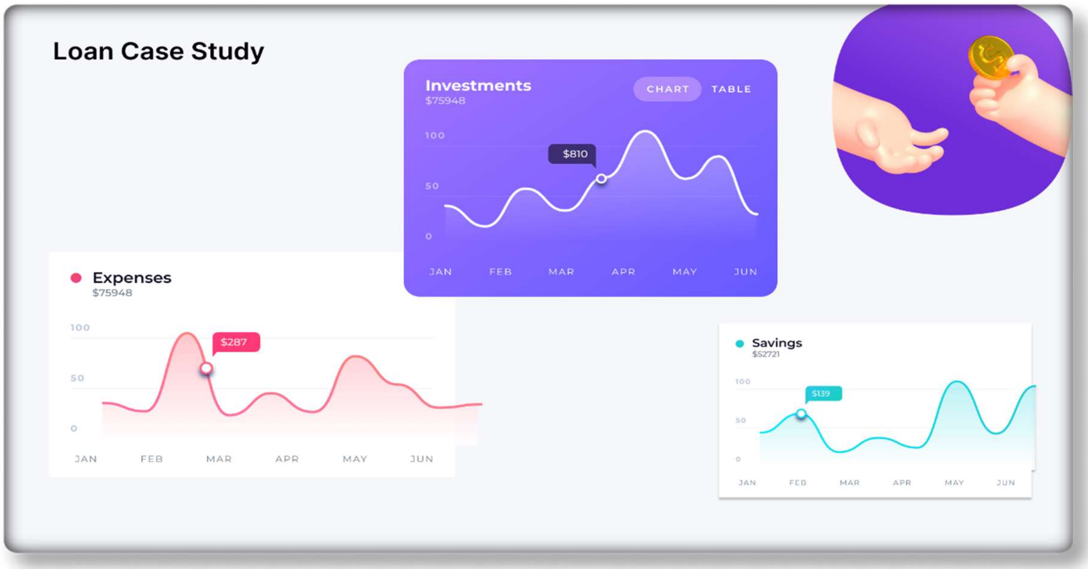
A

Live- Project Report

Bank Loan Case Study



Presented to – Trainity

Submitted by-

Anirudh Chaudhary

Project Description:

Conduct Exploratory Data Analysis (EDA) as a data analyst at a finance company specializing in lending loans to urban customers. The company faces a challenge of customers with insufficient credit history exploiting the system and defaulting on loans. The goal is to use EDA to analyze patterns in the data and ensure that qualified applicants are not rejected.

The dataset includes information on loan applications, categorized into customers with payment difficulties (late payments on installments) and those without payment issues. Four possible outcomes of a loan application are Approved, Canceled, Refused, and Unused Offer.

The business objectives are to identify patterns indicating if a customer will struggle with installment payments. This information can be used to make decisions such as denying loans, reducing loan amounts, or lending at higher interest rates to risky applicants. The company aims to understand key factors behind loan defaults for better decision-making in loan approval.

The context of risk analytics in banking and financial services is crucial to understanding the project, including the significance of various variables in predicting and mitigating loan default risks.

Approach:

To achieve the project goals, a structured analysis approach was adopted:

1. **Identify Missing Data and Handle Appropriately:**
   - Missing data was identified for all variables. Strategies like imputation (mean/median for numerical data and mode for categorical data) were applied where applicable. Irrelevant missing data was excluded.
   - **Visualization:** Bar charts were created to visualize the proportion of missing data across variables.

2. **Detect Outliers:**
   - Numerical variables were assessed for outliers using statistical methods like the Interquartile Range (IQR). Outliers were flagged for further investigation.
   - **Visualization:** Box plots and scatter plots highlighted the distribution of numerical variables and detected outliers.

3. **Analyze Data Imbalance:**
   - The dataset was evaluated for class imbalance, especially for the target variable (loan default status). The proportion of classes was calculated, and significant imbalances were noted.
   - **Visualization:** Pie charts and bar charts were used to depict the imbalance in target variable classes.

4. **Conduct Univariate, Segmented Univariate, and Bivariate Analysis:**
   - **Univariate Analysis:** Assessed the distribution of individual variables such as loan amount, income, and credit history.
   - **Segmented Univariate Analysis:** Compared distributions across different loan scenarios (e.g., approved vs. refused loans).
   - **Bivariate Analysis:** Explored relationships between key variables (e.g., income vs. loan amount) and their influence on loan default.
   - **Visualization:** Histograms, stacked bar charts, and scatter plots were created for comparisons.

5. **Identify Top Correlations for Different Scenarios:**
   - Correlation coefficients between variables and the target variable were calculated for each scenario. This helped identify key indicators of loan default.
   - **Visualization:** Correlation matrices and heatmaps highlighted the strongest relationships in each segment.

Tech-Stack Used

The primary tool for this project was **Microsoft Excel (Version 2022)** due to its robust analytical and visualization capabilities. Key features used:

1. **Data Cleaning Tools:** Functions like COUNT, ISBLANK, and IF for missing data handling.
2. **Statistical Analysis:** AVERAGE, MEDIAN, QUARTILE, and CORREL functions for descriptive and correlation analysis.
3. **Visualization Tools:** Bar charts, box plots, scatter plots, and correlation heatmaps.
4. **Conditional Formatting:** Highlighted outliers, data imbalances, and trends in the dataset.
5. **Pivot Tables:** Summarized segmented and bivariate insights.

❶

Data-Cleaning

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.

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.

EXCEL:

#TASK 1 (Data-Sorting, Cleaning)

To handle missing data I did the following:

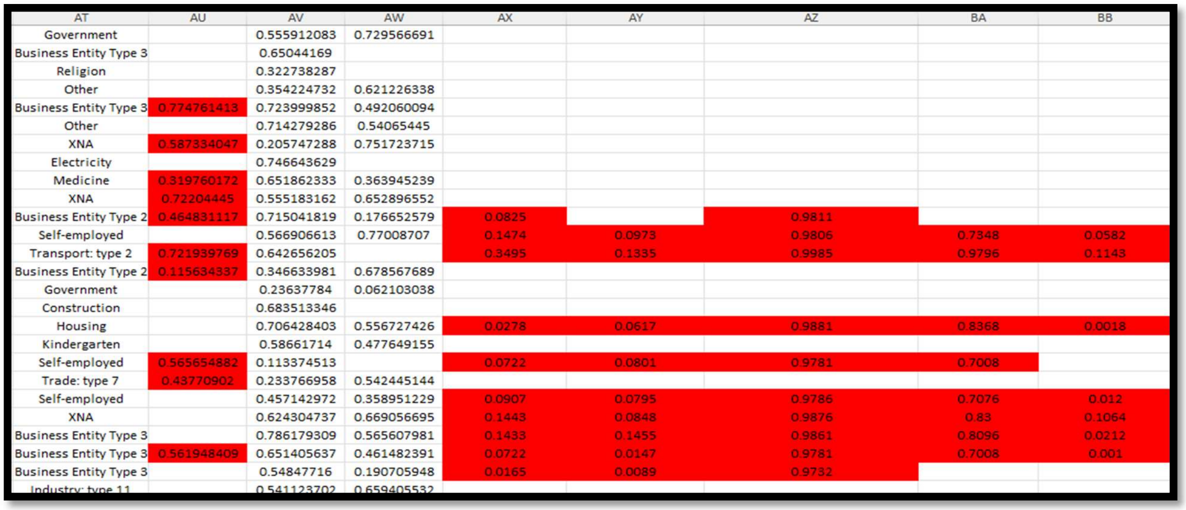
1.calculated the percentage of blank cells in a new row (50001) using the function

= 1-B2/$B$2

2.With the help of conditional formatting identified and deleted all the columns which had a percentage of missing cells more than 40%.

3. Filled all the missing cells with the median (row 50002) of that particular column (median as mean will be ineffective because of outliers).

After the cleaning the data was left with 73 columns and 50002 rows with 0 blank cells and no duplicates.



