Bank Loan Case Study Project

Project Description:

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their instalments. 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.

When a customer applies for a loan, there are four possible outcomes:

1. Approved: The company has approved the loan application.
2. Cancelled: The customer cancelled the application during the approval process.
3. Refused: The company rejected the loan.
4. Unused Offer: The loan was approved but the customer did not use it.

Approach:

I went through the Excel data provided by the Trainity Bank Loan Case Study and understood that there were columns related to the Bank loan in the dataset. Further, I understood the columns and their respective constraints to do the analysis. I was given a set of questions to solve as part of the analysis. By using the Microsoft Excel, I did solve the queries and provided the result as expected.

Tech-Stack Used:

Microsoft Excel 2021 – To answer the queries with the help of Excel formulas in the tool.

Insights:

Did the data cleaning like:

- Removing null values.
- Removed the columns which we don't use for the analysis.
- Removing the Duplicate rows.

Before the Data Cleaning the column number for the Excel:

application_data – 126, After cleaning now we have 77 columns.

previous_application – 37, After cleaning now we have 26 columns.

With the help of the Excel formulas, I found out many insights which include –

Task A - Identify Missing Data and Deal with it Appropriately:

Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

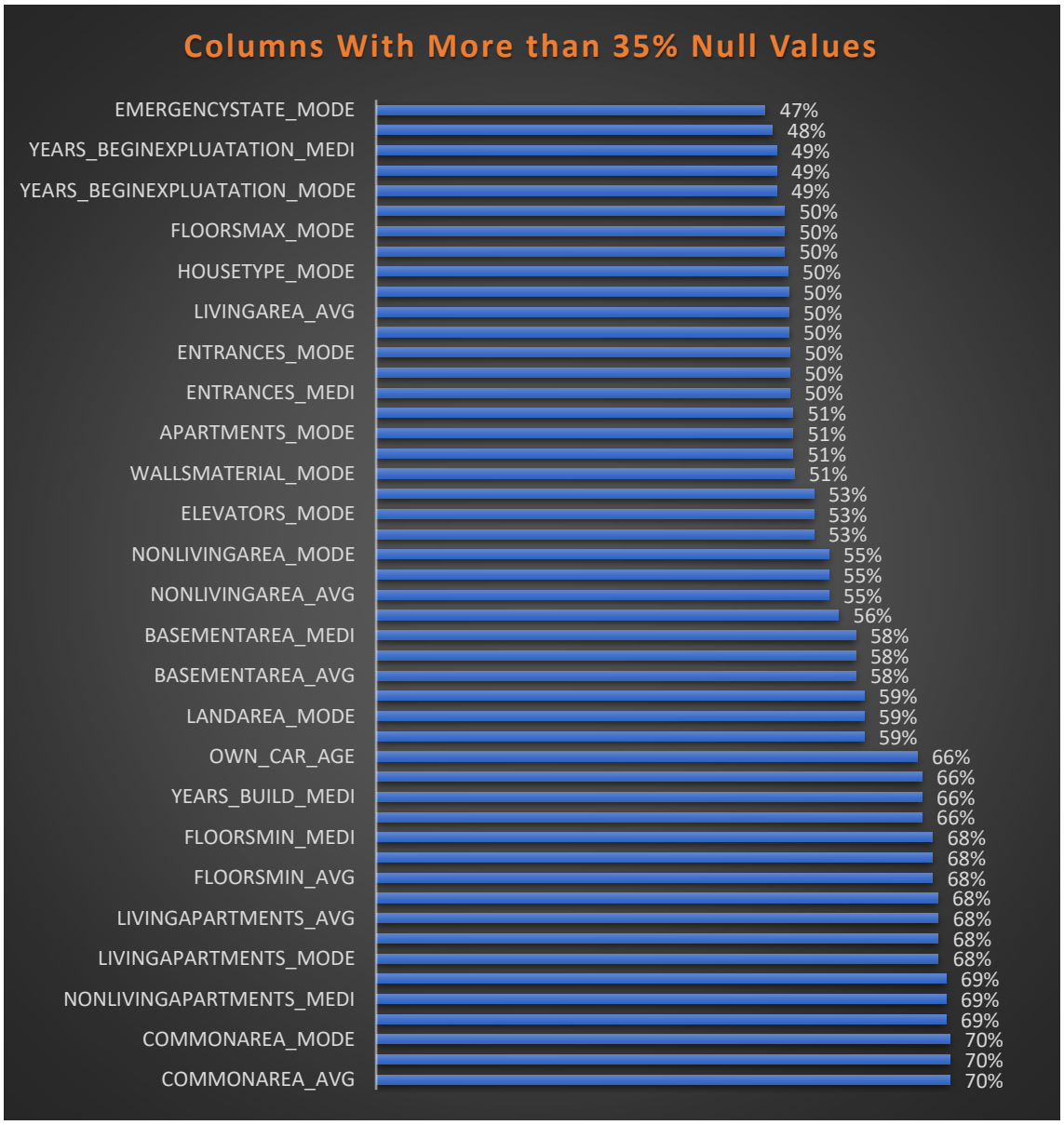
Used the formula counta formula to find out the total number of values in each column:

=COUNTA(B4:B50002)

Then to find out the Percentage of Null values used this formula:

