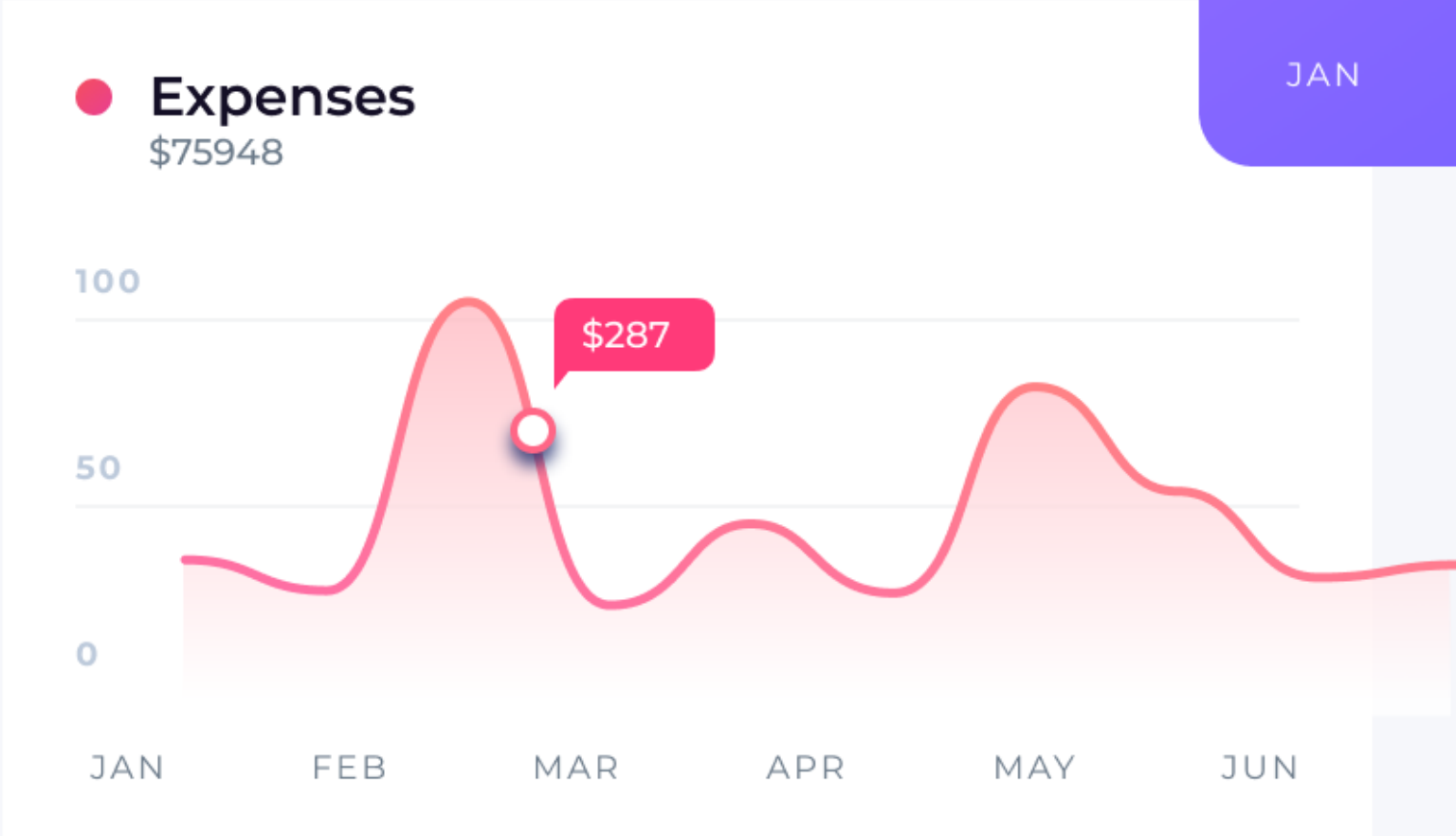
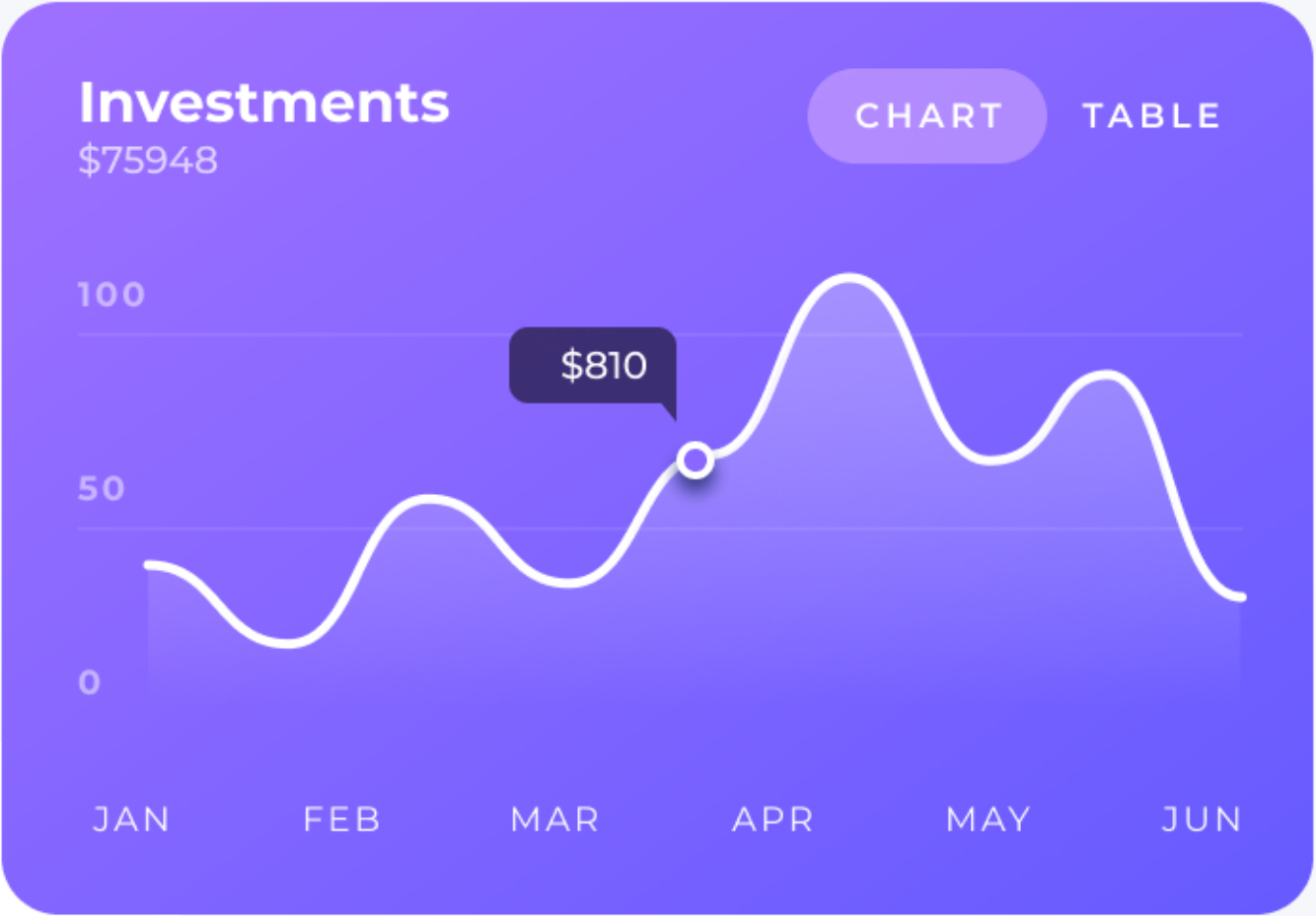


Bank Loan Case Study

By Suraj Beloshe

Powered by

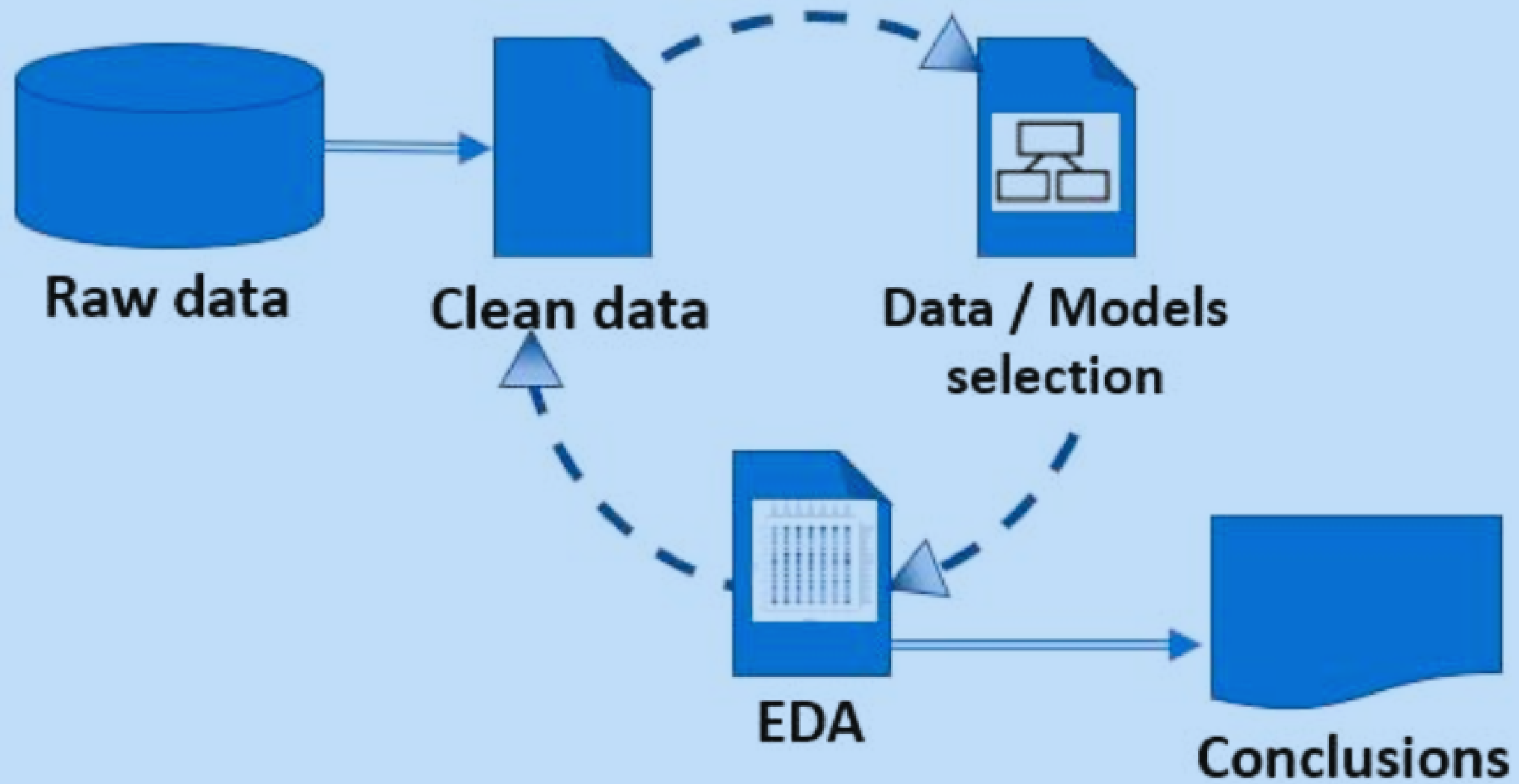


Project Description

- Lending money is the bread and butter of the banking business, but **selecting the correct borrowers** can often be a challenging task.
- This EDA project aims to **enhance the loan approval** process by analyzing customer loan applications and previous loan history.
- By assessing creditworthiness, identifying risk factors, and exploring loan approval rates, it seeks to make informed decisions when selecting borrowers.
- Gain a **understanding of risk analytics in the banking and financial services industry**, specifically how data is utilized to mitigate the potential risk.
- Insights gained from the analysis **will lead to actionable recommendations** for improving the loan approval process and minimizing default risks.
- The project's **ultimate goal is to optimize lending decisions**, contribute to the bank's success, and ensure the selection of the right borrowers.



Approach



Tech-Stack Used

Microsoft 365 Excel



- **Comprehensive Data Analysis Tools**
- **Seamless Integration**
- **Scalability and Accessibility**
- **Data Transformation Capabilities**
- **Advanced Visualization and Reporting**

Insights

Problem Statement

- **Challenges to Give Loan**
- **Two Type of Risk a) Loss of potential Client & b) Giving Loan to the Defaulter**
- **understand the driving factors**
- **Minimize the risk of losing money**
- **Creating Visuals**

Approach

- **Understand Dataset**
- **Work on insights to understand requirement of data**
- **Follow correct method of analysis**
- **Help to take decision with proper analysis**

Insights

Handling Missing Data

	Before		After	
File Name	Rows	Column	Rows	Column
Application Data	3,07,511	122	3,07,511	81
Previous Application	16,70,214	37	16,70,214	33

- By Using Power Query Reduce Size of Files
- Removed column more than 50% missing Value for both Data set as per thumb rule
- Check Duplicate - No Found
- Imputation need to done on the blank cells with column below 50% missing value
- Checked and highlighted them with conditional fomating

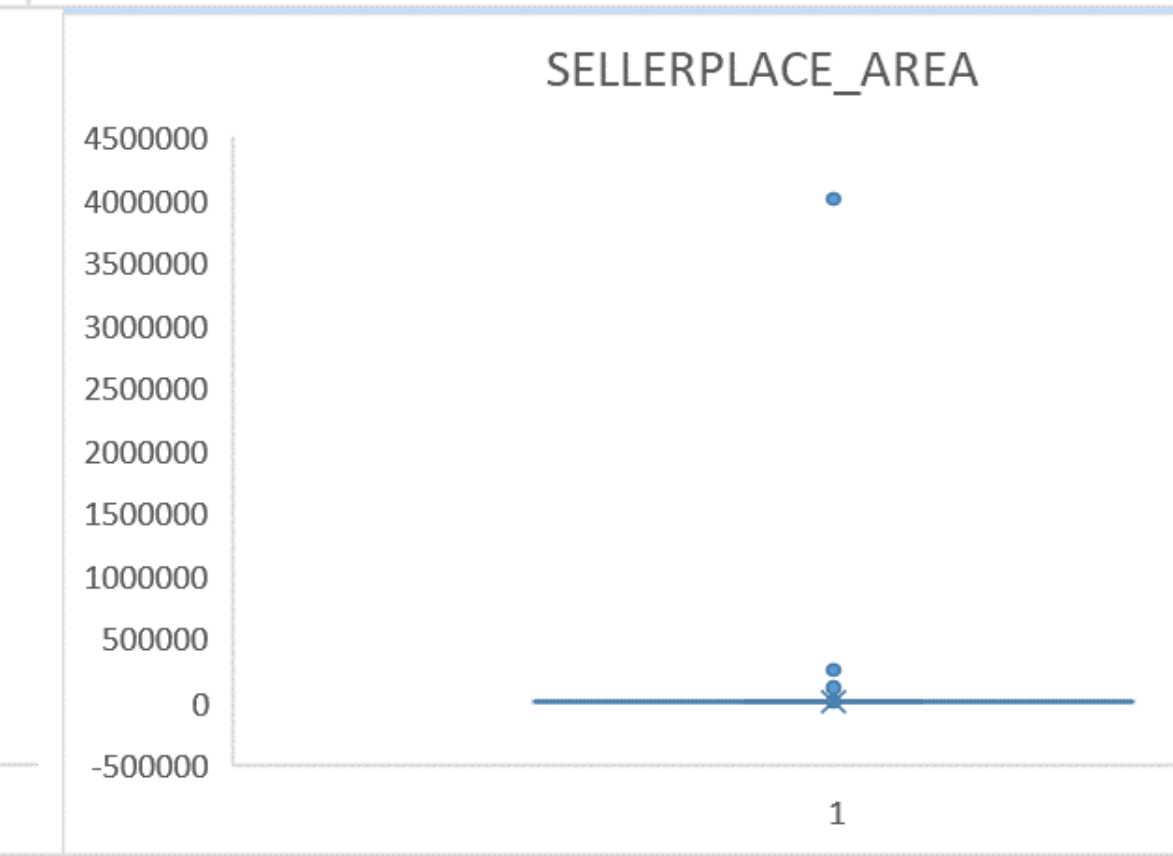
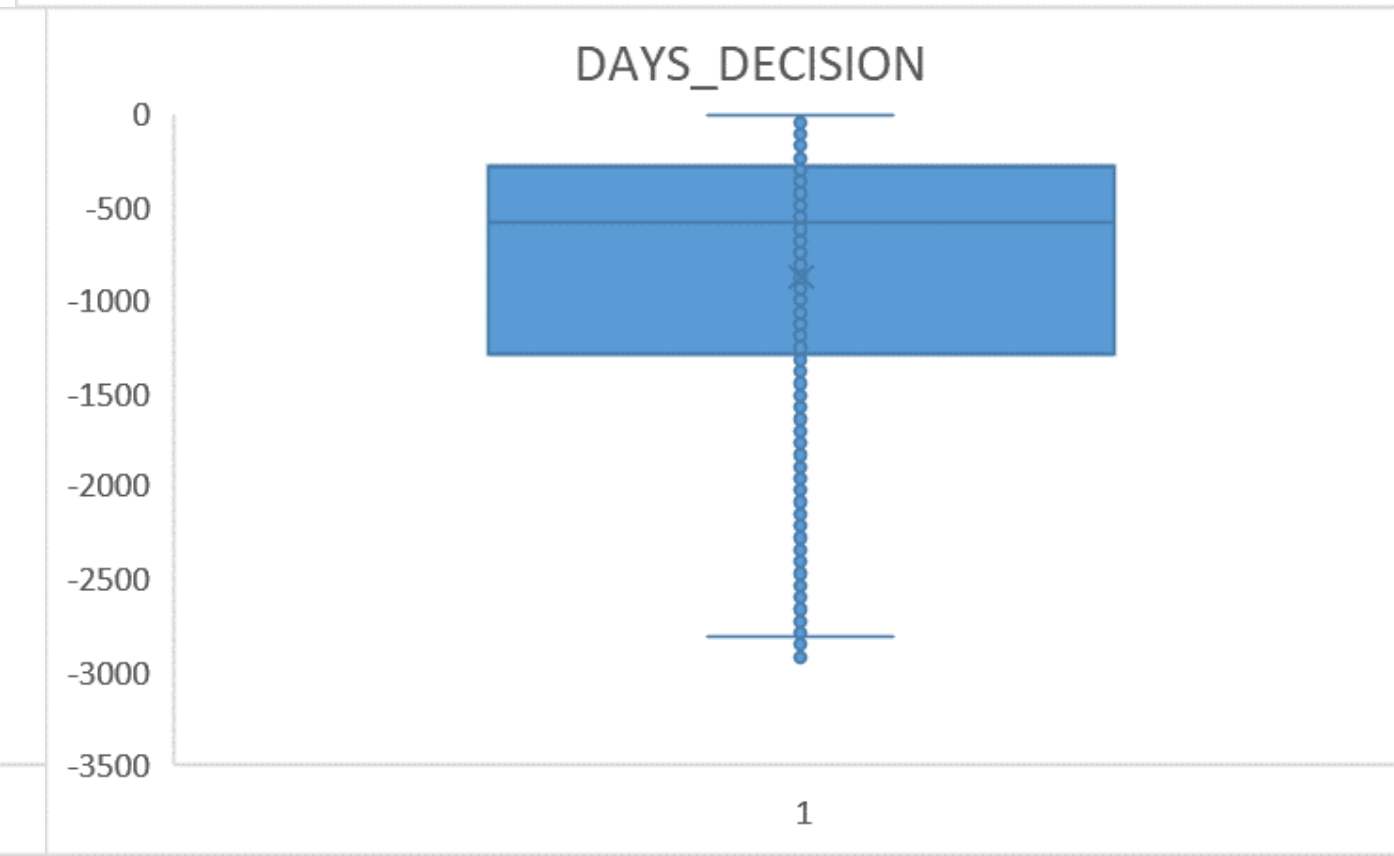
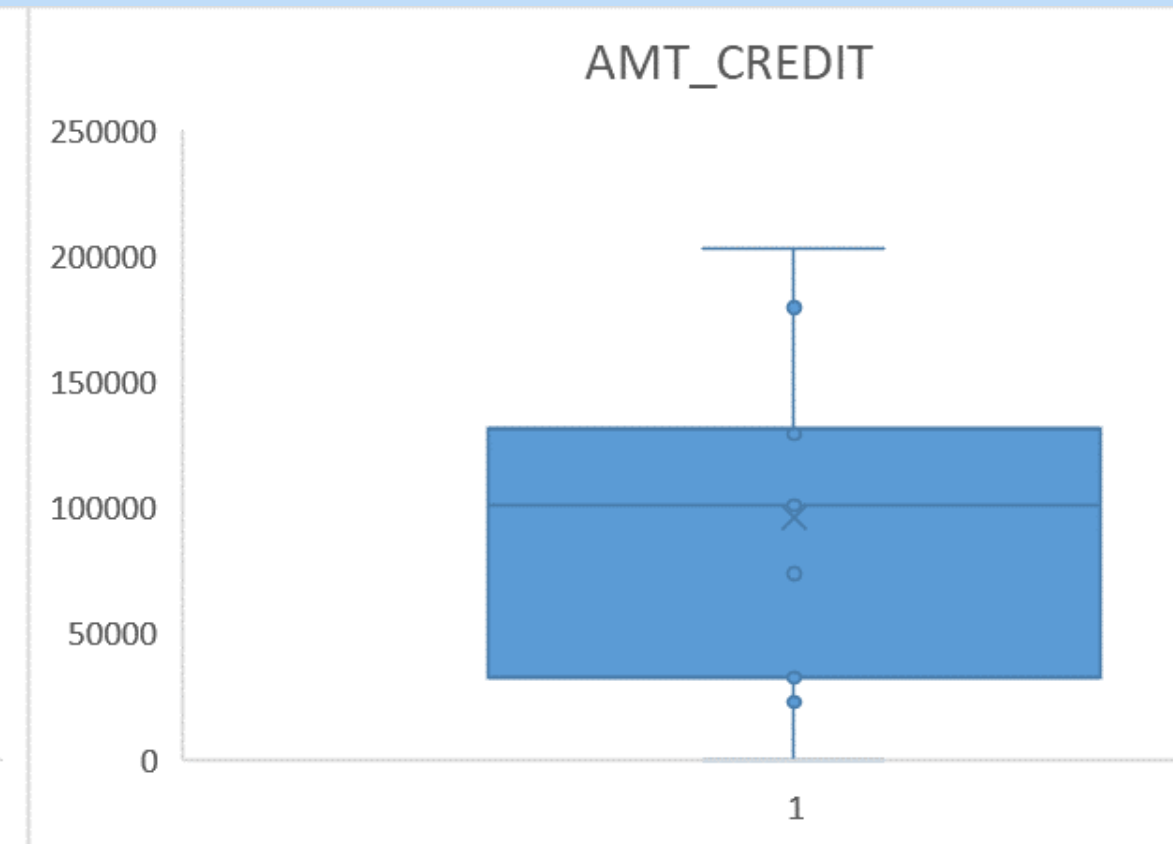
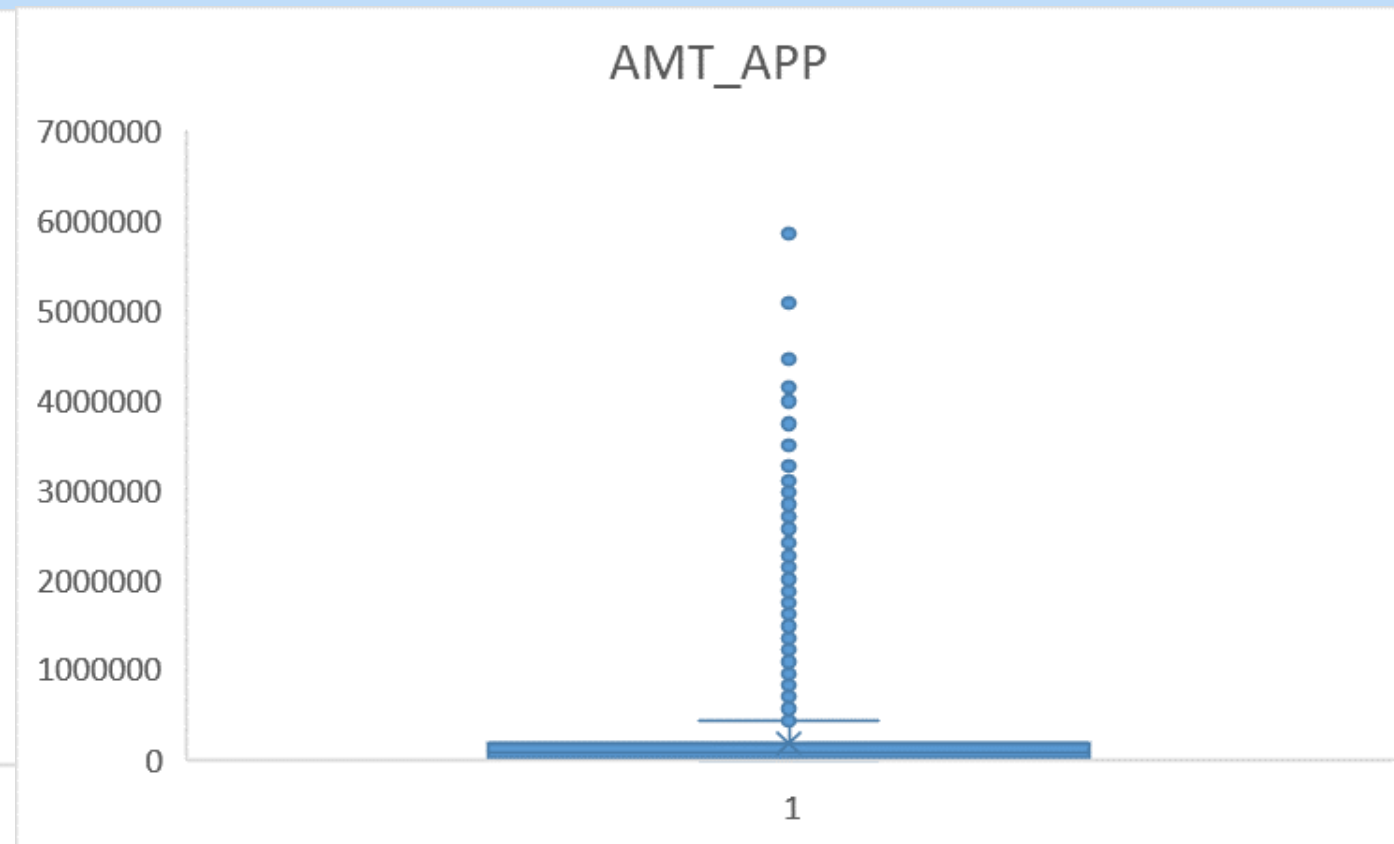
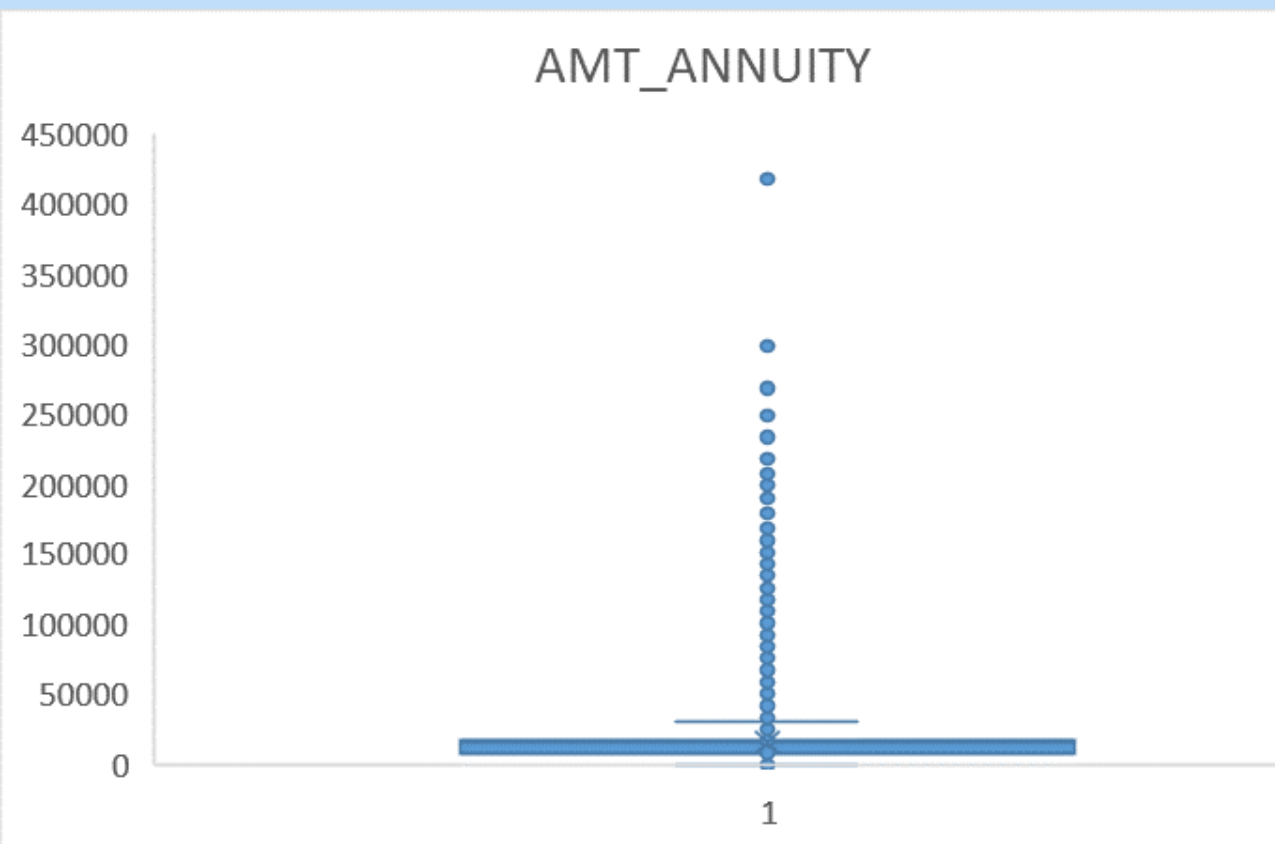
Insights

