Name:- Yash Rajendra Gaikwad
Data Analytics Trainee
Project 6:- Bank Loan Case Study.
Software Used:- Microsoft Excel.

❖ Analysis done on following Points:-

A) Identify Missing Data and Deal with it Appropriately:- As a data analyst, you come across missing dat in the loan application dataset. it is essential to handle missing data effectively to ensure the accuracy of the analysis.

Task:- Identify missing data in the dataset and decide on an appropriate method to deal with it usong excel built-in function and features.

B) Identify Outliners in the Dataset:- Outliners can significantly impact the analysis and distort the results. you need to identify outliner in the loan application dataset.

Task:- Detect and identift outliners in the dataset using Excel statistical function and features, focusing on numberical variables.

C) Analyze Data Imbalance:- Data imbalance can affect the accuracy of the analysis, especially for binary classification problems.

Task:- Determine if there is data imbalance in dataset and calculate ratio of data imbalance using Excel function.

D) Perform Univarite, Segmented Univariate and Bivariate Analysis:- To gain insight into the driving factors of loan default, it is important to conduct various analysis on cousumer and loan attributes.

Task:- Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variabledistributions for different scenarios, and bivariate analysis to explore relationship between variables and the target variable using Excel Functions and features.

E) Identify Top Correlations for Differenr Scenarios:- Understanding the correlation between variables and the target variable can provide insights into stromg indicators ofloan default.

Task:- Segment the dataset based on different Scenarious. and identify the top correlations for each segmented data using Excel Functions.

- **MICROSOFT EXCEL FILE:-** https://docs.google.com/spreadsheets/d/1tgmGcpx9cmGHbVpD6r4QAubecG-PgD_q/edit?usp=drive_link&ouid=103974361659264463652&rtpof=true&sd=true
- **VIDEO LINK:-**https://www.loom.com/share/71722adfb0424bb59daa57fcc3e03340?sid=0dbcd61a-5f08-4723-b32b-2e9e41d52d1a

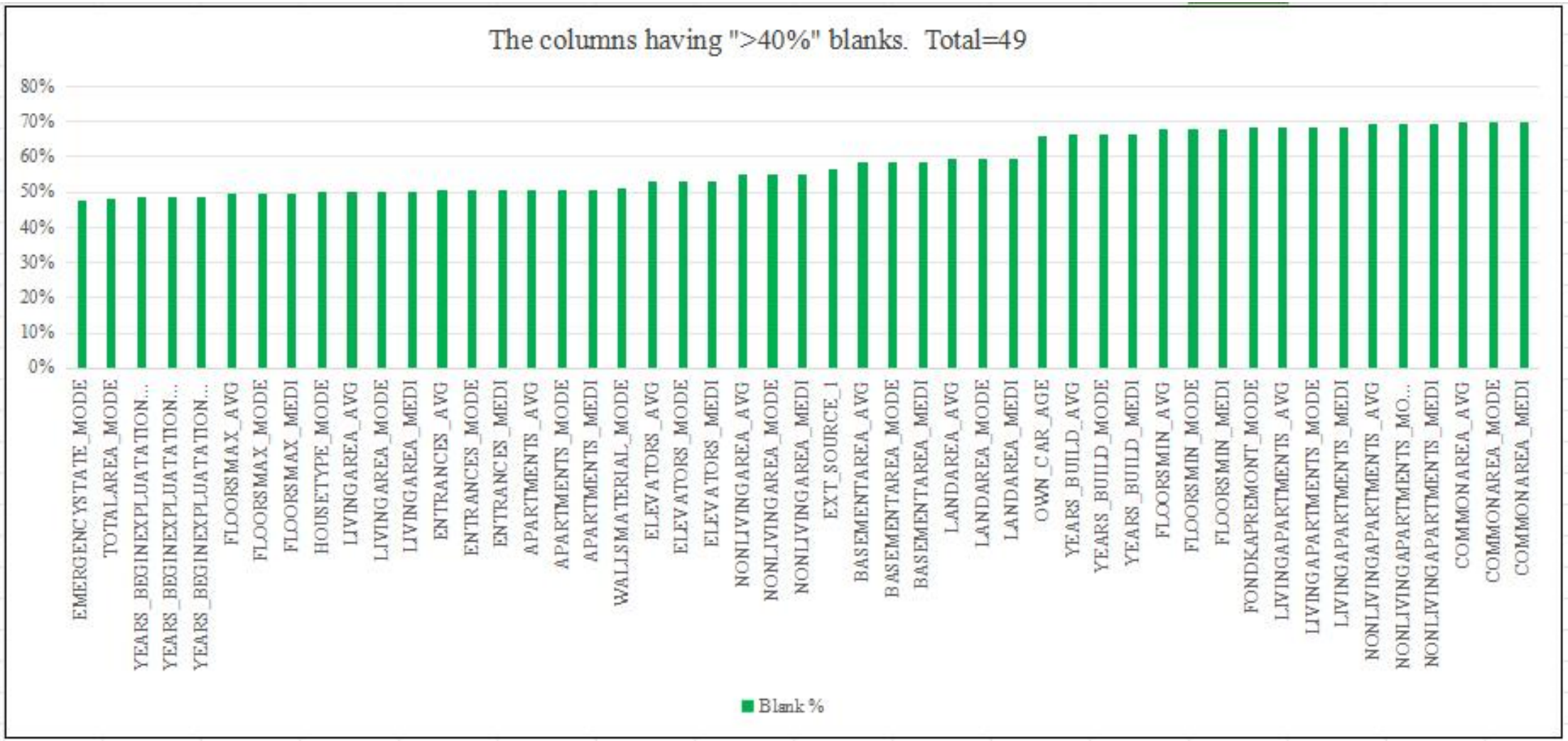
A) Identify Missing Data and Deal with it Appropriately:- As a data analyst, you come across missing dat in the loan application dataset. it is essential to handle missing data effectively to ensure the accuracy of the analysis.

Task:- Identify missing data in the dataset and decide on an appropriate method to deal with it usong excel built-in function and features.

- **Process:-**

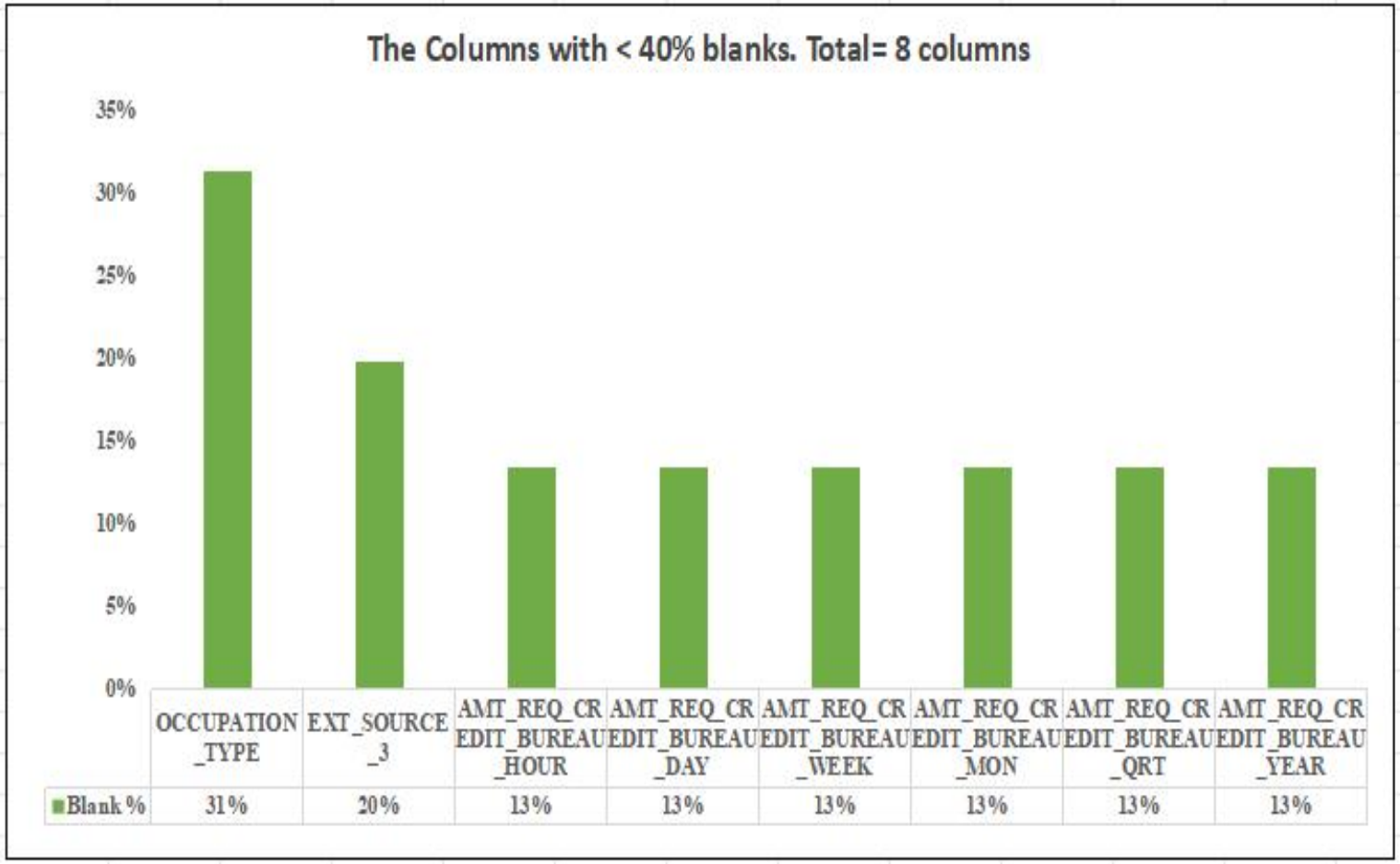
- First to count the blank values in column we use COUNTBLANK formula and find the blank value for the all columns.
- After that we calculate the percentage of blank value for every column. we simpally devide the blank value from the total number of values in the column. also convert that value into Percentage from the home tab.
- First, we drop the column which have more then 40% blank values.
- In column NAME_TYPE_SUIT we fill the missing values with “Unaccompained”.
- Also in column OCCUPATION_TYPE we fill the missing values with “Unknown”.
- For the columns with numbers we fill it with the Median value of the column.
- We use the Column chart for visualization for both chart.

