**Satvik Nagar 12017696**

**Lovely professional university**

**Problem Statement**

A person’s creditworthiness is often associated (conversely) with the likelihood they may default on loans. Here, we ask you to look at data on loan applicants, extract useful insights and guide a business towards decisions. We’re giving you anonymized data on about 1000 loan applications, along with a certain set of attributes about the applicant itself, and whether they were considered high risk. We’d like you to work your magic✨ on this.

1. How would you segment customers based on their risk (of default)

To segment customers based on their risk of default, we can use a combination of loan-specific and personal attributes that are likely to influence an applicant's ability to repay the loan. Some of the attributes that can be used to segment customers are:

Loan-specific attributes: loan amount, EMI rate, the purpose of the loan, number of existing loans at this bank, loan history

Personal attributes: age, gender, marital status, number of dependents, housing, years at current residence, employment status, savings account balance, balance in an existing bank account, whether the applicant has a co-applicant or guarantor

We can use clustering techniques to segment customers based on their risk of default. One approach could be to use k-means clustering to group customers into clusters based on their loan-specific and personal attributes. We can then assign each cluster a risk score based on the historical default rates of customers in that cluster.

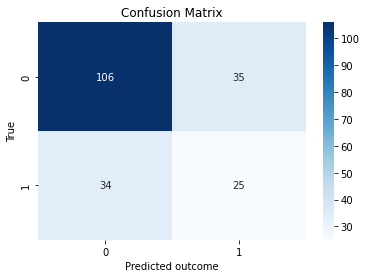
I have used the four attributes here for the cluster and to segment the customers

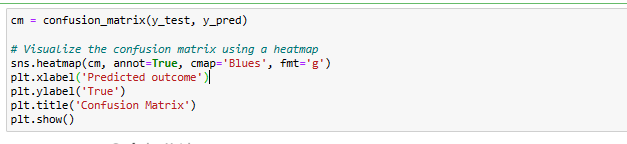
**'Primary\_applicant\_age\_in\_years','Principal\_loan\_amount','Number\_of\_existing\_loans\_at\_this\_bank','high\_risk\_applicant'**

Here’s the code for the cluster we made and the visualization of those cluster  


2. Which of these segments / sub-segments would you propose be approved?

### To determine which segments/sub-segments to approve, we need to balance the risk of default with the potential revenue from each loan. We can use the cost matrix provided to estimate the cost of approving or rejecting a loan application.





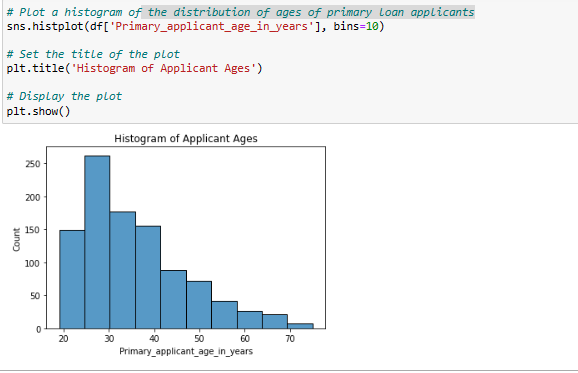
### In general, we should aim to approve loan applications from segments/sub-segments with a lower risk of default and a higher potential revenue. However, we should also consider factors such as the current economic conditions, the overall risk appetite of the organization, and the specific objectives of the loan program.

