Bank Loan Case Study

Project Description:

This project aims to analyze loan application data using Exploratory Data Analysis (EDA) to identify factors influencing loan default. The objective is to understand patterns that differentiate reliable borrowers from those at risk of default. By handling missing data, identifying outliers, analyzing data imbalance, and exploring correlations between variables, the project helps improve loan approval decisions and reduce financial risk. The approach leverages Excel functions for statistical analysis and visualizations to gain insights into customer and loan attributes.

Approach:

My approach involved cleaning the data by identifying and imputing missing values, detecting outliers using the Interquartile Range (IQR), and analyzing data imbalance with `COUNTIF`. I performed univariate, segmented, and bivariate analysis using Excel's statistical functions and pivot tables. Correlation analysis was conducted with `CORREL` to identify key indicators of loan default, and visual insights were presented through charts like bar, pie, scatter plots, and heatmaps. This method provided a comprehensive understanding of the factors influencing loan defaults.

Tech-Stack Used:

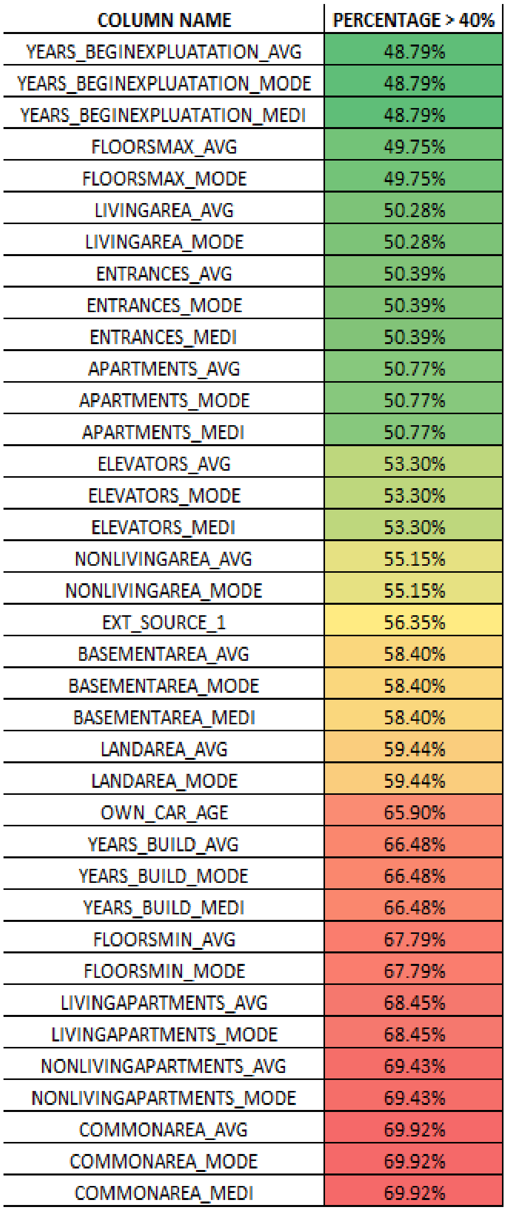
I utilized Microsoft Excel 2021 for data cleaning, analysis, and visualization. Key functions included `COUNTBLANK`, `AVERAGE`, and `MEDIAN` for handling missing data, along with `QUARTILE` and IQR calculations for outlier detection. To assess data imbalance, I employed `COUNTIF`, while `CORREL` facilitated correlation analysis. Additionally, pivot tables were used for segmented analysis, and various visualizations like bar charts, pie charts, and scatter plots were created to effectively present insights and patterns in the data.

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.

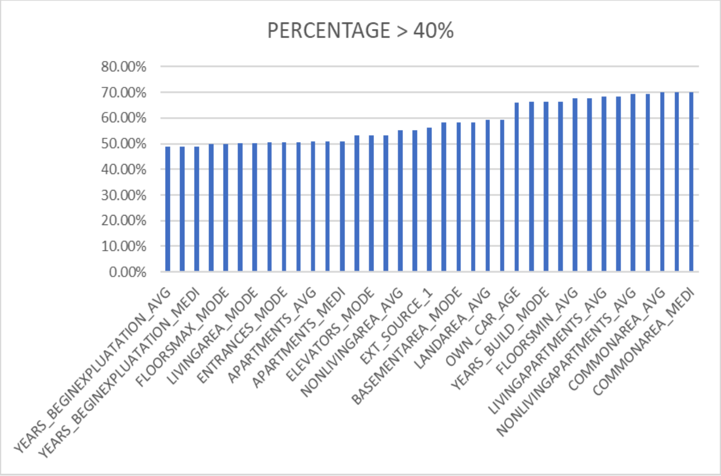
- **Task:** Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

Part 1: Null Values :-

These are the columns which have null values more than or equal to 40%.

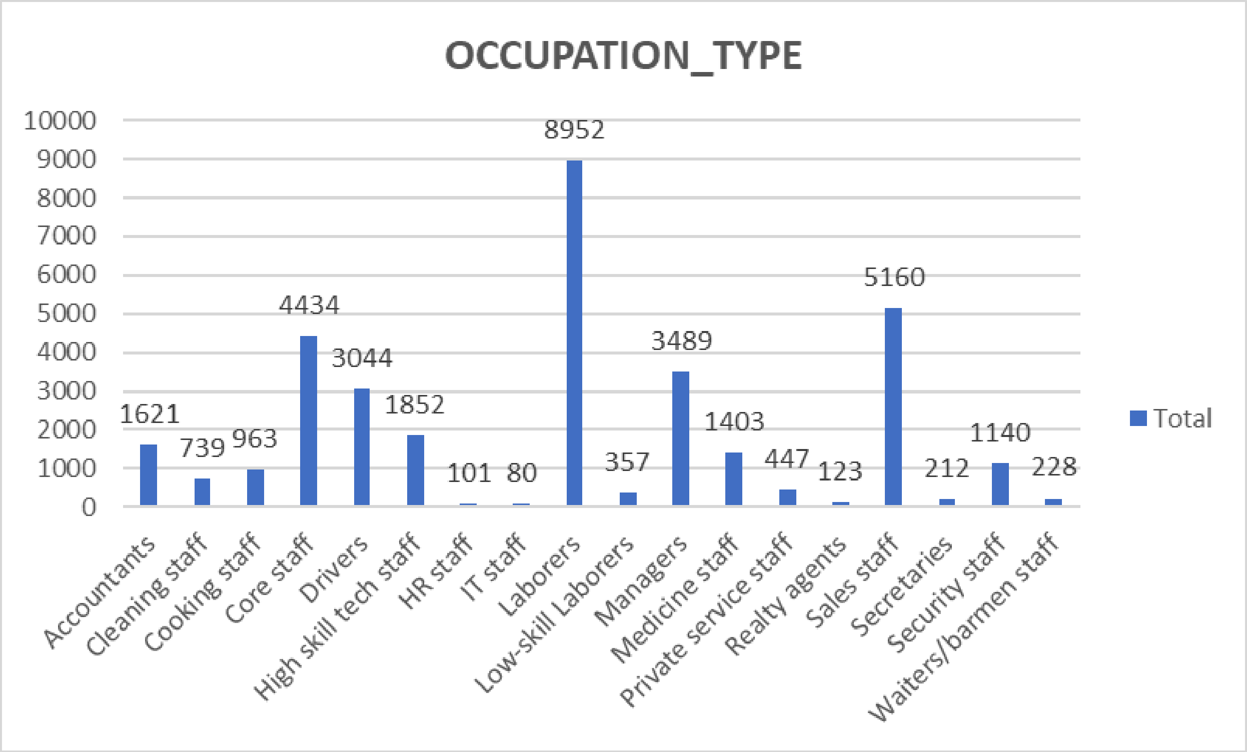


COLUMN NAME	PERCENTAGE > 40%
YEARS_BEGINEXPLUATATION_AVG	48.79%
YEARS_BEGINEXPLUATATION_MODE	48.79%
YEARS_BEGINEXPLUATATION_MEDI	48.79%
FLOORSMAX_AVG	49.75%
FLOORSMAX_MODE	49.75%
LIVINGAREA_AVG	50.28%
LIVINGAREA_MODE	50.28%
ENTRANCES_AVG	50.39%
ENTRANCES_MODE	50.39%
ENTRANCES_MEDI	50.39%
APARTMENTS_AVG	50.77%
APARTMENTS_MODE	50.77%
APARTMENTS_MEDI	50.77%
ELEVATORS_AVG	53.30%
ELEVATORS_MODE	53.30%
ELEVATORS_MEDI	53.30%
NONLIVINGAREA_AVG	55.15%
NONLIVINGAREA_MODE	55.15%
EXT_SOURCE_1	56.35%
BASEMENTAREA_AVG	58.40%
BASEMENTAREA_MODE	58.40%
BASEMENTAREA_MEDI	58.40%
LANDAREA_AVG	59.44%
LANDAREA_MODE	59.44%
OWN_CAR_AGE	65.90%
YEARS_BUILD_AVG	66.48%
YEARS_BUILD_MODE	66.48%
YEARS_BUILD_MEDI	66.48%
FLOORSMIN_AVG	67.79%
FLOORSMIN_MODE	67.79%
LIVINGAPARTMENTS_AVG	68.45%
LIVINGAPARTMENTS_MODE	68.45%
NONLIVINGAPARTMENTS_AVG	69.43%
NONLIVINGAPARTMENTS_MODE	69.43%
COMMONAREA_AVG	69.92%
COMMONAREA_MODE	69.92%
COMMONAREA_MEDI	69.92%



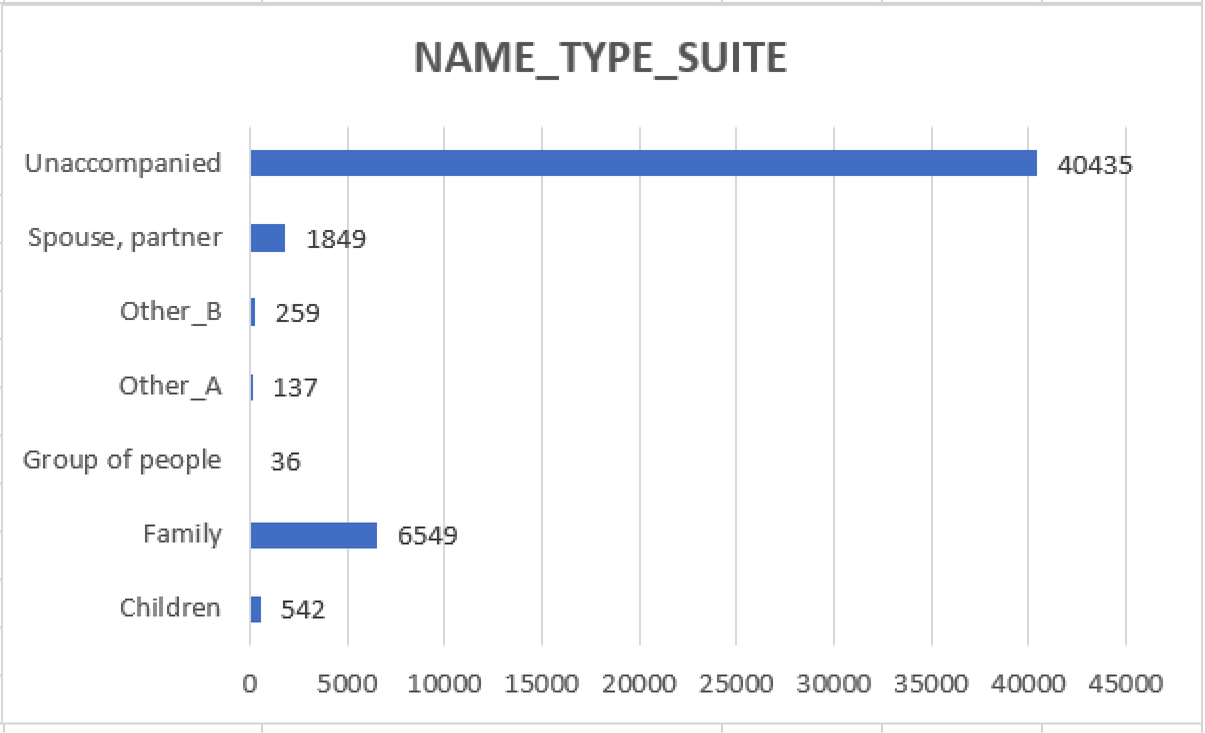
Part 2: Mode Imputation :-

1. OCCUPATION_TYPE



Most occurring variable is “Laborers" with a count of 8952.