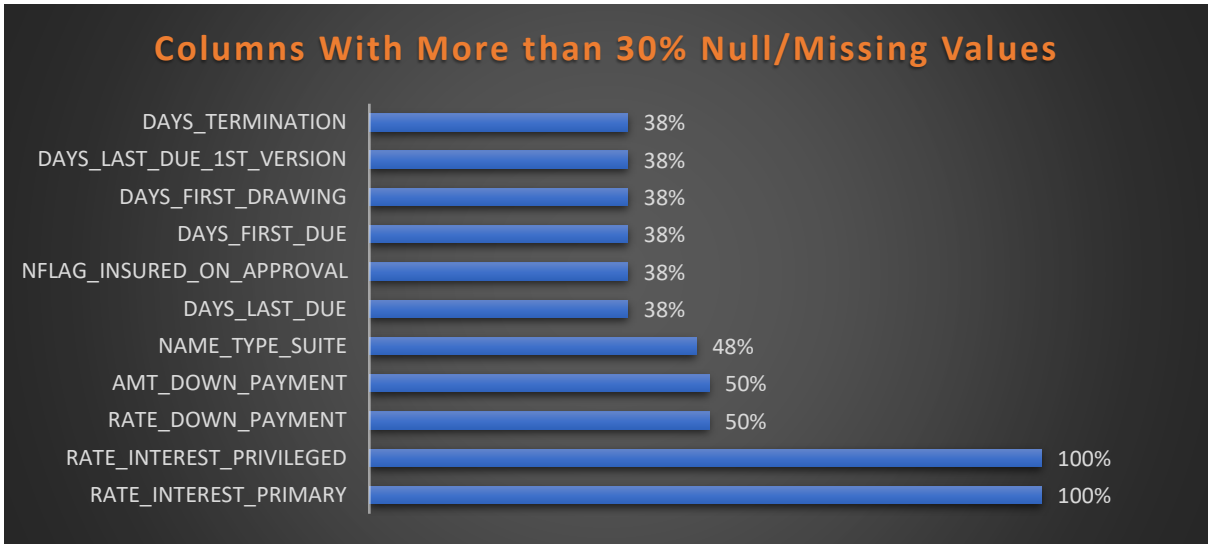
=1-B2/B2

Also decided to remove the columns which had more than 35% of null values since they don't contribute much to our analysis.



So, created a bar chart in order to find the columns which needs to be removed in the application_data so that we can do the analysis perfectly.

Similarly made the bar chart for the previous_application excel sheet for the columns which had more than 30% Null/Missing values:



Task B – Identify Outliers in the Dataset:

Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

For this task, we need to find out the Quartile, Inter Quartile Range (IQR), Upper Limit, Lower Limit. Hence, I used the excel inbuilt formulas such as:

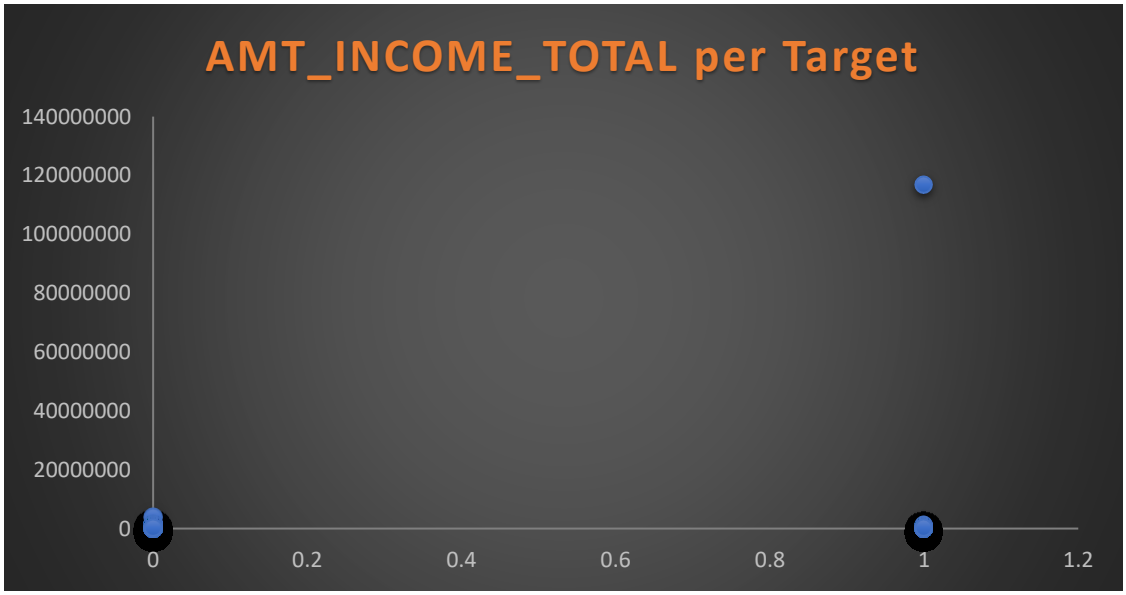
Quartile 1: =QUARTILE.INC(B2:B27319,1)

Quartile 3: =QUARTILE.INC(B2:B27319,3)

