Final Project-2

PROJECT 6

Bank Loan Case Study

RAKSHA NAYAK

Project Description: This Project "Bank Loan Case Study" helps to identify patterns that indicate if a customer will have difficulty paying their instalments. 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. We would be helping financial company to understand the **key factors behind loan default** so it can make better decisions about loan approval.

We would be using **Exploratory Data Analysis (EDA)** and **Risk Analytics** to analyse patterns in the data and ensure that capable applicants are not rejected by recognizing **False Positive and False Negative**.

Approach: There are 2 data files provided which are very huge. So, following would be the approach to follow.

1. Download and understanding the data files, their attributes and meanings.
2. Data Cleaning both the data files
3. Merging the data files for the analysis
4. Data Analysis: Analysing data and summarizing the insights and knowledge gained during the project. Also highlighting key findings and meaningful trends or patterns discovered.
5. Data Visualization: Plotting charts and graphs for better visualization
6. Results: Describing what we have achieved through the project and how it has contributed to our understanding of the Bank Loan Case Study.

Data Files understanding:

1st File: application_data.csv – Contains details about the current loan applications

- Total Columns: 122
- Total records: 49999

2nd File: previous_application.csv – Contains information about previous loan applications

- Total Columns: 37
- Total records: 49999

The dataset after analysis with answers, insights and visualization is,

previous_application dataset after cleaning,

https://docs.google.com/spreadsheets/d/1mmrtNDNblwXe7RRtr7iYWuNwCVIMpW4e/edit?usp=sharing&ouid=108154584635151678812&rtpof=true&sd=true

application_data dataset which consists all the tasks and merged file,

https://docs.google.com/spreadsheets/d/1hwEJ0oO13jqmuLt6ya0fmSjwh8NeVFqN/edit?usp=sharing&ouid=108154584635151678812&rtpof=true&sd=true

Tech stack used: Microsoft Excel Version 2407, 2019 – Excel is a spreadsheet editor developed by Microsoft. It features calculation or computation capabilities, graphing tools, pivot tables etc.

Data Analytics Tasks:

A. **Identify Missing Data and Deal with it Appropriately:** As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

- **Task:** Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.
- **Hint:** Utilize Excel functions like COUNT, ISBLANK, and IF to identify missing data. Consider using functions like AVERAGE or MEDIAN for imputation or other appropriate methods available in Excel.
- **Graph suggestion:** Create a bar chart or column chart to visualize the proportion of missing values for each variable.

1. **Omission/Dropping Columns/Removing unwanted columns:** To make analysis easy, we need to first understand what questions we are supposed to answer and which columns are needed to support it and if columns are **missing more than 40% of the values**. Once that is clear, we can proceed and drop columns which are not useful for that particular analysis in order to reduce the size of the dataset and avoid confusions.

In application_data file: We are dropping below columns with missing values more than 40%. Have used countblank() function of excel to calculate missing values for each column.

Output and Visual Representation:

Total Rows with values	49999
40% of total rows	19999.6
Total Columns below 40% of data/value:	49
Total columns above 40% of dat/values	73
Total Columns	122

Columns	OWN_CAR_AGE	EXT_SOURCE_1	APARTMENTS_AVG	BASEMENTAREA_AVG
Number of Blank rows	32950	28172	25385	29199
Status	Drop	Drop	Drop	Drop

Columns	YEARS_BEGINEXPLUATATION_AVG	YEARS_BUILD_AVG	COMMONAREA_AVG
Number of Blank rows	24394	33239	34960
Status	Drop	Drop	Drop

Columns	ELEVATORS_AVG	ENTRANCES_AVG	FLOORSMAX_AVG	FLOORSMIN_AVG
Number of Blank rows	26651	25195	24875	33894
Status	Drop	Drop	Drop	Drop

Columns	LANDAREA_AVG	LIVINGAPARTMENTS_AVG	LIVINGAREA_AVG
Number of Blank rows	29721	34226	25137
Status	Drop	Drop	Drop

Columns	NONLIVINGAPARTMENTS_AVG	NONLIVINGAREA_AVG	APARTMENTS_MODE
Number of Blank rows	34714	27572	25385
Status	Drop	Drop	Drop

Columns	BASEMENTAREA_MODE	YEARS_BEGINEXPLUATATION_MODE	YEARS_BUILD_MODE
Number of Blank rows	29199	24394	33239
Status	Drop	Drop	Drop

Columns	COMMONAREA_MODE	ELEVATORS_MODE	ENTRANCES_MODE
Number of Blank rows	34960	26651	25195
Status	Drop	Drop	Drop

Columns	FLOORSMAX_MODE	FLOORSMIN_MODE	LANDAREA_MODE	LIVINGAPARTMENTS_MODE
Number of Blank rows	24875	33894	29721	34226
Status	Drop	Drop	Drop	Drop

Columns	LIVINGAREA_MODE	NONLIVINGAPARTMENTS_MODE	NONLIVINGAREA_MODE
Number of Blank rows	25137	34714	27572
Status	Drop	Drop	Drop

Columns	APARTMENTS_MEDI	BASEMENTAREA_MEDI	YEARS_BEGINEXPLUATATION_MEDI
Number of Blank rows	25385	29199	24394
Status	Drop	Drop	Drop

Columns	YEARS_BUILD_MEDI	COMMONAREA_MEDI	ELEVATORS_MEDI	ENTRANCES_MEDI
Number of Blank rows	33239	34960	26651	25195
Status	Drop	Drop	Drop	Drop

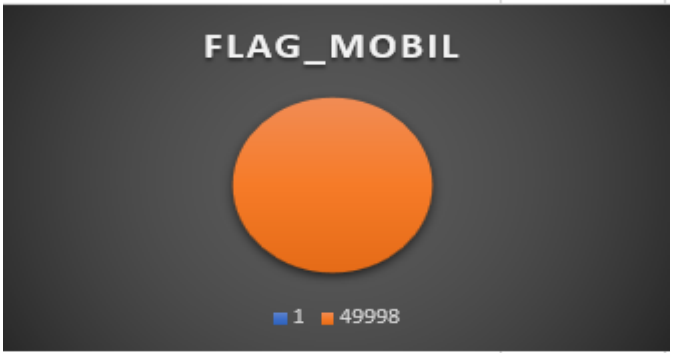
Columns	FLOORSMAX_MEDI	FLOORSMIN_MEDI	LANDAREA_MEDI	LIVINGAPARTMENTS_MEDI
Number of Blank rows	24875	33894	29721	34226
Status	Drop	Drop	Drop	Drop

Columns	LIVINGAREA_MEDI	NONLIVINGAPARTMENTS_MEDI	NONLIVINGAREA_MEDI
Number of Blank rows	25137	34714	27572
Status	Drop	Drop	Drop

Columns	FONDKAPREMONT_MODE	HOUSETYPE_MODE	TOTALAREA_MODE	WALLSMATERIAL_MODE	EMERGENCYSTATE_MODE
Number of Blank rows	34191	25075	24148	25459	23698
Status	Drop	Drop	Drop	Drop	Drop

Other Column to drop on careful observation and analysis, FLAG_MOBIL has all values as 1 and only 1 value as 0. So, we can drop this column as that is not going to be of much use.

FLAG_MOBIL	
1	49998
0	1



Other columns with FLAGS and their correlation with target: As some flags have weak correlation with target due to correlation value of 0. We can drop them. Dropping 25 columns and keeping FLAG_OWN_CAR and FLAG_OWN_REALTY, which might be useful in the analysis.

KEEPING THIS	
FLAG_OWN_CAR	
N	32949
Y	17050
Correlation of target and FLAG_EMAIL	-0.00131181
FLAG_EMAIL	
0	47216
1	2783
Correlation of target and FLAG_DOCUMENT_2	0.00975047
FLAG_DOCUMENT_2	
0	49997
1	2
Correlation of target and FLAG_DOCUMENT_5	-0.00308026
FLAG_DOCUMENT_5	
0	49214
1	785
Correlation of target and FLAG_DOCUMENT_8	-0.00867213
FLAG_DOCUMENT_8	
0	45961
1	4038
Correlation of target and FLAG_DOCUMENT_11	-0.00242768
FLAG_DOCUMENT_11	
0	49786
1	213
Correlation of target and FLAG_DOCUMENT_14	-0.00880393
FLAG_DOCUMENT_14	
0	49841
1	158
Correlation of target and FLAG_DOCUMENT_17	-0.00088208
FLAG_DOCUMENT_9	
0	49984
1	15
Correlation of target and FLAG_DOCUMENT_20	-0.00352584
FLAG_DOCUMENT_9	
0	49973
1	26

KEEPING THIS	
FLAG_OWN_REALTY	
N	15308
Y	34691
Correlation of target and FLAG_WORK_PHONE	0.006765545
FLAG_CONT_MOBILE	
0	101
1	49898
Correlation of target and FLAG_DOCUMENT_3	0.045050228
FLAG_DOCUMENT_3	
0	14387
1	35612
Correlation of target and FLAG_DOCUMENT_6	-0.003080257
FLAG_DOCUMENT_6	
0	45664
1	4335
Correlation of target and FLAG_DOCUMENT_9	-0.000990516
FLAG_DOCUMENT_9	
0	49815
1	184
Correlation of target and FLAG_DOCUMENT_12	0
FLAG_DOCUMENT_12	
0	49999
1	0
Correlation of target and FLAG_DOCUMENT_15	-0.008477638
FLAG_DOCUMENT_15	
0	49958
1	41
Correlation of target and FLAG_DOCUMENT_18	-0.00578216
FLAG_DOCUMENT_9	
0	49574
1	425
Correlation of target and FLAG_DOCUMENT_21	-0.005769845
FLAG_DOCUMENT_9	
0	49980
1	19

Correlation of target and FLAG_EMP_PHONE	0.04140843
FLAG_EMP_PHONE	
0	8926
1	41073
Correlation of target and FLAG_PHONE	-0.032679413
FLAG_PHONE	
0	36113
1	13886
Correlation of target and FLAG_DOCUMENT_4	-0.003970682
FLAG_DOCUMENT_4	
0	49990
1	9
Correlation of target and FLAG_DOCUMENT_7	-0.004389842
FLAG_DOCUMENT_7	
0	49988
1	11
Correlation of target and FLAG_DOCUMENT_10	-0.001323455
FLAG_DOCUMENT_10	
0	49998
1	1
Correlation of target and FLAG_DOCUMENT_13	-0.014224859
FLAG_DOCUMENT_13	
0	49838
1	161
Correlation of target and FLAG_DOCUMENT_16	-0.009107982
FLAG_DOCUMENT_16	
0	49498
1	501
Correlation of target and FLAG_DOCUMENT_19	0.000505091
FLAG_DOCUMENT_9	
0	49964
1	35
Correlation of target and FLAG_WORK_PHONE	0.021302134
FLAG_WORK_PHONE	
0	40036
1	9963

columns EXT_SOURCE_2 and EXT_SOURCE_3 have weak correlation (no linear correlation) with target. These 2 columns can be dropped.

Correlation of taget and EXT_SOURCE_2	-0.158424274
Correlation of taget and EXT_SOURCE_3	-0.181275965

So total columns now, 73-1(FLAG_MOBIL)-25-2(EXT_SOURCE) = 45

In previous_application file: We are dropping below columns with missing values more than 40%. Have used countblank() function of excel to calculate missing values for each column.

Total Rows with values	49999
40% of total rows	19999.6
Total Columns below 40% of data	5
Total columns above 40% of dat/v	32
Total Columns	37

Columns	AMT_DOWN_PAYMENT	RATE_DOWN_PAYMENT	RATE_INTEREST_PRIMARY
Number of Blank rows	25198	25198	49834
Status	Drop	Drop	Drop

Columns	RATE_INTEREST_PRIVILEGED	NAME_TYPE_SUITE
Number of Blank rows	49834	24243
Status	Drop	Drop

On further careful observation and analysis, we can drop few columns which have no importance in loans or repayments.

- WEEKDAY_APPR_PROCESS_START
- HOUR_APPR_PROCESS_START
- FLAG_LAST_APPL_PER_CONTRACT
- NFLAGS_LAST_APPL_IN_DAY

Few Unwanted Columns	
WEEKDAY_APPR_PROCESS_START	
HOUR_APPR_PROCESS_START	
FLAG_LAST_APPL_PER_CONTRACT	
NFLAG_LAST_APPL_IN_DAY	
Total columns after dropping above columns	28

2. **Handling Duplicate Values:** I used **conditional formatting** to check for duplicates and there are none present.

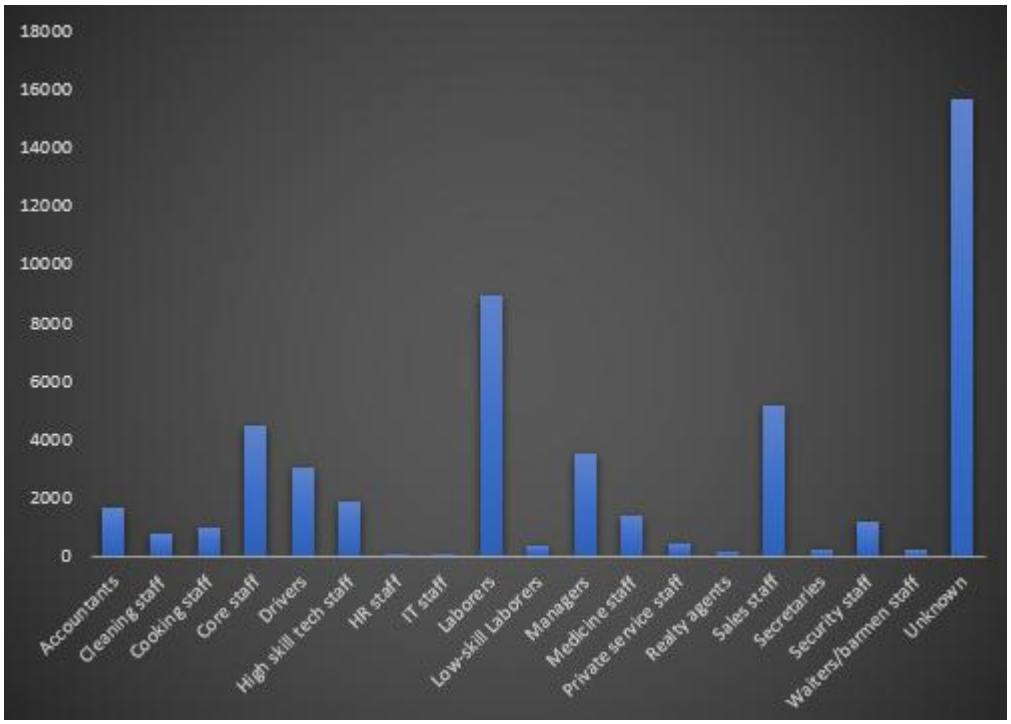
3. **Handling Missing Data**: We need to check if there are any missing values in the dataset. We can either use mean/median/mode based on outliers and type of data (categorical or numerical).

Output and Visual Representation:

In application_data file,

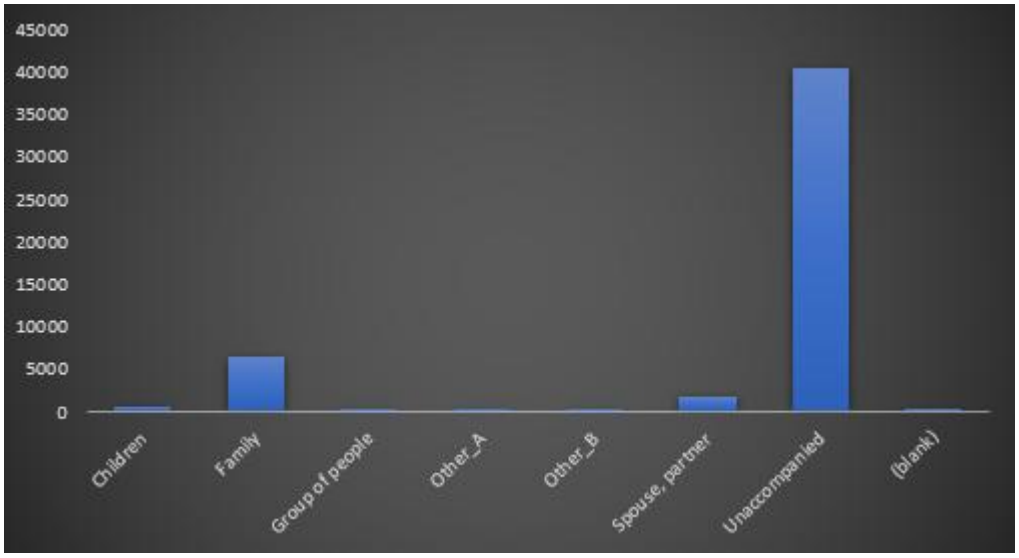
- Column OCCUPATION_TYPE and NAME_TYPE_SUITE has blank/null values, which can be filled **using MODE** as it consists categorical data.

OCCUPATION_TYPE	
Accountants	1621
Cleaning staff	739
Cooking staff	963
Core staff	4434
Drivers	3044
High skill tech staff	1852
HR staff	101
IT staff	80
Laborers	8952
Low-skill Laborers	357
Managers	3489
Medicine staff	1403
Private service staff	447
Realty agents	123
Sales staff	5160
Secretaries	212
Security staff	1140
Waiters/barmen staff	228
Unknown	15654



As OCCUPATION_TYPE has more of unknown values, we are filling up all blank values with “Unknown”.

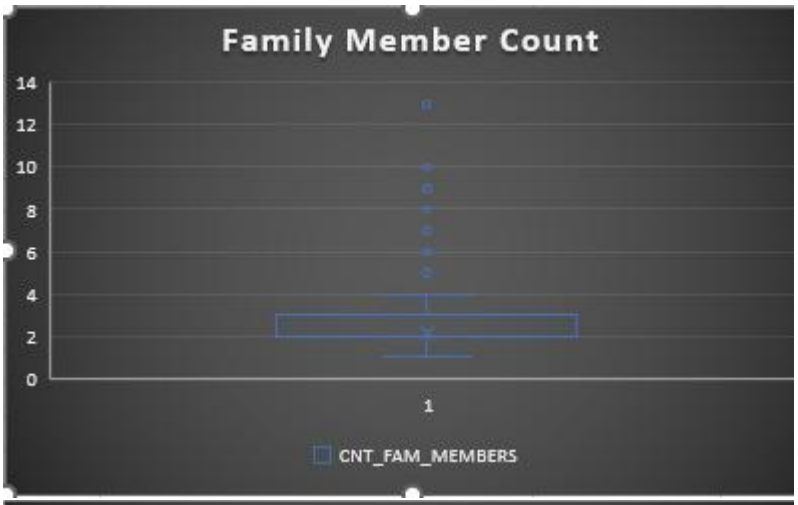
NAME_TYPE_SUITE	
Children	542
Family	6549
Group of people	36
Other_A	137
Other_B	259
Spouse, partner	1849
Unaccompanied	40435
(blank)	192

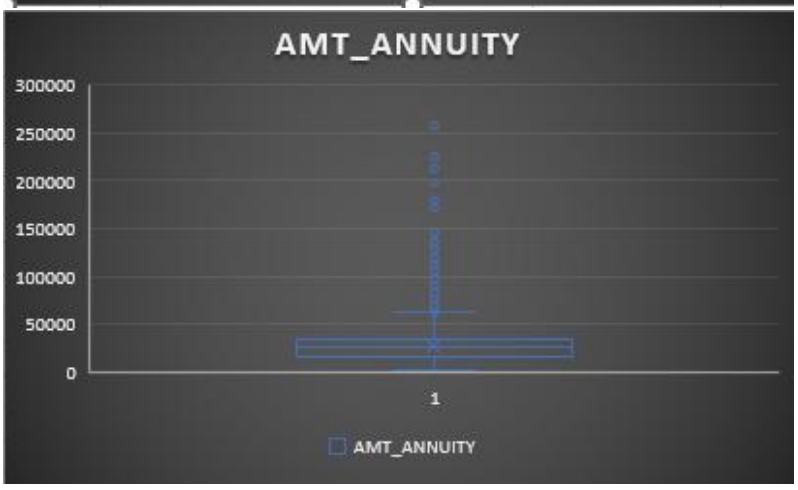


As NAME_TYPE_SUITE has more of Unaccompanied values, we are filling up all blank values with “Unaccompanied”.

