Sumiya Ganbaatar
Ishika Gupta

Project 2

Goal:

We use census data to predict whether or not someone has an annual income of more than $50,000.

Action Plan:

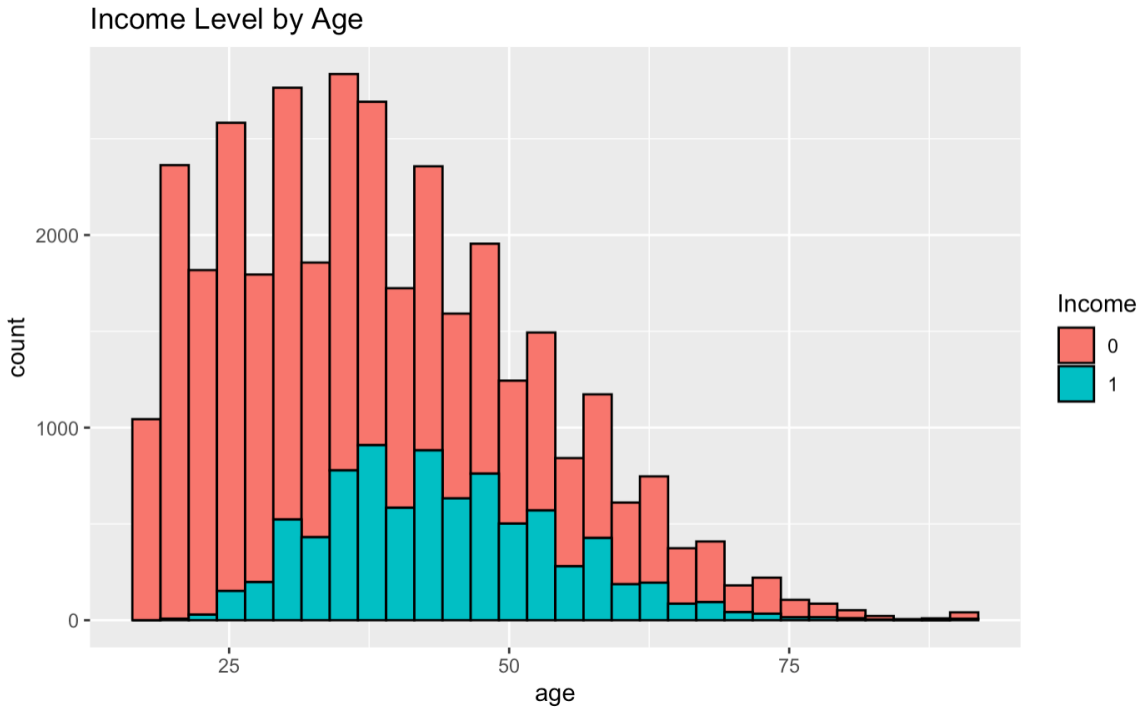
We will first explore the training data to understand the trends and representations of certain demographics in the dataset. We use these insights to select variables that are the most valuable to form models. We compare two models: logistic regression and random forest and use the better model to predict whether an individual would make more or less than $50,000 on the test data.

Exploratory Data Analysis:

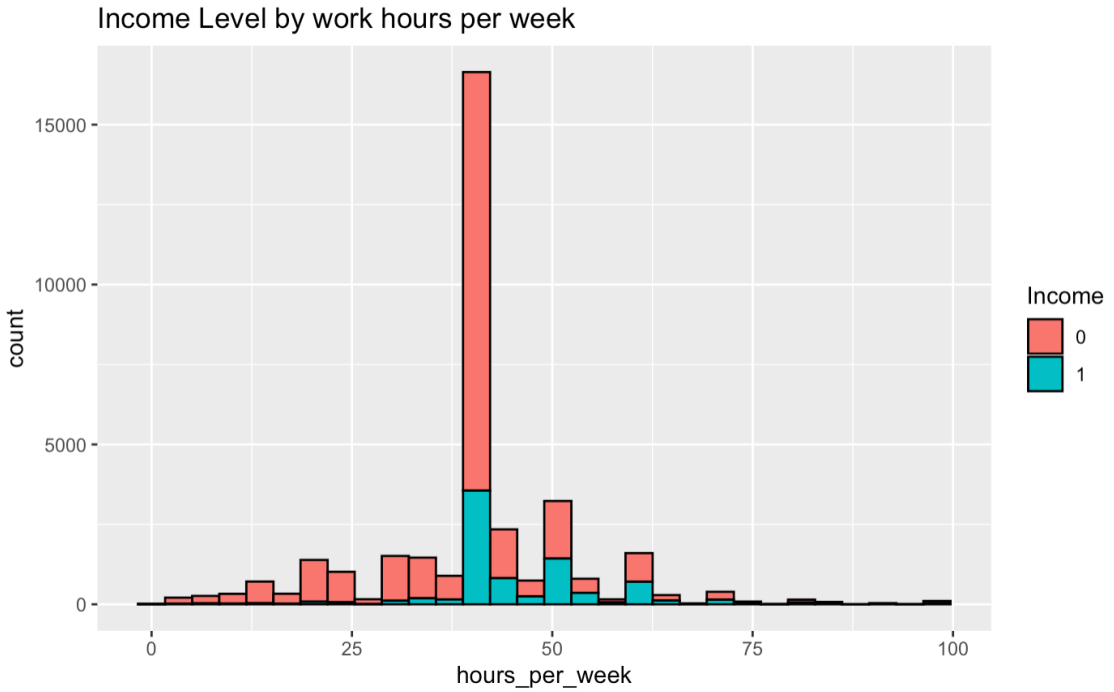
About the Dataset: The training dataset contains 35,000 rows representing unique individuals and 15 columns whereas the test dataset contains 13,840 rows, but only 14 columns since the “income” column have been removed.

- Some values in column 'income' have a period in the end. We decided to turn the ‘income’ variable into a dummy variable of 0’s and 1’s representing incomes less than or equal to $50,000 and more than $50,000 while accounting for values both with and without a period.
- Variable “education” is deleted because there is “education_num” which is a more precise indicator of education level.
- As the variable “relationship” is highly correlated with the variable “sex” and “marital_status”, it’s deleted due to the multicollinearity issue in our modeling.
- We ran the Pearson’s Correlation Coefficient test to find the correlation between "age", "fnlwgt", "education_num", "capital_gain", "capital_loss", "income", "hours_per_week".
- “Fnlwgt” is a weight that represents how common people with these exact age and racial demographics are in the United States is deleted because it has nothing to do with predicting an individual’s income as the correlation was found to be 0 as it can be seen in the figure below.

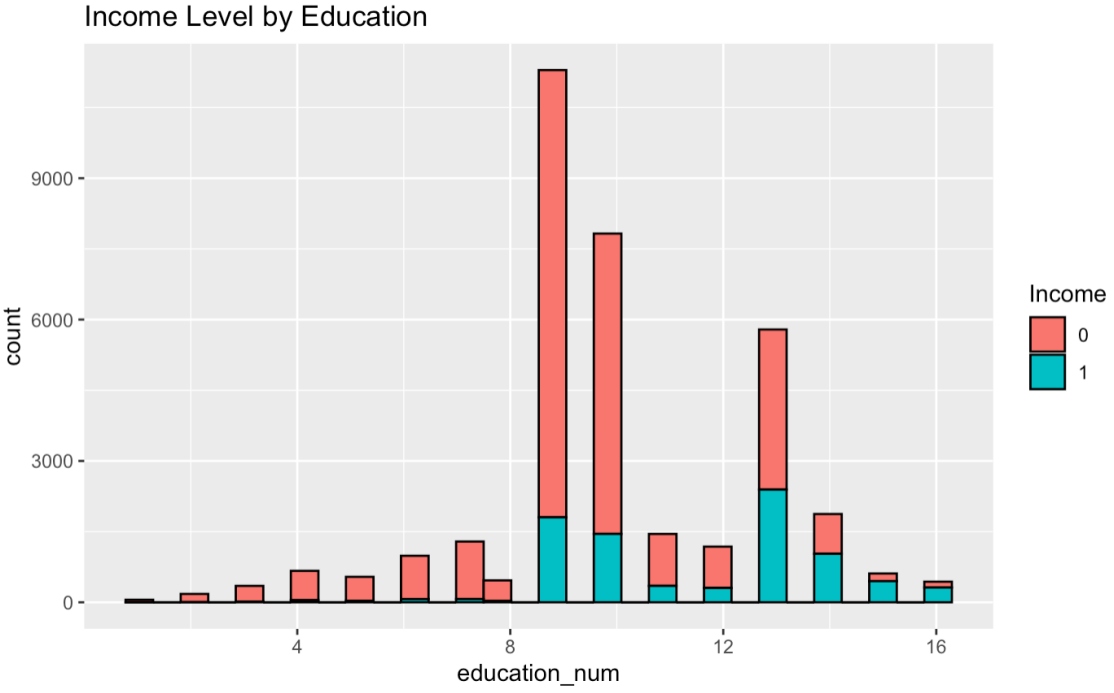
Age: The distribution of age grouped by income shows that the age distribution of those who make above $50,000 is more right-skewed than the age distribution of those who make less than $50,000. It means that middle-aged people tend to have higher incomes than other age groups. Therefore, the variable “age” should be included in the model as there is a correlation.