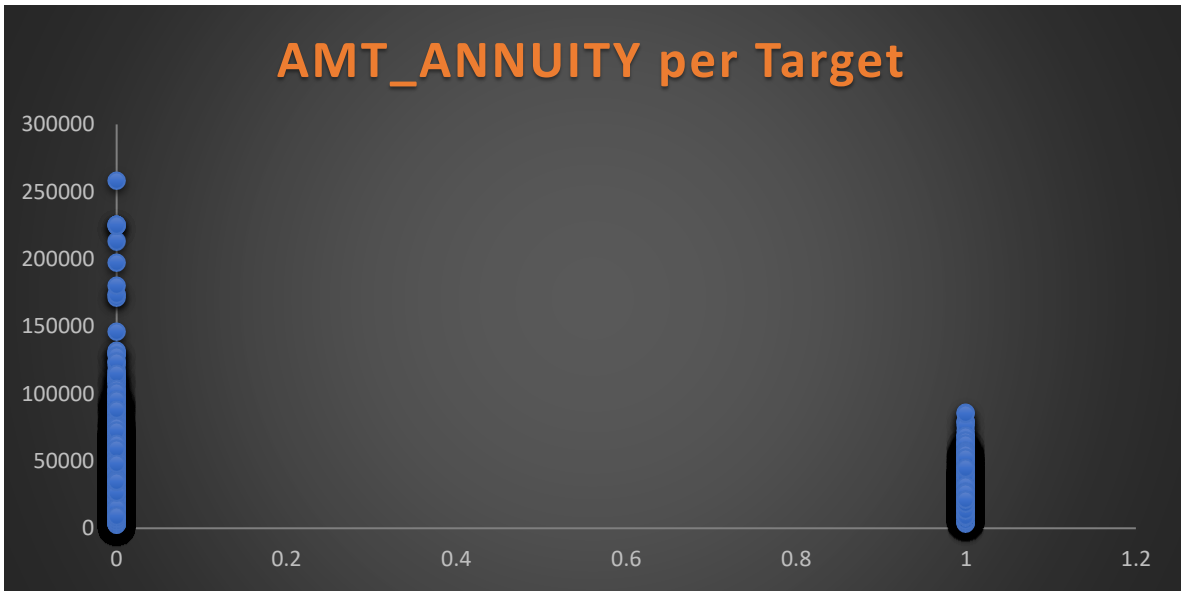
Inter Quartile: Q3 – Q1

Upper Limit: Q3 + (1.5 * IQR)

Lower Limit: Q3 - (1.5 * IQR)



Created a scatter plot to find out the outliers with Target and Total Income amount and could see that there is an outlier for the target 1. Likewise did the same to the column Annuity amount and could see that there are some outliers in target 0.



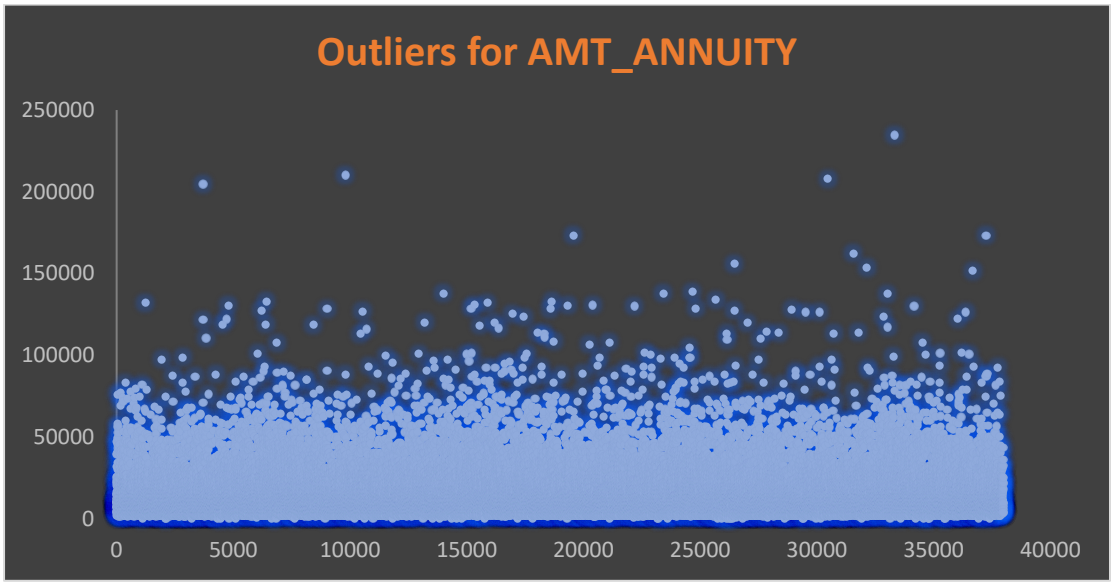
The Descriptive Analysis for Total Income Amount is as follows:

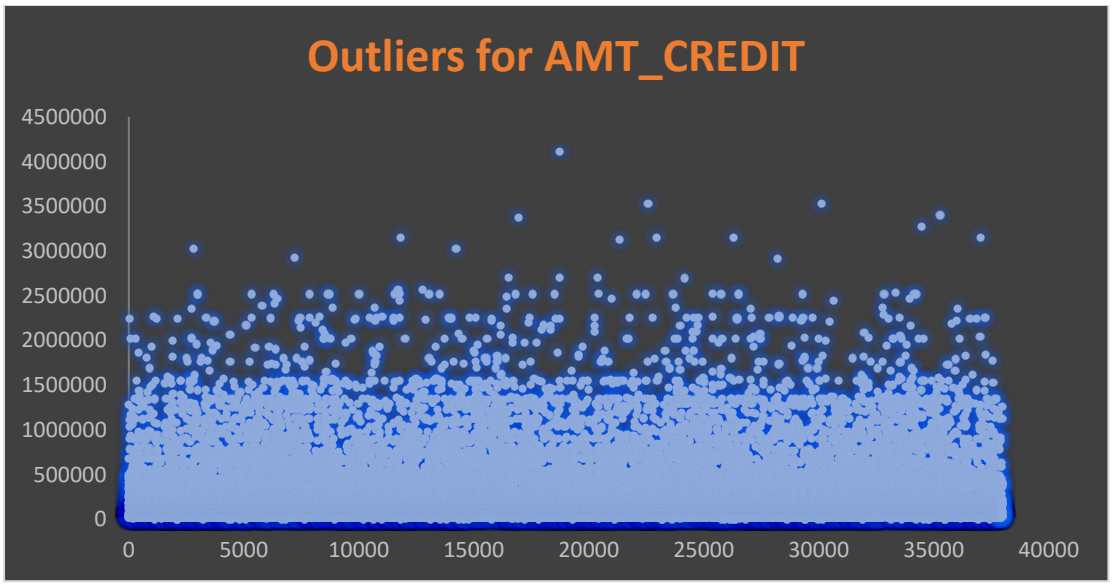
Mean	182906.4
Median	157500
Mode	135000
Std. Dev.	713802.238
Variance	5.09514E+11
Min	27000
Max	117000000
Count	27318

