Bank Loan Case Study

Project Description:

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. 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 late payment 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.

Approach:

First, understand the given dataset of the clients. Then clean the dataset finding the blanks and missing values and putting missing value using appropriate method (Mean, Median, and Mode), find the outliers, identify the imbalance in data set using Excel and plot graph using Pivot table.

Tech-Stack Used:

MS EXCEL 2016

Insights:

Some of my findings are:

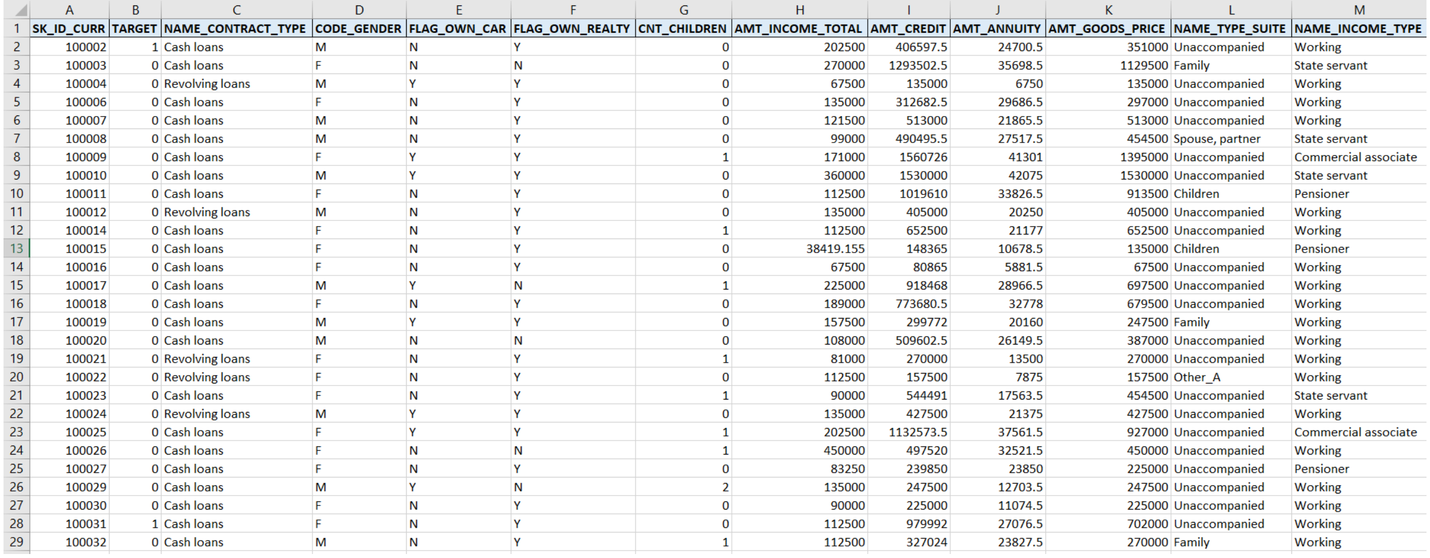
- People with medium income group are more likely to be defaulter's followed by low income groups.
- Around 52% of the total clients are working professionals. Among them only a small percentage of people comes under the category of defaulters (5%).

Result:

1) **Present the overall approach of the analysis. Mention the problem statement and the analysis approach briefly**
   This project focuses on applying EDA in a real business scenario. Apart from applying the techniques that are in the EDA module, we will also develop a basic understanding of risk

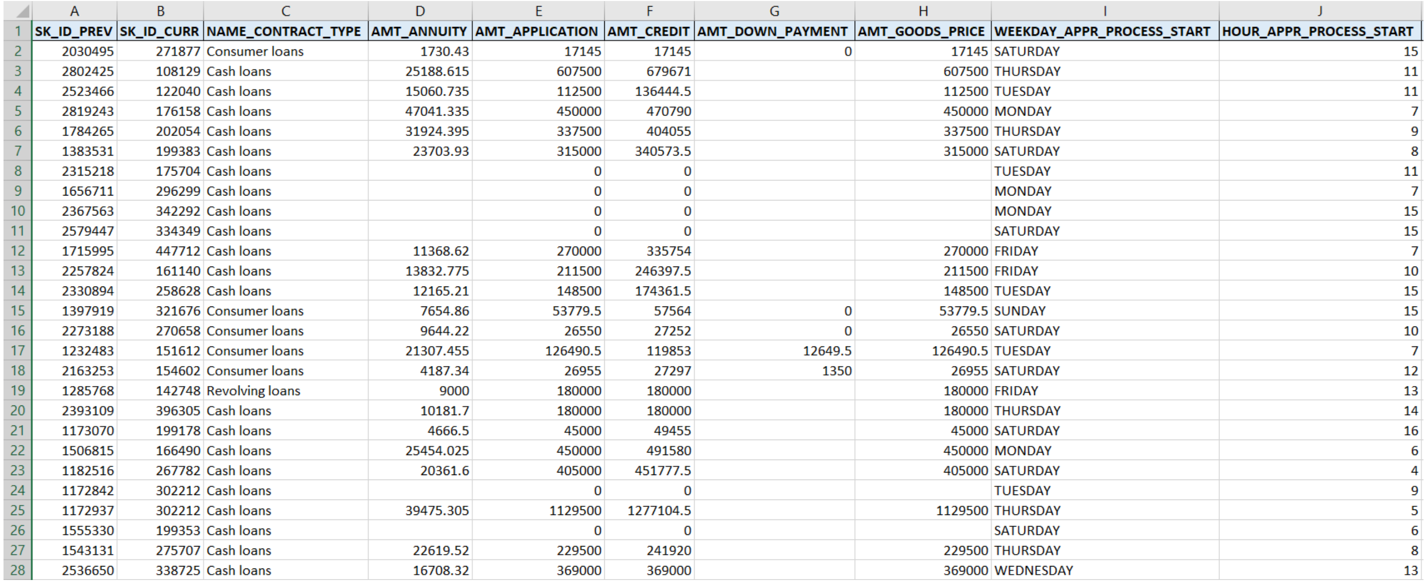
analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Importing Dataset



