Bank Loan Analysis

Project Description

In this project, I have to analyze the data provided by a bank about the applicants applying for a loan. The dataset contains all the information of the applicants such as their income, family members, credit, etc along with the target, that is whether or not they faced any problems making the installments for the loan. Based on this information I have to analyze the data and derive the necessary insights

Approach

For this project, I used the dataset provided by the Trainity team and loaded it into Excel. The data consists of many missing values and outliers thus we would have to perform data cleaning before we proceed. I have used various inbuilt formulas and data transformation techniques of Excel to derive the necessary insights. I have also used various graphs and charts for the visualization of data

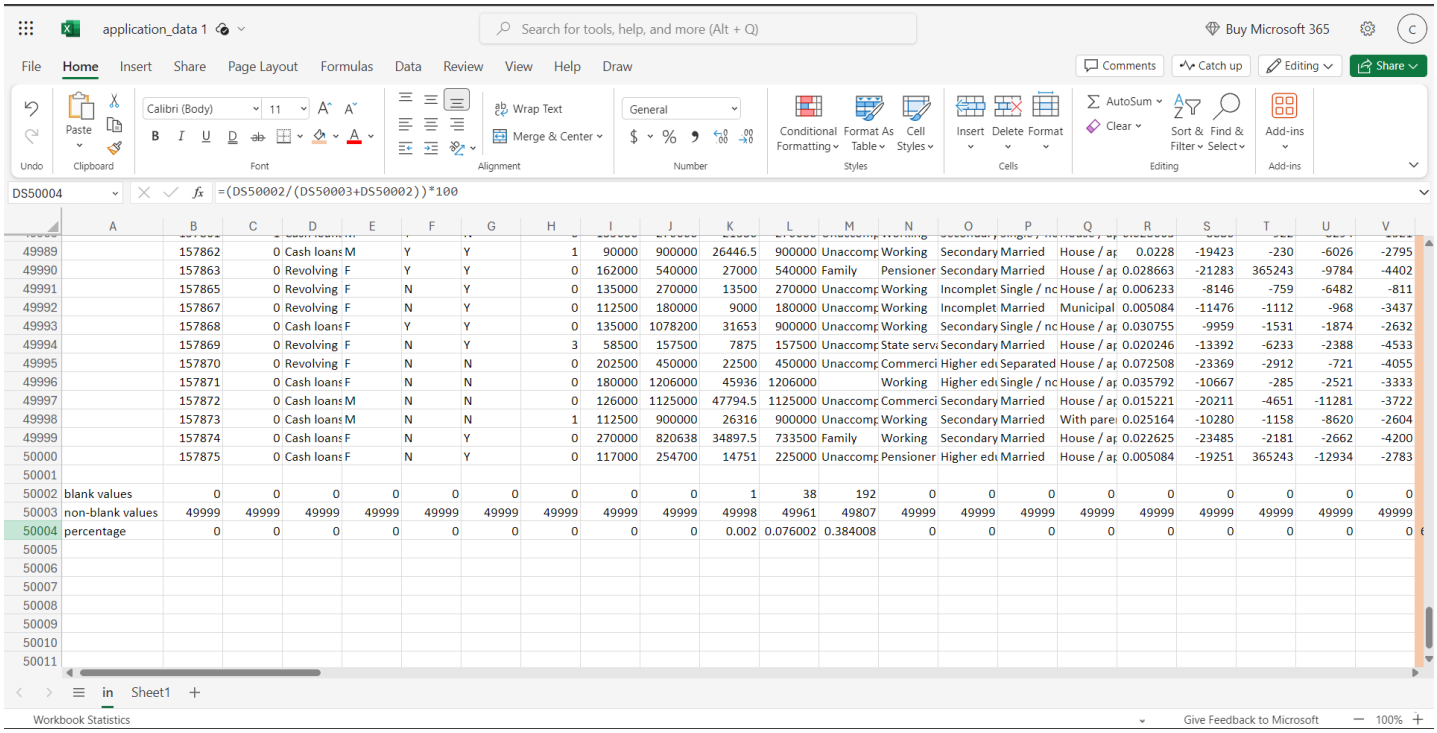
Tech-Stack Used

For this project, I have chosen Microsoft Excel as it is a powerful tool that offers numerous benefits for data analysis, business management, and personal use. Excel provides a wide range of built-in functions and formulas for mathematical, statistical, financial, and logical calculations. The use of filters and sorting mechanisms makes deriving insights easier

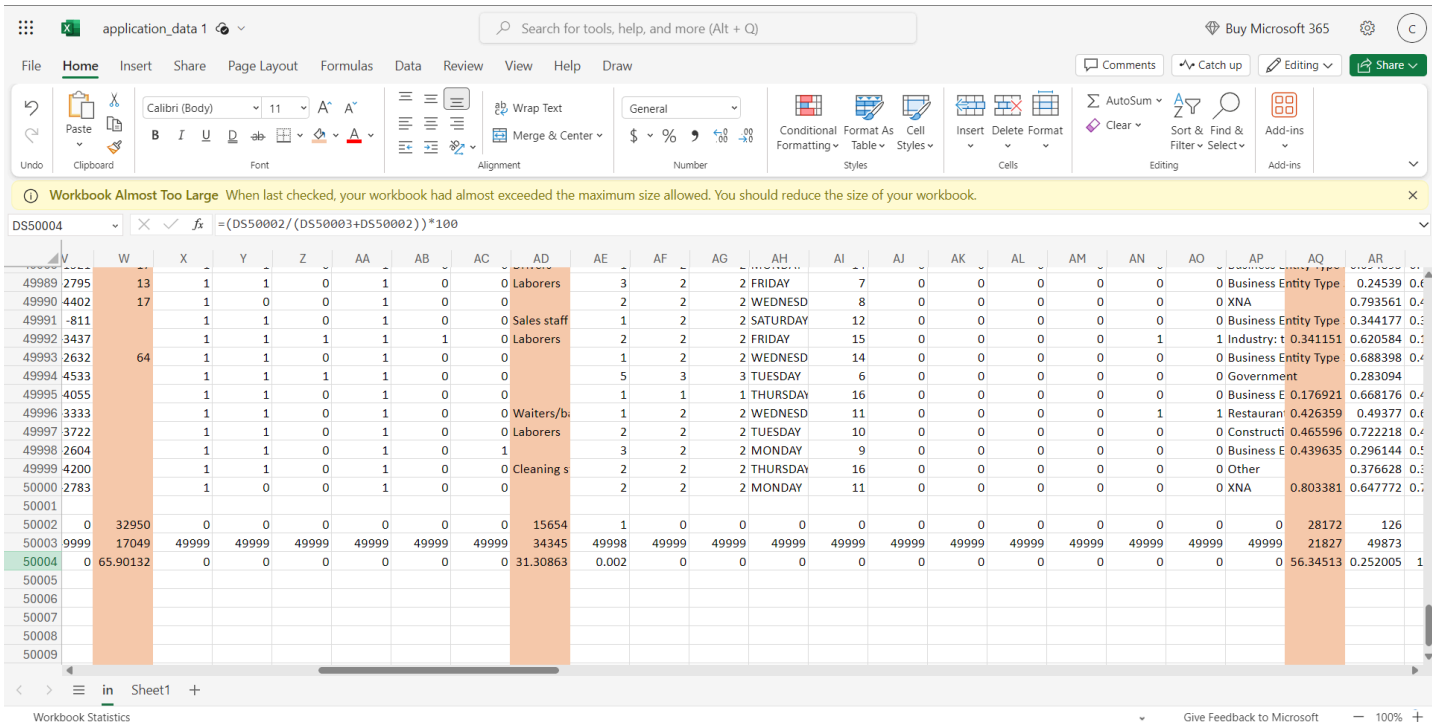
Insights

Task 1 : Identify Missing Data and Deal with it Appropriately:

We import the dataset into excel. Using the =COUNTBLANK function we find the number of blank values in each column. Columns in which the the blank values are greater than 30% are highlighted in red. These columns will be deleted and in the blank values in remaining columns will be replaced by the median value of the column found using the =MEDIAN function



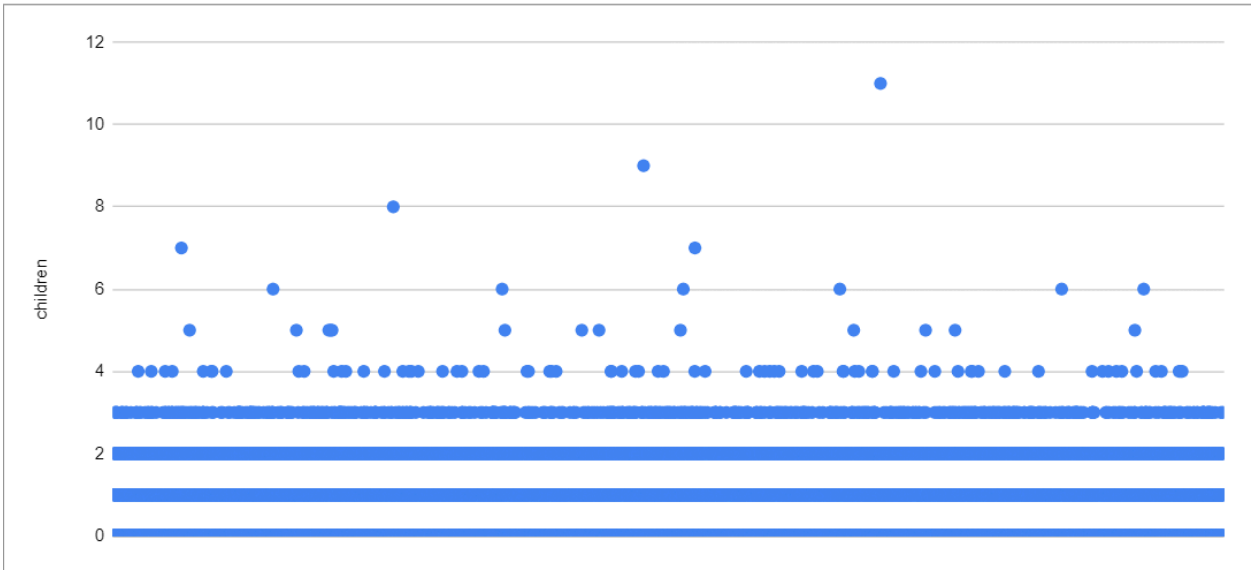
Percentage = (blank values / blank + non-blank values)*100



Task 2 : Identify Outliers in the Dataset

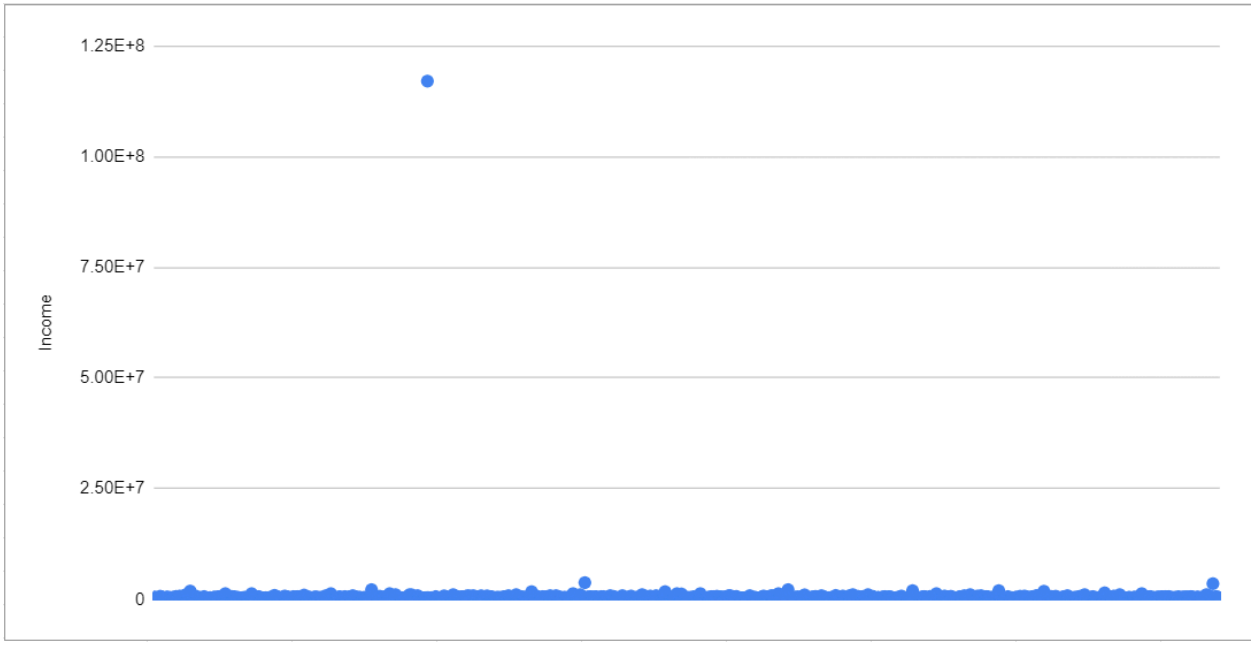
Formula used : =QUARTILE(array, quartile)

For CNT_CHILDREN		
Q1 =	0	
Q3=	1	
IQR = Q3 - Q1 =	1	
Upper bound =	Q3 + 1.5*IQR	2.5
Lower bound =	Q1 - 1.5*IQR	-1.5



