Bank Loan Case Study

- **Project Description:** The objective of this case study is to identify patterns that indicate a client's likelihood of having difficulty repaying their instalments. These insights can be leveraged to take preventive actions, such as denying loans, reducing the loan amount, or offering higher interest rates to risky applicants.
  1. **Clients with Payment Difficulties:** These clients had late payments exceeding X days for at least one of the first Y instalments of the loan.
  2. **Clients Without Payment Difficulties:** These are cases where all payments were made on time.

Loan Decisions: When a client applies for a loan, one of four decisions is typically made:

- **Approved:** The loan is approved and disbursed.
- **Cancelled:** The loan is cancelled by the client.
- **Refused:** The loan application is rejected by the lender.
- **Unused Offer:** The loan is approved, but the client does not proceed with it.

Analytical Goal:

By employing **Exploratory Data Analysis (EDA)**, we aim to understand how consumer demographics, financial attributes, and loan characteristics influence the likelihood of a loan default. Key variables that may impact this analysis include:

- **Consumer Attributes:**
  - Age
  - Income level
  - Employment status
  - Credit history
  - Number of dependents
- **Loan Attributes:**
  - Loan amount
  - Interest rate

- Loan duration
- Payment frequency
- Collateral provided (if any)

- **Approach:** The project was successfully executed through a well-structured approach involving data cleaning, EDA, feature engineering, predictive modelling, and visualization. This workflow allowed for the identification of key risk factors contributing to loan defaults and provided actionable insights for decision-making.

- **Tech-Stack**
  Microsoft Excel: For initial data review, performing quick descriptive analysis, and visualizing smaller datasets with pivot tables and charts.

  In this project, the two major data sheets, Application Data and Previous Application Data, are central to the analysis:
  1. **Application Data:**
     - Contains details about current loan applications, including consumer attributes (e.g., age, income) and loan characteristics (e.g., loan amount, interest rate).
  2. **Previous Application Data:**
     - Holds information about clients' past loan applications, including details such as approval status, previous payment behavior, and any loan defaults.
  3. **Merge Data:**
     - The combined dataset created by merging Application Data with Previous Application Data.
     - This merged dataset allows for a comprehensive view of each client's current and past loan behavior, improving the analysis of loan default risk.

- **Data Understanding**

  In this project first we understand the data which is important because of when we understand the data then and then we apply the operations. If the data is noisy and unclean and this data, they consist the blanks values like missing values then our output is not giving properly and our analysis getting worst so, for that we apply all operations for the cleaning the data.

- **Insights:**

A. Data Analytics Tasks: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

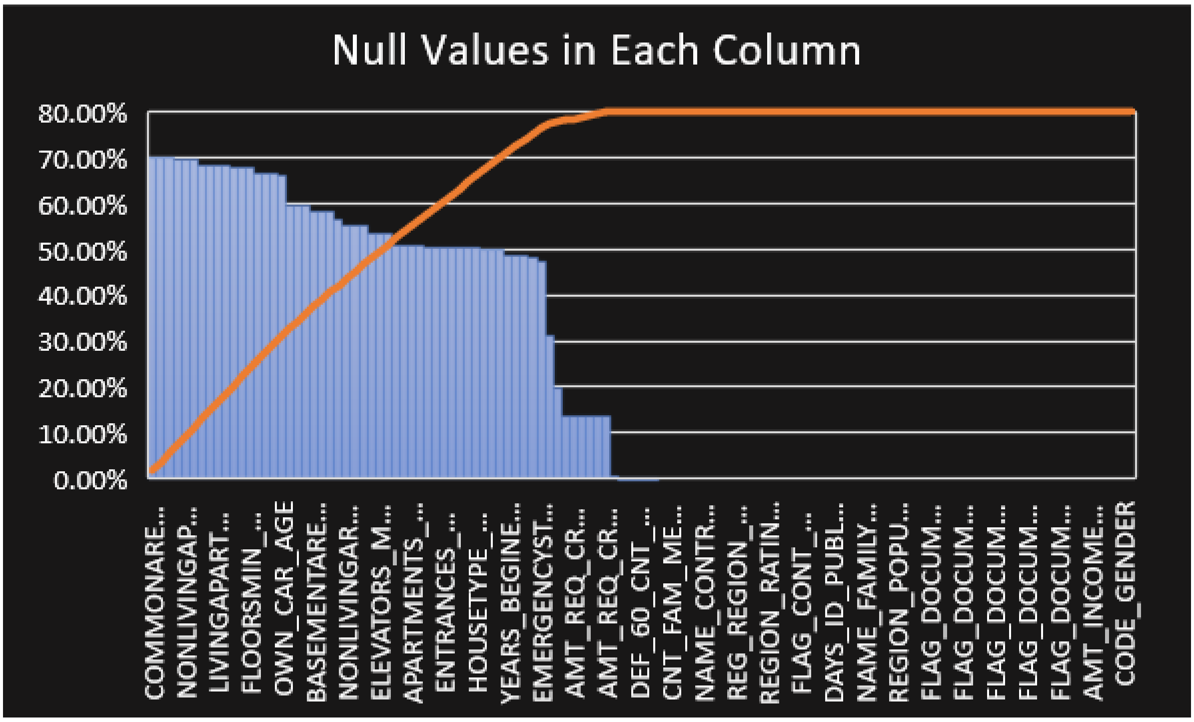
- ✓ **Conclusion:**
- ✍ **Application Data**

