US Adult Income

Springboard Capstone Project #2

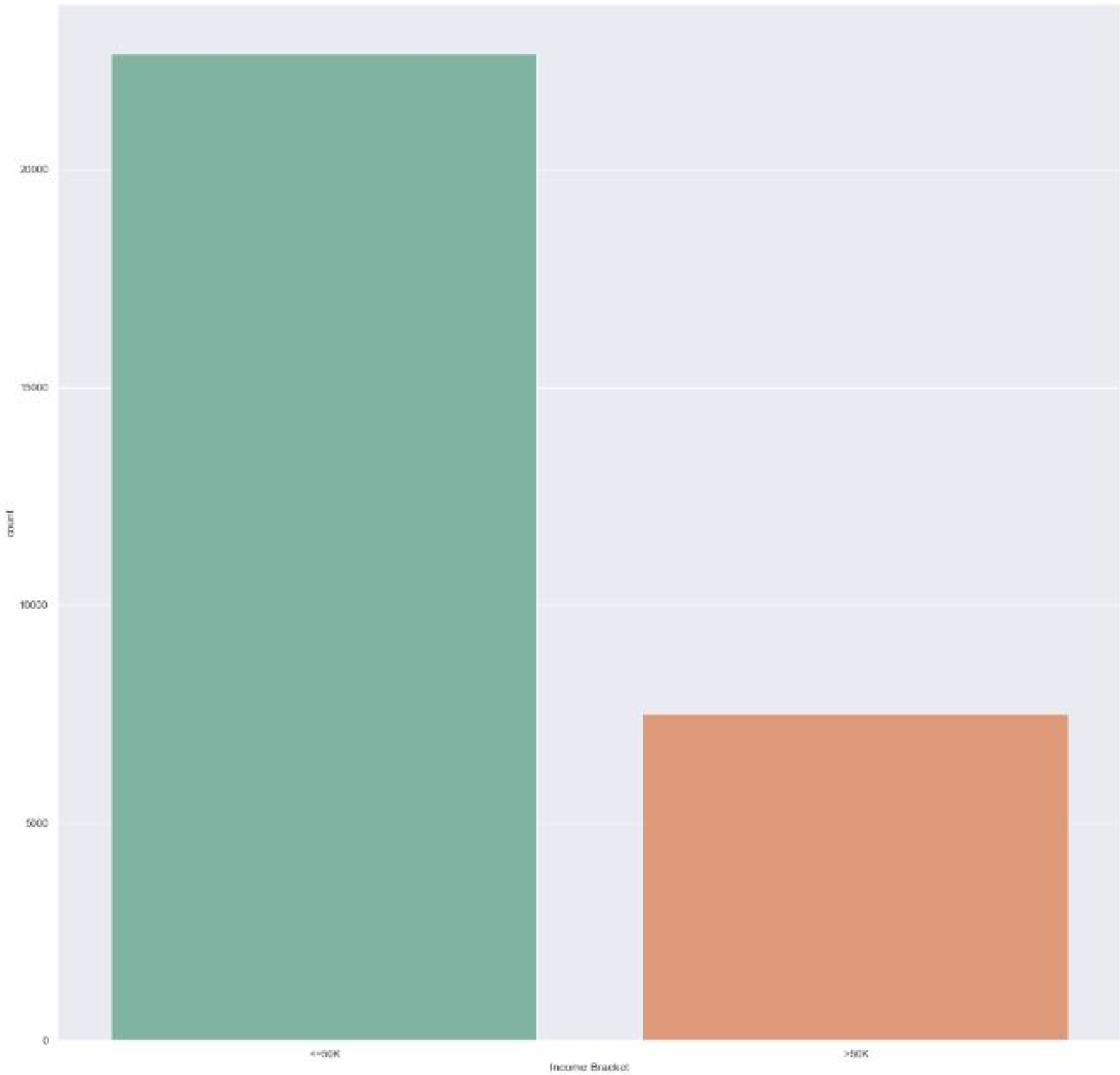
Introduction

The US Adult Census Dataset was retrieved by Barry Becker from the 1994 US Census Database. There are a total of 15 columns in this dataset, 14 of these variables will contribute whether that individual makes an income of ">50K" or "<=50K" in a given year. The objective is to predict the "Income Bracket" which has two different outcomes, ">50K" or "<=50K" and obtain a classifier with great accuracy. The following dataset can be found from kaggle, https://www.kaggle.com/johnolafenwa/us-census-data#adult-training.csv

