Bank Loan Case Study

Project Description:

This case study aims to give us an idea of applying EDA (Exploratory Data Analysis) in a real business scenario. In this case study, we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding:

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.

Role: we are working in a finance company which lends various types of loans to the customers.

Things to handle:

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company:

1. Approved: The company has approved loan application
2. Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
3. Refused: The company had rejected the loan (because the client does not meet their requirements etc.).
4. Unused Offer: Loan has been cancelled by the client but on different stages of the process.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Things we have to find out through the case study:

- Identification of such default applicants using EDA is the aim of this case study**.**

  In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.
- Cleaning the dataset, finding the outliers, data imbalance, univariate, bivariate analysis and the correlation for the client with the payment difficulty.

Approach:

Problem Statement:

The aim is to **identify patterns** which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. **Identification of such applicants using EDA** is the aim of this case study.

Analysis Approach:

- Downloading the dataset (Application_data & Previous_Application)
- Identifying the missing values and dealing with it.
- Identification of the Outliers in the data.
- Understanding the Ratio of Data imbalance in the given dataset.
- Finding out the correlation between client and target variables.
- Finally, Data Visualization with the help of Charts, Graphs.

Tech Stack Used:

I used

- **Microsoft® Excel® 2019 MSO (Version 2304 Build 16.0.16327.20200) 64-bit** which enables us to Clean, Format, Organize and Calculate the data in a spreadsheet.
- **Ms Word 2019** for the preparation of the document to be presented.

Insights:

Based on the achieved results I can conclude the following things:

- Most of the loans are given to applicants are in range 2-3lakhs.
- Most of the Loans are given to those who applied for an amount in the range of 1-1.5lakhs.
- Most of the loans are given to those who has income above 5 lakhs.
- Married people are the one's who pay their loans on time.
- Female are having high difficulties paying their loans.
- There is Data imbalance between the data with payments on time and data with payment difficulties.

Results: The detailed answers to the questions are below:

Task1: Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

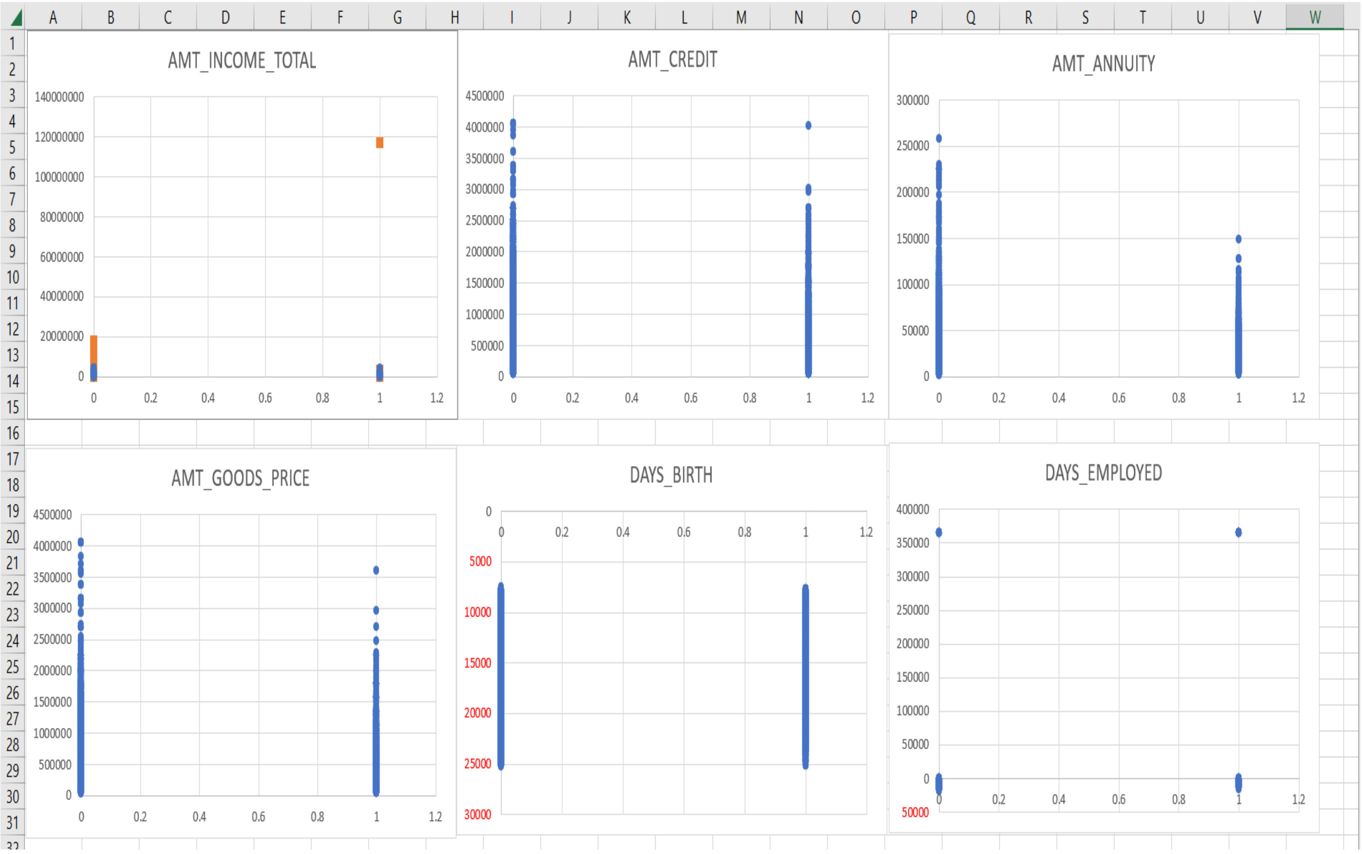
- I have used COUNTA function to find the total rows of each column.
- Then, I found the percentage of null values using the formula (Row counts of each column/Total no. of rows)
- Then, I dropped off all the columns with missing data greater than 40%.
- For columns with missing data less than 40%, I imputed with the median values of that column(numerical data),with mode values for categorical data.

- Converted all the negative values into positive values of columns like DAYS_ID_PUBLISH, DAYS_REGISTRATION, DAYS_DECISION, DAYS_FIRST_DUE, DAYS_LAST_DUE, DAYS_TERMINATION etc..
- Dropped all the columns which are not used for analysis.

Task2: Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier.

