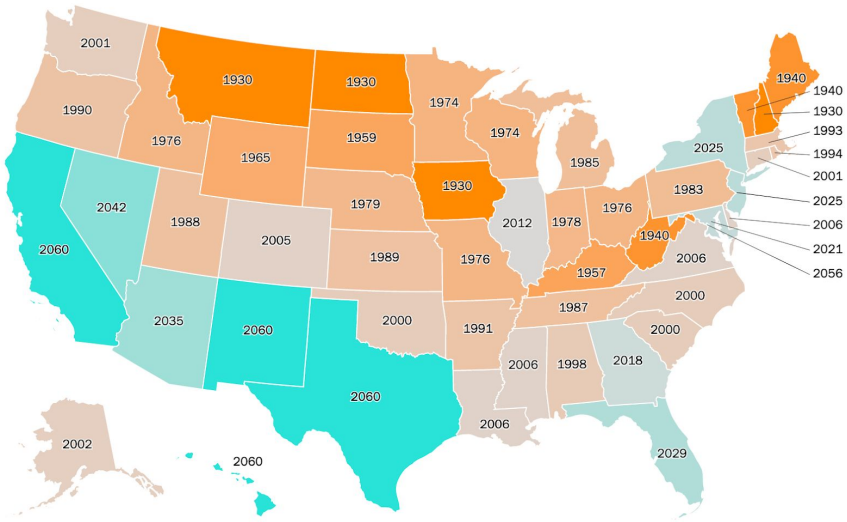
CSE 564 - VISUALIZATION AND VISUAL ANALYTICS

AMERICAN SOCIETY ANALYSIS: PROPOSAL

By: PRATIK PANTODE (112027432), ANUJ VERMA(112504481)



INTRODUCTION

In this project, we aim to build a dashboard to visualize the change in American society and demographics over time and also, with the help of the visualization, analyse the effects of these changes on various other factors of the society to produce interesting, statistically significant results that can be used for better governance by giving a basic, high level view into the American society.

DATASET

For our data, we are mostly looking at the United States Census data. We have used the wonderful tool “The National Historical Geographic Information System (NHGIS)” which provides easy access to summary tables and time series of population, housing, agriculture, and economic data for years from 1790 through the present. Currently, our high level goal

is to observe the effects of demographic changes on the society. NHGIS provides us with mostly economic data. We are also currently looking for other aspects, mainly health and education and plan to include them as well to study the effect of population on these variables. Our dataset is at a granularity level of state and county so we'll not be able to analyze census tracts.

STATEMENT OF NEED

In our opinion, government policies should keep up with the times and always be open to change in keeping with the changing human societies and their needs and concerns. Society is constantly evolving and we want to develop a way to visualize this evolution through time so as to make it easier for policy makers to heed to the changing needs of their county, state and county.

Our mission is to visualize the data in a way that pinpoints and brings forth the ways the demographic change affects the economics and other areas of the society and to show that these are important factors that need to be considered when designing policies and governance. But also, secondarily, we want to bring forth the idea that using regression and other statistical methods, future demographic data can be inferred and this data can use the same visualizations that we create to help those holding public offices to visualize in an intuitive way the needs of their community.

Hypothesis

Our entire visualization is build on the hypothesis that change in the demographics has economic as well as other societal ramifications. Broadly, the different aspects of demographics we have are:

1. Total Population
2. Population divided by Location (Urban/Rural)
3. Population divided by Sex
4. Population divided by Race (White/ African American/ American Indian/ Asian-Pacific Islander/ Two or More Races)

5. Population divided by Origin (Native/ Foreign)
6. Population of Immigrants based on Place of Birth
7. Population based on Sex and Educational Levels

And the different societal measures we have are these:

1. Population divided by Employment Status, Sex and Labor Force
2. Household Income (Median/ Binning)
3. Family Income
4. Per Capita Income
5. Population below poverty level
6. Number of Housing Units

Using the demographics and the different measures, we can visualize and test various hypothesis. For example, here are some high-level hypothesis that we can visualize:

1. Based on the number of people that have come into the state/county from different countries (#6), has the education levels increased or decreased(#7). These can answer questions like, is the education system coping with the influx of immigrants? Or is the population living below poverty line increased? This can give insights like, the government has not done enough to rehabilitate the immigrant population.
2. We can compare states based on the Number of Males/Females and visualize how it affects, say Family Income or number of people below poverty line?

