BANK LOAN CASE STUDY

-SOURAV PATTANAYAK

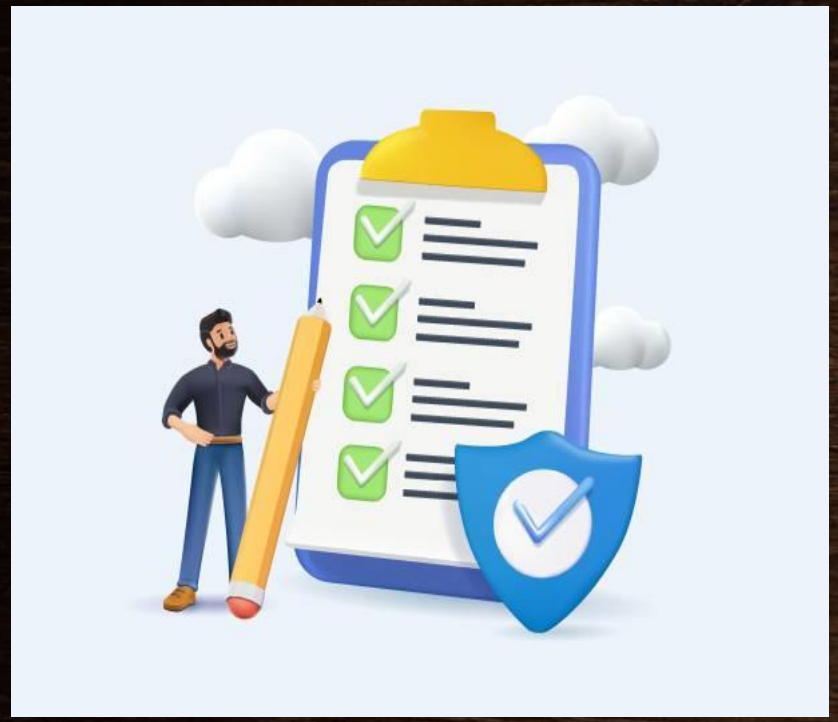


Table of Contents

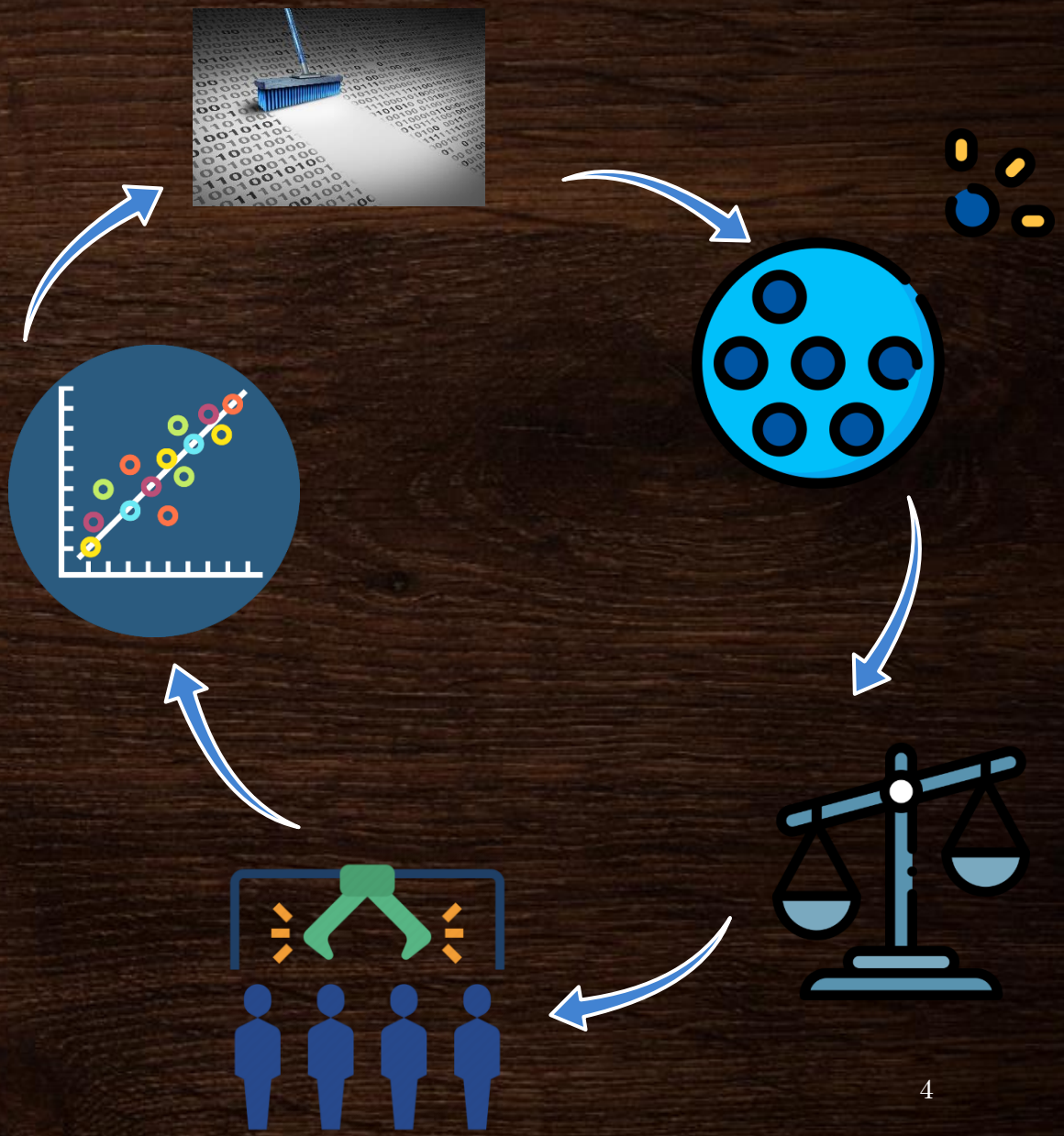
Project description

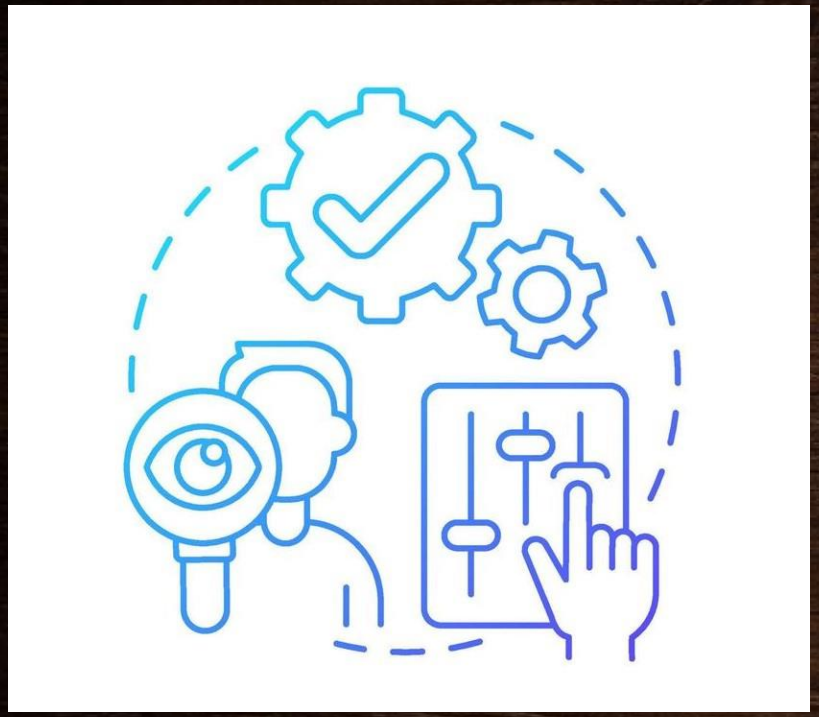
- In this project, we aim to analyze a loan application dataset to gain insights into factors influencing loan defaults.
- Our goal is to handle missing data, identify outliers, analyze data imbalance, and perform various statistical analyses.
- Accurate handling of data is crucial for ensuring the reliability of our analysis.
- Since some customers don't have a sufficient credit history, they default on their loans. I was provided with a dataset and my task was to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

Few charts and calculations (e.g.: Heatmaps, IQR calculations were not considered since the project was done in Excel only, and due to a huge dataset, Excel will lag provided some complicated calculations.

Project Objectives

- To explore and understand the loan application dataset.
- To Identify and deal with missing data effectively.
- To detect and handle outliers to ensure robust analysis.
- To analyze data imbalance to address issues in binary classification.
- To perform univariate, segmented univariate, and bivariate analyses for deeper insights.
- To Identify top correlations between variables and the target variable for different scenarios.





