CroStats — visualization of population in Croatia

V. Perković, F. Belac, K. Boroš, M. Kos, L. Humski, D. Pintar
University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia

***Abstract*—Population census is often a popular topic in everyday conversations. After extensive research, a conclusion was made that currently no services exist that offer a convenient, easily interpretable information about the census. Our solution for this problem is CroStats – a simple web interface for population census data visualization intended for general public. The central feature of the CroStats is its intuitive and simple graphical display. It analyzes the most important categories of the population census by county such as population, age, birth rate and mortality. All this information is displayed on an interactive map of the Republic of Croatia. Extra features offered are graphs of changes by year and interesting historical and demographic facts about counties that the broader audience may have interest in.**

Keywords*—*Visualization, population census, graphs, web interface, geography, demography.

I. INTRODUCTION

Population census data is a crucial aspect of understanding the demographic trends and patterns of a country. When it comes to understanding large amounts of data, people tend to rely on simple and clear visualizations. While various static visualizations may be informative and intuitive, they do not offer a way to easily adapt or expand on the depicted information.

In light of this issue, CroStats, a simple web interface for population census data visualization was developed.

The main focus of the application was to provide a user-friendly and intuitive graphical display of the most important categories of population census data, including population, age, birth rate, and mortality, presented on an interactive map of the Republic of Croatia. The web interface includes not only this graphical display but also graphs of changes by year and informative historical and demographic facts about counties, making the data more accessible and comprehensible to the general public.

Through the following sections, the paper will go over a brief history on population census, related work to CroStats, overview of the used technologies, a thorough overview of the application including examples of use and the application architecture, future work and a conclusion.

II. POPULATION CENSUS

The population census is designed to provide the government with insight into the number of residents and general demographic characteristics. It dates back far into history, with the first census being conducted approximately 2000 years ago by the Han dynasty in China. At that time, China had 57 million inhabitants, and today the number has grown significantly [1]. Much has changed since then, but population census is still used in almost all countries in the world. With the information collected through the population census, it is possible to track demographics and notice new demographic trends. In Croatia, the first population census was conducted by Emperor Joseph II in 1857 [2], and during the existence of the Republic of Croatia, three censuses have been conducted in 2001, 2011, and 2021, which is also the basis of CroStats. The data was collected from the State Bureau of Statistics [3] where it is available to everyone. Unfortunately, the data format is poorly designed, organized, and displayed. The data is altogether hard to read and makes it difficult for everyday people to navigate a large amount of numbers. Sorting the data logically was a challenge due to these shortcomings, but in the end it was successfully resolved.

III. RELATED WORK

Eurostat Regional Yearbook [4] is an example of a web interface that is available to everyone, displaying a graphical representation of Europe with data such as the median age by county and predictions for the year 2050. CroStats differs in the fact that it currently focuses on Croatia. It will also have various categories of information to display. The only widely used such program that we found is Eurostat.

IV. TECHNOLOGY

While developing CroStats, a variety of technologies were used to create an interactive and informative experience for users.

First and foremost, R Shiny [5] was used as the framework for the application. Shiny is a web application framework for the R programming language. It allows users to build interactive web applications and dashboards that can be easily shared online.

The package is based on a reactive programming model, which means that the user interface (UI) components automatically update when the underlying data changes.

Shiny applications consist of two parts: the user interface (UI) and the server logic. The UI defines the layout and components of the web application, such as buttons, sliders, and text boxes. The server logic defines the behaviour of the application, such as data processing,

analysis, and visualization. The Shiny package provides a variety of UI components and server functions that can be used to build complex applications. One of the main advantages of Shiny is its ability to handle large datasets and complex calculations in real time. The reactive programming model allows the application to respond quickly to user inputs, making it ideal for data exploration and visualization. Shiny also supports a wide range of data visualization libraries, such as ggplot2, plotly, and tmap.