- Below columns can be **imputed with median** as they consist outliers
  - AMT_GOODS_PRICE
  - AMT_REQ_CREDIT_BUREAU_HOUR
  - AMT_REQ_CREDIT_BUREAU_DAY
  - AMT_REQ_CREDIT_BUREAU_WEEK
  - AMT_REQ_CREDIT_BUREAU_MON
  - AMT_REQ_CREDIT_BUREAU_QRT
  - AMT_REQ_CREDIT_BUREAU_YEAR
  - OBS_30_CNT_SOCIAL_CIRCLE
  - OBS_60_CNT_SOCIAL_CIRCLE
  - DEF_30_CNT_SOCIAL_CIRCLE
  - DEF_60_CNT_SOCIAL_CIRCLE
  - AMT_ANNUITY
  - CNT_FAM_MEMBERS

COLUMNS	MEDIAN
AMT_GOODS_PRICE	450000
AMT_REQ_CREDIT_BUREAU_HOUR	0
AMT_REQ_CREDIT_BUREAU_DAY	0
AMT_REQ_CREDIT_BUREAU_WEEK	0
AMT_REQ_CREDIT_BUREAU_MON	0
AMT_REQ_CREDIT_BUREAU_QRT	0
AMT_REQ_CREDIT_BUREAU_YEAR	1
OBS_30_CNT_SOCIAL_CIRCLE	0
DEF_30_CNT_SOCIAL_CIRCLE	0
OBS_60_CNT_SOCIAL_CIRCLE	0
DEF_60_CNT_SOCIAL_CIRCLE	0
CNT_FAM_MEMBERS	2
AMT_ANNUITY	24939





We are using Median as the above columns have outliers and median is the best measure of central tendency when we have outliers.

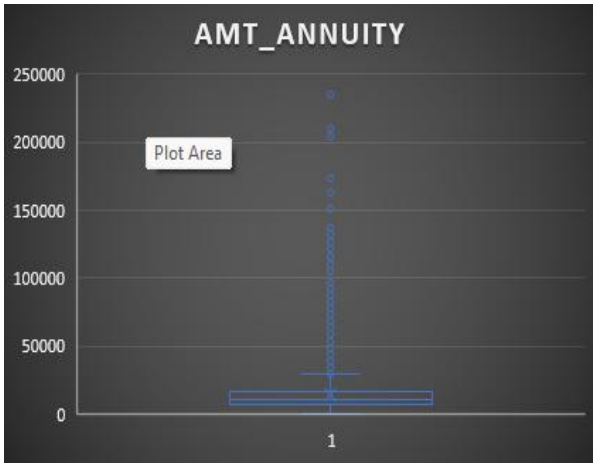
- Below columns can be imputed with mean as they consist no outliers after removing "-"from the values.
  - DAYS_LAST_PHONE_CHANGE

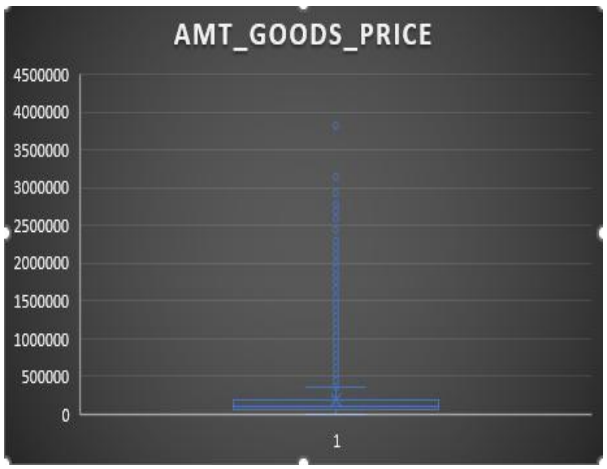
COLUMNS	MEAN
DAYS_LAST_PHONE_CHANGE	964

In previous_application file,

- Column NFLAG_INSURED_ON_APPROVAL has blank/null values, which can be filled using MODE as it consists only 0's and 1's.
- Below columns can be imputed with median as they have outliers
  - AMT_ANNUITY
  - AMT_GOODS_PRICE

COLUMN	MEDIAN
AMT_ANNUITY	10879.92
AMT_GOODS_PRICE	104017.5





- Column CNT_PAYMENT, DAYS_FIRST_DRAWING, DAYS_FIRST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, DAYS_TERMINATION, NFLAG_INSURED_ON_APPROVAL missing values are filled with 0 as their respective NAME_CONTRACT_STATUS shows either Refused or Cancelled or Unused.
- Column PRODUCT_COMBINATION missing values are filled with "Unknown"

4. **Error Correction/Rectification**:
    - In application_data file, below columns have **negative values** which can be corrected.
        - DAYS_BIRTH
        - DAYS_EMPLOYED
        - DAYS_REGISTRATION
        - DAYS_ID_PUBLISH
        - DAYS_LAST_PHONE_CHANGE
    - In previous_application file, below columns have **negative values** which can be corrected.
        - DAYS_DECISION
        - DAYS_FIRST DRAWING
        - DAYS_FIRST_DUE
        - DAYS_LAST_DUE_1ST_VERSION
        - DAYS_LAST_DUE
        - DAYS_TERMINATION

Insight: We have cleaned the data by managing missing values, dropping unwanted columns and handling errors. Summary of the cleaned data is as shown below,

The first dataset, application_data now consists of **45 columns**.
The second dataset, previous_application now consists of **28 columns**.

B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

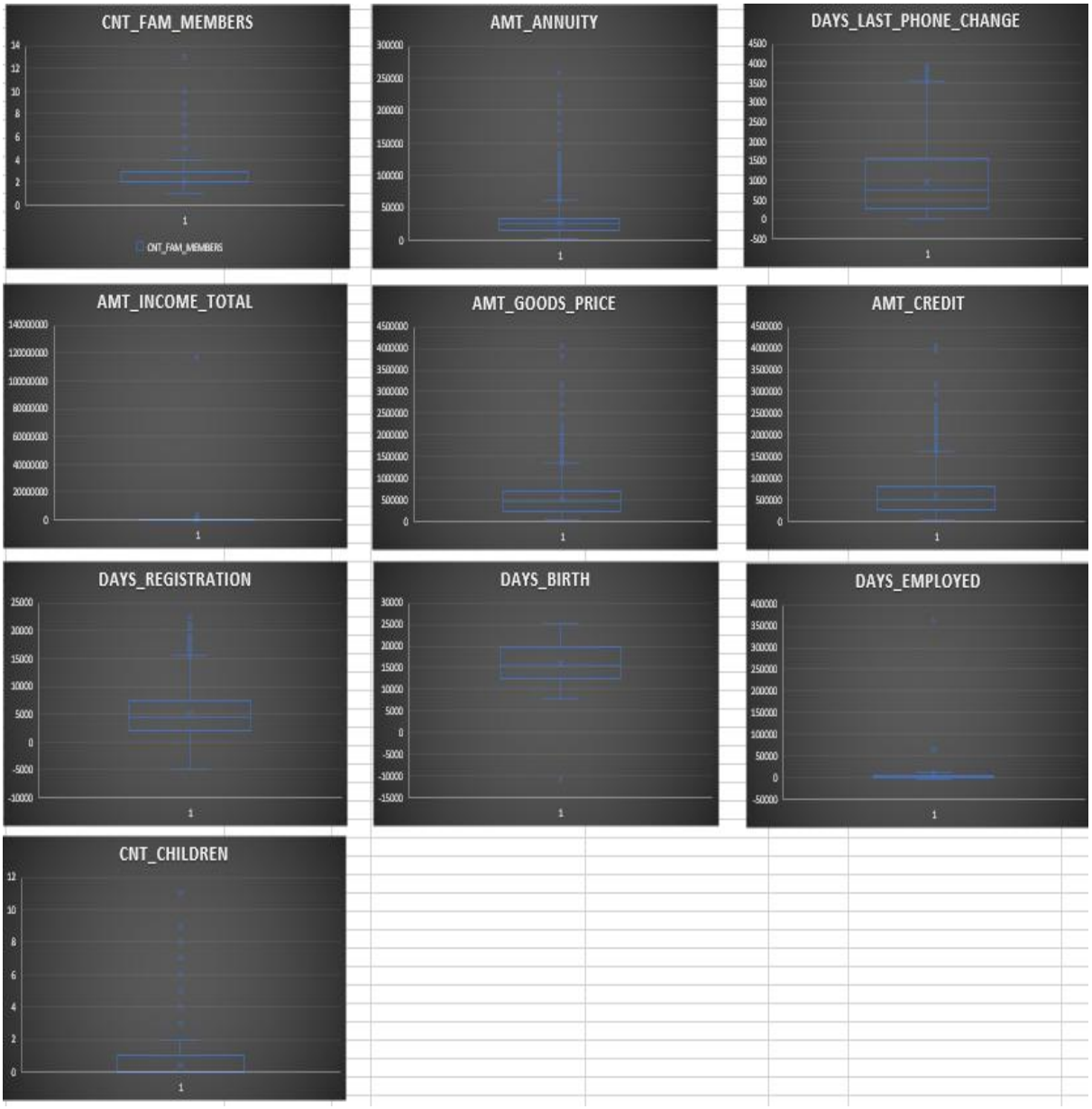
- Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

- Hint: Utilize Excel functions like QUARTILE, IQR, and conditional formatting to identify potential outliers. Consider applying thresholds or business rules to determine if the outliers are valid data points or require further investigation.
- Graph suggestion: Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers.

Output and Visual Representation:

In application_data file,

From the below box plots, we get to know that there are outliers for CNT_FAM_MEMBERS, AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, DAYS_REGISTRATION and others



COLUMN	Q1	Q3	IQR(Q3-Q1)	Lower Bound(Q1-1.5*IQR)	Upper Bound(Q3+1.5*IQR)
CNT_FAM_MEMBERS	2	3	1	0.5	4.5
AMT_ANNUITY	16456.5	34596	18139.5	-10752.75	61805.25
AMT_INCOME_TOTAL	112500	202500	90000	-22500	337500
AMT_GOODS_PRICE	238500	679500	441000	-423000	1341000
AMT_CREDIT	270000	808650	538650	-537975	1616625
DAYS_LAST_PHONE_CHANGE	270	1573	1303	-1684.5	3527.5
DAYS_REGISTRATION	1997.5	7463.5	5466	-6201.5	15662.5
DAYS_BIRTH	12378.5	19644	7265.5	1480.25	30542.25
DAYS_EMPLOYED	933	5718	4785	-6244.5	12895.5
CNT_CHILDREN	0	1	1	-1.5	2.5

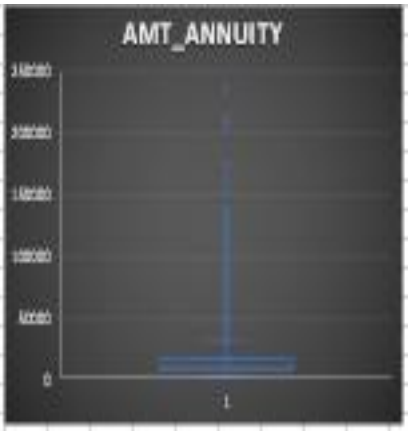
The value that falls above or below the upper bound and lower bound are considered as outliers.

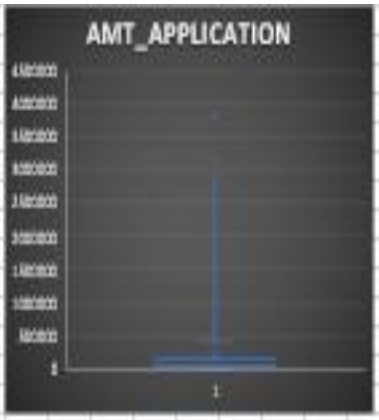
Insight:

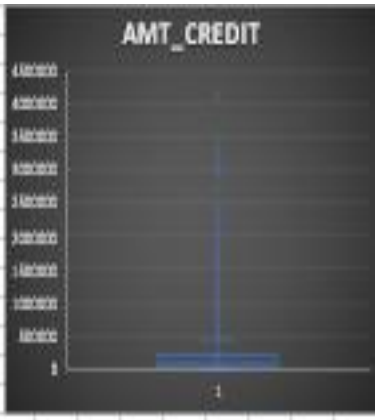
- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRODUCT, CNT_CHILDREN have some outliers.
- DAYS_BIRTH has no outlier, which means the data is reliable and accurate.
- DAYS_EMPLOYED has outlier value of around 350000 days, which is impossible and thus can be stated as an incorrect entry.
- CNT_FAM_MEMBERS has outlier value of more than 45000, which is impossible and thus can be stated as an incorrect entry.
- AMT_INCOME_TOTAL has outlier value of around 110000000, which is impossible and thus can be stated as an incorrect entry.
- DAYS_LAST_PHONE_CHANGE has few outliers.

In previous_application file,

From the below box plots, we get to know that there are outliers for AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE and CNT_PAYMENT.

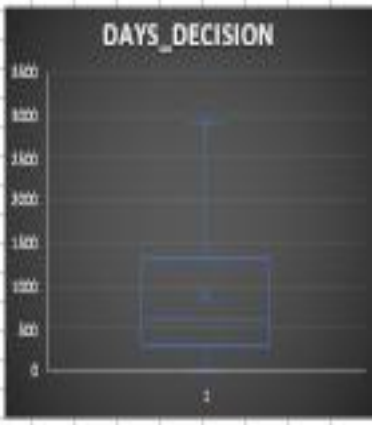

AMT_ANNUITY


AMT_APPLICATION

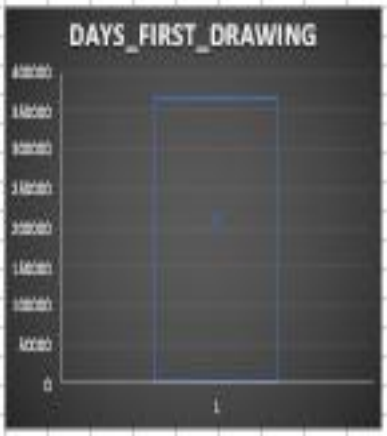

AMT_CREDIT


AMT_GOODS_PRICE


CNT_PAYMENT


DAYS_DECISION

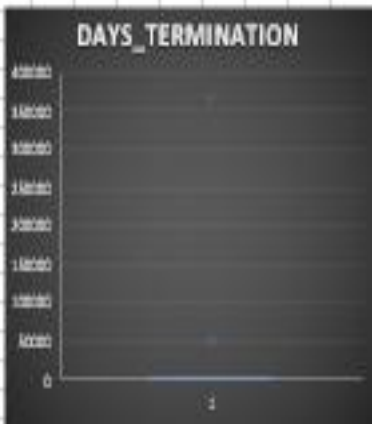

SELLERPLACE_AREA


DAYS_FIRST_DRAWING


DAYS_FIRST_DUE


DAYS_LAST_DUE_1ST_VERSION


DAYS_LAST_DUE


DAYS_TERMINATION

COLUMN	Q1	Q3	IQR(Q3-Q1)	Lower Bound(Q1-1.5*IQR)	Upper Bound(Q3+1.5*IQR)
AMT_ANNUITY	7189.74	16256.16	9066.42	-6409.89	29855.79
AMT_APPLICATION	22045.5	180000	157954.5	-214886.25	416931.75
AMT_CREDIT	26055	198105.8	172050.75	-232021.125	456181.875
AMT_GOODS_PRICE	63663.75	180000	116336.25	-110840.625	354504.375
CNT_PAYMENT	0	14	14	-21	35
DAYS_DECISION	292	16256.16	15964.16	-23654.24	40202.4
SELLERPLACE_AREA	1	100	99	-147.5	248.5
DAYS_FIRST_DRAWING	0	365243	365243	-547864.5	913107.5
DAYS_FIRST_DUE	0	1178	1178	-1767	2945
DAYS_LAST_DUE_1ST_VERSION	0	1032	1032	-1548	2580
DAYS_LAST_DUE_1ST_VERSION	0	1606	1606	-2409	4015
DAYS_TERMINATION	0	1634	1634	-2451	4085

The value that falls above or below the upper bound and lower bound are considered as outliers.

Insight:

- AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have large number of outliers.
- CNT_PAYMENT have few outliers.
- DAYS_DECISION has few outliers indicating that time taken for decision was more.

Data Summary: We have cleaned the data by managing missing values, errors and replacing the outliers with actual values.

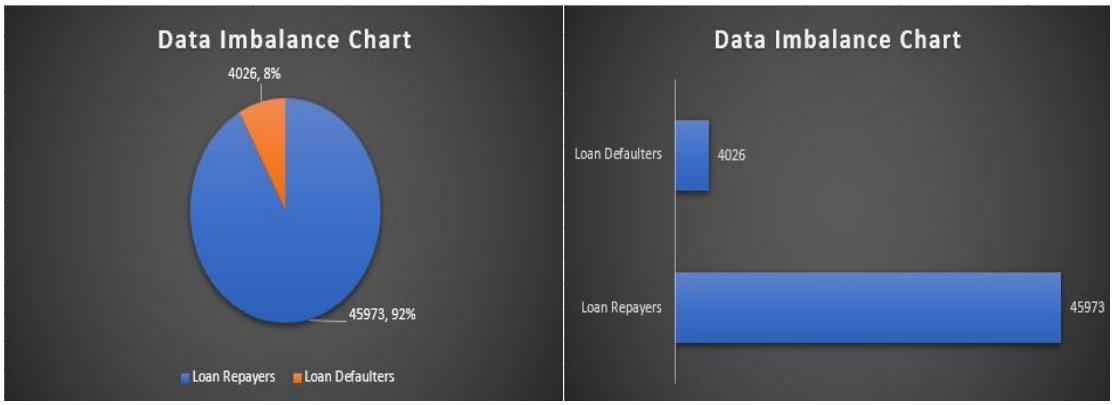
C. Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

- Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
- Hint: Utilize Excel functions like COUNTIF and SUM to calculate the proportions of each class. Compare the class frequencies to assess data imbalance.
- Graph suggestion: Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.

Output:

Loan Repayers	45973	91.94783896
Loan Defaulters	4026	8.052161043
Total	49999	

Visual Representation:



Approach: "TARGET" column has an entry of 2 values, 1 and 0.

- 1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample – Loan Defaulters
- 0 - all other cases – Loan Repayers

We will use excels countif function to get the count of loan repayers and defaulters and calculate percentage.

Total Loan Repayers = 45973 = 92%
Total Loan Defaulters = 4026 = 8%
Ratio of Loan Repayers to Loan Defaulters = 600: 53

Insight: The above ratio clearly indicates the data imbalance and can affect the accuracy of the analysis.

D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

- Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.
- Hint: Utilize Excel functions like COUNT, AVERAGE, MEDIAN, and statistical functions for descriptive analysis. Utilize Excel features like filters, sorting, and pivot tables for segmented and bivariate analysis.
- Graph suggestion: Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

Methodology Adopted: The output is different for categorical and numerical data columns.

- Pie chart, Column chart and Bar chart for describing the categorical variable
- Histogram, Column chart and Box plot for describing the numerical variable.

Approach: **Univariate Analysis** - Let's explore each variable in a data set, separately. Let's looks at the range of values, as well as the central tendency of the values, descriptive statistics and describe the pattern of response to the variable.

Segmented Univariate Analysis - Segmented univariate analysis is a data analysis technique that involves dividing data into segments and then analysing each segment separately

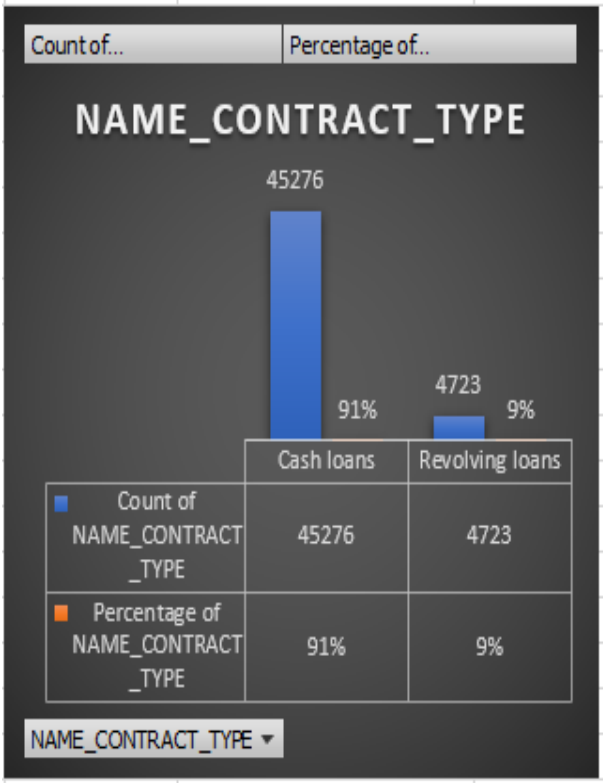
Bivariate Analysis - Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables, for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association.