Approach

- I have followed a systematic approach to analyze the loan application dataset.
- Done Data preprocessing and cleaning to ensure accuracy.
- Utilized Microsoft Excel and Power BI for Data Analysis and chart creation , Microsoft PowerPoint for create and present the presentation.
- Employed statistical functions and visualization techniques.

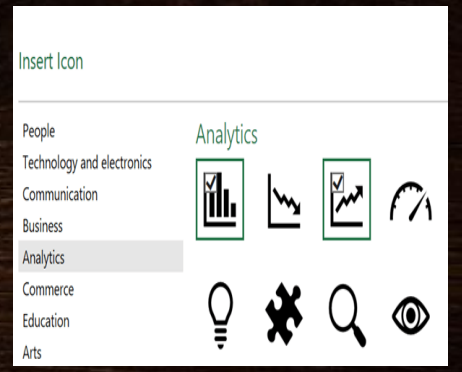
Methodology & Tech-stack used

A. Our analysis relies on Microsoft Excel 2016, PowerPoint, and Power BI.

B. Data processing techniques include handling missing data, outlier identification, and data balancing.

C. Statistical functions used for analysis: mean, median, correlation, and more.

Tech Stack



Insights-Data Cleaning

Data cleaning is the first and most important task to execute before proceeding to any kind of Data Analysis, in this project, the raw dataset had 122 columns and 307512 rows in a few columns, and in the rest of the columns it had 50000 rows, I had to remove rows above 50000 that are extra in a few columns and then the deletion of unnecessary columns and dealing with missing values were executed, this was done in three steps,

1. By Identifying the missing data in the dataset.

2. By Deleting columns with >30% null values and unnecessary columns.

3. By Imputing numerical columns using the median method(For columns<30% null values)

The deleted columns are:

1. FLAG MOBIL(Since only one 1 0 value and all are 1 value).

2. From FLAG DOCUMENT 2 to FLAG DOCUMENT 21(since these columns are unnecessary to the analyses).

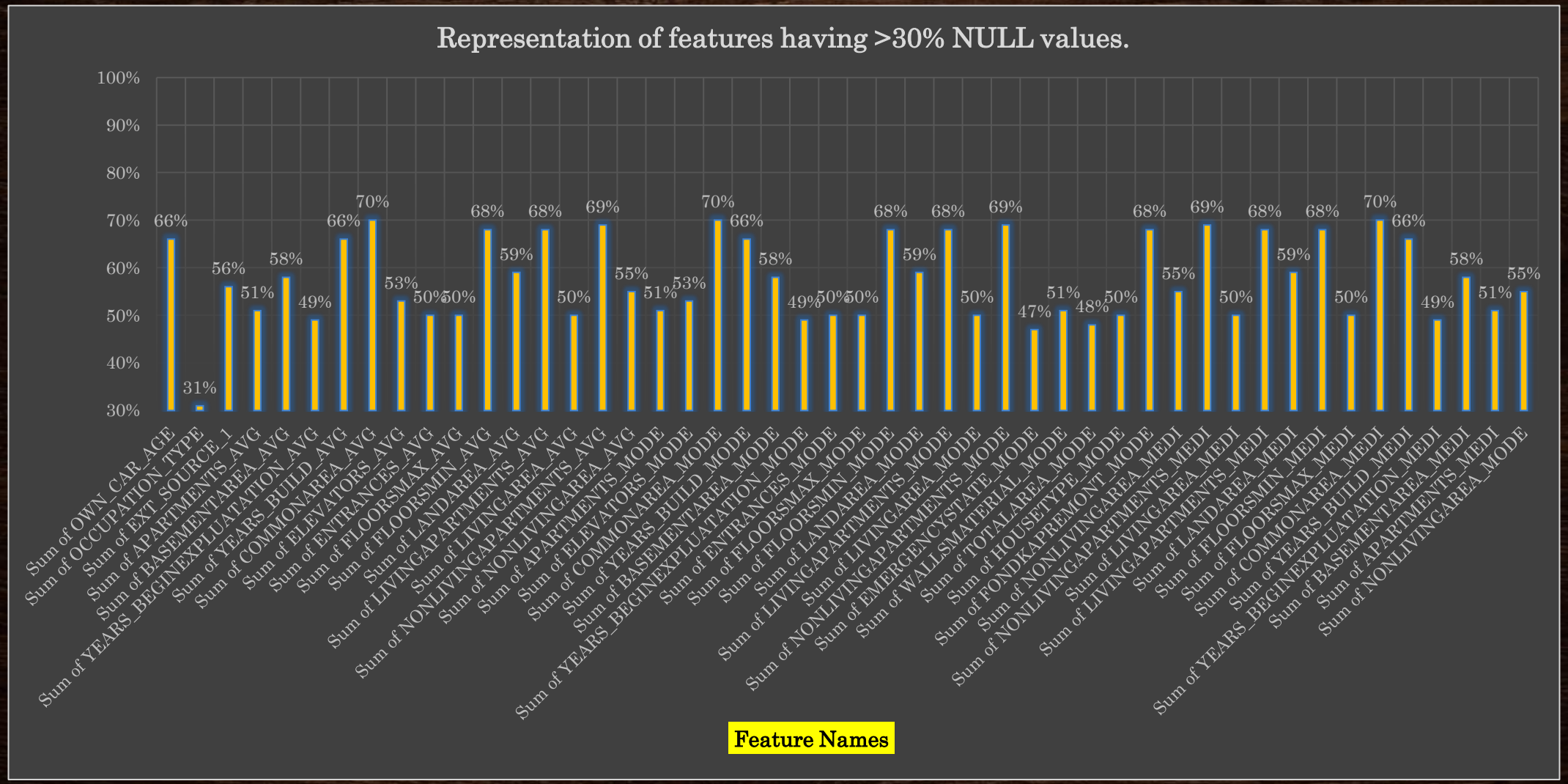
3. Median values were put in the blank cells for the features,

EXT_SOURCE_3,
AMT_REQ_CREDIT_BUREAU_HOUR,
AMT_REQ_CREDIT_BUREAU_DAY,
AMT_REQ_CREDIT_BUREAU_WEEK,
AMT_REQ_CREDIT_BUREAU_MON,
AMT_REQ_CREDIT_BUREAU_QRT,
AMT_REQ_CREDIT_BUREAU_YEAR

The formula used in computing the median:
=IF(ISBLANK(Reference cell), MEDIAN(Reference cell range), reference cell)

To go to the Excel worksheet, please click here:
https://docs.google.com/spreadsheets/d/1IL2HHUOEOFtU-4KwqyxX1sJeUOhtfQ_Y/edit?usp=sharing&ouid=101218975538808492645&rtpof=true&sd=true

Insights-Data Cleaning



Features with >30% Null values

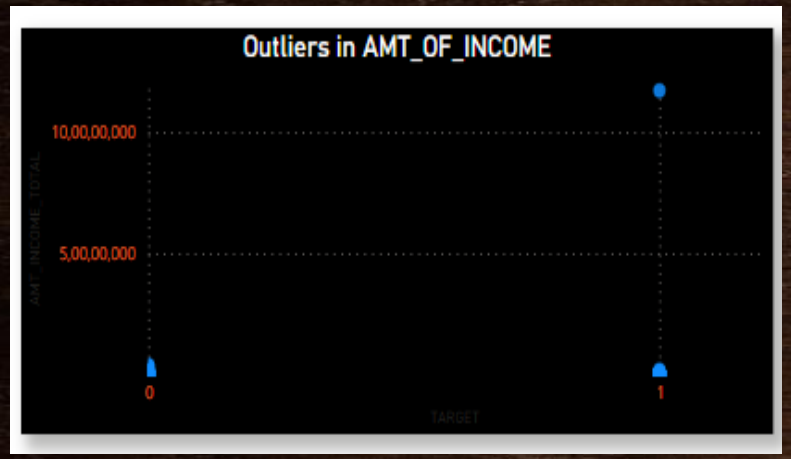
Insights-The Outliers

Focusing on numerical variables, outliers of the dataset were identified, the outliers were detected using the TARGET Feature with other numerical columns that help to understand the distribution of defaulters and conformists.

