SUMIT SHAMLAL CHAURE

**Bank Loan Case Study**

Trainity Project 6 (Final Project -2)

horizontal line



*Figure 1 -* [*https://trainity.link/data/project06*](https://trainity.link/img/data-project/Loan_Case.png)

***Note*** *– Docx files uploaded on drive gets converted to sheets so much of the connections and pivot relations are lost so if possible download the zip file or main file so that all the things are intact and graphs or pivots show data accordingly while I have attached individual questions the sheets might not show up everything so download the excel files not Gdocs file to see the analysis.*

*Important Links (*[*Tap Here*](#_Important_Links_:)*) (Drive* [*Folder*](https://drive.google.com/drive/folders/1eyK3giYVshfAOk0skO6fRiwGfh5QLEg5?usp=drive_link)*)*

***INDEX***

Contents

[Introduction](#_top) : ………………………………………………………………………………………………………………………………1

[Description …………………………………………………………………………………………………………………………………..3](#_Toc148130480)

[Problem Statement: 3](#_Toc148130481)

[Business Objectives: 3](#_Toc148130482)

[Requirements – 4](#_Toc148130483)

[Data Analytics Tasks : 8](#_Toc148130490)

[A) 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.* 8](#_Toc148130491)

[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.* 12](#_Toc148130492)

[C) : 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.* 16](#_Toc148130493)

[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.* 19](#_Toc148130494)

[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.* 27](#_Toc148130495)

[Important Links & Greetings : 30](#_Toc148130496)

# **Description**

## Problem Statement:

Imagine you're 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.

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

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

The dataset I will be working with contains information about loan applications ie. Current application or **application\_data.csv file**. It includes two types of scenarios:

1. 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.
2. All other cases: These are cases where the payment was made on time.

## 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.

# Requirements –

## 1) Project Description:

The main aim of the project is to find the use the knowledge of Descriptive statistics & make use of visualization tools to predict the loan payments and the factors that may otherwise hinder with the timely payment of debts. In order to do that we need find out necessary insights from excel sheet to grab the details and trends of the applicants like his income, lifestyle and factors that may affect his budgets and spending’s like family members and children’s count for which the client need to make spends in the long run also factors like the place where the applicant is planning to buy the house or utilizing the credit amount so as to make assumptions of his finances.

After looking at the data we plan a format to operate on the data, tools needed and charts that can be useful for trends etc. while keeping in the questions in mind. The descriptive analysis at various stages of the questions helps us to get the major trends and their effects on loan repayment analysis as well gain deep insights for deciding the factors for future loan approvals by analyzing trends according to income, profession, education , family members and region counts and taking into account the defaulting factors and other correlated factors that we gain from the overall dataset analysis.

## 2) Approach:

I first analyzed the data and looked for null values, blanks, duplicates and treated them using basic functions like delete cells, find & replace, remove blank rows , imputing with appropriate statistics like mean, mode, median values etc. For certain cell values I changed them to suit better with other values like – (age in years instead of days) etc. After confirming that the data has little to less outliers (single high income or credit limit etc.) and saved the raw data to work on with the operations. Also, I removed the most irrelevant columns from the dataset as the presence of such redundant data will not help us in the further analysis like few categorical columns had large amounts of blanks or numerical data with 50% and above missing values.

The analysis based on excel functions, Descriptive analysis, pivot tables has insights at the bottom of the screengrabs to let the others understand the aim of each analysis.

## 3) Tech-Stack Used:

***Excel*** – The basic data manipulation, handling and overall pivot charts and the statistics has been handled using MS Excel also the descriptive statistics and correlation matrix have been done on same.

***Google Sheets*** – Used to do the basic data manipulation and to get column stats (gif added) and also to verify some statistics before treating them.

***Word*** – The report is written in word/docx format using MS Word and then exported to pdf.

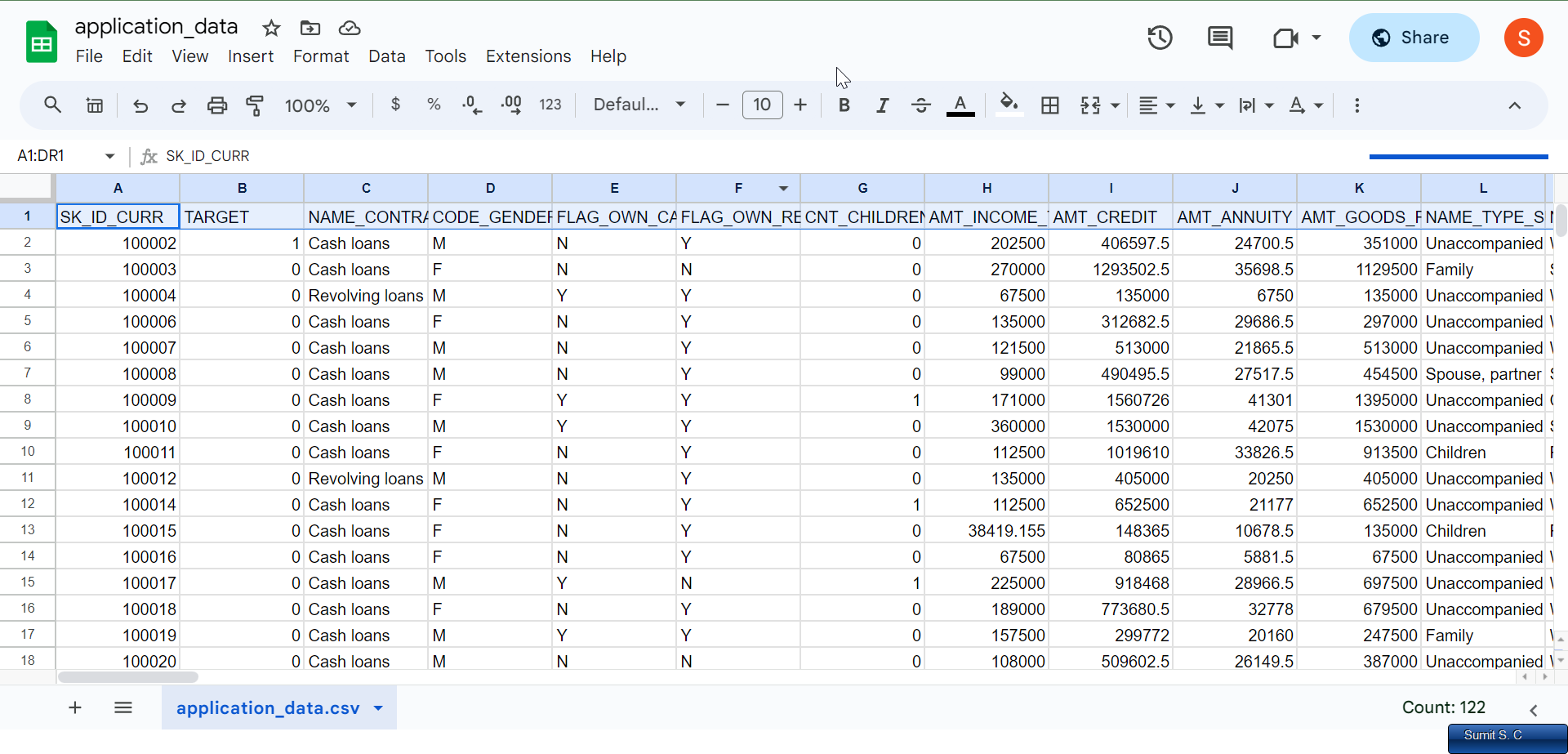
***Drive*** – To upload all the essential files attached in the report for reference & report upload.

***Loom*** – To make the video presentation of the analysis of each individual questions (links attached for the same below)

## 4) Insights:

I have done the analysis on **application\_data.csv** file which pertains to the current application details of the clients.

The summary for each query is given with the screenshot but to summarize the overall dataset I came to the conclusion that among the total data set we had **49999 data rows, 122 columns & a total of 4735935** cells in the given dataset.



After removing the blanks, duplicates(no duplicates were found in dataset) and adjusting the non-relevant columns we made the dataset to ***49999 rows, 45 columns & 2250000 cells*** for our final calculation. Few of the columns had no direct use in the analysis but I have kept them as the values were used in correlation calculation for factors impacting defaults.

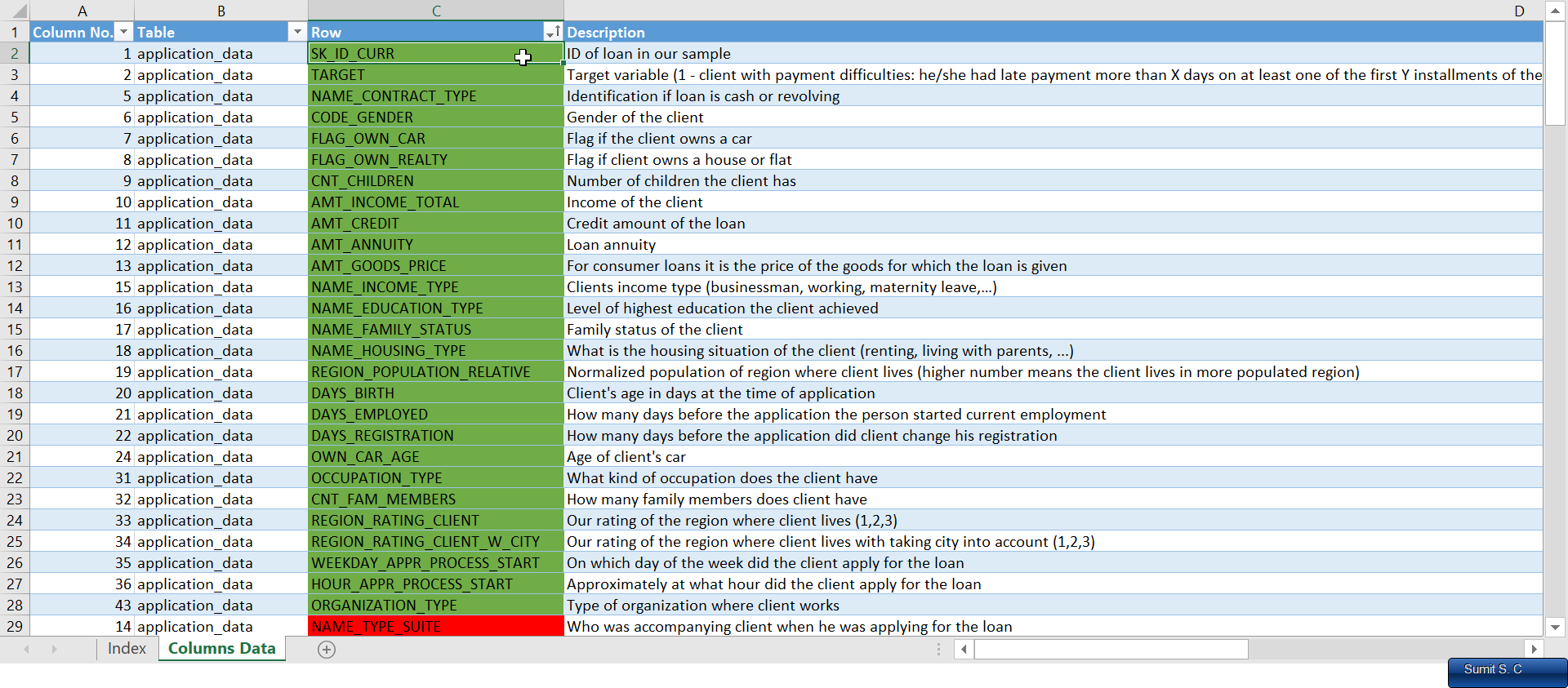


Figure 1 - Columns From application dataset to work (after cleaning)

For question 5 as we needed to calculate the correlation factor we only needed numerical columns in analysis so I made a separate cleaned data for same by removing the categorical columns and the columns which did not show any direct impact on loan payment as a factor so I trimmed it down to 15 or so columns to do correlation of the top factors that may hinder in the repayment process.

For more data insight on the questions look at the respective questions for screenshot, pivot charts and or pivot tables and descriptive statistics given at necessary places.

## 5) Result:

The Project has given me a good idea about the importance and vast variety of excel usage which helps us to look deep into plain numbers and generate a visually insightful data which can help business to gain knowledge and prepare for future as well as give out trends to focus on from the numerical data. The statistics section has helped me learn about the various concept which are useful for majority of the operation for handling and displaying basic charts and generate a meaningful dashboard as well as the use of Descriptive stats to get more deeper insights in the data. ***The overall result of the project has helped me gain knowledge about the real-life data operations that BFSI*** (Banking Financial services & Insurance) go through to analyze and decide whether to approve loans in future for similar types of clients or to change the amounts by checking for defaults, income and living standards of an individual.

# Data Cleaning Task:

As it is asked in the first question to do analysis and find out about the missing values and ways to handle them I will try to put all the necessary steps involved in data cleaning there itself instead of placing a separate block in the explanation.