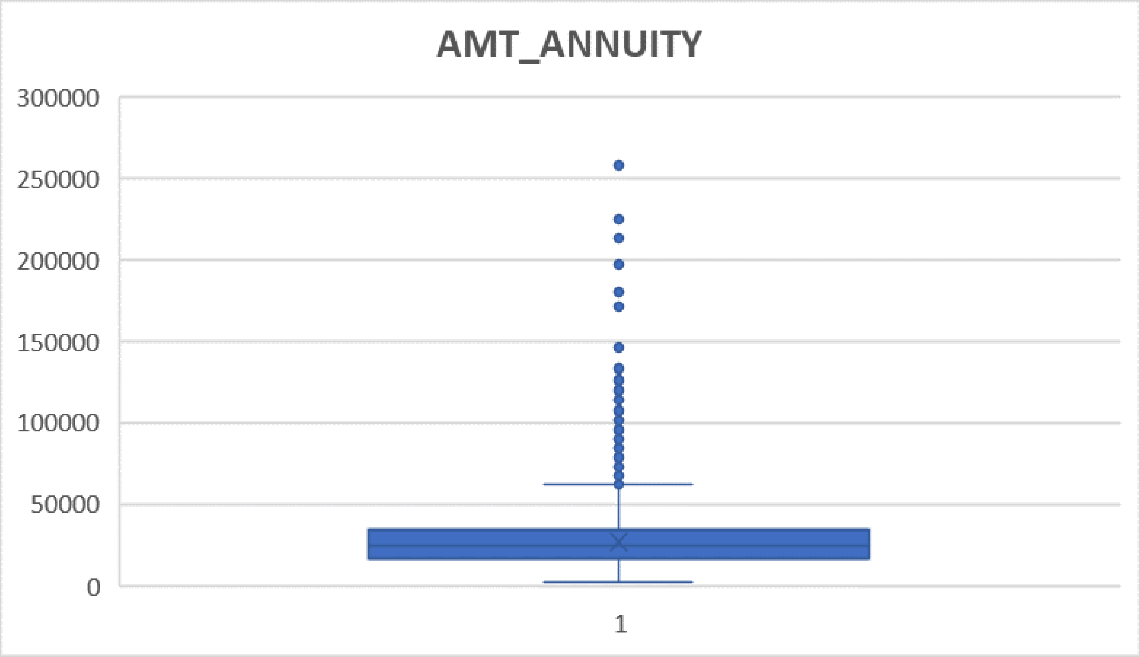
2. NAME_TYPE_SUIT



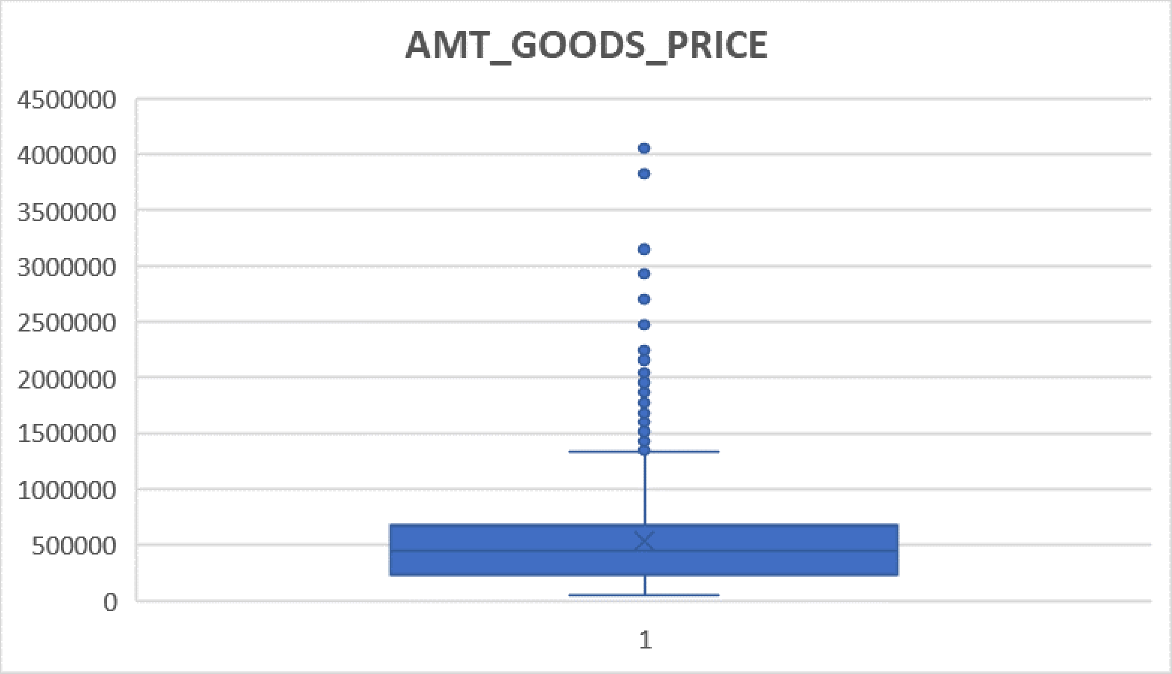
Most occurring variable is “Unaccompanied” with a count of 40435.

Part 3 : Median Imputation :-

1. AMT_ANNUITY



2. AMT_GOODS_PRICE



3. AMT_REQ_CREDIT_BUREAU_HOUR

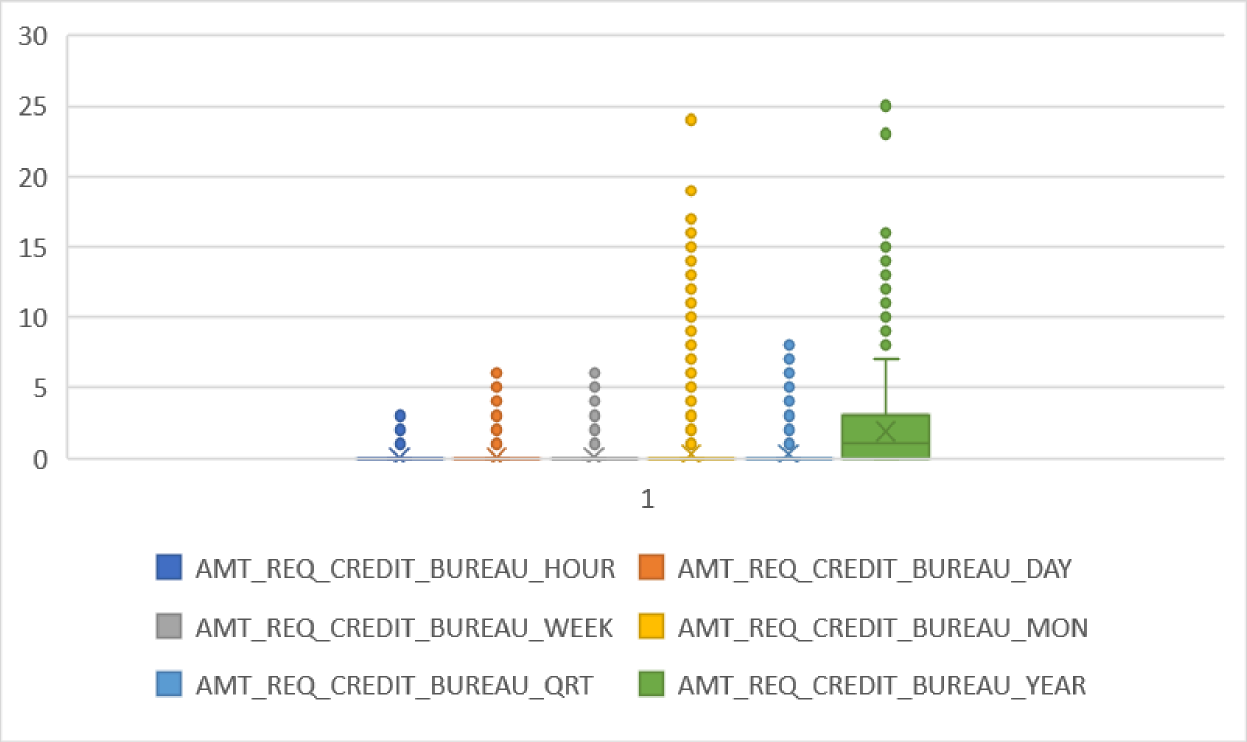
4. AMT_REQ_CREDIT_BUREAU_DAY

5. AMT_REQ_CREDIT_BUREAU_WEEK

6. AMT_REQ_CREDIT_BUREAU_MON

7. AMT_REQ_CREDIT_BUREAU_QRT

8. AMT_REQ_CREDIT_BUREAU_YEAR

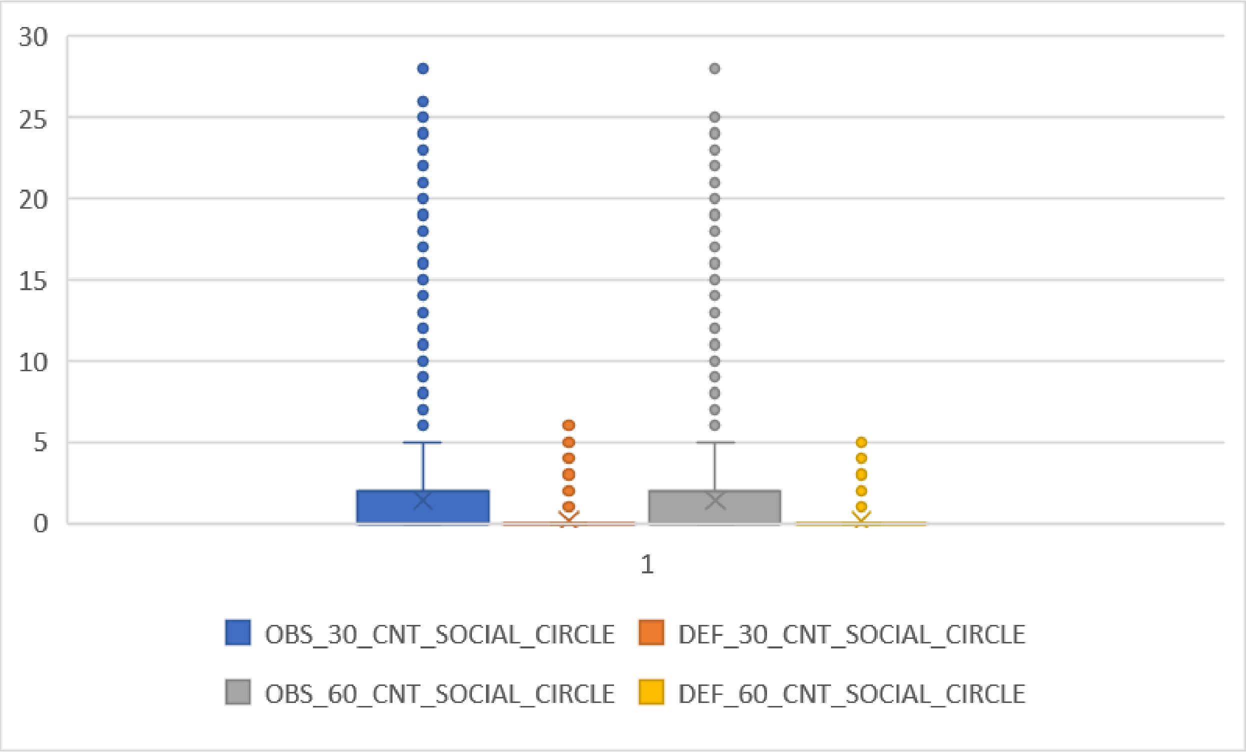


1. DEF_30_CNT_SOCIAL_CIRCLE

2. OBS_30_CNT_SOCIAL_CIRCLE

3. DEF_60_CNT_SOCIAL_CIRCLE

4. OBS_60_CNT_SOCIAL_CIRCLE



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.

Part 1:

Formulas:-

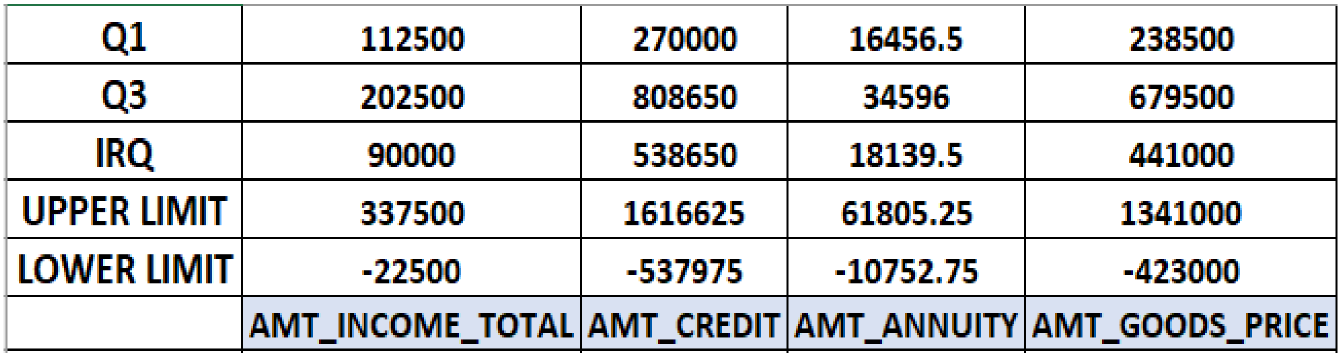
Quartile 1 : =QUARTILE.EXC(“range”,1)

Quartile 3 : =QUARTILE.EXC(“range”,3)

