Bank Loan Case Study

Project Description:

- This project 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, 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.
- When a client applies for a loan, there are 4 scenarios:
    1. Approved
    2. Cancelled
    3. Refused
    4. Unused offer

Approach:

- The data is quite huge with around 30lakhs rows and 122 columns in application_data which is the main file. So we need to check for the missing data and outliers in the dataset. This csv file contains all the information of the client at the time of application. The data is regarding if the client has difficulty in paying the loan.
- The second dataset is previous_application which contains information about the client's previous loan data. It tells us whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- Columns_description.csv describes the meaning of each variable.

A. EXPLORATORY DATA ANALYSIS FOR APPLICATION_DATA

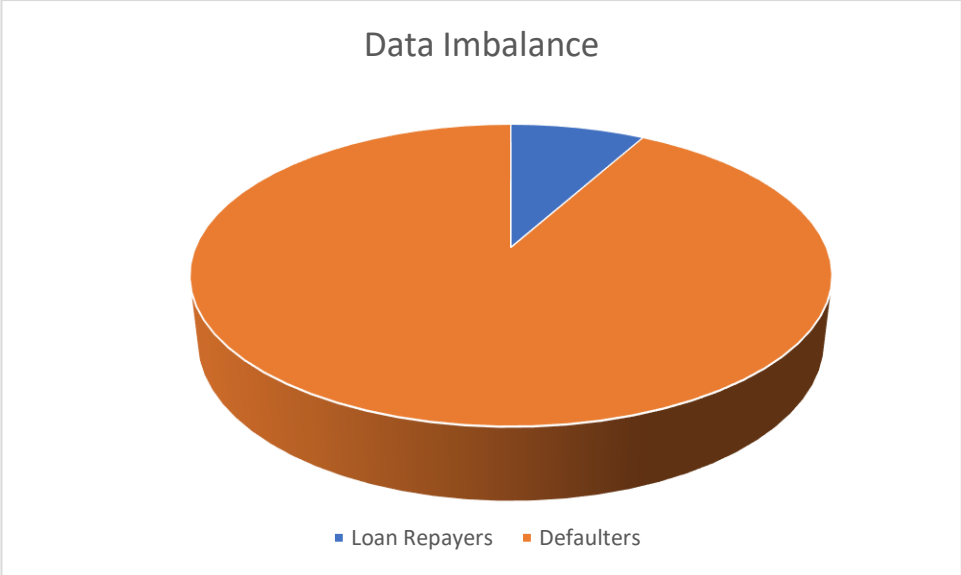
Before Cleaning
Columns 122
rows 307512

- We calculate the null values percentage using CountBlank() and there are 41 columns that have null values greater than 50% so we remove them from the dataset as imputation will not be a good option.
- For columns that have null values less than 50%, can be imputed. For numerical columns we use median/mean for imputation and for categorical variables we use mode as imputation method.
- column
- On further analysis, we found that "EXT_SOURCE_2", "EXT_SOURCE_3" has no correlation with the "TARGET" column.
- There is almost no correlation of 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE', 'FLAG_EMAIL' with the "TARGET" column.
- Upon further analysis, we find FLAG_MOBIL has all the values as 1 except 1 value which is 0. This Feature would not be of any good hence we decide to drop it as well.

- For 26 columns there are null values less than 50%. These are the features that needs to be retained.

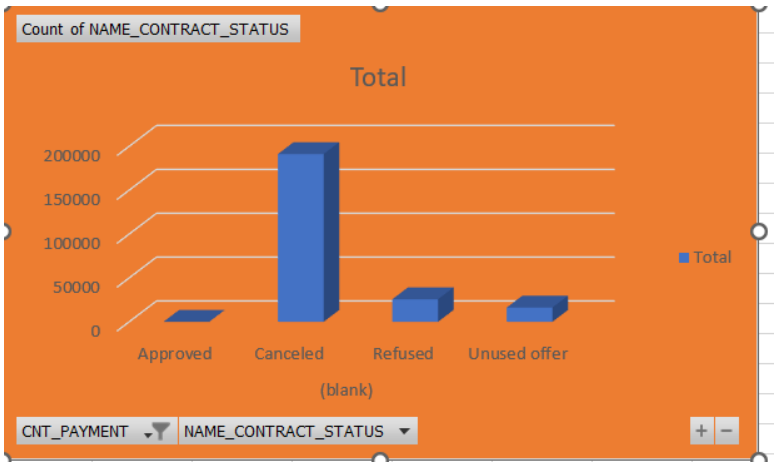
- OCCUPATION_TYPE
- OBS_30_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU_HOUR
- DEF_30_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU_DAY
- OBS_60_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU_WEEK
- DEF_60_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU_MON
- DAYS_LAST_PHONE_CHANGE
- AMT_REQ_CREDIT_BUREAU_QRT
- AMT_GOODS_PRICE
- AMT_REQ_CREDIT_BUREAU_YEAR
- AMT_ANNUITY
- NAME_TYPE_SUITE
- CNT_FAM_MEMBER
- We find the presence of outliers in AMT_INCOME_TOTAL where one of the outliers is 117000000 which is an extremely high salary to earn.
- We find no outliers in DAYS_BIRTH.
- We find a few outliers in DAYS_EMPLOYED where in we find people being employed for over 1000 Years That is impossible.
- We find a few outliers in AMT_GOODS_PRICE & AMT_CREDIT where in the amount is more than normal.
- We see a few outliers in DAYS_LAST_PHONE_CHANGE, where data suggests people using the same phone for almost 12 years which is a little too difficult with the advancement in the technology these days.
- we see a few outliers in CNT_CHILDREN which suggests there are a few people who have 19 children which is again not too realistic in today's day and age.
- Since there are less than 15 rows which are null from the below features so we drop the null values from these features--
- **AMT_ANNUITY** has a smaller number of null values (12). It can be imputed with mean.
- **CNT_FAM_MEMBERS**
- DAYS_LAST_PHONE_CHANGE
- **OCCUPATION_TYPE** has 96005 null values. Can be imputed by the category which is the most popular (Mode)
- **NAME_TYPE_SUIT – Mode Imputation**

