Bank Loan Case Study

Final Project – 2

Project Description: Analyzing data from a bank loan case study is the goal of this project. It's difficult for any finance company to lend different kinds of loans to urban customers.

Companies face two risks when a customer applies for a loan:

1. The firm suffers if the loan applicant can repay it but is denied approval.
2. The company suffers a loss of revenue if the loan is accepted but the applicant is unable to repay it.

By taking advantage of this some borrowers with insufficient credit histories fail to make loan payments.

The principal objective of this project is to discern trends that suggest if a client may experience trouble making their installment payments. Decisions on loan denial, loan amount reduction, and interest rate increase for high-risk borrowers can be made using this information. To make more informed decisions on loan approval, the company seeks to comprehend the primary causes of loan default.

There are four possible results when a customer applies for a loan:

Approved: The company has approved the loan application.

Cancelled: The customer canceled the application during the approval process.

Refused: The company rejected the loan.

Unused offer: The loan was approved but the customer did not use it.

To draw insights, I employ **Exploratory data analysis** to examine the trends and determine how loan and customer attributes affect the probability of default.

Approach:

I started by downloading the stakeholder-provided dataset. It has two datasets in it. 1. Application_data, 2. previous _application. I decided to use Excel as my data analysis tool for this project.

Second, after creating a copy of the data, began cleaning it. Eliminating columns that contain missing data by using the "count blank" function to

determine the percentage of blank shells in a column. then I dealt with the missing data using a suitable manner. after that, I began my analysis, gleaned significant insights, and then made recommendations.

Tech-Stack Used:

Microsoft Excel 2021 MSO(version 2403)

The Excel sheet is where all the analysis was completed. Excel is a tool for high-level visual summaries, trends, and patterns in data analysis. Additionally supplying pivot tables, charts, functions, and algorithms to support intelligent data interpretation.

Insights:

A. Identify Missing Data and Deal With it Appropriately

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.

Insights: The original dataset included 50,000 rows and 122 columns, many of which had missing values. So the dataset was trimmed down to 77 columns. This made the dataset easier to handle and more trustworthy for analysis.

Dealing with missing data

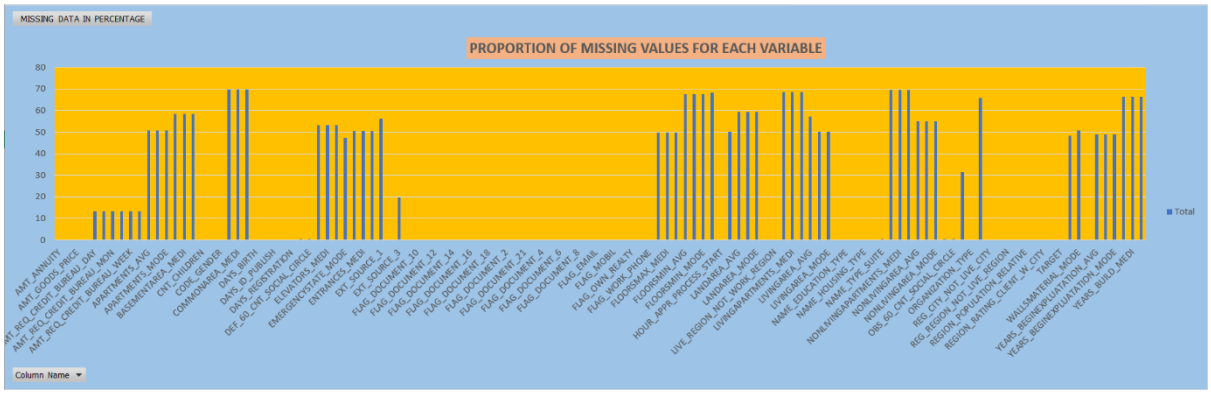
In cases where a column has fewer than 5% missing cells, I delete the rows containing the missing data. if a column has more than 5% missing cells and less than 30%, I perform imputations, replacing the mean in numerical columns and mode in categorical columns. If a column has more than 30% missing cells, I delete the entire column, assuming it is not essential to our analysis.

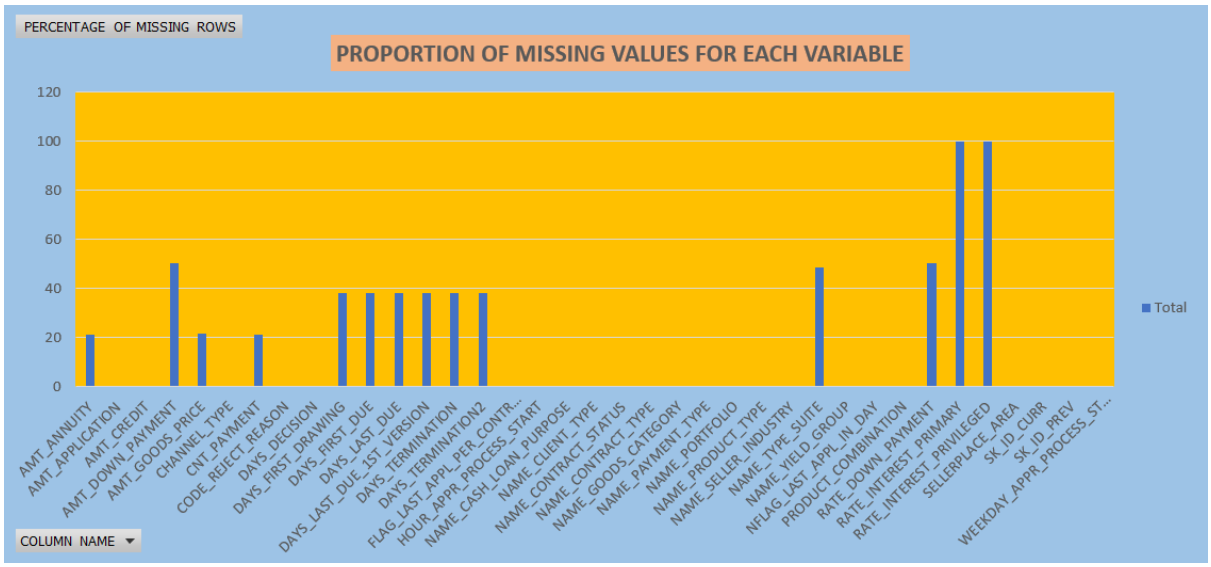
I colored – and highlighted the relevant column in the original dataset to help you understand data handling and cleaning.

I then looked for duplicate values. In the dataset, there are no duplicate values. I organized the Days_Birth column into Days_Birth_Years and like that some more columns. Additionally, I used the interquartile range approach while taking relevant variables into account to identify outliers. That's how I approached the data. Following this, I began understanding analysis to identify trends.

DEALING WITH MISSING DATA	
< 5%	columns with < 5% missing cell rows are deleted.
5% to 30%	columns with between 5% - 30% missing cell rows are replaced/imputed with mean and rounded it with nearest integer.
> 30%	columns with >30% missing cell are deleted .

Here missing data is identified and visually represented the proportion of missing values for each column with column charts for both data sets.





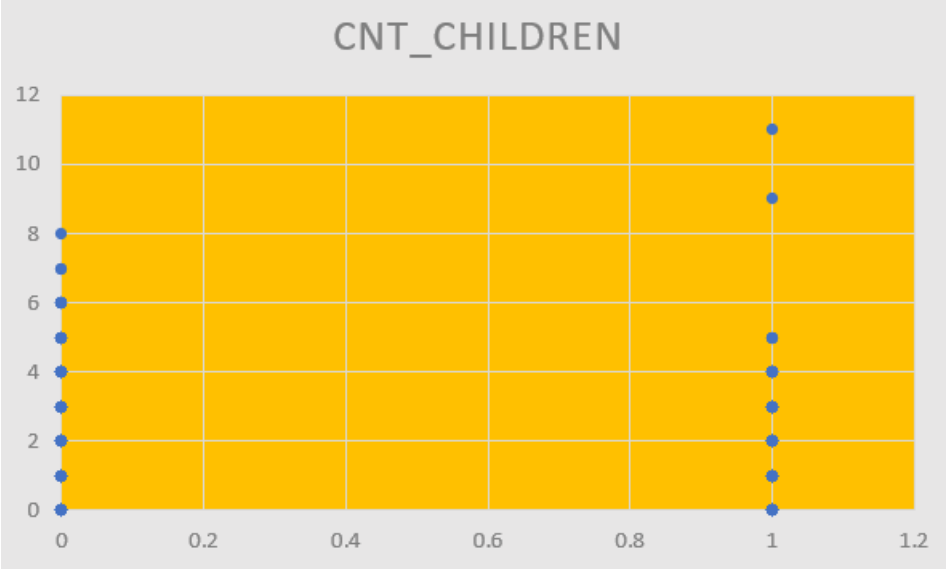
B. Identify outliers in the Dataset:

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

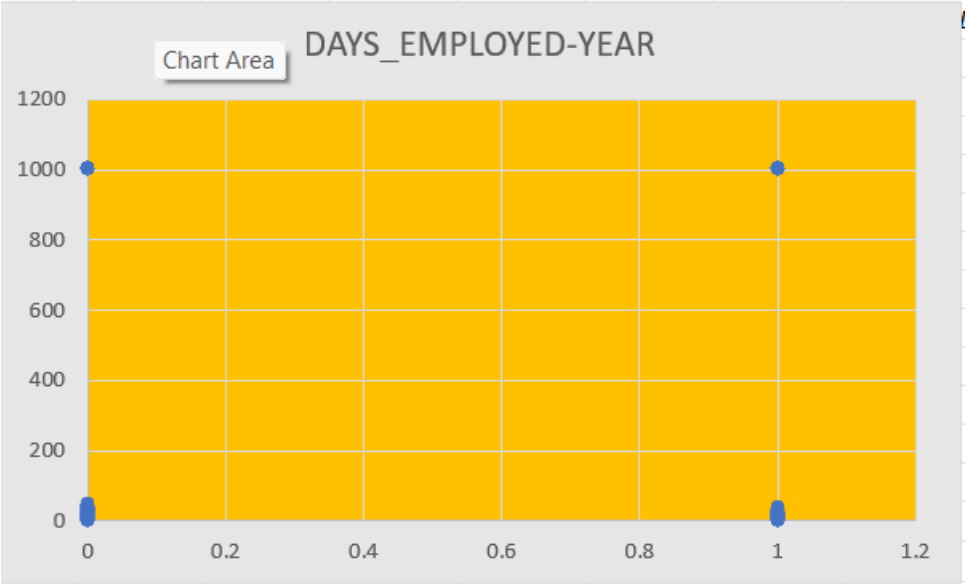
Insights:

- I use a scatter plot to show the numerical variable distribution in the dataset and draw attention to the presence of outliers.
- Within the provided dataset, I discovered two outliers in the column names “**CNT_CHILDREN”** and “**DAYS_EMPLOYED”.**

- Where the xy plotter shows that the applicant with target variable “1” has a maximum of 11 children, which is extremely unusual for a modern application. Whereas candidates with target variable “0” have a maximum of 8 children.

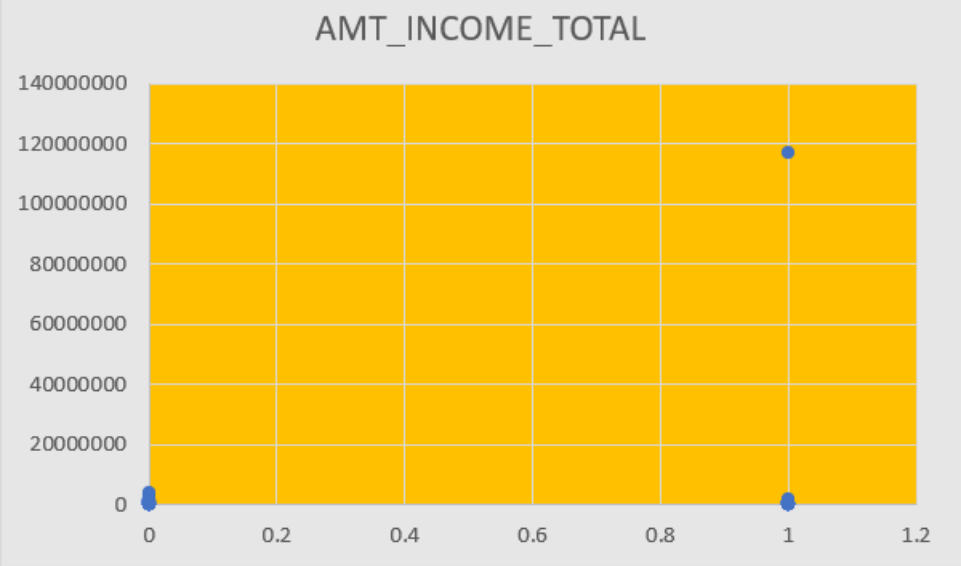


- Where days-employed shows that a few applicants have been employed for a **millennium**, and there are few candidates which, given that the majority of candidates have jobs that last **between 80-90 years**, is implausible.



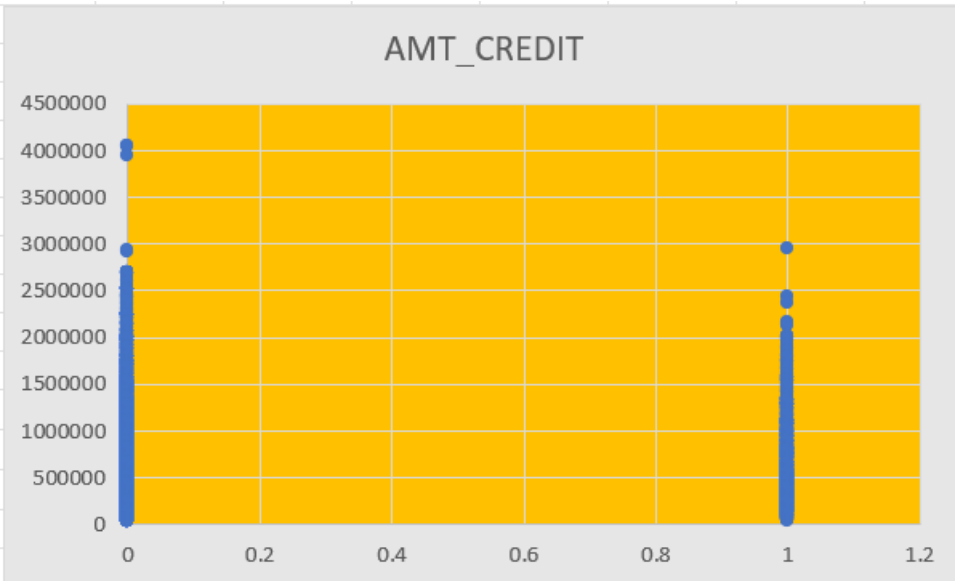
- Recognizing the outliers as invalid is crucial to ensuring the accuracy and reliability of the analytic results. In certain situations, further study and appropriate actions such as data cleaning, should be called for to maintain the integrity of the dataset.

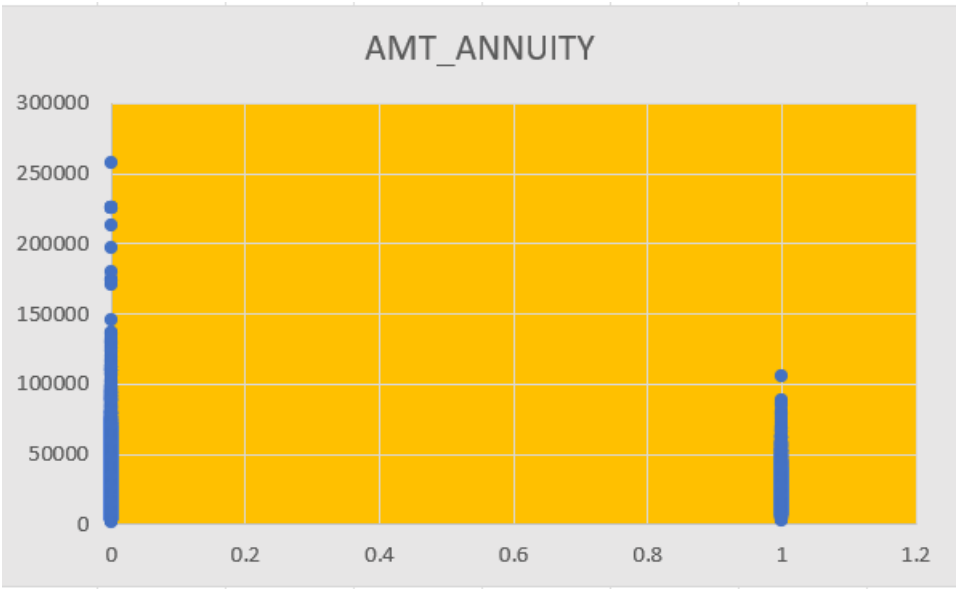
- Other outliers were also present, but in the context of business terminology, they are regarded as **legitimate data points** rather than outliers.



