# Biostats 625 Final Project - US 2019 Census Data

*https://github.com/Gongting811/625FinalProject*

Group 4 members: Ting Gong, Lap Sum Chan, Margaret Prentice

December 20, 2020

## 1. Introduction

The problem of income inequality has been of great concern in the recent years. Large differences in annual incomes may be the result of a combination of factors such as education level, age, gender, occupation, race, etc. This project aims to conduct a comprehensive analysis to highlight the key factors that are necessary in improving an individual's income. Such analysis would help to set focus on the important areas which can significantly improve the income levels of individuals and thus help to provide a guidance for the individual who wants to make some changes to improve their income level. After identifying the important predictors for income, several binary classifiers are trained to predict whether an individual's annual income in 2019 falls in the income category of either greater than or equal to 60,000 USD or less than 60,000 USD using the dataset extracted from the 2019 Census Bureau database.

## 2. Data Pre-processing

The demographic and income data in this report are from the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC) conducted by the US Census Bureau in 2019. The US Census Bureau collects data and publishes estimates on income and poverty each year to evaluate national economic trends as well as to understand their impact on the wellbeing of households, families, and individuals.

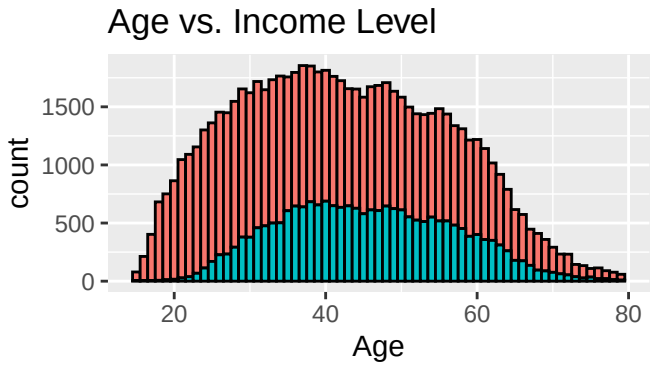
***Dataset Extraction***

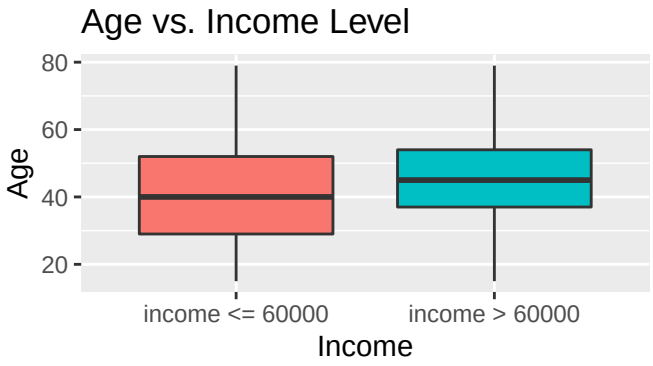
Our dataset were extracted from the original individual-level ASEC dataset, which contains 180101 individuals and 799 features in total. The biggest challenge was to encode all the categorical variables. Most of the variables are categorical ones, but they were all encoded numerically in the original dataset. Therefore, the first thing we did was to find them all by reading the documentations and encode them in the proper categorical form by hand. Next, we applied several conditions including `income` $> 0$, `age` $> 15$ and `age` $< 80$, `labor force status` == "working", `working hour` $> 0$ to the dataset. Then, to reduce the dimension and select related features from the raw data, we did the Backward Elimination with `income` variable as the response variable and all the remaining ones as covariates. As a result, a set of features were selected that contains demographic variables such as `age`, `race`, `sex`, `academic degree`, `marital status`, `ethnic`, and `region`; employment variables such as `labor force status`, `working hours`, `worker classes`, `major industry` and `major occupation`; health-related variables such as `total medical expenditures` and `health status`. Finally, we merged similar columns for several categorical variables, dropped all the outliers and eventually got our census19 dataset.

