Bank Loan Analysis

Agenda for application data.csv

- Data Handling
- Data Imbalance
- Segmented, Bivariate, Multivariate Analysis
- Correlation Analysis
- Result and Insights

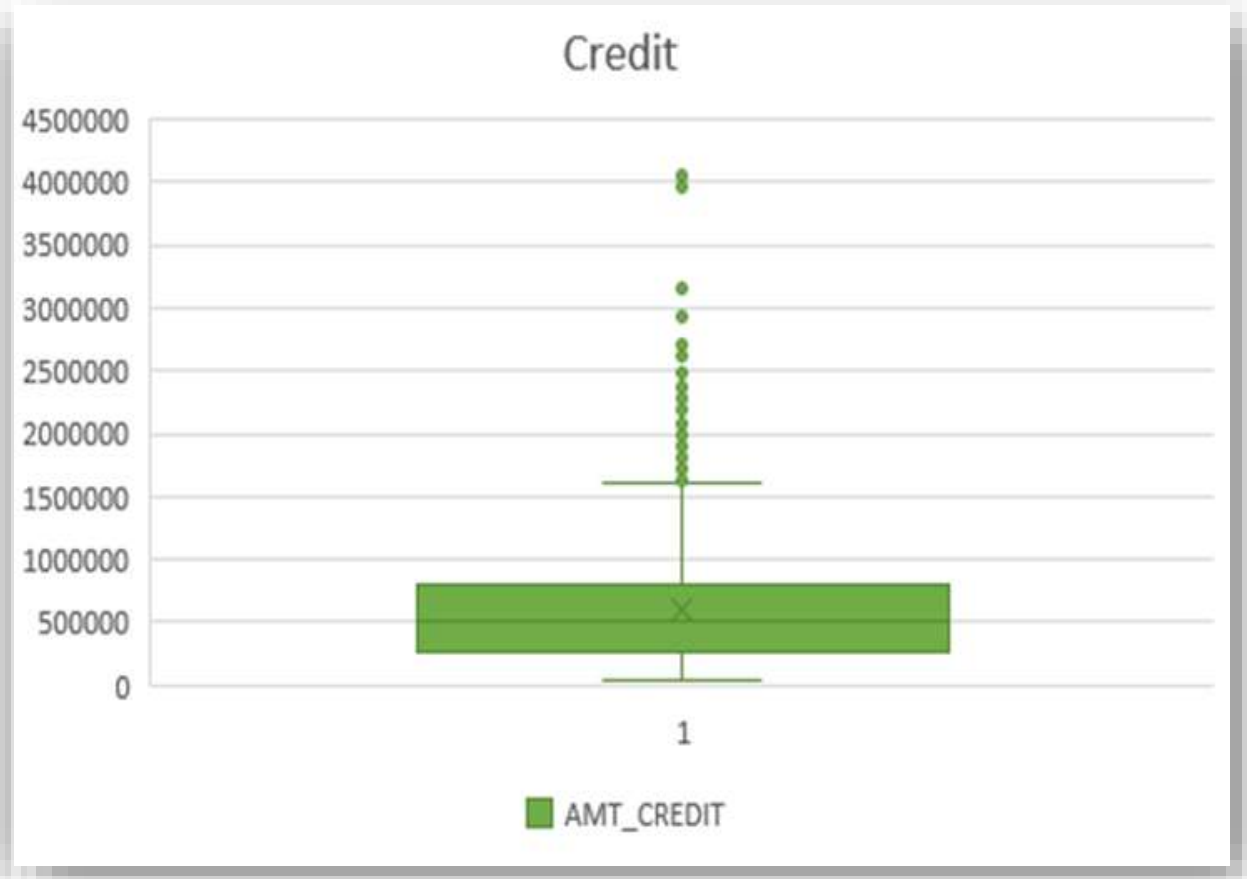
https://docs.google.com/spreadsheets/d/179zmVAFWvKm7oeU6ba8sZ9Eylsq9SVu4/edit?usp=sharing&ouid=104301423844572907298&rtpof=true&sd=true

Data Handling – Identifying Missing Data

- Amt_annuity, days_cell_change and cnt_family_members have one cell empty hence the blank rows are deleted
- For own_car_age null values percentage is greater than 30 percent hence deleted
- Amt_goods_price has 152 empty cells, hence median=45000 imputed in blank cells(=MEDIAN(K2:K49997))
- Ext_source_2 0.25% empty cells, hence average=0.513817582 imputed in blank cells(=COUNTBLANK(AQ2:AQ49997)/49996*100)
- Columns AC(31%), AS:CM(livelihood details)(64.2%) have blank percentage >30% hence dropped
- AMT req credit bureau ,OBS def cnt social circle ,name_type_suite are unwanted columns, hence dropped
- Duplicates were removed using Remove Duplicates from Data in table

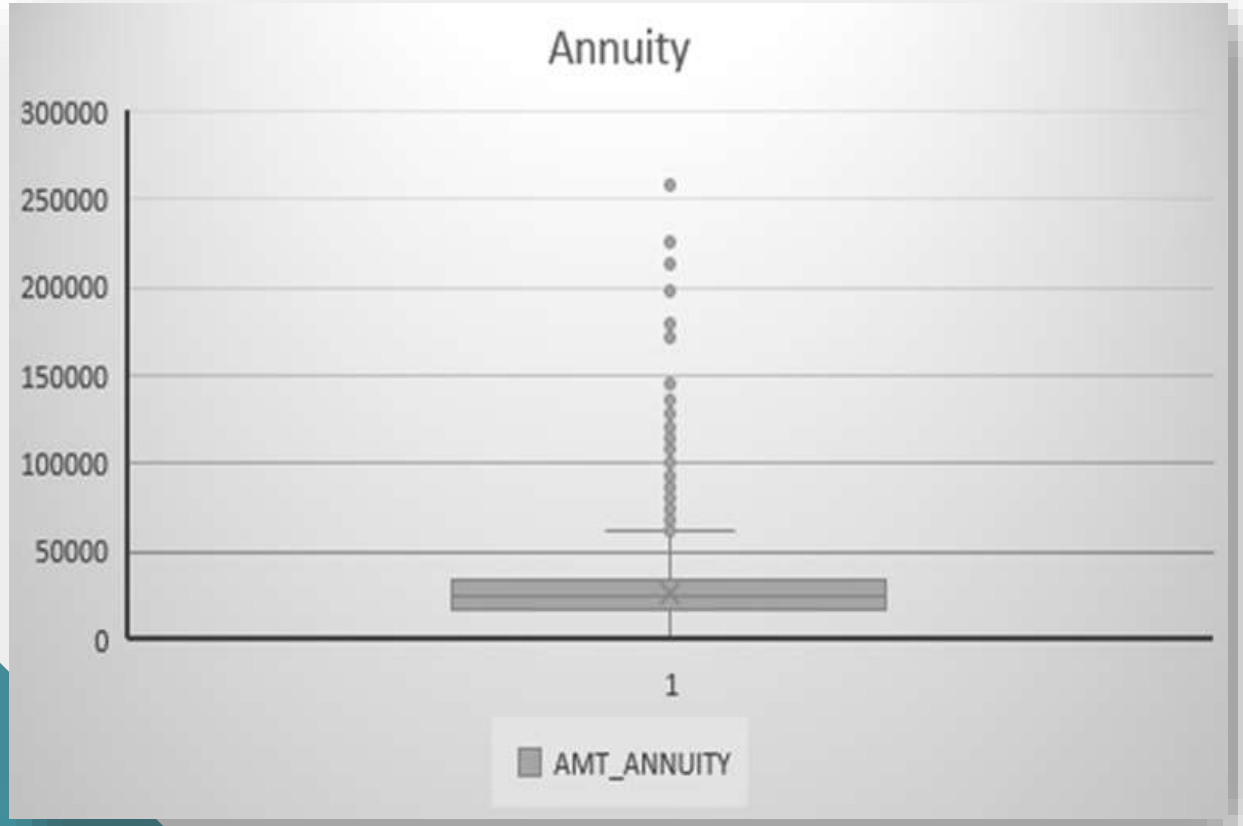
Data Handling – Identifying Outliers

- Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
- Outliers have to be removed before plotting graphs, to get balanced data
- Formula used:
- **=QUARTILE.INC(range,0)- lower quartile**
- **=QUARTILE.INC(range,1)- middle quartile**
- **=QUARTILE.INC(range,2)- middle quartile**
- **=QUARTILE.INC(range,3)- middle quartile**
- **=QUARTILE.INC(range,4)- upper quartile**



The outlier value is in the 4th quartile =QUARTILE.INC(H3:H49998,4) that is 117000000

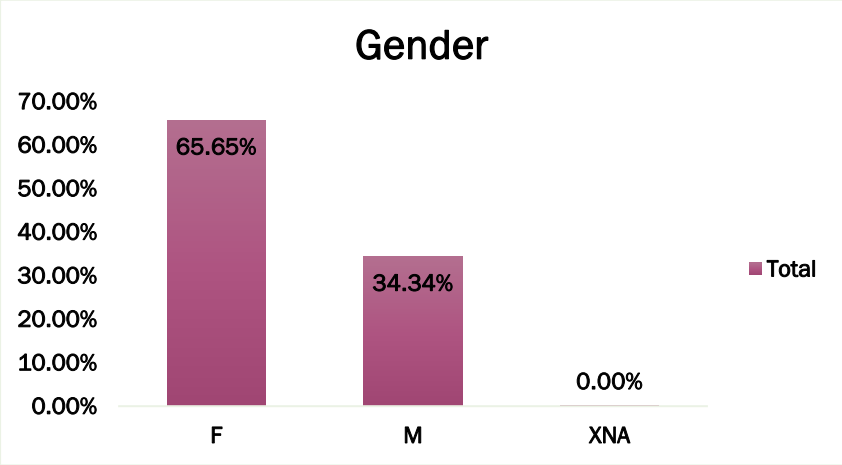
The outliers are between the 3rd to 4th quartile

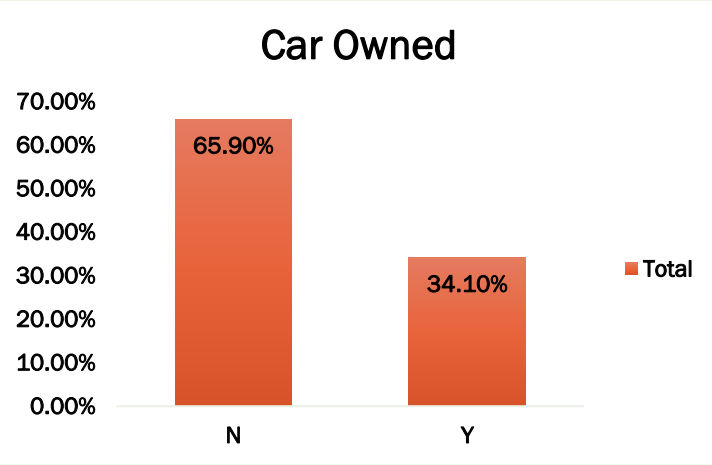


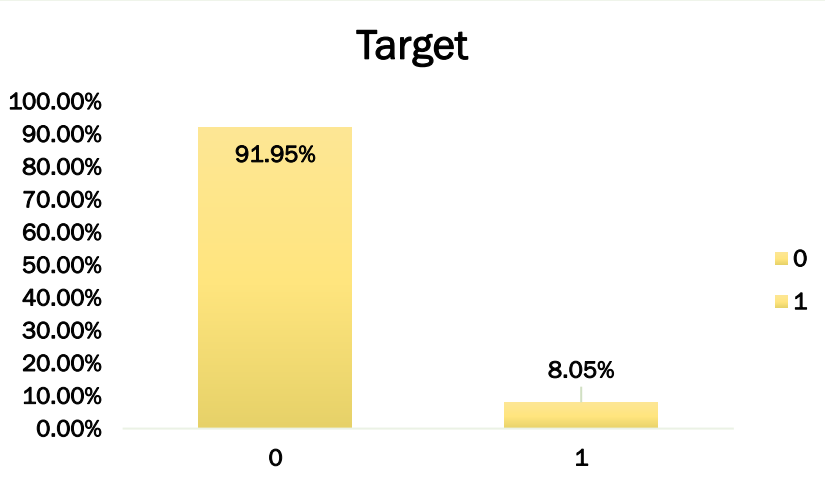
Some steady outlier values are between the 3rd and 4th quartile, while some are between 2 to 3 lakhs in 4th quartile

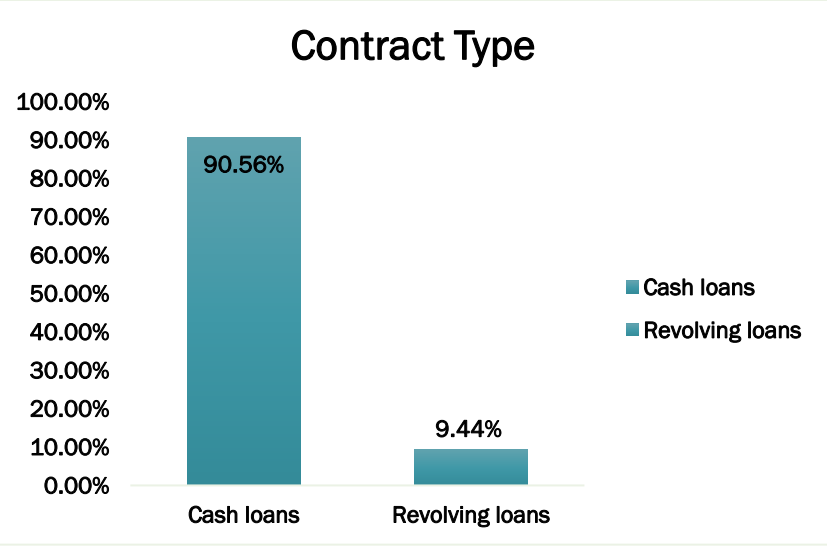
Some steady outliers are between the 3rd and 4th quartile while almost 5 outliers are in 4th quartile

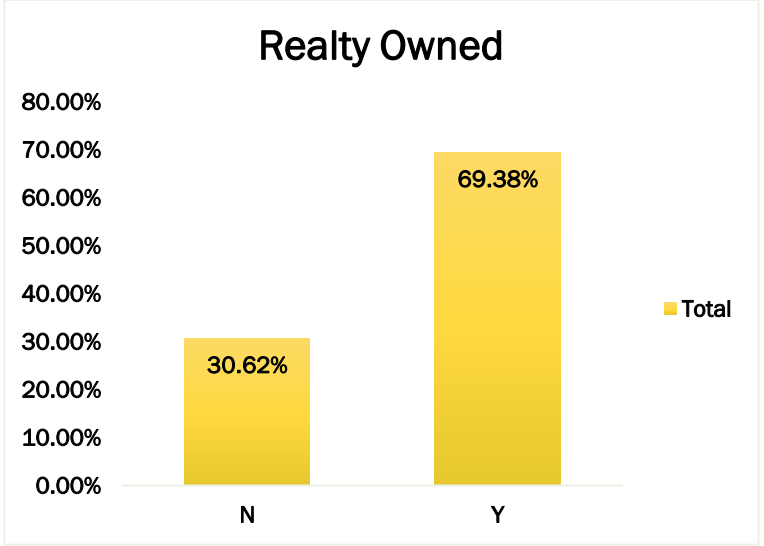
Data Imbalance and Univariate Analysis

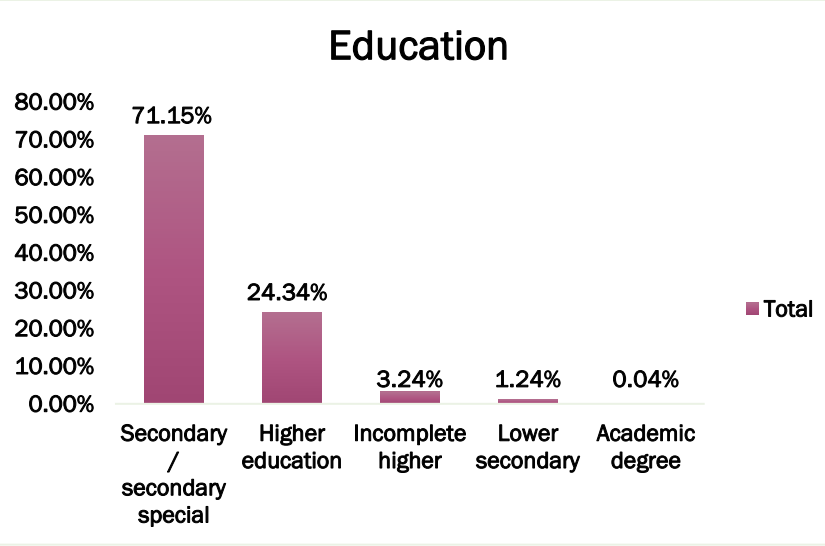








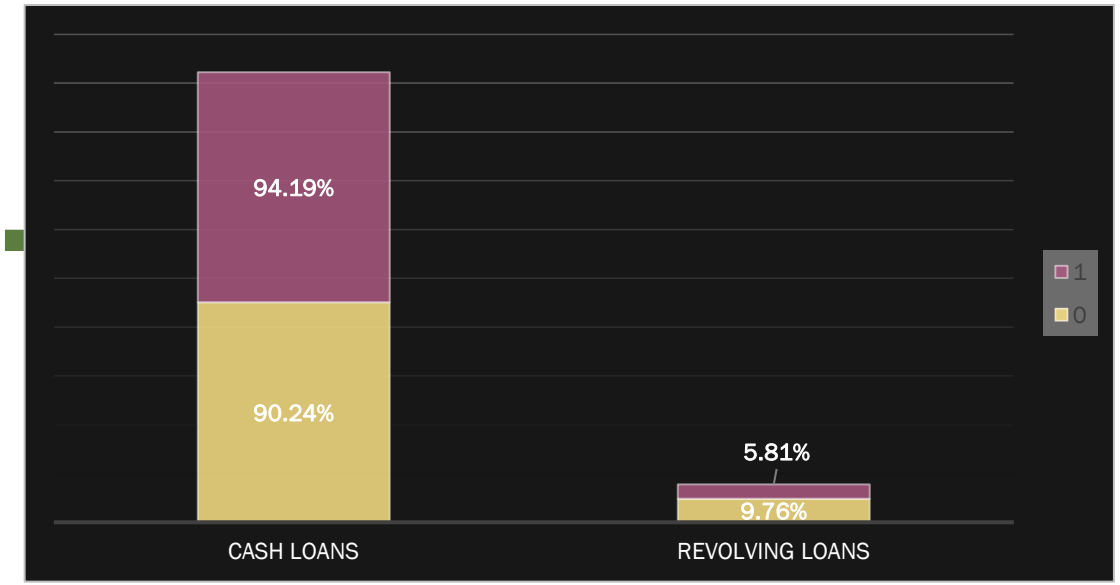


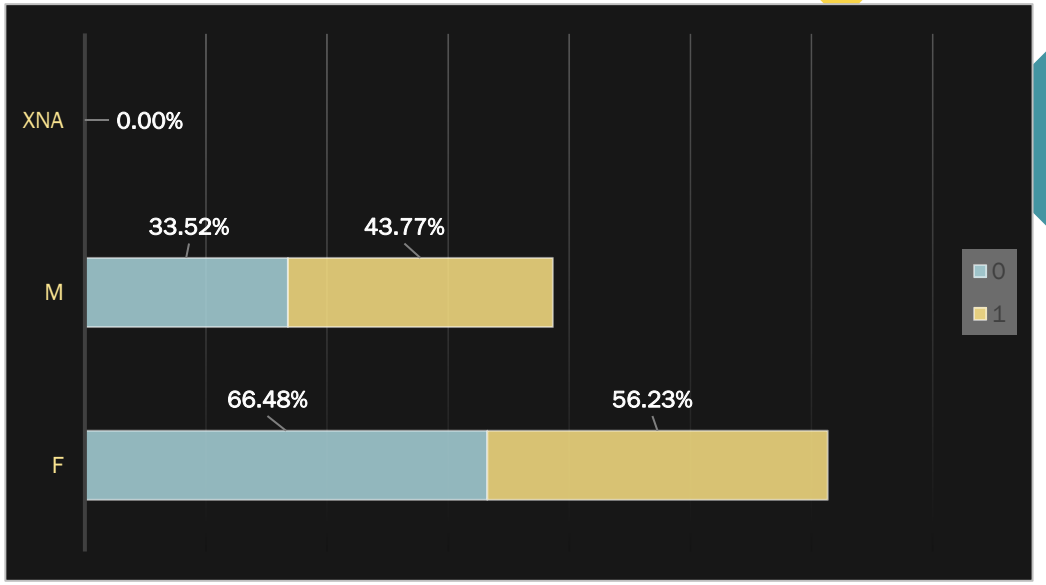


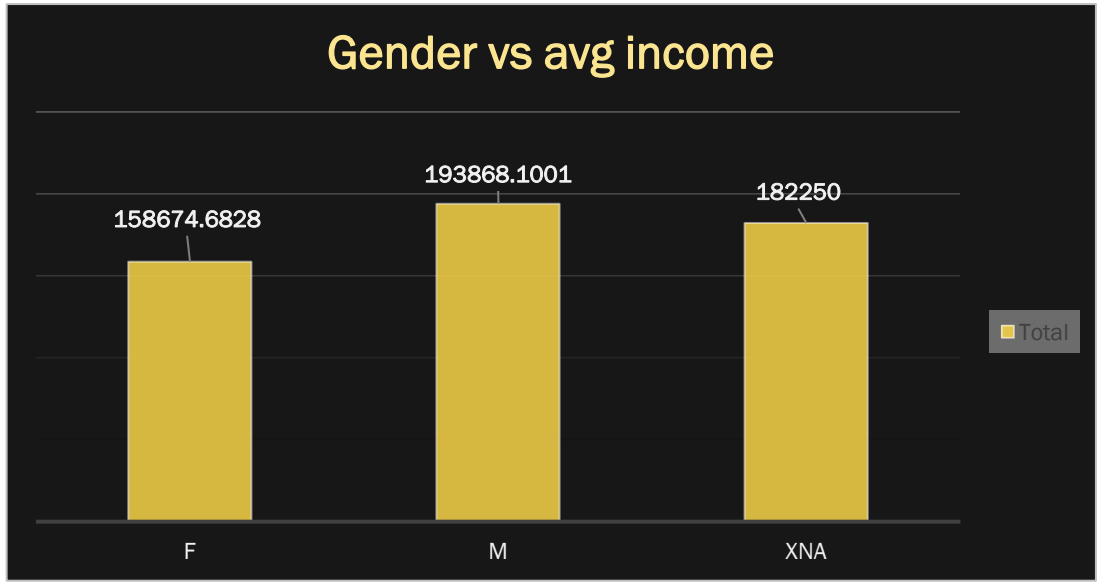
Insights

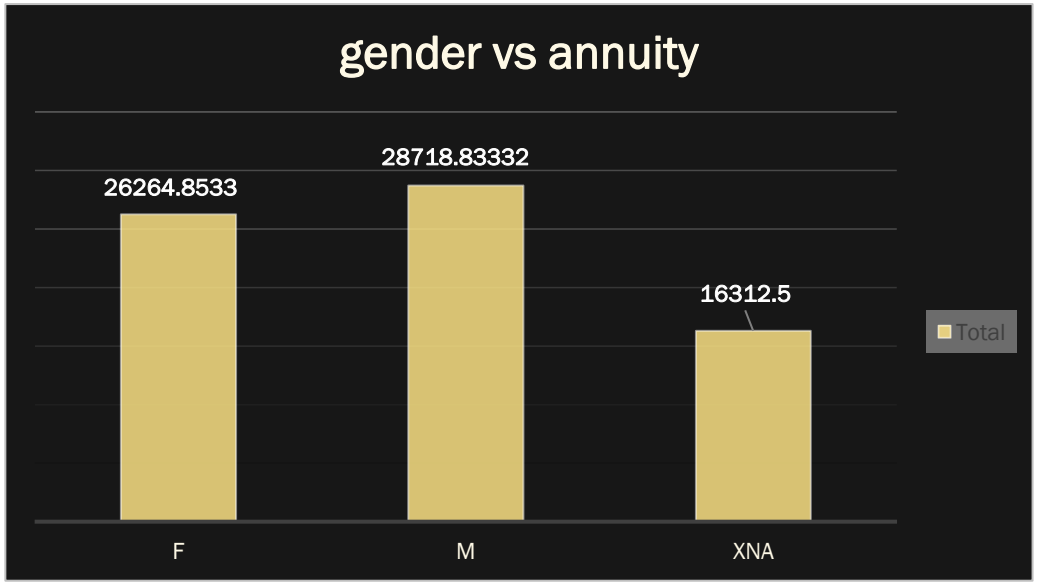
- Majority of population is female
- Majority of population own a car
- Majority of loans have been paid
- Most popular contract type is Cash Loans
- Majority of population own a realty
- Secondary/Secondary special education sector has taken the most number of loans