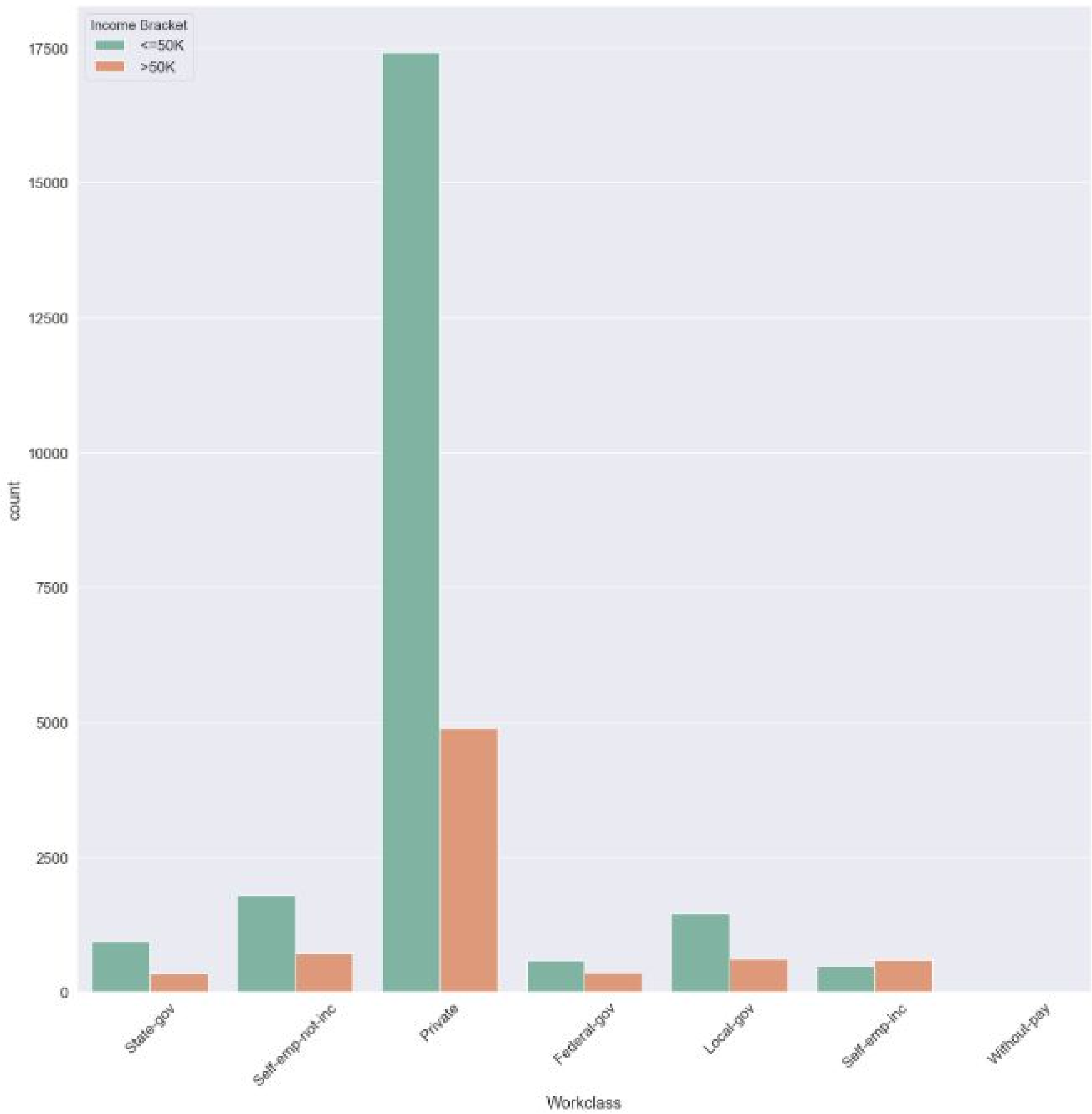
Throughout the project, there will be various tasks performed such as:

- Data Cleaning - To deal with missing data, in the following data there were data that had '?', this was converted to 'nan' and then dropped from the dataset.
- Data analysis - To better understand and conceptualize the data
- Machine Learning Models - Decision Tree Model, Logistic Regression Model, Random Forest Classifier Model and SVC Model will be used to determine the training accuracy.
- Model Tuning - GridSearchCV will be used to get a higher accuracy.

