BANK LOAN CASE STUDY

MIRRA. G

PROJECT DESCRIPTION:

The project focuses on analysing a bank loan application dataset to identify and handle missing data effectively using Microsoft Excel. The process includes identifying columns with high percentages of missing values, dropping irrelevant columns, and replacing missing values with appropriate substitutes. The aim is to ensure data accuracy for subsequent analysis. The project also involves exploring relationships between various factors and loan default rates through correlation analysis and data visualization. This comprehensive data cleaning and analysis approach enhances the reliability and insights derived from the dataset.

TECH STACK USED: Microsoft Excel

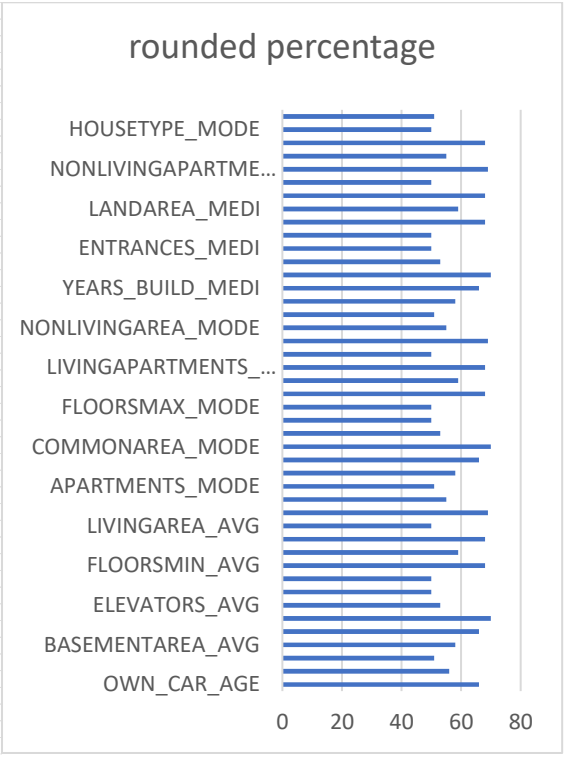
A) ***Identify Missing Data and Deal with it Appropriately*****: As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.**

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.**

Application dataset:

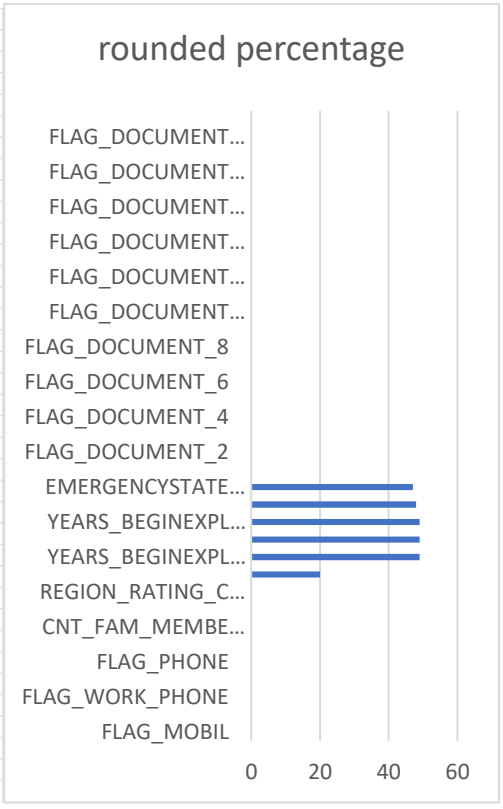
- We begin by identifying the number of null/blank values in each column and if the percentage of null/blank values if more than 50, we drop the column. The following columns are dropped on this criterion:

column_name	count of null	percentage of null	rounded percentage
OWN_CAR_AGE	32950	65.90131803	66
EXT_SOURCE_1	28172	56.3451269	56
APARTMENTS_AVG	25385	50.77101542	51
BASEMENTAREA_AVG	29199	58.39916798	58
YEARS_BUILD_AVG	33239	66.47932959	66
COMMONAREA_AVG	34960	69.92139843	70
ELEVATORS_AVG	26651	53.30306606	53
ENTRANCES_AVG	25195	50.39100782	50
FLOORSMAX_AVG	24875	49.75099502	50
FLOORSMIN_AVG	33894	67.78935579	68
LANDAREA_AVG	29721	59.44318886	59
LIVINGAPARTMENTS_AVG	34226	68.45336907	68
LIVINGAREA_AVG	25137	50.2750055	50
NONLIVINGAPARTMENTS_AVG	34714	69.42938859	69
NONLIVINGAREA_AVG	27572	55.1451029	55
APARTMENTS_MODE	25385	50.77101542	51
BASEMENTAREA_MODE	29199	58.39916798	58
YEARS_BUILD_MODE	33239	66.47932959	66
COMMONAREA_MODE	34960	69.92139843	70
ELEVATORS_MODE	26651	53.30306606	53
ENTRANCES_MODE	25195	50.39100782	50
FLOORSMAX_MODE	24875	49.75099502	50
FLOORSMIN_MODE	33894	67.78935579	68
LANDAREA_MODE	29721	59.44318886	59
LIVINGAPARTMENTS_MODE	34226	68.45336907	68
LIVINGAREA_MODE	25137	50.2750055	50
NONLIVINGAPARTMENTS_MODE	34714	69.42938859	69
NONLIVINGAREA_MODE	27572	55.1451029	55
APARTMENTS_MEDI	25385	50.77101542	51
BASEMENTAREA_MEDI	29199	58.39916798	58
YEARS_BUILD_MEDI	33239	66.47932959	66
COMMONAREA_MEDI	34960	69.92139843	70
ELEVATORS_MEDI	26651	53.30306606	53
ENTRANCES_MEDI	25195	50.39100782	50
FLOORSMAX_MEDI	24875	49.75099502	50
FLOORSMIN_MEDI	33894	67.78935579	68
LANDAREA_MEDI	29721	59.44318886	59
LIVINGAPARTMENTS_MEDI	34226	68.45336907	68
LIVINGAREA_MEDI	25137	50.2750055	50
NONLIVINGAPARTMENTS_MEDI	34714	69.42938859	69
NONLIVINGAREA_MEDI	27572	55.1451029	55
FONDKAPREMONT_MODE	34191	68.38336767	68
HOUSETYPE_MODE	25075	50.15100302	50
WALLSMATERIAL_MODE	25459	50.91901838	51



