Data Visualization Assignment

Application demonstrating population in Germany and its changes in the 21st century

Report

Authors:

Paul Borutta, Anna Chester Serra, Márton Bankó

Document History:

VERSION	DATE	DESCRIPTION
V1	06.01.2023	Creation of document
V2	13.01.2023	Completion of main sections
V3	15.01.2023	Finishing up details

Table of contents

1 Introduction

Data visualization is a powerful tool for understanding and analyzing complex data sets. It allows us to explore and discover patterns and trends in the data, and to communicate these insights to others in a clear and intuitive way. In this project, we set out to design and implement an interactive visualization tool that would allow us to examine the population dynamics of Germany. We used a dataset on the population of Germany to create visualizations that would help us understand and analyze various aspects of the population, such as the distribution of foreigners across different states, the trends in the male and female population over time and the distribution of different age groups.

To design and implement our visualization tool, we followed a process that involved several key steps, including: (1) identifying the purpose of the visualization and the key insights or questions that we wanted to understand or explore; (2) collecting and preprocessing the data that we would visualize; (3) selecting the appropriate visualization types and techniques to effectively communicate the insights or information contained in the data; (4) implementing the visualization, ensuring that it was visually appealing and easy to understand; and (5) evaluating the effectiveness of the visualization, both in terms of its ability to communicate the desired information and in terms of its usability and functionality.

In this report, we will describe the process of designing and implementing our data visualization tool in detail, highlighting the decisions that we made at each stage of the process and discussing the challenges and limitations that we encountered. Finally, we will present the results of our visualization and discuss the insights and conclusions that we were able to draw from the data.

2 Problem characterization

The population of Germany is a complex and dynamic system that is influenced by a range of factors, including demographic trends, economic conditions, and social and political changes. These factors can have significant impacts on the size, composition, and distribution of the population, and they can also shape the socio-economic outcomes and wellbeing of different groups within the population. Understanding the population dynamics of Germany is therefore important for a variety of purposes, including policy-making, planning, and research.

In this project, we sought to use data visualization to gain insights into the population of Germany and to answer the following specific questions:

1) **Percentage of foreigners per state**
   How did the percentage of foreigners in the population of Germany vary between the states in a given year, and were there any noticeable geographic patterns or trends in these differences?

2) **Male and Female Population over time**
   How have the male and female population of Germany changed over time, and how do they compare to each other?

3) **Distribution of age groups**
   In a given year, what was the distribution of the population of Germany across different age groups?

To answer these questions, datasets on the population of Germany, containing information on demographic characteristics were used. The data was obtained from the "Statistisches Bundesamt" (https://www.destatis.de/), which is the federal statistical office of Germany. It is responsible for collecting, processing, and publishing official statistics on a wide range of topics, including population, economy, education, and health. The datasets were taken from their online database, which provides access to a wide range of statistical data and publications.

3 Data and task abstractions

This section describes the abstract data and tasks categorizations that were considered for the visualization project and how they were selected. Data abstraction refers to the specific data sources and variables that were used, while task abstraction refers to the specific tasks or questions that should be addressed through the visualization. Understanding the "what" and "why" of the data and tasks is essential for designing and implementing effective visualizations, as it helps to identify the key insights or information that need to be communicated and to select the appropriate visualization techniques and types to effectively convey these insights. In the following paragraphs, the data and tasks that were used will be described, along with the reasoning behind these choices.

3.1 Data abstraction

3.1.1 Source dataset

As mentioned above, the data was obtained from the federal statistical office of Germany. The datasets of interest were found by searching the keywords of interest on their web page. The following two datasets were picked in order to answer the posed questions:

A) 12411-0042[1]: “Durchschnittliche Bevölkerung: Bundesländer, Jahre, Nationalität, Geschlecht”
translation: average population: states, years, nationality, gender

B) 12411-0012[2]: “Bevölkerung: Bundesländer, Stichtag, Altersjahre”
translation: population: states, reporting date, age in years

The first dataset (A) was used to address the first two topics (percentage of foreigners per state, male and female population over time). It contains information on the nationality and gender of the population of Germany for each year from 2000 to 2021 and for each of the 16 states in Germany. This results in a dataset with 352 rows in total.
The 9 columns of the dataset were manually read into the application (in english): German_Male, German_Female, German_Total, Foreign_Male, Foreign_Female, Foreign_Total, Total_Male, Total_Female, Total_Total.

The second dataset (B) was used to address the third topic (distribution of age groups). It contains information about the population of Germany per age group and state. Each group consists of the age in years (ranging from 0 to 90, additional rows: >90, total), which results in 92 rows per year. At download of the dataset the range of interest was also selected to be from 2000 to 2021, for consistency with the first dataset. This results in a dataset with 2024 rows in total.
The columns consist of the 16 states of Germany: Baden-Württemberg, Bayern, Berlin, Brandenburg, Bremen, Hamburg, Hessen, Mecklenburg-Vorpommern, Niedersachsen, Nordrhein-Westfalen, Rheinland-Pfalz, Saarland, Sachsen, Sachsen-Anhalt, Schleswig-Holstein and Thüringen.