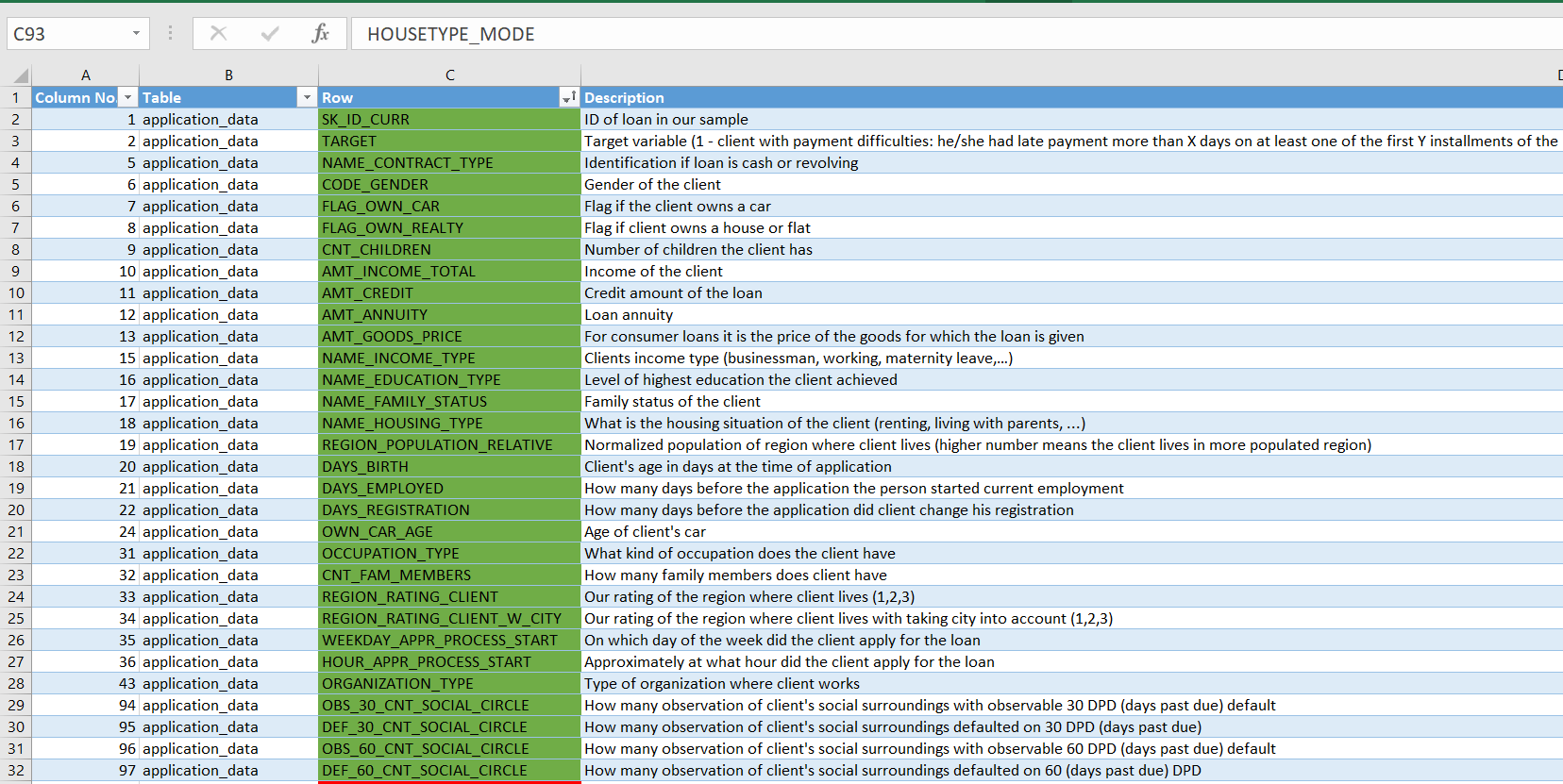
Will not explain much just have added the links of gif/process. (These were the basic steps to clean and adjust the data)

**Note:** *I have linked many files in the report like the separate excel sheet file for questions or High Quality image for graphs or GIF link to showcase cleaning process, also loom ppt links for reference and at the end the drive folder links. So the places highlighted by blue color and underline are hyperlink for sources which either gives the files,video or good quality of the screenshots and can be opened by tapping. Same in the main excel file I have linked the pages in index page for easy referencing so do download the xlsx files as it has pivot connections and other things while the files on google drive automatically converts to googledocs which loses graphs and pivot relations(individual questions might show error as data might be missing thing like that so download the* [*Main Excel sheet*](https://docs.google.com/spreadsheets/d/1-DphW5JffK6Y83PukCS6n775Td4rcS4P/edit?usp=drive_link&ouid=103173348283524847626&rtpof=true&sd=true) *or zip file)*

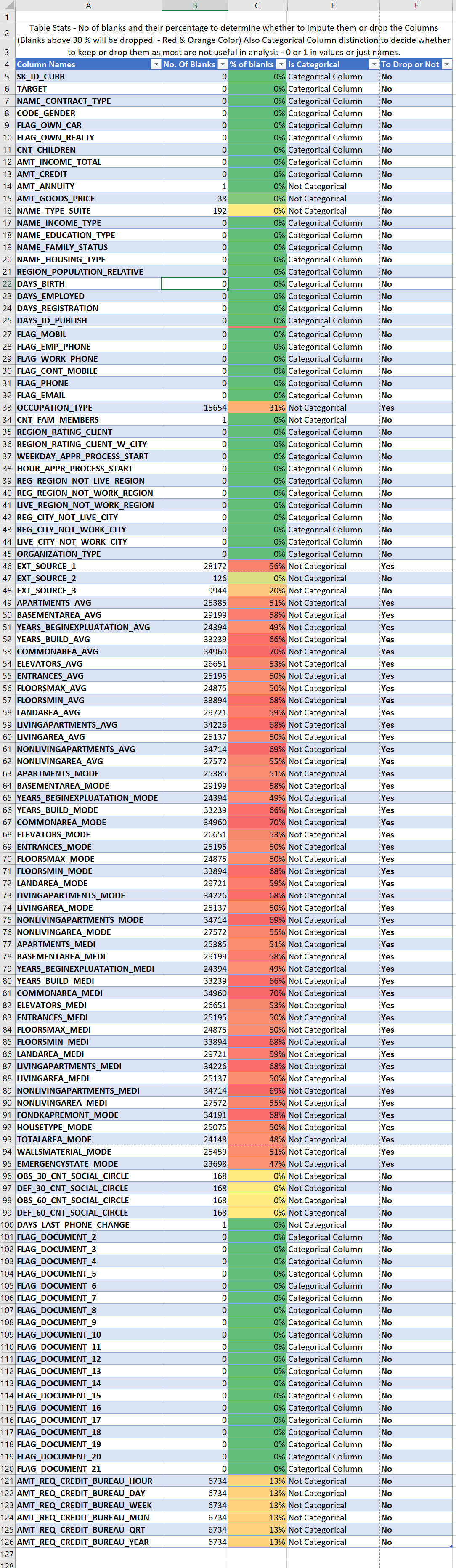
# Data Analytics Tasks :

## A) 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.



These are the remaining columns in our table after cleaning the missing data like blanks with deletion or appropriate imputation methods. Blanks and missing value shown below with a bar graph for comparison. (there are more columns which can not be fit in the screenshot)



HD Image 1 – [Pivot Table](https://mitsus.life-is-pa.in/6gLEi1jZs.png) HD Image 2 – [Gif of Pivot Table & Bar Graph](https://mitsus.life-is-pa.in/6gtt1oAxu.gif)

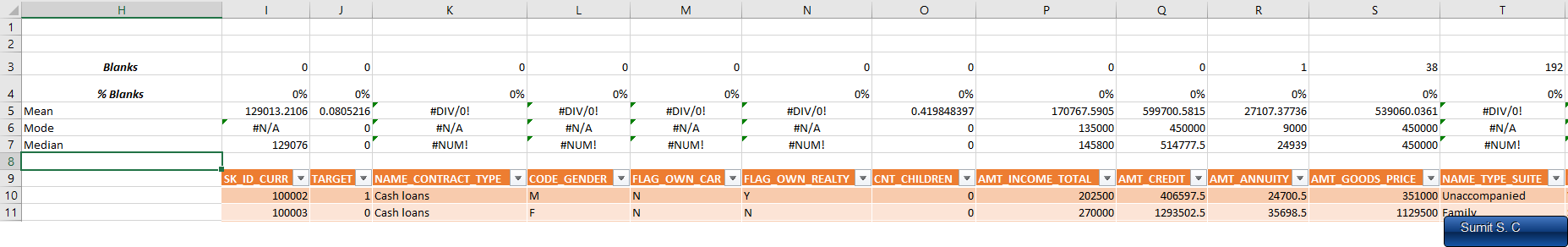
Since the Pivot table and Bar graph is lengthy I have captured a gif to show the image here also you can find the same in attached xls file. > [Gif](https://mitsus.life-is-pa.in/6gtt1oAxu.gif)

***Insights:***1. For missing value detection I first did a simple column statistics using google sheet to get info about the empty cells count of whole application dataset and understood the column relationship. For Duplicates & blanks we can use the data tabs feature to look for duplicates but there were none.

> [ColumnStats Gif](https://mitsus.life-is-pa.in/6g62wkHQ3.gif) ,

> [Duplicate Values](https://mitsus.life-is-pa.in/6g62wkHQ3.gif) , [Dupes](https://mitsus.life-is-pa.in/6gtt1oAxu.gif)

2. For missing values we used basic excel functions like **COUNT** (to get details about filled cells), **ISBLANK** or **COUNTBLANK** (to get the details about blank cells in a column). Using the counts of blank we determined the % of blanks for each column and then decided to treat the values *above 35%* and drop them for rest we did mean mode median imputation as applicable.

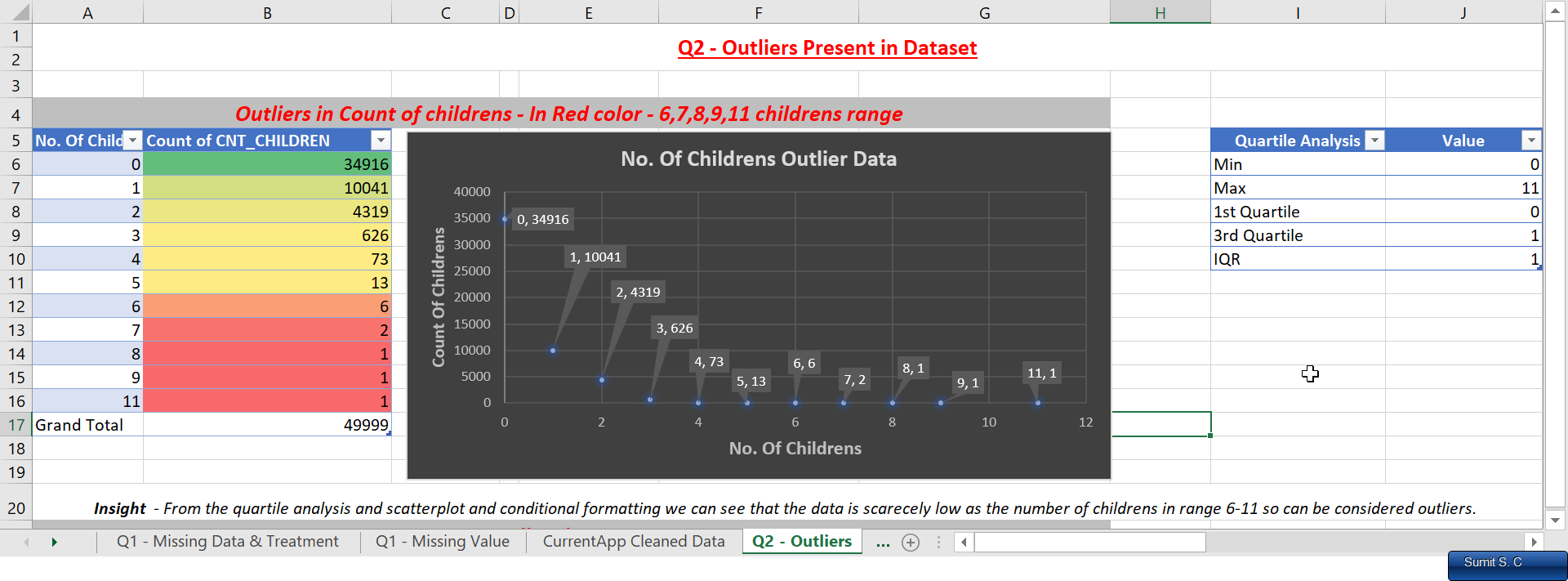


2 - For the categorical data directly doings stats will show some error like div by or n/a or num! but as most of them had no missing value that needed treatment below 35% so its fine to ignore them the formula was used to get the data for numerical missing data points calculation.

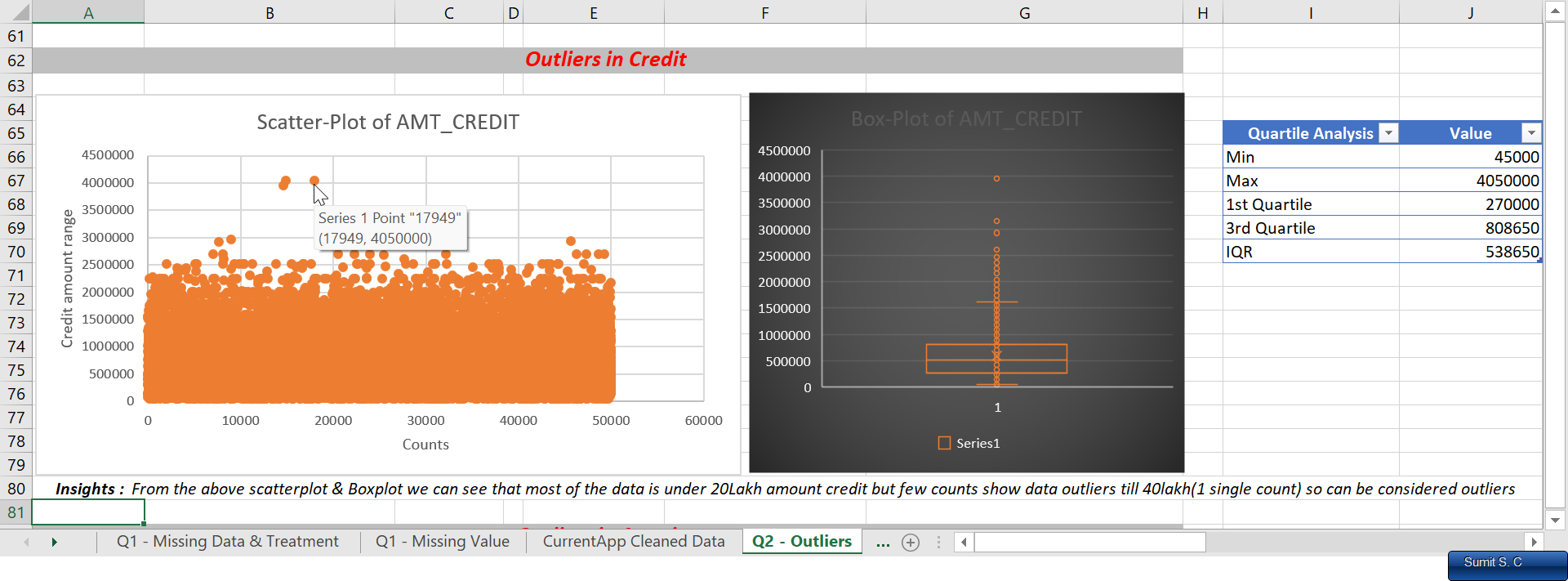
1. Looking at the Bar Graph of the missing data and doing some filtering using conditional formatting and **IF** formula we sorted the columns to be *dropped above 35%* (marked red in pivot table) and then the rest of the columns with missing values were treated and the final [application\_dataset file](https://docs.google.com/spreadsheets/d/1-LkOoaxAynku5tc9RiuQHg1NuHfU8Ako/edit?usp=drive_link&ouid=103173348283524847626&rtpof=true&sd=true) (Cleaned Dataset) was made to do analysis with necessary columns.
2. Files
   * + Q 1 [Excel File](https://docs.google.com/spreadsheets/d/1-7lBELisy6D5bxIddTK0rC8RY13DIn1E/edit?usp=drive_link&ouid=103173348283524847626&rtpof=true&sd=true), [Docs File](https://docs.google.com/spreadsheets/d/1Lifh1NQeOcFSKcdjSFAJMkFDzkW7sBmGNq8csUem5Y4/edit?usp=drive_link)
     + Loom Video Q1 – [Identifying Missing Values & Treating them.](https://www.loom.com/share/e46927a075c5454698b2cdd4428c94f4?sid=7245d958-9e44-4417-9c83-1ae0734fc521)