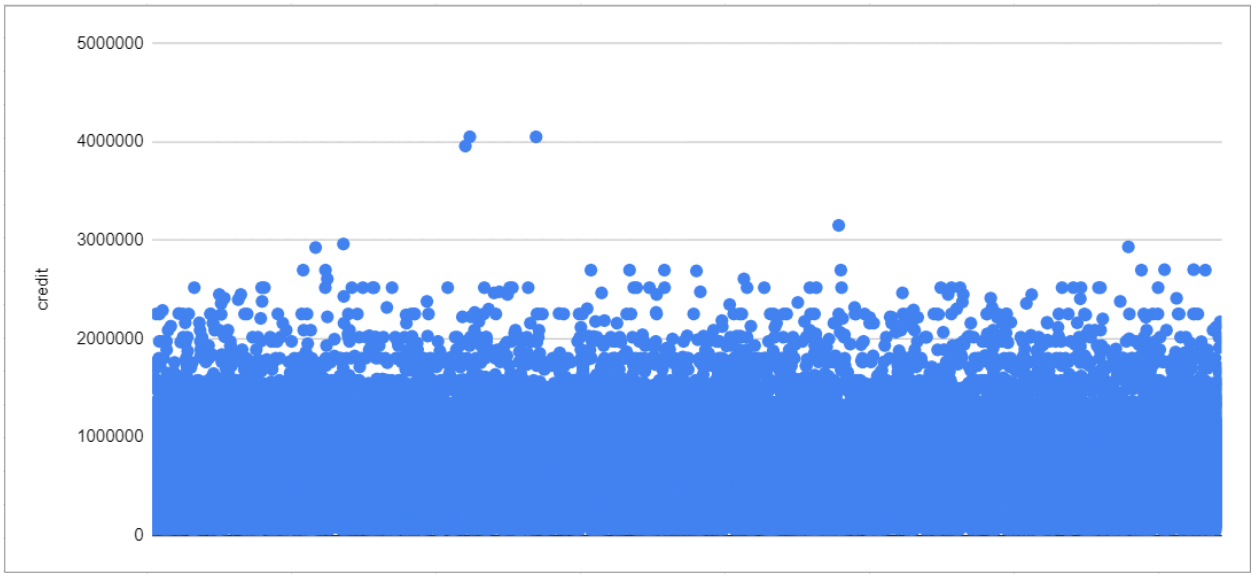
Hence we can see that applicants with number of children over 4 are considered to be outliers

for AMT_INCOME_TOTAL		
Q1 =	112500	
Q3=	202500	
IQR = Q3 - Q1 =	90000	
Upper bound =	Q3 + 1.5*IQR	337500
Lower bound =	Q1 - 1.5*IQR	-22500



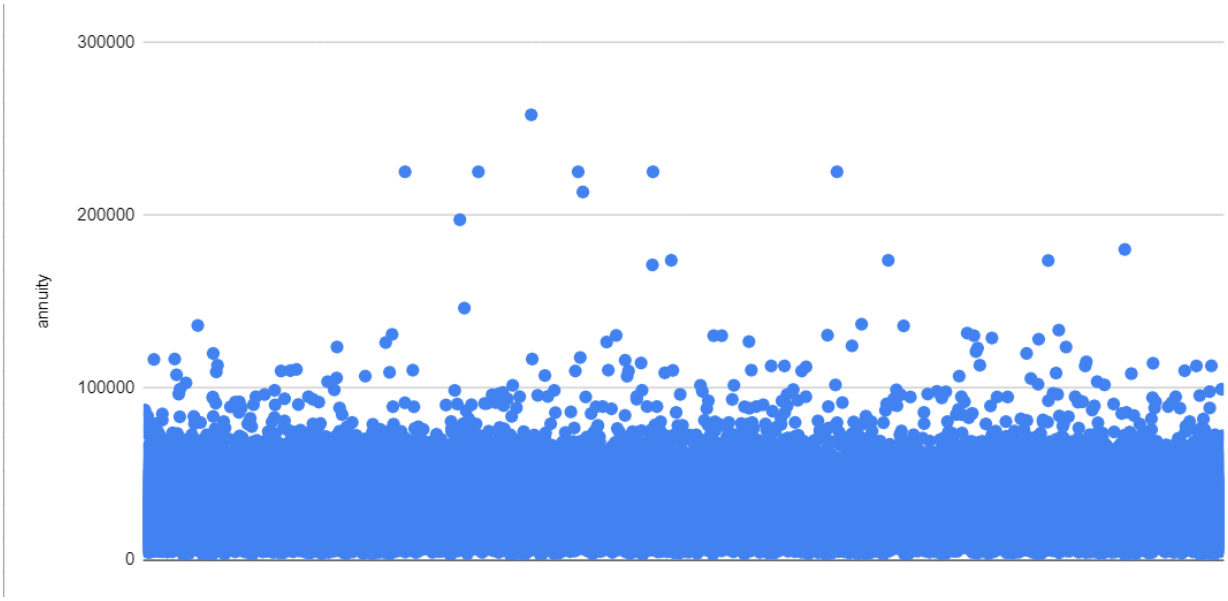
Here we can see a data point that deviates significantly from the rest of the dataset hence being classified as an outlier

for AMT_CREDIT		
Q1 =	270000	
Q3=	808650	
IQR = Q3 - Q1 =	538650	
Upper bound =	Q3 + 1.5*IQR	1616625
Lower bound =	Q1 - 1.5*IQR	-537975



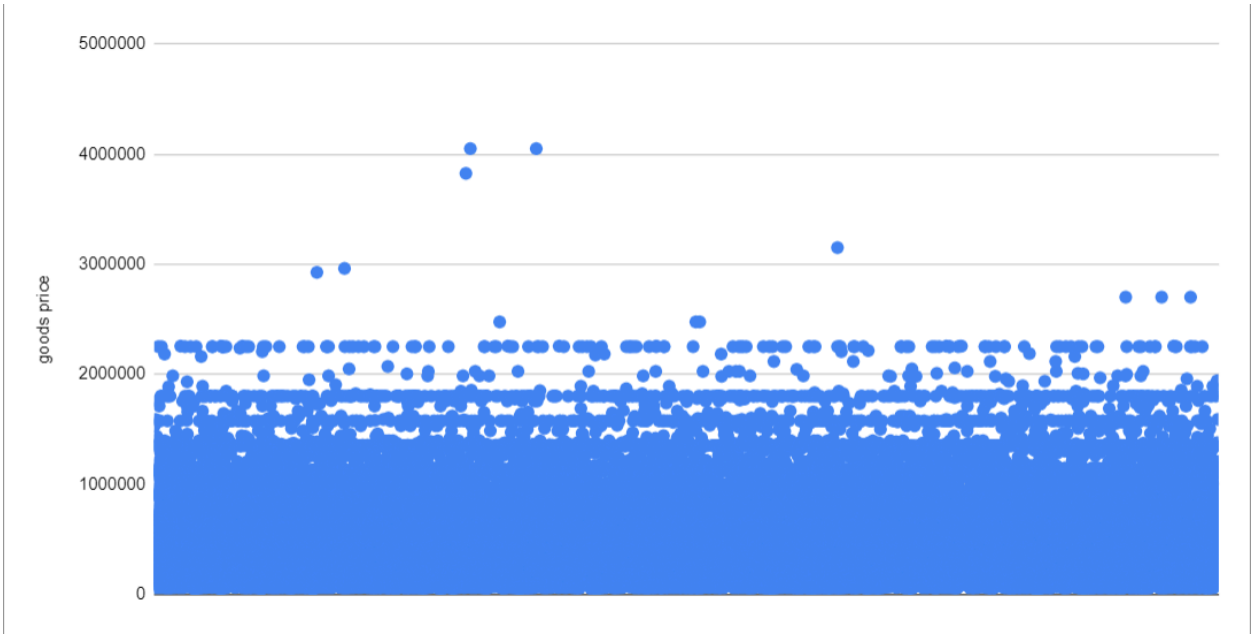
Thus the above scatter plot shows us the data points deviating significantly from the rest of the data. Although many points are classified as outliers, we need to do further analysis as these datapoints may be valid

