Bank Loan Case Study

Hyperlink of excel file:

https://docs.google.com/file/d/1tXW8VpRY1YRAJFb7Z8mM493sAbPeRDfZ/edit?usp=docslist_api&filetype=msexcel

NOTE: You need to download the above file to open

Hyperlink of video presentation:

https://drive.google.com/file/d/1rtb6SszzjGOuOWpLRmxVLQbps6hYsGXx/view?usp=drivesdk

Data Analytics Task:

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

I have identified missing values of each column by clicking on blanks and then decided to delete the columns with more than 30% null values and handle rest using descriptive statistics. Below is the snapshots of on how I handled nulls for each columns

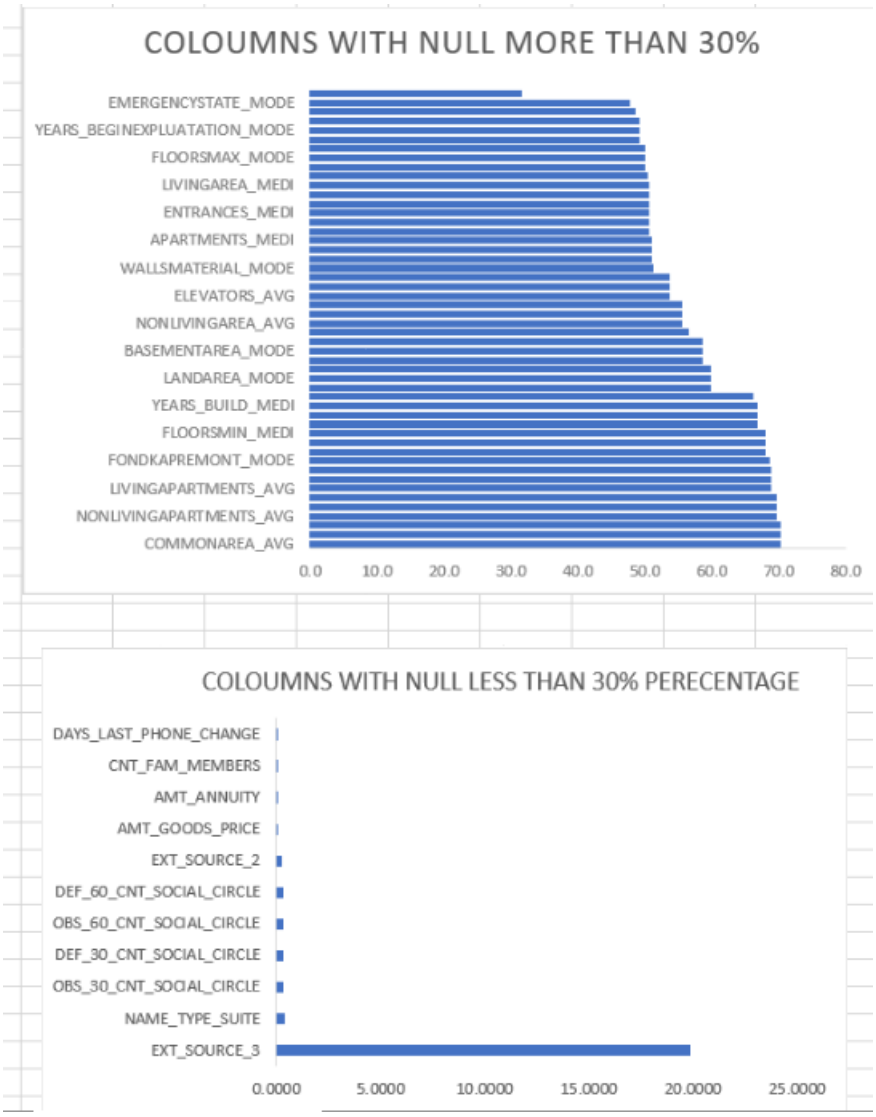
Row Labels	Count of NAME_TYPE_SUITE	
Children	542	
Family	6549	
Group of people	36	
Other_A	137	
Other_B	259	
Spouse, partner	1849	
Unaccompanied	40435	Unaccompanied
Grand Total	**49807**	

COLUMNS	MEAN	MEDIAN
AMT_ANNUITY	27107	24939
AMT_GOODS_PRICE	539060	450000
CNT_FAM_MEMBERS	2	2
EXT_SOURCE_2	1	0.56559
EXT_SOURCE_3	1	0.53528
OBS_30_CNT_SOCIAL_CIRCLE	1	0
DEF_30_CNT_SOCIAL_CIRCLE	0	0
OBS_60_CNT_SOCIAL_CIRCLE	1	0
DEF_60_CNT_SOCIAL_CIRCLE	0	0
DAYS_LAST_PHONE_CHANGE	964	755
AMT_REQ_CREDIT_BUREAU_HOUR	0	0
AMT_REQ_CREDIT_BUREAU_DAY	0	0
AMT_REQ_CREDIT_BUREAU_WEEK	0	0
AMT_REQ_CREDIT_BUREAU_MON	0	0
AMT_REQ_CREDIT_BUREAU_QRT	0	0
AMT_REQ_CREDIT_BUREAU_YEAR	2	1

Null values handling for the above columns will be using the highlighted values

Nulls of qualitative data set is handled using mode and data sets of quantitative data sets are handled using mean and median as show above . Between mean and median , best suited attribute for the column has been chosen as highlighted above.

Below is the graph showing the proportion of nulls categorized as above 30% and below 30%



All the columns in the above 30% category have been eliminated and rest was handled as mentioned above.

