# Bank Loan Case Study

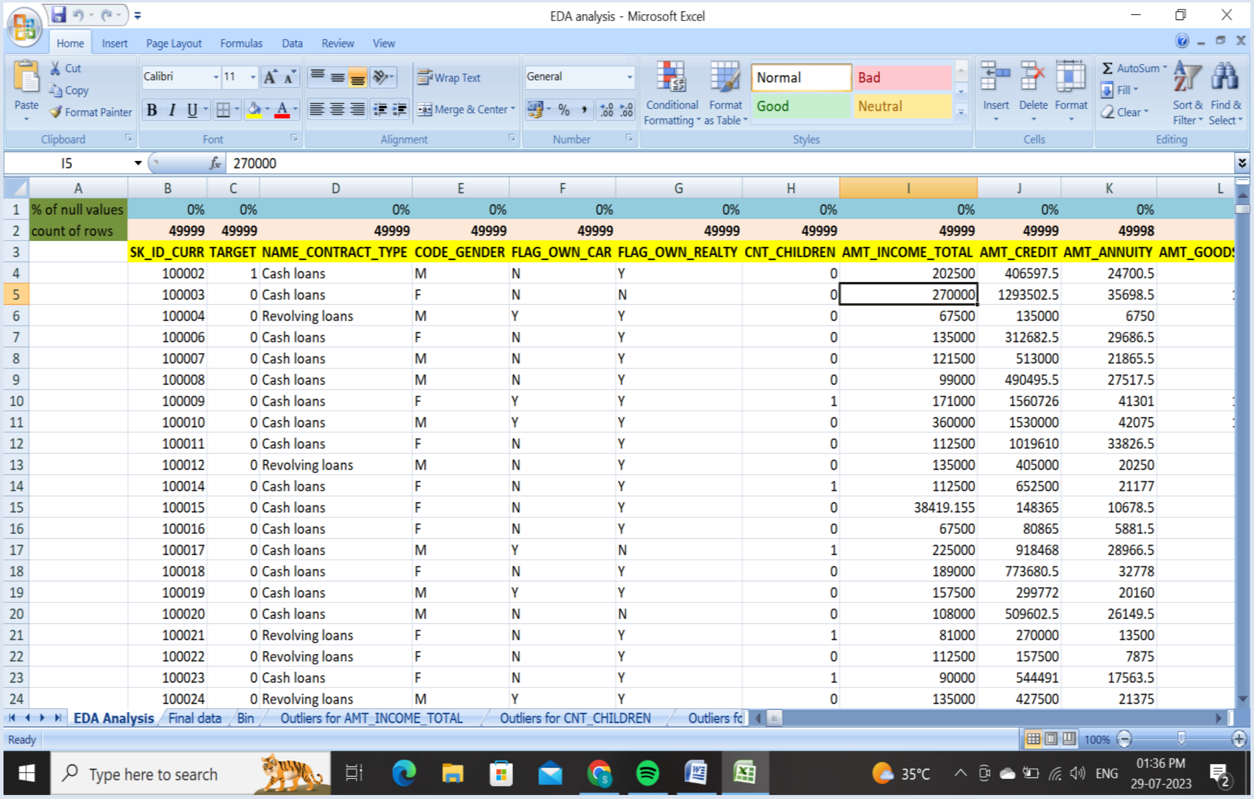
## Final Project-2

Sowmiya R | Data Analytics | 30th July 2023

# PROJECT DESCRIPTION

Imagine myself as a data analyst at a finance company that specializes in lending various types of loans to urban customers. Your company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans.

Your task is to use **Exploratory Data Analysis (EDA)** to analyze patterns in the data and ensure that capable applicants are not rejected.



I5 270000

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | % of null values | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | |
| 2 | count of rows | 49999 | 49999 | 49999 | 49999 | 49999 | 49999 | 49999 | 49999 | 49999 | 49998 | |
| 3 | | 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 |
| 4 | | 100002 | 1 | Cash loans | M | N | Y | 0 | 202500 | 406597.5 | 24700.5 | |
| 5 | | 100003 | 0 | Cash loans | F | N | N | 0 | 270000 | 1293502.5 | 35698.5 | |
| 6 | | 100004 | 0 | Revolving loans | M | Y | Y | 0 | 67500 | 135000 | 6750 | |
| 7 | | 100006 | 0 | Cash loans | F | N | Y | 0 | 135000 | 312682.5 | 29686.5 | |
| 8 | | 100007 | 0 | Cash loans | M | N | Y | 0 | 121500 | 513000 | 21865.5 | |
| 9 | | 100008 | 0 | Cash loans | M | N | Y | 0 | 99000 | 490495.5 | 27517.5 | |
| 10 | | 100009 | 0 | Cash loans | F | Y | Y | 1 | 171000 | 1560726 | 41301 | |
| 11 | | 100010 | 0 | Cash loans | M | Y | Y | 0 | 360000 | 1530000 | 42075 | |
| 12 | | 100011 | 0 | Cash loans | F | N | Y | 0 | 112500 | 1019610 | 33826.5 | |
| 13 | | 100012 | 0 | Revolving loans | M | N | Y | 0 | 135000 | 405000 | 20250 | |
| 14 | | 100014 | 0 | Cash loans | F | N | Y | 1 | 112500 | 652500 | 21177 | |
| 15 | | 100015 | 0 | Cash loans | F | N | Y | 0 | 38419.155 | 148365 | 10678.5 | |
| 16 | | 100016 | 0 | Cash loans | F | N | Y | 0 | 67500 | 80865 | 5881.5 | |
| 17 | | 100017 | 0 | Cash loans | M | Y | N | 1 | 225000 | 918468 | 28966.5 | |
| 18 | | 100018 | 0 | Cash loans | F | N | Y | 0 | 189000 | 773680.5 | 32778 | |
| 19 | | 100019 | 0 | Cash loans | M | Y | Y | 0 | 157500 | 299772 | 20160 | |
| 20 | | 100020 | 0 | Cash loans | M | N | N | 0 | 108000 | 509602.5 | 26149.5 | |
| 21 | | 100021 | 0 | Revolving loans | F | N | Y | 1 | 81000 | 270000 | 13500 | |
| 22 | | 100022 | 0 | Revolving loans | F | N | Y | 0 | 112500 | 157500 | 7875 | |
| 23 | | 100023 | 0 | Cash loans | F | N | Y | 1 | 90000 | 544491 | 17563.5 | |
| 24 | | 100024 | 0 | Revolving loans | M | Y | Y | 0 | 135000 | 427500 | 21375 | |

From above picture , can see a picture done a EDA analysis .I have attached a excel file which is included as a EDA analysis and final data….

https://docs.google.com/spreadsheets/d/1C9u00Xftuc0-4z5Tpg-zo4FtUPvboujw/edit?usp=drive_link&ouid=103768596710140113695&rtpof=true&sd=true

**When a customer applies for a loan, your company faces two risks:**

- If the applicant can repay the loan but is not approved, the company loses business.
- If the applicant cannot repay the loan and is approved, the company faces a financial loss.
- The dataset you'll be working with contains information about loan applications.

It includes two types of scenarios:

- **Customers with payment difficulties:** These are customers who had a late payment of more than X days on at least one of the first Y installments of the loan.
- **All other cases:** These are cases where the payment was made on time.

When a customer applies for a loan, there are four possible outcomes:

- **Approved:** The company has approved the loan application.
- **Cancelled:** The customer cancelled the application during the approval process.
- **Refused:** The company rejected the loan.
- **Unused Offer:** The loan was approved but the customer did not use it.

I need to be use the EDA analysis final dataset for the whole tasks of this project. Goal in this project is to use EDA to understand how customer attributes and loan attributes influence the likelihood of default.

### Business Objectives:

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.

**APPROACH**

- First of all , in this project they have provided with a 2 datasets which are **current application** and **previous application**. So I have started cleaning the datasets and grouped as one, removed blanks, unwanted and unrelated columns and so on.

- I thought that in the given datasets unneeded columns are present there so I thought to remove the useless columns which are not needed for risk analysis. I analyzed in given reference that helped me to analyze this risk assessment analysis.

- Next, I am supposed to find the outliers then removed the outliers, then I count the blanks of making it percentage how much the blanks are present so the percentage helped me to remove the unwanted columns and blanks by using excel **count functions.**

- By finding and cleaning all these data and moving on to tasks , my excel file named as **Final data** helped me to calculate the tasks and visually given a various insights like bar chart, box plot, column chart by make using excel pivot tables.

## TECH-STACK USED

I have used Microsoft excel 2011 and Tableau public . For the whole, MS Excel is been used for all calculating and analyzing the tasks and driven a multiple insights with a visual representation.

## INSIGHTS AND RESULTS

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.

**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.**

### i) Application.csv Dataset

In this task, I have analyzed a missing data and unwanted , unrelated columns and blanks . So I used a excel count functions to see a percentage of each and every columns calculated and kept as a separate sheet as **EDA analysis** analyzed every other columns with a percentage and comes with a final dataset that too I kept as a separate sheet as **Final Data** .

