Bank Loan Case Study

Project Description

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. 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.

The company wants to understand the key factors behind loan default so it can make better decisions about loan approval. I will be using Exploratory Data Analysis (EDA) to understand how customer attributes and loan attributes influence the likelihood of default.

Tech-Stack Used

To complete this project, I am using MS Excel 2021 as it is an excellent analysis tool to analyze small size datasets. Also, I will be using MS Power Point and MS Word to form a report.

Approach

Exploratory Data Analysis or EDA is used to take insights from the data. Data Scientists and Analysts try to find different patterns, relations, and anomalies in the data using some statistical graphs and other visualization techniques.

Now, let's understand the terms like cash loan and revolving loan. **Cash loan** is the type of loan where borrower gets the loan in cash. It can be given to an individual or a business. **Revolving loan** is the type of loan in which a lender lends money to a borrower upto an approved limit eg. credit cards, home equity etc.

The given data contains two dataset: application_data file which contains details about the current loan applications and previous_application file which contains information about previous loan applications. After understanding the given dataset, first I am going to:

1. Identify Missing Data and Deal with it Appropriately:

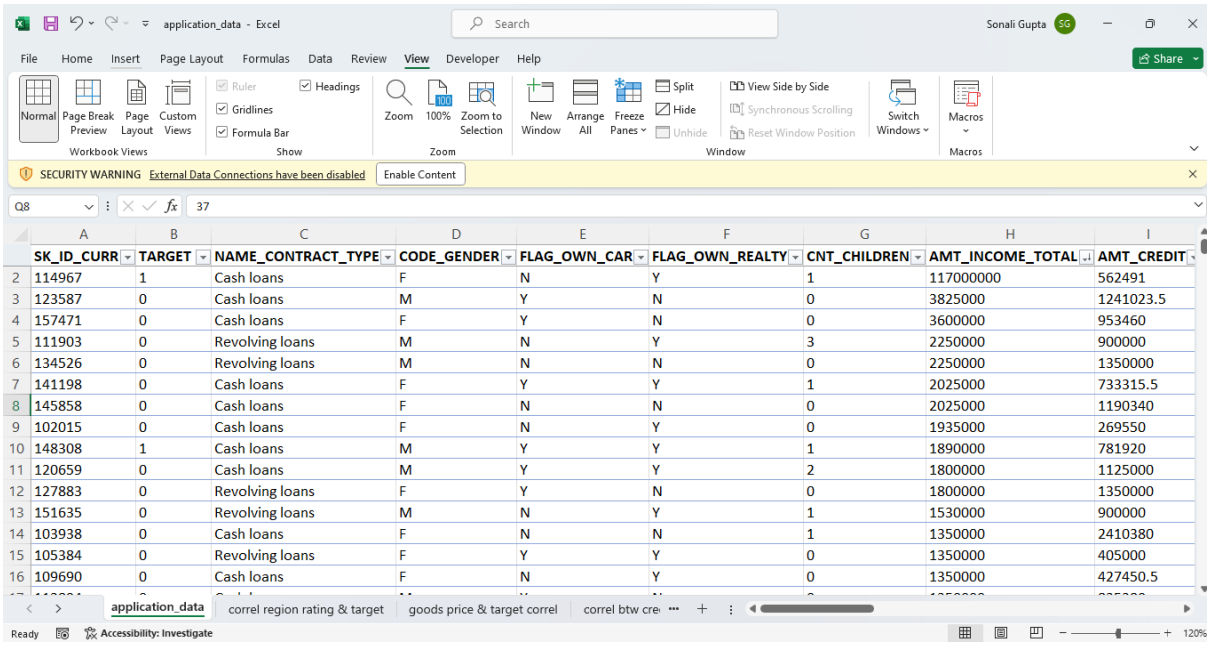
Before cleaning, the given file named application_data contains 50,000 rows and 122 columns in total and after cleaning the given file, it now contains 49961 rows and 22 columns. Also, earlier previous_aaplication file contained 50,000 rows and 38 columns, and after cleaning, it now contains 49992 rows and 22 columns.

- One column is a calculated column named age_of_applicant (in yrs) as the age of the applicant was given in days and by using =int(ABS(P2)/365) formula the age of the person in years has been derived in excel in application_data file.
- I have removed several rows that contains blank values.
- I have deleted a large number of unwanted columns that are not going to help with the analysis.

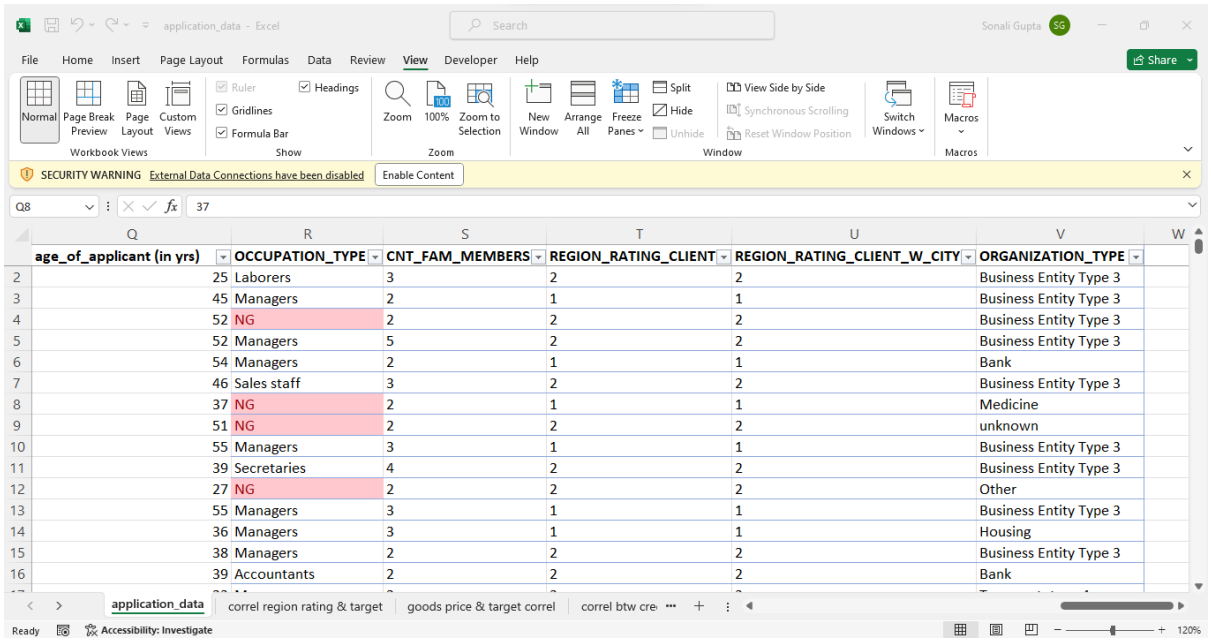
- I have also replaced house / apartment with house, secondary / special education degree with secondary education and single / not married with not married in application_data file.
- I have also used =ISBLANK() function to check for blanks.
- I have also replaced some blank values in column occupation_type with NG (not given) as deleting a large amount of blank rows can result in the loss of some important information required for the analysis. Same has been done in previous_application file.

Here are the screenshots for both:

Application_data

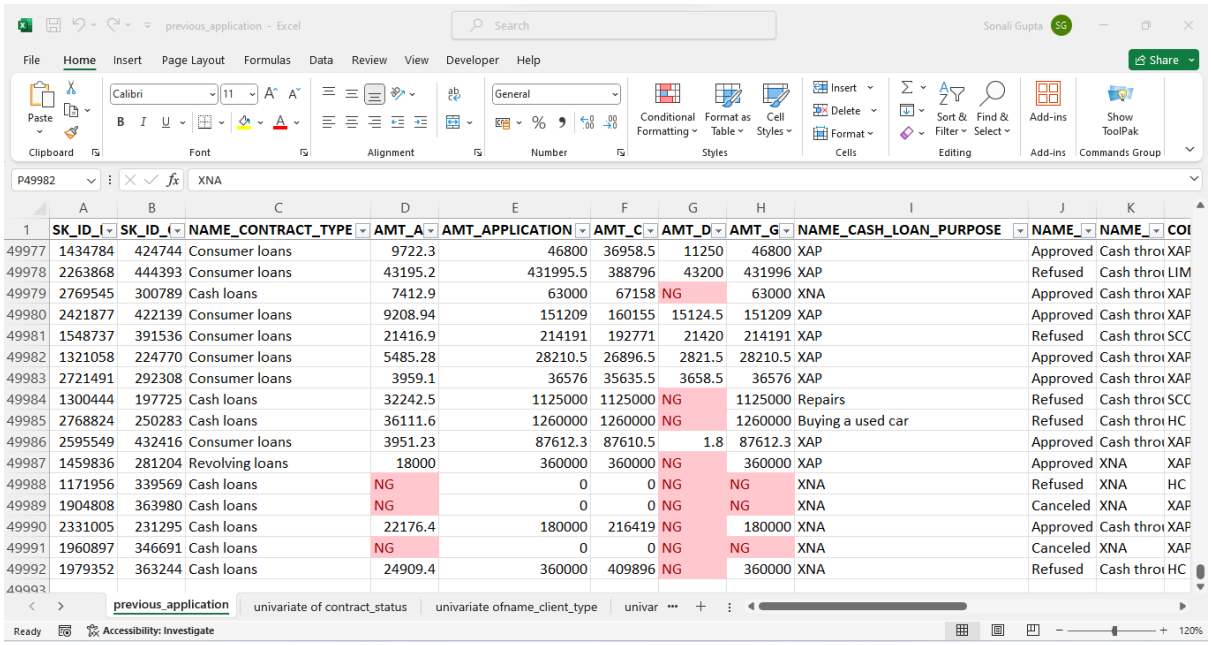


	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT
2	114967	1	Cash loans	F	N	Y	1	117000000	562491
3	123587	0	Cash loans	M	Y	N	0	3825000	1241023.5
4	157471	0	Cash loans	F	Y	N	0	3600000	953460
5	111903	0	Revolving loans	M	N	Y	3	2250000	900000
6	134526	0	Revolving loans	M	N	N	0	2250000	1350000
7	141198	0	Cash loans	F	Y	Y	1	2025000	733315.5
8	145858	0	Cash loans	F	N	N	0	2025000	1190340
9	102015	0	Cash loans	F	N	Y	0	1935000	269550
10	148308	1	Cash loans	M	Y	Y	1	1890000	781920
11	120659	0	Cash loans	M	Y	Y	2	1800000	1125000
12	127883	0	Revolving loans	F	Y	N	0	1800000	1350000
13	151635	0	Revolving loans	M	N	Y	1	1530000	900000
14	103938	0	Cash loans	F	N	N	1	1350000	2410380
15	105384	0	Revolving loans	F	Y	Y	0	1350000	405000
16	109690	0	Cash loans	F	N	Y	0	1350000	427450.5

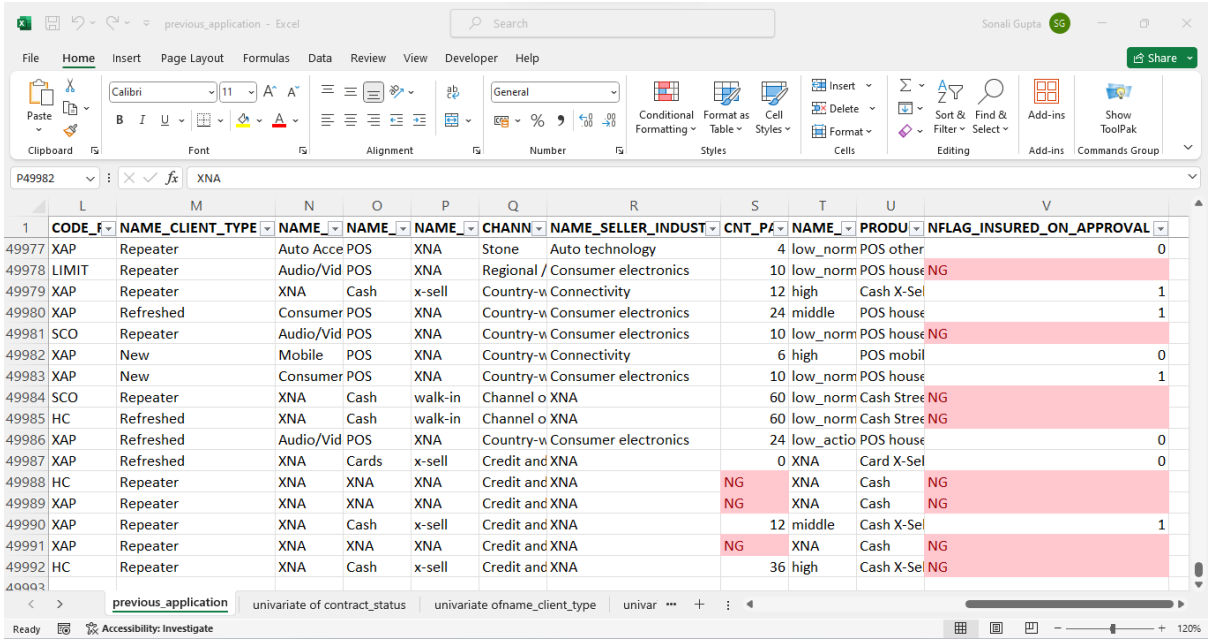


	age_of_applicant (in yrs)	OCCUPATION_TYPE	CNT_FAM_MEMBERS	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	ORGANIZATION_TYPE
2	25	Laborers	3	2	2	Business Entity Type 3
3	45	Managers	2	1	1	Business Entity Type 3
4	52	NG	2	2	2	Business Entity Type 3
5	52	Managers	5	2	2	Business Entity Type 3
6	54	Managers	2	1	1	Bank
7	46	Sales staff	3	2	2	Business Entity Type 3
8	37	NG	2	1	1	Medicine
9	51	NG	2	2	2	unknown
10	55	Managers	3	1	1	Business Entity Type 3
11	39	Secretaries	4	2	2	Business Entity Type 3
12	27	NG	2	2	2	Other
13	55	Managers	3	1	1	Business Entity Type 3
14	36	Managers	3	1	1	Housing
15	38	Managers	2	2	2	Business Entity Type 3
16	39	Accountants	2	2	2	Bank