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE	NAME_INCOME_TYPE
2	100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied	Working
3	100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servant
4	100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied	Working
5	100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied	Working
6	100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied	Working
7	100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner	State servant
8	100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied	Commercial associate
9	100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied	State servant
10	100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner
11	100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied	Working
12	100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied	Working
13	100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children	Pensioner
14	100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied	Working
15	100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied	Working
16	100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied	Working
17	100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family	Working
18	100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied	Working
19	100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied	Working
20	100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A	Working
21	100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied	State servant
22	100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied	Working
23	100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied	Commercial associate
24	100026	0	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied	Working
25	100027	0	Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccompanied	Pensioner
26	100029	0	Cash loans	M	Y	N	2	135000	247500	12703.5	247500	Unaccompanied	Working
27	100030	0	Cash loans	F	N	Y	0	90000	225000	11074.5	225000	Unaccompanied	Working
28	100031	1	Cash loans	F	N	Y	0	112500	979992	27076.5	702000	Unaccompanied	Working
29	100032	0	Cash loans	M	N	Y	1	112500	327024	23827.5	270000	Family	Working

Fig: Application_data



	A	B	C	D	E	F	G	H	I	J
1	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_APPR_PROCESS_START	HOUR_APPR_PROCESS_START
2	2030495	271877	Consumer loans	1730.43	17145	17145	0	17145	SATURDAY	15
3	2802425	108129	Cash loans	25188.615	607500	679671		607500	THURSDAY	11
4	2523466	122040	Cash loans	15060.735	112500	136444.5		112500	TUESDAY	11
5	2819243	176158	Cash loans	47041.335	450000	470790		450000	MONDAY	7
6	1784265	202054	Cash loans	31924.395	337500	404055		337500	THURSDAY	9
7	1383531	199383	Cash loans	23703.93	315000	340573.5		315000	SATURDAY	8
8	2315218	175704	Cash loans		0	0			TUESDAY	11
9	1656711	296299	Cash loans		0	0			MONDAY	7
10	2367563	342292	Cash loans		0	0			MONDAY	15
11	2579447	334349	Cash loans		0	0			SATURDAY	15
12	1715995	447712	Cash loans	11368.62	270000	335754		270000	FRIDAY	7
13	2257824	161140	Cash loans	13832.775	211500	246397.5		211500	FRIDAY	10
14	2330894	258628	Cash loans	12165.21	148500	174361.5		148500	TUESDAY	15
15	1397919	321676	Consumer loans	7654.86	53779.5	57564	0	53779.5	SUNDAY	15
16	2273188	270658	Consumer loans	9644.22	26550	27252	0	26550	SATURDAY	10
17	1232483	151612	Consumer loans	21307.455	126490.5	119853	12649.5	126490.5	TUESDAY	7
18	2163253	154602	Consumer loans	4187.34	26955	27297	1350	26955	SATURDAY	12
19	1285768	142748	Revolving loans	9000	180000	180000		180000	FRIDAY	13
20	2393109	396305	Cash loans	10181.7	180000	180000		180000	THURSDAY	14
21	1173070	199178	Cash loans	4666.5	45000	49455		45000	SATURDAY	16
22	1506815	166490	Cash loans	25454.025	450000	491580		450000	MONDAY	6
23	1182516	267782	Cash loans	20361.6	405000	451777.5		405000	SATURDAY	4
24	1172842	302212	Cash loans		0	0			TUESDAY	9
25	1172937	302212	Cash loans	39475.305	1129500	1277104.5		1129500	THURSDAY	5
26	1555330	199353	Cash loans		0	0			SATURDAY	6
27	1543131	275707	Cash loans	22619.52	229500	241920		229500	THURSDAY	8
28	2536650	338725	Cash loans	16708.32	369000	369000		369000	WEDNESDAY	13