- 'application data.csv' contains all of the client's information at the time of application. The information pertains to whether or not a client is having financial issues.
- 'previous application.csv' provides data from the client's previous loans. It indicates if the prior application was Accepted, Cancelled, Refused, or Unused.

**Categorical variables** (non-numerical variables)- person's occupation, education status.

**Numerical variables** - income, credit etc.,

The following are some of the categorical and numerical variables from the provided data set.

**Categorical variables :**

- Gender
- Name contract type

- Income type
- Education
- Housing type

**Numeric variables :**

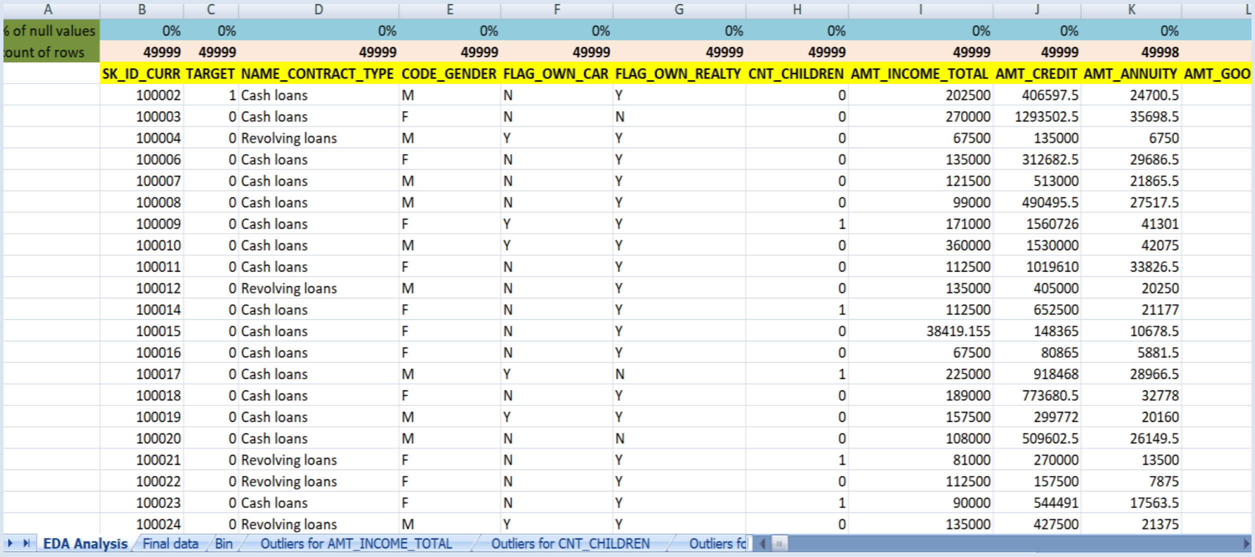
- Age
- Days Employed
- Amount Income
- Amount Annuity
- Amount Credit

There I have left with a 30 header rows and 307512 rows are left there. Below are these of 30 header row names listed here:

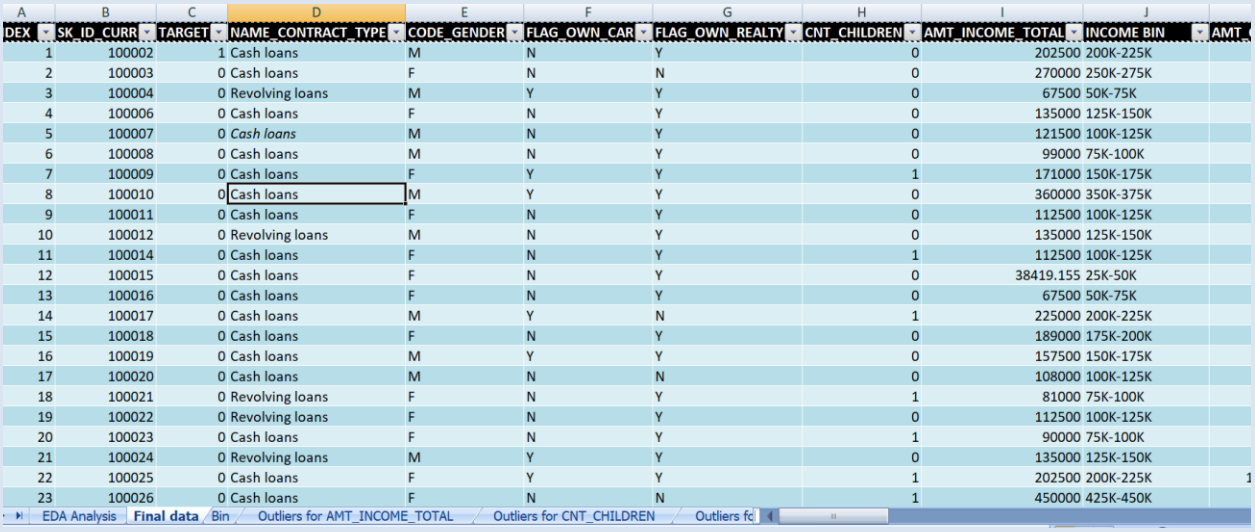
- **INDEX**
- **SK_ID_CURR**
- **TARGET**
- **NAME_CONTRACT_TYPE**
- **CODE_GENDER**
- **FLAG_OWN_CAR**
- **FLAG_OWN_REALTY**
- **CNT_CHILDREN**

- **AMT_INCOME_TOTAL**
- **INCOME BIN**
- **AMT_CREDIT**
- **CREDIT BIN**
- **AMT_ANNUITY**
- **AMT_GOODS_PRICE**
- **NAME_INCOME_TYPE**
- **NAME_EDUCATION_TYPE**
- **NAME_FAMILY_STATUS**
- **NAME_HOUSING_TYPE**
- **REGION_POPULATION_RELATIVE**
- **DAYS_BIRTH(yrs)**
- **DAYS_EMPLOYED(YRS)**
- **DAYS_REG (YRS)**
- **DAYS_ID_PUBLISH(YRS)**
- **FLAG_MOBIL**
- **FLAG_CONT_MOBILE**
- **CNT_FAM_MEMBERS**
- **REGION_RATING_CLIENT**
- **WEEKDAY_APPR_PROCESS_START**

- **HOUR_APPR_PROCESS_START**
- **ORGANIZATION_TYPE**



| A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|
| % of null values | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | |
| ount of rows | 49999 | 49999 | 49999 | 49999 | 49999 | 49999 | 49999 | 49999 | 49999 | 49998 | |
| | 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_GOO |
| | 100002 | 1 | Cash loans | M | N | Y | 0 | 202500 | 406597.5 | 24700.5 | |
| | 100003 | 0 | Cash loans | F | N | N | 0 | 270000 | 1293502.5 | 35698.5 | |
| | 100004 | 0 | Revolving loans | M | Y | Y | 0 | 67500 | 135000 | 6750 | |
| | 100006 | 0 | Cash loans | F | N | Y | 0 | 135000 | 312682.5 | 29686.5 | |
| | 100007 | 0 | Cash loans | M | N | Y | 0 | 121500 | 513000 | 21865.5 | |
| | 100008 | 0 | Cash loans | M | N | Y | 0 | 99000 | 490495.5 | 27517.5 | |
| | 100009 | 0 | Cash loans | F | Y | Y | 1 | 171000 | 1560726 | 41301 | |
| | 100010 | 0 | Cash loans | M | Y | Y | 0 | 360000 | 1530000 | 42075 | |
| | 100011 | 0 | Cash loans | F | N | Y | 0 | 112500 | 1019610 | 33826.5 | |
| | 100012 | 0 | Revolving loans | M | N | Y | 0 | 135000 | 405000 | 20250 | |
| | 100014 | 0 | Cash loans | F | N | Y | 1 | 112500 | 652500 | 21177 | |
| | 100015 | 0 | Cash loans | F | N | Y | 0 | 38419.155 | 148365 | 10678.5 | |
| | 100016 | 0 | Cash loans | F | N | Y | 0 | 67500 | 80865 | 5881.5 | |
| | 100017 | 0 | Cash loans | M | Y | N | 1 | 225000 | 918468 | 28966.5 | |
| | 100018 | 0 | Cash loans | F | N | Y | 0 | 189000 | 773680.5 | 32778 | |
| | 100019 | 0 | Cash loans | M | Y | Y | 0 | 157500 | 299772 | 20160 | |
| | 100020 | 0 | Cash loans | M | N | N | 0 | 108000 | 509602.5 | 26149.5 | |
| | 100021 | 0 | Revolving loans | F | N | Y | 1 | 81000 | 270000 | 13500 | |
| | 100022 | 0 | Revolving loans | F | N | Y | 0 | 112500 | 157500 | 7875 | |
| | 100023 | 0 | Cash loans | F | N | Y | 1 | 90000 | 544491 | 17563.5 | |
| | 100024 | 0 | Revolving loans | M | Y | Y | 0 | 135000 | 427500 | 21375 | |



