Adult Income Final Project

Executive Summary:

My original goal was to explore a random dataset and see if there were any significant relationships between the columns of that dataset. Originally, I had picked a NBA dataset to see if there were any significant relationships between the increasing number of points throughout NBA history and other factors such as Field Goal Attempts, 3 Pointers, etc. However, due to the dataset having a lot of missing values, the dataset was not fit for testing, and thus, I picked the Adult Income dataset due to how comprehensive the dataset was. Through EDA, I came to the conclusion that Adult Income has a variety of factors that could be potentially be influenced from the other columns in the data set. Eventually, after testing multiple relationships through both traditional classification methods such as bar graphs, and histograms, and linear regression models, I came to the conclusion that the income of a person was severely influenced by a person's wealth, education level, gender, marital status, etc. Furthermore, I trained and tested the dataset to see which method of learning was the best in terms of predicative modeling and I came to the conclusion that for this particular dataset, CART was the best method.

Summary of Learning:

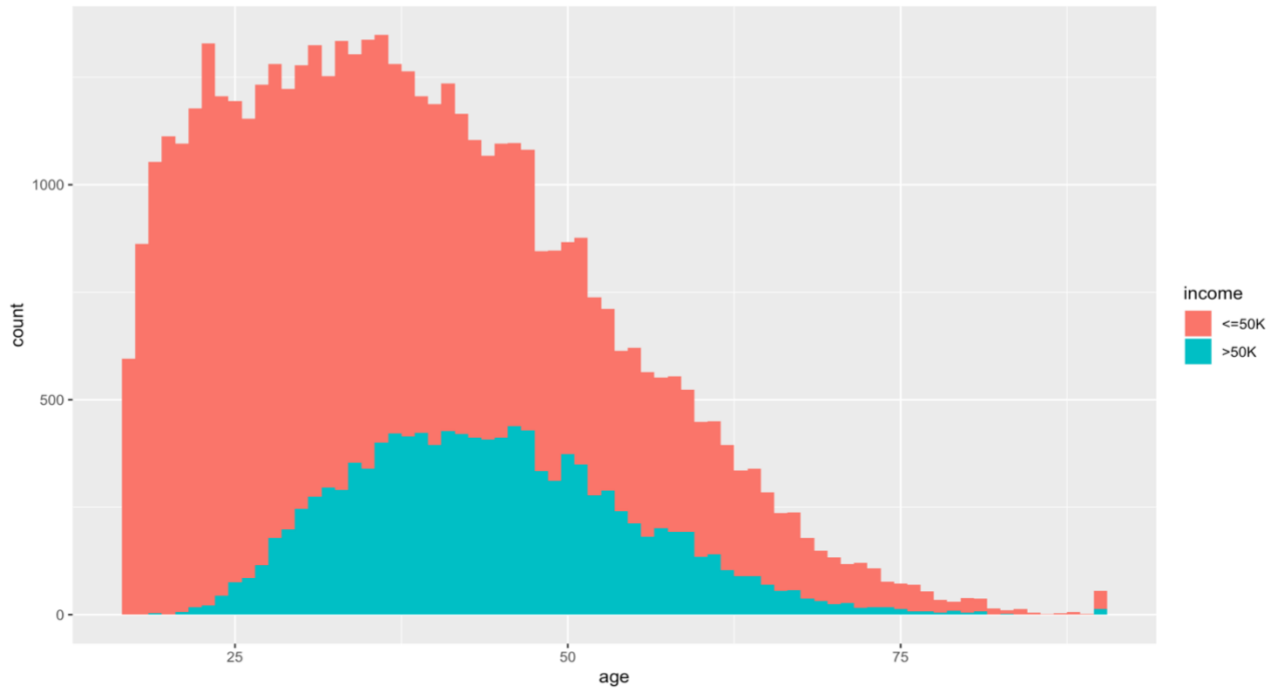
Throughout the course of this project, I have learned quite a bit, especially with regards to predicative modeling. I gained greater experience and understanding in Support Vector Machines, CART, linear regression, and the library, ggplot. I also learned how to use the function gsub to clean the data so that I could classify it easier and also gained a solid understanding of creating histograms and bar graphs with multiple variables.