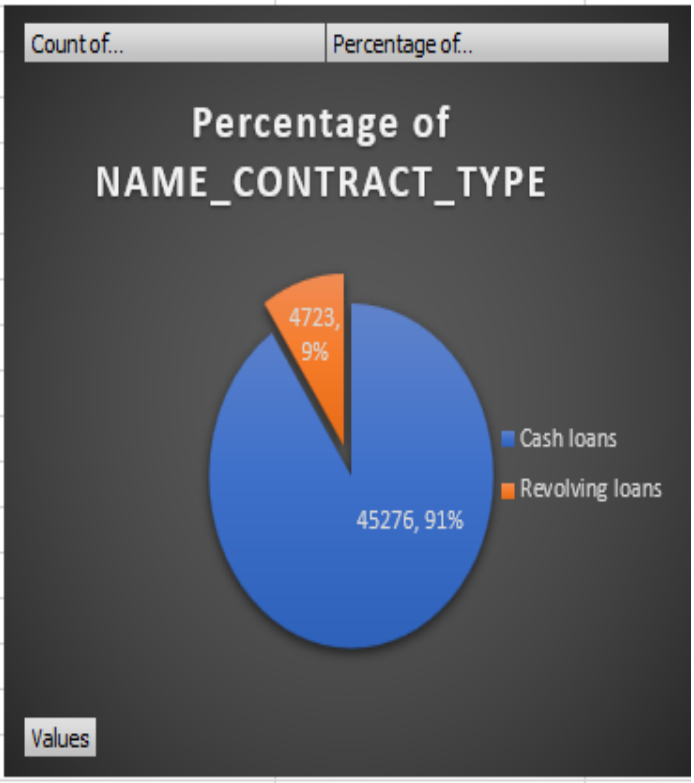
Let's merge application_data and previous_application together to continue with the easy analysis using Power Query.

Output and Visual Representation:

Univariate Analysis: NAME_CONTRACT_TYPE

CATEGORICAL VARIABLE ANALYSIS		
Row Labels	Count of NAME_CONTRA	Percentage of NAME_CONTRACT_TYPE
Cash loans	45276	91%
Revolving loan	4723	9%
Grand Total	**49999**	**100%**

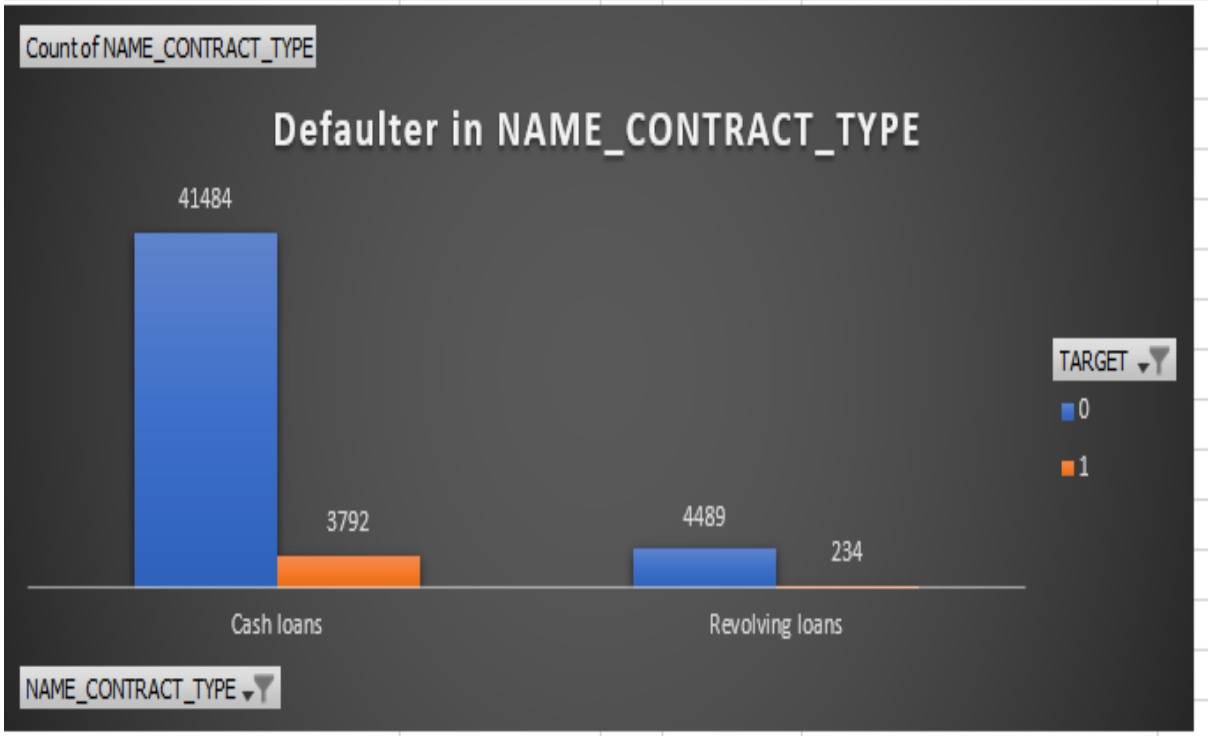




The number of cash loans taken are much higher than revolving loans

Segmented Univariate Analysis: NAME_CONTRACT_TYPE with TARGET

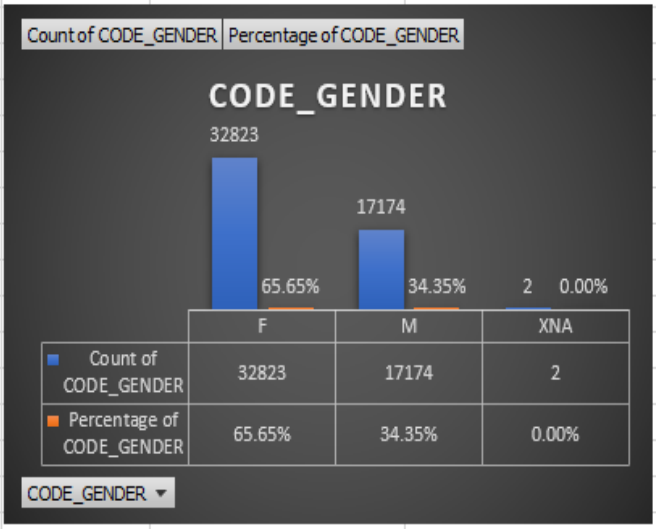
Count of NAME_CONTRACT_TYPE	Column Labels			Percentage of defaulters in
Row Labels	0	1	Grand Total	each category
Cash loans	41484	3792	45276	8%
Revolving loans	4489	234	4723	5%
Grand Total	**45973**	**4026**	**49999**	**13%**

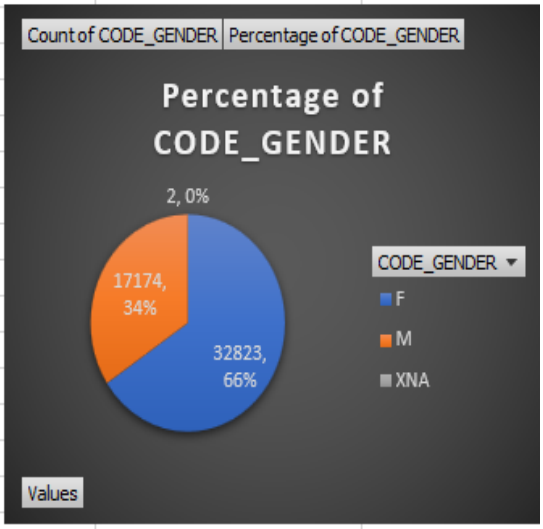


The default Percentage for cash loans is 8% and revolving loans is 5%

Univariate Analysis: CODE_GENDER

Row Labels	Count of CODE_GENDER	Percentage of CODE_GENDER
F	32823	65.65%
M	17174	34.35%
XNA	2	0.00%
Grand Total	**49999**	**100.00%**

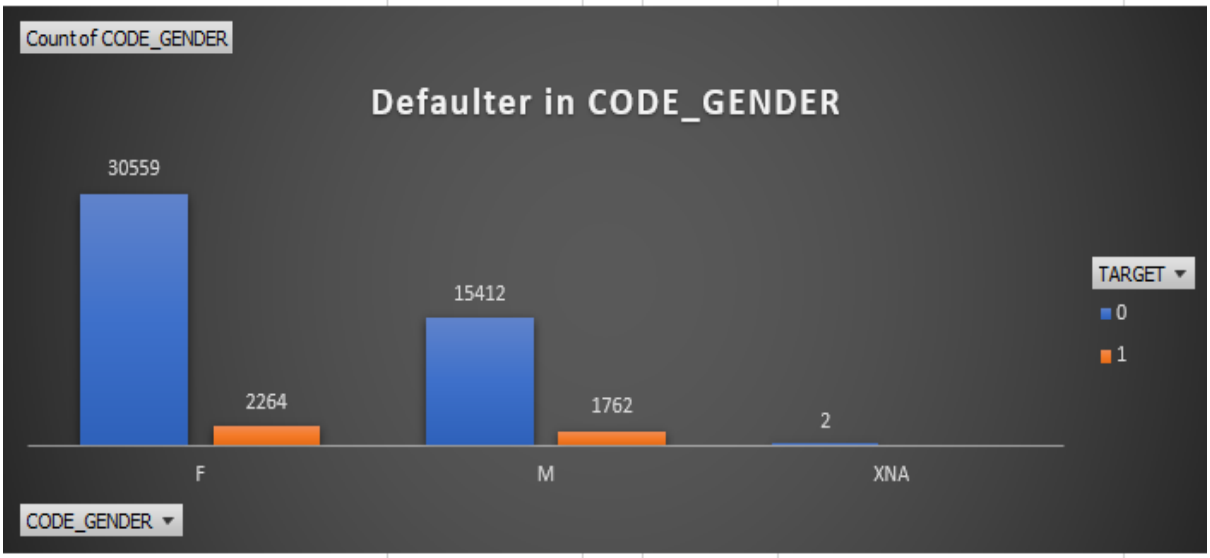




Female customers are more than the male customers

Segmented Univariate Analysis: CODE_GENDER with TARGET

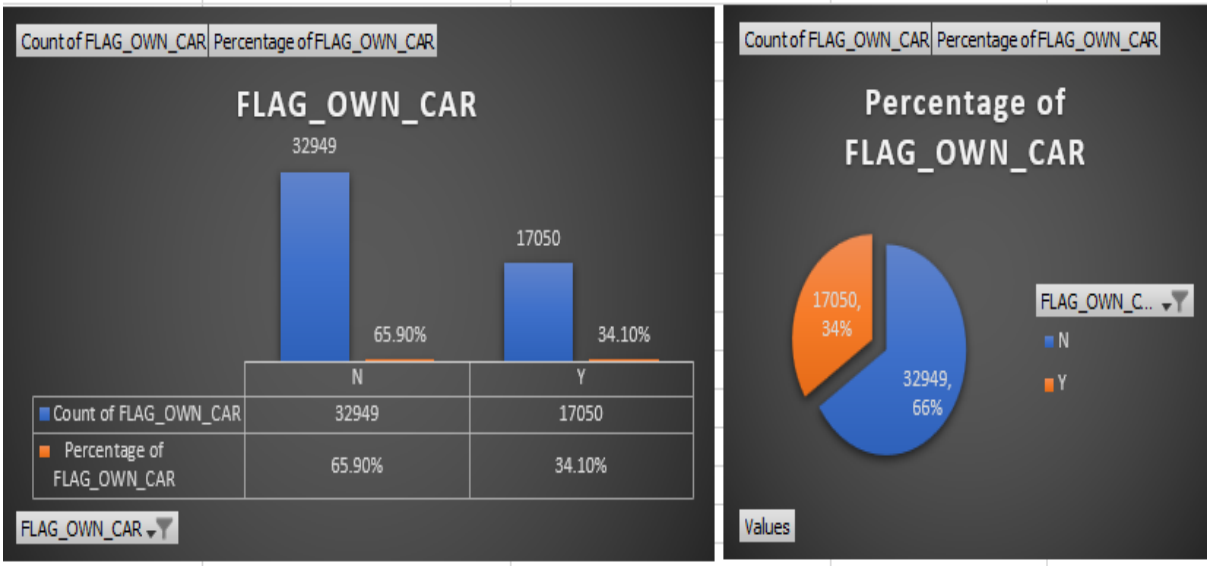
Count of CODE_GENDER	Column Labels			Percentage of defaulters in
Row Labels	0	1	Grand Total	each category
F	30559	2264	32823	7%
M	15412	1762	17174	10%
XNA	2		2	0%
Grand Total	**45973**	**4026**	**49999**	**17%**



The male percentage of defaulting/not returning loan is 10% and female percentage is 7%

Univariate Analysis: FLAG_OWN_CAR

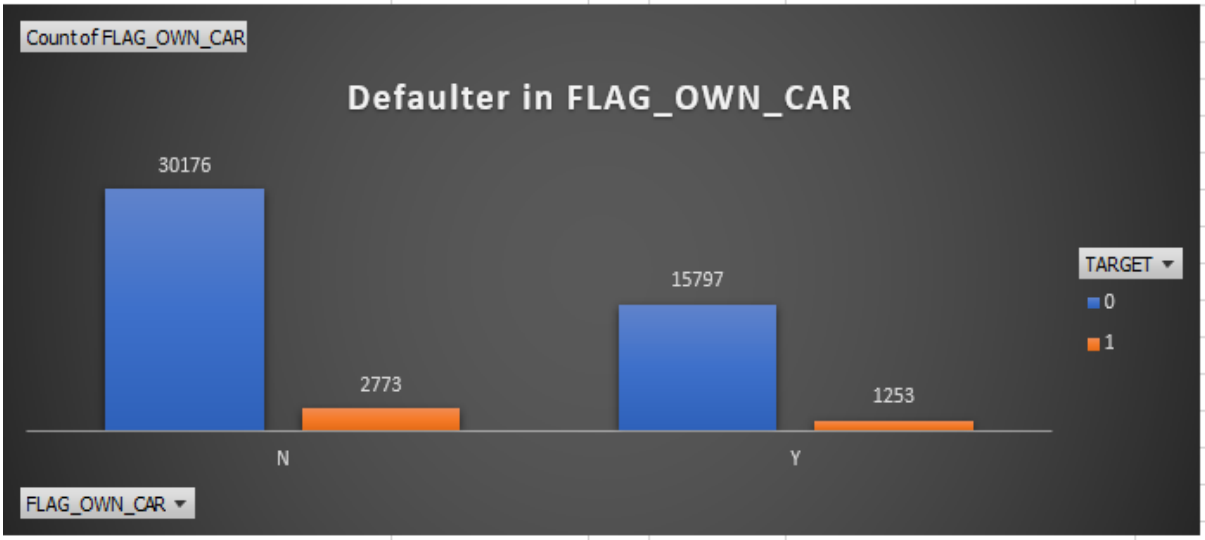
Row Labels	Count of FLAG_OWN_CAR	Percentage of FLAG_OWN_CAR
N	32949	65.90%
Y	17050	34.10%
Grand Total	**49999**	**100.00%**



The percentage of customers who doesn't own car are more than those who own.

Segmented Univariate Analysis: FLAG_OWN_CAR with TARGET

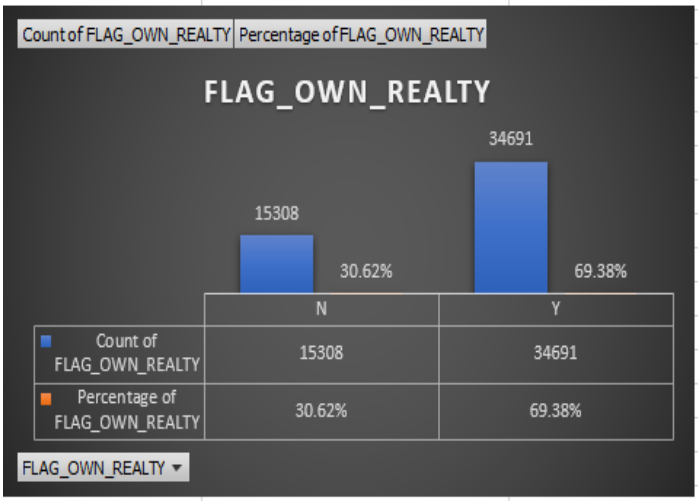
Count of FLAG_OWN_CAR	Column Labels			Percentage of defaulters in
Row Labels	0	1	Grand Total	each category
N	30176	2773	32949	8%
Y	15797	1253	17050	7%
Grand Total	**45973**	**4026**	**49999**	**16%**

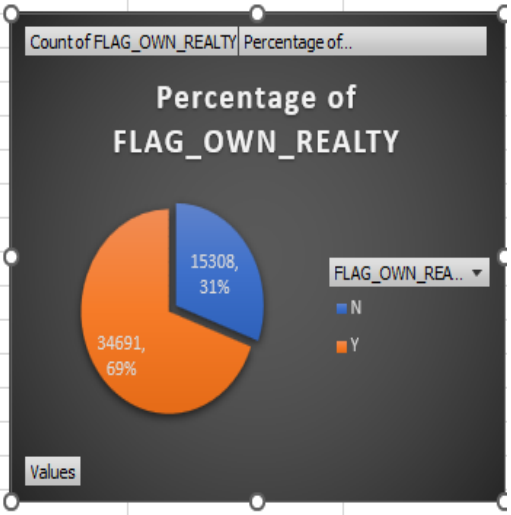


The customers who are defaulters, most of them do not own a car

Univariate Analysis: FLAG_OWN_REALTY

Row Labels	Count of FLAG_OWN_REALTY	Percentage of FLAG_OWN_REALTY
N	15308	30.62%
Y	34691	69.38%
Grand Total	**49999**	**100.00%**

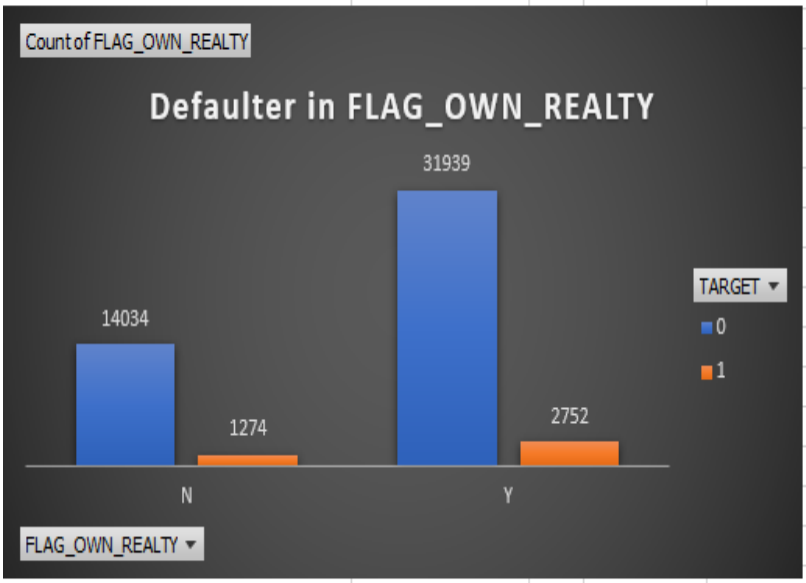




The customers who own real estate are more than those who don't own.

Segmented Univariate Analysis: FLAG_OWN_REALTY with TARGET

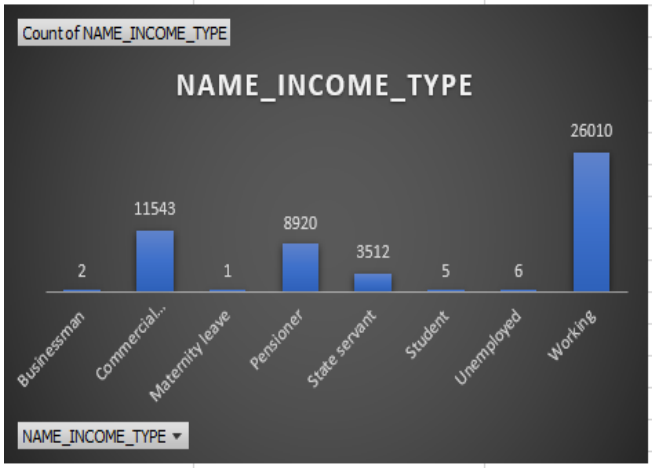
Count of FLAG_OWN_REALTY	Column Labels			Percentage of defaulters in
Row Labels	0	1	Grand Total	each category
N	14034	1274	15308	8%
Y	31939	2752	34691	8%
Grand Total	**45973**	**4026**	**49999**	**16%**

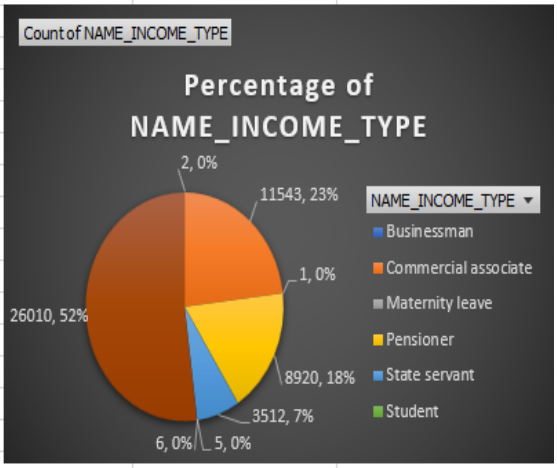


The defaulting rate of both, who owns real budget and doesn't own is almost same. Thus, we can conclude that there is no correlation between owing real estate and defaulting a loan.

