**Team Members-**

**1. Mohd Afraaz Firoz Khan Roll no: 220940325042**

**2.Shubham Bharat Chaudhari Roll no:220940325069**

**3.Mandar Manish Ghaisas Roll no:220940325041**

**4.Aishwarya Devdas Bhalbhar Roll no:220940325005**

**5.Rahul Vinayrao Joshi Roll no:220940325053**

**Problem Statement-**

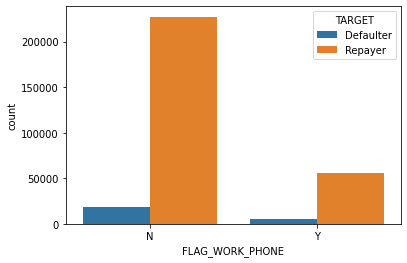
* If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
* If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.
* **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample,
* **All other cases:** All other cases when the payment is paid on time.
* Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.
* Identify the missing data and use appropriate methods to deal with it. (Remove columns/or replace it with an appropriate value)
* 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 ca1.**
* **Cleaning The Data:**
* **1) Upload the csv file and then see the top 5 rows and columns**
* **2)** we observe there are 307511 rows and 122 columns
* 3)Data types of each and every column is know over here
* Here there are int type as well as object and float type
* 4)Find the Statistical information about the data
* 5)Here we can see there are AMT\_ANNUITY has 12 null values which is the lowest
* 6)similarly NAME\_TYPE\_SUITE are 1292 Null values
* 7)OWN\_CAR\_AGE has 202929 Null values

**Handling Null Values**-

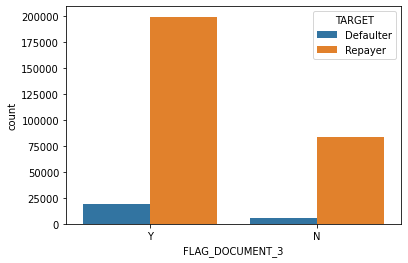
* 8) Here AMT\_ANNUITY, AMT\_GOODS\_PRICE, CNT\_FAM\_MEMBERS has the least null values, thus we can fill it with median
* 9)then we create a function to find null values for the dataframe
* 10)create a variable null\_vals for storing null columns having missing values more than 30%
* 11) These are the columns having more than 30% of null value decided to delete it.
* 12)Then Create the backup to, retrieve original dataset incase of emergency
* 13)Dropping the columns having more than 50% of Null values
* 14)Then from the columns dictionary we can conclude that only 'OCCUPATION\_TYPE','EXT\_SOURCE\_3 looks relevant to the TARGET column.
* thus we drop all other columns except 'OCCUPATION\_TYPE','EXT\_SOURCE\_3
* 15)now we will deal with null values more than 15%
* 16)We Check Correlation of EXT\_SOURCE\_3 , EXT\_SOURCE\_2.
* 17)Now Starting with EXT\_SOURCE\_3 , EXT\_SOURCE\_2. As they have normalised values, now we will understand the relation between these columns with TARGET column using a heatmap
* 18) Then putting irrelevant columns in variable "irrev"
* 19)As there doesn't seem to be correlatiob=n between the EXT\_SOURCE\_3 , EXT\_SOURCE\_2 and the target thus dropping these columns too
* 20)Then we put all the flag columns in a single list.therefore the amount of flag columns are 28.
* 21)we Start with EXT\_SOURCE\_3 , EXT\_SOURCE\_2. As they have normalised values, now we will understand the relation between these columns with TARGET column using a heatmap
* 22)then put irrelevant columns in variable "irrev"
* 23)As there doesn't seem to be correlatiob=n between the EXT\_SOURCE\_3 , EXT\_SOURCE\_2 and the target thus dropping these columns too
* Successfully handled null valueOccupation type being important factor, thus instead of dropping the rows of occupation type, we have decided to fill it with a common value
* Checking for claim-
* 24) after that replacing "0" as repayer and "1" as defaulter for TARGET column as stated in column description replacing "1" as Y being TRUE and "0" as N being False
* 25)we Successfully converted the 0 and 1 to Y and N

# **Performing BIVARIATE Analysis**

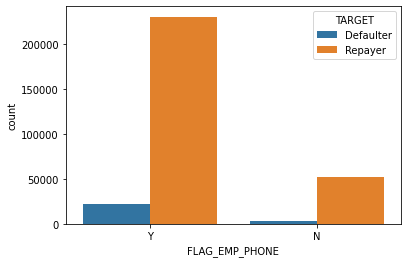
1. FLAG\_WORK\_PHONE vs TARGET
2. FLAG\_DOCUMENT\_3 vs TARGET
3. FLAG\_EMP\_PHONE vs TARGET



conclusion:We can see that People doesn't have a work phone seem to be able to pay as compared to people who have a work phone.



Conclusion is that There are more Repayers with the people having Document\_3



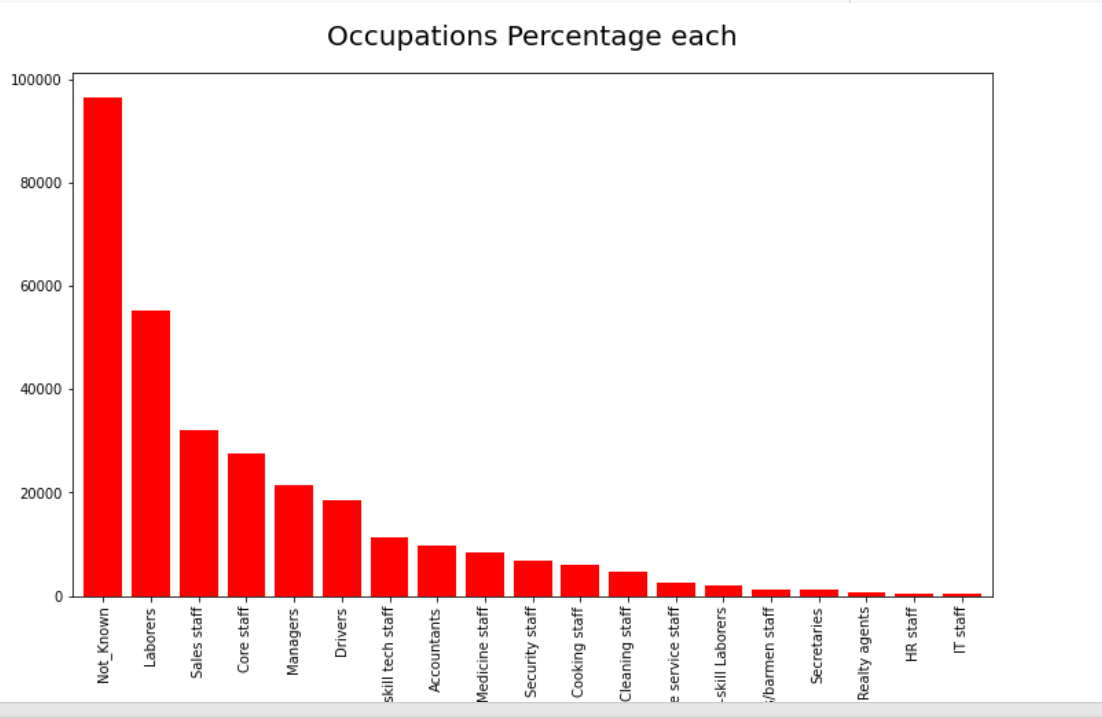
conclusion:We can conclude that Repairs are more where people have Phone

