1. **Objective**

The overall objective is to determine if someone earns more than 50. Specifically, the objective is to build a model that predicts if a person earns more than 50K.

The data for this is obtained from "https://archive.ics.uci.edu/ml/machine-learning-databases/adult/''.

2. **Exploration and Data Analysis**

2.1. Data Description

The data has 15 attributes in which 14 are features and one will be made the target set. A look through the data shows the data columns are not properly labelled as seen in figure 1. Thus, exploratory data analysis will involve renaming the columns by looking at the appropriate name given at the file's location. The corresponding object types show that there are a lot of object datatypes which will need to be converted to numbers during future engineering.

```
39                 int64
 State-gov        object
 77516             int64
 Bachelors        object
 13                int64
 Never-married    object
 Adm-clerical     object
 Not-in-family    object
 White            object
 Male             object
 2174              int64
 0                 int64
 40                int64
 United-States    object
 <=50K            object
dtype: object
```

Figure 1: Miss-labelled columns of the data set and corresponding data type.

From the file of the data, the column names are age, workclass, fnlwgt, education, education num, marital status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, country and salary. Table 1 shows the columns and corresponding data description.

Table 1: Column Name and Data Values of Data Set

Column Name	Data
age	continuous
workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
fnlwgt	continuous
education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
education-num	continuous

marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
sex	Female, Male
capital-gain	continuous
capital-loss	continuous
hours-per-week	continuous
native-country	country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
>50K, <=50K	

As can be seen the last column has two values >50K, <=50K which will be renamed as greater than 50K and less than or equal to 50K. This will be called 'salary' and will be set as the target variable (Y).

After renaming the columns, the new columns names are shown in Figure 2.

```
age                int64
workclass         object
fnlwgt             int64
education         object
education-num      int64
marital-status    object
occupation        object
relationship      object
race              object
sex               object
capital-gain       int64
capital-loss       int64
hours-per-week     int64
country           object
salary            object
dtype: object
```

Figure 2: Rename Column names and corresponding object type.

The number of values in each column is shown in Table 2. It shows that the amount of data in each colu mn is 32560 and there are no missing values.

Table 2: Information of the Data Set

```
Range Index: 32560 entries, 0 to 32559
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32560 non-null  int64
 1   workclass       32560 non-null  object
 2   fnlwgt          32560 non-null  int64
 3   education       32560 non-null  object
 4   education-num   32560 non-null  int64
 5   marital-status  32560 non-null  object
 6   occupation      32560 non-null  object
 7   relationship    32560 non-null  object
 8   race            32560 non-null  object
 9   sex             32560 non-null  object
 10  capital-gain    32560 non-null  int64
 11  capital-loss    32560 non-null  int64
 12  hours-per-week  32560 non-null  int64
 13  country         32560 non-null  object
 14  salary          32560 non-null  object
dtypes: int64(6), object(9)
```

2.2. Visualizations

For exploratory data analysis, some visualizations are done. These are shown in figures 3 through 5. Figures 3 shows the workclass variable. It can be seen that workers in the Private sector earn above 50K. Figure 4 shows that those with some form of education (high school, college, masters) earn more than 50K than those with other kinds of education. Figure 5 indicates that married couples who stay together earn more than 50k.

