BANK LOAN CASE STUDY

Description:

- In this project we are going to identify patterns which indicate if a client has difficulty in paying their installments which may be used for taking actions such as denying the loan , reducing the amount of loan ,lending (to risky applicants) at higher rate of interest.
- This will ensure that the customers capable of repaying the loan are not rejected.

Tech-Stack Used:

Microsoft Excel: Used for data cleaning, analysis, and visualization.

Project Approach:

Approach towards the project means Overall strategy that outlines how the project will be executed or done to achieve its Objectives.In this project first we need to import the dataset that provided and load that into Excel. After loading the data use specific functions and formulae in excel in order to achieve the required tasks in the project. If needed use charts to visualise the data in effective way.

We structured our analysis into five main steps:

1. Addressing Missing Data
2. Detecting Outliers
3. Assessing Data Imbalance
4. Conducting Univariate, Segmented Univariate, and Bivariate Analysis
5. Identifying Correlations to Loan Default.

Cleaning The Data

- Cleaning the data is an crucial step in any data analysis as it ensures the data is accurate, reliable and consistent .
- Without data cleaning ,Our data analysis will be inaccurate ,incomplete and inconsistent which can lead to serious consequences in decision making.
- There are some steps involved in Cleaning the data. Those are discussed below in detail

STEP 1 : Removing duplicates

First we need to check if there are any duplicates present in the primary key of the given data set. Here I took SK_ID_CURR as primary key because it's the unique value of the given data set.I removed duplicate by using "Remove Duplicates " feature in Excel by selecting category as SK_ID_CURR.

STEP 2: Finding Missing Values

Missing values are those which are not available in our dataset due to various reasons such as human errors or system issues.We should handle those values. First I found the count of null or missing values in each column using COUNTBLANK function in Excel and also calculated percentage of null values in each column.

STEP 3: Handling Missing Values

After finding the percentage of Missing values in each column and I deleted the columns which are having blanks percentage more than 40 %. Because the analysis wouldn't be fair if we are having majority values in columns are missing. So I used this Criteria to delete columns.

STEP 4:Removing Irrelevant Columns

There are many unnecessary columns present in our dataset .It is better to delete those columns as it helps to reduce the size of dataset ,improve efficiency of our analysis and eliminate irrelevant information.In the given data set columns giving normalised information about building where the client lives,apartment size ,living area, common area, number of elevators , number of entrances, number of floors etc..I deleted these columns and also majority of these columns having missing values.

STEP 5: Imputing with Mean/Median

We can replace the values in columns with less than 40 % with Mean or Median depending upon the category of the column.
Here , For continuous variable columns such as AMT_ANNUITY , EXT_SOURCE_2 , I found median of that columns in the Excel using “Median” function and then replaced null values using “Find and Replace” option by keeping null in “Find” and mean in “Replace”
For categorical columns such as NAME_TYPE_SUITE, OCCUPATION_TYPE, I found most repeated category. Then I replaced all missing values with Mode of that column.

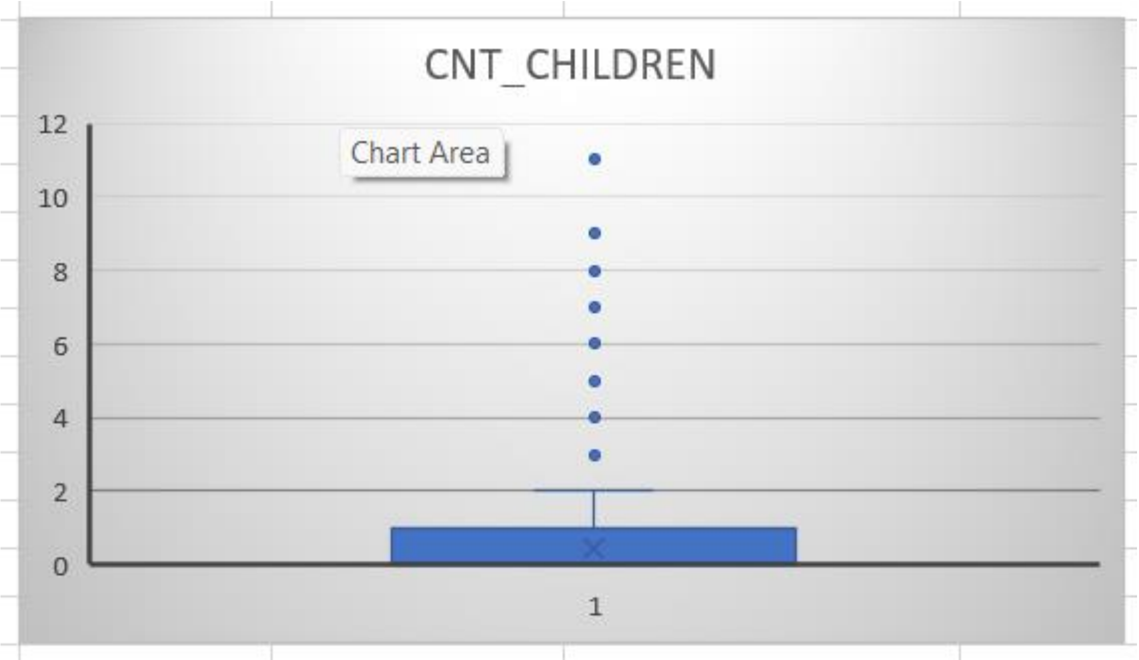
The Cleaned Dataset is provided below for reference:
https://docs.google.com/spreadsheets/d/1itD_XPW_DFPw7fU0ce5SeSe5Dh--LE5-/edit?usp=drive_link&ouid=107390593583715222805&rtpof=true&sd=true