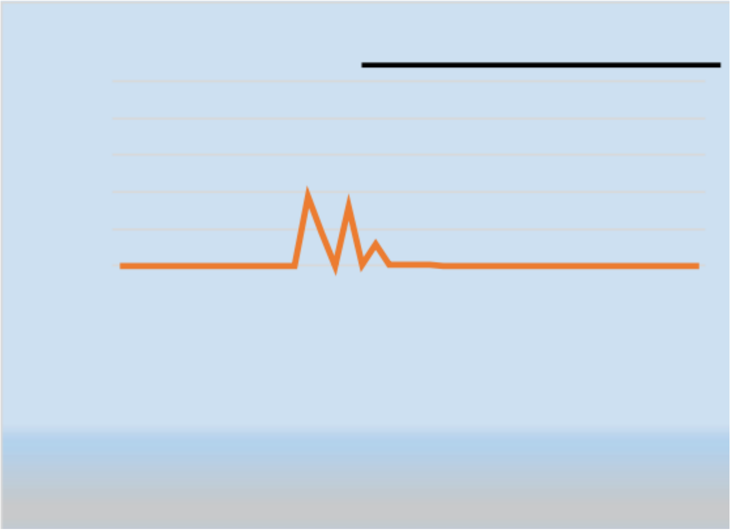
Fig: Previous_Application

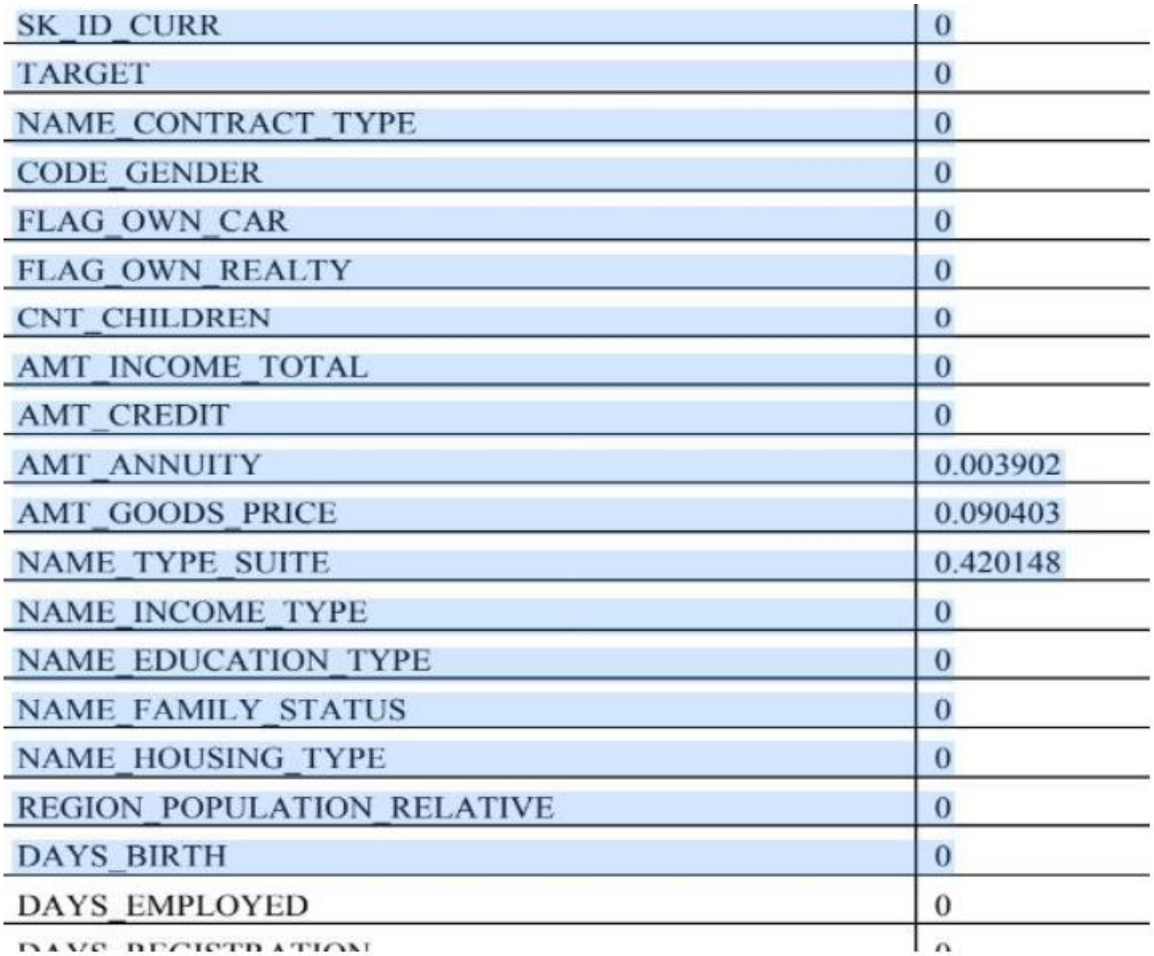
2) **Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)**

First, import the dataset into Excel. I understand that there are many missing values in the dataset. So count the missing value and find the percentage of missing value. If the percentage is greater than 50 percentage we drop that column.

Excel formula to find

=COUNTBLANK(DR2:DR307512)*100/(COUNT(DR2:DR307512)+COUNTBLANK(DR2:DR307512))



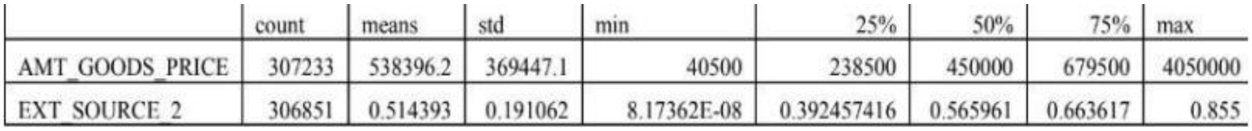


SK_ID_CURR	0
TARGET	0
NAME_CONTRACT_TYPE	0
CODE_GENDER	0
FLAG_OWN_CAR	0
FLAG_OWN_REALTY	0
CNT_CHILDREN	0
AMT_INCOME_TOTAL	0
AMT_CREDIT	0
AMT_ANNUITY	0.003902
AMT_GOODS_PRICE	0.090403
NAME_TYPE_SUITE	0.420148
NAME_INCOME_TYPE	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATUS	0
NAME_HOUSING_TYPE	0
REGION_POPULATION_RELATIVE	0
DAYS_BIRTH	0
DAYS_EMPLOYED	0
DAYS_REGISTRATION	0

We have inputted the below columns with mean and mode

- AMT_GOODS_PRICE (integer)
- NAME_TYPE_SUITE(Object)
- EXT_SOURCE_2(integer)

As shown in the below diagram, finding the mean and median for both AMT_GOODS_PROCE and EXT_SOURCE_2 and based on percentile variation between 25th and 75th, let's input the missing values by mean values of AMT_GOODS_PROCE and EXT_SOURCE_2 respectively.



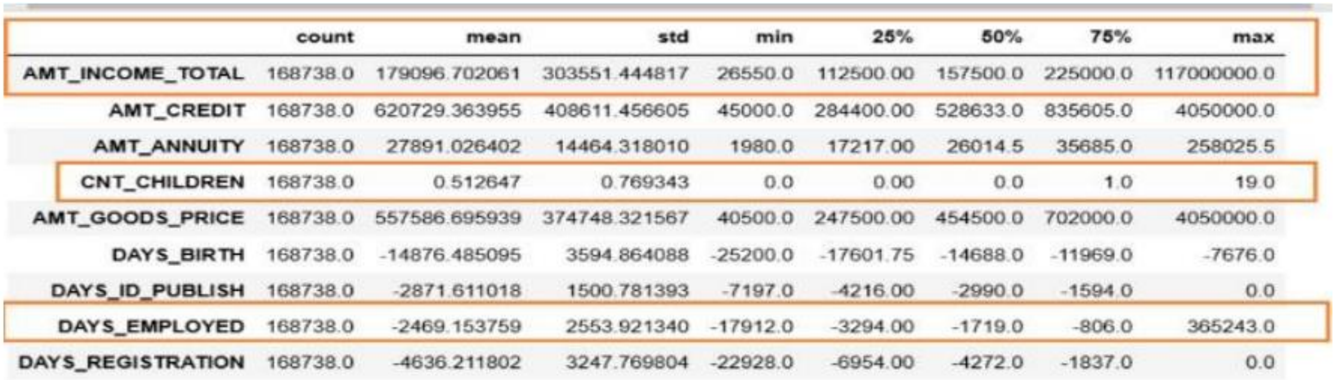
	count	means	std	min	25%	50%	75%	max
AMT_GOODS_PRICE	307233	538396.2	369447.1	40500	238500	450000	679500	4050000
EXT_SOURCE_2	306851	0.514393	0.191062	8.17362E-08	0.392457416	0.565961	0.663617	0.855

This is how we can deal with our dataset, we have to find whether the value which is missing has any impact on the dataset, if it doesn't, and then we can remove it. If it does impact then we have to find the type of variable it is, if its categorical data then mode, if not then we can use mean or median.