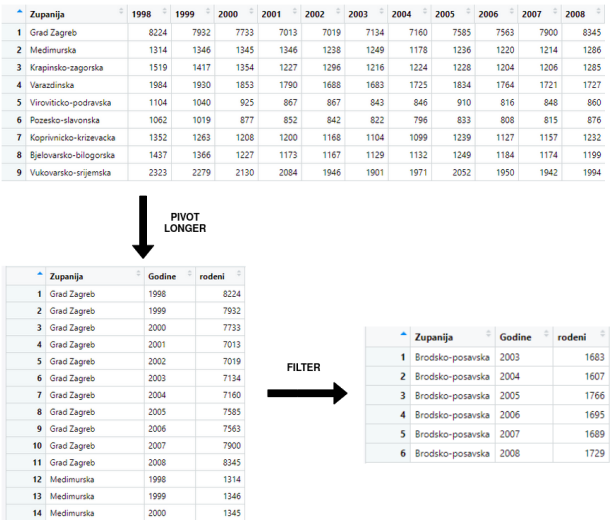
To display the data, several popular data visualization libraries in R were used. These included tmap [6], plotly [7], and ggplot2 [8].

Tmap is a package for R that provides a set of tools for creating thematic maps. Thematic maps are a type of data visualization that represents spatial data as colours, patterns, or symbols on a map. Tmap provides a simple syntax for creating thematic maps that can be customized with a variety of colour palettes, legends, and other design elements. One of the main advantages of tmap is its support for a wide range of map projections, including the popular Mercator and Robinson projections. Tmap also supports interactive map features such as zooming, panning, and tooltip displays, making it ideal for creating interactive data visualizations.

Plotly is a package for R that provides a set of tools for creating interactive data visualizations. Plotly allows users to create a wide range of charts and graphs, including scatter plots, bar charts, and line charts. The package also supports 3D plotting and animations. One of the main advantages of plotly is its ability to create interactive plots that can be easily shared online. Users can zoom, pan, and hover over data points to view more detailed information. Plotly also provides a wide range of customization options, including colour palettes, legends, and annotations.

Ggplot2 is a package for R that provides a set of tools for creating data visualizations using the grammar of graphics. The grammar of graphics is a theoretical framework for creating visualizations that consist of a set of rules for mapping data to visual elements, such as points, lines, and bars. One of the main advantages of ggplot2 is its ability to create complex visualizations with a few lines of code. The package provides a wide range of customization options, including colour palettes, legends, and themes. Ggplot2 also supports a wide range of data types, including categorical, continuous, and time-series data.

To prepare data for visualization, both the dplyr [9] and tidyr [10] libraries were used. The dplyr package provides a set of functions for manipulating and summarizing data, while the tidyr package is designed to reshape data into a tidy format for easier analysis and visualization. One of the functions from tidyr used in this analysis was "pivot_longer" (see figure 1), which is used to transform data from wide to long format, making it easier to work with and visualize. One of the main advantages of dplyr is its efficiency when handling large datasets. The package uses a set of optimized functions that can perform



	Zupanija	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
1	Grad Zagreb	8224	7932	7733	7013	7019	7134	7160	7585	7563	7900	8345
2	Medimurska	1314	1346	1345	1346	1238	1249	1178	1236	1220	1214	1286
3	Krapinsko-zagorska	1519	1417	1354	1227	1296	1216	1224	1228	1204	1206	1285
4	Varazdinska	1984	1930	1853	1790	1688	1683	1725	1834	1764	1721	1727
5	Viroviticko-podravska	1104	1040	925	867	867	843	846	910	816	848	860
6	Pozesko-slavonska	1062	1019	877	852	842	822	796	833	808	815	876
7	Koprivnicko-krizevacka	1352	1263	1208	1200	1168	1104	1099	1239	1127	1157	1232
8	Bjelovarsko-bilogorska	1437	1366	1227	1173	1167	1129	1132	1249	1184	1174	1199
9	Vukovarsko-srijemska	2323	2279	2130	2084	1946	1901	1971	2052	1950	1942	1994

	Zupanija	Godine	rodeni
1	Grad Zagreb	1998	8224
2	Grad Zagreb	1999	7932
3	Grad Zagreb	2000	7733
4	Grad Zagreb	2001	7013
5	Grad Zagreb	2002	7019
6	Grad Zagreb	2003	7134
7	Grad Zagreb	2004	7160
8	Grad Zagreb	2005	7585
9	Grad Zagreb	2006	7563
10	Grad Zagreb	2007	7900
11	Grad Zagreb	2008	8345
12	Medimurska	1998	1314
13	Medimurska	1999	1346
14	Medimurska	2000	1345

	Zupanija	Godine	rodeni
1	Brodsko-posavska	2003	1683
2	Brodsko-posavska	2004	1607
3	Brodsko-posavska	2005	1766
4	Brodsko-posavska	2006	1695
5	Brodsko-posavska	2007	1689
6	Brodsko-posavska	2008	1729

Fig. 1: Visualization of data wrangling

data manipulation operations much faster than the base R functions. Using the "filter" function from the dplyr package is more optimized for handling large datasets compared to the base R function "subset". This is because the "filter" function uses a lazy evaluation approach that only evaluates the necessary conditions to filter the data. In contrast, the "subset" function evaluates all conditions, which can be time-consuming with large datasets.

