CREDIT EDA ASSIGNMENT

Business Problem:

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.

Objective:

- To Perform EDA to Analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.
- 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, if he/she is likely to default, then approving the loan may lead to a financial loss for the company.

ANALYSIS ON APPLICATION DATASET

Data Inspection and Quality check on Application dataset

- Columns with null values more than 47 percent may give wrong insights, hence dropping them
- OCCUPATION_TYPE Column has 31% missing values, since its a categorical column, imputing the missing values with a unknown or others value
- EXTERNAL_SOURCE_3 numerical column with no outliers and there is not much difference between mean and median, hence we can impute with mean or median
- We could see that 99% of values or mode in the columns AMT_REQ_CREDIT_BUREAU_HOUR , AMT_REQ_CREDIT_BUREAU_DAY , AMT_REQ_CREDIT_BUREAU_WEEK ,AMT_REQ_CREDIT_BUREAU_MON , AMT_REQ_CREDIT_BUREAU_QRT , AMT_REQ_CREDIT_BUREAU_YEAR is 0.0, hence imputing these columns with the mode

Binning Numerical Columns

- Binning AMT_CREDIT Column, Conclusion: The Credit amount of the loan for most applicants is either low(200000 to 400000) or Very High(above 800000)

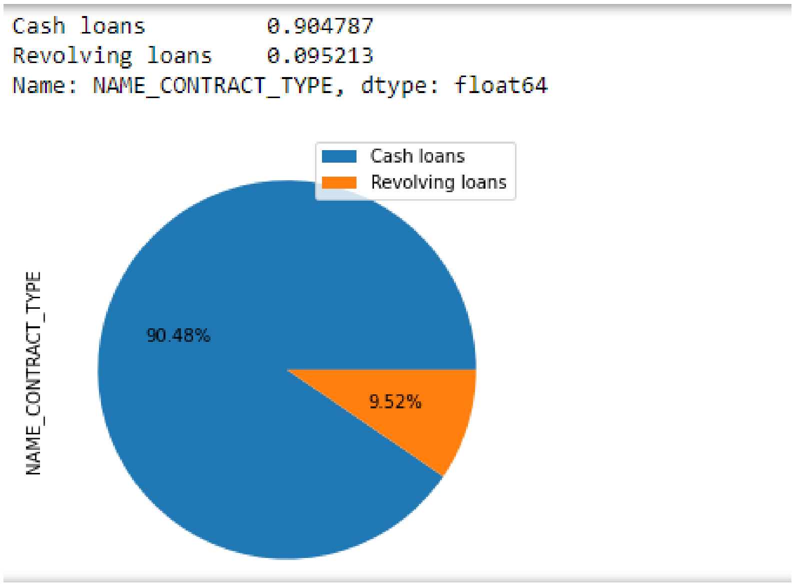
- Binning YEARS_BIRTH Column, Conclusion : Most of the Applicants are between 25-45 age group

Data Imbalance check on Target Column

- Conclusion: 1 out of Every 9 percent of applicants are defaulters

Univarient Analysis

- **Plot on Categorical columns**



```
Cash loans         0.904787
Revolving loans    0.095213
Name: NAME_CONTRACT_TYPE, dtype: float64
```