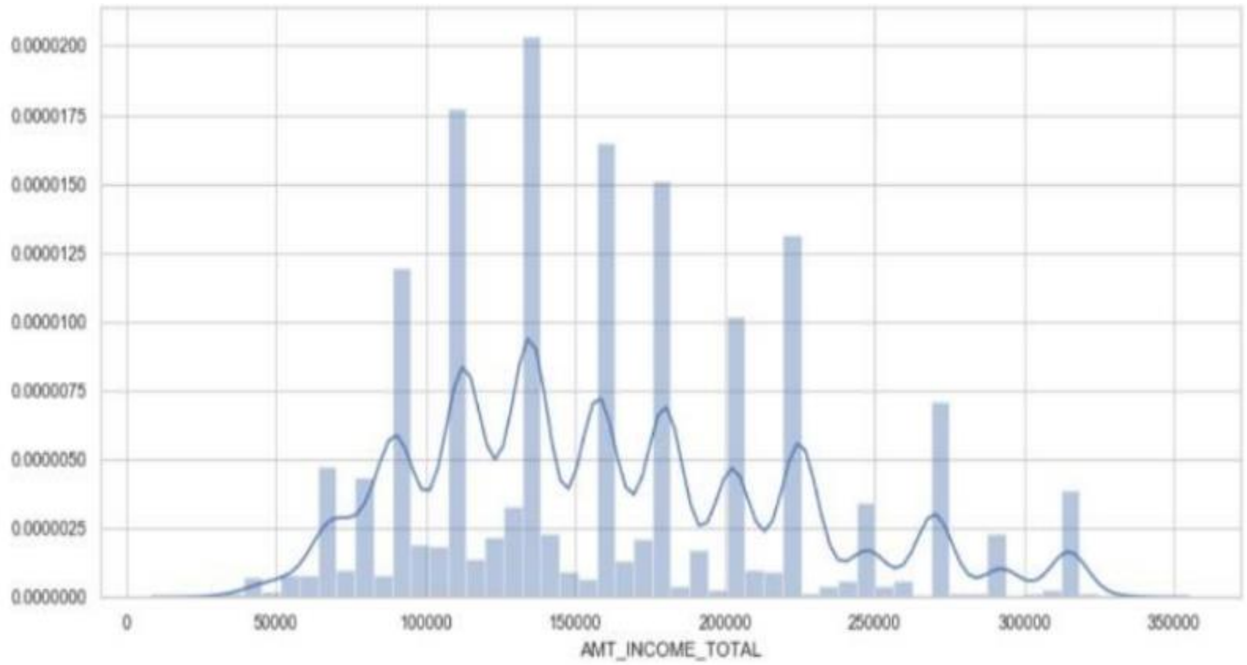
3) **Identify if there are outliers in the dataset. Also, mention why you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.**

Outliers are the data points that shows some deviations from other data points. In the given datasets there are some columns have outliers, some of these have no outliers. We have selected some random data frames, checking the Data frame for outlier values and analyze them

- Checking the outliers for columns and understanding the reason to mention that as an outlier.
- Here in our analysis to find out the outliers, we have considered few numerical columns and analyzed the statistics of them.
- If we observe the below screenshot, there are 3 columns with outlier values which are having a huge difference compared to the regular intervals of other values.

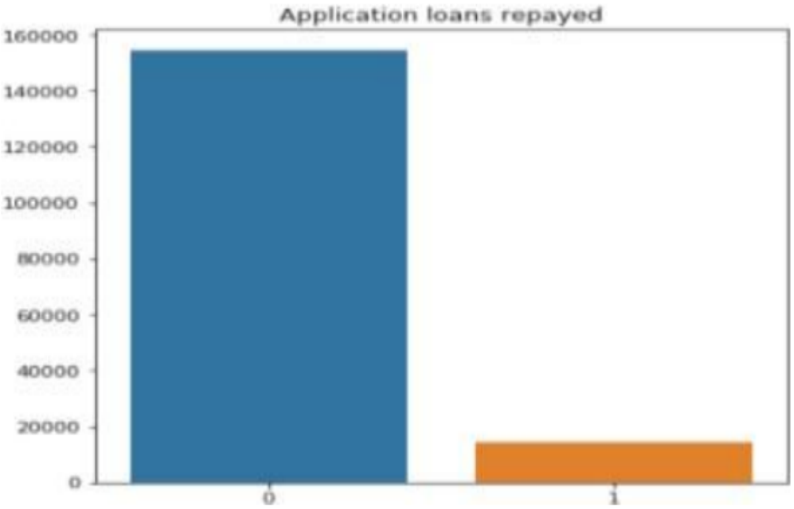


	count	mean	std	min	25%	50%	75%	max
AMT_INCOME_TOTAL	168738.0	179096.702061	303551.444817	26550.0	112500.00	157500.0	225000.0	117000000.0
AMT_CREDIT	168738.0	620729.363955	408611.456605	45000.0	284400.00	528633.0	835605.0	4050000.0
AMT_ANNUITY	168738.0	27891.026402	14464.318010	1980.0	17217.00	26014.5	35685.0	258025.5
CNT_CHILDREN	168738.0	0.512647	0.769343	0.0	0.00	0.0	1.0	19.0
AMT_GOODS_PRICE	168738.0	557586.695939	374748.321567	40500.0	247500.00	454500.0	702000.0	4050000.0
DAYS_BIRTH	168738.0	-14876.485095	3594.864088	-25200.0	-17601.75	-14688.0	-11969.0	-7676.0
DAYS_ID_PUBLISH	168738.0	-2871.611018	1500.781393	-7197.0	-4216.00	-2990.0	-1594.0	0.0
DAYS_EMPLOYED	168738.0	-2469.153759	2553.921340	-17912.0	-3294.00	-1719.0	-806.0	365243.0
DAYS_REGISTRATION	168738.0	-4636.211802	3247.769804	-22928.0	-6954.00	-4272.0	-1837.0	0.0

