Bank Loan Case Study

Sanjana Kumari Yadav

Agenda

- Identify Missing data with it Appropriately
- Identify Outliers in the dataset
- Analyze Data Imbalance
- Perform Univariate, Segmented Univariate and Bivariate Analysis
- Identify Top Correlation for Different Scenarios

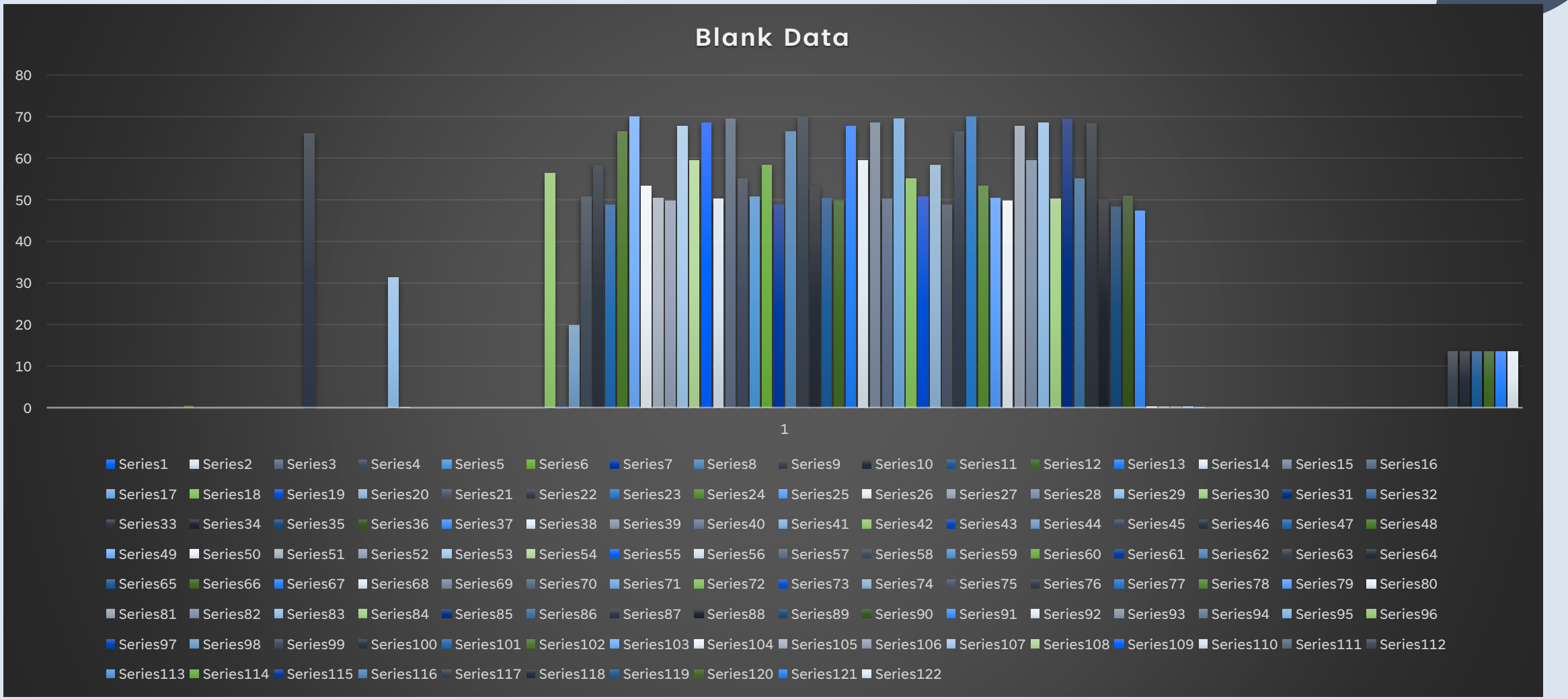
Identify Missing Data and Deal with it Appropriately

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

Solution: Here we use the countif function all over the column. We can delete that column whose value less than 34.33% as we considering median over mean.

