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.**

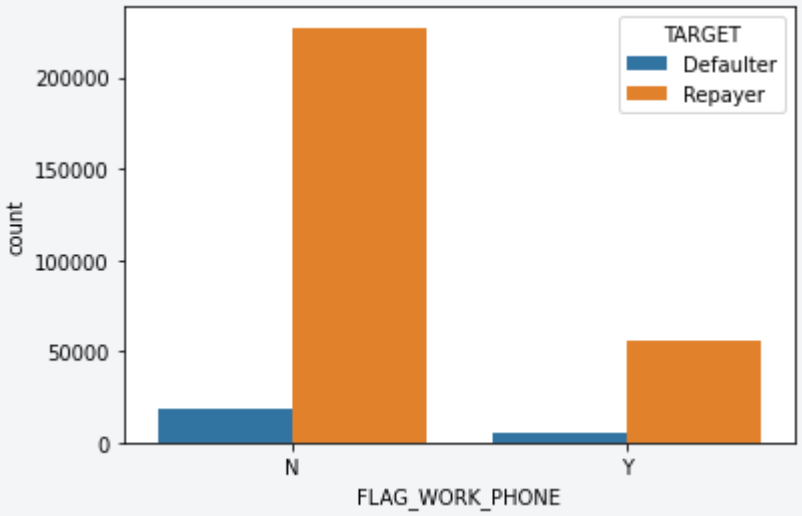
- **Cleaning The Data:**
-
- **1) Upload the csv file and then see the top 5 rows and columns**
- **2)** we observe there are 307511 rows and 122 columns
- 3) Data types of each and every column is know over here
- Here there are int type as well as object and float type
- 4) Find the Statistical information 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-

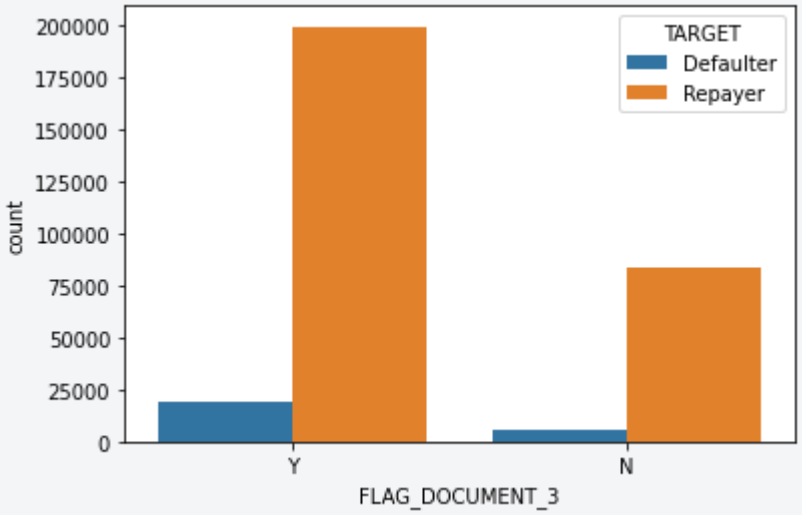
-
- 8) Here AMT_ANNUITY, AMT_GOODS_PRICE, CNT_FAM_MEMBERS has the least null values, thus we can fill it with median
- 9) then we create a function to find null values for the dataframe
- 10) create a variable null_vals for storing null columns having missing values more than 30%
- 11) These are the 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

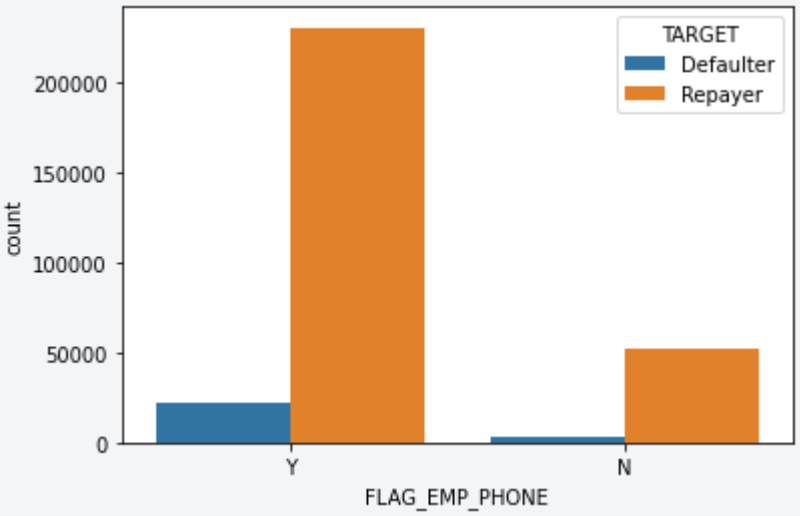
1. FLAG_WORK_PHONE vs TARGET
2. FLAG_DOCUMENT_3 vs TARGET
3. FLAG_EMP_PHONE vs TARGET



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



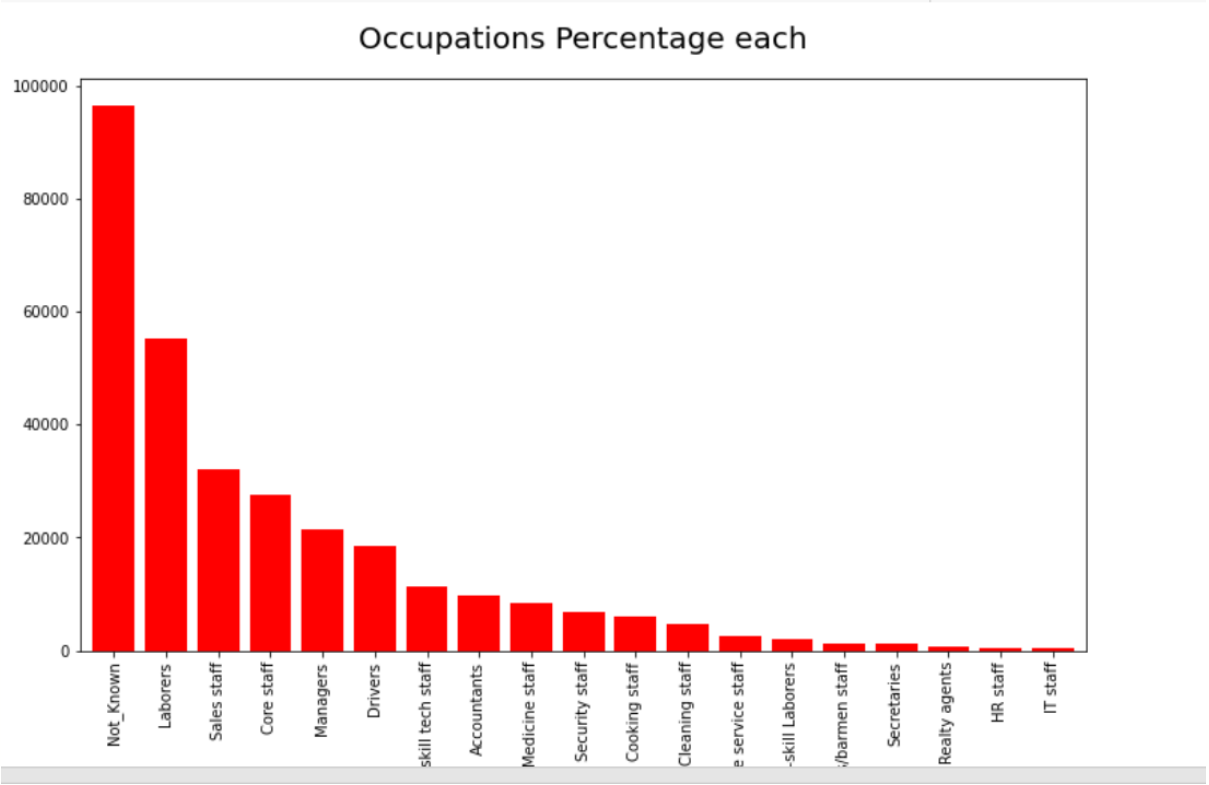
conclusion:We can conclude that Repairs are more where people have Phone

