# Bank Loan Case Study

Final Project-2

Yashi Gupta

# Project Description

As a data analyst at a finance company specializing in lending to urban customers, my role is to conduct **Exploratory Data Analysis (EDA)** to understand customer behavior and loan repayment patterns. The company faces a challenge where applicants with **insufficient credit history** sometimes exploit the system, leading to **loan defaults**. This increases financial risk and affects overall profitability. The goal of this project is to analyze loan application data to identify key indicators of default and help the company make better lending decisions.

### What We Are Going to Do

1. *Data Cleaning & Preparation*

    - Load and inspect the dataset to understand its structure.
    - Handle missing values, duplicates, and inconsistencies.
    - Perform data transformations to ensure usability.

2. *Exploratory Data Analysis (EDA)*

    - Analyze distributions of **customer demographics** (age, income, employment type, etc.).
    - Examine **loan characteristics** (amount, interest rates, tenure).
    - Compare **payment behavior** across different customer segments.
    - Identify patterns in customers who defaulted vs. those who paid on time.
    - Explore correlations between **loan approval outcomes** and customer attributes.

3. *Segmentation of Loan Applicants*

    - Classify customers into **high-risk** and **low-risk** categories.

- Evaluate how different loan types and amounts impact repayment behavior.
- Identify key risk factors that contribute to **late payments and defaults**.

*4. Visualization & Insights*

- Use **charts, graphs, and heatmaps** to present findings.
- Highlight trends in **loan defaults, approval rates, and customer profiles**.
- Provide recommendations based on data-driven insights.

5. Business Impact & Decision Support

- Develop strategies to **reduce loan default risk** (e.g., adjusting approval criteria, modifying loan terms, or implementing risk-based pricing).
- Ensure the company does not reject **creditworthy applicants unnecessarily**.
- Assist in optimizing lending policies to balance **profitability and risk management**.

By the end of this analysis, we will have a clearer understanding of **what factors influence loan defaults** and how the company can make more informed decisions to enhance its loan approval process.

# Approach

The dataset contains detailed records of loan applications, including customer demographics, financial background, and loan repayment history. Customers fall into two main categories:

1. **Customers with payment difficulties** – Those who have missed or delayed payments beyond a certain threshold.
2. **Customers without payment issues** – Those who have consistently paid on time.

Each loan application results in one of four outcomes:

- **Approved** – The loan is granted.
- **Cancelled** – The customer withdraws the application during the process.
- **Refused** – The application is rejected.
- **Unused Offer** – The loan is approved, but the customer does not utilize it.

The workflow consists of three main stages: **Data Cleaning, Data Analysis, and Data Visualization**.

### 1. Data Cleaning

Before performing any analysis, it is crucial to clean the dataset to ensure accuracy and reliability.

#### Step 1: Identifying and Handling Missing Data

- Used COUNTBLANK(range) to identify missing values in each column.
- Applied IF(ISBLANK(cell), "Missing", "Present") to flag missing data points.
- Dealt with missing values based on data type:
    - **Numerical data**: Replaced missing values with the **median** (MEDIAN(range)) to avoid skewness.
    - **Categorical data**: Filled missing values with the most frequently occurring category (MODE(range)).
- **Visualization**: Created a **bar chart** to show the proportion of missing values across different columns.

#### Step 2: Identifying and Handling Outliers

- Used **Interquartile Range (IQR) Method**:

    - Q1 = QUARTILE(range, 1), Q3 = QUARTILE(range, 3), IQR = Q3 - Q1.
    - Defined outliers as values outside **Q1 - 1.5*IQR** or **Q3 + 1.5*IQR**.
- Highlighted outliers using **Conditional Formatting** in Excel.
- **Visualization**: Created **box plots** to identify extreme values in loan amounts, applicant income, and credit history.

**2. Data Analysis**

This stage focuses on identifying trends, relationships, and patterns in loan default scenarios.

**Step 3: Analyzing Data Imbalance**

- Used COUNTIF(range, condition) to count occurrences of each loan status (Approved, Refused, etc.).
- Calculated the **class distribution percentage** to check for imbalance.
- **Visualization**:
    - **Pie chart** showing the distribution of loan approval statuses.
    - **Bar chart** comparing the proportion of defaulters vs. non-defaulters.

**Step 4: Univariate Analysis**

- Examined individual variable distributions using:
    - AVERAGE(range), MEDIAN(range), and STDEV.P(range) for numerical data.
    - **Pivot Tables** to summarize categorical variables.
- **Visualization**:
    - **Histograms** for income, loan amount, and credit history.
    - **Bar charts** for categorical variables like employment type and loan type.

**Step 5: Segmented Univariate Analysis**

- Compared variable distributions across different loan scenarios (approved vs. refused applicants).
- Used **Pivot Tables and Sorting** to segment data by customer profile.
- **Visualization**:
    - **Stacked bar charts** for comparing different applicant segments.
    - **Grouped bar charts** to analyze how income levels differ across loan statuses.

**Step 6: Bivariate Analysis**

- Explored relationships between two variables (e.g., income vs. default rate, loan amount vs. approval status).
- Used CORREL(range1, range2) to calculate correlation coefficients.
- **Visualization**:
    - **Scatter plots** for identifying trends in numerical data.
    - **Heatmaps (Conditional Formatting)** to highlight high correlations.

**Step 7: Identifying Top Correlations for Loan Default**

- Segmented dataset into two groups:
    - **Customers with payment difficulties**
    - **All other cases**
- Used CORREL(range1, range2) to rank top indicators of loan default.
- **Visualization**:
    - **Correlation matrix (Heatmap)** to display strongest predictors of default.
    - **Box plots** showing variations in top correlated features.

### 3. Data Visualization

To effectively communicate findings, multiple Excel visualization techniques were used:

- **Bar Charts & Pie Charts** → For categorical data distributions.
- **Box Plots** → For identifying outliers in income and loan amounts.
- **Histograms** → For numerical variable distributions.
- **Scatter Plots** → For relationships between two numerical variables.
- **Stacked Bar Charts** → For comparing different segments (approved vs. refused).

# Tech-Stack Used

**Microsoft Excel 2022**: Excel was used to perform the data analysis and calculations. The software's vast array of built-in functions allowed me to calculate descriptive statistics, correlation coefficients, and percentiles with ease. I also used its charting capabilities to visualize the relationships between different variables.

# Insights

Loan Application Analysis

1. Loan Approval & Rejection Trends

- Out of 43,344 loan applications, about 38.6% were approved, while 58.9% were rejected due to strict credit policies.
- The most common loan types were cash loans and repair loans, but large loans (like home loans) faced more rejections.
- Consumer loans had the highest approval rate, while loans for paying off other loans had the lowest.

2. Who is More Likely to Default?

The data reveals key factors that influence whether a client will default on a loan:
✅ Lower risk (less likely to default):

- Older clients with more experience → They tend to manage finances better.
- Higher education levels → College-educated clients default less than those with only secondary education.
- Female clients → Women default less than men.
- Corporate employees → Safer bet than labor workers.
- Clients from Region 1 → Lowest default rate, safest for banks.
- Older clients take larger loans but default less, making them highly profitable and low-risk.

❌ Higher risk (more likely to default):

- Younger clients with fewer years of experience.
- Lower education levels (secondary or lower).
- Male clients default more than females.

- Labor-class clients.
- Clients from Region 3 → Highest percentage of defaulters. Banks should apply stricter loan policies for this region.

3. General Loan Insights

- Loan approval rates depend on purpose – loans for repairs and urgent needs get approved more than home loans.
- Higher-income individuals apply for fewer loans.
- Married individuals take out more loans than unmarried ones.
- Most clients are aged 35-50, but younger clients (26-31) default the most.
- Clients asking for loans above ₹3,500 are more likely to get denied.

4. What Should Banks Do?

🔹 Prioritize older, educated, and corporate clients – they are safer and bring better profits.
🔹 Be stricter with younger, less educated, and labor-class applicants – they have a higher risk of default.
🔹 Region-based policies – stricter rules for Region 3 clients, while clients from Region 1 are the safest bet.
🔹 Encourage responsible borrowing – since younger clients default more, financial literacy programs could help reduce risk.

Older, educated, and corporate clients are the safest bets. Younger, less-educated, and labor-class clients are riskier. Women default less than men. Banks should be stricter with Region 3 applicants and offer more loans to experienced clients.

# CONCLUSION

## PROJECT SUMMARY

In this project, we analyzed two datasets: the **application dataset** and the **previous application dataset**. These not only helped us understand the current status of individuals applying for loans but also provided insights into their previous loan history. The dataset was very large, making it challenging to manipulate and perform queries on, but it provided valuable insights. We learned that **data imbalance** can cause serious problems while analyzing data—for example, since the **labor profession** had a higher representation, it dominated both **default and non-default categories**, affecting analysis accuracy. We also explored **correlations between different columns** and identified key factors influencing loan defaults and approvals. **Older, educated, and corporate clients are the safest bets**, while **younger, less-experienced individuals from Region 3 have a higher default risk**. Women tend to default less than men. Based on these findings, the bank should implement **stricter credit policies for high-risk clients**, **prioritize profitable, low-risk clients**, and **enhance data management** to improve decision-making and reduce default rates.

# Tasks

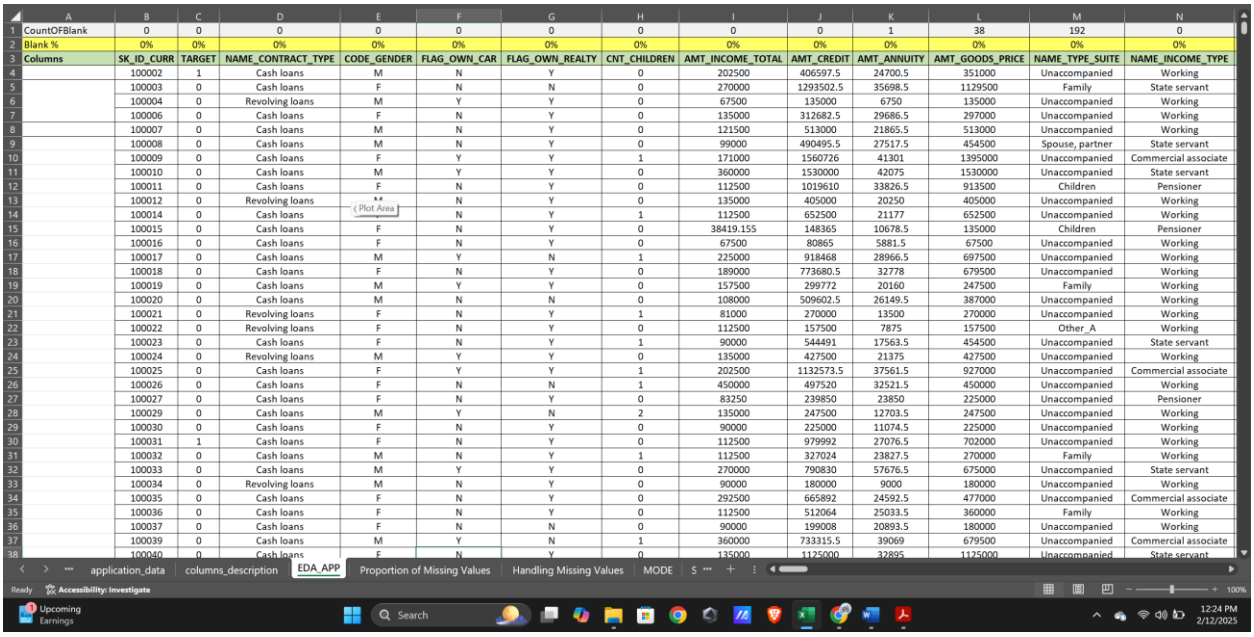
A. **Identify Missing Data and Deal with it Appropriately:** As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

- **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.
- **Hint:** Utilize Excel functions like COUNT, ISBLANK, and IF to identify missing data. Consider using functions like AVERAGE or MEDIAN for imputation or other appropriate methods available in Excel.
- **Graph suggestion:** Create a bar chart or column chart to visualize the proportion of missing values for each variable.

1)WE WIL COUNT THE NO OF BLANKS CELLS PER COLUMN USING COUNT OF BLANK AS SHOWN IN THE IMAGE

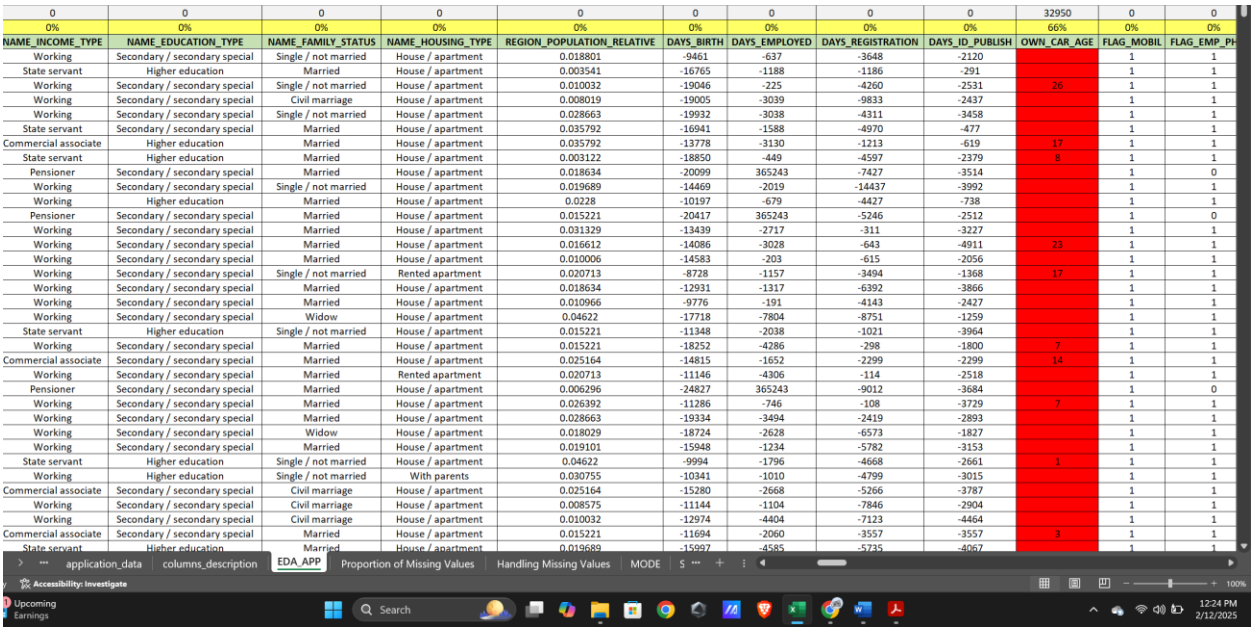
2)ALL COLUMNS HAVING BLANKS GREATER THAN 30 ARE HIGHLIGHTED IN RED

3)THIS WILL HELP US ELLIMINATE MISSING DATA,IF MORE AMOUNT OF CELLS ARE EMPTY HAN FILLED WE WILL REMOVE THAT SAID COLUMN.



