MODULE 6

AUGUST 2024

BANK LOAN CASE STUDY

FINAL PROJECT-2

PROJECT REPORT

Gunisha Chopra

DATA ANALYST

BANK LOAN CASE STUDY

FINAL PROJECT- 2

PROJECT DESCRIPTION

The Loan-providing companies find it hard to give loans to people due to their insufficient or non-existent credit history. I'm working as a data analyst for a finance company specialising in lending various loans to urban customers. I used Exploratory Data Analysis (EDA) to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected and help us understand how customer attributes and loan attributes influence the likelihood of default. This will also help us to develop a basic understanding of risk analytics in banking and financial services using the EDA module and understand how data is used to minimize the risk of losing money while lending to customers.

This Project aims to identify patterns that indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected thus preventing us from losing business. Identification of such applicants using EDA is the aim of this case study.

The company wants to understand the driving factors (or driver variables) behind loan default, i.e., the variables that are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

APPROACH

I have been provided with the following datasets:

1. `application_data.csv` contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
2. `previous_application.csv` contains information about the client's previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.
3. `columns_descrption.csv` is a data dictionary that describes the meaning of the variables.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- The client with payment difficulties: he/she had a late payment of more than X days on at least one of the first Y instalments of the loan in our sample
- All other cases: All other cases when the payment is paid on time.

When a client applies for a loan, four types of decisions could be taken by the client/company:

1. Approved: The company has approved the loan application
2. Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk the client received worse pricing which he did not want.
3. Refused: The company had rejected the loan (because the client does not meet their requirements etc.).
4. Unused Offer: The loan has been cancelled by the client but in different stages of the process.

We have used the below approach for deriving the insights:

• The required libraries needed for data cleansing and visualisation are imported.

• We have done the data cleansing for columns wherever necessary and dropped the columns with the majority of data as NA. Outliers are identified and handled wherever possible. Data imbalance is checked.

• Created new columns as per the requirements

• Analysis of the relevant Categorical/numerical is done and insights are derived

• Current and Previous application data is done to derive insights based on bank Approval loan status.

TECH-STACK USED:

MICROSOFT EXCEL 2019

TASKS AND INSIGHTS:

A.) IDENTIFY MISSING DATA AND DEAL WITH IT APPROPRIATELY:

As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

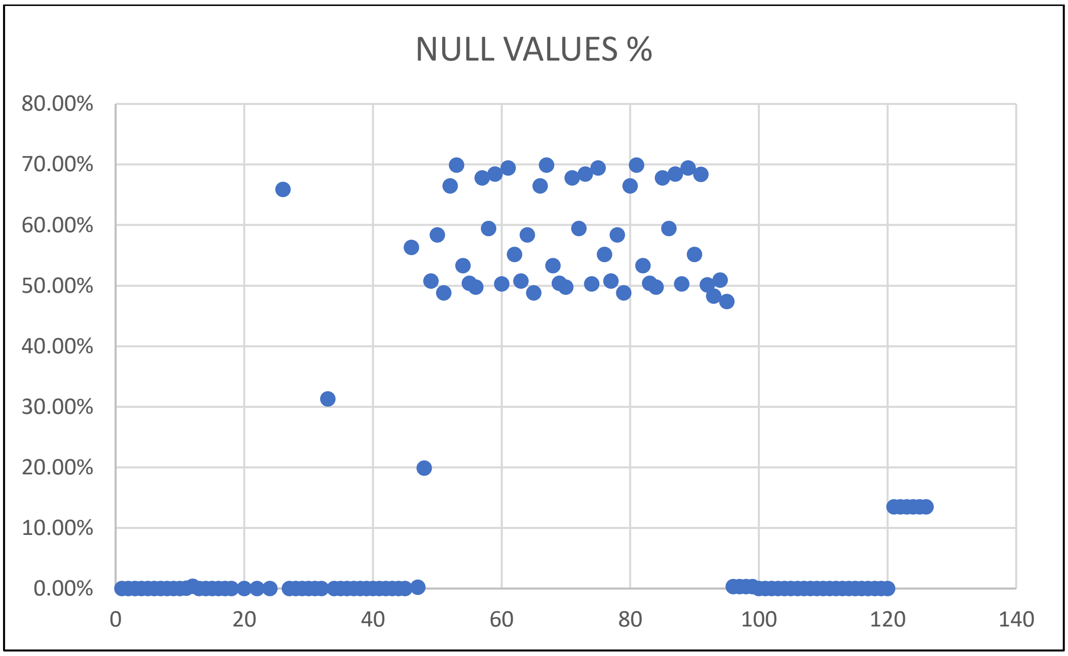
#**TASK**: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel's built-in functions and features.

#**INSIGHTS:** The original raw data has 122 variables scattered over 50000 rows. Among these are several missing values.

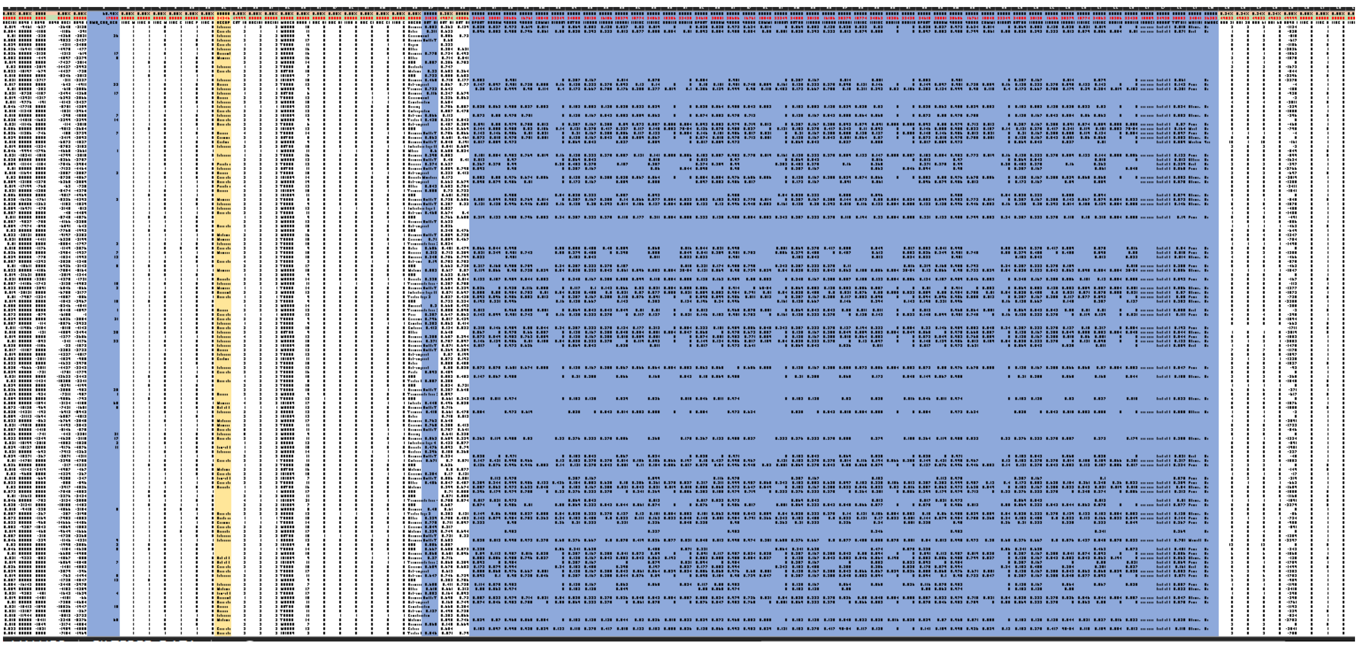
COLUMN COUNTS BEFORE FILTER:

ROW COUNT BEFORE FILTERING:

We then proceed to find the count in each column and the null values it may contain.

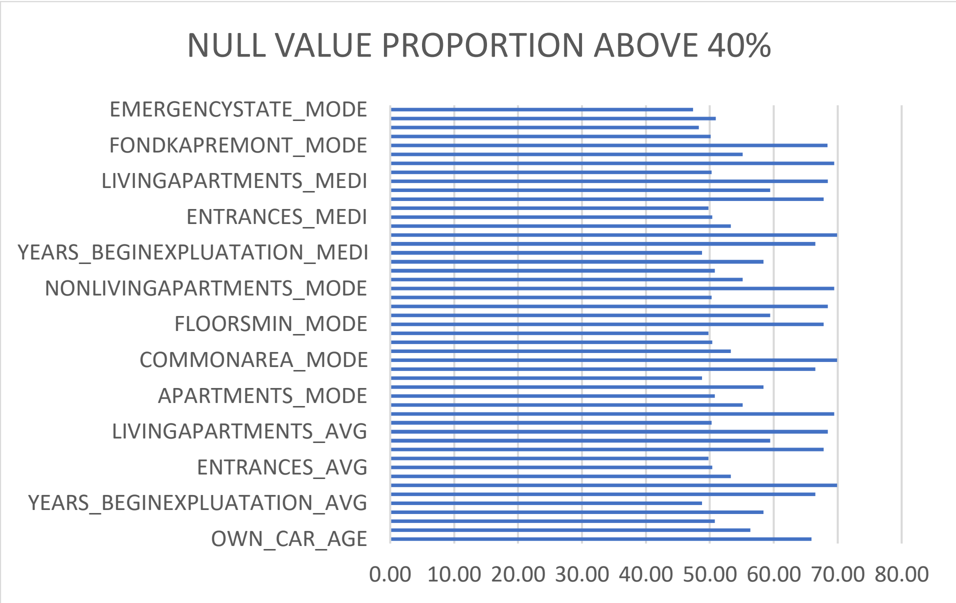


Most of the null values are between the 40-60% slot. So we proceed to mark them to make our further analysis easier.

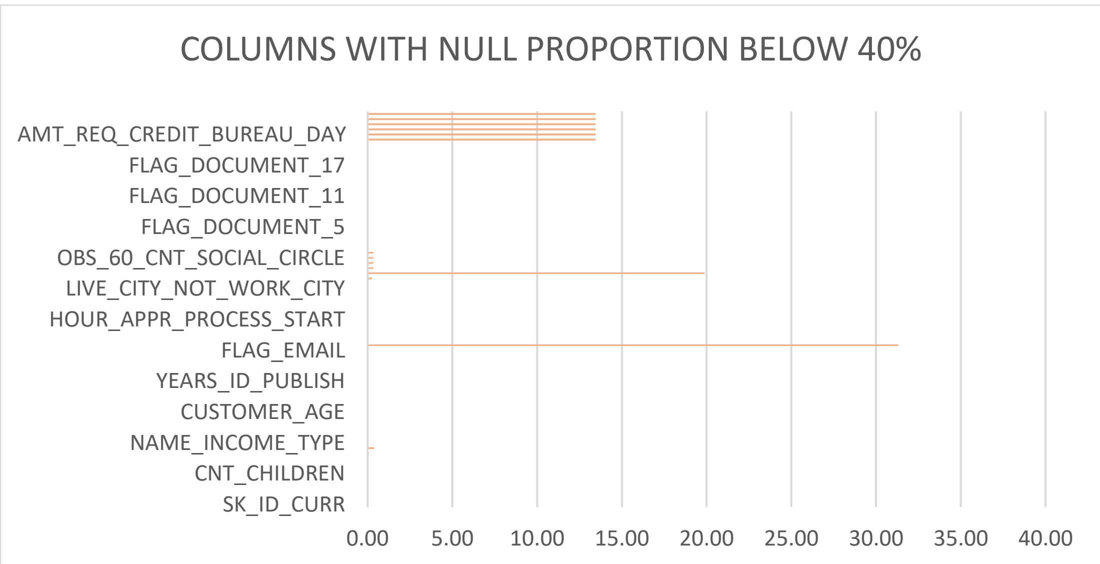


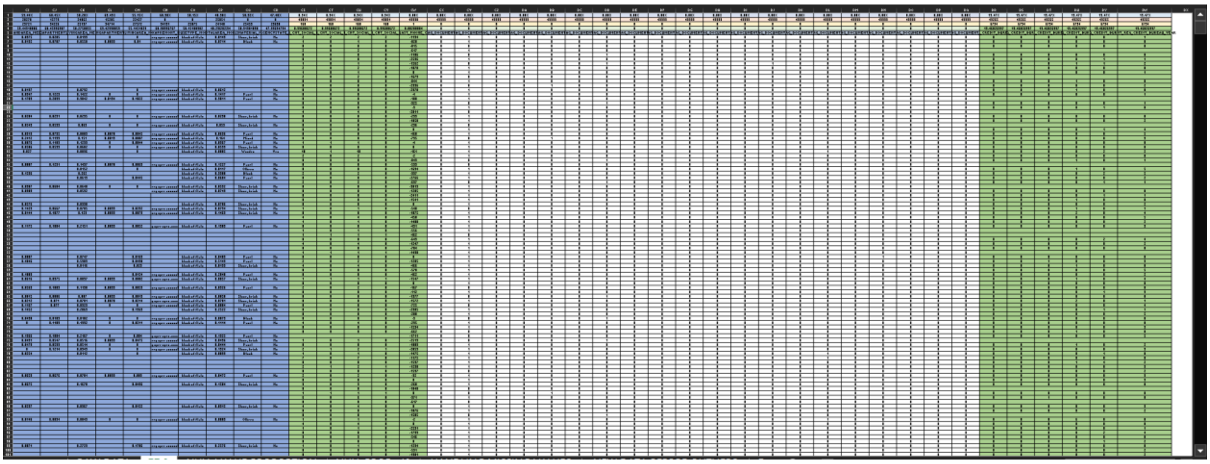
We can see a huge BLUE REGION which indicates columns with above 40% null value percentage.

