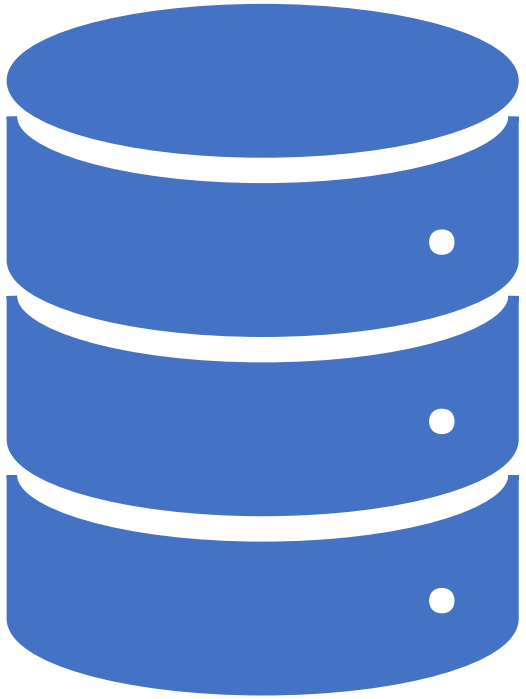
Case Study: Bank Loan & Risk Analysis

- Loan-providing companies find it difficult to provide loans and to identify the defaulters.

- Risks

1. Approving loans to those who do not pay leads to financial loss.
2. Not approving loans to those who are likely to pay back lead to business loss.



Loan application data

Scenarios:

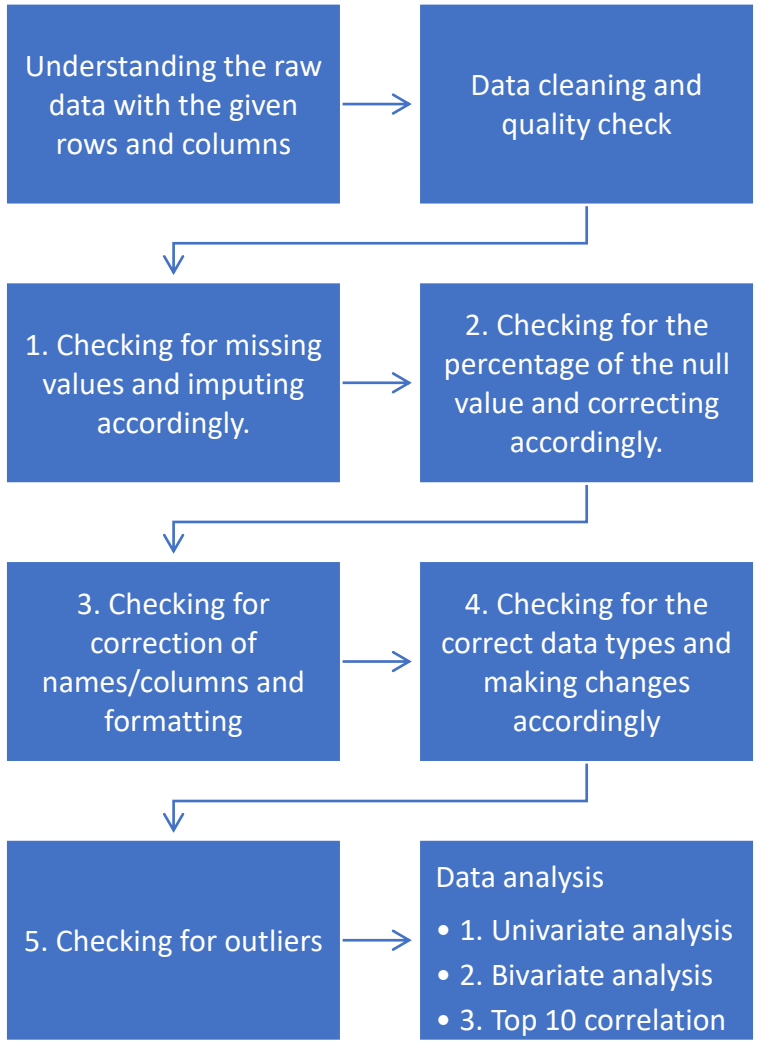
1. Client with payment difficulties: late payment made more than x days on at least one of the first installments y of the loan.

2. All other cases: payments made on time.

Loan application status

- **Approved:** The Company has approved the loan application
- **Cancelled:** The client canceled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.
- **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- **Unused offer:** Loan has been canceled by the client but at different stages of the process.

Process Flow



Understanding Raw Data

First import the required libraries

Load the data set using padas read_csv

Check for the data frame info() which will provide information on columns, not null and data types

By using df.shape we can get the number of rows and columns

By using describe() on the data frame we get the understanding of data with a lot of statistics like mean, median, standard deviation, percentiles of data, and mode.

From the above steps we try to first understand the basics of data

Data Cleaning

- First check for the missing values i.e Null values in the data frame
- Always convert to percentages of null values of the data frame for better analysis.
- The given application_data.csv had columns with null values greater than 40 percent. This clearly gives an idea that these columns we need to drop these for better analysis of data.
- Always recheck the data frame after dropping columns.
- Check for the remaining null values percent

1. Categorial column: Try to check for the most used value with mode and replace it accordingly in the place of null values.

2. Numerical column: Try to replace with mean if it sounds appropriate.

Data Cleaning Cont…

- Check for NAN or NaN or nan present in the columns and based on the analysis we can replace them with constant values to these rows.
- Check for negative numerical data and need to do analysis on further what kind of data it is based on.
- Check for correction of wrong data in columns or rows and correct it accordingly e.g. : Ram or ram or RaM
- Check for data types of columns and change them accordingly to the correct dtype.

Approach:

1. Get the columns with dtype as an “object” and try to convert it with astype(). This will not be a good idea as it can lead to wrong analysis.

2. Find the unique values of the columns and then based on the certain higher values we can come to decision on whether it’s a categorial or numerical type.

Out[10]:

OWN_CAR_AGE	65.990810
EXT_SOURCE_1	56.381073
APARTMENTS_AVG	50.749729
BASEMENTAREA_AVG	58.515956
YEARS_BEGINEXPLUATATION_AVG	48.781019
YEARS_BUILD_AVG	66.497784
COMMONAREA_AVG	69.872297
ELEVATORS_AVG	53.295980
ENTRANCES_AVG	50.348768
FLOORSMAX_AVG	49.760822
FLOORSMIN_AVG	67.848630
LANDAREA_AVG	59.376738
LIVINGAPARTMENTS_AVG	68.354953
LIVINGAREA_AVG	50.193326
NONLIVINGAPARTMENTS_AVG	69.432963
NONLIVINGAREA_AVG	55.179164
APARTMENTS_MODE	50.749729
BASEMENTAREA_MODE	58.515956
YEARS_BEGINEXPLUATATION_MODE	48.781019
YEARS_BUILD_MODE	66.497784
COMMONAREA_MODE	69.872297
ELEVATORS_MODE	53.295980
ENTRANCE	