And the descriptive analysis for Annuity income is:

Mean	28001
Median	26145
Mode	9000
Std. Dev.	14637.12784
Variance	214245511.3
Min	2754
Max	258026
Count	27318

And for the previous_application data:





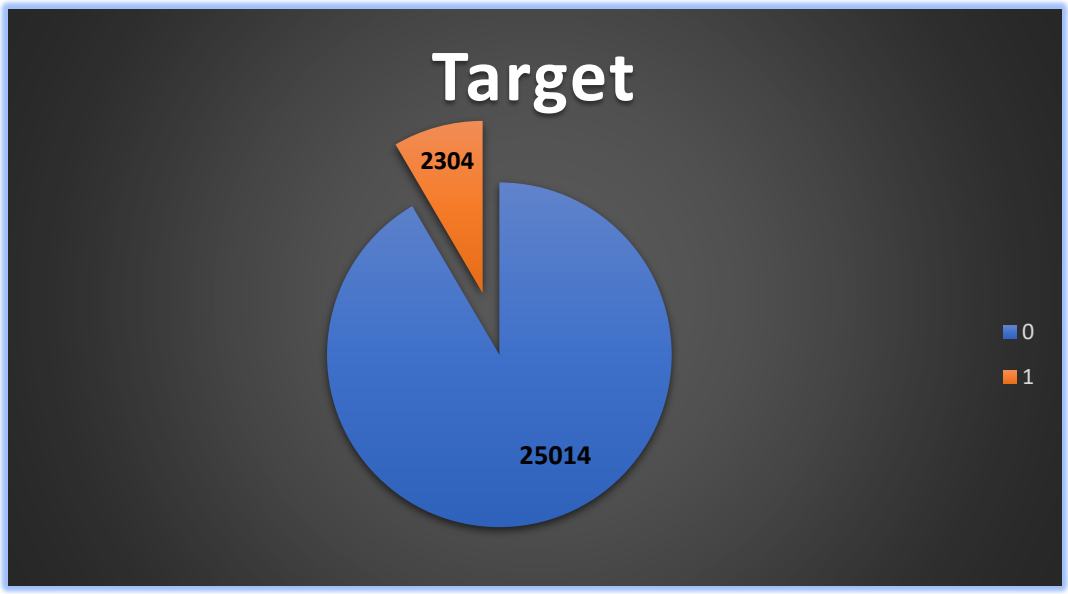
Found the count of the occurrences of the columns Annuity amount and Credit Amount and plotted the Scatter plot to find the outliers.

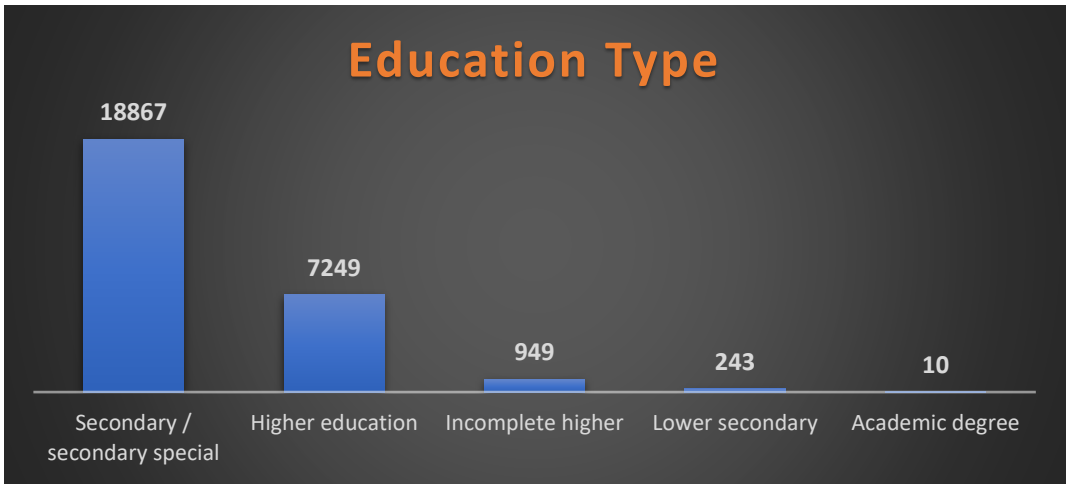
Task C – Analyse Data Imbalance:

Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

Used the formulas countif for the target and found there was a huge Data imbalance in the column Target.