❖ **The columns having More then 40% of blank:-** There are total 49 columns which have more then 40% of blank values. * We drop the columns which have more then 40% blanks



❖ **The columns having less then 40% blanks:-** There are total 8 columns which have less then 40% of blank values.

The Columns with < 40% blanks. Total= 8 columns	
Columns	Blank
OCCUPATION_TYPE	31%
EXT_SOURCE_3	20%
AMT_REQ_CREDIT_BUREAU_HOUR	13%
AMT_REQ_CREDIT_BUREAU_DAY	13%
AMT_REQ_CREDIT_BUREAU_WEEK	13%
AMT_REQ_CREDIT_BUREAU_MON	13%
AMT_REQ_CREDIT_BUREAU_QRT	13%
AMT_REQ_CREDIT_BUREAU_YEAR	13%



B) Identify Outliners in the Dataset:- Outliners can significantly impact the analysis and distort the results. you need to identify outliner in the loan application dataset.

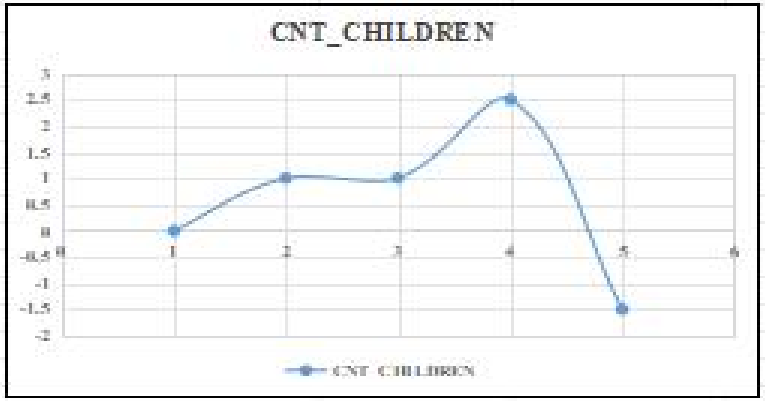
Task:- Detect and identift outliners in the dataset using Excel statistical function and features, focusing on numberical variables.

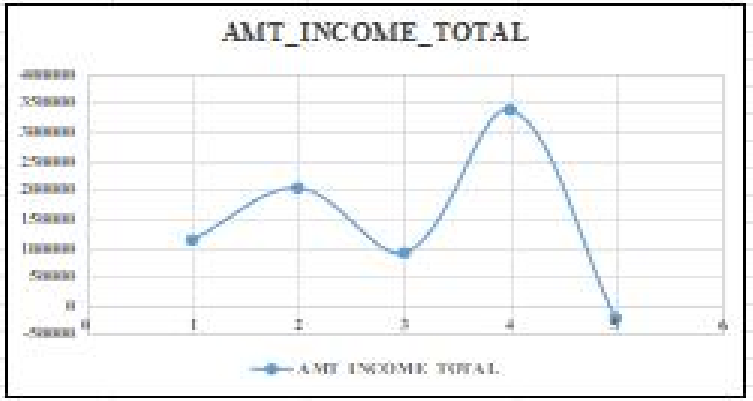
- **Process:-**

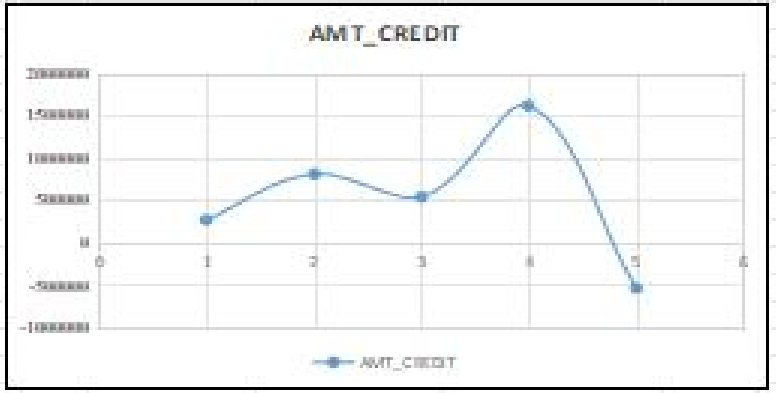
- First we take data on another sheet, then we use the QUARTILE formula for the 1st quart and call it Q1.
- After that we use QUARTILE formula for 3rd quart and call it Q2.
- After that we find Inter Quartial Range (IQR). IQR=Q3-Q1
- Lastly we find the Upper Limit and Lower limit.

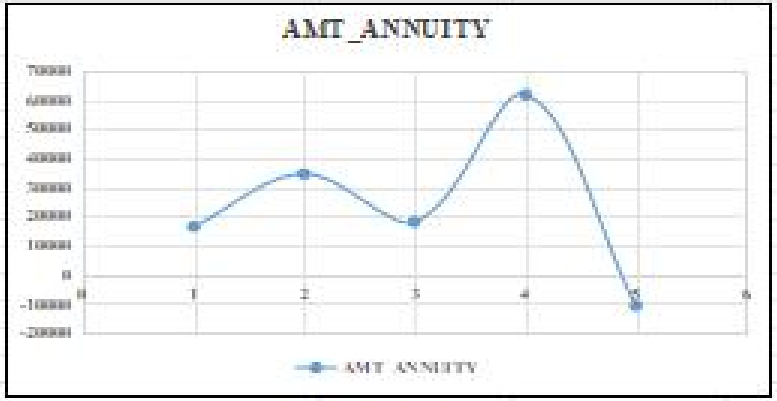
  UPPER LIMIT= Q3+(1.5*IQR) LOWER LIMIT=Q1-(1.5*IQR)
- We use the Scatter Plot chart for Visualization.

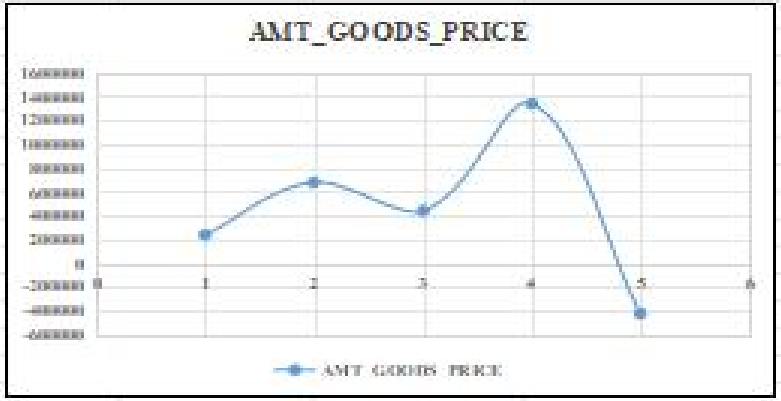
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION
Q1	0	112500	270000	16456.5	238500	0.010006	33.91369863	2.556164384	5.473972603
Q3	1	202500	808650	34596	679500	0.028663	53.81917808	15.66575342	20.44794521
IQR = Q3-Q1	1	90000	538650	18139.5	441000	0.018657	19.90547945	13.10958904	14.9739726
Upper Limit= Q3+(1.5*IQR)	2.5	337500	1616625	61805.25	1341000	0.0566485	83.67739726	35.33013699	42.90890411
Lower Limit=Q1-(1.5*IQR)	-1.5	-22500	-537975	-10752.75	-423000	-0.0179795	4.055479452	-17.10821918	-16.9869863

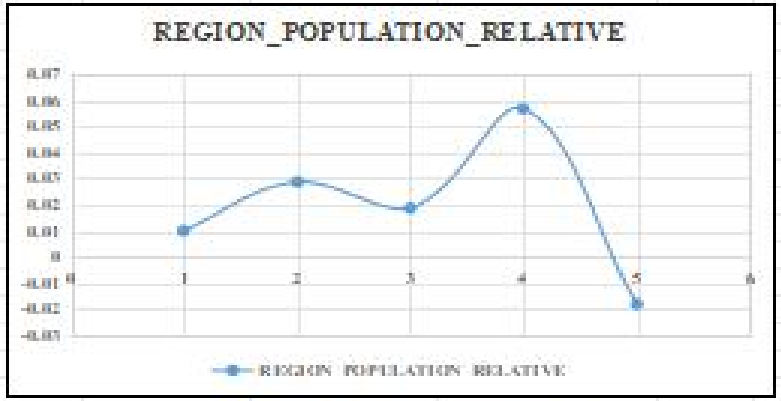

CNT_CHILDREN
CNT_CHILDREN


AMT_INCOME_TOTAL
AMT_INCOME_TOTAL


AMT_CREDIT
AMT_CREDIT


AMT_ANNUITY
AMT_ANNUITY


AMT_GOODS_PRICE
AMT_GOODS_PRICE


REGION_POPULATION_RELATIVE
REGION_POPULATION_RELATIVE


YEARS_BIRTH
YEARS_BIRTH


YEARS_EMPLOYED
YEARS_EMPLOYED


YEARS_REGISTRATION
YEARS_REGISTRATION

C) Analyze Data Imbalance:- Data imbalance can affect the accuracy of the analysis, especially for binary classification problems.