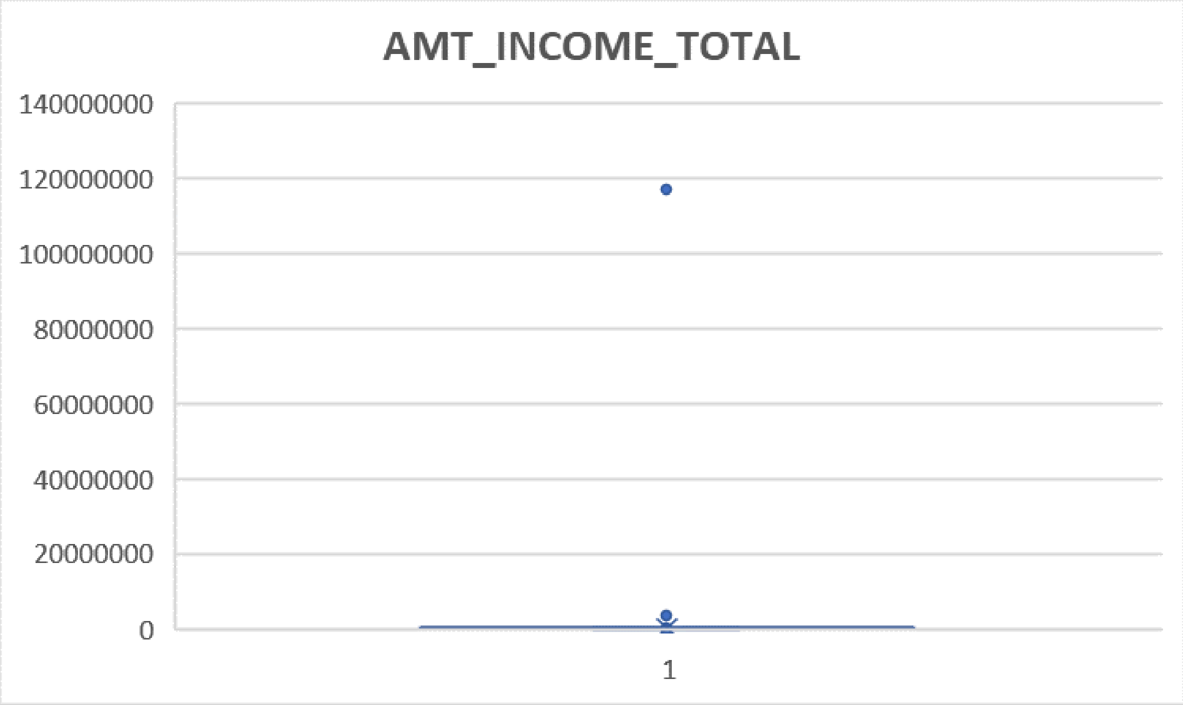
IQR = Quartile 3 - Quartile 1

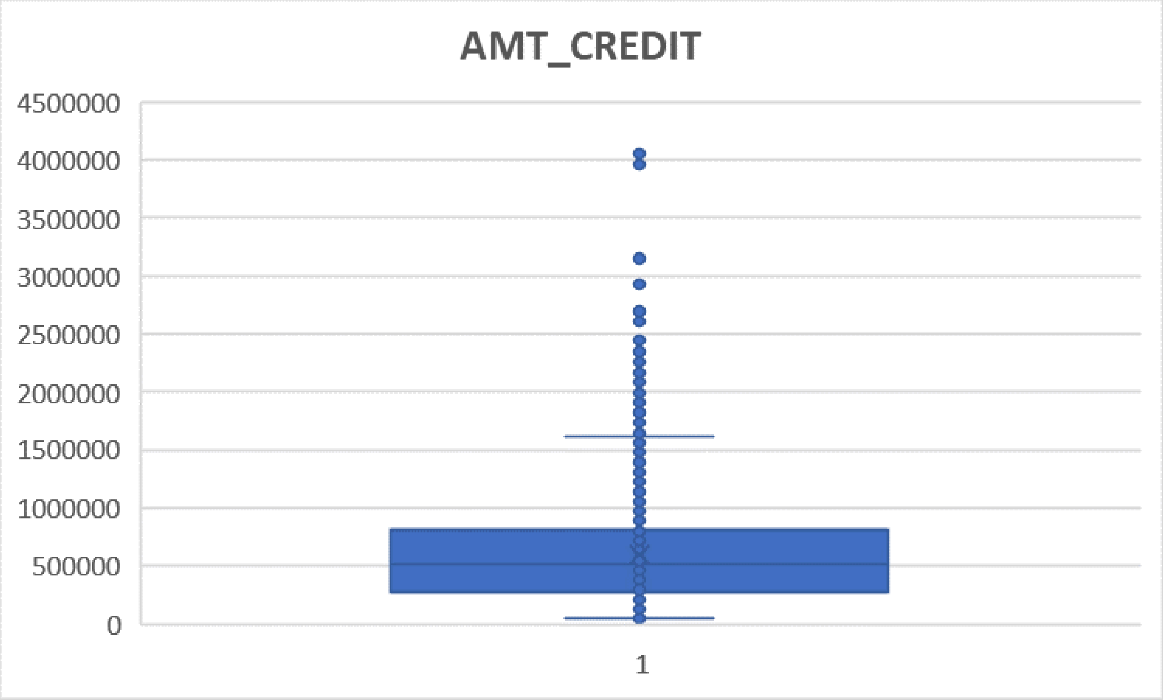
Upper Limit = Quartile 3 + 1.5*IQR

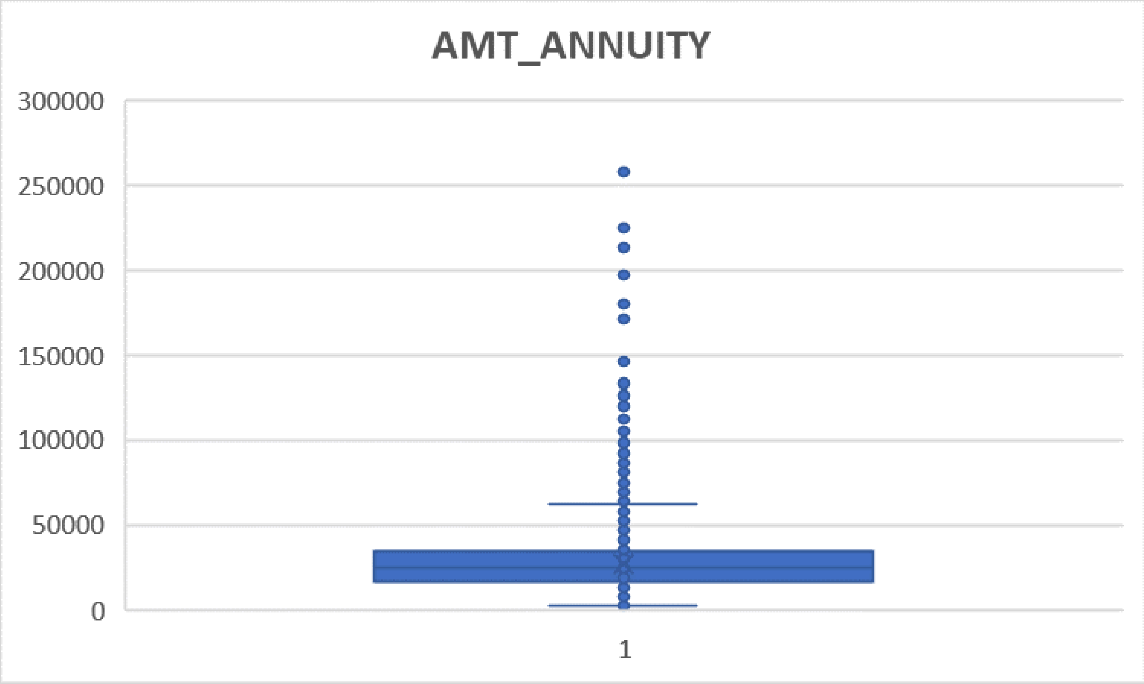
Lower Limit = Quartile 1 – 1.5IQR

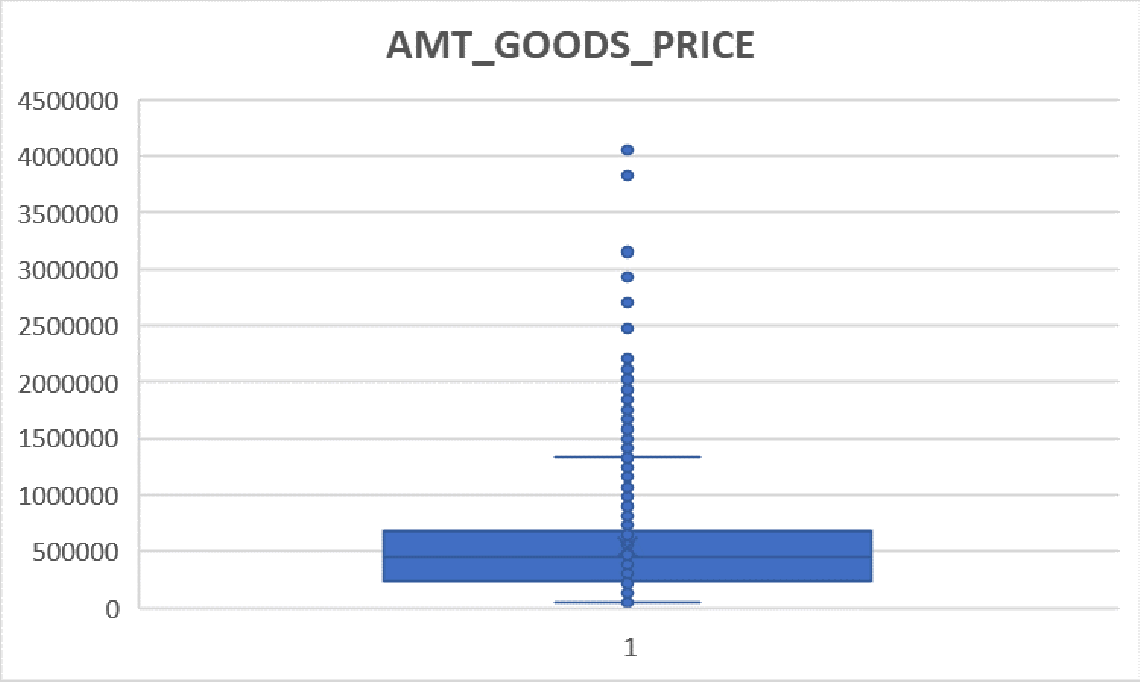


Q1	112500	270000	16456.5	238500
Q3	202500	808650	34596	679500
IRQ	90000	538650	18139.5	441000
UPPER LIMIT	337500	1616625	61805.25	1341000
LOWER LIMIT	-22500	-537975	-10752.75	-423000
	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE


AMT_INCOME_TOTAL
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
1

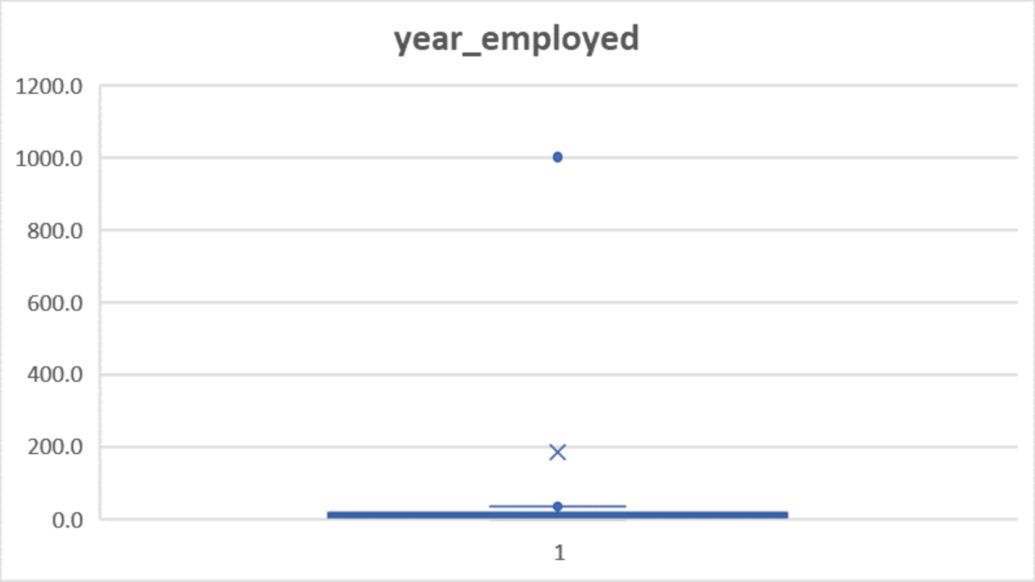

AMT_CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1


AMT_ANNUITY
300000
250000
200000
150000
100000
50000
0
1


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

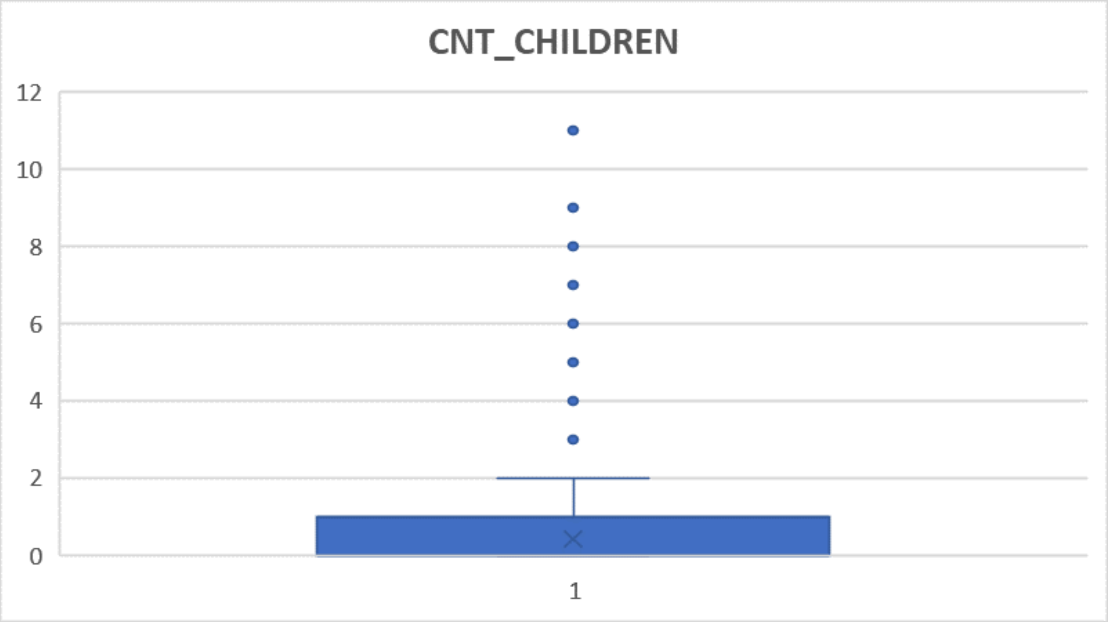
Part 2:

1. year_employed



In column “year_employed” we can see people being employed for 1001 yrs which is not possible.

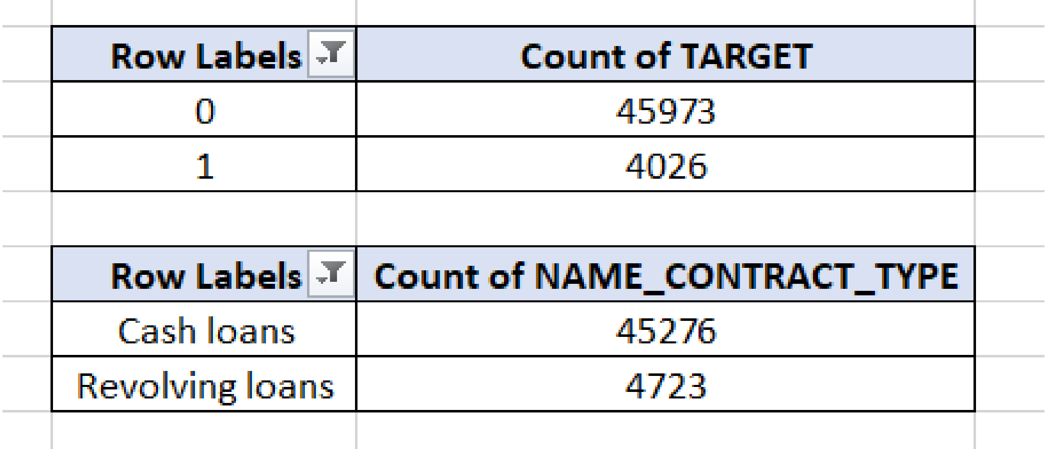
2. CNT_CHILDREN



Column “CNT_CHILDREN” shows people are having 11 children which is impossible in today’s age.