Previous_application

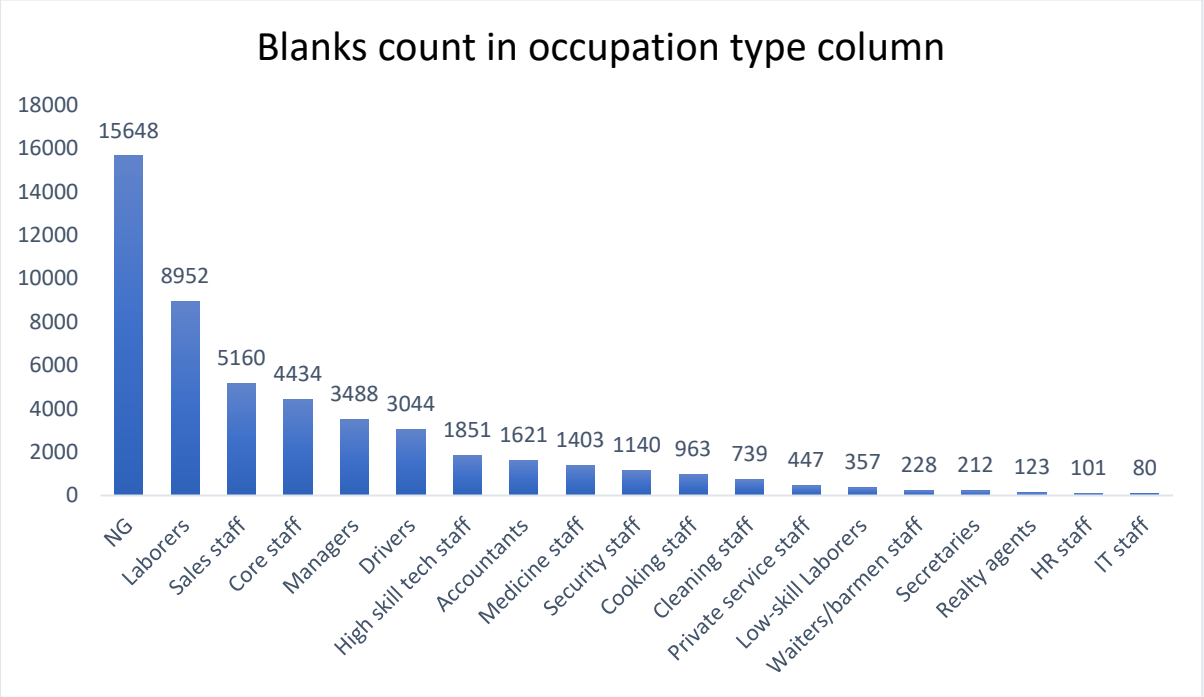


	SK_ID_(	SK_ID_(	NAME_CONTRACT_TYPE	AMT_A	AMT_APPLICATION	AMT_C	AMT_D	AMT_G	NAME_CASH_LOAN_PURPOSE	NAME_	NAME_	COI
49977	1434784	424744	Consumer loans	9722.3	46800	36958.5	11250	46800	XAP	Approved	Cash throu	XAP
49978	2263868	444393	Consumer loans	43195.2	431995.5	388796	43200	431996	XAP	Refused	Cash throu	LIM
49979	2769545	300789	Cash loans	7412.9	63000	67158	NG	63000	XNA	Approved	Cash throu	XAP
49980	2421877	422139	Consumer loans	9208.94	151209	160155	15124.5	151209	XAP	Approved	Cash throu	XAP
49981	1548737	391536	Consumer loans	21416.9	214191	192771	21420	214191	XAP	Refused	Cash throu	SCC
49982	1321058	224770	Consumer loans	5485.28	28210.5	26896.5	2821.5	28210.5	XAP	Approved	Cash throu	XAP
49983	2721491	292308	Consumer loans	3959.1	36576	35635.5	3658.5	36576	XAP	Approved	Cash throu	XAP
49984	1300444	197725	Cash loans	32242.5	1125000	1125000	NG	1125000	Repairs	Refused	Cash throu	SCC
49985	2768824	250283	Cash loans	36111.6	1260000	1260000	NG	1260000	Buying a used car	Refused	Cash throu	HC
49986	2595549	432416	Consumer loans	3951.23	87612.3	87610.5	1.8	87612.3	XAP	Approved	Cash throu	XAP
49987	1459836	281204	Revolving loans	18000	360000	360000	NG	360000	XAP	Approved	XNA	XAP
49988	1171956	339569	Cash loans	NG	0	0	NG	NG	XNA	Refused	XNA	HC
49989	1904808	363980	Cash loans	NG	0	0	NG	NG	XNA	Canceled	XNA	XAP
49990	2331005	231295	Cash loans	22176.4	180000	216419	NG	180000	XNA	Approved	Cash throu	XAP
49991	1960897	346691	Cash loans	NG	0	0	NG	NG	XNA	Canceled	XNA	XAP
49992	1979352	363244	Cash loans	24909.4	360000	409896	NG	360000	XNA	Refused	Cash throu	HC



	L	M	N	O	P	Q	R	S	T	U	V
1	CODE_F	NAME_CLIENT_TYPE	NAME_	NAME_	NAME_	CHANN	NAME_SELLER_INDUST	CNT_PA	NAME_	PRODU	NFLAG_INSURED_ON_APPROVAL
49977	XAP	Repeater	Auto Acce	POS	XNA	Stone	Auto technology	4	low_norm	POS other	0
49978	LIMIT	Repeater	Audio/Vid	POS	XNA	Regional /	Consumer electronics	10	low_norm	POS house	NG
49979	XAP	Repeater	XNA	Cash	x-sell	Country-w	Connectivity	12	high	Cash X-Sel	1
49980	XAP	Refreshed	Consumer	POS	XNA	Country-w	Consumer electronics	24	middle	POS house	1
49981	SCO	Repeater	Audio/Vid	POS	XNA	Country-w	Consumer electronics	10	low_norm	POS house	NG
49982	XAP	New	Mobile	POS	XNA	Country-w	Connectivity	6	high	POS mobil	0
49983	XAP	New	Consumer	POS	XNA	Country-w	Consumer electronics	10	low_norm	POS house	1
49984	SCO	Repeater	XNA	Cash	walk-in	Channel o	XNA	60	low_norm	Cash Stree	NG
49985	HC	Refreshed	XNA	Cash	walk-in	Channel o	XNA	60	low_norm	Cash Stree	NG
49986	XAP	Refreshed	Audio/Vid	POS	XNA	Country-w	Consumer electronics	24	low_actio	POS house	0
49987	XAP	Refreshed	XNA	Cards	x-sell	Credit and	XNA	0	XNA	Card X-Sel	0
49988	HC	Repeater	XNA	XNA	XNA	Credit and	XNA	NG	XNA	Cash	NG
49989	XAP	Repeater	XNA	XNA	XNA	Credit and	XNA	NG	XNA	Cash	NG
49990	XAP	Repeater	XNA	Cash	x-sell	Credit and	XNA	12	middle	Cash X-Sel	1
49991	XAP	Repeater	XNA	XNA	XNA	Credit and	XNA	NG	XNA	Cash	NG
49992	HC	Repeater	XNA	Cash	x-sell	Credit and	XNA	36	high	Cash X-Sel	NG
49993											

Here is the chart showing the count of blanks in the name_occupation_type column:



The above chart is showing the number of blanks which I have replaced with NG (not given) in the name_occupation_type column as this column is going to be used for analysis and removing such huge amount of rows from the dataset can result into a skewed analysis.

2. Identify Outliers in the Dataset

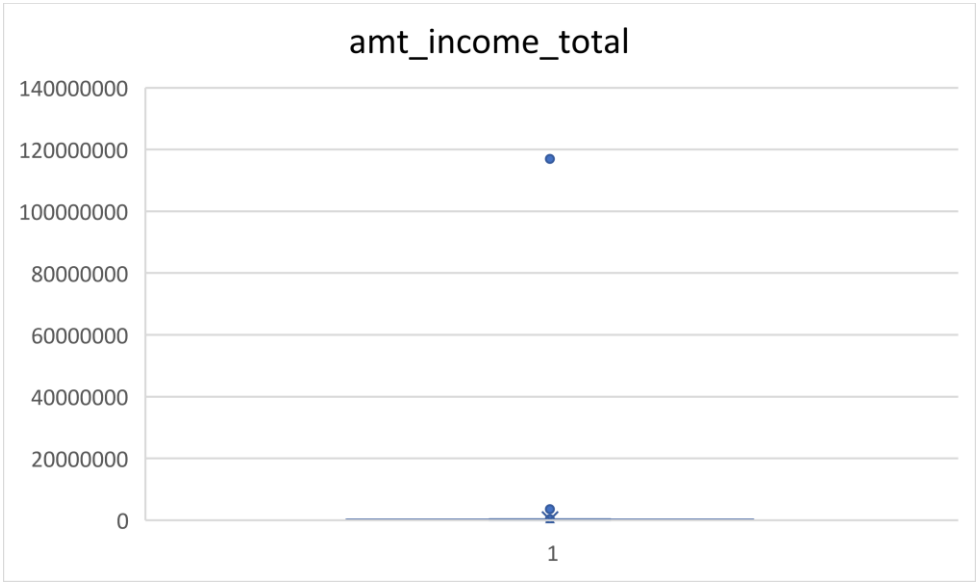
Outliers can significantly impact the analysis and distort the results. They can only be identified in columns that contains numerical values like amt_credit, amt_annuity, amt_income_total. I have used quartile, IQR and average functions in excel to identify outliers in application_data.

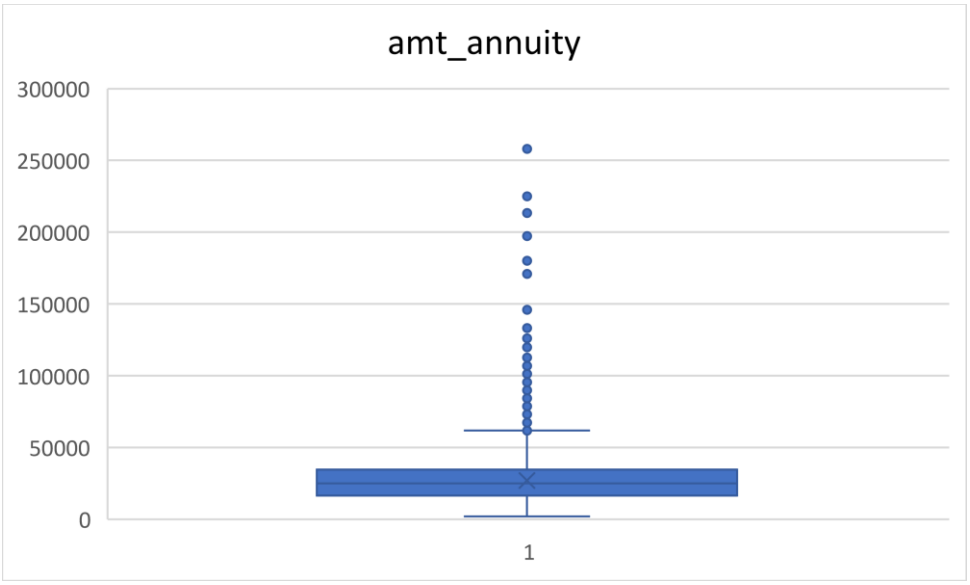
For amt_income_total	
First quartile =	112500
Third quartile =	202500
Median of total income =	146250
max total income =	117000000
min total income =	25650
Interquartile range (IQR) =	90000
Lower outlier limit =	-22500
Upper outlier limit =	337500
Average total income =	170786
Number of rows =	49960