Task:- Determine if there is data imbalance in dataset and calculate ratio of data imbalance using Excel function.

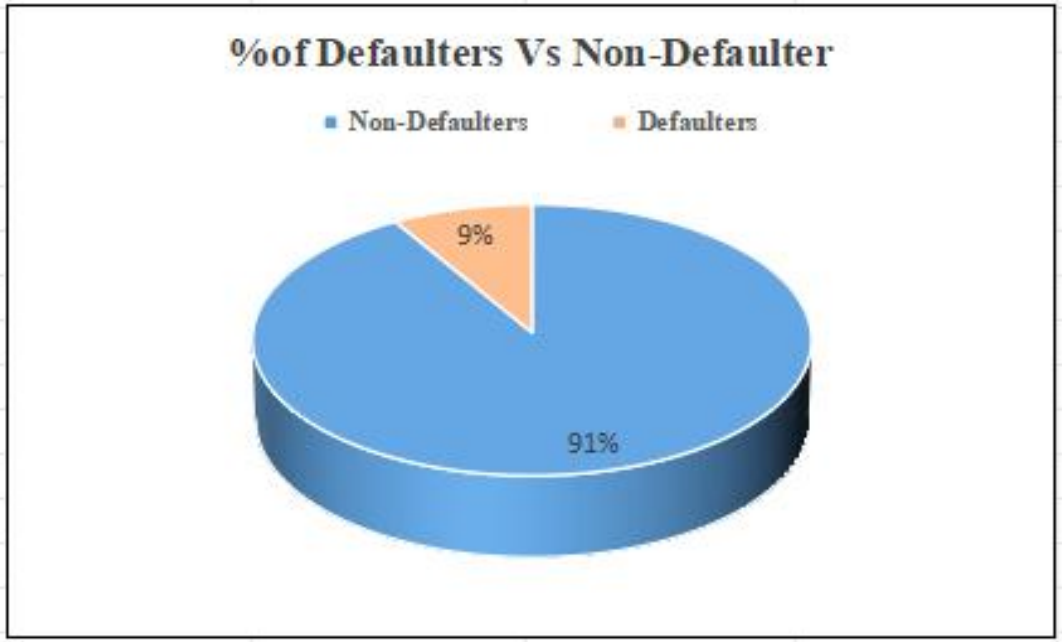
- **Process:-**

- For finding data imbalance we use the Target column. so, we copy column to another sheet.
- We use the COUNTIFS formula to find the Defaulter and Non-Defaulter.
- In target collume 0 indicates Non-Defaulters customers and 1 indicate Defaulter customers.
- We can also use Pivot table. placing Target in Rows and the Count of Target into Values.
- Now, we find the maximum and minimum value in count of target colume by using MAX and MIN formula.
- After that we devide the MIN by MAX and we will get the Ration of Imbalance.
- Lastly we use the Pie Chart to visulization of chart.

❖ Result:- The Ratio Of Imbalance is 0.09.

	Target	Count Of Targ	% of Target
Non-Defaulters	0	37549	91%
Defaulters	1	3520	9%
	Total	**41069**	**100%**

Max	37549
Min	3520
Ratio Of Imbalance (Min/Max)	0.09



D) Perform Univarite, Segmented Univariate and Bivariate Analysis:- To gain insight into the driving factors of loan default, it is important to conduct various analysis on cousumer and loan attributes.

Task:- Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variabledistributions for different scenarios, and bivariate analysis to explore relationship between variables and the target variable using Excel Functions and features.

- **Process:-**
- **Univarite Analysis:-**

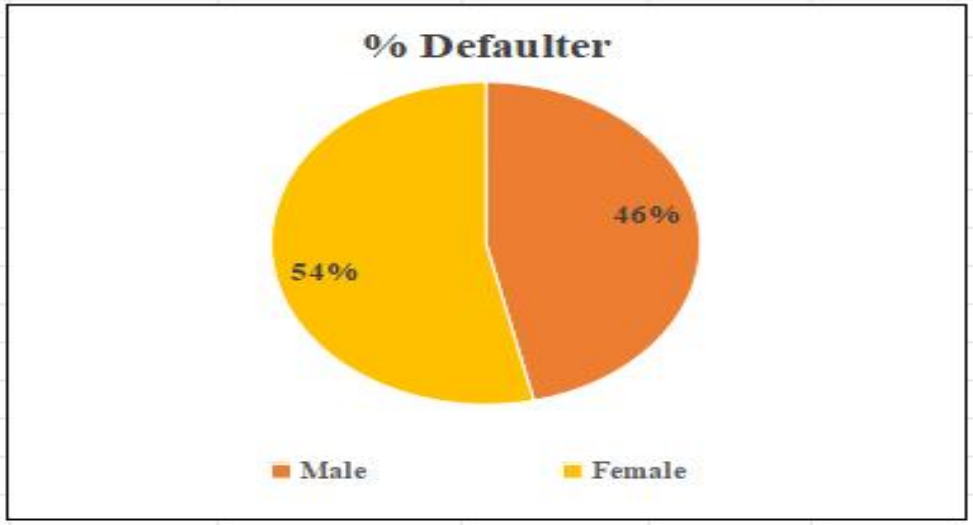
1) Gender Analysis:- *The more number of Females taken lone then the Males but % Defaulters in female is more.

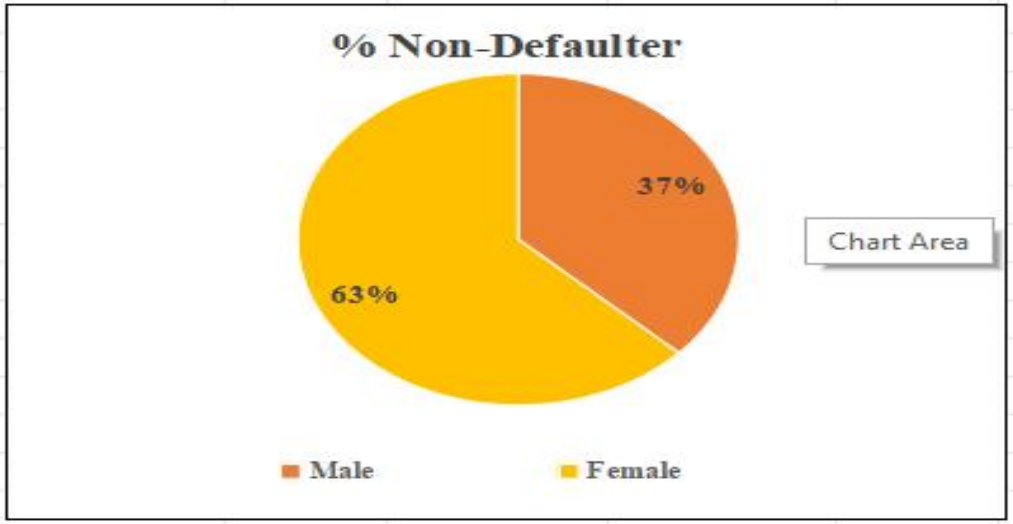
*Male have less % of Defaulters.