This sheet consists the 122 columns which is very huge for our analysis we won't all the columns so we remove the columns which the consist the large percent of missing data we give the criteria for that like greater than 30% we remove these columns. So, we get only 75 columns which they consist less than 30% missing data.

Less Than 30% Missing data

SK_ID_CURR
TARGET
NAME_CONTRACT_TYPE
CODE_GENDER
FLAG_OWN_CAR
FLAG_OWN_REALTY
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
NAME_TYPE_SUITE
NAME_INCOME_TYPE
NAME_EDUCATION_TYPE
NAME_FAMILY_STATUS
NAME_HOUSING_TYPE
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_EMPLOYED(YRS)
DAYS_REGISTRATION
DAYS_REGISTRATION(YRS)
DAYS_ID_PUBLISH
DAYS_ID_PUBLISH(YRS)
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE

FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
WEEKDAY_APPR_PROCESS_START
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
ORGANIZATION_TYPE
EXT_SOURCE_2
EXT_SOURCE_3
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
FLAG_DOCUMENT_2
FLAG_DOCUMENT_3
FLAG_DOCUMENT_4
FLAG_DOCUMENT_5
FLAG_DOCUMENT_6
FLAG_DOCUMENT_7
FLAG_DOCUMENT_8
FLAG_DOCUMENT_9
FLAG_DOCUMENT_10
FLAG_DOCUMENT_11
FLAG_DOCUMENT_12
FLAG_DOCUMENT_13
FLAG_DOCUMENT_14
FLAG_DOCUMENT_15
FLAG_DOCUMENT_16
FLAG_DOCUMENT_17
FLAG_DOCUMENT_18
FLAG_DOCUMENT_19
FLAG_DOCUMENT_20
FLAG_DOCUMENT_21
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



