TRAINITY TASK-5

Bank Loan Case Study

PRESENTED BY- ADITYA PALANDE

Email id - adityap.works@gmail.com

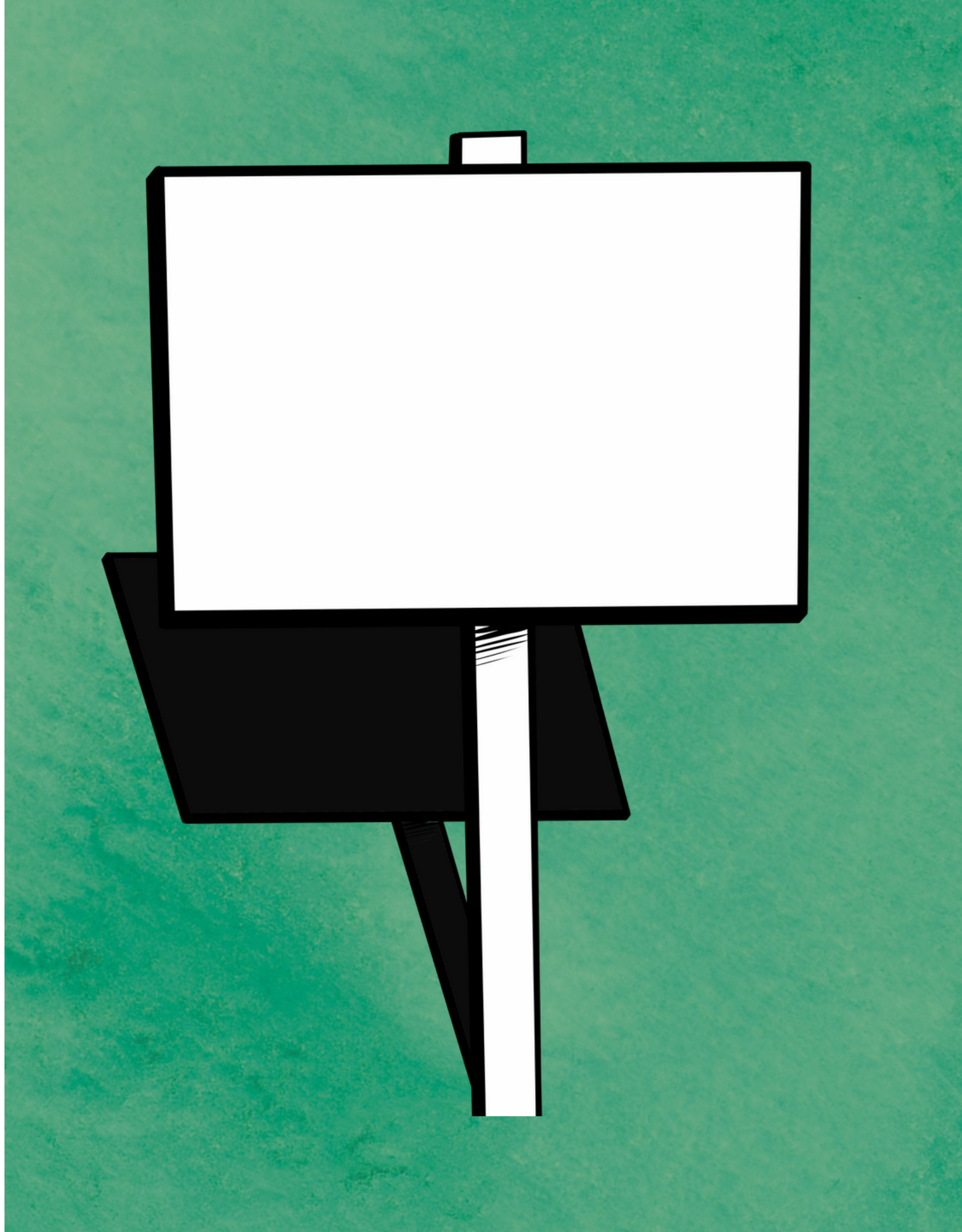
Contents

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

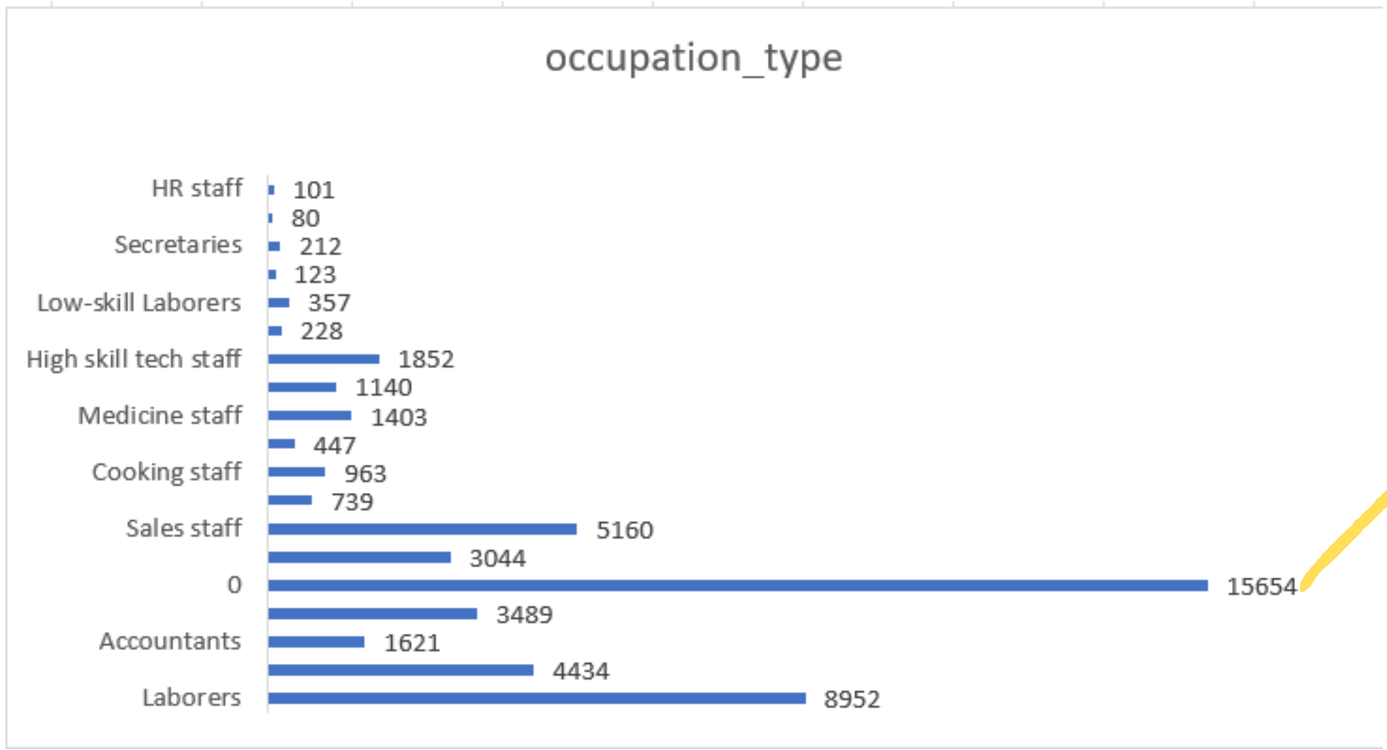
To identify blanks in each column/row, I used **COUNTA()** function to calculate the count of non-empty cells and subtracted the result from the total number of cells.

Missing values were handled using two approaches:

1. **Deleting** rows/columns with maximum blanks
2. **Replacing** blanks with appropriate values.



1. Columns/rows containing more blanks than actual data (i.e columns/rows with more than 50% blank cells) were deleted.
2. Remaining empty cells were replaced by either median or mode of the respective column.