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:

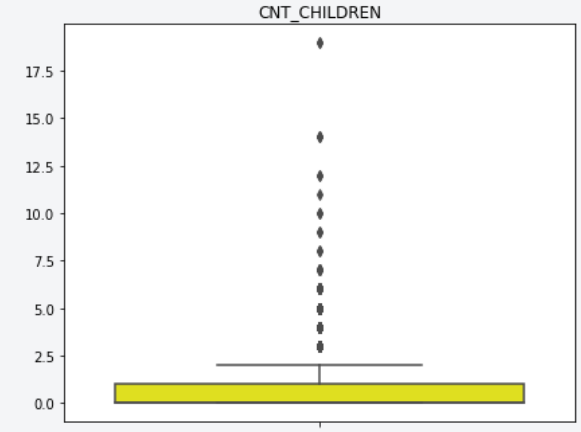
```
df.describe()
```

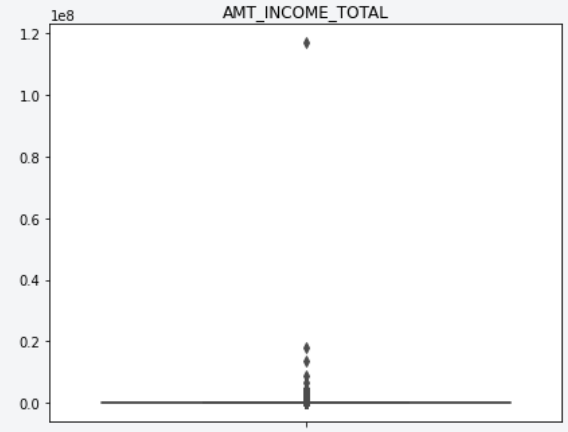
	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLO
count	307511.000000	307511.000000	307511.000000	3.075110e+05	3.075110e+05	307511.000000	3.075110e+05	307511.000000	307511.000000	307511.000
mean	278180.518577	0.080729	0.417052	1.687979e+05	5.990260e+05	27108.487841	5.383163e+05	0.020868	-16036.995067	63815.045
std	102790.175348	0.272419	0.722121	2.371231e+05	4.024908e+05	14493.461065	3.692890e+05	0.013831	4363.988632	141275.766
min	100002.000000	0.000000	0.000000	2.565000e+04	4.500000e+04	1615.500000	4.050000e+04	0.000290	-25229.000000	-17912.000
25%	189145.500000	0.000000	0.000000	1.125000e+05	2.700000e+05	16524.000000	2.385000e+05	0.010006	-19682.000000	-2760.000
50%	278202.000000	0.000000	0.000000	1.471500e+05	5.135310e+05	24903.000000	4.500000e+05	0.018850	-15750.000000	-1213.000
75%	367142.500000	0.000000	1.000000	2.025000e+05	8.086500e+05	34596.000000	6.795000e+05	0.028663	-12413.000000	-289.000
max	456255.000000	1.000000	19.000000	1.170000e+08	4.050000e+06	258025.500000	4.050000e+06	0.072508	-7489.000000	365243.000

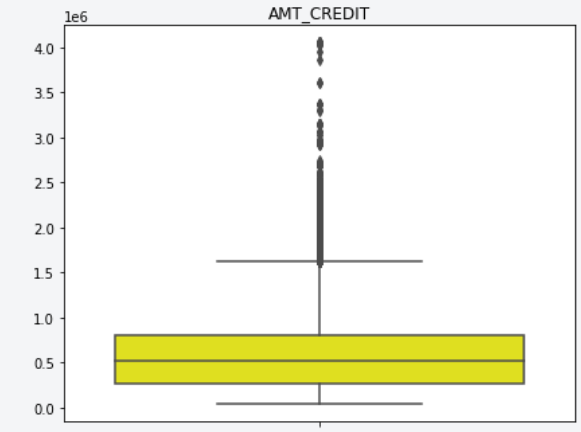
8 rows × 36 columns

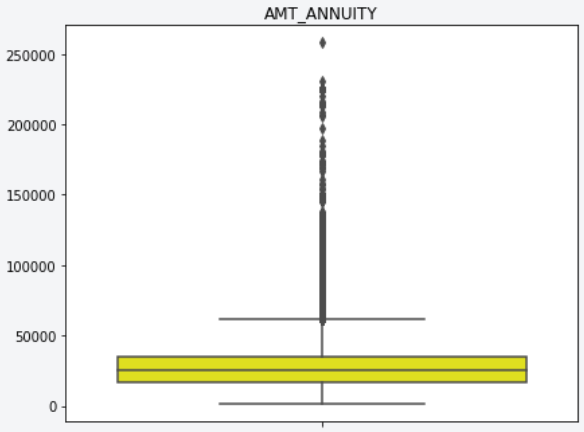
- **From the above command we find out**

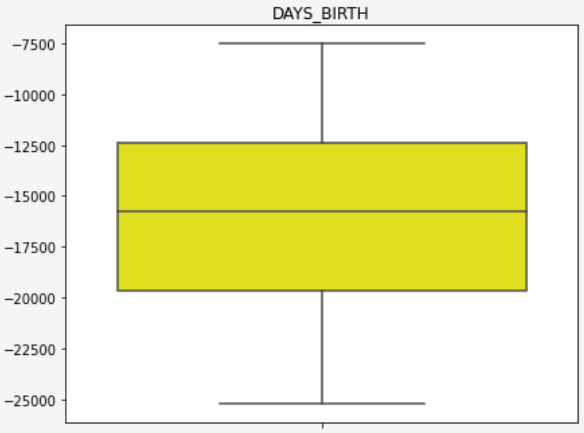
1. count
2. mean
3. std
4. 25%
5. 50%
6. 75%
7. max