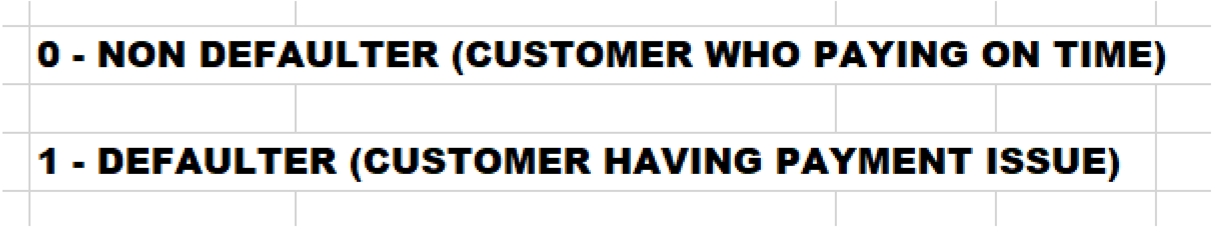
C. Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

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



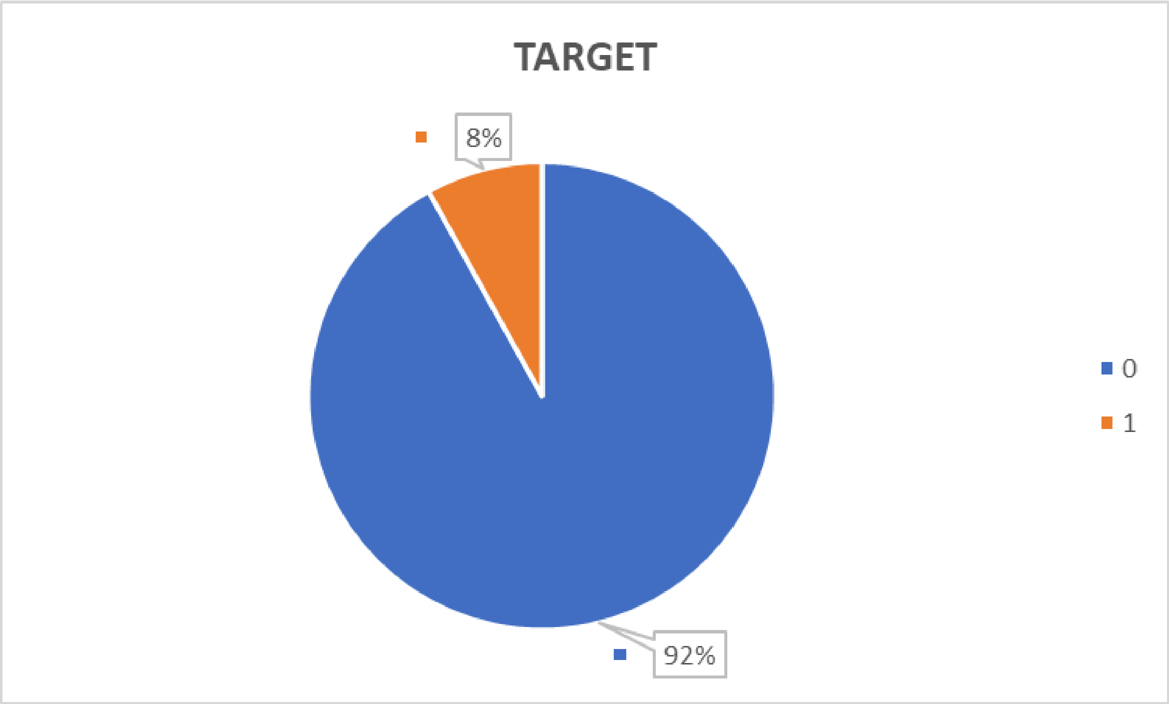
Row Labels	Count of TARGET
0	45973
1	4026

Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	45276
Revolving loans	4723

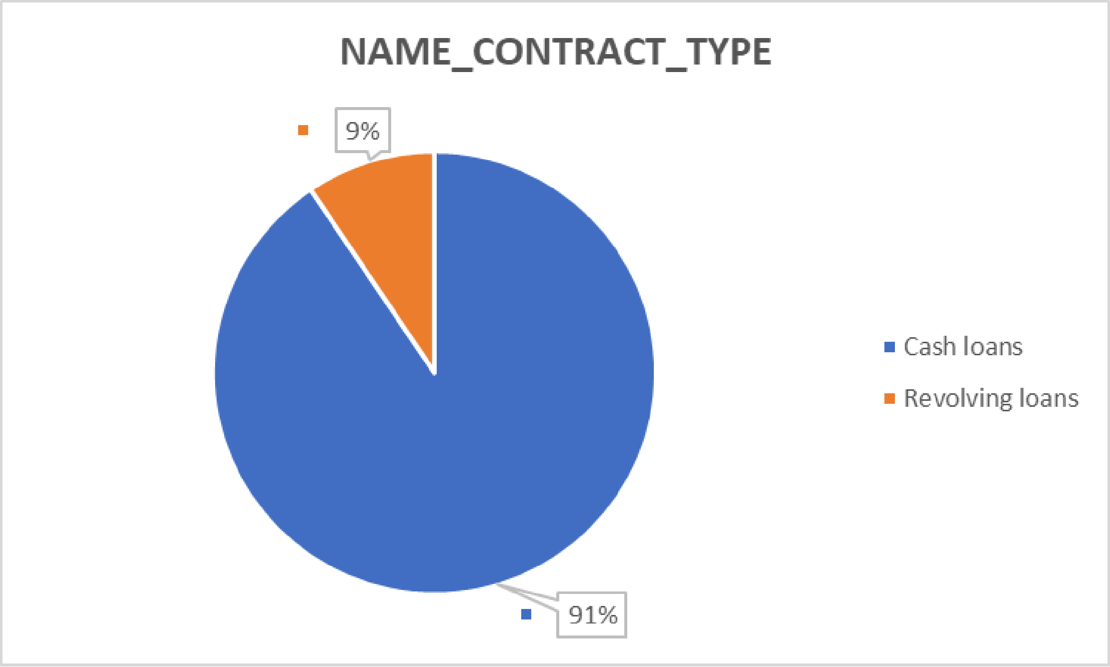


0 - NON DEFAULTER (CUSTOMER WHO PAYING ON TIME)

1 - DEFAULTER (CUSTOMER HAVING PAYMENT ISSUE)



Almost 92% of clients repay loans on time. Whereas, 8% of clients are defaulters.



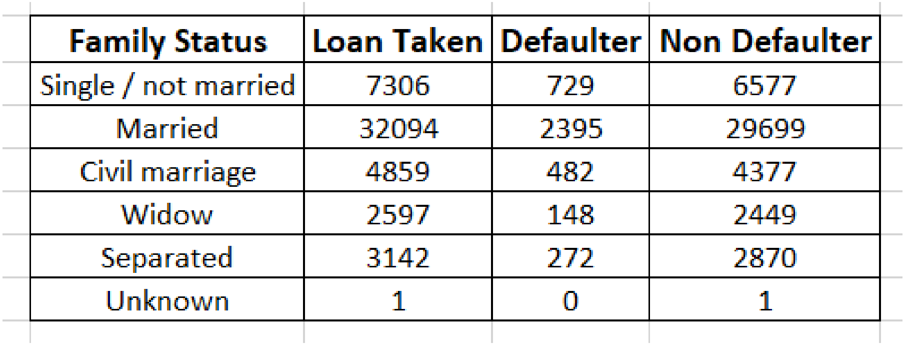
91% of clients applied for cash loans. And, 9% applied for revolving loans.

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.

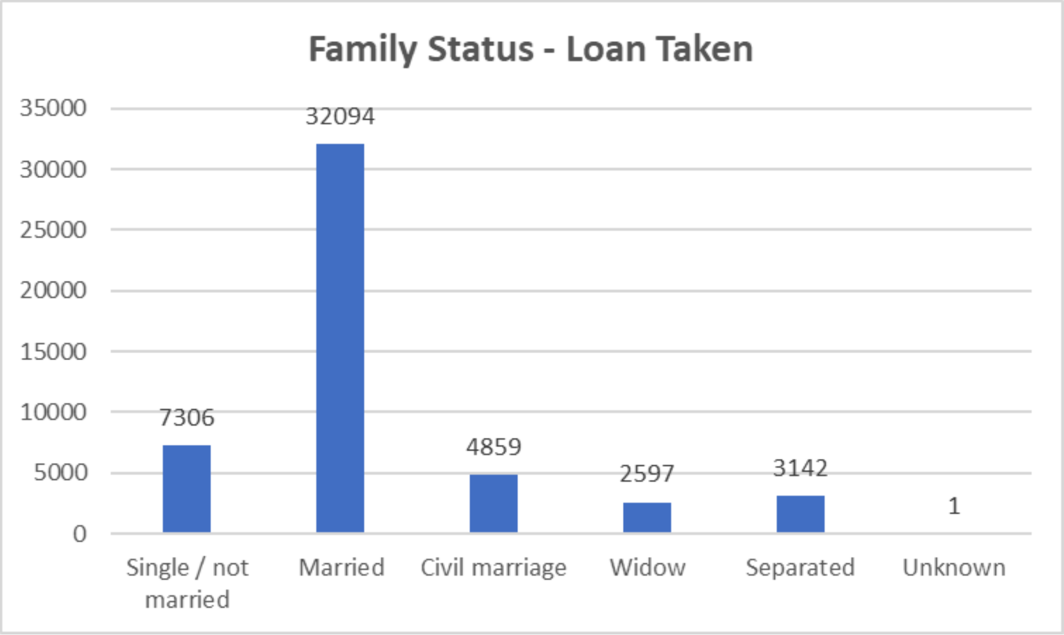
- **Task:** Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

Univariate:

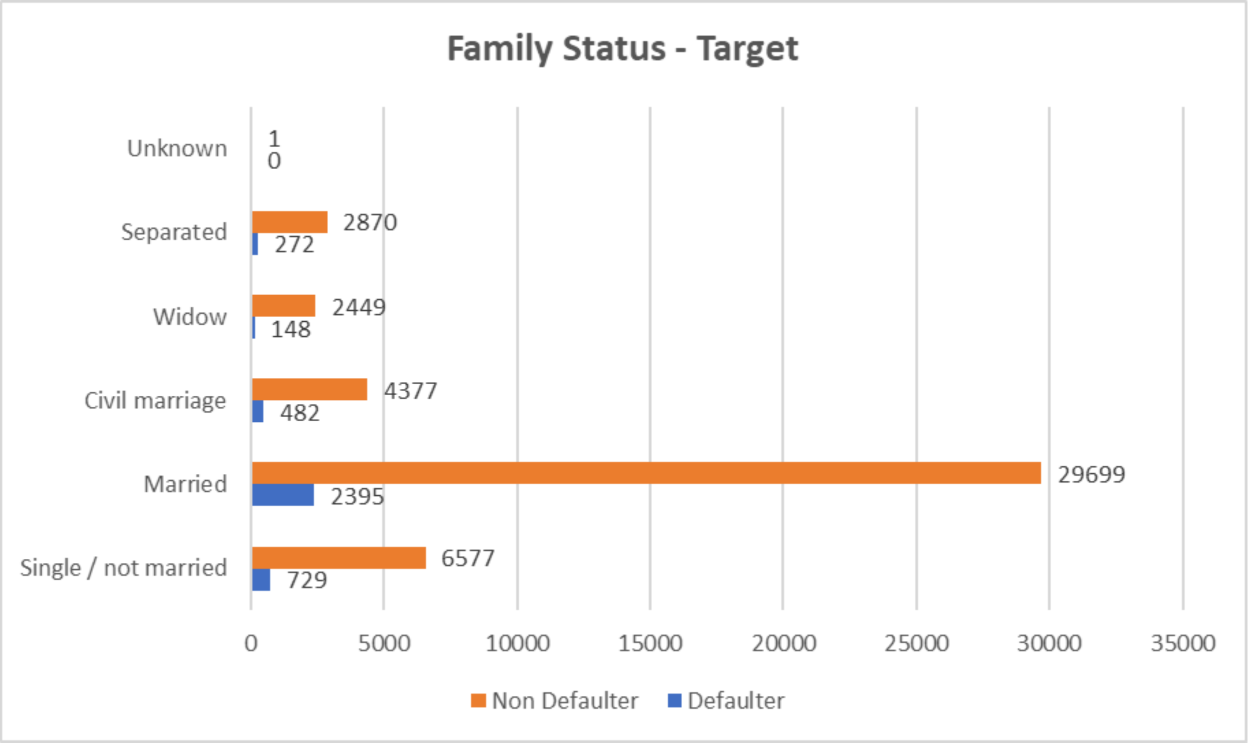
1.



Family Status	Loan Taken	Defaulter	Non Defaulter
Single / not married	7306	729	6577
Married	32094	2395	29699
Civil marriage	4859	482	4377
Widow	2597	148	2449
Separated	3142	272	2870
Unknown	1	0	1

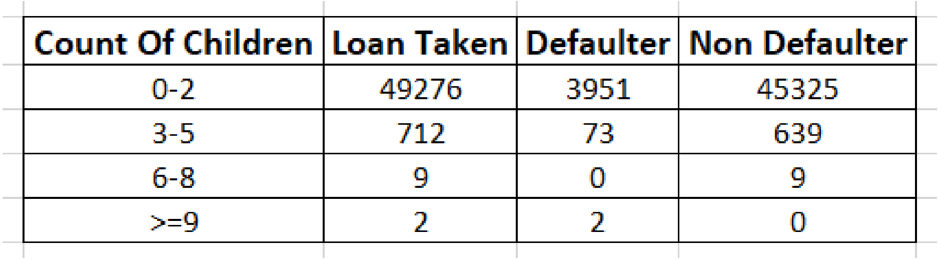


We can observe that most clients who took the loan are married, then followed by single and so on.

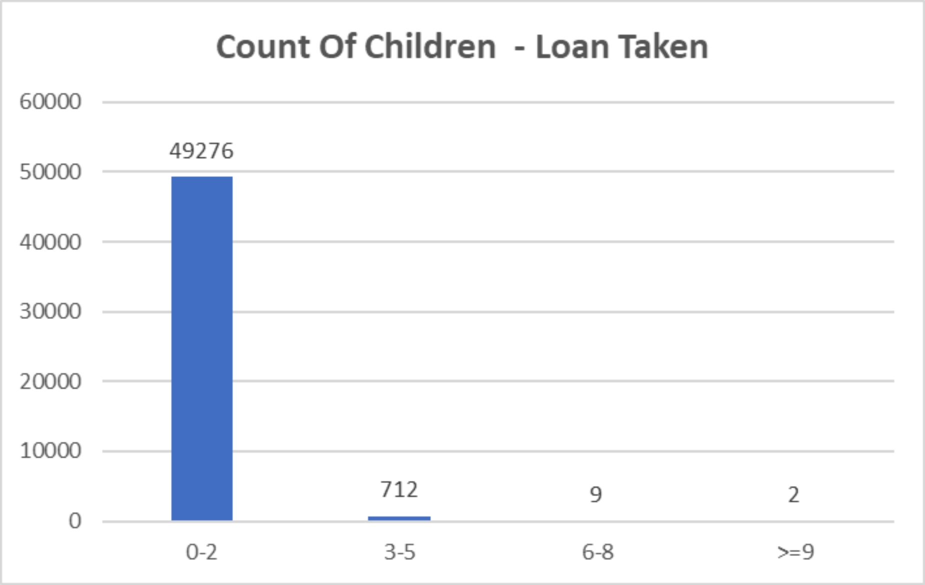


Also, most clients who repay their loans on time are married then followed by singles.

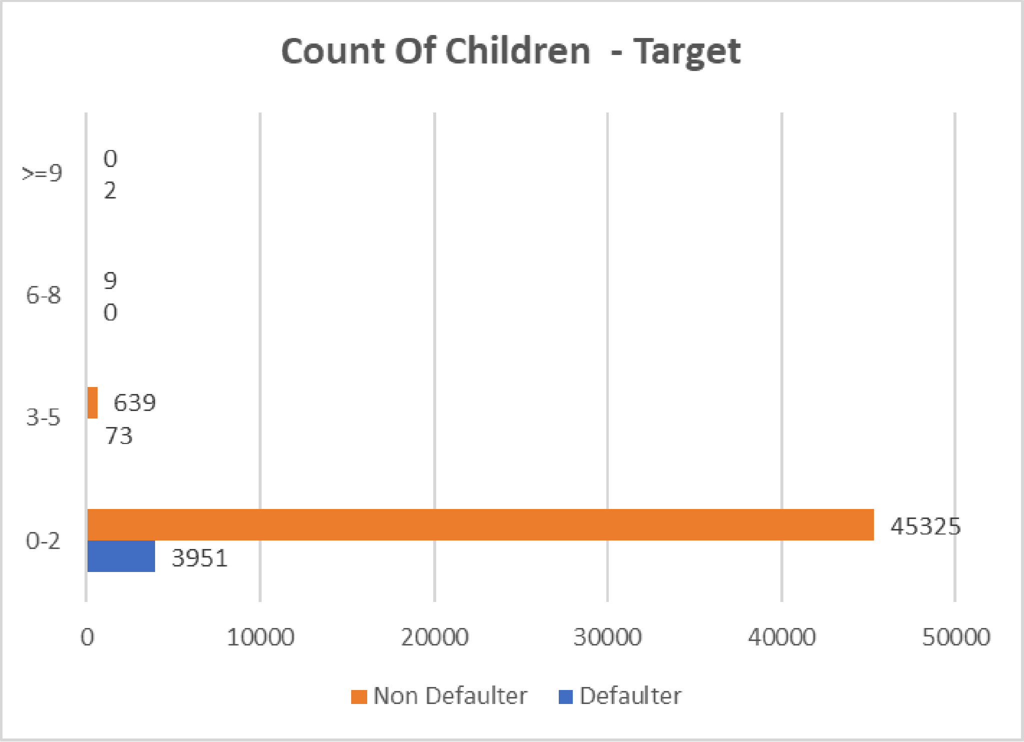
2.



