Bank Loan Case Study

Project Description

This project is useful for finance companies that give various types of loan to customers. Company faces two types of customers, first of all the customer who had a late payment of more than X days on at least one of the first Y installments of the loan means these are defaulters and second customers who had made payment on time means these are non defaulters.

By default If the company approved the defaulter's loan application, it faces both business and finance loss. So the main purpose of this project is to use Exploratory Data Analysis(EDA) to analyze patterns in the data and to ensure that capable applicants are not rejected.

With the help of this analysis the company can better understand the key factors behind loan default so it can make better decisions about loan approval.

There are five types of data analytics tasks :

1. Identify missing data and deal with it appropriately
2. Identify outliers in the dataset
3. Analyze data imbalance
4. Perform univariate, univariate segmented and bivariate analysis
5. Identify top correlation between different variables.

Approach

For this project, I apply the following approach:-

1. First of all i download the dataset which is available on the trainity platform. It contains 3 types of file attachment.

previous_application.csv: Contains information about previous loan applications.

application_data.csv: Provides details about the current loan applications.

columns_description.csv: Describes the columns present in the other datasets, explaining what each column represents.

2. After analyzing the dataset I started working on this project by the use of excel. The first and second task is related to data cleaning activity. It contains identifying missing data and outliers and handling it appropriately. After identifying the columns that have the missing data more than 30% , I drop all these columns and separate the dataset after task 1.

3. After that i perform other analysis on the dataset that are mentioned in the project details like data imbalance, univariate, bivariate , segmented univariate analysis and so on with the help of excel's and statistics functions and formulas(mean,countifs,unique,percentage,absolute,table,pivot table),visualization tools and all other features that are available on excel.

Tech stack used

Microsoft Excel : Since the project is totally based on excels and statistics, therefore I use ms excel. I use excel's different tools , different tabs, editing tools, formulas like mean,countifs,unique,percentage,absolute, sort and filter option , table design options, pivot table, visualization tools eg. column and bar chart, pie chart, histogram, scatter plot and many more. They are easy to access and use.

Insights

The insights and knowledge gained during the project:

This project help me to expand my statistics knowledge for eg. exploratory data analysis like Descriptive statistics, different formulas and functions like mean, median,countifs, absolute,unique, max and min functions etc. data cleaning activity helps me to preprocess the data to make it valuable for data analysis

Also it expands my excel skills like how to use tables, pivot tables, sort and filtering options, conditional formatting, data visualization tools help me how to convert the data into column and bar chart, scatter plot , histogram etc and all other tools and techniques like to design tables , editing options and use of different tabs and many more that are available on ms excel.

In this project the following observations and meaningful trends are covered:-

A. Identify Missing Data and Deal with it Appropriately:

- **Task:** Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

- **Primary File**: **application_data.csv**
- **Reference File**: **columns_description.csv**

By using the COUNTBLANK function, I calculate the blank values for each column.

Result:-

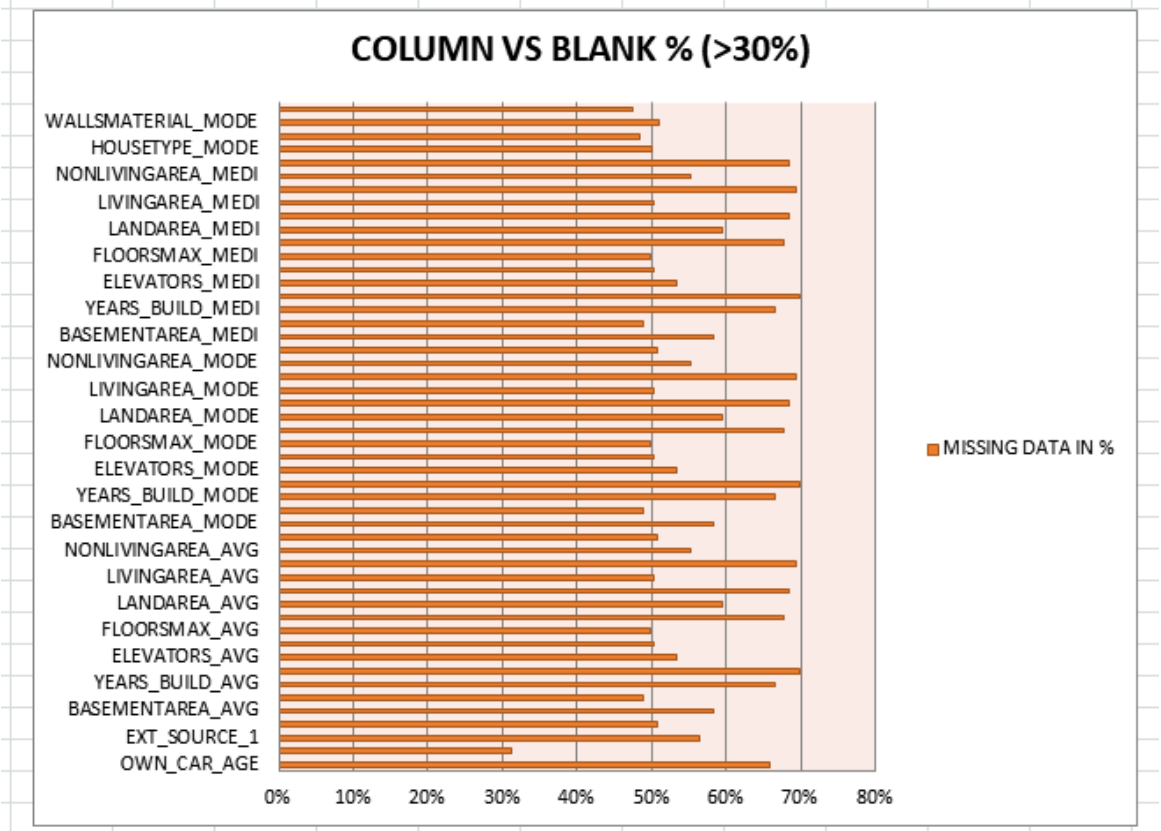
E1 =COUNTBLANK(E4:E50002)

	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY
1	0	28172	126	9944	25385	29199	24394	33239	34960	266
2	0%	56%	0%	20%	51%	58%	49%	66%	70%	5
3	ORGANIZATION_TYPE	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3	APARTMENTS_AVG	BASEMENTAREA_AVG	YEARS_BEGINEXPLUATATION_AVG	YEARS_BUILD_AVG	COMMONAREA_AVG	ELEVATORS_AV
4	Business Entity Type 3	0.083036967	0.262948593	0.13937578	0.0247	0.0369	0.9722	0.6192	0.0143	
5	School	0.311267311	0.622245775		0.0959	0.0529	0.9851	0.796	0.0605	0
6	Government		0.555912083	0.729566691						
7	Business Entity Type 3		0.65044169							
8	Religion		0.322738287							
9	Other		0.354224732	0.621226338						
10	Business Entity Type 3	0.774761413	0.723999852	0.492060094						
11	Other		0.714279286	0.54065445						
12	XNA	0.587334047	0.205747288	0.751723715						
13	Electricity		0.746643629							
14	Medicine	0.319760172	0.651862333	0.363945239						
15	XNA	0.72204445	0.555183162	0.652896552						
16	Business Entity Type 2	0.464831117	0.715041819	0.176652579	0.0825		0.9811			
17	Self-employed		0.566906613	0.77008707	0.1474	0.0973	0.9806	0.7348	0.0582	0
18	Transport: type 2	0.721939769	0.642656205		0.3495	0.1335	0.9985	0.9796	0.1143	
19	Business Entity Type 2	0.115634337	0.346633981	0.678567689						
20	Government		0.23637784	0.062103038						
21	Construction		0.683513346							
22	Housing		0.706428403	0.556727426	0.0278	0.0617	0.9881	0.8368	0.0018	
23	Kindergarten		0.58661714	0.477649155						
24	Self-employed	0.565654882	0.113374513		0.0722	0.0801	0.9781	0.7008		
25	Trade: type 7	0.43770902	0.233766958	0.542445144						

After that I separate the columns that have missing values more than 30% and less than 30% with the help of appropriate charts.