| A | B | C | D | E | F | G | H | I | J | |
|---|---|---|---|---|---|---|---|---|---|---|
| DEX | SK_ID_CURR | TARGET | NAME_CONTRACT_TYPE | CODE_GENDER | FLAG_OWN_CAR | FLAG_OWN_REALTY | CNT_CHILDREN | AMT_INCOME_TOTAL | INCOME BIN | AMT_ |
| 1 | 100002 | 1 | Cash loans | M | N | Y | 0 | 202500 | 200K-225K | |
| 2 | 100003 | 0 | Cash loans | F | N | N | 0 | 270000 | 250K-275K | |
| 3 | 100004 | 0 | Revolving loans | M | Y | Y | 0 | 67500 | 50K-75K | |
| 4 | 100006 | 0 | Cash loans | F | N | Y | 0 | 135000 | 125K-150K | |
| 5 | 100007 | 0 | *Cash loans* | M | N | Y | 0 | 121500 | 100K-125K | |
| 6 | 100008 | 0 | Cash loans | M | N | Y | 0 | 99000 | 75K-100K | |
| 7 | 100009 | 0 | Cash loans | F | Y | Y | 1 | 171000 | 150K-175K | |
| 8 | 100010 | 0 | Cash loans | M | Y | Y | 0 | 360000 | 350K-375K | |
| 9 | 100011 | 0 | Cash loans | F | N | Y | 0 | 112500 | 100K-125K | |
| 10 | 100012 | 0 | Revolving loans | M | N | Y | 0 | 135000 | 125K-150K | |
| 11 | 100014 | 0 | Cash loans | F | N | Y | 1 | 112500 | 100K-125K | |
| 12 | 100015 | 0 | Cash loans | F | N | Y | 0 | 38419.155 | 25K-50K | |
| 13 | 100016 | 0 | Cash loans | F | N | Y | 0 | 67500 | 50K-75K | |
| 14 | 100017 | 0 | Cash loans | M | Y | N | 1 | 225000 | 200K-225K | |
| 15 | 100018 | 0 | Cash loans | F | N | Y | 0 | 189000 | 175K-200K | |
| 16 | 100019 | 0 | Cash loans | M | Y | Y | 0 | 157500 | 150K-175K | |
| 17 | 100020 | 0 | Cash loans | M | N | N | 0 | 108000 | 100K-125K | |
| 18 | 100021 | 0 | Revolving loans | F | N | Y | 1 | 81000 | 75K-100K | |
| 19 | 100022 | 0 | Revolving loans | F | N | Y | 0 | 112500 | 100K-125K | |
| 20 | 100023 | 0 | Cash loans | F | N | Y | 1 | 90000 | 75K-100K | |
| 21 | 100024 | 0 | Revolving loans | M | Y | Y | 0 | 135000 | 125K-150K | |
| 22 | 100025 | 0 | Cash loans | F | Y | Y | 1 | 202500 | 200K-225K | 1 |
| 23 | 100026 | 0 | Cash loans | F | N | N | 1 | 450000 | 425K-450K | |

From the above picture, I have represented the EDA analysis by percentage values and final dataset which are cleaned are pasted above as picture. Also then I have included a excel file as **EDA analysis** and **Final data**. This can be referred as in this file…

https://docs.google.com/spreadsheets/d/1RwBFqWdMVj8QUFEmBDAnmrCg8qgD0uwC/edit?usp=drive_link

ii) **Previous application.csv dataset**

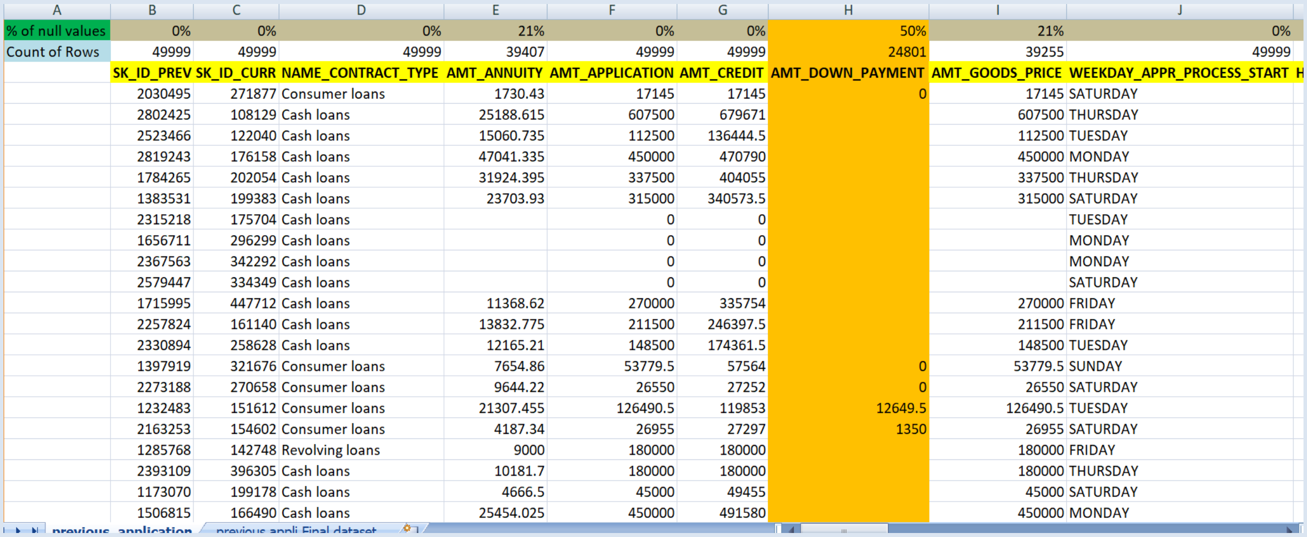
'previous application.csv' provides data from the client's previous loans. It indicates if the prior application was Accepted, Cancelled, Refused, or Unused.

From this previous application dataset , I have started cleaning the dataset by removing every blanks on calculating in excel itself, then I found a percentage every other header columns and removed which were unwanted and the percentage is above 25%.

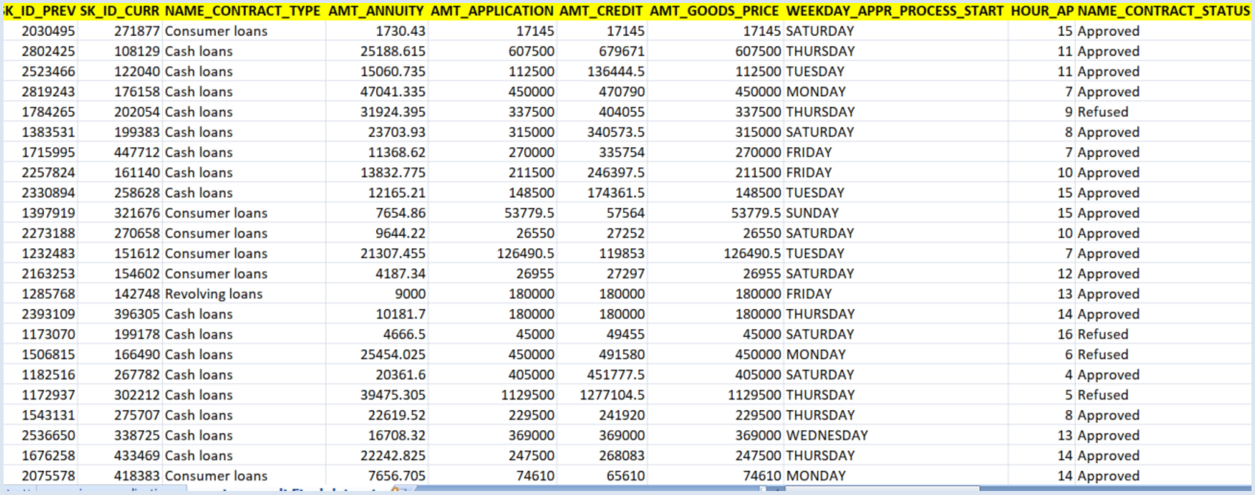
So finally I arrived at **39256** entries are left and 17 header rows. They are:

- 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
- NAME_CONTRACT_STATUS
- DAYS_DECISION
- NAME_PAYMENT_TYPE
- CODE_REJECT_REASON
- NAME_CLIENT_TYPE
- NAME_GOODS_CATEGORY
- CNT_PAYMENT
- PRODUCT_COMBINATION



| A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|
| % of null values | 0% | 0% | 0% | 21% | 0% | 0% | 50% | 21% | 0% |
| Count of Rows | 49999 | 49999 | 49999 | 39407 | 49999 | 49999 | 24801 | 39255 | 49999 |
| | SK_ID_PREV | SK_ID_CURR | NAME_CONTRACT_TYPE | AMT_ANNUITY | AMT_APPLICATION | AMT_CREDIT | AMT_DOWN_PAYMENT | AMT_GOODS_PRICE | WEEKDAY_APPR_PROCESS_START |
| | 2030495 | 271877 | Consumer loans | 1730.43 | 17145 | 17145 | 0 | 17145 | SATURDAY |
| | 2802425 | 108129 | Cash loans | 25188.615 | 607500 | 679671 | | 607500 | THURSDAY |
| | 2523466 | 122040 | Cash loans | 15060.735 | 112500 | 136444.5 | | 112500 | TUESDAY |
| | 2819243 | 176158 | Cash loans | 47041.335 | 450000 | 470790 | | 450000 | MONDAY |
| | 1784265 | 202054 | Cash loans | 31924.395 | 337500 | 404055 | | 337500 | THURSDAY |
| | 1383531 | 199383 | Cash loans | 23703.93 | 315000 | 340573.5 | | 315000 | SATURDAY |
| | 2315218 | 175704 | Cash loans | | 0 | 0 | | | TUESDAY |
| | 1656711 | 296299 | Cash loans | | 0 | 0 | | | MONDAY |
| | 2367563 | 342292 | Cash loans | | 0 | 0 | | | MONDAY |
| | 2579447 | 334349 | Cash loans | | 0 | 0 | | | SATURDAY |
| | 1715995 | 447712 | Cash loans | 11368.62 | 270000 | 335754 | | 270000 | FRIDAY |
| | 2257824 | 161140 | Cash loans | 13832.775 | 211500 | 246397.5 | | 211500 | FRIDAY |
| | 2330894 | 258628 | Cash loans | 12165.21 | 148500 | 174361.5 | | 148500 | TUESDAY |
| | 1397919 | 321676 | Consumer loans | 7654.86 | 53779.5 | 57564 | 0 | 53779.5 | SUNDAY |
| | 2273188 | 270658 | Consumer loans | 9644.22 | 26550 | 27252 | 0 | 26550 | SATURDAY |
| | 1232483 | 151612 | Consumer loans | 21307.455 | 126490.5 | 119853 | 12649.5 | 126490.5 | TUESDAY |
| | 2163253 | 154602 | Consumer loans | 4187.34 | 26955 | 27297 | 1350 | 26955 | SATURDAY |
| | 1285768 | 142748 | Revolving loans | 9000 | 180000 | 180000 | | 180000 | FRIDAY |
| | 2393109 | 396305 | Cash loans | 10181.7 | 180000 | 180000 | | 180000 | THURSDAY |
| | 1173070 | 199178 | Cash loans | 4666.5 | 45000 | 49455 | | 45000 | SATURDAY |
| | 1506815 | 166490 | Cash loans | 25454.025 | 450000 | 491580 | | 450000 | MONDAY |



| K_ID_PREV | SK_ID_CURR | NAME_CONTRACT_TYPE | AMT_ANNUITY | AMT_APPLICATION | AMT_CREDIT | AMT_GOODS_PRICE | WEEKDAY_APPR_PROCESS_START | HOUR_AP | NAME_CONTRACT_STATUS |
|---|---|---|---|---|---|---|---|---|---|
| 2030495 | 271877 | Consumer loans | 1730.43 | 17145 | 17145 | 17145 | SATURDAY | 15 | Approved |
| 2802425 | 108129 | Cash loans | 25188.615 | 607500 | 679671 | 607500 | THURSDAY | 11 | Approved |
| 2523466 | 122040 | Cash loans | 15060.735 | 112500 | 136444.5 | 112500 | TUESDAY | 11 | Approved |
| 2819243 | 176158 | Cash loans | 47041.335 | 450000 | 470790 | 450000 | MONDAY | 7 | Approved |
| 1784265 | 202054 | Cash loans | 31924.395 | 337500 | 404055 | 337500 | THURSDAY | 9 | Refused |
| 1383531 | 199383 | Cash loans | 23703.93 | 315000 | 340573.5 | 315000 | SATURDAY | 8 | Approved |
| 1715995 | 447712 | Cash loans | 11368.62 | 270000 | 335754 | 270000 | FRIDAY | 7 | Approved |
| 2257824 | 161140 | Cash loans | 13832.775 | 211500 | 246397.5 | 211500 | FRIDAY | 10 | Approved |
| 2330894 | 258628 | Cash loans | 12165.21 | 148500 | 174361.5 | 148500 | TUESDAY | 15 | Approved |
| 1397919 | 321676 | Consumer loans | 7654.86 | 53779.5 | 57564 | 53779.5 | SUNDAY | 15 | Approved |
| 2273188 | 270658 | Consumer loans | 9644.22 | 26550 | 27252 | 26550 | SATURDAY | 10 | Approved |
| 1232483 | 151612 | Consumer loans | 21307.455 | 126490.5 | 119853 | 126490.5 | TUESDAY | 7 | Approved |
| 2163253 | 154602 | Consumer loans | 4187.34 | 26955 | 27297 | 26955 | SATURDAY | 12 | Approved |
| 1285768 | 142748 | Revolving loans | 9000 | 180000 | 180000 | 180000 | FRIDAY | 13 | Approved |
| 2393109 | 396305 | Cash loans | 10181.7 | 180000 | 180000 | 180000 | THURSDAY | 14 | Approved |
| 1173070 | 199178 | Cash loans | 4666.5 | 45000 | 49455 | 45000 | SATURDAY | 16 | Refused |
| 1506815 | 166490 | Cash loans | 25454.025 | 450000 | 491580 | 450000 | MONDAY | 6 | Refused |
| 1182516 | 267782 | Cash loans | 20361.6 | 405000 | 451777.5 | 405000 | SATURDAY | 4 | Approved |
| 1172937 | 302212 | Cash loans | 39475.305 | 1129500 | 1277104.5 | 1129500 | THURSDAY | 5 | Refused |
| 1543131 | 275707 | Cash loans | 22619.52 | 229500 | 241920 | 229500 | THURSDAY | 8 | Approved |
| 2536650 | 338725 | Cash loans | 16708.32 | 369000 | 369000 | 369000 | WEDNESDAY | 13 | Approved |
| 1676258 | 433469 | Cash loans | 22242.825 | 247500 | 268083 | 247500 | THURSDAY | 14 | Approved |
| 2075578 | 418383 | Consumer loans | 7656.705 | 74610 | 65610 | 74610 | MONDAY | 14 | Approved |

From the above 2 pictures , I had shown that counting the number of null rows with the percentage of every other columns and resulted to be 39256 entries and 17 header rows. Also I have attached a excel file which was added as a **previous_application** and **Previous appli final dataset** (link) :

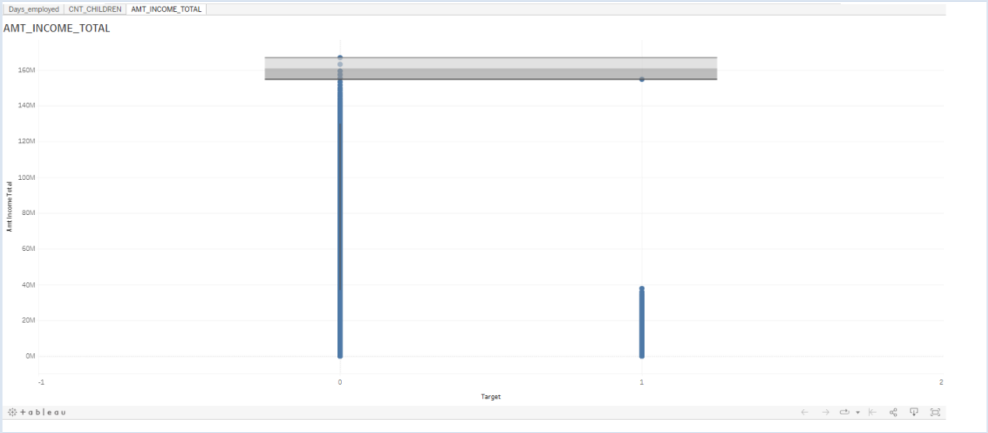
https://docs.google.com/spreadsheets/d/1gNqPVbP9FZ_OTC3O-jzlB18tFZelwVAl/edit?usp=drive_link

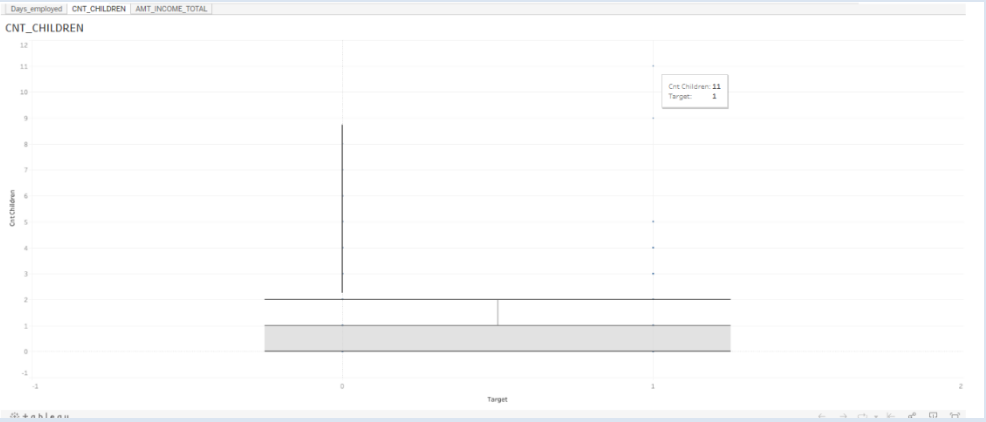
2. **Identify Outliers in the Dataset:** Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.**Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.**