Imputation of the Missing Data

- Imputation refers to the process of estimating or filling in missing values in a dataset
- Imputation methods aim to replace missing values with plausible estimates based on the available information in the dataset
- Application Data :- There were 26 Columns Needs imputation
- Previous Application :- There were 12 Columns Needs imputation
- Divided data in two category : 1. Continuous numerical and 2. Categorical
- Find out Mean and Median for Continuous numerical data and imputed with required value
- Categorical Value imputed with Mode of the data
- Drop Column :- DAYS_FIRST_DRAWING column from Previous application has insignificant value 365243 which was not relatable hence dropped this column

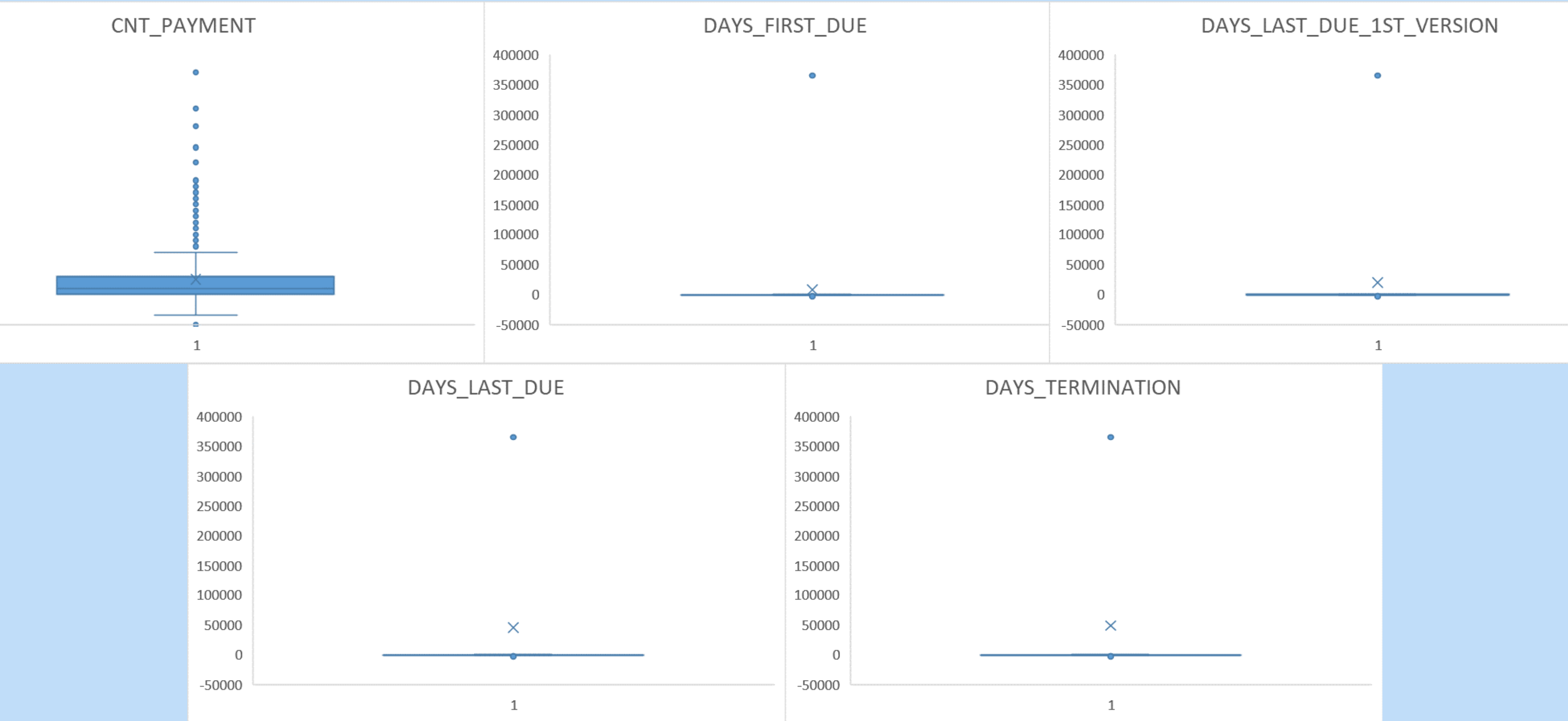
Insights

Outliers : Previous Application 1 of 2



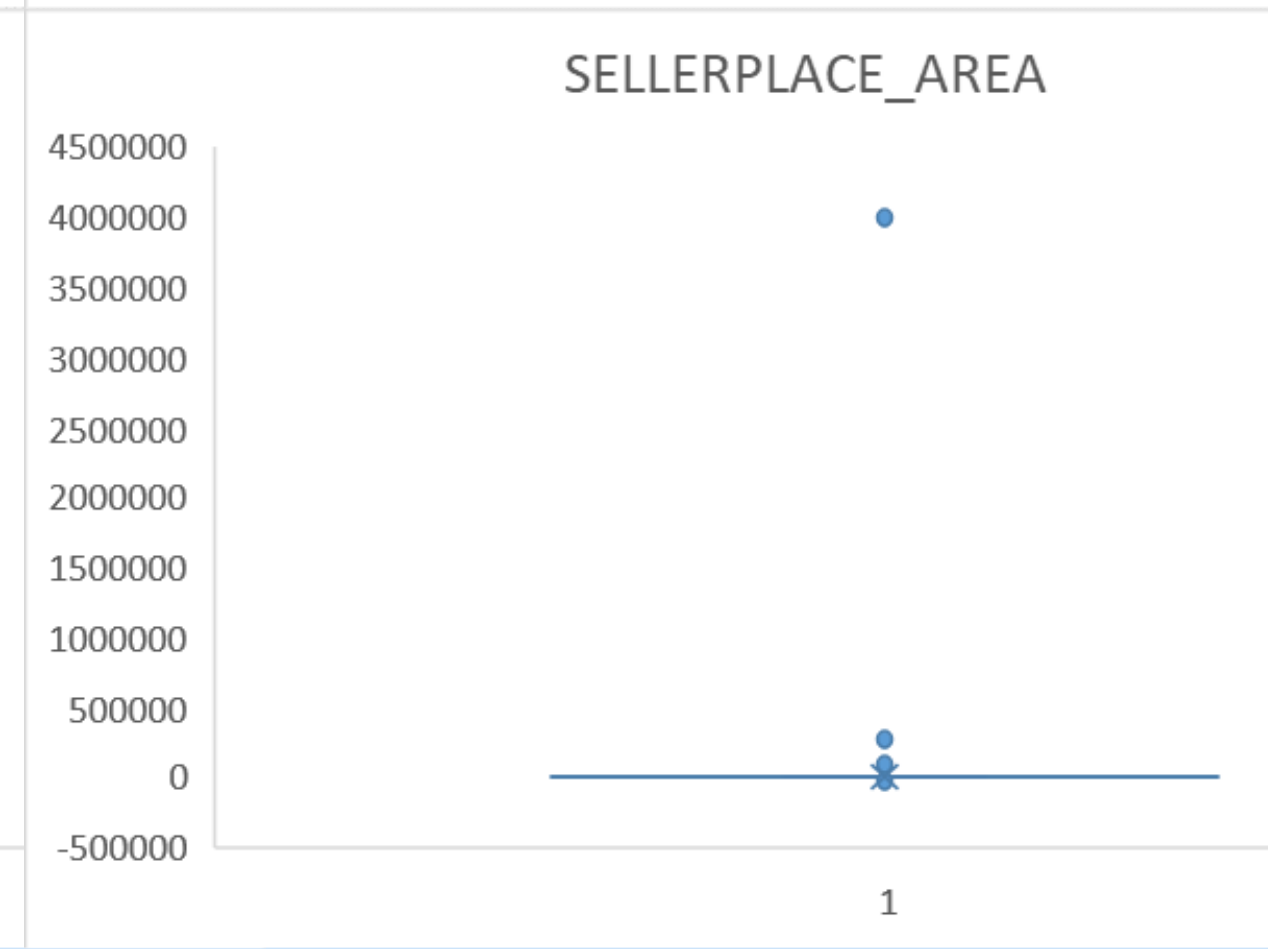
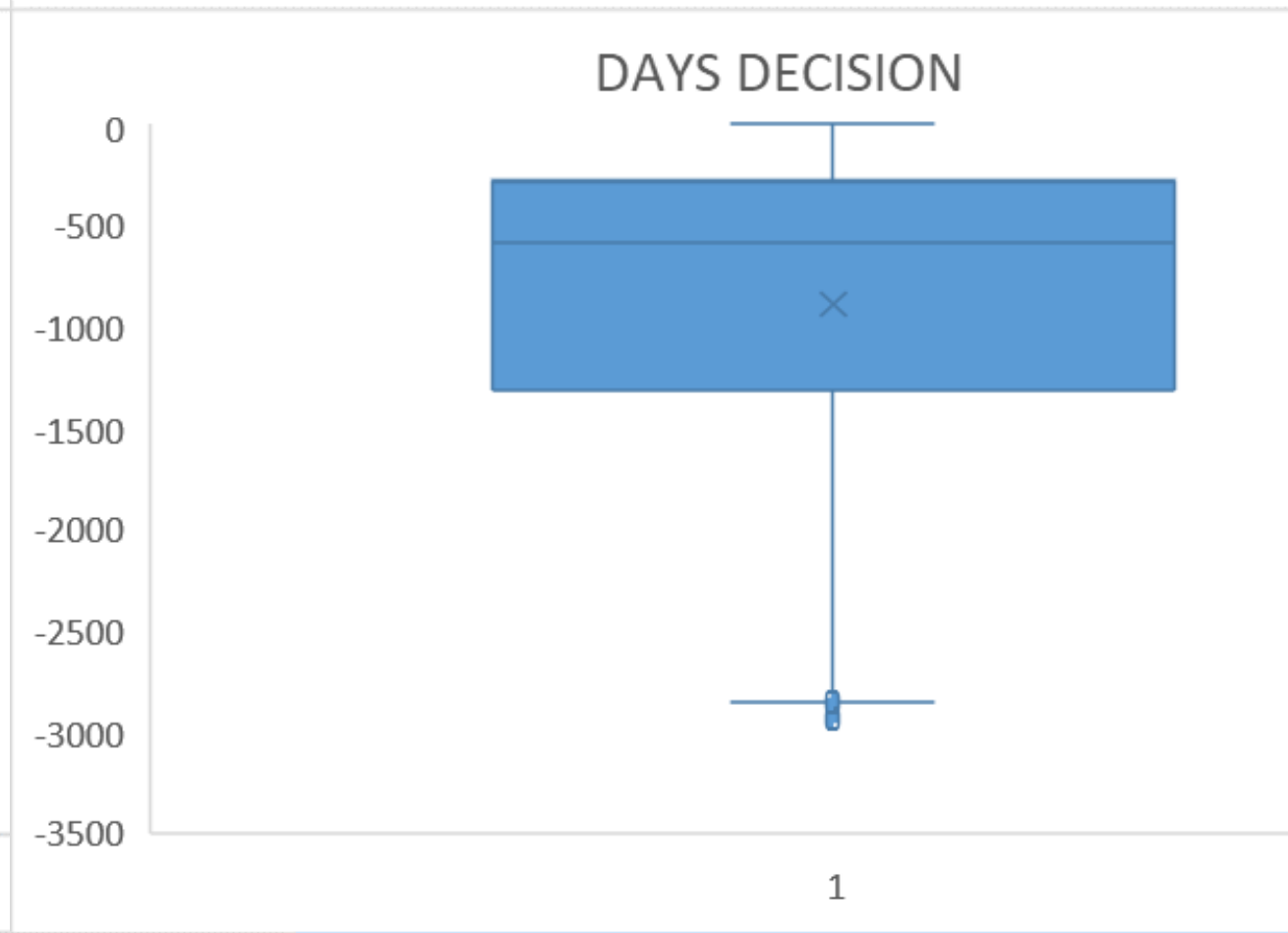
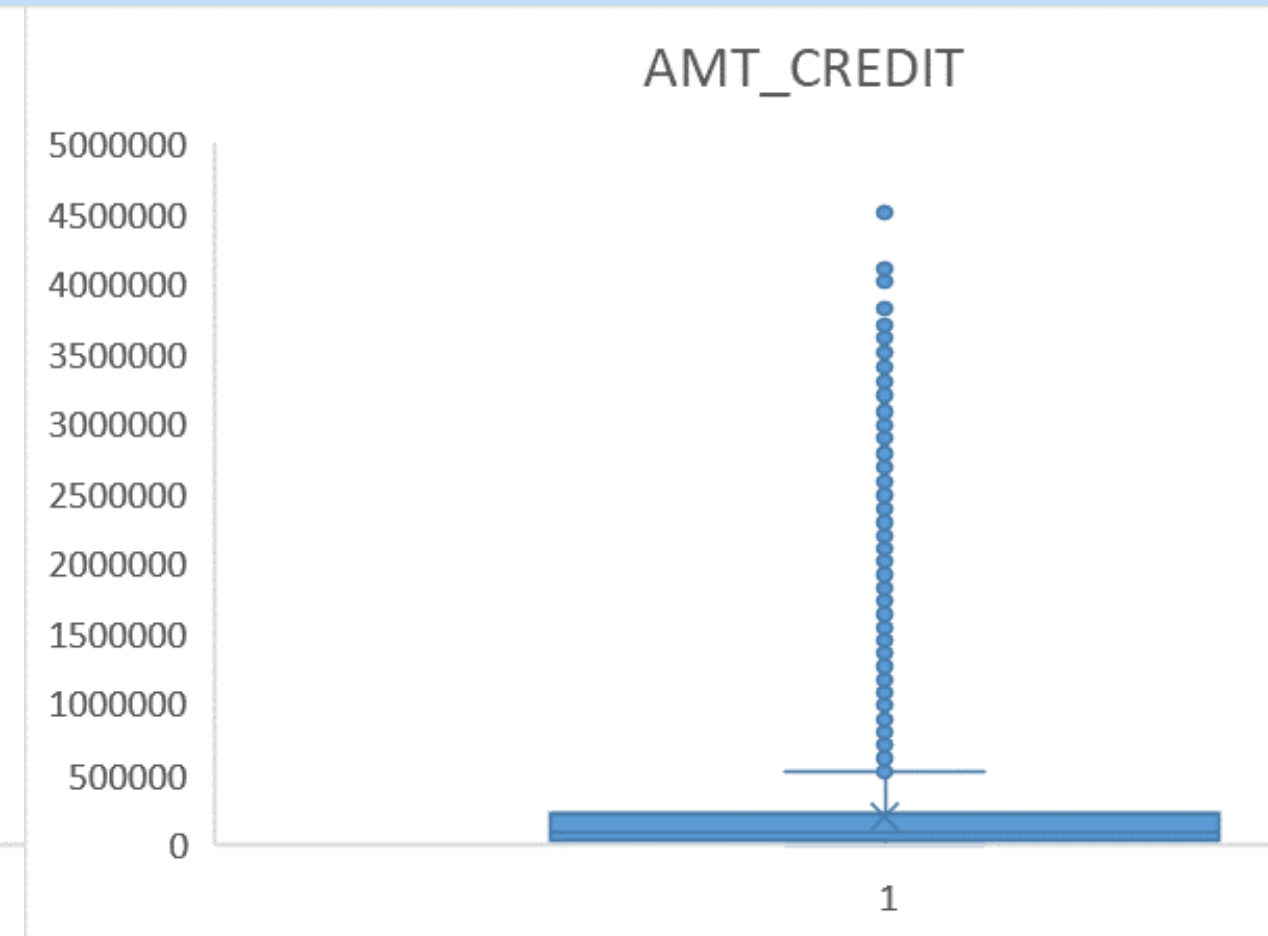
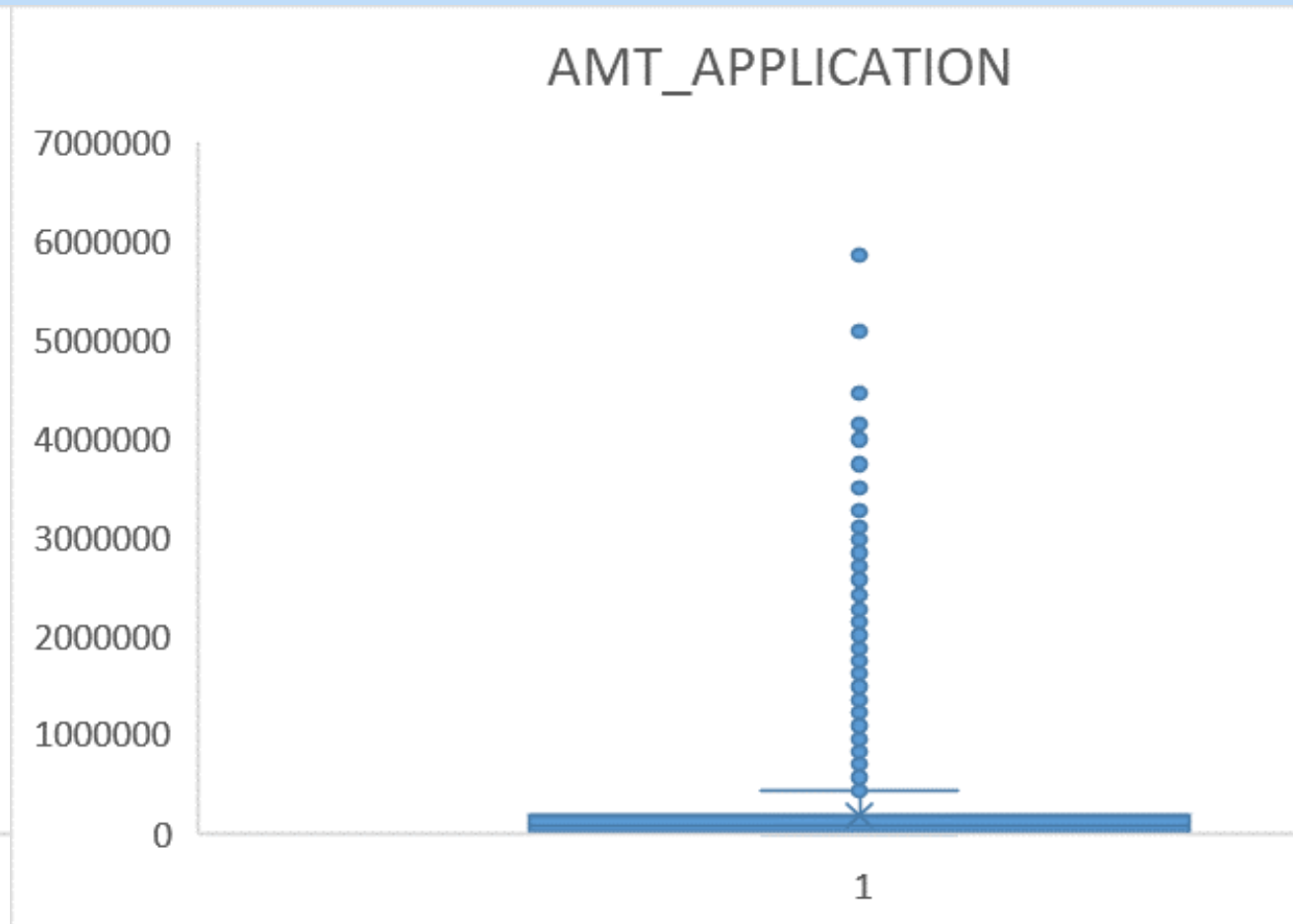
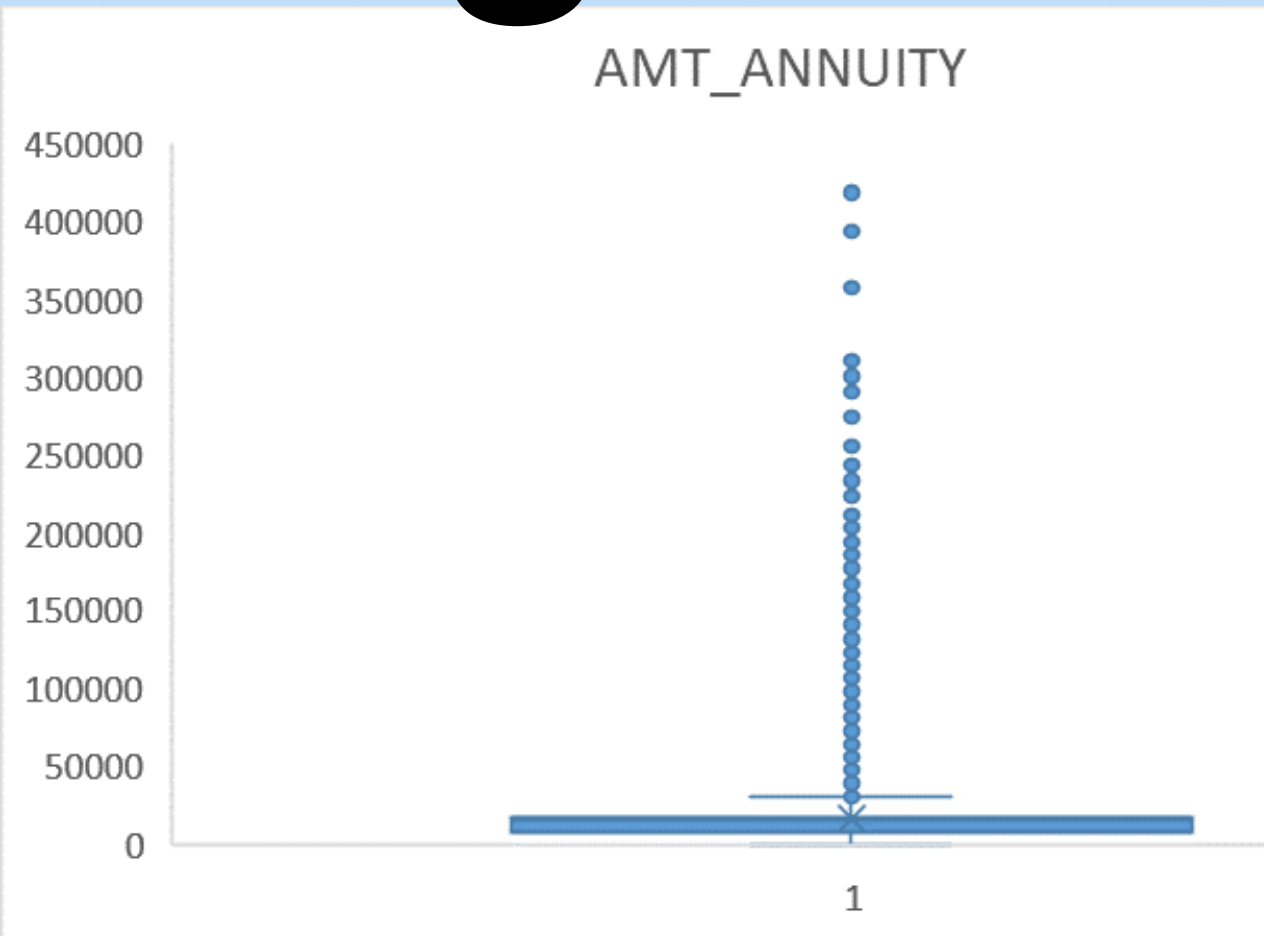
Insights