Data Imbalance



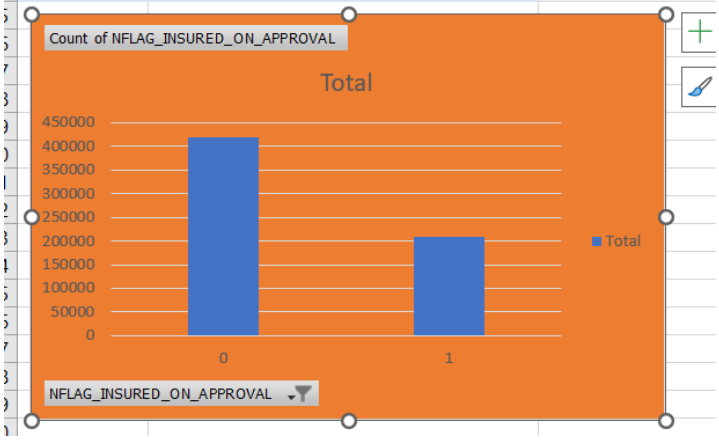
Previous_Application.csv

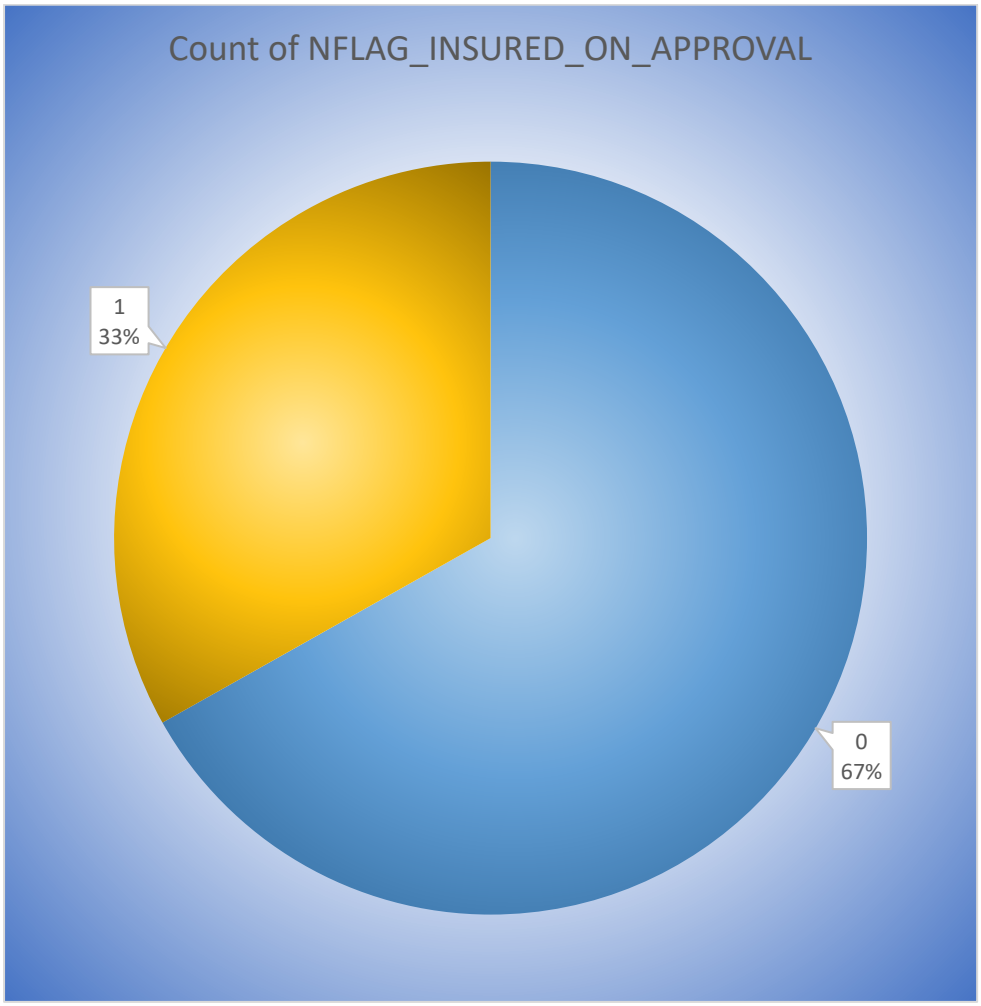
- There are 1670214 rows in the dataset where as Excel has a Max limit of 1048576 rows and as per the project requirement we are supposed to use only Excel for Analysis. Hence we'd be limited to the use of 1048576 rows
- There are 4 columns with more than 50% null values, so we drop them.
  1. RATE_INTEREST_PRIMARY
  2. RATE_INTEREST_PRIVILEGED
  3. AMT_DOWN_PAYMENT
  4. RATE_DOWN_PAYMENT
- Further we see that more unnecessary columns can be removed---
  1. NAME_TYPE_SUITE
  2. WEEKDAY_APPR_PROCESS_START
  3. HOUR_APPR_PROCESS_START
  4. FLAG_LAST_APPL_PER_CONTRACT
  5. NFLAG_LAST_APPL_IN_DAY
- Product_combination columns can be removed as it has very less % of nulls.
- Median imputation of AMT_GOODS_PRICE and AMT_ANNUITY
- Contract status for blank CNT_PAYMENT
- majority of the Contract Status were either cancelled or refused it makes more sense replacing them with 0 rather than Mean/ Median as their term of previous credits would be 0 if the loan was not taken or rejected



- Mode Imputation of NFLAG_INSURED_ON_APPROVAL

Row Labels	Count of NFLAG_INSURED_ON_APPROVAL
0	419710
1	208157
Grand Total	**627867**