Identifying Outliers

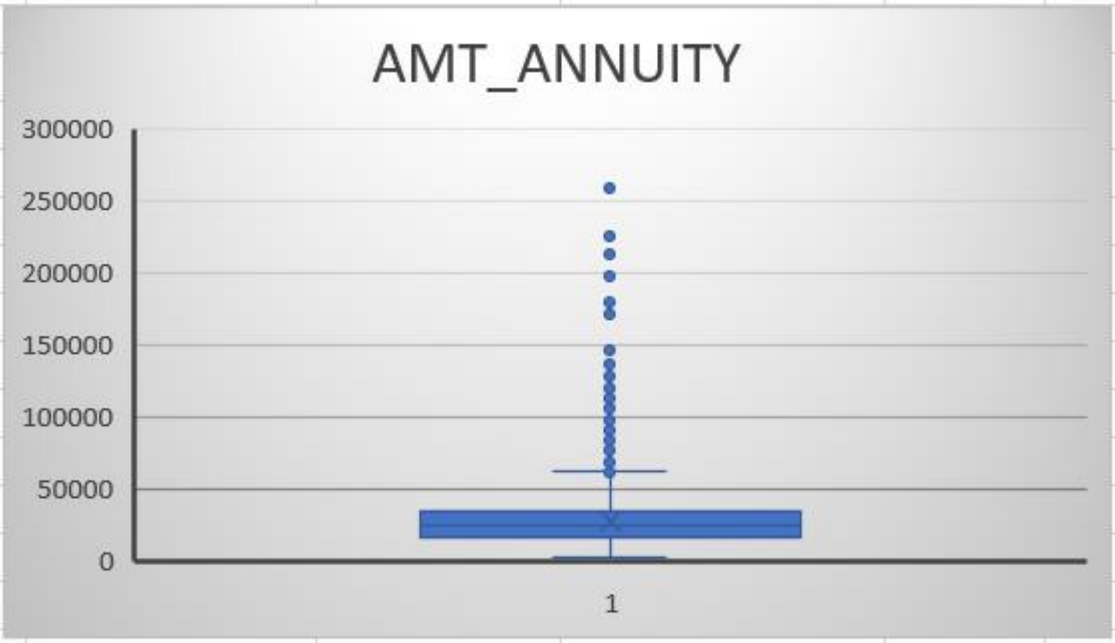
- We know that the Outliers are the data points that lie far away from the rest of the data points due to some measurement errors , data entry errors and some other natural errors.
- These can affect the statistical analysis of the data as they reduce accuracy of models.I found those outliers using Box and Whisker plot .
- Steps to find Outliers using Tukey’s method:
    1. Finding 1st Quartile Q1 and 3rd Quartile Q3
    2. Finding Inner Quarter Range(IQR)

3. Finding Upper Bound(Q3+(1.5*IQR)
4. Finding Lower Bound(Q1-(1.5*IQR)
5. Now, any data point above Upper Bound or Below Lower Bound Considered as Outlier

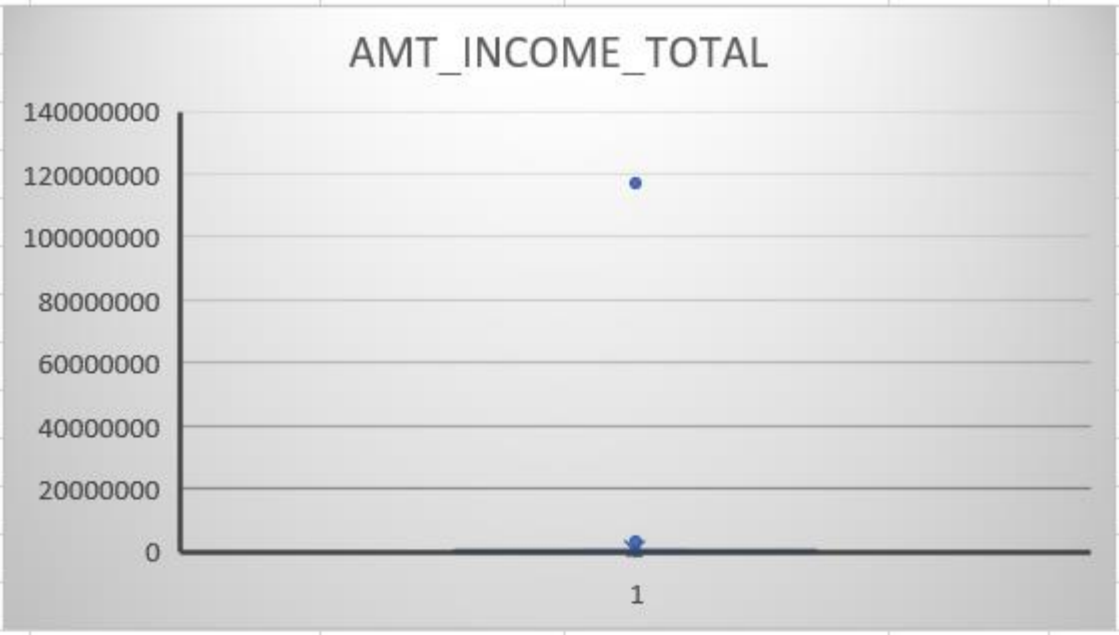
Below , I plotted some Box-Whisker Plots to find there are any outliers present in the data.



