trainity

PROJECT 6

BY ASHISH KUMAR SAMANTARAY

Ashish Kumar Samantaray

B.Tech, Computer Science and Engineering

ashish.kumar.samantaray2003@gmail.com

7205691104

HYPERLINK

OF EXCEL SHEET

Click here to get the working file

PROJECT DESCRIPTION:

This project requires us 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.

TECH STACK USED:

MICROSOFT EXCEL
CANVA FOR CREATING PPT

I chose **Microsoft Excel** because it is thw most convenient spreadhseet and can be used efficiently to view statistics and analyse the data set given very quickly.

I chose **Canva** so as to make my PPT look more visually appealing.

insights AHEAD

WITH DETAILED APPROACH AND OUTPUT AND FORMULA BOX
(GRAPH IF ASKED)

First I added both the work books in the same working file so that it would be easier for me to get transferred to another sheet and not open a whole new workbook for doing other task of the other sheet. I made this possible by importing data as a csv file and then formatting it as a table to make it possible for me to make a pivot table for the same.

TASK **A**

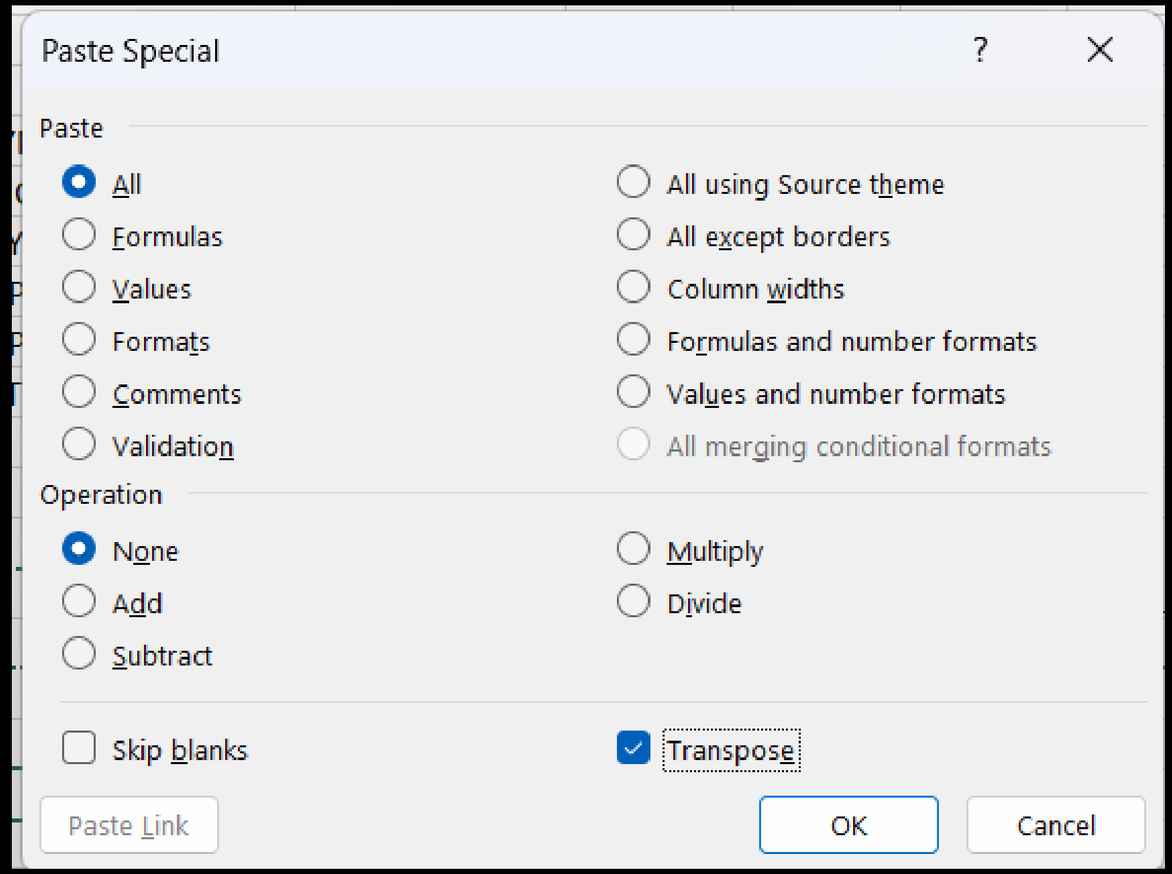
Missing Data Analysis

TASK A

First we used the **countblank** formula so as to find the number of BLANKS in each field of the tables given to us as the dataset. The formula for the same is given below.

Since formula was created for one field, it was extended to other columns and later the table consisting of count of blanks and name of the field was copied and pasted as the transpose of the table and formatted as a table so as to get the headers. Finally we carried out the Graph operations with the tables. Next are the tables along with the graphs.

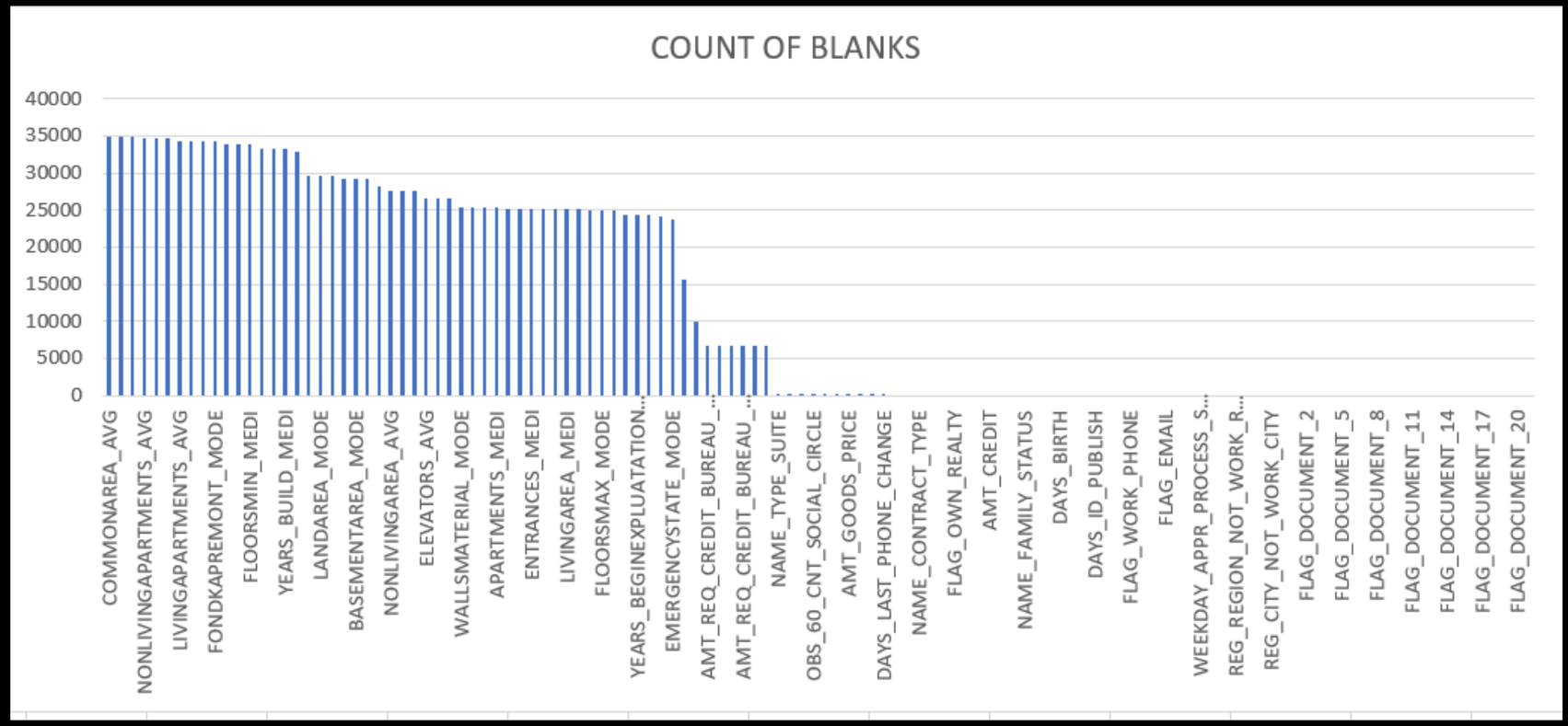
AMT_REQ_CREDIT_BUREAU_HOUR
=COUNTBLANK(DM2:DM50000)



TASK A

FIELDS	COUNT OF BLANKS	PERCENTAGE
COMMONAREA_AVG	34960	69.92139843
COMMONAREA_MODE	34960	69.92139843
COMMONAREA_MEDI	34960	69.92139843
NONLIVINGAPARTMENTS_AVG	34714	69.42938859
NONLIVINGAPARTMENTS_MODE	34714	69.42938859
NONLIVINGAPARTMENTS_MEDI	34714	69.42938859
LIVINGAPARTMENTS_AVG	34226	68.45336907
LIVINGAPARTMENTS_MODE	34226	68.45336907
LIVINGAPARTMENTS_MEDI	34226	68.45336907
FONDKAPREMONT_MODE	34191	68.38336767
FLOORSMIN_AVG	33894	67.78935579
FLOORSMIN_MODE	33894	67.78935579
FLOORSMIN_MEDI	33894	67.78935579
YEARS_BUILD_AVG	33239	66.47932959
YEARS_BUILD_MODE	33239	66.47932959
YEARS_BUILD_MEDI	33239	66.47932959
OWN_CAR_AGE	32950	65.90131803
LANDAREA_AVG	29721	59.44318886
LANDAREA_MODE	29721	59.44318886
LANDAREA_MEDI	29721	59.44318886
BASEMENTAREA_AVG	29199	58.39916798
BASEMENTAREA_MODE	29199	58.39916798
BASEMENTAREA_MEDI	29199	58.39916798
EXT_SOURCE_1	28172	56.3451269
NONLIVINGAREA_AVG	27572	55.1451029
NONLIVINGAREA_MODE	27572	55.1451029
NONLIVINGAREA_MEDI	27572	55.1451029
ELEVATORS_AVG	26651	53.30306606

***Full Table in the Working file



TASK A

After analysing the table informing the percentage of null values in the dataset, we decided that the fields having more than or equal to **50** percent are to be dropped because if they are manipulated, then the overall dataset may be wrongly influenced through the highly present median values; fields having null values less than **1** percent can be ignore as they are negligible and wont affect the data set so much and the **0s** being left out as it is. The remaining are to be filled with appropriate data.