=(COUNTBLANK(DM2:DM50000)/COUNT($A$2:$A$50000))*100

Blank space Data

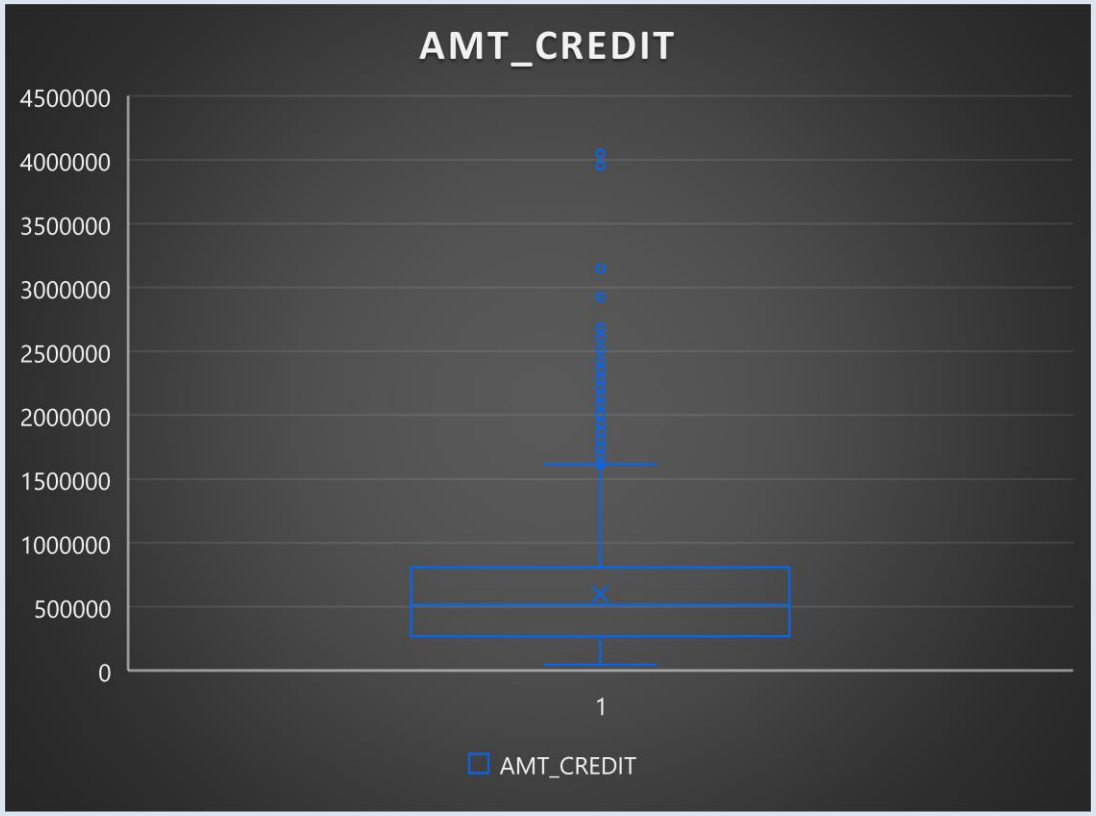


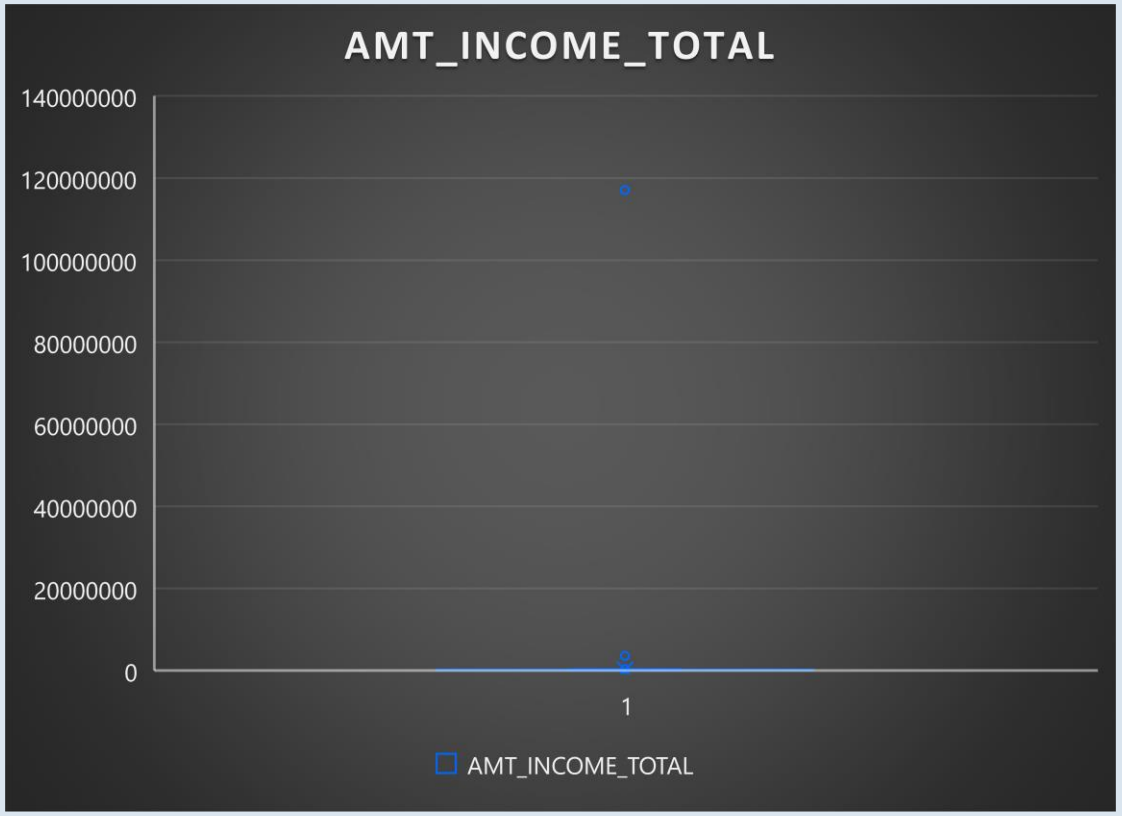
Identify Outliers in the Dataset

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

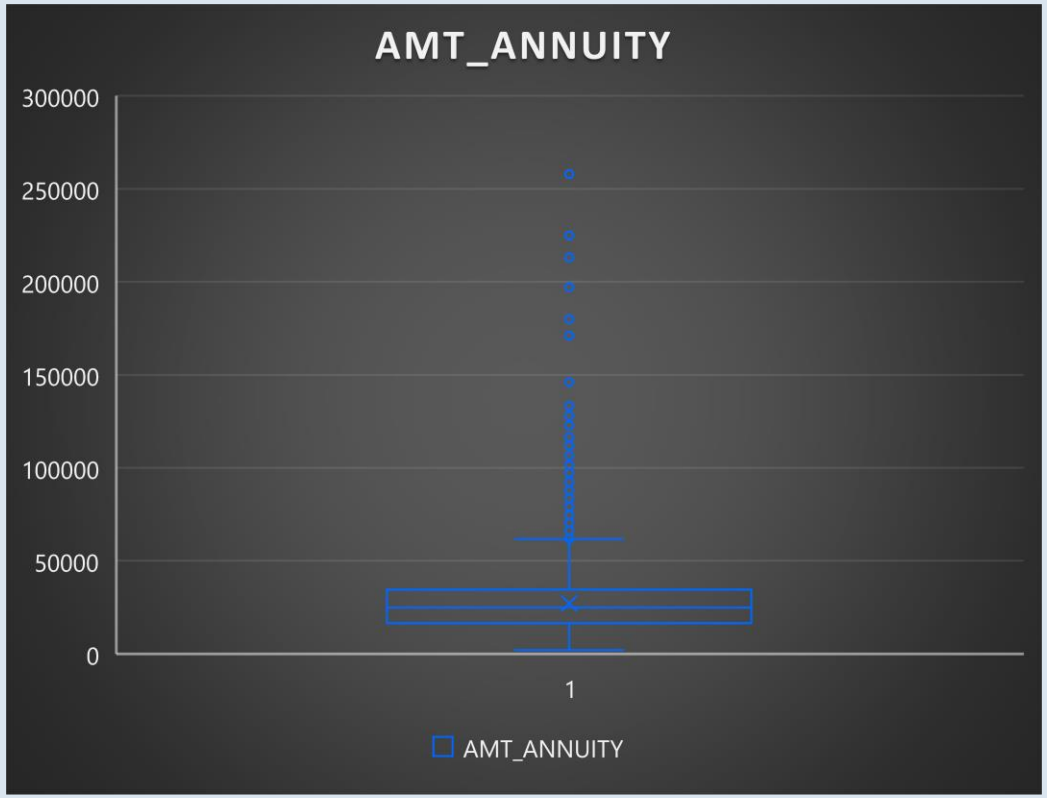
Solution: Here I plotted a box graph using the data with AMD_INCOME_TOTAL , AMT_CREDIT , CNT_FAM_MEMBERS , AMT_ANNUITY , AMT_GOODS_PRICE.

