-Rukmini Annadata (rukmini2k2@gmail.com)

Bank loan case study

Project Description:

Imagine yourself in the role of a data analyst within a financial institution specializing in the provision of various loan products to an urban customer base. The company grapples with a significant challenge - a portion of its clients, particularly those with limited credit histories, exploit this vulnerability, resulting in loan defaults. Your critical mission involves the application of Exploratory Data Analysis (EDA) techniques to delve deep into the dataset, uncover meaningful insights, and implement strategies to ensure that creditworthy applicants are not unfairly denied access to loans.

In this scenario, your analytical skills are crucial in identifying underlying trends and anomalies within the data. By doing so, you can contribute to the development of more robust credit assessment models, thereby striking a balance between mitigating risks associated with lending to customers with limited credit history and ensuring that deserving individuals are not unjustly rejected from loan opportunities. Ultimately, the objective is to enhance the lending process's accuracy, fairness, and efficiency, aligning it more closely with the financial institution's goals of providing accessible financial solutions to urban customers.

Approach:

Steps include:

Understanding data → Pre-processing -> Merging datasets -> Analysis & visualizations -> Result

Within the scope of this case study, we encountered two extensive datasets: one pertaining to the current loan applications and another to the previous loan applications. These datasets contained numerous extraneous columns and a considerable amount of missing data, rendering them impractical for risk assessment purposes. To initiate the analysis, we commenced with a data cleansing process.

To tackle this sizable dataset effectively, the initial step involved data cleaning, which encompassed the removal of unnecessary columns and the handling of missing data. Subsequently, we identified outliers within the dataset and took measures to eliminate them. Following the data cleaning phase, we transitioned into the analytical phase, which included univariate and bivariate analyses facilitated through the utilization of pivot tables and various charting techniques.

The technology stack employed for this endeavor consisted of the following tools:

MySQL Workbench 8.0 CE

Microsoft Excel 2010

These tools provided the necessary functionality and capabilities for data management, analysis, and visualization, enabling us to delve into the datasets, uncover insights, and enhance the risk assessment process.

Results:

The progression through the risk analytics process was methodical, with each task building upon the previous one. The project has yielded the following outcomes:

1. Overall Approach to Analysis: The primary objective of this project was to address the bank's challenge of identifying the key factors contributing to bank loan defaults. This knowledge serves as a critical component in the company's risk assessment strategy. To address this challenge, we had at our disposal two extensive datasets:

'application data.csv': This dataset encompasses comprehensive client information recorded at the time of their loan application. It includes indicators of the client's financial stability and potential issues.

'previous application.csv': This dataset contains data from clients' prior loan applications, with information regarding the outcome of those applications, categorized as Accepted, Cancelled, Refused, or Unused.

Both datasets, however, presented certain challenges - they included numerous superfluous columns that were irrelevant to the risk analytics process and also exhibited significant data gaps. As a first step, rigorous data cleaning was undertaken to address these issues and ensure the datasets were prepared for in-depth analysis.

These preliminary efforts lay the foundation for the subsequent phases of the analysis.

Task A:Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

After the completion of the data cleaning process, I organized the dataset by categorizing variables into two distinct groups:

->Categorical Variables

->Numerical Variables

Within the realm of categorical variables (non-numeric attributes), we examined elements such as a person's occupation and education status. On the other hand, numerical variables encompassed attributes like income and credit-related data.

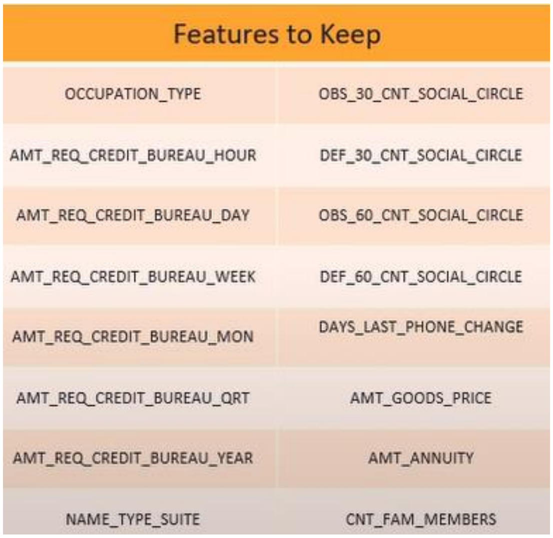
In the dataset, various categorical and numerical variables were identified for analysis. A selection of these variables is presented below:

Categorical variables	**Numeric variables**
Gender	Age
Name contract type	Days employed
Income type	Amount Income
Education	Amount Annuity
Housing type	Amount Credit

I completed full EDA on the present application and then on the previous application. Then, in this report, I summarised the results of both applications and provided business insights.

