# Bank loan Case Study

## Project Description:

As an employee of a consumer finance company, my task is to analyse patterns in the data to ensure that the loans are provided to the applicants who are capable of repaying them, instead of rejecting them based on insufficient or non-existent credit history.

This can be achieved by conducting an Exploratory Data Analysis (EDA) on the available data. EDA involves looking at the data from various perspectives, identifying patterns, and gaining insights into the data. This will help the company reduce the risk of default and provide loans to those who are more likely to repay them.

## Approach:

The first step is to clean the available data by removing any irrelevant columns and deleting any outliers. This ensures that the analysis is based on relevant and accurate data.

Conducting EDA is crucial to identify patterns and gain insights into the data. this involves analysing the data from various perspectives, such as using pivot tables, charts, outlier analysis to identify the factors that influence loan repayment. Based on the insights gained from the EDA the company can provide loans to applicants who are more likely to repay them, reduce the risk of default, and optimize its loan approval process.

## Tech Stack Used:

- Excel
- MySQL
- Microsoft Word
- Google Chrome

# Overall Approach of the Analysis

The problem statement of this case study is to use EDA to analyse loan application data and identify patterns that indicate if a client has difficulty paying their instalments. The aim is to minimize the risk of losing money while lending to customers by ensuring that the loans are provided to the applicants who are capable of repaying them, instead of rejecting them based on insufficient or non-existent credit history.

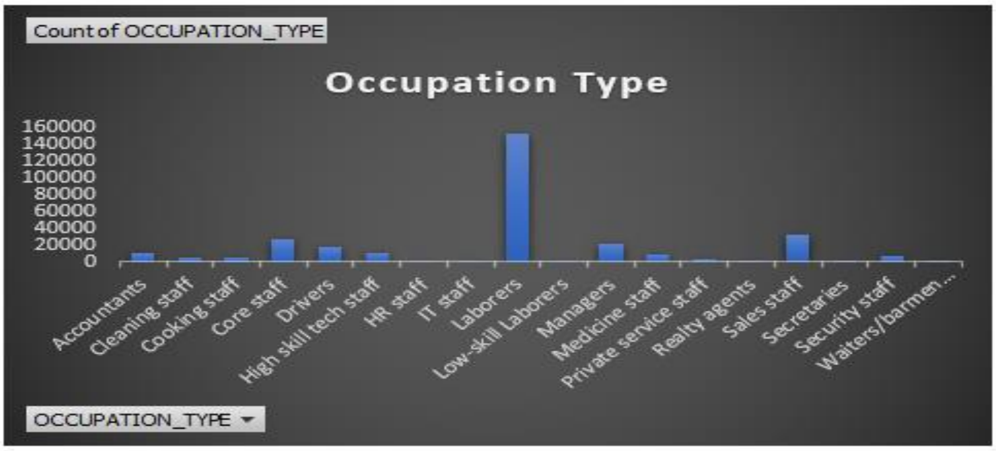
The analysis approach involves cleaning the data, identifying and removing outliers, and performing analysis using pivot tables and charts. The objective is to identify the driving factors behind loan default and utilize this knowledge for portfolio and risk assessment.

Additionally, independent research on risk analytics has been done to understand the types of variables and their significance.

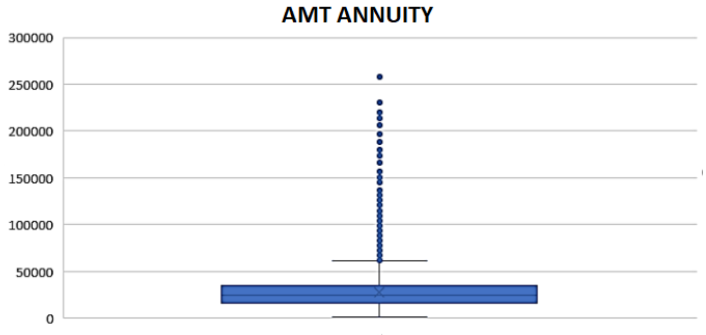
The data sets used in this analysis include **'application_data.csv'**, **'previous_application.csv'**, and **'columns_descrption.csv'**. The applications data contains all the information of the client at the time of application, the data is about whether a client has payment difficulties, while the previous applications contains information about the client's previous loan data, including whether the previous application had been approved, cancelled, refused, or an unused offer.

## Identifying the missing data and use appropriate method to deal with it.

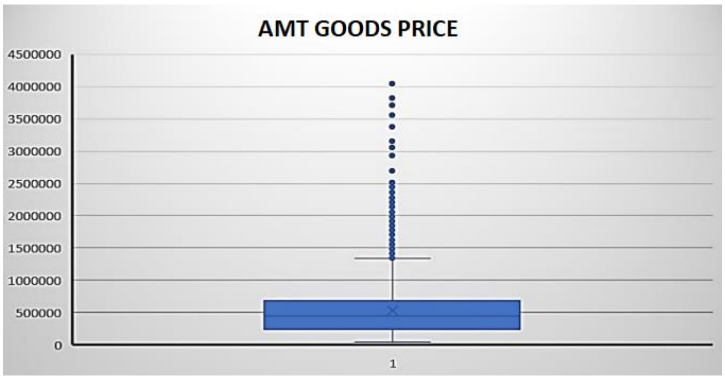
- Replacing the blank values in occupation type with highest occurring categorical variable



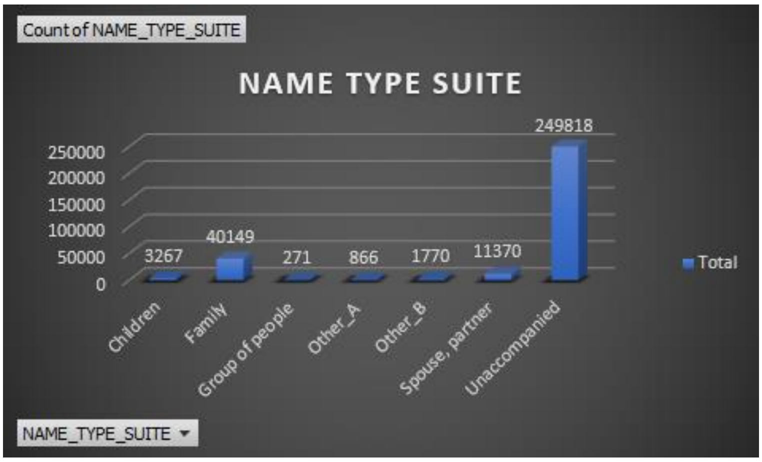
- Replacing the amt annuity blank cells with the median value of the column as there exists outliers in the amt annuity column



- Replacing the amt goods price blank cells with the median value of the column as there exists outliers in the amt goods price



- Replacing the black cells in the name type suite column with the highest occurring categorical variable

