Name: - VARUN GANDHI (Data Analytics Trainee)  
E-Mail :- [varun.poseidon84@gmail.com](mailto:varun.poseidon84@gmail.com)  
Contact :- +91 9465433841

**Bank Loan Case Study** / PROJECT – 6

**DESCRIPTION :-­­­­**

In this project we are going to operate on various excel sheets containing data for loan applications and also previous data records. We are going to focus on what factors influence the approval or refusal towards a loan. Our task is to use exploratory data analysis (EDA) to analyze patterns in the data and ensure that capable applications are not rejected no matter what. This is done to ensure the transparency in the system and also improves the reputation of the bank.

Our major concerns in this project will be:

* Data cleansing: This step involves preprocessing of the dataset to make it suitable for data analysis. Proceeding without cleaning the data will result in inaccurate outcomes. This includes eradicating any missing values, removing duplicates and converting data types if necessary.
* Data Analysis : This step involves analysing various factors that contribute to the approval, cancellation, refusal or unused offer condition towards a loan application. Our major objective will be to identify the patterns that indicate if a customer will have difficulty paying their installments. This information can be further uses to make decisions on a person’s loan application whether to deny, reduce or accept the loan.

**PROJECT APPROACH :- ­­­­­­­­­­­­­**

* Our approach will simply consist of performing various operations on the dataset.
* Our first step is to find the discrepancies in the dataset like missing values in rows and columns which can interfere in accuracy.
* Next, we need to perform various operations like identifying missing values and finding appropriate methods to deal with them, detecting and removing outliers and visualize the results through different graphs and charts.

**TECH-STACK USED :-**

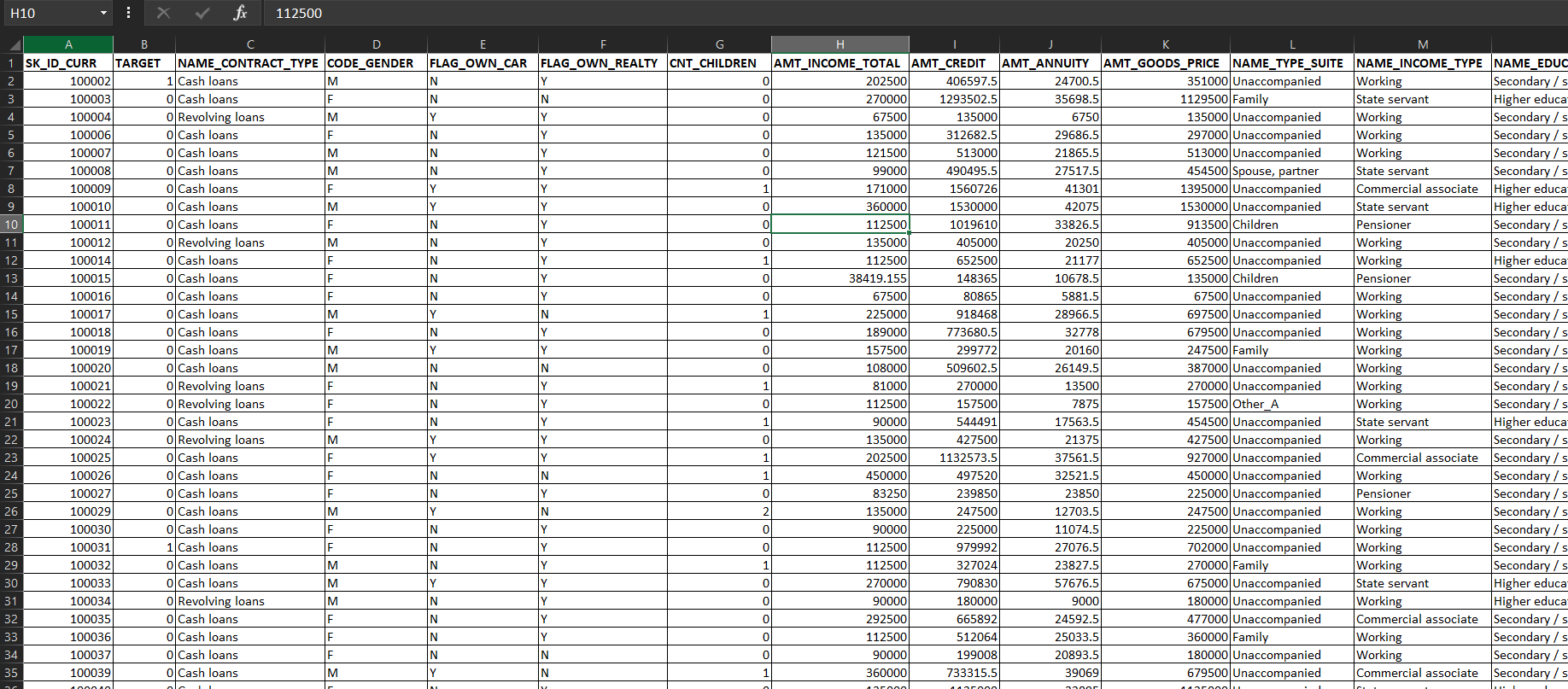


**INSIGHTS :-**

|  |  |
| --- | --- |
| Identifying Missing data and dealing with it appropriately | It is extremely crucial for a data analyst to handle the missing values to ensure the precision and accuracy of results. Our objective will be to identify the missing data in the dataset and decide on an appropriate method to deal with it. |
| Identifying Outliers and removing them | Outliers are just like exceptions in a dataset which need to be removed to ensure the accuracy of the data. Outliers can significantly impact the data analysis and also distort the results. Our task will be to detect and identify outliers in the dataset. |
| 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. Our task is to determine if there Is data imbalance in the loan application dataset and also calculate the ration of data imbalance using excel functions. |
| Perform Univariate, Segmented Univariate and Bivariate Analysis | To gain useful insights into the driving factors of the loan default, it is important to conduct various analysis on consumers and loan attributes. |
| Identify top correlations for different scenarios | Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default. Our task is to segment the dataset based on different scenarios and identifying the top correlations for each segmented data. |

