CASE STUDY

- Rohitchandran
Ravichandran

INTRODUCTION

We will look into a comprehensive analysis of bank loan applications and explore the factors that influence loan approval decisions.

This study provides valuable insights into how financial institutions can optimize their lending processes while mitigating risks effectively.

As we all know, banks play a crucial role in the economy by providing financial assistance to individuals and businesses.

The process of granting loans is a critical aspect of a bank's operations and requires careful consideration of various factors.

Understanding the loan approval process and identifying key variables that influence these decisions is of utmost importance for banks to make informed and data-driven lending decisions.

Tech Stack : Microsoft Excel 2016

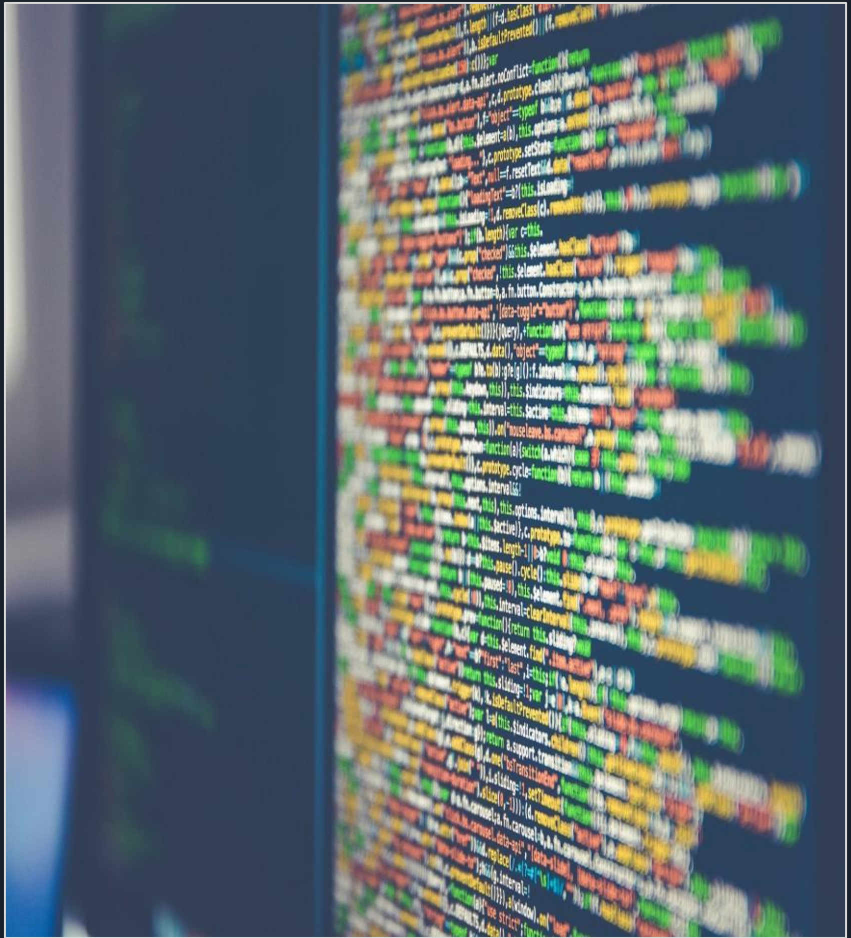
Objective

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

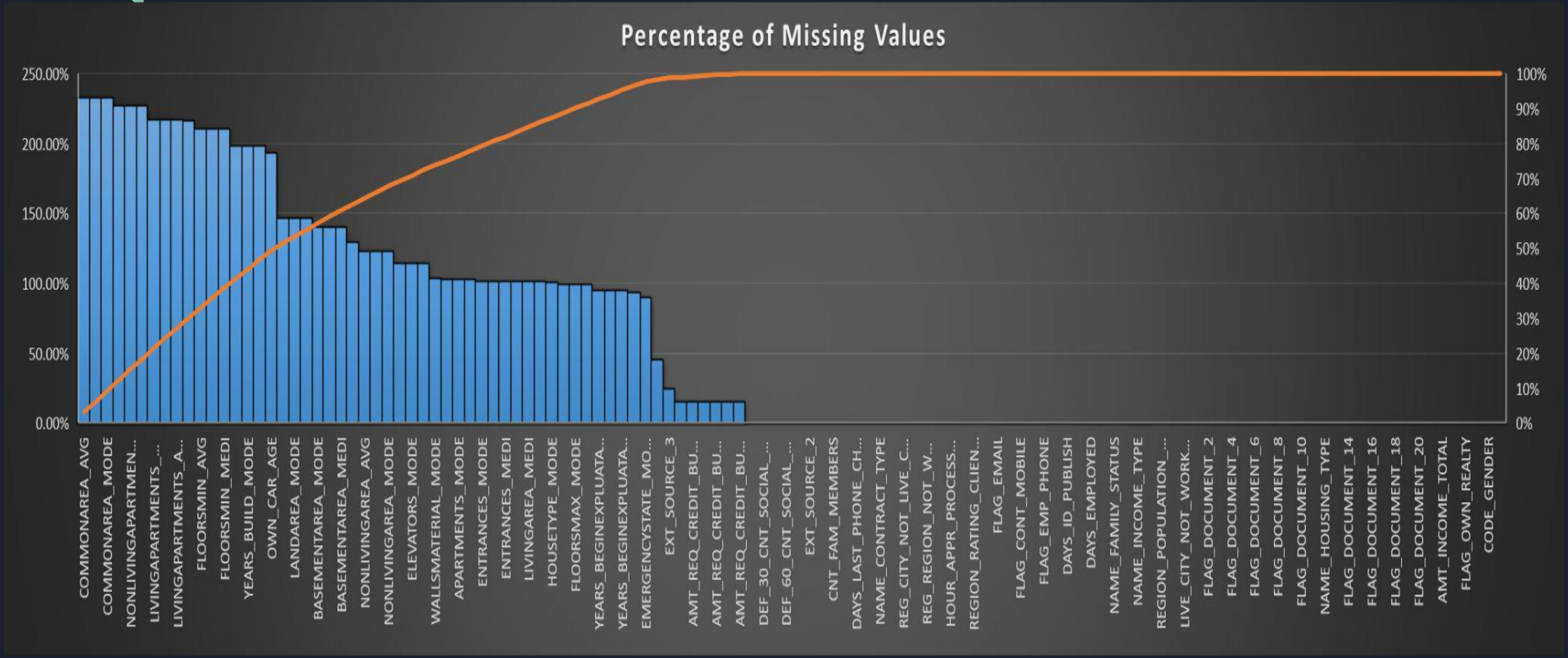
Data Cleaning

The application_data csv file had around 307512 records but the SK_ID_CURR column along with a few important columns had only 49999 records.

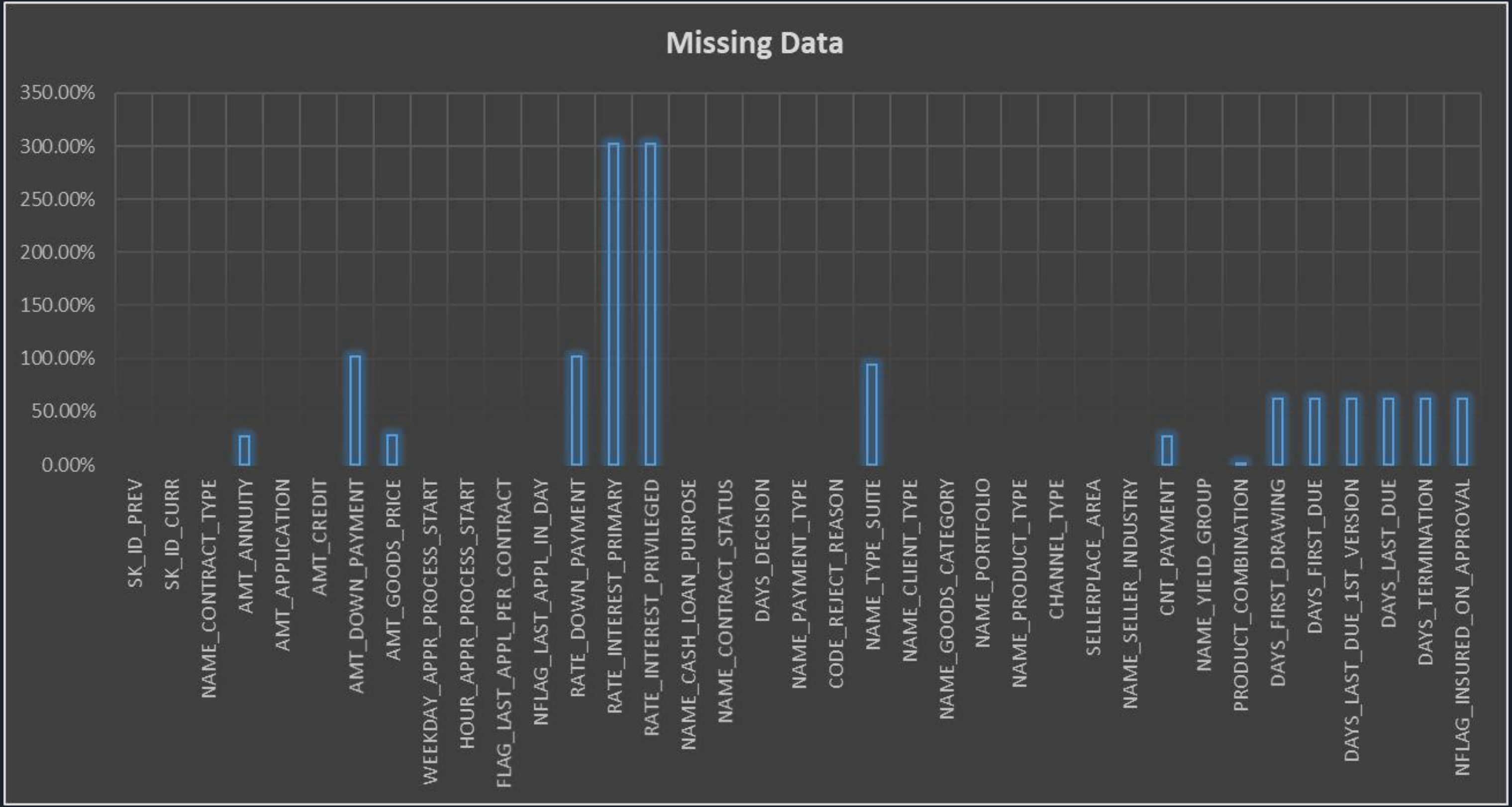
Without the id of the loan or the name_contract(identification of loan), The other columns are not relevant. Therefore i deleted all the records which didn’t have SK_ID_CURR and Name_Contract.



