BANK LOAN CASE STUDY

-by Akhilesh Mishra

Project Description:

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Objectives :

- analysis
- missing data
- Outliers
- data imbalance
- univariate, segmented univariate, bivariate analysis
- Correlation

Approach :

Data Collection- Create Excel table of the dataset from source which includes movie title,
IMDb score, genre, director, budget, gross, etc.

Data Preparation- Perform thorough data cleaning and preparation tasks to ensure the dataset is accurate, consistent, and ready for analysis.

Performing Analysis- Create table of the dataset to extract required and requested insights, utilizing
advanced Excel functions, statistics to uncover hidden patterns and relationships within the data.

Visualization- Create dashboard to present the extracted data.

Tech-Stack :

Excel 2019- We employed Excel as our visualization tool to create stunning charts, graphs,
and dashboard. Its extensive library of chart types, customization options, and interactive features allowed us to create compelling visualizations that effectively conveyed the key insights.

To Find :

- Analysis
- Missing Data
- Outliers
- Data Imbalance
- Univariate, Segmented Univariate, Bivariate analysis
- Correlation

Analysis :

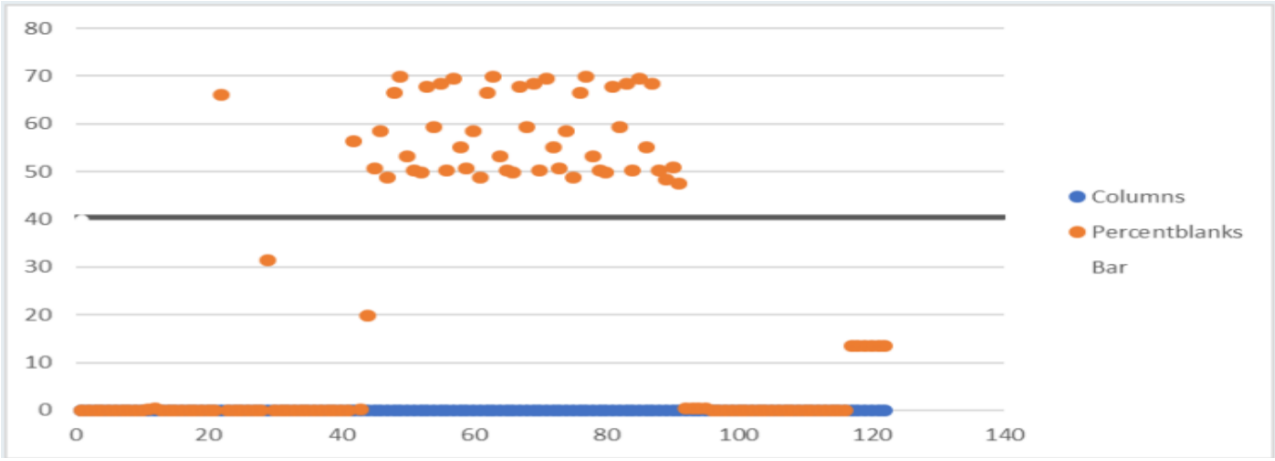
We will carefully go over the both Microsoft Excel spreadsheets files application_data
and previous_application to look for and remove any unnecessary or redundant information, including
excess columns and rows. Data cleansing techniques will be used as necessary.
The datasets will also be examined for outliers to identify any skewness in the pertinent columns. For
accurate visualization and insightful analysis, this stage is crucial.
Additionally, the occurrence of data imbalance will be investigated.
To study the relationships between various variables and pinpoint the factors that cause loan defaults,
various forms of analysis will be used.
Key insights can be discovered using statistical techniques, correlation analysis, and other strategies.

Visualizations will be used to further investigate the connections, providing a better understanding of the patterns and connections in the data.

Data Cleaning:

Removing Duplicates & Blanks:

More than 30% of the cells in the dataset are empty. There are about 30% of useful columns in the dataset.
There are two methods for dealing with these columns. Although I may opt to eliminate these columns, doing so might result in the loss of information I need for my research. However, adding the missing values might distort the statistics. Columns with over 40% of blanks were removed. If the columns should be preserved, it depends on how well the issue statement is understood, how useful the variable is, and how much data there is overall. There are still several columns with just a few missing values; we will leave these columns alone.

