Bank Loan Case Study

Project Description

This project's primary goal is to find trends that suggest if a client would struggle to make their installment payments. This information can be used to make decisions such as denying the loan, reducing the amount of the loan, or lending at a higher interest rate to risky applicants. To improve its decision-making about loan approval, the organization seeks to understand the main factors contributing to loan default.

Approach

Understanding Data: Before analysis, it is important to understand the data. Look at the structure of data and get a sense of overall content, This helps in identifying any potential issues or challenges that need to be addressed.

Cleaning Data: Once any issue is identified in data it is important to resolve and clean the data for better analysis. This includes handling missing or incomplete data, identifying and handling outliers, and correcting text mistakes.

Analyzing Data: Once the data is cleaned it is analyzed using various formulas and functions for finding out insights and answering the business questions.

Visualizing Data: After doing analysis the insights are visualized using various charts and plots, so that insights can be understood by stakeholders and better business decisions can be taken.

Tech stack

Used Microsoft Excel 2016 for cleaning, analyzing, and visualizing data. And used Google Docs to create reports.

Data Cleaning:

This is the most difficult and crucial step of any Data analysis project. The process included in this step varies from question to question and Dataset to Dataset.
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.

Insights

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.

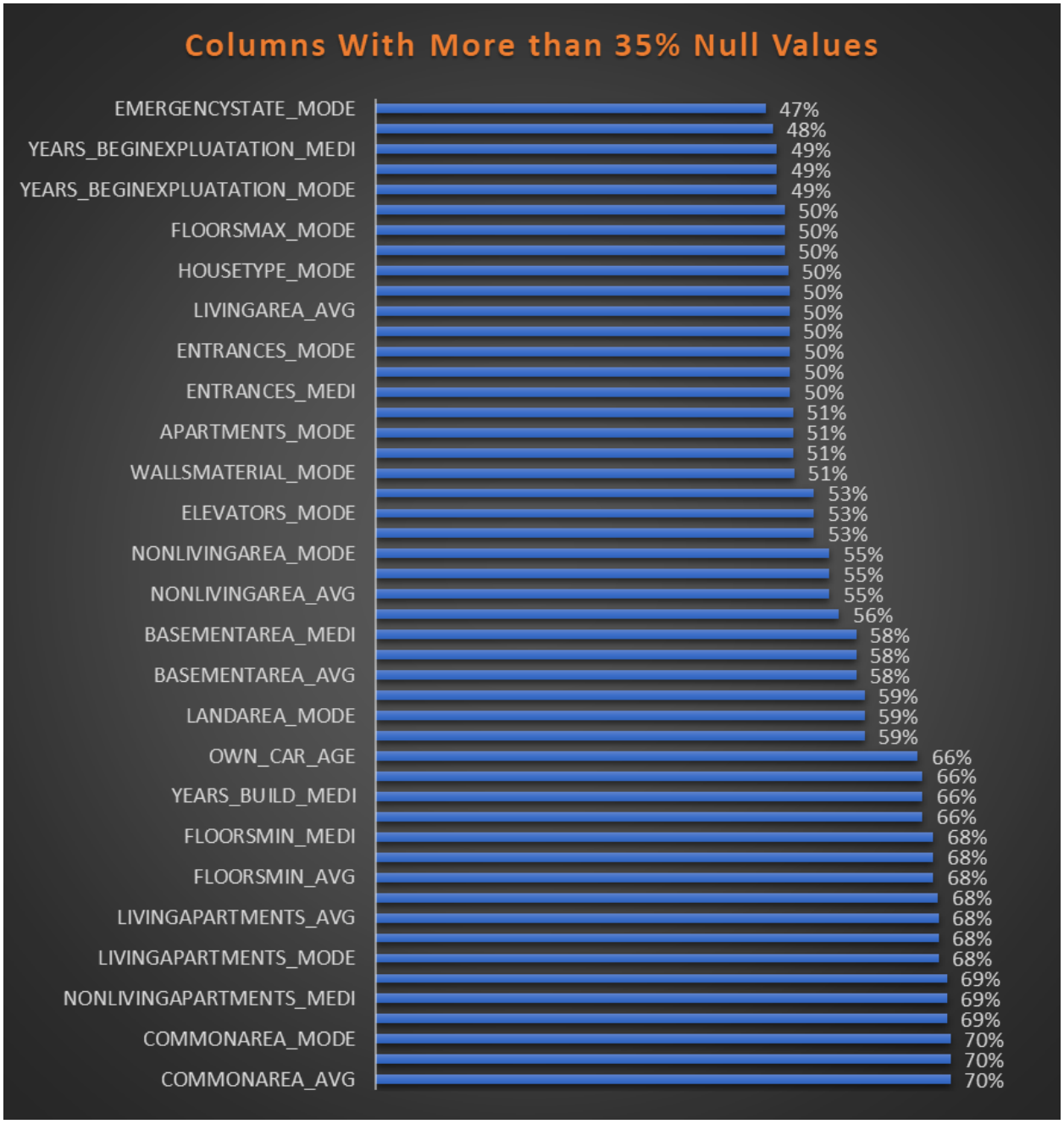
Used the formula CountA formula to find out the total number of values in each column:

=COUNTA(B4:B50002)