B. **Identify Outliers in the Dataset:**

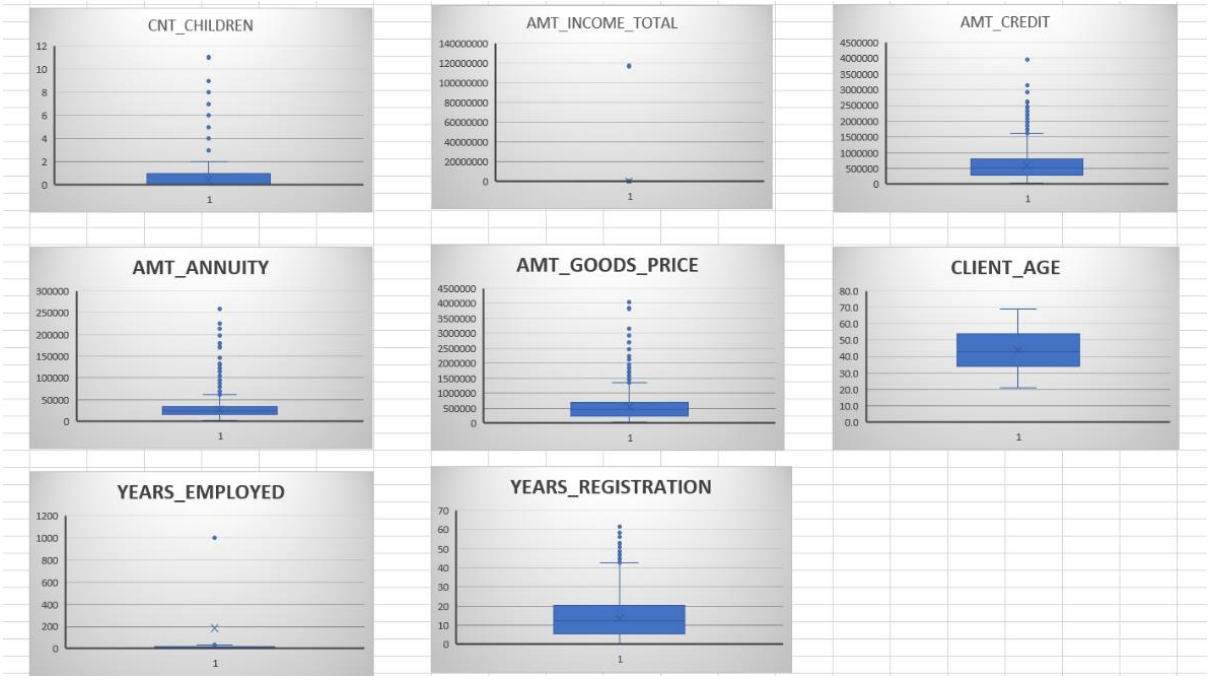
Below is the snapshot of identifying outliers and eliminating it

Possible Columns to have outliers are

COLUMNS	Q1	Q3	IQR	UPPER LMIT	LOWER LIMIT
CNT_CHILDREN	0	1	1	2.5	0
AMT_CREDIT	270000	808650	538650	1616625	0
AMT_ANNUITY	16456.5	34596	18139.5	61805.25	0
AMT_GOODS_PRICE	238500	679500	441000	1341000	0
CLIENT_AGE	33.915	53.82	19.905	83.6775	4.0575
YEARS_EMPLOYED	2.6	15.7	13.1	35.35	0
YEARS_REGISTRATION	5.5	20.4	14.9	42.75	0

These above columns are possible columns to have outliers, I calculated Q1,Q3,IQR and upper and lower limit for each of these columns using excel functions and filtered these columns to have values between upper and lower limit, and handled all those whose values

which don't come under the limit by deleting the most which would cause a problem in the later analysis



Above are graphs showing outliers on each selected column before elimination using a bow plot, as we can see only client age is exempted from having an outlier and rest of the columns do contain outliers.

C. **Analyse Data Imbalance:**

There is data imbalance in the dataset and the ratio of data imbalance is given below

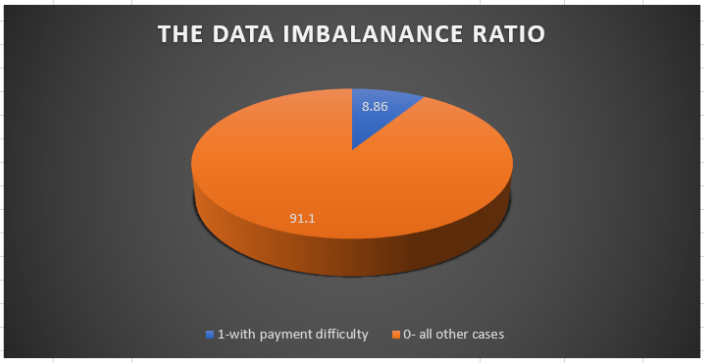
THE DATA IMBALANANCE RATIO

Row Labels	Count of TARGET
0	35065
1	3409
Grand Total	38474

1-with payment difficulty	8.86
0- all other cases	91.1

NOTE

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 loan in our sample, 0 - all other cases



The above picture shows data imbalance between non-defaulters and defaulters with ratio of 92.1 is to 8.86. There is total of 3409 clients having payment difficulties out of 38474 total clients .Pivot table was used to calculate the above in which took row as Target and values as count of target to arrive at the above result.

D. **Performing Univariate, Segmented Univariate, and Bivariate Analysis:**

NOTE: All the calculations and methods for below will be explained in detail in the video presentation.