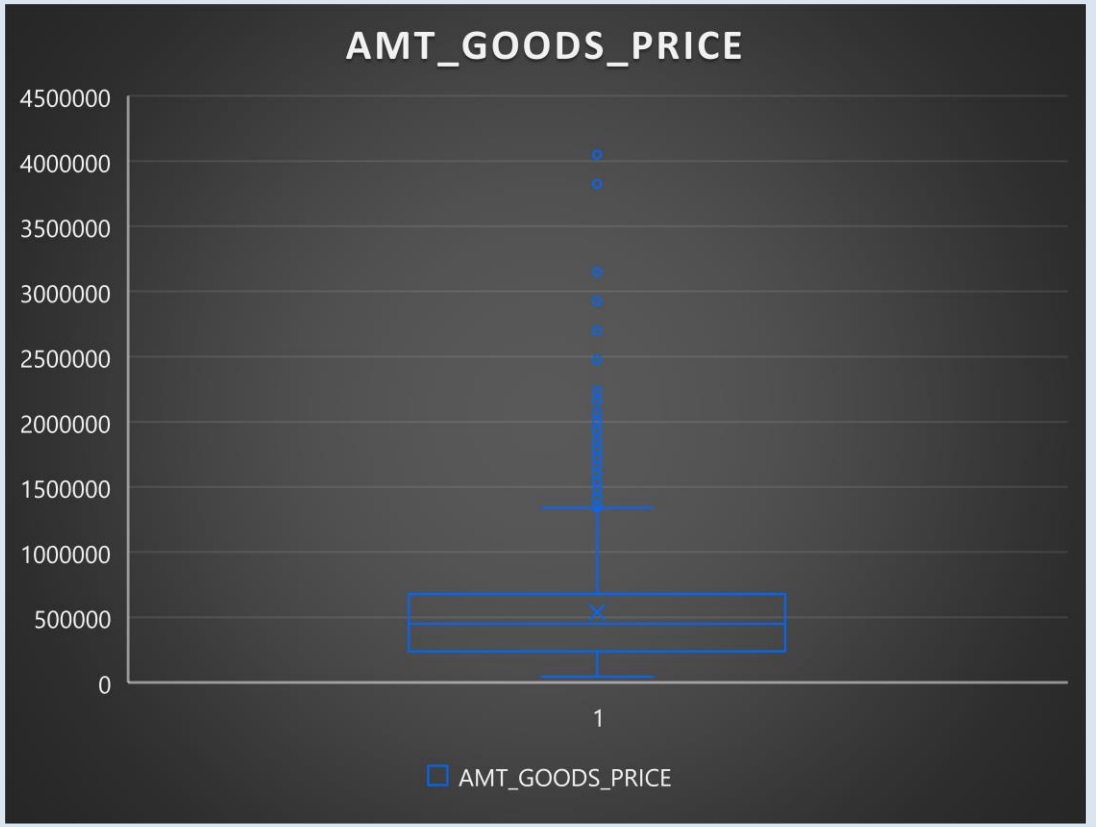
Outliers in the Dataset



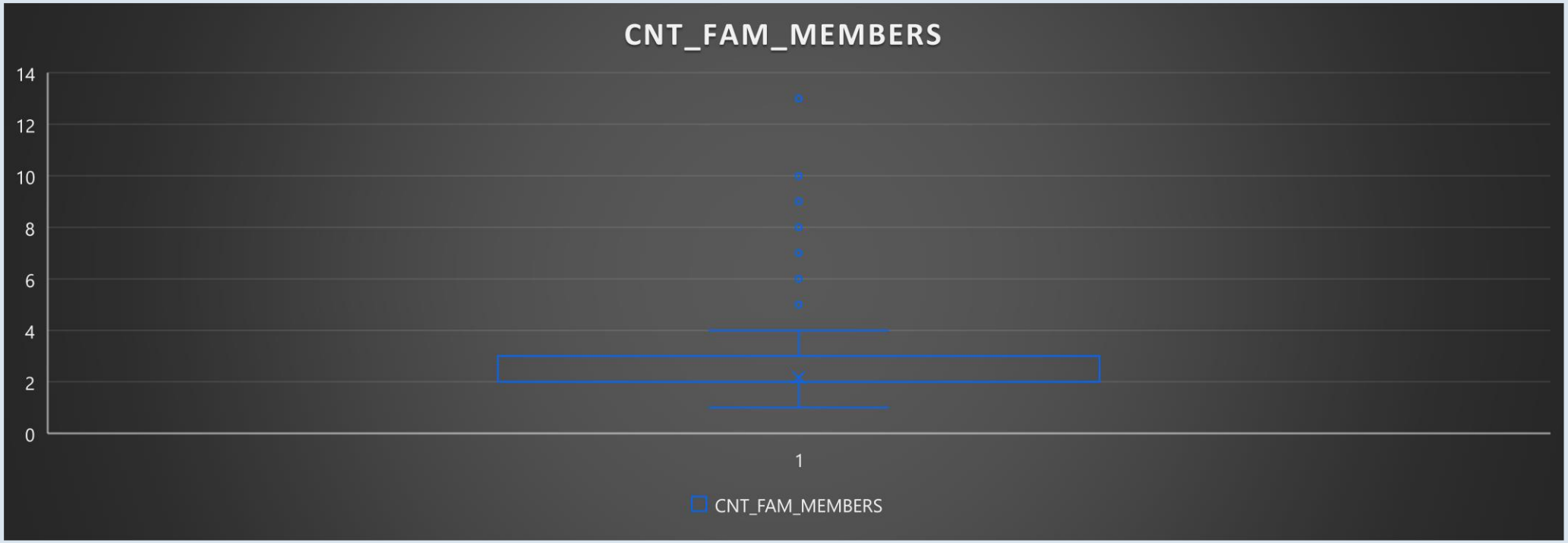


Outliers in the Dataset





Outliers in the Dataset

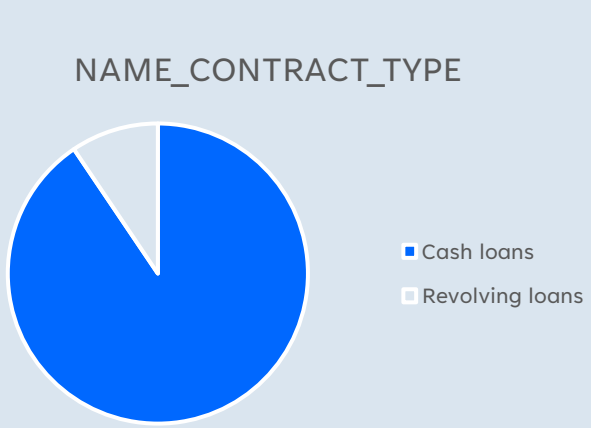


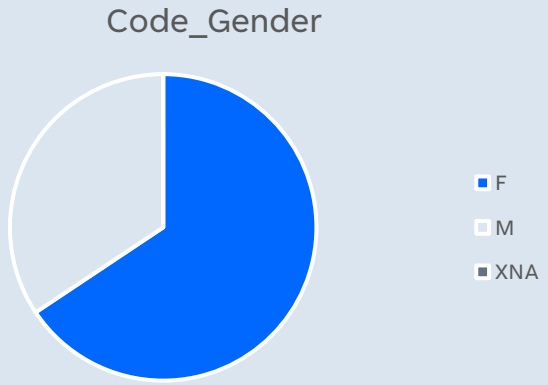
Analyse Data Imbalance

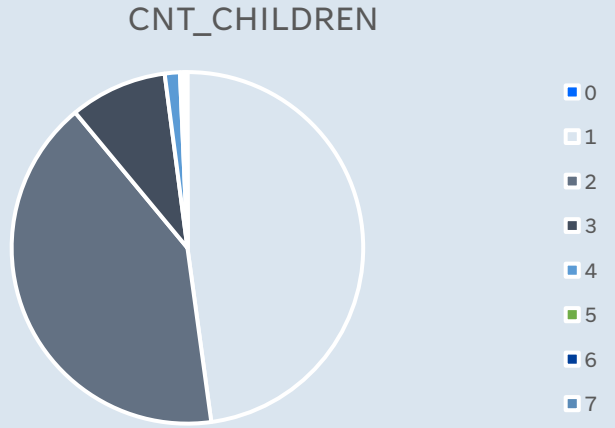
Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

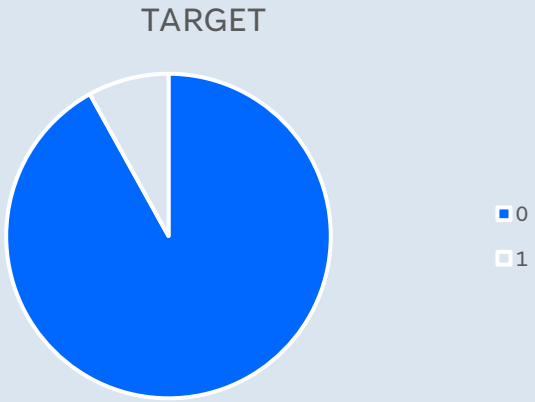