26)Here FLAG\_DOCUMENT\_3 FLAG\_WORK\_PHONE FLAG\_EMP\_PHONE SEEMS TO CORRELATE WITH THE DATA THUS KEEPING IT REMOVING OTHERS then after dropping irrelevant flag terms from the data

# Univariate Analysis -> Looking into Occupation

# Finding percentage of people belonging to each occupation

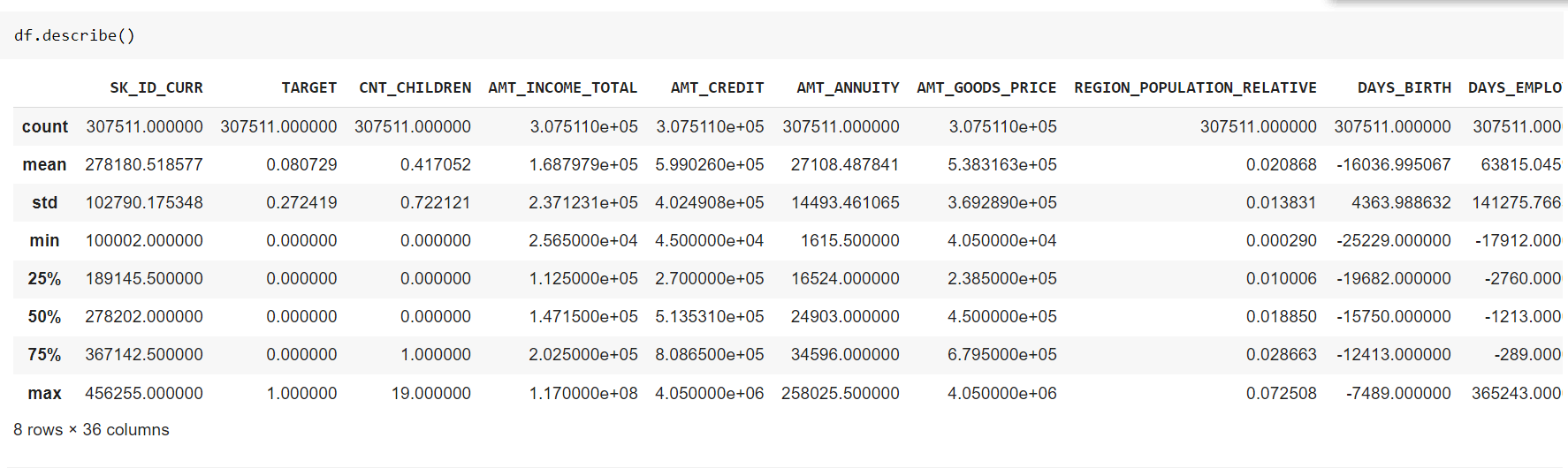
* 27) Then after Plotting a percentage graph having each category of "OCCUPATION\_TYPE".This graph represents the amount of people in each profession, here we can see that the amount of people working as Labourers are More as compared to others.
* 28) Occupation type being an important factor, thus instead of dropping the rows of occupation type, we have decided to fill it with a common value.



Conclusion:

* As we can conclude, the majority of people's occupation is unknown, dropping such rows would have resulted in shortage of dataset, thus a common value is fair enough to be replaced.This graph represent the amount of people in each profession, here we can see that amount of people working as a Labours are more as compared to others.
* **Looking For Outliers-**

Checking for the five point summary:



* **From the above command we find out**

**1.count**

**2.mean**

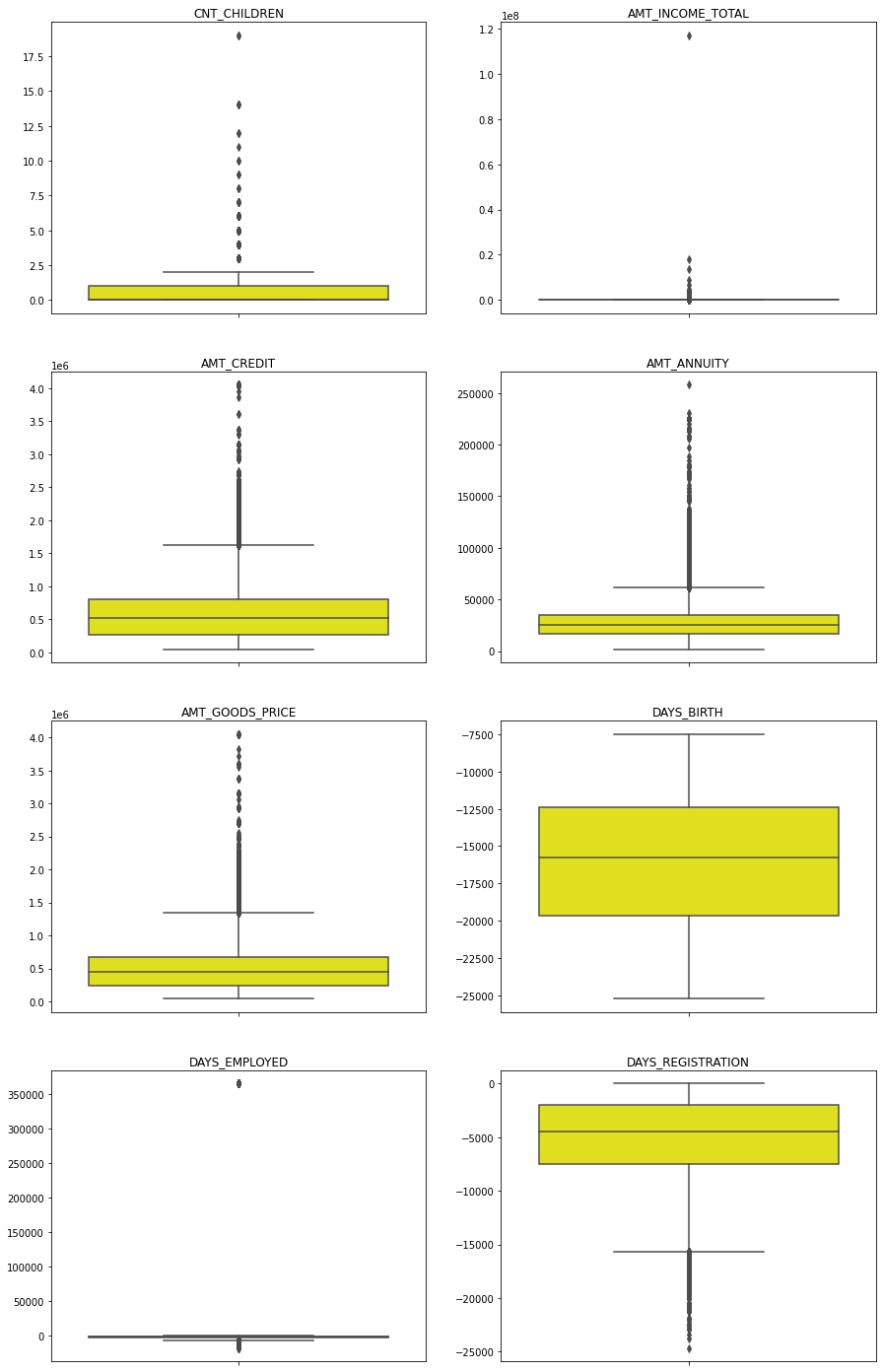
**3.std**

**4.25%**

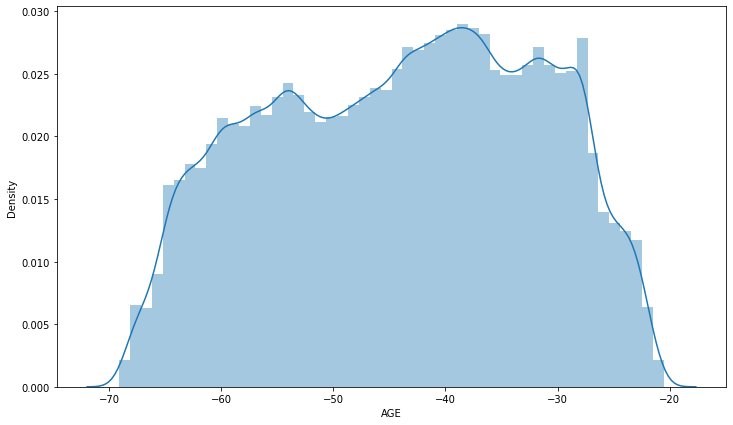
**5.50%**

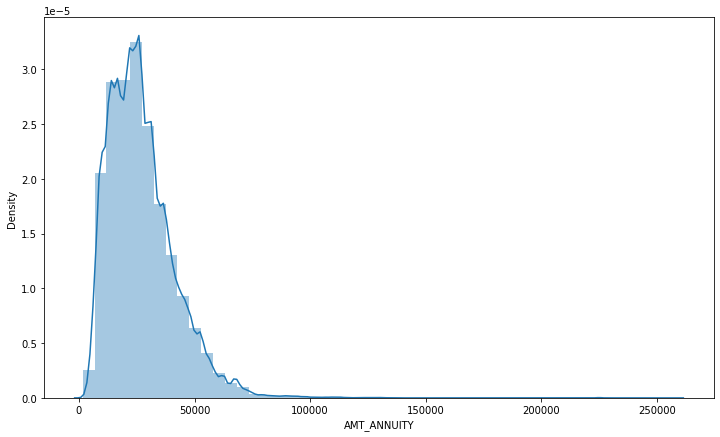
**6.75%**

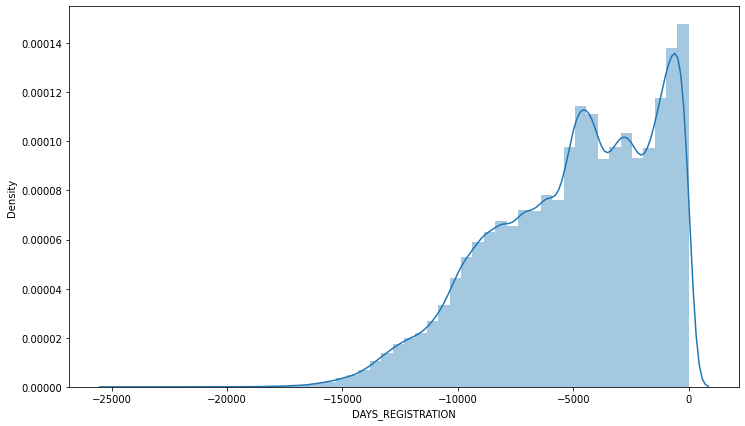
**7.max**

* ****
* **Observations:**
* AMT\_ANNUITY, AMT\_CREDIT, AMT\_GOODS\_PRICE,CNT\_CHILDREN have some number of outliers.
* AMT\_INCOME\_TOTAL has a huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
* **DAYS\_BIRTH shows no outliers so the data is reliable**
* DAYS\_EMPLOYEED HAS AN OUTLIER AT 350000 I.E 350000/365 = 958.9041095890411 years which is not possible, hence it can be a
* Chance variation.

# Data imbalance

* **Identify if there is data imbalance in the data. Find the ratio of data imbalance-**
* 
* Checking for variation in particular column-

* 
* After plotting the displot we can conclude that AMT\_ANNUITY is right skewed graph which means it is positive



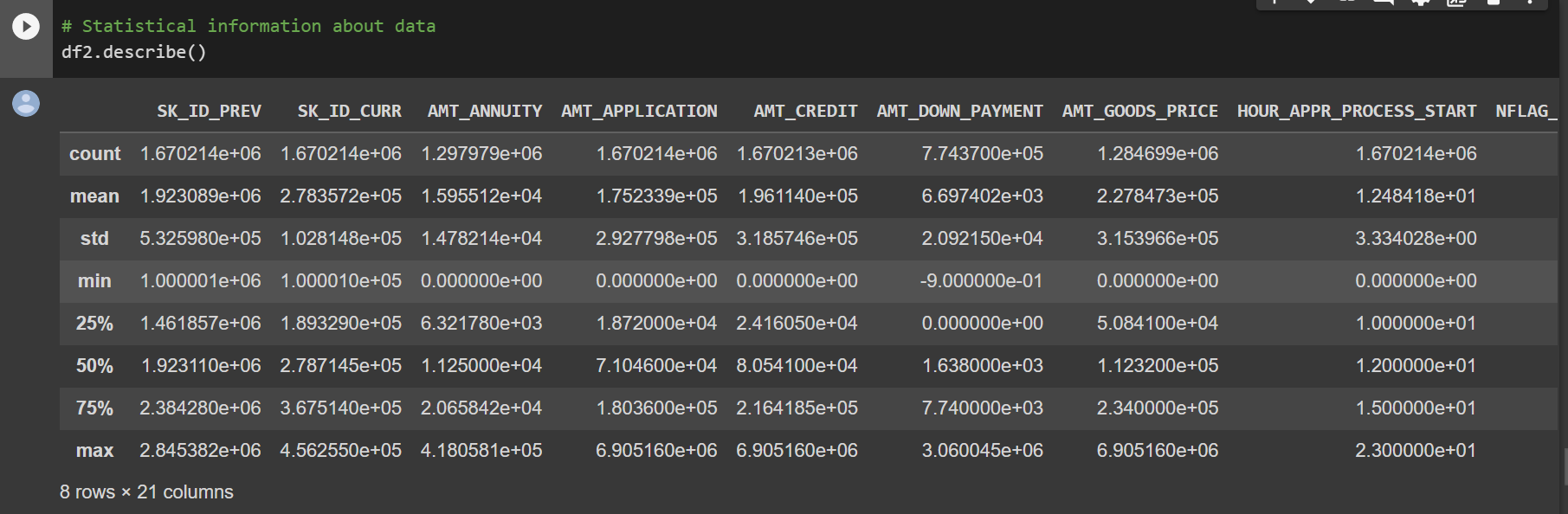
* After plotting the displot we can conclude that DAYS\_REGISTRATION is left skewed graph which means it is Negative