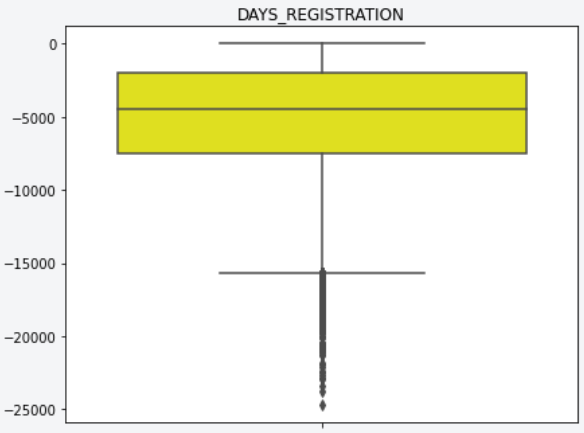










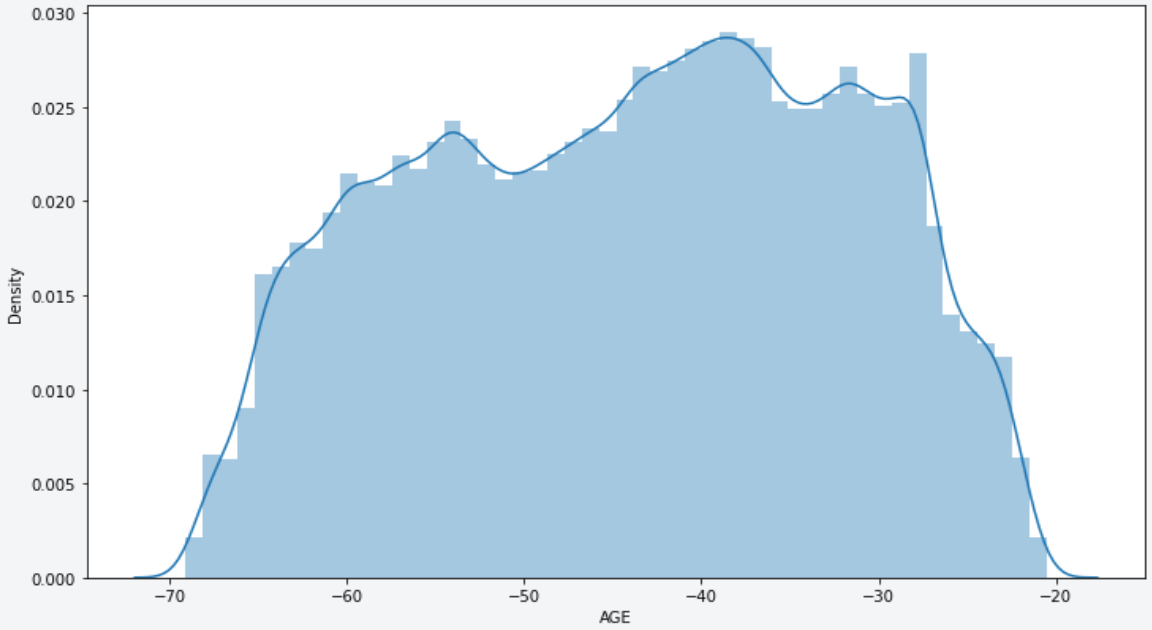


-
-
- **Observations:**

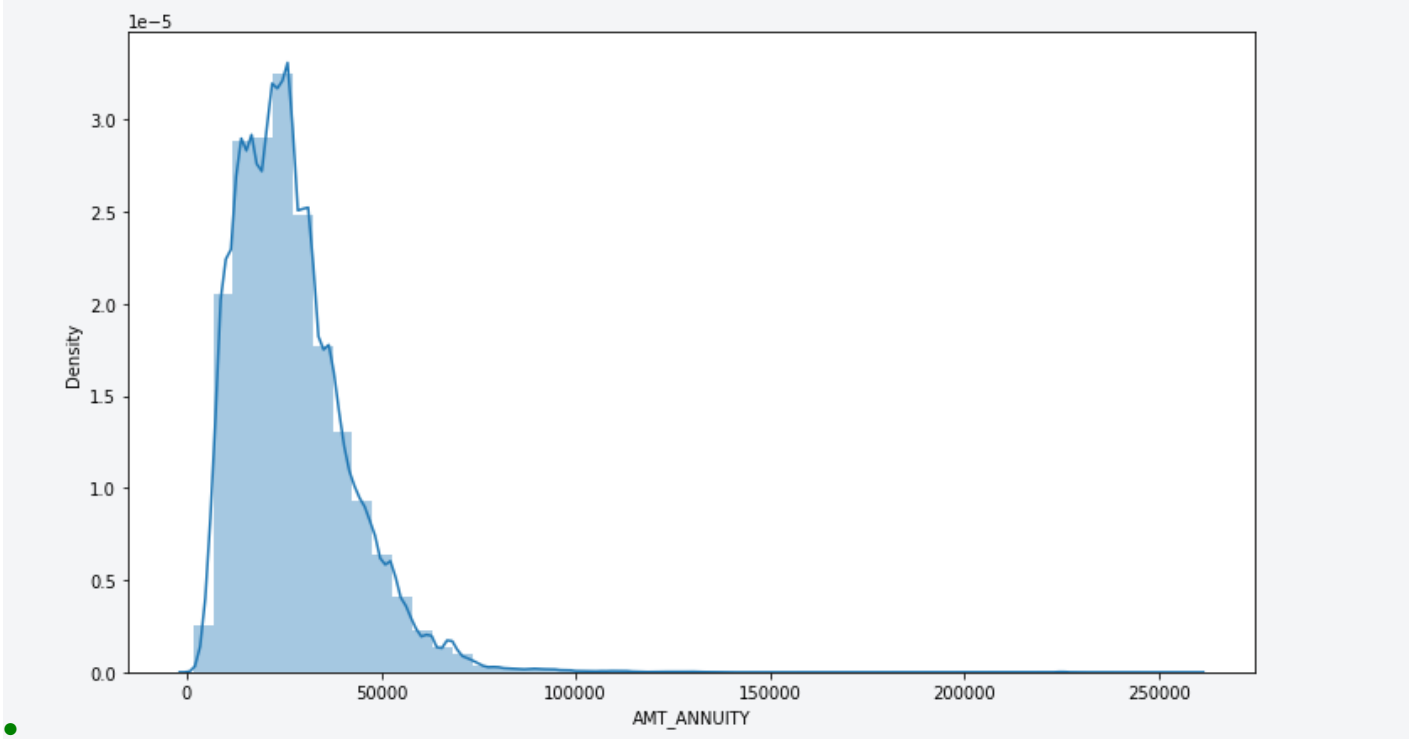
- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
- AMT_INCOME_TOTAL has a huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
- **DAYS_BIRTH shows no outliers so the data is reliable**
- DAYS_EMPLOYEED 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

