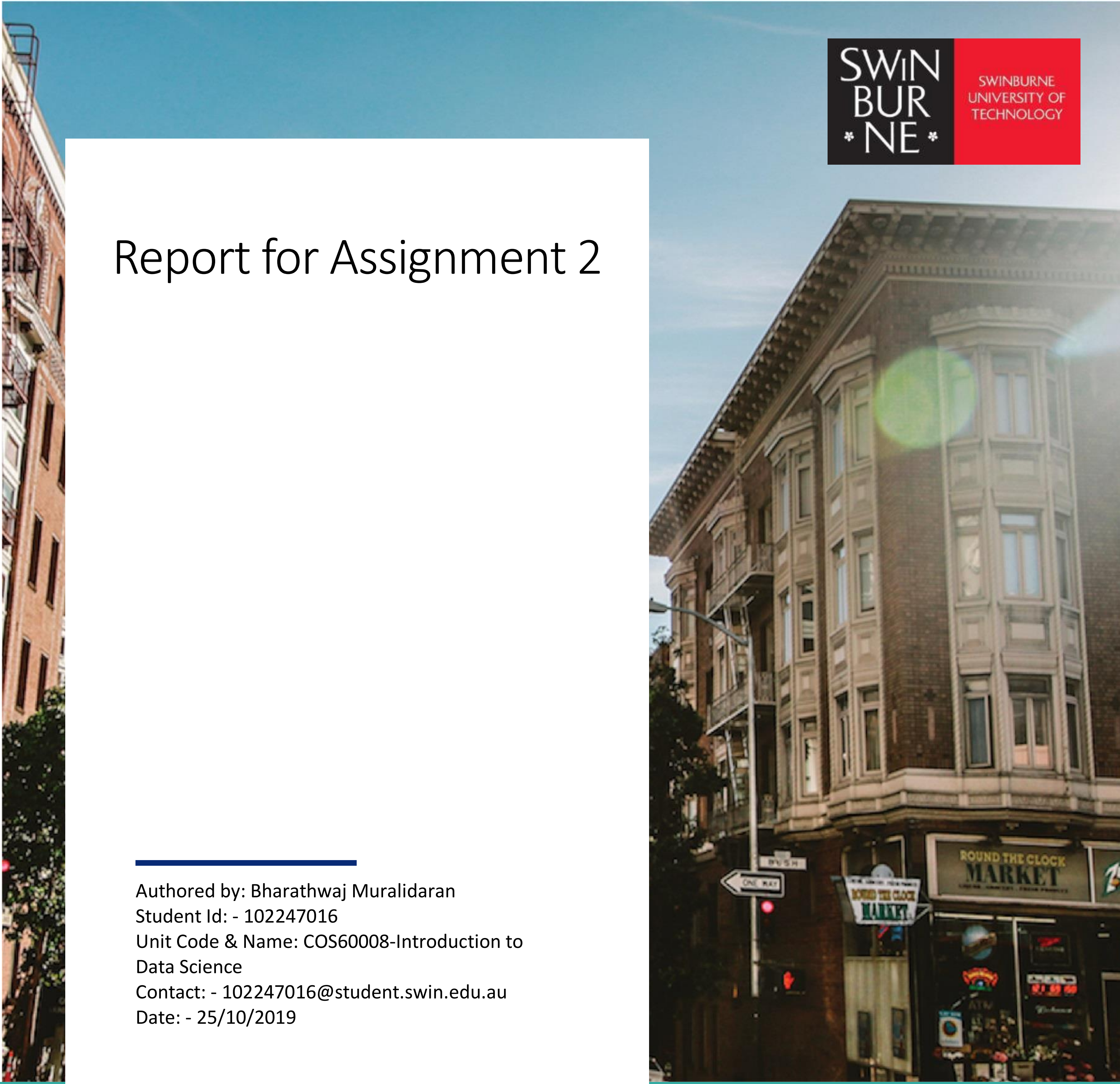


# Report for Assignment 2

Authored by: Bharathwaj Muralidaran
Student Id: - 102247016
Unit Code & Name: COS60008-Introduction to Data Science
Contact: - 102247016@student.swin.edu.au
Date: - 25/10/2019

# Executive Summary

The main aim of this assignment is to formulate a problem and find the appropriate predictive solution by performing modelling techniques. The report is related to the Adult dataset which deals with the census data of USA. The agenda of selecting the Census dataset is to identify the total number of categories which are responsible for earning salary more than 50,000. The report will explain the techniques used for data preparation which involves data cleaning and data transformation. It also addresses important insights by performing exploratory data analysis. The report will also cover the best modelling technique which can be applied on the Census Datasets.