## 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.*

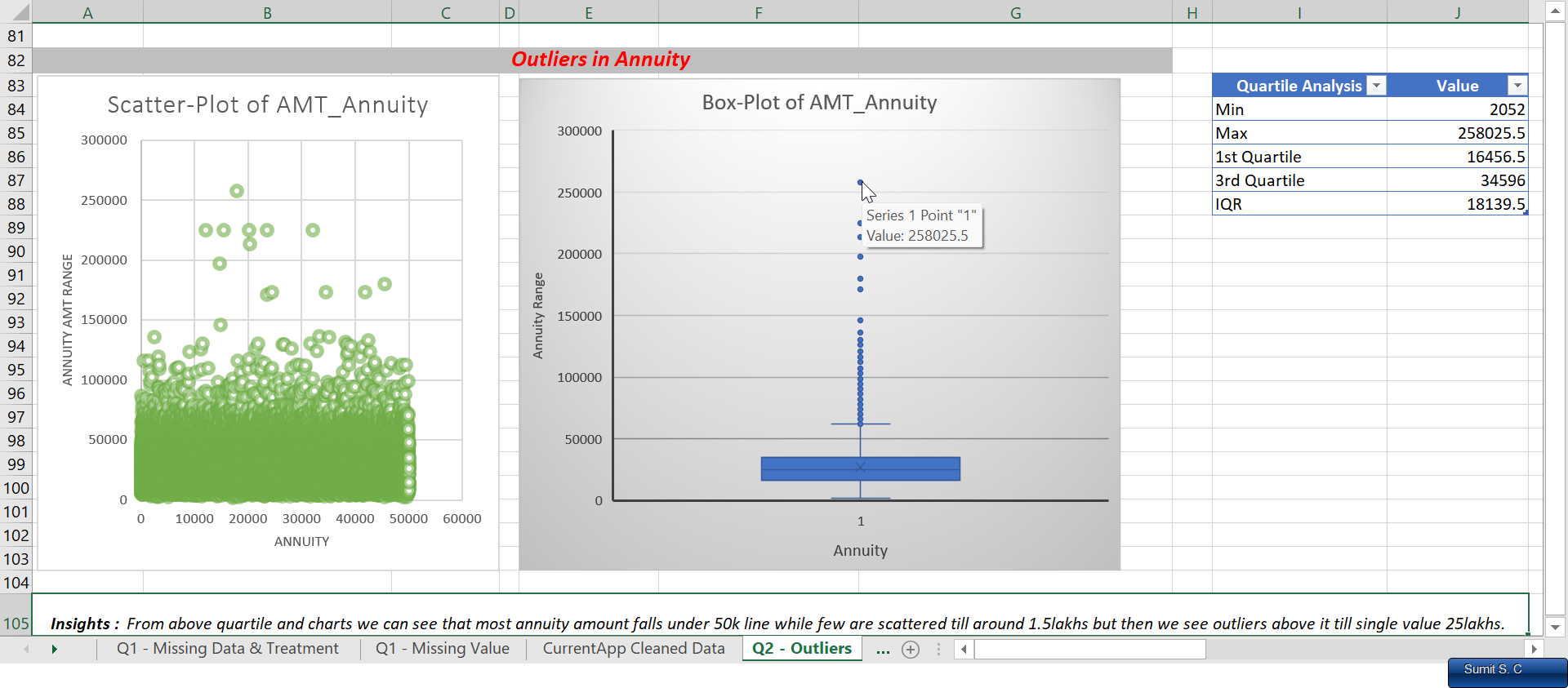
* *Task***:** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.



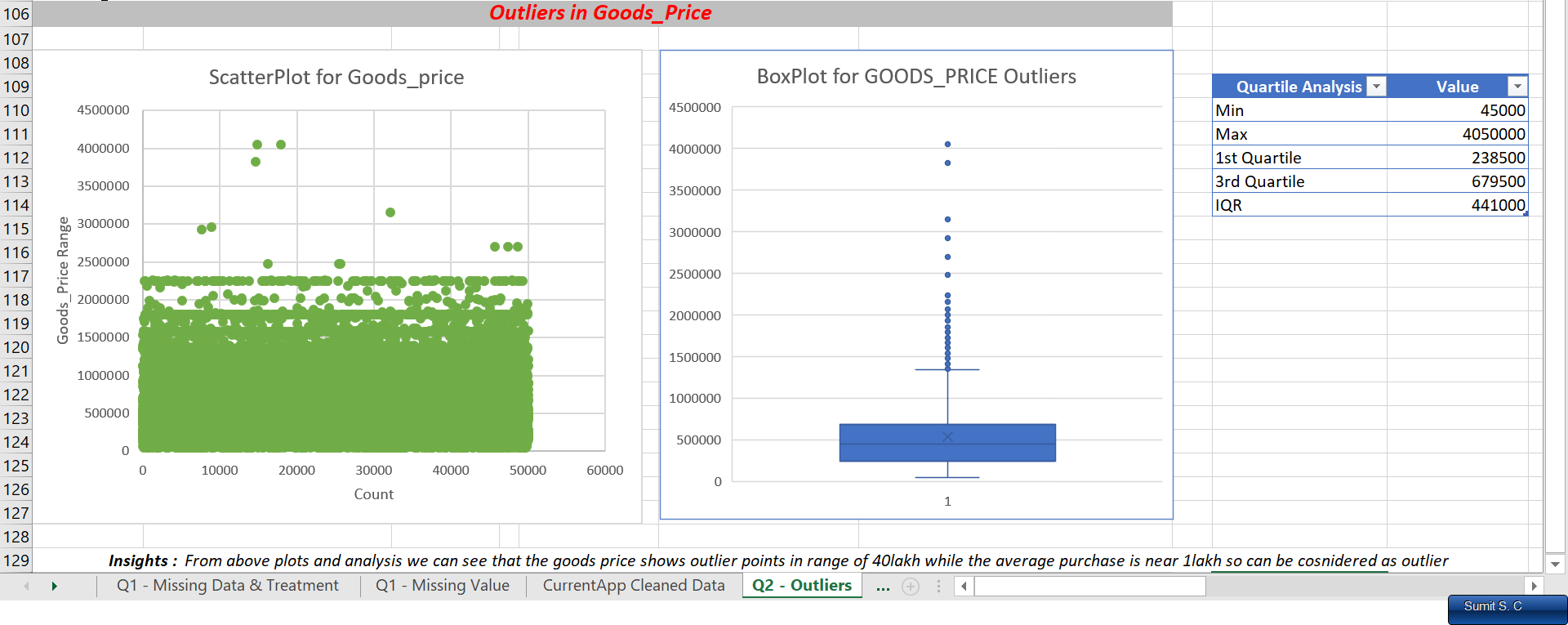
[Outlier 1](https://mitsus.life-is-pa.in/6gLRk9x7H.png) - In No. Of childrens



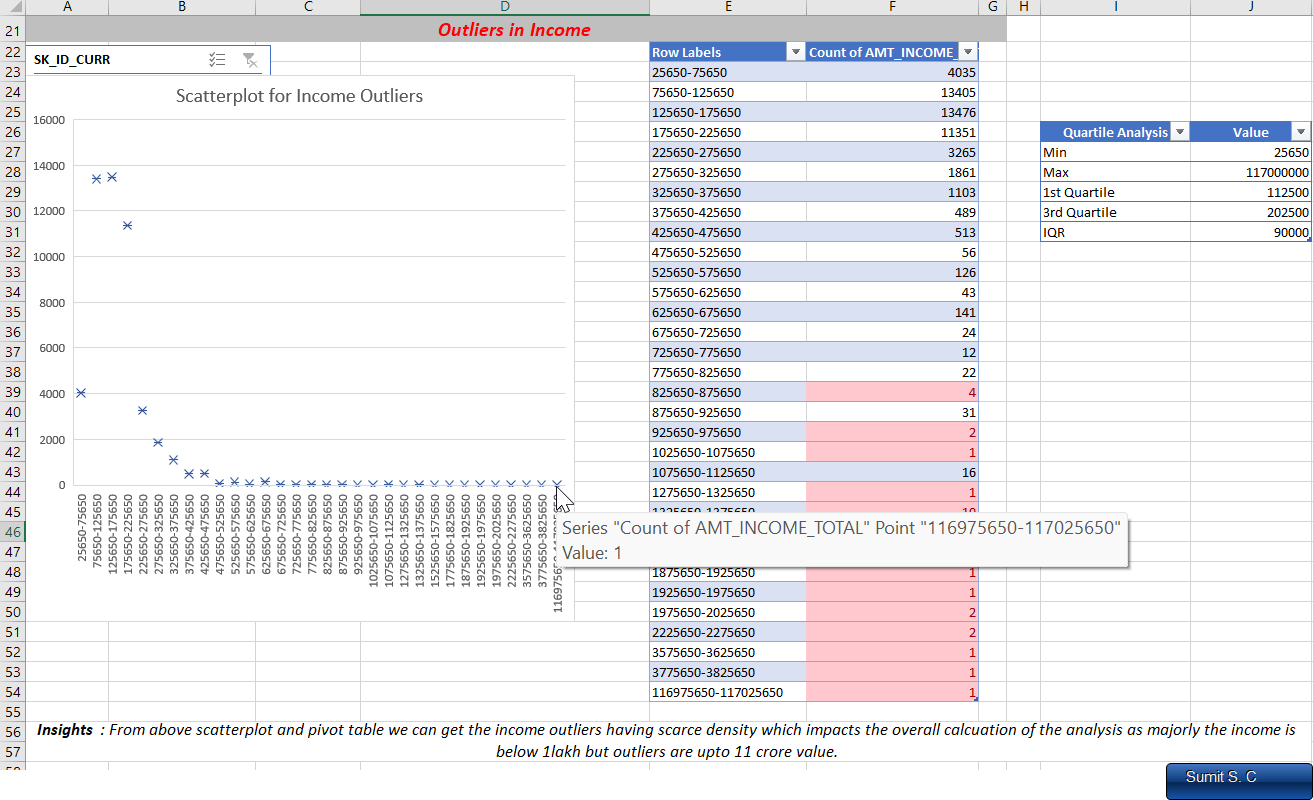
[Outlier](https://mitsus.life-is-pa.in/6gM0prrEF.png) 2 - In Credit\_AMT



[Outlier](https://mitsus.life-is-pa.in/6gM1BfXT1.png) 3 In Annuity Amount

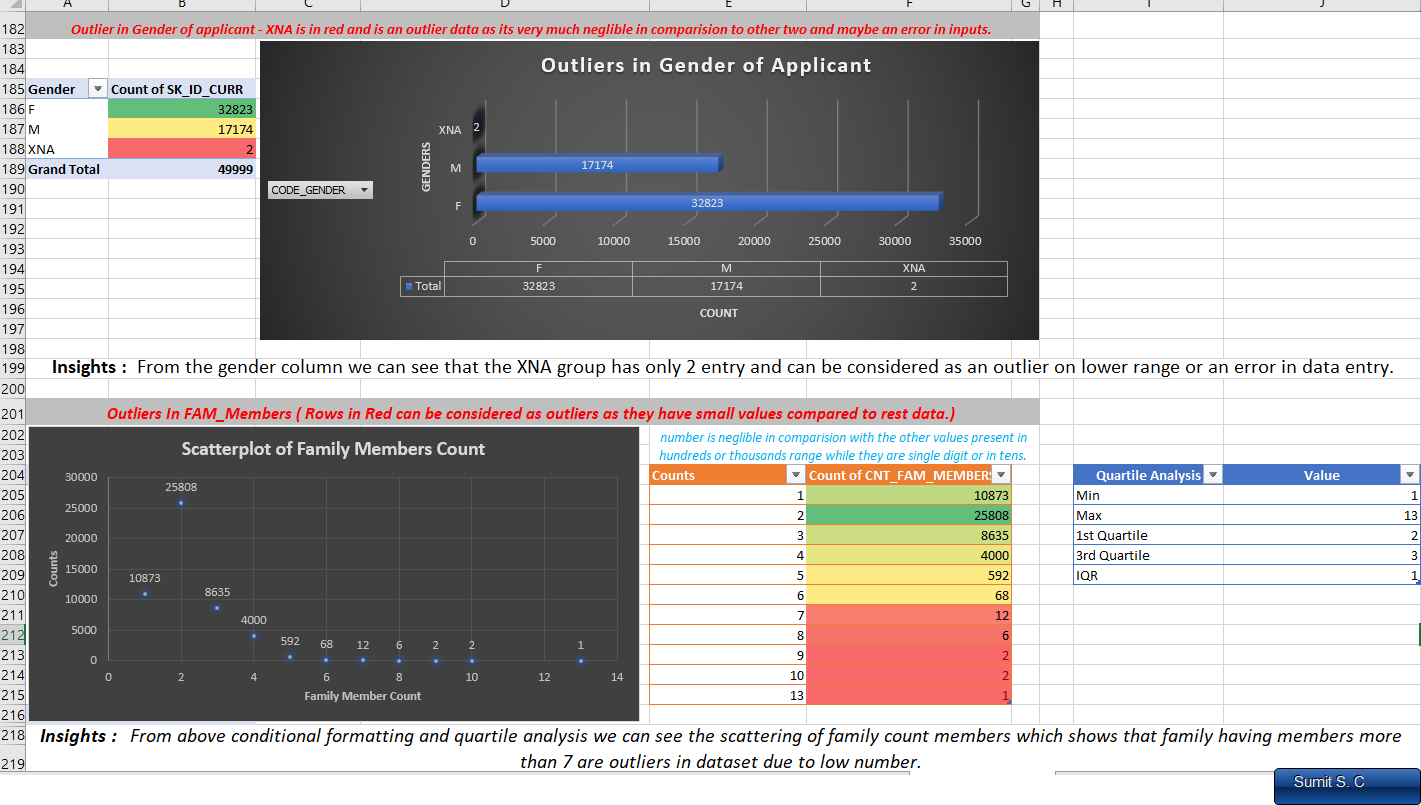


[Outlier](https://mitsus.life-is-pa.in/6gM2XJqKy.png) 4 - In Goods\_Price Column

[Outlier](https://mitsus.life-is-pa.in/6gM7VzXIY.png) 5 - Outliers In income



[Outlier](https://mitsus.life-is-pa.in/6gM8U4LED.png) 6 - Outliers in Tenure (In employment years and from application day)