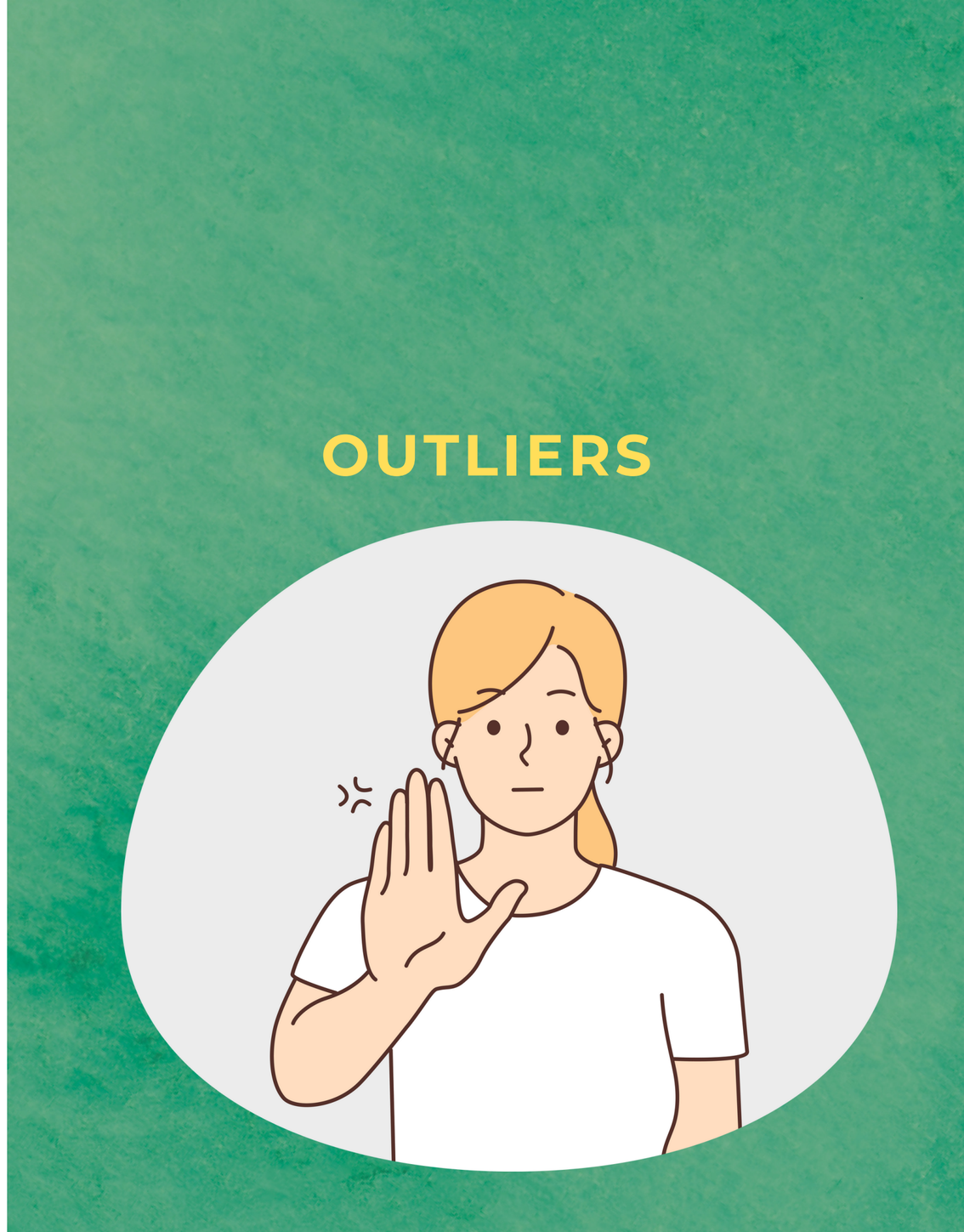
[Link to my excel sheet]

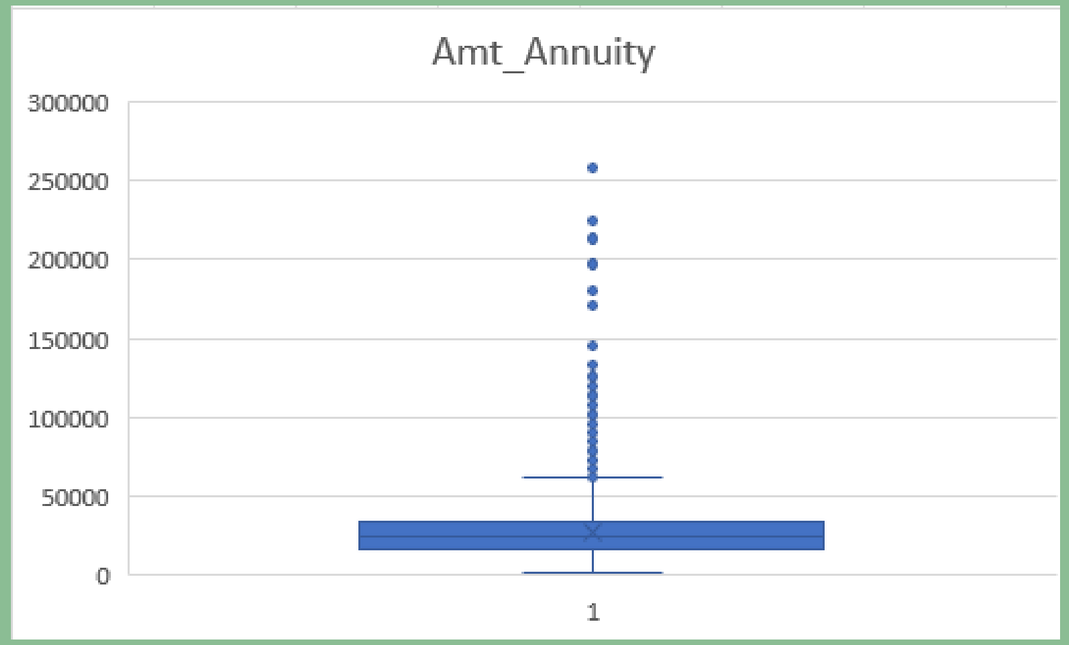
Occupation_type column in application dataset has 15654 blank cells contributing to 31% of blank cells

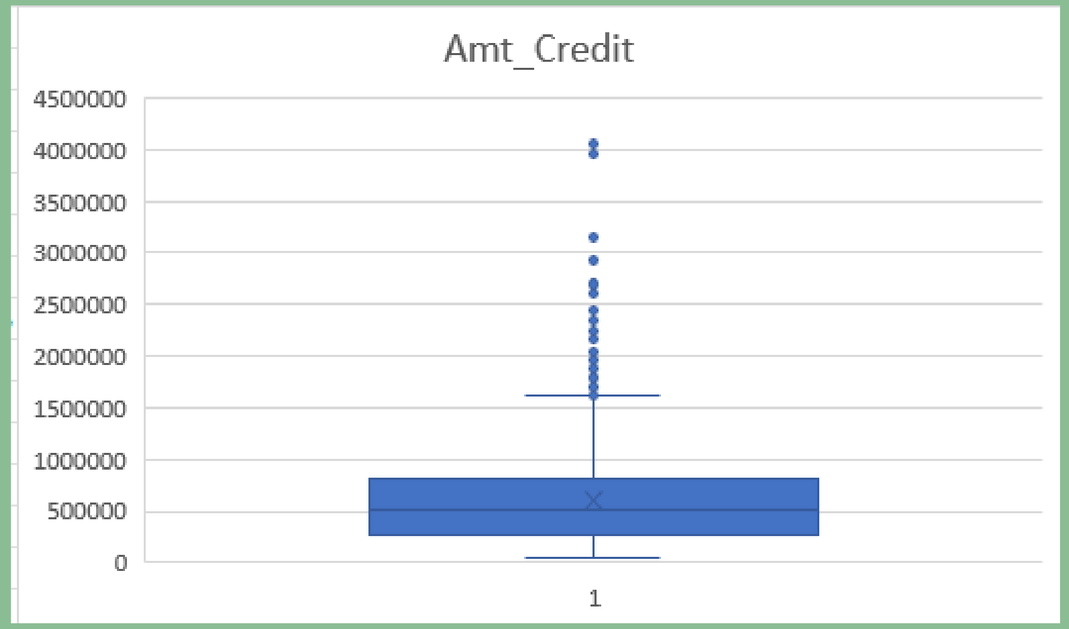
Task B: Identify Outliers in the Dataset

Ouliers were found using **QUARTILE(), IQR()** functions and plotting box plots for each variable. Every value greater than the maximum and lesser than the minimum were considered outliers.

