Project 6 – Bank Loan Case Study

Report By – Siva Sankari H

1.Introduction

A bank loan is a debt that a person called borrower owes to a bank. Its an agreement between the borrower and the bank about a certain amount of money that the borrower will borrow and then pay back in specific interval at a specific interest rate.

In this project we will be working with a finance company that specializes in lending various types of loans to urban customers. The main aim here is to identify the patterns that indicate if a customer will have difficulty paying their instalments. This information can be used to make decisions such a denying the loan, reducing the loan amount or lending it in a higher interest rate to risky applicants.

1.1 Project Description

In this project we will be working with a finance company that specializes in lending various types of loans to urban customers. The company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. Our task is to use *Exploratory Data Analysis (EDA)* to analyse patterns in the data and ensure that capable applicants are not rejected.

2. Pre read –

When a person applies for a loan, the banks will go through the details given by the applicant and based on the applicant's credit worthiness they can either deny or approve it. When a customer applies for a loan, there are four possible outcomes:

a) Approved: The company has approved the loan application.
b) Cancelled: The customer cancelled the application during the approval process.
c) Refused: The company rejected the loan.
d) Unused Offer: The loan was approved but the customer did not use it.

Analysing the applications received is very important because –

a) If the applicant can repay the loan but is not approved, the company loses business.
b) If the applicant cannot repay the loan and is approved, the company faces a financial loss.

Both these situations will lead into serious situations. As an analyst our goal is to understand how customer attributes and loan attributes influence the likelihood of default and make decisions accordingly.

Before moving understanding a bit about risk analytics will help in better understanding of the loan process. Risk analytics allows businesses to measure and manage risk. Risk analysis involves the following steps –

1. Risk identification – identifying potential risks that could negatively impact business. These could be financial, operational or strategic in nature.
2. Risk Assessment – analysing the risk identified based on their potential impact and likely hood of occurrence. This helps in prioritising risk.
3. Risk Quantification – quantifying the potential loss. This could be financial, reputational or any other measurable loss.
4. Risk mitigation – based on assessment and quantification, strategies are developed to mitigate the identified risks.
5. Risk monitoring - identified risks are continuously monitored and reassessed. This helps in identifying any changes in the risk profile and taking necessary actions in a timely manner.

3. Dataset Description

Dataset given has 3 files.

*File a) – **previous_application.csv**:* Contains information about previous loan applications.

*File b) – **application_data.csv**:* Provides details about the current loan applications.

*File c)– **columns_description.csv**:* Describes the columns present in the other datasets, explaining what each column represents.

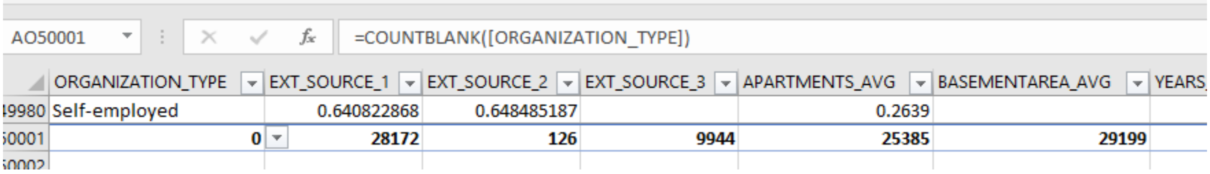
We will be working with **File b)** for our data analysis tasks. This file has 122 columns with 50000 entries. This file contains important information about our client like their contact info, place of residence, annuity amount paid, loan amount quoted and credited, goods under review, basic description of the goods and many such.

4. Data Analytics Tasks – Approach, Analysis and Insights gathered.

Task A: Identify Missing Data and Deal with it Appropriately:

Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

<u>Excel Function used –</u> We have used ***countblank*** function to count the number of blank entries in our dataset. We have total 122 columns. Countblank function on the next immediate row of each column will provide the blank row counts. Example is given below –