Segmented Analysis

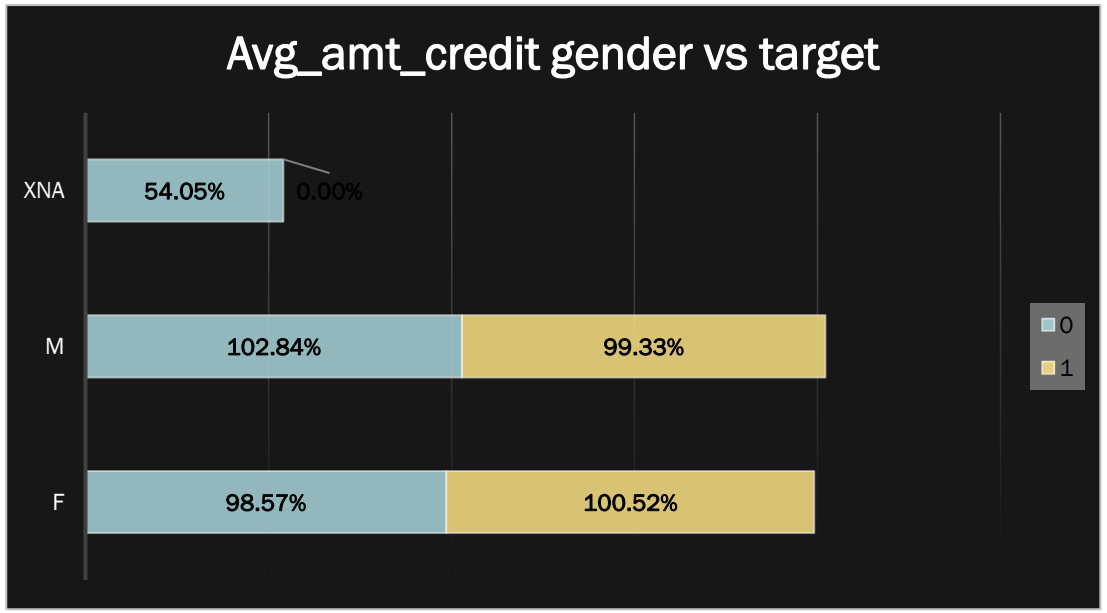


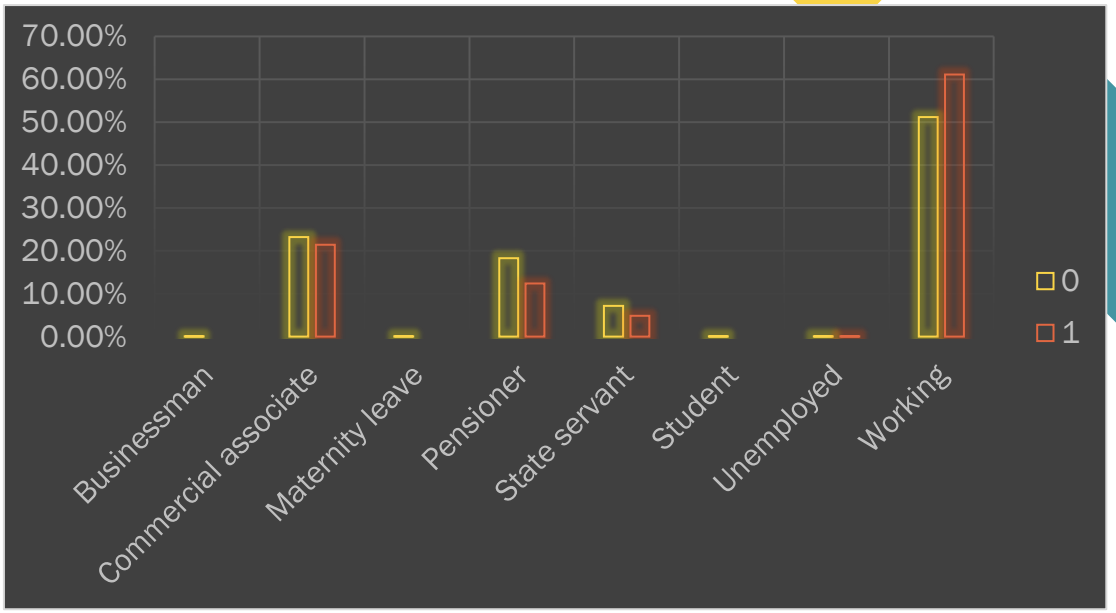


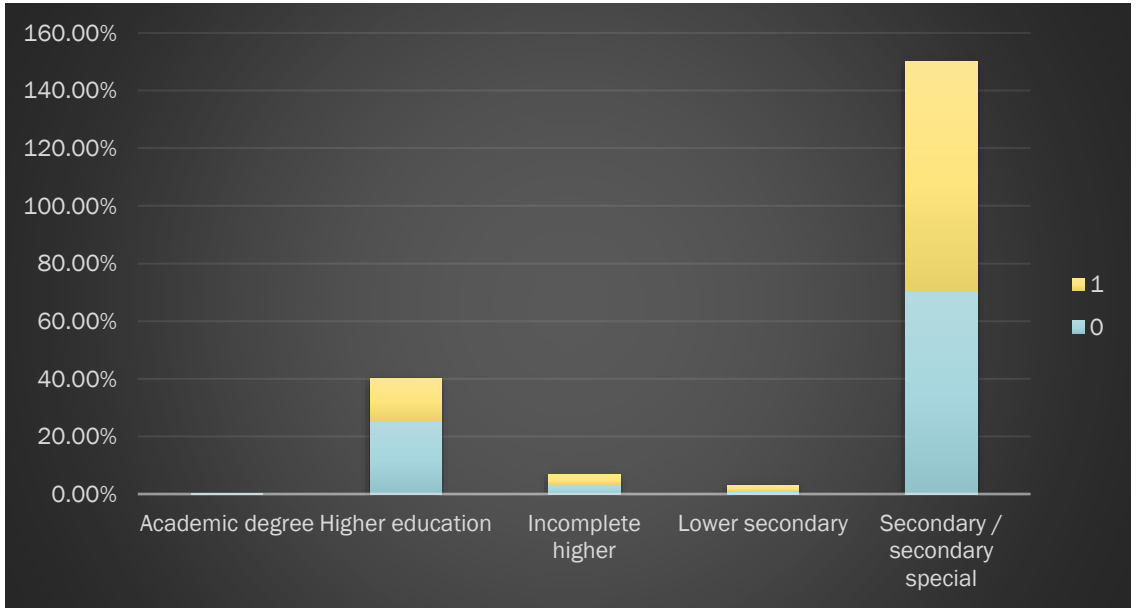


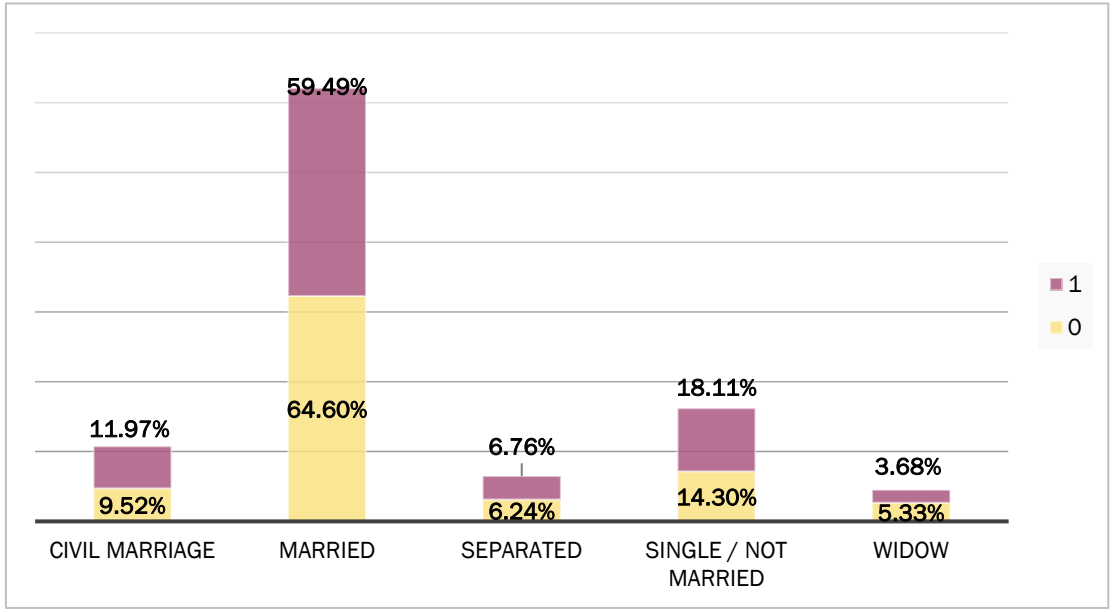


Segmented Analysis

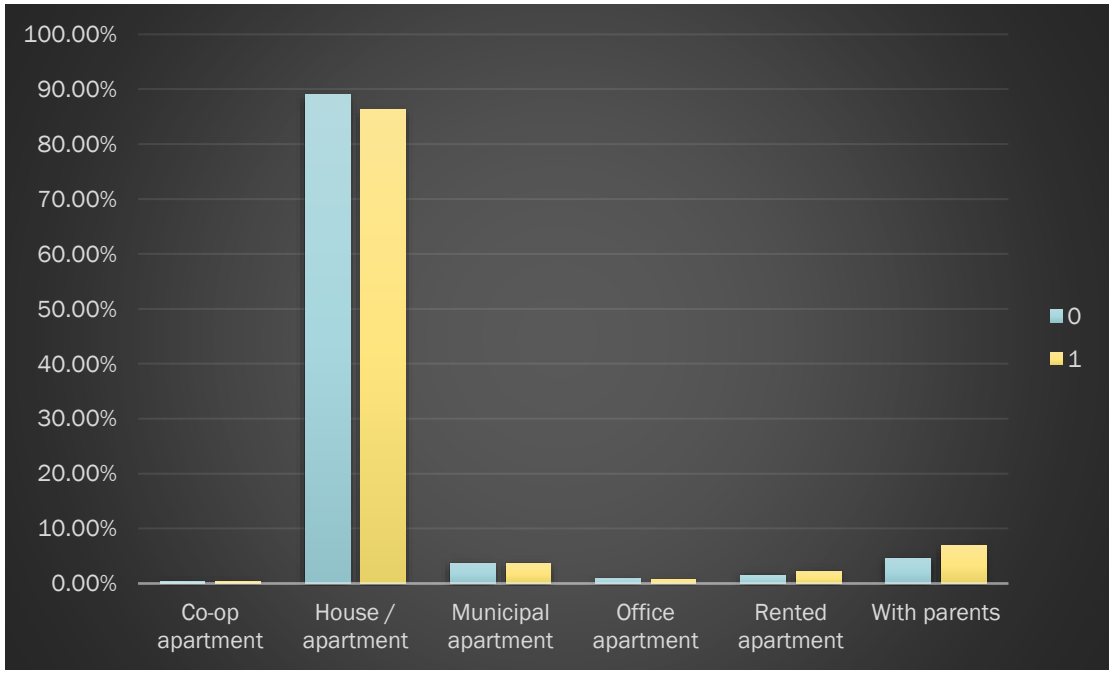


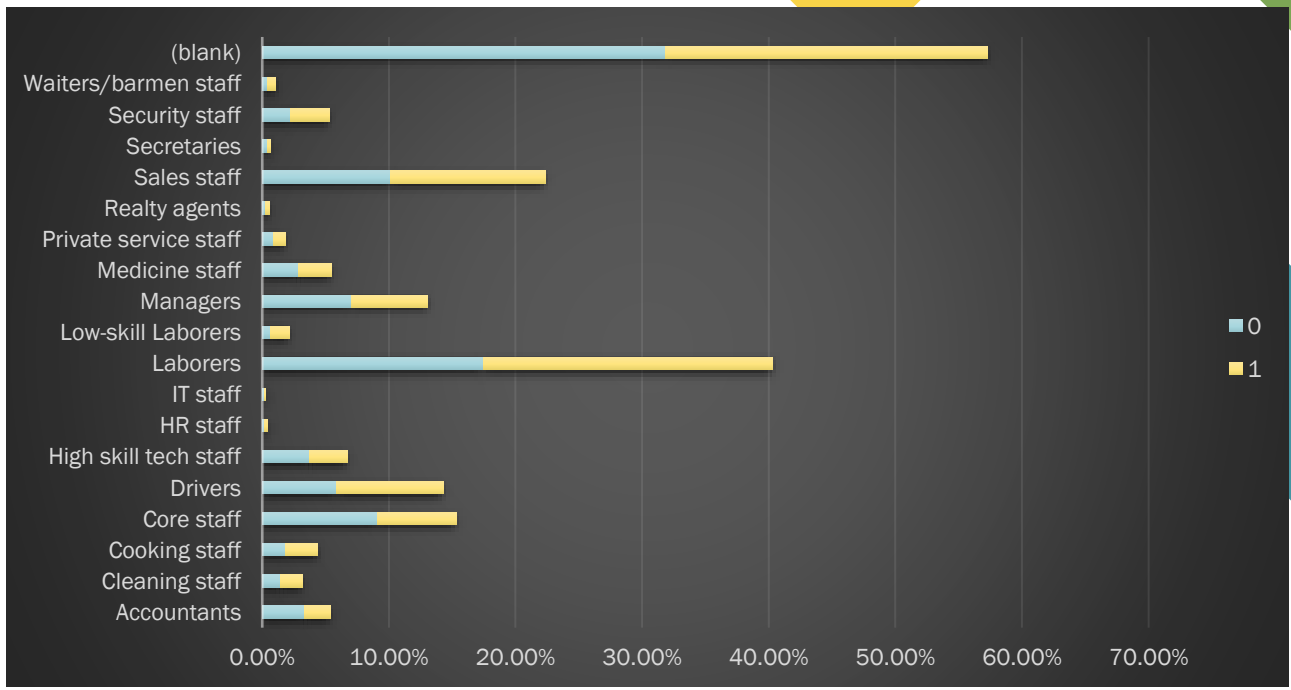


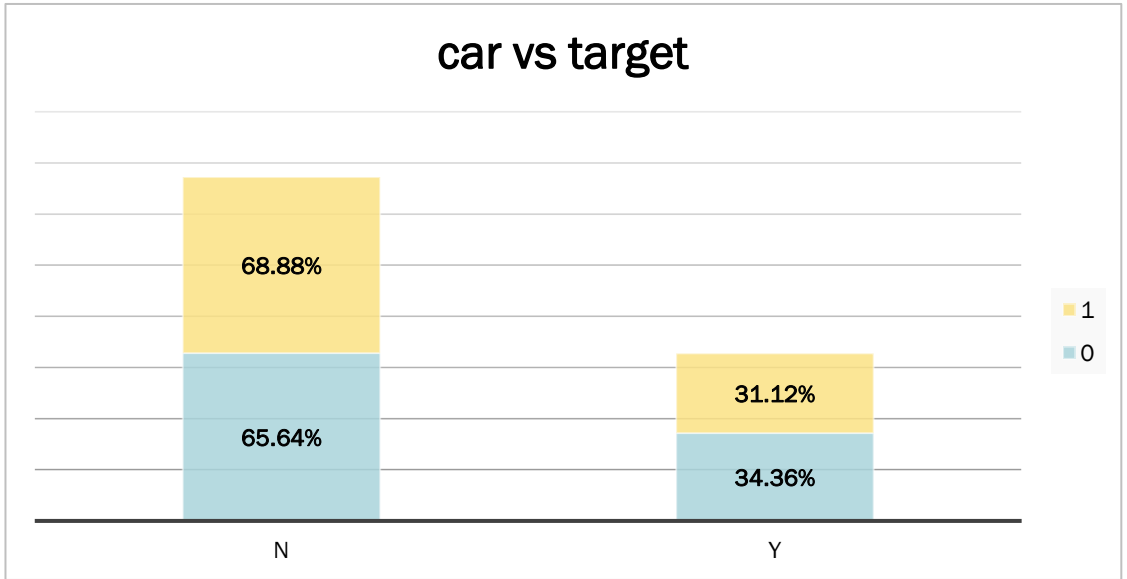


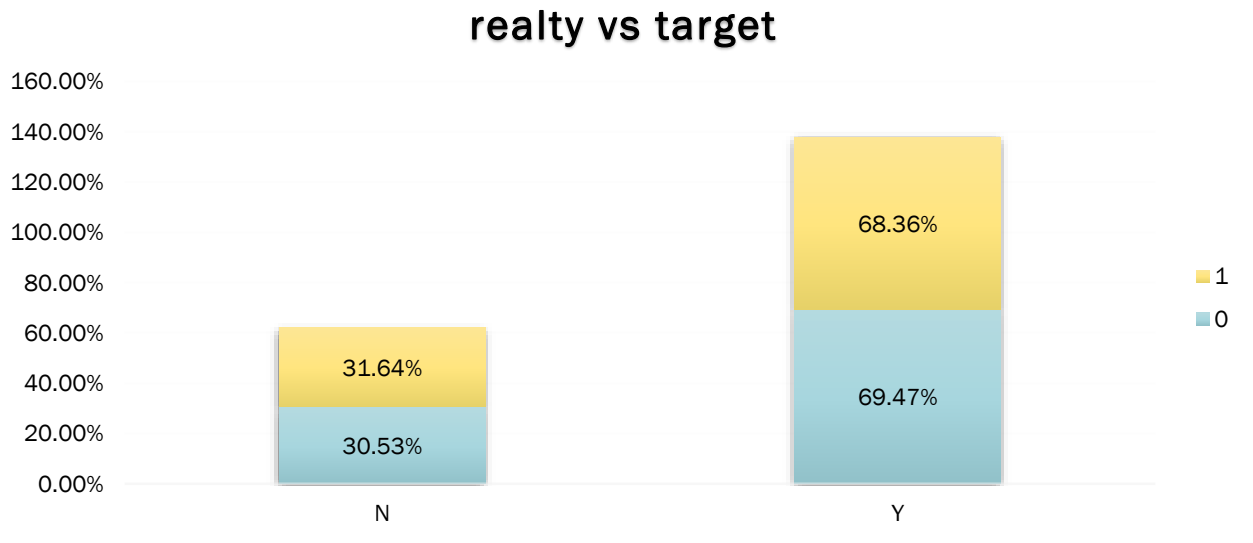


Segmented Analysis

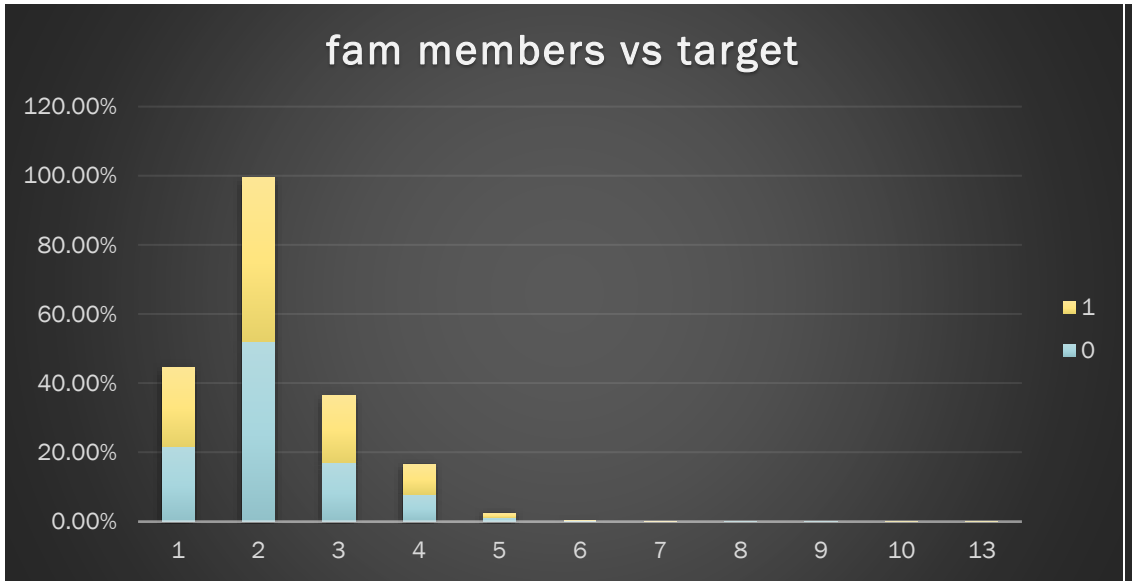


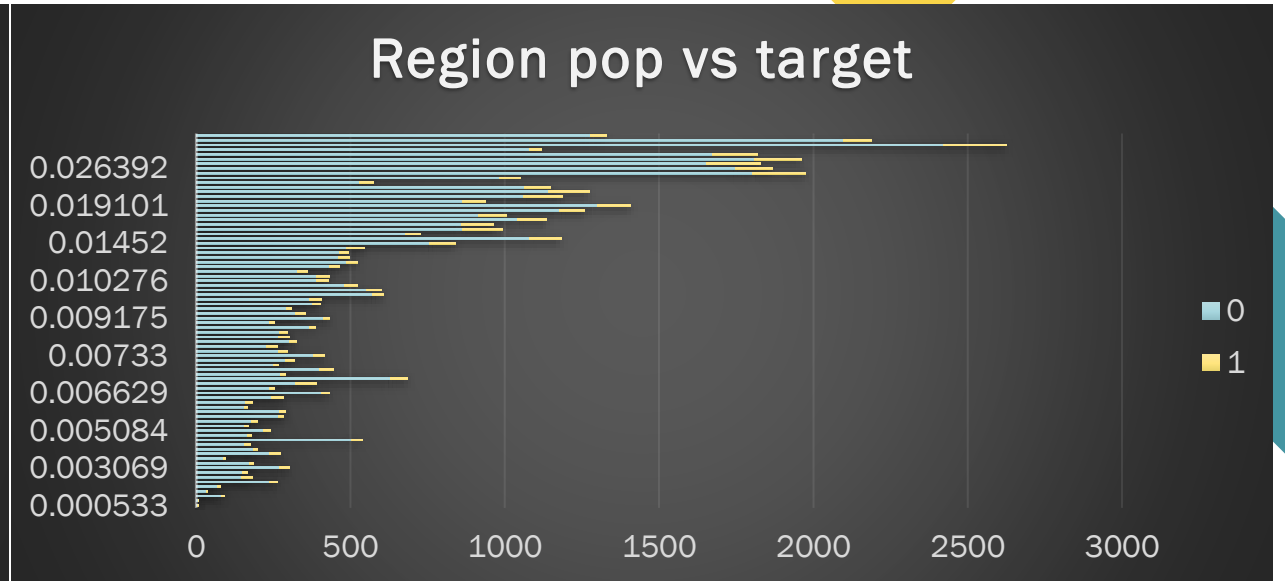






Segmented Analysis

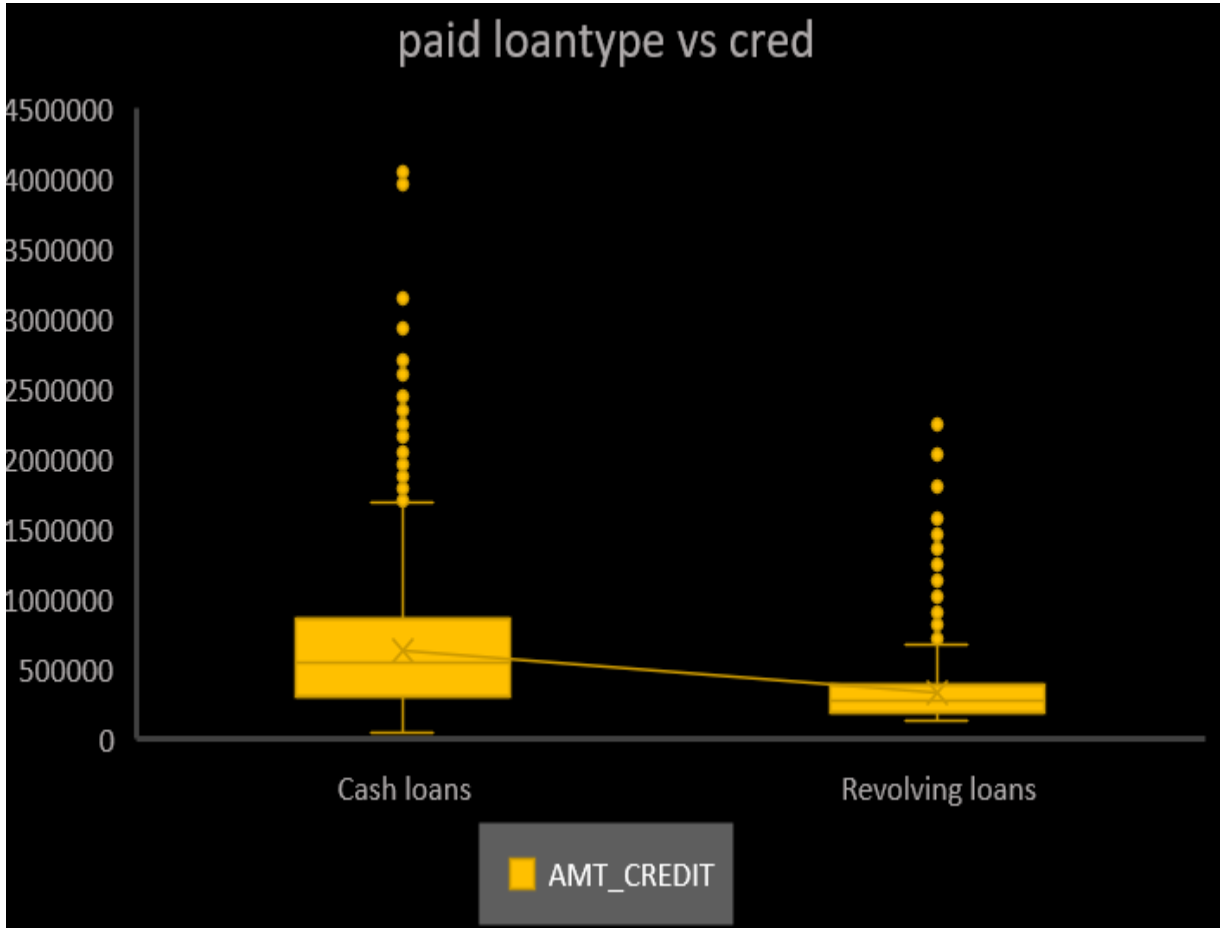


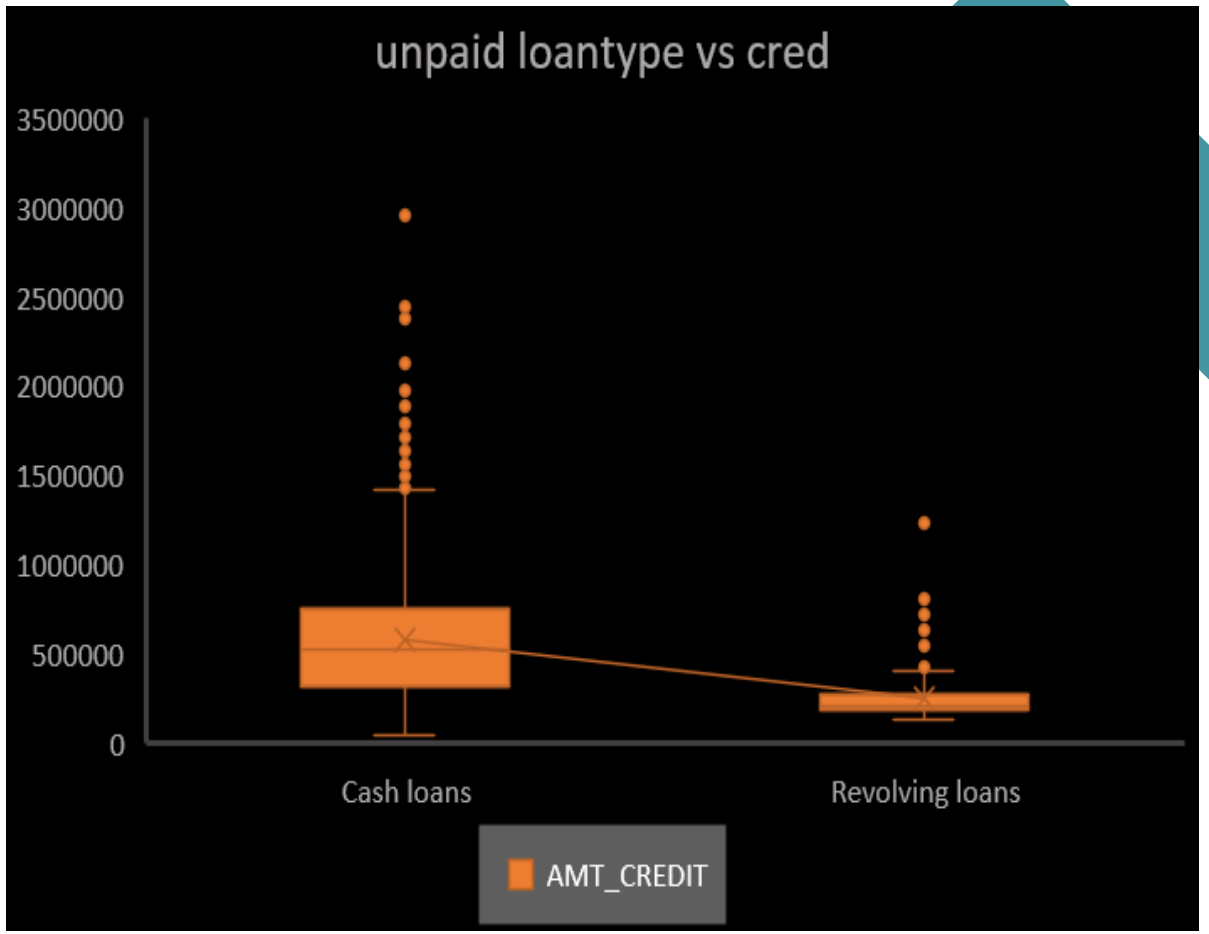


Insights

- Number of unpaid Cash Loans are relatively higher by only 4% to paid loans
- Females pay back most of their loans