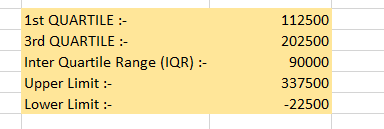
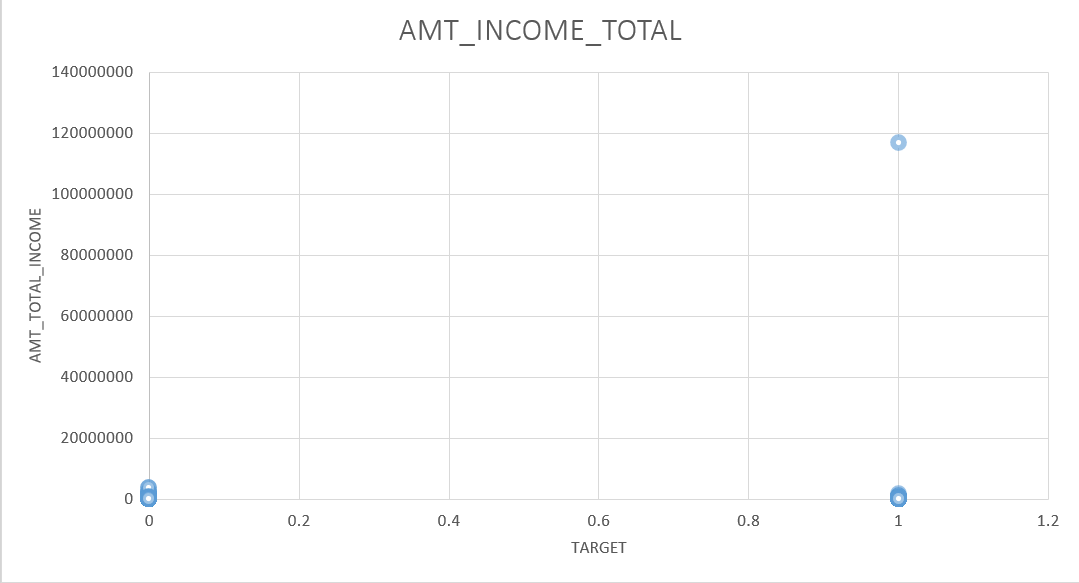
**ANALYSIS :-**

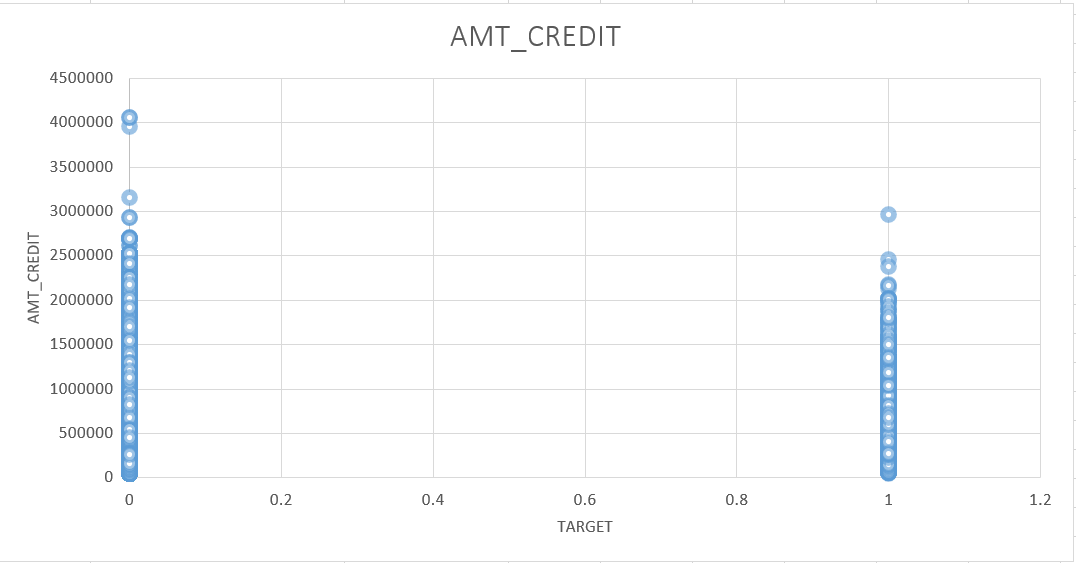
**PERFORMING OPERATIONS ON APPLICATION\_DATA**