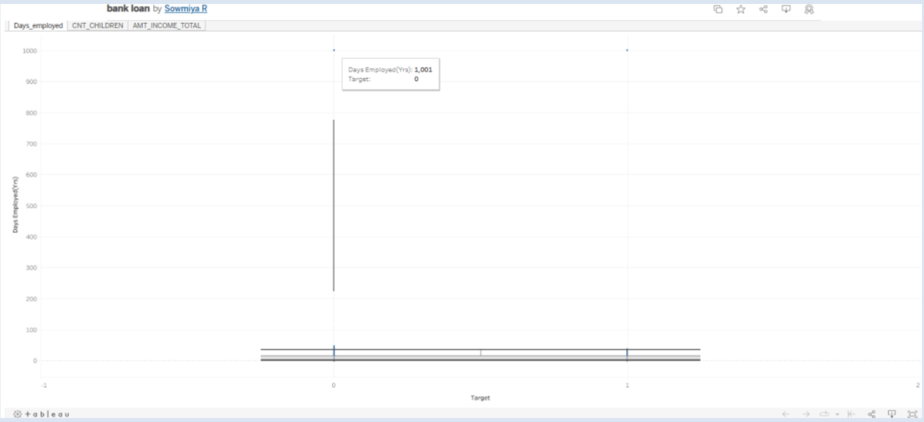
i) Outliers are values within a dataset that vary greatly from the others—they're either much larger, or significantly smaller. Outliers may indicate variabilities in a measurement, experimental errors, or a novelty.

Also then, I have found in this application dataset there are much more outliers which are calculated as the quartile ranges that is Q1, Q3, inner Quartile range, upper limit, lower limit using excel quartile functions itself.

Through then **AMT_INCOME_TOTAL, CNT_CHILDREN, DAYS_EMPLOYED(yrs)** has some number of outliers.







I have used Tableau for finding out the outliers box plot which are visually represented in the above 3 images .

https://public.tableau.com/app/profile/sowmiya.r1850/viz/bankloan_16905603122070/AMT_INCOME_TOTAL?publish=yes

From above link I have attached is that Tableau Public which I used to make a box plots data driven insights.

**Quartile 1**
87750

**Quartile 3**
248273

**Inter Quartile Range**
160523

**Upper Limit**
489057

**Lower Limit**
328534

This is the findings of outliers quartile ranges for amt_income_total which are shown above.

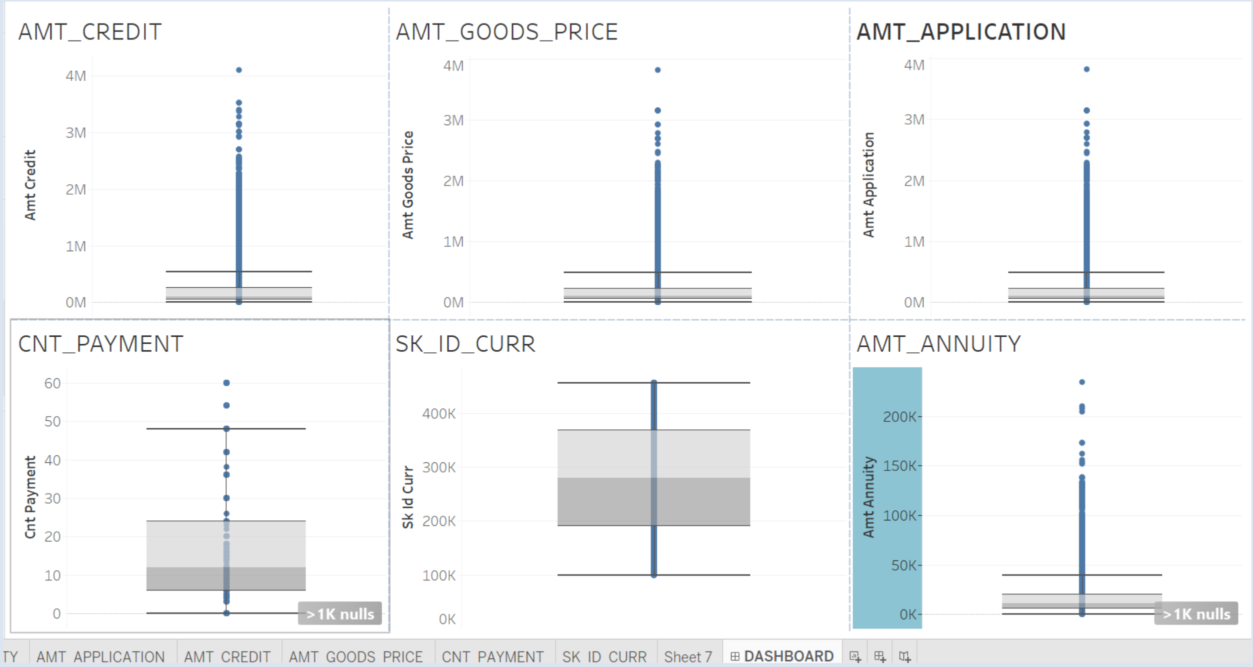
I have attached a excel file which is included , the calculations of statistics that is mean, median, mode and so on.

https://docs.google.com/spreadsheets/d/15E-puu5RNJL3Xcsi1iz3NF9XLfPaMH68/edit?usp=drive_link

ii) Also the Previous_application.csv dataset, I have found the outliers for some of the columns those are:

**AMT_ANNUITY AMT_APPLICATION AMT_CREDIT AMT_GOODS_PRICE CNT_PAYMENT SK_ID_CURR**

With this column names , I have found out the outliers with the techstack of Tableau Public .



From the above dashboard is then ouliers of amt_credit, amt_goods_price, amt_application, cnt_payment, sk_id_curr, amt_annuity and inferred that number of outliers with the relation between the previous application.

Here I have attached my excel link:

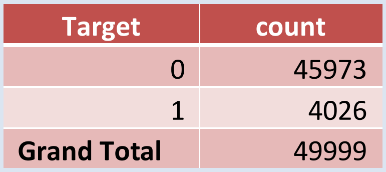
https://docs.google.com/spreadsheets/d/1gNqPVbP9FZ_OTC3O-jzlB18tFZelwVAl/edit?usp=drive_link

Tableau Public Link: PRE_APPLICATION | Tableau Public

3. **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.

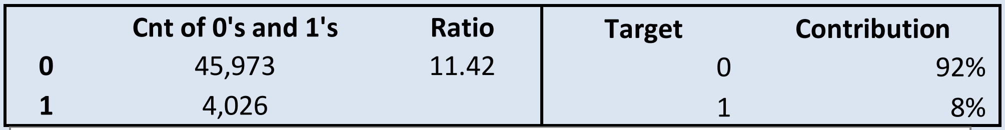
**Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.**

In this data imbalance tasks, I have inferred that accuracy of data in the representation of visual charts as counting off the applicants which are grouped as **0** and **1** .

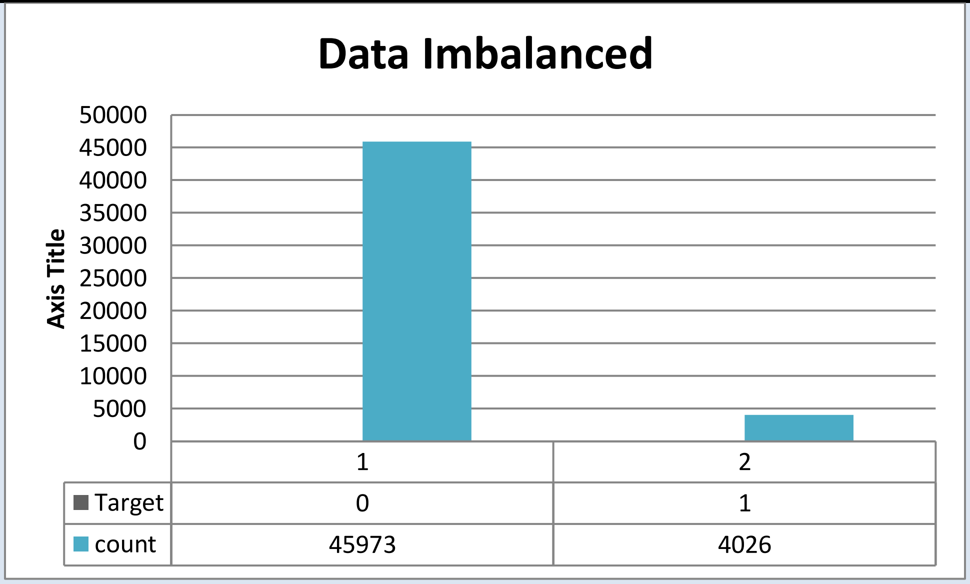


| Target | count |
|---|---|
| 0 | 45973 |
| 1 | 4026 |
| **Grand Total** | 49999 |

The above table says that the total number of applicants as Target (0 and 1) in the count which is total of 49999.



| | Cnt of 0's and 1's | Ratio | Target | Contribution |
|---|---|---|---|---|
| **0** | 45,973 | 11.42 | 0 | 92% |
| **1** | 4,026 | | 1 | 8% |



From the above chart represents the data imbalancement of target as 0 and 1 (applicants) and the count of those referred to the repayment of loan status.