Null Values >40 %

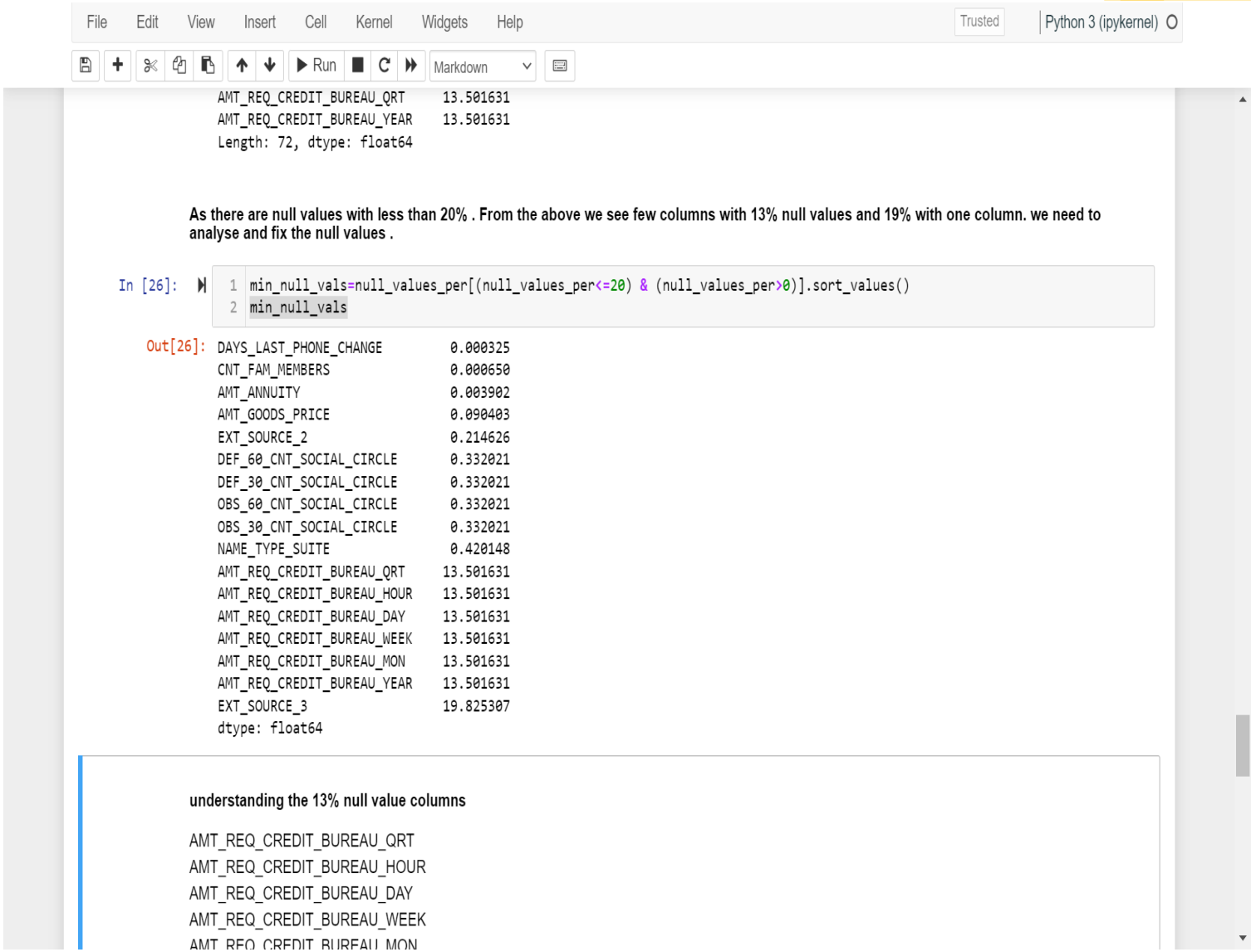
- We see the null values with greater than 40 % and we will drop those columns to avoid data irregularities.

Null values < 20 %

We see there are still columns with null values 13% and 19% .

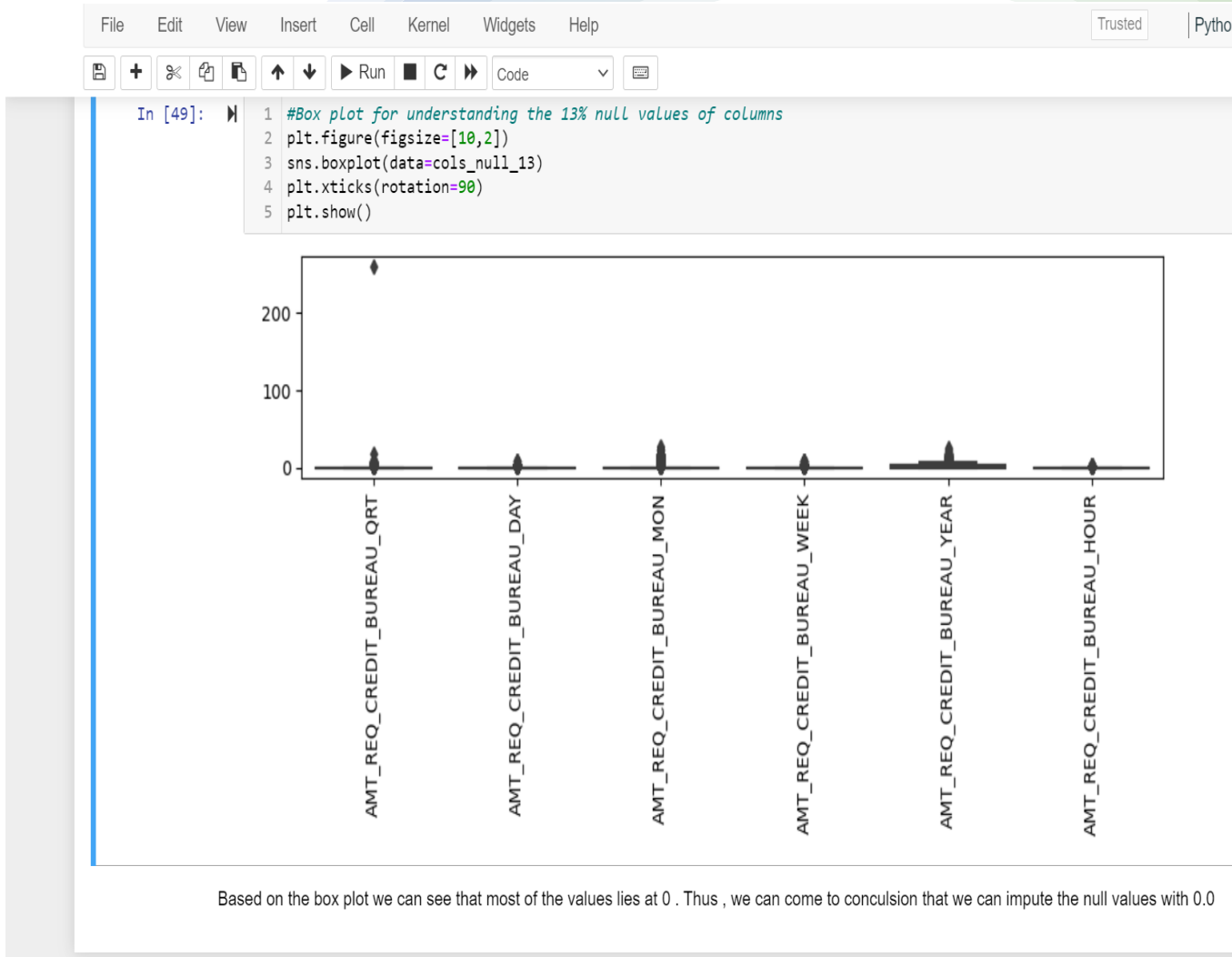
We check for the mode of the columns and impute the null values with their mode

Finally, we recheck if there are null values in those columns.



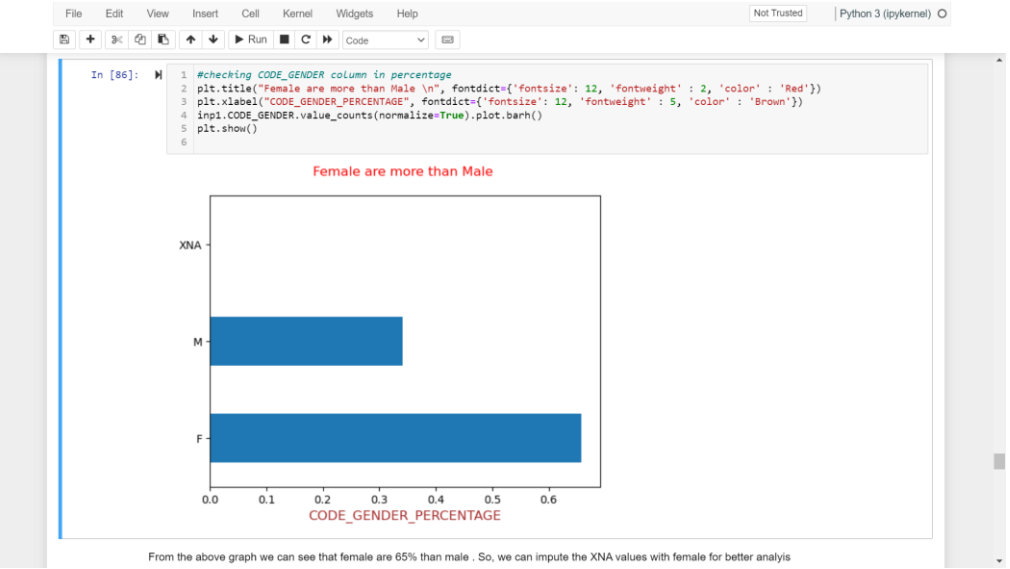
Imputing null values with 0.0 based on the box plot

