

**ANL252 Python for Data Analytics**

**July 2022 Presentation**

**ECA**

|  |  |
| --- | --- |
| TG: | 09 |
| Student Name: | Shawn Ng Zhen Xiang (Q2172585) |

**Question 1**

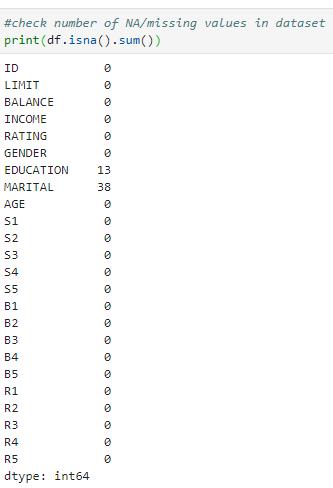
|  |  |
| --- | --- |
| **Categorial Values** | **Numerical Values** |
| ID | Limit |
| Gender | Balance |
| Education | Income |
| Marital | Age |
| S(n) (Repayment Status) | B(n) (Billable Amount) |
| Rating (ordinal) | R(n) (Repayment Amount) |

**Question 2**

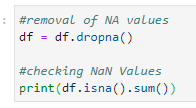
1. Cleaning Data - Check for missing data, Dropping outlying data, Removal of duplicates
2. Readjusting datatypes
3. Binning data values
4. Reducing number of categories
5. Train test split and standardization(however this would be done later for regression)

**Cleaning Data**

1. There were a number of NA values for education and marital values:



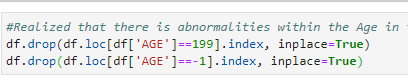
Dropping NaN Values so it does not skew the data

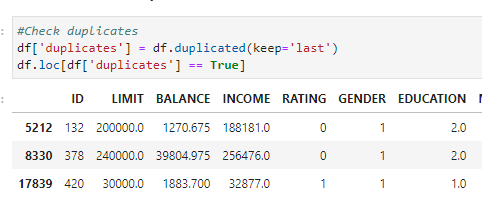


1. Abnormalities with age values with min max data



Removal of abnormal data age 199 and -1

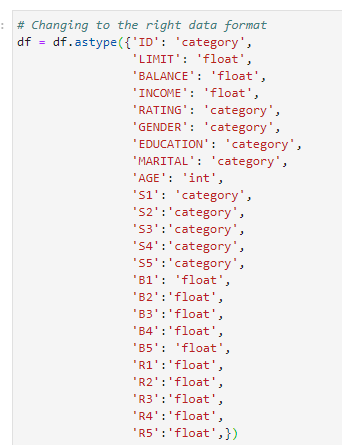


1. Removal of duplicate rows from duplicate checks  
   



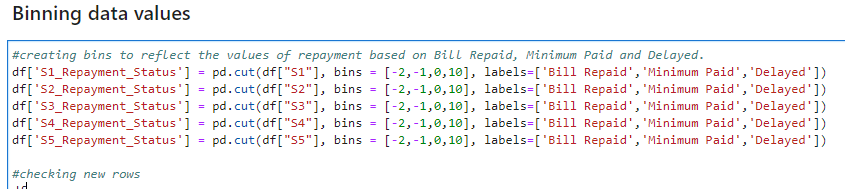
After the initial checks to clean data we are able to convert data types into different forms.