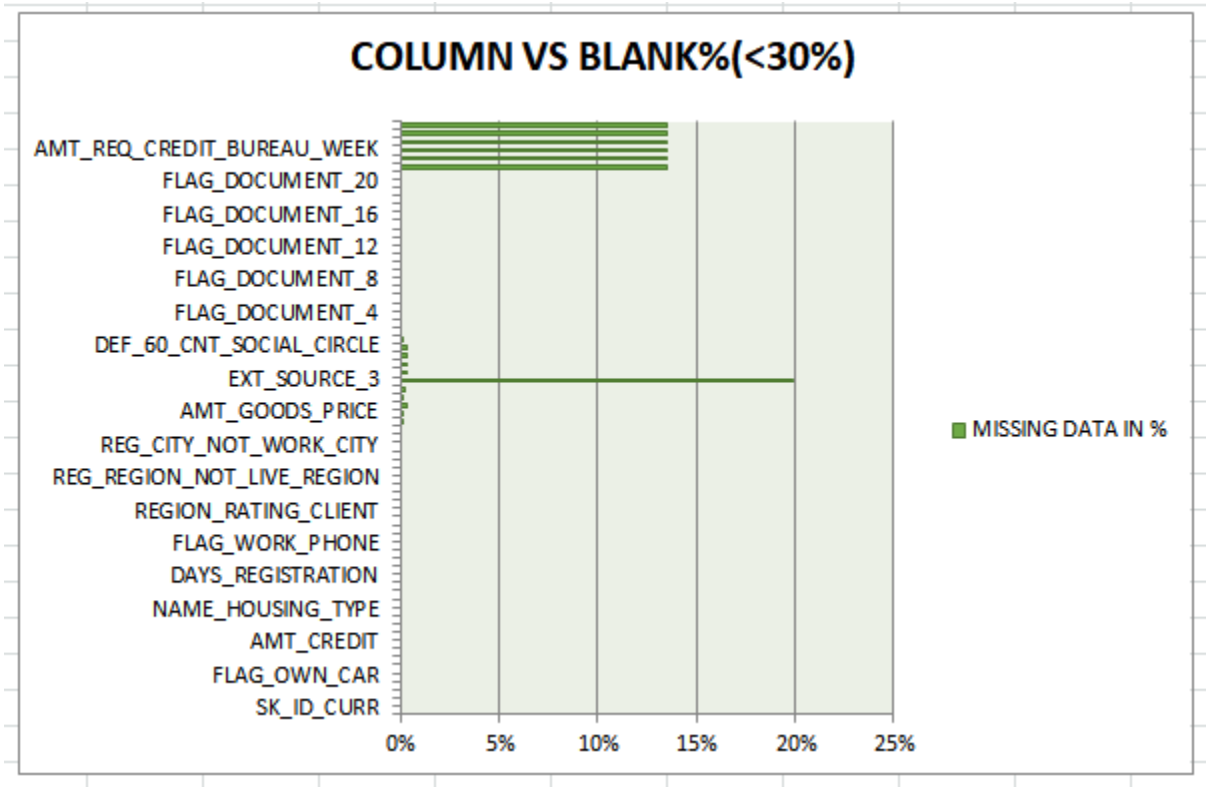
Result:-

1	Missing data more than 30%	
2		
3		
4	**COLUMN NAME**	**MISSING DATA IN %**
5	OWN_CAR_AGE	66%
6	OCCUPATION_TYPE	31%
7	EXT_SOURCE_1	56%
8	APARTMENTS_AVG	51%
9	BASEMENTAREA_AVG	58%
10	YEARS_BEGINEXPLUATATION_AVG	49%
11	YEARS_BUILD_AVG	66%
12	COMMONAREA_AVG	70%
13	ELEVATORS_AVG	53%
14	ENTRANCES_AVG	50%
15	FLOORSMAX_AVG	50%
16	FLOORSMIN_AVG	68%
17	LANDAREA_AVG	59%
18	LIVINGAPARTMENTS_AVG	68%
19	LIVINGAREA_AVG	50%
20	NONLIVINGAPARTMENTS_AVG	69%
21	NONLIVINGAREA_AVG	55%
22	APARTMENTS_MODE	51%
23	BASEMENTAREA_MODE	58%
24	YEARS_BEGINEXPLUATATION_MODE	49%
25	YEARS_BUILD_MODE	66%



Missing data less than 30%

COLUMN NAME	MISSING DATA IN %
SK_ID_CURR	0%
TARGET	0%
NAME_CONTRACT_TYPE	0%
CODE_GENDER	0%
FLAG_OWN_CAR	0%
FLAG_OWN_REALTY	0%
CNT_CHILDREN	0%
AMT_INCOME_TOTAL	0%
AMT_CREDIT	0%
NAME_INCOME_TYPE	0%
NAME_EDUCATION_TYPE	0%
NAME_FAMILY_STATUS	0%
NAME_HOUSING_TYPE	0%
REGION_POPULATION_RELATIVE	0%
DAYS_BIRTH	0%
DAYS_EMPLOYED	0%
DAYS_REGISTRATION	0%
DAYS_ID_PUBLISH	0%
FLAG_MOBIL	0%
FLAG_EMP_PHONE	0%
FLAG_WORK_PHONE	0%
FLAG_CONT_MOBILE	0%
FLAG_PHONE	0%
FLAG_EMAIL	0%



After that the numerical columns(the columns that have missing data less than 30%) missing data are treated by mean imputation, i replace the blank values with the average value of the data for a particular column and the blank values of the categorical column is filled with 'not defined'.

Then I drop all the columns that have missing values more than 30% and separate the dataset with the columns that have less than 30% missing data and treat by mean imputation.

One more thing I did during this task is to convert the negative values of some columns like age, registration days , id publish days etc into positive values and also convert the data that are in days format into year format with the help of absolute function.

For the further analysis, I use this cleaned dataset.

B. Identify Outliers in the Dataset:

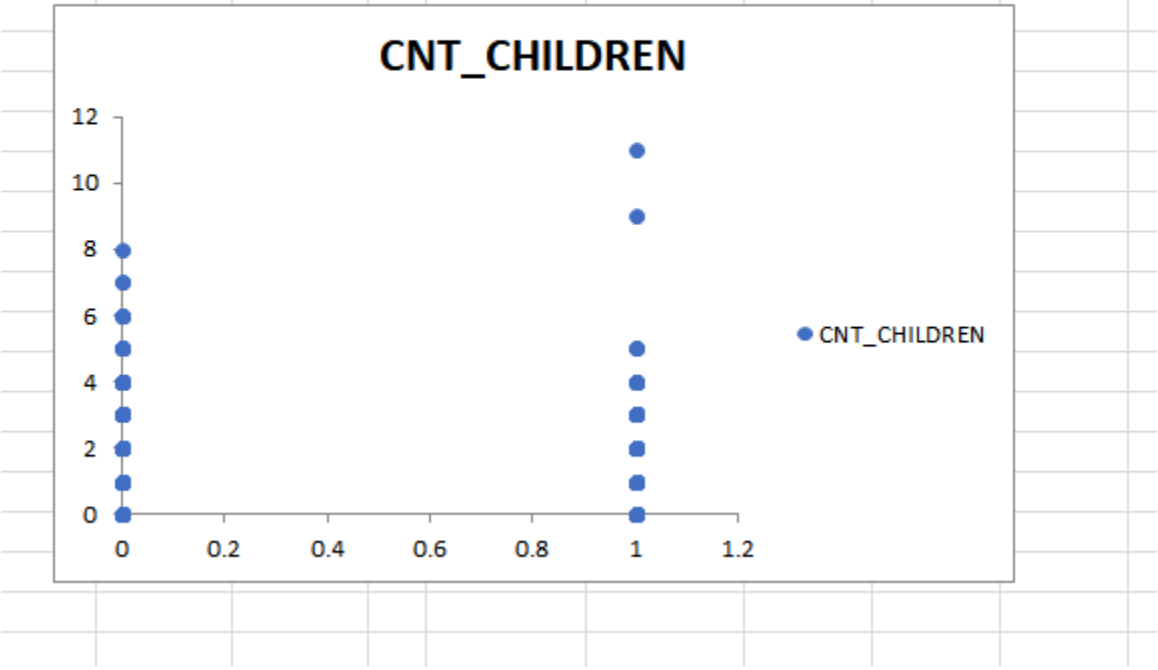
- **Task:** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