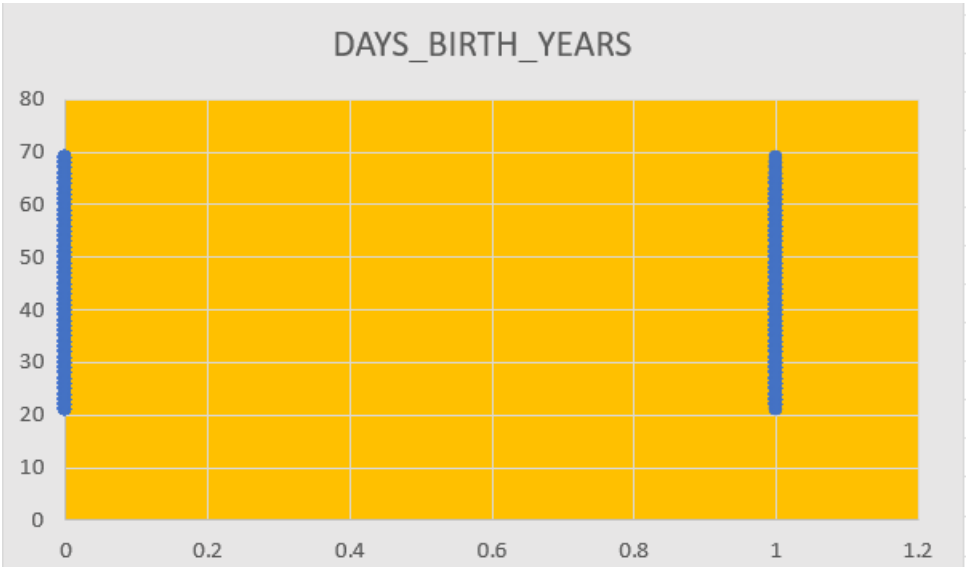
In the above chart variable target1 has an outlier. Here one of the applicants got an income total of 120000000. It is a huge difference for other applicants whose income is below 4000000 lacs. but it is a **valid data point**.

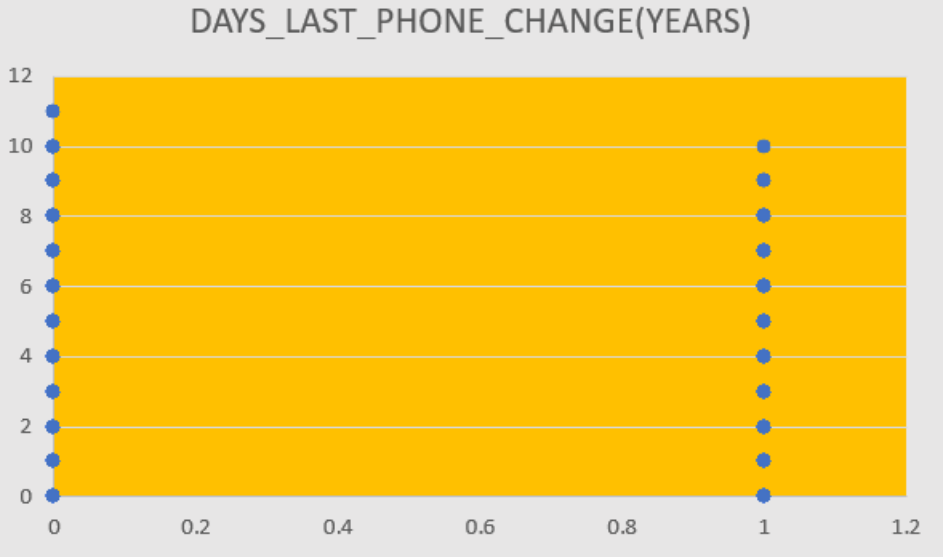
Below are some of the other attributes showing outliers.

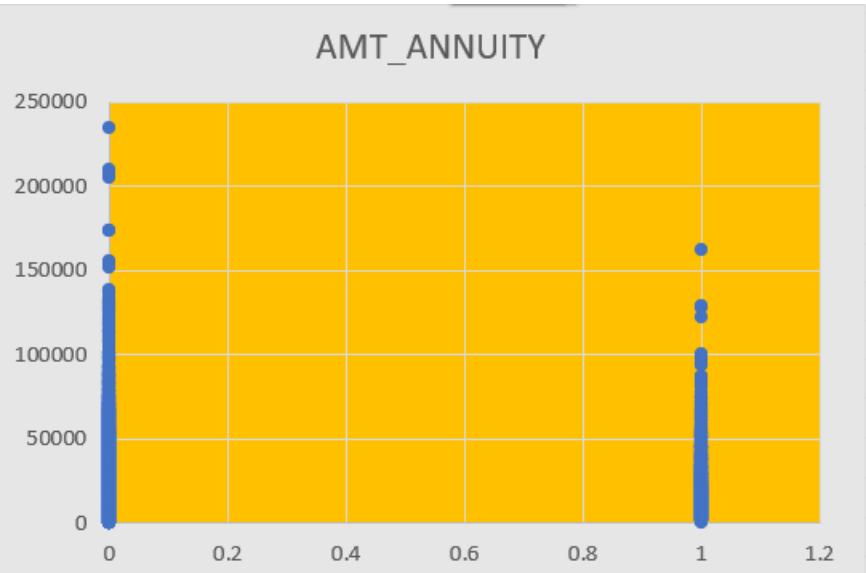


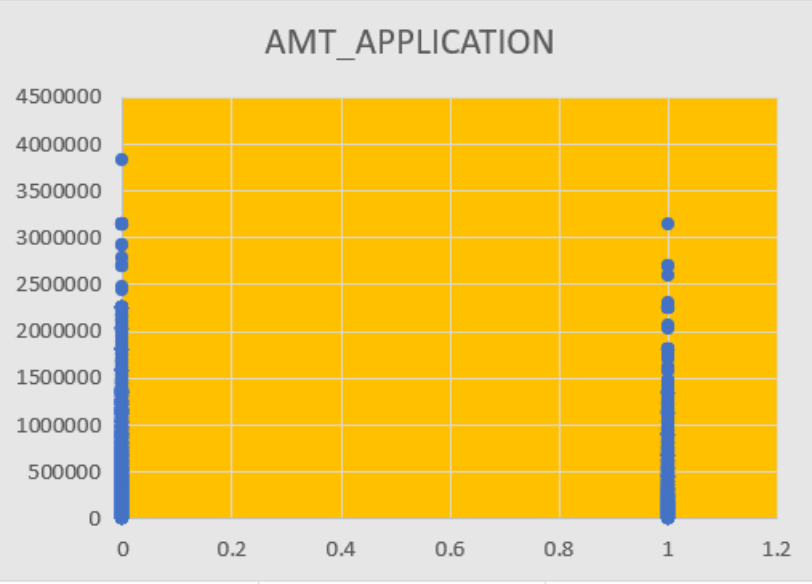

AMT_ANNUITY
300000
250000
200000
150000
100000
50000
0
0
0.2
0.4
0.6
0.8
1
1.2

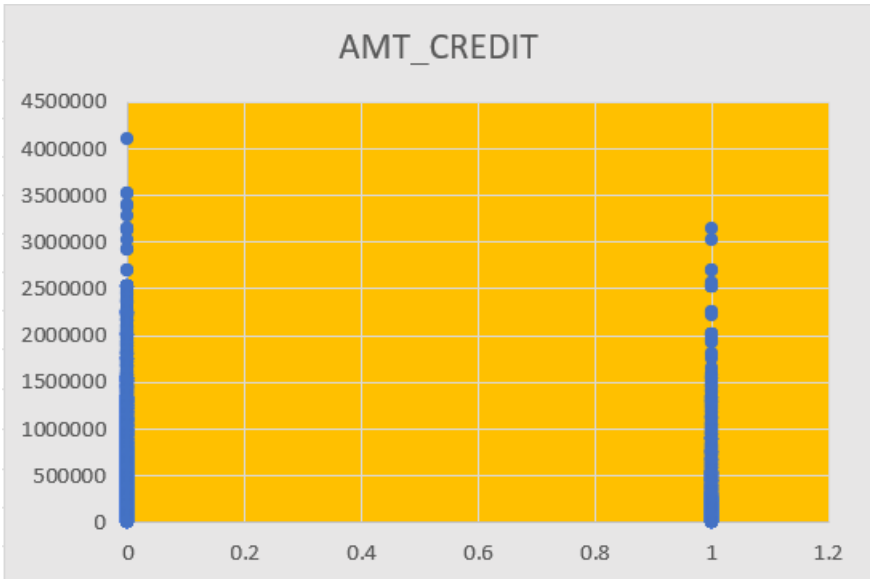

AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
0.2
0.4
0.6
0.8
1
1.2


DAYS_BIRTH_YEARS
80
70
60
50
40
30
20
10
0
0
0.2
0.4
0.6
0.8
1
1.2

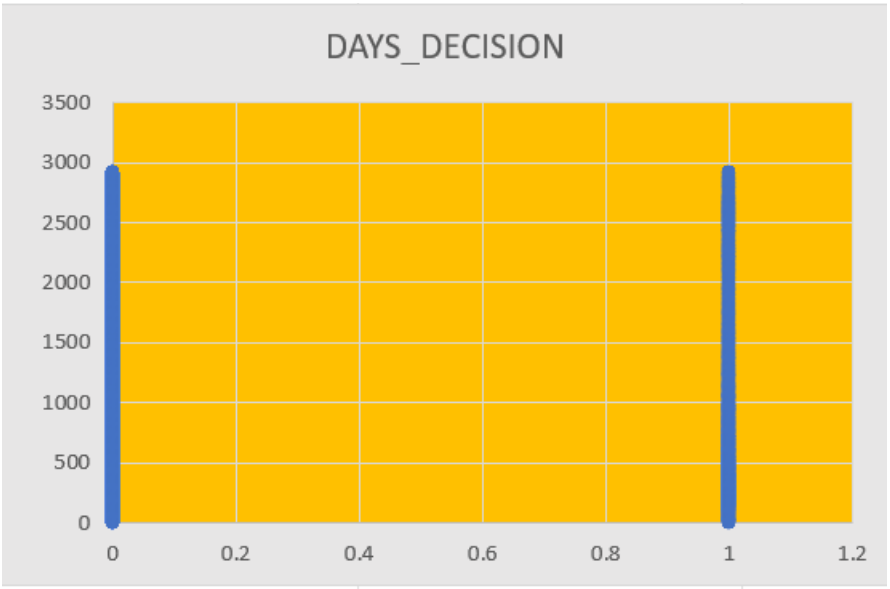

DAYS_LAST_PHONE_CHANGE(YEARS)
12
10
8
6
4
2
0
0
0.2
0.4
0.6
0.8
1
1.2

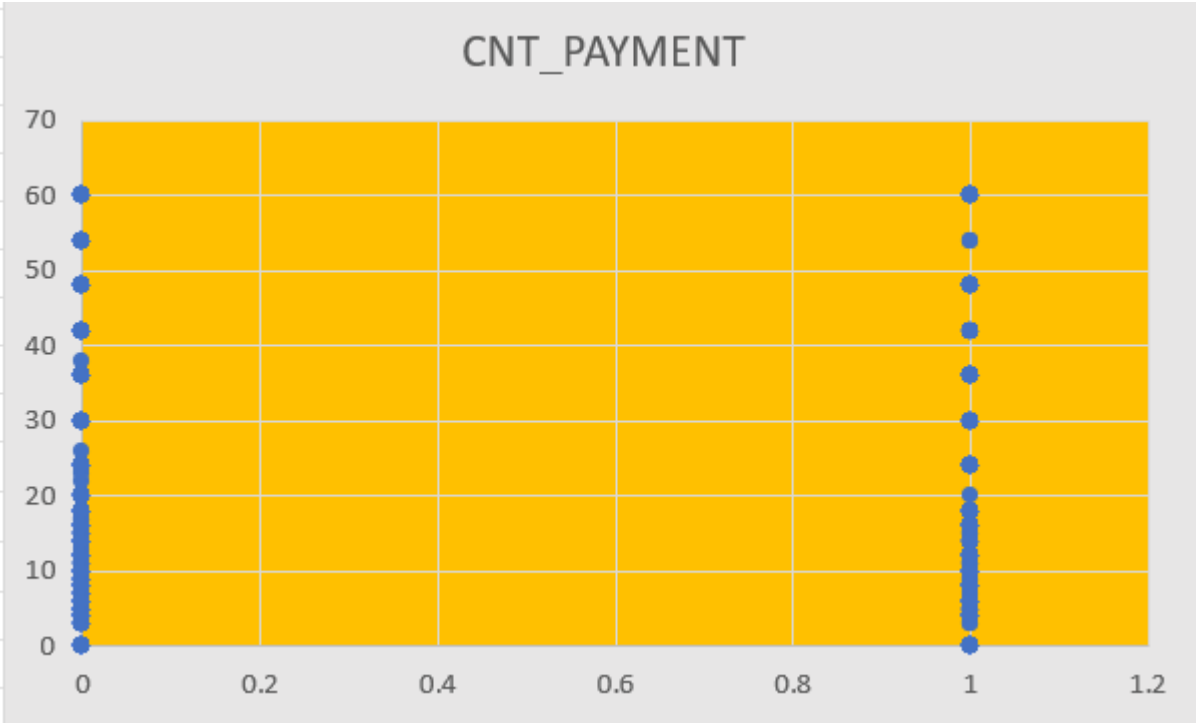

AMT_ANNUITY
250000
200000
150000
100000
50000
0
0
0.2
0.4
0.6
0.8
1
1.2


AMT_APPLICATION
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
0.2
0.4
0.6
0.8
1
1.2


AMT_CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
0.2
0.4
0.6
0.8
1
1.2


AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
0.2
0.4
0.6
0.8
1
1.2

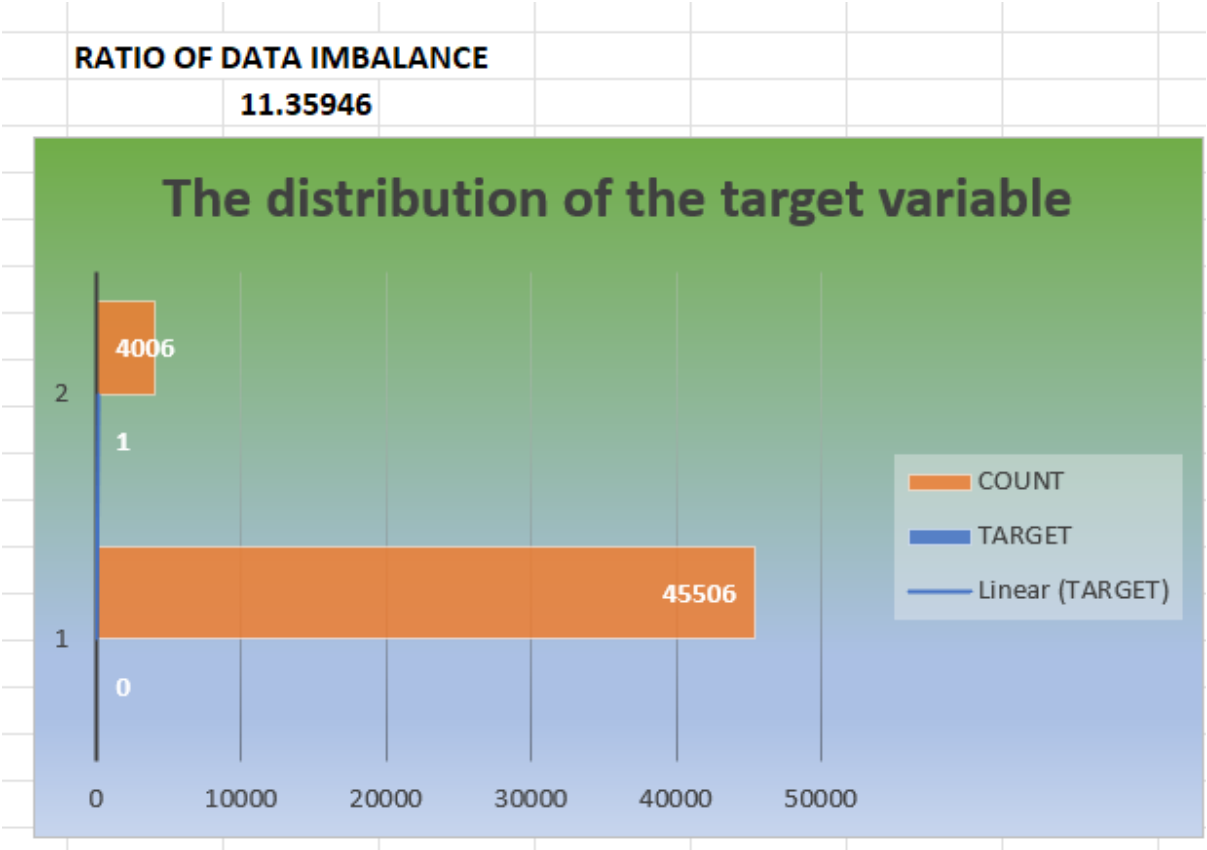

DAYS_DECISION
3500
3000
2500
2000
1500
1000
500
0
0
0.2
0.4
0.6
0.8
1
1.2



C.Analyze Data Imbalance:

Task: Determine if there is a data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

TARGE	COUN	PERCENTAGE
0	45506	91.90903215
1	4006	8.090967846
Total	49512	



Insights:

The ratio of all applicants experiencing payment issues (i.e. target 1) to all the applicants making timely payments (i.e. target 0) displayed in the above chart is **11.36** of the 45506 applicants, and **92%** of them submit their applicants on time, creating a majority class. Conversely, **8 %** of applicants (4006) experience difficulties, creating a minor class.