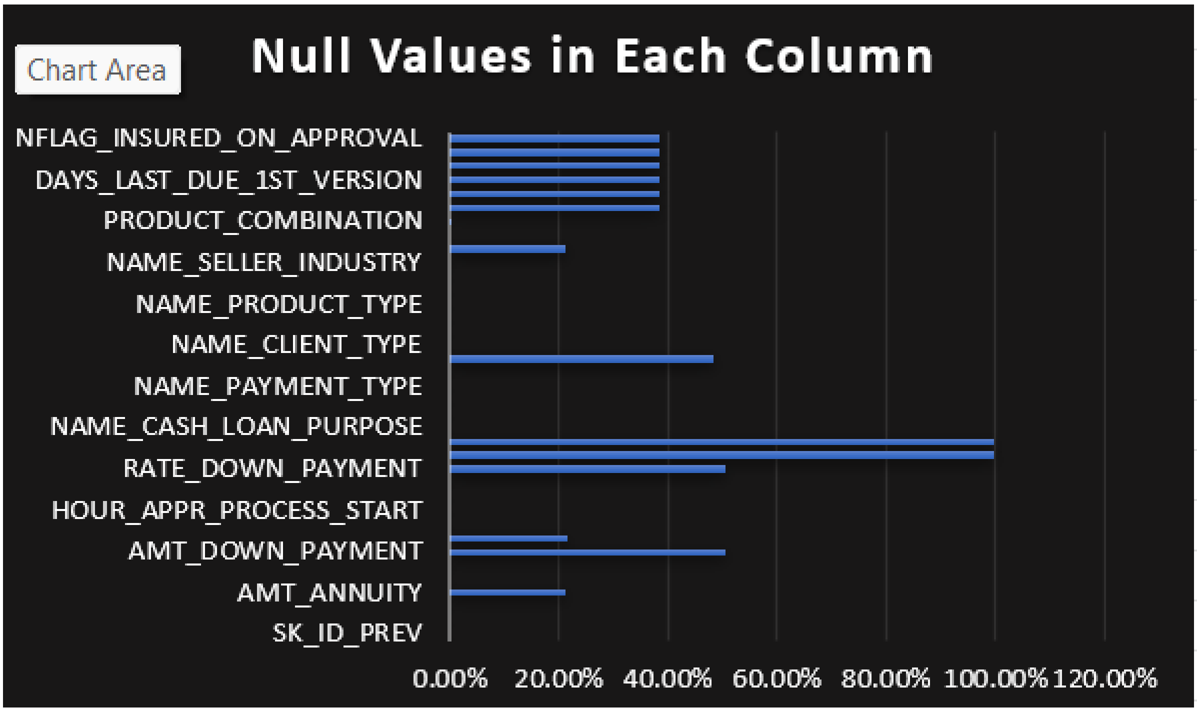
✍ Previous Application Data

This sheet consists the 37columns which is very huge for our analysis we won't all the columns so we remove the columns which the consist the large percent of missing data we give the criteria for that like greater than 30% we remove these columns. So, we get only 26 columns which they consist less than 30% missing data.

Less Than 30% Missing data

SK_ID_PREV
SK_ID_CURR
NAME_CONTRACT_TYPE
AMT_ANNUITY
AMT_APPLICATION
AMT_CREDIT
AMT_GOODS_PRICE
WEEKDAY_APPR_PROCESS_START
HOUR_APPR_PROCESS_START
FLAG_LAST_APPL_PER_CONTRACT
NFLAG_LAST_APPL_IN_DAY
NAME_CASH_LOAN_PURPOSE
NAME_CONTRACT_STATUS
DAYS_DECISION
NAME_PAYMENT_TYPE

CODE_REJECT_REASON
NAME_CLIENT_TYPE
NAME_GOODS_CATEGORY
NAME_PORTFOLIO
NAME_PRODUCT_TYPE
CHANNEL_TYPE
SELLERPLACE_AREA
NAME_SELLER_INDUSTRY
CNT_PAYMENT
NAME_YIELD_GROUP
PRODUCT_COMBINATION



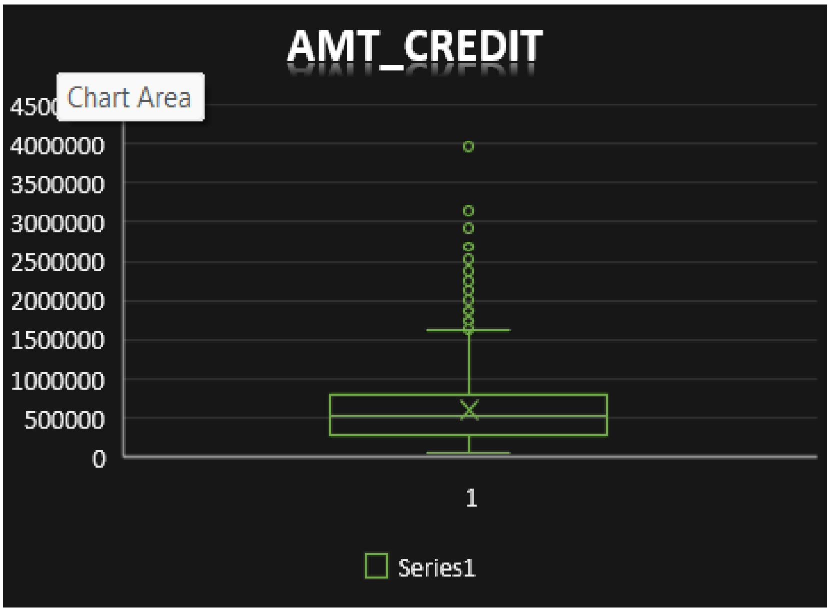
B. Identify Outliers in the Dataset: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

- ✓ **Conclusion:** Data points that fall above the upper limit are classified as upper outliers, while those below the lower limit are classified as lower outliers. These outliers deviate significantly from the central distribution of the dataset and may warrant further investigation or exclusion depending on the analysis objective.

