**Problem Statement:**

* 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.
* Identify if there is data imbalance in the data. Find the ratio of data imbalance.
* Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.
* 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, Target**. And if you have to find the top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, 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.
* Include visualizations and summarize the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the **clients with payment difficulties with all other cases.**

**Understanding, Preparing and observing the data:**

* On the importing and reading the application data we find that we have the 307511 records of the Details of the client loan application i.e. application\_data.csv (record data set)
* On the importing and reading the previous data we find that we have the 1670214 records of the Details of the loan details stored by the company i.e. previous\_data.csv (record data set)
* In Application data we have 122 columns and in previous data we have 37 columns
* On checking the data of the null value percentage for above 50 percent we decide that those are irrelevant data are we drop those columns 41 columns from the application data and 4 from previous data.
* Now on the basis of correlation factor we observed that the Ext columns and Flag columns are not related to target column data. So we are dropping the columns of that from application data.
* We are filling the data of column name CNT\_FAM\_MEMBERS, OCCUPATION\_TYPE, NAME\_TYPE\_SUITE with **mode values** and AMT\_ANNUITY with **mean value** and AMT\_REQ\_CREDIT\_BUREAU, AMT\_GOODS\_PRICE with **median values**
* Changing the days column with positive values as some values are in negative.
* While converting the continuous data to range values we find that particular category has some outliers which are shown below:
  1. AMT\_GOODS\_PRICE 4050000.0
  2. 'AMT\_INCOME\_TOTAL' 117000000.0
  3. 'AMT\_CREDIT' 4050000.0
  4. 'AMT\_ANNUITY' 258025.5
  5. 'DAYS\_EMPLOYED' 365243
  6. 'DAYS\_BIRTH' 7489

**Observation:**

* On the basis of count plot we have decided that cash loans are more than revolving loans.
* On the basis of countplot we can see that for different loan purpose for different contract status.

**Conclusion:**

* People having income less than or equals to 1 ml, are more like to take loans out of which who are taking loan of less than 1.5 million, coudl turn out to be defaulters. we can target income below 1 million and loan maount greater than 1.5 million1. Bank should target the customers having low income i.e. below 1 ml2. 80-90% of the customer who were prev. canceled/refused, are repayers. Bank can do the analysis and can consider to give loan to these segments
* For the above data we can use **Binary Logistic Regression model** as the response is **TARGET** column which is in binary form and remaining predictors values are in continuous as well as categorical data.