Bank Loan Case Analysis

Project Description:

The main objective of this project is to use EDA to identify patterns in loan application data that indicate if a customer is likely to have difficulty repaying their instalments. By understanding these patterns, the company aims to make informed decisions about loan approval to minimize the risk of default while ensuring capable applicants are not rejected.

Approach:

Data collection and Understanding: obtain the loan application dataset and explore it to understand its structure, features and datatypes. Identify the target variable and distribution of each feature to gain insights into the data.

Exploratory Data Analysis (EDA): Explore the relationship between customer attributes and loan attributes. Identify potential correlations and patterns using scatter plots, histograms, box plots, and correlation matrices.

Statistical Analysis: Conduct statistical tests to assess the significance of relationships between variables and loan default. use measures such as correlation coefficients to quantify relationships and identify significant factors.

Interpretation and Reporting: interpret the findings from EDA and model results to iy factors influencing loan default. Provide actionable insights and recommendation to the finance company for making decisions about loan approval.

Tech Stack use:

MS Excel- A spreadsheet editor software used by professionals and businesses to enter data in a table format, perform data manipulations, computations, modelling, advanced analytics, plot graphs, etc

Hyperlink

https://drive.google.com/file/d/13cAtcKt5bDexKhC16iU4qN_CI4tphySt/view?usp=sharing
https://drive.google.com/file/d/1Vp2AV5KB1E1UjfuXQLAOkslVowmXNFdQ/view?usp=sharing

Task 1: Identify missing data and deal with it appropriately

To start with I have use countblank function to check for blank cells in all the column, after that I have dropped columns which are not necessary for the analysis and have blank values greater than 40% percentage.
Some values can be imputed using other relevant columns with help of statistical functions like **AVERAGE, MEDIAN, MODE, IF, ISBLANK, INDEX, MATCH** etc
In application data

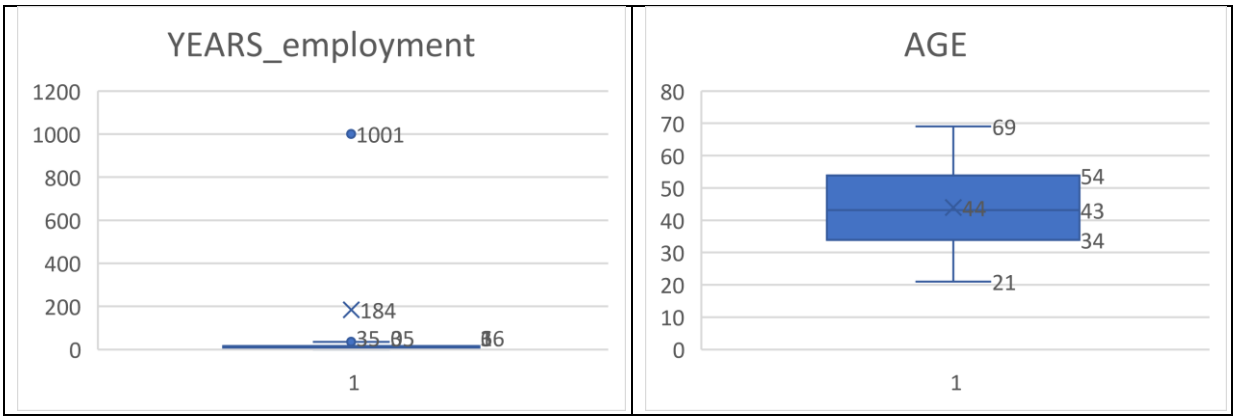
AMT_ANNUITY has one missing value and is replaced with the average value from the column
AMT_DOWN_PAYMENT is replaced with corresponding AMT_Credit values
Column CODE_GENDER has two wrong values replacing it with F for female as they are the most common in the column.
AMT_GOOD_PRICE has 38 missing values replacing them with corresponding AMT_CREDIT values as amount credited is the value of the goods purchased.
NAME_TYPE_SUITE has 192 missing values they are replaced with the most common word "Unaccompanied'
=INDEX(M6:M50004,MODE(IFERROR(MATCH(M6:M500004,M6:M50004,0),0)))
the column EMPLOYMENT_TYPE has 15k empty values and they are replaced with values corresponding with most common words for INCOME_TYPE category
=INDEX(AD3:AD50000,MODE(IFERROR(MATCH(AD3:AD50000,AD3:AD50000,0),0)))
and blank values are replaced with Laborers.