Univariate Analysis: NAME_INCOME_TYPE

Row Labels	Count of NAME_INCOME_TYPE	Percentage
Businessman	2	0%
Commercial associate	11543	23%
Maternity leave	1	0%
Pensioner	8920	18%
State servant	3512	7%
Student	5	0%
Unemployed	6	0%
Working	26010	52%
Grand Total	**49999**	100%

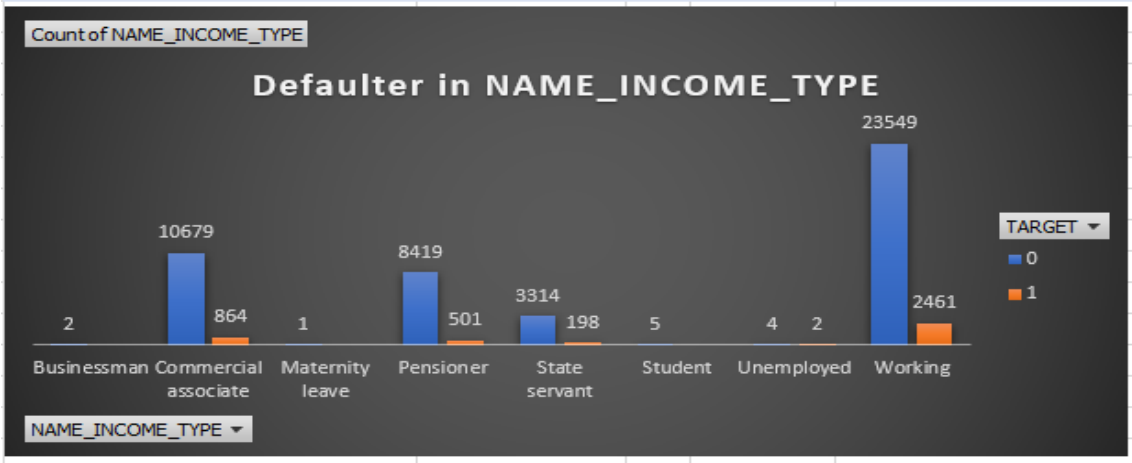




More than 50% of customers who have taken loan belong to working class

Segmented Univariate Analysis: NAME_INCOME_TYPE with TARGET

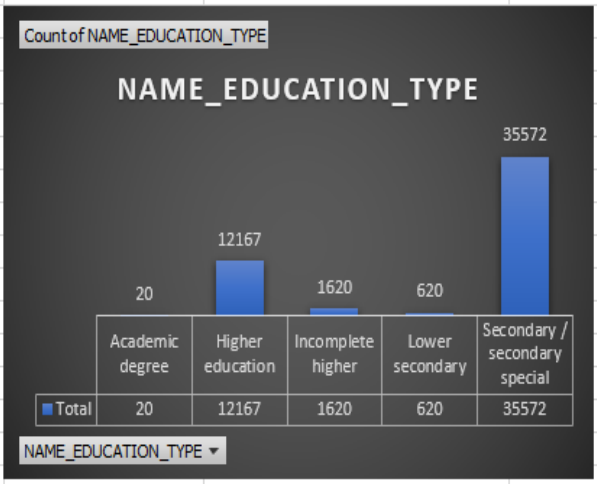
Count of NAME_INCOME_TYPE / Row Labels	Column Labels 0	1	Grand Total	Percentage of defaulters in each category
Businessman	2		2	0%
Commercial associate	10679	864	11543	7%
Maternity leave	1		1	0%
Pensioner	8419	501	8920	6%
State servant	3314	198	3512	6%
Student	5		5	0%
Unemployed	4	2	6	33%
Working	23549	2461	26010	9%
Grand Total	**45973**	**4026**	**49999**	62%

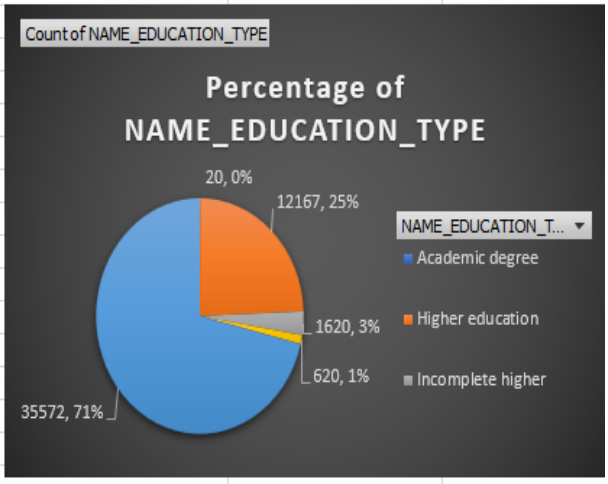


Majority of defaulters of loan are Unemployed followed by working.

Univariate Analysis: NAME_EDUCATION_TYPE

Row Labels	Count of NAME_EDUCATION_TYP	Percentage
Academic degree	20	0%
Higher education	12167	24%
Incomplete higher	1620	3%
Lower secondary	620	1%
Secondary / secondary	35572	71%
Grand Total	**49999**	100%

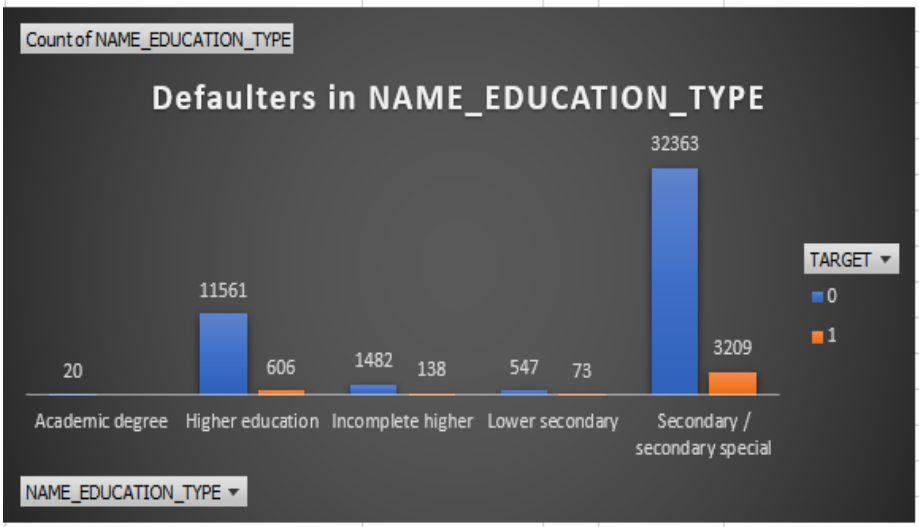




Majority of customers have Secondary/Secondary special education followed by Higher education

Segmented Univariate Analysis: NAME_EDUCATION_TYPE with TARGET

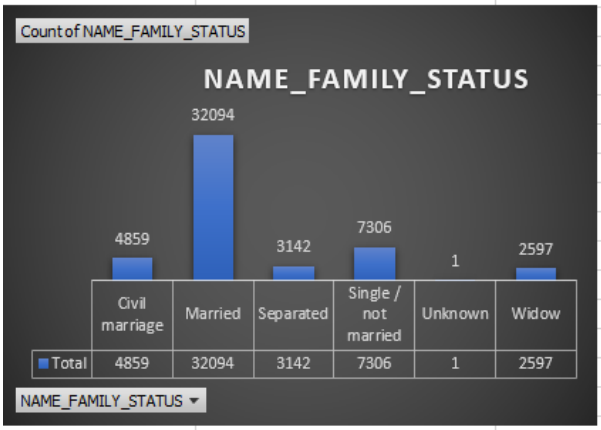
Count of NAME_EDUCATION_TYPE	Column Labels			Percentage of defaulters in
Row Labels	0	1	Grand Total	each category
Academic degree	20		20	0%
Higher education	11561	606	12167	5%
Incomplete higher	1482	138	1620	9%
Lower secondary	547	73	620	12%
Secondary / secondary special	32363	3209	35572	9%
Grand Total	**45973**	**4026**	**49999**	**34%**

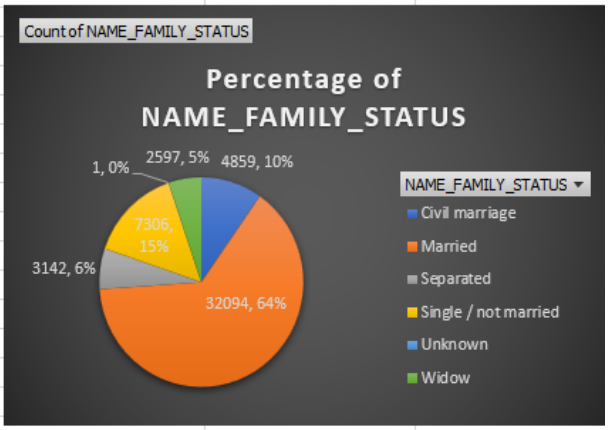


12% defaulters have lower secondary degree followed by Incomplete higher and Secondary special.
The academic degree holders are least likely to default

Univariate Analysis: NAME_FAMILY_STATUS

Row Labels	Count of NAME_FAMILY_STATUS	Percentage
Civil marriage	4859	10%
Married	32094	64%
Separated	3142	6%
Single / not married	7306	15%
Unknown	1	0%
Widow	2597	5%
Grand Total	**49999**	100%

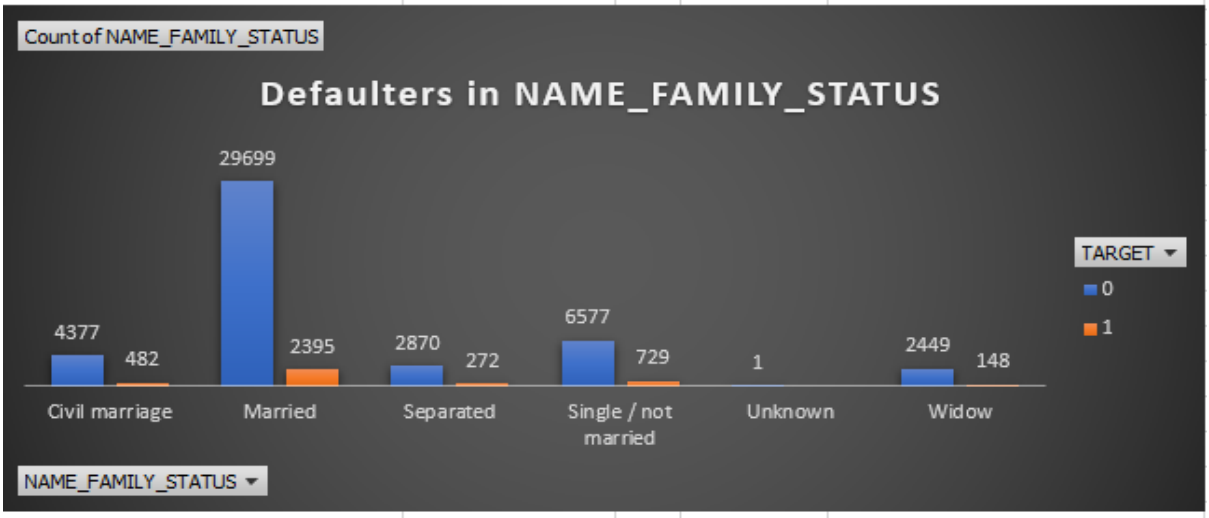




Majority of customers who have taken loans are married followed by single

Segmented Univariate Analysis: NAME_FAMILY_STATUS with TARGET

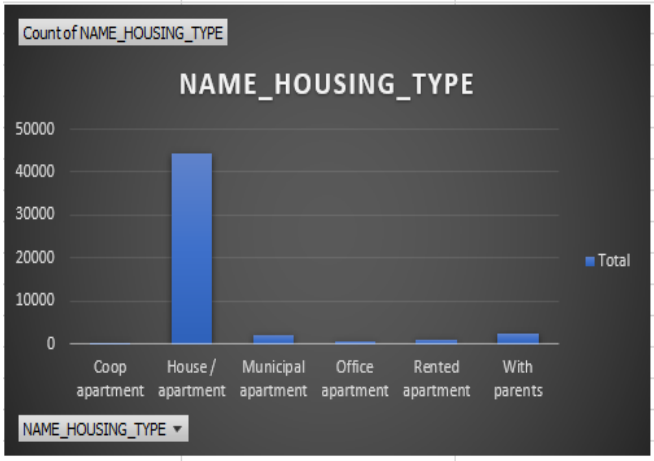
Count of NAME_FAMILY_STATUS / Row Labels	Column Labels 0	1	Grand Total	Percentage of defaulters in each category
Civil marriage	4377	482	4859	10%
Married	29699	2395	32094	7%
Separated	2870	272	3142	9%
Single / not married	6577	729	7306	10%
Unknown	1		1	0%
Widow	2449	148	2597	6%
Grand Total	**45973**	**4026**	**49999**	42%

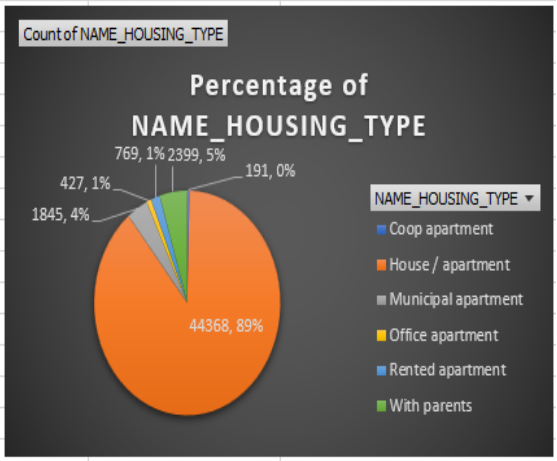


Civil Marriage and Single/not married have highest percentage of defaulting a loan.
Widow have least percent of defaulting

Univariate Analysis: NAME_HOUSING_TYPE

Row Labels	Count of NAME_HOUSING_TYPE	Percentage
Coop apartment	191	0%
House / apartment	44368	89%
Municipal apartment	1845	4%
Office apartment	427	1%
Rented apartment	769	2%
With parents	2399	5%
Grand Total	**49999**	100%

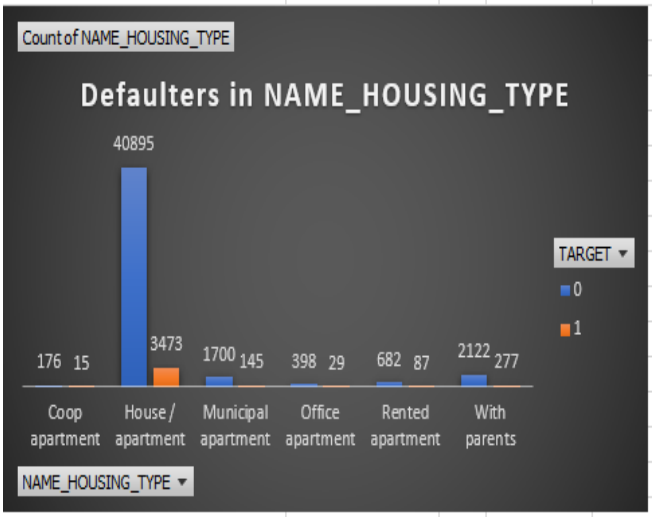




Majority of almost 89% of customers live in House/apartment

Segmented Univariate Analysis: NAME_HOUSING_TYPE with TARGET

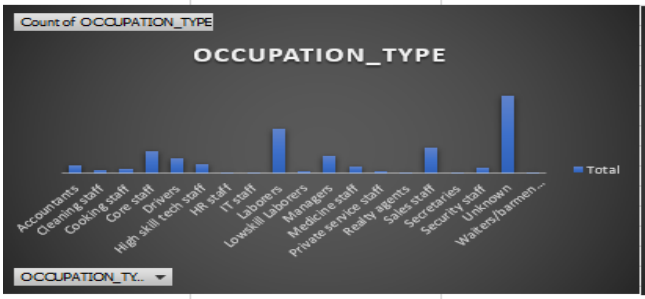
Count of NAME_HOUSING_TYPE / Row Labels	Column Labels 0	1	Grand Total	Percentage of defaulters in each category
Coop apartment	176	15	191	8%
House / apartment	40895	3473	44368	8%
Municipal apartment	1700	145	1845	8%
Office apartment	398	29	427	7%
Rented apartment	682	87	769	11%
With parents	2122	277	2399	12%
Grand Total	**45973**	**4026**	**49999**	53%

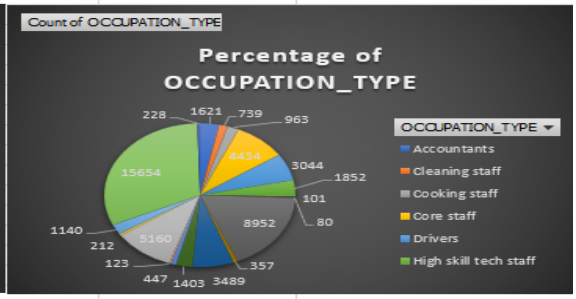


Person living with parents closely followed by those staying in rented apartments have higher probability of defaulting.