Count Of Children	Loan Taken	Defaulter	Non Defaulter
0-2	49276	3951	45325
3-5	712	73	639
6-8	9	0	9
>=9	2	2	0

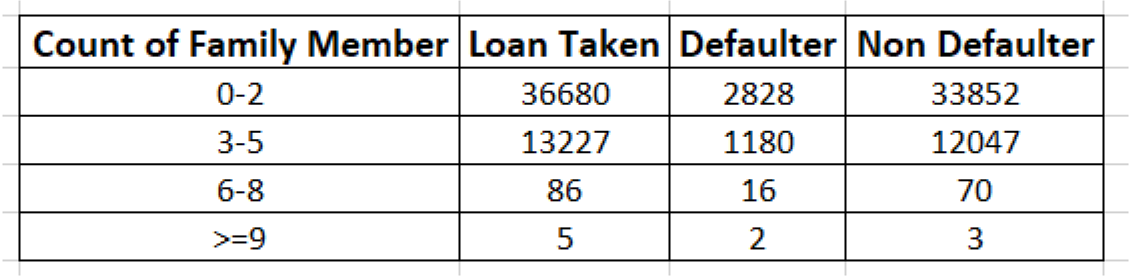


This analysis shows that those who have 0 - 2 children are likely to take out loans.

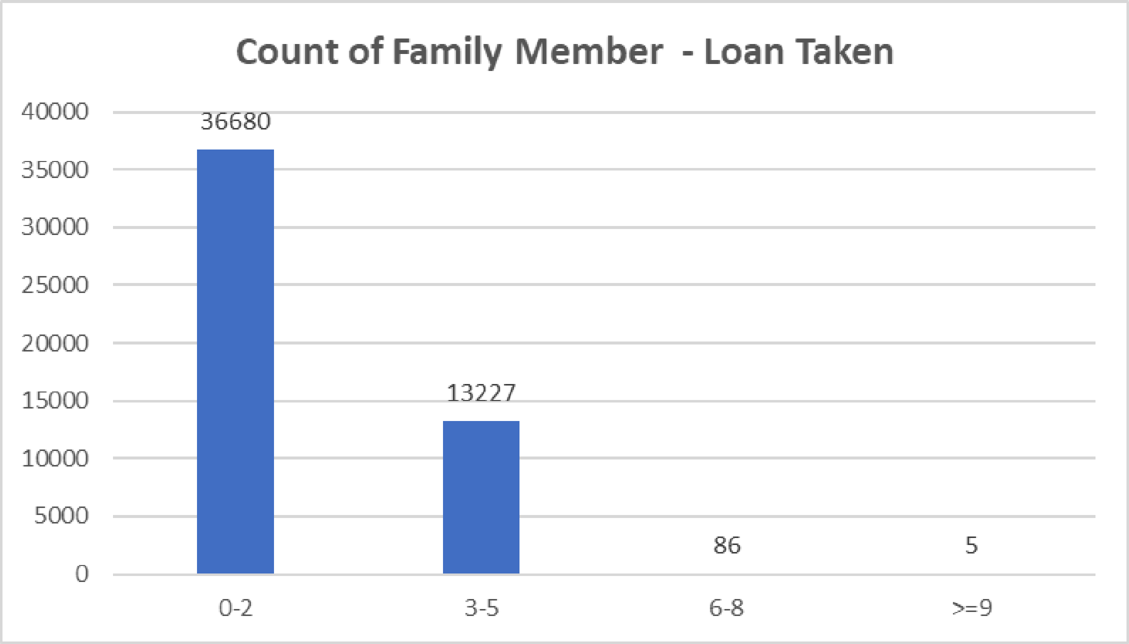


Similarly, clients having 0 - 2 children repay their loans on time.

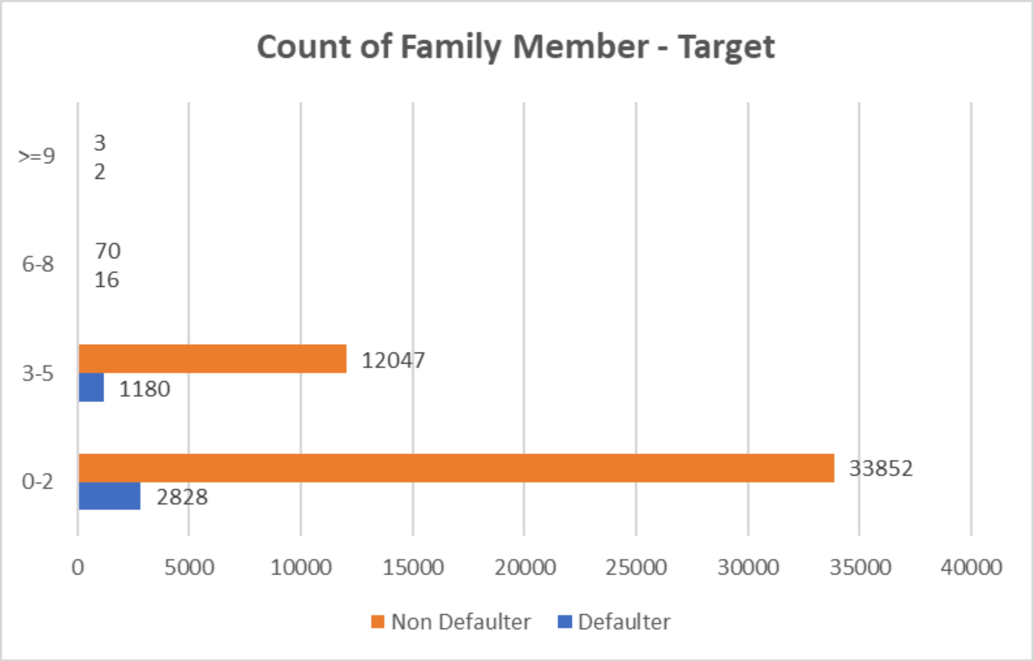
3.



Count of Family Member	Loan Taken	Defaulter	Non Defaulter
0-2	36680	2828	33852
3-5	13227	1180	12047
6-8	86	16	70
>=9	5	2	3

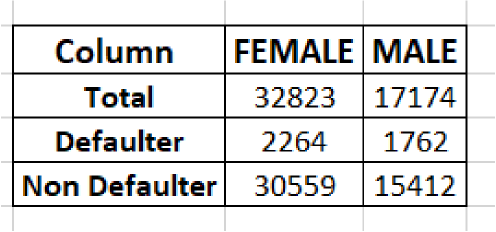


Clients having 0 - 2 members in their family are likely to take out loans.

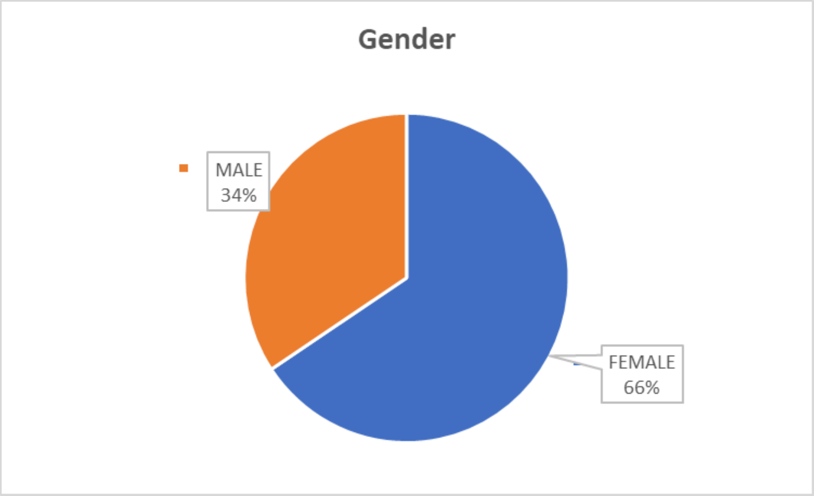


Similarly, clients with 0 - 2 members in their family repay loans on time followed by clients with 3 - 5 members.

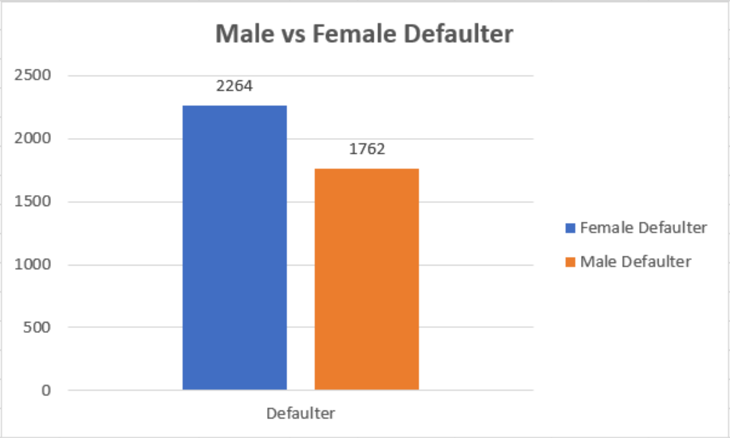
4.

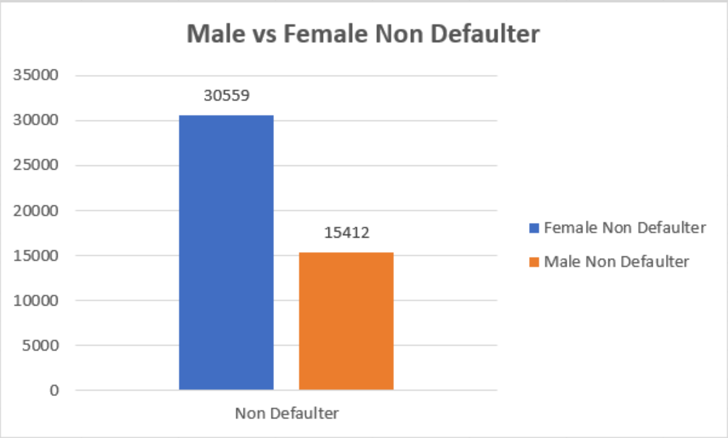


Column	FEMALE	MALE
Total	32823	17174
Defaulter	2264	1762
Non Defaulter	30559	15412

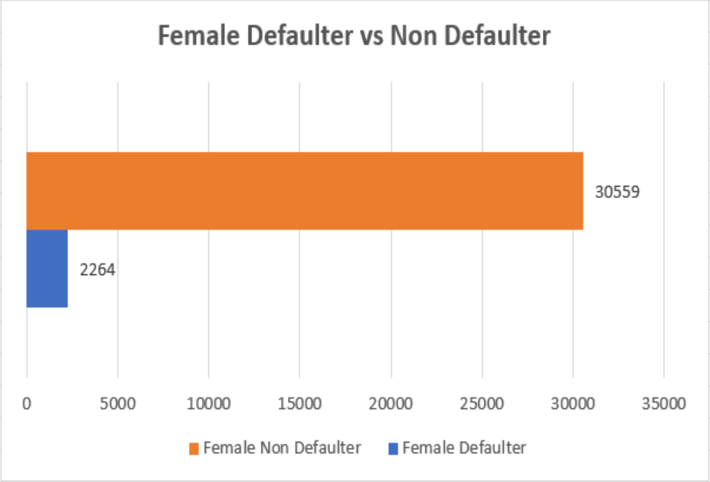


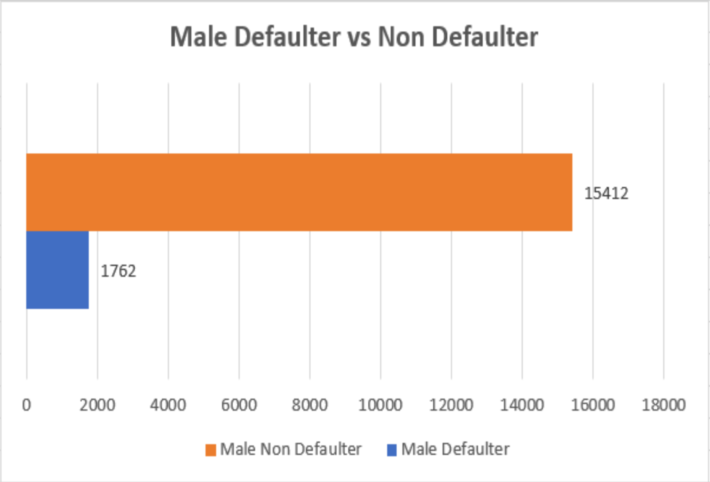
More females are likely to take loans compared to male. 66% female and 34% male.


Male vs Female Defaulter


Male vs Female Non Defaulter

In both cases females lead because they took out more loans compared to male.