After dealing with missing values above 20% , I used median and average for imputation for numerical variables which has less than 20% blank data.



Doing the same procedure for previous_application.csv, we get this column chart that shows proportion of missing values for each variable.

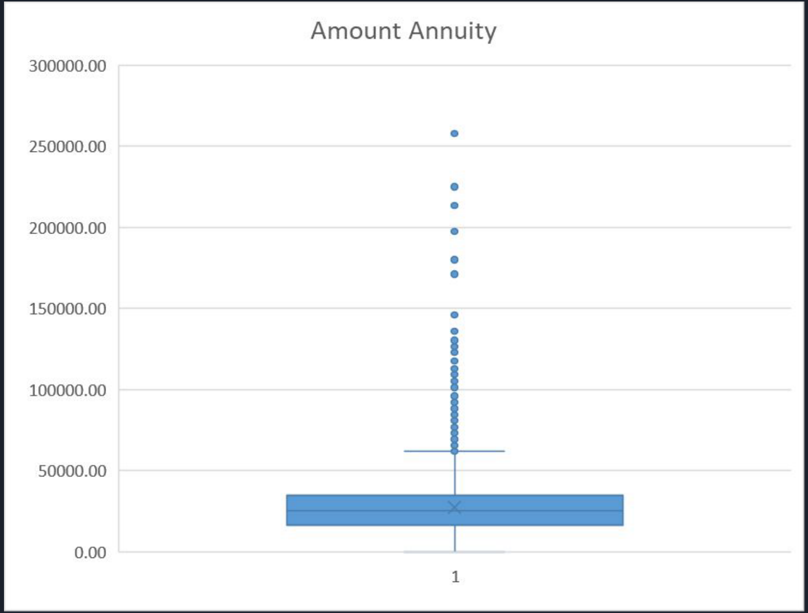


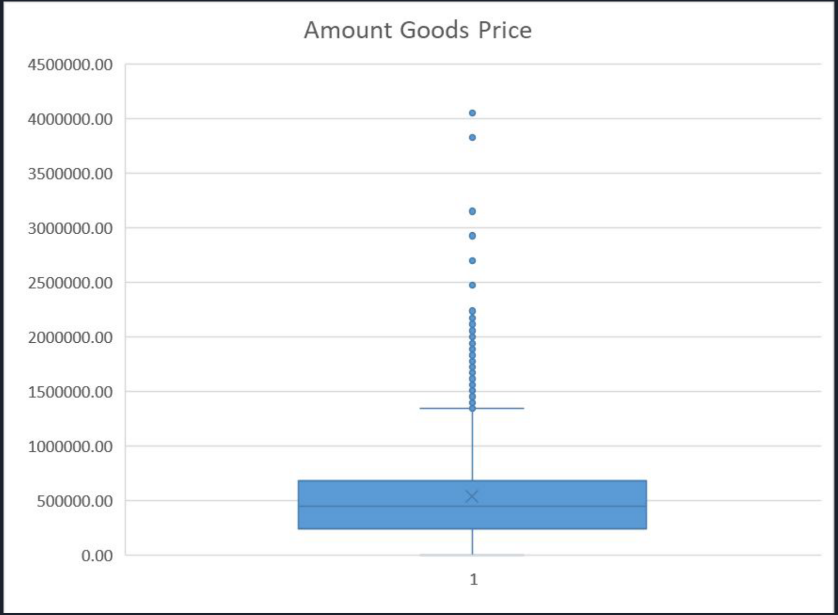
Identify Outliers in the Dataset

- Outliers are data points that deviate significantly from the rest of the data in a dataset.
- They are values that are unusually high or low compared to the majority of the data points.
- Also, outliers might be valid data points, so it's important to carefully consider their significance before making any decisions about their treatment.
- Outliers can only be found in numerical variables . Here are few columns which can contain outliers.
- AMT_INCOME_TOTAL , DAYS_EMPLOYED , AMT_CREDIT , AMT_ANNUITY, AMT_GOODS_PRICE , CNT_FAM_MEMBERS - These columns have potential outliers. (application_data.csv)

In both the amount annuity and amount goods price, we can see that there are a lot of outliers

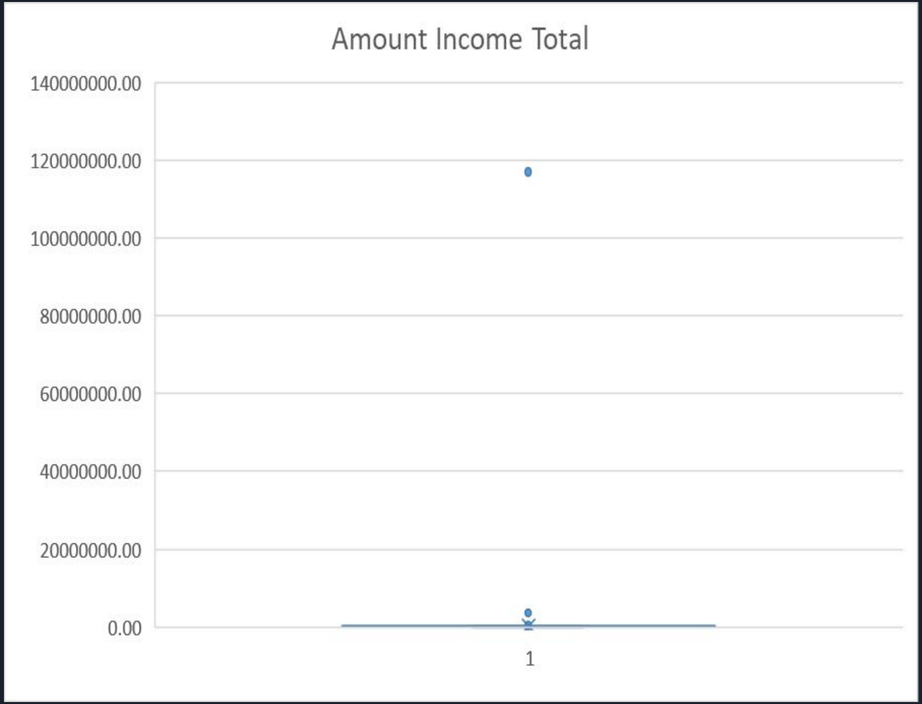
It shows that few people have presented goods valued more than 20 lakhs for getting their consumer loan.

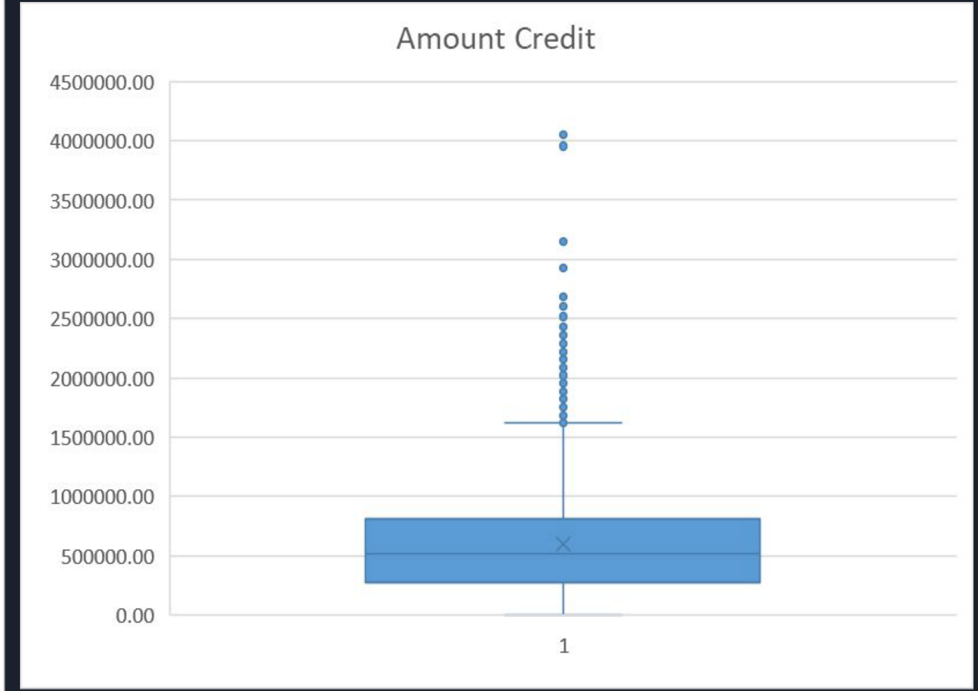




Amount income total has a big outlier which is likely to be an incorrect entry as a customer with over 10 crores of annual income wouldn’t be getting a loan for 40 lakhs or below.