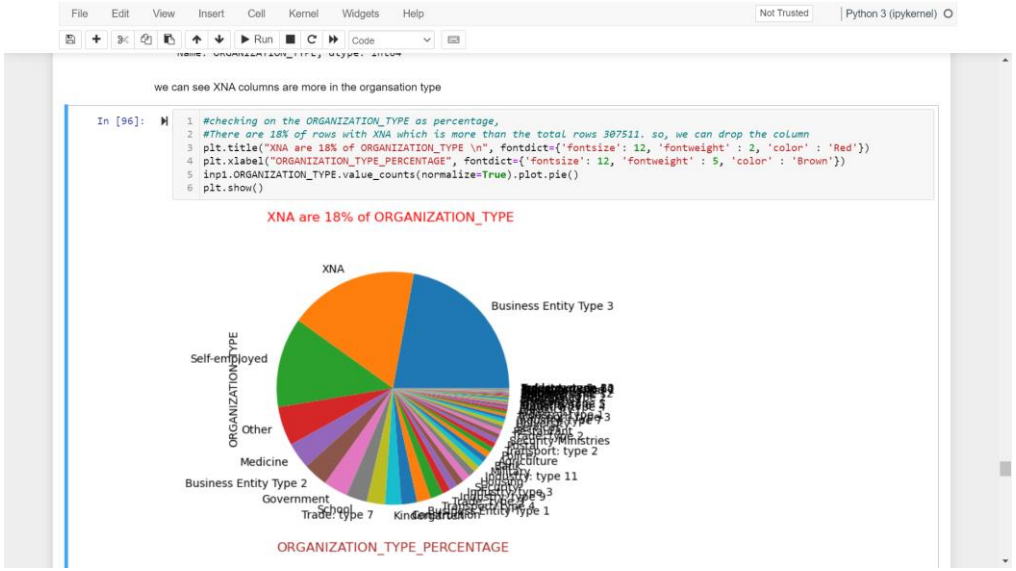
- We see these columns have 13% of null values
- AMT_REQ_CREDIT_BUREAU_YEAR
- AMT_REQ_CREDIT_BUREAU_QRT
- AMT_REQ_CREDIT_BUREAU_MON
- AMT_REQ_CREDIT_BUREAU_WEEK
- AMT_REQ_CREDIT_BUREAU_DAY
- AMT_REQ_CREDIT_BUREAU_HOUR



Imputing Categorial column

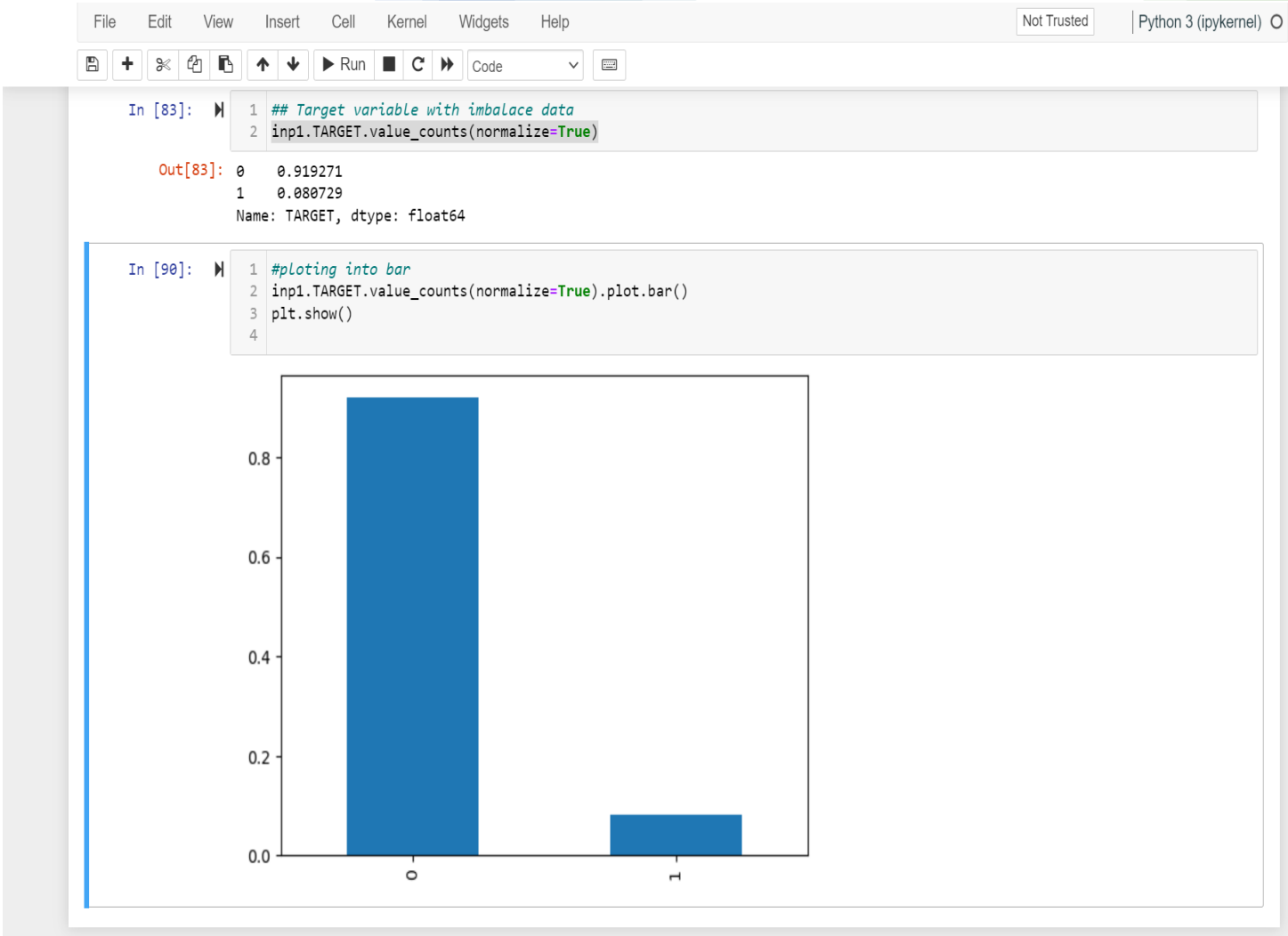
- The CODE_GENDER with XNA values imputed with F and ORGANIZATION_TYPE has XNA values with 18% so those rows will be dropped.





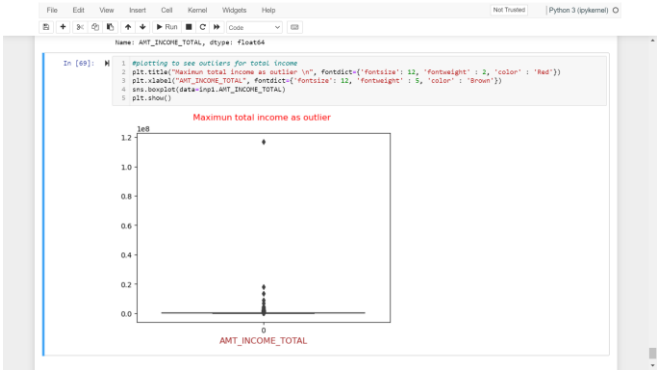
Imbalance Data

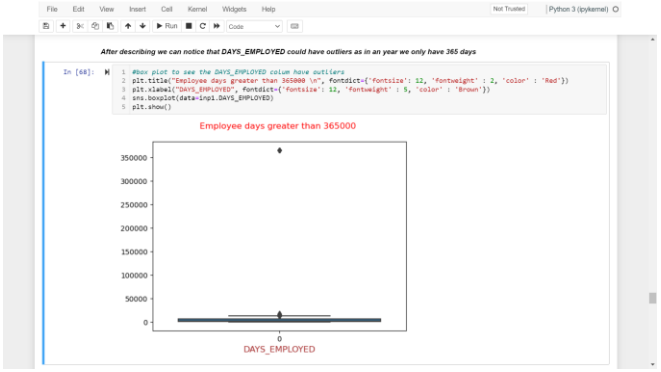
- The Target column has an imbalance of data based on the plot of the bar graph.