In previous_application:
AMT_ANNUITY is imputed with 0 for Unused offer, Canceled, Refused on NAME_CONTRACT_STATUS as these conditions mean the loan amount wasn't approved meaning no annuity, using the formula =**IF(ISBLANK(D6),IF(OR(F6=0,G6=0,Q6="Unused offer",Q6="Canceled",Q6="Refused"),0,""),D6)** and named the new column AMT_ANNUITY1
AMT_DOWN_PAYMENT is imputed using =**IF(ISBLANK(H6),IF(OR(Q6="Unused offer",Q6="Canceled",Q6="Refused",C6="Cash loans",C6="Revolving loans"),0,""),H6)**
AMT_GOODS_PRICE using =**IF(ISBLANK(J6),IF(OR(F6=0,G6=0),0,""),J6) where F6 and G6 are AMT_APPLICATION and AMT_CREDIT**
NAME_TYPE_SUITE is imputed with "Unaccompanied" for null values
CNT_PAYMENT is imputed with 0 as the corresponding NAME_CONTRACT_STATUS were either Unused offer, Canceled, Refused. =**IF(ISBLANK(AD6),IF(OR(Q6="Unused offer",Q6="Canceled",Q6="Refused"),0,""),AD6)**

Task 2 : Identify outliers in dataset

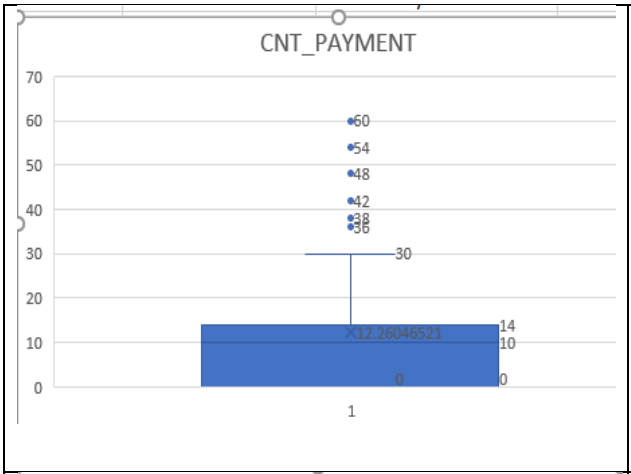
Using QUARTILE and IQR and conditional formatting the outliers in the datast are as follows

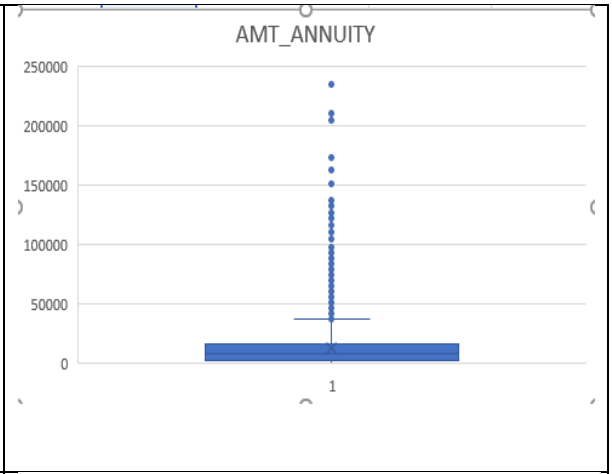
IQR=QUARTILE(Range,Q3)- QUARTILE(Range,Q1)
Higher Bound= QUARTILE(Range,Q3)+IQR*1.5
Lower Bound= QUARTILE(Range,Q1)-IQR*1.5
In application_data
AGE has no outliers, DAYS_EMPLOYED converted to YEARS_EMPLOYED has outlier value to be 1001 months which is not possible.
AMT_INCOME_TOTAL has many outlier that are very higher than the upper bound indicating few borrowers have high income
AMT_GOODS_PRICE,AMT_ANNUITY, AMT_CREDIT also have high no of outliers.

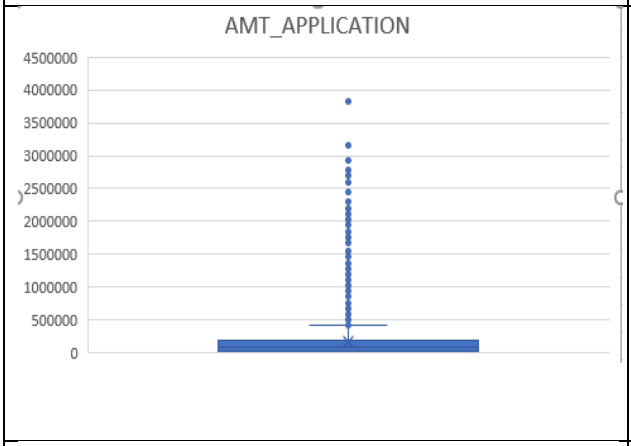


Previous application:

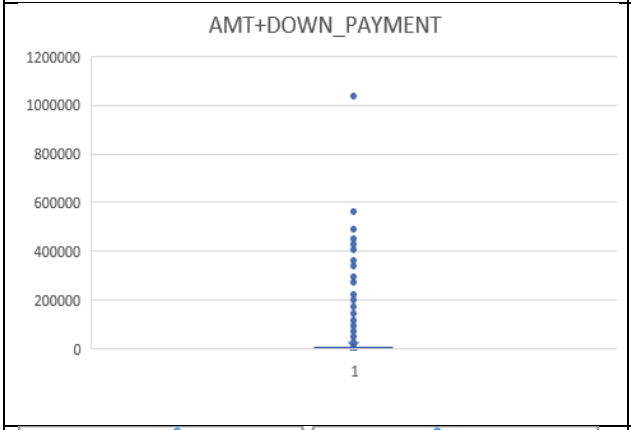
AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE have huge no of outliers wheras CNT_PAYMENT and DAYS_DECISION_MONTHS has very minimal outliers.


CNT_PAYMENT
70
60
50
40
30
20
10
0
60
54
48
42
38
36
30
12.26046521
14
10
0
0
1

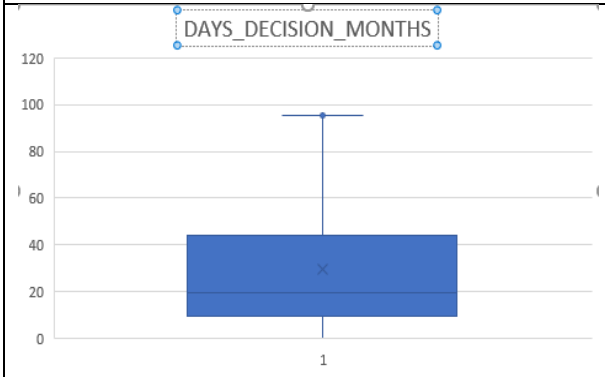

AMT_ANNUITY
250000
200000
150000
100000
50000
0
1


AMT_APPLICATION
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0


AMT_CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1


AMT+DOWN_PAYMENT
1200000
1000000
800000
600000
400000
200000
0
1

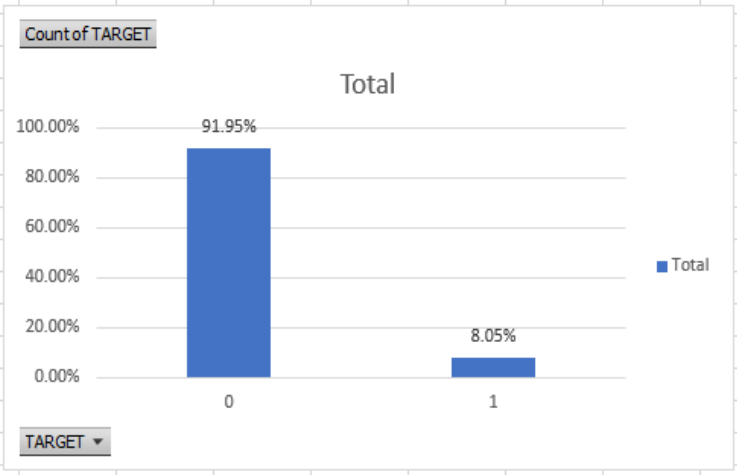

AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1


DAYS_DECISION_MONTHS
120
100
80
60
40
20
0
1

Task 3: Analyze data imbalance

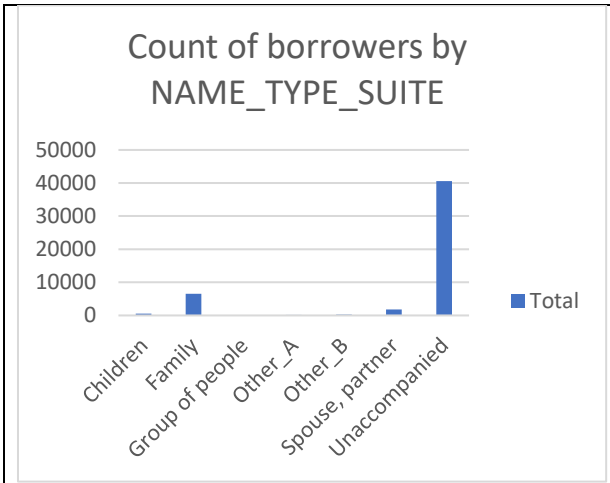
We can see the data is highly imbalanced as the percentage of defaulters compared to non - defaulters is very less, only 8% of total population of the dataset are defaulters. The remaining population had no difficulty paying off their loans regularly.
Hence the further analysis can be biased towards non-defaulters.

