector. It can be mentioned that the processing of this data is the general responsibility of the government's analysts and decision-makers and in this way, the analysis of the data for the benefits ofthe citizens should be consistently analyzed and observed. It can be concluded that the existence of a huge amount of data requires a very extensive and large analysis that needs to process large amounts of data from which many things can be learned about the situation in society. With these analyzes and decisions about future strategies and policies, the life of the population can be greatly improved inall aspects of living, such as income, health system, social and pension insurance, employment, the standard of living, product prices, etc. Many activities are known through the analysis of large amounts of data that are aggregated and processed every day by institutions and sent to statistical institutions, such as Eurostat. Data analysis with data visualization is the most preferred and popular method of data analysis that helps decision makers and analysts to perceive situations realistically and take measures to solve emerging problems and improve situations. Since VDA itself is a complex process that allows one to see the big picture and then do on-demand analytics, the MyDataApp Python web application was created for the purpose of VDA on public sector data and demonstrated innovative methods for analysis of big data.Desktop usage of the application is with limited data processing possibilities but the usage for VDA from some URL do not limit the data amount for analysis. The MyDataApp offers many advantages, among which: easy and simple user interface, available user manual for the web application, easy data search procedure, fast reading of data and their display in a table for easier orientation, easy selection of graph type, ability to manipulate charts, ability to ignore empty columns or columns with NULL values, listing all options for each specific chart, automatic display of all columns from the selected chart, enabling communication with individuals and companies via email directly from the application. Using this application, data visualizations with several different techniques have been made. This research is not limited only to these techniques and methods and provides wide possibilities, especially with usage of big data analysis from some URL location. In this paper, we focused on the VDA for desktop usage of MyDataApp.

As future directions, we can state that we will work on improving the solution by increasing its computational power and expanding the amount of data that can be visualized using URL location, as well as obtaining new types of visualizations that are more expressive compared to the standard known visualizations. Knowing the metadata for data deeply will provide better data representation and understanding of visualization results. Many different data types have to be extracted, transformed and loaded in order to prepare for visualization aiming to gain better VDA and data representation, especially when big data has to be taken into consideration. This task will be also one of our future works.

References:

[1] C. Brooke, “Big Data: What Is It and How Does It Work?”: https://www.business2community.com/big-data/big-data-what-is-it-and-how-does-it-work-02265540last accessed on 14.09.2022.

[2] N. Bikakis, “Big data Visualization Tools”: https://arxiv.org/abs/1801.08336last accessed on 14.09.2022.

[3] R. Munne´, “Big Data in the Public Sector”: https://link.springer.com/content/pdf/10.1007%2F978-3-319-21569-3_11.pdf last accessed on 14.09.2022.

[4] A. Syed Mohd, G. Noopur, N. Krishna Gopal, L. Kumar Rakesh, “Big data visualization: Tools and challenges”, https://ieeexplore.ieee.org/abstract/document/7918044, last accessed on 31.08.2022.

[5] O. Jones, A. Evans, J. McQueen, T. Bradtke, “Big data, big outcomes: how analytics can transform public services and improve citizens’ lives”file:///C:/Users/user/Downloads/ey-future-of-gov-digital-analytics-report.pdf, last accessed on 10.09.2022.

[6] OECD, “The application of data in the public sector to generate public value”: https://www.oecd-ilibrary.org/sites/1ab27217-en/index.html?itemId=/content/component/1ab27217-en, last accessed on 10.09.2022.

[7] R. Agrawal, A. Kadadi, X. Dai, F. Andres, “Challenges and opportunities with Big data visualization”: https://dl.acm.org/doi/abs/10.1145/2857218.2857256last accessed on 31.08.2022.

[8] E. Yur'evich Gorodov, V. Vasil'evich Gubarev, “Analytical review of data visualization methods in application to Big data”: https://dl.acm.org/doi/abs/10.1155/2013/969458last accessed on 10.09.2022.

[9] Python Pandas Tutorial, W3C, https://www.w3schools.com/python/pandas/default.asp last accessed on 18.9.2022.

[10] A. L'Heureux, K. Grolinger, H. F. El Yamany, M. A. M. Capretz, Machine learning with big data: Challenges and Approaches, IEEE Access, 2017, PP(99):1-1, https://www.researchgate.net/publication/316448042_Machine_Learning_With_Big_Data_Challenges_and_Approaches.

[11] B. Shneiderman, "The Big Picture for Big data”: Visualization https://www.researchgate.net/publication/260211777_The_Big_Picture_for_Big_Data_Visualization, last accessed on29.01.2022.

[12] N. Tyagi, “What is Vital Statistics? Types, Uses and Examples”: https://www.analyticssteps.com/blogs/what-vital-statistics-types-uses-exampleslast accessed on 18.9.2022

[13] R. A. Israel, “Vital Statistics, Overview”: http://www.medicine.mcgill.ca/epidemiology/hanley/c609/Material/VitalStatisticsEoB.pdf, last accessed on 18.9.2022.

[14] Big data in the public sector – Challenges and Importance, https://ebusinesstalks.com/big-data-in-public-sector-challenges-importance/ last accessed on01.09.2022.