Data Report: U.S Unemployment Analysis

Naihe Xiao

2023-04-27

Contents

Introduction

Background

The unemployment rate is considered one of the most important economic indicators. It measures the share of workers in the labor force who do not currently have a job but are actively looking for work. However, the issue of unemployment is a significant concern that has garnered considerable attention in the United States, particularly in the wake of the pandemic.

Despite efforts to promote equal opportunities for all, disparities still exist between various demographic groups. According to BLS reports(https://www.bls.gov/opub/reports/race-and-ethnicity/2021/home.htm), the unemployment rate averaged 8.6 percent for Blacks, 8.2 percent for American Indians and Alaska Natives, 6.9 percent for Native Hawaiians and Other Pacific Islanders, 5.0 percent for Asians, and 4.7 percent for Whites.

Research Question

This study aims to perform a comprehensive investigation of employment data and its relationship with pertinent factors in the United States. Analyzing the underlying reasons for employment imbalances can provide insights into possible remedies for achieving a more equitable and fair society. The research question under scrutiny is whether a correlation exists between the unemployment rate and factors such as income or commute time.

By conducting this analysis, we hope to shed light on the complex interplay of factors that influence employment in the United States. Through this understanding, we can work towards creating more effective policies and initiatives that promote greater economic and social equity for all Americans.

Note about interactive plots

We created 3 interactive plots during our analysis, which can not be displayed in this report. Please refer to our website for more information here.

Methods

Data source

Our data come from the DP03 and DP05 tables of the American Community Survey 5-year estimates, which provide detailed information on social, economic, and housing characteristics of the United States population at various geographic levels. Our objective is to explore the relationship between unemployment and variables such as gender, race, and income, as well as identify other underlying patterns that may impact the employment rate.

Data cleaning and wrangling

We utilized the Kaggle API to retrieve data in zip format and subsequently transformed it into R tables. Our dataset consists of census tract level observations, where a census tract is a designated

geographic region for conducting a census. Typically, a county contains multiple tracts, and they serve as the smallest territorial unit for population data collection and dissemination in many countries. In the United States, census tracts are further divided into block groups and census blocks. After merging data from 2015 and 2017, our dataset encompasses 147,774 tracts, each with 38 variables.

Select relevant variables and deal with missing values

The variables that are pertinent to our analysis have been identified, and their definitions along with the summary of missing values are presented in the table below:

Table 1: Selected columns statistics

	Column Name	Definition	Data Type	Missing Values
TractId	'TractId'	Census Tract ID	Numeric	0
State	'State'	State name	String	0
County	'County'	County name	String	0
TotalPop	'TotalPop'	Total population	String	0
Men	'Men'	Number of men	String	0
Women	'Women'	Number of women	String	0
Hispanic	'Hispanic'	Percentage of Hispanic population	Numeric	**1372**
White	'White'	Percentage of White population	Numeric	**1372**
Black	'Black'	Percentage of Black population	Numeric	**1372**
Native	'Native'	Percentage of Native American population	Numeric	**1372**
Asian	'Asian'	Percentage of Asian population	Numeric	**1372**
Pacific	'Pacific'	Percentage of Pacific Islander population	Numeric	**1372**
Income	'Income'	Median household income	Numeric	**2199**
Unemployment	'Unemployment'	Unemployment rate	Numeric	**1600**
year	'year'	Year	Numeric	0
MeanCommute	'MeanCommute'	Mean Commute Time in Minutes	Numeric	**1880**

According to the summary table provided, we discovered that the dataset had a total of 6102 missing values. Specifically, the race columns had 1372 missing values each, while the income and unemployment columns had 2199 and 1600 missing values, respectively. To ensure the integrity and reliability of the analysis, we decided to eliminate all missing values in the dataset, given the significance of the aforementioned columns. Given the size of the dataset, we deemed this approach appropriate and not likely to significantly impact the analysis results. After removing the missing values, the dataset consisted of 145571 rows and 15 columns, allowing for a thorough and rigorous analysis.

Data issues

To ensure the accuracy and validity of the unemployment rate column, we conducted a thorough examination of the dataset by analyzing the observations with the highest and lowest unemployment rates. Our analysis revealed that there were two instances with unusually high unemployment rates (91.9 and 100) and 638 instances with abnormally low unemployment rates (0). To address this

issue and prevent any potential inaccuracies, we opted to remove these extreme values. As a result, the dataset was reduced to 145154 tracts, which allowed for a more reliable and accurate analysis.