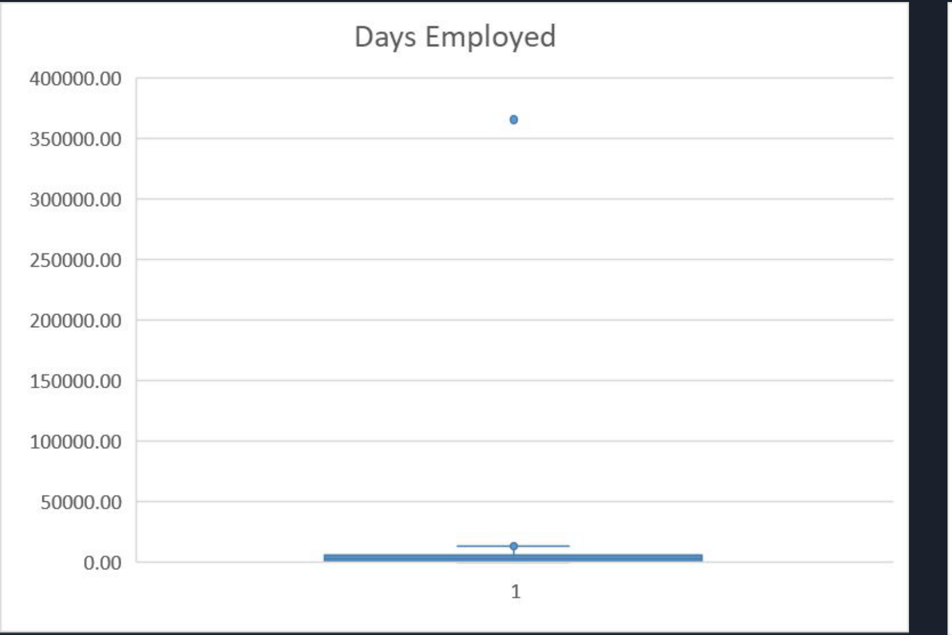
We can see that lot of people opted for loans above 15 lakhs which makes them outliers.

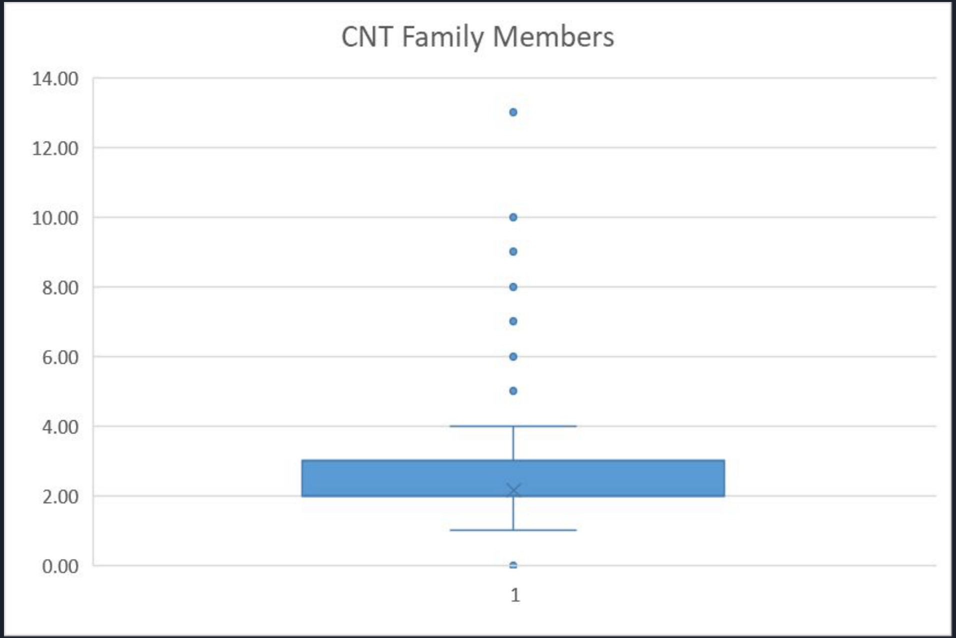




Days employed column has entries with 365243 days, that's roughly 1000 years which is impossible. Henceforth , there are invalid entries.

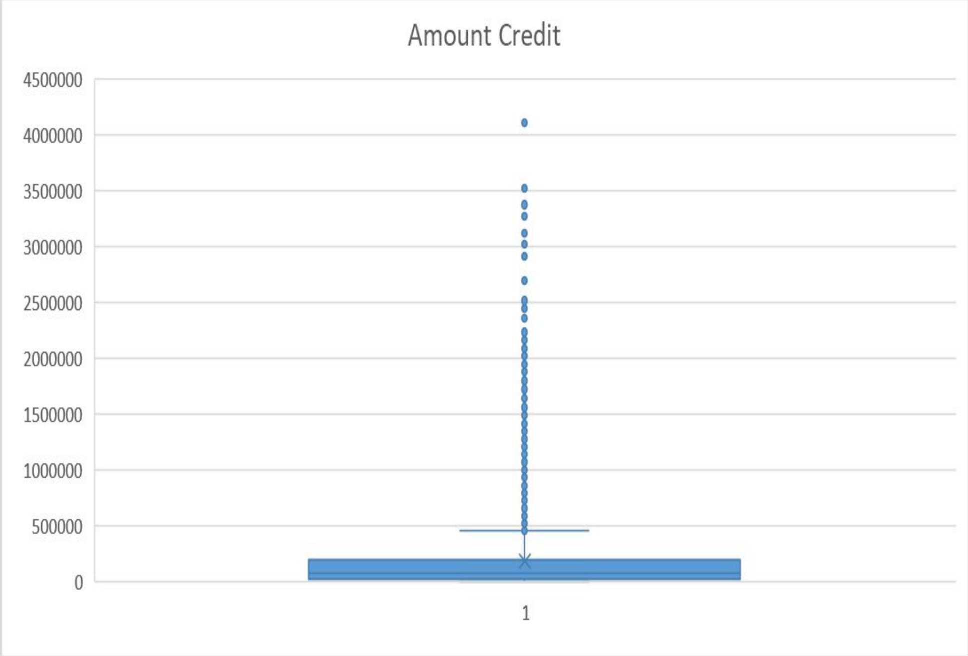
CNT family members has an entry over 13 members which can be possible if it's a joint family.

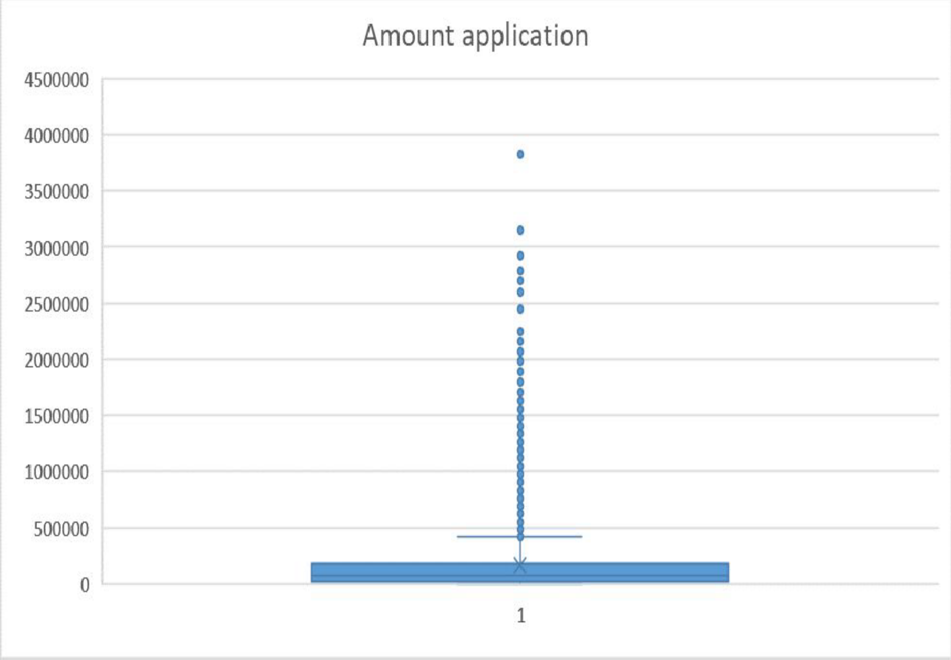




In Previous_application dataset, we can see that the amount credit and amount application columns have a large number of outliers.

