Assignment 5 - Covid-19

Project Overview

Group-11

Members: Abdul Ahad Ayaz, Sayalee Chavan, Suganthi Jaganathan, Bhagyashree Sanjay Borade

This document provides the overview of the completed project, containing the description of dataset, the End Users that we are targeting, visualization techniques and the interactions that have been used, other code details and the overall justification on meeting the goals.

DATA

There are three multivariate datasets that we are going to use **covid_de.csv**, **demgraphics_de.csv** and **counties_population.csv**. First dataset **covid_de.csv** provides the daily update of Covid-19 cases of different states and counties of Germany based on gender and age group. Second dataset **demgraphics_de.csv** is the supporting dataset for the first dataset as it contains the population of different states of Germany based on gender and age group. Third dataset **counties_population.csv** describes the population of each county in NRW state. The reason of choosing these dataset upon other dataset is that it satisfies the requirement of End User described below.

Data Source

- **covid_de.csv**
  - https://www.kaggle.com/headsortails/covid19-tracking-germany?select=covid_de.csv (covid_de.csv is refined dataset based on RKI data)
  - https://npgeo-corona-npgeo-de.hub.arcgis.com (Collected by Robert Koch Institute)
- **demographic_de.csv**
  - https://www.kaggle.com/headsortails/covid19-tracking-germany?select=demographics_de.csv
  - https://www-genesis.destatis.de/genesis/online/data?operation=sprachwechsel&language=en (Collected by Statistisches Bundesamt)
- **counties_population.csv**
  - https://www.citypopulation.de/en/germany/admin/05__nordrhein_westfalen/

Dataset Type

Dataset	Spatial vs Non-Spatial Data	Point, Scalar, Vector Data
covid_de.csv	Non-Spatial Data	Point Data
demographic_de.csv	Non-Spatial Data	Point Data
counties_population.csv	Non-Spatial Data	Point Data

Dataset Characteristics

The first dataset **covid_de.csv** consist of following columns:

Column Name	Data Type	Description
state	Nominal	It contains the name of German states that are total 16 in number. Example Baden-Wuerttemberg, Nordrhein-Westfalen, Hessen, Bayern etc.
county	Nominal	It contains the name of counties of each state label with LandKries(LK)/StadtKries(SK). Example SK Koeln, SK Hamm, LK Paderborn, LK Guetersloh in Nordrhein-Westfalen state.
age_group	Ordinal	It consists of different age categories such as 0-4, 5-14, 15-34, 35-59, 60-79, 80-99 and empty field NA with age unknown.
gender	Ordinal	It consist of gender M for male and F for female and NA for gender unknown.
date	Nominal	It consist of date when the data is updated.
cases	Quantitative	It consist of per day count of confirmed cases of Covid-19.
deaths	Quantitative	It consits of per day count of death due to Covid-19
recovered	Quantitative	It consists of per day count of recovered patients from Covid-19.

The Second file **demographics_de.csv** that we are going to use for our project consist of following columns:

Column Name	Data Type	Description
state	Nominal	It contains the name of German states that are total 16 in number. Example Baden-Wuerttemberg, Nordrhein-Westfalen, Hessen, Bayern etc.
gender	Ordinal	It consist of gender such as male and female.
age_group	Ordinal	It consists of different age categories such as 0-4, 5-14, 15-34, 35-59, 60-79 and 80-99.
population	Quantitative	It consists of population categorized by state, gender and age.

The Third file **counties_population.csv** that we are going to use for our project consist of following columns:

Column Name	Data Type	Description
county	Nominal	It contains the name of counties of each state label with LandKries(LK)/StadtKries(SK). Example SK Koeln, SK Hamm, LK Paderborn, LK Guetersloh in Nordrhein-Westfalen state.
population	Quantitative	It consists of population categorized by counties

Data Modeling

- **covid_de.csv** is described as Entity E_7^P, where P represents that dataset has Point Data and 7 represents the dimensions.
- **demographics_de.csv** is described as Entity E_3^P, where P represents that dataset has Point Data and 3 represents the dimensions.
- **counties_population.csv** is described as Entity E_3^P, where P represents that dataset has Point Data and 2 represents the dimensions.

USER AND TASK

Users are the ones for whom the application or the concept has been designed. Meanwhile, tasks are used to express the target of our visualization.
They describe the purpose of an application or a project.

User

The Dashboard has been exclusively designed for the State Govt of North-Rhein-Westphalia(NRW), Germany.

Task

We have defined a set of visualization tasks from which the user could be benefitted on the following aspects:

- To know the Covid-19 infections explicitly with respect to every County in NRW and its growth/decrease.
- To know the infection rate with respect to children (aged 0 to 4 or 5 to 14) on the after effect of the operations of kindergarten/schools.
- To know the spread of infection with every age group such as 0 to 4, 35 to 59 and find the most vulnerable age group.
- Daily status of the of Covid-19 interms of total cases, recovered cases and deaths of the state.
- To monitor the spread of Covid-19 in each county with respect to the population.
- The above information would help the user to resume/halt the kindergarten/schools operations, revise restrictions on workplaces/shops/markets and further to decide upon the lockdown phases for a county.
- To know the major policies taken to reduce no of cases daywise can help government to take majors.