Models establishment

We implement two types of models: the regression models and the classification models. For regression models, we choose to implement a generalized linear model, a mixed effect model with random intercept, and two mixed effect models with random slope. The specific descriptions of the four models are provided below:

- The first model is a generalized linear model containing every variable in the dataset, since we have several variables which are count data.
- The second model is a random intercept model to examine the relationship between unemployment and the predictors Income and MeanCommute with a random intercept for each State. This model accounts for the potential correlation between observations within the same State, allowing for more accurate estimation of the fixed effects of Income and MeanCommute on Unemployment.
- The third and fourth model are also mixed-effects model with Income and MeanCommute as predictors, respectively, but with a random slope which allows the relationship between the variables to vary across States, capturing heterogeneity in the effect of Income and MeanCommute on Unemployment across different regions.

For classification, we want to predict whether a region is highly unemployed or operating just fine. To evaluate this, we create a new binary variable "Jobless" which is 1 if the unemployment rate in that tract is higher than the mean unemployment rate, and 0 otherwise. We create a decision tree and a random forest to conduct the classification, and examine if there is a difference between their abilities.

Tools used

We used library httr to use kaggle API to obtain the data, library kableExtra to generate tables, libraries ggplot2, plotly, ggpubr, viridis to create figures, libraries data.table, tidyverse, dplyr, and tidyr for basic data wrangling, and libraries lmerTest, rpart, randomForest, gbm, and lme4 for establishing models.