D.Perform Univariate, Segmented Univariate, and Bivariate Analysis:

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.

Univariate, Segmented univariate analysis:

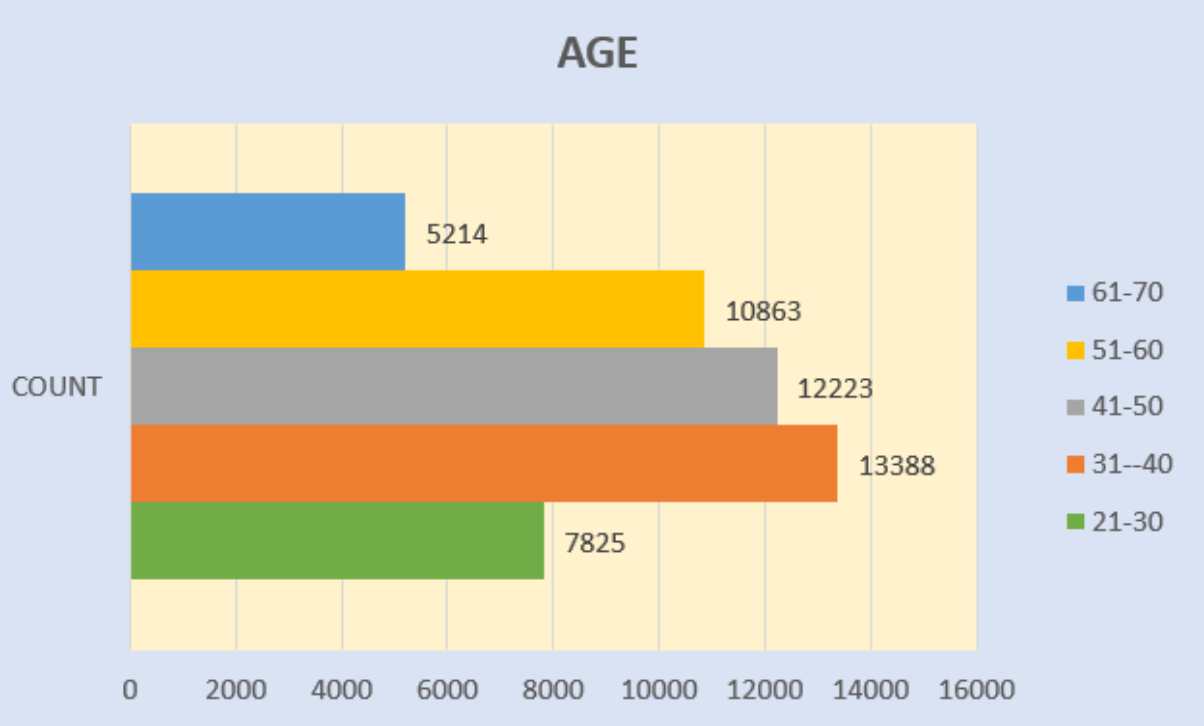
univariate, by definition, only takes into account one data variable when doing analysis. When a data variable is evaluated in subsets, it is referred to as segmented analysis. This type of analysis is highly beneficial as it allows

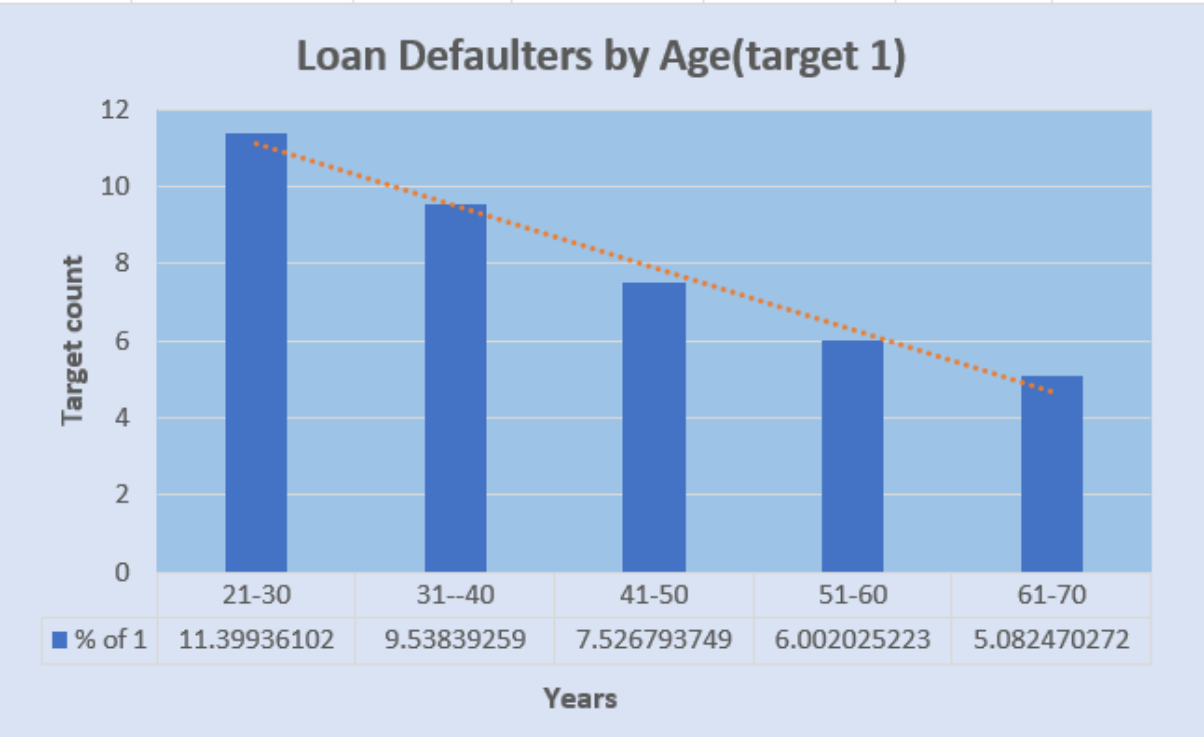
for the display of pattern change metrics across several segments of the same variable.

The below graphs show that univariate and segmented univariate for the same column names.

AGE:

CLASS BINS	COUNT	TARGET 0	TARGET 1	% of 0	% of 1
21-30	7825	6933	892	88.6006	11.3994
31--40	13388	12111	1277	90.4616	9.53839
41-50	12223	11303	920	92.4732	7.52679
51-60	10863	10211	652	93.998	6.00203
61-70	5214	4949	265	94.9175	5.08247

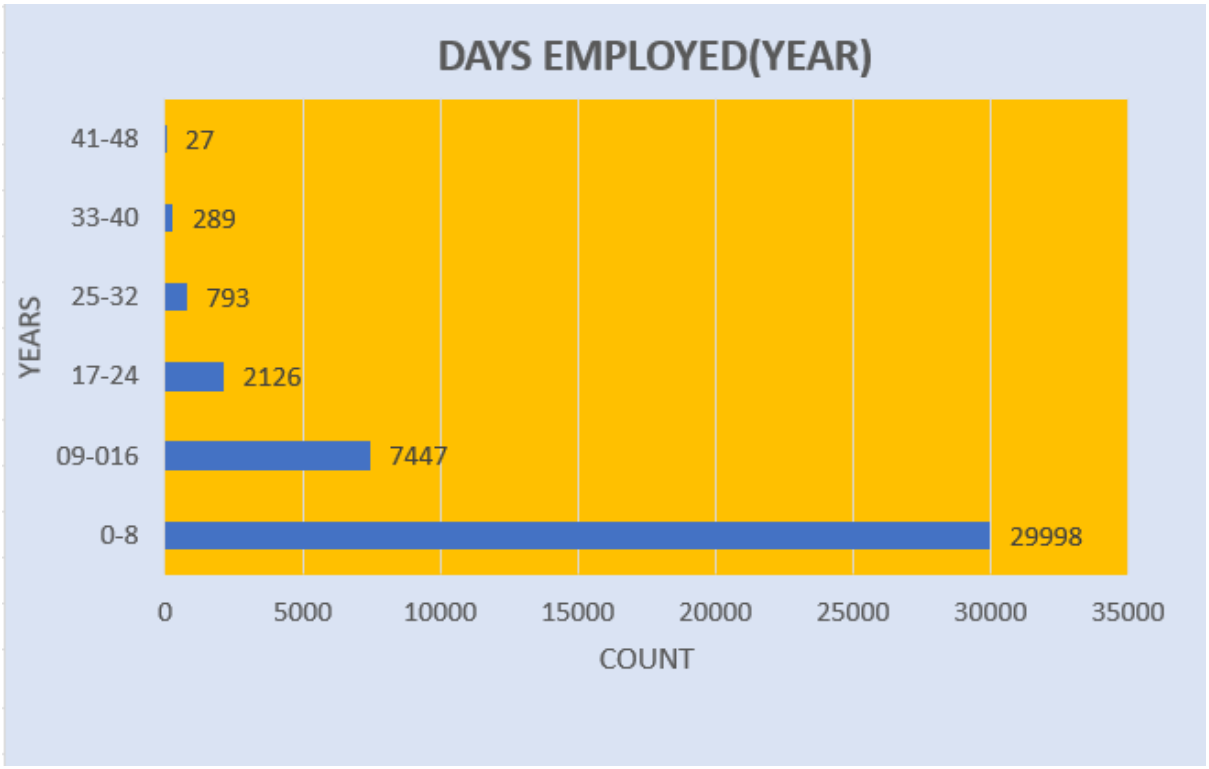


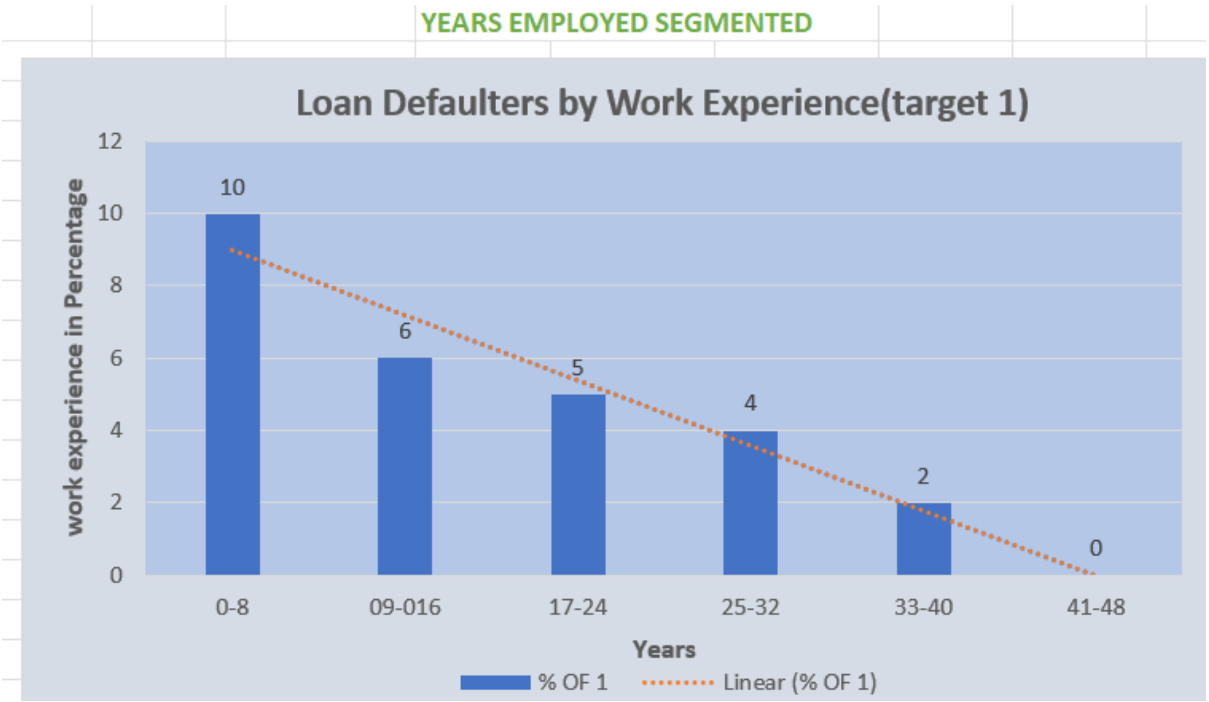


- The majority of the loan applicants were in the age group between 31-40 years.
- The majority of loan defaulters (target 1)are in the age group between 21-30 years. So, in the above graph, it clearly says that as the age increases loan defaulters percentage decreases.

Days_Eemployed:

CLASS BINS	COUNT	TARGET 0	TARGET 1	% OF 0	% OF 1
0-8	29998	27056	2942	90	10
09-016	7447	7026	421	94	6
17-24	2126	2019	107	95	5
25-32	793	763	30	96	4
33-40	289	283	6	98	2
41-48	27	27	0	100	0