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CountOFBlank | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 38 | 192 | 0 |
| 2 | Blank % | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| 3 | Columns | 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 |
| 4 | | 100002 | 1 | Cash loans | M | N | Y | 0 | 202500 | 406597.5 | 24700.5 | 351000 | Unaccompanied | Working |
| 5 | | 100003 | 0 | Cash loans | F | N | N | 0 | 270000 | 1293502.5 | 35698.5 | 1129500 | Family | State servant |
| 6 | | 100004 | 0 | Revolving loans | M | Y | Y | 0 | 67500 | 135000 | 6750 | 135000 | Unaccompanied | Working |
| 7 | | 100006 | 0 | Cash loans | F | N | Y | 0 | 135000 | 312682.5 | 29686.5 | 297000 | Unaccompanied | Working |
| 8 | | 100007 | 0 | Cash loans | M | N | Y | 0 | 121500 | 513000 | 21865.5 | 513000 | Unaccompanied | Working |
| 9 | | 100008 | 0 | Cash loans | M | N | Y | 0 | 99000 | 490495.5 | 27517.5 | 454500 | Spouse, partner | State servant |
| 10 | | 100009 | 0 | Cash loans | F | Y | Y | 1 | 171000 | 1560726 | 41301 | 1395000 | Unaccompanied | Commercial associate |
| 11 | | 100010 | 0 | Cash loans | M | Y | Y | 0 | 360000 | 1530000 | 42075 | 1530000 | Unaccompanied | State servant |
| 12 | | 100011 | 0 | Cash loans | F | N | Y | 0 | 112500 | 1019610 | 33826.5 | 913500 | Children | Pensioner |
| 13 | | 100012 | 0 | Revolving loans | M | N | Y | 0 | 135000 | 405000 | 20250 | 405000 | Unaccompanied | Working |
| 14 | | 100014 | 0 | Cash loans | F | N | Y | 1 | 112500 | 652500 | 21177 | 652500 | Unaccompanied | Working |
| 15 | | 100015 | 0 | Cash loans | F | N | Y | 0 | 38419.155 | 148365 | 10678.5 | 135000 | Children | Pensioner |
| 16 | | 100016 | 0 | Cash loans | F | N | Y | 0 | 67500 | 80865 | 5881.5 | 67500 | Unaccompanied | Working |
| 17 | | 100017 | 0 | Cash loans | M | Y | N | 1 | 225000 | 918468 | 28966.5 | 697500 | Unaccompanied | Working |
| 18 | | 100018 | 0 | Cash loans | F | N | Y | 0 | 189000 | 773680.5 | 32778 | 679500 | Unaccompanied | Working |
| 19 | | 100019 | 0 | Cash loans | M | Y | Y | 0 | 157500 | 299772 | 20160 | 247500 | Family | Working |
| 20 | | 100020 | 0 | Cash loans | M | N | N | 0 | 108000 | 509602.5 | 26149.5 | 387000 | Unaccompanied | Working |
| 21 | | 100021 | 0 | Revolving loans | F | N | Y | 1 | 81000 | 270000 | 13500 | 270000 | Unaccompanied | Working |
| 22 | | 100022 | 0 | Revolving loans | F | N | Y | 0 | 112500 | 157500 | 7875 | 157500 | Other_A | Working |
| 23 | | 100023 | 0 | Cash loans | F | N | Y | 1 | 90000 | 544491 | 17563.5 | 454500 | Unaccompanied | State servant |
| 24 | | 100024 | 0 | Revolving loans | M | Y | Y | 0 | 135000 | 427500 | 21375 | 427500 | Unaccompanied | Working |
| 25 | | 100025 | 0 | Cash loans | F | Y | Y | 1 | 202500 | 1132573.5 | 37561.5 | 927000 | Unaccompanied | Commercial associate |
| 26 | | 100026 | 0 | Cash loans | F | N | N | 1 | 450000 | 497520 | 32521.5 | 450000 | Unaccompanied | Working |
| 27 | | 100027 | 0 | Cash loans | F | N | Y | 0 | 83250 | 239850 | 23850 | 225000 | Unaccompanied | Pensioner |
| 28 | | 100029 | 0 | Cash loans | M | Y | N | 2 | 135000 | 247500 | 12703.5 | 247500 | Unaccompanied | Working |
| 29 | | 100030 | 0 | Cash loans | F | N | Y | 0 | 90000 | 225000 | 11074.5 | 225000 | Unaccompanied | Working |
| 30 | | 100031 | 1 | Cash loans | F | N | Y | 0 | 112500 | 979992 | 27076.5 | 702000 | Unaccompanied | Working |
| 31 | | 100032 | 0 | Cash loans | M | N | Y | 1 | 112500 | 327024 | 23827.5 | 270000 | Family | Working |
| 32 | | 100033 | 0 | Cash loans | M | Y | Y | 0 | 270000 | 790830 | 57676.5 | 675000 | Unaccompanied | State servant |
| 33 | | 100034 | 0 | Revolving loans | M | N | Y | 0 | 90000 | 180000 | 9000 | 180000 | Unaccompanied | Working |
| 34 | | 100035 | 0 | Cash loans | F | N | Y | 0 | 292500 | 665892 | 24592.5 | 477000 | Unaccompanied | Commercial associate |
| 35 | | 100036 | 0 | Cash loans | F | N | Y | 0 | 112500 | 512064 | 25033.5 | 360000 | Family | Working |
| 36 | | 100037 | 0 | Cash loans | F | N | N | 0 | 90000 | 199008 | 20893.5 | 180000 | Unaccompanied | Working |
| 37 | | 100039 | 0 | Cash loans | M | Y | N | 1 | 360000 | 733315.5 | 39069 | 679500 | Unaccompanied | Commercial associate |
| 38 | | 100040 | 0 | Cash loans | F | N | Y | 0 | 135000 | 1125000 | 32895 | 1125000 | Unaccompanied | State servant |

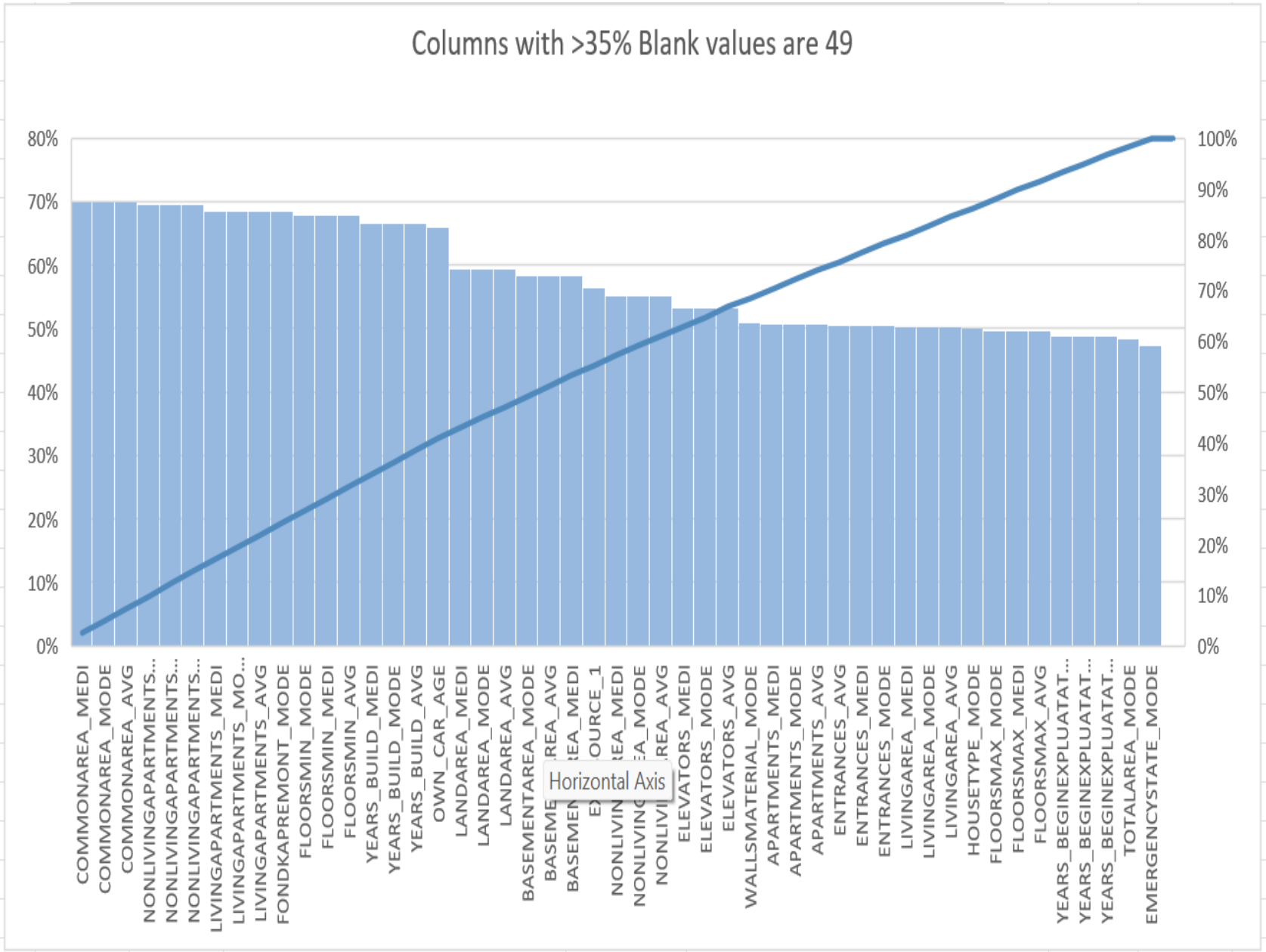




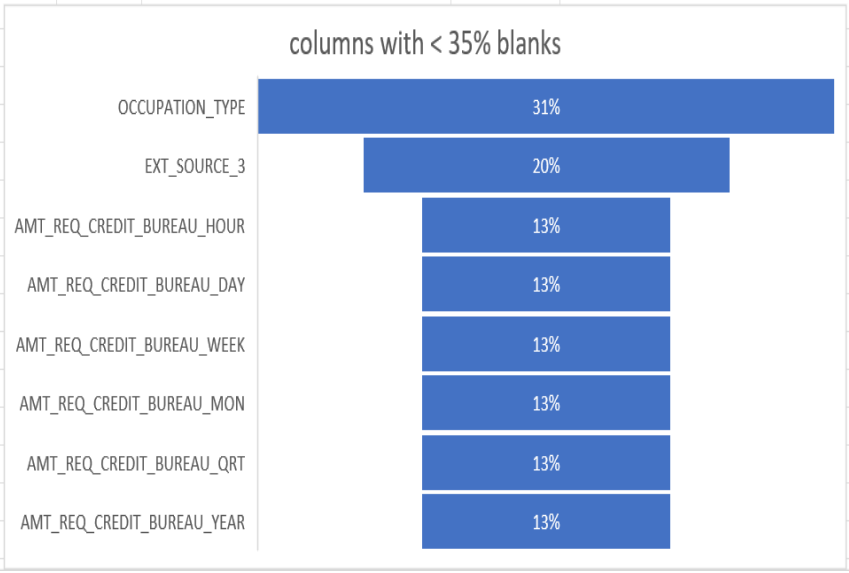
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 32950 | 0 | 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 66% | 0% | 0% |
| NAME_INCOME_TYPE | NAME_EDUCATION_TYPE | NAME_FAMILY_STATUS | NAME_HOUSING_TYPE | REGION_POPULATION_RELATIVE | DAYS_BIRTH | DAYS_EMPLOYED | DAYS_REGISTRATION | DAYS_ID_PUBLISH | OWN_CAR_AGE | FLAG_MOBIL | FLAG_EMP_PH |
| Working | Secondary / secondary special | Single / not married | House / apartment | 0.018801 | -9461 | -637 | -3648 | -2120 | | 1 | 1 |
| State servant | Higher education | Married | House / apartment | 0.003541 | -16765 | -1188 | -1186 | -291 | | 1 | 1 |
| Working | Secondary / secondary special | Single / not married | House / apartment | 0.010032 | -19046 | -225 | -4260 | -2531 | 26 | 1 | 1 |
| Working | Secondary / secondary special | Civil marriage | House / apartment | 0.008019 | -19005 | -3039 | -9833 | -2437 | | 1 | 1 |
| Working | Secondary / secondary special | Single / not married | House / apartment | 0.028663 | -19932 | -3038 | -4311 | -3458 | | 1 | 1 |
| State servant | Secondary / secondary special | Married | House / apartment | 0.035792 | -16941 | -1588 | -4970 | -477 | | 1 | 1 |
| Commercial associate | Higher education | Married | House / apartment | 0.035792 | -13778 | -3130 | -1213 | -619 | 17 | 1 | 1 |
| State servant | Higher education | Married | House / apartment | 0.003122 | -18850 | -449 | -4597 | -2379 | 8 | 1 | 1 |
| Pensioner | Secondary / secondary special | Married | House / apartment | 0.018634 | -20099 | 365243 | -7427 | -3514 | | 1 | 0 |
| Working | Secondary / secondary special | Single / not married | House / apartment | 0.019689 | -14469 | -2019 | -14437 | -3992 | | 1 | 1 |
| Working | Higher education | Married | House / apartment | 0.0228 | -10197 | -679 | -4427 | -738 | | 1 | 1 |
| Pensioner | Secondary / secondary special | Married | House / apartment | 0.015221 | -20417 | 365243 | -5246 | -2512 | | 1 | 0 |
| Working | Secondary / secondary special | Married | House / apartment | 0.031329 | -13439 | -2717 | -311 | -3227 | | 1 | 1 |
| Working | Secondary / secondary special | Married | House / apartment | 0.016612 | -14086 | -3028 | -643 | -4911 | 23 | 1 | 1 |
| Working | Secondary / secondary special | Married | House / apartment | 0.010006 | -14583 | -203 | -615 | -2056 | | 1 | 1 |
| Working | Secondary / secondary special | Single / not married | Rented apartment | 0.020713 | -8728 | -1157 | -3494 | -1368 | 17 | 1 | 1 |
| Working | Secondary / secondary special | Married | House / apartment | 0.018634 | -12931 | -1317 | -6392 | -3866 | | 1 | 1 |
| Working | Secondary / secondary special | Married | House / apartment | 0.010966 | -9776 | -191 | -4143 | -2427 | | 1 | 1 |
| Working | Secondary / secondary special | Widow | House / apartment | 0.04622 | -17718 | -7804 | -8751 | -1259 | | 1 | 1 |
| State servant | Higher education | Single / not married | House / apartment | 0.015221 | -11348 | -2038 | -1021 | -3964 | | 1 | 1 |
| Working | Secondary / secondary special | Married | House / apartment | 0.015221 | -18252 | -4286 | -298 | -1800 | 7 | 1 | 1 |
| Commercial associate | Secondary / secondary special | Married | House / apartment | 0.025164 | -14815 | -1652 | -2299 | -2299 | 14 | 1 | 1 |
| Working | Secondary / secondary special | Married | Rented apartment | 0.020713 | -11146 | -4306 | -114 | -2518 | | 1 | 1 |
| Pensioner | Secondary / secondary special | Married | House / apartment | 0.006296 | -24827 | 365243 | -9012 | -3684 | | 1 | 0 |
| Working | Secondary / secondary special | Married | House / apartment | 0.026392 | -11286 | -746 | -108 | -3729 | 7 | 1 | 1 |
| Working | Secondary / secondary special | Married | House / apartment | 0.028663 | -19334 | -3494 | -2419 | -2893 | | 1 | 1 |
| Working | Secondary / secondary special | Widow | House / apartment | 0.018029 | -18724 | -2628 | -6573 | -1827 | | 1 | 1 |
| Working | Secondary / secondary special | Married | House / apartment | 0.019101 | -15948 | -1234 | -5782 | -3153 | | 1 | 1 |
| State servant | Higher education | Single / not married | House / apartment | 0.04622 | -9994 | -1796 | -4668 | -2661 | 1 | 1 | 1 |
| Working | Higher education | Single / not married | With parents | 0.030755 | -10341 | -1010 | -4799 | -3015 | | 1 | 1 |
| Commercial associate | Secondary / secondary special | Civil marriage | House / apartment | 0.025164 | -15280 | -2668 | -5266 | -3787 | | 1 | 1 |
| Working | Secondary / secondary special | Civil marriage | House / apartment | 0.008575 | -11144 | -1104 | -7846 | -2904 | | 1 | 1 |
| Working | Secondary / secondary special | Civil marriage | House / apartment | 0.010032 | -12974 | -4404 | -7123 | -4464 | | 1 | 1 |
| Commercial associate | Secondary / secondary special | Married | House / apartment | 0.015221 | -11694 | -2060 | -3557 | -3557 | 3 | 1 | 1 |
| State servant | Higher education | Married | House / apartment | 0.019689 | -15997 | -4585 | -5735 | -4067 | | 1 | 1 |