Outliers : Previous Application 1 of 2



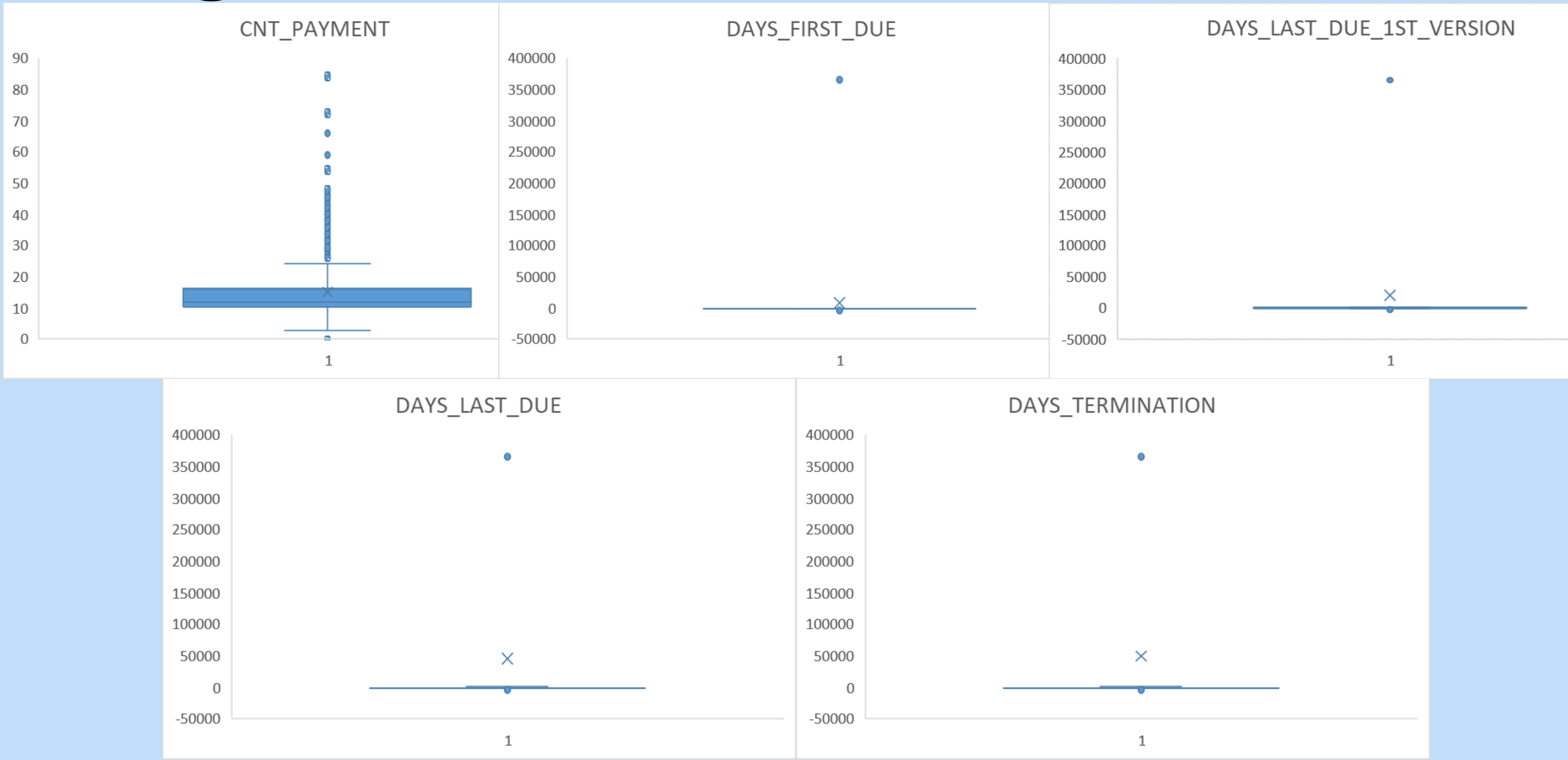
Insights

Outliers : Previous Application 2 of 2



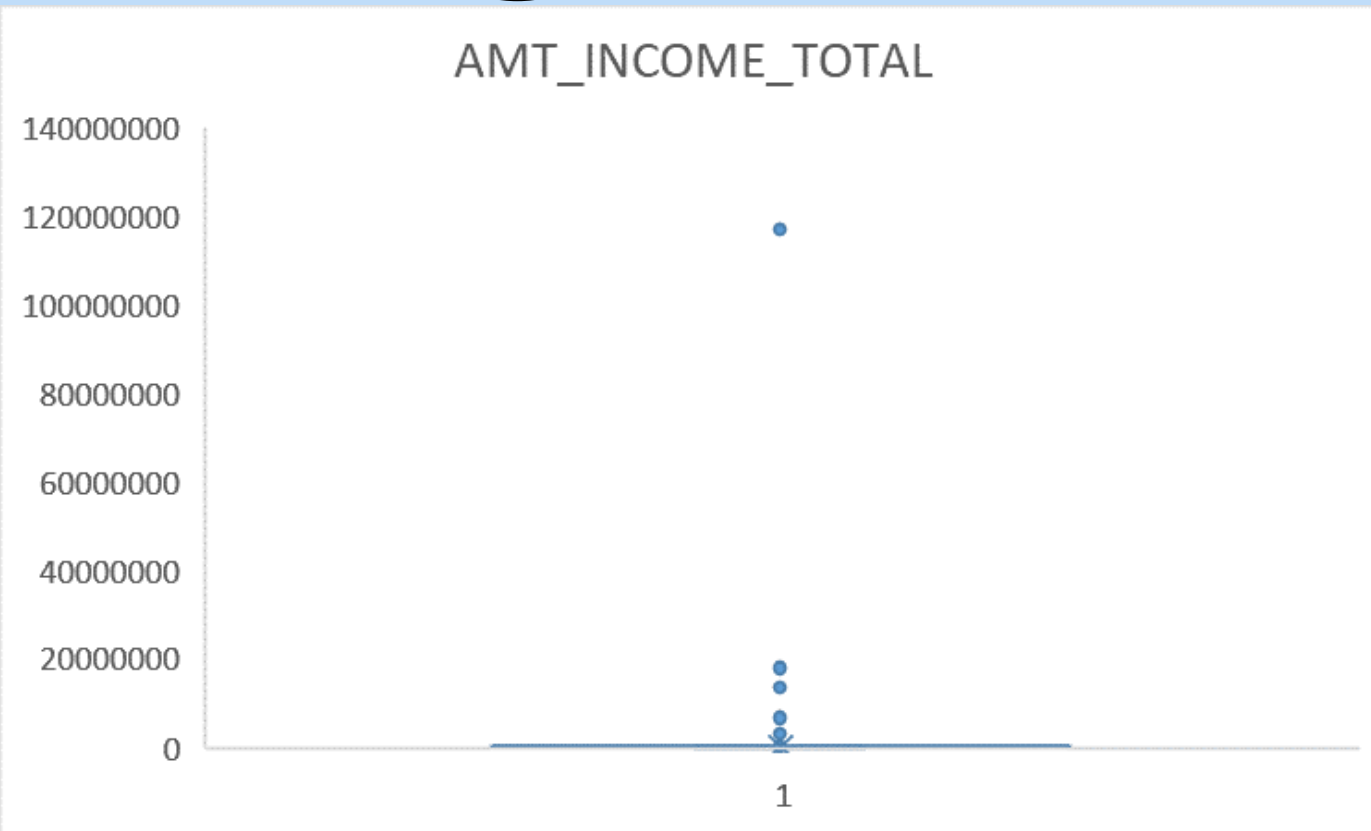
Insights

Outliers : Previous Application 2 of 2

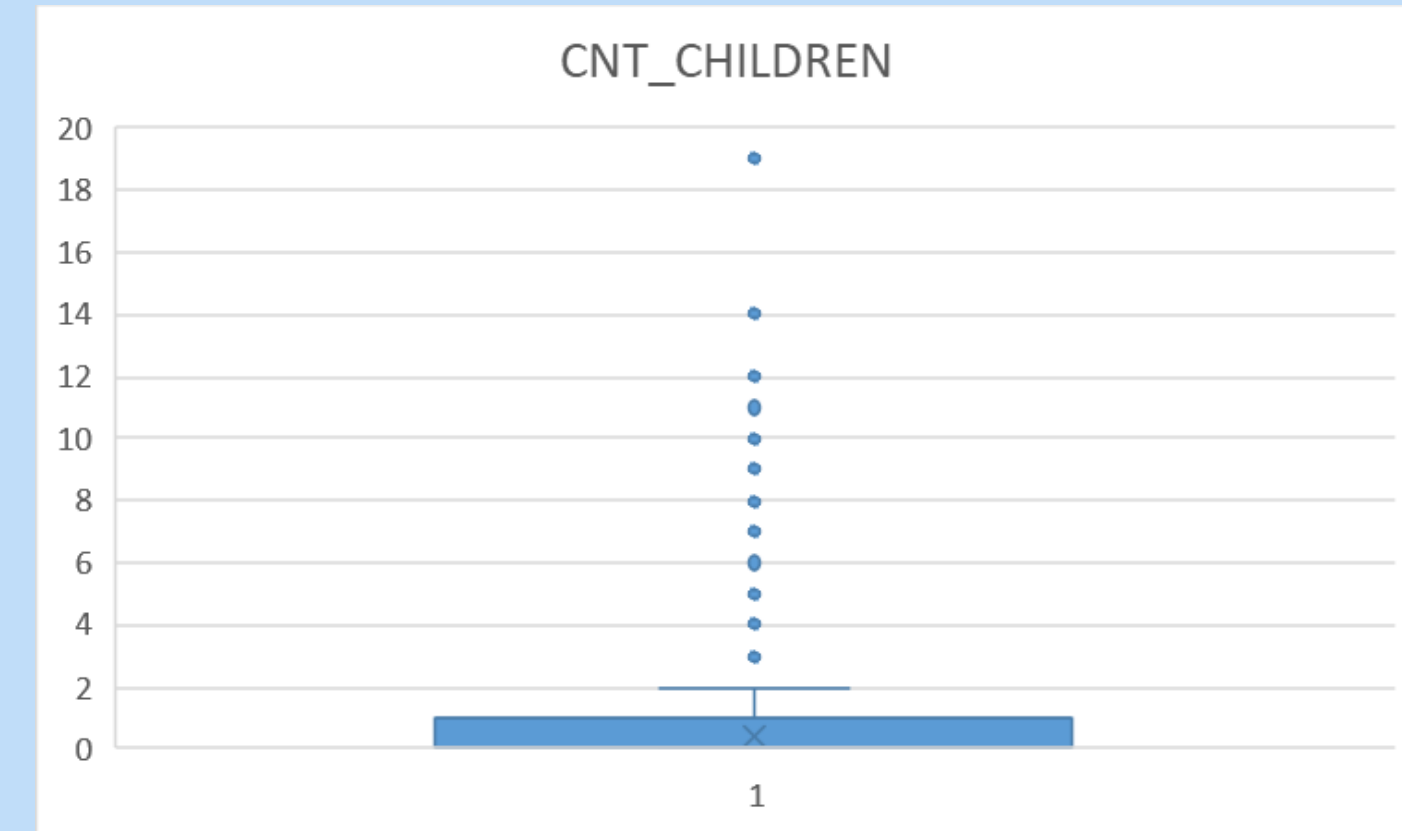


Insights

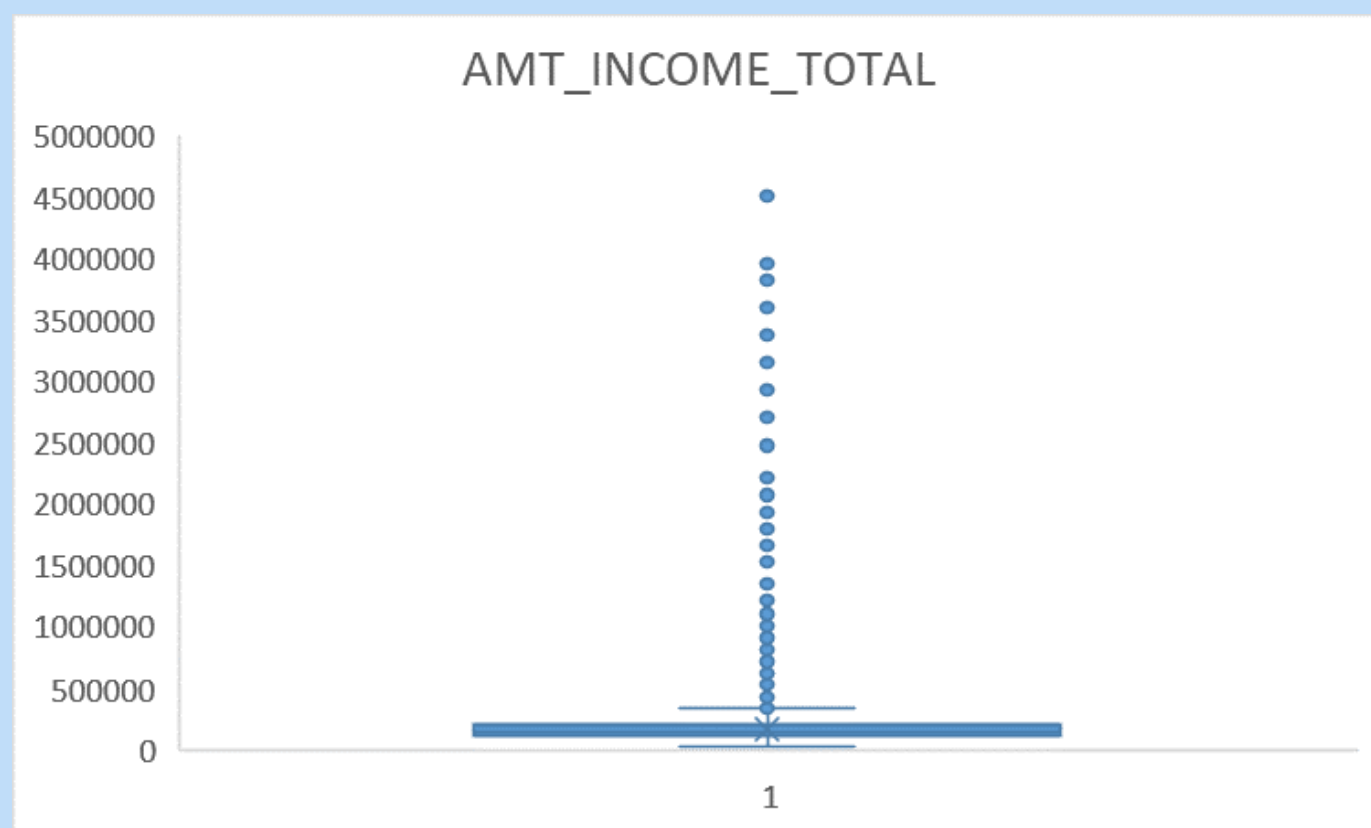
Outliers : Application Data



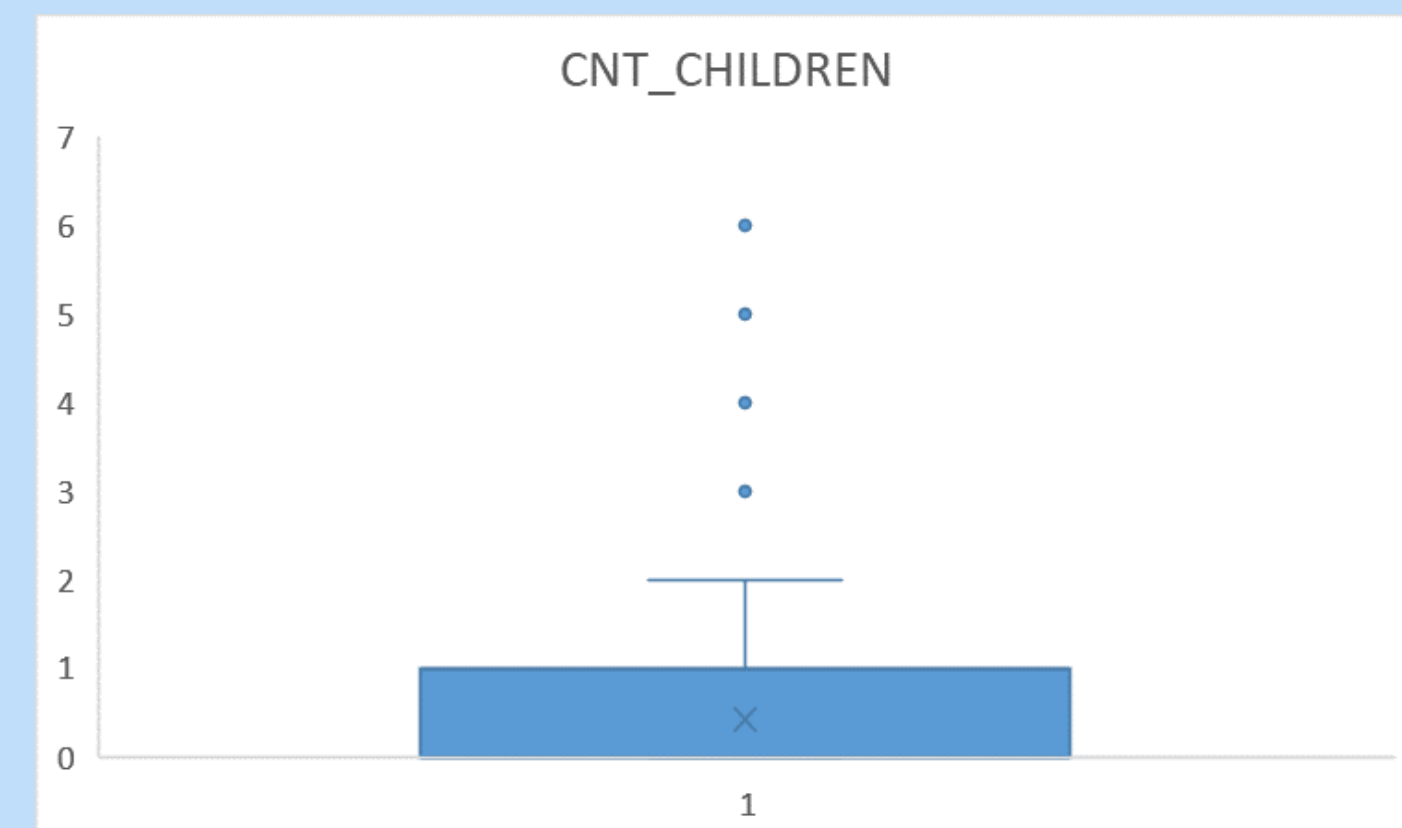
With Outlier



Removed 26 Rows