Conclusion: Insights on Categorical columns

1. NAME_CONTRACT_TYPE - More application have Cash loans than Revolving loans

2. CODE_GENDER - Number of Female applicants are twice than that of male applicants

3. FLAG_OWN_CAR - Most(70%) of the applicants do not own a car

4. FLAG_OWN_REALTY - Most(70%) of the applicants do not own a house

5. NAME_TYPE_SUITE - Most(81%) of the applicants are Unaccompanied

6. NAME_INCOME_TYPE - Most(51%) of the applicants are earning their income from Work

7. NAME_EDUCATION_TYPE - 71% of the applicants have completed Secondary / secondary special education

8. NAME_FAMILY_STATUS - 63% of the applicants are married

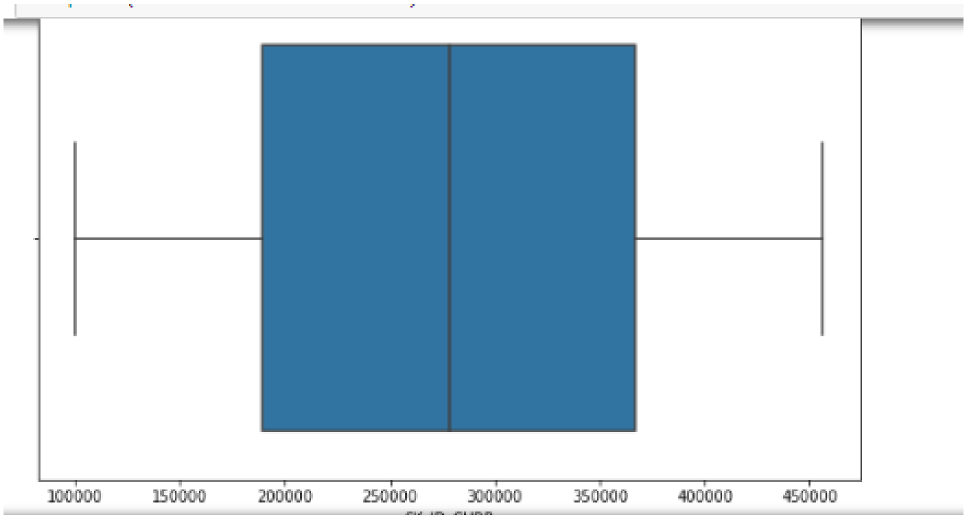
9. NAME_HOUSING_TYPE - 88% of the housing type of applicants are House/apartment

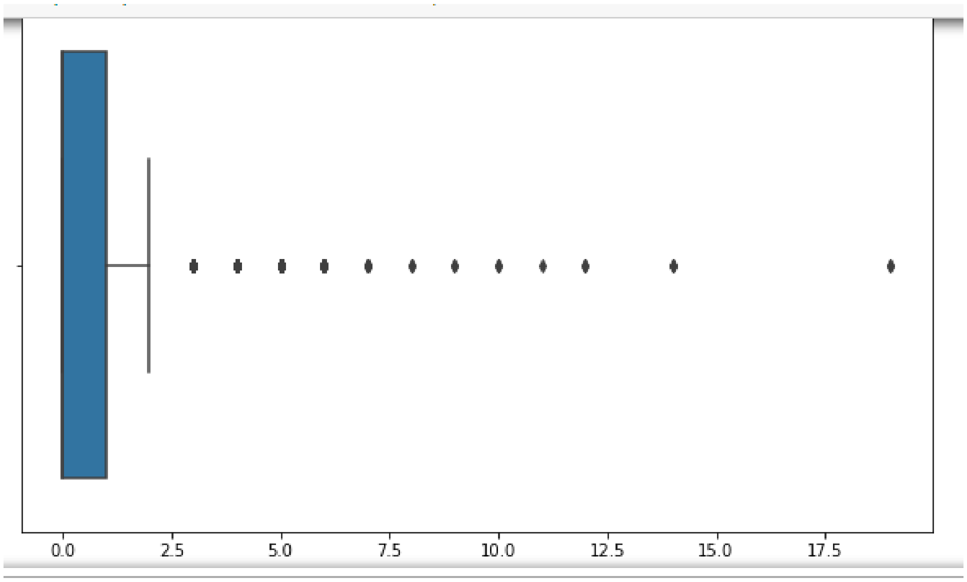
10. OCCUPATION_TYPE - Most(31%) of the applicants have other Occupation type

11. WEEKDAY_APPR_PROCESS_START- Most of the applicant have applied the loan on Tuseday

12. ORGANIZATION_TYPE - Most of the Organization type of employees are Business Entity Type 3

Plot on Numerical columns





Conclusion: Few Columns are with outliers are below

1. AMT_INCOME_TOTAL Column has a few outliers and there is a huge difference between the 99th percentile and the max value, also we could see huge variation in mean and median due to outliers

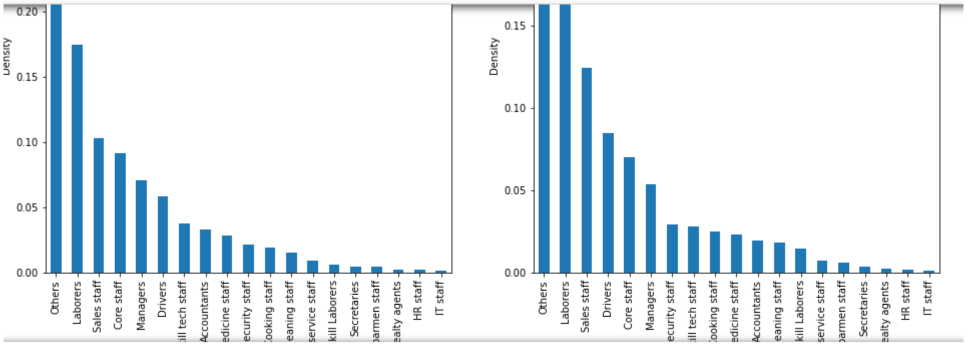
2. AMT_CREDIT Column has a few outliers and there is a huge difference between the 99th percentile and the max value, also we could see huge variation in mean and median due to outliers