Univariate Analysis: OCCUPATION_TYPE

Row Labels	Count of OCCUPATION_TYPE	Percentage
Accountants	1621	3%
Cleaning staff	739	1%
Cooking staff	963	2%
Core staff	4434	9%
Drivers	3044	6%
High skill tech staff	1852	4%
HR staff	101	0%
IT staff	80	0%
Laborers	8952	18%
Lowskill Laborers	357	1%
Managers	3489	7%
Medicine staff	1403	3%
Private service staff	447	1%
Realty agents	123	0%
Sales staff	5160	10%
Secretaries	212	0%
Security staff	1140	2%
Unknown	15654	31%
Waiters/barmen staff	228	0%
Grand Total	**49999**	**100%**

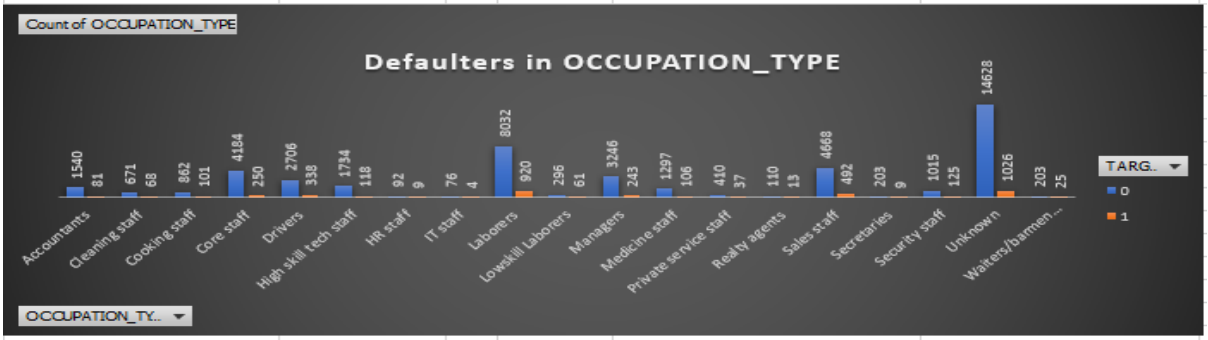




Most of the loans are taken by people whose occupation is unknown and followed by Laborers

Segmented Univariate Analysis: OCCUPATION_TYPE with TARGET

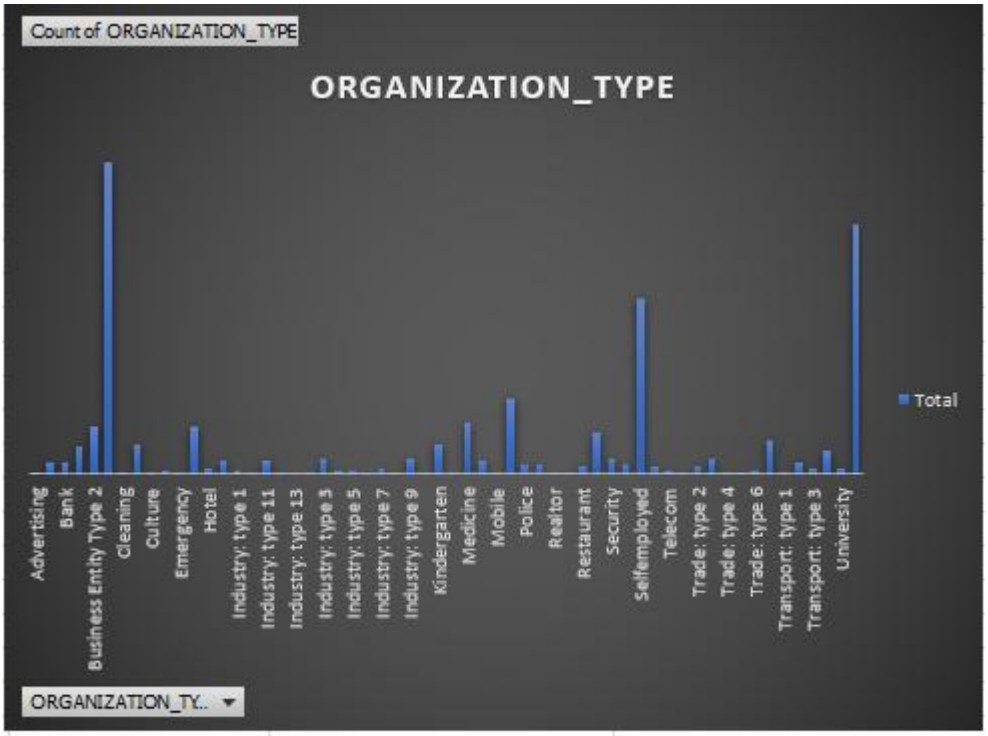
Count of OCCUPATION_TYPE / Row Labels	Column Labels 0	1	Grand Total	Percentage of defaulters in each category
Accountants	1540	81	1621	5%
Cleaning staff	671	68	739	9%
Cooking staff	862	101	963	10%
Core staff	4184	250	4434	6%
Drivers	2706	338	3044	11%
High skill tech staff	1734	118	1852	6%
HR staff	92	9	101	9%
IT staff	76	4	80	5%
Laborers	8032	920	8952	10%
Lowskill Laborers	296	61	357	17%
Managers	3246	243	3489	7%
Medicine staff	1297	106	1403	8%
Private service staff	410	37	447	8%
Realty agents	110	13	123	11%
Sales staff	4668	492	5160	10%
Secretaries	203	9	212	4%
Security staff	1015	125	1140	11%
Unknown	14628	1026	15654	7%
Waiters/barmen staff	203	25	228	11%
Grand Total	**45973**	**4026**	**49999**	**165%**



The category with highest percentage of defaulters is Lowskilled Laborers followed by Drivers, Realty agents and Security staff.
The category with least percentage of defaulters is Secretaries followed by Accountants

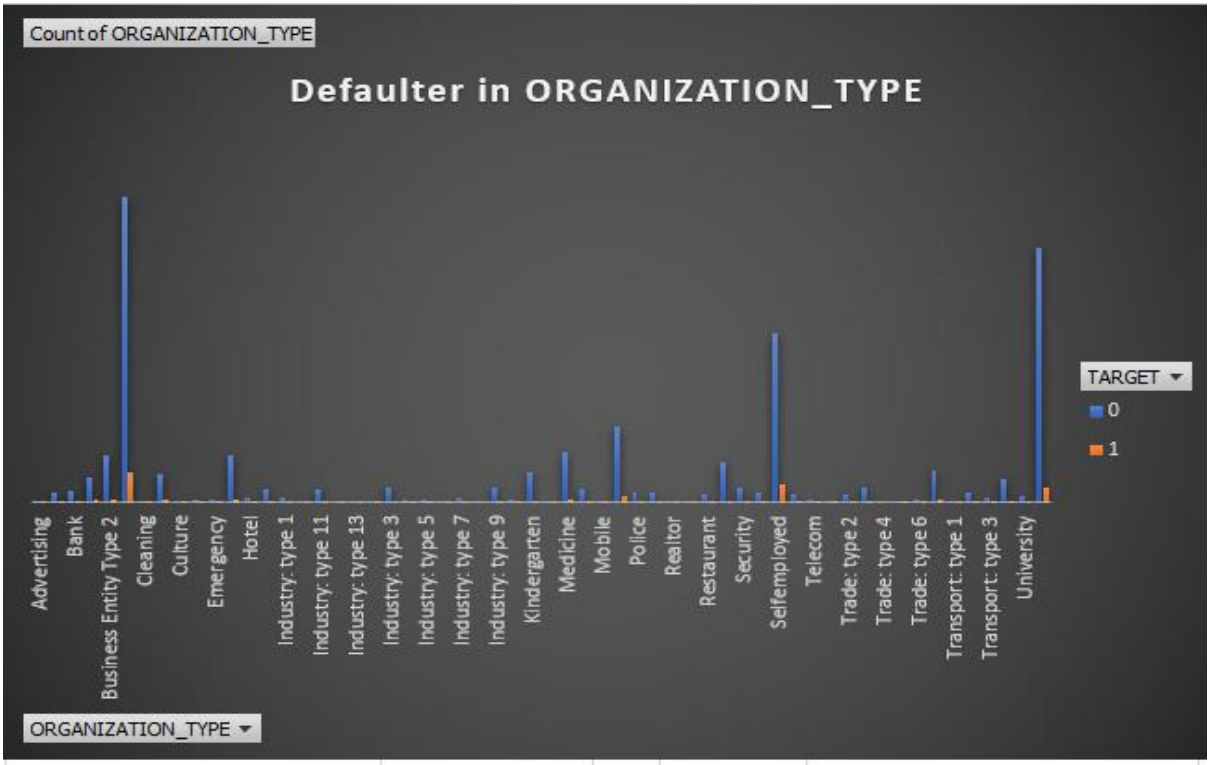
Univariate Analysis: ORGANIZATION_TYPE

Row Labels	Count of ORGANIZATION_TYPE	Percentage
Advertising	68	0%
Agriculture	392	1%
Bank	435	1%
Business Entity Type 1	953	2%
Business Entity Type 2	1704	3%
Business Entity Type 3	11101	22%
Cleaning	40	0%
Construction	1066	2%
Culture	64	0%
Electricity	147	0%
Emergency	93	0%
Government	1716	3%
Hotel	182	0%
Housing	489	1%
Industry: type 1	159	0%
Industry: type 10	21	0%
Industry: type 11	489	1%
Industry: type 12	53	0%
Industry: type 13	15	0%
Industry: type 2	78	0%
Industry: type 3	542	1%
Industry: type 4	140	0%
Industry: type 5	103	0%
Industry: type 6	12	0%
Industry: type 7	209	0%
Industry: type 8	8	0%
Industry: type 9	537	1%
Insurance	89	0%
Kindergarten	1090	2%
Legal Services	44	0%
Medicine	1817	4%
Military	458	1%
Mobile	56	0%
Other	2717	5%
Police	366	1%
Postal	370	1%
Realtor	61	0%
Religion	14	0%
Restaurant	289	1%
School	1450	3%
Security	550	1%
Security Ministries	331	1%
Selfemployed	6240	12%
Services	284	1%
Telecom	106	0%
Trade: type 1	66	0%
Trade: type 2	307	1%
Trade: type 3	550	1%
Trade: type 4	8	0%
Trade: type 5	8	0%
Trade: type 6	108	0%
Trade: type 7	1210	2%
Transport: type 1	28	0%
Transport: type 2	392	1%
Transport: type 3	191	0%
Transport: type 4	837	2%
University	222	0%
XNA	8924	18%
Grand Total	**49999**	100%



Most of the loan application is from Business Entity Type 3
Almost 18% of loan applicants organization type is unknown.

Segmented Univariate Analysis: ORGANIZATION_TYPE with TARGET

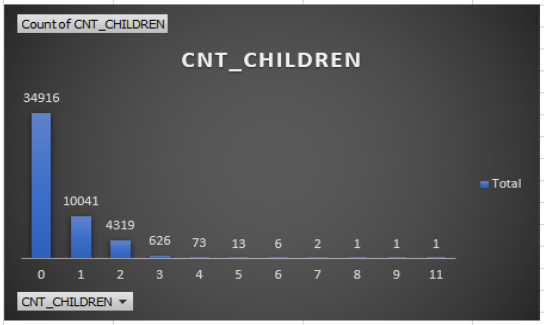


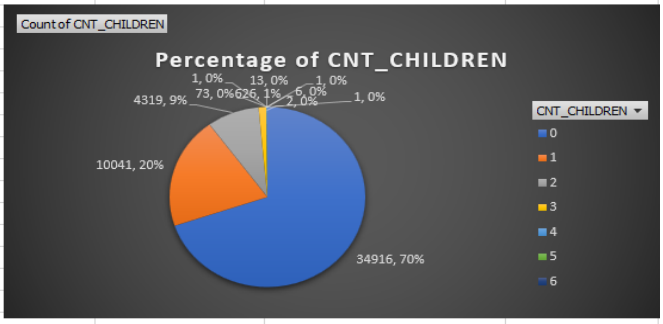
Organization with highest percent of defaulter is Industry: type 13 with 27%
Organization withleast percent of defaulter is Industry: type 8 and Trade: type 4

Count of ORGANIZATION_TYPE	Column Labels			Percentage of defaulters in
Row Labels	0	1	Grand Total	each category
Advertising	61	7	68	10%
Agriculture	341	51	392	13%
Bank	408	27	435	6%
Business Entity Type 1	865	88	953	9%
Business Entity Type 2	1571	133	1704	8%
Business Entity Type 3	10087	1014	11101	9%
Cleaning	37	3	40	8%
Construction	958	108	1066	10%
Culture	62	2	64	3%
Electricity	134	13	147	9%
Emergency	86	7	93	8%
Government	1592	124	1716	7%
Hotel	169	13	182	7%
Housing	447	42	489	9%
Industry: type 1	140	19	159	12%
Industry: type 10	20	1	21	5%
Industry: type 11	461	28	489	6%
Industry: type 12	50	3	53	6%
Industry: type 13	11	4	15	27%
Industry: type 2	68	10	78	13%
Industry: type 3	491	51	542	9%
Industry: type 4	125	15	140	11%
Industry: type 5	96	7	103	7%
Industry: type 6	11	1	12	8%
Industry: type 7	190	19	209	9%
Industry: type 8	8		8	0%
Industry: type 9	496	41	537	8%
Insurance	82	7	89	8%
Kindergarten	1024	66	1090	6%
Legal Services	40	4	44	9%
Medicine	1687	130	1817	7%
Military	432	26	458	6%
Mobile	52	4	56	7%
Other	2509	208	2717	8%
Police	348	18	366	5%
Postal	343	27	370	7%
Realtor	54	7	61	11%
Religion	13	1	14	7%
Restaurant	257	32	289	11%
School	1372	78	1450	5%
Security	488	62	550	11%
Security Ministries	315	16	331	5%
Selfemployed	5612	628	6240	10%
Services	260	24	284	8%
Telecom	98	8	106	8%
Trade: type 1	61	5	66	8%
Trade: type 2	286	21	307	7%
Trade: type 3	490	60	550	11%
Trade: type 4	8		8	0%
Trade: type 5	7	1	8	13%
Trade: type 6	105	3	108	3%
Trade: type 7	1090	120	1210	10%
Transport: type 1	26	2	28	7%
Transport: type 2	359	33	392	8%
Transport: type 3	166	25	191	13%
Transport: type 4	770	67	837	8%
University	213	9	222	4%
XNA	8421	503	8924	6%
Grand Total	**45973**	**4026**	**49999**	

Univariate Analysis: CNT_CHILDREN

	NUMERICAL VARIABLE ANALYSIS					
Row Labels	Count of CNT_CHILDREN	Percentage	Mean	Median	Variance	Standard Deviation
0	34916	70%	0.419848397	0	0.524231818	0.724038548
1	10041	20%				
2	4319	9%				
3	626	1%				
4	73	0%				
5	13	0%				
6	6	0%				
7	2	0%				
8	1	0%				
9	1	0%				
11	1	0%				
Grand Total	49999	100%				

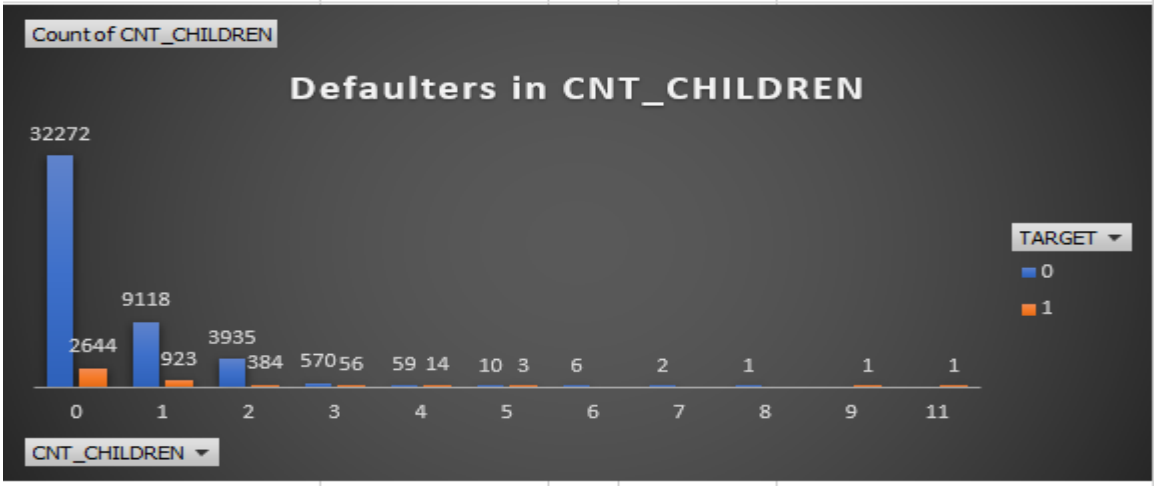




Most of the loan applicants do not have any child.

Segmented Univariate Analysis: CNT_CHILDREN with TARGET

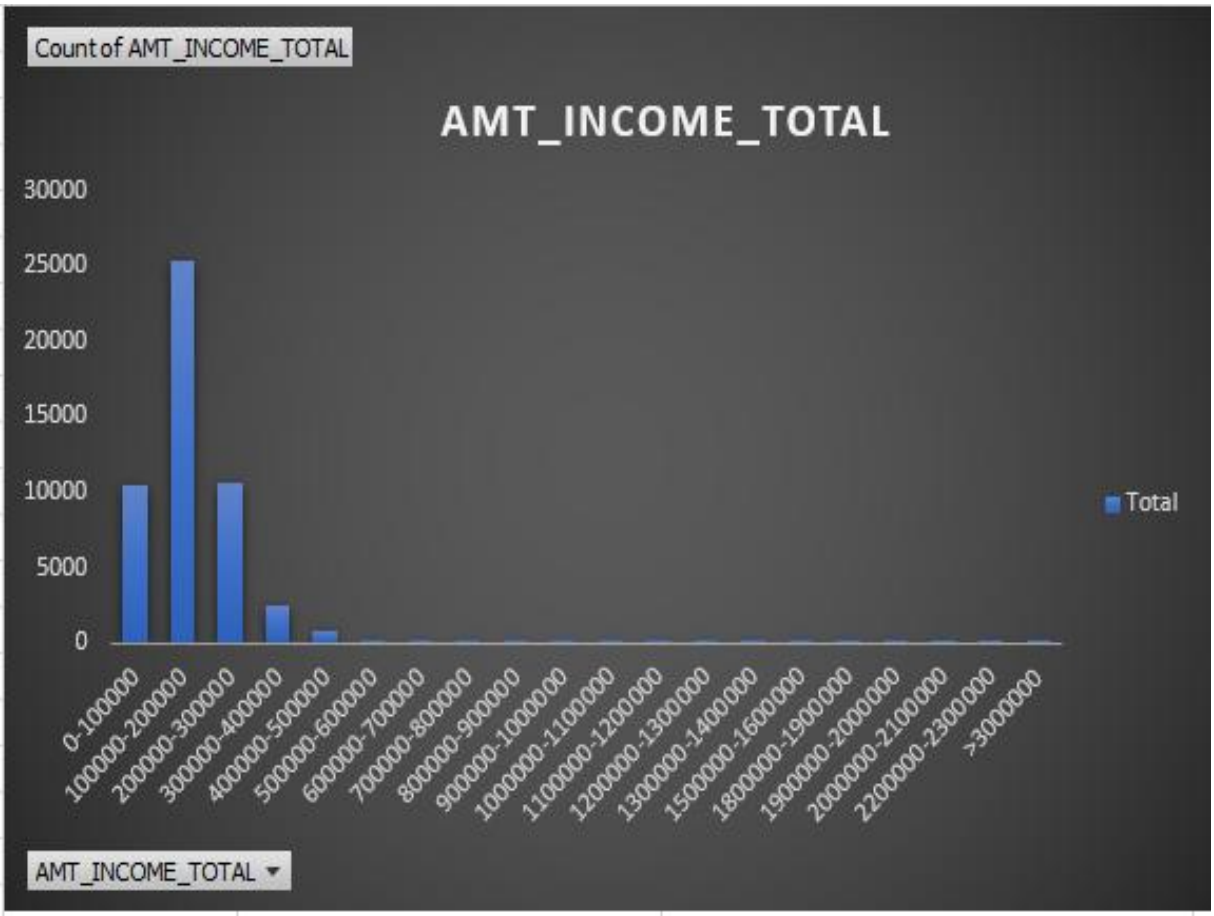
Count of CNT_CHILDREN	Column Labels			Percentage of defaulters in
Row Labels	0	1	Grand Total	each category
0	32272	2644	34916	8%
1	9118	923	10041	9%
2	3935	384	4319	9%
3	570	56	626	9%
4	59	14	73	19%
5	10	3	13	23%
6	6		6	0%
7	2		2	0%
8	1		1	0%
9		1	1	100%
11		1	1	100%
Grand Total	45973	4026	49999	



Customers with 9 or more than 9 children have higher(100%) rate of defaulting

Univariate Analysis: AMT_INCOME_TOTAL

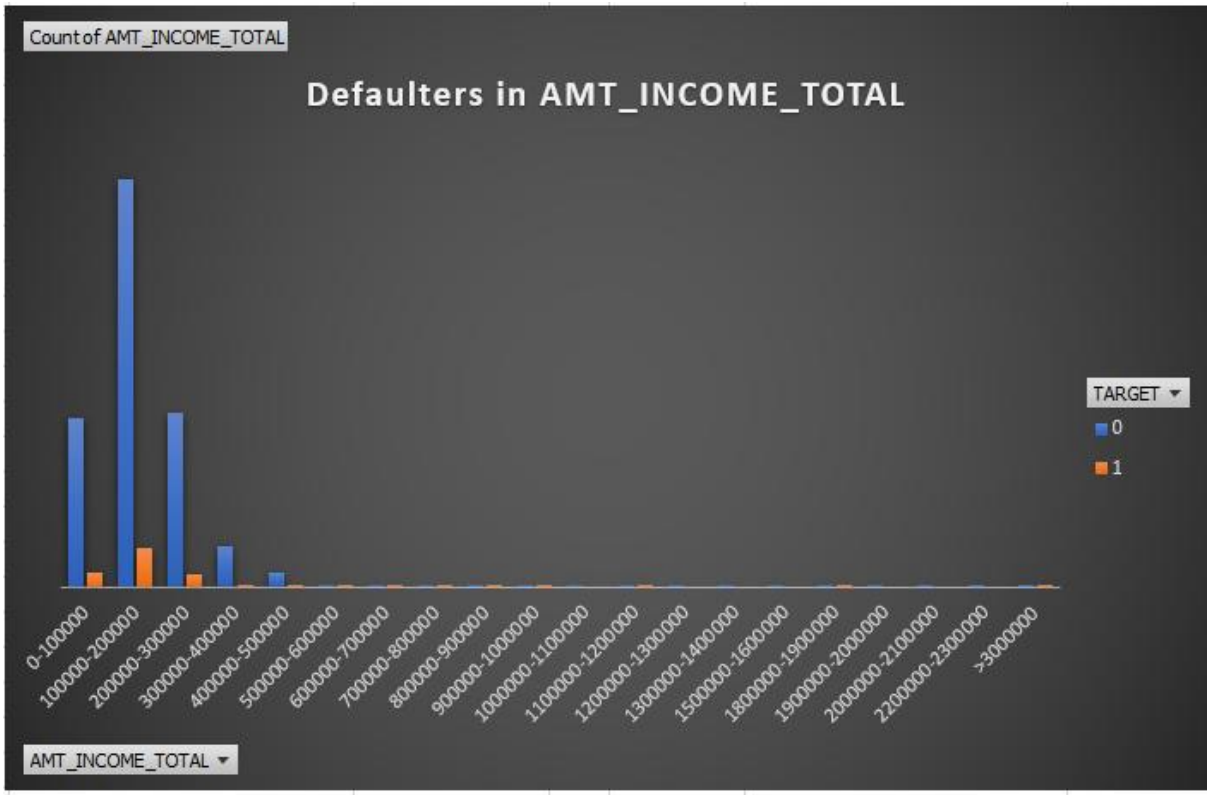
Row Labels	Count of AMT_INCOME_TOTAL	Percentage	Mean	Median	Variance	Standard Deviatio	min	max
0-100000	10392	21%	170767.5905	145800	2.82832E+11	531819.0951	25650	117000000
100000-200000	25260	51%						
200000-300000	10606	21%						
300000-400000	2438	5%						
400000-500000	849	2%						
500000-600000	167	0%						
600000-700000	157	0%						
700000-800000	33	0%						
800000-900000	26	0%						
900000-1000000	31	0%						
1000000-1100000	4	0%						
1100000-1200000	13	0%						
1200000-1300000	1	0%						
1300000-1400000	10	0%						
1500000-1600000	1	0%						
1800000-1900000	3	0%						
1900000-2000000	1	0%						
2000000-2100000	2	0%						
2200000-2300000	2	0%						
>3000000	3	0%						
Grand Total	49999	100%						



Majority of the applicants have total income of less than 3 lakhs
Maximum applicants fall in a range of 1 lakh to 2 lakh income class.