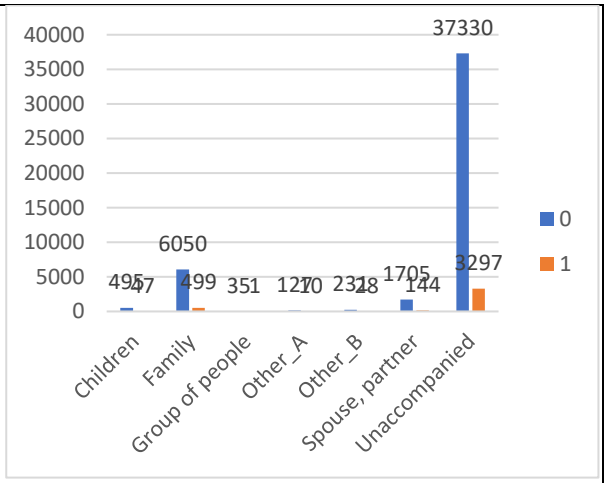
Row Labels	Count of TARGET
0	91.95%
1	8.05%
Grand Total	**100.00%**

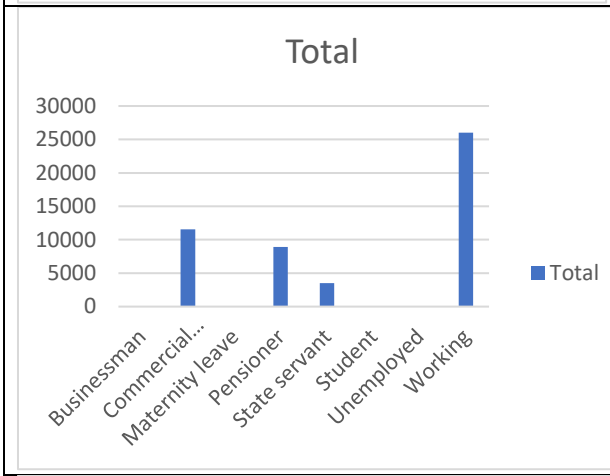


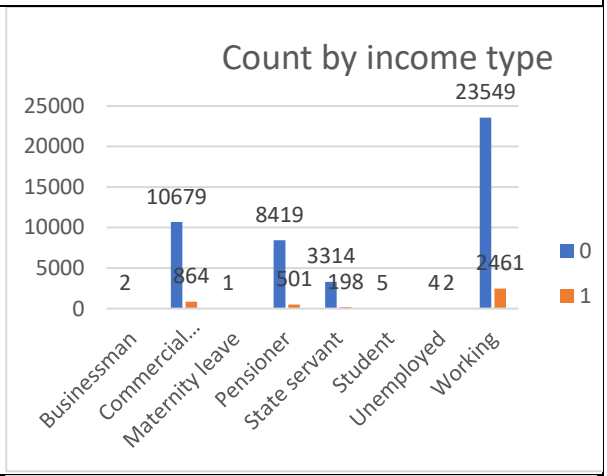
Task 4: Perform Univariate, segmented univariate and bivariate analysis

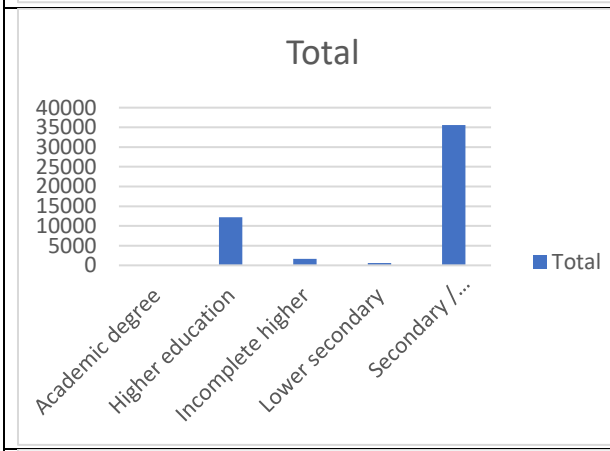
For **Univariate analysis** Most of the borrowers applied for cash loans and very few are of revolving loans.
Percentage of female borrowers are higher than male borrowers
High no of applicants were married followed by single parent
Most of the applicants had working income type while second highest applicants were from commercial associate income type
The age group of majority of borrowers were 30-40 while least were 60-70
Majority of borrowers were not accompanied during the loan application followed by few borrowers being accompanied by their family.
For **Bivariate analysis** the trend follows same as above for both defaulters and repayers which we can see from below graphs.