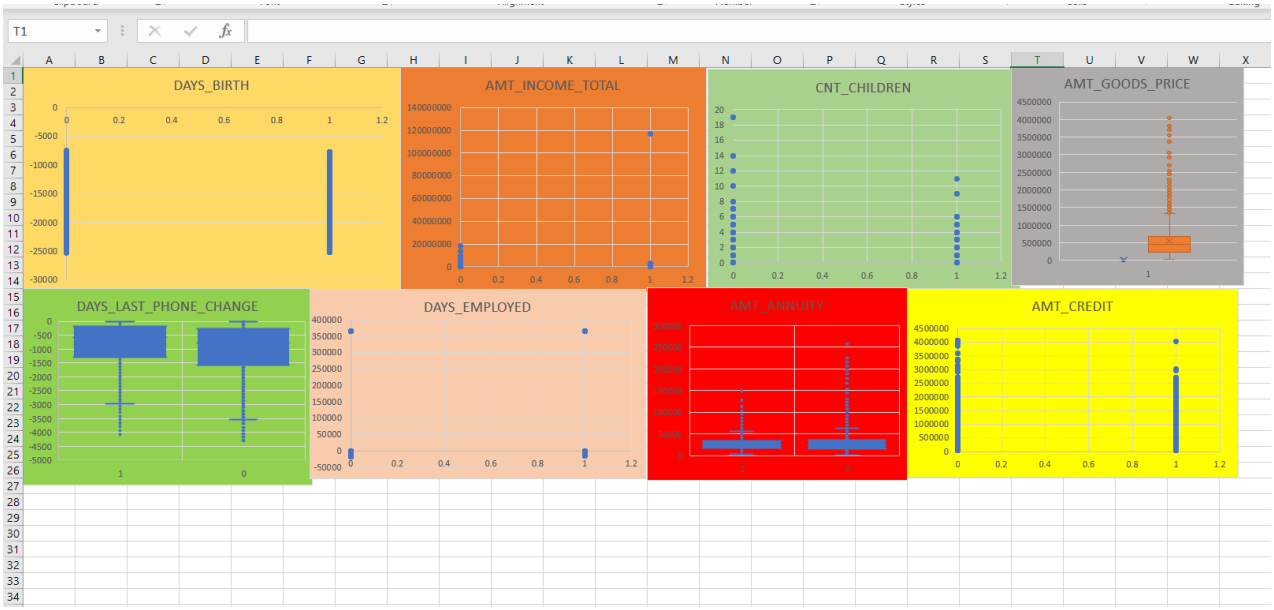




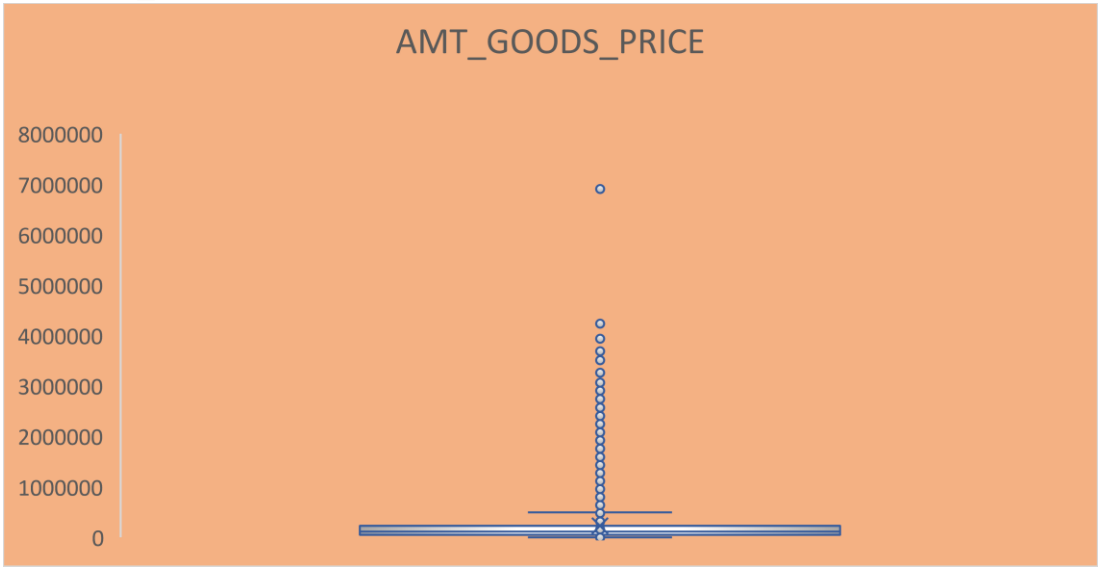
OUTLIERS FOR APPLICATION AND PREVIOUS DATA:

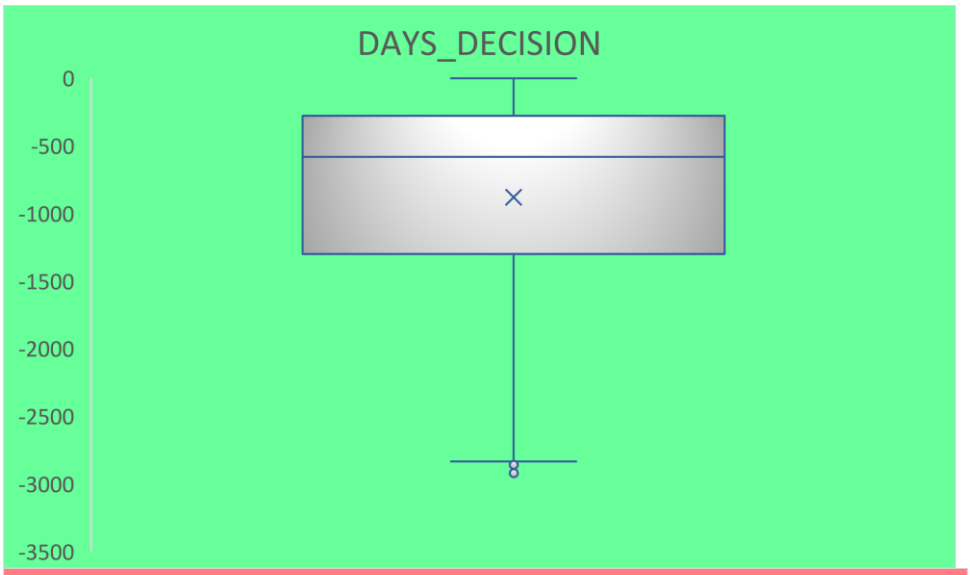
- **IQR** method and **UPPER & LOWER Bound** formulas have been used to find if the given columns contain any outliers.
- Outliers are those values which are Less Than LOWER BOUND and Greater Than UPPER BOUND. The above excel calculations clearly suggest that all the four columns have outliers in the upper bound. The same can be confirmed by Box and Whisker chart as shown by an example below.
- We find a few outliers in DAYS_EMPLOYED where we find people being employed for over 1000 Years. Which is impossible.
- We find a few outliers in AMT_GOODS_PRICE & AMT_CREDIT where in the amount is more than normal.
- We see a few outliers in DAYS_LAST_PHONE_CHANGE, where data suggests people using the same phone for almost 12 years which is a little too difficult with the advancement in the technology these days.