  This method is particularly useful for detecting extreme values in skewed data distributions, as it is based on the IQR, which is resistant to the influence of outliers themselves.

This Is for the Column AMT_CREDIT

AMT_CREDIT	
Quartile 1	270000
Quartile 2/ Median	514777.5
Quartile 3	808650
Inter Quartile Range	538650
upper limit	1616625
lower limit	-537975



This is for AMT_GOODS_PRICE

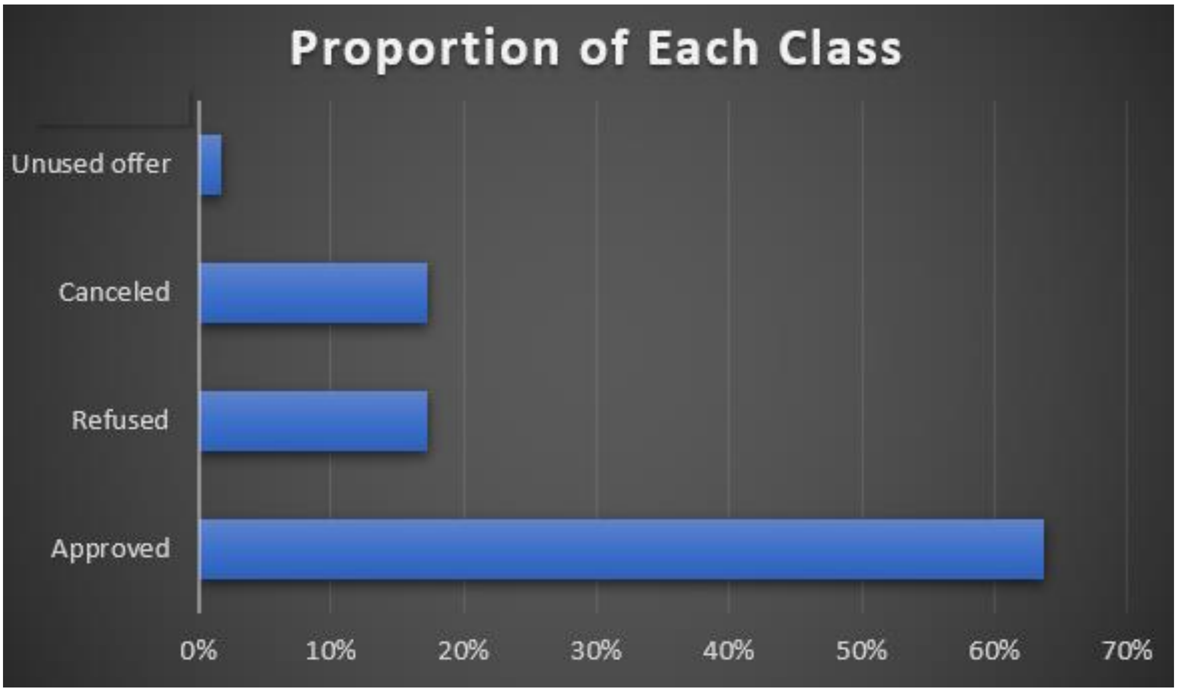
AMT_GOODS_PRICE	
Quartile 1	238500
Quartile 2/ Median	450000
Quartile 3	679500
Inter Quartile Range	441000
upper limit	1341000
lower limit	-423000