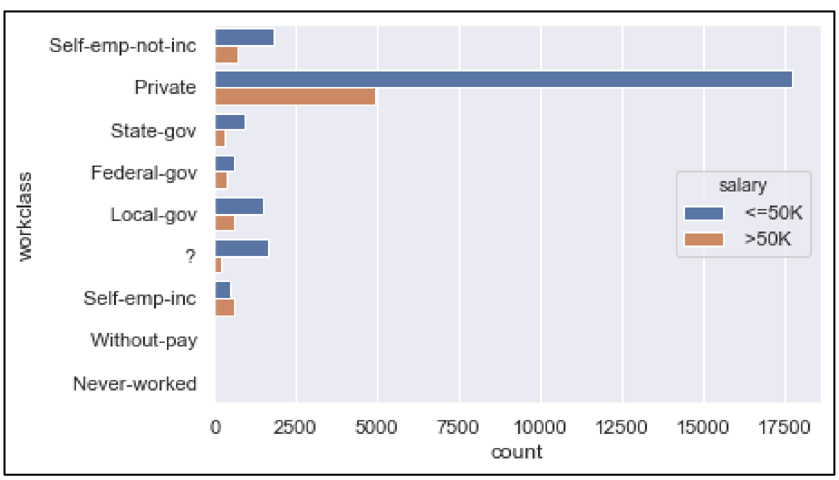

Figure 3: Worker's earnings per class

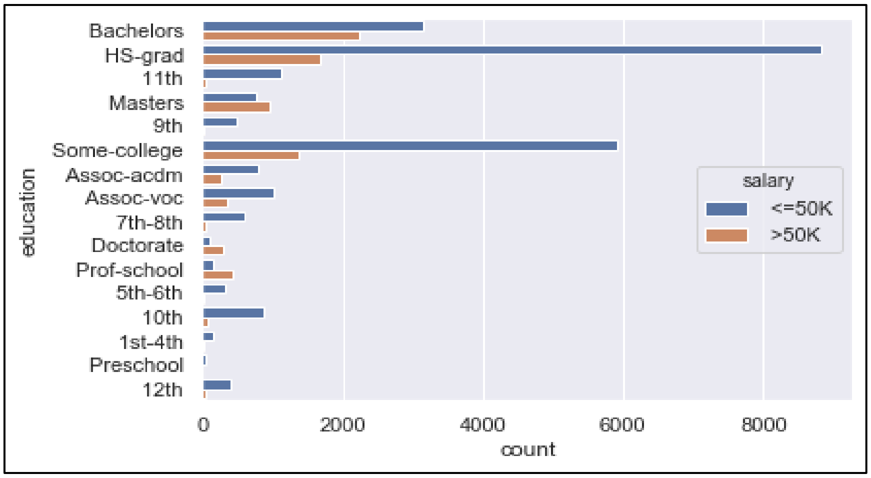

Figure 4: Earning per Education

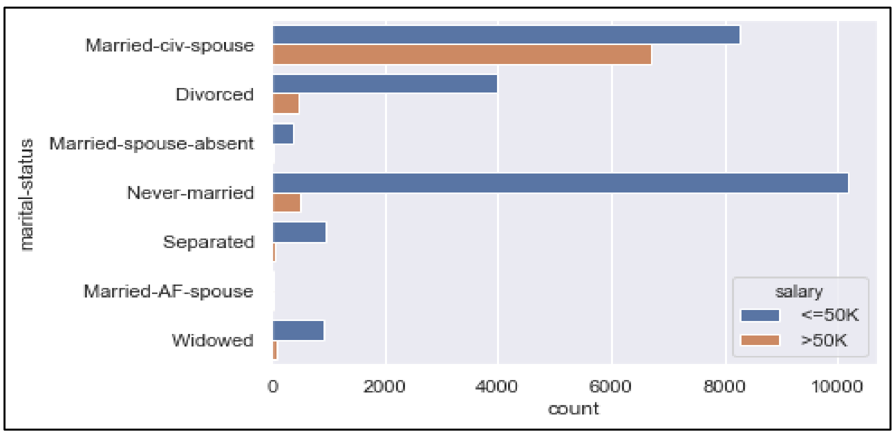

Figure 5: Earnings based on Marital Status

Other visualizations show that those who earn more than 50k are workers such as executive managers, professional specialties, those in sales and craft-repairs (Figure 6); husbands (Figure 7); whites (Figure 8); males (Figure 9) and stay in the United States (Figure 10).