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

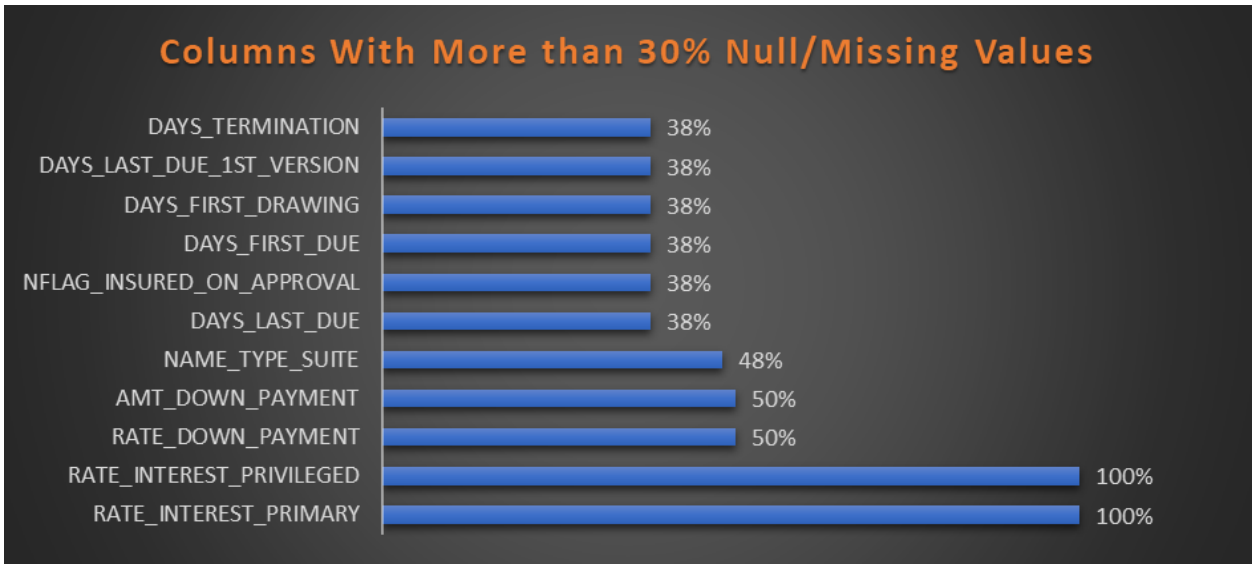
=1-B2/B2

Also decided to remove the columns that had more than **35%** of null values since they don't contribute much to our analysis and for the columns which have less than 35% of null values i have used **mean,median** for numerical data, and **mode** for categorical data.



Bar chart to find the columns which need to be removed in the application_data

Similarly, the previous_application columns which had more than **30%** Null/Missing values were dropped



Bar chart to find the columns which need to be removed in the pervious_application

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, I used Quartile, Inter Quartile Range (IQR), Upper Limit, and Lower Limit Excel's inbuilt formulas

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

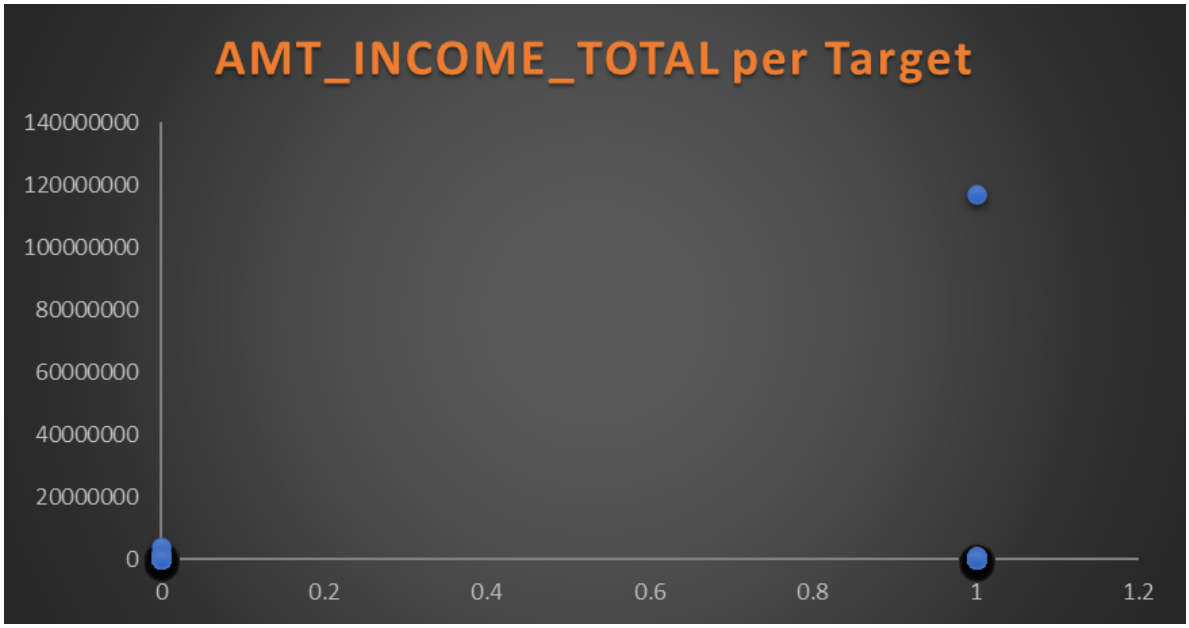
Quartile 3: =QUARTILE.INC(B2:B27319,3)

