CA682: Data Visualization Assignment

Student ID	23260024 23266282
Student name:	Purva Prasad Gosavi Vaishnavi Murlidhar Kulkarni
Student email	purva.gosavi2@mail.dcu.ie vaishnavi.kulkarni2@mail.dcu.ie

Mortality Rate of Different Heart Disease

ABSTRACT

According to the American Heart Association news, two third of the people who have heart attacks are undiagnosed and 84% patients lead to sudden death. (Williamson, 2022) Our topic, "Mortality Rate of Different Heart Diseases," delves into crucial aspects of heart health, utilising a variety of explorations for mortality rates associated with various heart diseases. By Employing PowerBI, the project constructs a compelling stacked bar chart where each colour represents a distinct heart disease topic. Beyond aesthetic appeal, the chart serves as an insightful tool for uncovering patterns and trends in heart disease data. The focus on mortality rates provides a poignant narrative, highlighting the urgency of addressing various dimensions of heart diseases. Through a visually accessible format, the project aims to convey a compelling story that fosters awareness, instigates meaningful discussions, and supports well-informed decision-making within the domain of cardiovascular health. Across the years, the visualisation showcased fluctuations in mortality, offering insights into the dynamic nature of cardiovascular health along with the prevention line telling the how much remained undiagnosed and whether the primary health care.

1. DATASET

Rates and Trends in Heart Disease and Stroke Mortality dataset is taken from the Centers for Disease Control and Prevention website and captured by the Division for Heart Disease and Stroke Prevention in the USA. The dataset's source is the National Vital Statistics System.

This dataset captures the rates and the patterns of mortality due to heart diseases and strokes. It contains the data related to heart disease from 1999 to 2019, which runs over 20 years. This dataset is a combination of two timelines from 1999 to 2009 and 2009 to 2019 intervals. These trends and rates are derived through the Bayesian spatiotemporal model for the smoothing process. It also has a category called class, which determines the cause of death by heart

disease, coronary heart disease, heart failure, or stroke. The dataset can be retrieved through Centres for Disease Control and Prevention website is are follows: https://data.cdc.gov/Heart-Disease-Stroke-Prevention/Rates-and-Trends-in-Heart-Disease-and-Stroke-Morta/7b9s-s8ck

Talking about the size of the data, it contains 5.77 million rows with columns in it. Attributes like year, location, morality rate, topic, prevention numbers, location ID, and others can be noted through this dataset. Also, its categories the data by topic like different types of heart diseases .This file was retrieved from the Centres for Disease Control and Prevention website in CSV format. Its size is 12.75Kb. (Centers for Disease Control and Prevention, 2023)

This dataset possesses three out of four qualities of big data that is volume, velocity, and variety:

1. Volume:

The dataset has 5.77 million rows of data in it. The volume of the data acts as a significant feature of big data.

2. Variety:

That dataset comprises 21 attributes which are various plaintext like location, numeric data like rates, and categorical data like topic. The variety refers to different types of data types present inside the data.

3. Velocity:

The dataset contains data from 1999 to 2019, moreover 20-year period, and it provides us with information on rates and trends in heart disease and stroke mortality.

2. DATA EXPLORATION, PROCESSING, CLEANING

To prepare the dataset we split our dataset into 4 parts for cleaning purposes, such that each spilt dataset had 1.44 million rows in it as our dataset contained 5.77 million rows, we first decided to split it then cleaned that data accordingly. We decided to choose pandas for the complete cleaning process of our dataset. Firstly, we tried loading our dataset in google collab but due to the size of data even after splitting data was large and collab was consuming time to load the data, therefore we decided to go with VSCode so that we can create a folder and add the CSV file of the dataset in that folder and run our python code to clean the data.

Initially, we loaded the CSV file into Pandas dataframe in python. As the code was executing in our local machine, we gave the path of our folder in which the CSV file is located. Once the data got loaded in Pandas dataframe we tried displaying the initial information of the dataset to provide a concise summary of the Data Frame including the number of non-null values in each column of the dataset. We also removed the duplicate rows from the Data Frame. Once the required cleaning was done, we printed the immediate information about the dataset. After the cleaning process was done then a new CSV file was generated in the folder which had cleaned the dataset.

Once all the four parts of the split dataset were cleaned, we combined those files in a one CSV file making it a final cleaned dataset. Before cleaning the dataset, we found out that there were many NA values in it which were removed after the cleaning process. Also, we filtered out many negative values from the dataset which would have affected our visualisation as our aim is mortality rate of different heart disease. These steps were crucial for ensuring that the data used for visualisation is free from missing values, duplicates while providing a more accurate representation of data.