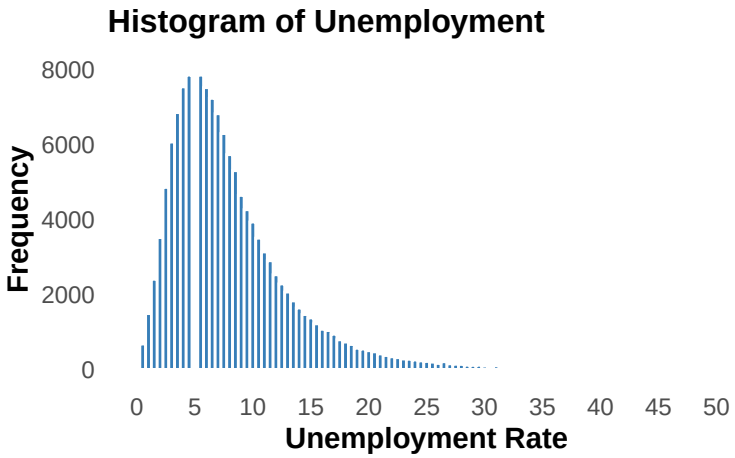
Results

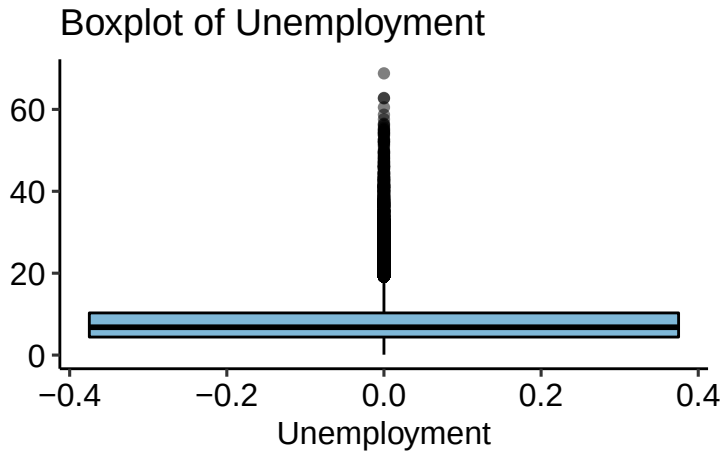
Exploratory plots and tables

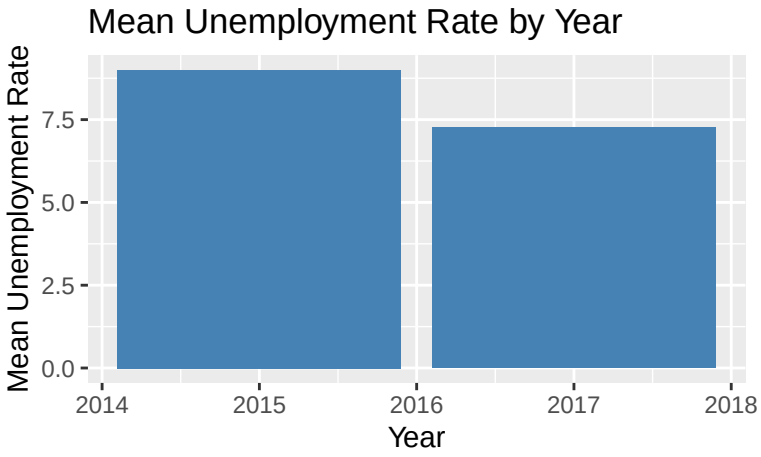
To gain initial insights into the relationship between unemployment and our variables of interest, we will generate some visualizations.

Unemployment Visualizations

To gain initial insights into the distribution of our target variable, unemployment, we will generate a histogram, box plot, bar plot, and map visualizations. The map will display the unemployment rate across different states in the United States.







- The histogram indicates that the distribution of the unemployment rate column is highly skewed and has a truncated right tail.
- According to the boxplot, the median of the unemployment rate is around 8 percent, with most values falling between 0 to 10 percent. However, the dataset contains a significant number of outliers, which constitute 4.6 percent of the entire table. Despite the potential value of outliers in revealing important patterns, we opted to retain all observations in the analysis.
- The barplot reveals a decline in the unemployment rate between 2015 (8.75%) and 2017 (6.4%), suggesting a positive trend in the U.S. labor market.
- The map visualizes the distribution of the unemployment rate across states, which appears to be relatively uniform, ranging from 0 to 20 percent. Notably, states in the middle region exhibit lower unemployment rates than those in the southern and southeastern regions. Additionally, South Dakota stands out for having a significantly higher unemployment rate compared to other states.

Income and Unemployment

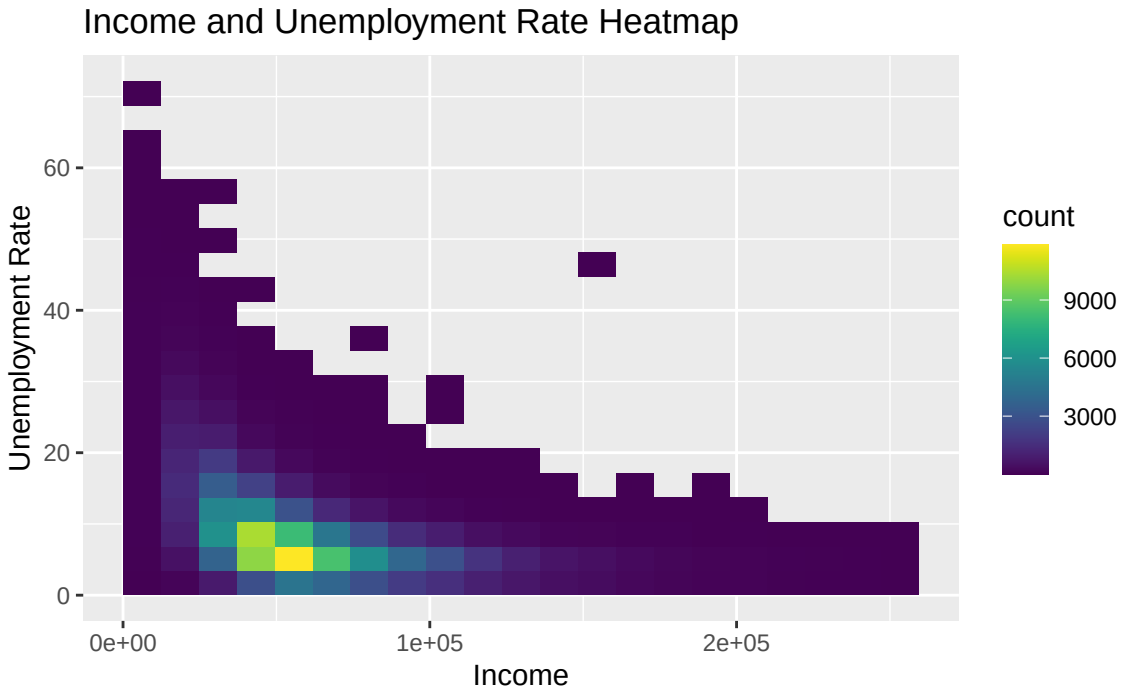
Intuition tells us that regions with higher average income tend to have lower unemployment rate. Let's take a look at a scatterplot between them to check if our intuition aligns with the reality:

The plot shows the relationship between the income and the log unemployment rate(since the histogram of unemployment shows heavy skewness) for the top 100 most populated states in the data frame, where each point contains information about the state, mean commute time, number of men and number of women in the tract, scaled by the total population of the state. This plot illustrate some interesting features:

- There is one tract with particularly low unemployment rate, which is in ada county in Idaho.
- As the income increases, the unemployment rate seems to be a slight tendency of decreasing, especially when income is in the range from 40k to 100k.
- There does not seem to be a relationship between the population of the tract and the unemployment rate according to the size of the balls: there are small balls and large balls in both high and low unemployment rate areas.
- California seems to contain most counties with the highest unemployment rates, while Texas contains most counties with relatively low unemployment rates.

- States including Texas, Florida, and California contain the highest number of the most populated tracts.

Now we create a simple heatmap to showcase the relationship between income and unemployment rate directly.

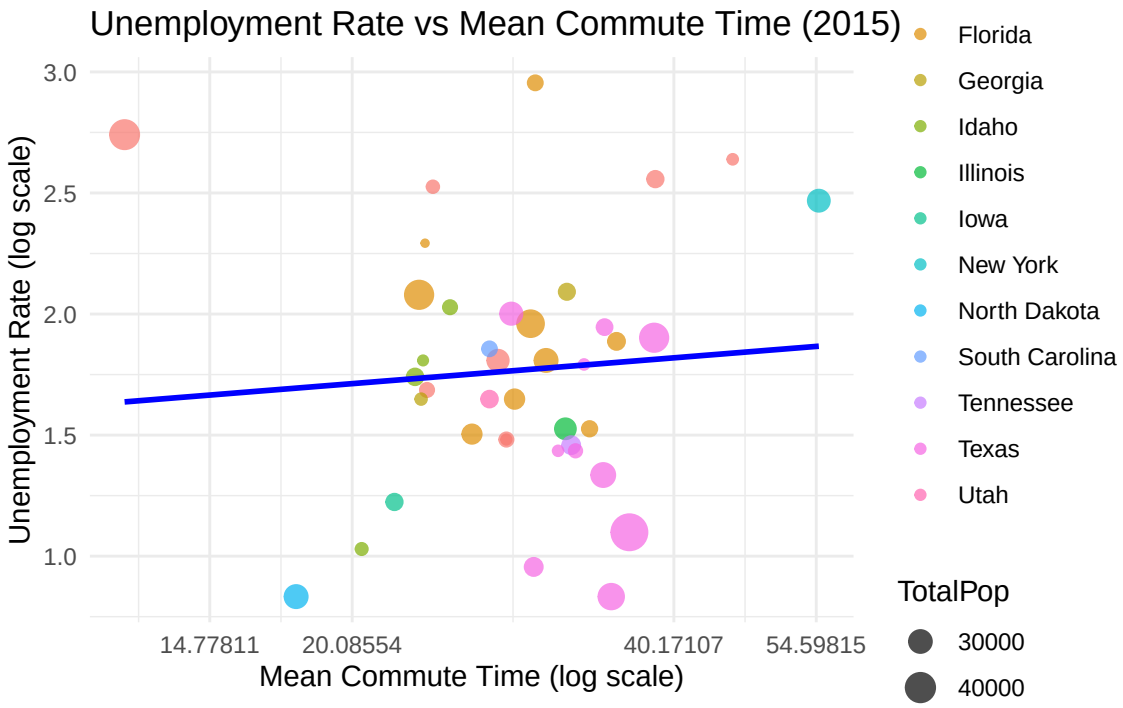


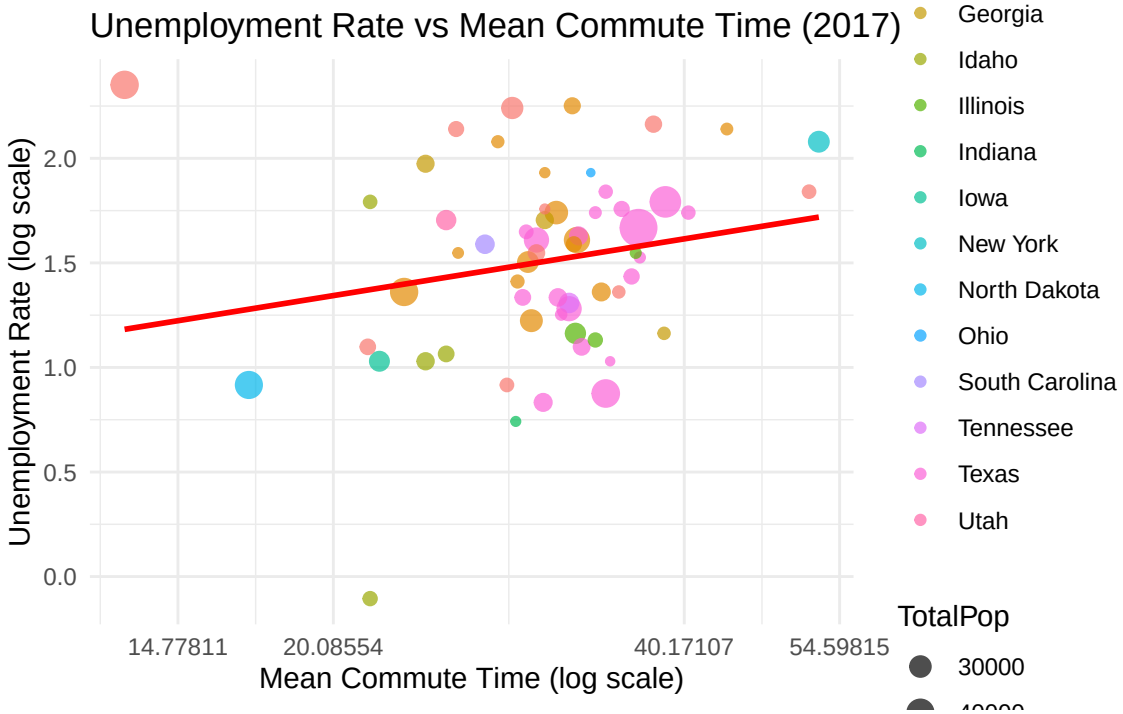
The heatmap visualization indicates a noticeable pattern of a negative correlation between the unemployment rate and income level. Moreover, the majority of the population lies within the income range of 0 to 100,000 dollars and an unemployment rate range of 0 to 20 percent.