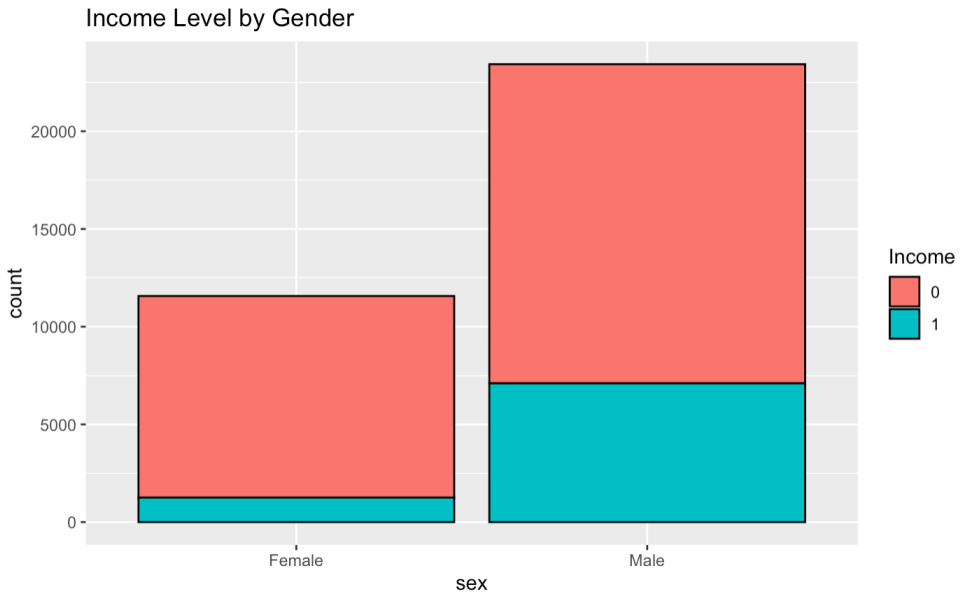
Work hours per week: The histogram of work hours per week grouped by income shows that the chance of having a higher income than $50,000 increases as the number of work hours per week increases. Therefore, the variable “hours_per_week” is also correlated to income and hence should be included in the model.



Years of education: The relationship between years of education and income level shows that Income increases as the years of education increase. Observations with less than 8 years of education barely have an annual income of greater than $50,000.