Ouliers in each column are represented by <light red filled cells with dark red font>. This was done by conditional formatting.

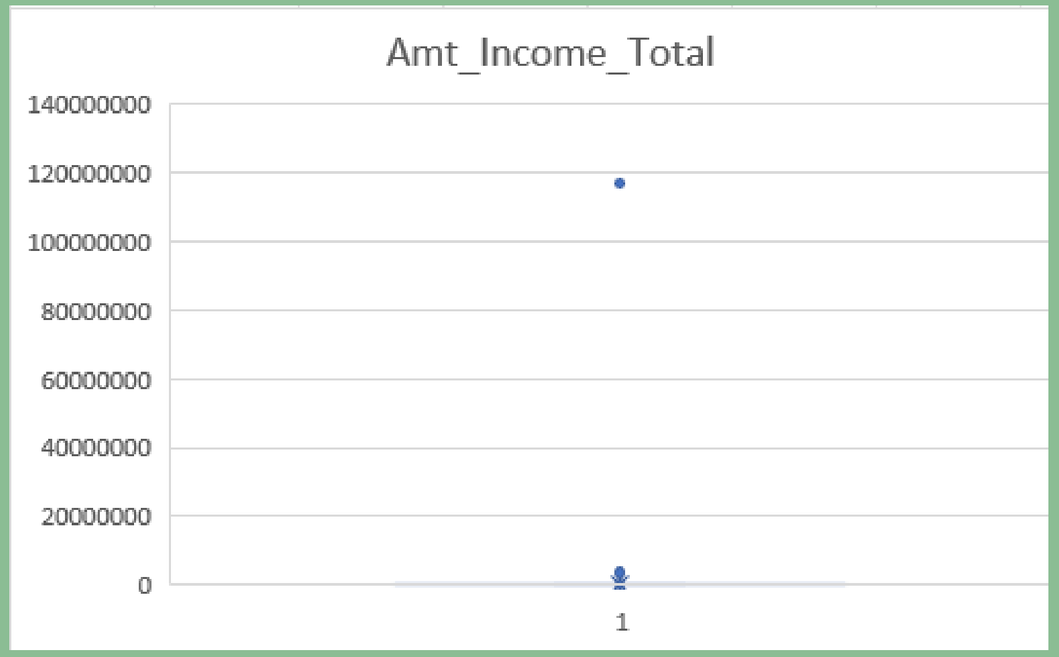






AMT_CREDIT	AMT_ANNUITY
254700	25321.5
876078	49050
2250000	83515.5
1040985	30568.5
684054	77494.5
1256400	36864
720000	28552.5
270000	13500
1024740	55719
1102500	32364
544491	16047
405000	20250
912240	30276
679671	28926
180000	9000
364896	19926
312768	20353.5
1288350	37800
781920	50148
1483231.5	51687
590337	28530
101880	10827
161730	13833
447768	35505
495216	26995.5
402939	19381.5
808650	23643
76410	4513.5
90000	6529.5
270000	14134.5