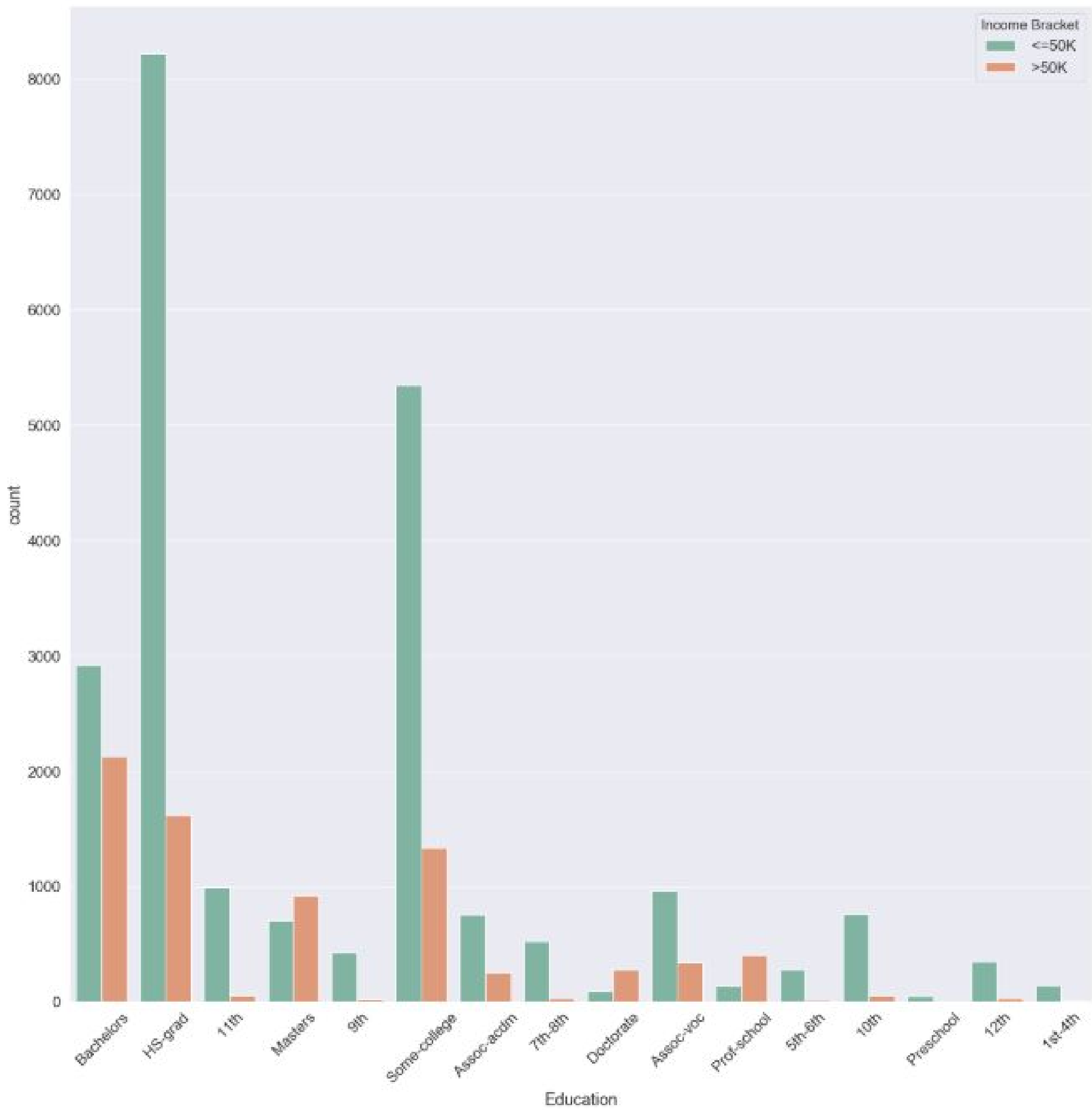
Data Analysis



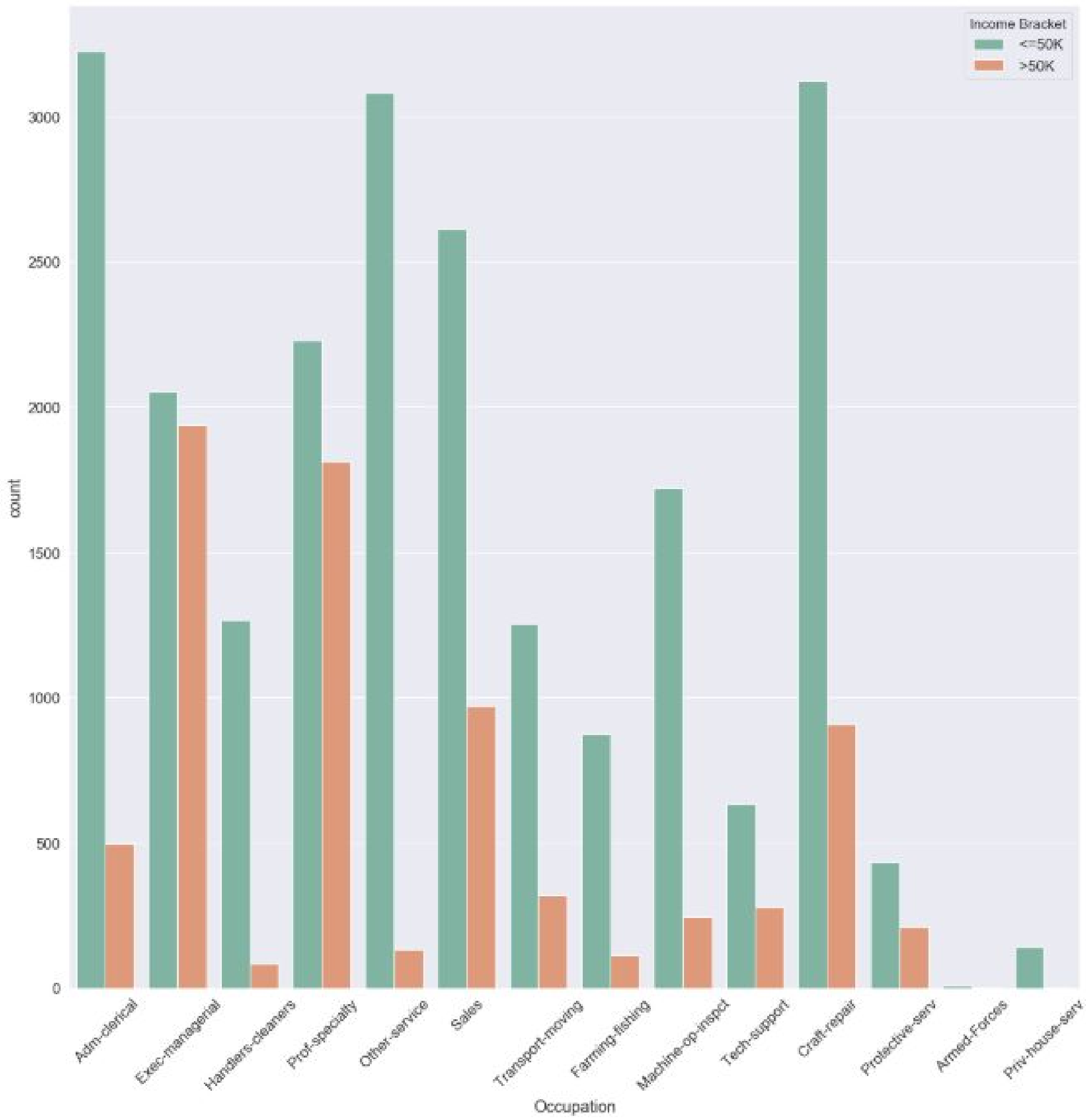
The figure above shows that there are far more individuals making <=50K than in comparison to >50K. In this data set, less than half make >50K.



