SUMIT SHAMLAL CHAURE

Bank Loan Case Study

Trainity Project 6 (Final Project -2)

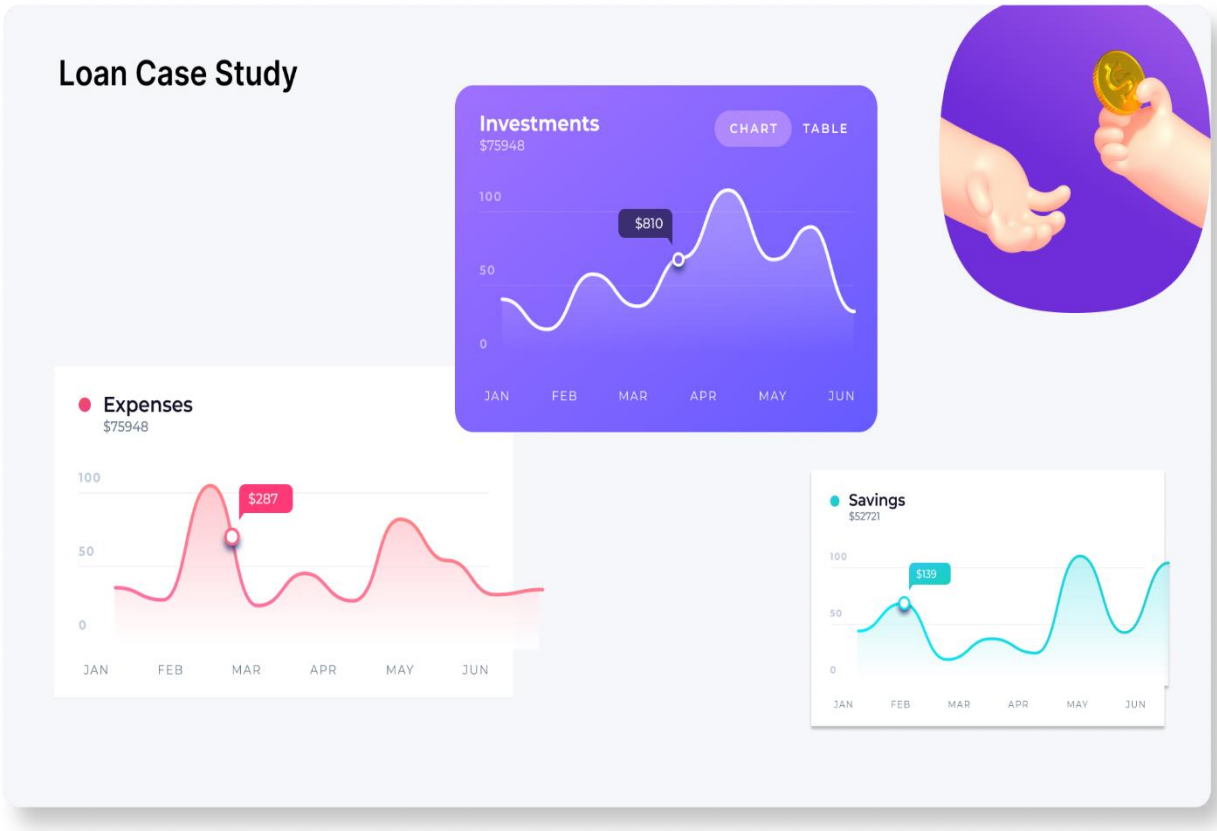


Figure 1 - https://trainity.link/data/project06

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) (Drive Folder)

INDEX

Contents

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.

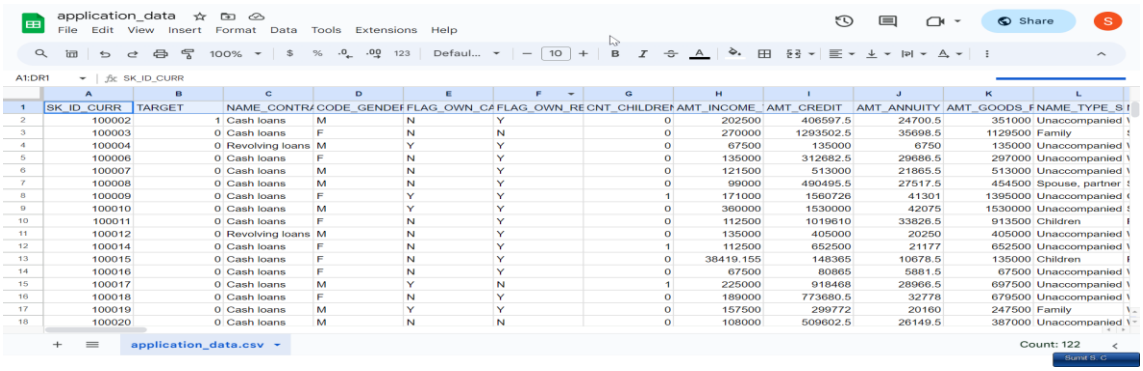
Drive – To upload all the essential files attached in the report for reference & report upload.

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

4) Insights:

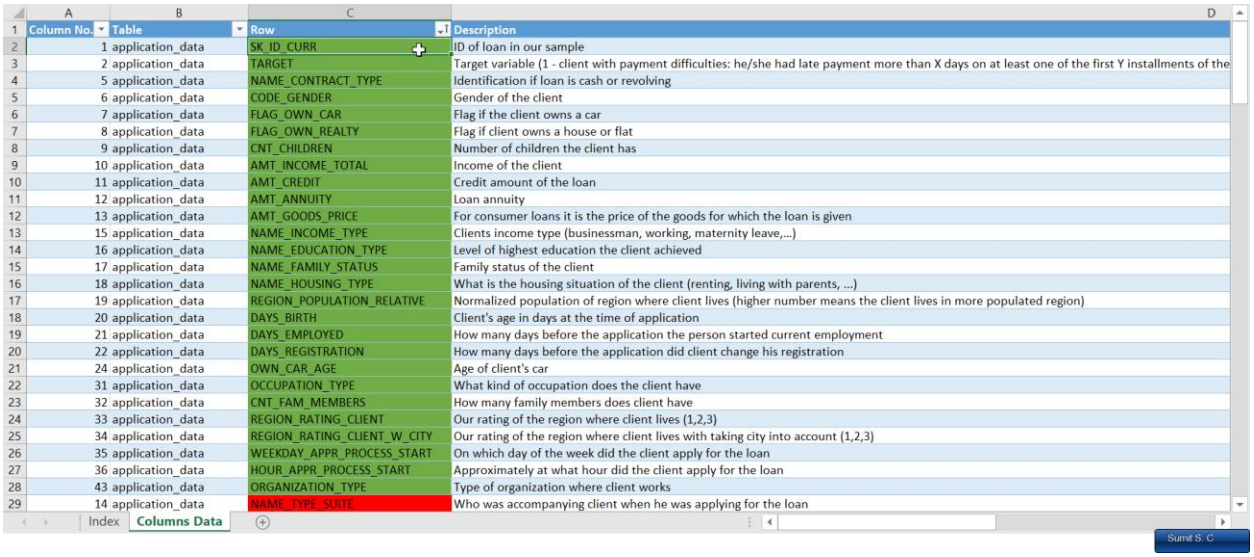
I have done the analysis on **application_data.csv** file which pertains to the current application details of the clients.

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.



	SK_ID_CURR	TARGET	NAME_CONTRA	CODE_GENDE	FLAG_OWN_CA	FLAG_OWN_RE	CNT_CHILDRE	AMT_INCOME_	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_F	NAME_TYPE_S
2	100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied
3	100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family
4	100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied
5	100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied
6	100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied
7	100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner
8	100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied
9	100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied
10	100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children
11	100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied
12	100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied
13	100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children
14	100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied
15	100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied
16	100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied
17	100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family
18	100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied

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.



Column No.	Table	Row	Description
1	application_data	SK_ID_CURR	ID of loan in our sample
2	application_data	TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the
5	application_data	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
6	application_data	CODE_GENDER	Gender of the client
7	application_data	FLAG_OWN_CAR	Flag if the client owns a car
8	application_data	FLAG_OWN_REALTY	Flag if client owns a house or flat
9	application_data	CNT_CHILDREN	Number of children the client has
10	application_data	AMT_INCOME_TOTAL	Income of the client
11	application_data	AMT_CREDIT	Credit amount of the loan
12	application_data	AMT_ANNUITY	Loan annuity
13	application_data	AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
15	application_data	NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,...)
16	application_data	NAME_EDUCATION_TYPE	Level of highest education the client achieved
17	application_data	NAME_FAMILY_STATUS	Family status of the client
18	application_data	NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)
19	application_data	REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more populated region)
20	application_data	DAYS_BIRTH	Client's age in days at the time of application
21	application_data	DAYS_EMPLOYED	How many days before the application the person started current employment
22	application_data	DAYS_REGISTRATION	How many days before the application did client change his registration
24	application_data	OWN_CAR_AGE	Age of client's car
31	application_data	OCCUPATION_TYPE	What kind of occupation does the client have
32	application_data	CNT_FAM_MEMBERS	How many family members does client have
33	application_data	REGION_RATING_CLIENT	Our rating of the region where client lives (1,2,3)
34	application_data	REGION_RATING_CLIENT_W_CITY	Our rating of the region where client lives with taking city into account (1,2,3)
35	application_data	WEEKDAY_APPR_PROCESS_START	On which day of the week did the client apply for the loan
36	application_data	HOUR_APPR_PROCESS_START	Approximately at what hour did the client apply for the loan
43	application_data	ORGANIZATION_TYPE	Type of organization where client works
14	application_data	NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan

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 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.