***Exploratory Data Analysis***

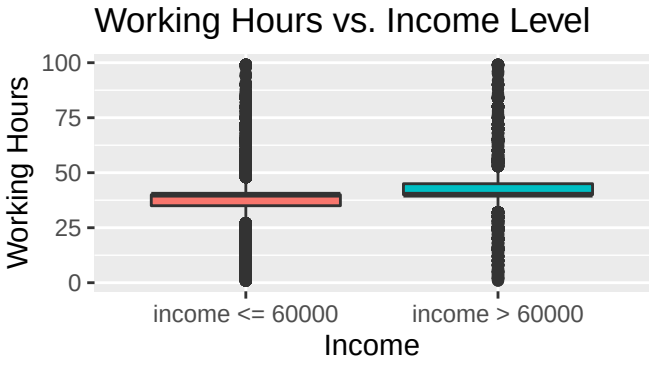
First, we plot a histogram and a boxplot of `age` versus `income level`. We can see from the graphs below that the `age` variable has a wide range and variability, and the percentage of people who make above $60000 peaks out at roughly 35% between ages 35 and 60. The boxplot shows that individuals who have higher

incomes tend to be older than those who don't. This implies that `age` might be a good predictor of `income level`.

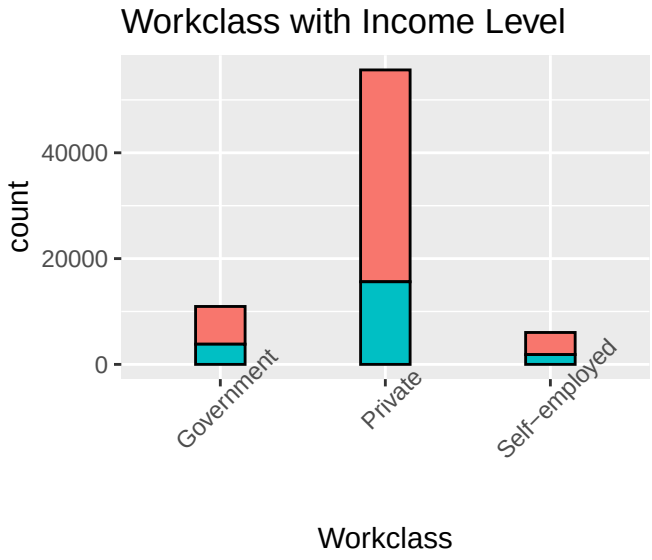


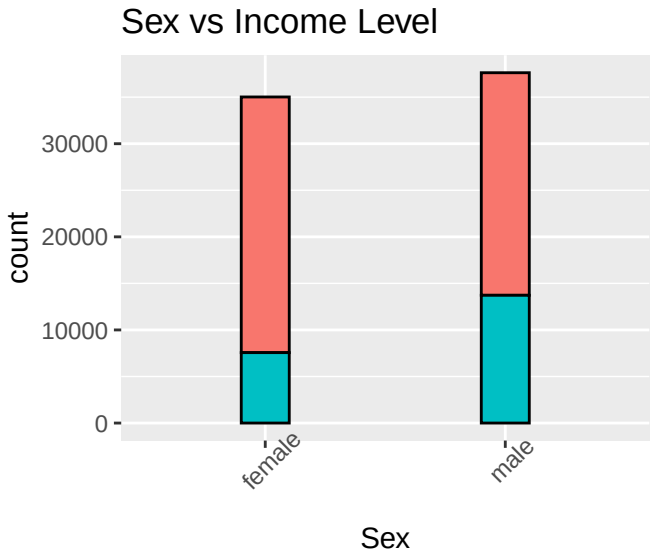



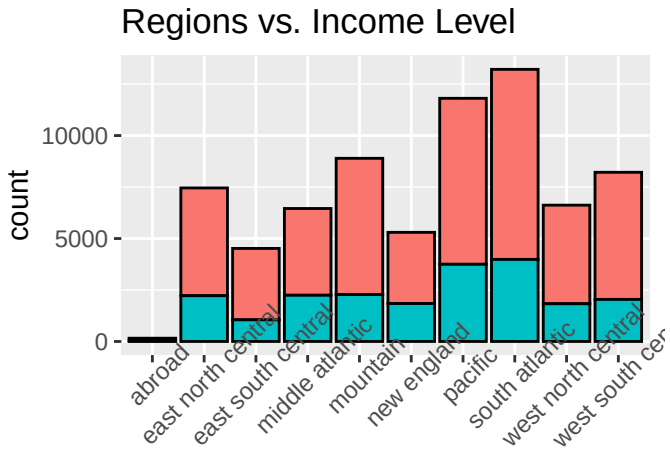

Next, we plot a histogram and a boxplot of `working hours` (per week) versus `income level`. It is shown from the graph that the highest frequency of `working hours` (per week) occurs at around 35-45 hours. The boxplot shows that individuals who have higher incomes tend to work longer than those who don't.