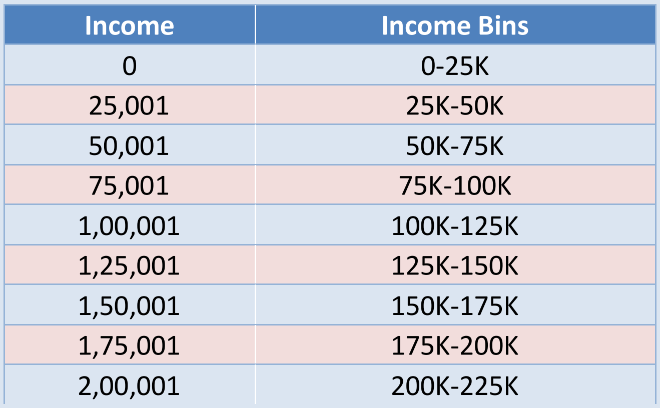
Here I have attached a excel link:

https://docs.google.com/spreadsheets/d/1mPqQiRTEdbybe5B11JT04IYUpEMBBTdx/edit?usp=drive_link

4. **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.**

In this univariate and bivariate analysis , I had found out the class intervals of income bins and credit bins .



| Income | Income Bins |
|---|---|
| 0 | 0-25K |
| 25,001 | 25K-50K |
| 50,001 | 50K-75K |
| 75,001 | 75K-100K |
| 1,00,001 | 100K-125K |
| 1,25,001 | 125K-150K |
| 1,50,001 | 150K-175K |
| 1,75,001 | 175K-200K |
| 2,00,001 | 200K-225K |

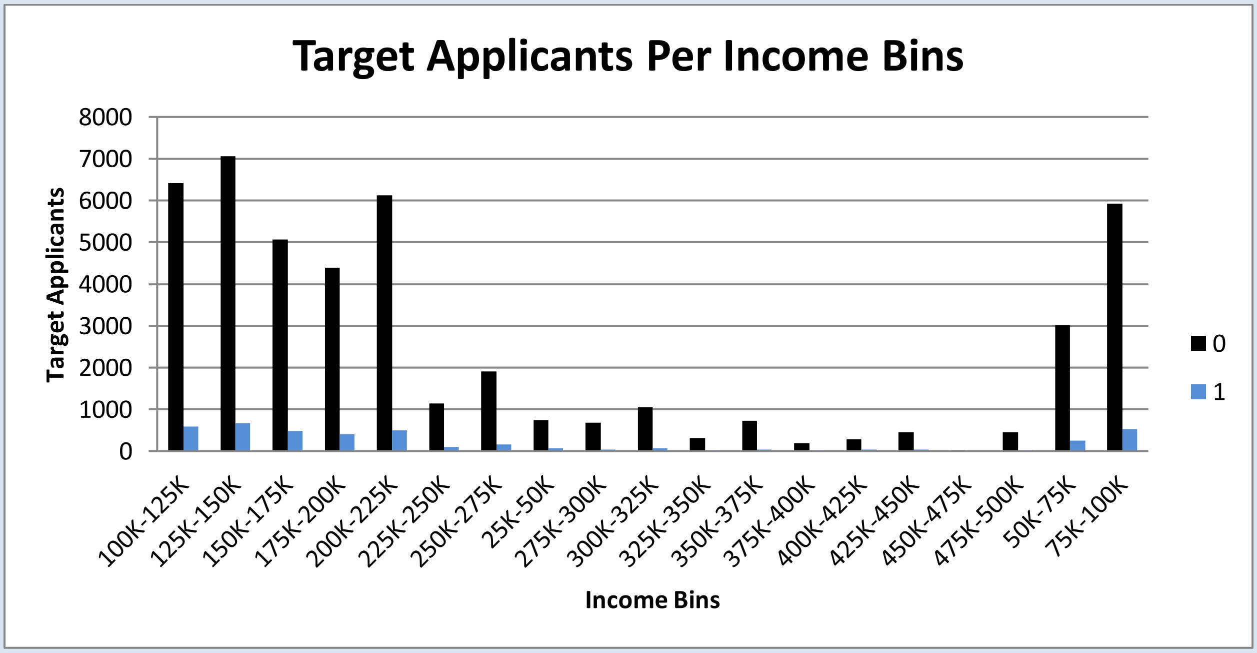
| | |
|---|---|
| 2,25,001 | 225K-250K |
| 2,50,001 | 250K-275K |
| 2,75,001 | 275K-300K |
| 3,00,001 | 300K-325K |
| 3,25,001 | 325K-350K |
| 3,50,001 | 350K-375K |
| 3,75,001 | 375K-400K |
| 4,00,001 | 400K-425K |
| 4,25,001 | 425K-450K |
| 4,50,001 | 450K-475K |
| 4,75,001 | 475K-500K |
| 5,00,001 | 5 Lacs and above |

| Credit | Credit Bins |
|---|---|
| 0 | 0 - 1.5 Lacs |
| 1,50,001 | 1.5 Lacs - 2 Lacs |
| 2,00,001 | 2 Lacs - 2.5 Lacs |
| 2,50,001 | 2.5 Lacs - 3 Lacs |
| 3,00,001 | 3 Lacs - 3.5 Lacs |
| 3,50,001 | 3.5 Lacs - 4 Lacs |
| 4,00,001 | 4 Lacs - 4.5 Lacs |
| 4,50,001 | 4.5 Lacs - 5 Lacs |
| 5,00,001 | 5 Lacs - 5.5 Lacs |
| 5,50,001 | 5.5 Lacs  - 6 Lacs |
| 6,00,001 | 6 Lacs  - 6.5 Lacs |
| 6,50,001 | 6.5 Lacs - 7 Lacs |
| 7,00,001 | 7 Lacs - 7.5 Lacs |
| 7,50,001 | 7.5 Lacs - 8 Lacs |
| 8,00,001 | 8 Lacs - 8.5 Lacs |
| 8,50,001 | 8.5 Lacs - 9 Lacs |
| 9,00,001 | 9 Lacs  and above |

These are the class intervals which are to found in the univariate , segmented univariate, bivariate analysis.

i) So for the analysis of **segmented univariate**, I have taken target as applicants , income_bins and amt_income_total then I converted this table as pivot table and then filtered through the class intervals of segmented univariate analysis and calculated with respect to the applicants who were getting the salary of above mentioned class intervals , through this I have analyzed a salary whoever can pay the loan or repay.

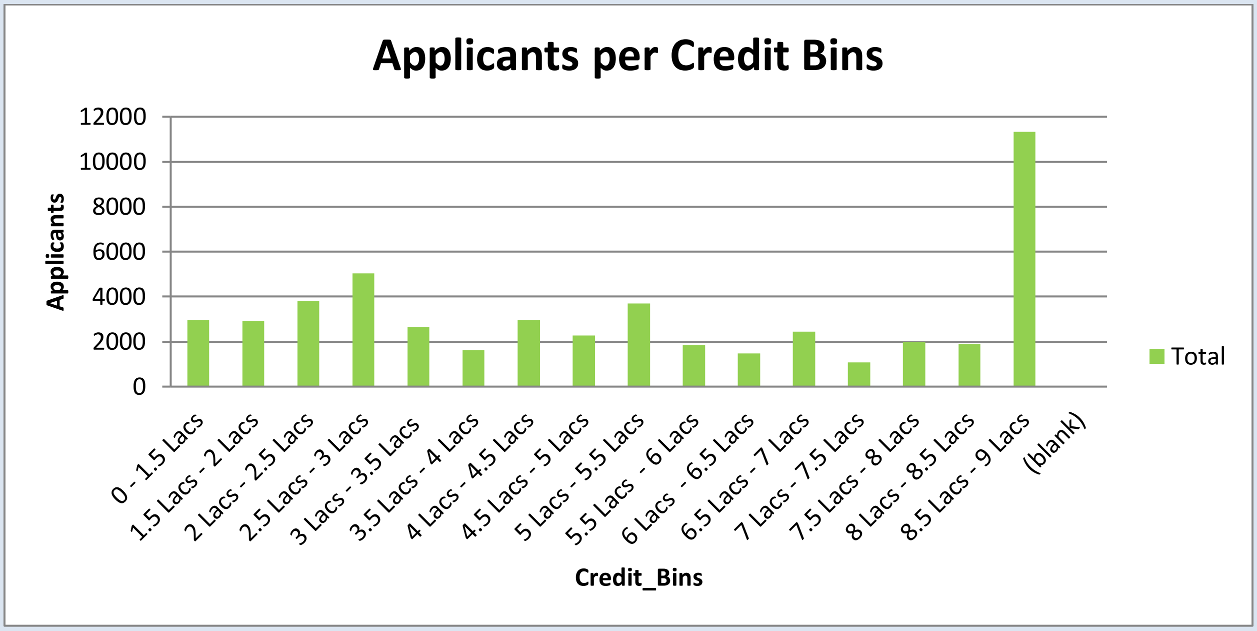
## SEGMENTED UNIVARIATE



ii) So for the analysis of **univariate** , I have taken Applicants, Amt_credit, Credit_bin in this. Next I grouped this as tables in the excel and performed a pivot table calculation for filtering the tables and then through the class intervals of credit bins and calculated with respect to the applicants who has the highest score of credit bins. So with this I have accomplished a univariate analysis.