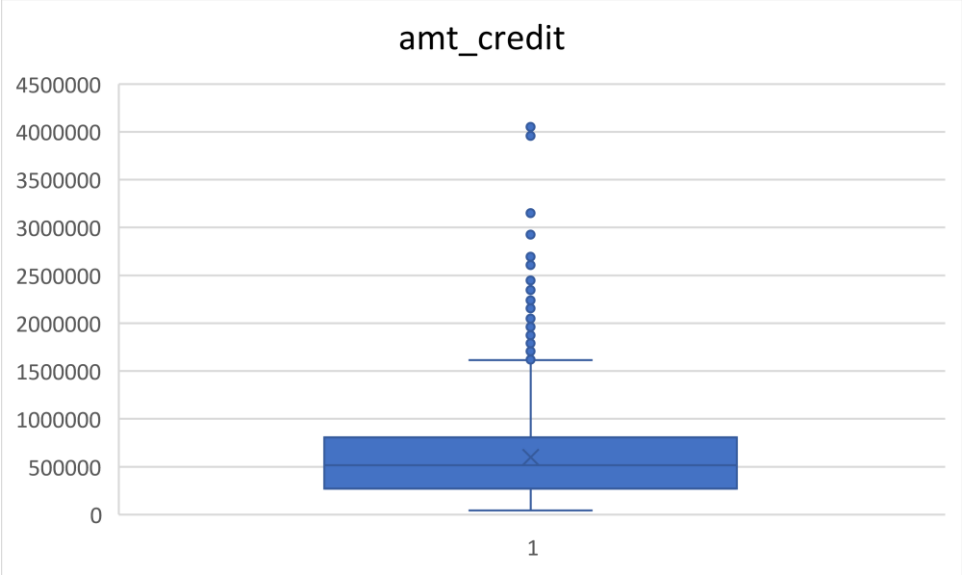
For amt_annuity	
First quartile =	16486.9
Third quartile =	34596
Median =	24939
max =	258026
min =	2052
Interquartile range (IQR) =	18109.1
Lower outlier limit =	-10676.8
Upper outlier limit =	61759.7
Average =	27115.1
Number of rows =	49960

For amt_credit	
First quartile =	270000
Third quartile =	808650
Median =	515529
max =	4050000
min =	45000
Interquartile range (IQR) =	538650
Lower outlier limit =	-537975
Upper outlier limit =	1616625
Average =	599903
Number of rows =	49960

Box plot for all three numerical values from application_data:







Insights Gained:

The blue colored dots in all the above box plot is showing that outliers are present in the amt_income_total, amt_credit and amt_annuity. Also, from the above statistical measures for these three columns, few things has been derived like:

- For amt_income_total, values lying below -22500 will be outliers and the values lying above 337500 will be outliers.
- For amt_annuity, values lying below -10676.8 will be outliers and the values lying above 61759.7 will be outliers.
- For amt_credit, values lying below -537975 will be outliers and the values lying above 1616625 will be outliers.

3. Analyze Data Imbalance:

Data imbalance means when the data is classified into skewed class proportions and can affect the accuracy of the analysis, especially for binary classification problems. Below are the pivot tables showing the skewed class distributions for various variables along with target variable.

Row Labels	Count of TARGET
0	45936
1	4024
Grand Total	**49960**

Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	45275
Revolving loans	4685
Grand Total	49960

Row Labels	Count of CODE_GENDER
F	32800
M	17158
XNA	2
Grand Total	49960

Row Labels	Count of NAME_INCOME_TYPE
Working	25985
Commercial associate	11536
Pensioner	8916
State servant	3509
Unemployed	6
Student	5
Businessman	2
Maternity leave	1
Grand Total	**49960**

Row Labels	Count of NAME_EDUCATION_TYPE
Secondary education	35547
Higher education	12156
Incomplete higher	1619
Lower secondary	618
Academic degree	20
Grand Total	49960

Row Labels	Count of NAME_FAMILY_STATUS
Married	32069
not married	7298
Civil marriage	4855
Separated	3142
Widow	2596
Grand Total	**49960**

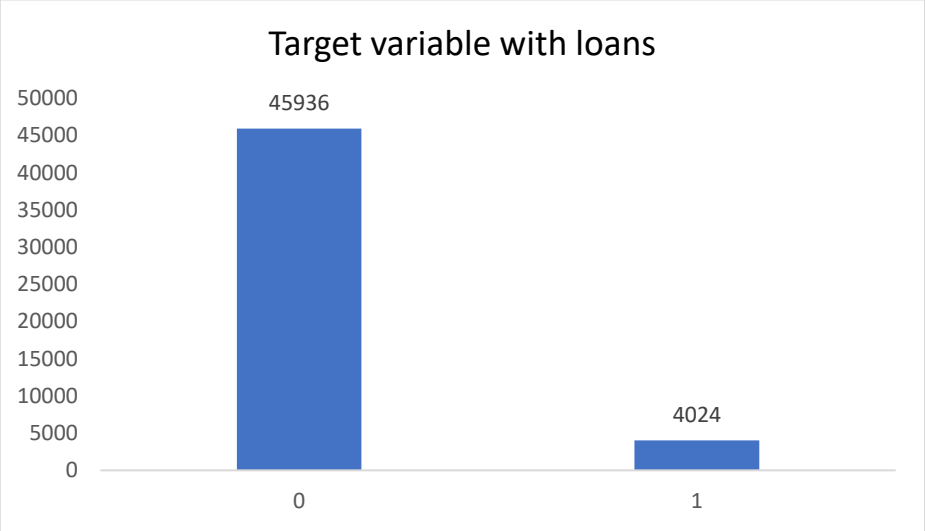
Row Labels	Count of NAME_HOUSING_TYPE
House	44334
With parents	2398
Municipal apartment	1843
Rented apartment	768
Office apartment	427
Co-op apartment	190
Grand Total	**49960**

Row Labels	Count of OCCUPATION_TYPE
Laborers	8944
Sales staff	5157
Core staff	4428
Managers	3482
Drivers	3042
High skill tech staff	1851
Accountants	1620
Medicine staff	1403
Security staff	1137
Cooking staff	963
Cleaning staff	739
Private service staff	446
Low-skill Laborers	357
Waiters/barmen staff	227
Secretaries	212
Realty agents	123
HR staff	101
IT staff	80
Grand Total	**34312**

Row Labels	Sum of REGION_RATING_CLIENT
2	73868
3	23418
1	5220
Grand Total	102506

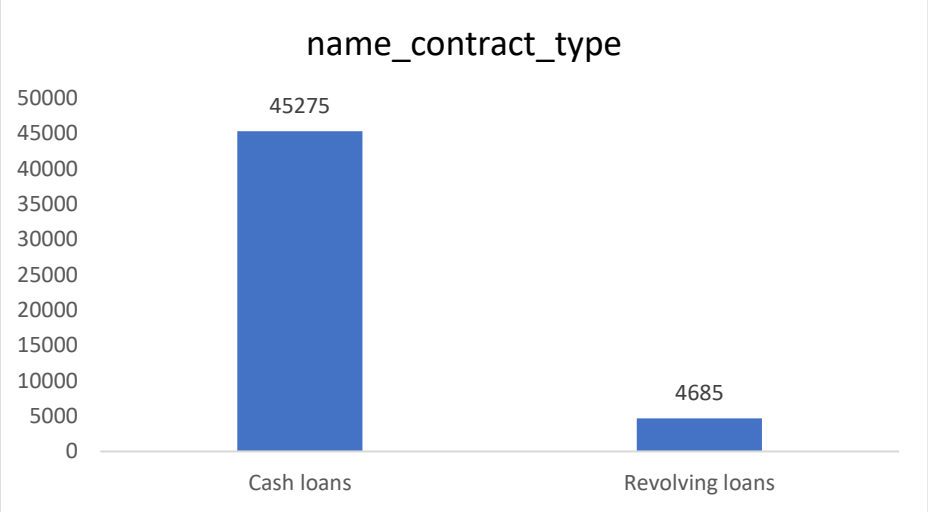
Row Labels	Count of ORGANIZATION_TYPE
Business Entity Type 3	11096
unknown	8920
Self-employed	6236
Other	2714
Medicine	1817
Government	1714
Business Entity Type 2	1702
School	1446
Trade: type 7	1209
Kindergarten	1089
Grand Total	**37943**

Graphical Representation of data imbalance:



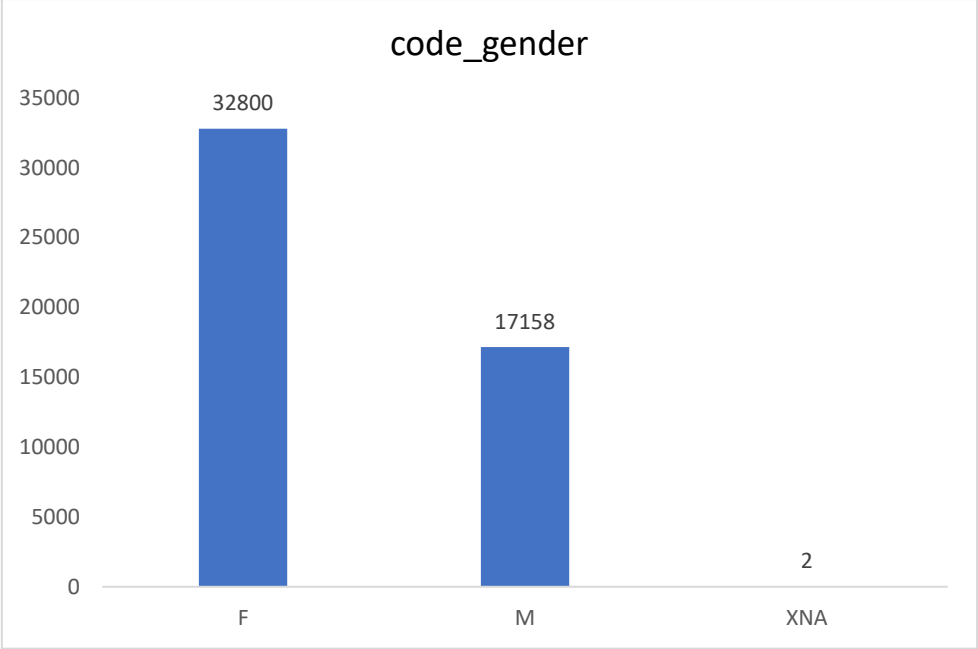
Insight Gained:

The above bar chart is showing the distribution of target variable which is showing that 0 means the payment of loan done on time is more than 1 means the late payment of loan.



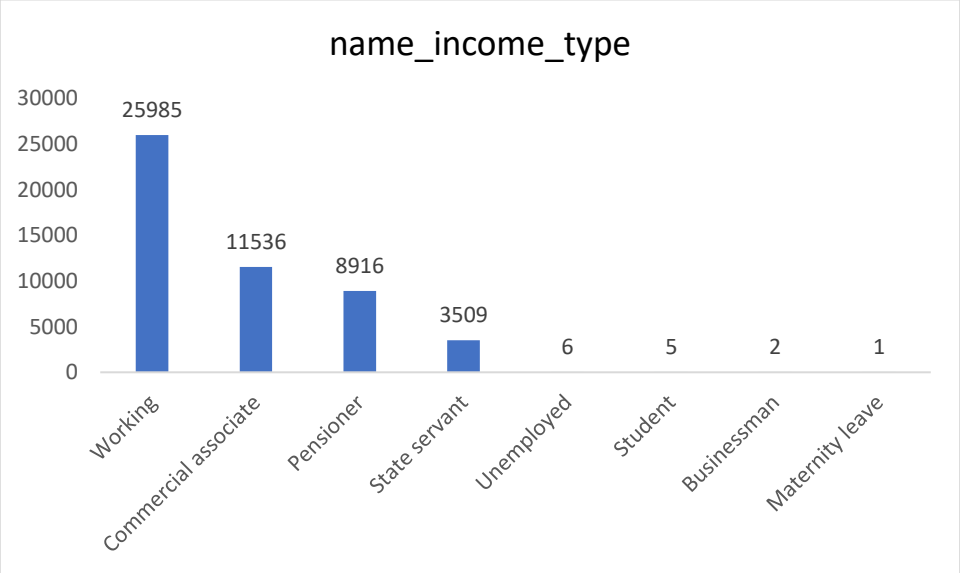
Insight Gained:

The above bar chart is showing that the number of cash loans requested by customers is higher that revolving loans.