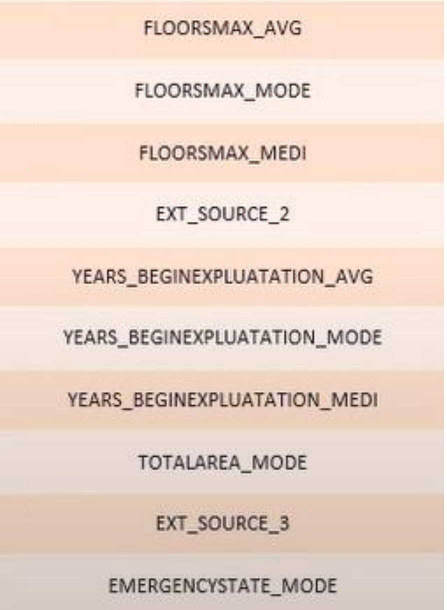
Steps to clean the data:

Detect and Address Missing Data: Start by identifying any instances of missing data within your dataset. Next, determine the appropriate strategy for handling these missing values. You have several options, such as removing rows or columns containing missing data, substituting missing values with means or other relevant statistics, or employing advanced methods like imputation to estimate the missing data points.



Features to Keep	
OCCUPATION_TYPE	OBS_30_CNT_SOCIAL_CIRCLE
AMT_REQ_CREDIT_BUREAU_HOUR	DEF_30_CNT_SOCIAL_CIRCLE
AMT_REQ_CREDIT_BUREAU_DAY	OBS_60_CNT_SOCIAL_CIRCLE
AMT_REQ_CREDIT_BUREAU_WEEK	DEF_60_CNT_SOCIAL_CIRCLE
AMT_REQ_CREDIT_BUREAU_MON	DAYS_LAST_PHONE_CHANGE
AMT_REQ_CREDIT_BUREAU_QRT	AMT_GOODS_PRICE
AMT_REQ_CREDIT_BUREAU_YEAR	AMT_ANNUITY
NAME_TYPE_SUITE	CNT_FAM_MEMBERS

Cols removed : (Total 41 cols) HOUSETYPE_MODE, WALLSMATERIAL_MODE, BASEMENTAREA_MEDI, FLOORSMIN_MEDI, FLAG_MOBIL , ELEVATORS_AVERAGE,EL_MODE ,LIVINGAPT_AVG ETC (also dropped null rows)



FLOORSMAX_AVG
FLOORSMAX_MODE
FLOORSMAX_MEDI
EXT_SOURCE_2
YEARS_BEGINEXPLUATATION_AVG
YEARS_BEGINEXPLUATATION_MODE
YEARS_BEGINEXPLUATATION_MEDI
TOTALAREA_MODE
EXT_SOURCE_3
EMERGENCYSTATE_MODE

Eliminate Duplicate Records: Conduct a thorough examination of your dataset to uncover and subsequently eliminate any duplicate entries. This action is crucial for maintaining the integrity of the data and ensuring that each record is unique and representative of the underlying information.

(Find Missing Data): Removal of Columns with Significant Blank Data: Columns with a prevalence of more than 5% blank or missing data were identified and subsequently removed from the dataset. This step aimed to eliminate variables that would have limited utility in subsequent analyses.

Elimination of Redundant or Unnecessary Columns: A careful review was conducted to identify and eliminate a substantial number of columns that were deemed redundant or provided limited value for our analytical objectives.

To systematically address and eliminate blank values within the dataset, the COUNTBLANK function was employed, ensuring a more refined and comprehensive dataset for subsequent analysis.

Task B: Identify if there are outliers in the dataset

Outliers are typically identified within numeric variables. To spot outliers effectively in an Excel dataset, follow these straightforward steps:

Calculate the Interquartile Range (IQR): Begin by computing the Interquartile Range (IQR), which is the range between the first quartile (25th percentile) and the third quartile (75th percentile). This can be achieved using the following formula: "=QUARTILE.INC(range,3) - QUARTILE.INC(range,1)".