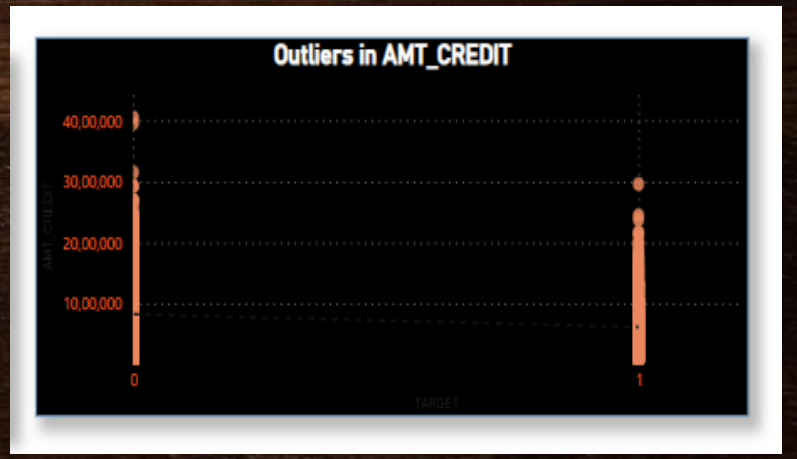
The outliers were detected against the TARGET feature with features:

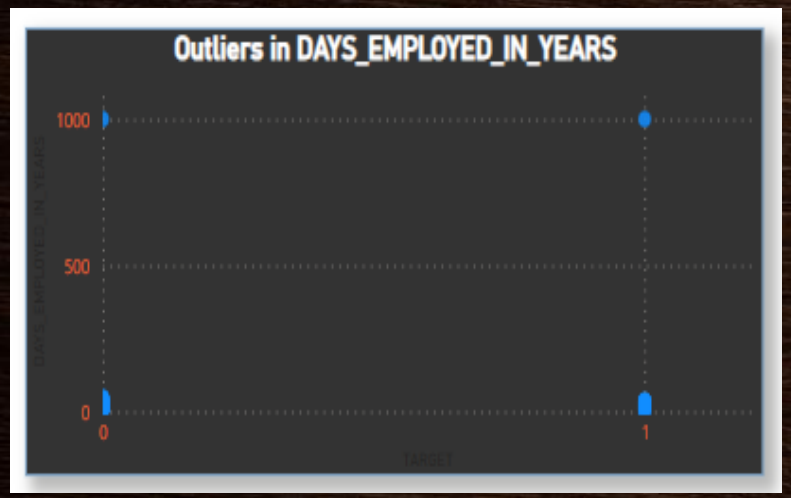
- *CNT_OF_CHILDERN*
- *AMT_OF_INCOME*
- *AMT_CREDIT*
- *DAYS_EMPLOYED_IN YEARS*

Insights-The Outliers



Outliers can be calculated manually by Calculating Q1, Q3, and IQR separately in a new feature.

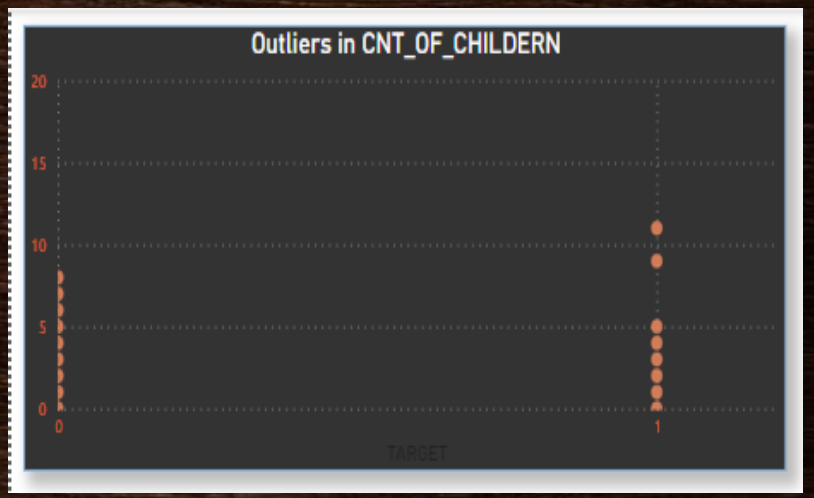




Q1: ***=QUARTILE(A2:A50000, 1)***
Q3:=QUARTILE(A2:A50000, 3)

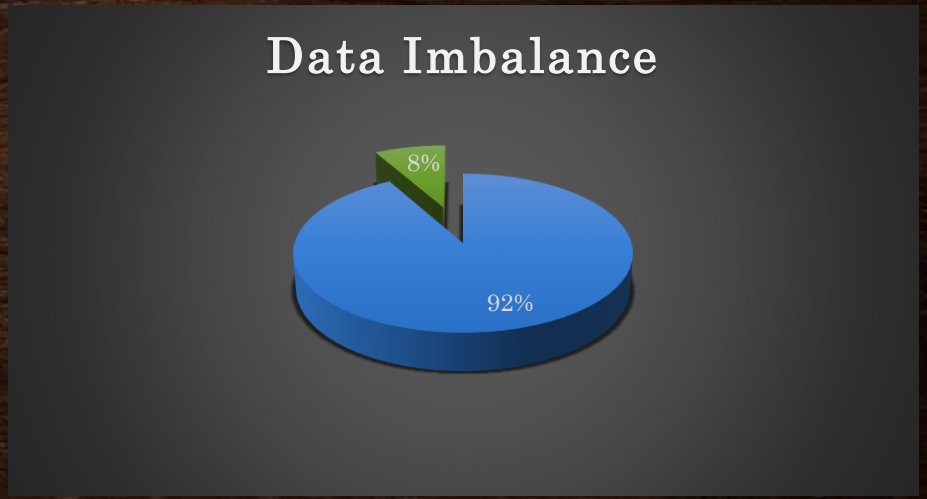
IQR: = (Q3-Q1)

Lower bound: =Q1-1.5*IQR
Upper bound: =Q3+1.5*IQR



Insights-Data imbalance

- Data imbalance can affect the accuracy of the analysis, hugely.
- In the given dataset the TARGET feature was the choice for calculating the Data imbalance since this feature actually identifies the basic information on customers(0 for customers who are not defaulters in payment and 1 for the customers who were defaulters in payment)
- A pivot table of the TARGET feature was created and then the counts of both the values(0 and 1) were calculated followed by the Pie-chart.
- Out of 49999 values, only 4026 values are 1 and others are 0.



Feature	Count of TARGET
0	45973
1	4026
Grand Total	49999

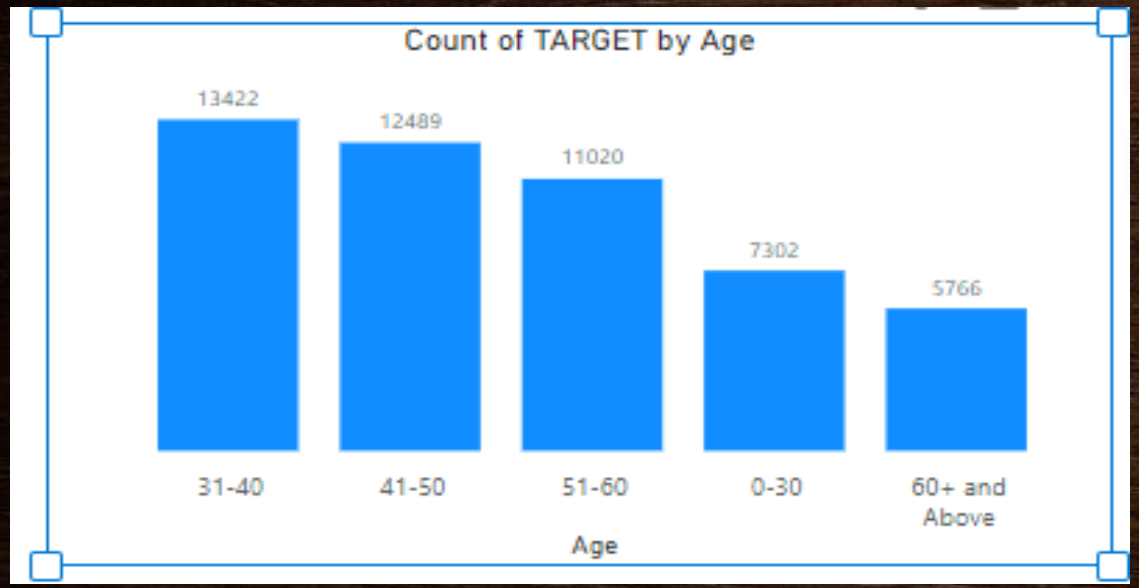
Insights-Univariate and Segmented Analysis