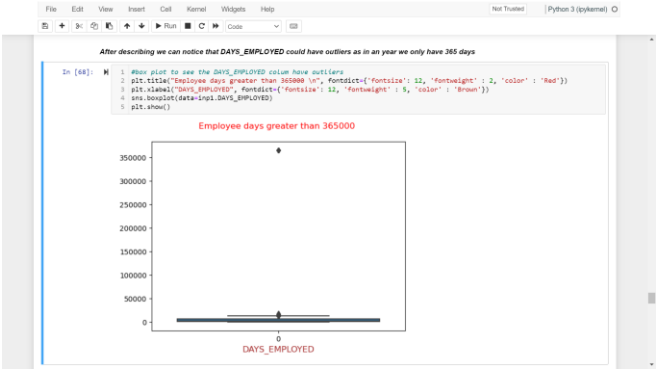


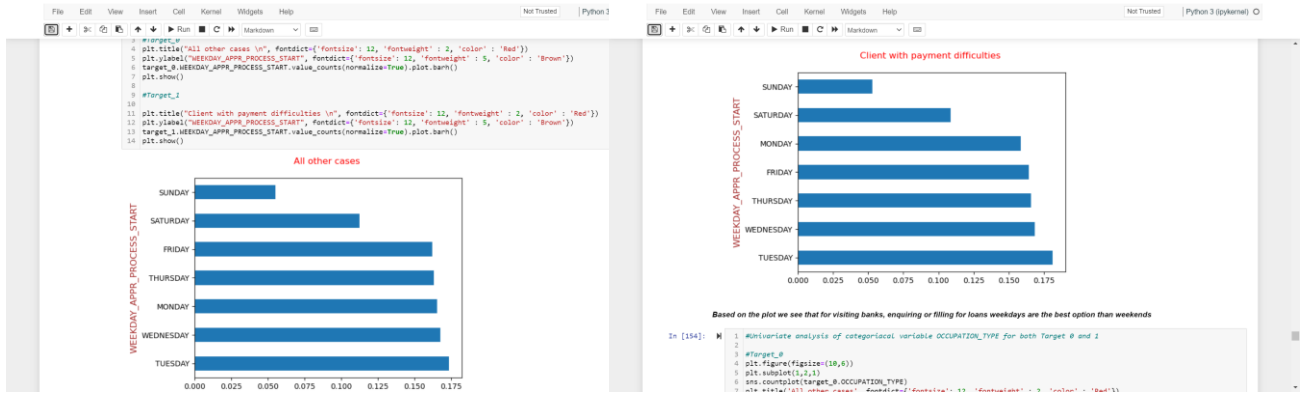
Outliers

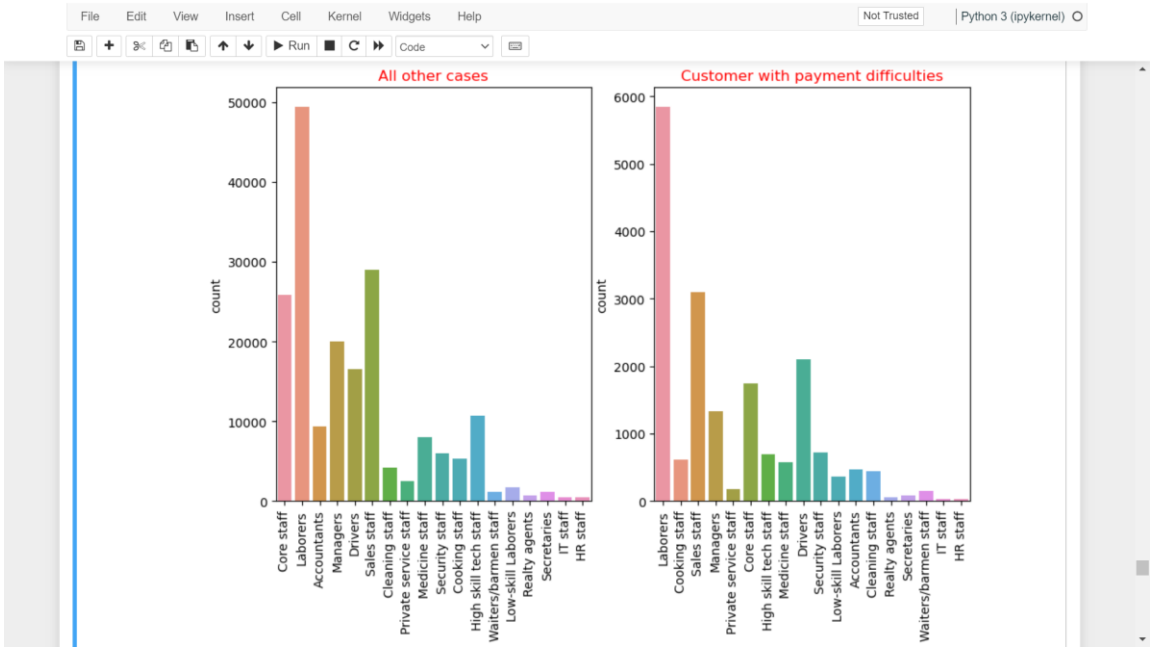
- There are a few columns AMT_INCOME_TOTAL,AMT_CREDIT CNT_CHILDREN,DAYS_EMPLOYED that are having outliers









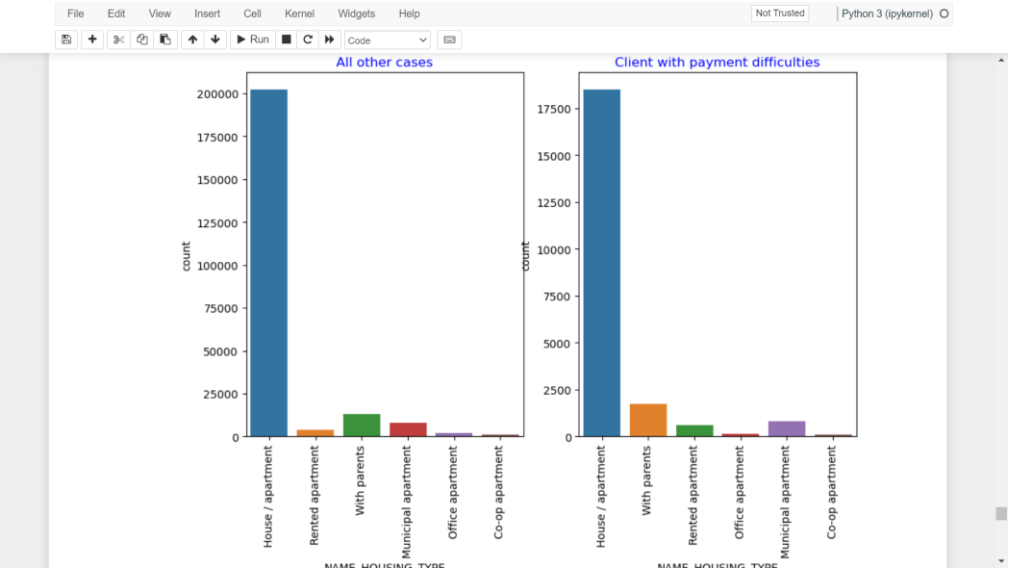


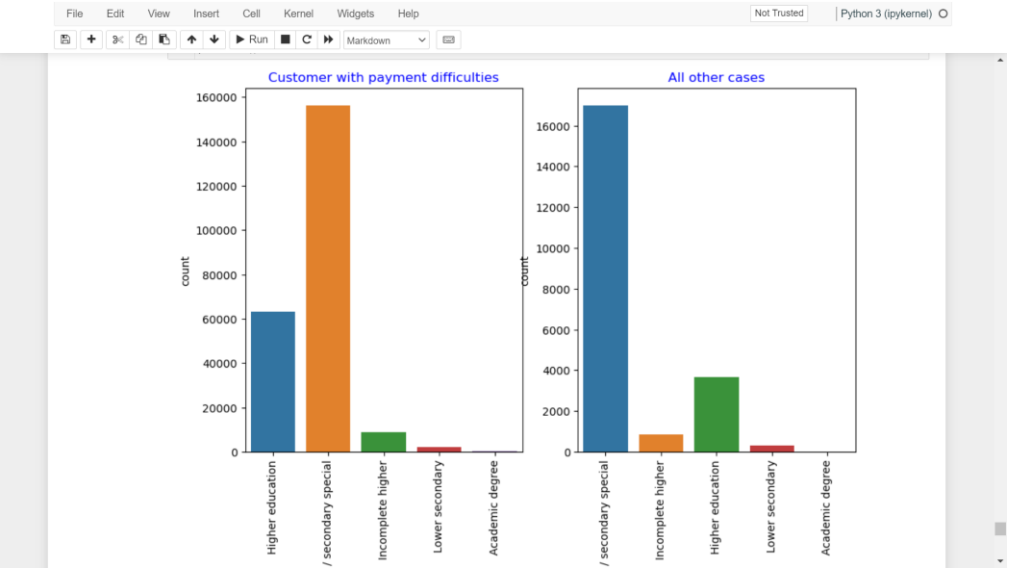
Univariate Analysis Categorical column

- Based on the first plot WEEKDAY_APPR_PROCESS_START we see that for visiting banks, enquiring, or filling for loan weekdays are the best option over weekends
- Based on the second plot OCCUPATION_TYPE we can see that most of the loans are taken by people with occupation as Laborers and the second highest of people for loans are with occupation Sales staff. However, both kind of occupation people can be defaulters too.