AT	AU	AV	AW	AX	AY	AZ	BA	BB
Government		0.555912083	0.729566691					
Business Entity Type 3		0.65044169						
Religion		0.322738287						
Other		0.354224732	0.621226338					
Business Entity Type 3	0.774761413	0.723999852	0.492060094					
Other		0.714279286	0.54065445					
XNA	0.587334047	0.205747288	0.751723715					
Electricity		0.746643629						
Medicine	0.319760172	0.651862333	0.363945239					
XNA	0.72204445	0.555183162	0.652896552					
Business Entity Type 2	0.464831117	0.715041819	0.176652579	0.0825		0.9811		
Self-employed		0.566906613	0.77008707	0.1474	0.0973	0.9806	0.7348	0.0582
Transport: type 2	0.721939769	0.642656205		0.3495	0.1335	0.9985	0.9796	0.1143
Business Entity Type 2	0.115634337	0.346633981	0.678567689					
Government		0.23637784	0.062103038					
Construction		0.683513346						
Housing		0.706428403	0.556727426	0.0278	0.0617	0.9881	0.8368	0.0018
Kindergarten		0.58661714	0.477649155					
Self-employed	0.565654882	0.113374513		0.0722	0.0801	0.9781	0.7008	
Trade: type 7	0.43770902	0.233766958	0.542445144					
Self-employed		0.457142972	0.358951229	0.0907	0.0795	0.9786	0.7076	0.012
XNA		0.624304737	0.669056695	0.1443	0.0848	0.9876	0.83	0.1064
Business Entity Type 3		0.786179309	0.565607981	0.1433	0.1455	0.9861	0.8096	0.0212
Business Entity Type 3	0.561948409	0.651405637	0.461482391	0.0722	0.0147	0.9781	0.7008	0.001
Business Entity Type 3		0.54847716	0.190705948	0.0165	0.0089	0.9732		
Industry: type 11		0.541123702	0.659405532					

2

Outlier Analysis

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.

STEPS:

#TASK 2 (Identify Outliers)

To identify the outliers the quartile function was used as the following:

1.calculated the first and third quartile using the function =QUARTILE(ARRAY,1) and =

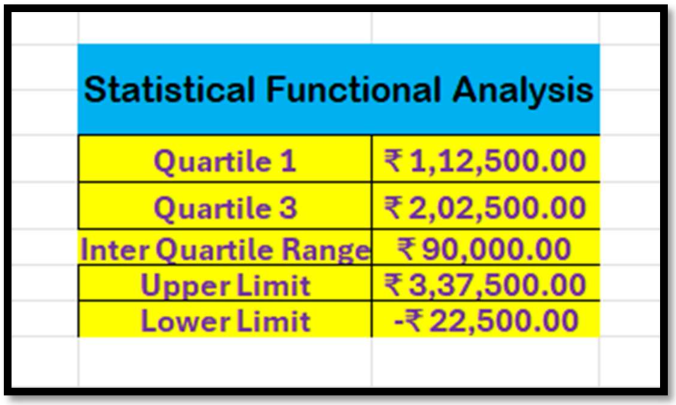
QUARTILE(ARRAY,3)

2. Calculated the inter quartile range(IQR) by subtracting the first quarter from the third quarter
3. Calculated the lower and upper bound using the formula lower bound = Q1 - 1.5*IQR, upper bound = Q3 + 1.5*IQR.

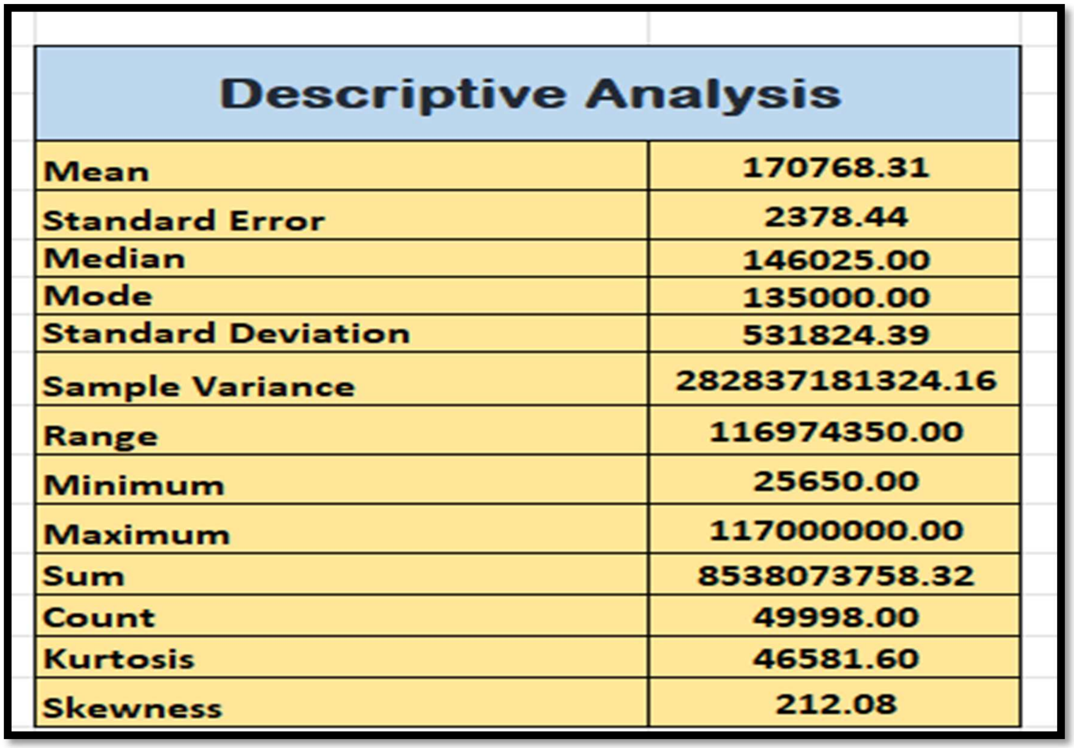
4. Another column was created to check if the values in the previous column lie between the range of upper bound and lower bound which will be true and false if the value is an outlier.

Graphs: the scatter plots here are shown to visualize the outlier(took 15000 rows as excel was freezing for a large number of rows).

Insight:

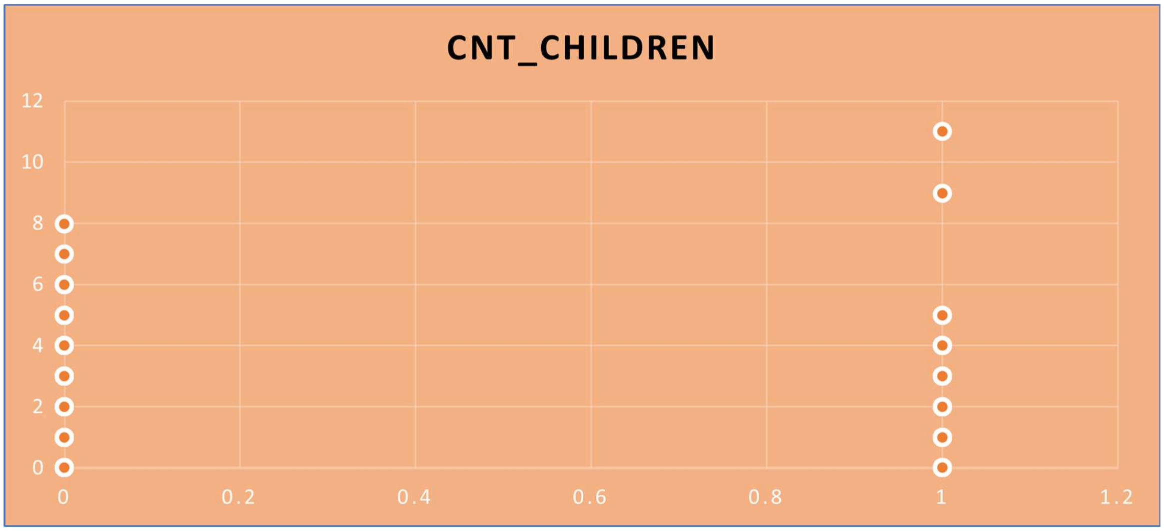


Statistical Functional Analysis	
Quartile 1	₹1,12,500.00
Quartile 3	₹2,02,500.00
Inter Quartile Range	₹90,000.00
Upper Limit	₹3,37,500.00
Lower Limit	-₹22,500.00