| Columns | Blanks | Blank % |
|---|---|---|
| SK_ID_CURR | 0 | 0% |
| TARGET | 0 | 0% |
| NAME_CONTRACT_TYPE | 0 | 0% |
| CODE_GENDER | 0 | 0% |
| FLAG_OWN_CAR | 0 | 0% |
| FLAG_OWN_REALTY | 0 | 0% |
| CNT_CHILDREN | 0 | 0% |
| AMT_INCOME_TOTAL | 0 | 0% |
| AMT_CREDIT | 0 | 0% |
| AMT_ANNUITY | 1 | 0% |
| AMT_GOODS_PRICE | 38 | 0% |
| NAME_TYPE_SUITE | 192 | 0% |
| NAME_INCOME_TYPE | 0 | 0% |
| NAME_EDUCATION_TYPE | 0 | 0% |
| NAME_FAMILY_STATUS | 0 | 0% |
| NAME_HOUSING_TYPE | 0 | 0% |
| REGION_POPULATION_RELATIVE | 0 | 0% |
| DAYS_BIRTH | 0 | 0% |
| DAYS_EMPLOYED | 0 | 0% |
| DAYS_REGISTRATION | 0 | 0% |
| DAYS_ID_PUBLISH | 0 | 0% |
| EMERGENCYSTATE_MODE | 23698 | 47% |
| FLAG_MOBIL | 0 | 0% |
| FLAG_EMP_PHONE | 0 | 0% |
| FLAG_WORK_PHONE | 0 | 0% |
| FLAG_CONT_MOBILE | 0 | 0% |
| FLAG_PHONE | 0 | 0% |
| FLAG_EMAIL | 0 | 0% |
| OCCUPATION_TYPE | 15654 | 31% |
| CNT_FAM_MEMBERS | 1 | 0% |
| REGION_RATING_CLIENT | 0 | 0% |
| REGION_RATING_CLIENT_W_CITY | 0 | 0% |
| WEEKDAY_APPR_PROCESS_START | 0 | 0% |
| HOUR_APPR_PROCESS_START | 0 | 0% |
| REG_REGION_NOT_LIVE_REGION | 0 | 0% |
| REG_REGION_NOT_WORK_REGION | 0 | 0% |
| LIVE_REGION_NOT_WORK_REGION | 0 | 0% |

| Columns with >35% blanks are 49 | |
|---|---|
| Column Name | Blank % |
| EMERGENCYSTATE_MODE | 47% |
| TOTALAREA_MODE | 48% |
| YEARS_BEGINEXPLUATATION_AVG | 49% |
| YEARS_BEGINEXPLUATATION_MODE | 49% |
| YEARS_BEGINEXPLUATATION_MEDI | 49% |
| FLOORSMAX_AVG | 50% |
| FLOORSMAX_MODE | 50% |
| FLOORSMAX_MEDI | 50% |
| HOUSETYPE_MODE | 50% |
| LIVINGAREA_AVG | 50% |
| LIVINGAREA_MODE | 50% |
| LIVINGAREA_MEDI | 50% |
| ENTRANCES_AVG | 50% |
| ENTRANCES_MODE | 50% |
| ENTRANCES_MEDI | 50% |
| APARTMENTS_AVG | 51% |
| APARTMENTS_MODE | 51% |
| APARTMENTS_MEDI | 51% |
| WALLSMATERIAL_MODE | 51% |
| ELEVATORS_AVG | 53% |
| ELEVATORS_MODE | 53% |
| ELEVATORS_MEDI | 53% |
| NONLIVINGAREA_AVG | 55% |
| NONLIVINGAREA_MODE | 55% |
| NONLIVINGAREA_MEDI | 55% |
| EXT_SOURCE_1 | 56% |
| BASEMENTAREA_AVG | 58% |
| BASEMENTAREA_MODE | 58% |
| BASEMENTAREA_MEDI | 58% |
| LANDAREA_AVG | 59% |
| LANDAREA_MODE | 59% |
| LANDAREA_MEDI | 59% |
| OWN_CAR_AGE | 66% |
| YEARS_BUILD_AVG | 66% |
| YEARS_BUILD_MODE | 66% |
| YEARS_BUILD_MEDI | 66% |


Columns with >35% Blank values are 49
80%
70%
60%
50%
40%
30%
20%
10%
0%
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
COMMONAREA_MEDI
COMMONAREA_MODE
COMMONAREA_AVG
NONLIVINGAPARTMENTS...
NONLIVINGAPARTMENTS...
NONLIVINGAPARTMENTS...
LIVINGAPARTMENTS_MEDI
LIVINGAPARTMENTS_MO...
LIVINGAPARTMENTS_AVG
FONDKAPREMONT_MODE
FLOORSMIN_MODE
FLOORSMIN_MEDI
FLOORSMIN_AVG
YEARS_BUILD_MEDI
YEARS_BUILD_MODE
YEARS_BUILD_AVG
OWN_CAR_AGE
LANDAREA_MEDI
LANDAREA_MODE
LANDAREA_AVG
BASEMENTAREA_MODE
BASEME REA_AVG
BASEMEN EA_MEDI
E OURCE_1
NONLIVIN EA_MEDI
NONLIVING A_MODE
NONLIVI REA_AVG
ELEVATORS_MEDI
ELEVATORS_MODE
ELEVATORS_AVG
WALLSMATERIAL_MODE
APARTMENTS_MEDI
APARTMENTS_MODE
APARTMENTS_AVG
ENTRANCES_AVG
ENTRANCES_MEDI
ENTRANCES_MODE
LIVINGAREA_MEDI
LIVINGAREA_MODE
LIVINGAREA_AVG
HOUSETYPE_MODE
FLOORSMAX_MODE
FLOORSMAX_MEDI
FLOORSMAX_AVG
YEARS_BEGINEXPLUATAT...
YEARS_BEGINEXPLUATAT...
YEARS_BEGINEXPLUATAT...
TOTALAREA_MODE
EMERGENCYSTATE_MODE
Horizontal Axis

| Columns with < 35% blank values are 8 | |
|---|---|
| Column Name | Blank % |
| OCCUPATION_TYPE | 31% |
| EXT_SOURCE_3 | 20% |
| AMT_REQ_CREDIT_BUREAU_HOUR | 13% |
| AMT_REQ_CREDIT_BUREAU_DAY | 13% |
| AMT_REQ_CREDIT_BUREAU_WEEK | 13% |
| AMT_REQ_CREDIT_BUREAU_MON | 13% |
| AMT_REQ_CREDIT_BUREAU_QRT | 13% |
| AMT_REQ_CREDIT_BUREAU_YEAR | 13% |

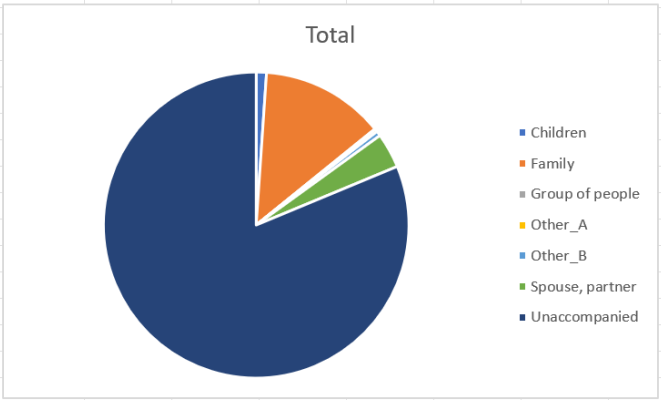


HANDLING HISSING VALUES

1) WE FIND MEAN,MEDIAN AND MODE FOR THE COLUMNS AND FILL THE BLANK CELLS WITH THE RESPECTIVE VALUES
2) IF NUMEICAL VALUE- MEDIAN
3) IF CATEGORICAL VALUE – MODE

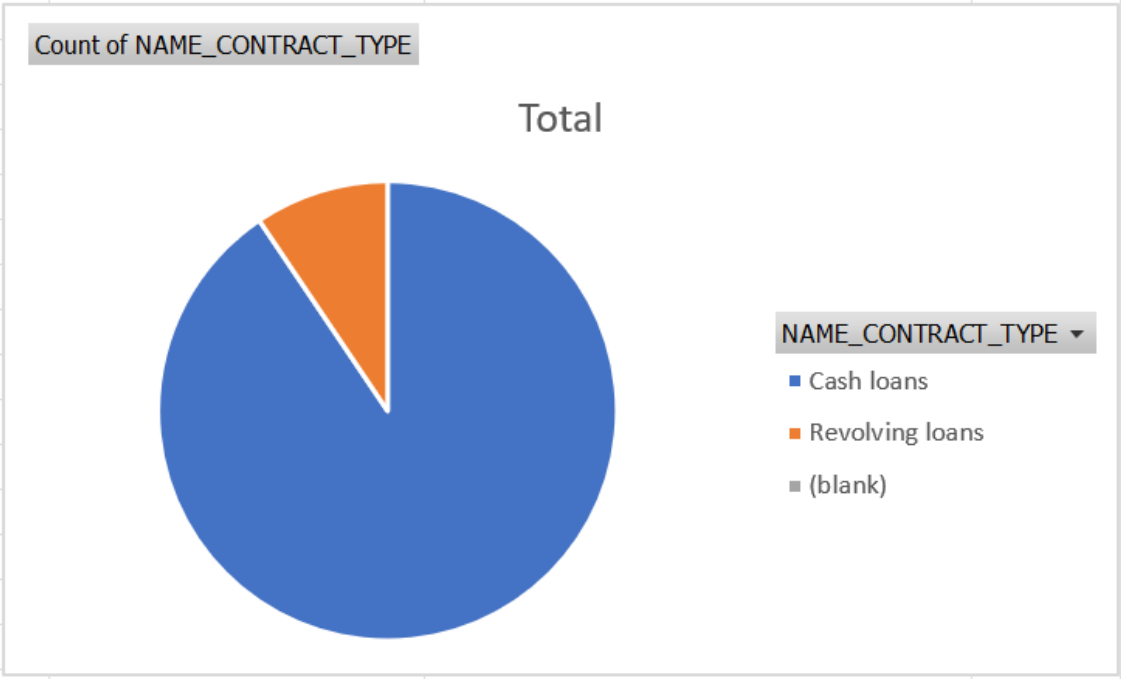
| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MEDIAN | 127652 | 0 | #NUM! | #NUM! | #NUM! | #NUM! | 0 | 144000 | 513531 | 24939 | 450000 | #NUM! | #NUM! | #NUM! |
| | SK_ID_CURR | TARGET | NAME_CONTRACT_TYPE | CODE_GENDER | FLAG_OWN_CAR | FLAG_OWN_REALTY | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | NAME_TYPE_SUITE | NAME_INCOME_TYPE | NAME_EDUCATION_TYPE |
| | 100002 | 1 | Cash loans | M | N | Y | 0 | 202500 | 406597.5 | 24700.5 | 351000 | Unaccompanied | Working | Secondary / secondary speci |
| | 100003 | 0 | Cash loans | F | N | N | 0 | 270000 | 1293502.5 | 35698.5 | 1129500 | Family | State servant | Higher education |
| | 100004 | 0 | Revolving loans | M | Y | Y | 0 | 67500 | 135000 | 6750 | 135000 | Unaccompanied | Working | Secondary / secondary speci |
| | 100006 | 0 | Cash loans | F | N | Y | 0 | 135000 | 312682.5 | 29686.5 | 297000 | Unaccompanied | Working | Secondary / secondary speci |
| | 100007 | 0 | Cash loans | M | N | Y | 0 | 121500 | 513000 | 21865.5 | 513000 | Unaccompanied | Working | Secondary / secondary speci |
| | 100008 | 0 | Cash loans | M | N | Y | 0 | 99000 | 490495.5 | 27517.5 | 454500 | Spouse, partner | State servant | Secondary / secondary speci |
| | 100009 | 0 | Cash loans | F | Y | Y | 1 | 171000 | 1560726 | 41301 | 1395000 | Unaccompanied | Commercial associate | Higher education |
| | 100010 | 0 | Cash loans | M | Y | Y | 0 | 360000 | 1530000 | 42075 | 1530000 | Unaccompanied | State servant | Higher education |
| | 100011 | 0 | Cash loans | F | N | Y | 0 | 112500 | 1019610 | 33826.5 | 913500 | Children | Pensioner | Secondary / secondary speci |
| | 100012 | 0 | Revolving loans | M | N | Y | 0 | 135000 | 405000 | 20250 | 405000 | Unaccompanied | Working | Secondary / secondary speci |
| | 100014 | 0 | Cash loans | F | N | Y | 1 | 112500 | 652500 | 21177 | 652500 | Unaccompanied | Working | Higher education |

| Row Labels | Count of NAME_TYPE_SUITE |
|---|---|
| Children | 542 |
| Family | 6549 |
| Group of people | 36 |
| Other_A | 137 |
| Other_B | 259 |
| Spouse, partner | 1849 |
| Unaccompanied | 40627 |
| Grand Total | 49999 |




**"Unaccompanied" is counted most of the time hence Mode of this column is "Unaccompanied"**

| Row Labels | Count of NAME_CONTRACT_TYPE |
|---|---|
| Cash loans | 45276 |
| Revolving loans | 4723 |
| (blank) | |
| **Grand Total** | **49999** |




**MODE FOR THIS IS CASH LOANS**

**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.
- **Hint:** Utilize Excel functions like QUARTILE, IQR, and conditional formatting to identify potential outliers. Consider applying thresholds or business rules to determine if the outliers are valid data points or require further investigation.
- **Graph suggestion:** Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers.