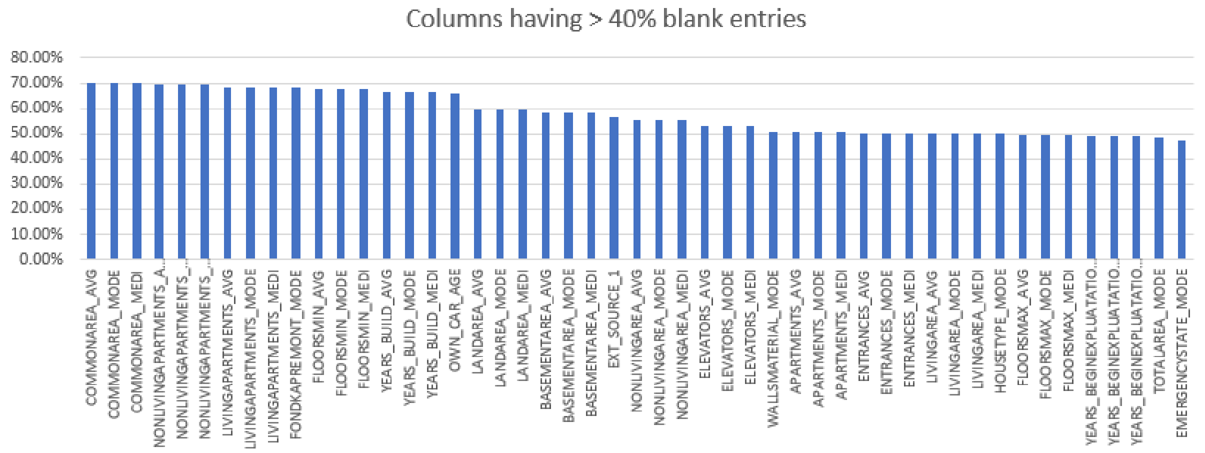


AO50001 =COUNTBLANK([ORGANIZATION_TYPE])

	ORGANIZATION_TYPE	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3	APARTMENTS_AVG	BASEMENTAREA_AVG	YEARS_
49980	Self-employed	0.640822868	0.648485187		0.2639		
50001	0	28172	126	9944	25385	29199	
50002							

For our ease, we will be deleting irrelevant columns from our dataset. From the list taken we will be removing all the columns having more than 40% blank entries. We have 49 such columns. Aside from this we will also be removing irrelevant columns. We are now finally left with 39 columns and 50000 rows.

The plot below shows the distribution of columns having greater than 40 % empty entry.



The selected 39 columns also have blank entries which we are filling in using ***mean imputation*** method. The data are distributed normally hence we are filling in the missing values with the mean of that column. This ensures that there is no much variance introduced by the missing values.

Modified dataset is attached with the report in the section **"Drive Link".**

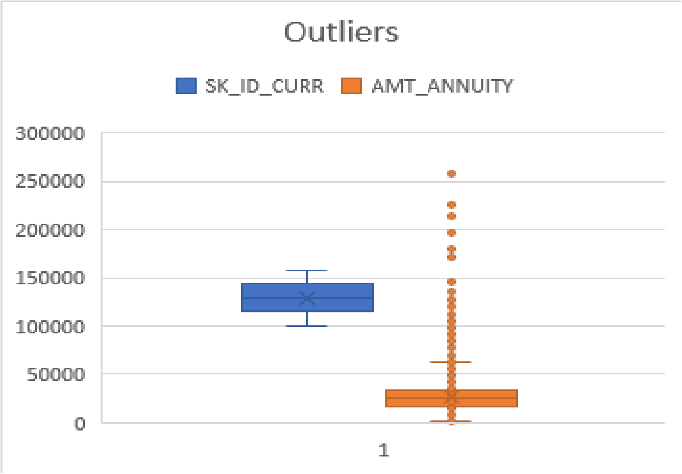
B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

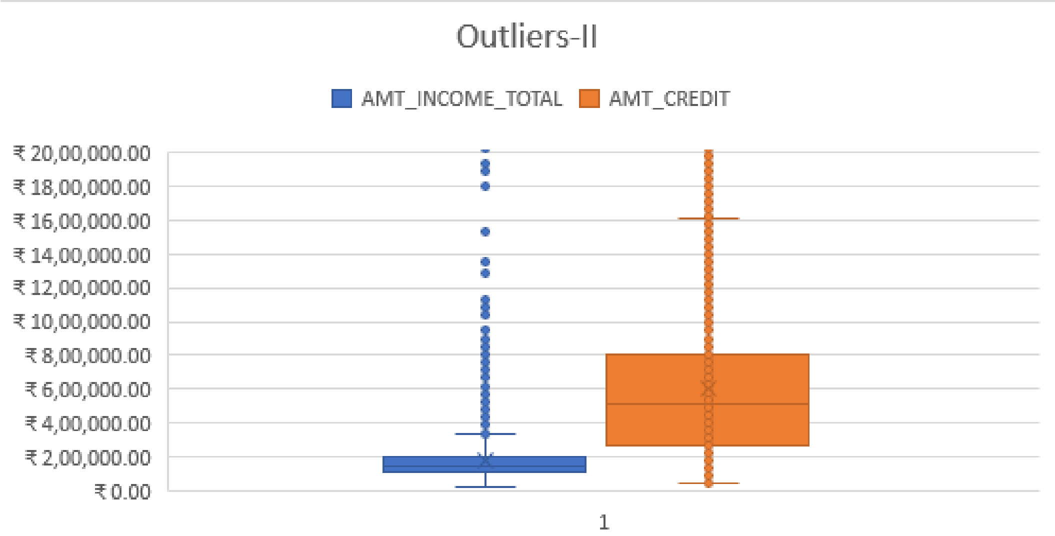
Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