- **Primary File**: **application_data.csv**

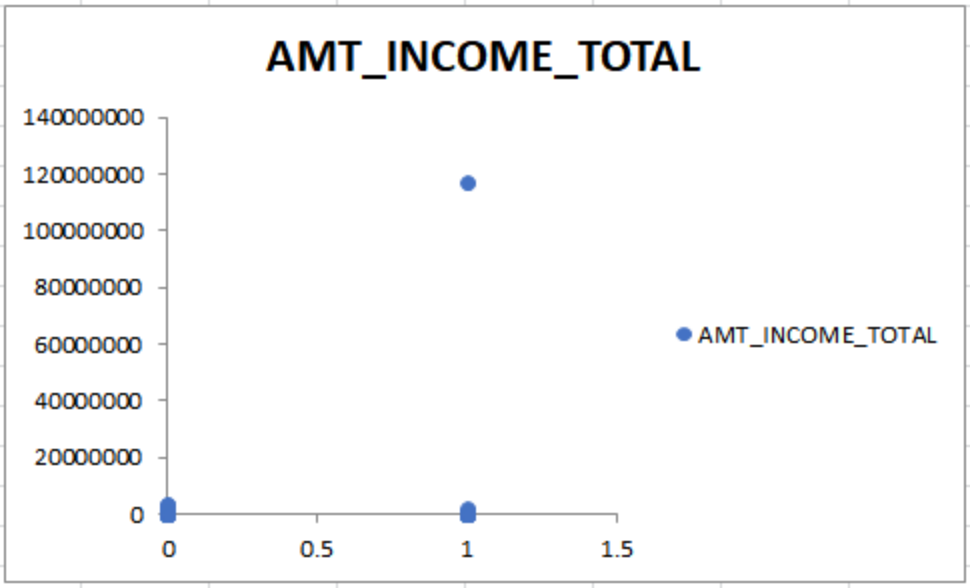
To identify the outliers from the data I convert the data into a scatter plot after calculating quartile 1,quartile 3, IQR, upper limit and lower limit. The scatterplot highlights the outliers appropriately.

Result:

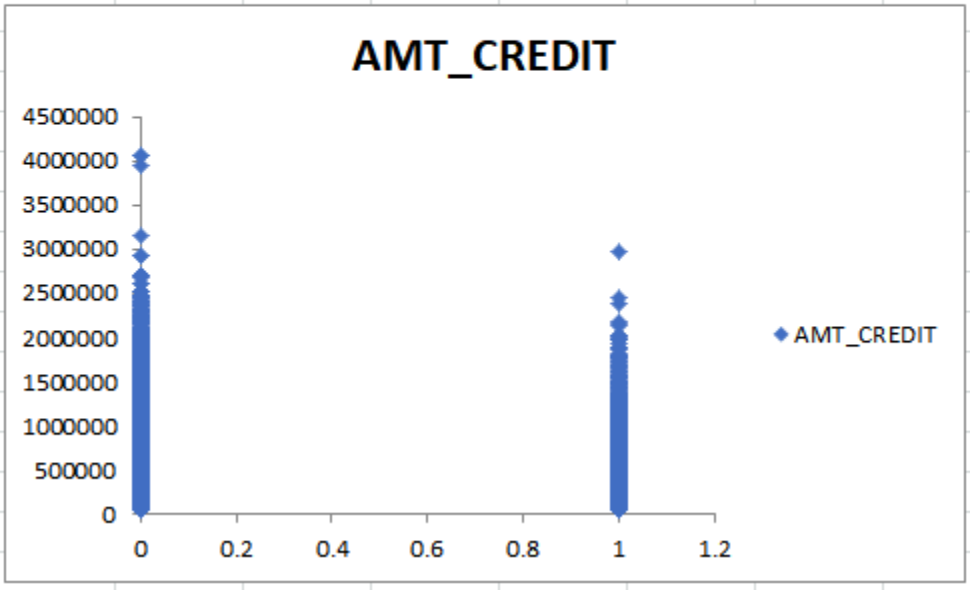
Quartile 1	Quartile 3	IQR	Upper Limit	Lower Limit
0	1	1	2.5	-1.5



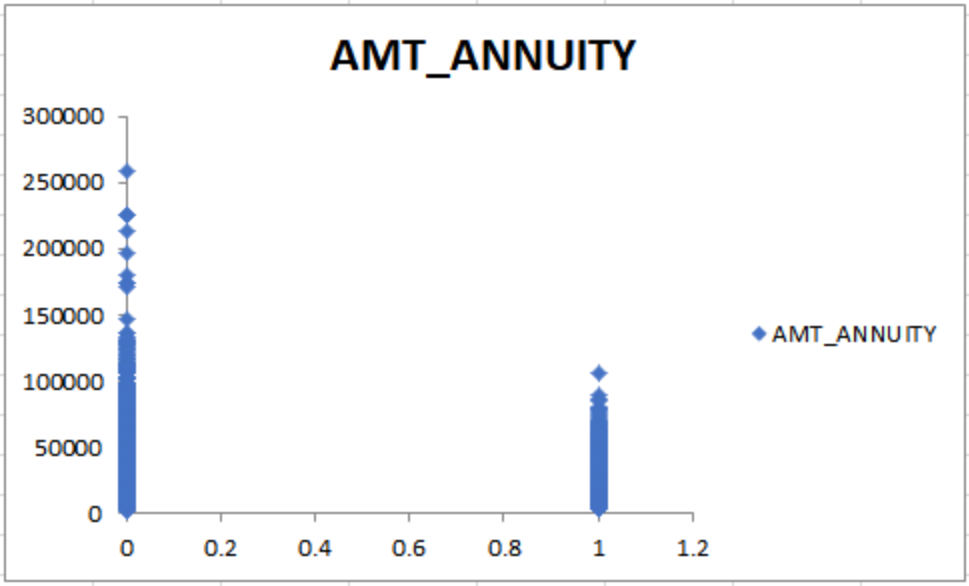
Quartile 1	Quartile 3	IQR	Upper Lim	Lower Limit
112500	202500	90000	337500	-22500



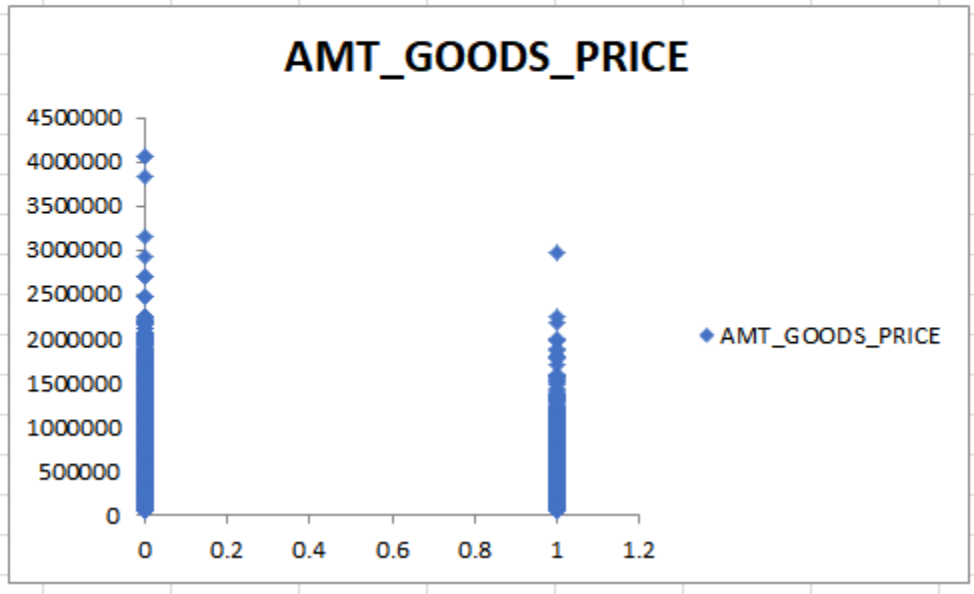
Quartile 1	Quartile 3	IQR	Upper Lim	Lower Limit
270000	808650	538650	1616625	-537975