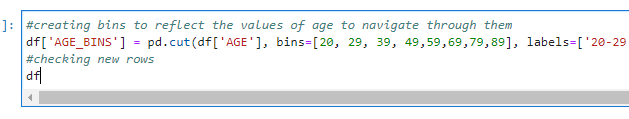
**Readjusting data types**

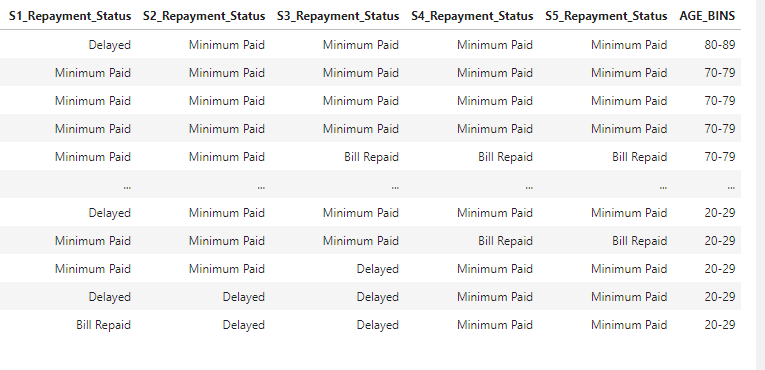


Data types are adjusted to the correct sort of values to be able to process them as graphs, categories, instead of strings and objects which would not allow you to do calculations and processing.

**Binning data values**







Reducing category types of S1, S2, S3, S4, S5 into different categories to be able to understand if users simply delated payment, paid minimal amounts or entire bill repaid.

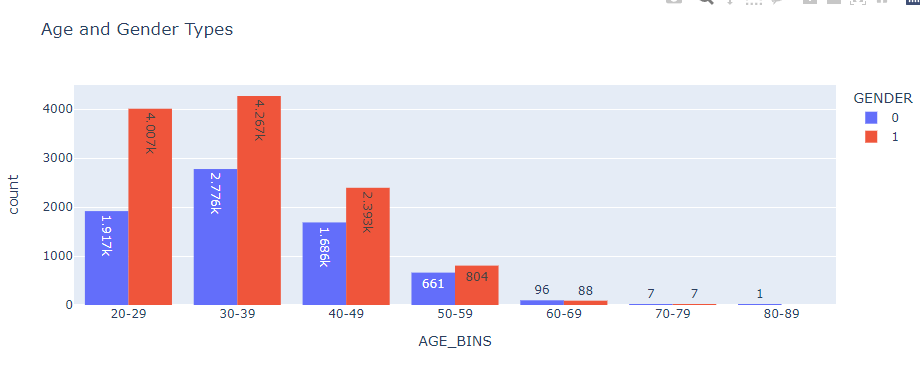
Also reduced the ages into bins so that we can do histograms easier to understand the different age groups and what their patterns are.

**Reducing categories**

****

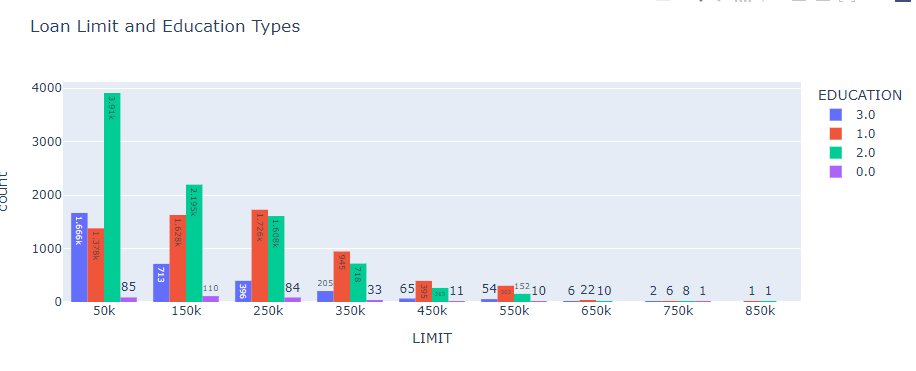
Reduced the categories of marital, S1-S5. Reason being is to simplify the data set, with different marital data being split into simply single or married, as well as the payment amounts categories into simple values such as bill repaid, minimum sum paid and delayed instead of having a number of delayed months.