This makes it ideal for data analysis and visualization tasks that involve large datasets. Dplyr provides a set of functions that can be used for data manipulation and summarization tasks, such as filtering, selecting, grouping, and aggregating data.

In order to load and handle the data related to drawing the map, the sf library [11] was used. The sf package in R is a popular library for working with spatial data. It provides a set of functions and classes for manipulating, visualizing, and analyzing geospatial data in R. One of the main advantages of sf is its ability to handle a variety of spatial data formats, including shapefiles, GeoJSON files, and other common GIS data formats such as KML, TIGER and GML. It provides a set of functions for reading and writing spatial data from these formats, as well as for performing common spatial operations such as spatial joins and intersections.

Besides the technologies above, ShinyThemes [12] is a package that provides a variety of premade themes that can be easily applied to a Shiny app. It was used to add visual flair to the app. Lastly, Git [13] and Gitlab [14] were used for version control throughout the development process. Version control is an essential tool for any software development project, and it allowed for keeping track of changes made to the codebase, collaboration with other team members, and easily roll back changes if needed.

There are other data dashboarding frameworks available besides Shiny, such as Dash [15], Bokeh [16], and Streamlit [17]. However, Shiny was superior to the other

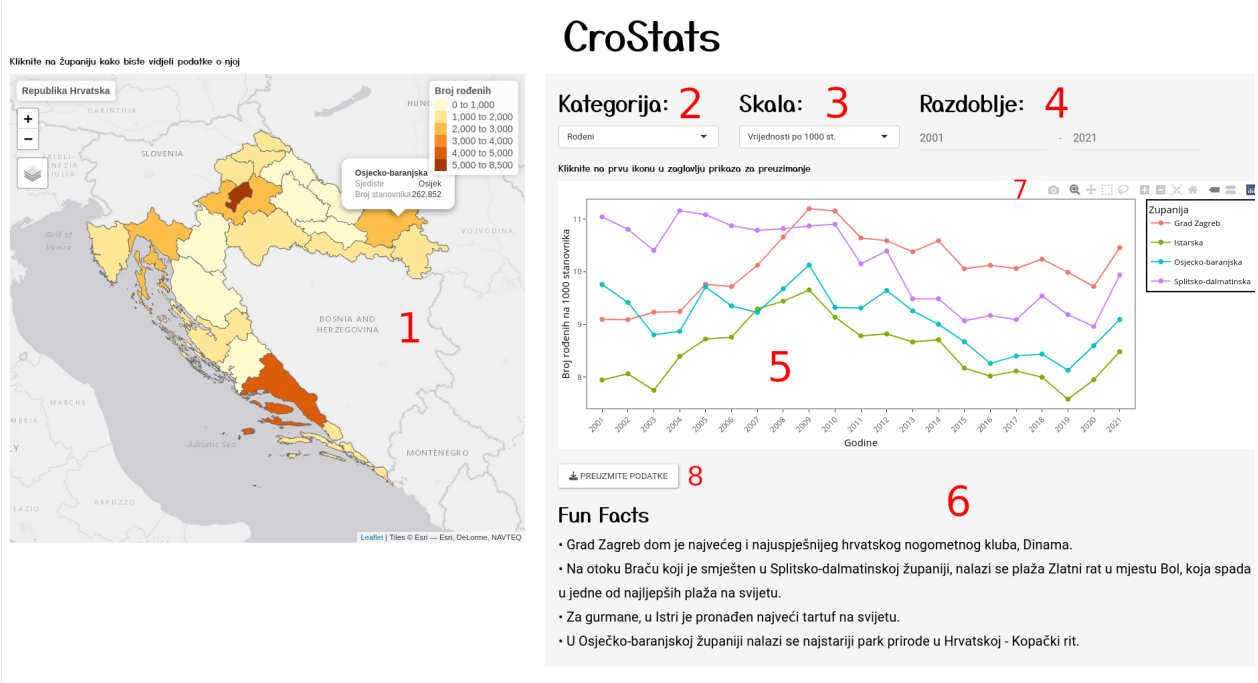


Fig. 2: CroStats app interface

options because it is the most seamlessly integrated with R ecosystem. Streamlit, Dash, and Bokeh are Python-based dashboarding solutions, running on the Tornado [18] and Flask [19] web frameworks.