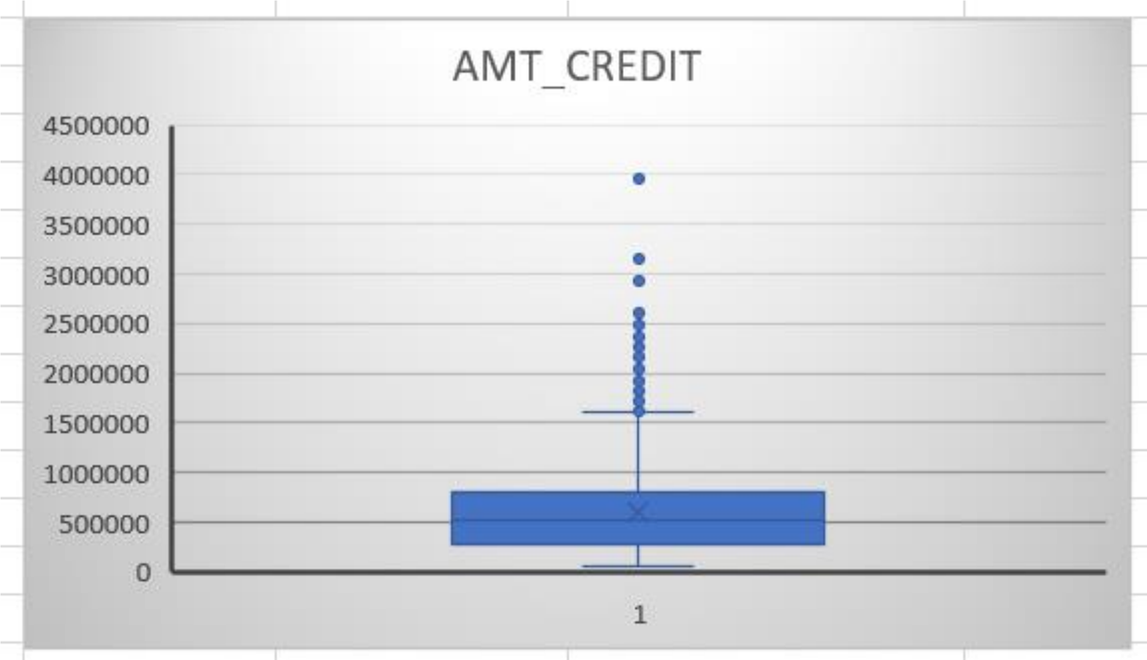
In CNT_CHILDREN, maximum count is 11 which is not acceptable. It has to replace with the median value.



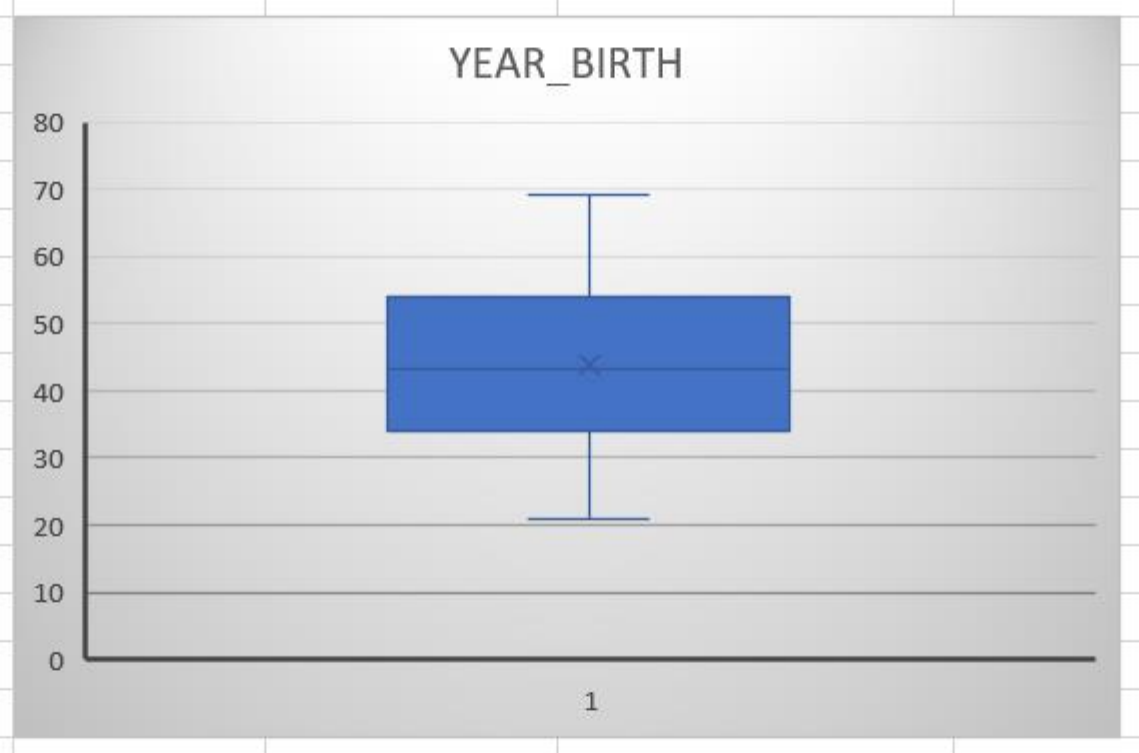
First outlier is in AMT_ANNUITY which is greater than 250000 this outlier is replaced with 24939 median of AMT_ANNUITY.



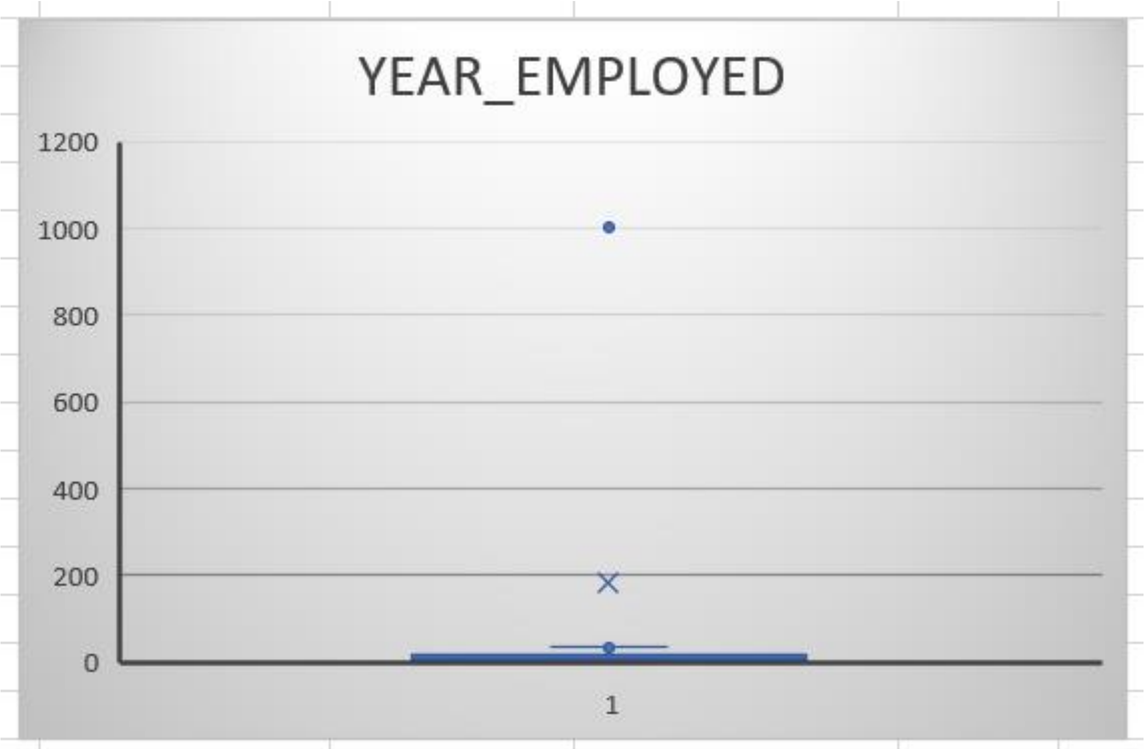
Here we can observe that there is huge difference between the 25%, 50% and 75% quartile and this is due to presence of outliers. But since the amount of total income varies from person to person we will not remove the outliers.