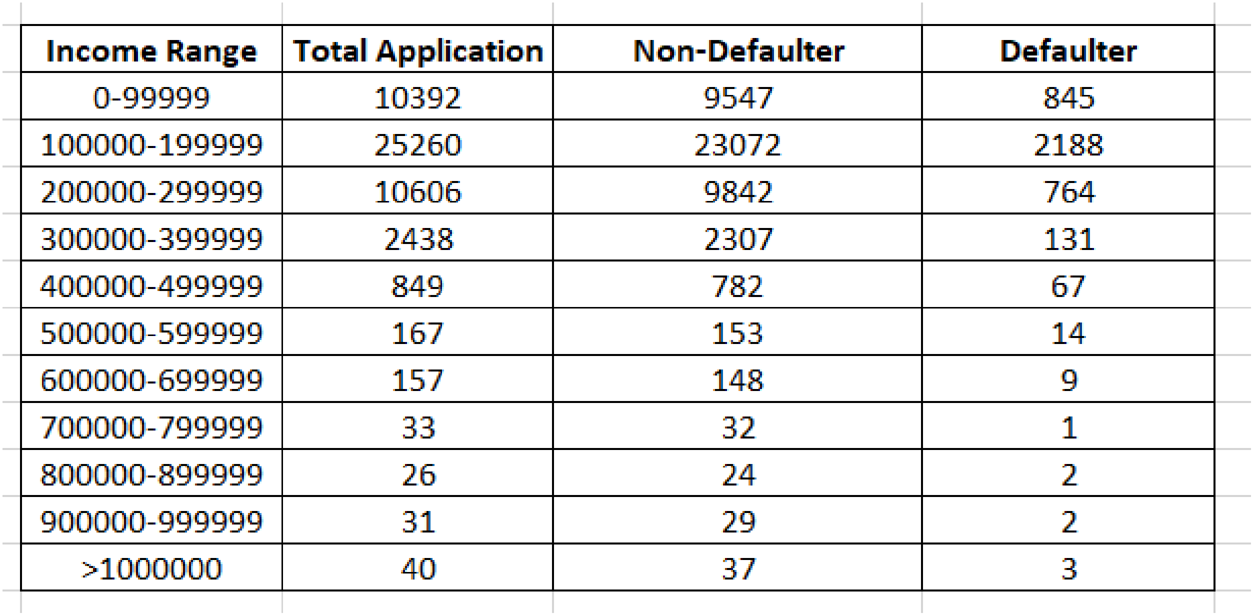

Female Defaulter vs Non Defaulter


Male Defaulter vs Non Defaulter

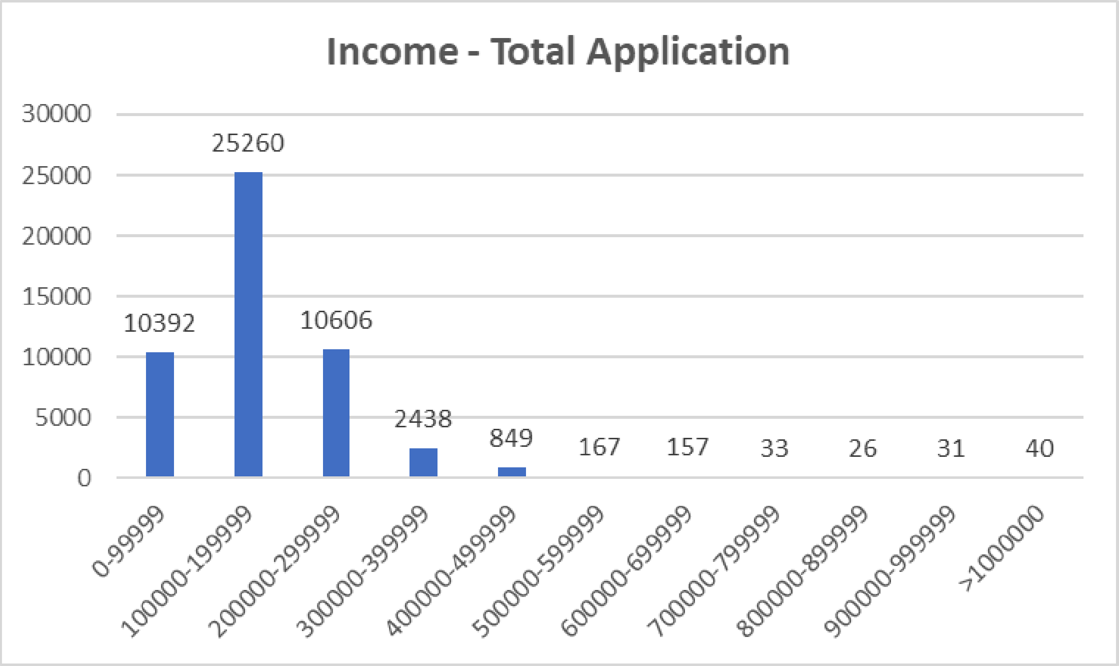
Here we can observe that most of the clients (male and female) paid their loans on time.

Segmented Univariate:

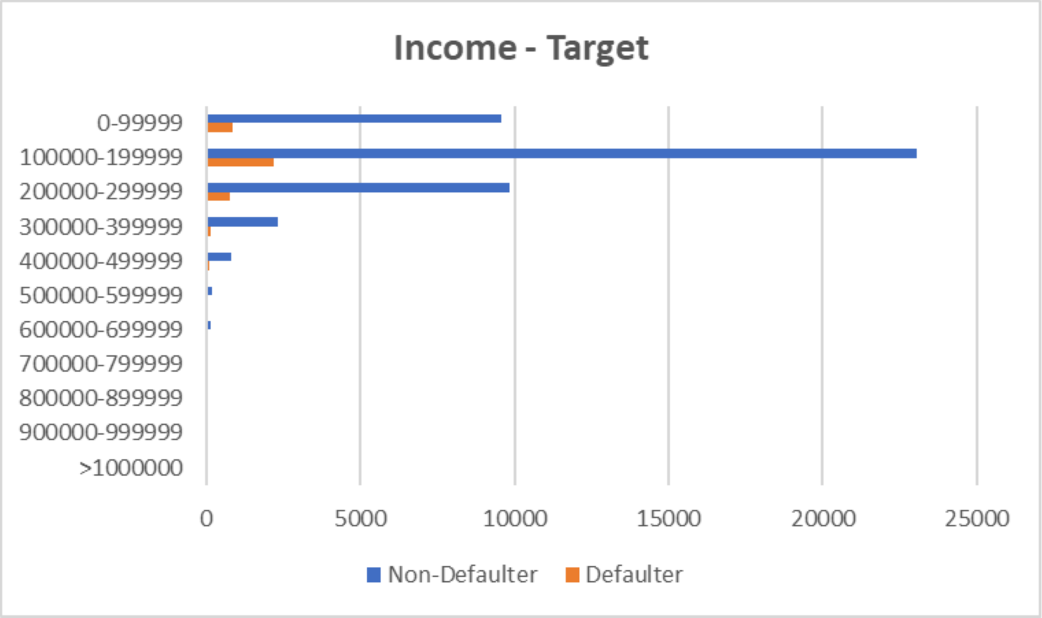
1.



Income Range	Total Application	Non-Defaulter	Defaulter
0-99999	10392	9547	845
100000-199999	25260	23072	2188
200000-299999	10606	9842	764
300000-399999	2438	2307	131
400000-499999	849	782	67
500000-599999	167	153	14
600000-699999	157	148	9
700000-799999	33	32	1
800000-899999	26	24	2
900000-999999	31	29	2
>1000000	40	37	3

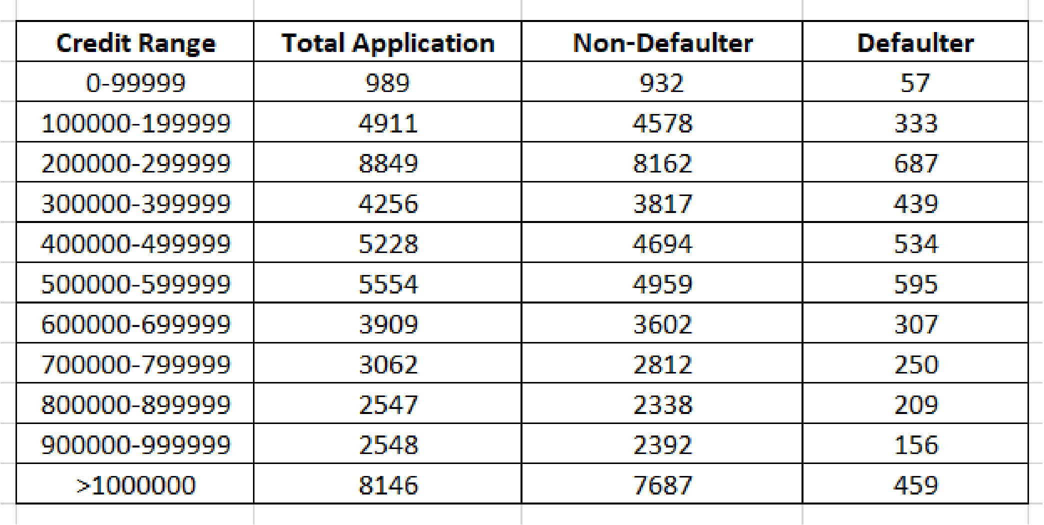


Most clients having income of 100,000 - 199,999 are likely to take out loans followed by the income range of 200,000 - 299,999 and 0 - 99,999.

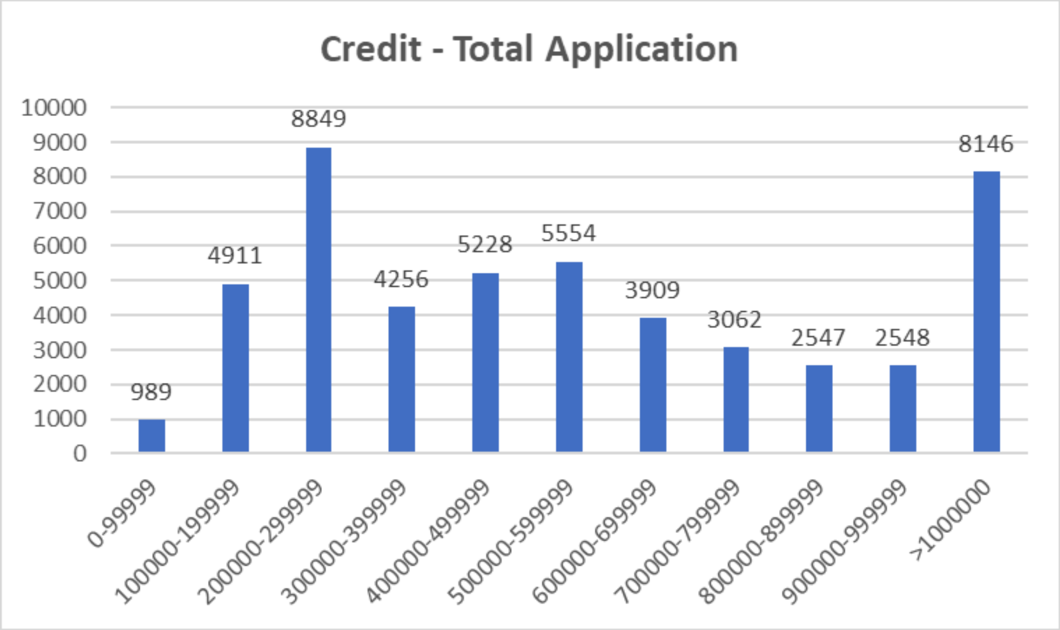


In most cases we can observe that 80% - 90% of times clients are repaying their loans on time.

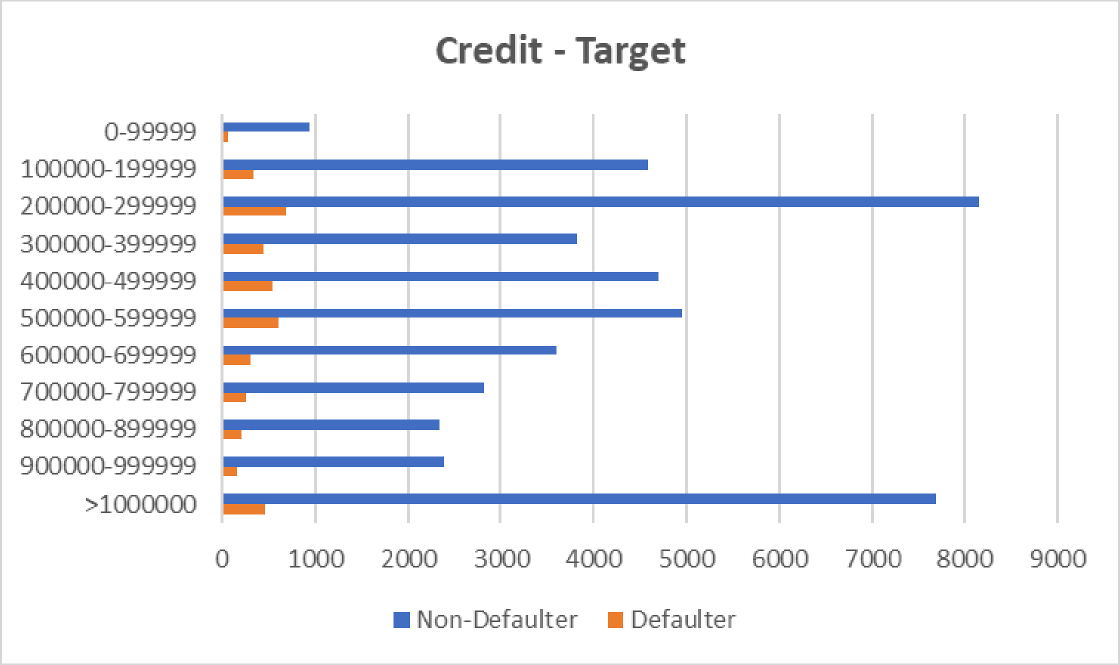
2.



Credit Range	Total Application	Non-Defaulter	Defaulter
0-99999	989	932	57
100000-199999	4911	4578	333
200000-299999	8849	8162	687
300000-399999	4256	3817	439
400000-499999	5228	4694	534
500000-599999	5554	4959	595
600000-699999	3909	3602	307
700000-799999	3062	2812	250
800000-899999	2547	2338	209
900000-999999	2548	2392	156
>1000000	8146	7687	459

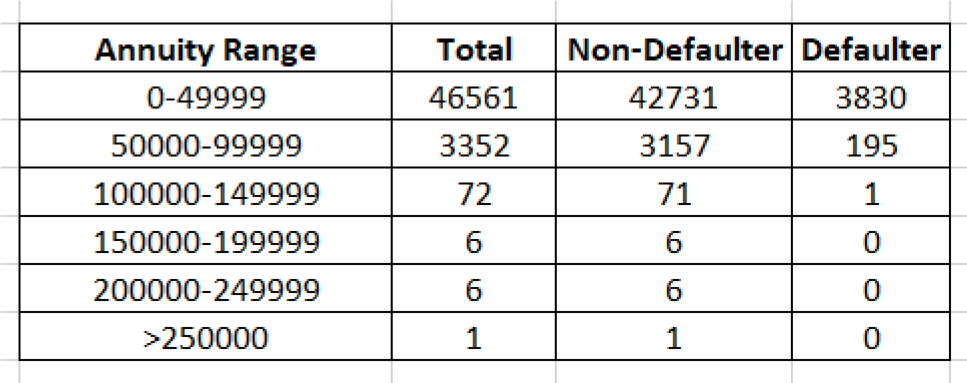


Clients having credit range from 200,000 - 299,999 have taken more loans followed by clients with credit range of more than 1,000,000.