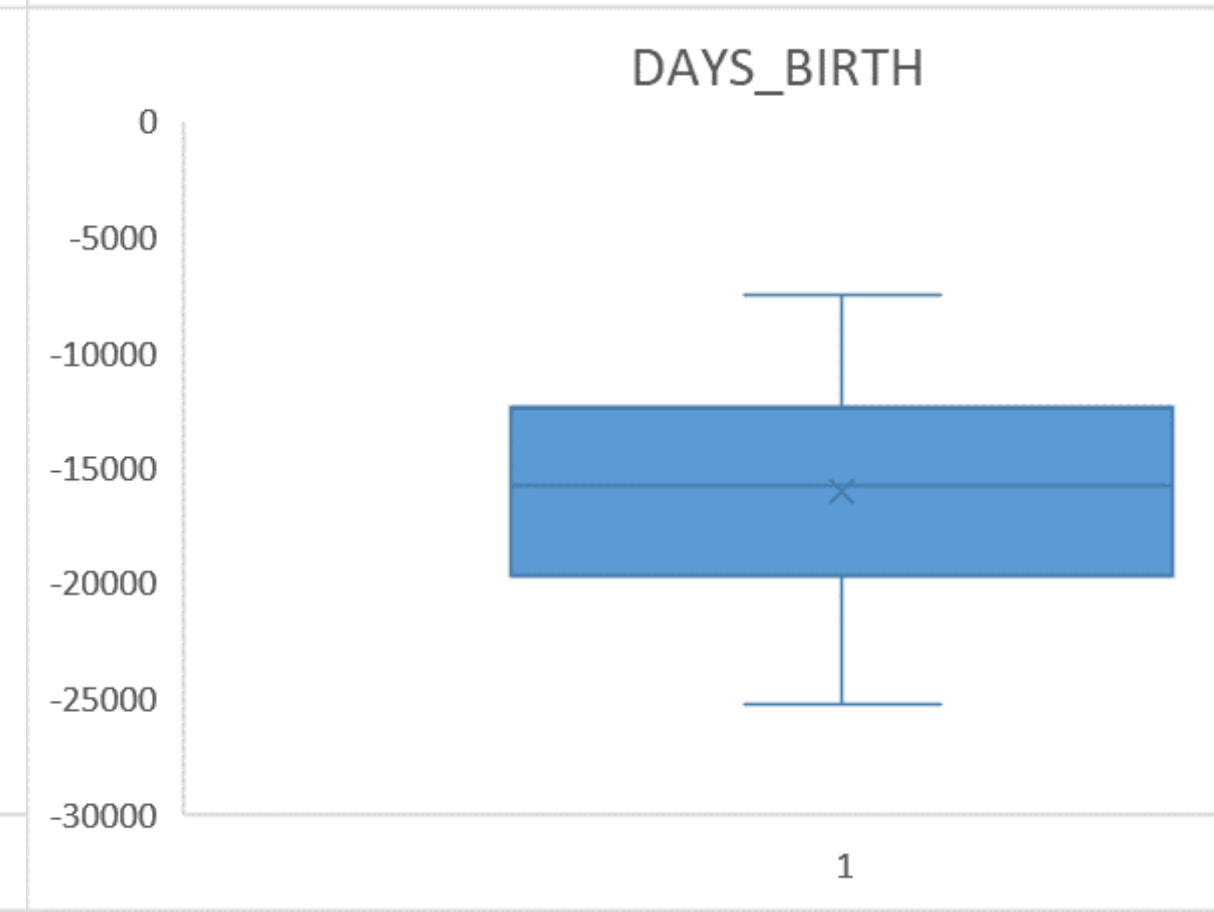
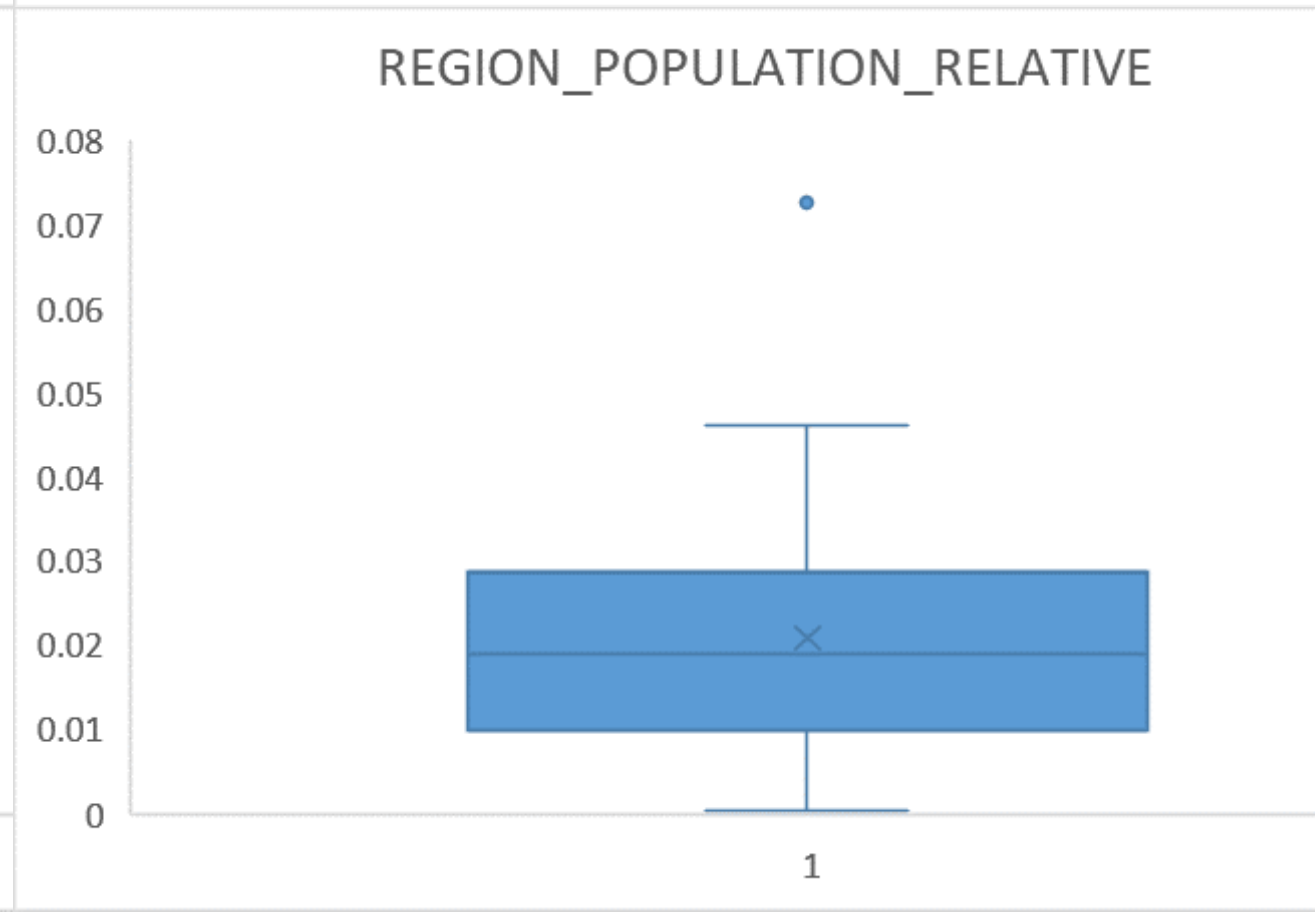
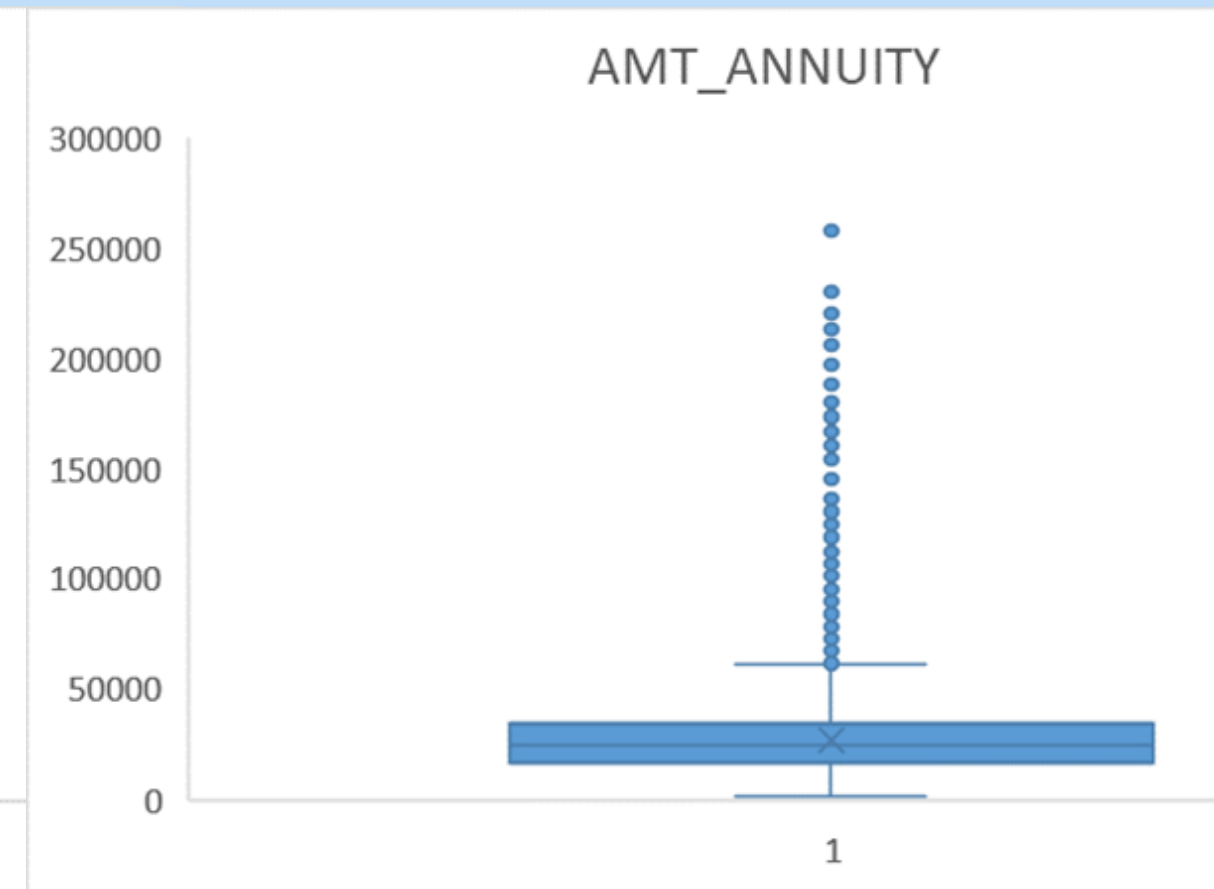
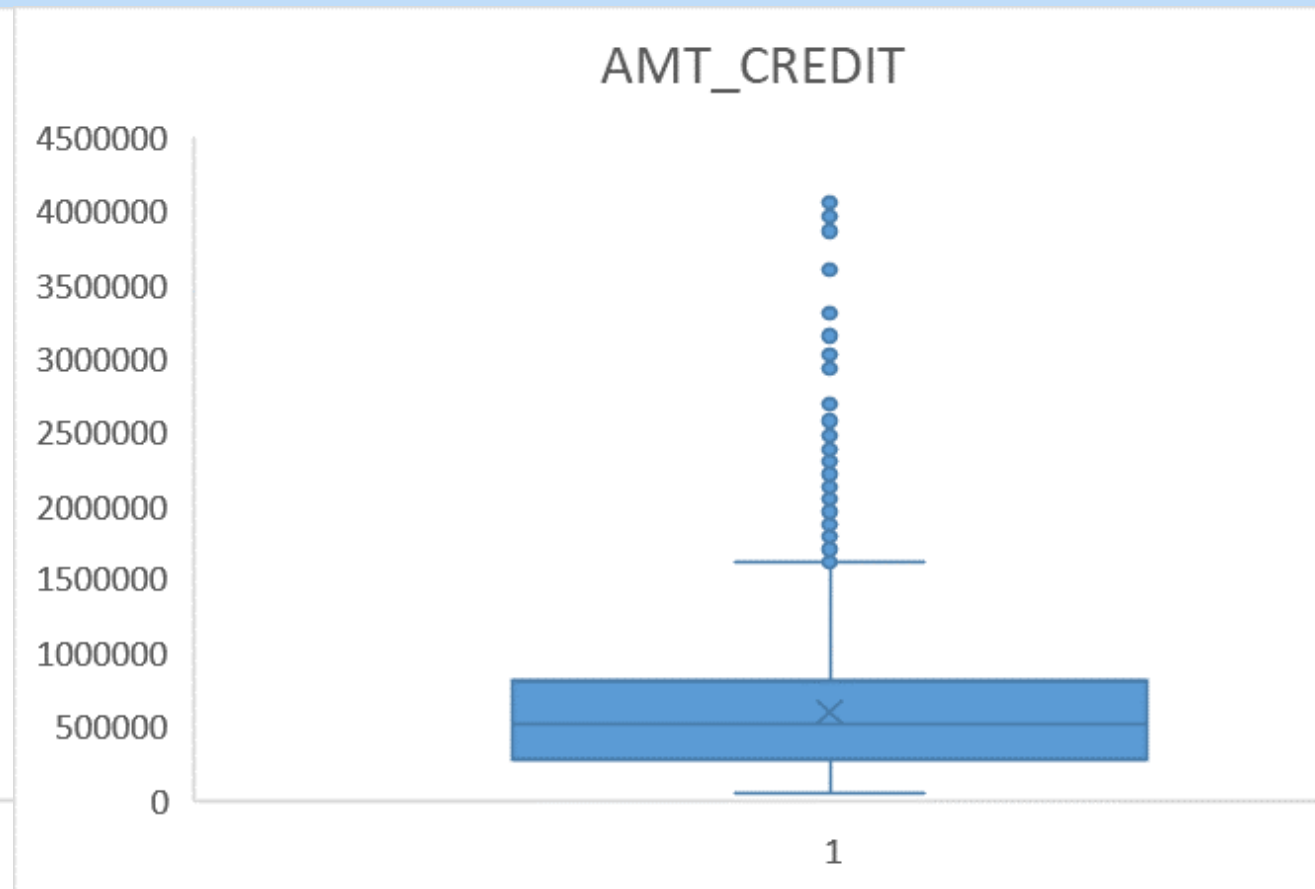
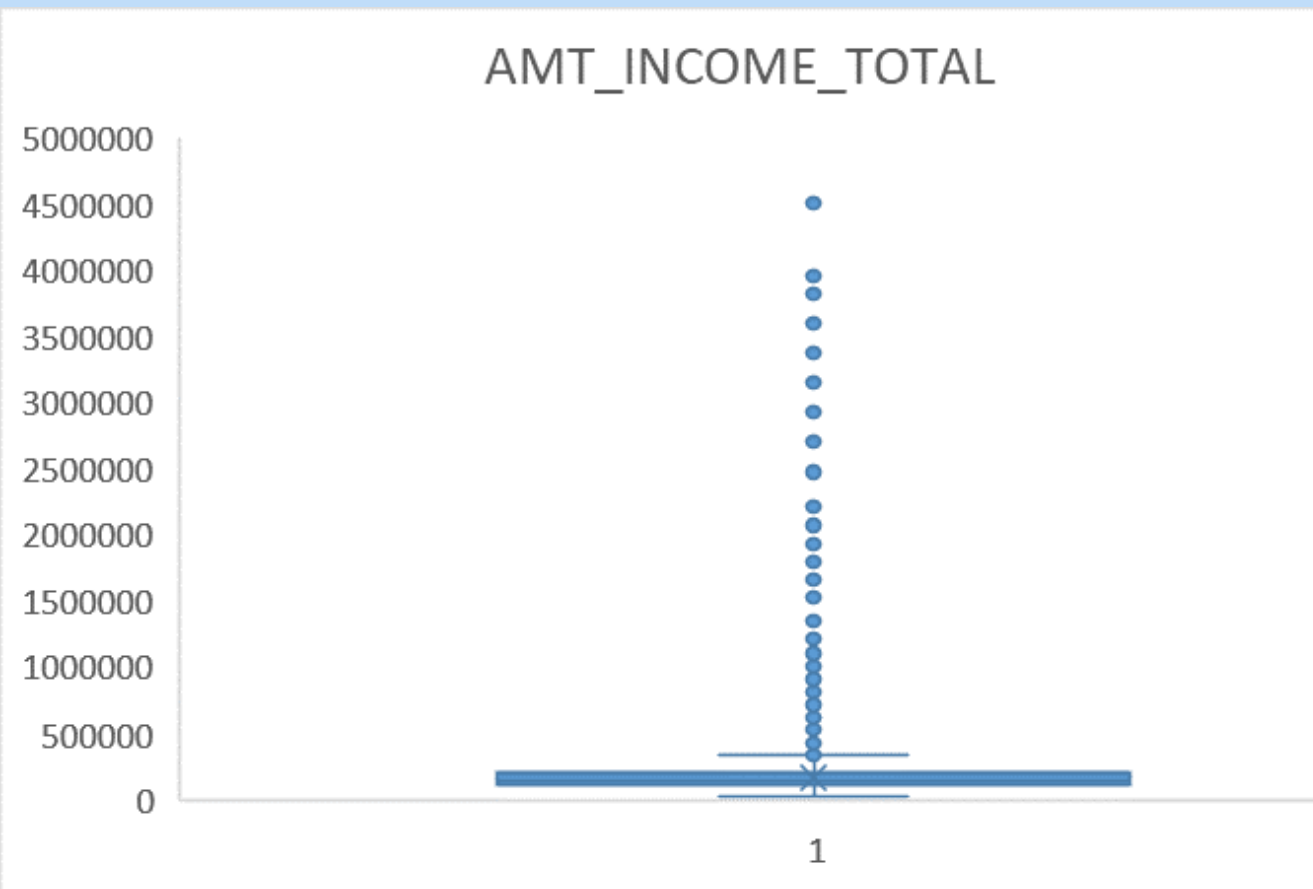


Without Outlier



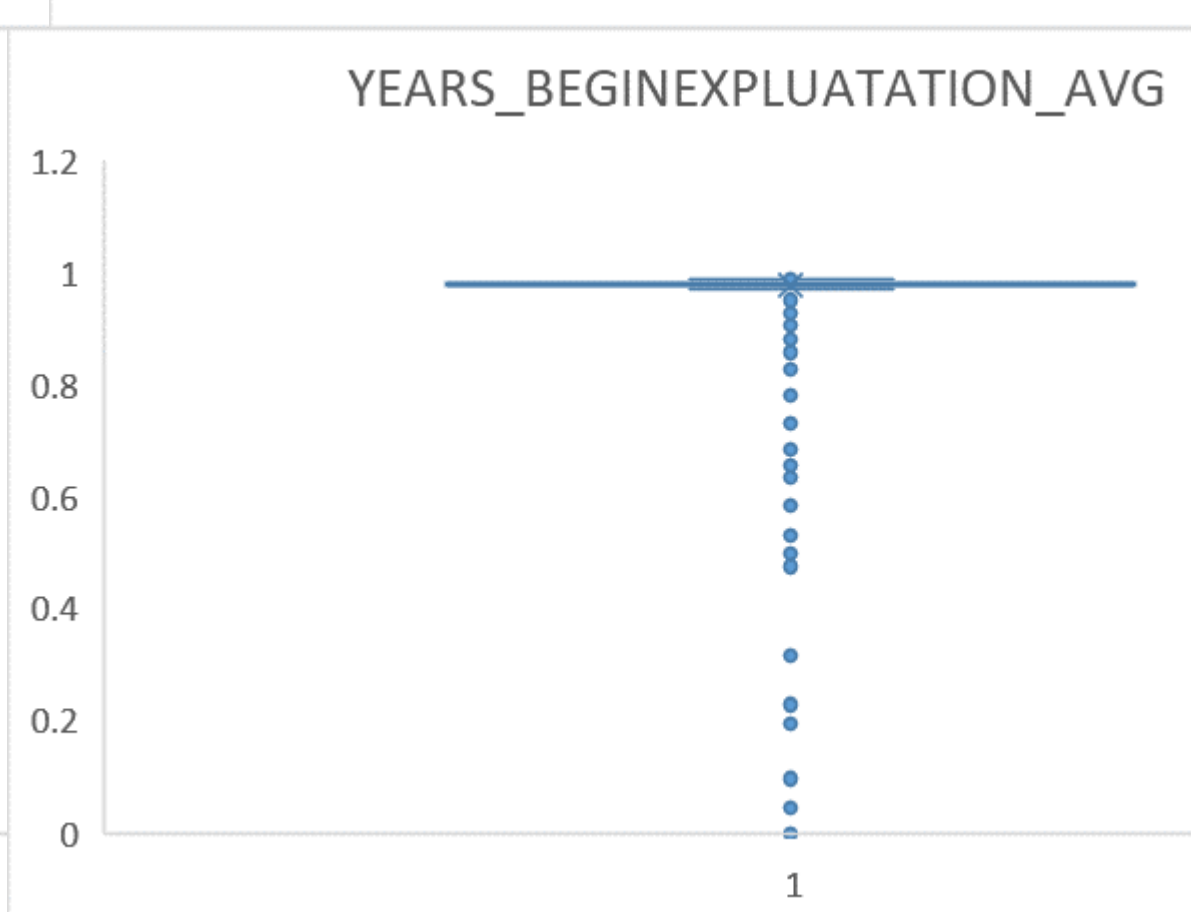
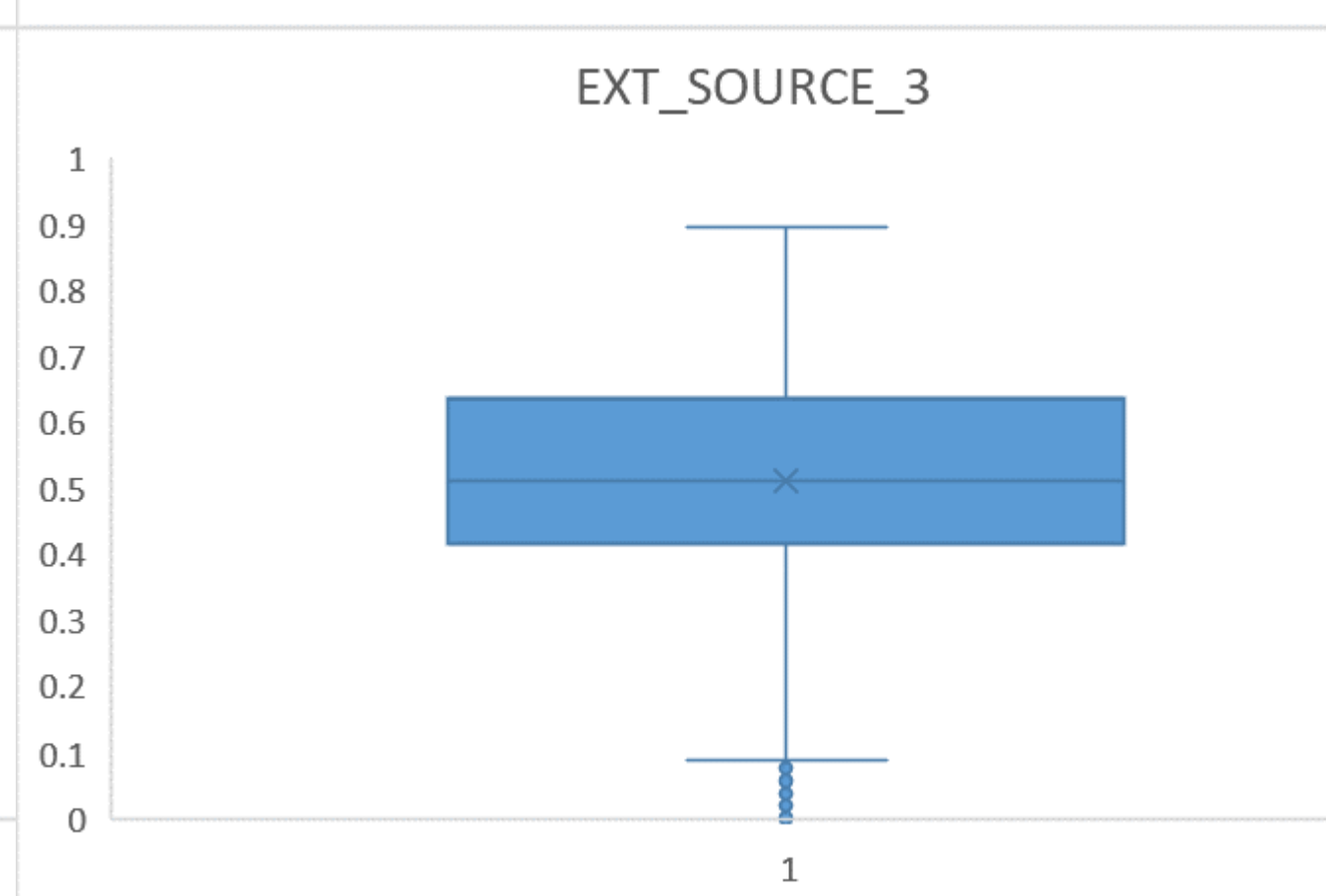
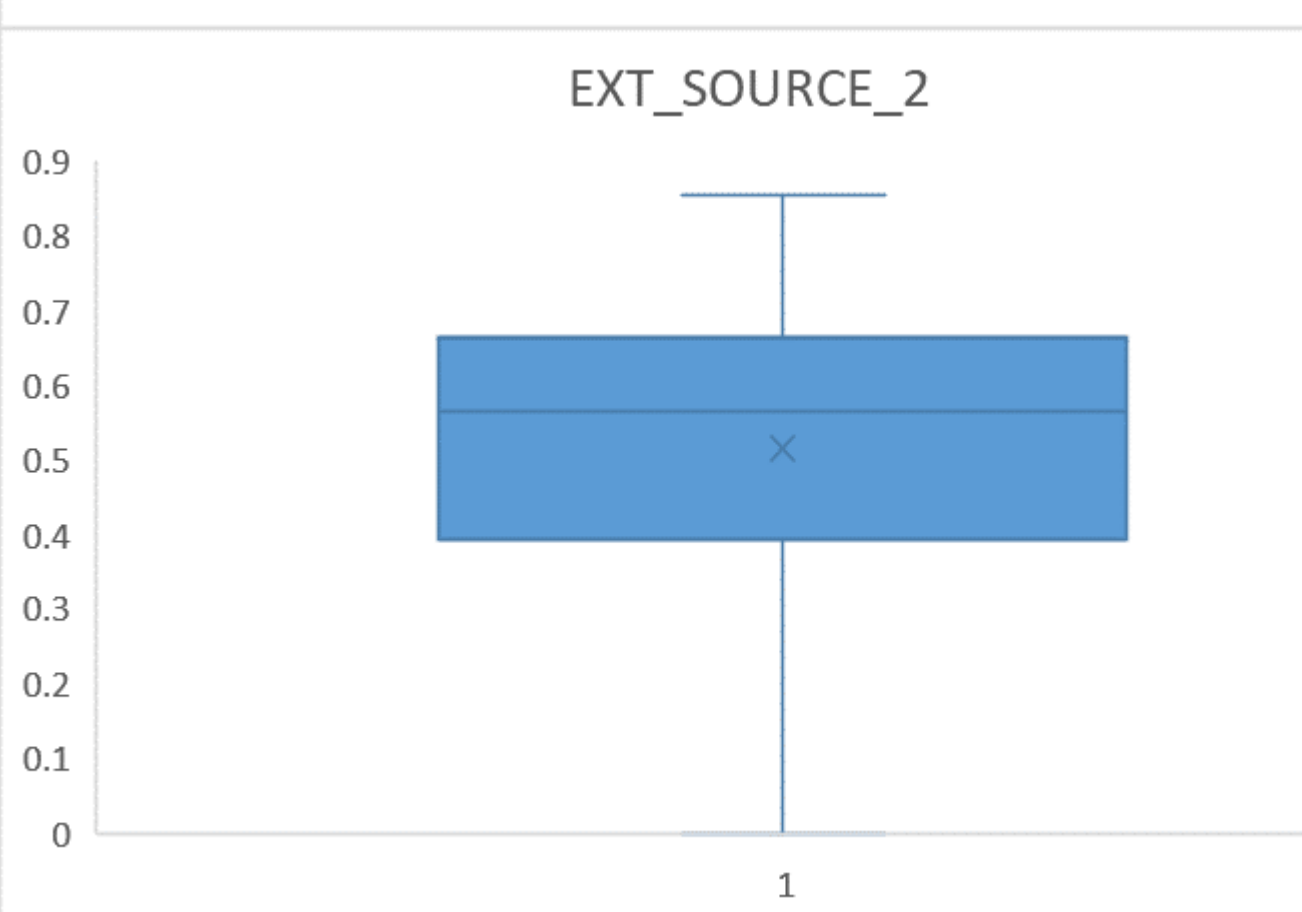
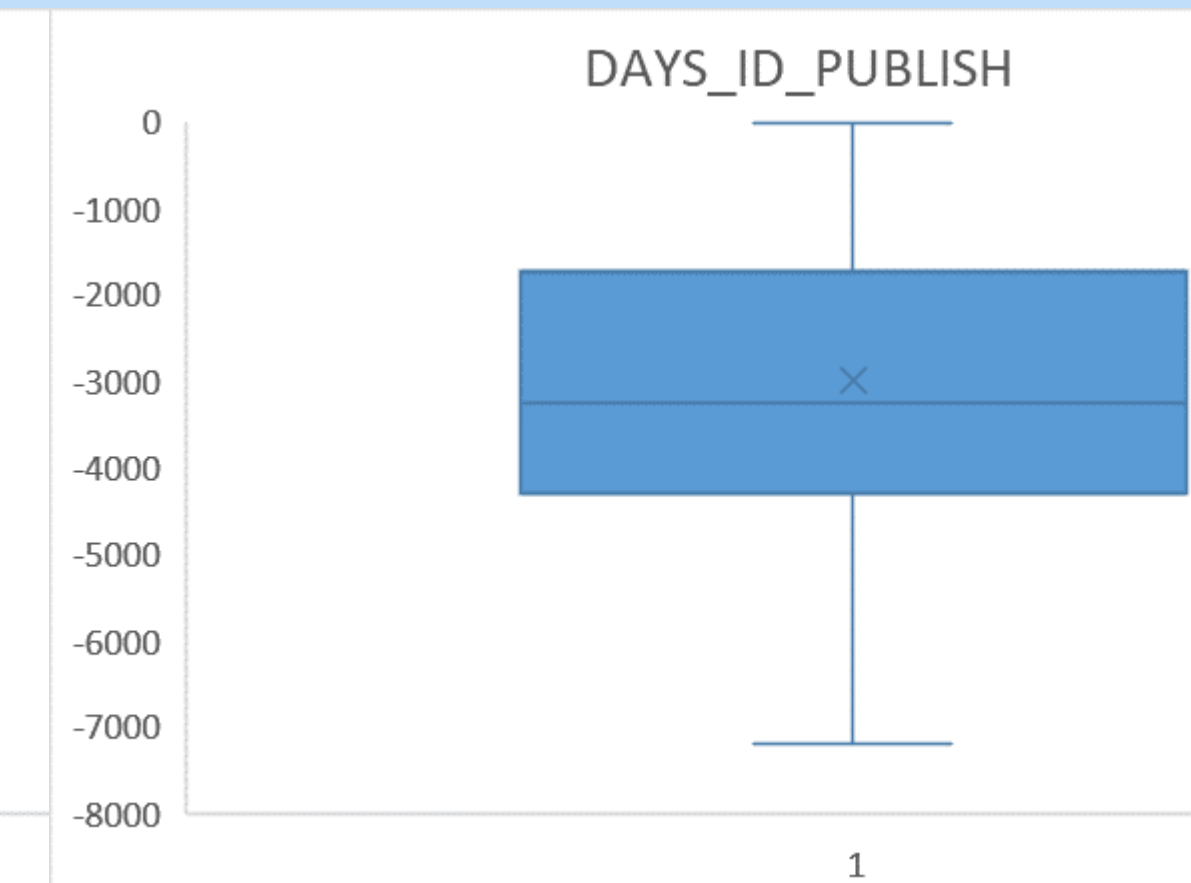
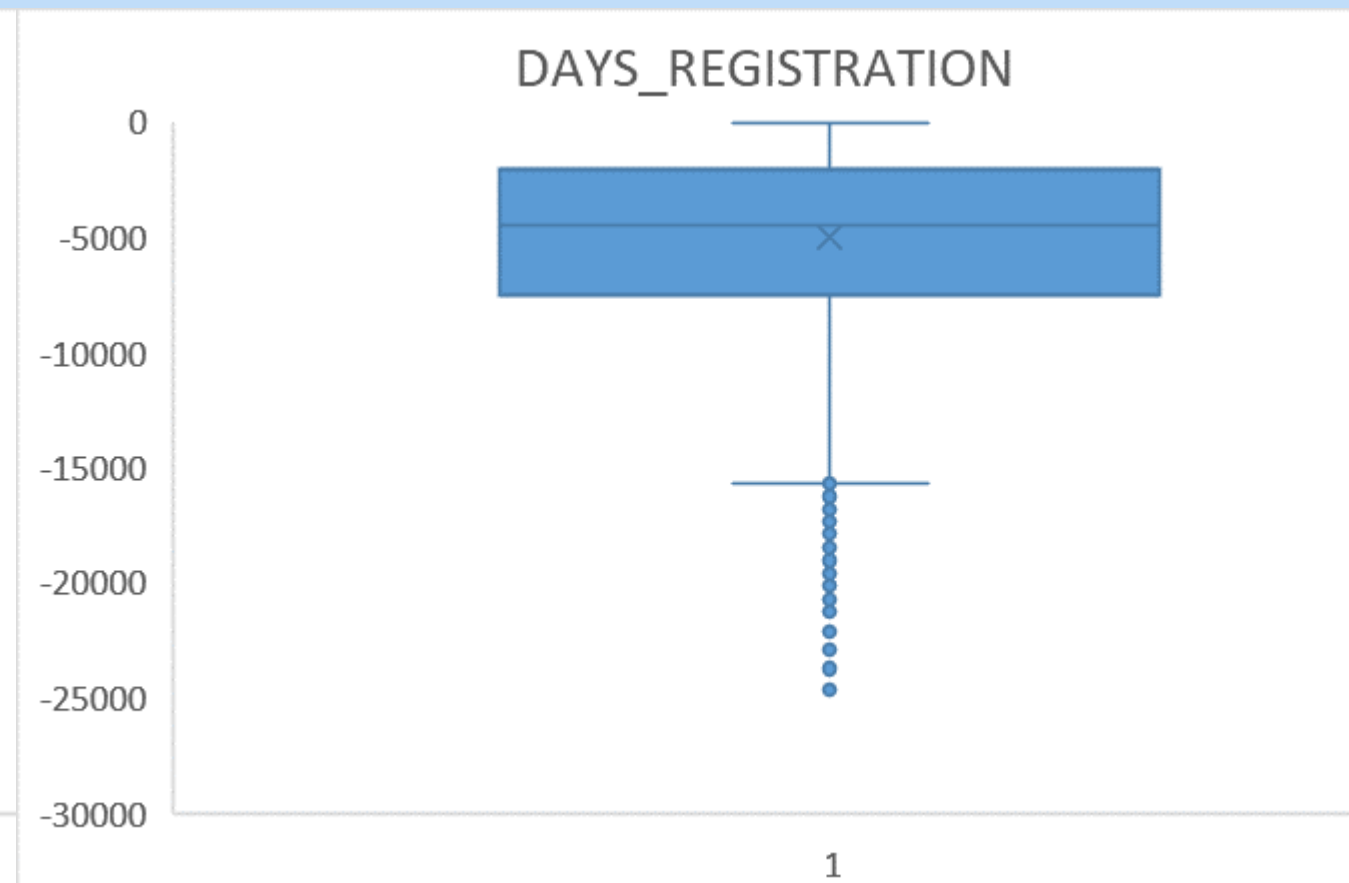
Insights