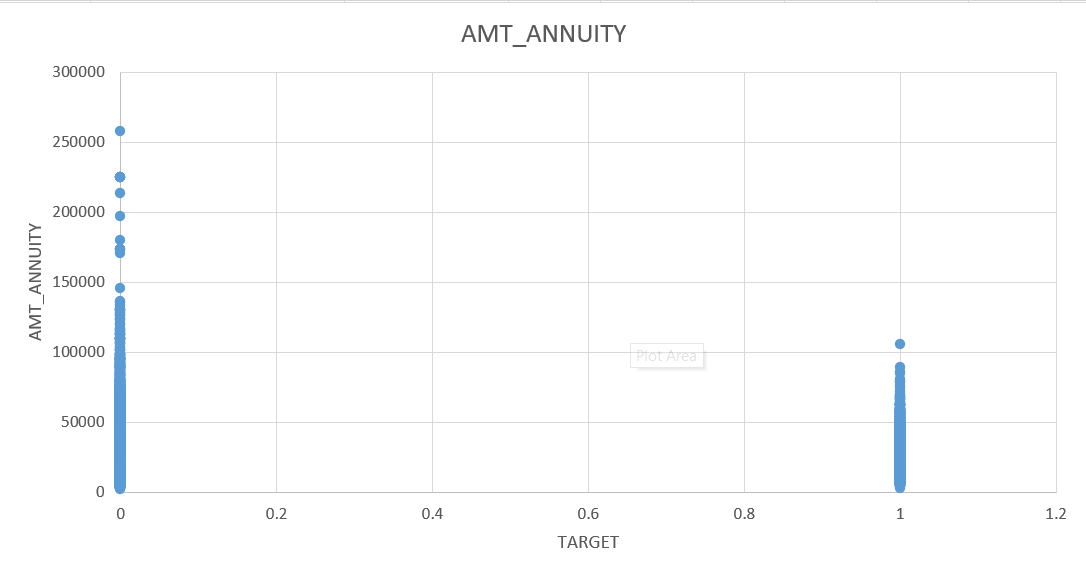
1. **Task cleansing :**
2. **Firstly we have removed all the duplicate values in correspondence with the ‘SK\_ID\_CURR’ column.**
3. **Next, we found out the columns having more than 50% null entries by using the following formula : 100-((B2/50000)\*100).**
4. **And we removed the columns which exceeded more than 50% null values.**
5. **Outlier detection/identification :**
6. **Our approach is to calculate the 1st quartile, 3rd quartile, interquartile range (IQR), upper limit and the lower limit.**
7. **Next, we create a scatter plot to visualize the outlier using target and amount\_total\_income column.**
8. **And creating scatter plot to find the outlier using the other two columns namely target and CNT\_children.**

**1st quartile formula : =QUARTILE.INC(C:C, 1)  
2nd quartile formula :** **=QUARTILE.INC(C:C, 3)**

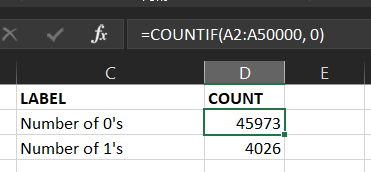
**Interquartile Range formula : =K3-K2**

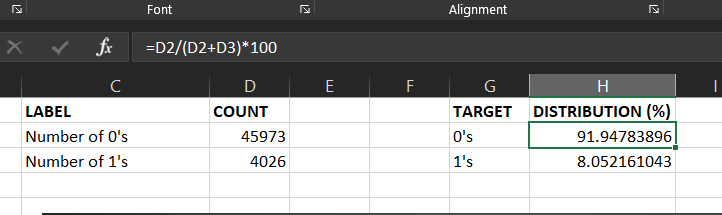
**Upper Limit : =K3+(1.5\*K4)  
Lower Limit : =K2-(1.5\*K4)**

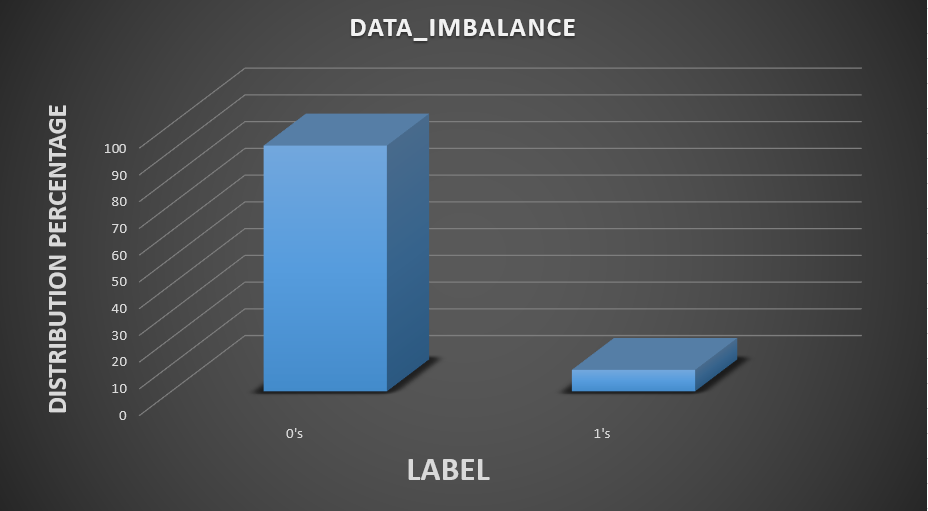




1. **Analysing Data Imbalance :**
2. **Our major approach here will be to count the number of 0’s and 1’s present in the ‘target’ column.**
3. **Next, we will plot a pivot/graph to visualize the distribution of imbalanced data.**