Solution: For finding imbalance in the given data, we use pivot table and make pie chart to represent the ratio of data imbalance and we also use count function.

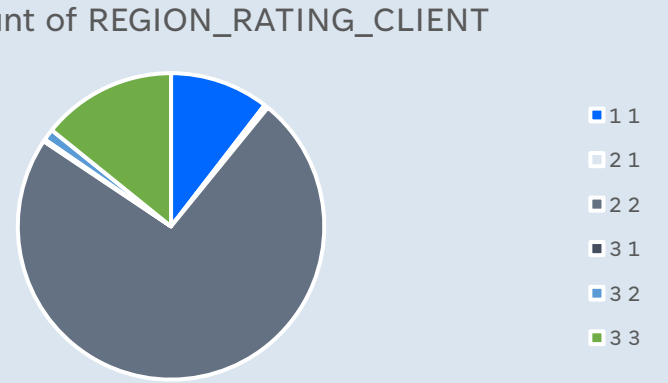
Analyse Data Imbalance











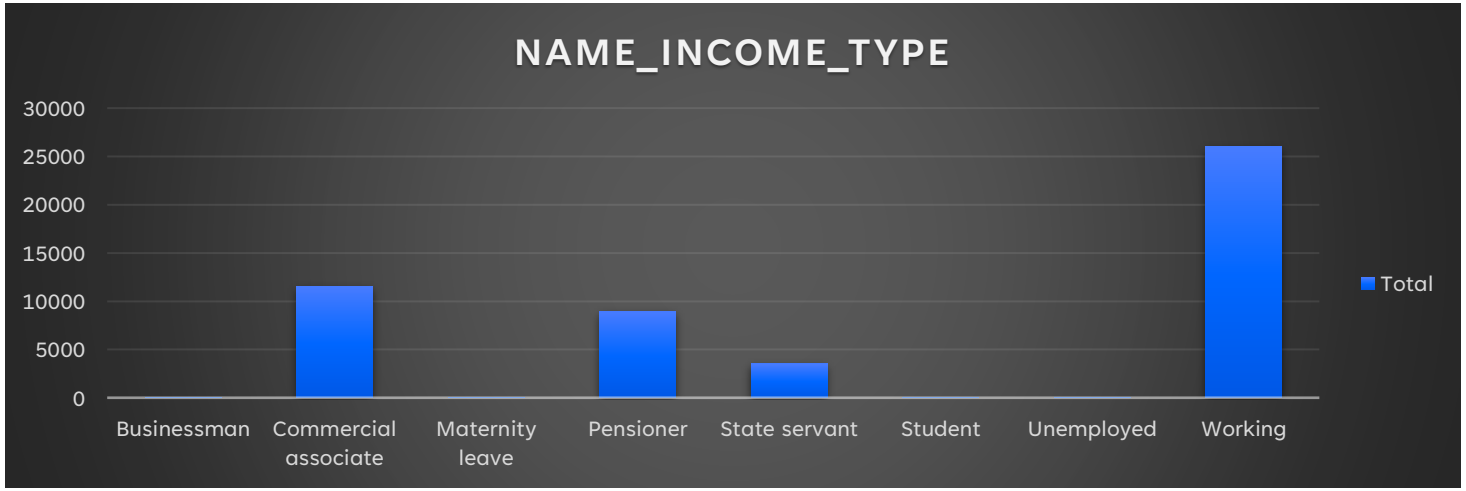
Perform Univariate, Segmented Univariate, and Bivariate Analysis

Task: 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.