To summarize this section, we discovered that there is indeed a negative relationship between annual salary and unemployment rate in the U.S, which we will consolidate later with a regression model.

Commute Time and Unemployment

We will use an interactive scatterplot to showcase the relationship between the 2 variables in both years, and two simple scatterplots to inspect if there a difference between these two variable's correlation in different years.





Again, the plots focus on the top 100 most populated states. The first plot provides us with the following information:

- Most state's average commute time is in the range of 20 to 40 minutes.
- According to the fitted smoothed line, there are no monotone relationship between the mean commute time and unemployment rate. In fact, as the mean commute time increases from

0 to 18 minutes, the unemployment decreases; it fluctuates when the mean commute time increases from 20 minutes to 35 minutes, and from then on increases as the mean commute time increases.

Different from the first visualization, the following two scatterplots show a positive correlation between the mean commute time and the unemployment rate in both years, but there are 2 notable differences:

- the fitted line for 2015 is flatter than that of 2017, indicating a stronger association between the variables in 2017
- the examples are more scattered in 2015, and relatively concentrated around the fitted line in 2017.

We can not make any conclusions yet from the visualizations we created in this section. Hopefully we can obtain a more solid result in the modeling section.

Correlation Test

We now conduct a correlation test between the two pairs of variables: (unemployment rate, income) and (unemployment rate, commute time) to acquire some quantitative results.

