**Bank Loan Case Study**

**Project Description:** In this project I was working as a data analyst at a finance company that specializes in lending various types of loans to urban customers. My company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. My task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

1. **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.

**My 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.

1. **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.

**My Task:** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

1. **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.

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

1. **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.

**My 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.

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

**My 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.

**Approach:** First Igone through dataset to know all the columns present in the table. Then I saw all the questions and thought of functions which could be used to answer each question. After that I applied those functions and found the answer to each question and plotted the graph wherever was required.

**Tech-Stack Used:** The software used for the project is Microsoft Excel 365. It is used to run the functions and get answers of each question. It is also used to plot the graphs.

**Insights:**

1. Identify Missing Data and Deal with it Appropriately:

Function:-

First, I found percentage of blank data in each column. I deleted all columns in which percentage of blank data was more than 25%.

Then I filled blank cells as follow:

=IF(ISBLANK(application\_data!K2),MEDIAN(application\_data!K$2:K$50000),application\_data!K2)

Output:-

<https://docs.google.com/spreadsheets/d/1OZwBS2eC2zznkfd8_TQyGbaM46sVeKNZ/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true>

|  |  |
| --- | --- |
| **Column** | **Percentage\_of\_black\_cells** |
| SK\_ID\_PREV | 0 |
| SK\_ID\_CURR | 0 |
| NAME\_CONTRACT\_TYPE | 0 |
| AMT\_ANNUITY | 21.18442369 |
| AMT\_APPLICATION | 0 |
| AMT\_CREDIT | 0 |
| AMT\_DOWN\_PAYMENT | 50.39700794 |
| AMT\_GOODS\_PRICE | 21.48842977 |
| WEEKDAY\_APPR\_PROCESS\_START | 0 |
| HOUR\_APPR\_PROCESS\_START | 0 |
| FLAG\_LAST\_APPL\_PER\_CONTRACT | 0 |
| NFLAG\_LAST\_APPL\_IN\_DAY | 0 |
| RATE\_DOWN\_PAYMENT | 50.39700794 |
| RATE\_INTEREST\_PRIMARY | 99.6699934 |
| RATE\_INTEREST\_PRIVILEGED | 99.6699934 |
| NAME\_CASH\_LOAN\_PURPOSE | 0 |
| NAME\_CONTRACT\_STATUS | 0 |
| DAYS\_DECISION | 0 |
| NAME\_PAYMENT\_TYPE | 0 |
| CODE\_REJECT\_REASON | 0 |
| NAME\_TYPE\_SUITE | 48.48696974 |
| NAME\_CLIENT\_TYPE | 0 |
| NAME\_GOODS\_CATEGORY | 0 |
| NAME\_PORTFOLIO | 0 |
| NAME\_PRODUCT\_TYPE | 0 |
| CHANNEL\_TYPE | 0 |
| SELLERPLACE\_AREA | 0 |
| NAME\_SELLER\_INDUSTRY | 0 |
| CNT\_PAYMENT | 21.18442369 |
| NAME\_YIELD\_GROUP | 0 |
| PRODUCT\_COMBINATION | 0.01600032 |
| DAYS\_FIRST\_DRAWING | 38.32076642 |
| DAYS\_FIRST\_DUE | 38.32076642 |
| DAYS\_LAST\_DUE\_1ST\_VERSION | 38.32076642 |
| DAYS\_LAST\_DUE | 38.32076642 |
| DAYS\_TERMINATION | 38.32076642 |
| NFLAG\_INSURED\_ON\_APPROVAL | 38.32076642 |

Graph:

1. Identify Outliers in the Dataset:

Function:-

=QUARTILE.EXC(H2:H50000,1)

=QUARTILE.EXC(H2:H50000,3)

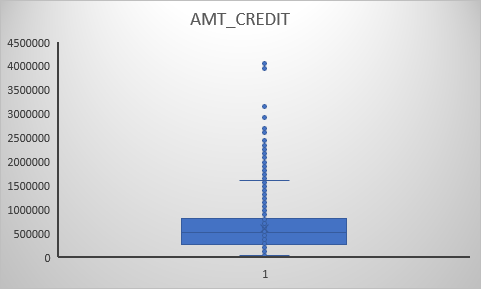
Then I used conditional formatting to highlight the values which are less than lower bound and higher than upper bound.