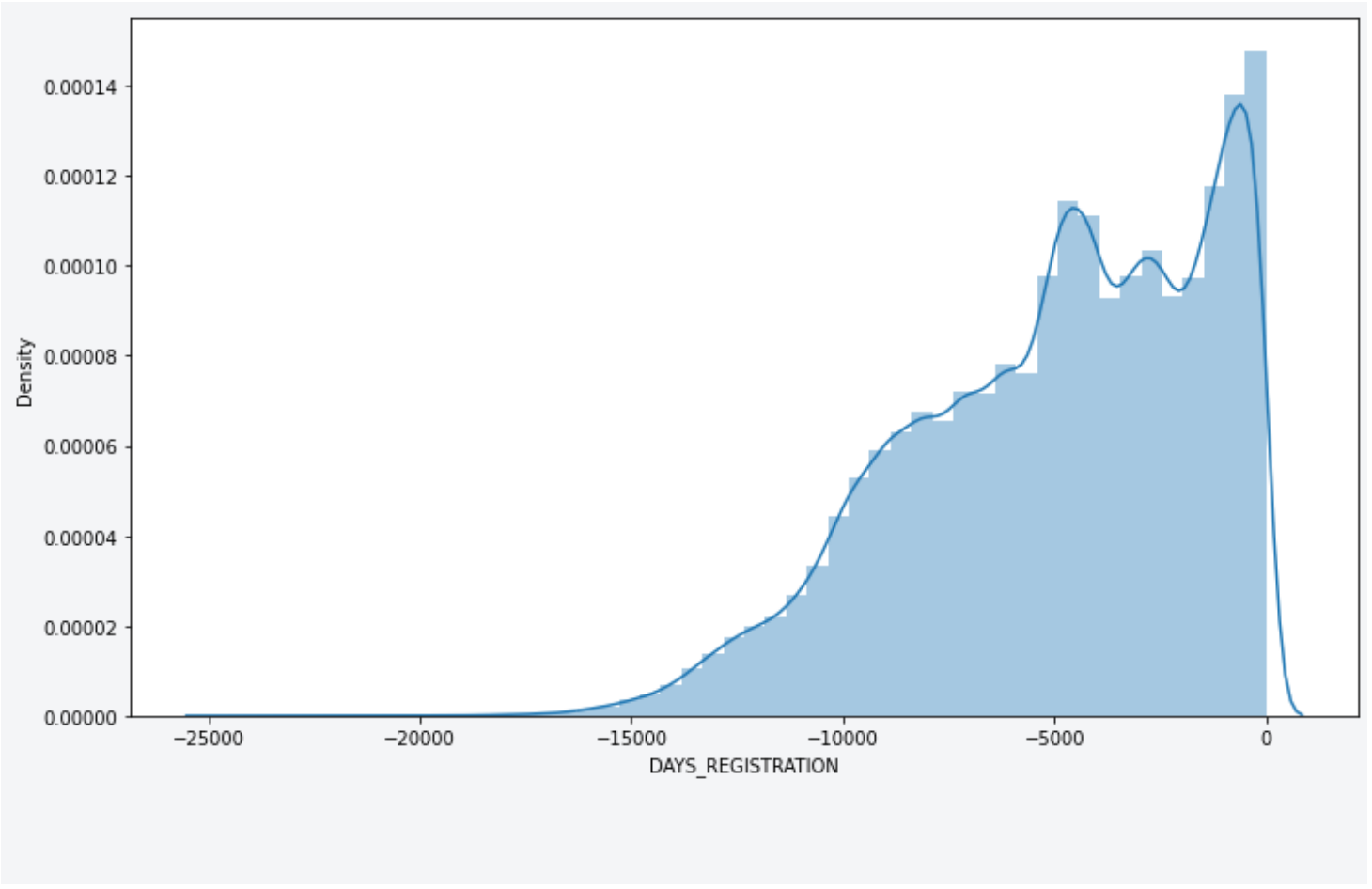
- **Identify if there is data imbalance in the data. Find the ratio of data imbalance-**



- 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

```
# Statistical information about data
df2.describe()
```

	SK_ID_PREV	SK_ID_CURR	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	HOUR_APPR_PROCESS_START	NFLAG_
count	1.670214e+06	1.670214e+06	1.297979e+06	1.670214e+06	1.670213e+06	7.743700e+05	1.284699e+06	1.670214e+06	
mean	1.923089e+06	2.783572e+05	1.595512e+04	1.752339e+05	1.961140e+05	6.697402e+03	2.278473e+05	1.248418e+01	
std	5.325980e+05	1.028148e+05	1.478214e+04	2.927798e+05	3.185746e+05	2.092150e+04	3.153966e+05	3.334028e+00	
min	1.000001e+06	1.000010e+05	0.000000e+00	0.000000e+00	0.000000e+00	-9.000000e-01	0.000000e+00	0.000000e+00	
25%	1.461857e+06	1.893290e+05	6.321780e+03	1.872000e+04	2.416050e+04	0.000000e+00	5.084100e+04	1.000000e+01	
50%	1.923110e+06	2.787145e+05	1.125000e+04	7.104600e+04	8.054100e+04	1.638000e+03	1.123200e+05	1.200000e+01	
75%	2.384280e+06	3.675140e+05	2.065842e+04	1.803600e+05	2.164185e+05	7.740000e+03	2.340000e+05	1.500000e+01	
max	2.845382e+06	4.562550e+05	4.180581e+05	6.905160e+06	6.905160e+06	3.060045e+06	6.905160e+06	2.300000e+01	