- Average annuity and income of males is higher than other genders
- Working class apply for most loans and unpaid loans are relatively higher
- Academic degrees opt for least loans and Married people have most loan applications
- Labourers and people who live in a house/apartment have most loan applications(mostly paid)
- People who own a realty, not own a car have paid majority of their loans
- A family of 2 are able to pay majority of their loans
- Business Entity and University workers are able to pay back majority of their loans

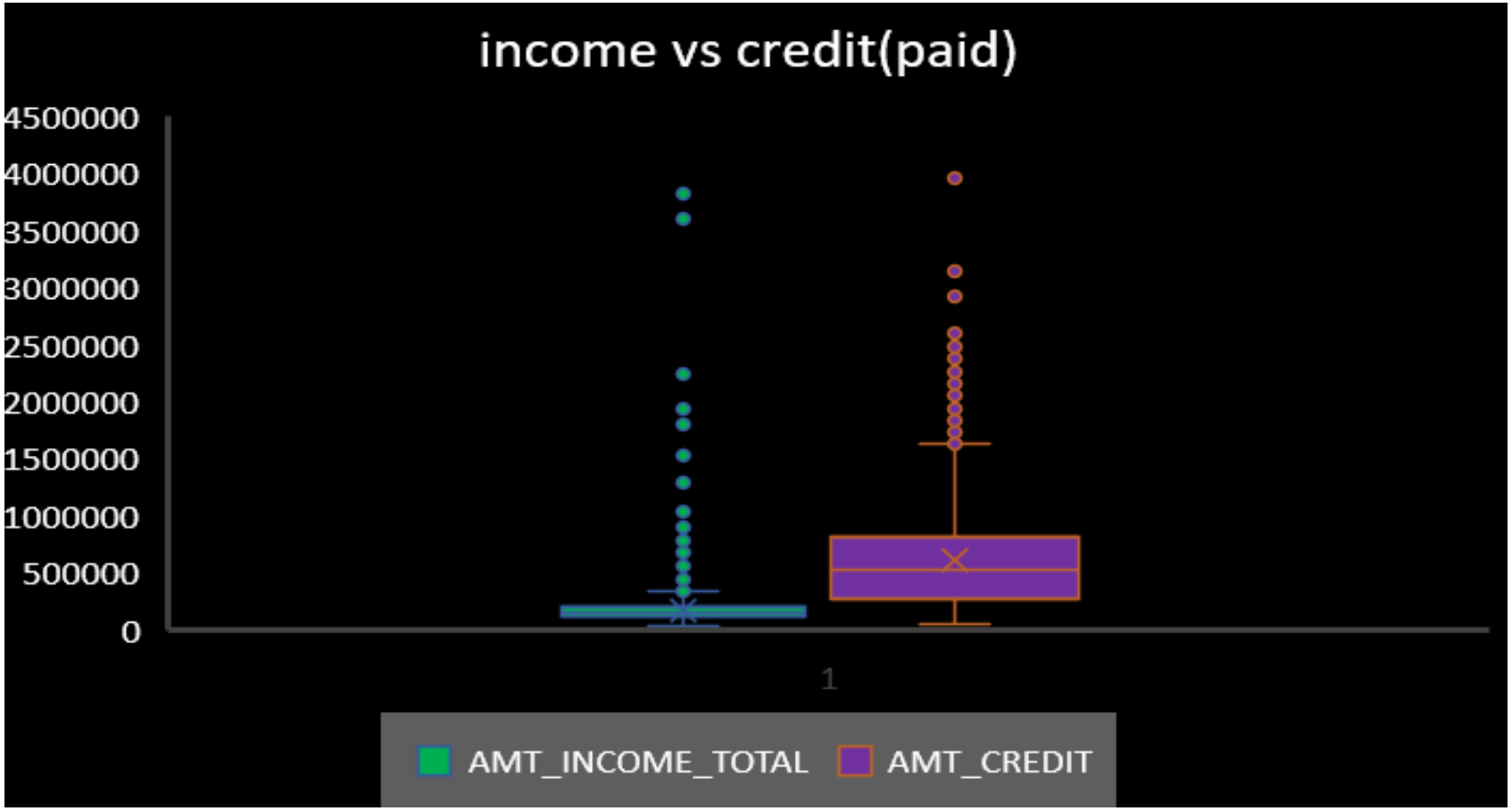
Bivariate Analysis





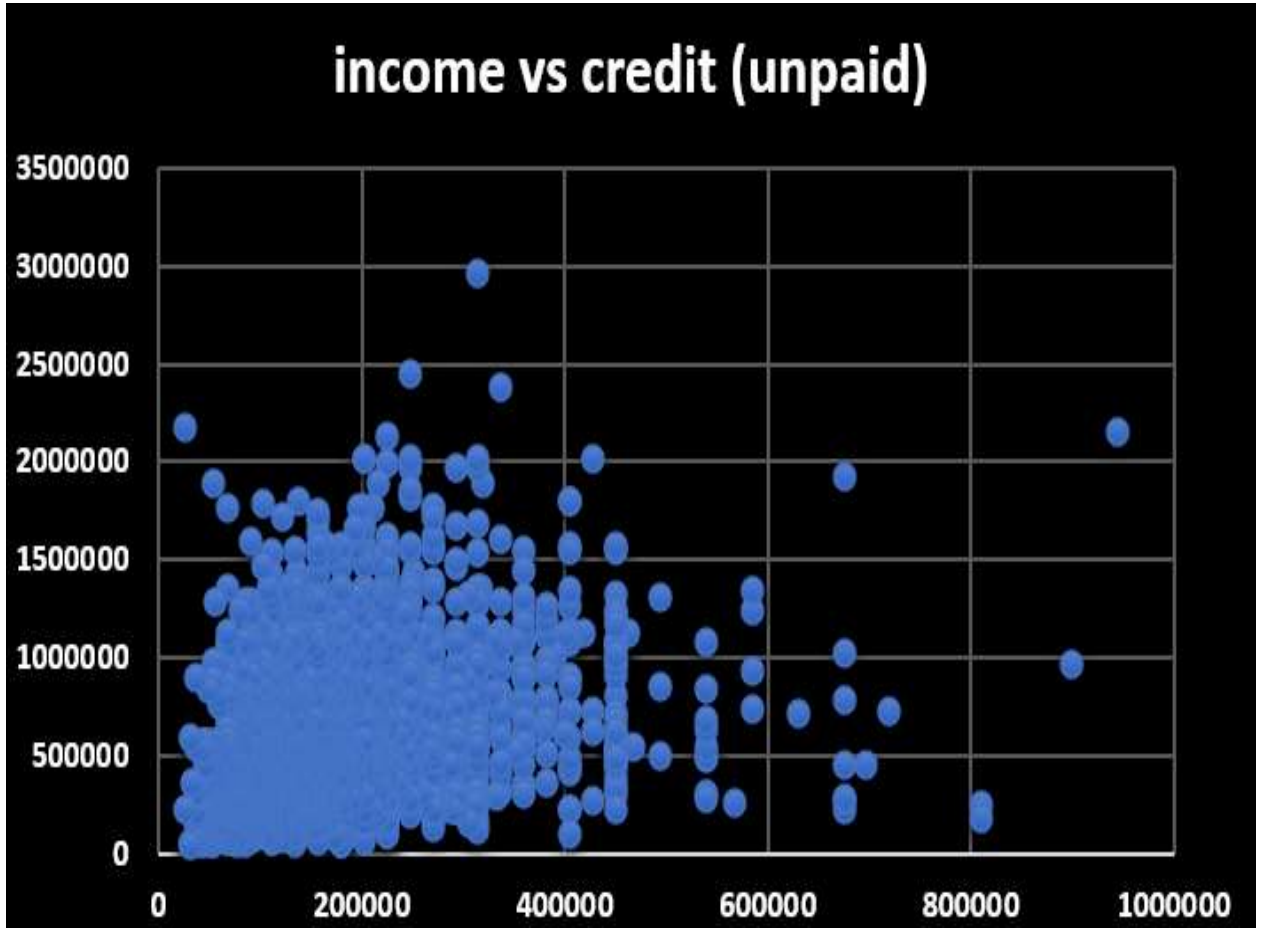
The average of paid and unpaid Cash Loans are much higher than Revolving Loans