Therefore, we proceed to delete these columns which might cause hindrances in our analysis.



But we are not done with missing data just yet, even though we have detected the columns with more than 40% missing data, we cannot ignore the rest of the missing data that is not as big as 40% but close or less.





We have already marked the columns with more than 40% null values “blue”.

We proceed to mark other columns with missing data with “green”.

#HANDLING THE MISSING VALUES:

We extract the green columns and divide them into two parts:

- Numerical data
- Categorical data

NUMERICAL DATA COLUMNS:

AMT_ANNUITY

AMT_GOODS_PRICE

CNT_FAM_MEMBERS

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_CHANGE

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

We proceed to calculate the skewness, mean and median of each column, find all the blank cells and then process to fill them with the 'median'.

CATEGORICAL DATA COLUMNS:

OCCUPATION_TYPE

NAME_TYPE_SUITE

We proceed to find the count of each variable of the NAME_TYPE_SUITE column using the COUNTIF function.

Unaccompanied	40435
Family	6549
Children	542
Spouse, partner	1849
other_A	137
Other_B	259
Group of people	36

Then we proceed by filling any blank cells with "unaccompanied".

For the OCCUPATION_TYPE column, we fill the blank cells with "unknown".

Now that we have handled the missing values in our below 40% (green) columns, we proceed to delete all the blue columns and get a final count.

COLUMN COUNT AFTER FILTER: Count: 73

ROW COUNT AFTER FILTER: Count: 41071

B.) IDENTIFY OUTLIERS IN THE DATASET:

Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

#TASK: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

COLUMNS THAT MIGHT CONTAIN OUTLIERS:

CNT_CHILDREN

AMT_INCOME_TOTAL

AMT_CREDIT

AMT_ANNUITY

AMT_GOODS_PRICE

REGION_POPULATION_RELATIVE

CUSTOMER_AGE

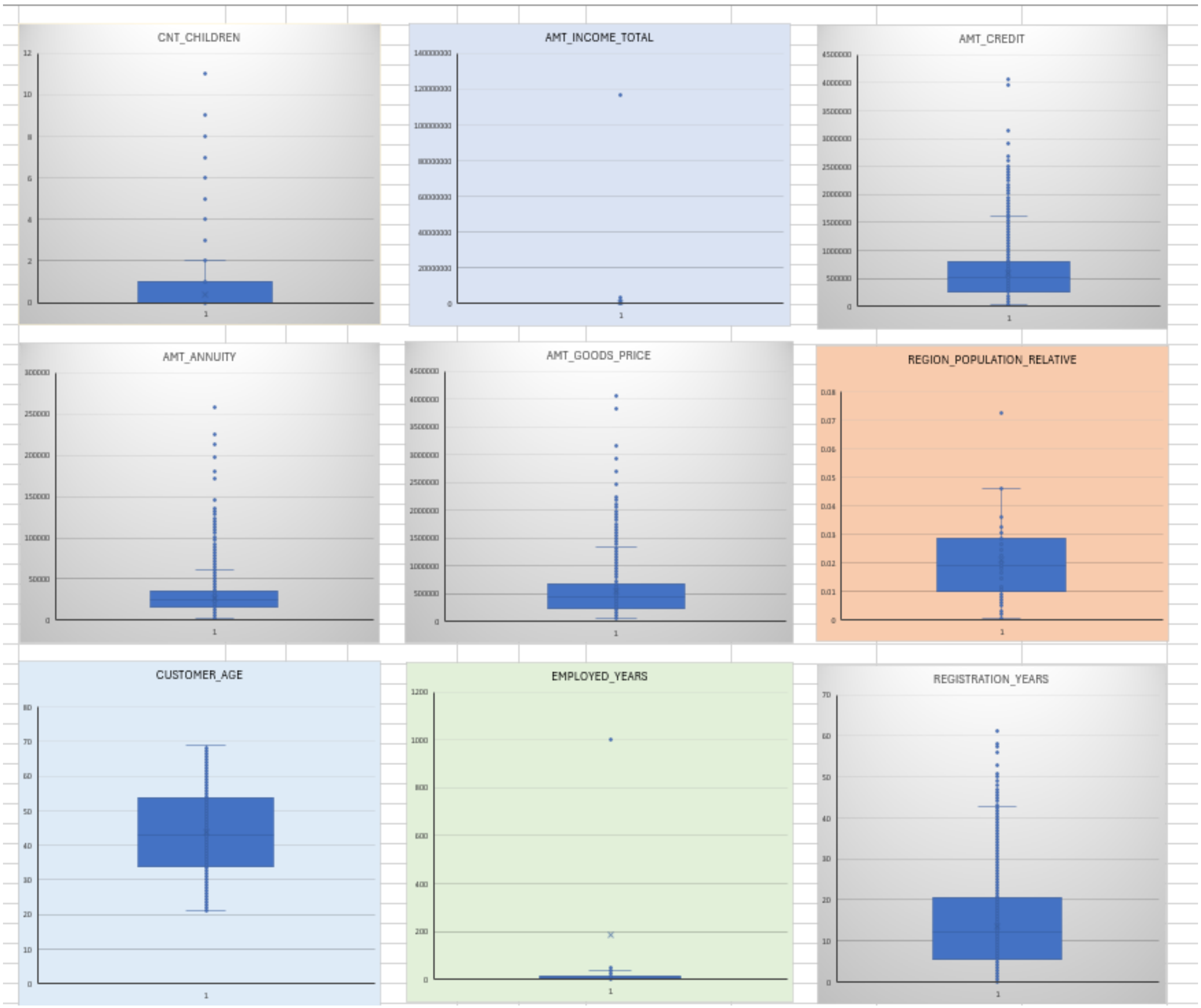
EMPLOYED_YEARS

REGISTRATION_YEARS

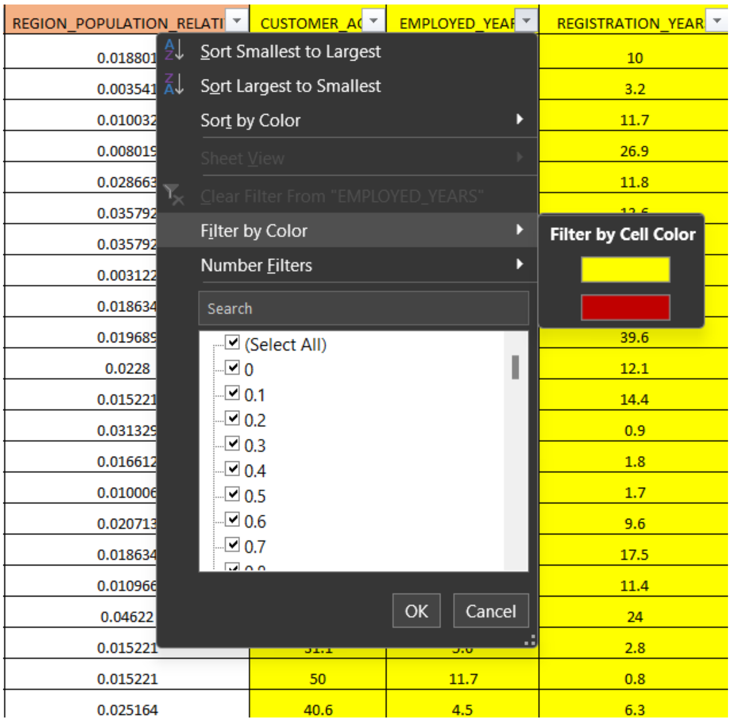
WE PROCEED TO FIND QUARTILE 1, QUARTILE 3, INTER QUARTILE, UPPER LIMIT AND LOWER LIMIT FOR THESE COLUMNS.

	QUARTILE 1	QUARTILE 3	INTER QUARTILE RANGE	UPPER LIMIT	LOWER LIMIT
CNT_CHILDREN	0	1	1	2.5	-1.5
AMT_INCOME_TOTAL	112500	202500	90000	337500	-22500
AMT_CREDIT	270000	808650	538650	1616625	-537975
AMT_ANNUITY	16456.5	34596	18139.5	61805.25	-10752.75
AMT_GOODS_PRICE	238500	679500	441000	1341000	-423000
REGION_POPULATION_RELATIVE	0.010	0.028663	0.018657	0.0566485	-0.0179795
CUSTOMER_AGE	33.900	53.8	19.9	83.65	4.05
EMPLOYED_YEARS	2.600	15.7	13.1	35.35	-17.05
REGUSTRATION_YEARS	5.500	20.4	14.9	42.75	-16.85

#For better understanding, we proceed to make box plots for each column:



##We then proceed to extract the outlier by colour in each column, for easier access and filtering process of the outliers simpler. We chose 'RED'.



Analysis for each column:

→CNT_CHILDREN:

We see that one of the clients has 11 children, which is not possible in today's economy and hence it is an outlier.

→AMT_INCOME_TOTAL

We notice one of the clients earning 117,000,000 (one hundred seventeen million) which we mark as an outlier.

→AMT_CREDIT

We notice several values indicate that the credit on the loan is over 4,000,000 so we mark them as an outlier.

→AMT_ANNUITY

Several columns indicate loan annuity to be more than 250,000, we mark it as an outlier.

→AMT_GOODS_PRICE

The price of the goods for which the loan is given seems to be more than 4,000,000 in some cells, thus we mark them as outliers.

→REGION_POPULATION_RELATIVE

The normalized population of the region where the client lives (a higher number means the client lives in a populated region. We have an outlier at a value of 0.072508 hence an outlier.

→CUSTOMER_AGE

There doesn't seem to be any outlier in the customer age column.

→EMPLOYED_YEARS

It is physically impossible for a person to work for 1000 years. Hence that indicates 1000 years are our outliers.

→REGISTRATION_YEARS

This column shows the number of years before the application that the client changed his/her registration and the cells showing 61 years are an outlier.

We now move to filter out the data by colour RED and then process to delete sheet rows and get the final data after outliers have been dealt with.

C.) ANALYSE DATA IMBALANCE:

Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

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

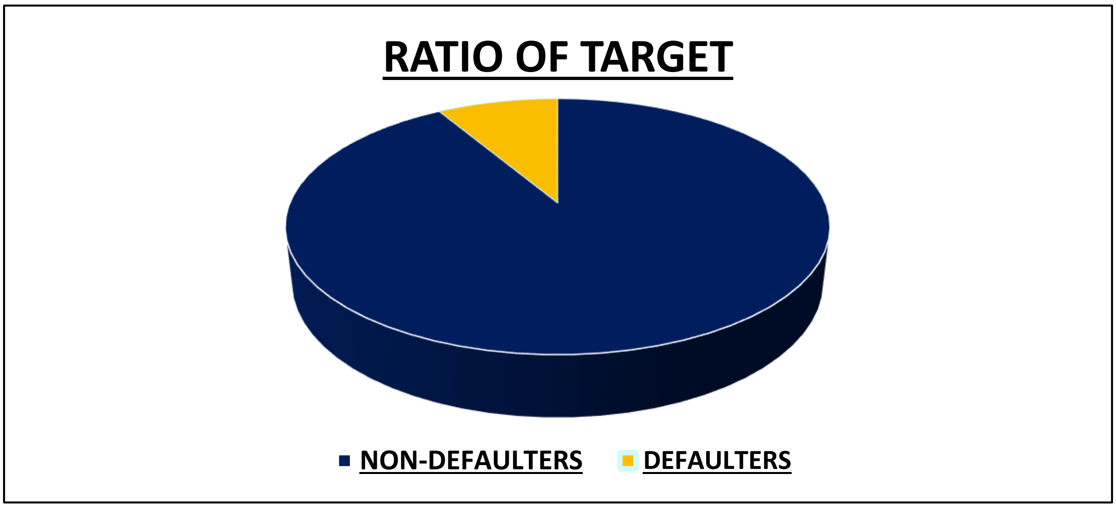
In the TARGET column, we have two kinds of variables: ‘0’ and ‘1’ where:

- ‘1’ means clients with payment difficulties or DEFAULTERS: he/she has late payment more than X days on at least one of the first Y instalments of the loan in our sample.
- ‘0’ stands for all the other cases: these are the cases where payments were made on time or NON-DEFAULTERS.

#As we talked about before, if our bank gets more of these defaulters, it will harm our business. We analyse if there is in fact an imbalance in our dataset.

CLIENTS	ROW LABELS	COUNT OF TARGET	
NON-DEFAULTERS	0	37549	
DEFAULTERS	1	3521	
	GRAND TOTAL	**41070**	

TARGET	PERCENTAGE OF TARGET	RATIO OF TARGET	RATIO OF DATA IMBALANCE
NON-DEFAULTERS	91%	0.91427	0.094
DEFAULTERS	9%	0.08573	0.086



We see a clear huge difference between defaulters and Non-defaulters i.e., there are more non-defaulters as compared to defaulters so we can tell that our business is booming.