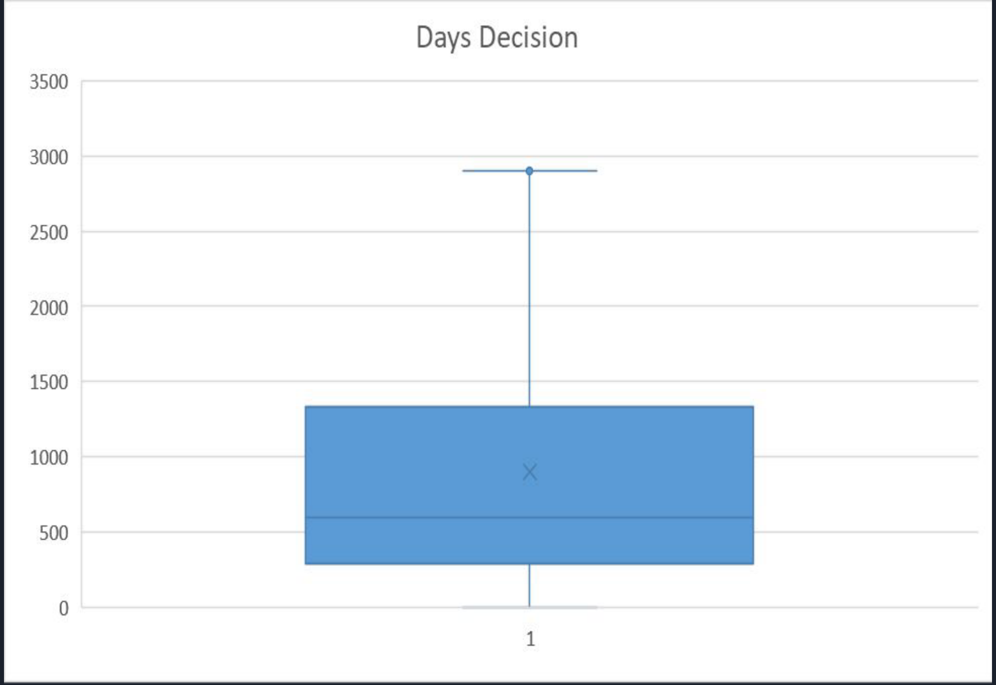
It shows that many consumers took a loan above 10 lakhs.

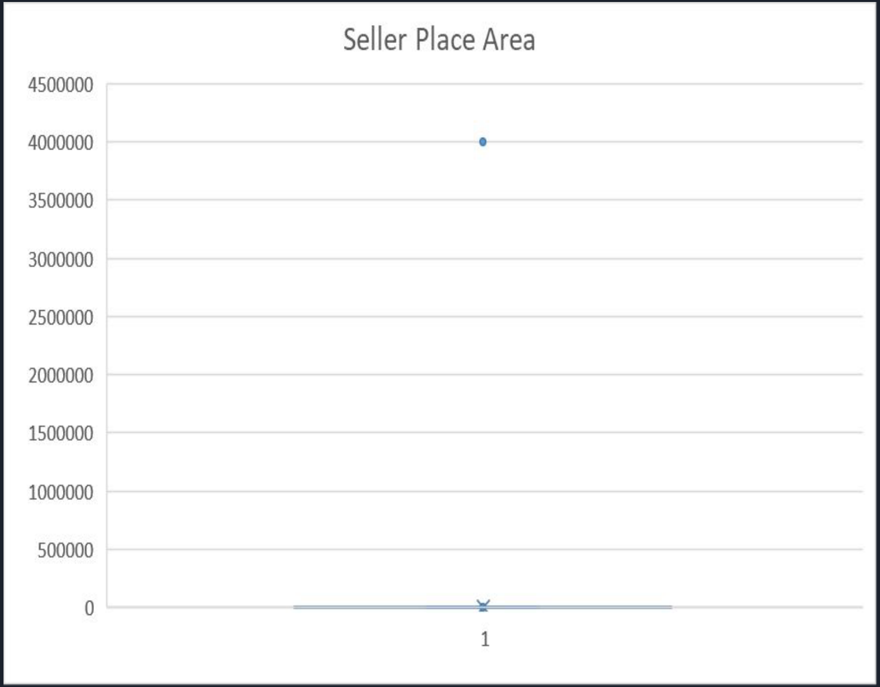




Days Decision has no anomalies which means the data is reliable.

Seller place area which is highly unlikely as the consumer usually takes loan near the location of the bank instead of being that far away.





Analyze Data Imbalance

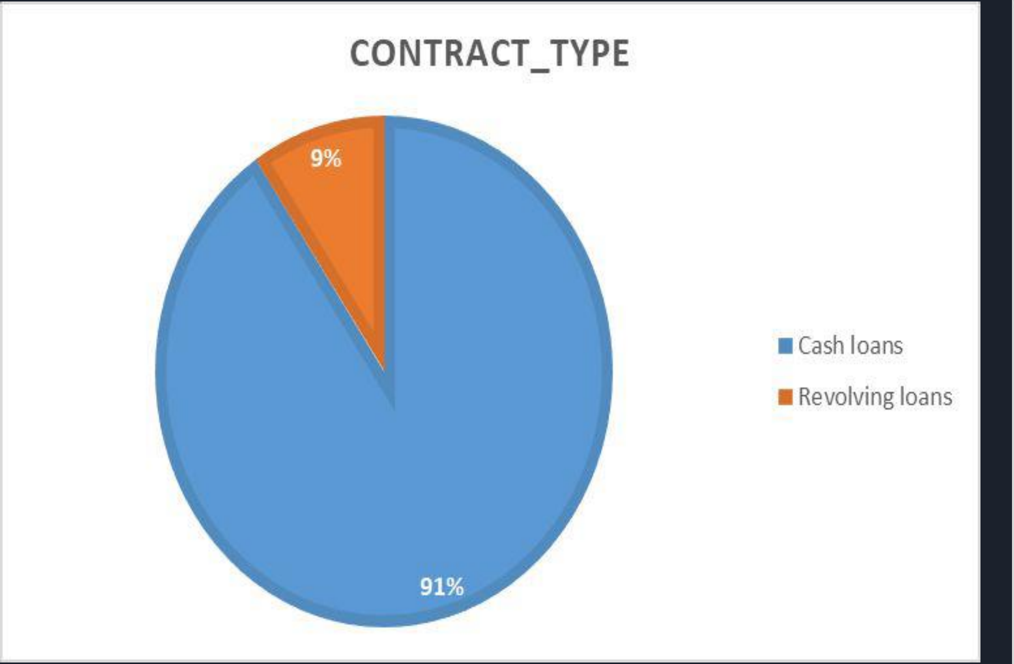
To determine if there is data imbalance in a loan application dataset, you'll need to check the distribution of the target variable.

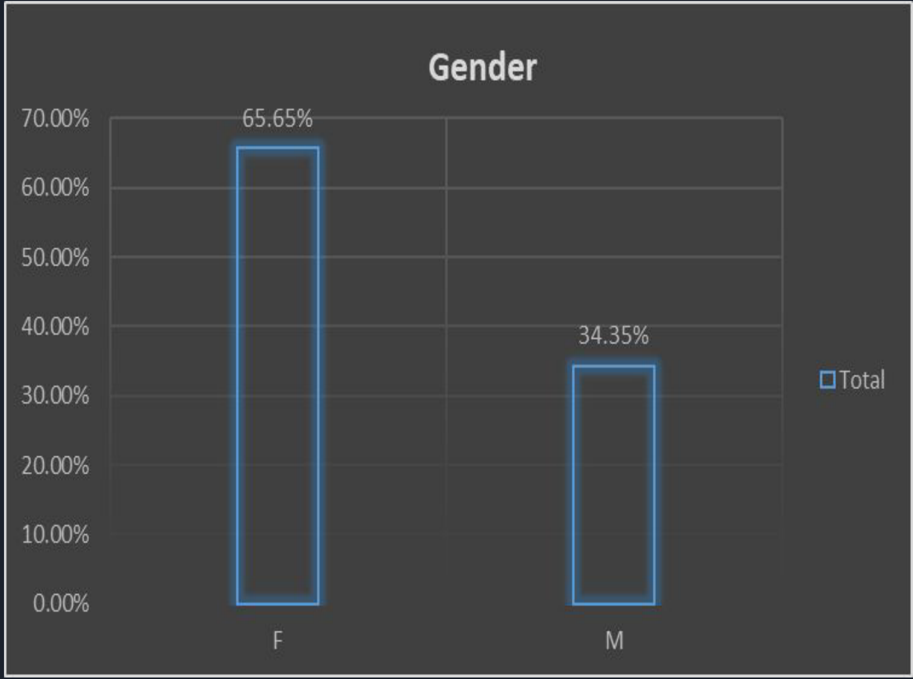
Data imbalance can occur on categorical variables and numerical variables which can be divided into sub groups.

We can calculate the ratio of data imbalance by comparing the counts of the two classes.

In Contract Type, we can see that revolving loans take up the majority which leads to an data imbalance in the ratio of 9.5 :: 1 to cash loans.

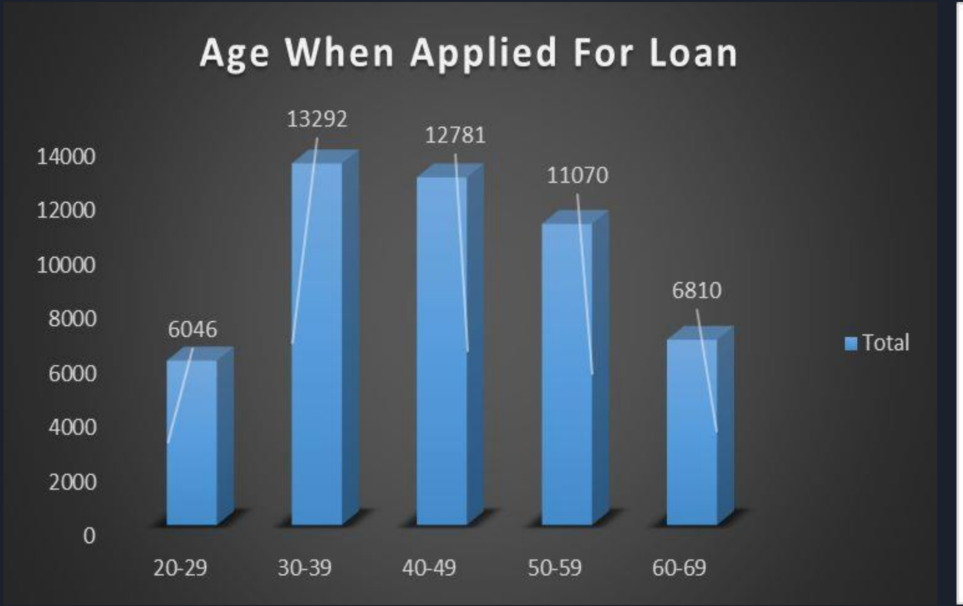
In Gender , we can see that the number of female loan applicants are nearly as double as the number of male applicants causing the ratio to be 1.9 :: 1.