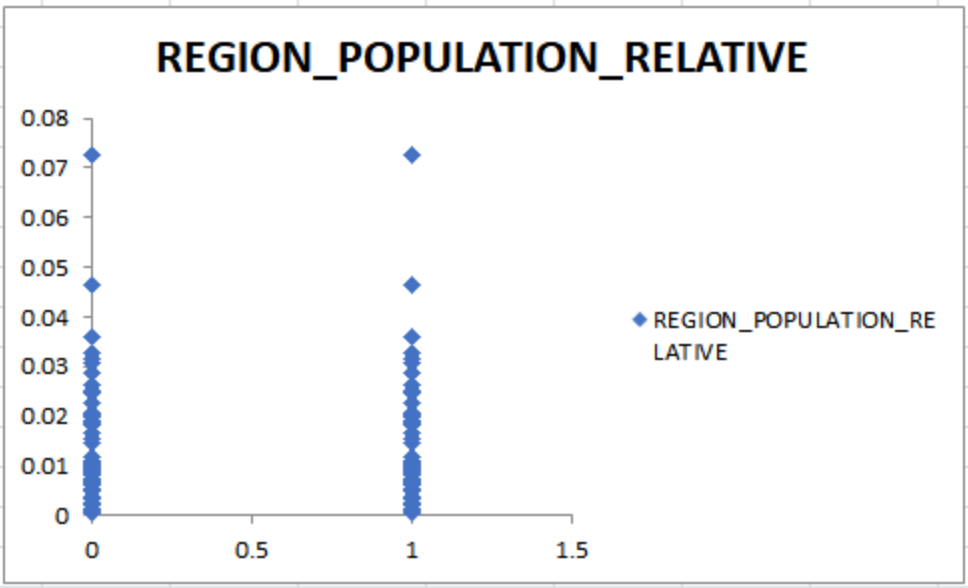
Quartile 1	Quartile 3	IQR	Upper Lim	Lower Limit
16456.5	34596	18139.5	61805.25	-10752.8



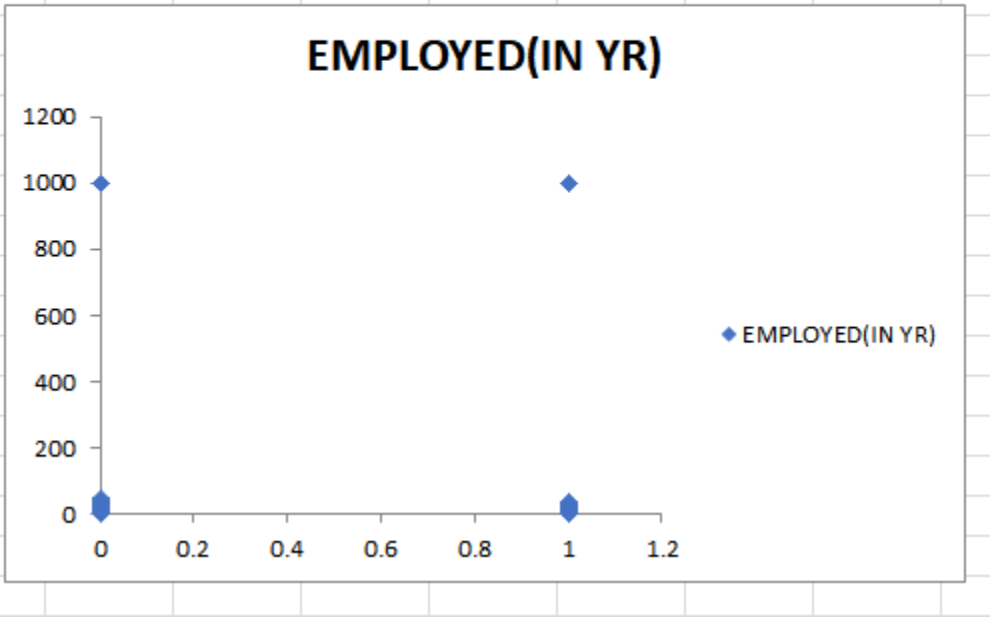
Quartile 1	Quartile 3	IQR	Upper Lim	Lower Limit
238500	679500	441000	1341000	-423000



Quartile 1	Quartile 3	IQR	Upper Lim	Lower Limit
0.010006	0.028663	0.018657	0.056649	-0.01798



Quartile 1	Quartile 3	IQR	Upper Lim	Lower Limit
2.556164	15.66575	13.10959	35.33014	-17.1082



C. Analyze Data Imbalance:

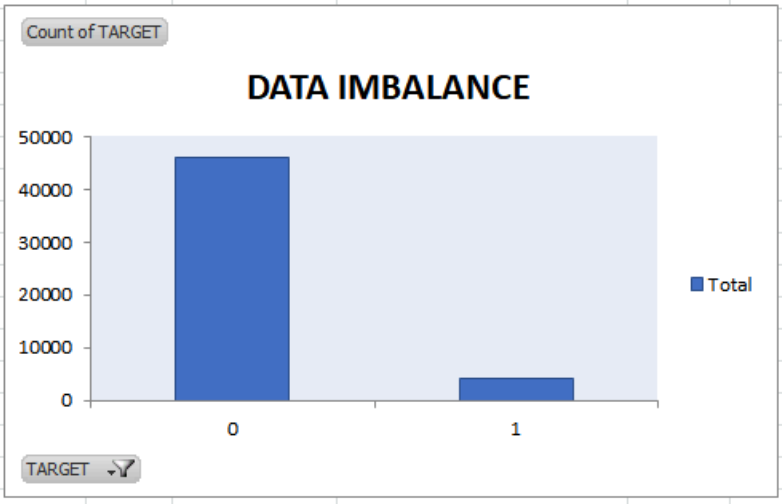
- **Task:** Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

- **Primary File**: **application_data.csv**

Result:

Row Labels	Count of TARGET
0	45973
1	4026
Grand Total	49999

TARGET	PERCENTAGE(%)	RATIO OF DATA IMBALANCE
0	92%	11.42
1	8%	



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

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

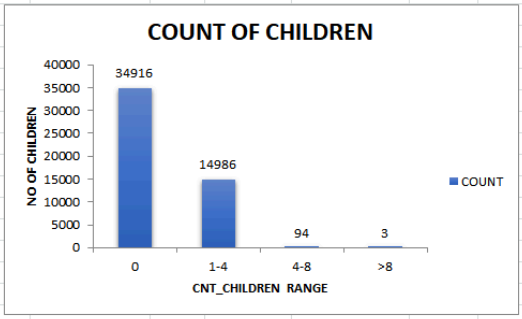
- **Primary File: application_data.csv**

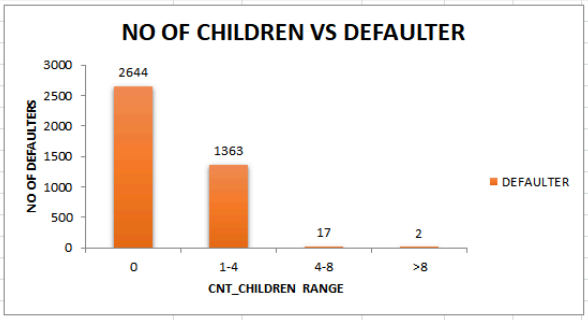
UNIVARIATE ANALYSIS

Result:

	NO OF CHILDREN	COUNT
11	0	34916
0	1-4	14986
	4-8	94
	>8	3

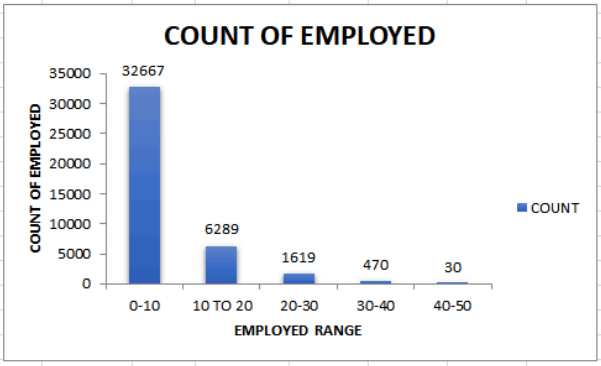
NO OF CHILDREN	DEFAULTER
0	2644
1-4	1363
4-8	17
>8	2