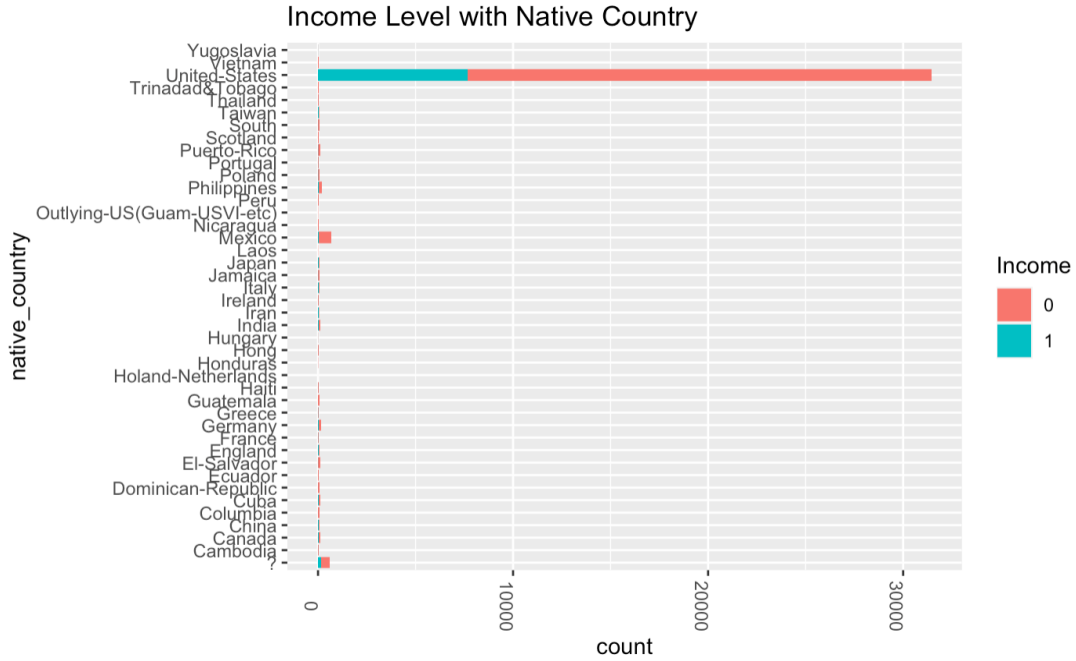


Gender: The relationship between gender and income shows that there is a huge disparity of income between males and females which means it is one of the significant explanatory variables.



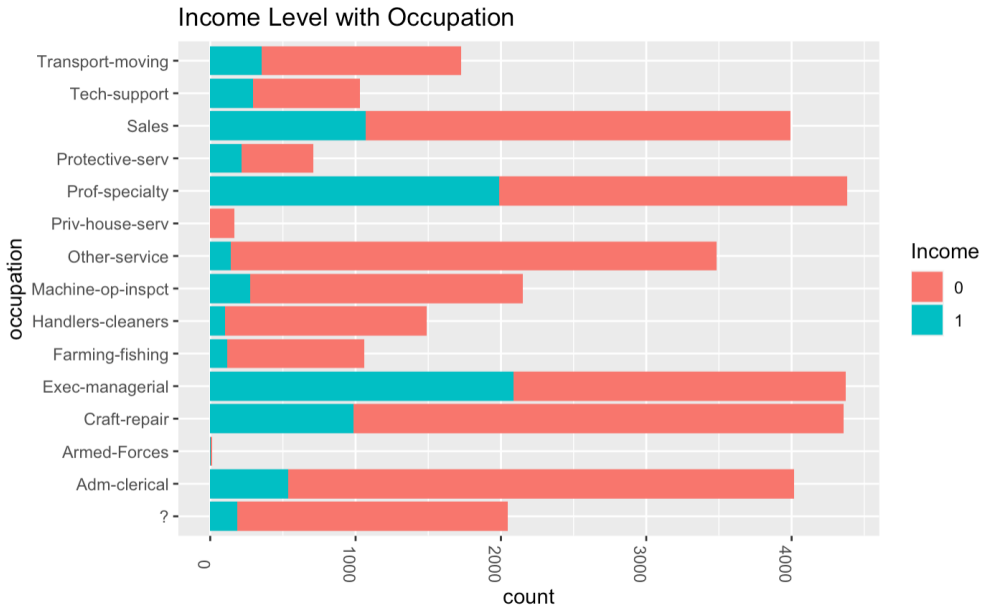
Capital gain and capital loss: The variables 'capital_gain' and 'capital_loss' are two continuous variables that show income and loss from the sale of capital assets. Although most observations have zero capital_gain and/or capital_loss, these variables are somewhat correlated with the variable income.

Native country: The variable 'native_country' displays high skewness as most observations are from the United States as we can see in the figure below. Hence, this variable is not included in our predictive model.

