Team details :

Section : F

Name and SRN :

Pramatha Gajanan Bhat PES1UG19CS339

Rajeshwari R PES1UG19CS375

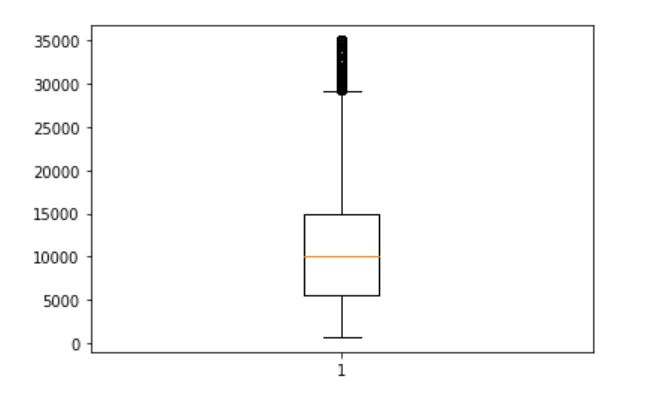
Ramya C PES1UG19CS379

Rushab Prakash Kulkarni PES1UG19CS399

Abstract :

This project is based on the banking sector domain. The purpose of this project is to automate loan eligibility process based on customer profile provided through application form. The chosen dataset is stored in an excel file. Problem statement is the exploratory data analysis of loan dataset involving data cleaning, data preparation, graph visualization of data using python libraries (numpy, pandas, matplotlib, seaborn, sklearn ) and hypothesis testing on loan amount. We define the dataset, it’s columns for data analysis. Categorical and Numerical columns are found. We proceed to data cleaning, where we remove null values and fill the empty box with median, if it’s left skewed. With mean if it’s normal. Then the data is visualized using various graphs (Histogram, Boxplot, Scatterplot, Bar Chart).

Histogram, boxplot, scatterplot for numerical columns and bar chart for categorical columns. In the next step, unwanted outliers are to be filtered for numerical columns. Note, only those outliers are unwanted and removed, whose removal would not affect the dataset. If the number of outliers is significantly large, such outliers are not considered as unwanted and are retained. Values of each numerical column is normalized, to remove redundant data from table in order to improve storage efficiency, scalability and data integrity. Hypothesis testing is done to test specific predictions. We have used p value approach for hypothesis testing, also z test and population proportion. This is quite the summary of our mini project.

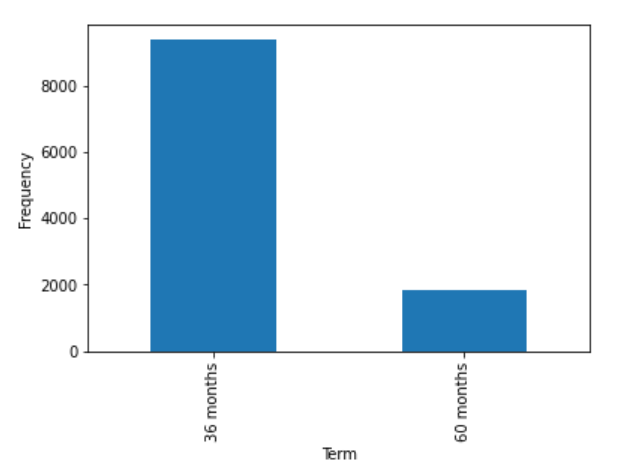
Exploratory Data Analysis :

There are 15 columns including numerical and categorical.

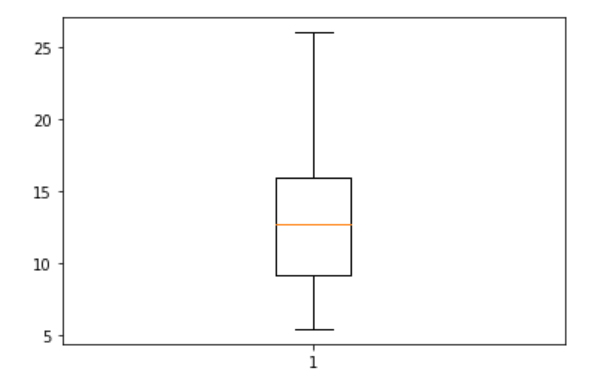
1. loan\_amnt :

Amount of loan applied for, by the borrower. From the boxplot, minimum is at around 1 and maximum at around 29000. Q1 is around 5500, Median is around 10500 and Q3 around 15000. There are considerable number of unwanted outliers.