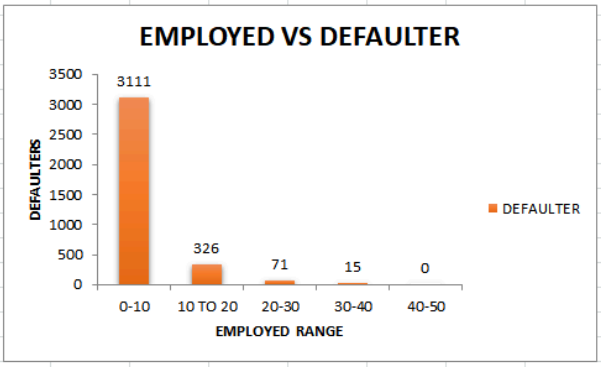




EMPLOYED	COUNT
0-10	32667
10 TO 20	6289
20-30	1619
30-40	470
40-50	30

EMPLOYE	DEFAULTER
0-10	3111
10 TO 20	326
20-30	71
30-40	15
40-50	0

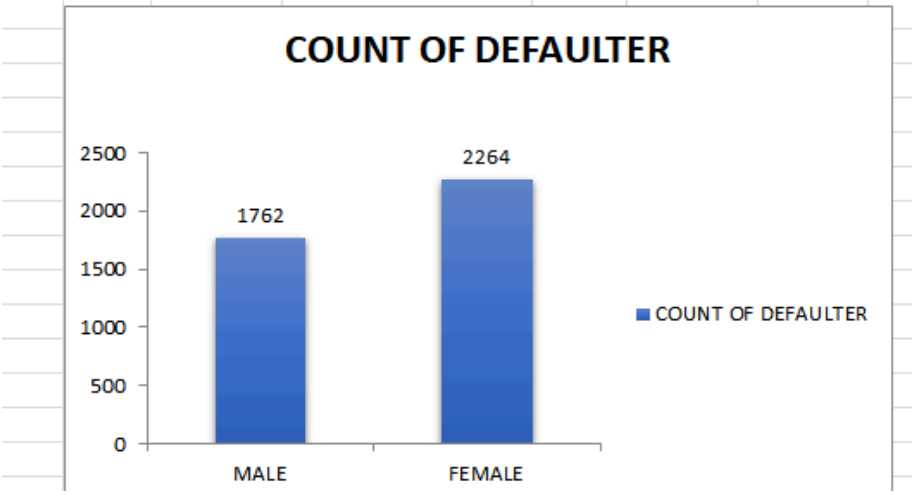




MALE	
COUNT	DEFAULTER
17174	1762

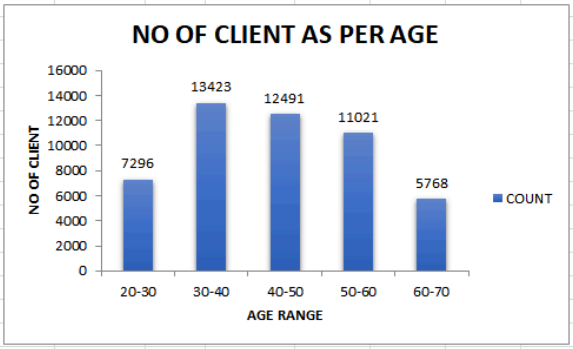
FEMALE	
COUNT	DEFAULTER
32823	2264

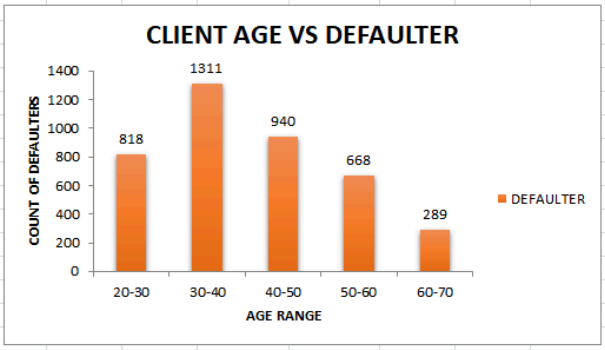
GENDER	COUNT OF DEFAULTER
MALE	1762
FEMALE	2264



	CLIENT AGE	COUNT
21.0411	20-30	7296
68.99726	30-40	13423
	40-50	12491
	50-60	11021
	60-70	5768

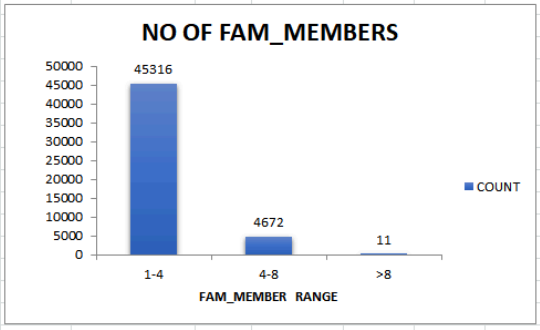
CLIENT AGE	DEFAULTER
20-30	818
30-40	1311
40-50	940
50-60	668
60-70	289

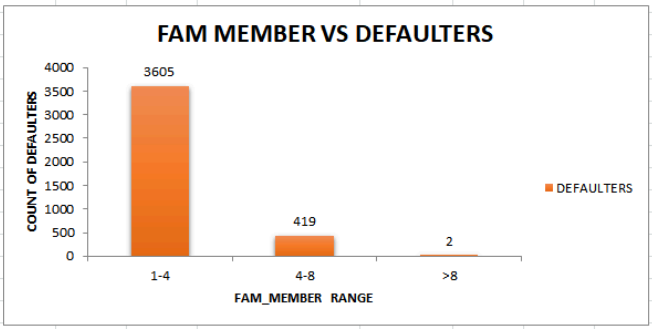




	NO OF FAM MEM	COUNT
13	1-4	45316
1	4-8	4672
	>8	11

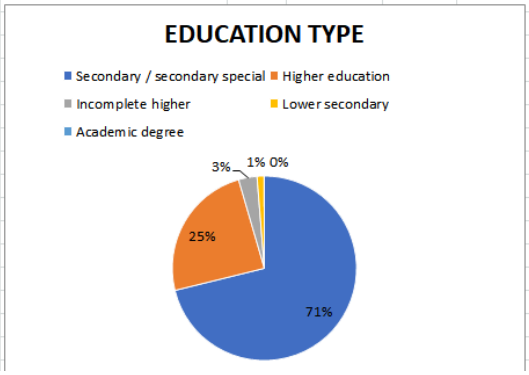
NO OF FAM MEMBERS	DEFAULTERS
1-4	3605
4-8	419
>8	2

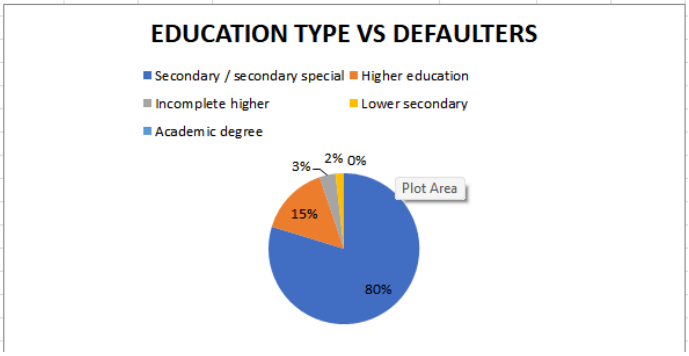




NAME_EDUCATION_TYPE	COUNT
Secondary / secondary special	35572
Higher education	12167
Incomplete higher	1620
Lower secondary	620
Academic degree	20

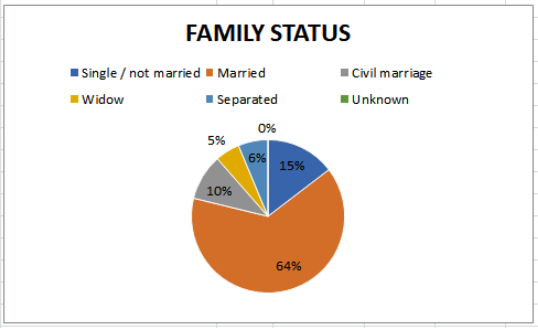
NAME_EDUCATION_TYPE	DEFAULTERS
Secondary / secondary special	3209
Higher education	606
Incomplete higher	138
Lower secondary	73
Academic degree	0