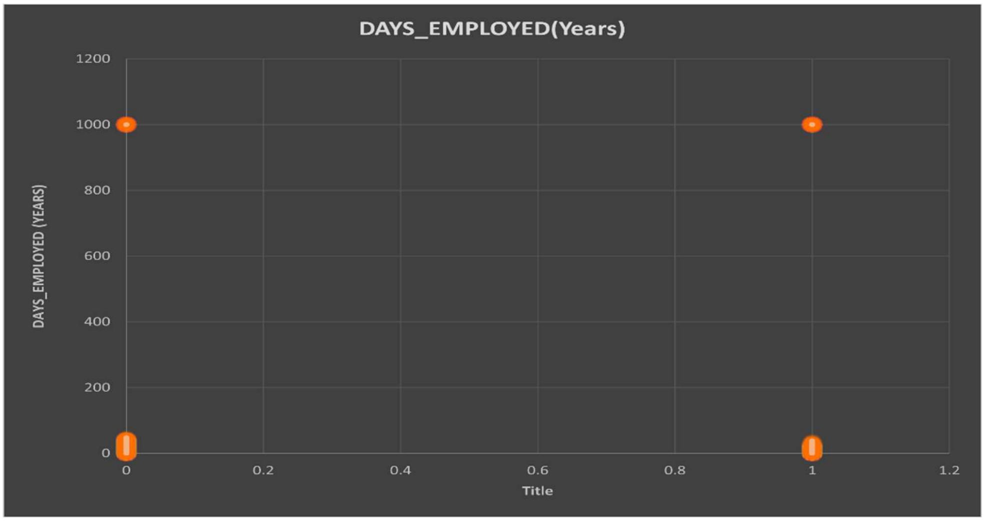


Descriptive Analysis	
Mean	170768.31
Standard Error	2378.44
Median	146025.00
Mode	135000.00
Standard Deviation	531824.39
Sample Variance	282837181324.16
Range	116974350.00
Minimum	25650.00
Maximum	117000000.00
Sum	8538073758.32
Count	49998.00
Kurtosis	46581.60
Skewness	212.08

#Outliers for CNT_CHILDREN



# Outliers for Days_Employed



3

Data-imbalance

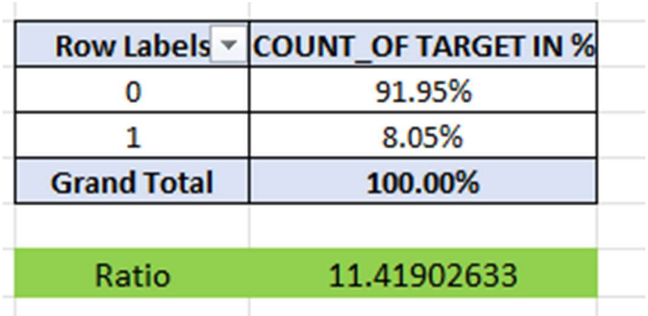
C) **Analyze Data Imbalance:** Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

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

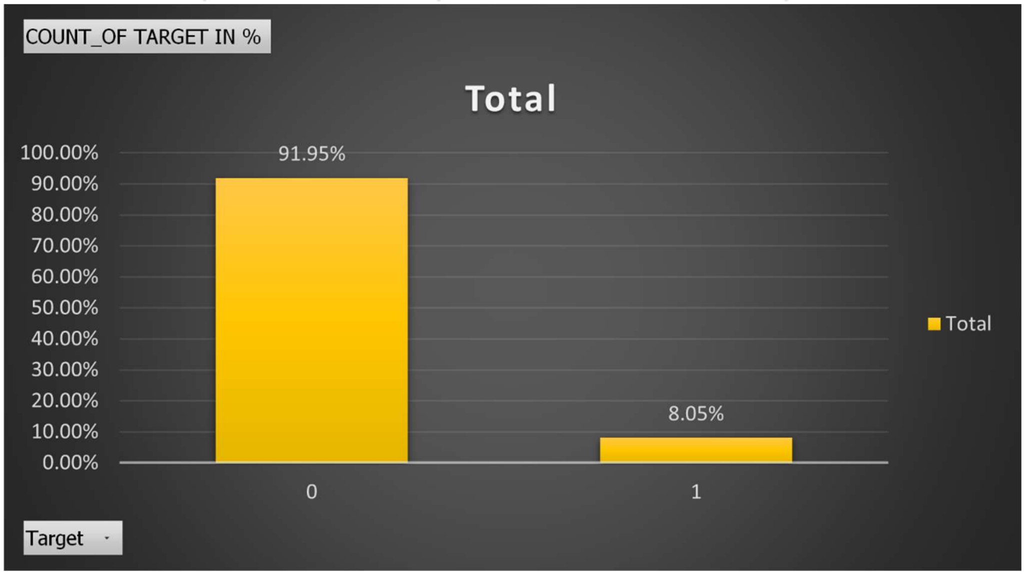
STEPS: **#Task3 Identifying Data-imbalance**

To check the data imbalance in the dataset I created different pivot tables for columns in which the data imbalance was to be checked and generated column charts for each of them in a different sheet.

EXCEL Result:



Row Labels	COUNT_OF TARGET IN %
0	91.95%
1	8.05%
Grand Total	**100.00%**
Ratio	11.41902633



Findings:

Most of the people had paid instalments on time comparatively few had difficulties.

4

Univariate Analysis

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.

2) **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.

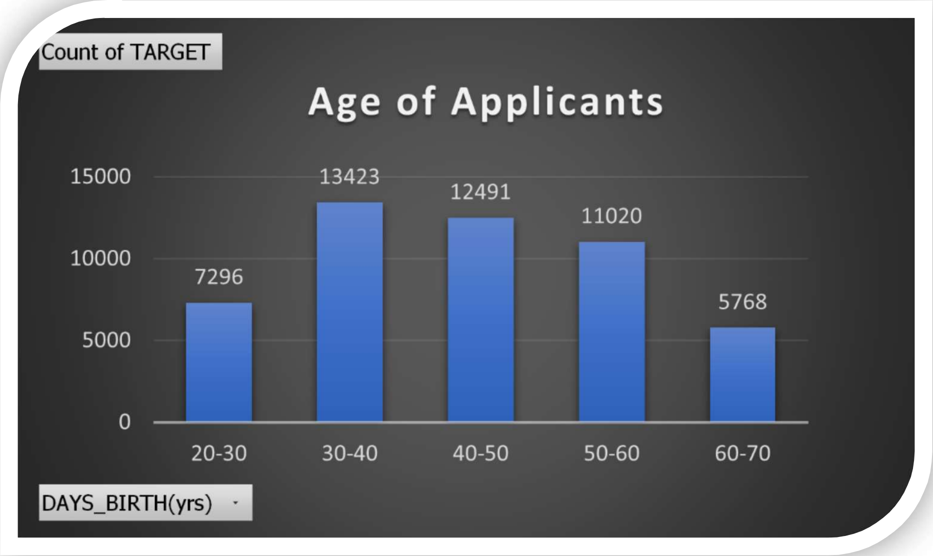
STEPS:

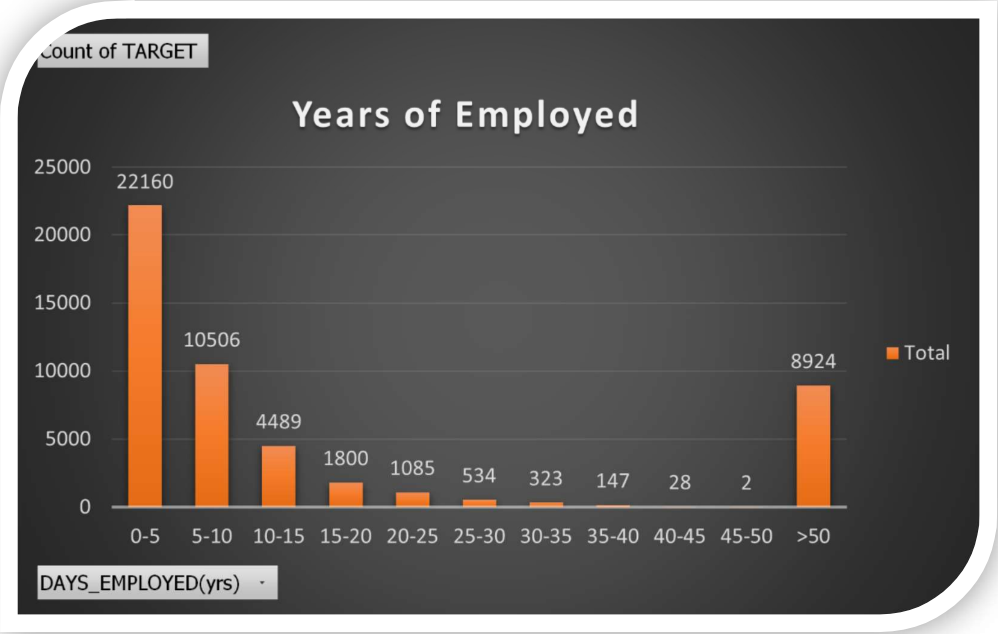
#TASK 4 (Univariate Analysis)

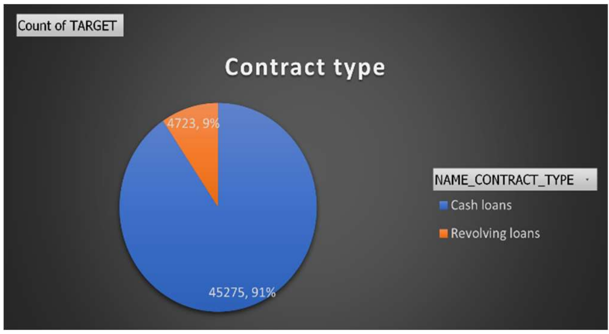
- **Univariate analysis:** In this type of analysis data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
- **Segmented univariate analysis:** segmented univariate analysis is an extension of univariate analysis as Segmented analysis here means that the data variable is analyzed in subsets(as ranges).
- I generated the frequency distribution histogram by creating classes (from max, min), bins and using data analytics option>histogram>input range,output range, chart output.