Integration of R Shiny, tmap, plotly, and ggplot2 libraries with the dplyr package enabled the development of a functional and visually appealing application for data visualization and analysis. These technologies provided a range of features and customization options that allowed for the manipulation and summarization of data, filtration and selection of relevant data points, and creation of informative and visually appealing maps and graphics. The result was an application that provided an intuitive user experience and showcased the data in a visually appealing and informative way.

V. APPLICATION OVERVIEW

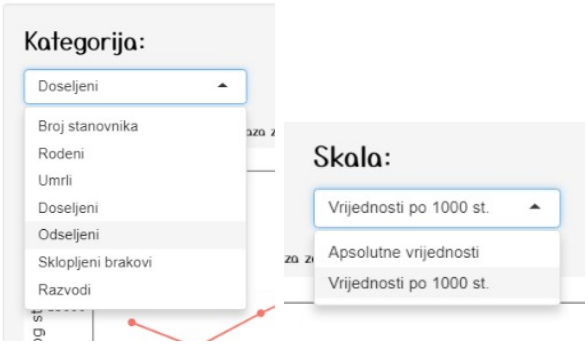
The CroStats application was realized as a web interface.

The application's interface features a map of the Republic of Croatia on the left side (see figure 2, mark 1), while on the right side, there are two drop-down menus for category and scale selection and a range selection for a specific period (see figure 2, marks 2, 3, 4 and figure 3). Scatter plots of the selected county can be viewed below the drop-down menus and range selection (see figure 2, mark 5), along with historical and demographic facts about the selected county (see figure 2, mark 6).

The application allows for the selection of any county on the map by clicking with a mouse. The selected category, scale and period for the selected county will then be displayed on a scatter plot. Up to four counties can be selected for comparison, and four scatter plots with accompanying facts will be displayed. The map of the Republic of Croatia will be re-rendered and coloured based on the selected category each time a different category is chosen (Number of inhabitants, Inhabited population, Migrated population). If a fifth county is selected, the earliest county selected will be replaced, so that there is always a display of up to four county trends on the graph with accompanying facts.

The application offers the functionality to save the graph of user's interest (see figure 2, mark 7 and figure 4) and data (see figure 2, mark 8) to their device storage so that the users can download them locally as a CSV file containing the current category.

Downloading graphs and data gives the users the flexibility to customize the information in their reports, presentations, and projects.



(a) Category selection (b) Scale selection

Fig. 3: Selection

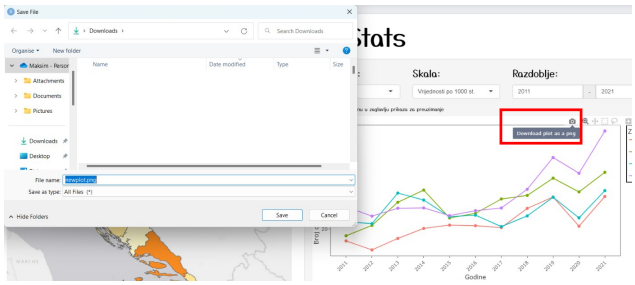


Fig. 4: Graph download

The application provides a platform for examining population data and gaining new insights for anyone with a curious interest in Croatia's population patterns.

A. Architecture

The architecture of the system is designed to support the creation of an interactive web application using the Shiny framework (see figure 5). The Shiny application consists of two main components: the Shiny server and the Shiny UI. The Shiny server is responsible for processing user requests, executing R code, and returning the results to the user's web browser. It handles all the logic and calculations required for data analysis and visualization. The Shiny UI is responsible for defining the layout and visual elements of the web application. It allows users to interact with the data and provides a graphical interface for displaying the results of the analysis. The UI is defined using R code, which creates HTML, CSS, and JavaScript code that is sent to the user's web browser. Data for processing and visualization are sourced from an Excel spreadsheet.