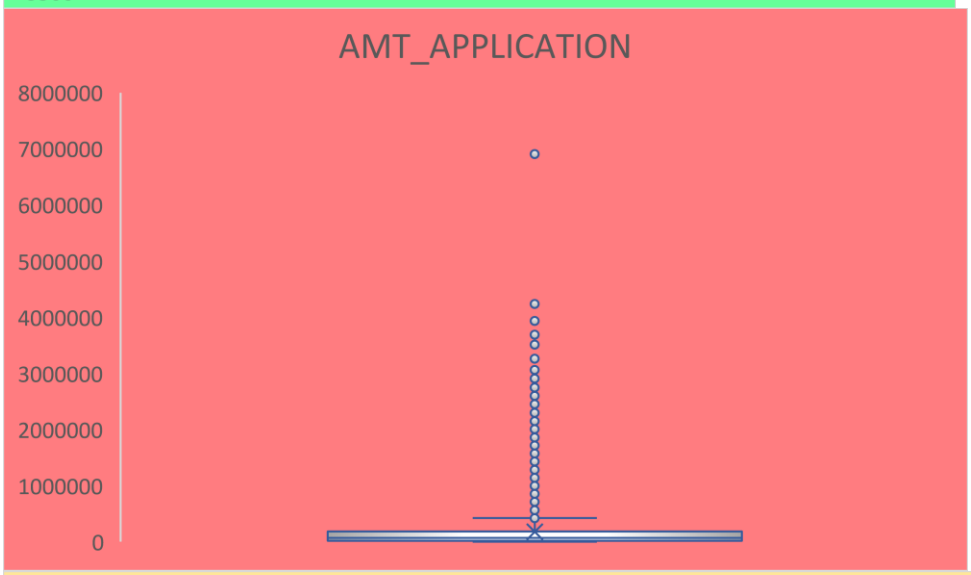
Application_data

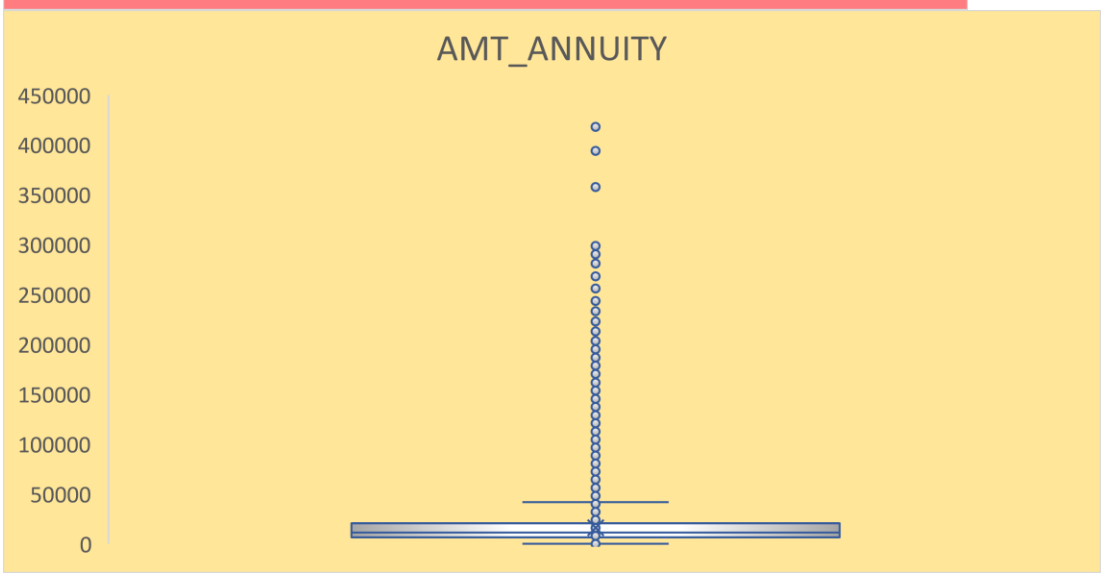


Previous_data



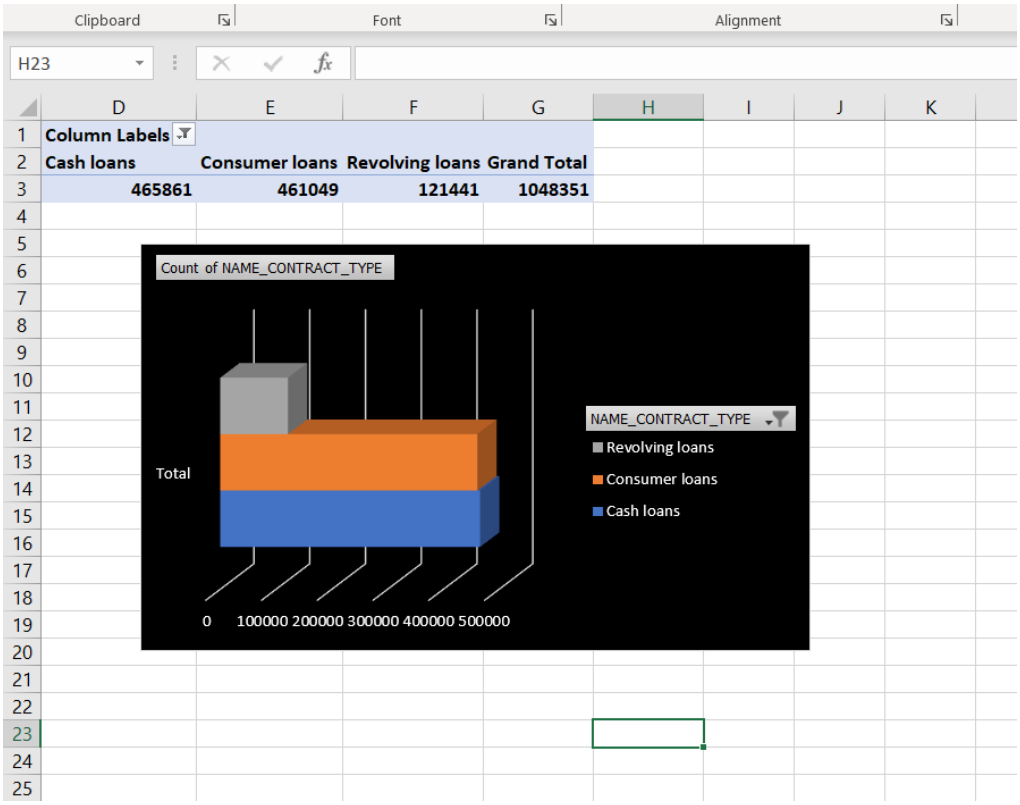


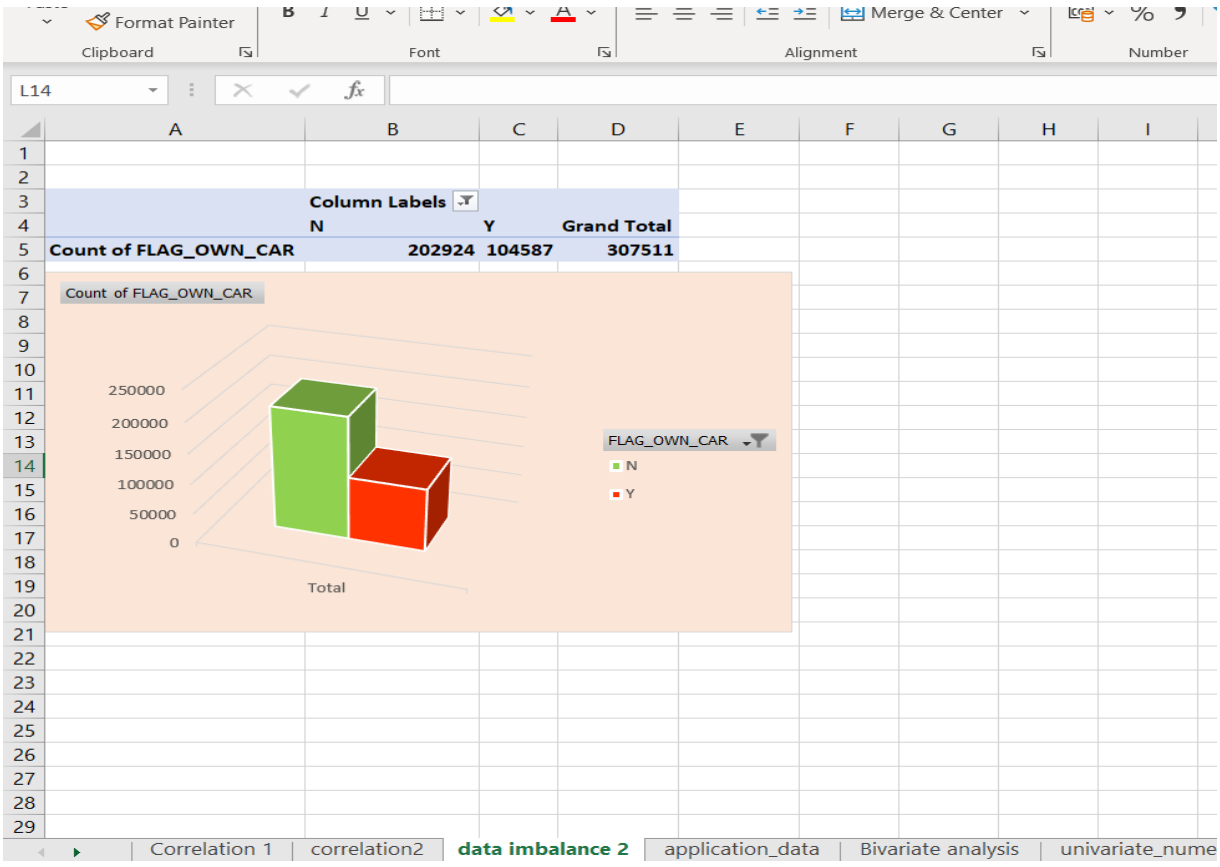