Segmented Univariate Analysis: AMT_INCOME_TOTAL with TARGET

Count of AMT_INCOME_TOTAL Row Labels	Column Labels 0	1	Grand Total	Percentage of defaulters in each category
0-100000	9547	845	10392	8%
100000-200000	23072	2188	25260	9%
200000-300000	9842	764	10606	7%
300000-400000	2307	131	2438	5%
400000-500000	782	67	849	8%
500000-600000	153	14	167	8%
600000-700000	148	9	157	6%
700000-800000	32	1	33	3%
800000-900000	24	2	26	8%
900000-1000000	29	2	31	6%
1000000-1100000	4		4	0%
1100000-1200000	12	1	13	8%
1200000-1300000	1		1	0%
1300000-1400000	10		10	0%
1500000-1600000	1		1	0%
1800000-1900000	2	1	3	33%
1900000-2000000	1		1	0%
2000000-2100000	2		2	0%
2200000-2300000	2		2	0%
>3000000	2	1	3	33%
Grand Total	**45973**	**4026**	**49999**	

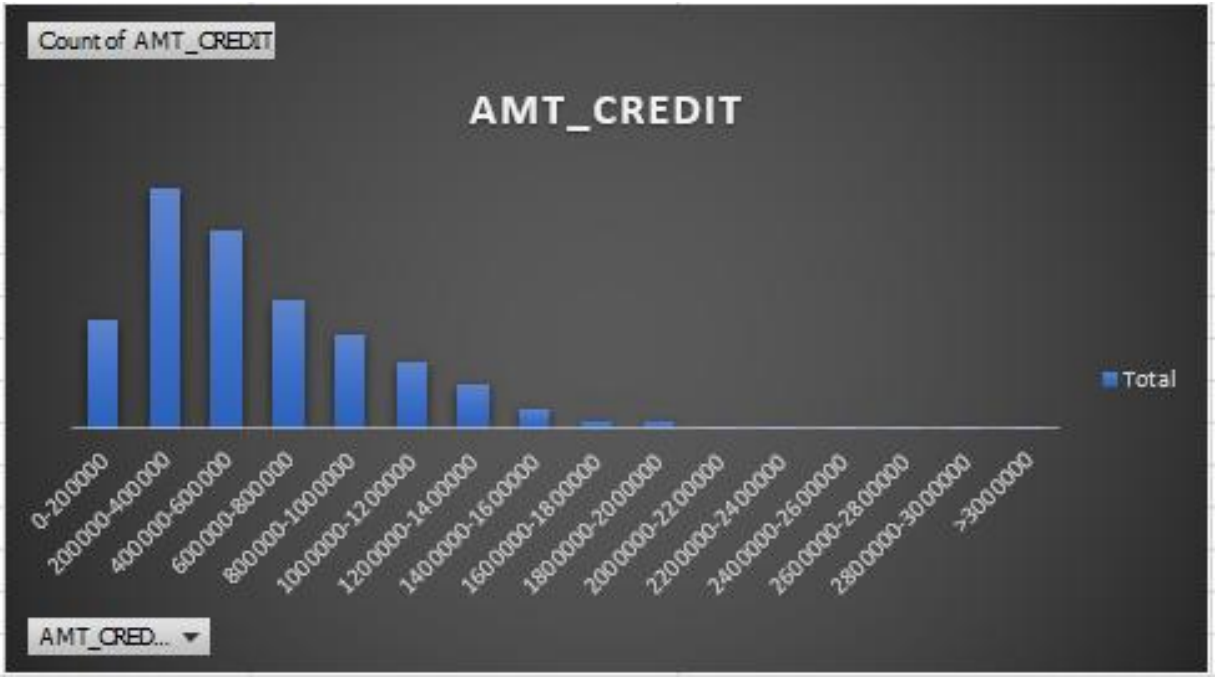


Applicants with less than 2 lakh has high probability of defaulting CONSIDERING the defaulting count

Applicants with greater than 12 lakh income are less likely to default.

Univariate Analysis: AMT_CREDIT

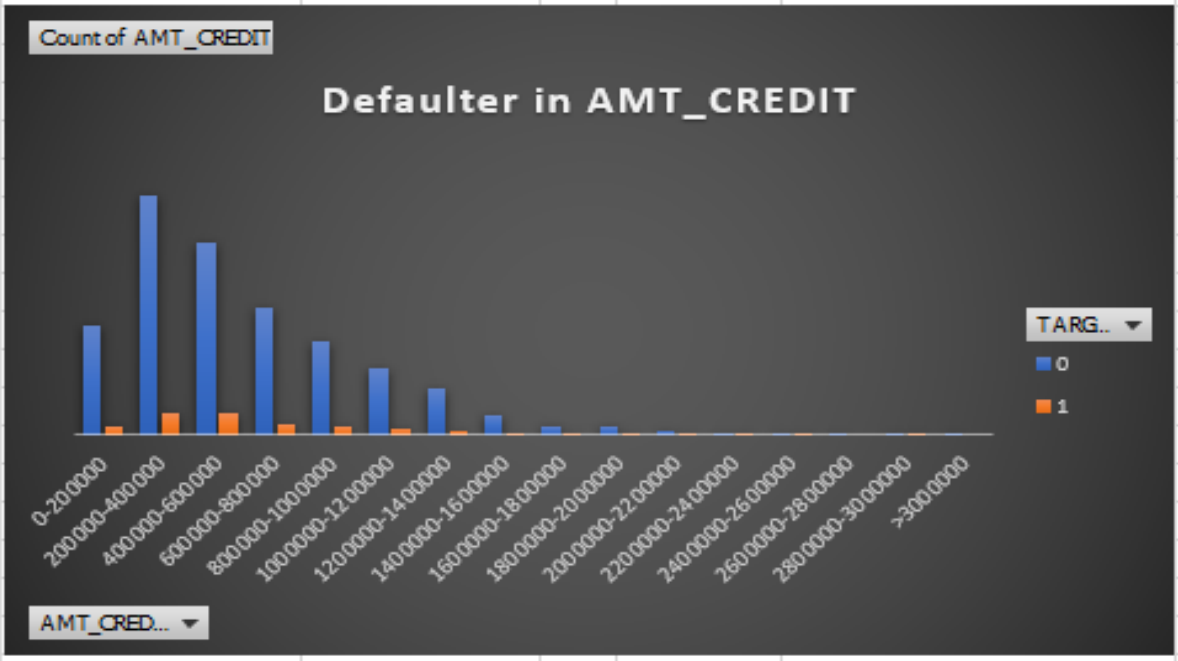
Row Labels	Count of AMT_CREDIT	Percentage	Mean	Median	Variance	Standard Deviati	min	max
0-200000	5900	12%	599700.5815	514777.5	1.61938E+11	402415.4339	45000	4050000
200000-400000	13105	26%						
400000-600000	10782	22%						
600000-800000	6971	14%						
800000-1000000	5095	10%						
1000000-1200000	3615	7%						
1200000-1400000	2408	5%						
1400000-1600000	1038	2%						
1600000-1800000	399	1%						
1800000-2000000	363	1%						
2000000-2200000	151	0%						
2200000-2400000	103	0%						
2400000-2600000	49	0%						
2600000-2800000	13	0%						
2800000-3000000	3	0%						
>3000000	4	0%						
Grand Total	49999	100%						



Maximum loan amount taken is between 2 to 4 lakhs
Minimum loan amount taken is above 20 lakhs

Segmented Univariate Analysis: AMT_CREDIT with TARGET

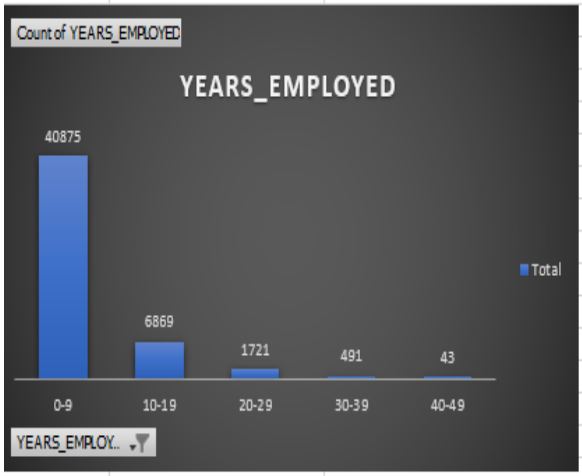
Count of AMT_CREDIT Row Labels	Column Labels 0	1	Grand Total	Percentage of defaulters in each category
0-200000	5510	390	5900	7%
200000-400000	11979	1126	13105	9%
400000-600000	9653	1129	10782	10%
600000-800000	6414	557	6971	8%
800000-1000000	4730	365	5095	7%
1000000-1200000	3369	246	3615	7%
1200000-1400000	2284	124	2408	5%
1400000-1600000	993	45	1038	4%
1600000-1800000	379	20	399	5%
1800000-2000000	350	13	363	4%
2000000-2200000	143	8	151	5%
2200000-2400000	102	1	103	1%
2400000-2600000	48	1	49	2%
2600000-2800000	13		13	0%
2800000-3000000	2	1	3	33%
>3000000	4		4	0%
Grand Total	45973	4026	49999	



Highest defaulters are those who have taken loans between the range of 4 to 6 lakhs

Univariate Analysis: YEARS_EMPLOYED

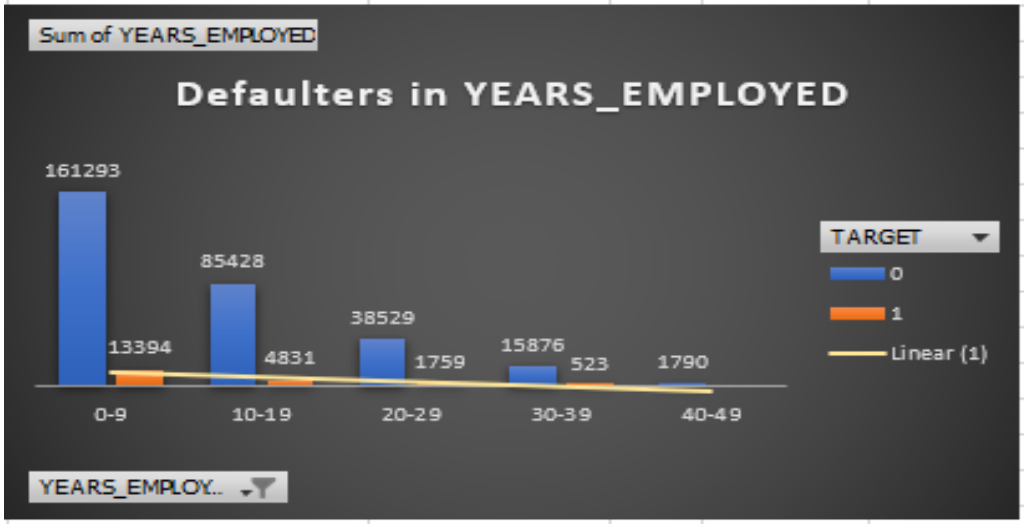
Row Labels	Count of YEARS_EMPLOYED	Percentage	Mean	Median	Variance	Standard Deviati	min	max
0-9	40875	82%	6.468589372	6	34.2350978	5.851076636	0	48
10-19	6869	14%						
20-29	1721	3%						
30-39	491	1%						
40-49	43	0%						
Grand Total	49999	100%						



Majority of loan applicants have 0 to 9 years of experience

Segmented Univariate Analysis: YEARS_EMPLOYED

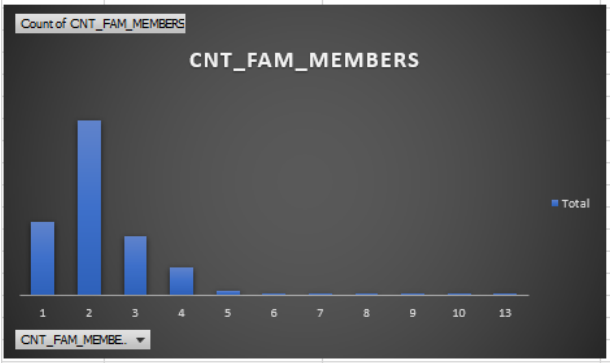
Sum of YEARS_EMPLOYED Row Labels	Column Labels 0	1	Grand Total	Percentage of defaulters in each category
0-9	161293	13394	174687	8%
10-19	85428	4831	90259	5%
20-29	38529	1759	40288	4%
30-39	15876	523	16399	3%
40-49	1790		1790	0%
Grand Total	302916	20507	323423	



The applicants with 0 to 9 years of experience have high rate of defaulting a loan
The applicants with 40 to 49 years of experience have low rate of defaulting a loan
With increase in experience, defaulting rate is gradually decreasing.

Univariate Analysis: CNT_FAM_MEMBERS

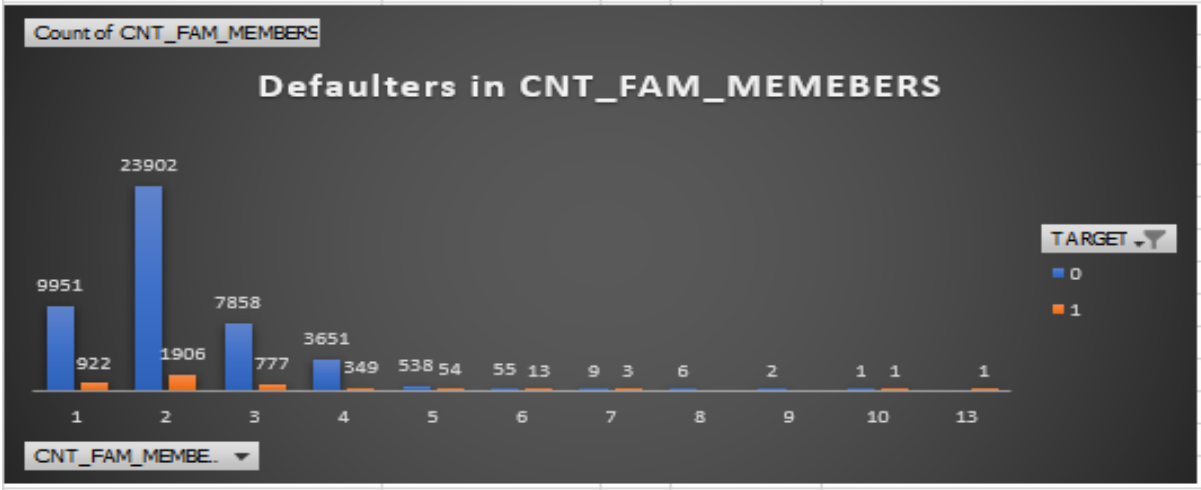
Row Labels	Count of CNT_FAM_MEMBERS	Percentage	Mean	Median	Variance	Standard Deviation	min	max
1	10873	22%	2.158943179	2	0.830510792	0.911323648	1	13
2	25808	52%						
3	8635	17%						
4	4000	8%						
5	592	1%						
6	68	0%						
7	12	0%						
8	6	0%						
9	2	0%						
10	2	0%						
13	1	0%						
Grand Total	49999	100%						



Most of the loan applicants family is of 2

Segmented Univariate Analysis: CNT_FAM_MEMBERS

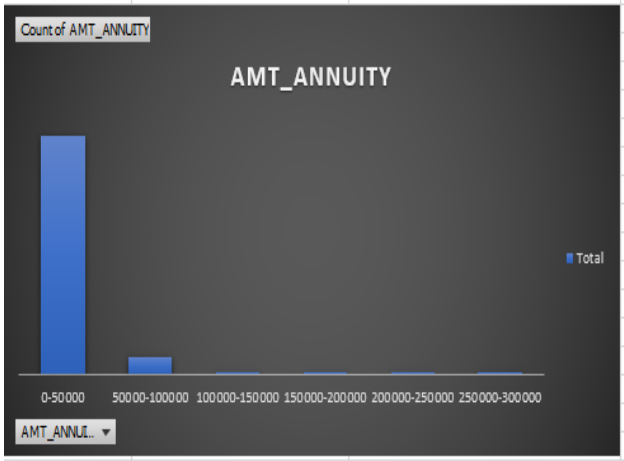
Count of CNT_FAM_MEMBERS / Row Labels	Column Labels 0	1	Grand Total	Percentage of defaulters in each category
1	9951	922	10873	8%
2	23902	1906	25808	7%
3	7858	777	8635	9%
4	3651	349	4000	9%
5	538	54	592	9%
6	55	13	68	19%
7	9	3	12	25%
8	6		6	0%
9	2		2	0%
10	1	1	2	50%
13		1	1	100%
Grand Total	45973	4026	49999	



Family of more number of people tend to default more, maybe due to more constraints and need

Univariate Analysis: AMT_ANNUITY

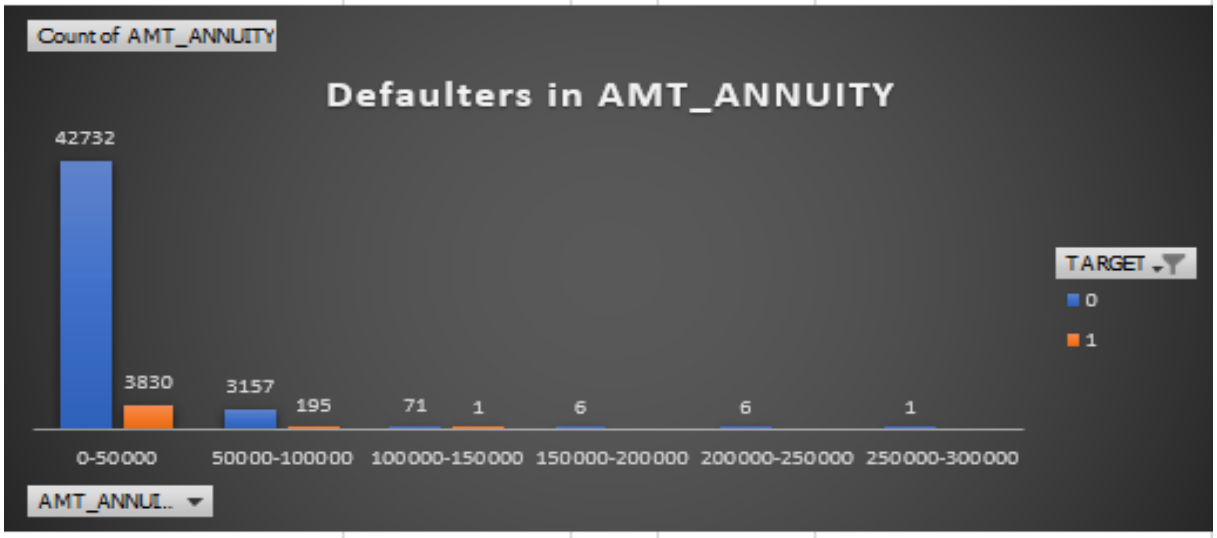
Row Labels	Count of AMT_ANNUITY	Percentage	Mean	Median	Variance	Standard Deviation	min	max
0-50000	46562	93%	27107.33399	24939	212075202.9	14562.80203	2052	258025.5
50000-100000	3352	7%						
100000-150000	72	0%						
150000-200000	6	0%						
200000-250000	6	0%						
250000-300000	1	0%						
Grand Total	49999	100%						