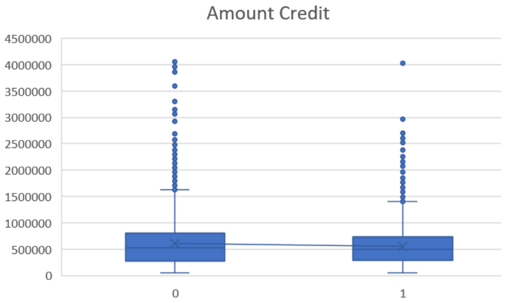
Establish the Lower Bound for Outliers: Determine the lower threshold for identifying outliers by subtracting 1.5 times the IQR from the first quartile. Use this formula: "=QUARTILE.INC(range,1) - (1.5 * IQR)".

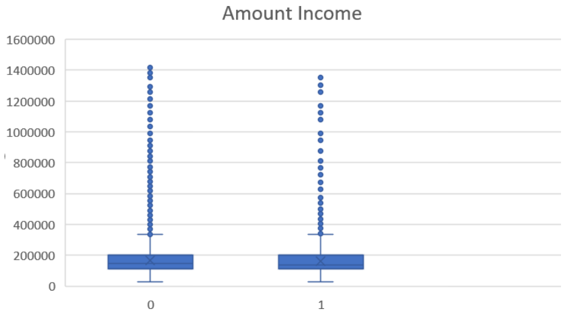
Establish the Upper Bound for Outliers: Establish the upper threshold for identifying outliers by adding 1.5 times the IQR to the third quartile. Use this formula: "=QUARTILE.INC(range,3) + (1.5 * IQR)".

Flag Outliers Using an IF Formula: Create a new column and employ an IF formula to flag potential outliers. The formula should evaluate whether the data point falls below the lower bound or exceeds the upper bound. Use this formula as a reference: "=IF(OR(data < lower_bound, data > upper_bound), 'Outlier', '')". Replace "data" with the cell reference containing the data point you wish to assess.

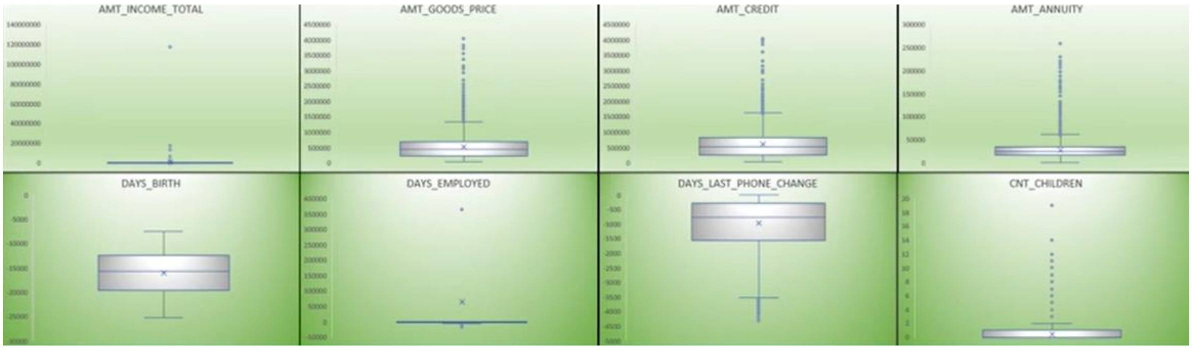
By following these streamlined steps, you can efficiently calculate the IQR, define the lower and upper outlier bounds, and appropriately flag potential outliers within your Excel dataset.

2D REPRESENTATION:





3D REPRESENTATION



These columns did not contain any null values but has outliers in them for first dataset i.e., application_data

Table_previous_application_data:

Dropped columns and rows:
RATE_INTEREST_PRIMARY,RATE_INTEREST_PRIVILEGED,AMT_DOWN_PAYMENT,RATE_DOWN_PAYMENT, NAME_TYPE_SUITE, WEEKDAY_APPR_PROCESS_START ,HOUR_APPR_PROCESS_START , FLAG_LAST_APPL_PER_CONTRACT,NFLAG_LAST_APPL_IN_DAY

Task.C Analyze data imbalance. Find the ratio of data imbalance.