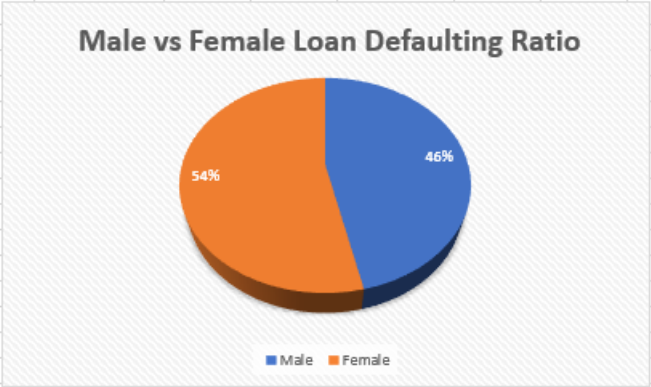
A. Univariate Analysis:

Univariate analysis is the analysis of a single variable to describe its distribution, central tendency, and variability using summary statistics like mean, median, mode, and visualizations such as histograms or box plots. Below are snapshots of all univariate analysis

TOTAL MALE	14435
TOTAL FEMALE	24035
TOTAL	38470

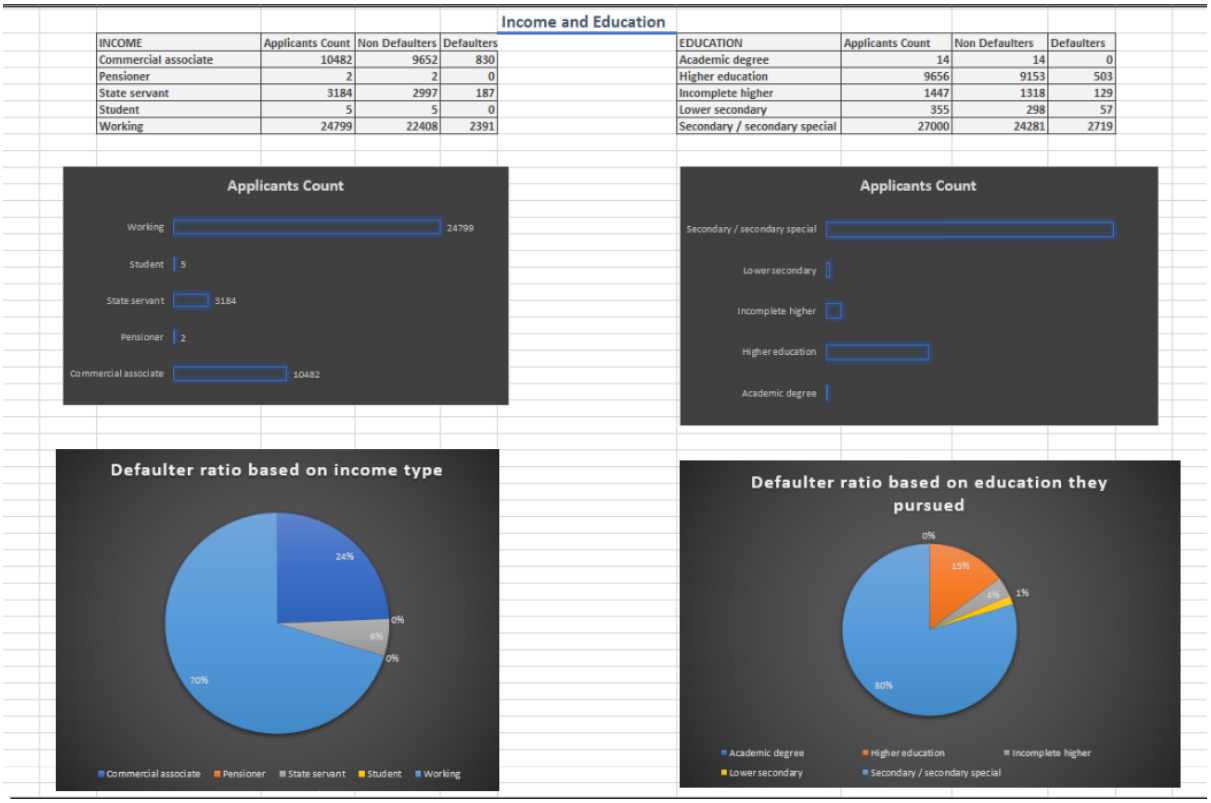
MALE		Female	
Non Defaulter	Defaulter	Non Defauter	Defaulter
12851	1584	22211	1824
89.03%	10.97%	92.41%	7.59%

Male vs Female Loan Defaulting Ratio		
Male	Female	Total
1584	1824	3408
46.48%	53.52%	