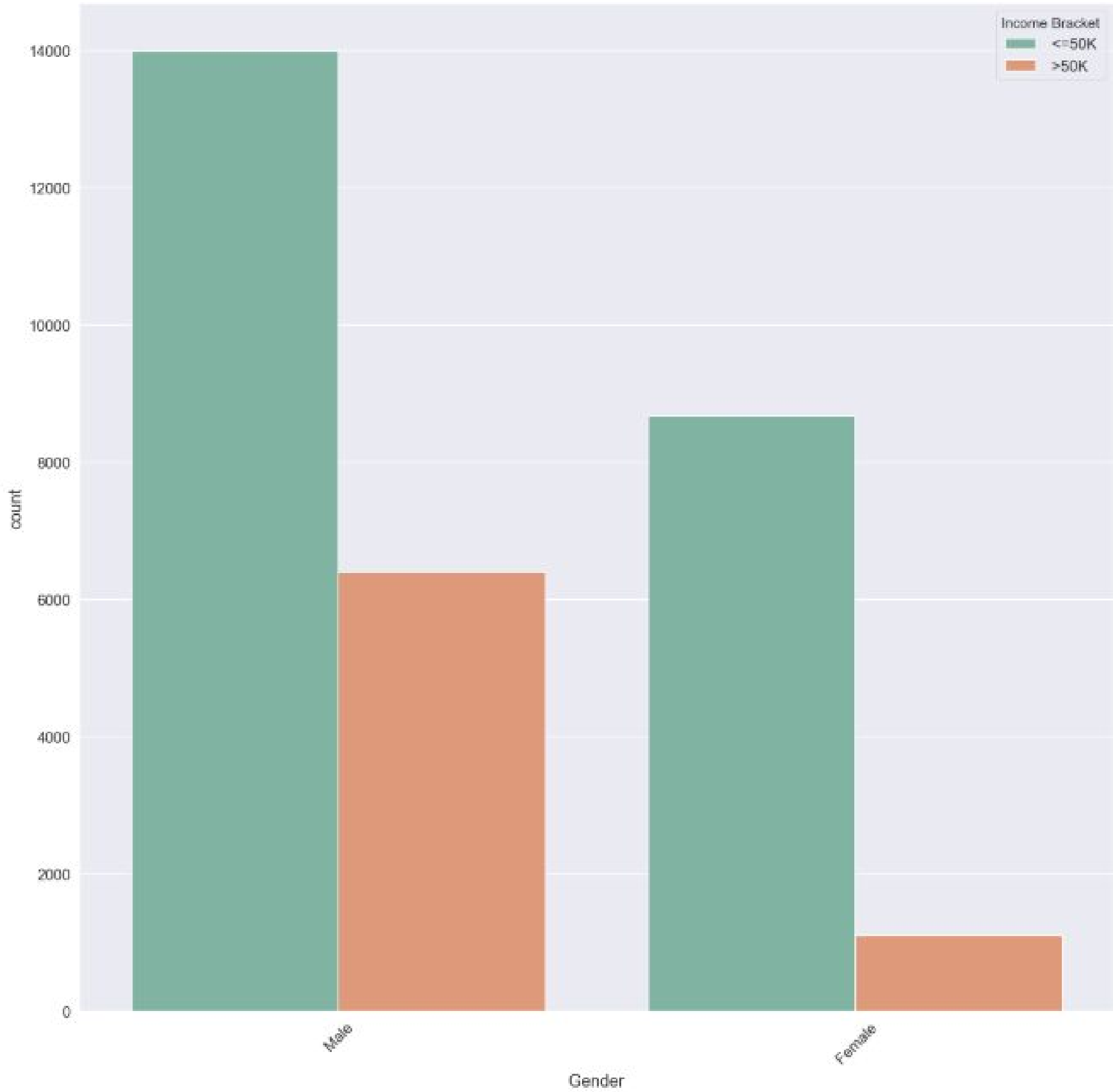
The figure above shows that most people are working in the private sector. Self employed workers have more individuals making >50K. In all the other working classes, there is a huge gap between <=50K and >50K, most people in these sectors make <=50K.