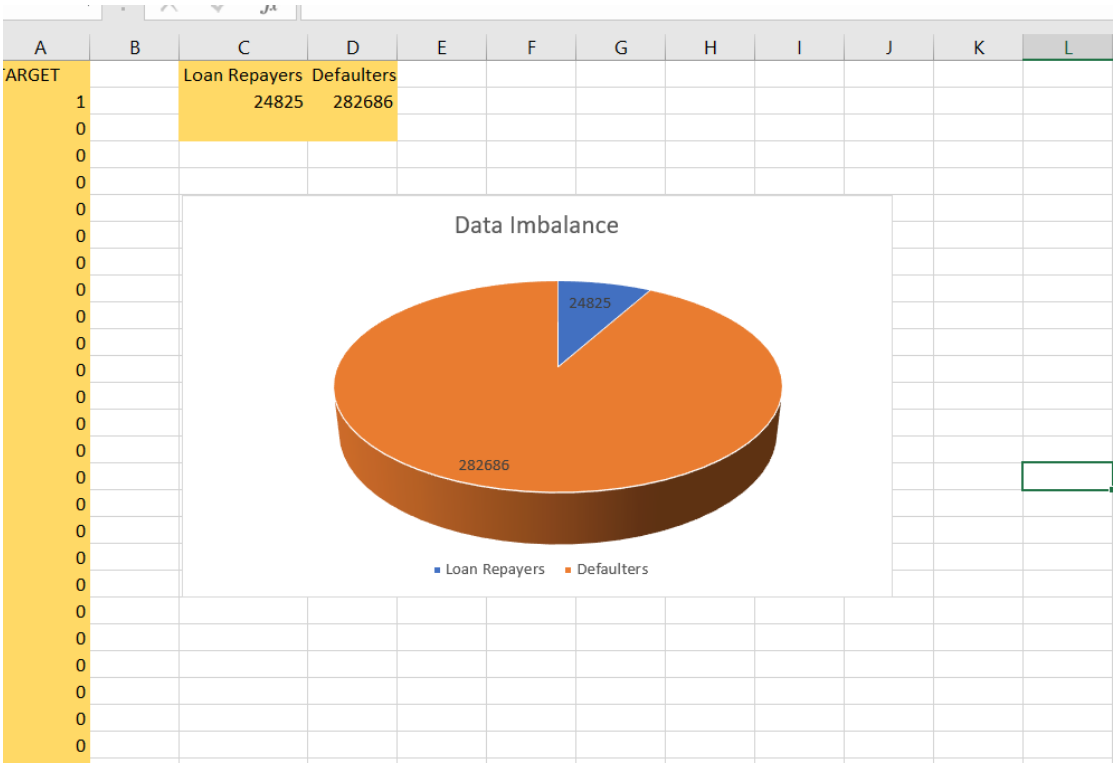


The rest of the outliers are in the excel sheet. Median Imputation for AMT_ANNUITY and AMT_GOODS_PRICE as we have outliers present.

