Trainity Project 6

Project Description

The project involves analyzing a loan application dataset for a finance company specializing in lending various types of loans to urban customers. The primary objectives are to identify patterns that indicate if a customer is likely to have difficulty paying their installments and to understand the key factors behind loan default. The task was to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

Approach

Approach for the project :

1) Handling Missing Data -

- Functions (e.g., COUNT, ISBLANK) were used to identify missing data for each variable. Columns with more than 30% blank cells were deleted. For the remaining columns, missing values were imputed using appropriate methods.
- Bar Chart was used for showing the proportion of blanks of each variable.

2) Identifying Outliers -

- Functions like QUARTILE and IQR were applied to calculate quartiles and the interquartile range (IQR) for numerical variables.
- Conditional formatting was used to identify and visually highlight potential outliers. Scatter plot was used for visualization.

3) Data Imbalance -

- COUNTIF function was used to calculate the proportions of each class in the target variable.
- The imbalance ratio was calculated to assess data imbalance. Bar chart was used to show the data imbalance.

4) Univariate, Segmented Univariate, and Bivariate Analysis -

- Univariate analysis was done using education type, family status variables. The frequency of each category was determined. Bar charts were created to visualize the distributions of variables.
- For Segmented Univariate Analysis, data was filtered into subsets based on different scenarios, such as customers with payment difficulties. Tree map was created to visualize the analysis.
- Bivariate Analysis:Pivot tables were created to analyze the relationships between variables(gender, income type) and the target variable. Column charts were generated to visualize these relationships.

5) Correlations -

- The data is filtered for targets 0 and 1.
- The correlation between different variables for specific target variables(0 and 1) was calculated using the CORREL function.

Tech-Stack Used

Microsoft Excel - Excel was the primary software used for data manipulation, analysis, and visualization. Its functions and features were employed for calculations, data cleaning, and creating visualizations. Pivot tables were used for univariate, bivariate analysis. COUNTA, COUNTIF, AVERAGE, MEDIAN, QUARTILE, CORREL were used to perform calculations. Various charts like Column chart, Pie chart, Scatter plot, etc were used for visualization.

Insights

Key findings and patterns discovered:

A. Identify Missing Data and Deal with it Appropriately:

- Missing data was identified in the loan application dataset. The not required columns had more blank cells overall and were removed.
- Visualizing the proportion of missing values for each variable showed that the important/necessary columns had no or very less blanks.

B. Identify Outliers in the Dataset:

- Outliers were detected in Amt_Income_Total variable using the interquartile range (IQR) method.
- Outliers can significantly impact data analysis. Outliers were found mostly more than the upper bound. Scatter plot showed the distribution of data and outliers present.

C. Analyze Data Imbalance:

- Data imbalance was identified, where the majority of loan applicants were all other cases, while a smaller proportion were clients with payment difficulties.
- The bar chart showed that target 1 had only 4026 applicants while target 0 had 45973 applicants.

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

- Univariate analysis helped in understanding the distribution of individual variables, such as education_type and family_status.
- Segmented univariate analysis revealed differences in variable distributions for loan type, target(customers with payment difficulties and all other cases).
- Bivariate analysis showed the relationship between variables(gender,income type) and the target variable.
- The use of bar charts enhanced the visualization of variable distributions, while pivot tables were used in segmented and bivariate analysis.

E. Identify Top Correlations for Different Scenarios:

- For Target 0 - Top correlation was between AMT_INCOME_TOTAL & AMT_CREDIT
- For Target 1 - Top correlation was between DAYS_REGISTRATION(years) & DAYS_EMPLOYED(years).

Result

Through this project, I learned how to effectively manage missing data, identify outliers, and deal with data imbalances, ensuring the dataset's quality and integrity. I also learned how to explore relationships between variables using pivot tables and scatter plots for insights. It helped me gain practical experience in data analysis and equipped me with valuable skills applicable in the finance industry.

