

**ANL252 : ECA**

**Ng Rui Feng Jonathan**

**SUSS PI number: Y2210848**

**Date:5/9/2022**

**Question one**

**List the categorical and numeric variables in the dataset.**

The categorical variables in the dataset include education and marital whereas the numerical values include ID, LIMIT, BALANCE, INCOME, RATING, GENDER, EDUCATION, MARITAL, AGE, S1, S2, S3, S4, S5, B1, B2, B3, B4, B5, R1, R2, R4, and R5.

**Question two**

**Conduct four data pre-processing tasks for the analysis of the data, explaining results obtained.**

1. **Data quality assessment.** I conducted quality assessment on the data to check the various data types, the number of observations (number of columns and rows), first five columns and last five columns of the dataset. There are 18,769 observations and
2. **Descriptive statistics evaluation.** Ialso conducted some descriptive tasks in the data, where I obtained the descriptive statistics of each column. The descriptive statistics include count, mean, standard deviation, minimum value, maximum value, median, 1st and 3rd quartiles of the data for each column name. These statistics are crucial as a data pre-processing task since they help in gaining a deeper insight of the data in the various columns.
3. **Data cleaning.** I checked the missing values in the dataset since working with missing values for analysis can give incorrect analysis. In the dataset, there were 51 missing values, where 13 of them were from ‘Education’ column while 38 values were from ‘marital’.
4. **Data transformation.** I transformed the missing values by adding 0 value into all the missing values. This would help in avoiding confusing and incorrect analysis during the next steps of the project.

**Question three**

**Articulate five relevant insights of the data, with supporting visualization for each insight.**

1. From the histogram of age of the customers, it is evident that most of customers’ age ranges between 25 and 45 years. The customers’ attendance decreases with increasing in age as per the data illustrated in the histogram.
2. A plot of a histogram is also used to display the balances for the customers in the credit facility. From the histogram, we can deduce that most of the customers have low balances or no balances ranging from 0 to 40000. This indicates that the credit facility does not have many customers holding their money, thus, the facility is not running at a loss at the moment.
3. The other histogram shows the distribution of the limit for the customers in the credit facility. From the graph, most of the customers have low limit for credit. This means that the customers cannot request high credit from the facility. Most of the customers have credit limit ranging from 100000 to 200000. However, there are some other customers that cannot access credit facilities from the facility, since they have a limit of 0. Customers with a limit of above 400000 are quite few, which means that the credit facility is able to manage the credits owed by their customers and avoid customers from defaulting their payments.
4. Using a correlation heat map plot, we can also plot a correlation between the variables LIMIT, BALANCE, INCOME, GENDER, and AGE. Through the correlation, we can understand how the variables are related to each other. Variables whose value is closer to 0 show that there is no linear trend between the variables. Those closer to a correlation of 1 indicate that they are more positively correlated (as one variable increases, the other variable increases also). Variables closer to -1 show that as one variable increases, the other variable decreases. From the heat map, we can observe that income has a higher correlation with limit, meaning that those with higher income levels have high limit. Age and income have a high positive correlation as well, implying that the higher the income, the older a customer. Income and balance also have a slightly higher correlation. On the other hand, Age and gender have negative correlation, meaning that age and gender do to have any correlation.
5. The scatter plots show the relationships between the various variables. Variables such as Income and limit have a linear relationship as shown in the scatter plot. This means that a customer’s limit is determined by their income levels. Age and income also show a smooth linear relationship, implying that the older an individual is, the higher their income. Balance and income also show a slightly fair relationship.

**Question 4**

**Perform linear regression modelling to predict variable, B1, explaining the approach taken, including any further data pre-processing.**

The approach taken in performing the linear regression modelling to predict variable B1 was to first of all, import the required packages/libraries for regression modelling. Secondly I defined the linear model so as to use the defined model in the other stages. Another step that I undertook was to add a constant, which would essentially help in getting the intercept. The next step was to set the variable to be predicted (B1) and split the dataset into training and testing datasets, with a ratio of 4:6. Thereafter, I created a linear regression object. Reshaping was also part of the process, which is a pre-processing task in building the regression model. I predicted the model using variable B1 as required and then plotted a regression line of the same. I also fitted the regression line using OLS and printed the parameters. Lastly, I displayed a summary of the results as shown in the output.

**Question 5**

**State the linear regression equation and explain key insights from the results obtained in question 4.**

The linear regression equation

**y= mx + c,**

Where **m** is the coefficient of **x**, and **c** is the intercept.

Equation would be B1= 1976.53 – 5.29\*Balance

(The above equation is rounded off to 2 d.p.)

R-squared value is 0.907, implying that 90.7% of the variance in B1 can be explained by Balance or the regression model.

The coefficient for B1 is 1976.53 (2 d.p.), P-value is close to zero, showing that the coefficient is statistically significant.