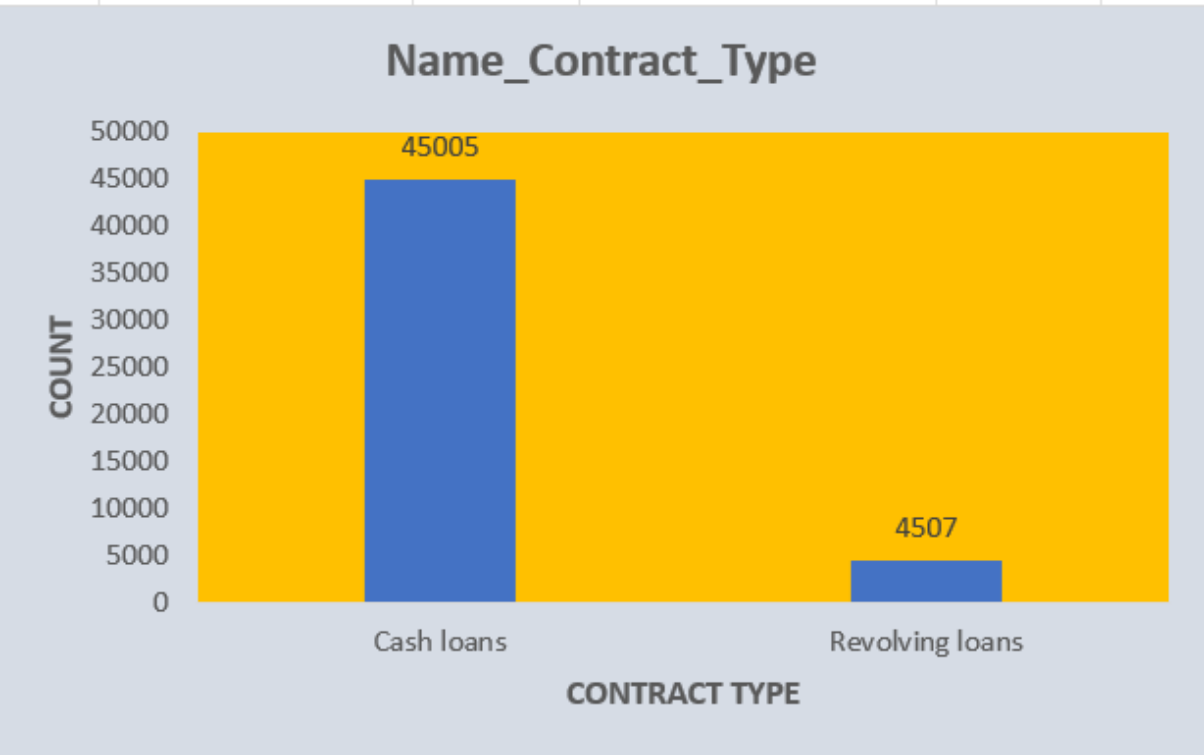


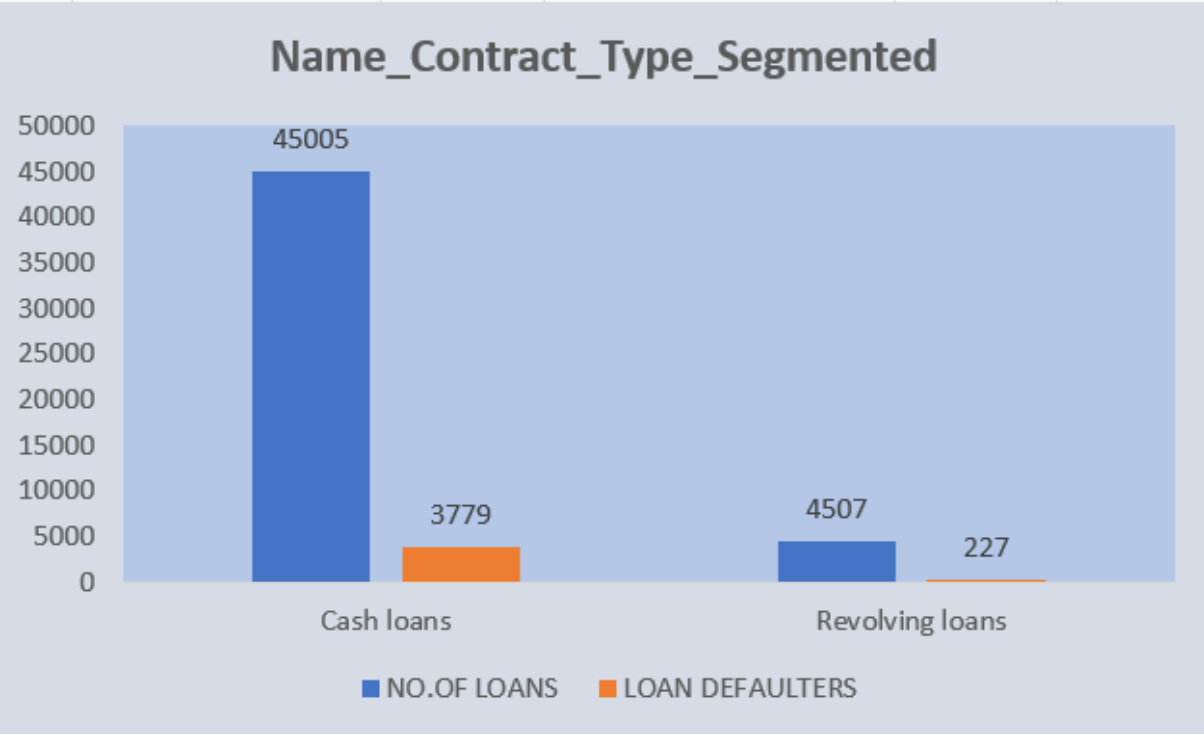


- Majority of loan applicants were employed for up to 8 years. Although most applicants worked for 0-8 years and also these are the highest loan defaulters percentage.
- It clearly says when the working years increase the loan defaulters decrease. And also the number of applicants applying for loans decreases.

Name_Contract_Type:

CONTRACT TYPE	NO.OF LOANS	LOAN DEFAULTERS
Cash loans	45005	3779
Revolving loans	4507	227

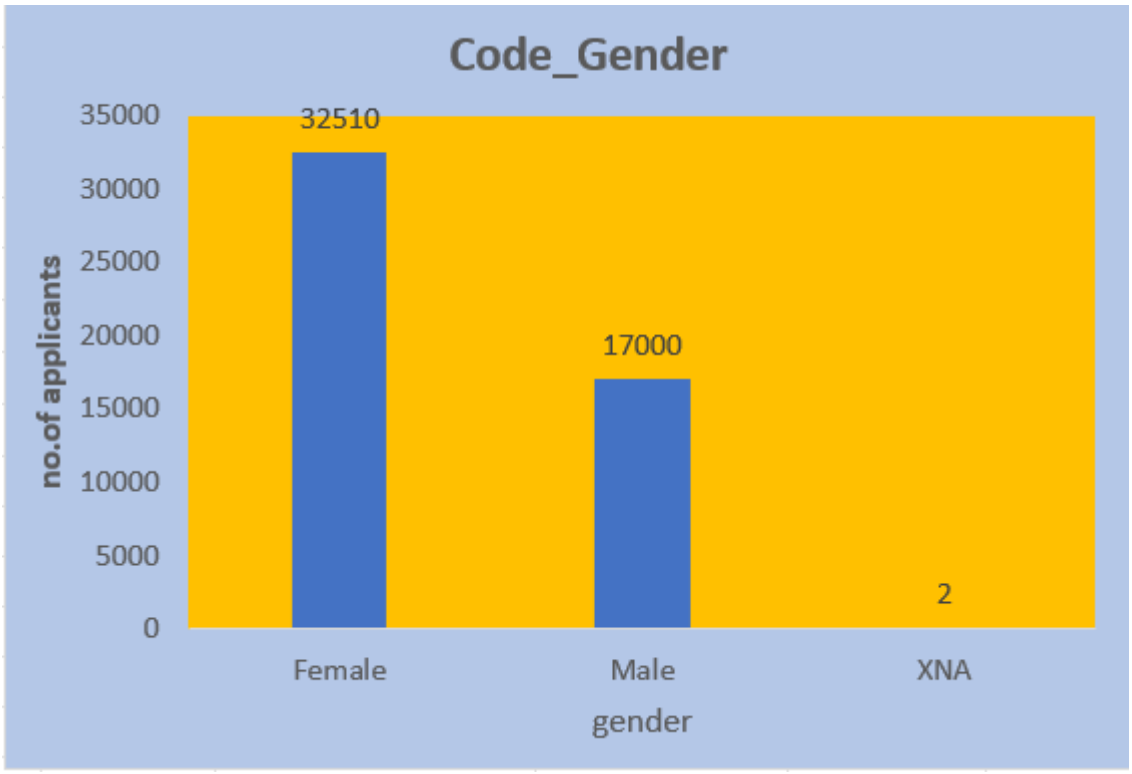


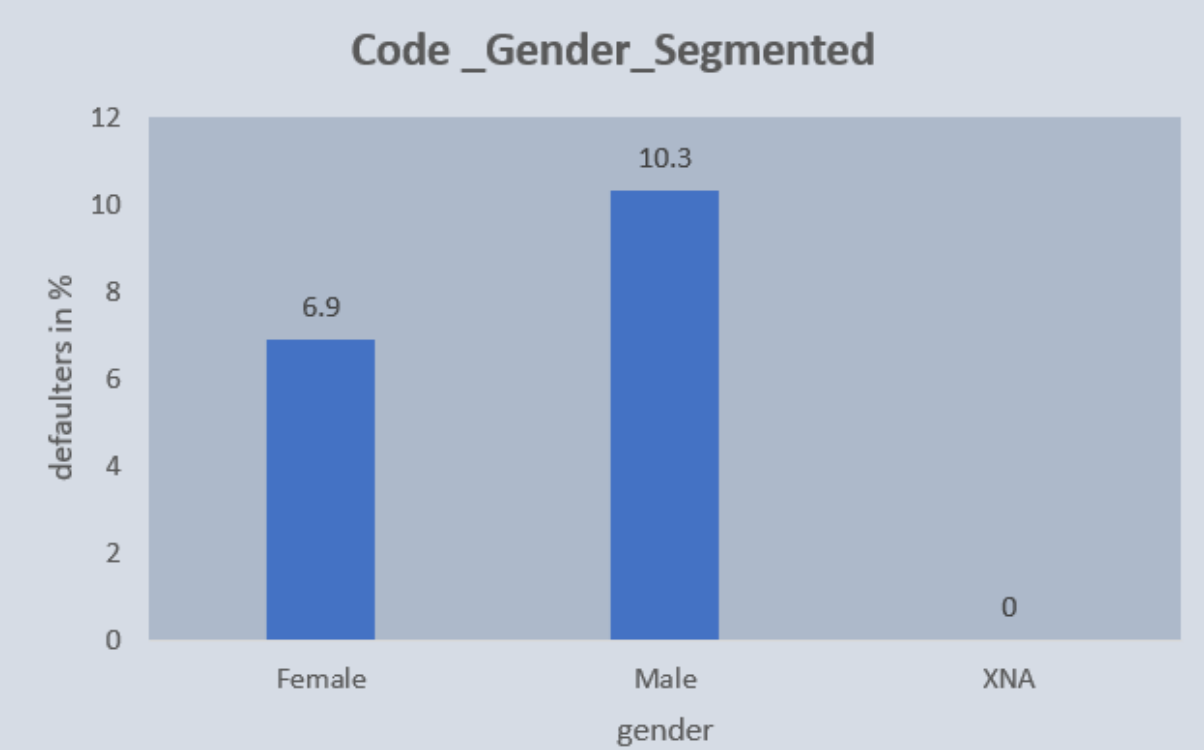


- The number of applicants for contract type is high for cash loans and low for revolving loans.
- Number of loans and loan defaulters are directly proportional to Name_Contract_Type.

Code_Gender:

Gender	no.of Applicants	Loan Default
Female	32510	6.9
Male	17000	10.3
XNA	2	0

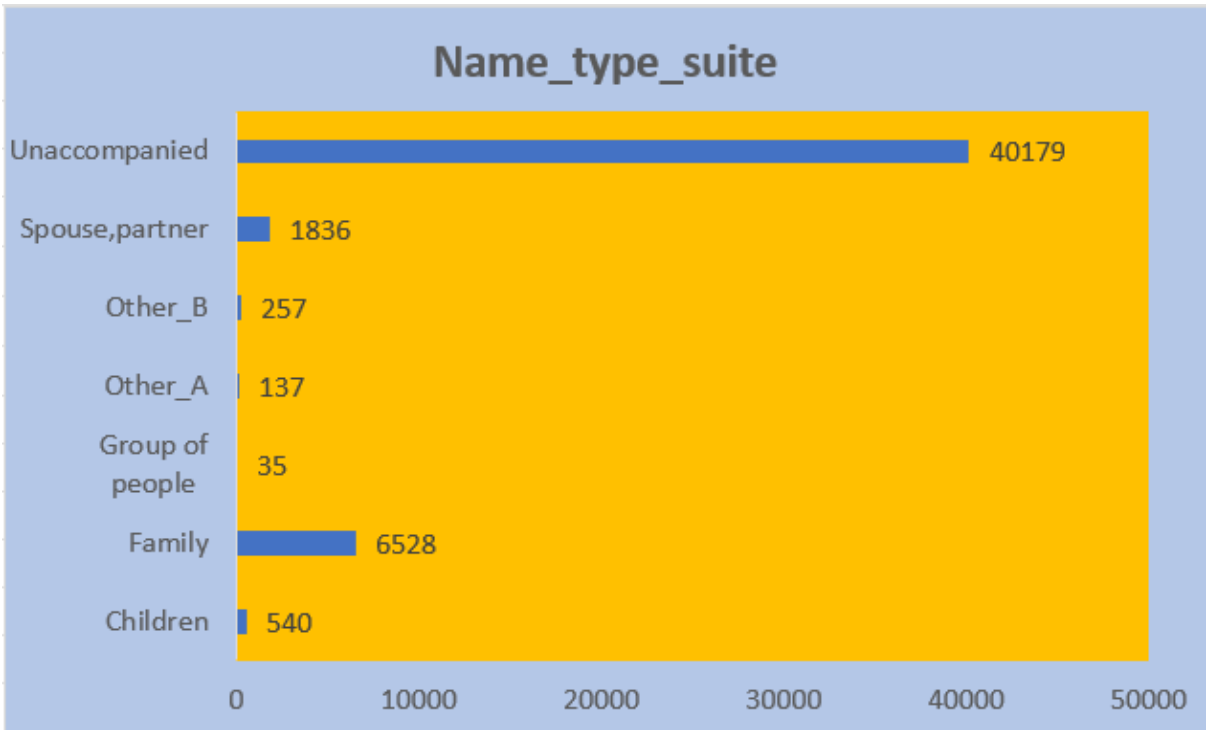


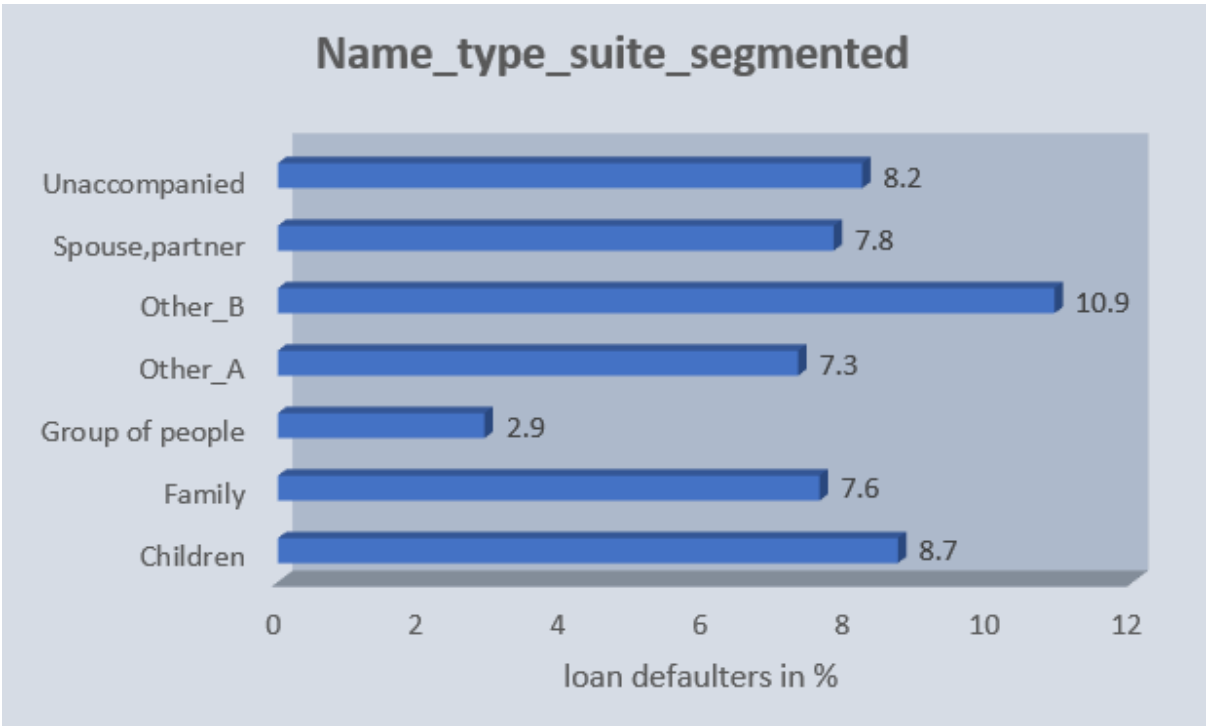


- Majority of loan applicants are from females. But when comes to loan defaulters females made payments on time compared to males.