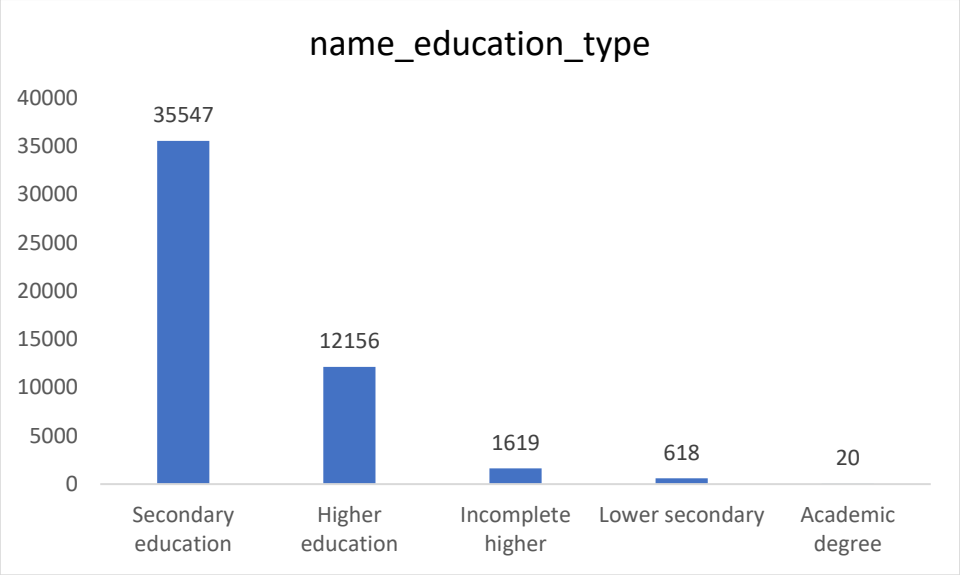
Insight Gained:

The above chart is showing that more loans are applied by more females than males.



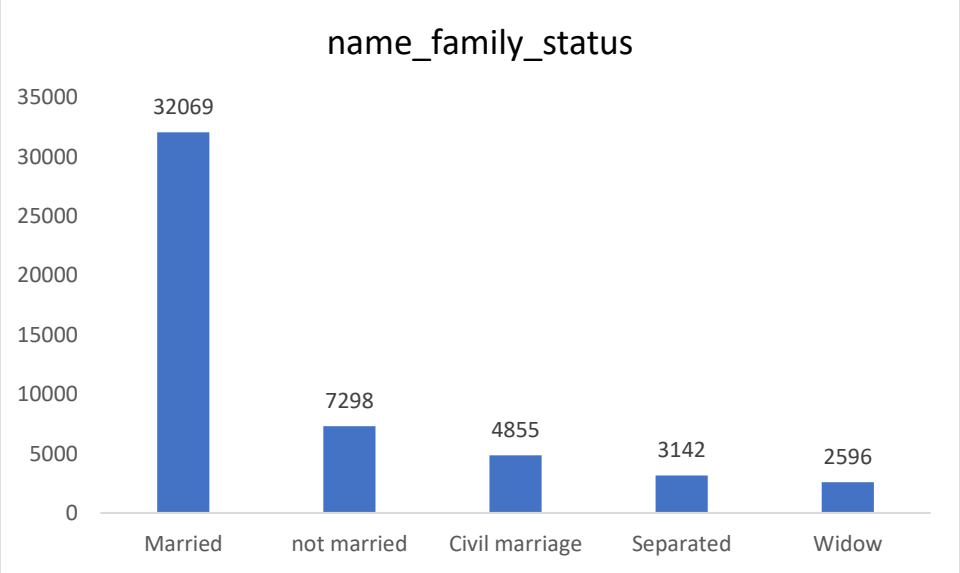
Insight Gained:

The above chart is showing that maximum number of people who applied for loans are working professionals.



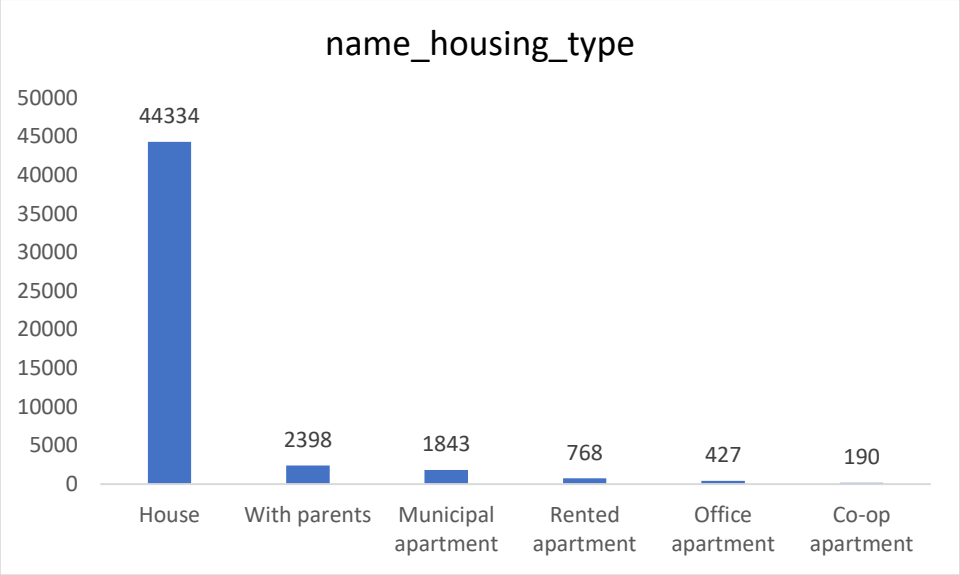
Insight Gained:

The above chart is showing that the maximum number of loans are applied by people who have secondary education.



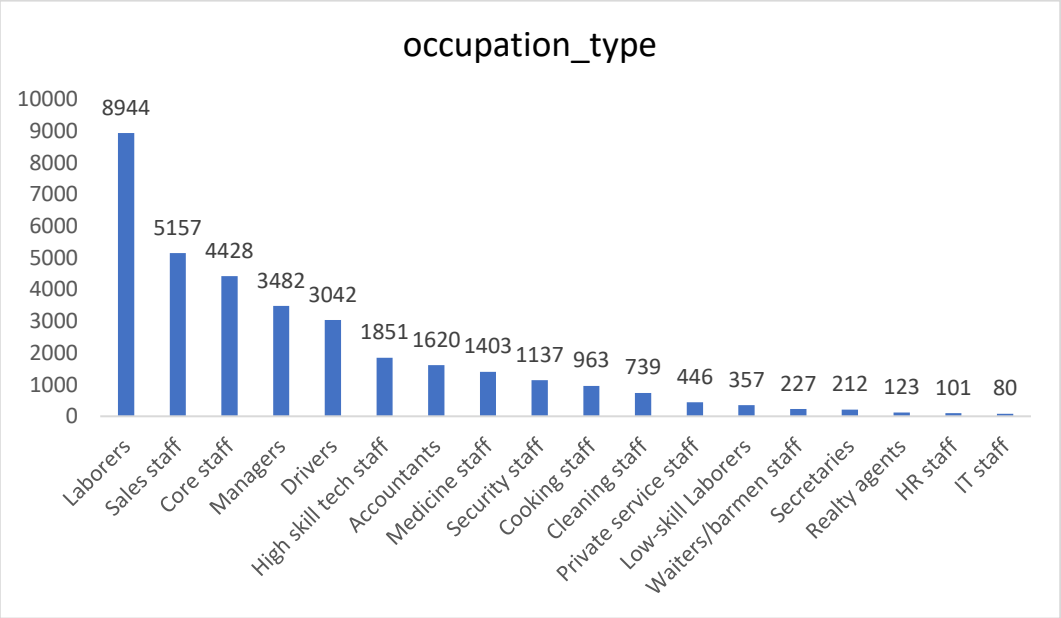
Insight Gained:

The above chart is showing that maximum loans are applied by married people.



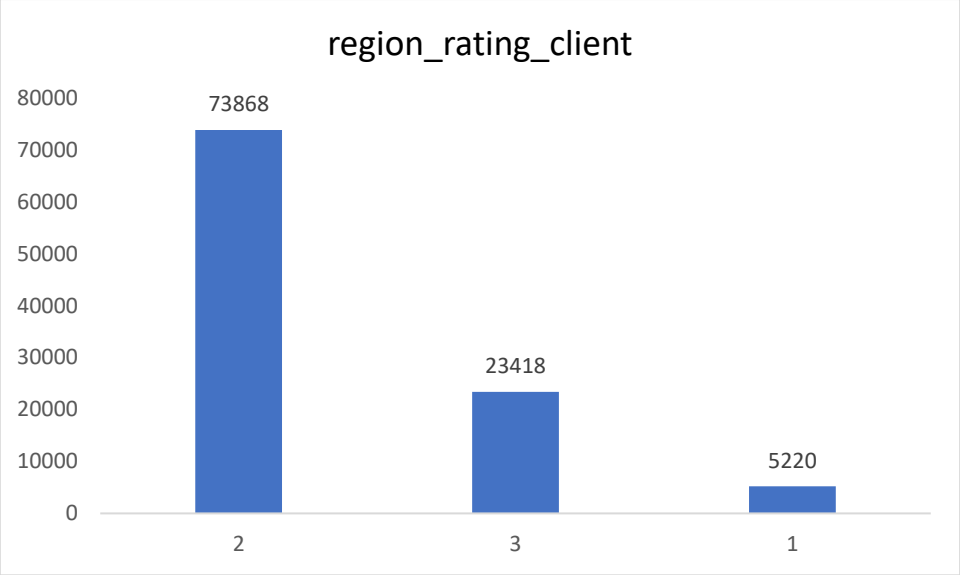
Insight Gained:

The above chart is showing that maximum number of loans are applied by the people with a house.



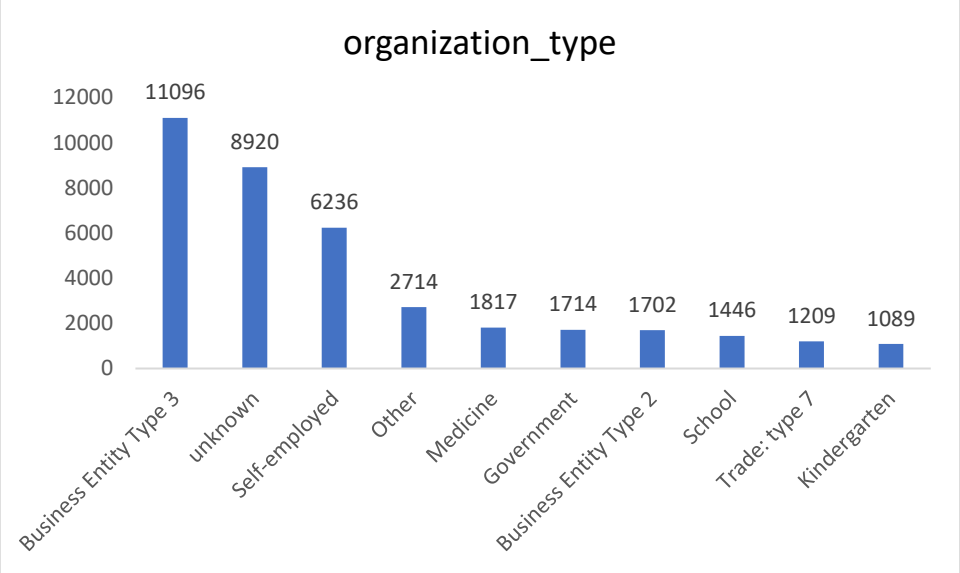
Insight Gained:

The above chart is showing that the maximum number of loans are applied by laborers than any other occupation type people.



Insight Gained:

The above chart is showing that the maximum number of loans are applied by the people living in the region with rating 2.



Insight Gained:

The above chart is showing that maximum number of loans are applied by people with organization type business entity type 3.

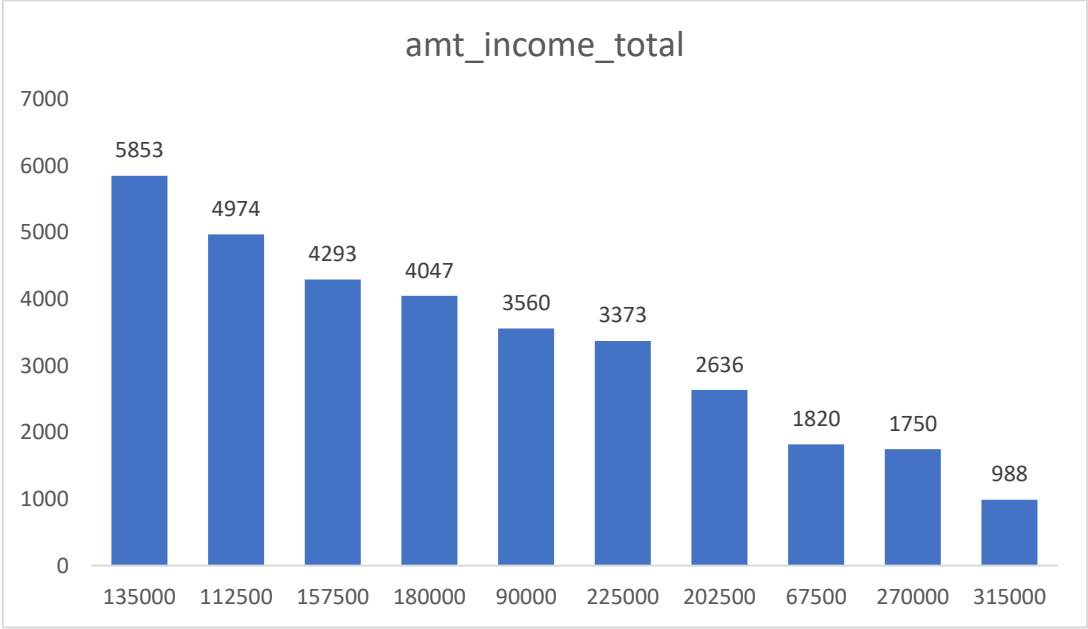
4. Perform Univariate, Segmented Univariate, and Bivariate Analysis

Univariate analysis is done to understand the distribution of a single variable. It includes:

❖ The below pivot table is showing the distribution of total income of first 10 applicants

Row Labels	Count of AMT_INCOME_TOTAL
135000	5853
112500	4974
157500	4293
180000	4047
90000	3560
225000	3373
202500	2636
67500	1820
270000	1750
315000	988
Grand Total	**33294**

Graphical Representation:



Insight Gained:

The above bar chart is clearly showing that the applicants with total income of 135000 has more financial capacity to repay loans. Also, the customers with more salary like 315000 are less likely to apply for loans.