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

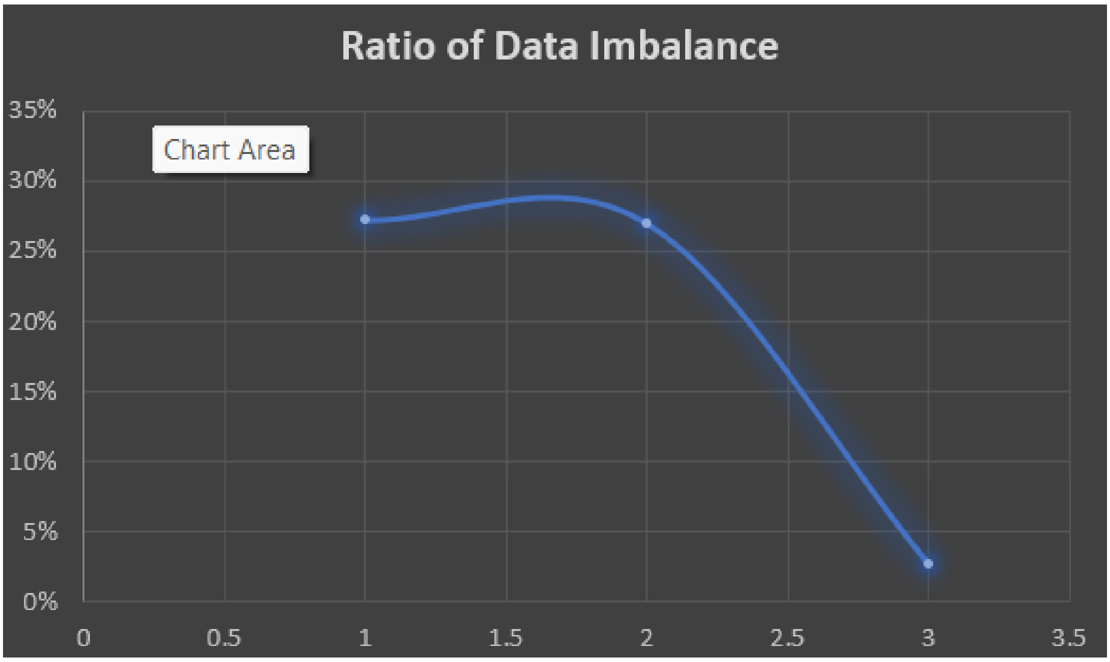
- ✓ **Conclusion:** By analyzing the distribution of categories within a specific attribute, we can identify imbalanced classes and take appropriate measures to mitigate their impact. This step is crucial for ensuring that models do not disproportionately Favor the majority class, thereby enhancing the robustness of the analysis or predictive modeling.

Count the Occurrences of Each Class	
Approved	31885
Refused	8660
Canceled	8595
Unused offer	859

Proportion of Each Class	
Approved	64%
Refused	17%
Canceled	17%
Unused offer	2%



Ratio of Data Imbalance	
Refused	27%
Canceled	27%
Unused offer	3%



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

✓ **Conclusion:** In this analysis, we performed three types of statistical evaluations — Univariate Analysis, Segmented Univariate Analysis, and Bivariate Analysis — to extract meaningful insights from the dataset and assess variable distributions, relationships, and their impact on the target variable.

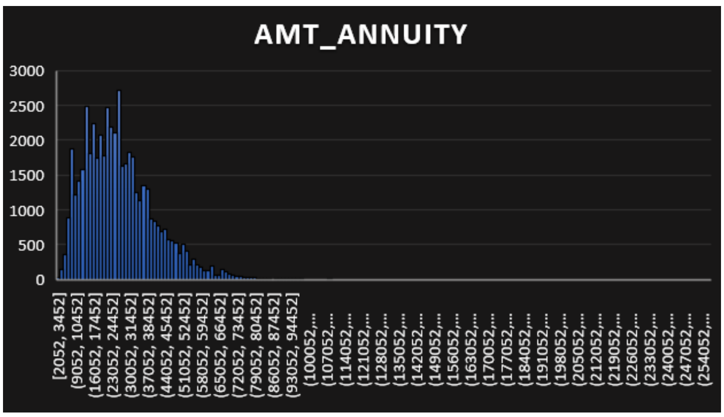
1. Univariate Analysis:

Objective: To examine the distribution of individual variables independently.