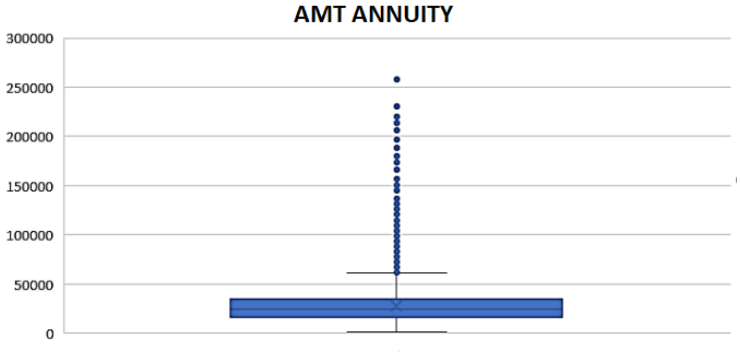


## Insights:

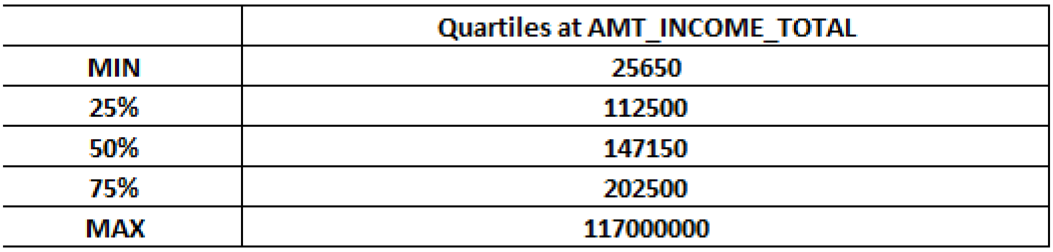
- Highest occurring categorical variable in **Occupation_type** is 'Laborers'
- Median value of AMT_ANNUITY is **24903**
- Median value of AMT_GOODS_PRICE is **450000**
- Highest occurring categorical variable in Name_type_suite is **'Unaccompained'**
- Highest occurring categorical variable in Organisation_type is '**Business Entity 3**'

# Outlier

- First Outlier is in AMT ANNUITY which is greater than 250000 this outlier is replaced with 24903 median of AMT_ANNUITY

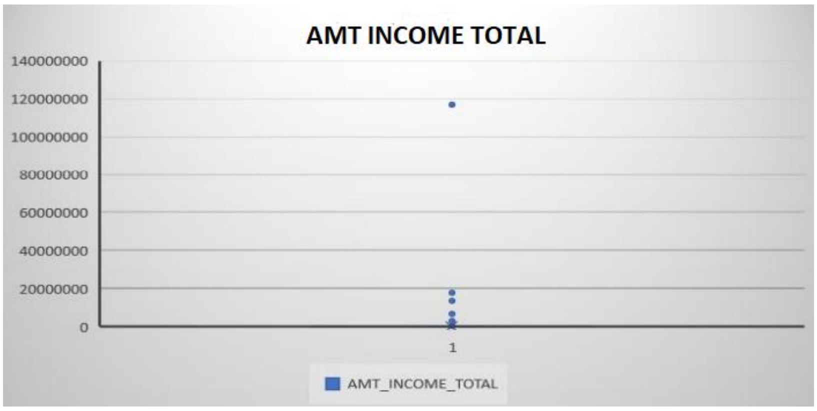


- In the application dataset we can observe that there is a huge difference between the 25%, 50%, and 75% quartile and this is due to the presence of the outliers
- But since the amount of the total income varies from person to person we will not remove the outliers

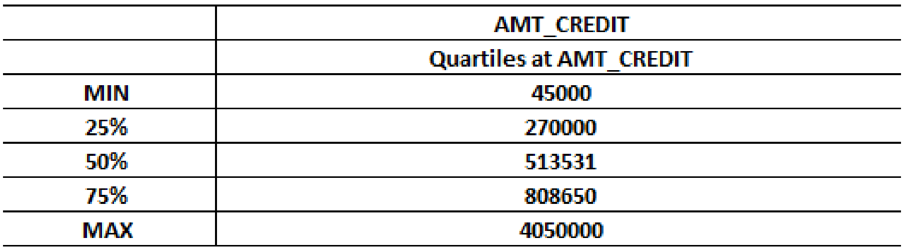


| | Quartiles at AMT_INCOME_TOTAL |
|---|---|
| MIN | 25650 |
| 25% | 112500 |
| 50% | 147150 |
| 75% | 202500 |
| MAX | 117000000 |

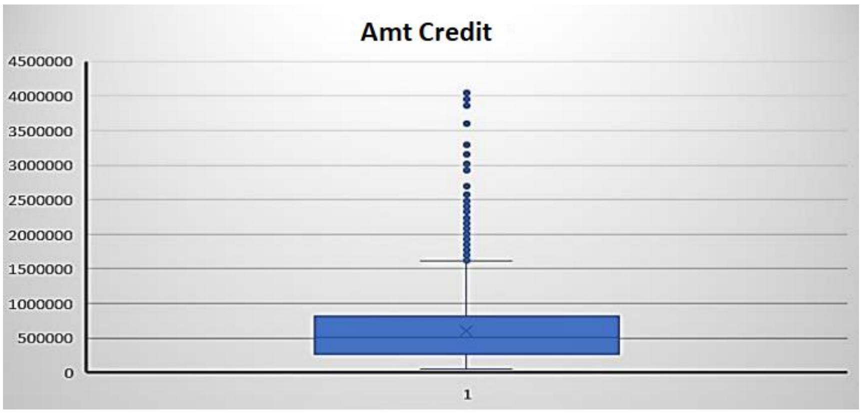
- Outliers at extreme points i.e max 1.700x10^8



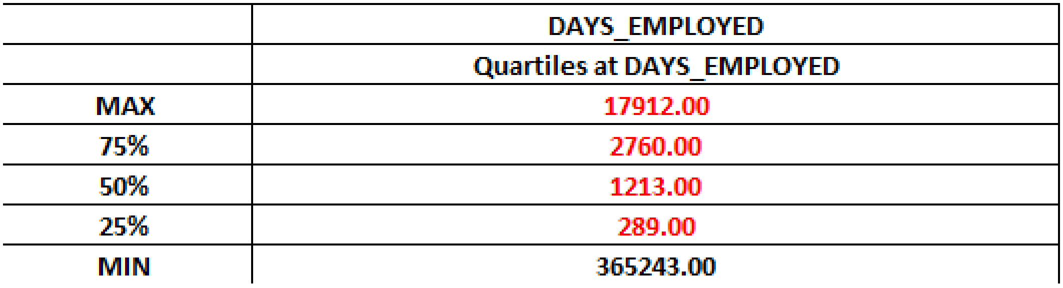
- From the Amt credit column data it is clear that outliers lie in the 98% and near max side of the box plot Also there is a significant difference between the 75% quartile and the max value and this is due the presence of the outliers
- But since the amount of credit varies from person to person we will not remove the outliers