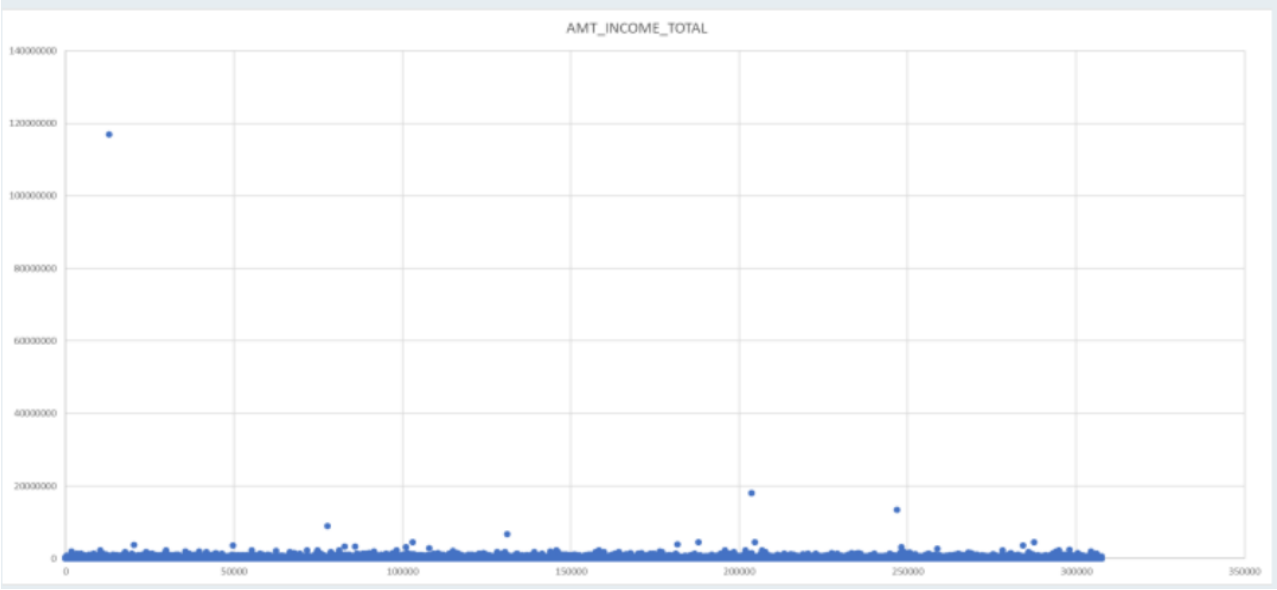


Identifying Outliers:

We have to find the outliers in the datasets. Graphical analysis can help identify outliers in a dataset by examining points as isolated points in a scatterplot.

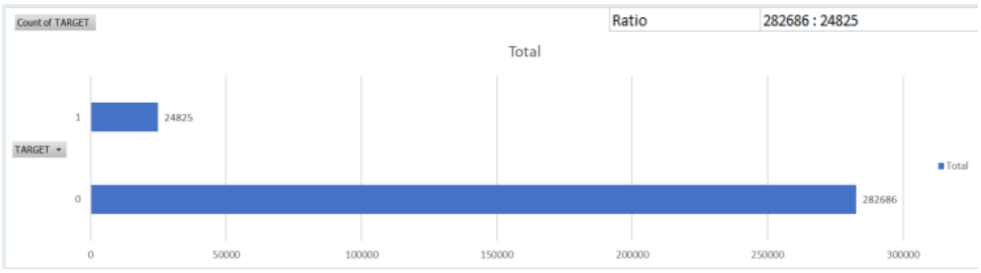
Analysis:

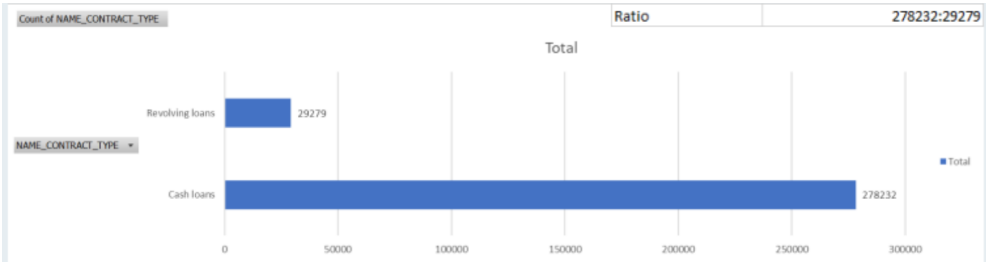
Making use of the scatter plotted graphs we can see that there are many outliers in the amt_income_total column. We other important columns to identify the outliers.
From the graph we can see that the most incomes are below 2Cr and one of the data points is around 11Cr this makes the data skewed.