❖ The below pivot chart is showing the distribution of the types of loans requested in application_data file and in previous_application file

For application_data

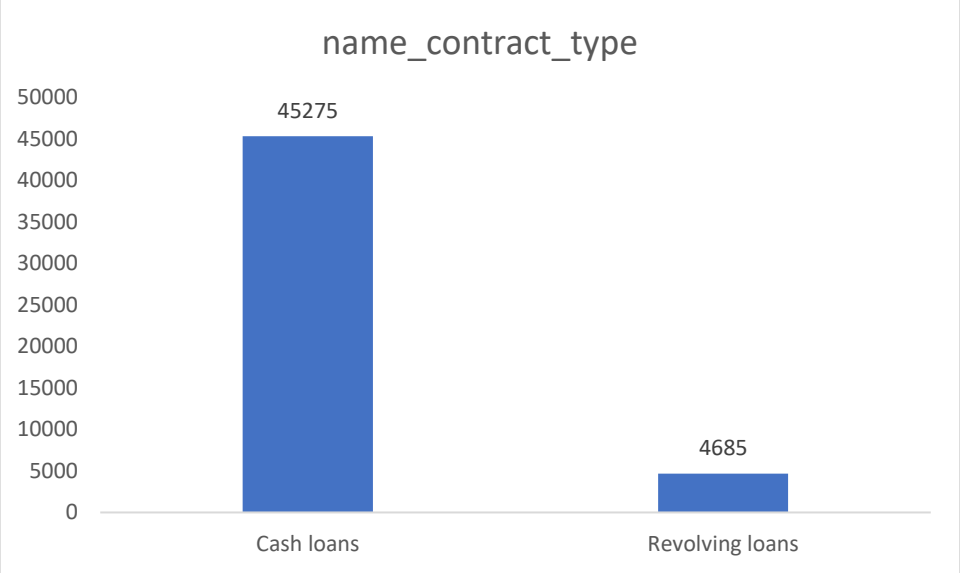
Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	45275
Revolving loans	4685
Grand Total	**49960**

For previous_application

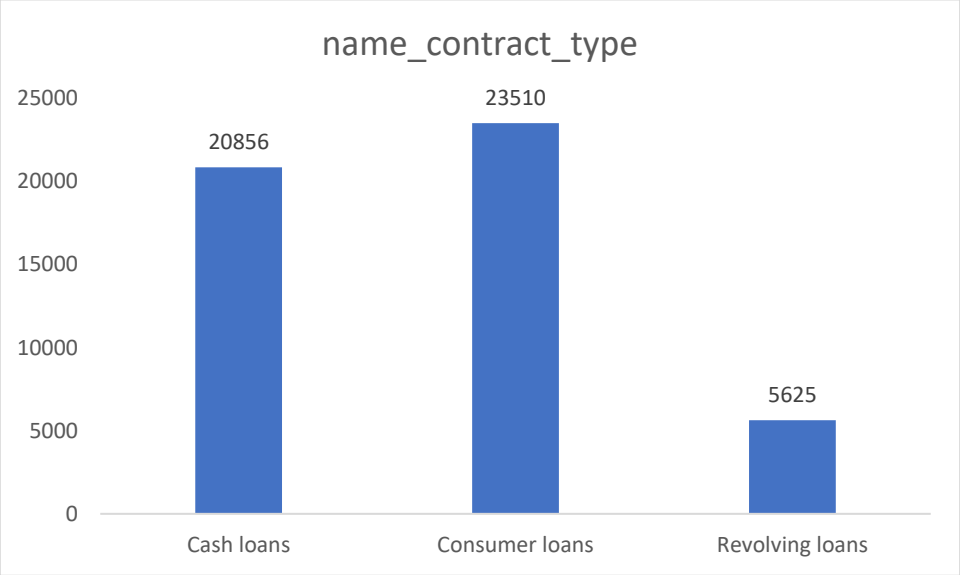
Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	20856
Consumer loans	23510
Revolving loans	5625
Grand Total	**49991**

Graphical Representations:

For application_data



For previous_application



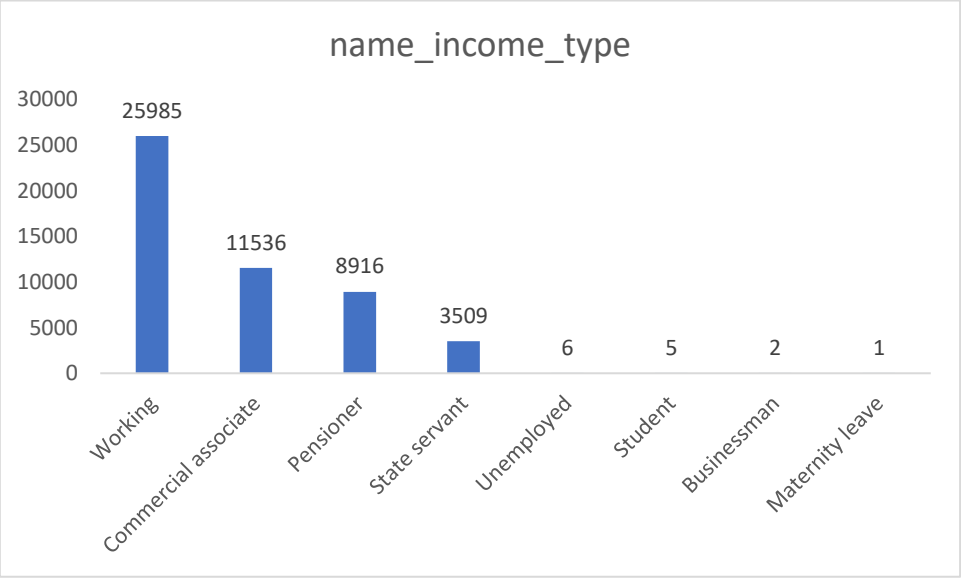
Insights Gained:

- From 1st bar chart, it is clear that only two types of loans are applied for by the applicants out of which cash loans are the most applied type of loan.
- From 2nd bar chart, it is clear that three types of loans are applied by the applicants previously, out of which consumer loans was the most applied type of loan.

❖ Below is the pivot table showing the distribution of employments of current applicants:

Row Labels	Count of NAME_INCOME_TYPE
Working	25985
Commercial associate	11536
Pensioner	8916
State servant	3509
Unemployed	6
Student	5
Businessman	2
Maternity leave	1
Grand Total	**49960**

Graphical Representation:



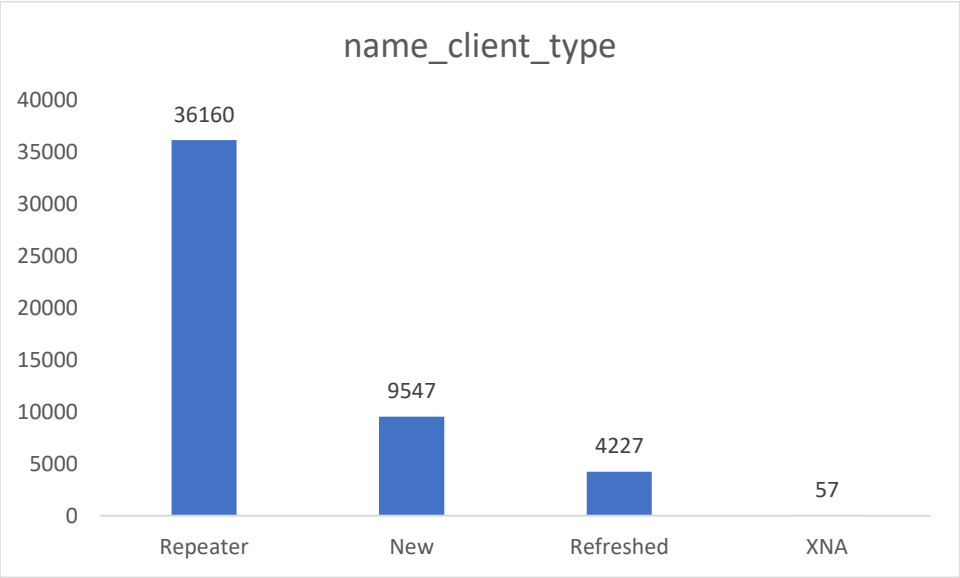
Insight Gained:

As from the above chart, it is clear that most of the loans are applied by working professionals which means the stable income source. This means that applicants who are working professionals have more probability of getting their loan approved.

❖ Below is the pivot table showing the previous applicant's nature:

Row Labels	Count of NAME_CLIENT_TYPE
Repeater	36160
New	9547
Refreshed	4227
XNA	57
Grand Total	**49991**

Graphical Representation:



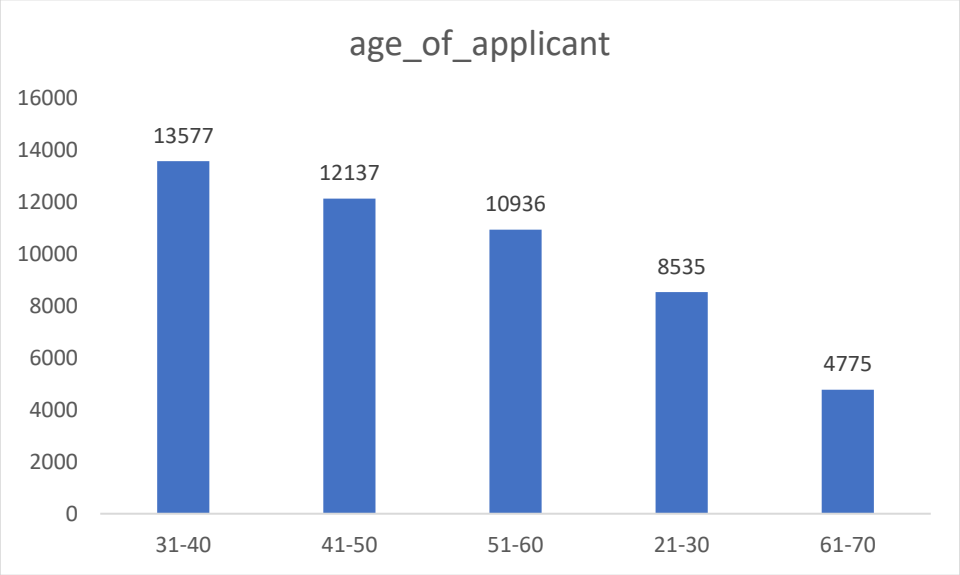
Insight Gained:

After the analysis, it has been found that numbers of people who applied for loan repeatedly is high than the others which means that there is a possibility that the people who apply for loans repeatedly can be loan defaulters.

❖ Below is the pivot table showing the age distribution loan applicants:

Row Labels	Count of age_of_applicant (in yrs)
31-40	13577
41-50	12137
51-60	10936
21-30	8535
61-70	4775
Grand Total	**49960**

Graphical Representation:



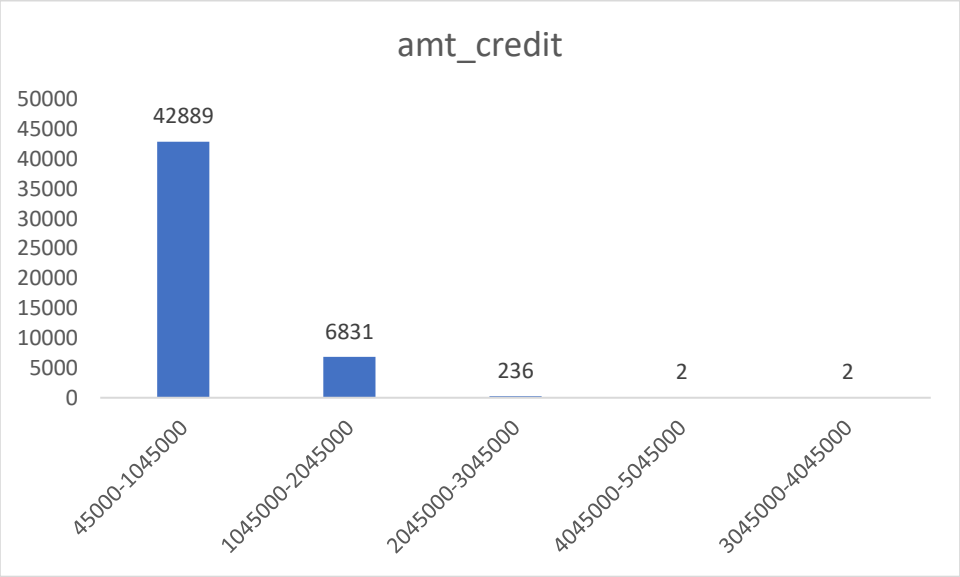
Insight Gained:

The above bar chart is clearly showing that mostly of the loans are applied by people with age between 31 and 40 years.