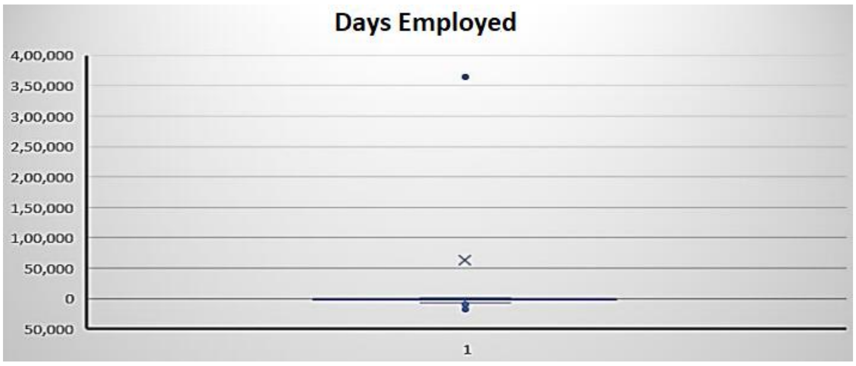
| | AMT_CREDIT |
|---|---|
| | Quartiles at AMT_CREDIT |
| MIN | 45000 |
| 25% | 270000 |
| 50% | 513531 |
| 75% | 808650 |
| MAX | 4050000 |



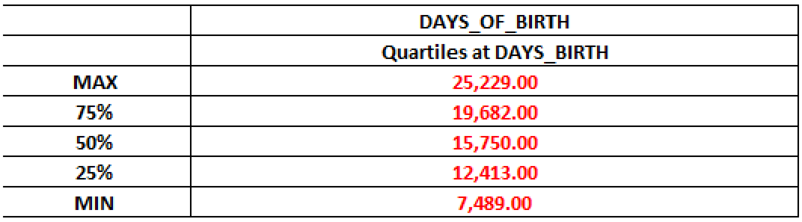
- From the days employed there exists only one outlier i.e + or -365243 we will replace it with median 1213.



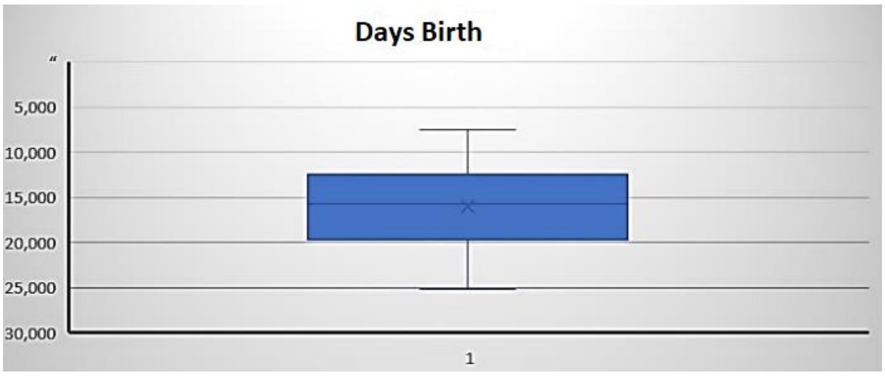
| | DAYS_EMPLOYED |
|---|---|
| | Quartiles at DAYS_EMPLOYED |
| MAX | 17912.00 |
| 75% | 2760.00 |
| 50% | 1213.00 |
| 25% | 289.00 |
| MIN | 365243.00 |



- The data from the days_birth is well distributed



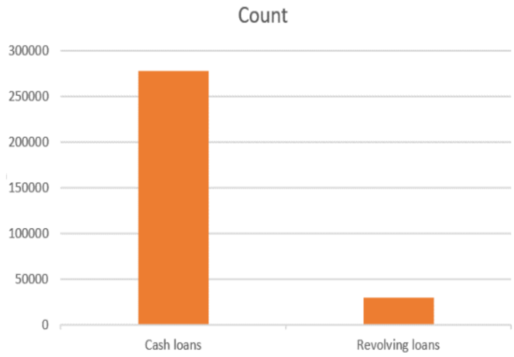
| | DAYS_OF_BIRTH |
|---|---|
| | Quartiles at DAYS_BIRTH |
| MAX | 25,229.00 |
| 75% | 19,682.00 |
| 50% | 15,750.00 |
| 25% | 12,413.00 |
| MIN | 7,489.00 |

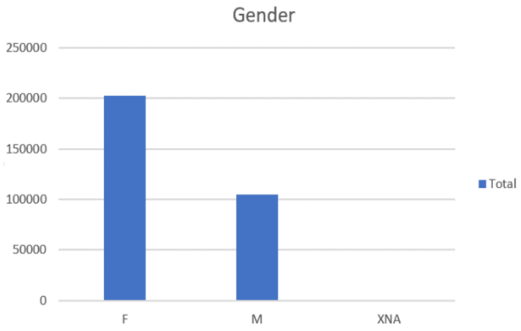


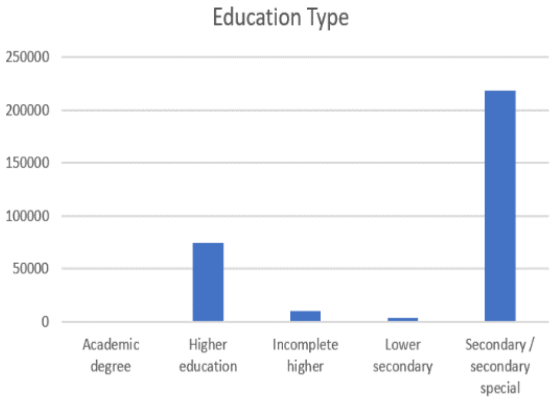
## Data Imbalance

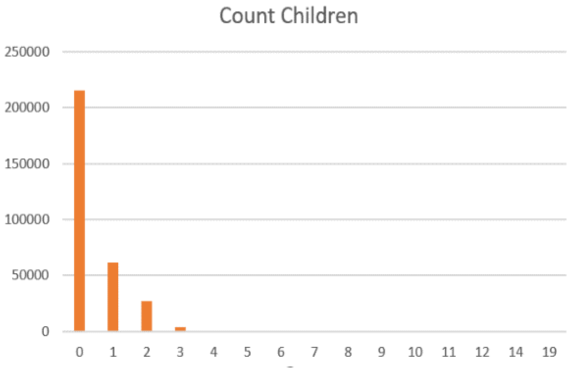
- Data imbalance occurs when data is disseminated in anunequal manner. I plotted data imbalance using Pivot charts.

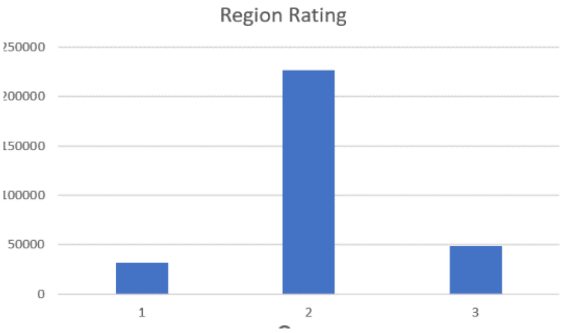
### Name Contract Type:

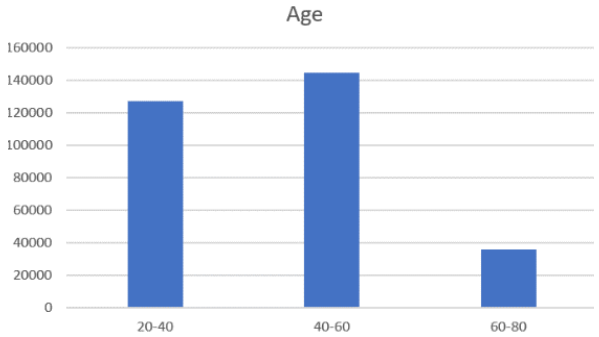




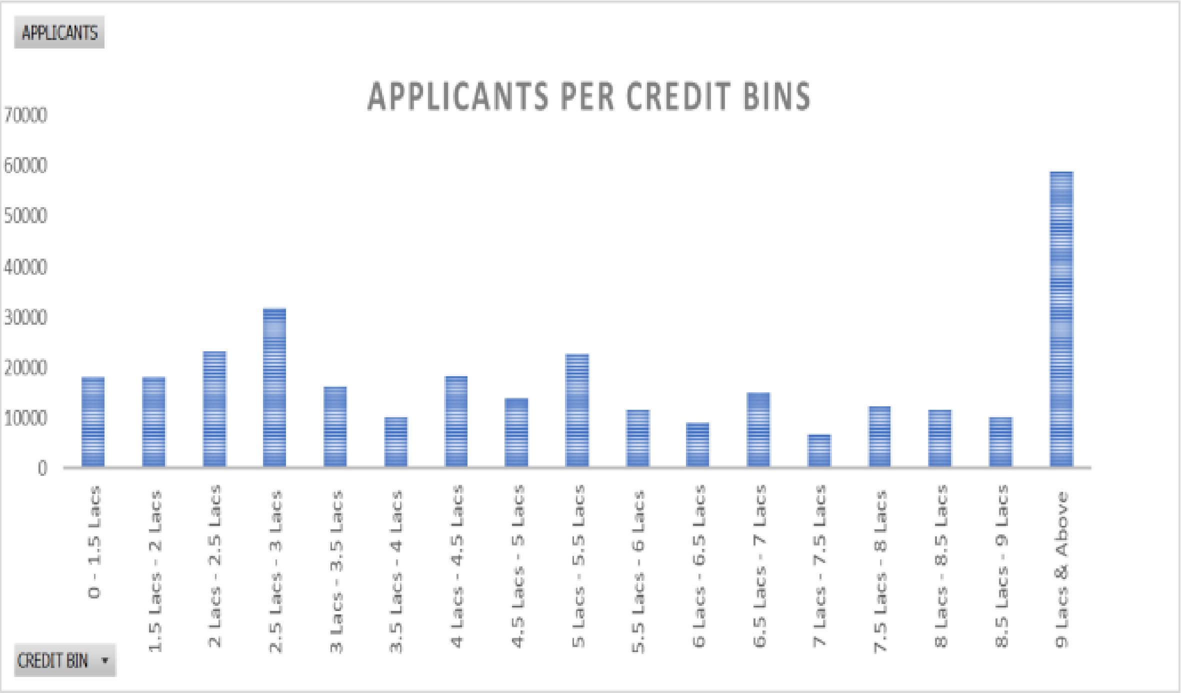






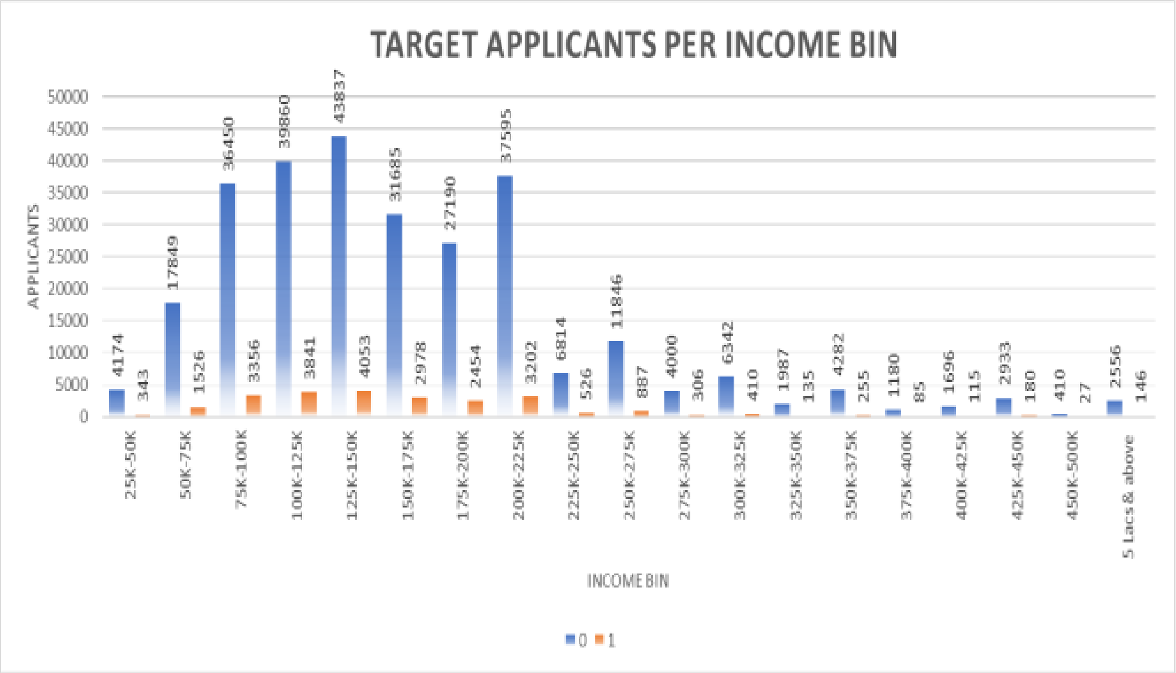


## Univariate Analysis



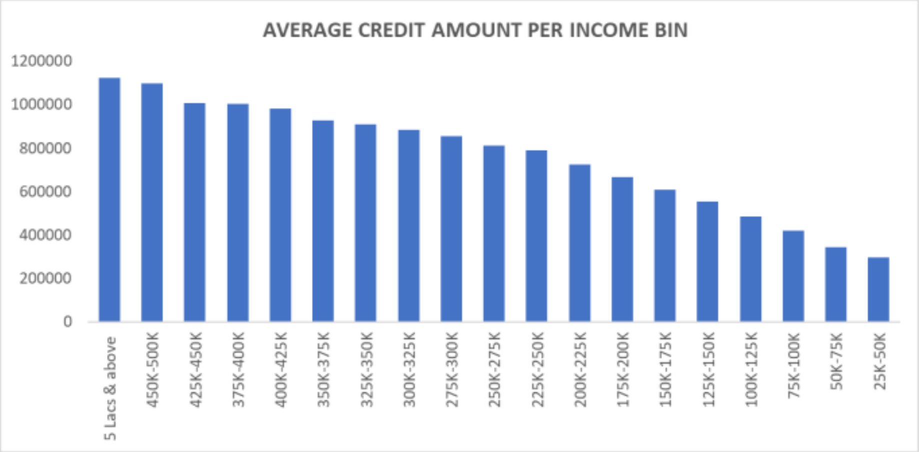
- Univariate Analysis focuses on examining data that involves a single variable. Its purpose is to describe the data and identify patterns without exploring causes or relationships. The provided graph exemplifies univariate analysis by displaying the distribution of applicants based on their loan amounts (AMT_CREDIT) grouped into different credit ranges. The graph reveals that a significant number of applicants were granted loans in the credit range of 9 Lacs and above.