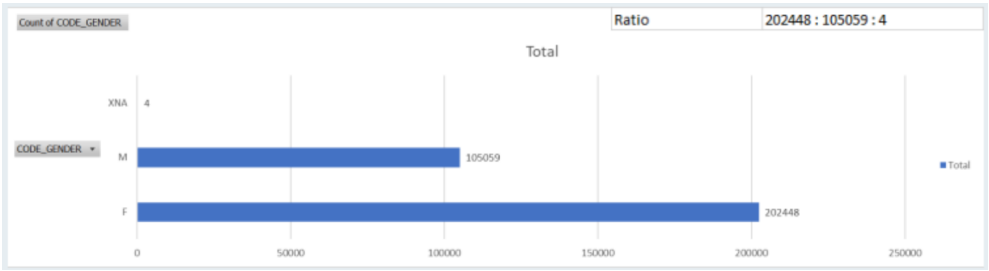


Data Imbalance:

We need to find the imbalance in the data . We can get this by using pivot table and chart, to find ratios we have to calculate GCD then finding the ratio.







Analysis:

Let’s create a new pivot table for each of the columns we are checking for imbalance. Lets check for the amount of applicants with target, contract type, gender, etc columns.

From the charts we can see the imbalance in target, contract type, gender data. Its clear that there are more applicants that payback on time, most applicants want cash loans and most applicants are women.

Univariate:

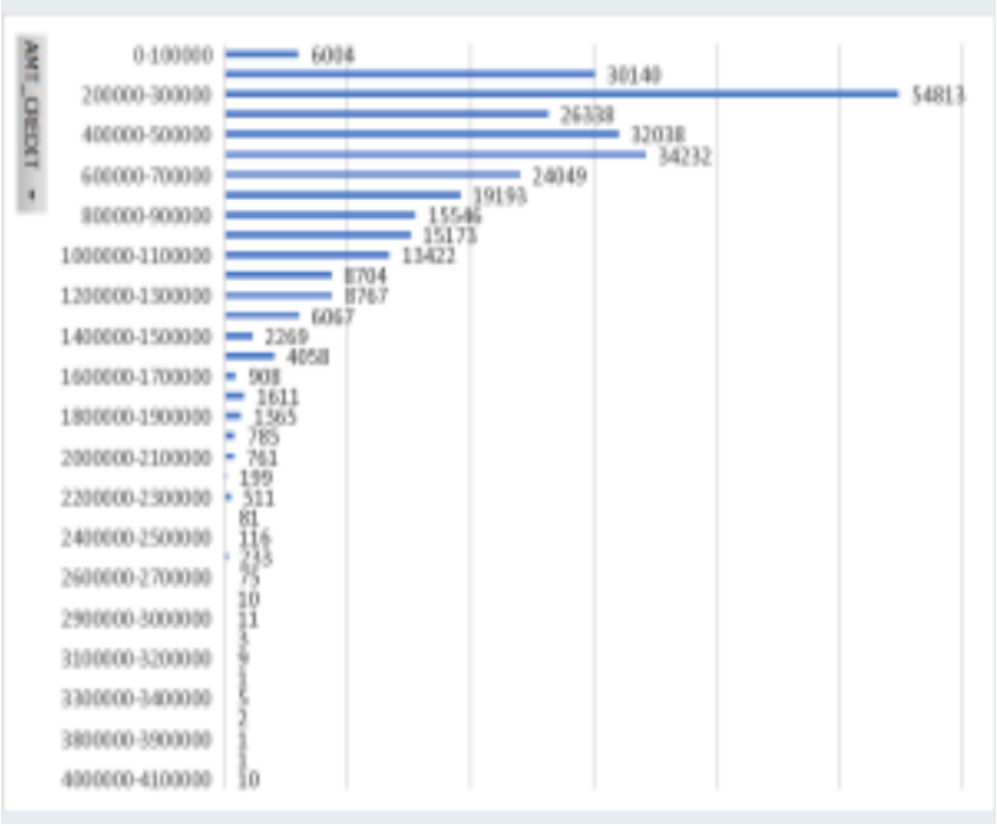
Analyzing one variable at a time is known as univariate analysis.