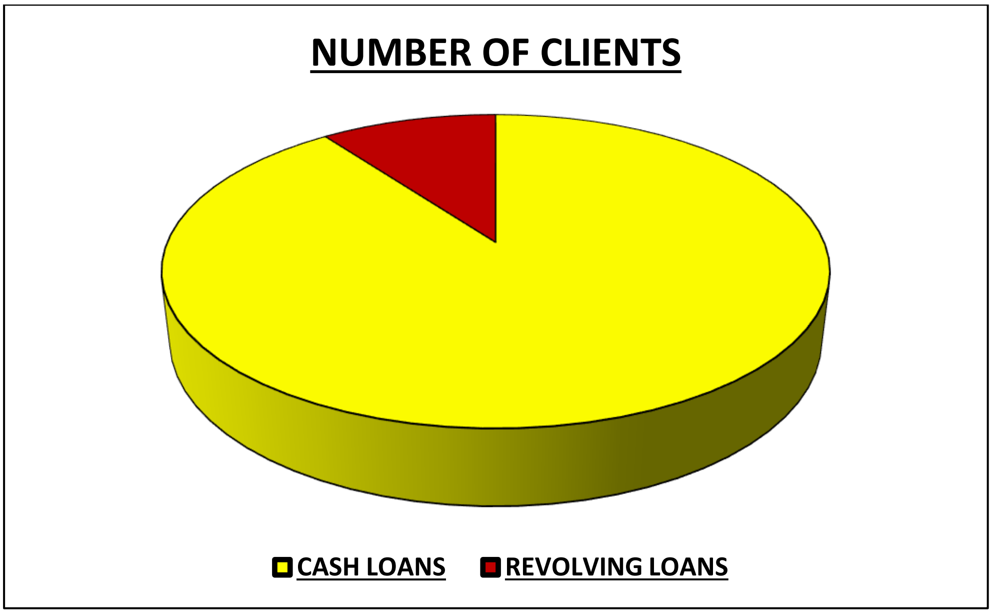
Now that we have confirmed that we have a good clientele when it comes to loan difficulties, we move on to the type of loans we are providing.

There are two types of loans we can see in our dataset:

- CASH LOANS
- REVOLVING LOANS

We proceed to analyse the imbalance between the two:

LOAN TYPE	NO. OF CLIENTS	PERCENTAGE OF LOAN CLIENTS
CASH LOANS	36893	90%
REVOLVING LOANS	4177	10%
GRAND TOTAL	41070	100%



We see an imbalance between the two types of loans we have i.e., <u>Cash loans are way more applicable than revolving loans.</u>

#INSIGHT:

D.) PERFORM UNIVARIATE, SEGMENTED UNIVARIATE, AND BIVARIATE ANALYSIS:

To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

#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.

CLIENT INFORMATION:

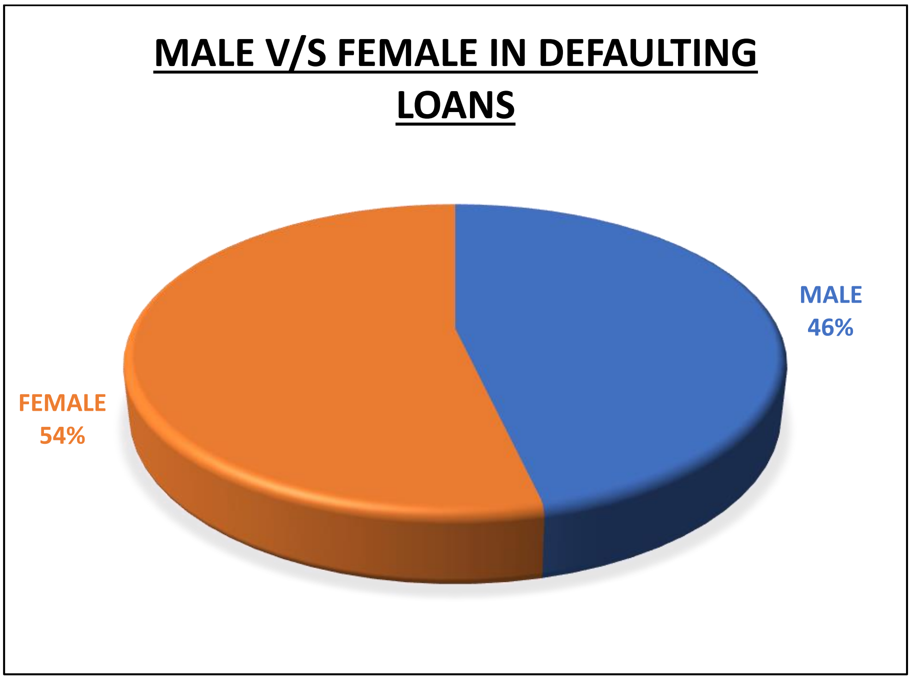
GENDER OF THE CLIENT: CODE_GENDER

When we analyse the client's information, we move on to the gender of the client and figure out how many defaulters are men and how many of them are men just for a better understanding of our clientele.

MALE	NON DEFAULTERS	DEFAULTERS	TOTAL
COUNT	13897	1634	15531
PERCENTAGE	89%	11%	100%

FEMALE	NON DEFAULTERS	DEFAULTERS	TOTAL
COUNT	23650	1887	25537
PERCENTAGE	93%	7%	100%

RATIO	OF MALE TO FEMALE DEFAULTING LOAN		
	MALE	FEMALE	TOTAL
COUNT	1633	1887	3520
PERCENTAGE	46%	54%	100%



#We see that 54% of the defaulters are FEMALES and 46% of them are MALES.

CLIENT'S AGE: CUSTOMER_AGE

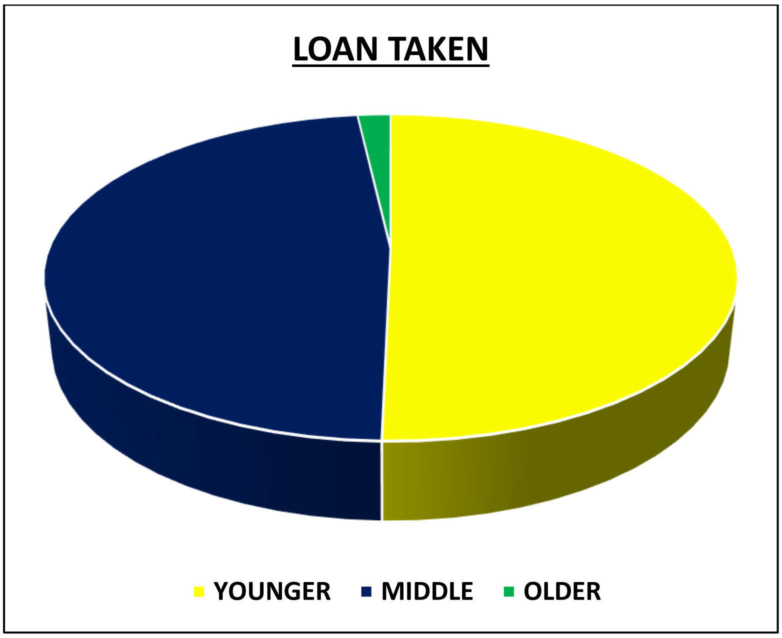
We first convert the client's age which is given in days into years by dividing it by 365. Then we perform further analysis by categorizing them into:

- 20-40 → YOUNGER
- 40-60 → MIDDLE
- >60 → OLDER

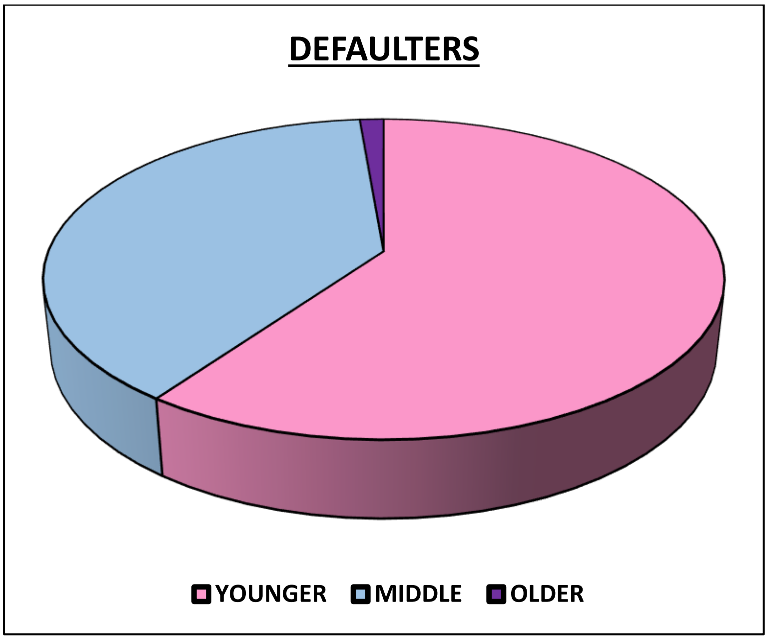
Then we analyse how the loan defaulters have been distributed throughout these categories by using the COUNTIFS function.

AGE	CATEGORY	DEFAULTERS	NON DEFAULTERS	LOAN TAKEN
20-40	YOUNGER	2108	18438	20546
40-60	MIDDLE	1366	18161	19527
>60	OLDER	47	698	745

We can see that most of the loans are taken by clients in age groups 20-40. The reason for this can be that the younger generation tends to take several loans for many aspects of life like education loans, or loans for cars or houses etc. closely followed by middle-aged clients and a very low percentage for the older generation which we can assume is because they have probably already paid off most of their loans. Here is a graphical representation for the same:



We further go on to analyse the number of defaulters in all age groups.



We see most of the defaulters tend to be young people with an overwhelmingly large percentage.

WHETHER CLIENT OWNS A CAR OR NOT: FLAG_OWN_CAR

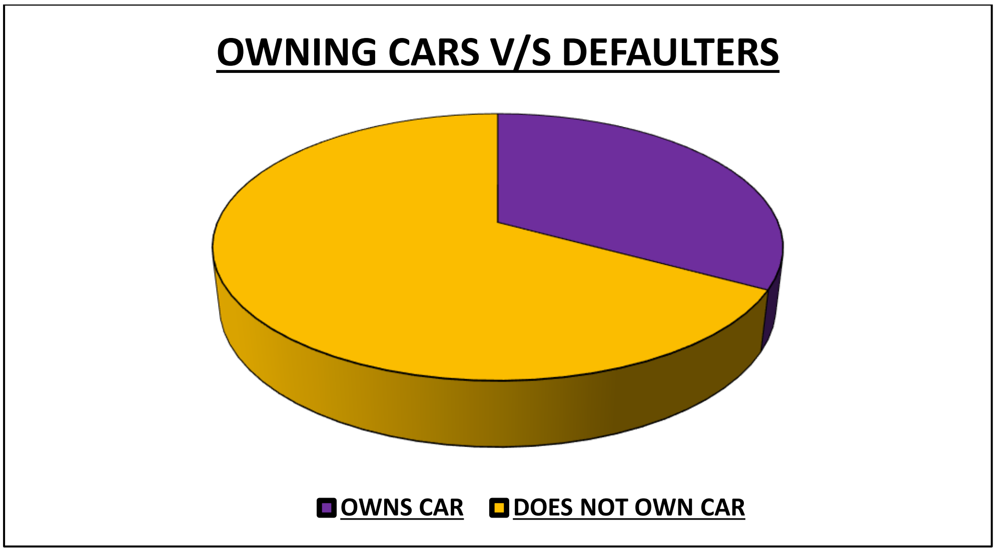
We analyse how our client being an owner of a car can affect our loan application:

OWNS CAR	DEFAULTERS	NON-DEFAULTERS	TOTAL
COUNT	1160	14232	15392
PERCENTAGE	8%	92%	100%

DOES NOT OWN CAR	DEFAULTERS	NON DEFAULTERS	TOTAL
COUNT	2361	23317	25678
PERCENTAGE	9%	91%	100%

RATIO	OWNS CAR	DOES NOT OWN CAR	TOTAL DEFAULTERS
COUNT	1160	2361	3521
PERCENTAGE	33%	67%	100%

Graphical representation for the same:



We see that clients who do not own a car tend to have more difficulty when it comes to paying off a loan.