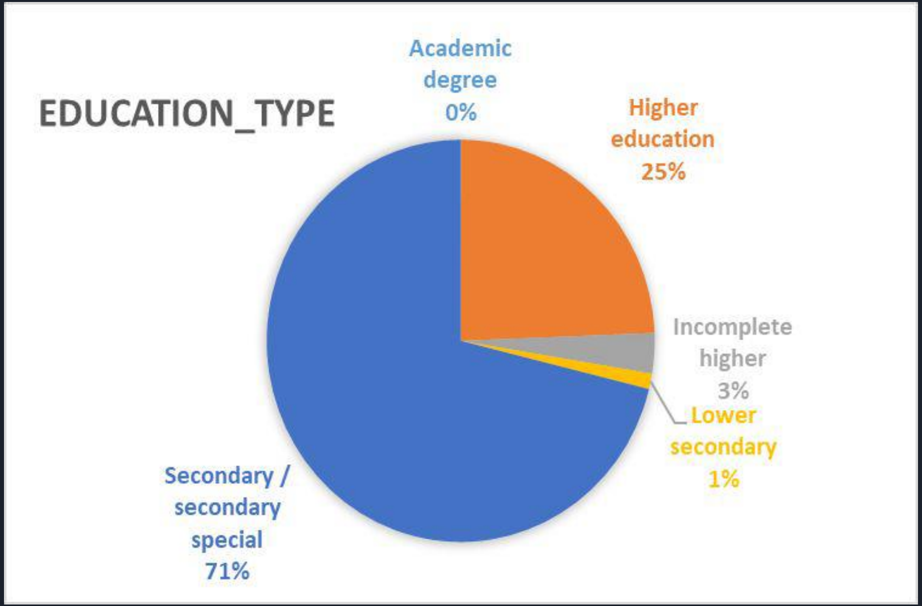




Taking the count of each age at which the consumer applied would be tedious. Therefore i divided them into subgroups to see the number the people applied for loan in each age group. Majority are from 30 to 60.

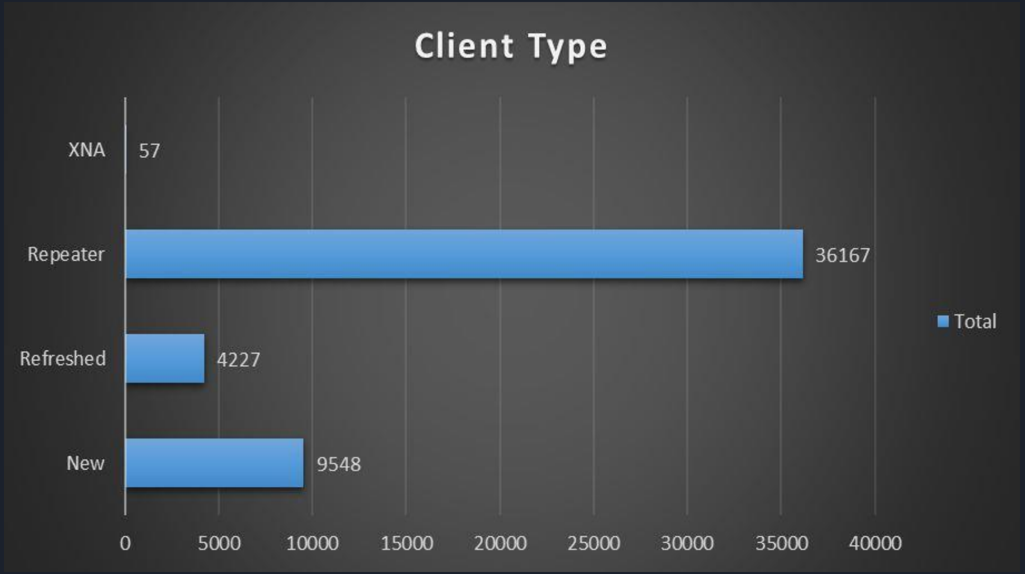
In Education , we can see that 71% of people who applied for loan completed secondary/ secondary special.

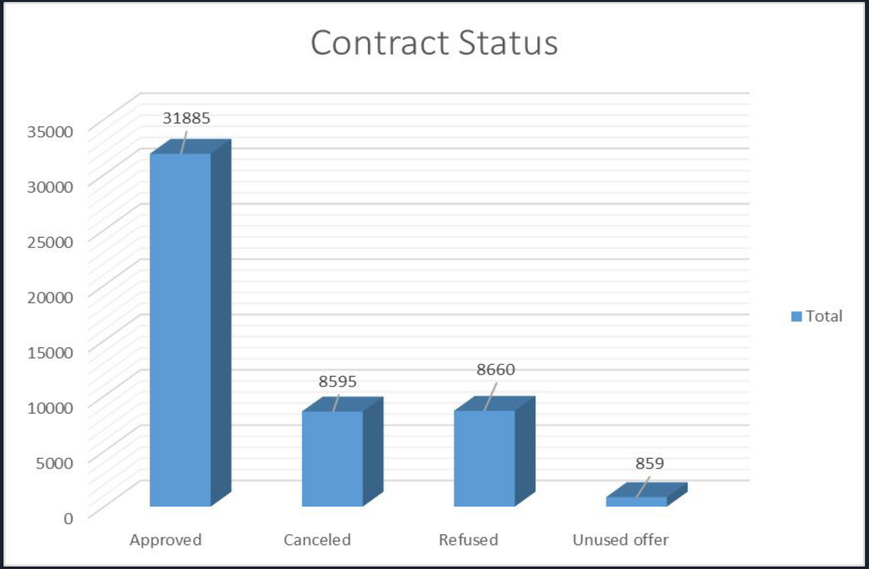




In previous_application.csv, we can see that most of the clients who applied for loan are repeaters(people who have previously applied for loans).

As per the contract status, Majority of the clients loan request has been approved(63%) while a few clients have refused the offer and a less than 1000 people have unused offer.





Perform Univariate, Segmented Univariate, and Bivariate Analysis

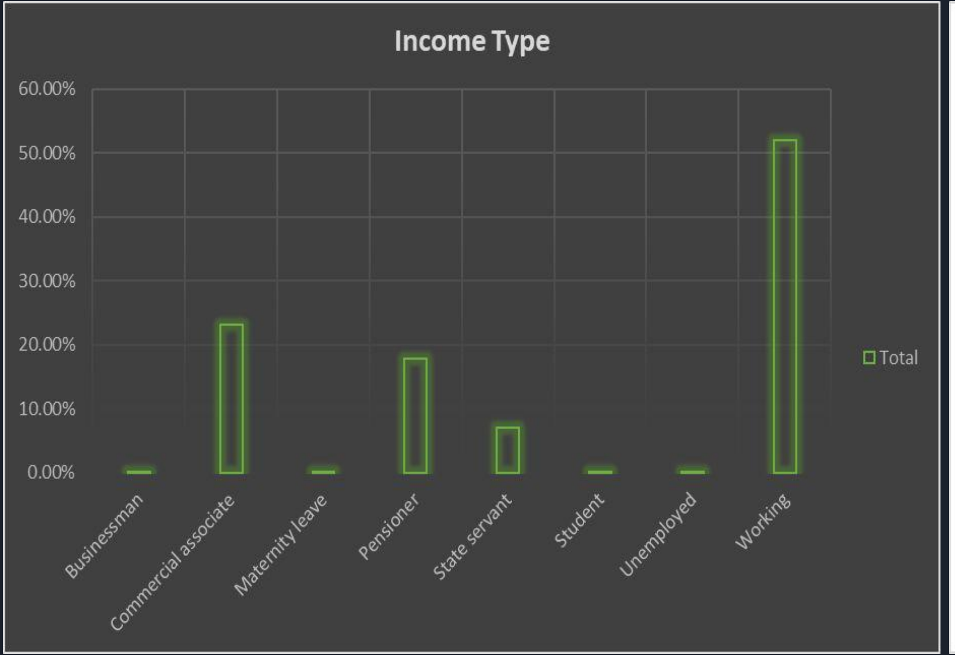
It is important to conduct various analyses on consumer and loan attributes.