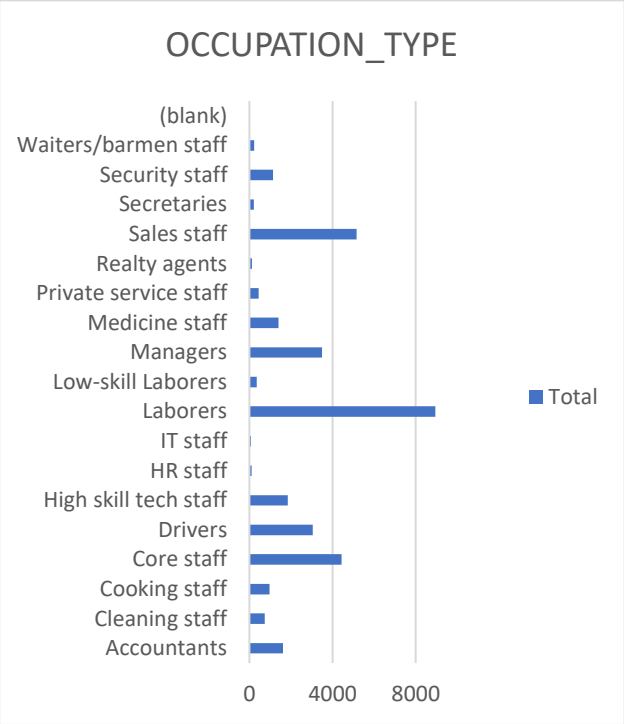
- We then identify the columns that would be unrelated to the analysis we are about to perform. Based on this, the following columns are dropped:

column_name	count of null	percentage of null	rounded percentage
FLAG_MOBIL	0	0	0
FLAG_EMP_PHONE	0	0	0
FLAG_WORK_PHONE	0	0	0
FLAG_CONT_MOBILE	0	0	0
FLAG_PHONE	0	0	0
FLAG_EMAIL	0	0	0
CNT_FAM_MEMBERS	1	0.00200004	0
REGION_RATING_CLIENT	0	0	0
REGION_RATING_CLIENT_W_CITY	0	0	0
EXT_SOURCE_3	9944	19.88839777	20
YEARS_BEGINEXPLUATATION_AVG	24394	48.78897578	49
YEARS_BEGINEXPLUATATION_MODE	24394	48.78897578	49
YEARS_BEGINEXPLUATATION_MEDI	24394	48.78897578	49
TOTALAREA_MODE	24148	48.29696594	48
EMERGENCYSTATE_MODE	23698	47.39694794	47
DAYS_LAST_PHONE_CHANGE	1	0.00200004	0
FLAG_DOCUMENT_2	0	0	0
FLAG_DOCUMENT_3	0	0	0
FLAG_DOCUMENT_4	0	0	0
FLAG_DOCUMENT_5	0	0	0
FLAG_DOCUMENT_6	0	0	0
FLAG_DOCUMENT_7	0	0	0
FLAG_DOCUMENT_8	0	0	0
FLAG_DOCUMENT_9	0	0	0
FLAG_DOCUMENT_10	0	0	0
FLAG_DOCUMENT_11	0	0	0
FLAG_DOCUMENT_12	0	0	0
FLAG_DOCUMENT_13	0	0	0
FLAG_DOCUMENT_14	0	0	0
FLAG_DOCUMENT_15	0	0	0
FLAG_DOCUMENT_16	0	0	0
FLAG_DOCUMENT_17	0	0	0
FLAG_DOCUMENT_18	0	0	0
FLAG_DOCUMENT_19	0	0	0
FLAG_DOCUMENT_20	0	0	0
FLAG_DOCUMENT_21	0	0	0