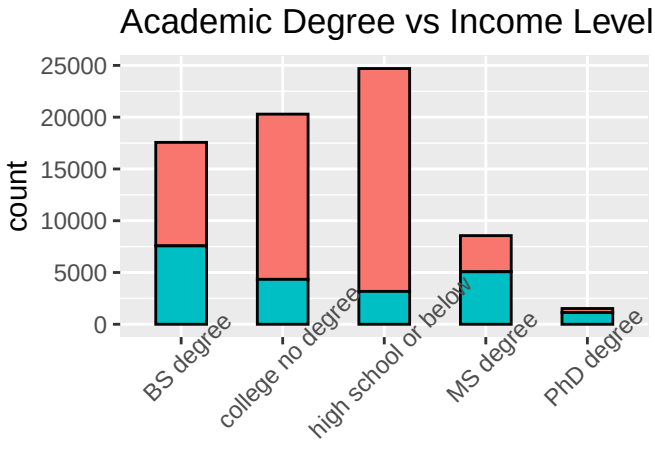




After visualizing the distribution of `income level` versus the above two continuous variables `age` and `working hours`, we then explored some categorical variables. It turned out that the following four variables `sex`, `work class`, `academic degree` and `region` are all likely to be good predictors.






Lastly, we split the dataset into 60%, 20% and 20% for training, validation and testing respectively after the EDA.

## 3. Methodology

For each of the models, we selected the best performing model based on their accuracy performance on the validation set.

**Logistic Regression**

Several models and interaction terms were considered for logistic regression. The final model included all features settled upon during the features extraction, as well as interaction terms between `sex` and `marital status` and between `worker classes` and `major industry`. These interaction terms increased the accuracy of the model on the validation set as well as increased the area under the curve (AUC) of the receiver operating characteristic (ROC) curve as compared to the base logistic regression model with all features and no interaction terms.

**Random Forest**

For the random forest model, it was based on the `randomForest` package and we varied the number of variables randomly sampled as candidates at each split (`mtry`). We tried `mtry` from 2 to 10, and `mtry` = 4 had the best accuracy in the validation set. Notice that the difference in accuracy between the model with highest accuracy (`mtry` = 4) and that of lowest accuracy (`mtry` = 9) was only about 0.007 or 0.7%.

**XGBoost**

We used the `xgboost` package for the XGBoost tree booster model. We tried a wide range of tuning parameters, which included the learning rate `eta` = 0.025, 0.05, 0.1 and 0.3, the maximum depth of a tree `max_depth` from 2 to 6, and varied the number of boosting iterations from 300 to 1000. The model with learning rate 0.025, maximum depth of a tree = 6 and number of boosting itearions = 300 was chosen as it had the lowest validation error (18.63%) amongst all the parameter grids we searched. In other words, the XGBoost model achieved an accuracy of 81.37% in the validation set.

## 4. Result

***Feature Importance*** Based on our training data, we found that the `major occupation` feature was most important with about 0.036 mean decrease in accuracy when `major occupation` is permuted after training

and before prediction, while the `health status` feature was least important with about 0.0005 mean decrease in accuracy when it is permuted according to the results of our random forest model. The `academic degree` and `working hours` were second and third most important, both with about 0.023 mean decrease in accuracy. `age`, `total medical expenditures`, `major industry`, and `sex` each had mean decrease in accuracy ranging from about 0.016 to 0.012, while all other features had mean decrease in accuracy less than 0.01.

***Model Comparison*** Finally, we applied the previously learned models to the test data to get a relatively unbiased evaluation of each model. The prediction accuracy of logistic regression, random forest and XGBoost models on the test data were 80.04%, 80.06% and 80.98%, respectively. Similar result could also be seen in the ROC curve comparing the three methods. We can see that there is a huge overlap between the logistic regression and random forest and XGBoost has only a tiny margin above the two methods. Indeed, the corresponding AUC for logistic regression, random forest and XGBoost are 0.8476, 0.8450 and 0.8593. Notice that random forest had an even worse performance when measured using AUC.

## 5. Conclusion

In this project found that occupation, academic degree and working hours were the most important factors determining whether one's income is above or below 60,000 USD. This suggests that people can if people are not earning 60,000 USD, they can consider getting a higher education, increase their working hours or even change their jobs wherever applicable in order to reach the goal of more than 60,000 USD. In terms of prediction performance, XGBoost only performed slightly better than random forest or logistic regression. It was surprising that random forest performed very similarly to logistic regression in this income prediction problem.