Univariate Analysis Categorical column

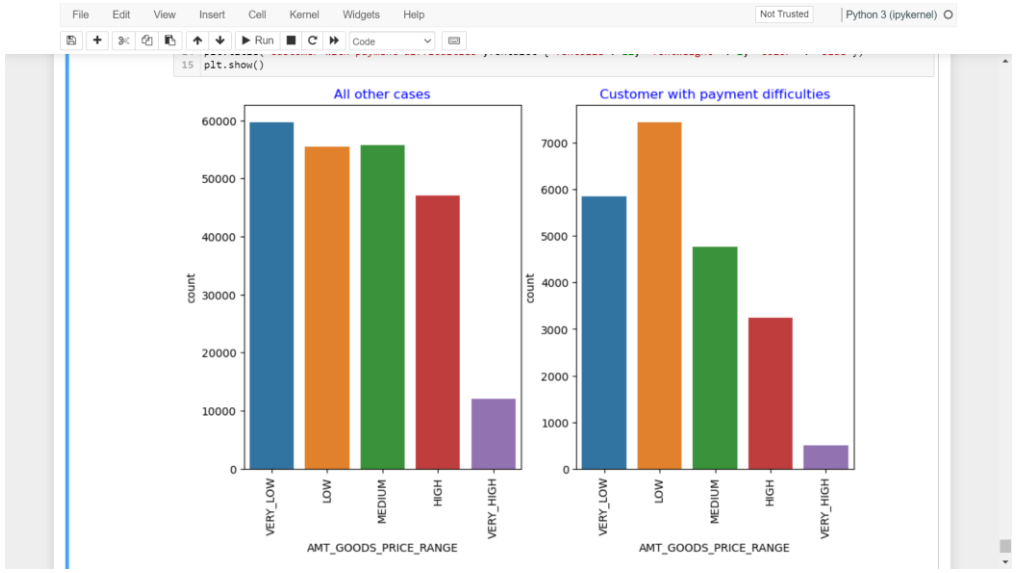
- Based on the first plot NAME_HOUSING_TYPE we can see that most of the clients living in their own house/apartment tend to be defaulters and loan payers.
- Based on the second plot NAME_EDUCATION_TYPE we see that people with education type as secondary/secondary special are high in both cases .

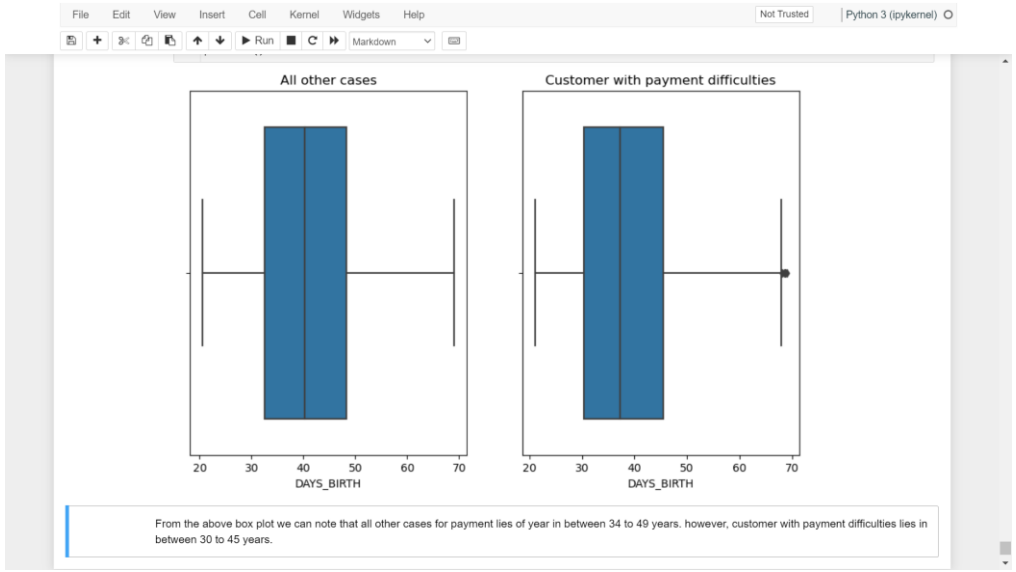




Univariate Numerical Variable

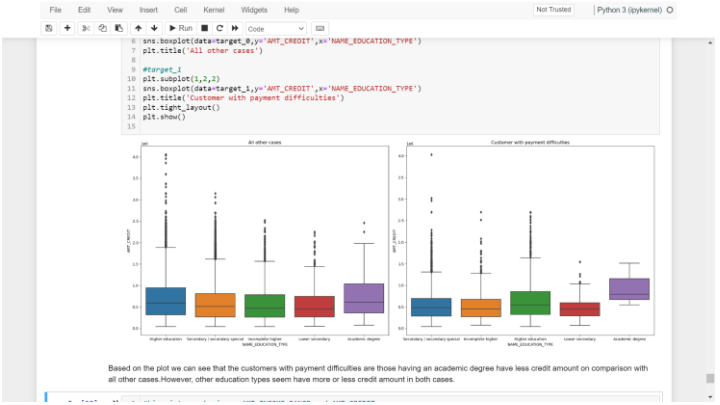
- Based on the first plot AMT_GOODS_PRICE_RANGE we can see the clients with payment difficulties for consumer loans along with goods price belongs under the Low category. On the other hand, the cases with other clients for consumer loans fall under the Very_Low category.
- Based on the second plot DAYS_BIRTH we can note that all other cases for payment lie in the year between 34 to 49 years. However, the customer with payment difficulties lies between 30 to 45 years.

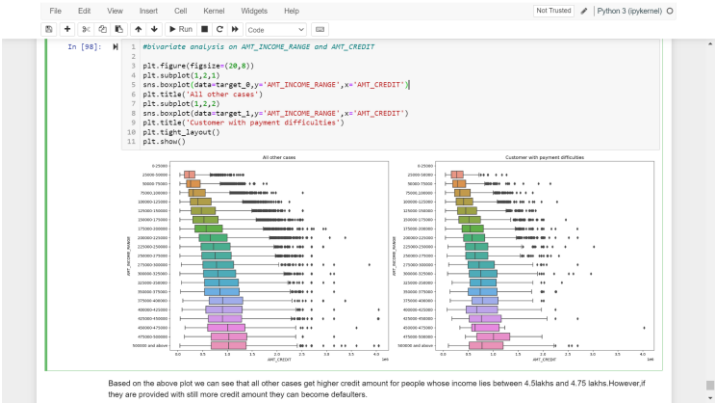




Bivariate Analysis

- Based on the First plot ***AMT_CREDIT Vs NAME_EDUCATION_TYPE*** we can see that the customers with payment difficulties are those having an academic degree and have less credit amount in comparison with all other cases. However, other education types seem to have more or fewer credit amounts in both cases.
- Based on the Second plot ***AMT_INCOME_RANGE Vs AMT_CREDIT*** we can see that all other cases get higher credit amounts for people whose income lies between 4.5lakhs and 4.75 lakhs. However, if they are provided with still more credit amount they can become defaulters.
- Based on the Third plot ***AMT_CREDIT Vs OCCUPATION_TYPE*** we can see that credit amount was given more to the clients with occupation as Accounts, IT staff, High skilled tech staff, and Realty agents.