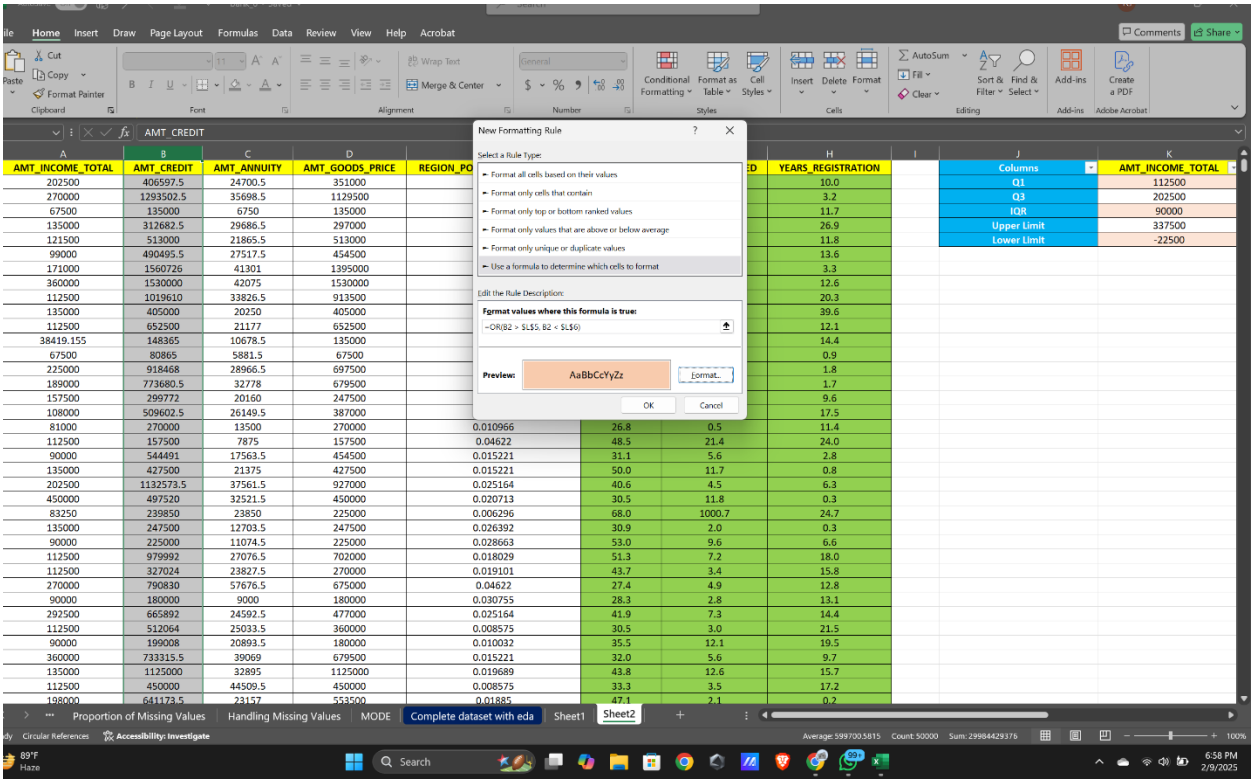
THE FOLLOWING STEPS WERE EXECUTED TO PERFORM THE GIVEN TASK-

Step 1: Calculate Quartiles and IQR

1. Calculate Q1 (25th percentile) using the formula:
   `=QUARTILE.INC(B2:B1000, 1)`
2. Calculate Q3 (75th percentile) using the formula:
   `=QUARTILE.INC(B2:B1000, 3)`
3. Calculate the Interquartile Range (IQR):
   `=Q3 - Q1`
4. Calculate the Lower Bound:
   `=Q1 - 1.5 * IQR`
5. Calculate the Upper Bound:
   `=Q3 + 1.5 * IQR`

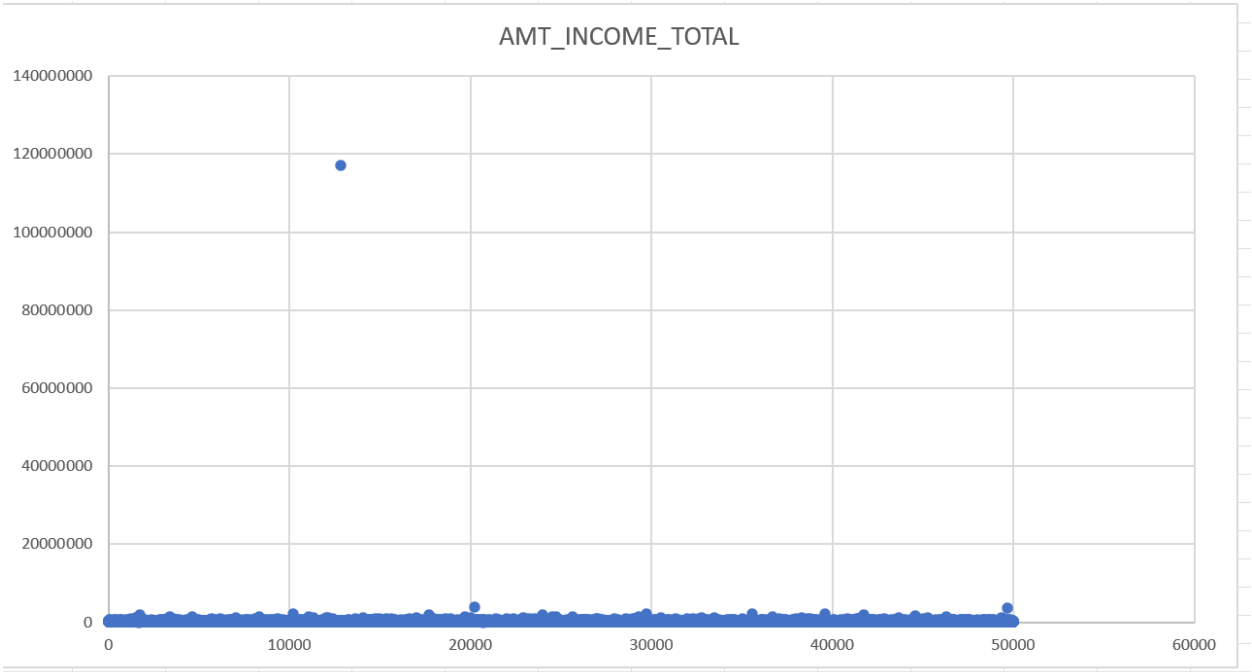
## Step 2: Flag Outliers

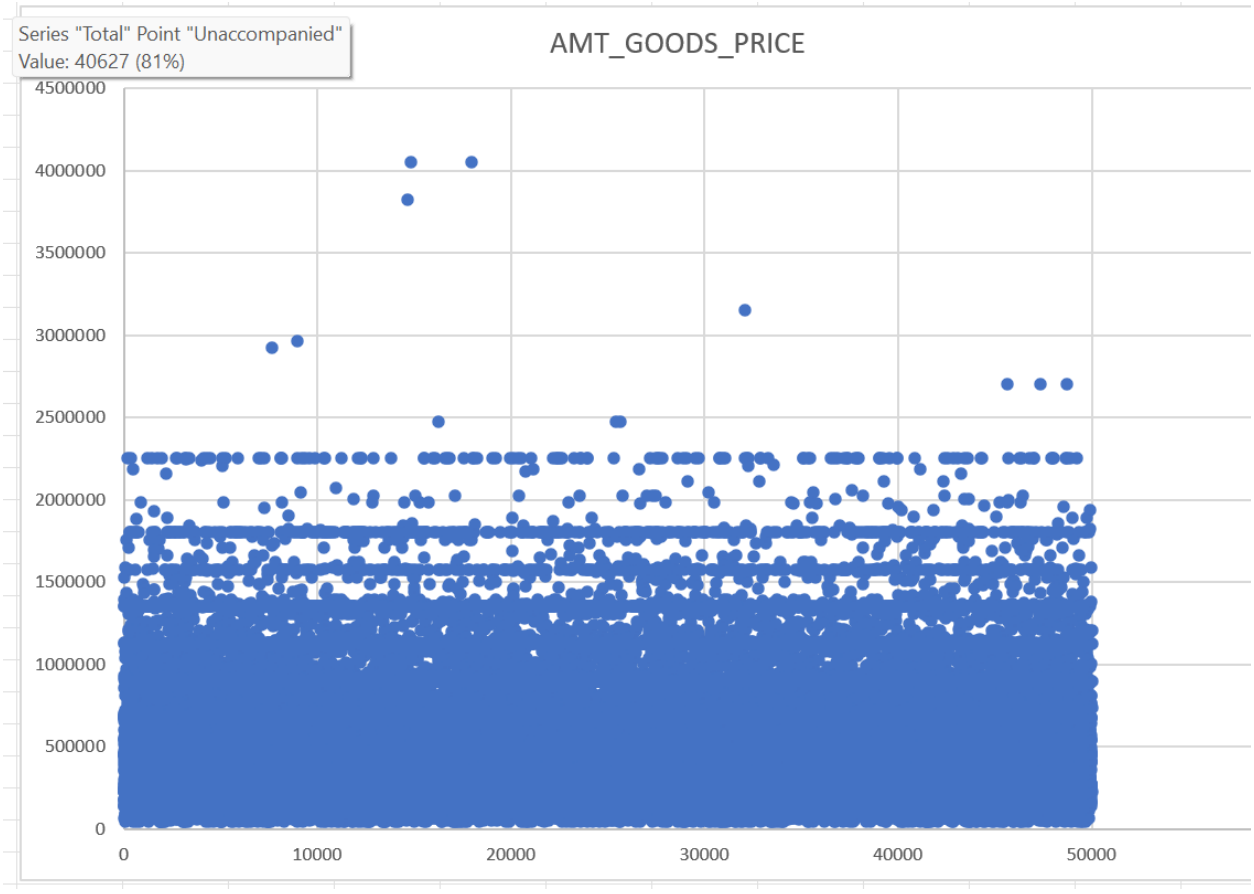
1. Apply Conditional Formatting:
    - Select the data range (e.g., B2:B1000).
    - Go to Home → Conditional Formatting → New Rule.
    - Choose “Use a formula to determine which cells to format.”
    - Enter the formula: =OR(B2 < Lower_Bound, B2 > Upper_Bound)
    - Click Format, choose a color, and press OK.