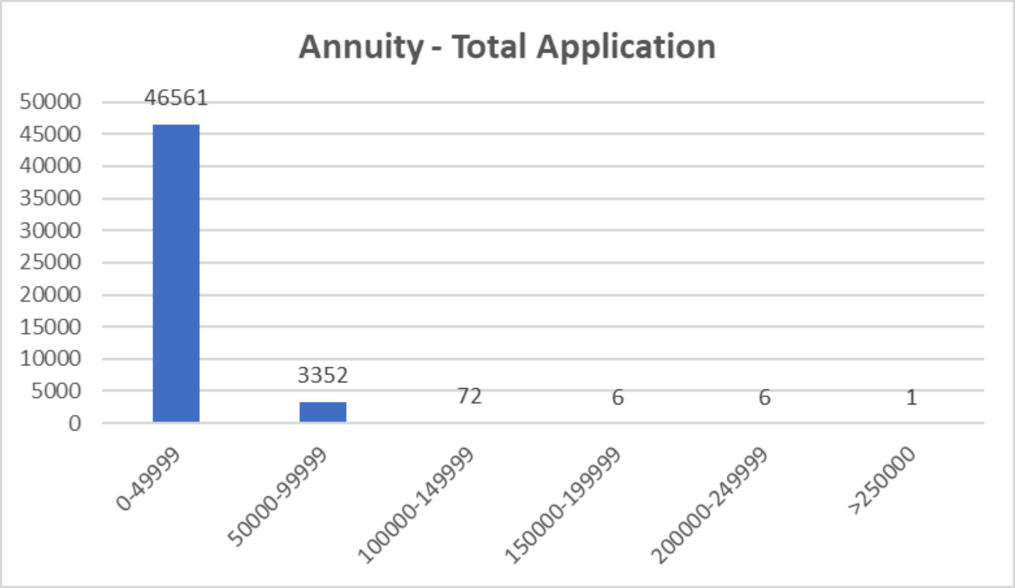


Similarly, the same credit ranges lead here. With most of them being non defaulters.

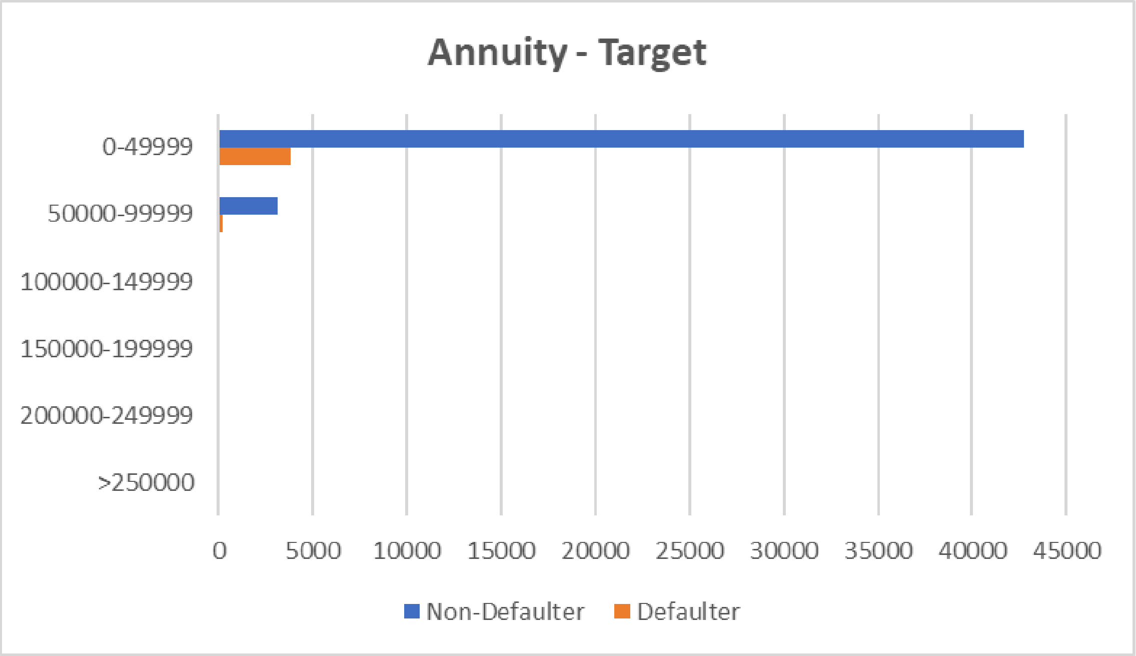
3.



Annuity Range	Total	Non-Defaulter	Defaulter
0-49999	46561	42731	3830
50000-99999	3352	3157	195
100000-149999	72	71	1
150000-199999	6	6	0
200000-249999	6	6	0
>250000	1	1	0



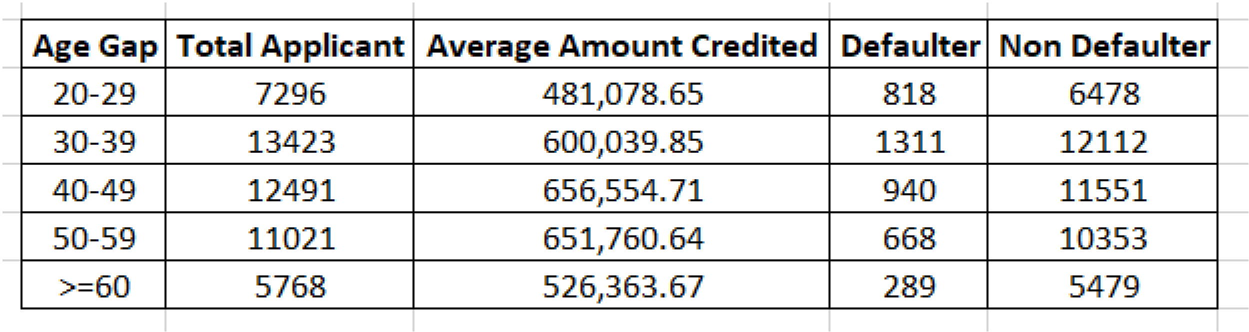
Annuity range of 0 - 49,999 have most clients with the count of 46561. And the least count is of 1 with an annuity more than 250,000.



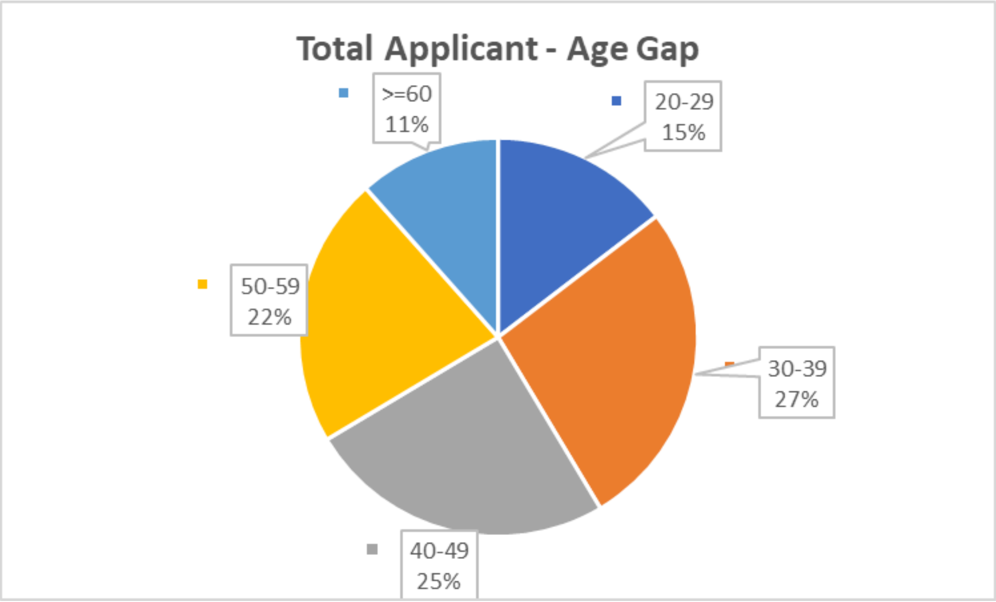
Majority of the clients repay their loans on time.

Bivariate:

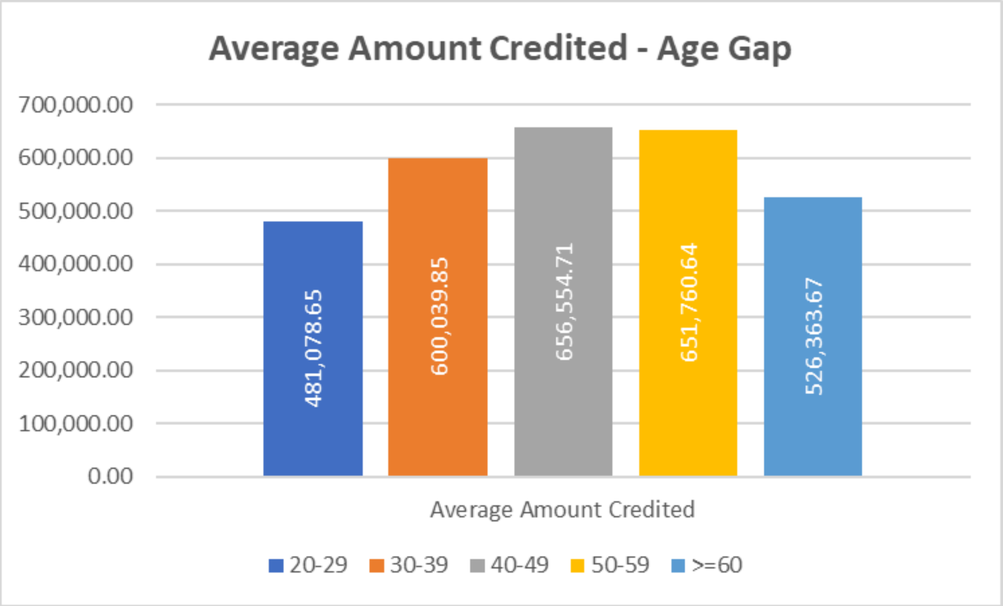
1.



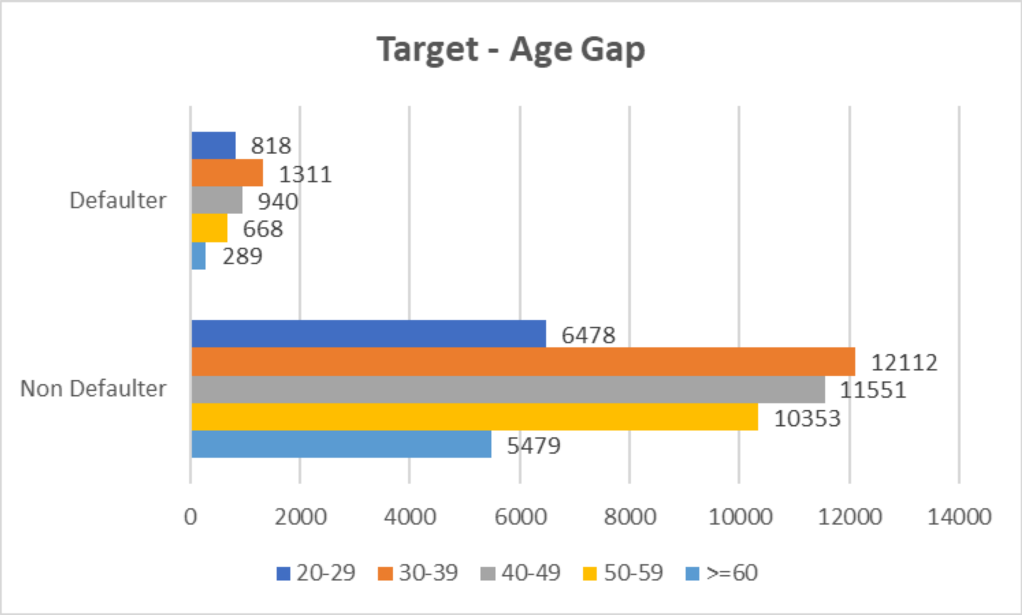
Age Gap	Total Applicant	Average Amount Credited	Defaulter	Non Defaulter
20-29	7296	481,078.65	818	6478
30-39	13423	600,039.85	1311	12112
40-49	12491	656,554.71	940	11551
50-59	11021	651,760.64	668	10353
>=60	5768	526,363.67	289	5479



Majority of the clients are from the age range of 30 - 39 then closely followed by 40 - 49 and 50 - 59, with the least number of clients from the age range of more than or equal to 60.

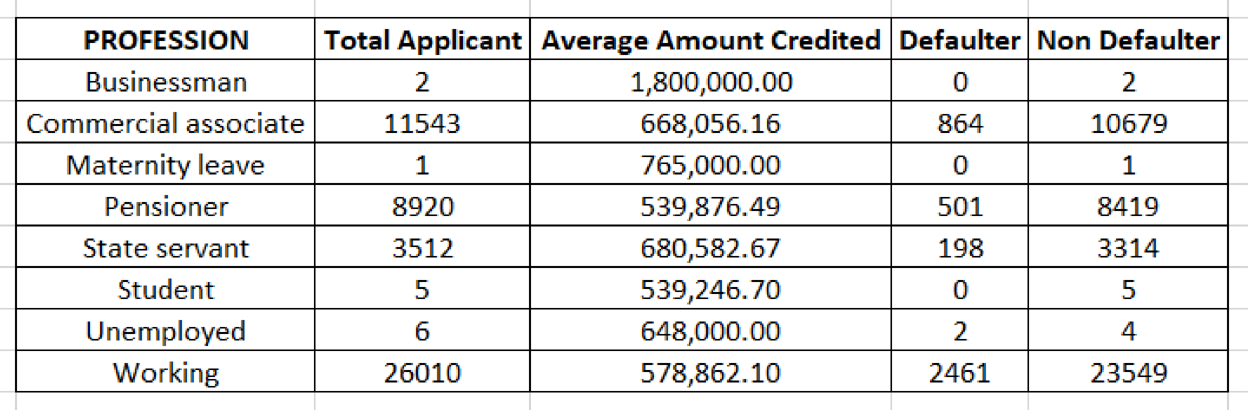


Age group of 40 - 49 have the most amount credited. Least is 20 - 29.

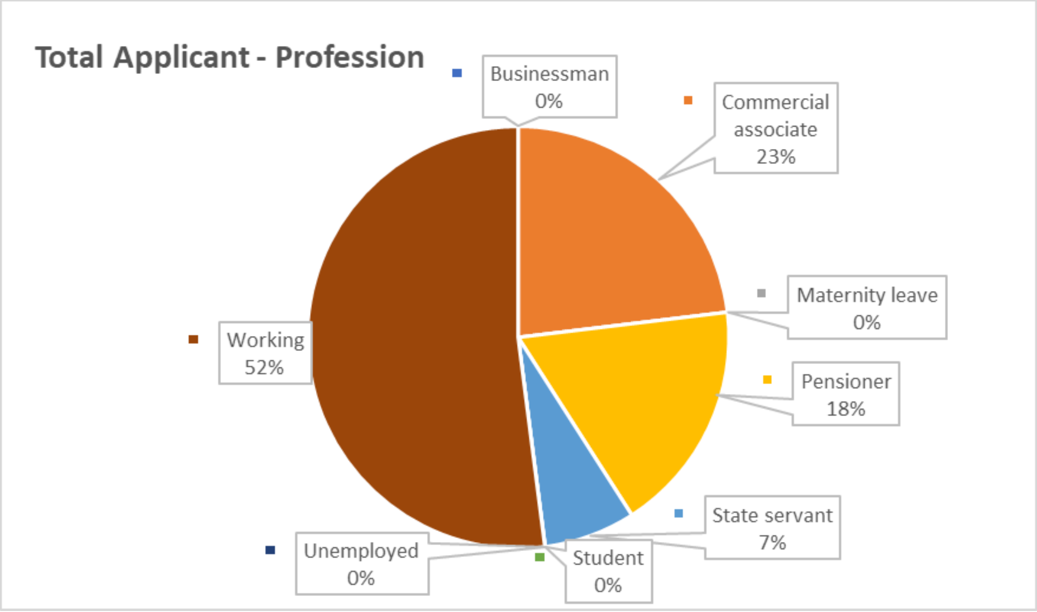


All age groups are likely to repay their loans on time. Highest being 30 - 39.