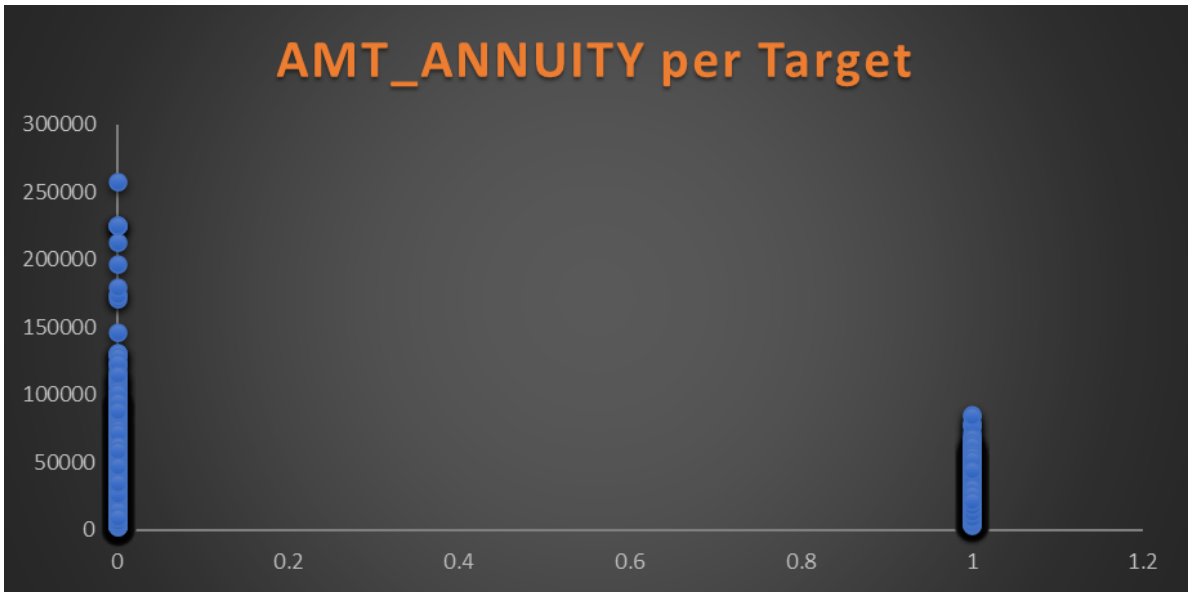
Inter Quartile: Q3 – Q1

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 Target 1. Likewise did the same to the column Annuity amount and could see that there are some outliers in target 0.


AMT_INCOME_TOTAL per Target
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
0
0.2
0.4
0.6
0.8
1
1.2


AMT_ANNUITY per Target
300000
250000
200000
150000
100000
50000
0
0
0.2
0.4
0.6
0.8
1
1.2

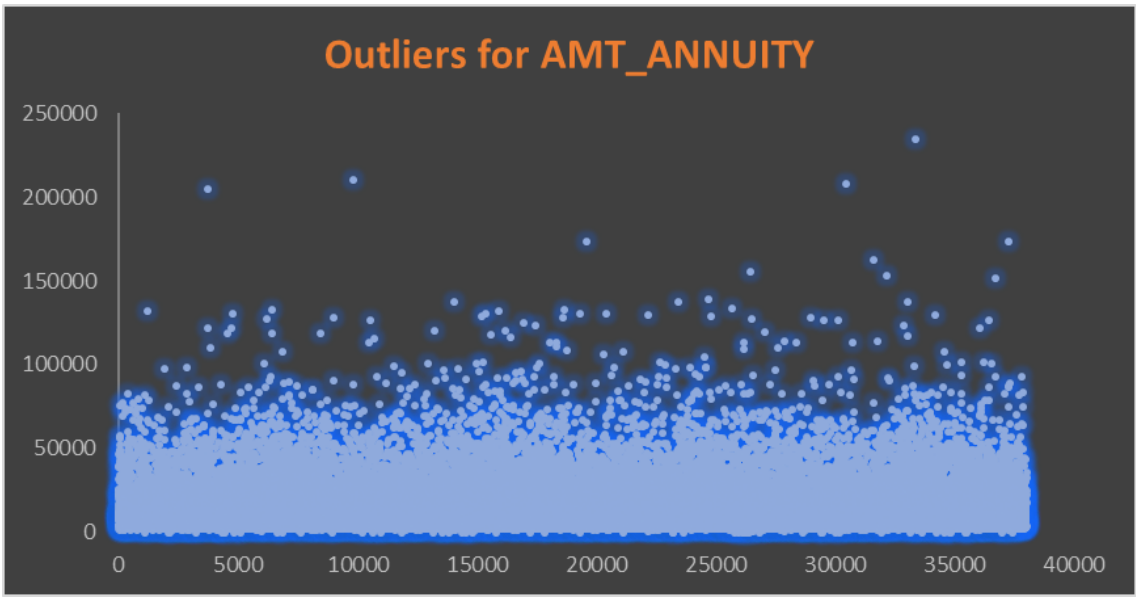
Descriptive Analysis for Total Income Amount:

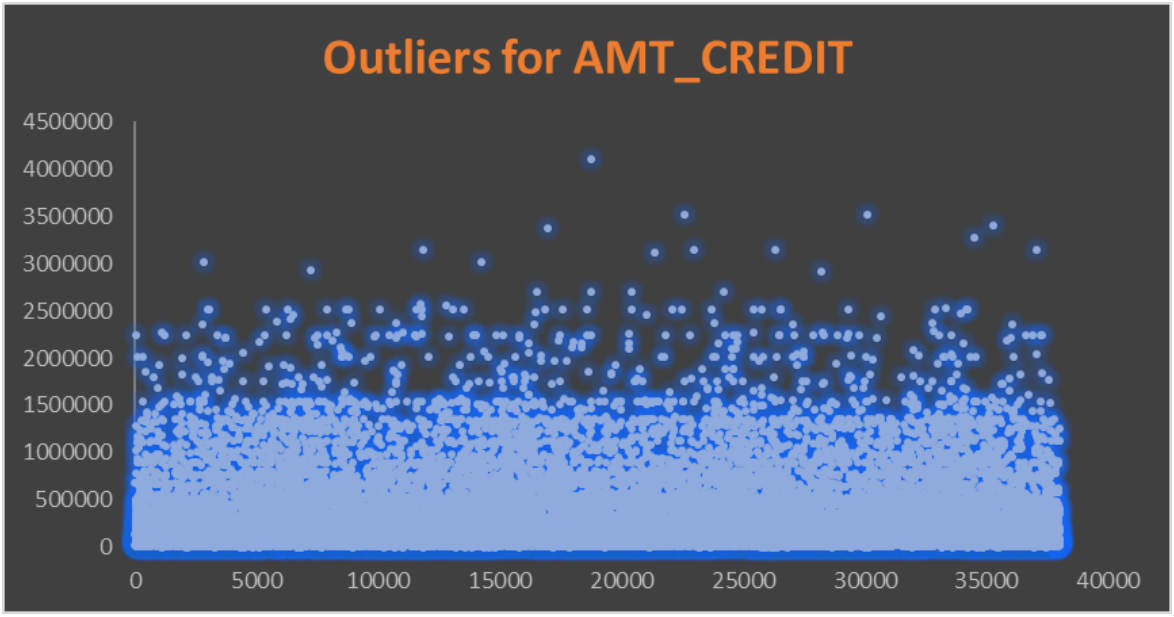
Mean	182906.4
Median	157500
Mode	135000
Std. Dev.	713802.238
Variance	5.09514E+11
Min	27000
Max	117000000
Count	27318