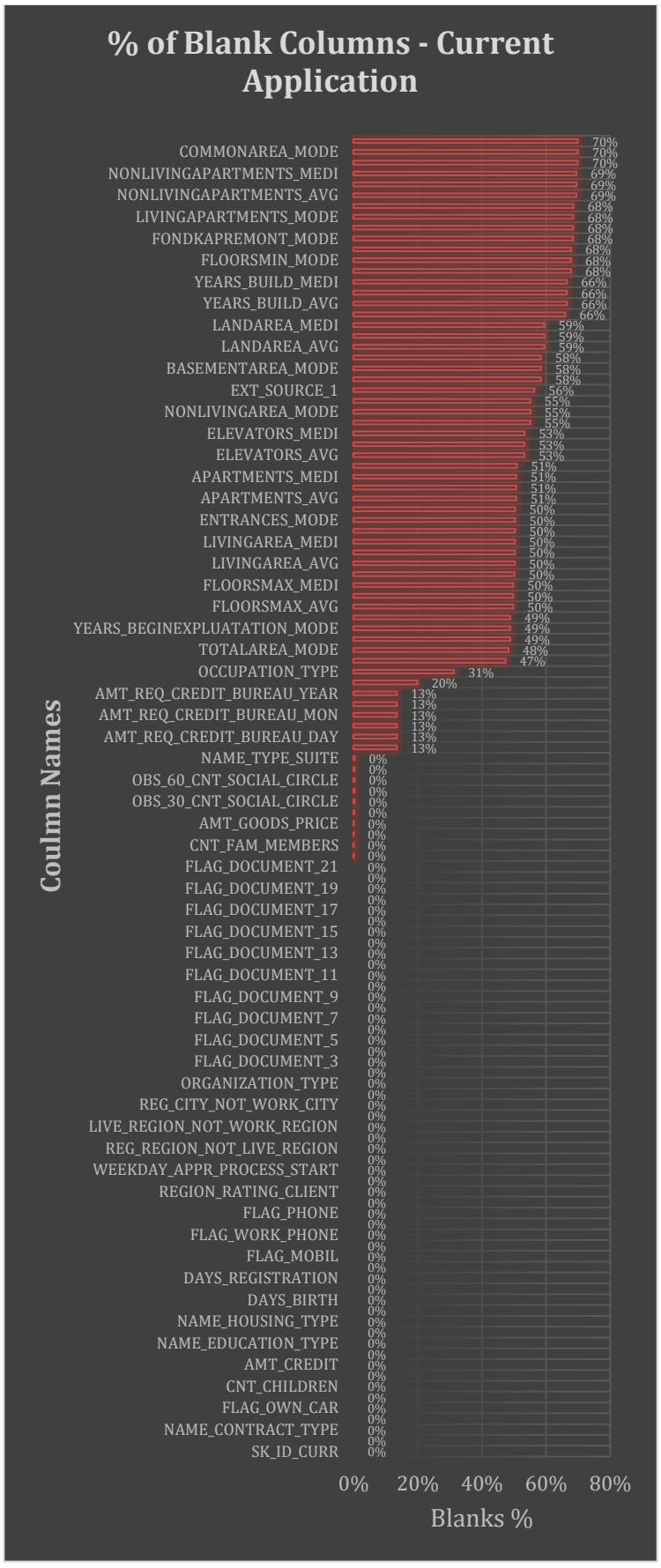
C93 | HOUSETYPE_MODE

	A	B	C	D
1	Column No	Table	Row	Description
2	1	application_data	SK_ID_CURR	ID of loan in our sample
3	2	application_data	TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the
4	5	application_data	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
5	6	application_data	CODE_GENDER	Gender of the client
6	7	application_data	FLAG_OWN_CAR	Flag if the client owns a car
7	8	application_data	FLAG_OWN_REALTY	Flag if client owns a house or flat
8	9	application_data	CNT_CHILDREN	Number of children the client has
9	10	application_data	AMT_INCOME_TOTAL	Income of the client
10	11	application_data	AMT_CREDIT	Credit amount of the loan
11	12	application_data	AMT_ANNUITY	Loan annuity
12	13	application_data	AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
13	15	application_data	NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,...)
14	16	application_data	NAME_EDUCATION_TYPE	Level of highest education the client achieved
15	17	application_data	NAME_FAMILY_STATUS	Family status of the client
16	18	application_data	NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)
17	19	application_data	REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more populated region)
18	20	application_data	DAYS_BIRTH	Client's age in days at the time of application
19	21	application_data	DAYS_EMPLOYED	How many days before the application the person started current employment
20	22	application_data	DAYS_REGISTRATION	How many days before the application did client change his registration
21	24	application_data	OWN_CAR_AGE	Age of client's car
22	31	application_data	OCCUPATION_TYPE	What kind of occupation does the client have
23	32	application_data	CNT_FAM_MEMBERS	How many family members does client have
24	33	application_data	REGION_RATING_CLIENT	Our rating of the region where client lives (1,2,3)
25	34	application_data	REGION_RATING_CLIENT_W_CITY	Our rating of the region where client lives with taking city into account (1,2,3)
26	35	application_data	WEEKDAY_APPR_PROCESS_START	On which day of the week did the client apply for the loan
27	36	application_data	HOUR_APPR_PROCESS_START	Approximately at what hour did the client apply for the loan
28	43	application_data	ORGANIZATION_TYPE	Type of organization where client works
29	94	application_data	OBS_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 30 DPD (days past due) default
30	95	application_data	DEF_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 30 DPD (days past due)
31	96	application_data	OBS_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 60 DPD (days past due) default
32	97	application_data	DEF_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 60 (days past due) DPD

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)

Table Stats - No of blanks and their percentage to determine whether to impute them or drop the Columns (Blanks above 30 % will be dropped - Red & Orange Color) Also Categorical Column distinction to decide whether to keep or drop them as most are not useful in analysis - 0 or 1 in values or just names.

Column Names	No. Of Blanks	% of blanks	Is Categorical	To Drop or Not
SK_ID_CURR	0	0%	Categorical Column	No
TARGET	0	0%	Categorical Column	No
NAME_CONTRACT_TYPE	0	0%	Categorical Column	No
CODE_GENDER	0	0%	Categorical Column	No
FLAG_OWN_CAR	0	0%	Categorical Column	No
FLAG_OWN_REALTY	0	0%	Categorical Column	No
CNT_CHILDREN	0	0%	Categorical Column	No
AMT_INCOME_TOTAL	0	0%	Categorical Column	No
AMT_CREDIT	0	0%	Categorical Column	No
AMT_ANNUITY	1	0%	Not Categorical	No
AMT_GOODS_PRICE	38	0%	Not Categorical	No
NAME_TYPE_SUITE	192	0%	Not Categorical	No
NAME_INCOME_TYPE	0	0%	Categorical Column	No
NAME_EDUCATION_TYPE	0	0%	Categorical Column	No
NAME_FAMILY_STATUS	0	0%	Categorical Column	No
NAME_HOUSING_TYPE	0	0%	Categorical Column	No
REGION_POPULATION_RELATIVE	0	0%	Categorical Column	No
DAYS_BIRTH	0	0%	Categorical Column	No
DAYS_EMPLOYED	0	0%	Categorical Column	No
DAYS_REGISTRATION	0	0%	Categorical Column	No
DAYS_ID_PUBLISH	0	0%	Categorical Column	No
FLAG_MOBIL	0	0%	Categorical Column	No
FLAG_EMP_PHONE	0	0%	Categorical Column	No
FLAG_WORK_PHONE	0	0%	Categorical Column	No
FLAG_CONT_MOBILE	0	0%	Categorical Column	No
FLAG_PHONE	0	0%	Categorical Column	No
FLAG_EMAIL	0	0%	Categorical Column	No
OCCUPATION_TYPE	15654	31%	Not Categorical	Yes
CNT_FAM_MEMBERS	1	0%	Not Categorical	No
REGION_RATING_CLIENT	0	0%	Categorical Column	No
REGION_RATING_CLIENT_W_CITY	0	0%	Categorical Column	No
WEEKDAY_APPR_PROCESS_START	0	0%	Categorical Column	No
HOUR_APPR_PROCESS_START	0	0%	Categorical Column	No
REG_REGION_NOT_LIVE_REGION	0	0%	Categorical Column	No
REG_REGION_NOT_WORK_REGION	0	0%	Categorical Column	No
LIVE_REGION_NOT_WORK_REGION	0	0%	Categorical Column	No
REG_CITY_NOT_LIVE_CITY	0	0%	Categorical Column	No
REG_CITY_NOT_WORK_CITY	0	0%	Categorical Column	No
LIVE_CITY_NOT_WORK_CITY	0	0%	Categorical Column	No
ORGANIZATION_TYPE	0	0%	Categorical Column	No
EXT_SOURCE_1	28172	56%	Not Categorical	Yes
EXT_SOURCE_2	126	0%	Not Categorical	No
EXT_SOURCE_3	9944	20%	Not Categorical	No
APARTMENTS_AVG	25385	51%	Not Categorical	Yes
BASEMENTAREA_AVG	29199	58%	Not Categorical	Yes
YEARS_BEGINEXPLUATATION_AVG	24394	49%	Not Categorical	Yes
YEARS_BUILD_AVG	33239	66%	Not Categorical	Yes
COMMONAREA_AVG	34960	70%	Not Categorical	Yes
ELEVATORS_AVG	26651	53%	Not Categorical	Yes
ENTRANCES_AVG	25195	50%	Not Categorical	Yes
FLOORSMAX_AVG	24875	50%	Not Categorical	Yes
FLOORSMIN_AVG	33894	68%	Not Categorical	Yes
LANDAREA_AVG	29721	59%	Not Categorical	Yes
LIVINGAPARTMENTS_AVG	34226	68%	Not Categorical	Yes
LIVINGAREA_AVG	25137	50%	Not Categorical	Yes
NONLIVINGAPARTMENTS_AVG	34714	69%	Not Categorical	Yes
NONLIVINGAREA_AVG	27572	55%	Not Categorical	Yes
APARTMENTS_MODE	25385	51%	Not Categorical	Yes
BASEMENTAREA_MODE	29199	58%	Not Categorical	Yes
YEARS_BEGINEXPLUATATION_MODE	24394	49%	Not Categorical	Yes
YEARS_BUILD_MODE	33239	66%	Not Categorical	Yes
COMMONAREA_MODE	34960	70%	Not Categorical	Yes
ELEVATORS_MODE	26651	53%	Not Categorical	Yes
ENTRANCES_MODE	25195	50%	Not Categorical	Yes
FLOORSMAX_MODE	24875	50%	Not Categorical	Yes
FLOORSMIN_MODE	33894	68%	Not Categorical	Yes
LANDAREA_MODE	29721	59%	Not Categorical	Yes
LIVINGAPARTMENTS_MODE	34226	68%	Not Categorical	Yes
LIVINGAREA_MODE	25137	50%	Not Categorical	Yes
NONLIVINGAPARTMENTS_MODE	34714	69%	Not Categorical	Yes
NONLIVINGAREA_MODE	27572	55%	Not Categorical	Yes
APARTMENTS_MEDI	25385	51%	Not Categorical	Yes
BASEMENTAREA_MEDI	29199	58%	Not Categorical	Yes
YEARS_BEGINEXPLUATATION_MEDI	24394	49%	Not Categorical	Yes
YEARS_BUILD_MEDI	33239	66%	Not Categorical	Yes
COMMONAREA_MEDI	34960	70%	Not Categorical	Yes
ELEVATORS_MEDI	26651	53%	Not Categorical	Yes
ENTRANCES_MEDI	25195	50%	Not Categorical	Yes
FLOORSMAX_MEDI	24875	50%	Not Categorical	Yes
FLOORSMIN_MEDI	33894	68%	Not Categorical	Yes
LANDAREA_MEDI	29721	59%	Not Categorical	Yes
LIVINGAPARTMENTS_MEDI	34226	68%	Not Categorical	Yes
LIVINGAREA_MEDI	25137	50%	Not Categorical	Yes
NONLIVINGAPARTMENTS_MEDI	34714	69%	Not Categorical	Yes
NONLIVINGAREA_MEDI	27572	55%	Not Categorical	Yes
FONDKAPREMONT_MODE	34191	68%	Not Categorical	Yes
HOUSETYPE_MODE	25075	50%	Not Categorical	Yes
TOTALAREA_MODE	24148	48%	Not Categorical	Yes
WALLSMATERIAL_MODE	25459	51%	Not Categorical	Yes
EMERGENCYSTATE_MODE	23698	47%	Not Categorical	Yes
OBS_30_CNT_SOCIAL_CIRCLE	168	0%	Not Categorical	No
DEF_30_CNT_SOCIAL_CIRCLE	168	0%	Not Categorical	No
OBS_60_CNT_SOCIAL_CIRCLE	168	0%	Not Categorical	No
DEF_60_CNT_SOCIAL_CIRCLE	168	0%	Not Categorical	No
DAYS_LAST_PHONE_CHANGE	1	0%	Not Categorical	No
FLAG_DOCUMENT_2	0	0%	Categorical Column	No
FLAG_DOCUMENT_3	0	0%	Categorical Column	No
FLAG_DOCUMENT_4	0	0%	Categorical Column	No
FLAG_DOCUMENT_5	0	0%	Categorical Column	No
FLAG_DOCUMENT_6	0	0%	Categorical Column	No
FLAG_DOCUMENT_7	0	0%	Categorical Column	No
FLAG_DOCUMENT_8	0	0%	Categorical Column	No
FLAG_DOCUMENT_9	0	0%	Categorical Column	No
FLAG_DOCUMENT_10	0	0%	Categorical Column	No
FLAG_DOCUMENT_11	0	0%	Categorical Column	No
FLAG_DOCUMENT_12	0	0%	Categorical Column	No
FLAG_DOCUMENT_13	0	0%	Categorical Column	No
FLAG_DOCUMENT_14	0	0%	Categorical Column	No
FLAG_DOCUMENT_15	0	0%	Categorical Column	No
FLAG_DOCUMENT_16	0	0%	Categorical Column	No
FLAG_DOCUMENT_17	0	0%	Categorical Column	No
FLAG_DOCUMENT_18	0	0%	Categorical Column	No
FLAG_DOCUMENT_19	0	0%	Categorical Column	No
FLAG_DOCUMENT_20	0	0%	Categorical Column	No
FLAG_DOCUMENT_21	0	0%	Categorical Column	No
AMT_REQ_CREDIT_BUREAU_HOUR	6734	13%	Not Categorical	No
AMT_REQ_CREDIT_BUREAU_DAY	6734	13%	Not Categorical	No
AMT_REQ_CREDIT_BUREAU_WEEK	6734	13%	Not Categorical	No
AMT_REQ_CREDIT_BUREAU_MON	6734	13%	Not Categorical	No
AMT_REQ_CREDIT_BUREAU_QRT	6734	13%	Not Categorical	No
AMT_REQ_CREDIT_BUREAU_YEAR	6734	13%	Not Categorical	No



HD Image 1 – Pivot Table *HD Image 2 – Gif of Pivot Table & Bar Graph*

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

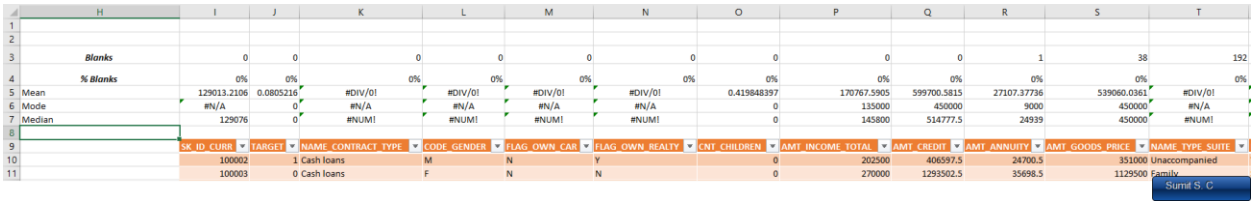
Insights:

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 ,

> Duplicate Values , Dupes

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.