NAME_TYPE_SUITE	192	0.38400768
OBS_30_CNT_SOCIAL_CIRCLE	168	0.33600672
DEF_30_CNT_SOCIAL_CIRCLE	168	0.33600672
OBS_60_CNT_SOCIAL_CIRCLE	168	0.33600672
DEF_60_CNT_SOCIAL_CIRCLE	168	0.33600672
EXT_SOURCE_2	126	0.25200504
AMT_GOODS_PRICE	38	0.07600152
AMT_ANNUITY	1	0.00200004
CNT_FAM_MEMBERS	1	0.00200004

Light Green for lower percentage (Undisturbed)

Light yellow for Mid values (Manipulated)

FLOORSMAX_AVG	24875	49.75099502
FLOORSMAX_MODE	24875	49.75099502
FLOORSMAX_MEDI	24875	49.75099502
YEARS_BEGINEXPLUATATION_AVG	24394	48.78897578
YEARS_BEGINEXPLUATATION_MODE	24394	48.78897578
YEARS_BEGINEXPLUATATION_MEDI	24394	48.78897578
TOTALAREA_MODE	24148	48.29696594
EMERGENCYSTATE_MODE	23698	47.39694794
OCCUPATION_TYPE	15654	31.30862617

Full Table in Working file

FIELDS	COUNT OF BLANK	PERCENTAGE
COMMONAREA_AVG	34960	69.92139843
COMMONAREA_MODE	34960	69.92139843
COMMONAREA_MEDI	34960	69.92139843
NONLIVINGAPARTMENTS_AVG	34714	69.42938859
NONLIVINGAPARTMENTS_MODE	34714	69.42938859
NONLIVINGAPARTMENTS_MEDI	34714	69.42938859
LIVINGAPARTMENTS_AVG	34226	68.45336907
LIVINGAPARTMENTS_MODE	34226	68.45336907

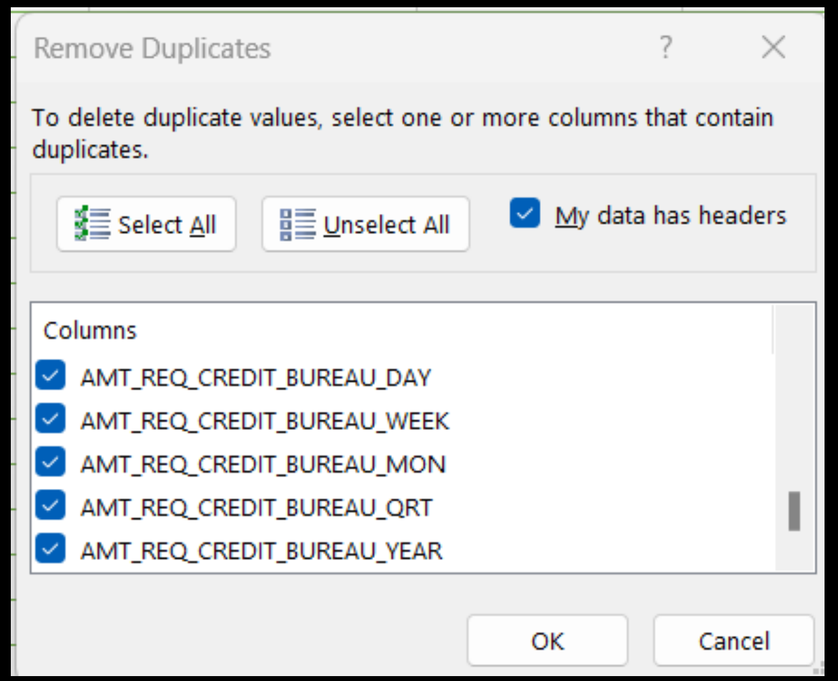
Red for high percentage (Dropped)

Green for 0 category (Undisturbed)

SK_ID_CURR	0	0
TARGET	0	0
NAME_CONTRACT_TYPE	0	0
CODE_GENDER	0	0
FLAG_OWN_CAR	0	0
FLAG_OWN_REALTY	0	0
CNT_CHILDREN	0	0
AMT_INCOME_TOTAL	0	0
AMT_CREDIT	0	0

TASK A

Additionally, We also remove the duplicates by using the inbuilt function of excel of removing duplicate rows (If found).



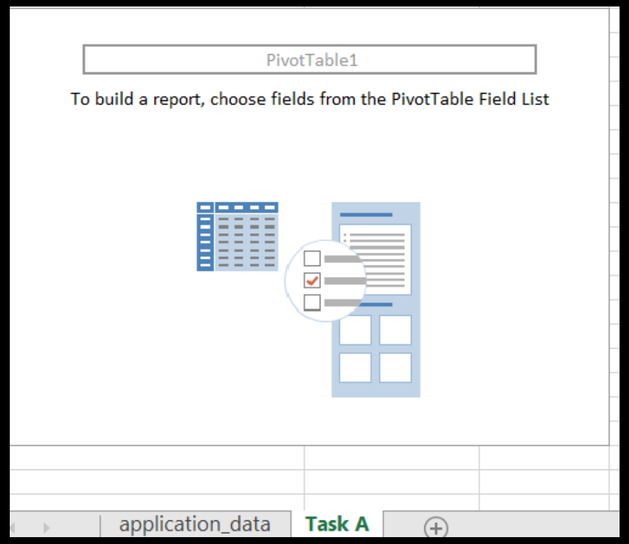
TASK A

Further, there are some columns such as **Total_Area_Mode**,**Floor_Max**, **EmergencyState_Mode** , etc which are to be dropped essentially so as to make the dataset more compact and so as to make it easily being analysed and further the fields are out of the scope of the analysis.

FLOORSMAX_AVG	24875	49.75099502
FLOORSMAX_MODE	24875	49.75099502
FLOORSMAX_MEDI	24875	49.75099502
YEARS_BEGINEXPLUATATION_AVG	24394	48.78897578
YEARS_BEGINEXPLUATATION_MODE	24394	48.78897578
YEARS_BEGINEXPLUATATION_MEDI	24394	48.78897578
TOTALAREA_MODE	24148	48.29696594

TASK A

Then a Pivot Table is created now to calculate the mean and median and mode of the various fields so that we could work upon replacing the fields with appropriate data so that dataset is not affected and is filled perfectly.



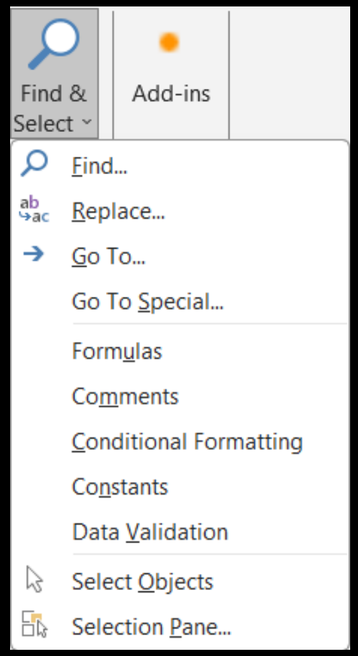
TASK A

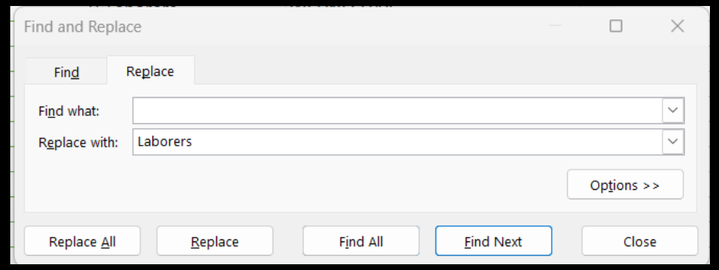
Row Labels	Count of OCCUPATION_TYPE
Laborers	8952
Sales staff	5160
Core staff	4434
Managers	3489
Drivers	3044
High skill tech staff	1852
Accountants	1621
Medicine staff	1403
Security staff	1140
Cooking staff	963
Cleaning staff	739
Private service staff	447
Low-skill Laborers	357
Waiters/barmen staff	228
Secretaries	212
Realty agents	123
HR staff	101
IT staff	80
(blank)	
Grand Total	**34345**

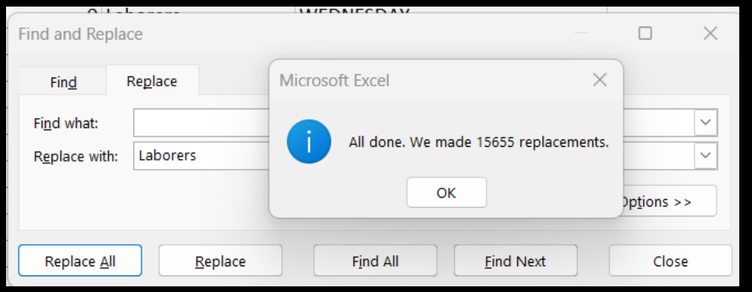
Then we categorize the occupation type and then we count the number of occupation type in total and observe that the labourers is basically the **mode** of the occupation type.

Thus, we can replace the blanks with labourers so that the data set get stretched to the mode and does get outlied.

TASK A







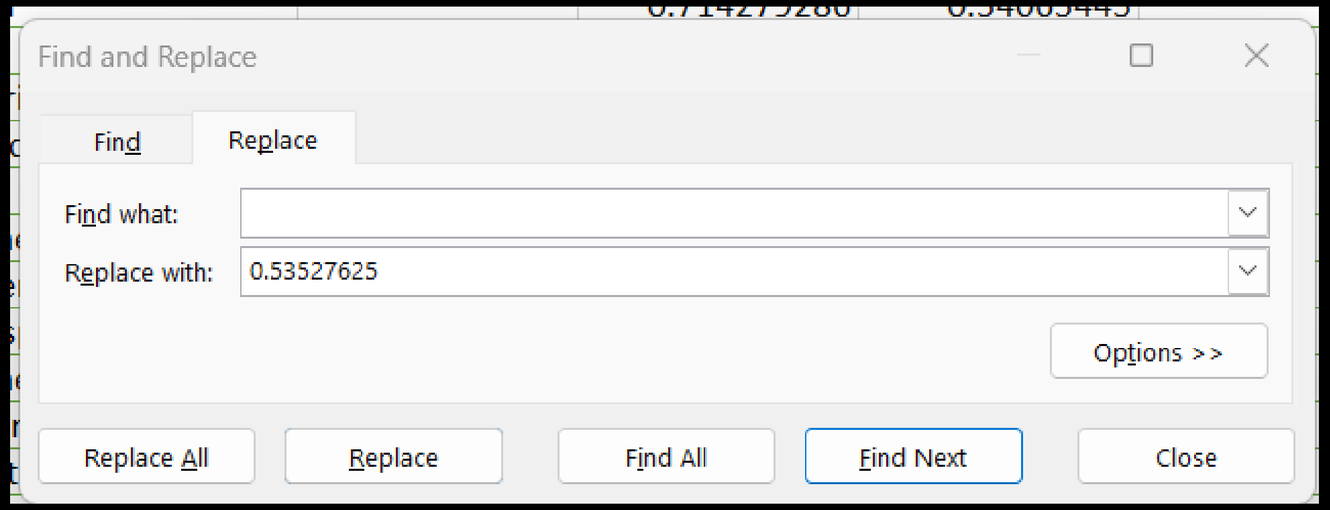
Now we replace the **blanks** with the **laborers** value by clickling find and select in editing section of home tab and then click on replace and then leaving the *find what* blank and then writing laborers with *replace with* section and click on replace all.

TASK A

Now we calculated the median of EXT_SOURCE_3 and decided to replace it throughout the blanks of that field.

```
=MEDIAN(AN2:AN50000)
```

Then we performed the **replacement** function with the whole column and the blanks were finally replaced with the median of the field. Similarly we worked out with other field eliminating the blank cells.



TASK B

Outliers Analysis

TASK B

First we created the pivot table and get the row label as the SK_ID and the sum of AMT_CREDIT. Further after creating the pivot table, the table was copied and values were pasted as it was not possible to create a scatter plot with the pivot table. Now we created a scatter plot as well as box pot to get the extent of scattering of the data points during the analysis of the data set.

Next are the tables and the pivot table and the box plot and scattered plot along with the marked outliers.