Descriptive analysis for Annuity income:

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

for the previous_application data:



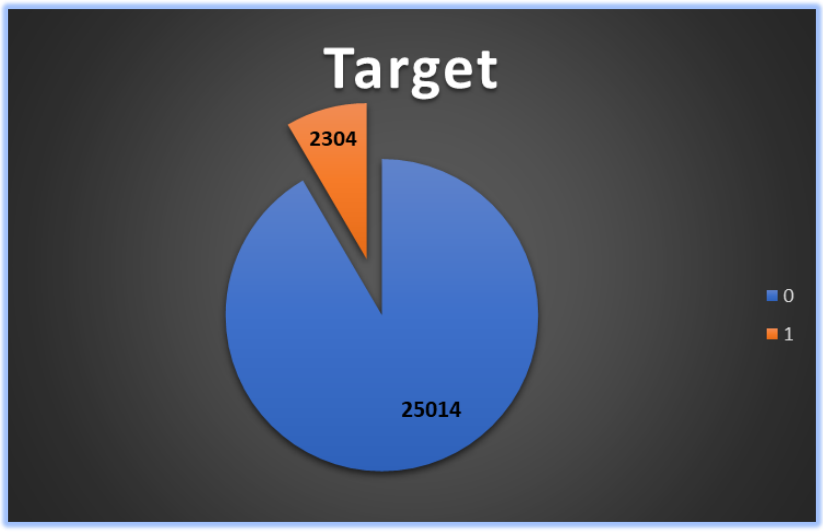


Found the count of the occurrences of the columns Annuity amount and Credit Amount and plotted the Scatter plot to find the outliers.

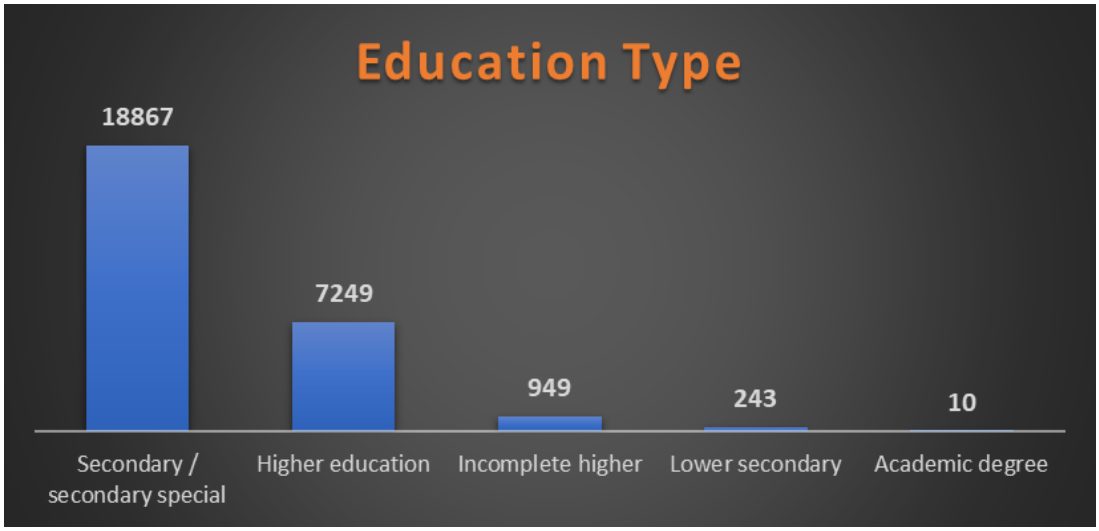
C) Analyse Data Imbalance: Determine if there is a data imbalance in the loan application dataset and calculate the ratio of data imbalance.

Used the formula **CountIF** for the target and found there was a huge Data imbalance in the column Target. Around **92%** of all the customers pay on time and only **8%** of customers default thus making data extremely imbalanced.

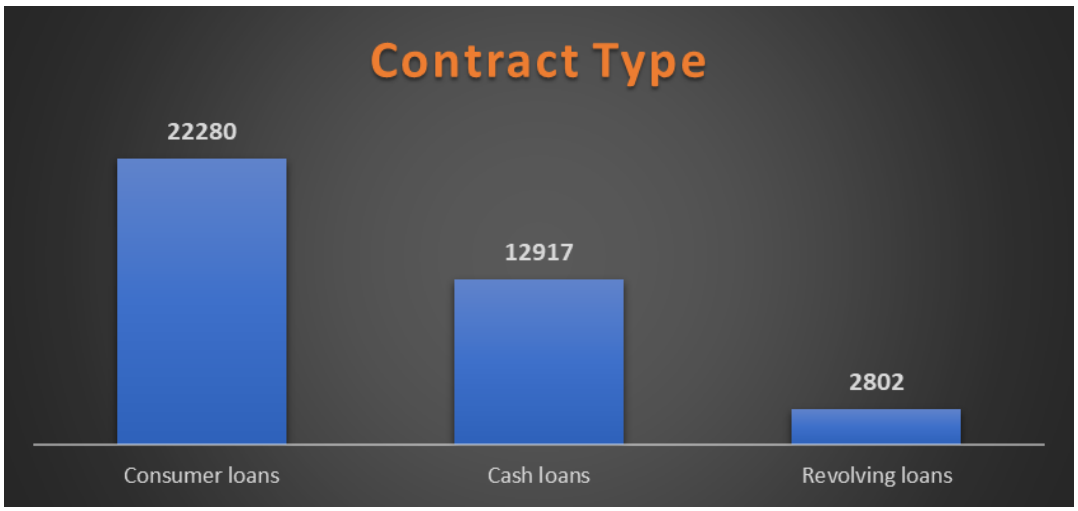
Target	Count of Targets	Contribution
0	25014	92%
1	2303	8%