Summary:

A) **Dataset nationality and gender**: 352 rows, 9 columns
→ used to answer question 1 and question 2

B) **Dataset age groups**: 2024 rows, 16 columns
→ used to answer question 3

3.1.2 Data type categories

Now, let’s discuss the data types of both datasets used to investigate the original questions. Both datasets used in this visualization project are of attribute data type, which is measurable and recordable. Dataset A contains information on the nationality and gender

[1] https://www-genesis.destatis.de/genesis//online?operation=table&code=12411-0012&bypass=true&levelindex=0&levelid=1673201094367#abreadcrumb

[2] https://www-genesis.destatis.de/genesis//online?operation=table&code=12411-0042&bypass=true&levelindex=0&levelid=1673200920096#abreadcrumb

of the population of Germany for each year from 2000 to 2021 and for each of the 16 states in Germany. It is considered a grid data type, as a strategy for sampling can be rooted in the geometric layout of the states. Dataset B contains information about the population of Germany per age group and state. Both deal with numbers representing the sum of people, so they can be classified as ordered quantitative data. The used datasets are simply structured tables. They can additionally be classified as time-varying datasets, since in both cases one of the key attributes is year. The topic “Percentage of foreigners per state” shifts the focus from exact numbers to relative values. Therefore, the addition of a new column containing percentage information is going to be necessary. In addition to the two mentioned datasets, geometric data of the map of the states of Germany is needed for this topic.

There are cases where a range of a feature is too large to be visualized at every single value, and grouping/categorizing is needed. For example, age can be represented as age groups, which makes it more readable than having a category for every single age in years. It also makes more sense if the analysis does not intend to distinguish between two groups of individuals in slightly different ages. Another example is when measurement is in percentages, and an applicable slicing of percentage-ranges makes visibility better.

3.2 Task abstractions

In data visualization, task abstractions refer to the different levels of understanding or analysis that a user may need to perform when working with a dataset. These can include simple tasks such as identifying a single data point or trend, to more complex tasks such as comparing multiple data sets or identifying patterns and relationships. Task abstractions can also refer to the different ways in which data can be presented, such as through a graph, chart, or map. The goal of data visualization is to effectively communicate information and insights at the appropriate level of abstraction for the intended audience and purpose.

The task abstractions of the three topics all intend discovery and presentation of the data. The visualizations serve to show the different aspects of the population of Germany and any patterns or trends that may be present. They can generate new hypotheses and insights for the user. A high quality display can also be used as information to third parties.

In terms of search, all tasks are considered a "lookup" as they have known target and location. The queries for these tasks fall under the category of summarization and comparison, as they provide a summary of the targets within the data and compare them to each other. Regarding the first topic, it compares the percentage of foreigners between the states. For the second topic, it compares the male and female population. The third topic compares the size of different age groups to each other. The following outlines the task abstractions in more detail for each topic individually.

1) Percentage of foreigners per state

The task abstraction of this question is complex. It requires the user to understand the percentage of foreigners in the population of each state in Germany for a given year, and also compare the percentages of foreigners across different states to identify variations. This visualization would not only consume information, but also produce it, since the percentage value is included in the dataset, but needs to be calculated.
The type of search can be considered a "location" task, since the user knows what to look for, but not exactly where. The user should be able to find noticeable geographic patterns in the distribution of foreigners across Germany, and possibly compare the relationship of variations with the changes happening each year. As mentioned above, it can also be considered a "lookup" search if the user is interested in the percentage in a specific state and does not intend to find general patterns. The target of the question is to look at all the data and discover trends, outliers or visual features, although, if the user intends to, one attribute can be singled out and focused on its values.
To accomplish this, the user would likely need to use multiple visualization techniques, such as a map to show the geographic distribution of foreigners. They would also likely need to interact with the data in various ways, such as filtering or sorting the data to focus on time periods or percentages of certain sizes.

2) Male and Female Population over time

This question involves the understanding of the population of Germany by gender over time, and the comparison of them. It includes the recognition of patterns shown by the graphics of genders individually, but also looking for trends in the scope of the two, possibly adding the entire population as a third factor.
The target of this task is many attributes, and to bring light to possible similarities, dependencies, or correlation.
To achieve this, the use of visualization techniques such as line or area chart to show the population of males and females over time is needed. The interaction with the data is to be focused on the filtering of specific time periods.