3. AMT_ANNUITY Column has a few outliers and there is a huge difference between the 99th percentile and the max value, also we could see significant variation in mean and median due to outliers

4. AMT_GOODS_PRICE Column has a few outliers and there is a huge difference between the 99th percentile and the max value, also we could see significant variation in mean and median due to outliers

5. REGION_POPULATION_RELATIVE Column has a one outliers and there not much difference between mean and median

Univarient Analysis on Columns with Target 0 and 1

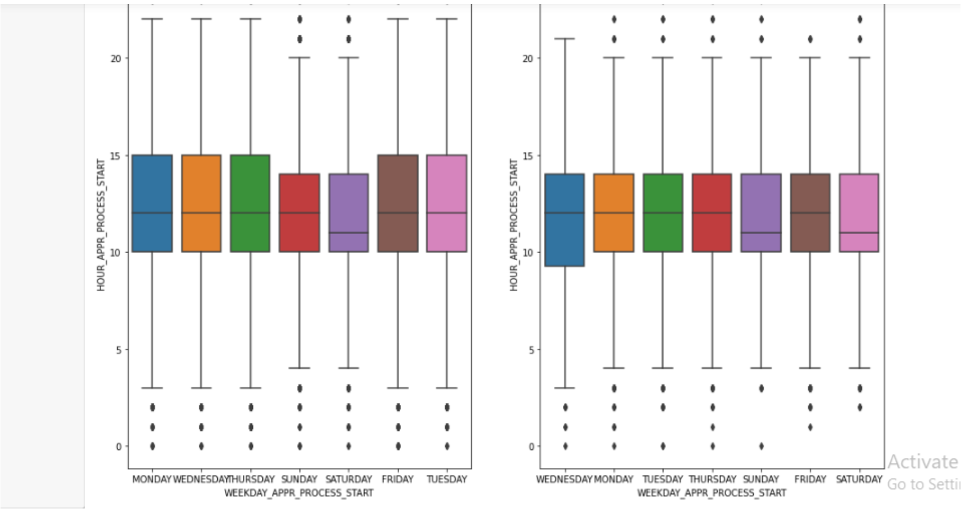


Conclusion >> Below are the column insights

1. NAME_CONTRACT TYPE- The Applicants are receiving more of Cash loans than Revolving loans both for Target 0 and 1
2. CODE_GENDER - Number of Female applicants are twice than that of male applicants both for Target 0 and 1
3. FLAG_OWN_CAR - Most(70%) of the applicants do not own a car both for Target 0 and 1
4. FLAG_OWN_REALTY - Most(70%) of the applicants do not own a house both for Target 0 and 1
5. NAME_TYPE_SUITE - Most(81%) of the applicants are Unaccompanied both for Target 0 and 1
6. NAME_INCOME_TYPE - For both Target 0 and 1, Most(51%) of the applicants are earning their income from Work
7. NAME_EDUCATION_TYPE - For both Target 0 and 1, almost 71% of the applicants have completed Secondary / secondary special education
8. NAME_FAMILY_STATUS - 63% of the applicants are married for both Target 0 and 1
9. NAME_HOUSING_TYPE - 88% of the housing type of applicants are House/apartment for both Target 0 and 1
10. OCCUPATION_TYPE - Most(31%) of the applicants have other Occupation type, are non defaulters and Laborere,Sales staff,Drivers and core staff are not able to repay the loan on time
11. WEEKDAY_APPR_PROCESS_START- Most of the applicant have applied the loan on Tuseday and the least on Sunday
12. ORGANIZATION_TYPE - Most of the Applicants are working in Business Entity Type 3, Self Employed and other Organization type

Bivarient & Multivarient Analysis

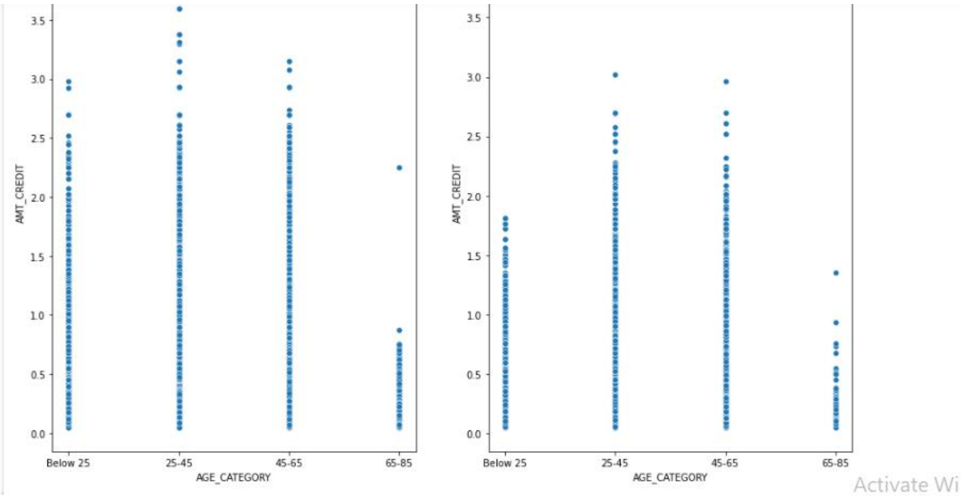
Bivarient Analysis between WEEKDAY_APPR_PROCESS_START VS HOUR_APPR_PROCESS_START¶