The spreadsheet contains multiple sheets, each of which represents a different data source. The application uses R packages such as readxl [20] to import the data into the application. The Shiny framework leverages a variety of packages and tools within the R ecosystem, including ggplot2 and plotly, to create interactive and responsive web applications. Overall, a scalable and adaptable solution for data processing and visualization is intended to be provided by the system architecture. By leveraging the power of the Shiny framework and the R programming language, the application can easily adapt to changing data and analytical requirements. A simple and effective method of organizing data for analysis is the ability to source data from various sheets within a single Excel spreadsheet.

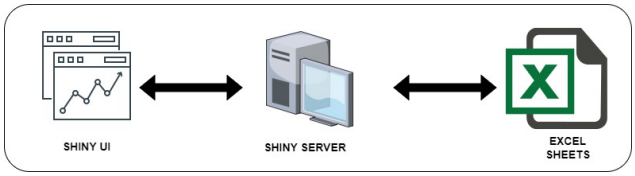


Fig. 5: Application architecture

B. Examples of use

Due to the lightweight interface that the application CroStats provides, the users should feel fairly comfortable when using it for the first time.

There are several ways to use the application. The main ones are finding the exact values of the categories for a certain year, comparing the absolute values of several counties and comparing the counties based on the values per thousand inhabitants. The user can compare up to four counties so that the view doesn't get crowded.

Firstly, an example where the user just wants to find out the population of a certain county in a certain year. Since that is the default layout, all the user has to do is hover over the point that coresponds to the wanted year (see figure 6).

Secondly, if the user wanted to check out more than one county, the user could just choose more counties by clicking on the map as described in the applications hint placed nearby. Deselection is made by clicking on the selected county. All of the above works the same for all the other categories prompted in the drop down menu under "Kategorija:". Figures 6 and 7 showcase the implementation of these examples.

In order to compare counties it should be noted that it's not a good idea to base the comparison on the absolute values. The number of inhabitants plays a big role if the counties vastly differ from one another. Because of that, scale is incorporated into the application. Aside from absolute values, the application provides values per thousand inhabitants. This way a user can confidently compare counties without having to worry about the size of each of the counties (see figure 2).

And while the user plays with the settings of the interface, historical and demographic facts are generated at the bottom to give a few insights about the selected counties.

VI. FUTURE WORK

For future work, even though Shiny made the development of the interface possible and intuitive, the lack of versatility in hosting options and overall design options resulted in considering scaling the application using a different technology.

One technology that is particularly interesting is D3.js [21]. D3.js, or Data-Driven Documents, is a JavaScript library that allows for the creation of interactive data visualizations on the web. It is a powerful tool that can be used to create a wide range of visualizations, from simple charts and graphs to more complex and interactive maps and animations.

There are several reasons why using D3.js for future work is considered. First, D3.js is a widely used and well-established technology that has a large and active community of developers. This means that there are many resources and tutorials available for learning and working with the library. Additionally, D3.js has a wide range of