# Univariate Segmented Analysis



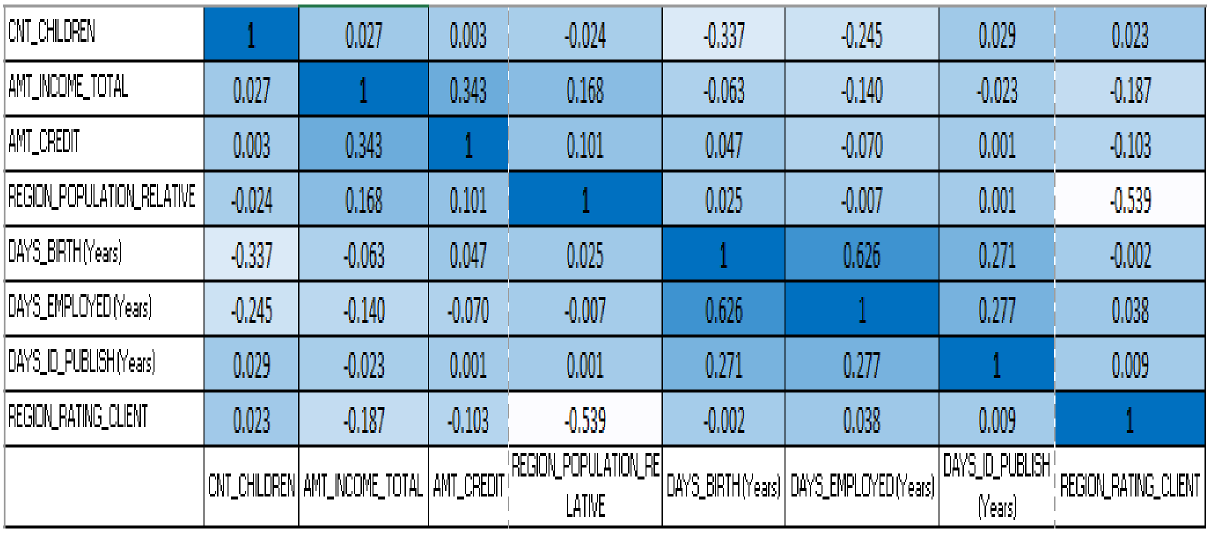
- Univariate Analysis involves analysing data with a single variable, while segmented analysis focuses on examining subsets within that variable. The provided graph demonstrates segmented univariate analysis by depicting the count of applicants (0 & 1) based on their total income (AMT_TOTAL_INCOME) grouped into different income ranges. The graph highlights that there are very few applicants with a total income of 50 Lacs and above who are facing payment difficulties, which may explain their challenges in making payments. Additionally, the majority of applicants (0,1) fall within the income range of 1.25 Lacs to 1.5 Lacs, but there are instances of applicants within this range who are experiencing payment difficulties.

# Bivariate Analysis



- Bivariate Analysis involves analysing data that consists of two variables and focuses on understanding their relationship and potential causes. The graph provided illustrates a bivariate analysis between AMT_CREDIT and AMT_TOTAL_INCOME. The graph demonstrates that applicants with higher income levels were offered higher loan amounts, indicating a positive correlation between these two variables. This suggests that there is a proportional relationship between income and loan amount, where higher income applicants tend to receive larger loans.

## Corelation for applicants with payment made on time



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| CNT_CHILDREN | 1 | 0.027 | 0.003 | -0.024 | -0.337 | -0.245 | 0.029 | 0.023 |
| AMT_INCOME_TOTAL | 0.027 | 1 | 0.343 | 0.168 | -0.063 | -0.140 | -0.023 | -0.187 |
| AMT_CREDIT | 0.003 | 0.343 | 1 | 0.101 | 0.047 | -0.070 | 0.001 | -0.103 |
| REGION_POPULATION_RELATIVE | -0.024 | 0.168 | 0.101 | 1 | 0.025 | -0.007 | 0.001 | -0.539 |
| DAYS_BIRTH (Years) | -0.337 | -0.063 | 0.047 | 0.025 | 1 | 0.626 | 0.271 | -0.002 |
| DAYS_EMPLOYED (Years) | -0.245 | -0.140 | -0.070 | -0.007 | 0.626 | 1 | 0.277 | 0.038 |
| DAYS_ID_PUBLISH (Years) | 0.029 | -0.023 | 0.001 | 0.001 | 0.271 | 0.277 | 1 | 0.009 |
| REGION_RATING_CLIENT | 0.023 | -0.187 | -0.103 | -0.539 | -0.002 | 0.038 | 0.009 | 1 |
| | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | REGION_POPULATION_RELATIVE | DAYS_BIRTH (Years) | DAYS_EMPLOYED (Years) | DAYS_ID_PUBLISH (Years) | REGION_RATING_CLIENT |

The analysis focuses on correlations for applicants who have made payments on time, represented by the target (0). The accompanying heat map visually represents these correlations between different variables. The colour scheme used ranges from blue to white, with blue indicating stronger correlations and white representing weaker correlations.

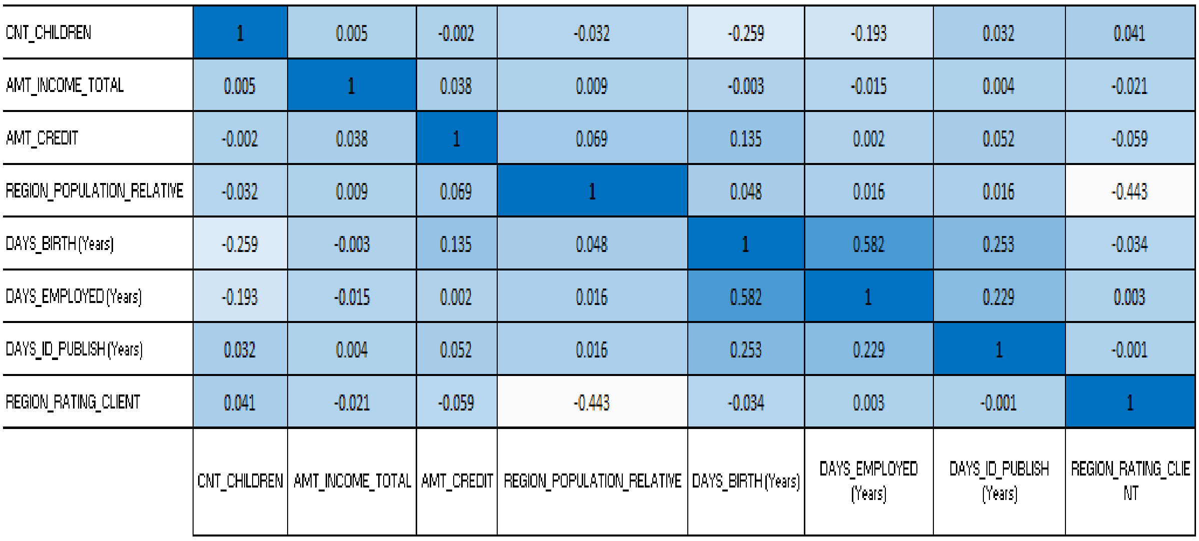
The key correlations observed in the analysis are as follows:

1. The variable "AMT_TOTAL_INCOME" shows a significant correlation with "AMT_CREDIT."

2. "DAYS_EMPLOYED" correlates closely with "DAYS_BIRTH."