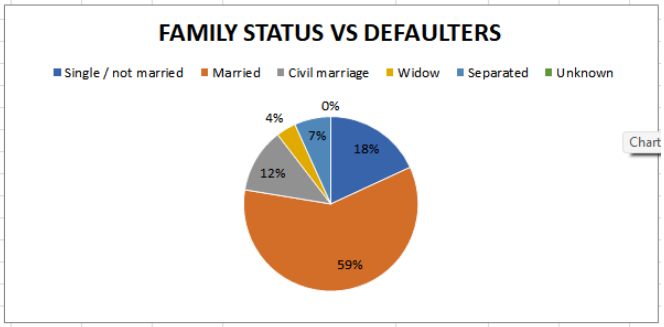




NAME_FAMILY_STATUS	COUNT
Single / not married	7306
Married	32094
Civil marriage	4859
Widow	2597
Separated	3142
Unknown	1

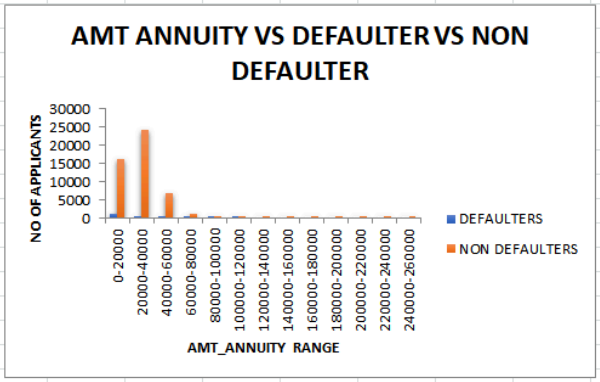
NAME_FAMILY_STATUS	DEFAULTERS
Single / not married	729
Married	2395
Civil marriage	482
Widow	148
Separated	272
Unknown	0



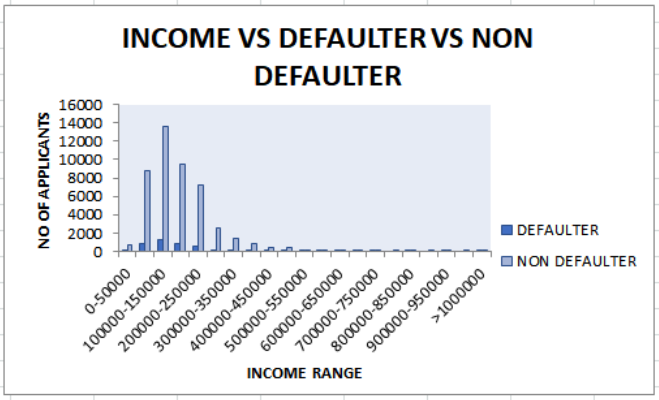


SEGMENTED UNIVARIATE ANALYSIS

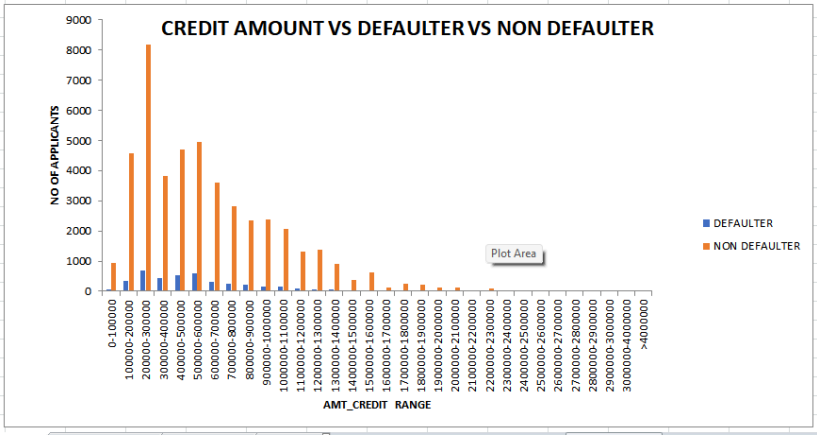
25.5	AMT_ANNUITY	APPLICANTS	DEFAULTERS	NON DEFAULTERS
2052	0-20000	17291	1297	15994
	20000-40000	24607	474	24133
	40000-60000	6792	44	6748
	60000-80000	1062	4	1058
	80000-100000	162	1	161
	100000-120000	50	1	49
	120000-140000	21	0	21
	140000-160000	1	0	1
	160000-180000	4	0	4
	180000-200000	2	0	2
	200000-220000	1	0	1
	220000-240000	5	0	5
	240000-260000	1	0	1



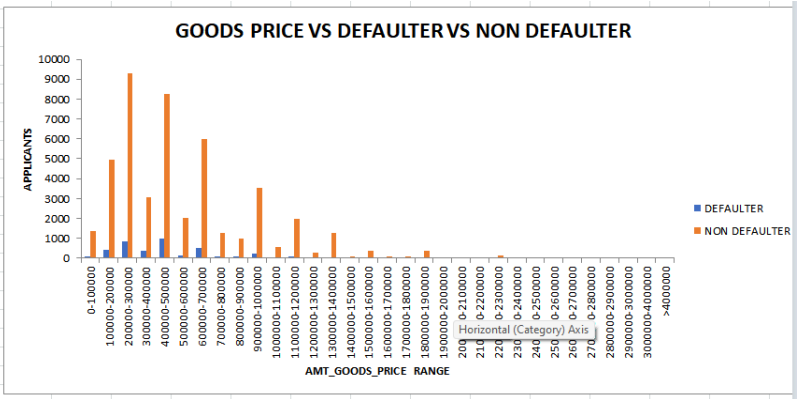
INCOME	APPLICANTS	DEFAULTER	NON DEFAULTER
0-50000	804	63	741
50000-100000	9588	782	8806
100000-150000	14852	1298	13554
150000-200000	10408	890	9518
200000-250000	7818	576	7242
250000-300000	2788	188	2600
300000-350000	1481	83	1398
350000-400000	957	48	909
400000-450000	393	30	363
450000-500000	456	37	419
500000-550000	124	9	115
550000-600000	43	5	38
600000-650000	40	1	39
650000-700000	117	8	109
700000-750000	22	1	21
750000-800000	11	0	11
800000-850000	21	2	19
850000-900000	5	0	5
900000-950000	30	2	28
950000-1000000	1	0	1
>1000000	40	3	37



AMT_CREDIT	APPLICANTS	DEFAULTER	NON DEFAULTER
0-100000	989	57	932
100000-200000	4911	333	4578
200000-300000	8849	687	8162
300000-400000	4256	439	3817
400000-500000	5228	534	4694
500000-600000	5554	595	4959
600000-700000	3909	307	3602
700000-800000	3062	250	2812
800000-900000	2547	209	2338
900000-1000000	2548	156	2392
1000000-1100000	2219	162	2057
1100000-1200000	1396	84	1312
1200000-1300000	1463	75	1388
1300000-1400000	945	49	896
1400000-1500000	389	17	372
1500000-1600000	649	28	621
1600000-1700000	144	9	135
1700000-1800000	255	11	244
1800000-1900000	241	8	233
1900000-2000000	122	5	117
2000000-2100000	115	5	110
2100000-2200000	36	3	33
2200000-2300000	89	0	89

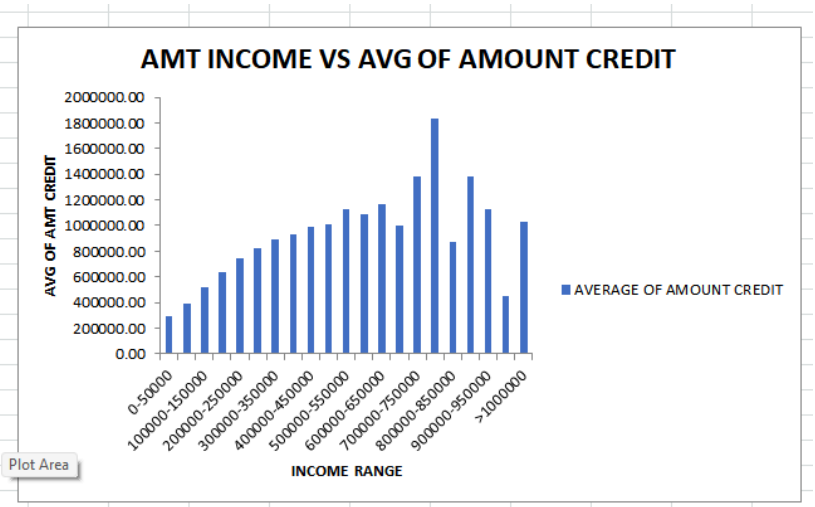


AMT_GOODS_PRICE	APPLICANTS	DEFAULTER	NON DEFAULTER
0-100000	1434	83	1351
100000-200000	5372	400	4972
200000-300000	10170	865	9305
300000-400000	3430	370	3060
400000-500000	9221	976	8245
500000-600000	2150	142	2008
600000-700000	6507	521	5986
700000-800000	1349	95	1254
800000-900000	1040	76	964
900000-1000000	3780	233	3547
1000000-1100000	576	33	543
1100000-1200000	2082	107	1975
1200000-1300000	314	16	298
1300000-1400000	1349	59	1290
1400000-1500000	102	3	99
1500000-1600000	396	23	373
1600000-1700000	69	0	69
1700000-1800000	78	3	75
1800000-1900000	376	15	361
1900000-2000000	29	3	26
2000000-2100000	21	0	21