# Introduction

The Census dataset is collected from the US Census Bureau database which comprises of 32,561 entries of data. The report is divided into three sections of the task were each task refer the important steps of data science like data preparation, data transformation, exploratory data analysis and data modelling. The first task will be focusing on finding ambiguous data and address the necessary methods to clean the data. The second task will focus on exploratory data analysis where we find some important insights based on comparing different categories in the dataset. The final task will distinguish the major classification technique to identify which technique would be the best in performing effective model.
The Dataset: -
The Census dataset of US Census Bureau has 9 categorical column and 6 numerical columns(http://archive.ics.uci.edu/ml/datasets/Census+Income). Each column describe the associated data and they are as follows: -
Age: - The column comprises data related to age of individuals.

Work class: - The column determines the employment of different individuals.
State-gov, Self-emp-not-inc, Private, Federal-gov, Local-gov, nan, Self-emp-inc, Without-pay, Never-worked.

Fnlwgt: - The column represents the total number of people represent the specific category and it is termed as final weight.

Education: - The column comprises data about the education level of individuals.

- Bachelors, HS-grad, 11th, Masters, 9th, Some-college, Assoc-acdm, Assoc-voc, 7th-8th, Doctorate, Prof-school,5th-6th, 10th, 1st-4th, Preschool, $12^{th}$.

Education-num:- The column has data related to the education level in numerical form.

Marital-status: - The current marital status of the individual.

- Never-married, Married-civ-spouse, Divorced, Married-spouse-absent, Separated, Married-AF-spouse, Widowed.

Occupation: - The column specifies the occupation or field of work of each individual.

- Adm-clerical, Exec-managerial, Handlers-cleaners,Prof-specialty, Other-service, Sales, Craft-repair,Transport-moving, Farming-fishing, Machine-op-inspct,Tech-support, Protective-serv, Armed-Forces,Priv-house-serv.

Relationship:- The data represent the current relationship of each individual.

- Not-in-family, Husband, Wife, Own-child, Unmarried, Other-relative.

Race:- describe the race of an individual.

- White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo,Other

Sex:- The gender of each individual.

- Male, Female

Capital_gain: - Total capital gain of an individual.

Capital_loss: - Total capital loss of an individual.

Hours_per_week: - Total hours each individual work in a week.

Native_country: - The native country of everyone.

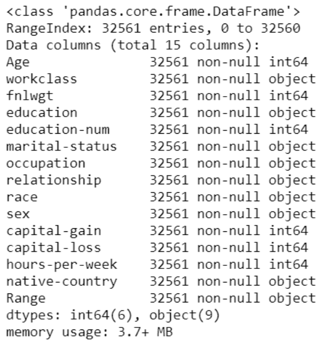
- (United-States, Cuba, Jamaica, India, Mexico, South,Puerto-Rico, Honduras, England, Canada, Germany, Iran,Philippines, Italy, Poland, Columbia, Cambodia,Thailand, Ecuador, Laos, Taiwan, Haiti, Portugal,Dominican-Republic, El-Salvador, France, Guatemala,China, Japan, Yugoslavia, Peru,Outlying-US(Guam-USVI-etc), Scotland, Trinadad&Tobago,Greece, Nicaragua, Vietnam, Hong, Ireland, Hungary,Holand-Netherlands).

Range: - The range specifies the label of each individuals having salary < =or > 50,000.

- <=50K, >50K.

# Task 1 -Problem formulation, Data Acquisition and Preparation

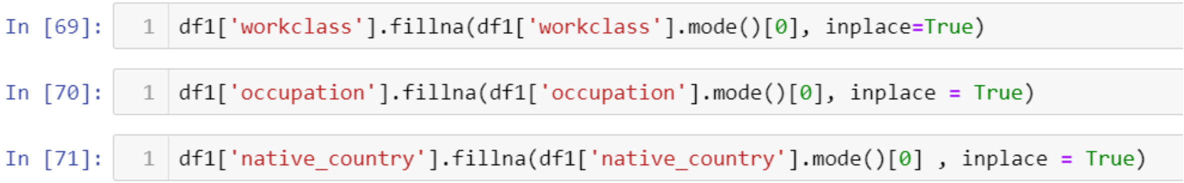
In this task, we identified the appropriate data which satisfies the specified condition and the choice was Census Dataset belongs to USA Census Bureau.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
Age                 32561 non-null int64
workclass           32561 non-null object
fnlwgt              32561 non-null int64
education           32561 non-null object
education-num       32561 non-null int64
marital-status      32561 non-null object
occupation          32561 non-null object
relationship        32561 non-null object
race                32561 non-null object
sex                 32561 non-null object
capital-gain        32561 non-null int64
capital-loss        32561 non-null int64
hours-per-week      32561 non-null int64
native-country      32561 non-null object
Range               32561 non-null object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

Structure of Adult Dataset

The above dataset was loaded into the data frame to perform data cleaning. The dataset had duplicate Values, around 24 duplicate values were found and dropped from the original datasets. The next step was to strip the whitespaces from data of categorical column. After performing stripping of whitespace, the next most important finding was some data from then datasets had missing data which were represented as '?'. The work class, occupation and native country column had missing values which were first replaced with the null value and then these null values were replaced by most occurring data of that column.