TASK B

Pivot table

Row Labels	Sum of AMT_CREDIT
100002	406597.5
100003	1293502.5
100004	135000
100006	312682.5
100007	513000
100008	490495.5
100009	1560726
100010	1530000
100011	1019610
100012	405000
100014	652500
100015	148365
100016	80865
100017	918468
100018	773680.5
100019	299772
100020	509602.5
100021	270000
100022	157500
100023	544491
100024	427500
100025	1132573.5
100026	497520
100027	239850
100029	247500
100030	225000
100031	979992
100032	327024
100033	790830
100034	180000
100035	665892
100036	512064

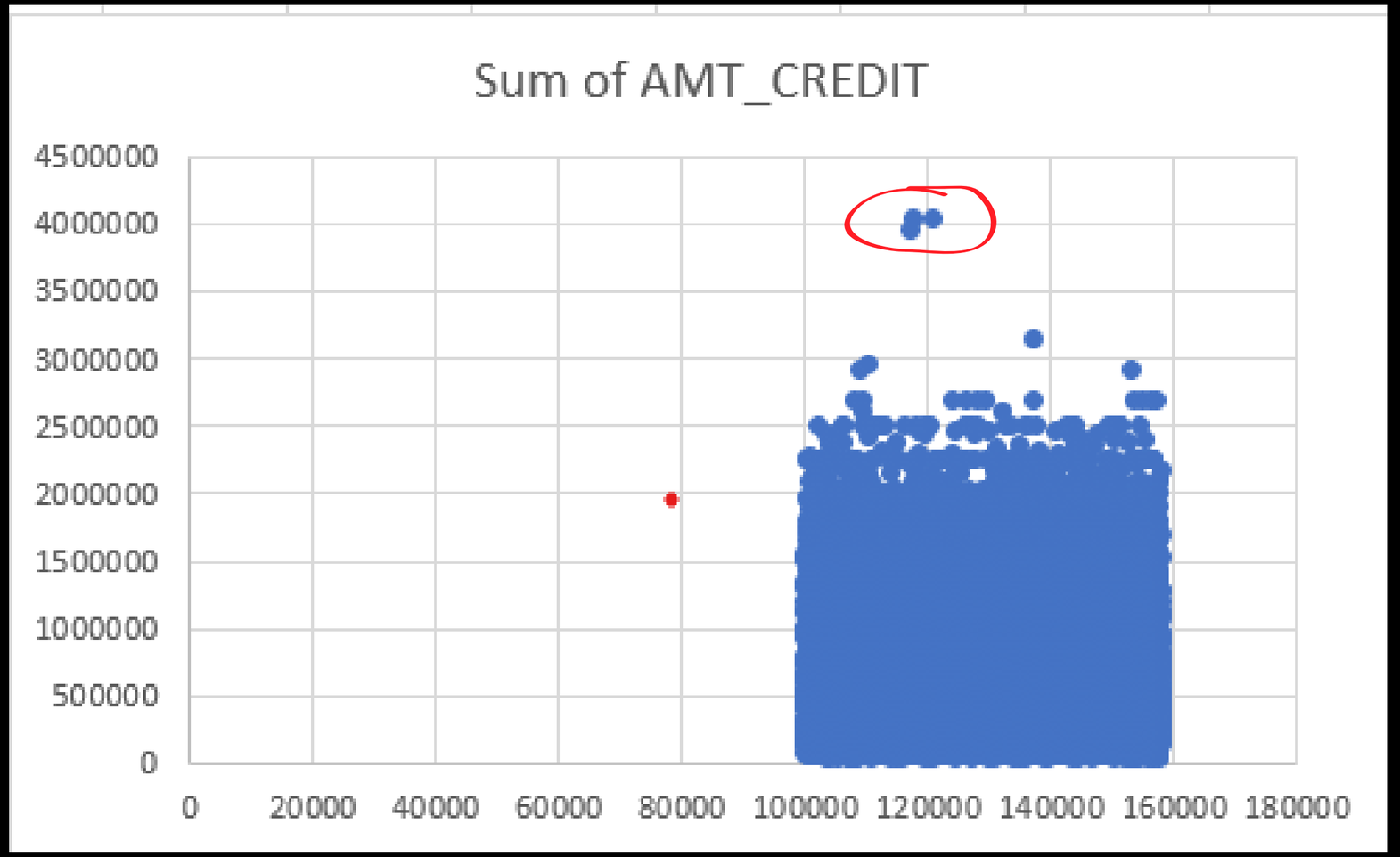
Copied table

Row Labels	Sum of AMT_CREDIT
100002	406597.5
100003	1293502.5
100004	135000
100006	312682.5
100007	513000
100008	490495.5
100009	1560726
100010	1530000
100011	1019610
100012	405000
100014	652500
100015	148365
100016	80865
100017	918468
100018	773680.5
100019	299772
100020	509602.5
100021	270000
100022	157500
100023	544491
100024	427500
100025	1132573.5
100026	497520
100027	239850
100029	247500
100030	225000
100031	979992
100032	327024
100033	790830
100034	180000
100035	665892
100036	512064

