CSE564: Visualization and Visual Analytics

Final Project Report

by

Rutvik Parekh

112687483

INTRODUCTION

This project aims to put forward an objective to perform visual analytics and then to develop various standard and advanced **visualization techniques** and a **clean, interactive, and subtle dashboard** to better understand the data related to various social and economic parameters of the human population in the United States. These socio-economic parameters are analysed at discrete granular levels such as at the level of a particular geographic location, city, zip-code, and state.The parameters to be visualized are an eclectic mix.

a. Mean, median and standard deviation of individual **income**,
b. Average **area of land** owned by each individual,
c. Average **population**, average **male** and **female population**,
d. mean, median and standard deviation **rent** paid,
e. **debt**,
f. **mortgage**,
g. percentage of male and female having **high school degree**, its mean, median and standard deviation
h. mean, median and standard deviation of **family income**

BACKGROUND

Deriving meaning out of socio-economic parameters is extremely important for any society, whether it is developed, developing or an underdeveloped economy. Doing a comprehensive analysis of all parameters related to social and economic life of the population is crucial for the administrators in the government or any private institution such as hospitals, schools, non-profit organizations to devise instrumental and effective schemes and policies so that there is a proper resource allocation and everybody gets what is needed for them.

Let's take the recent example of the **COVID19 pandemic**. Let's say we are in the **office of the New York state government** and we need to **allocate fundings for all hospitals**. If we know where the people of low-income groups are located, then we can send more funding to that particular area so that we can give free care to the needy. Also, knowing the population density of an area helps to allocate funds accordingly. The average age of male and female population will also be impactful here as trends have shown that COVID19 fatalities depend a lot on age. They can easily do so by looking at the visualizations and find out the information pretty easily. This dashboard will also be very useful for **Trump** to allocate a proper share of fundings of the **$2.2 trillion stimulus package**.

In the private sector, let's take an example of a **car manufacturer**. Suppose, he wants to maximise his profits. He will want to **target** areas with **high net worth individuals**. He can have a look at visualizations and will **target areas with very high average income**. Other factors which he can look at is **debt**. People with **high debt** will **not want to buy** a **car** right now. So, he will look at areas which have people with **low debt**. It will also be interesting to look at the **correlation** between all these **socio-economic factors**.

This interactive **visualization and analytics dashboard** can also be used for **research purposes**. Suppose a student is doing data science research about society and the economy and how it affects a larger scheme of things. He can use this project to derive interesting insights related to the data which will help him further in his analysis.

As we can see, there are plenty of applications of this visualization dashboard and hence there is a need for this project.

PROBLEM STATEMENT

If we go **searching for interesting and useful datasets** for **socio-economic study** of the population of the United States, we will **not get a proper, comprehensive** bundle which will contain everything for doing a detailed analysis and deriving meaning out of it. As a result, we need to find data in bits and pieces and then combine them to build a new dataset containing merged entries. Even if we get a complete dataset, we will **not get**

such a dashboard which will contain many visualization tools that will help us see the true nature of the data.

Hence, it becomes a necessity to develop an interactive dashboard containing visualization alternatives so that it can be useful for so many purposes.

DATASET

The dataset used is from **kaggle**. It contains data related to **financial** and **social** parameters as described above. Currently, it has around **80 columns** and **39,000 records** out of which many are **not useful for the analysis and visualization**. We reduced the number of columns corresponding to the parameters (described above) used in our project. Further, data engineering needs to be done using python. In data engineering, main tasks are data pre-processing and data cleaning. We imputed the data using imputation techniques such as **random value imputation**, **kNN imputation, mean imputation, outlier imputation.** Rows containing null values are removed.