**Dataset -> Previous Application:**

importing second dataset

name = previous\_application.csv

* Observation:
* Dataset has 1670214 and 37 columns



* **Observation:**
* #Some columns has 99.64 % of null values
* #To verify the claim
* df2['RATE\_INTEREST\_PRIVILEGED'].isnull().sum()
* There are only 4 columns where the null values are more than 50%
* dropping null columns having missing values more than 50%
* creating a variable p\_null\_col\_50 for storing null columns having missing values more than 15%
* These are the rows that have more than 15% null values

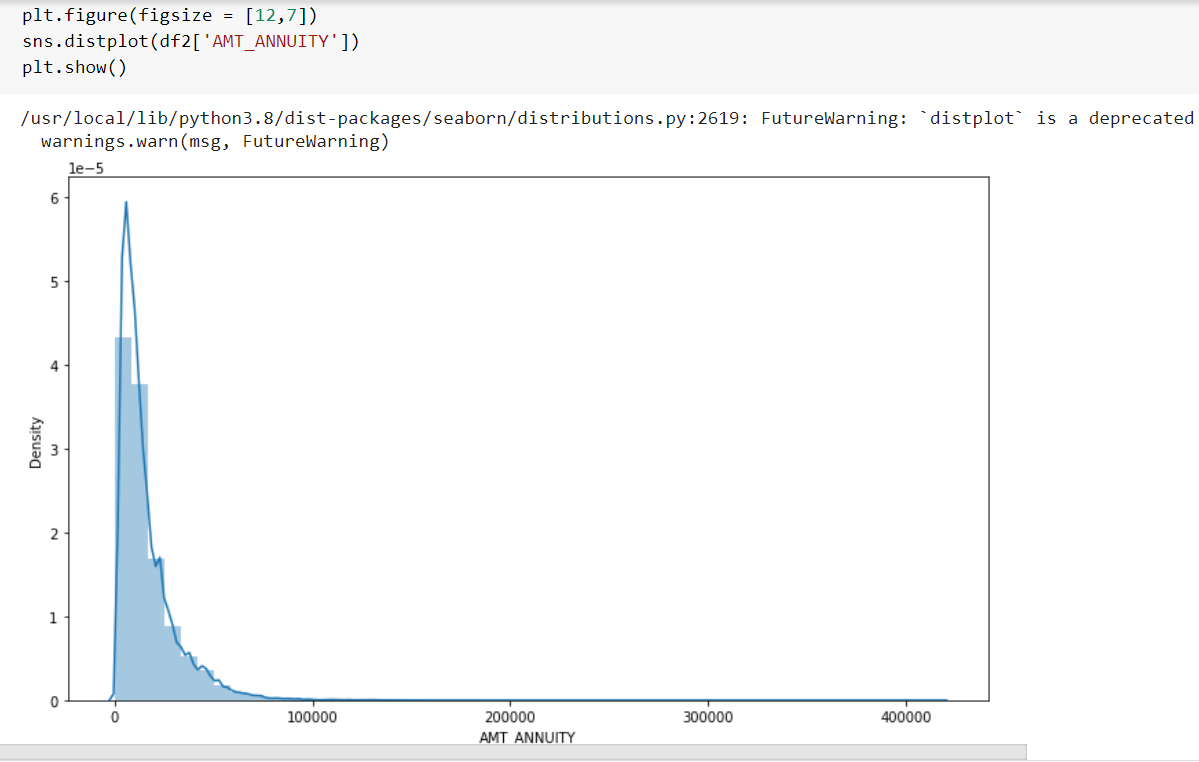
**Checking for the null values of**

AMT\_GOODS\_PRICE