TASK B

Red mark are the outliers

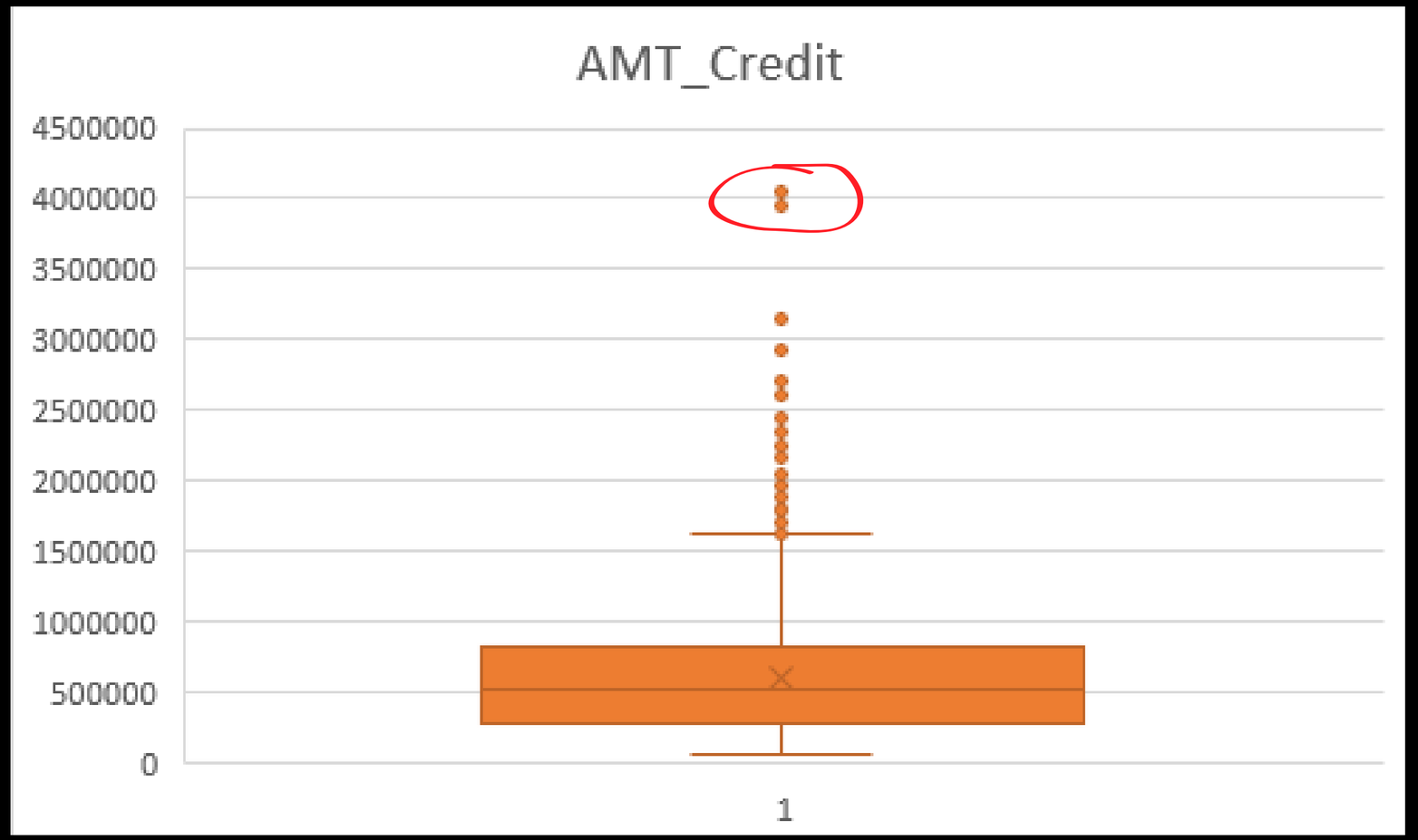
Scatter Plot



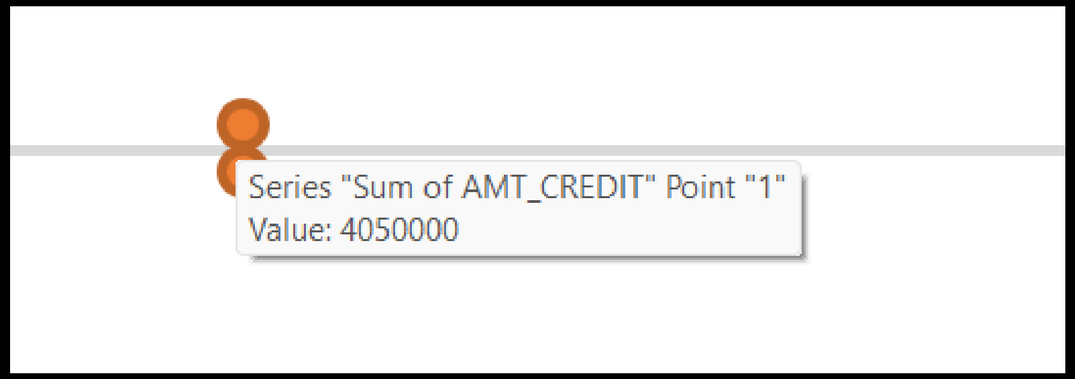
TASK B

Red mark are the outliers

Box Plot

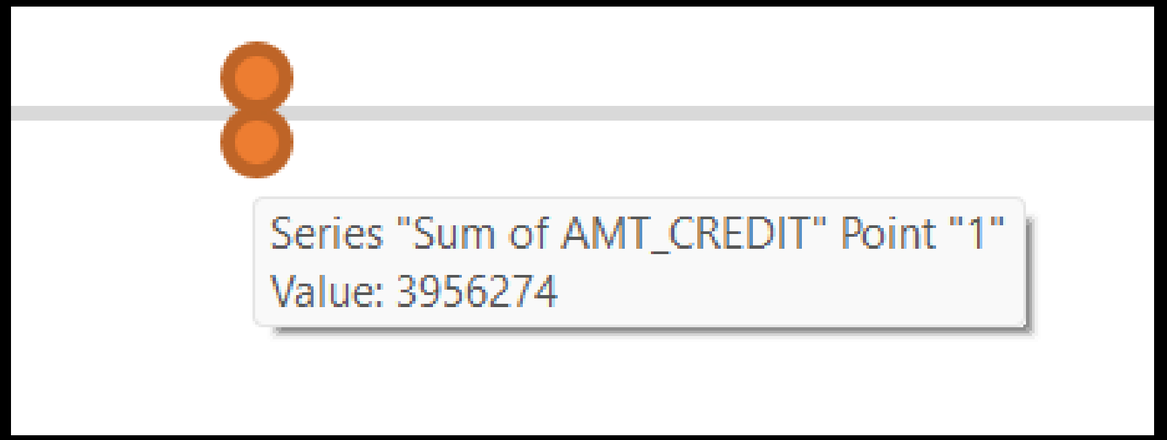


TASK B



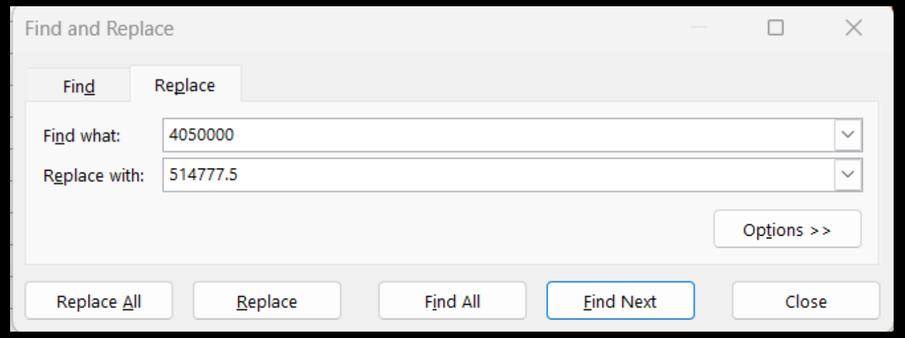
Outlier Value 1

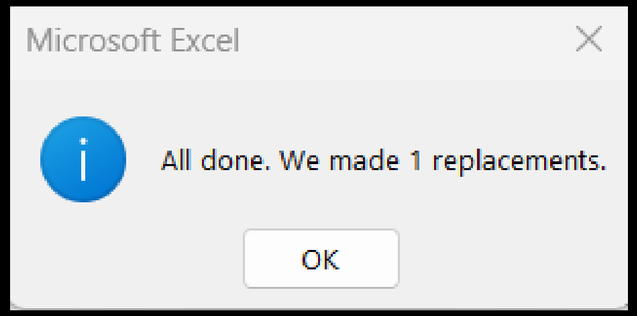
Outlier Value 2

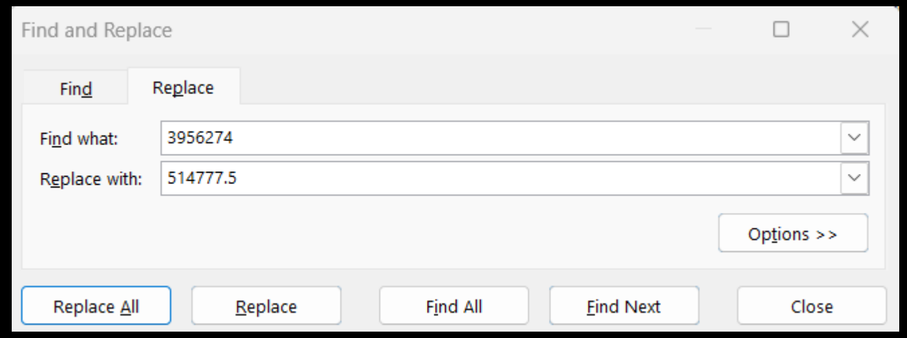


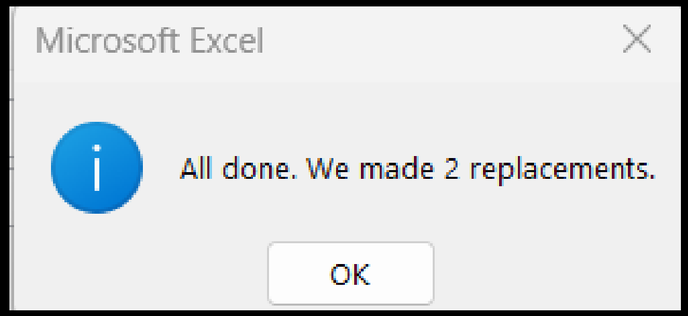
TASK B

Now as we identified the **outliers**, we decided to get the values replaced so as to get the dataset settled and keep the median mode and mean of the dataset consistent.





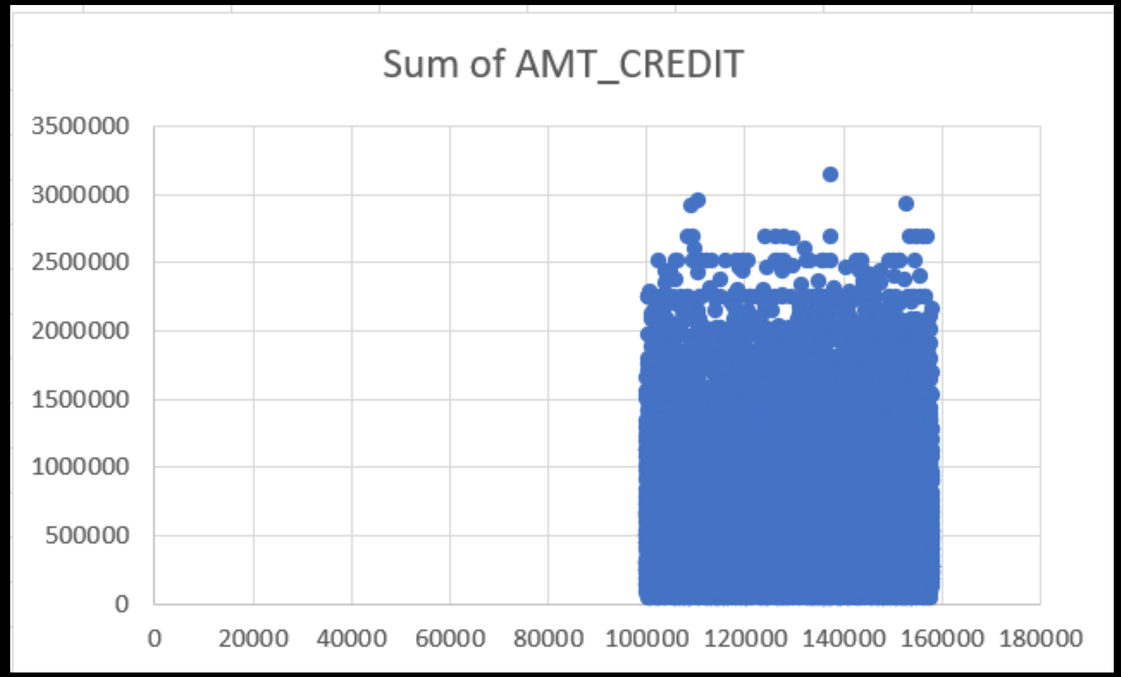


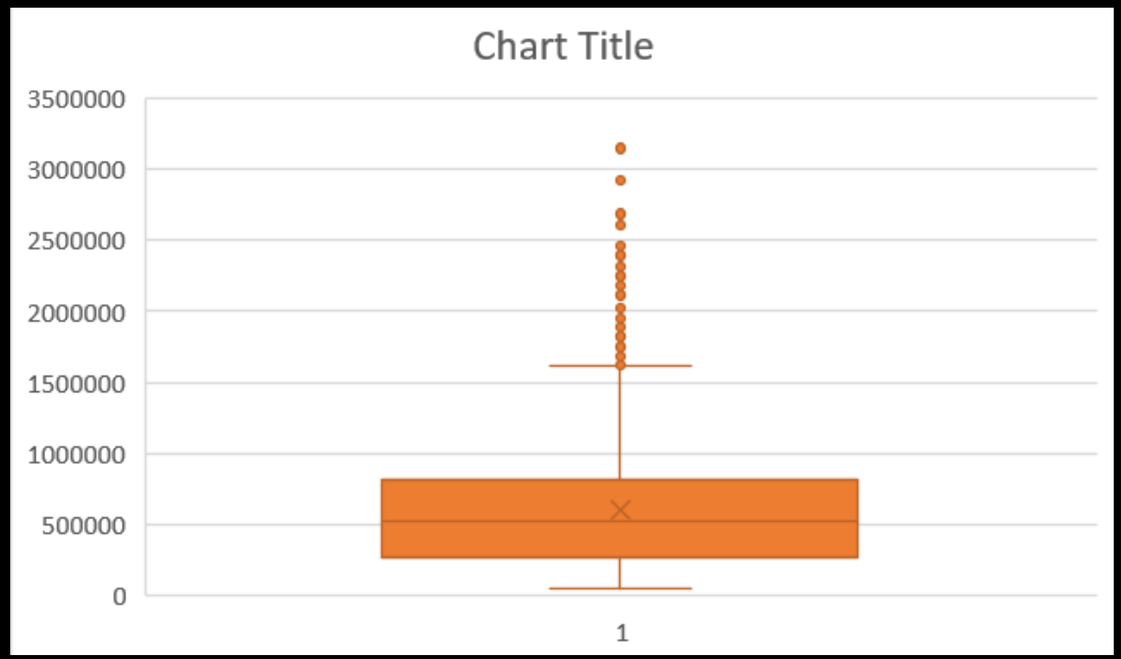


TASK B

Row Labels	Sum of AMT_CREDIT
100002	406597.5
100003	1293502.5
100004	135000
100006	312682.5
100007	513000
100008	490495.5
100009	1560726
100010	1530000
100011	1019610
100012	405000
100014	652500
100015	148365
100016	80865
100017	918468
100018	773680.5
100019	299772
100020	509602.5
100021	270000
100022	157500
100023	544491
100024	427500
100025	1132573.5

Updated Tables and Graphs after replacing **outliers**





TASK B

Then we performed the **quartile ANALYSIS** so as to divide the data into four quart parts.

Below are the quart table and the table before modification.
(Before changing outliers)

QUARTILE ANALYSIS	
MIN	45000
1ST	270000
2ND	514778
3RD	808650
MAX	4050000

Row Labels	Sum of AMT_CREDIT
117337	4050000
120926	4050000
117085	3956274
137220	3150000
110403	2961000
152866	2931660
108906	2925000
154804	2700000
156385	2700000
108224	2695500
109450	2695500

127306	808650
127401	808650
127405	808650
127416	808650
127548	808650
127698	808650
127704	808650
127722	808650
127890	808650
127899	808650
127902	808650

114537	294322.5
115245	294322.5
117095	294322.5
118533	294322.5
118783	294322.5
120585	294322.5
125621	294322.5
134898	294322.5
135098	294322.5

139173	45000
143626	45000
144432	45000
146605	45000
150491	45000
152058	45000
152670	45000
153225	45000
153845	45000

TASK B

Below are the quart table and the table after modification. (after changing outliers)

***FULL DATA TABLE IN WORKFILE

QUARTILE ANALYSIS	
MIN	45000
1ST	270000
2ND	514778
3RD	808650
MAX	3150000

Row Labels	Sum of AMT_CREDIT
100002	406597.5
100003	1293502.5
100004	135000
100006	312682.5
100007	513000
100008	490495.5
100009	1560726
100010	1530000
100011	1019610
100012	405000
100014	652500
100015	148365
100016	80865
100017	918468
100018	773680.5
100019	299772
100020	509602.5
100021	270000
100022	157500
100023	544491
100024	427500
100025	1132573.5
100026	497520
100027	239850
100029	247500
100030	225000
100031	979992

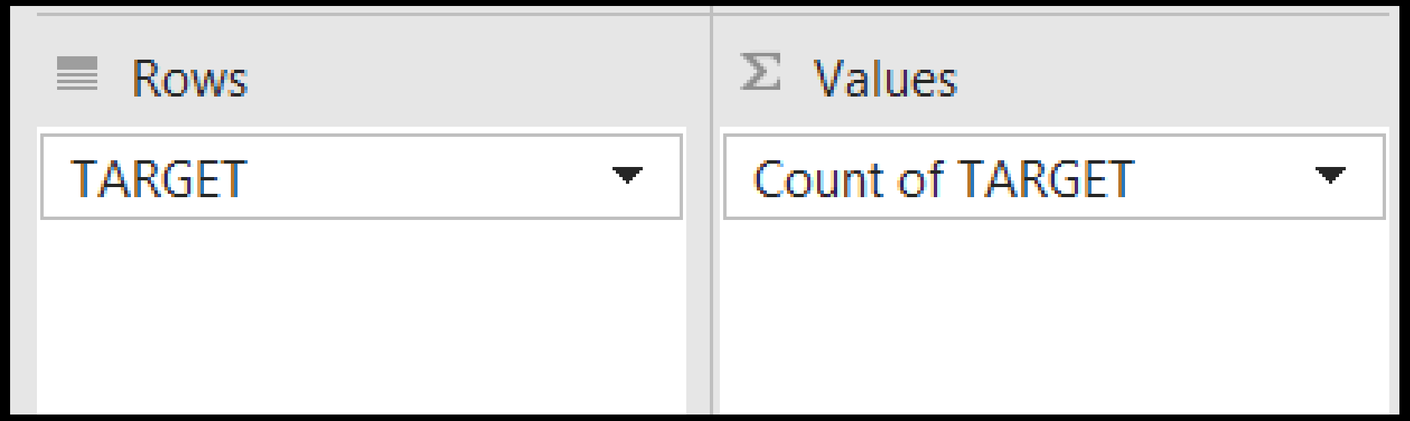
TASK C

Data Imbalance Analysis

TASK C

First we created a pivot table so that we could easily analyse or classify a particular type of data. Here we have to take the Target Variable as the field as mentioned in the task heading and perform the task according to the outcome in pivot table.