## 3. What other insights can you share about the general creditworthiness of these segments?

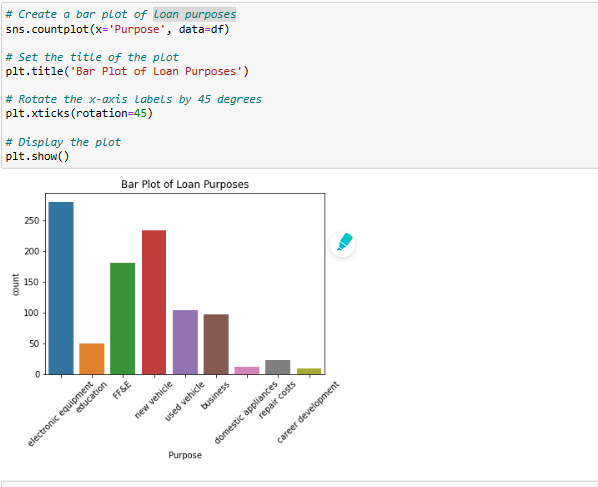
# We can use descriptive statistics and data visualization techniques to gain insights into the general creditworthiness of each segment. For example, we can calculate the average loan amount, EMI rate, and default rate for each segment. We can also plot histograms or box plots to visualize the distribution of loan-specific and personal attributes within each segment.

### Some possible insights that we can derive from the data are:

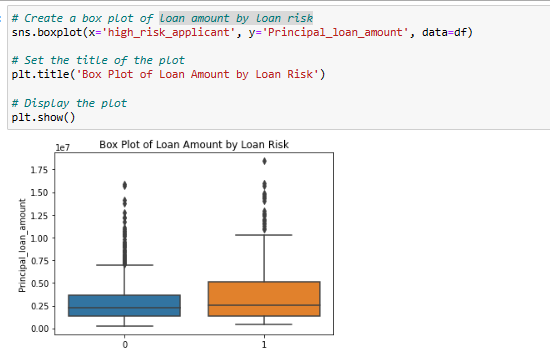
### the distribution of ages of primary loan applicants



### Customers with a higher number of existing loans at this bank tend to have a higher risk of default and what the purpose of the loan.



### loan amount by loan risk



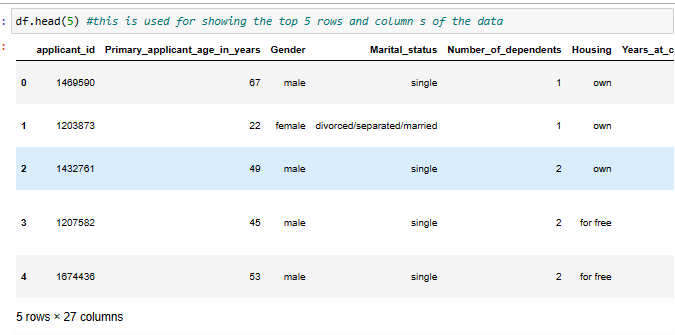
## 4. Tell us what your observations were on the data itself (completeness, skews) and how you would treat any anomalies (for eg - missing data)

Before analyzing the data, we need to check for completeness, consistency, and accuracy. Some observations on the data are:

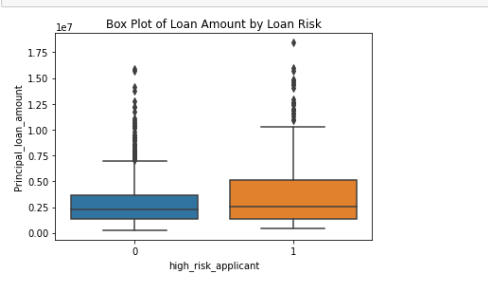
The applicant table has 1,000 rows and 14 columns, while the loan table has 1,000 rows and 11 columns. There is a one-to-one mapping between the applicant\_id in the applicant table and the applicant\_id in the loan table.

Some columns have missing data, such as Has\_been\_employed\_for\_at\_least, Has\_been\_employed\_for\_at\_most, and Savings\_account\_balance.

Some columns have values that are not consistent with the data dictionary, such as Foreign\_worker, which should be a string but has numeric values.