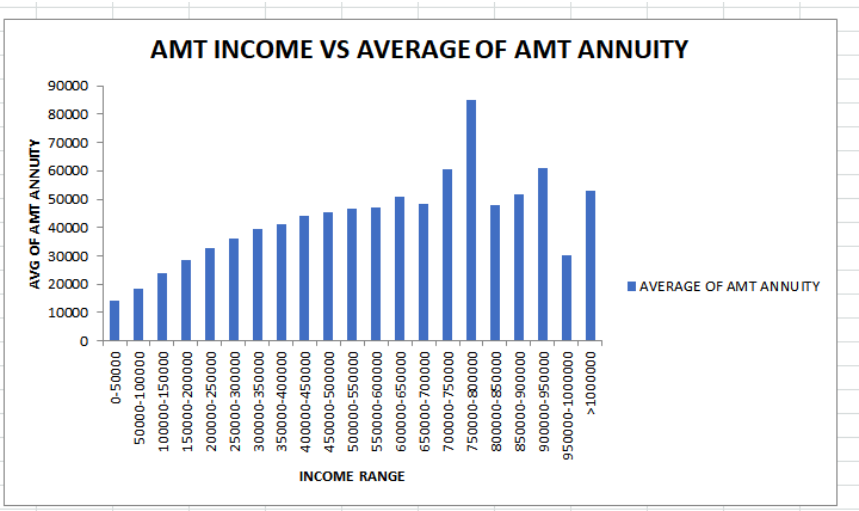


BIVARIATE ANALYSIS

INCOME RANGE	AVERAGE OF AMOUNT CREDIT
0-50000	297752.08
50000-100000	393033.34
100000-150000	520073.66
150000-200000	632290.90
200000-250000	741970.00
250000-300000	826106.65
300000-350000	892307.68
350000-400000	932609.93
400000-450000	993467.84
450000-500000	1011895.84
500000-550000	1124198.82
550000-600000	1091708.16
600000-650000	1165433.74
650000-700000	1001836.27
700000-750000	1386633.68
750000-800000	1836769.09
800000-850000	876760.07
850000-900000	1380720.60
900000-950000	1132219.80
950000-1000000	450000.00
>1000000	1030422.71



INCOME RANGE	AVERAGE OF AMT ANNUITY
0-50000	14086.09701
50000-100000	18478.04427
100000-150000	23850.11907
150000-200000	28489.43595
200000-250000	32838.22276
250000-300000	36130.31438
300000-350000	39291.98244
350000-400000	41170.55172
400000-450000	44015.24427
450000-500000	45523.59868
500000-550000	46640.68548
550000-600000	47158.95349
600000-650000	50893.2
650000-700000	48305
700000-750000	60653.45455
750000-800000	84841.36364
800000-850000	47799.85714
850000-900000	51605.1
900000-950000	60760.35
950000-1000000	30073.5
>1000000	52976.5875

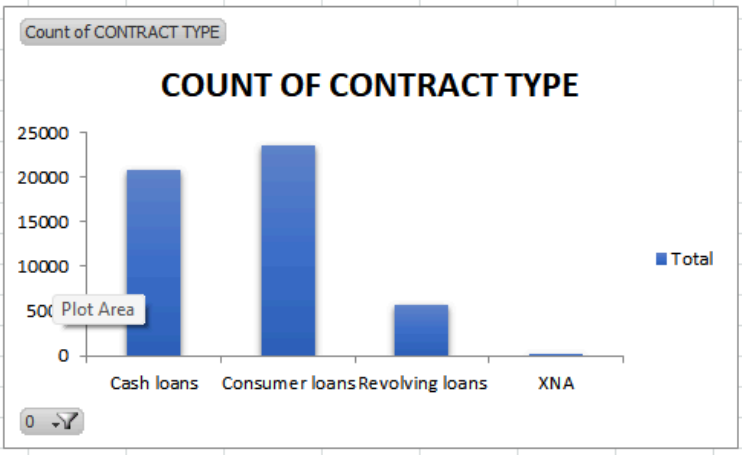


- **Secondary File: previous_application.csv**

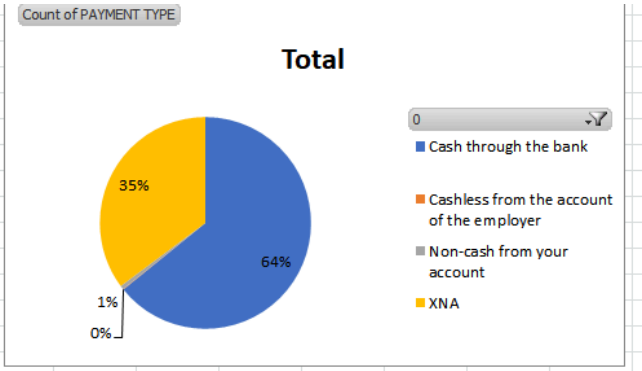
UNIVARIATE ANALYSIS

Result:

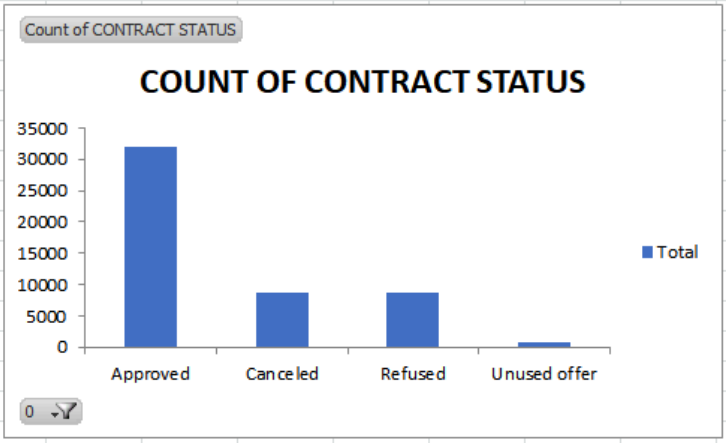
Row Labels	Count of CONTRACT TYPE
Cash loans	20856
Consumer loans	23510
Revolving loans	5625
XNA	8
Grand Total	**49999**



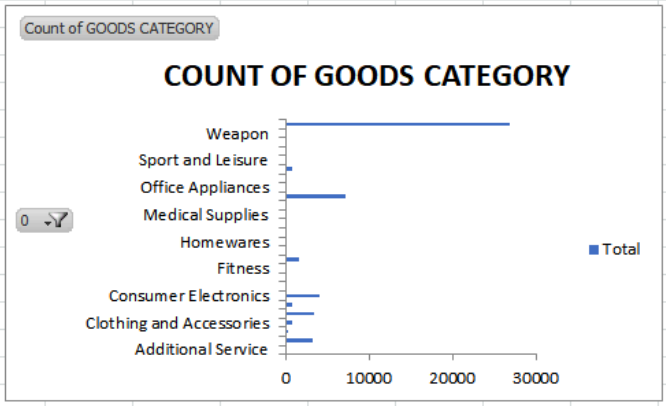
Row Labels	Count of PAYMENT TYPE
Cash through the bank	32089
Cashless from the account of the employer	35
Non-cash from your account	286
XNA	17589
Grand Total	**49999**