3. "REGION_POPULATION_RELATIVE" exhibits a notable correlation with "AMT_INCOME_TOTAL."

# Corelation for applicants with payment difficulties



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| CNT_CHILDREN | 1 | 0.005 | -0.002 | -0.032 | -0.259 | -0.193 | 0.032 | 0.041 |
| AMT_INCOME_TOTAL | 0.005 | 1 | 0.038 | 0.009 | -0.003 | -0.015 | 0.004 | -0.021 |
| AMT_CREDIT | -0.002 | 0.038 | 1 | 0.069 | 0.135 | 0.002 | 0.052 | -0.059 |
| REGION_POPULATION_RELATIVE | -0.032 | 0.009 | 0.069 | 1 | 0.048 | 0.016 | 0.016 | -0.443 |
| DAYS_BIRTH (Years) | -0.259 | -0.003 | 0.135 | 0.048 | 1 | 0.582 | 0.253 | -0.034 |
| DAYS_EMPLOYED (Years) | -0.193 | -0.015 | 0.002 | 0.016 | 0.582 | 1 | 0.229 | 0.003 |
| DAYS_ID_PUBLISH (Years) | 0.032 | 0.004 | 0.052 | 0.016 | 0.253 | 0.229 | 1 | -0.001 |
| REGION_RATING_CLIENT | 0.041 | -0.021 | -0.059 | -0.443 | -0.034 | 0.003 | -0.001 | 1 |
| | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | REGION_POPULATION_RELATIVE | DAYS_BIRTH (Years) | DAYS_EMPLOYED (Years) | DAYS_ID_PUBLISH (Years) | REGION_RATING_CLIENT |

The analysis focuses on correlations among different variables for applicants who do not encounter payment difficulties, indicated by the target (0). The corresponding heat map visually represents these correlations using a colour scheme ranging from blue to white. blue hues signify stronger correlations, while white hues represent weaker correlations.

Noteworthy correlations observed in the analysis include:

1. A significant correlation between "AMT_TOTAL_INCOME" and "AMT_CREDIT."
2. A close correlation between "DAYS_EMPLOYED" and "DAYS_BIRTH."
3. An observable correlation between "REGION_POPULATION_RELATIVE" and "AMT_INCOME_TOTAL."

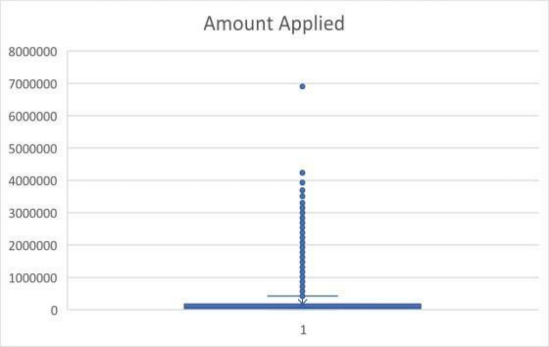
# Previous Applications

## Data Cleaning:

- Column removal- I used the COUNTBLANK function to determine the number of blanks in a column, and if it exceeds 5% then, I eliminated it.
- I removed following columns as they were no use to the analysis.
    - HOUR_APPR_PROCESS_START
    - WEEKDAY_APPR_PROCESS_START_PREV
    - FLAG_LAST_APPL_PER_CONTRACT
    - NFLAG_LAST_APPL_IN_DAY
    - SK_ID_CURR
    - WEEKDAY_APPR_PROCESS_START
- Removed the rows with the values 'XNA' & 'XAP' from the column: NAME_TYPE_SUITE

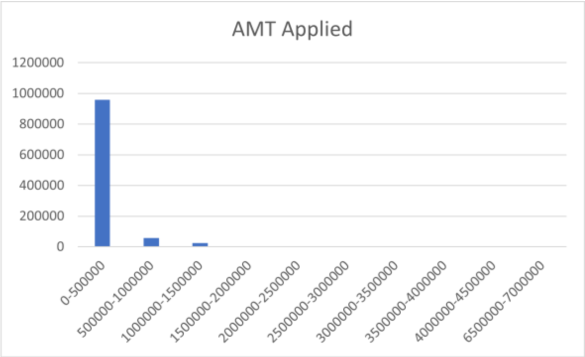
## Finding Outliers:

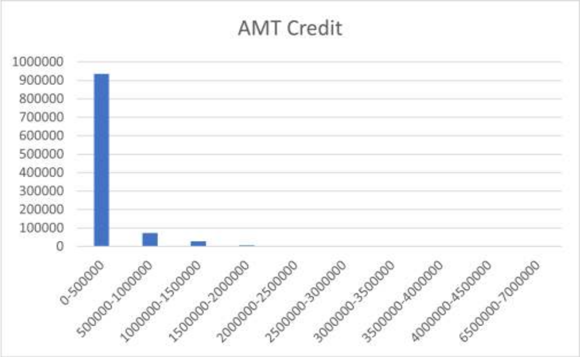
- Replacing the amt annuity blank values with its median 21340

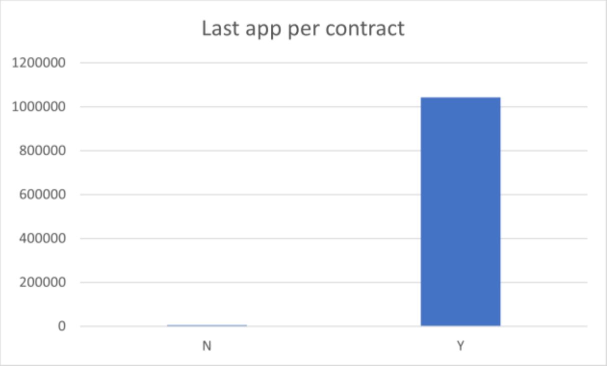


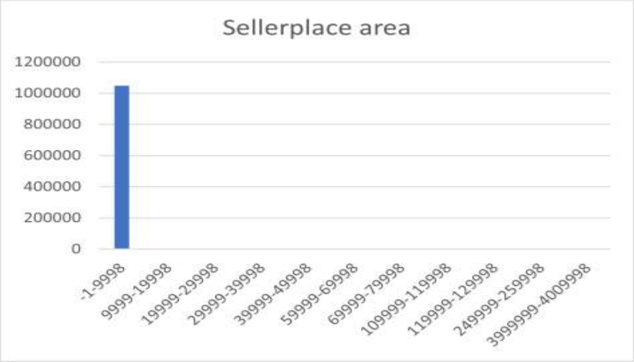
## Data Imbalance:

- Below are the columns where data is unevenly distributed.