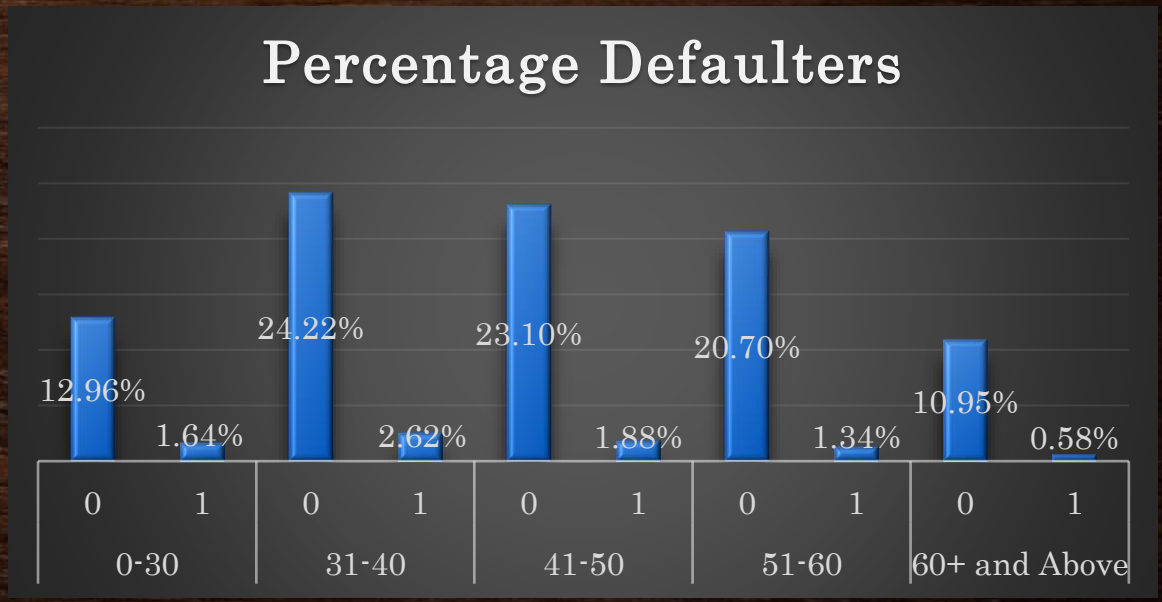
- To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.
- On targets, I have Performed univariate analysis to understand variable distributions.
- Utilized segmentation for comparing variable distributions in different scenarios.
- Presented results using insightful column charts and line charts.

I have done Univariate and segmented analysis on Features:

- Age(bins) and Segmented age(The count of targets in percentage)
- Years employed and segmented Years Employed
- Code of gender and segmented Count of gender
- Name type suite vs Segmented name type suite
- Education type vs segmented education type
- Occupation type and occupation type segmented

Insights-Univariate and Segmented Analysis

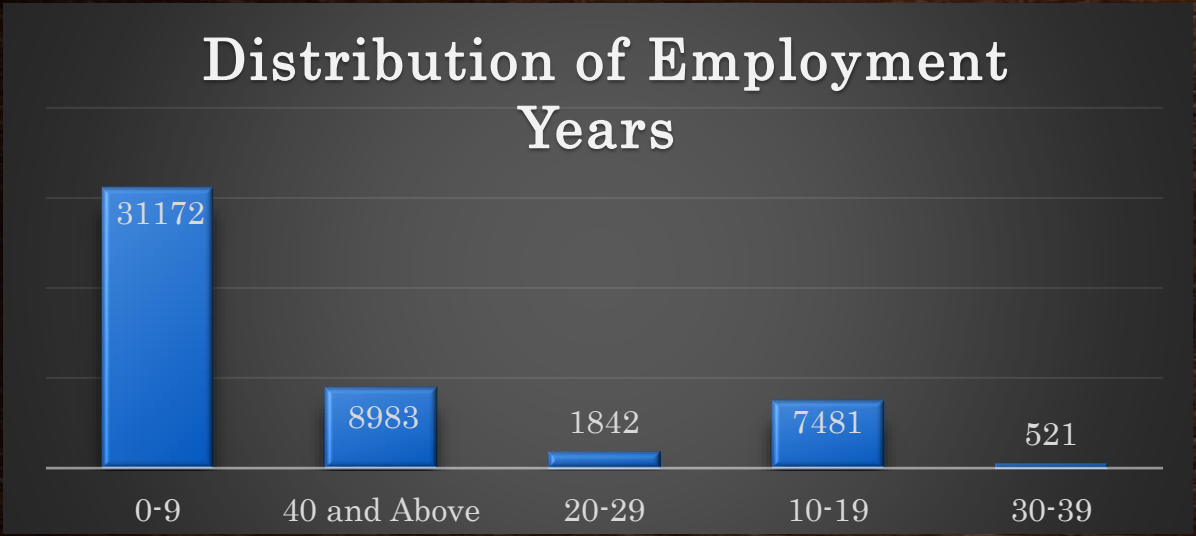


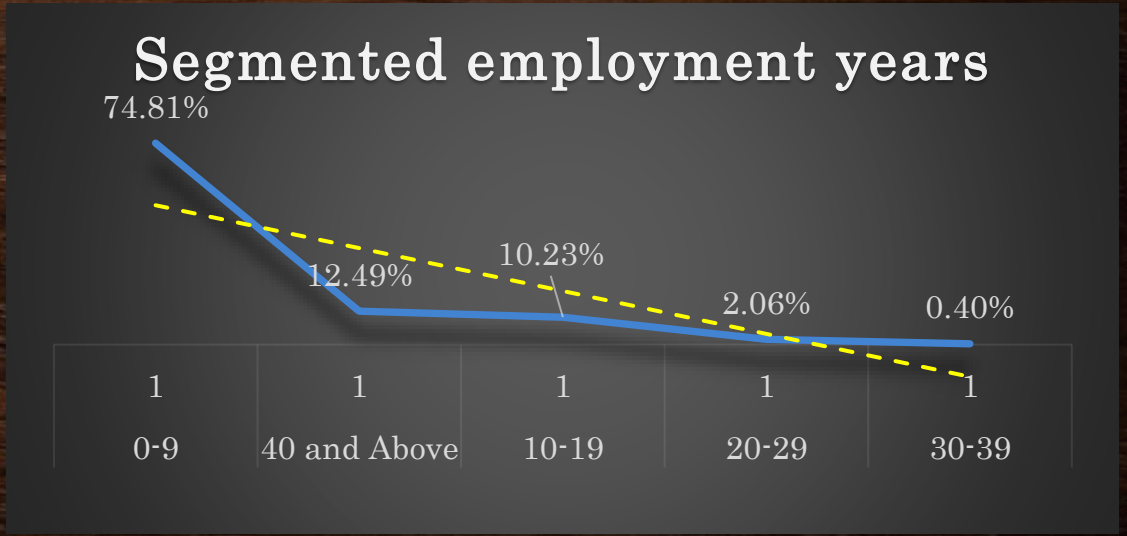


The maximum number of customers take loan reside in the age range of 31-40

The highest number of defaulters:31-40. The lowest number of defaulters: 60+ and above