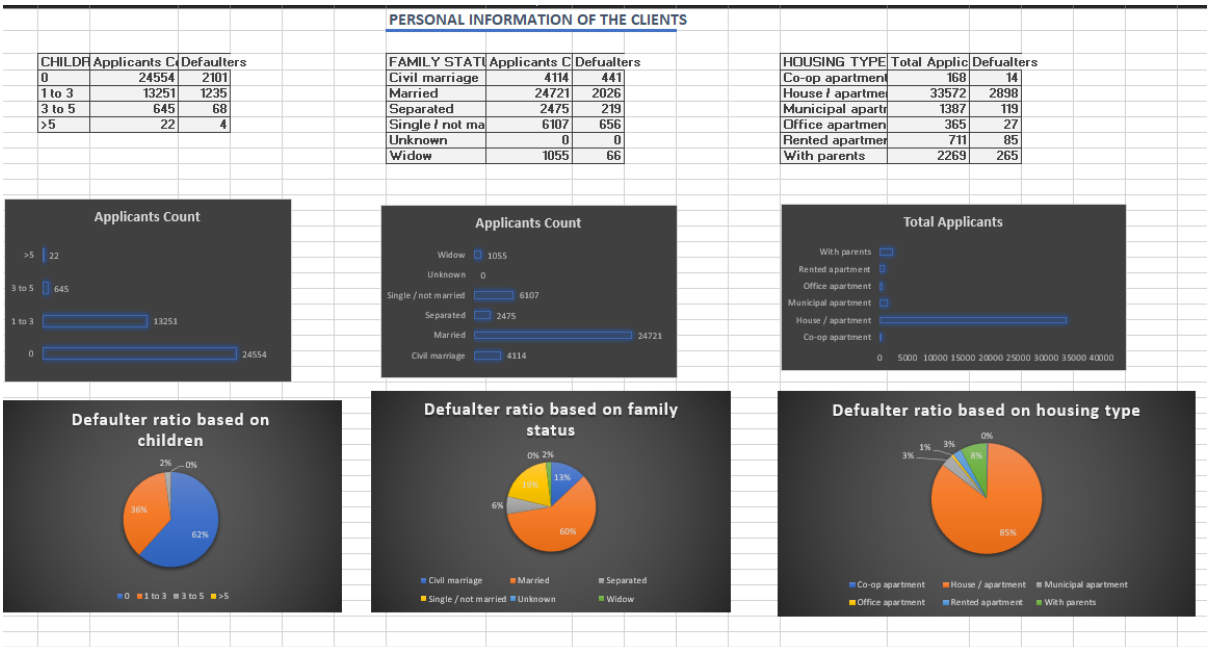
Above snapshot gives the univariate analysis of male to female defaulting ratio. It was calculated initially by taking the count of male and female using excel functions and later manipulated the result to arrive at above answer.

INSIGHT: It can be seen that Male is leading over female in defaulting loan even though the count of male is lesser than female. It is obvious that, there are more chances of a male defaulting the loan and granting loan to male can be avoided or should carefully examined before defaulting.



Above image shows the applicants count and defaulter ratio based on income type and education background. All the above result was calculated by filtering and using excel functions which will be explained in detail in the video presentation.

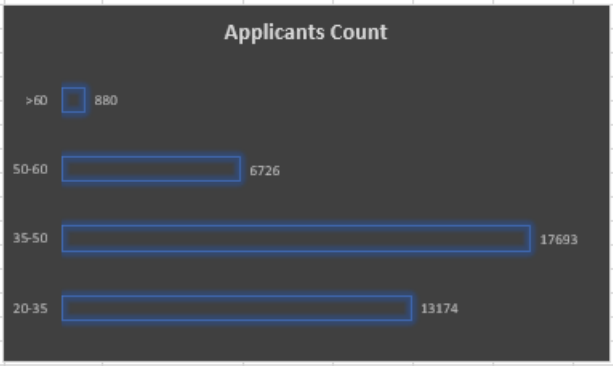
INSIGHT: Working and Secondary/ Secondary Special categories have the highest loan takers and commercial associate and Secondary/Secondary Special lead the default ratio. Therefore lending loans to this category can be avoided.

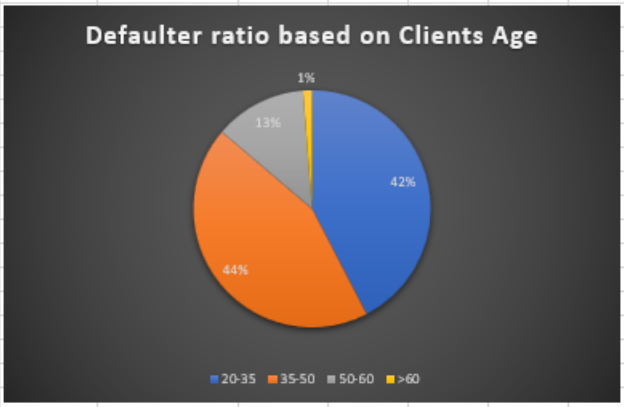


Above is the univariate analysis of client based on client's personal information.

INSIGHT: Clients with zero children, who are married and who owns a house or apartment are more likely to take loans and more likely to default the loan as show above

Clients_A	Applicants Co	Defaulters
20-35	13174	1445
35-50	17693	1494
50-60	6726	427
>60	880	42





Above is the univariate analysis based on client's age

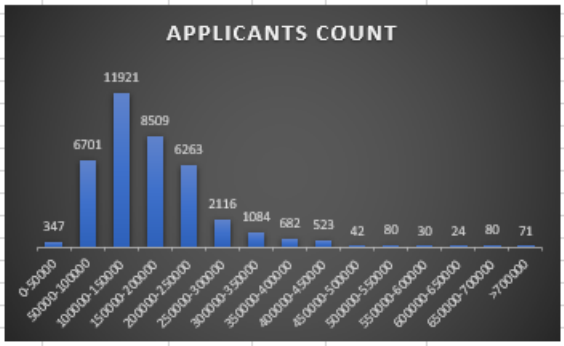
INSIGHT: Clients with age group between 30 to 50 are more likely to take loans and are more likely to default.