From the chart it is clear that outliers lie in the 98% and near max side of the box plot. Also there is a significant difference between the 75% quartile and the max value and this is due the presence of the outliers. But since the amount of credit varies from person to person we will not remove the outliers.



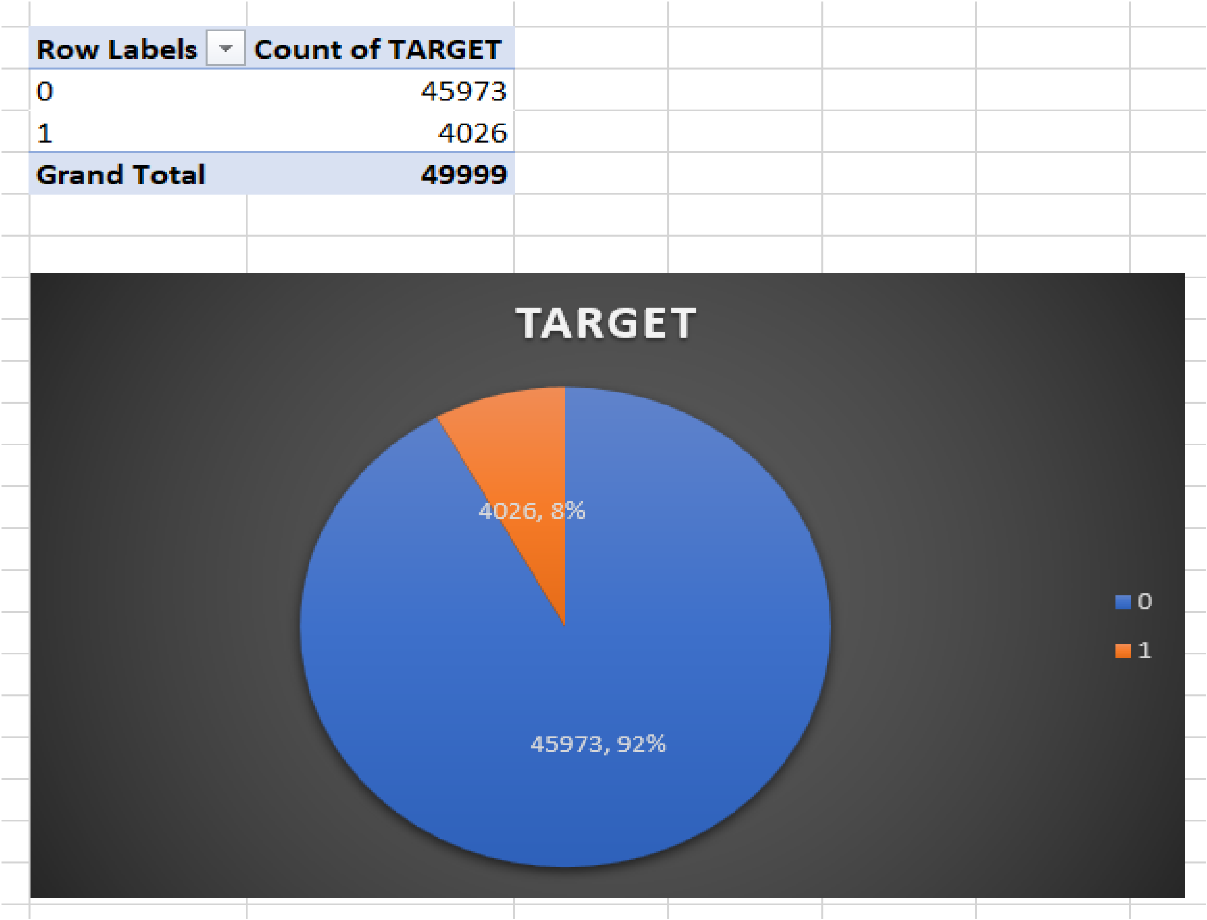
As seen from the boxplot it is clear that there are no outliers. The data of YEARS_BIRTH is well distributed.



There exists only 1 outlier. So, replace it with median value 6.07.