Insights-Univariate and Segmented Analysis

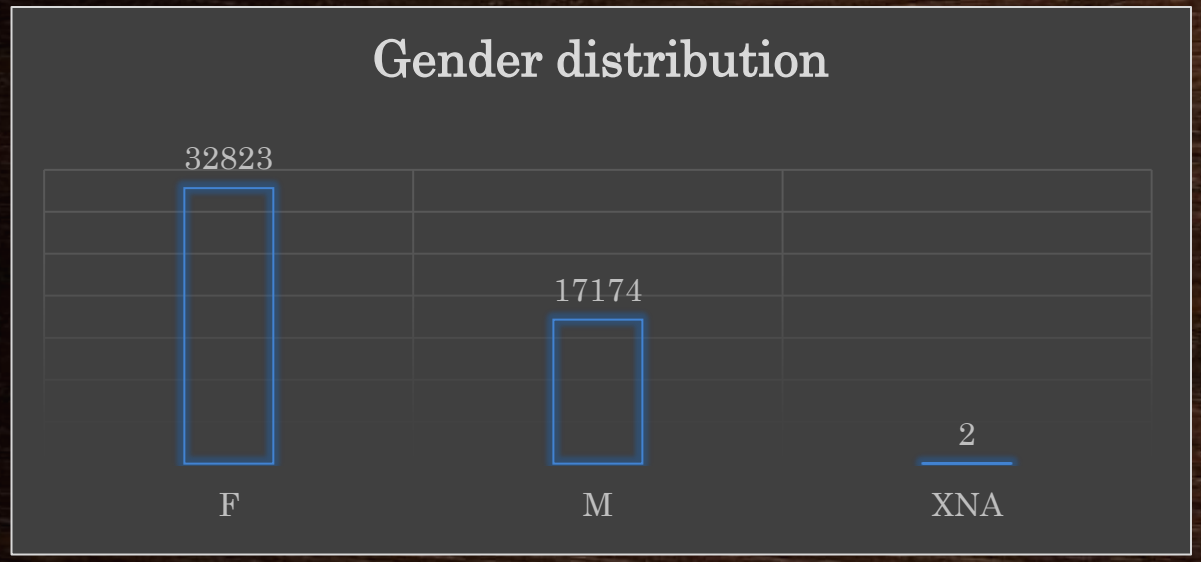


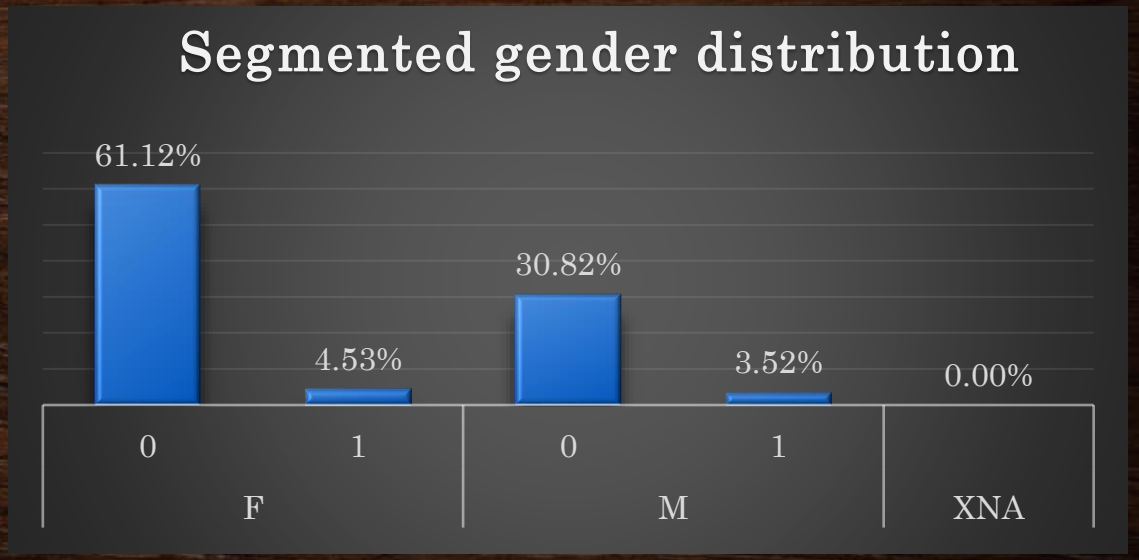


The maximum number of customers who take loans reside in 0-9 years of experience

The highest number of defaulters:0-9 years
The lowest number of defaulters: 30-39 years.

Insights-Univariate and Segmented Analysis

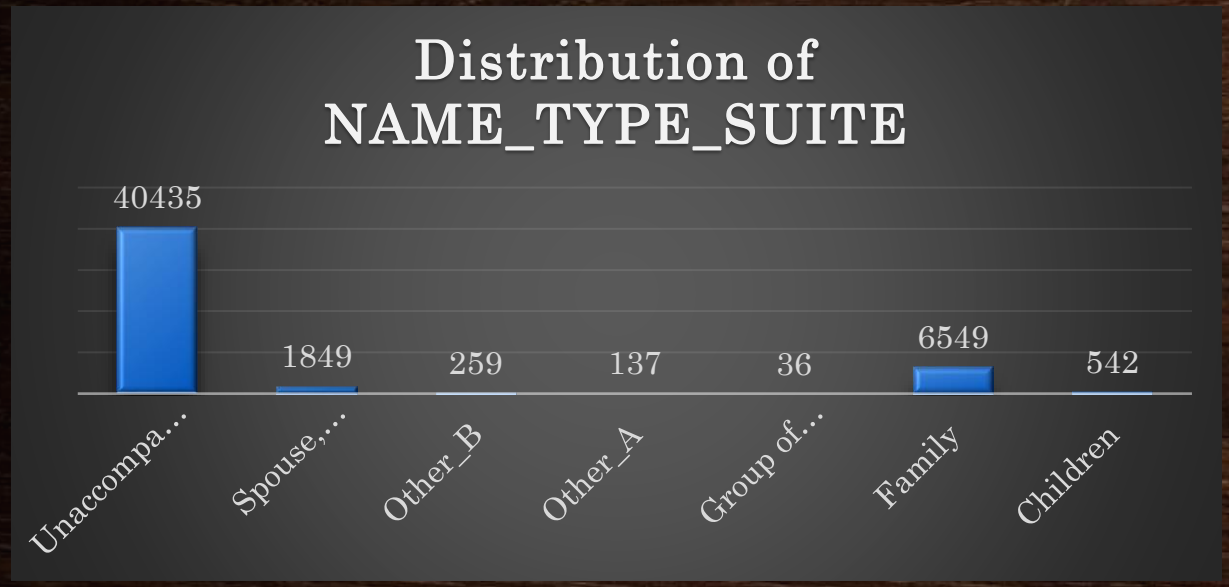


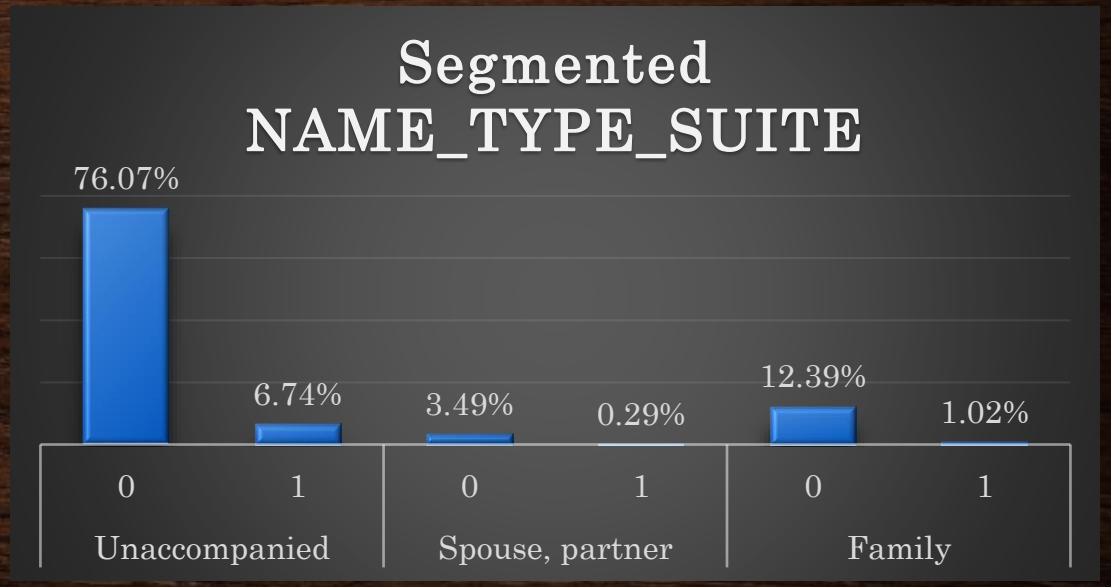


The Number of female customers is much more than other gender(s)

Female customers have more number of defaulters compared with other gender(s)