VISUALIZATION TECHNIQUES

Data is preprocessed by acquiring only the NRW state data and its relevant columns. The data was polished then (handling bad data)and is utilized for transformation and graphical representation . It will be visualized using various interactive visualization techniques to have a broader overview of correlation amongst data which will help to deduce significant information.

Techniques

The visualization techniques that will be used are as follows:

- **Time-oriented visualization technique**
  - Multi-series Line chart (for multivariate data) : x-axis represents "Dates" and y-axis represents "total_cases". Graphical attribute (glyph) denotes categories of cases (confirmed, recovered, death) legends.
- **Region-based technique**
  - Bar chart: x-axis represents "age group" and y-axis represents "population". Also, the data categorization is done with respect to number of cases. Hover represents the population, and total cases against the age group.
  - Grouped Bar chart (for multivariate data): x-axis represents "Counties" and y-axis represents "cases". Each bar in a group represent the Confirmed, recovered and death cases with respect to every County.
    The legends support to view the data categorically such as confirmed cases in all the Counties, similarly recovered cases in Counties and so on.
    Hover represents the count for each category.
- **Multivariate data visualization technique**
  - Bubble chart: x-axis represents population and y-axis represents "No. of cases". Size of the bubble depends on the count of cases with respect to every County which is added to legends. Hover represents the count cases and population.

INTERACTIONS

Interaction in visualization techniques provides more understanding of visual representation. User can do the selection, filtering for switching between different visualization techniques or choosing data variables. Interaction for visualization techniques specified as following:

Interaction Operators

- **Filtering:** (Purpose : Navigation) The user can select a particular range for days mentioned on the x-axis as well as able to see cases in the range of selected days. Filtering has been done for all the graphs by dropdown selection and checking/unchecking the legends".
- **Selection :** (Purpose : Selection) Selection can be used for visualizing different data variables by selecting/unselecting legend-based checkboxes and selection would be one or multiple. This interaction operator can be used in Multi-Series Line Chart where the user can do selection based on "No. of confirmed cases", "No. of recovered cases" and "No. of deaths". For Bar Chart and Grouped Bar Chart, the user can select between Bar Chart and Grouped Bar Chart in dropdown. For Grouped Bar Chart, visualization can be customized by selecting checkboxes represents "No. of confirmed cases", "No. of recovered cases" and "No. of deaths". For Bubble Chart, User can customize visualization by selecting the checkbox which consists of "Total confirmed cases", "No. of recovered cases", "No. of deaths" parameters.
- **Hover :** (Purpose : Distortion)This interactive technique reveals detailed data point information by moving the mouse cursor over data value space which will be included in all visualization techniques mentioned above. For eg, hovering on scatter data points reveals information about no of cases represented by y-axis and age group represented by x-axis.

All the above interaction operators have been added to each and every individual component.

Interaction Operands

The operands are the data values or records mapped to a particular field such as County or Age_group. In all the represented visualizations, age_group is the interaction operand in bar graph and scatter plot, then county in grouped bar graph and bubble chart, and date in the multi line series.

Interaction Space

- **Data value space**
  Data value space is utilized which is performed by the slider to highlight on specific data values. The categorical data will be highlighted on clicking the Legends (single or multiple is possible). Hover highlights data specific to x-axis.
  Purpose : To view a particular data range.
- **Visualization Structure space**
  Visualization Structure space is utilized which is again performed by the slider.
  Purpose : To Move or navigate.

LIBRARIES

The following libraries were used,

- Pandas
- NumPy
- Plotly
- Matplotlib

PARTICIPANTS

We focussed to have equal participation of all the team members. Tasks were fairly distributed and the below table depicts the responsibilities,

Participant	Concept	Visualization and Interaction
Abdul Ahad Ayaz	Dataset	Bar plot, Integration/Documentation
Bhagyashree Sanjay Borade	Visualization Techniques	Multi line series plot, Scatter plot/Integration/Documentation
Sayalee Chavan	Interaction	Bubble plot, Integration/Documentation
Suganthi Jaganathan	User and Tasks	Grouped Bar plot, Integration/Documentation

CONCEPT

We have tried to reach the goals mentioned in the concept document. We have tried to use the same visualization techniques and representations. However, while polishing data, visualizing and achieving interactions, we faced challenges interms of data and visualization libaries. The steps we took to takle these challenges are as follows:

- Added additional task, to monitor the spread of Covid-19 against population of the county.
- We have added additional dataset **counties_population.csv** to support the tasks.
- 3D plots have been changed to 2D plots for making it more pleasant and appealing.
- We faced issues while integrating all visualization techniques so range sliders does not exist in final version.
- We have added dropdown so that user can select visualization based on age group , county, population, and date in single platform.
- We have implemented scatter plot but does not include in the final version due to very less categories of age group data.