❖ Below is the pivot chart showing the distribution of amount of credit asked by people mostly:

Row Labels	Count of age_of_applicant (in yrs)
31-40	13577
41-50	12137
51-60	10936
21-30	8535
61-70	4775
Grand Total	**49960**

Graphical representation:



Insight Gained:

After analysis, it is clear that the amount of credit asked by most people ranges from 45000 to 1045000.

❖ After the analysis, it has been found that most of the loans are applied by people who are married (the graphical representation has been provided in data imbalance section).

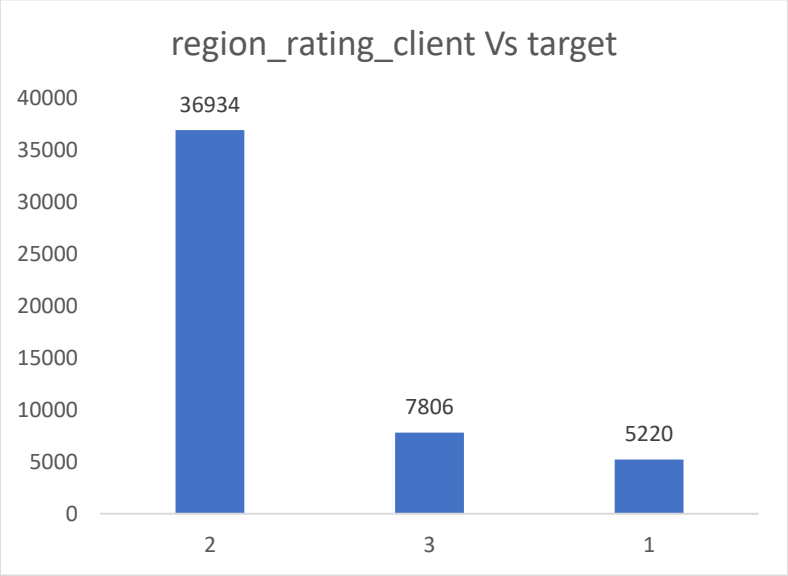
Bivariate analysis involves the examination and analysis of the relationship between two variables. The goal is to identify patterns, correlations, or dependencies between the chosen variables. It includes:

- ❖ Region rating client Vs target variable

Pivot Table:

Row Labels	Count of TARGET
2	36934
3	7806
1	5220
Grand Total	**49960**

Graphical Representation:



Insight Gained:

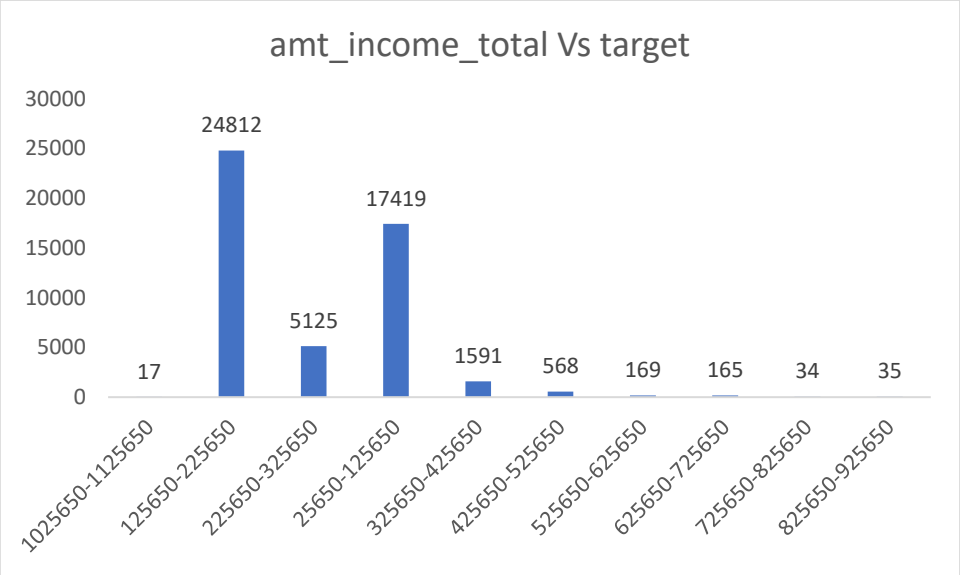
As from the above chart, it is clear that maximum loans applied by the applicants with region ratings 2 will most likely have more defaulters.

❖ Total income Vs Target

Pivot table:

Row Labels	Count of TARGET
1025650-1125650	17
125650-225650	24812
225650-325650	5125
25650-125650	17419
325650-425650	1591
425650-525650	568
525650-625650	169
625650-725650	165
725650-825650	34
825650-925650	35
Grand Total	**49935**

Graphical representation



Insight Gained:

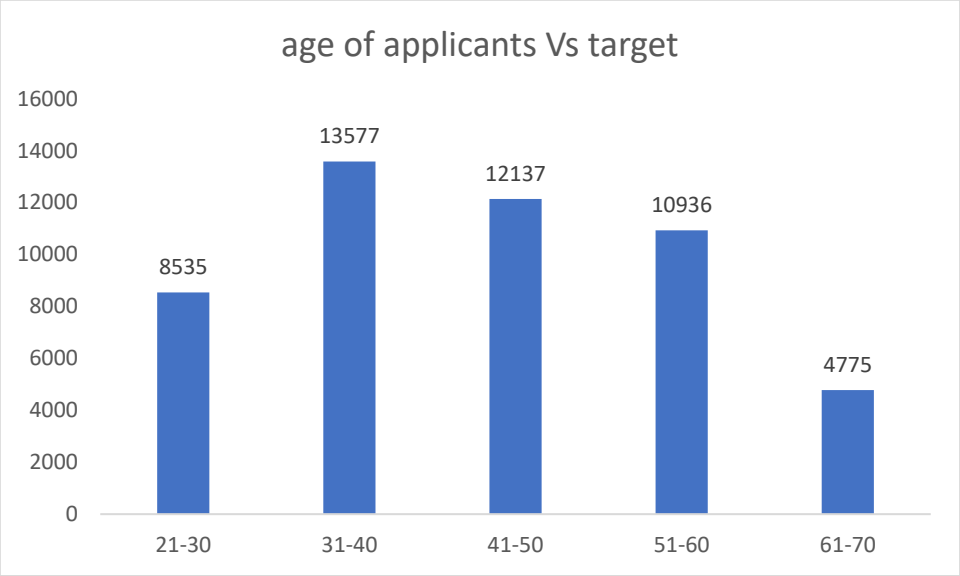
After the analysis, it has been observed that people with low income will most likely be defaulters.

❖ Age of applicants Vs target

Pivot table:

Row Labels	Count of TARGET
21-30	8535
31-40	13577
41-50	12137
51-60	10936
61-70	4775
Grand Total	**49960**

Graphical representation:



Insight Gained:

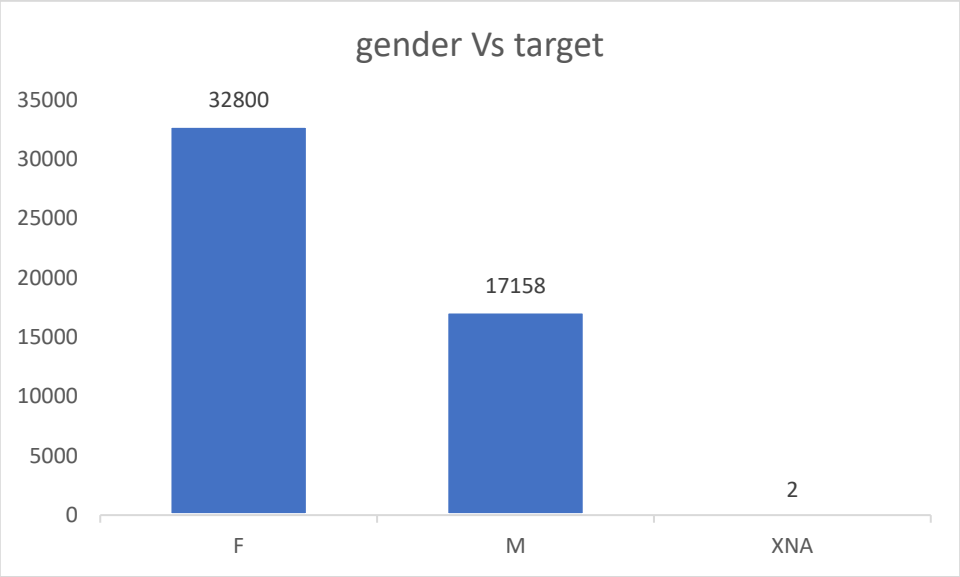
From the above analysis, it is clear that young people are most likely to be defaulters and defaulters are decreasing with increase in age.

❖ Gender of the applicants Vs Target

Pivot table:

Row Labels	Count of TARGET
F	32800
M	17158
XNA	2
Grand Total	**49960**

Graphical representation



Insight Gained:

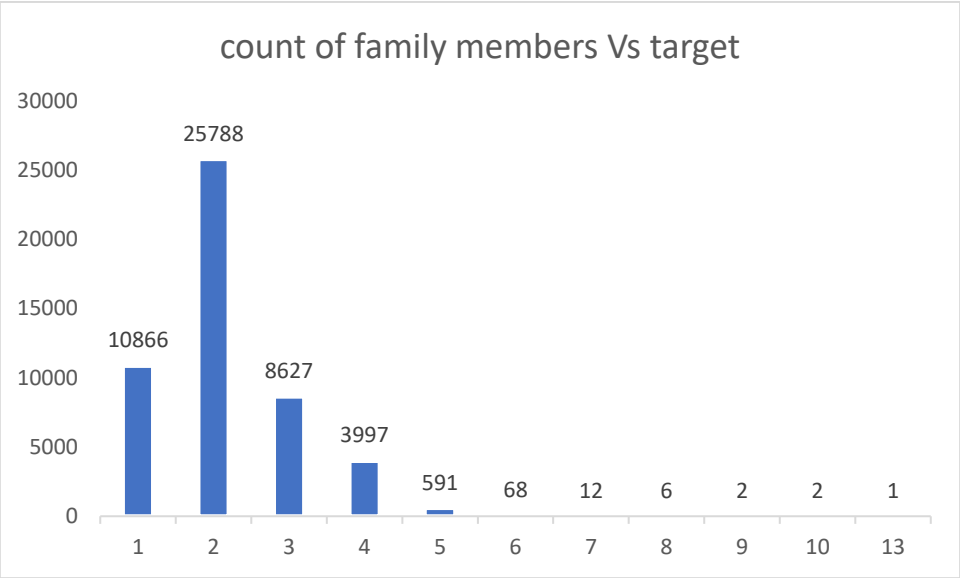
After the analysis, it has been found that females are less likely to be defaults than males.

❖ Family members Vs target

Pivot table:

Row Labels	Count of TARGET
1	10866
2	25788
3	8627
4	3997
5	591
6	68
7	12
8	6
9	2
10	2
13	1
Grand Total	**49960**

Graphical representation:



Insight Gained:

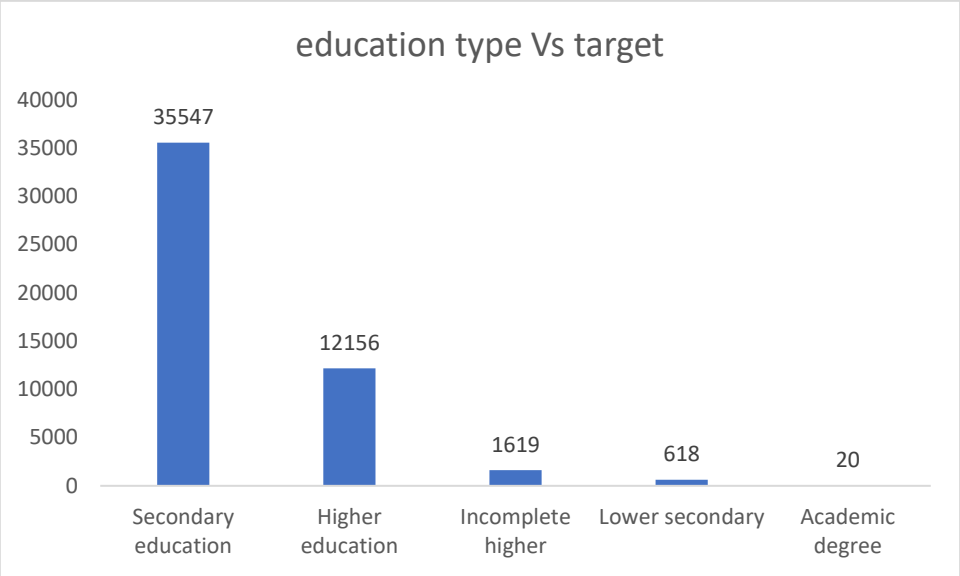
After the analysis, it has been found that people with more than 5 family members will most likely to be defaulters.