Insights-Univariate and Segmented Analysis



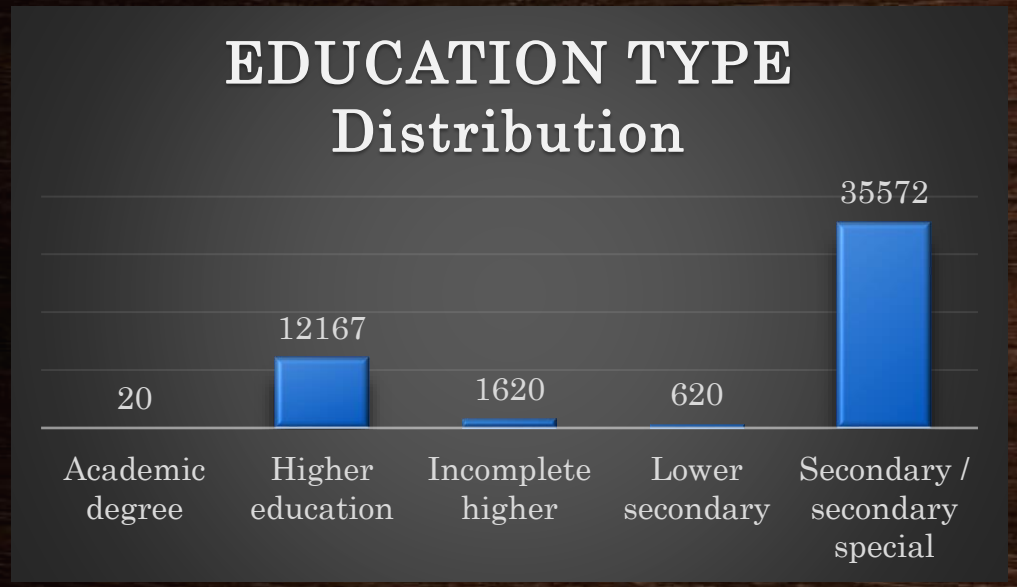


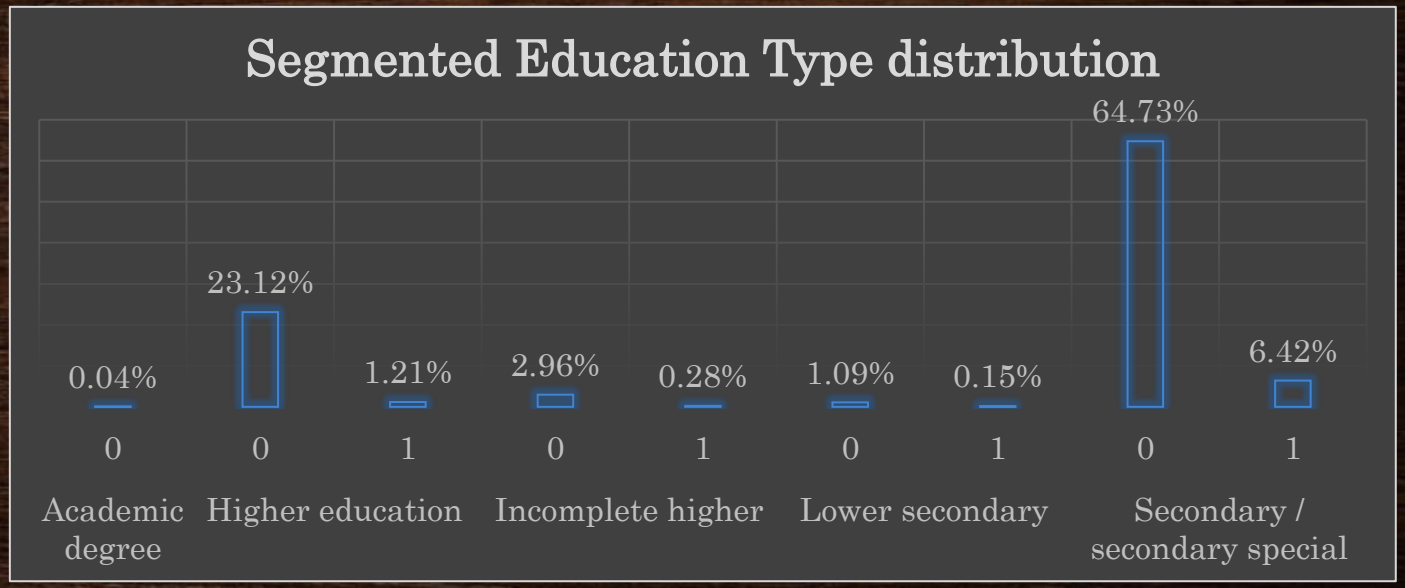
The majority of customers were unaccompanied at the time of issuing loans to themselves from the bank.

Maximum number of defaulters: ***Unaccompanied people***

The minimum number of defaulters: Customers with ***spouses, and partners****.*

Insights-Univariate and Segmented Analysis



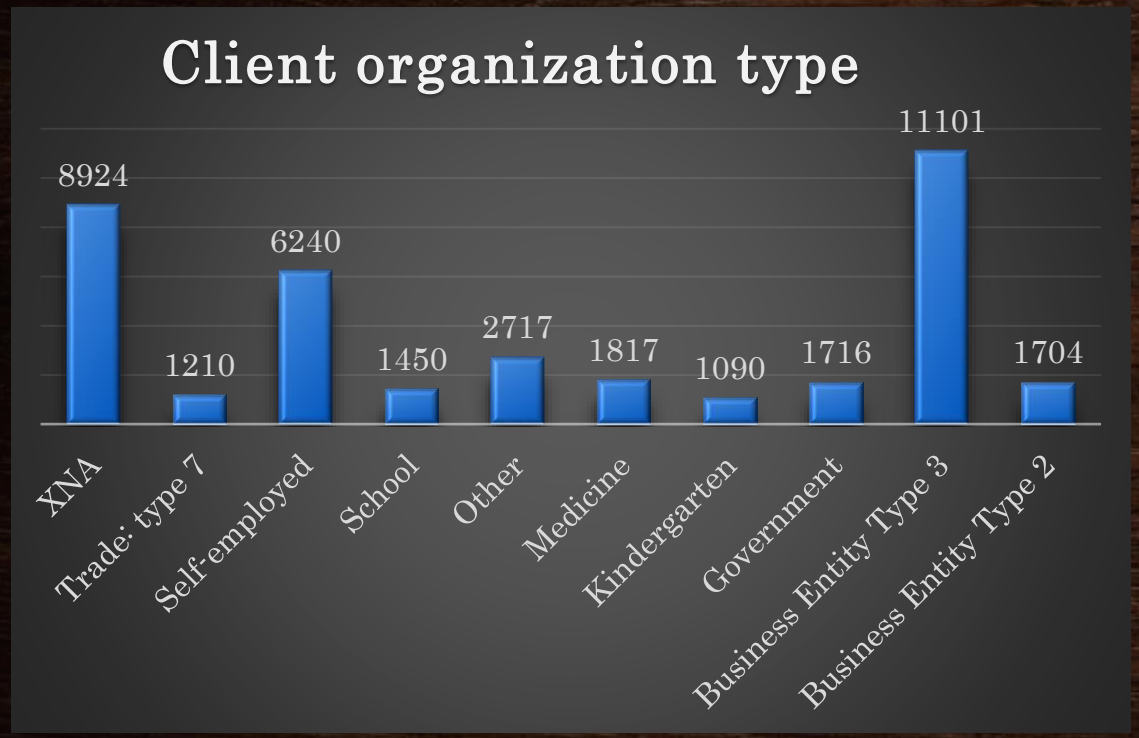


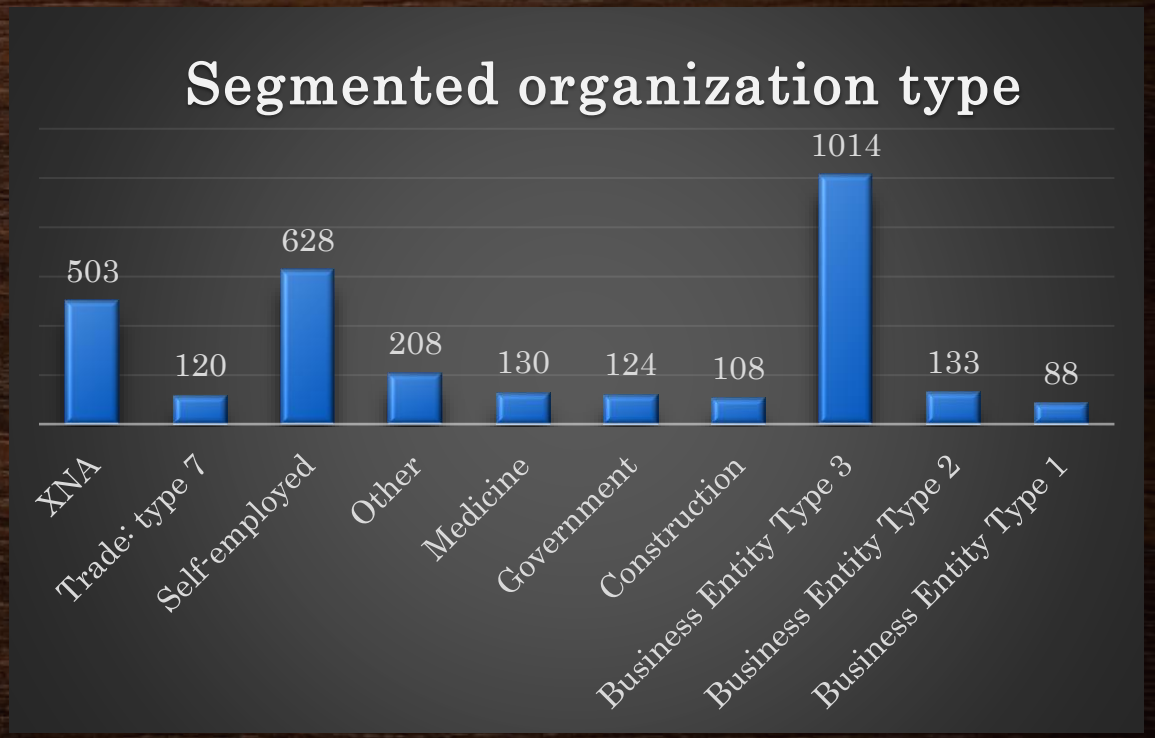
The majority of customers completed their secondary/special secondary examination at the time of issuing loans to themselves from the bank.

The maximum number of defaulters: ***People with academic degrees.***

The minimum number of defaulters: Customers with secondary/special secondary degrees.