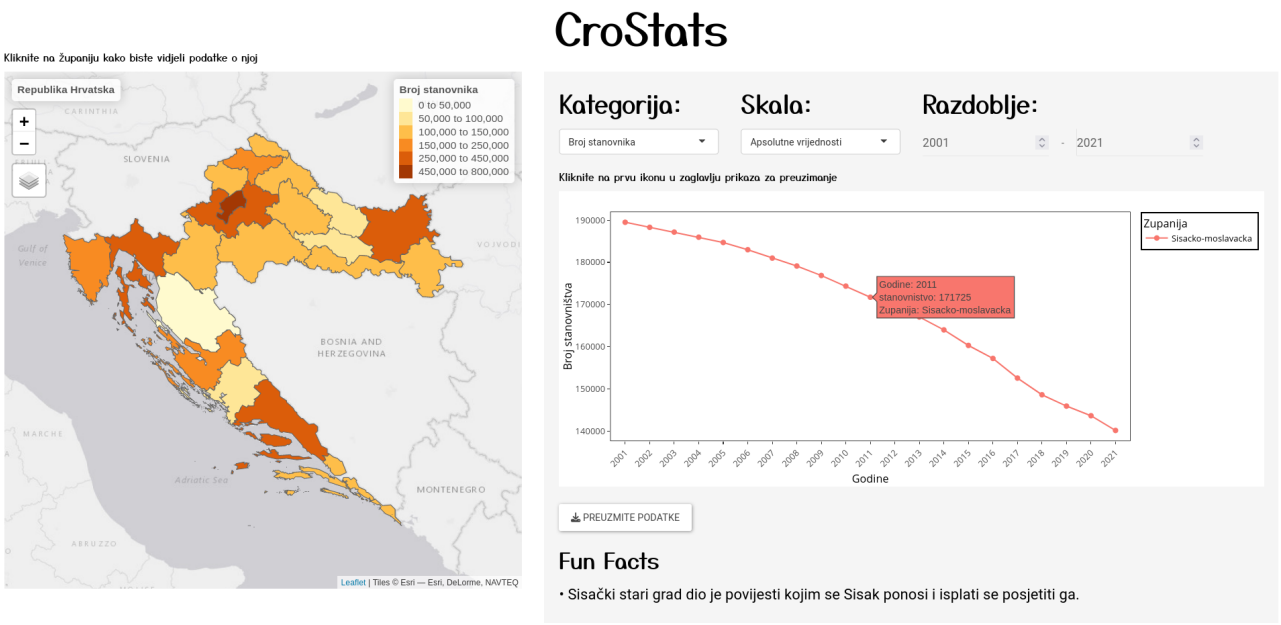


Fig. 6: CroStats example 1

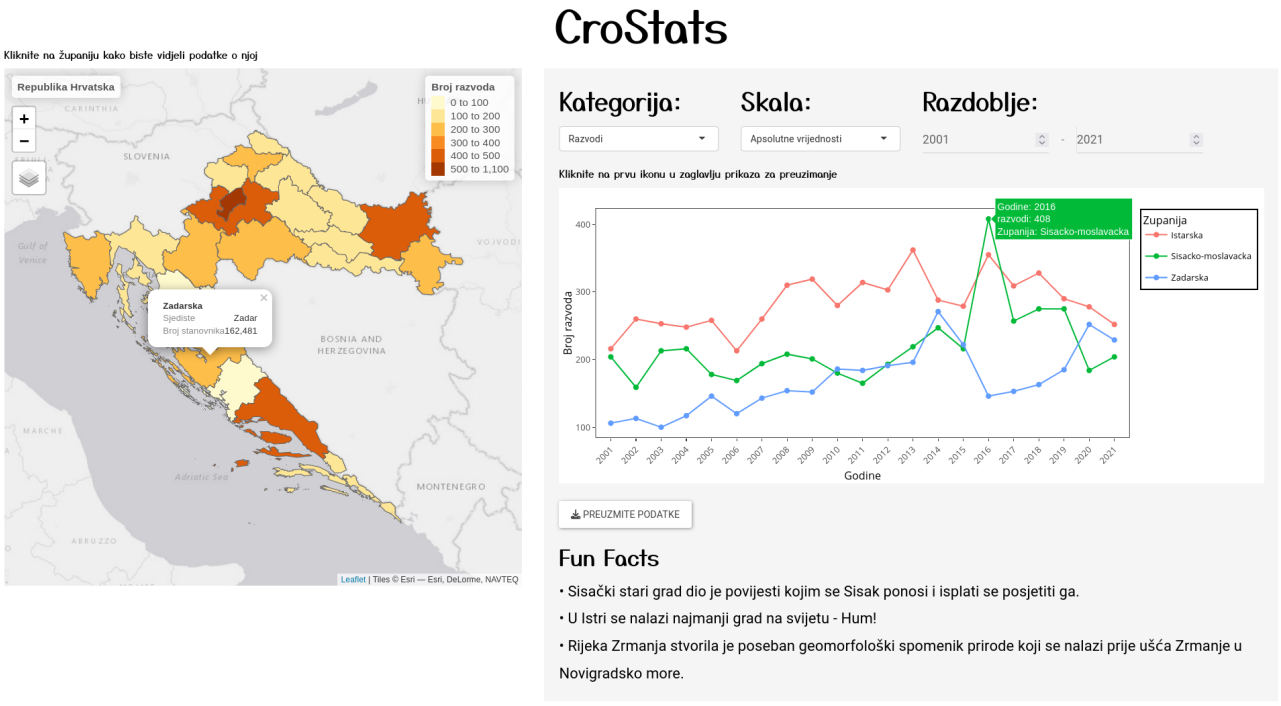


Fig. 7: CroStats example 2

built-in functions and tools that can be used to create advanced visualizations and animations that are not possible with R Shiny. There is also a significant advantage when interactivity is taken into account. Animations can be added to visualizations, which can help the data become more engaging and easier to understand. R Shiny has a more streamlined way of displaying visualizations but the big drawback is loss of flexiblity in shaping interactivity. Another advantage of D3.js is that it is built on JavaScript, which is a widely-used programming language that is supported by most web browsers.