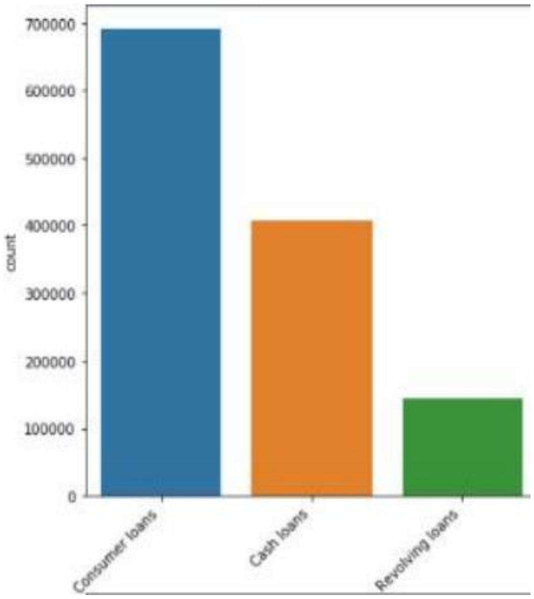


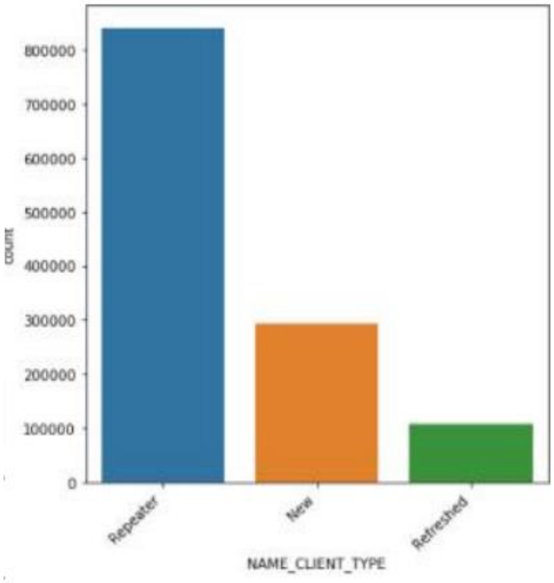
Income ranges from 25k to 300k.There are few spikes in between, this is the plot we get after removing outliers.

4) **Identify if there is data imbalance in the data. Find the ratio of data imbalance.**
We See a the imbalance between target type 1 and 0.Ratio is of 91.5: 8.45



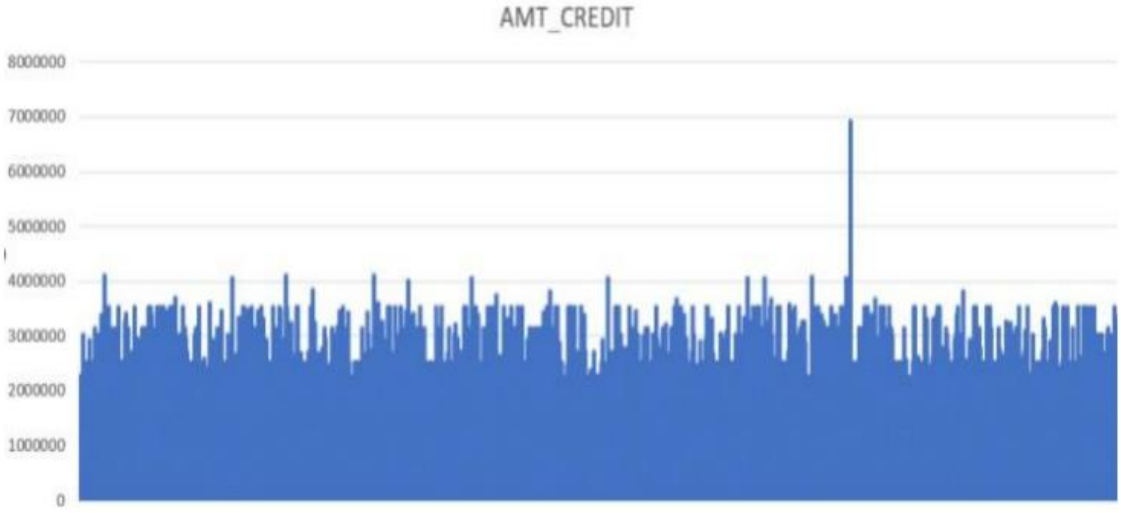
We divide the dataset into two subsets based on Target variable. I.e. Target=0 and Target=1.