Also did the same to the column Education Type



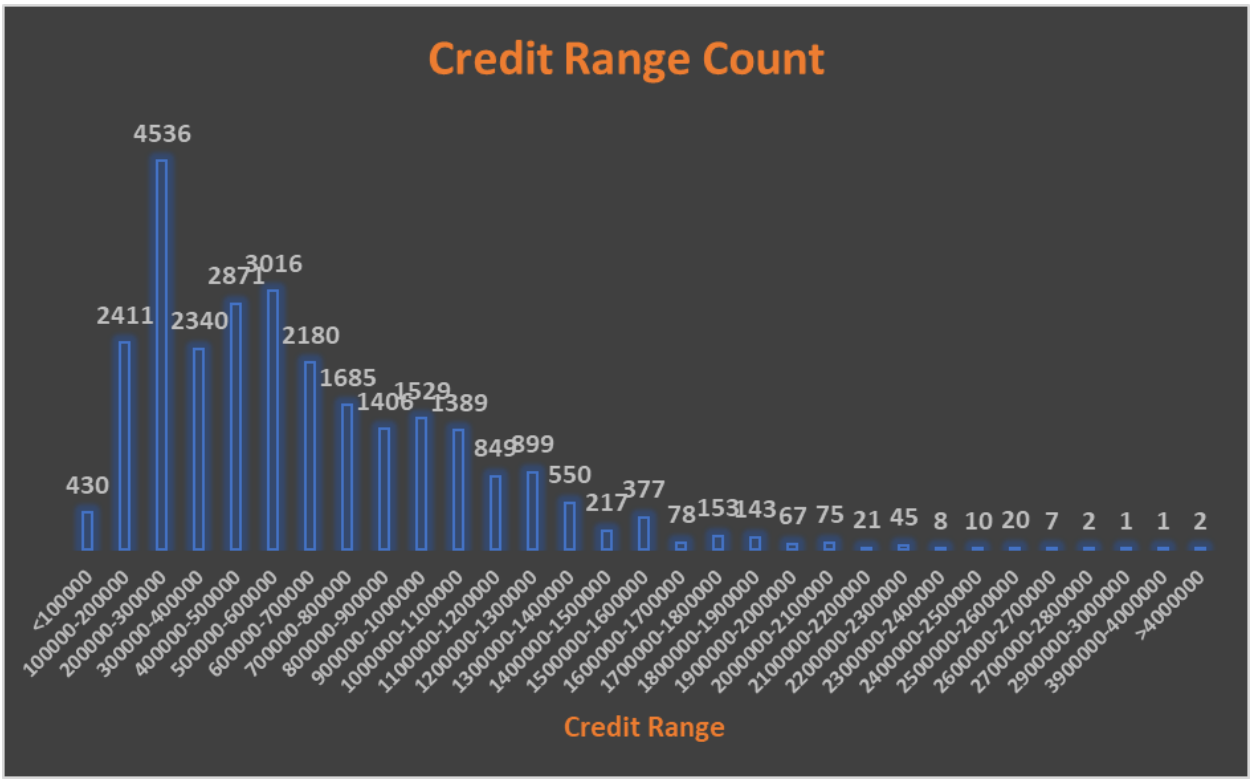
Did the same analysis for the previous application data and found out the data imbalance in the column contract type.



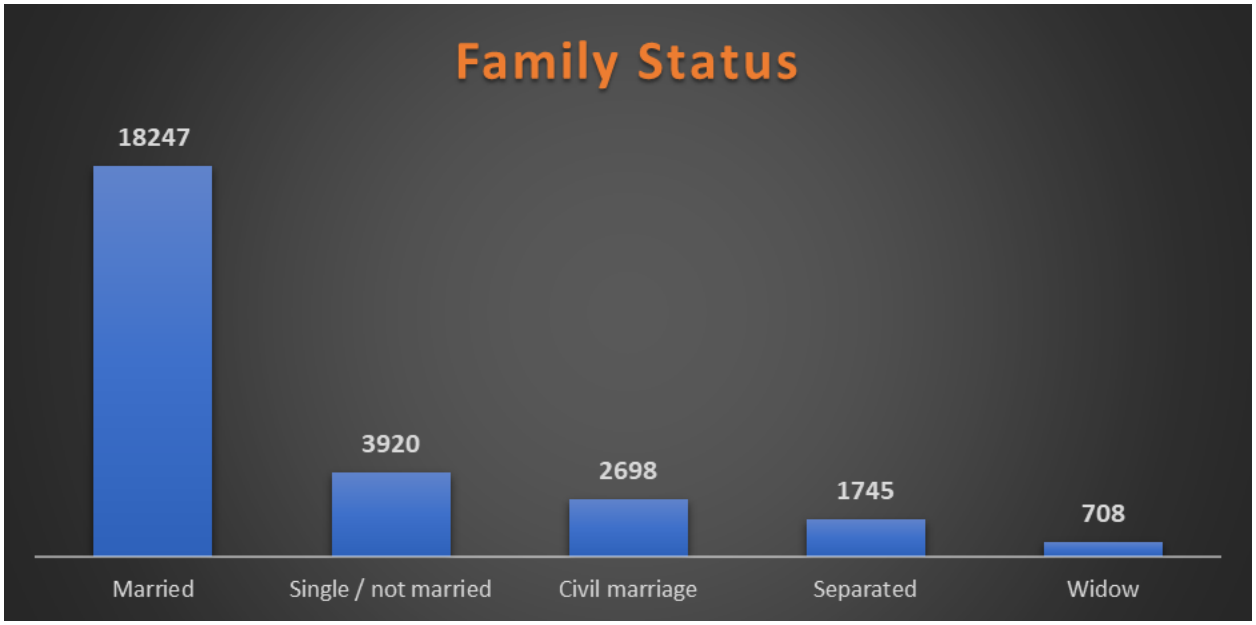
From the Column chart, we can see that the **Consumer Loans** are more in number than the Cash loans and revolving loans.