**Question 3**

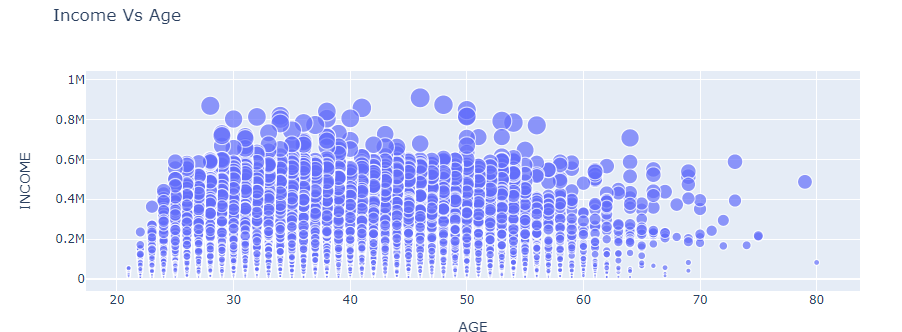
****

Gender 0 = Male, Gender 1 = Female.

Based on the data set, it seems that there are more females in the data set compared to males, and majority of the data belonging to younger audiences, particularly from the 20s-50s.

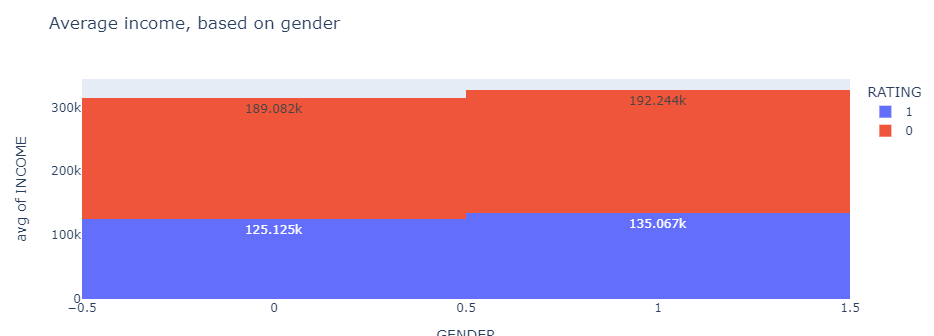
****

From the Data above (0: Others, 1: Postgraduate, 2: Tertiary, 3: High School), we can determine that majority of the loan limits change, with the smaller amounts having a higher number of tertiary graduates, and larger loan limits leaning towards post graduates i.e. 250k-550k.

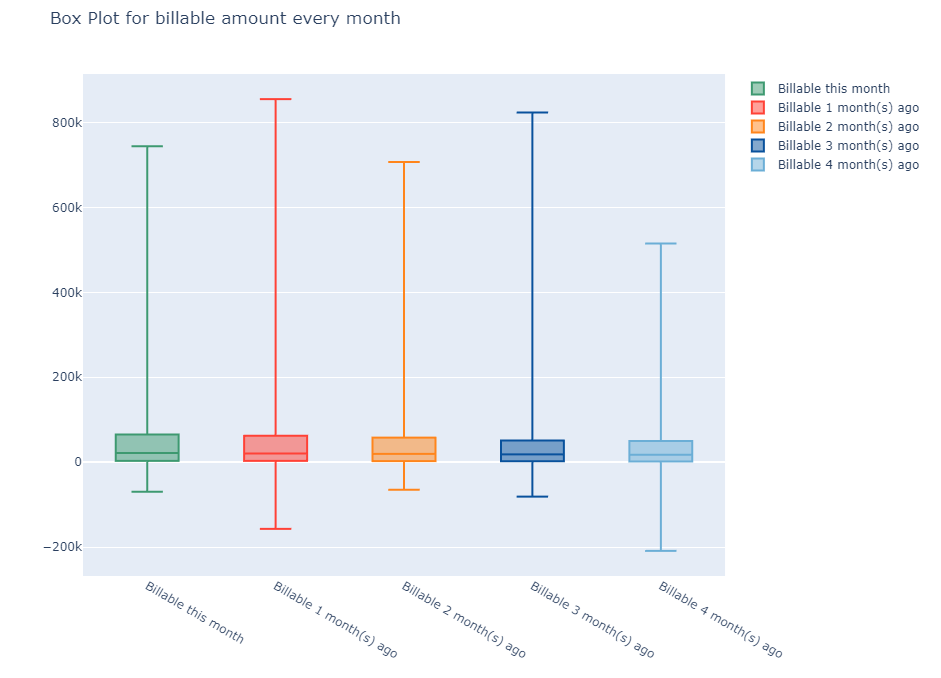
****

Income Vs Age

Scatter plot to determine the income VS the age of customers, the size of the plots are set by the limit of which the customers can borrow from the bank, showing data that the higher their income the higher the limit they could potentially borrow. The general age that has large bubble clusters are somewhere between the 30s to 60s, where most of the larger bubble clusters are available, it also shows us that the later the age the fewer the number of customers and it seems to have higher income customers from 60-70.