Outliers : Application Data



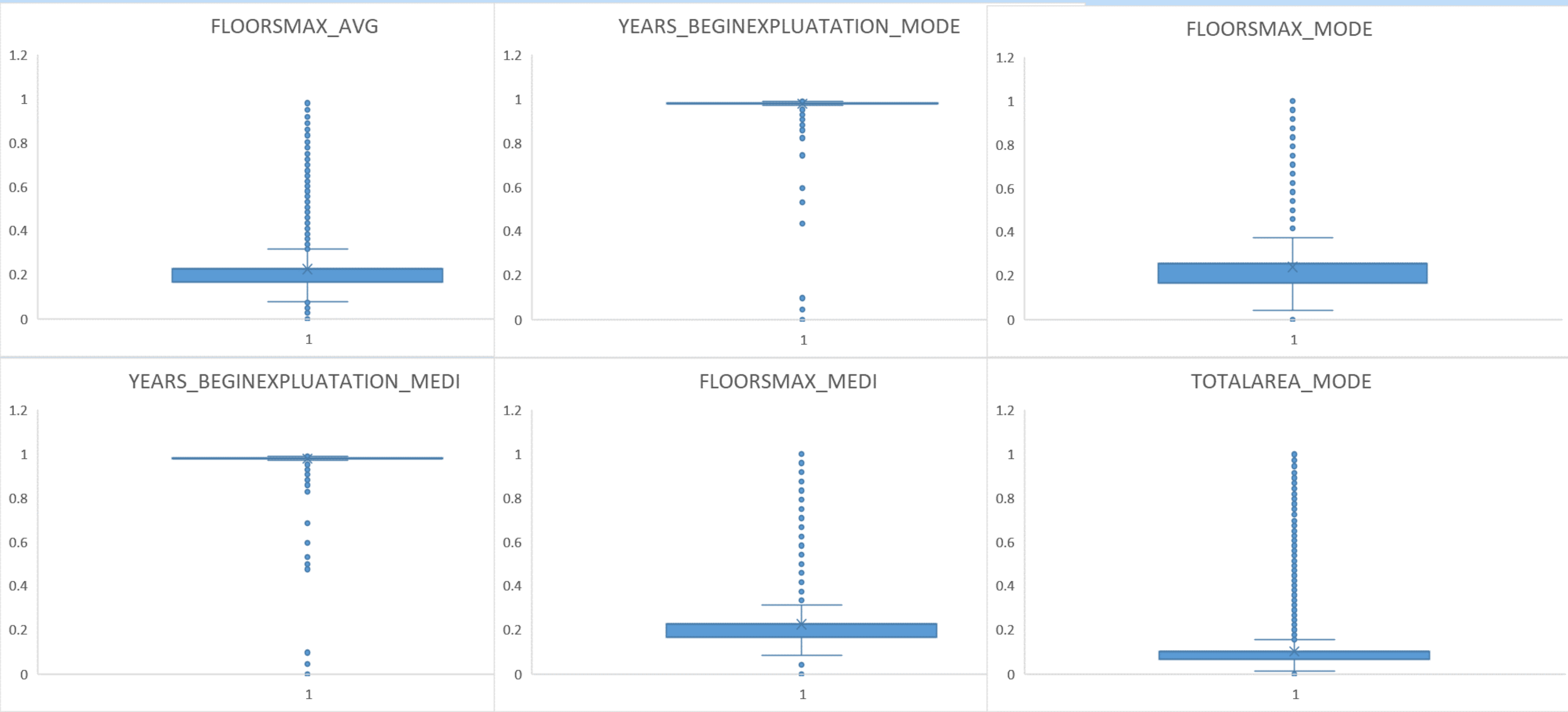
Insights

Outliers : Application Data



Insights

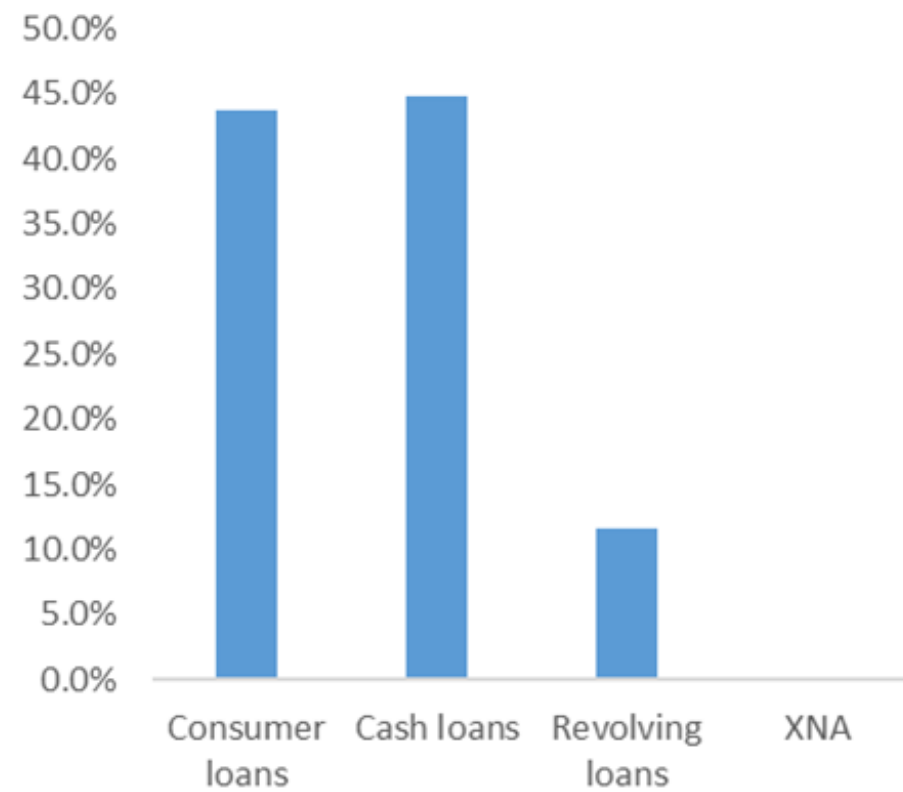
Outliers : Application Data



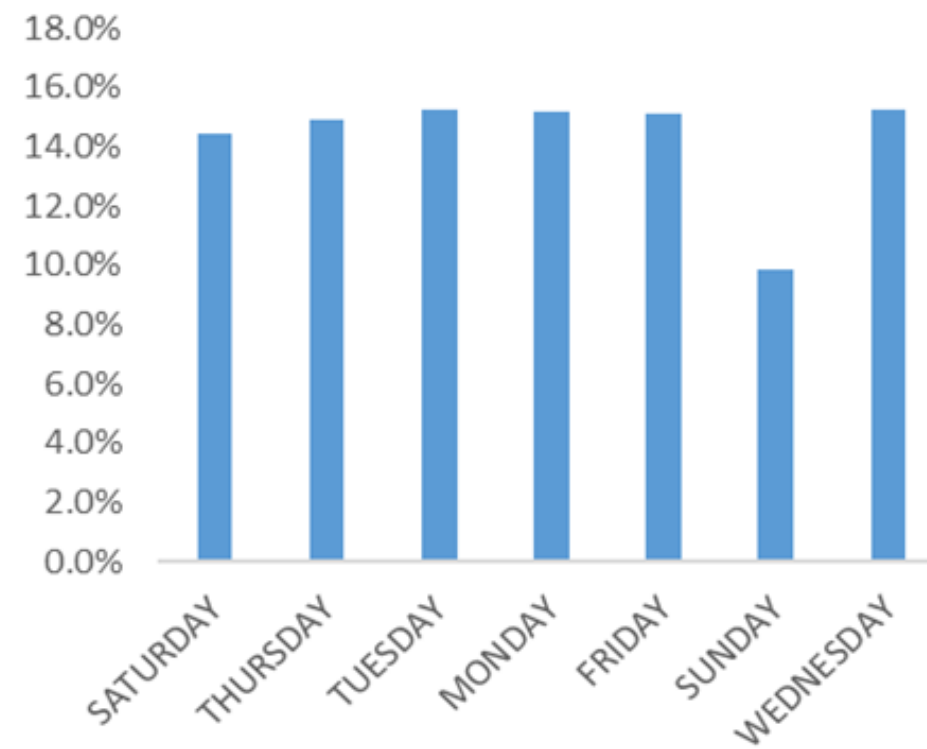
Insights

Data Imbalance:- Previous Application

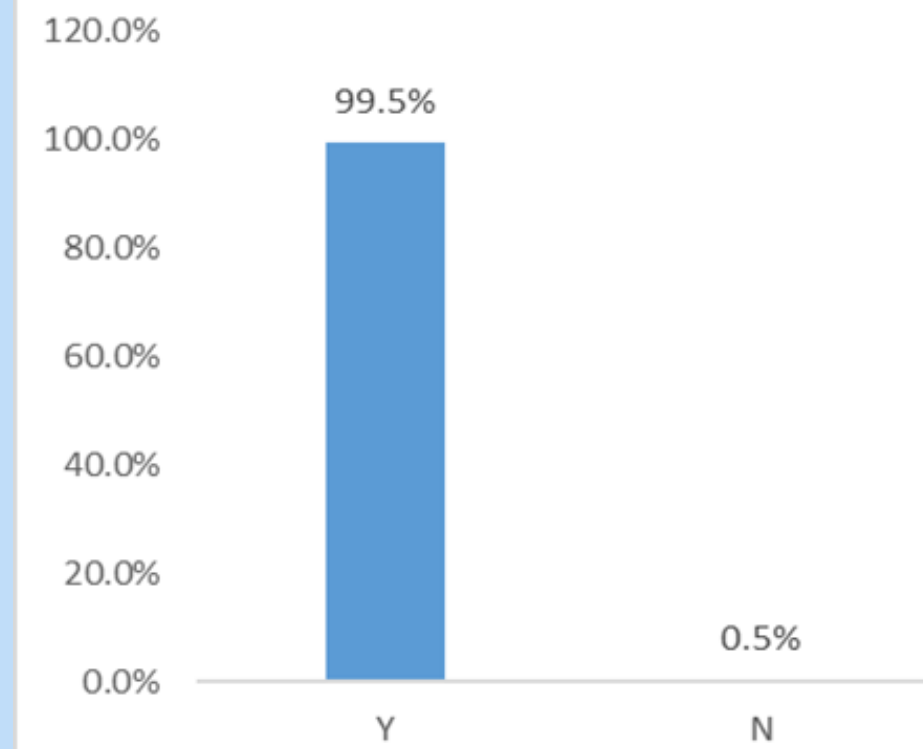
NAME_CONTRACT_TYPE



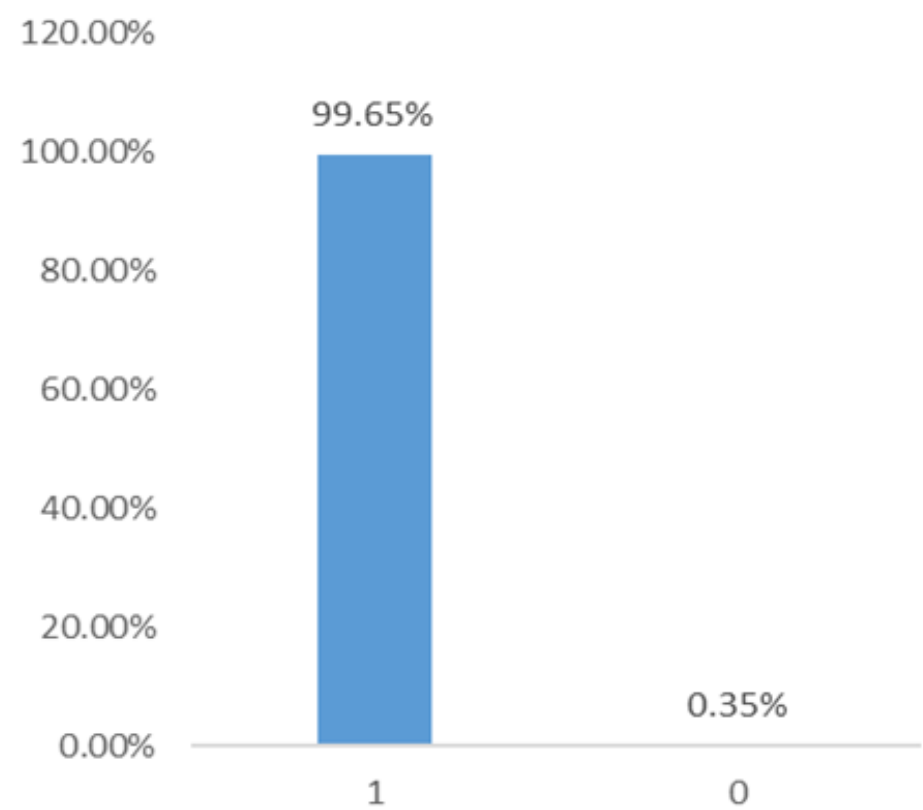
WEEKDAY_APPR_PROCESS_ST
ART



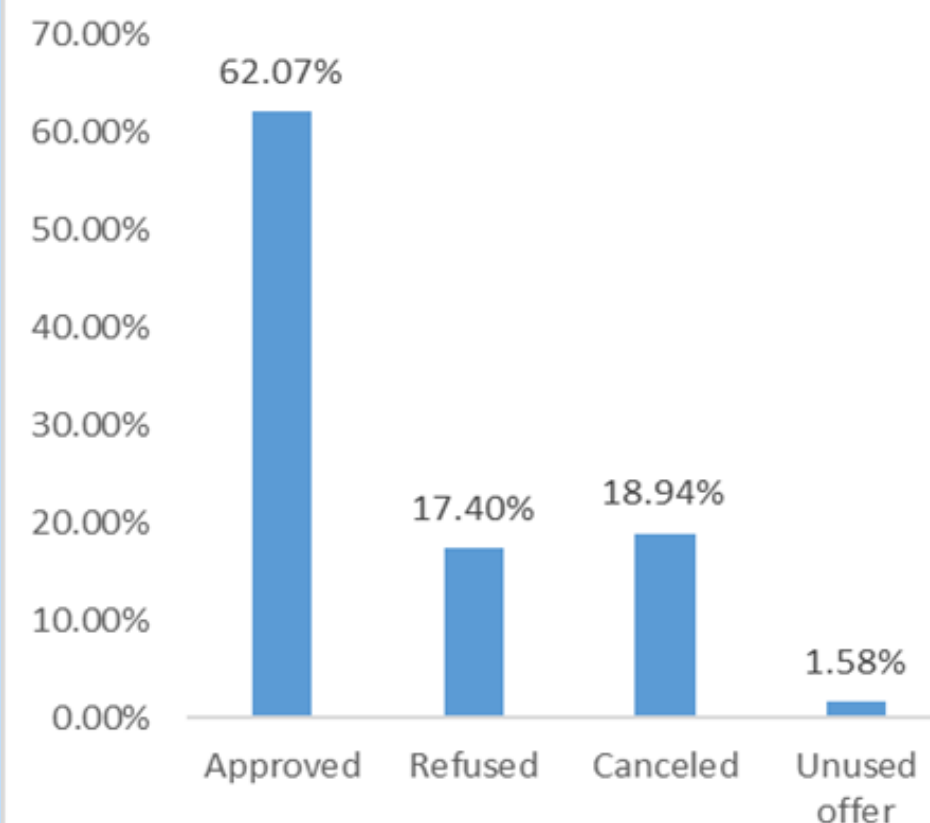
FLAG_LAST_APPL_PER_CONTR
ACT



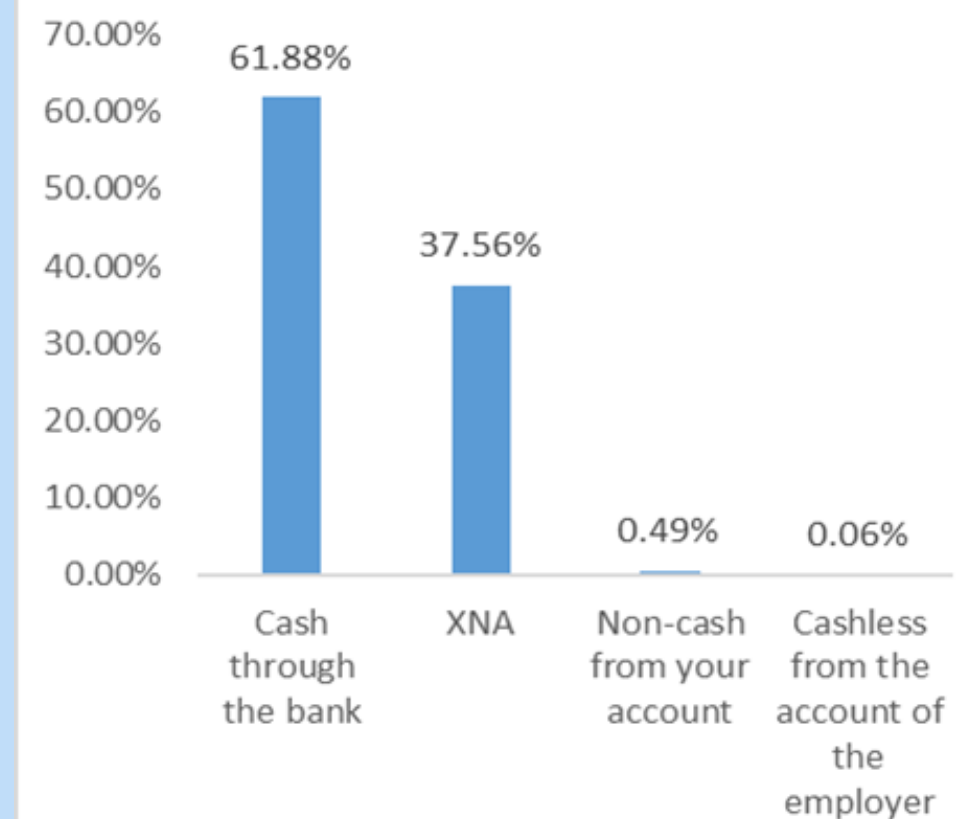
NFLAG_LAST_APPL_IN_DAY



NAME_CONTRACT_STATUS

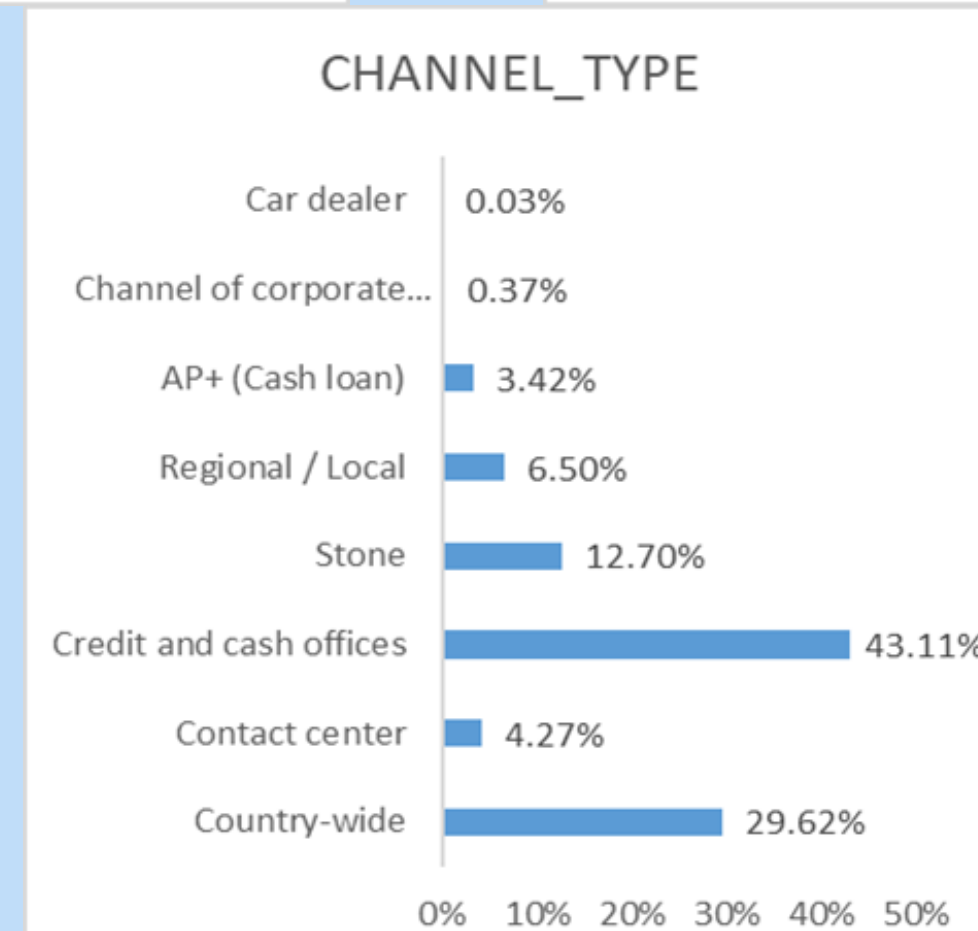
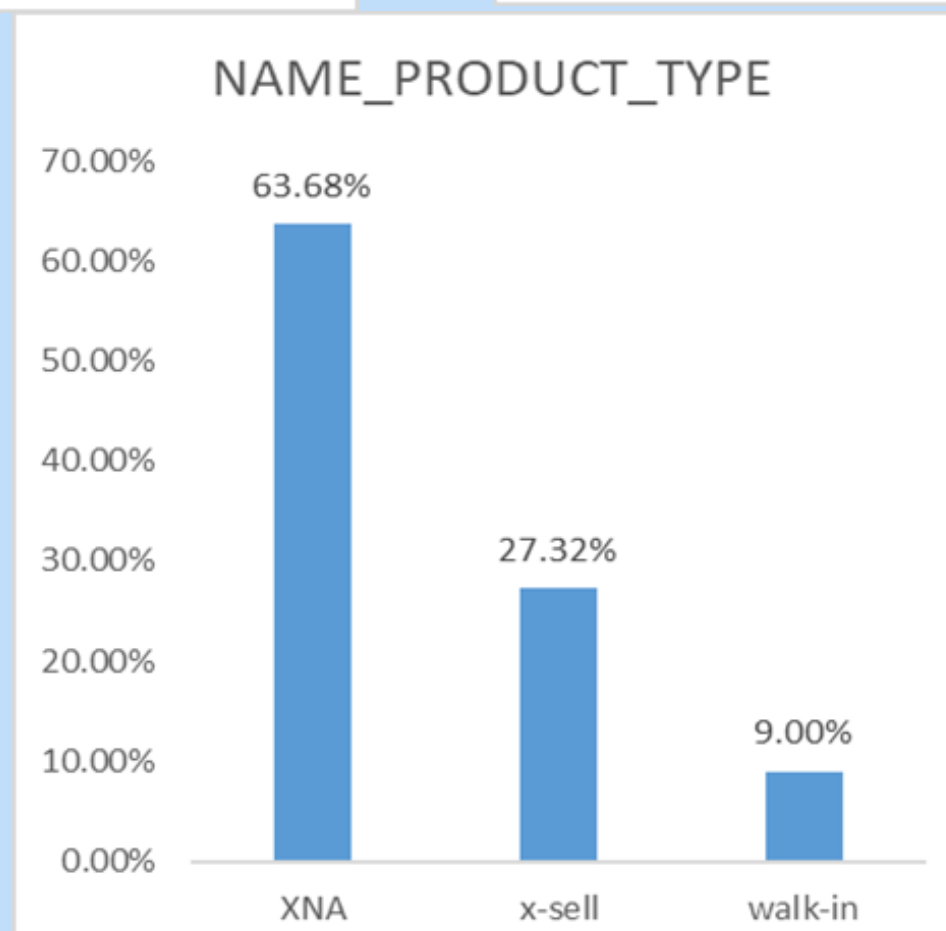
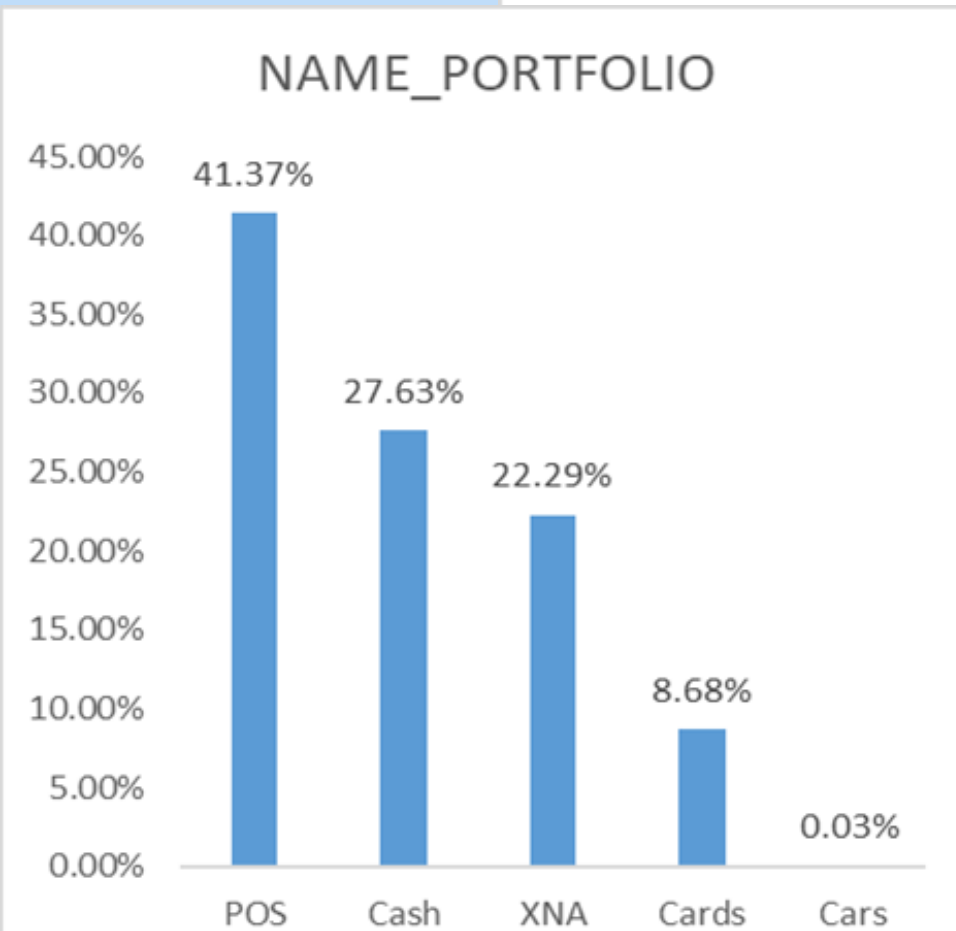
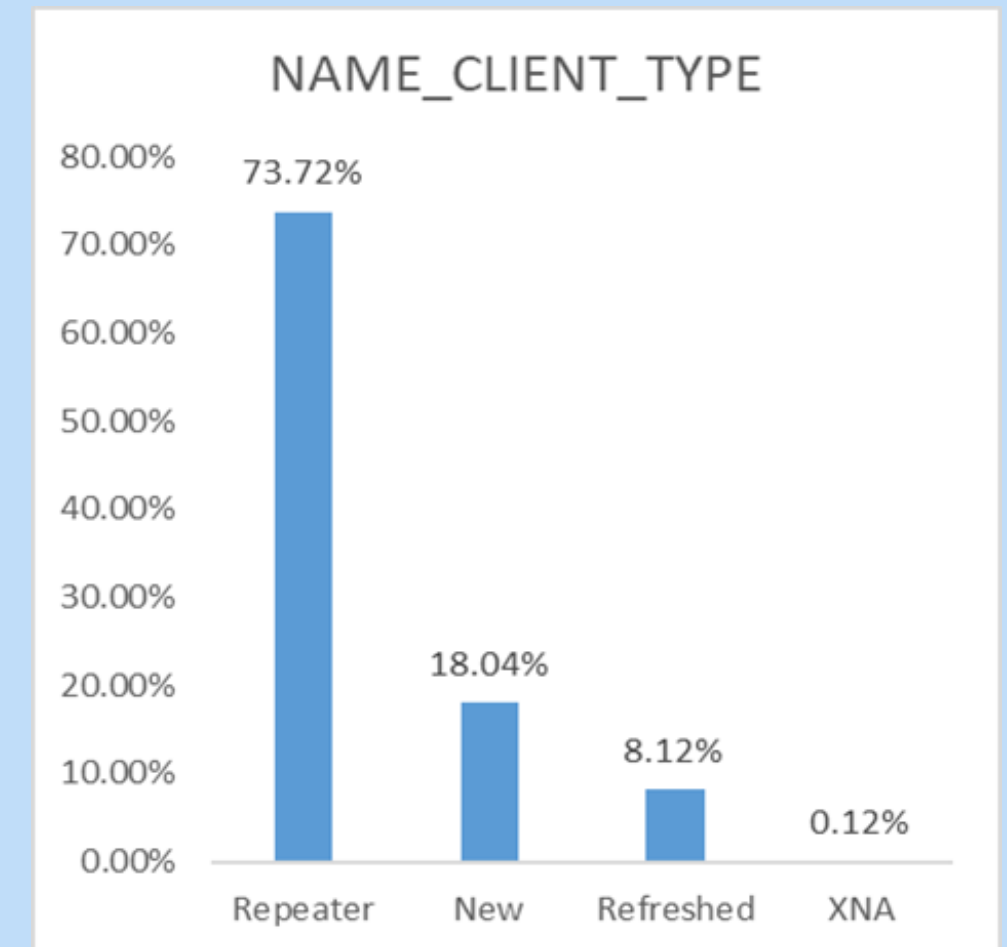
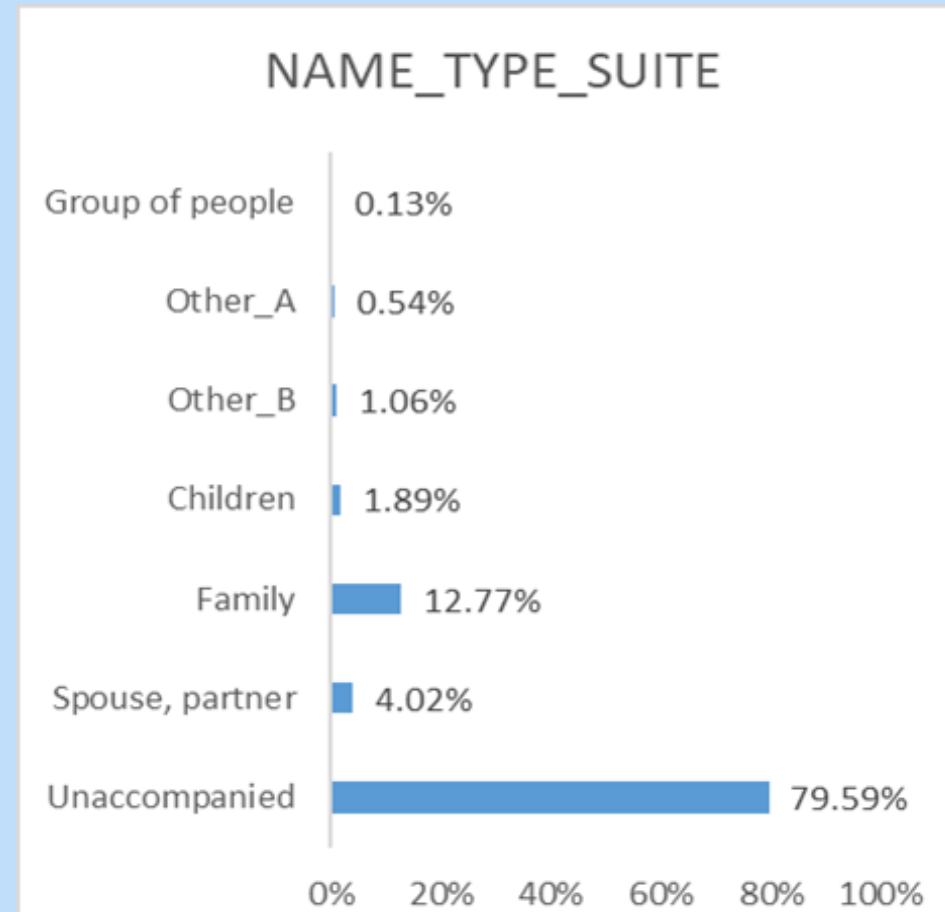
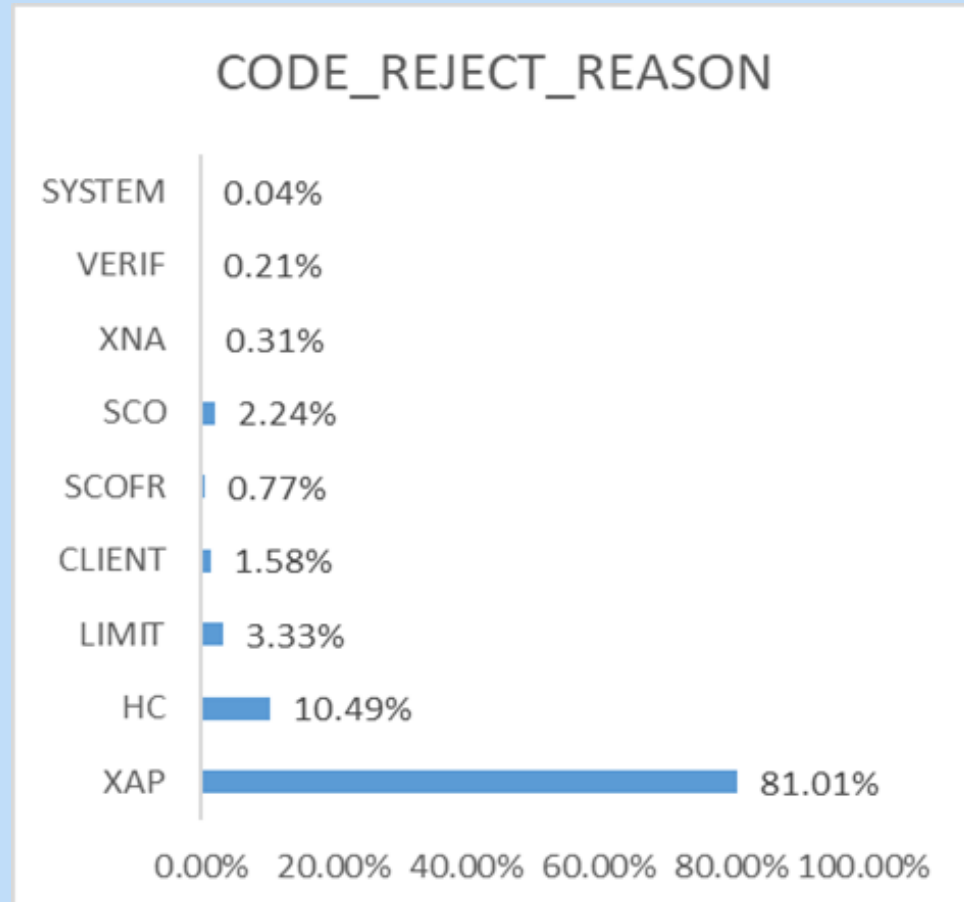