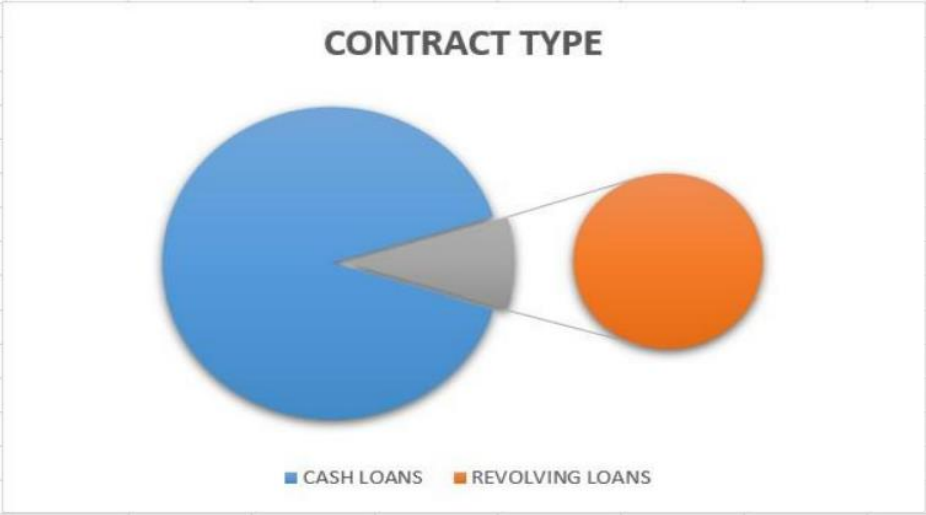




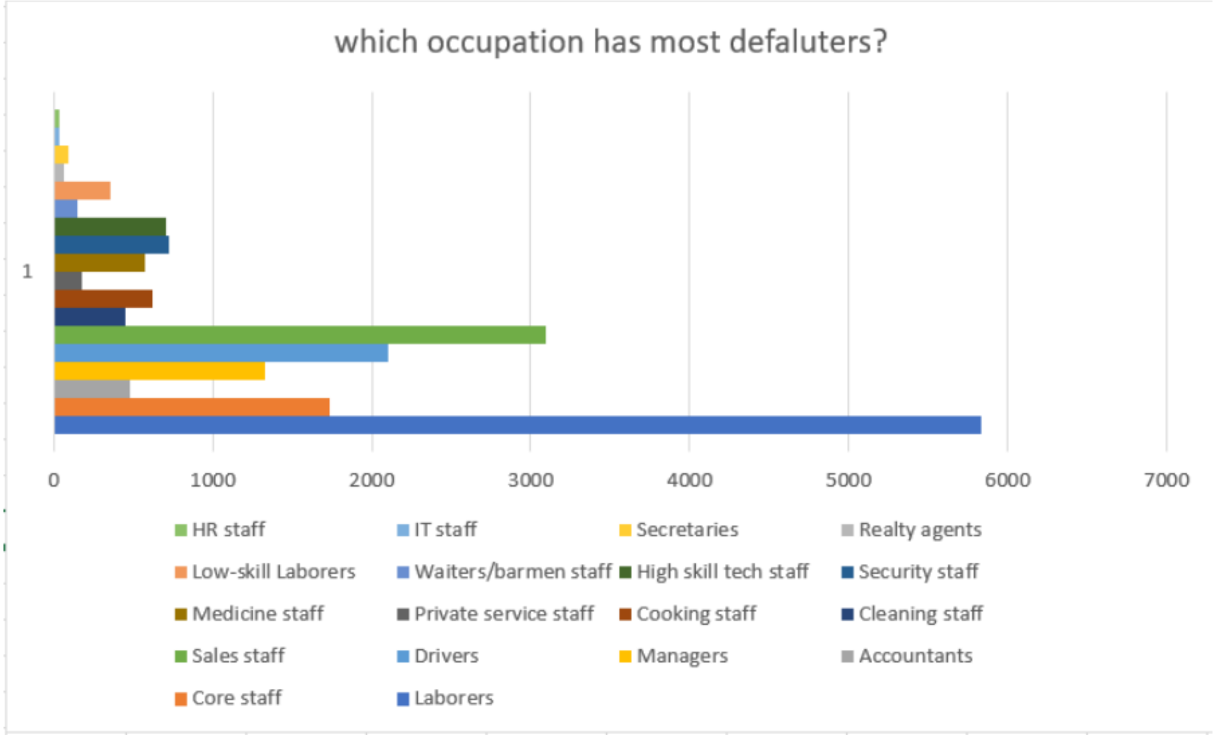
5) **Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.**

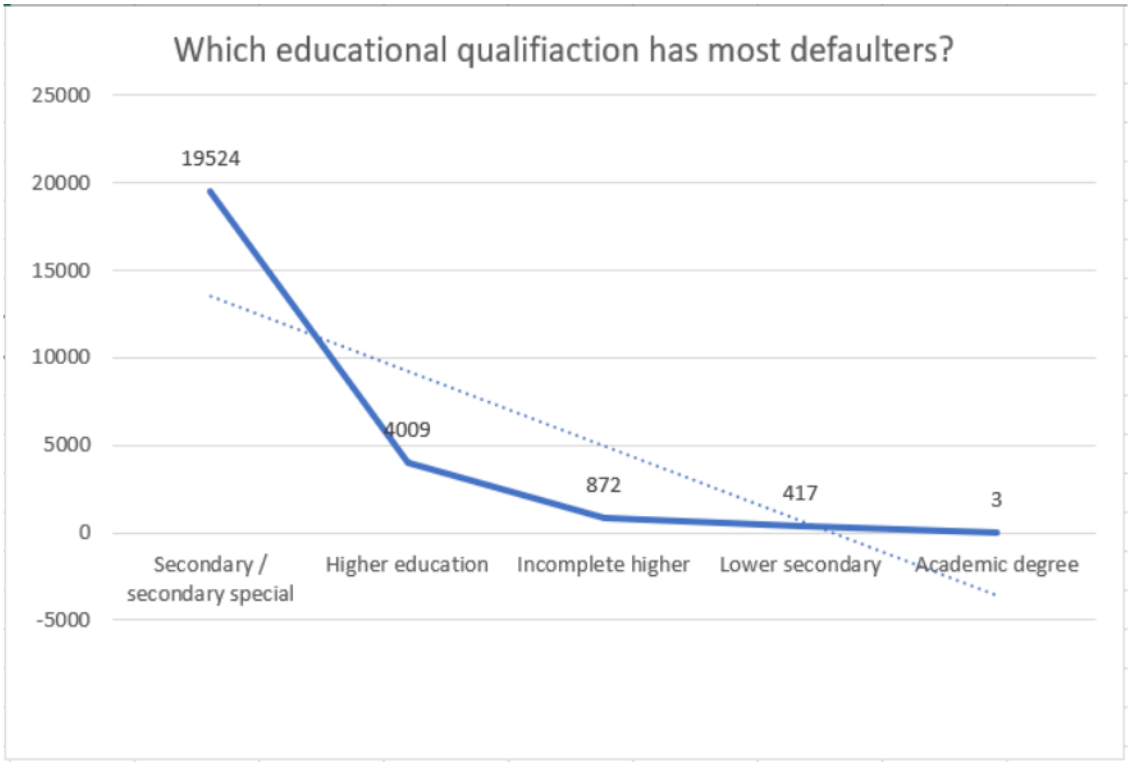
The results of univariate, segmented and bivariate analysis using visual means are as follows:



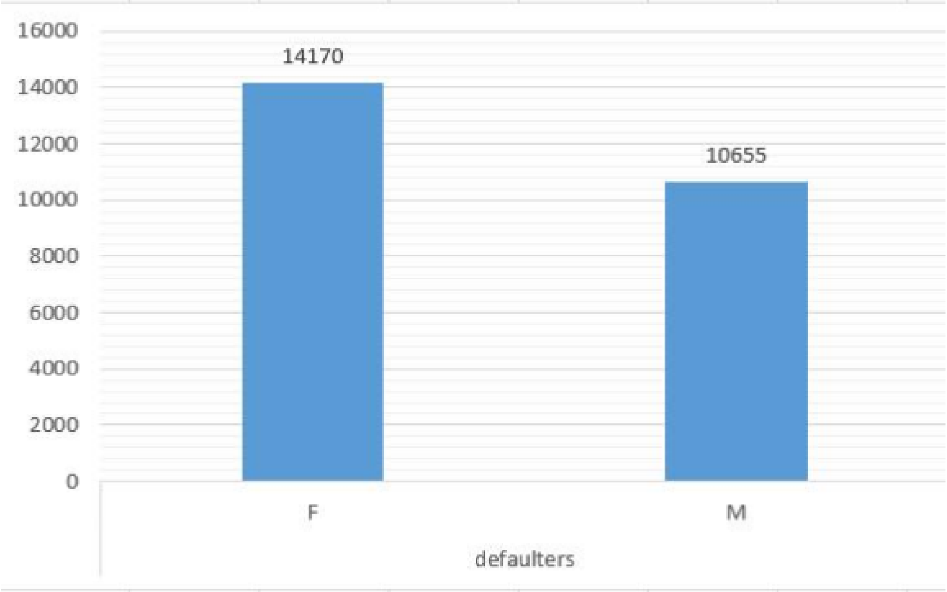


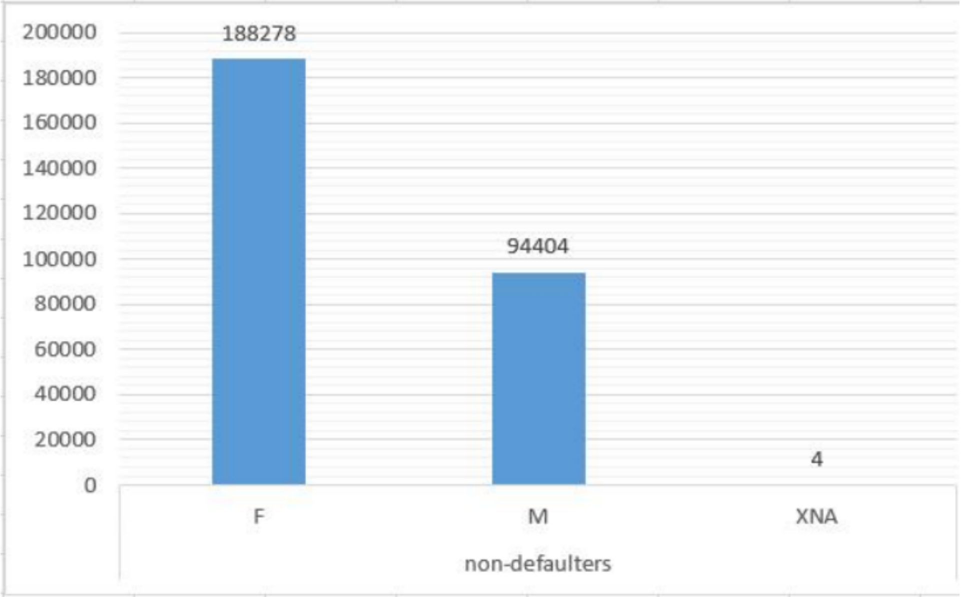
6) **Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, and Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, and Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.**