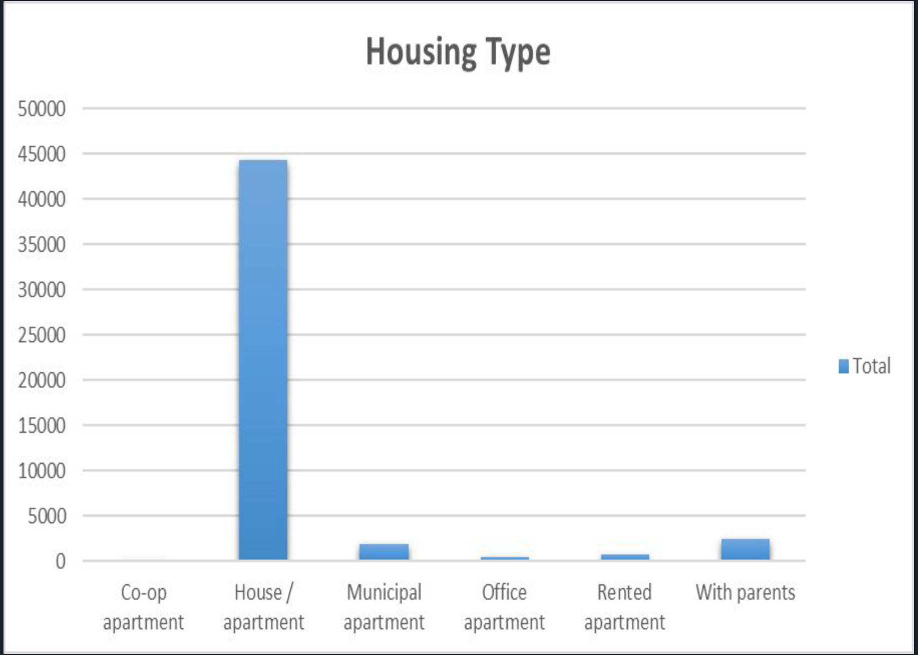
To do the analysis, various functions are created to analyse the dataset according to the required analysis (univariate/segmented univariate/ bivariate).

Univariate Analysis

Univariate analysis is taking a single parameter and comparing its values.

For categorical variables like these two, the results are best shown in bar charts or histograms.

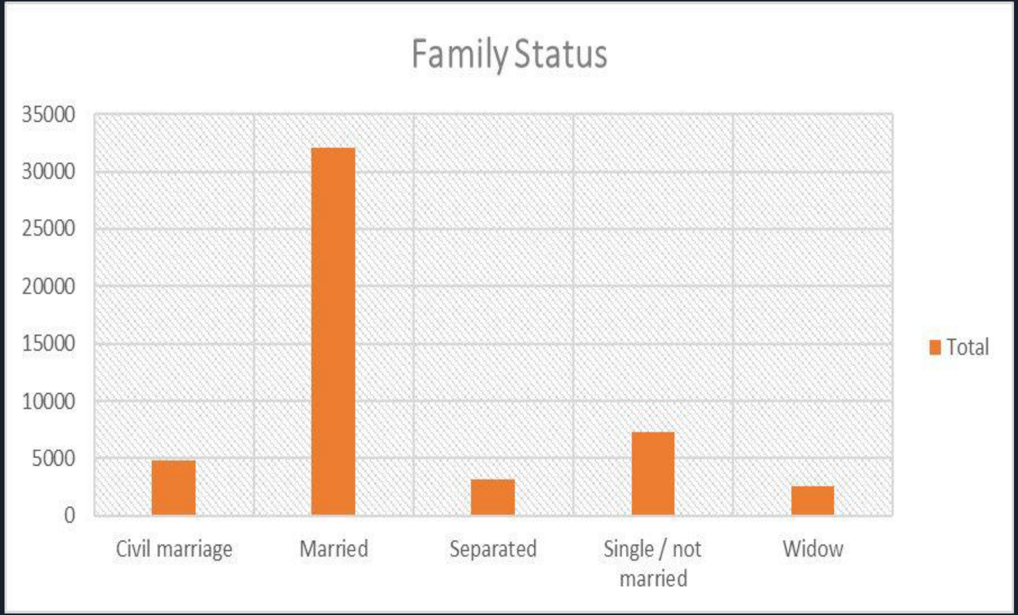




From the three charts , we can conclude that majority of the people who request for a loan are working employees who own a house/ apartment and most likely to be married.

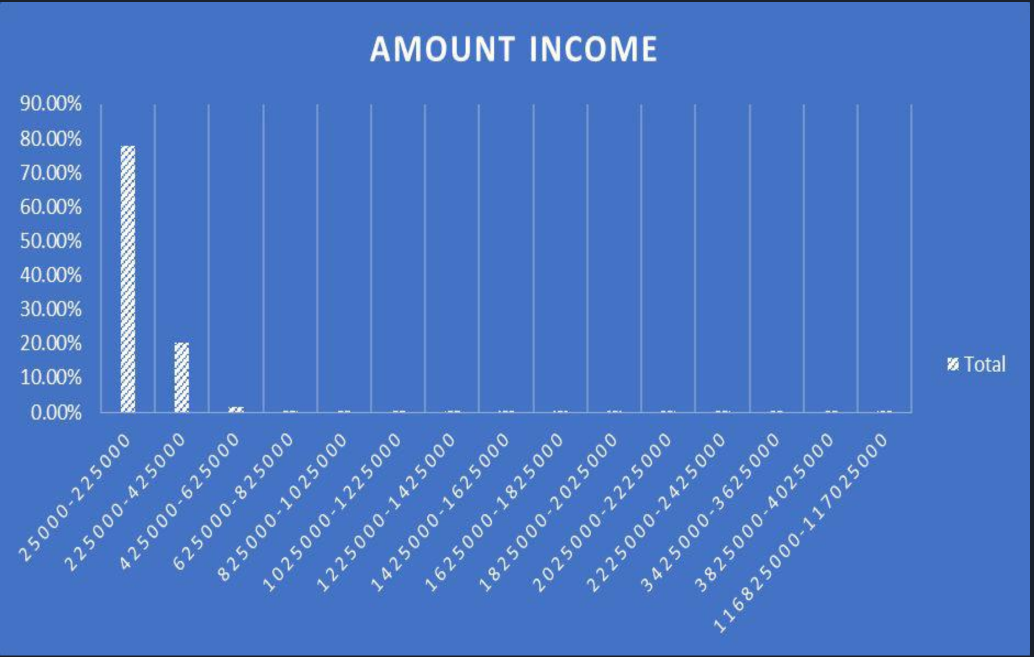
This can mean that they are opting for loan as they are going to start a family. Banks approve these loans more often as they have their house/apartment as collateral.

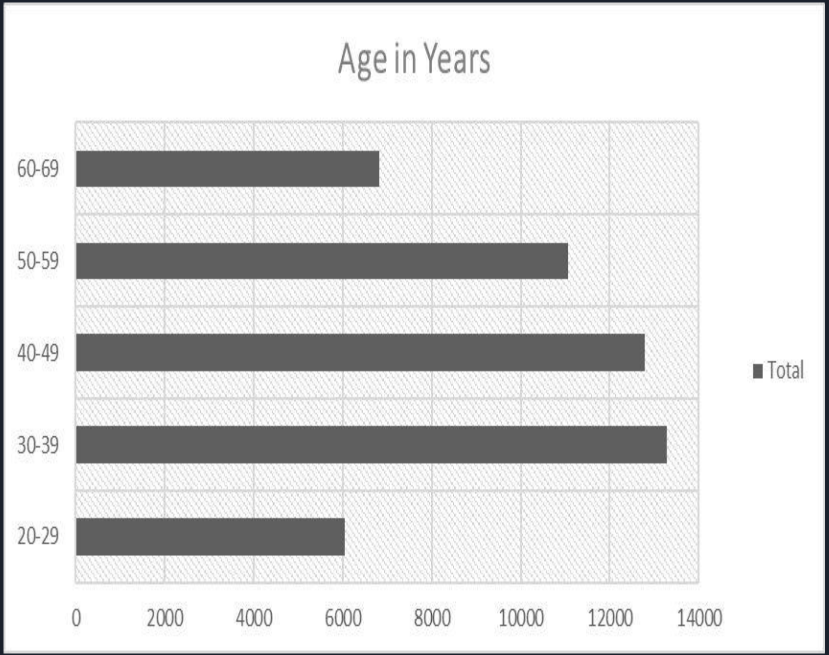
Students also opt for loans for their education although its minimal.