NAME_PAYMENT_TYPE



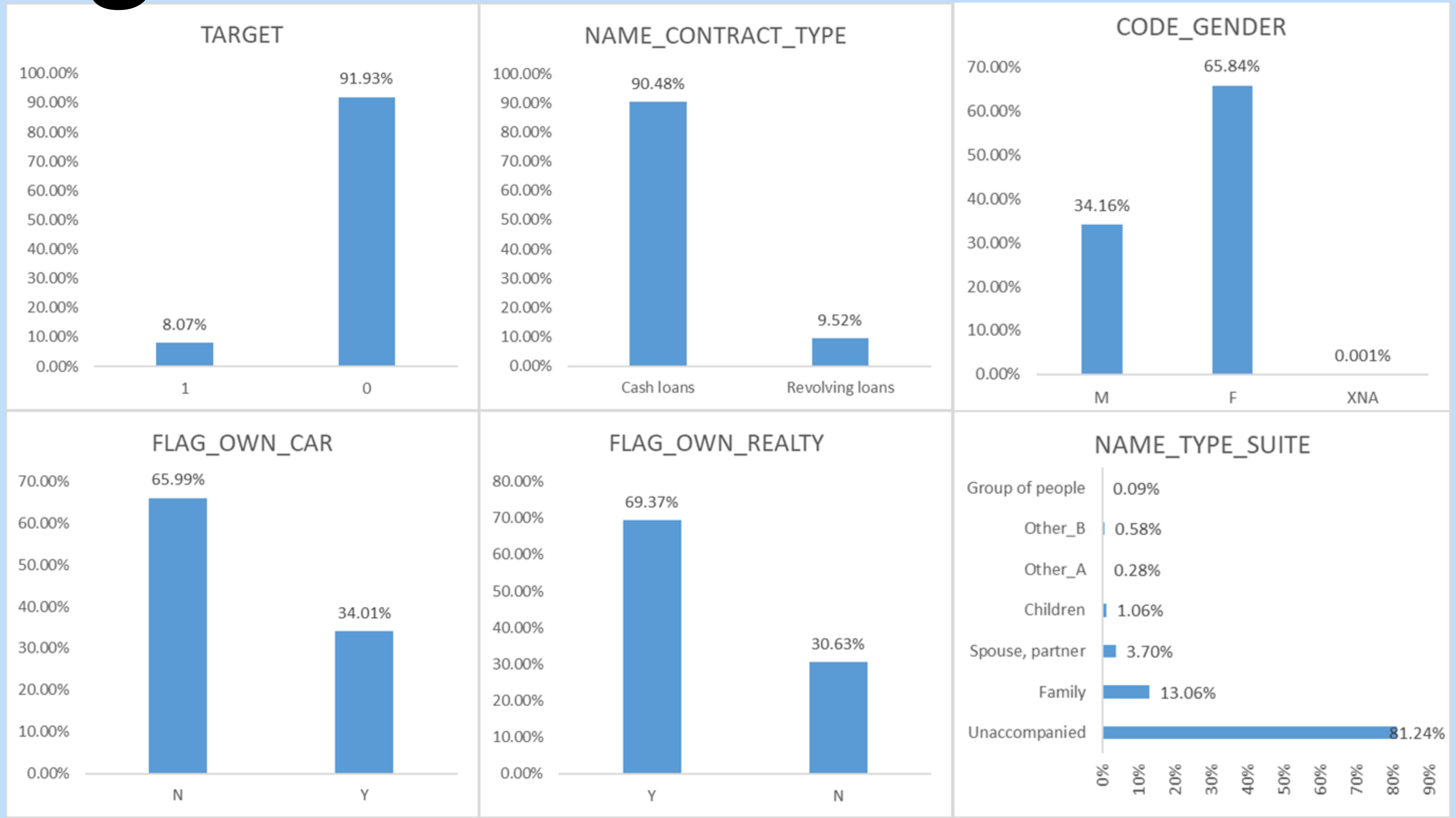
Insights

Data Imbalance:- Previous Application



Insights

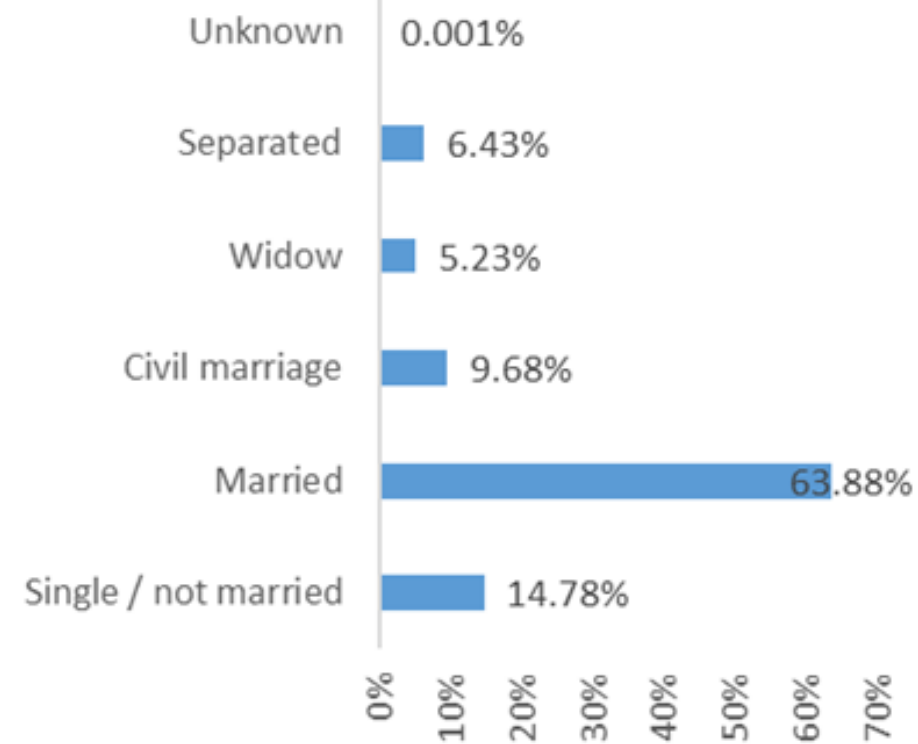
Data Imbalance:- Application Data



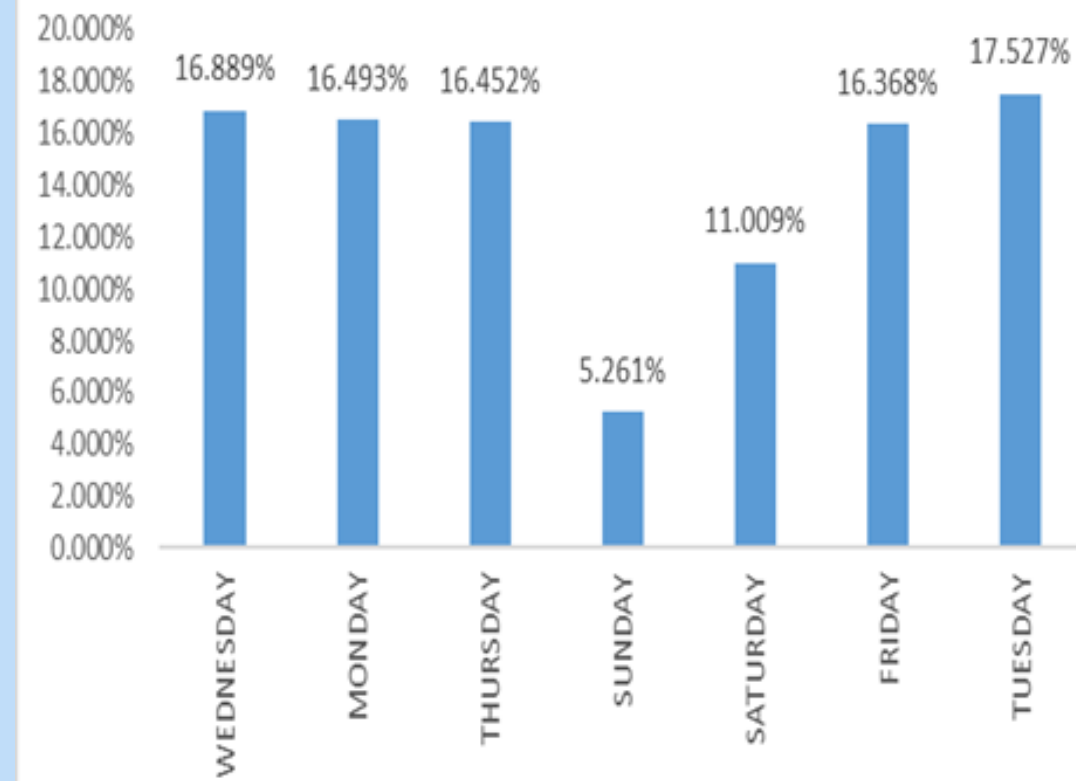
Insights

Data Imbalance:- Application Data

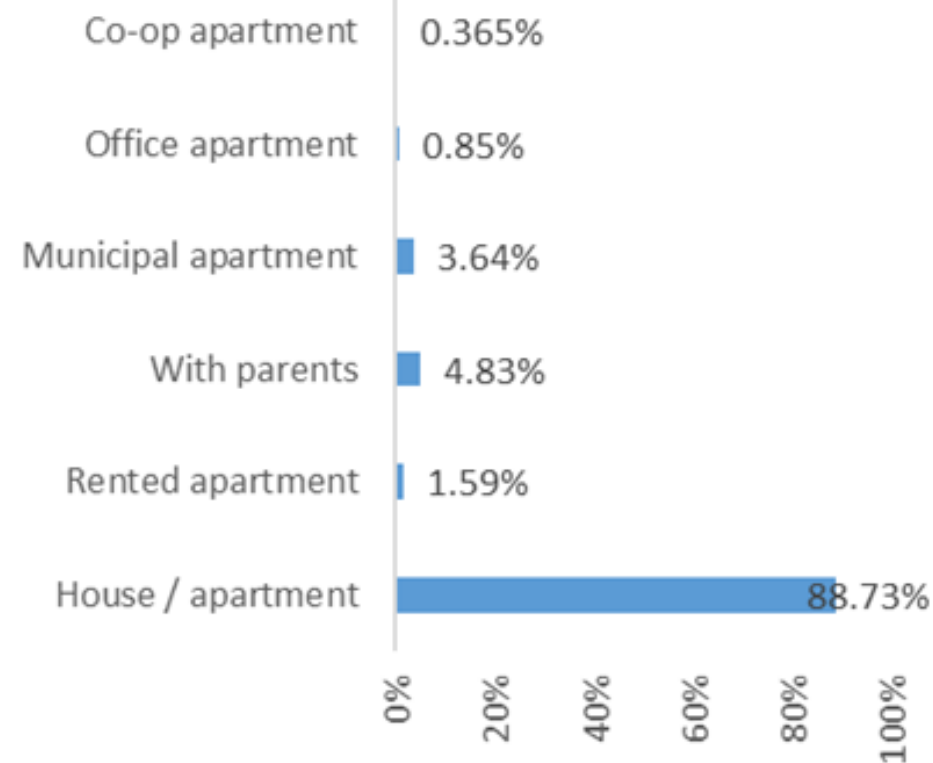
NAME_FAMILY_STATUS



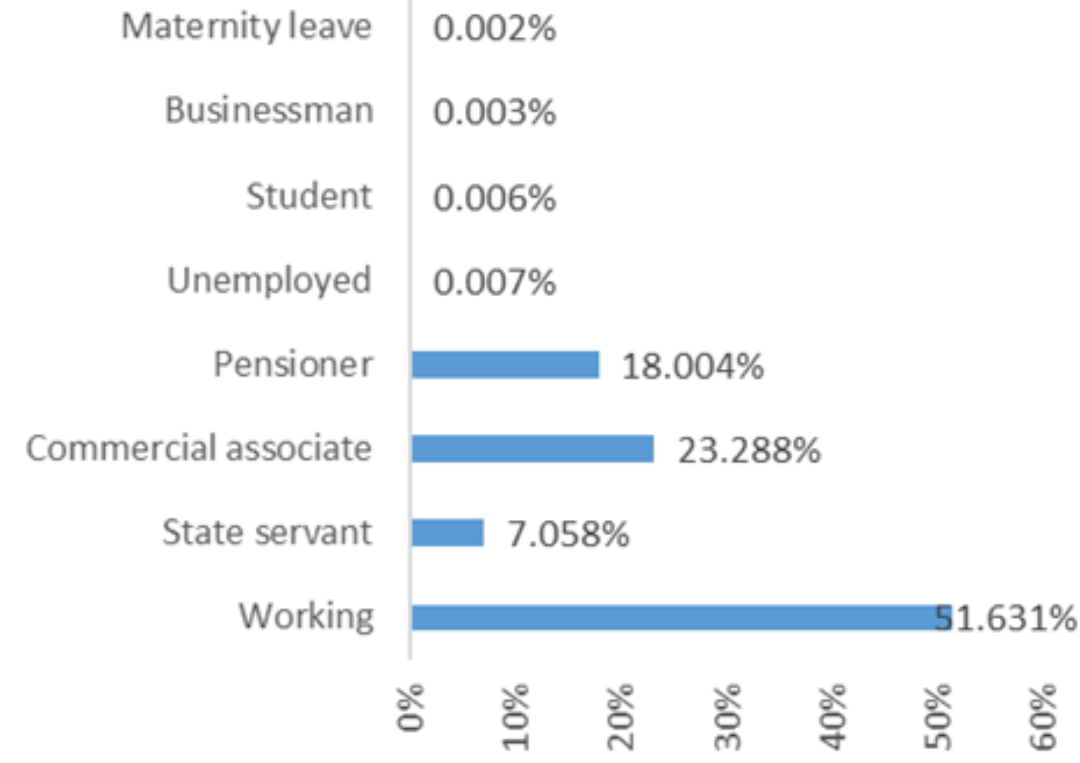
WEEKDAY_APPR_PROCESS_START



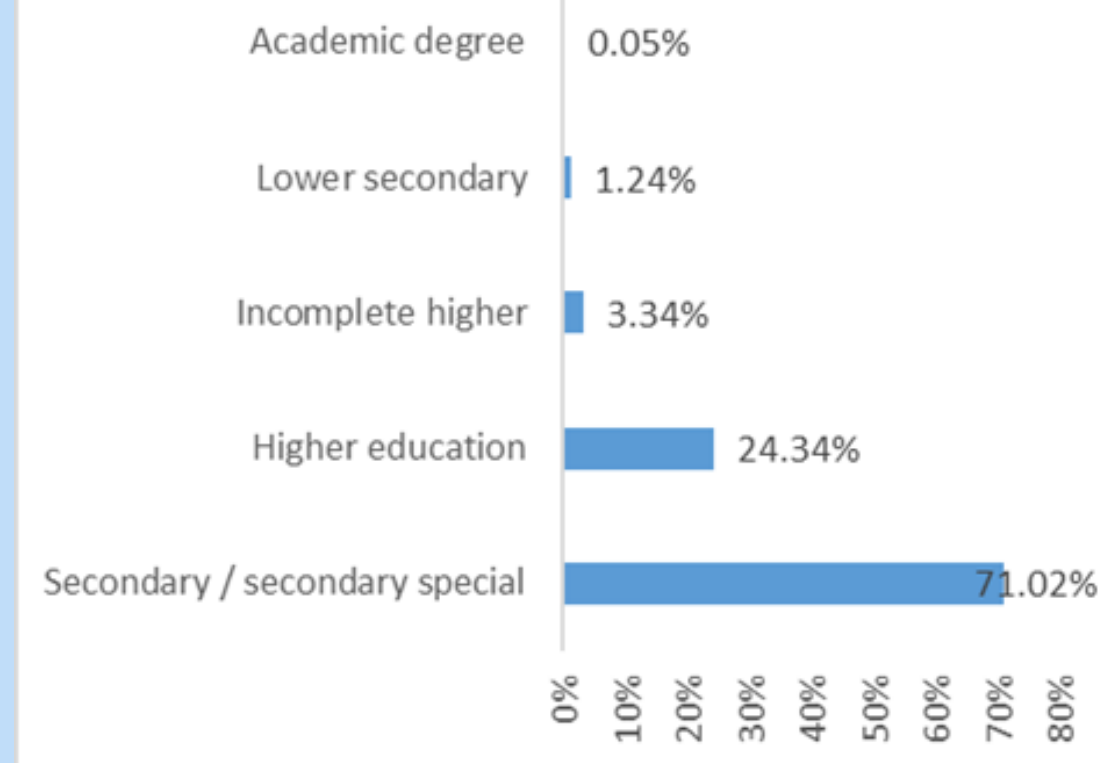
NAME_HOUSING_TYPE



NAME_INCOME_TYPE



NAME_EDUCATION_TYPE



Univariate Analysis:

- **Univariate analysis:** Simplest data analysis with one variable.
- **Purpose:** Describe data, find patterns.
- **Business application:** Understand attribute distribution (e.g., customer age, product price, sales volume).
- **Benefits:** Insights for decision-making, strategic planning

Segmented Univariate Analysis:

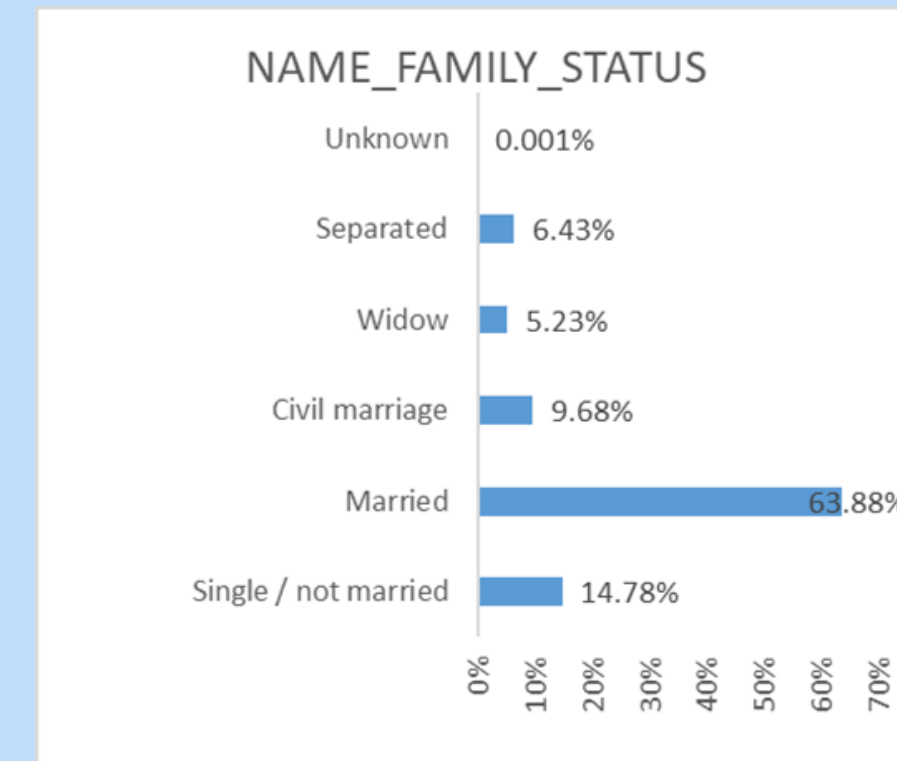
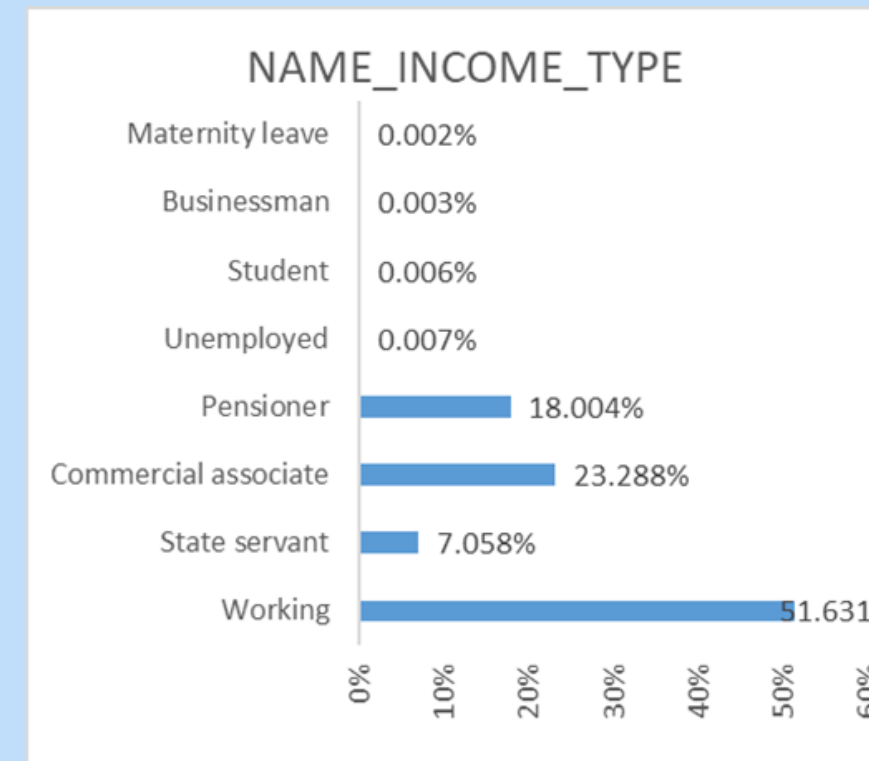
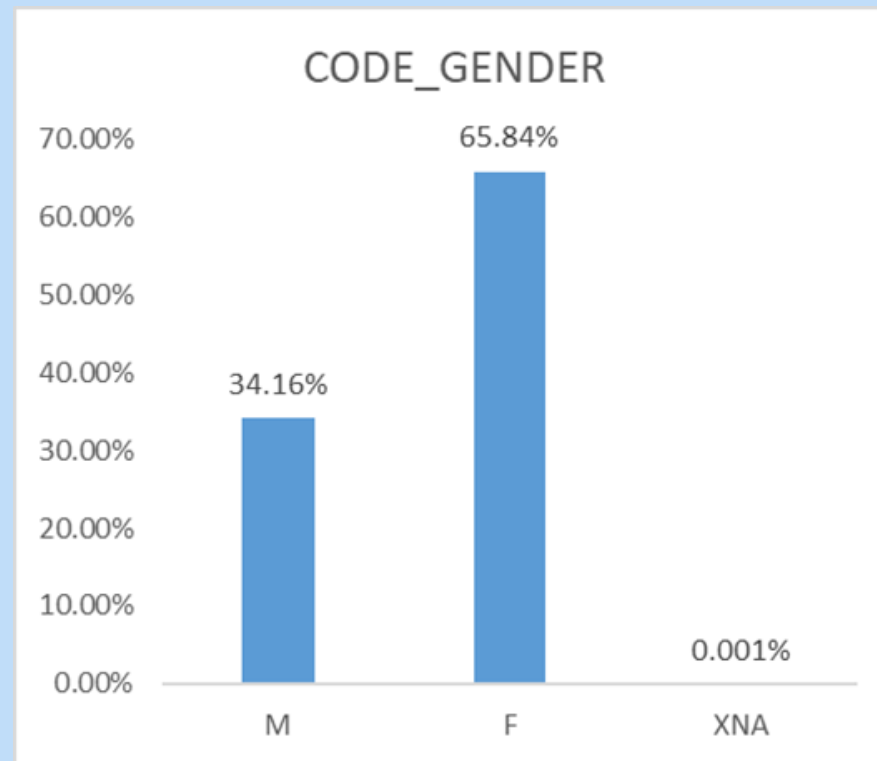
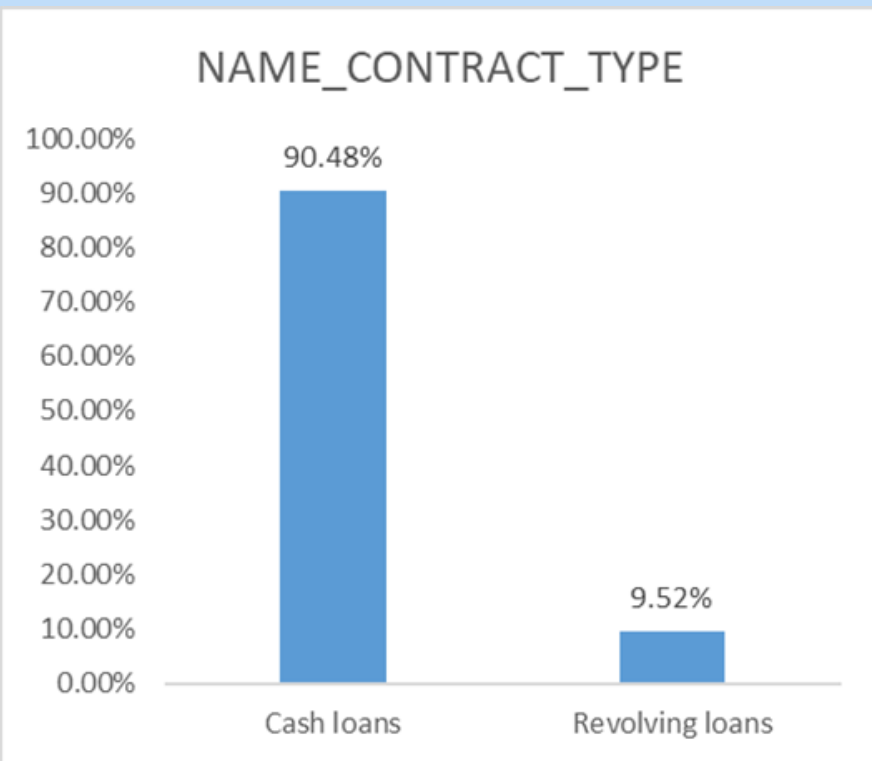
- **Segmented univariate analysis:** Analyzing data within different segments or categories.
- **Purpose:** Understand distribution patterns within specific segments.
- **Business application:** Segmenting data by product categories, geographical regions, etc.
- **Benefits:** Identify trends and patterns unique to specific segments, enhance targeted decision-making and strategic planning.

Bivariate Analysis:

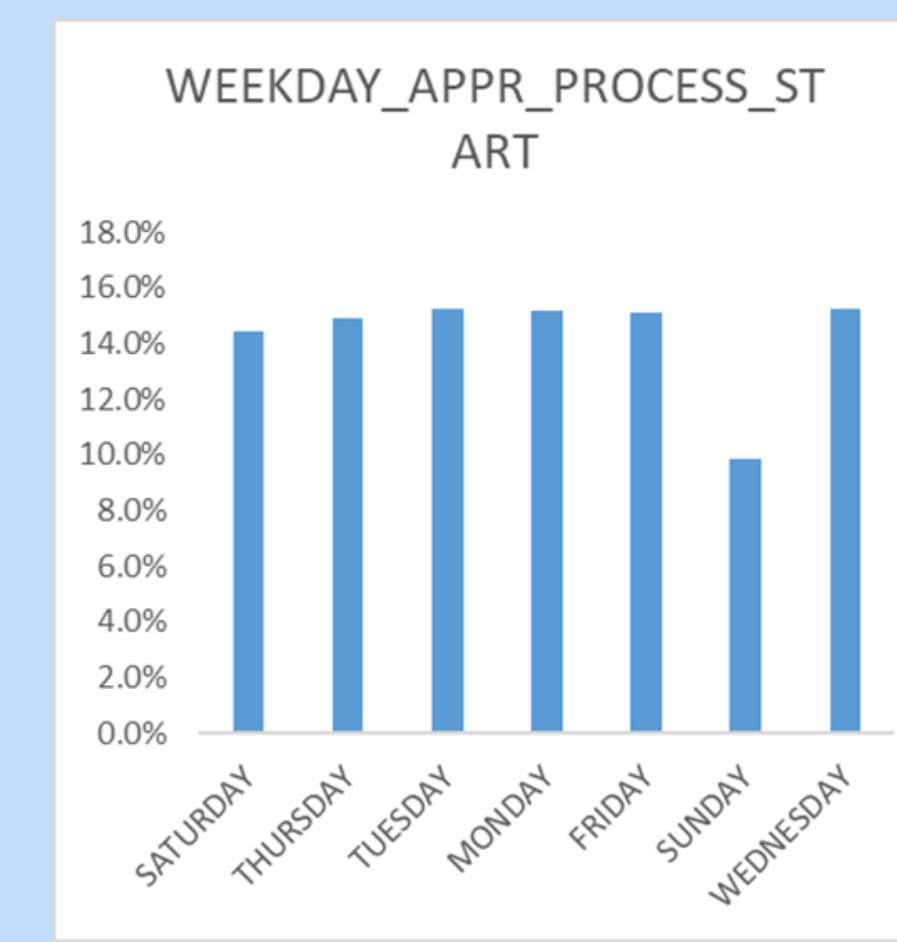
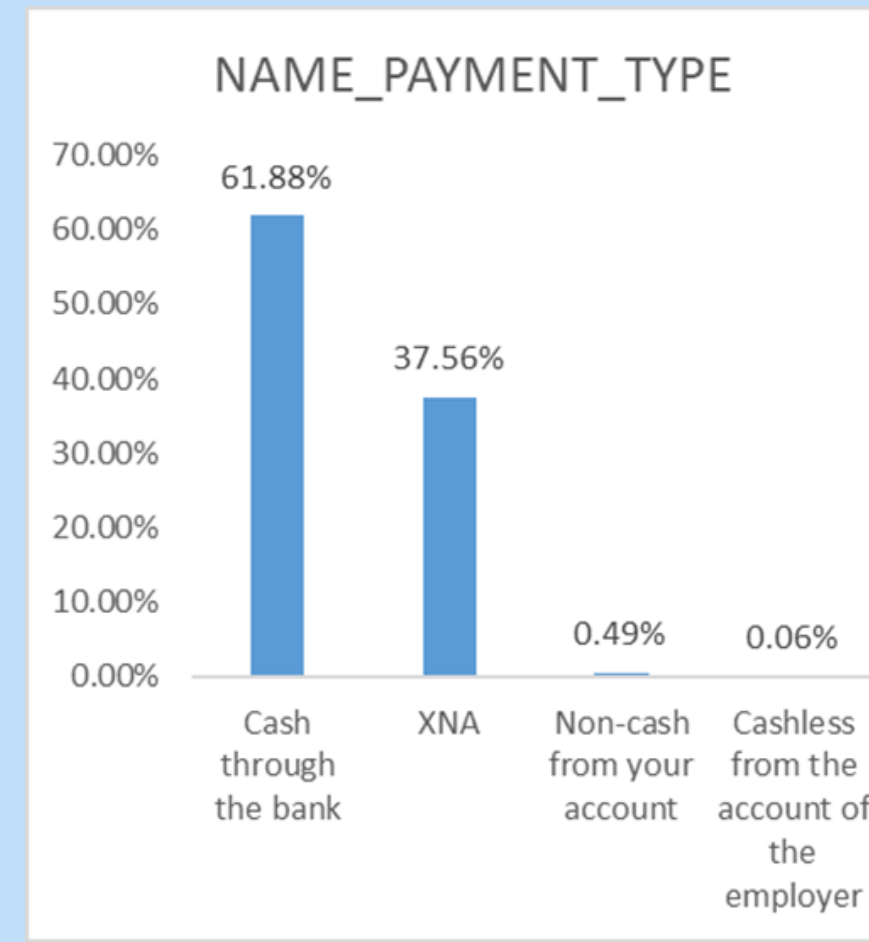
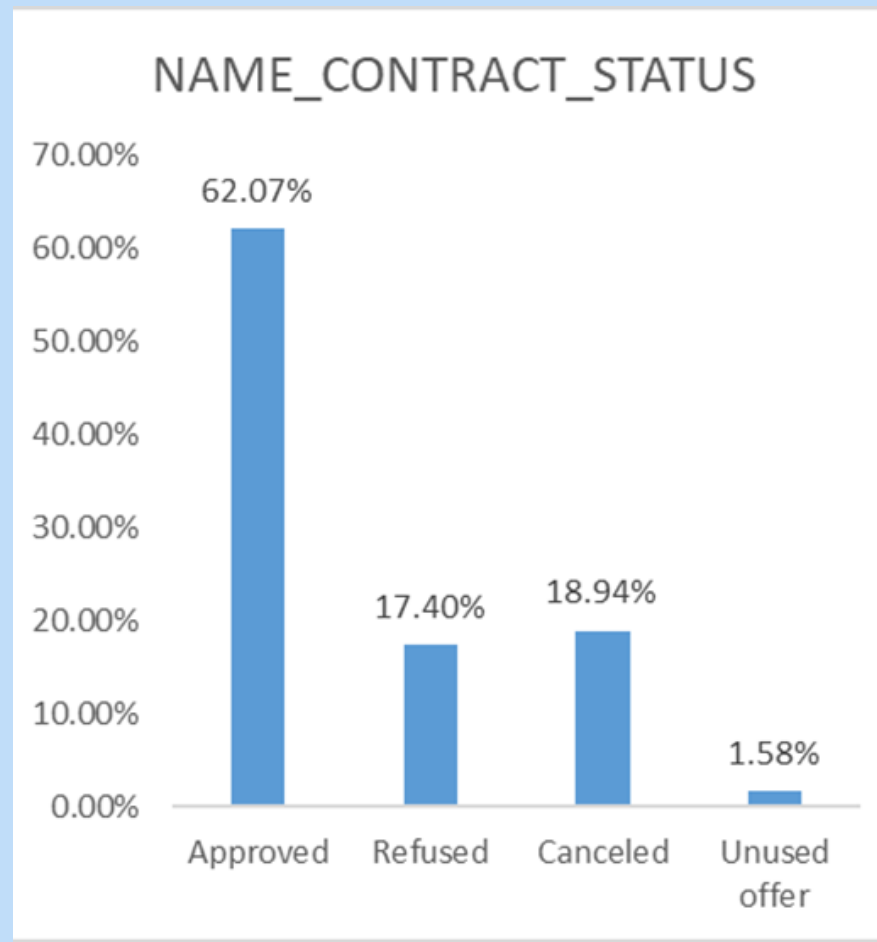
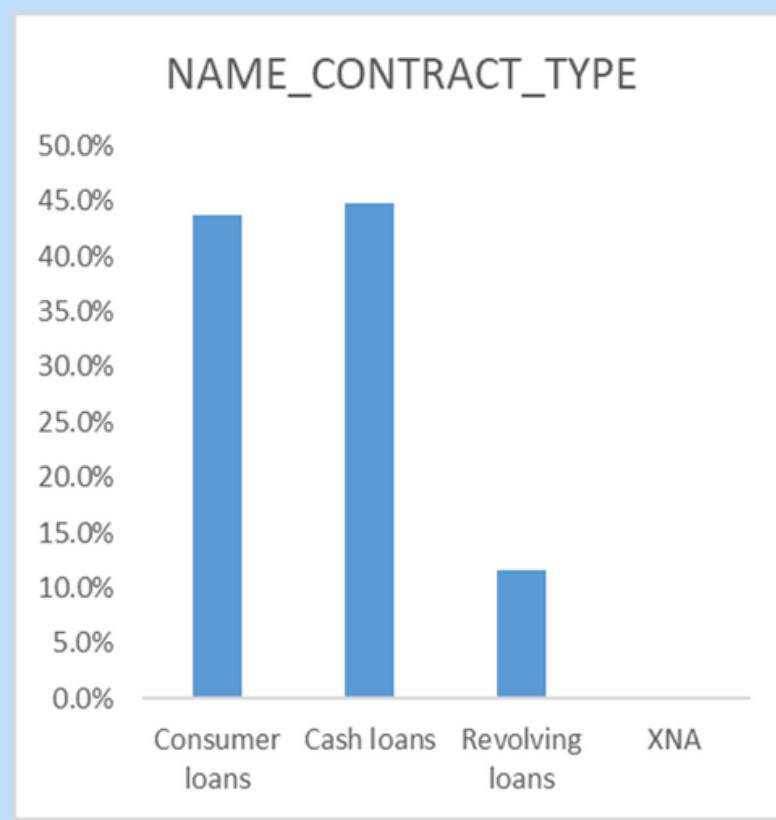
- **Bivariate analysis:** Analyzing two variables to understand their relationship.
- **Purpose:** Investigate causes and relationships between the variables.
- **Business application:** Examining relationships like advertising spend and sales revenue, customer age and product preference.
- **Benefits:** Uncover insights on the impact and correlation between variables, inform marketing strategies and customer targeting.

Insights

Application Data

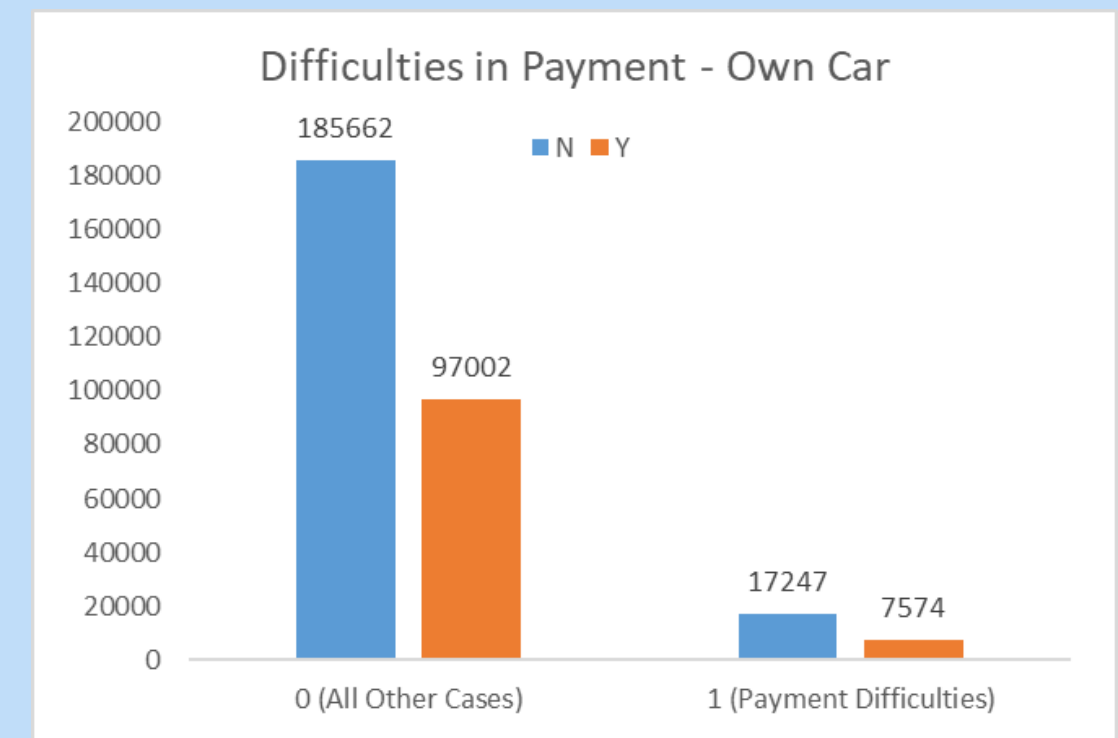
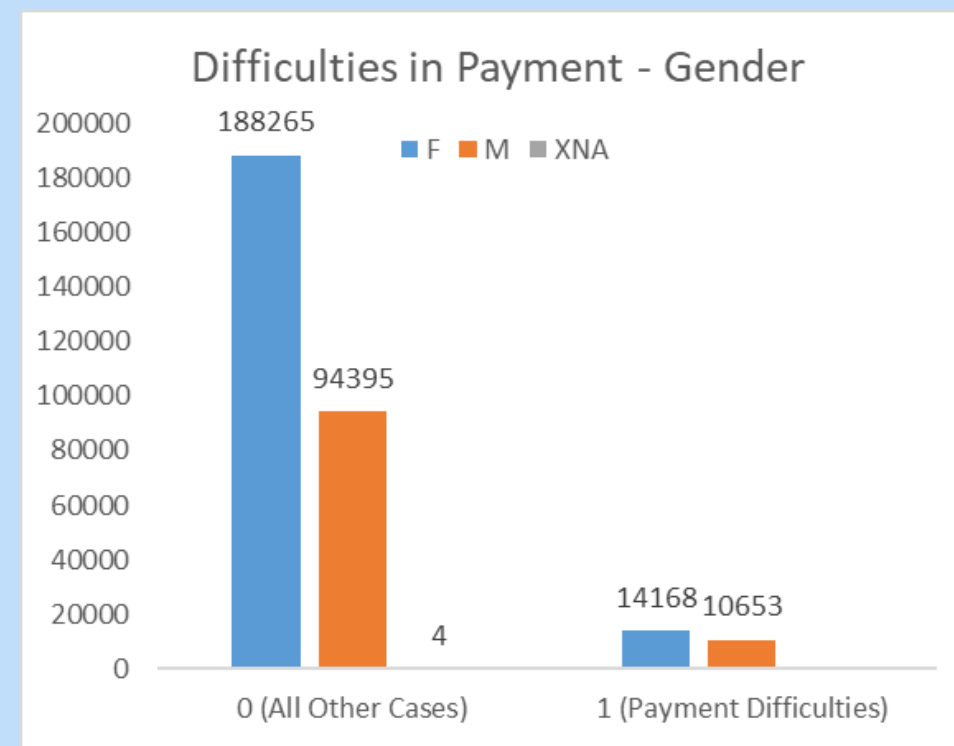
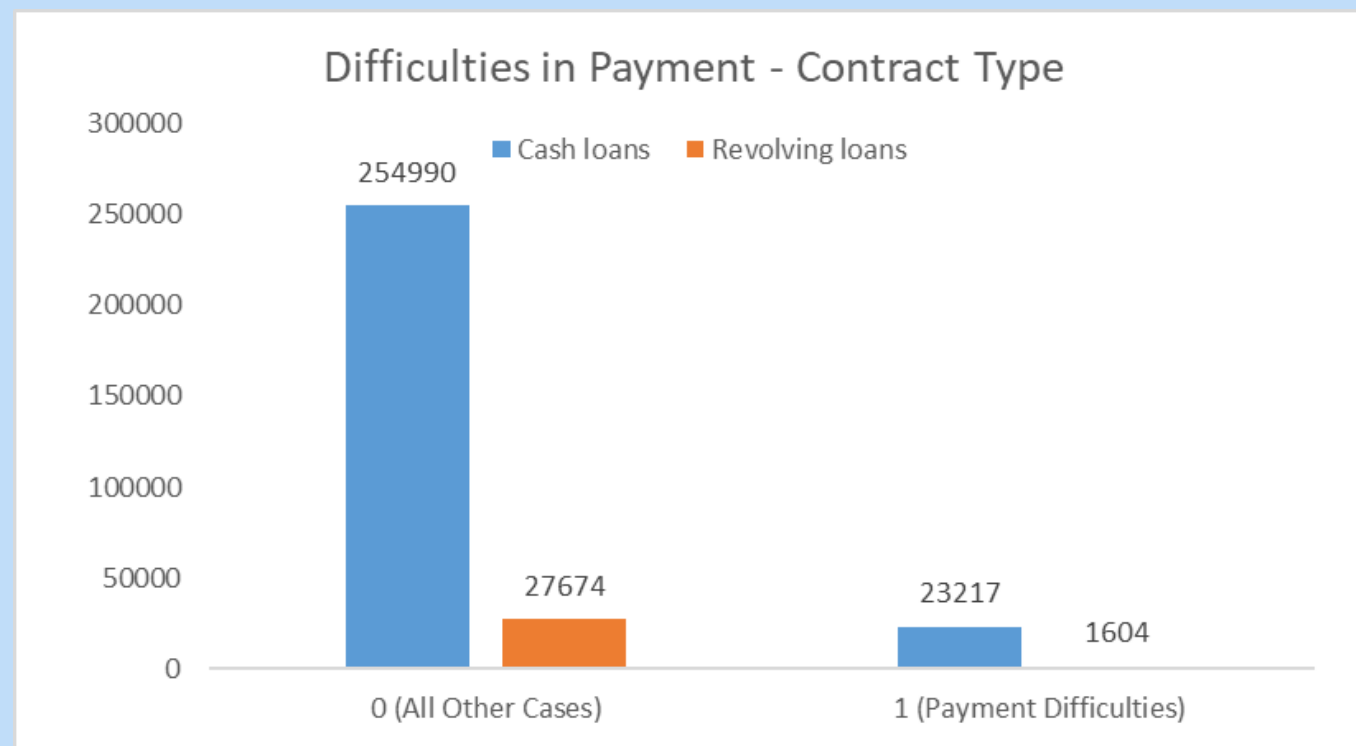