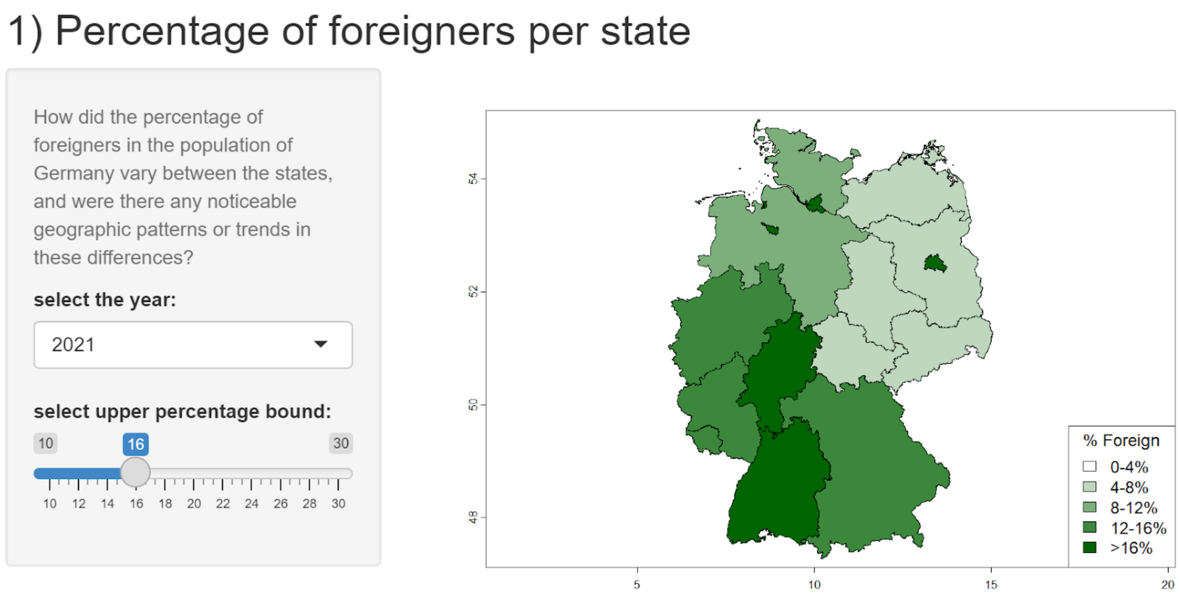
3) Distribution of age groups

The task abstraction of this question entails the grasping of the population of Germany across different age groups for a given year, interpreting the distribution of the population across different age groups and recognizing motives in the proportion of population in each age group.
The best way to reach this would need to use visualization techniques such as a bar chart, or a histogram. The user would interact with the data by selecting between data collected in various points of time. The target here is the entire data.

4 Interaction & visual encoding

This level determines the specific design choice for creating and manipulating the visual representations of the abstract data types that have been selected in the upper abstraction level, guided by the abstract data types and tasks identified in the previous section.

Considering interactions, in all questions to be answered we can see time as a main feature: they consist of observations to be made over time or at a specific year. Therefore the user should be able to pick the year or time period of interest in all three cases. The visualizations are displayed in the following figures 1-3. The left box contains the question, which the visualization intends to answer, followed by the interactive input options for the user. The right side of each layout contains the plot with the visualization.



1) Percentage of foreigners per state

How did the percentage of foreigners in the population of Germany vary between the states, and were there any noticeable geographic patterns or trends in these differences?

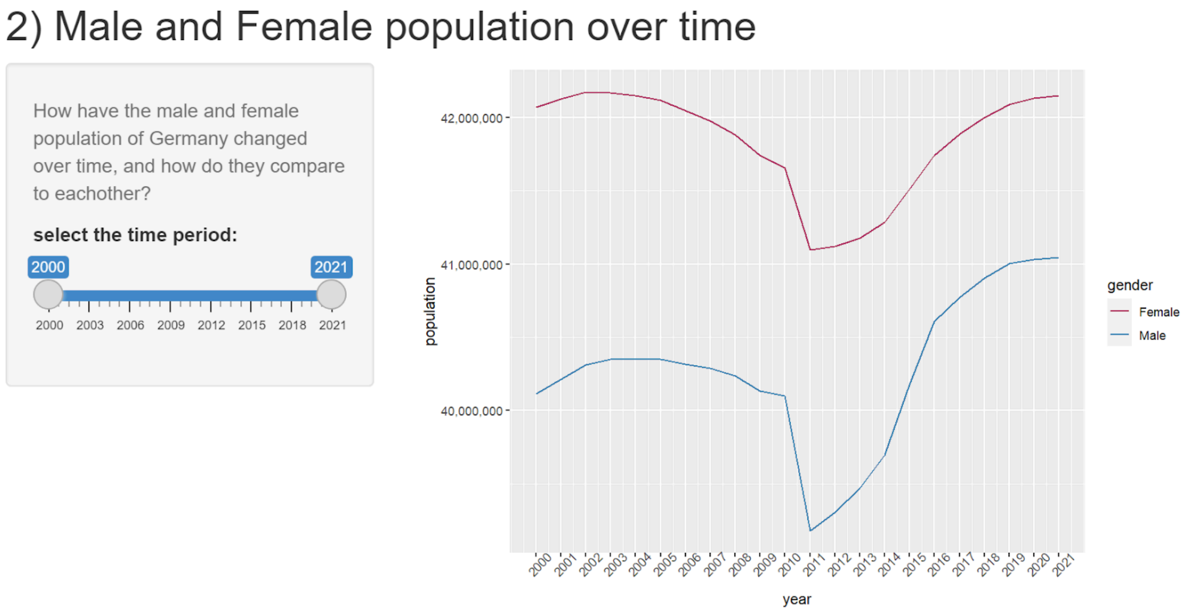
select the year: 2021

select upper percentage bound: 16

Figure 1.

Visual encoding: Answering the first question involves displaying the percentage of foreigners in the population of Germany for each state in a given year, using a choropleth map of Germany. A choropleth map is a type of map that uses color to encode quantitative data. It is a particularly appropriate choice for this task because we have geographic geometry data and we want to display one quantitative attribute per region. The given geometry is used for area mark boundaries. This allows the user to see the spatial distribution of the data and to identify any geographic patterns or trends. This is especially relevant in this case, because one of the goals of the visualization is to explore whether there are any noticeable differences between east and west Germany in terms of the percentage of foreigners in the population. By using a map, the user can easily see the geographical location of each state and how the data varies across different regions.