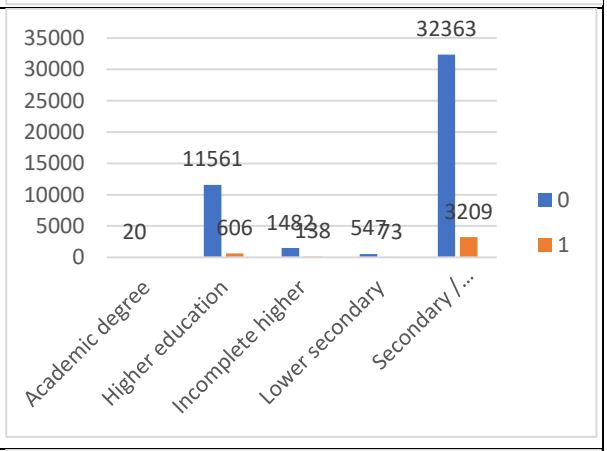

Count of borrowers by NAME_TYPE_SUITE
50000
40000
30000
20000
10000
0
Children
Family
Group of people
Other_A
Other_B
Spouse, partner
Unaccompanied
Total

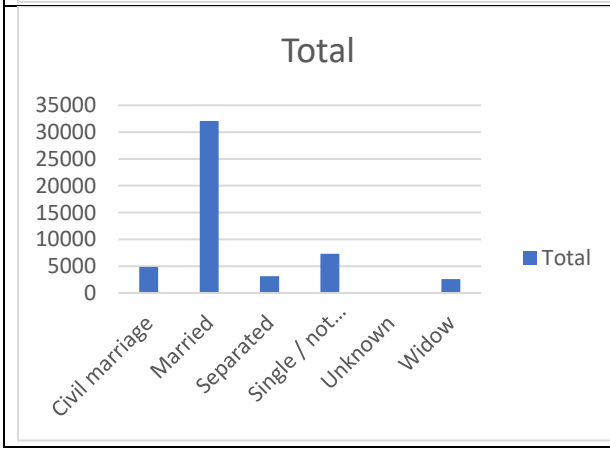

40000
35000
30000
25000
20000
15000
10000
5000
0
37330
6050
3297
1705
495
47
499
351
144
0
1
Children
Family
Group of people
Other_A
Other_B
Spouse, partner
Unaccompanied

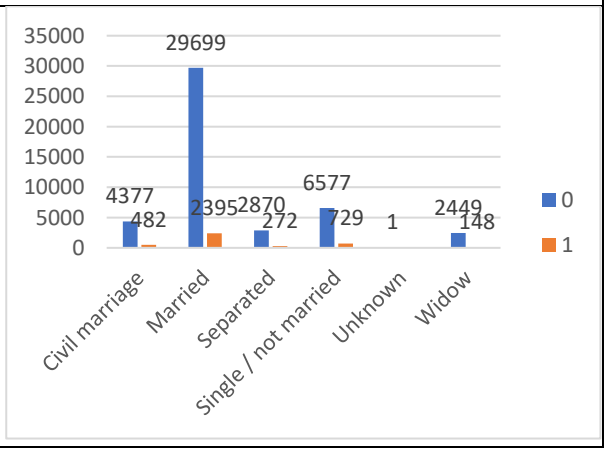

Total
30000
25000
20000
15000
10000
5000
0
Businessman
Commercial...
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
Total


Count by income type
25000
20000
15000
10000
5000
0
23549
10679
8419
3314
2461
2
864
1
501
198
5
42
0
1
Businessman
Commercial...
Maternity leave
Pensioner
State servant
Student
Unemployed
Working

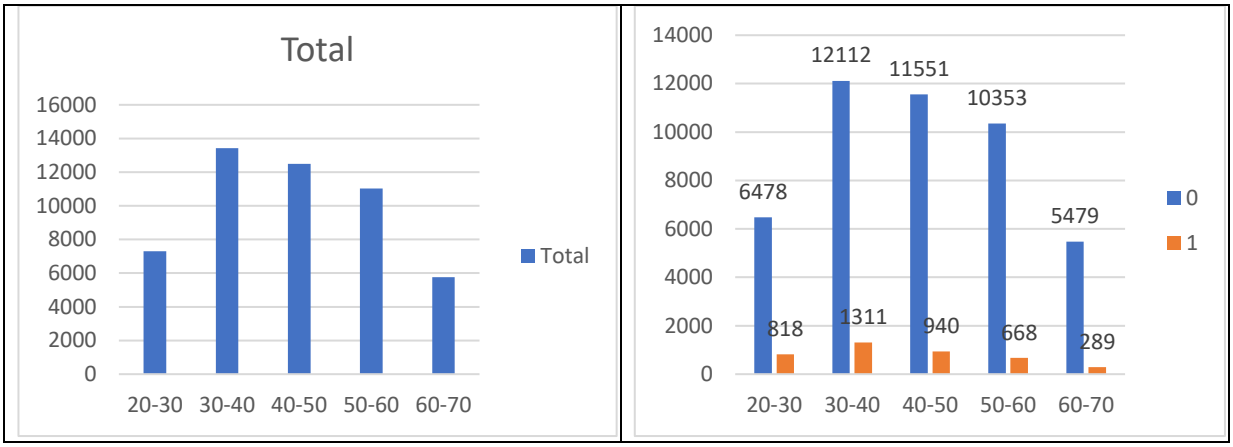

Total
40000
35000
30000
25000
20000
15000
10000
5000
0
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / ...
Total


35000
30000
25000
20000
15000
10000
5000
0
32363
11561
3209
20
606
1482
138
547
73
0
1
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / ...

