Interactive Visualization of US Census and Demographic data

CSCI-55200 Spring 2018

Aakarsh Nadella and Prem Chand Avanigadda

Under the Guidance
of

Dr. SHIAOFEN FANG
Professor and Department Chair
Department of Computer and Information Science
Indiana University-Purdue University Indianapolis

April 30, 2018

CONTENTS

1. Abstract

Huge data is being generated every day from various parts across the world. Analyzing this vast amount of data, gives us useful underlying insights which in turn helps us to find approaches of decision making to elevate productivity. There are various mechanisms to analyze the data such as visualization, descriptive and predictive analytics etc. Visualization helps us to better understand the data and to make inferences. In this project we will be working on United States Census Demographic data to analyze each state according to the ethnicity, type of work, how people commute, and other statistics related to state. Also, we shall find insights from the data such as the states with highest poverty rate, employment rate, which state has the highest people employed in professional jobs, states carrying out more constructional activities etc.

2. Motivation

As we are foreign nationals, we are completely unaware of the demographic details of United States. United States is a country with high development, creating more jobs, the second largest exporter of goods and is also one among the largest economies in the world. It is also known for its diverse population with people migrating from all around the world, comprising various ethnic groups and mixed culture. All these factors lead us to know in detail about each state in various categories such as population, gender, race, average income, unemployment rate etc.

3. Introduction

The United States Census Bureau [1] is a government agency that deals with the statistics of the American population and economy. This agency organizes the census for every ten years and releases data to the government and to the public. Based on population released in this data, seats are reserved for House of Representatives for each state. An estimate of censuses is carried out in between this decennial period by American Community Survey, which is in turn used to make necessary improvements to transportation, establishing new schools, housing, public utilities and health facilities. This data can be further used by government to implement new policies to reduce poverty and unemployment by establishing industries in these regions, which creates new job opportunities. One of the focus in this project is about employment type whether it is public work, private work, self-employed or family work and the sector they work in. This gives us work sector popular in these states and find a way to enhance the work culture environment to improve quality and quantity of the product. Also, we shall find out the backward

states based on poverty and unemployment so that appropriate steps would be taken to improve economic condition of the people of these states.

4. Dataset Description

The Dataset used for this project is US Census Demographic data took from DP03 and DP05 tables of "2015 American Community Survey 5-year estimate", which is published in Kaggle for public use.

Dataset containing 74002 records with 37 attributes each. The following section explains each attribute in detail.

4.1. Attributes in detail

1. CensusTract: This is unique ID for each record to distinguish from other records.
2. State: Name of state
3. County: Name of county
4. TotalPop: It is Total population of each county with respect to the CensusTract
5. Men: Total number of men
6. Women: Total number of Women
7. Hispanic: Percentage of Hispanic or Latino in total population
8. White: Percentage of White in total population
9. Black: Percentage of Black in total population
10. Native: Percentage of Native American or Native Alaskan in total population
11. Asian: Percentage of Asian population in total population
12. Pacific: Percentage of Native Hawaiian or Pacific Islander
13. Citizen: Number of people having American Citizenship
14. Income: Median household income of that county
15. IncomeErr: Median household income error
16. IncomePerCap: Income per capita of that county
17. IncomePerCapErr: Income per capita error
18. Poverty: Percentage of population under poverty
19. ChildPoverty: Percentage of children under poverty level
20. Professional: Percentage of total employees involved in management, business, science and arts.
21. Service: Percentage involved in service jobs

22. Office: Percentage involved in sales and office jobs
23. Construction: Percentage involved in construction jobs
24. Production: Percentage involved in production, transportation and material movement.
25. Drive: Percentage of people commuting alone in car, van and truck to work
26. Carpool: percentage of carpool in car, van and truck
27. Transit: Percentage commuting through public transportation
28. Walk: Percentage of people walking to work
29. OtherTransp: percentage using other ways to commute
30. WorkAtHome: Percent working at home
31. MeanCommute: mean commute time to work (Minutes)
32. Employed: Percentage of people employed (16+)
33. PrivateWork: Percent employed in private sectors
34. PubicWork: Percent employed in Public jobs
35. SelfEmployed: percentage of people self- employed
36. Familywork: percent of unpaid family workers
37. Unemployment: Percentage of unemployment

5. Data Preprocessing

We need to prepare data to our project requirements to achieve targeted results and good insights without any misleading information. Our main concentration is to draw the useful inferences only for the states in USA, but we had data focusing on each county. So, we need to aggregate the 74002 records to each state respectively, firstly we removed CensusTract and County attributes which decreases the diversity in our data, then we used Microsoft Access to group the records based on State and then aggregated them. We got 56 records having aggregated data of US states and territories.