Gender	Non-Defaulter	Defaulter
Male	13898	1634
Female	23649	1886
Total	37547	3520

% Defaulter	
Male	Female
46%	54%

% Non-Defaulter	
Male	Female
37%	63%



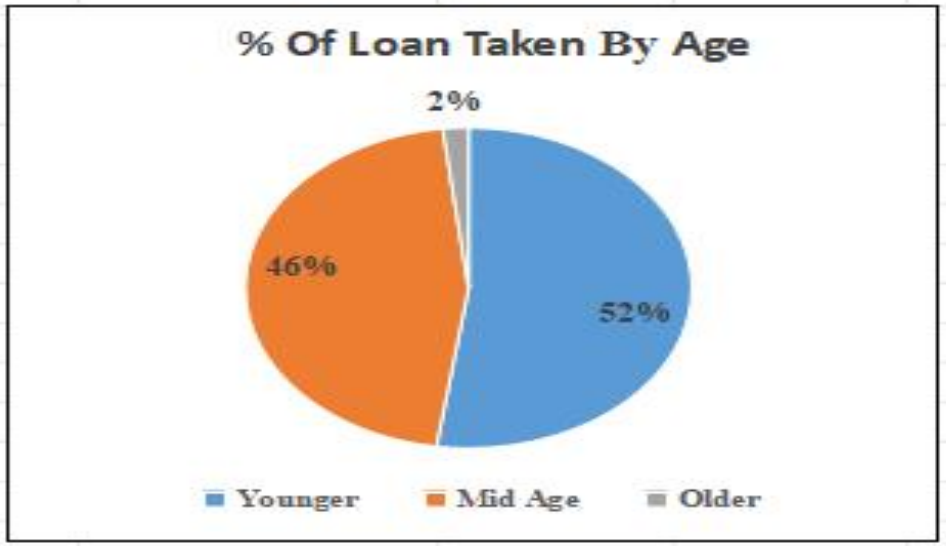


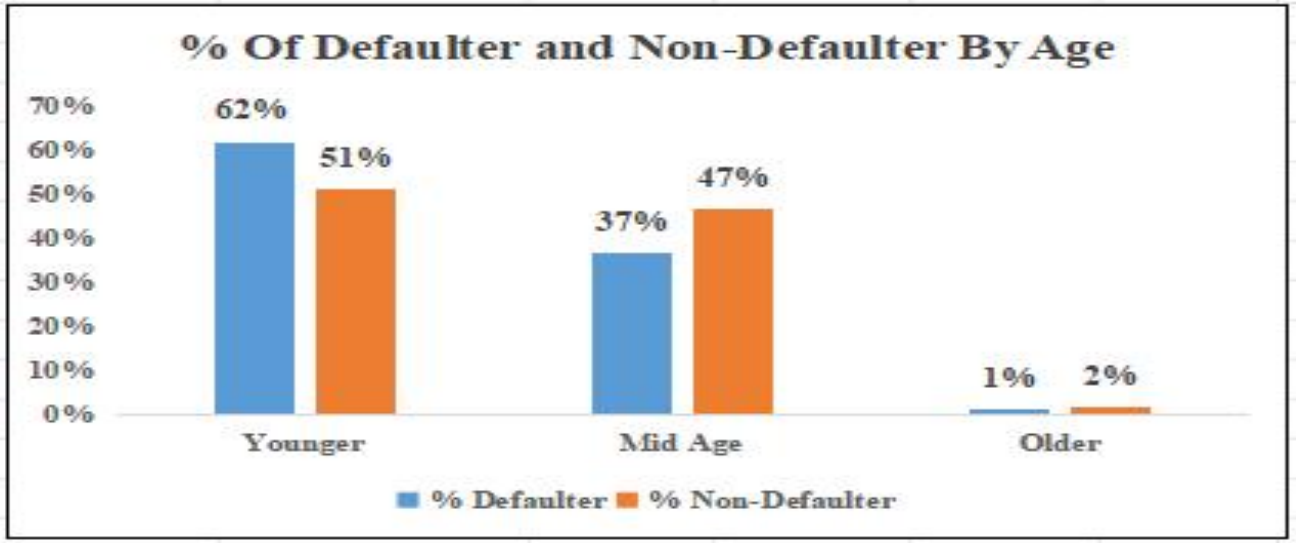
2) Age Analysis:- *The macimum number of loan is taken by Younger people, followed by Mid Age then Older.
*Also the % of Defaulter is large in Younger people.
*There is very mininun % of lone taken Old people.

Age Group	Younger	Mid Age	Older
Age Range(Years)	20-40	41-60	>61
Total	20616	18074	725
Defaulters	2114	1254	40
Non-Defaulters	18502	16820	688

% Loan Taken By Age		
Younger	Mid Age	Older
52%	46%	2%

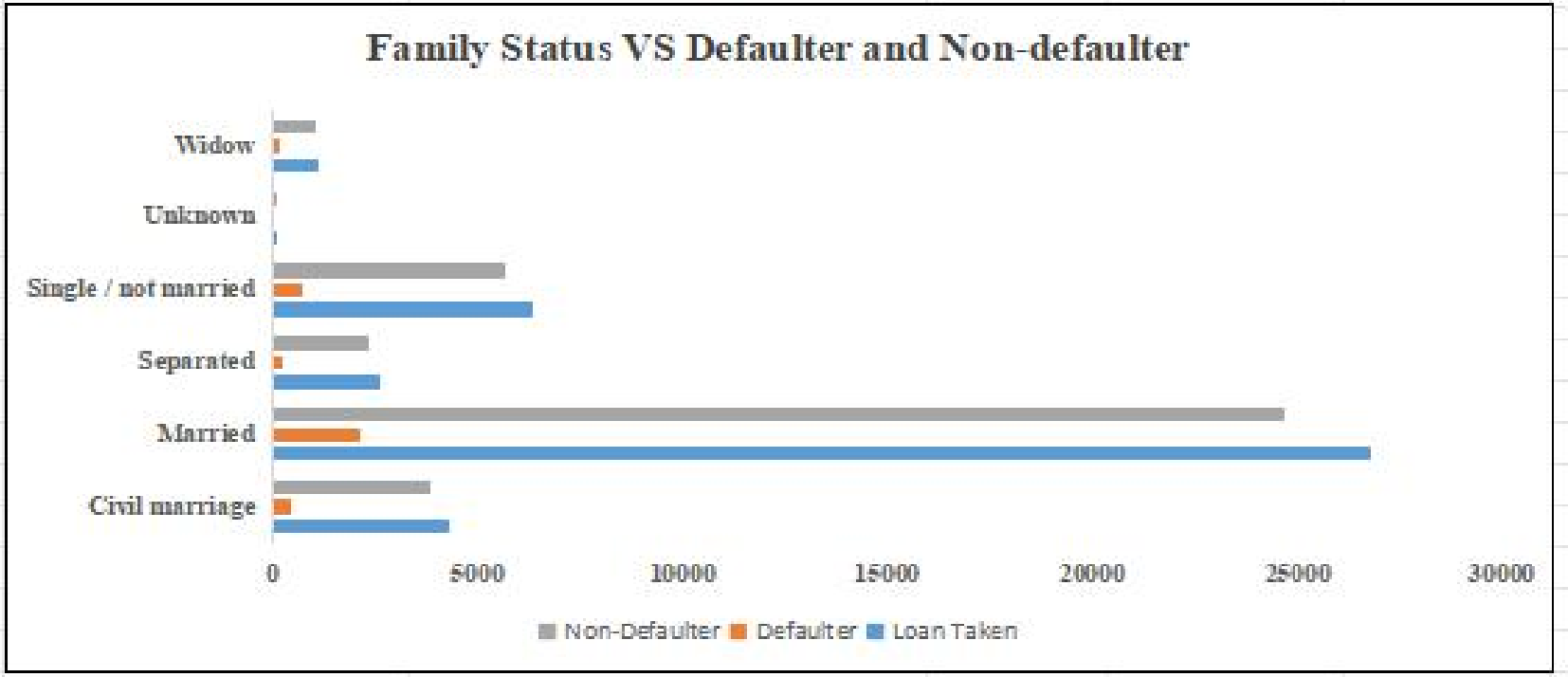
	Younger	Mid Age	Older
% Defaulter	62%	37%	1%
% Non-Defaulter	51%	47%	2%





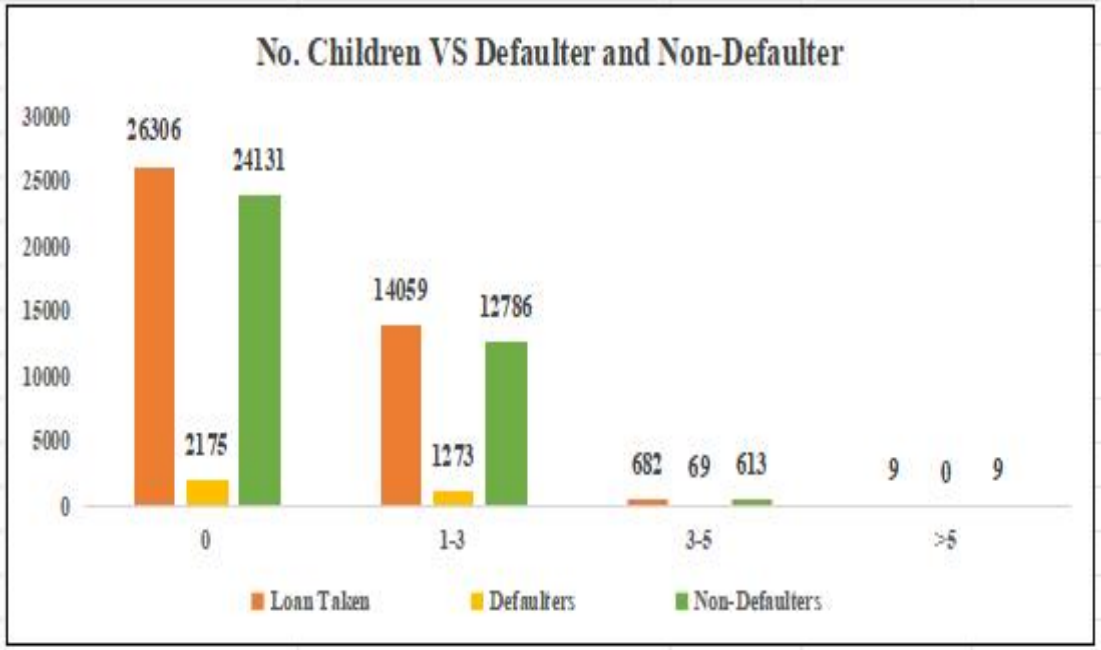
3) Family Status Analysis:- *The % Defaulter is more in the Married and Seperated peoples.
*The widows have very less % of Defaulter.