We are still trying to get more data for comparing the counties/states, mostly in the area of healthcare and quality of life.

PROJECT DESCRIPTION

We plan to show as well as visualize statistics pertaining to demographics and societal attributes on a per county and on a per state level. This would likely show which states and counties are doing better than others and in what areas or which groups of people are doing better and which groups need help. We also want to mention if the results found by us are statistically significant or not.

Approach

Firstly, we will be working with the data a bit to do data cleaning and filling in the missing values, of which there are many. For examples, for some states, population information starts from 1790 and for others, it starts later. After data cleaning and bringing data into an uniform structure, we first want to think about which analysis are more important than others and focus our attention on those. In other words, we will try to reduce the data to a subset that we feel contains most of the interesting hypothesis. We will be dividing our project into two types of visualizations. In the first part, we just plan to visualize how the population demographics have changed over time. In the second part, we plan on visualizing the effects of these changes on the society.

Implementation

We will use Python (and data science libraries of Python like Pandas, skLearn, scipy etc.) for the backend processing of data. Flask will be used as a framework for running the application. For the frontend, we'll use D3.js, JavaScript/JQuery, HTML and CSS . For creating interactive maps, we are looking at javascript libraries like DC.js, Crossfilter.js, Leaflet.js or D3-geo. We are also planning to implement bootstrap for making our dashboard more responsive, as it is equipped with responsive layout and 12-column grid system that helps dynamically adjust the website to a suitable screen resolution. This will enable us to run our dashboard from any device like mobile or tablet, etc.

CONCLUSION

In conclusion, we can say that we are hoping our visualization can help us with some interesting analysis and some meaningful, statistically significant results that can be used in the real world to help with governance and bringing to forefront counties/states that are failing to provide a good economic, educational and healthcare opportunities to its people. Some example results that we are hoping to uncover are:

1. In New York State, as the number of people residing in New York State that were born in Europe has increased, so has the Urban population. This has the implication that if a state has primarily a rural population and wants to urbanize themselves,

they have to provide incentives to people of European descent to move there as they help in urbanization.

2. In Suffolk County, the number of female with a higher education level has led to (or correlated to) an increase in Per capita income. Another county may look at this and provide scholarships to women.

REFERENCES

Steven Manson, Jonathan Schroeder, David Van Riper, and Steven Ruggles. *IPUMS National Historical Geographic Information System: Version 13.0* [Database]. Minneapolis: University of Minnesota. 2018. http://doi.org/10.18128/D050.V13.0

AMERICAN SOCIETY ANALYSIS: PRELIMINARY REPORT

Building on the proposal, we have so far achieved the following:

1. Data Cleaning for Demographic Data
2. Developed the responsive layout for the dashboard using bootstrap, which is accessible from any device.
3. Visualized the Demographic Data for Race and Sex for each state of USA.

DATA CLEANING

The dataset for state wise division consisted of multiple columns separated yearwise for different categories. On top of that there was some old data related to previous states that are apparently no longer there with state names such “Alaska Territory”. These needed to be integrated with the current state “Alaska”. Also, for each of the demographic attributes (Gener, Race, Urban/Rural, Native/Foreign Born), the ratios were calculated because population over time will increase, it is the ratio that is more interesting and representative of the change. Number of Females increasing over time is not as interesting as the number of females per 100 males increasing. The below diagram shows ratios of the different races.

```
11 df.shape
11 (51, 46)

12 W, AA, AI, APAC
12 ('B18AA2010', 'B18AB2010', 'B18AC2010', 'B18AD2010')

13 statecols = statecols[1:]
   No output

14 df.drop(columns=statecols, inplace=True)
   No output

15 df.head()
15
```

	STATE	W_Ratio_1970	AA_Ratio_1970	AI_Ratio_1970	APAC_Ratio_1970	Total_Population_1970	W_Ratio_1
1	Alabama	0.735688	0.262318	0.000709	0.001284	3444165	0.738723
2	Alaska	0.788220	0.029666	0.054184	0.127930	300382	0.783078
4	Arizona	0.906289	0.030123	0.054104	0.009484	1770900	0.899295
6	Arkansas	0.814183	0.183251	0.001047	0.001519	1923295	0.827848
8	California	0.890137	0.070172	0.004562	0.035129	19953134	0.844180

5 rows × 26 columns

DEMOGRAPHIC DATA VISUALIZATION

We have categorised the demographic data into various attributes like Sex, Race, Age, Rural/Urban, Native/Foreigner and Immigrant Origin. We have so far completed the visualisation of “Sex” and “Race” for each state and the entire country for a particular year. Also, we are displaying the total population of the entire country which helps us to analyse how the population has increased and decreased over the years, mainly from 1790 to 2012. and its effect on a particular attribute.