Name_Type_Suite:

Name type suite	Applicants	Loan defaulters in %
Children	540	8.7
Family	6528	7.6
Group of people	35	2.9
Other_A	137	7.3
Other_B	257	10.9
Spouse,partner	1836	7.8
Unaccompanied	40179	8.2

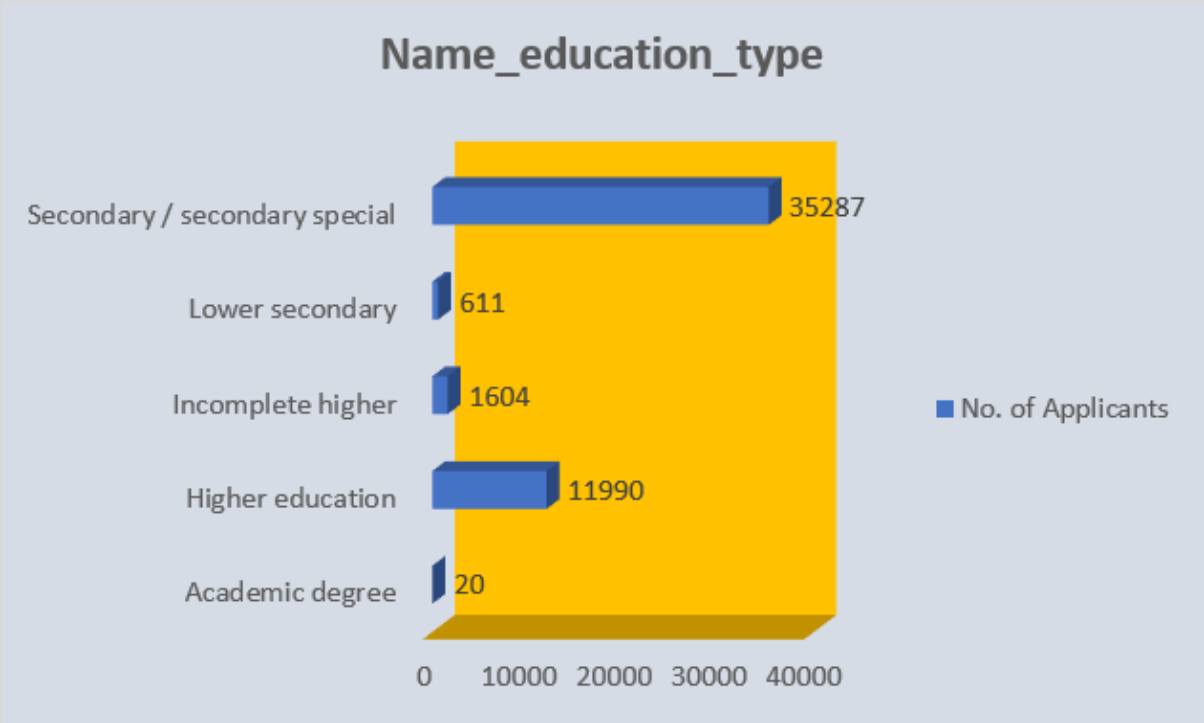


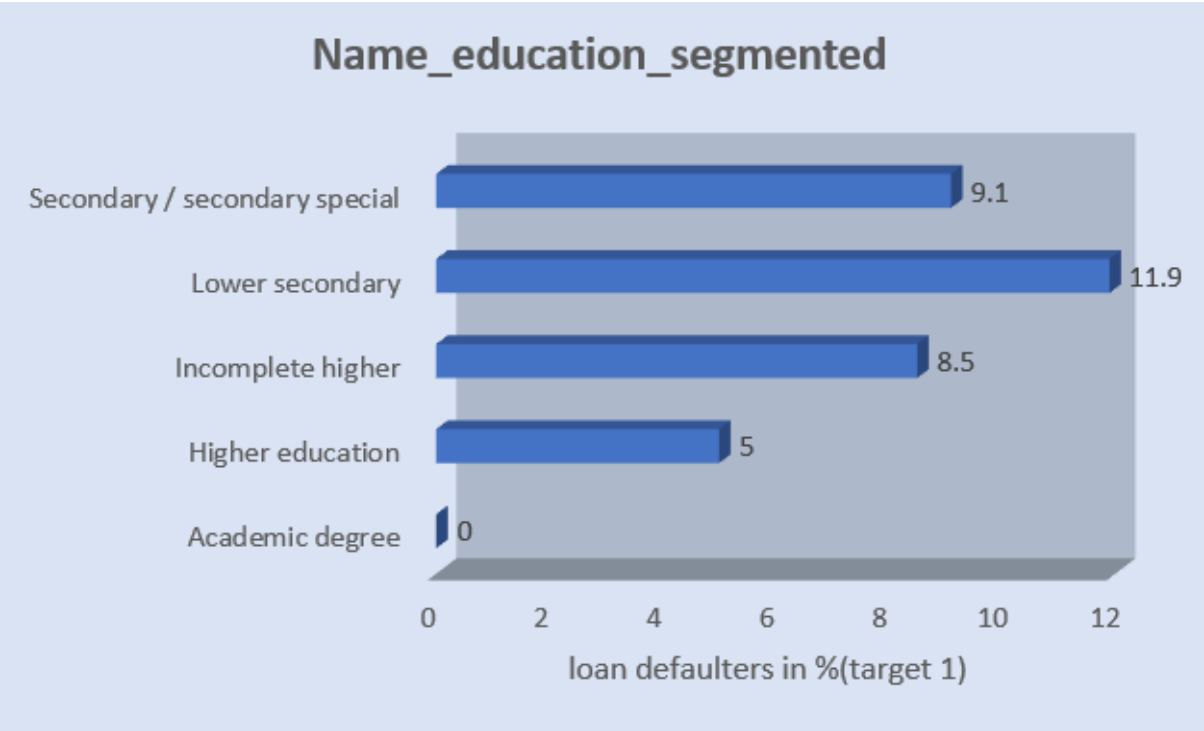


- Number of applicants from unaccompanied people is greater and low for group of people.
- When name_type segmented, loan defaulters are high from Other_B and low from Group of people.

Name_Education_Type:

Education type	No. of Applicants	Loan defaulters in %
Academic degree	20	0
Higher education	11990	5
Incomplete higher	1604	8.5
Lower secondary	611	11.9
Secondary / secondary special	35287	9.1

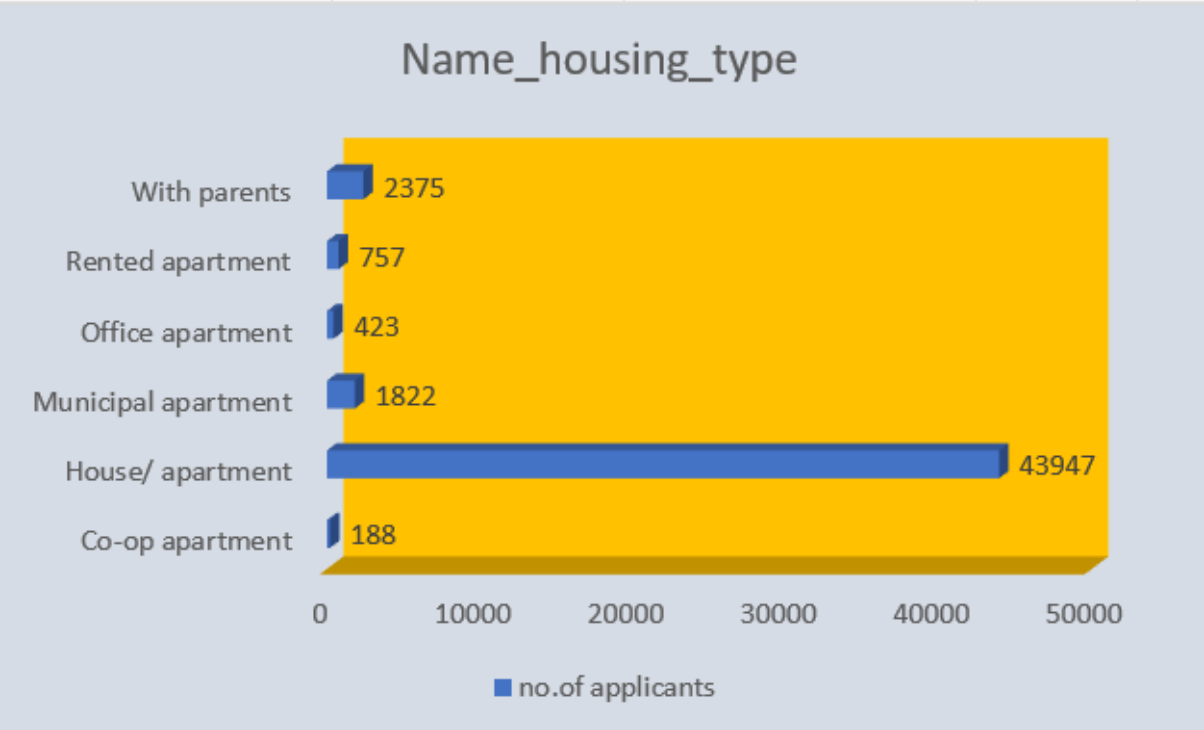


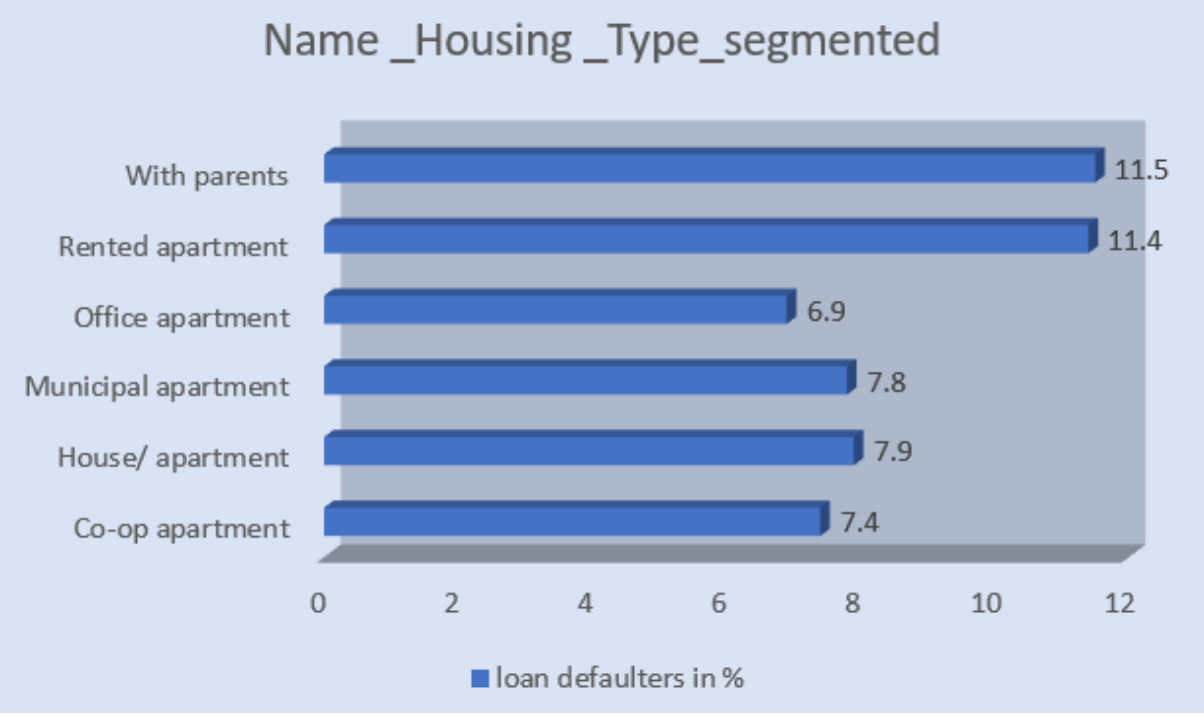


- Those whose education is for secondary/secondary special have most applicants for loans.
- Those who studied lower-secondary have the highest number of loan defaulters percentage.

Name_Housing_Type:

Name housing type	no.of applicants	loan defaulters in %
Co-op apartment	188	7.4
House/ apartment	43947	7.9
Municipal apartment	1822	7.8
Office apartment	423	6.9
Rented apartment	757	11.4
With parents	2375	11.5

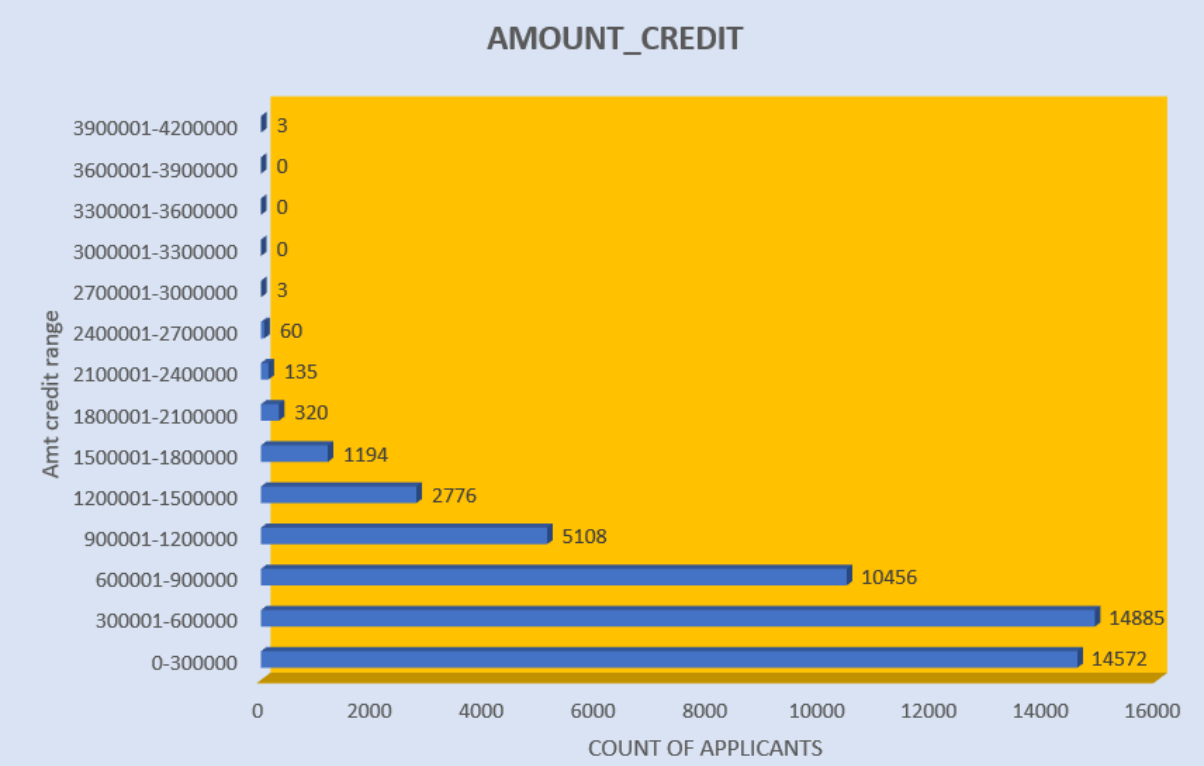


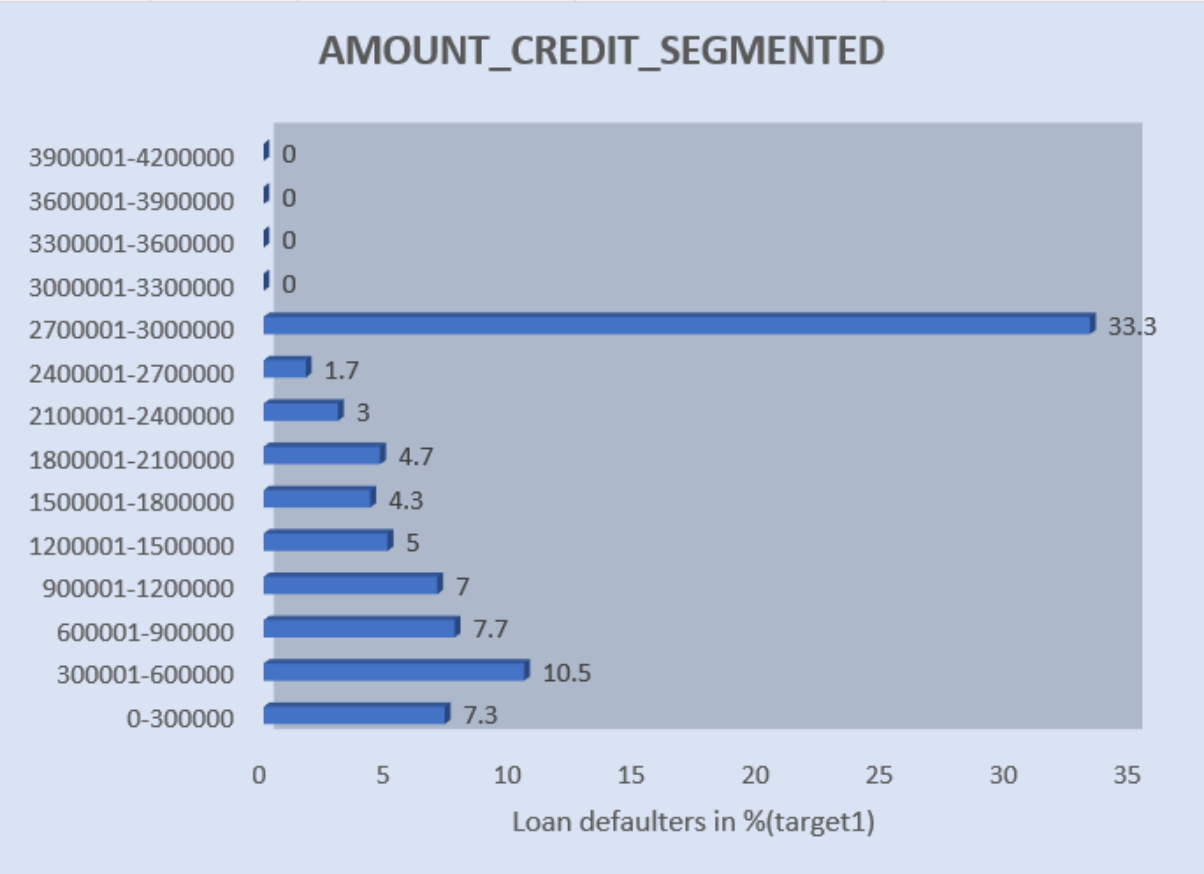


- Those who own a house/apartment are most likely to apply for a loan and coming to loan defaulters those who live with parents and in rented apartments have the highest loan defaulters percentage.