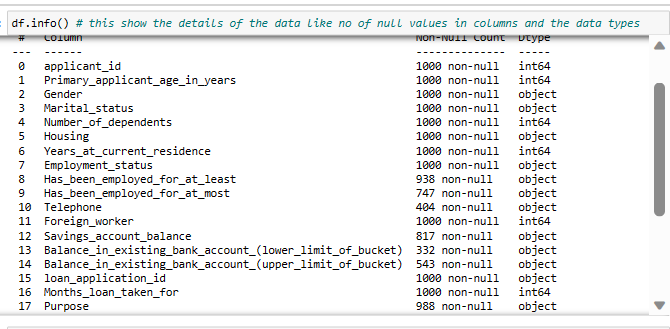


Some columns have values that are skewed or have outliers, such as Months\_loan\_taken\_for, Principal\_loan\_amount, and EMI\_rate\_in\_percentage\_of\_disposable\_income.



The above box plot shows the outliers the point that are outside the inter quartile range

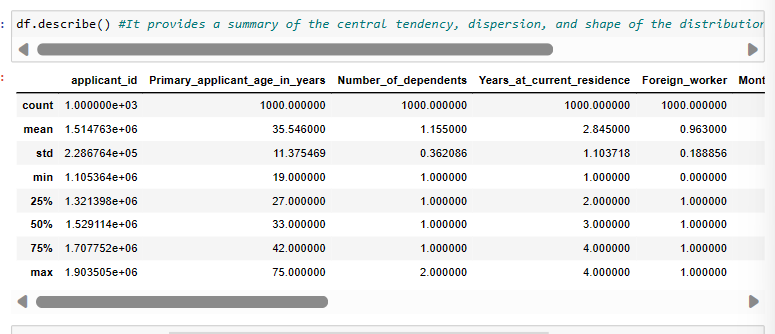
Here’s the all the information about he datatype null count present in the coloumns



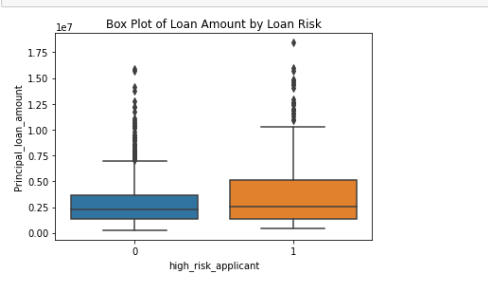
## To treat any anomalies in the data, we can use the following approaches:

### For missing data, we can use imputation techniques such as mean, median, or mode imputation, or use more advanced imputation techniques such as KNN imputation or multiple imputation. We need to assess the extent of missing data and the appropriateness of the imputation technique for each column.

Here below we have calculated the mean mode median for KNN imputation

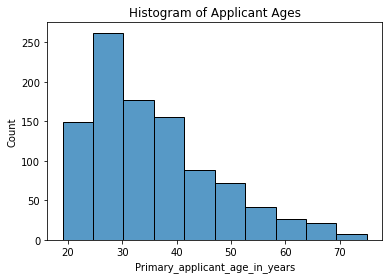


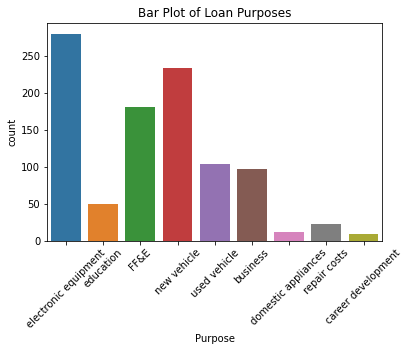
### For skewed or outlier data, we can use normalization or standardization techniques to scale the data. We can also use techniques such as winsorization or trimming to remove outliers. We need to consider the impact of these techniques on the analysis and the interpretation of the results.