Analysis of “Sex Ratio”

The sex ratio is the ratio of males to females in the population (normalized to 100).We are visualizing the sex ratio of each state and the entire country for a particular year. Also, how the sex ratio has changed for each state and the country over the years. After visualisation of the data, we found that New York state is having one of the lowest sex ratio over the years. Also, sex ratio of the entire country has increased from 94.47 in 1970 to 96.70 in 2010.

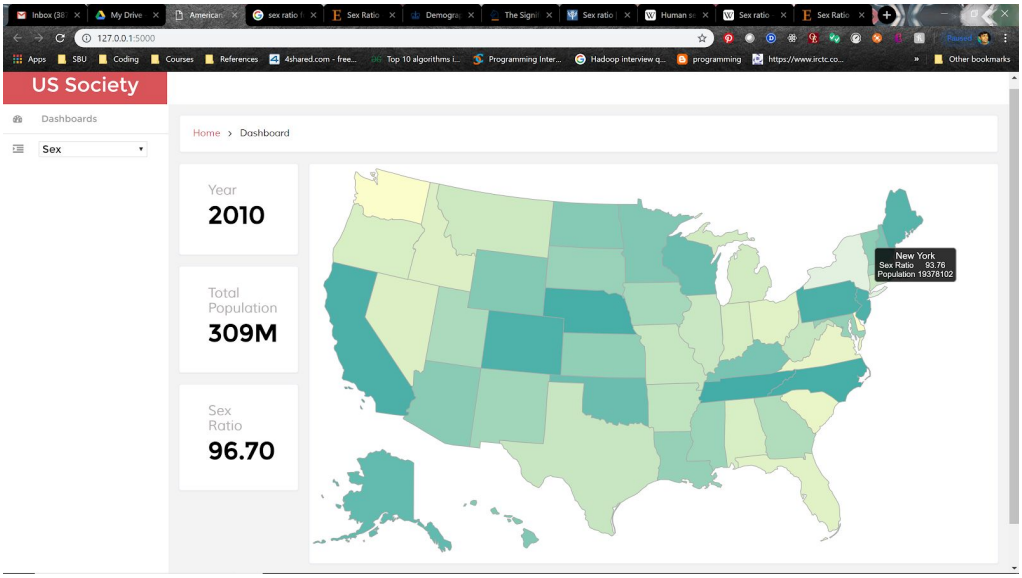


Fig : Dashboard displaying Sex Ratio and Total Population for year 2010 for the country. Sex ratio and population of each state is also displayed on hover over each state in the US map.

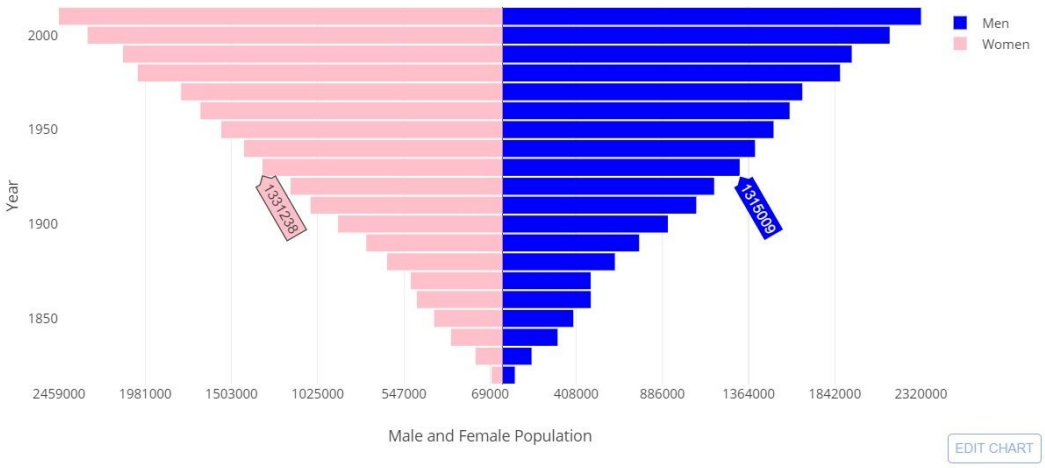


Fig : Population Pyramid Chart for the state of Alabama (made using python plotly library). Planning to implement these analysis for each state in d3 in the final version.