8 rows × 21 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

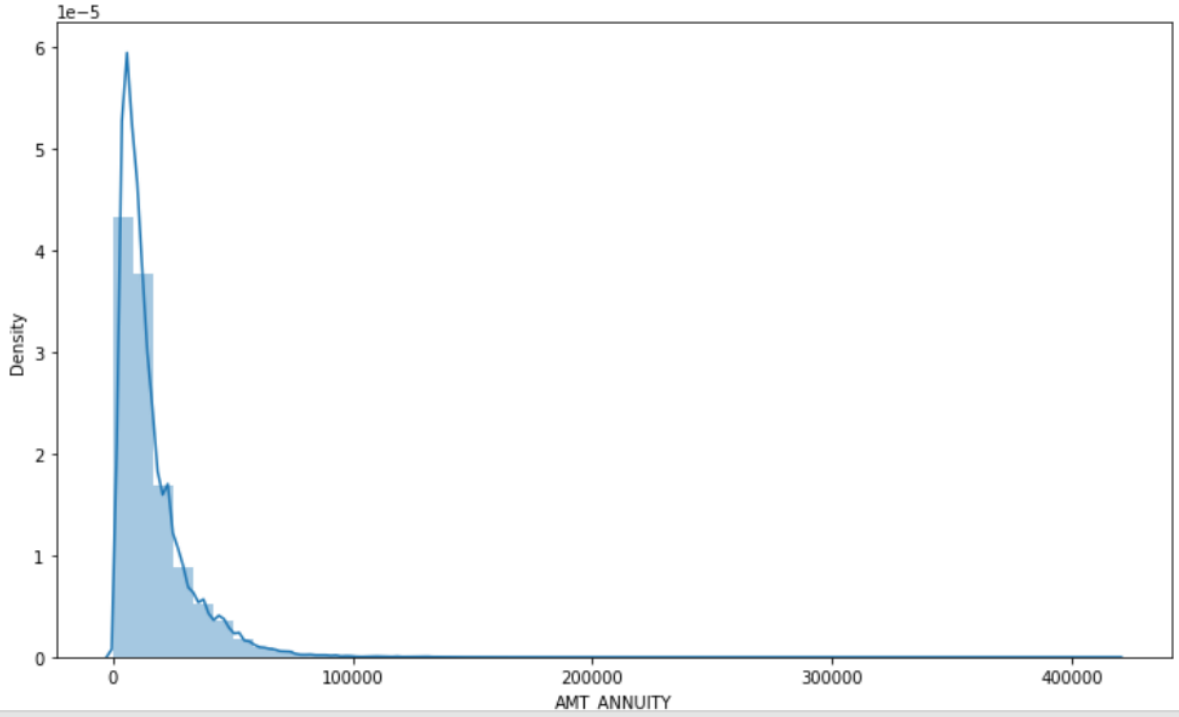
AMT_ANNUITY

CNT_PAYMENT

- Listing down columns which are not required

```
plt.figure(figsize = [12,7])
sns.distplot(df2['AMT_ANNUITY'])
plt.show()
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated
  warnings.warn(msg, FutureWarning)
```

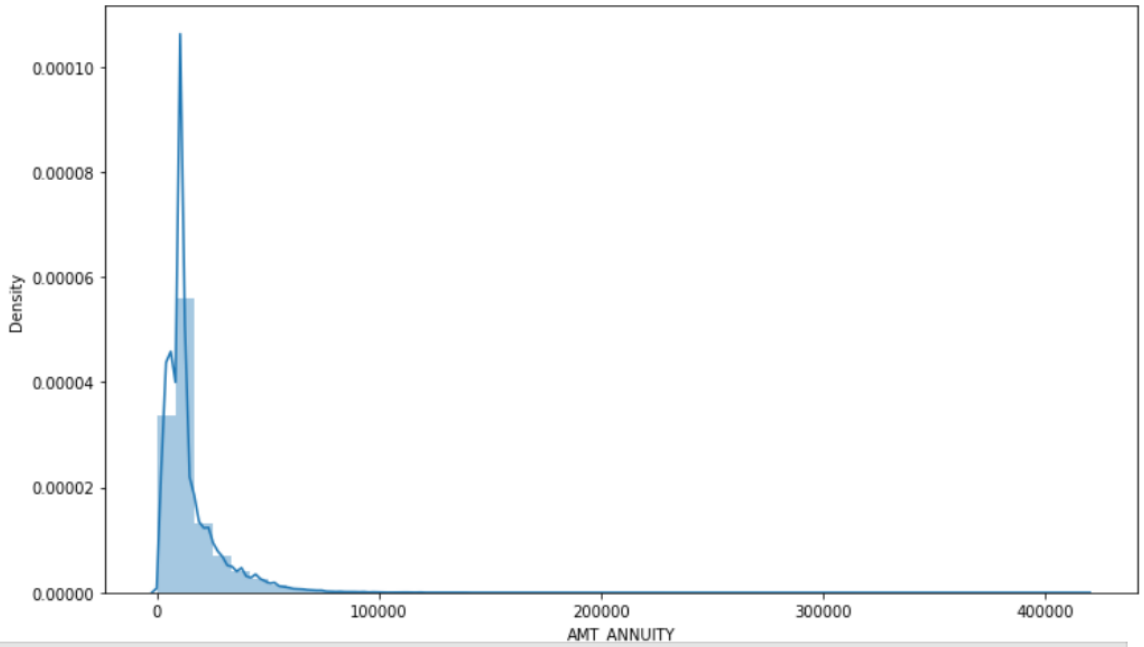


There is a single peak at the left side of the distribution and it indicates the presence of outliers and Hence imputing with mean would not be the right approach and hence imputing with median.then after replacing null values with median

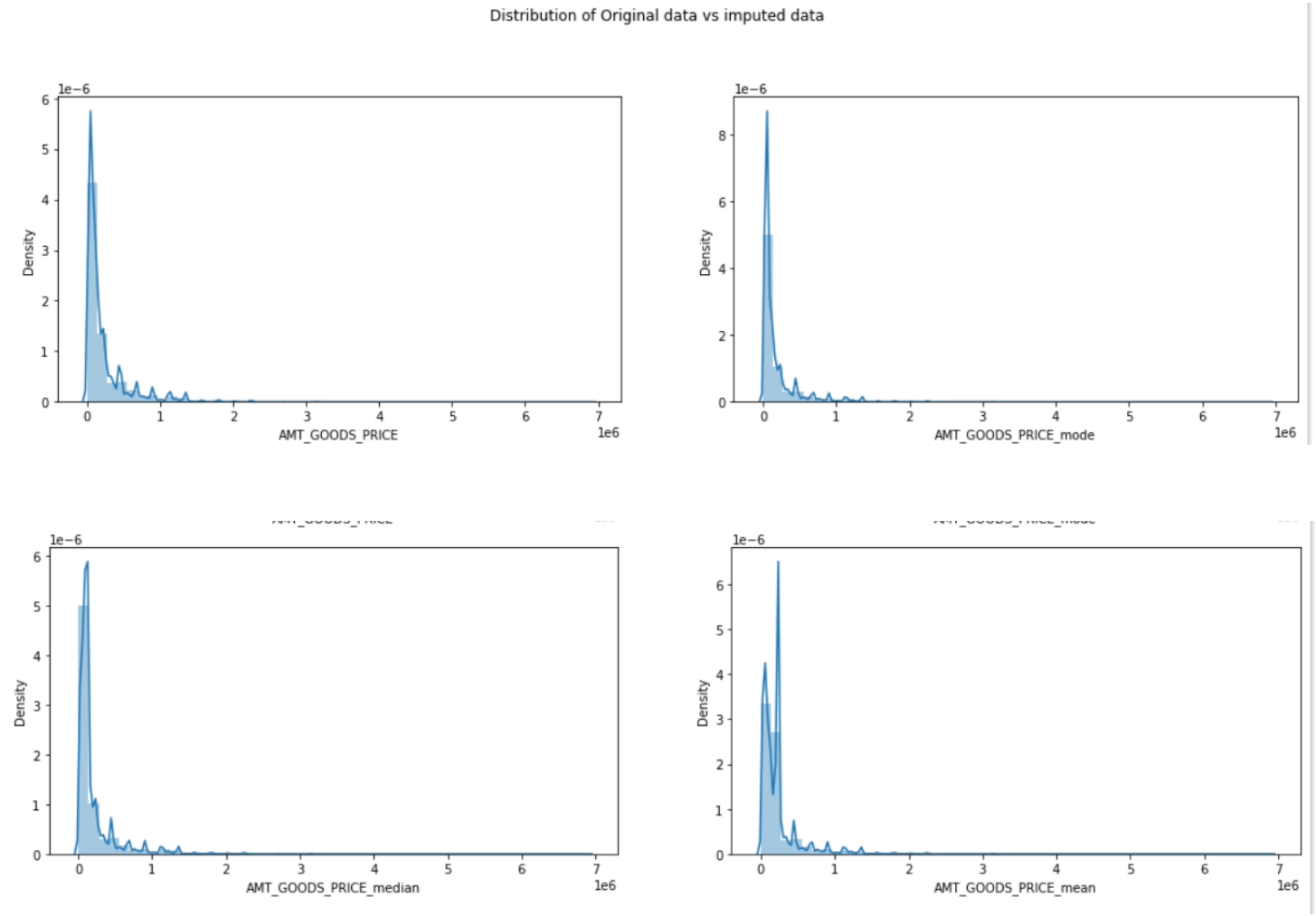
#replacing null values with median

```
plt.figure(figsize = [12,7])
sns.distplot(df2['AMT_ANNUITY'])
plt.show()
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function
  warnings.warn(msg, FutureWarning)
```