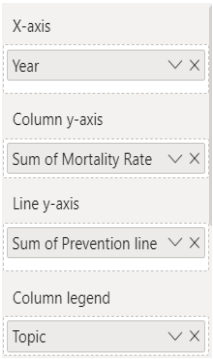
There were columns named *footnote* and *per_value* which did not have any data in it. They were blank therefore we removed them. After completing all this process our data was successfully cleaned and prepared for the final visualisation.

In curating the attributes for visualisation, we strategically opted for the *'Year'* attribute to capture temporal trends in heart disease rates, vital for discerning patterns over time. The inclusion of *'Confidence_limit_high'* serves to integrate a measure of uncertainty, allowing a nuanced representation of the potential range within which true values lie. Finally, the focal point, *'Data_Value,'* was chosen for direct visualisation, providing a clear insight into heart disease metrics. We used the *'Topic'* attribute to create a visually distinct stacked bar chart in PowerBI. Each topic is represented by a different colour, making it easy to see the distribution of heart disease metrics.

3. VISUALISATION

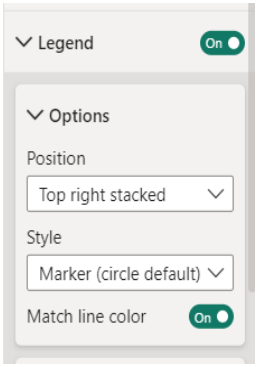
For building our visualisation, we have selected four attributes: Year, Mortality Rate, Prevention Line, and the topic. As we are creating a visualisation on PowerBI of the trend analysis it could have been a line chart or bar chart but as we have different types of heart disease in the topic attribute and for showing the prevention line we opted for the line and stacked bar graph. To showcase the trend analysis, we added the Year in the X-axis and Mortality Rate in the Y-axis. In the Stacked Bar Columns field, we added the Topic as the field contains different types of heart diseases. Creating the line that shows us the prevention of the disease could have been done in addition to the line of the Y- axis field.

Following is the screenshot of the different values imported in the fields in the Build Visual Tab.



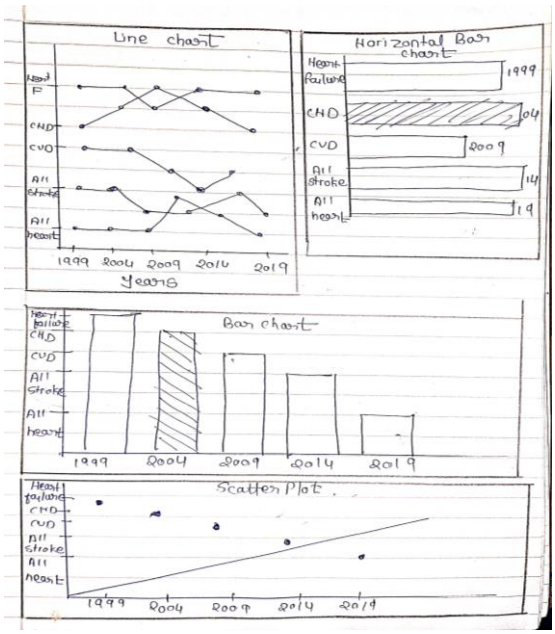
By following the Evergreen's Data Visualization (Stephanie Evergreen, 2016) Checklist, we decided to add the Title and Subtitle in the Upper left corner, from the Format Visualise tab in the General panel.

For aligning the legend on the top right corner, we use the visual format tab from the Visual panel of PowerBI,