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.

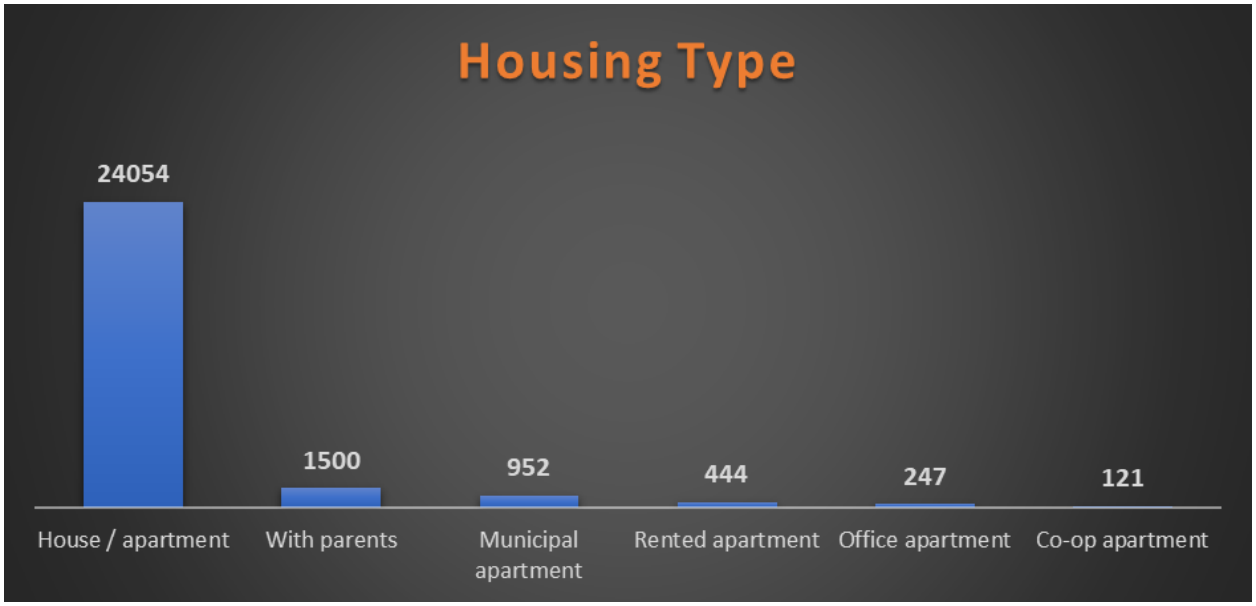
Univariate and Segmented Univariate Analysis:



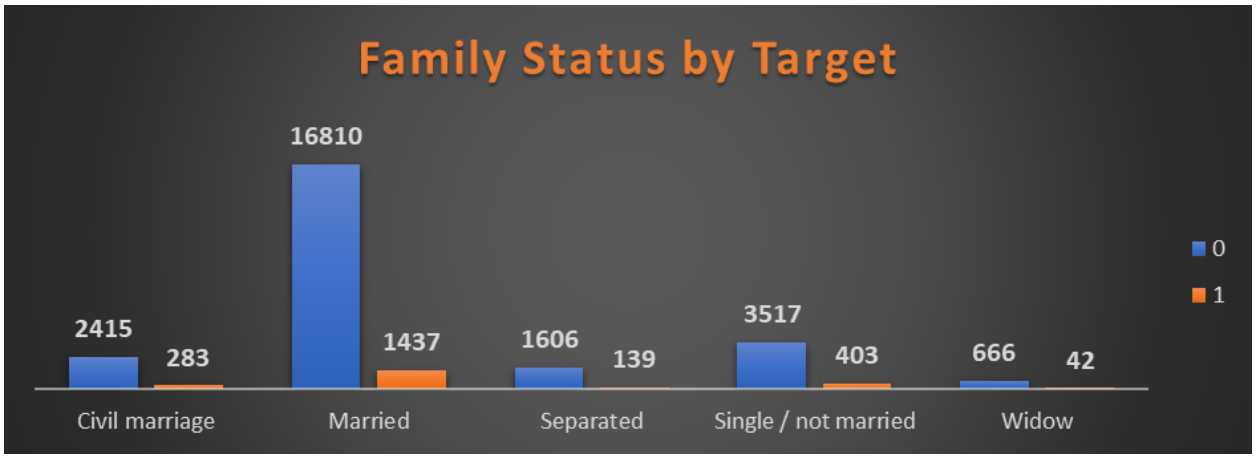
From the above Column chart, we can see that the adults within the credit range **1 lakh to 14 lakhs** group tend to take loans more than the other credit ranges.

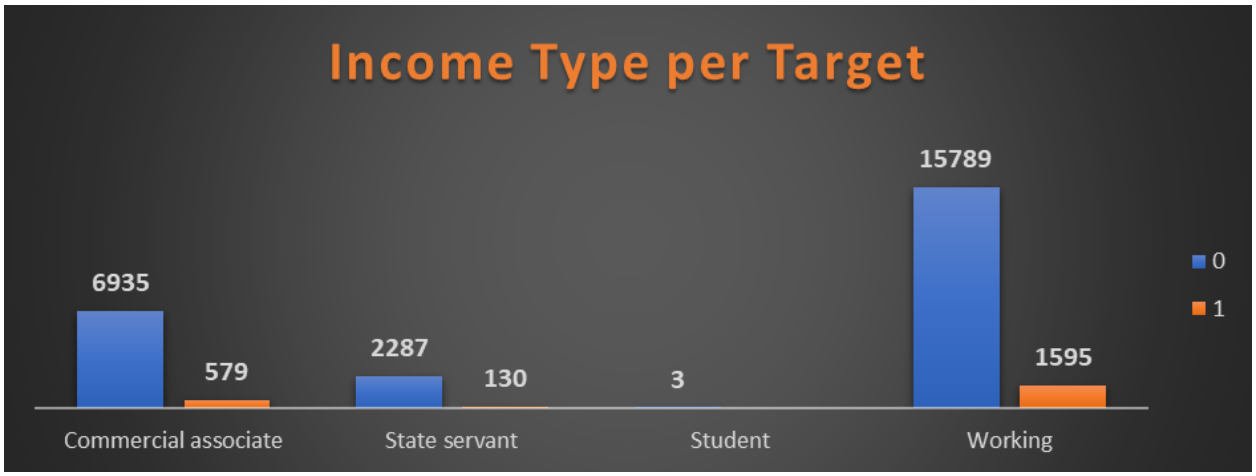


From the above column chart, it is evident that a greater number of adults are taking loans from the **married–-family status** as they do have some additional possibility to take a loan than the other segments.