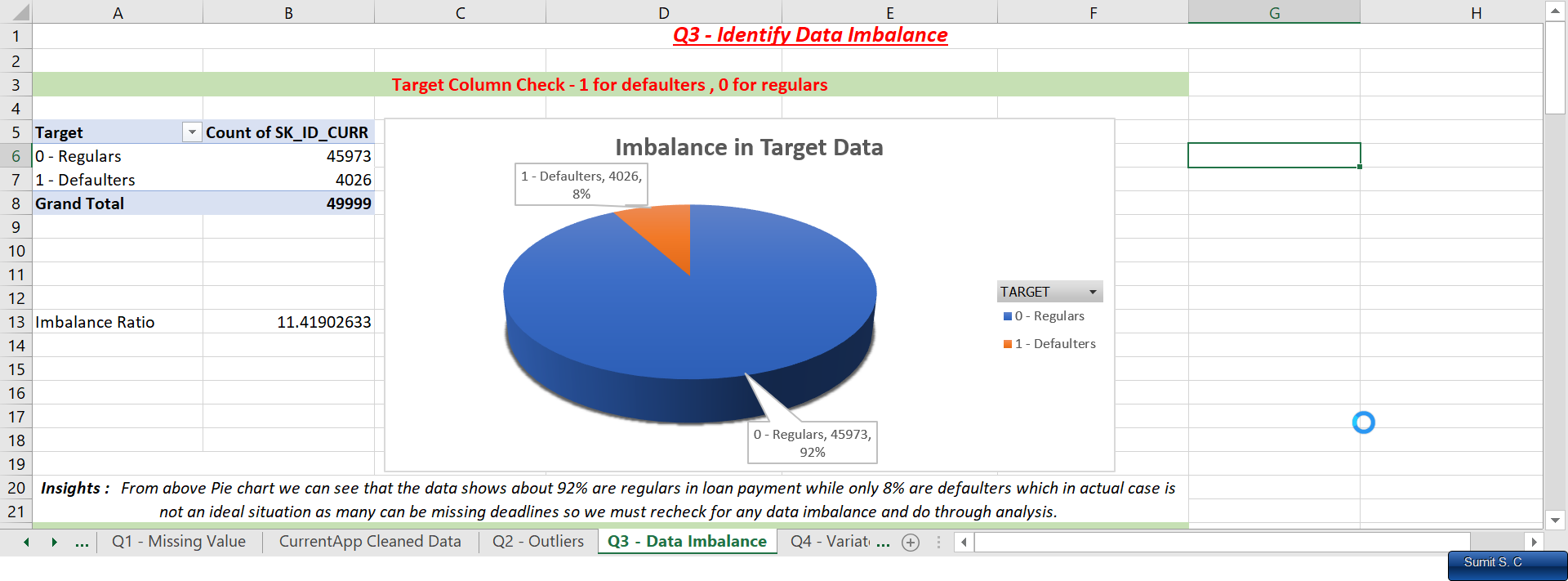
[Outlier](https://mitsus.life-is-pa.in/6gMdECAkl.png) 7 - Inside Gender distribution & Family members count

***Insights:***

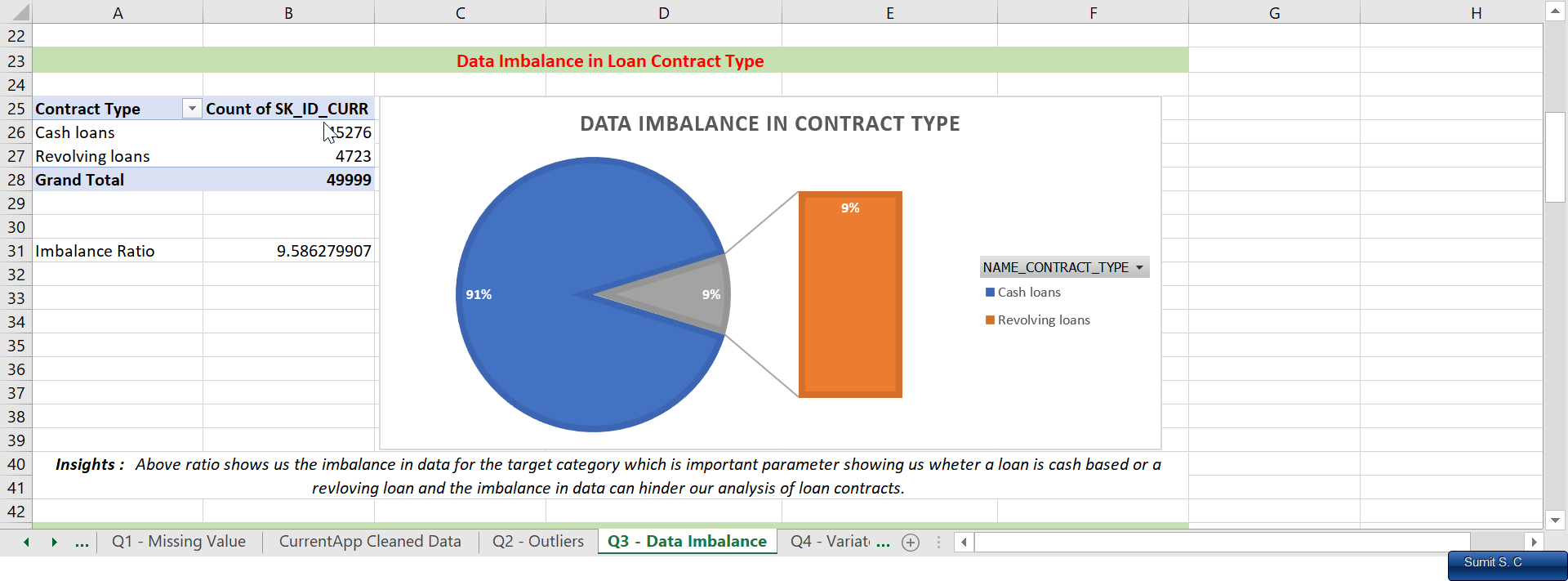
1. For every outlier I have tried giving the insight in the screenshot but to summarize the overall data we have found out that many outliers data are present in the dataset that needs to be checked or appropriately dealt with to make the analysis and overall future predictions about customer loan repayments and factors affecting their loans.
2. Certain factors like income and annuity shows us outlier points because 1 or 2 data entry shows a huge number of loan amount compared to rest of the data , same with the outliers present in tenure of loan and employment factor analysis would help us gain knowledge about trends in different category of profession and income range and corresponding habits etc so these outliers should be dealt accordingly to make analysis more righteous.
3. Few other outlier points are mentioned in screenshot and in excel report so you can find more insights from there.
4. Links
   * + Q2 Excel sheet (Individual file crashes the excel program so look into [main excel file](https://docs.google.com/spreadsheets/d/1-DphW5JffK6Y83PukCS6n775Td4rcS4P/edit?usp=drive_link&ouid=103173348283524847626&rtpof=true&sd=true) and go to Q2 tab)
     + Loom Video Explanation [(Identifying Outliers)](https://www.loom.com/share/96dc64ddbcc14528a138e757a44d2e7c?sid=62997e36-28e3-4f48-a870-83744e3bc403)

## C) 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.*

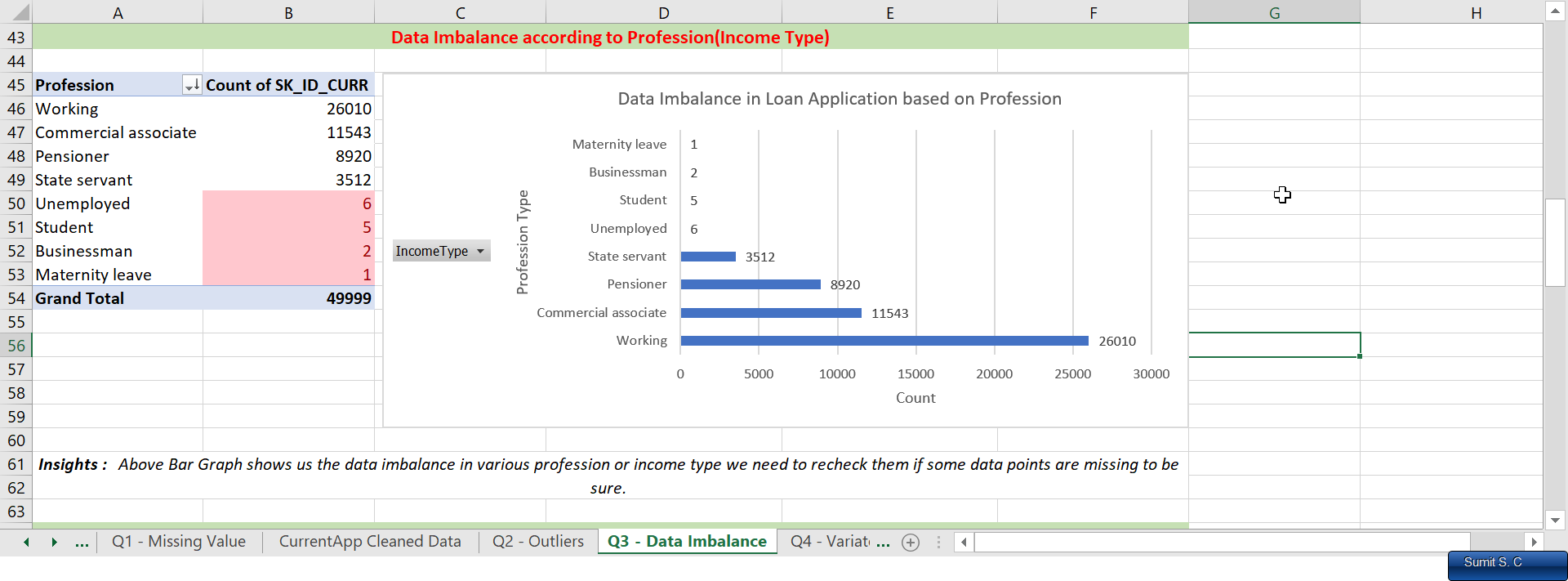
* *Task***:** Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.



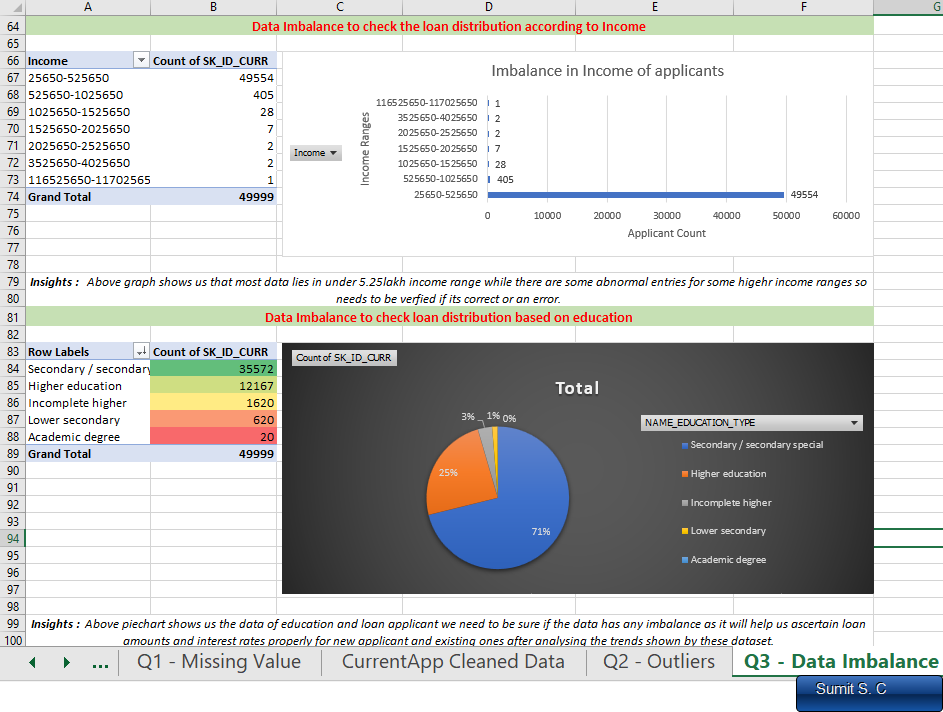
Data Imbalance 1 - [In target Column](https://mitsus.life-is-pa.in/6gMnitLY9.png)



Data Imbalance 2 - [In Target Category](https://mitsus.life-is-pa.in/6gMramOyN.png)



Data Imbalance 3 - In Income Type<https://mitsus.life-is-pa.in/6gMrRvcDW.png>



Data Imbalance 4 - [In education & Income range](https://mitsus.life-is-pa.in/6gMt6QSTC.png)

***Insights:***

1. From the above analysis which we did using pivot table and basic functions like **COUNT, COUNTIF & SUM** we got the data imbalance.
2. For better results and analysis we must recheck the data if the given data inputs are valid and are not having error will help us to get insights which will help us determine certain factors like age, profession, income and contract types which might hamper the payment or our overall analysis if some data are incorrect.
3. Data imbalance and its ratio helps us to prevent any unwanted calculation or change the analysis due to certain factors not being entered or taken correctly. It act as an indicator for certain values like yes or no , 1 – 0 columns which while entering or copying might have some human error and which might impact overall dataset working.
4. Files