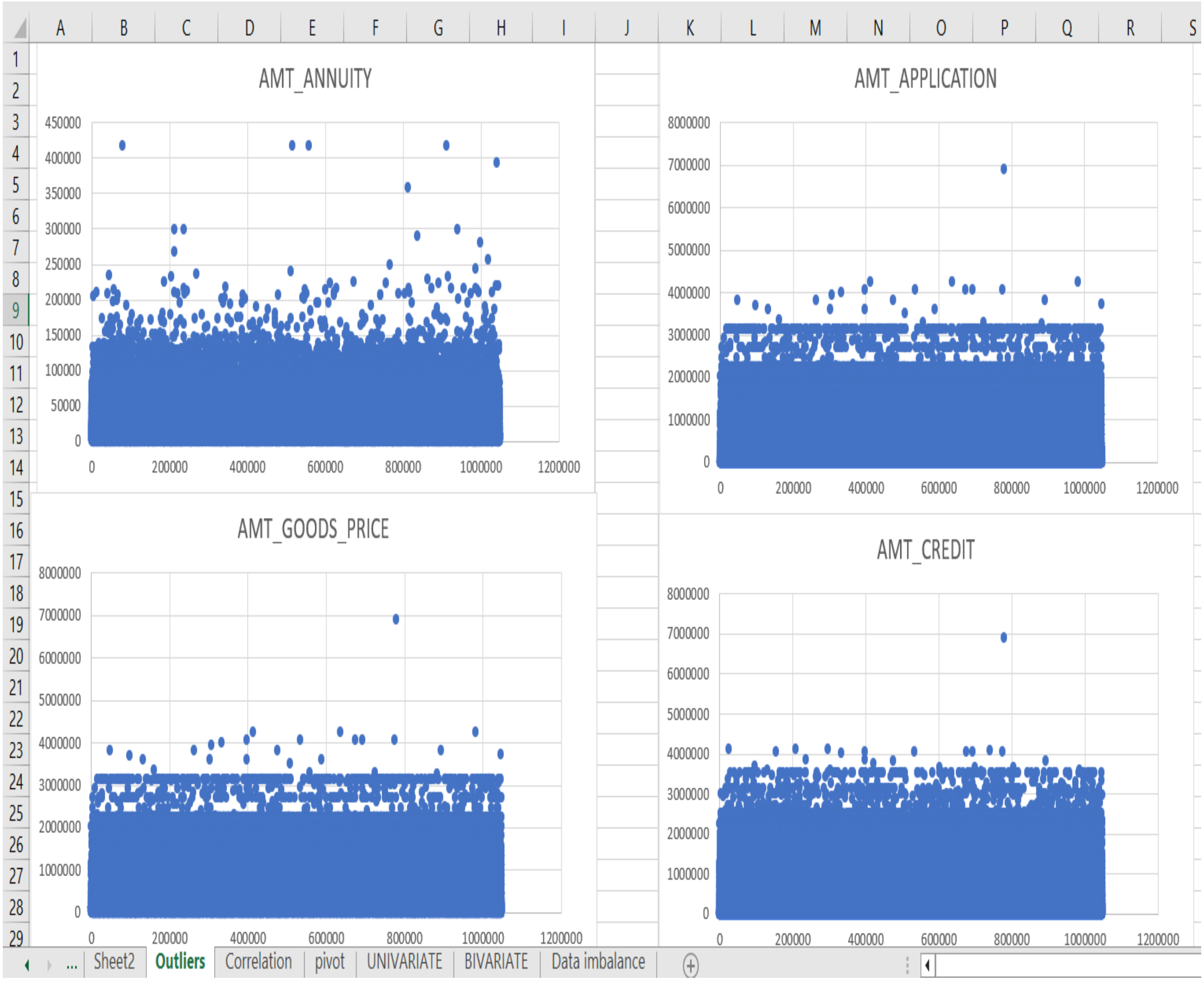
Application_data:

- There are some outliers in the AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE.
- There is an outlier in AMT_INCOME_TOTAL whose income is far greater than remaining (11cr).
- There are no outliers in DAYS_BIRTH which is reliable for analysis.



Previous_application :

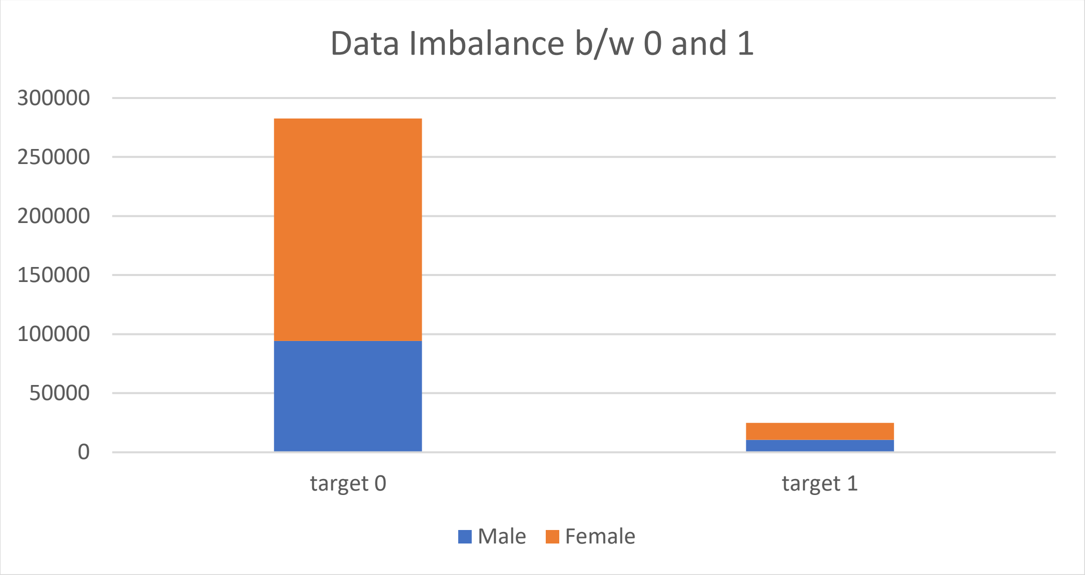
- There are huge outliers in AMT_ANNUITY column data.
- There are some outliers in AMT_APPLICATION, AMT_GOODS_PRICE, AMT_CREDIT columns data.

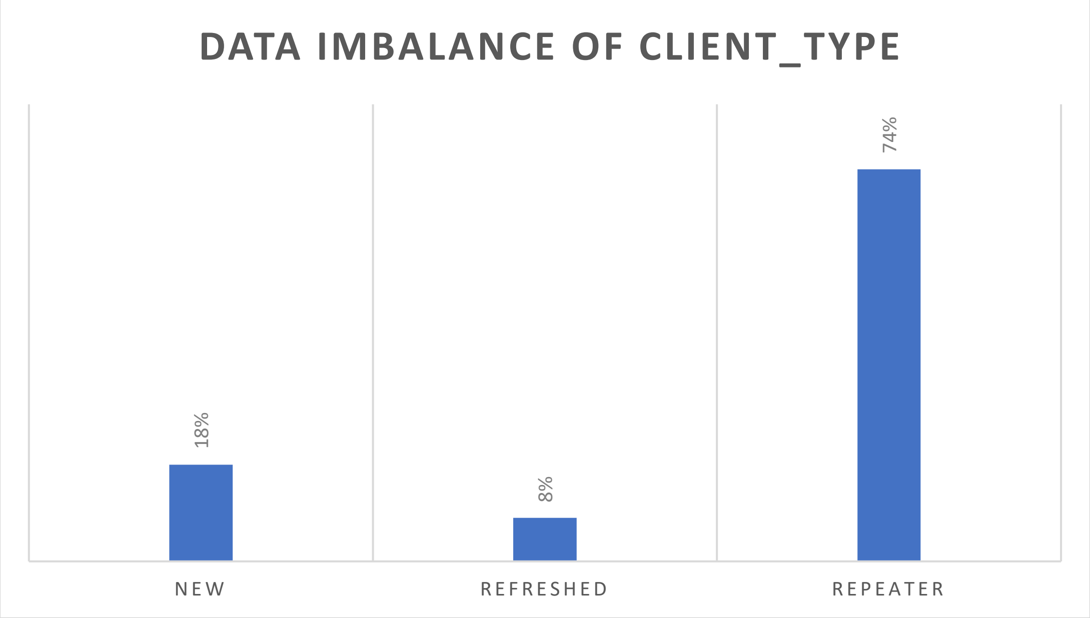


Task3: Identify if there is data imbalance in the data. Find the ratio of data imbalance.