for AMT_ANNUITY		
Q1 =	16456.5	
Q3=	34596	
IQR = Q3 - Q1 =	18139.5	
Upper bound =	Q3 + 1.5*IQR	61805.25
Lower bound =	Q1 - 1.5*IQR	-10752.8



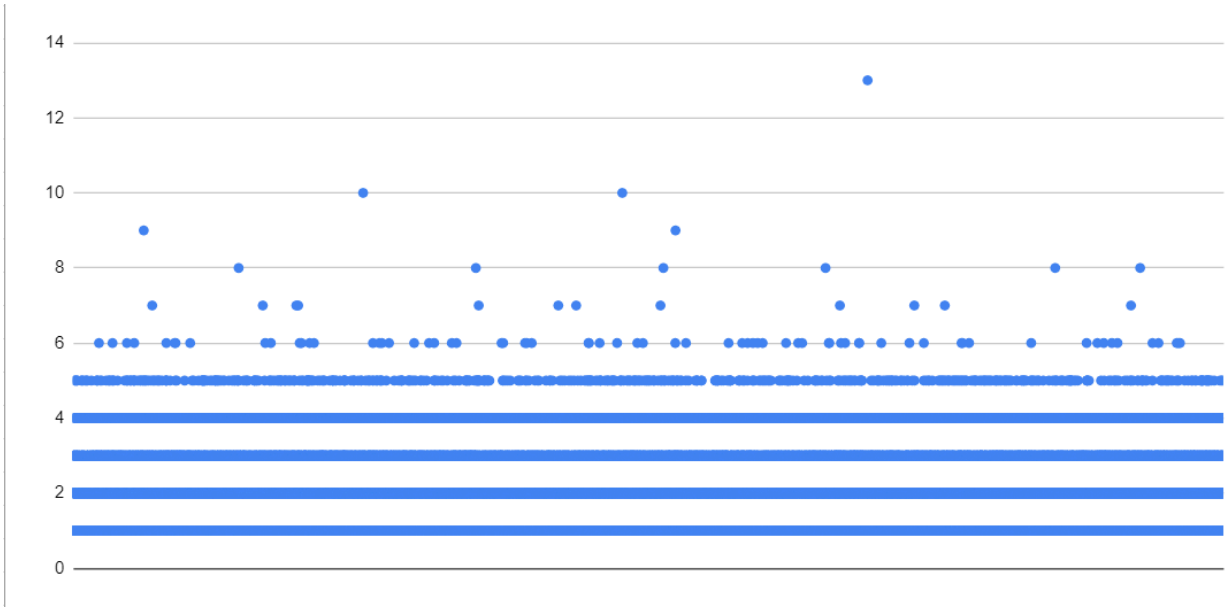
Thus the above scatter plot shows us the data points deviating significantly from the rest of the data. Although many points are classified as outliers, we need to do further analysis as these datapoints may be valid

for AMT_GOODS_PRICE		
Q1 =	238500	
Q3=	679500	
IQR = Q3 - Q1 =	441000	
Upper bound =	Q3 + 1.5*IQR	1341000
Lower bound =	Q1 - 1.5*IQR	-423000



Thus the above scatter plot shows us the data points deviating significantly from the rest of the data

for CNT_FAMILY_MEMBERS		
Q1 =	2	
Q3=	3	
IQR = Q3 - Q1 =	1	
Upper bound =	Q3 + 1.5*IQR	4.5
Lower bound =	Q1 - 1.5*IQR	0.5

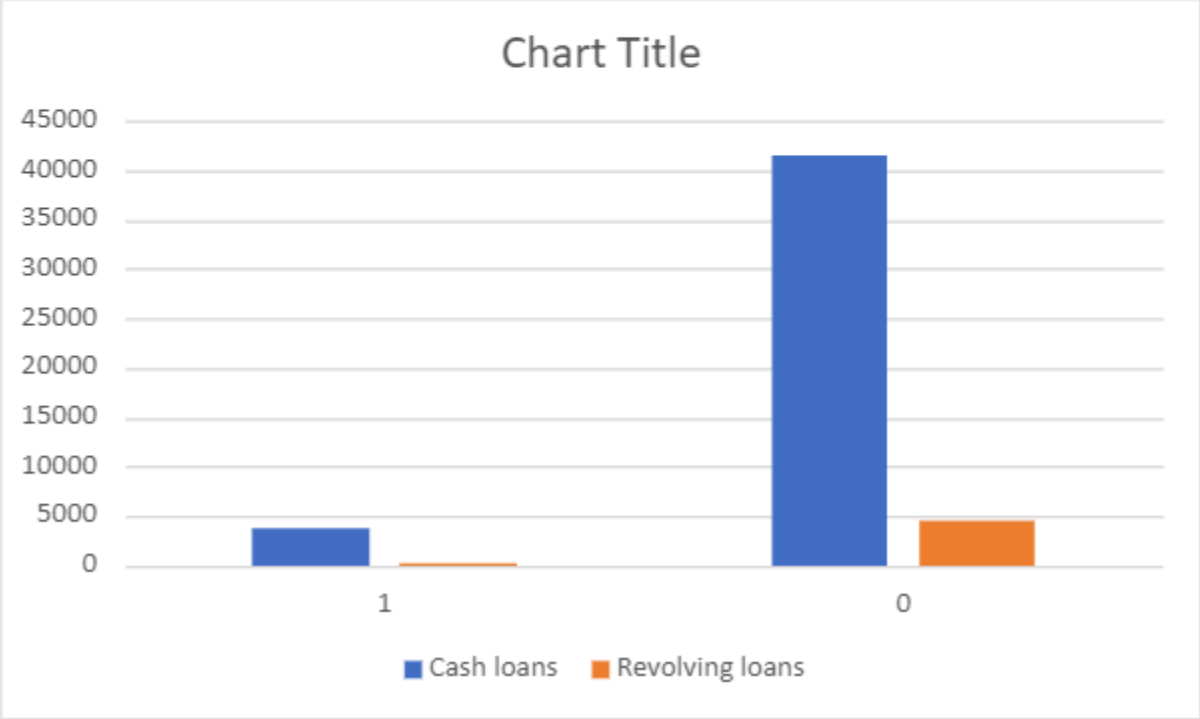


Thus the above scatter plot shows us the data points deviating significantly from the rest of the data