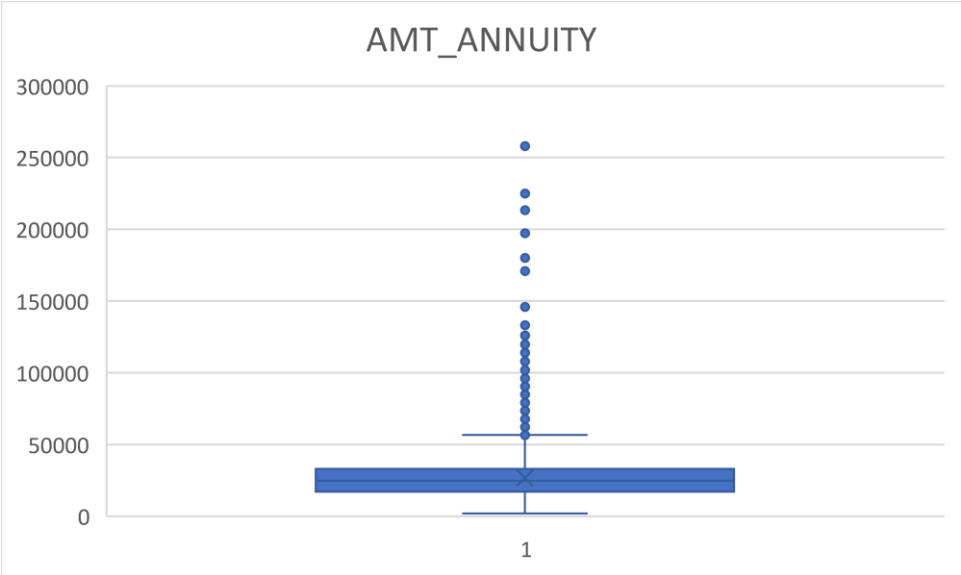


- We then replace the blank rows of the 'OCCUPATION_TYPE' column with the value 'Laborers' as that is the most occurring value in the column.

Row Labels	Count of 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
(blank)	



- We replace the blanks in 'AMT_ANNUITY' with the median of the column as there are outliers present.



median of AMT_ANNUITY	24700.5

- We replace the blanks in 'AMT_GOODS_PRICE' using the median of the column as it contains outliers.

median of AMT_GOODS_PRICE	450000

Previous application dataset:

- We drop the following columns as they are not relevant to our analysis: WEEKDAY_APPR_PROCESS_START, HOUR_APPR_PROCESS_START, FLAG_LAST_APPL_PER_CONTRACT, NFLAG_LAST_APPL_IN_DAY

- We drop the columns following columns as they have more than 50% of null values.

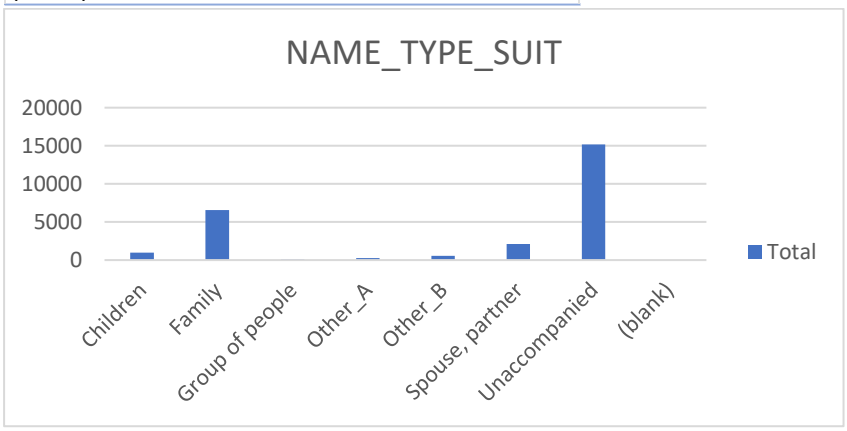
columns	count of nu	percentage of nu	rounded percentag
AMT_DOWN_PAYMENT	25198	50.39700794	50
RATE_DOWN_PAYMENT	25198	50.39700794	50
RATE_INTEREST_PRIMARY	49834	99.6699934	100
RATE_INTEREST_PRIVILEGED	49834	99.6699934	100

- We replace the blanks in the 'AMT_ANNUITY' column with the median of the column.

median of AMT_ANNUITY	10879.92

- We replace the blanks in the 'NAME_TYPE_SUIT' with the most occurring value in the column 'Unaccompanied'.

Row Labels	Count of NAME_TYPE_SUITE
Children	993
Family	6581
Group of people	76
Other_A	262
Other_B	551
Spouse, partner	2098
Unaccompanied	15195
(blank)	

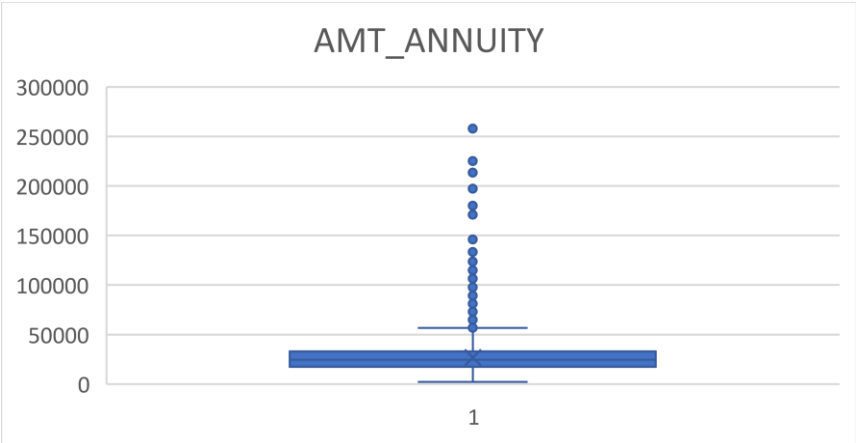


B) ***Identify Outliers in the Dataset:*** **Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.**