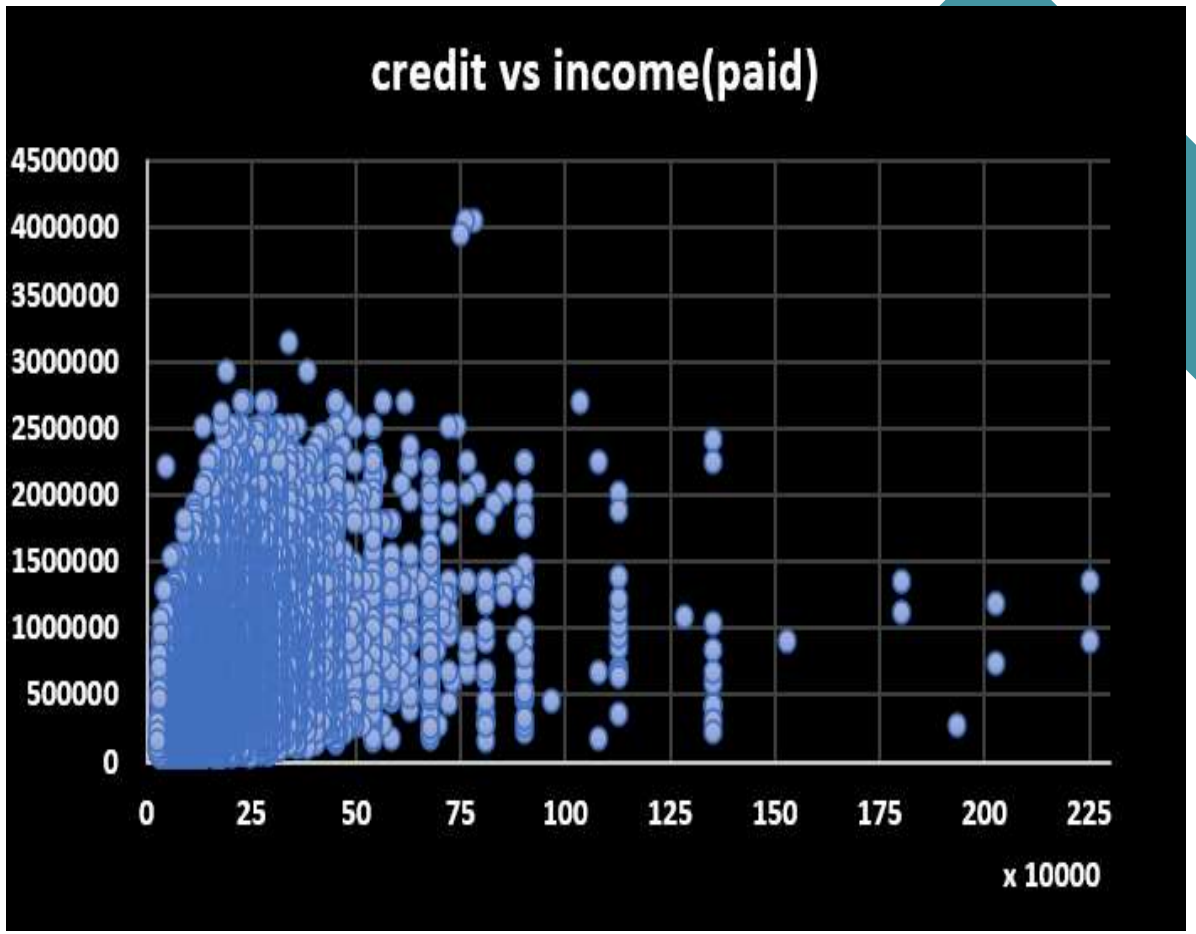
Bivariate Analysis



Performers apply for loan much higher than their income

Bivariate Analysis

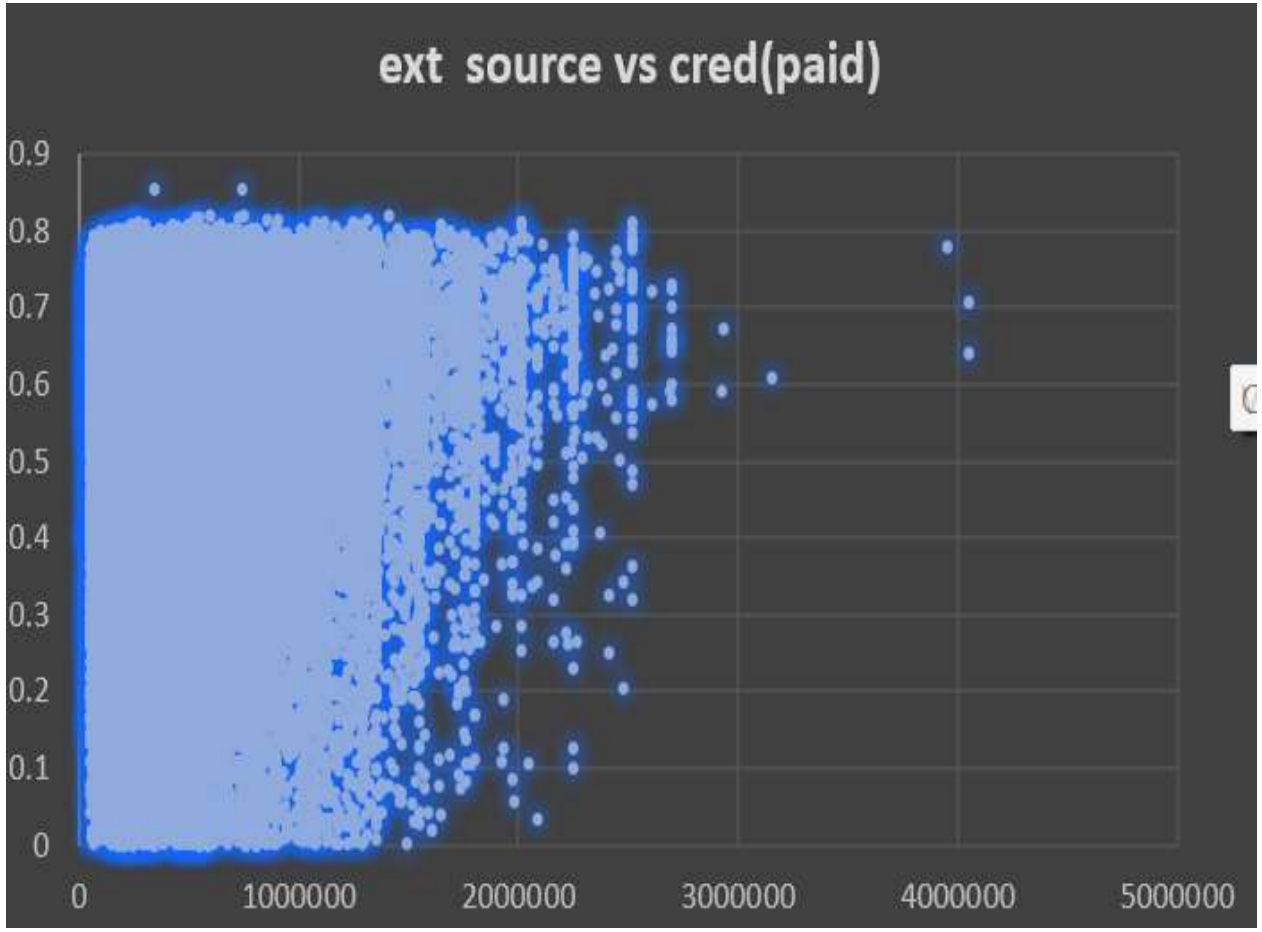


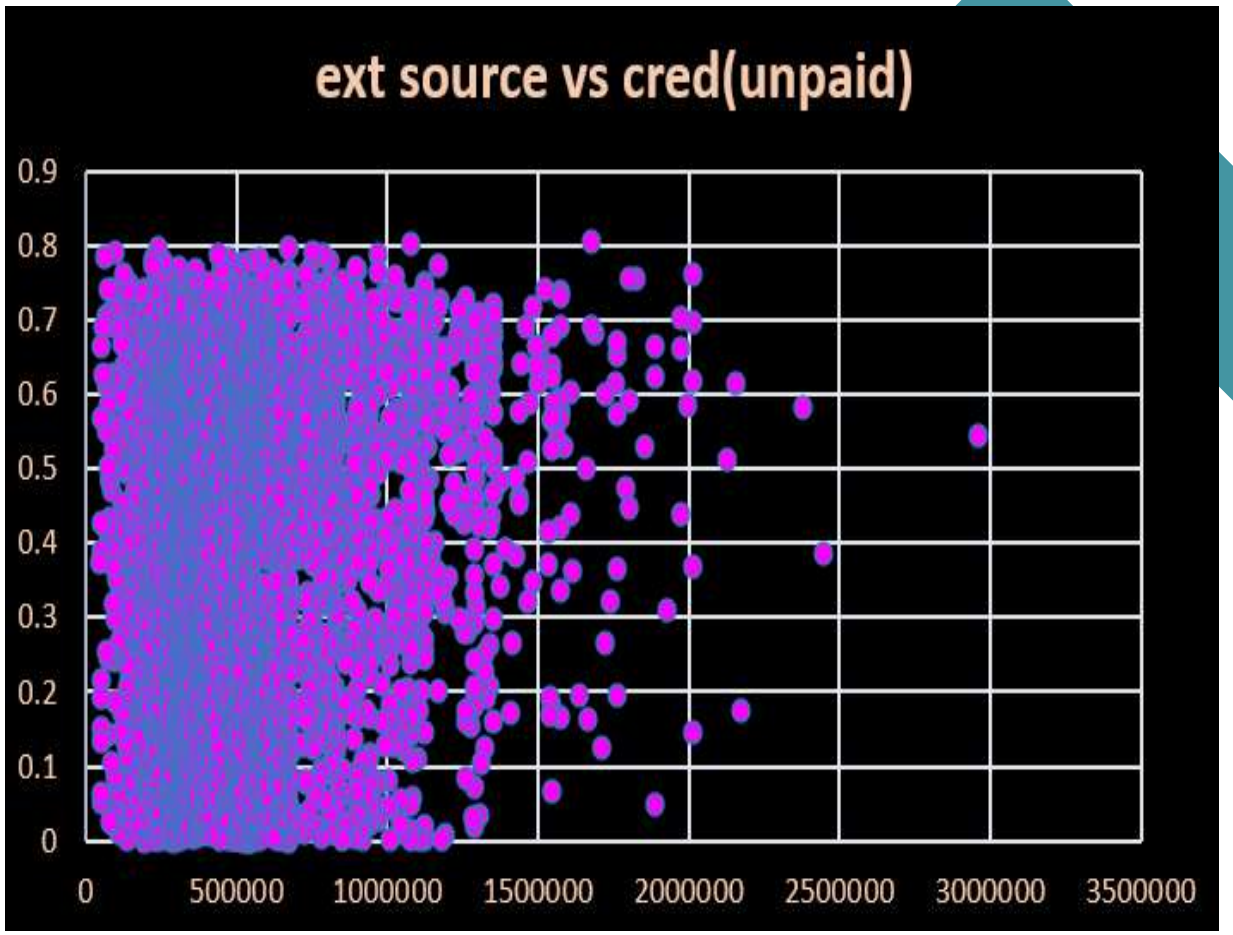


Most Defaulters and Performers with income lesser than 3 Lakhs have applied for loan

Most of the outliers lie beyond the 4th quartile of income

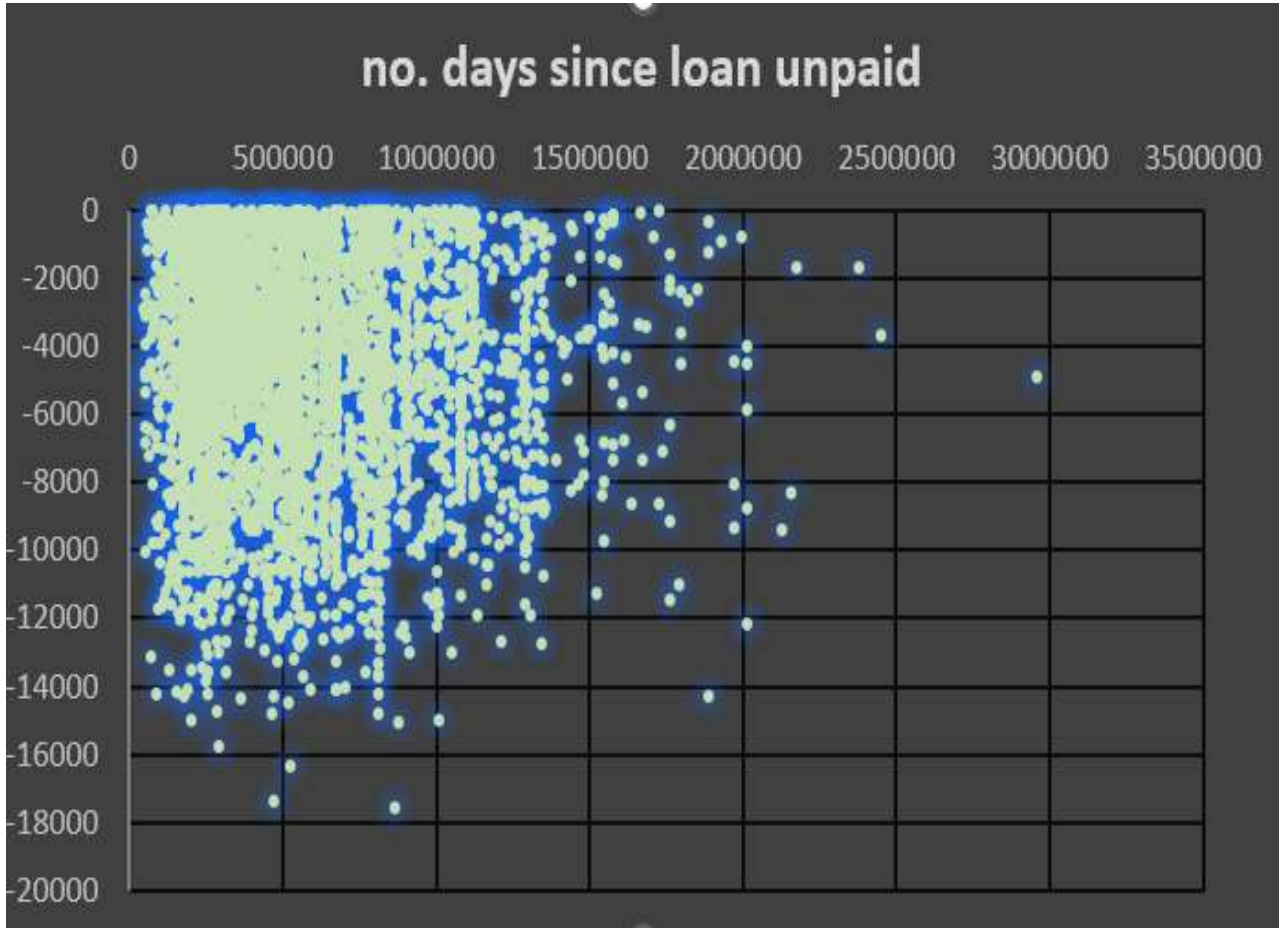
Bivariate Analysis





Most External sources of finance are offered for loans less than 2 Million

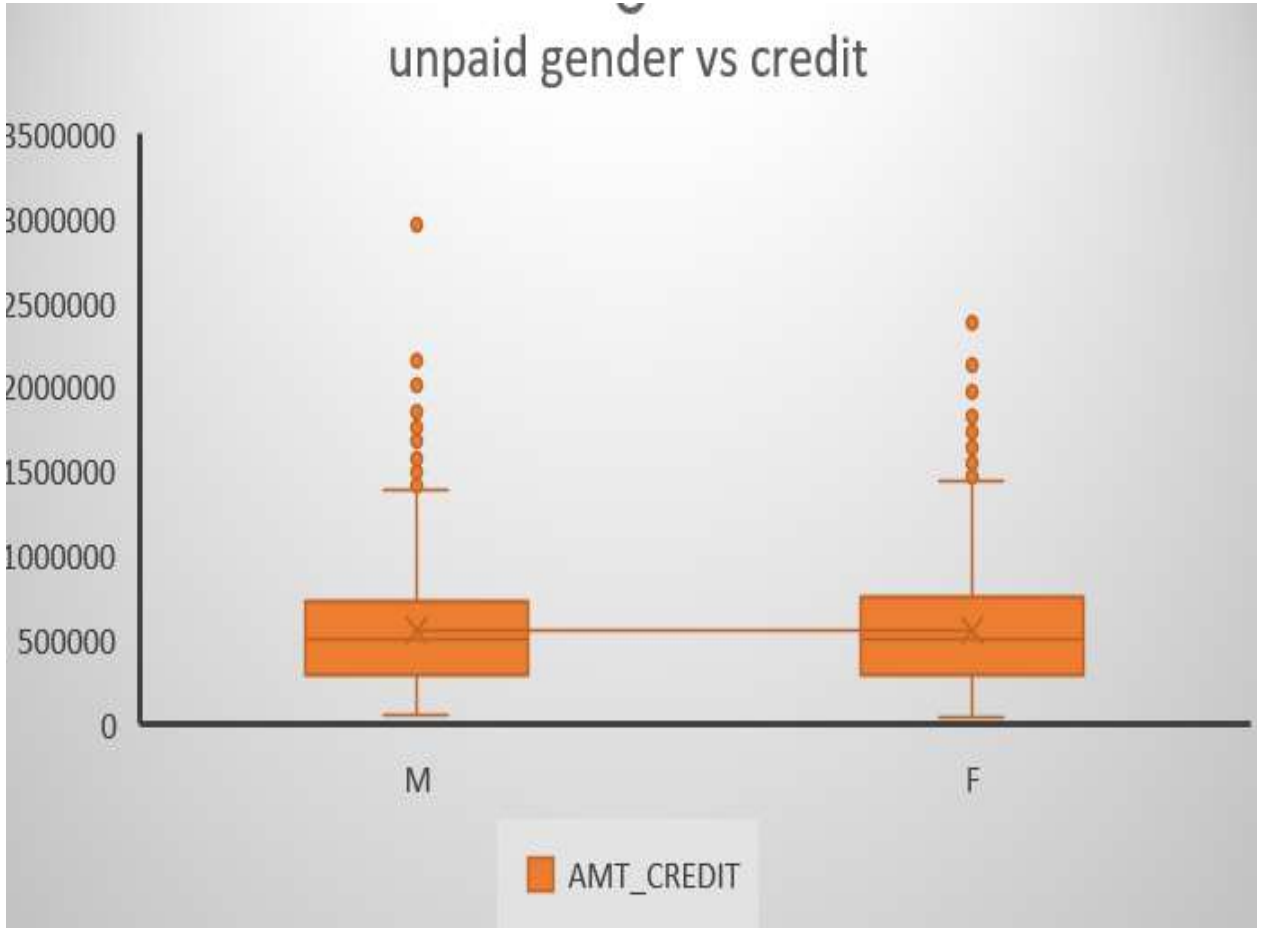
Bivariate Analysis

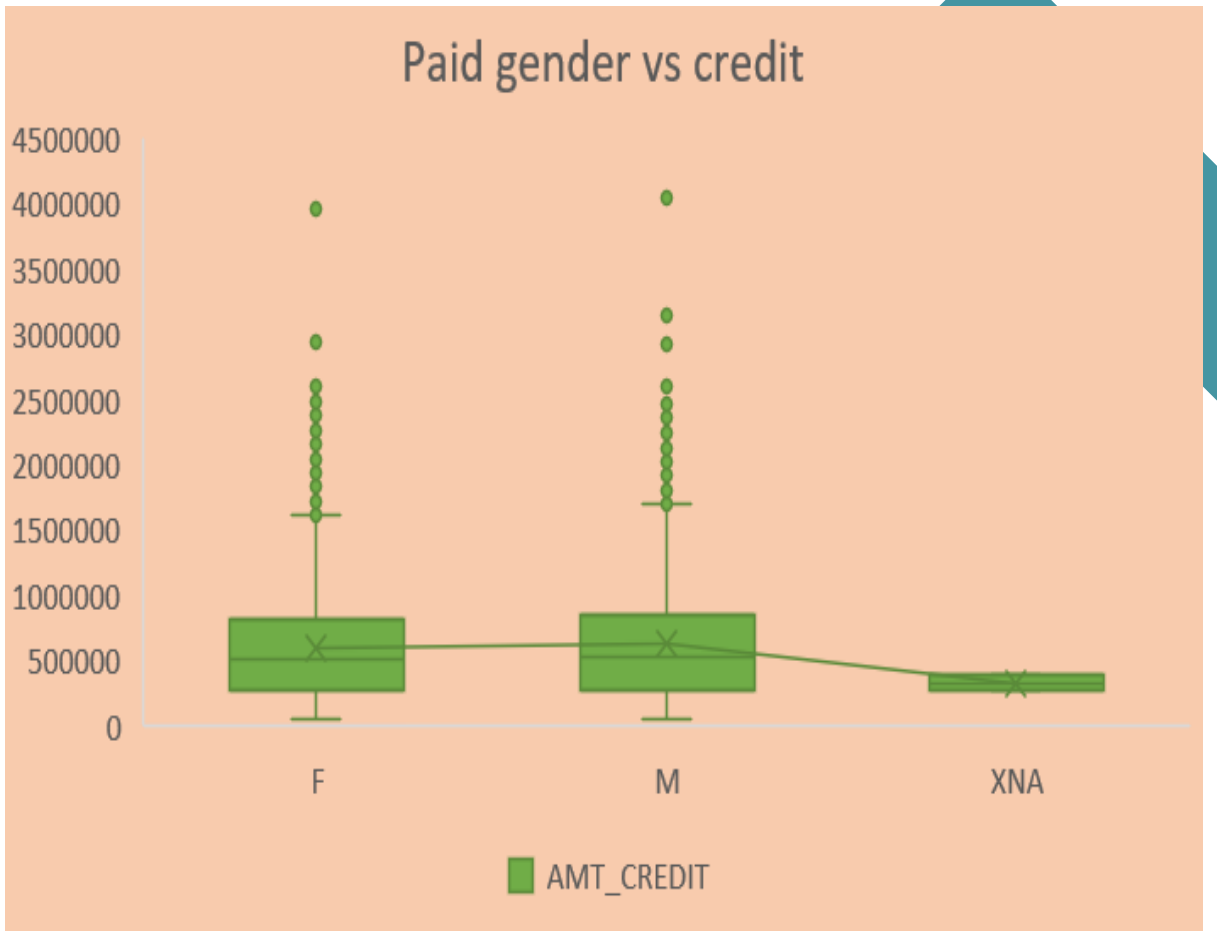


Most Performers changed their application before 15,000 days of applying for loan

Most Defaulters changed their application 12,000 days before applying for loan

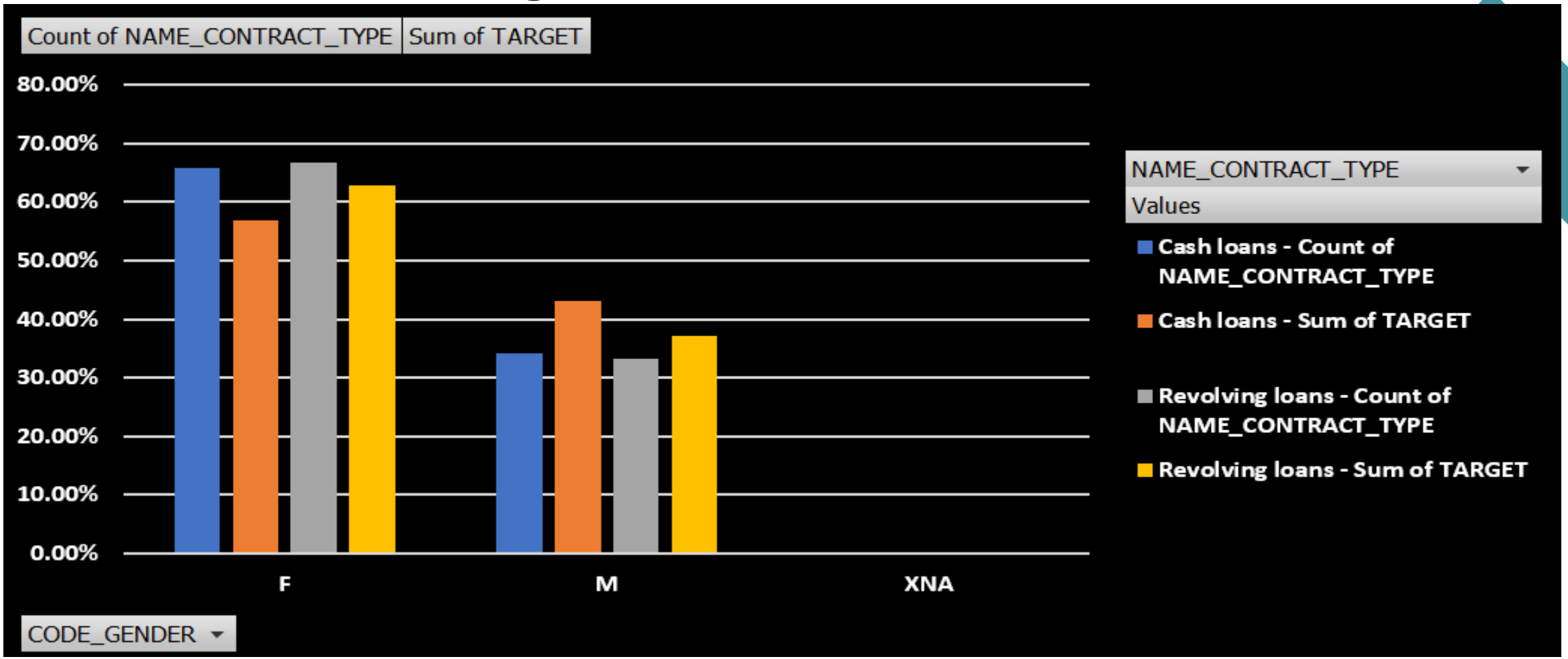
Bivariate Analysis





Female defaulters and Male performers are a majority for an average loan amount of more than 5 Lakhs

Bivariate Analysis



Females are able to pay more of Revolving Loans compared to Cash Loans

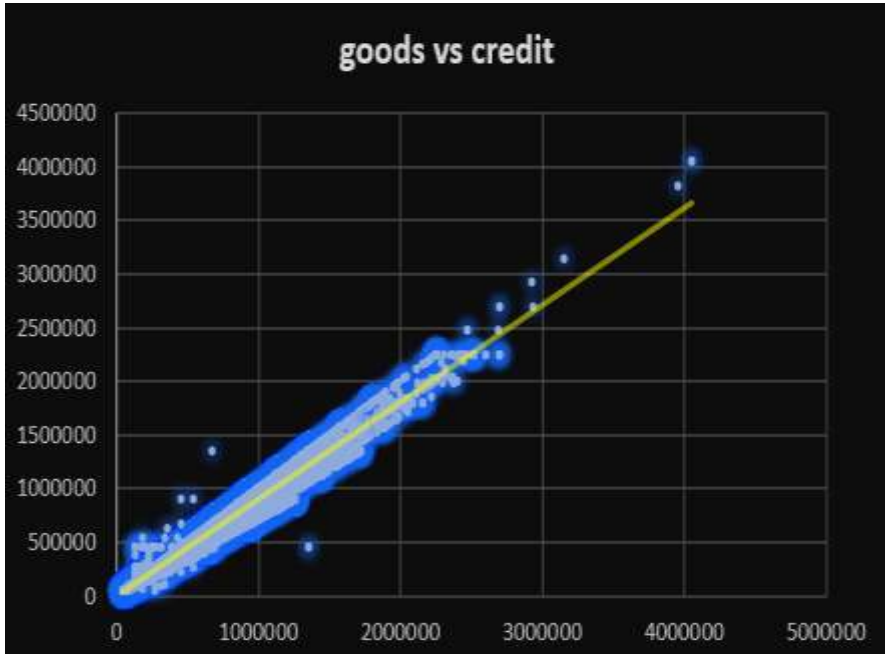
Males are able to pay back more of Cash Loans than Revolving Loans