AMT\_ANNUITY

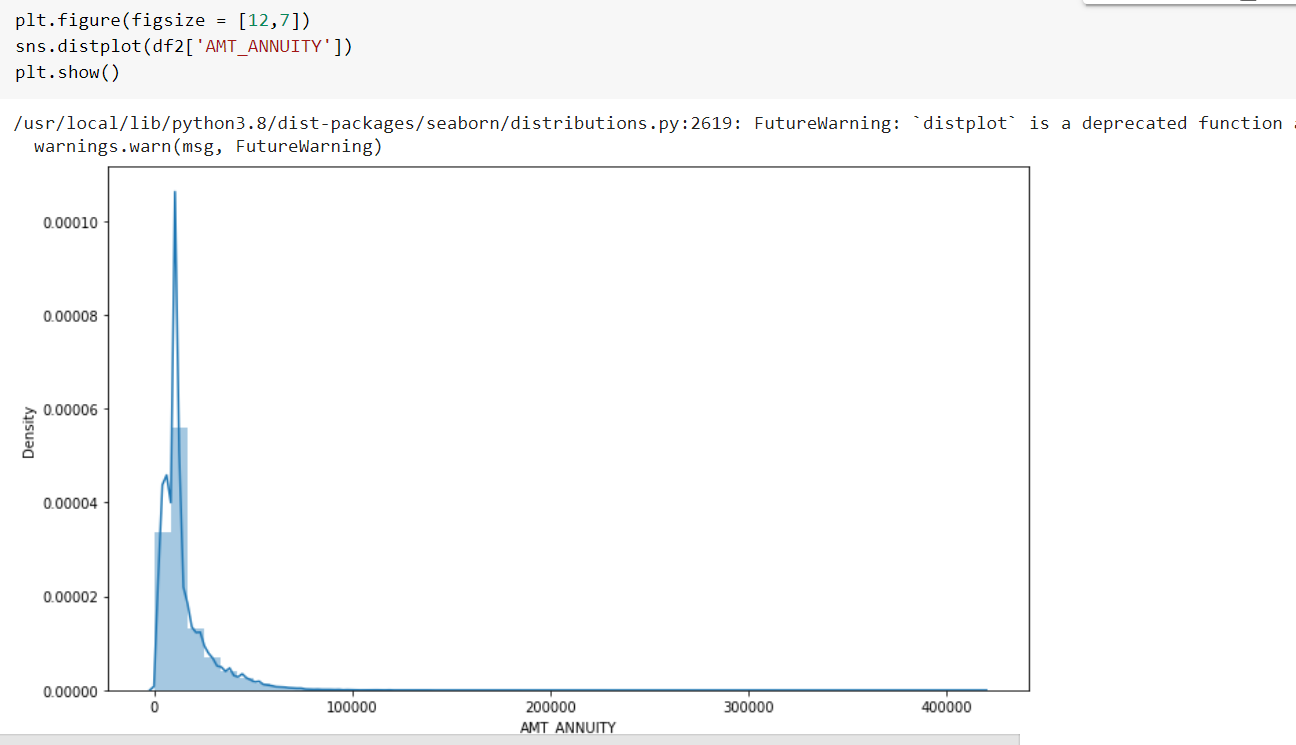
CNT\_PAYMENT

* Listing down columns which are not required

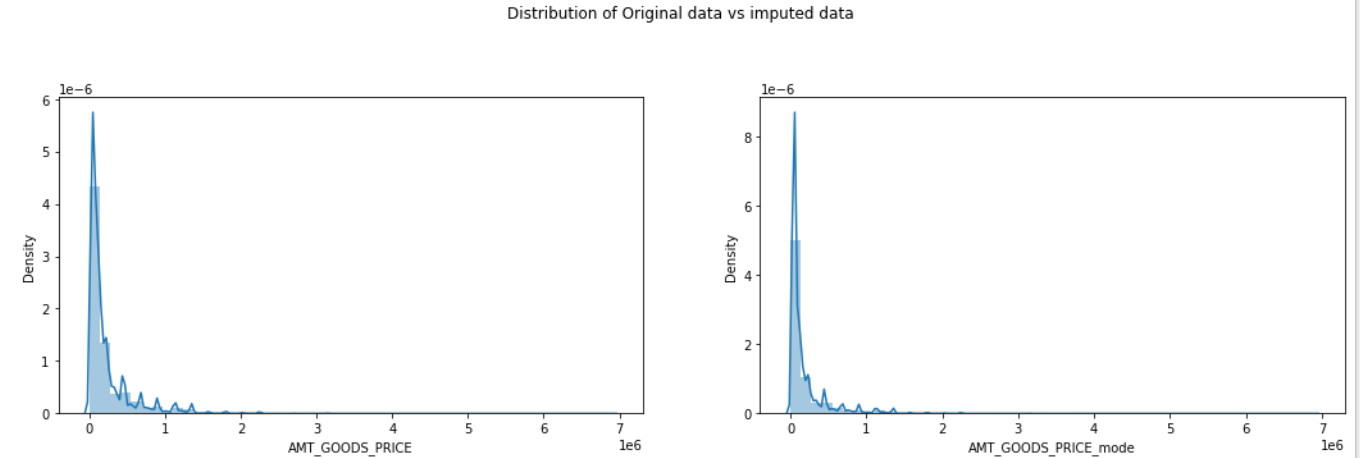
****

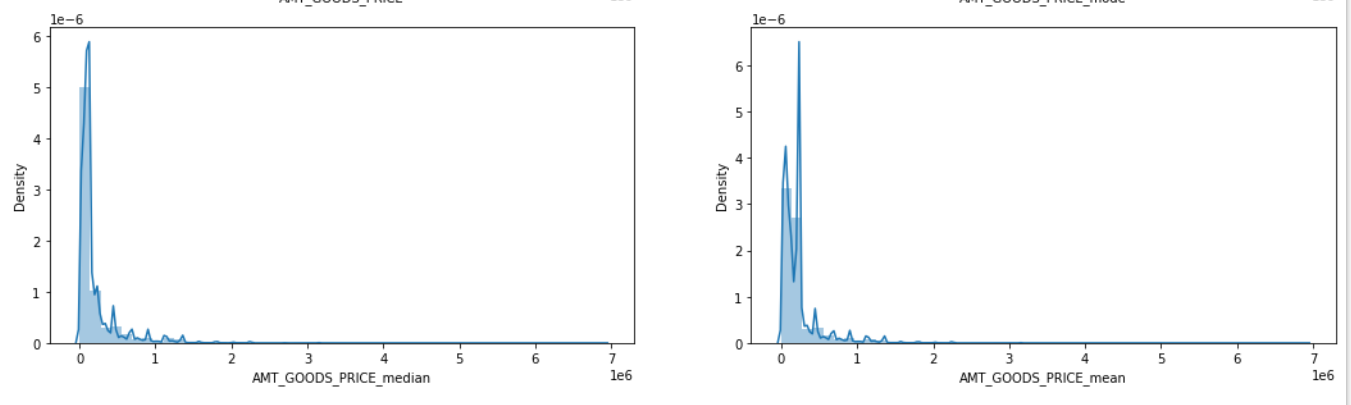
**There is a single peak at the left side of the distribution and it indicates the presence of outliers and Hence imputing with mean would not be the right approach and hence imputing with median.then after replacing null values with median**Rahul Joshi

**#replacing null values with median**

****

**After replacing the null values with median values we plotted the distplot for AMT\_ANNUITY with density and variance is found to be less.**