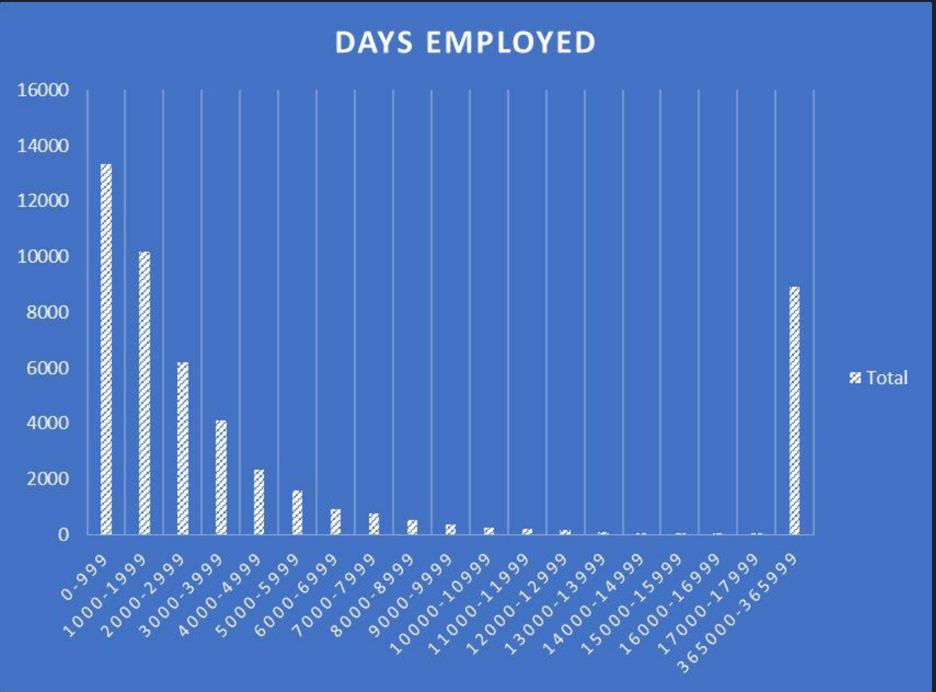
Segmented Univariate Analysis

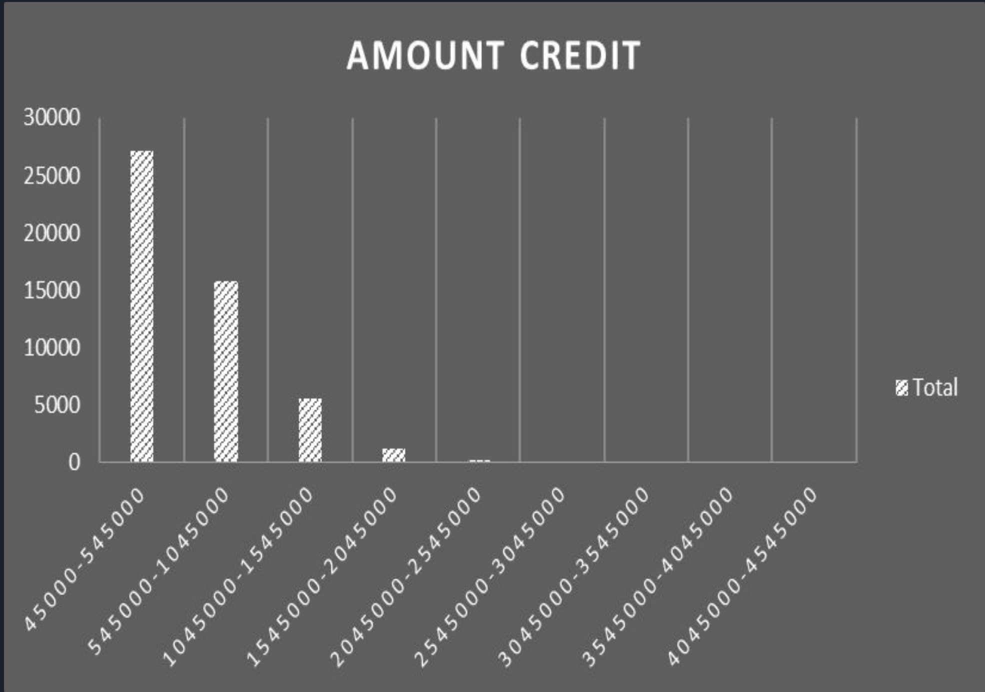
In segmented univariate analysis, you conduct univariate analysis within specific segments or categories of a variable. This helps us to see the patterns in each category.





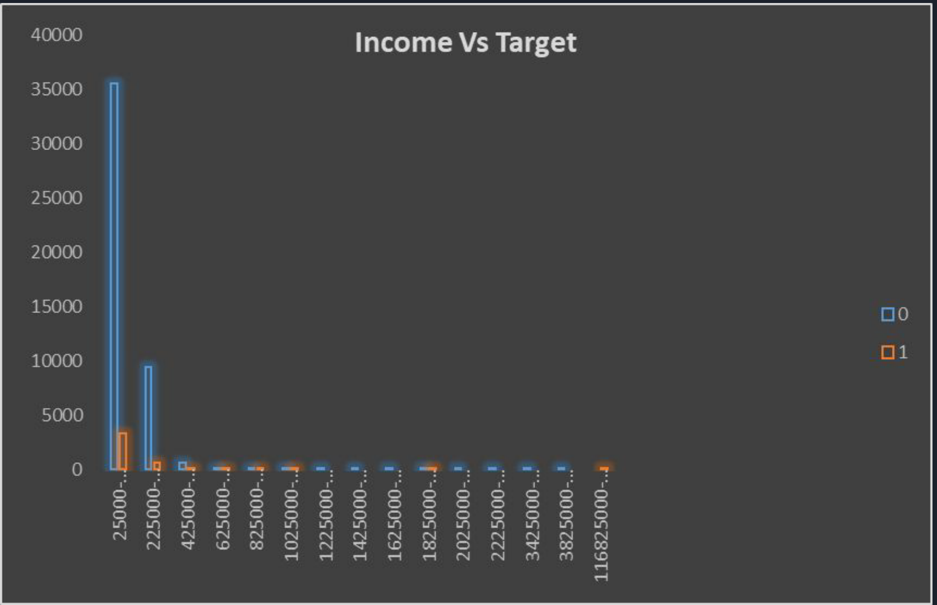
From these charts , we can conclude that people with higher incomes are less likely to apply for loans. Banks usually credit amounts ranging from 45000 to 15 lakhs. Excluding the incorrect entries in days employed, majority of the people apply for loans within 7 years of their employment.

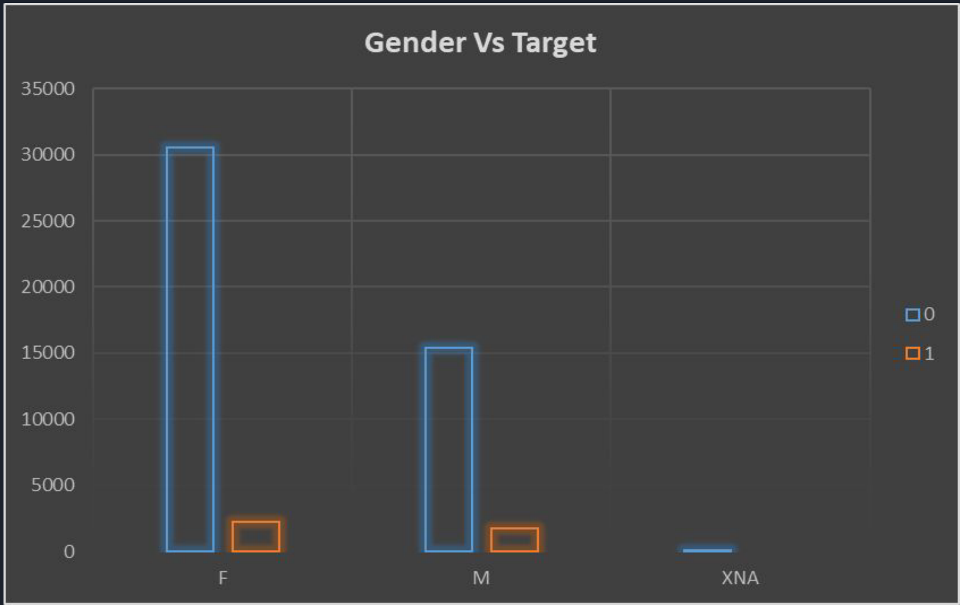




Bivariate Analysis

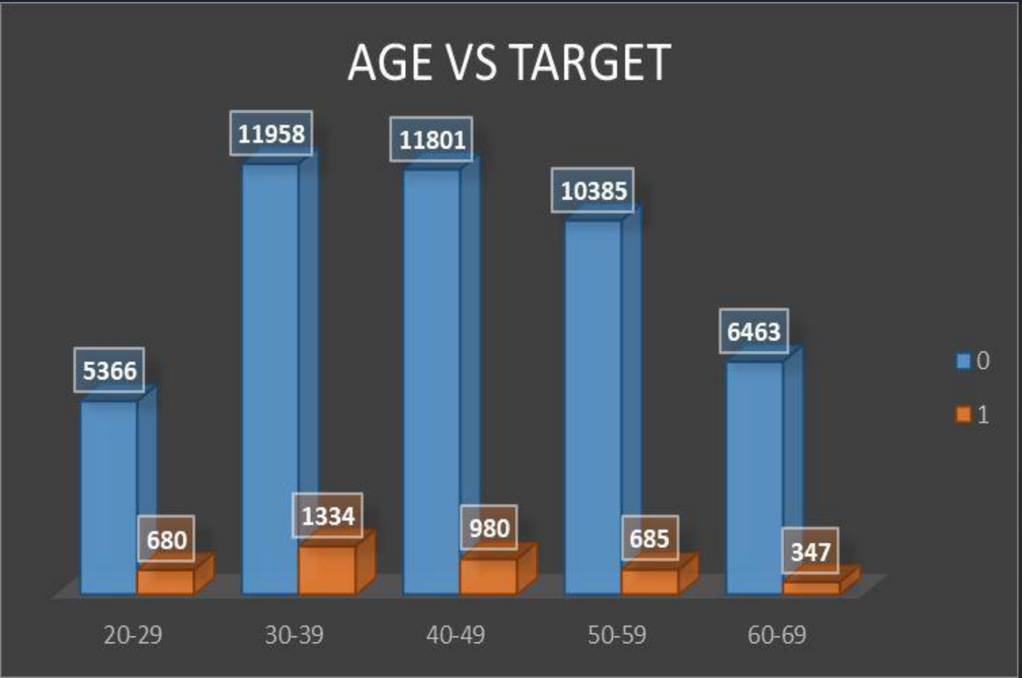
In bivariate analysis, we compare two or more parameters to see the correlation between them. Since this is a loan case study, we will be focusing on the defaulters based on different parameters. From the graphs below, we can deduct that defaulters are present in the salary range of 5 lakhs and below and the percentage of defaulters is higher in females than males.