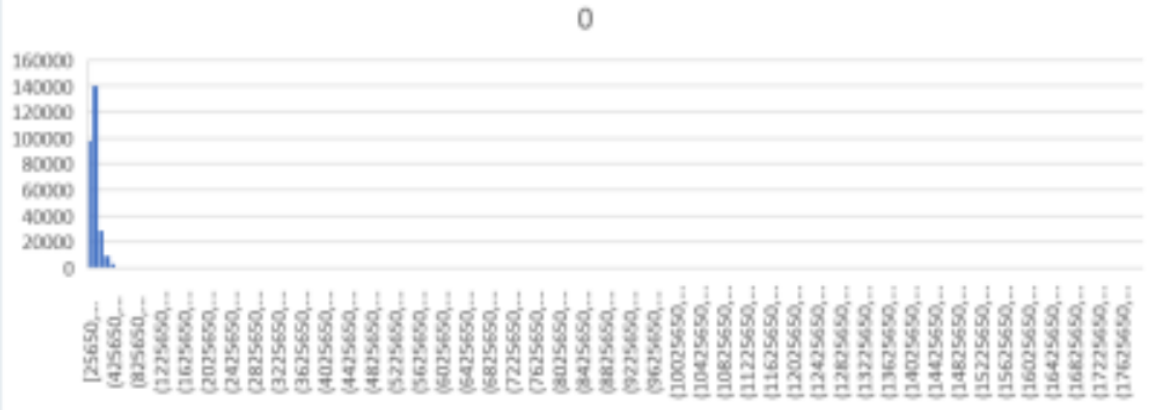
Following are some translations of the findings of the univariate study into business terms:

We can ascertain the distribution of each category and comprehend the frequency of occurrence for categorical variables, such as income and credit.

Column1	
Mean	168797.9193
Standard Er	427.6058332
Median	147150
Mode	135000
Standard De	237123.1463
Sample Var	56227386501
Kurtosis	191786.5544
Skewness	391.5596541
Range	116974350
Minimum	25650
Maximum	117000000
Sum	51907216961
Count	307511

Column1	
Mean	599026
Standard I	725.8144
Median	513531
Mode	450000
Standard I	402490.8
Sample V	1.62E+11
Kurtosis	1.934041
Skewness	1.234778
Range	4005000
Minimum	45000
Maximum	4050000
Sum	1.84E+11
Count	307511





Analysis:

Lets create graphs for Incomes, Credits and also for incomes with target 0 and target 1.

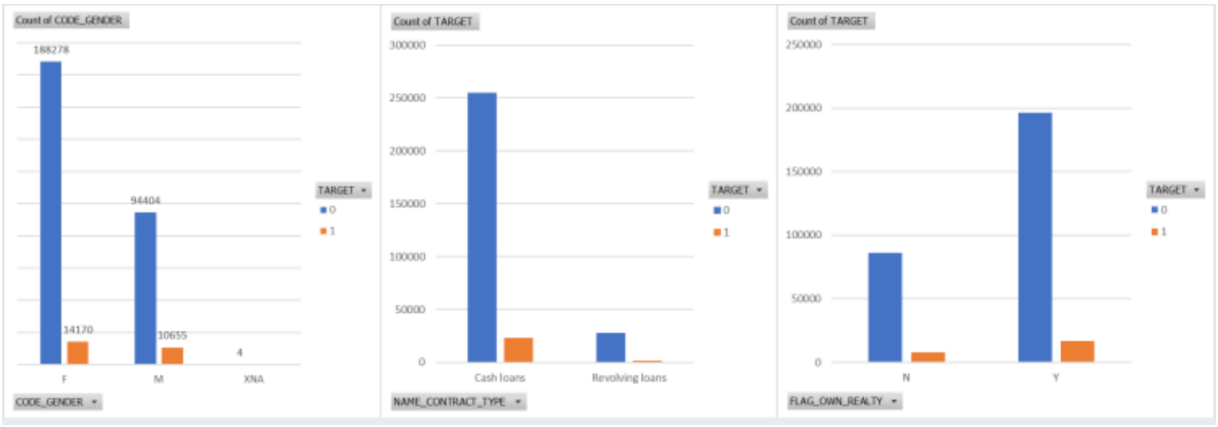
In the charts we see incomes are more in 1L-2L bin, and loan requests for 2L-3L are the highest.