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

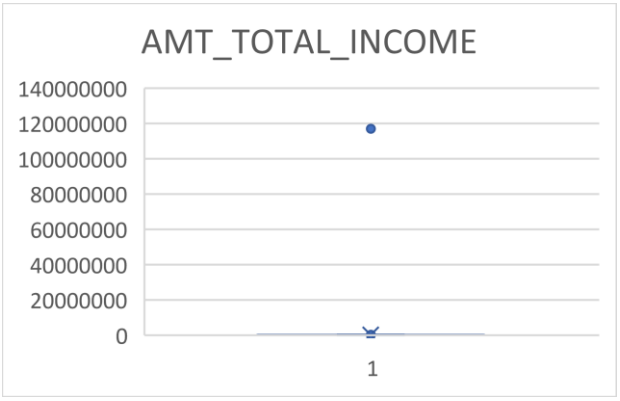
Application data:

- We replace the outlier that is greater than 250000 in the 'AMT_ANNUITY' column with the median value of the column.

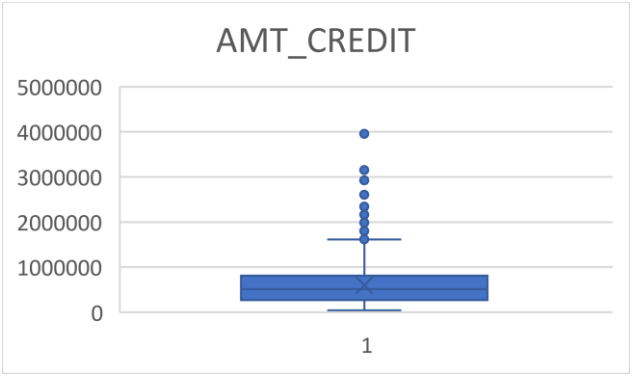


- We do not remove/modify the outliers in the columns ‘AMT_INCOME_TOTAL’ and ‘AMT_CREDIT’ as they vary from person to person.

Quartiles of AMT_INCOME_TOTAL	
min	25650
25%	112500
50%	145800
75%	202500
max	117000000

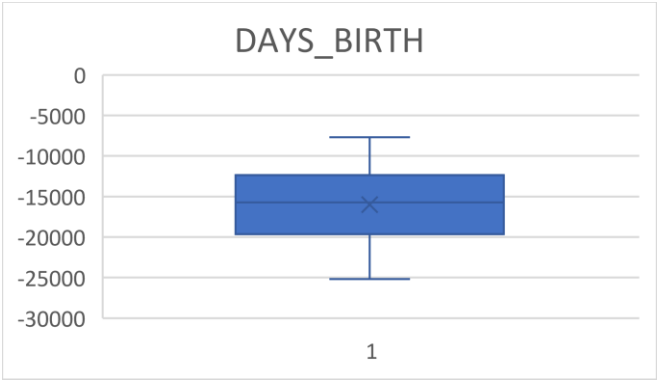


Quartiles of AMT_CREDIT	
min	45000
25%	270000
50%	514777.5
75%	808650
max	4050000

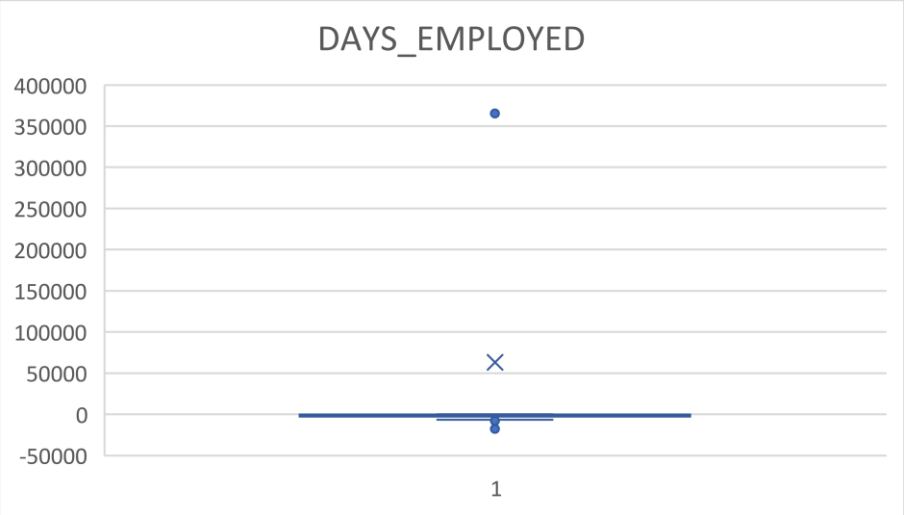


- We can see that the ‘DAYS_BIRTH’ column is well distributed and doesn’t have any outliers.