2. term :

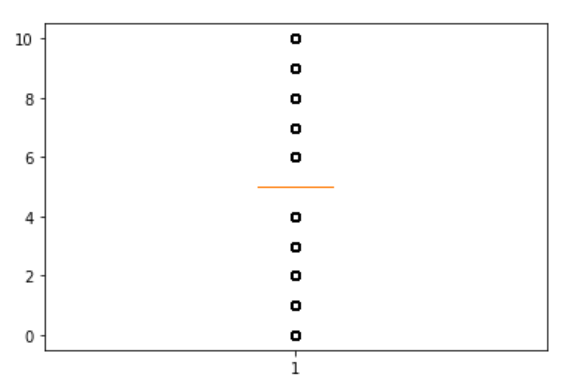
Number of payments on the loan. Either 36 or 60 months 

From the bar chart, it can be seen that frequency of 36 months term is more than that of 60 months term

3. int\_rate :

Interest rate on the loan

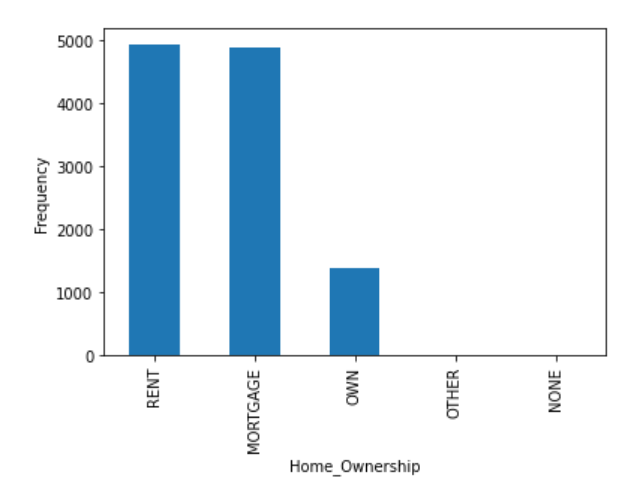
Q1 is at 9, Median is at around 12, Q3 is at 16.5. There are no outliers for this distribution.



4. emp\_length :

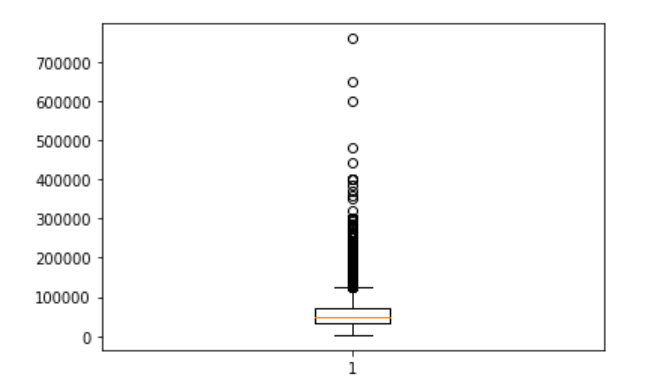
Employment length of the applicant

There are quite a number of outliers, but since they are widely distributed and form major part of the graph, they are to be considered. Median is at 5