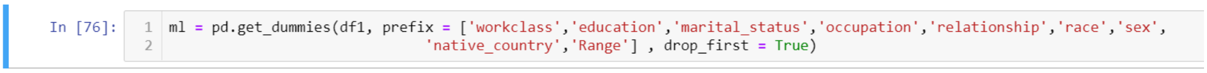
```
In [69]:  df1['workclass'].fillna(df1['workclass'].mode()[0], inplace=True)

In [70]:  df1['occupation'].fillna(df1['occupation'].mode()[0], inplace = True)

In [71]:  df1['native_country'].fillna(df1['native_country'].mode()[0] , inplace = True)
```

Replacing the null value with the mode value of that column

The next step was to perform the feature engineering where we performed the one-hot code which converted all categorical column into 0 and 1. The data obtained from one hot code is then split into data and target for modelling.



```
In [76]:  ml = pd.get_dummies(df1, prefix = ['workclass','education','marital_status','occupation','relationship','race','sex',
                                       'native_country','Range'] , drop_first = True)
```

One-hot code method

# Task 2- Exploratory Data Analytics

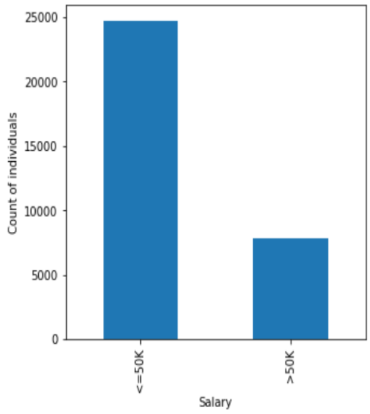


Fig 1: - Total number of salary count

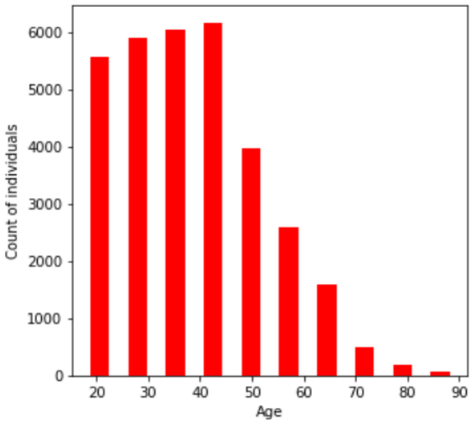


Fig 2: -Total number of individuals in the different age group

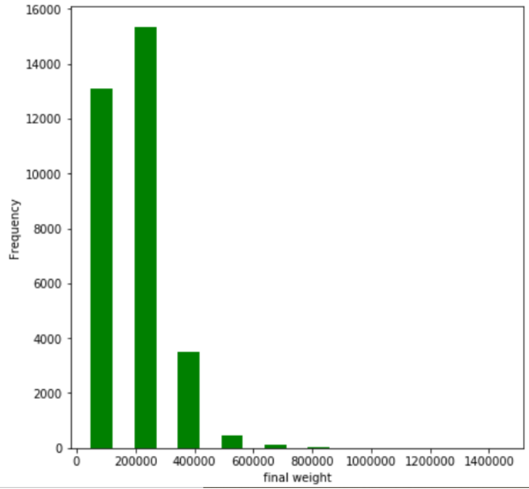


Fig 3: - Total number of individuals in a specific category

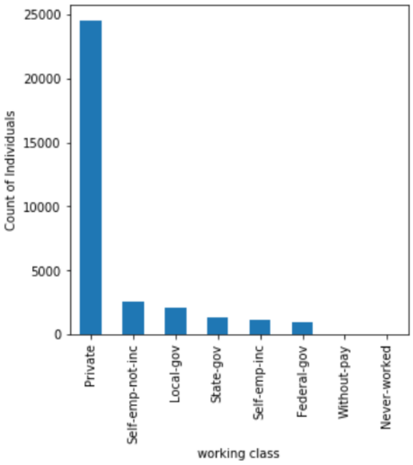


Fig 4: -Total number of people in different working class

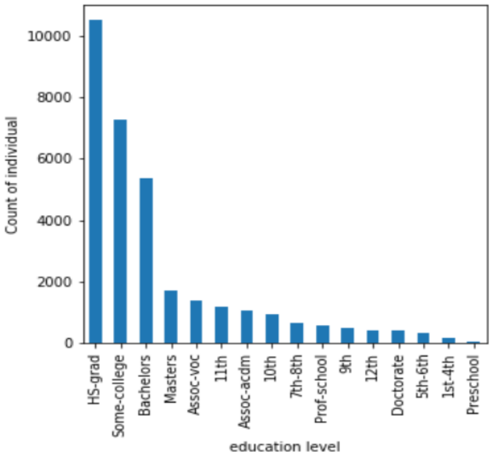


Fig 5: - Total number of people from different education level