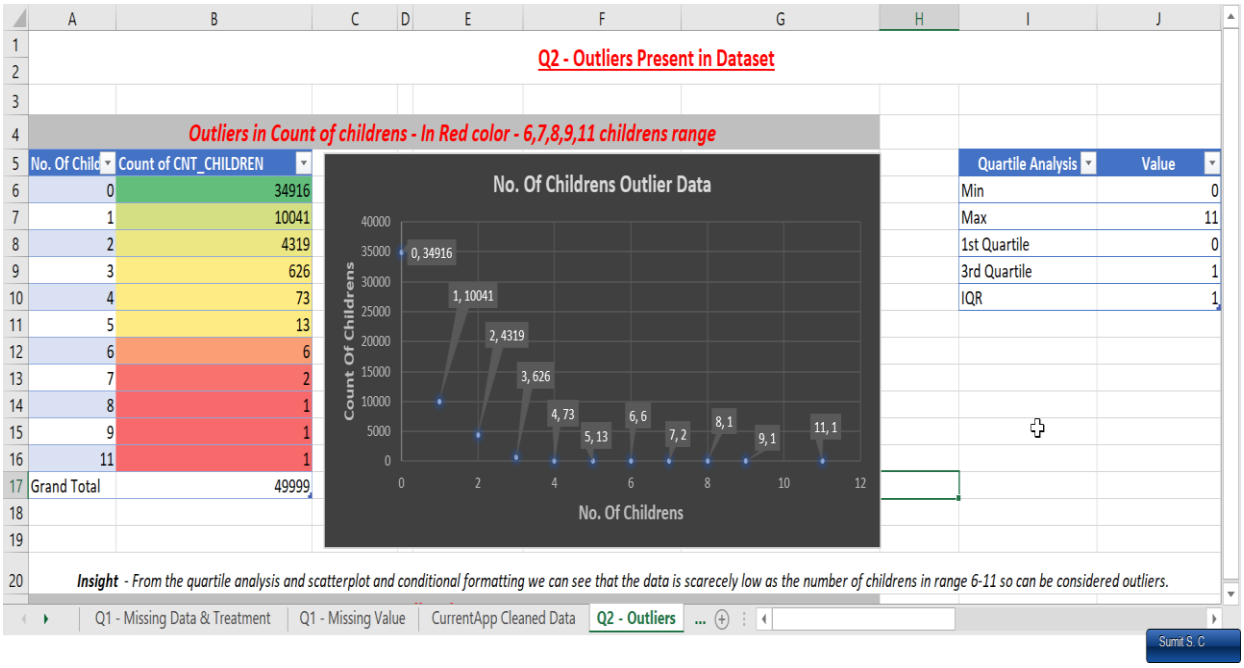
	H	I	J	K	L	M	N	O	P	Q	R	S	T
1													
2													
3	*Blanks*	0	0	0	0	0	0	0	0	0	1	38	192
4	*% Blanks*	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
5	Mean	129013.2106	0.0805216	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	0.419848397	170767.5905	599700.5815	27107.37736	539060.0361	#DIV/0!
6	Mode	#N/A	0	#N/A	#N/A	#N/A	#N/A	0	135000	450000	9000	450000	#N/A
7	Median	129076	0	#NUM!	#NUM!	#NUM!	#NUM!	0	145800	514777.5	24939	450000	#NUM!
8													
9		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_PRICE	NAME_TYPE_SUITE
10		100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied
11		100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family

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.

3. 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 (Cleaned Dataset) was made to do analysis with necessary columns.
4. Files
    - Q 1 Excel File, Docs File
    - Loom Video Q1 – Identifying Missing Values & Treating them.

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.*

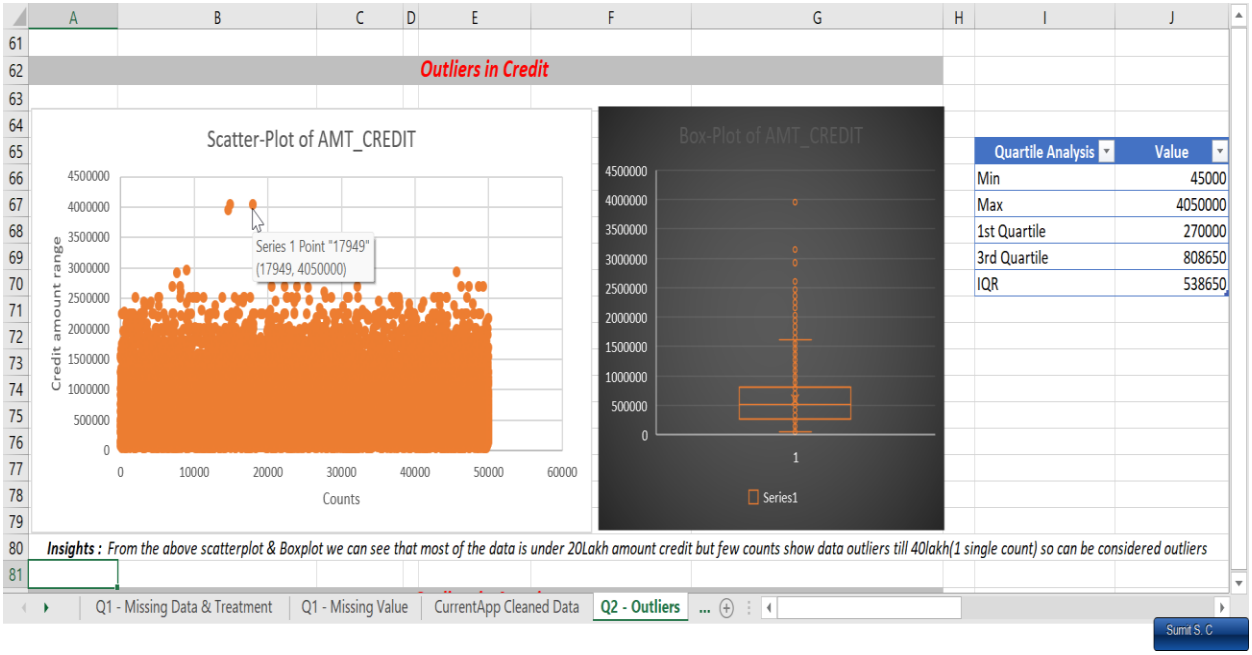
- *Task***:** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.



No. Of Childrens	Count of CNT_CHILDREN
0	34916
1	10041
2	4319
3	626
4	73
5	13
6	6
7	2
8	1
9	1
11	1
Grand Total	49999

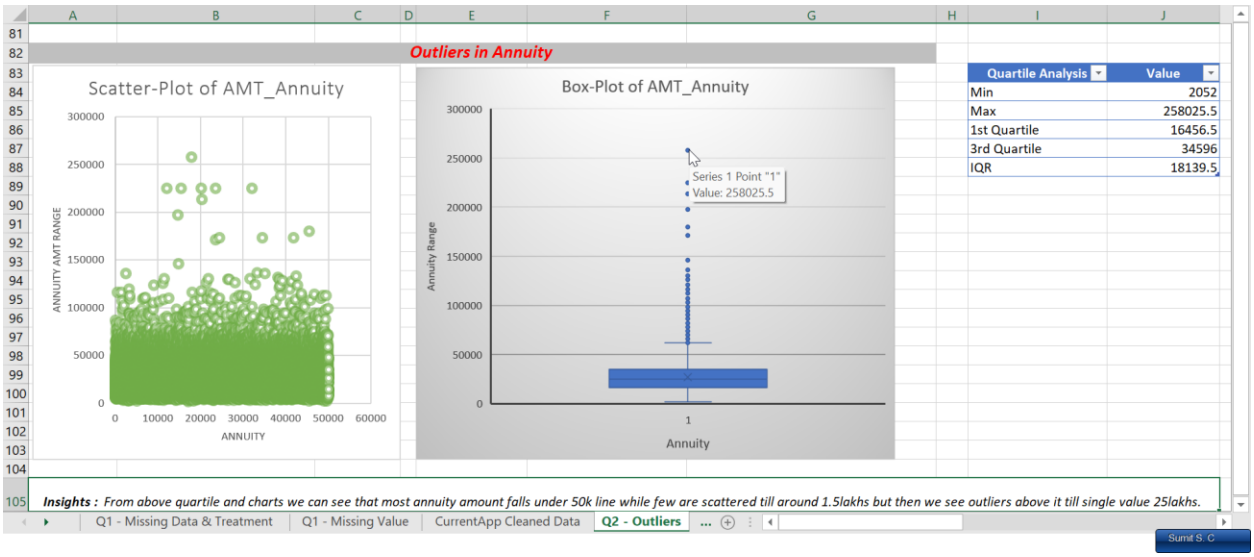
Quartile Analysis	Value
Min	0
Max	11
1st Quartile	0
3rd Quartile	1
IQR	1

[Outlier 1](#) - *In No. Of childrens*

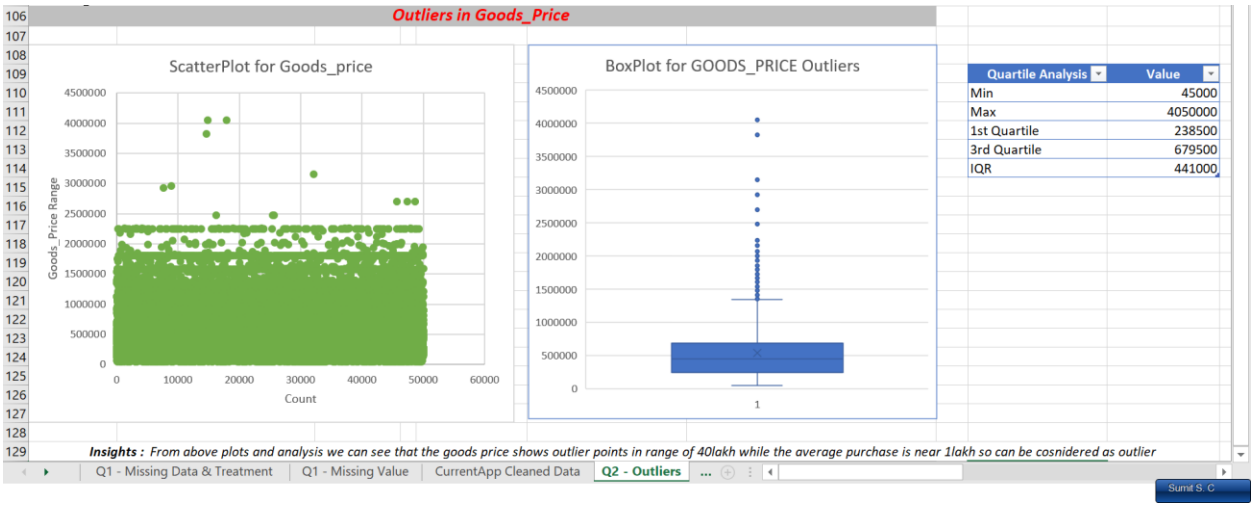


Quartile Analysis	Value
Min	45000
Max	4050000
1st Quartile	270000
3rd Quartile	808650
IQR	538650