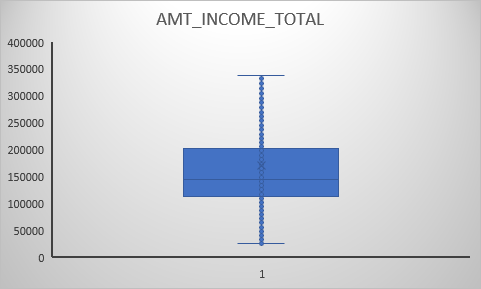
Output:-

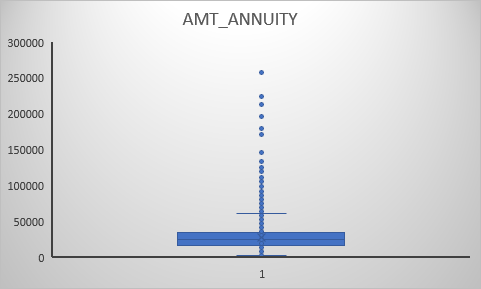
<https://docs.google.com/spreadsheets/d/1JYdPdrTVUm-1J8ab20tFfIjHCyqwYMkv/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true>

|  |  |  |  |
| --- | --- | --- | --- |
|  | **AMT\_INCOME\_TOTAL** | **AMT\_CREDIT** | **AMT\_ANNUITY** |
| lower bound | 112500 | 270000 | 16456.5 |
| upper bound | 202500 | 808650 | 34596 |
|  |  |  |  |

Graph:







1. Analyze Data Imbalance:

Function:-

First, I found the age of people using the following function:

=ROUND(R2:R50000/365\*-1,0)

Then I found occurrences of each unique elements in different column.

=UNIQUE(C2:C50000)

=COUNTIF(C$2:C$50000,BX2)

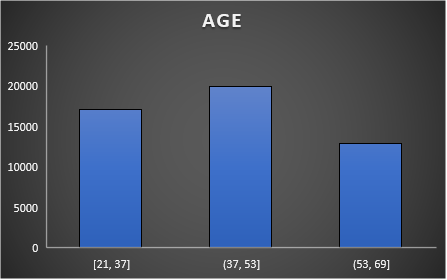
After that I plotted the graphs.

Output:-

<https://docs.google.com/spreadsheets/d/1qaTJqUwWq5NDogJFvwFH9y_-6ote5GFk/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true>

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **NAME\_CONTRACT\_TYPE** | **Occurrence** | **GENDER** | **Occurrence** | **EDUCATION\_TYPE** | **Occurrence** |
| Cash loans | 45276 | M | 17174 | Secondary / secondary special | 35572 |
| Revolving loans | 4723 | F | 32823 | Higher education | 12167 |
|  |  | XNA | 2 | Incomplete higher | 1620 |
|  |  |  |  | Lower secondary | 620 |
|  |  |  |  | Academic degree | 20 |

Graph:



1. Perform Univariate, Segmented Univariate, and Bivariate Analysis:

Function:-

=AVERAGE(H2:H50000)

=MEDIAN(H2:H50000)

=MODE(H2:H50000)

Then I found unique values for HOUSING\_TYPE, FAMILY\_STATUS, INCOME\_TYPE and occurrences of each unique value.

=UNIQUE(P2:P50000)

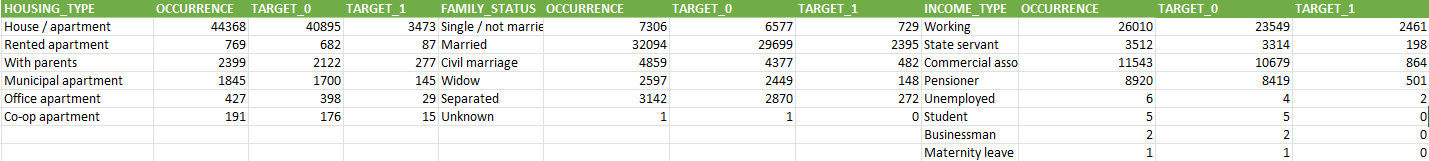
=COUNTIF(P$2:P$50000,CD2)

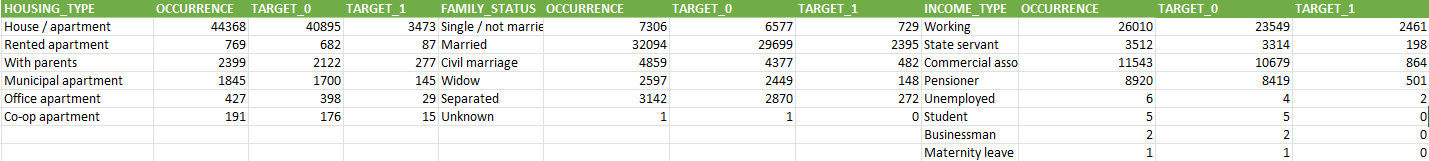
Then I plotted graphs for different variables