Excel sheet link - Excel_P6.xlsx

Data Analytics Tasks:

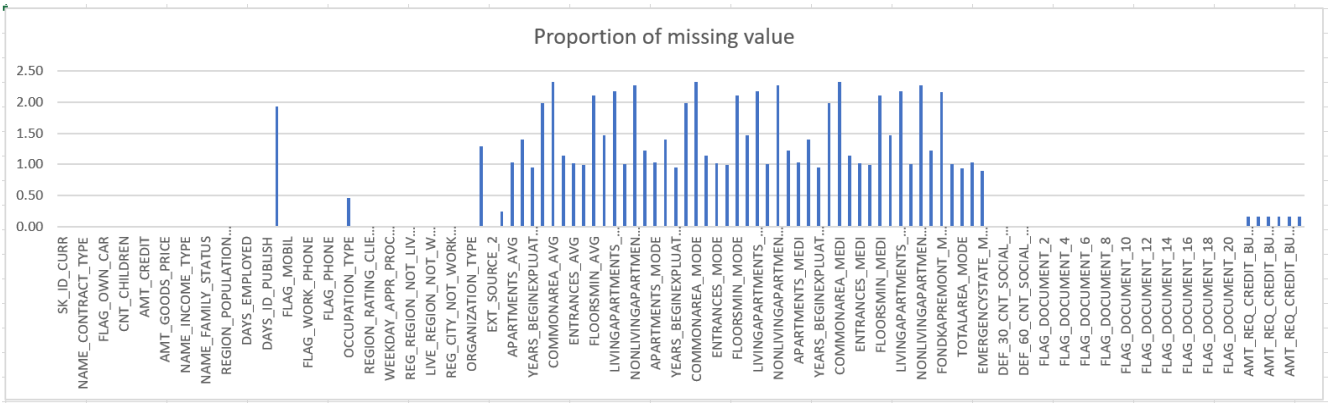
1) Task: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

The count, blanks, percent of blanks, proportion of missing values were calculated for each variable. The variables/columns having blanks more than 30%(highlighted in red) were removed.

Count	49999	49999	49999	49999	49999	49999	49999	49999	49999	49998	49961	49807
Blanks	0	0	0	0	0	0	0	0	0	1	38	192
Blanks %	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.08%	0.38%
Proportion n	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

49999	49999	49999	49999	49999	49999	17049	49999	49999	49999	49999	49999	49999	34345	49998
0	0	0	0	0	0	32950	0	0	0	0	0	0	15654	1
0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	65.90%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	31.31%	0.00%
0.00	0.00	0.00	0.00	0.00	0.00	1.93	0.00	0.00	0.00	0.00	0.00	0.00	0.46	0.00

Graph for proportion of missing values of each variable :



Final dataset :

SK_ID_CU	TARGET	NAME_CO	CODE_GE	FLAG_OW	FLAG_OW	CNT_CHIL	AMT_INC	AMT_CRE	AMT_ANI	AMT_GO	NAME_TY	NAME_IN	NAME_ED	NAME_FA	NAME_H
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompar	Working	Secondary /	Single / not r	House / a
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servan	Higher educ	Married	House / a
100004	0	Revolving lo	M	Y	Y	0	67500	135000	6750	135000	Unaccompar	Working	Secondary /	Single / not r	House / a
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompar	Working	Secondary /	Civil marriag	House / a
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompar	Working	Secondary /	Single / not r	House / a
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, part	State servan	Secondary /	Married	House / a
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompar	Commercial	Higher educ	Married	House / a
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompar	State servan	Higher educ	Married	House / a
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondary /	Married	House / a
100012	0	Revolving lo	M	N	Y	0	135000	405000	20250	405000	Unaccompar	Working	Secondary /	Single / not r	House / a
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompar	Working	Higher educ	Married	House / a
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children	Pensioner	Secondary /	Married	House / a
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompar	Working	Secondary /	Married	House / a
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompar	Working	Secondary /	Married	House / a
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompar	Working	Secondary /	Married	House / a

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

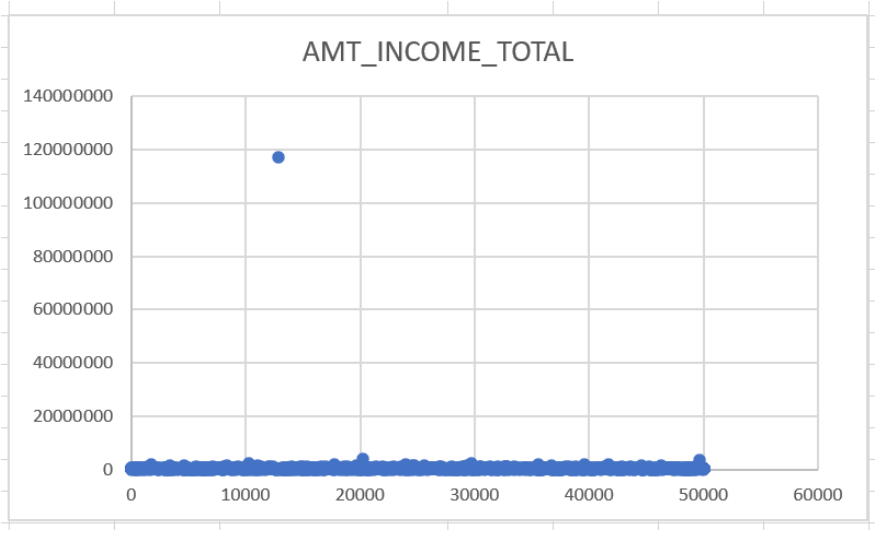
The AMT_INCOME_TOTAL is used and the quartile, inter quartile range, lower and upper bounds are calculated.