Interaction: The user can select the year of interest for the map, between 2000 and 2021. In addition, the user can also select the upper bound for the percentage range of the map, which allows adjustment of the stepsize for the five color categories. This is useful for the comparison of states with similar levels of foreign residents. The default upper bound is set to 16%, which yields the five categories visible in figure 1. When the user sets the upper bound, for example to 20%, these 5 categories are shown in the map: <5%, 5-10%, 10-15%, 15-20%, <20%.



2) Male and Female population over time

How have the male and female population of Germany changed over time, and how do they compare to eachother?

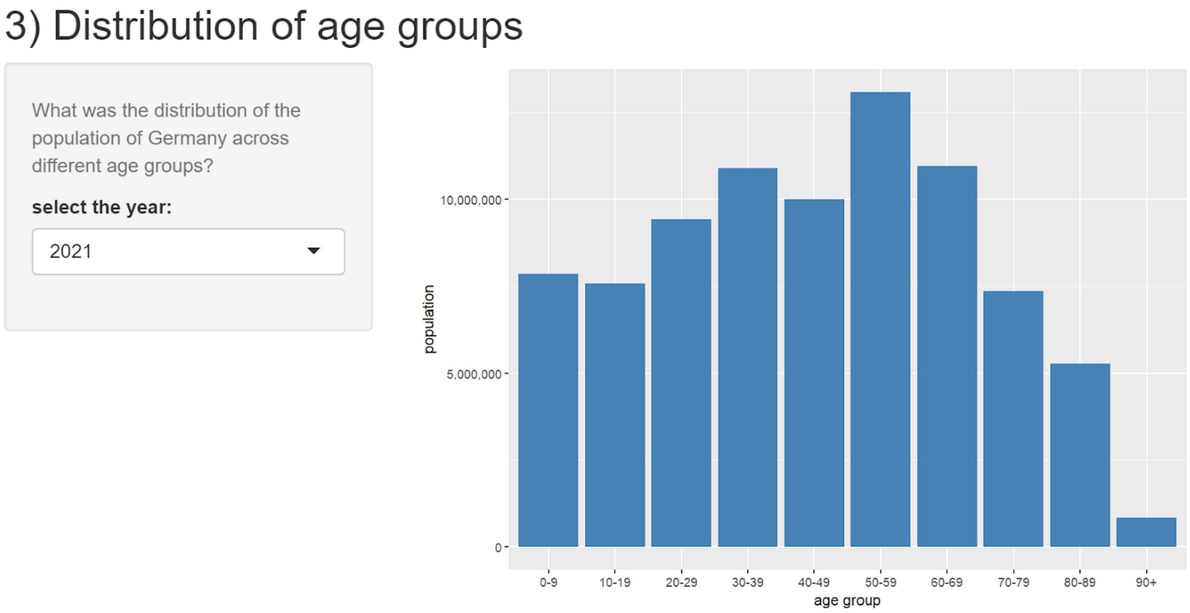
select the time period:

2000 – 2021

Figure 2.

Visual encoding: This involves visualizing the changes in the male and female population of Germany over time, and comparing them to each other. To achieve this, a line chart was used, which is well-suited for displaying data that is related to time. The line chart displays two lines, one for the male population and one for the female population. This way they are easily comparable. This design choice was motivated by the aim to enable the user to gain insights into the trends and patterns in the male and female populations of Germany and to understand how they compare to each other.

Interaction: The user can explore different time periods of interest. This creates a subset of Dataset A that only contains the entries of the years within the given time period. The default time period is set between 2000 and 2021, which is the entire range of the dataset.



3) Distribution of age groups

What was the distribution of the population of Germany across different age groups?

select the year:

2021

Figure 3.

Visual encoding: For addressing the third topic, a bar chart was used. This design choice was made to enable the user to see the distribution of the population by age group. The bar chart allows the user to see the relative proportions of the population in each age group and to compare them to each other. By showing one bar for each group, the user can see the distribution of the population across all of the age groups. There are 10 age groups in total, each containing a range of 10 years, except for the last group which includes all people over the age of 90.

Interaction: The user can select the year of interest. The selection automatically picks the subset of Dataset B that only includes the entries with the given year. The default is set to year 2021.

5 Algorithmic implementation

The section describes the technical details and steps taken to create the data visualizations for the population dynamics of Germany. It is intended to provide a clear understanding of the methods and processes used to create each of the visualizations, as well as any limitations or considerations that should be taken into account when interpreting the results. The visualization was created using the R programming language and the Shiny library. It is a web application that allows the user to interact with the data. The desired layout was created using three sidebarLayout elements, each consisting of a SidebarPanel containing the description and the user input (left) and a MainPanel containing the visualization (right).

1) Percentage of foreigners per state

To address the first topic, the dataset A is preprocessed by selecting only the columns of interest, which are the year (user input), the total number of foreigners and the total population (per state). Additionally, the percentage of foreigners per state was calculated dividing the number of foreigners in each state by the total population in that state. The resulting percentage was included in a new column "`foreign_pct`".

The user selects the year of interest from a dropdown menu, which is implemented with the "`selectInput`" function and the upper percentage bound on a slide bar, implemented with "`sliderInput`" function.