Family Status	Loan Taken	Defaulter	Non-Defaulter	% Defaulter
Civil marriage	4277	453	3824	8%
Married	26758	2103	24655	11%
Separated	2578	225	2353	11%
Single / not married	6352	672	5680	9%
Unknown	1	0	1	0%
Widow	1103	67	1036	6%

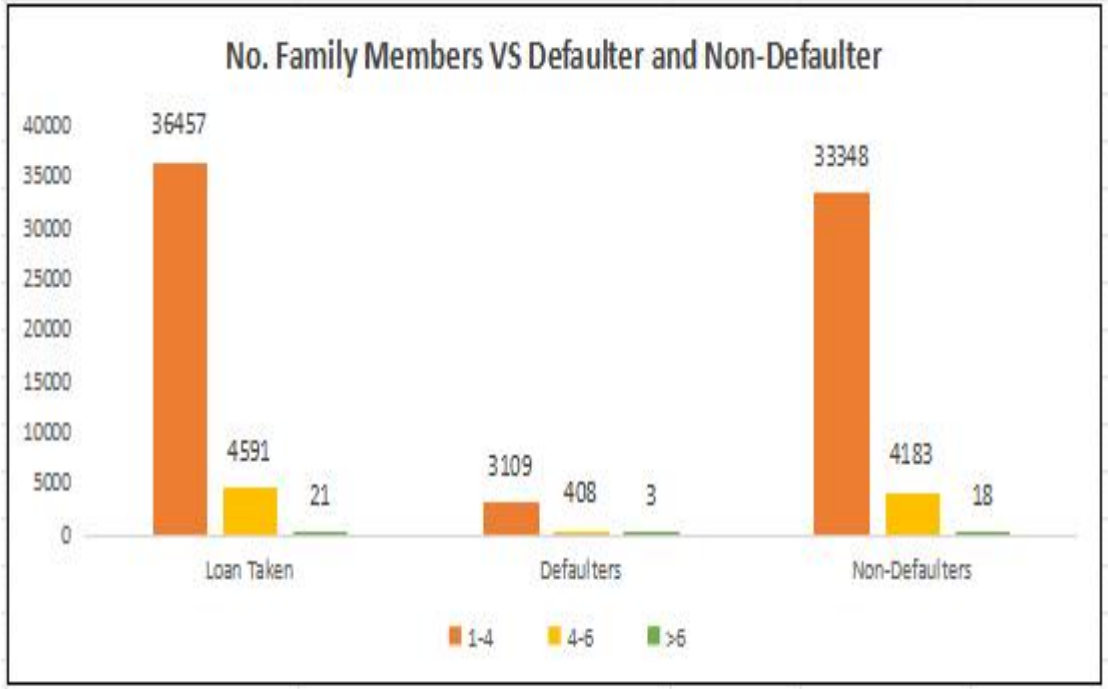


* The more number of people have taken lone who have no children and 1-3 childrens.
* The % of Defaulter is more who have 3-5 childrens.
*The number of lone taken by Small families are more also there % of Defaulter is less.

No. of Children	Loan Taken	Defaulters	Non-Defaulters	% Defaulters
0	26306	2175	24131	8%
1-3	14059	1273	12786	9%
3-5	682	69	613	10%
>5	9	0	9	0%



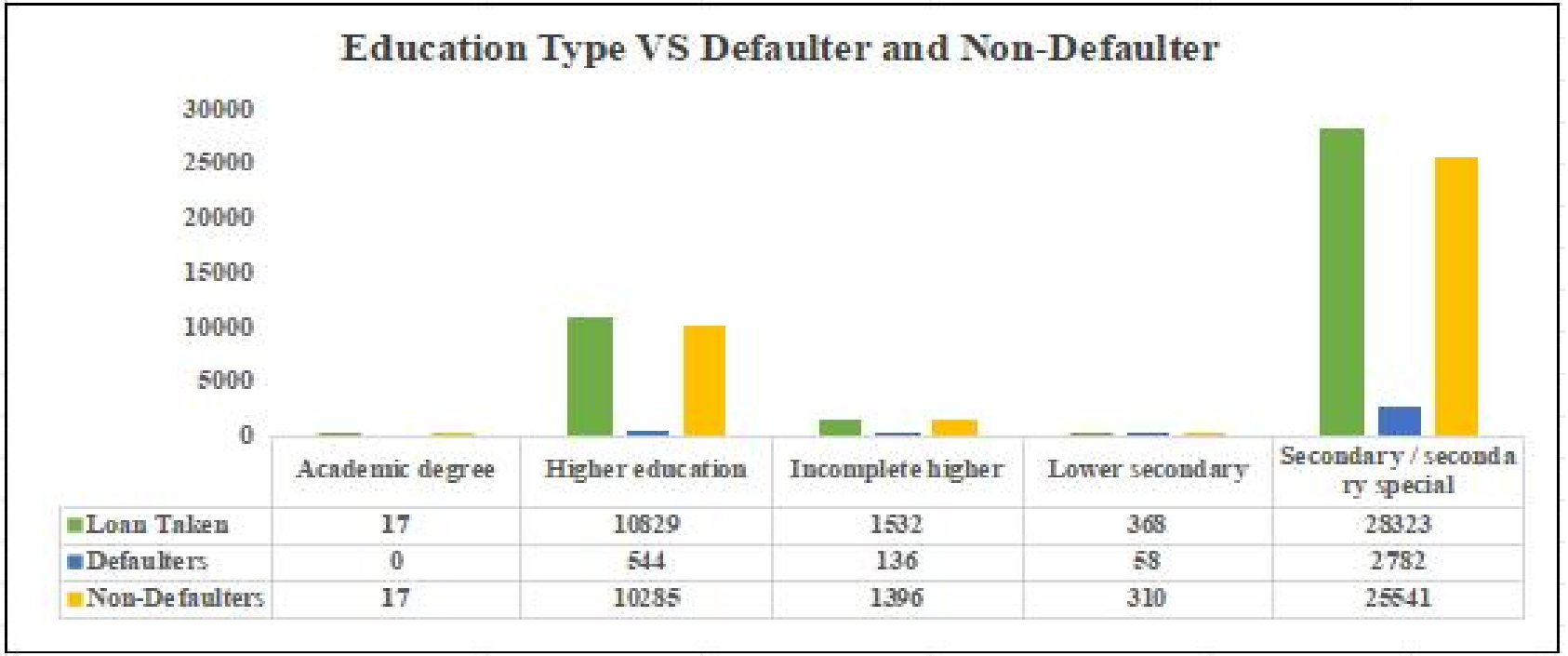
No. of Family Members	Loan Taken	Defaulters	Non-Defaulters	% Defaulters
1-4	36457	3109	33348	9%
4-6	4591	408	4183	9%
>6	21	3	18	14%



4) Education Analysis:- * There are large number of people taken lone whoes education is Secondary/Secondary Special education.

*The people with Lower Secondary education have more % of Defaulters.