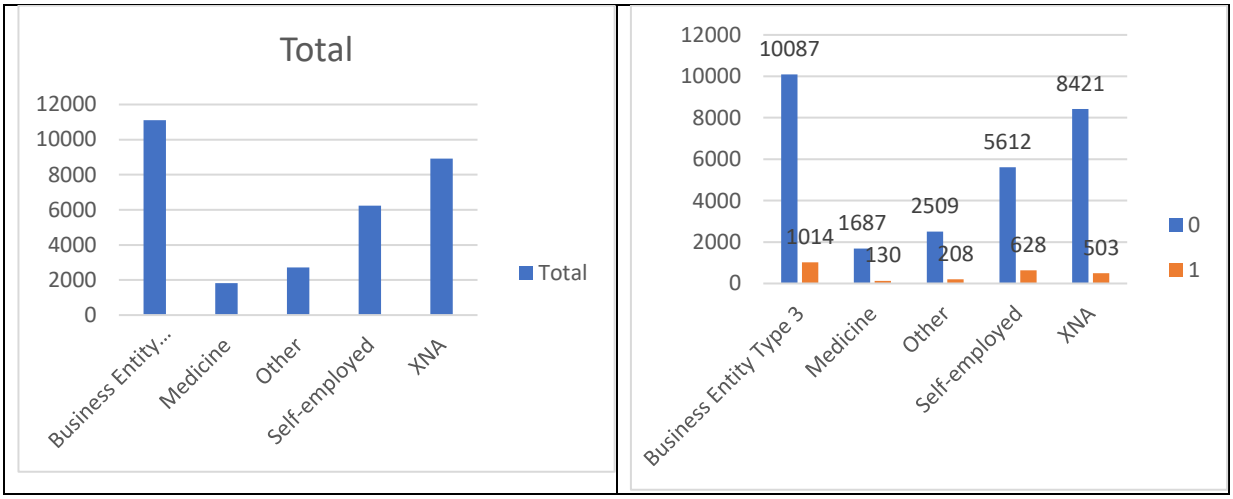

Total
35000
30000
25000
20000
15000
10000
5000
0
Civil marriage
Married
Separated
Single / not...
Unknown
Widow
Total

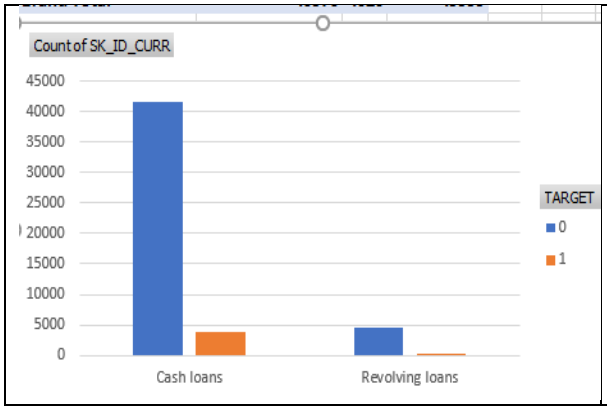

35000
30000
25000
20000
15000
10000
5000
0
29699
6577
4377
482
2395
2870
272
729
1
2449
148
0
1
Civil marriage
Married
Separated
Single / not married
Unknown
Widow

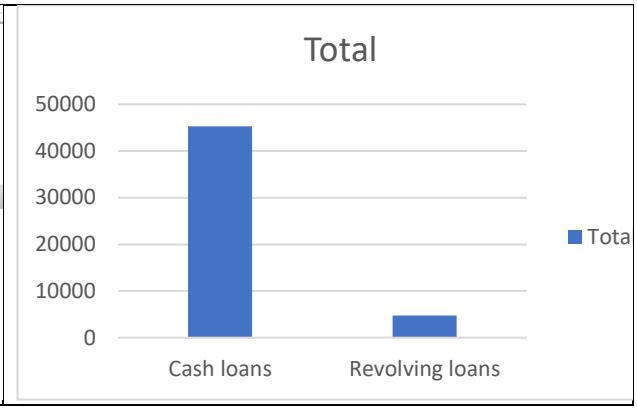
From the charts below on finding segmented univariate and bivariate we can see age 30-40 group has most borrowers for non-defaulters followed by age group 40-50 and thirdly 50-60, whereas for defaulters age group 30-40 group has most borrowers followed by age group 40-50 and thirdly 20-30