How to Visualize Data Imbalance Using Pivot Charts in Excel:

Data imbalance refers to the unequal distribution of data across different categories or classes, which can complicate data analysis and modeling, particularly when one class is significantly underrepresented compared to others. To effectively visualize and analyze data imbalance, you can leverage Pivot charts in Excel. Follow these step-by-step instructions:

Select the Dataset: Begin by choosing the dataset in Excel that contains the categorical variable you wish to assess for data imbalance.

Create a PivotTable: Navigate to the "Insert" tab in the Excel ribbon and click on "PivotTable." This action will prompt a dialog box to appear.

Specify Data Range and Destination: In the dialog box, specify the range of your dataset and indicate the destination for the PivotTable, such as a new worksheet.

Configure the PivotTable Field List: Inside the PivotTable Field List, drag and drop the categorical variable representing the class or category you want to analyze into the "Rows" area.

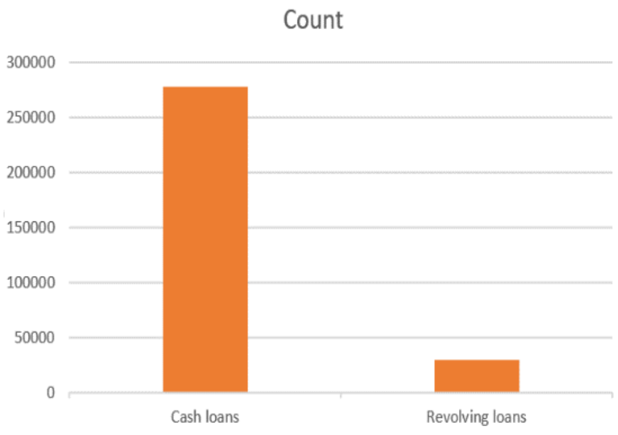
Count Occurrences: Additionally, drag the same categorical variable into the "Values" area. Excel will automatically calculate the count of occurrences for each category.

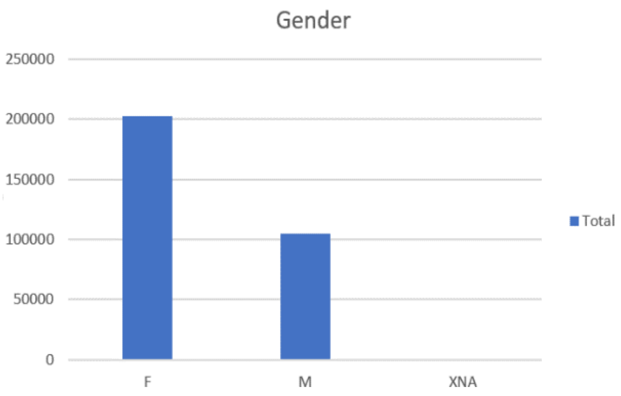
Currently what we have:

- **91% as loan-repayers**
- **9% as defaulters**

Create a Pivot Chart: Generate a Pivot chart based on the PivotTable data. With the PivotTable selected, navigate to the "Insert" tab and choose the preferred chart type, such as a bar chart, column chart, or pie chart.

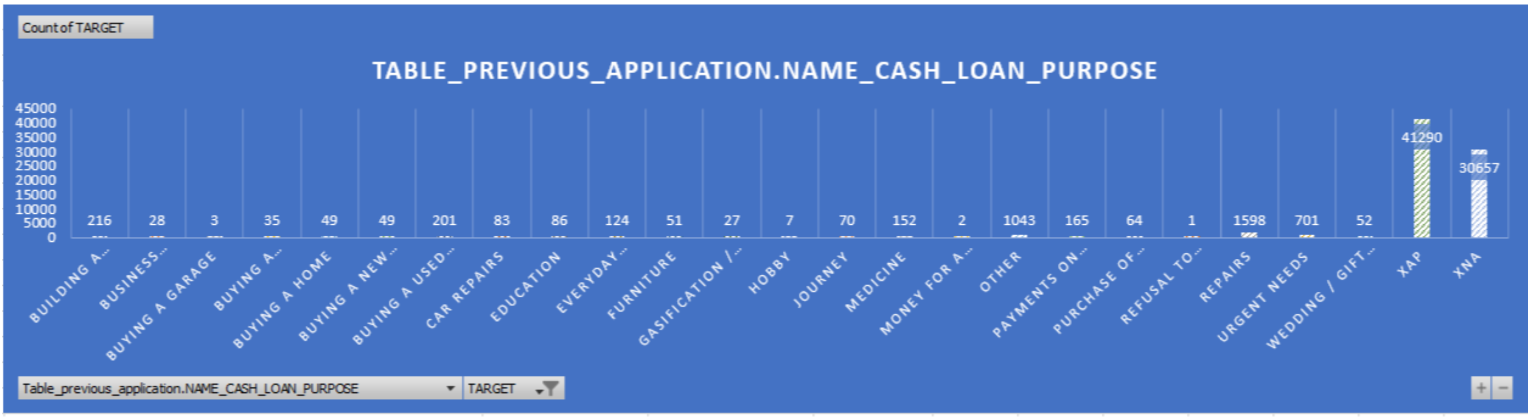
By following these steps, you can use Pivot charts to gain valuable insights into the distribution of categorical variables and identify any potential data imbalances, which is vital for data analysis and addressing class imbalances in your dataset