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

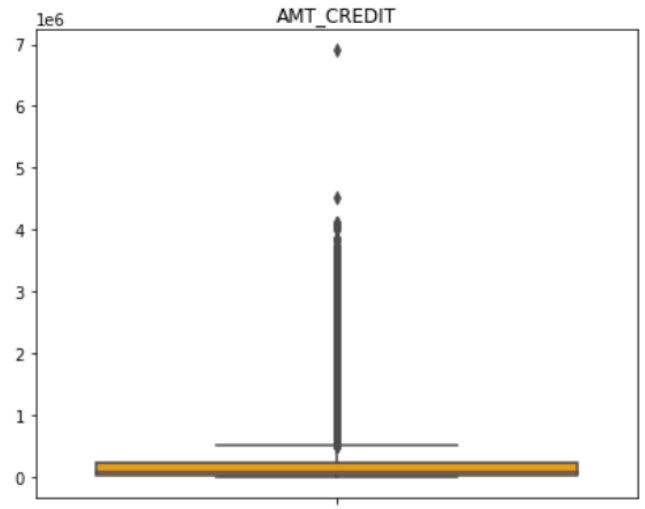
ImputingFinding Outliers null values with mode

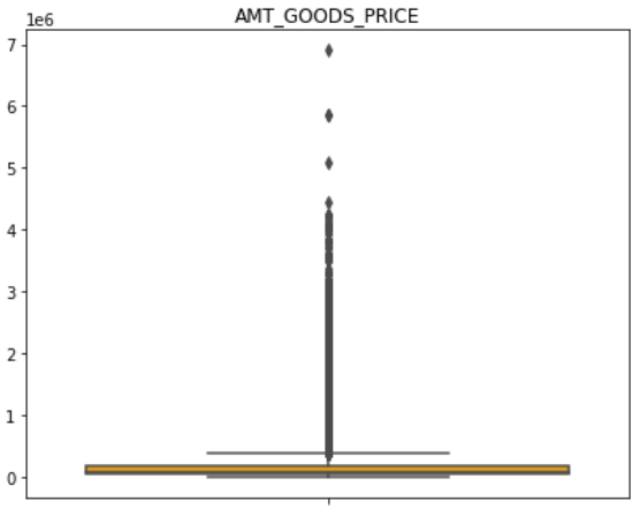
successfully handled all the null values in AMT_GOODS_PRICE column

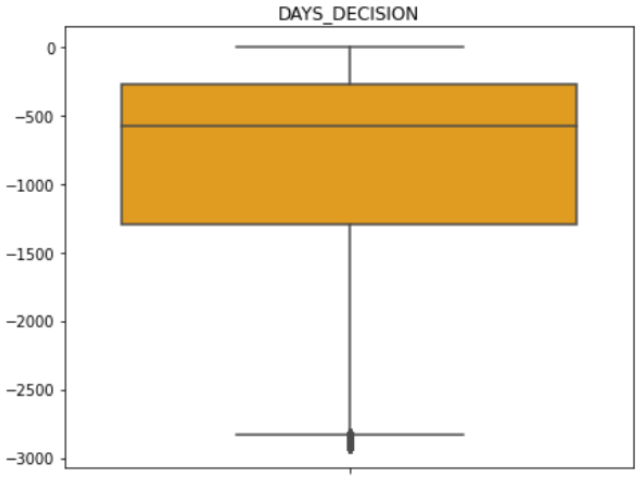
Finding Outliers

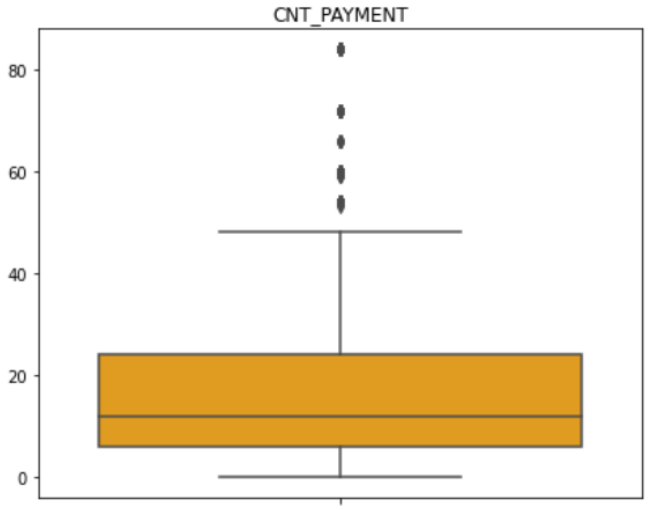
```
df2.describe()
```

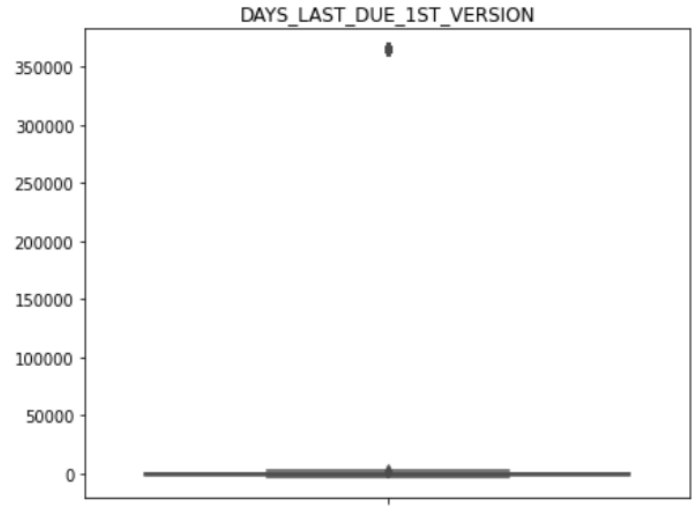
	SK_ID_PREV	SK_ID_CURR	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	DAYS_DECISION	SELLERPLACE_AREA
count	1.670214e+06	1.670214e+06	1.670214e+06	1.670214e+06	1.670213e+06	1.670214e+06	1.670214e+06	1.670214e+06
mean	1.923089e+06	2.783572e+05	1.490651e+04	1.752339e+05	1.961140e+05	1.856429e+05	-8.806797e+02	3.139511e+02
std	5.325980e+05	1.028148e+05	1.317751e+04	2.927798e+05	3.185746e+05	2.871413e+05	7.790997e+02	7.127443e+03
min	1.000001e+06	1.000010e+05	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	-2.922000e+03	-1.000000e+00
25%	1.461857e+06	1.893290e+05	7.547096e+03	1.872000e+04	2.416050e+04	4.500000e+04	-1.300000e+03	-1.000000e+00
50%	1.923110e+06	2.787145e+05	1.125000e+04	7.104600e+04	8.054100e+04	7.105050e+04	-5.810000e+02	3.000000e+00
75%	2.384280e+06	3.675140e+05	1.682403e+04	1.803600e+05	2.164185e+05	1.804050e+05	-2.800000e+02	8.200000e+01
max	2.845382e+06	4.562550e+05	4.180581e+05	6.905160e+06	6.905160e+06	6.905160e+06	-1.000000e+00	4.000000e+06