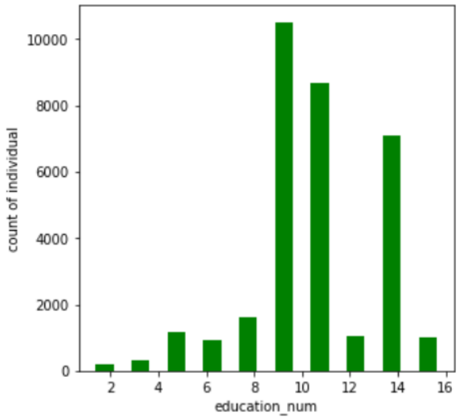


Fig 6: - Total number of people from different education levels in number

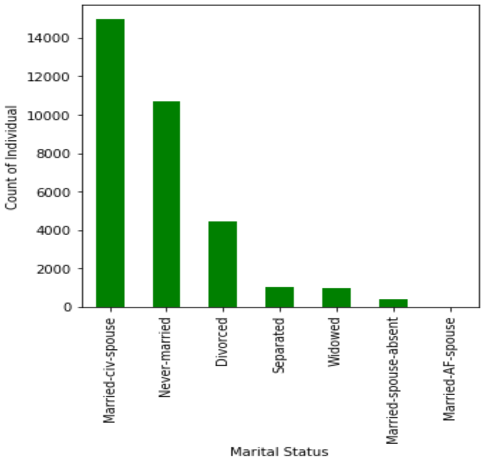


Fig 7: - Total number of people having different Marital status

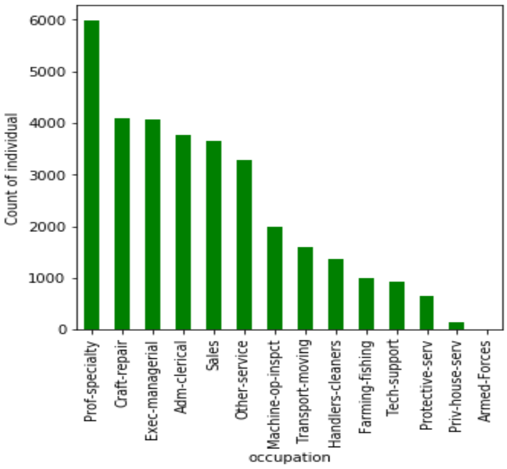


Fig 8: - Total number of people from different Occupation

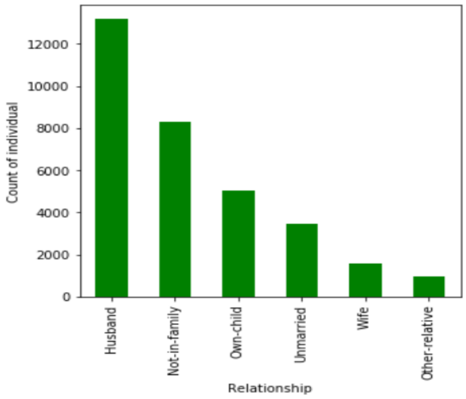


Fig 9: - Total number of people having different Relationship status

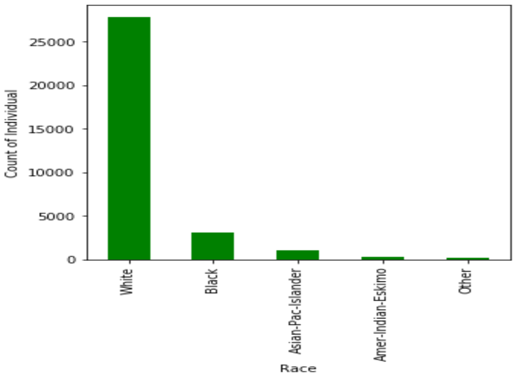


Fig 10: - Total number of people having different Race

One the basis of analysis based on individual columns from fig 1 we can say that there are comparatively a smaller number of individuals earning more than 50,000 and from fig 2 we understand the major amount of people fall under the category of age group of 25 to 45. The fig 3 and fig 4 conclude that about 0.2 million individual falls under the category of work class as private. From fig 5 and 6 we can say that most of the individuals are basically are High School graduated. Most of the population has occupation as professional specialist (fig 8) and they live with their spouse (fig 7). The population also has highest number of white people as compared to another race (fig 10). The population also has more working husbands as compared to working wife in a family (fig 9).