| CREDIT_BINS | APPLICANTS |
|---|---|
| 0 - 1.5 Lacs | 2964 |
| 1.5 Lacs - 2 Lacs | 2936 |
| 2 Lacs - 2.5 Lacs | 3822 |
| 2.5 Lacs - 3 Lacs | 5027 |
| 3 Lacs - 3.5 Lacs | 2634 |
| 3.5 Lacs - 4 Lacs | 1622 |
| 4 Lacs - 4.5 Lacs | 2960 |
| 4.5 Lacs - 5 Lacs | 2268 |
| 5 Lacs - 5.5 Lacs | 3708 |
| 5.5 Lacs  - 6 Lacs | 1846 |
| 6 Lacs  - 6.5 Lacs | 1464 |
| 6.5 Lacs - 7 Lacs | 2445 |
| 7 Lacs - 7.5 Lacs | 1066 |
| 7.5 Lacs - 8 Lacs | 1996 |
| 8 Lacs - 8.5 Lacs | 1911 |
| 8.5 Lacs - 9 Lacs | 11330 |
| (blank) | |
| **Grand Total** | **49999** |

# UNIVARIATE ANALYSIS

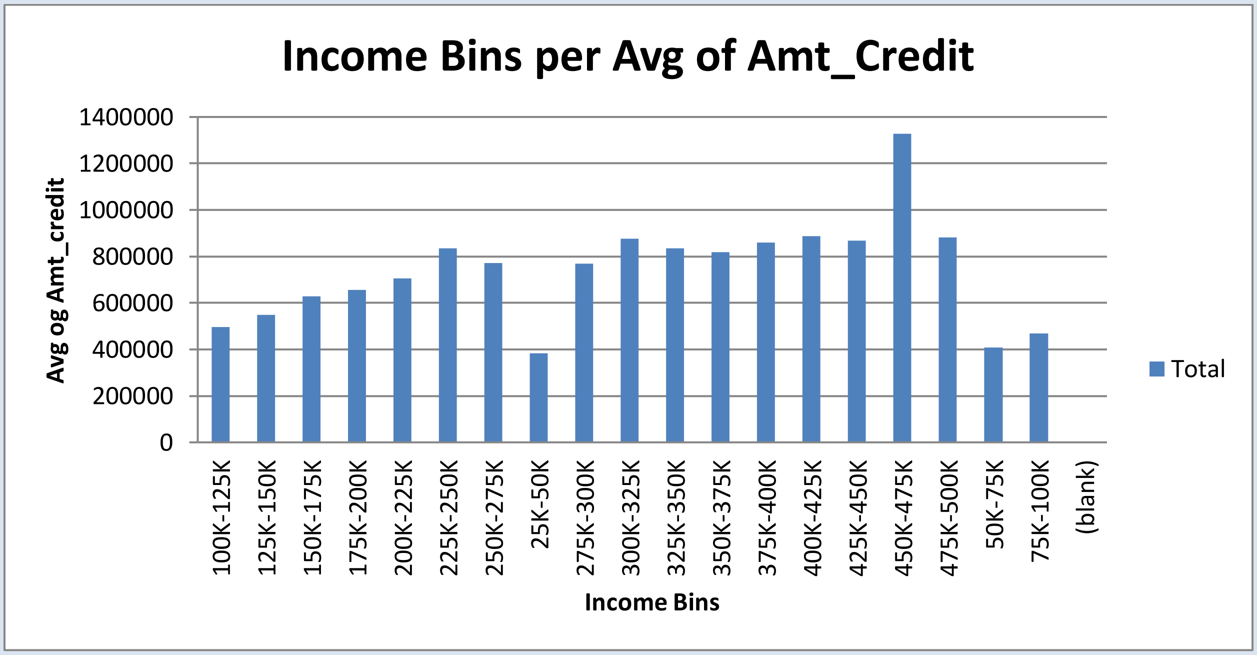


From the above chart , I analyzed that the credit bins is the score of applicants pay the loan or not.

iii) For the **Bivariate analysis** , I have found with both the intervals of credit bins and amt_income. Then I have created a table and the pivoted in the excel for filtering those class intervals of credit bins with respect to the average of amount credits. So with this I have inferred with a column chart and pivot table filtering for the bivariate analysis.

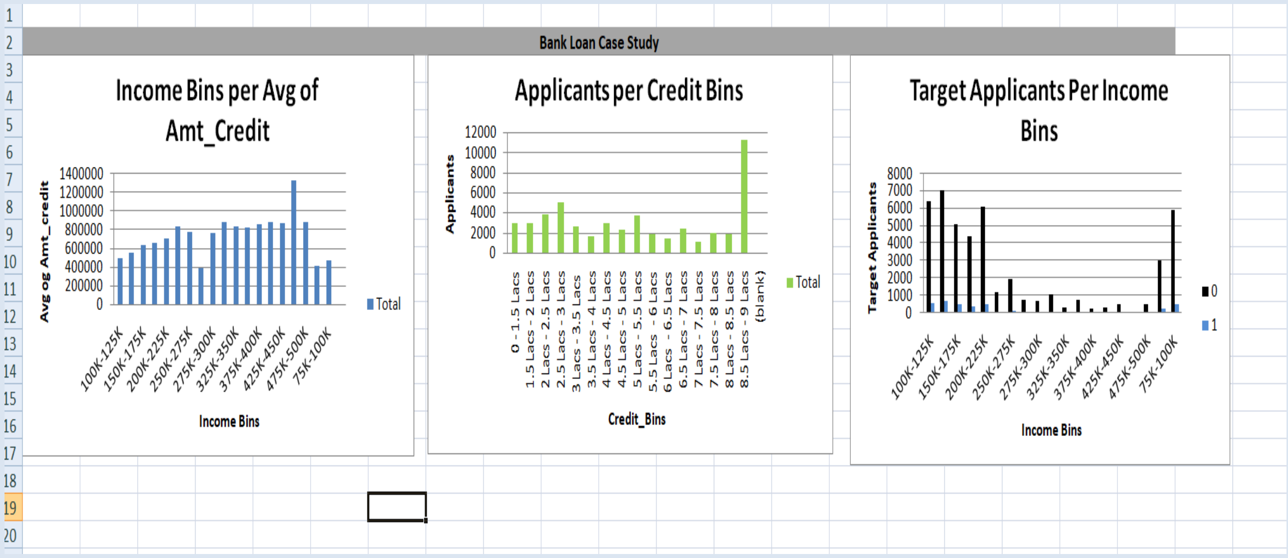
| INCOME_BIN | Average of AMT_CREDIT |
|---|---|
| 100K-125K | 496518.7668 |
| 125K-150K | 548318.112 |
| 150K-175K | 629686.4004 |
| 175K-200K | 654776.5956 |
| 200K-225K | 704078.6153 |
| 225K-250K | 834911.416 |
| 250K-275K | 772695.1337 |
| 25K-50K | 383814.5844 |
| 275K-300K | 768453.375 |
| 300K-325K | 876861.8617 |
| 325K-350K | 836116.3929 |
| 350K-375K | 817785.4932 |
| 375K-400K | 858555.45 |
| 400K-425K | 886368.5526 |
| 425K-450K | 868797.4355 |
| 450K-475K | 1327648.5 |
| 475K-500K | 882098.1486 |
| 50K-75K | 408129.1212 |
| 75K-100K | 469470.5171 |
| (blank) | |
| **Grand Total** | **603931.5578** |

# BIVARIATE ANALYSIS



From the above insights, I have inferred as the income bins with respect to the average of amt_credits and also I have attached a excel files that included everything.