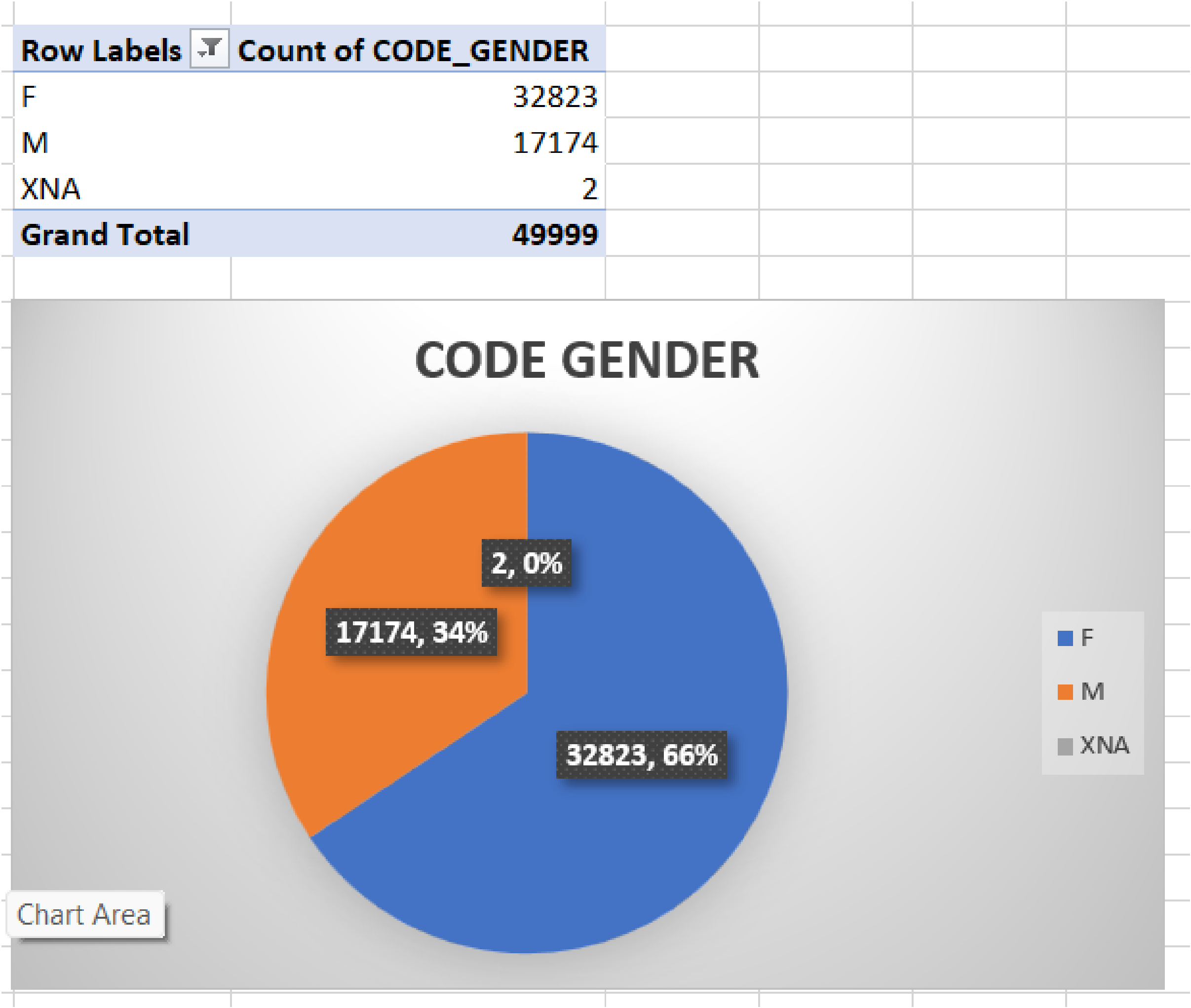
Data Imbalance

- Data Imbalance refers to situation where the classes in classification problem are not equally represented in the dataset.
- Imbalanced datasets can lead to poor performance in data analysis because data model may be biased towards majority of classes and have difficulty in identifying minority classes.
- We can find the data imbalances by creating the pivot chart of Histogram in excel.



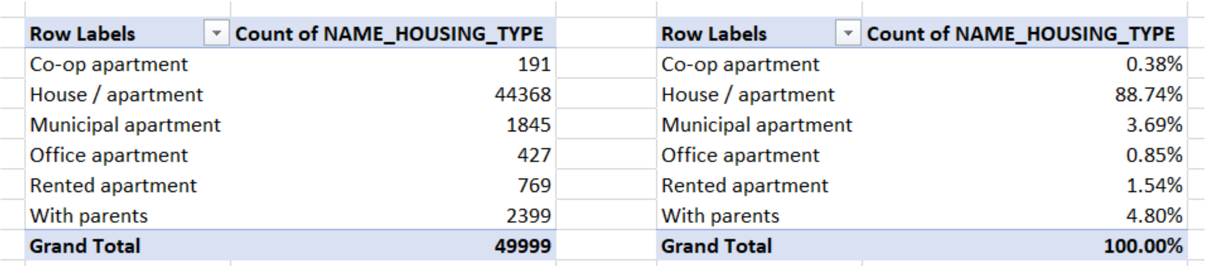
Row Labels	Count of TARGET
0	45973
1	4026
Grand Total	**49999**

The Target Variable Pie chart shows that almost 92% of the total clients had no problem during payment while 8% of the clients had some or the other problem



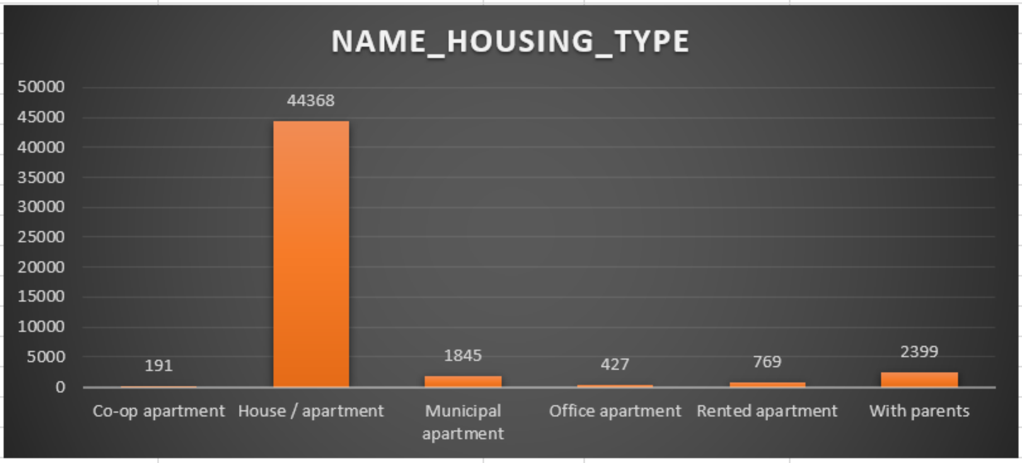
Row Labels	Count of CODE_GENDER
F	32823
M	17174
XNA	2
Grand Total	**49999**