Conclusion:

1. The Bank operates between 10am to 3pm except for Saturday and Sunday, its between 10am to 2pm.
2. We can observe that around 11:30am to 12pm around 50% of Customers visit the branch for loan application on all the days except for Saturday where the time is between 10am to 11am for both Target 0 and 1
3. The loan defaulters have applied for the loan between 9:30am-10am and 2pm where as the applicants who repay the loan on time have applied for the loan between 10am to 3pm

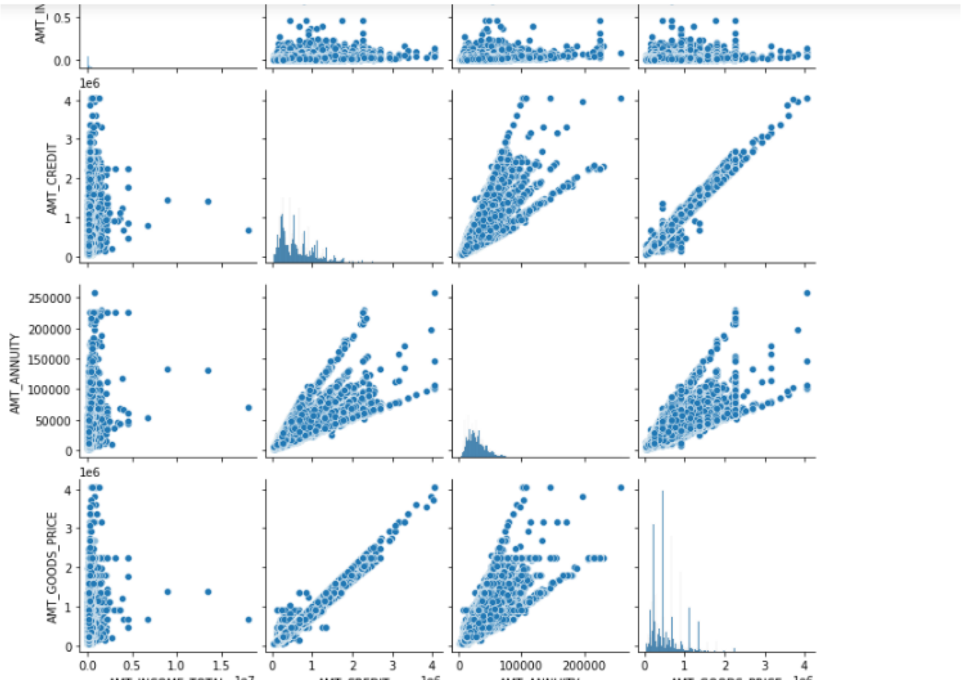
Bivarient Analysis between AGE_CATEGORY VS AMT_CREDIT



Conclusion:

1. The applicants between age group 25 to 65 have Credit amount of the loan less than 2500000 and are able to repay the loan properly
2. The applicants with less than 100000 Credit amount are with age group greater than 65 may be considered as loan defaulters
3. Most applicants who have Credit amount of the loan less than 1700000 are loan defaulters with 25 and less age