Other section of this project gives quick analysis of Top 10 states in terms of economy, poverty, employment, ethnicity and other aspects. To get this analyzed data, we wrote SQL queries on aggregated data in MS ACCESS.

6. Visualization techniques

Census data having mostly numerical values and our targeted goal is to compare states and draw inferences between them. So, we need to choose visualization techniques such that they can handle numerical data efficiently without any loss of information, and they can communicate user through actionable insights and useful in informed decision making. We find Geo visualization, donut chart, bar chart and 3D donut chart best describe our data by following best practices mentioned above.

a. Geo Visualization: Graphical representation of data linked with geographical information beneath. We used D3 inbuilt functions like d3.geo.path() to visualize our data.
b. Bar chart: In this visualization, value is represented as height or length of the bar, we used concept in pareto charts to arrange the bars in descending order from left to right.
c. Donut chart: they are like pie charts, but we can use the space inside the donut to represent the type of attributes visualized, these are more efficient in using area when compared to pie chart.
d. 3D donut chart: These visualizations gives effect of three dimension to donut chart, which are more effective to understand the data.

7. Implementation and Results

To implement the following visualizations, we have used D3.js, which is a library in javascript to implement dynamic and interactive visualizations. In addition, we also used HTML and CSS for making the web page stylish and ordered.

7.1. Geo-visualization

As our dataset contains the different attributes relative to each state, it would be a better way to visualize the data on a map. For this we used an existing JSON file containing Id's, names, latitudes and longitudes of of each state. We had synchronized our dataset to this JSON file using state Id's. The heart of our visualization is linking of our dataset to this JSON file and the rest of the part became easier as everything from now on will be inter-related.

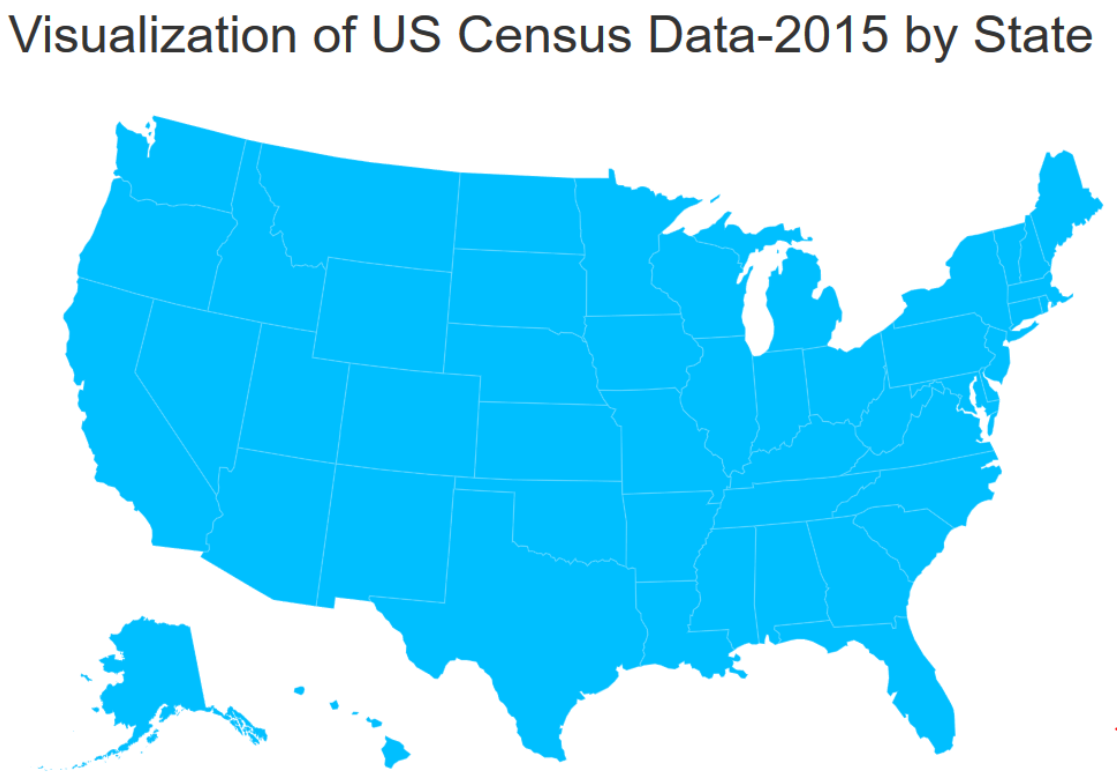

Figure 1: Initial Map

We have provided a tooltip functionality to this map, once the user hover mouse over state of his choice, that state will be highlighted in a different color and essential details such as name of the state, it's total population and average income will be displayed in a text box. This can be illustrated in the figure 2.