WHETHER CLIENT OWNS REALTY: FLAG_OWN_REALTY

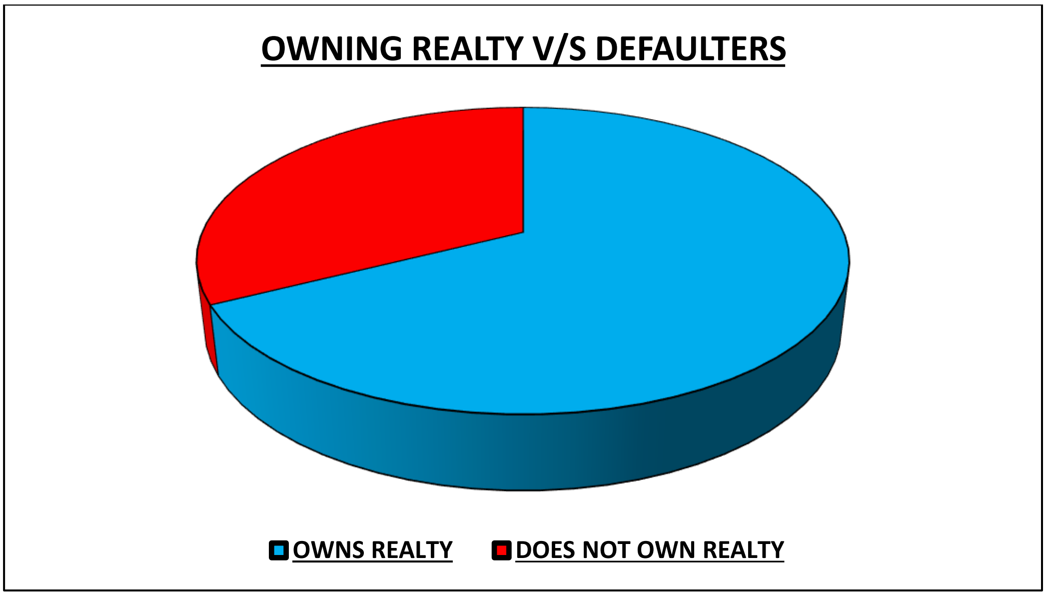
We analyse whether our client owns a house or a flat and how it affects our loan process.

OWNS REALTY	DEFAULTERS	NON DEFAULTERS	TOTAL
COUNT	2379	25561	27940
PERCENTAGE	9%	91%	100%

DOES NOT OWN REALTY	DEFAULTERS	NON DEFAULTERS	TOTAL
COUNT	1142	11988	13130
PERCENTAGE	9%	91%	100%

RATIO	OWNS REALTY	DOES NOT OWN REALTY	TOTAL DEFAULTERS
COUNT	2379	1142	3521
PERCENTAGE	68%	32%	100%

Graphical representation:

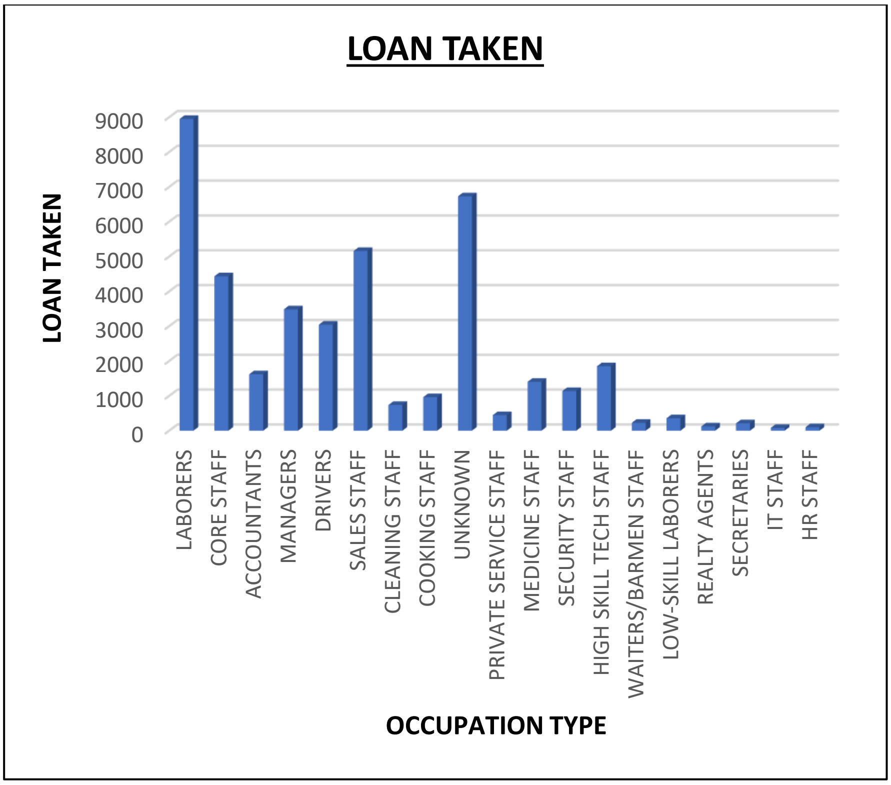


We see that people who own real estate tend to have more difficulty paying off their debts.

CLIENT'S OCCUPATION: OCCUPATION_TYPE

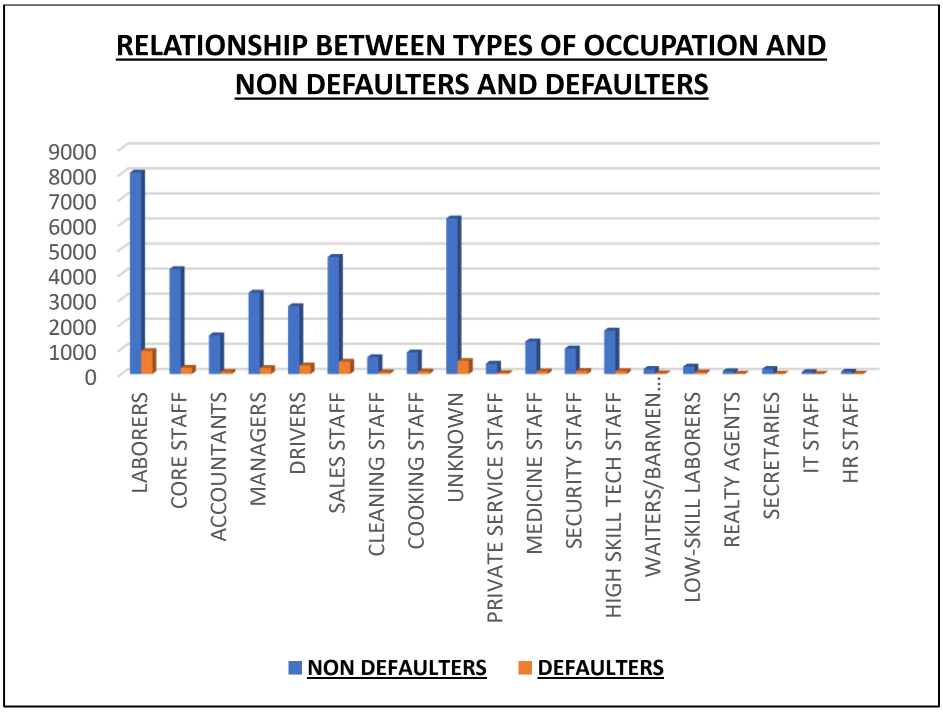
We analyse what kind of clientele we have based on their occupation and how this may affect their payment difficulties.

OCCUPATION	NON DEFAULTERS	DEFAULTERS	LOAN TAKEN
LABORERS	8031	919	8950
CORE STAFF	4184	250	4434
ACCOUNTANTS	1540	81	1621
MANAGERS	3244	242	3486
DRIVERS	2706	338	3044
SALES STAFF	4668	492	5160
CLEANING STAFF	671	68	739
COOKING STAFF	862	101	963
UNKNOWN	6207	523	6730
PRIVATE SERVICE STAFF	410	37	447
MEDICINE STAFF	1297	106	1403
SECURITY STAFF	1015	125	1140
HIGH SKILL TECH STAFF	1734	118	1852
WAITERS/BARMEN STAFF	203	25	228
LOW-SKILL LABORERS	296	61	357
REALTY AGENTS	110	13	123
SECRETARIES	203	9	212
IT STAFF	76	4	80
HR STAFF	92	9	101



We see that most of these loans are taken out by Laborers which is probably because these labourers may need the bank's help for their daily life things.

Since we can imagine that these labourers don't have a steady income, they tend to comprise most of the defaulting clientele.

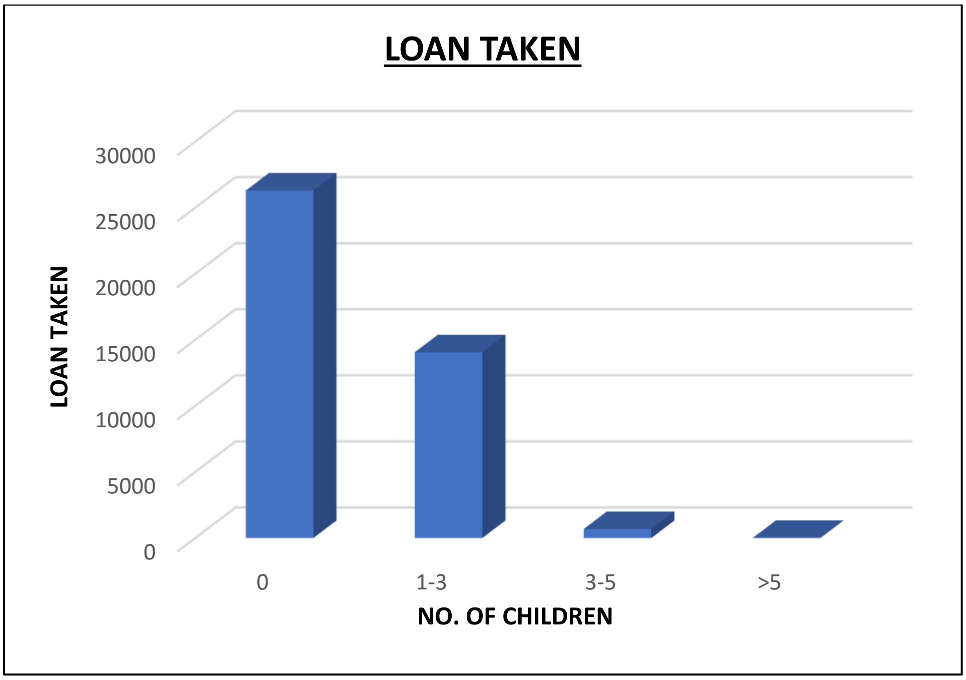


NUMBER OF CHILDREN THE CLIENT HAS: CNT_CHILDREN

We first categorize the number of children into:

- 0
- 1-3
- 3-5
- >5

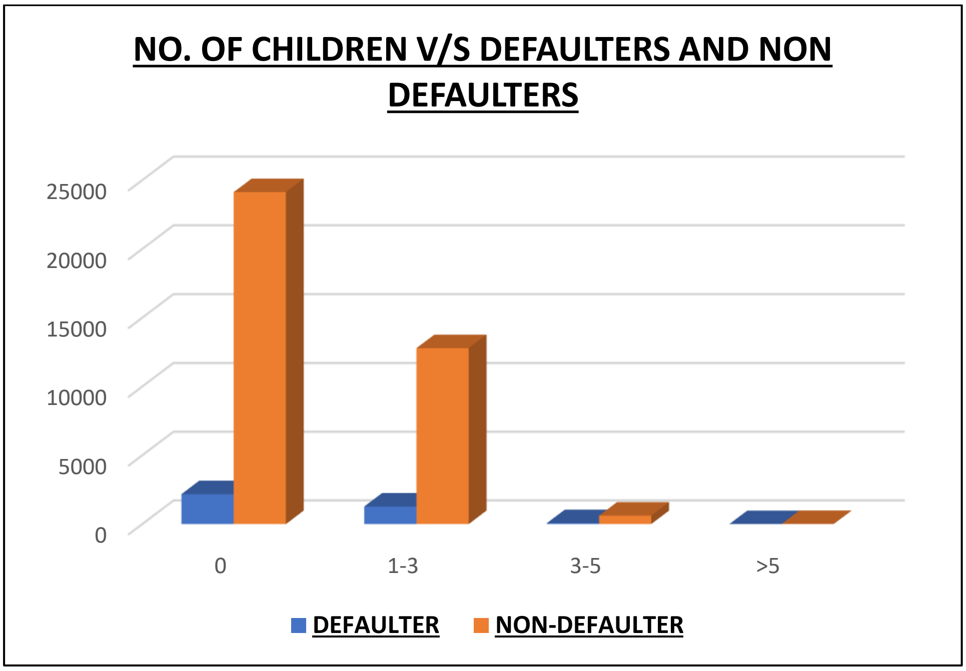
And further, analyse the loans taken by such clients and then figure out defaulters and non-defaulters:



Surprisingly, most of these loans are taken out by people with 0 children. We would imagine that people with more children would take out several loans for the kids.

People in this group also happen to consist large amount of defaulters as well. It could mean that these people may be young or even students who do not have a family yet and are taking loans for themselves and then facing difficulties during repayment.