Most people pay an annuity of below 50,000 for the credit loan

Segmented Univariate Analysis: AMT_ANNUITY

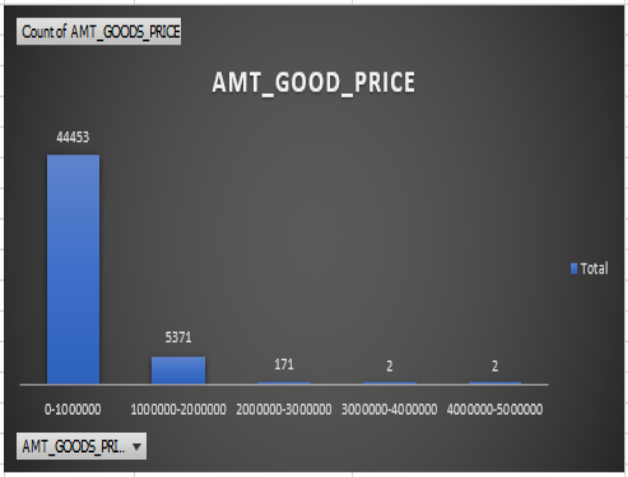
Count of AMT_ANNUITY	Column Labels			Percentage of defaulters in
Row Labels	0	1	Grand Total	each category
0-50000	42732	3830	46562	8%
50000-100000	3157	195	3352	6%
100000-150000	71	1	72	1%
150000-200000	6		6	0%
200000-250000	6		6	0%
250000-300000	1		1	0%
Grand Total	45973	4026	49999	



Most of the default is done by people who pay an anuity of below 50,000

Univariate Analysis: AMT_GOOD_PRICE

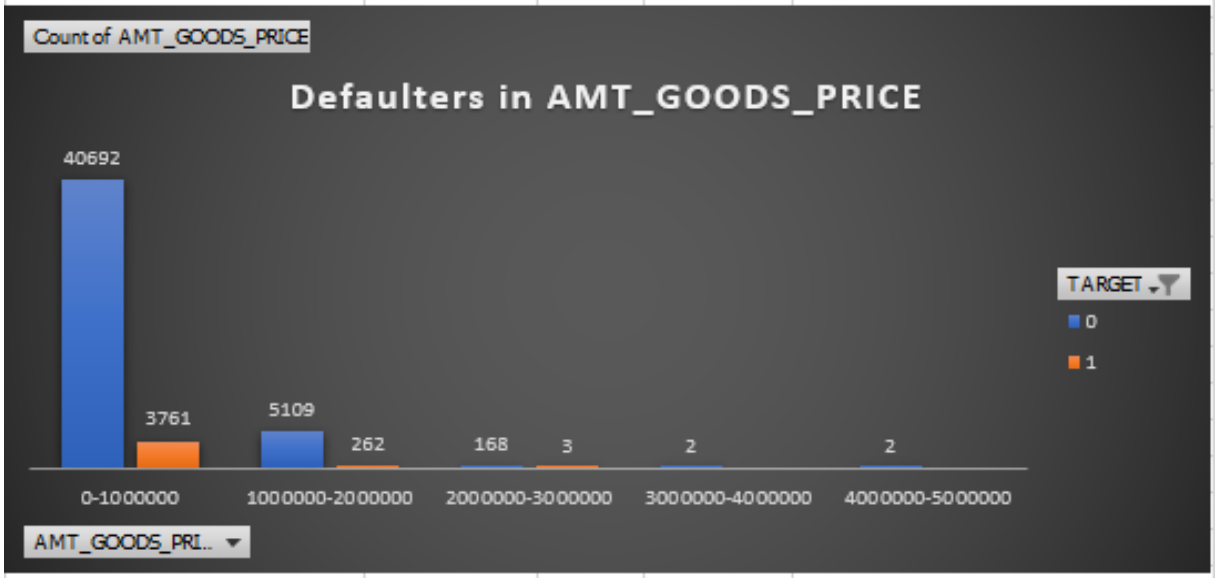
Row Labels	Count of AMT_GOODS_PRICE	Percentage	Mean	Median	Variance	Standard Deviati	min	max
0-1000000	44453	89%	538992.3491	450000	1.36693E+11	369720.8225	45000	4050000
1000000-2000000	5371	11%						
2000000-3000000	171	0%						
3000000-4000000	2	0%						
4000000-5000000	2	0%						
Grand Total	49999	100%						



Most loans are granted for goods that cost below 10 lakhs

Segmented Univariate Analysis: AMT_GOOD_PRICE

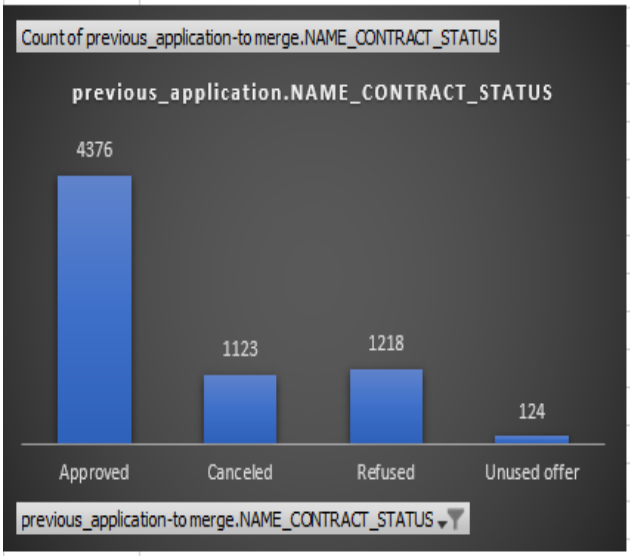
Count of AMT_GOODS_PRICE	Column Lab			Percentage of defaulters in
Row Labels	0	1	Grand Total	each category
0-1000000	40692	3761	44453	8%
1000000-2000000	5109	262	5371	5%
2000000-3000000	168	3	171	2%
3000000-4000000	2		2	0%
4000000-5000000	2		2	0%
Grand Total	45973	4026	49999	

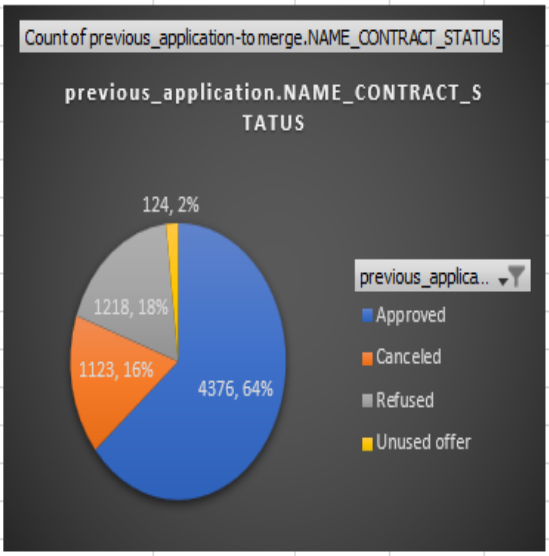


The highest default is done by a person who has taken a loan for goods that cost below 10 lakh

Univariate Analysis: previous_application.NAME_CONTRACT_TYPE

Univariate Analysis		
Row Labels	Count of previous_application-to merge.NAME_CONTRACT_STATUS	Percentage
Approved	4376	64%
Canceled	1123	16%
Refused	1218	18%
Unused offer	124	2%
Grand Total	**6841**	100%

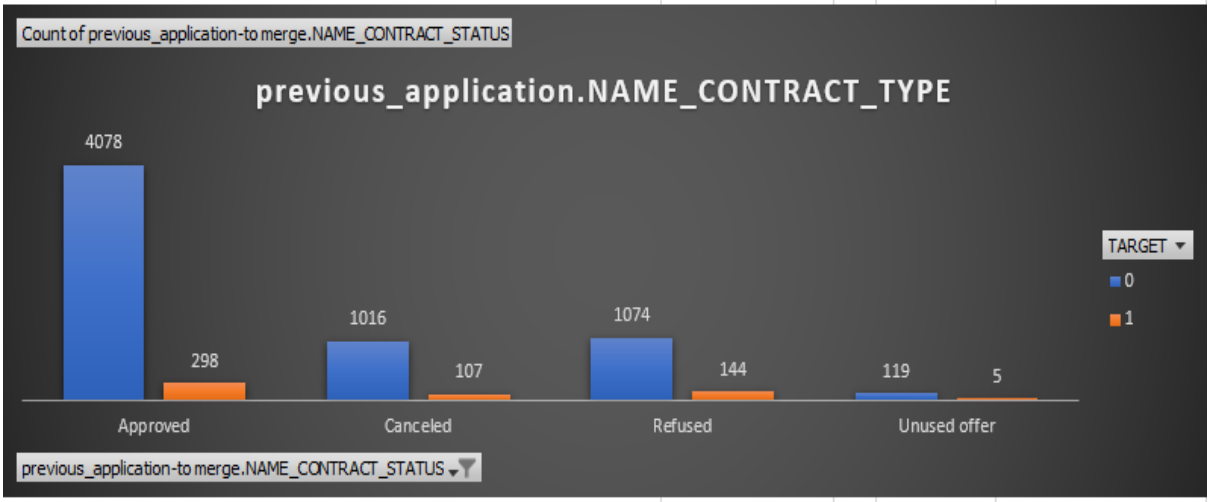




64% of the previous loans were approved

Segmented Univariate Analysis: previous_application.NAME_CONTRACT_TYPE

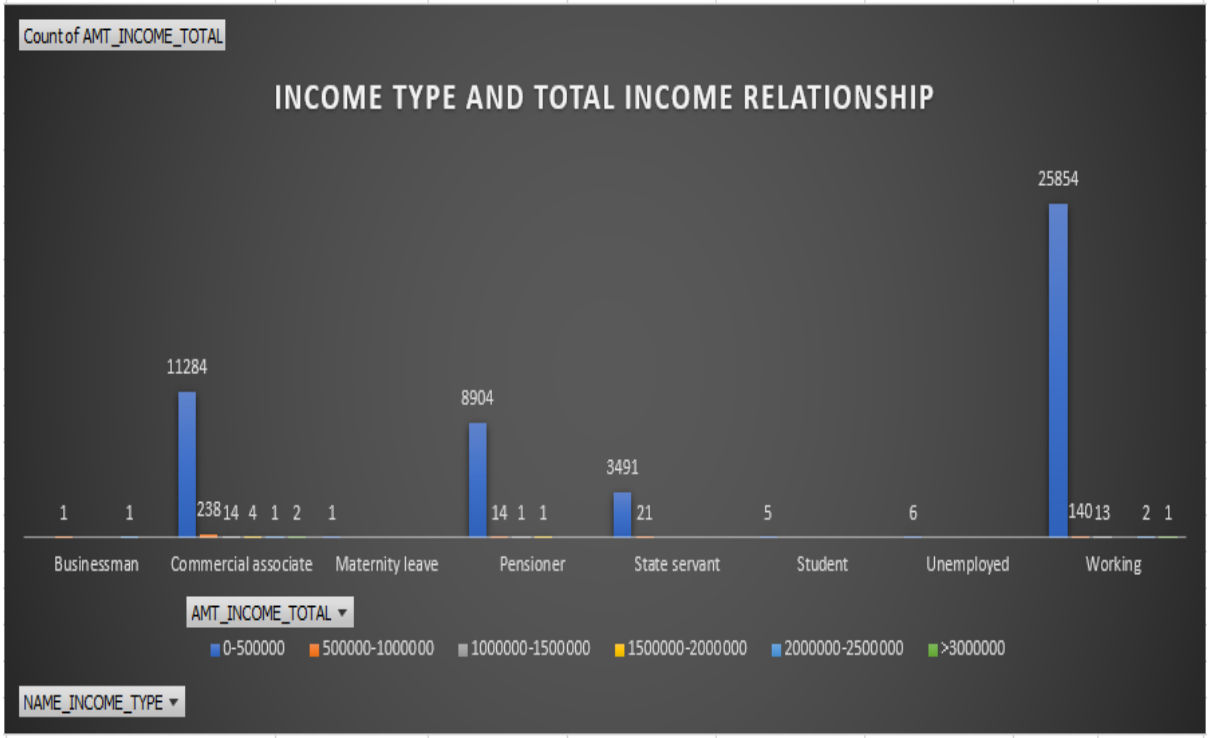
Count of previous_application-to merge.NAME_CONTRACT_STATUS	Column Labels			Percentage of
Row Labels	0	1	Grand Total	each category
Approved	4078	298	4376	7%
Canceled	1016	107	1123	10%
Refused	1074	144	1218	12%
Unused offer	119	5	124	4%
Grand Total	**6287**	**554**	**6841**	



Most of the defaulters previous loan applications were refused

Relationship between NAME_INCOME_TYPE and AMT_INCOME_TYPE

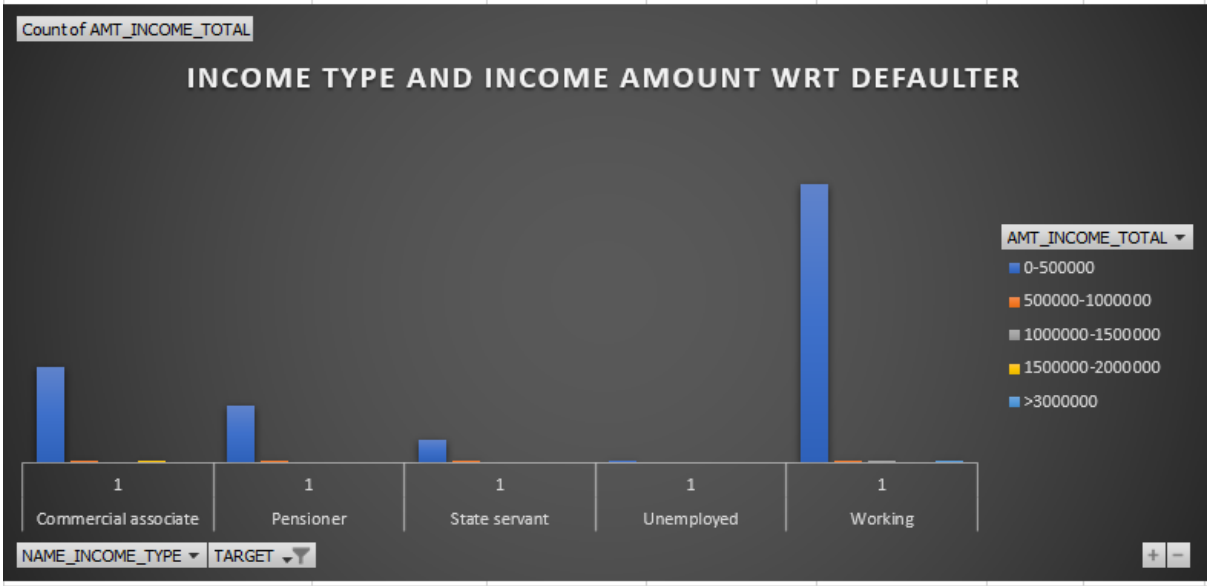
Bivariate Analysis							
Count of AMT_INCOME_TOTAL	Column Labels						
Row Labels	0-500000	500000-1000000	1000000-1500000	1500000-2000000	2000000-2500000	>3000000	Grand Total
Businessman		1			1		2
Commercial associate	11284	238	14	4	1	2	11543
Maternity leave	1						1
Pensioner	8904	14	1	1			8920
State servant	3491	21					3512
Student	5						5
Unemployed	6						6
Working	25854	140	13		2	1	26010
Grand Total	**49545**	**414**	**28**	**5**	**4**	**3**	**49999**



Most of the commercial associates and working class earns between 5 lakhs to 30 lakhs
Most of the working class and businessman earn between 5 lakh to 25 lakhs
Those on maternity leave, pensions and state servants have an income of within 5 lakhs

NAME_INCOME_TYPE and AMT_INCOME_TYPE wrt Defaulter

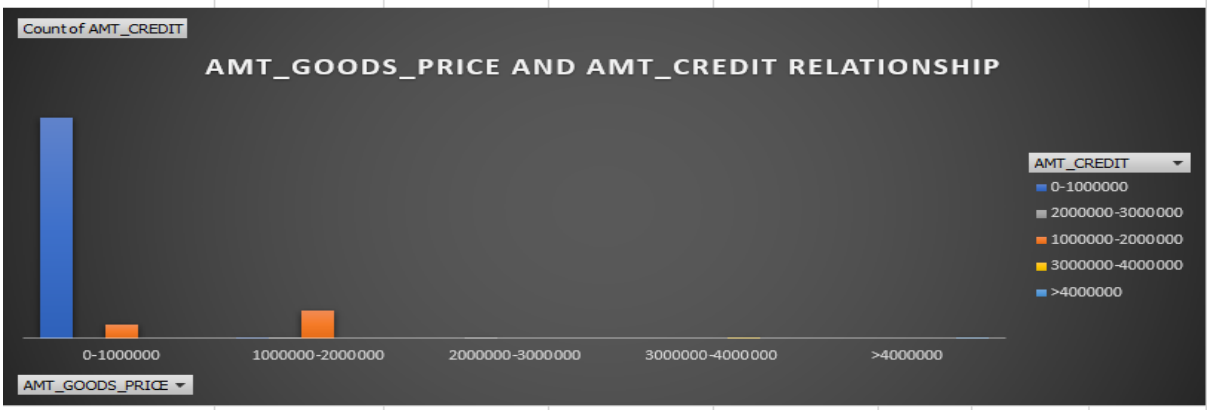
Count of AMT_INCOME_TOTAL	Column Labels					
Row Labels	0-500000	500000-1000000	1000000-1500000	1500000-2000000	>3000000	Grand Total
Commercial associate	**848**	**15**		**1**		**864**
1	848	15		1		864
Pensioner	**500**	**1**				**501**
1	500	1				501
State servant	**196**	**2**				**198**
1	196	2				198
Unemployed	**2**					**2**
1	2					2
Working	**2449**	**10**	**1**		**1**	**2461**
1	2449	10	1		1	2461
Grand Total	**3995**	**28**	**1**	**1**	**1**	**4026**



According to the above visual, Working class with a income of 0-5 lakhs are the highest defaulters.