Table 2: Correlation test results

Variable1	Variable2	Method	Correlation	P.value
Income	Unemployment	Pearson	-0.4691188	0
Income	Unemployment	Spearman	-0.5466294	0
Mean Commute	Unemployment	Pearson	0.0905117	0
Mean Commute	Unemployment	Spearman	0.0657963	0

The results of the correlation test are promising. Since we are using the pearson correlation test which assumes the variables to be normally distributed, we use a log transformation on the variable unemployment. The table reveals that, no matter what the method is, there is a median negative correlation between income and unemployment rate, and a slightly negative correlation between mean commute time and unemployment rate. Notice that all p-values are 0, indicating that the results are statistically significant.

Modeling

The modeling outputs are summarized below.

Regression model

The outputs of the four models explained in the methods section are presented in the tables below

Table 3: Summarization of Model Coefficients for Variable Income

Model	RMSE	Income_Coefficient	Income_pvalue
generalized linear model	4.293045	-5.97e-05	0.0000000
linear mixed effects model(random intercept)	4.525520	-9.58e-05	0.0000000
linear mixed effects model(random slope)	5.184477	-1.13e-04	0.4413404

Table 4: Summarization of Model Coefficients for Variable MeanCommute

Model	RMSE	Commute_Coefficient	Commute_pvalue
generalized linear model	4.293045	0.0748860	0.0000000
linear mixed effects model(random intercept)	4.525520	0.0992608	0.0000000
linear mixed effects model(random slope)	5.184477	0.0064013	0.7646992

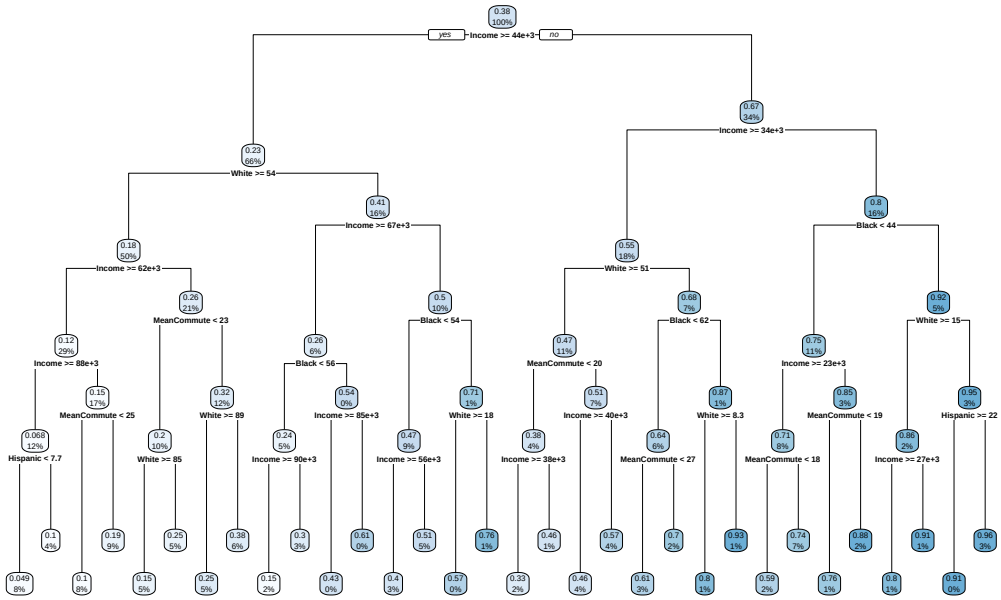
The table for income indicates that there is indeed a slightly negative correlation between income and unemployment rate; specifically, the three models suggest that as the median annual salary in a region increases by 1 dollar, the average unemployment rate in that region decreases by 0.0000597, 0.0000958, 0.000113 percents, respectively.

On the other hand, the commute table implies a positive correlation between the mean commute time and the unemployment rate; specifically, the three models suggest that as the mean commute time in a region increases by 1 minute, the average unemployment rate in that region increases by 0.0749, 0.0993, and 0.0064 percents, respectively. Meanwhile, all models yield small root-mean-square errors and small p value(expect for the random slope models), hence the results can be considered significant.