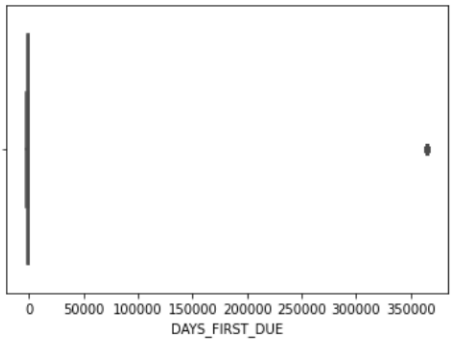







```
sns.boxplot(df2['DAYS_FIRST_DUE'], orient = "h", color = "orange")
```

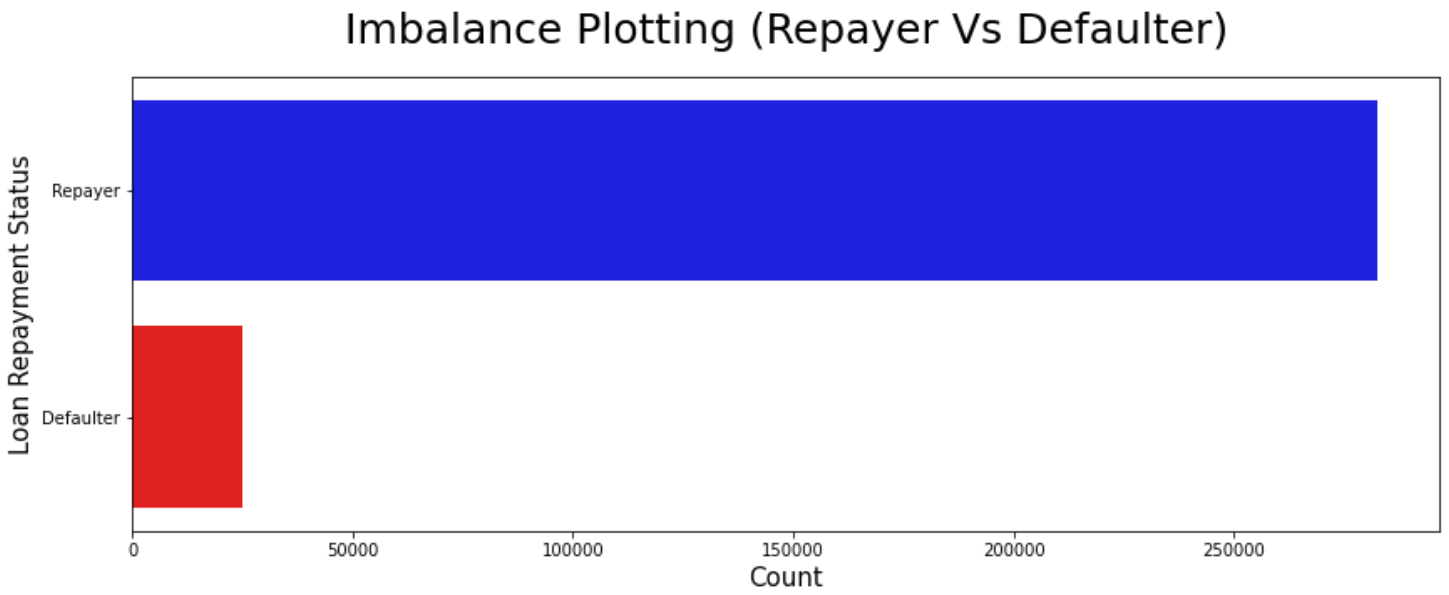
```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version
  warnings.warn(
<matplotlib.axes._subplots.AxesSubplot at 0x7fc1c1e950d0>
```



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



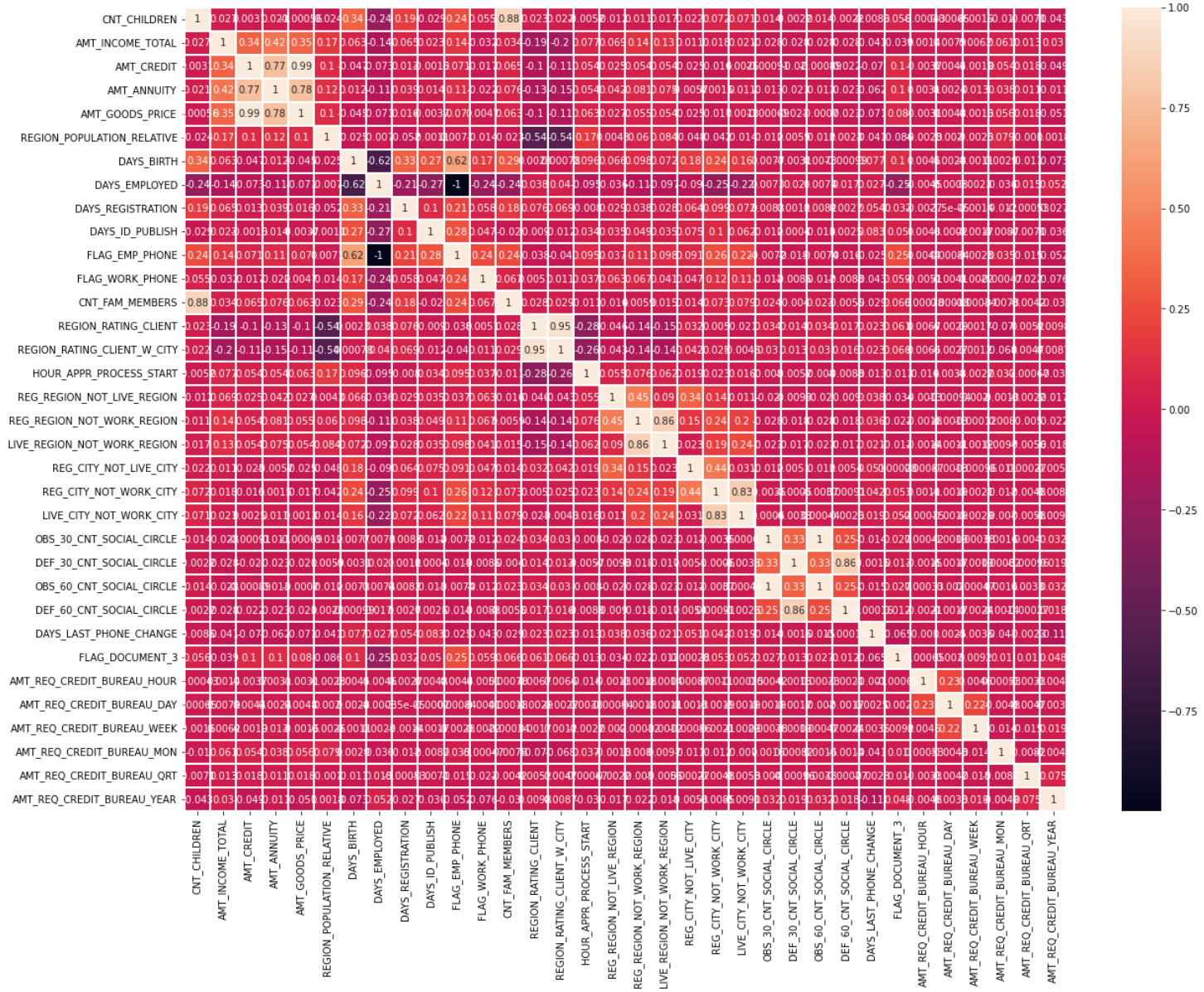
It can be seen that Businessman income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a Businessman could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs