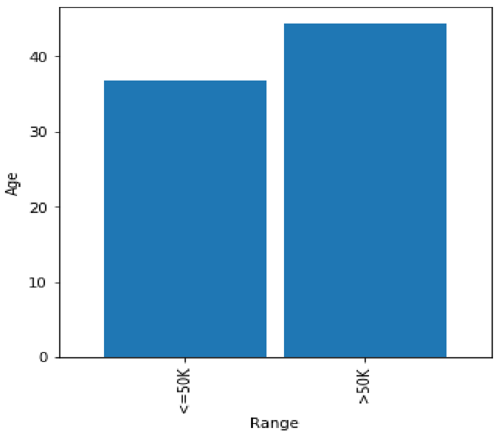


Fig 11: - Average age of people earning more than 50K

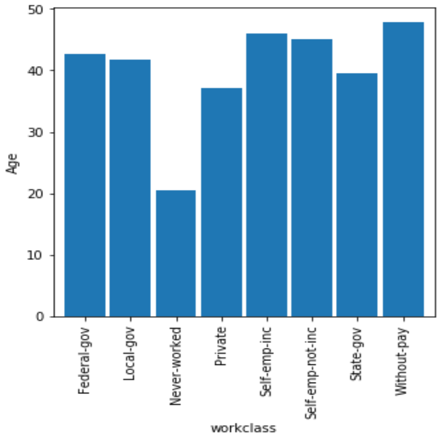


Fig 12: - Average age of people working in different work class

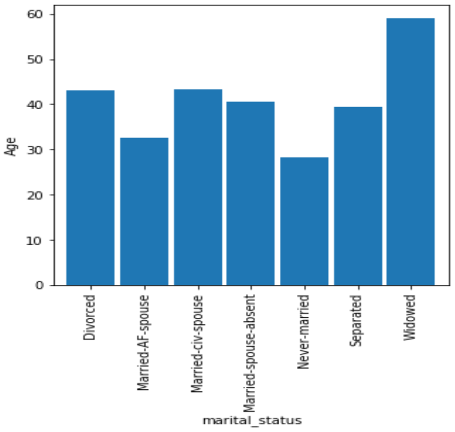


Fig 13: - Average age of people having different marital status

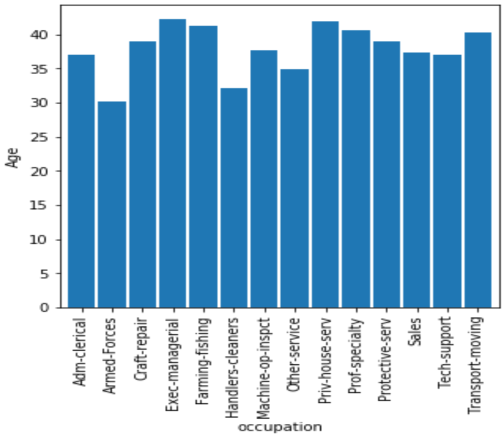


Fig 14: - Average age of people from different occupation

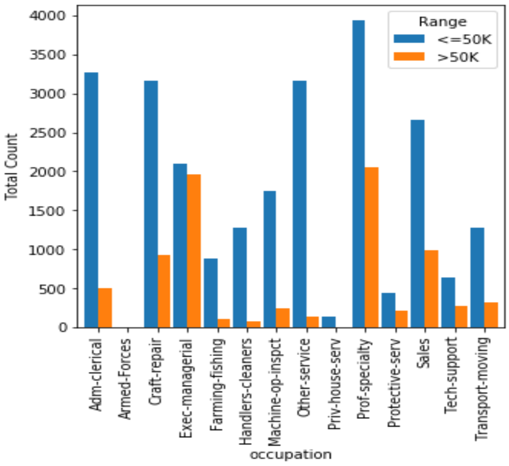


Fig 15: - Total count of people from different occupation earning 50K

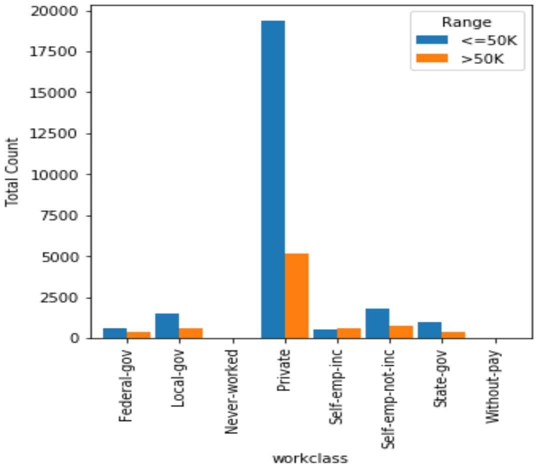


Fig 16: - Total count of people from different work class earning 50K

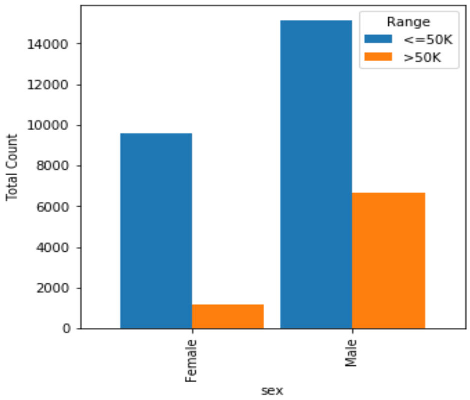


Fig 17: - Comparison between the gender and their earning

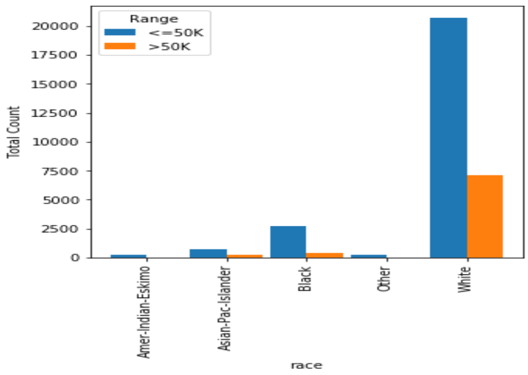


Fig 18: - Comparison between the race and their earning.

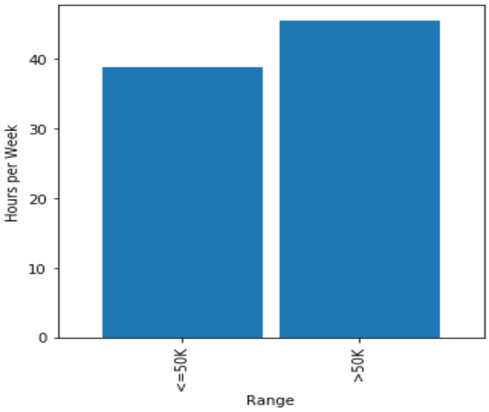


Fig 19: - Earnings based on the total number of hours people work per week

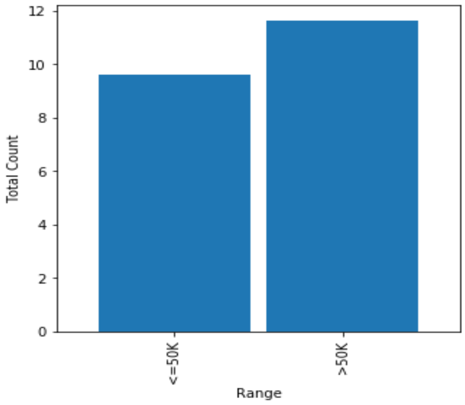


Fig 20: - Earnings based on average education level.

In this task, we formulated a problem where we explored the whole dataset to find what factors that determine whether an individual can earn more or less than salary 50,000?

The individual with an average age of more than 40 has salary greater than 50,000(fig 11) and they all belong to some important work-class like federal government and local government (fig 12). The individual having age more than 40 has a marital status as divorced or widowed (fig 13) and most of them have an occupation of executive manager and private house servicing (fig 14).It has been observed that large amount of individual is working in an occupation as professional specialist but only half of the total count can make it earn more than 50,000(fig 15).The private sector individual earn more as compared to other sectors people and more male gender individual earn more than 50,000 (fig 16 & 17).The individual having grade more than 12 i.e high school graduation and work more than 40 hours per week earns more than 50,000(fig 19 and 20).

So, from the above graphs, we can conclude our hypothesis that individuals holding degree of high school graduation and working in private sector with an occupation of executive manager can expect to earn more than 50,000 at the age of 40.