Excel function used – We have used QUARTILE (ARRAY, VALUE) function to measure the outliers. Value = 0 – gives minimum value of the range, value = 1 – gives 1st quartile (25th percentile), value = 2 – gives median or middle, value = 4 – gives 2nd quartile (75th percentile).

We are only considering the numerical values and columns having that. Separating categorical and numerical values, quartile is determined. The quartile determined for each column in is attached with the modified excel sheet.

We have calculated mean, first quartile, median, 3rd quartile, maximum value, inter-quartile range (IQR), and the outlier points. The data points falling below **Mean-1.5IQR** and above **Mean+1.5IQR** will be the outlier points. Box plot representation for some of the columns highlighting outliers is shown below.





C. Analyse 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.

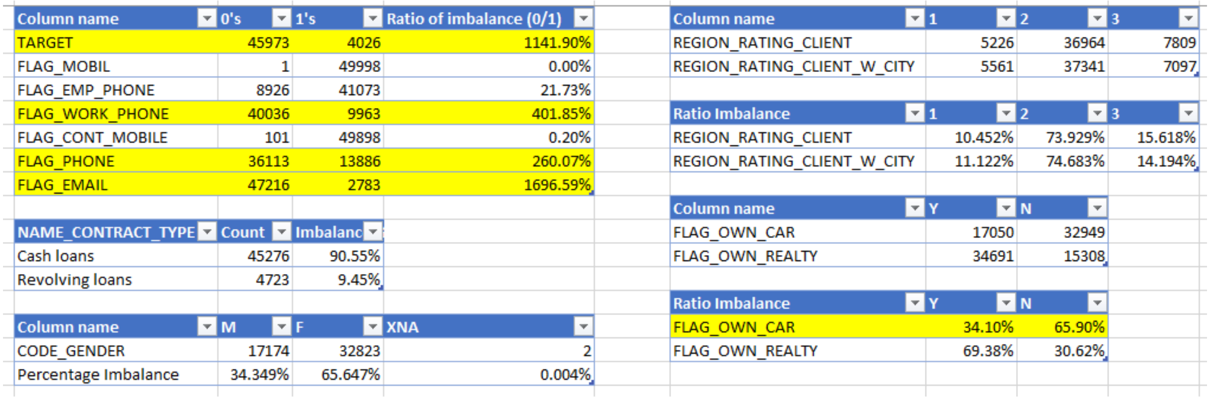
Data Imbalance refers to the case where negative proportions (or class) is significantly larger than the positive proportion. This is a commonly occurring problem in binary classifications.

We have 13 such columns which collects binary data. The calculations for imbalance is shown in the figure below.

Observations –

1. It can be seen that negative class is significantly larger than the positive class for columns "target" and flag variables – work phone, phone, email – indicating imbalance.
2. This means that our dataset does not have significant information on the applicants contact details even though overall contact details are available.

3. All the applicants have provided their primary contact info and have updated the same in case of any changes.
4. Cash loans – new loans play significant role in the company finance pool.
5. Most of the loans are applied by the female applicants.
6. Tier 2 regions have applied for majority of the loans.
7. In case of goods/ assets – cash loans have been applied the most.