Quartiles of DAYS_BIRTH	
min	-25184
25%	-19644
50%	-15731
75%	-12378.5
max	-7680

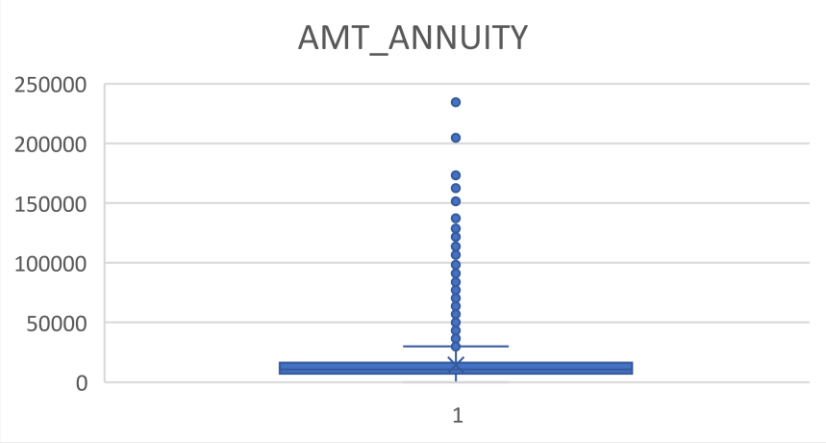


- There exists one outlier in the column ‘DAYS_EMPLOYED’ which we replace with the median of the column.



Previous application data:

- We replace the outliers greater than 200000 in the 'AMT_ANNUITY' column with the median



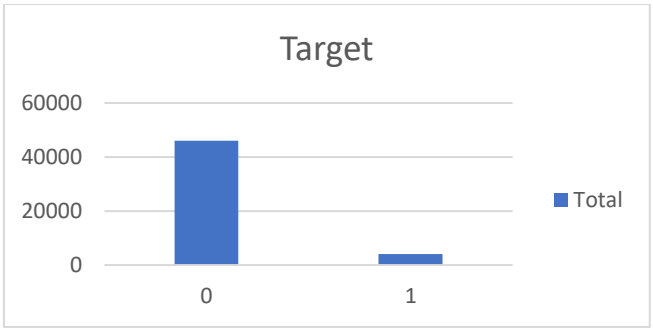
C) ***Analyze Data Imbalance:*** **Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.**

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

Application data:

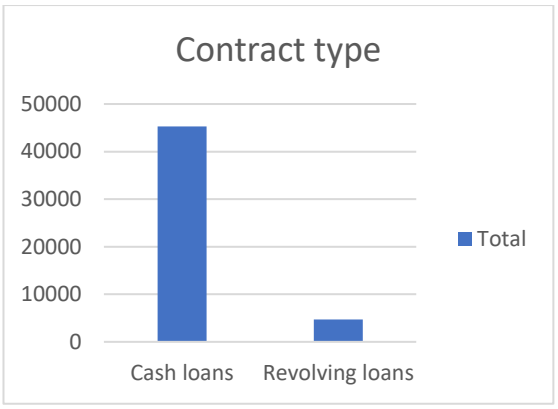
- We can see that most of them had paid instalments on time whereas few had difficulties using the pivot table we built.

Row Labels	Count of TARGET
0	45973
1	4026
Grand Total	**49999**



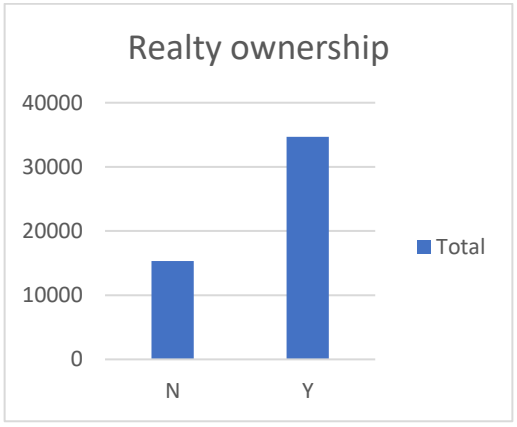
- We can observe that there is a higher number of cash loans than revolving loans among clients.

Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	45276
Revolving loans	4723
Grand Total	**49999**

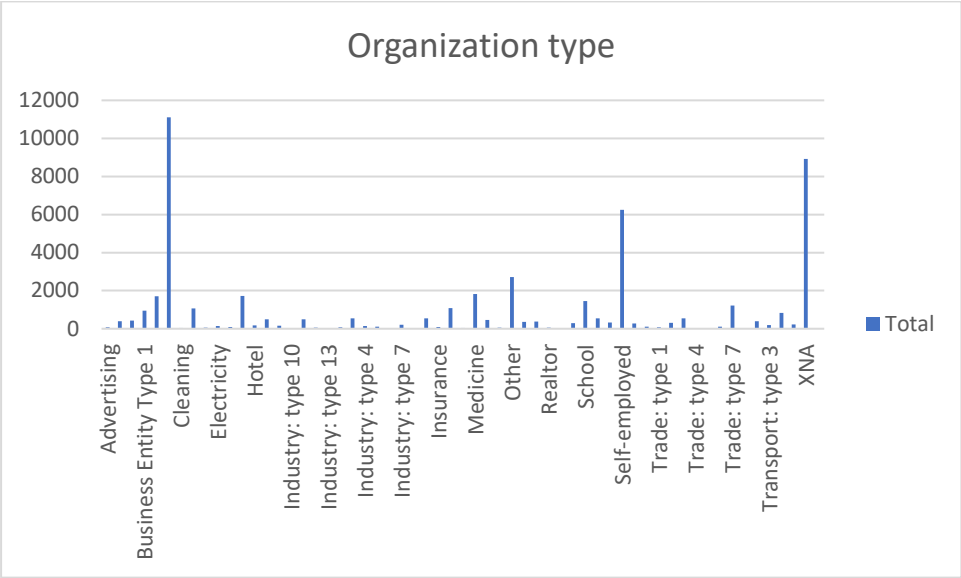


- From the below pivot table we can see that the majority are realty owners.