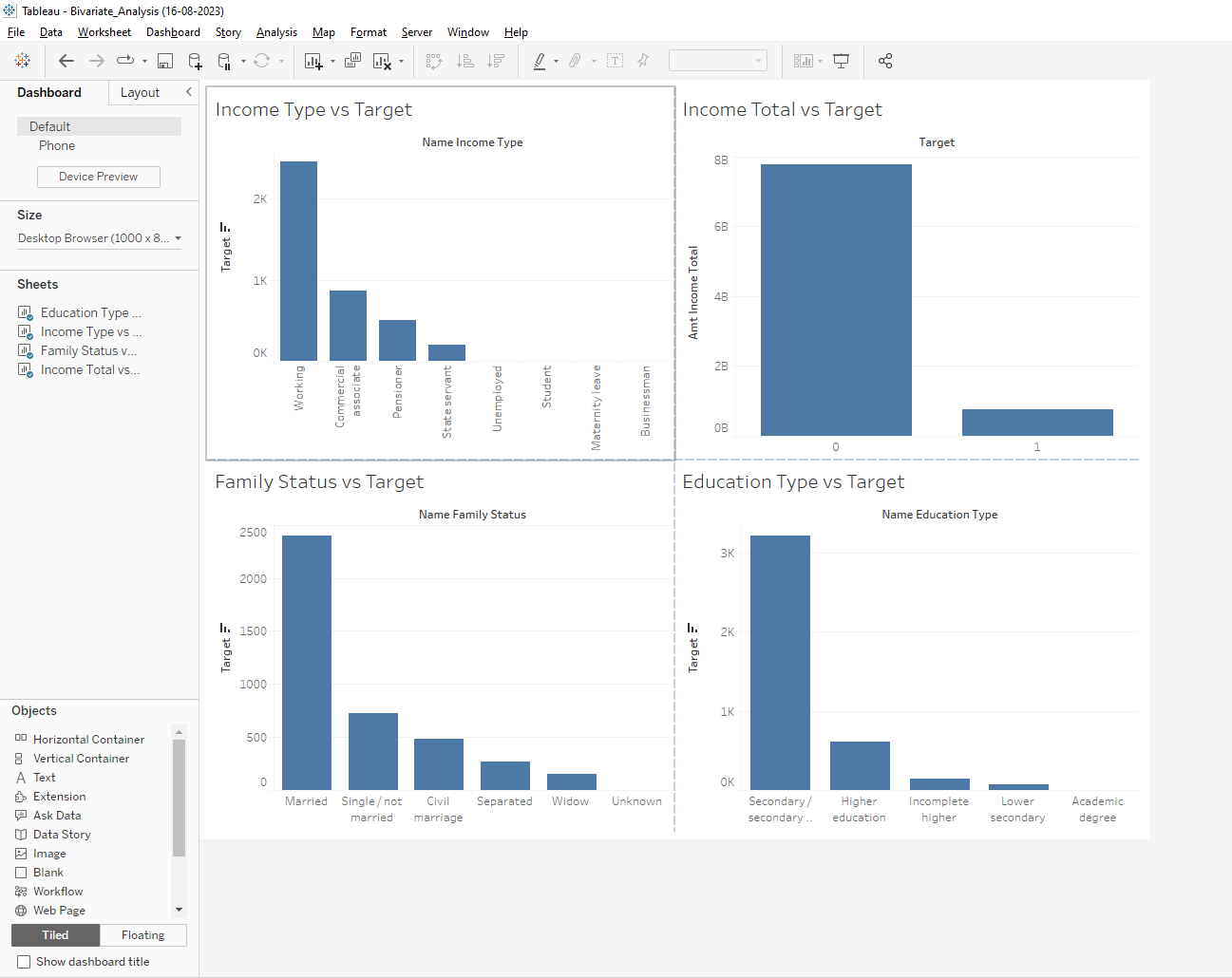




1. Perform Univariate, Segmented Univariate and Bivariate Analysis **:**
2. Bivariate Analysis : As the name suggests, this refers to a relationship between any two variables present in the dataset. The major goal of bivariate analysis is to understand how changes in one of the variables is directly/indirectly associated with changes in another variable. We have used ‘TABLEAU’ for this process to make the data analysis easier. This technique further provides us with the basic trends and patterns between the two variables.

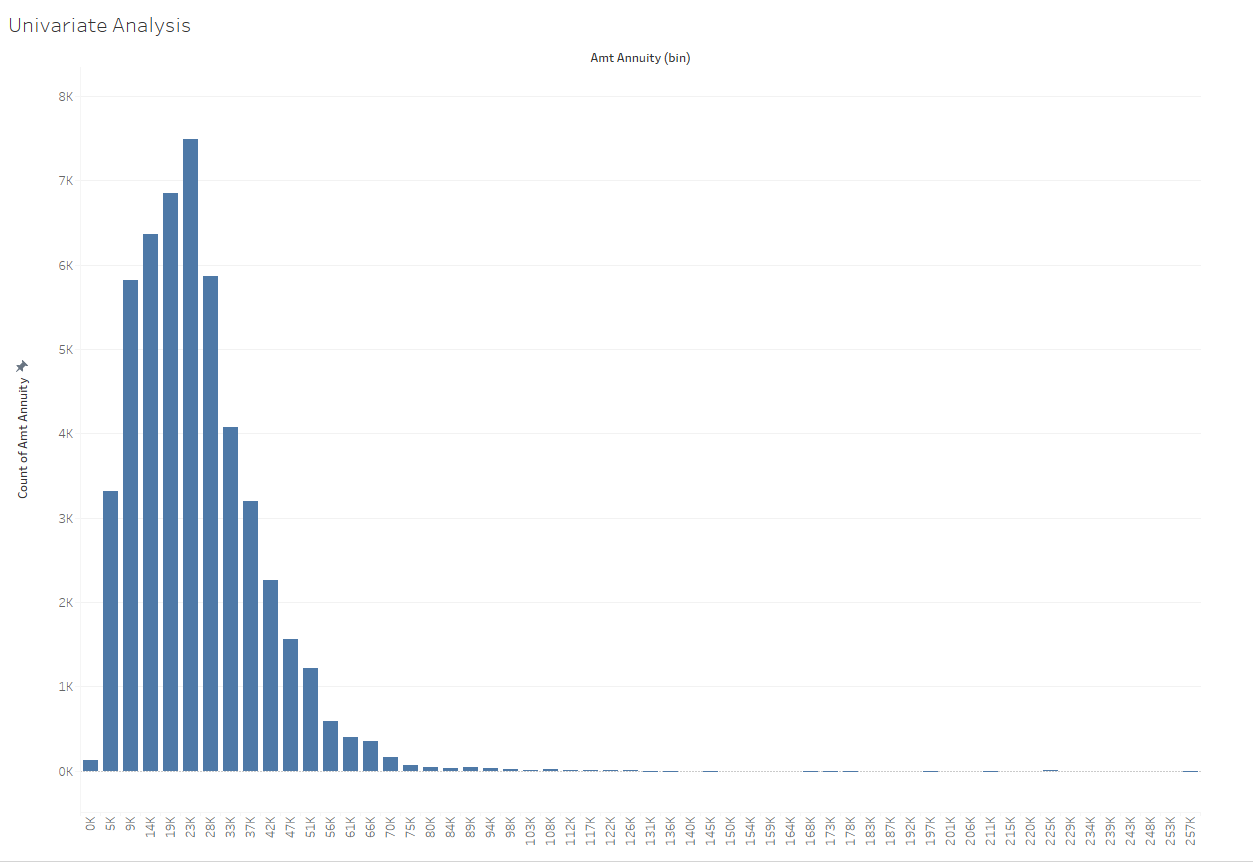
Various relationships we have created between variables are :-