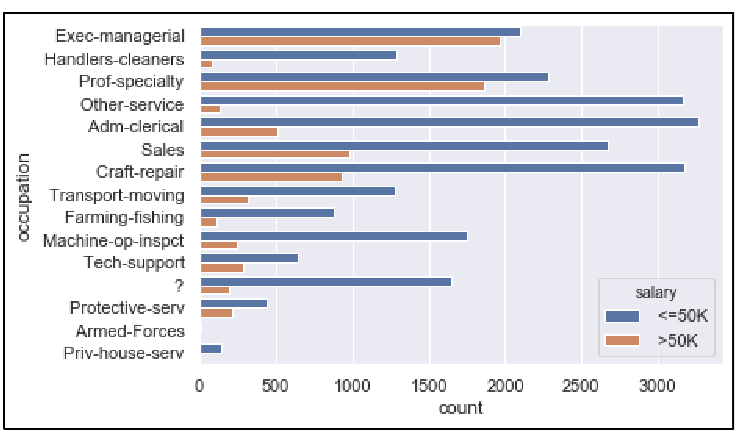

Figure 6: Earnings of Different Occupations

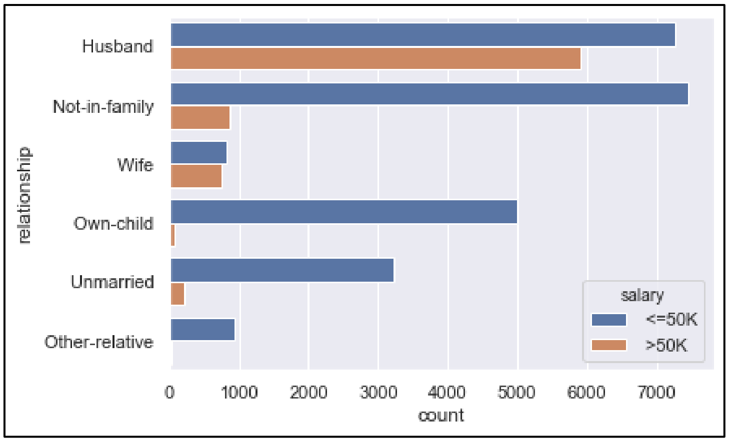

Figure 7: Earnings based on Relationship

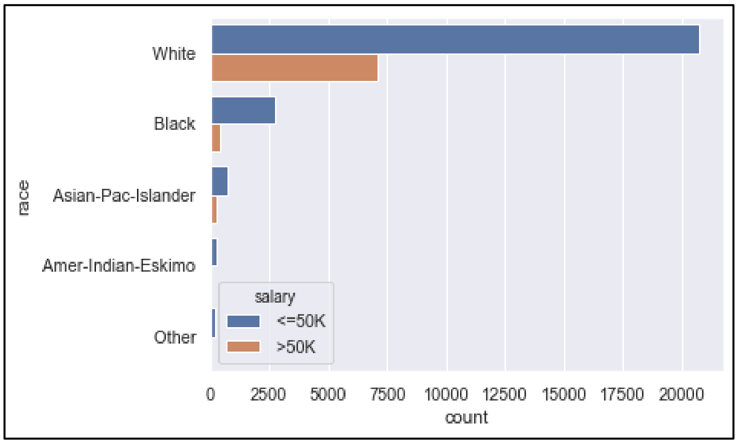

Figure 8: Earnings of Different Races

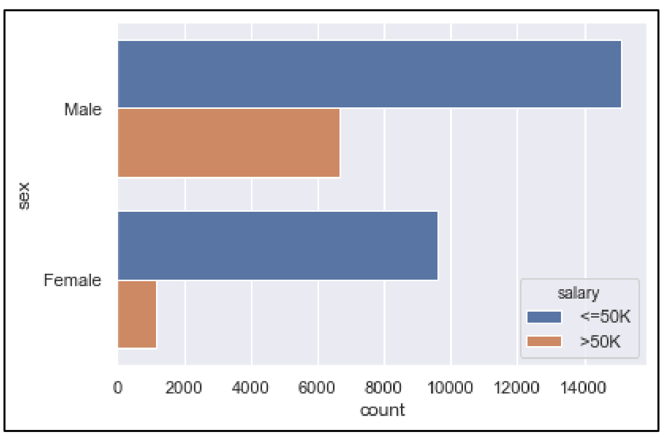

Figure 9: Earnings of Different Sex

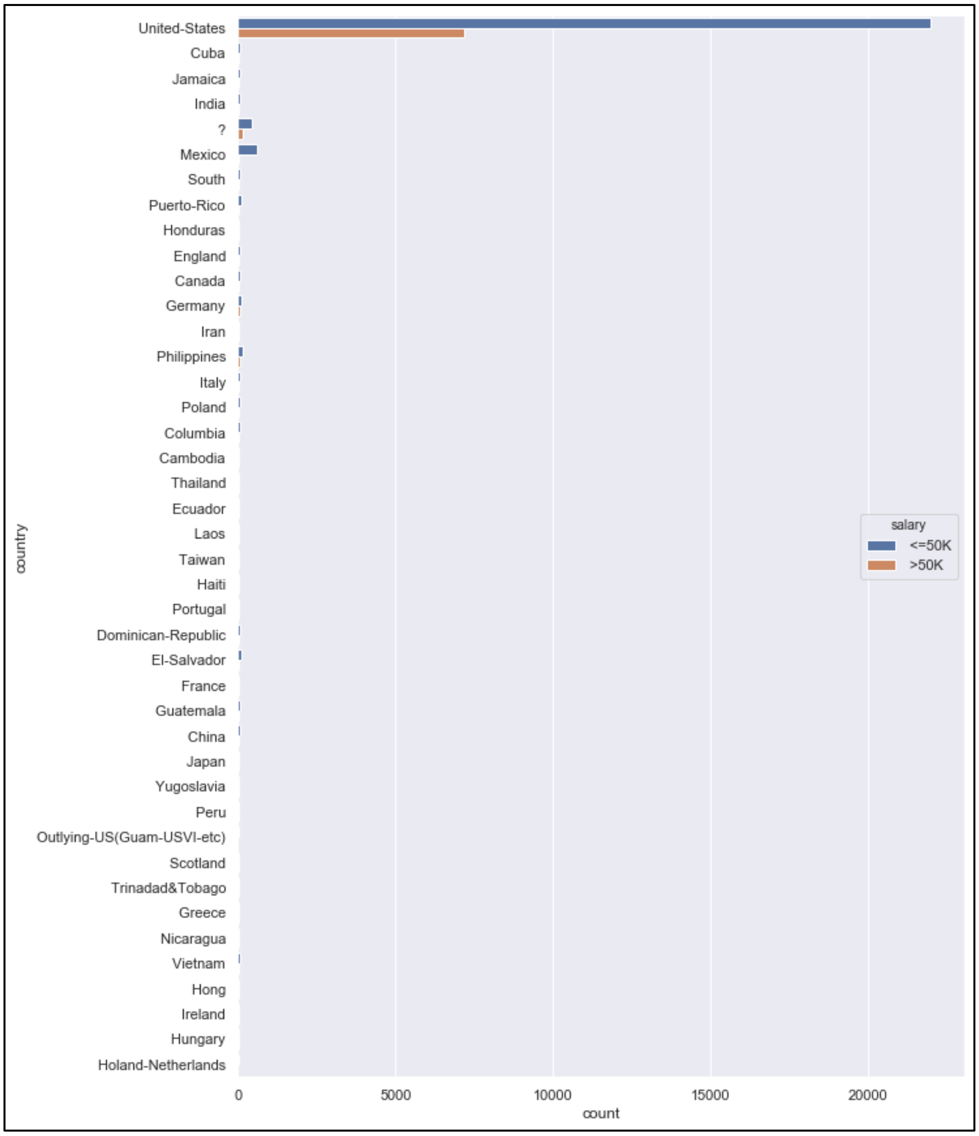

Figure 10: Earnings in Different Countries.

3. Feature Engineering

The feature engineering will focus on converting the non-categorical values to numerical ones. The salary column will be converted using label encoder while the 'one-hot-encoder' using 'get dummies' method is used for the other columns.

Furthermore, the columns are separated to X (independent) variables and y (dependent) variable where the salary column becomes the y (dependent variable) and other columns are the independent variables. Thus, salary is predicted using the other variables.

Table 3: View of Data Set before Feature Engineering

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	country	salary
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
1	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
2	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
3	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
4	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K

4. Pre-Modelling

The data is split into training and testing set to ensure that there is an even ratio of the variable to be predicted is the same in both the training and testing set. After application, the ratio is the same in both the test and training sets as seen in Table 4.