DataSet:

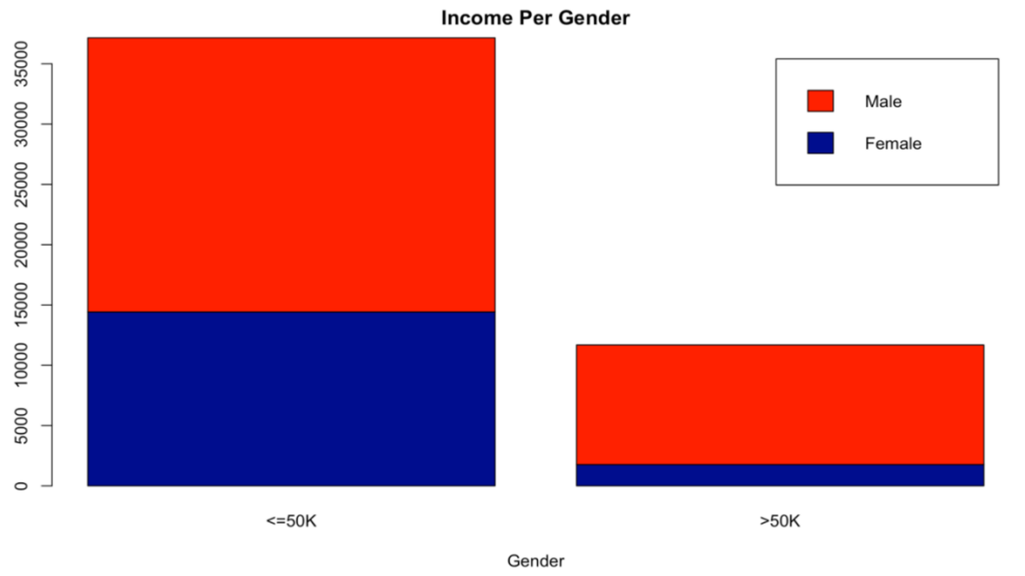
I used the AdultIncome.csv dataset from Kaggle and downloaded the dataset and read the csv file through RStudio. The dataset describes the income of each person and is divided into two classes(<50K or >=50K). The dataset also includes other demographic information such as an individual's education level, age, gender, occupation, etc.