Univariate Analysis

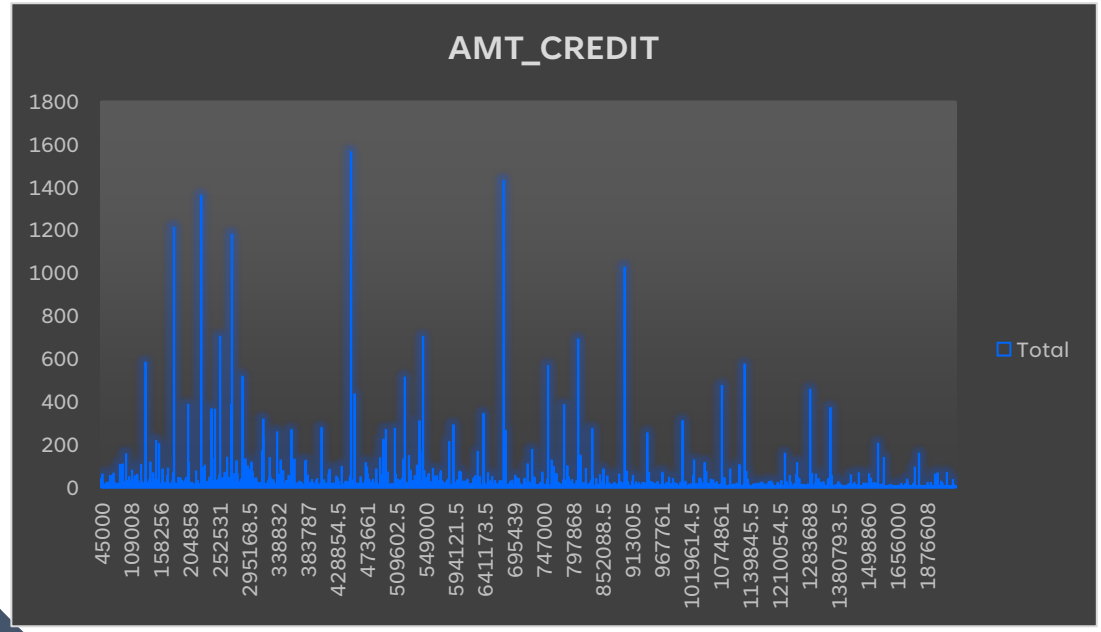
Univariate analysis focuses on examining and describing individual variables in isolation.

- Individuals with higher incomes trend to be less inclined to apply for loan. The income under the range of 25000 to 275000 will be the higher income.