The calculated percentage values for the given year, along with selected the upper percentage bound for the categories, serve as input for the implemented function that maps the data to the corresponding states on a choropleth map:

The function (`percent_map`, in helpers.R) generates a vector of fill colors for the map using the colorRampPalette function. It creates five shades, ranging from white to dark green. Next, the function calculates the cut-off points for the percentage (input) and assigns each state to a range based on the percentage of foreigners in that state. The function then loads the geographic data for the states of Germany using the readRDS function. This rds data was obtained from https://gadm.org/download_country.html. The data is plotted on a map using the plot function. The map is colored according to the percentage of foreigners in each state using the col argument in the plot function. A legend is added to the map, which shows the range of percentages for each color. This allows the user to easily understand the meaning of the colors on the map. Darker colors indicate higher percentages.

2) Male and Female Population over time

To address this topic, the dataset A was preprocessed by selecting only the columns of interest, which were the total number of males, females, and year. The dataset was then grouped by year and all columns were summed up to obtain the values for the entire population of Germany, instead of per state. These preprocessing steps were performed using "`dplyr`" functions, which are a set of tools for working with data frames in R.

The user selects the time period of interest from a two-sided slide bar, which was implemented with the "sliderInput" function. The values range from 2000 to 2021 and the default covers the entire range.

According to this input, the data is transformed by creating a new dataset that contains only the selected time period. The data is then reshaped from wide format to long format so that it can be used in the plot function. The data is mapped to the ggplot function ("`ggplot2`" library) to create a line chart, using "`geom_line`". The x-axis represents the years, the y-axis represents the population and the color represents the gender. Maroon represents Female and steelblue represents Male. The two lines are represented in one line chart for comparison.

3) Distribution of age groups

To address this topic, the dataset B was preprocessed to only contain rows without missing values. Following the transformation of the date column to only keep the year, the rows containing total aggregated data were removed, as they are not helpful for achievement of the goal. After that the columns containing data for the different states were merged into a single column. Then, the ages listed were grouped into age groups (0-9, 10-19, …, 90<), which is followed by the aggregation of population by the combinations of given year and age group. Finally, unused columns were dropped, leaving the table with columns directly used in representation.

The user selects the year of interest from a drop-down menu with values between 2000 and 2021. This was implemented using "`selectInput`". The default value is 2021.

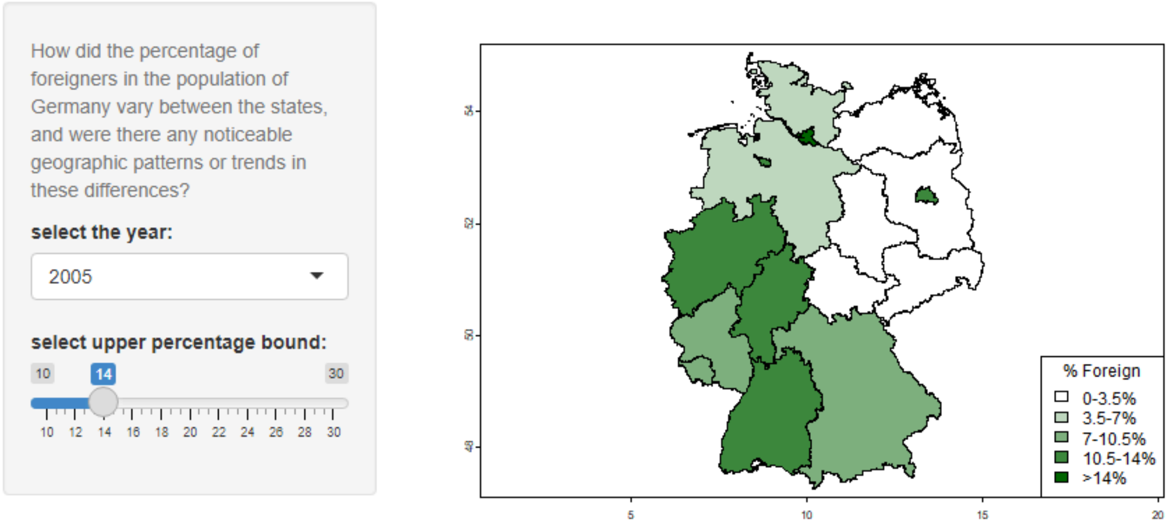
Analogous to the second implementation, a subset of the data is created according to the user input and the plot is created using the "`ggplot`" function. The bar chart showing the different age groups is generated using "`geom_bar`". The x-axis shows the ten age group category and the y-axis shows the population.

6 Evaluation

The evaluation section of this report aims to assess the effectiveness and efficiency of the visualizations created to gain insights into the population dynamics of Germany. Both qualitative and quantitative methods were used to gather feedback and measure the success of the visualizations in communicating the intended information and insights. The evaluation results will provide an understanding of the strengths and weaknesses of the visualizations and suggest areas for improvement. This section assesses the visualizations' ability to effectively convey the key insights and information, as well as their ease of use and efficiency for the user.

1) Percentage of foreigners per state

The evaluation of the first topic visualization showed that the visualization effectively communicates the insights and information intended to be conveyed. The map-based format allows for easy identification of patterns and trends in the data, making it easy for users to understand the distribution of foreigners across the states of Germany. Through the visualization, the most eye-catching insight that users gain is that the states of former East Germany have considerably less foreigners than the states of former West Germany. This highlights a demographic difference between the two regions and could be an interesting point of further exploration. By changing the visualization throughout the years, the user can also see the trend that the percentage of foreigners increases over the years. Additionally, it stands out that the following states have a comparably high number of foreigners: Berlin, Hamburg, Bremen, Baden-Württemberg and Hessen. The interactive elements, such as the year and percentage range selectors, were found to be user-friendly and efficient, allowing users to easily filter and explore the data.