Data Source: https://www.kaggle.com/wenruliu/adult-income-dataset

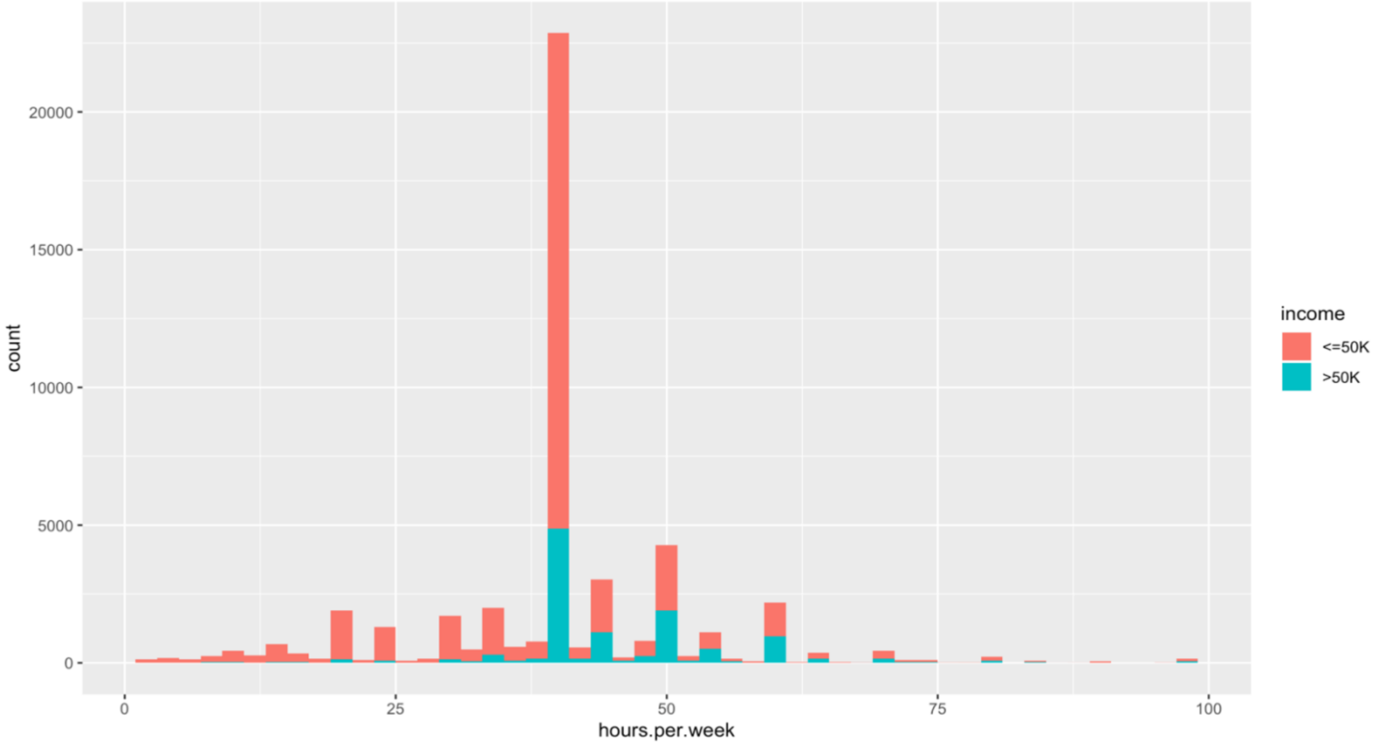
Exploratory Data Analysis



Generally, there were a lot more people who earned less than 50K than people who earned 50K. As people got older, there were more people who earned 50K, but after they turned 50, there were less people who earned 50K.

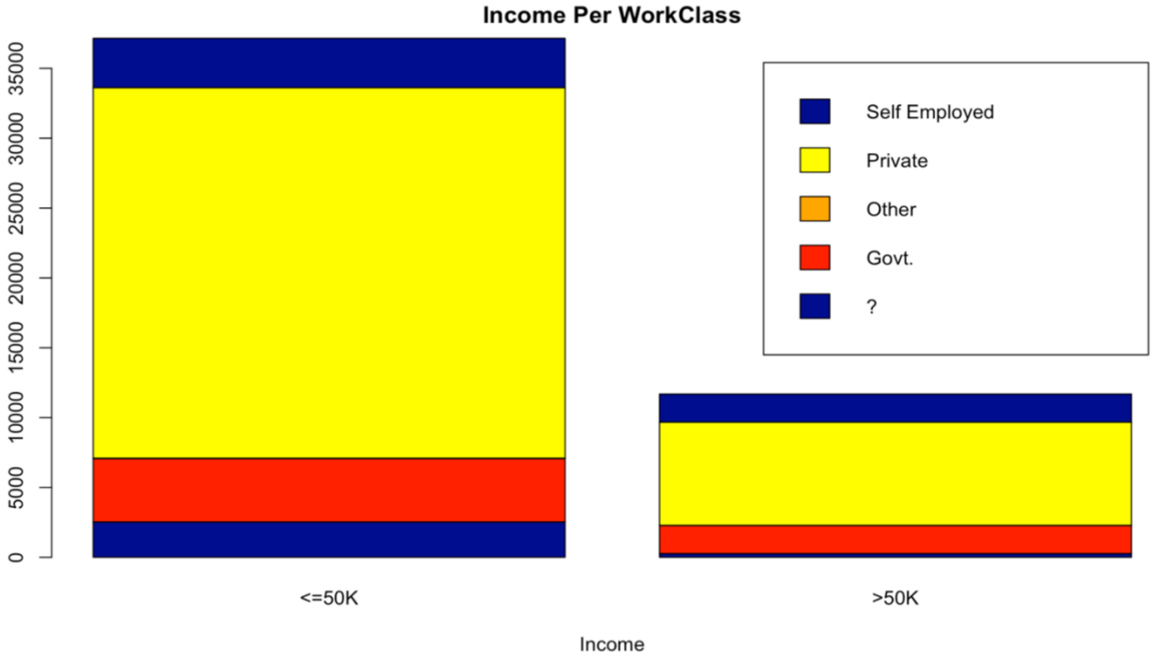


Men earn more than women for both levels of income. There is a much bigger disparity though in men earning more than 50K over women.