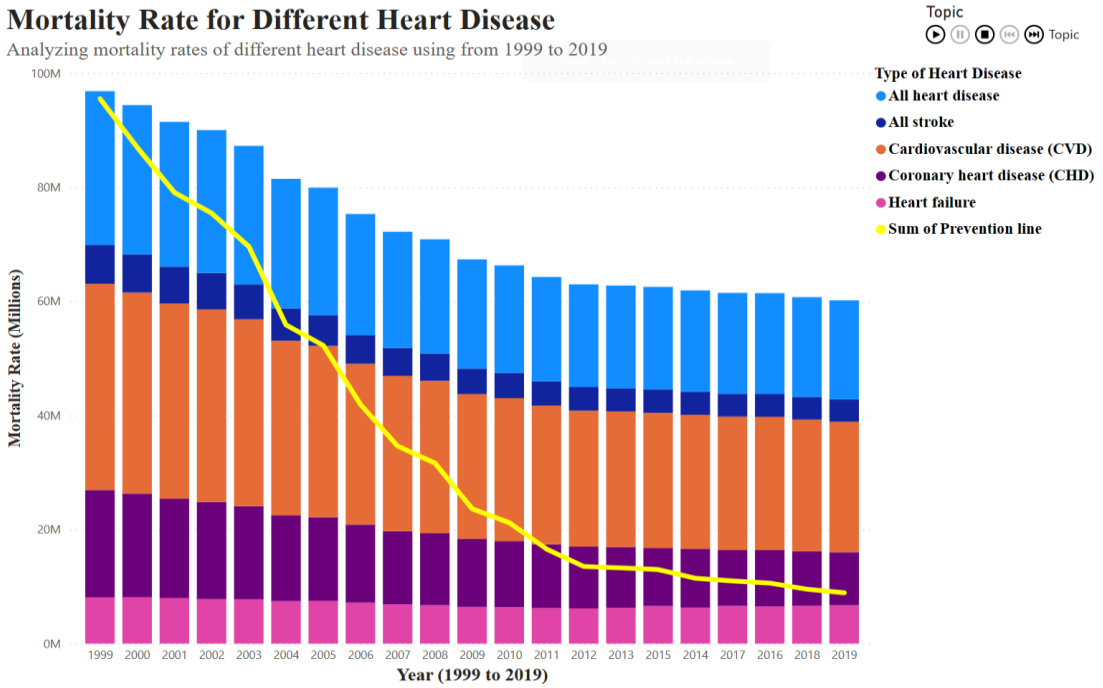


As the sketch of our visualisation is concerned first, we tried sketching four options along with the findings such as line chart, horizontal bar chart, bar chart and scatter plot by keeping our finding constant for all the four charts. While doing the visualisation there were many errors and misinterpretation while performing on these charts, and we were not fully satisfied by the overall look these charts when shown in the visualisation chart and these charts were not able to give justice to the finding which we wanted therefore in opting for a stacked bar chart over other sketched charts, we considered the inherent characteristics of the data and the insights we intended to convey. The stacked bar chart excels in illustrating the breakdown of a whole into different categories, aligning well with our dataset featuring various heart diseases across different years.

Sketch of our visualization



Visualization on PowerBI



Our aim here was to showcase various heart diseases, years, and mortality rates along with prevention that could have been done. Stacking different heart diseases in the stacked bar graph was one of the best options and to represent the prevention of every disease we need a line that runs over the various heart diseases. Thus, we have opted for a line and stacked bar chart such that it shows the relative difference between the type of heart disease and its prevention line. As we are showing trend analysis over a 20-year period it's a temporal type of chart.

Present your design choices - justify your use of colour, shapes, marks, layout, structure, font, labels, etc. referring to books or articles as necessary.

In crafting the visualisation, our choices were intentional, blending aesthetics with clarity. The stacked bar chart was chosen for its ability to vividly portray the prevalence of different heart diseases across the years.

The Colour Palette we choose to represent different heart disease is as follows:

- **Dodger Blue (All Heart Disease):** Conveys a sense of trust and stability, aligning with the overarching theme of heart diseases.
- **Royal Blue (All Stroke):** Represents focus and reliability, emphasising the gravity of stroke-related data.
- **Burnt Sienna (Cardiovascular Disease):** A warm, earthy tone symbolising vitality and caution in the realm of cardiovascular health.
- **Deep Magenta (Coronary Heart Disease):** The rich magenta evokes a sense of gravity, apt for representing coronary heart diseases.
- **Cerise (Heart Failure):** A bright cerise shade conveys hope and determination in the context of heart failure emphasising a balance that maximises the data-ink ratio for optimal perception.

- **Yellow (Prevention Line):** As we are using too many colours in the stacked bar column, we needed a contrast colour to counter all other five colours in one go.

Each colour was selected to distinctly represent a specific category of heart disease, ensuring easy identification.

The inclusion of a line denoting the cumulative impact of prevention measures adds a crucial layer to the data, aligning with Stephen Few's principles of information design that advocate for a holistic narrative within a single visualisation. The decision to place years on the x-axis and mortality rates in million on the y-axis adheres to the principles of good chart design outlined by William S. Cleveland. This arrangement facilitates a logical flow of information and aids in straightforward interpretation. The choice of Times New Roman font was driven by considerations of readability and classic elegance, aligning with Robert Bringhurst's principles in "The Elements of Typographic Style."

Strategically placing the dataset title on the left and a colour legend on the right follows **Gestalt principles**, fostering a clear association between elements. The addition of a subtitle aims to provide context to the graph's purpose, following **Alberto Cairo's guidance** on effective storytelling through data visualisation. The article “A Complete Guide to Stacked Bar Charts” gave us the overall of how to use the stacked bar graph and what colour scheme should be used and how ordering of category levels should be done. (Yi, 2022)For adding the colours into our visualization, we referred the “Coloring Charts in Power BI” (Singh, 2020)articles, it shows the steps how to add different colours in the stacked bar column ,which we didn’t know in the begin.