****

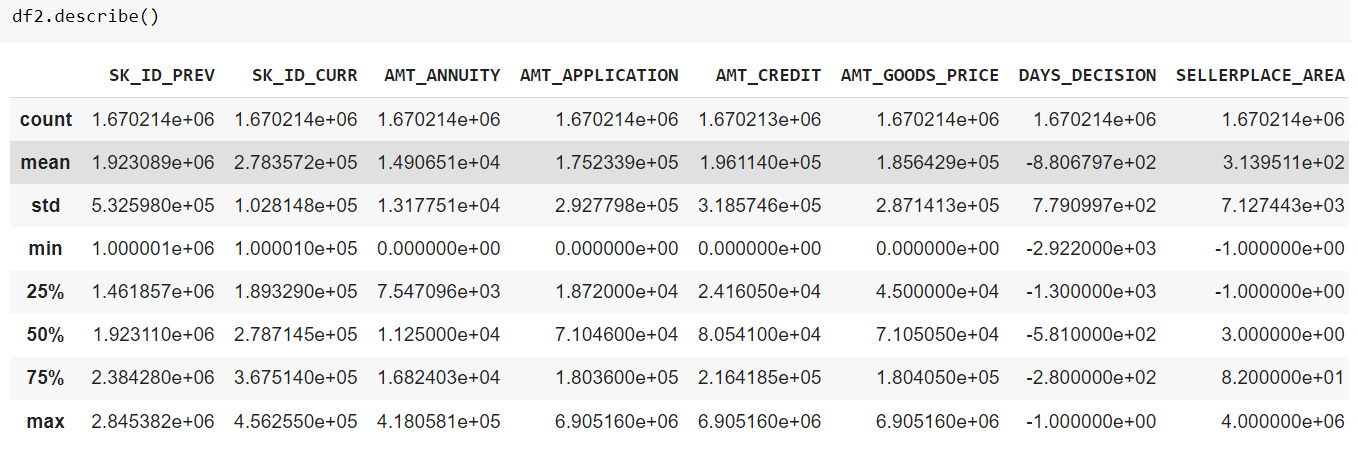
****

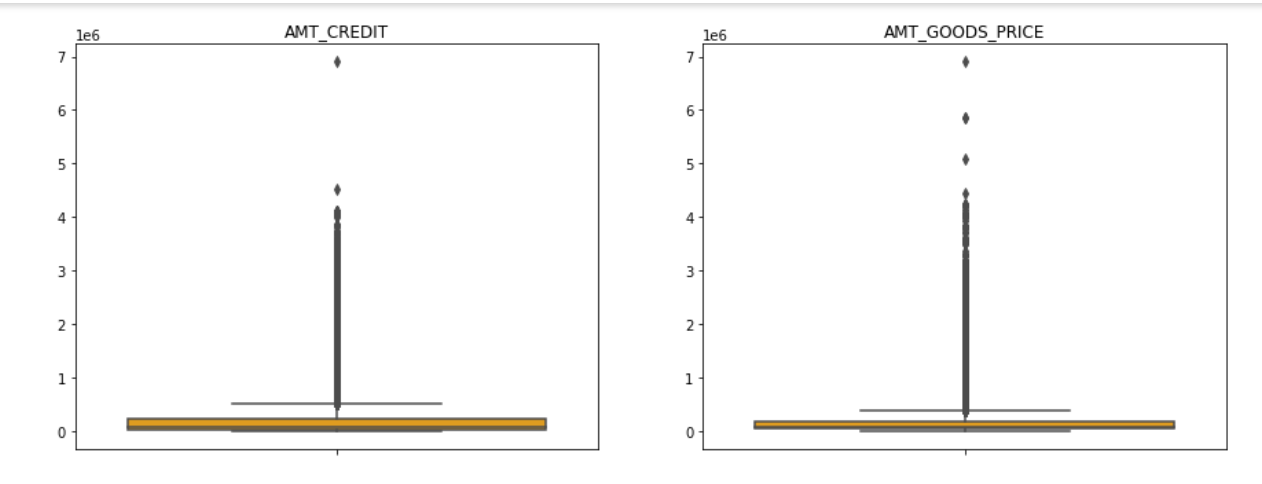
**The original distribution is closer with the distribution of data imputed with mode in this case, thus will impute mode for missing values**

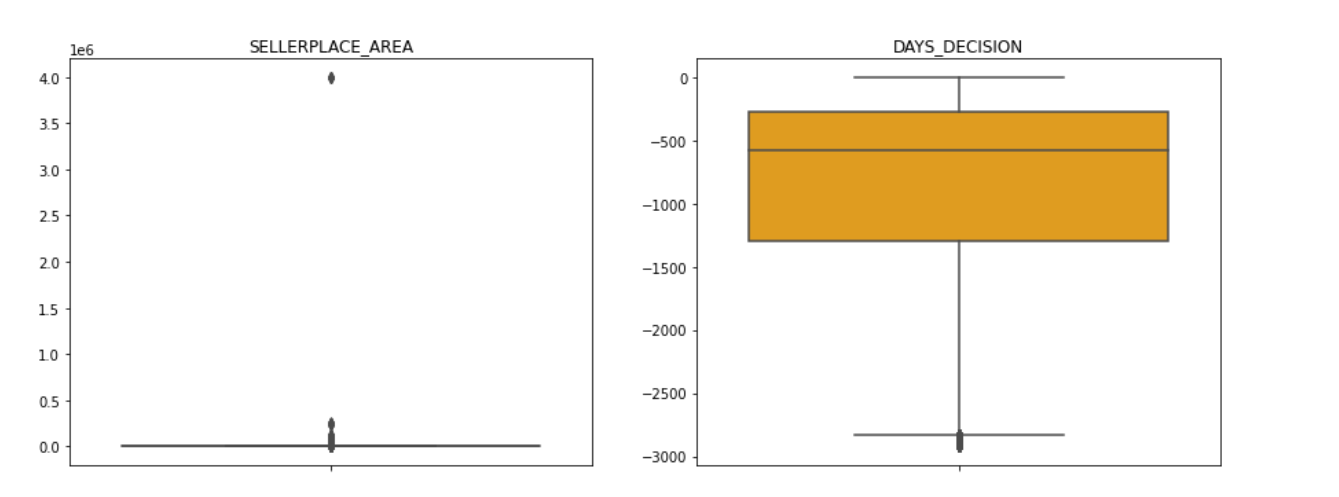
**ImputingFinding Outliers null values with mode**

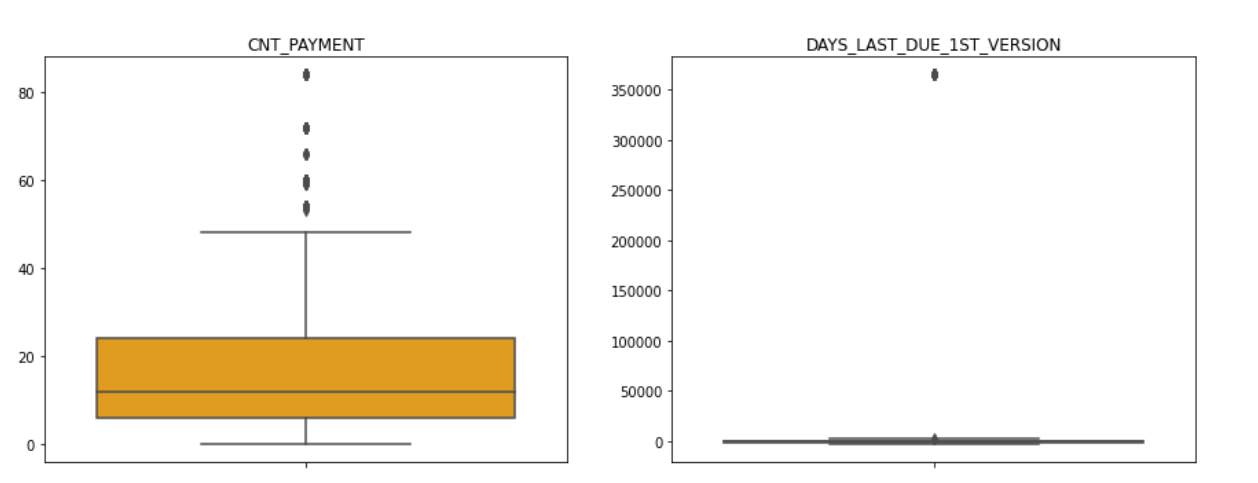
**successfully handled all the null values in AMT\_GOODS\_PRICE column**