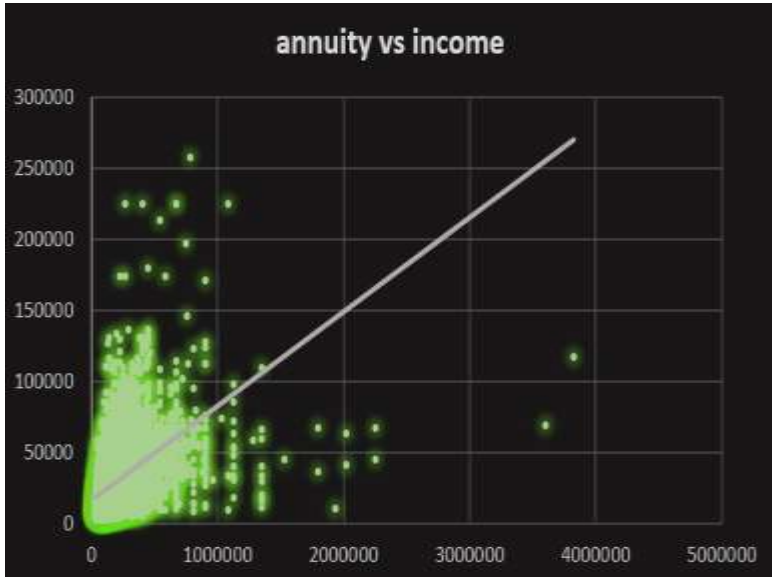
Correlation matrix

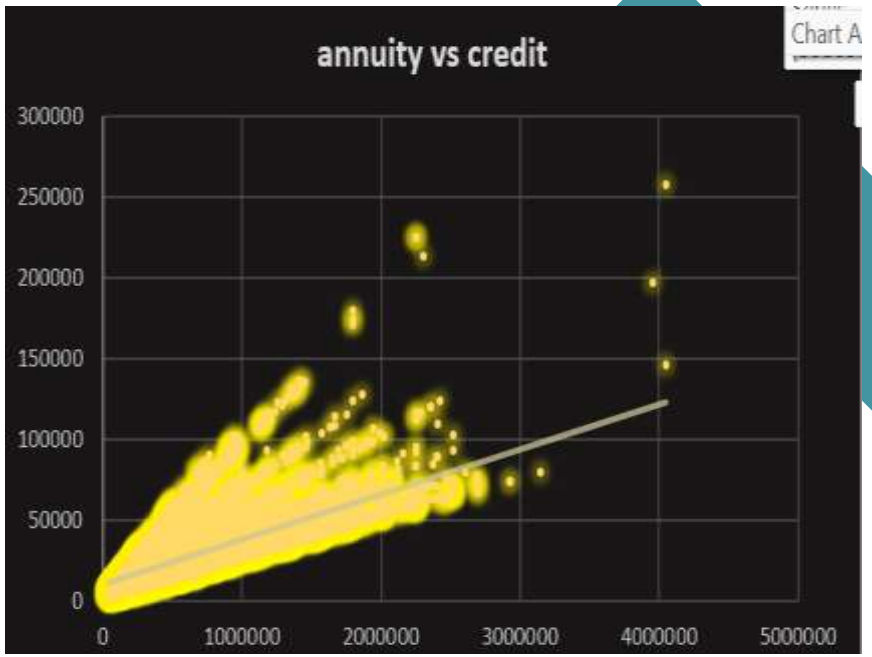
Column1	TARGET	CNT_CHILDR	T_INCOME_	MT_CRI	MT_ANN	GOOI	ULA	AYS_BI	S_EMP	REGIS	_ID_I	FAM_ME	PR_PRO	T_SOUR	T_PHO
TARGET	1.000														
CNT_CHILDREN	0.026	1.000													
AMT_INCOME_TOTAL	0.011	0.010	1.000												
AMT_CREDIT	-0.032	0.005	0.069	1.000											
AMT_ANNUITY	-0.012	0.026	0.083	0.770	1.000										
AMT_GOODS_PRICE	-0.041	0.000	0.070	0.987	0.774	1.000									
REGION_POPULATION_RELATIVE	-0.041	-0.026	0.030	0.095	0.115	0.099	1.000								
DAYS_BIRTH	0.077	0.329	0.016	-0.059	0.008	-0.058	-0.032	1.000							
DAYS_EMPLOYED	-0.040	-0.240	-0.032	-0.070	-0.110	-0.068	-0.004	-0.614	1.000						
DAYS_REGISTRATION	0.042	0.181	0.010	0.003	0.033	0.006	-0.059	0.334	-0.205	1.000					
DAYS_ID_PUBLISH	0.047	-0.032	0.004	-0.012	0.007	-0.014	-0.004	0.271	-0.270	0.104	1.000				
CNT_FAM_MEMBERS	0.013	0.880	0.011	0.064	0.077	0.062	-0.023	0.277	-0.230	0.170	-0.026	1.000			
HOUR_APPR_PROCESS_START	-0.032	-0.006	0.018	0.057	0.053	0.066	0.168	0.091	-0.088	-0.008	0.034	-0.012	1.000		
EXT_SOURCE_2	-0.158	-0.018	0.020	0.138	0.129	0.147	0.201	-0.094	-0.026	-0.061	-0.047	0.003	0.157	1.000	
DAYS_LAST_PHONE_CHANGE	0.056	-0.002	-0.005	-0.076	-0.067	-0.080	-0.048	0.080	0.028	0.052	0.091	-0.023	-0.018	-0.192	1.000

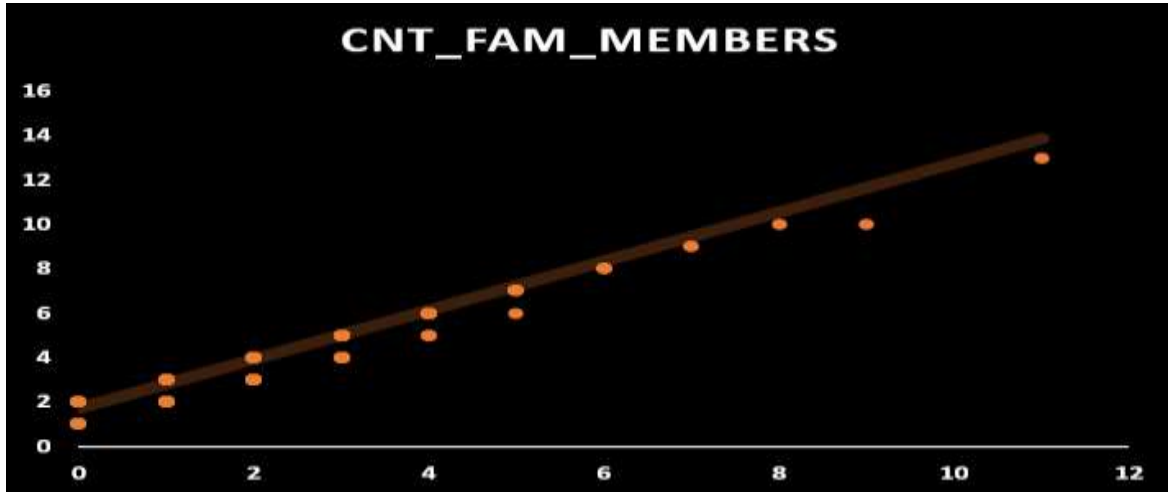
Correlation

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
AMT_INCOME_TOTAL	1			
AMT_CREDIT	0.377985089	1		
AMT_ANNUITY	0.451148293	0.77077712	1	
AMT_GOODS_PRICE	0.384621454	0.987001704	0.775843488	1









Agenda for previous_application.csv

- Data Handling
- Data Imbalance
- Segmented, Bivariate, Multivariate Analysis
- Correlation Analysis
- Result and Insights

Data Handling – Identifying Missing Data

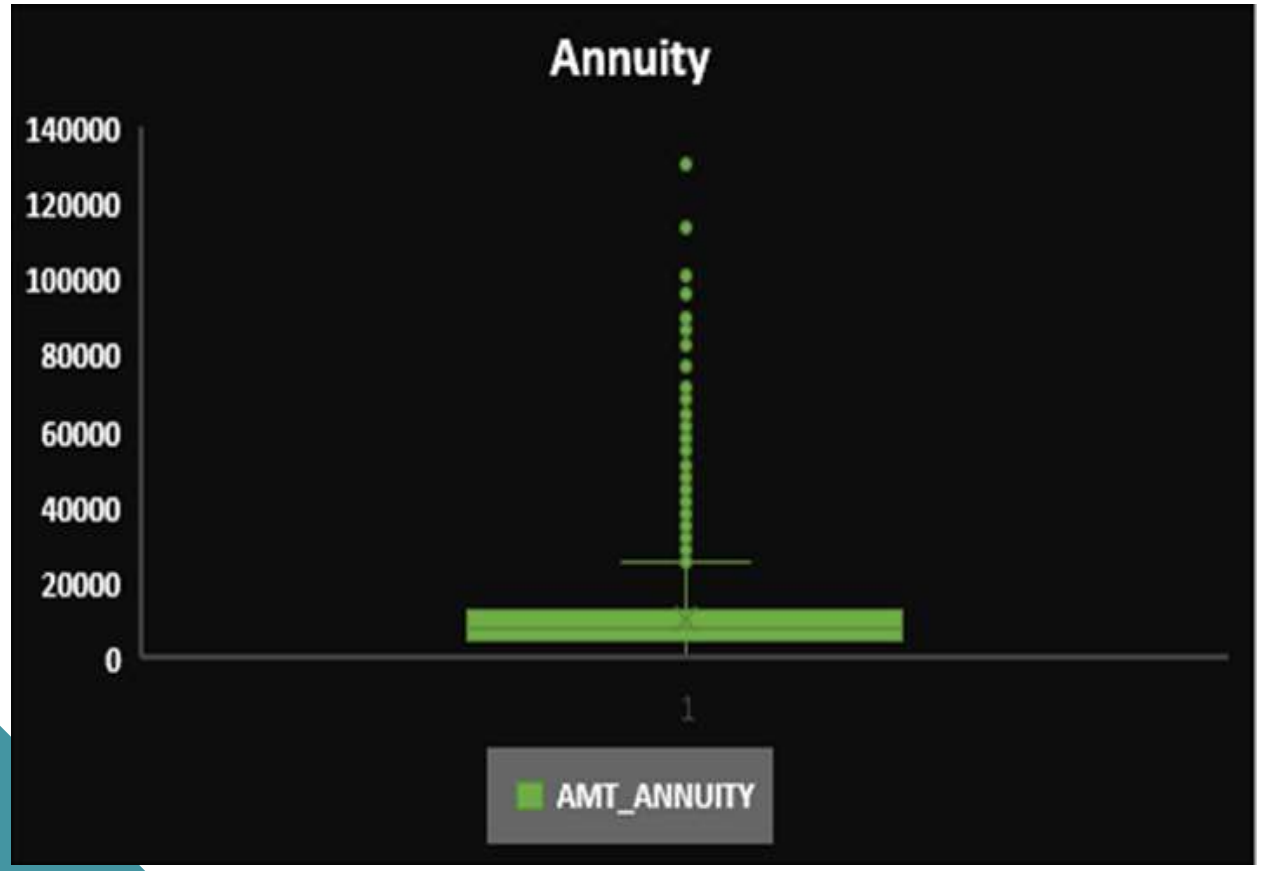
AF to AK

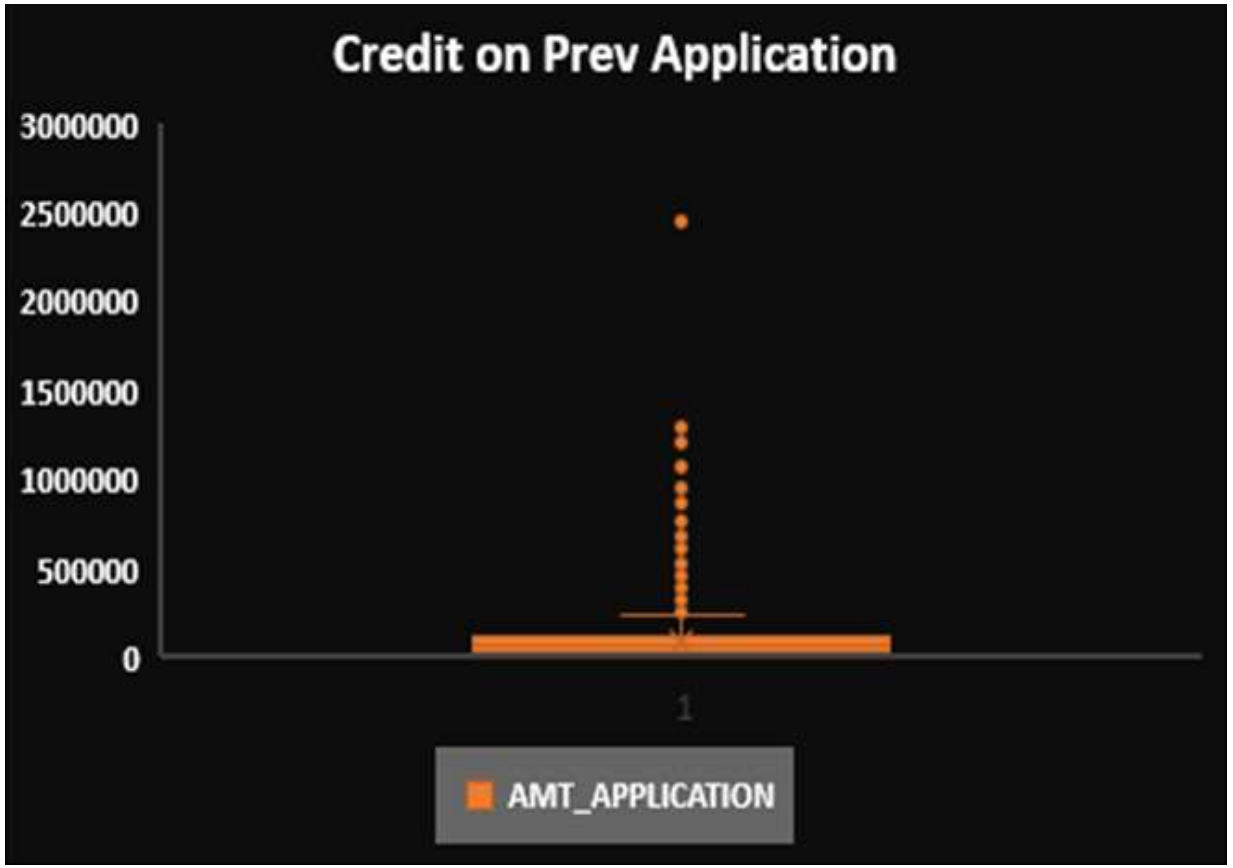
DAYS_FIRST_DRAWING
DAYS_FIRST_DUE
DAYS_LAST_DUE_1ST_VERSION
DAYS_LAST_DUE
DAYS_TERMINATION
NFLAG_INSURED_ON_APPROVAL

- Amt_Annuity had more than 21.8% empty rows which were removed(=Countblank(range/number of cells*100)
- Name_Type_Suit and Product_combination had blanks more than 30% for which empty rows were removed
- Rate_Primary and Rate_privilege columns had more than 99% blanks, hence these columns were deleted
- Down_payment,Rate_down_Pyt,Amt_Goods_Price had empty rows which were removed
- 0.16% of product combination was blank changed to XNA, as corresponding column product type also had XNA
- Columns AF to AK had 38% blank rows, which were unwanted columns hence columns removed
- Duplicates were removed using Remove Duplicates from Data in table

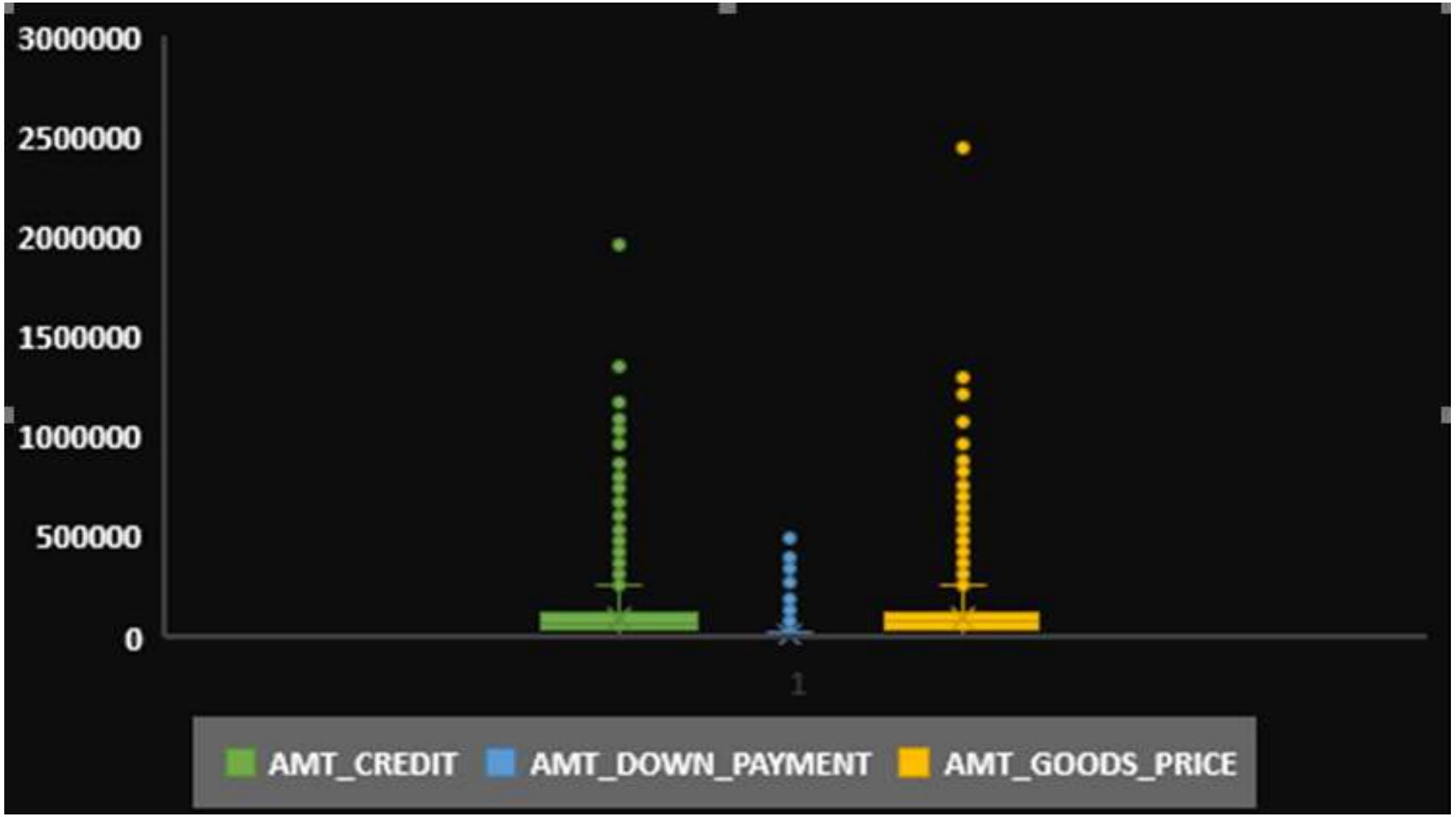
Data Handling – Identifying Outliers

- Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
- Outliers have to be removed before plotting graphs, to get balanced data
- Formula used:
- **=QUARTILE.INC(range,0)- lower quartile**
- **=QUARTILE.INC(range,1)- middle quartile**
- **=QUARTILE.INC(range,2)- middle quartile**
- **=QUARTILE.INC(range,3)- middle quartile**
- **=QUARTILE.INC(range,4)- upper quartile**





Most outliers are in the 3rd quartile and few others lie in the 4th quartile

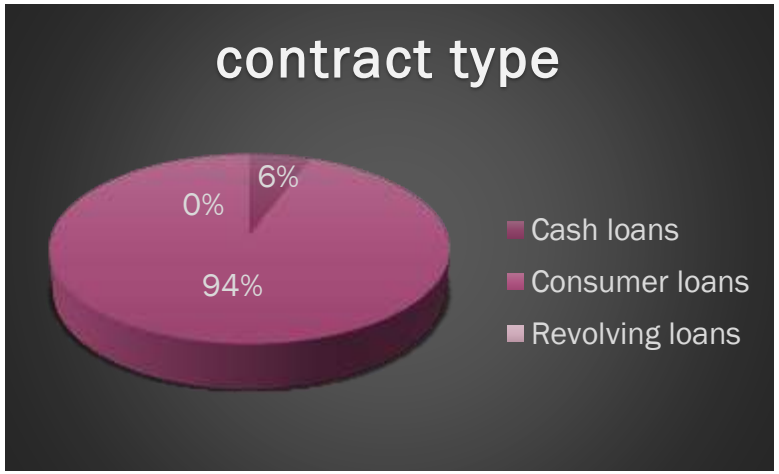


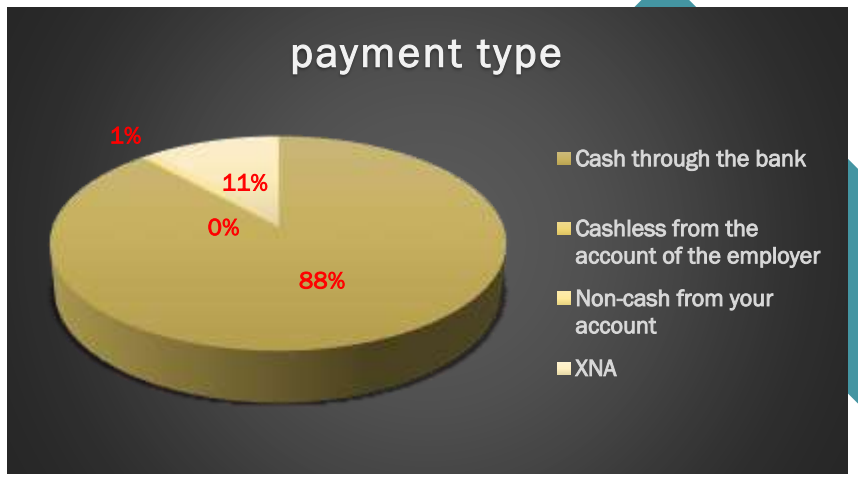
For Amt_credit most outliers lie in the range of 12 Lakhs

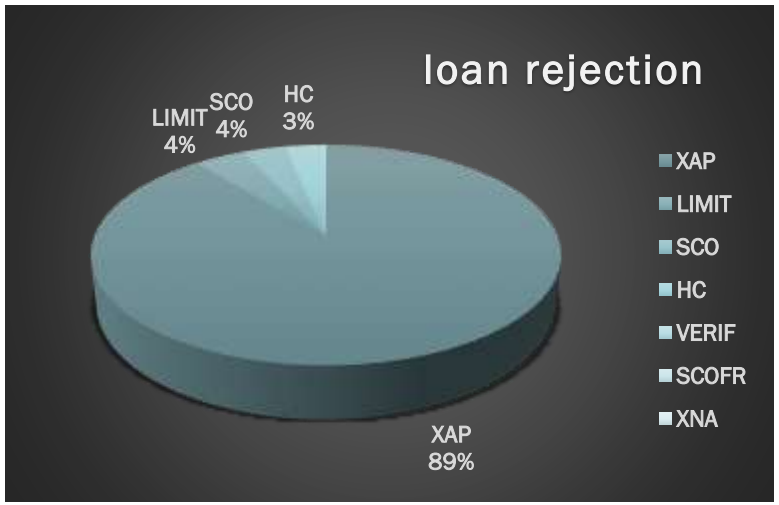
For Amt_Down_Payment most outliers are between 1 and 8 Lakhs

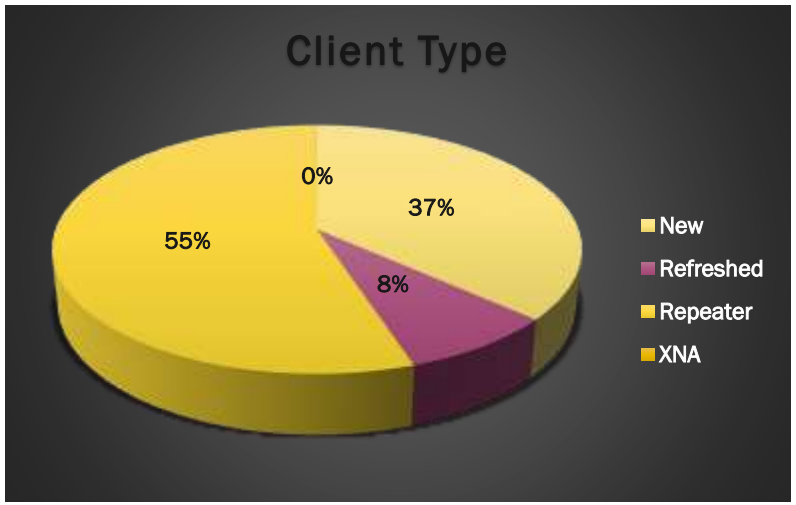
For Amt_Goods_Price most outliers are between 2 to 10 Lakhs