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

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

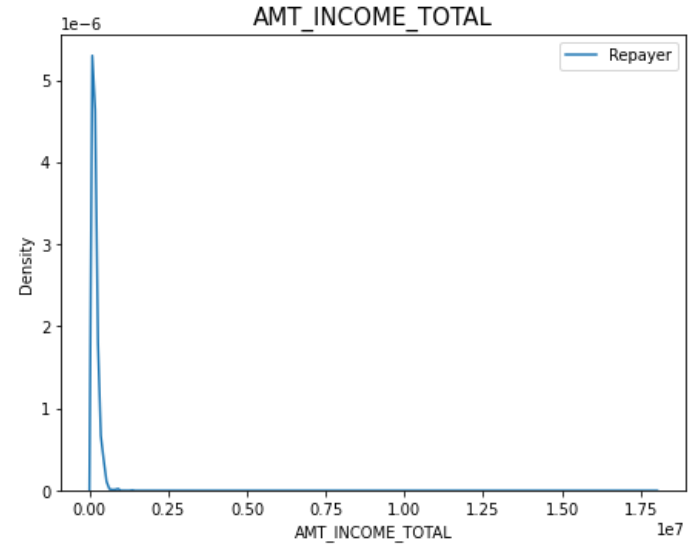
```
corr_df_repayer = corr_repayer.where(np.triu(np.ones(corr_repayer.shape),k=1).astype(np.bool)).u
```

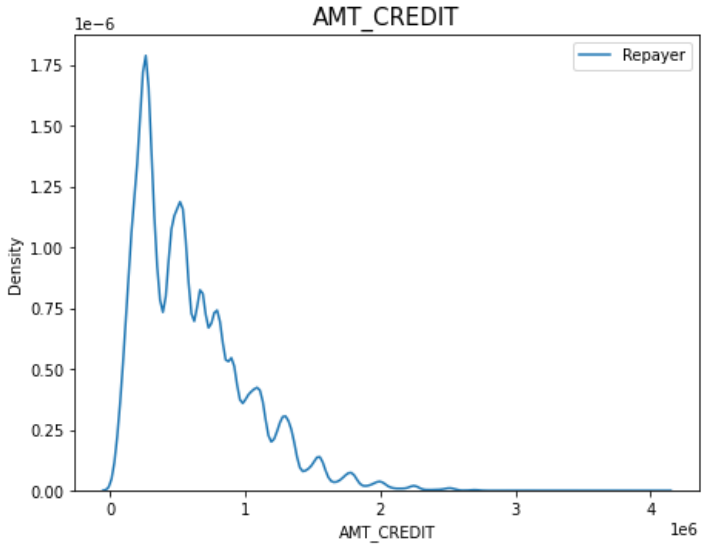
	VAR1	VAR2	Correlation
347	FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999758
838	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508
138	AMT_GOODS_PRICE	AMT_CREDIT	0.987022
489	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
408	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
629	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
873	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859332
734	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381
139	AMT_GOODS_PRICE	AMT_ANNUITY	0.776421
104	AMT_ANNUITY	AMT_CREDIT	0.771297

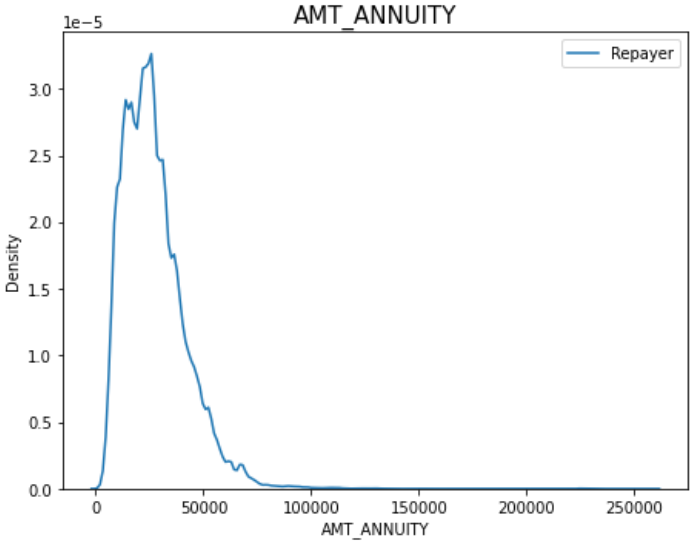


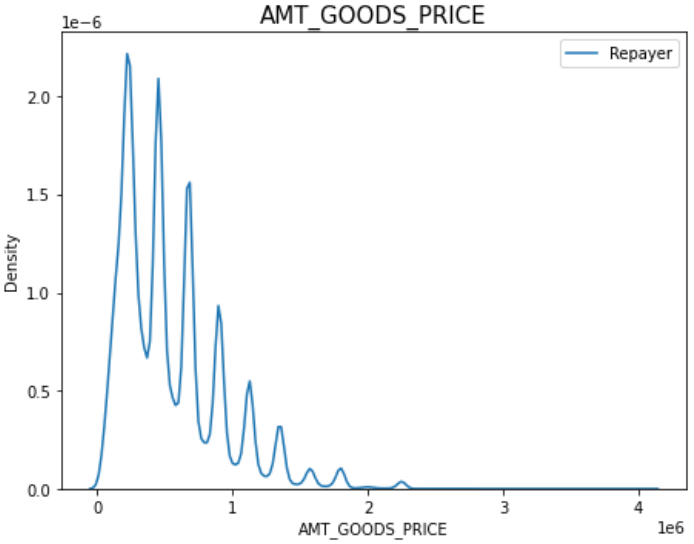
Plotting the numerical columns related to amount as distribution plot to see density

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.