Here is a graphical representation for the same:



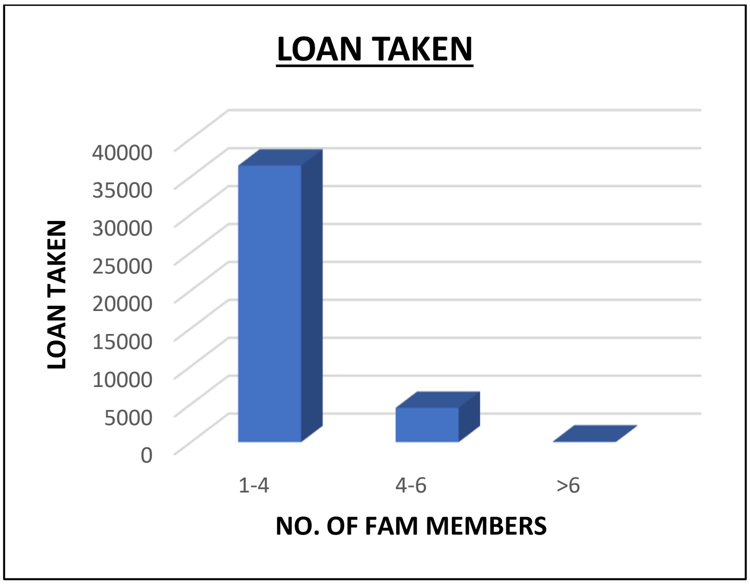
#<u>NUMBER OF FAMILY MEMBERS THE CLIENT HAS</u>: CNT_FAM_MEMBERS

Number of members in a family. Clients with larger families (more children or dependents) might face more payment difficulties.

For this analysis, we have divided our family member data into the following categories:

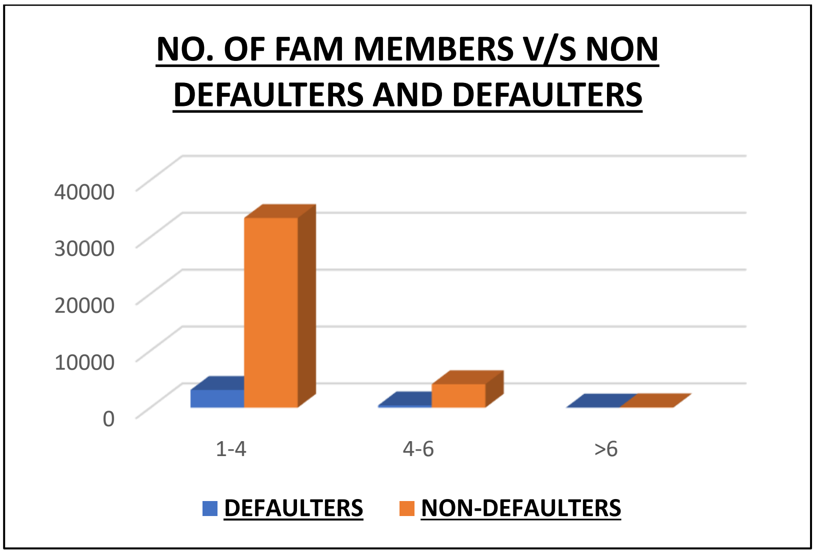
- 1-4
- 4-6
- >6

We analyse how many defaulters these categories contain and how much loan is being taken by these clients.



We see that most of the loans are taken by people in the first category i.e., 1-4 family members with an overwhelmingly high rate.

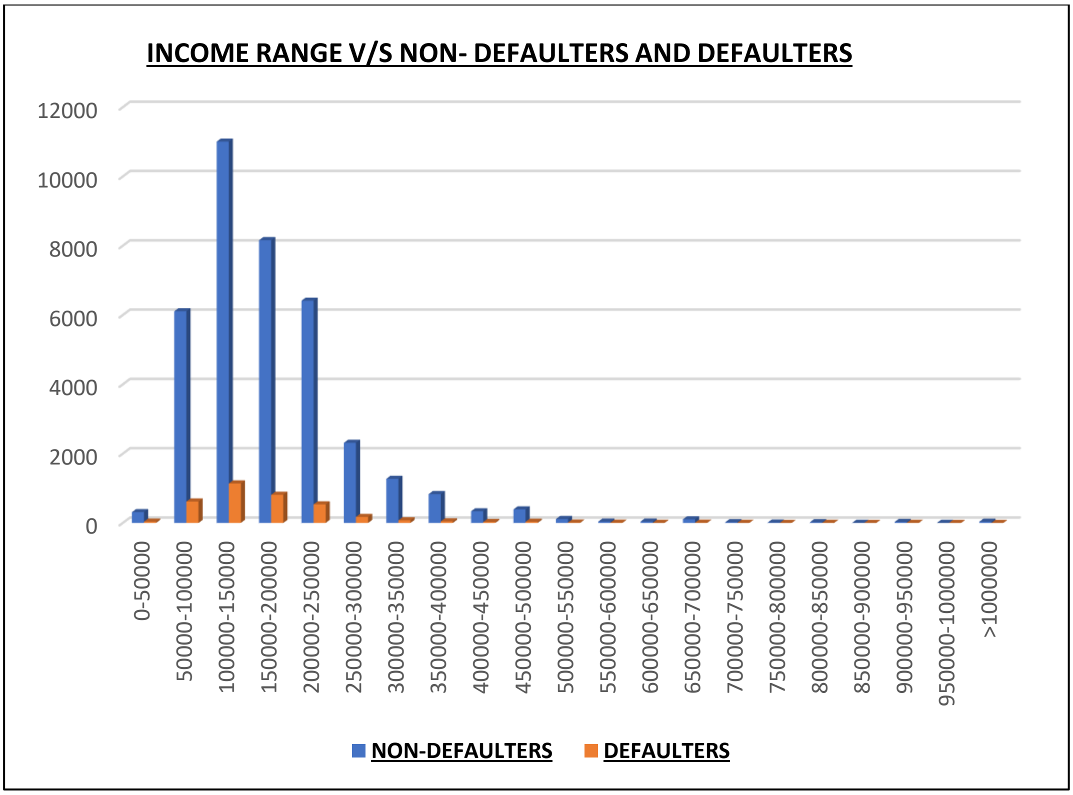
Naturally, most of the defaulters will also lie in this category. Here is a graphical representation of the same:

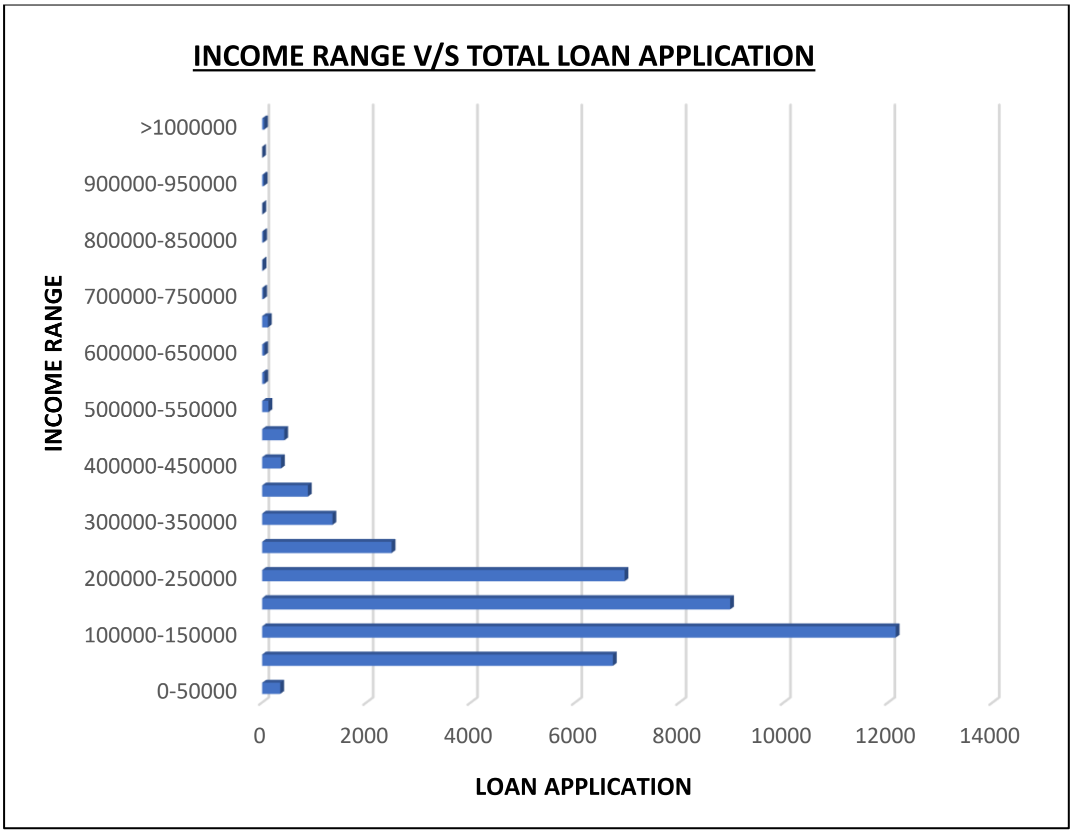


INCOME OF THE CLIENT AND DEFAULTERS: AMT_INCOME_TOTAL

We figure out the income range and which range has the most defaulters and non-defaulters.

INCOME_RANGE	NON-DEFAULTERS	DEFAULTERS	TOTAL
0-50000	310	37	347
50000-100000	6109	617	6726
100000-150000	11006	1139	12145
150000-200000	8161	812	8973
200000-250000	6414	536	6950
250000-300000	2312	171	2483
300000-350000	1273	76	1349
350000-400000	831	45	876
400000-450000	335	26	361
450000-500000	391	33	424
500000-550000	112	9	121
550000-600000	37	5	42
600000-650000	38	1	39
650000-700000	104	7	111
700000-750000	20	1	21
750000-800000	9	0	9
800000-850000	19	2	21
850000-900000	4	0	4
900000-950000	27	2	29
950000-1000000	1	0	1
>1000000	35	0	37



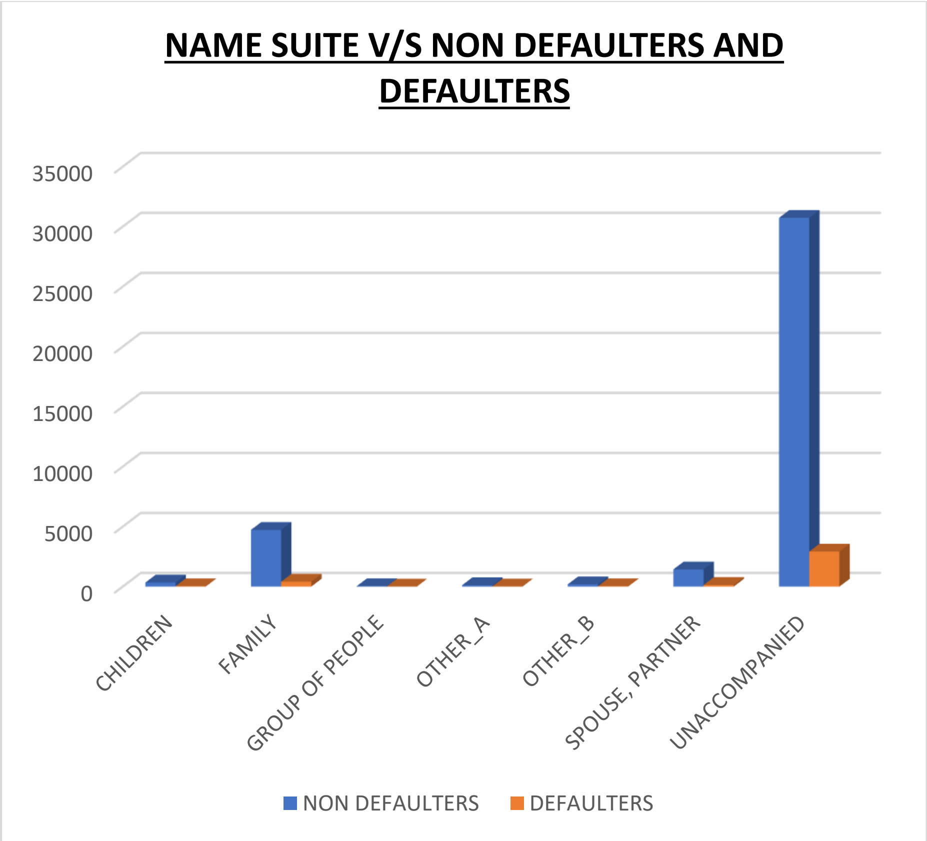


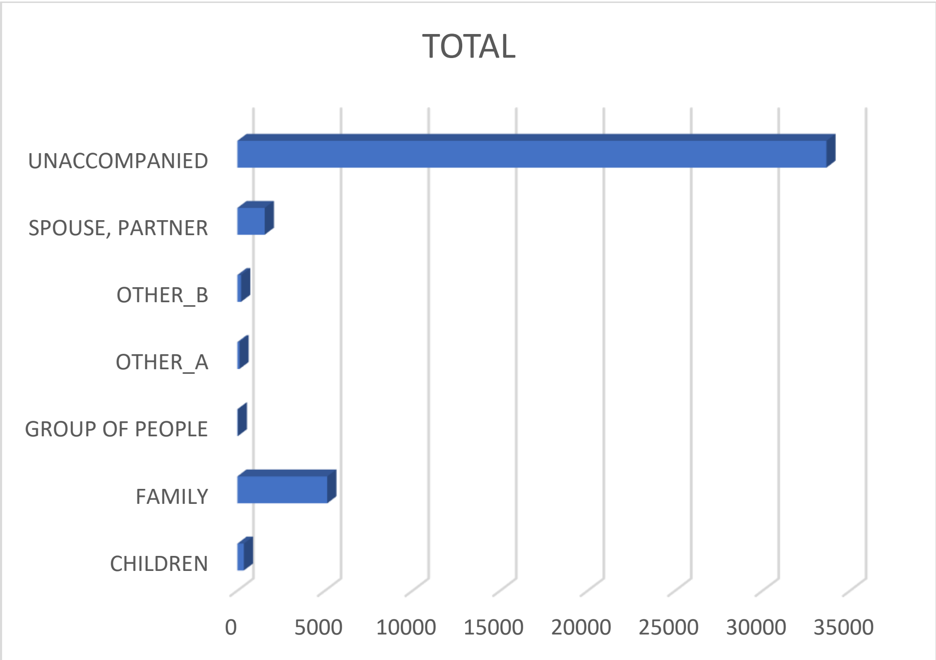
We see that most of the loans were taken out by clients who may be earning somewhere between 100,000-150,000 income range and most of the defaulters and non-defaulters are also in the income range as well.