Column name	0's	1's	Ratio of imbalance (0/1)
TARGET	45973	4026	1141.90%
FLAG_MOBIL	1	49998	0.00%
FLAG_EMP_PHONE	8926	41073	21.73%
FLAG_WORK_PHONE	40036	9963	401.85%
FLAG_CONT_MOBILE	101	49898	0.20%
FLAG_PHONE	36113	13886	260.07%
FLAG_EMAIL	47216	2783	1696.59%

NAME_CONTRACT_TYPE	Count	Imbalanc
Cash loans	45276	90.55%
Revolving loans	4723	9.45%

Column name	M	F	XNA
CODE_GENDER	17174	32823	2
Percentage Imbalance	34.349%	65.647%	0.004%

Column name	1	2	3
REGION_RATING_CLIENT	5226	36964	7809
REGION_RATING_CLIENT_W_CITY	5561	37341	7097

Ratio Imbalance	1	2	3
REGION_RATING_CLIENT	10.452%	73.929%	15.618%
REGION_RATING_CLIENT_W_CITY	11.122%	74.683%	14.194%

Column name	Y	N
FLAG_OWN_CAR	17050	32949
FLAG_OWN_REALTY	34691	15308

Ratio Imbalance	Y	N
FLAG_OWN_CAR	34.10%	65.90%
FLAG_OWN_REALTY	69.38%	30.62%

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.

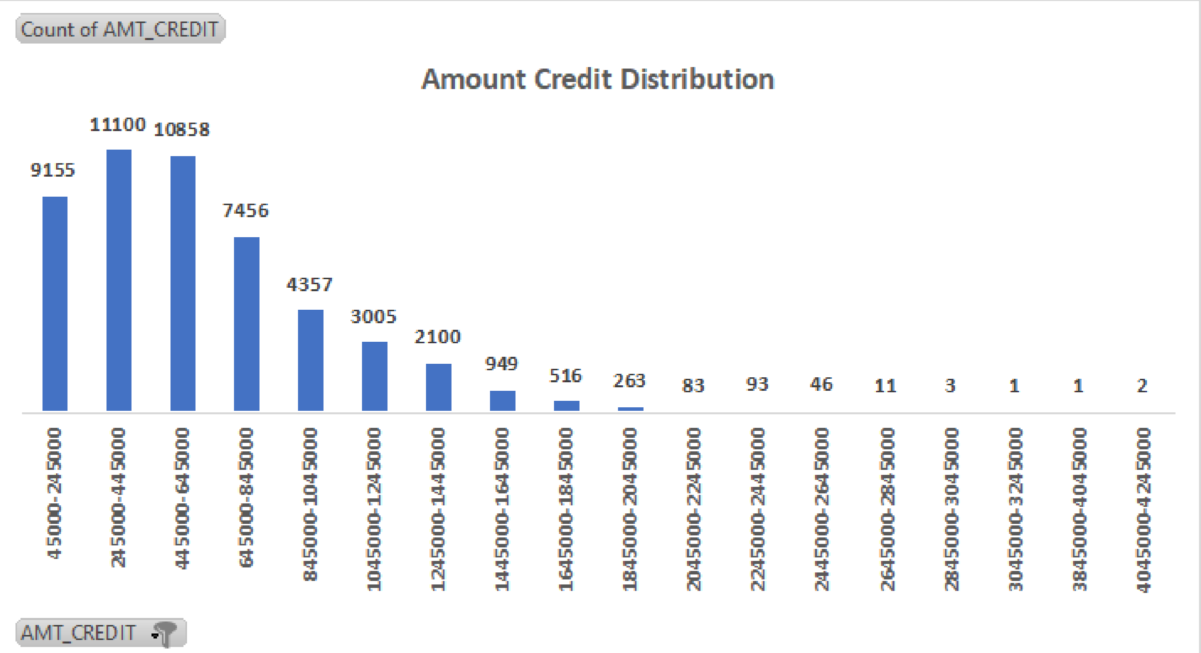
Excel function used - COUNT, AVERAGE, MEDIAN, and statistical functions for descriptive analysis.

Univariate, Bivariate and Segmented Univariate analysis are all used in statistics to describe the number of variables involved in data and the purpose of analysis.