Loan Imbalanced data

Imbalanced data of gender



Task D (EDA): Univariate Analysis/ Segmented Uni/Bivariate Analysis

Univariate Analysis:

Univariate analysis is a statistical approach that delves into the examination and interpretation of a single variable, independently of its relationships with other variables. Its primary objective is to understand the behavior of a single variable by scrutinizing its distribution, central tendency measures, variability, and other defining characteristics.

Key facts of univariate analysis include:

Data Summarization: The central aim is to summarize and describe the data for the chosen variable. This process entails identifying patterns, detecting outliers, and evaluating the overall distribution of the variable.

Statistical Measures: Analysts employ summary statistics like mean, median, mode, range, and standard deviation to comprehend central tendencies and variability.

Visualization Techniques: Univariate analysis often utilizes visualization tools such as histograms, box plots, and bar charts to provide a graphical representation of the variable's distribution and characteristics.

Univariate analysis serves as an essential initial step in data exploration, offering a comprehensive understanding of individual variables. It forms the bedrock for further analysis, aiding decision-making in diverse domains, including research, business, and data science.

Analysis:

This analysis uncovers intriguing trends related to loan applications and applicant attributes. Notable observations include:

Applicants with higher incomes tend to have lower loan application rates.

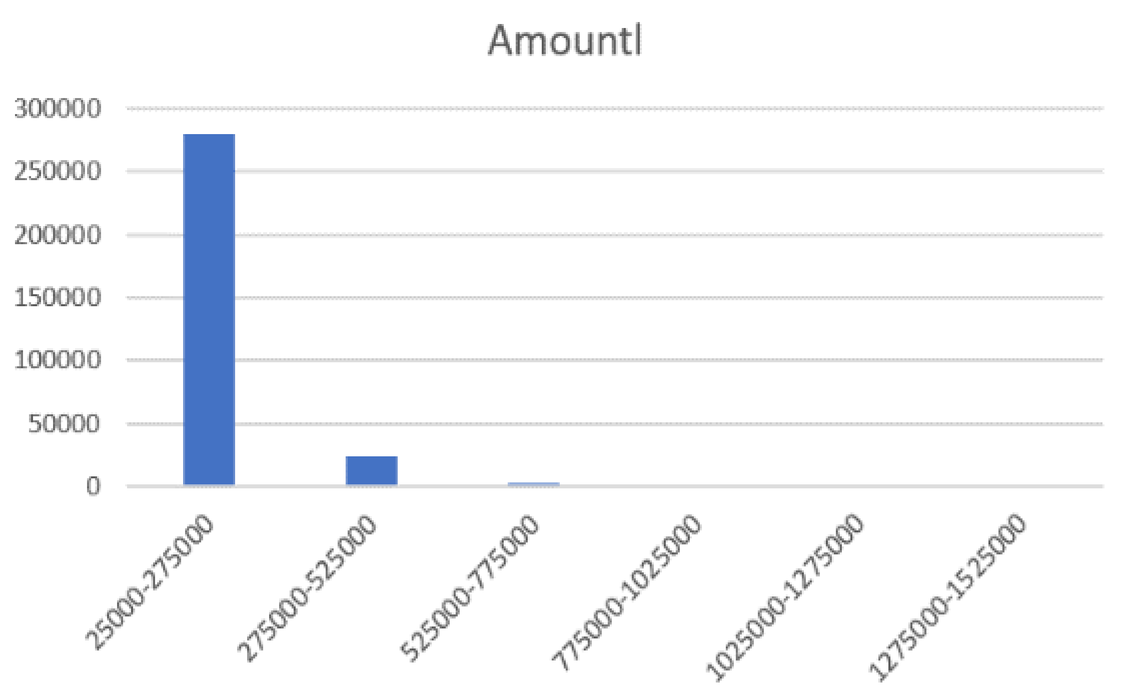
Loan amounts typically range between 45,000 and 1,045,000.