Table 4: Training Value Counts In both Train and Tests Sets of Salary

Train		Test	
0	0.75917		0.75917
1	0.24083		0.24083

Furthermore, the independent variables are scaled.

5. Modelling

Since the objective is to predict a categorical variable, the logistic regression, logistic regression with lasso regression (L1) and Random Forests are used. The accuracy, precision, recall, F1 and AUC scores are used for performance measurement of the models.

Performance Metrics	Scores		
	Logistic Regression	Logistic Regression with Lassa (L1) Regularization	Random Forests
Accuracy	0.857801	0.857801	0.858825
Precision	0.752756	0.752756	0.743372
Recall	0.609694	0.609694	0.631803
F1	0.673714	0.673714	0.683061
AUC	0.910556	0.910556	0.908223

The Random Forest Model has the highest accuracy and the highest F1 values. However, in terms of precision. Recall and AUC, the logistic regression models (with and without regularization) have higher values and are similar.

The Random Forest Model is chosen since it has higher F1 values which incorporates both precision and recall and also has higher accuracy.

To have an idea of the influence of each independent variables in predicting if salary is greater than 50k, the feature importance is carried out and it shows that fnlwgt, age, capital-gain, marital status, hours-per-week have higher influence on if income is greater than 50k.

Observations/Suggestions

The random forest is a better predictor with 85.88% accuracy and in this data set features such as age, capital gains, marital status, hours of work per week, education level (num) have higher influence on predicting if a person earns more than 50K.