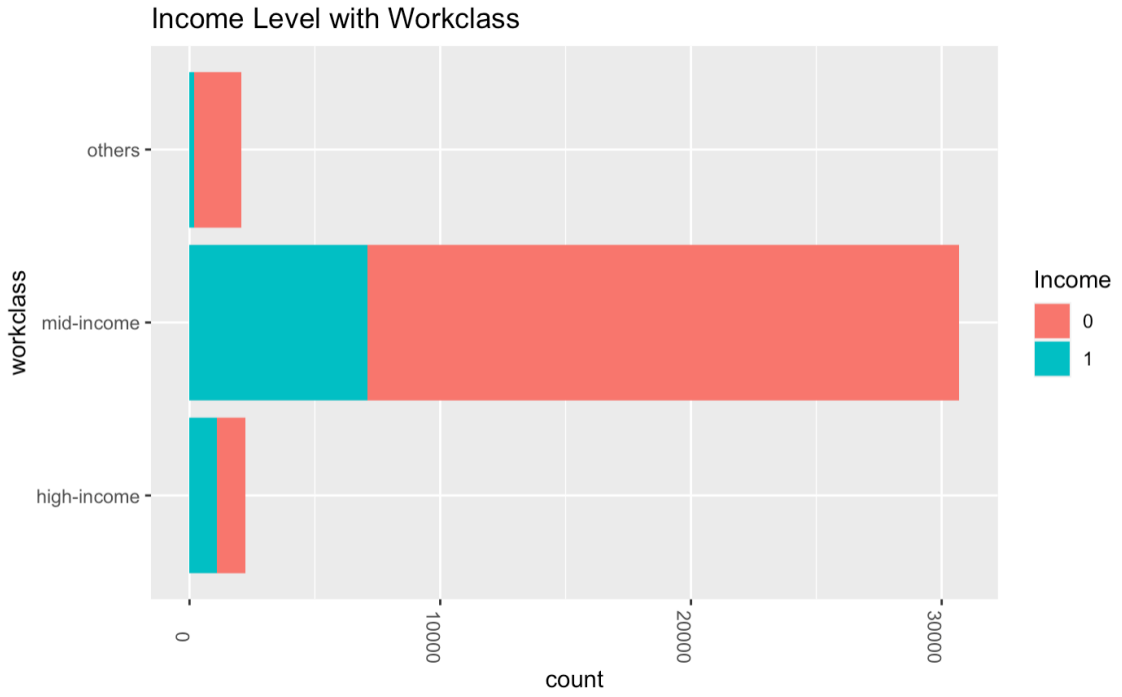


Feature Engineering:

