Group 15 Atindra Bandi, Marie Gleichauf, Hope Knopf, Clay Mason

U.S. Census Income Data Analysis

Introduction

Rampant income inequality in the US is undoubtedly one of the biggest challenges facing the U.S. today. In a society where the top 1% of the population earn more than 20% of all income[1] and the wealthiest 10% of households own 76% of US wealth[2], the social climate warrants extensive study. Income and wealth inequality mean less economic growth for the US because those who make and own less cannot contribute to the economy as much as they could if there was less disparity. Identifying some of the underlying causes and addressing them with policy changes could benefit the U.S. enormously. We posited that educational achievement played a substantive role in predicting income levels and poverty. The Bureau of Labor Statistics indicates that the higher the level of education, the lower the unemployment rate, and the higher the income level. This suggests that higher educational attainment is in demand and thus wages are rising for those positions, while jobs for lower levels of education are not in demand and have stagnant wages[3]. Furthermore, individuals with higher income levels tend to see greater increases in their income as they progress through their career, while lower income workers see less dramatic growth[4]. To put it simply, the rich and educated get richer, and the poor and less educated barely move. The dichotomy of this system has continued to perpetuate income inequality. We therefore decided to investigate why some Americans make considerably more than others by examining the relationships of variables such as educational level, age, and occupation taken from the 1994 census on whether an individual made above or below $50K that year. We then looked at 2015 American Community Survey (ACS) data on income and unemployment on a regional basis and to consider how we could expand on and contextualize our results.

Project Goals

1. Data: For our project, we explored US Census Data from 1994, found on Kaggle.com. We wanted to explore the effects of different variables on income in order to better explain income differences in the US. Additionally, we collected more recent data for contextual information on the changes in demographics from 1994 to 2015. The 2015 data is also from Kaggle.com. In the 1994 dataset, we had over 30,000 data points and 14 features to work with. The variables included race, ethnicity, national origin, education level, marital status, gender and age. Additionally, the 2015 data set obtained from

the American Community Survey (ACS) included income, ethnicity, poverty and unemployment distributions across all the counties in every state. Our goal was to determine which variable are most influential in determining a person's income, what income inequality looks like across the U.S., and how these things have changed in the last two decades.

2. Importance: Income inequality has been steadily rising in the United States and this trend has far reaching impacts. We wanted to investigate the interactions between income inequality, education level, and unemployment as these are all tied together and constitute some of the most persistent and widespread issues facing the nation today. We feel our study is important because understanding the factors affecting income level could inform concrete policy changes to address these issues. For example, instituting policies that subsidized education for those with historically less access, policies like tax incentives for companies to create more entry and mid-level positions that require less advanced levels of education, and policies that incentivize companies to expand to regions with higher unemployment levels, could all help to alleviate the disparity.

Exploratory Analysis

a. **1994 data set:**

Before starting the exploratory analysis, we did some data cleaning. We removed rows that had missing values, and looked for strong correlations among the 14 variables in the 1994 data set. We primarily did this through making pair plots of the variables in Python. We found that 75% of our data included people making less than 50K and the data had a higher proportion of males (67.6%), clearly leading to a biased sample. Our data was also about 86% white, which we initially thought skewed our data. However, looking at other historical data sources from the U.S. census website confirmed that this makeup, when white and Hispanic were grouped in the same category, was consistent with the actual population in 1994.

When we looked at the age variable, we found a range between 17 and 90 years old, with most people falling between about 25 and 45 years. We found that those under 20 and over 70 rarely have instances of income >= 50K (Figure 1). For those between 20 and 70, there was a large variance in the percentage of people making above or below 50K. When looking at education, we found that the most individuals had only high school education, many people had some college, many had a bachelor's degree, and very few had a masters or doctorate (Figure 2). As you might expect, those who went to professional school like law school or med school, had the lowest chance of making <50K, followed by those with doctorate degrees and master's degrees.