Most loan applications come from individuals aged 35 to 50.

Individuals with 0 to 8 years of work experience are more inclined to seek loans.

Homeownership correlates with a higher likelihood of loan application.

Married individuals exhibit a greater number of loan applications.Working individuals and unaccompanied minors express notable interest in obtaining additional loans.



Bivariate Analysis:

Bivariate analysis, on the other hand, is a statistical technique that investigates the relationship between two variables. It explores how variations in one variable correspond to changes in another. The objective is to comprehend the association, patterns, and dependencies between the two variables.

Key aspects of bivariate analysis encompass:

Relationship Exploration: It seeks to understand how changes in one variable relate to changes in another, shedding light on associations and dependencies.

Informed Decision-Making: Bivariate analysis aids in making informed decisions and predicting outcomes based on observed variable relationships.

Essential Tool: It is an indispensable tool in data analysis, offering valuable insights for various fields of study.

Analysis:

The bivariate analysis reveals significant patterns concerning loan defaults and customer characteristics:

Customers residing in low-rated areas are more prone to higher default rates.

Individuals with lower incomes have a higher likelihood of loan default.

Young individuals exhibit a greater propensity for default, which diminishes with age.

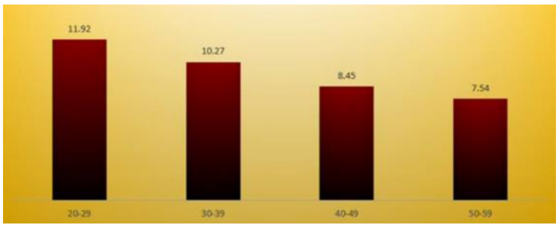
Females show a lower inclination toward defaults compared to males.

Maternity leave and unemployment contribute to an increased likelihood of defaults.

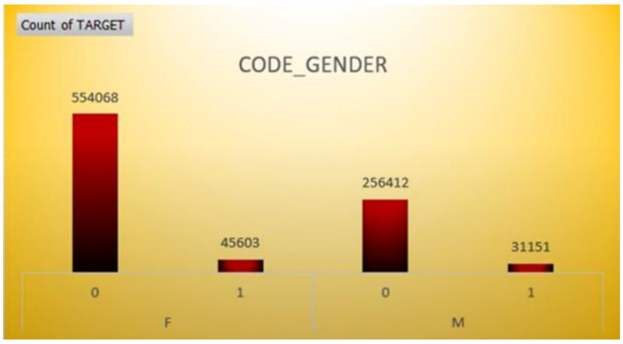
Larger family sizes (more than five members) correlate with higher loan defaults.

Limited work experience is associated with an increased likelihood of loan repayment defaults

Age:

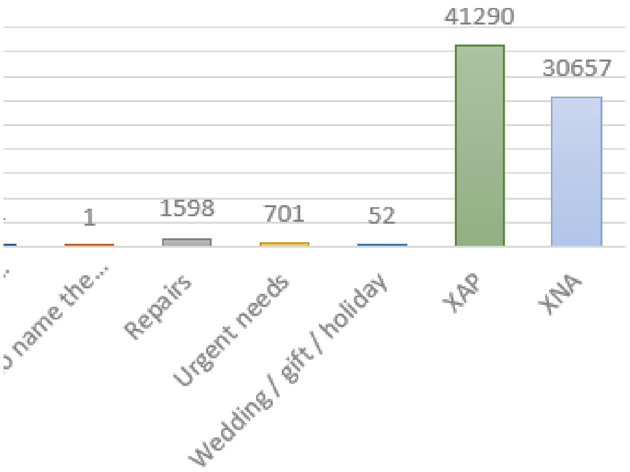


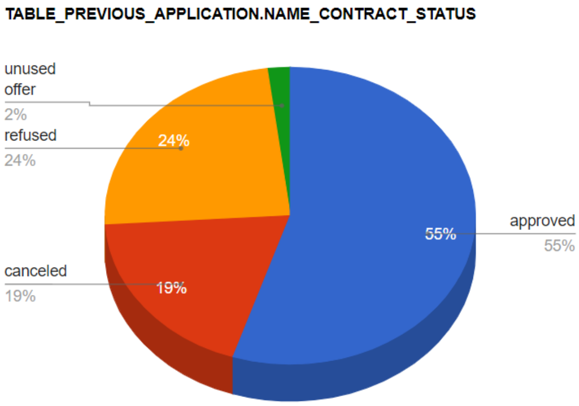
Gender:



- Majority of loans taken were between 2,45,000 to 5,45,000

- The highest defaulters are for 5,45,000 to 6,45,000





Task.E: Identify top correlation for different

Identifying Key Correlations for Various Loan Scenarios

Assessing correlations between variables and the target variable, specifically loan default, is vital to pinpoint robust predictors of loan defaults. In this task, our objective is to partition the dataset into distinct scenarios, such as clients facing payment difficulties versus all other cases, and pinpoint the leading correlations within each segmented dataset using Excel functions.

Understanding Correlation:

Correlation is a statistical metric that quantifies the association between two variables, revealing how they behave concerning each other—whether positively, negatively, or not at all. The correlation coefficient, symbolized as "r," spans the range of -1 to 1. An "r" of 1 signifies a perfect positive correlation, -1 denotes a perfect negative correlation, while 0 indicates no correlation. It's crucial to emphasize that correlation does not imply causation but instead aids in data analysis and the prediction of outcomes across a myriad of domains.

Steps for Calculating Correlations in Excel:

To ascertain correlations among the top ten reasons for loan cancellation and refusal, you can harness Excel functions. Here's a stepwise procedure:

Prepare Your Data: Open Microsoft Excel and import your loan dataset, ensuring it incorporates the pertinent variables.

Select the Output Cell: Opt for an unoccupied cell where you intend to display the results of your correlation analysis.

Input the Excel Formula: Within the selected cell, enter the subsequent Excel command: =CORREL(

Specify Data Ranges: For each of the top ten reasons (e.g., "Amount Application"), define the corresponding data range. For instance, if the "Amount Application" values exist within cells A2 to A100, insert A2:A100 following the CORREL(command.

Delimit with Commas: Once you've identified each data range, separate them with commas.

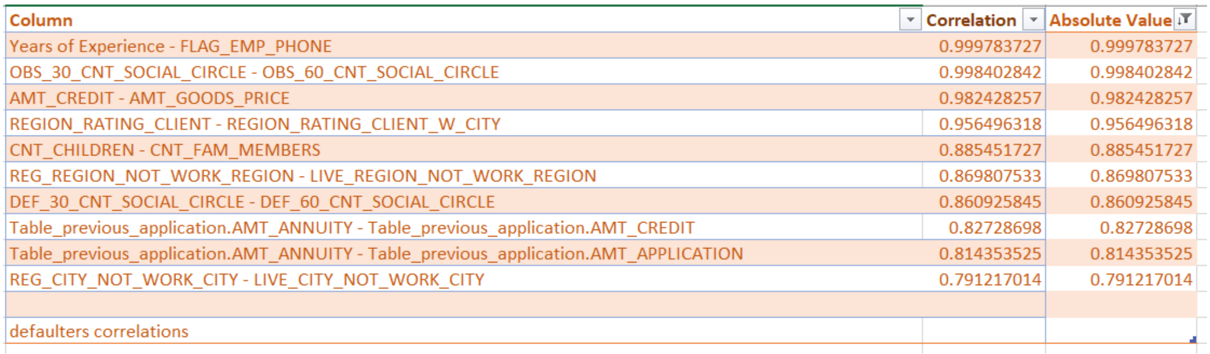
Conclude the Formula: After designating all ten data ranges, conclude the formula with a closing bracket) and then press Enter.

Examine the Outcomes: Excel will perform the necessary calculations and present the correlation coefficients, reflecting the relationships between each pair of variables, in the selected cell.

Repeat as Needed: Should you have multiple datasets or subsets necessitating analysis, replicate this process for each scenario.