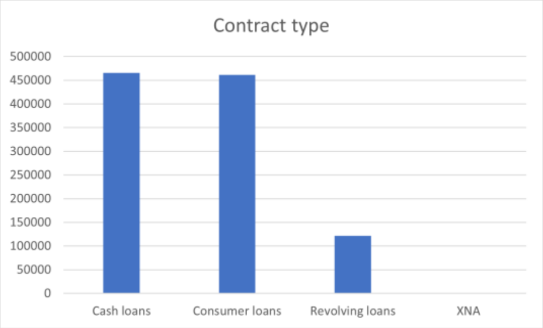


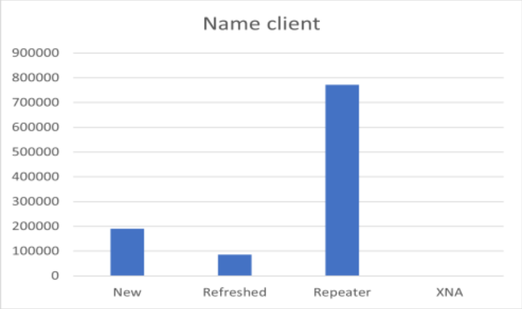


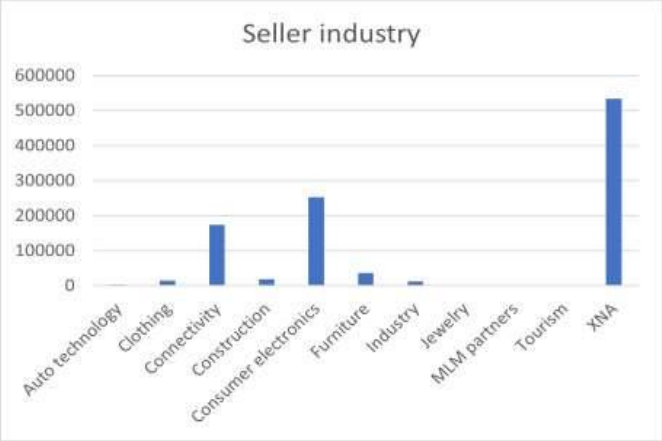
## EDA (Exploratory Data Analysis)

### Univariate Analysis:

- The analysis of the univariate data reveals that cash and consumer loans are the preferred payment methods among customers.
- The majority of clients are repeat customers, indicating a positive customer retention rate.
- There has been an increase in loan applications within the past ten months, and consumer gadgets appear to be a popular loan purpose among applicants.