Amount_Credit:

Amt_credit_bins	count of applicants	LOAN DEFAULTERS IN %
0-300000	14572	7.3
300001-600000	14885	10.5
600001-900000	10456	7.7
900001-1200000	5108	7
1200001-1500000	2776	5
1500001-1800000	1194	4.3
1800001-2100000	320	4.7
2100001-2400000	135	3
2400001-2700000	60	1.7
2700001-3000000	3	33.3
3000001-3300000	0	0
3300001-3600000	0	0
3600001-3900000	0	0
3900001-4200000	3	0

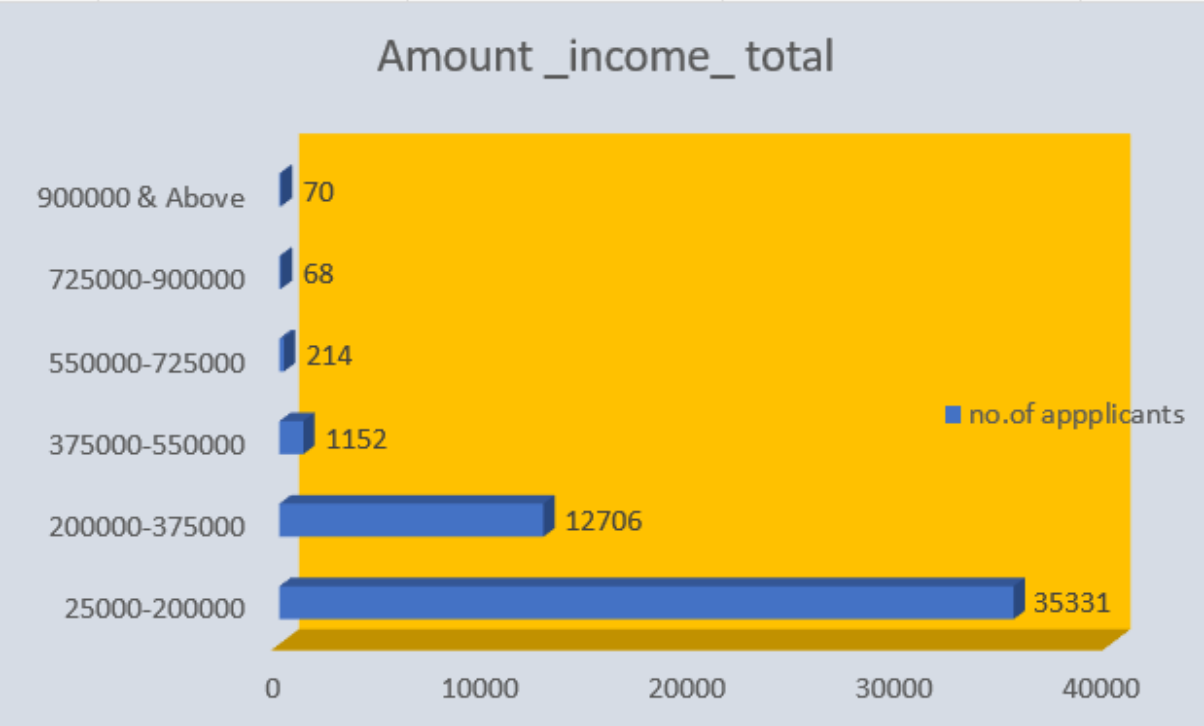


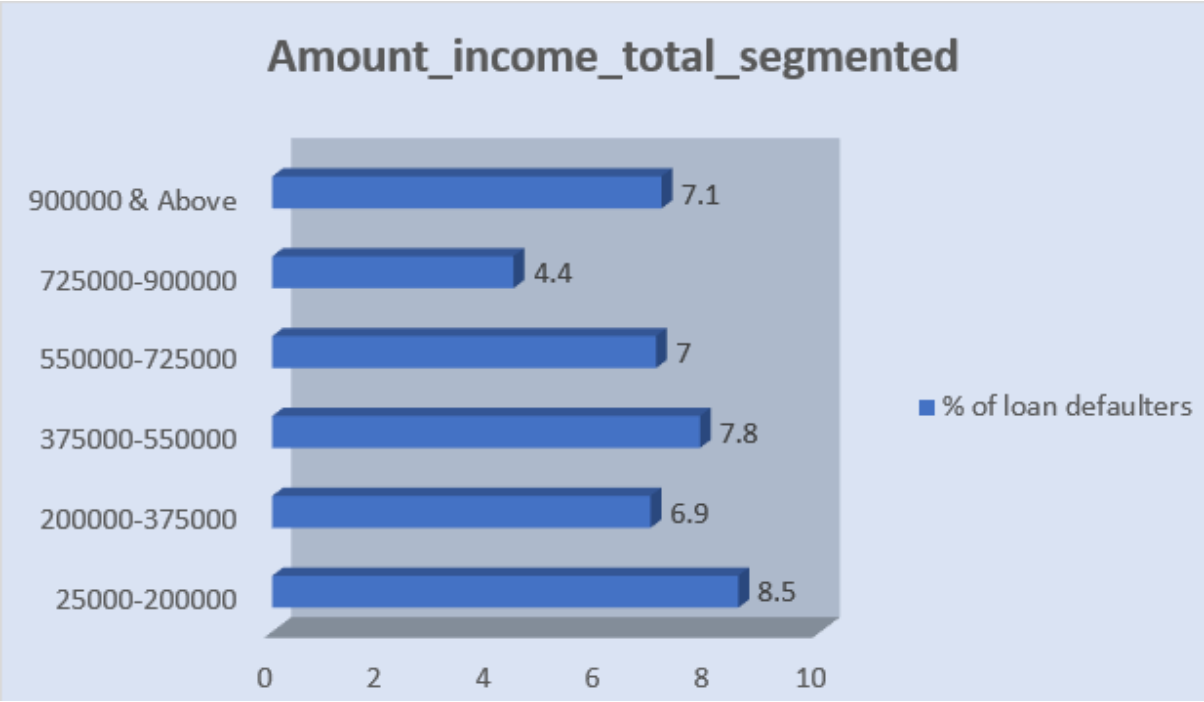


- Those who got an amount of credit between 3 lack - 6 lack applied for loans more.
- Those amount credit is between 27lack – 30 lack has the high loan defaulters.

Amount_Income:

Amt income bins	no.of appplicants	% of loan defaulters
25000-200000	35331	8.5
200000-375000	12706	6.9
375000-550000	1152	7.8
550000-725000	214	7
725000-900000	68	4.4
900000 & Above	70	7.1

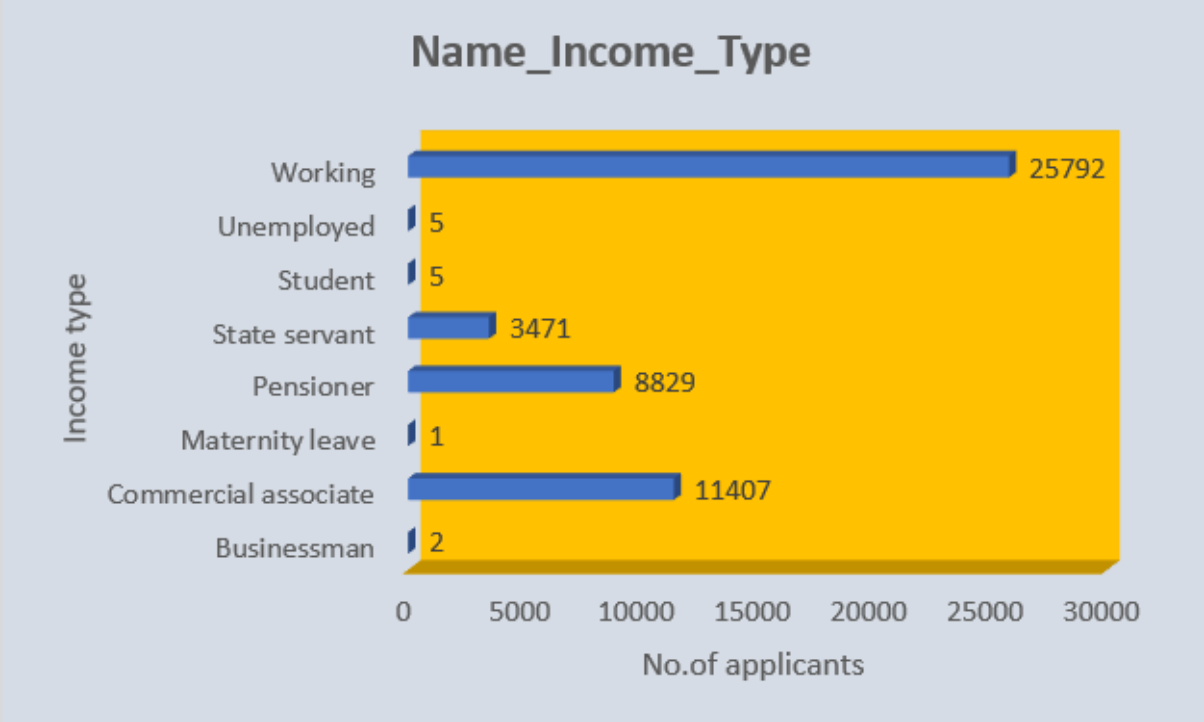


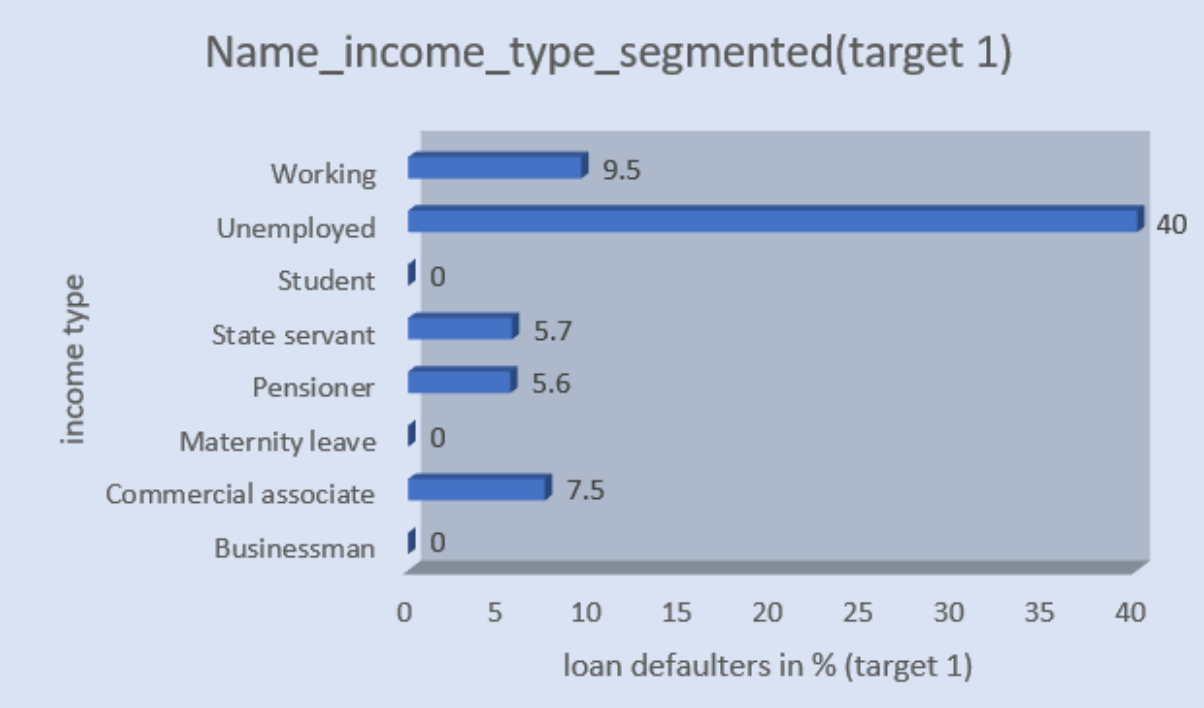


- Those who have an income total between 2500-200000 are applying for loans is the highest number. and also highest loan defaulters are present in this income range only.

Name_Income_Type:

Income type	count of applicants	% of loan defaulters
Businessman	2	0
Commercial associate	11407	7.5
Maternity leave	1	0
Pensioner	8829	5.6
State servant	3471	5.7
Student	5	0
Unemployed	5	40
Working	25792	9.5