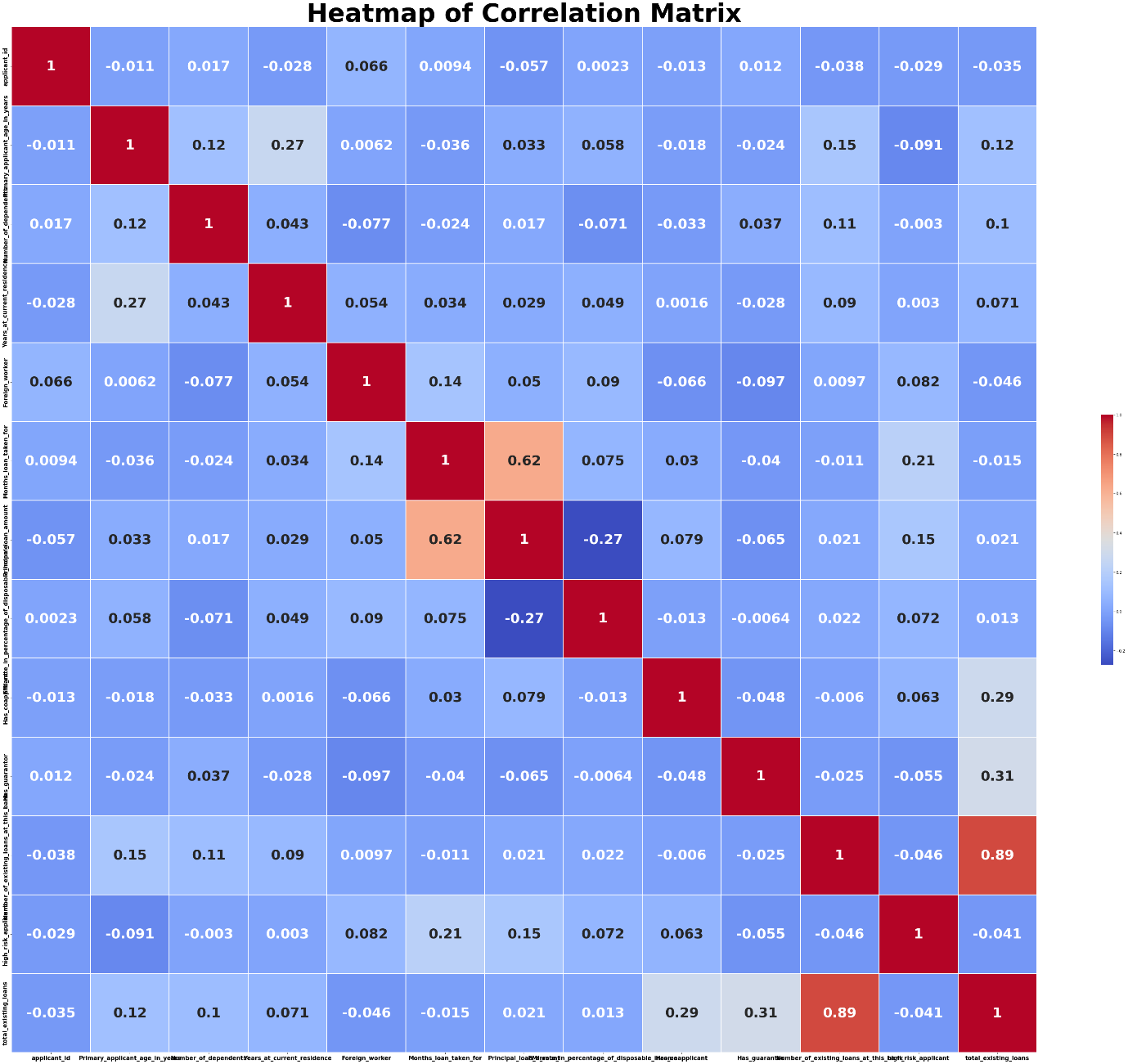


The above visualization show the outliers and remove

### We can also use exploratory data analysis (EDA) techniques such as visualization and statistical tests to identify anomalies and assess the quality of the data. EDA can also help us to identify relationships and patterns in the data that can guide further analysis.







The above show the correlation between all the columns in the data