❖ Educational Qualifications of applicant Vs target

Pivot table:

Row Labels	Count of TARGET
Secondary education	35547
Higher education	12156
Incomplete higher	1619
Lower secondary	618
Academic degree	20
Grand Total	**49960**

Graphical Representation:



Insight Gained:

After the analysis, it has been found that as most of the loans are applied by customers with secondary education, then it means that customers with low educational qualification will most likely to be defaulters.

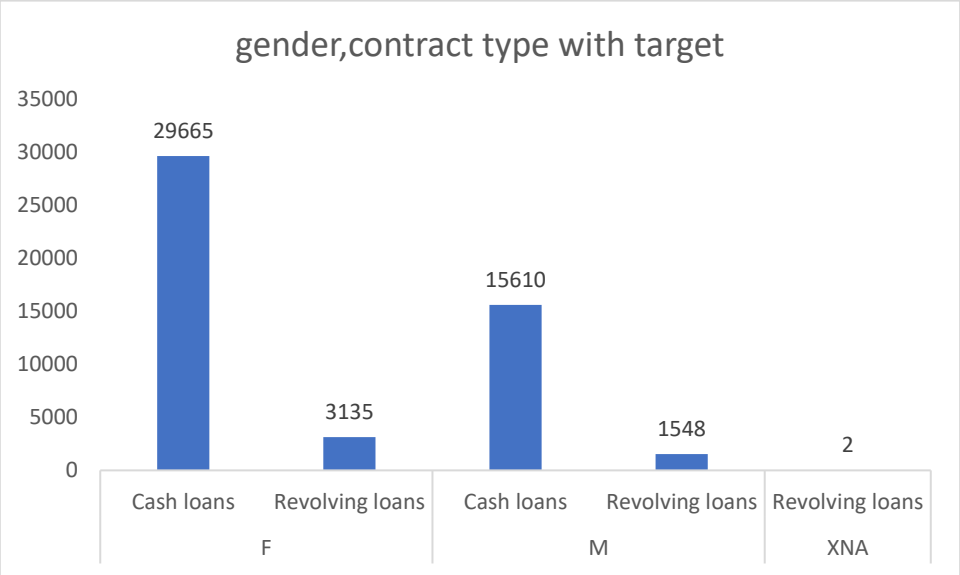
Segmented univariate analysis refers to the examination of individual variables within distinct subgroups or segments of a dataset. This approach allows for a more granular understanding of how different factors impact the variable of interest. By conducting segmented univariate analysis based on various factors, data analysts can uncover insights that might be overlooked in a broad analysis. It includes:

- Gender, contract type with target

Pivot table

Row Labels	Count of TARGET
F	**32800**
Cash loans	29665
Revolving loans	3135
M	**17158**
Cash loans	15610
Revolving loans	1548
XNA	**2**
Revolving loans	2
Grand Total	**49960**

Graphical Representation:



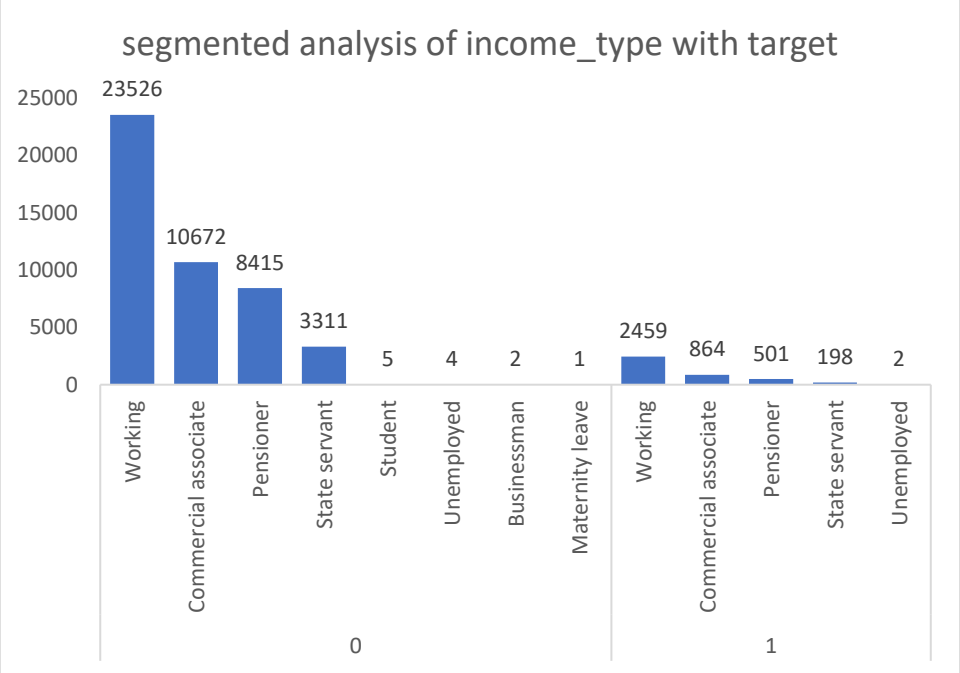
The above chart is showing the breakdown of gender of applicants and type of loans applied by them with the target variable.

❖ Income type with target

Pivot table:

Row Labels	Count of NAME_INCOME_TYPE
0	**45936**
Working	23526
Commercial associate	10672
Pensioner	8415
State servant	3311
Student	5
Unemployed	4
Businessman	2
Maternity leave	1
1	**4024**
Working	2459
Commercial associate	864
Pensioner	501
State servant	198
Unemployed	2
Grand Total	**49960**

Graphical Representation:



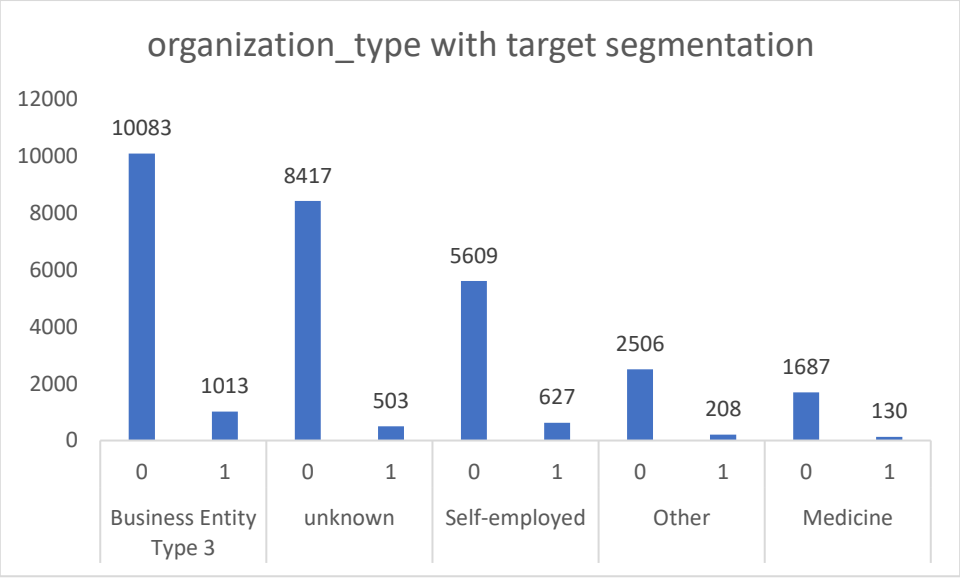
The above chart is showing the breakdown of income type of applicants with the target variable.

❖ Organization type with target

The below pivot table is showing the breakdown of organization type with target for top 5 professions only:

Row Labels	Count of TARGET
Business Entity Type 3	**11096**
0	10083
1	1013
unknown	**8920**
0	8417
1	503
Self-employed	**6236**
0	5609
1	627
Other	**2714**
0	2506
1	208
Medicine	**1817**
0	1687
1	130
Grand Total	**30783**

Graphical Representation:



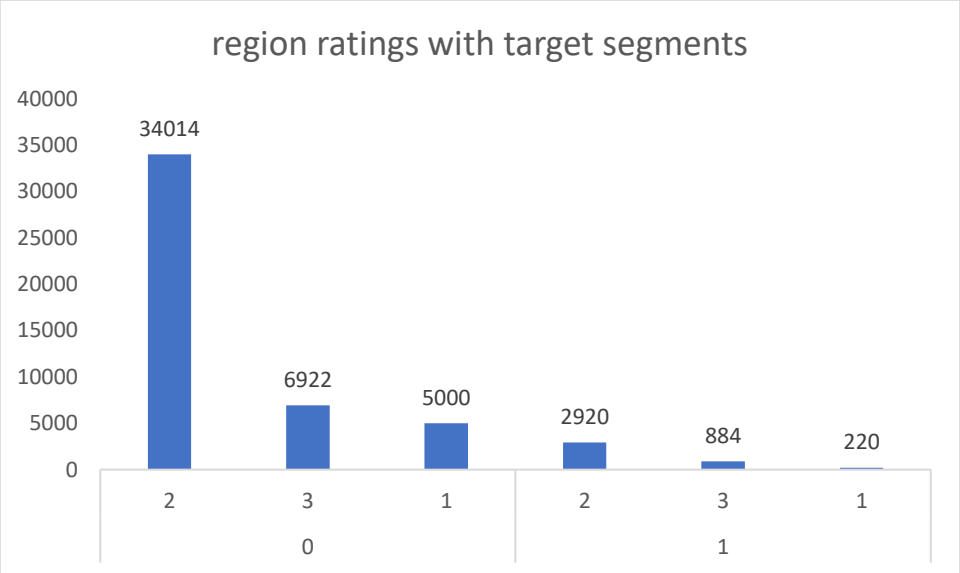
The above chart is showing the breakdown of type of employment of the applicants with target.

❖ Ratings of a region with target

Pivot table:

Row Labels	Count of TARGET
0	**45936**
2	34014
3	6922
1	5000
1	**4024**
2	2920
3	884
1	220
Grand Total	**49960**

Graphical Representation:



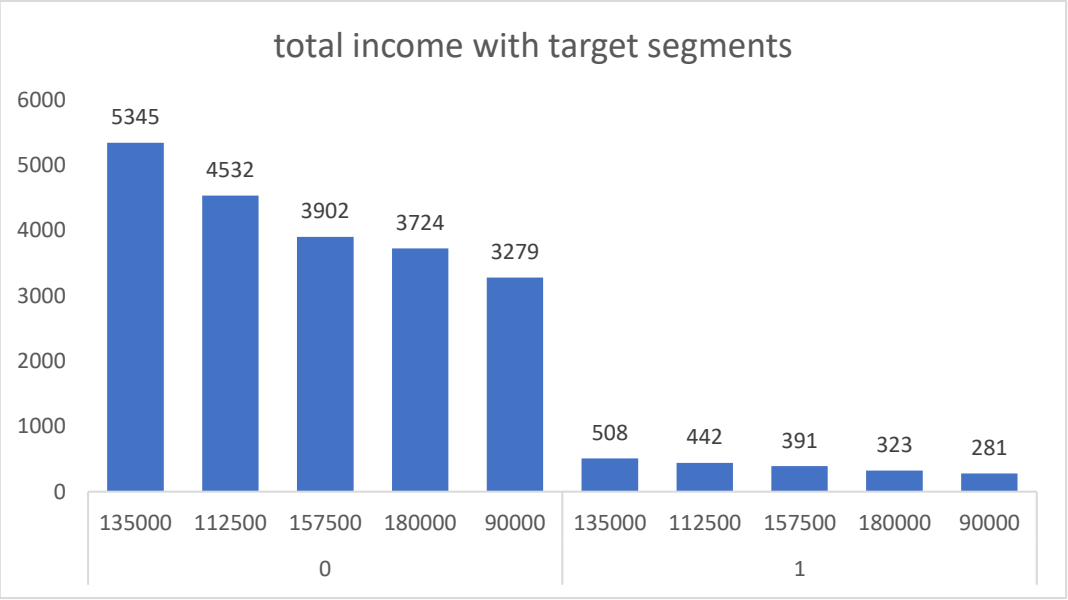
The above chart is showing the breakdown of ratings of region of applicants with target variable.

❖ Total Income of the applicants with target

Pivot Table:

Row Labels	Count of AMT_INCOME_TOTAL
0	**20782**
135000	5345
112500	4532
157500	3902
180000	3724
90000	3279
1	**1945**
135000	508
112500	442
157500	391
180000	323
90000	281
Grand Total	**22727**

Graphical Representation:



The above chart is showing the breakdown of total income of applicants with target variables.