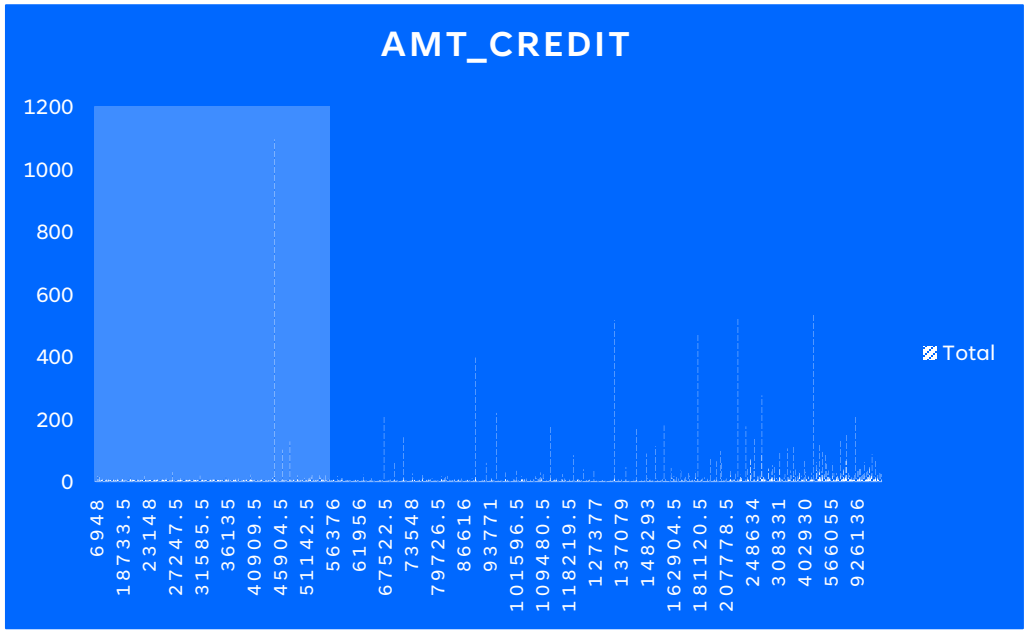
Univariate Analysis

- Bank loan fall within the range of 45000 and 1045000



Recent data

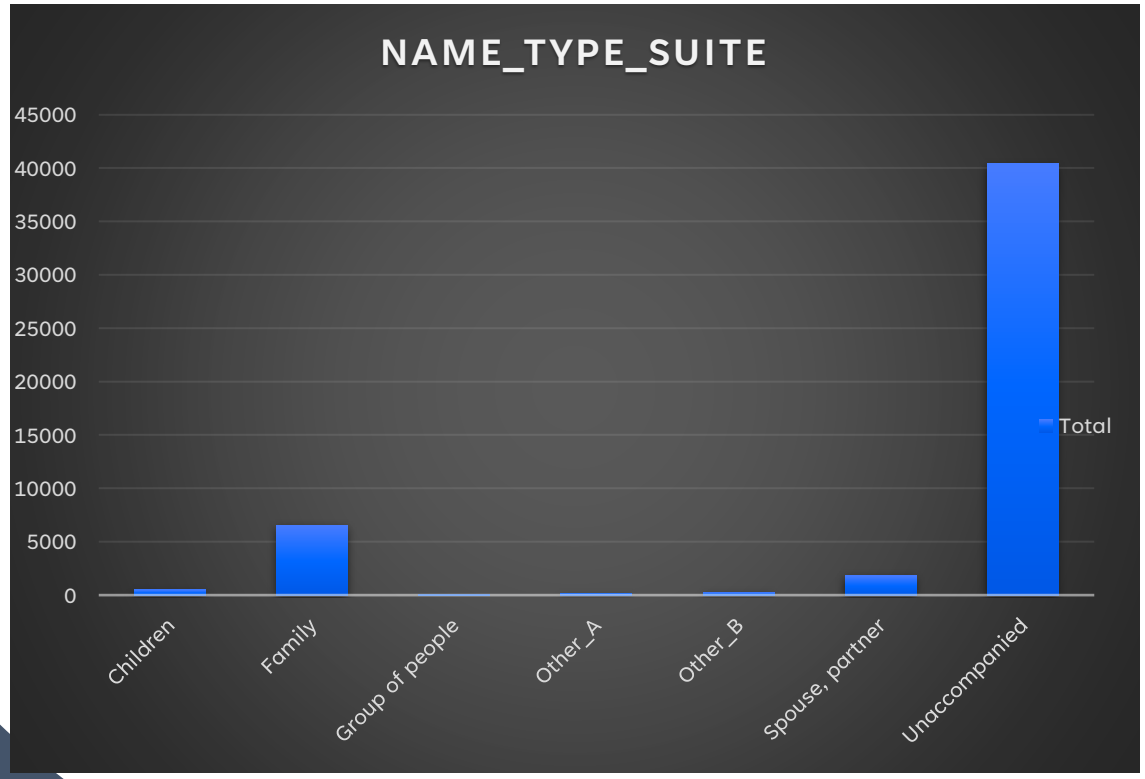
Recent data



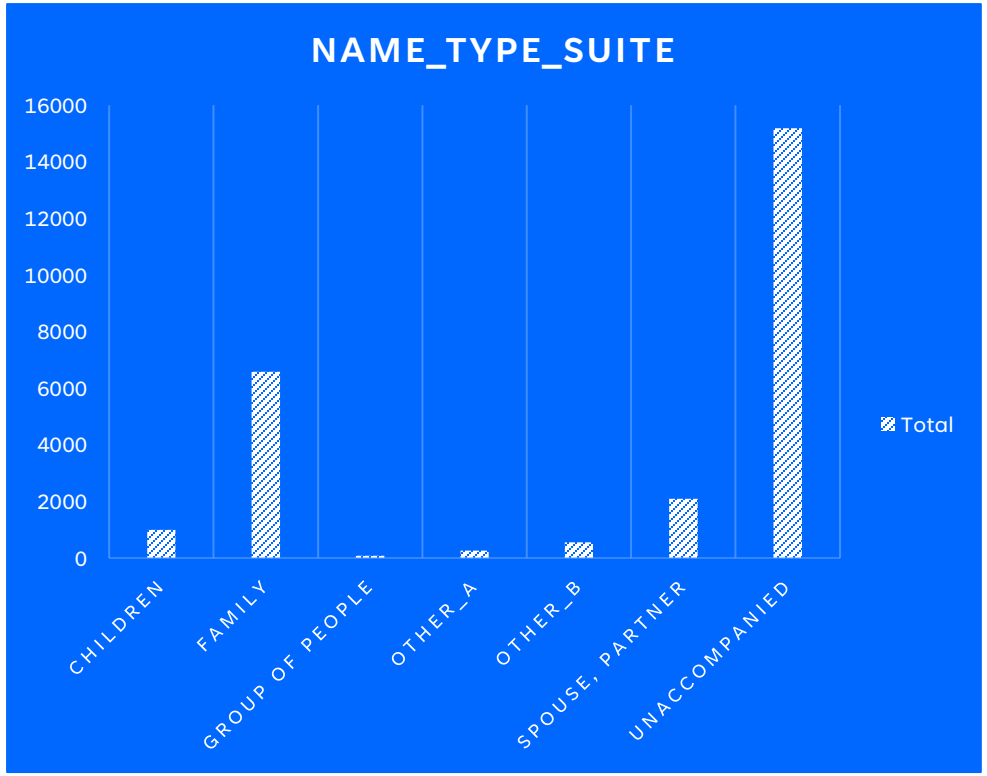
Previous data

Univariate Analysis

- Unaccompanied people highly applied for the loan.



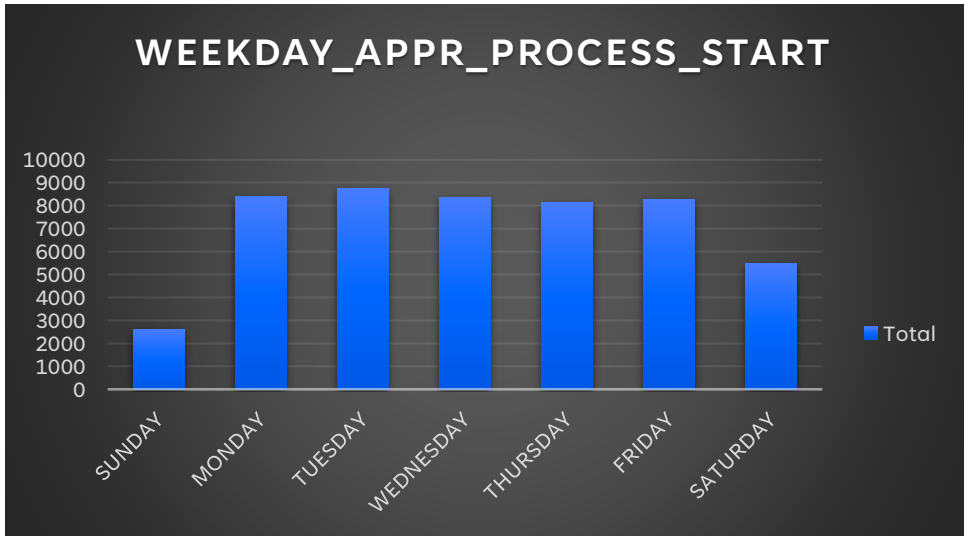
Recent data



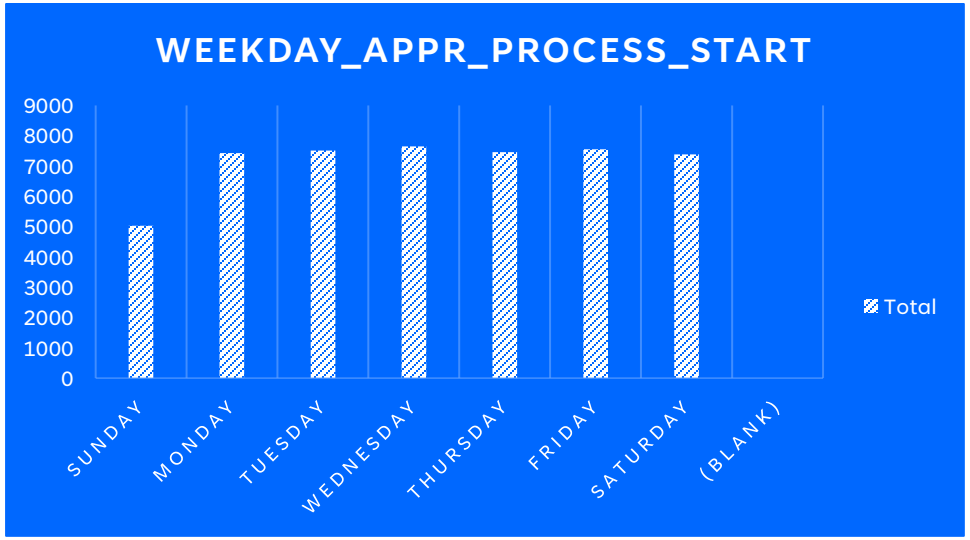
Previous data

Univariate Analysis

On Sunday, people apply less for loan as compare to ordinary days.