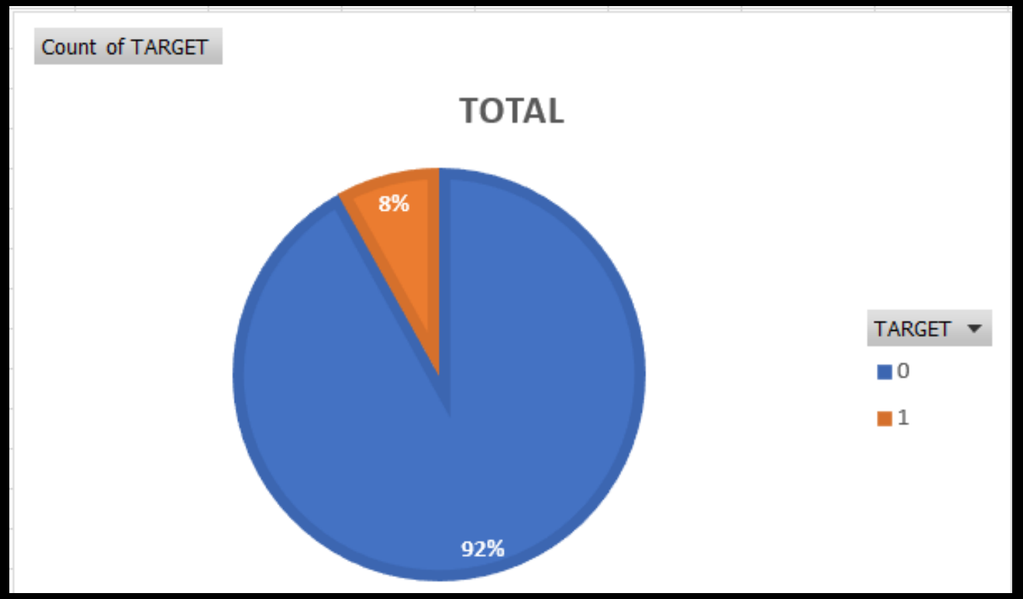
Thus we selected Target in row and values respectively as seen below.

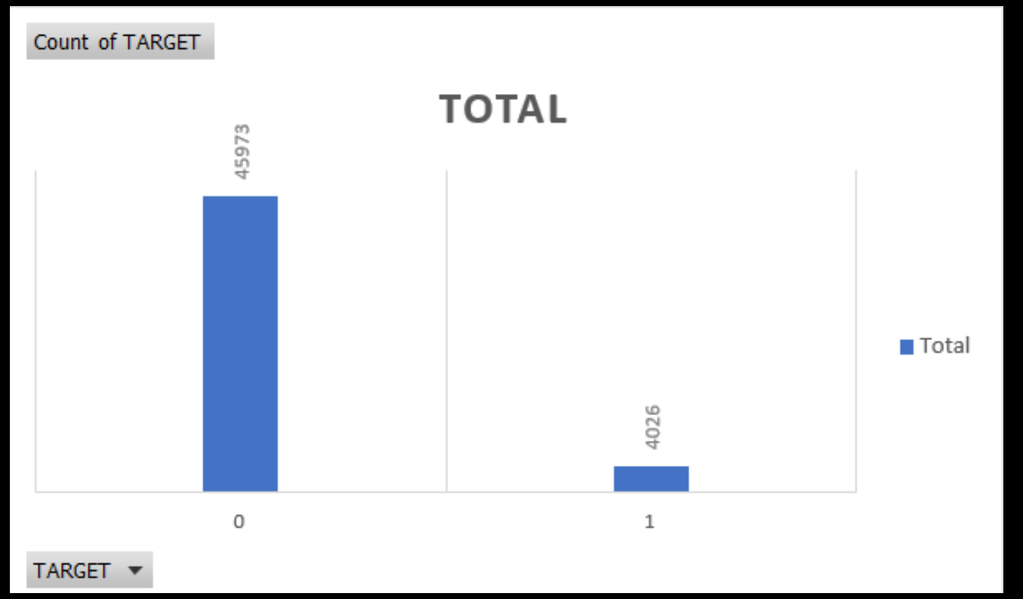


TASK C

Then we created a pie chart as well as bargraph for the table we received in the pivot table. Below are the Pie chart, Bar Graph and the relative table.

Row Labels	Count of TARGET	Percenta
0	45973	91.94783896
1	4026	8.052161043
Grand Total	**49999**	





TASK C

By seeing the ratio of the imbalance in the dataset of target analysis, It is not a suitable field to be used as a representative. Thus we chose Name Family Status as the represtative field now.

RATIO
45973:4026

Now we take **Name Family Status** as the dataset and construct the bargraph and piechart.

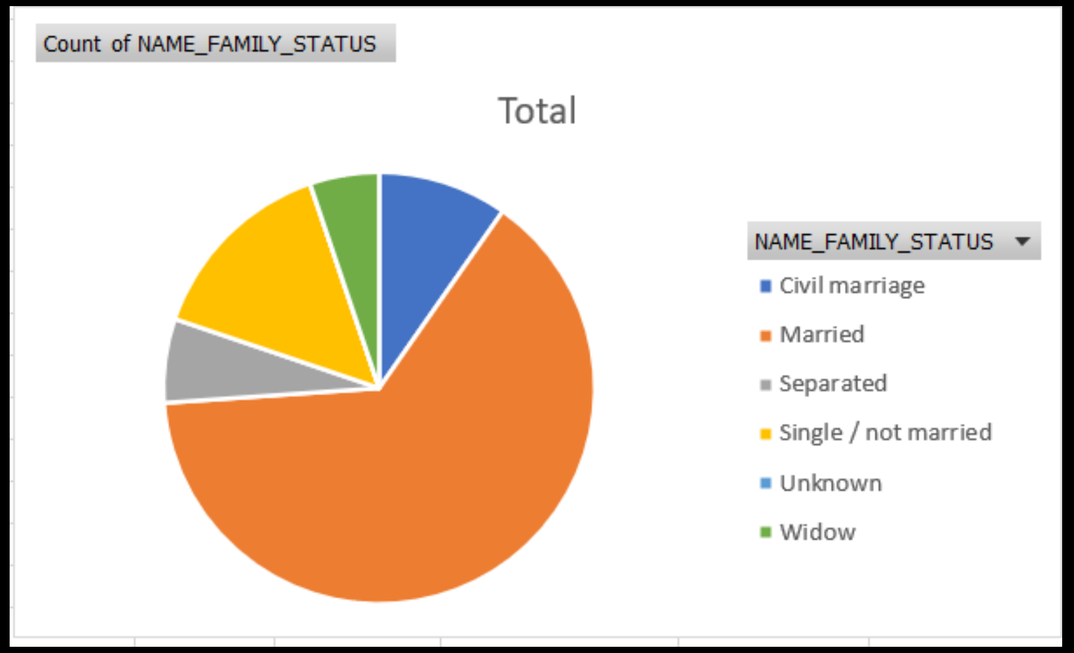
Graph and table is in the next page.

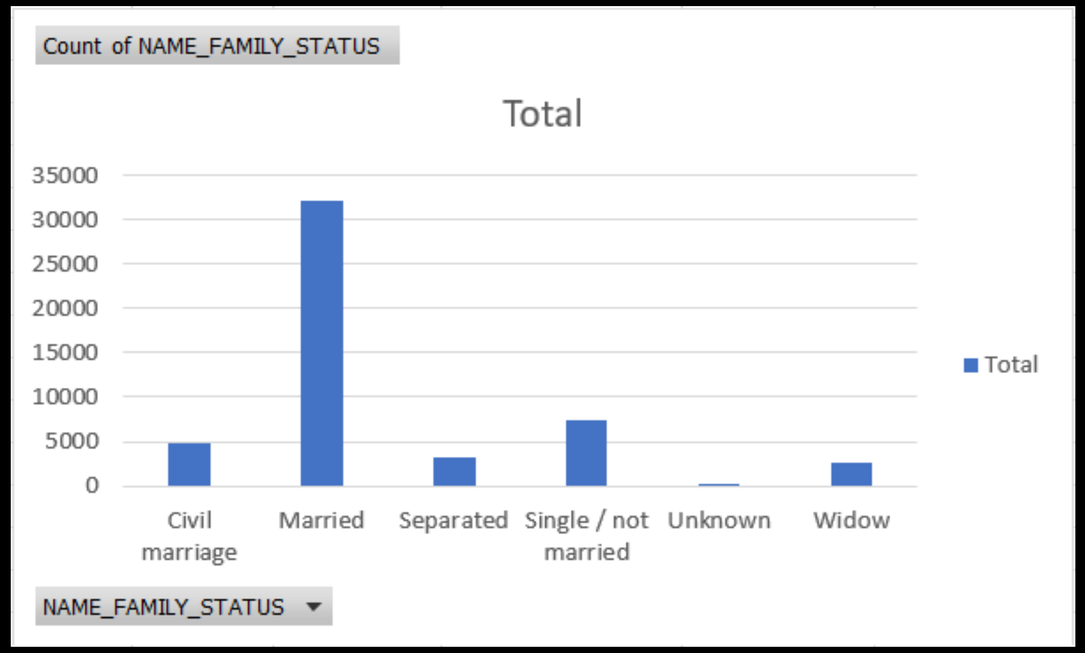
TASK C

Pie Chart and Bar Graph and Table
for
Name Family Status

Classification is Diversified thus suitable for Analysis. (though not binary)

Row Labels	Count of NAME_FAMILY_STATUS
Civil marriage	4859
Married	32094
Separated	3142
Single / not married	7306
Unknown	1
Widow	2597
Grand Total	**49999**