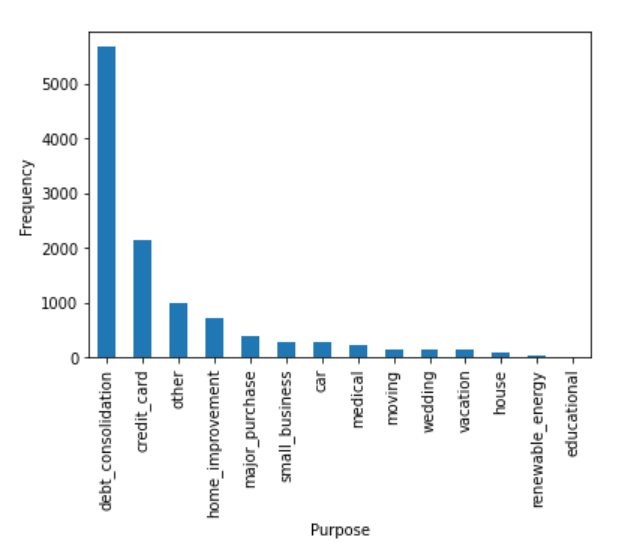
5. home\_ownership :

Around 4900 people have rented home , 4850 have mortgaged their ownership, and 1200 people own a house.



6. annual\_inc :

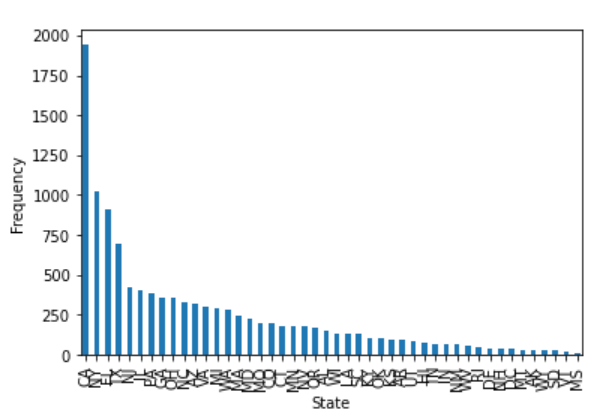
Reported annual income of the borrower. From the boxplot, median is seen at 50,000. A lot of people have annual income more than 1 lakh



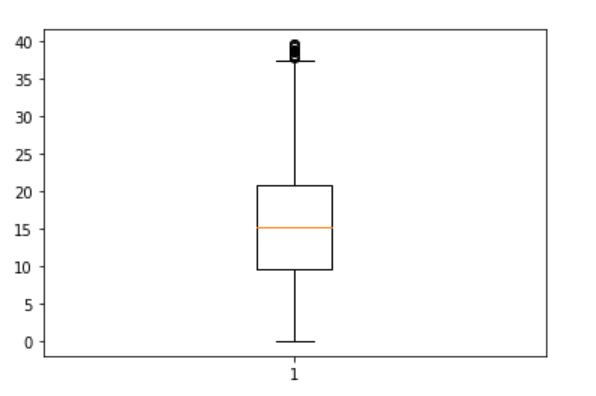
7. purpose :

A category provided by borrower for loan request.

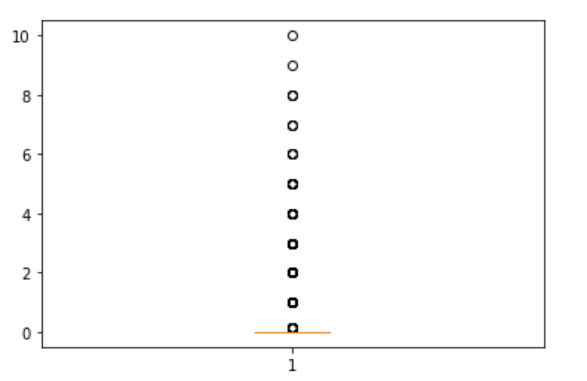
This is a bar chart for the distribution.

8. addr\_state :

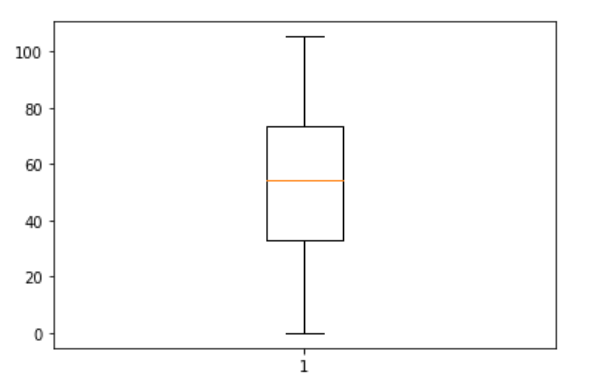
State provided by borrower in loan application.