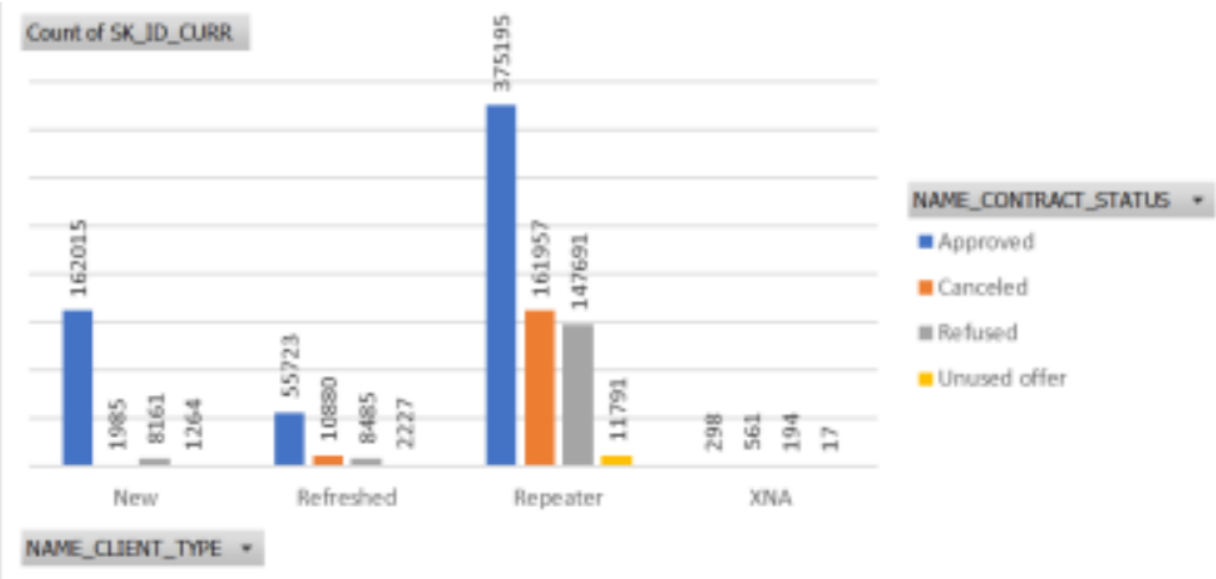
Segmented Univariate:

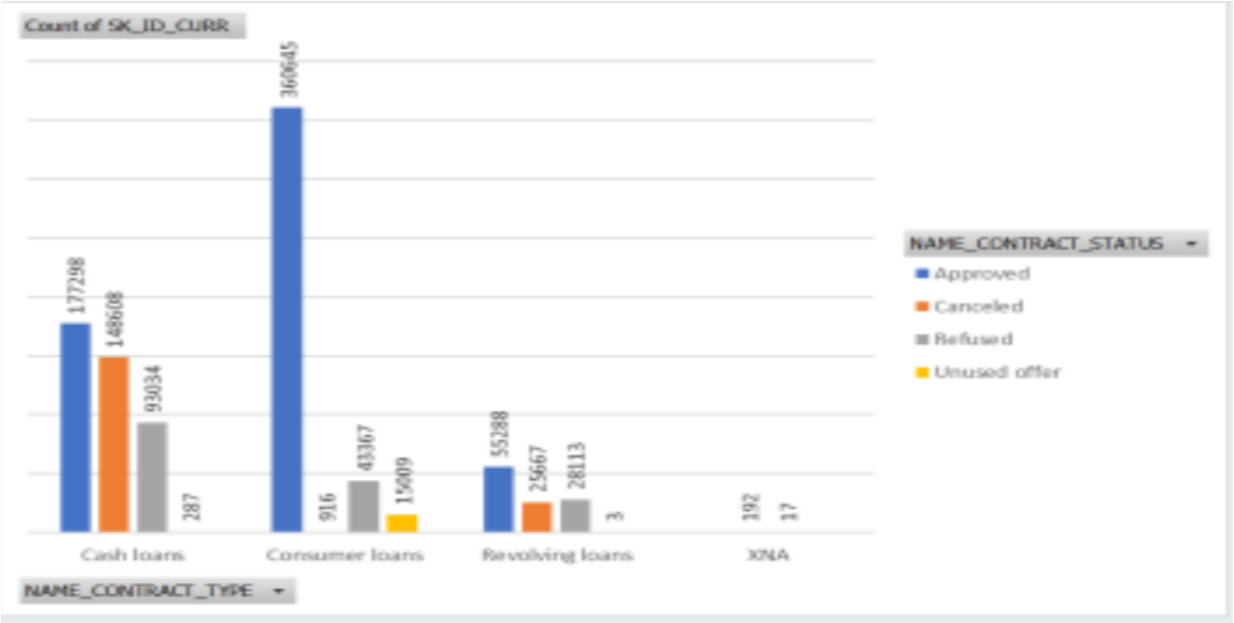
For segmented univariate analysis, it entails analyzing the distribution of a single categorical variable independently for multiple groups, allowing for concentrated insights and comparisons within each group. This strategy aids in identifying distinct trends and variations in data depending on various categories.