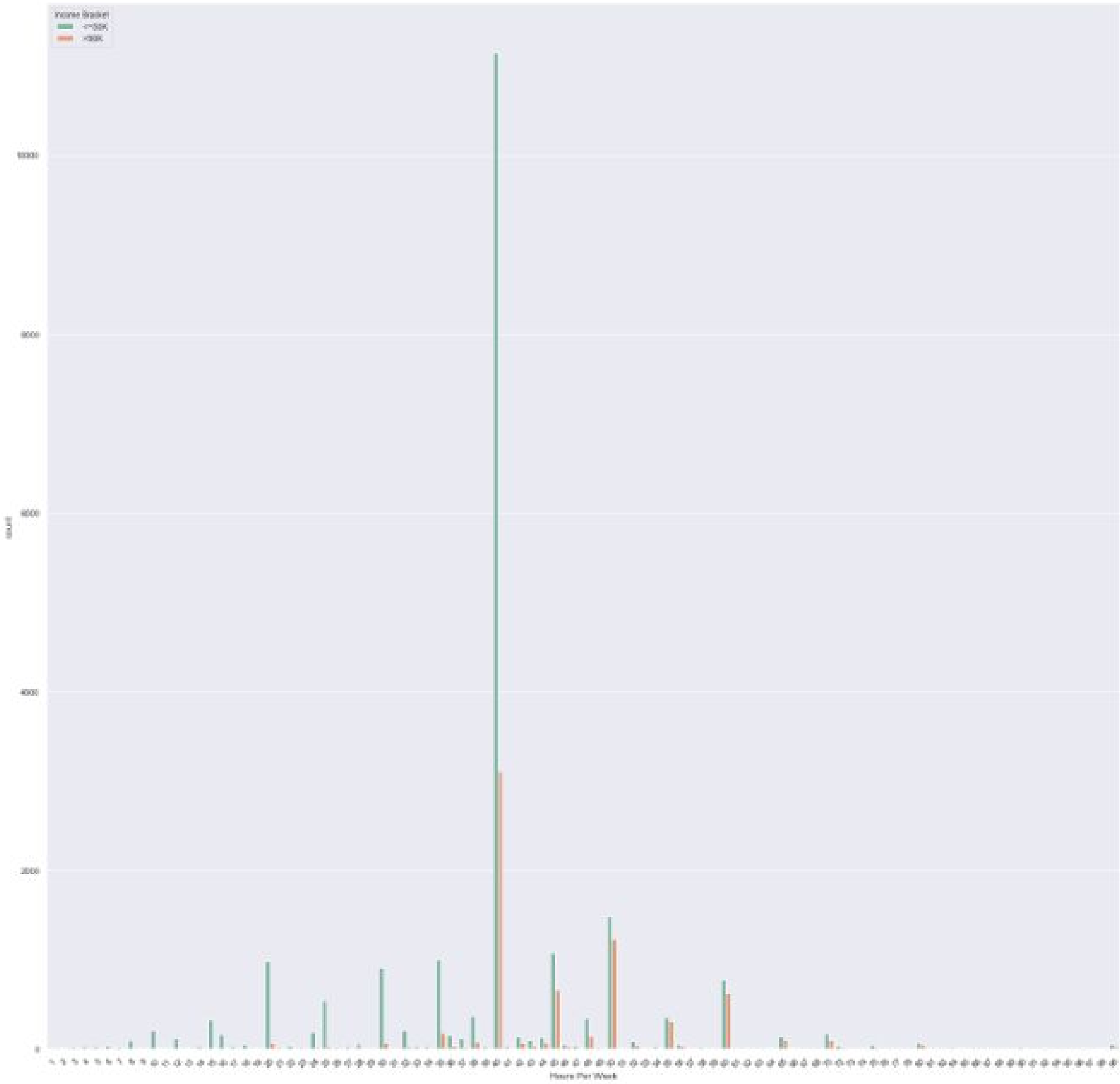
From the figure above, most individuals from this dataset have an education of highschool or more. In comparison, there are more individuals making >50K than <=50K with an education of masters, doctorate and prof-school.