* Education Type VS Target
* Income Type VS Target
* Family Status VS Target
* Total Income VS Target



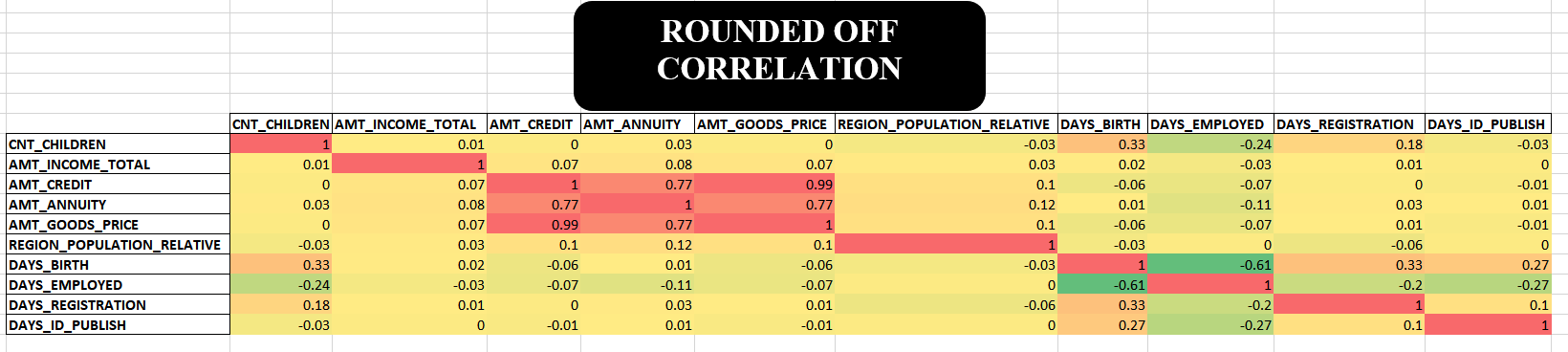
**For a detailed analysis : Link to the workbook :** [**Bivariate Analysis**](https://drive.google.com/file/d/1_kkzAJmqsjP5upxXV0cY1VDX5BEgpqEy/view?usp=sharing)

1. Univariate Analysis :- This is a statistical method which is used to analyse the distribution and characteristics of a single variable in a dataset. This involves calculating the data’s different components such as mean, median, mode, range, variance, etc. We have used ‘TABLEAU’ for this process too to make the data analysis easier.

* Moving on to univariate analysis : We will plot count of Amt Annuity as rows.
* Next, we will convert a copy of Amt Annuity to bins with a bin size of ‘4680’.
* Now plot the Amt Annuity (Bin) onto the column bar.

For detailed analysis, please refer to the univariate analysis workbook :- [Univariate Analysis](https://drive.google.com/file/d/1JdyIHJI1mN5L8LQhFUlMqQUT94pSdPBY/view?usp=drive_link)

1. Identification of top correlations :- Major factors that influence the payments made on time are CNT\_CHIILDREN, AMT\_INCOME\_TOTAL, AMT\_CREDIT, AMT\_ANNNUITY, etc.

* The approach we will follow for this is to compare the correlations between different variables with a CORREL chart.
* The chart with the rounded off values that define the correlation between the major variables :

Link to the final excel file : Application\_data :- [Application\_data\_excelsheet](https://docs.google.com/spreadsheets/d/11dtFVBEL3dqqPH_gM1AXJnMn9t3ilBTy/edit?usp=drive_link&ouid=114249730076600449568&rtpof=true&sd=true)

Now, let’s move on to our net analysis which is based on the next dataset named **previous\_data** which contains the details of the previous applications of the customers :-

1. Identification of missing or inappropriate data :-

* So, first of all we need to identify the unnecessary columns having too many missing values and drop them.