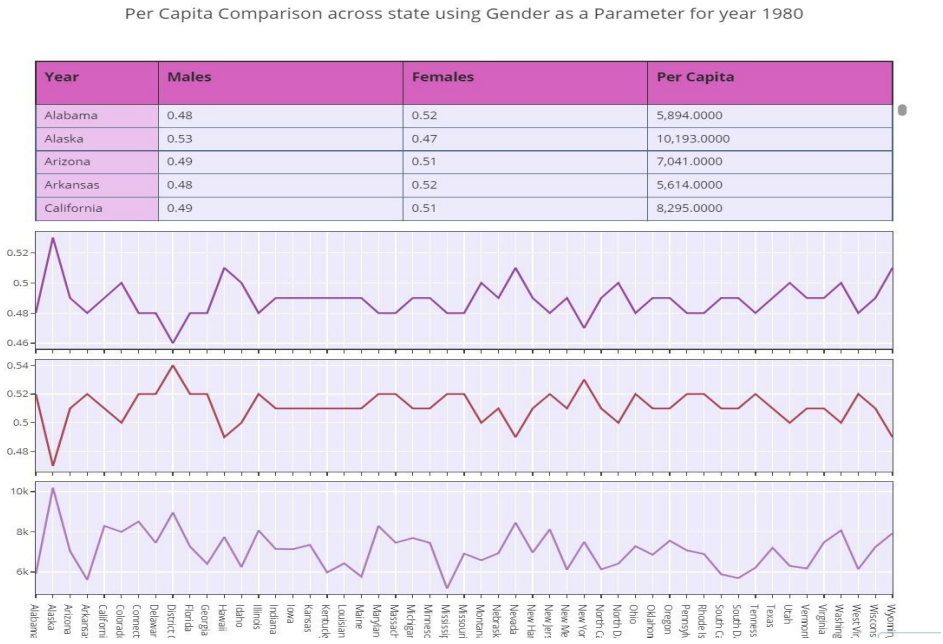


Fig : Population Pyramid Chart for all states for the year 1980 (made using python plotly library). Planning to implement for the analysis over a period of time in d3 in the final version.

Analysis of "Race Ratios"

We have categorised the ratios of different races for a particular state into four categories:

1. Black or African American Ratio (Af-Am ratio)

2. American Indian and Alaska Native Ratio (Am-Ind ratio)
3. Asian and Pacific Islander Ratio (As-Pc Ratio)
4. White Ratio (White Ratio)

These ratios helps us to find the diversity of each state. Also, how the migration of people from different races has changed over the years for each state and the entire country. This kind of information also helps us to find out about the states which have renowned universities or industries. For example, we found that state like California or Texas has more people from different races over the years.