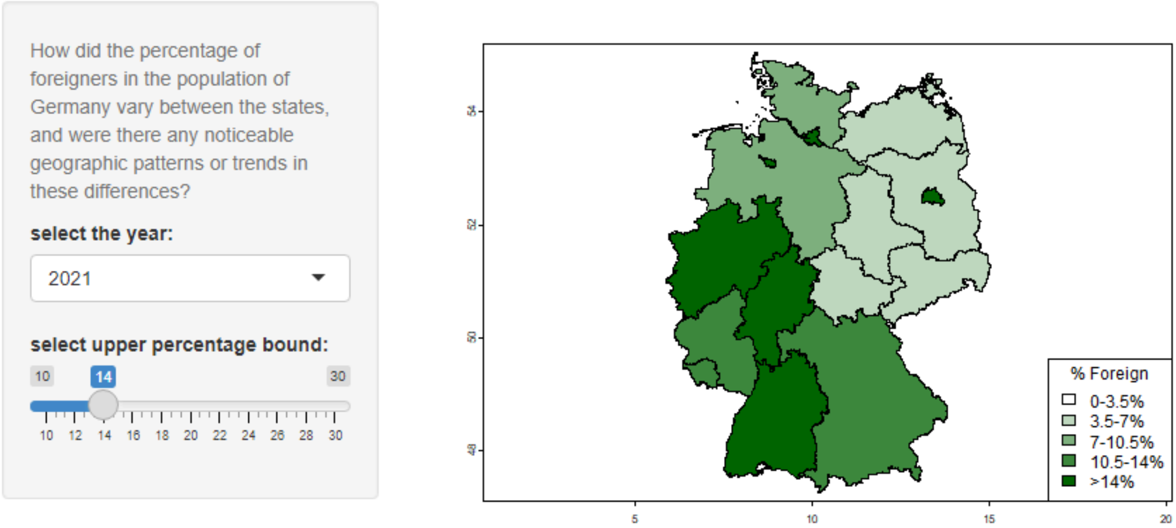
How did the percentage of foreigners in the population of Germany vary between the states, and were there any noticeable geographic patterns or trends in these differences?

select the year:

2005

select upper percentage bound: 14

Figure 4: Foreigner population in 2005



How did the percentage of foreigners in the population of Germany vary between the states, and were there any noticeable geographic patterns or trends in these differences?

select the year:

2021

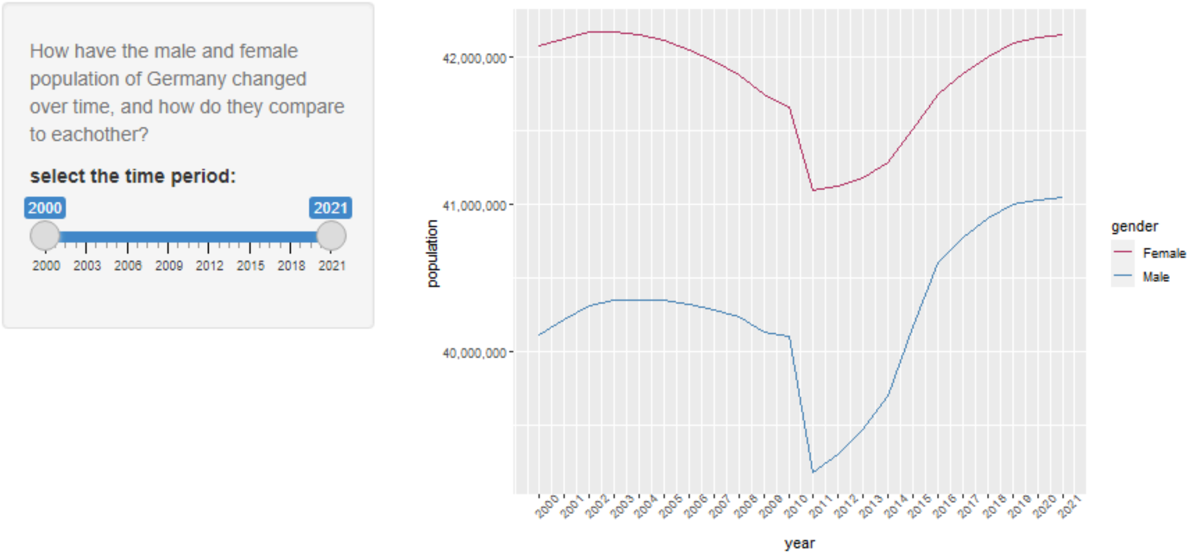
select upper percentage bound: 14

Figure 5: Foreigner population in 2021

However, user testing yielded that more information on the map would be desirable. For example, the exact percentage of foreigners in each state or a way of visualizing the names of the states for users that are not completely familiar with the geography of Germany could be useful and make the data more actionable. Additionally, adding a temporal dimension to the visualization in order to show the development over time would be preferable, as it would allow to see trends and patterns over the years.