Quartile 1	112500
Quartile 3	202500
Interquartile range	90000
Lower bound	-22500
Upper bound	337500

Conditional formatting is used to highlight the outliers in the data.

130500
360000
54000
540000
76500
225000
81000
180000
67500
81000
360000
540000
180000

Scatter plot is created to visualize the distribution of numerical variables and outliers.



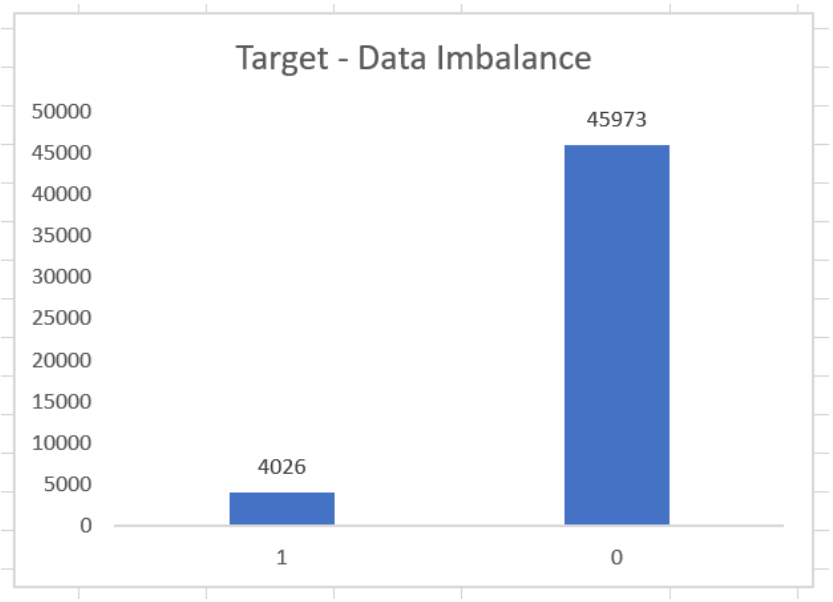
3) Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

The target variable is used to find data imbalance and count, imbalance ratio is calculated.

Target	Count
1	4026
0	45973

Imbalance Ratio
0.087573141

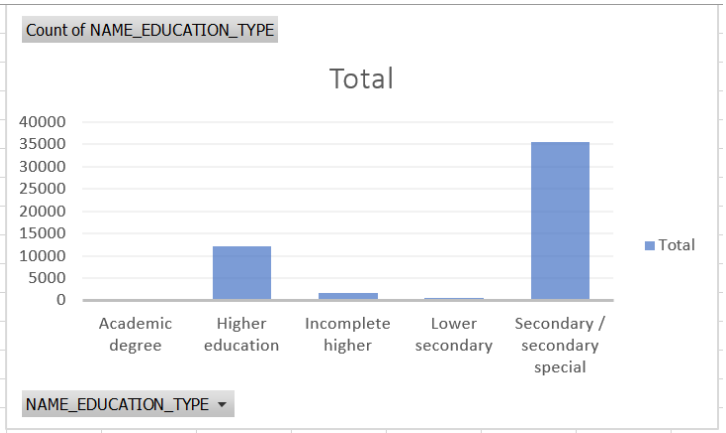
The bar chart is made to visualize the distribution of the target variable and highlight the imbalance.