In the case of occupation, there's a somewhat uniform distribution among occupations, excluding the armed forces and protective services (Figure 3). However, occupations such as managerial executives, sales positions, and specialty professionals were the most positively correlated with income >=50K while occupations such as farming and service workers were negatively correlated. When we looked at hours worked, the greatest majority of people worked around 40 hours a week, which makes sense given that 40 hours is the standard in most U.S. jobs. We also saw that those who worked less than 40 hours a week were less likely to make >=50K and those that worked more than 40 hours a week were more likely to make >=50K.

b. **2015 ACS data set:**

After analyzing some of the potentially interesting variables in the 1994 census data, we considered the 2015 ACS data that contained more regional variables to contextualize our findings and see how demographics had changed. We had 3220 rows and 37 columns of data containing demographic information such as population, average income, unemployment rate, poverty rate, ethnicity, and gender distribution across all the counties in every state. We again first cleansed the data by removing the counties which had missing values and excluding the District of Columbia in our analysis.

On aggregating the total population across all counties within a state, we found that California is the most populous state, followed by Texas and New York. To get a better understanding of the population across different states, we plotted a density heat-map of state populations (Figure 4). We also analyzed interesting differences in the population of different counties within these top states (Figure 5). Los Angeles was the largest county by population in California; the largest county in Texas was Harris county, which contains Houston. One interesting finding is that LA's population in 2015 was greater than that of the 8th largest state in the US (Figure 6).

We further looked at the gender and ethnic distributions across the U.S. to better understand how have the demographics changed over time. We found that the Asian and black populations have increased the most over the last 20 years, while the white population declined slightly (Figure 7). Lastly, we looked at the income and unemployment densities across all the states and tried to understand the relation between income and unemployment (Figures 8, 9, 10). We found that income and unemployment have a negative correlation, i.e. higher rates of unemployment lead to lower average income. However, this is not always the case. We see that while California has one of the highest average incomes, it also has very high unemployment rates. One potential explanation to this could be a higher income variability and inequality in California, with a large amount of people in the top 1% of incomes, particularly in LA and San Francisco, and a large amount of poverty as well.

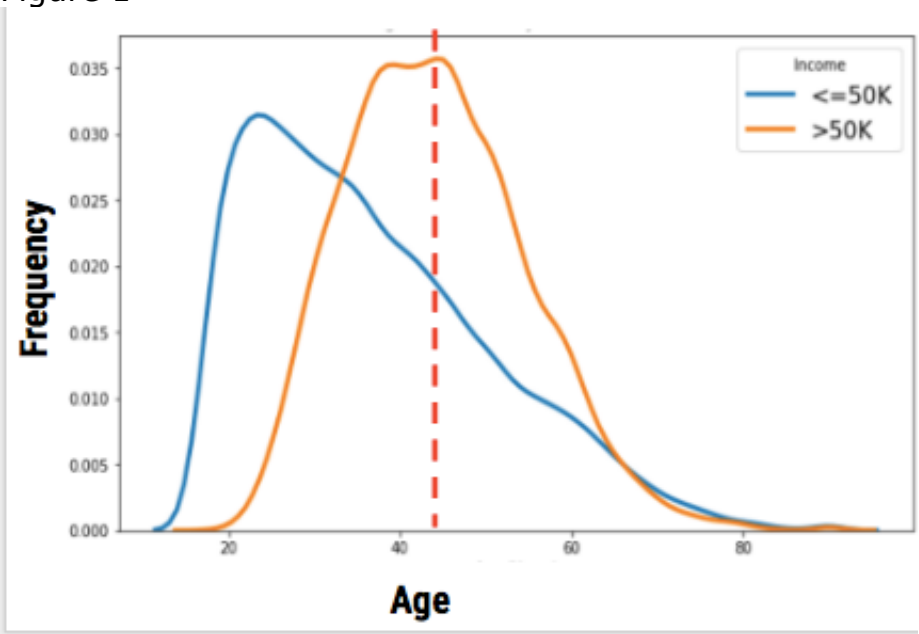
Solutions and Insights

We looked at Naive Bayes and Decision tree models on our data before deciding to focus on logistic regression, which gave us 84% accuracy both in sample and out of sample. We found education, age, marital status, and occupation to be most significant variables in classifying a person as making above or below 50K. These variables were all positively correlated with making >50K. Higher levels of education lead to higher incomes, and occupations like managerial executive and business owner showed the highest correlation. There was also a negative correlation between income and marital status, with those who were unmarried generally making less than 50K. However, this is most likely due to the interaction between age and marital status, in that those who are unmarried are younger on average and likely earlier in their careers and thus making lower salaries.