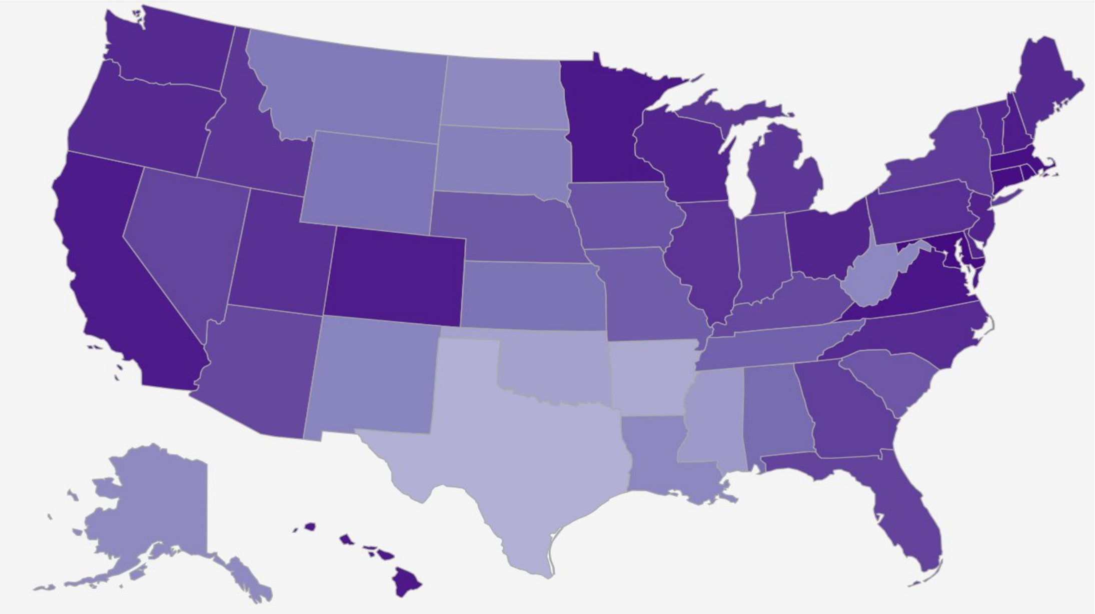
METHODS

Various visualization methods were used to successfully interact with the data and find some meaning out of it.

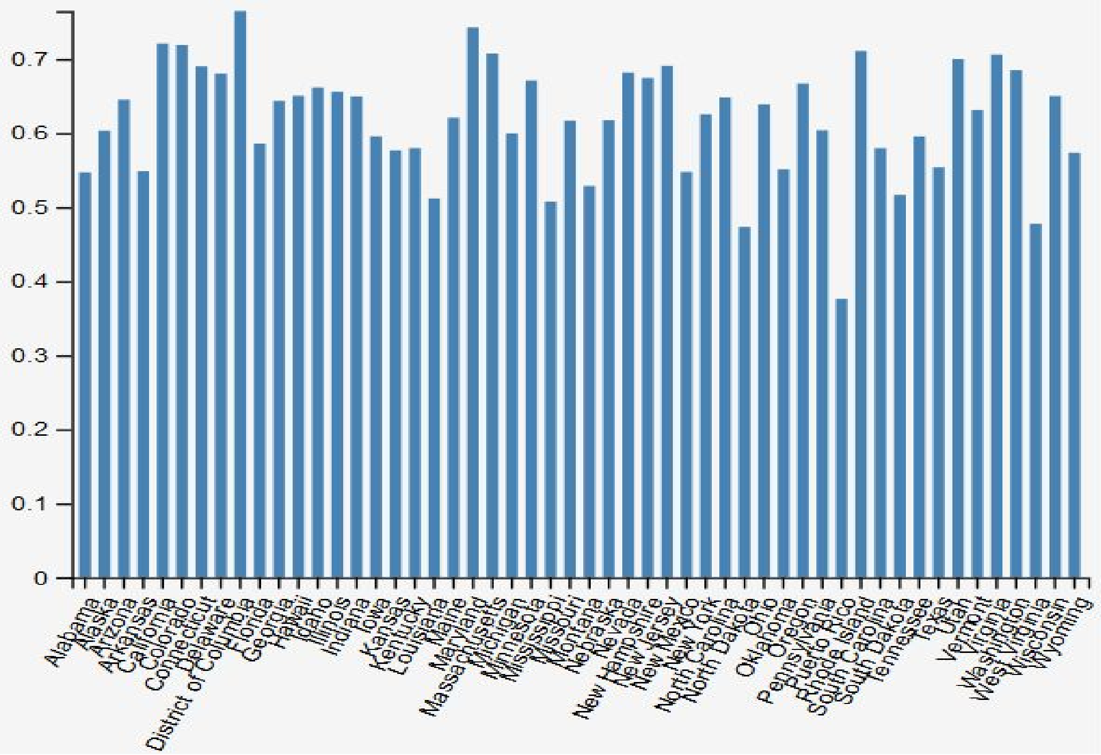
The techniques that were used for visualization are as follows:

1) Choropleth maps: You can select a parameter filter and then the map will get updated accordingly. The map also has a tooltip feature in it.
2) Bar graphs: The bar graph will reflect the parameter selected.
3) Scatterplot: This section contains the scatterplot between the most important parameters to analyze. The most important parameters are: debt, high school degree, mean rent and monthly home cost.
4) Radar charts: We can analyze two different states across the given parameters in a radar chart.
5) Percent Circles: This is a very interesting and innovative visualization technique in which we can compare any state with any state across any filter.