In Application_data, There is an data imbalance ratio of 11.39 : 1 between Target 0 (Clients who pay on time) and Target 1 (Clients with payment difficulties).

In Previous_application, There is a data imbalance between different client types as Repeaters data is 74%, new applicants data is 18%, Refreshed data is just 8%.

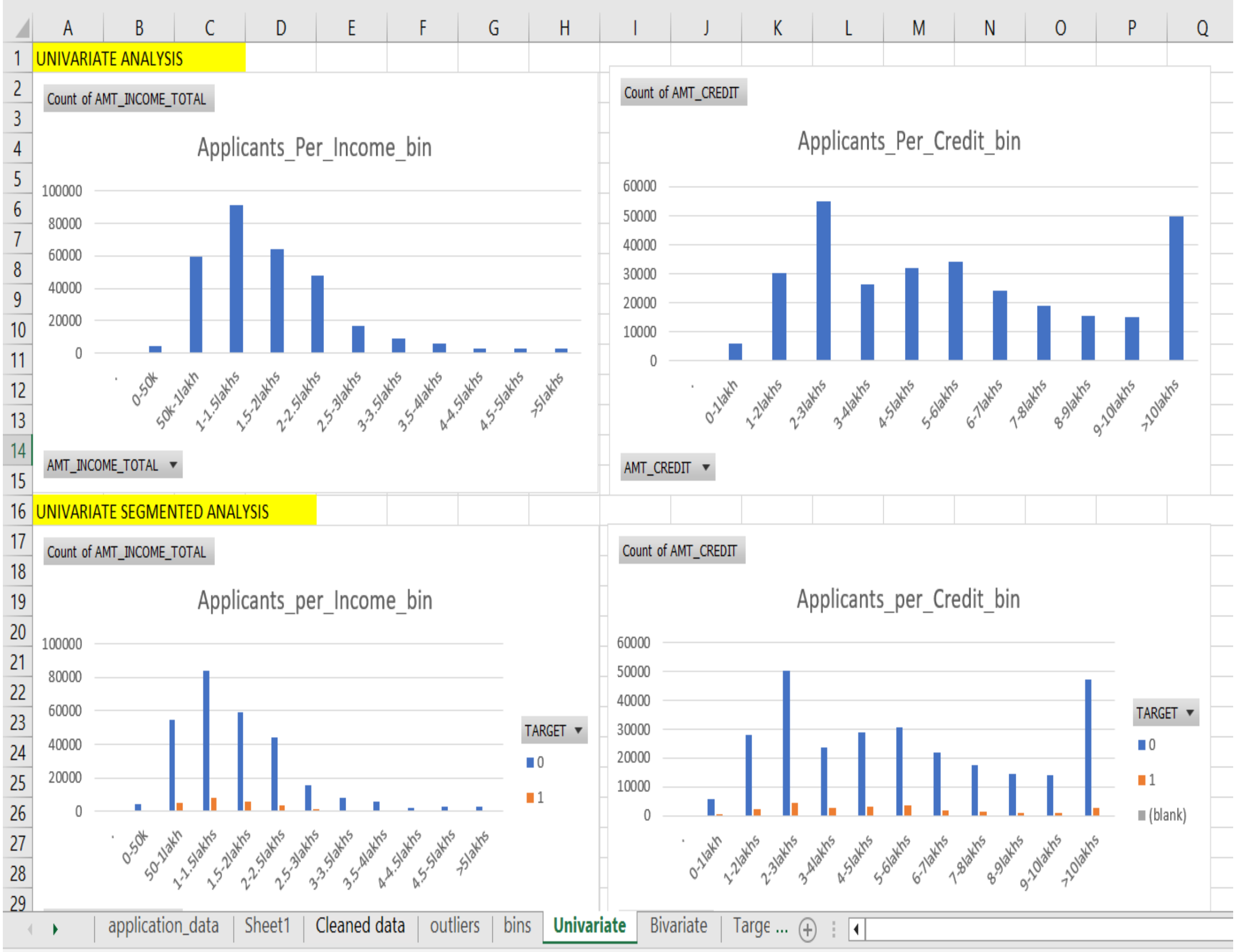


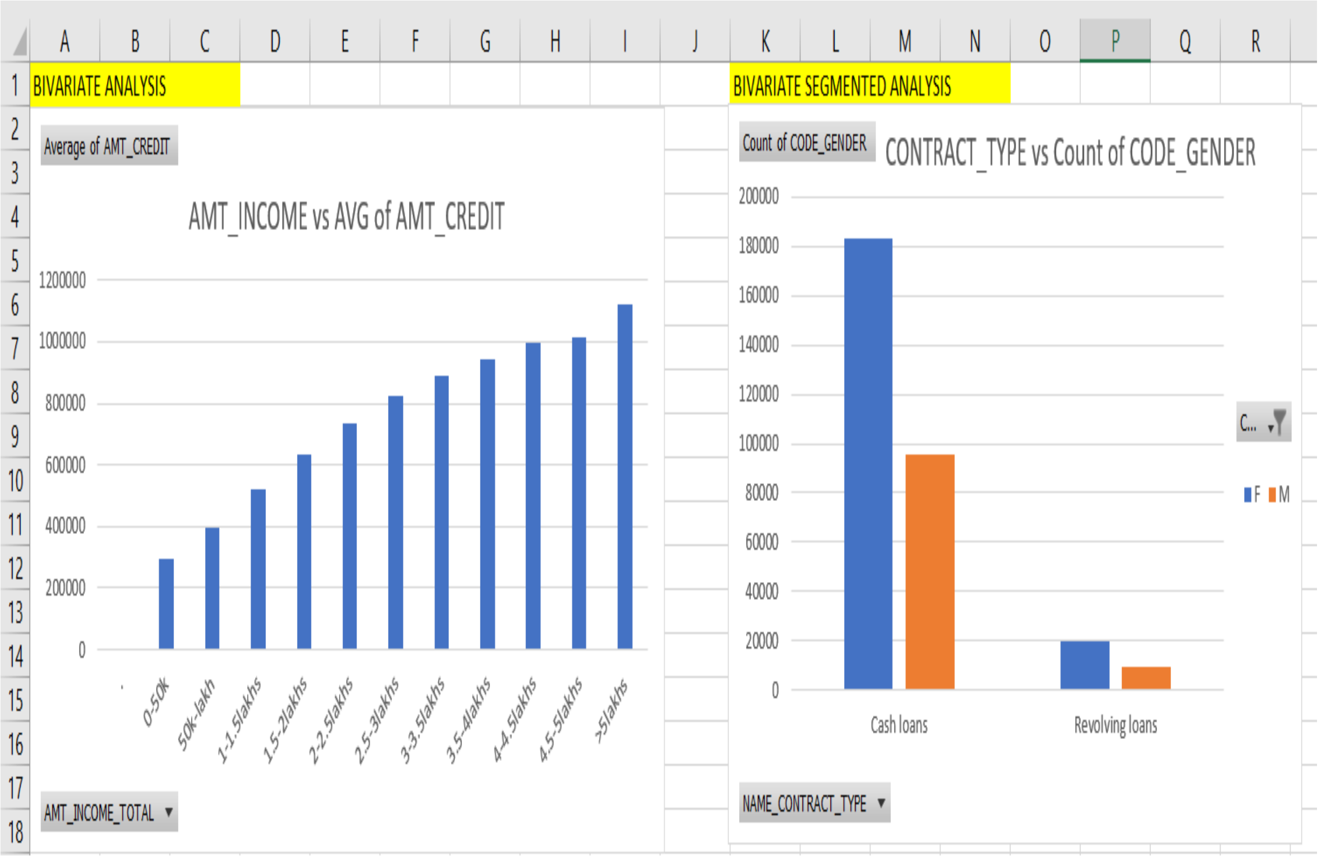


Task4: Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

In Application_data:

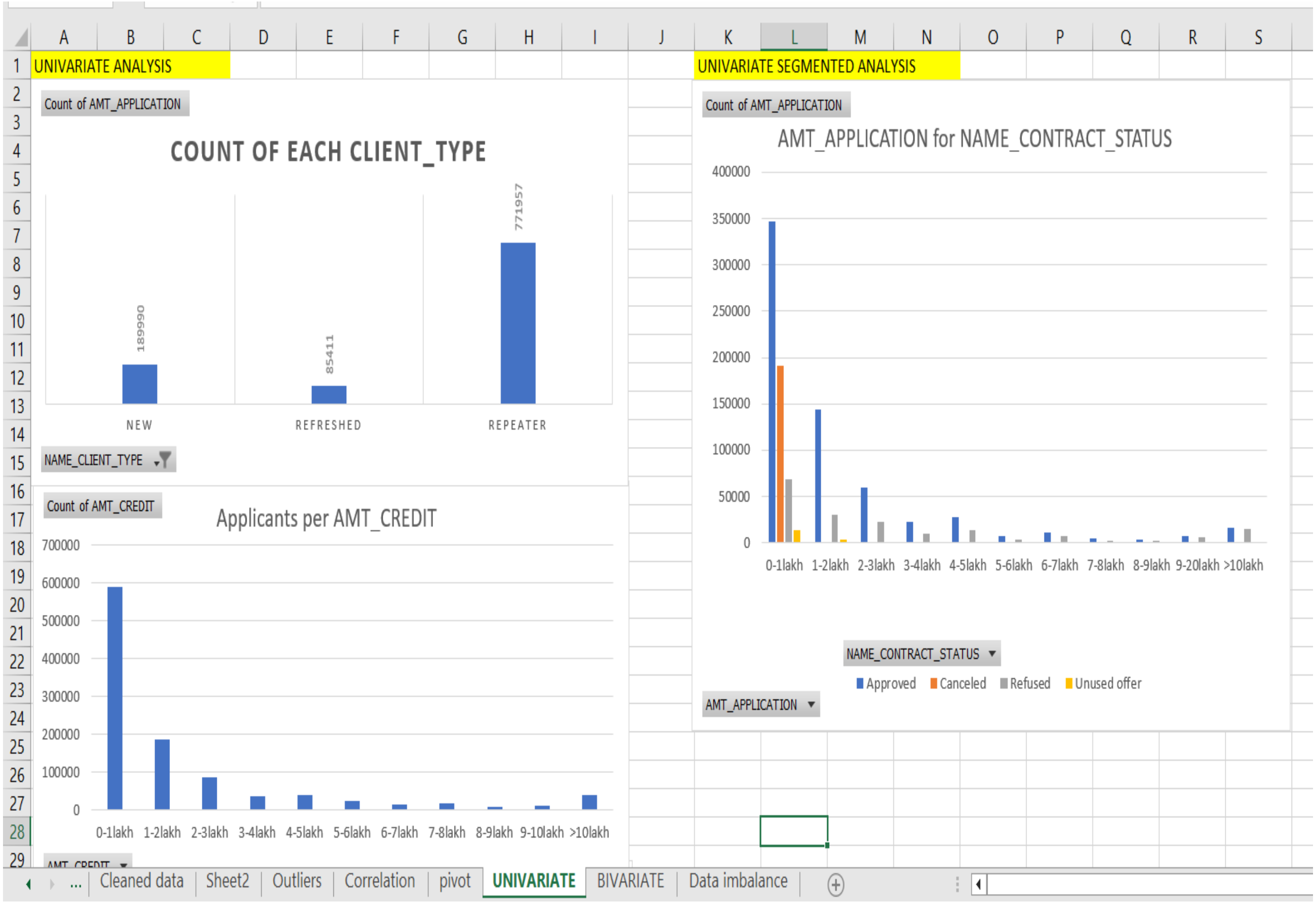
- Most of the people who are applying with AMT_APPLICATION(amount specified while applying for loan) in the range of 1-1.5lakhs.
- Maximum people got the credit amount (loan amount) in the range of 2-3lakhs and next maximum people got the loan above 10 lakhs.
- The loans are mostly given to those whose income(INCOME_TOTAL) is above 5lakhs.
- Maximum loans that are given are Cash loans. Very few Revolving loans.



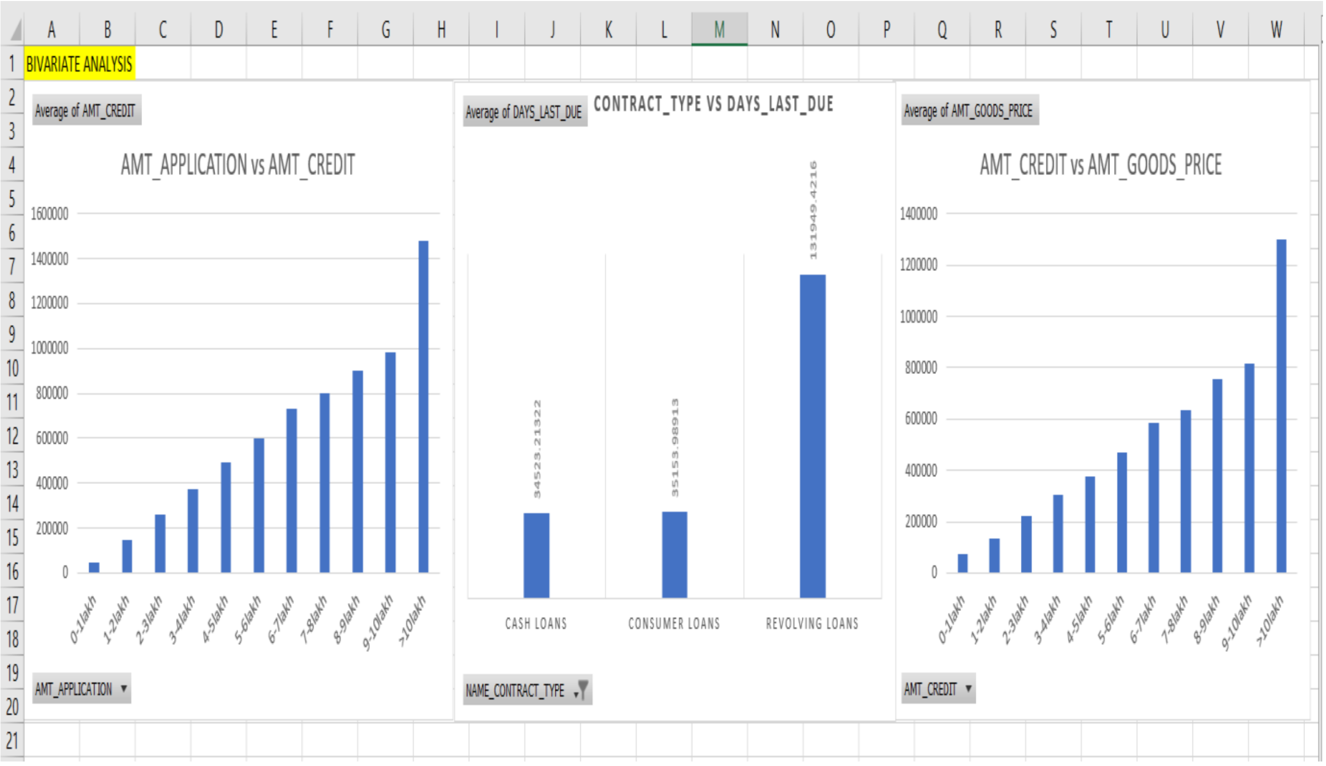


In Previous_application:

- Most of the applicants are repeaters (those who already took the loan and paid).
- The loan amount given to the people is mostly in the range of 0-1lakh.
- The last dues are high in case of revolving loans.
- Those who applied for the amount greater than 10 lakhs has got an average credit amount of 14lakhs.



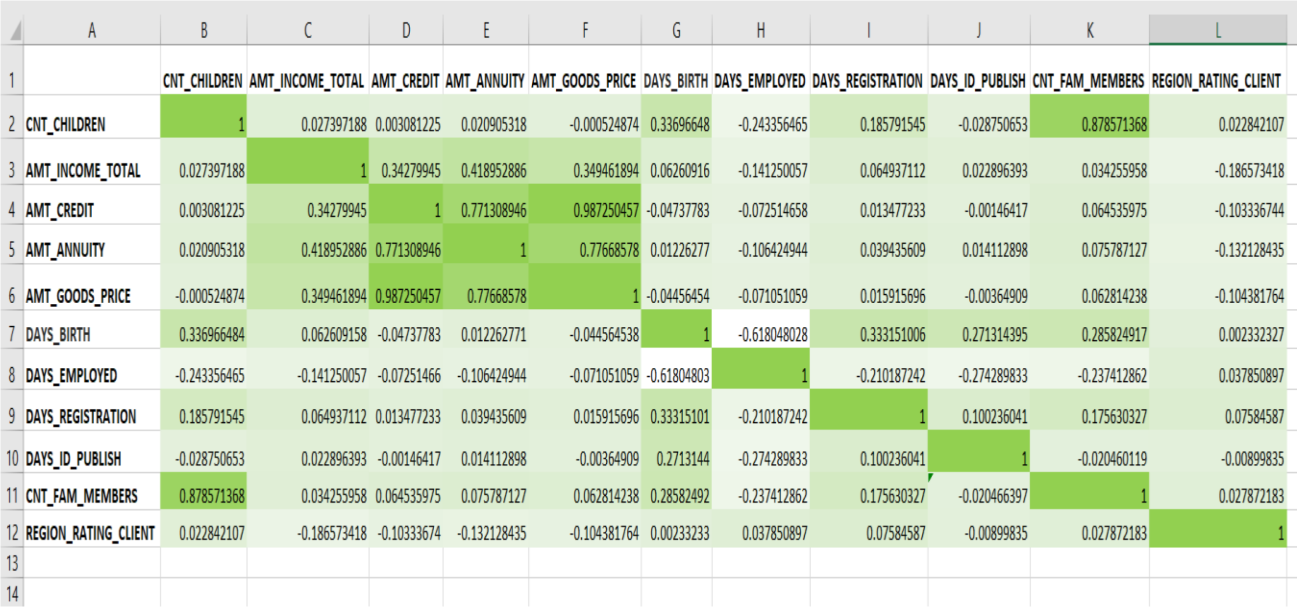
UNIVARIATE ANALYSIS
Count of AMT_APPLICATION
COUNT OF EACH CLIENT_TYPE
189990
85411
771957
NEW
REFRESHED
REPEATER
NAME_CLIENT_TYPE
Count of AMT_CREDIT
Applicants per AMT_CREDIT
700000
600000
500000
400000
300000
200000
100000
0
0-1lakh
1-2lakh
2-3lakh
3-4lakh
4-5lakh
5-6lakh
6-7lakh
7-8lakh
8-9lakh
9-10lakh
>10lakh
AMT_CREDIT
UNIVARIATE SEGMENTED ANALYSIS
Count of AMT_APPLICATION
AMT_APPLICATION for NAME_CONTRACT_STATUS
400000
350000
300000
250000
200000
150000
100000
50000
0
0-1lakh
1-2lakh
2-3lakh
3-4lakh
4-5lakh
5-6lakh
6-7lakh
7-8lakh
8-9lakh
9-20lakh
>10lakh
NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer
AMT_APPLICATION
Cleaned data
Sheet2
Outliers
Correlation
pivot
UNIVARIATE
BIVARIATE
Data imbalance



BIVARIATE ANALYSIS
Average of AMT_CREDIT
AMT_APPLICATION vs AMT_CREDIT
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
0-1lakh
1-2lakh
2-3lakh
3-4lakh
4-5lakh
5-6lakh
6-7lakh
7-8lakh
8-9lakh
9-10lakh
>10lakh
AMT_APPLICATION
Average of DAYS_LAST_DUE
CONTRACT_TYPE VS DAYS_LAST_DUE
34523.21322
35153.98913
131949.4216
CASH LOANS
CONSUMER LOANS
REVOLVING LOANS
NAME_CONTRACT_TYPE
Average of AMT_GOODS_PRICE
AMT_CREDIT vs AMT_GOODS_PRICE
1400000
1200000
1000000
800000
600000
400000
200000
0
0-1lakh
1-2lakh
2-3lakh
3-4lakh
4-5lakh
5-6lakh
6-7lakh
7-8lakh
8-9lakh
9-10lakh
>10lakh
AMT_CREDIT