TASK D

Univariate, Segmented Univariate, and Bivariate Analysis

TASK D

UNIVARIATE ANALYSIS

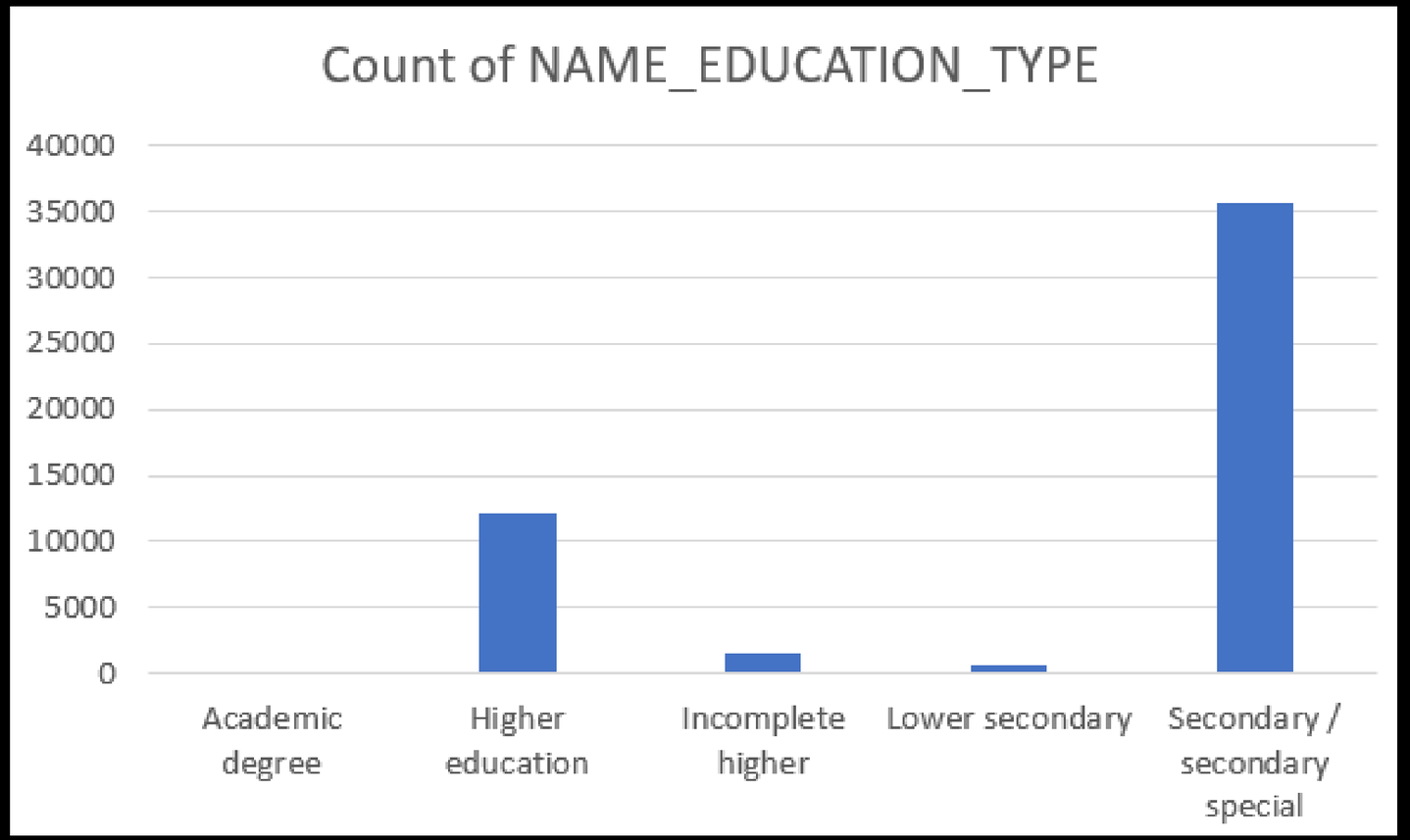
As the name suggests, **Univariate analysis** explores one variable in a data set, separately.

Next are examples of three univariate analysis performed in the working file of our data set.
(WorkfileAttached at the beginning)

TASK D

UniVariate Analysis I

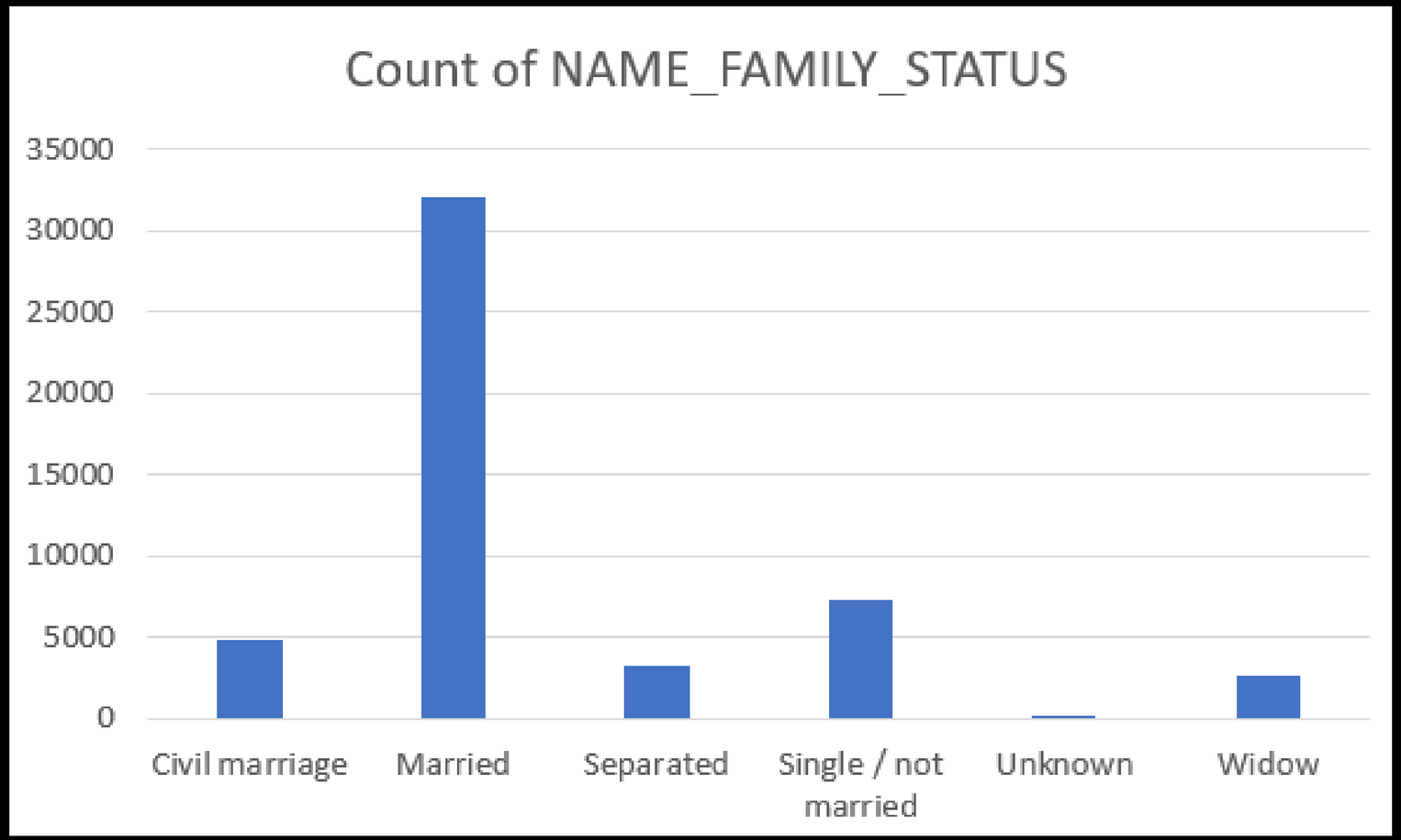
Row Labels	Count of NAME_EDUCATION_TYPE
Academic degree	20
Higher education	12167
Incomplete higher	1620
Lower secondary	620
Secondary / secondary special	35572



TASK D

UniVariate Analysis II

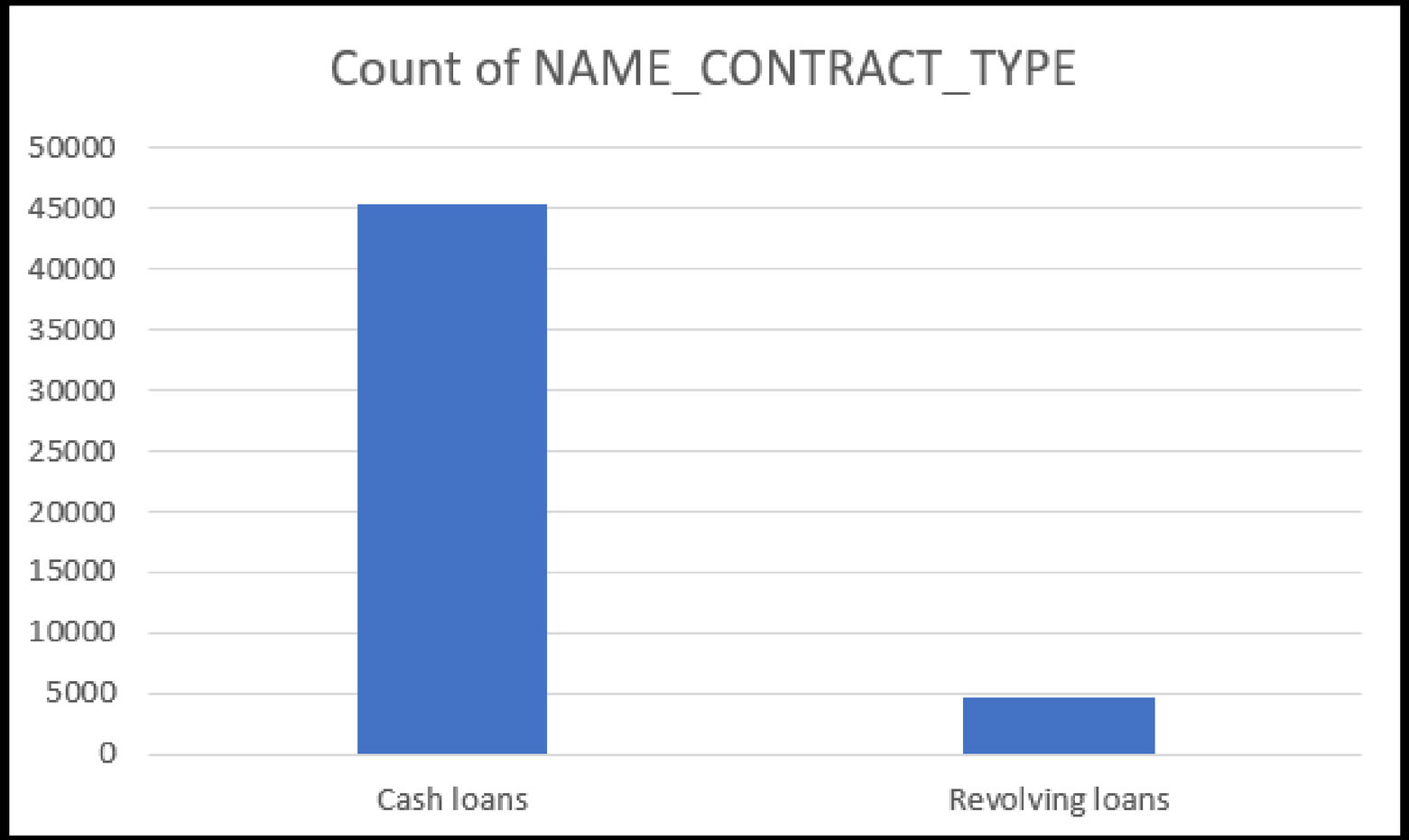
Row Labels	Count of NAME_FAMILY_STATUS
Civil marriage	4859
Married	32094
Separated	3142
Single / not married	7306
Unknown	1
Widow	2597