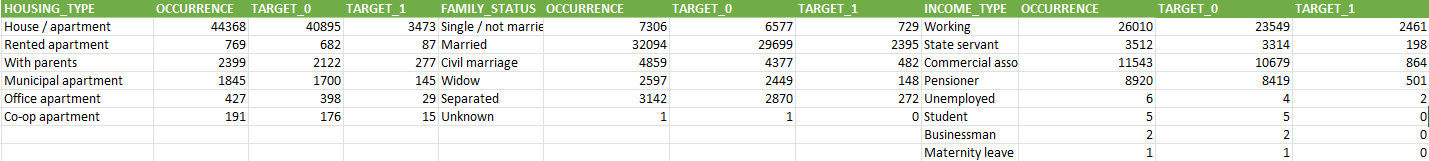
Output:-

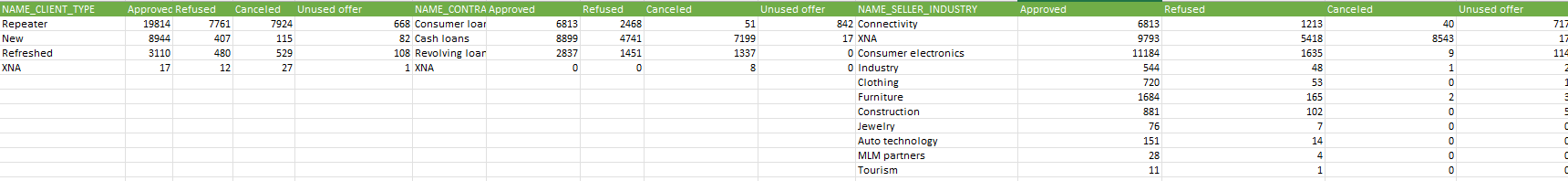
<https://docs.google.com/spreadsheets/d/1zK1mjbyOD_t4vjcyNqw5a7TYdXmONp7k/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true>

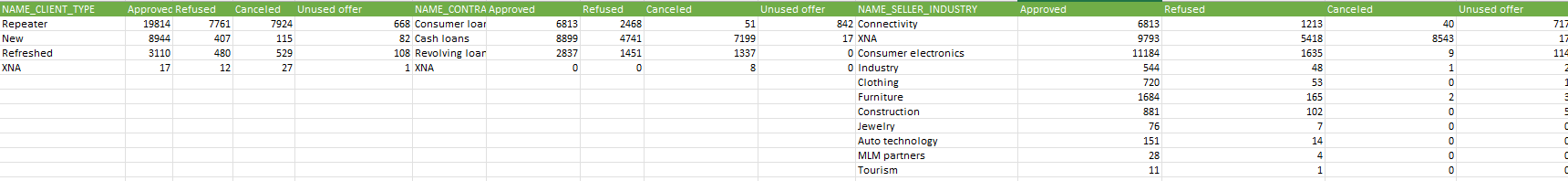
|  |  |  |  |
| --- | --- | --- | --- |
|  | **AMT\_INCOME\_TOTAL** | **AMT\_CREDIT** | **AGE** |
| AVERAGE | 170767.5905 | 599700.5815 | 43.8975 |
| MEDIAN | 145800 | 514777.5 | 43 |
| MODE | 135000 | 450000 | 39 |