We identified the following limitations of this data set: a) The sample was not representative of the US population, b) The data was 20 years old, c) The sample did not include everyone living/working in the US (such as legal immigrants), d) it did not include actual income values and instead was a threshold of income. We aimed to correct for a) and b) in our analysis of the 2015 dataset. After contextualizing our findings from the 2015 ACS data with respect to the income levels across different geographical locations, we concluded that education and occupation are the most important predictors of income. Further, above average income does not always mean unemployment is low for a state or region. As we saw with California, high income variability and inequality can lead to both high average income and high unemployment. As discussed in our introduction, this analysis could be useful in determining social and economic policy to reduce economic inequality and poverty in the U.S. As economic inequality continues to rise, we believe it is more important than ever to dig deeper into the data on these issues and to inform decision making from a data-driven perspective.

Plots and Tables

Figure 1



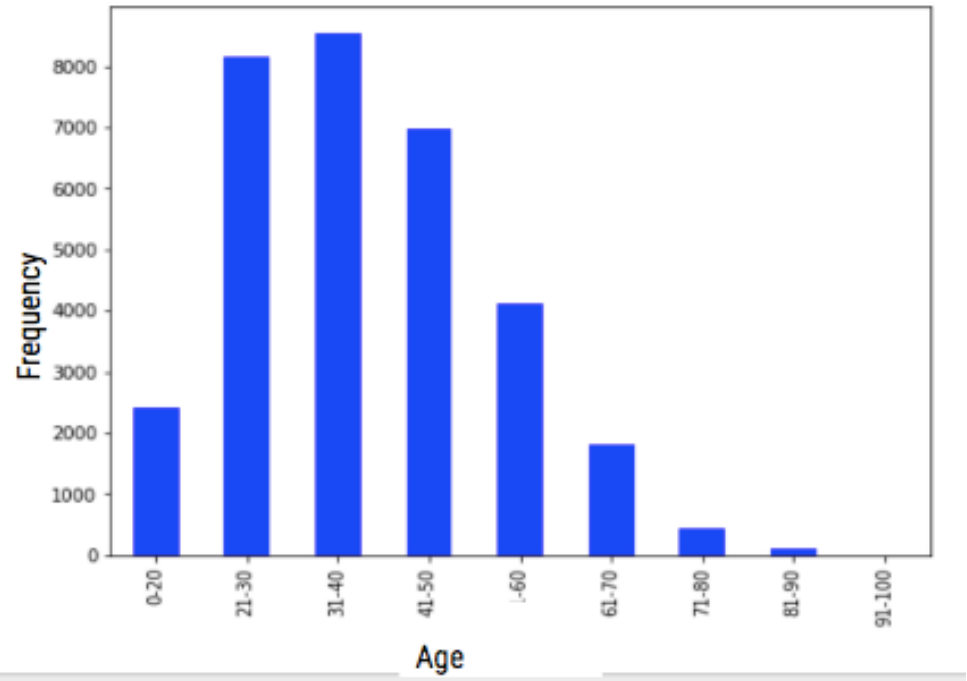


Figure 2