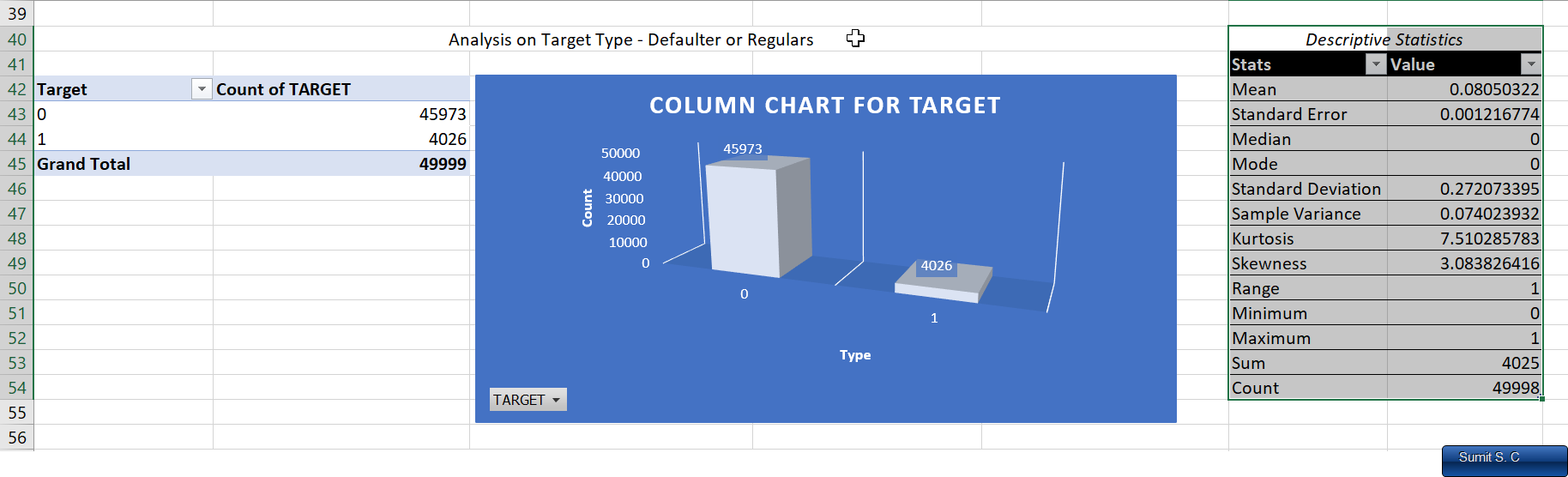
- Q3 [Excel File](https://docs.google.com/spreadsheets/d/1-95D5HtaqNnzDHDWj2YS-c0ScPCSZhQn/edit?usp=drive_link&ouid=103173348283524847626&rtpof=true&sd=true), [Gdocs File](https://docs.google.com/spreadsheets/d/1l73mjZwpfieK9EiE5Dbm9VGmj5iUFgpPa0CjK-9x6DM/edit?usp=drive_link)

- Loom Video PPT [Q3 – Identify the data imbalance](https://www.loom.com/share/09a09785538944e48a28c0b2be7a993f?sid=4ed46581-6576-4795-8b54-27c2f3cf0fea)

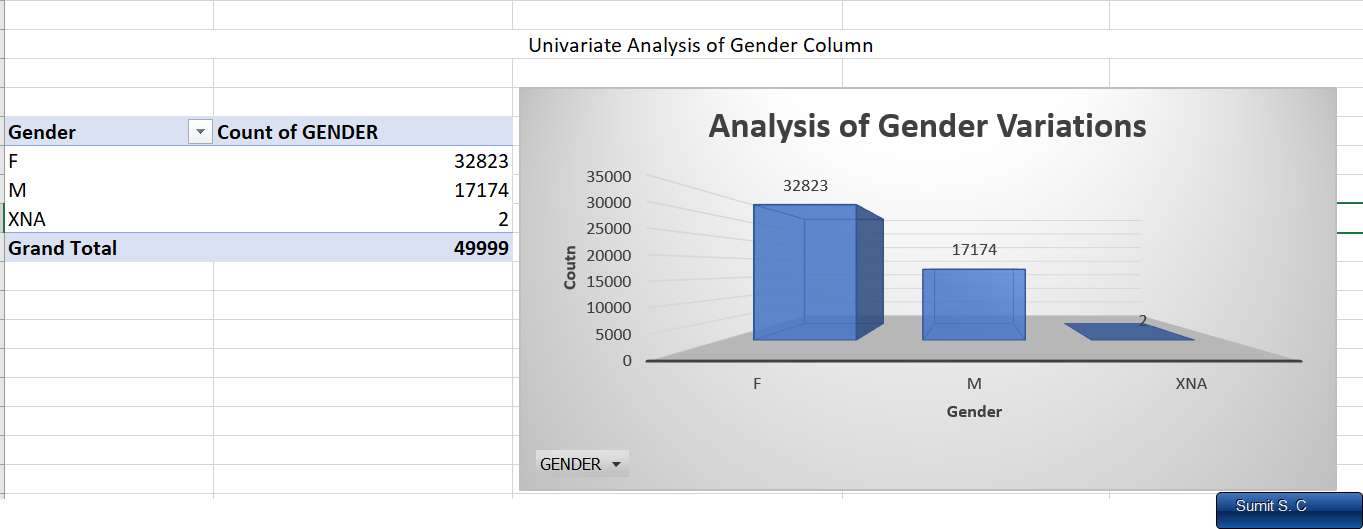
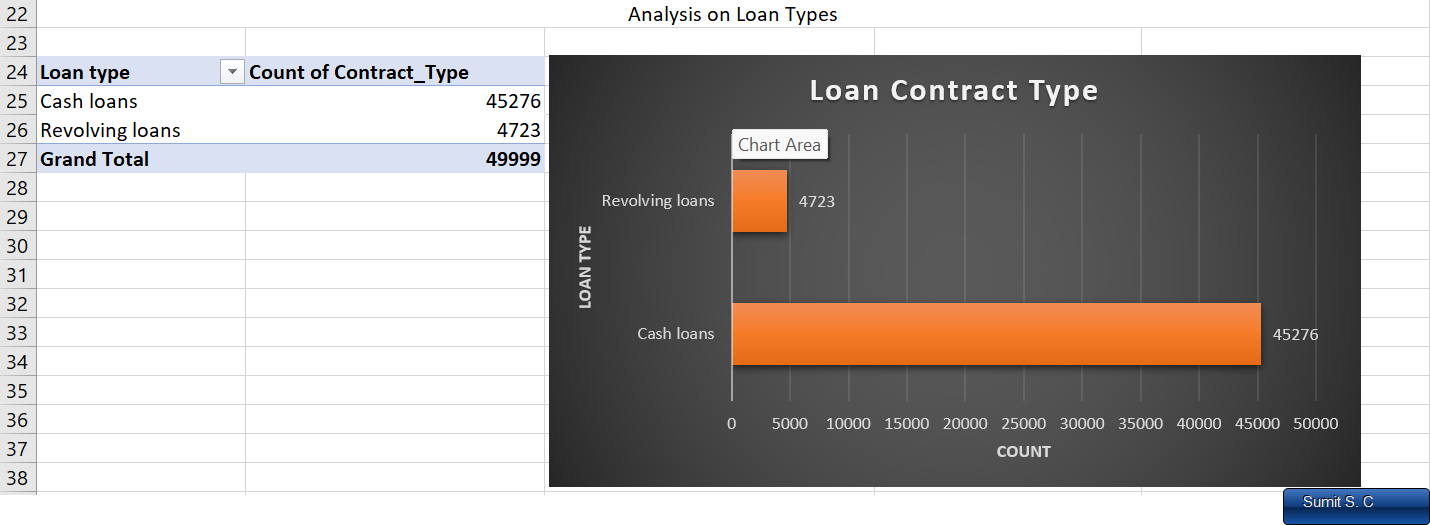
## D) Perform Univariate, Segmented Univariate, and Bivariate Analysis: *To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.*

* *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.

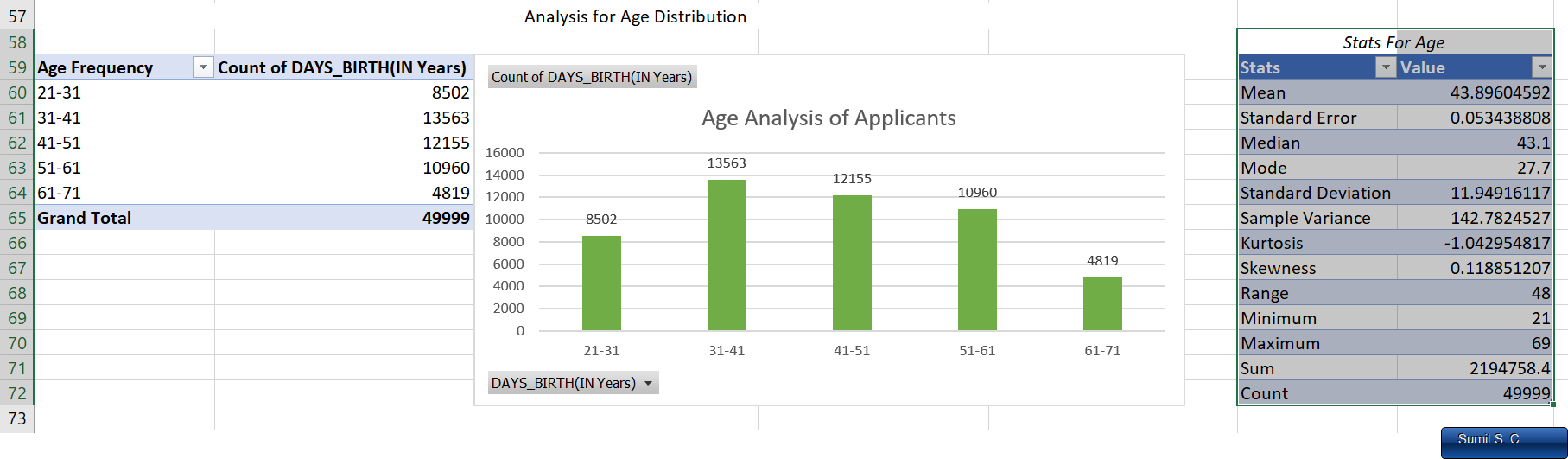
Univariate Analysis – When we examine and analyse a single variable in isolation and understand the distribution summary stats using descriptive stats and try to gain information its called as univariate analysis.



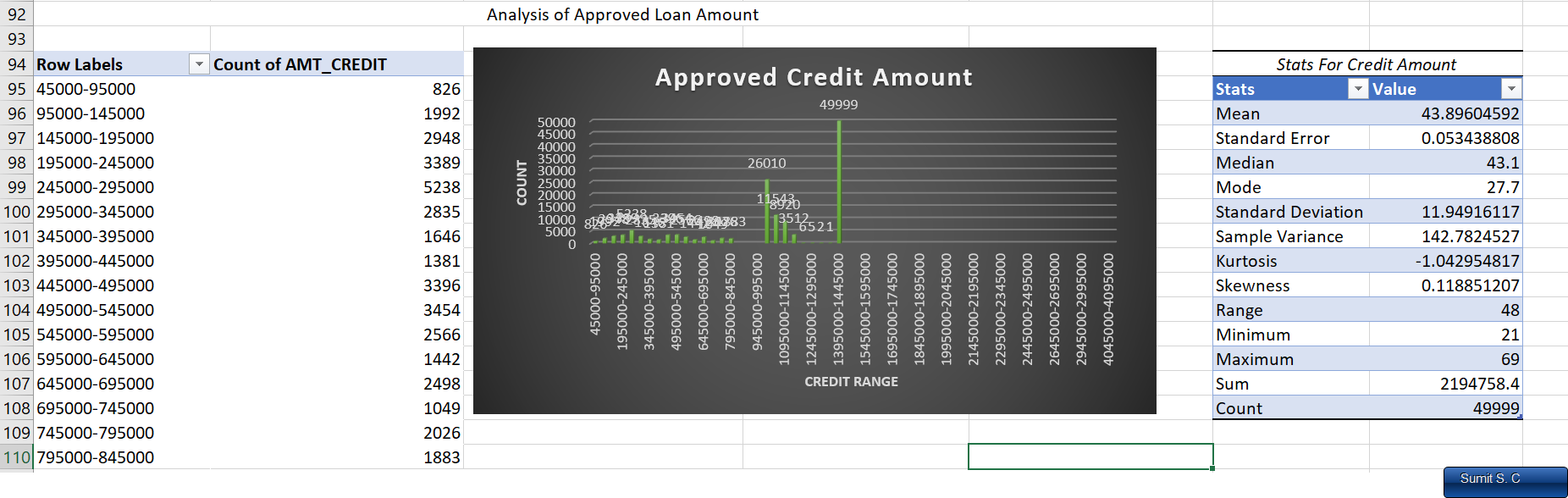
Univariate 1 - [On Target](https://mitsus.life-is-pa.in/6gMRkufHR.png)

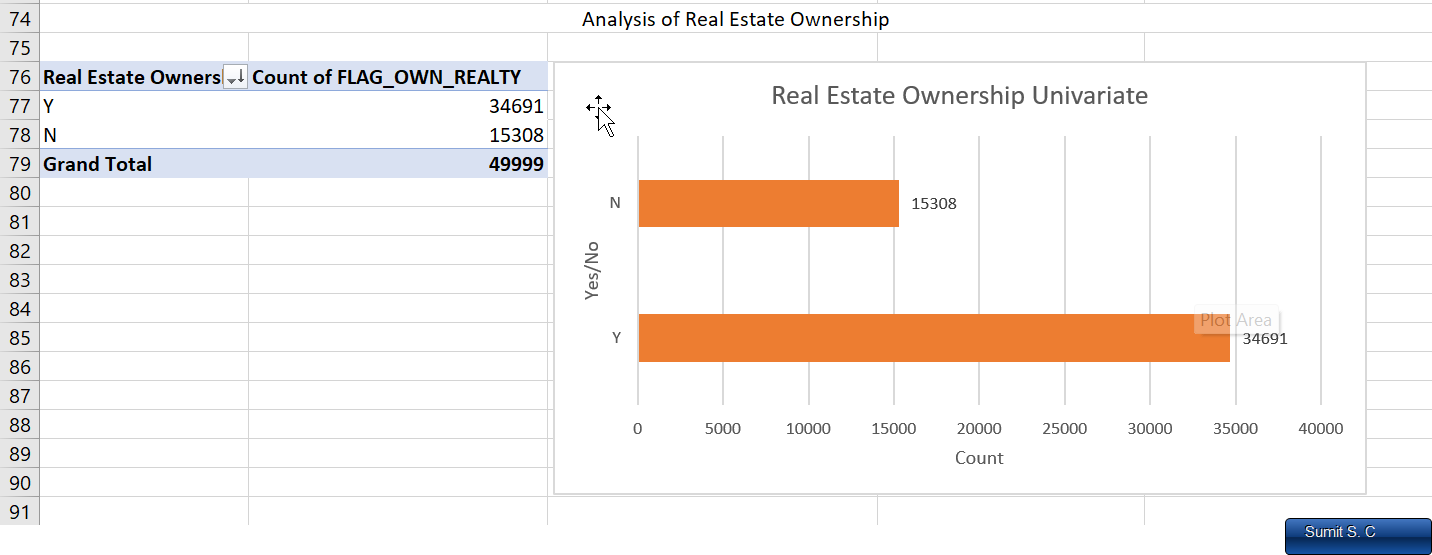
Univariate 2 [On Gender](https://mitsus.life-is-pa.in/6gMMKkCU2.png) Univariate 3 [On Loan type](https://mitsus.life-is-pa.in/6gMRIIHoY.png)