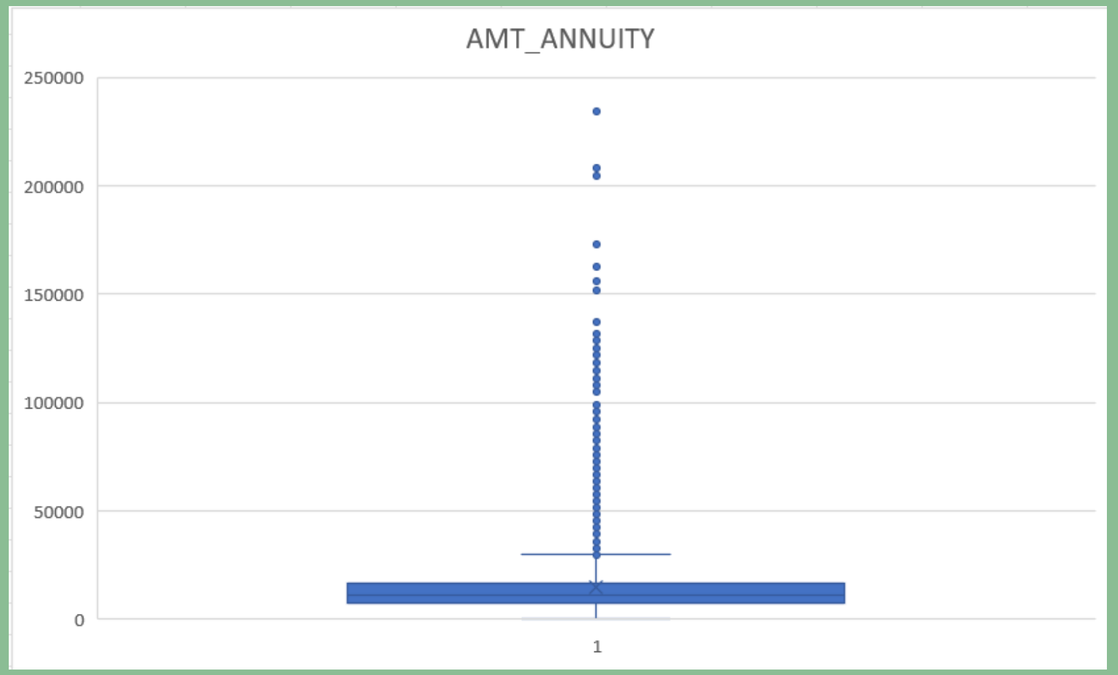
APPLICATION DATASET



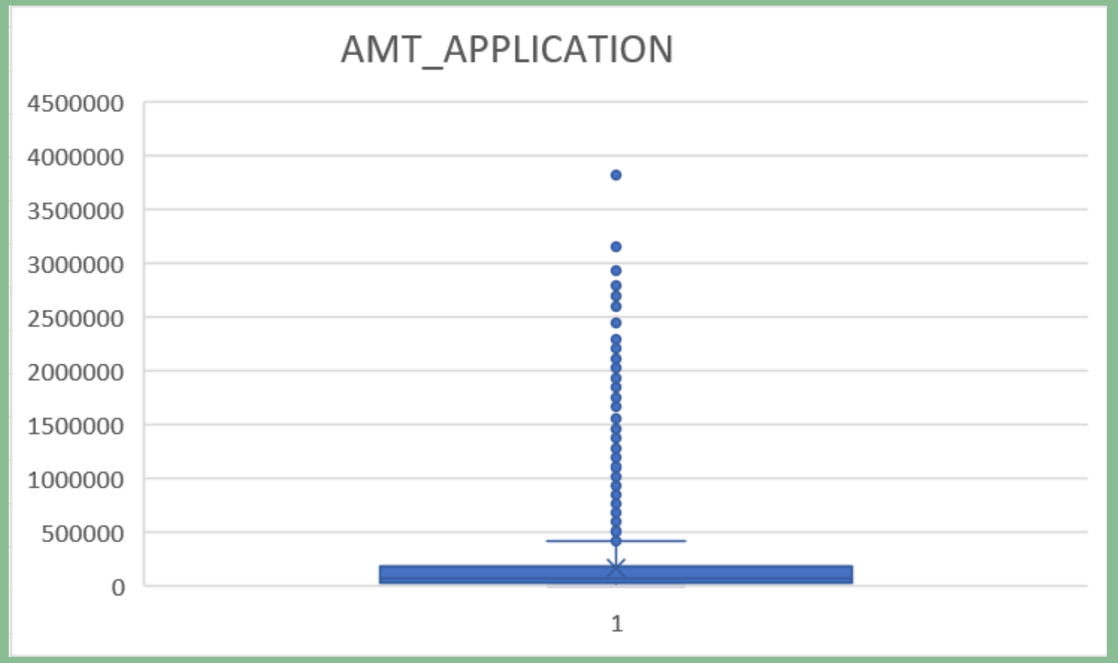
AMT_GOODS_PRICE	AMT_INCOME_TOTAL
351000	202500
1129500	270000
135000	67500
297000	135000
513000	121500
454500	99000
1395000	171000
1530000	360000
913500	112500
405000	135000
652500	112500
135000	38419.155
67500	67500
697500	225000
679500	189000
247500	157500
387000	108000
270000	81000
157500	112500
454500	90000
427500	135000
927000	202500
450000	450000
225000	83250
247500	135000

APPLICATION DATASET

PREVIOUS APPLICATION

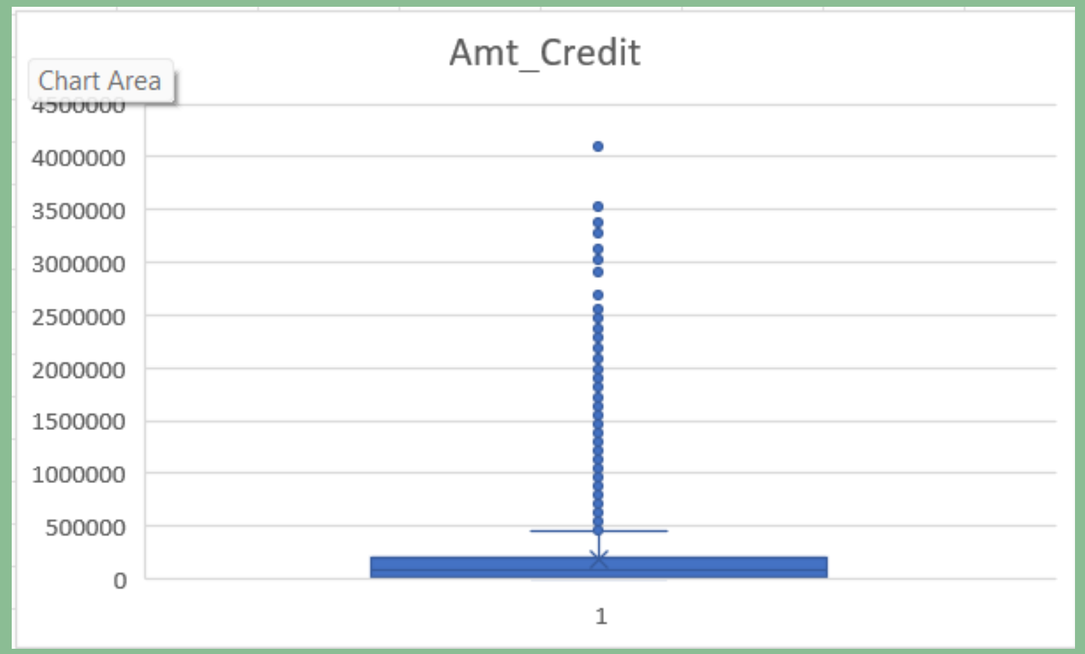


min	0
25%	7189.74
Median	10879.92
75%	16256.16
max	29855.79
IQR	9066.42



PREVIOUS APPLICATION

AMT_ANNUITY	AMT_APPLICATION
1730.43	17145
25188.615	607500
15060.735	112500
47041.335	450000
31924.395	337500
23703.93	315000
10879.92	0
10879.92	0
10879.92	0



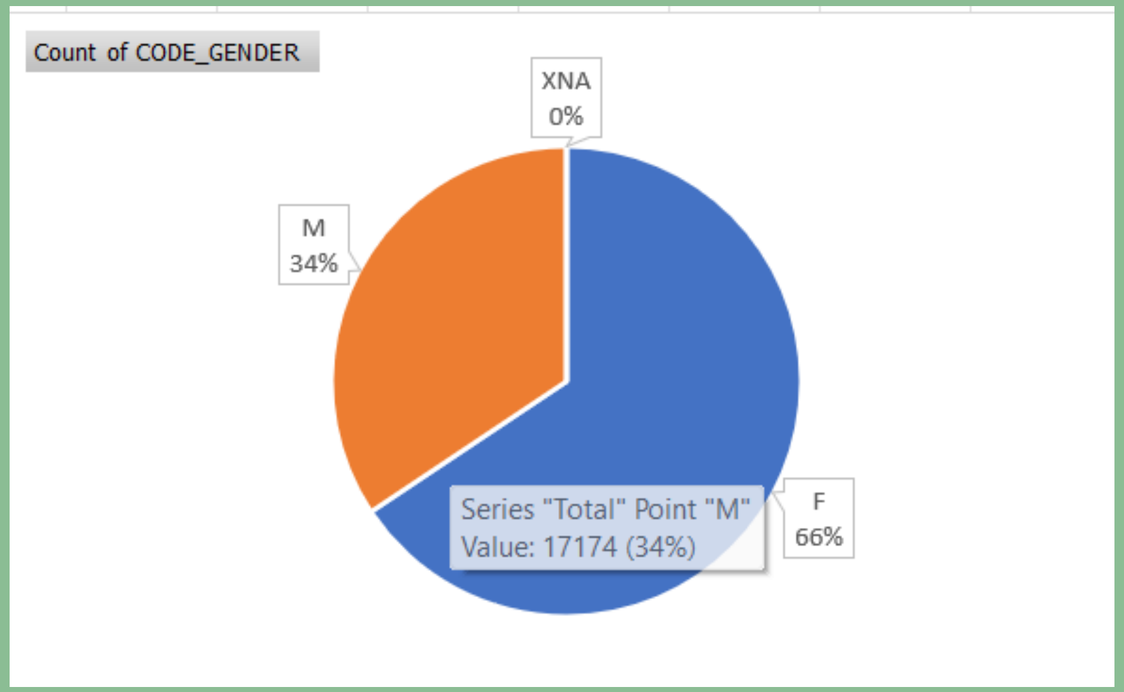
AMT_CREDIT	AMT_GOODS_PRICE
17145	17145
679671	607500
136444.5	112500
470790	450000
404055	337500
340573.5	315000
0	104017.5
0	104017.5
0	104017.5
0	104017.5

Task C: Analyze Data Imbalance

An **imbalanced dataset** refers to a situation in which the distribution of classes or categories within the dataset is uneven. This **data imbalance ratio** will give you an indication of the data imbalance between the two classes. If the ratio is close to 1, it indicates a balanced dataset.