9. dti :

A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan , divided by the borrowers self reported monthly income. Q1 is 11, Median is 15.5, Q3 is 21.

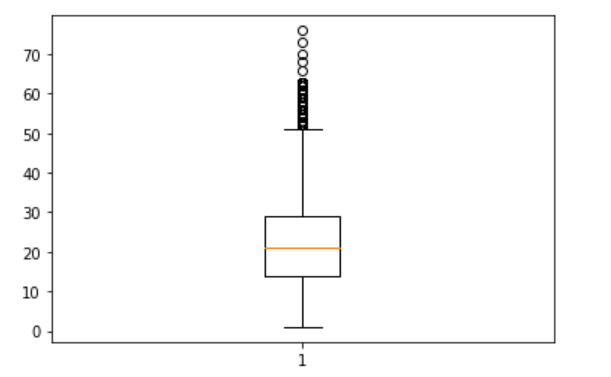
10. delinq\_2yrs :

The number of 30+ days past-due incidences of delinquency in the borrowers credit file for the past 2 years. Median can be seen at 0.1

11. revol\_util :

revolving line utilization rate/ the amount of credit the borrower is using relative to all available revolving credit.

Median is at 56, Q1 at 37, Q3 at 76. No outliers.

12. Total\_acc :

The total number of credit lines currently in borrowers credit file. Median is at 21, Q1 at 14, Q3 at 28. There are quite a number of outliers.

Hypothesis Testing

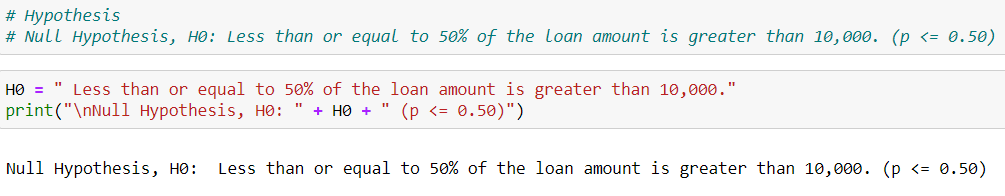
Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter.

The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

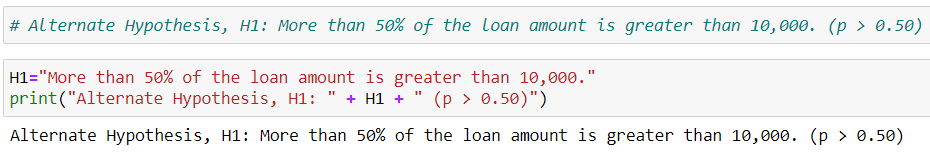
Here we have used the p-value approach for hypothesis testing

Since the data sample is large,we have used z-test.Here population proportion is used.

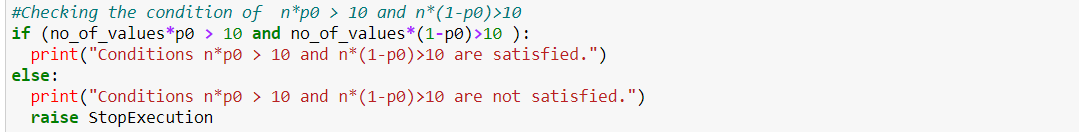
Null Hypothesis(H0): Less than or equal to 50% of the loan amount is greater than 10,000. (p <= 0.50).



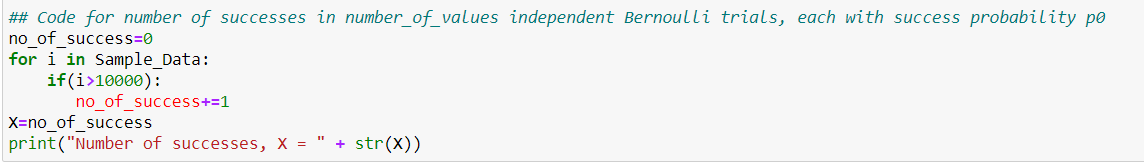
Alternate Hypothesis(H1):More than 50% of the loan amount is greater than 10,000. (p > 0.50).