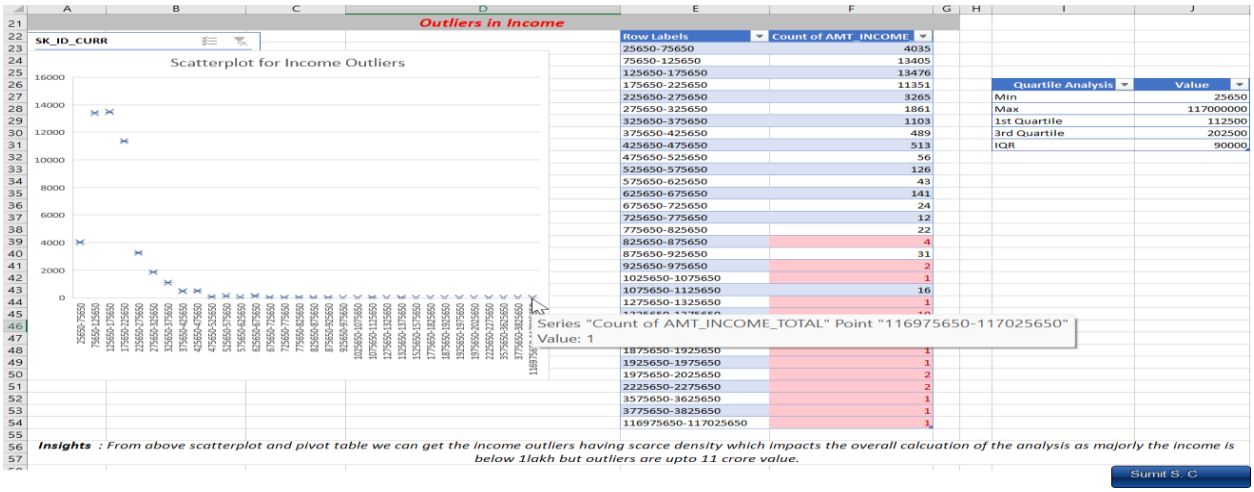
[Outlier 2](#) - *In Credit_AMT*



Outlier 3 In Annuity Amount

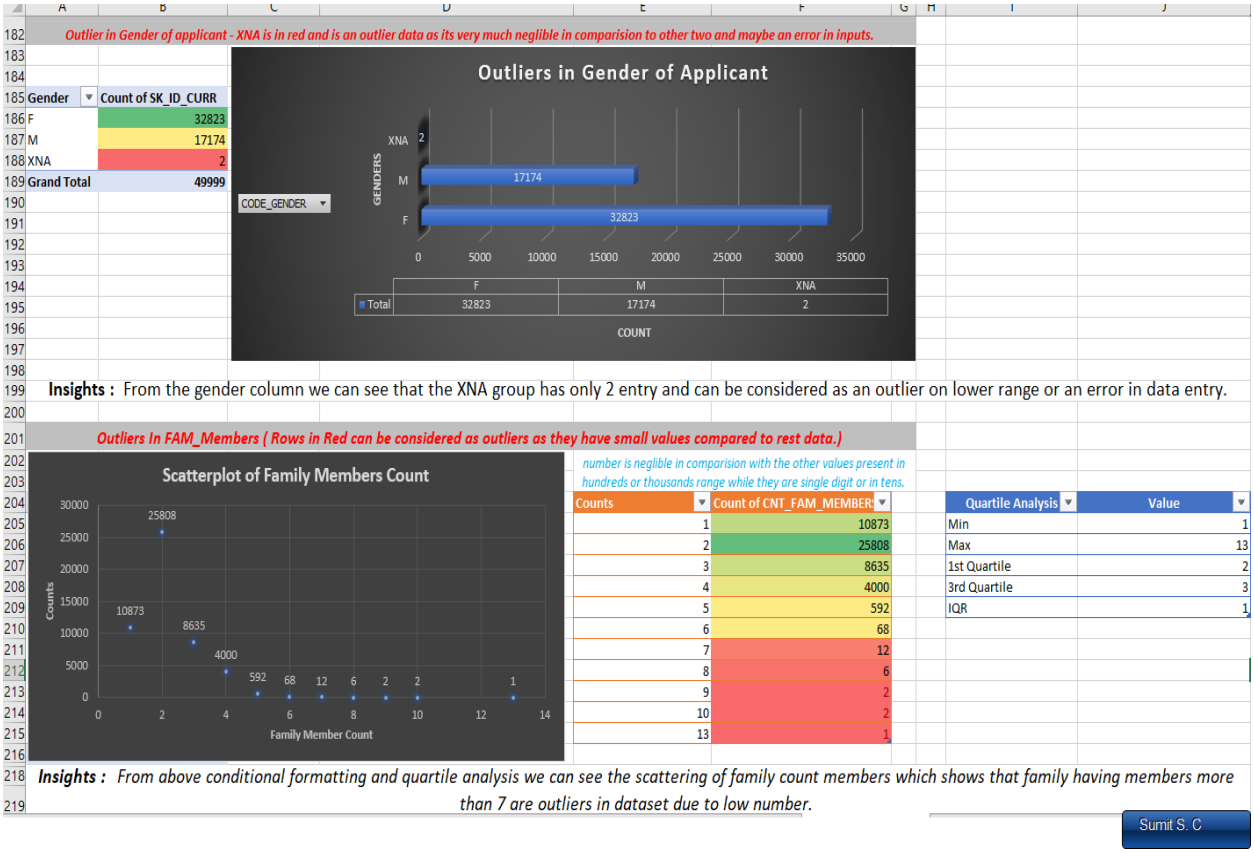


Outlier 4 - In Goods_Price Column



Outlier 5 - Outliers In income

Outlier 6 - Outliers in Tenure (In employment years and from application day)



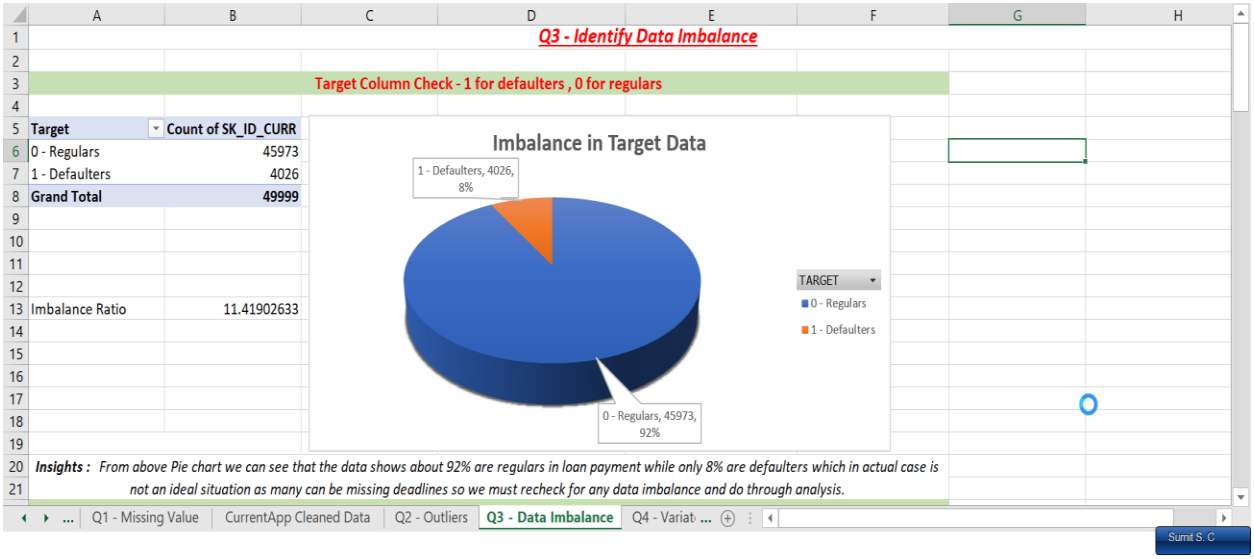
Outlier 7 - Inside Gender distribution & Family members count

Insights:

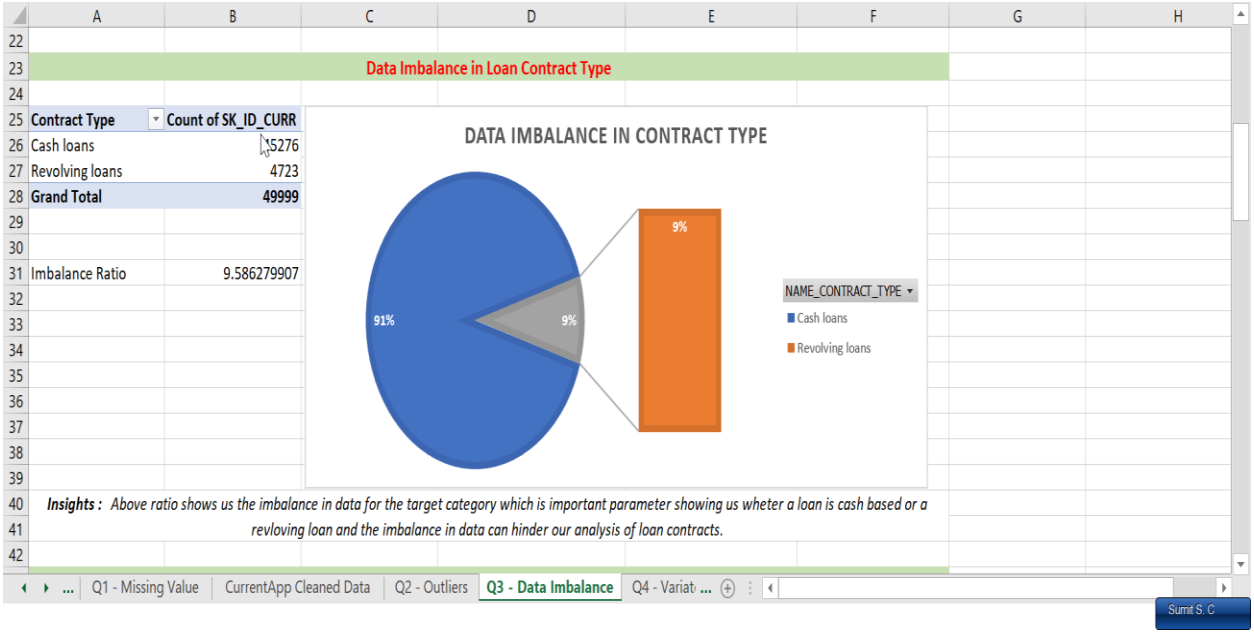
1. For every outlier I have tried giving the insight in the screenshot but to summarize the overall data we have found out that many outliers data are present in the dataset that needs to be checked or appropriately dealt with to make the analysis and overall future predictions about customer loan repayments 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 and go to Q2 tab)
    - Loom Video Explanation (Identifying Outliers)

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.*

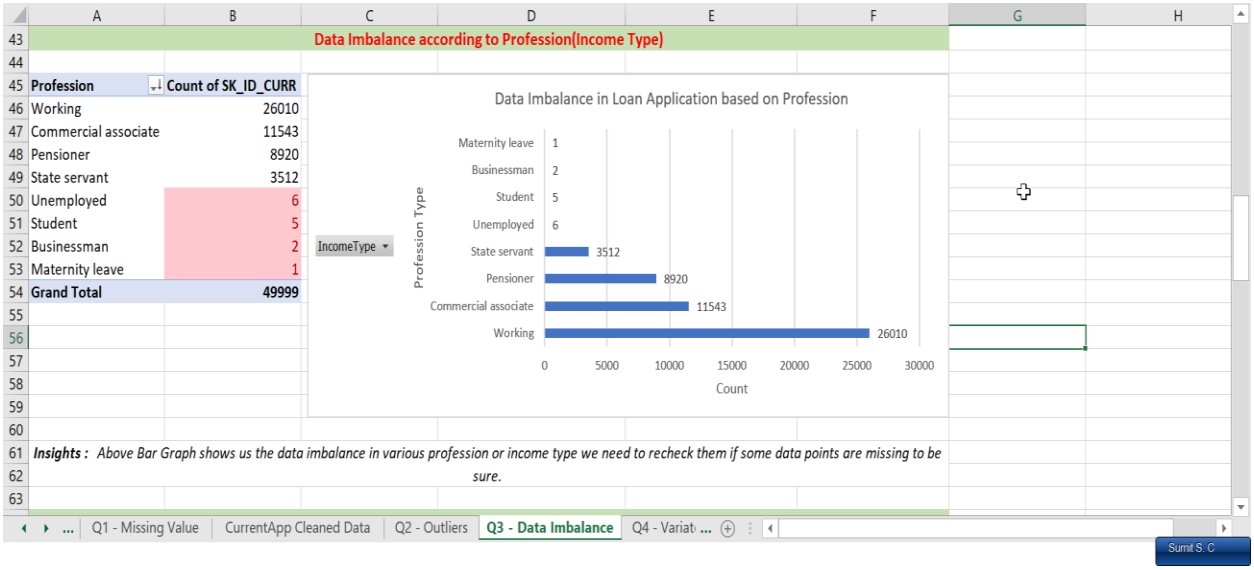
- *Task***:** Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.