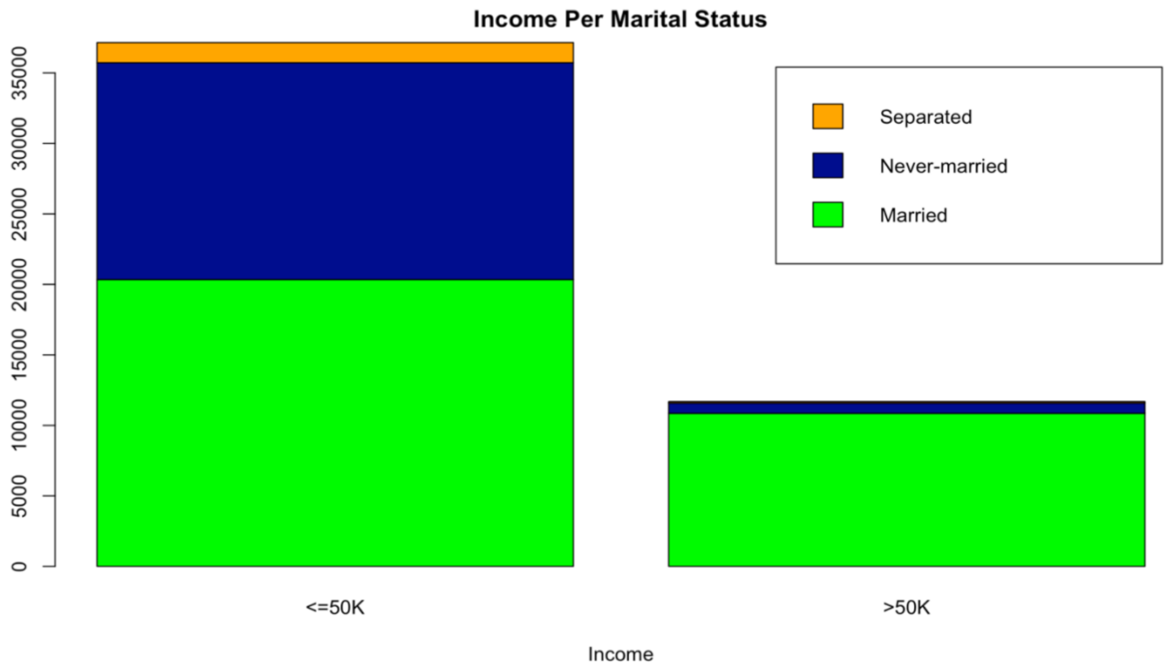


Histogram shows the distribution between the hours per week and income level. No real relationship derived from the graph in terms of income, but most people in the dataset worked around 36 hours per week.