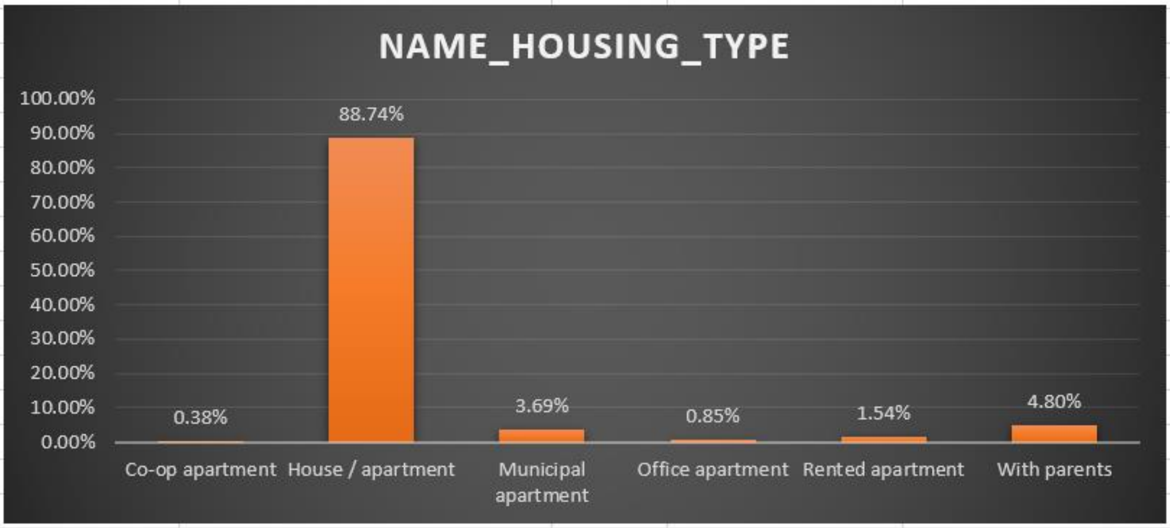
From the GENDER_VARIABLE pie chart we can infer that almost 66% of the clients are female and 34% of the clients are Male The 2 of the appicants have gender as XNA which can be ignored



Row Labels	Count of NAME_HOUSING_TYPE
Co-op apartment	191
House / apartment	44368
Municipal apartment	1845
Office apartment	427
Rented apartment	769
With parents	2399
Grand Total	**49999**

Row Labels	Count of NAME_HOUSING_TYPE
Co-op apartment	0.38%
House / apartment	88.74%
Municipal apartment	3.69%
Office apartment	0.85%
Rented apartment	1.54%
With parents	4.80%
Grand Total	**100.00%**

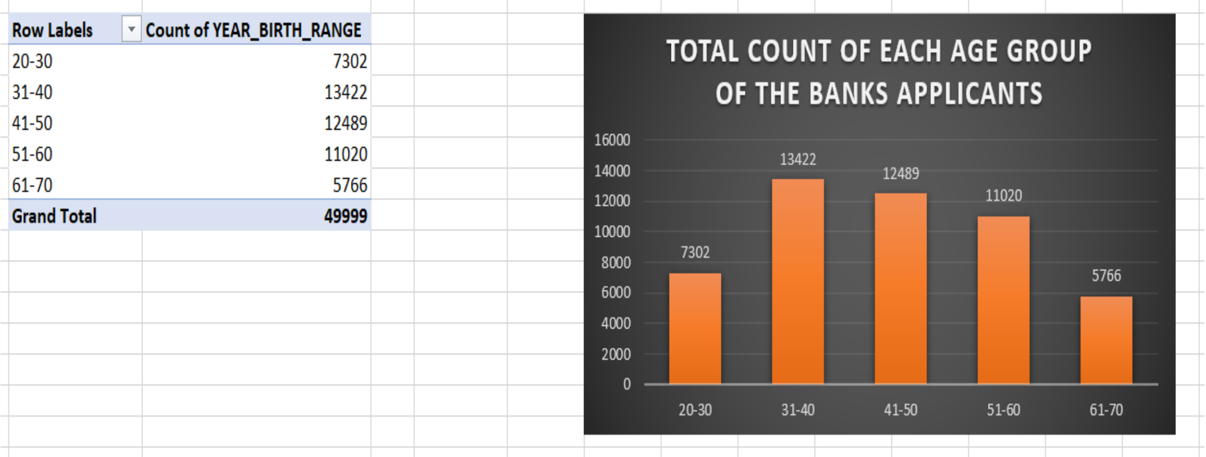




From the bar graphs of count and percentage The bank can target those groups who do not have their own apartment i.e. the bank may consider the people living in Co-op apartment, Municipal Apartment, Rented Apartment and people living with their parents.