PROPERTY DETAILS:

WHO WAS ACCOMPANYING THE CLIENT WHEN HE WAS APPLYING FOR A LOAN: NAME_TYPE_SUITE

NAME_TYPE_SUITE	NON DEFAULTERS	DEFAULTERS	TOTAL
CHILDREN	340	31	371
FAMILY	4723	406	5129
GROUP OF PEOPLE	26	1	27
OTHER_A	113	10	123
OTHER_B	187	26	213
SPOUSE, PARTNER	1429	132	1561
UNACCOMPANIED	30731	2915	33646



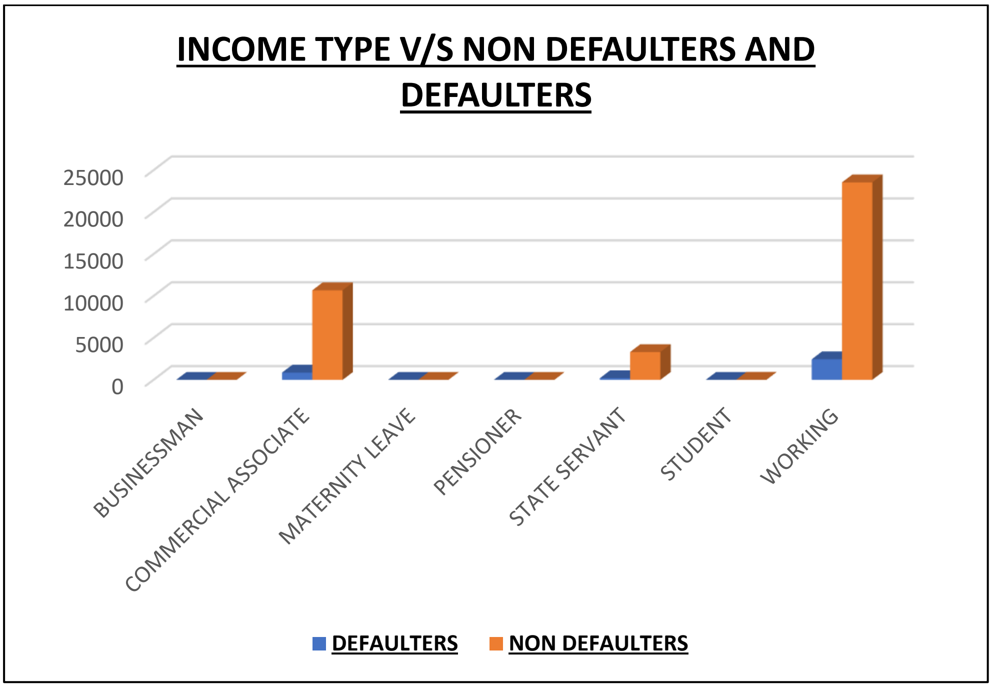


We see that most people come by themselves when applyingfor a loan.

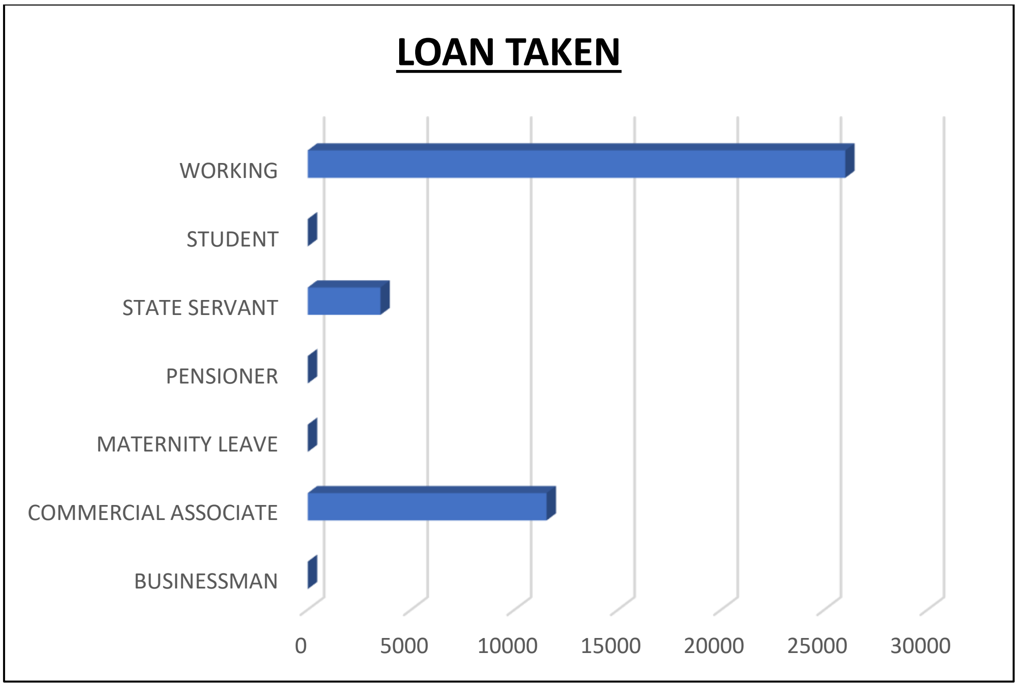
CLIENT'S INCOME TYPE: NAME_INCOME_TYPE

We analyse where our client gets his/her income flow from (businessman, pensioner, student, etc)

INCOME TYPE	DEFAULTERS	NON DEFAULTERS	LOAN TAKEN
BUSINESSMAN	0	2	2
COMMERCIAL ASSOCIATE	864	10677	11541
MATERNITY LEAVE	0	1	1
PENSIONER	0	2	2
STATE SERVANT	198	3314	3512
STUDENT	0	5	5
WORKING	2459	23548	26007



We see that working clients tend to apply for more loans and pay off their debts well. This can be due to a steady income. In contrast, businessmen tend to take less loans from banks.



CLIENT'S EDUCATION: NAME_EDUCATION_TYPE

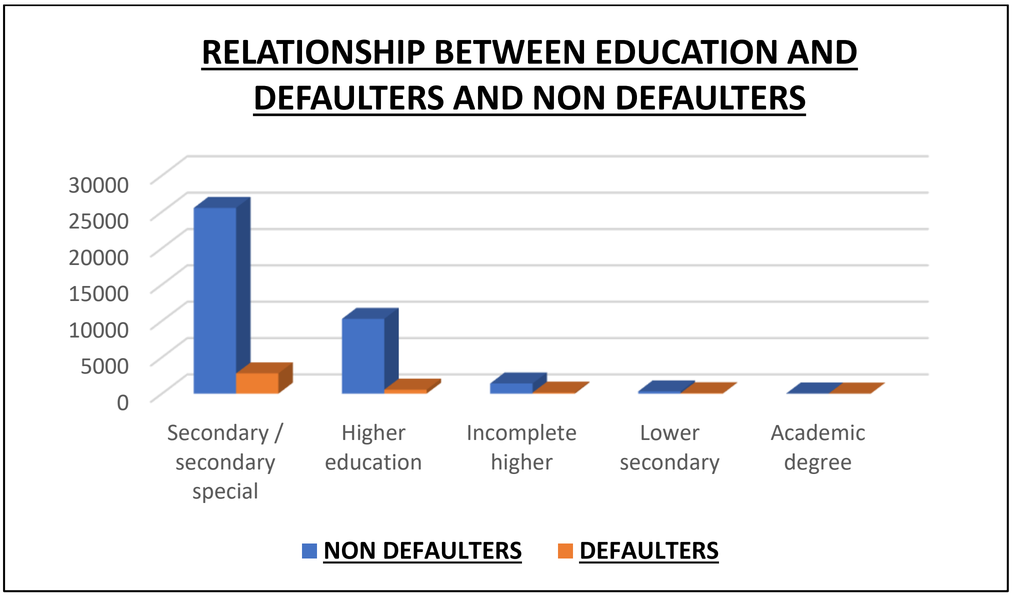
The highest level of education the client has had and then the loans provided to them.

EDUCATION	LOAN TAKEN	NON DEFAULTERS	DEFAULTERS
Secondary / secondary special	28322	25540	2782
Higher education	10831	10286	545
Incomplete higher	1532	1396	136
Lower secondary	368	310	58
Academic degree	17	17	0

#We can see that clients who have had education up to secondary level tend to take more loans, we can assume it's due to their plans to pursue higher education.

These clients also tend to be the biggest part of both defaulting and non-defaulting clientele.

Graphical representation:

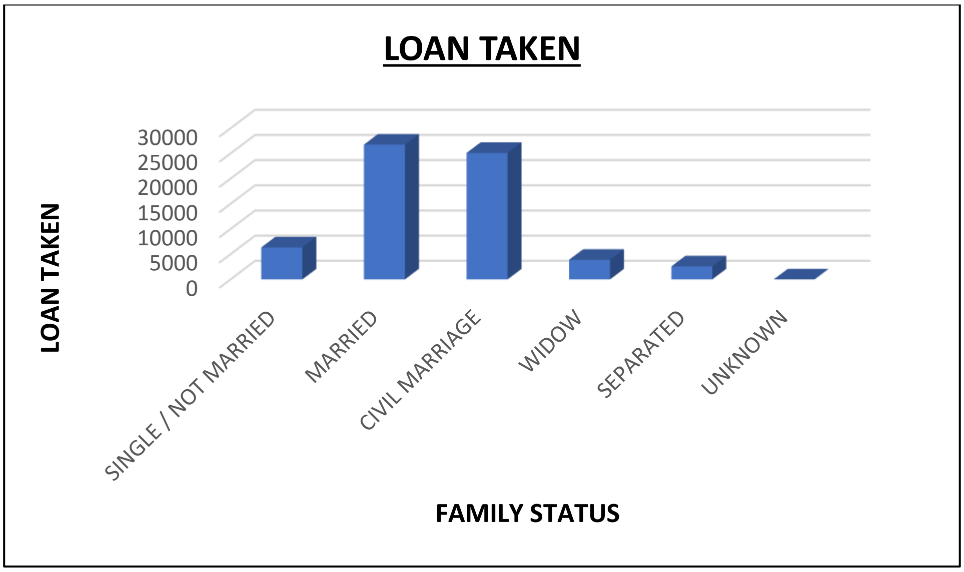


FAMILY STATUS OF THE CLIENT: NAME_FAMILY_STATUS

We analyse how family status (whether the client is single, married, widowed, etc) can affect payment difficulties.

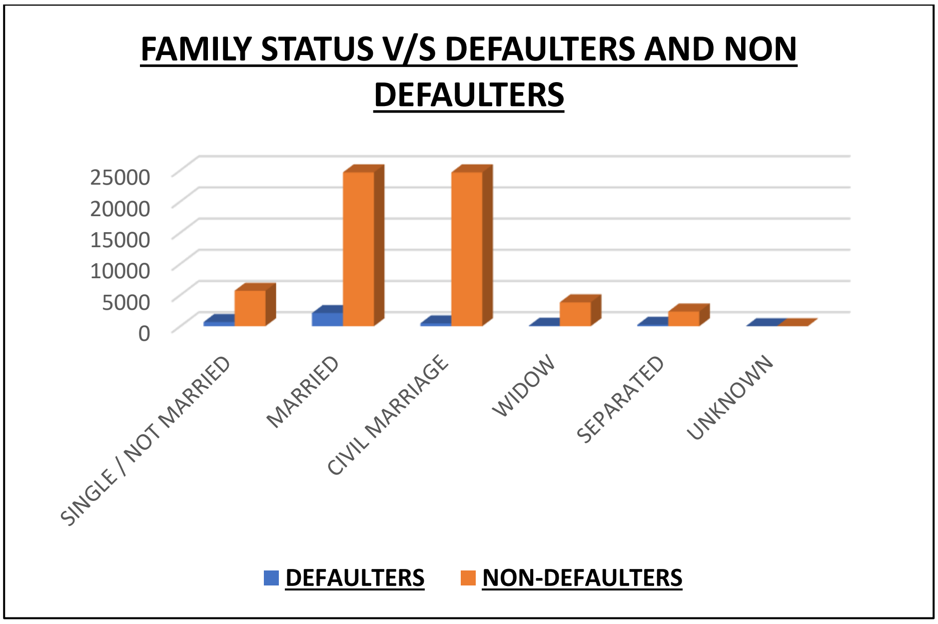
FAMILY STATUS	LOAN TAKEN	DEFAULTERS	NON-DEFAULTERS
SINGLE / NOT MARRIED	6352	673	5679
MARRIED	26759	2103	24656
CIVIL MARRIAGE	25109	453	24656
WIDOW	3891	67	3824
SEPARATED	2578	225	2353
UNKNOWN	1	0	1

We see that married clients take the highest amount of loans and they also happen to be the highest non-defaulters among the categories. Followed closely by clients in civil marriages.