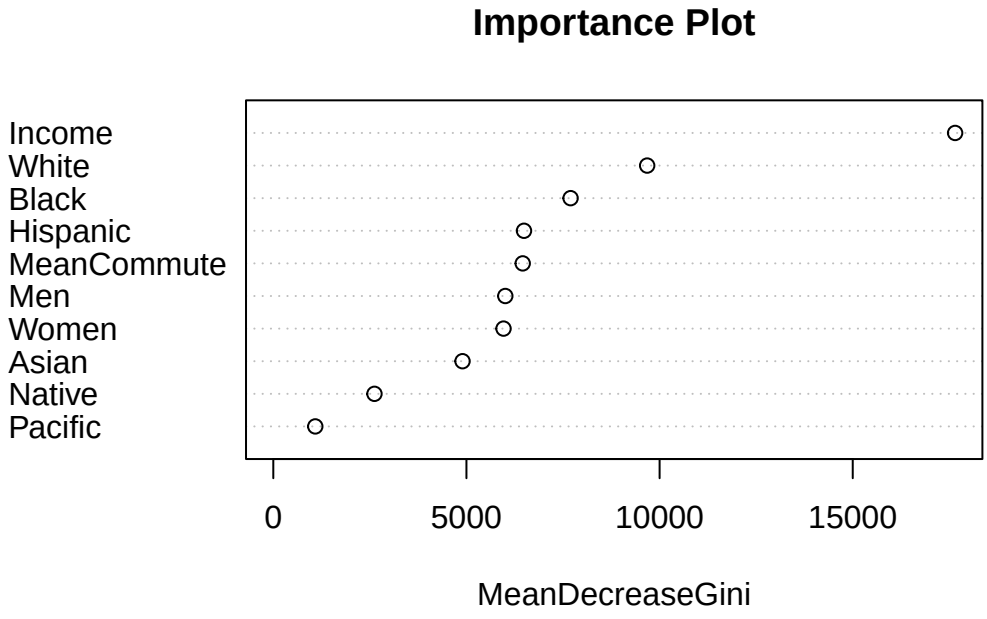
Classification model

The outputs of the two classification models are illustrated below:

Decision Tree



Random Forest



- The decision tree chooses to make the first split based on whether the annual salary is higher than 44000 or not; it then splits on variables white and income again. The depth of the

decision tree is 6, with 32 terminal nodes; the percent of examples falling into each terminal node is approximately even, without any node containing more than 10 percent of the entire population.

- The importance plot above shows that income is a very strong predictor of unemployment, while commute time is relatively weaker. An interesting point is that the number of white and black people in the region are also strong indicators of unemployment rate, while the importance of both genders are not outstanding.
- The mean squared error for the decision tree and the random forest are 0.1674157 and 1.0000207, respectively, so both model are yielding excellent performances. Notice that the decision tree may have a smaller MSE due to overfitting, since we are computing the MSE using our original data as new data.

Conclusions and Summary

Conclusion

The present study leverages the American Community Survey 5-year estimates dataset to investigate the correlation between unemployment rate and various factors, such as income and commute time. Through the utilization of interactive visualizations and tables, our analysis reveals a potential negative correlation between income and unemployment rate, as well as a positive correlation between commute time and unemployment rate. These findings are further validated by the construction of four models, all of which exhibit promising results.

In addition, we employed a decision tree and a random forest to classify regions that may be experiencing unemployment issues, both of which achieved state-of-the-art performances. Overall, this study contributes to the understanding of the relationship between unemployment rate and various socio-economic factors in the U.S. The results may be utilized to inform policy decisions aimed at reducing unemployment rates and promoting economic growth.

Limitations and Improvements

This study has some limitations that need to be acknowledged.

- Firstly, all the data used in this study were collected prior to the outbreak of the COVID-19 pandemic, which has had a profound impact on the economic landscape and labor market in the United States. Therefore, the current unemployment situation may be significantly different from what is presented in our study, and our results and models may not be directly applicable to the current societal context.
- Secondly, the visualizations created in this study only utilized data from the 100 most populated states, which could potentially affect the accuracy and generalizability of our findings. Although this approach improved the interpretability and aesthetics of our graphs, it may not provide the most representative reflection of the variable relationships.

These limitations should be taken into account when interpreting the results and conclusions of this study, and further research is warranted to investigate the impact of the COVID-19 pandemic on the relationships between unemployment, income, and commute time in the United States.