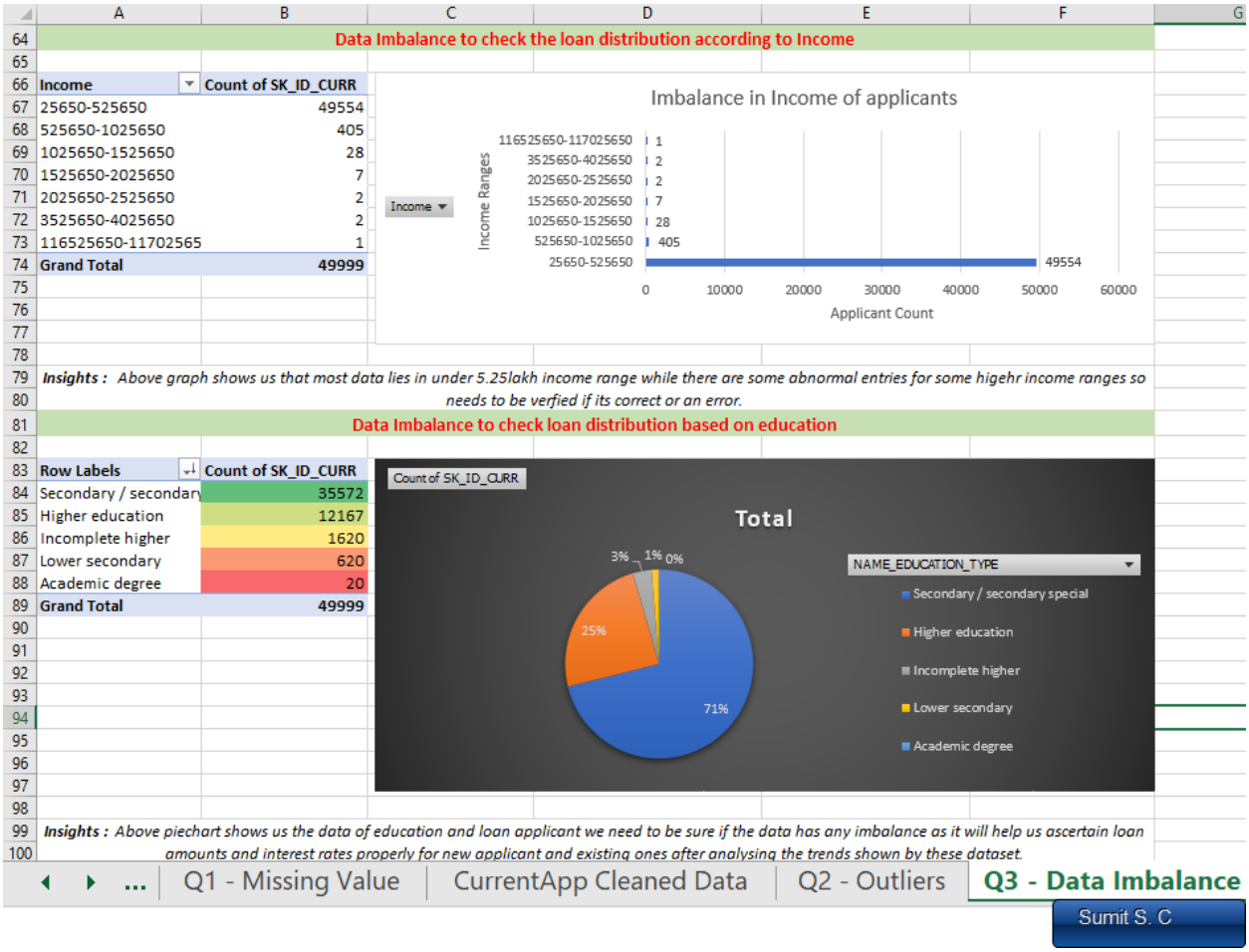
Data Imbalance 1 - In target Column



Data Imbalance 2 - In Target Category



Data Imbalance 3 - In Income Typehttps://mitsus.life-is-pa.in/6gMrRvcDW.png



Data Imbalance 4 - [In education & Income range]

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

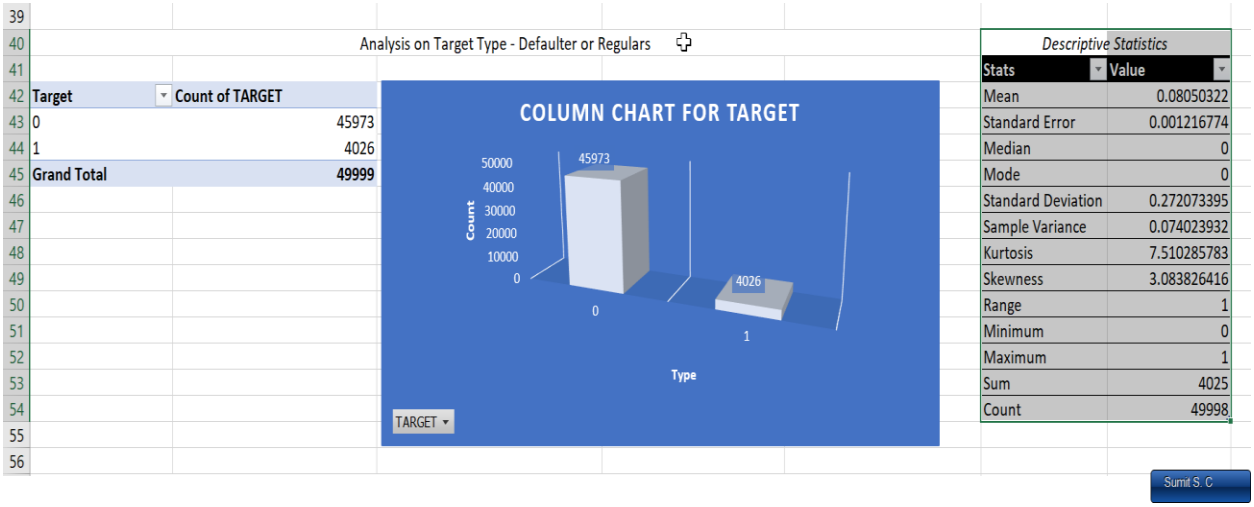
   - Q3 Excel File, Gdocs File

   - Loom Video PPT Q3 – Identify the data imbalance

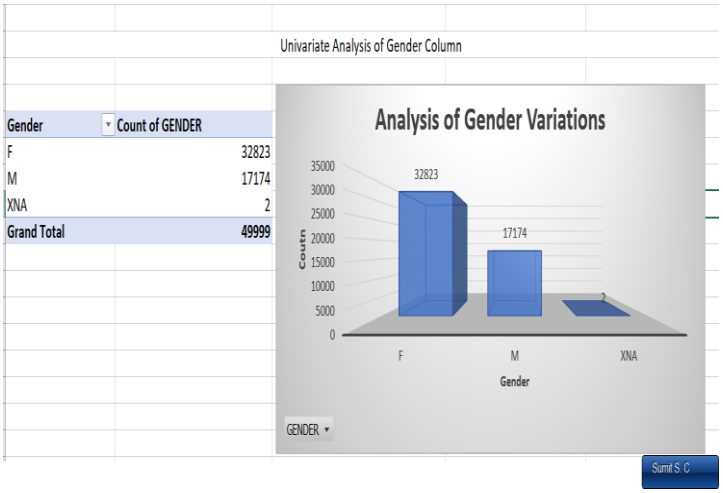
D) Perform Univariate, Segmented Univariate, and Bivariate Analysis: *To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.*

- *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.

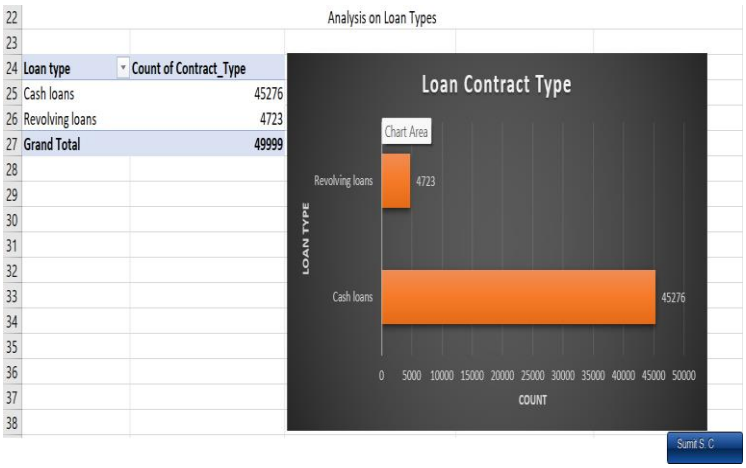
  Univariate Analysis – When we examine and analyse a single variable in isolation and understand the distribution summary stats using descriptive stats and try to gain information its called as univariate analysis.



Univariate 1 - On Target



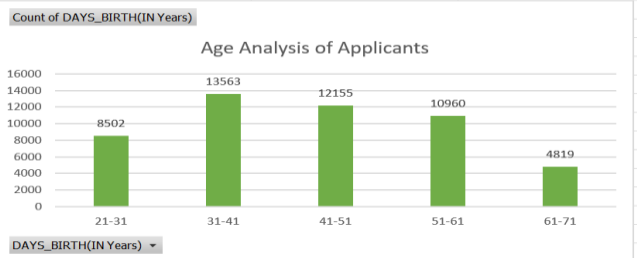
Univariate 2 On Gender



Univariate 3 On Loan type

Analysis for Age Distribution

Age Frequency	Count of DAYS_BIRTH(IN Years)
21-31	8502
31-41	13563
41-51	12155
51-61	10960
61-71	4819
Grand Total	49999



Stats For Age

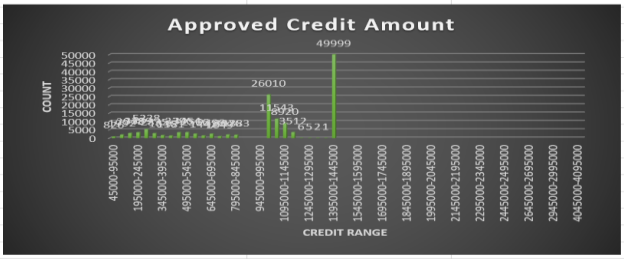
Stats	Value
Mean	43.89604592
Standard Error	0.053438808
Median	43.1
Mode	27.7
Standard Deviation	11.94916117
Sample Variance	142.7824527
Kurtosis	-1.042954817
Skewness	0.118851207
Range	48
Minimum	21
Maximum	69
Sum	2194758.4
Count	49999

Sumit S. C

Univariate 3 On Age

Analysis of Approved Loan Amount

Row Labels	Count of AMT_CREDIT
45000-95000	826
95000-145000	1992
145000-195000	2948
195000-245000	3389
245000-295000	5238
295000-345000	2835
345000-395000	1646
395000-445000	1381
445000-495000	3396
495000-545000	3454
545000-595000	2566
595000-645000	1442
645000-695000	2498
695000-745000	1049
745000-795000	2026
795000-845000	1883



Stats For Credit Amount

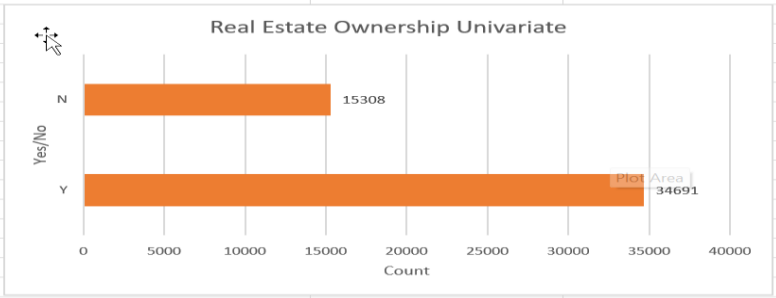
Stats	Value
Mean	43.89604592
Standard Error	0.053438808
Median	43.1
Mode	27.7
Standard Deviation	11.94916117
Sample Variance	142.7824527
Kurtosis	-1.042954817
Skewness	0.118851207
Range	48
Minimum	21
Maximum	69
Sum	2194758.4
Count	49999

Sumit S. C

Univariate 4 On Loan Amount

Analysis of Real Estate Ownership

Real Estate Owners	Count of FLAG_OWN_REALTY
Y	34691
N	15308
Grand Total	49999

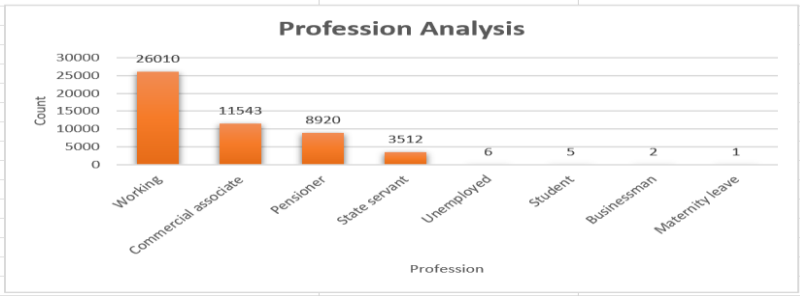


Sumit S. C

Univariate 5 on estate ownership

Univariate Analysis of Profession/Income Source of Applicants

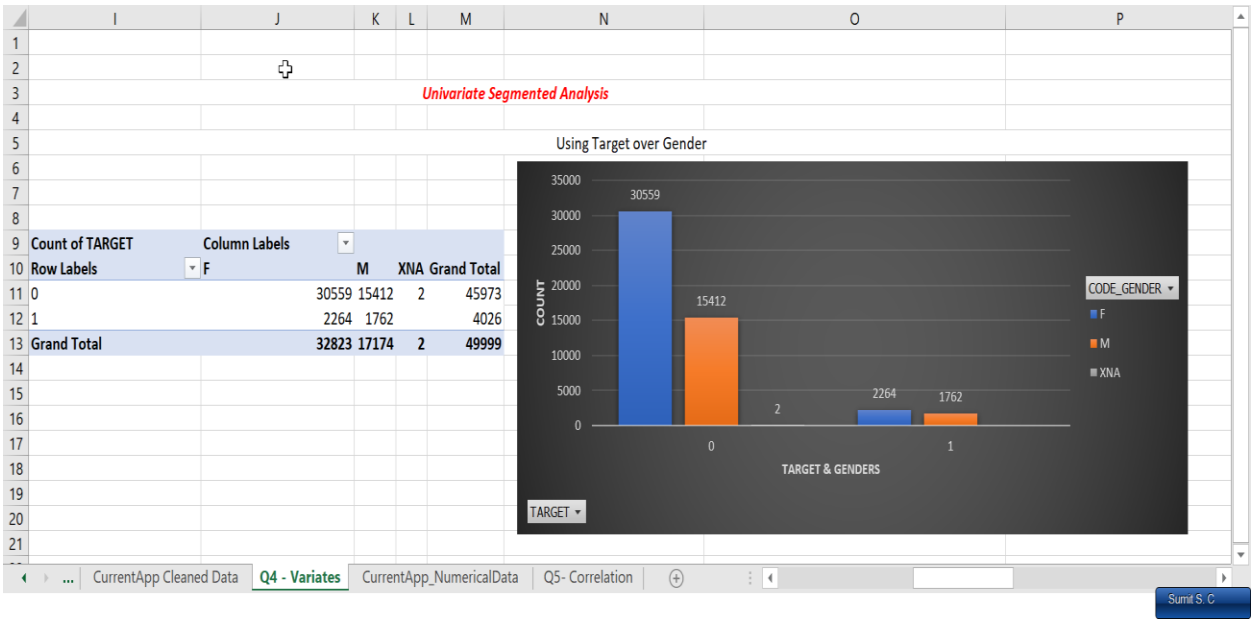
Profession	Count of NAME_INCOME_TYPE
Working	26010
Commercial associate	11543
Pensioner	8920
State servant	3512
Unemployed	6
Student	5
Businessman	2
Maternity leave	1
Grand Total	49999