Insights-Univariate and Segmented Analysis

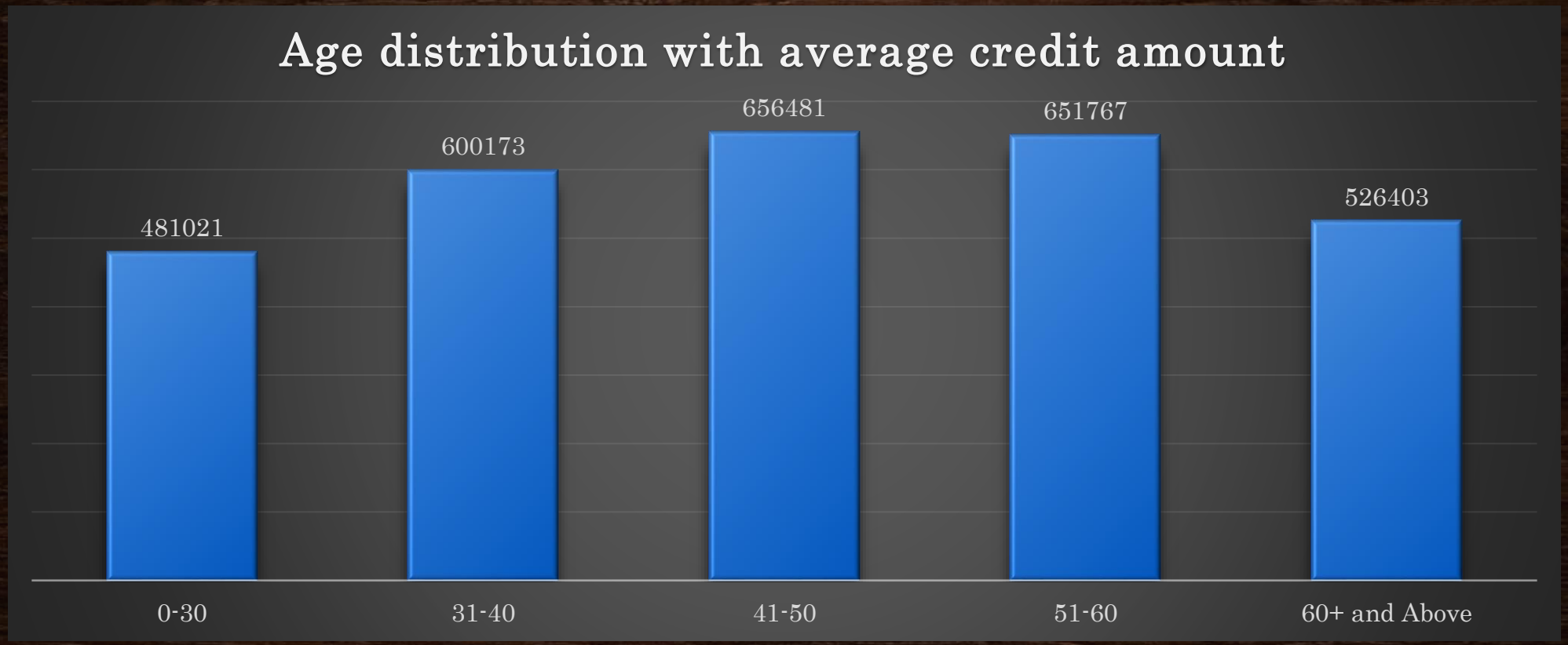




The majority of customers were from business entity-3 at the time of issuing loans to themselves from the bank.

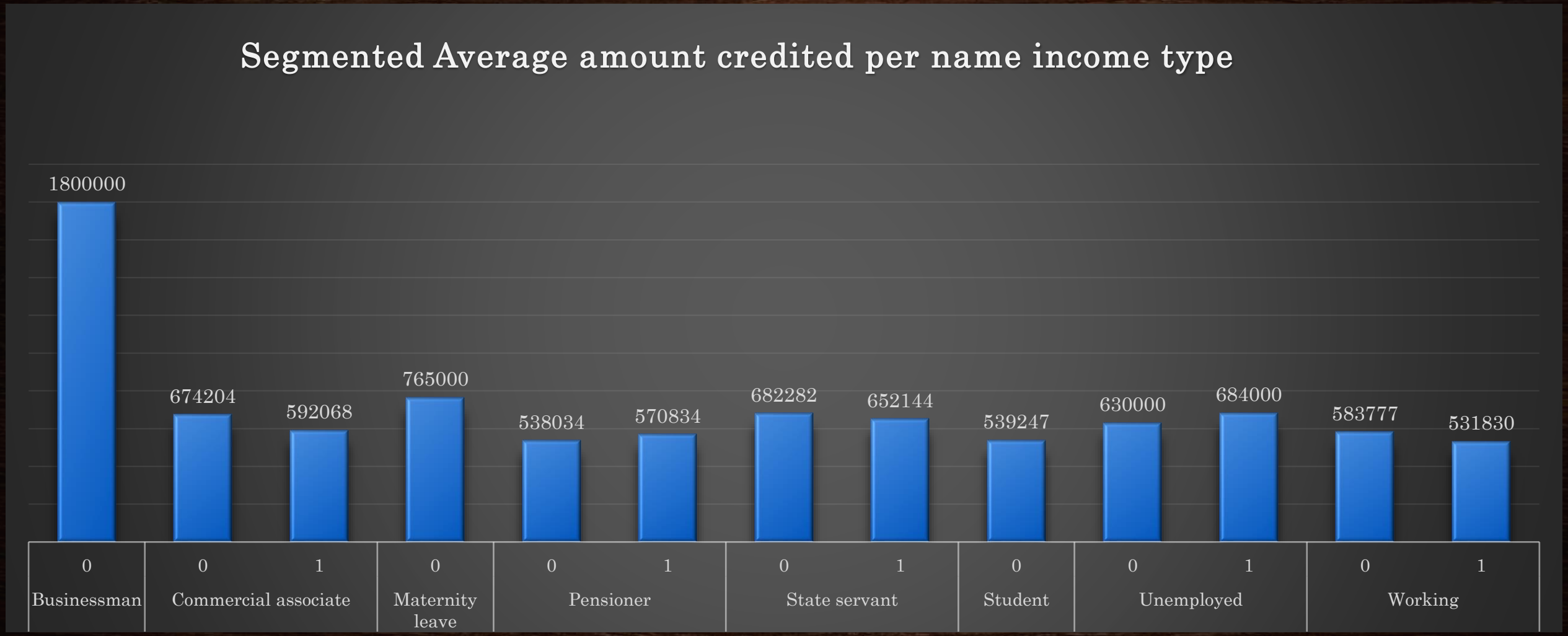
Maximum number of defaulters: Business entity-3
The minimum number of defaulters: Business entity -1.

Insights-The Bivariate Analysis



From the bivariate analysis, it is seen that customers in the age range of 41-50 have credited the highest amount of average credit.

Insights-The Bivariate Analysis



From the bivariate analysis, it is seen that customers who were businessmen, under maternity, and students were no defaulters whereas unemployed customers were the highest number of defaulters.

Insights-The correlations

- Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.
- IN this project I have segmented the dataset to find top correlations between variables and the target variable.
- Calculated correlation coefficients using Excel functions.
- Highlighted top 10 correlations with correlation matrices.

1. I have used Excel's Data Analysis feature to find correlation matrices.

2. The formula used to find the correlation for the specific features:

=CORREL(Sheet1!L2:L4027,Sheet1!Q2:Q4027

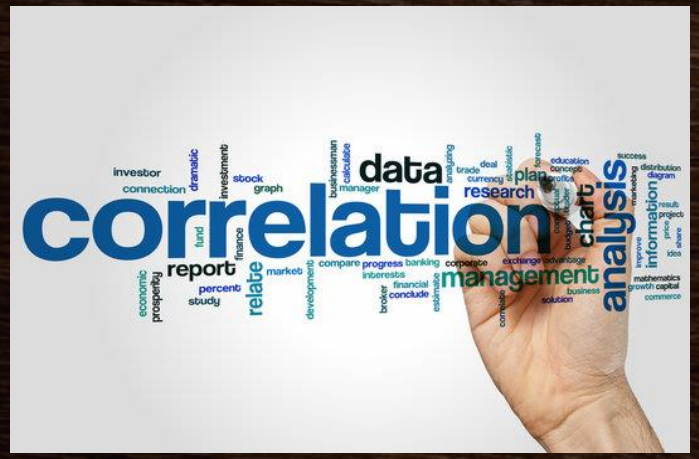
I have executed correlation analysis on Features:

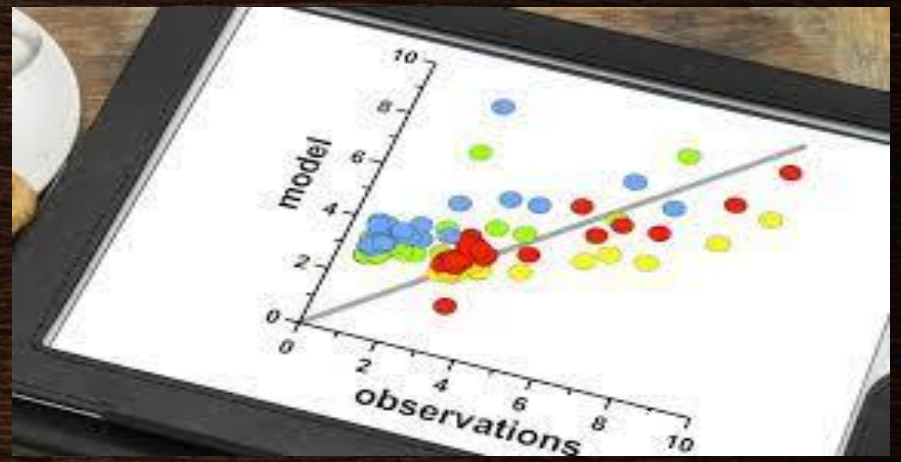
- YEARS OF EXPERIENCE WITH FLAG_EMP_PHONE
- OBS_30_CNT_SOCIAL_CIRCLE WITH OBS_60_CNT_SOCIAL_CIRCLE
- AMT_CREDIT WITH AMT_GOODS_PRICE
- REGION_RATING_CLIENT WITH REGION_RATING_CLIENT_WITH_CITY
- CNT_CHILDERN WITH CNT_FAMILY_MEMBERS
- REG_REGION_NOT_WORK_REGION WITH LIVE_REGION_NOT_WORK_REGION
- DEF_30_CNT_SOCIAL_CIRCLE WITH DEF_60_CNT_SOCIAL_CIRCLE
- REG_CITY_NOT_WORK_CITY WITH LIVE_CITY_NOT_WORK_CITY
- AMT_ANNUITY WITH AMT_GOODS_PRICE
- AMT_CREDIT WITH AMT_ANNUITY.

Insights-The correlations

Please follow the hyperlink below to refer to the correlation matrices

Columns





Insights-The correlations-Top 10 values

1. Years of experience-
FLAG_EMP_PHONE :

-0.999899227065333

2.OBS_30_CNT_SOCIAL_CIRCLE-
OBS_60_CNT_SOCIAL_CIRCLE:

0.998064837349622

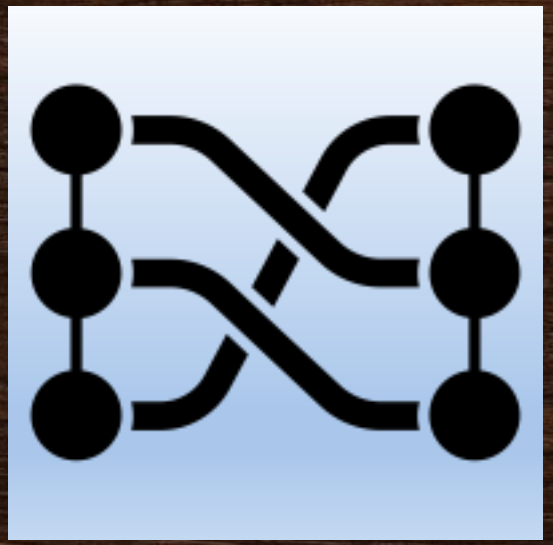
3.AMT_CREDIT-AMT_GOODS_PRICE:

0.982432317758075

4.REGION_RATING_CLIENT-
REGION_RATING_CLIENT_WITH_CITY:

0.950768898851766

5.CNT_CHILDERN -
CNT_FAMILY_MEMBERS:
0.892521874588344



.

6.REG_REGION_NOT_WORK_REGION-
LIVE_REGION_NOT_WORK_REGION:

0.806743885821316

7.DEF_30_CNT_SOCIAL_CIRCLE-
DEF_60_CNT_SOCIAL_CIRCLE:

0.890496347984347

8.REG_CITY_NOT_WORK_CITY-
LIVE_CITY_NOT_WORK_CITY;

0.783754676418725

9.AMT_ANNUITY-AMT_GOODS_PRICE:

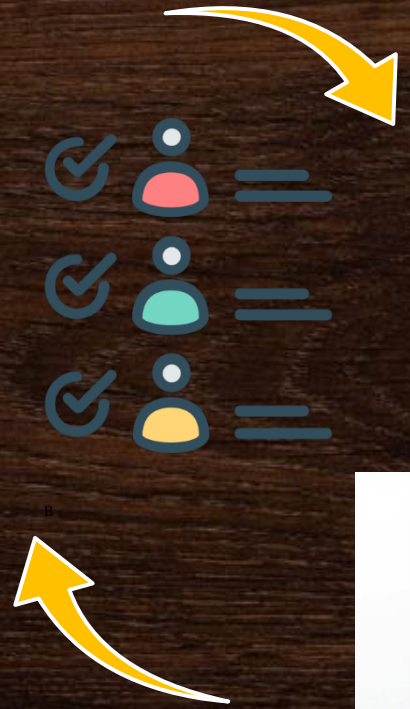
0.749705184075139

10.AMT_CREDIT-AMT_ANNUITY:

0.749665201431795

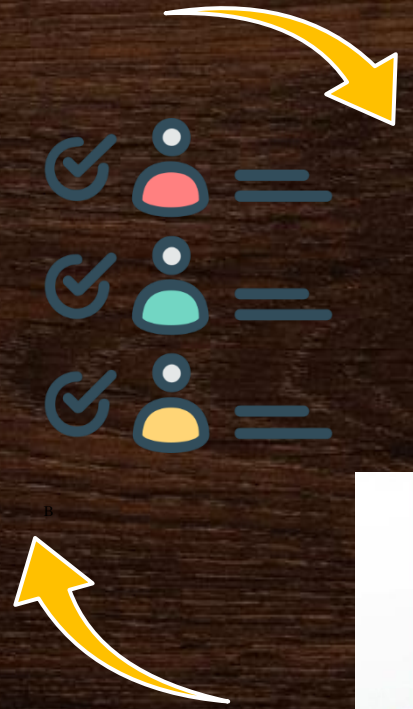
Key findings

- During the Data Cleaning, Removed 48 columns with >30% null values; filled missing values using the median.
- The Outliers were observed in income, children, and credit amount, especially for clients with payment difficulties (TARGET = 1).
- The Dataset was Highly imbalanced with only 8% facing payment difficulties.
- From the charts it is evident that Bank prefers loans to senior clients (31-40 age) with fewer defaulters aged 60+.
- Newly employed clients have higher default rates, posing risks.
- There are Slightly higher default rates in females than in males but compared with the number of loans taken, Females are way behind.
- The Secondary education clients are the largest loan takers.
- Business Entity Type 3 has the most loan takers, and Type 1 lowest default rate.
- There exist Strong correlations between several features, aiding risk assessment and decisions.



Conclusions

- The bank needs to address outliers and payment difficulties to minimize risks associated with defaulting clients.
- The imbalanced dataset necessitates the use of appropriate techniques to handle class imbalance during modeling.
- The bank's preference for granting loans to comparatively senior clients seems reasonable due to potential stability in financial backgrounds, resulting in the chart.
- Clients with limited employment experience may require closer scrutiny during loan approval.
- Further investigation is needed to understand why male clients show slightly higher default rates when compared with the total number of loans taken.
- Focusing on clients with secondary education might be beneficial as they form the largest customer segment.
- Business Entity Type 3 clients may require special attention due to their high loan uptake.

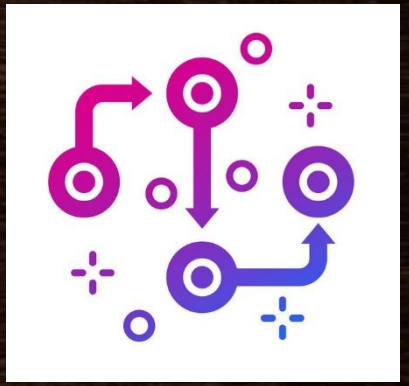


Conclusions-Correlations.

The correlations speak much about the necessary steps the Bank can take, considering the high correlation values

The Bank should

- Verify phone information for less experienced clients.
- Monitor high social interaction observations closely.
- Review the loan policies considering goods prices.
- Analyze regional ratings consistently.
- Address discrepancies in registered locations.
- Investigate clients with social circle defaults.
- Align loan annuity with goods prices.
- Optimize loan terms based on correlation with annuity payments.



Achievements

Being a statistical and pure Data Analysis project, this has helped me a lot in learning and applying, there are a few areas I have learned to work on,

- **Thorough Analysis:** As the first step, I have explored loan data for default insights.
- **Effective Data Handling:** Learned to manage missing values and outliers.
- **Balanced Analysis:** Got familiar with addressing data imbalance for classification.
- **Strong Correlations:** Got familiar with correlation methods and how to identify the key indicators for defaults.
- **Impactful Visualization:** Learned to create clear charts for better understanding, especially for complicated calculations.
- **Tech Expertise:** Added proficiency in Excel, PowerPoint, and Power BI.
- **Business Optimization:** Learned to deal with, work with, and Analyse real-world complex data for further improvements.

THANK YOU