Excel file: https://docs.google.com/spreadsheets/d/1YQXLFoFK-jtF4oVWNMxCQ-RK49-6bTTl/edit?usp=drive_link



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

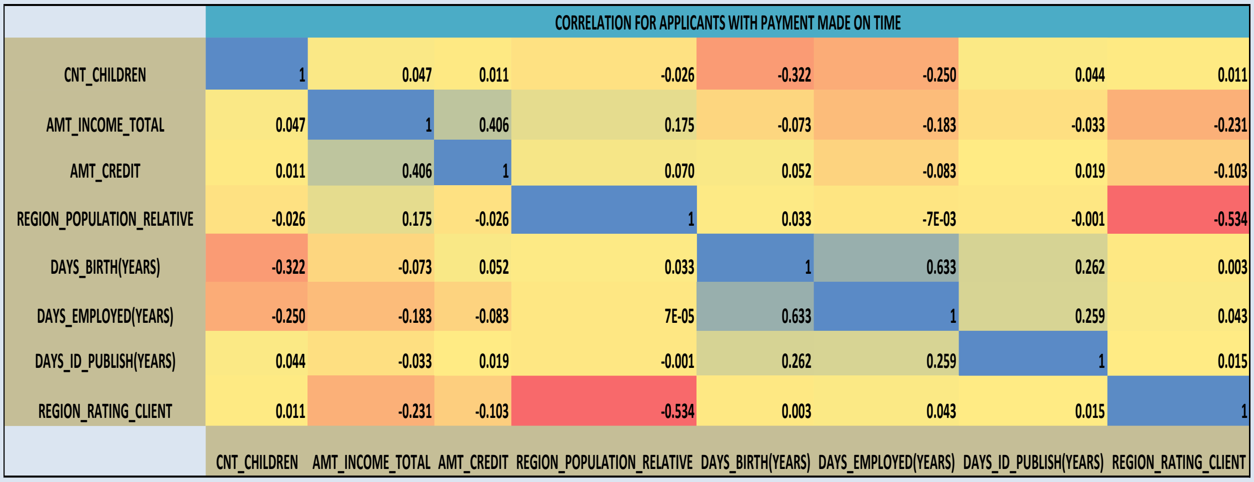
Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default. **Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other**

**cases) and identify the top correlations for each segmented data using Excel functions.**

For this task , I have been finding out the correlation between different scenarios that the client is facing any issue or difficulties with all other cases . So then I will be finding out different scenarios of each and every other to correlations in the excel functions.

i) So I have taken up of **target 0** as one correlation. Then I included these columns alone from the final dataset and performed correlation between them.

TARGET, CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, REGION_POPULATION_RELATIVE, DAYS_BIRTH(yrs), DAYS_EMPLOYED(YRS), DAYS_ID_PUBLISH(YRS), REGION_RATING_CLIENT



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | CORRELATION FOR APPLICANTS WITH PAYMENT MADE ON TIME | | | | | | | |
| CNT_CHILDREN | 1 | 0.047 | 0.011 | -0.026 | -0.322 | -0.250 | 0.044 | 0.011 |
| AMT_INCOME_TOTAL | 0.047 | 1 | 0.406 | 0.175 | -0.073 | -0.183 | -0.033 | -0.231 |
| AMT_CREDIT | 0.011 | 0.406 | 1 | 0.070 | 0.052 | -0.083 | 0.019 | -0.103 |
| REGION_POPULATION_RELATIVE | -0.026 | 0.175 | -0.026 | 1 | 0.033 | -7E-03 | -0.001 | -0.534 |
| DAYS_BIRTH(YEARS) | -0.322 | -0.073 | 0.052 | 0.033 | 1 | 0.633 | 0.262 | 0.003 |
| DAYS_EMPLOYED(YEARS) | -0.250 | -0.183 | -0.083 | 7E-05 | 0.633 | 1 | 0.259 | 0.043 |
| DAYS_ID_PUBLISH(YEARS) | 0.044 | -0.033 | 0.019 | -0.001 | 0.262 | 0.259 | 1 | 0.015 |
| REGION_RATING_CLIENT | 0.011 | -0.231 | -0.103 | -0.534 | 0.003 | 0.043 | 0.015 | 1 |
| | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | REGION_POPULATION_RELATIVE | DAYS_BIRTH(YEARS) | DAYS_EMPLOYED(YEARS) | DAYS_ID_PUBLISH(YEARS) | REGION_RATING_CLIENT |

So from the above selected columns I have correlated and inferred to create a heat map for this **Target 0.** In the above heat map itself says the answer that is the correlation of each and every different scenarios.

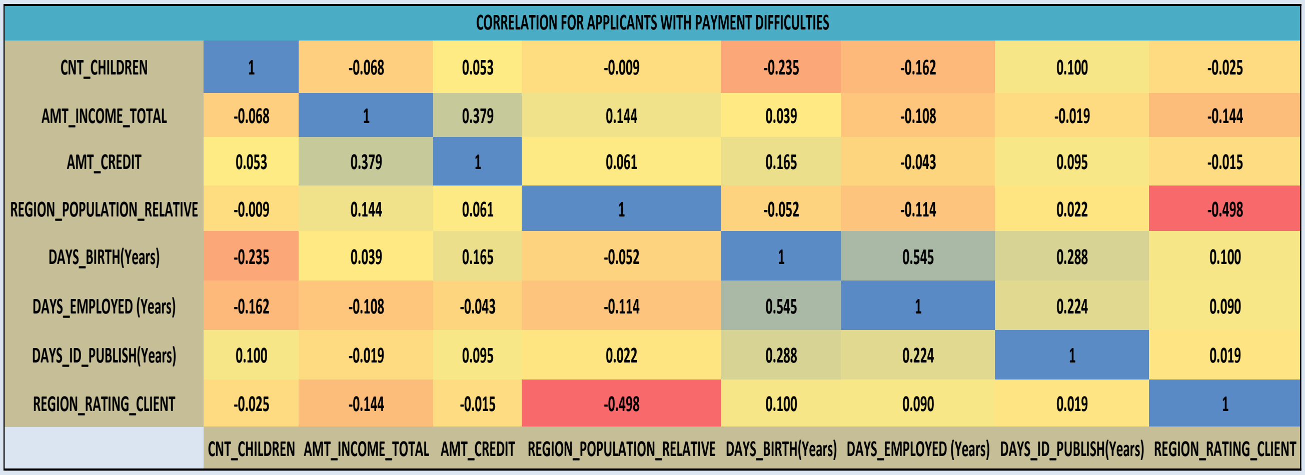
ii) For the **Target 1** correlation, the same as above the calculations as including the columns of particular needed headed rows and performing calculation of correlation of different scenarios.

TARGET, CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT,

REGION_POPULATION_RELATIVE, DAYS_BIRTH(yrs), DAYS_EMPLOYED(YRS),

DAYS_ID_PUBLISH(YRS), REGION_RATING_CLIENT

With this columns I have inferred a **correlation of target 1**in the kind of heat maps.



CORRELATION FOR APPLICANTS WITH PAYMENT DIFFICULTIES

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| CNT_CHILDREN | 1 | -0.068 | 0.053 | -0.009 | -0.235 | -0.162 | 0.100 | -0.025 |
| AMT_INCOME_TOTAL | -0.068 | 1 | 0.379 | 0.144 | 0.039 | -0.108 | -0.019 | -0.144 |
| AMT_CREDIT | 0.053 | 0.379 | 1 | 0.061 | 0.165 | -0.043 | 0.095 | -0.015 |
| REGION_POPULATION_RELATIVE | -0.009 | 0.144 | 0.061 | 1 | -0.052 | -0.114 | 0.022 | -0.498 |
| DAYS_BIRTH(Years) | -0.235 | 0.039 | 0.165 | -0.052 | 1 | 0.545 | 0.288 | 0.100 |
| DAYS_EMPLOYED (Years) | -0.162 | -0.108 | -0.043 | -0.114 | 0.545 | 1 | 0.224 | 0.090 |
| DAYS_ID_PUBLISH(Years) | 0.100 | -0.019 | 0.095 | 0.022 | 0.288 | 0.224 | 1 | 0.019 |
| REGION_RATING_CLIENT | -0.025 | -0.144 | -0.015 | -0.498 | 0.100 | 0.090 | 0.019 | 1 |
| | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | REGION_POPULATION_RELATIVE | DAYS_BIRTH(Years) | DAYS_EMPLOYED (Years) | DAYS_ID_PUBLISH(Years) | REGION_RATING_CLIENT |

From the above heat map , I have analyzed that the correlation between the different scenarios of **correlation for target 1** and above map itself calculated for this tasks for correlation of each and every other loan scenarios.

Excel link:
https://docs.google.com/spreadsheets/d/10_SMG1RDWcuLIeCy2KHP3WP3KVy5xKXz/edit?usp=drive_link

## RESULTS

- In this project , I have inferred that the given datasets of application_data and previous_data is been used for the whole tasks using the excel functions to create a charts, tables, pivot tables and so on.

- I have applied EDA analysis technique and understand the excel functions of calculating percentage of null rows in the real business scenarios.
- I learned risk analysis in this case study by making visual insights in banking and financial services, Then this project is for me challenging for the whole analyzed datasets by implementation of correlations, data imbalance , outliers factors in this.

## DRIVE LINK

https://drive.google.com/drive/folders/1BPXUmps2cb8Q4bD1HmBHuwsYSGuSYyl5?usp=drive_link