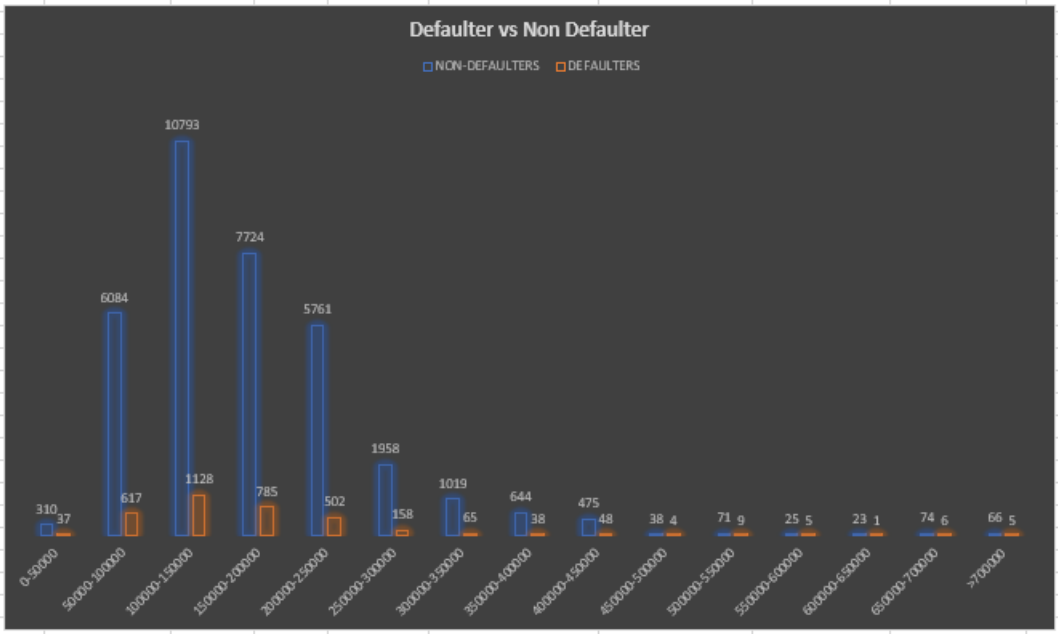
B. Segmented Univariate Analysis:

Segmented univariate analysis examines a single variable across different subgroups or categories. As given below

MAX	2250000
MIN	25650
TOTAL Applic	38473

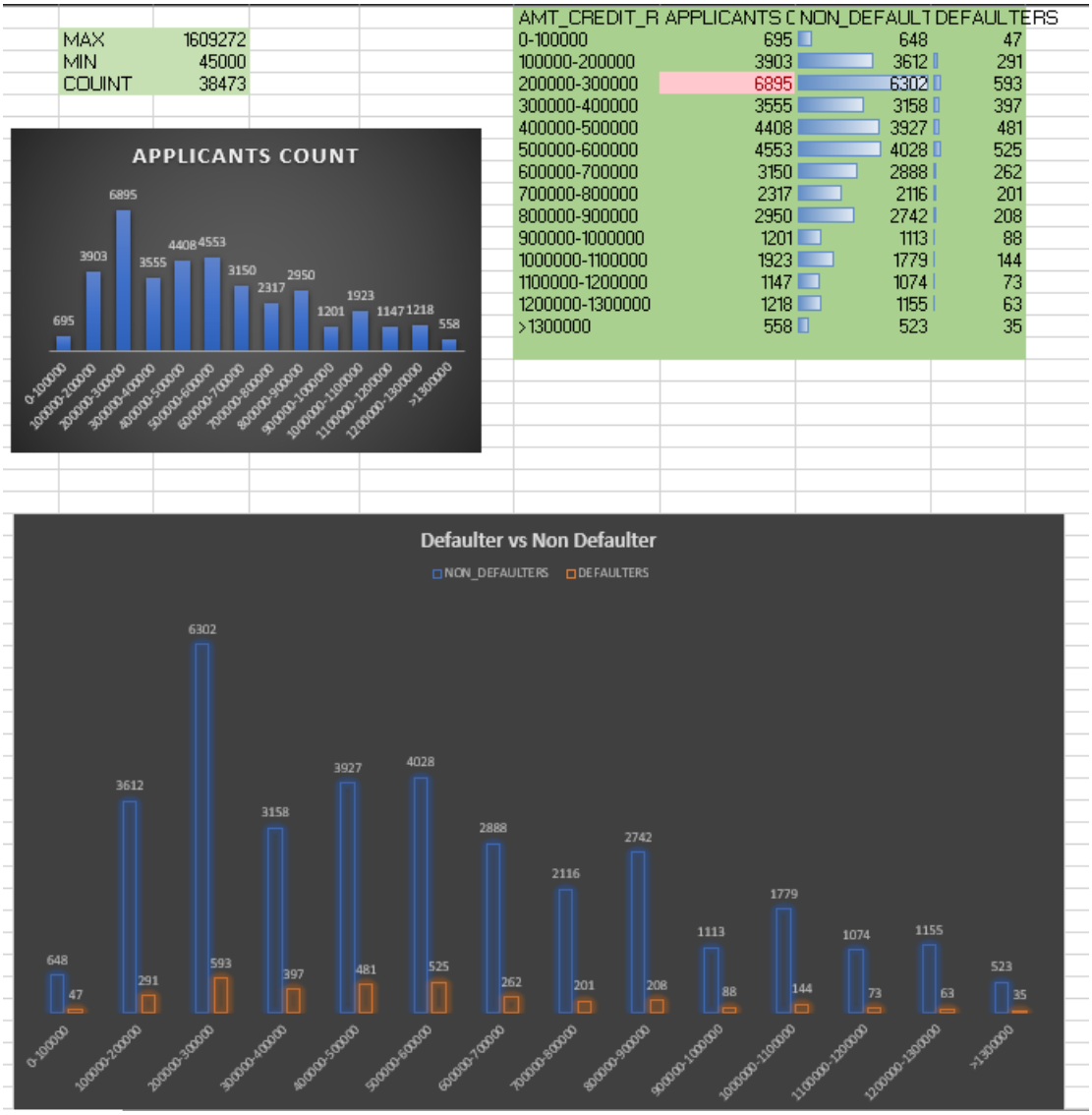
AMT_INCOME_R	APPLICANTS C	NON-DEFAULT	DEFAULTERS
0-50000	347	310	37
50000-100000	6701	6084	617
100000-150000	11921	10793	1128
150000-200000	8509	7724	785
200000-250000	6263	5761	502
250000-300000	2116	1958	158
300000-350000	1084	1019	65
350000-400000	682	644	38
400000-450000	523	475	48
450000-500000	42	38	4
500000-550000	80	71	9
550000-600000	30	25	5
600000-650000	24	23	1
650000-700000	80	74	6
>700000	71	66	5





