Salary of a Person

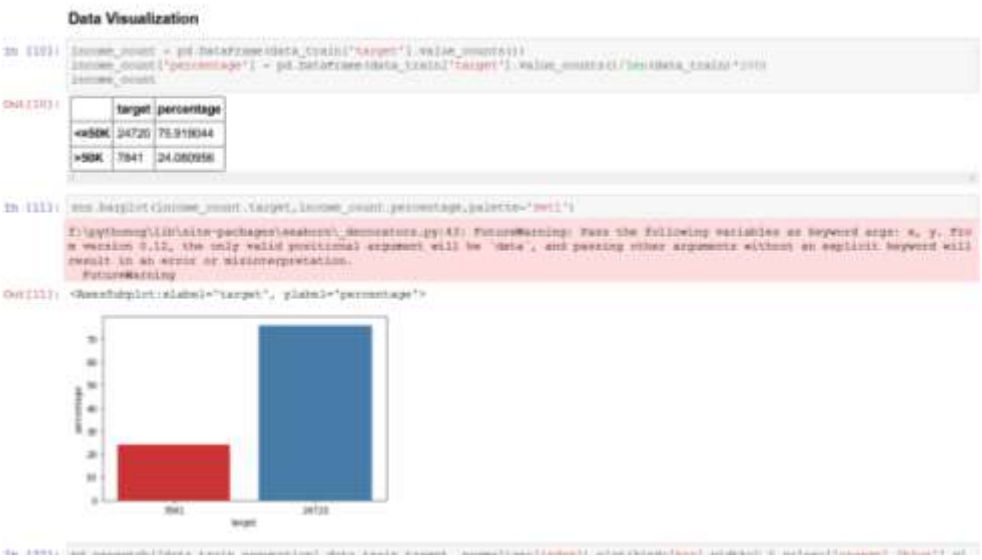
Name: Subash.S
Roll.No:1833052

Problem Statement 1

1. Import the necessary libraries like pandas, Numpy, Seaborn etc.
2. Reading the data with the help of pandas
3. Removing the column names of dataset so that it is easy to use
4. Separating the dataset into numeric dataset and non-numeric dataset
5. Summarising data

Interpretations:

1) There are 32561 rows and 15 columns.Out of which 9 are categorical and 6 are numerical
2) Columns are ['age', 'workclass', 'fnlwgt', 'education', 'education_num', 'marital_status', 'occupation', 'relationship', 'race', 'sex','capital_gain', 'capital_loss', 'hours_per_week', 'native_country', 'target'] present in dataset
3) 3 columns that have missing values. we can see that Occupation, workclass and native_country columns have few missing values
4) There is no strong correlation between variables
5) Almost 75% of people might get below 50K of salary.



6) Feature engineering
   a) Replacing missing values with mode
   b) Converting categorical variables into numerical ones
7) Feature Selection and Test-train split
   a) Splitting the training and testing set from both train and test datasets.
8) Model Fitting

   a) Here,we use Logistic Regression as dependant variable(Target) is discrete.

b) Model:
   i) Output:0/1
   ii) Hypothesis => Z = WX + B
   iii) Activation function => Sigmoid (0,1)
   iv) Decision boundary => threshold = 0.5 (1 if y >0.5, 0 if y <0.5)

9) Predicted the class 1 or 0
10) RMSE is 0.44
11) Confusion matrix is build based on test set and predicted test set. Classification report is generated. Accuracy is around 80%.