Row Labels	Count of FLAG_OWN_REALTY
N	15308
Y	34691
Grand Total	**49999**



- Most applicants have business entities or are self-employed.



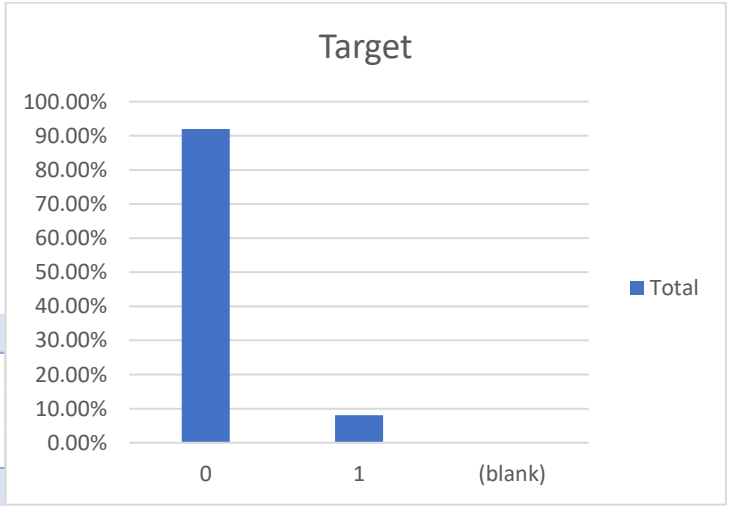
D) *Perform Univariate, Segmented Univariate, and Bivariate Analysis:* To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

***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.**

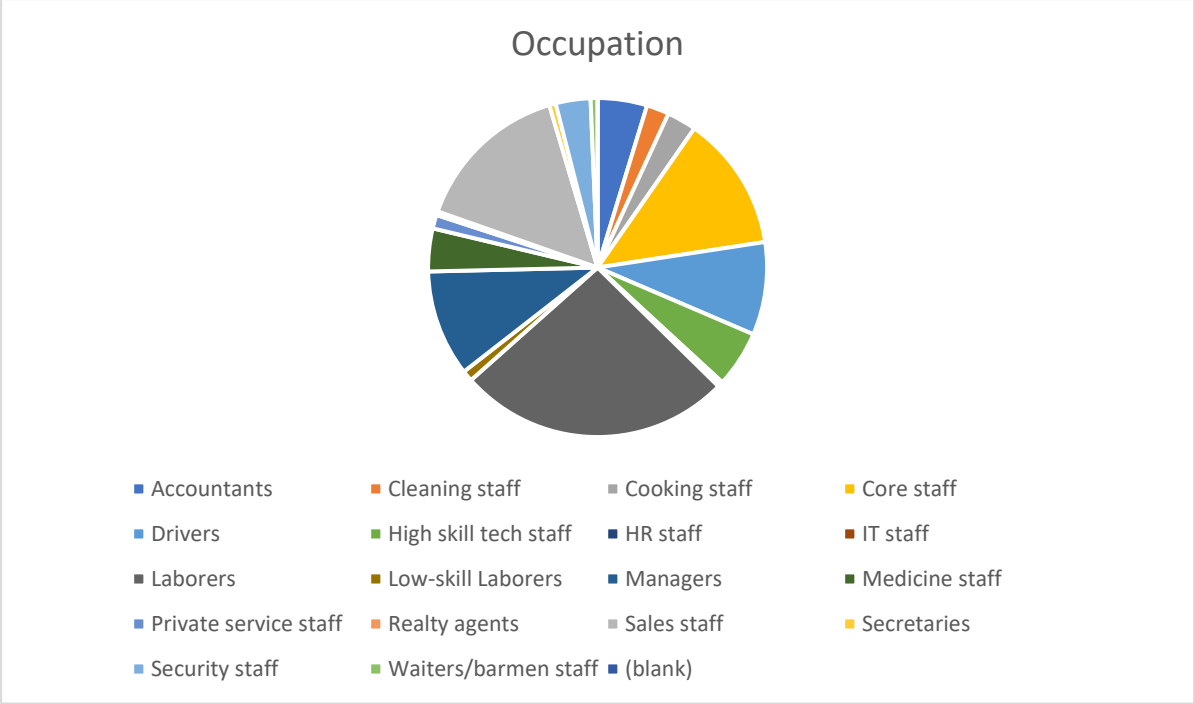
Univariate analysis:

- We can see that the percentage of loan defaulters is around 8% and non-defaulters is almost 92%.

Row Labels	Count of TARGET
0	91.95%
1	8.05%
(blank)	0.00%
Grand Total	**100.00%**

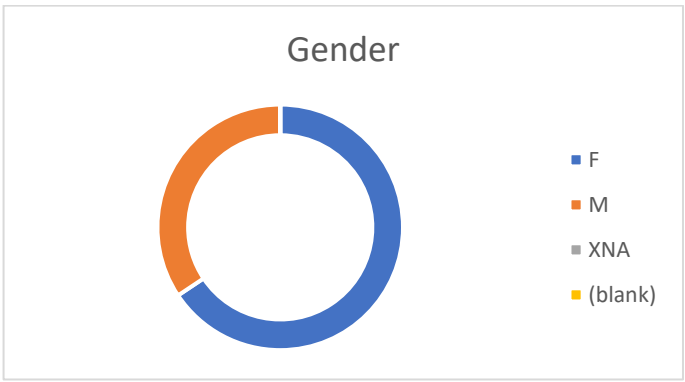


- We can see that the most of the clients are labourers followed by sales staff and core staff.