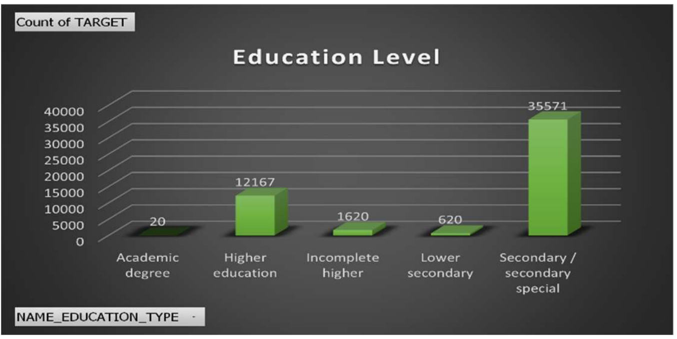
EXCEL Result:

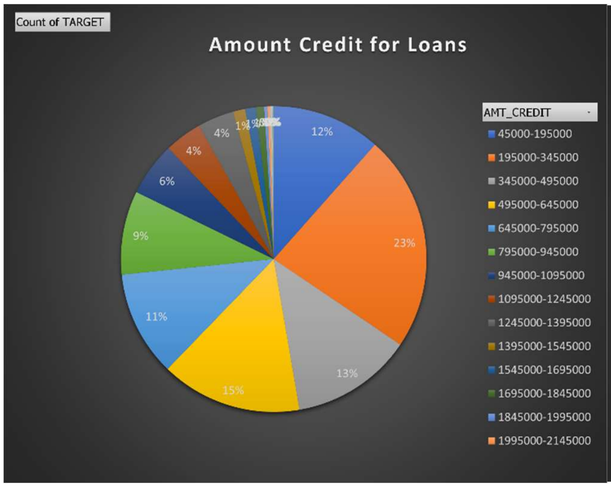
Univariate Analysis

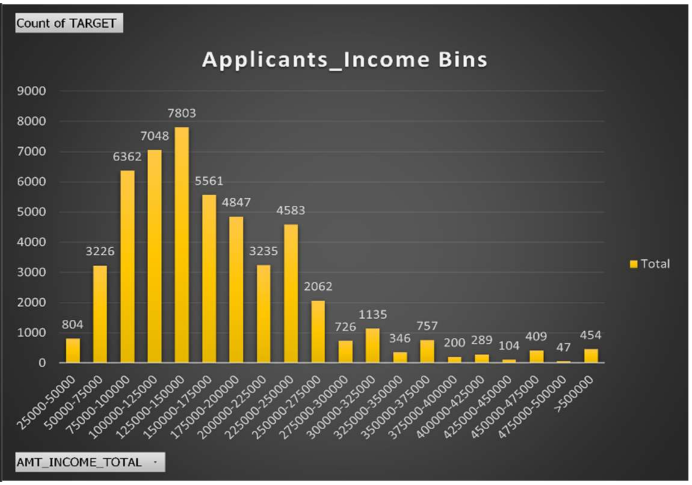


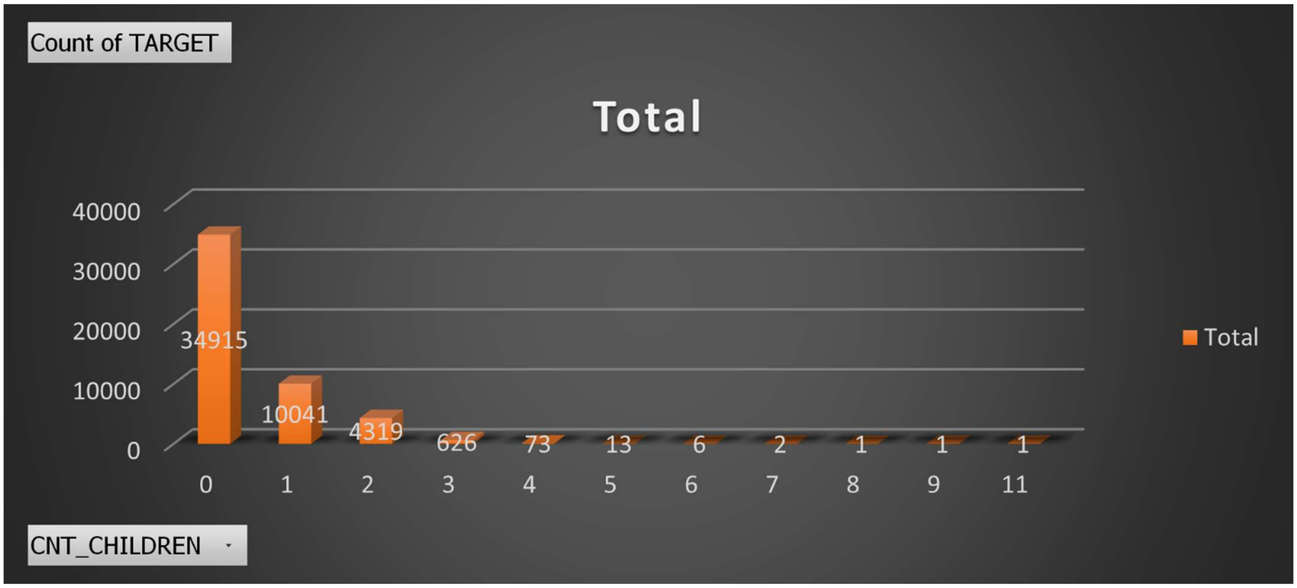


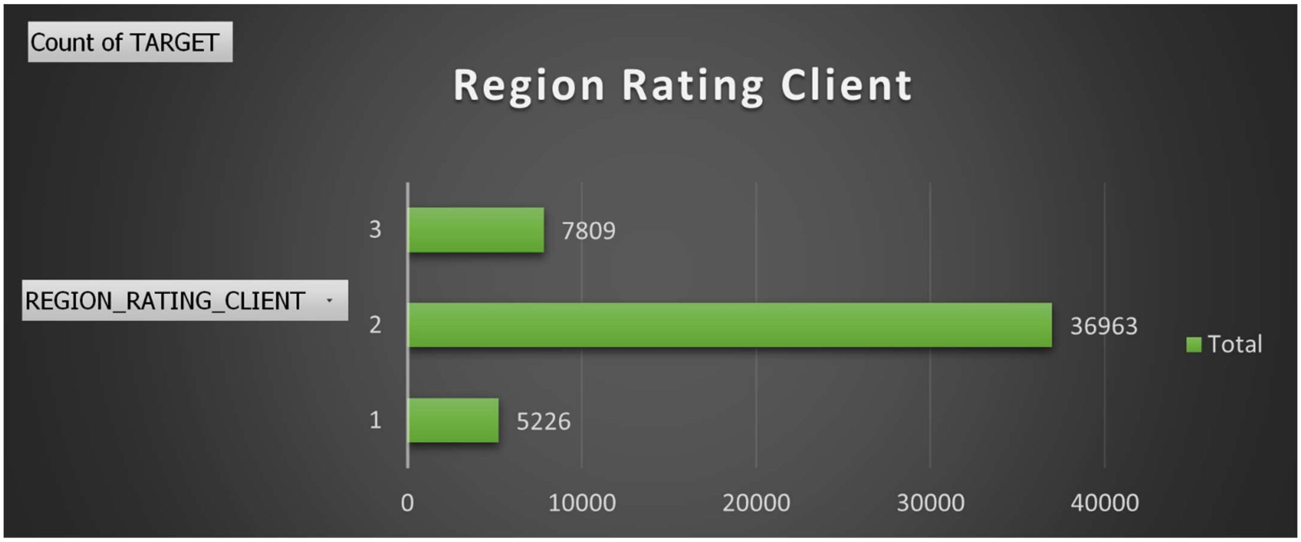

Count of TARGET
Contract type
4723, 9%
45275, 91%
NAME_CONTRACT_TYPE
Cash loans
Revolving loans