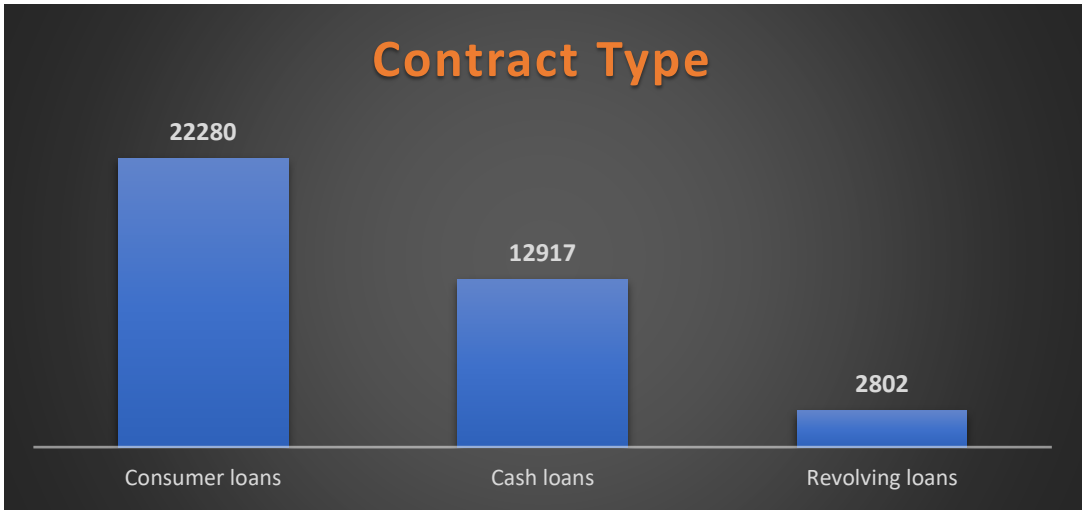
Also did the same to the column Education Type to find out if any Data Imbalance is there in this column and the charts for the columns are as below:





From the above Column chart, it is evident that the target 1 is so less when compared to the target value 0.

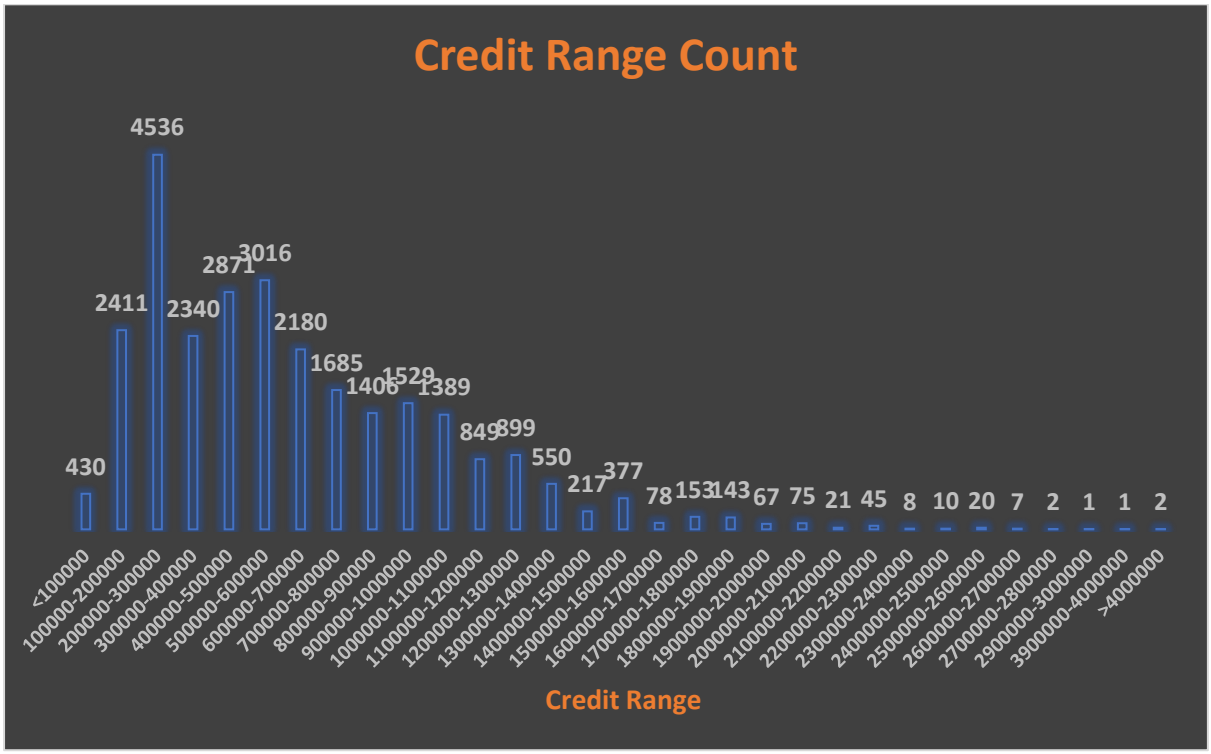
Did the same analysis for the previous application data and found out the data imbalance in the column contract type.