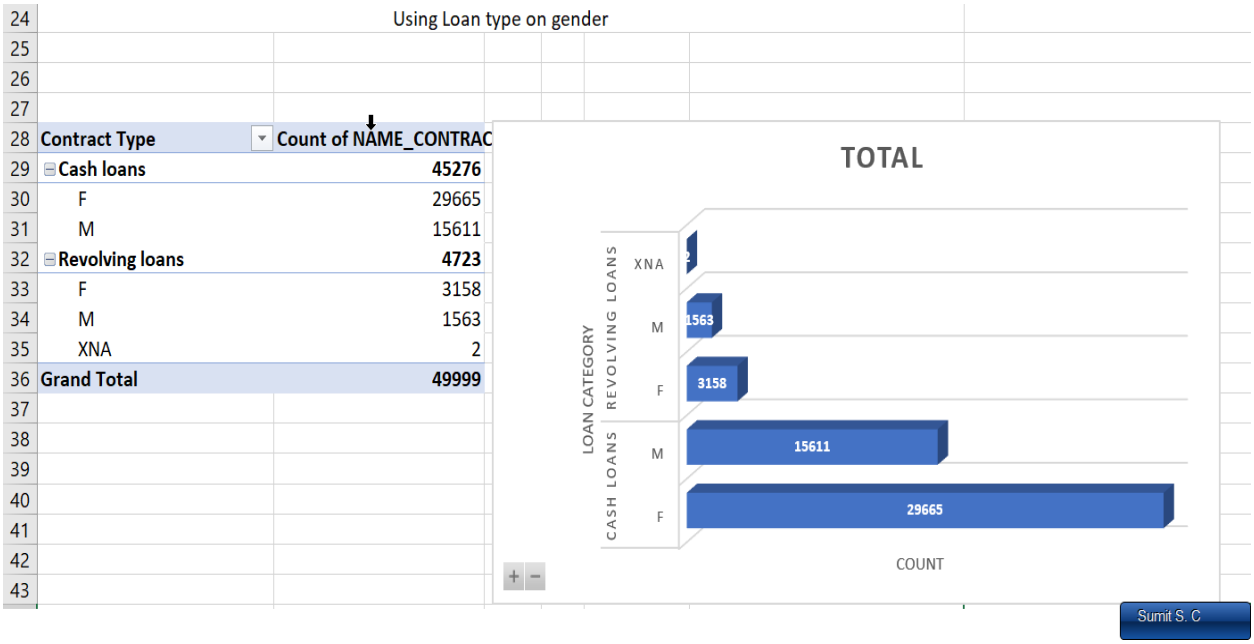
Sumit S. C

Univariate 6 On Income Type

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.*



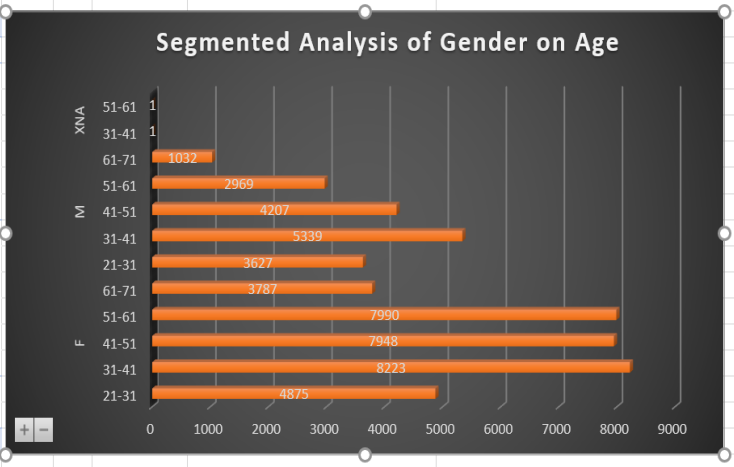
Univariate Segmented 1 - [On Target & Gender]



Univariate Segmented 2 - [On Loan type & gender]

Segmented Analysis on Age & Gender

Gender & Age Frequency	Count of CODE_GENDER
F	32823
21-31	4875
31-41	8223
41-51	7948
51-61	7990
61-71	3787
M	17174
21-31	3627
31-41	5339
41-51	4207
51-61	2969
61-71	1032
XNA	2
31-41	1
51-61	1
Grand Total	49999

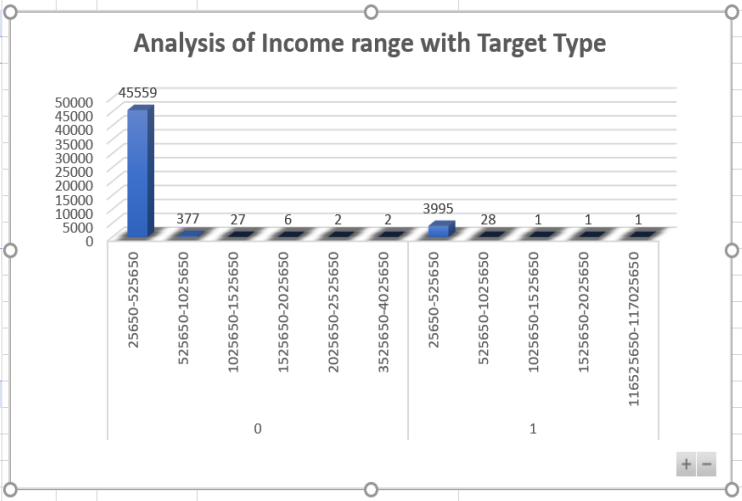


Sumit S. C

Univariate Segmented 3 - On Gender & Age

Segmented Analysis of Target type and Income

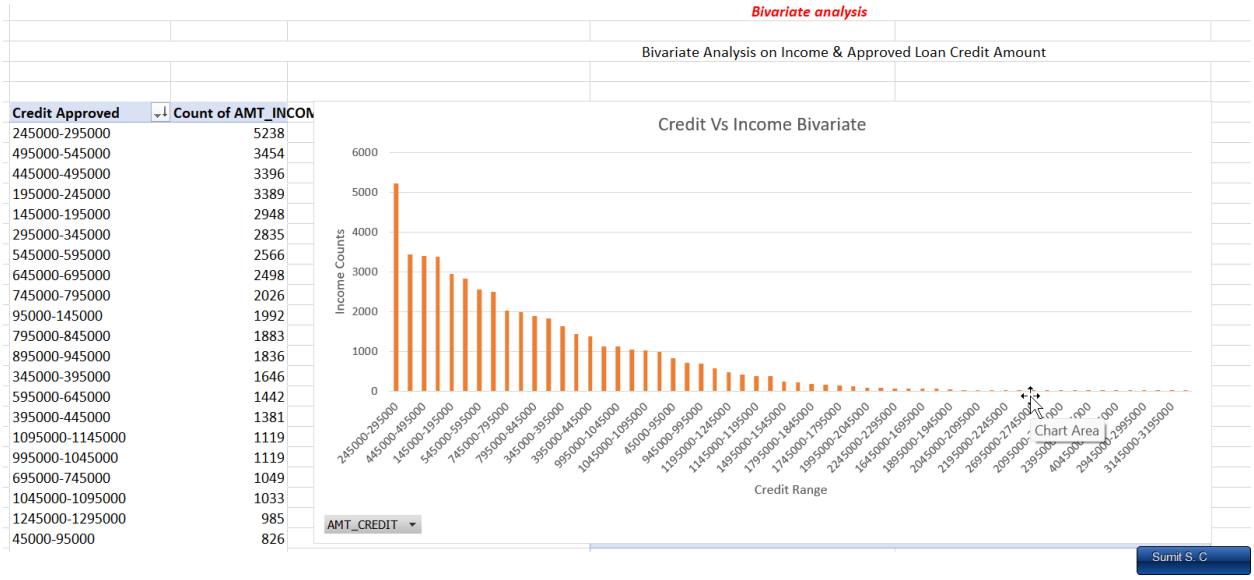
Income Range & Target	Count of TARGET
0	45973
25650-525650	45559
525650-1025650	377
1025650-1525650	27
1525650-2025650	6
2025650-2525650	2
3525650-4025650	2
1	4026
25650-525650	3995
525650-1025650	28
1025650-1525650	1
1525650-2025650	1
116525650-117025650	1
Grand Total	49999



Sumit S. C

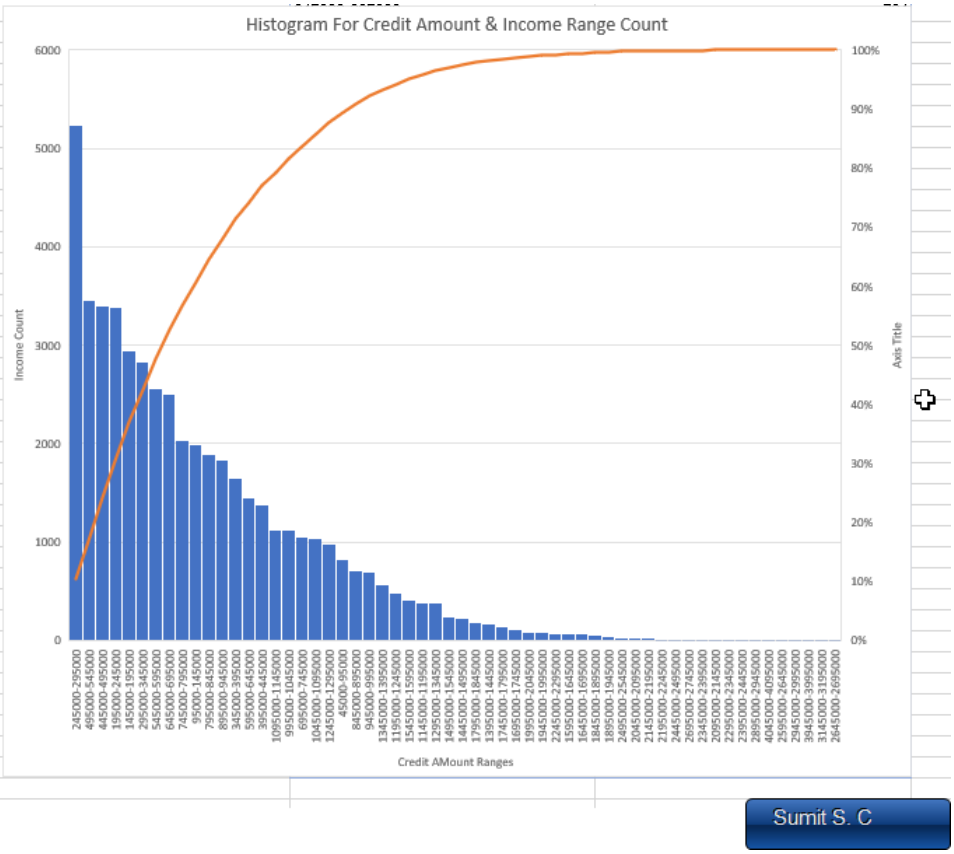
Univariate Segmented 4 - On target & Income Range