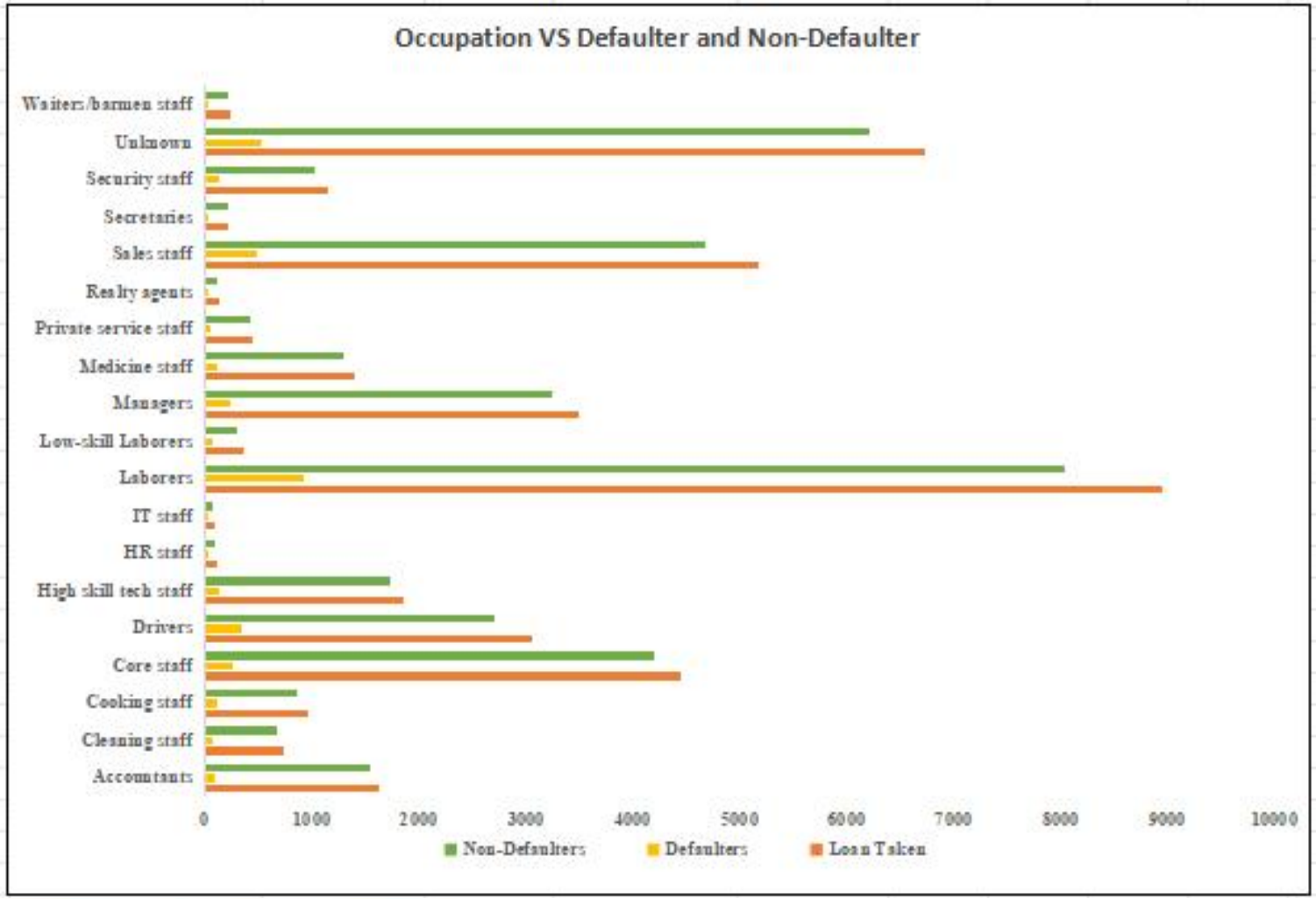
Education Type	Loan Taken	Defaulters	Non-Defaulters	% Defaulters
Academic degree	17	0	17	0%
Higher education	10829	544	10285	5%
Incomplete higher	1532	136	1396	9%
Lower secondary	368	58	310	16%
Secondary / secondary special	28323	2782	25541	10%



5) Occupation Analysis:-

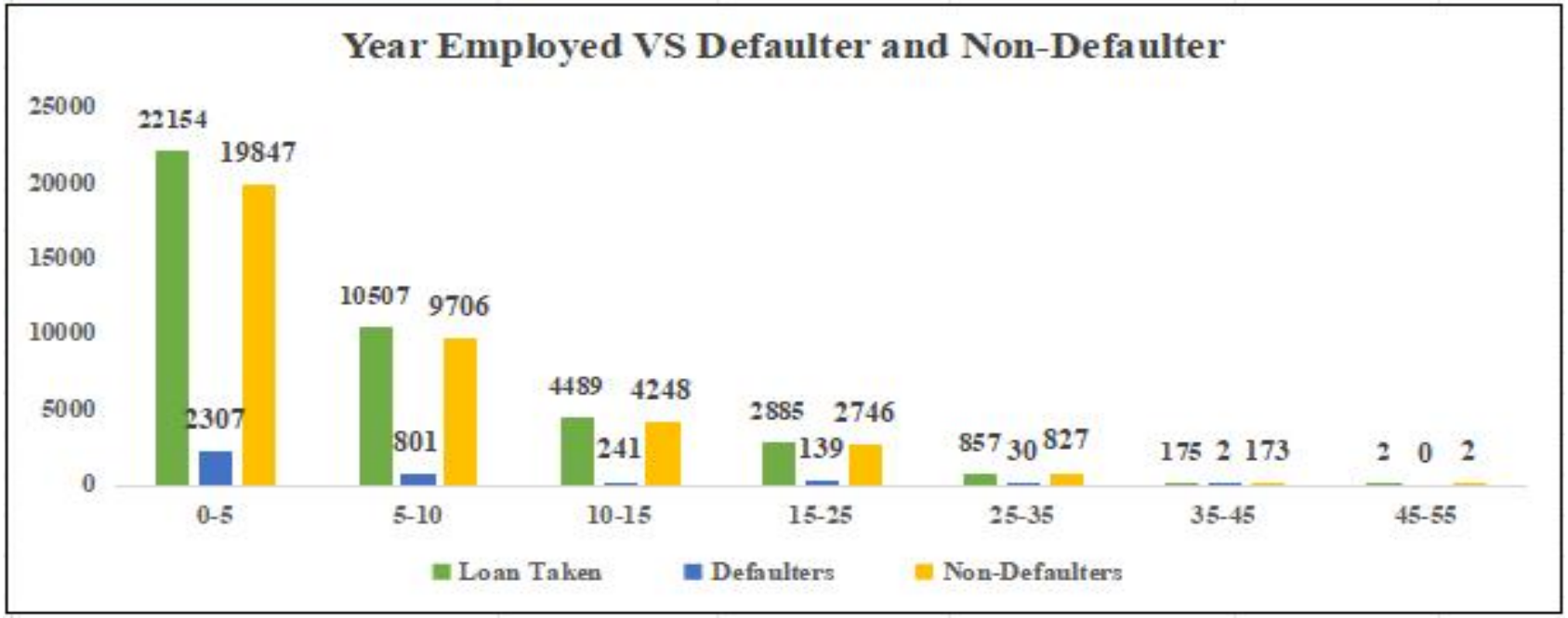
*The maximum number of customers are Laborers.
*The % of Defaulters are maximun in Low-skilled Laborers.

Occupation Type	Loan Taken	Defaulters	Non-Defaulters	% Defaulters
Accountants	1621	81	1540	5%
Cleaning staff	739	68	671	9%
Cooking staff	963	101	862	10%
Core staff	4434	250	4184	6%
Drivers	3044	338	2706	11%
High skill tech staff	1852	118	1734	6%
HR staff	101	9	92	9%
IT staff	80	4	76	5%
Laborers	8951	919	8032	10%
Low-skill Laborers	357	61	296	17%
Managers	3485	242	3243	7%
Medicine staff	1403	106	1297	8%
Private service staff	447	37	410	8%
Realty agents	123	13	110	11%
Sales staff	5159	491	4668	10%
Secretaries	212	9	203	4%
Security staff	1140	125	1015	11%
Unknown	6730	523	6207	8%
Waiters/barmen staff	228	25	203	11%



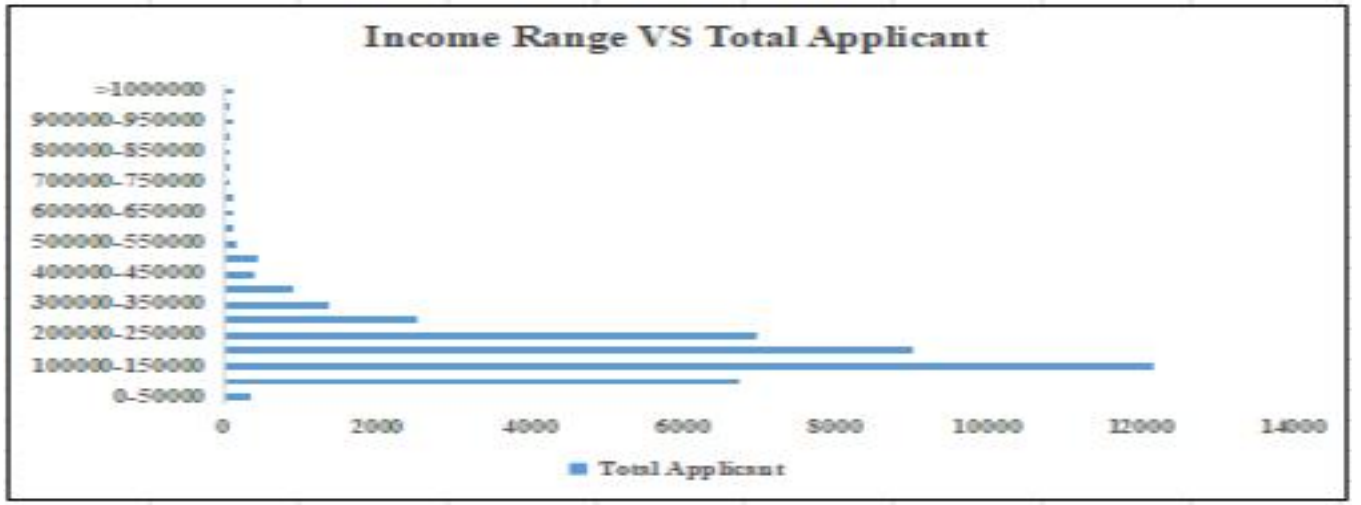
* There are more number of people with 5 or less then 5 years of experiance and they have maximun % of Defaulters,