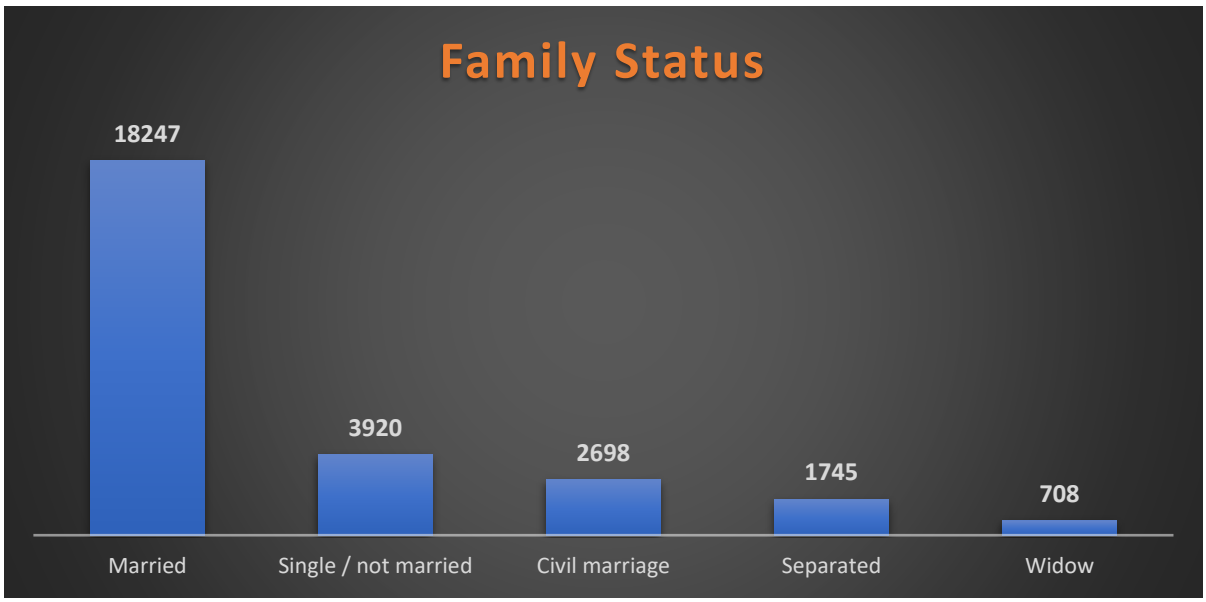
From the Column chart we could see that the Consumer Loans are more in number than the Cash loans and revolving loans.

Task D – Perform Univariate, Segmented Univariate, and Bivariate Analysis:

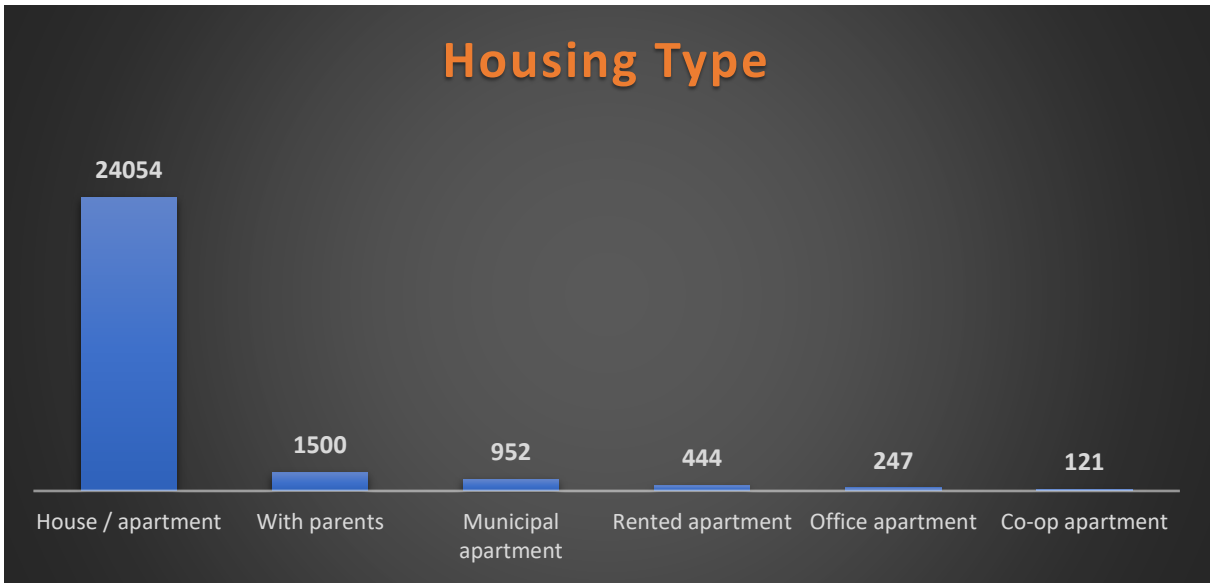
Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.



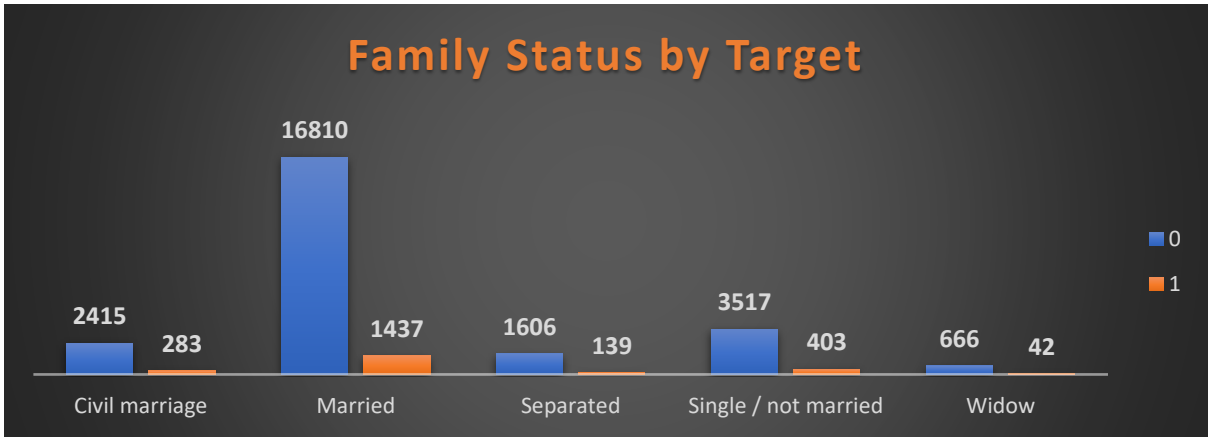
From the above Column chart, we could see that the adults within the credit range 1 lakh to 14 lakhs group tend to take the loan more than the other credit ranges.