Attributes used: State, TotalPop, Income

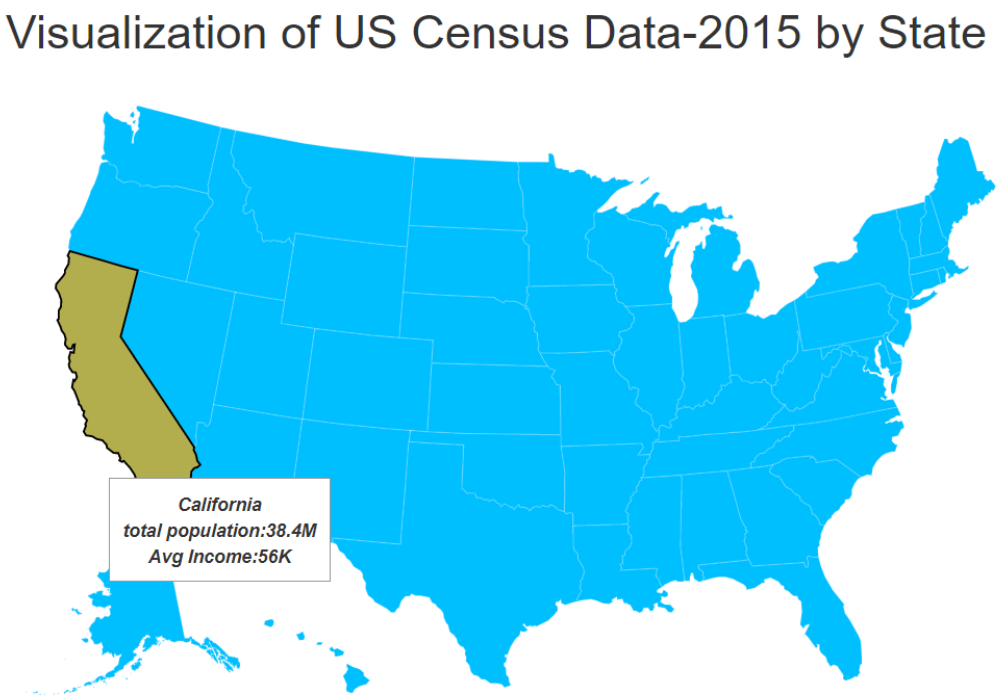

Figure 2: US Map with essential details

Once the user clicks on state of his choice other visualizations such as Donut Chart, Bar Chart and 3D-Donut will pop up showing the details of that state on which mouse is clicked.

7.2. Donut Chart

We have implemented 2 different donut charts pertaining to the state which user clicks. The first chart shows the people's choice to commute to work. There are five different ways to commute such as drive, carpool, transit, walk and other transport. A legend is created assigning different colors to each category and, tooltip functionality is implemented to show the percentage occupied by each category in donut as shown in figure 3. From this visualization, it is observed that most people use their own vehicles for commuting to work and the least preferred one is the transit. Government can take necessary steps to make people aware of the traffic problems and pollution caused due to this. This can be reduced if more people use transit, but at the first step government needs to ensure that transit covers all the populated places and workplaces very frequently.

Attributes used: State, Drive, Carpool, Transit, Walk, OtherTransport

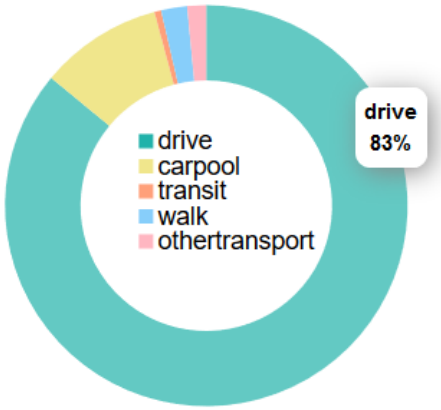

Figure 3: Mode of Transport

The second chart, figure 4 shows the proportion of male and female population in that particular state distinguished by two different colors, it is observed that most of the states have almost equal number of males and females which is a good sign in a gender equality society.