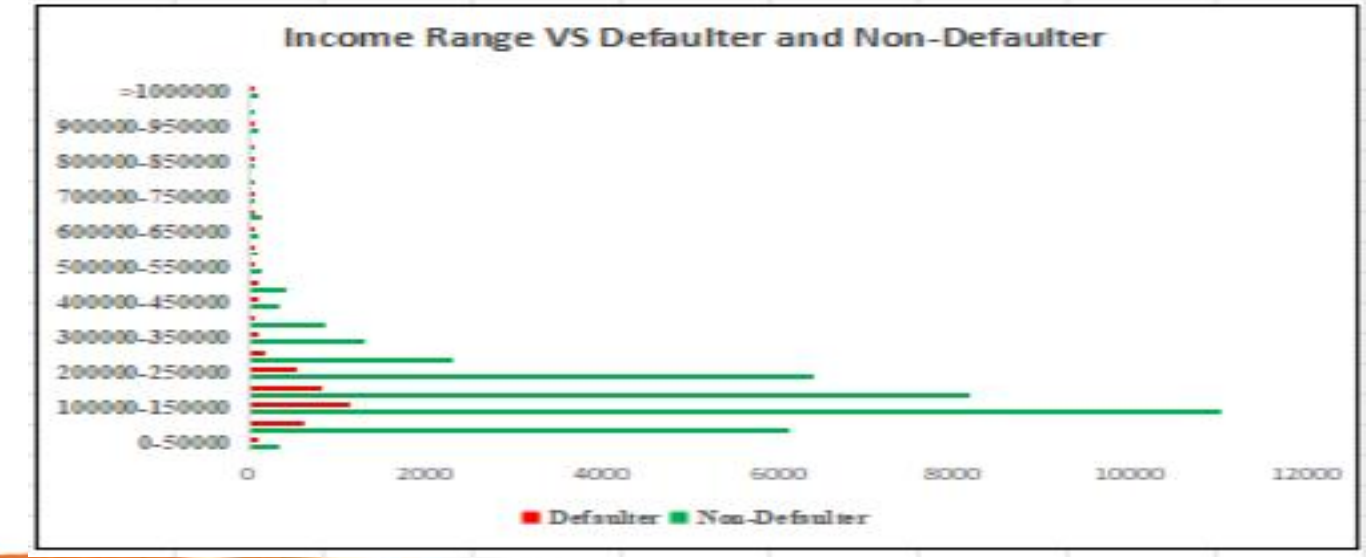
Year Employed	Loan Taken	Defaulters	Non-Defaulters	% Defaulters
0-5	22154	2307	19847	10%
5-10	10507	801	9706	8%
10-15	4489	241	4248	5%
15-25	2885	139	2746	5%
25-35	857	30	827	4%
35-45	175	2	173	1%
45-55	2	0	2	0%



❖ Segmented Univariate Analysis:- 1) Income Range:- *The most number of customer have Income between 50000-250000. also the number of Defalters are more too.

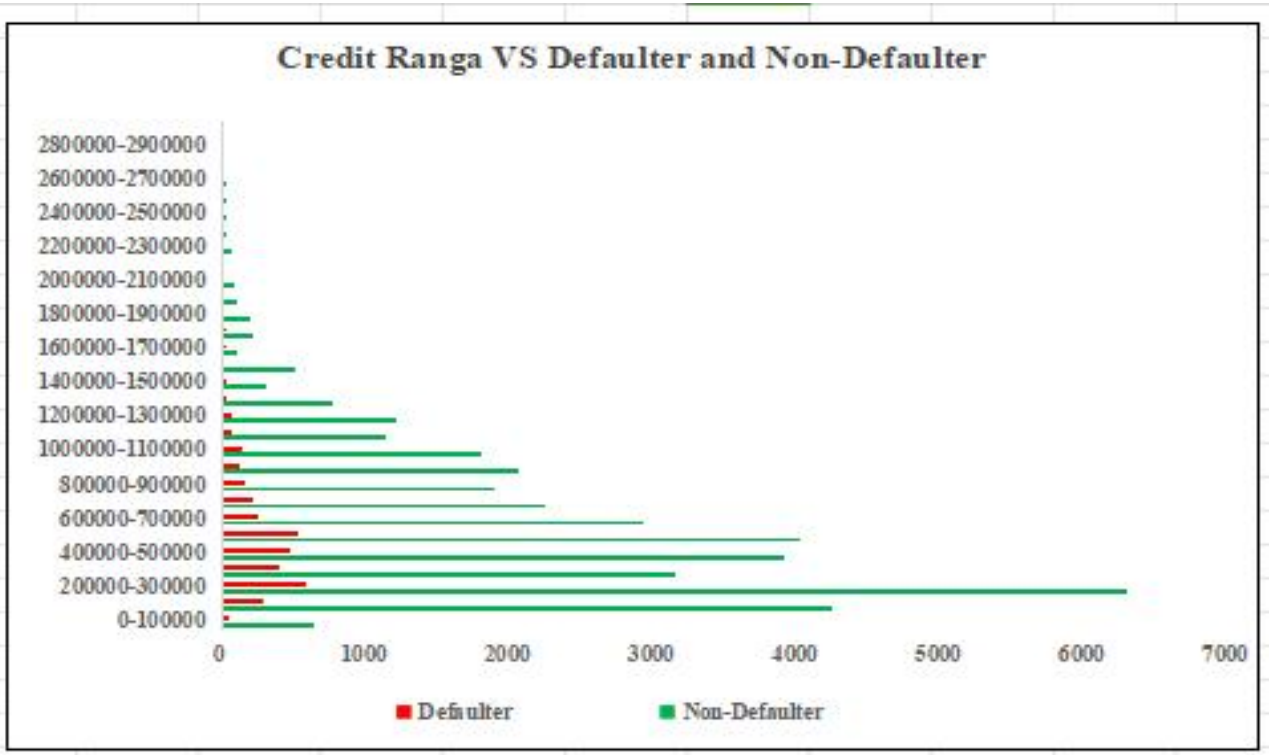
Income Range	Total Applicant	Non-Defaulter	Defaulter
0-50000	347	310	37
50000-100000	6726	6109	617
100000-150000	12145	11006	1139
150000-200000	8973	8162	811
200000-250000	6950	6414	536
250000-300000	2484	2313	171
300000-350000	1349	1273	76
350000-400000	876	831	45
400000-450000	361	335	26
450000-500000	424	391	33
500000-550000	121	112	9
550000-600000	42	37	5
600000-650000	39	38	1
650000-700000	111	104	7
700000-750000	20	19	1
750000-800000	9	9	0
800000-850000	21	19	2
850000-900000	4	4	0
900000-950000	29	27	2
950000-1000000	1	1	0
>1000000	37	35	2





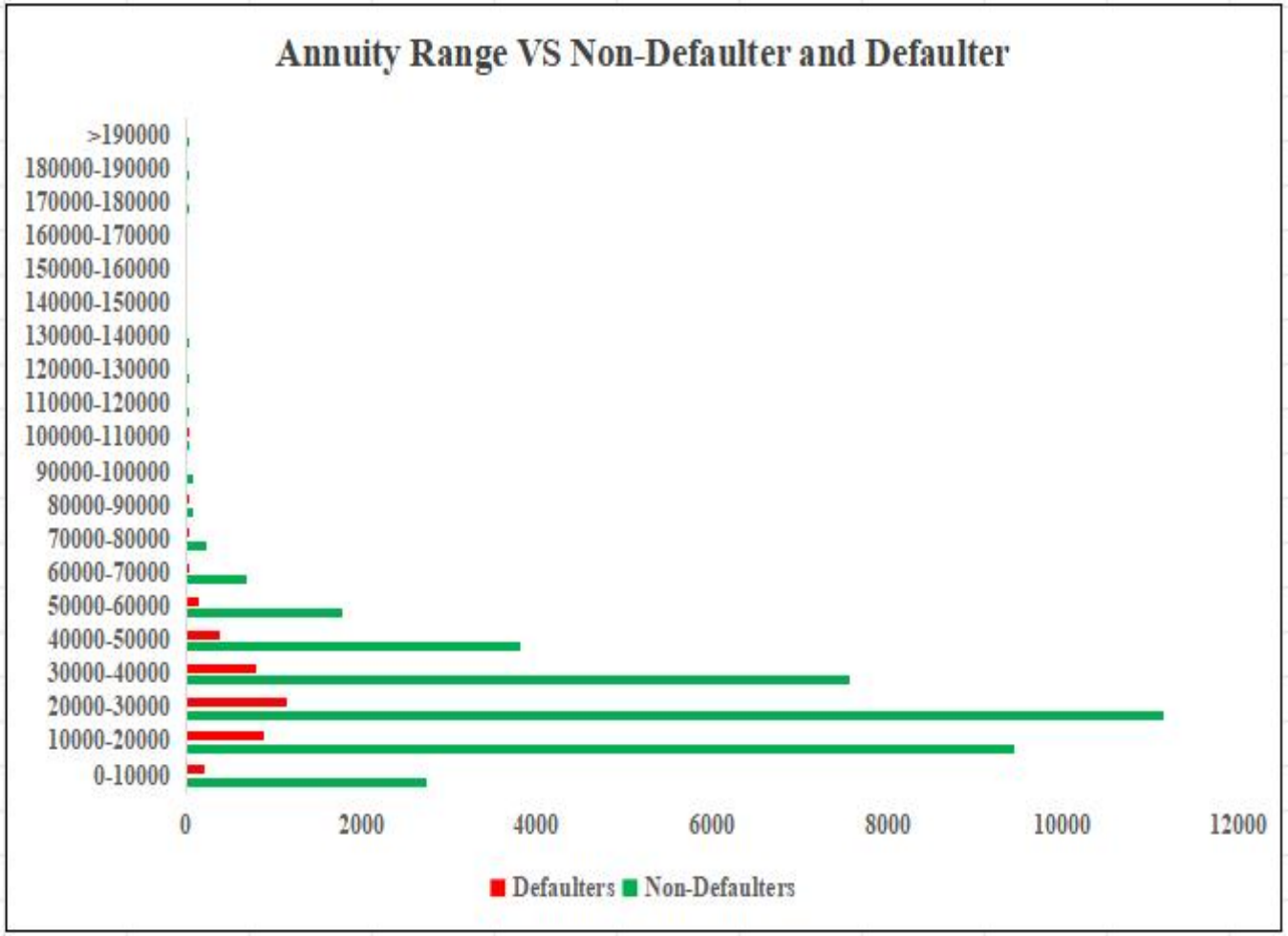
2) Credit Range:- * Most of the customers have Credit range between 200000-300000 also they have less number of Defaulters.