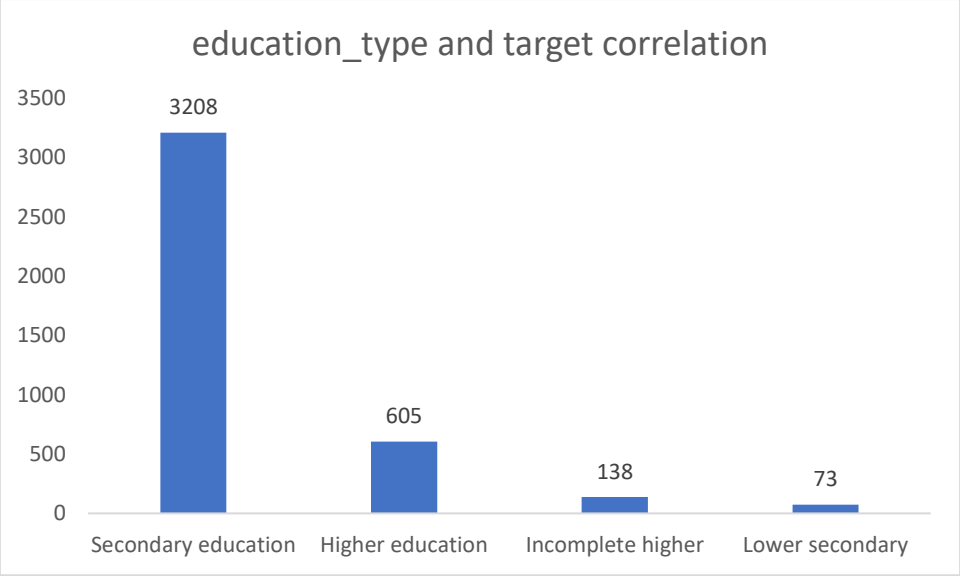
5. Identify top correlations for clients with payment difficulties and all other cases with the target variable.

❖ Correlation between applicant's qualification and target variable

Pivot table:

TARGET	1
Row Labels	**Count of TARGET**
Secondary education	3208
Higher education	605
Incomplete higher	138
Lower secondary	73
Grand Total	**4024**

Graphical Representation:



Insight Gained:

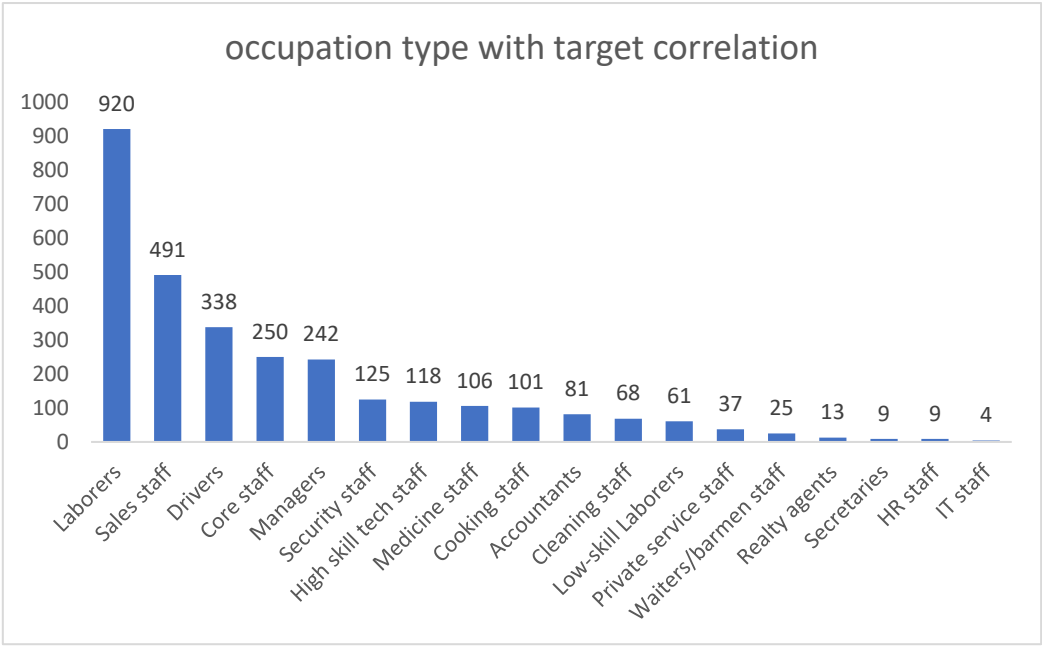
After the analysis, it has been found that most likely the defaulters will be the people with secondary education.

❖ Correlation between occupation type and target

Pivot table:

TARGET	1
Row Labels	**Count of TARGET**
Laborers	920
Sales staff	491
Drivers	338
Core staff	250
Managers	242
Security staff	125
High skill tech staff	118
Medicine staff	106
Cooking staff	101
Accountants	81
Cleaning staff	68
Low-skill Laborers	61
Private service staff	37
Waiters/barmen staff	25
Realty agents	13
Secretaries	9
HR staff	9

Graphical Representation:



Insight Gained:

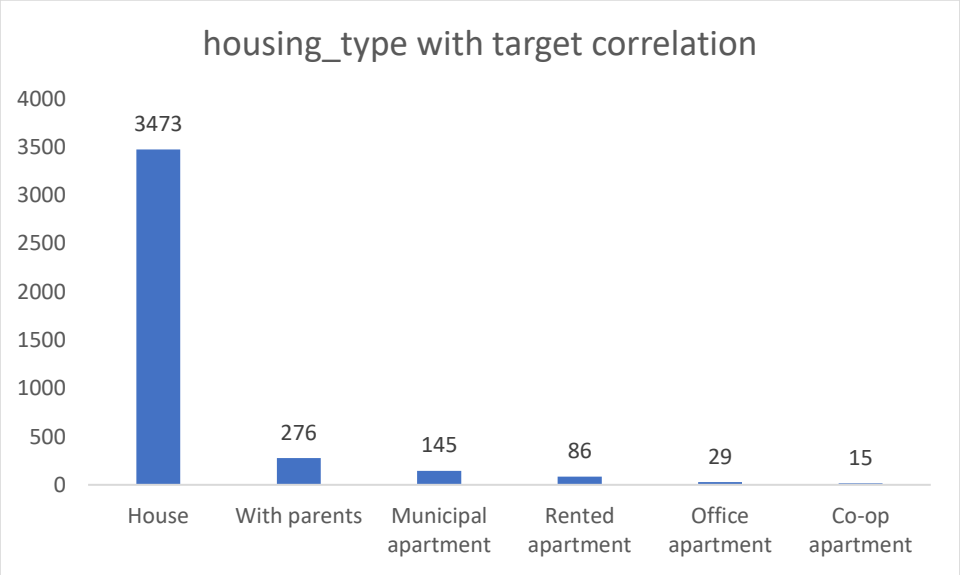
After the analysis, it has been found that most likely the defaulters will be the people who are laborers. I have not included NG (with whom I have replaced blanks in this column).

❖ Correlation between housing type of applicants and target

Pivot table:

TARGET	1
Row Labels	**Count of TARGET**
House	3473
With parents	276
Municipal apartment	145
Rented apartment	86
Office apartment	29
Co-op apartment	15
Grand Total	**4024**

Graphical representation:



Insight Gained:

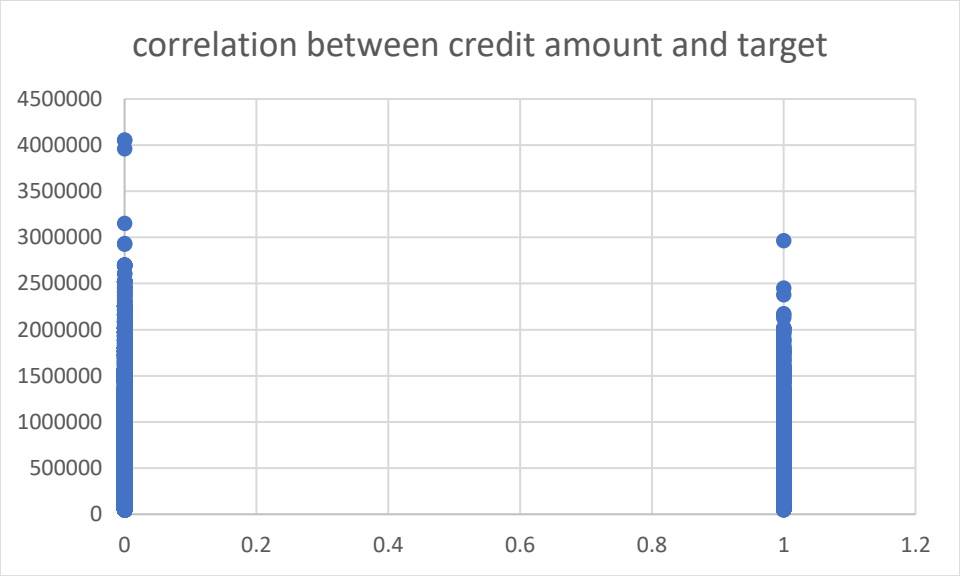
After the analysis, it has been found that people who have house will most likely to be defaulters. Although, for the people with other housing types, the possibility of them being defaulters are very low as per the above chart.

❖ Correlation between credit and target

I have used CORREL function in excel to find the correlation coefficient between amount of credit and target. The correlation coefficient is -0.03246. This means that a very weak negative correlation exists between them. Graphical representation will provide more clarity on their relationship.

correlation between credit amount and target =	-0.03246

Graphical representation:



Insight Gained:

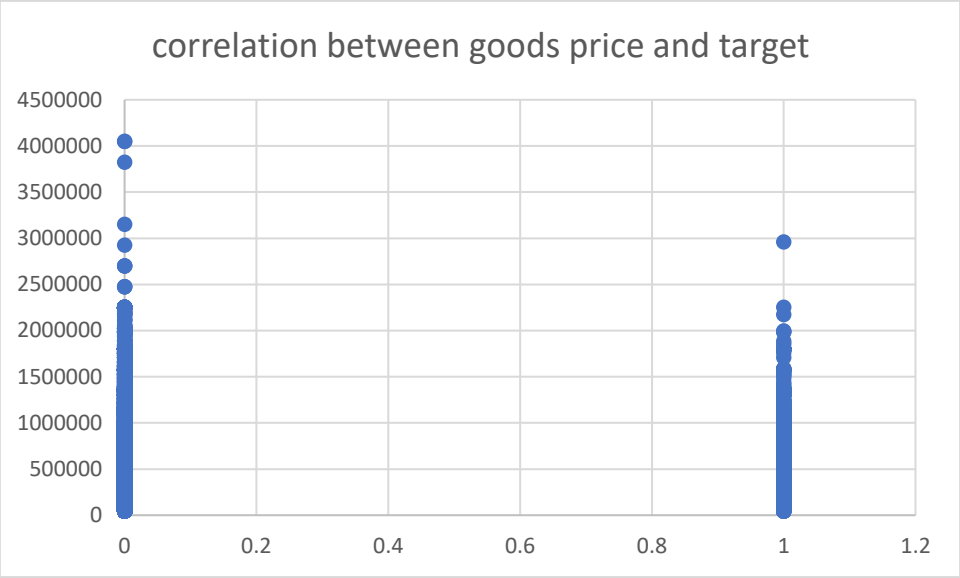
After plotting the scatterplot between amount of credit and target, it is clear that no linear relationship exists between them.

❖ Correlation between goods price and target

By using CORREL function, I have calculated correlation coefficient between goods price and target and it is -0.04131. This means a very weak negative relationship exist between them. Graphical representation will provide more clarity on their relationship.

Correlation between goods price and target =	-0.04131

Graphical representation:



Insight Gained:

As per the above scatterplot, no linear relationship exists between goods price and target.

Insights:

After doing the analysis, various insights has been found like:

1) After analyzing the data, we can predict whether a person will be able to repay the loan or not.
2) The people who are most likely to face problems in repaying the loan will be laborers.
3) Most likely the defaulters will be the people with secondary education.
4) It has been found that people who have house will most likely be defaulters (maybe due to EMI etc).
5) It has been found that people who are married will most likely be defaulters(maybe due to house loan, consumer loan etc).

Result

1) By doing an analysis of a loan application data, I have learned how to handle missing data. I learned that not every missing values needs to be removed from the dataset, as it can result in the loss of some important information required for the analysis.
2) Identify outliers present in the dataset. I also learned that after careful considerations and understanding of the context, we should decide whether to remove outliers from the data or keep them. In this case, I have kept them as they provided me a more comprehensive understanding of the loan scenarios.
3) This project helps me to understand the importance of data imbalance. It helps me to understand the skewed distributions of data for various variables in the given dataset.
4) It introduced me to the terms like univariate analysis, bivariate analysis and segmented univariate analysis and their importance.

Overall, this project level up my knowledge on excel functions and visualizations that can be used to draw various insights from a financial data and helped to mitigate potential risks.

Drive Link

https://drive.google.com/drive/folders/1T7wnR0BTeMw8aCMTxppX3lffyzER-Ztl?usp=sharing