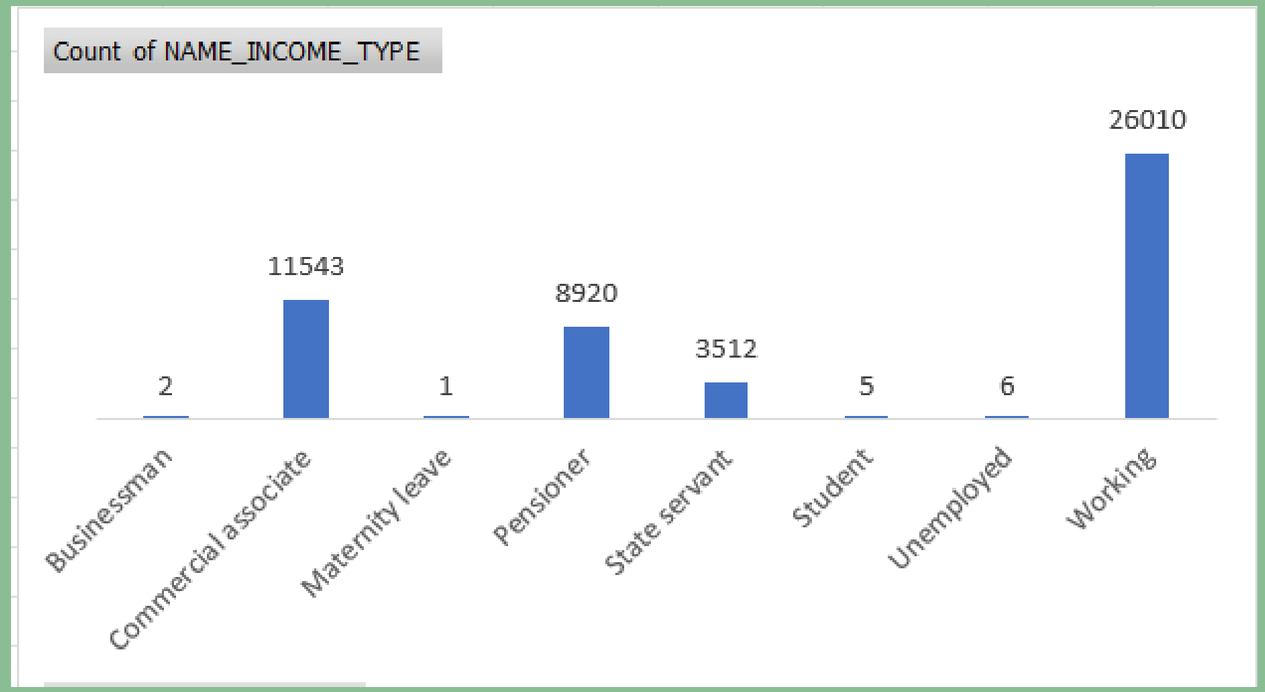
For example, if the counts of "Class 1" and "Class 2" are in cells C2 and C3 respectively, you would use the formula =**C2/C3** to calculate the ratio.

Row Labels	Count of CODE_GENDER
F	32823
M	17174
XNA	2
Grand Total	**49999**



Data imbalance ratio
0.5232

Row Labels	Count of NAME_INCOME_TYPE
Businessman	2
Commercial associate	11543
Maternity leave	1
Pensioner	8920
State servant	3512
Student	5
Unemployed	6
Working	26010



Data imbalance ratio
0.00004

Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis

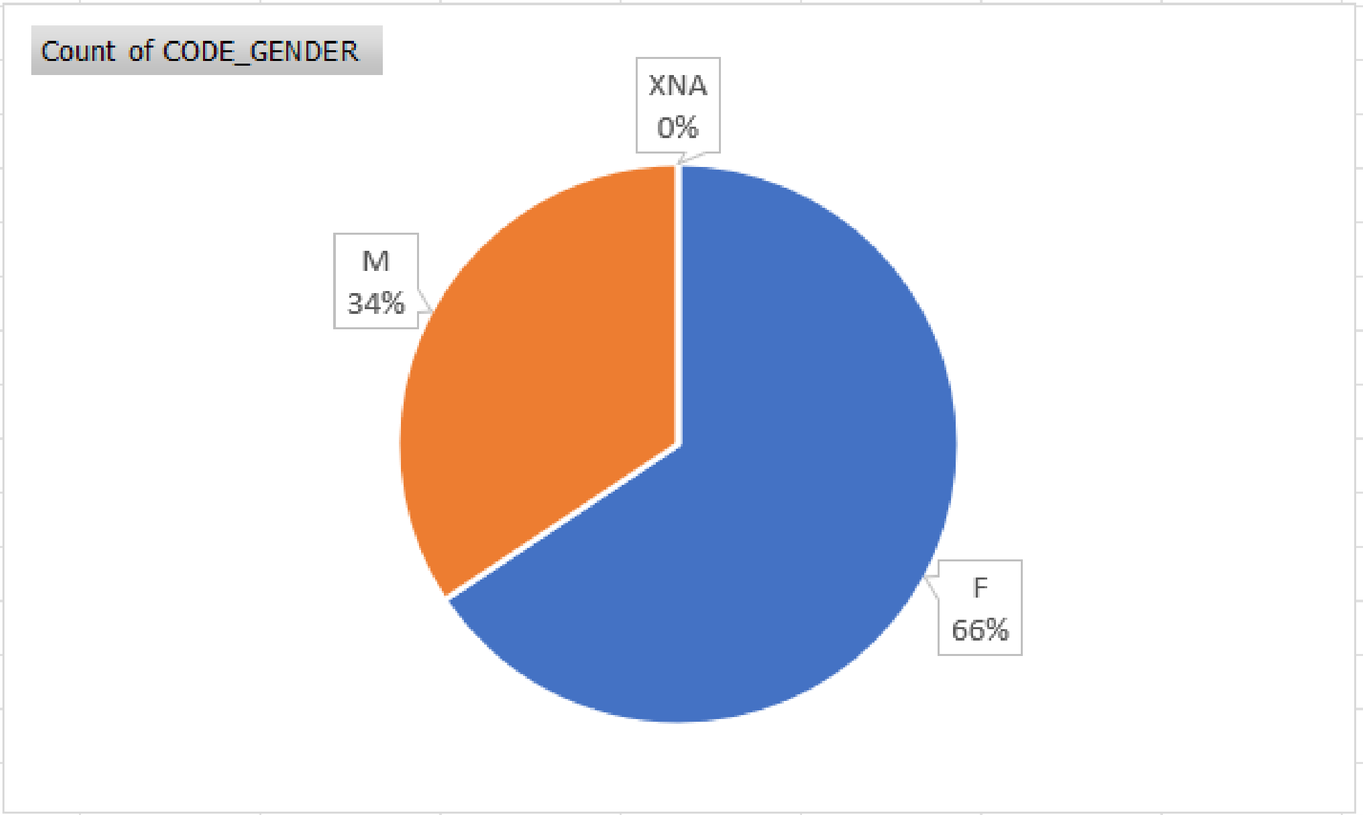
Univariate analysis focuses on understanding individual variables. - **Bivariate** analysis examines relationships between two variables

Histograms, bar charts, or box plots have been created to visualize the distributions of variables.