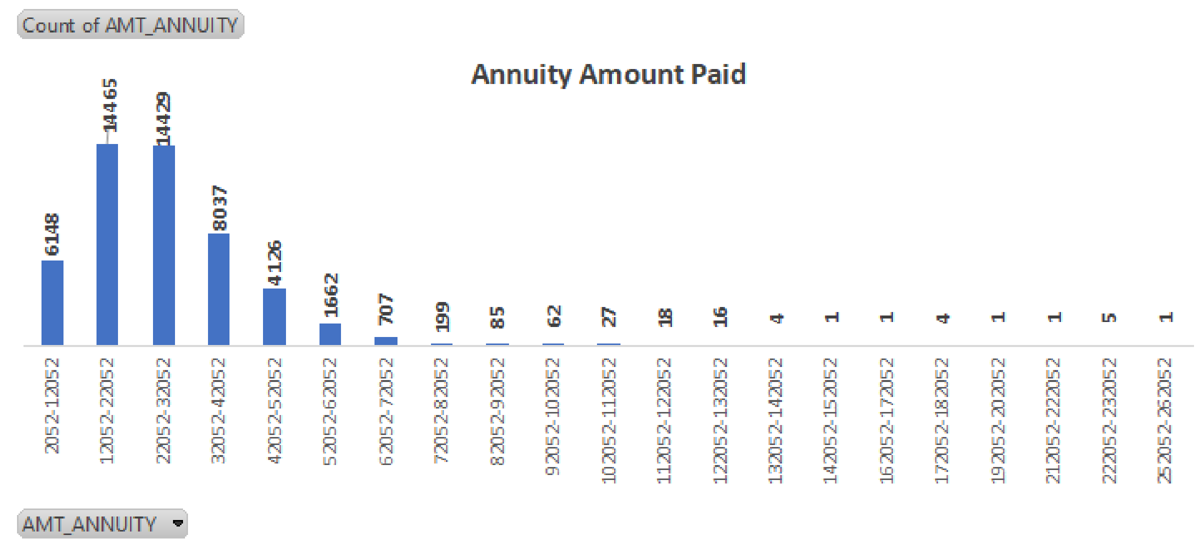
Univariate Analysis – We need to perform this analysis on individual variables to understand their spread. Uni – means single. This analysis can be done by simple statistical descriptions. We can understand the spread of the data, central tendency and the range of values. This method does highlight any correlation between variables, cannot be used to identify the patterns.

Insights derived –

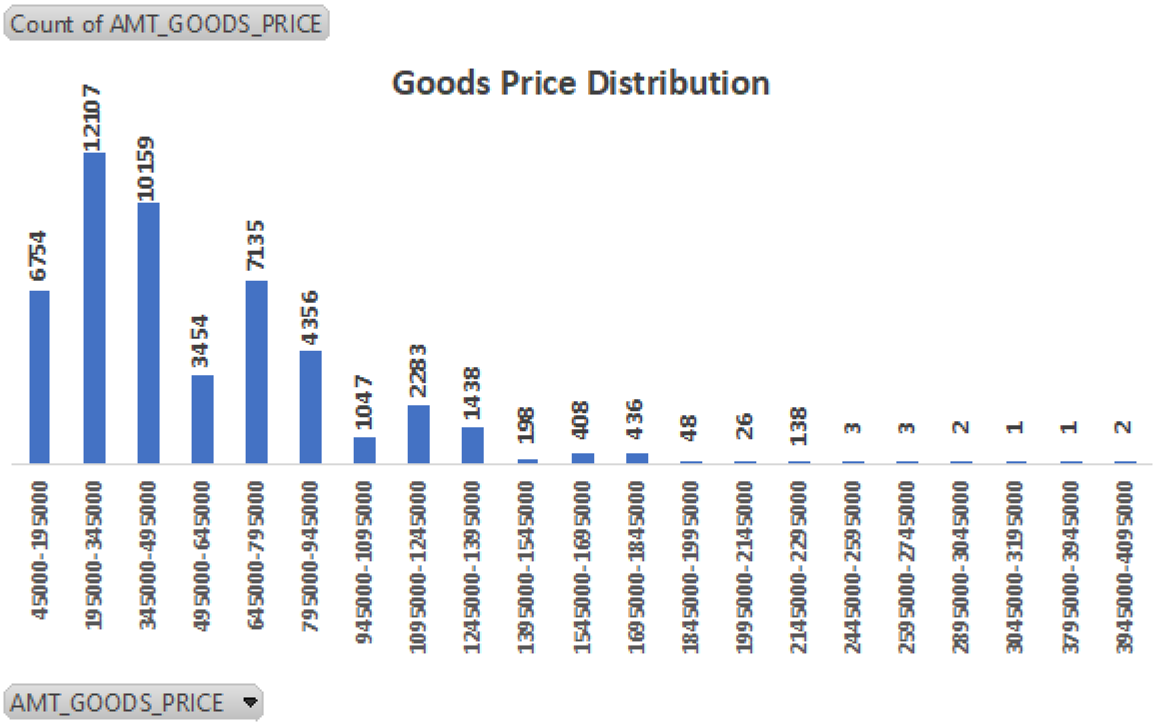
1. Amount Credit Column – we have come to a conclusion that the maximum loan amount credited falls within the range of ₹ 2,45,000 - ₹ 44,5,000 INR.
2. Applicants are more likely to receive easier loan approval when the requested loan amount falls within the range of ₹ 45,000 to ₹ 12,45,000.