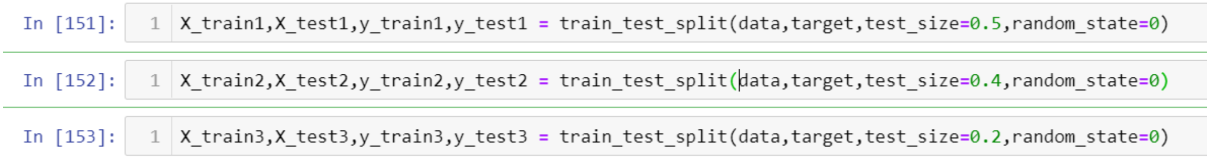
## Task 3 – Data Modelling

### Step 3.1

We used the method of train_test_split which splits the given data into training, testing and further splits the data into training and validation sets.

The training set of data is used for data modelling and validation set is used especially for hyperparameter tuning and test sets will evaluate the performance of data in real-world scenarios.

The data and the target is passed in the train_test_split and splitting the data into three different splits i.e 50-50,60-40 and 80-20 by keeping the random state as 0.



```
In [151]: X_train1,X_test1,y_train1,y_test1 = train_test_split(data,target,test_size=0.5,random_state=0)

In [152]: X_train2,X_test2,y_train2,y_test2 = train_test_split(data,target,test_size=0.4,random_state=0)

In [153]: X_train3,X_test3,y_train3,y_test3 = train_test_split(data,target,test_size=0.2,random_state=0)
```

### Step 3.2

**KNN Modelling**

It is a supervised learning technique and, in this method,, the object is arranged to a specific class which has common K nearest neighbours. In this method we found the best K value through performing cross-validation score to attain the best accuracy. The cross-validation score was set to 10 because of huge dataset. We performed KNN on three different data splits to find the best accuracy score. The performance of the KNN is measured by accuracy, confusion matrix, precision, recall and F1 score.

**Decision Tree**

It is supervised learning which represents tree-structured model, it predicts the value of the target variable by learning decision rules from the training data. We use the same method of cross-validation score to find the maximum depth of the decision tree for different splits.

**Measuring factors for models**: -

**Accuracy**: - Accuracy is one method which is used to evaluate the classification models. It is a method used to predict whether we got the model right.

Accuracy = TP+TN/(TP+TN+FP+FN)

$TP$ = True Positives, $TN$ = True Negatives, $FP$ = False Positives, and $FN$ = False Negatives

**Confusion Matrix**: - Confusion matrix is often showing the summary of prediction results on a classification problem. The number of correct and incorrect value is demonstrated using a matrix.

| | Class 1 Predicted | Class 2 Predicted |
|---|---|---|
| Class 1 Actual | TP | FN |
| Class 2 Actual | FP | TN |

**Precision**:- It calculates the total True positive by sum of true positive and false negative to specify the precision of the selected modelling technique.

Precision = TP/(TP+FN)

**Recall**:- It calculates the total True positive by the sum of true positive and false negative to specify the recall value of selected modelling technique

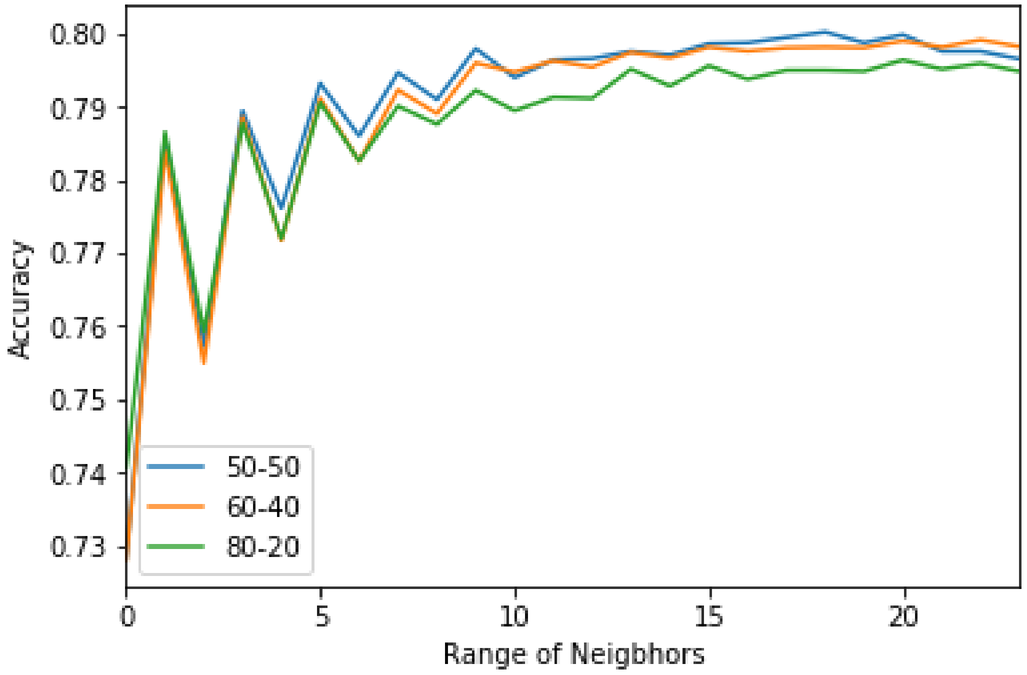
Recall = TP/ (TP + FN)

**F1 Score**:-It is the weighted average of precision and recall. It takes both true positive and false negative into its consideration.