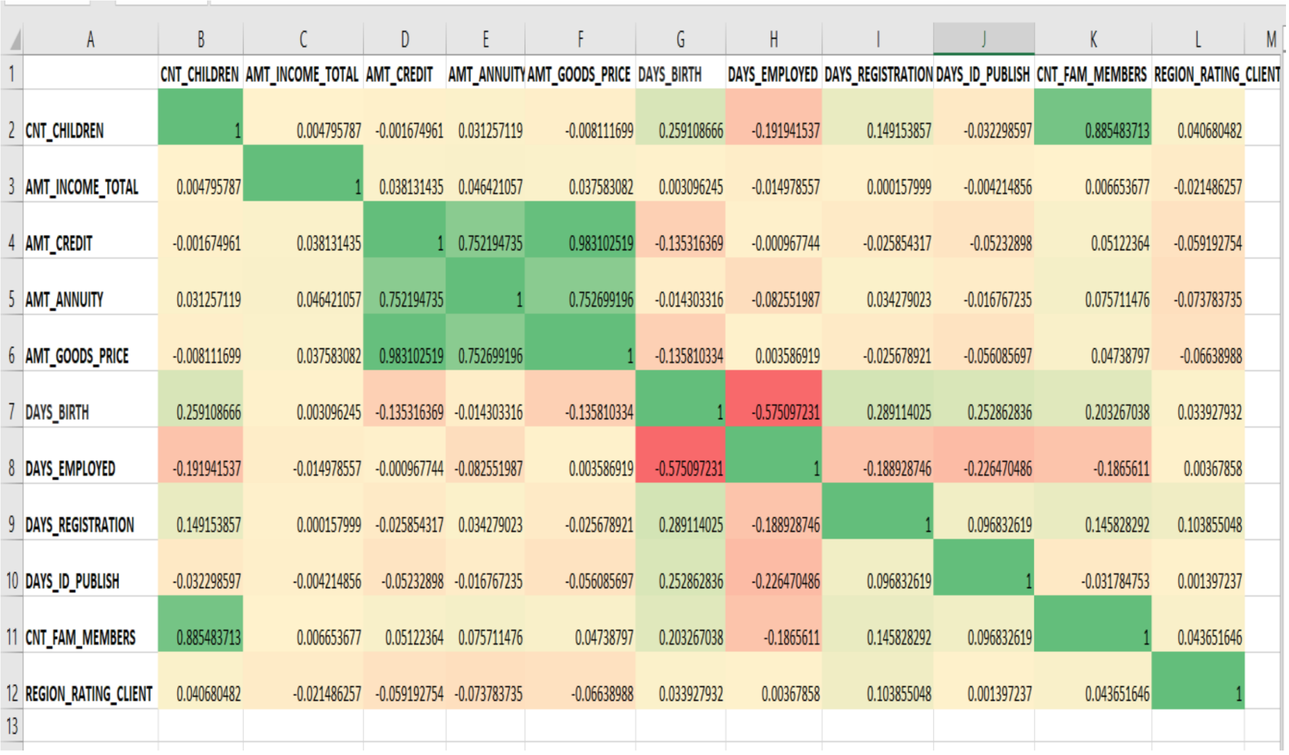
Task5: Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Find if any insight is there. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.

Client with Payments on time:



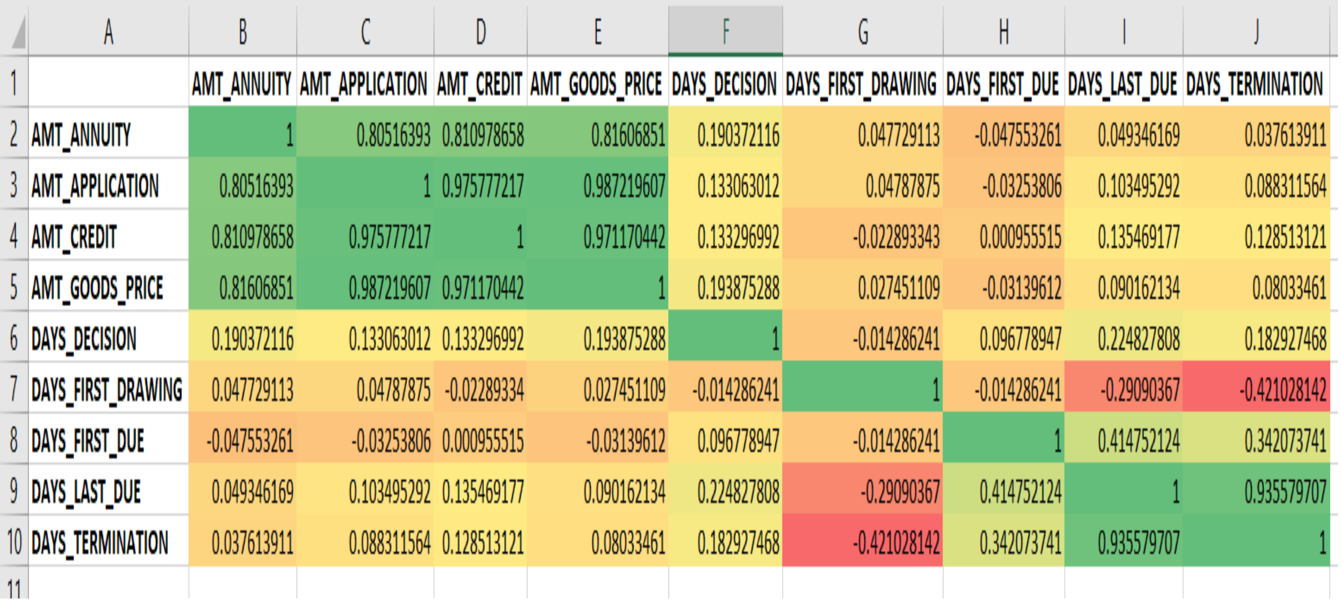
	A	B	C	D	E	F	G	H	I	J	K	L
1		CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	REGION_RATING_CLIENT
2	CNT_CHILDREN	1	0.027397188	0.003081225	0.020905318	-0.000524874	0.33696648	-0.243356465	0.185791545	-0.028750653	0.878571368	0.022842107
3	AMT_INCOME_TOTAL	0.027397188	1	0.34279945	0.418952886	0.349461894	0.06260916	-0.141250057	0.064937112	0.022896393	0.034255958	-0.186573418
4	AMT_CREDIT	0.003081225	0.34279945	1	0.771308946	0.987250457	-0.04737783	-0.072514658	0.013477233	-0.00146417	0.064535975	-0.103336744
5	AMT_ANNUITY	0.020905318	0.418952886	0.771308946	1	0.77668578	0.01226277	-0.106424944	0.039435609	0.014112898	0.075787127	-0.132128435
6	AMT_GOODS_PRICE	-0.000524874	0.349461894	0.987250457	0.77668578	1	-0.04456454	-0.071051059	0.015915696	-0.00364909	0.062814238	-0.104381764
7	DAYS_BIRTH	0.336966484	0.062609158	-0.04737783	0.012262771	-0.044564538	1	-0.618048028	0.333151006	0.271314395	0.285824917	0.002332327
8	DAYS_EMPLOYED	-0.243356465	-0.141250057	-0.07251466	-0.106424944	-0.071051059	-0.61804803	1	-0.210187242	-0.274289833	-0.237412862	0.037850897
9	DAYS_REGISTRATION	0.185791545	0.064937112	0.013477233	0.039435609	0.015915696	0.33315101	-0.210187242	1	0.100236041	0.175630327	0.07584587
10	DAYS_ID_PUBLISH	-0.028750653	0.022896393	-0.00146417	0.014112898	-0.00364909	0.2713144	-0.274289833	0.100236041	1	-0.020460119	-0.00899835
11	CNT_FAM_MEMBERS	0.878571368	0.034255958	0.064535975	0.075787127	0.062814238	0.28582492	-0.237412862	0.175630327	-0.020466397	1	0.027872183
12	REGION_RATING_CLIENT	0.022842107	-0.186573418	-0.10333674	-0.132128435	-0.104381764	0.00233233	0.037850897	0.07584587	-0.00899835	0.027872183	1
13												
14												