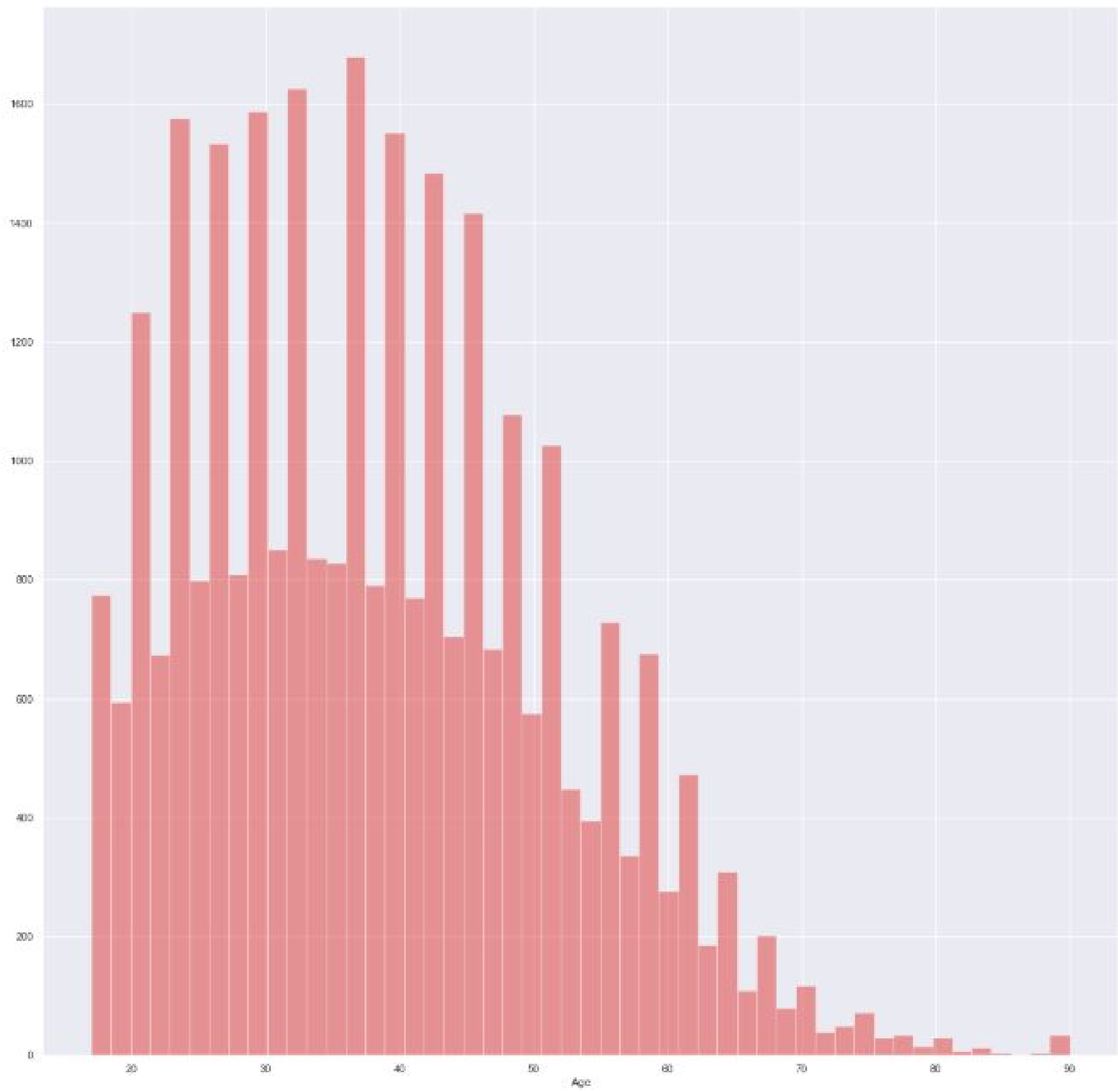
The figure above shows the histogram of the different occupations in the dataset for incomes <=50K and >50K. Handlers-cleaners and Priv-house-serv have the greatest difference of income. Exec-managerial has the least difference as there are far more people in this occupation making >50K. From all of the occupations shown, not a single occupation has a ratio where there are more individuals making >50K than <=50K.