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.*

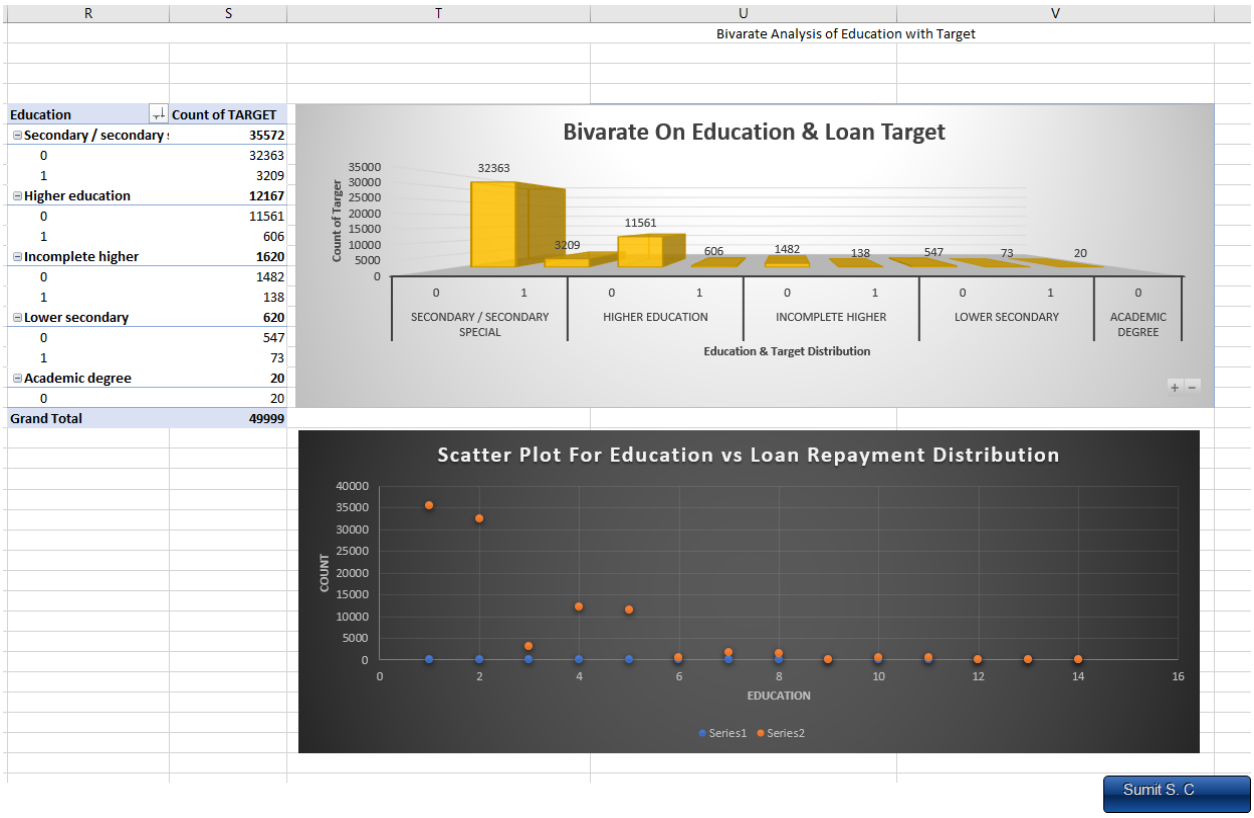


Bivariate 1 [Column Bar GraphOn Credit amount & Credit approved]

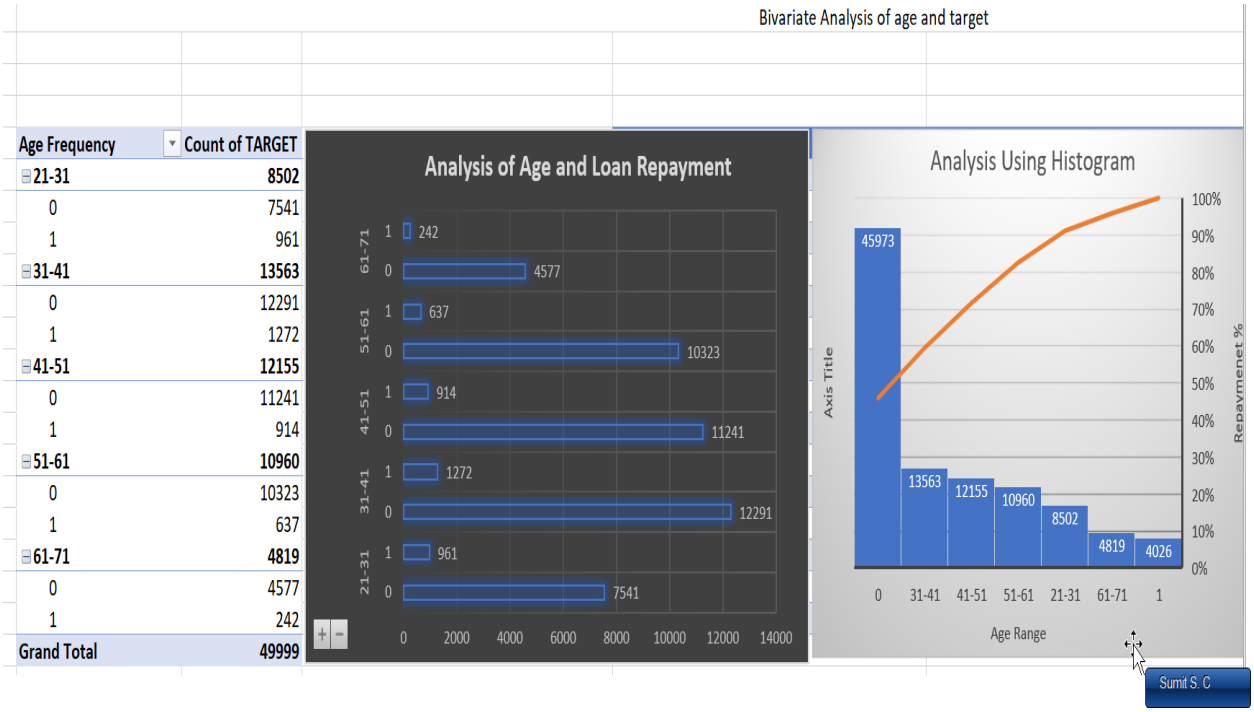
845000-895000	704
945000-995000	698
1345000-1395000	567
1195000-1245000	477
1545000-1595000	411
1145000-1195000	376
1295000-1345000	376
1495000-1545000	239
1445000-1495000	227
1795000-1845000	184
1395000-1445000	172
1745000-1795000	144
1695000-1745000	117
1995000-2045000	85
1945000-1995000	79
2245000-2295000	73
1595000-1645000	72
1645000-1695000	71
1845000-1895000	58
1895000-1945000	41
2495000-2545000	32
2045000-2095000	32
2145000-2195000	28
2195000-2245000	15
2445000-2495000	12
2695000-2745000	10
2345000-2395000	8
2095000-2145000	8
2295000-2345000	6
2395000-2445000	6
2895000-2945000	2
4045000-4095000	2
2595000-2645000	2
2945000-2995000	1
3945000-3995000	1
3145000-3195000	1
2645000-2695000	1
Grand Total	**49999**



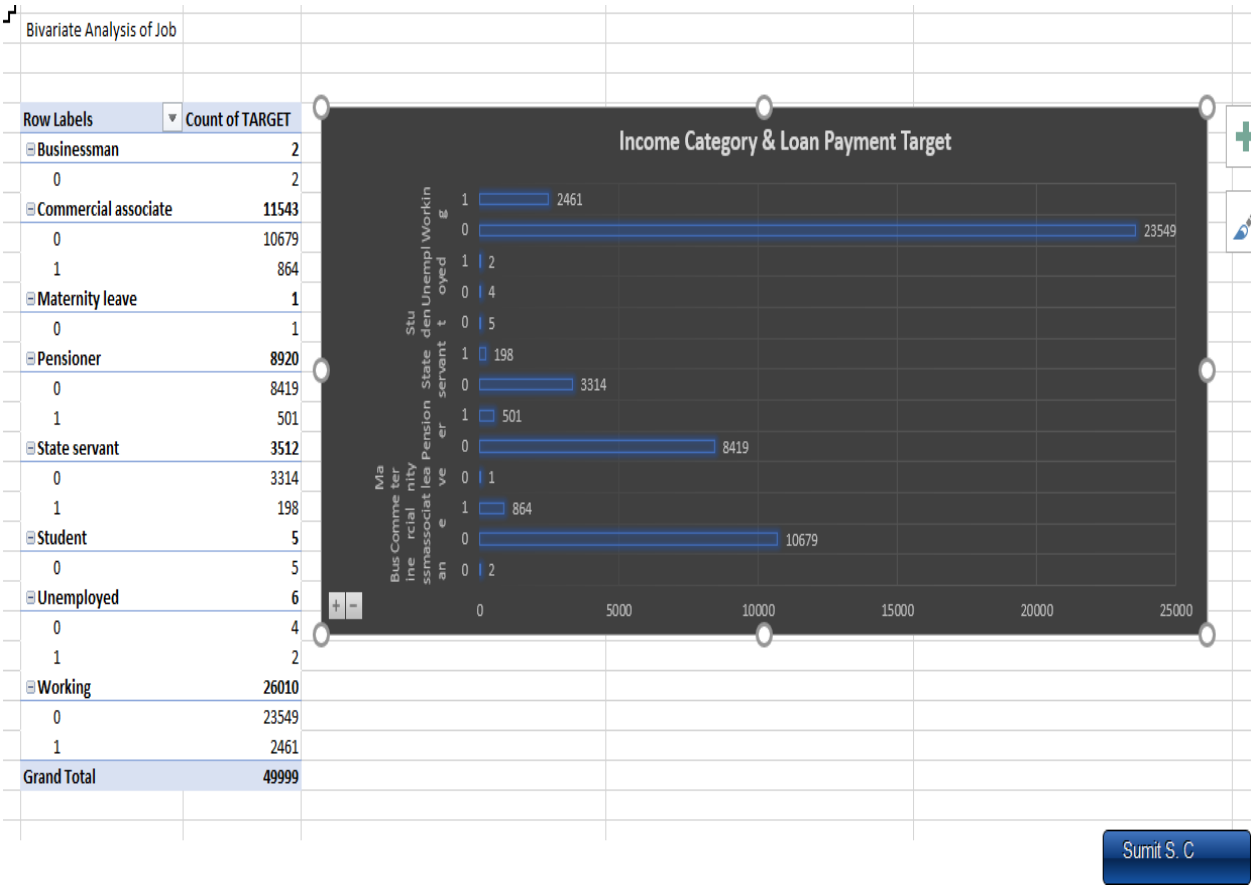
Bivariate 2 [Histogram on above set]



Bivariate 3 - [On education and target counts (Column & Scatterplot distribution)]



Bivariate 4 - [On age & Target (Histogram & Bar plot)]



Bivariate 5 - [On Profession & Target type (Bar Plotting)]

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

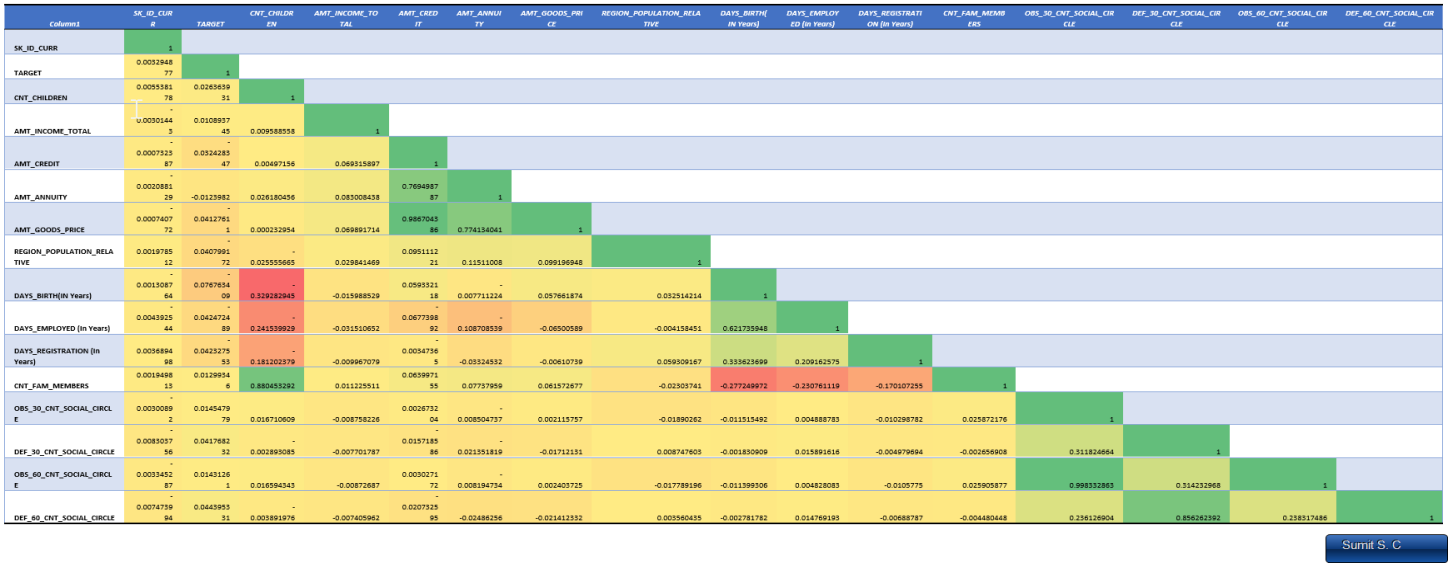
    - Q4 Excel Link, Gdocs Links

    - Loom Video PPT Q4 – Variate analysis

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.