Univariate Analysis

Row Labels	Count of CODE_GENDER
F	32823
M	17174
XNA	2
Grand Total	**49999**

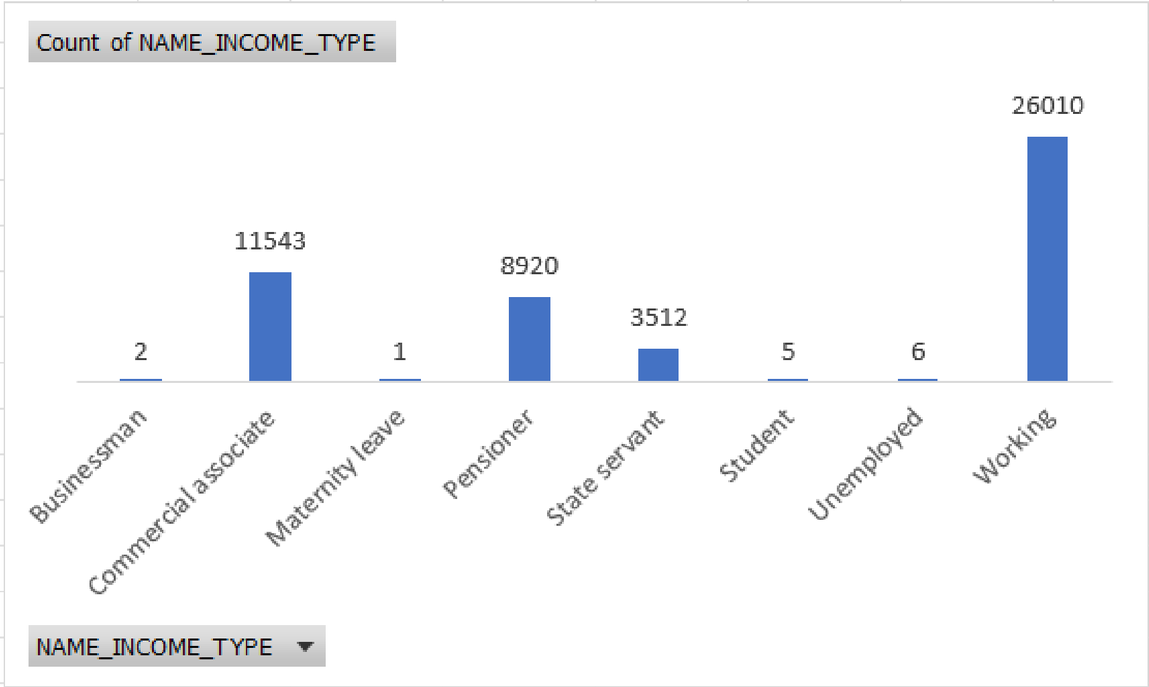
Univariate analysis- There are more female applicants than male applicants



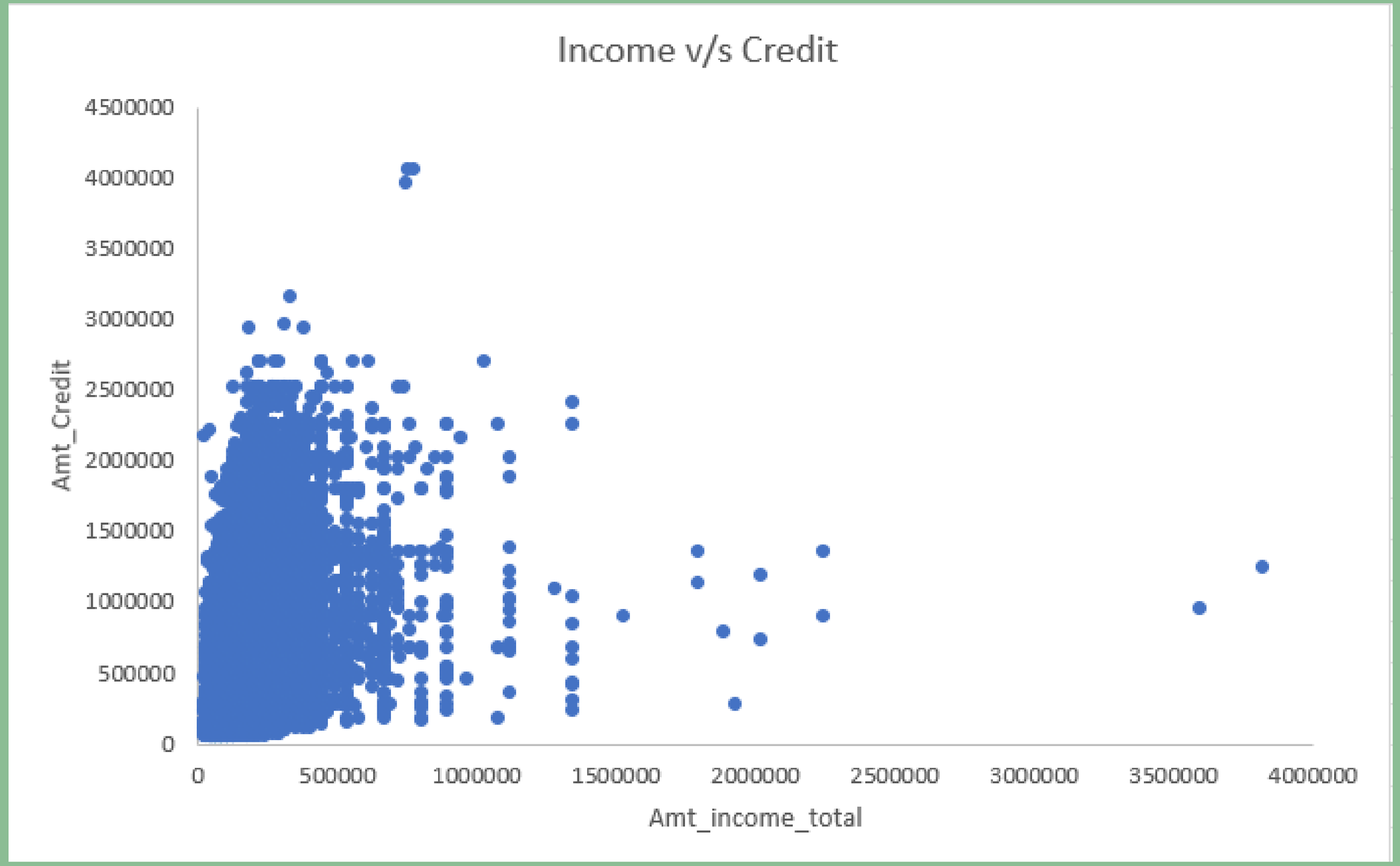
Univariate Analysis

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

Most of the applicants are working.