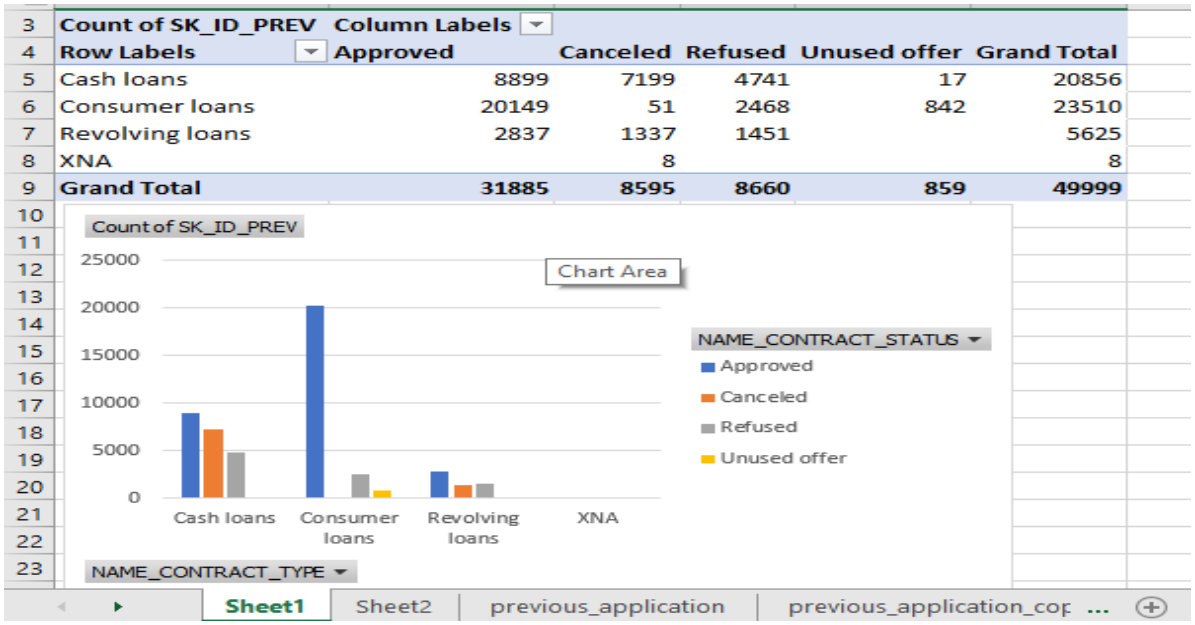
From the charts below on finding segmented univariate and bivariate we can see organization type business entity type3 has most borrowers for non-defaulters followed by XNA and thirdly self-employed, whereas for defaulters business entity type3 has most borrowers followed by self-employed and thirdly XNA







From the previous application we can see there are more borrowers from cash loans followed by consumer loans that were rejected and high no of consumer loans were approved followed by cash loans.



Count of SK_ID_PREV	Column Labels				
Row Labels	Approved	Canceled	Refused	Unused offer	Grand Total
Cash loans	8899	7199	4741	17	20856
Consumer loans	20149	51	2468	842	23510
Revolving loans	2837	1337	1451		5625
XNA		8			8
Grand Total	31885	8595	8660	859	49999

Task 5: Identify top correlations for different scenerios

Defaulters :

AMT_CREDIT	AMT_GOODS_PRICE	0.982435
AMT_CREDIT	AMT_ANNUITY	0.749665
AMT_ANNUITY	AMT_INCOME_TOTAL	0.018005
CNT_CHILDREN	CNT_FAMILY	0.892522
AMT_ANNUTY	AMT_GOODS_PRICE	0.749801
AMT_INCOME+TOTAL	AMT_CREDIT	0.015271
AGE	EMPLOYED	0.588243
AGE	AMT_INCOME_TOTAL	-0.00903
YEARS_EMPLOYED	AMT_INCOME_TOTAL	-0.01176
YEARS_EMPLOYED	AMT_CREDIT	0.018782
AGE	AMT_CREDIT	0.142506

AMT_CREDIT and AMT_GOODS_PRICE are positively correlated with correlation coefficient very close to 1