Task 3 : Analyze Data Imbalance

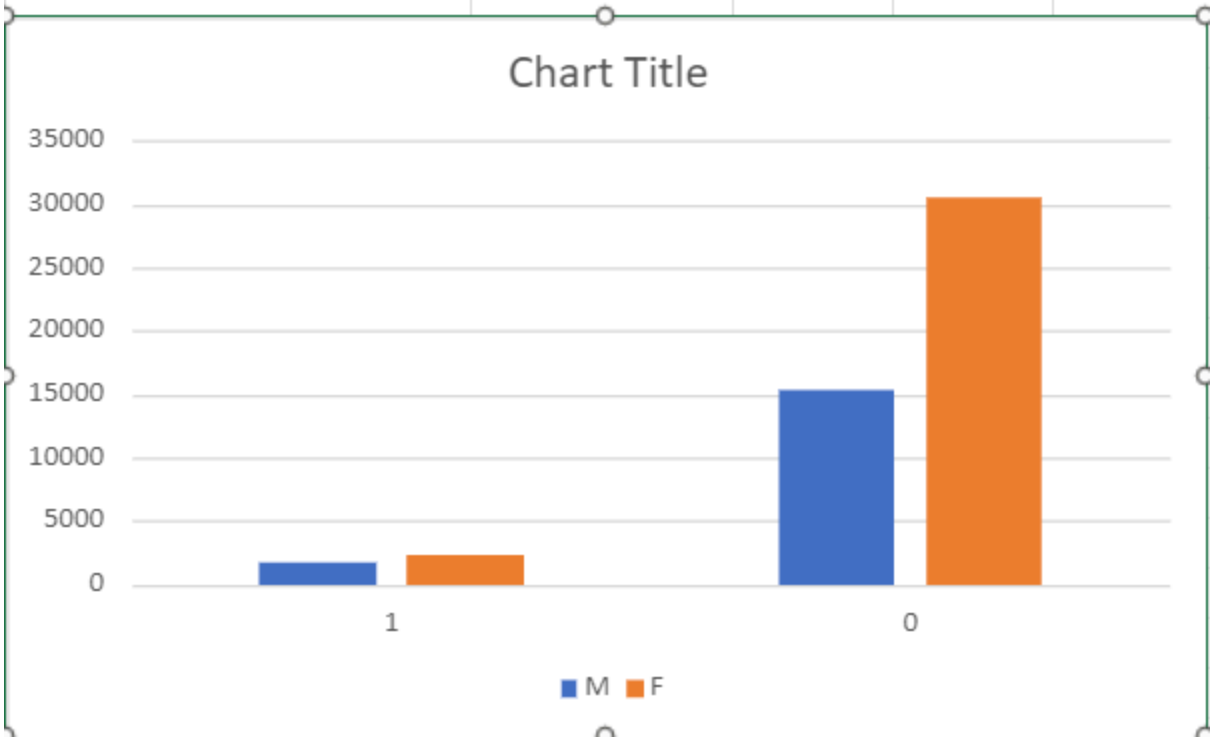
Formula used : =COUNTIFS(range1, condition1, range2, condition2)

For Contract Type			
	1	0	total
Cash loans	3792	41484	45276
Revolving loans	234	4489	4723
total	4026	45973	49999



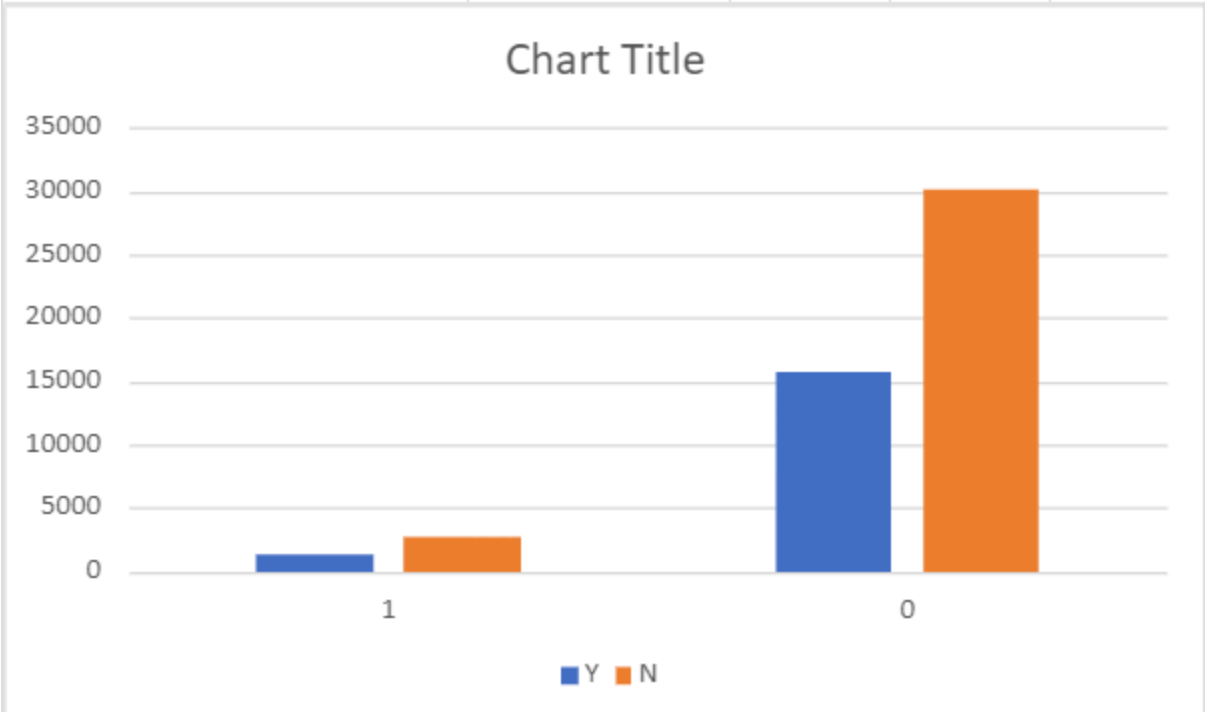
For contract type, we can see that there is a huge data imbalance as most of the applicants who had payment difficulties as most of them had contacts of cash loans. This trend carries over to applicants who didnt have any problems as most of them also had contracts with cash loans

for Gender			
	1	0	total
M	1762	15412	17174
F	2264	30559	32823
total	4026	45971	49999