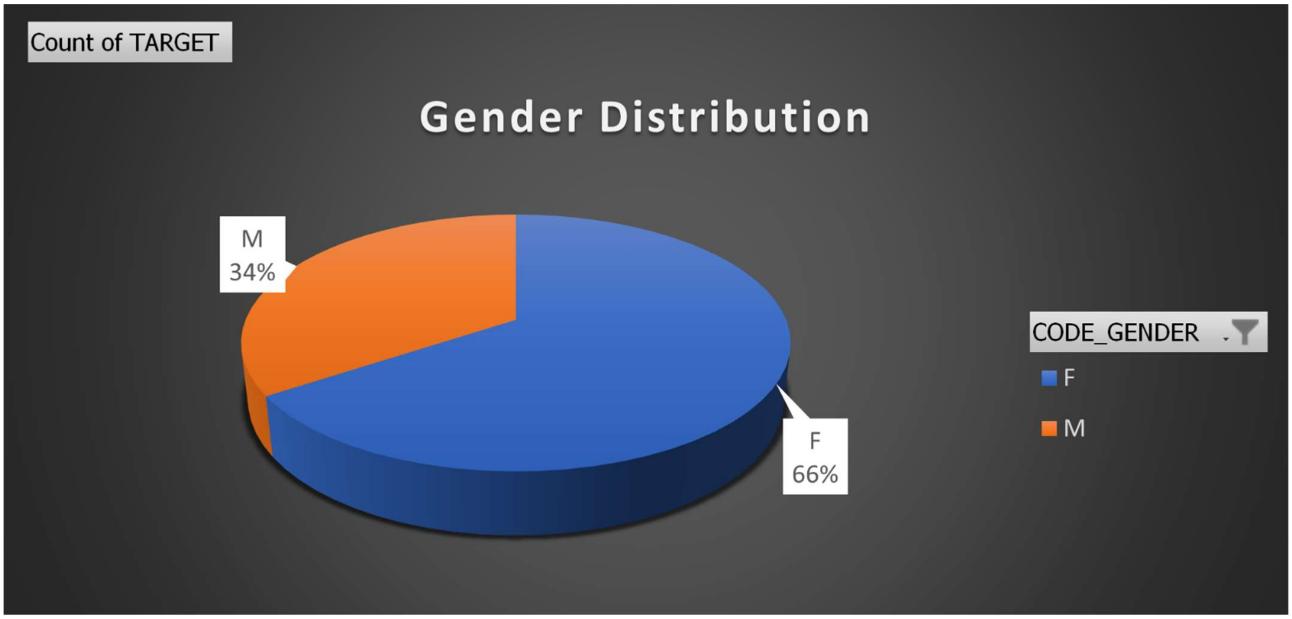

Count of TARGET
Education Level
40000
35000
30000
25000
20000
15000
10000
5000
0
20
12167
1620
620
35571
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
NAME_EDUCATION_TYPE

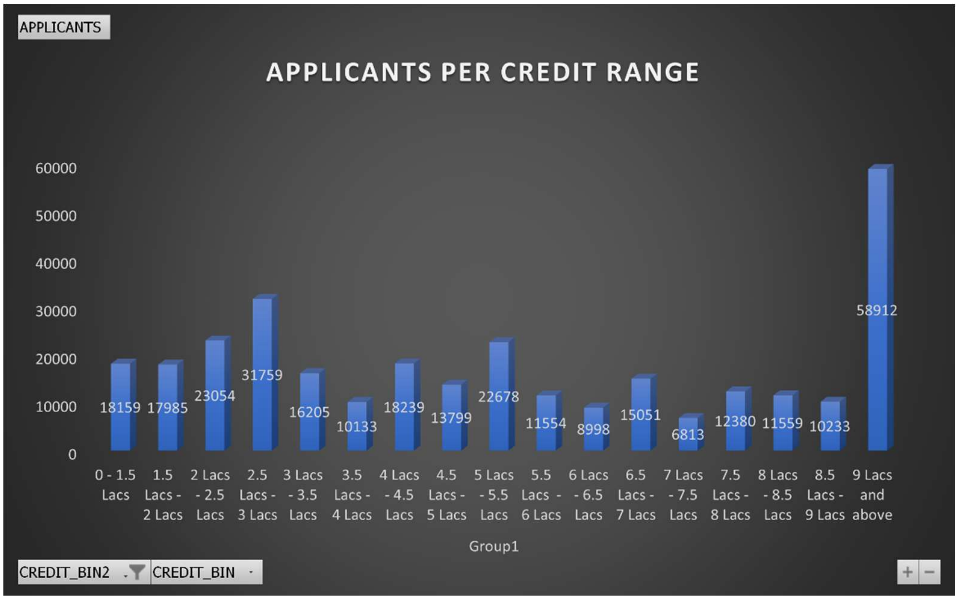

Count of TARGET
Amount Credit for Loans
12%
23%
13%
15%
11%
9%
6%
4%
4%
1%
AMT_CREDIT
45000-195000
195000-345000
345000-495000
495000-645000
645000-795000
795000-945000
945000-1095000
1095000-1245000
1245000-1395000
1395000-1545000
1545000-1695000
1695000-1845000
1845000-1995000
1995000-2145000


Count of TARGET
Applicants_Income Bins
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
804
3226
6362
7048
7803
5561
4847
3235
4583
2062
726
1135
346
757
200
289
104
409
47
454
25000-50000
50000-75000
75000-100000
100000-125000
125000-150000
150000-175000
175000-200000
200000-225000
225000-250000
250000-275000
275000-300000
300000-325000
325000-350000
350000-375000
375000-400000
400000-425000
425000-450000
450000-475000
475000-500000
>500000
Total
AMT_INCOME_TOTAL


Count of TARGET
Total
40000
30000
20000
10000
0
34915
10041
4319
626
73
13
6
2
1
1
1
0
1
2
3
4
5
6
7
8
9
11
Total
CNT_CHILDREN


Count of TARGET
Region Rating Client
REGION_RATING_CLIENT
3
2
1
7809
36963
5226
0
10000
20000
30000
40000
🟩 Total