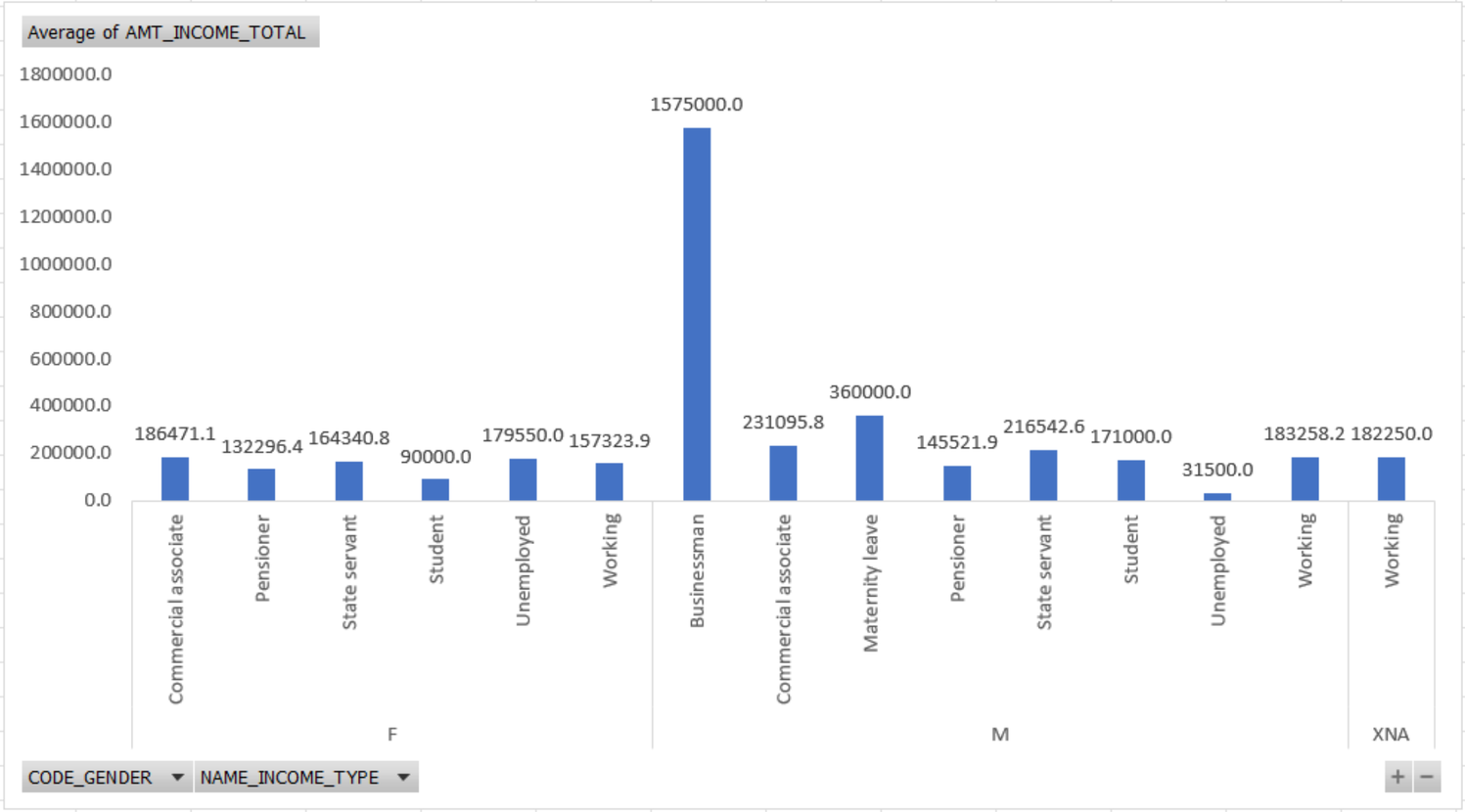


Bivariate Analysis



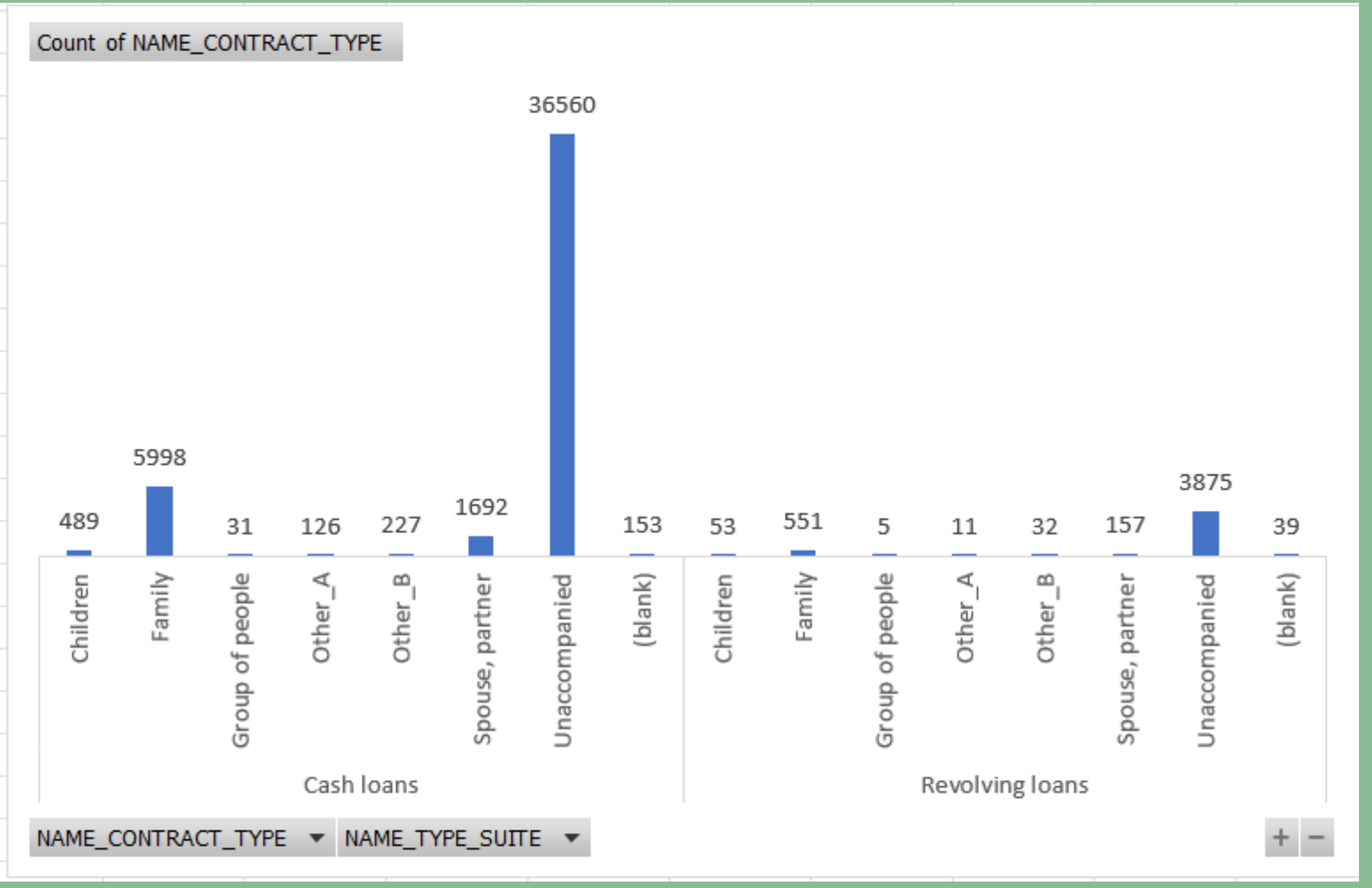
Bivariate Analysis

Row Labels	Average of AMT_INCOME_TOTAL
F	**158674.7**
Commercial associate	186471.1
Pensioner	132296.4
State servant	164340.8
Student	90000.0
Unemployed	179550.0
Working	157323.9
M	**193878.3**
Businessman	1575000.0
Commercial associate	231095.8
Maternity leave	360000.0
Pensioner	145521.9
State servant	216542.6
Student	171000.0
Unemployed	31500.0
Working	183258.2
XNA	**182250.0**
Working	182250.0
Grand Total	**170767.6**