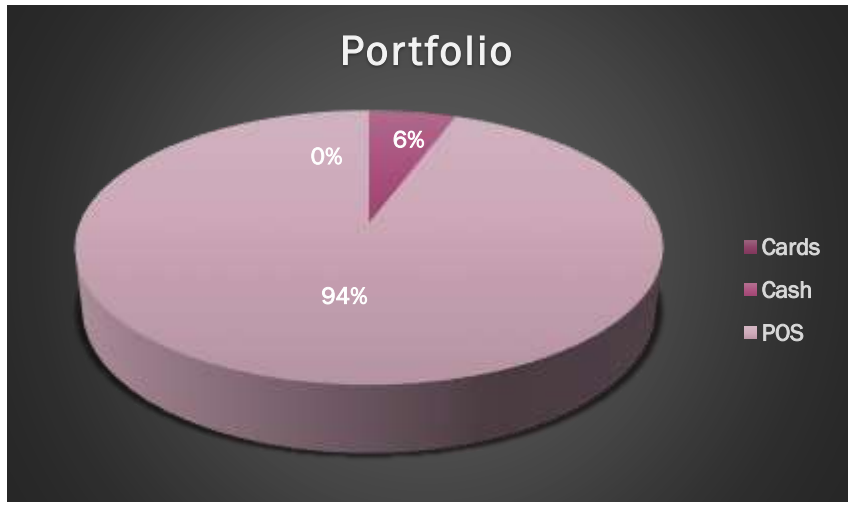
Data Imbalance and Univariate Analysis



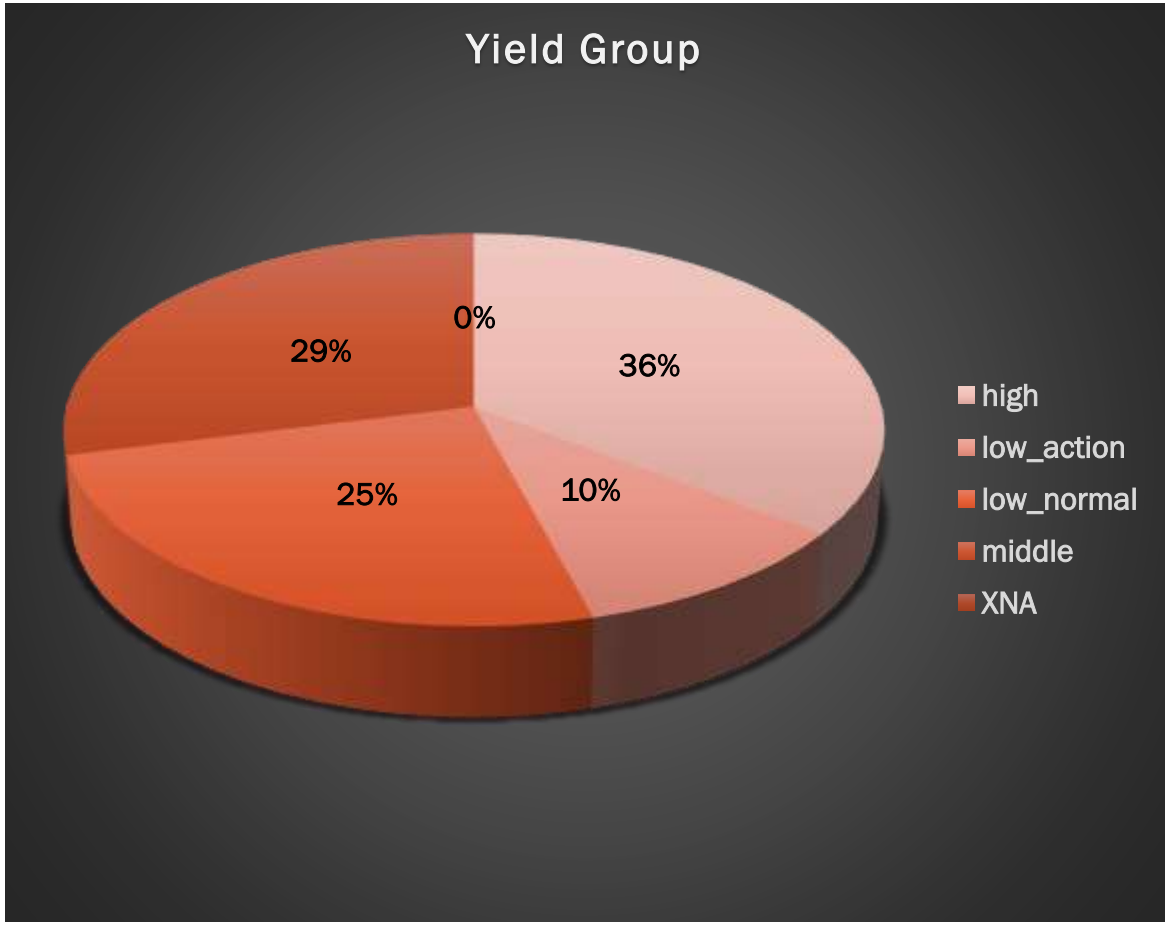


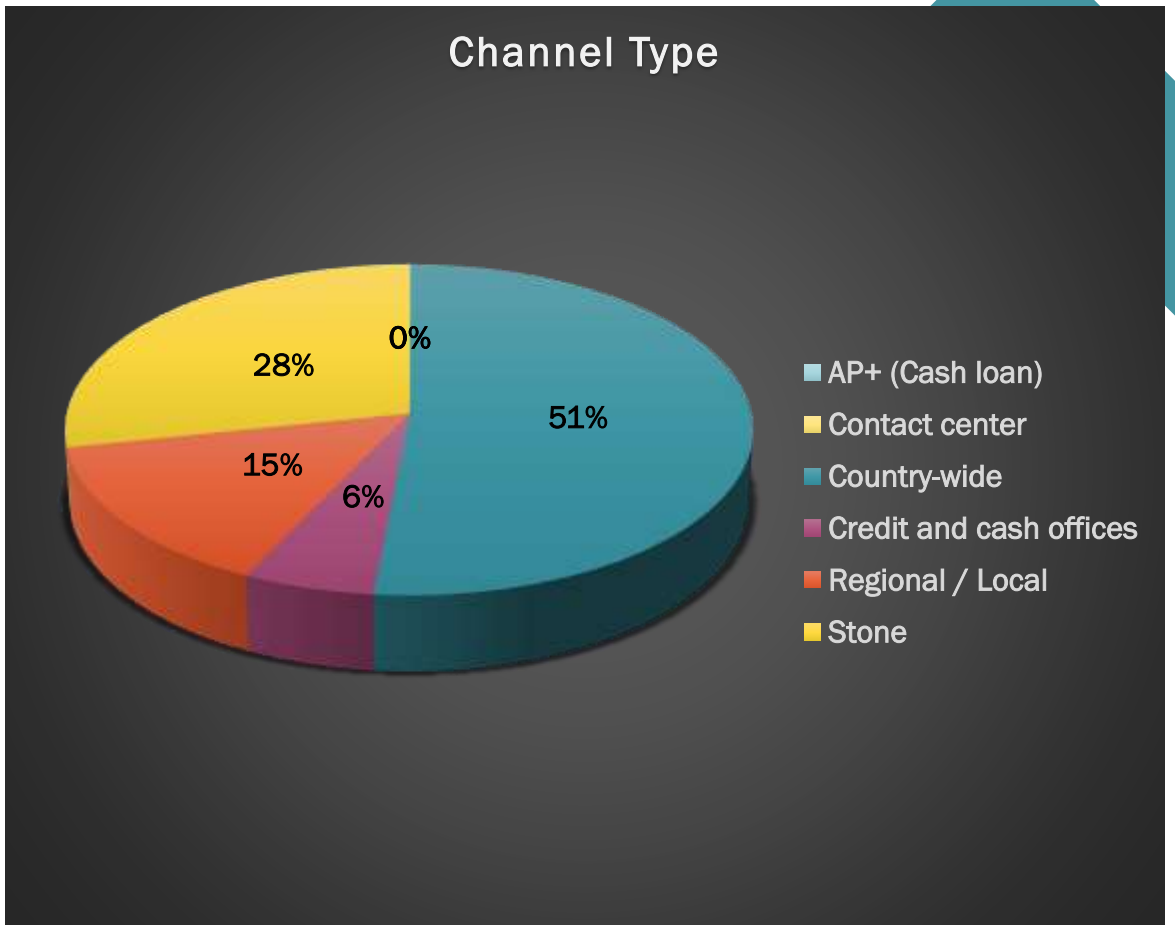






Data Imbalance and Univariate Analysis

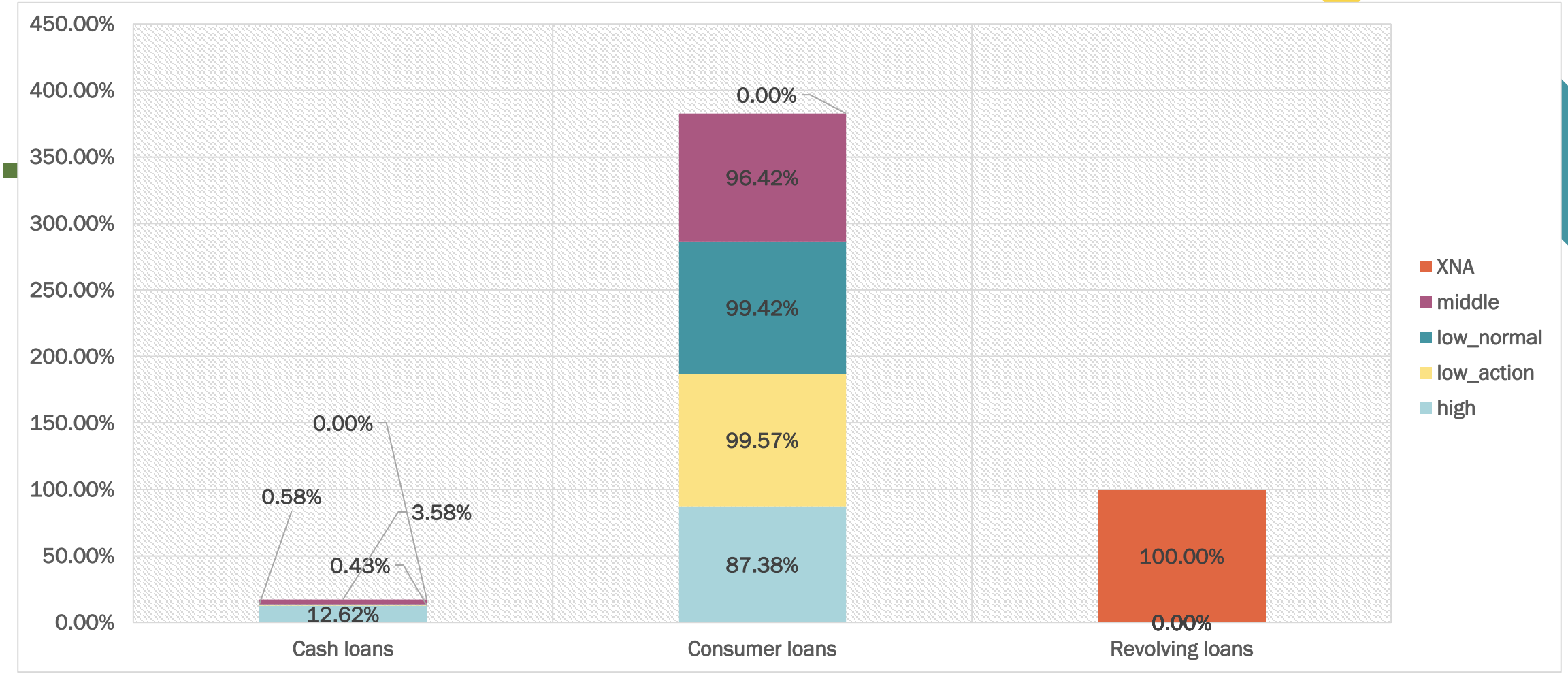




Insights

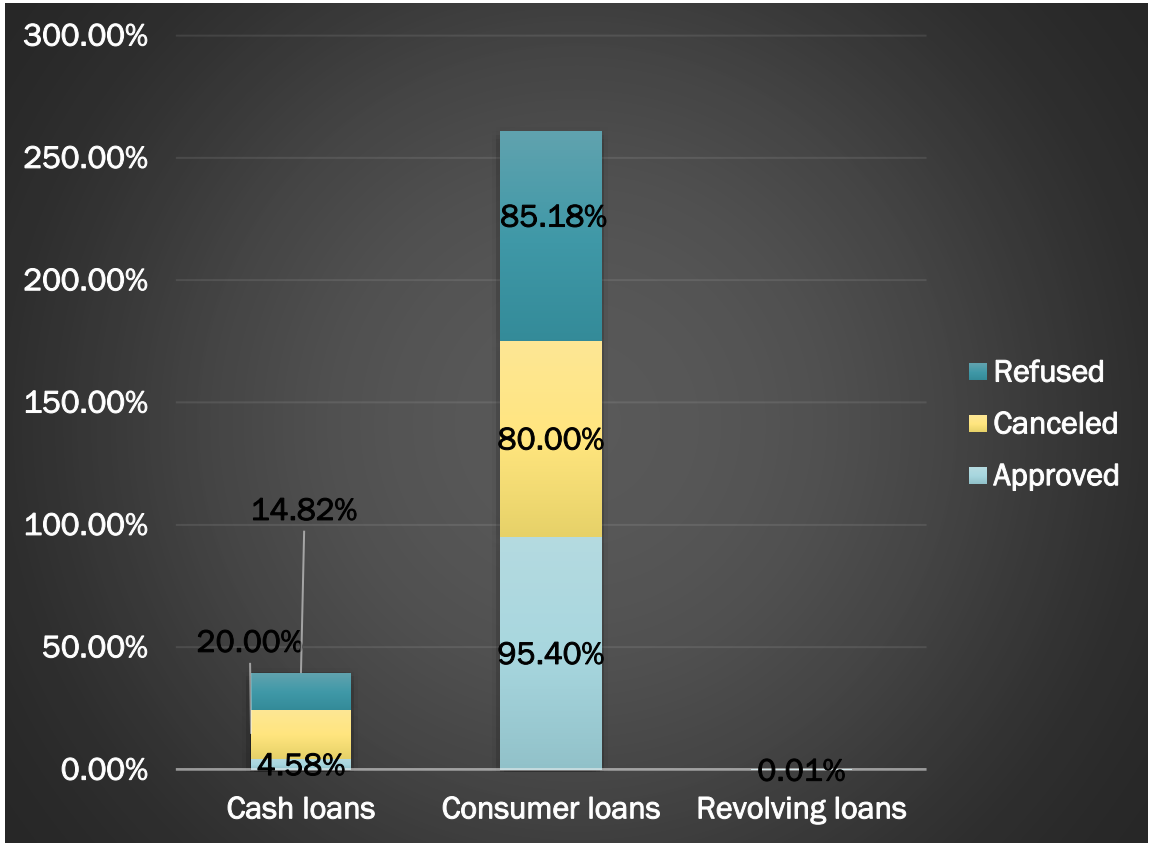
- Majority of previous applications were for Consumer Loans
- Most of the loans were approved and yield is mostly high
- Most of the loans were paid through bank by cash
- Loan rejections were mostly of XAP
- Majority of clients were Repeaters with POS type of Loan Portfolio
- Loans are generally received from Country-wide

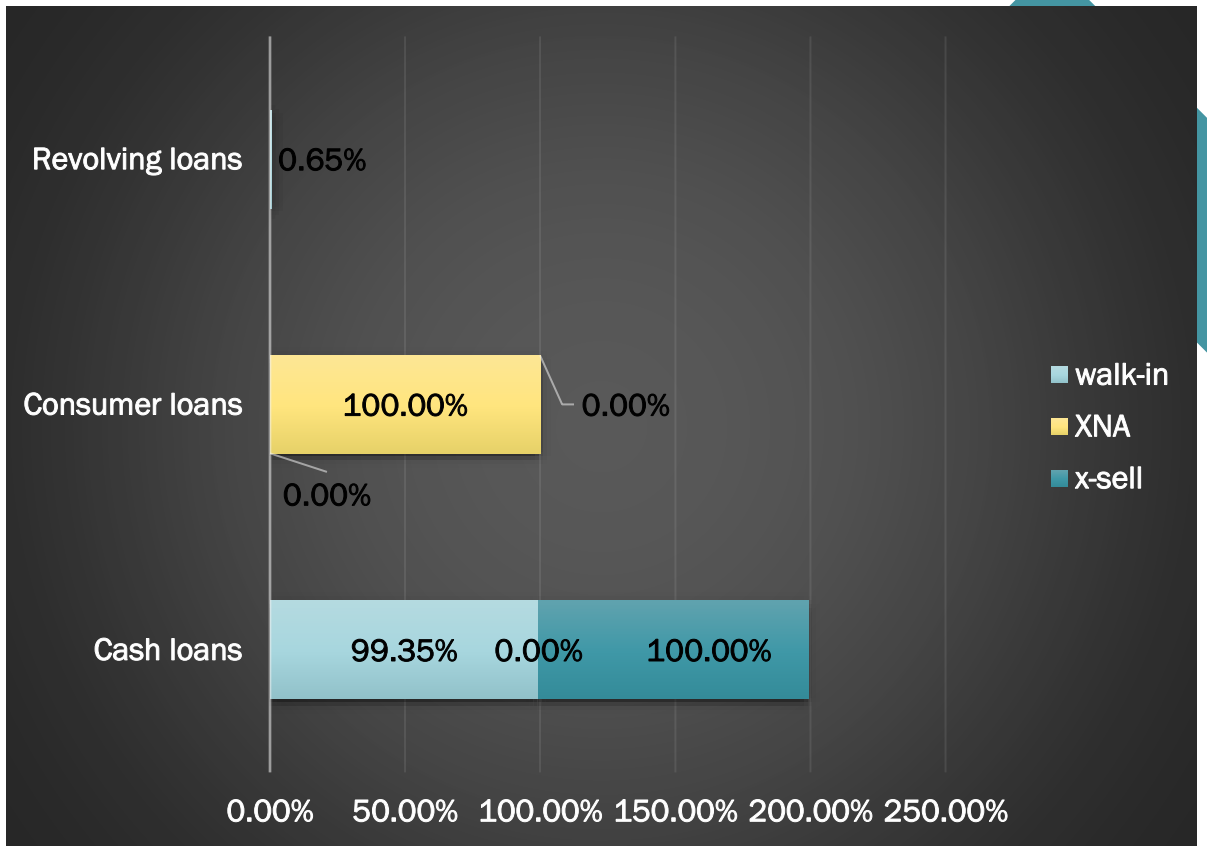
Segmented Analysis



Revolving Loans have a yield of XNA and Consumer Loans have mostly low_action , low_normal and middle yields

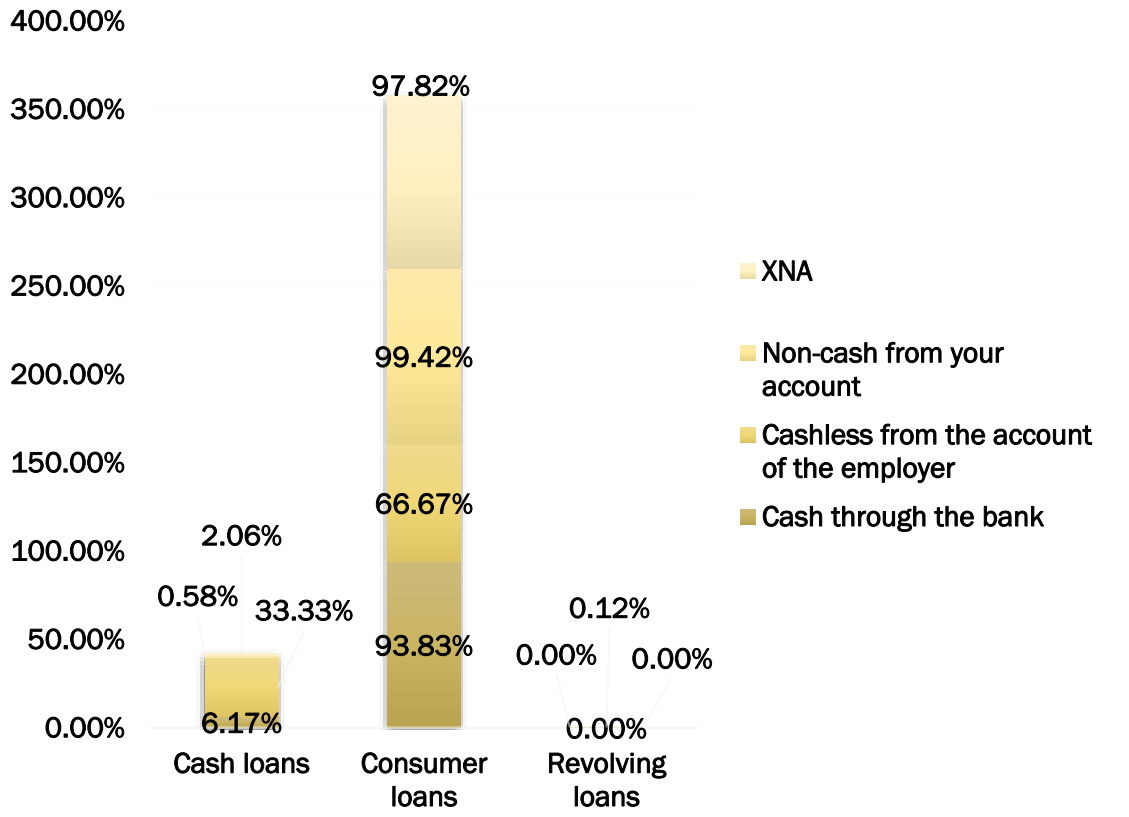
Segmented Analysis

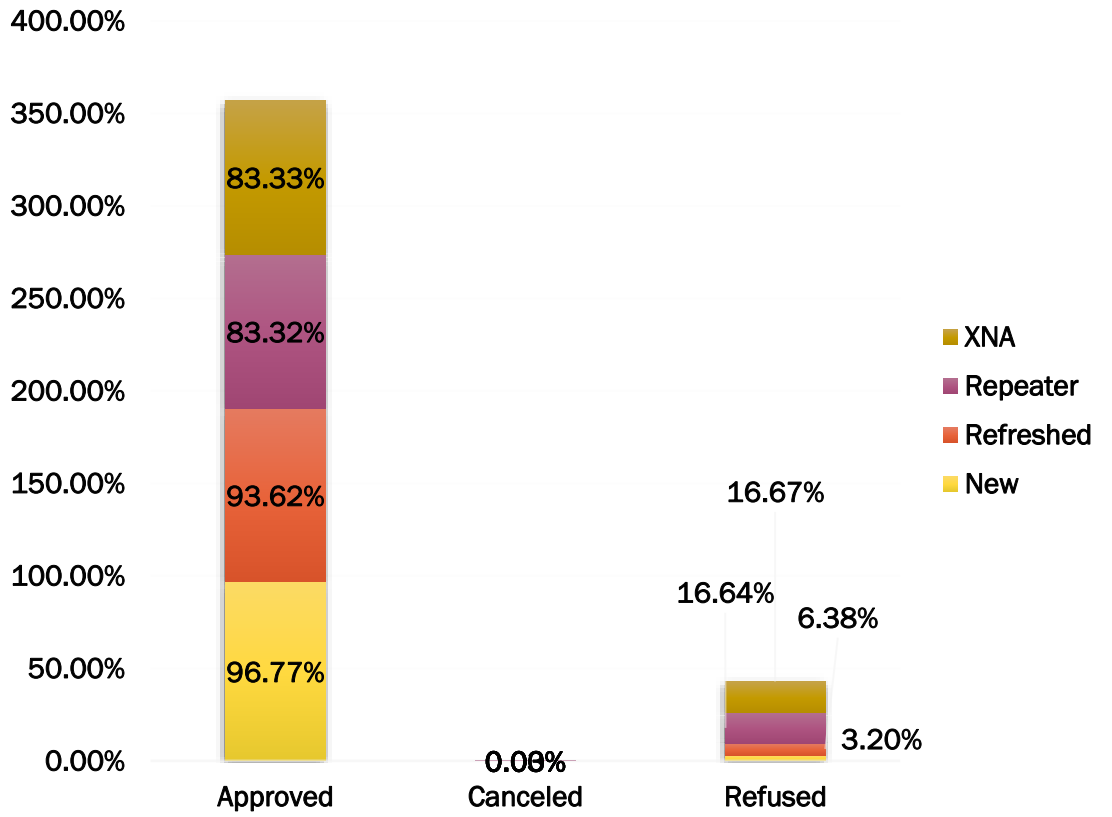




Majority of approved loans were Consumer Type and product through Cash Loans is mostly walk-in and for Consumer Loans-XNA.