Pair plot of Amount Columns for Target 0



Conclusion: For Applicants who are able to replay the loan on time

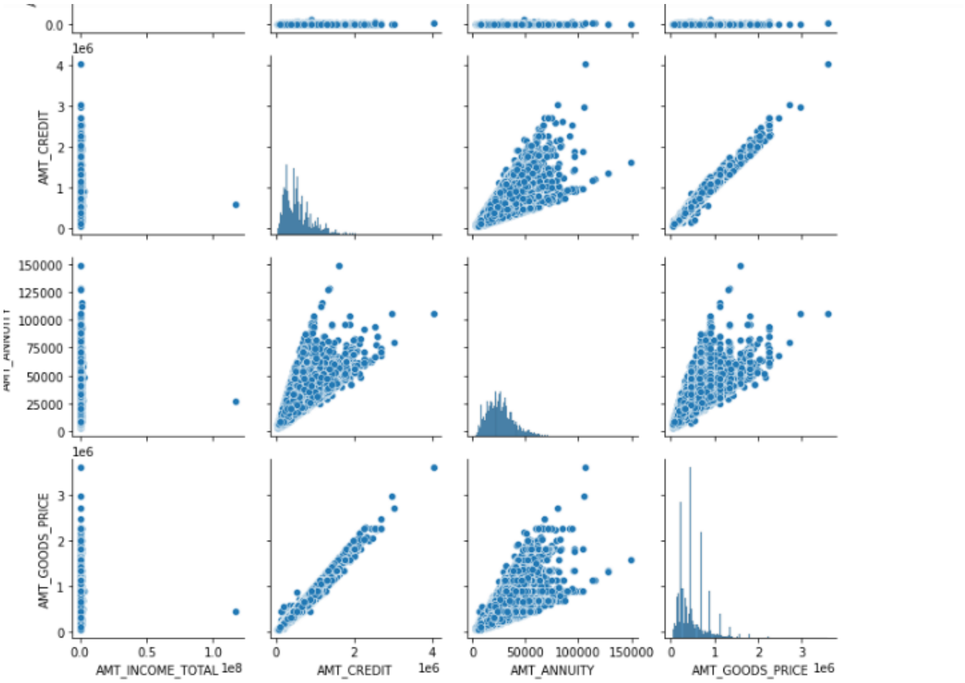
1. AMT_CREDIT Increases or varies linearly with AMT_GOODS_PRICE and AMT_CREDIT Increases with AMT_ANNUITY

2. AMT_ANNUITY Increases with Increases in AMT_GOODS_PRICE and AMT_Credit

3. AMT_GOODS_PRICE Increases with Increases in AMT_Credit and AMT_ANNUITY

4. AMT_INCOME_TOTAL has a drastic Increase with slight increase in AMT_CREDIT,AMT_ANNUITY,AMT_GOODS_PRICE

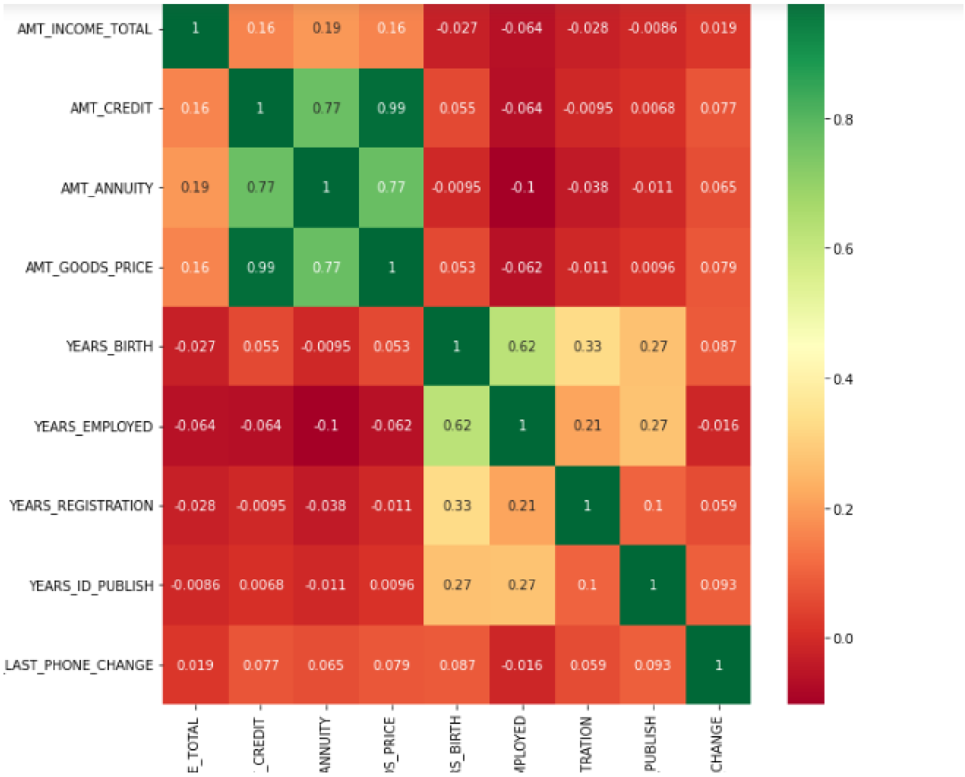
Pair plot of Amount Columns for Target 1