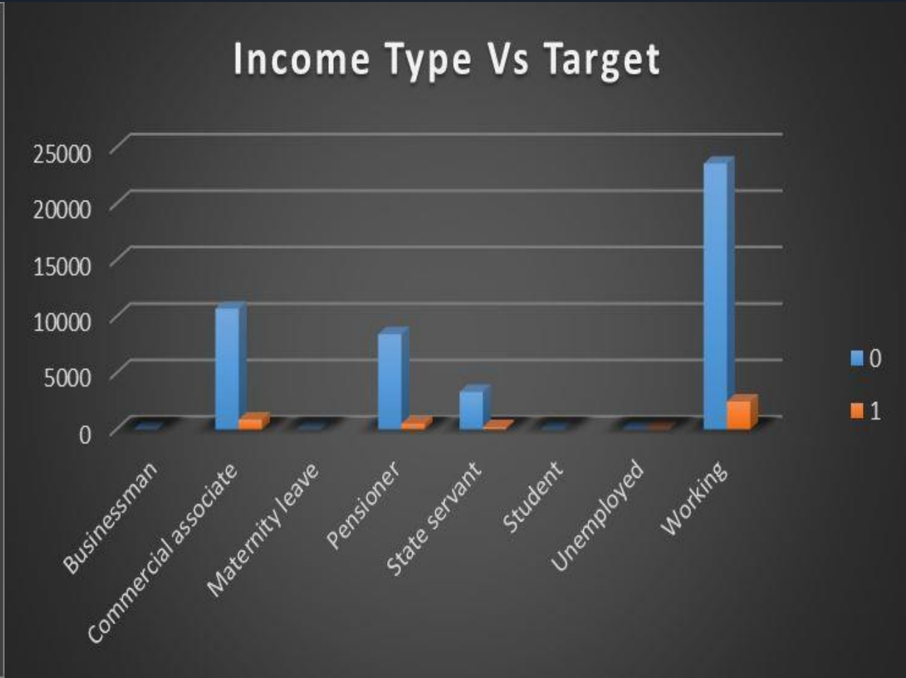




People in the age group 30-39 have the most defaulters but if we look it through the percentage of each category, 20-29 are having the most defaulters(11.25%).

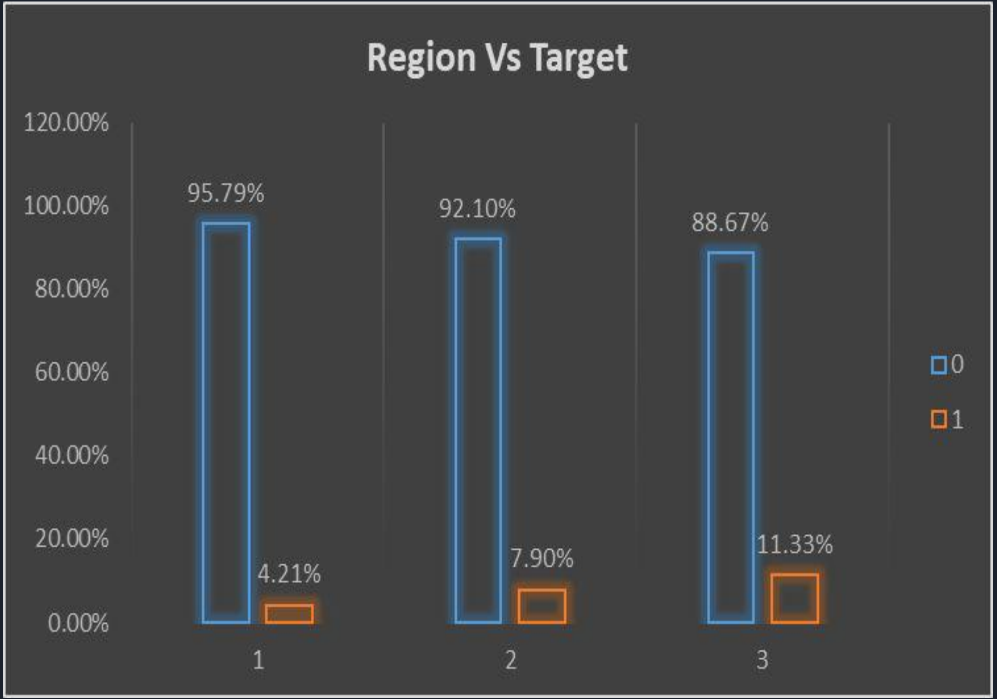
Unemployed people are not offered loans in most cases. In the rarest of cases where it happens, 33% of people are defaulters. When we look at the majority of people, Working people have the most defaulters.

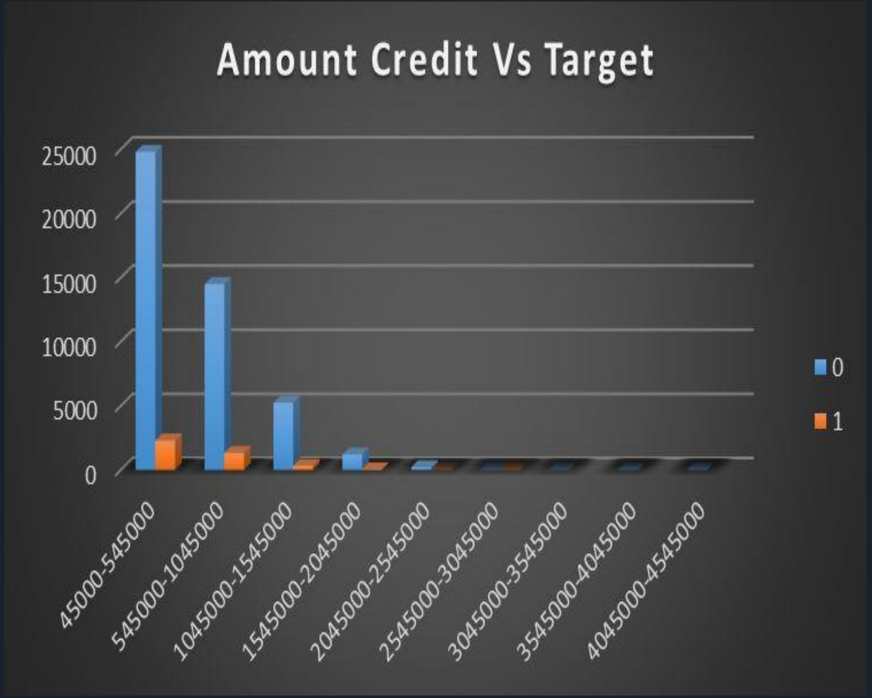




Region plays a vital role when it comes to defaulters. People who are region 3 tend to have more defaulters than the other regions.

People who gets credit amount below 10 lakhs show the most defaulters percentage than any other credit range.





Identify Top Correlations for Different Scenarios

Correlation is a statistical measure that helps us understand the relationship between two or more variables. It quantifies the degree to which changes in one variable correspond to changes in another.

In this Project, we will find the correlation between different variables to find the top indicators of loan default in each scenario.

	TARGET	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAY_Birth (In Years)	DAYS_EMPLOYED	DAYS_REGISTRATION
TARGET	1							
AMT_INCOME_TOTAL	0.010895624	1						
AMT_CREDIT	-0.032418778	0.069323857	1					
AMT_ANNUITY	-0.012387637	0.0830175	0.769510642	1				
AMT_GOODS_PRICE	-0.041296974	0.069893382	0.986944287	0.774444931	1			
DAY_Birth (In Years)	-0.076759978	-0.015986081	0.059401534	-0.00760104	0.05763803	1		
DAYS_EMPLOYED	-0.042469067	-0.03150712	-0.067723342	-0.108687629	-0.065043855	0.621616436	1	
DAYS_REGISTRATION	-0.042333443	-0.009943104	-0.003406347	-0.033164567	-0.006059604	0.333747537	0.209181122	

Top Correlation Variables

- Amounts_Goods_Price to Amount_Credit
- Amount_Annuity to Amount_Credit
- Amount_Goods_Price to Amount_Annuity
- Days_Employed to Days_Birth

I have used red colour for negative correlations fading to white for neutral and fading to blue for positive correlation.

From the correlation matrix, we can see the variables with most correlation values.

Insights For Repayers

- People with academic degrees have lesser defaulters.
- People living in region 1 are most likely to pay their loans on time.
- Businessmen and Students are less likely to be defaulters.
- People with higher income (i.e more than 10 lakhs) have a high on-time payment rate.
- People aged 45 and above are less likely to default.

Insights for Defaulters

- People with lower education(secondary or lower secondary) tend to default.
- Men are most likely to default than women.
- People aged below 40 have a high probability of defaulting.
- Certain industries such as Industry Type 13 (26%), Agriculture(13%), Industry Type 2(12%) have high defaulters rate. It is best to avoid providing high interest loans to these applicants.
- People who have more children have a higher tendency to default.

Summary

- Loans are recommended for clients whose previous application is approved.
- Applicants who are unemployed are high risk candidates.
- People with high income are less likely to default compared to people with low income.
- Applicants with more than 5 years of work experience are more trustworthy when it comes to approval.
- Young applicants and applicants whose previous application is denied are to be avoided.
- Female applicants are more likely to pay their due on time compared to male applicants.

Conclusion

In conclusion, this project has been a valuable journey into the world of data analytics, offering a hands-on experience that significantly enhanced my understanding of the subject. Through various tasks and analyses performed on the loan application dataset, I've not only honed my Excel skills but also gained critical insights into the intricacies of working with real-world data.

This project deepened my knowledge of data analytics by providing practical exposure to key techniques and tools, including pivot tables, correlation matrices, and various Excel functions. It reinforced the importance of data preprocessing, outlier detection, and understanding data distributions. These skills will undoubtedly be invaluable in future data analysis endeavors, and I am better equipped to harness the potential of data to derive meaningful insights.

Links for the Excel Files:
https://drive.google.com/drive/folders/1fqXihh_MSvaVdItfbtYjo2RgI-J8R5TY?usp=sharing