DATA IMBALANCED for both data:

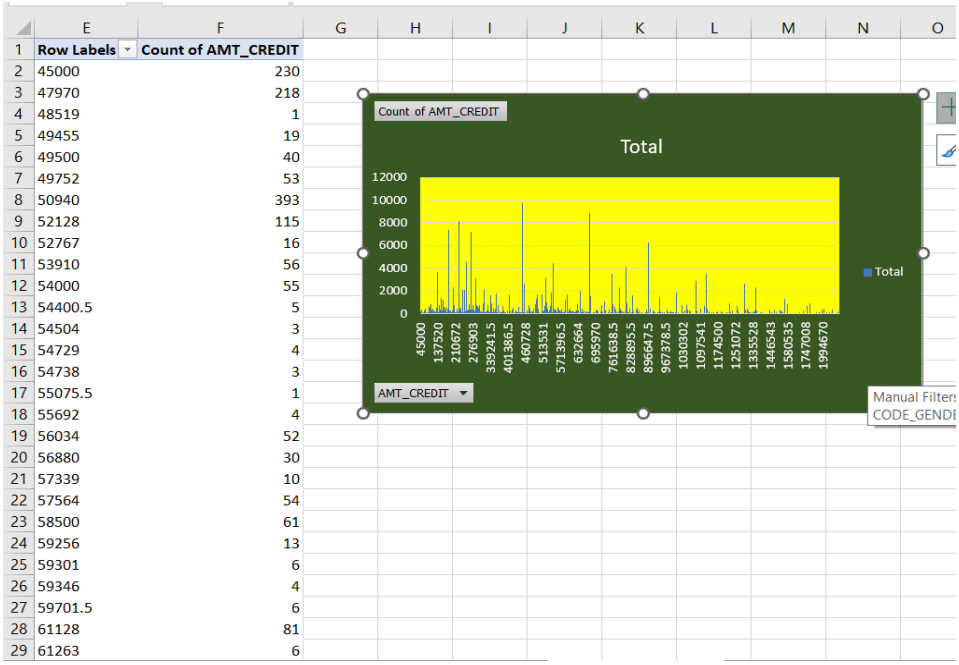


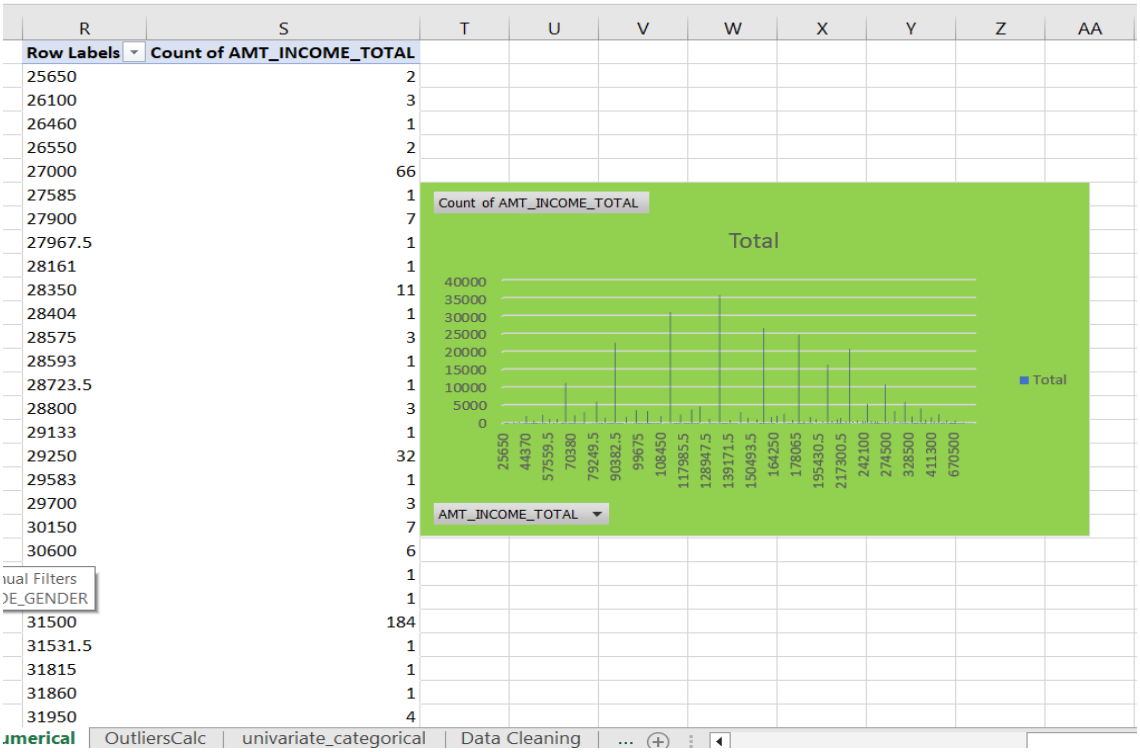




Results of univariate, segmented variate and bivariate analysis.