#we see that most loans are taken by married people.

#here is a graphical representation of how many defaulters and nom defaulters may consist of these categories.



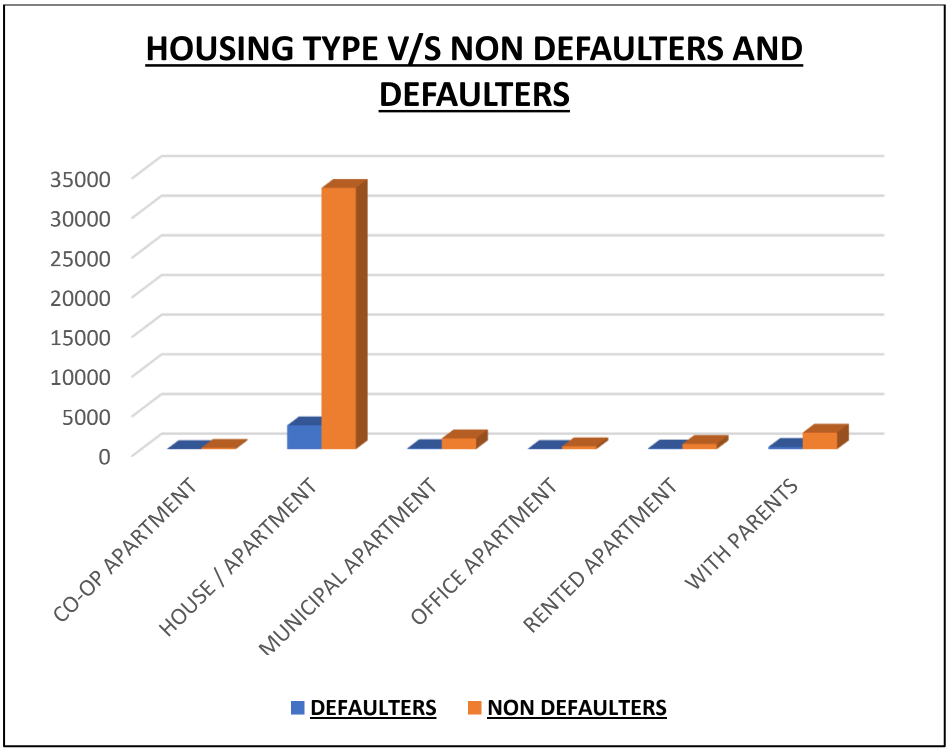
HOUSING SITUATION: NAME_HOUSING_TYPE

We analyse what is the housing situation of the client (renting, living with parents, etc)

We further go on to analyse the defaulters from these categories:

HOUSING TYPE	DEFAULTERS	NON DEFAULTERS	LOAN TAKEN
CO-OP APARTMENT	14	161	175
HOUSE / APARTMENT	2997	32937	35934
MUNICIPAL APARTMENT	122	1358	1480
OFFICE APARTMENT	27	356	383
RENTED APARTMENT	87	653	740
WITH PARENTS	274	2084	2358

We see that most clients live in a house/apartment and tend to take the highest amount of loans, the reason for that could be they may be getting EMIs on their houses and their apartments.



These clients also happen to be pretty good at paying their debts off.

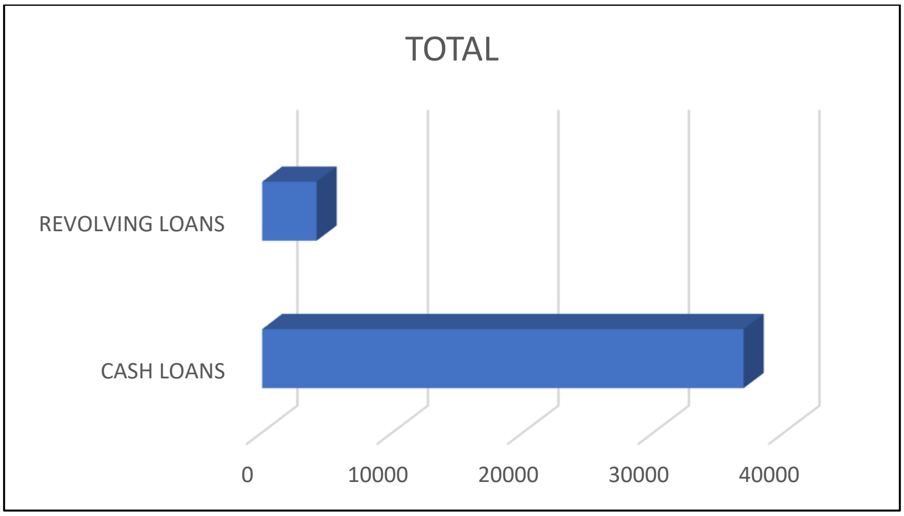
CONTRACT INFORMATION:

IDENTIFICATION IF THE LOAN IS CASH OR REVOLVING:

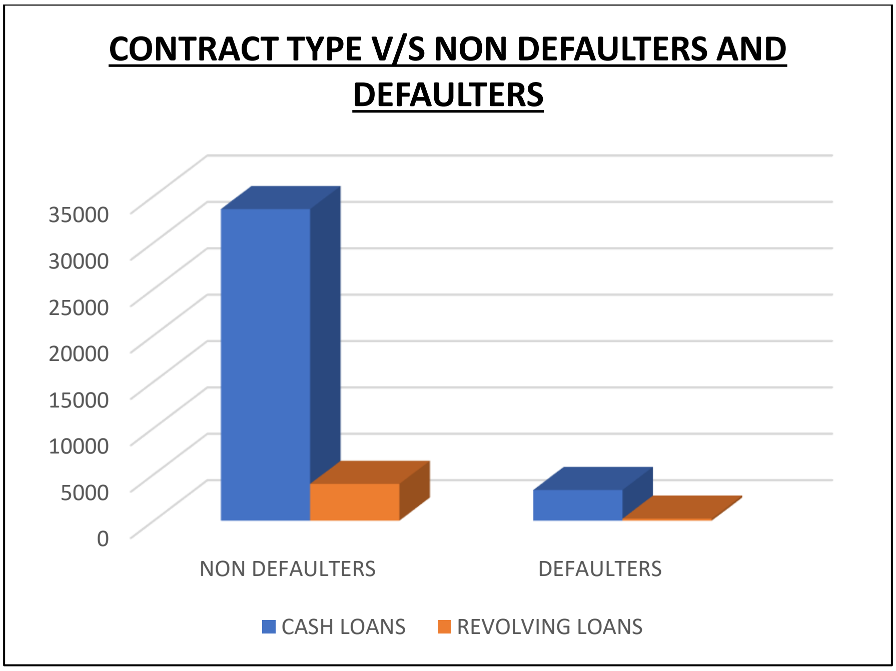
NAME_CONTRACT_TYPE

CONTRACT TYPE	NON DEFAULTERS	DEFAULTERS	TOTAL
CASH LOANS	33586	3307	36893
REVOLVING LOANS	3963	214	4177

As we discussed while calculating the imbalance in our data most loan applications are made for cash loans. Here is a graphical representation of the same:



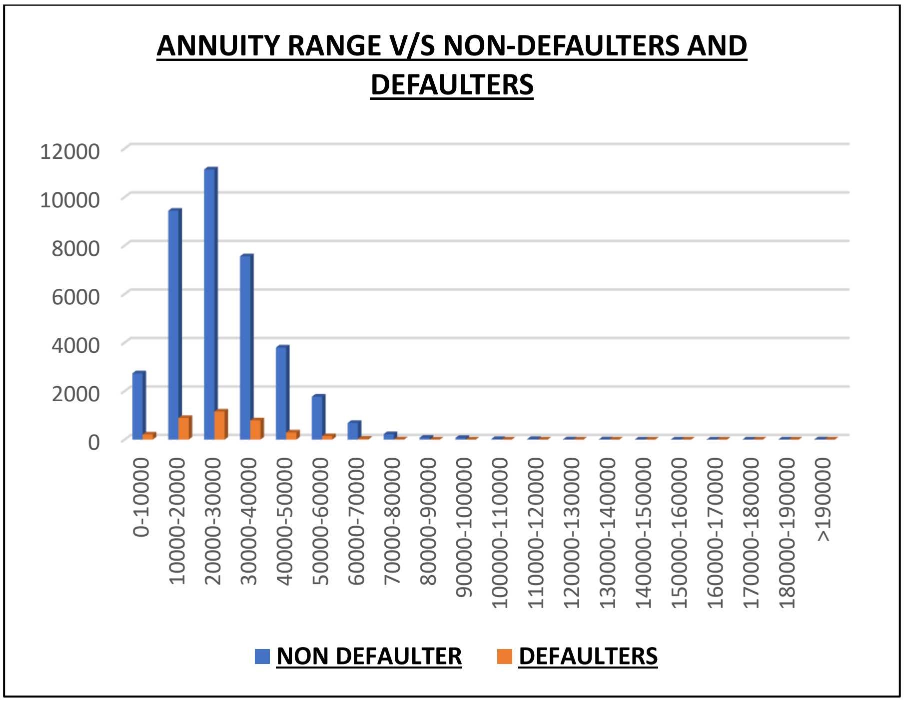
now we proceed to see the number of defaulters in each contract type.



LOAN ANNUITY AND DEFAULTERS: AMT_ANNUITY

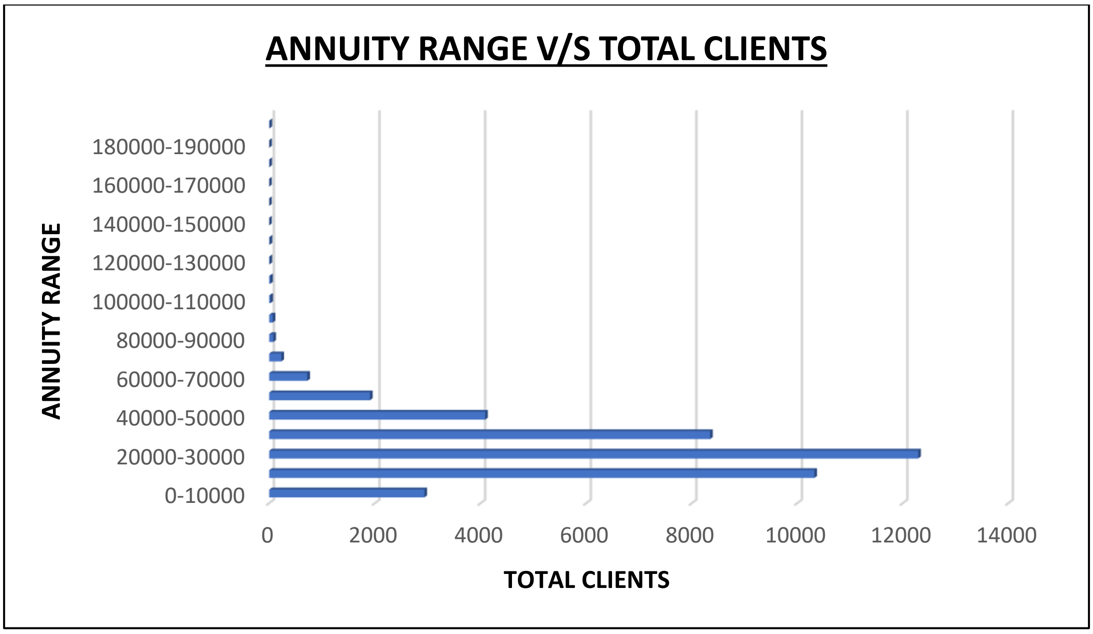
We figured out if there were any defaulters when it came to loan annuities or constant repayment instalments.

ANNUITY RANGE	NON DEFAULTER	DEFAULTERS	TOTAL
0-10000	2730	209	2939
10000-20000	9436	889	10325
20000-30000	11138	1156	12294
30000-40000	7559	790	8349
40000-50000	3797	291	4088
50000-60000	1769	139	1908
60000-70000	688	34	722
70000-80000	222	8	230
80000-90000	74	4	78
90000-100000	67	0	67
100000-110000	21	1	22
110000-120000	20	0	20
120000-130000	7	0	7
130000-140000	8	0	8
140000-150000	0	0	0
150000-160000	0	0	0
160000-170000	0	0	0
170000-180000	4	0	4
180000-190000	1	0	1
>190000	7	0	7



We see that most of the non-defaulters and defaulters come in the annuity range of 20,000-30,000.

#We further see that most of the loans were taken out in the 20k-30k range only :



CREDIT AMOUNT OF THE LOAN: AMT_CREDIT

We figured out how many loan applications were made for certain credit amounts. For that, we first set certain ranges and then use the COUNTIFS function to calculate all defaulters and non-defaulters in each respective range and finally calculate the total number of applications.