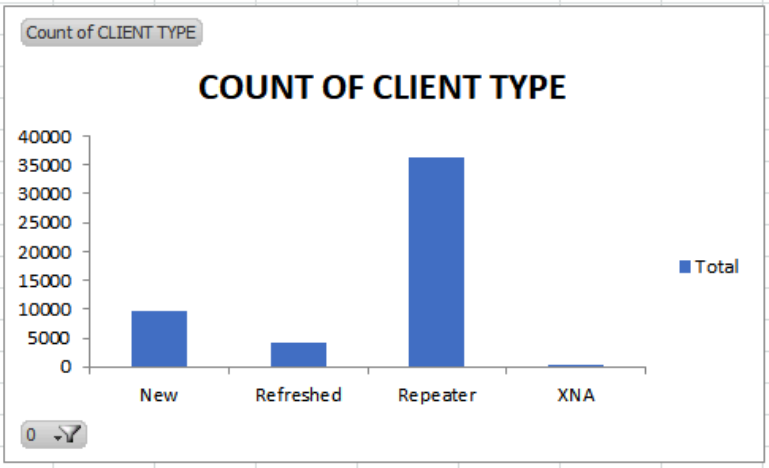
Row Labels	Count of CONTRACT STATUS
Approved	31885
Canceled	8595
Refused	8660
Unused offer	859
Grand Total	**49999**



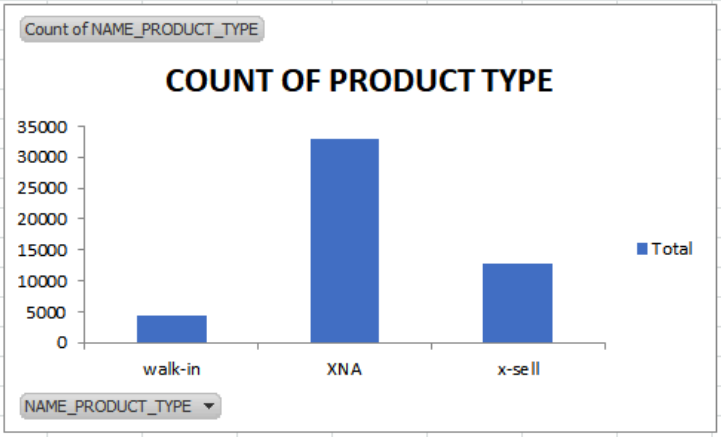
Row Labels	Count of GOODS CATEGORY
Additional Service	3
Audio/Video	3254
Auto Accessories	248
Clothing and Accessories	765
Computers	3344
Construction Materials	843
Consumer Electronics	4067
Direct Sales	13
Education	7
Fitness	9
Furniture	1696
Gardening	83
Homewares	154
Insurance	3
Jewelry	168
Medical Supplies	141
Medicine	56
Mobile	7149
Office Appliances	82
Other	68
Photo / Cinema Equipment	818
Sport and Leisure	85
Tourism	55
Vehicles	100



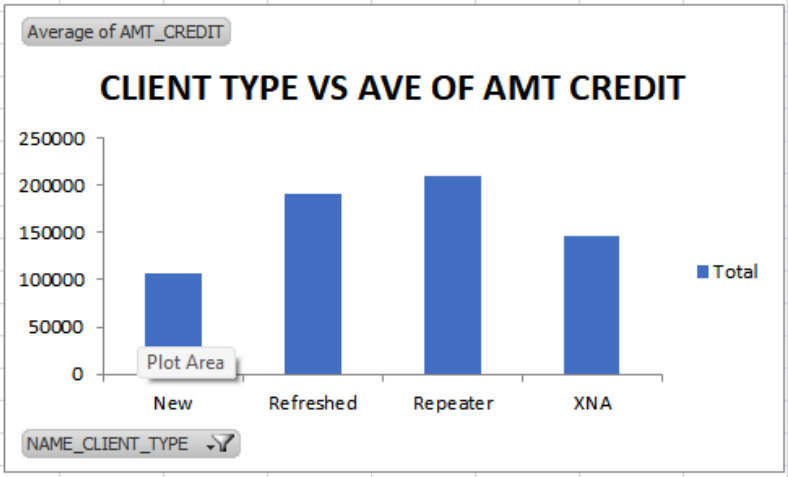
Row Labels	Count of CLIENT TYPE
New	9548
Refreshed	4227
Repeater	36167
XNA	57
Grand Total	**49999**



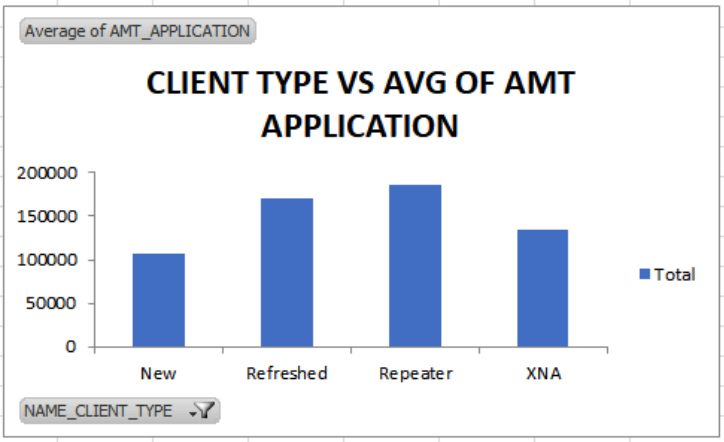
Row Labels	Count of NAME_PRODUCT_TYPE
walk-in	4447
XNA	32872
x-sell	12680
Grand Total	**49999**



CLIENT TYPE	Average of AMT_CREDIT
New	106535.8283
Refreshed	190349.3485
Repeater	210048.1329
XNA	146212.7092
Grand Total	**188542.8855**



CLIENT TYPE	Average of AMT_APPLICATION
New	106637.81
Refreshed	171221.8678
Repeater	185109.2583
XNA	134642.9724
Grand Total	**168892.4546**



E. Identify Top Correlations for Different Scenarios:

- **Task:** Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

- **Primary File**: **application_data.csv**

In this task I separate the numerical columns based on different scenarios like the dataset thats column's values contains only 0 target and named it 'dataset_target0' and so on for target 1, named it 'dataset_target1'.

This is done by the 'Advanced filter' option in excel.

Then I calculate correlation coefficients between variables and the target variable within each segment by using the CORREL function. After that, Ranking the correlations to identify the top indicators of loan default for each scenario for defaulters and non defaulters.

With the help of ‘conditional formatting option’ in excel i highlight the lower, mid point and highest values using 3 color scales. Also Highlight the top correlated variables for each scenario (the values between 0.6 to 0.99) using purple color.

Result:

TOP CORRELATION FOR TARGET 1 (DEFAULTERS)

VARIABLE 1	VARIABLE 2	CORRELATION
AMT_CREDIT	AMT_ANNUITY	0.749665201
AMT_CREDIT	AMT_GOODS_PRICE	0.982130206
AMT_ANNUITY	AMT_GOODS_PRICE	0.74932991
CNT_CHILDREN	CNT_FAM_MEMBERS	0.892521875
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950768899
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.806743886
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.783754676
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.89051161

TOP CORRELATION FOR TARGET 0 (NON DEFAULTER)

VARIABLE 1	VARIABLE 2	CORRELATION
AMT_CREDIT	AMT_ANNUITY	0.770771802
AMT_CREDIT	AMT_GOODS_PRICE	0.986904954
AMT_ANNUITY	AMT_GOODS_PRICE	0.775727492
CNT_CHILDREN	CNT_FAM_MEMBERS	0.879238049
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950468157
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.861374946
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.825358079
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.850995792

Result

In this project, I gained knowledge about how data analytics is useful in the field of financial sector like by analyzing different data patterns we can find out defaulters and non defaulters applicants and make better decisions about loan approval so that the financial institute doesn't face any type of loss.

We can also manage applicants data by finding outliers, data imbalance etc. I have also gained experience for data analysis using statistical knowledge and excel's tools and techniques. Through this I have learnt to apply my data analytics skills in solving real life problems.

Excel sheet Link

Bank loan case study sheet
Click here

Previous application_univariate analysis sheet (Secondary file for task D)
Click here

Video Presentation Link

You can view the presentation through this link
Click here