Previous Application

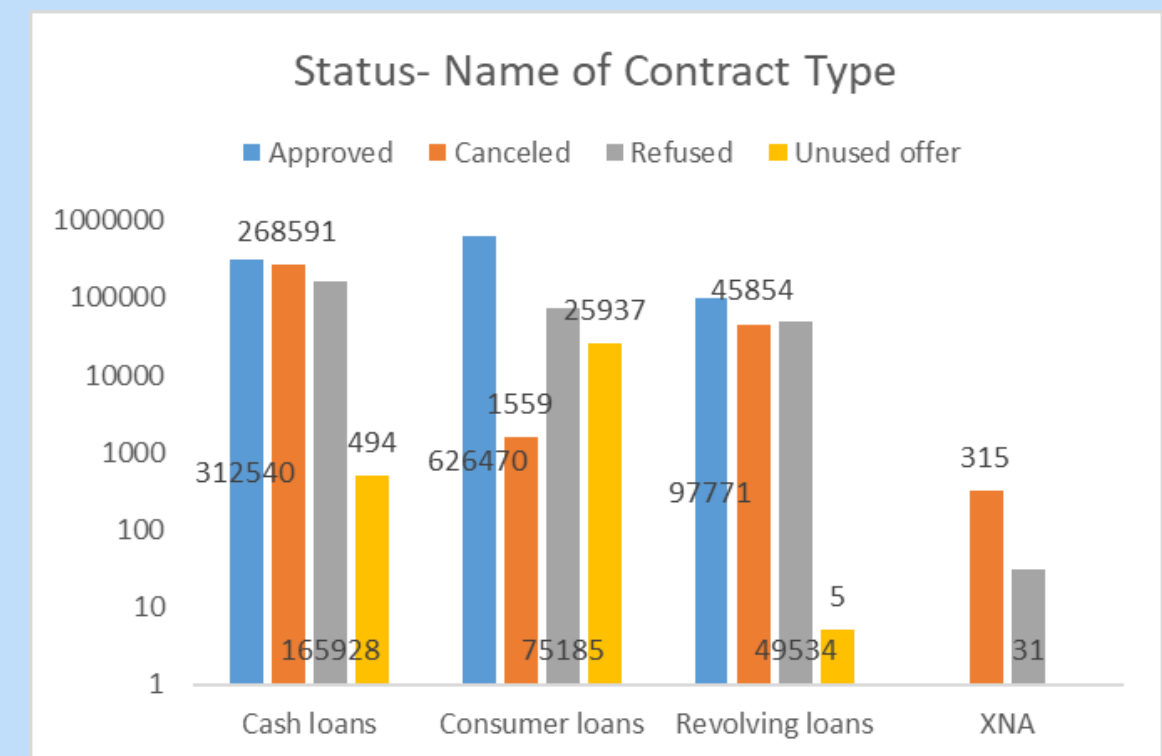
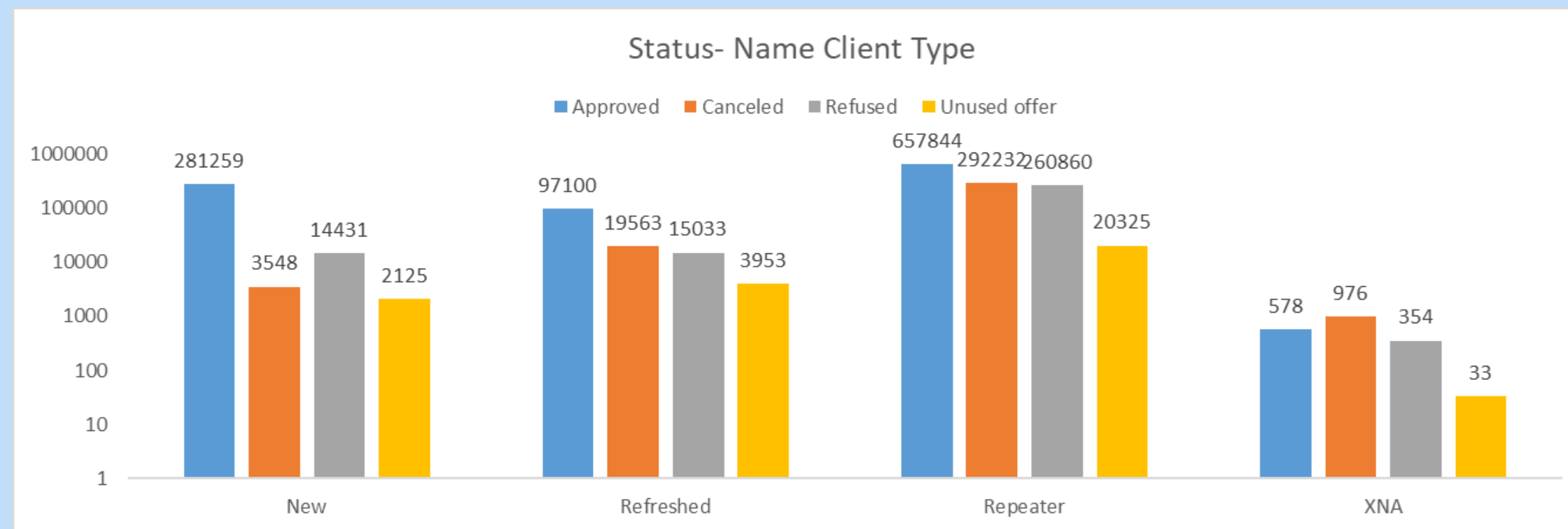


Insights

Application Data

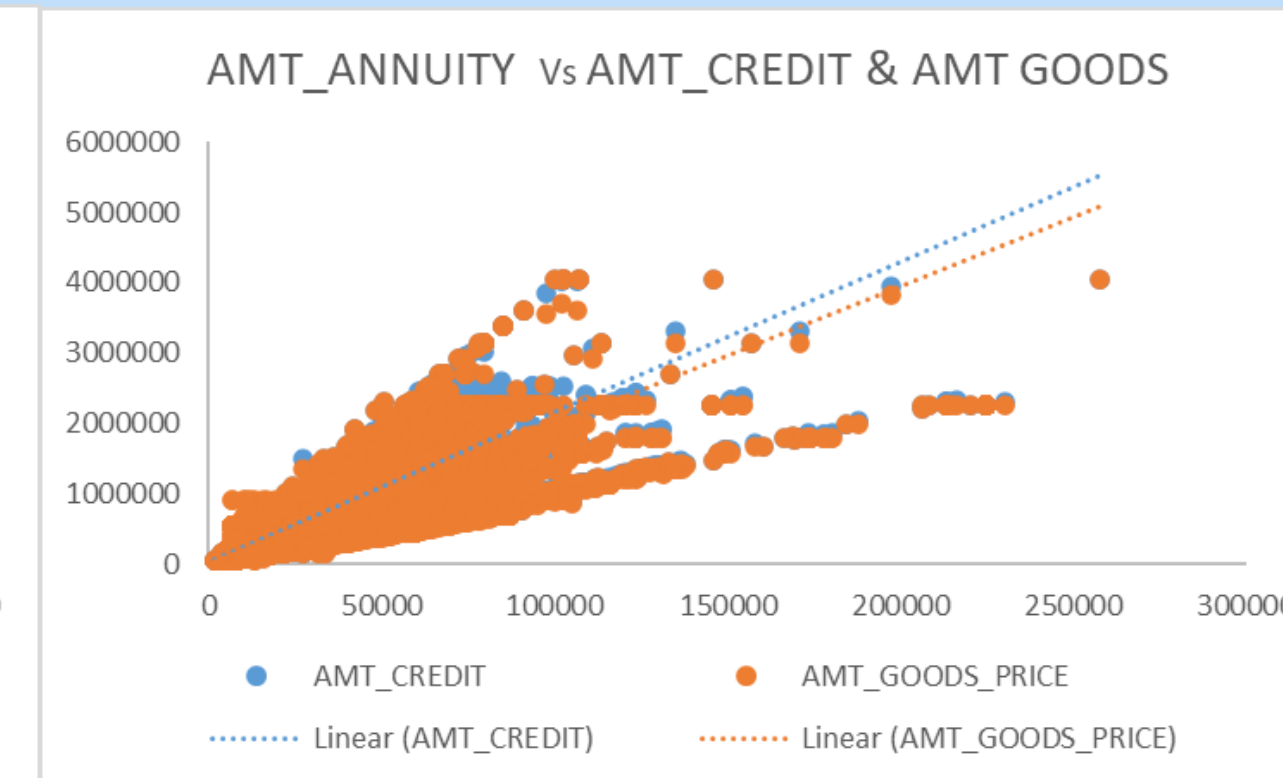
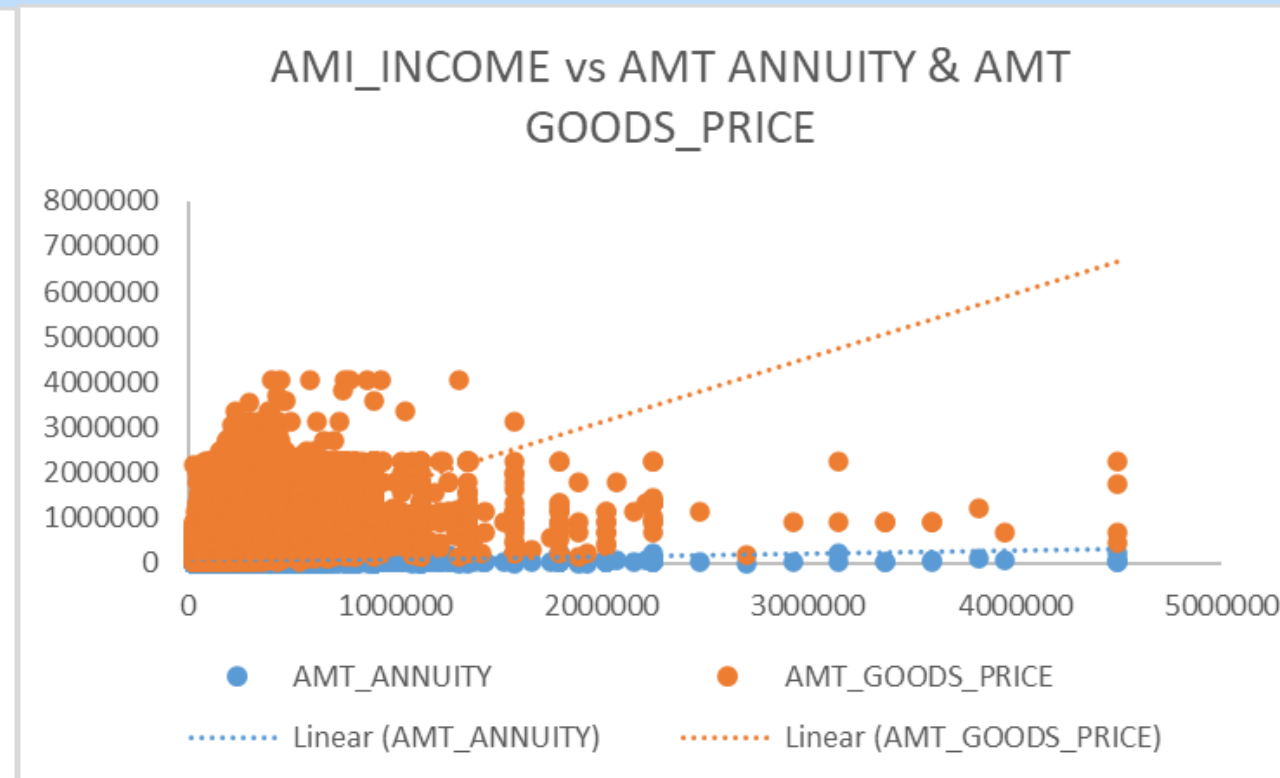
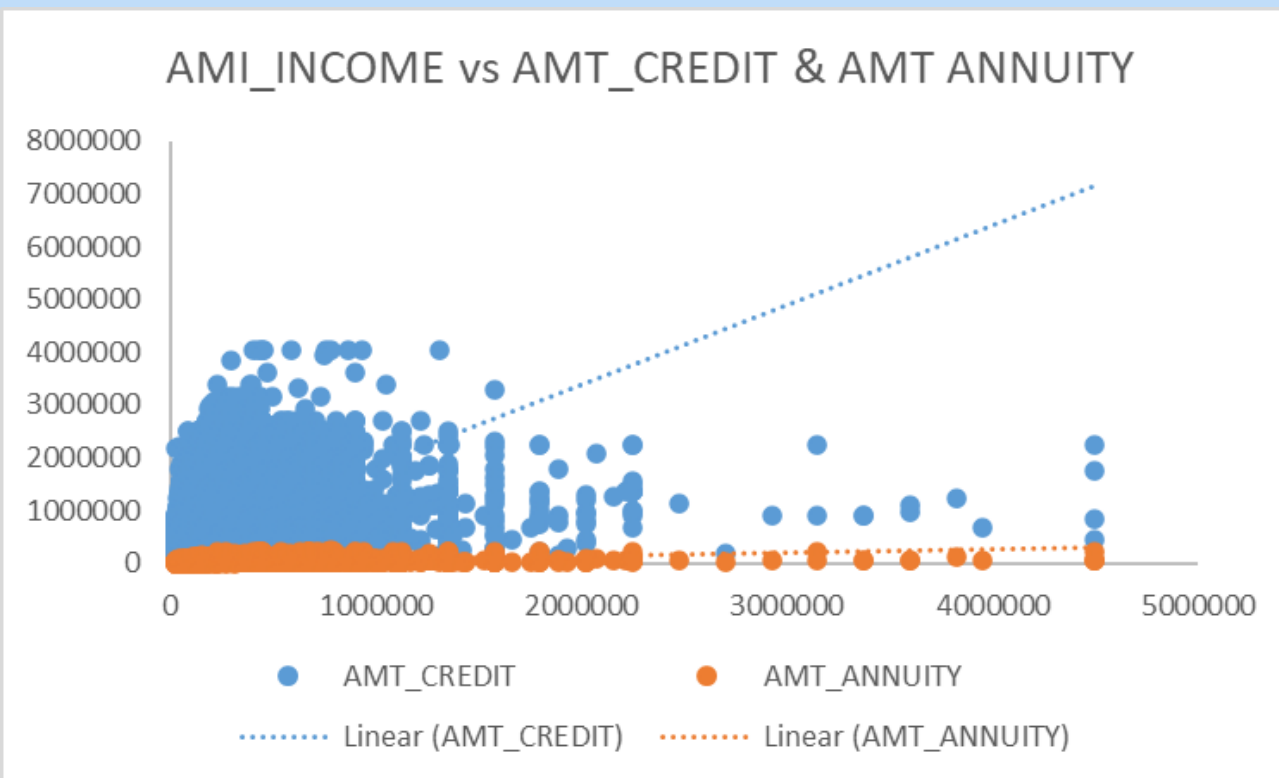


Previous Application

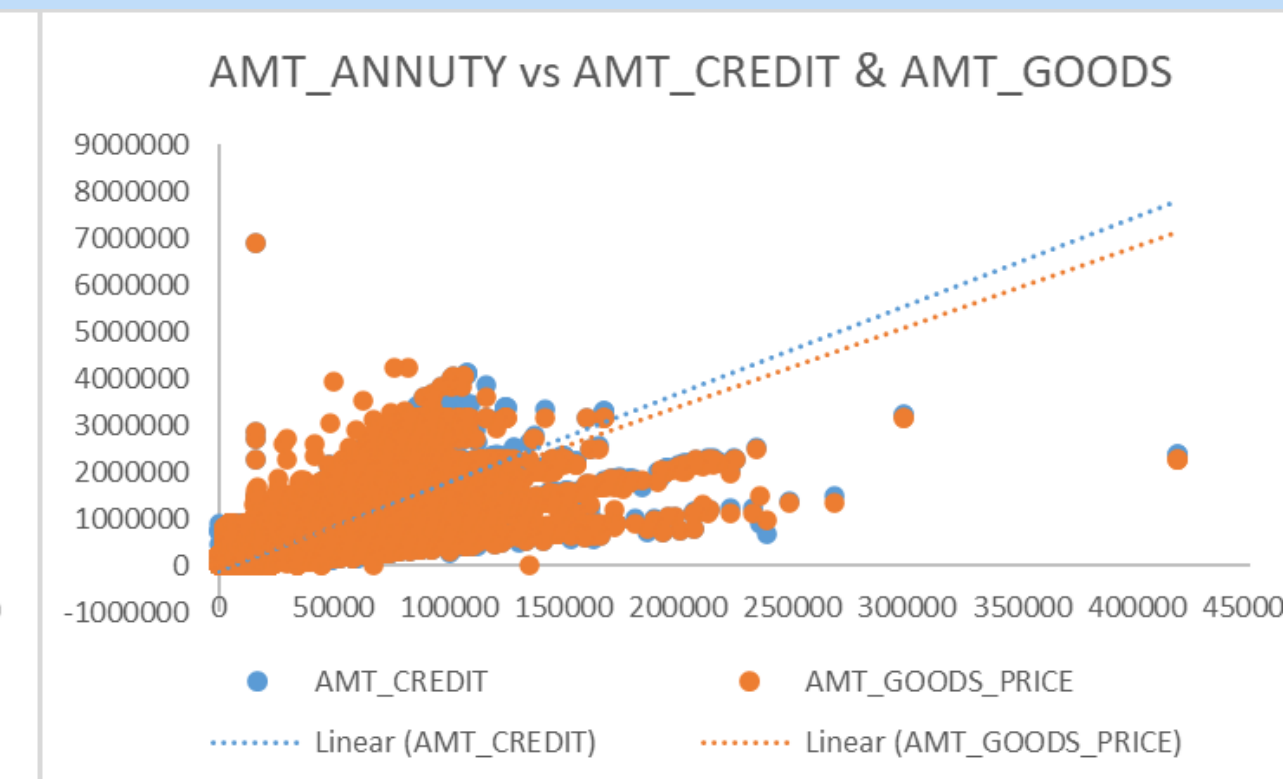
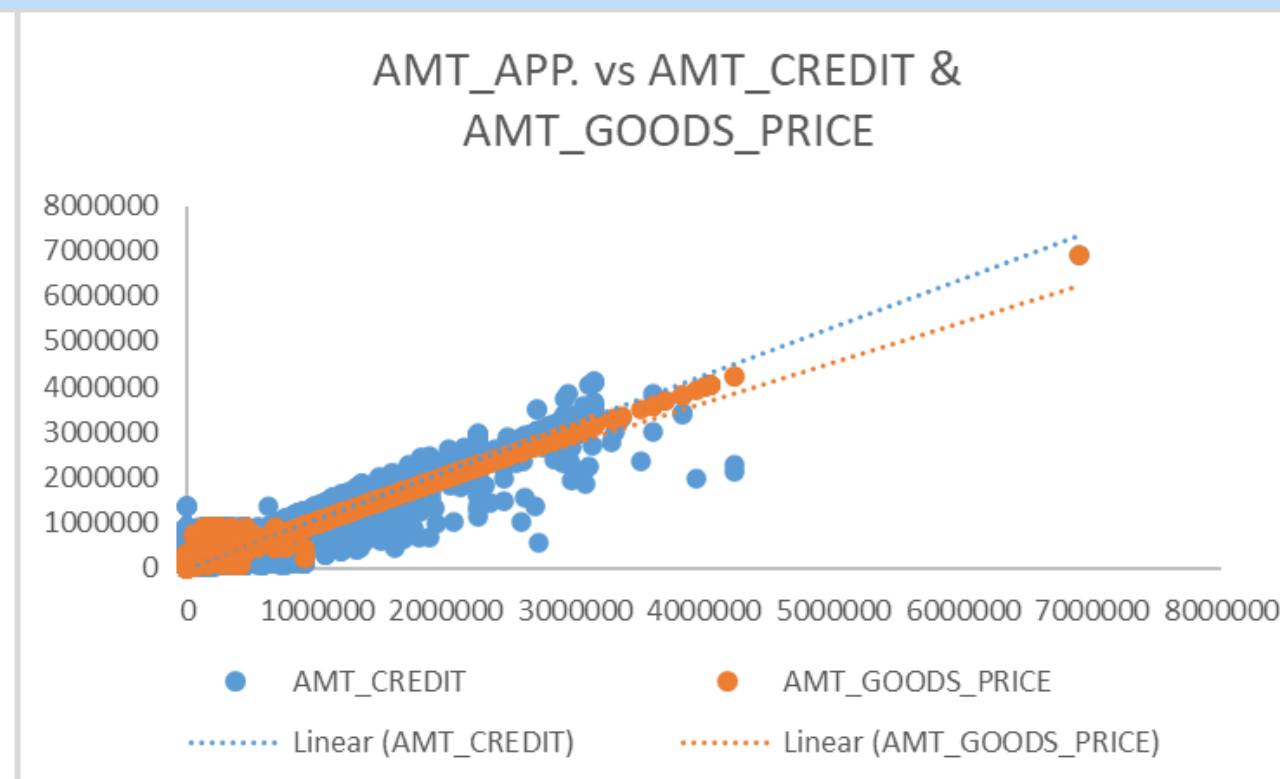
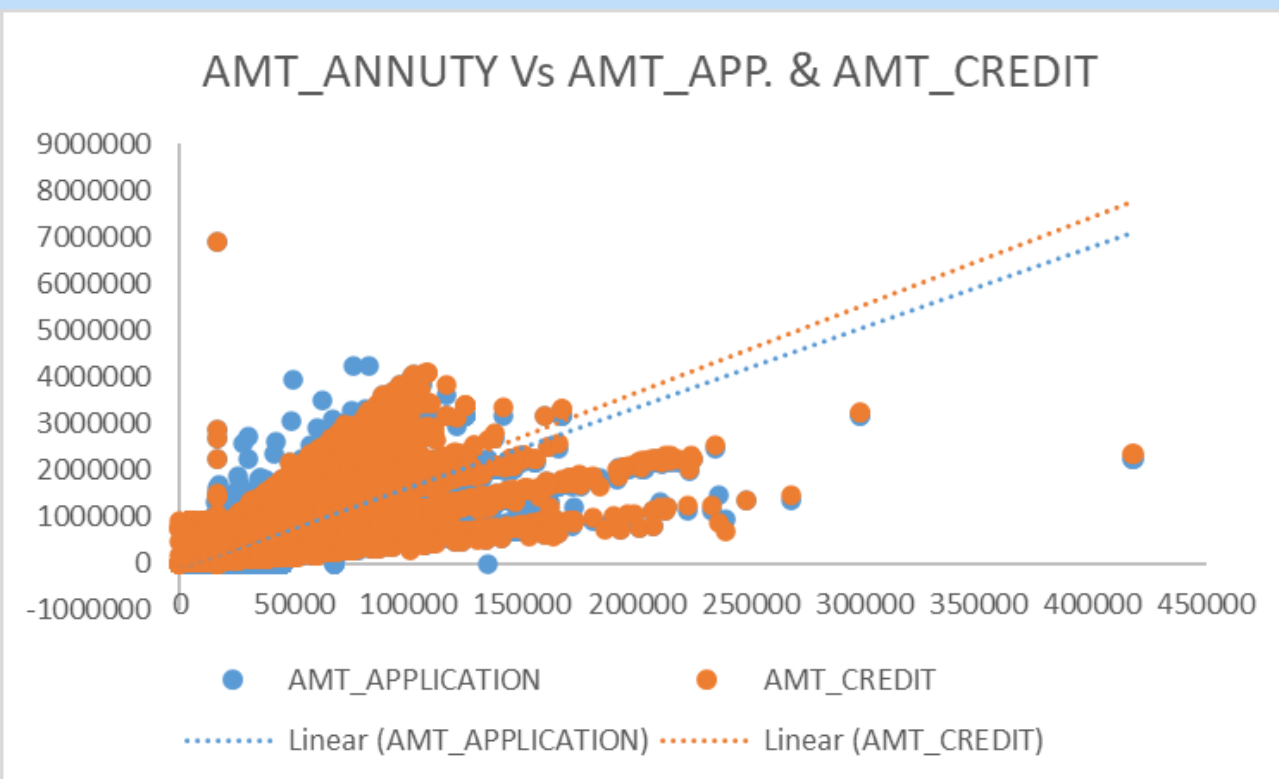


Insights

Application Data



Previous Application



Insights

Correlation

- Correlation is a statistical measure that quantifies the relationship between two variables.
- It provides a numerical value, called the correlation coefficient, which indicates the strength and direction of the relationship.
- The correlation coefficient ranges from -1 to 1, where 1 represents a perfect positive correlation, -1 represents a perfect negative correlation, and 0 indicates no correlation.
- Correlation analysis is beneficial in identifying patterns, trends, and associations between variables.
- Correlation does not imply causation, meaning that a strong correlation between two variables does not necessarily indicate a cause-and-effect relationship.

	AMT_INCOME _TOTAL	AMT_CRED IT	AMT_ANN UITY	AMT_GOODS _PRICE	DAYS_BIRT H	DAYS_EMPL OYED	DAYS_REGIS TRATION	DAYS_ID_P UBLISH	EXT_SOURC E_2	EXT_SOURCE _3
AMT_INCOME_TOTAL	1	0.3737784	0.4512858	0.380578016	0.0620884	-0.1510885	0.06834703	0.0229584	0.15240144	-0.0618298
AMT_CREDIT	0.373778428	1	0.7701792	0.986587779	-0.0554194	-0.0668474	0.00963079	-0.0065718	0.13113346	0.039203242
AMT_ANNUITY	0.451285761	0.7701792	1	0.774707228	0.00945502	-0.1043293	0.03853192	0.0112934	0.12569408	0.027276659
AMT_GOODS_PRICE	0.380578016	0.9865878	0.7747072	1	-0.0533985	-0.0648364	0.01157149	-0.0092567	0.13921034	0.042955846
DAYS_BIRTH	0.062088397	-0.055419	0.009455	-0.05339852	1	-0.6158666	0.33190564	0.2727149	-0.0918793	-0.18126148
DAYS_EMPLOYED	-0.151088471	-0.066847	-0.104329	-0.06483635	-0.6158666	1	-0.21022726	-0.2723883	-0.0207403	0.101504456
DAYS_REGISTRATION	0.068347027	0.0096308	0.0385319	0.011571494	0.33190564	-0.2102273	1	0.1019044	-0.0598375	-0.09624586
DAYS_ID_PUBLISH	0.022958446	-0.006572	0.0112934	-0.00925673	0.27271489	-0.2723883	0.10190441	1	-0.05089	-0.11626699
EXT_SOURCE_2	0.152401438	0.1311335	0.1256941	0.139210345	-0.0918793	-0.0207403	-0.05983751	-0.05089	1	0.096649871
EXT_SOURCE_3	-0.061829796	0.0392032	0.0272767	0.042955846	-0.1812615	0.10150446	-0.09624586	-0.116267	0.09664987	1

Observations

Univariate Analysis:

- Higher applications received for cash loan instead of revolving loan
- Female applicants are more than Male Applicants make the data imbalance and biased results.
- Out of given applicants majority of applicants don't own a car but owns a realty
- Majority of the time its observed that applicant is "Unaccompanied"
- It seems that applicant under "Maternity Leave, Businessman, Student and unemployed are less actively apply for the loan than working followed by commercial associate pensioner and state servant

Segmented Univariate Analysis:

- Consumer loans are most likely to get approved than cash loan and revolving loan.
- Maximum Applicant preferred payment cash through the bank
- Repeater are most likely to get loan approval on there loan application
- POS portfolio has major role in sale of loan with less number of cancelation

Bivariate Analysis:

- Amount of Goods price and amount credit is highly corelated with each other
- Days employed and Day from birth are negatively corelated with each other

Result

- Microsoft Excel has very powerful tool know as Power Query and Power Pivot which helps with analysis of Large data set
- Understanding the structures of data is important before doing any type of data cleaning process
- When it comes to a Data Cleaning need to keep focus on how results going to help business to improve not need to work on unnecessary data if its not giving any conclusion to business
- This project help to understand how to perform Exploratory Data Analysis and its important. also what are the key things to keep in mind while doing EDA.
- Data was related to banks loan business got clear idea how does the bank works on the data in depth before proving approval to the loan applicant
- Overall Data Analysis is a key of every business the clean data we have the clear insights we can get to improve business.

Attachment:- Link for project file

Thank You..