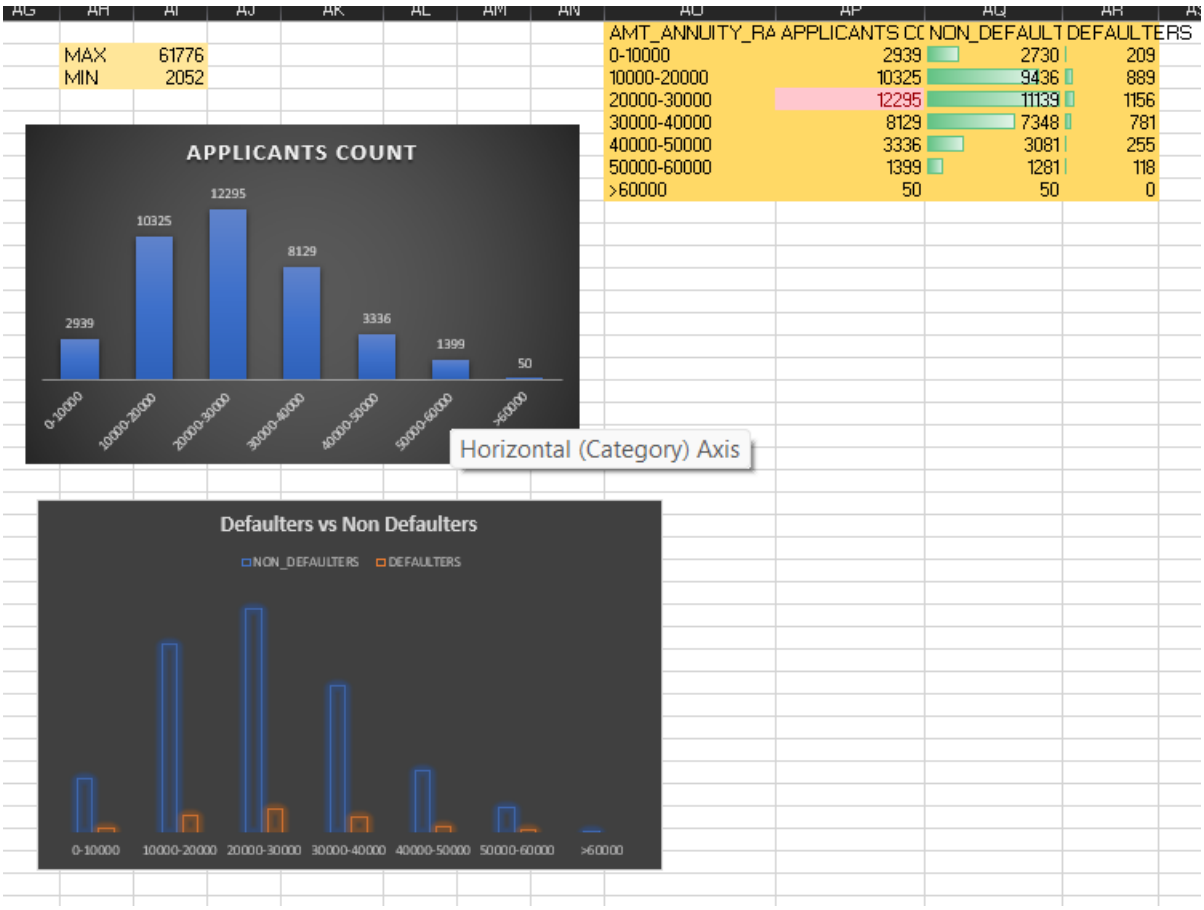
Above image shows the count and Non-defaulter to defaulter ratio based on client's income

INSIGHT: Clients with low income say less than 4,50,000 are more likely to take loans and some of them are likely to default the same.



Above is the segmented analysis of clients based on the total loan amount credit.

INSIGHT: There are more chances of client's wanting loan in the range 2 to 3 lakh and clients with loan range 5 to 6 lakh are more likely to default compared all the category.



Above is the segmented univariate analysis based on amount annuity of clients

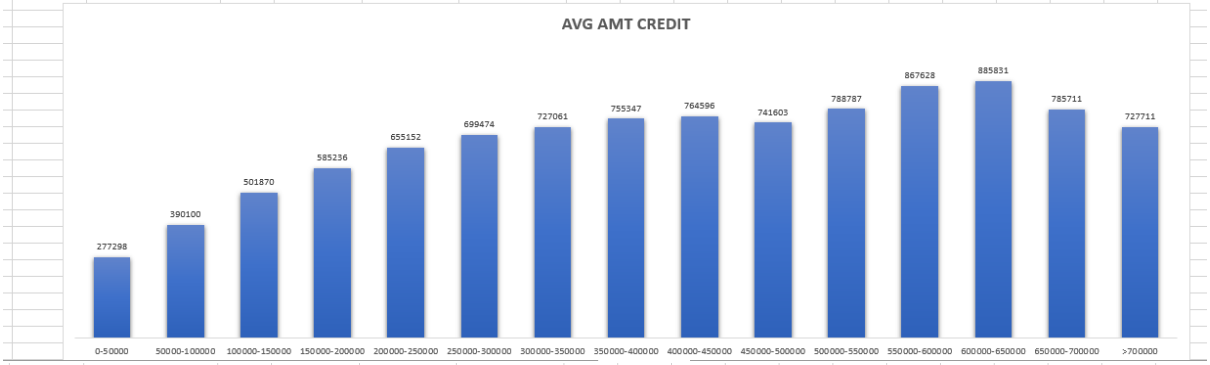
INSIGHT: Expected annual annuity of a client taking loan will be between 20000 to 30000.