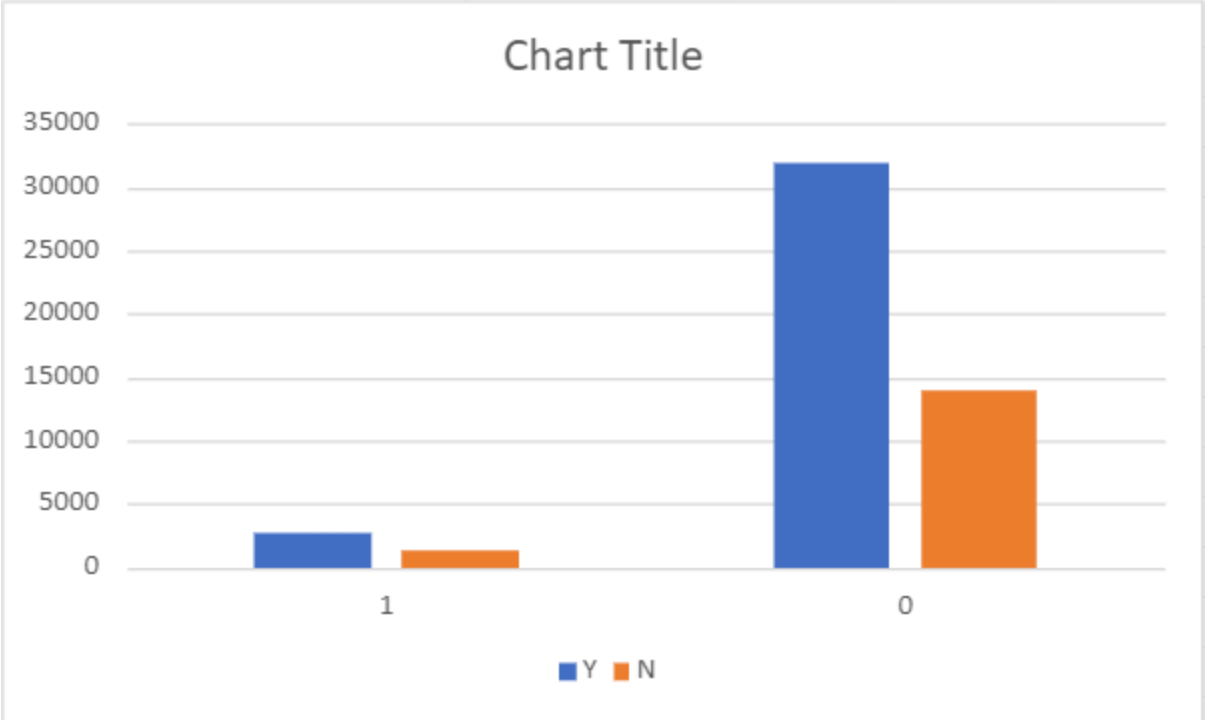
For gender we can see that there is slight imbalance as most of the applicants who faced problems are female. This trend carries over for applicants who didnt face any problems but with a slightly greater imbalance

For Owns Car			
	1	0	total
Y	1253	15797	17050
N	2773	30176	32949
total	4026	45973	49999



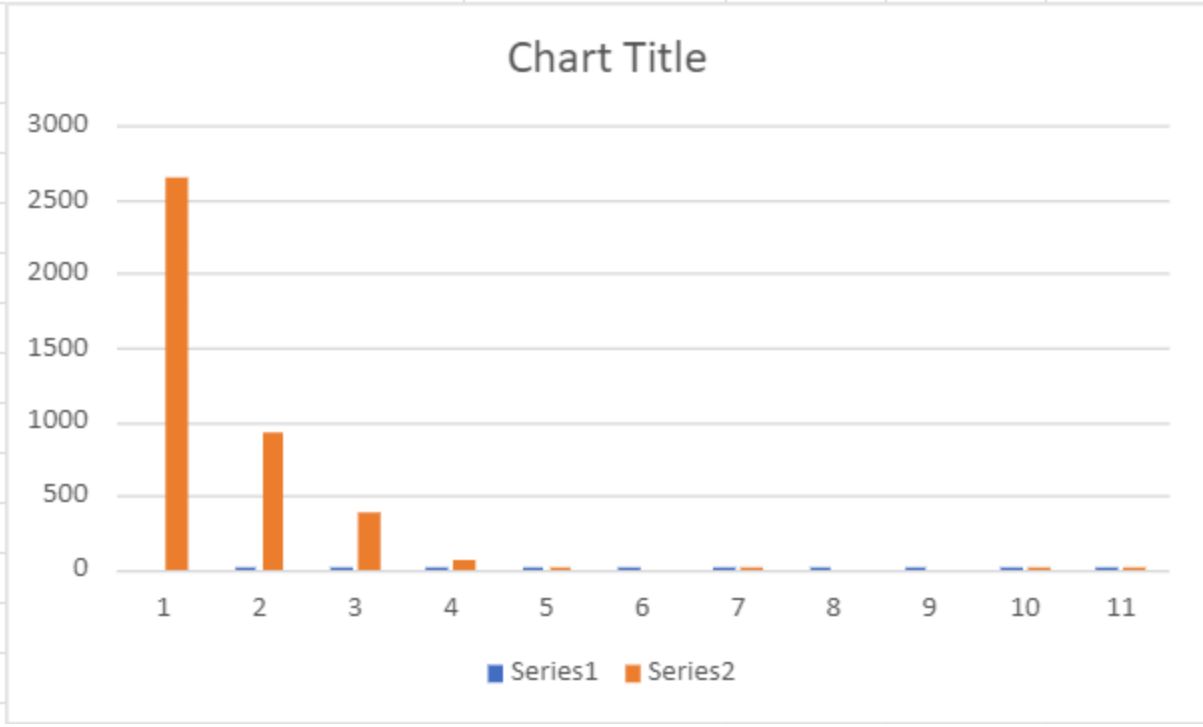
For Owns car there is significant imbalance as most of the applicants who faced problems didnt own a car.This trend carries over for applicants who didnt face problems

	for owns realty		
	1	0	total
Y	2752	31939	34691
N	1274	14034	15308
total	4026	45973	49999