Segmented Analysis

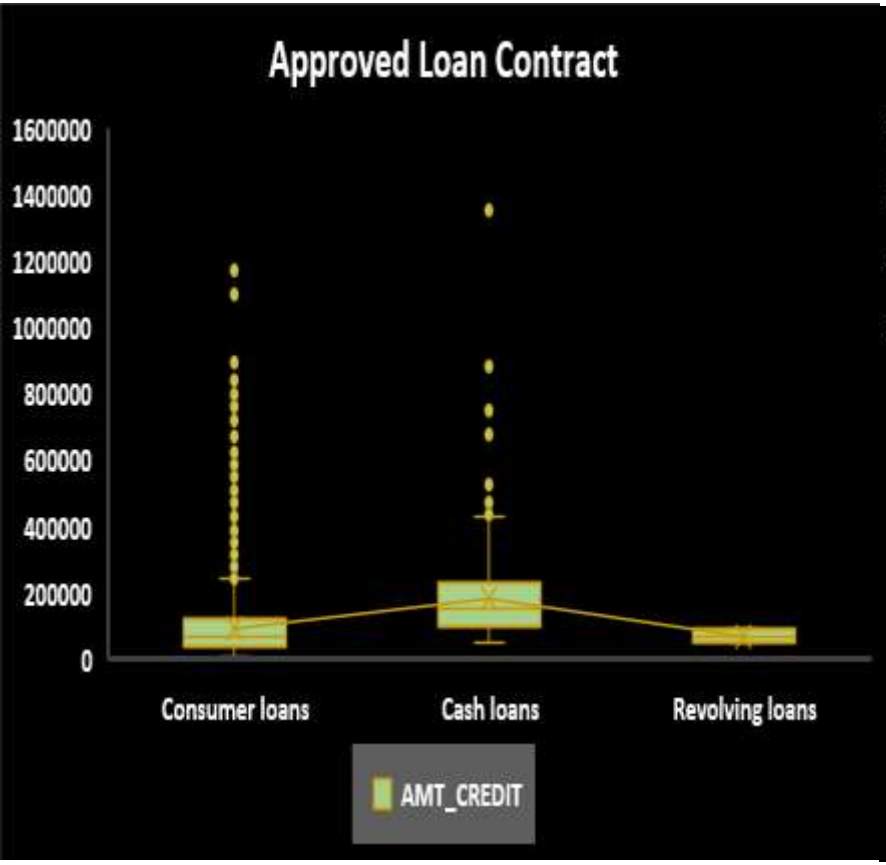


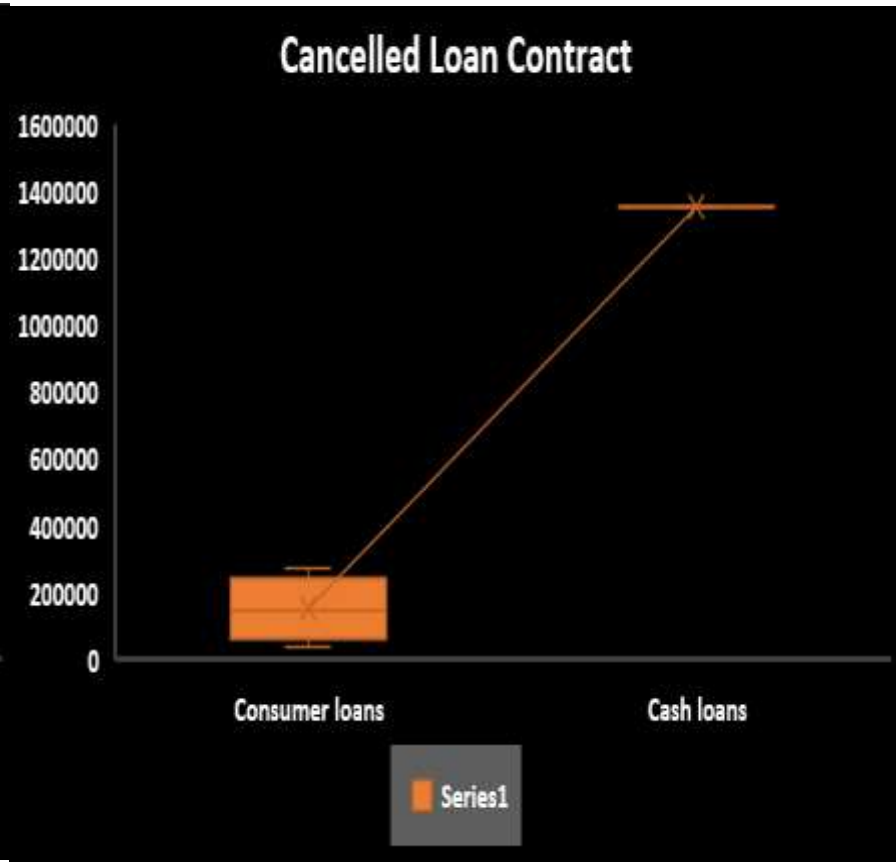


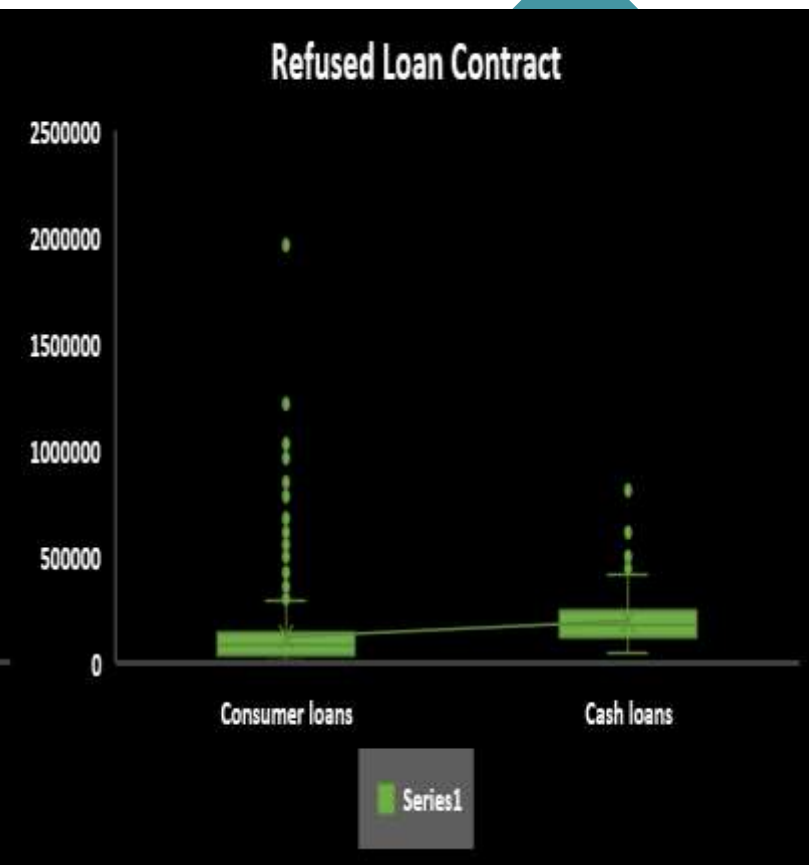
Majority of Consumer Loans were paid as Non-Cash from account,XNA and Cash through bank.

Rate of Loan Approval for freshers and Refreshers is high

Bivariate Analysis



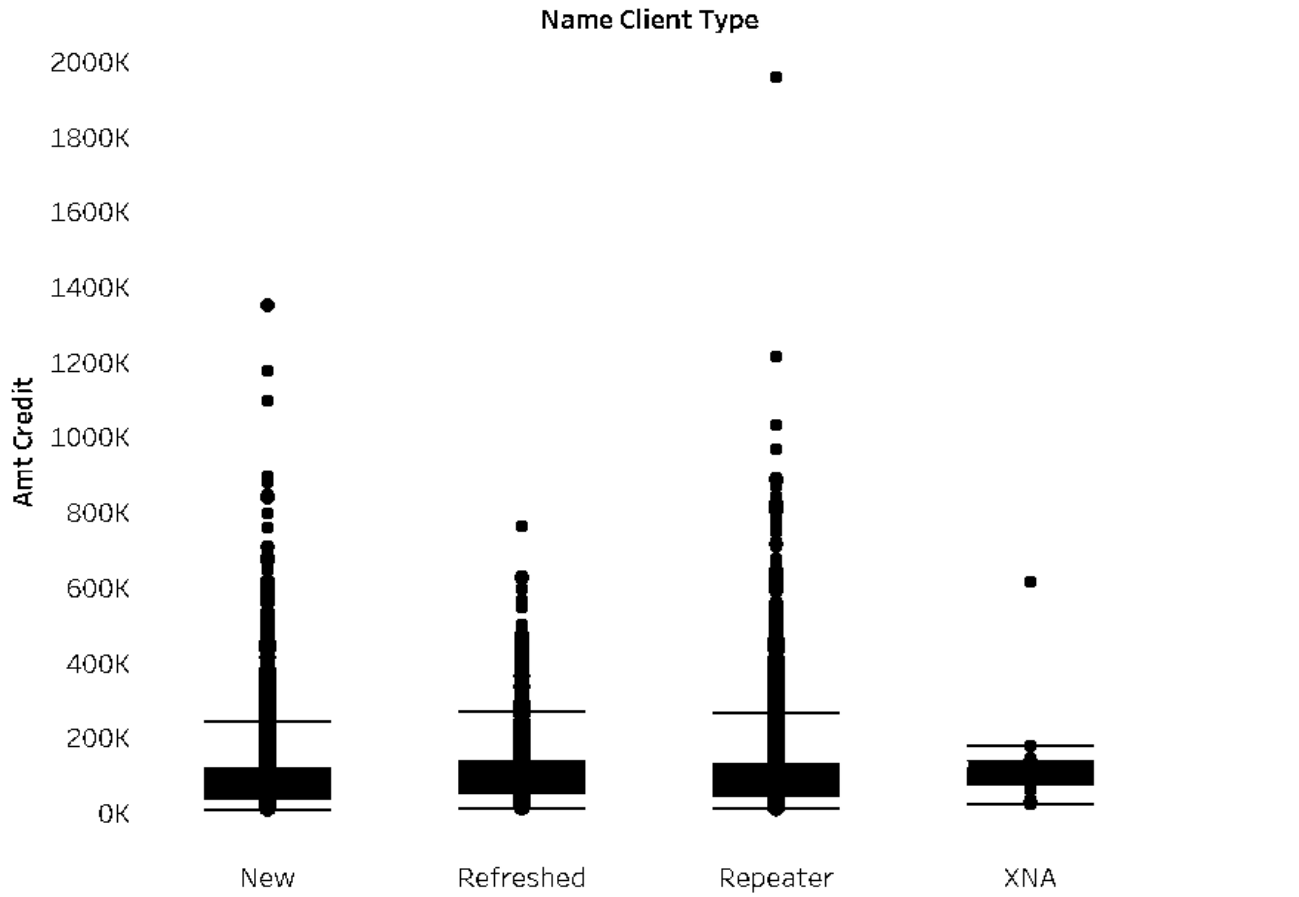


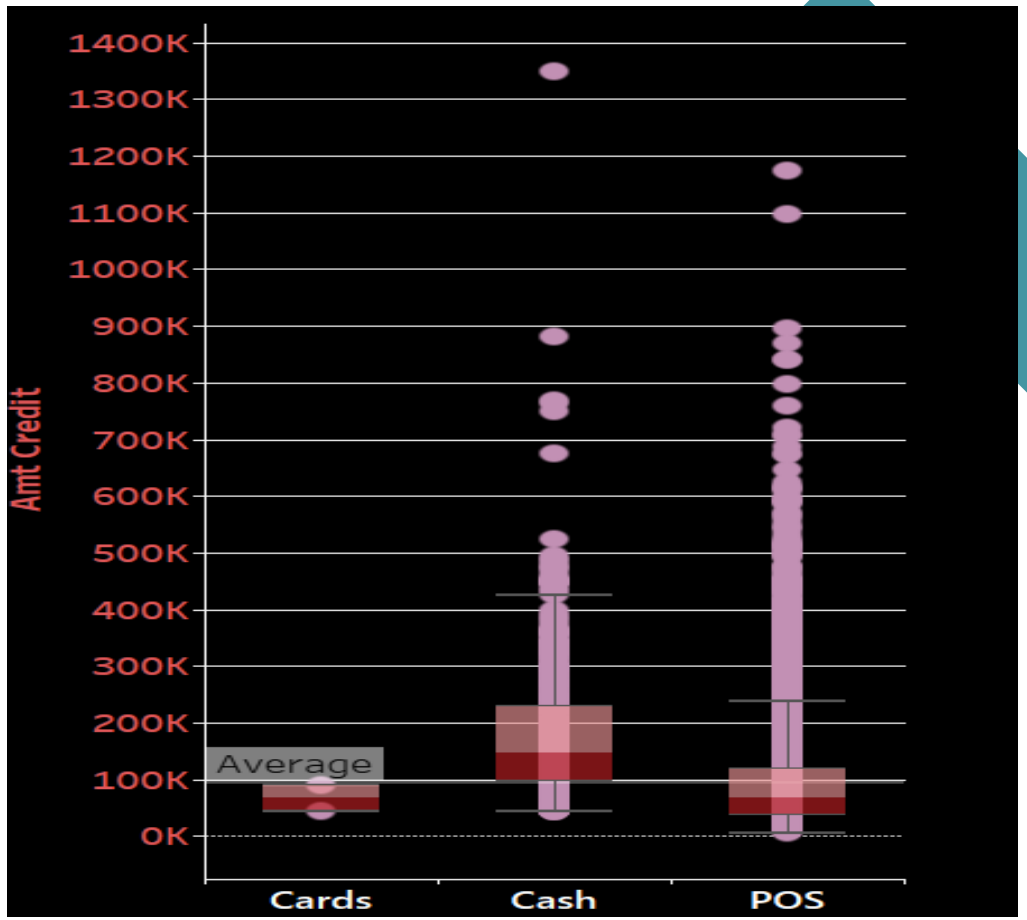


Cash loans have a high rate of Approval,Refusal and Cancellation

Outliers are between the 3rd and 4th quartile for loan amount between 2 and 10 Lakhs

Bivariate Analysis

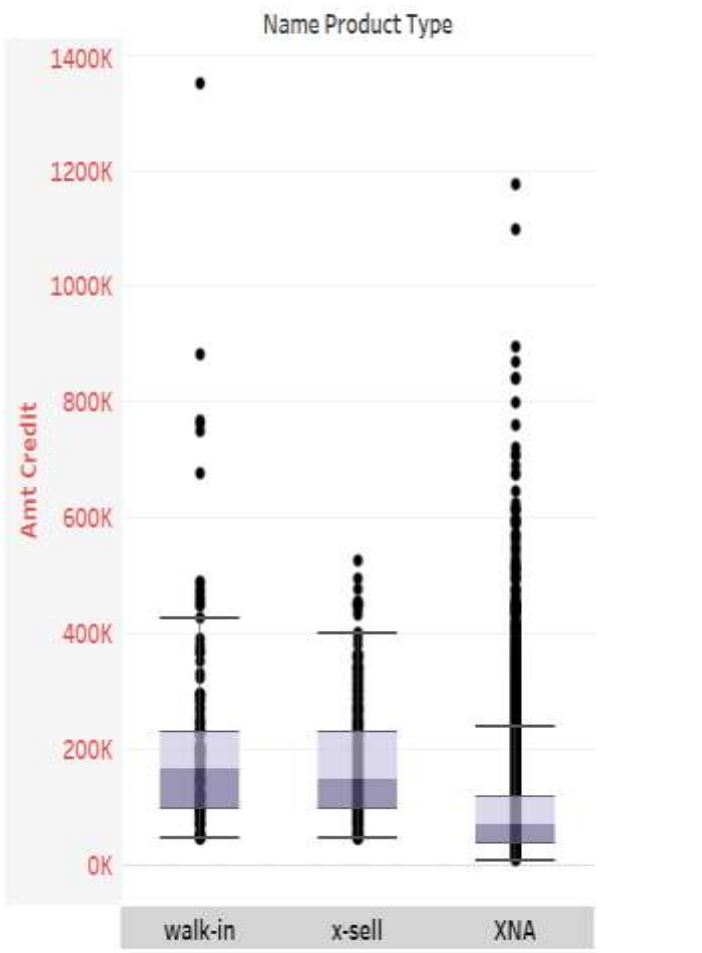




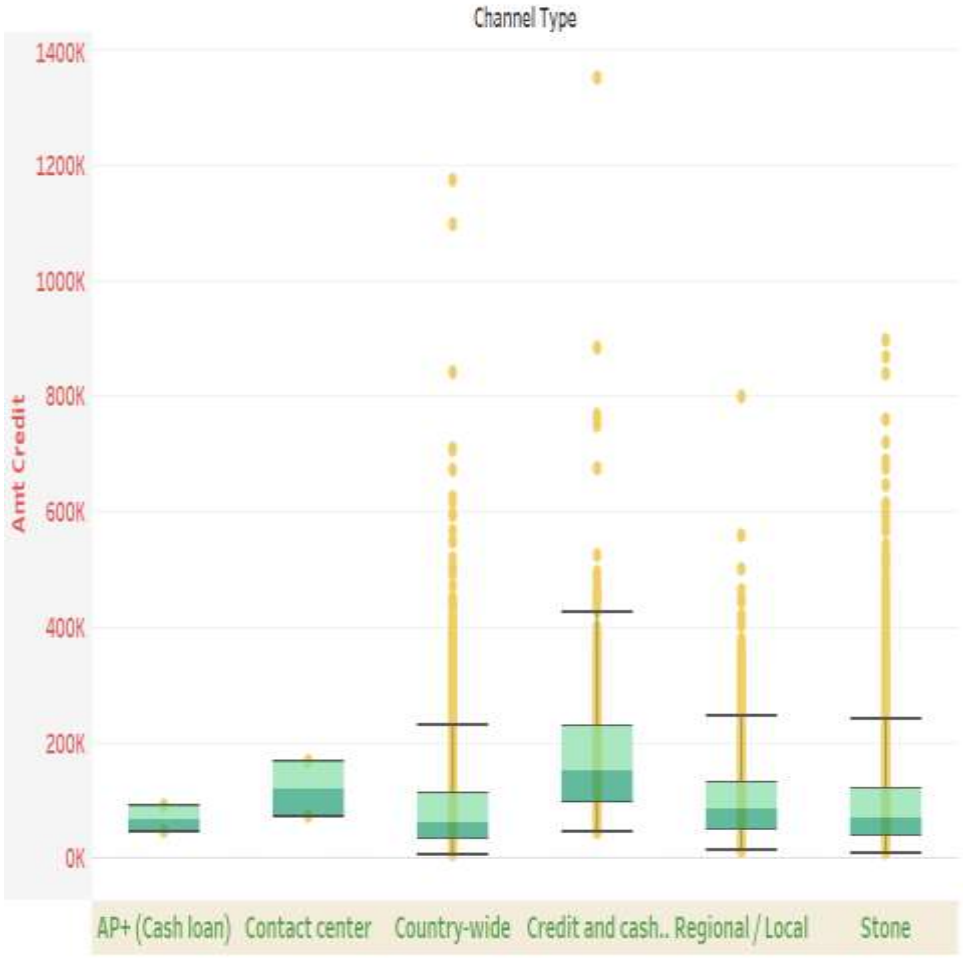
Freshers and Repeaters have applied for most Loans