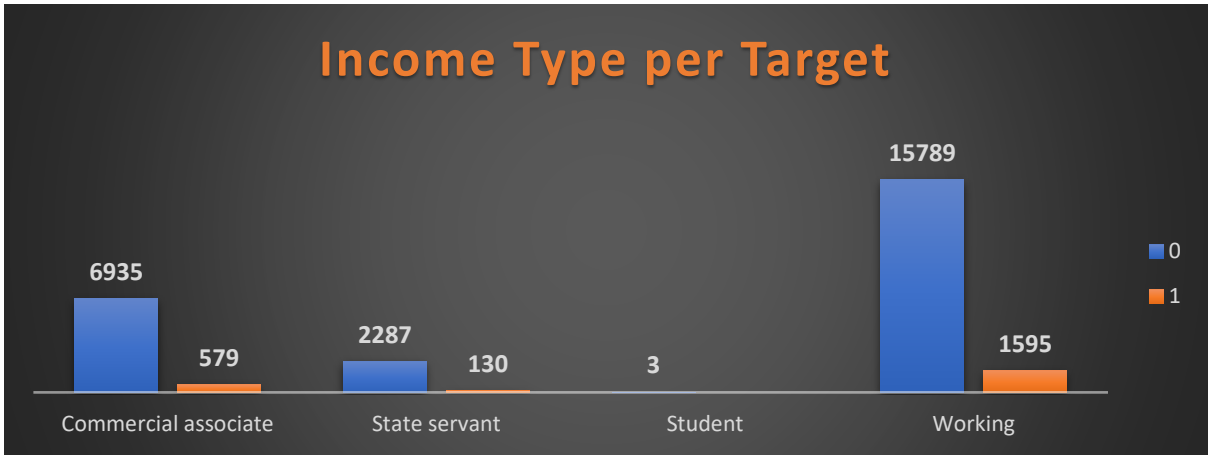


From the above column chart, it is evident that a greater number of adults are taking loan from the Married – family status as they do have some additional possibility to take a loan than the other segments.