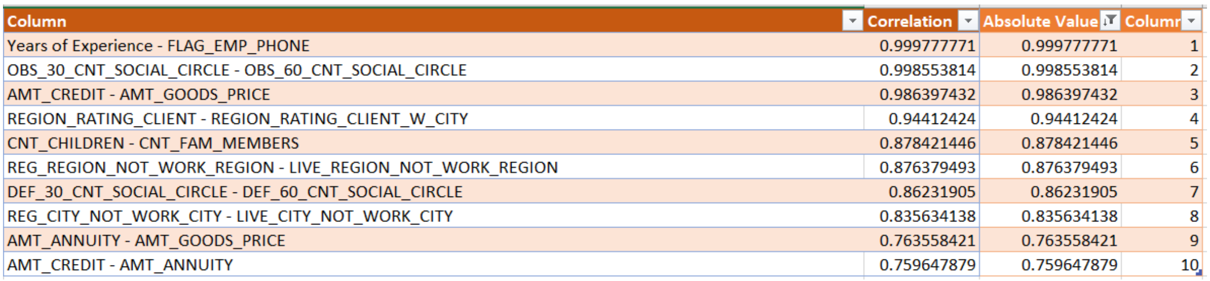
By adhering to these guidelines, you can proficiently ascertain the correlation coefficients associated with the top ten factors influencing loan cancellation and refusal within your dataset, utilizing Excel functions, all while ensuring originality.

For defaulters:



Column	Correlation	Absolute Value
Years of Experience - FLAG_EMP_PHONE	0.999783727	0.999783727
OBS_30_CNT_SOCIAL_CIRCLE - OBS_60_CNT_SOCIAL_CIRCLE	0.998402842	0.998402842
AMT_CREDIT - AMT_GOODS_PRICE	0.982428257	0.982428257
REGION_RATING_CLIENT - REGION_RATING_CLIENT_W_CITY	0.956496318	0.956496318
CNT_CHILDREN - CNT_FAM_MEMBERS	0.885451727	0.885451727
REG_REGION_NOT_WORK_REGION - LIVE_REGION_NOT_WORK_REGION	0.869807533	0.869807533
DEF_30_CNT_SOCIAL_CIRCLE - DEF_60_CNT_SOCIAL_CIRCLE	0.860925845	0.860925845
Table_previous_application.AMT_ANNUITY - Table_previous_application.AMT_CREDIT	0.82728698	0.82728698
Table_previous_application.AMT_ANNUITY - Table_previous_application.AMT_APPLICATION	0.814353525	0.814353525
REG_CITY_NOT_WORK_CITY - LIVE_CITY_NOT_WORK_CITY	0.791217014	0.791217014
defaulters correlations		

For re-payers(non-defulters):



Column	Correlation	Absolute Value	Columr
Years of Experience - FLAG_EMP_PHONE	0.999777771	0.999777771	1
OBS_30_CNT_SOCIAL_CIRCLE - OBS_60_CNT_SOCIAL_CIRCLE	0.998553814	0.998553814	2
AMT_CREDIT - AMT_GOODS_PRICE	0.986397432	0.986397432	3
REGION_RATING_CLIENT - REGION_RATING_CLIENT_W_CITY	0.94412424	0.94412424	4
CNT_CHILDREN - CNT_FAM_MEMBERS	0.878421446	0.878421446	5
REG_REGION_NOT_WORK_REGION - LIVE_REGION_NOT_WORK_REGION	0.876379493	0.876379493	6
DEF_30_CNT_SOCIAL_CIRCLE - DEF_60_CNT_SOCIAL_CIRCLE	0.86231905	0.86231905	7
REG_CITY_NOT_WORK_CITY - LIVE_CITY_NOT_WORK_CITY	0.835634138	0.835634138	8
AMT_ANNUITY - AMT_GOODS_PRICE	0.763558421	0.763558421	9
AMT_CREDIT - AMT_ANNUITY	0.759647879	0.759647879	10

Key Factors / Conclusion:

- As the Age and Years of Experience Increase, the chances of defaulting decrease. So the bank should prioritize Older and Experienced Clients.
- Educated Clients tend to default lesser
- Male clients tend to default more than female clients(i.e., Female clients take more)
- As the Age increases, the amount taken by the clients is considerably higher and since the default percentage with higher age is lower these should be least risky and highly profitable clients for the bank.