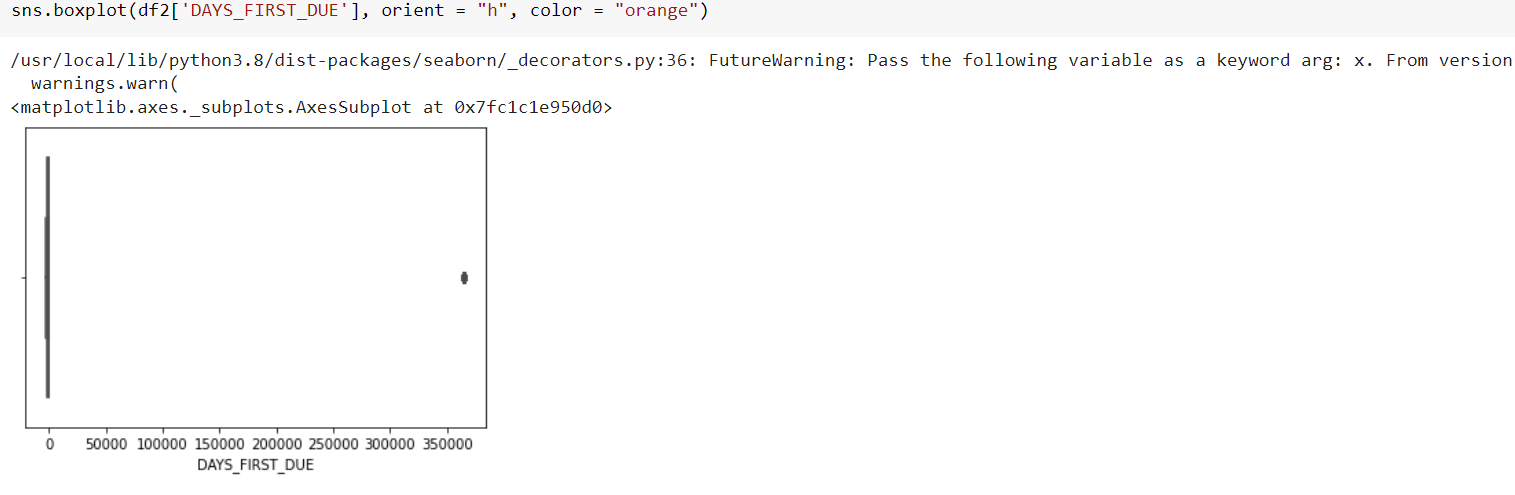
# Finding Outliers

****

****

****

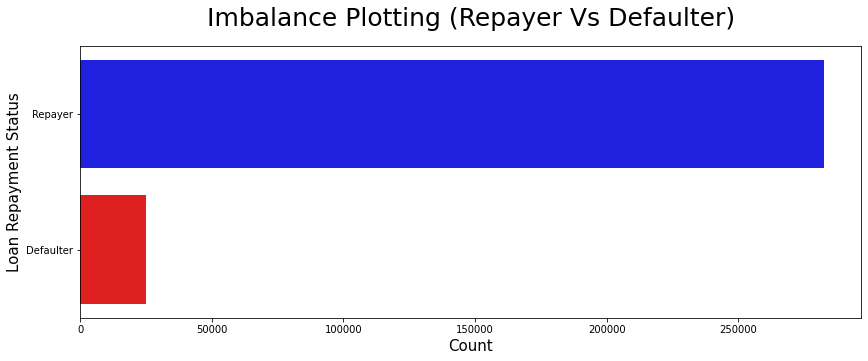
****

****

**can be seen that in previous application data**

**AMT\_ANNUITY, AMT\_APPLICATION, AMT\_CREDIT, AMT\_GOODS\_PRICE, SELLERPLACE\_AREA have a huge number of outliers. CNT\_PAYMENT has few outlier values. DAYS\_DECISION has little number of outliers indicating that these previous application decisions were taken long back.**