Univariate Analysis

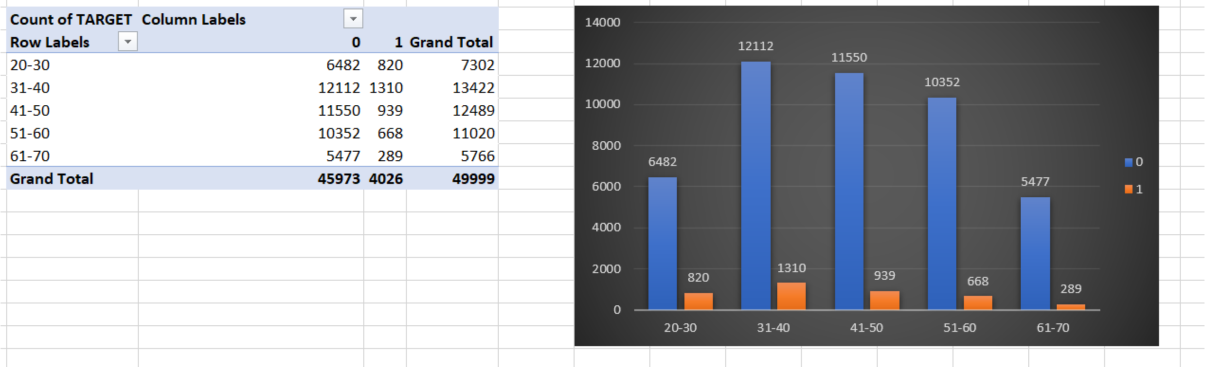
- ➢ Univariate analysis is statistical method to analyze the data with one variable.
- ➢ It involves the examining the distribution of single variable and deriving insights from it.



Row Labels	Count of YEAR_BIRTH_RANGE
20-30	7302
31-40	13422
41-50	12489
51-60	11020
61-70	5766
Grand Total	**49999**

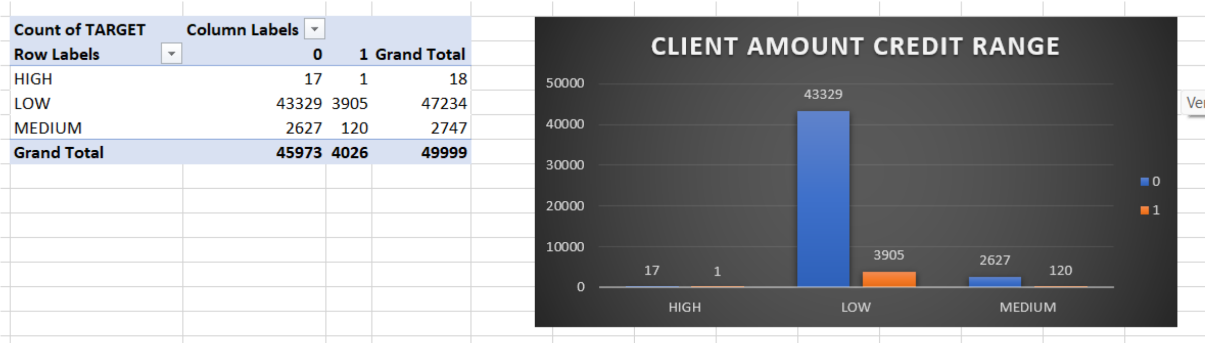
TOTAL COUNT OF EACH AGE GROUP OF THE BANKS APPLICANTS

From the above bar plot we can infer that most of the applicants belong to the Age Group ’31-40’.



Count of TARGET	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
20-30	6482	820	7302
31-40	12112	1310	13422
41-50	11550	939	12489
51-60	10352	668	11020
61-70	5477	289	5766
Grand Total	**45973**	**4026**	**49999**

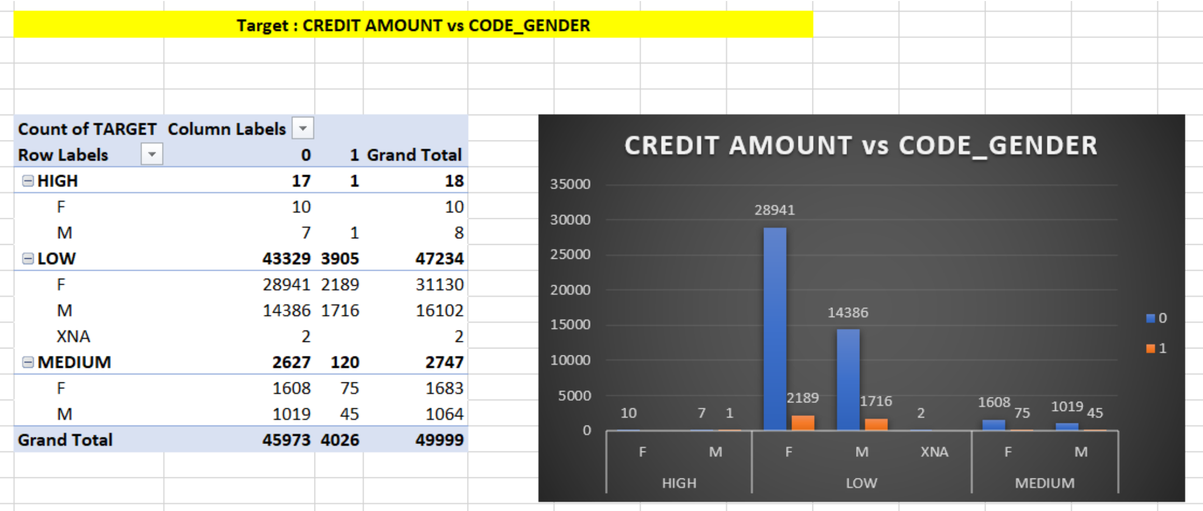
From the above Bar plot we can infer that clients/applicants in the Age Group ’31-40’ are having the highest number when it comes to doing/returning Payment to Banks.



Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
HIGH	17	1	18
LOW	43329	3905	47234
MEDIUM	2627	120	2747
Grand Total	**45973**	**4026**	**49999**

- ➢ From the above Bar plot we can infer that clients belonging to ‘Low’ income range have the highest count when it comes to clients with no payment issues.
- ➢ From the above Bar plot we can infer that clients belonging to ‘Low’ income range have the highest count when it comes to clients with payment issues.