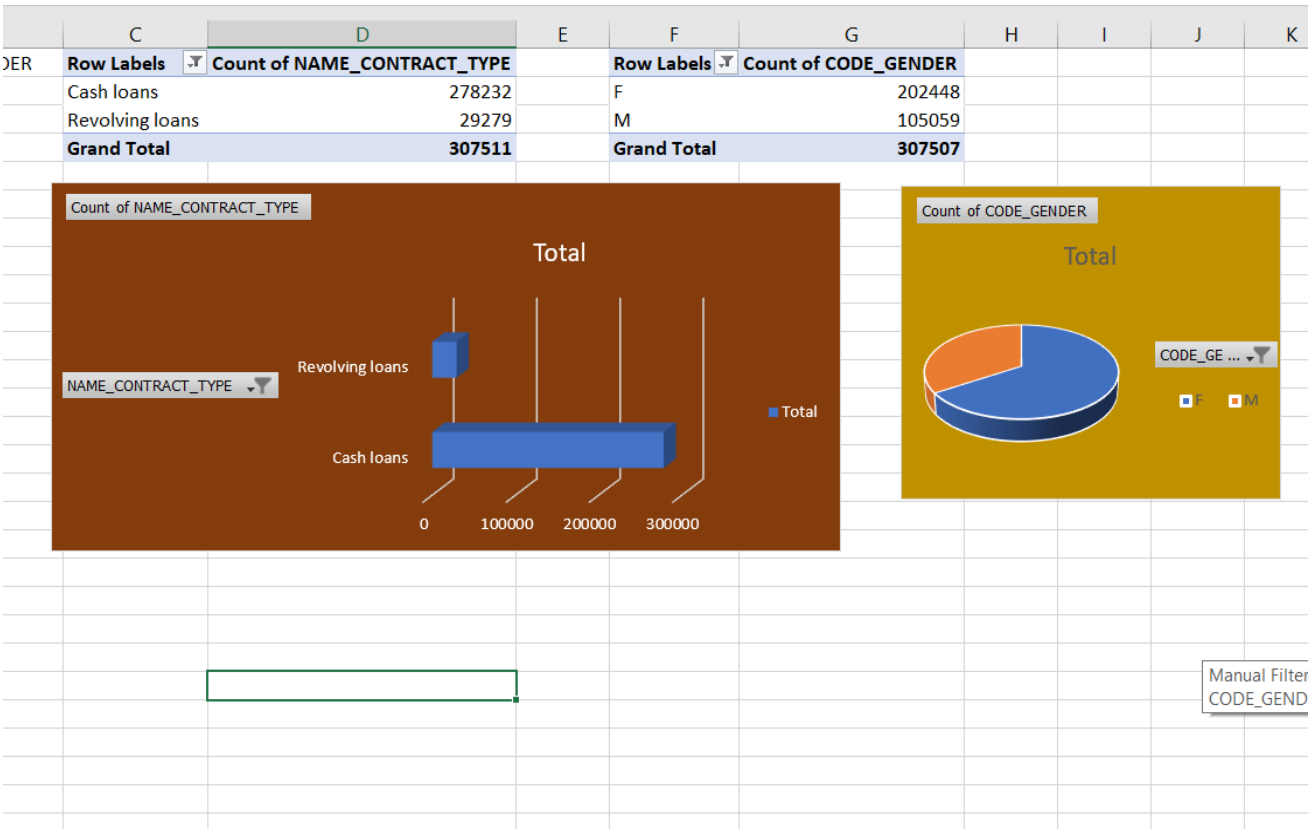
Univariate Numerical Analysis.





Row Labels	Count of AMT_INCOME_TOTAL
25650	2
26100	3
26460	1
26550	2
27000	66
27585	1
27900	7
27967.5	1
28161	1
28350	11
28404	1
28575	3
28593	1
28723.5	1
28800	3
29133	1
29250	32
29583	1
29700	3
30150	7
30600	6
	1
	1
31500	184
31531.5	1
31815	1
31860	1
31950	4

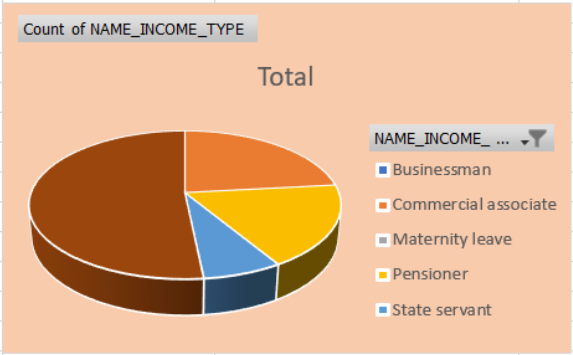
Univariate Categorical Analysis



Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	278232
Revolving loans	29279
Grand Total	**307511**

Row Labels	Count of CODE_GENDER
F	202448
M	105059
Grand Total	**307507**

NAME_INCOME_TYPE	Row Labels	Count of NAME_INCOME_TYPE
Working	Businessman	10
State servant	Commercial associate	71617
Working	Maternity leave	5
Working	Pensioner	55362
Working	State servant	21703
State servant	Student	18
Commercial associate	Unemployed	22
State servant	Working	158774
Pensioner	**Grand Total**	**307511**
Working		
Working		
Pensioner		
Working		
Working		
Working		
Working		
Working		
Working		
State servant		
Working		
Commercial associate		
Working		
Pensioner		
Working		
Working		
Working		
Working		



ita | Bivariate analysis | univariate_numerical | OutliersCalc | univariate_categorical | Data Cleani

ate

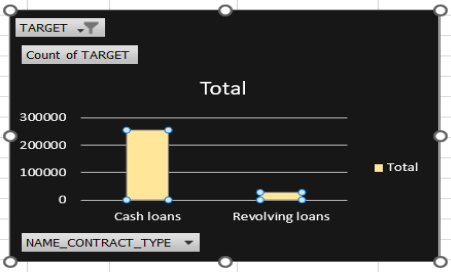
Segmented Univariate Analysis

Chart 2 | =SERIES(SegmentedUnivariate!$E$3,SegmentedUnivariate!$D$4:$D$6,SegmentedUnivar

TARGET	NAME_CONTRACT_TYPE
1	Cash loans
0	Cash loans
0	Revolving loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Revolving loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Revolving loans
0	Revolving loans
0	Cash loans
0	Revolving loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans

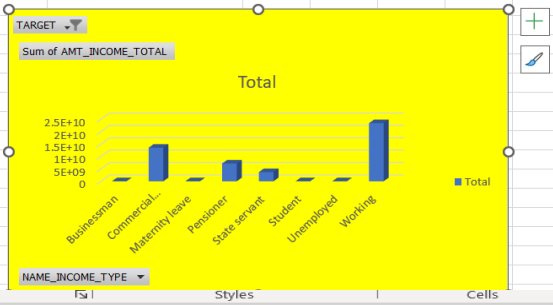
TARGET	0

Row Labels	Count of TARGET
Cash loans	255011
Revolving loans	27675
Grand Total	**282686**



Bivariate Analysis

	C	D	E	F	G	H	I	J	K	L	M
1	NAME_INCOME_TYPE		TARGET	0							
2	Working										
3	State servant		**Row Labels**	**Sum of AMT_INCOME_TOTAL**							
4	Working		Businessman	6525000							
5	Working		Commercial associate	13526206822							
6	Working		Maternity leave	585000							
7	State servant		Pensioner	7147217546							
8	Commercial associate		State servant	3695126175							
9	State servant		Student	3069000							
10	Pensioner		Unemployed	1855800							
11	Working		Working	23415319653							
12	Working		**Grand Total**	**47795904996**							
13	Pensioner										
14	Working										
15	Working										
16	Working										
17	Working										
18	Working										
19	Working										
20	Working										
21	State servant										
22	Working										
23	Commercial associate										
24	Working										
25	Pensioner										
26	Working										
27	Working										
28	Working										