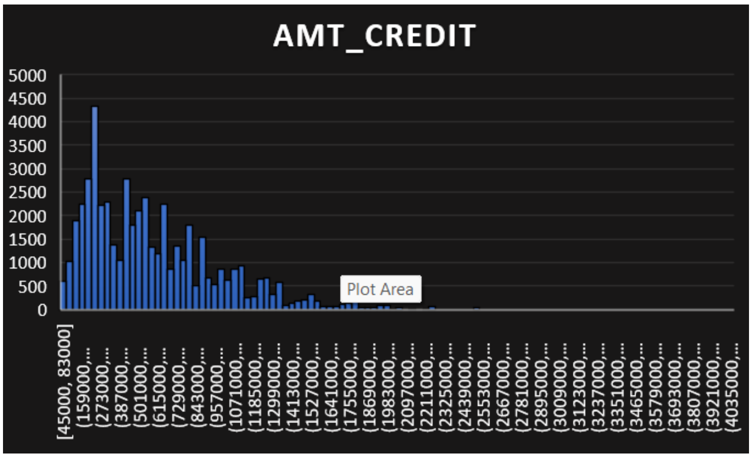
- We calculated measures like **mean, median, mode, variance, standard deviation**, and **range** to understand the central tendency, spread, and shape of the distribution for each variable.
- For categorical variables, we analyzed the **frequency distribution** to identify the most common categories.
- We used histograms and bar charts to visually represent the distribution of variables, enabling us to identify skewness, outliers, and data spread.

✓ **Conclusion**: This analysis helped us identify the **distribution pattern** of each variable, uncover outliers, and determine whether the data was normally distributed or skewed. These insights are critical for further analysis, especially for predictive modeling.

AMT_ANNUITY	
Mean	27107.37736
Median	24939
Mode	9000
Variance	212075108.9
Standard Deviation	14562.7988
Max	258025.5
Min	2052
Range	255973.5



AMT_CREDIT	
Number of loans	49999
Average loan amou	599701
Median loan amour	514778
Standard deviation	402411



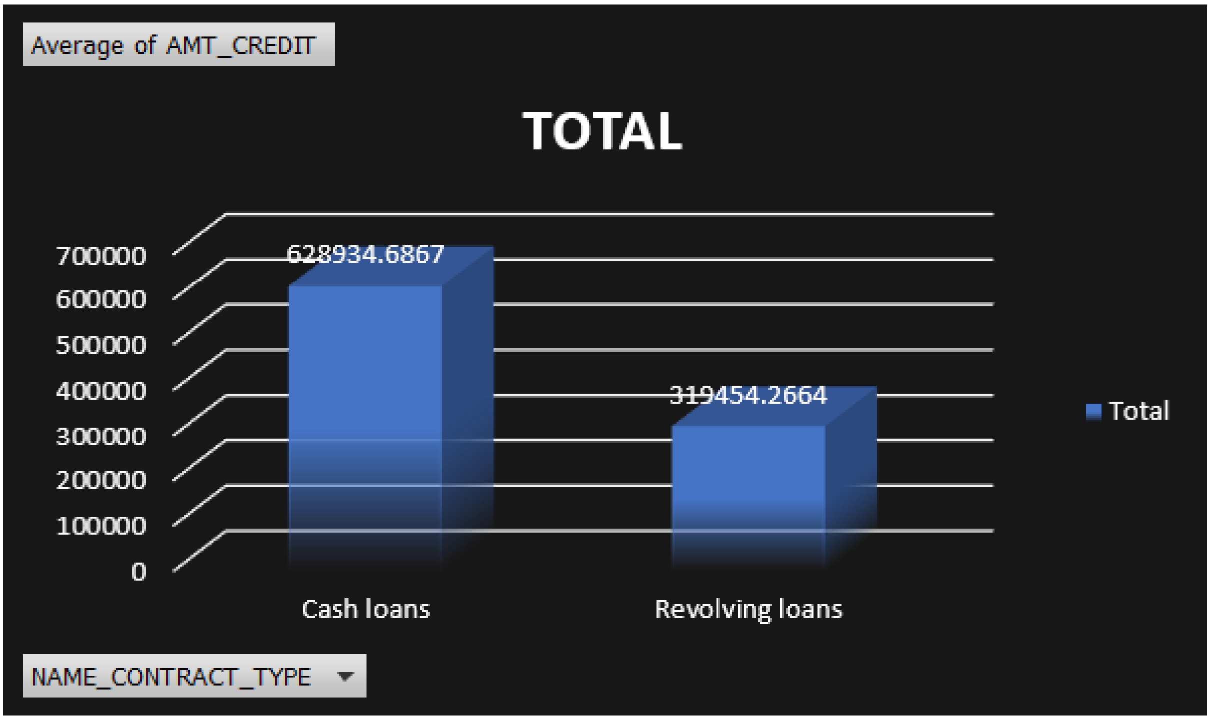
2. Segmented Univariate Analysis:

Objective: To compare the distribution of a specific variable across different **subgroups or categories** (e.g., different segments or conditions).