Bivariate Analysis



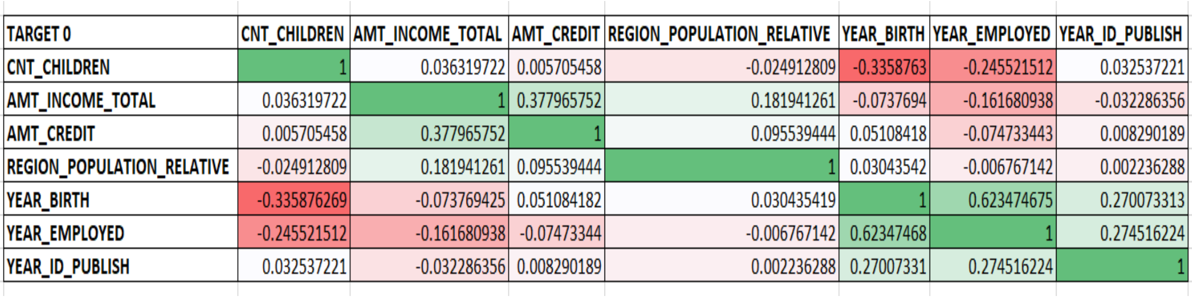
Target : CREDIT AMOUNT vs CODE_GENDER

Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
⊟**HIGH**	**17**	**1**	**18**
F	10		10
M	7	1	8
⊟**LOW**	**43329**	**3905**	**47234**
F	28941	2189	31130
M	14386	1716	16102
XNA	2		2
⊟**MEDIUM**	**2627**	**120**	**2747**
F	1608	75	1683
M	1019	45	1064
Grand Total	**45973**	**4026**	**49999**

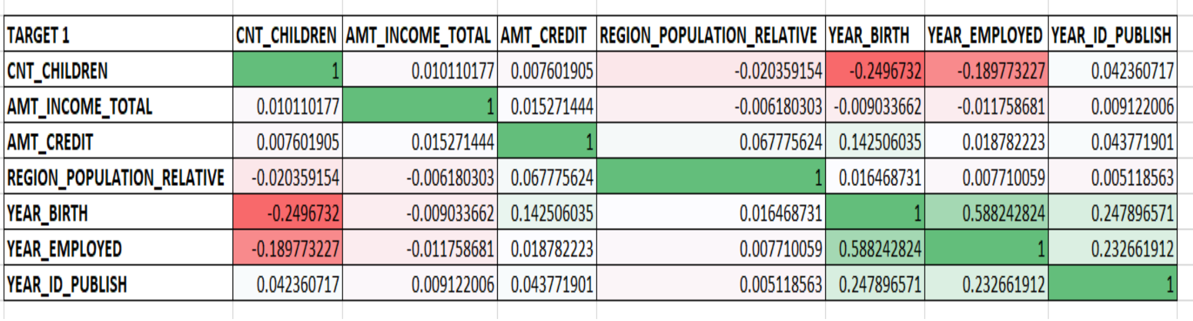
- ➢ From the above Bar plot we can infer that Females belonging to Low credit range are the highest number of clients with no payment issues.
- ➢ From the above Bar plot we can infer that Females belonging to Low credit range are the highest number of clients with payment issues.

Correlation Analysis

- Correlation analysis is Statistical method used to measure strength and direction of the relationship between two variables.
- Correlation Coefficient is the measure of linear relationship between two variables.The value of correlation coefficient ranges from -1 to +1.



TARGET 0	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	YEAR_BIRTH	YEAR_EMPLOYED	YEAR_ID_PUBLISH
CNT_CHILDREN	1	0.036319722	0.005705458	-0.024912809	-0.3358763	-0.245521512	0.032537221
AMT_INCOME_TOTAL	0.036319722	1	0.377965752	0.181941261	-0.0737694	-0.161680938	-0.032286356
AMT_CREDIT	0.005705458	0.377965752	1	0.095539444	0.05108418	-0.074733443	0.008290189
REGION_POPULATION_RELATIVE	-0.024912809	0.181941261	0.095539444	1	0.03043542	-0.006767142	0.002236288
YEAR_BIRTH	-0.335876269	-0.073769425	0.051084182	0.030435419	1	0.623474675	0.270073313
YEAR_EMPLOYED	-0.245521512	-0.161680938	-0.07473344	-0.006767142	0.62347468	1	0.274516224
YEAR_ID_PUBLISH	0.032537221	-0.032286356	0.008290189	0.002236288	0.27007331	0.274516224	1



TARGET 1	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	YEAR_BIRTH	YEAR_EMPLOYED	YEAR_ID_PUBLISH
CNT_CHILDREN	1	0.010110177	0.007601905	-0.020359154	-0.2496732	-0.189773227	0.042360717
AMT_INCOME_TOTAL	0.010110177	1	0.015271444	-0.006180303	-0.009033662	-0.011758681	0.009122006
AMT_CREDIT	0.007601905	0.015271444	1	0.067775624	0.142506035	0.018782223	0.043771901
REGION_POPULATION_RELATIVE	-0.020359154	-0.006180303	0.067775624	1	0.016468731	0.007710059	0.005118563
YEAR_BIRTH	-0.2496732	-0.009033662	0.142506035	0.016468731	1	0.588242824	0.247896571
YEAR_EMPLOYED	-0.189773227	-0.011758681	0.018782223	0.007710059	0.588242824	1	0.232661912
YEAR_ID_PUBLISH	0.042360717	0.009122006	0.043771901	0.005118563	0.247896571	0.232661912	1

The Results Dataset Link:-

https://docs.google.com/spreadsheets/d/1itD_XPW_DFPw7fU0ce5SeSe5Dh--LE5-/edit?usp=drive_link&ouid=107390593583715222805&rtpof=true&sd=true

Conclusion:

From this BANK LOAN CASE STUDY , I learned how the data is being analyzed in Banking sector to give loans and asses the risks in them.I gained valuble insights from this Case Study by analyzing historical loan data and customer characteristics and I may be able to develop models to predict the likelihood of loan defaults.I also knew about the "Exploratory Data Analysis" and it's importance in data analysis.

- Clients who hold Academic degree are highly capable of paying loans in time.
- Female Clients are more capable than Male Clients in repaying the loans.
- Clients who taken Revolving loans are facing less difficulty in repaying the loan when compared to clients who took other type of loans.
- Clients who fall in the Age Group 31-40 have the highest count for paying off their loans on time followed by the clients who fall in the Age Groups 41-60.
- Students and Businessmen are getting no difficulties in re paying the loans.