Credit range	Total Applicant	Non-Defaulter	Defaulter
0-100000	695	648	47
100000-200000	3903	4260	291
200000-300000	6895	6302	593
300000-400000	3555	3158	397
400000-500000	4408	3927	480
500000-600000	4560	4033	527
600000-700000	3209	2943	266
700000-800000	2473	2260	213
800000-900000	2079	1912	167
900000-1000000	2209	2074	135
1000000-1100000	1954	1807	147
1100000-1200000	1219	1143	76
1200000-1300000	1279	1214	65
1300000-1400000	822	779	43
1400000-1500000	320	306	14
1500000-1600000	537	517	20
1600000-1700000	125	116	9
1700000-1800000	229	220	9
1800000-1900000	215	209	6
1900000-2000000	107	102	5
2000000-2100000	96	92	4
2100000-2200000	28	26	2
2200000-2300000	79	79	0
2300000-2400000	13	12	1
2400000-2500000	14	13	1
2500000-2600000	29	29	0
2600000-2700000	11	11	0
2700000-2800000	2	2	0
2800000-2900000	0	0	0
2900000-3000000	3	2	1



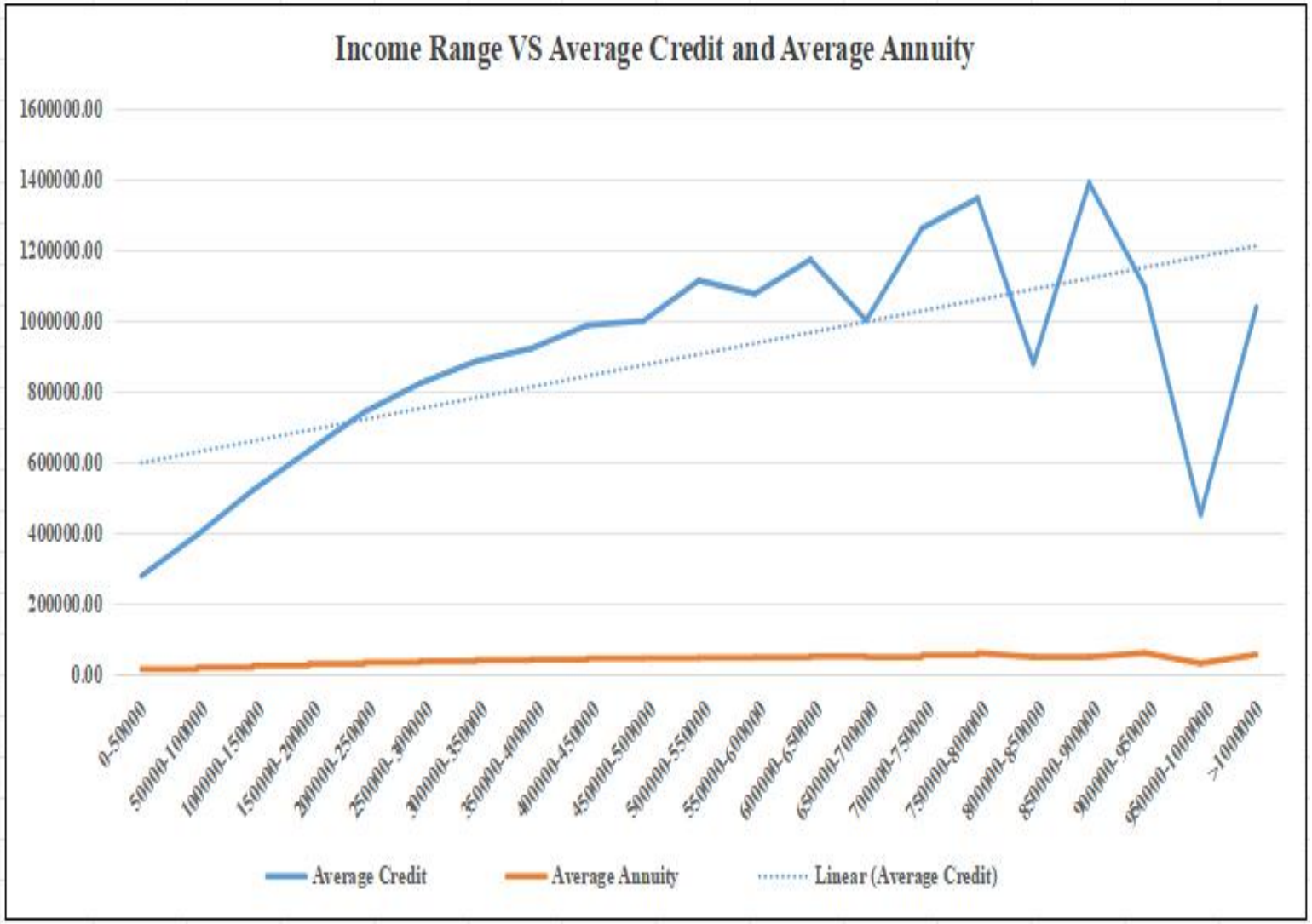
3) Annuity Range:- * The most of the peoples have Annuity range between 10000-40000 also they have approximately 10% - 15% of Defaulters.

Annuity Range	Total Applicant	Non-Defaulters	Defaulters
0-10000	2939	2730	209
10000-20000	10325	9436	889
20000-30000	12294	11139	1155
30000-40000	8348	7559	789
40000-50000	4088	3797	391
50000-60000	1908	1769	139
60000-70000	722	688	34
70000-80000	230	222	8
80000-90000	78	74	4
90000-100000	67	67	0
100000-110000	22	21	1
110000-120000	20	20	0
120000-130000	7	7	0
130000-140000	8	8	0
140000-150000	0	0	0
150000-160000	0	0	0
160000-170000	0	0	0
170000-180000	4	4	0
180000-190000	1	1	0
>190000	6	6	0



❖ Bivariate Analysis:- *The Trendline shows that, if the Income Range is increased then the Average Credit also increaseed .The average annuity is also gradually increasing.

Income Range	Average Credit	Average Annuity
0-50000	277298.05	13640.21
50000-100000	393349.85	18704.74
100000-150000	519709.47	24009.44
150000-200000	630878.77	28654.87
200000-250000	740833.61	33007.89
250000-300000	821826.37	36100.31
300000-350000	884090.21	39301.63
350000-400000	920791.10	40810.46
400000-450000	985704.88	44097.51
450000-500000	997944.62	44784.76
500000-550000	1112433.21	45984.46
550000-600000	1074844.07	46788.96
600000-650000	1171325.88	49857.69
650000-700000	1000031.84	47379.41
700000-750000	1259983.35	53482.28
750000-800000	1344940.00	58803.00
800000-850000	876760.07	47799.86
850000-900000	1388400.75	47631.38
900000-950000	1093675.66	58976.22
950000-1000000	450000.00	30073.50
>1000000	1037054.80	54840.04



E) Identify Top Correlations for Differenr Scenarios:- Understanding the correlation between variables and the target variable can provide insights into stromg indicators ofloan default.

Task:- Segment the dataset based on different Scenarious. and identify the top correlations for each segmented data using Excel Functions.

Top 5 Correlation (Non-Defaulters)		
Variable 1	Variable 2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998
AMT_GOODS_PRICE	AMT_CREDIT	0.986
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.948
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.853

Top 5 Correlation (Defaulters)		
Variable 1	Variable 2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998
AMT_GOODS_PRICE	AMT_CREDIT	0.982
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.951
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.891
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.806

- **MICROSOFT EXCEL FILE:-** https://docs.google.com/spreadsheets/d/1tgmGcpx9cmGHbVpD6r4QAubecG-PgD_q/edit?usp=drive_link&ouid=103974361659264463652&rtpof=true&sd=true
- **VIDEO LINK:-**https://www.loom.com/share/71722adfb0424bb59daa57fcc3e03340?sid=0dbcd61a-5f08-4723-b32b-2e9e41d52d1a

END