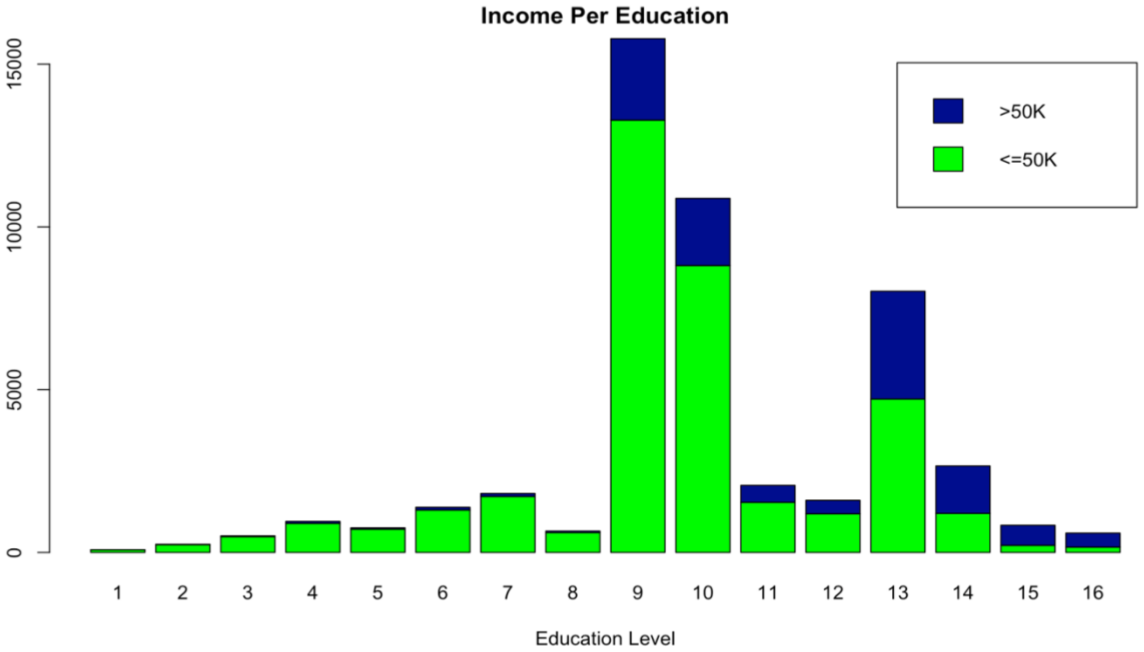
In the graphs below, I used the gsub function to help clean the data into fewer rows so that I could compare trends over more broad categories instead of multiple categories which probably would have led to the visualizations becoming unclear.



This bar graph shows that the most people work under the private sector. Not many people work under Other, if any at all, and people who work in the Govt. are more likely to earn over 50K compared to people who work in the private sector.



The bar graph above shows the Income Per Marital Status. People who are married are much likely to earn over 50K than people who are separated or have never married. This is due to the fact that most people who have never married are generally the ones who are young, and don’t typically earn that much.



The histogram above shows that people who have not been given a higher form of education typically do not make that much money. The people who have earned 50K are the ones who have completed a higher level of education and are more likely to get a higher earning job.

Models:

Before, I built my regression model, I set aside 80% of my Adult dataset to train and 20% of it to test my models. Then, I made sure to take to clean my null values with the complete.cases() function. After this, I built a regression model to see if there was a significant relationship between the Income and other Categorical fields. I used a binomial regression due to the fact that Income only had two factors(>= 50K or <50K). Through this, I learned that income had a significant relationship with occupation, age, education, work class, etc. This meant that the income of a person was directly affected by a person's job status, marital status, education level, etc. Next, I built a Polynomial Support Vector Machine to help predict the income level.

Through this, I found out that the training error is low, which means that it could be a result of my data overfitting due to a large training set with multiple parameters. Through a confusion Matrix, my model performance is quite high as its accuracy is .8275 and since my Kappa statistic(compares the accuracy of the system to the accuracy of a random system) is .4642, the accuracy of my predicative model favors comparably to a random system. My Positive Class is <50K due to the fact that there are more people who earn less than 50K over people who earn more than 50K. I also built a linear Support Vector Machine(which is another form of supervised learning). Again, I found out that the training error is low, which means that it could be a result of my data overfitting due to a large training set with multiple parameters. However, when I ran a confusion matrix, its accuracy was .83(which is derived from a sum of the false negatives and true positives divided by the total) and my Kappa statistic was .4715. This means that in this particular instance, the linear SVM model is the better predicative model compared to my polynomial SVM model. The last model I built was the CART model. It is obtained by partitioning the data recursively and fitting a prediction model within each partition. After this happens, a classification tree is grown (which is useful here since I am classifying income amongst a slew of variables). Through this, I found out that my accuracy value is actually .8436, and my kappa statistic is .5184. Out of the three models, this is the most accurate one yet, and thus, is the best model to use out of the three I have created.