TASK D

UniVariate Analysis III

Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	45276
Revolving loans	4723



TASK D

BIVARIATE ANALYSIS

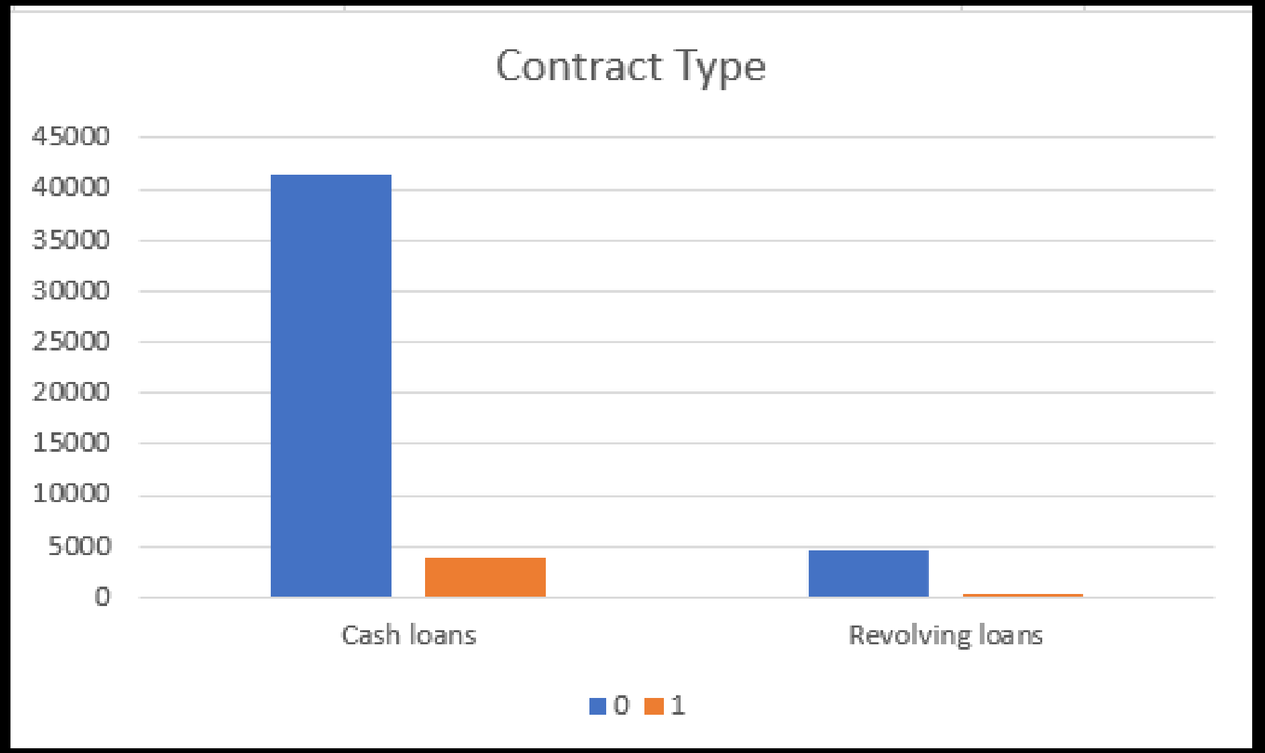
Bivariate analysis is stated to be an analysis of any **concurrent** relation between two variables or attributes.

Next are examples of three bivariate analysis performed in the working file of our data set.
(WorkfileAttached at the beginning)

TASK D

BiVariate Analysis I

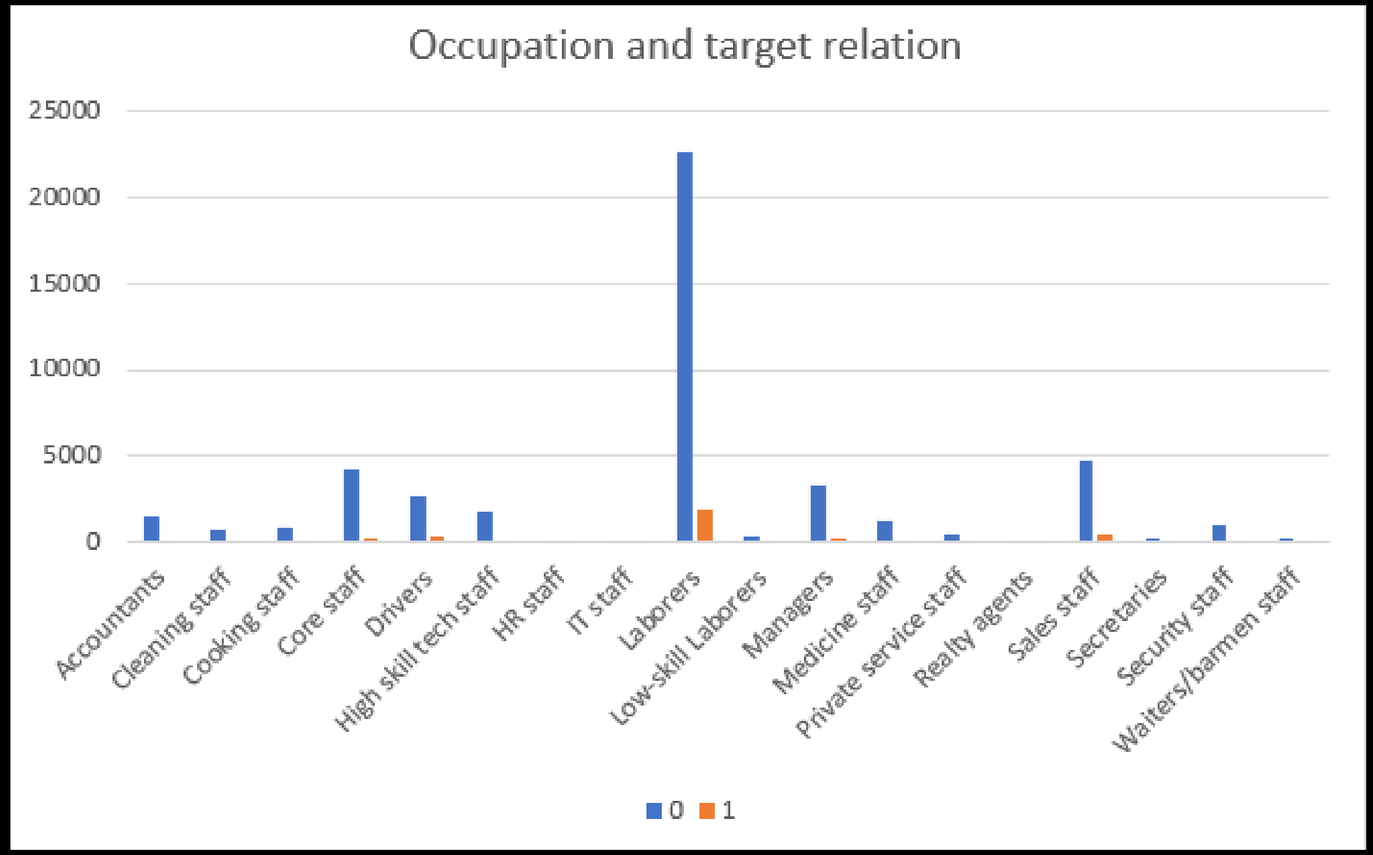
Row Labels	0	1	Grand Tota
Cash loans	41484	3792	45276
Revolving loans	4489	234	4723



TASK D

BiVariate Analysis II

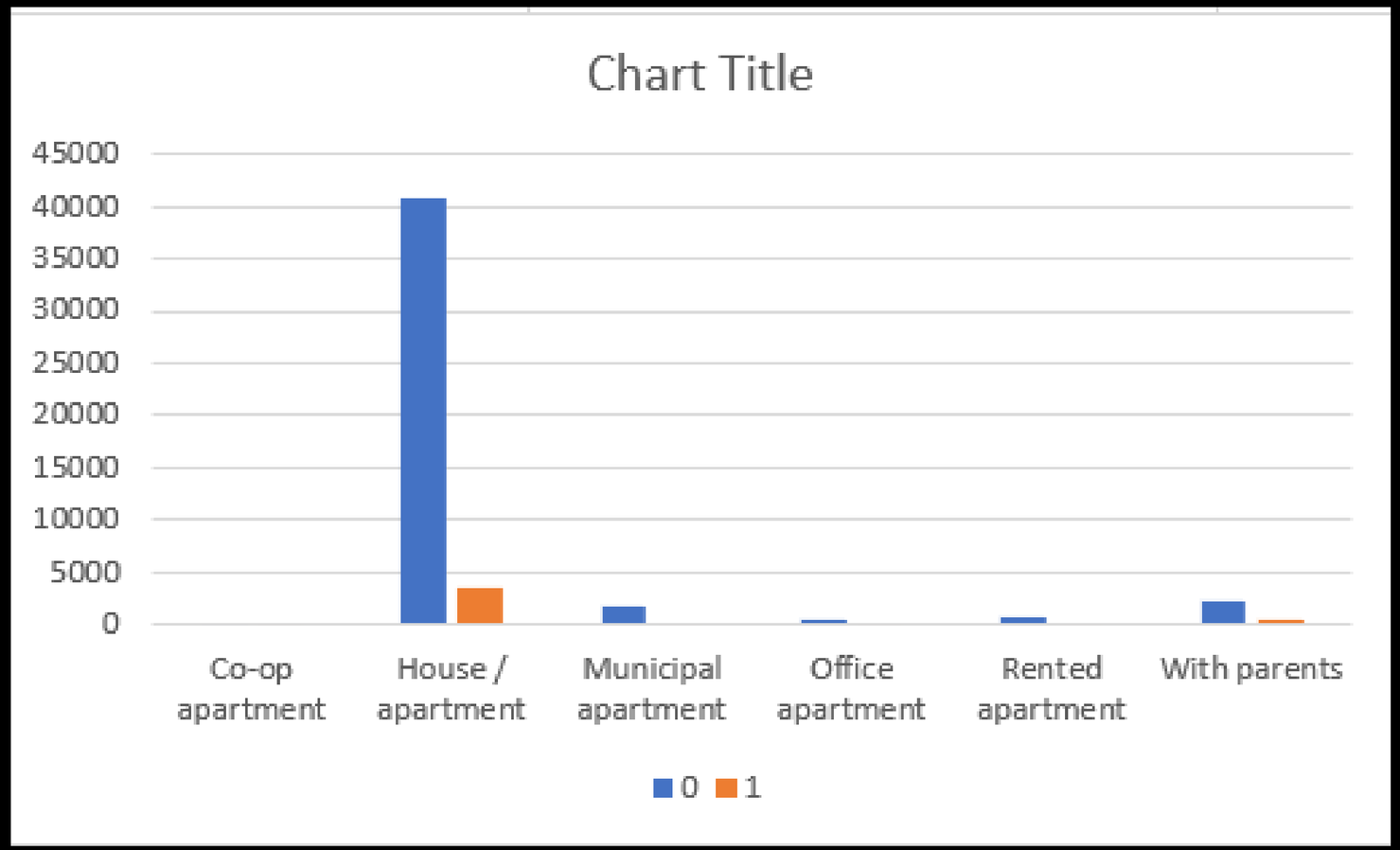
Row Labels	0	1
Accountants	1540	81
Cleaning staff	671	68
Cooking staff	862	101
Core staff	4184	250
Drivers	2706	338
High skill tech staff	1734	118
HR staff	92	9
IT staff	76	4
Laborers	22660	1946
Low-skill Laborers	296	61
Managers	3246	243
Medicine staff	1297	106
Private service staff	410	37
Realty agents	110	13
Sales staff	4668	492
Secretaries	203	9
Security staff	1015	125
Waiters/barmen staff	203	25



