

**ANL252**

**Python for Data Analytics**

**End-of-Course Assignment (ECA)**

**July Semester 2022**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **Jeffrey Lim Dao Fong** | **Y2111639** |

**Tutorial Group : T09**

**Instructor’s Name : Mr Kumar Munish**

**Submission Date : 5 Sept 2022**

**Q1)**

|  |  |
| --- | --- |
| In [1] | # Importing libraries  import numpy as np  import pandas as pd  import matplotlib.pyplot as plt  import seaborn as sns  from sklearn.linear\_model import LinearRegression  # Reading in dataset  df = pd.read\_csv("ECA\_data.csv")  df.head() |

**Categorical variables**

From the data, (1) Rating (2) Gender (3) Education (4) Marital and (5) S(n) are categorical variables.

**Numerical Variables**

Whereas (1) Limit (2) Balance (3) Income (5) Age (6) B(n) (7) R(n) are numerical variables.

**Q2)**

|  |  |
| --- | --- |
| In [2] | # 1st Task is to remove any white spaces before & after the column names  df.columns = df.columns.str.strip()  df.head() |
| In [3] | # 2nd Task is to check the data types and observe that they match the intended data types  print(f"Shape of the dataset = {df.shape} \n")  print(f"Data types are below where 'object' indicates a string type: ")  print(df.dtypes) |
| In [4] | # we observe that 'R3' is indicated as an 'object' due to some values  # having the '$' and ',' sign  df['R3'] = df['R3'].str.strip('$').replace(',','', regex=True).astype(str).astype(int)  print(df['R3'].dtypes) |
| In [5] | # 3rd task is to check the data for missing values  print(f"Number of missing values for each column:")  print(df.isnull().sum()) |
| In [6] | # since we have an abundance of data (18769 data), we will drop all rows with missing values  df.dropna(inplace=True)  print(f"Number of missing values for each column:")  print(df.isnull().sum()) |
| In [7] | # 4th task is to check for any weird values in age  df['AGE'].value\_counts() |
| In [8] | # we see that there are 5 counts of '-1' and '199' each  # we shall remove them as well  df.drop(df[df['AGE'] == 199].index, inplace = True)  df.drop(df[df['AGE'] == -1].index, inplace = True)  df['AGE'].value\_counts() |

**Q3)**

|  |  |
| --- | --- |
| In [9] | plt.figure(figsize = (17,12))  fig\_1 = sns.boxplot(x ='GENDER', y ='INCOME', data = df)  plt.title('Figure 1: Boxplot of GENDER by INCOME', fontsize = 15)  plt.xlabel('GENDER', fontsize = 10)  plt.ylabel('INCOME', fontsize = 10)  plt.show() |
| Out [9] |  |

From Figure 1, we can see that the median, IQR and minimum for males and females are almost the same. However, when we further examine the outliers, we can conclude the male has a slightly higher outlier maximum as compared to the females. The males have a higher maximum value according to the boxplot as compared to the female as well.

|  |  |
| --- | --- |
| In [10] | plt.figure(figsize = (18,12))  fig\_2 = sns.scatterplot(x ='INCOME', y ='BALANCE', hue='GENDER', data = df)  plt.title('Figure 2: Scatterplot of INCOME by BALANCE', fontsize = 15)  plt.xlabel('INCOME', fontsize = 10)  plt.ylabel('BALANCE', fontsize = 10)  plt.show() |
| Out [10] |  |

From Figure 2, we can observe that the distribution is somewhat linearly distributed for both male and female. As the income increases, the balance that they have would increase as well. It is also fair to say that there is a fair number of people who have a higher income but have no savings as well. This indicates that they do not save the income that they generate.

|  |  |
| --- | --- |
| In [11] | plt.figure(figsize = (17,12))  fig\_3 = sns.barplot(x ='EDUCATION', y ='INCOME', hue='GENDER', data = df)  plt.title('Figure 3: Boxplot of EDUCATION by INCOME', fontsize = 15)  plt.xlabel('EDUCATION', fontsize = 10)  plt.ylabel('INCOME', fontsize = 10)  plt.show() |
| Out [11] |  |

From Figure 3, The data suggests that those that have a postgraduate qualification has the highest income. Qualifications stated under 'Other' has the 2nd highest income followed by tertiary and high school. Also, females, in all education level other than those having postgraduate qualifications, have higher income than males.

|  |  |
| --- | --- |
| In [12] | plt.figure(figsize = (18,12))  fig\_4 = sns.scatterplot(x ='INCOME', y ='LIMIT', hue='GENDER', data = df)  plt.title('Figure 4: Scatterplot of INCOME by LIMIT', fontsize = 15)  plt.xlabel('INCOME', fontsize = 10)  plt.ylabel('LIMIT', fontsize = 10)  plt.show() |
| Out [12] |  |

From Figure 4, the data suggests that the limit is directly proportionate to the income they have. This applies for both male and female.

|  |  |
| --- | --- |
| In [13] | plt.figure(figsize = (17,15))  fig\_5 = sns.boxplot(x ='MARITAL', y ='INCOME', hue='GENDER', data = df)  plt.title('Figure 5: Boxplot of MARITAL by INCOME', fontsize = 15)  plt.xlabel('MARITAL', fontsize = 10)  plt.ylabel('INCOME', fontsize = 10)  plt.show() |
| Out [13] |  |

From Figure 5, we can observe that the average single man makes more income than the average single woman. They also have a higher income than men who are married or categorised as 'others'. Women on the other hand, on average, have the same income as when they are single or married. However, women who are categorised as 'others' has lesser income than those single or married.

**Q4)**

|  |  |
| --- | --- |
| In [14] | X = df.loc[:, 'B1'].values.reshape(-1, 1) # values converts it into a numpy array  Y = df.loc[:, 'B2'].values.reshape(-1, 1) # -1 means that calculate the dimension of rows, but have 1 column  linear\_regressor = LinearRegression() # create object for the class  linear\_regressor.fit(X, Y) # perform linear regression  Y\_pred = linear\_regressor.predict(X) # make predictions |
| In [15] | plt.scatter(X, Y)  plt.plot(X, Y\_pred, color='red')  plt.title('Figure 6: B1 VS B2', fontsize = 10)  plt.show() |

Using the data, we have plotted B1 against B2 to find the linear regression model. In Figure 6, we can see the red line which describes the regression line.

**Q5)**

|  |  |
| --- | --- |
| In [16] | print(f'The regression equation is y = {linear\_regressor.coef\_}x + {linear\_regressor.intercept\_}.') |

The regression equation is y = [[0.91800757]]x + [2284.93864928].

Key insights obtained from Q4

1) The key insights obtained from the results in Question 4 are as follows

2) There is linear relationship between B1 and B2

3) There is a positive correleation from the independent variables (B1) and the dependent variables in (B2)

4) In the regression model, it is evident that there are more residues on the lower end of the model and outliers (i.e., between 300000 and 20000) this indicates that the data is not evenly distributed.