E	F	G	H	I	J	K
K_ID_CURR		NAME_CONTRACT_TYPE	Cash loans			
100002		TARGET	0			
100003						
100004		**Sum of AMT_ANNUITY**	**Column Labels**			
100006		**Row Labels**	**F**	**M**	**Grand Total**	
100007		100003	35698.5		35698.5	
100008		100006	29686.5		29686.5	
100009		100007		21865.5	21865.5	
100010		100008		27517.5	27517.5	
100011		100009	41301		41301	
100012		100010		42075	42075	
100014		100011	33826.5		33826.5	
100015		100014	21177		21177	
100016		100015	10678.5		10678.5	
100017		100016	5881.5		5881.5	
100018		100017		28966.5	28966.5	
100019		100018	32778		32778	
100020		100019		20160	20160	
100021		100020		26149.5	26149.5	
100022		100023	17563.5		17563.5	
100023		100025	37561.5		37561.5	
100024		100026	32521.5		32521.5	
100025		100027	23850		23850	
100026		100029		12703.5	12703.5	
100027		100030	11074.5		11074.5	
100029		100032		23827.5	23827.5	
100030		100033		57676.5	57676.5	
100031		100035	24592.5		24592.5	
100032		100036	25033.5		25033.5	

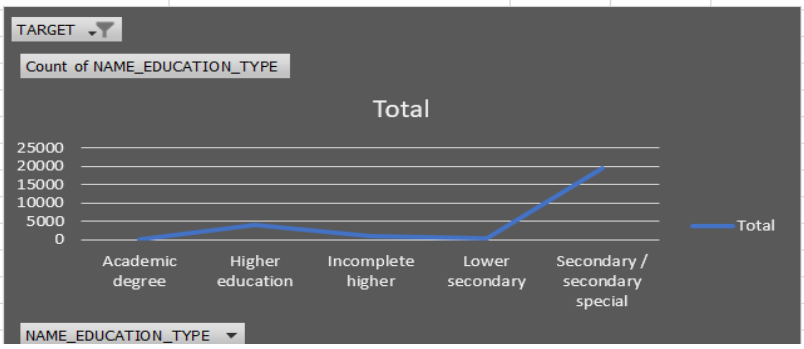
100016 (SK_ID_CURR)
Row: 100016

SegmentedUnivariate | Bivariate

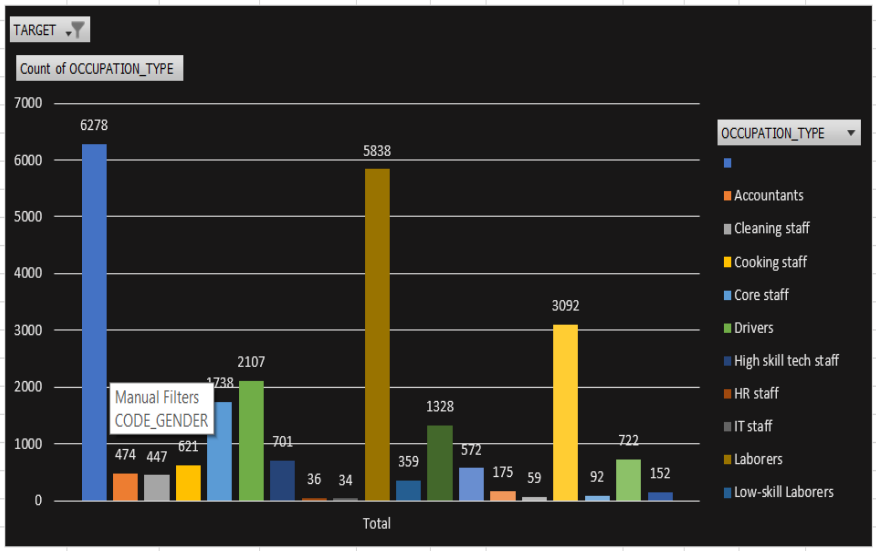
E	F	G	H	I	J	K
_ID_CURR		NAME_CONTRACT_TYPE	Revolving loans			
100002		TARGET	0			
100003						
100004		Sum of AMT_ANNUITY	Column Labels			
100006		Row Labels	F	M	Grand Total	
100007		100004		6750	6750	
100008		100012		20250	20250	
100009		100021	13500		13500	
100010		100022	7875		7875	
100011		100024		21375	21375	
100012		100034		9000	9000	
100014		100046		27000	27000	
100015		100052	9000		9000	
100016		100058	6750		6750	
100017		100068		12375	12375	
100018		100079		13500	13500	
100019		100080	22500		22500	
100020		100088	6750		6750	
100021		100095	6750		6750	
100022		100098		13500	13500	
100023		100119	9000		9000	
100024		100126		9000	9000	
100025		100129	6750		6750	
100026		100134	9000		9000	
100027		100140	33750		33750	
100029		100143		13500	13500	
100030		100154		9000	9000	
100031		100174	11250		11250	
100032		100182	6750		6750	

Correlation----------

A	B
TARGET	1
Row Labels	**Count of NAME_EDUCATION_TYPE**
Academic degree	3
Higher education	4009
Incomplete higher	872
Lower secondary	417
Secondary / secondary special	19524
Grand Total	**24825**



TARGET	1											
	Column Labels											
		Accountants	Cleaning staff	Cooking staff	Core staff	Drivers	High skill tech staff	HR staff	IT staff	Laborers	Low-skill Laborers	Managers
Count of OCCUPATION_TYPE	6278	474	447	621	1738	2107	701	36	34	5838	359	1328



TECH STACK-USED : Microsoft excel 2016

Insights:

- **As the age and experience increases, the chances of defaulting increases**
- **Educated clients tend to default less as compared to people who are less educated**
- **Corporate cleints are a safer choice as compared to labour class**
- **Male clients tend to default more**
- **As the age increases, the amount taken by the clients is higher.**

Conclusion:

This project involved extensive use of Excel. The major challenge was working with such huge data. This project helped me understand how to work with huge datasets. This helped me understand how 2 datasets are merged to analyze the details. The dataset involved a lot of missing data and outliers, handling them was a task and this project helped me understand what to how and why of handling the outliers and Null values. The project also helped me discover new add-ins such as data analyze