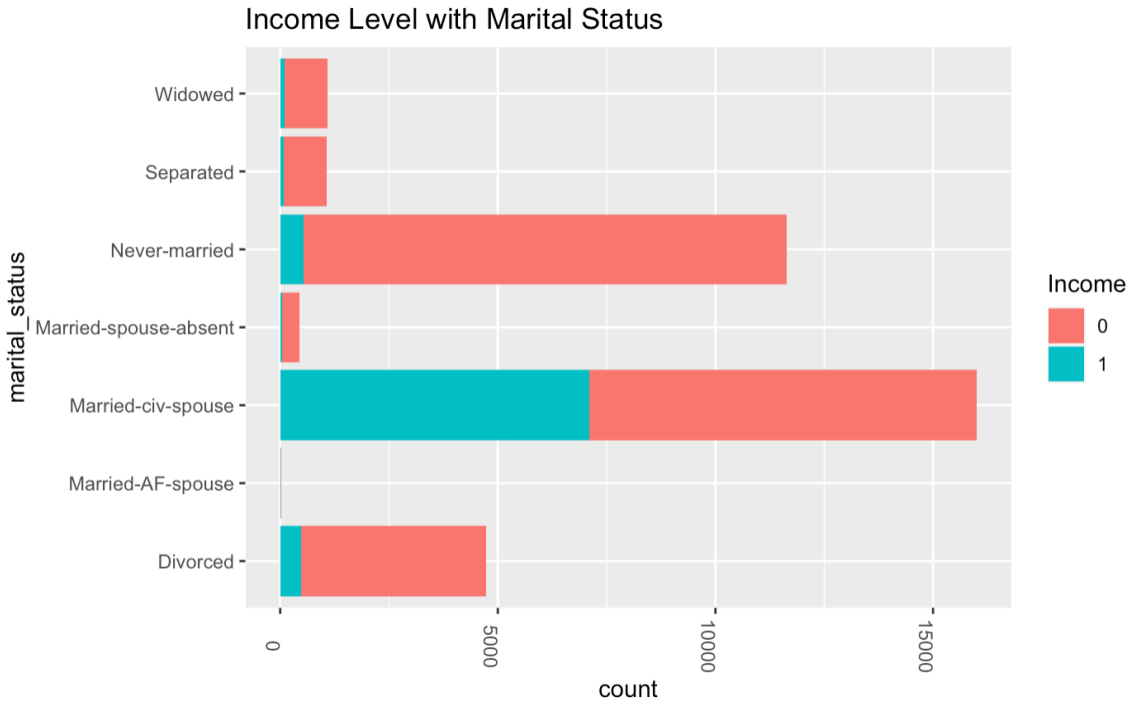
Occupation: Based on the percentages of people with income more than $50k for each occupation, we divided the variable 'occupation' into categories such as high_income (Exec-Managerial, Prof-specialty), middle_income (Tech-support, Transport moving, Protective service, Sales, Adm-clerical, Craft-repair), low_income (Farming-fishing, Handlers-cleaners, Machine-op-inspct, Priv-house-serv) and others (Armed-Forces, Other-service, ?). The relationship between income and occupation before and after simplifying the values is shown in the bar plots below:

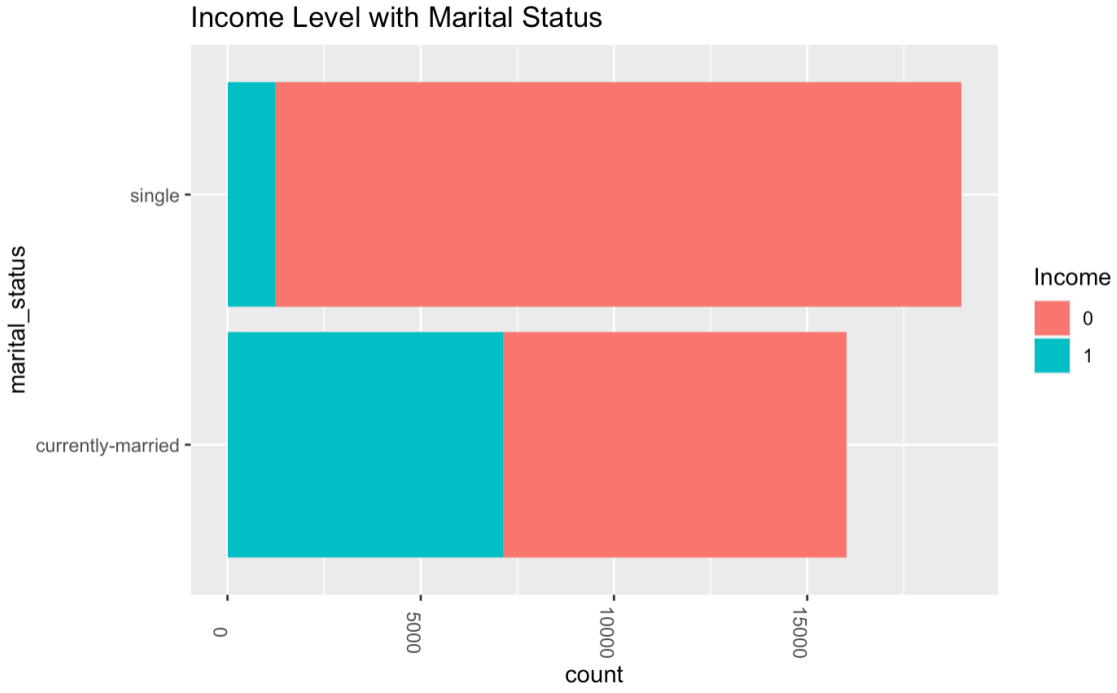


Workclass: Based on the percentages of people with income more than $50k for each work class, we divided the variable 'workclass' into categories such as high_income (Federal-gov, Self-emp-inc), middle_income (Local-gov, Private, Self-emp-not-inc, State-gov, Adm-clerical, Craft-repair), and others (Without-pay, Never-worked, ?). The relationship between income and work class before and after simplifying the values is shown in the bar plots below:

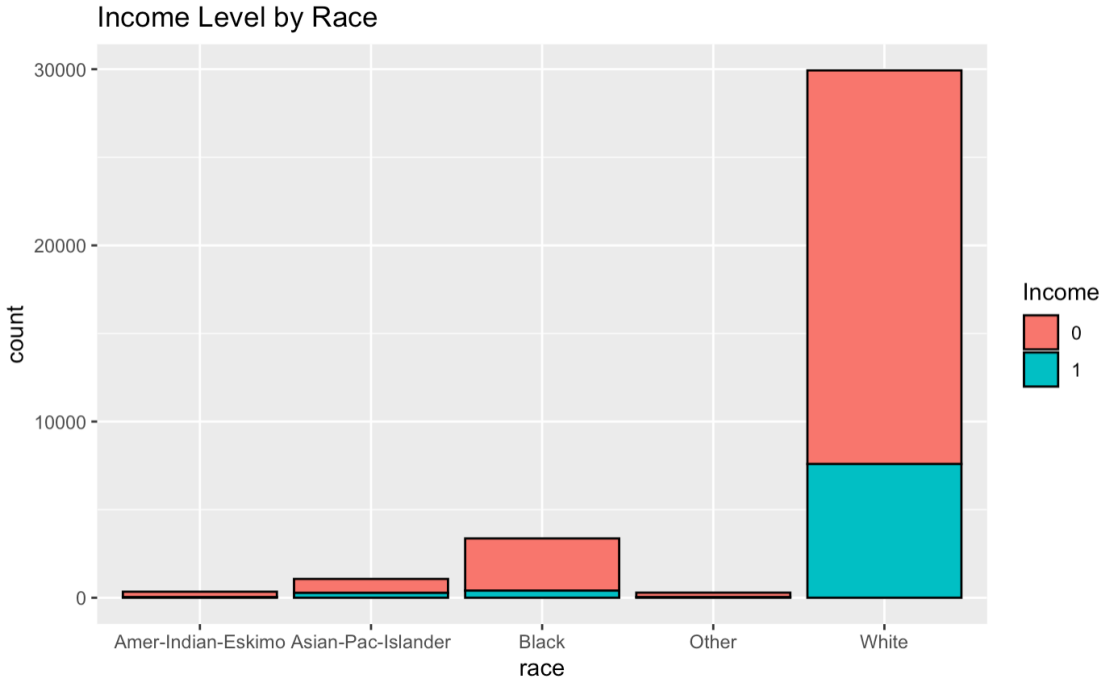


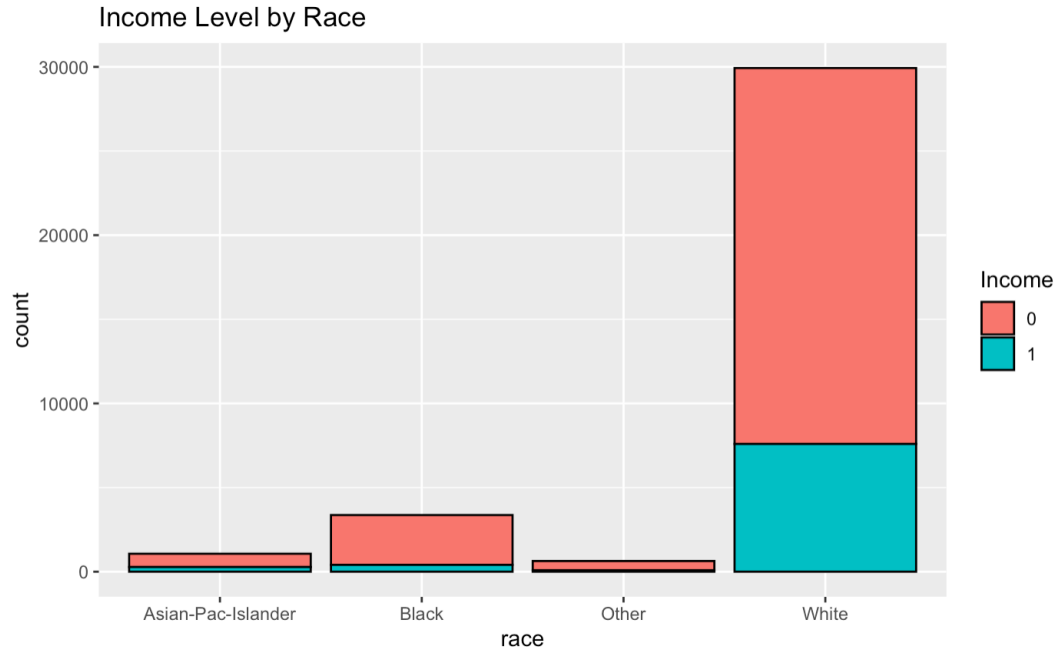
Marital status: The relationship between income and marital status shows that most of the people who are married are making more than $50,000 a year. To facilitate our analysis, we divided the variable ‘marital_status’ into categories such as single (Divorced, Never-married, Separated, Widowed, Married-spouse-absent) and currently-married (Married-civ-spouse, Married-AF-spouse). The relationship between income and marital_status before and after simplifying the values is shown in the bar plots below:





Race: We decided to combine “Amer-Indian-Eskimo” with “Other” because we have a very small number of observations and the percentages of people who make more than $50’000 are almost the same.





Modeling & Result:

We split the training dataset to train(70%) and test(30%) to evaluate our models. We used a logistic regression model and a random forest model to predict the income. Our exploratory data analysis showed that these variables are effective to predict income: age, workclass, education_num, marital_status, occupation, race, sex, capital_gain, capital_loss, and hours_per_week.