3. Annuity Amount – it can be seen that maximum annuity paid by a borrower falls in the range ₹ 12,052 - ₹ 22,052. Annuity refers to the payments (usually of equal amounts) made to a bank over a specified period of time.
4. The bank calculates the annuity payments based on the principal loan amount, interest rate, and loan term. These fixed payments help ensure that the borrower can repay the loan in manageable amounts.



5. Goods Price Distribution – Goods Price here refers to the price of the goods/asset for which loan is applied for. It can be seen that loan is applied for a asset with price range between ₹ 195000 - ₹ 345000.



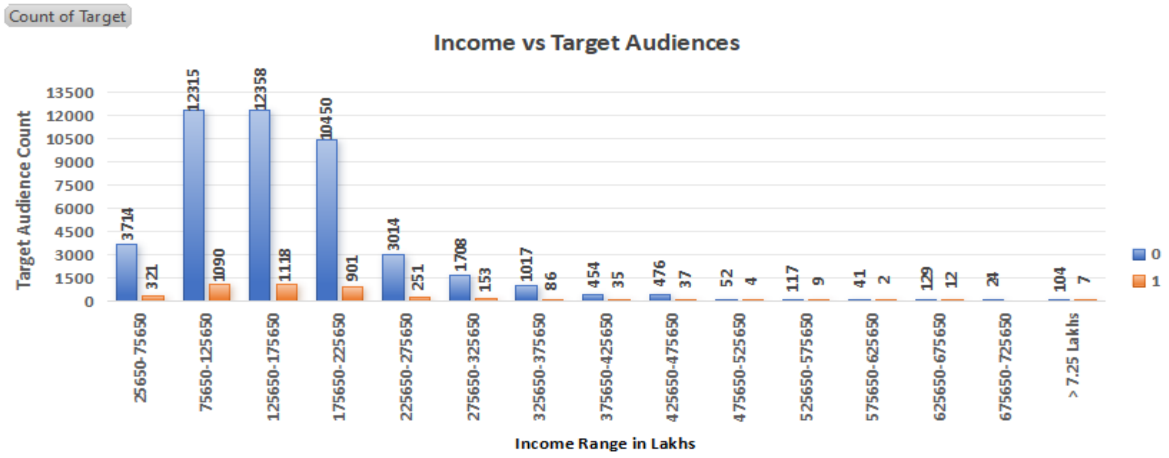
Bivariate Analysis – Analysis of any **concurrent relation between two variables** or attributes. This study explores the relationship of two variables as well as the depth of this relationship to figure out if there are any discrepancies between two variables and any causes of this difference. Some of the examples are percentage table, scatter plot, etc.

(All these analyses are done on target = 0 audience as we are more interested payment defaulters.)