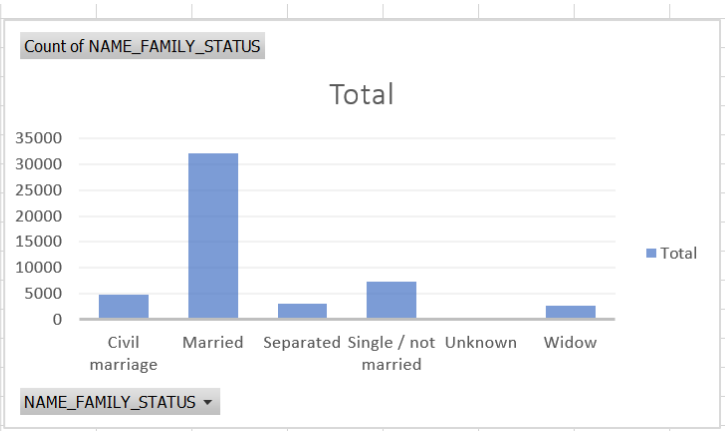
4) Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

a) Univariate analysis for Education Type and Family Status

Row Labels	Count of NAME_EDUCATION_TYPE
Academic degree	20
Higher education	12167
Incomplete higher	1620
Lower secondary	620
Secondary / seconda	35572
Grand Total	**49999**

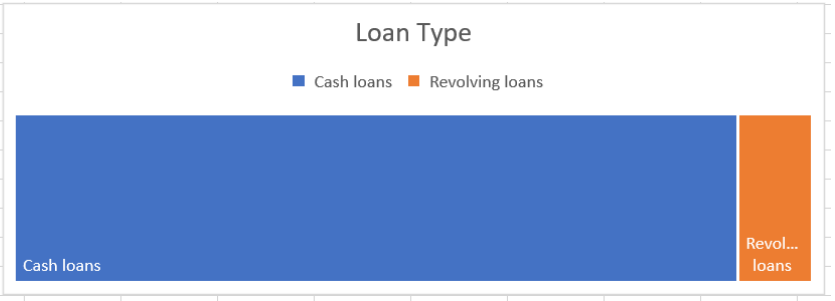


Row Labels	Count of NAME_FAMILY_STATUS
Civil marriage	4859
Married	32094
Separated	3142
Single / not married	7306
Unknown	1
Widow	2597
Grand Total	**49999**



b) Segmented univariate analysis for Contract type and Target

Cash loans	45276	90.55381
Revolving loans	4723	9.446189

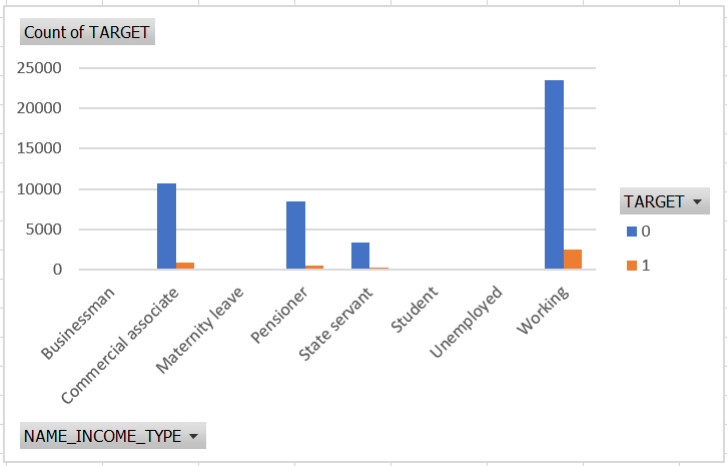


Target	
Target 0	45973
Target 1	4026

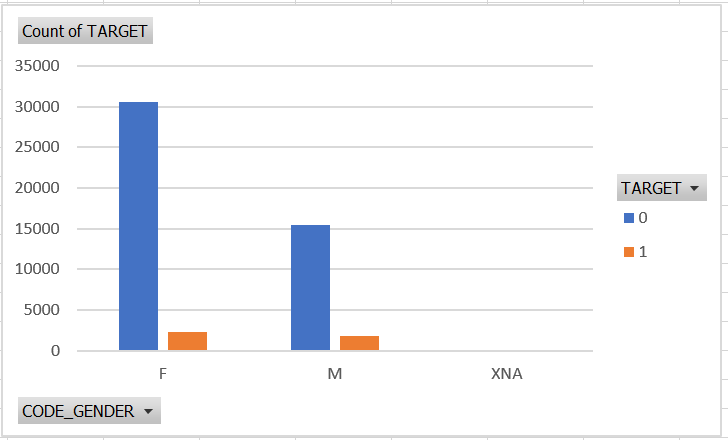
c) Bivariate analysis between target and income type and between target and gender.