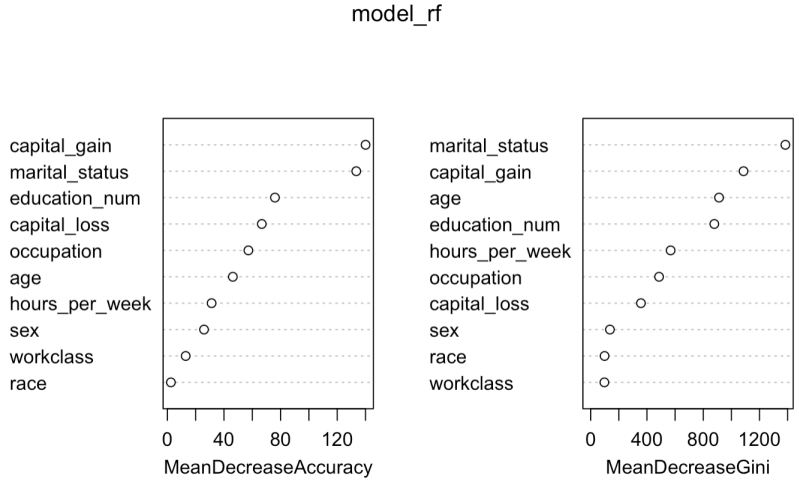
Logistic Regression: The accuracy of the logistic regression model is 85.1% and the confusion matrix is shown below when the cutoff value is optimal.

	Actual income < $50k	Actual income > $50k
Predicted income < $50k	7487	1009
Predicted income > $50k	562	1442

Random Forest: The number of trees was chosen based on the accuracy of the model on the test data, and the highest accuracy is 86.5%. The confusion matrix of the random forest model is shown below.

	Actual income < $50k	Actual income > $50k
Predicted income < $50k	7569	922
Predicted income > $50k	480	1529

The variable importance plot shows that “capital_gain”,”marital_status”, and “education_num” are the most important explanatory variables in our random forest model.



Conclusion:

After visualizing each variable and its effect on income, we applied Machine Learning tools to make predictions about whether an individual's annual income would be more than $50,000. The Random Forest model worked the best out of the two models due to the lower misclassification rate.