Clients with payment difficulties:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	REGION_RATING_CLIENT	
2	CNT_CHILDREN	1	0.004795787	-0.001674961	0.031257119	-0.008111699	0.259108666	-0.191941537	0.149153857	-0.032298597	0.885483713	0.040680482	
3	AMT_INCOME_TOTAL	0.004795787	1	0.038131435	0.046421057	0.037583082	0.003096245	-0.014978557	0.000157999	-0.004214856	0.006653677	-0.021486257	
4	AMT_CREDIT	-0.001674961	0.038131435	1	0.752194735	0.983102519	-0.135316369	-0.000967744	-0.025854317	-0.05232898	0.05122364	-0.059192754	
5	AMT_ANNUITY	0.031257119	0.046421057	0.752194735	1	0.752699196	-0.014303316	-0.082551987	0.034279023	-0.016767235	0.075711476	-0.073783735	
6	AMT_GOODS_PRICE	-0.008111699	0.037583082	0.983102519	0.752699196	1	-0.135810334	0.003586919	-0.025678921	-0.056085697	0.04738797	-0.06638988	
7	DAYS_BIRTH	0.259108666	0.003096245	-0.135316369	-0.014303316	-0.135810334	1	-0.575097231	0.289114025	0.252862836	0.203267038	0.033927932	
8	DAYS_EMPLOYED	-0.191941537	-0.014978557	-0.000967744	-0.082551987	0.003586919	-0.575097231	1	-0.188928746	-0.226470486	-0.1865611	0.00367858	
9	DAYS_REGISTRATION	0.149153857	0.000157999	-0.025854317	0.034279023	-0.025678921	0.289114025	-0.188928746	1	0.096832619	0.145828292	0.103855048	
10	DAYS_ID_PUBLISH	-0.032298597	-0.004214856	-0.05232898	-0.016767235	-0.056085697	0.252862836	-0.226470486	0.096832619	1	-0.031784753	0.001397237	
11	CNT_FAM_MEMBERS	0.885483713	0.006653677	0.05122364	0.075711476	0.04738797	0.203267038	-0.1865611	0.145828292	0.096832619	1	0.043651646	
12	REGION_RATING_CLIENT	0.040680482	-0.021486257	-0.059192754	-0.073783735	-0.06638988	0.033927932	0.00367858	0.103855048	0.001397237	0.043651646	1	
13													

Previous_application correlation:



	A	B	C	D	E	F	G	H	I	J
1		AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	DAYS_DECISION	DAYS_FIRST_DRAWING	DAYS_FIRST_DUE	DAYS_LAST_DUE	DAYS_TERMINATION
2	AMT_ANNUITY	1	0.80516393	0.810978658	0.81606851	0.190372116	0.047729113	-0.047553261	0.049346169	0.037613911
3	AMT_APPLICATION	0.80516393	1	0.975777217	0.987219607	0.133063012	0.04787875	-0.03253806	0.103495292	0.088311564
4	AMT_CREDIT	0.810978658	0.975777217	1	0.971170442	0.133296992	-0.022893343	0.000955515	0.135469177	0.128513121
5	AMT_GOODS_PRICE	0.81606851	0.987219607	0.971170442	1	0.193875288	0.027451109	-0.03139612	0.090162134	0.08033461
6	DAYS_DECISION	0.190372116	0.133063012	0.133296992	0.193875288	1	-0.014286241	0.096778947	0.224827808	0.182927468
7	DAYS_FIRST_DRAWING	0.047729113	0.04787875	-0.02289334	0.027451109	-0.014286241	1	-0.014286241	-0.29090367	-0.421028142
8	DAYS_FIRST_DUE	-0.047553261	-0.03253806	0.000955515	-0.03139612	0.096778947	-0.014286241	1	0.414752124	0.342073741
9	DAYS_LAST_DUE	0.049346169	0.103495292	0.135469177	0.090162134	0.224827808	-0.29090367	0.414752124	1	0.935579707
10	DAYS_TERMINATION	0.037613911	0.088311564	0.128513121	0.08033461	0.182927468	-0.421028142	0.342073741	0.935579707	1
11										

The Top correlations from all the correlation tables are:

From Target 1:

AMT_CREDIT and AMT_GOODS_PRICE	0.983102519
CNT_FAM_MEMBERS and CNT_CHILDREN	0.885483713
AMT_ANNUITY and AMT_GOODS_PRICE	0.752699196
AMT_CREDIT Aand AMT_ANNUITY	0.752194735
DAYS_BIRTH and DAYS_REGISTRATION	0.289114025

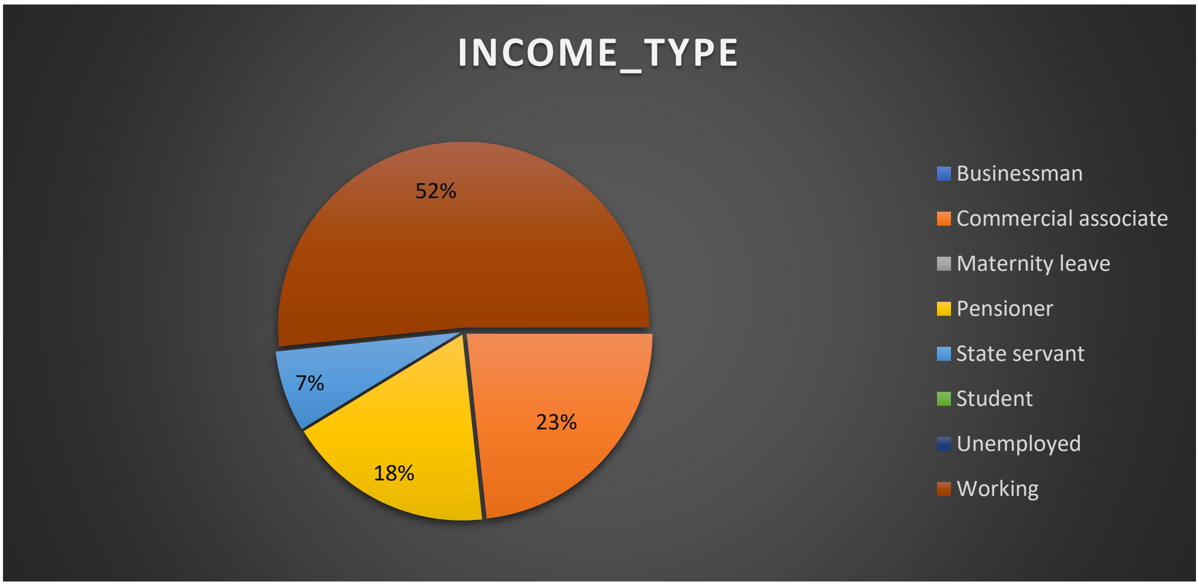
From Target 2:

AMT_CREDIT and AMT_GOODS_PRICE	0.98725046
CNT_CHILDREN and CNT_FAM_MEMBERS	0.87857137
AMT_GOODS_PRICE and AMT_ANNUITY	0.77668578
AMT_CREDIT and AMT_ANNUITY	0.77130895
AMT_INCOME_TOTAL and AMT_ANNUITY	0.41895289