- Majority of the clients are women constituting about 65% whereas the remaining 35% are men.

Row Labels	Count of CODE_GENDER
F	65.65%
M	34.35%
XNA	0.00%
(blank)	0.00%
Grand Total	**100.00%**

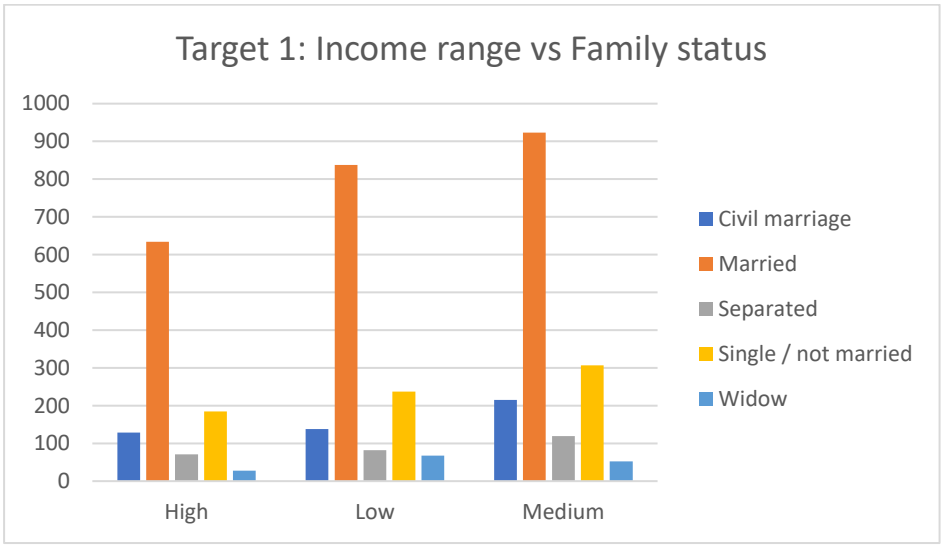


Bivariate analysis:

We build two more columns in the application data 'TOTAL_INCOME_RANGE' and 'CREDIT_RANGE' that categorizes the columns 'AMT_TOTAL_INCOME' and 'AMT_CREDIT' as High, Medium and Low.

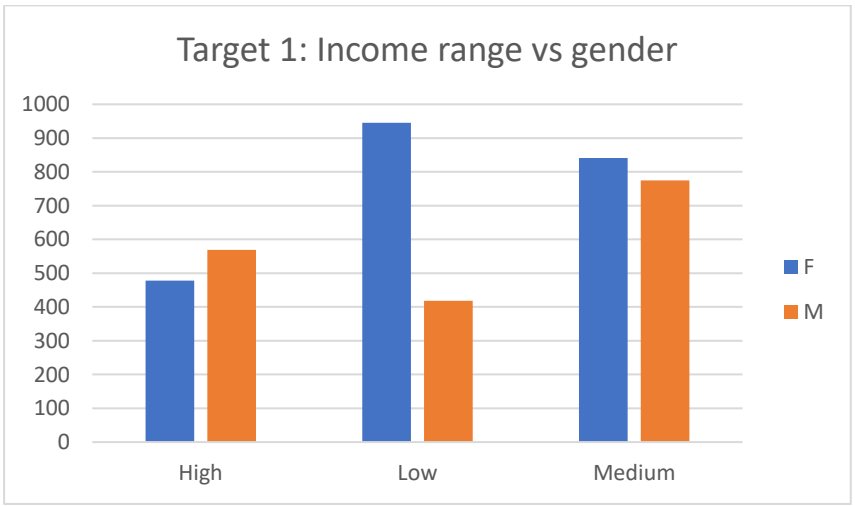
- From the below table, we can see that most clients with medium to low income and have family status as married have payment issues.

TARGET	1					
Count of NAME_FAMILY_STATUS	**Column Labels**					
Row Labels	**Civil marriage**	**Married**	**Separated**	**Single / not married**	**Widow**	**Grand Total**
High	129	634	71	185	28	1047
Low	138	838	82	237	68	1363
Medium	215	923	119	307	52	1616
Grand Total	**482**	**2395**	**272**	**729**	**148**	**4026**