To get the animation of our visualise, we have installed a new visual on PowerBi that it Play-Axis. Play axis help to get interactivity without any user interaction. For our visualise, we have added a “topic” attribute on the play axis such that it shows different types of heart disease along with the mortality rate and would have been “prevented line” for over 20 years.

The animation helps us in communicating the trendline analysis information such as what amount of people would have prevented various heart attacks if the disease could have been detected at an earlier stage in their lives or given proper care. This animation is played from upward to downwards direction along with the prevention line changing along the various heart attack.

4. CONCLUSION

In conclusion, our exploration into heart disease trends using PowerBI visualisations has been instrumental in unravelling the intricate patterns of cardiovascular health over 1999-2019. The line and stacked bar charts, with their colour differentiations and clear axes, provided a comprehensive view of how various heart conditions contribute to mortality rates across different years and could be prevented if proper detection and care was taken represented by the line in the graph. For our visualisation we have used **PowerBI** tool, with its diverse features and user-friendly interface, we effectively interpreted our data in a meaningful visualisation, enhancing our capabilities to extract valuable insights from the data. We have also imported a **Play Axis** on the PowerBi which is not default available of the PowerBi, which helped us to animate the visualisation.

Through our visualisation using PowerBI, we gained insightful perspectives on the trends in heart disease over the years 1999-2019. The line and stacked bar charts effectively showcased the distribution of different types of heart diseases with contributions to the overall mortality rates. The clear colour differentiations allowed for easy interpretation like coronary heart disease heart failure and others.

Overall, the outcomes of our visualisations significantly contributed for trendline analysis in heart disease, can as follows:

1. Surviving rate of all heart disease is low
2. The Survival rate of all strokes is the least.
3. The significant finding is that the prevention of heart failure is improving by the mid-2012.

The aspects that we think can be improved upon is understanding of various different data visualisation tools and how they work. Initially we faced a lot of problems in understanding the tool, which was a bit time consuming for us along with that having information about how these visualisation tools work would have improved. This will enable us to have a variety of visualisations.

We were unable to change the colour of the stacked bar column as per the colour palette we wanted to, then we searched it on Google and found PowerBI's Documentation. We wanted to make our visualise animation but we didn't know how to understand we watched PowerBi tutorial of the YouTube. Thus, we wanted the animation on our visualise.

As it was a group assignment, we both have contributed equally throughout, starting with finding of dataset which was done by Purva after that sketching the basic estimated design was completed by Vaishnavi, Cleaning of dataset using python and visualisation in PowerBI was equally contributed by both of us and finally for the report we made a google document where we can uniformly edit it accordingly. (cube, 2020)

REFERENCES

1. Centers for Disease Control and Prevention, 2023. *Rates and Trends in Heart Disease and Stroke Mortality Among US Adults (35+) by County, Age Group, Race/Ethnicity, and Sex – 2000-2019.* Atlanta(Georgia): Centers for Disease Control and Prevention.

2. cube, G. i. a., 2020. [Online]
   Available at: https://youtu.be/NvgAehGPqqI?si=UjoMsDz67Hl9viqU
   [Accessed 27 Nov 2023].

3. DataScienceRoadMap, 2023. [Online]
   Available at: https://youtu.be/k47KXaY7_rc?si=brbKvctOgnVyVpHx
   [Accessed 21 Nov 2023].

4. GHUCCTS, 2021. [Online]
   Available at: https://youtu.be/6iO5X_BjJ-4?si=-xevRg_PwWlkMZUQ
   [Accessed 24 Nov 2023].

5. Simplilearn, 2018. [Online]
   Available at: https://youtu.be/fO7g0pnWaRA?si=tFdfxjNPqcN6M4Q3
   [Accessed 25 Nov 2023].

6. Singh, D., 2020. *Coloring Charts in Power BI.* Mumbai: PluralSight.

7. Stephanie Evergreen, A. K. E., 2016. *Evergreen Data Visualization Checklist.* .:Evergreen Data.

8. Williamson, L., 2022. *Undiagnosed heart disease may be common in people with heart attacks not caused by clots.* Dallas(Texas): American Heart Association News.

9. Yi, M., 2022. *A Complete Guide to Stacked Bar Charts.* .:CHARTIO.

IMPORTANT LINKS:

COLAB:

https://colab.research.google.com/drive/1oqzcY-kQKSf2s5vmk953_KUvZKq9LJhB?usp=sharing

SCREENCAST LINK:

https://drive.google.com/file/d/1yGGZlPYy10I5teg3B8OZuiNvIhM-HqUv/view?usp=sharing