- We divided the data into different groups based on a **categorical feature** (such as gender, region, or income level) and performed univariate analysis for each subgroup.
- Using Excel's **pivot tables**, **filters**, and **charts**, we compared the distribution of key variables across these segments.
- For each subgroup, we evaluated statistics like the **mean, median, and standard deviation** to see how the distribution changes across different categories.

✓ **Conclusion**: Segmented univariate analysis allowed us to **compare variable behavior** under different conditions. This identified segments with significant variation or trends (e.g., higher income groups have different spending habits) and revealed key drivers for different target outcomes.

CONTRACT_TYPE	Average of AMT_CREDIT
Cash loans	628934.6867
Revolving loans	319454.2664
Grand Total	**599700.5815**



3. Bivariate Analysis:

Objective: To explore relationships between two variables, especially focusing on how **independent variables** relate to the **target variable**.

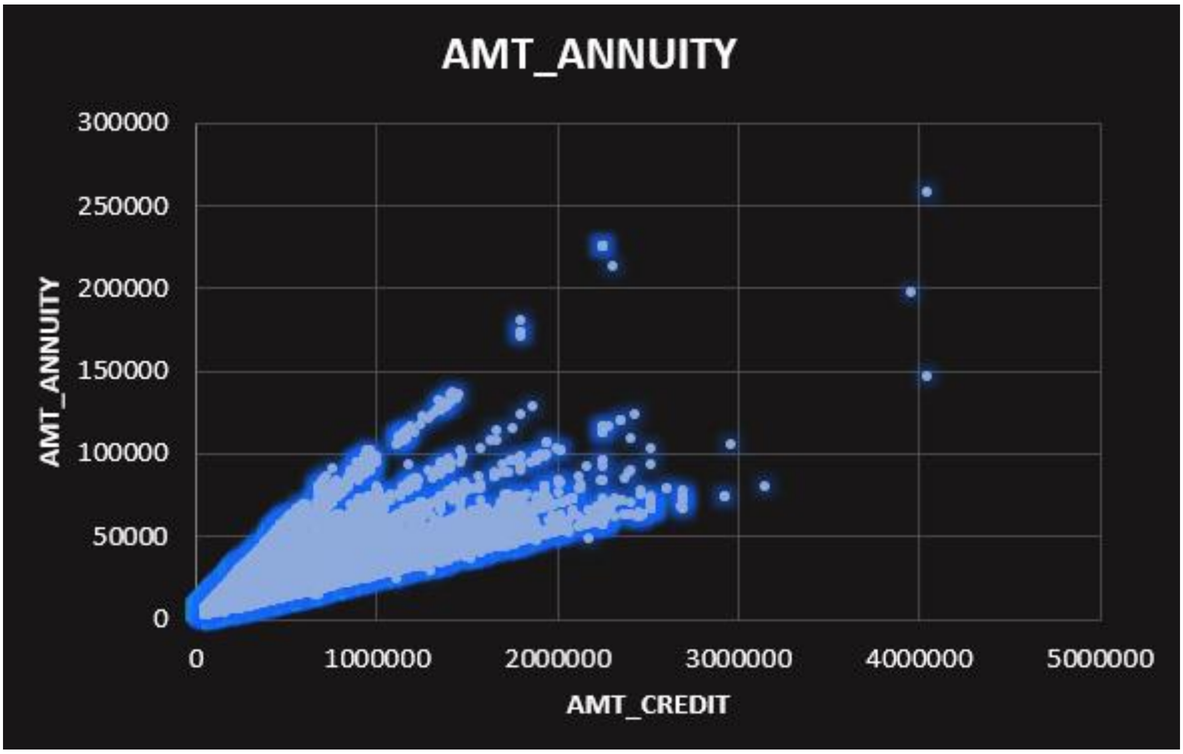
- We used **scatter plots, correlation coefficients, and pivot tables** to study relationships between continuous variables.
- **Correlation analysis** was performed to quantify the strength and direction of relationships between numerical variables. We used Pearson's correlation coefficient to identify **positive, negative, or no correlation**.
- For categorical variables, we used **cross-tabulation** and **stacked bar charts** to examine associations between different categories and the target variable.