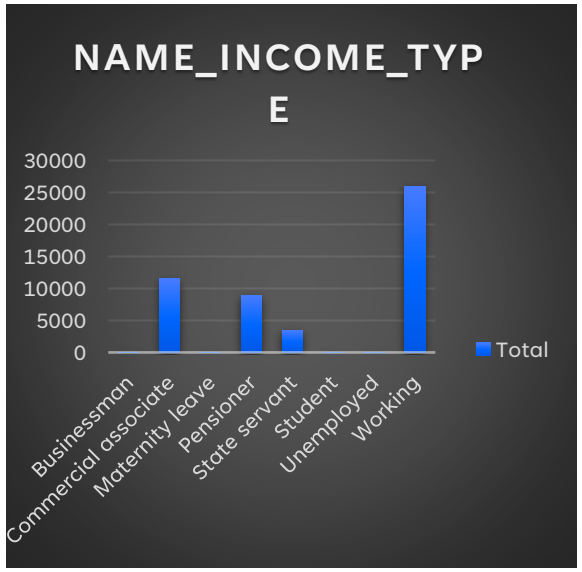
Recent data

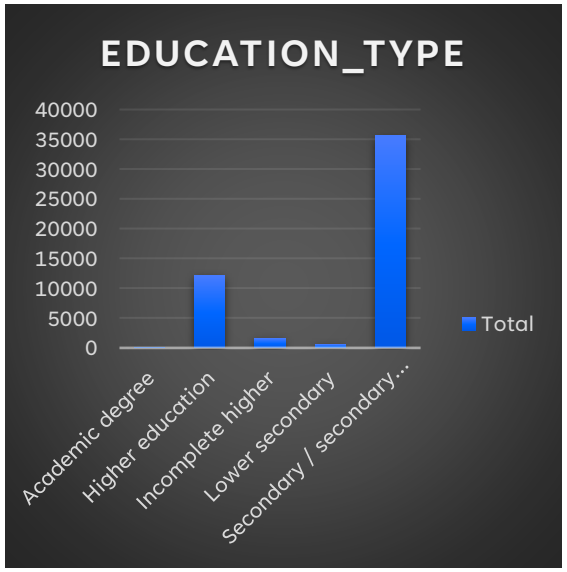


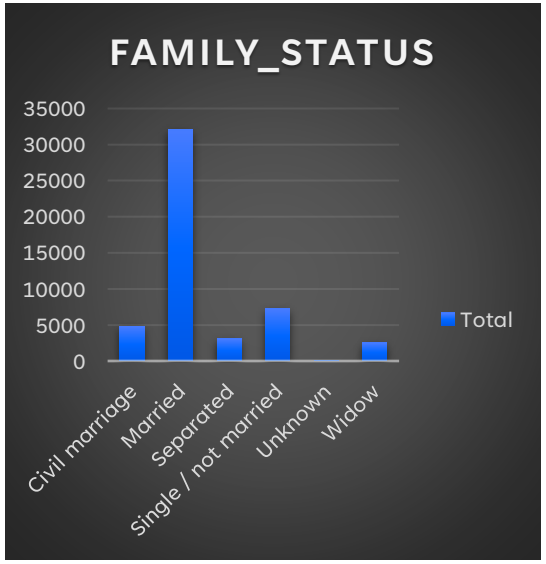
Previous data

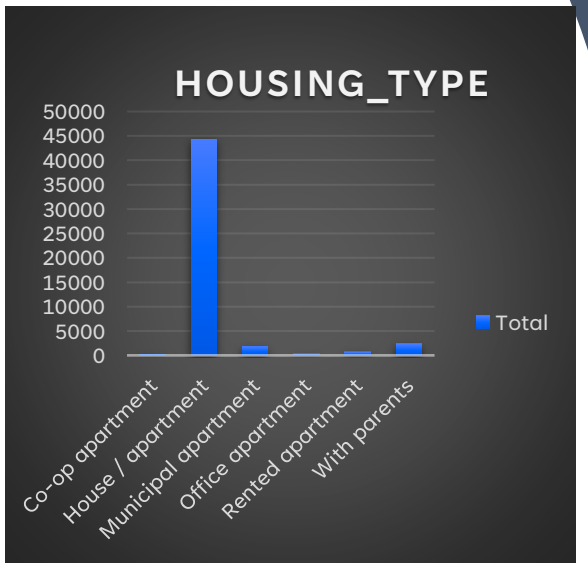
Univariate Analysis

The rest of the parameter of loan are represents with the help of graph.









Segmented Univariate Analysis

It is the extension of univariate analysis that involve splitting the dataset into specific segments.

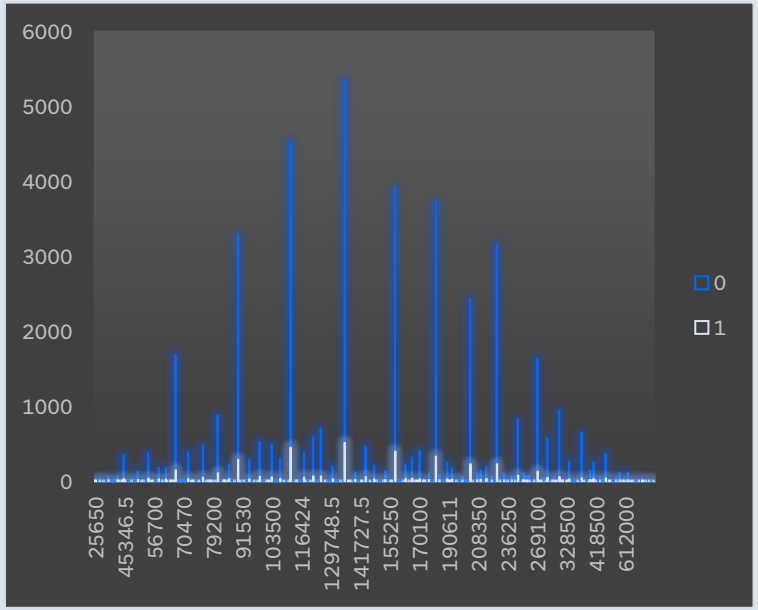
Bivariate Analysis

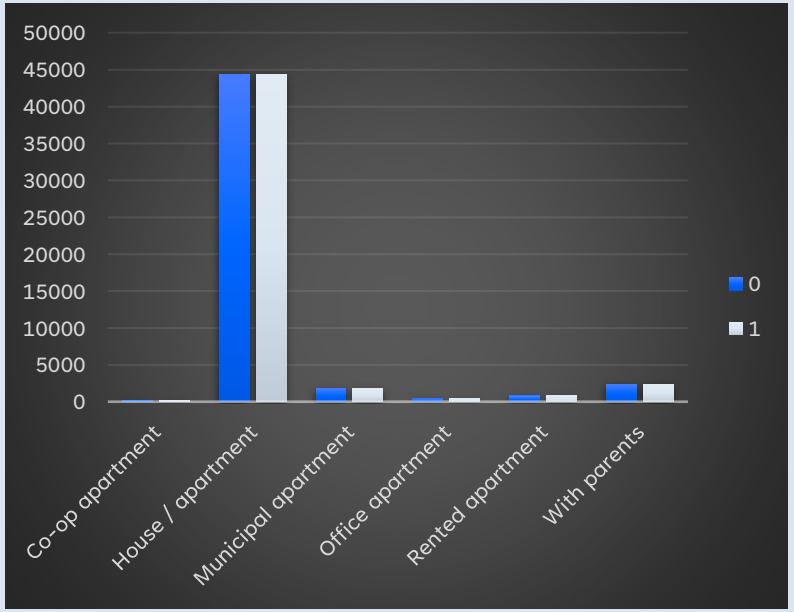
Bivariate Analysis examine the relationship between two variable simultaneously. It focuses on exploring the association, correlation, dependency between two variables and helps understand how changes in one variable are related to change in the other.