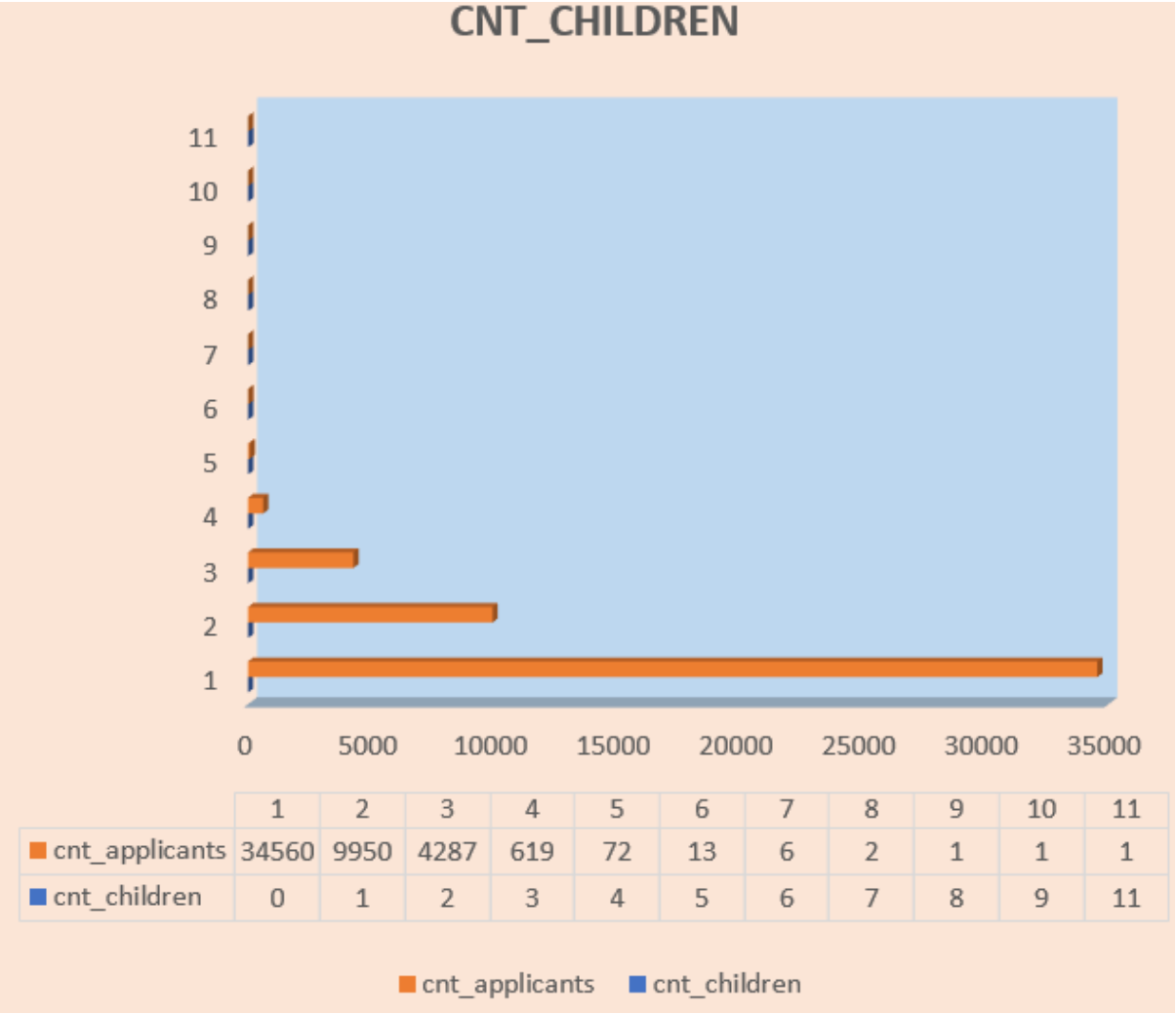




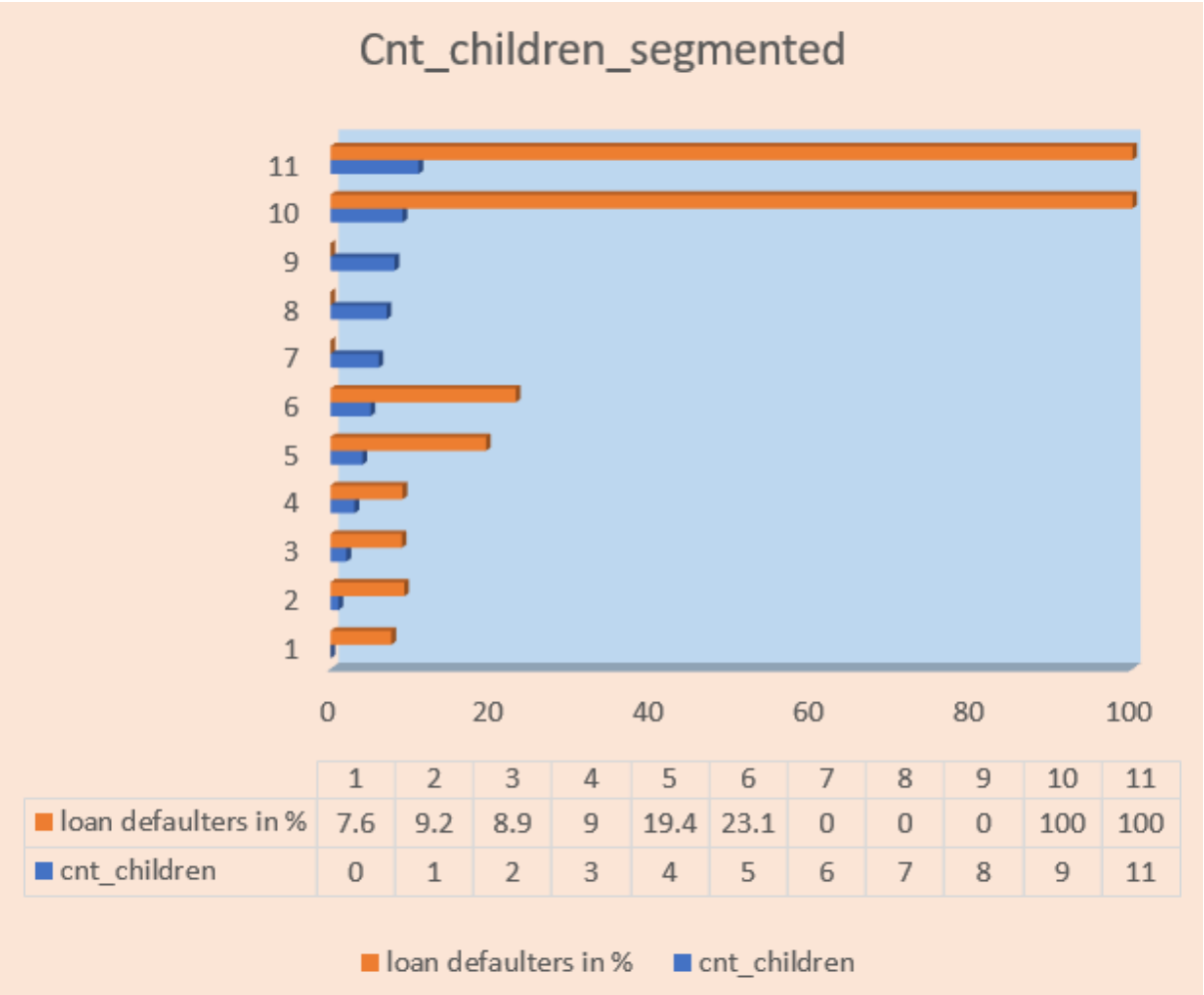
- The highest number of loan applicants belongs to those who are working and the high loan defaulters rate is for unemployed income type.

Cnt_Children:

cnt_children	cnt_applicants	loan defaulters in %
0	34560	7.6
1	9950	9.2
2	4287	8.9
3	619	9
4	72	19.4
5	13	23.1
6	6	0
7	2	0
8	1	0
9	1	100
11	1	100



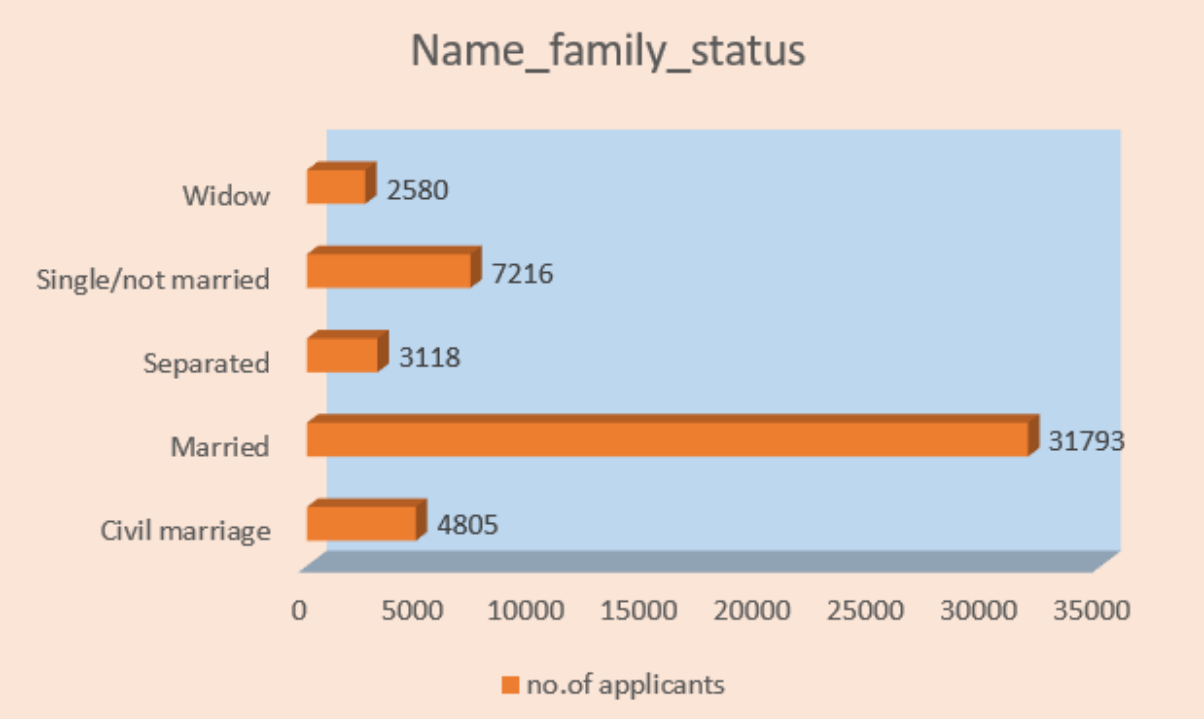
	1	2	3	4	5	6	7	8	9	10	11
cnt_applicants	34560	9950	4287	619	72	13	6	2	1	1	1
cnt_children	0	1	2	3	4	5	6	7	8	9	11

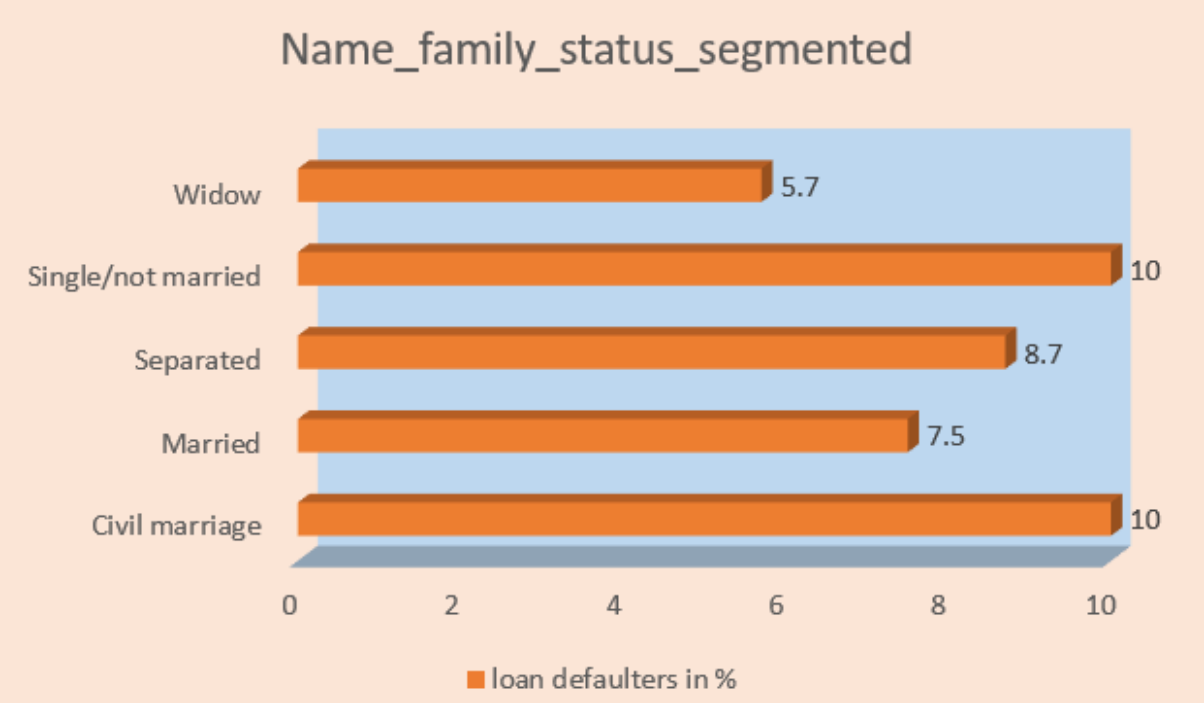


- CNT_Children is inversely proportional to the number of applicants
- A family who has 10 and 11 children is struggling to pay the loan on time they don't pay the loan amount regularly.

Name_Family_Status:

Family status	no.of applicants	loan defaulters in %
Civil marriage	4805	10
Married	31793	7.5
Separated	3118	8.7
Single/not married	7216	10
Widow	2580	5.7

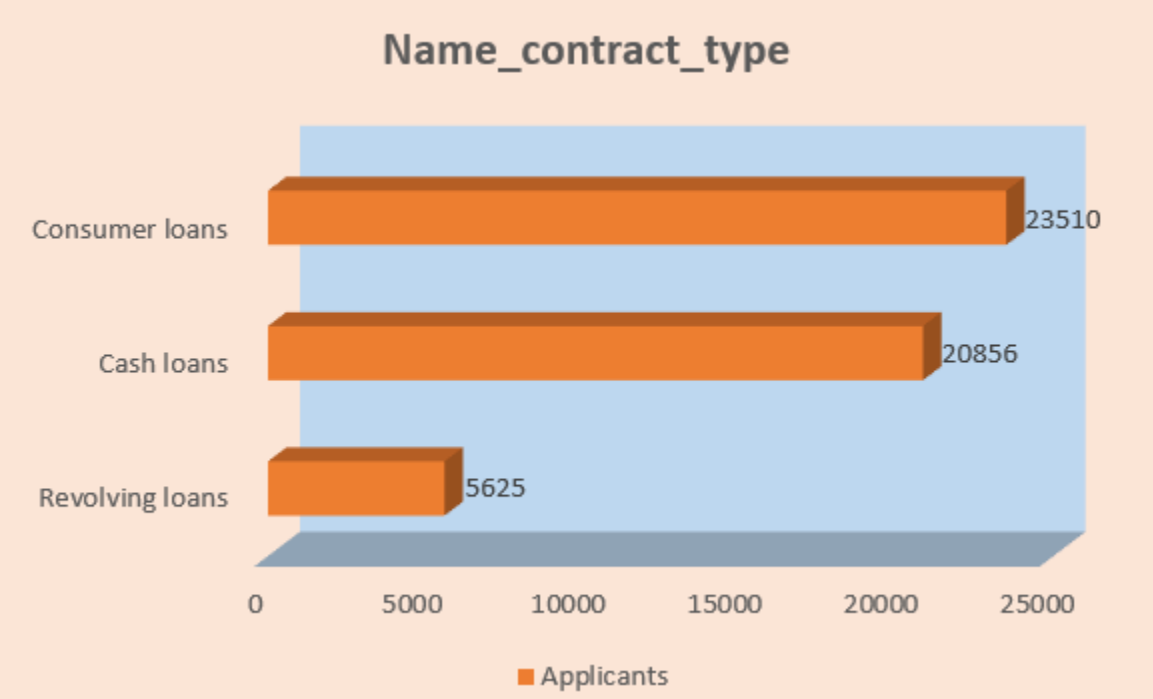




- Married people applying for loans is high compared to other relationship statuses.
- Those who are single/not married and civil marriage people have the highest loan defaulters rate.

PREVIOUS_ APPILCATION (UNIVARIENT ANALYSIS)

Contract_type	Applicants	Approved	Canceled	Refused	Unused offer
Revolving loans	5625	50.4	23.8	25.8	0
Cash loans	20856	42.7	34.5	22.7	0.1
Consumer loans	23510	85.7	0.2	10.5	3.6



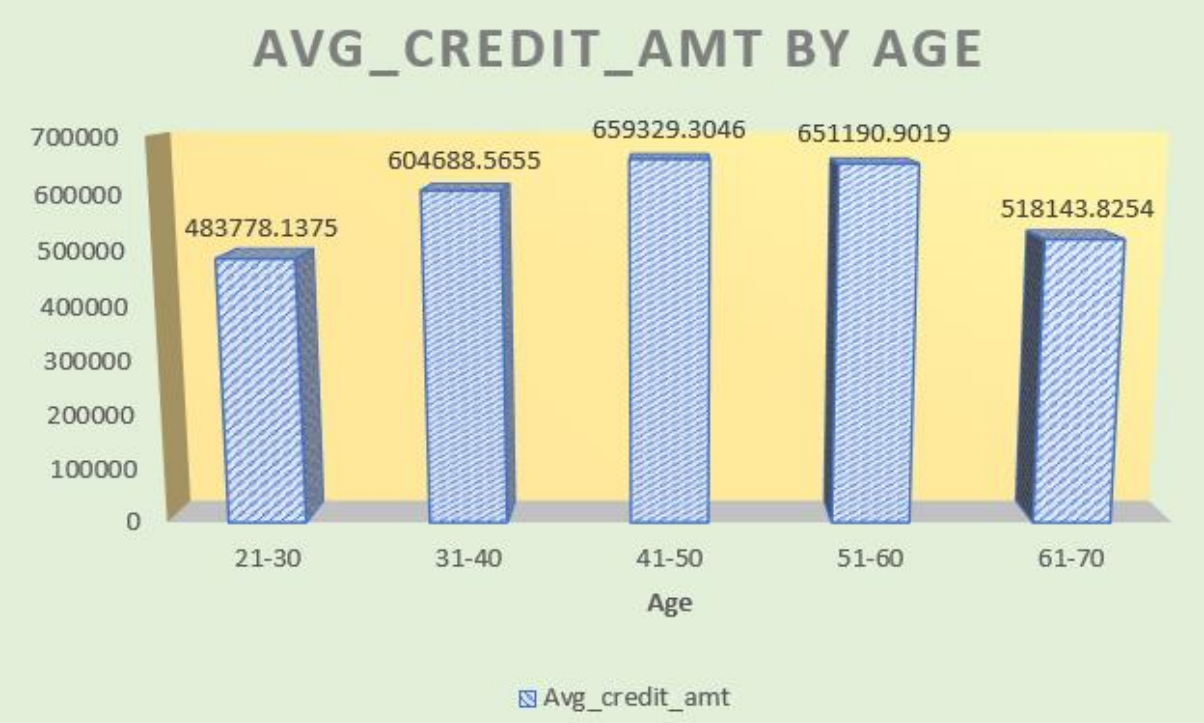
- For the previous application dataset consumer loans have the highest number of loans followed by cash loans and revolving loans.