- We also explored **relationships between numerical and categorical variables** by comparing group means and visualizing the differences through box plots.

Conclusion: Bivariate analysis highlighted important relationships between the variables and the target outcome. For example, we identified which variables have the strongest correlation with the target variable, indicating potential predictive features. Additionally, it revealed any significant **interaction effects** or **dependencies** between different variables, helping guide feature selection for more complex modeling.

Correlation between loan amount and annuity (strong positive correlation).

Which is being **76.95%**



E. **Identify Top Correlations for Different Scenarios:** 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.

- ✓ **Conclusion:** In this analysis, we segmented the dataset based on different scenarios, such as clients with payment difficulties and all other cases, to identify the top correlations within each segment. This approach helps in understanding how variables behave differently under specific conditions and reveals important relationships that may not be evident when analyzing the entire dataset as a whole.

Column	Correlation
AMT_GOODS_PRICE	0.98694373017159
REGION_RATING_CLIENT_W_CITY	0.950710179345493
LIVE_REGION_NOT_WORK_REGION	0.857141676571727
LIVE_CITY_NOT_WORK_CITY	0.82158378872428
Previous AMT_APPLICATION	0.812970626257039
Previous AMT_CREDIT	0.975771048697365
Previous AMT_GOODS_PRICE	0.993495986186513

- **Specific Relationships**

Loan Amount vs. Goods Price
98.7%
Loan Amount vs. Annuity Amount
76.95%
Application Amount vs. Credit Amount
97.58%

Column1	AMT_CREDI	AMT_ANNUIT	AMT_GOODS_PRIC	AMT_CREDIT	AMT_APPLICATIO
AMT_CREDIT	1				
AMT_ANNUITY	0.769498914	1			
AMT_GOODS_PRICE	0.98694373	0.774433947	1		
AMT_CREDIT	0.002306407	0.002788987	0.003423709	1	
AMT_APPLICATION	0.002719648	0.002917718	0.003696807	0.975771049	1

➤ Key Insights

1. **Cash loans** (with a higher average) are likely used for **larger financial commitments**, while **revolving loans** (with a lower average) are used for **short-term or recurring financial needs**. The distinction between the two types of loans is important for financial institutions when assessing risk, understanding customer behavior, and designing loan products tailored to specific financial situations.
2. Female borrowers (32,823) significantly outnumber male borrowers (17,174), representing nearly double the number of male clients. This shows that the financial institution's customer base is skewed heavily toward women.
3. Male borrowers make up a smaller share of the total dataset, suggesting either lower engagement with the loan products or differences in financial behavior or access to credit.
4. The gender distribution is **imbalanced**, with females being the dominant group. This could influence product design, marketing strategies, and financial services tailored to female clients.
5. The near-perfect correlation between application and credit amounts shows that the institution's credit approval process is well-calibrated to borrower requests, ensuring **efficient processing** and **appropriate loan approval**.

➤ Result

This project required extensive use of Microsoft Excel to process and analyze large datasets. A significant challenge was managing the volume of data, which provided valuable insights into handling large-scale datasets efficiently. The project involved merging two distinct datasets to perform in-depth analyses, improving my understanding of data integration techniques.

Dealing with missing data and outliers was a critical aspect of the project. I learned the technical approaches for identifying, handling, and imputing missing values, as well as detecting and addressing outliers using statistical methods. This project enhanced my ability to apply data-cleaning techniques, understanding the "what," "how," and "why" behind handling outliers and null values to ensure the integrity of the analysis.

Additionally, I explored advanced features in Excel, such as the Data Analysis Toolpak and other add-ins, which significantly improved the efficiency and depth of the analysis. This hands-on experience solidified my knowledge of data preprocessing, merging, and analysis in Excel.

Excel Sheet link:
https://docs.google.com/spreadsheets/d/1SKSumdO0yNE6i55QBbQ57UxZfY9nSW8j/edit?usp=sharing&ouid=116406143301160000153&rtpof=true&sd=true

Thank You