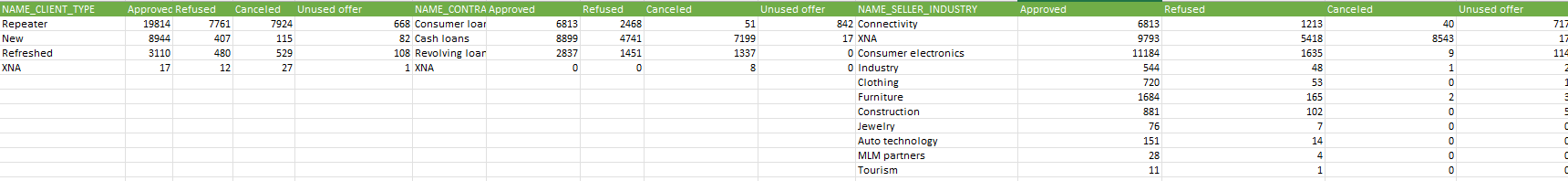






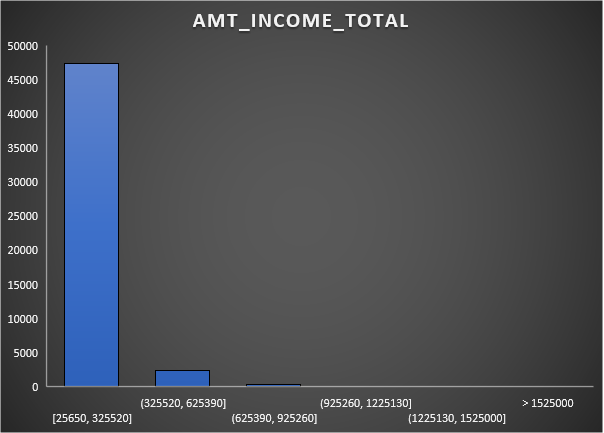


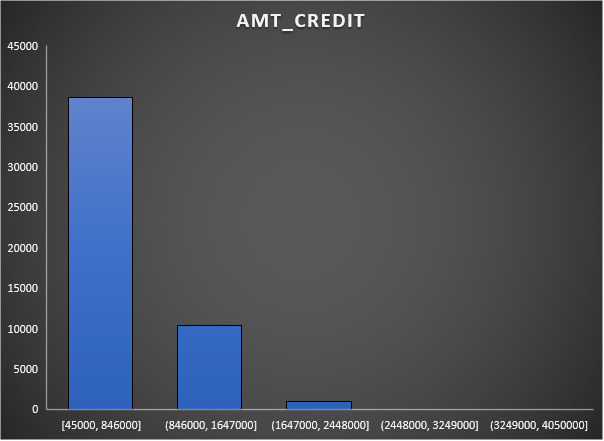


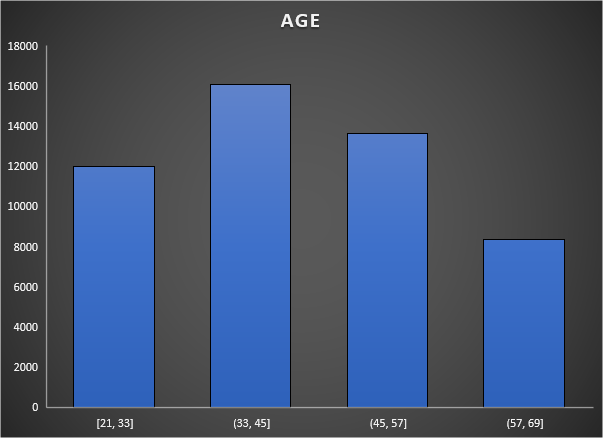


Graph:

Univariate







Bivariate

1. Identify Top Correlations for Different Scenarios:

Function:-

First, I found correlation between target and various columns by using following function:

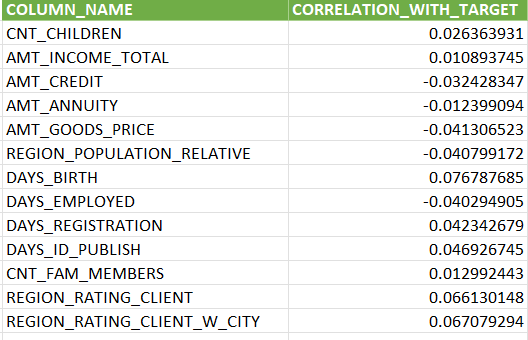
=CORREL(G2:G50000,B2:B50000)

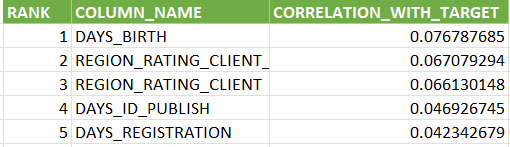
Then I found top five correlation among them:

=INDEX(SORTBY(BX2:BY15,BY2:BY15,-1),SEQUENCE(5),{1,2})

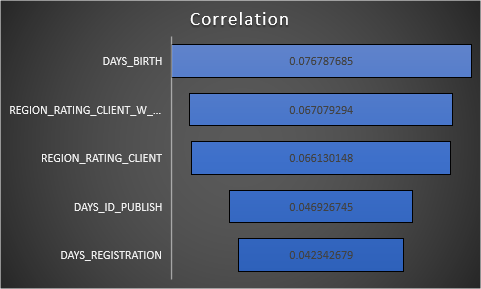
Output:-

<https://docs.google.com/spreadsheets/d/1048NbwR9YLTf3777YbH3AvcKg0dcThUF/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true>





Graph:



**Results:**

1. Identify Outliers in the Dataset:

There are many outliers in the data.

1. Analyze Data Imbalance:

Data is imbalance in most of the columns.

1. Perform Univariate, Segmented Univariate, and Bivariate Analysis:

People who are married, have low salary and live in house/apartment are most likely to take loan.

1. Identify Top Correlations for Different Scenarios:

Highest correlation with TARGET is of DAYS\_BIRTH .