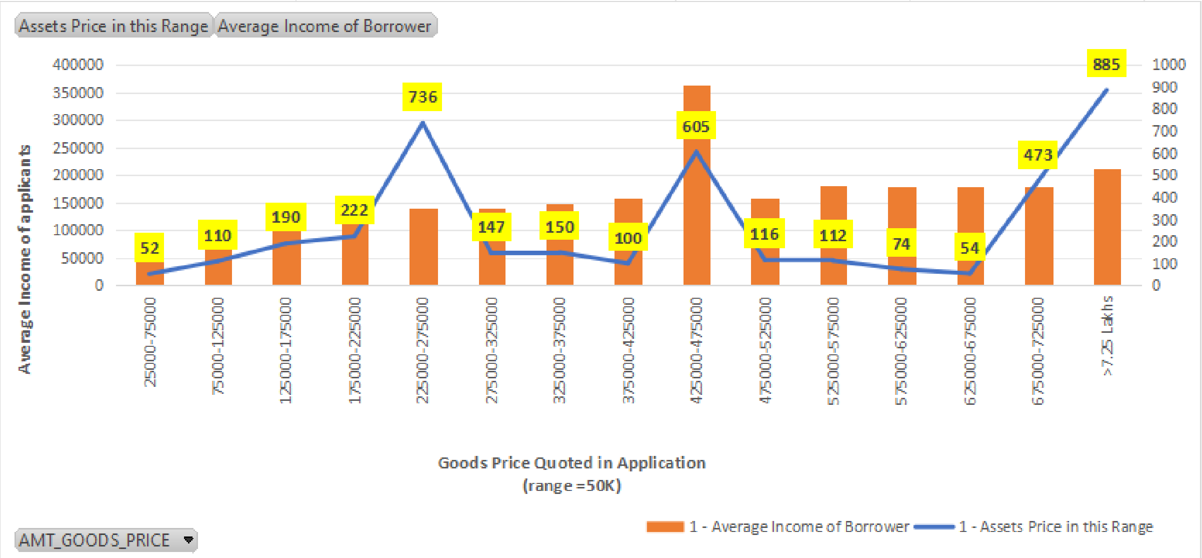
Insights derived –

The plot below shows Income projected vs Target Audience plot. Target value = 0 indicates clients having payment difficulty.

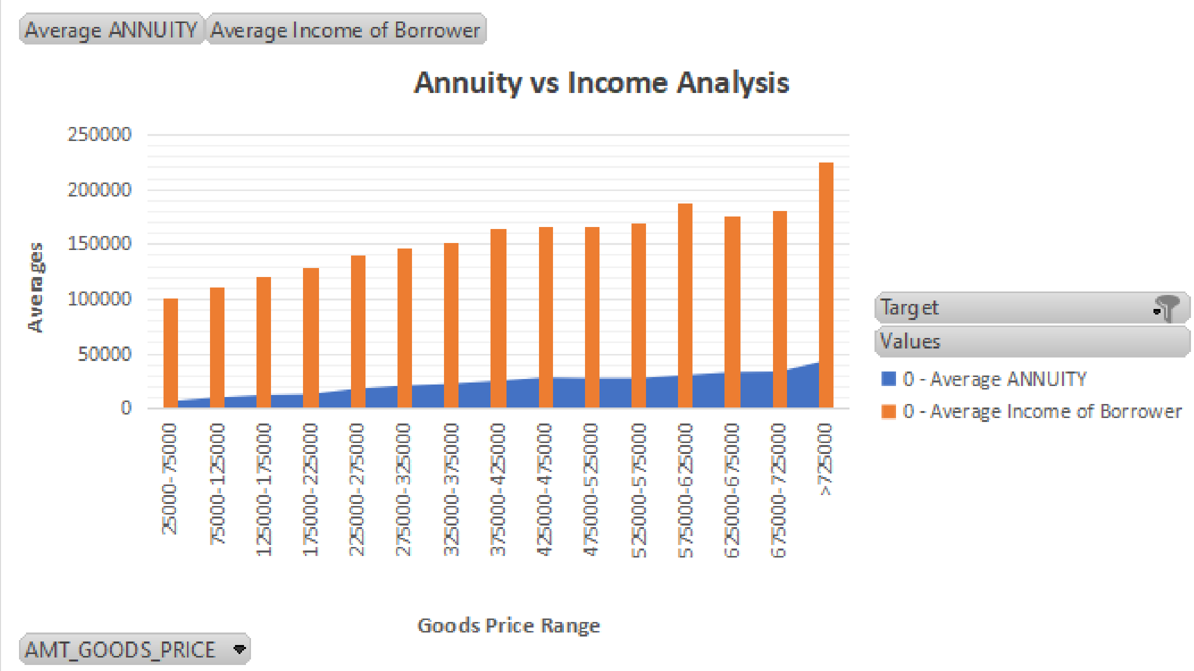
1. All the income category faces payment difficulties. Lot of factors may be a reason for this. Some may be, the price of the asset for which loan is applied for. Other financial and social commitments, etc.



2. The plot below shows the goods price projected vs the average income of applicants.
3. The primary axis – x indicates the goods price quoted in the loan applications and the y axis indicates average income of the applicants falling under this price range.
4. Secondary axis – blue line – indicates the number of assets falling under this income range. This means – we have 52 assets falling in price range ₹ 25,000 – ₹ 75,000. The average income in this category is near ₹ 50,000. Which mean, 52 applicants having average salary in ₹ 50000 margin have quoted for goods of price close to ₹ 75,000.