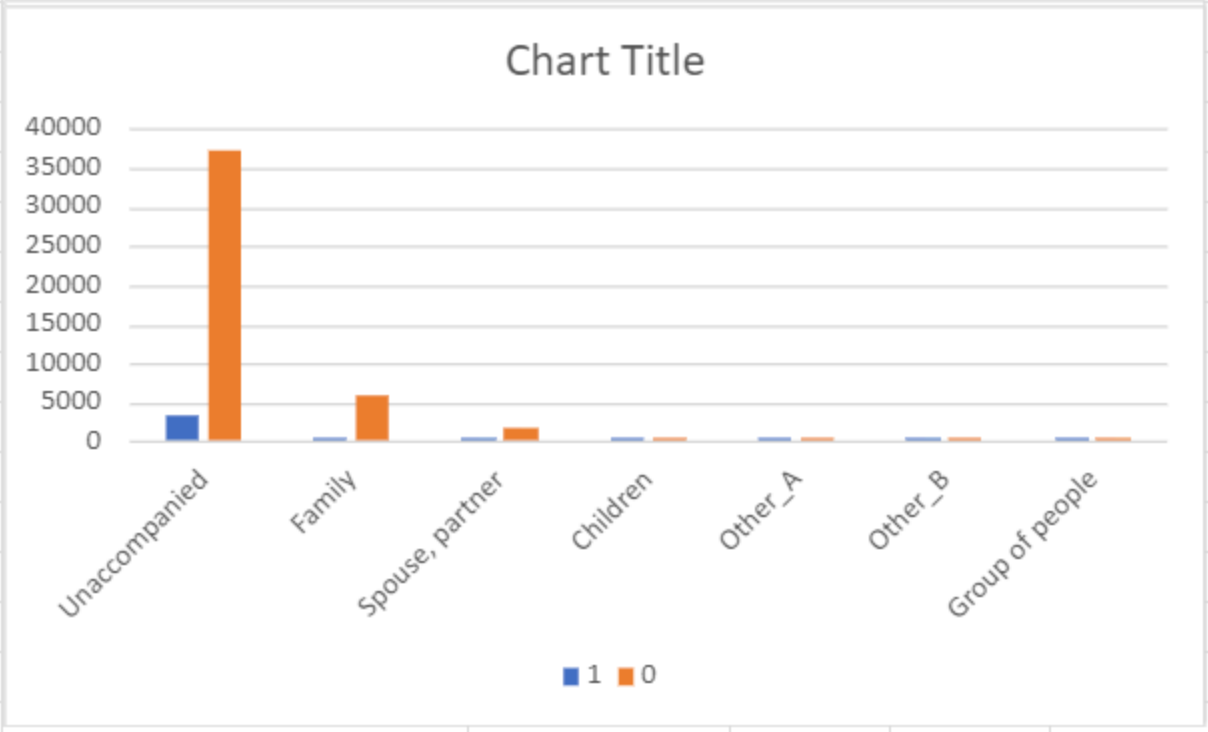
For owns realty we can see significant imbalance as most of the applicants owned some realty

for Cnt_Children			
	1	0	total
0	2644	32272	34916
1	923	923	1846
2	384	3935	4319
3	56	56	112
4	14	59	73
7	0	0	0
5	3	10	13
6	0	0	0
8	0	1	1
9	1	1	2
11	1	0	1
total	4026	37257	49999



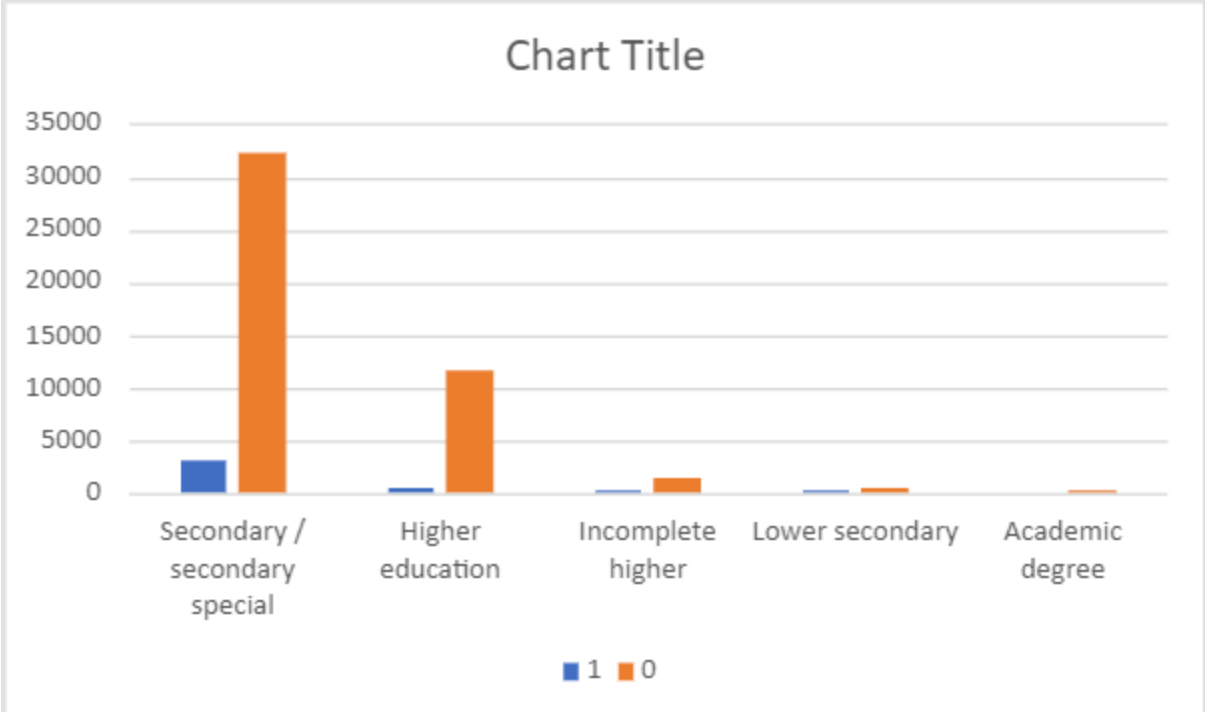
Therefore for Cnt_Children there is heavy imbalance in the data as most of the applicants have zero children

For Type_suite			
	1	0	total
Unaccompanied	3297	37330	40627
Family	499	6050	6549
Spouse, partner	144	1705	1849
Children	47	495	542
Other_A	10	127	137
Other_B	28	231	259
Group of people	1	35	36
total	4026	45973	49999



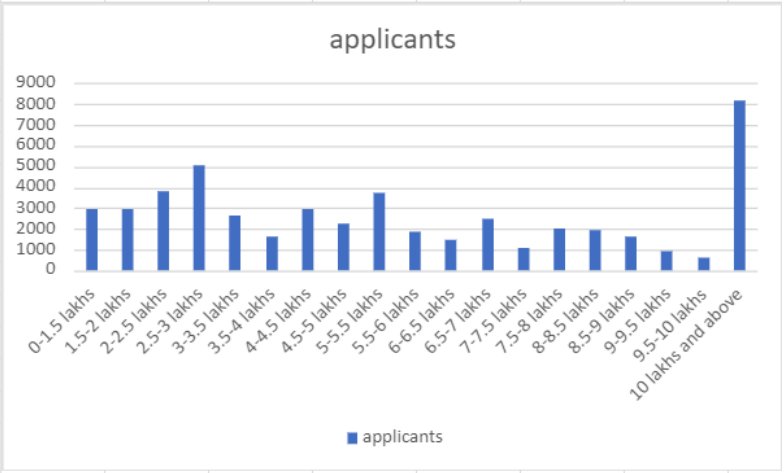
Therefore most of the applicants are unaccompanied

for education type			
	1	0	total
Secondary / secondary special	3209	32363	35572
Higher education	606	11561	12167
Incomplete higher	138	1482	1620
Lower secondary	73	547	620
Academic degree	0	20	20
total	4026	45973	49999