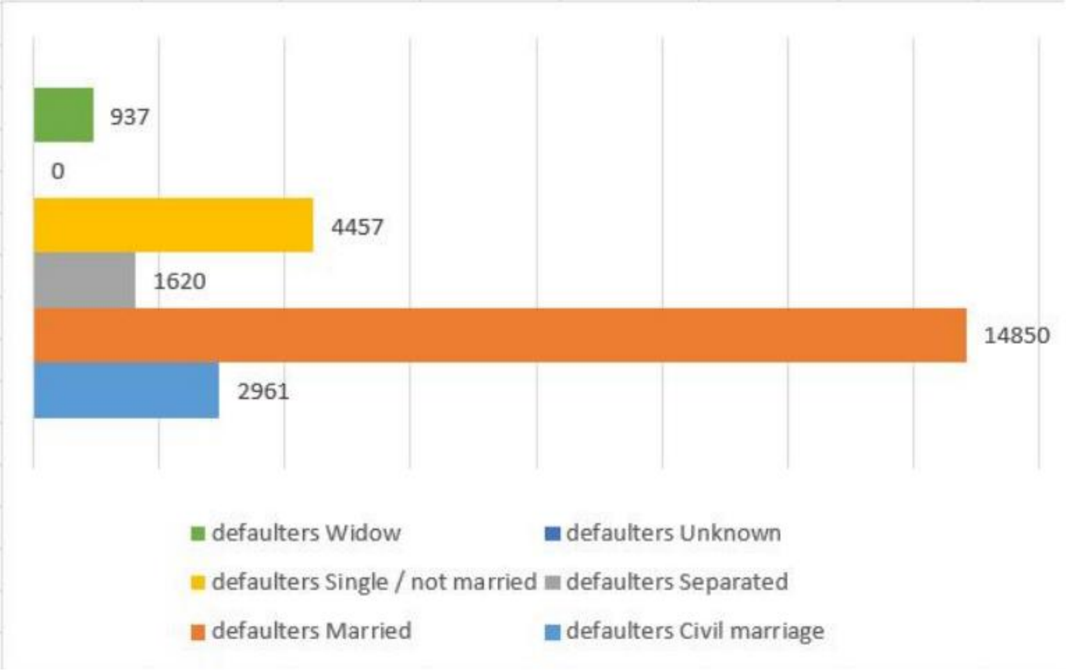


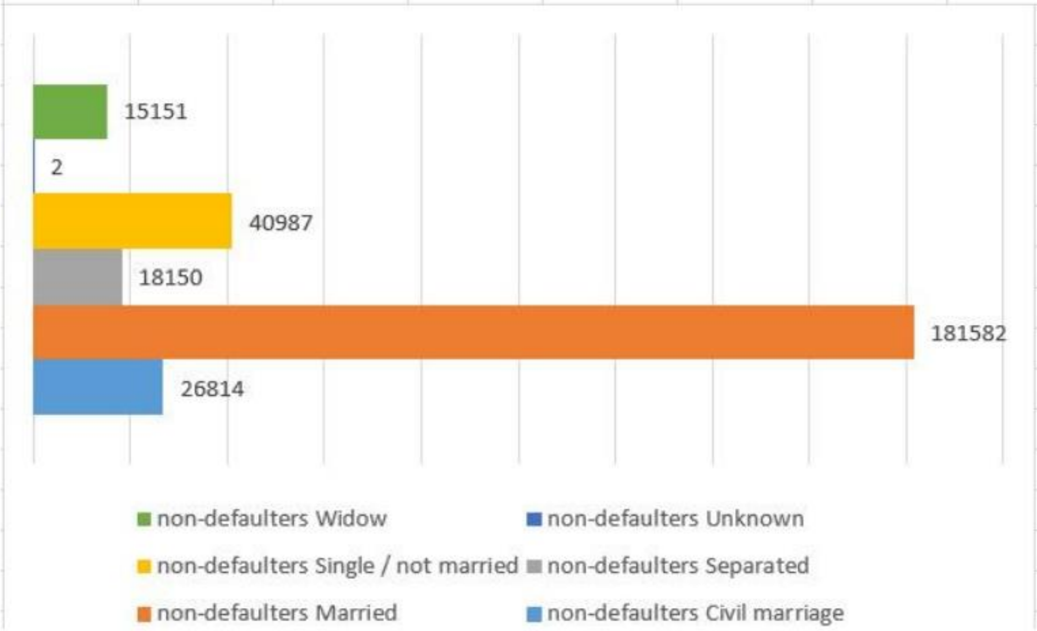
- **Univariate analysis (defaulters &non-defaulters)**
  Based on gender



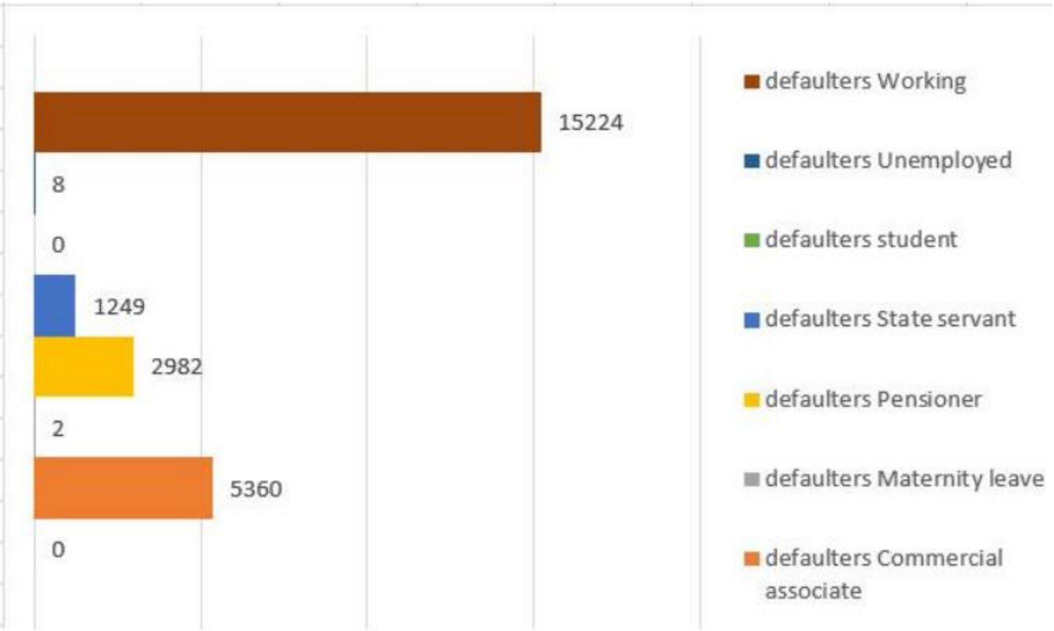


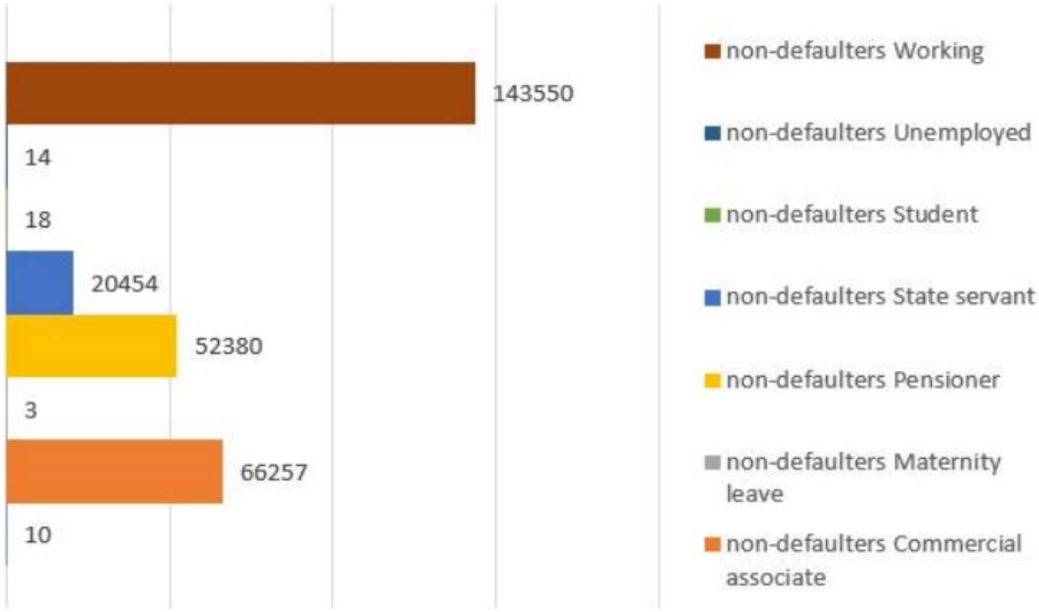
Based on family status





<u>Based on income type</u>





7) **Conclusions:**

After performing the analysis, we can rectify whether a client will repay the loan or not. Also, the people who are likely to face problem in loan repayment are laborers .Also people with Secondary /secondary special education might face problem in loan repayment. Moreover, those who are living in House/apartment are facing difficulty in loan repayment (may be because of extra home loan, EMIs and so on).people opting for cash loan faces difficulty in doing the same.

Dataset: https://drive.google.com/file/d/16dk4_PY9Fxx82f1K2-fyJhwUmsR56yDH/view?usp=sharing

Thank You