## Project Description

Use EDA to understand how customer attributes and loan attributes influence the likelihood of default. 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

**Download the dataset:** Begin by downloading the dataset provided for the project. Ensure that you have access to Microsoft Excel for data analysis.

**Perform Analysis:** Used Excel to perform the analysis and answer the questions mentioned in the project details

**Submit a Report:** Create a report (PDF/PPT) to present findings to the leadership team.

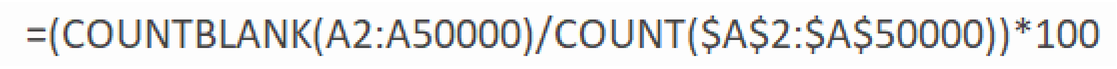
## Tech-Stack Used

I used Microsoft Excel 2019 to complete this project.

## Insights

**Identify Missing Data and Deal with it Appropriately:** It is essential to handle missing data effectively to ensure the accuracy of the analysis.

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 first calculated the blank data using the formula :



```
=(COUNTBLANK(A2:A50000)/COUNT($A$2:$A$50000))*100
```

I deleted the column having more than 30% blanks as it will not give me accurate information and in all other columns I replaced the blanks with median.

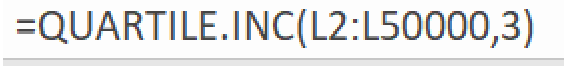
**Identify Outliers in the Dataset:** Outliers can significantly impact the analysis and distort the results.

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

I calculated Q1,Q3,IQR,Upper Bound ,Lower Bound using which we can find out what are the outliers.

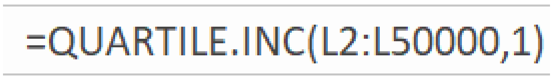
IQR = Q3-Q1 . Upper Bound = Q3 + (1.5*IQR) . Lower Bound = Q1 – (1.5*IQR).

Q3 =



```
=QUARTILE.INC(L2:L50000,3)
```

Q1=



```
=QUARTILE.INC(L2:L50000,1)
```

**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.

Used pivot tables to analyse every variable for data imbalance.

**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.

Used pivot tables to perform univariate, segmented univariate and bivariate analysis with graphs.

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

In application_data I set target to 1 and then find the correlation to all other numerical columns.

In previous_application we filter data for refused,cancelled,unused offer and then find correlation with all other numerical fields.

Top 5 correlation in previous_application

| | | |
|---|---|---|
| AMT_GOODS_PRICE | *AMT_APPLICATION* | 0.989495 |
| AMT_CREDIT | *AMT_APPLICATION* | 0.98655 |
| AMT_GOODS_PRICE | *AMT_CREDIT* | 0.977429 |
| AMT_GOODS_PRICE | *AMT_ANNUITY* | 0.856335 |
| AMT_CREDIT | *AMT_ANNUITY* | 0.852998 |

## Result

I found outliers in the data using the Q1,Q3,IQR and made graphs to visualize it . We found that people having income bracket of 100000-200000 have taken the most loans.

People living in a house or apartment have taken most of the loans. People with Secondary / secondary special academic background have taken most loans.

Working class people take most number of loans.

In previous application most people have taken consumer loans. Most people have Name of seller industry as Consumer Electronics. Most of the people who took loans are repeaters.