Attributes Used: State, Male, Female

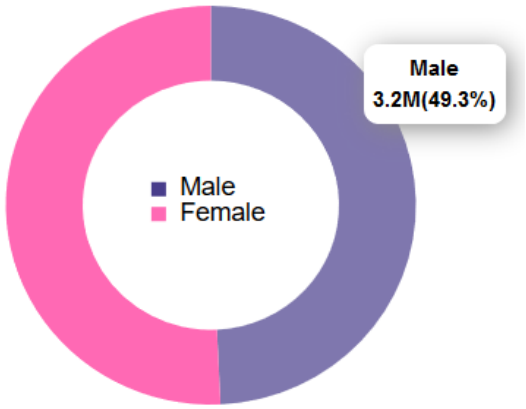

Figure 4: Gender Proportion

7.3. Bar Chart

Using this visualization technique, it will be easier to understand people's choice of work and in which sector state had more employment. Employment sector is classified into five different categories. Once the user moves mouse over a bar, the corresponding bar will be highlighted in a different color and the value corresponding to that bar will be displayed in a text box as in Figure 5.

Attributes Used: Professional, Service, Office, Construction, Production.

Figure 5: Employment Stream

7.4. 3D-Donut

This is another form of visualization used for each state to know the type of work people of that state are associated in. This type of work is broadly categorized into 4 different categories. 3D - Donut is used to visualize the portion occupied by each category of donut for each state. 4 different colors are used to differentiate each category by defining a legend and the percentage contributed by each category is placed on that portion of donut itself. It is observed people are mostly associated in private work and the least is family work.

Attributes Used: PublicWork, PrivateWork, FamilyWork and SelfEmployed

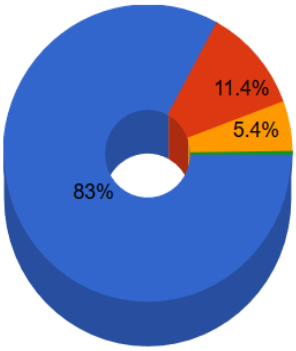

Figure 6: Type of Work

7.5. Barchart with drop-down filters

Visualization is further made interactive by providing a dropdown filter. The major part of analysis can be done here. 3 different filters are placed for development index, ethnicity and type of work. Development index filter contains Income, unemployment rate, poverty rate, child-

poverty rate, employment rate and Income per capita. Ethnicity filter contains Hispanic, Asian, White, Black, Native and Pacific. Type of Work filter contains public work, private work self-employed, family work, production, construction, professional and service.

When you select a category from any one of the filters, the top 10 states pertaining to that category will be displayed in a bar chart. When you hover the mouse over each bar, the value corresponding to that bar will be displayed in a text box. The below figure shows the top 10 states with Hispanic population. Puerto Rico tops the list. It is to be recollected that this territory was once a part of Spanish empire until the famous Spanish-American war in 1898. Later it was occupied by United States but the traces of most of the people there are still from Spain.