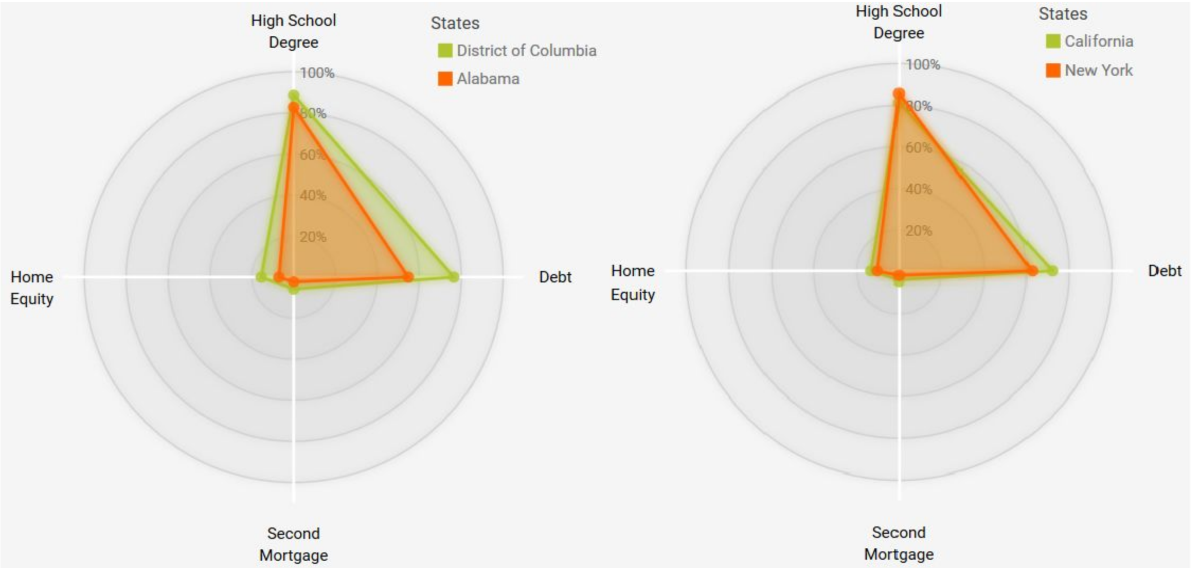
The below shows the representation of all the visualization techniques used:



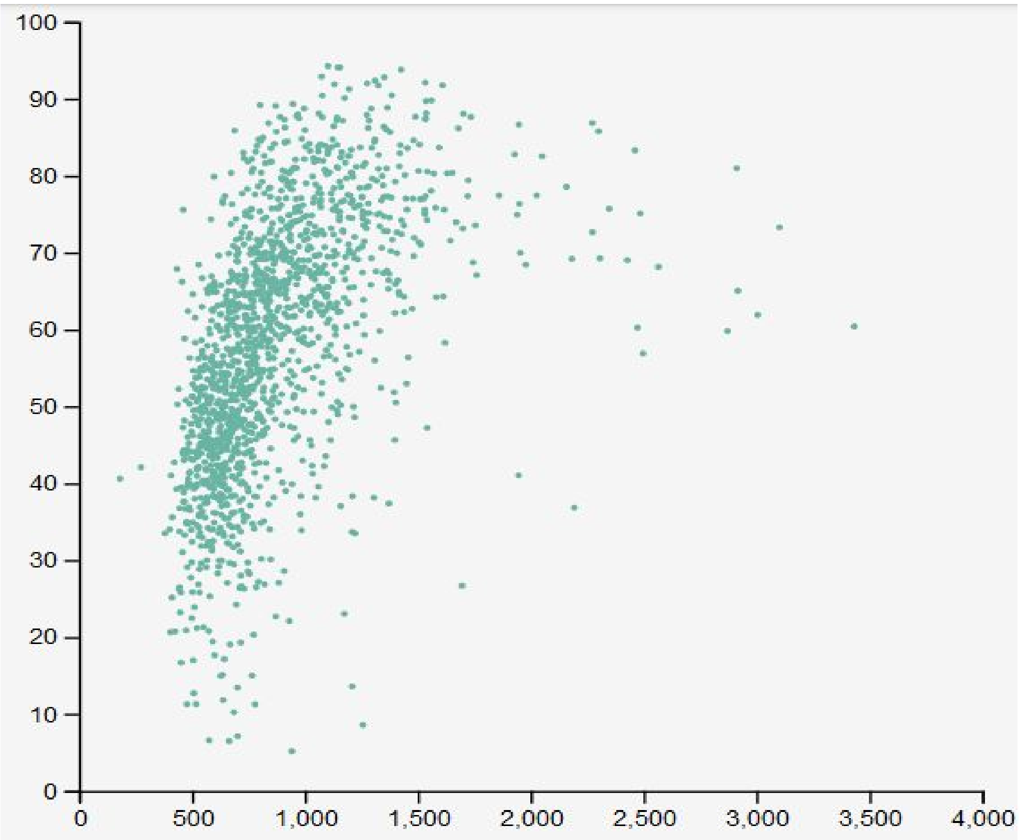
Choropleth map demonstration for home equity loan



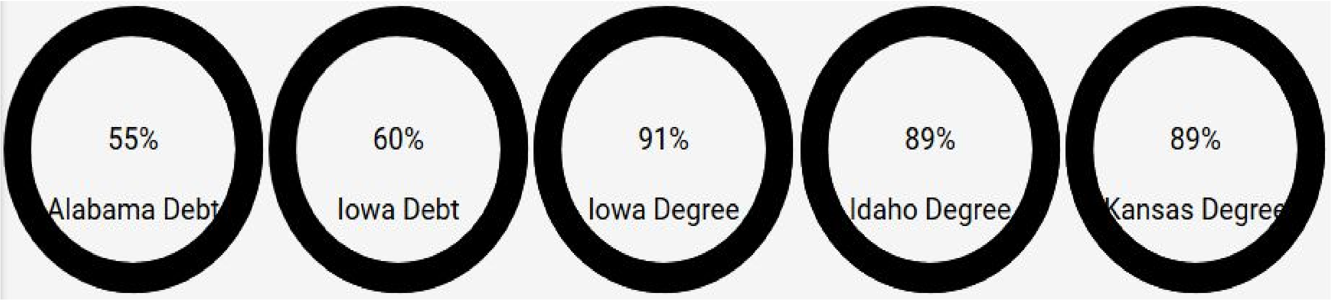
Bar graph demonstration for the percentage of debt statewise



Radar chart demonstration: For the states of District of Columbia, Alabama, California and New York



Scatterplot for Debt % vs Mean Rent



Percent Circle Representation

FINDINGS

By using the above visualization techniques for the said dataset, we got to know many interesting and amazing findings about the data which can be used by various private organizations and government agencies in their plans.

Some of the insights are as follows:

1) Average rent across the US is $1054
2) High School Degree percentage is positively correlated with debt taken in a family
3) The middle part of the US has among the highest percentages of high school degrees
4) Average rent is the highest in the western and eastern parts of the US, especially in the states of California and New York
5) Monthly home cost is the highest across stretches of the eastern US
6) The middle part of the US has the highest amount of area of land and water. Alaska is an outlier here with an insane amount of both of the resources
7) 63% of people in the US are under some kind of debt
8) 86% of people in the US has at least a high school degree
9) There are 14.3 million households across the US with a home equity loan
10) 3% households across the US have taken a second mortgage loan
11) The average monthly home cost added with the mortgage in the United States
12) 51% of people across the US are married
13) The divorce rate across the US is 10%

TECHNOLOGIES USED

The following technologies were extensively used for this project:

1) javascript
2) HTML
3) CSS
4) d3.js
5) python
6) flask
7) Bootstrap template

CONCLUSION

We conclude that the socio-economic parameters of the US were comprehensively, visually and interactively analyzed using standard and non-standard visualization techniques to extract some concrete findings and meaning out of the data