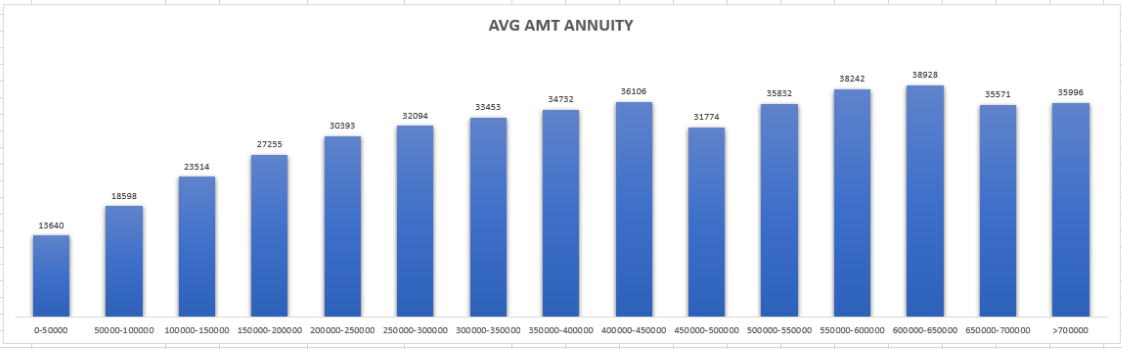
C. Bivariate Analysis:

Bivariate analysis involves the examination of two variables to explore their relationship or correlation. It helps identify patterns, associations, or differences between the two variables using techniques such as scatter plots, correlation coefficients, and cross-tabulations. Below is the snapshot of bivariate analysis.

MIN INCOME	25650

Income-Range	AVG AMT CREDIT	AVG AMT ANNUITY
0-50000	277298	13640
50000-100000	390100	18598
100000-150000	501870	23514
150000-200000	585236	27255
200000-250000	655152	30393
250000-300000	699474	32094
300000-350000	727061	33453
350000-400000	755347	34732
400000-450000	764596	36106
450000-500000	741603	31774
500000-550000	788787	35832
550000-600000	867628	38242
600000-650000	885831	38928
650000-700000	785711	35571
>700000	727711	35996



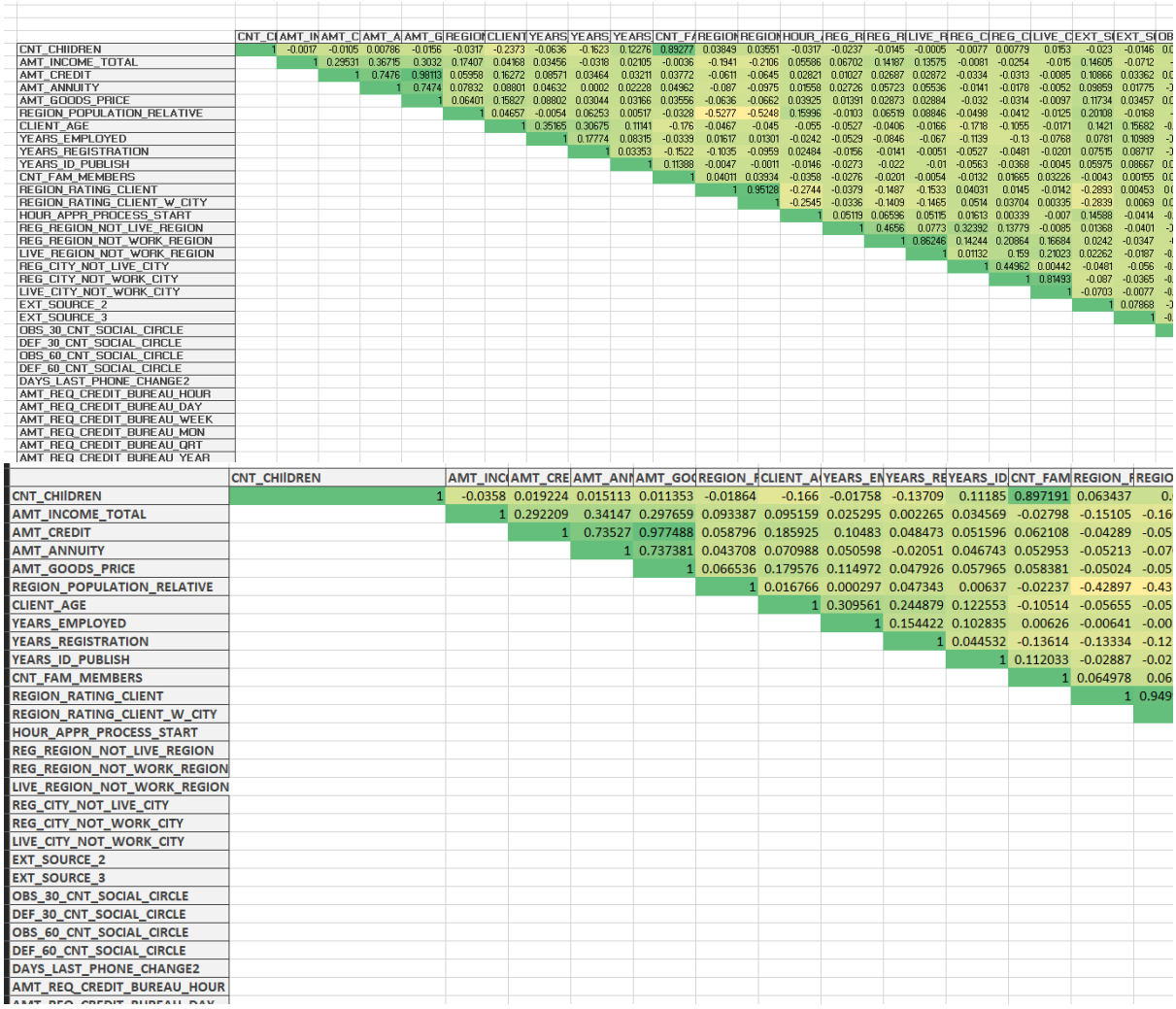


Here I conducted a bivariate analysis to understand the relationship between income ranges and two key variables: average amount of credit and average amount of annuity.