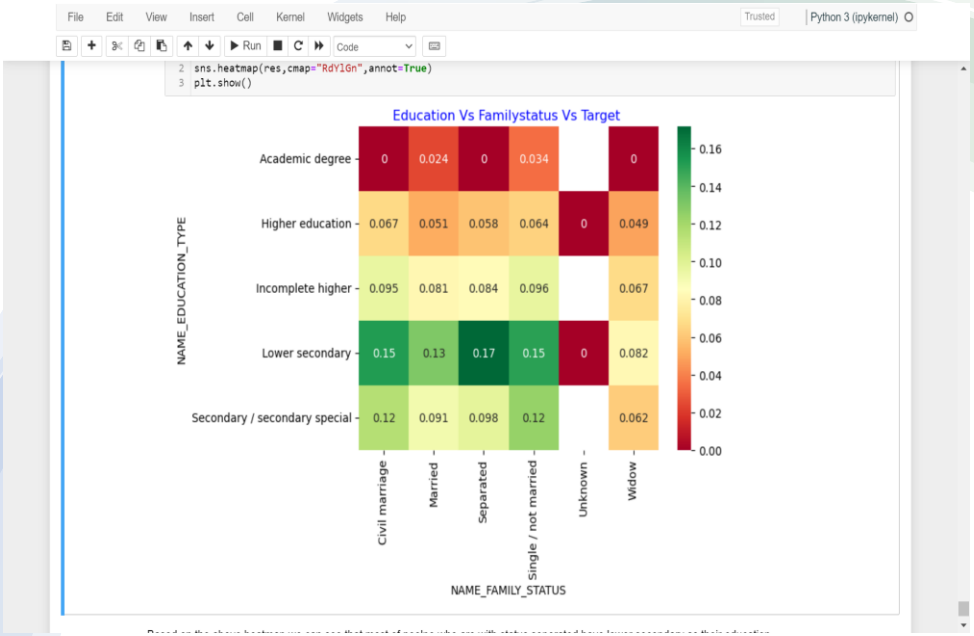
Multivariate Analysis

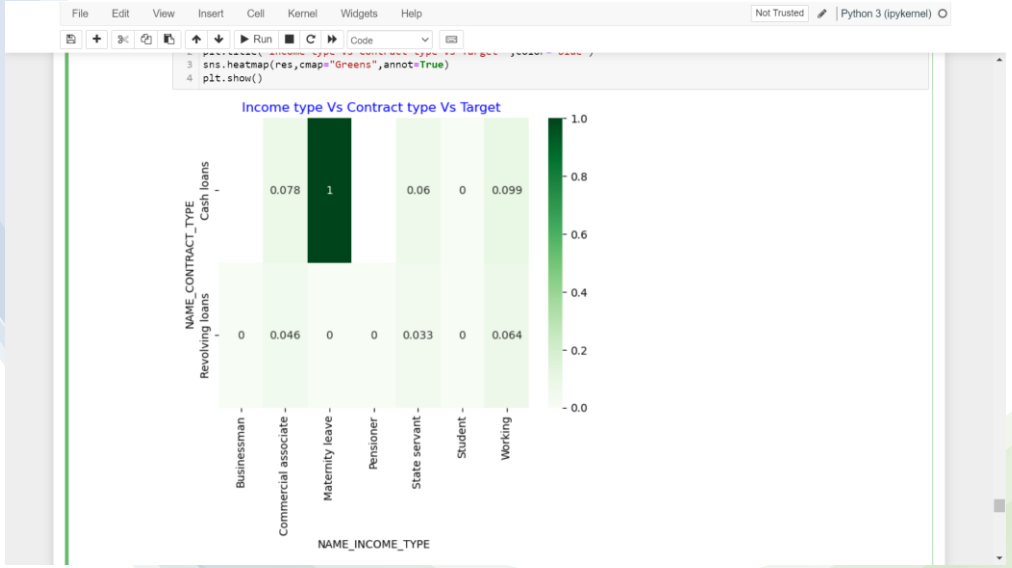
Education Vs Family status Vs Target

- Based on the above heatmap we can see that most of the people who are with status separated have lower secondary as their education.

Income type Vs Contract type Vs Target

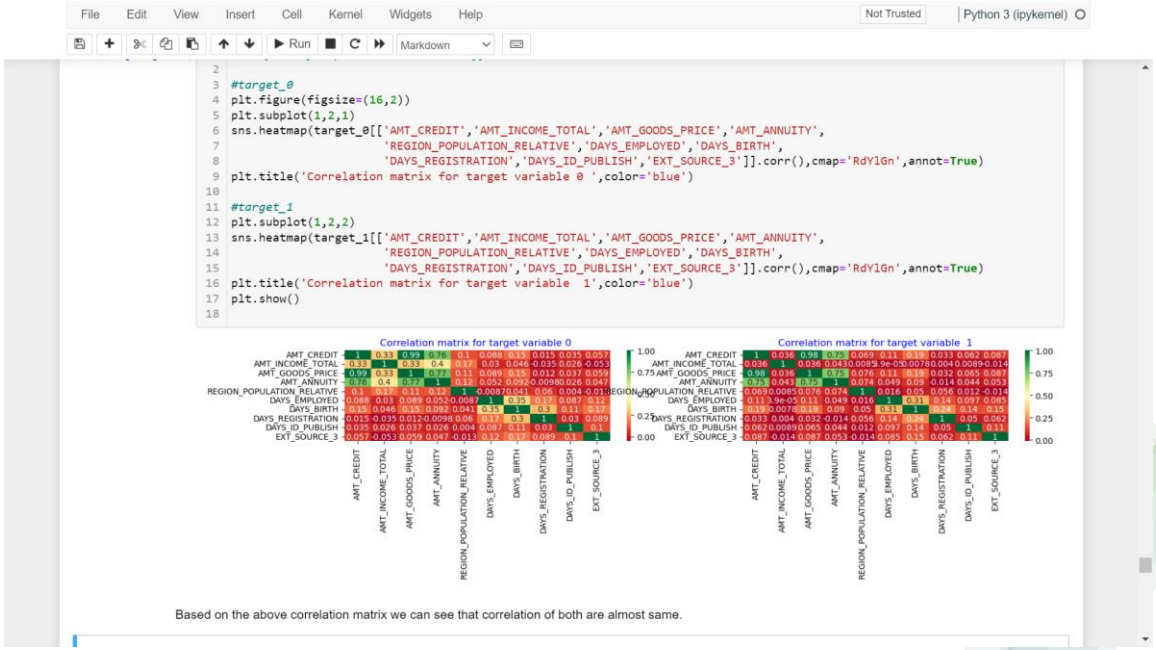
- From the plot we can see that Working, State servants, and Commercial associates are higher in default percentage. The maternity category can be a problem with high repayment.