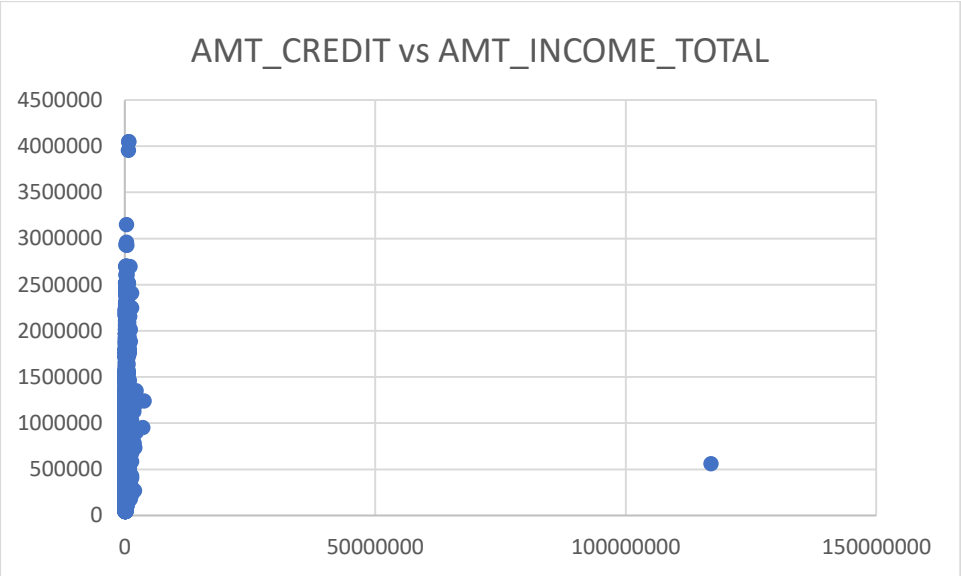
AMT_ANNUITY vs AMT_GOODS_PRICE and AMT_CREDIT vs AMT_ANNUITY both have almost same correlation values and are positively related with correlation coefficient of 0.74

CNT_CHILDREN and CNT_FAMILY_MEMBERS are positively correlated as most of the borrowers were married followed by single parent

AGE and years employed are positively correlated which is valid as more age means more years of experience

AMT_CREDIT vs AMT_INCOME_TOTAL and AMT_ANNUITY vs AMT_INCOME_TOTAL are not related at all with similar correlation coefficient of 0.015

Below scatter plot for AMT_CREDIT vs AMT_INCOME_TOTAL shows no correlation and the datapoints are all sitting at one corner.



Non-Defaulters :

AMT_CREDIT	AMT_GOODS_PRICE	0.987247
AMT_CREDIT	AMT_ANNUITY	0.770772
AMT_ANNUITY	AMT_INCOME_TOTAL	0.451136
CNT_CHILDREN	CNT_FAMILY	0.879239
AMT_ANNUTY	AMT_GOODS_PRICE	0.776313
AMT_INCOME+TOTAL	AMT_CREDIT	0.377966
AGE	EMPLOYED	0.623475
AGE	AMT_INCOME_TOTAL	-0.07377
YEARS_EMPLOYED	AMT_INCOME_TOTAL	-0.16168
YEARS_EMPLOYED	AMT_CREDIT	-0.07473
AGE	AMT_INCOME_TOTAL	0.051084

AMT_CREDIT and AMT_GOODS_PRICE are positively correlated with correlation coefficient very close to 1

AMT_ANNUITY vs AMT_GOODS_PRICE and AMT_CREDIT vs AMT_ANNUITY both have almost same correlation values and are positively related with correlation coefficient of 0.7

CNT_CHILDREN and CNT_FAMILY_MEMBERS are positively correlated as most of the borrowers were married followed by single parent

AGE and years employed are positively correlated which is valid as more age means more years of experience

AMT_CREDIT vs AMT_INCOME_TOTAL and AMT_ANNUITY vs AMT_INCOME_TOTAL are positively related with lower correlation coefficient meaning greater the income greater the credit also means borrowers with higher income are no-defaulters.

AGE vs AMT_CREDIT shows correlation coefficient of 0.1 meaning they are not related and by greater age doesn't mean the credit limit can be increased.

YEARS_EMPLOYED vs AMT_CREDIT are also not correlated

KEY TAKEAWAYS: from the Segmented bivariate analysis when features are analysed against the target variable charts we could see for both defaulters and non-defaulters the trends are same for all features this is because of the data imbalance we saw earlier, as 92% of the total dataset containers non-defaulters the result tends to be biased towards non-defaulters

Ex: the analysis for occupation type both was defaulters and non-defaulters Business entity type 3 category had the highest borrowers, does this mean borrowers of this occupation are likely to default or repay the loan?

Hence if we only correct the data imbalance the analysis can be unbiased.

Result:

Data cleaning and pre-processing: the importance of data cleaning and preprocessing in preparing the dataset for analysis, including handling missing values using descriptive statistics and advance formulae and functions, removing duplicates, converting calculated columns.

Exploratory Data Analysis: EDA techniques such as univariate, segmented univariate and bivariate analysis were performed to gain insights into the distribution of individual variables, compare variable distributions across different scenarios, and explore relationships between variables and the target variable, this helped identify patterns, trends, and correlations within the data.

Correlation Analysis: learn how to perform correlation analysis to identify relationships between different variables and understand their impact on loan defaults.

Decision Support: learn how to translate data analysis findings into actionable recommendations to support decision-making facilitating decision-making processes related to loan approval and risk management.

Overall, the analysis aims to provide a comprehensive understanding of the dynamics behind loan defaults, leveraging data-driven insights to inform decision-making in finance domain mitigating risks.