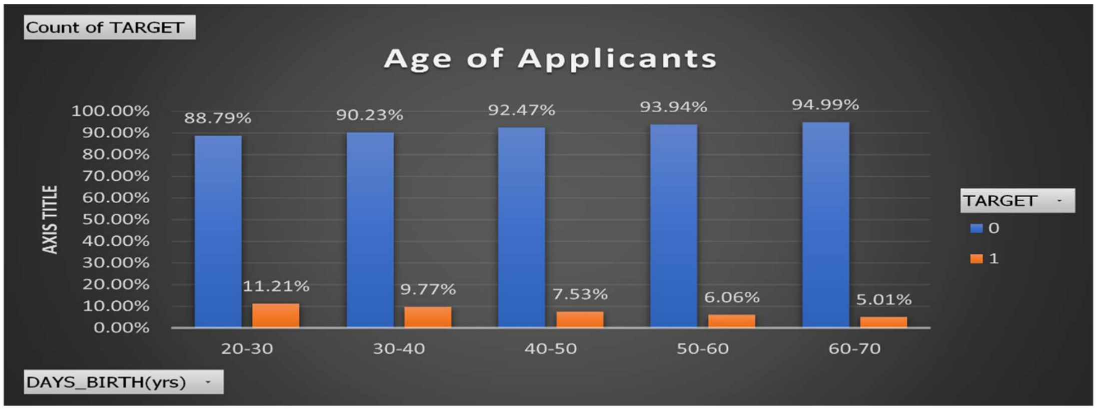

Count of TARGET
Gender Distribution
M
34%
F
66%
CODE_GENDER
🟦 F
🟧 M

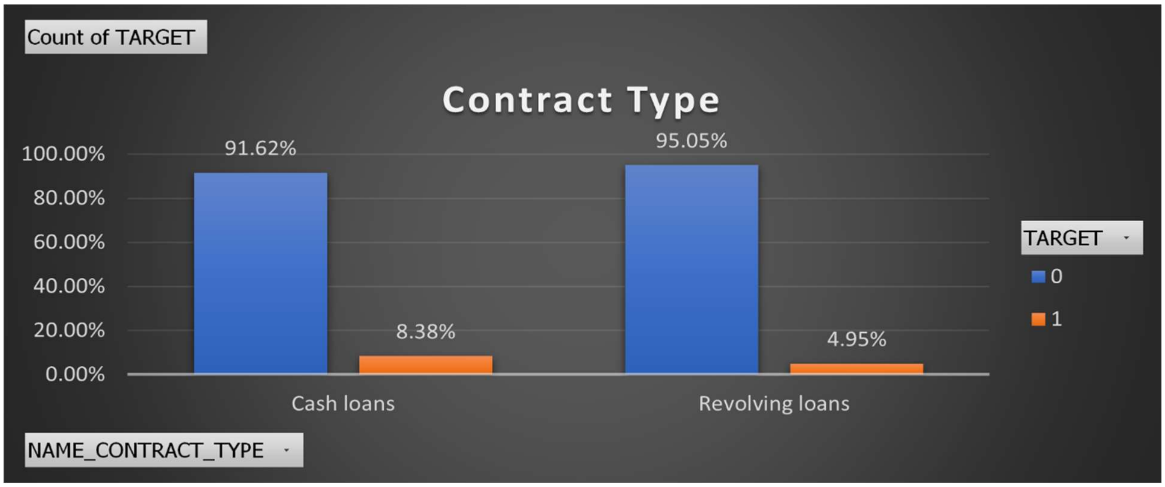

APPLICANTS
APPLICANTS PER CREDIT RANGE
60000
50000
40000
30000
20000
10000
0
18159
17985
23054
31759
16205
10133
18239
13799
22678
11554
8998
15051
6813
12380
11559
10233
58912
0 - 1.5 Lacs
1.5 Lacs - 2 Lacs
2 Lacs - 2.5 Lacs
2.5 Lacs - 3 Lacs
3 Lacs - 3.5 Lacs
3.5 Lacs - 4 Lacs
4 Lacs - 4.5 Lacs
4.5 Lacs - 5 Lacs
5 Lacs - 5.5 Lacs
5.5 Lacs - 6 Lacs
6 Lacs - 6.5 Lacs
6.5 Lacs - 7 Lacs
7 Lacs - 7.5 Lacs
7.5 Lacs - 8 Lacs
8 Lacs - 8.5 Lacs
8.5 Lacs - 9 Lacs
9 Lacs and above
Group1
CREDIT_BIN2
CREDIT_BIN

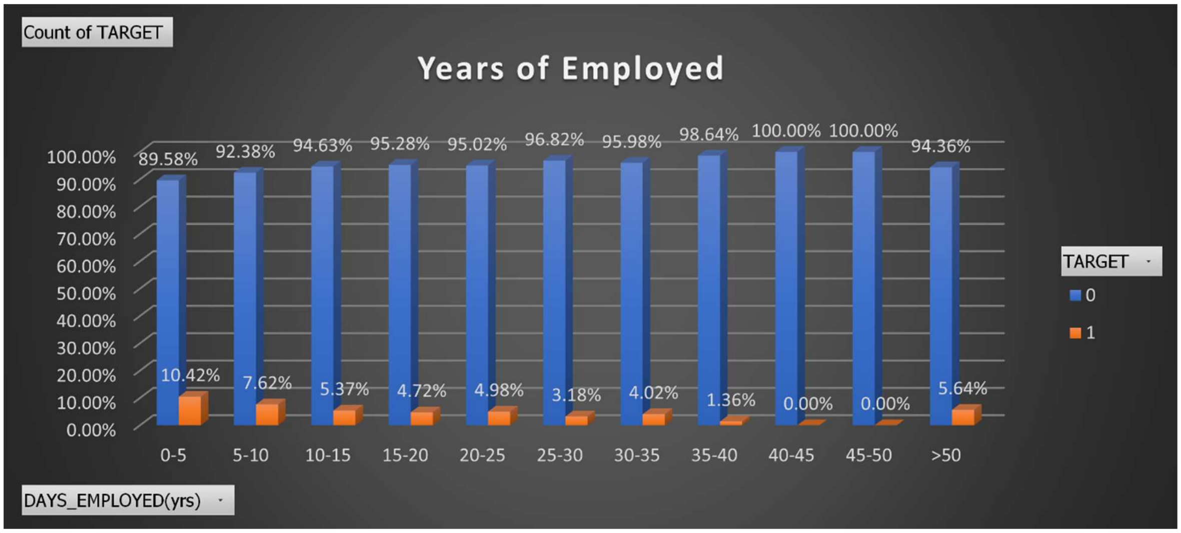
Insight:

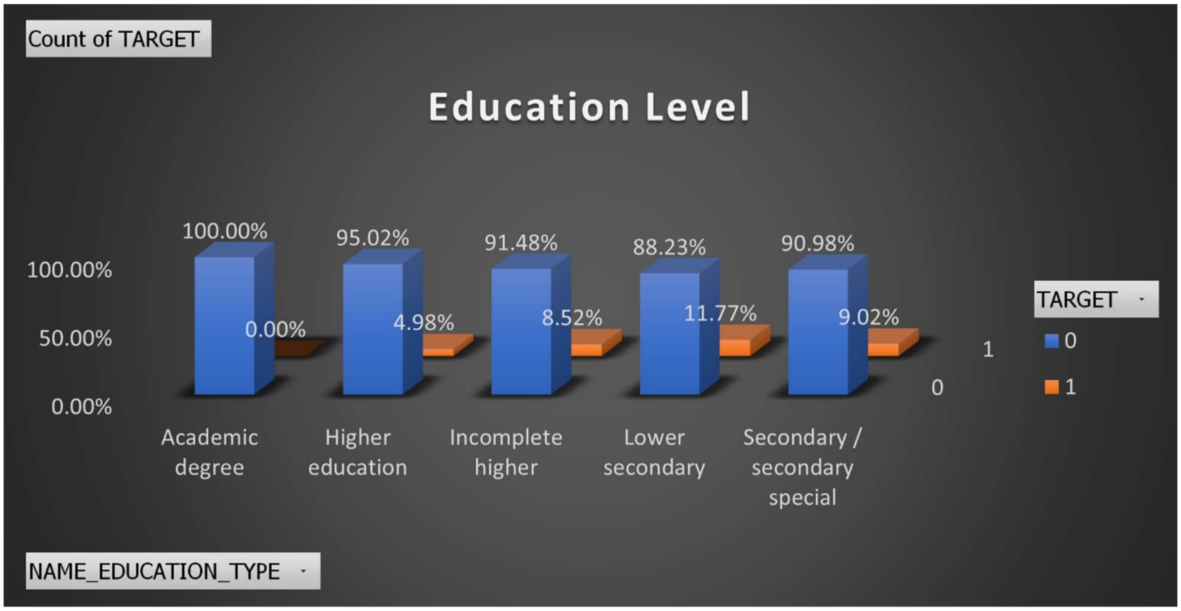
- **a.** Most of the loans opted are CASH -LOANS
- **b.** Max Income bins fall under 1,25,000 – 1,50,000
- **c.** Most of the loans opted by employees fall under 0-5 years of employment.
- **d.** Most of the loans opted by Females