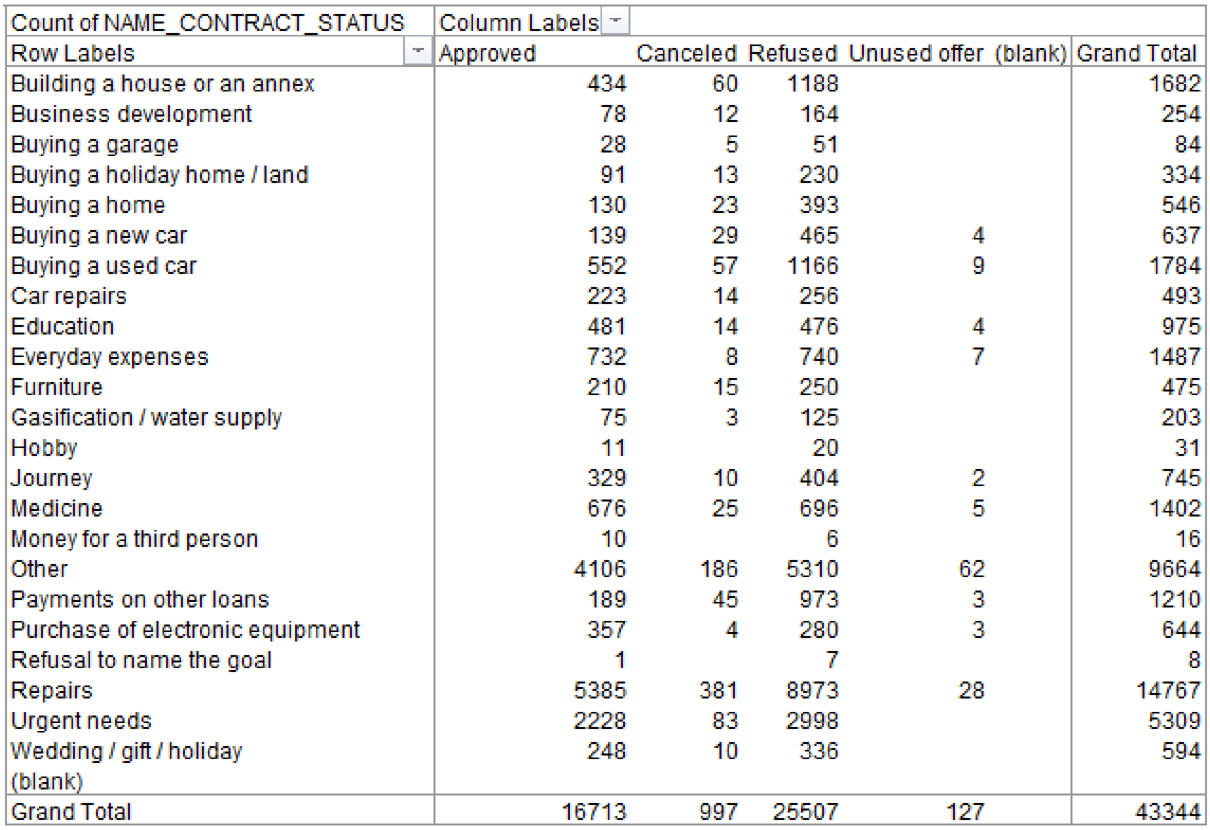




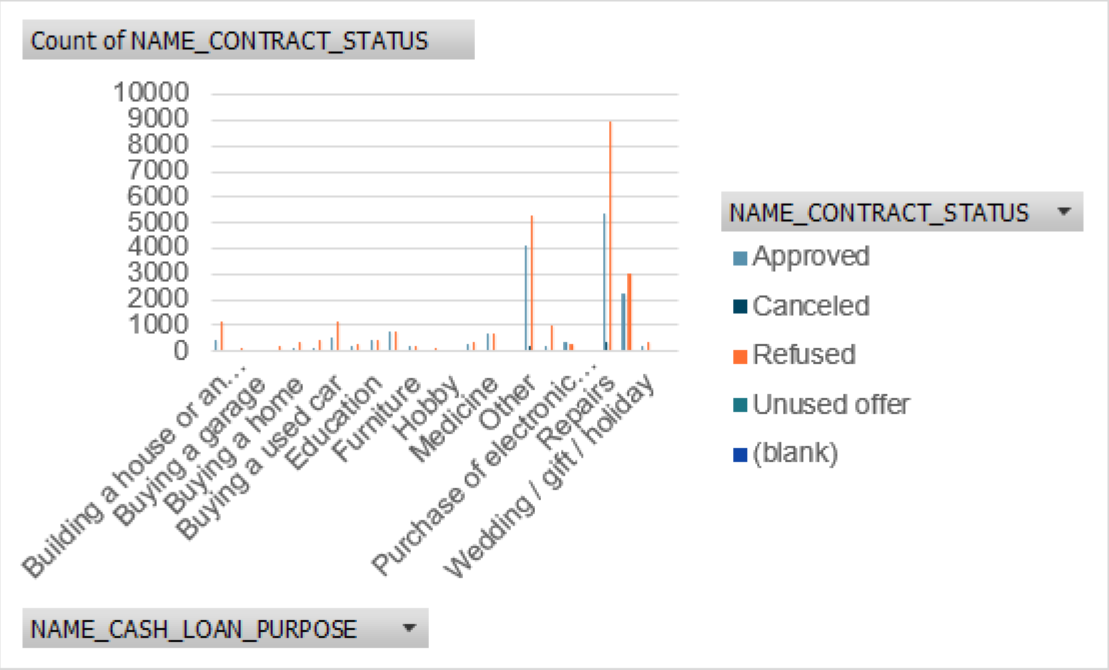


## Distribution of name contract status:

- The bivariate analysis revealed interesting insights. Loan applicants requesting amounts over Rs. 350,000 faced higher denial rates.
- Credit and cash agency loans had a higher cancellation rate. New clients experienced a higher loan approval rate. Car loans had lower approval rates. Loans associated with MLM partnerships were more likely to be cancelled.
- Consumer loans had high approval rates and low cancellations. Loans in the first selling place area group had a higher cancellation rate. Repeat applicants within a 10-month period had higher loan cancellations.
- Walk-in loans had a higher refusal rate. These findings provide valuable information for risk assessment in the loan approval process.



| Count of NAME_CONTRACT_STATUS | Column Labels | | | | | |
|---|---|---|---|---|---|---|
| Row Labels | Approved | Canceled | Refused | Unused offer | (blank) | Grand Total |
| Building a house or an annex | 434 | 60 | 1188 | | | 1682 |
| Business development | 78 | 12 | 164 | | | 254 |
| Buying a garage | 28 | 5 | 51 | | | 84 |
| Buying a holiday home / land | 91 | 13 | 230 | | | 334 |
| Buying a home | 130 | 23 | 393 | | | 546 |
| Buying a new car | 139 | 29 | 465 | 4 | | 637 |
| Buying a used car | 552 | 57 | 1166 | 9 | | 1784 |
| Car repairs | 223 | 14 | 256 | | | 493 |
| Education | 481 | 14 | 476 | 4 | | 975 |
| Everyday expenses | 732 | 8 | 740 | 7 | | 1487 |
| Furniture | 210 | 15 | 250 | | | 475 |
| Gasification / water supply | 75 | 3 | 125 | | | 203 |
| Hobby | 11 | | 20 | | | 31 |
| Journey | 329 | 10 | 404 | 2 | | 745 |
| Medicine | 676 | 25 | 696 | 5 | | 1402 |
| Money for a third person | 10 | | 6 | | | 16 |
| Other | 4106 | 186 | 5310 | 62 | | 9664 |
| Payments on other loans | 189 | 45 | 973 | 3 | | 1210 |
| Purchase of electronic equipment | 357 | 4 | 280 | 3 | | 644 |
| Refusal to name the goal | 1 | | 7 | | | 8 |
| Repairs | 5385 | 381 | 8973 | 28 | | 14767 |
| Urgent needs | 2228 | 83 | 2998 | | | 5309 |
| Wedding / gift / holiday | 248 | 10 | 336 | | | 594 |
| (blank) | | | | | | |
| Grand Total | 16713 | 997 | 25507 | 127 | | 43344 |

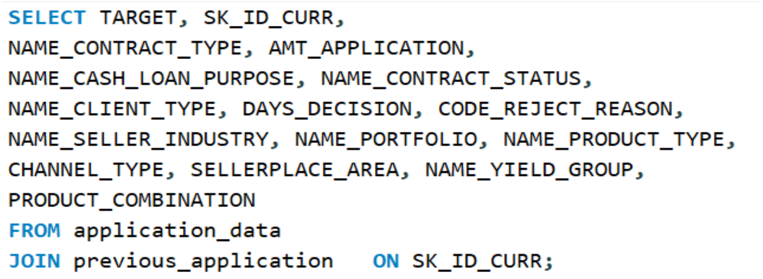


**Finding Correlations:** Top ten reasons for loan cancellation and refusal

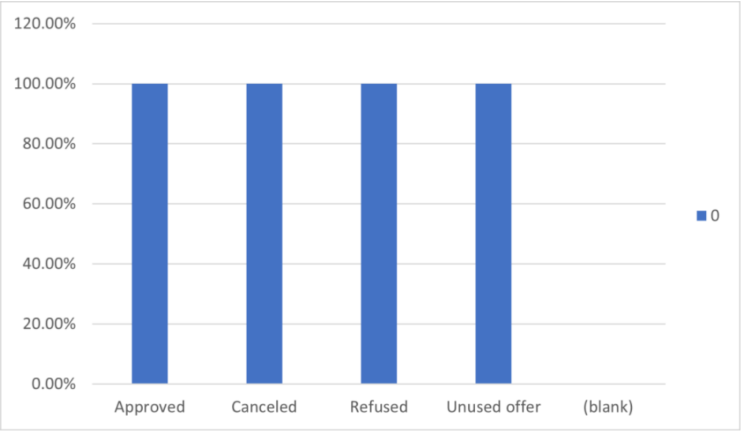
1. Amount Application
2. Cash loan Purpose
3. Goods Category
4. Product Combination
5. Product type
6. Channel type
7. Months Decision
8. Contract type
9. Client type
10. Payment type

**Combining two sheets:**

- I joined target column with the previous application data – I used mySQL for this



```
SELECT TARGET, SK_ID_CURR,
NAME_CONTRACT_TYPE, AMT_APPLICATION,
NAME_CASH_LOAN_PURPOSE, NAME_CONTRACT_STATUS,
NAME_CLIENT_TYPE, DAYS_DECISION, CODE_REJECT_REASON,
NAME_SELLER_INDUSTRY, NAME_PORTFOLIO, NAME_PRODUCT_TYPE,
CHANNEL_TYPE, SELLERPLACE_AREA, NAME_YIELD_GROUP,
PRODUCT_COMBINATION
FROM application_data
JOIN previous_application   ON SK_ID_CURR;
```



- Clients who have applied for previous loans have no defaults in currentloan.

## Summary:

Based on the analysis conducted, the following conclusions can be drawn:

- Highly recommended loan groups include previous application approved clients, married individuals, senior clients, more educated clients, high-income customers, and customers with strong work experience.

- High-risk loan groups consist of unemployed individuals, youth clients, customers with previously denied applications, low-income customers, individuals with insufficient external sources, customers with little work experience, customers on maternity leave, and customers with larger family sizes.

In summary, our EDA process involved cleaning the data, performing univariate and bivariate analysis, identifying correlations, and combining datasets. The insights gained from this analysis can aid in risk assessment and inform decision-making for loan approvals and risk mitigation strategies in the banking and financial services sector.

## Here are the links for the working csv files of the project:

**Application data:** “Application_data.csv”

**Previous data:** “Previous_data.csv”

## Thank you

- Sravan B