Unnecessary columns :-

AMT\_DOWN\_PAYMENT

WEEKDAY\_APPR\_PROCESS\_START

HOUR\_APPR\_PROCESS\_START

FLAG\_LAST\_APPL\_PER\_CONTRACT

NFLAG\_LAST\_APPL\_IN\_DAY

RATE\_DOWN\_PAYMENT

RATE\_INTEREST\_PRIMARY

RATE\_INTEREST\_PRIVILEGED

SELLERPLACE\_AREA

NAME\_SELLER\_INDUSTRY

DAYS\_FIRST\_DUE

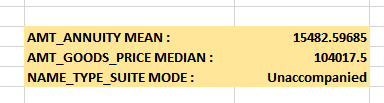
DAYS\_LAST\_DUE\_1ST\_VERSION

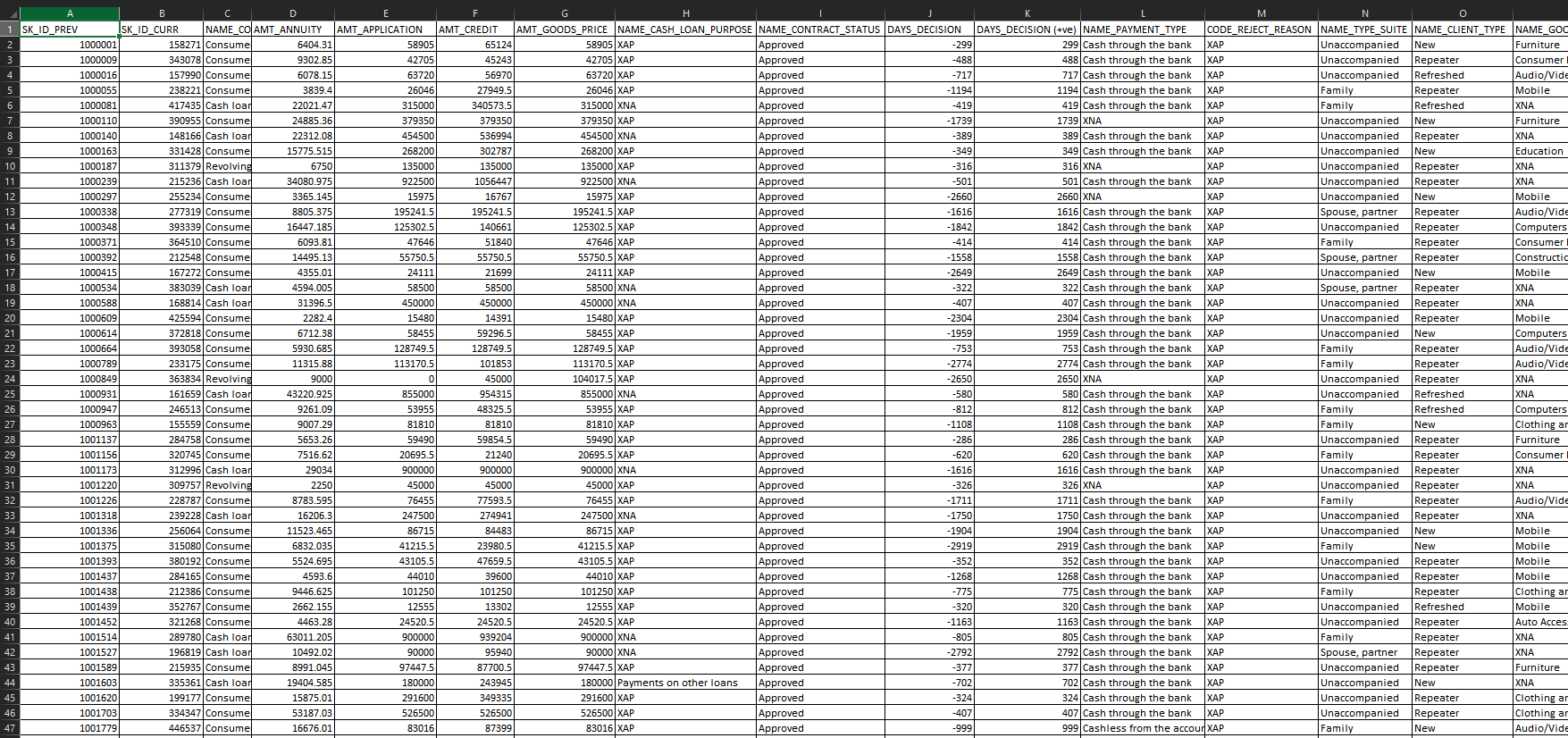
DAYS\_LAST\_DUE

DAYS\_TERMINATION

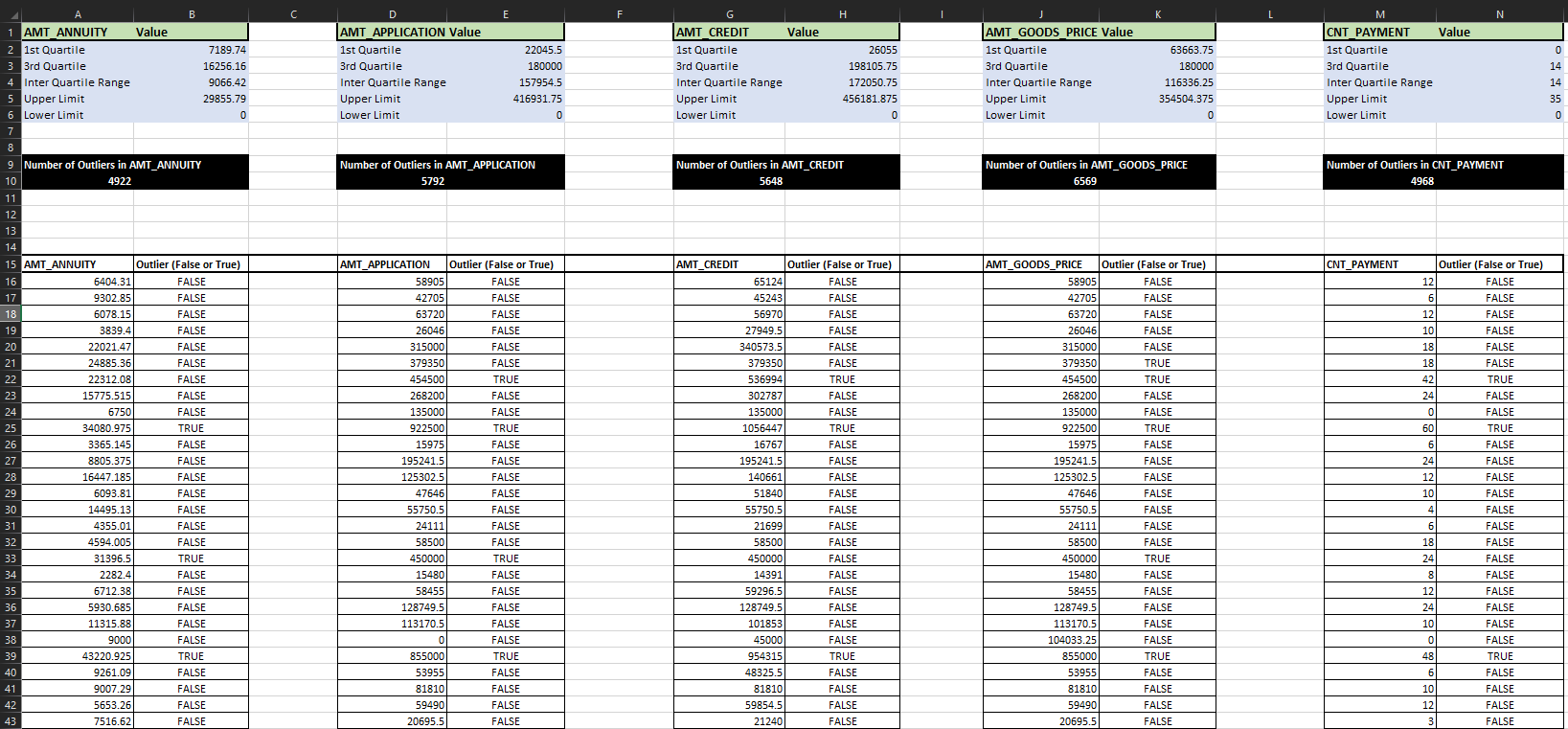
NFLAG\_INSURED\_ON\_APPROVAL

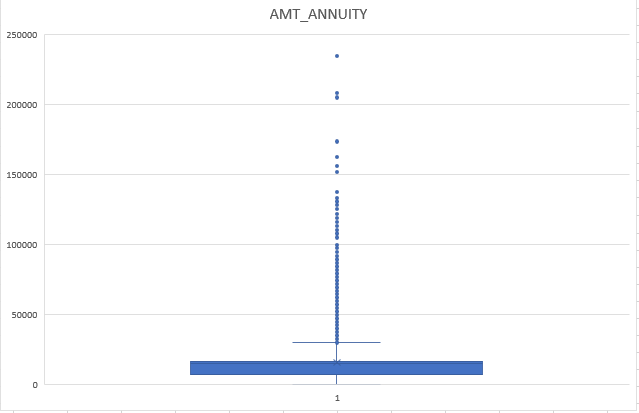