2) Male and Female Population over time

The visualization for the second topic also communicates the insights and information as intended. The line chart format allows for easy identification of patterns and trends in the data, making it easy for users to understand the change of the male and female population over time and how they compare to each other. The interactive element, such as the time period selector, was found to be user-friendly and efficient, allowing users to easily filter and explore the data. Through the visualization, users were able to gain an overview over the time periods with increasing and decreasing population. These insights include a slight increase of the population between 2000 and 2003, a decrease of the population between 2003 and 2010, a pretty pronounced drop in 2011, and a steady increase since 2011. Furthermore, the user can observe that, in 2000, there were about 2 million more women than men in the population and in 2021, only about 1 million more women than men, revealing a trend towards a more balanced gender ratio.



How have the male and female population of Germany changed over time, and how do they compare to eachother?

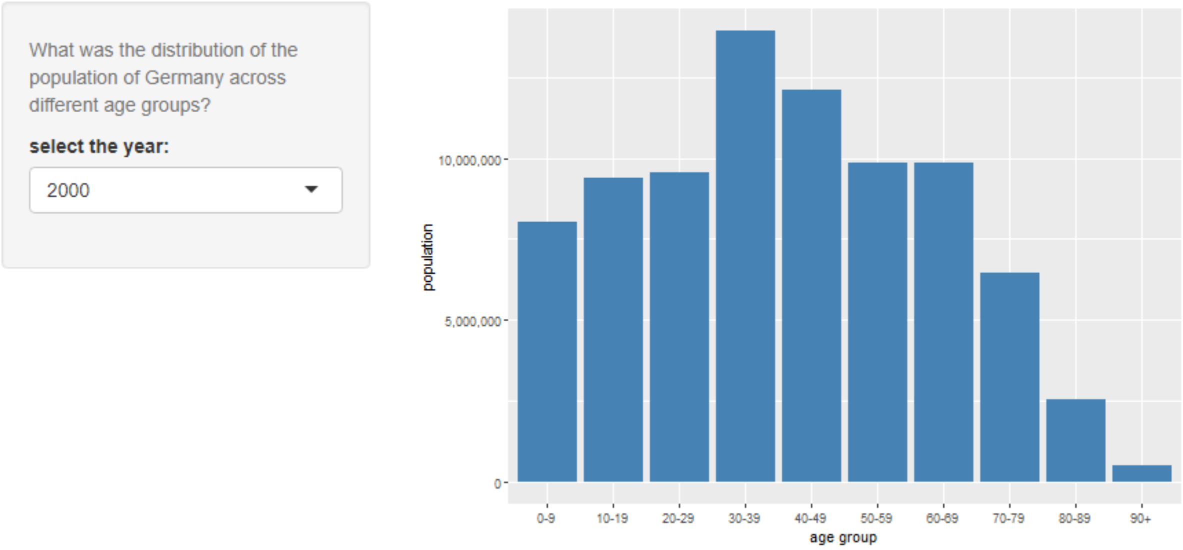
select the time period: 2000 – 2021

Figure 6: Changes of population of genders between 2000-2021

However, feedback from user testing suggested that providing more context for the data, for example, by showing the total population size as an additional label for each displayed year, would be favorable.

3) Distribution of age groups

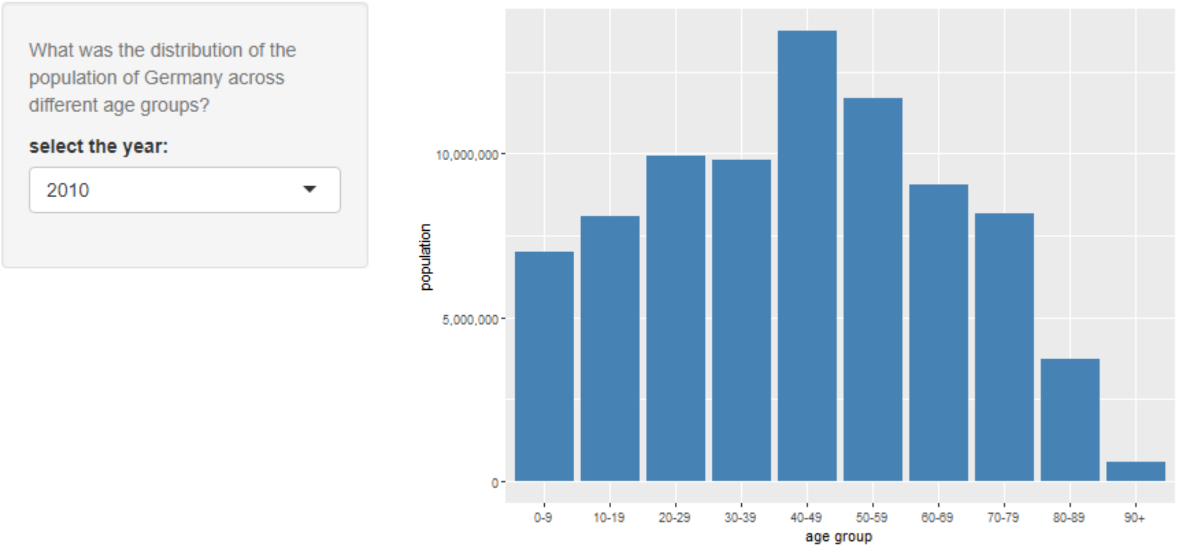
The visualization graphic created for the third topic also adequately conveys the data and information needed to answer the questions regarding this matter. The bar chart is a simple but powerful tool to present the values of each age group, highlighting the differences between them, and help recognize underlying patterns behind the numbers. The dropdown list, which was also used in the previous graphics as an interactive element, was effective in this case as well, allowing the user to change between which year’s data should be represented. With this, the benefit of measuring and contrasting the differences of each sample is yielded. We can observe a slight ‘aging’ in the population of Germany. As the years progressed, the evidently most populated age group has slowly shifted to older age groups, from the people aging between 30-39 showed as largest community in the year 2000, to it gradually changing to the 50-59 age group. This demonstrates the widely experienced motive of countries in Europe, where cultures go through a similar case of their population maturing with less presence of the younger generations.