The structure of the data could also be implemented in a better way. Instead of Excel spreadsheet, the data could be imported into a database that communicates with the app. It would then be extracted and used for the visualization. That could simplify the process and make the app faster which would definitely be an improvement.

As far as features go, the number of different graphs that are displayed on the app would be expanded. Further data analysis combined with new insights for what is possible with other frameworks revealed that data can be visualized in many ways. Some of them are not traditional but they can easily provide the same, if not more intuition on what the data is trying to tell.

There would be rules set in place for which column means what, and if the users were motivated enough to clean their data, the users could use the tool for their own (real or imaginary) demographic. Of course, a user might want to store the data for later analysis, and adding a data persistence layer should be pretty simple because of available infrastructure as a service (IaaS) solutions like Microsoft Azure [22]. The app would then connect to a web service that would handle user authentication and authorization, and then serve the saved data from a database. Both the web service and the database would be handled by the cloud provider, which would enable the developers to quickly build the business logic.

VII. Conclusion

CroStats is a great tool for visualizing data from population census. It is based on a user-friendly web interface. The main idea was to build a robust tool that helps users gain insights into population trends and patterns. The use of interactive graphs and maps in the web interface improves the user experience as it provides a dynamic way to explore data. Using that kind of data presentation enables users to better understand the demographic data of Croatia. Furthermore, the simplicity of the web interface makes it accessible to a broader audience, regardless of the technical background.

Acknowledgment

This research has been supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS).

References

[1] "history," Mar 2023. [Online]. Available: https://www.abs.gov.au/websitedbs/d3310114.nsf/51c9a3d36edfd0dfca256acb00118404/eadaffffb171cab6ca257161000a78d7!OpenDocument
[2] "Cro history," Mar 2023. [Online]. Available: https://www.croatian-genealogy.com/educational-articles/19th-century-census-source-croatian-genealogical-data/
[3] "The state bureau of statistics," Feb. 2023. [Online]. Available: https://dzs.gov.hr/
[4] "Eurostat regional yearbook," Feb 2023. [Online]. Available: https://ec.europa.eu/statistical-atlas/viewer/
[5] "R shiny," Feb 2023. [Online]. Available: https://shiny.rstudio.com/
[6] "tmap," Feb. 2023. [Online]. Available: https://cran.r-project.org/web/packages/tmap/vignettes/tmap-getstarted.html
[7] "plotly," Feb. 2023. [Online]. Available: https://plotly.com/
[8] "ggplot2," Feb. 2023. [Online]. Available: https://ggplot2.tidyverse.org/
[9] "dplyr," Feb. 2023. [Online]. Available: https://dplyr.tidyverse.org/
[10] "tidyr," Mar 2023. [Online]. Available: https://cran.r-project.org/web/packages/tidyr/index.html
[11] "sf," Feb. 2023. [Online]. Available: https://cran.r-project.org/web/packages/sf/index.html
[12] "Shinythemes," Feb. 2023. [Online]. Available: https://rstudio.github.io/shinythemes/
[13] "Git," Feb. 2023. [Online]. Available: https://git-scm.com/
[14] "Gitlab," Feb. 2023. [Online]. Available: https://gitlab.com/
[15] "dash," Mar 2023. [Online]. Available: https://plotly.com/dash/
[16] "bokeh," Mar 2023. [Online]. Available: https://bokeh.org/
[17] "streamlit," Mar 2023. [Online]. Available: https://streamlit.io/
[18] "tornado," Mar 2023. [Online]. Available: https://www.tornadoweb.org/en/stable/
[19] "flask," Mar 2023. [Online]. Available: https://flask.palletsprojects.com/en/2.2.x/
[20] "readxl," Mar. 2023. [Online]. Available: https://readxl.tidyverse.org/
[21] "Data-driven documents," Feb. 2023. [Online]. Available: https://d3js.org/
[22] "azure," Mar 2023. [Online]. Available: https://azure.microsoft.com/en-us