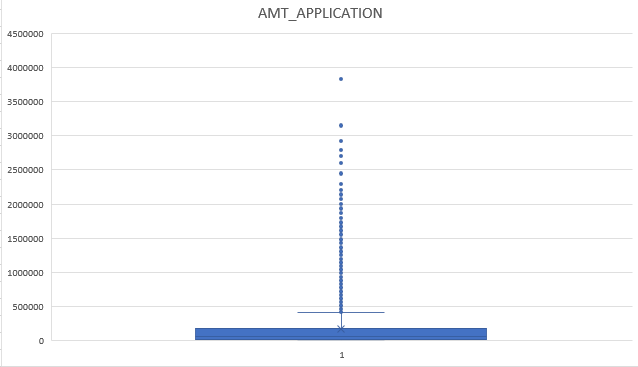
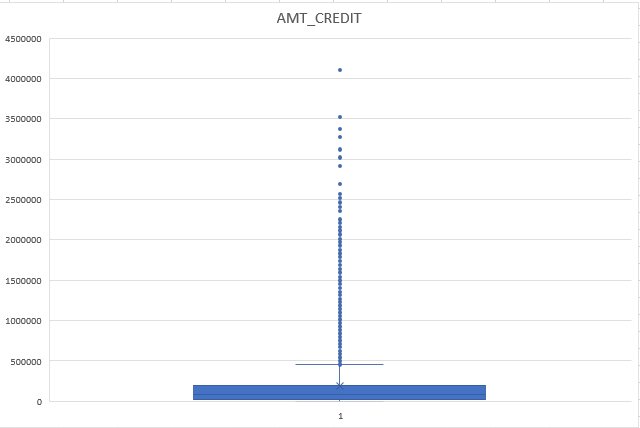
* Next, we will replace all the missing values present in AMT\_ANNUITY, AMT\_GOODS\_PRICE and NAME\_TYPE\_SUITE column with the following values :-

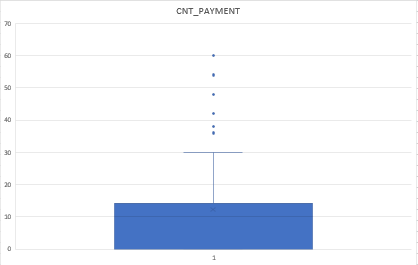


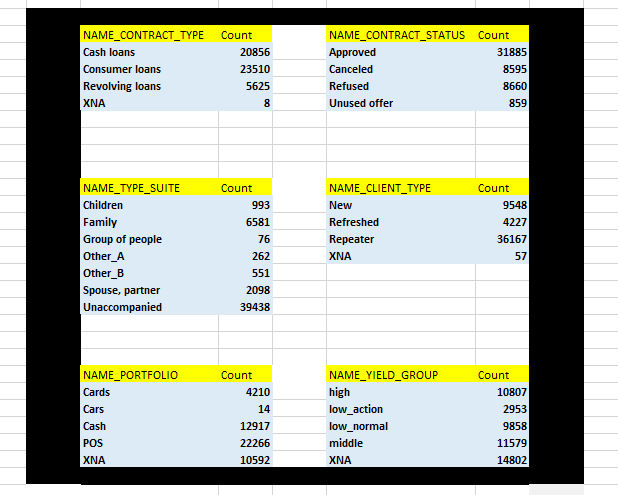
* Now, to fix the days column we need to multiply the values with -1 to convert the negative days into positive values.
* Now, that we have cleaned the dataset and also replaced the necessary values. Let’s proceed with further analysis:-