Conclusion: For Applicants who are unable to replay the loan on time¶

1. AMT_CREDIT Increases or varies linearly with AMT_GOODS_PRICE and AMT_CREDIT Increases with AMT_ANNUITY
2. AMT_ANNUITY Increases with Increases in AMT_GOODS_PRICE and AMT_Credit
3. AMT_GOODS_PRICE Increases with Increases in AMT_Credit and AMT_ANNUITY
4. AMT_INCOME_TOTAL has a drastic Increase with slight increase in AMT_CREDIT,AMT_ANNUITY,AMT_GOODS_PRICE

Co-relation between Numerical Columns

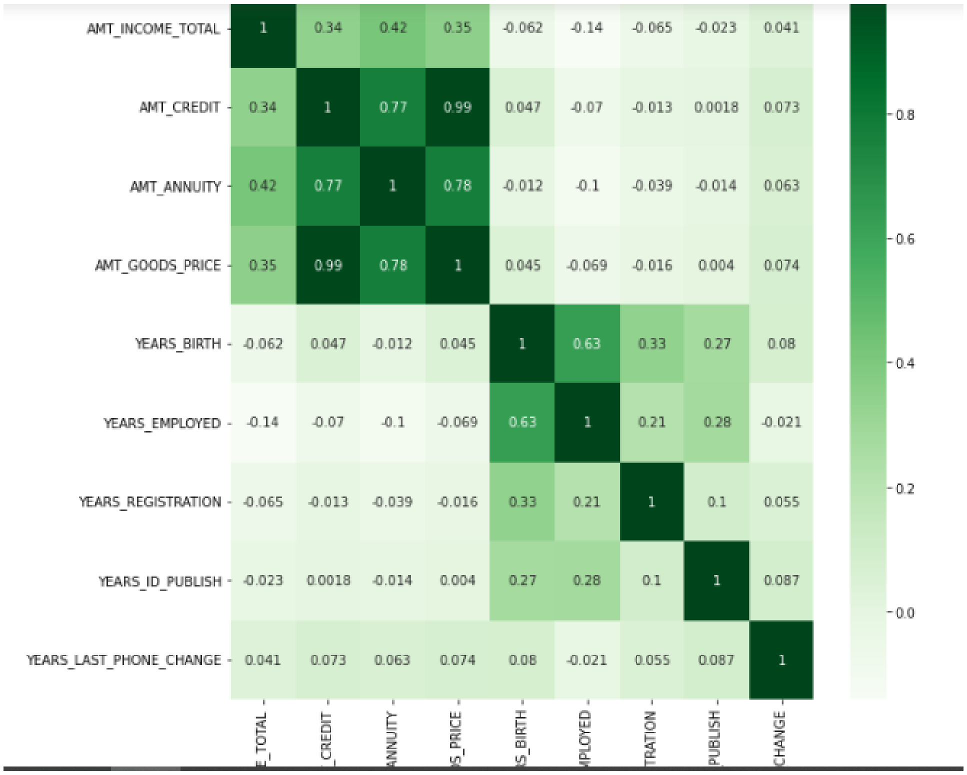


Conclusion:

1. AMT_INCOME_TOTAL - It has a positive corelation index of 0.16,0.19,0.16 with AMT_CREDIT,AMT_ANNUITY,AMT_GOODS_PRICE respectively.
2. AMT_CREDIT - Is has negative coreltaion index of 0.064 with YEARS_EMPLOYED and positive coreltaion index of 0.99,0.77 with AMT_GOODS_PRICE, AMT_ANNUITY respectively.
3. AMT_ANNUITY - Is has negative coreltaion index of 0.1 with YEARS_EMPLOYED and positive coreltaion index of 0.77 with AMT_CREDIT
4. AMT_GOODS_PRICE - It has a positive corelation with AMT_CREDIT,AMT_ANNUITY
5. YEARS_BIRTH - It has a positive corelation with YEARS_EMPLOYED, AMT_GOODS_PRICE and negative coreltaion with AMT_ANNUITY,AMT_INCOME_TOTAL
6. YEARS_EMPLOYED - Is has negative coreltaion index of 0.1 with AMT_ANNUITY and has a positive corelation with YEARS_REGISTRATION, YEARS_ID_PUBLISH
7. YEARS_REGISTRATION - It has a positive corelation with YEARS_ID_PUBLISH, YEARS_BIRTH, YEARS_EMPLOYED
8. YEARS_ID_PUBLISH - It has a positive corelation with YEARS_REGISTRATION and negative coreltaion with AMT_INCOME_TOTAL,AMT_ANNUITY