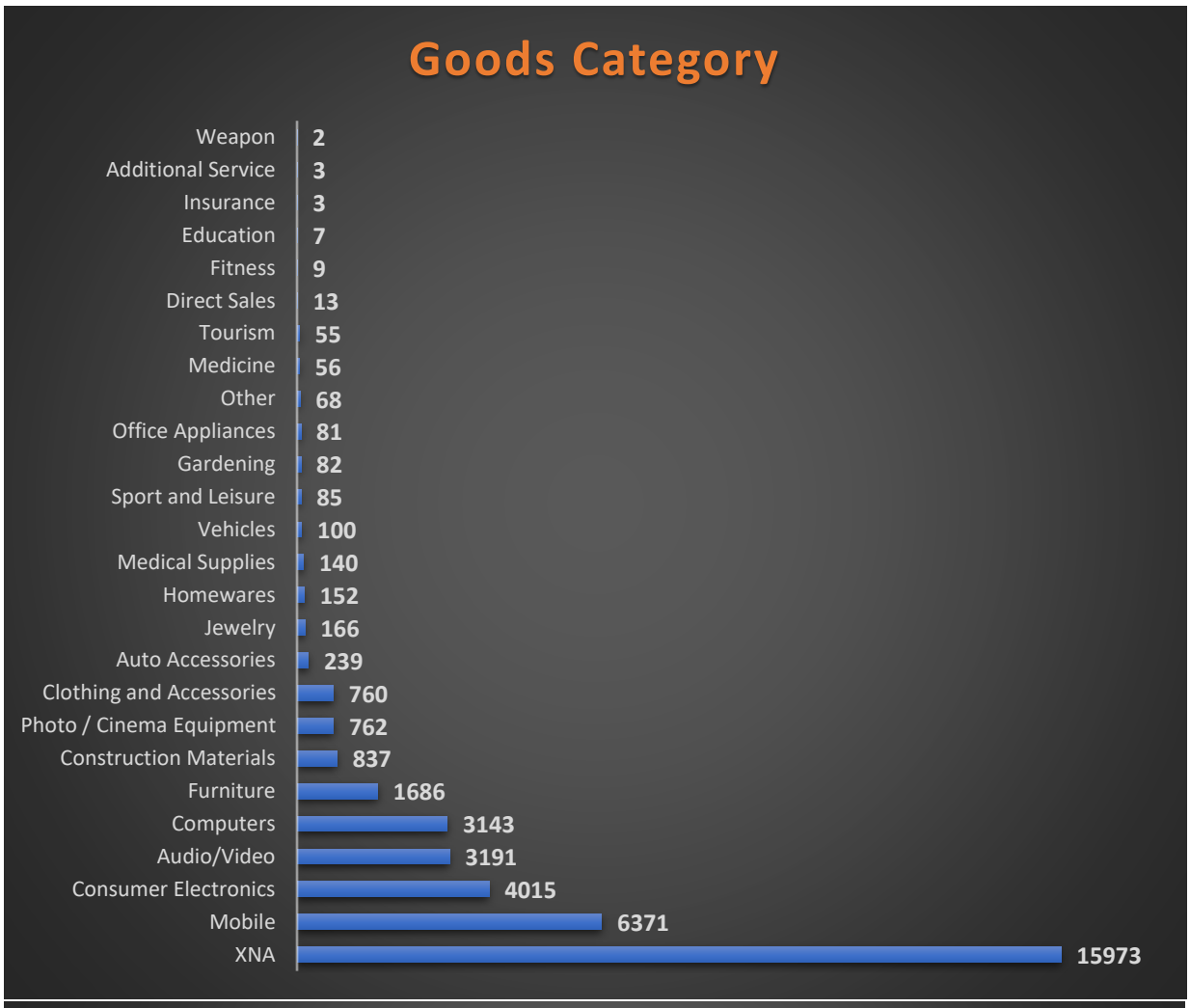


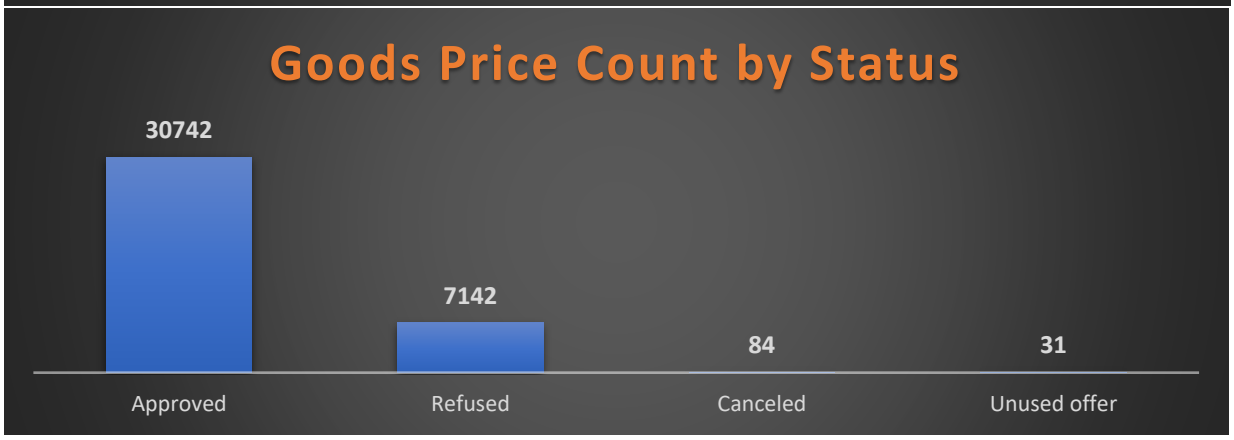
With respect to the housing type, we can say that the adults who live in the House/Apartment tend to take the loan than the people living in other range.





Did the bivariate analysis with the target column compared to the Family status and income type.





In the previous application data, did the analysis for the column's Goods price and Credit amount, and could see that the Approved status is more and the most goods priced bought is XNA.

We can come to the conclusion that adults who are working and who are married have a higher chance to take the loan.

Task E – Identify Top Correlations for Different Scenarios:

Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

Column Names	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	Days_Birth_Yrs	Days_Employed_Yrs	Days_Id_Publish_Yrs	REGION_RATING_CLIENT
CNT_CHILDREN	1	-0.004911747	-0.016013487	-0.026197114	-0.2557575	-0.070309609	0.129553223	0.035395697
AMT_INCOME_TOT	-0.004911747	1	0.365279441	0.179846628	0.054125487	0.026999244	0.014928611	-0.20806829
AMT_CREDIT	-0.016013487	0.365279441	1	0.093268561	0.160451708	0.089594989	0.034060456	-0.107726011
REGION_POPULATI	-0.026197114	0.179846628	0.093268561	1	0.044620857	-0.010415056	0.000656732	-0.523154439
Days_Birth_Yrs	-0.2557575	0.054125487	0.160451708	0.044620857	1	0.345551383	0.072472675	-0.045952464
Days_Employed_Yr	-0.070309609	0.026999244	0.089594989	-0.010415056	0.345551383	1	0.064595883	0.017965584
Days_Id_Publish_Y	0.129553223	0.014928611	0.034060456	0.000656732	0.072472675	0.064595883	1	-0.002768905
REGION_RATING_C	0.035395697	-0.20806829	-0.107726011	-0.523154439	-0.045952464	0.017965584	-0.002768905	1

Columns	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	CNT_PAYMENT
AMT_ANNUITY	1	0.825528512	0.818147682	0.825564271	0.394100654
AMT_APPLICATION	0.825528512	1	0.993466353	0.999901663	0.663825921
AMT_CREDIT	0.818147682	0.993466353	1	0.993444101	0.69334727
AMT_GOODS_PRICE	0.825564271	0.999901663	0.993444101	1	0.663684765
CNT_PAYMENT	0.394100654	0.663825921	0.69334727	0.663684765	1

Result:

Through this project I was able to understand the formulas being used in the Excel which can be used to find the Correlation and various charts on how to use them. I got used to the Excel formulas and how to convert the Raw Data into meaningful insights. And the steps which I used are – cleansing the data and using the formulas to find the desired outcome and also learnt how to convert the data into a visualized chart so that the insights can be drawn within seconds by seeing the graphs instead of searching the whole data.

As a result, we could summarize as there is higher possibility for the adults who fall in the category:

1. Married
2. Educated
3. Strong Work Experience
4. Previously Approved Clients

The people who don't tend to take loan falls in the category:

1. Unemployed
2. Youth
3. Less Work Experience
4. Previously Unapproved Clients

I have achieved the end result and I think I have contributed my full support into the Analysis. I hope this project helps the Analysis and it achieves what it was tend to achieve.

Hyperlink for the Excel sheet:

Bank Loan Case Study Excel File Link