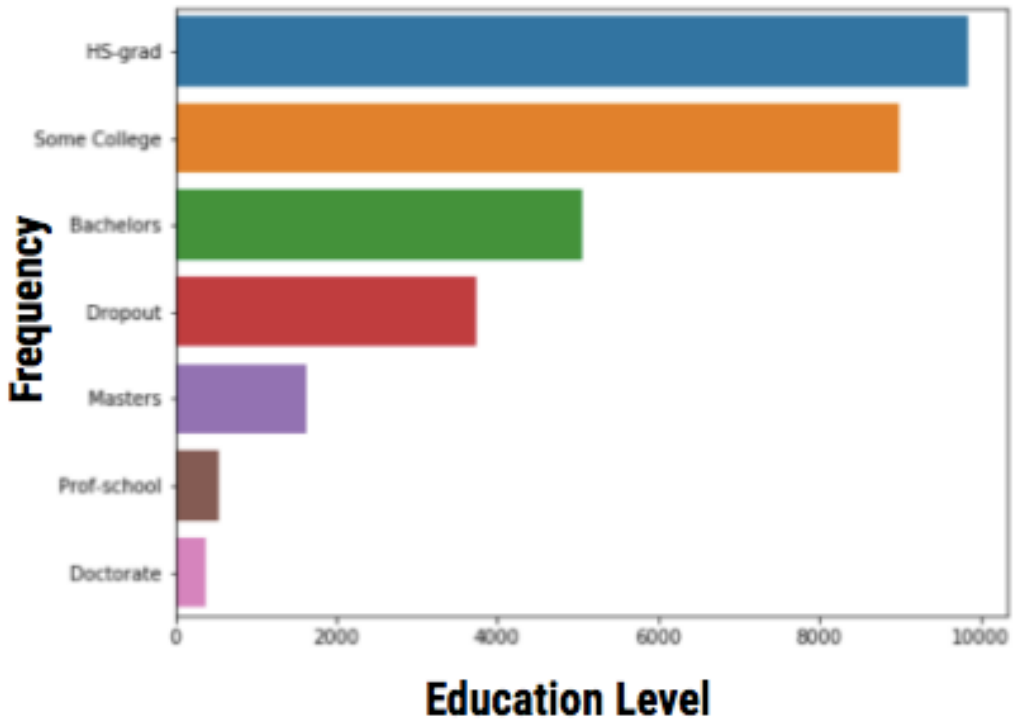


Figure 3

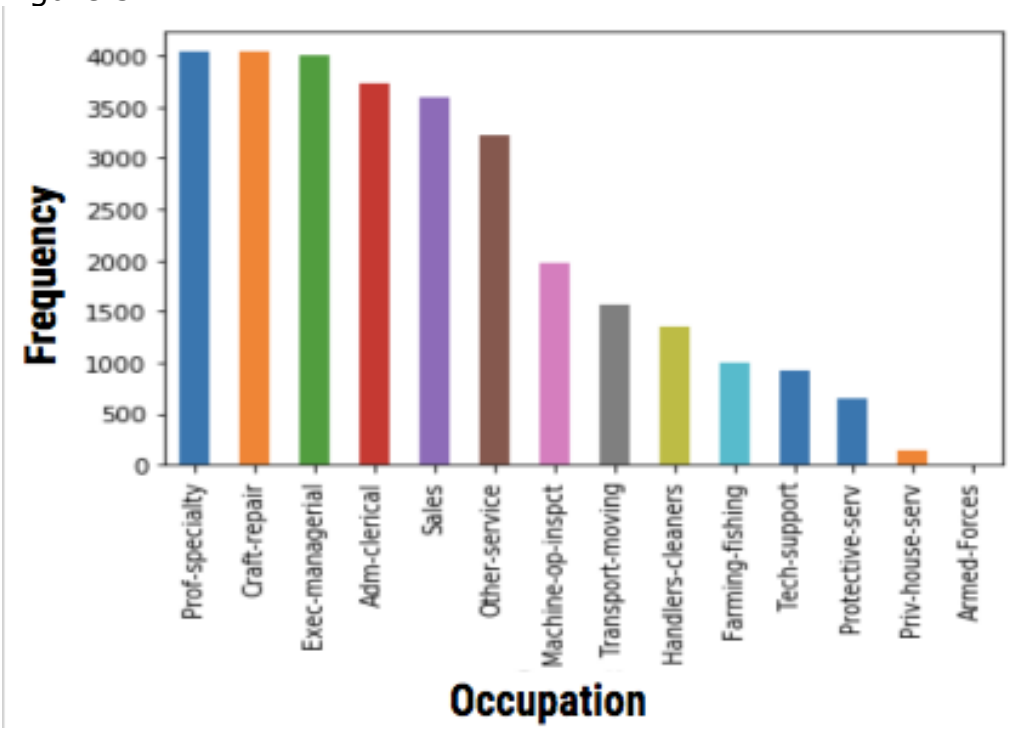


Figure 4

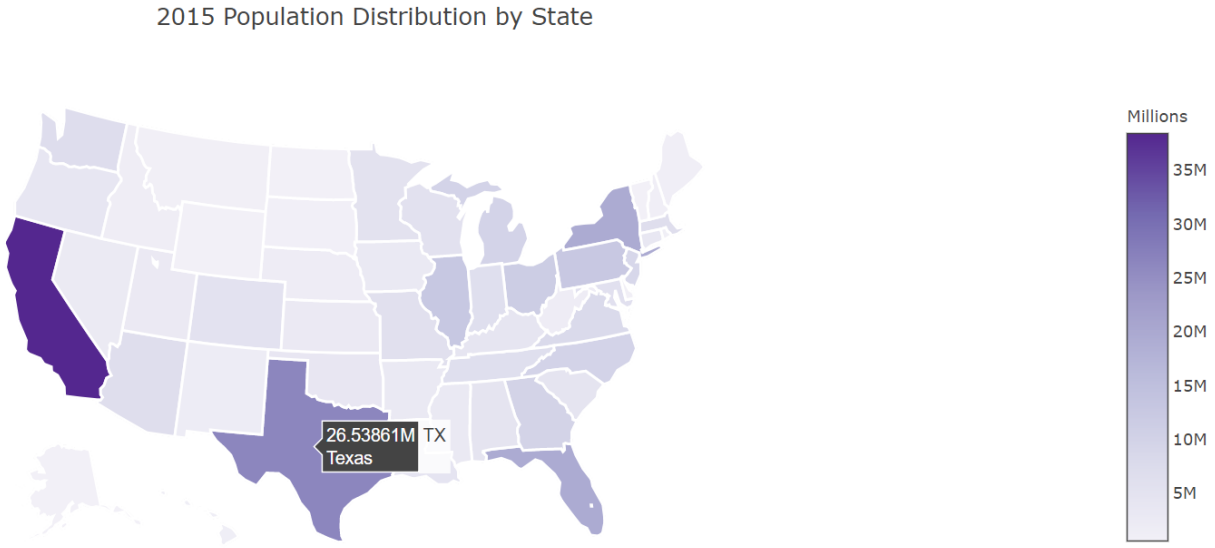


Figure 5

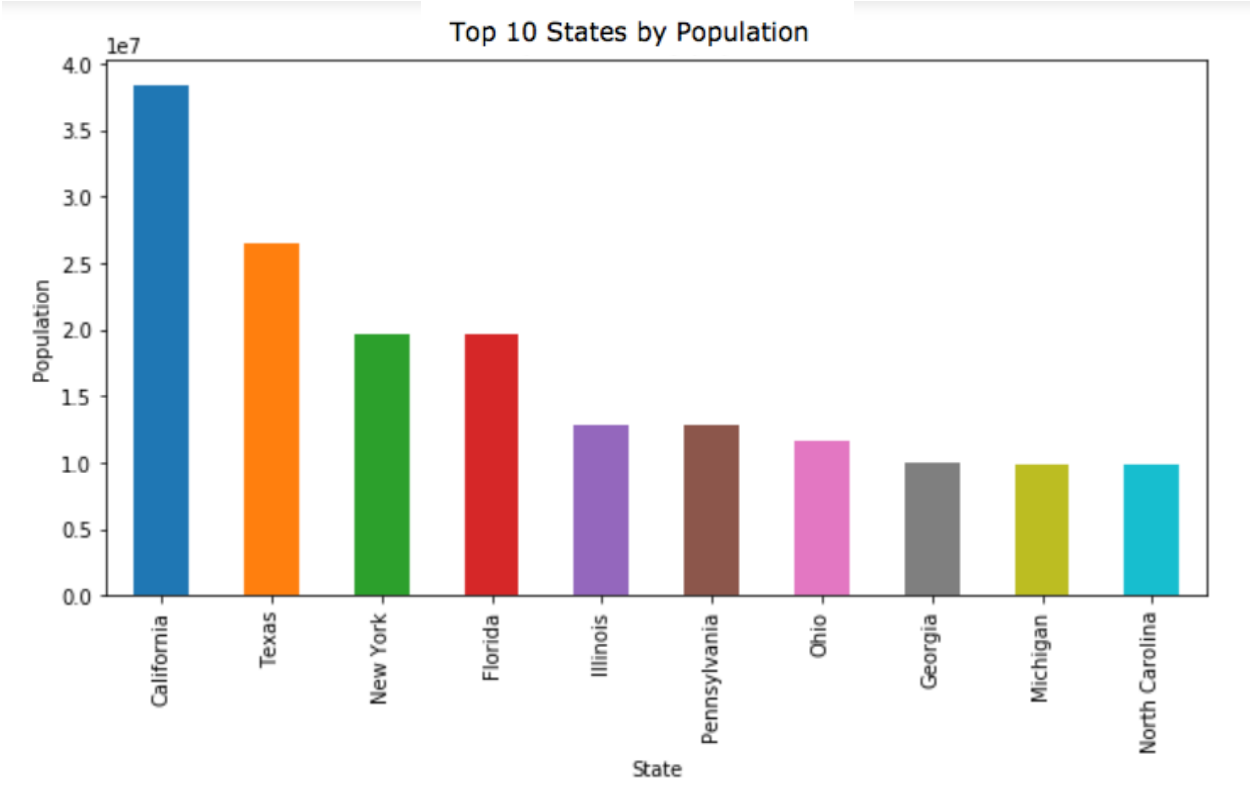


Figure 6

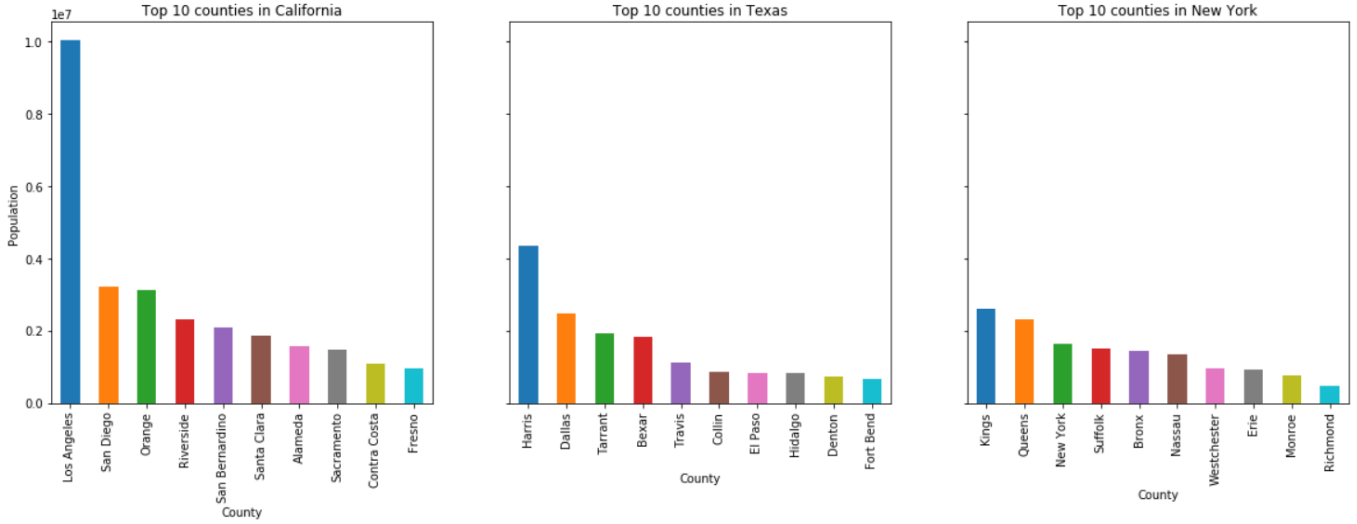


Figure 7

Race	1994	2015	Δ
White	86.0%	81.6% (63.1%)	-4.4
Hispanic	X	18.5%	X
Black	9.2%	12.4%	+3.2
Asian-Pac-Islander	3.0%	5.4%	+2.4
American-Indian_Eskimo	1.0%	0.7%	-0.3
Other	0.8%	X	X