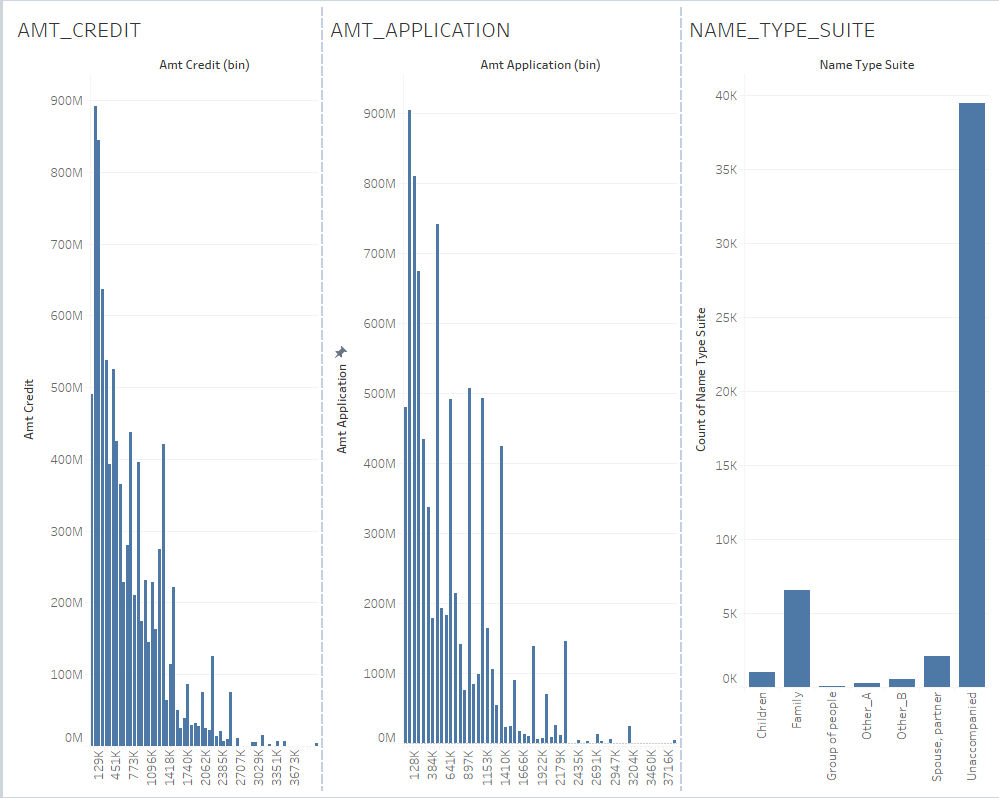
1. Detecting Outliers in the dataset :





1. Data Imbalance :-
2. Univariate Analysis :- This is a statistical method which is used to analyse the distribution and characteristics of a single variable in a dataset. This involves calculating the data’s different components such as mean, median, mode, range, variance, etc. We have used ‘TABLEAU’ for this process too to make the data analysis easier.

* Moving on to univariate analysis : We will plot count of Amt\_Credit, Amt\_Application and Name\_type\_suite as rows.
* Next, we will convert a copy of Amt\_Credit, Amt\_Application and Name\_type\_suite to bins with a bin size of recommended size.
* Now plot the (Bin) quantities onto the column bar.

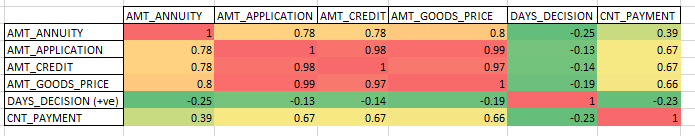


For detailed analysis, please refer to the following tableau workbook :- [Univariate\_Analysis\_Previous\_Applications\_dataset](https://drive.google.com/file/d/1Bqq9MTC9Lfc8LiWfFBnBShSitfmqdoI9/view?usp=drive_link)

1. Identification of top correlations :- Major factors that influence the correlation are :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| AMT\_ANNUITY | AMT\_APPLICATION | AMT\_CREDIT | AMT\_GOODS\_PRICE | DAYS\_DECISION | CNT\_PAYMENT |

* The approach we will follow for this is to compare the correlations between different variables with a CORREL chart.
* The chart with the rounded off values that define the correlation between the major variables :



Applications\_data Excel sheet :-[Applications\_data\_final\_sheet](https://docs.google.com/spreadsheets/d/11dtFVBEL3dqqPH_gM1AXJnMn9t3ilBTy/edit?usp=sharing&ouid=114249730076600449568&rtpof=true&sd=true)

Previous\_applications Excel sheet :- [Previous\_Applications\_Final\_Sheet](https://docs.google.com/spreadsheets/d/1VFD5QFvaEkOAKhNEvlF2rHwTDvXa-X8A/edit?usp=drive_link&ouid=114249730076600449568&rtpof=true&sd=true)

**Learning Outcomes:**

1. **Learnt about advanced excel techniques and tools.**
2. **Learnt about tableau functions and graph visualization.**
3. **Learnt about outliers and strategies to handle them.**
4. **Learnt about Pivot, different types of visualizations graphs and charts.**

**I would like to thank TRAINITY for giving me this wonderful opportunity to learn so much from a single project. Creating a project on BANK LOAN CASE STUDY was a wonderful experience. I got to learn so many things about excel and tableau tools and useful formulas, brainstormed to find solutions whenever errors bulged in and learnt a lot about advanced excel.  
  
The best part of the project was live implementations. I was able to implement all the excel functions and derive various results successfully in the end.**