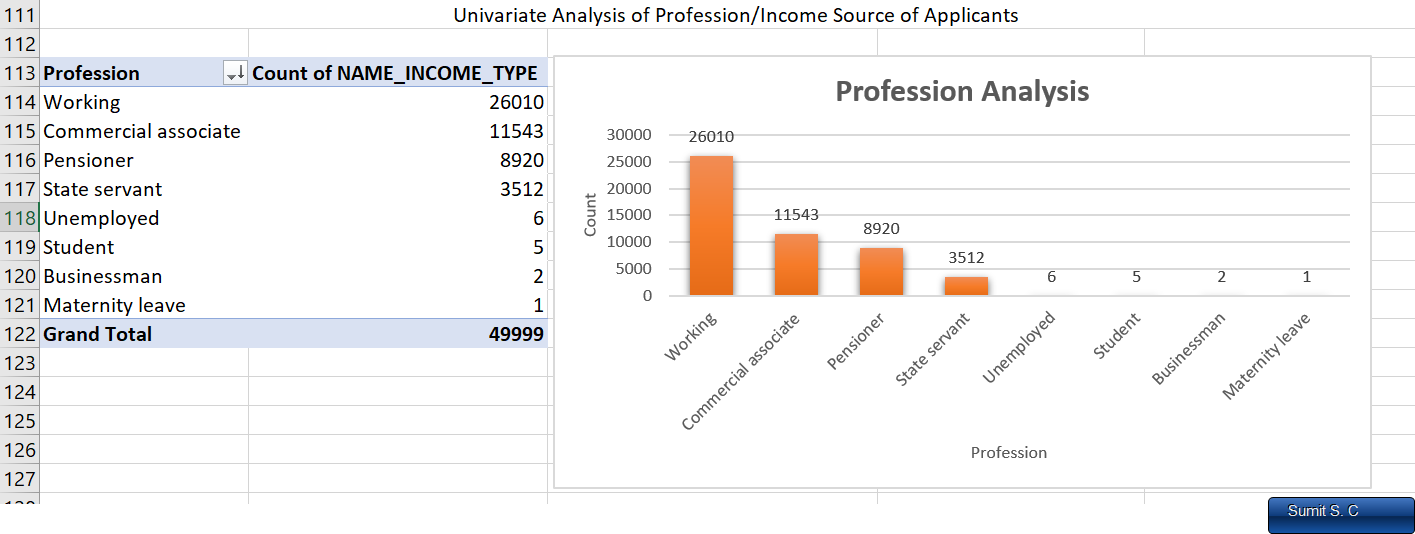
Univariate 3 [On Age](https://mitsus.life-is-pa.in/6gMTjTQst.png)



Univariate 4 [On Loan Amount](https://mitsus.life-is-pa.in/6gMTHAJD4.png)

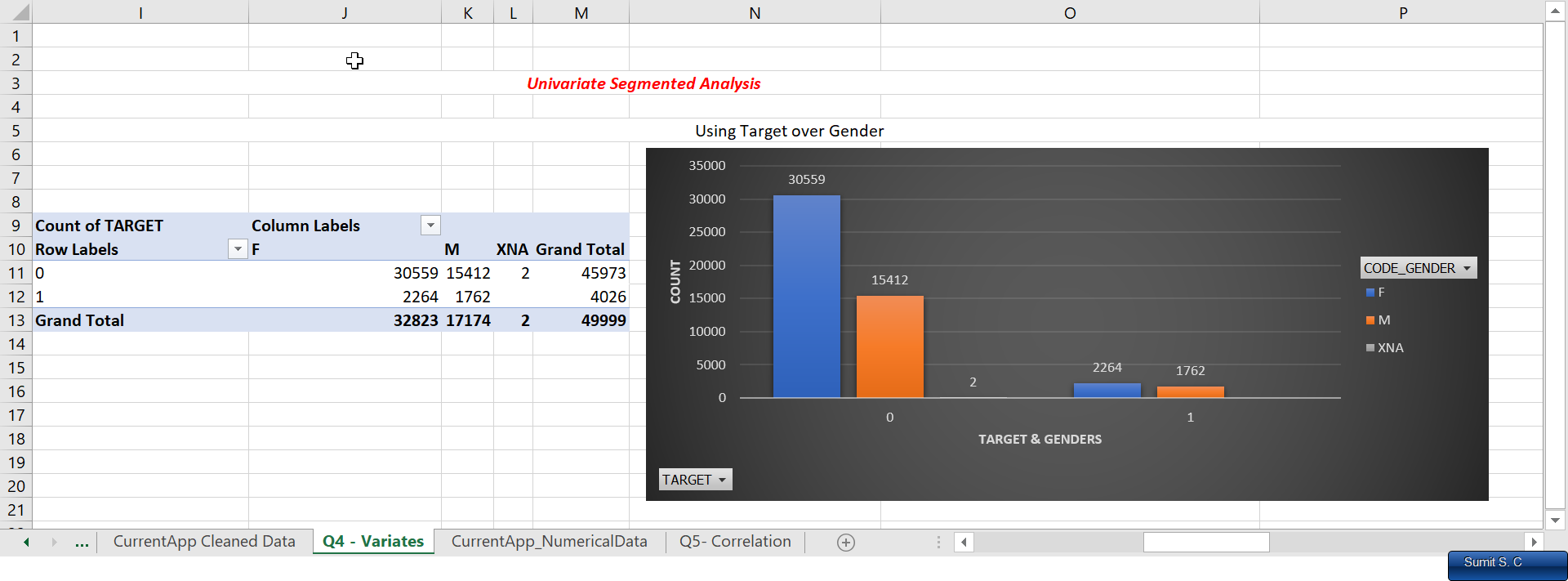


Univariate 5 [on estate ownership](https://mitsus.life-is-pa.in/6gMV0osQ3.png)

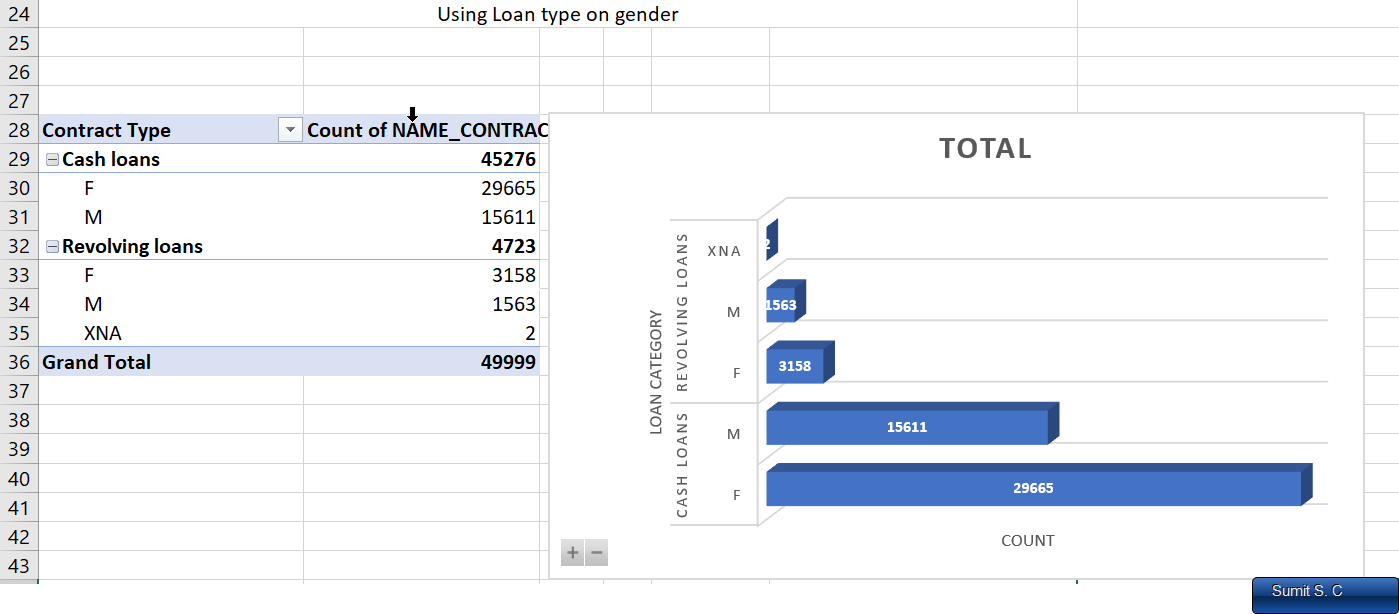


Univariate 6 [On Income Type](https://mitsus.life-is-pa.in/6gMVl9QzS.png)

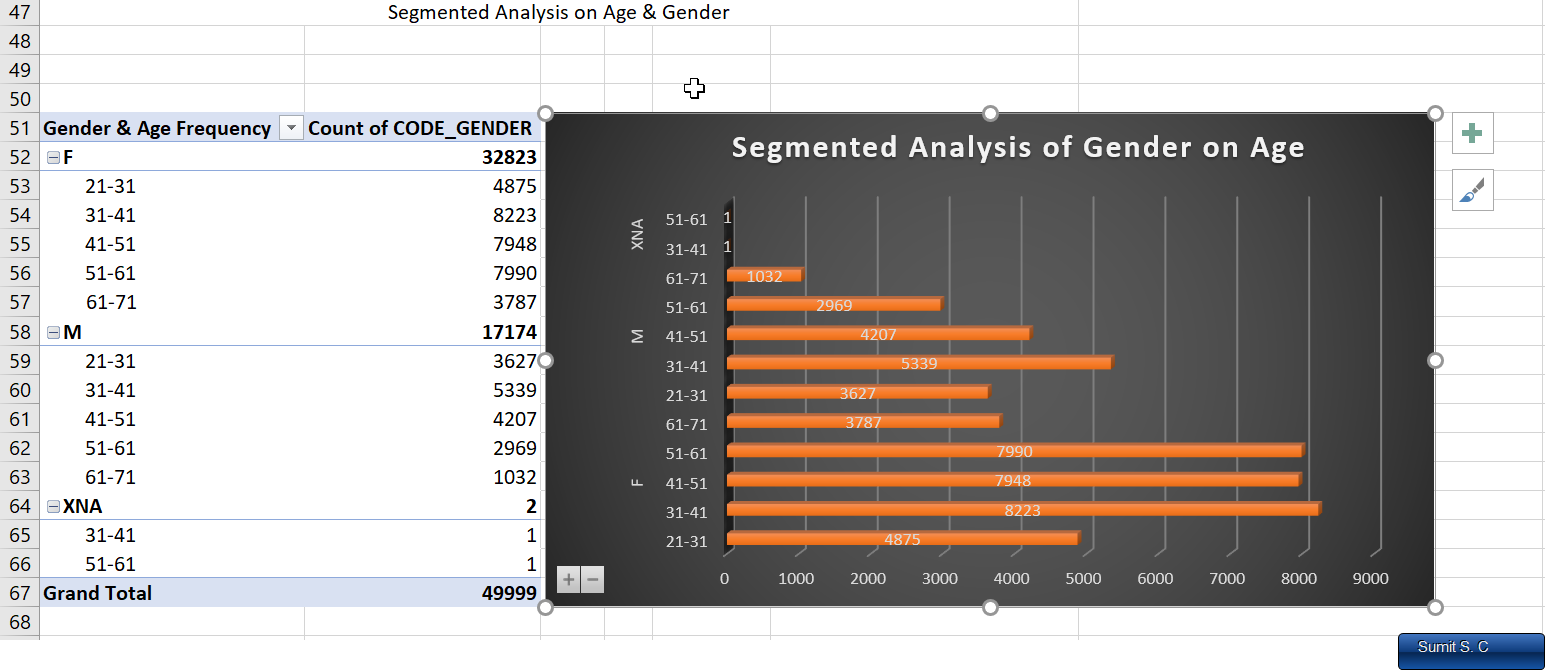
**Segmented Univariate Analysis** – When we examine and analyse a single variable with different segments or subgroups defined by another variable and understand the distribution summary stats using descriptive stats and try to gain information its called as segmented univariate analysis.



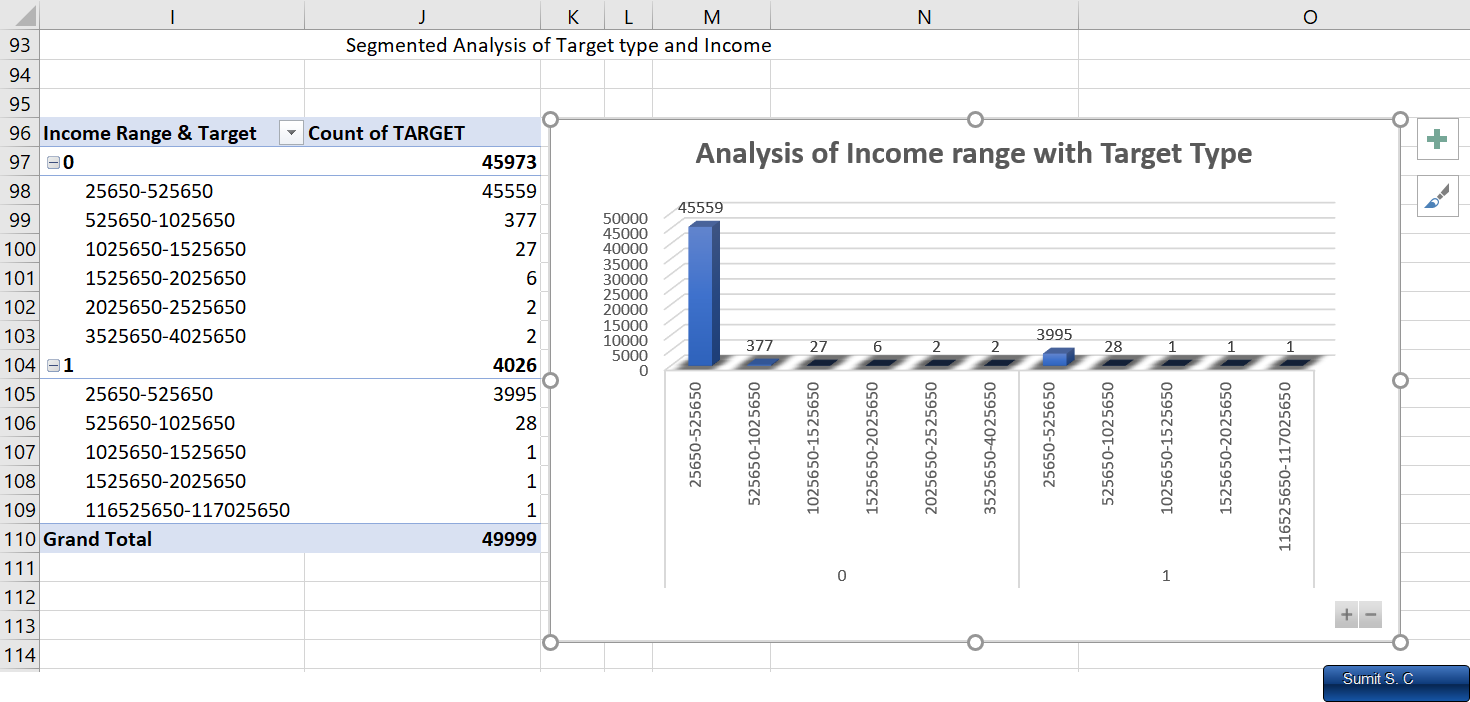
Univariate Segmented 1 - [On Target & Gender](https://mitsus.life-is-pa.in/6gMXQwjeY.png)