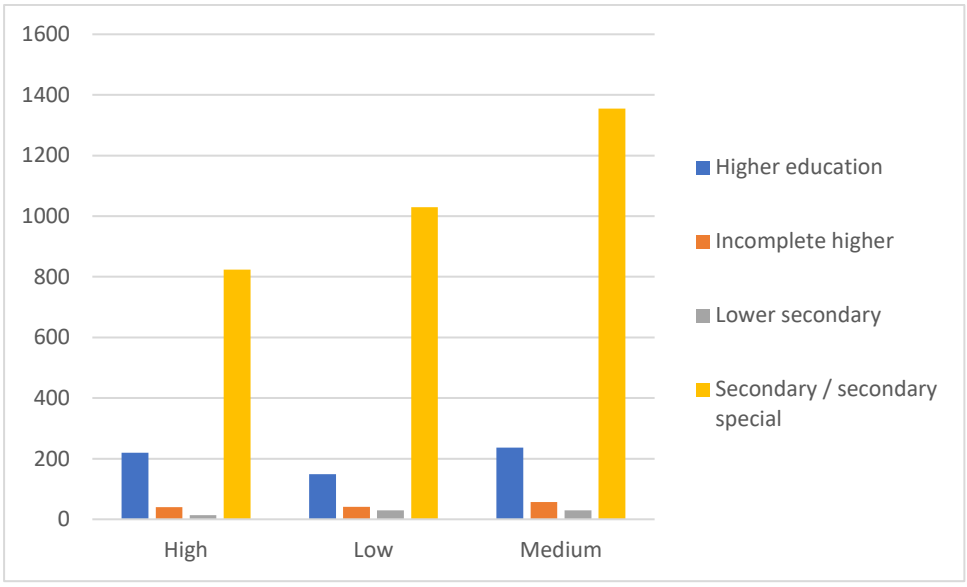
- Male clients with low income have difficulties in payments whereas in the medium income range, the gender of the client doesn't seem to have an impact.

TARGET	1		
Count of CODE_GENDER	**Column Labels**		
Row Labels	**F**	**M**	**Grand Total**
High	478	569	1047
Low	945	418	1363
Medium	841	775	1616
Grand Total	**2264**	**1762**	**4026**



- Clients with medium credit range and an education type of Secondary / Secondary special are those with maximum payment issues.

TARGET	1				
Count of NAME_EDUCATION_TYPE	**Column Labels**				
Row Labels	**Higher education**	**Incomplete higher**	**Lower secondary**	**Secondary / secondary special**	**Grand Total**
High	220	40	14	824	1098
Low	149	41	30	1030	1250
Medium	237	57	29	1355	1678
Grand Total	**606**	**138**	**73**	**3209**	**4026**



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.**

We calculate correlation coefficients, correlation matrix, and plot scatter plots between variables and the target variable.

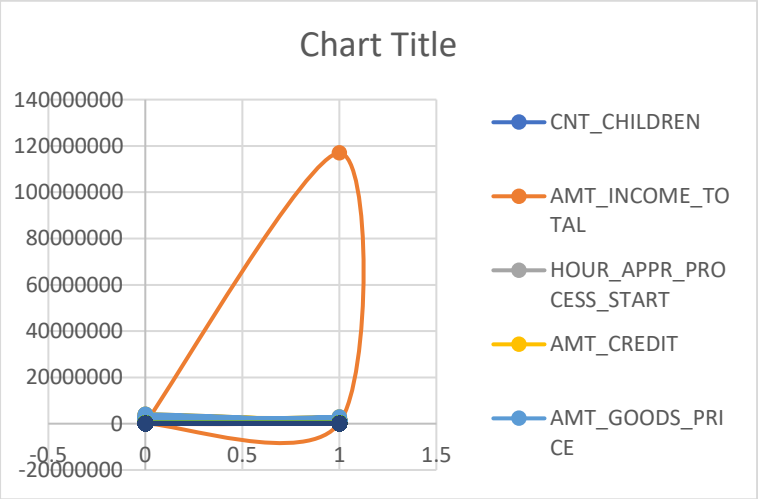
Correlation coefficients:

COLUMN AGAINST TARGET	CORRELATION COEFFICIEN
CNT_CHILDREN	0.026363931
AMT_INCOME_TOTAL	0.010893745
HOUR_APPR_PROCESS_START	-0.032036463
AMT_CREDIT	-0.032428347
DAYS_EMPLOYED	-0.040281269
AMT_GOODS_PRICE	-0.04127611
EXT_SOURCE_2	-0.158424274

Correlation Matrix:

	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	HOUR_APPR_PROCESS_START	AMT_CREDIT	AMT_GOODS_PRICE	DAYS_EMPLOYED	EXT_SOURCE_2
TARGET	1							
CNT_CHILDREN	0.026363931	1						
AMT_INCOME_TOTAL	0.010893745	0.009588558	1					
HOUR_APPR_PROCESS_START	-0.032036463	-0.006253862	0.01846417	1				
AMT_CREDIT	-0.032428347	0.00497156	0.069315897	0.056676981	1			
AMT_GOODS_PRICE	-0.04127611	0.000232954	0.069891714	0.065891636	0.986704386	1		
DAYS_EMPLOYED	-0.040281269	-0.239673381	-0.031603988	-0.08796449	-0.070416099	-0.067738221	1	
EXT_SOURCE_2	-0.158424274	-0.017641055	0.019517645	0.157147521	0.138125321	0.146862491	-0.026153993	1

Graph:



LINKS TO EXCEL SHEETS:

Application data:

https://docs.google.com/spreadsheets/d/125iMJEo3NYenwISJD4wYxd2SwFnQ8Rq3/edit?usp=sharing&ouid=101623027215720977577&rtpof=true&sd=true

Previous application data:

https://docs.google.com/spreadsheets/d/1STEkGcRsrvmF6QCR9-31VTK-goaNv-PE/edit?usp=sharing&ouid=101623027215720977577&rtpof=true&sd=true

RESULT:

Hence, we have implemented all the tasks given as a part of the Bank Loan Case Study project.