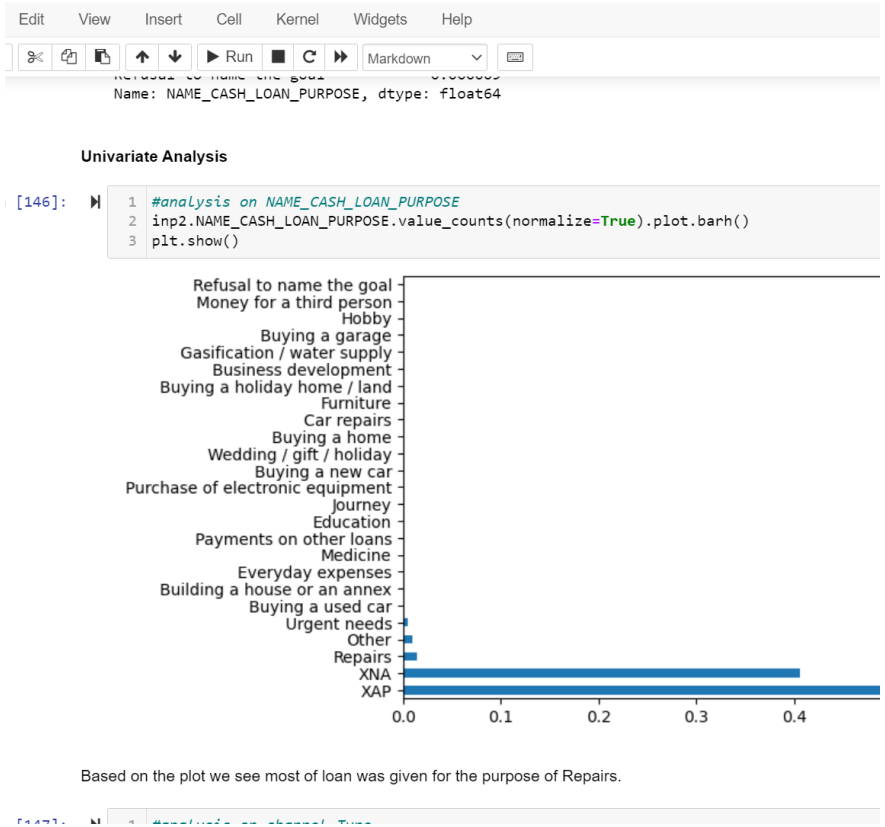
Correlation of top 10 columns

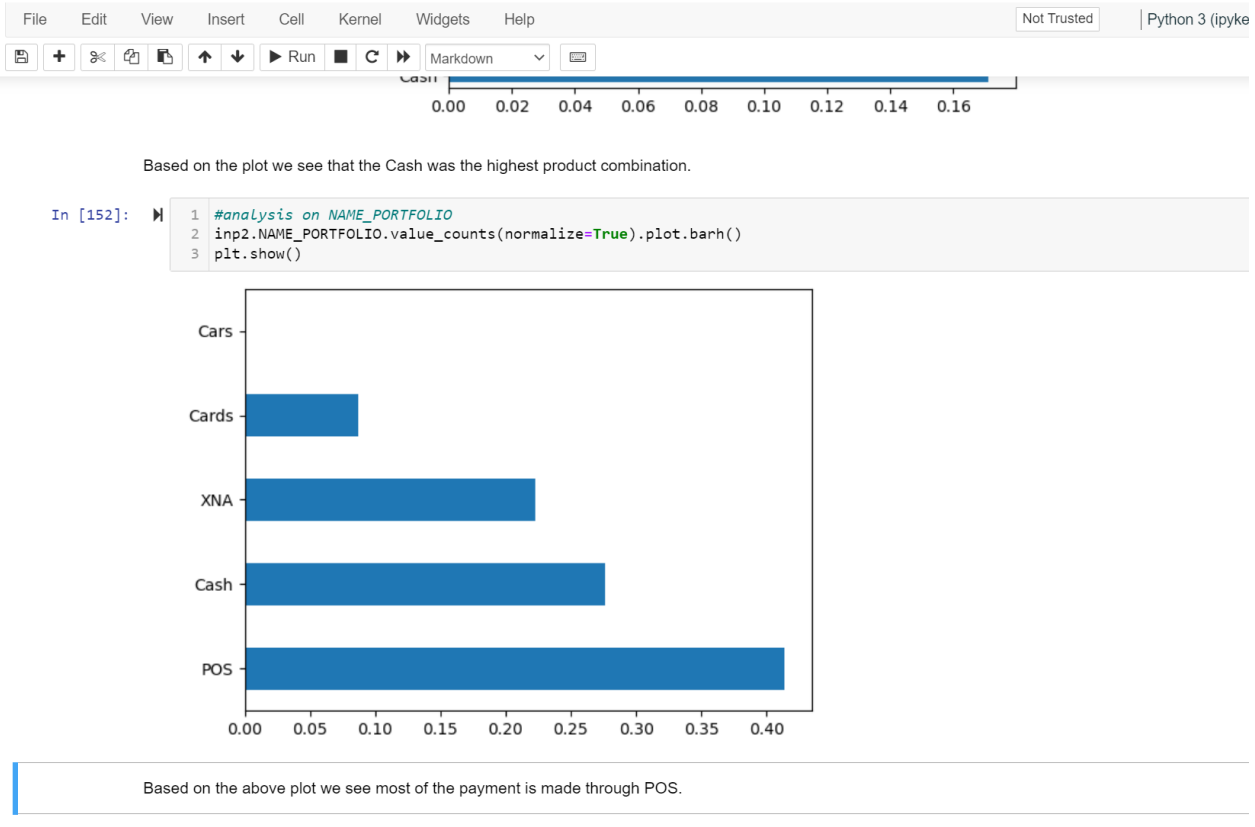
- Based on the correlation matrix we can see that correlation of both target_0 and tatget_1 is almost the same for a few columns. Based on the above correlation matrix we can see that correlation of both is almost the same. Thus, AMT_CREDIT and AMT_GOODS_PRICE are highly correlated.



Univariate Analysis on previous Data

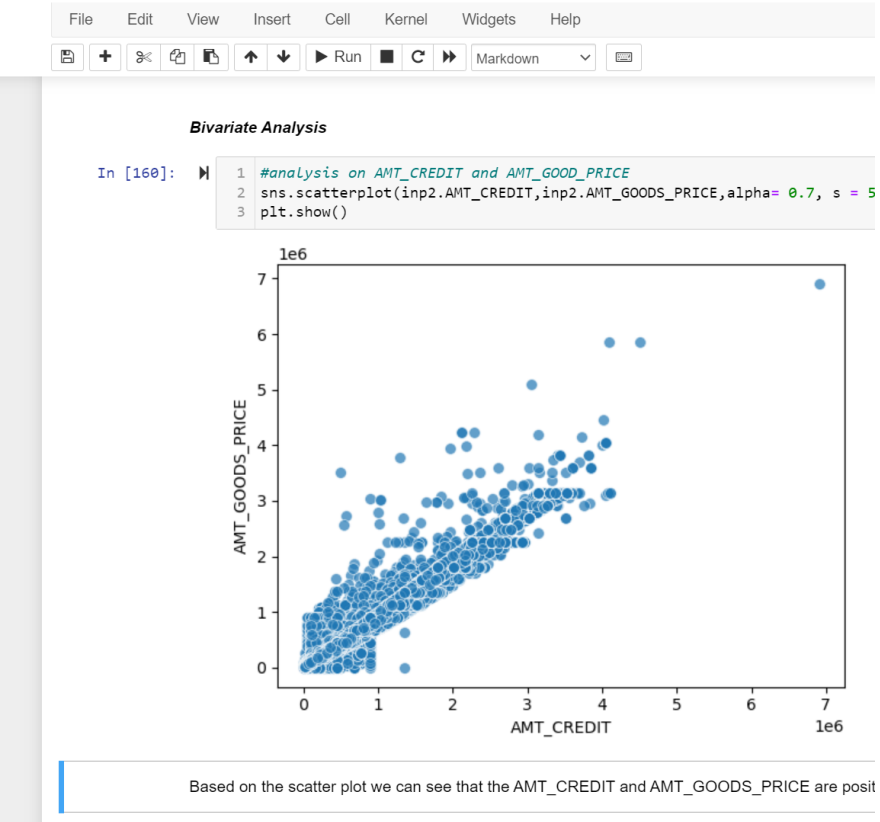
- Based on the First plot we can see most of the loan was given for the purpose on Repairs.
- Based on the Second plot we can see most of the payment made through POS.