AMT_GOODS_PRICE AND AMT_CREDIT RELATIONSHIP

Count of AMT_CREDIT	Column Labels					
Row Labels	0-1000000	1000000-2000000	2000000-3000000	3000000-4000000	>4000000	Grand Total
0-1000000	41852	2601				44453
1000000-2000000	1	5222	148			5371
2000000-3000000			171			171
3000000-4000000				2		2
>4000000					2	2
Grand Total	**41853**	**7823**	**319**	**2**	**2**	**49999**

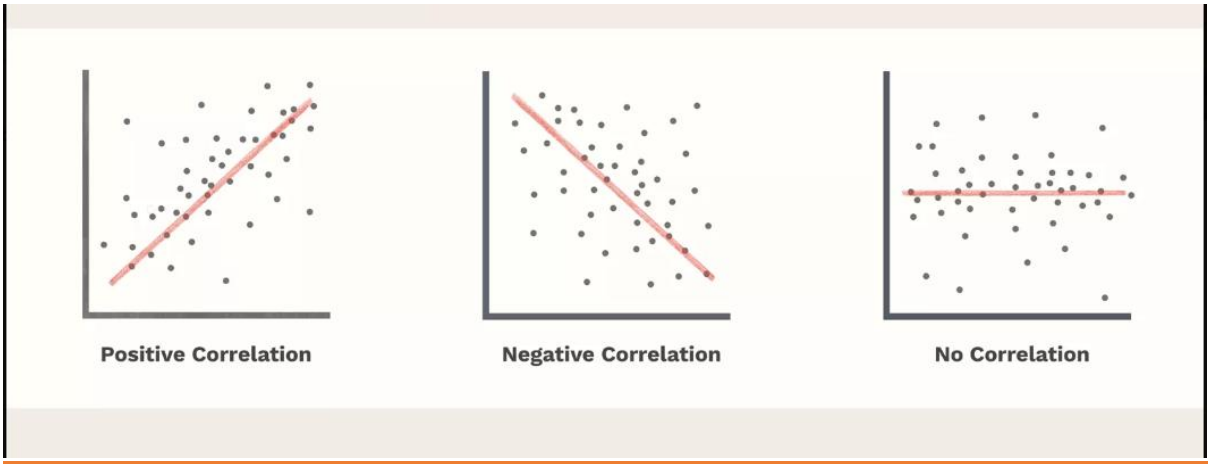


More amount is credited to goods that cost within 10 lakhs

E. Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

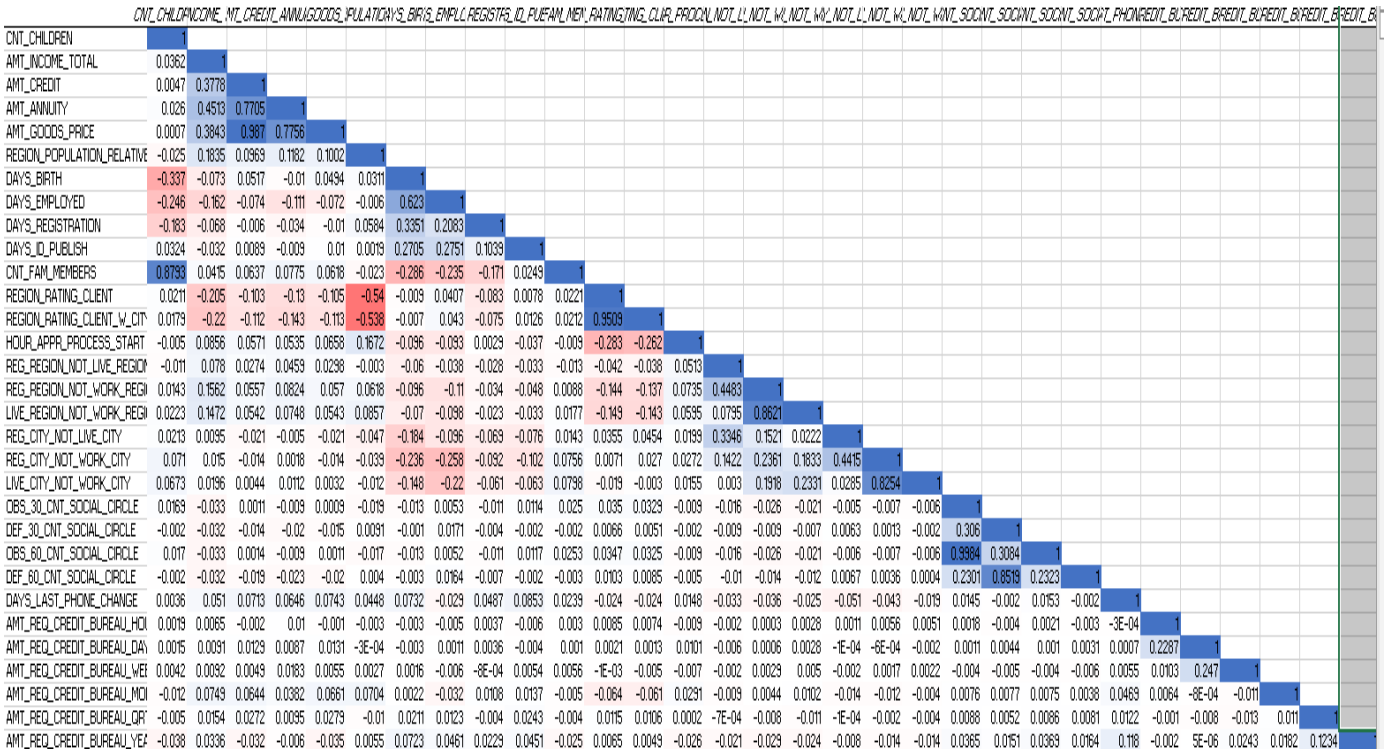
- Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.
- Hint: Utilize Excel functions like CORREL to calculate correlation coefficients between variables and the target variable within each segment. Rank the correlations to identify the top indicators of loan default for each scenario.
- Graph suggestion: Create correlation matrices or heatmaps to visualize the correlations between variables within each segment. Highlight the top correlated variables for each scenario using different colors or shading.

  Note: and A correlation coefficient is a number between -1 and 1 that tells us the strength and direction of a relationship between variables, tells us how similar the measurements of two or more variables are across a dataset.
  - 1 value indicates a **perfect positive correlation** – all data points align in a straight line
  - -1 value indicates a **perfect negative/inverse correlation** - all data points align in a straight line
  - 0 value indicates **no linear relationship** or a weak correlation.
  - Closer to 0 – weaker correlation
  - Closer to 1 or -1 – stronger correlation



Output and Visual Representation:

Correlation of variables for Repayers

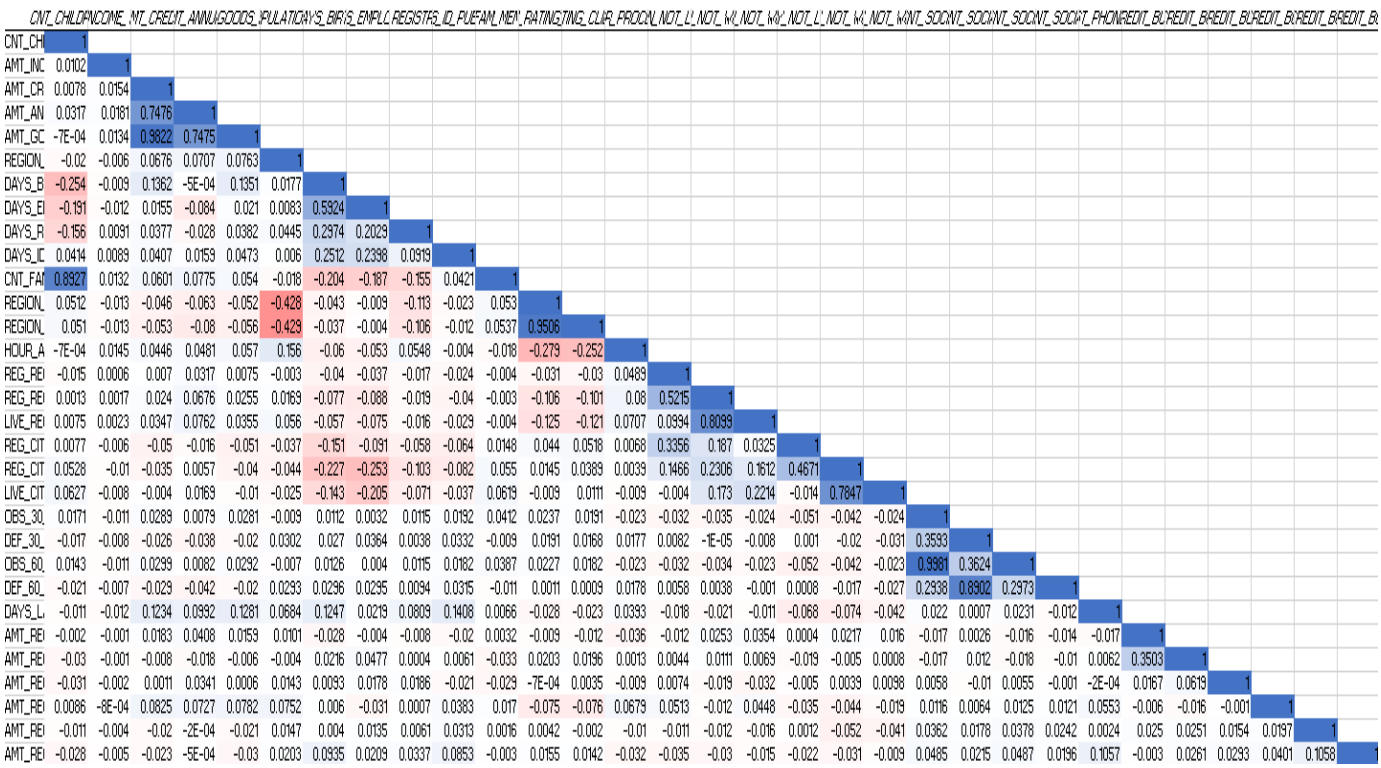


	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	HOUR_APPR_PROCESS_START	REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	DAYS_LAST_PHONE_CHANGE	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1																														
AMT_INCOME_TOTAL	0.0362	1																													
AMT_CREDIT	0.0047	0.3778	1																												
AMT_ANNUITY	0.026	0.4513	0.7705	1																											
AMT_GOODS_PRICE	0.0007	0.3843	0.987	0.7756	1																										
REGION_POPULATION_RELATIVE	-0.025	0.1835	0.0969	0.1182	0.1002	1																									
DAYS_BIRTH	-0.337	-0.073	0.0517	-0.01	0.0494	0.0311	1																								
DAYS_EMPLOYED	-0.246	-0.162	-0.074	-0.111	-0.072	-0.006	0.623	1																							
DAYS_REGISTRATION	-0.183	-0.068	-0.006	-0.034	-0.01	0.0584	0.3351	0.2083	1																						
DAYS_ID_PUBLISH	0.0324	-0.032	0.0089	-0.009	0.01	0.0019	0.2705	0.2751	0.1039	1																					
CNT_FAM_MEMBERS	0.8793	0.0415	0.0637	0.0775	0.0618	-0.023	-0.286	-0.235	-0.171	0.0249	1																				
REGION_RATING_CLIENT	0.0211	-0.205	-0.103	-0.13	-0.105	-0.54	-0.009	0.0407	-0.083	0.0078	0.0221	1																			
REGION_RATING_CLIENT_W_CITY	0.0179	-0.22	-0.112	-0.143	-0.113	-0.538	-0.007	0.043	-0.075	0.0126	0.0212	0.9509	1																		
HOUR_APPR_PROCESS_START	-0.005	0.0856	0.0571	0.0535	0.0658	0.1672	-0.096	-0.093	0.0029	-0.037	-0.003	-0.283	-0.262	1																	
REG_REGION_NOT_LIVE_REGION	-0.011	0.078	0.0274	0.0459	0.0298	-0.003	-0.06	-0.038	-0.028	-0.033	-0.013	-0.042	-0.038	0.0513	1																
REG_REGION_NOT_WORK_REGION	0.0143	0.1562	0.0557	0.0824	0.057	0.0618	-0.096	-0.11	-0.034	-0.048	0.0088	-0.144	-0.137	0.0735	0.4483	1															
LIVE_REGION_NOT_WORK_REGION	0.0223	0.1472	0.0542	0.0748	0.0543	0.0857	-0.07	-0.098	-0.023	-0.033	0.0177	-0.149	-0.143	0.0595	0.0795	0.8621	1														
REG_CITY_NOT_LIVE_CITY	0.0213	0.0095	-0.021	-0.005	-0.021	-0.047	-0.184	-0.096	-0.069	-0.076	0.0143	0.0355	0.0454	0.0139	0.3346	0.1521	0.0222	1													
REG_CITY_NOT_WORK_CITY	0.071	0.015	-0.014	0.0018	-0.014	-0.039	-0.236	-0.258	-0.092	-0.102	0.0756	0.0071	0.027	0.0272	0.1422	0.2361	0.1833	0.4415	1												
LIVE_CITY_NOT_WORK_CITY	0.0673	0.0196	0.0044	0.0112	0.0032	-0.012	-0.148	-0.22	-0.061	-0.063	0.0798	-0.019	-0.003	0.0155	0.003	0.1918	0.2331	0.0285	0.8254	1											
OBS_30_CNT_SOCIAL_CIRCLE	0.0169	-0.033	0.0011	-0.009	0.0009	-0.019	-0.013	0.0053	-0.011	0.0114	0.025	0.035	0.0329	-0.009	-0.016	-0.026	-0.021	-0.005	-0.007	-0.006	1										
DEF_30_CNT_SOCIAL_CIRCLE	-0.002	-0.032	-0.014	-0.02	-0.015	0.0091	-0.001	0.0171	-0.004	-0.002	-0.002	0.0066	0.0051	-0.002	-0.009	-0.009	-0.007	0.0063	0.0013	-0.002	0.306	1									
OBS_60_CNT_SOCIAL_CIRCLE	0.017	-0.033	0.0014	-0.009	0.0011	-0.017	-0.013	0.0052	-0.011	0.0117	0.0253	0.0347	0.0325	-0.009	-0.016	-0.026	-0.021	-0.006	-0.007	-0.006	0.9984	0.3084	1								
DEF_60_CNT_SOCIAL_CIRCLE	-0.002	-0.032	-0.019	-0.023	-0.02	0.004	-0.003	0.0164	-0.007	-0.002	-0.003	0.0103	0.0085	-0.005	-0.01	-0.014	-0.012	0.0067	0.0036	0.0004	0.2301	0.8519	0.2323	1							
DAYS_LAST_PHONE_CHANGE	0.0036	0.051	0.0713	0.0646	0.0743	0.0448	0.0732	-0.029	0.0487	0.0853	0.0239	-0.024	-0.024	0.0148	-0.033	-0.036	-0.025	-0.051	-0.043	-0.019	0.0145	-0.002	0.0153	-0.002	1						
AMT_REQ_CREDIT_BUREAU_HOUR	0.0019	0.0065	-0.002	0.01	-0.001	-0.003	-0.003	-0.005	0.0037	-0.006	0.003	0.0085	0.0074	-0.009	-0.002	0.0003	0.0028	0.0011	0.0056	0.0051	0.0018	-0.004	0.0021	-0.003	-3E-04	1					
AMT_REQ_CREDIT_BUREAU_DAY	0.0015	0.0091	0.0129	0.0087	0.0131	-3E-04	-0.003	0.0011	0.0036	-0.004	0.001	0.0021	0.0013	0.0101	-0.006	0.0006	0.0028	-1E-04	-6E-04	-0.002	0.0011	0.0044	0.001	0.0031	0.0007	0.2287	1				
AMT_REQ_CREDIT_BUREAU_WEEK	0.0042	0.0032	0.0049	0.0183	0.0055	0.0027	0.0016	-0.006	-8E-04	0.0054	0.0056	-1E-03	-0.005	-0.007	-0.002	0.0029	0.005	-0.002	0.0017	0.0022	-0.004	-0.005	-0.004	-0.006	0.0055	0.0103	0.247	1			
AMT_REQ_CREDIT_BUREAU_MON	-0.012	0.0749	0.0644	0.0382	0.0661	0.0704	0.0022	-0.032	0.0108	0.0137	-0.005	-0.064	-0.061	0.0291	-0.009	0.0044	0.0102	-0.014	-0.012	-0.004	0.0076	0.0077	0.0075	0.0038	0.0469	0.0064	-8E-04	-0.011	1		
AMT_REQ_CREDIT_BUREAU_QRT	-0.005	0.0154	0.0272	0.0095	0.0279	-0.01	0.0211	0.0123	-0.004	0.0243	-0.004	0.0115	0.0106	0.0002	-7E-04	-0.008	-0.011	-1E-04	-0.002	-0.004	0.0088	0.0052	0.0086	0.0081	0.0122	-0.001	-0.008	-0.013	0.011	1	
AMT_REQ_CREDIT_BUREAU_YEAR	-0.038	0.0336	-0.032	-0.006	-0.035	0.0055	0.0723	0.0461	0.0229	0.0451	-0.025	0.0065	0.0049	-0.026	-0.021	-0.029	-0.024	-0.008	-0.014	-0.014	0.0365	0.0151	0.0369	0.0164	0.118	-0.002	5E-06	0.0243	0.0182	0.1234

Top 10 Correlations for Repayers

Sl.No	Correlated Columns	Correlation Value
1	OBS_60_CNT_SOCIAL_CIRCLE and OBS_30_CNT_SOCIAL_CIRCLE	0.998371233
2	previous_application-to merge.AMT_GOODS_PRICE and previous_application-to merge.AMT_APPLICATION	0.989047437
3	AMT_GOODS_PRICE and AMT_CREDIT	0.986992533
4	previous_application-to merge.AMT_CREDIT and previous_application-to merge.AMT_APPLICATION	0.974409547
5	REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT	0.950868745
6	CNT_FAM_MEMBERS and CNT_CHILDREN	0.87927628
7	LIVE_REGION_NOT_WORK_REGION and REG_REGION_NOT_WORK_REGION	0.862137383
8	LIVE_CITY_NOT_WORK_CITY and REG_CITY_NOT_WORK_CITY	0.825395977
9	AMT_GOODS_PRICE and AMT_ANNUITY	0.775584263
10	AMT_ANNUITY and AMT_CREDIT	0.77052209

Correlation of variables for Defaulters:



Top 10 Correlations for Defaulters		
Sl.No	Correlated Columns	Correlatio
1	OBS_60_CNT_SOCIAL_CIRCLE and OBS_30_CNT_SOCIAL_CIRCLE	0.9980871
2	previous_application-to merge.AMT_GOODS_PRICE and previous_application-to merge.AMT_APPLICATION	0.9867197
3	AMT_GOODS_PRICE and AMT_CREDIT	0.9822434
4	previous_application-to merge.AMT_CREDIT and previous_application-to merge.AMT_APPLICATION	0.9739561
5	REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT	0.9505936
6	CNT_FAM_MEMBERS and CNT_CHILDREN	0.8927315
7	LIVE_REGION_NOT_WORK_REGION and REG_REGION_NOT_WORK_REGION	0.8099127
8	LIVE_CITY_NOT_WORK_CITY and REG_CITY_NOT_WORK_CITY	0.7847459
9	AMT_ANNUITY and AMT_CREDIT	0.7476435
10	AMT_GOODS_PRICE and AMT_ANNUITY	0.7475234

Key Hypothesis/Insight: Few key hypothesis after analysis to detect possible defaulters are,

- CODE_GENDER: Male customers tend to default more than females
- FLAG_OWN_CAR: Customers who do not own a car tend to default more
- CNT_CHILDREN: Customers with 9 or more than 9 children have higher (100%) rate of defaulting
- AMT_INCOME_TOTAL: Applicants with less than 2 lakh income has high probability of defaulting
- AMT_CREDIT: Highest defaulters are those who have taken loans between the range of 4 to 6 lakhs
- AMT_ANNUITY: Most of the default is done by people who pay an annuity of below 50,000
- AMT_GOOD_PRICE: The highest default is done by a person who has taken a loan for goods that cost below 10 lakhs
- NAME_INCOME_TYPE: Customers who are unemployed tend to default more
- NAME_EDUCATION_TYPE: Customers with lower secondary, incomplete higher and secondary special education tend to default more
- NAME_FAMILY_STATUS: Customers who are single or had civil marriage tend to default more. Widows are the safest as they have low default rate.
- NAME_HOUSING_TYPE: Customers living with parents or in rented apartments are more likely to default.
- YEARS_EMPLOYED: The applicants with 0 to 9 years of experience have high rate of defaulting a loan
- OCCUPATION_TYPE: Avoid granting loans to low skilled labourers, drivers, Realty agents and Security staff as they tend to default more.
- CNT_FAM_MEMBERS: Family of a greater number of people tend to default more, maybe due to more constraints and need

Solutions to mitigate the risk of defaulting

- Avoid approving loans to customers who satisfy above criteria
- Grant loan at a higher interest rate

- Reducing the amount of loan

Result: This project helped me to understand

- To work on multiple huge dataset and data clean them.
- The process of data cleaning. First start with deleting duplicate values using “Remove Duplicate” option of Excel. Then drop unwanted columns followed by deleting rows with many missing values. Also check if missing values can be fetched and retained with the help of internet. Finally, look for errors and correct them.
- Helped me understand how to find the outlier and use best strategy to remove or replace or keep them depending on the situation.
- Has made me think and apply logic and use the best out of them with proper explanation so that the audience can relate with it.
- Have used “Pivot Tables” widely in this project and
- This project helped me to think and explore various factors to analyse the patterns and trends that will help a bank to improve their business by keeping the defaulters away.
- How to retrieve/extract the answers and what technique or formulas to use to achieve it.
- Has made me think like an analyst, considering what makes the entire thing perfect with relevant and valuable insights.
- Along with approach and output, visual representation plays a very important role for faster understanding of analysis and this project has made me more confident and aware about various options that are available.
- I have gained knowledge about univariate, segmented univariate and bivariate analysis.
- Got a chance to use “DATA ANALYSIS” tool of excel for correlation analysis.
- Very importantly, worked on “**Risk Analysis”** to analyse patterns in the data and ensure that capable applicants are not rejected by recognizing **False Positives and False Negatives.**
- Risk analysis helped me in identifying and assessing risks that may jeopardize banks success

Links: My excel worksheet link with different sheets for each task,

previous_application dataset after cleaning,

https://docs.google.com/spreadsheets/d/1mmrtNDNblwXe7RRtr7iYWuNwCVIMpW4e/edit?usp=sharing&ouid=108154584635151678812&rtpof=true&sd=true

application_data dataset which consists all the tasks and merged file,

https://docs.google.com/spreadsheets/d/1hwEJ0oO13jqmuLt6ya0fmSjwh8NeVFqN/edit?usp=sharing&ouid=108154584635151678812&rtpof=true&sd=true

You can connect with me on LinkedIn account,

https://www.linkedin.com/in/raksha-nayak-41578738/

Loom Video link,

https://www.loom.com/share/fdf2d0bff6f645f7b389cfcf609be96c?sid=bb6fc88e-e5d7-456a-b246-b184ea7f2216