TASK D

BiVariate Analysis III

Row Labels	0	1
Co-op apartment	176	15
House / apartment	40895	3473
Municipal apartment	1700	145
Office apartment	398	29
Rented apartment	682	87
With parents	2122	277



TASK D

SEGMENTED BIVARIATE ANALYSIS

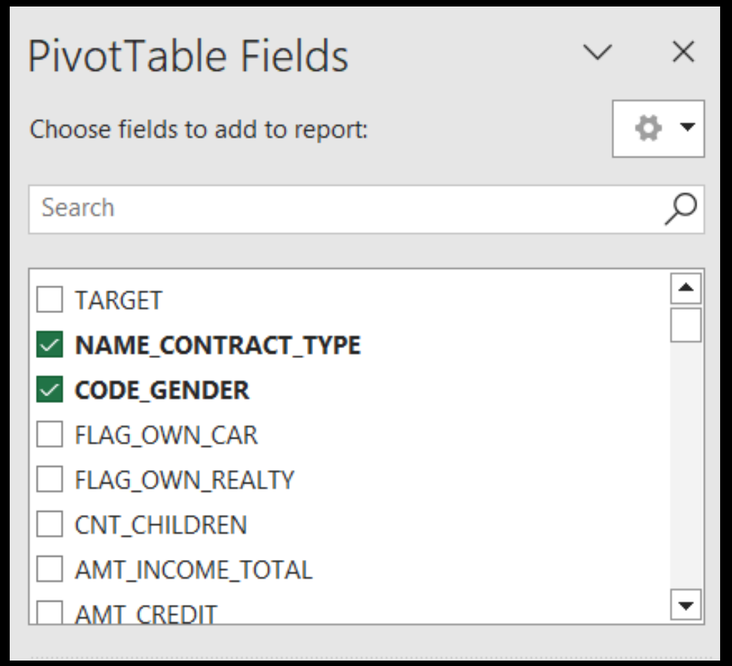
Segmented Bivariate analysis is one of the simplest form of visualization to analyze data.

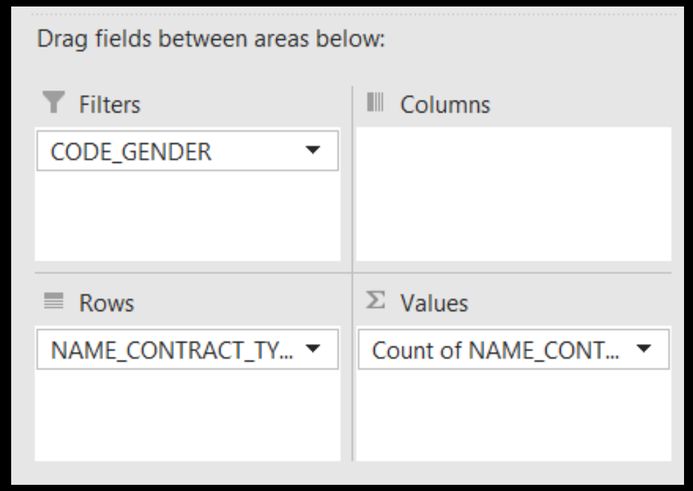
Next is an example of segmented univariate analysis performed in the working file of our data set.
(WorkfileAttached at the beginning)

TASK D

Segmented UniVariate Analysis

We chose the variable as the **name contract type** and then decided to segment it on the basis of the **code gender**, for which we kept the name contract type in row value and the code gender in column value.





TASK D

Segmented UniVariate Analysis

Then we filtered out the results on the basis of the gender on the filter section of the pivot table. And below are the tables we obtained.

Row Labels	Count of NAME_CONTRACT_TYPE
Revolving loans	2
Cash loans	0

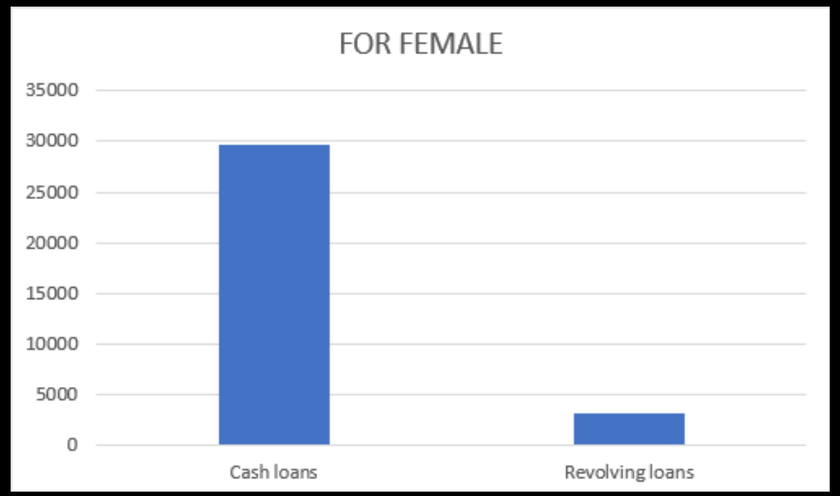
Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	15611
Revolving loans	1563
Grand Total	**17174**

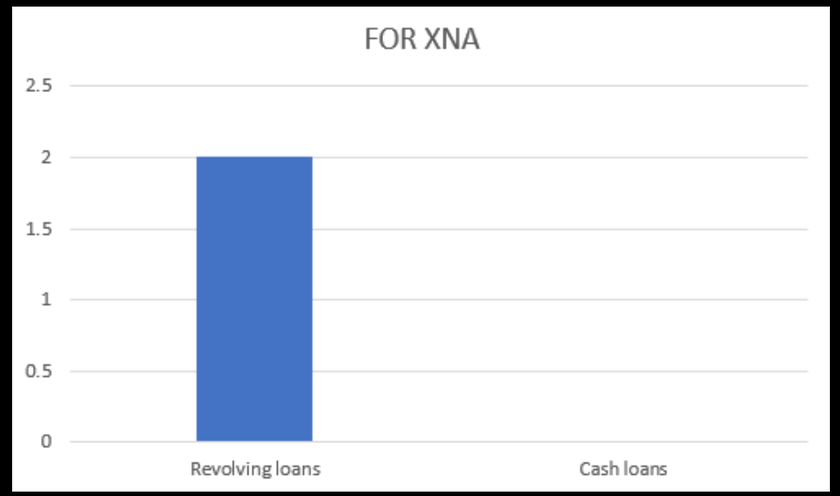
Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	29665
Revolving loans	3158
Grand Total	**32823**

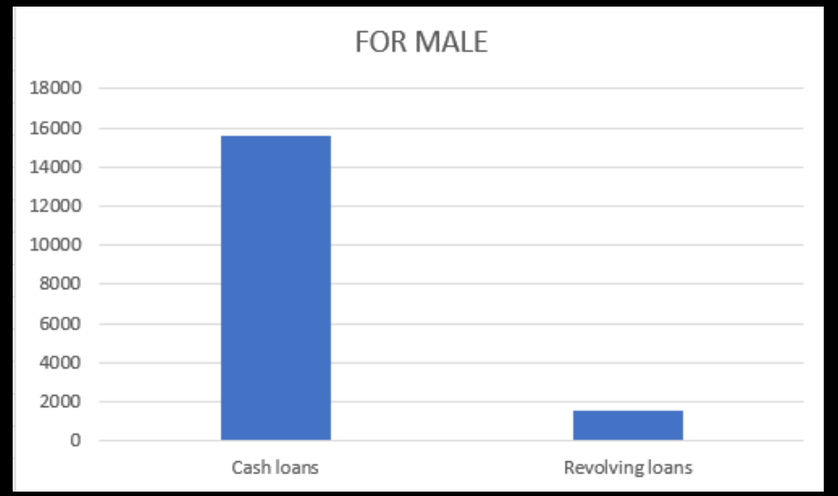
TASK D

Segmented UniVariate Analysis

Below are the graphs for the segmented analysis of the **NAME_CONTRACT_TYPE** on the basis of **CODE_GENDER**.







TASK E

Correlations Analysis

TASK E

For a correlation table, first we need to select the target variable. Here our target variable is **TARGET** as seen from the dataset. Now we correlate **TARGET** with every possible field.

But there is one limitation, we cant correlate them with the field having non numeric values. Thus we have correlated them with the ones having numeric values.

Following is the formula we used (We kept the TARGET field as constant array1 to provided in formulas of **correl** for every other field except TARGET itself.

```
=CORREL($B$1:$B$50000,BK1:BK50000)
```

TASK E

Then a table was obtained consisting of correlation of target with every other field.

***Full Table in Working file

Field with which correl is done w.r.t. Target Variabl	Correl Value
SK_ID_CURR	0.003294877
CNT_CHILDREN	0.026363931
AMT_INCOME_TOTAL	0.010893745
AMT_CREDIT	-0.032343834
AMT_ANNUITY	-0.012399094
AMT_GOODS_PRICE	-0.041306523
REGION_POPULATION_RELATIVE	-0.040799172
DAYS_BIRTH	0.076787685
DAYS_EMPLOYED	-0.040294905
DAYS_REGISTRATION	0.042342679
DAYS_ID_PUBLISH	0.046926745
FLAG_MOBIL	0.001323455
FLAG_EMP_PHONE	0.04140843
FLAG_WORK_PHONE	0.021302134
FLAG_CONT_MOBILE	0.006765545
FLAG_PHONE	-0.032679413
FLAG_EMAIL	-0.001311805
HOUR_APPR_PROCESS_START	-0.032036463
REG_REGION_NOT_LIVE_REGION	0.009438717
REG_REGION_NOT_WORK_REGION	-0.001006443
LIVE_REGION_NOT_WORK_REGION	-0.005497852
REG_CITY_NOT_LIVE_CITY	0.0387731
REG_CITY_NOT_WORK_CITY	0.048450787
LIVE_CITY_NOT_WORK_CITY	0.032261323
EXT_SOURCE_2	-0.158424274

TASK E

Then we arranged the correlations from largest to smallest and obtained the top 10n correlations.

***Full Table in Working file
(Only Top 10 is shown here)

Field with which correl is done w.r.t. Target Variabl	Correl Value
DAYS_BIRTH	0.076787685
DAYS_LAST_PHONE_CHANGE	0.05613673 5
REG_CITY_NOT_WORK_CITY	0.048450787
DAYS_ID_PUBLISH	0.046926745
FLAG_DOCUMENT_3	0.045050228
DEF_60_CNT_SOCIAL_CIRCLE	0.044259774
DAYS_REGISTRATION	0.042342679
DEF_30_CNT_SOCIAL_CIRCLE	0.041603087
FLAG_EMP_PHONE	0.04140843
REG_CITY_NOT_LIVE_CITY	0.0387731

RESULT

I found this project on risk analytics in banking and financial services to be immensely valuable and insightful. It provided me with a practical understanding of how real-world data is analyzed in the financial sector using tools like Excel. By exploring the key factors that indicate customer difficulties in paying installments, I gained a deeper insight into risk assessment in lending. This project's focus on identifying patterns for loan default allowed me to appreciate the importance of data-driven decision-making in the industry. I learned how financial institutions can use data analysis to make informed choices, such as denying loans to high-risk applicants or adjusting loan terms based on risk profiles. Overall, this project not only enhanced my Excel skills but also deepened my understanding of risk management and analytics in the banking sector.

Ashish Kumar Samantaray

B.Tech, Computer Science and Engineering

ashish.kumar.samantaray2003@gmail.com

7205691104