The figure above shows the histogram for males and females in the dataset for incomes <=50K and >50K. There are far less females in ratio making >50K in comparison to males. All together the dataset has far more males than females.

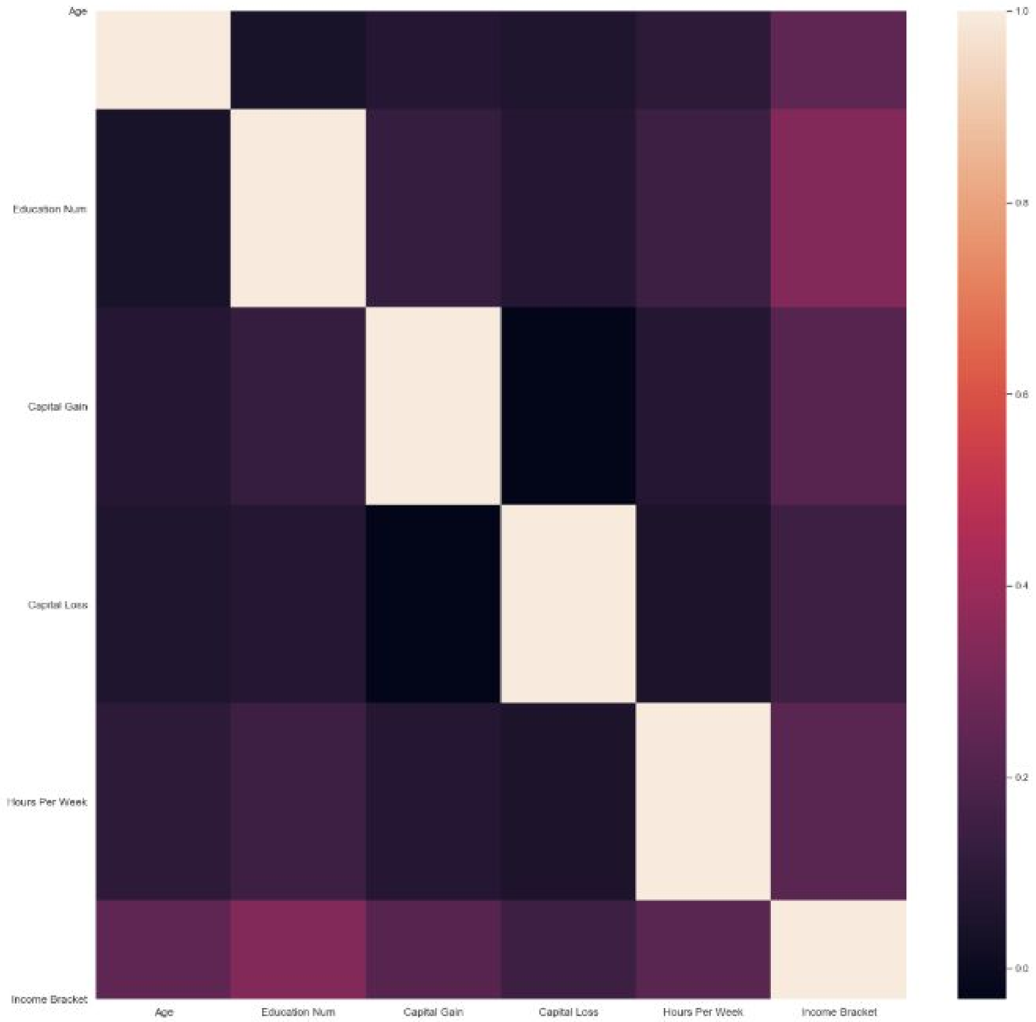


The figure above shows the histogram of the "Hours Per Week" in the dataset for incomes <=50K and >50K. Most individuals are working 40 hours per week. A large amount of individuals who make >50K seem to work 40 hours or more per week.



The figure above shows the histogram of the different ages in the dataset. There is a wide age gap in this dataset, from 17 years old to 90 years old.

	Age	Education Num	Capital Gain	Capital Loss	Hours Per Week	Income Bracket
Age	1.000000	0.043526	0.080154	0.060165	0.101599	0.241998
Education Num	0.043526	1.000000	0.124416	0.079646	0.152522	0.335286
Capital Gain	0.080154	0.124416	1.000000	-0.032229	0.080432	0.221196
Capital Loss	0.060165	0.079646	-0.032229	1.000000	0.052417	0.150053
Hours Per Week	0.101599	0.152522	0.080432	0.052417	1.000000	0.229480
Income Bracket	0.241998	0.335286	0.221196	0.150053	0.229480	1.000000



The table and figure above shows a correlation matrix of the numerical variables in the dataset. The correlation matrix is geared towards the Income Bracket and its correlation towards the other numerical variables. Education Num has the highest correlation while fnlwgt has the lowest correlation.

Machine Learning

The objective of this project is to predict the "Income Bracket" using binary classification. I will be using decision tree model, logistic regression model, random forest classifier model and SVC model to evaluate which of the following will give the best accuracy. For preprocessing, the data was split into 80 and 20 percent for training and test data. Before using the different machine learning models I scaled the features using standard scaler and made the categorical variables into dummy variables. The table below shows the four different machine learning models used and its accuracy.

Decision Tree Model
Accuracy: 0.8085529587270015
Logistic Regression Model
Accuracy: 0.8102105088678933
Random Forest Classifier Model
Accuracy: 0.8143543842201226
SVC Model
Accuracy: 0.8219791148682247

Model Tuning

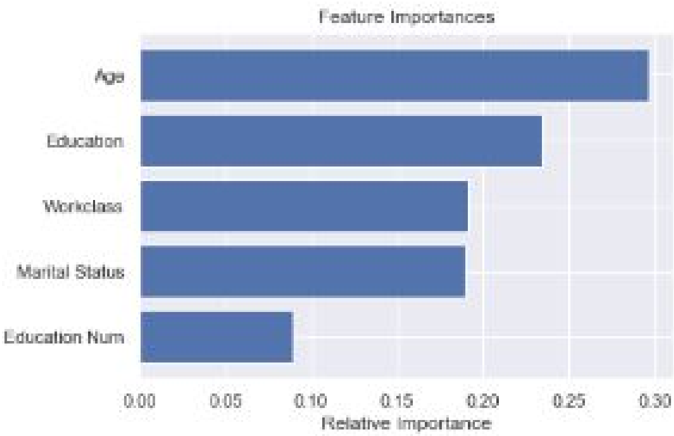
SVC model had the best accuracy out of the four models, as a result I tuned this model using GridSearchCV.

Before Model Tuning	After Model Tuning
Accuracy: 0.8219791148682247	Accuracy: 0.8252942151500083

Classification Report

	precision	recall	f1-score	support
0	0.84	0.95	0.89	4503
1	0.76	0.45	0.57	1530
accuracy			0.83	6033
macro avg	0.80	0.70	0.73	6033
weighted avg	0.82	0.83	0.81	6033

Feature Importance



The figure above shows the top 20 important features, these features provide the most predictive power on the dataset.

Conclusion

From the feature importances, Age had the most predictive power on the dataset. Out of the following classification models used, the SVC Model gave the best accuracy. After tuning the model using GridSearchCV I received an accuracy of 0.825. The age of an individual has the highest predictive power from the dataset.

Recommendations & Future Improvements

Only four models were used, by using more models it will help to analyze which of the models will give a better accuracy. I also recommend using imblearn for future work, this will help over-sampling data such as the income bracket which has far more individual making <=50K than in comparison to >50K.