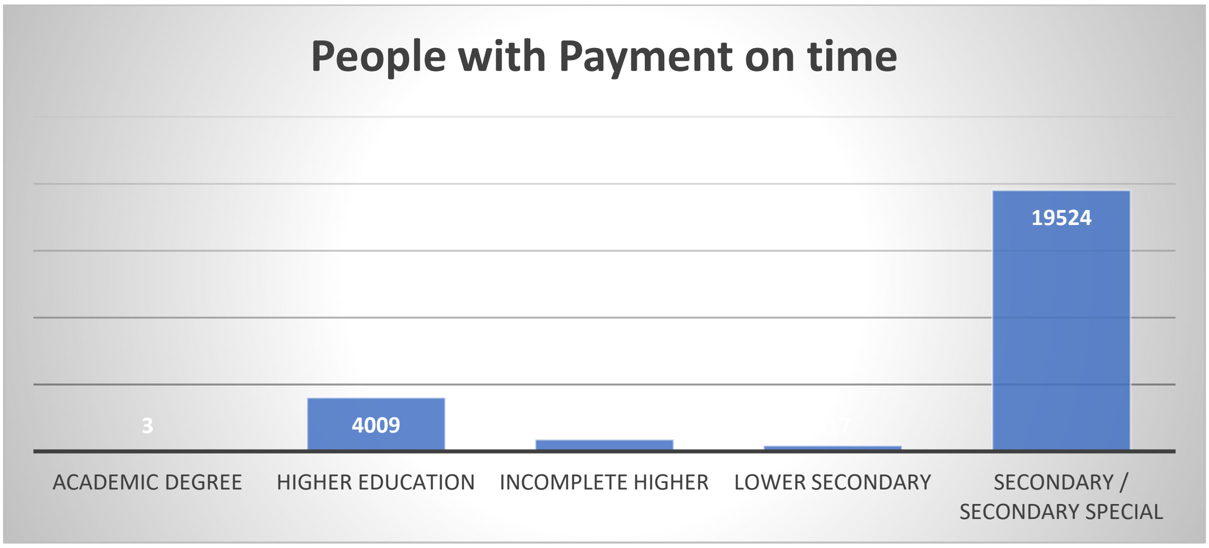
From Previous_application:

AMT_APPLICATION & AMT_GOODS_PRICE	0.987219607
AMT_APPLICATION & AMT_CREDIT	0.975777217
AMT_CREDIT & AMT_GOODS_PRICE	0.971170442
AMT_ANNUITY & AMT_GOODS_PRICE	0.81606851
AMT_CREDIT & AMT_ANNUITY	0.810978658

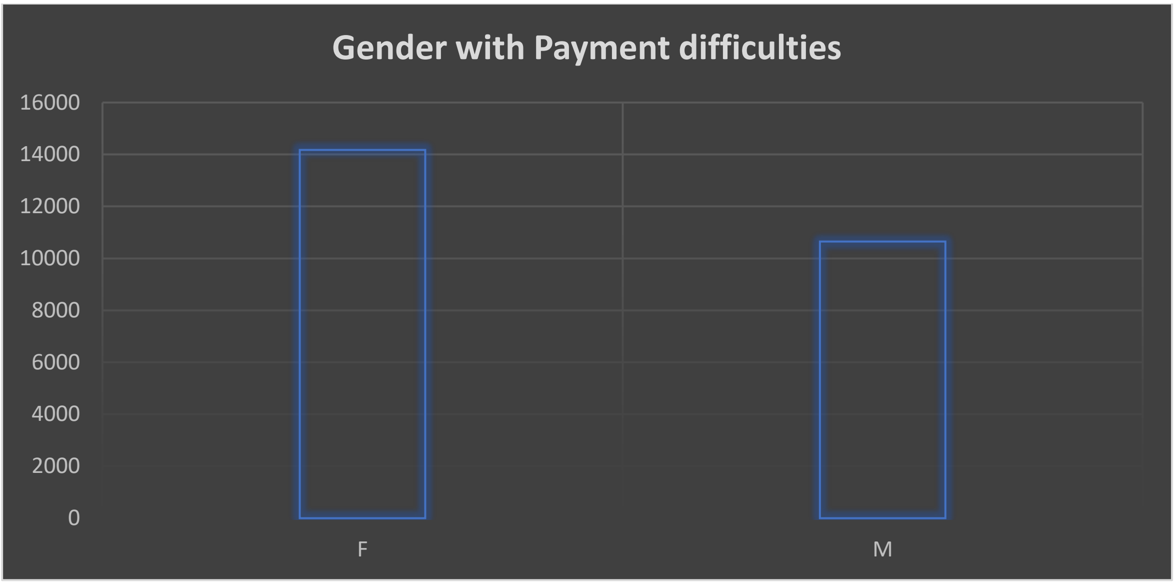
Task6: Include visualizations and summarize the most important results in the presentation.



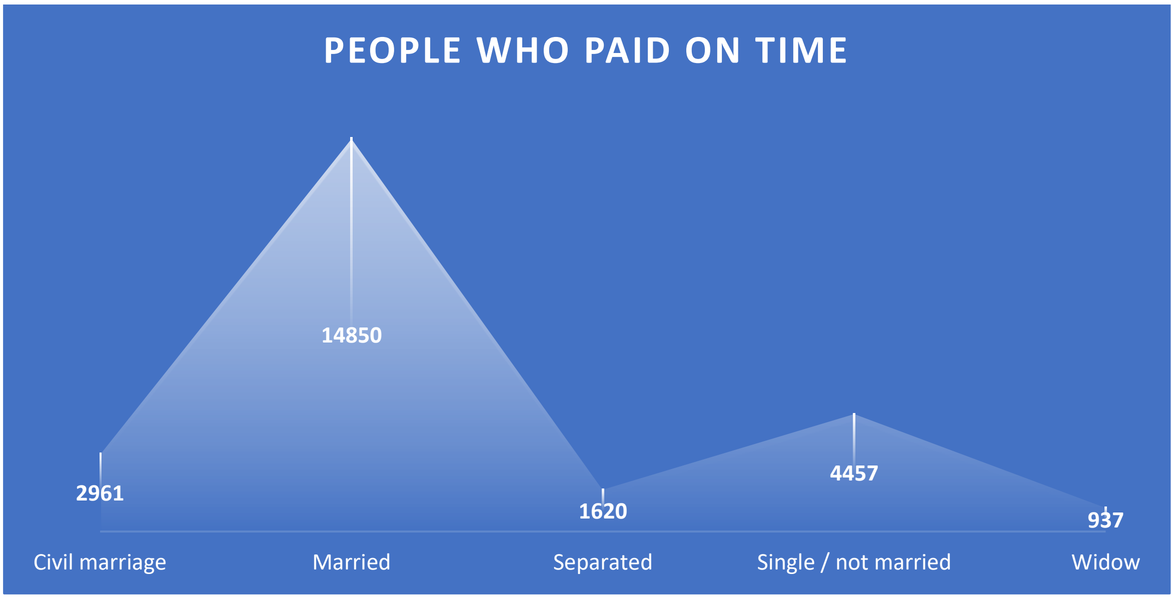
- From above chart, we can see Most of the applicants (52%) who are applying for loans are Bussinessman.



- From above chart, we can say people with qualification of secondary / secondary special pay their loans on time.



- From above chart, we can say that Female are the ones with higher defaults.



- From above chart, we can say that Married people pay their loans on time when compared to single,widow,separated categories.

MY EXCEL FILES:

https://docs.google.com/spreadsheets/d/1ZqglQi9zhT1tzXxyR9r2OexHk6dgLPO9/edit?usp=share_link&ouid=104755012826368900391&rtpof=true&sd=true

https://docs.google.com/spreadsheets/d/13AN6vNQMXtE2UlJxoBpFsmVv4H7BtC9P/edit?usp=share_link&ouid=104755012826368900391&rtpof=true&sd=true