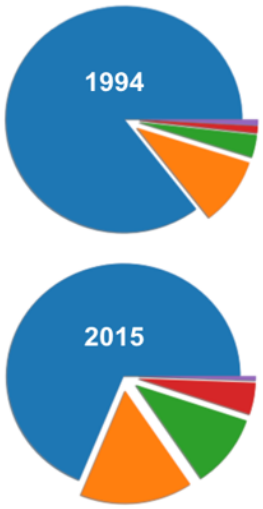


Figure 8

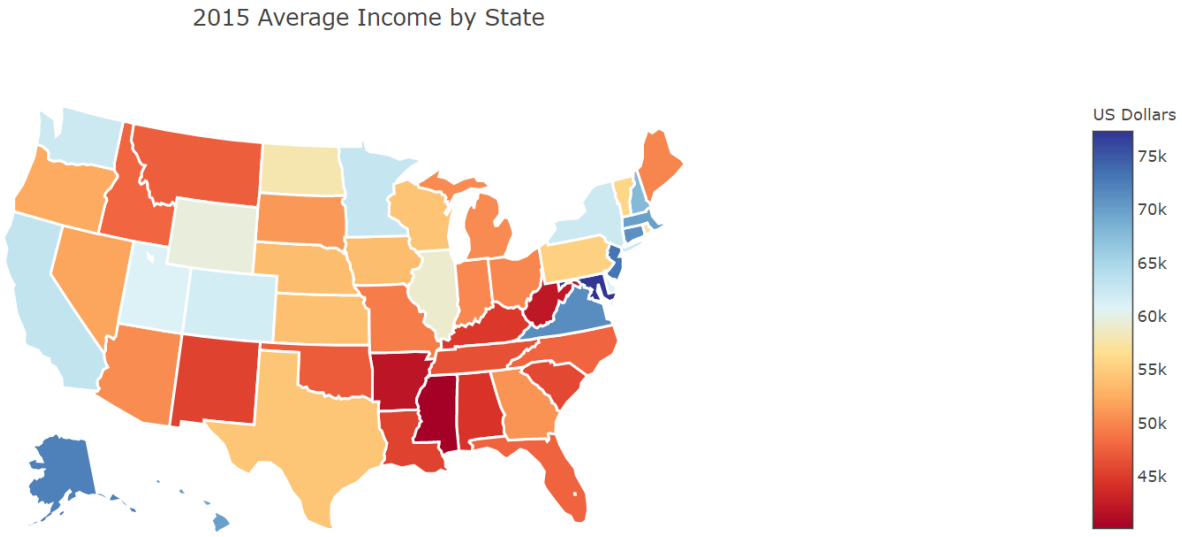


Figure 9

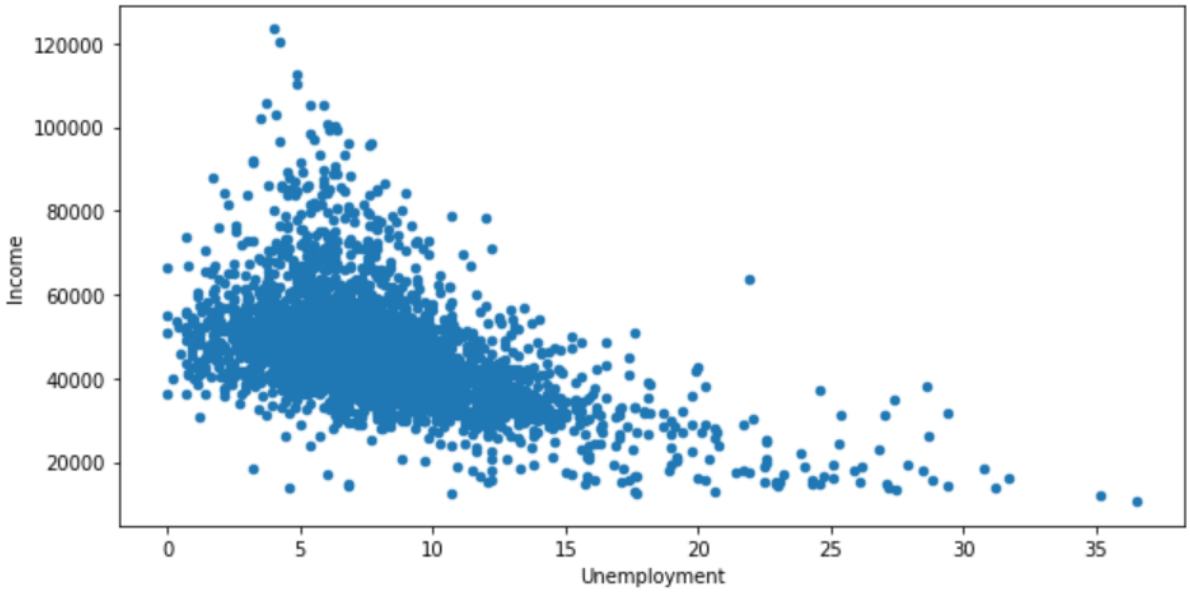
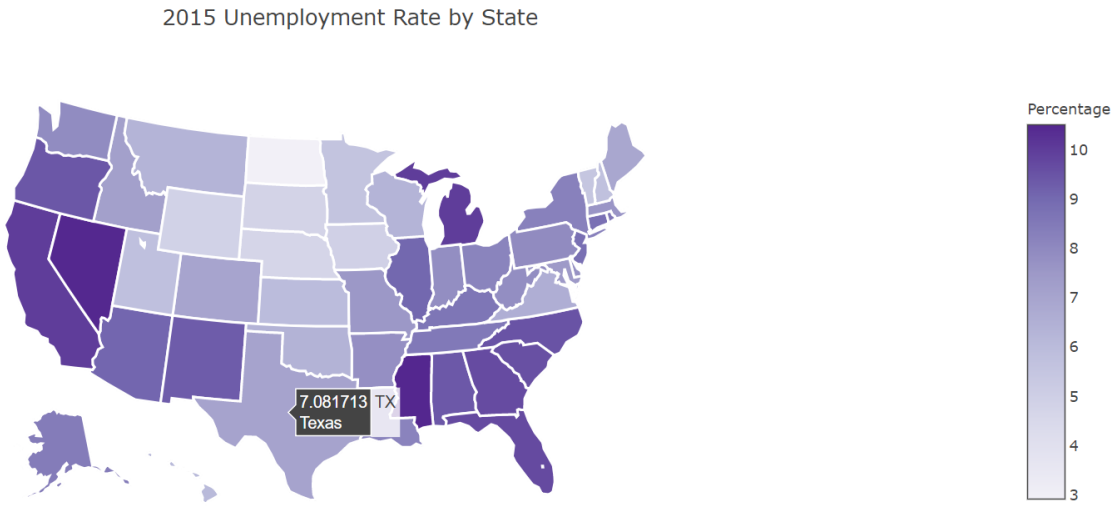


Figure 10



Sources

1.http://money.cnn.com/2016/12/22/news/economy/us-inequality-worse/

2. https://www.washingtonpost.com/news/wonk/wp/2015/05/21/the-top-10-of-americans-own-76-of-the-stuff-and-its-dragging-our-economy-down/?utm_term=.cf24bf757522

3. https://www.huffingtonpost.com/steven-strauss/the-connection-between-ed_b_1066401.html

4. https://dqydj.com/income-change-career-income-increase-age/

5. https://www.bls.gov/emp/chart-unemployment-earnings-education.htm

6. https://www.census.gov/population/estimates/nation/intfile3-1.txt

7.https://www.census.gov/prod/1/pop/profile/95/p23-189.pdf