Univariate Segmented 2 - [On Loan type & gender](https://mitsus.life-is-pa.in/6gMYddcpz.png)

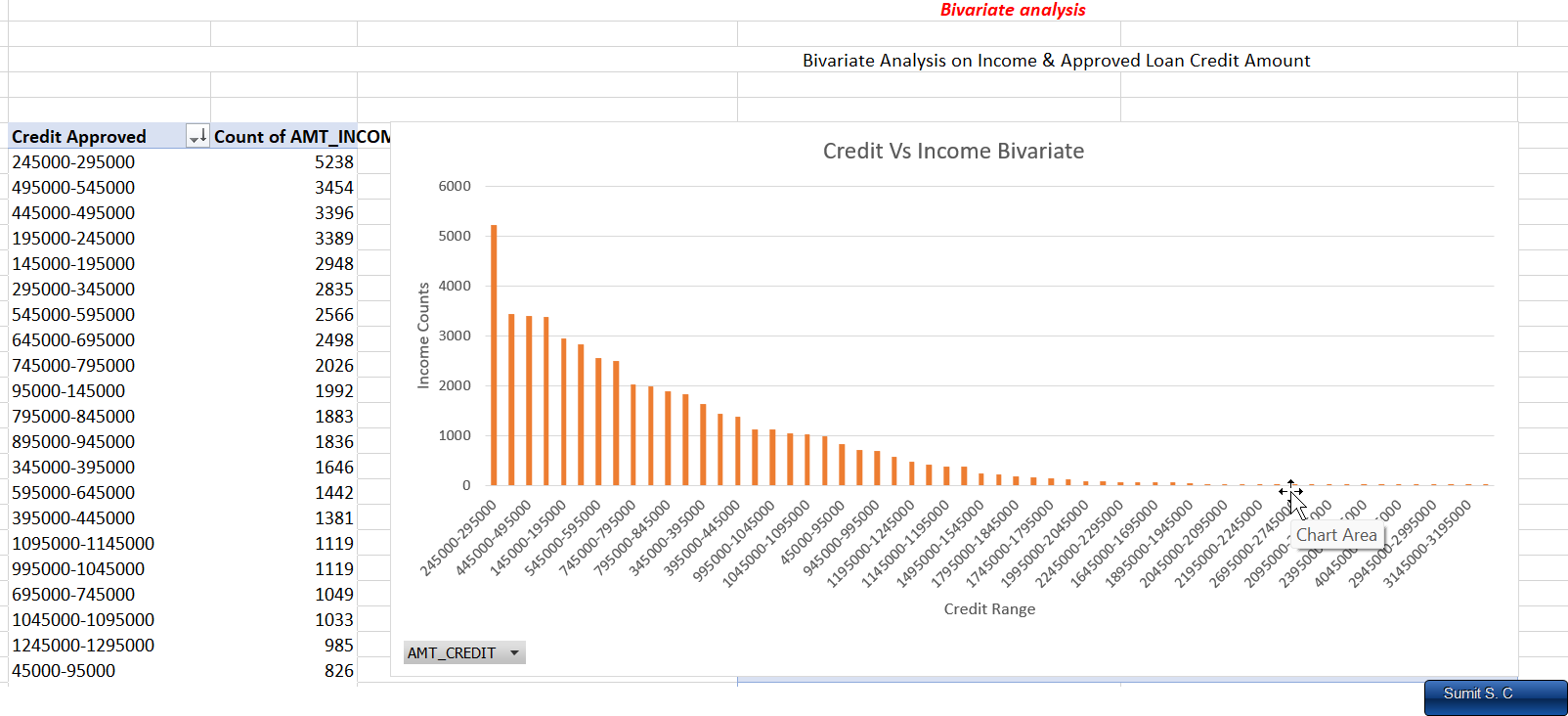


Univariate Segmented 3 - [On Gender & Age](https://mitsus.life-is-pa.in/6gMYEDWiJ.png)

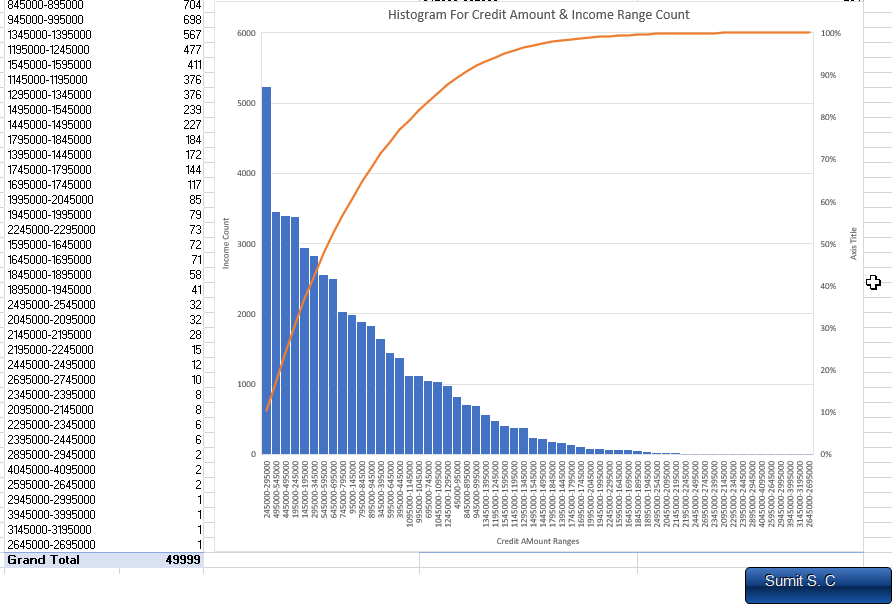


Univariate Segmented 4 - [On target & Income Range](https://mitsus.life-is-pa.in/6gMYWd0Ru.png)

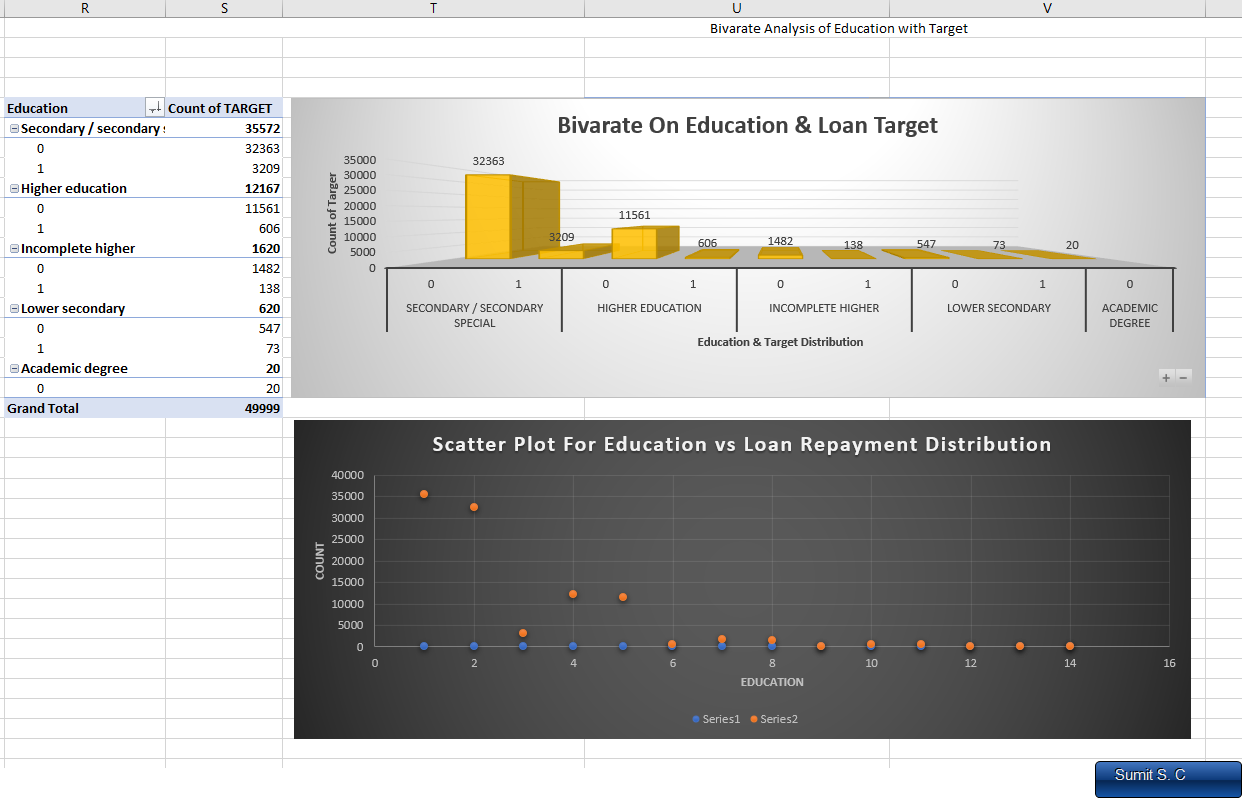
**Bivariate Analysis** – When we examine and analyse two variables simultaneously and understand the distribution summary stats using descriptive stats and try to gain information its called as bivariate analysis. The purpose is to identify the relationship between 2 variables to understand the patterns that change sthe data corresponding to other data in comparisons.



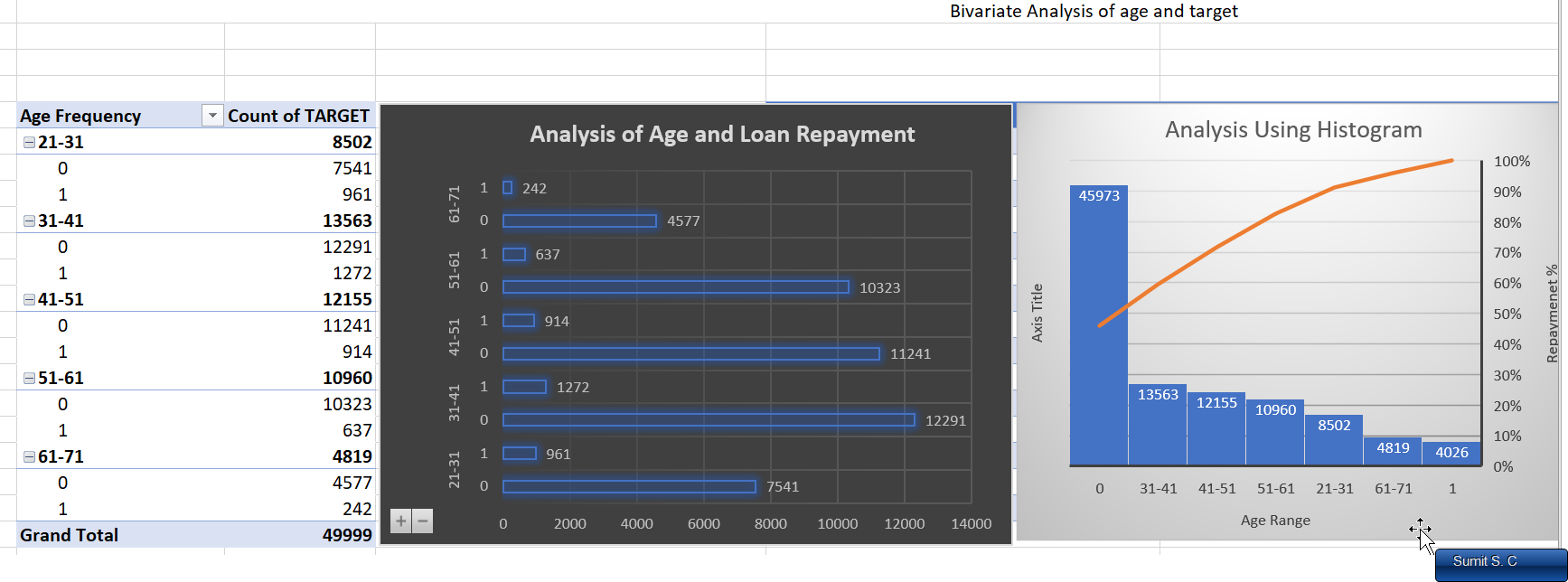
Bivariate 1 [Column Bar GraphOn Credit amount & Credit approved](https://mitsus.life-is-pa.in/6gM_uJqfs.png)



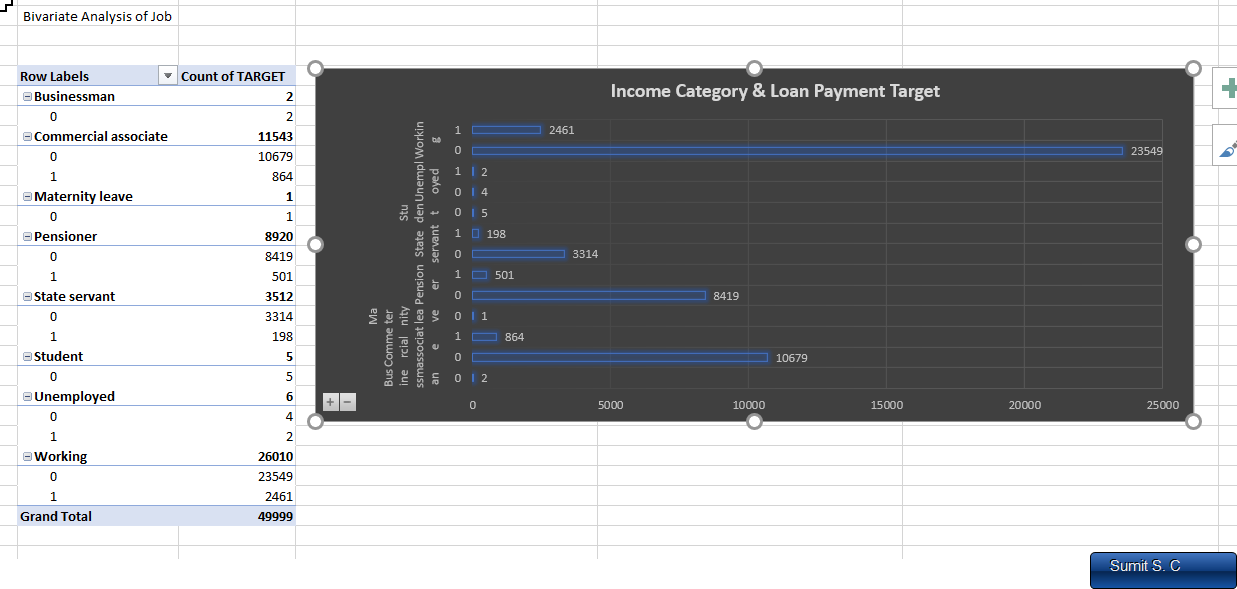
Bivariate 2 [Histogram on above set](https://mitsus.life-is-pa.in/6gM_RTLTy.png)



Bivariate 3 - [On education and target counts (Column & Scatterplot distribution)](https://mitsus.life-is-pa.in/6gN0nUKS2.png)



Bivariate 4 - [On age & Target ( Histogram & Bar plot)](https://mitsus.life-is-pa.in/6gN0PktLc.png)



Bivariate 5 - [On Profession & Target type (Bar Plotting)](https://mitsus.life-is-pa.in/6gN1K6rYj.png)

*Note –* During taking the screenshots I missed the Descriptive stats of Bivariate & Segmented analysis for numerical data which are calculated in the excel sheet.

