CS5044 - Information Visualization

220018772

22 February 2023

1 Introduction

The Pantheon dataset, which consists of 11,341 biographies of well-known people from around the globe, was examined for this report.The datasets, which was produced by Yu et al.[1],contains important details like historical personality domain, industry and the number of Wikipedia page views. This report's objectives are to offer insights into the datasets through visualization and to critically assess accordingly. These insights include gender status, domain occupation status, geographical overview, industry-wise overview, Top L star, Top HPI, and top 10 by page views.

2 Description of the Visualization

For this analysis,a horizontal bar chart was used to show the relationship between the name and their Historical Popularity Index (HPI), L- star, and the number of page views on the English Wikipedia pages of the biography. This allowed us to analyze the number of views on the English Wikipedia biography pages is represented on the x-axis, along with the HPI, which is a measure of a person's historical significance and L-stay,and the significant personality is represented on the y-axis.

In addition to that, We also used a horizontal bar chart to showcase the top 10 wise Trends based on the number of biographies in each industry and their relations to Domain, Country , Occupation, and Gender. The x-axis represents the Industries type and the y-axis is the Number of names associated with that industry. Moreover, this overall dashboard also displays the geographical Map, Domain Status, and gender status in respective shapes.

The horizontal bar is chosen because of its interactive elements allow users to compare the multiple biographies with other elements in clear and concise way.This, in turn, makes it easy to analyze each biography and extract insights from the data.

To accurately represent the gender distribution , we deployed a packed bubble chart that utilized different shapes to represent male and female individuals. This made it easy to quickly and accurately identify the gender breakdown of the data at a glance.

Table[1] provides a list of all attributes represented in the visualization, their attribute type, and their representation via visual variables.

Table 1: Attributes Represented in the Visualization

Attribute	Attribute Type	Visual Variable
L Star	Quantitative	X-axis
HPI	Quantitative	X-axis
Page Views (English)	Quantitative	X-axis
Name	Categorical	Y-Axis
Name (count)	Categorical	Y-Axis
Industry	Categorical	X-Axis
Occupation	Categorical	Color Packed bubble
Domain	Categorical	Pie chart
Gender	Categorical	Shape

Figure[1] shows a screenshot of Dashboard.

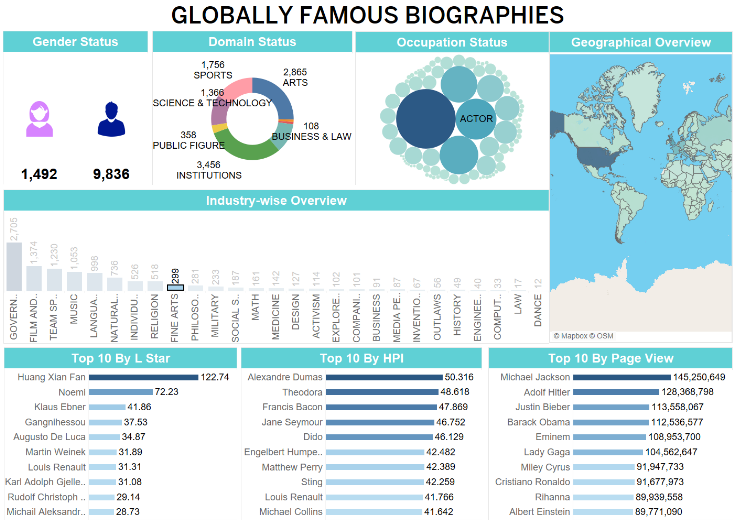

Figure 1: Screenshot of the Famous People Biography Dashboard

3 Insights from the Visualization

The visualization offers several insights from the dataset. One interesting finding is that there is no clear positive correlation between HPI and page views, suggesting that other factors may influence a person's online popularity on Wikipedia.This in return suggests that while historical significance is an important factor in determining page views, it is not the only factor. Other factors, such as contemporary relevance, may also play a role in determining page views on Wikipedia. It is worth noting that the dataset also reveals that people who are historically significant in some countries like India are more likely to have higher page views on Wikipedia, indicating cultural and geographic factors also plays some role.

This visualization also highlights the overrepresentation of certain occupations like politicians, actors, and writers, while also revealing the underrepresentation of women and unknown individuals from diverse backgrounds. .
This might cause biases in the historical record and the way in which historical significance is measured.Looking at the top names in terms of L star, HPI, and page views, we can also gain insights from the visualization. For example, Huang Zian Fan, Noemi, and Klaus Ebner are among the top 10 names according to L star, whereas Francis Bacon, Theodora, and Alexandre Dumas are amng the top 10 names according to HPI. Michael Jackson, Adolf Hitler, and Justin Beiber are among the top 10 individuals by page views.

The chart also breaks down the industries associated with the most important historical figures, with government, film and theatre, and team sports being the most well-represented. However, there are significant imbalances in the representation of certain industries like engineering, computer science, law, and dance. These imbalances may be due to societal biases and inequalities in the distribution of power and influence.

4 Critical Discussion

The limitations of the dataset impacted on the conclusion drawn from the insights gained from the visualization. The missing data and errors in the country, continent, and unknown name detail are the classification that can bring biases into the dataset, which can result in conclusions that are incomplete or inaccurate. Therefore, it is important to make sure the dataset has been completely scrubbed and cleaned.

Furthermore, another limitation of visualization techniques like packed bubble chart is that it provides a high-level overview of the data set. While this can be useful in some cases, it can also obscure important details that might be necessary to fully understand the relationships between variables in the data. But by using multiple visualization techniques and statistical methods, we can develop a more complete picture of the data and gain insights that might not be apparent from a single visualization technique alone.

Despite all these drawbacks, the Pantheon dataset offers useful information for comprehending the distribution of historical importance across the globe and the person's contribution to history. Moreover, all the insights gained from the visualization should be interpreted with caution. Although the visualization can imply connections between factors, it cannot establish a correlation. Therefore, it is essential to carry out additional research to confirm the findings and find any possible confounding variables that might affect the relationship between the variables.

5 Conclusion

In conclusion, the Pantheon dataset has provided a provided valuable insights into the biographies of famous people from around the globe. Through the use of horizontal bar charts, packed bubble charts, and pie charts, we were able to analyze and present the data in an interactive and clear manner. The visualization revealed interesting findings, such as the lack of correlation between HPI and page views, the underrepresentation of certain groups in the dataset, and the strong representation of politicians, actors, and writers among the most notable individuals. The visualization also highlighted the most well-represented industries in the dataset, which include the government, film and theatre, and team sports. Overall, the visualization serves as a useful tool for researchers, historians, and anyone interested in gaining insights into the biographies of famous people throughout history.

6 Reference

1.Yu, A. Z., et al. (2016). Pantheon 1.0, a manually verified dataset of globally famous biographies.Scientific Data 2:150075. doi: 10.1038/sdata.2015.75
2.Munzner, T. (2014). Visualization Analysis and Design. CRC Press.

Word Count: 978