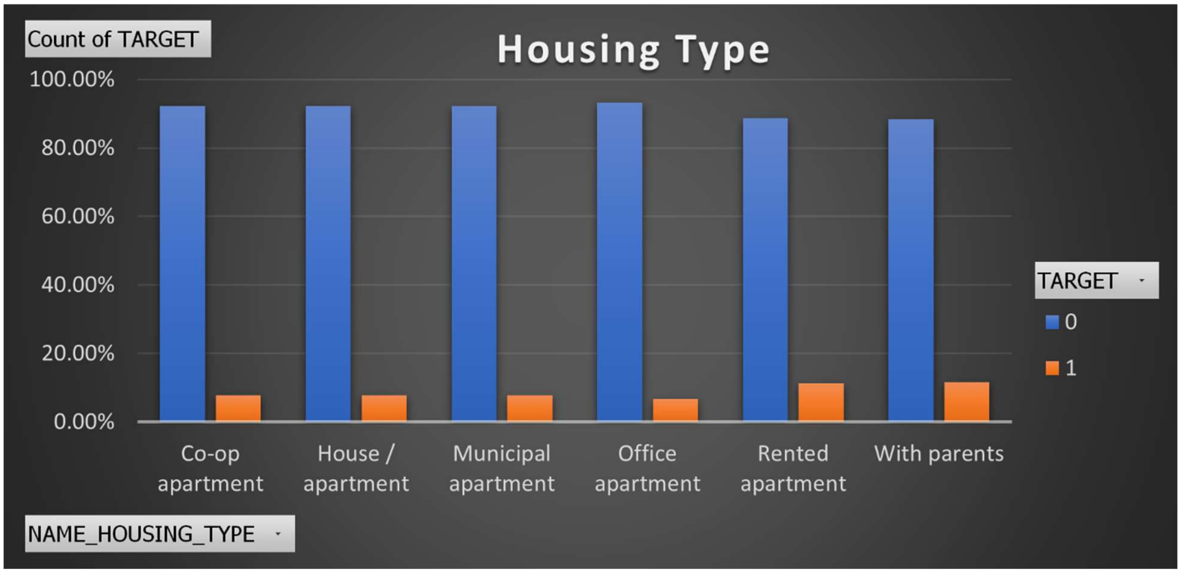
Segmented Variation Analysis

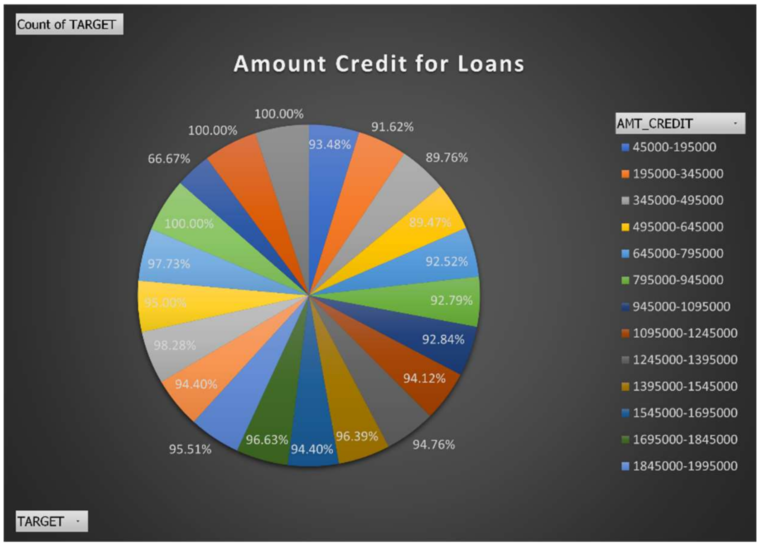


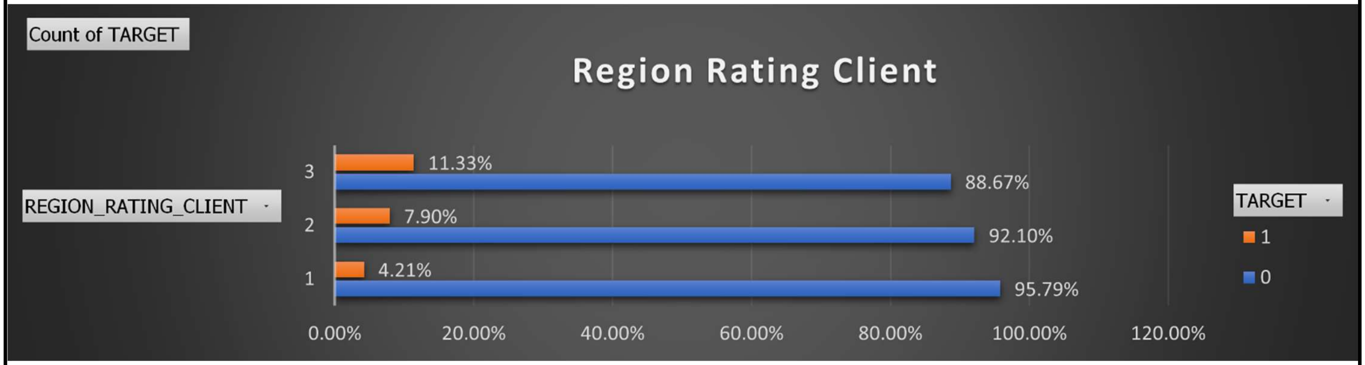


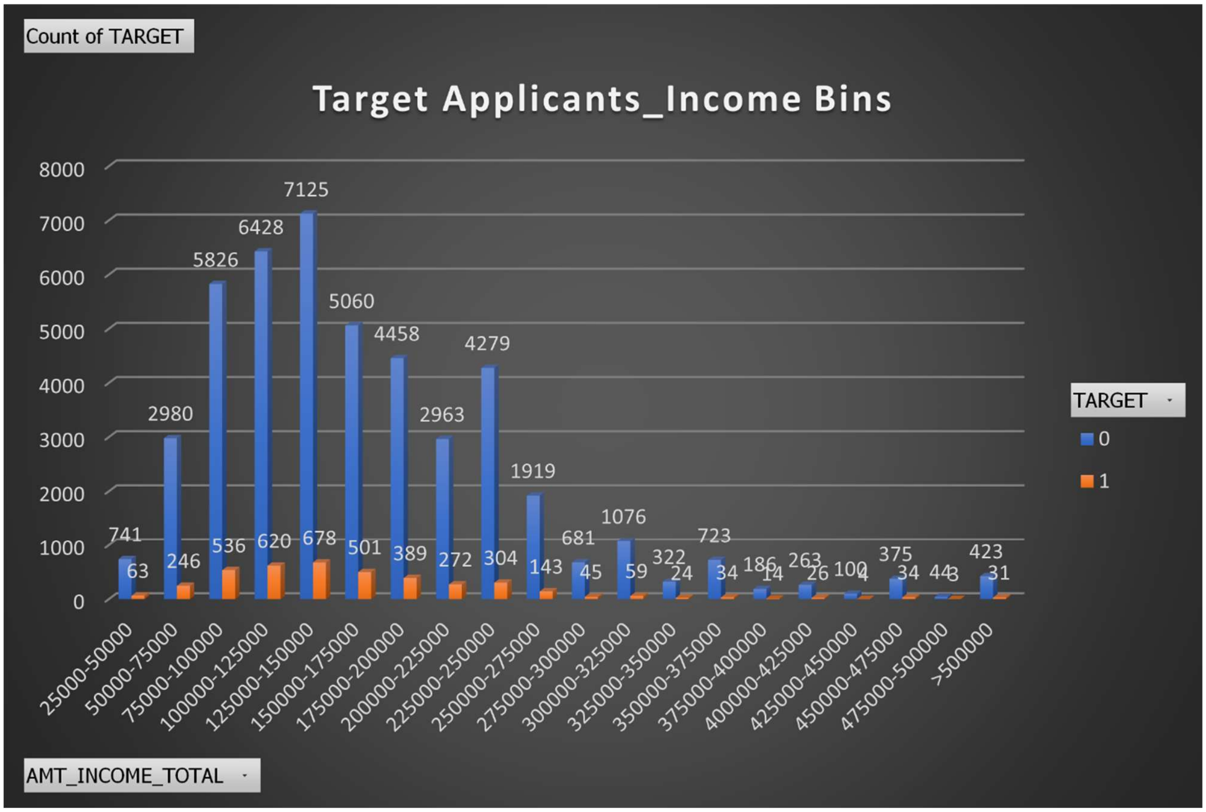

Count of TARGET
Years of Employed
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
89.58%
92.38%
94.63%
95.28%
95.02%
96.82%
95.98%
98.64%
100.00%
100.00%
94.36%
10.42%
7.62%
5.37%
4.72%
4.98%
3.18%
4.02%
1.36%
0.00%
0.00%
5.64%
0-5
5-10
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
>50
TARGET
🟦 0
🟧 1
DAYS_EMPLOYED(yrs)


Count of TARGET
Education Level
100.00%
50.00%
0.00%
100.00%
95.02%
91.48%
88.23%
90.98%
0.00%
4.98%
8.52%
11.77%
9.02%
1
0
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
TARGET
🟦 0
🟧 1
NAME_EDUCATION_TYPE


Count of TARGET
Housing Type
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents
TARGET
🟦 0
🟧 1
NAME_HOUSING_TYPE


Count of TARGET
Amount Credit for Loans
100.00%
91.62%
93.48%
100.00%
89.76%
66.67%
89.47%
100.00%
92.52%
97.73%
92.79%
95.00%
92.84%
98.28%
94.12%
94.40%
96.39%
94.76%
95.51%
96.63%
94.40%
AMT_CREDIT
45000-195000
195000-345000
345000-495000
495000-645000
645000-795000
795000-945000
945000-1095000
1095000-1245000
1245000-1395000
1395000-1545000
1545000-1695000
1695000-1845000
1845000-1995000
TARGET


Count of TARGET
Region Rating Client
REGION_RATING_CLIENT
3
11.33%
88.67%
2
7.90%
92.10%
1
4.21%
95.79%
0.00%
20.00%
40.00%
60.00%
80.00%
100.00%
120.00%
TARGET
1
0


Count of TARGET
Target Applicants_Income Bins
8000
7000
6000
5000
4000
3000
2000
1000
0
741
63
2980
246
5826
536
6428
620
7125
678
5060
501
4458
389
2963
272
4279
304
1919
143
681
45
1076
59
322
24
723
34
186
14
263
26
100
4
375
34
44
3
423
31
25000-50000
50000-75000
75000-100000
100000-125000
125000-150000
150000-175000
175000-200000
200000-225000
225000-250000
250000-275000
275000-300000
300000-325000
325000-350000
350000-375000
375000-400000
400000-425000
425000-450000
450000-475000
475000-500000
>500000
TARGET
0
1
AMT_INCOME_TOTAL