AMT_CREDIT	NON DEFAULTERS	DEFAULTERS	TOTAL
0-100000	648	47	695
100000-200000	3612	291	3903
200000-300000	6301	593	6894
300000-400000	3158	397	3555
400000-500000	3927	481	4408
500000-600000	4033	527	4560
600000-700000	2943	267	3210
700000-800000	2260	213	2473
800000-900000	1911	167	2078
900000-1000000	2074	135	2209
1000000-1100000	1807	147	1954
1100000-1200000	1143	76	1219
1200000-1300000	1214	65	1279
1300000-1400000	779	43	822
1400000-1500000	306	14	320
1500000-1600000	517	20	537
1600000-1700000	116	9	125
1700000-1800000	220	9	229
1800000-1900000	209	6	215
1900000-2000000	102	5	107
2000000-2100000	92	4	96
2100000-2200000	26	2	28
2200000-2300000	79	0	79
>2300000	71	3	74

CLIENT'S EMPLOYMENT: EMPLOYED_YEARS

First, we convert the days before the application the person started their current employment into years.

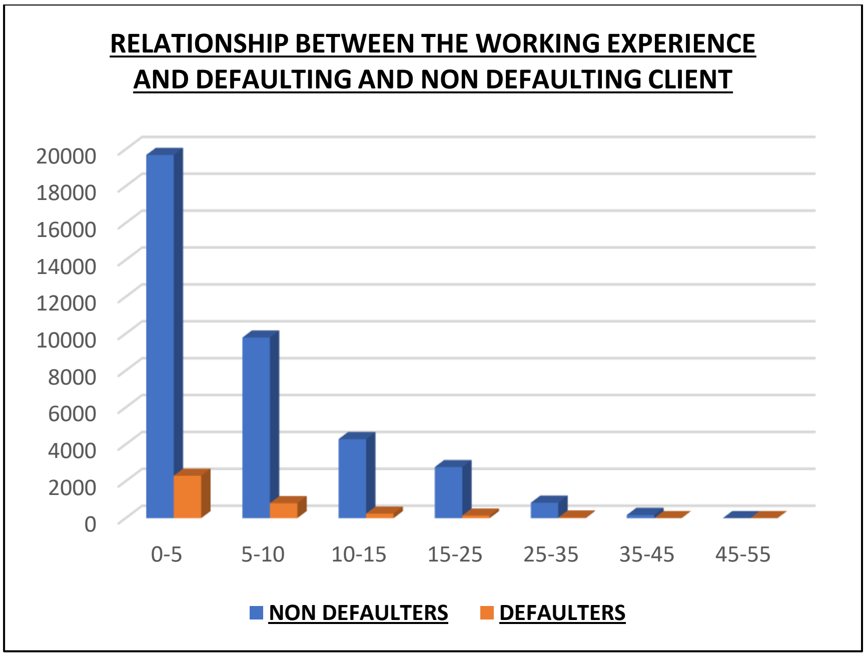
And then further analyse how that affects the loan application process.

We then proceed to divide these years into the following categories:

- 0-5
- 5-10
- 10-15
- 15-25
- 25-35
- 35-45

We find defaulters and non-defaulters in these categories:

EMPLOYED_YEARS	LOAN TAKEN	NON DEFAULTERS	DEFAULTERS
0-5	21990	19693	2297
5-10	10604	9793	811
10-15	4525	4284	241
15-25	2906	2767	139
25-35	867	836	31
35-45	176	174	2
45-55	2	2	0

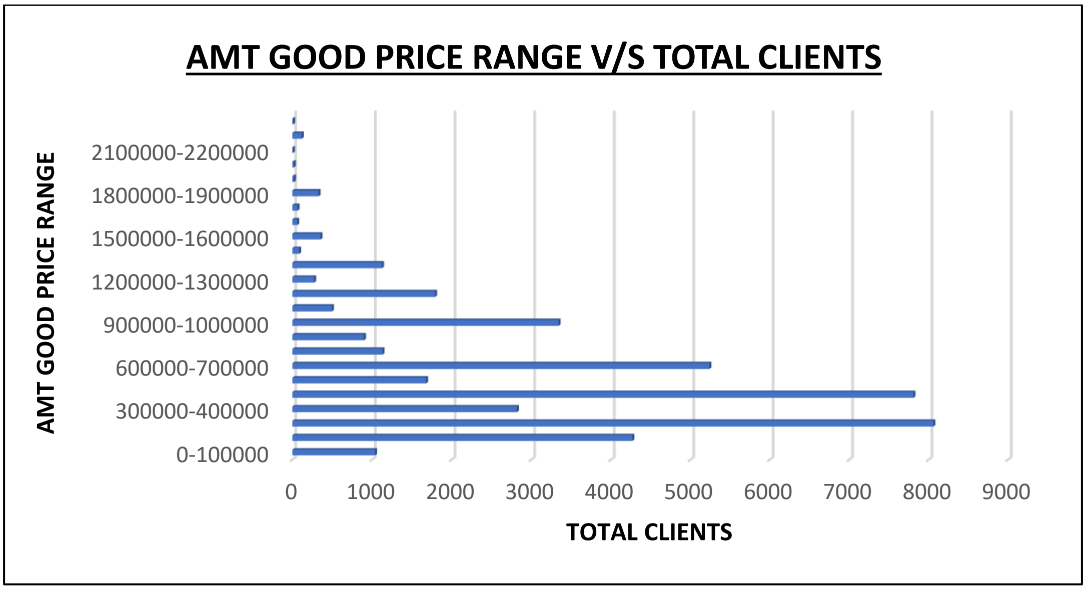


Most loans are taken by people in the initial stage of their employment and they also happen to be good about repayment of their loans.

<u>FOR CONSUMER LOANS IT IS THE PRICE OF GOODS FOR WHICH THE LOAN IS GIVEN:</u> AMT_GOODS_PRICE

AMT GOOD PRICE RANGE	NON DEFAULTERS	DEFAULTERS	TOTAL
0-100000	971	68	1039
100000-200000	3923	348	4271
200000-300000	7295	766	8061
300000-400000	2500	321	2821
400000-500000	6927	883	7810
500000-600000	1558	122	1680
600000-700000	4809	432	5241
700000-800000	1054	79	1133
800000-900000	836	64	900
900000-1000000	3134	211	3345
1000000-1100000	469	29	498
1100000-1200000	1697	96	1793
1200000-1300000	259	15	274
1300000-1400000	1083	44	1127
1400000-1500000	83	3	86
1500000-1600000	334	21	355
1600000-1700000	62	0	62
1700000-1800000	65	2	67
1800000-1900000	318	12	330
1900000-2000000	18	3	21
2000000-2100000	19	0	19
2100000-2200000	7	0	7
2200000-2300000	118	1	119
>2300000	9	1	10

#To analyse this, we first set certain ranges and proceed to calculate defaulters and non-defaulters to get the total count of applications made for a certain price range.



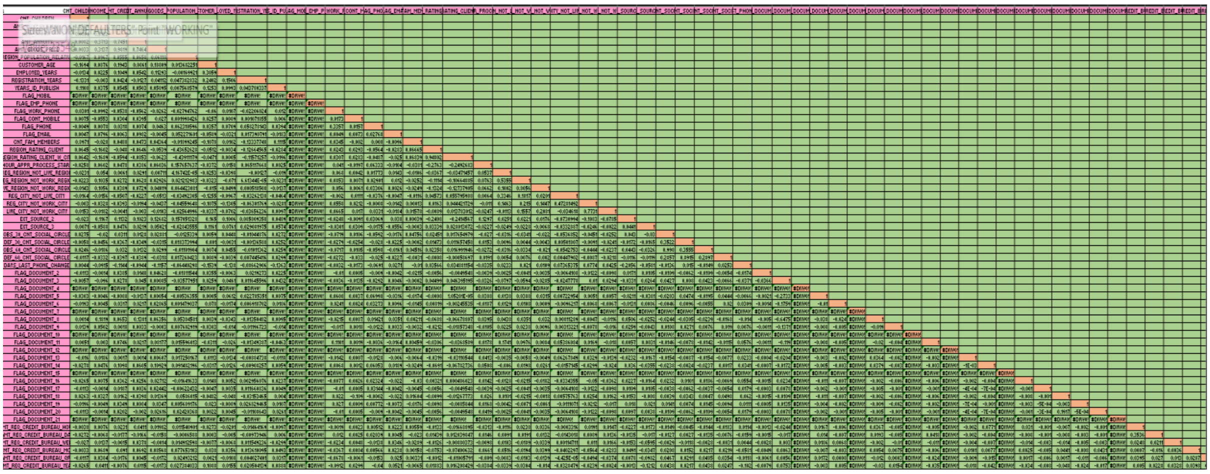
#most loans are taken out for goods lying in the 3,00,000-4,00,000 price range.

E.) IDENTIFY TOP CORRELATIONS FOR DIFFERENT SCENARIOS:

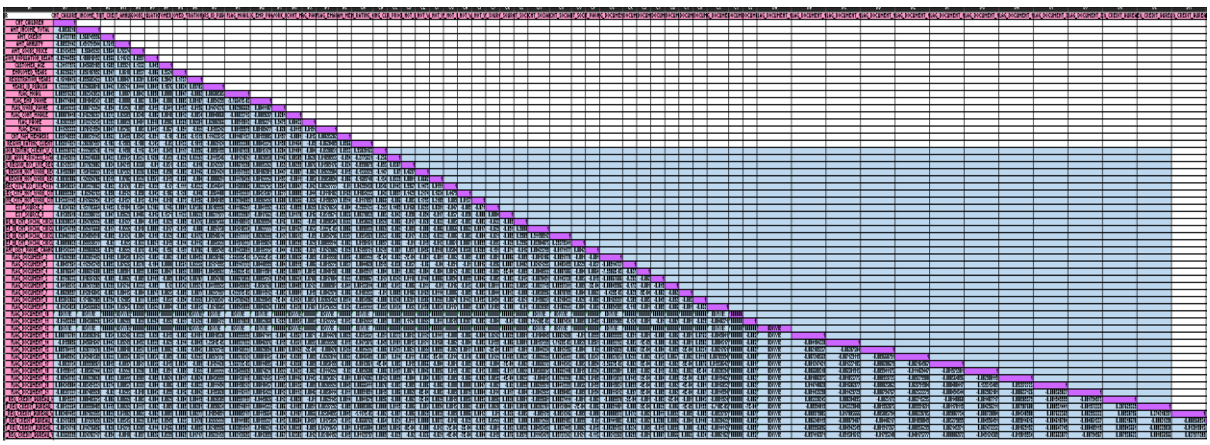
Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

#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.

CORRELATION AND DEFAULTERS



CORRELATION AND NON DEFAULTERS



TOP CORRELATIONS

- DEFAULTERS:

	DEFAULTERS		
RANK	VARIABLE 1	VARIABLE 2	CORRELATION
1	AMT_GOODS_PRICE	AMT_CREDIT	0.981928143
2	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.948020808
3	CNT_FAM_MEMBERS	CNT_CHILDREN	0.895600339
4	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.891467244
5	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.805583225
6	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.773107352
7	AMT_ANNUITY	AMT_GOODS_PRICE	0.746422447
8	AMT_ANNUITY	AMT_CREDIT	0.745132112

- NON-DEFAULTERS:

	NON-DEFAULTERS		
RANK	VARIABLE 1	VARIABLE 2	CORRELATION
1	AMT_GOODS_PRICE	AMT_CREDIT	0.98635817
2	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950286525
3	CNT_CHILDREN	CNT_FAM_MEMBERS	0.893735596
4	REG_REGION_NOT_WORK_REGION -	LIVE_REGION_NOT_WORK_REGION	0.860167703
5	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.853040752
6	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.815604978
7	REGION_RATING_CLIENT	AMT_GOODS_PRICE	0.765201743
8	AMT_ANNUITY	AMT_GOODS_PRICE	0.765201743
9	AMT_CREDIT	AMT_ANNUITY	0.760827873

Here's, an understanding of the relationships between different variables in the dataset. Strong correlations (close to 1 or -1) indicate a strong linear relationship, while weak correlations (close to 0) indicate a weak or no linear relationship. These insights can be used to further analyse the dataset and understand the underlying patterns or dependencies between variables.

CONCLUSION

This project is very useful and knowledgeable for deep learning of Excel. In this project, I have learnt many new concepts like finding outliners with the help of inter-quartile function and how to make a correlation heat map and matrix, gaining knowledge from it. I learned more about how to analyse and gain insights from graphs.

This project gave me a good hand in EDA analysis. I had performed all the tasks that I was supposed to do.
I had developed a good understanding of the domain, and a little about risk analytics – understanding the types of variables and their significance should be enough).

-END-

HYPERLINK TO EXCEL FILE:
https://docs.google.com/spreadsheets/d/1UKAvtKpiy7gcPKdzEY2QPgoSMcxG_e3S/edit?usp=sharing&ouid=102683227032029211056&rtpof=true&sd=true

HYPERLINK TO PPT:
https://drive.google.com/file/d/1_vb782OOyr59yHt5THl8K-uXHlbrUsRn/view?usp=sharing

HYPERLINK TO VIDEO SUBMISSION:
https://drive.google.com/file/d/1NJGVbqEDrNl2CHO3_J4kyfrxuzJtAhy_/view?usp=sharing