Bivariate

Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
Businessman	2		2
Commercial assoc	10679	864	11543
Maternity leave	1		1
Pensioner	8419	501	8920
State servant	3314	198	3512
Student	5		5
Unemployed	4	2	6
Working	23549	2461	26010
Grand Total	**45973**	**4026**	**49999**



Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
F	30559	2264	32823
M	15412	1762	17174
XNA	2		2
Grand Total	**45973**	**4026**	**49999**



5) 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.

Segmented data for Target 0

Target 0					
CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	DAYS_EMPLOYED(Years)	DAYS_REGISTRATION(Years)	REGION_RATING_CLIENT
0	270000	1293502.5	3.3	3.2	1
0	67500	135000	0.6	11.7	2
0	135000	312682.5	8.3	26.9	2
0	121500	513000	8.3	11.8	2
0	99000	490495.5	4.4	13.6	2
1	171000	1560726	8.6	3.3	2
0	360000	1530000	1.2	12.6	3
0	112500	1019610	1000.7	20.3	2
0	135000	405000	5.5	39.6	2
1	112500	652500	1.9	12.1	2
0	38419.155	148365	1000.7	14.4	2
0	67500	80865	7.4	0.9	2
1	225000	918468	8.3	1.8	2
0	189000	773680.5	0.6	1.7	2
0	157500	299772	3.2	9.6	3
0	108000	509602.5	3.6	17.5	2
1	81000	270000	0.5	11.4	2

Correlations between variables for Target 0

CORRELATION - TARGET 0						
CNT_CHILDREN	1.00	0.04	0.01	-0.25	-0.18	0.02
AMT_INCOME_TOTAL	0.04	1.00	0.38	-0.16	-0.07	-0.21
AMT_CREDIT	0.01	0.38	1.00	-0.07	-0.01	-0.10
DAYS_EMPLOYED(Years)	-0.25	-0.16	-0.07	1.00	0.21	0.04
DAYS_REGISTRATION(Years)	-0.18	-0.07	-0.01	0.21	1.00	-0.08
REGION_RATING_CLIENT	0.02	-0.21	-0.10	0.04	-0.08	1.00
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	DAYS_EMPLOYED(Years)	DAYS_REGISTRATION(Yea	REGION_RATING_CLIENT

Top Indicators - Target 0	Correalation
AMT_INCOME_TOTAL & AMT_CREDIT	0.38
DAYS_REGISTRATION(Years) & DAYS_EMPLOYED(Years)	0.21

Segmented data for Target 1

Target 1					
CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	DAYS_EMPLOYED(Years)	DAYS_REGISTRATION(Years)	REGION_RATING_CLIENT
0	202500	406597.5	1.7	10.0	2
0	112500	979992	7.2	18.0	3
0	202500	1193580	3.5	3.2	2
0	135000	288873	9.9	0.1	3
0	81000	252000	1000.7	14.8	2
0	315000	953460	5.5	13.2	2
1	157500	723996	0.7	1.1	2
0	292500	675000	0.5	14.4	2
0	157500	245619	21.0	2.1	2
0	111915	225000	0.4	7.0	2
3	180000	540000	2.8	2.1	2
1	202500	436032	0.3	4.7	1
0	135000	495216	0.4	18.5	2
0	157500	1710000	25.4	2.2	2
0	73341	135000	0.4	8.0	2
1	121500	263686.5	1.2	9.8	2

Correlations between variables for Target 1

CORRELATION - TARGET 1						
CNT_CHILDREN	1.00	0.01	0.01	-0.19	-0.15	0.06
AMT_INCOME_TOTAL	0.01	1.00	0.02	-0.01	0.01	-0.01
AMT_CREDIT	0.01	0.02	1.00	0.02	0.04	-0.05
DAYS_EMPLOYED(Years)	-0.19	-0.01	0.02	1.00	0.19	-0.01
DAYS_REGISTRATION(Years)	-0.15	0.01	0.04	0.19	1.00	-0.12
REGION_RATING_CLIENT	0.06	-0.01	-0.05	-0.01	-0.12	1.00
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	DAYS_EMPLOYED(Years)	DAYS_REGISTRATION(Years)	REGION_RATING_CLIENT

Top Indicators - Target 1	Correalation
DAYS_REGISTRATION(Years) & DAYS_EMPLOYED(Years)	0.19
REGION_RATING_CLIENT & CNT_CHILDREN	0.06