***Insights:***

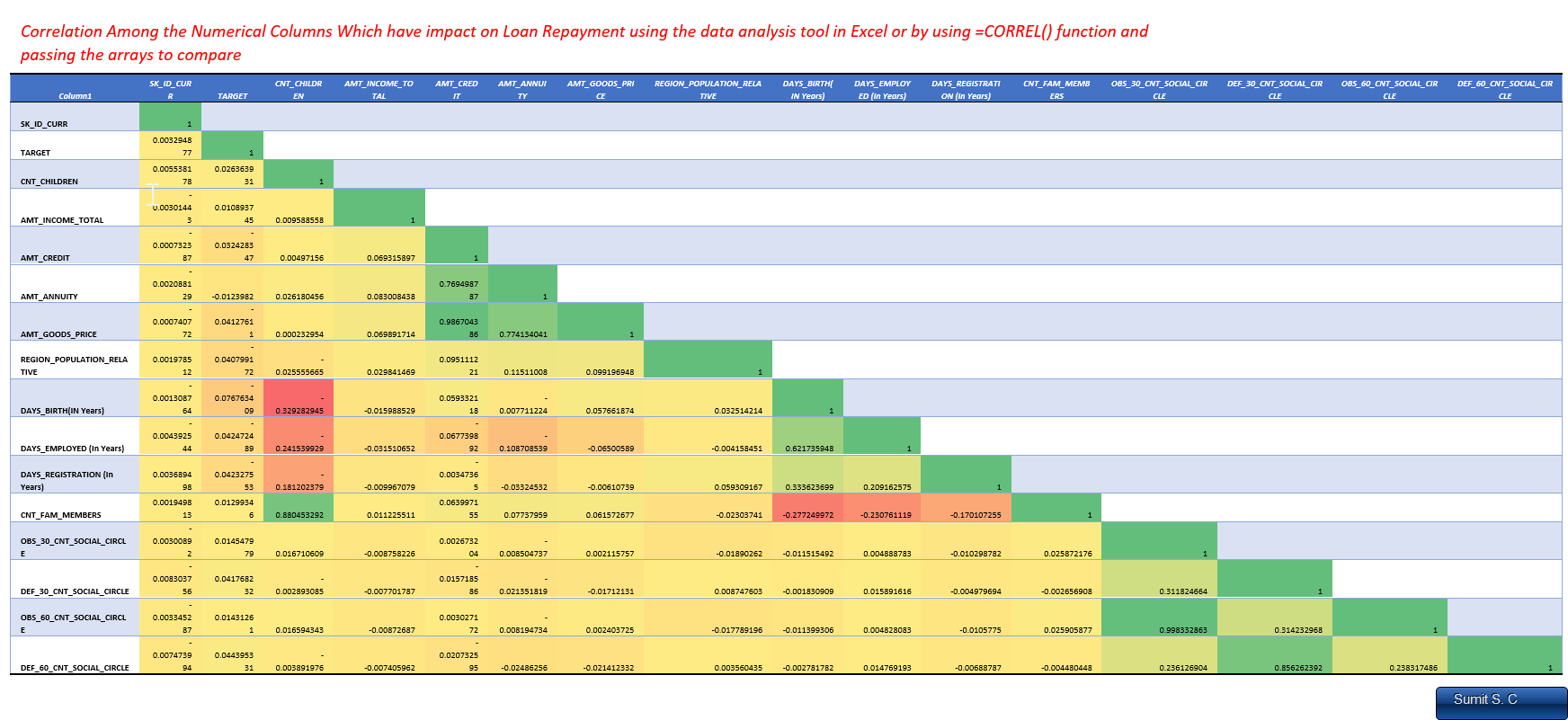
1. In the univariate analysis we have selected a single column and checked its decsriptive stats and plotted a graph to see its distribution.
2. For univariate analysis we basically take a single variable into consideration and check the other variables that impacts each other using scatter and other plots for reference but all are individuals.
3. I have placed few univariate analyses above for reference.
4. In Segmented univariate analysis we basically divide the single variable and further deeply analysis it says on target type on gender roles or say by loan type and drill the column for more better insights on the column.
5. In segmented analysis we basically get more deeper insights of the category like just gender column gets divided into male & female for analysis , loan type gets into cash or revolving thus giving us insight about loan repayment and categories more.
6. In Bivariate analysis we compare two variables and compare them on a single feature to see the impacts of the comparison on particular analysis.
7. Say in our above analysis we prepared histogram and scatterplots/Boxplots to show variations that arises when we compare different entity and its overall impact (here on loan repayment).
8. For more insights one can go through the excel files if I have missed something in screenshot or in report because these questions is a long one and writing everything will not be feasible.
9. Links

- Q4 [Excel Link](https://docs.google.com/spreadsheets/d/1-AJkZq0nkVR65sDWys5iEgxLvYg88aH7/edit?usp=drive_link&ouid=103173348283524847626&rtpof=true&sd=true), [Gdocs Links](https://docs.google.com/spreadsheets/d/10QF7rI-uIgZW8ygrighYHdAjgN0bckuNCjm-8I411u4/edit?usp=drive_link)

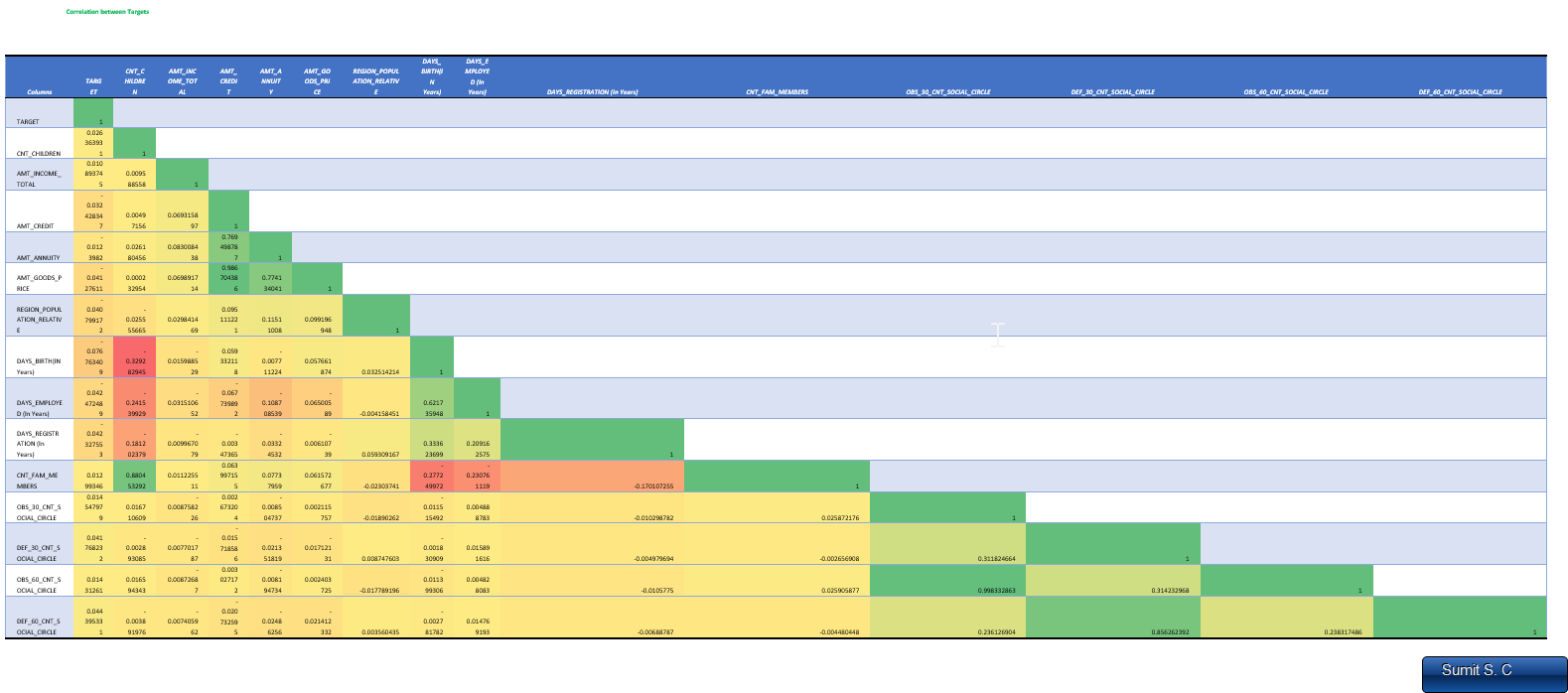
- Loom Video PPT [Q4 – Variate analysis](https://www.loom.com/share/b8865a6932a04d079f6d297c640bfaad?sid=4434e9f7-ebaf-4cb2-9363-5f2435e054cf)

## 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.*

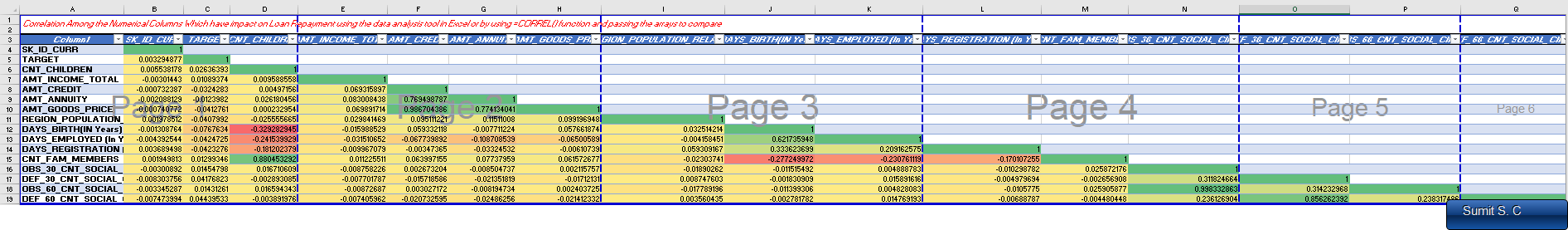
* *Task***:** Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.



Correlation of factors 1 [HD Image](https://mitsus.life-is-pa.in/6gMAzfyem.png) (Considering ID)



Correlation of factors 2 [HD Image](https://mitsus.life-is-pa.in/6gMHBiAKY.png) (Removing ID)



***Insights***:

1. We have used the Correlation analysis present in data > data analysis tab which helps us to choose the correlation factor for the selected column range or called array of numbers. Correlations are basically done by taking the 1st column into account and then identifying the changes in other columns with respect to 1st one. For our analysis we have taking target type for the correlation if we need to do more analysis we can select the specific columns such that we can make the correlation between them.
2. In correlation comparison we do the comparison of each column with oneself and with other column in question and thus a matrix is formed of the same length as the column numbers and the diagonal is equal to 1 (since relation of same column comparison.)
3. The same result can be achieved with making column and then applying the **CORREL** function on the columns to compare one by one which will be time consuming process. A histogram or heatmap like plot is plotted to get the relevance of data. In Excel the lower triangle gets filled while upper remains blank as it’s the transposed values for columns if we make it in python or BI tools they might fill the same on other sides.
4. Do note that for correlation we need numbers so we need to remove any categorical value columns first. So, I cleaned up the dataset further by removing the categorical column and the numerical column which does not directly impact on the loan behaviors.
5. From the above analysis we have found out about the top 15 factors that might impact the loan defaults as these factors play a role in individuals spending and savings habit etc.
6. Links
   * + Q5 [Excel File](https://docs.google.com/spreadsheets/d/1-7byFwmpEvL88klrtQMPhpoztFWHEpUw/edit?usp=drive_link) , [Gdocs Files](https://docs.google.com/spreadsheets/d/1j4PtdN8RZbIIp0u4gO0Wyq6fbswbA9Y0r7kD5DmSkpg/edit?usp=drive_link)
     + Loom Video PPT [Q5 – Find the Correlation of scenario](https://www.loom.com/share/b8865a6932a04d079f6d297c640bfaad?sid=39f9da59-3a82-4d72-858c-147c10a56280)

# Important Links :

[Drive Folder Link](https://drive.google.com/drive/folders/1eyK3giYVshfAOk0skO6fRiwGfh5QLEg5)

[Individual Excel Sheets](https://drive.google.com/drive/folders/1IhVg5n_djCNxr2yJaPi_VHMbr7gcdShS) (Individual Questions File)

[Final Excel sheet](https://drive.google.com/drive/folders/1GzxZegYTGY197iZqy0Iqxgz_cutDY9g1) (cleaned dataset & all questions also present on drive folder page)

Word File Link & Pdf File (Will Be in the [drive folder](https://drive.google.com/drive/folders/1U6AMwzb1OyZAee9u5rDsA4A1AcXzOjSD) – can’t add before I upload the file)

[Video Presentation(Loom Folder 6 Videos)](https://loom.com/share/folder/6d8bf79fb6874547bb8b7839e0d319c6)

[Zip Link](https://drive.google.com/file/d/1-Wabffd0O22ycK7s-_qrE0qN7Hlqsoyq/view?usp=drive_link) (Download these to extract the pdf and excel sheet files)

**Thank You**