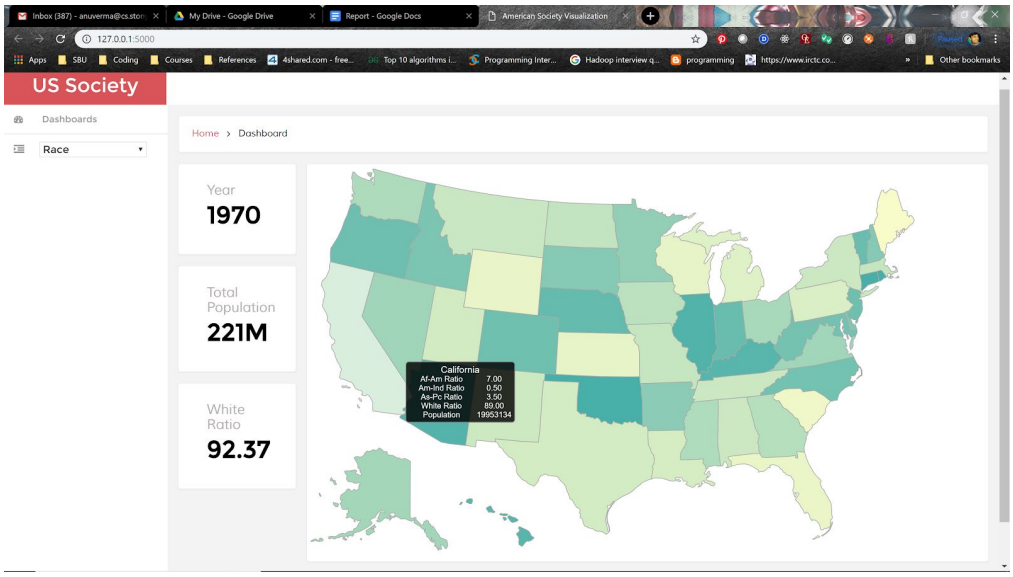


Fig: Dashboard displaying White ratio(signifies that USA is dominated by White people in terms of population) and Total Population of the country. Ratio of different races is also displayed for each state on hover over.

REMAINING WORK

- ❏ Similar visualisation as displayed in dashboard screenshot is remaining for other attributes like Age, Rural/Urban, Native/Foreigner and Immigrant Origin using USA map.
- ❏ We will be using regression models to show the relationship between dependent and explanatory variables like Race and per capita income on a scatterplot.
- ❏ We have to analyse the data for other attributes like per capita income, household by income, Total Households, Families over the years. We are planning to use the

bubble chart to visualize the same, where the the size of the bubbles will represent the population with respect to Sex, Race, Age, Rural/Urban, Native/Foreigner and Immigrant Origin.

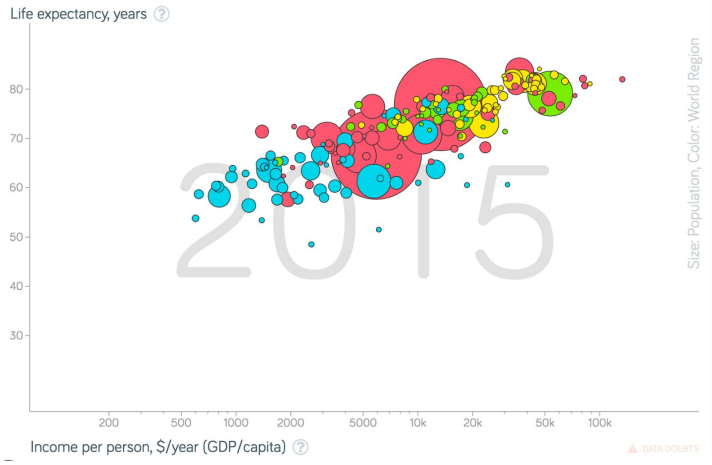


Fig: Mock bubble chart to handle above mentioned functionality

- We are also planning to visualize the data using parallel coordinates for each state for a particular year over a period of time considering various attributes like Male ratio, female ratio, race,age, Rural/Urban, Native/Foreigner and Immigrant Origin.

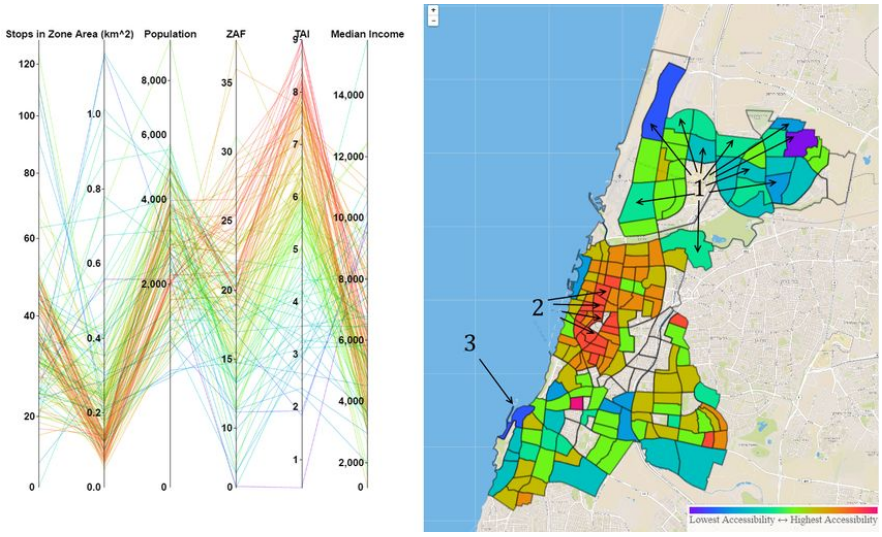


Fig: Mock Parallel coordinates for each state considering various attributes