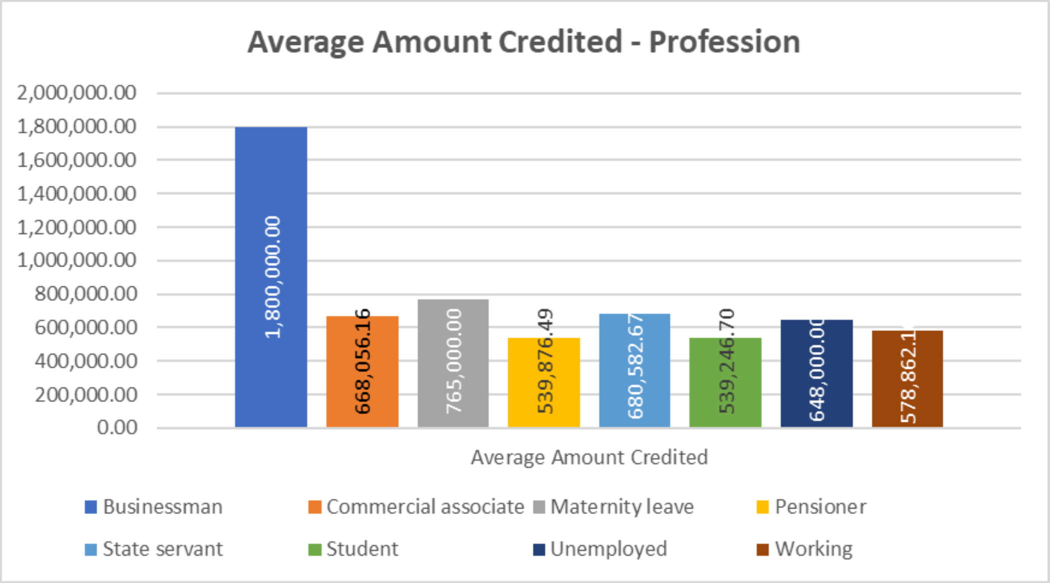
2.



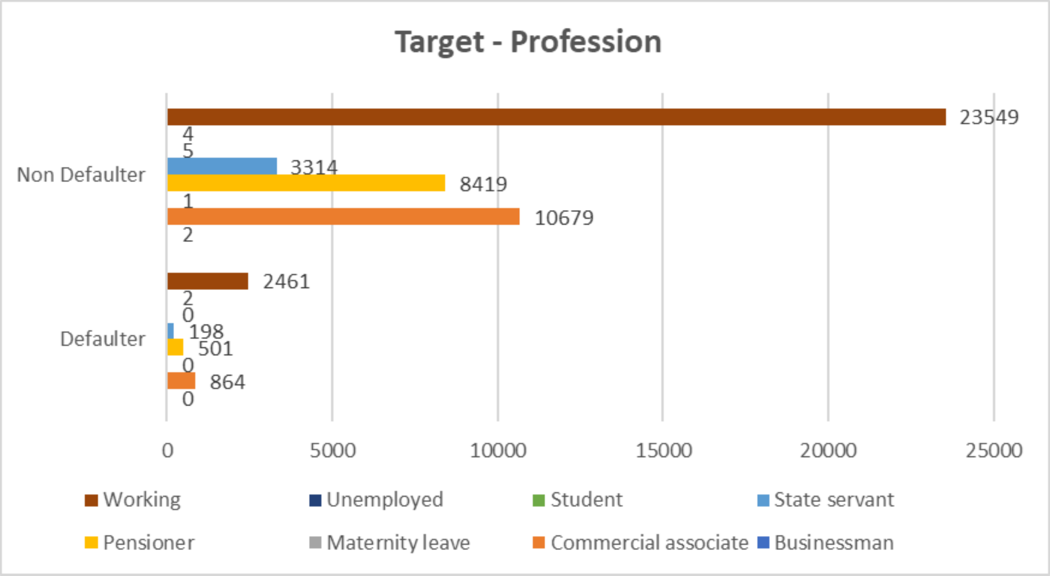
PROFESSION	Total Applicant	Average Amount Credited	Defaulter	Non Defaulter
Businessman	2	1,800,000.00	0	2
Commercial associate	11543	668,056.16	864	10679
Maternity leave	1	765,000.00	0	1
Pensioner	8920	539,876.49	501	8419
State servant	3512	680,582.67	198	3314
Student	5	539,246.70	0	5
Unemployed	6	648,000.00	2	4
Working	26010	578,862.10	2461	23549



Most applicants are Working with 52%. The least are Businessman, Maternity Leave, Student and Unemployed.



Most amount credited to a client is Businessman.



Majority of the clients are non defaulters with most being Working.

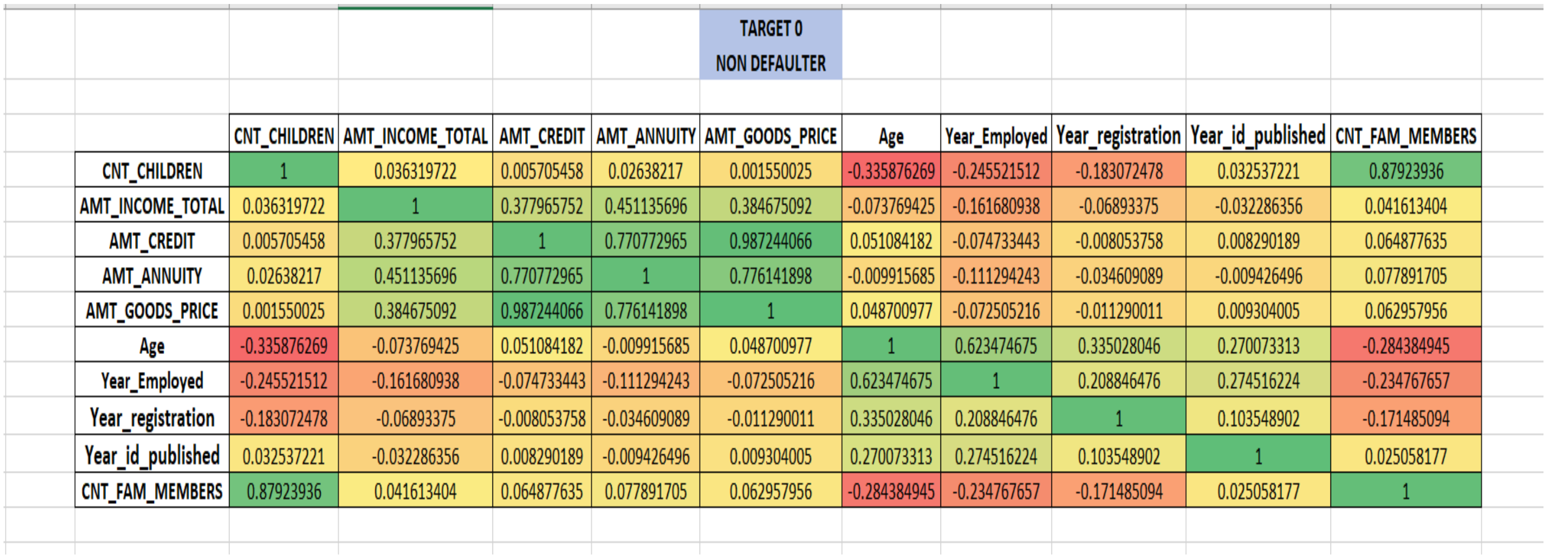
E. Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

- **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.

Correlation measures the statistical relationship between two variables, indicating how changes in one variable are associated with changes in the other. The `CORREL` function calculates the Pearson correlation coefficient, which ranges from -1 to 1:

- 1 indicates a perfect positive correlation,
- -1 indicates a perfect negative correlation,
- 0 means no correlation.

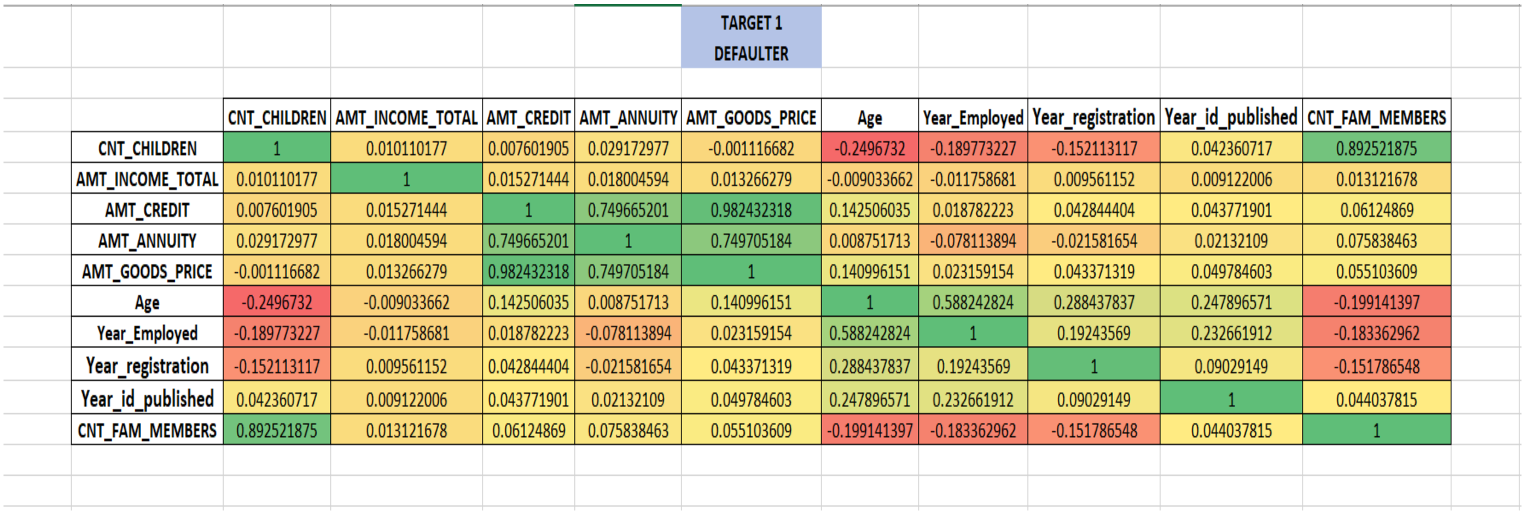
Target 0, Non Defaulter:



TARGET 0
NON DEFAULTER

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	Age	Year_Employed	Year_registration	Year_id_published	CNT_FAM_MEMBERS
CNT_CHILDREN	1	0.036319722	0.005705458	0.02638217	0.001550025	-0.335876269	-0.245521512	-0.183072478	0.032537221	0.87923936
AMT_INCOME_TOTAL	0.036319722	1	0.377965752	0.451135696	0.384675092	-0.073769425	-0.161680938	-0.06893375	-0.032286356	0.041613404
AMT_CREDIT	0.005705458	0.377965752	1	0.770772965	0.987244066	0.051084182	-0.074733443	-0.008053758	0.008290189	0.064877635
AMT_ANNUITY	0.02638217	0.451135696	0.770772965	1	0.776141898	-0.009915685	-0.111294243	-0.034609089	-0.009426496	0.077891705
AMT_GOODS_PRICE	0.001550025	0.384675092	0.987244066	0.776141898	1	0.048700977	-0.072505216	-0.011290011	0.009304005	0.062957956
Age	-0.335876269	-0.073769425	0.051084182	-0.009915685	0.048700977	1	0.623474675	0.335028046	0.270073313	-0.284384945
Year_Employed	-0.245521512	-0.161680938	-0.074733443	-0.111294243	-0.072505216	0.623474675	1	0.208846476	0.274516224	-0.234767657
Year_registration	-0.183072478	-0.06893375	-0.008053758	-0.034609089	-0.011290011	0.335028046	0.208846476	1	0.103548902	-0.171485094
Year_id_published	0.032537221	-0.032286356	0.008290189	-0.009426496	0.009304005	0.270073313	0.274516224	0.103548902	1	0.025058177
CNT_FAM_MEMBERS	0.87923936	0.041613404	0.064877635	0.077891705	0.062957956	-0.284384945	-0.234767657	-0.171485094	0.025058177	1

Target 1, Defaulter:



TARGET 1
DEFAULTER

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	Age	Year_Employed	Year_registration	Year_id_published	CNT_FAM_MEMBERS
CNT_CHILDREN	1	0.010110177	0.007601905	0.029172977	-0.001116682	-0.2496732	-0.189773227	-0.152113117	0.042360717	0.892521875
AMT_INCOME_TOTAL	0.010110177	1	0.015271444	0.018004594	0.013266279	-0.009033662	-0.011758681	0.009561152	0.009122006	0.013121678
AMT_CREDIT	0.007601905	0.015271444	1	0.749665201	0.982432318	0.142506035	0.018782223	0.042844404	0.043771901	0.06124869
AMT_ANNUITY	0.029172977	0.018004594	0.749665201	1	0.749705184	0.008751713	-0.078113894	-0.021581654	0.02132109	0.075838463
AMT_GOODS_PRICE	-0.001116682	0.013266279	0.982432318	0.749705184	1	0.140996151	0.023159154	0.043371319	0.049784603	0.055103609
Age	-0.2496732	-0.009033662	0.142506035	0.008751713	0.140996151	1	0.588242824	0.288437837	0.247896571	-0.199141397
Year_Employed	-0.189773227	-0.011758681	0.018782223	-0.078113894	0.023159154	0.588242824	1	0.19243569	0.232661912	-0.183362962
Year_registration	-0.152113117	0.009561152	0.042844404	-0.021581654	0.043371319	0.288437837	0.19243569	1	0.09029149	-0.151786548
Year_id_published	0.042360717	0.009122006	0.043771901	0.02132109	0.049784603	0.247896571	0.232661912	0.09029149	1	0.044037815
CNT_FAM_MEMBERS	0.892521875	0.013121678	0.06124869	0.075838463	0.055103609	-0.199141397	-0.183362962	-0.151786548	0.044037815	1

Result:

Through this project, I successfully identified key factors that influence loan default, such as customer attributes and loan characteristics, by performing Exploratory Data Analysis (EDA). I handled missing data, detected outliers, and addressed data imbalance, which enhanced the accuracy of the analysis. The correlation analysis revealed strong indicators of loan default, helping identify risky applicants. This project deepened my understanding of how data-driven insights can improve decision-making in loan approval, helping banks mitigate financial risks and improve customer targeting.

Excel File Link: project_6_excel.xlsb