Analysis:
Let’s use pivot table to find the insights.

From the charts it is clear that most females, cash loans, realty owners payback. Most repeater client and consumer loans get their loans approved, highest rejections are in repeater clients and cash loans.



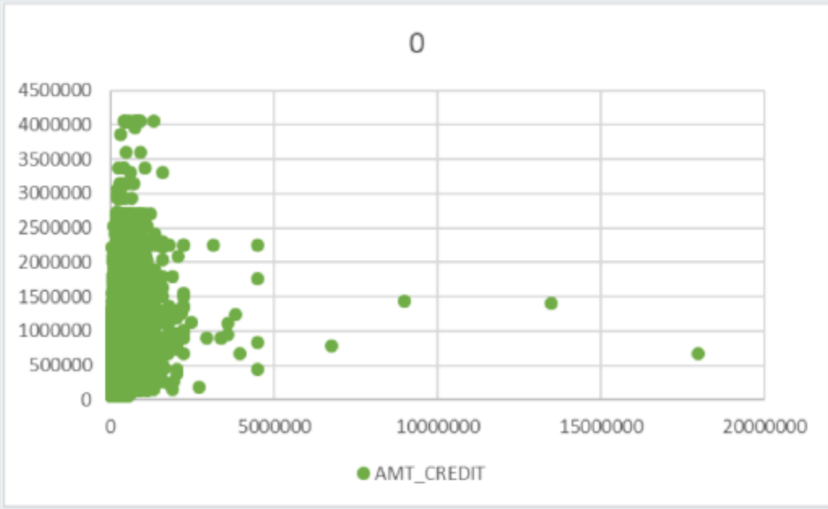


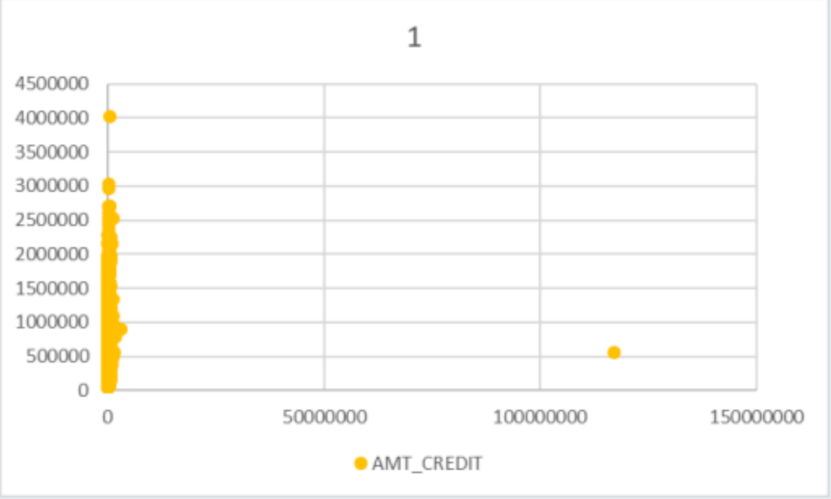


Bivariate analysis:

Bivariate analysis looks at the relationship between two variables to see if there are any connections or patterns. It facilitates understanding of how changes in

one variable affect the other, giving useful insights for data analysis and decision-making.