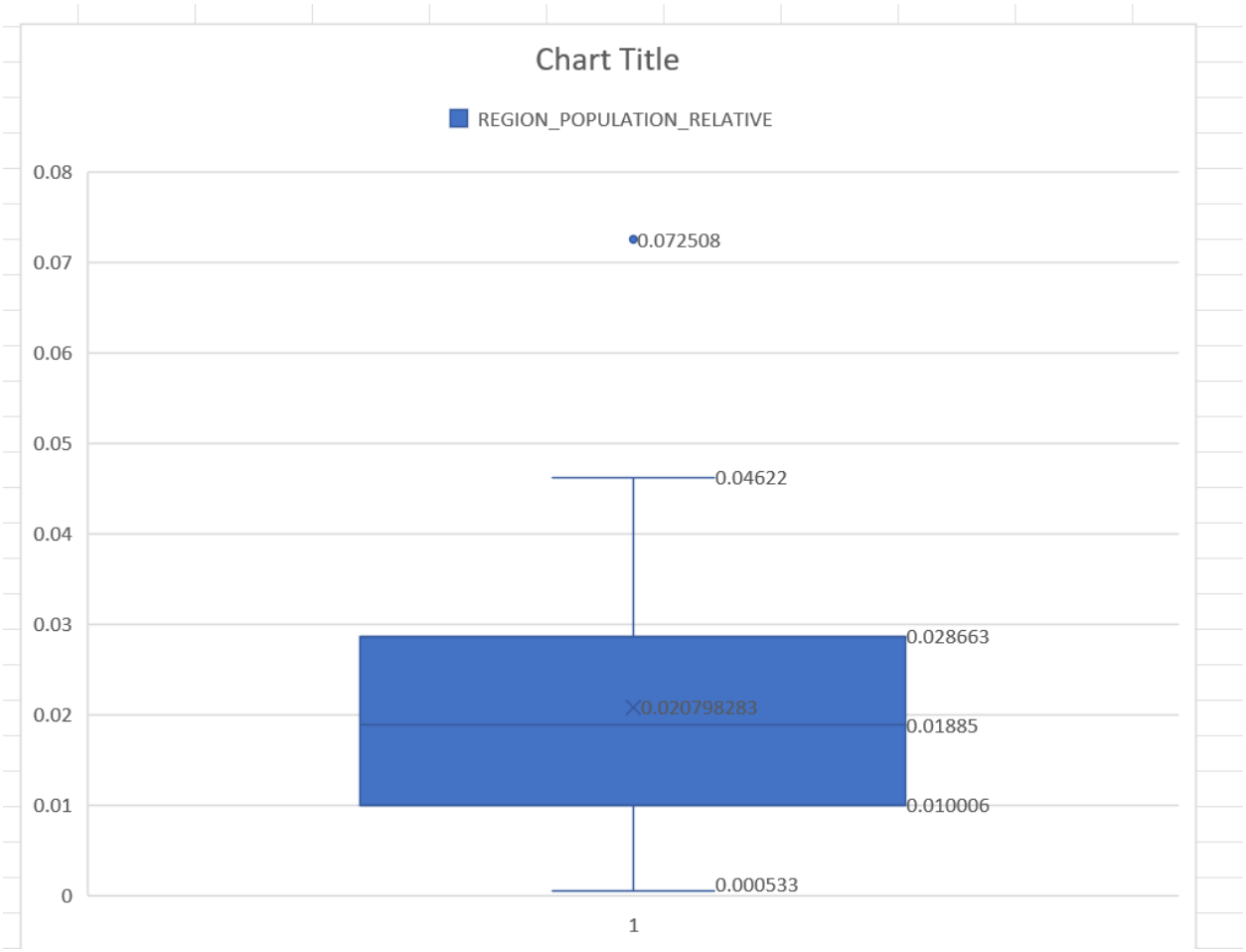


| AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | REGION_POPULATION_RELATIVE | YEARS_BIRTH | YEARS_EMPLOYED | YEARS_REGISTRATION |
|---|---|---|---|---|---|---|---|
| 202500 | 406597.5 | 24700.5 | 351000 | 0.018801 | 25.9 | 1.7 | 10.0 |
| 270000 | 1293502.5 | 35698.5 | 1129500 | 0.003541 | 45.9 | 3.3 | 3.2 |
| 67500 | 135000 | 6750 | 135000 | 0.010032 | 52.2 | 0.6 | 11.7 |
| 135000 | 312682.5 | 29686.5 | 297000 | 0.008019 | 52.1 | 8.3 | 26.9 |
| 121500 | 513000 | 21865.5 | 513000 | 0.028663 | 54.6 | 8.3 | 11.8 |
| 99000 | 490495.5 | 27517.5 | 454500 | 0.035792 | 46.4 | 4.4 | 13.6 |
| 171000 | 1560726 | 41301 | 1395000 | 0.035792 | 37.7 | 8.6 | 3.3 |
| 360000 | 1530000 | 42075 | 1530000 | 0.003122 | 51.6 | 1.2 | 12.6 |
| 112500 | 1019610 | 33826.5 | 913500 | 0.018634 | 55.1 | 1000.7 | 20.3 |
| 135000 | 405000 | 20250 | 405000 | 0.019689 | 39.6 | 5.5 | 39.6 |
| 112500 | 652500 | 21177 | 652500 | 0.0228 | 27.9 | 1.9 | 12.1 |
| 38419.155 | 148365 | 10678.5 | 135000 | 0.015221 | 55.9 | 1000.7 | 14.4 |
| 67500 | 80865 | 5881.5 | 67500 | 0.031329 | 36.8 | 7.4 | 0.9 |
| 225000 | 918468 | 28966.5 | 697500 | 0.016612 | 38.6 | 8.3 | 1.8 |
| 189000 | 773680.5 | 32778 | 679500 | 0.010006 | 40.0 | 0.6 | 1.7 |
| 157500 | 299772 | 20160 | 247500 | 0.020713 | 23.9 | 3.2 | 9.6 |
| 108000 | 509602.5 | 26149.5 | 387000 | 0.018634 | 35.4 | 3.6 | 17.5 |
| 81000 | 270000 | 13500 | 270000 | 0.010966 | 26.8 | 0.5 | 11.4 |
| 112500 | 157500 | 7875 | 157500 | 0.04622 | 48.5 | 21.4 | 24.0 |
| 90000 | 544491 | 17563.5 | 454500 | 0.015221 | 31.1 | 5.6 | 2.8 |
| 135000 | 427500 | 21375 | 427500 | 0.015221 | 50.0 | 11.7 | 0.8 |
| 202500 | 1132573.5 | 37561.5 | 927000 | 0.025164 | 40.6 | 4.5 | 6.3 |
| 450000 | 497520 | 32521.5 | 450000 | 0.020713 | 30.5 | 11.8 | 0.3 |
| 83250 | 239850 | 23850 | 225000 | 0.006296 | 68.0 | 1000.7 | 24.7 |
| 135000 | 247500 | 12703.5 | 247500 | 0.026392 | 30.9 | 2.0 | 0.3 |
| 90000 | 225000 | 11074.5 | 225000 | 0.028663 | 53.0 | 9.6 | 6.6 |
| 112500 | 979992 | 27076.5 | 702000 | 0.018029 | 51.3 | 7.2 | 18.0 |
| 112500 | 327024 | 23827.5 | 270000 | 0.019101 | 43.7 | 3.4 | 15.8 |
| 270000 | 790830 | 57676.5 | 675000 | 0.04622 | 27.4 | 4.9 | 12.8 |
| 90000 | 180000 | 9000 | 180000 | 0.030755 | 28.3 | 2.8 | 13.1 |
| 292500 | 665892 | 24592.5 | 477000 | 0.025164 | 41.9 | 7.3 | 14.4 |
| 112500 | 512064 | 25033.5 | 360000 | 0.008575 | 30.5 | 3.0 | 21.5 |
| 90000 | 199008 | 20893.5 | 180000 | 0.010032 | 35.5 | 12.1 | 19.5 |
| 360000 | 733315.5 | 39069 | 679500 | 0.015221 | 32.0 | 5.6 | 9.7 |
| 135000 | 1125000 | 32895 | 1125000 | 0.019689 | 43.8 | 12.6 | 15.7 |
| 112500 | 450000 | 44509.5 | 450000 | 0.008575 | 33.3 | 3.5 | 17.2 |
| 198000 | 641173.5 | 23157 | 553500 | 0.01885 | 47.1 | 2.1 | 0.2 |

| Columns | AMT_INCOME_TOTAL | AMT_CREDI | AMT_ANNUITY | AMT_GOODS_PRIC | REGION_POPULATION_RELATIV | YEARS_BIRTI | YEARS_EMPLOYEI | YEARS_REGISTRATIO |
|---|---|---|---|---|---|---|---|---|
| Q1 | 112500 | 270000 | 16456.5 | 238500 | 0.010006 | 33.91232877 | 2.556164384 | 5.473972603 |
| Q3 | 202500 | 808650 | 34596 | 679500 | 0.028663 | 53.81917808 | 15.66575342 | 20.44931507 |
| IQR | 90000 | 538650 | 18139.5 | 441000 | 0.018657 | 19.90684932 | 13.10958904 | 14.97534247 |
| Upper Limit | 337500 | 1616625 | 61805.25 | 1341000 | 0.0566485 | 83.67945205 | 35.33013699 | 42.91232877 |
| Lower Limit | -22500 | -537975 | -10752.75 | -423000 | -0.0179795 | 4.052054795 | -17.10821918 | -16.9890411 |


AMT_INCOME_TOTAL
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
0
10000
20000
30000
40000
50000
60000


Series "Total" Point "Unaccompanied"
Value: 40627 (81%)
AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
10000
20000
30000
40000
50000




VISUAL REPRESENTATION OF OUTLIERS

**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.

- **Task:** Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
- **Hint:** Utilize Excel functions like COUNTIF and SUM to calculate the proportions of each class. Compare the class frequencies to assess data imbalance.
- **Graph suggestion:** Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.

Class imbalance occurs when there are more observations in one class than in another. This is a common problem in machine learning.

Explanation

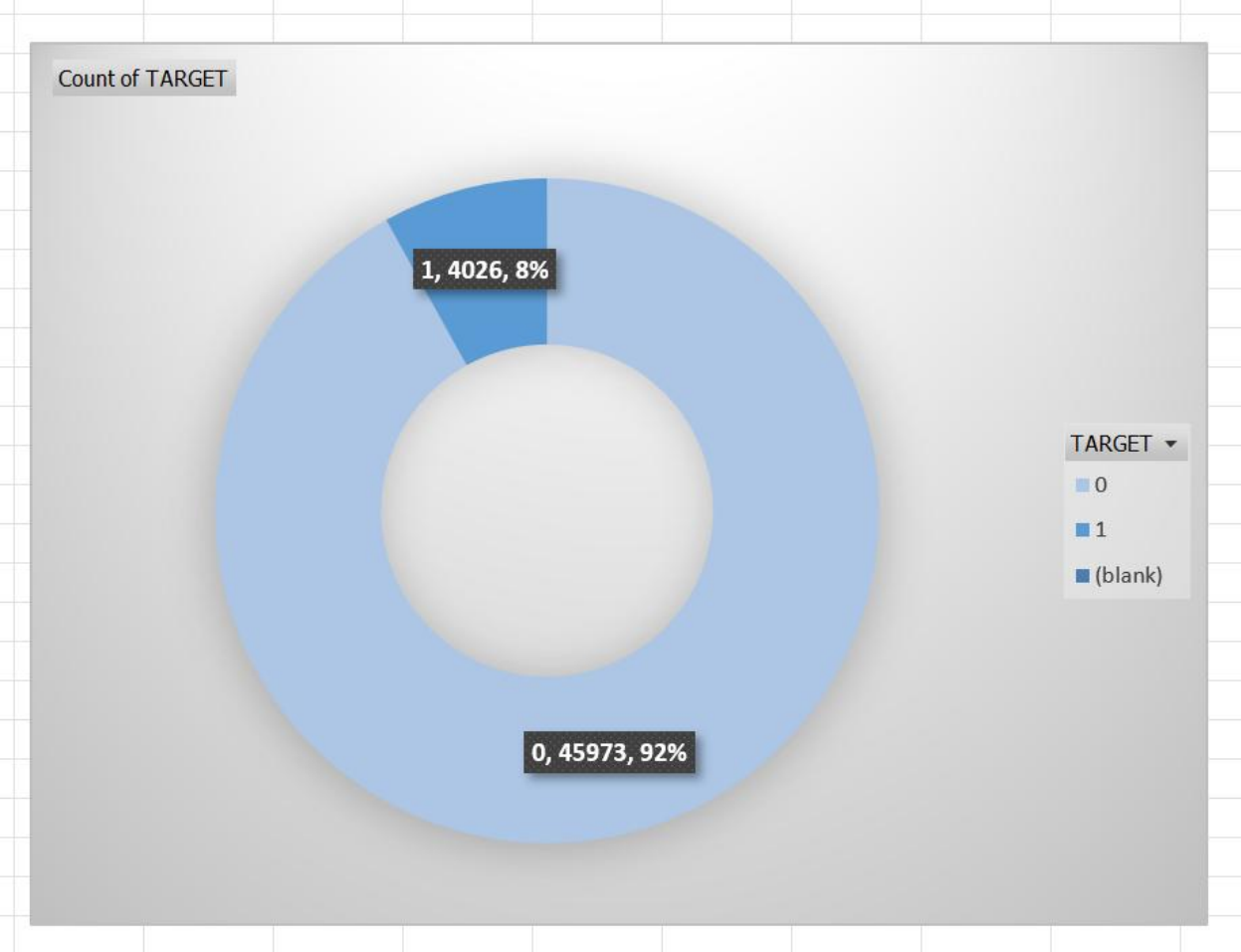
The class with more observations is called the majority class.

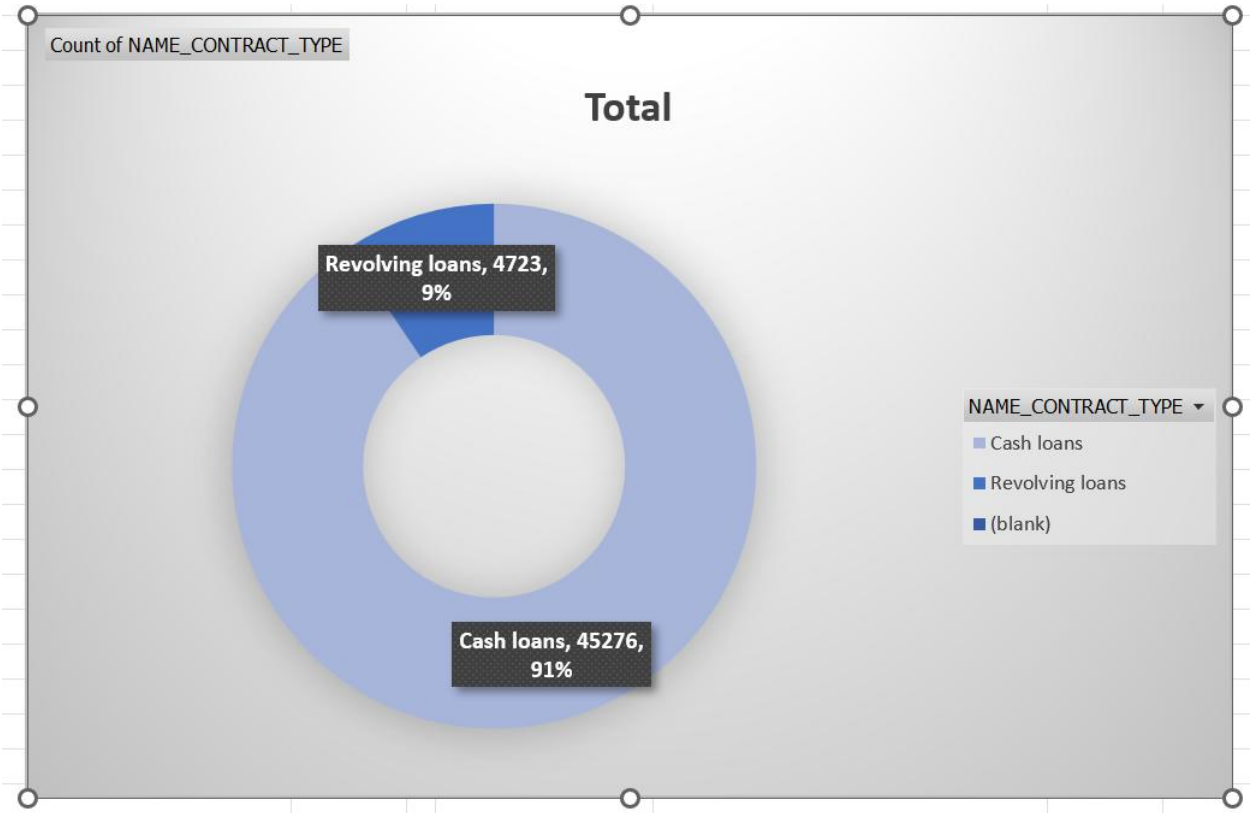
The class with fewer observations is called the minority class.

In a balanced dataset, the number of positive and negative labels is roughly equal.

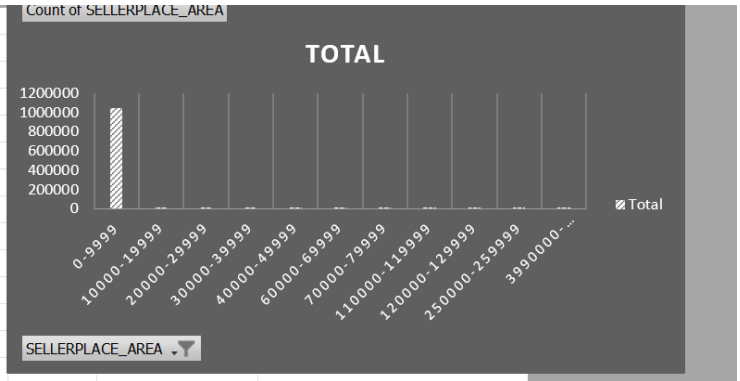
Class imbalance can affect model accuracy.

We have calculated data imbalance and its ratio using pivot tables on columns then visualizing then using pie chart and bar graphs.