Now check the condition of n\*p0 > 10 and n\*(1-p0)>10 is satisfied or not.



If satisfied:calculate the number of successes from the total length, independent Bernoulli trials each having a success probability of p0.

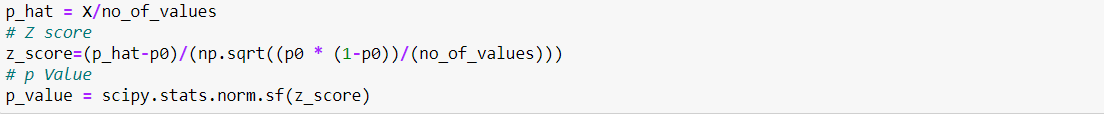


Further determine if the test is one tailed test or a two tailed test and allocate alpha value accordingly. The test is one tailed because Ha has only greater than condition.Hence alpha will be 0.05.

Now calculate p\_hat by using p\_hat=X/no\_of\_values where X is the number of successes.

Then find z\_score by using the formula z\_score=(p\_hat-p0)/(np.sqrt((p0 \* (1-p0))/(no\_of\_values)))

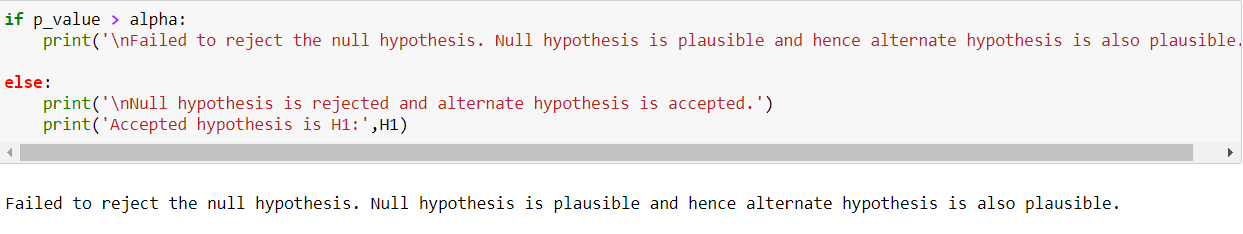
and further find the p\_value by the formula p\_value = scipy.stats.norm.sf(z\_score).



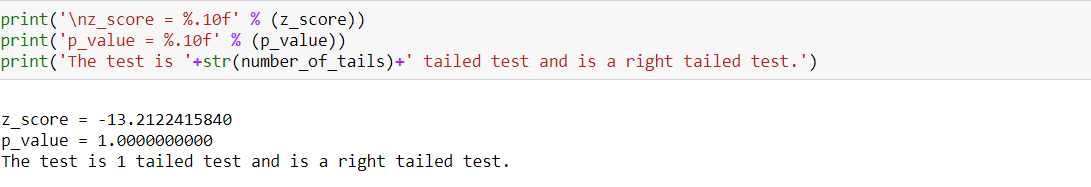
Now check whether the p\_value is greater than alpha or p\_value is less than alpha.

If p\_value is greater than alpha, then the null hypothesis cannot be rejected. Null hypothesis is plausible and hence alternate hypothesis is also plausible.

Else Null hypothesis is rejected and alternate hypothesis is accepted.



print z\_score ,p\_value and conclude about the test



Results and Discussion:

According to the problem statement, we concluded that the Null hypothesis is plausible and hence alternate hypothesis is also plausible.

we have used appropriate methods for data cleaning,Then we used various graphical representations to analyse the data set.Further we computed mean and variance for each column fused various methods to normalise the data and also used graphical analysis to check whether the data was normalised or not.

Then further we found out the correlation between all the parameters of the dataset by correlation coefficient