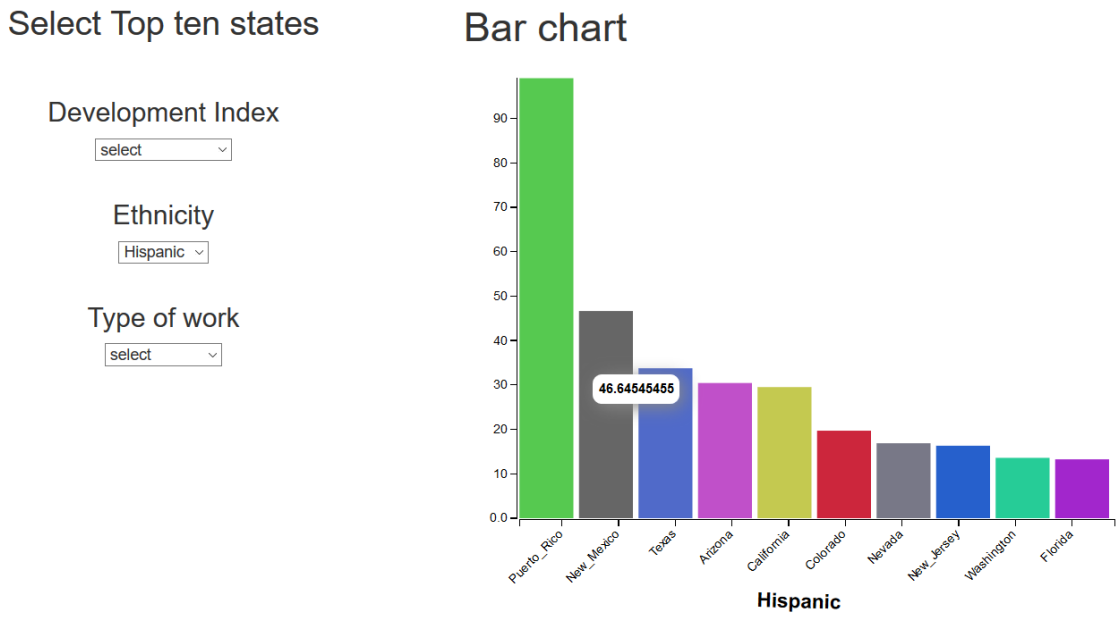

Figure 7: Top 7 Hispanic populated states

7.6. Dashboard

We created dashboard to ease the context switching and to show all relevant visualizations at the same place. It is help full to hold the all visualizations at one place which results dynamic experience to user. Positioning and size of visualizations on dashboard is challenging because user need to know where to make changes to view dynamic content. So, we made sure the size of geo visualization is more than other visualizations to grab the

attention of user to map, which is also very important to switch states and to see dynamic visualizations based on relative state data.

We created Nav-bar to switch between geo-visualization and Top ten states, “about” navigates to page with project details.

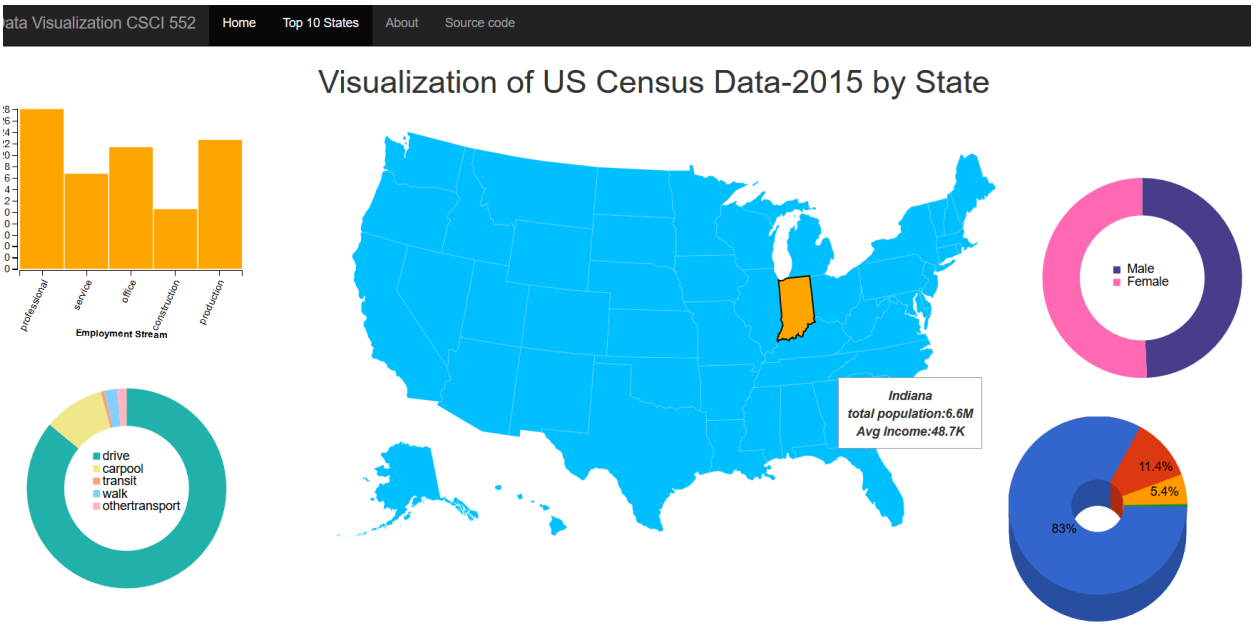

Figure 8: Dashboard

8. Future work

In this project our concentration is only at the state level. We would like to extend the implementation of this work to county level and provide more interactions to the user, which will be helpful to identify underdeveloped counties.

9. Conclusion

From the visualizations it is observed that Puerto Rico, Mississippi, Georgia and Arizona are the territory/states having highest poverty, unemployment and child poverty rate. Government should take necessary steps to boost up the economy of these territory/states by creating new jobs in these areas. Employment and Income is found to be highest in the states of District of Columbia, New Jersey and Connecticut due to more concentration of professional

and service jobs. Private Work and Production are highest in the state of Indiana whereas lot of construction activity is going on in the state of Nevada which is evident in the projects such as Tesla Gigafactory which is of worth $1.3 billion, and roads and infrastructure development.

10. References

1. https://en.wikipedia.org/wiki/United_States_Census_Bureau
2. https://github.com/d3/d3
3. http://bl.ocks.org/michellechandra/0b2ce4923dc9b580992