5. The plot below shows the Annuity vs Income Analysis against target = 0. We are only interested in clients having payment difficulties. Bar plot represents the average income of the borrower. Area plot indicates the average annuity to be paid. Maximum annuity ever paid by a client is around ₹ 50,000.



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.

Excel functions used - CORREL to calculate correlation coefficients between variables and the target variable within each segment.

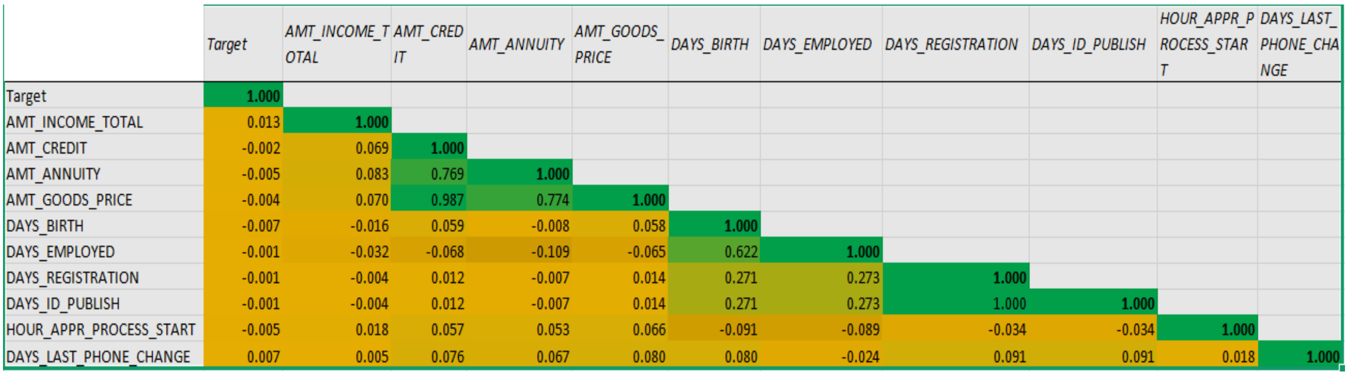
The figure below shows the correlation matrix generated via excel.

1. Green highlighted cells indicate positive correlation.
2. Yellow indicates negative correlation.
3. Light shaded region indicates values closer to 0 - means no correlation.

Positive correlation indicates linear relationship between variables. Matrix value = 1 shows self-correlation of a variable.

From the plot it can be seen that –

a) Amount credited has 59% correlation with annuity amount paid and 97% correlation with goods price. This means 59% of the annuity amount paid can be explained by credit amount factor and 97% of goods price can be explained by credit amount.



	Target	*AMT_INCOME_TOTAL*	*AMT_CREDIT*	*AMT_ANNUITY*	*AMT_GOODS_PRICE*	*DAYS_BIRTH*	*DAYS_EMPLOYED*	*DAYS_REGISTRATION*	*DAYS_ID_PUBLISH*	*HOUR_APPR_PROCESS_START*	*DAYS_LAST_PHONE_CHANGE*
Target	1.000										
AMT_INCOME_TOTAL	0.013	1.000									
AMT_CREDIT	-0.002	0.069	1.000								
AMT_ANNUITY	-0.005	0.083	0.769	1.000							
AMT_GOODS_PRICE	-0.004	0.070	0.987	0.774	1.000						
DAYS_BIRTH	-0.007	-0.016	0.059	-0.008	0.058	1.000					
DAYS_EMPLOYED	-0.001	-0.032	-0.068	-0.109	-0.065	0.622	1.000				
DAYS_REGISTRATION	-0.001	-0.004	0.012	-0.007	0.014	0.271	0.273	1.000			
DAYS_ID_PUBLISH	-0.001	-0.004	0.012	-0.007	0.014	0.271	0.273	1.000	1.000		
HOUR_APPR_PROCESS_START	-0.005	0.018	0.057	0.053	0.066	-0.091	-0.089	-0.034	-0.034	1.000	
DAYS_LAST_PHONE_CHANGE	0.007	0.005	0.076	0.067	0.080	0.080	-0.024	0.091	0.091	0.018	1.000

3. Tech Stack Used

The goal in this project is to use EDA to understand how customer attributes and loan attributes influence the likelihood of default. We have used Excel's COUNTIF, SUMIF, AVERAGE, MEDIAN, MODE, CORREL, IF to analyse the dataset and gain insights. We have used Bar, Column, Pie and Box plot to visualise.

4. Drive Link

Bank Laon DataSet Solutions

Bank Laon Presenattion