Bivariate analysis :

When two variables are analyzed to find their correlations, this is referred to as bivariate analysis.

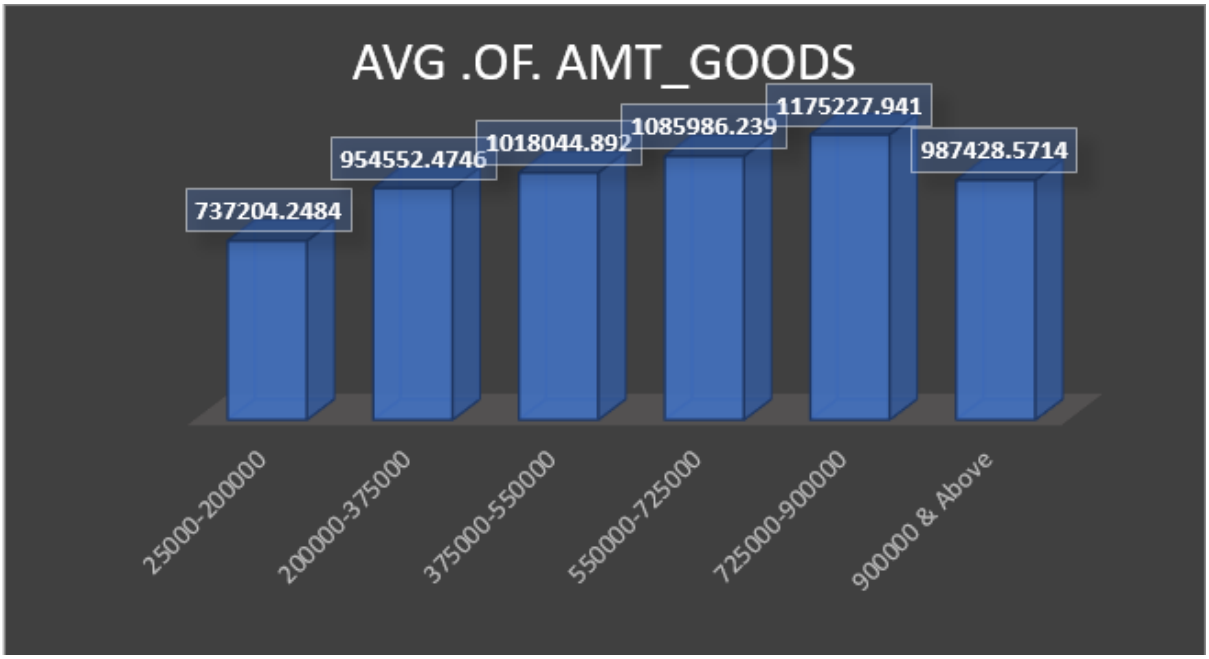
Avg_Credit_Amt_By Age:

Age	Avg_credit_amt
21-30	483778.1375
31-40	604688.5655
41-50	659329.3046
51-60	651190.9019
61-70	518143.8254



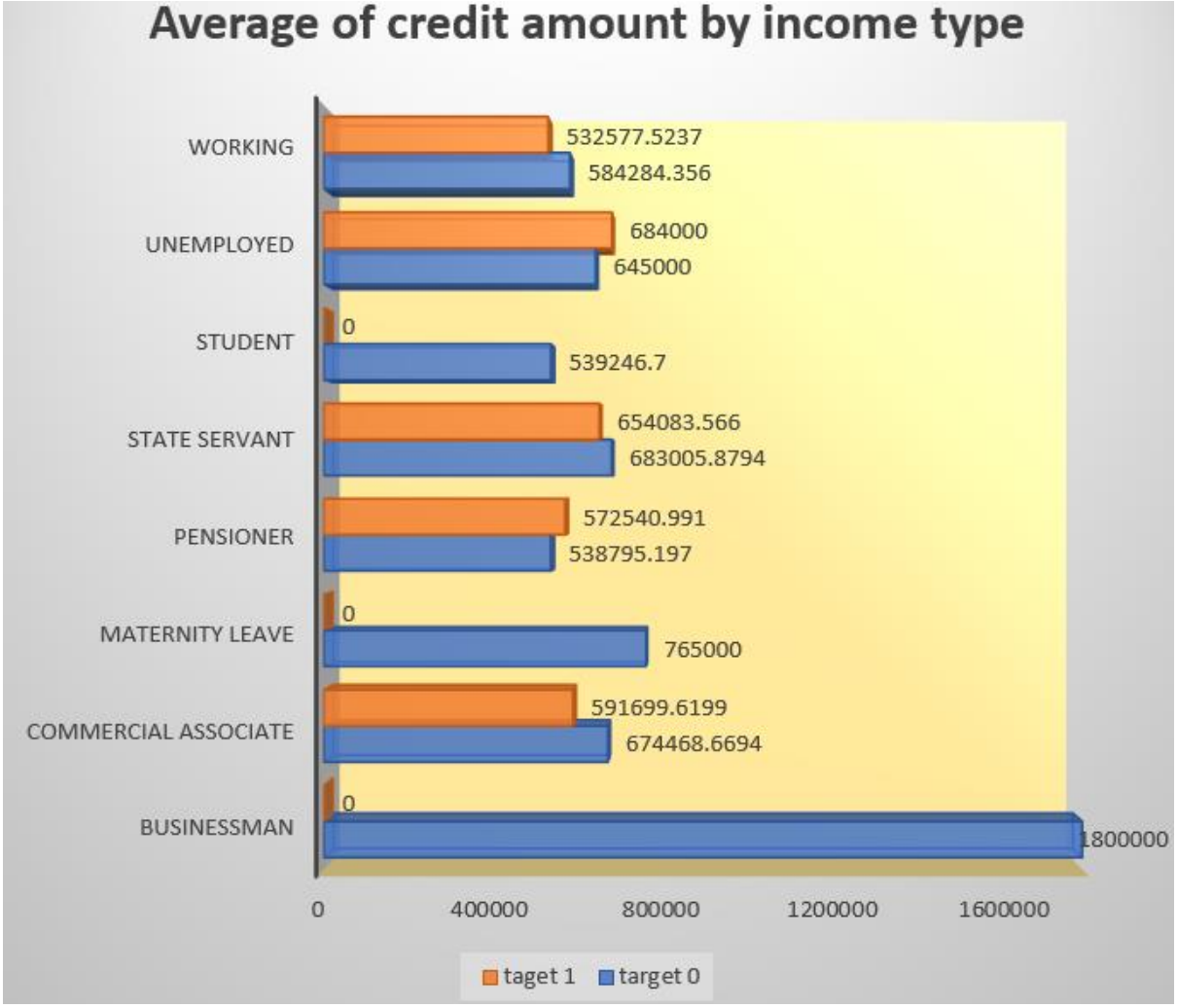
- The above graph shows the relationship between age and the amount credited.
- The avg_credit_amt by age is highest for the age group between 41-50 years followed by 51-60 years.

Income bins	Avg .of. Amt_goods
25000-200000	737204.2484
200000-375000	954552.4746
375000-550000	1018044.892
550000-725000	1085986.239
725000-900000	1175227.941
900000 & Above	987428.5714



- The above graph shows the relationship between amount_income and amount _goods_price.
- The average avg_amt_goods_price is for income bins between 725000 - 900000.

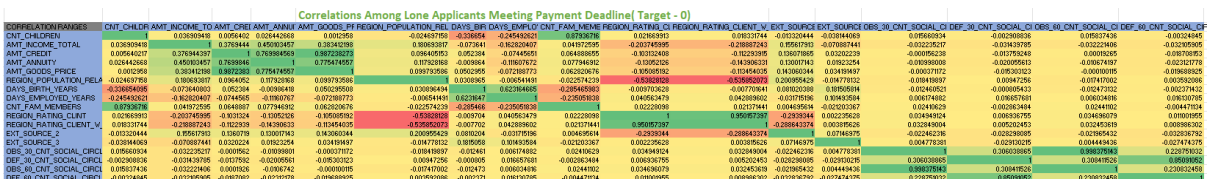
Income _type	Avg_credit_amt(target 0)	Ave_credit _amt(target 1)
Businessman	1800000	#DIV/0!
Commercial associate	674468.6694	591699.6199
Maternity leave	765000	#DIV/0!
Pensioner	538795.197	572540.991
State servant	683005.8794	654083.566
Student	539246.7	#DIV/0!
Unemployed	645000	684000
Working	584284.356	532577.5237



- The above graph shows the relationship between income_type and credit_amount(arget 0 and target 1).

E. Identify Top Correlations for Different Scenarios:

- **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.



	CORRELATIONS AMONG LOAN APPLICANTS WHO FAIL TO MEET PAYMENT OBLIGATIONS(TARGET - 1)																
CORRELATION RANGES	CNT_CHILI	AMT_INC(	AMT_CRE	AMT_ANN	AMT_GOC	REGION_F	DAYS_BIR	DAYS_EMI	CNT_FAM	REGION_F	REGION_F	EXT_SOUF	EXT_SOUF	OBS_30_C	DEF_30_C	OBS_60_C	DEF_60_C
CNT_CHILDREN	1	0.01011	0.00743	0.02928	-0.001	-0.0207	-0.2501	-0.1905	0.8928	0.05607	0.05539	-0.0155	-0.0132	0.01807	-0.0134	0.01522	-0.0182
AMT_INCOME_TOTAL	0.01011	1	0.01515	0.01791	0.01317	-0.0062	-0.0084	-0.0117	0.01317	-0.0129	-0.0127	-0.0162	-0.0262	-0.0114	-0.008	-0.0113	-0.0068
AMT_CREDIT	0.00743	0.01515	1	0.74927	0.98243	0.06563	0.14509	0.01991	0.06193	-0.0432	-0.051	0.11986	0.05059	0.03267	-0.0258	0.03356	-0.0298
AMT_ANNUITY	0.02928	0.01791	0.74927	1	0.74956	0.0701	0.01059	-0.0774	0.07632	-0.0601	-0.0779	0.11491	0.02044	0.01212	-0.0367	0.01229	-0.0427
AMT_GOODS_PRICE	-0.001	0.01317	0.98243	0.74956	1	0.07529	0.14332	0.02426	0.05606	-0.05	-0.0553	0.1339	0.05245	0.03212	-0.0197	0.03327	-0.0212
REGION_POPULATION_RELATIVE	-0.0207	-0.0062	0.06563	0.0701	0.07529	1	0.01829	0.00929	-0.0181	-0.4286	-0.4303	0.15889	-0.0215	-0.0086	0.02572	-0.0069	0.02452
DAYS_BIRTH_YEARS	-0.2501	-0.0084	0.14509	0.01059	0.14332	0.01829	1	0.58824	-0.1997	-0.0471	-0.0402	0.11125	0.14037	0.01101	0.02072	0.01236	0.02558
DAYS_EMPLOYED_YEARS	-0.1905	-0.0117	0.01991	-0.0774	0.02426	0.00929	0.58824	1	-0.1833	-0.0107	-0.0056	-0.0171	0.08379	0.00537	0.03035	0.00604	0.02439
CNT_FAM_MEMBERS	0.8928	0.01317	0.06193	0.07632	0.05606	-0.0181	-0.1997	-0.1833	1	0.05775	0.0585	0.00587	-0.0205	0.04002	-0.0066	0.03744	-0.0089
REGION_RATING_CLINT	0.05607	-0.0129	-0.0432	-0.0601	-0.05	-0.4286	-0.0471	-0.0107	0.05775	1	0.95055	-0.2387	0.02212	0.02552	0.01687	0.02478	7.7E-05
REGION_RATING_CLIENT_W_CITY	0.05539	-0.0127	-0.051	-0.0779	-0.0553	-0.4303	-0.0402	-0.0056	0.0585	0.95055	1	-0.2389	0.01868	0.02063	0.01508	0.02012	0.00072
EXT_SOURCE_2	-0.0155	-0.0162	0.11986	0.11491	0.1339	0.15889	0.11125	-0.0171	0.00587	-0.2387	-0.2389	1	0.04958	0.04884	-0.0112	0.05002	-0.0081
EXT_SOURCE_3	-0.0132	-0.0262	0.05059	0.02044	0.05245	-0.0215	0.14037	0.08379	-0.0205	0.02212	0.01868	0.04958	1	-0.0242	-0.0181	-0.027	-0.0111
OBS_30_CNT_SOCIAL_CIRCLE	0.01807	-0.0114	0.03267	0.01212	0.03212	-0.0086	0.01101	0.00537	0.04002	0.02552	0.02063	0.04884	-0.0242	1	0.3647	0.99808	0.2976
DEF_30_CNT_SOCIAL_CIRCLE	-0.0134	-0.008	-0.0258	-0.0367	-0.0197	0.02572	0.02072	0.03035	-0.0066	0.01687	0.01508	-0.0112	-0.0181	0.3647	1	0.36765	0.89034
OBS_60_CNT_SOCIAL_CIRCLE	0.01522	-0.0113	0.03356	0.01229	0.03327	-0.0069	0.01236	0.00604	0.03744	0.02478	0.02012	0.05002	-0.027	0.99808	0.36765	1	0.30104
DEF_60_CNT_SOCIAL_CIRCLE	-0.0182	-0.0068	-0.0298	-0.0427	-0.0212	0.02452	0.02558	0.02439	-0.0089	7.7E-05	0.00072	-0.0081	-0.0111	0.2976	0.89034	0.30104	1

Insights:

The above tables show the correlations for different scenarios.

- Correlations among loan applicants meeting payment deadline(target 0).
- Correlations among loan applicants who fail to meet payment obligations (target 1). For this, I use a conditional formatting style with a green-yellow-red color scale. Here green represents a strong correlation while red color represents a weak correlation

The top 10 correlations for **target 0** are:

- OBS_60_CNT SOCIAL_CIRCLE and OBS_30_CNT SOCIAL_CIRCLE
- AMT_GOODS_PRICE and AMT_CREDIT
- REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT
- CNT_CHILDREN and CNT_FAMILY_MEMBERS
- DEF_60_CNT SOCIAL_CIRCLE and DEF_30_CNT SOCIAL_CIRCLE
- AMT_ANNUITY and AMT_GOODS_PRICE
- AMT_ANNUITY and AMT_CREDIT
- DAYS_BIRTH and CNT_CHILDREN
- OBS_60_CNT SOCIAL_CIRCLE and DEF_30_CNT SOCIAL_CIRCLE
- DEF_30_CNT SOCIAL_CIRCLE and OBS_30_CNT SOCIAL_CIRCLE

The top 10 correlations for **target 1** are:

- OBS_60_CNT SOCIAL_CIRCLE and OBS_30_CNT SOCIAL_CIRCLE
- AMT_GOODS_PRICE and AMT_CREDIT
- REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT
- CNT_CHILDREN and CNT_FAMILY_MEMBERS
- DEF_60_CNT SOCIAL_CIRCLE and DEF_30_CNT SOCIAL_CIRCLE
- AMT_ANNUITY and AMT_GOODS_PRICE
- AMT_ANNUITY and AMT_CREDIT
- DAYS_BIRTH and CNT_CHILDREN

- OBS_60_CNT SOCIAL_CIRCLE and DEF_30_CNT SOCIAL_CIRCLE
- DEF_30_CNT SOCIAL_CIRCLE and OBS_30_CNT SOCIAL_CIRCLE

RESULT:

The project offered helpful details on data analysis containing information about loan applications. With Excel's features and functions, I learned how to handle missing data, spot outliers, find data imbalances, perform statistical analyses like univariate, segmented, and bivariate analysis, and also identify correlations. I was able to gain a thorough understanding of the variables causing loan default by closely examining the relationships between different variables and loan default through this analysis. To forecast business decisions and manage risks, it is imperative to acquire this knowledge. The efficiency of data-driven techniques for reducing default risks and accelerating the loan approval process was also learned by this project.

Hyperlink for Excel workbook:

P6-BANK LOAN CASE STUDY.xlsx

Loom video:

https://www.loom.com/share/163d01b801a643c6a4e659fe91d4c368?sid=b68c6d15-a57c-40e3-8c32-c42ac176abb4