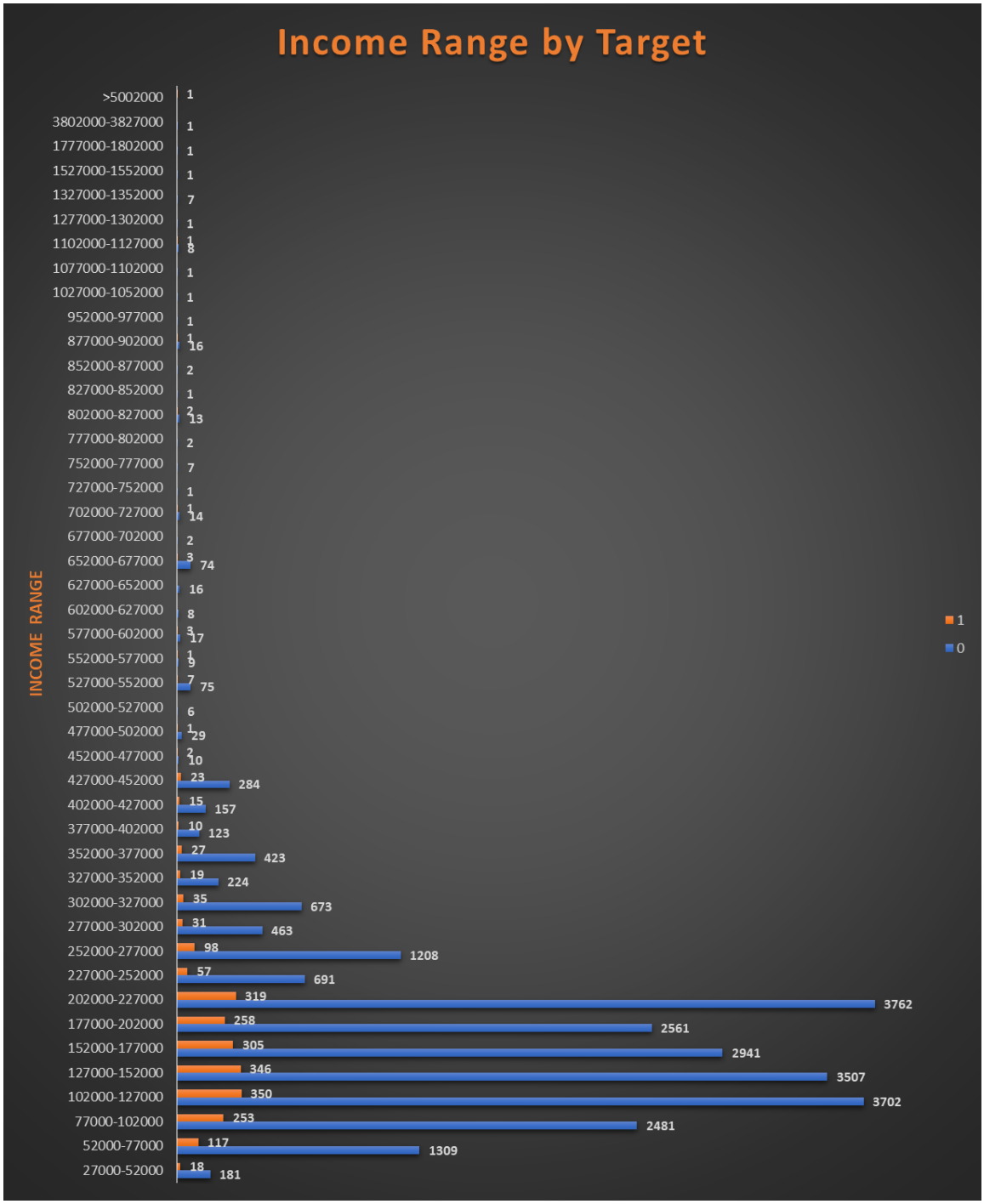


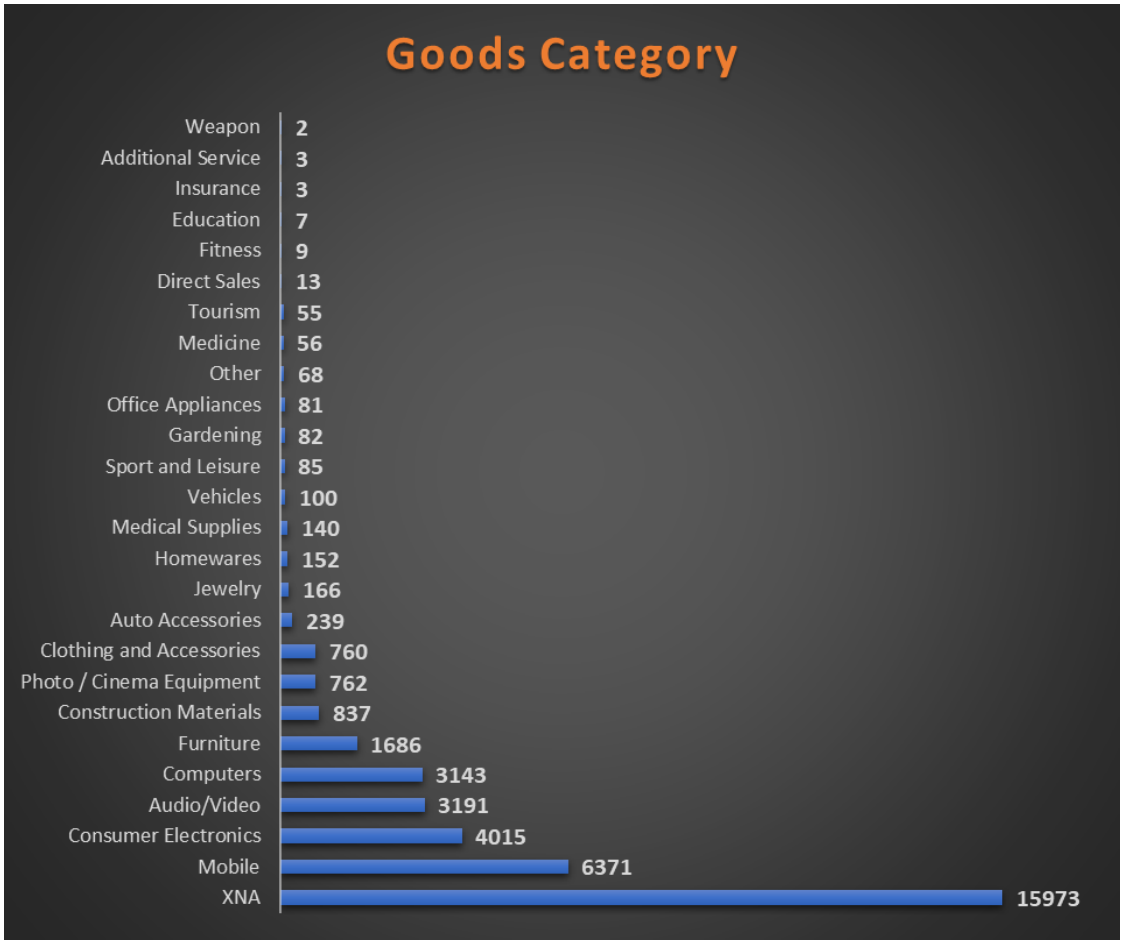
With respect to the housing type, we can say that the adults who live in the **House/Apartment** tend to take the loan than the people living in other ranges.

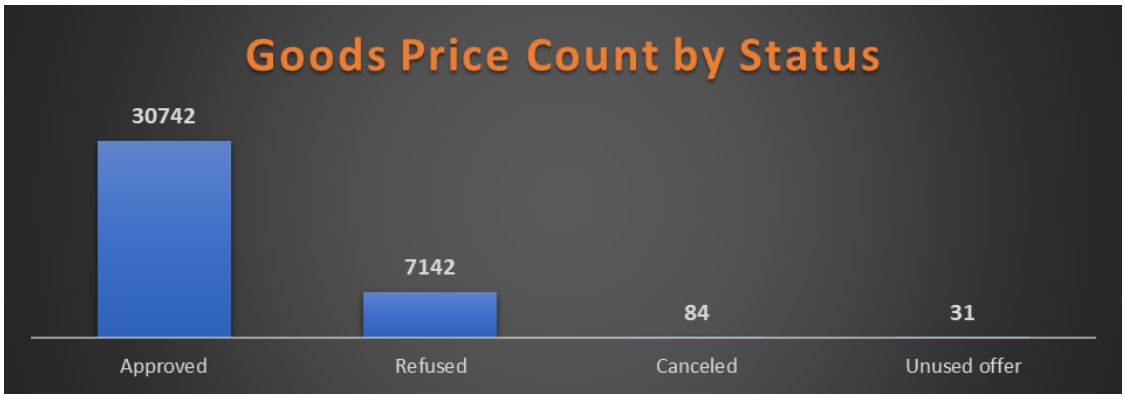




Bivariate analysis with the target column compared to the Family status and income type:


Income Range by Target
INCOME RANGE
>5002000 1
3802000-3827000 1
1777000-1802000 1
1527000-1552000 1
1327000-1352000 7
1277000-1302000 1
1102000-1127000 1 8
1077000-1102000 1
1027000-1052000 1
952000-977000 1
877000-902000 1 16
852000-877000 2
827000-852000 1
802000-827000 2 13
777000-802000 2
752000-777000 7
727000-752000 1
702000-727000 1 14
677000-702000 2
652000-677000 3 74
627000-652000 16
602000-627000 8
577000-602000 3 17
552000-577000 1 9
527000-552000 7 75
502000-527000 6
477000-502000 1 29
452000-477000 2 10
427000-452000 23 284
402000-427000 15 157
377000-402000 10 123
352000-377000 27 423
327000-352000 19 224
302000-327000 35 673
277000-302000 31 463
252000-277000 98 1208
227000-252000 57 691
202000-227000 319 3762
177000-202000 258 2561
152000-177000 305 2941
127000-152000 346 3507
102000-127000 350 3702
77000-102000 253 2481
52000-77000 117 1309
27000-52000 18 181
1
0





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 conclude that adults who are **working** and who are **married** have a higher chance of taking a loan.

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.

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_POPULAT	-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

We collected, cleaned, and analyzed data to gather useful insights and visualize them for better understanding.

Conclusions from insights:

A) We could summarize that there is a higher possibility for adults who take loans fall into the category:

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

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

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

B) Our data is extremely imbalanced with **92%** of customers paying on time and only **8%** of the default.

C) Customers within the credit range **1 lakh to 14 lakhs** group tend to take loans more than the other credit ranges. Maximum Customers take loans within the credit range of **2 - 3 lakh** with a count of **4536.**

D) Customers who are **married** and have a family are taking more loans.

E) Customers who live in a House/Apartment tend to take more loans.

F) The most good-priced bought is **XNA**.

Excel file:
Bank Loan Case Study Trainity Project 6(Jaskirat Singh)).xlsx – application_data
previous_application.xlsx

Jaskirat Singh