5

Top Correlations

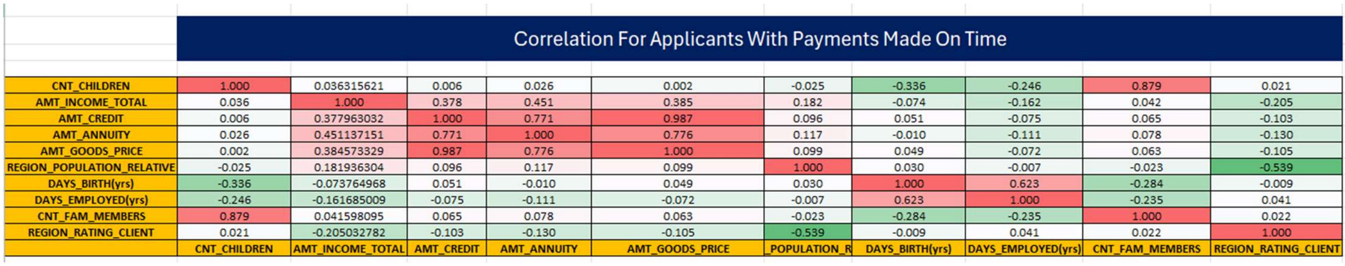
Different positions within a company often have different tiers or levels.

1) **<u>Task</u>:** 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.

<u>STEPS</u>:

To calculate the correlation of different scenarios (scenarios with numeric data) i copied columns with numeric data to a different sheet and calculated their correlation matrix using data>data analytics>correlation.

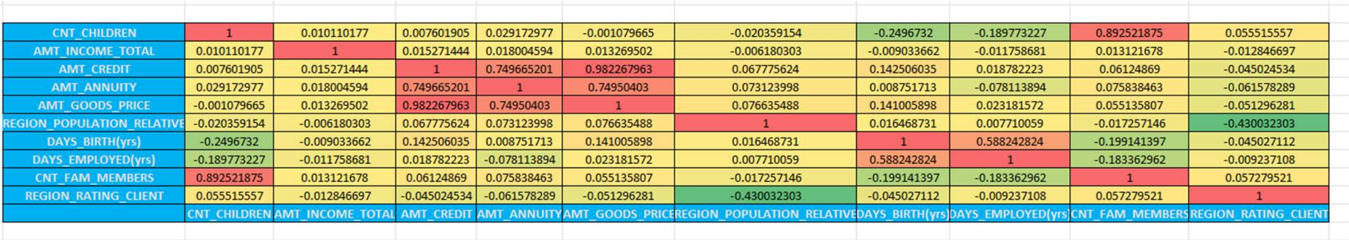
Correlation For Applicants With Payments Made On Time



Correlation For Applicants With Payments Made On Time

CNT_CHILDREN	1.000	0.036315621	0.006	0.026	0.002	-0.025	-0.336	-0.246	0.879	0.021
AMT_INCOME_TOTAL	0.036	1.000	0.378	0.451	0.385	0.182	-0.074	-0.162	0.042	-0.205
AMT_CREDIT	0.006	0.377963032	1.000	0.771	0.987	0.096	0.051	-0.075	0.065	-0.103
AMT_ANNUITY	0.026	0.451137151	0.771	1.000	0.776	0.117	-0.010	-0.111	0.078	-0.130
AMT_GOODS_PRICE	0.002	0.384573329	0.987	0.776	1.000	0.099	0.049	-0.072	0.063	-0.105
REGION_POPULATION_RELATIVE	-0.025	0.181936304	0.096	0.117	0.099	1.000	0.030	-0.007	-0.023	-0.539
DAYS_BIRTH(yrs)	-0.336	-0.073764968	0.051	-0.010	0.049	0.030	1.000	0.623	-0.284	-0.009
DAYS_EMPLOYED(yrs)	-0.246	-0.161685009	-0.075	-0.111	-0.072	-0.007	0.623	1.000	-0.235	0.041
CNT_FAM_MEMBERS	0.879	0.041598095	0.065	0.078	0.063	-0.023	-0.284	-0.235	1.000	0.022
REGION_RATING_CLIENT	0.021	-0.205032782	-0.103	-0.130	-0.105	-0.539	-0.009	0.041	0.022	1.000
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	_POPULATION_R	DAYS_BIRTH(yrs)	DAYS_EMPLOYED(yrs)	CNT_FAM_MEMBERS	REGION_RATING_CLIENT

Correlation For Applicants With Payment Difficulties



CNT_CHILDREN	1	0.010110177	0.007601905	0.029172977	-0.001079665	-0.020359154	-0.2496732	-0.189773227	0.892521875	0.055515557
AMT_INCOME_TOTAL	0.010110177	1	0.015271444	0.018004594	0.013269502	-0.006180303	-0.009033662	-0.011758681	0.013121678	-0.012846697
AMT_CREDIT	0.007601905	0.015271444	1	0.749665201	0.982267963	0.067775624	0.142506035	0.018782223	0.06124869	-0.045024534
AMT_ANNUITY	0.029172977	0.018004594	0.749665201	1	0.74950403	0.073123998	0.008751713	-0.078113894	0.075838463	-0.061578289
AMT_GOODS_PRICE	-0.001079665	0.013269502	0.982267963	0.74950403	1	0.076635488	0.141005898	0.023181572	0.055135807	-0.051296281
REGION_POPULATION_RELATIVE	-0.020359154	-0.006180303	0.067775624	0.073123998	0.076635488	1	0.016468731	0.007710059	-0.017257146	-0.430032303
DAYS_BIRTH(yrs)	-0.2496732	-0.009033662	0.142506035	0.008751713	0.141005898	0.016468731	1	0.588242824	-0.199141397	-0.045027112
DAYS_EMPLOYED(yrs)	-0.189773227	-0.011758681	0.018782223	-0.078113894	0.023181572	0.007710059	0.588242824	1	-0.183362962	-0.009237108
CNT_FAM_MEMBERS	0.892521875	0.013121678	0.06124869	0.075838463	0.055135807	-0.017257146	-0.199141397	-0.183362962	1	0.057279521
REGION_RATING_CLIENT	0.055515557	-0.012846697	-0.045024534	-0.061578289	-0.051296281	-0.430032303	-0.045027112	-0.009237108	0.057279521	1
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH(yrs)	DAYS_EMPLOYED(yrs	CNT_FAM_MEMBERS	REGION_RATING_CLIENT

• Conclusion:

The Bank Loan Case Study highlights the critical role of **Exploratory Data Analysis (EDA)** in addressing financial risks associated with loan approvals. By leveraging structured analysis techniques and Excel's advanced functionalities, the project successfully identified patterns and key factors influencing loan defaults.

The analysis revealed that variables like **loan amount, credit history, customer income, and dependents** significantly impact the likelihood of default. For instance, customers with higher incomes were less likely to default, while those with no or limited credit history posed a greater risk. Outliers in income and loan amounts indicated extreme cases that could skew decision-making if not properly managed.

Moreover, the study emphasized the presence of **data imbalance** in the target variable (e.g., a higher proportion of approved loans compared to rejected or defaulted ones). This imbalance, if left unaddressed, could affect the accuracy of predictive models or decision frameworks. Visualizations, such as pie charts and bar graphs, effectively illustrated these patterns, aiding in clear and actionable interpretations.

The segmented univariate and bivariate analyses provided deeper insights into customer and loan attributes across different scenarios, such as **approved loans, refused loans, and payment difficulties.** These findings highlighted relationships, such as the strong correlation between loan amount and payment delays, enabling the company to tailor lending strategies.

From a business perspective, this project equips the company with valuable insights to enhance its loan approval process. By identifying high-risk applicants early, the organization can:

1. **Minimize financial losses** by adjusting loan terms or denying risky applications.
2. **Optimize revenue opportunities** by ensuring capable applicants are not rejected unnecessarily.
3. **Implement targeted interest rates** for high-risk profiles to mitigate potential defaults.

In conclusion, the project underscores the importance of data-driven decision-making in the finance sector. By addressing challenges like missing data, outliers, and data imbalance, and by identifying key predictors of loan defaults, the company can refine its risk assessment processes. This not only reduces losses but also ensures sustainable growth by maintaining customer satisfaction and trust.

Link to Excel Sheet:

https://docs.google.com/spreadsheets/d/1s8ENJO1ky-MLubMopT8hMr83fN1gnr-W/edit?usp=sharing&ouid=103428047773693985368&rtpof=true&sd=true

END