9. YEARS_LAST_PHONE_CHANGE - It has negative coreltaion with YEARS_EMPLOYED and positive corelation with AMT_GOODS_PRICE, YEARS_ID_PUBLISH

Co-relation for Numerical columns for Target 0

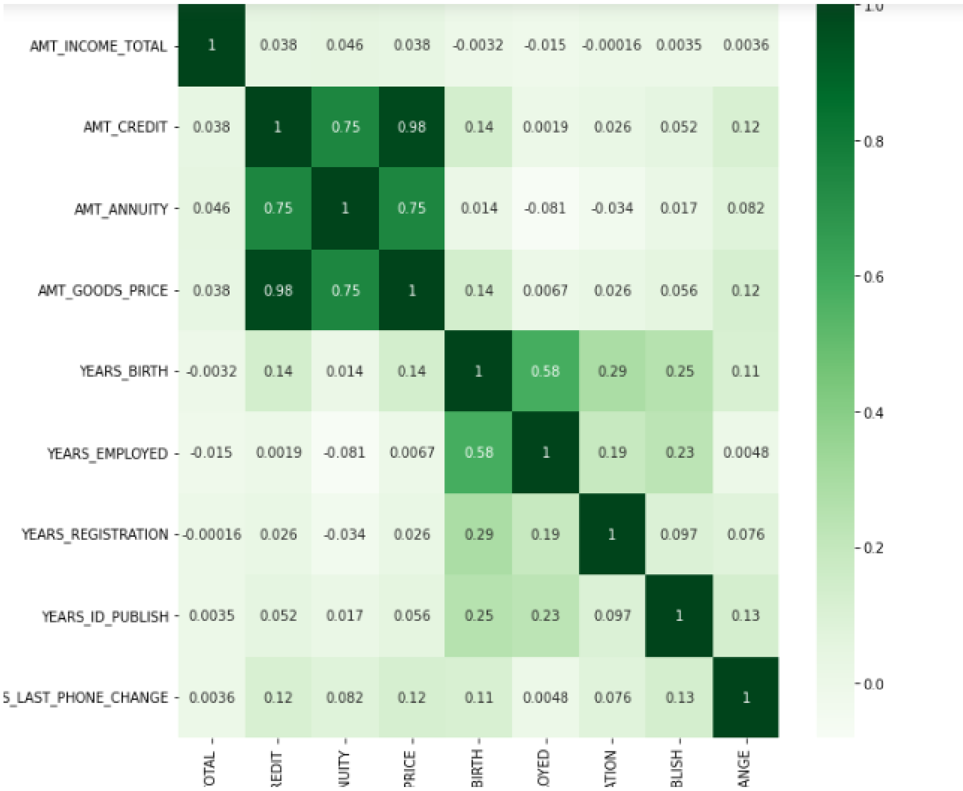


Conclusion:

1. AMT_INCOME_TOTAL - It has a positive corelation index of 0.34,0.42,0.35 with AMT_CREDIT,AMT_ANNUITY,AMT_GOODS_PRICE respectively and Negative with most of the other Year columns
2. AMT_CREDIT - Is has a strong positive coreltaion index of 0.99,0.77 with AMT_GOODS_PRICE, AMT_ANNUITY respectively.

3. AMT_ANNUITY - Is has positive coreltaion index of 0.77,0.78 with AMT_CREDIT,AMT_GOODS_PRICE respectively and Negative with most of the other Year columns
4. AMT_GOODS_PRICE - It has a strong positive corelation index 0.78,0.99 with AMT_ANNUITY, AMT_CREDIT

Co-relation for Numerical columns for Target 1



	OTAL	REDIT	NUITY	PRICE	BIRTH	OYED	ATION	BLISH	ANGE
AMT_INCOME_TOTAL	1	0.038	0.046	0.038	-0.0032	-0.015	-0.00016	0.0035	0.0036
AMT_CREDIT	0.038	1	0.75	0.98	0.14	0.0019	0.026	0.052	0.12
AMT_ANNUITY	0.046	0.75	1	0.75	0.014	-0.081	-0.034	0.017	0.082
AMT_GOODS_PRICE	0.038	0.98	0.75	1	0.14	0.0067	0.026	0.056	0.12
YEARS_BIRTH	-0.0032	0.14	0.014	0.14	1	0.58	0.29	0.25	0.11
YEARS_EMPLOYED	-0.015	0.0019	-0.081	0.0067	0.58	1	0.19	0.23	0.0048
YEARS_REGISTRATION	-0.00016	0.026	-0.034	0.026	0.29	0.19	1	0.097	0.076
YEARS_ID_PUBLISH	0.0035	0.052	0.017	0.056	0.25	0.23	0.097	1	0.13
5_LAST_PHONE_CHANGE	0.0036	0.12	0.082	0.12	0.11	0.0048	0.076	0.13	1