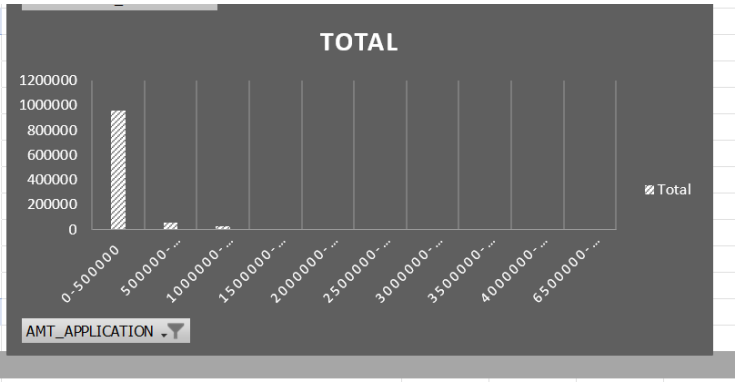




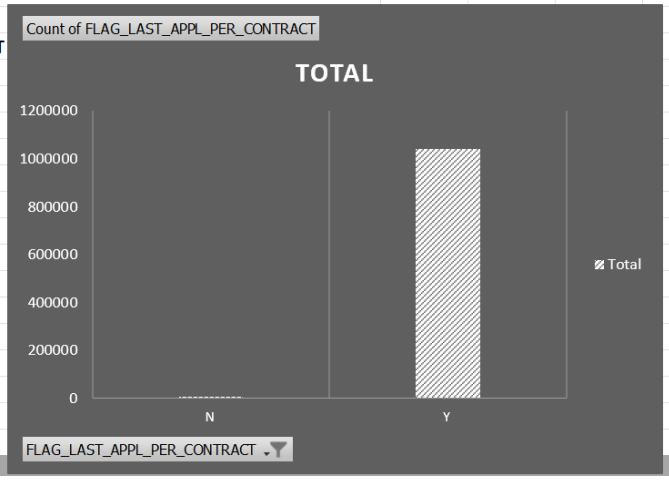

| Row Labels | Count of SELLERPLACE_AREA |
|---|---|
| 0-9999 | 1046970 |
| 10000-19999 | 698 |
| 20000-29999 | 7 |
| 30000-39999 | 627 |
| 40000-49999 | 31 |
| 60000-69999 | 6 |
| 70000-79999 | 219 |
| 110000-119999 | 1 |
| 120000-129999 | 3 |
| 250000-259999 | 8 |
| 3990000-4000000 | 4 |
| **Grand Total** | **1048574** |



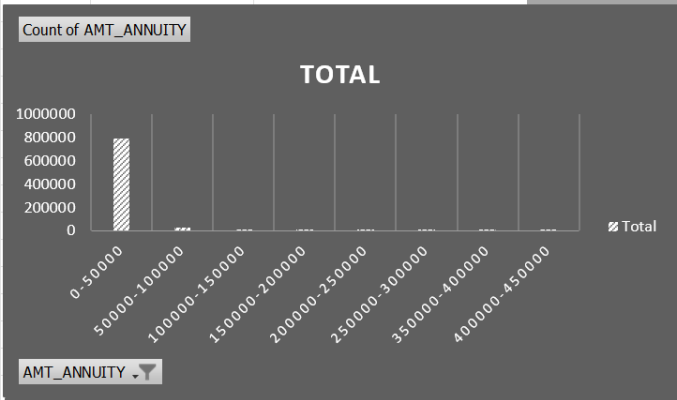
| Row Labels | Count of AMT_APPLICATION |
|---|---|
| 0-500000 | 957615 |
| 500000-1000000 | 58212 |
| 1000000-1500000 | 25535 |
| 1500000-2000000 | 4314 |
| 2000000-2500000 | 2395 |
| 2500000-3000000 | 211 |
| 3000000-3500000 | 270 |
| 3500000-4000000 | 12 |
| 4000000-4500000 | 9 |
| 6500000-7000000 | 1 |
| **Grand Total** | **1048574** |



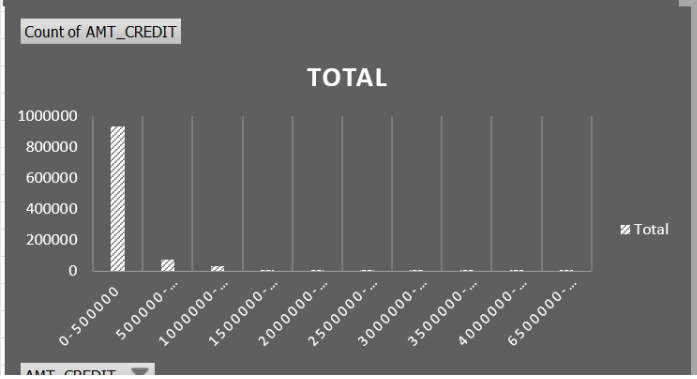

| Row Labels | Count of FLAG_LAST_APPL_PER_CONTRACT |
|---|---|
| N | 5372 |
| Y | 1043202 |
| **Grand Total** | **1048574** |




| Row Labels | Count of AMT_ANNUITY |
|---|---|
| 0-50000 | 786610 |
| 50000-100000 | 27254 |
| 100000-150000 | 1493 |
| 150000-200000 | 141 |
| 200000-250000 | 54 |
| 250000-300000 | 7 |
| 350000-400000 | 2 |
| 400000-450000 | 4 |
| **Grand Total** | **815565** |




| Row Labels | Count of AMT_CREDIT |
|---|---|
| 0-500000 | 935511 |
| 500000-1000000 | 73644 |
| 1000000-1500000 | 28609 |
| 1500000-2000000 | 6851 |
| 2000000-2500000 | 2909 |
| 2500000-3000000 | 677 |
| 3000000-3500000 | 235 |
| 3500000-4000000 | 126 |
| 4000000-4500000 | 11 |
| 6500000-7000000 | 1 |
| **Grand Total** | **1048574** |



**D. 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.

- **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.
- **Hint:** Utilize Excel functions like COUNT, AVERAGE, MEDIAN, and statistical functions for descriptive analysis. Utilize Excel features like filters, sorting, and pivot tables for segmented and bivariate analysis.
- **Graph suggestion:** Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

In data analysis, "univariate" means analyzing a single variable at a time, "segmented univariate" refers to analyzing a single variable but broken down into different groups or segments within the data, and "bivariate" means examining the relationship between two variables simultaneously.

UNIVARIATE ANALYSIS

WE ARE SUPPOSE TO CALCULATE MEAN MEDIAN STD DEVIATION MIN AND MAX VALUES FOR THE COLUMNS

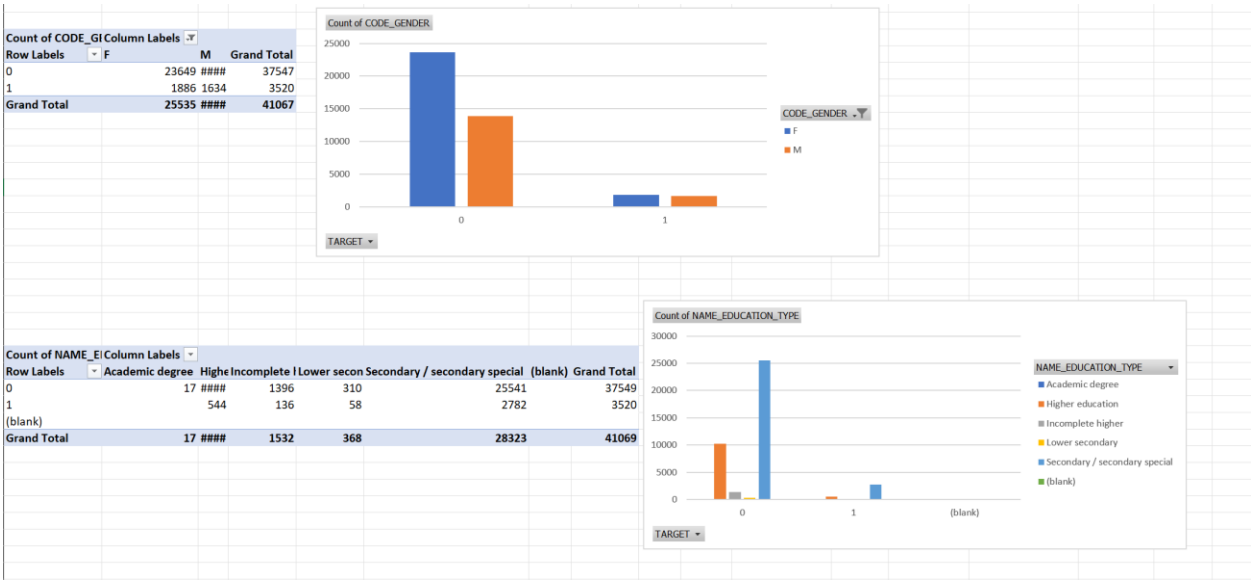
1. **Objective**: Summarize each variable's statistics.
2. **Formula**: Use Excel functions like *AVERAGE*, *STDEV*, *MIN*, *MAX*, and *COUNT*.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| | | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE |
| 2 | MEAN | 170767.5905 | 599700.5815 | 27107.33399 | 538992.3491 |
| 3 | MEDIAN | 145800 | 514777.5 | 24939 | 450000 |
| 4 | STD DEV | 531813.7768 | 402411.4096 | 14562.6564 | 369717.1252 |
| 5 | MIN | 25650 | 45000 | 2052 | 45000 |
| 6 | MAX | 117000000 | 4050000 | 258025.5 | 4050000 |
| 7 | COUNT | 49999 | 49999 | 49999 | 49999 |
| 8 | | | | | |
| 9 | | 427500 | 814041 | 23800.5 | 679500 |
| 10 | | 99000 | 808650 | 26217 | 675000 |
| 11 | | 90000 | 545040 | 26640 | 450000 |
| 12 | | 225000 | 720000 | 21051 | 720000 |
| 13 | | 67500 | 254700 | 14220 | 225000 |
| 14 | | 90000 | 592560 | 25105.5 | 450000 |
| 15 | | 54000 | 71955 | 7245 | 67500 |
| 16 | | 157500 | 225000 | 26703 | 225000 |
| 17 | | 81000 | 539100 | 27652.5 | 450000 |
| 18 | | 58500 | 675000 | 21775.5 | 675000 |

Segmented Univariate Analysis

**What We Do in Segmented Univariate Analysis**

- **Group by a Category:**
  We divide the data into segments based on a categorical variable. In your case, the TARGET variable (with values 0 for non-defaults and 1 for defaults) is used to segment the data.
- **Analyze One Variable at a Time:**
  For each segment, we examine individual variables—such as **AMT_INCOME_TOTAL**, **AMT_CREDIT**, **AMT_ANNUITY**, and **YEARS_BIRTH**—to understand their distribution, central tendency (mean, median), variability (standard deviation), and range (min, max).
- **Compare Statistics Across Groups:**
  By looking at the descriptive statistics for each segment separately, you can compare how a variable behaves across different groups. For example, you might compare the average income between those who defaulted and those who did not.

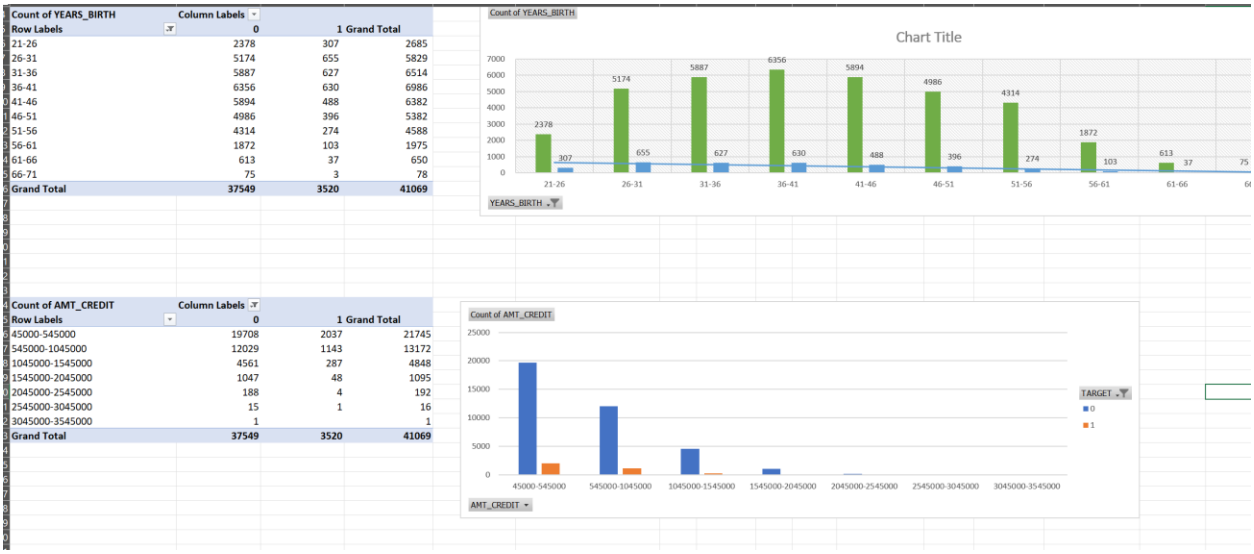


| Count of CODE_GI | Column Labels | | |
|---|---|---|---|
| Row Labels | F | M | Grand Total |
| 0 | 23649 | #### | 37547 |
| 1 | 1886 | 1634 | 3520 |
| Grand Total | 25535 | #### | 41067 |



| Count of NAME_EI | Column Labels | | | | | |
|---|---|---|---|---|---|---|
| Row Labels | Academic degree | Highe | Incomplete l | Lower secon | Secondary / secondary special | (blank) | Grand Total |
| 0 | 17 | #### | 1396 | 310 | 25541 | | 37549 |
| 1 | | 544 | 136 | 58 | 2782 | | 3520 |
| (blank) | | | | | | | |
| Grand Total | 17 | #### | 1532 | 368 | 28323 | | 41069 |





| Count of YEARS_BIRTH | Column Labels | | |
|---|---|---|---|
| Row Labels | 0 | 1 | Grand Total |
| 21-26 | 2378 | 307 | 2685 |
| 26-31 | 5174 | 655 | 5829 |
| 31-36 | 5887 | 627 | 6514 |
| 36-41 | 6356 | 630 | 6986 |
| 41-46 | 5894 | 488 | 6382 |
| 46-51 | 4986 | 396 | 5382 |
| 51-56 | 4314 | 274 | 4588 |
| 56-61 | 1872 | 103 | 1975 |
| 61-66 | 613 | 37 | 650 |
| 66-71 | 75 | 3 | 78 |
| Grand Total | 37549 | 3520 | 41069 |