F1 score = 2*(Recall*Precision)/ (Recall + Precision)

We used the technique of classification report to demonstrate the precision,recall,f1 score of both modelling techniques.

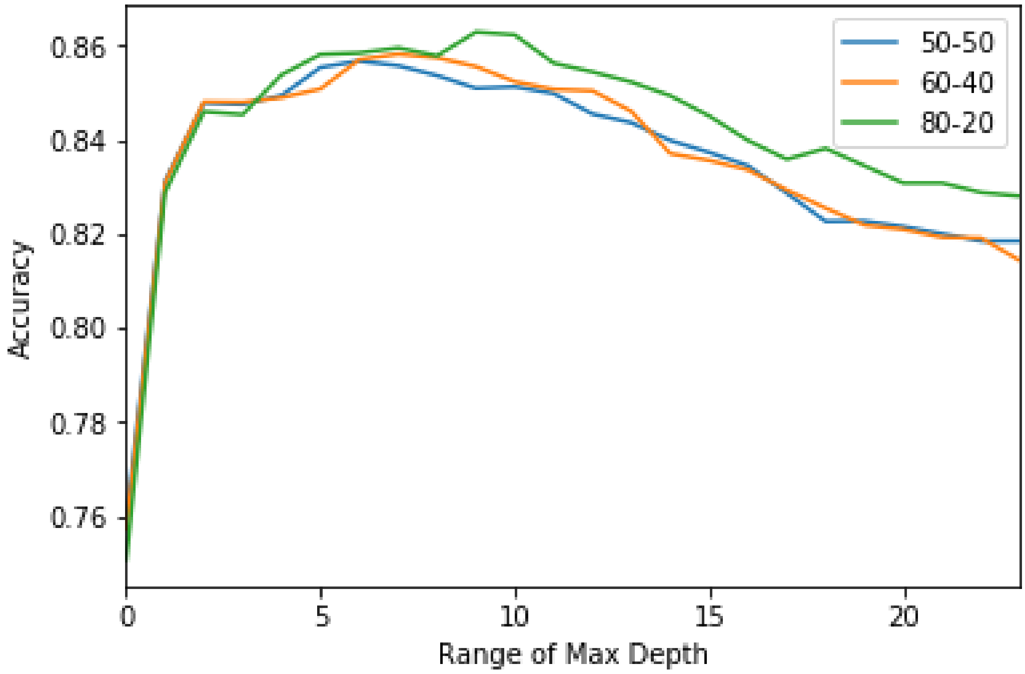
**Step 3.3**



Graphical representation of the accuracy rate of three splits.

| Splits | Accuracy |
|---|---|
| 50-50 | 0.7975905095580552 |
| 60-40 | 0.7991548213599693 |
| 80-20 | 0.7956361401352182 |

The result of KNN performed on three different split says that the accuracy obtained through 60-40 split is more as compared to other splits and precision rate of 0.79 for income <=50 and 0.88 for income > 50.



Graphical Representation of Decision tree accuracy

| Splits | Accuracy |
|---|---|
| 50-50 | 0.8567213719343537 |
| 60-40 | 0.857164809834806 |
| 80-20 | 0.8584818684695759 |

The result obtained through applying the decision tree on three different splits were 80-20 split has more accuracy rate of 0.858 as compared with other splits. The precision rate of 0.89 for income <= 50 and 0.65 for income > 50.

So, we can conclude that Decision Tree is the best model as compared to KNN because the model achieves the highest amount of accuracy while performing decision tree on three different splits as compared to the accuracy attained by KNN method.

## Discussion and Conclusion

In this report, we had performed the important steps of data preparation and explore the data to find the insights which can satisfy our hypothesis. The exploratory data analysis helps us to identify the key insights. We determined a hypothesis which basically finds the best categories which are responsible for an individual to earn more than 50K per year. Through exploratory data analysis we found certain important categories which were responsible to predict the outcome. So, to support the hypothesis, we chose two different predictive models from the classification. The classification models that we applied were KNN and decision tree while comparing both the models we found that KNN attained the accuracy of 0.799 with the k best of 23 but through decision tree the accuracy of 0.858 was attained with maximum depth of 7. The learning outcome from these two-modelling techniques is that better accuracy and precision rate determines the best algorithm for modelling.