Portfolio Loans have a higher rate than Cash or Card

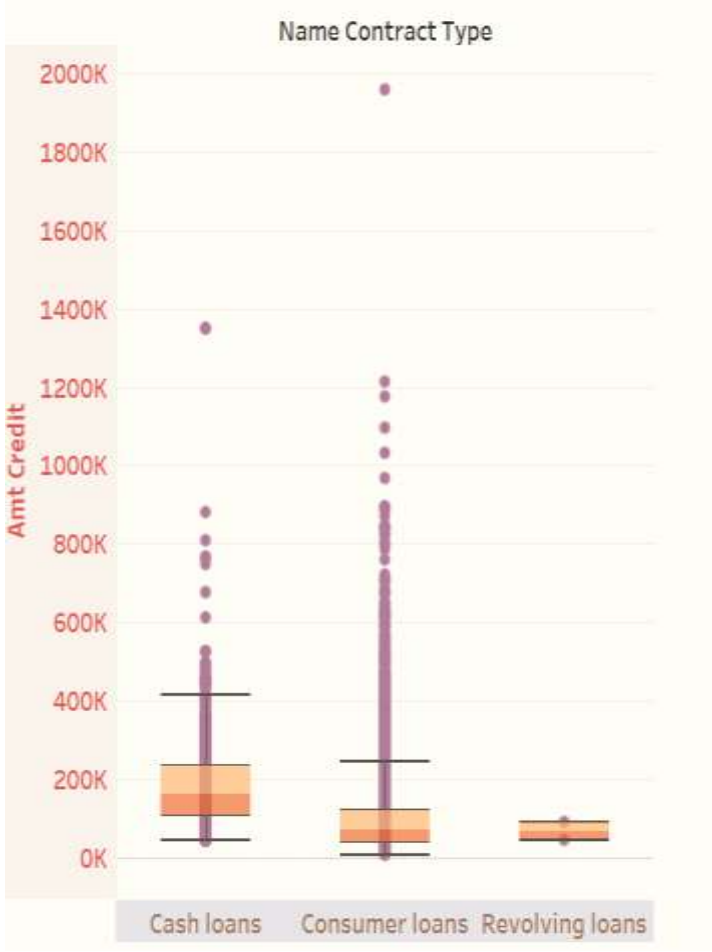
Bivariate Analysis



Amount of credit for approved products



Credit based on Channel Type

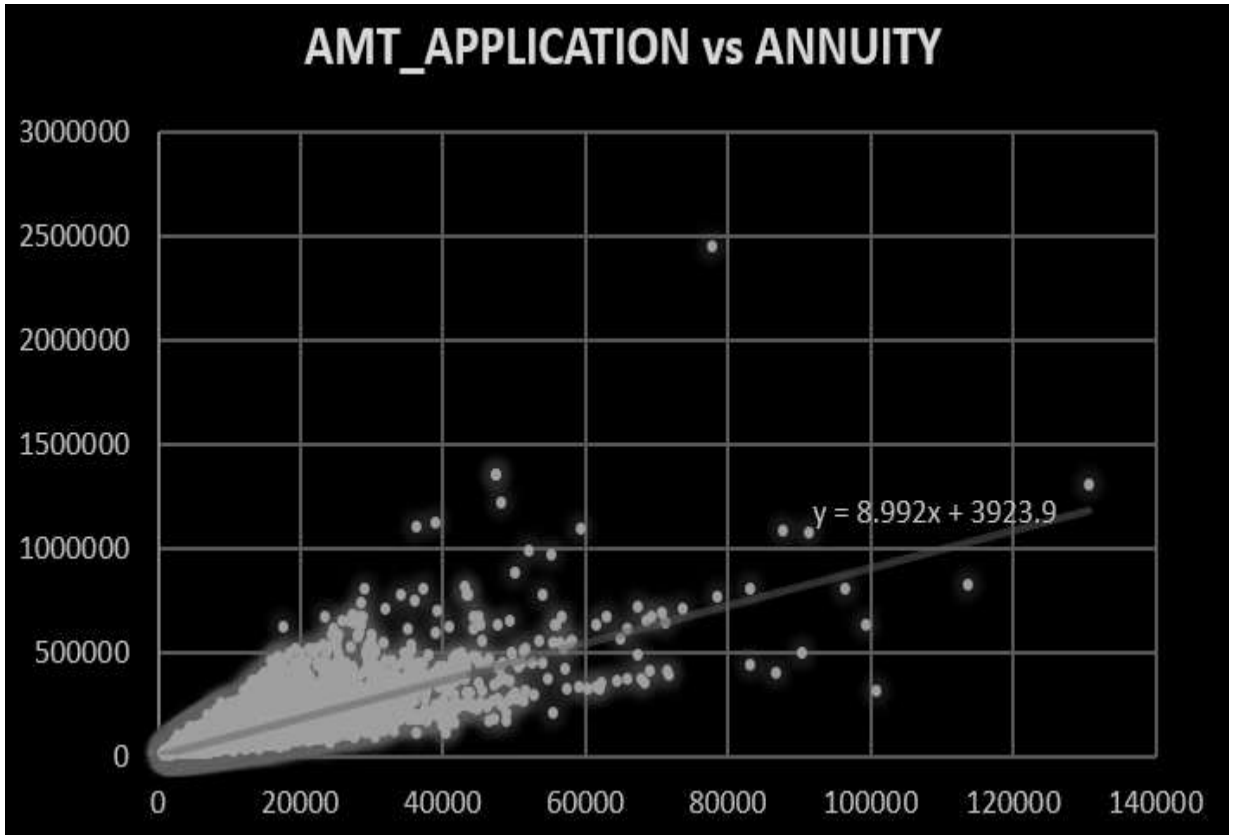


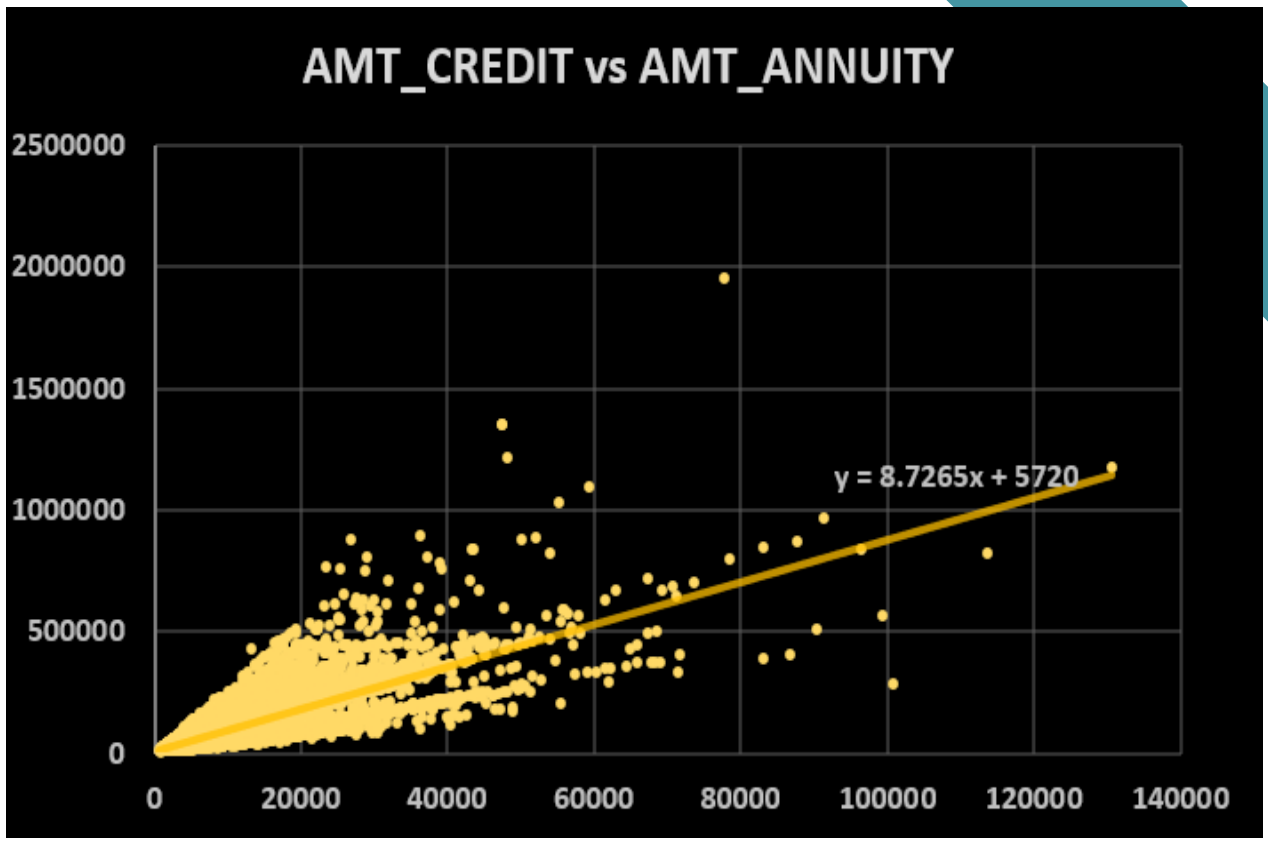
Credit based on Loan Contracts

Correlation matrix

Column1	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	RATE_DOWN_PAYMENT
AMT_ANNUITY	1					
AMT_APPLICATION	0.79628125	1				
AMT_CREDIT	0.78067728	0.980370344	1			
AMT_DOWN_PAYMENT	0.296790995	0.38770408	0.220581324	1		
AMT_GOODS_PRICE	0.79628125	1	0.980370344	0.38770408	1	
RATE_DOWN_PAYMENT	-0.097924973	-0.099787541	-0.224319462	0.617272836	-0.099787541	1

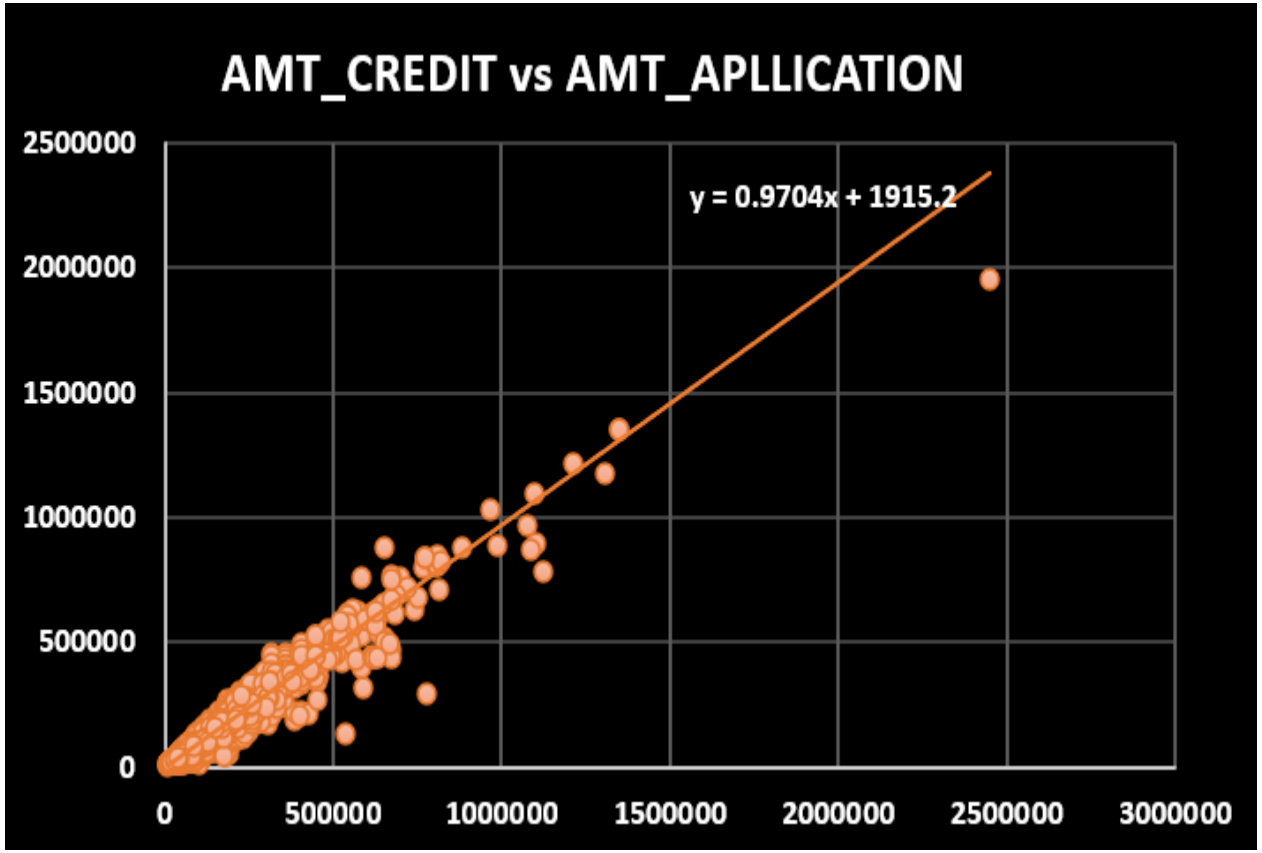
Correlation

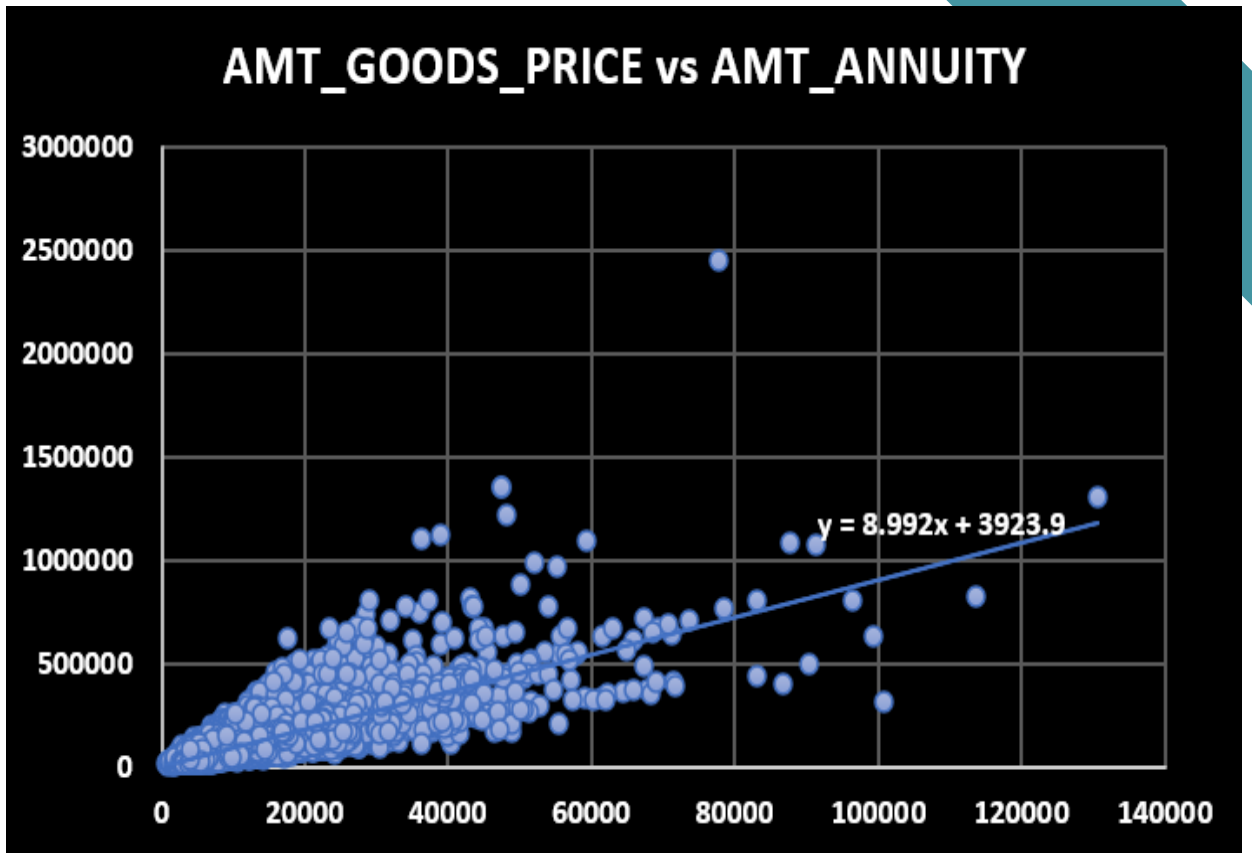




Amt_Annuity has a strong correlation of 0.79 with Credit and Amt_Application, so the X-axis increases with the Y-axis. Straight trendline indicates performance

Correlation

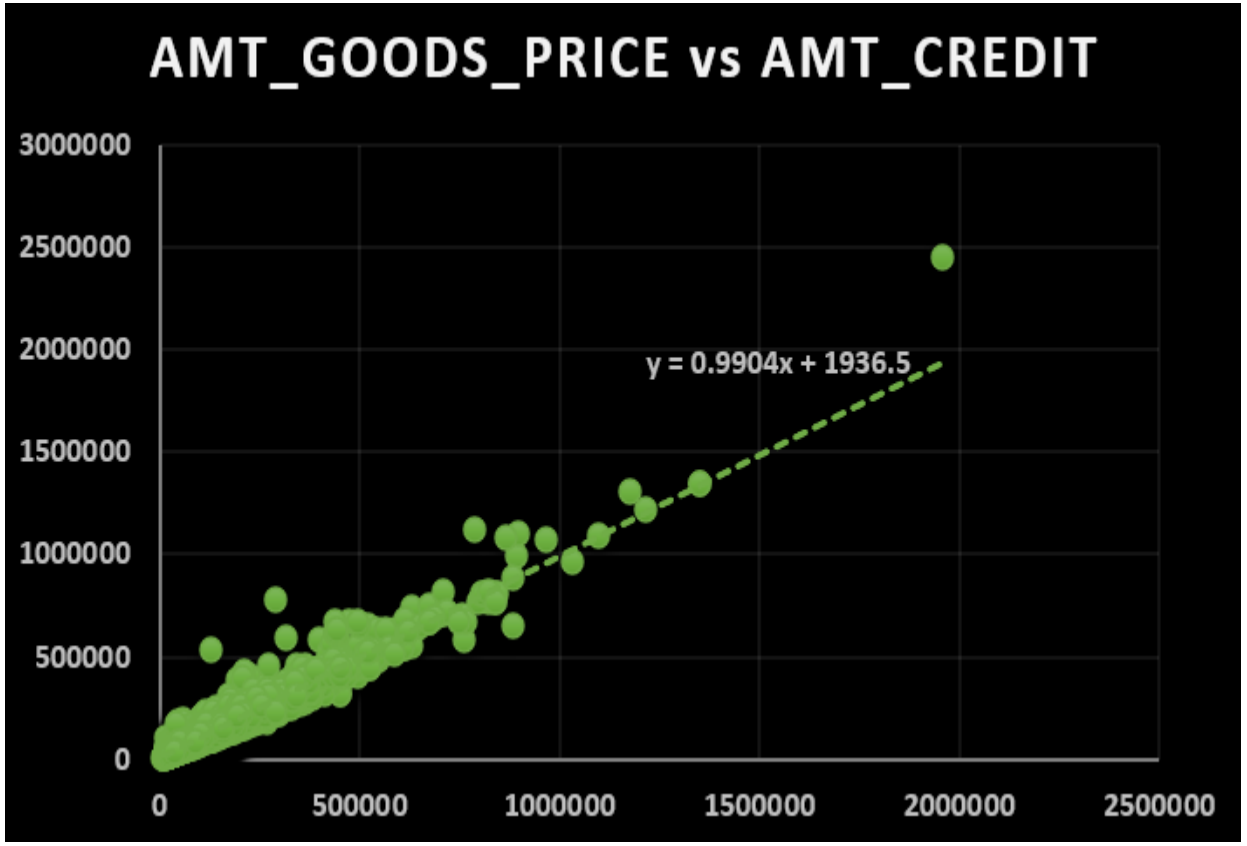


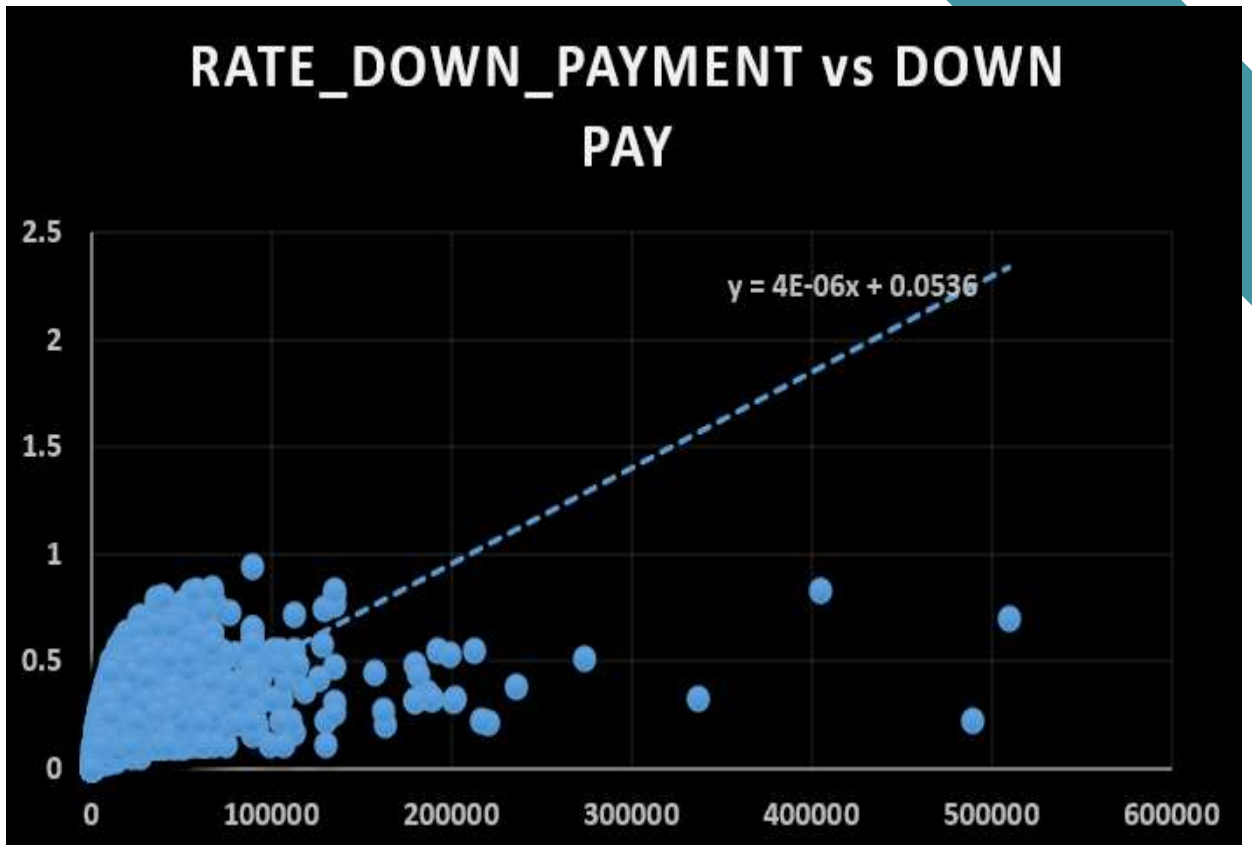


Strong correlation of 0.98 with Credit and Application, 0.79 with Goods_Price and Annuity.
So the X-axis increases with the Y-axis. Straight trendline indicates performance

Closer the coefficient to 1, more the slope of trendline

Correlation





Strong correlation of 0.98 with Credit and Goods_Price. Medium correlation of 0.61 with Down_Payment and Rate of Down_Payment. So the X-axis increases with the Y-axis. Straight trendline indicates performance

Closer the coefficient to 1, more the slope of trendline

Hypothesis

1. **Demographic Insights:**
   -The majority of the population is female.
   - Most people own a car.
   - Majority own real estate.
   - Married individuals apply for the most loans.
   - A family of 2 tends to pay off the majority of their loans.

2. Loan Repayment
   - Most loans have been paid back, especially by females.
   - Those who own real estate but not a car have a higher tendency to pay back their loans.
   - Business entity and university workers have a high rate of loan repayment.

3. Loan Application Insights
   - Consumer loans were the most common type of loan applied for previously.
   - The working class applies for the most loans, with a higher proportion of unpaid loans.
   - Academic degree holders opt for loans the least.

Hypothesis

4.Loan Processing and Types:

- Cash loans are the most popular contract type.
- Loan rejections are predominantly for XAP.
- Loans are mostly approved, with a high yield and paid through banks.

5.Sector-Specific Insights:

- The secondary/secondary special education sector has taken the most loans.
- Repeat clients with POS loan portfolios are common.

6.Geographical and Institutional Insights:

- Loans are predominantly received country-wide.

7.Financial and Employment Insights:

- Average annuity and income are higher among males.
- Labourers and homeowners (house/apartment) have the highest number of loan applications, mostly paid back.

Overall Review

- Most people whose current loan is greater than previous loan are defaulters(credit difference compared by joining tables in Tableau)
- More income, more credit
- Average income of males > females
- Females, Businessmen and Academic Degree holders are most capable of paying back their loans
- Previous applicants preferred Consumer Loans compared to Current Applicants who prefer Cash Loans
- Rate of approval for Freshers and Refreshers is high
- Portfolio loans with walk-in products have high credit
- Widows and people who live with parents are incapable of paying loans
- Family of 2 members are most efficient performers
- Consumer loans have high yields
- Unemployed and clients on Maternity leaves have applied for least loans

Result

- The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. I have learnt that 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. Through Exploratory Data Analysis we can understand the key factors behind loan default to make better decisions about loan approval.

Thank you

Shivaani Dushyanth
shivzlantern09@gmail.com