Analysis:

Making use of amount credit and income we performed the Bivariate analysis in target values.

From the charts we see that applicants with high income are ore likely to payback compared to low income applicants having problems paying back.

Correlations:

We have to find the correlations between the attributes. Lets use the correl(c function in excel to find it. We found correlations by both the correl(c and Correlations in Excel Data Analysis Tools.

Target=	0									
	CNT_CHILDRI	INCOME_T	MT_CREDI	MT_ANNUI	GOODS_P	OPULATIO	EARS_BIR	TRS_EMPLO	RS_ID_PUB	RATING
CNT_CHIL	1									
AMT_INC	0.027397	1								
AMT_CRE	0.003081	0.342799	1							
AMT_ANN	0.020905	0.418953	0.771309	1						
AMT_GOC	-0.00052	0.349462	0.98725	0.776686	1					
REGION_F	-0.02436	0.167851	0.100604	0.120988	0.103827	1				
YEARS_BIF	-0.33685	-0.0626	0.047426	-0.0122	0.044601	0.025168	1			
YEARS_EM	-0.24517	-0.14039	-0.0701	-0.10497	-0.06861	-0.0072	0.625824	1		
YEARS_ID_	0.028569	-0.02284	0.001405	-0.01407	0.003587	0.001193	0.271571	0.27592	1	
REGION_R	0.022842	-0.18657	-0.10334	-0.13213	-0.10438	-0.539	-0.00231	0.03833	0.008992	1

Target=	1									
	CNT_CHILDRI	INCOME_T	MT_CREDI	MT_ANNUI	GOODS_P	OPULATIO	EARS_BIR	TRS_EMPLO	RS_ID_PUB	RATING
CNT_CHIL	1									
AMT_INC	0.004796	1								
AMT_CRE	-0.00167	0.038131	1							
AMT_ANN	0.031257	0.046421	0.752195	1						
AMT_GOC	-0.00811	0.037583	0.983103	0.752699	1					
REGION_F	-0.03197	0.009135	0.069161	0.07169	0.076049	1				
YEARS_BIF	-0.25891	-0.00287	0.135318	0.014249	0.135744	0.048294	1			
YEARS_EM	-0.19286	-0.01497	0.001933	-0.0812	0.006647	0.015543	0.581765	1		
YEARS_ID_	0.032679	0.004242	0.052461	0.016781	0.056084	0.016089	0.25279	0.228181	1	
REGION_R	0.04068	-0.02149	-0.05919	-0.07378	-0.06639	-0.44324	-0.03409	0.003486	-0.00206	1

Analysis:

Creating a new dataset for target is 0
and 1 and finding correlations between them.

From the chart it is clear that correlation
between amt goods and amt credit is the
highest and lowest is between region
population relative and region rating client.

Results:

Removing Duplicates & Blanks :

- Columns with more than 40% blanks were removed, but those with minimal blanks were retained

Identifying Outliers :

- By employing graphical analysis, we identified outliers in the "amt_income_total" column
- Most incomes were below 2Cr, while one data point was around 11Cr, leading to data skewness

Data Imbalance :

- We saw more on-time paybacks, a preference for cash loans, and a higher number of female applicant

Univariate Analysis :

- Incomes were concentrated in the 1L-2L range, and loan requests were highest in the 2L-3L range

Segmented Univariate :

- Most females, cash loans, and realty owners exhibited higher payback rates, while repeater clients and cash loans faced more rejections

Bivariate Analysis:

- Applicants with higher incomes were more likely to payback their loans compared to those with lower

Correlations :

- The highest correlation existed between "amt_goods" and "amt_credit," while the lowest was between "region_population_relative" and "region_rating_client."

Conclusion:

Finally, in this data analysis project, we delved into various aspects of the dataset, unearthing valuable insights through different analytical methods.
In conclusion, our data analysis endeavors provided crucial insights for informed decision-making. By addressing data quality issues, identifying outliers, detecting imbalances, and exploring variable relationships, we gained a comprehensive understanding of the dataset. These findings serve as a valuable foundation for future research and decision-making in various domains.