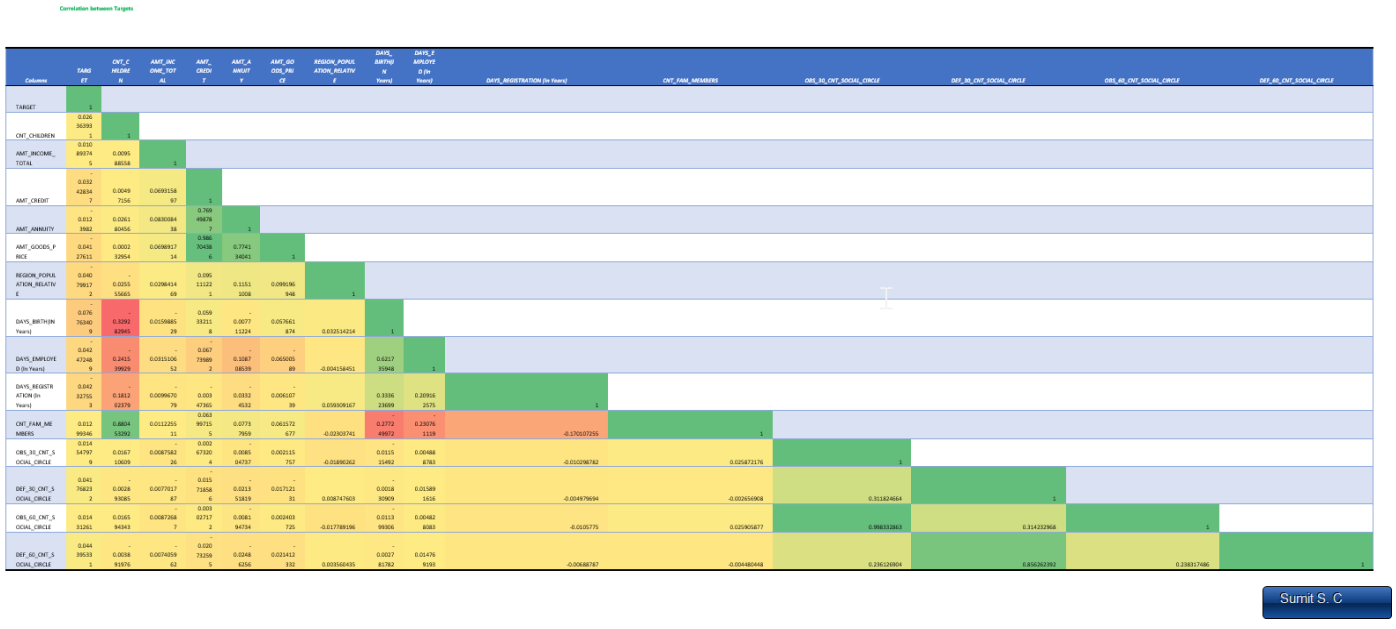
Correlation Among the Numerical Columns Which have impact on Loan Repayment using the data analysis tool in Excel or by using =CORREL() function and passing the arrays to compare



Column1	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH(IN Years)	DAYS_EMPLOYED (In Years)	DAYS_REGISTRATION (In Years)	CNT_FAM_MEMBERS	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE
SK_ID_CURR	1															
TARGET	0.0032948	1														
CNT_CHILDREN	0.0055581	0.0263639	1													
AMT_INCOME_TOTAL	-0.0030144	0.0108937	0.009588558	1												
AMT_CREDIT	-0.0007523	0.0324283	0.00497156	0.069315897	1											
AMT_ANNUITY	-0.0020881	-0.0123982	0.026180456	0.083008458	0.7694987	1										
AMT_GOODS_PRICE	-0.0007407	0.0412761	0.000232954	0.069891714	0.9867043	0.774134041	1									
REGION_POPULATION_RELATIVE	0.0019785	-0.0407991	-0.025555665	0.029841469	0.0951112	0.11511008	0.099196948	1								
DAYS_BIRTH(IN Years)	-0.0013087	0.0767634	-0.329282945	-0.015988529	0.0593521	-0.007711224	0.057661874	0.032514214	1							
DAYS_EMPLOYED (In Years)	-0.0043925	-0.0424724	0.241539929	-0.031510652	-0.0677398	-0.108708539	-0.06500589	-0.004158451	0.621735948	1						
DAYS_REGISTRATION (In Years)	0.0036894	-0.0425275	0.181202579	-0.009967079	-0.0054736	-0.03324532	-0.00810739	0.059309167	0.333623699	0.209162575	1					
CNT_FAM_MEMBERS	0.0019498	0.0129934	0.880453292	0.011225511	0.0659871	0.07757959	0.061572677	-0.02303741	-0.277249972	-0.230761119	-0.170107255	1				
OBS_30_CNT_SOCIAL_CIRCLE	-0.0030089	0.0145479	0.016710609	-0.008758226	-0.0026732	-0.008504737	-0.002115757	-0.01890262	-0.011515492	0.004888783	-0.010298782	0.025872176	1			
DEF_30_CNT_SOCIAL_CIRCLE	-0.0083057	0.0417682	0.002893085	-0.007701787	-0.0157185	-0.021351819	-0.01712151	0.008747603	-0.001830909	0.015891618	-0.004979694	-0.002656908	0.311824664	1		
OBS_60_CNT_SOCIAL_CIRCLE	-0.0033452	0.0143126	0.016594343	-0.00872687	-0.0030271	-0.008194734	-0.002405725	-0.017789196	-0.011399306	0.004828083	-0.0105775	0.025905877	0.998332863	0.314232968	1	
DEF_60_CNT_SOCIAL_CIRCLE	-0.0074759	0.0445953	0.003891976	-0.007405962	-0.0207525	-0.02486256	-0.021412552	0.005560435	-0.002781782	0.014769193	-0.00688787	-0.004480448	0.236126904	0.856262392	0.238517486	1

Sumit S. C

Correlation of factors 1 HD Image (Considering ID)

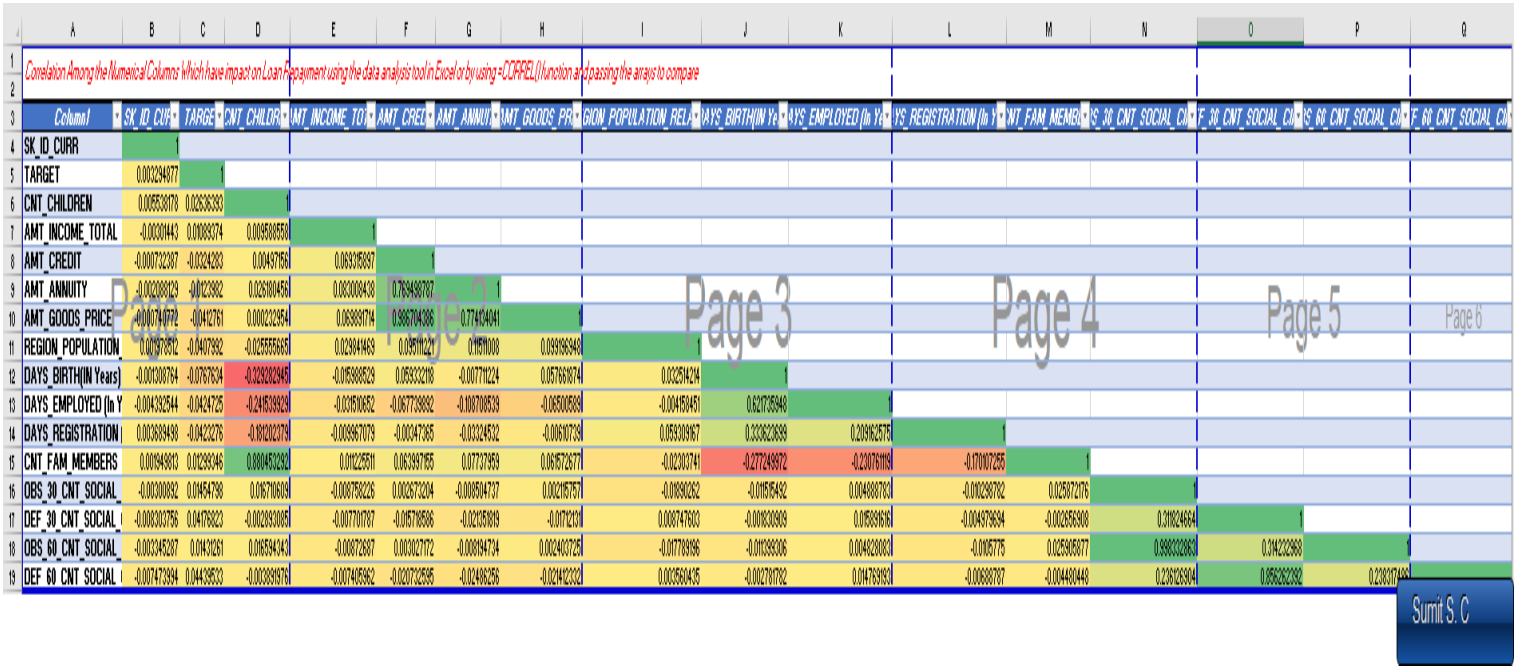


Correlation between Targets

Column	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH(IN Years)	DAYS_EMPLOYED (In Years)	DAYS_REGISTRATION (In Years)	CNT_FAM_MEMBERS	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE
TARGET	1														
CNT_CHILDREN	0.026	1													
AMT_INCOME_TOTAL	0.010	0.0095	1												
AMT_CREDIT	0.032	0.0049	0.0693158	1											
AMT_ANNUITY	0.012	0.0261	0.0830884	0.769	1										
AMT_GOODS_PRICE	0.041	0.0002	0.0698917	0.986	0.7741	1									
REGION_POPULATION_RELATIVE	0.040	-0.0255	0.0298414	0.095	0.1151	0.099196	1								
DAYS_BIRTH(IN Years)	-0.076	0.3292	0.0159885	-0.059	-0.0077	0.057661	0.032514214	1							
DAYS_EMPLOYED (In Years)	0.042	-0.2415	0.0315106	0.067	-0.1087	-0.065005	-0.004158451	0.6217	1						
DAYS_REGISTRATION (In Years)	0.042	-0.1812	0.0099670	-0.003	-0.0332	-0.006107	0.059309167	0.3336	0.20916	1					
CNT_FAM_MEMBERS	0.012	0.8804	0.0112255	0.063	0.0773	0.061572	-0.02303741	-0.2772	-0.23076	-0.170107255	1				
OBS_30_CNT_SOCIAL_CIRCLE	0.014	-0.0167	0.0087582	-0.020	-0.0085	-0.002115	-0.01890262	-0.0115	-0.00488	0.025872176	0.025872176	1			
DEF_30_CNT_SOCIAL_CIRCLE	0.041	-0.0028	0.0077017	-0.015	-0.0213	-0.017121	0.008747603	-0.0018	-0.01589	-0.004979694	-0.002656908	0.311824664	1		
OBS_60_CNT_SOCIAL_CIRCLE	0.014	-0.0165	0.0087368	-0.003	-0.0081	-0.002403	-0.017789196	-0.0113	-0.00482	-0.0105775	0.025905877	0.998332863	0.314232968	1	
DEF_60_CNT_SOCIAL_CIRCLE	0.044	-0.0038	0.0074059	-0.020	-0.0248	-0.021412	0.003560435	-0.0027	-0.01476	-0.00688787	-0.004480448	0.236126904	0.856262392	0.238517486	1

Sumit S. C

Correlation of factors 2 HD Image (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 , Gdocs Files
    - Loom Video PPT Q5 – Find the Correlation of scenario

Important Links :

Drive Folder Link

Individual Excel Sheets (Individual Questions File)

Final Excel sheet (cleaned dataset & all questions also present on drive folder page)

Word File Link & Pdf File (Will Be in the drive folder – can't add before I upload the file)

Video Presentation(Loom Folder 6 Videos)

Zip Link (Download these to extract the pdf and excel sheet files)

Thank You