INSIGHT: The amount of credit increases as the income increases and amount annuity follows the same pattern , which increases as the income increses.

E. Identify Top Correlations for Different Scenarios:

Below are the glimpses of all the co relations calculated categorized as Non-defaulter and Defaulter



	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	CLIENT_AGE	YEARS_EMPLOYED	YEARS_REGISTRATION	YEARS_ID_PUBLISH	CNT_FAM_MEMBERS	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	HOUR_APPR_PROCESS_START	REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	EXT_SOURCE_2	EXT_SOURCE_3	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	DAYS_LAST_PHONE_CHANGE2	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1	-0.0358	0.019224	0.015113	0.011353	-0.01864	-0.166	-0.01758	-0.13709	0.11185	0.897191	0.063437	0.																				
AMT_INCOME_TOTAL		1	0.292209	0.34147	0.297659	0.093387	0.095159	0.025295	0.002265	0.034569	-0.02798	-0.15105	-0.16																				
AMT_CREDIT			1	0.73527	0.977488	0.058796	0.185925	0.10483	0.048473	0.051596	0.062108	-0.04289	-0.05																				
AMT_ANNUITY				1	0.737381	0.043708	0.070988	0.050598	-0.02051	0.046743	0.052953	-0.05213	-0.07																				
AMT_GOODS_PRICE					1	0.066536	0.179576	0.114972	0.047926	0.057965	0.058381	-0.05024	-0.05																				
REGION_POPULATION_RELATIVE						1	0.016766	0.000297	0.047343	0.00637	-0.02237	-0.42897	-0.43																				
CLIENT_AGE							1	0.309561	0.244879	0.122553	-0.10514	-0.05655	-0.05																				
YEARS_EMPLOYED								1	0.154422	0.102835	0.00626	-0.00641	-0.00																				
YEARS_REGISTRATION									1	0.044532	-0.13614	-0.13334	-0.12																				
YEARS_ID_PUBLISH										1	0.112033	-0.02887	-0.02																				
CNT_FAM_MEMBERS											1	0.064978	0.06																				
REGION_RATING_CLIENT												1	0.949																				
REGION_RATING_CLIENT_W_CITY																																	
HOUR_APPR_PROCESS_START																																	
REG_REGION_NOT_LIVE_REGION																																	
REG_REGION_NOT_WORK_REGION																																	
LIVE_REGION_NOT_WORK_REGION																																	
REG_CITY_NOT_LIVE_CITY																																	
REG_CITY_NOT_WORK_CITY																																	
LIVE_CITY_NOT_WORK_CITY																																	
EXT_SOURCE_2																																	
EXT_SOURCE_3																																	
OBS_30_CNT_SOCIAL_CIRCLE																																	
DEF_30_CNT_SOCIAL_CIRCLE																																	
OBS_60_CNT_SOCIAL_CIRCLE																																	
DEF_60_CNT_SOCIAL_CIRCLE																																	
DAYS_LAST_PHONE_CHANGE2																																	
AMT_REQ_CREDIT_BUREAU_HOUR																																	
AMT_REQ_CREDIT_BUREAU_DAY																																	

Above is the calculation of corelations to all the numeric columns in the final dataset. The above calculation was calculated using the excel function COREL. Conditional formatting was used to highlight the top corelations and below is the top co relation from each section.

Top Corelations

Non Defaulter			Defaulter		
Variable1	Variable2	Corelation	Variable1	Variable2	Corelation
CNT_CHILDREN	CNT_FAM_MEMBERS	0.892769	CNT_CHILDREN	CNT_FAM_MEMBERS	0.897191
AMT_CREDIT	AMT_ANNUITY	0.7476	AMT_CREDIT	AMT_ANNUITY	0.73527
AMT_CREDIT	AMT_GOODS_PRICE	0.91883	AMT_CREDIT	AMT_GOODS_PRICE	0.977488
AMT_ANNUITY	AMT_GOODS_PRICE	0.7474	AMT_ANNUITY	AMT_GOODS_PRICE	0.737381
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.951283	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.949972
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.86246	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.804259
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.81493	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.769521
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998352	OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.997992
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.855422	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.892084

INSIGHT: The defaulter group tends to show stronger correlations in areas related to financial behaviour and defaults (e.g., credit and goods price correlation). In contrast, non-defaulters show higher correlations in regional factors and social circle metrics. This suggests that external factors, such as location and social environment, may play a larger role in their financial behaviour. Meanwhile, defaulters have tighter financial links, especially in terms of credit and goods pricing.

Project Description: The project is based on real life bank loan dataset, which was aimed to prevent loss to the bank if given because of defaulting. Basically, bank or company should be profited by lending or not lending loan.

Approach: I initially read the question in detail and understood the problem before executing anything, once I understood the question, I used my knowledge in excel and statistics to execute the same. Excel functions and tools were used.

Tech-Stack Used: Microsoft Excel 2019

Insight: Insights are given above in the required tasks

Result: I was able improve my skills in excel, statistics and problem solving. It has helped improve my patience, consistency and perseverance which eventually helped me grow overall as data analyst.