Bivariate Analysis

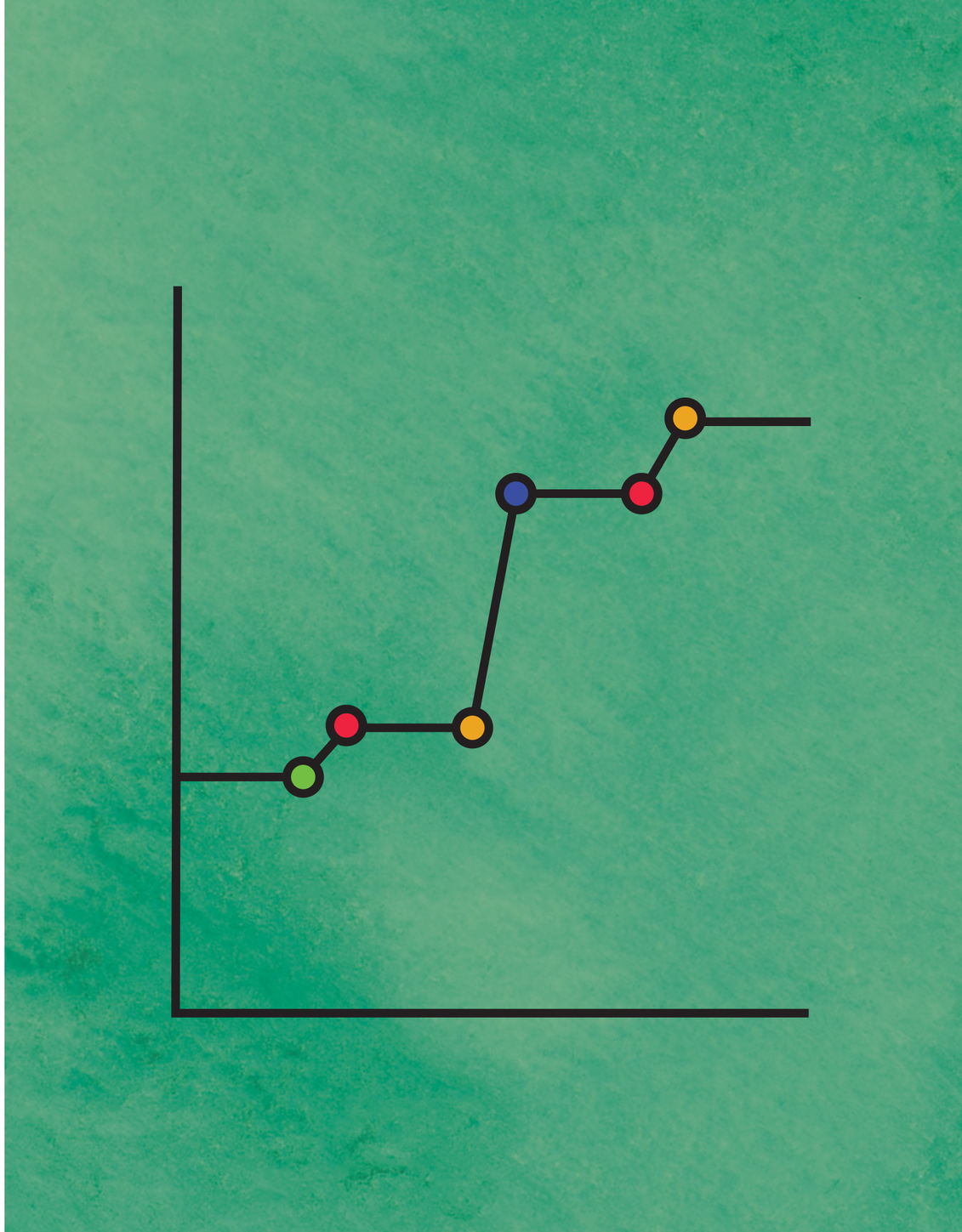
Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	**45276**
Children	489
Family	5998
Group of people	31
Other_A	126
Other_B	227
Spouse, partner	1692
Unaccompanied	36560
(blank)	153
Revolving loans	**4723**
Children	53
Family	551
Group of people	5
Other_A	11
Other_B	32
Spouse, partner	157
Unaccompanied	3875
(blank)	39
Grand Total	**49999**



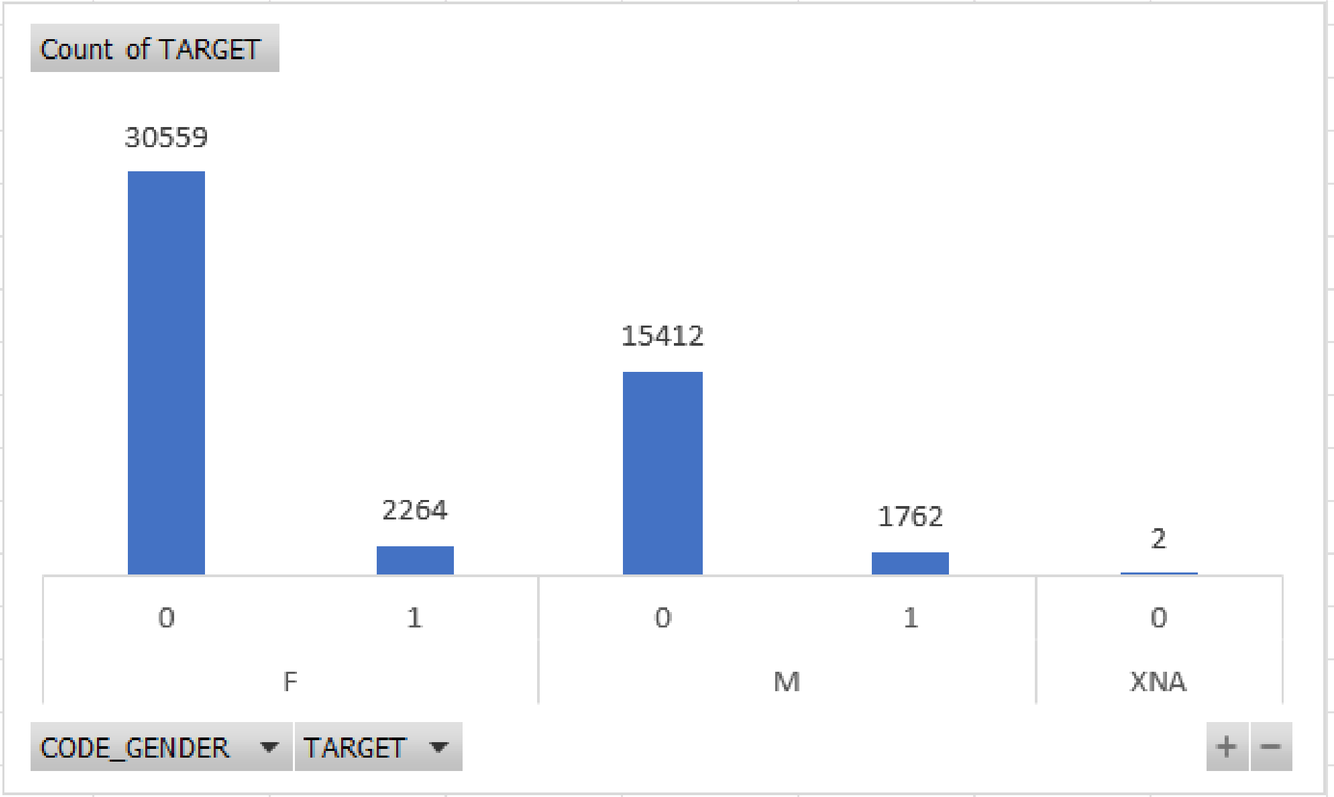
Task E: Identify Top Correlations for Different Scenarios

Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

Heatmaps have been created to visualize the correlations between variables within each segment. The top correlated variables for each scenario have been highlighted using different colors.



Row Labels	Count of TARGET	
F	**32823**	
0	30559	
1	2264	
M	**17174**	
0	15412	
1	1762	
XNA	**2**	
0	2	
Grand Total	**49999**	
ratio of female defaulters		0.068976
ratio of male defaulters		0.102597



Ratio of male defaulters is more than that of female defaulters.

Income_type / defaulter	0	1	Total	Ratio of defaulters
Working	23549	2461	26010	0.094617455
State servant	3314	198	3512	0.056378132
Commercial associate	10679	864	11543	0.074850559
Pensioner	8419	501	8920	0.056165919
Unemployed	4	2	6	0.333333333
Student	5	0	5	0
Businessman	2	0	2	0
Maternity leave	1	0	1	0

Link to my working excel sheet and video presentation:

Excel Sheet

Video

CONCLUSION

The tasks were performed using Microsoft Excel. The tasks not only helped in understanding excel tools but also allowed me to get hands-on experience by solving real-life examples. Through these tasks insights could be drawn and strategies couls be made.