What was the distribution of the population of Germany across different age groups?

select the year: 2000

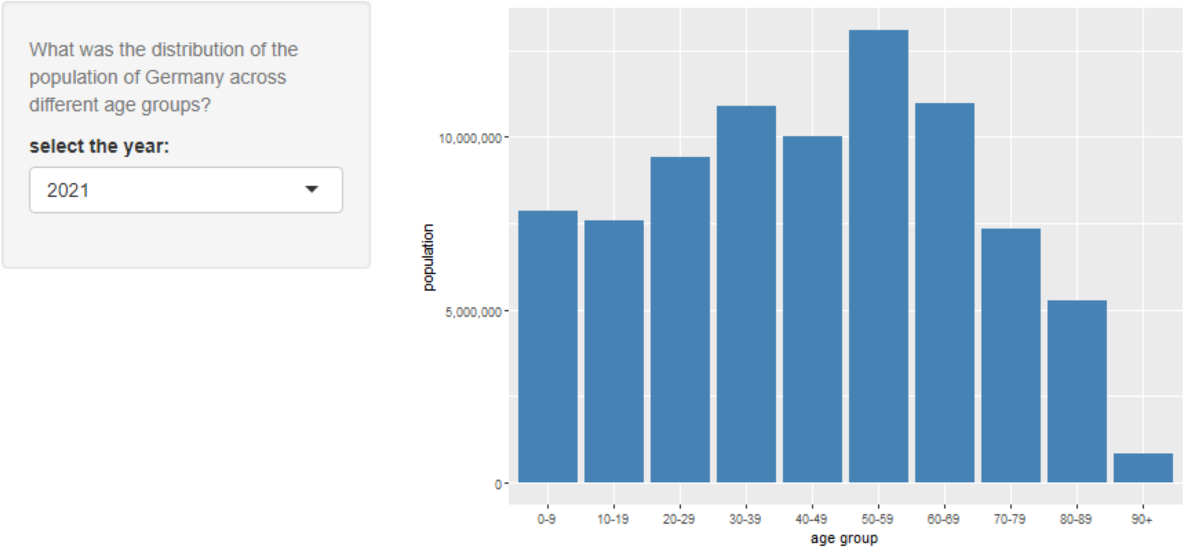
Figure 7: Age distribution in 2000



What was the distribution of the population of Germany across different age groups?

select the year: 2010

Figure 8: Age distribution in 2010



What was the distribution of the population of Germany across different age groups?

select the year: 2021

Figure 9: Age distribution in 2021

As a retrospection, the comparison of data between the years could be improved by allowing the user to select multiple years to visualize, and the option of showing them clearly and close proximity to each other.

7 How to run the program

The application is structured in the following way, in a folder called **germany_population**:

- **data**:
  - **gadm**: File containing information about the geographic layout of Germany and used in the visualization of the first question
  - **age.csv:** Dataset B about the age of the population of Germany, in CSV format
  - **natonality_gender.csv:** Dataset A about the nationality and gender of the population of Germany, in CSV format
- **app.R:** The file holding the main functions of the program, executing the steps off including necessary libraries, data loading and preprocessing, assembling the UI of the tool, and registering the appropriate functions to draw out the graphs in an interactive way
- **helpers.R:** An R file with a helper function assisting with the display and the color coding of the different states in the map graphic

All other files hold meta information about the project or git.

To run the program:

1. Download the ZIP containing our application, and extract it to a folder of you choice

2. Install the following libraries on you R environment: **shiny, RColorBrewer, leaflet, dplyr, tidyr, ggplot2, geodata**
3. Run the app with RStudio, or with the console command:
   `runApp`("<Path to folder containing the file "**app.R**">")

   (e.g.: `runApp("C:/Path/to/germany_population")`)

4. While the app is running, you can interact with the diagrams through the interactive dropdown menus or sliders

The application is also published on shinyapps.io, it can be found in the following URL:
https://bankomarton.shinyapps.io/germany_population/

8 Conclusion

Through this project, we have demonstrated the effectiveness of data visualization in examining the population dynamics of Germany. We have successfully designed and implemented an interactive visualization tool that allows us to explore and discover patterns and trends in the population data, such as the distribution of foreigners across different states, the trends in the male and female population over time, and the distribution of different age groups.

We have followed a systematic process that involves key steps, including identifying the purpose of the visualization, collecting and preprocessing the data, selecting the appropriate visualization types, implementing the visualization, and evaluating its effectiveness. We have highlighted the decisions we made and discussed the challenges and limitations we encountered during the process.

Overall, the visualization tool that we created can effectively communicate the insights and information contained in the data, and it is visually appealing and easy to understand. We were able to draw meaningful conclusions from the population data, which can help support decision making and policy-making in Germany. We hope that this project will provide valuable insights and inspire further research in data visualization and population dynamics.