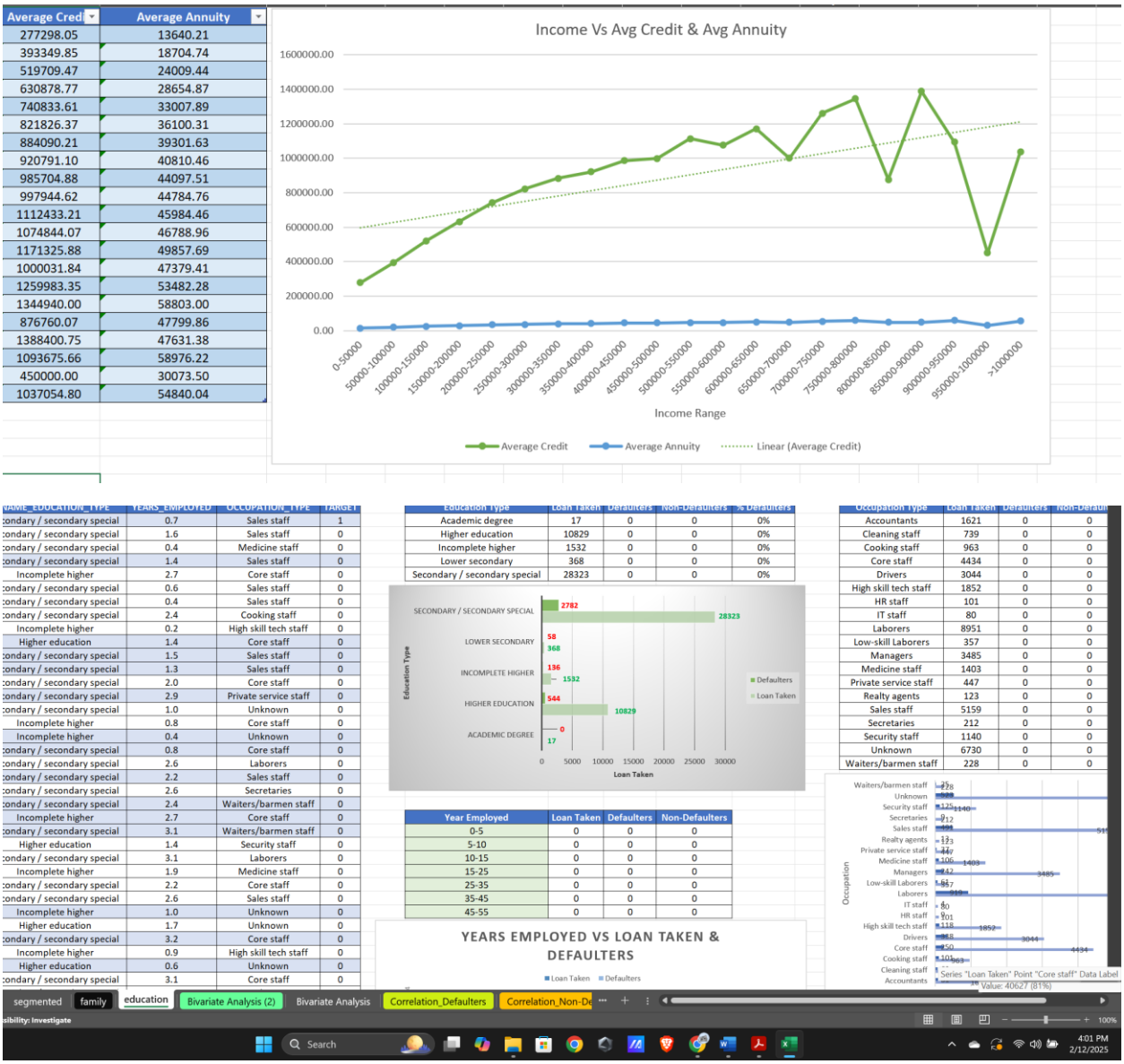
| Count of AMT_CREDIT | Column Labels | | |
|---|---|---|---|
| Row Labels | 0 | 1 | Grand Total |
| 45000-545000 | 19708 | 2037 | 21745 |
| 545000-1045000 | 12029 | 1143 | 13172 |
| 1045000-1545000 | 4561 | 287 | 4848 |
| 1545000-2045000 | 1047 | 48 | 1095 |
| 2045000-2545000 | 188 | 4 | 192 |
| 2545000-3045000 | 15 | 1 | 16 |
| 3045000-3545000 | 1 | | 1 |
| Grand Total | 37549 | 3520 | 41069 |



We have compare multiple columns with target column to check and compare payment difficulties.

Bivariate analysis

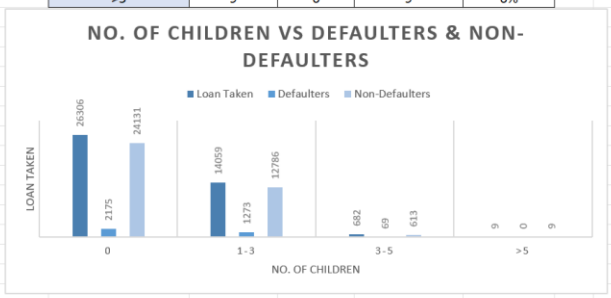
Bivariate analysis helps us understand the relationship between two variables. Since our dataset involves loan application data, this analysis will help us explore relationships between numeric and categorical variables to identify trends or risk factors.

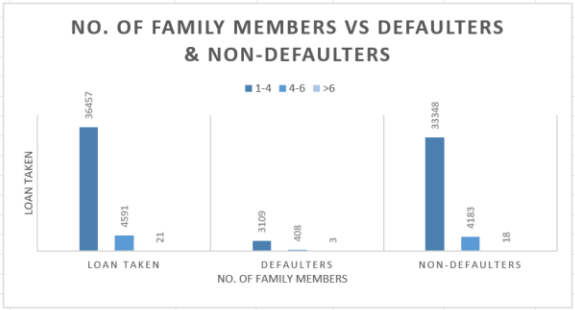


## Analysis Based On Family Status

| No. of Children | Loan Taken | Defaulters | Non-Defaulters | % Defaulters |
|---|---|---|---|---|
| 0 | 26306 | 2175 | 24131 | 8% |
| 1-3 | 14059 | 1273 | 12786 | 9% |
| 3-5 | 682 | 69 | 613 | 10% |
| >5 | 9 | 0 | 9 | 0% |

| No. of Family Members | Loan Taken | Defaulters | Non-Defaulters | % Defaulters |
|---|---|---|---|---|
| 1-4 | 36457 | 3109 | 33348 | 9% |
| 4-6 | 4591 | 408 | 4183 | 9% |
| >6 | 21 | 3 | 18 | 14% |

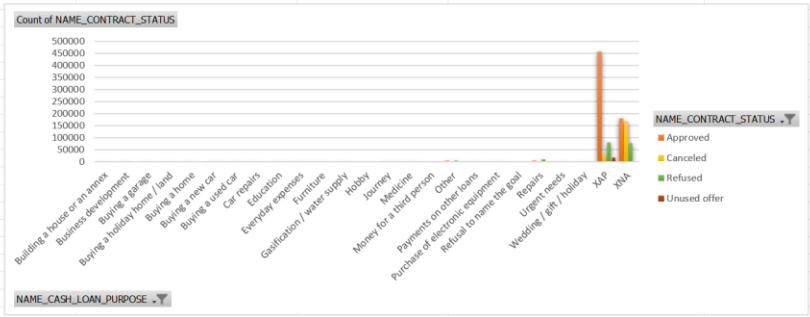



| Family Status | Loan Taken | Defaulters | Non-Defaulters | %Defaulters |
|---|---|---|---|---|
| Married | 26758 | 2103 | 24655 | 8% |
| Single / not married | 6352 | 672 | 5680 | 11% |
| Civil marriage | 4277 | 453 | 3824 | 11% |
| Separated | 2578 | 225 | 2353 | 9% |
| Widow | 1103 | 67 | 1036 | 6% |
| Unknown | 1 | 0 | 1 | 0% |


Series "Tot Horizontal (Category) Axis
Value: 40627 (81%)

| Refused | Unused offer | Grand Total |
|---|---|---|
| 1188 | | 1682 |
| 164 | | 254 |
| 51 | | 84 |
| 230 | | 334 |
| 393 | | 546 |
| 465 | 4 | 637 |
| 1166 | 9 | 1784 |
| 256 | | 493 |
| 476 | 4 | 975 |
| 740 | 7 | 1487 |
| 250 | | 475 |
| 125 | | 203 |
| 20 | | 31 |
| 404 | 2 | 745 |
| 696 | 5 | 1402 |
| 6 | | 16 |
| 5310 | 62 | 9664 |
| 973 | 3 | 1210 |
| 280 | 3 | 644 |
| 7 | | 8 |
| 8973 | 28 | 14767 |
| 2998 | | 5309 |
| 336 | | 594 |
| 78886 | 16465 | 582713 |
| 77690 | 183 | 422517 |
| **182083** | **16775** | **1048574** |




| Refused | Unused offer | Grand Total |
|---|---|---|
| 147329 | 583 | 786610 |
| 8632 | | 27254 |
| 384 | | 1493 |

E. **Identify Top Correlations for Different Scenarios:** Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

- **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.
- **Hint:** Utilize Excel functions like CORREL to calculate correlation coefficients between variables and the target variable within each segment. Rank the correlations to identify the top indicators of loan default for each scenario.
- **Graph suggestion:** Create correlation matrices or heatmaps to visualize the correlations between variables within each segment. Highlight the top correlated variables for each scenario using different colors or shading.

The CORREL function in Excel is used to calculate the correlation coefficient between two data sets. The correlation coefficient measures the strength and direction of the linear relationship between two variables.

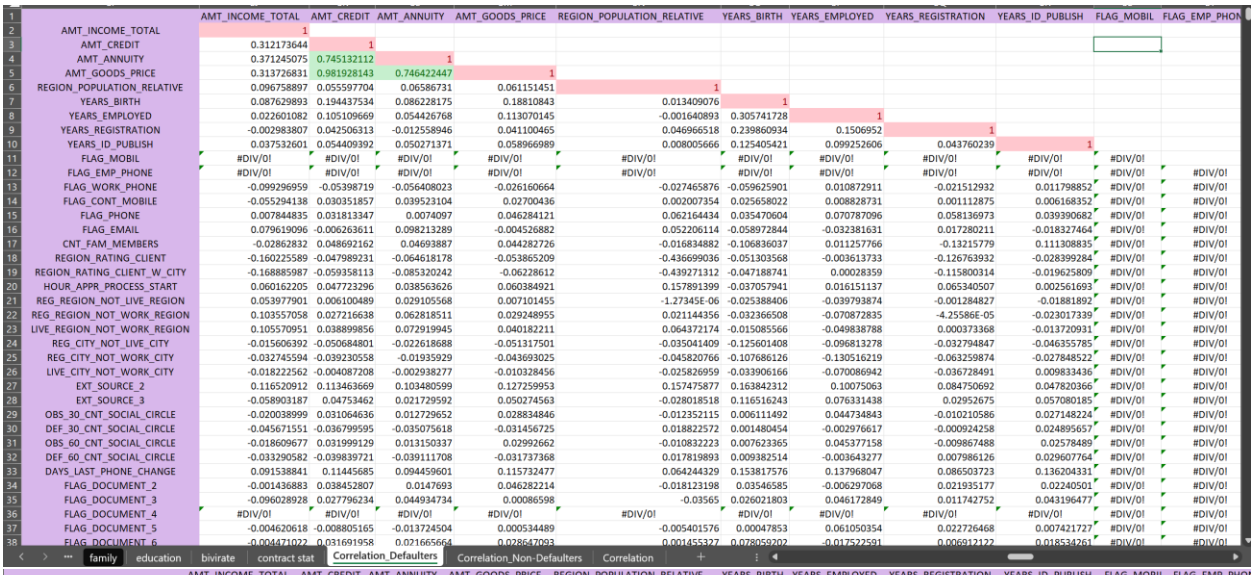
Formula-

=CORREL(array1, array2)

Steps-

1. Separate the data based on defaulters and no defaulters that is 1 and 0 respectively.

2. Then use =CORREL(array1, array2)  function to get the correlation between different columns in the data set for both defaulters and non defaulters.



| | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | REGION_POPULATION_RELATIVE | YEARS_BIRTH | YEARS_EMPLOYED | YEARS_REGISTRATION | YEARS_ID_PUBLISH | FLAG_MOBIL | FLAG_EMP_PHON |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AMT_INCOME_TOTAL | 1 | | | | | | | | | | |
| AMT_CREDIT | 0.312173644 | 1 | | | | | | | | | |
| AMT_ANNUITY | 0.371245075 | 0.745132112 | 1 | | | | | | | | |
| AMT_GOODS_PRICE | 0.313726831 | 0.981928143 | 0.746422447 | 1 | | | | | | | |
| REGION_POPULATION_RELATIVE | 0.096758897 | 0.055597704 | 0.06586731 | 0.061151451 | 1 | | | | | | |
| YEARS_BIRTH | 0.087629893 | 0.194437534 | 0.086228175 | 0.18810843 | 0.013409076 | 1 | | | | | |
| YEARS_EMPLOYED | 0.022601082 | 0.105109669 | 0.054426768 | 0.113070145 | -0.001640893 | 0.305741728 | 1 | | | | |
| YEARS_REGISTRATION | -0.002983807 | 0.042506313 | -0.012558946 | 0.041100465 | 0.046966518 | 0.239860934 | 0.1506952 | 1 | | | |
| YEARS_ID_PUBLISH | 0.037532601 | 0.054409392 | 0.050271371 | 0.058966989 | 0.008005666 | 0.125405421 | 0.099252606 | 0.043760239 | 1 | | |
| FLAG_MOBIL | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | |
| FLAG_EMP_PHONE | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! |
| FLAG_WORK_PHONE | -0.099296959 | -0.05398719 | -0.056408023 | -0.026160664 | -0.027465876 | -0.059625901 | 0.010872911 | -0.021512932 | 0.011798852 | #DIV/0! | #DIV/0! |
| FLAG_CONT_MOBILE | -0.055294138 | 0.030351857 | 0.039523104 | 0.02700436 | 0.002007354 | 0.025658022 | 0.008828731 | 0.001112875 | 0.006168352 | #DIV/0! | #DIV/0! |
| FLAG_PHONE | 0.007844835 | 0.031813347 | 0.0074097 | 0.046284121 | 0.062164434 | 0.035470604 | 0.070787096 | 0.058136973 | 0.039390682 | #DIV/0! | #DIV/0! |
| FLAG_EMAIL | 0.079619096 | -0.006263611 | 0.098213289 | -0.004526882 | 0.052206114 | -0.058972844 | -0.032381631 | 0.017280211 | -0.018327464 | #DIV/0! | #DIV/0! |
| CNT_FAM_MEMBERS | -0.02862832 | 0.048692162 | 0.04693887 | 0.044282726 | -0.016834882 | -0.106836037 | 0.011257766 | -0.13215779 | 0.111308835 | #DIV/0! | #DIV/0! |
| REGION_RATING_CLIENT | -0.160225589 | -0.047989231 | -0.064618178 | -0.053865209 | -0.436699036 | -0.051303568 | -0.003613733 | -0.126763932 | -0.028399284 | #DIV/0! | #DIV/0! |
| REGION_RATING_CLIENT_W_CITY | -0.168885987 | -0.059358113 | -0.085320242 | -0.06228612 | -0.439271312 | -0.047188741 | 0.00028359 | -0.115800314 | -0.019625809 | #DIV/0! | #DIV/0! |
| HOUR_APPR_PROCESS_START | 0.060162205 | 0.047723296 | 0.038563626 | 0.060384921 | 0.157891399 | -0.037057941 | 0.016151137 | 0.065340507 | 0.002561693 | #DIV/0! | #DIV/0! |
| REG_REGION_NOT_LIVE_REGION | 0.053977901 | 0.006100489 | 0.029105568 | 0.007101455 | -1.27345E-06 | -0.025388406 | -0.039793874 | -0.001284827 | -0.01881892 | #DIV/0! | #DIV/0! |
| REG_REGION_NOT_WORK_REGION | 0.103557058 | 0.027216638 | 0.062818511 | 0.029248955 | 0.021144356 | -0.032366508 | -0.070872835 | -4.25586E-05 | -0.023017339 | #DIV/0! | #DIV/0! |
| LIVE_REGION_NOT_WORK_REGION | 0.105570951 | 0.038899856 | 0.072919945 | 0.040182211 | 0.064372174 | -0.015085566 | -0.049838788 | 0.000373368 | -0.013720931 | #DIV/0! | #DIV/0! |
| REG_CITY_NOT_LIVE_CITY | -0.015606392 | -0.050684801 | -0.022618688 | -0.051317501 | -0.035041409 | -0.125601408 | -0.096813278 | -0.032794847 | -0.046355785 | #DIV/0! | #DIV/0! |
| REG_CITY_NOT_WORK_CITY | -0.032745594 | -0.039230558 | -0.01935929 | -0.043693025 | -0.045820766 | -0.107686126 | -0.130516219 | -0.063259874 | -0.027848522 | #DIV/0! | #DIV/0! |
| LIVE_CITY_NOT_WORK_CITY | -0.018222562 | -0.004087208 | -0.002938277 | -0.010328456 | -0.025826959 | -0.033906166 | -0.070086942 | -0.036728491 | 0.009833436 | #DIV/0! | #DIV/0! |
| EXT_SOURCE_2 | 0.116520912 | 0.113463669 | 0.103480599 | 0.127259953 | 0.157475877 | 0.163842312 | 0.10075063 | 0.084750692 | 0.047820366 | #DIV/0! | #DIV/0! |
| EXT_SOURCE_3 | -0.058903187 | 0.04753462 | 0.021729592 | 0.050274563 | -0.028018518 | 0.116516243 | 0.076331438 | 0.02952675 | 0.057080185 | #DIV/0! | #DIV/0! |
| OBS_30_CNT_SOCIAL_CIRCLE | -0.020038999 | 0.031064636 | 0.012729652 | 0.028834846 | -0.012352115 | 0.006111492 | 0.044734843 | -0.010210586 | 0.027148224 | #DIV/0! | #DIV/0! |
| DEF_30_CNT_SOCIAL_CIRCLE | -0.045671551 | -0.036799595 | -0.035075618 | -0.031456725 | 0.018822572 | 0.001480454 | -0.002976617 | -0.000924258 | 0.024895657 | #DIV/0! | #DIV/0! |
| OBS_60_CNT_SOCIAL_CIRCLE | -0.018609677 | 0.031999129 | 0.013150337 | 0.02992662 | -0.010832223 | 0.007623365 | 0.045377158 | -0.009867488 | 0.02578489 | #DIV/0! | #DIV/0! |
| DEF_60_CNT_SOCIAL_CIRCLE | -0.033290582 | -0.039839721 | -0.039111708 | -0.031737368 | 0.017819893 | 0.009382514 | -0.003643277 | 0.007986126 | 0.029607764 | #DIV/0! | #DIV/0! |
| DAYS_LAST_PHONE_CHANGE | 0.091538841 | 0.11445685 | 0.094459601 | 0.115732477 | 0.064244329 | 0.153817576 | 0.137968047 | 0.086503723 | 0.136204331 | #DIV/0! | #DIV/0! |
| FLAG_DOCUMENT_2 | -0.001436883 | 0.038452807 | 0.0147693 | 0.046282214 | -0.018123198 | 0.03546585 | -0.006297068 | 0.021935177 | 0.02240501 | #DIV/0! | #DIV/0! |
| FLAG_DOCUMENT_3 | -0.096028928 | 0.027796234 | 0.044934734 | 0.00086598 | -0.03565 | 0.026021803 | 0.046172849 | 0.011742752 | 0.043196477 | #DIV/0! | #DIV/0! |
| FLAG_DOCUMENT_4 | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! |
| FLAG_DOCUMENT_5 | -0.004620618 | -0.008805165 | -0.013724504 | 0.000534489 | -0.005401576 | 0.00047853 | 0.061050354 | 0.022726468 | 0.007421727 | #DIV/0! | #DIV/0! |
| FLAG_DOCUMENT_6 | -0.004471022 | 0.031691958 | 0.021665664 | 0.028647093 | 0.001455327 | 0.078059202 | -0.017522591 | 0.006912122 | 0.018534261 | #DIV/0! | #DIV/0! |

family | education | bivirate | contract stat | Correlation_Defaulters | Correlation_Non-Defaulters | Correlation

| | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | REGION_POPULATION_RELATIVE | YEARS_BIRTH | YEARS_EMPLOYED | YEARS_REGISTRATION | YEARS_ID_PUBLISH | FLAG_MOBIL | FLAG_EMP_PHON |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AMT_INCOME_TOTAL | 1 | | | | | | | | | | |
| AMT_CREDIT | 0.360011781 | 1 | | | | | | | | | |
| AMT_ANNUITY | 0.431026919 | 0.760827873 | 1 | | | | | | | | |
| AMT_GOODS_PRICE | 0.367729007 | 0.98635817 | 0.765201743 | 1 | | | | | | | |
| REGION_POPULATION_RELATIVE | 0.188785867 | 0.09654213 | 0.116108426 | 0.099641681 | 1 | | | | | | |
| YEARS_BIRTH | 0.049536299 | 0.160878908 | 0.099056566 | 0.155143369 | 0.048987416 | 1 | | | | | |
| YEARS_EMPLOYED | 0.036221572 | 0.094943177 | 0.06213425 | 0.095918555 | -0.005606334 | 0.352389434 | 1 | | | | |
| YEARS_REGISTRATION | -0.033500961 | 0.024294688 | 0.000851932 | 0.020392259 | 0.064621189 | 0.304733514 | 0.175692735 | 1 | | | |
| YEARS_ID_PUBLISH | 0.023115928 | 0.044246818 | 0.03222207 | 0.044495693 | 0.004355924 | 0.107692262 | 0.08250215 | 0.037023396 | 1 | | |
| FLAG_MOBIL | 0.00254311 | 0.004263478 | 0.00070066 | 0.004159523 | 0.003829664 | 0.008815734 | 0.004667144 | -0.000181687 | 0.005949554 | 1 | |
| FLAG_EMP_PHONE | 0.001840188 | -0.003321741 | -0.000809968 | -0.002120288 | 0.003952304 | -0.008240593 | 0.000482146 | 0.001072225 | -0.009423469 | -3.76647E-05 | |
| FLAG_WORK_PHONE | -0.081009347 | -0.034281389 | -0.053428591 | -0.00965972 | -0.019062959 | -0.040974482 | 0.013544183 | -0.013564483 | 0.014735981 | 0.002906452 | 0.00411 |
| FLAG_CONT_MOBILE | -0.016304619 | 0.027480203 | 0.026055469 | 0.024835936 | -0.006171892 | 0.00180144 | 0.001232114 | -0.005354334 | 0.000519242 | -0.00025715 | -0.00036 |
| FLAG_PHONE | 0.011981698 | 0.022918689 | 0.008119793 | 0.039816451 | 0.091760336 | 0.038540329 | 0.056860072 | 0.065065531 | 0.028924793 | 0.003197907 | -0.00362 |
| FLAG_EMAIL | 0.078193988 | 0.00476983 | 0.057543405 | 0.005050654 | 0.041495511 | -0.067389587 | -0.03088492 | -0.022046867 | -0.013549504 | 0.00133979 | 0.00189 |
| CNT_FAM_MEMBERS | -0.000842483 | 0.036171529 | 0.042829942 | 0.034141373 | -0.031006544 | -0.180331567 | -0.03596343 | -0.15121213 | 0.114272484 | 0.001406748 | 0.00198 |
| REGION_RATING_CLIENT | -0.206983514 | -0.105744089 | -0.13049337 | -0.108059985 | -0.544721325 | -0.049623826 | 0.015487867 | -0.101320619 | -0.006932647 | 0.000322541 | 0.00045 |
| REGION_RATING_CLIENT_W_CITY | -0.222990386 | -0.1145176 | -0.144012979 | -0.11611621 | -0.540963979 | -0.048821639 | 0.011631131 | -0.093658548 | -0.003029496 | 0.000107595 | 0.00015 |
| HOUR_APPR_PROCESS_START | 0.062155852 | 0.042415202 | 0.032978892 | 0.052014707 | 0.169741541 | -0.05249584 | -0.022972212 | 0.022382729 | -0.011940947 | -0.001218541 | -0.00283 |
| REG_REGION_NOT_LIVE_REGION | 0.07707556 | 0.024088354 | 0.042248024 | 0.02685031 | -0.010152917 | -0.050584554 | -0.051638535 | -0.018019288 | -0.024217535 | 0.000673598 | 0.00095 |
| REG_REGION_NOT_WORK_REGION | 0.150267056 | 0.052029517 | 0.076135793 | 0.053710538 | 0.069312685 | -0.037906108 | -0.084713495 | -0.014658288 | -0.020159836 | 0.001317332 | 0.00186 |
| LIVE_REGION_NOT_WORK_REGION | 0.142728268 | 0.052033499 | 0.070788918 | 0.052444038 | 0.096078968 | -0.013270406 | -0.067876651 | -0.004016713 | -0.00809165 | 0.001170453 | 0.00165 |
| REG_CITY_NOT_LIVE_CITY | -0.005244989 | -0.032441358 | -0.017797182 | -0.031000334 | -0.052726245 | -0.169778772 | -0.110955527 | -0.052540252 | -0.054539507 | 0.001601051 | 0.00227 |
| REG_CITY_NOT_WORK_CITY | -0.029395219 | -0.038075471 | -0.03108411 | -0.038105486 | -0.045446779 | -0.105586448 | -0.128027403 | -0.048038224 | -0.036453026 | 0.00319281 | 0.00451 |
| LIVE_CITY_NOT_WORK_CITY | -0.017596755 | -0.014437677 | -0.015525242 | -0.015274782 | -0.014286923 | -0.018314441 | -0.075450092 | -0.019570201 | -0.00396603 | 0.002704832 | 0.00382 |
| EXT_SOURCE_2 | 0.157592079 | 0.143329983 | 0.131789222 | 0.150147083 | 0.210410741 | 0.145947751 | 0.079985787 | 0.075749806 | 0.061074607 | -0.001112609 | -0.00413 |
| EXT_SOURCE_3 | -0.055446702 | 0.046896373 | 0.036043041 | 0.048434991 | -0.015944632 | 0.157059546 | 0.112506111 | 0.086551796 | 0.086764284 | -0.000559366 | -0.00176 |
| OBS_30_CNT_SOCIAL_CIRCLE | -0.034628833 | -0.003083082 | -0.012532838 | -0.003651531 | -0.019413993 | -0.022913038 | -0.002616498 | -0.017619762 | 0.009077835 | 0.003188915 | 0.00289 |
| DEF_30_CNT_SOCIAL_CIRCLE | -0.032942435 | -0.016637802 | -0.022537261 | -0.018242423 | 0.008881307 | -0.016976929 | -0.013442007 | -0.008043937 | -0.009179422 | 0.001610554 | 0.00227 |
| OBS_60_CNT_SOCIAL_CIRCLE | -0.034754311 | -0.002758599 | -0.012217567 | -0.00340683 | -0.018358055 | -0.022686332 | -0.002316044 | -0.017610276 | 0.009490502 | 0.003177775 | 0.00286 |
| DEF_60_CNT_SOCIAL_CIRCLE | -0.033893862 | -0.020419982 | -0.024979398 | -0.021464651 | 0.00209051 | -0.0190726 | -0.014226927 | -0.011579482 | -0.009590063 | 0.001370527 | 0.00193 |
| DAYS_LAST_PHONE_CHANGE | 0.039981388 | 0.072724145 | 0.065528991 | 0.076329889 | 0.046497327 | 0.135575222 | 0.137202222 | 0.070541372 | 0.108975026 | 0.006126366 | -0.00135 |
| FLAG_DOCUMENT_2 | -0.002091503 | 0.010328028 | 0.00458807 | 0.012078696 | -0.005133667 | -0.001874196 | -0.002975966 | 0.004347061 | 0.005253938 | 2.66326E-05 | 3.76647 |
| FLAG_DOCUMENT_3 | -0.112120419 | 0.084011213 | 0.076960043 | 0.058286399 | -0.103877667 | 0.080901319 | 0.056075182 | 0.022517612 | 0.021733936 | 0.009147272 | 0.00440 |
| FLAG_DOCUMENT_4 | -0.000629495 | 0.00392641 | 0.003826322 | 0.00392409 | 0.006600493 | 0.0046708 | 0.003172034 | 0.008809352 | 0.004319347 | 7.99062E-05 | 0.00011 |
| FLAG_DOCUMENT_5 | 0.010285533 | -0.008857849 | -0.006896385 | -0.002665309 | 0.014266335 | -0.002673407 | 0.004538925 | 0.007832391 | 0.003479546 | 0.000676859 | 0.00095 |
| FLAG_DOCUMENT_6 | -0.007960752 | 0.023951297 | 0.012196585 | 0.022567668 | -0.002797636 | 0.120021504 | 0.054185158 | 0.039204737 | 0.041311219 | 0.000499859 | -0.03732 |

| B | C | D | E |
|---|---|---|---|

| Top 5 Correlation (Non-Defaulters) | | |
|---|---|---|
| **Variable 1** | **Variable 2** | **Correlation** |
| OBS_60_CNT_SOCIAL_CIRCLE | OBS_30_CNT_SOCIAL_CIRCLE | 0.998 |
| AMT_GOODS_PRICE | AMT_CREDIT | 0.986 |
| REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.948 |
| LIVE_REGION_NOT_WORK_REGION | REG_REGION_NOT_WORK_REGION | 0.861 |
| DEF_60_CNT_SOCIAL_CIRCLE | DEF_30_CNT_SOCIAL_CIRCLE | 0.853 |

| Top 5 Correlation (Defaulters) | | |
|---|---|---|
| **Variable 1** | **Variable 2** | **Correlation** |
| OBS_60_CNT_SOCIAL_CIRCLE | OBS_30_CNT_SOCIAL_CIRCLE | 0.998 |
| AMT_GOODS_PRICE | AMT_CREDIT | 0.982 |
| REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.951 |
| DEF_60_CNT_SOCIAL_CIRCLE | DEF_30_CNT_SOCIAL_CIRCLE | 0.891 |
| LIVE_REGION_NOT_WORK_REGION | REG_REGION_NOT_WORK_REGION | 0.806 |

# LINKS

## EXCEL SHEET

https://docs.google.com/spreadsheets/d/1v9-xxUrDClQv0AKGV6XI7dska5lKu5Y3/edit?usp=sharing&ouid=108286913145936487778&rtpof=true&sd=true

## LOOM VIDEO

https://www.loom.com/share/0bb4d6886c954c4081a76613f925af74?sid=23bc2763-2581-4188-9c76-86a824dc7a84