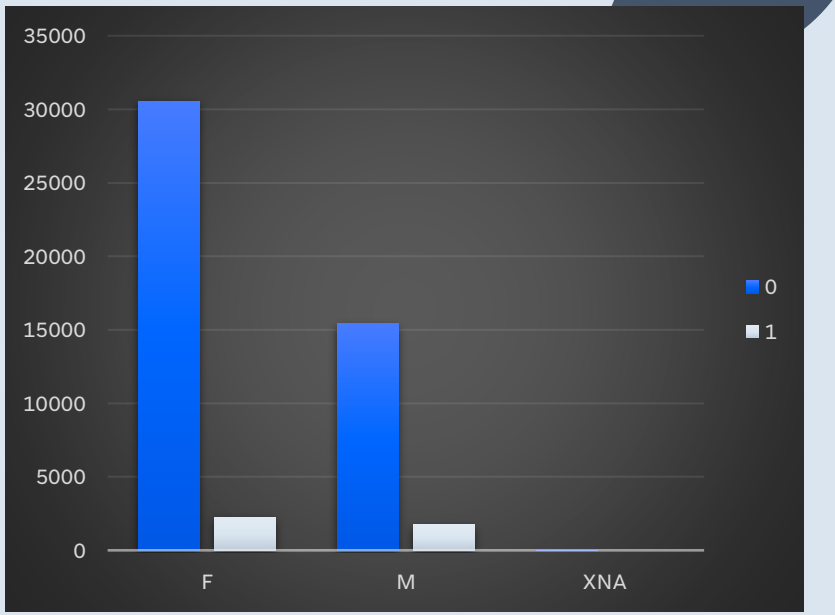
Outcomes:

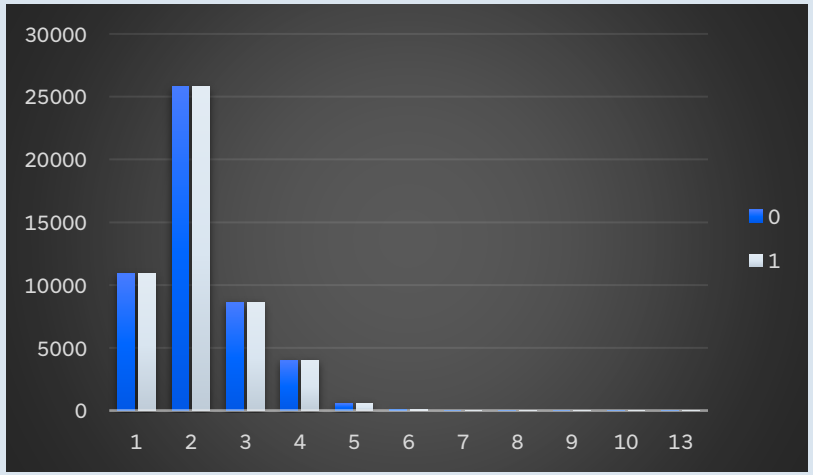
- Individuals with lower incomes exhibits a greater likelihood of defaulting.
- Female customers display a lower inclination toward default compared to males.
- Customers with more than five family members are less likely to default on their bank loans.
- Customers with secondary qualifications are more prone to loans defaults.
- Customers taken loans most for housing.

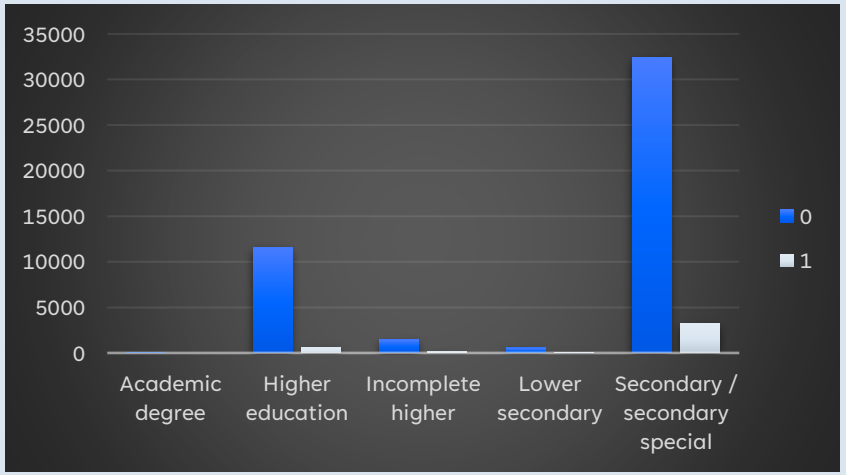
Bivariate Analysis











Identify Top Correlations for Different Scenarios

Task: 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.

Here, I use the inbuild data analyser for find corelation between then

https://drive.google.com/file/d/1K5xjkIM7K83AMUuRKPsD4Et92Jfpm892/view?usp=drive_link

CORRELATION

VAR1	VAR2	CORRELATION
AMT_CREDIT	AMT_GOODS_PRICE	0.986943730194452
CNT_CHILDREN	CNT_FAM_MEMBERS	0.880454498084003
DAYS_BIRTH	DAYS_EMPOYED	0.613553972166506
DAYS_BIRTH	DAYS_REGISTRATION	0.33363250873219
AMT_CREDIT	AMT_ANNUITY	0.33363250873219
DAYS_BIRTH	DAYS_ID_PUBLISH	0.270825141291616
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950710179345493
OBS_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.31141087635516
OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.237930380809405
AMT_ANNUITY	AMT_GOODS_PRICES	0.77443394658966

Resources

Here is the link of all the spread sheets :
https://docs.google.com/spreadsheets/d/1ILbRmidNrpBggjLj9VtubLrRarK3m8S7/edit?usp=drive_link&ouid=101206229307191695705&rtpof=true&sd=true

Thank you