Task 4 : Perform Univariate, Segmented Univariate, and Bivariate Analysis
Formula used : =COUNTIFS(range1, condition1, range2, condition2)

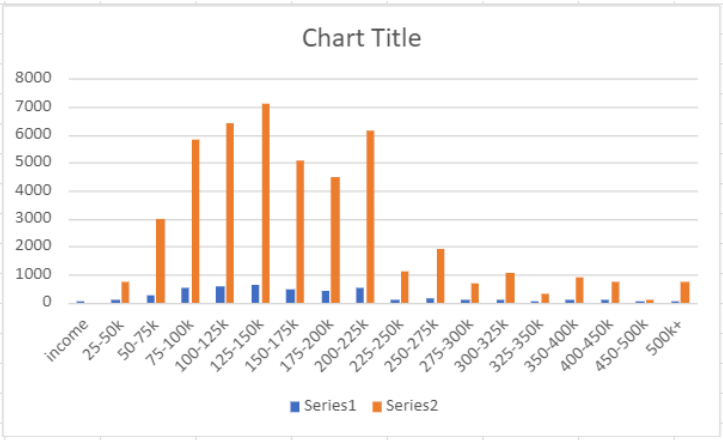
credit	applicants
0-1.5 lakhs	2964
1.5-2 lakhs	2936
2-2.5 lakhs	3822
2.5-3 lakhs	5027
3-3.5 lakhs	2633
3.5-4 lakhs	1622
4-4.5 lakhs	2959
4.5-5 lakhs	2267
5-5.5 lakhs	3707
5.5-6 lakhs	1846
6-6.5 lakhs	1464
6.5-7 lakhs	2445
7-7.5 lakhs	1066
7.5-8 lakhs	1996
8-8.5 lakhs	1911
8.5-9 lakhs	1660
9-9.5 lakhs	905
9.5-10 lakhs	618
10 lakhs and above	8146



Univariate analysis

Univariate analysis involves the examination of a single variable. It provides simple summaries about the data and the measures of the variable.

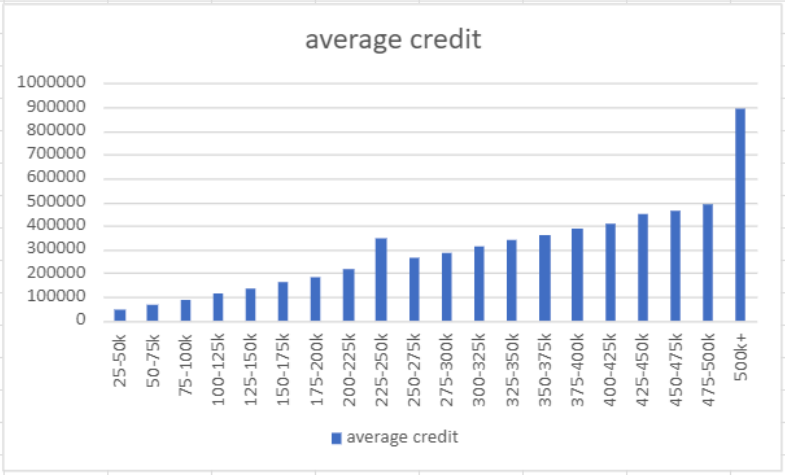
income	1	0
25-50k	63	741
50-75k	241	2980
75-100k	502	5826
100-125k	547	6428
125-150k	615	7126
150-175k	482	5060
175-200k	388	4458
200-225k	527	6121
225-250k	97	1121
250-275k	164	1919
275-300k	62	681
300-325k	111	1076
325-350k	25	322
350-400k	72	909
400-450k	66	719
450-500k	7	63
500k+	36	741



Segmented univariate analysis

Segmented univariate analysis involves analyzing a single variable but within segments or groups of data.

income	average credit
25-50k	47061.2
50-75k	62879.80452
75-100k	88693.00382
100-125k	111021.9838
125-150k	136831.4893
150-175k	160463.6664
175-200k	182799.8441
200-225k	217958.6997
225-250k	347496.6338
250-275k	264380.7461
275-300k	287377.0838
300-325k	312710.0058
325-350k	336954.8093
350-375k	361477.7835
375-400k	386082.3425
400-425k	410188.3026
425-450k	446472.3089
450-475k	461853.5374
475-500k	491111.0577
500k+	895488.9482



Bivariate analysis

Bivariate analysis examines the relationship between two variables. It helps in understanding the association or correlation between them.

Task 5 : Identify Top Correlations for Different Scenarios

Formula used : =CORREL(array1, array2)

Correlation analysis for applicants who didnt face any problems

	cnt children	total income	credit	cnt fam member	region population relative	region rating
cnt children	1	0.009588558	0.00497156	0.880453292	-0.025555665	0.025913889
total income	0.009588558	1	0.069315897	0.011225511	0.029841469	-0.038188511
credit	0.00497156	0.069315897	1	0.063997155	0.095111221	-0.100507425
cnt fam member	0.880453292	0.011225511	0.063997155	1	-0.02303741	0.025985394
region population relative	-0.025555665	0.029841469	0.095111221	-0.02303741	1	-0.532667302
region rating	0.025913889	-0.038188511	-0.100507425	0.025985394	-0.532667302	1

Correlation analysis for applicants who faced problems

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	CNT_FAM_MEMBER	REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.009588558	0.00497156	0.880453292	-0.025555665	0.025913889
AMT_INCOME_TOTAL	0.009588558	1	0.069315897	0.011225511	0.029841469	-0.038188511
AMT_CREDIT	0.00497156	0.069315897	1	0.063997155	0.095111221	-0.100507425
CNT_FAM_MEMBER	0.880453292	0.011225511	0.063997155	1	-0.02303741	0.025985394
REGION_POPULATION_RELATIVE	-0.025555665	0.029841469	0.095111221	-0.02303741	1	-0.532667302
REGION_RATING_CLIENT	0.025913889	-0.038188511	-0.100507425	0.025985394	-0.532667302	1

Blue represents high correlation between the data while red represents low or negative correlation

Results

Hence we were able to leverage microsoft excels data analytics functionalities for Bank Loan Analysis and extracted the required insights