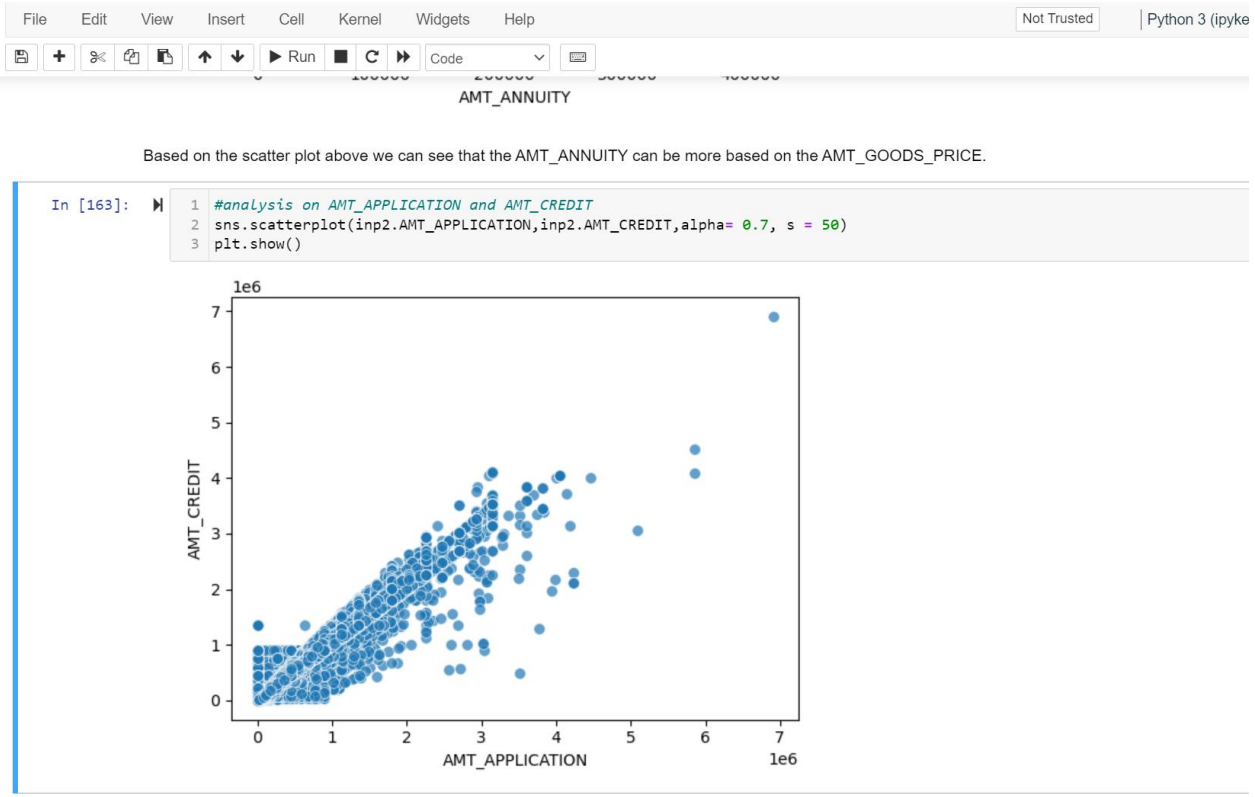




Bivariate analysis on previous application

- Based on the First scatter plot we can see that the AMT_CREDIT and AMT_GOODS_PRICE are positively correlated.
- Based on the second scatter plot we can see that the AMT_APPLICATION and AMT_CREDIT most of the loan amount was given as per client needs.





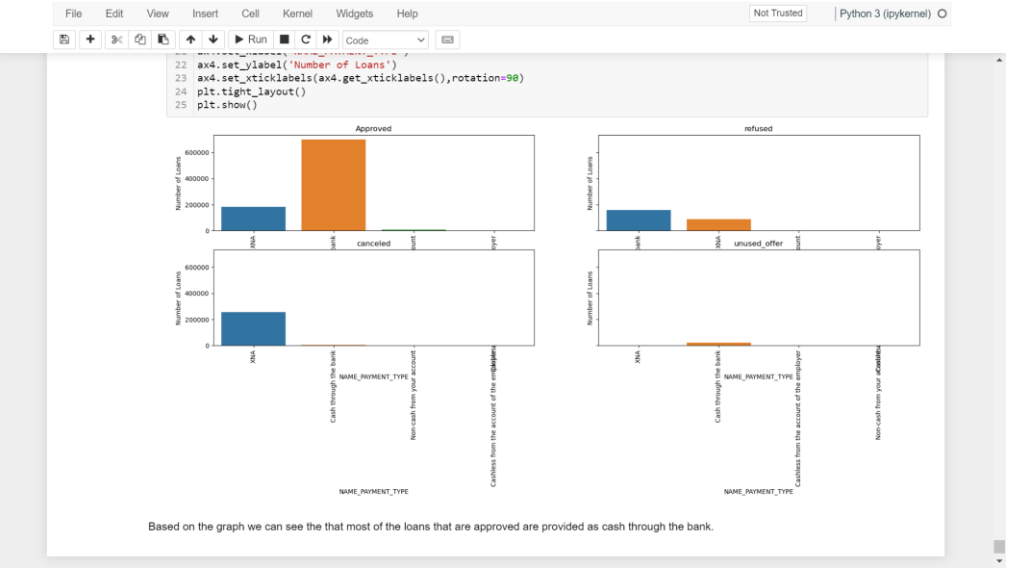
Analysis with Merged Data frame

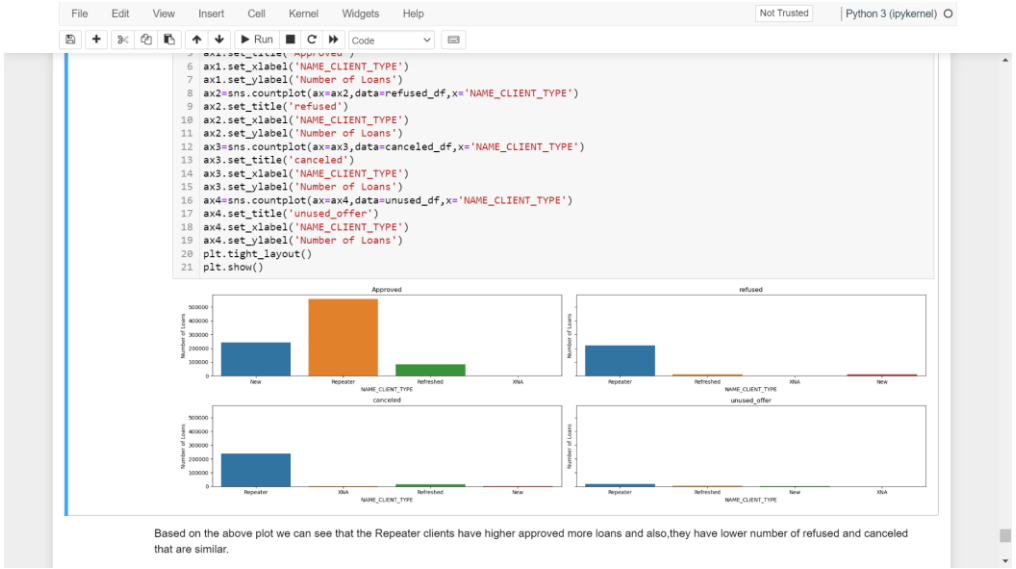
Analysis with NAME_CONTRACT_STATUS and NAME_PAYMENT_TYPE

- Based on the First plot we can see that most of the loans that are approved are provided as cash through the bank.

Analysis with NAME_CONTRACT_STATUS and NAME_CLIENT_TYPE

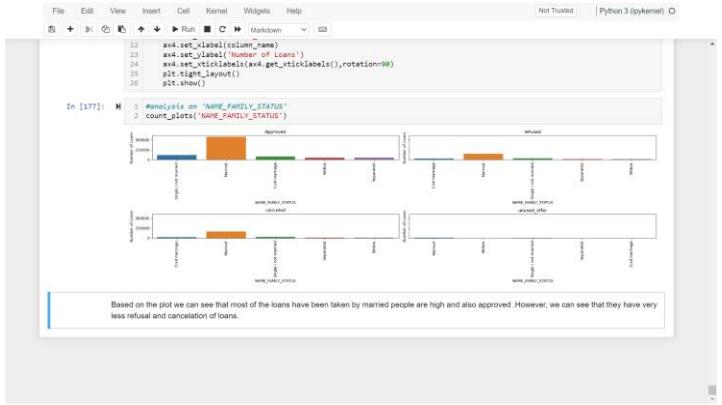
- Based on the Second plot we can see that the Repeater clients have higher approved more loans and also, they have a lower number of refused and canceled that are similar.

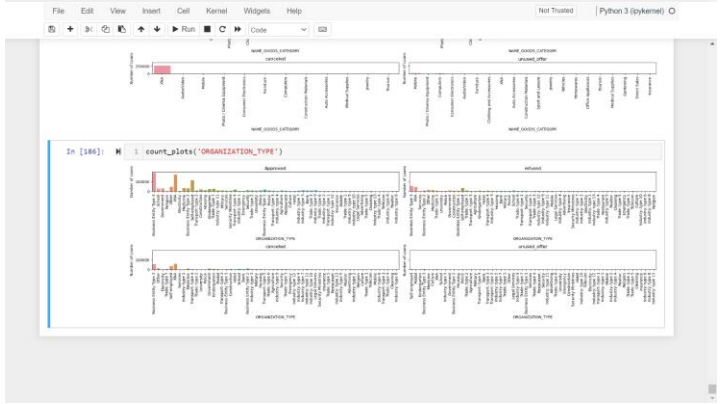




Merged Analysis Cont…

- Based on the First plot we can see that most of the loans that have been taken by married people are high and also approved. However, we can see that they have very less refusals and cancelations of loans.
- Based on the Second Plot we can see that most of the loans were taken for mobile, consumer electronics, and computers that are being approved on the same.
- Based on the Third Plot we can see that most of the loans were given to clients under organization Business entity type: 3, Self-employed, and Medicine that are being approved on the same.





Conclusion

- Based on the two data set we observe that

1. Females are with a higher percentage of applying for more loans.
2. Most of the loans are approved to the customers working under organization type ‘Business Entity type:3’
3. Banks should focus less on income type ‘Working’ as they are having the most number of unsuccessful payments.
4. Loan Purpose on ‘Repair’ having the highest number of unsuccessful repayments.
5. Housing Type 'With Parents' having the least number of unsuccessful repayments.
6. Most provide loans to clients having higher education type as Secondary/ Secondary special with marital status Married.