****

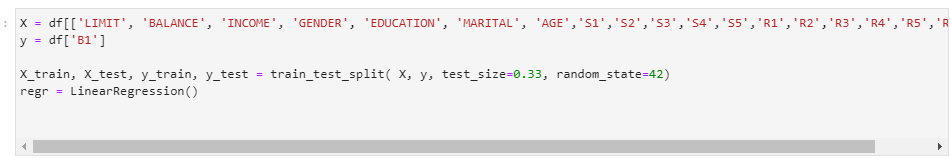
Comparing the average income of male to female with (0: Male, 1: Female), where the average income of females seem to be higher than which of the males. The general ratings seem quite evenly split between the two, with ratings (0: Good, 1: Bad), with more customers rating the credit facility good with those having lower income rating it worse than those with higher income.

****

From the graph we are able to determine that across the month the general median, interquartile ranges of billable amounts is about the same, what varies is the max amount billable and the minimal billable across the different months, starting with the current month to the billable amounts 4 months ago it seems that there are users who have seem to borrowed more recently. This may not be a good thing if the users are not able to pay off the debts which may end up as bad debts, it may be more worth while to increase the number of users within the quartiles so as to not over leverage with fewer customers that are borrowing larger amounts.

**Question 4**

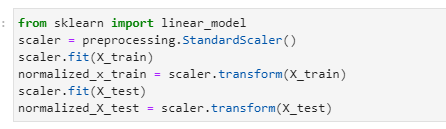
The approach is to conduct a train test split with the data, we want to be able to evaluate the performance of the machine learning algorithm when we use a linear regression model so as to test if the prediction of the data is accurate.



Here we choose to split the test size to .33, keeping two thirds for training and the final third for testing.

Y would be the dependent variable which is B1 in this case.

We then use the standard scaler which removes the mean and scales each feature/variable to unit variance. The reason why is because the data set differs greatly between its ranges, and to effectively compare each variable we need to standardize the data, so as to determine if the regression based on the variables affect the dependent variable.





The above shows the data being fitted into the regression model, once that is done, y\_pred is used to predict the values of B1 with the regression model.

**Question 5**

The linear regression equation is

Y = 50871.53 + 7.74306513e+02(LIMIT) + 3.35364162e+04 (BALANCE) -8.23288432e+02(INCOME) -5.79994299e+01(GENDER) -4.79028548e+01(EDUCATION) + 3.98690030e+02(MARITAL) + 2.82338220e+02(AGE) -1.36848753e+03(S1) + 1.75053627e+03(S2) -2.61058262e+02(S3) -2.13899717e+01(S4) -1.10010920e+02(S5) + 4.05547347e+04(B2) -1.66638984e+03(B3) -1.54119845e+03(B4) + 2.46409207e+03(B5) -7.14775441e+03(R1) + 3.05643623e+03 (R2) 8.23620236e+02(R3) -9.94278458e+02(R4) + 2.24653409e+02(R5) -3.57756537e+01(Rating)

Where the constant is the Y intercept, followed by the coefficients above from the regression model.

The regression score of the test data is higher than the regression score of the training data, and both are close to 1 which determines a good level of accuracy. The standard deviation across the values are also 1. With R square the coefficient of determination being at .95 which is close to 1 meaning that the data above is a good fit of the variance of independent variables to the dependent variable. With that we are able to predict the values of Y with the estimated values for X based on the above equation.

The positive values above directly add to the potential of the Y value however the negative values above would reduce the Y value that would possibly be predicted based on the weightage of dependency of the independent variables to the dependent variable.