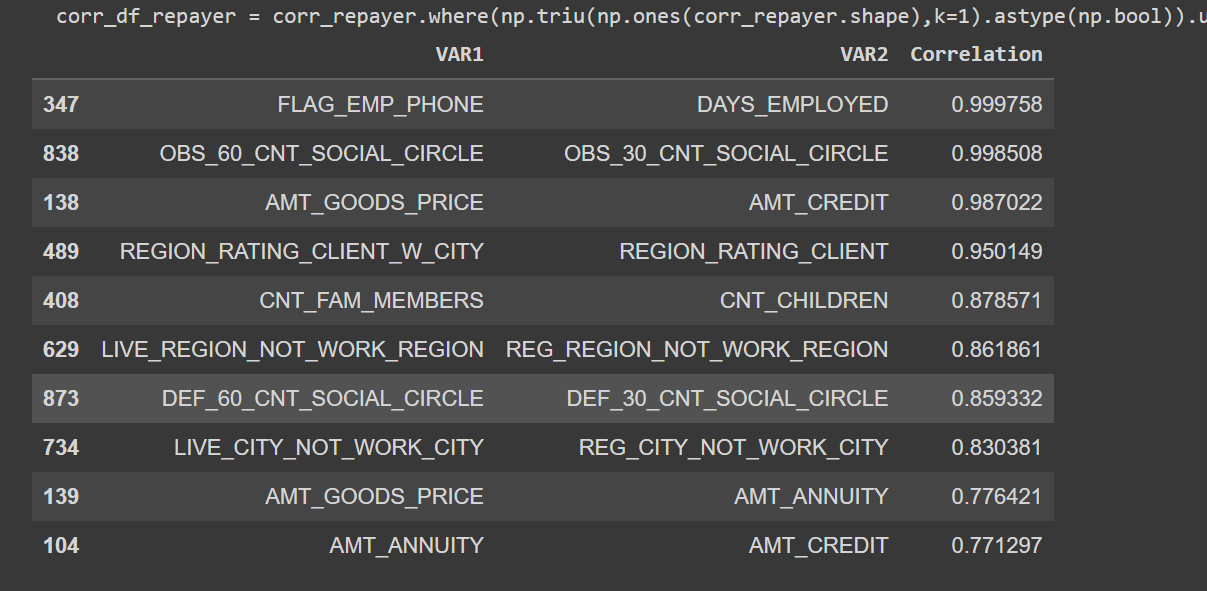
# Imbalance Data

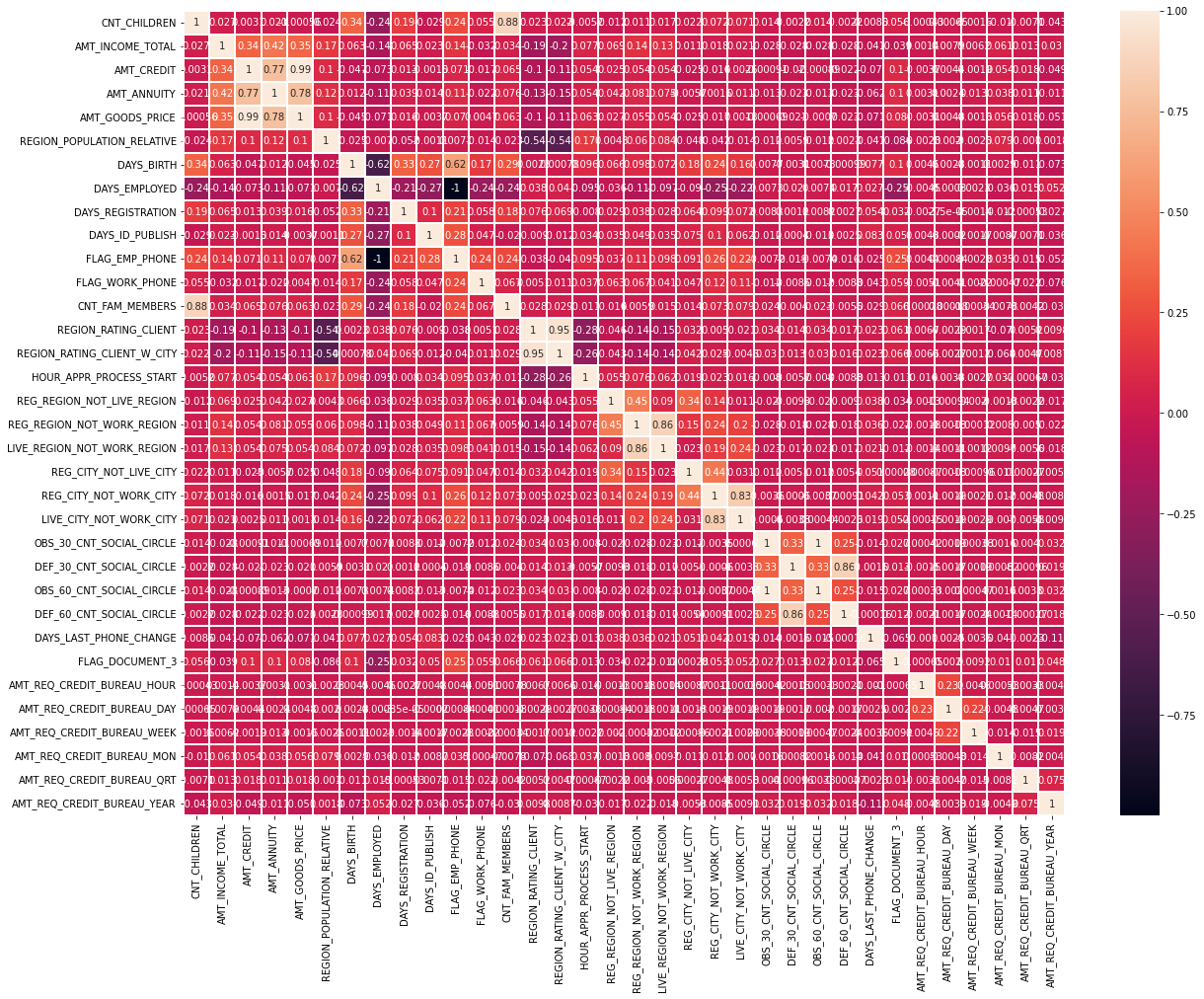
****

**It can be seen that Businessman income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a Businessman could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs**

**#plotting heat map to see linear correlation among Repayers**

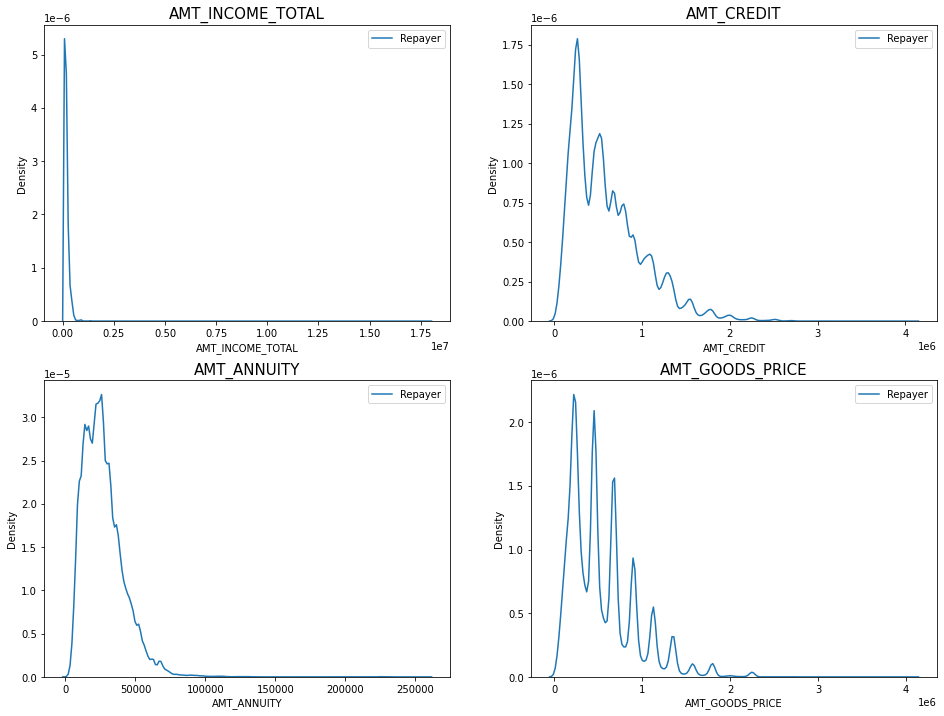
**Getting top 10 correlation for the Repairs dataframe**

****

****

**Plotting the numerical columns related to amount as distribution plot to see density**

**sns.distplot(Defaulter\_df[i[1]], hist=False,label ="Defaulter")**

****

**Observation:**

**Most no of loans are given for goods price below 10 lakhs Most people pay annuity below 50K for the credit loan Credit amount of the loan is mostly less than 10 lakhs The repayers and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision**

**Machine Models can be used are:**

**On basis of the database, we conclude that, Supervised Machine Learning Models can**

**Binary Logistic Linear Regression**

**Details:**

**As the output of the model is in 0 and 1 as repayer and Defaulter, Binary Regression can predict the output as either customer can be Repayer or defaulter based on its data given.**

**Decision Tree**

**Details:**

**Decision tree Can predict the output, with internal nodes representing the whole dataset and leaf nodes being the predictions.**

**Random Forest**

**Details:**

**Combination of various Decision trees can be used for prediction using Random Forest, However the the computation time required to train such Random Forest Model is greater as compared to Decision tree but the prediction and the accuracy will be greater than descision tree.**

**K-NN Classifier**

**Details:**

**K-NN is the K-Nearest Neighbour Classifier, keeps track of similarity in the dataset and classifies new data by comparing it with already present data.**

**Conclusion:**

**1. Banks Should Focus on the People having FLAG\_DOCUMENT\_3, they seem to be the majority in Repayers as compared to others.**

**2. The percentage of IT professions applying for loans is less than as compared to other professions**

**Thus, Bank can focus on other professions rather than IT sectors**

**3. As we can see, Counts of females is more than as compared to males in repayers**

**4. Banks can focus more on married people because the Count of Married people is more than as compared to others in repayers.**