Conclusion:

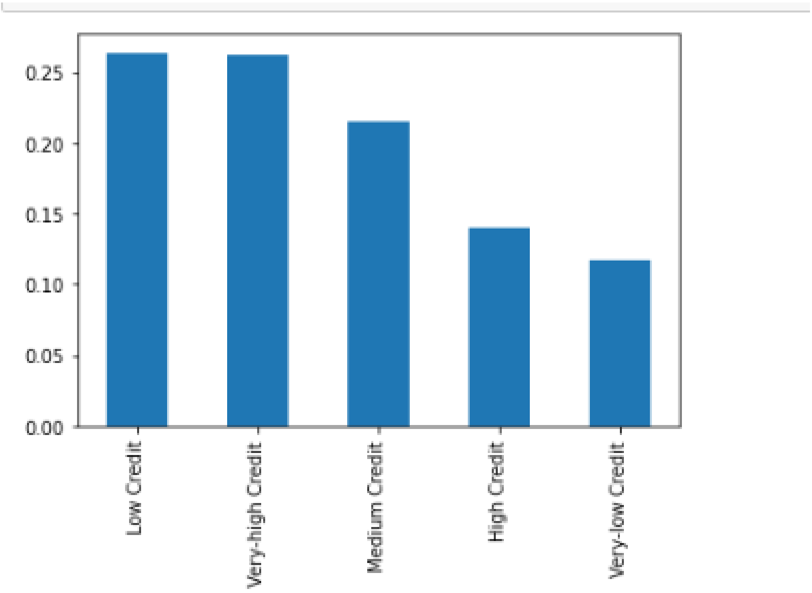
1. AMT_INCOME_TOTAL - It is less correlated with AMT_CREDIT,AMT_ANNUITY,AMT_GOODS_PRICE respectively
2. AMT_CREDIT - Is has a strong positive coreltaion index of 0.98,0.75 with AMT_GOODS_PRICE, AMT_ANNUITY respectively and also positive corelation with other Year Columns

3. AMT_ANNUITY - Is has positive coreltaion index of 0.75 with AMT_CREDIT,AMT_GOODS_PRICE and Negative with YEAR_EMPLOYED,YEAR_REGISTRATION
4. AMT_GOODS_PRICE - It has a strong positive corelation index 0.75,0.98 with AMT_ANNUITY, AMT_CREDIT and weak positive corelation with other Year columns

Analysis on Previous Application CSV

- Columns with null values more than 49 percent may give wrong insights, hence dropping them
- Binning of continuous variables

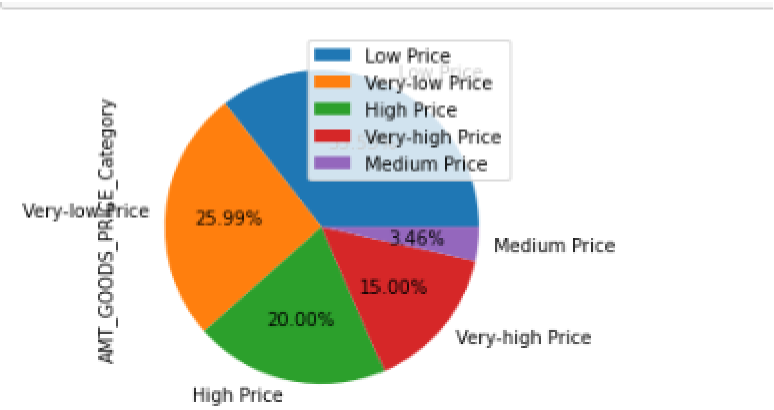
Binning AMT_CREDIT Column



Conclusion:

- The Credit amount of the loan for most applicants is either low(200000 to 400000) or Very High(above 800000)

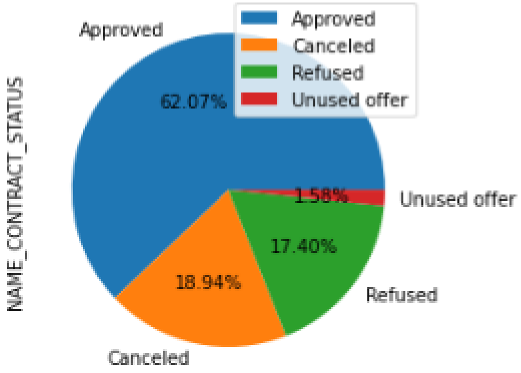
Binning AMT_GOODS_PRICE Column



Conclusion:

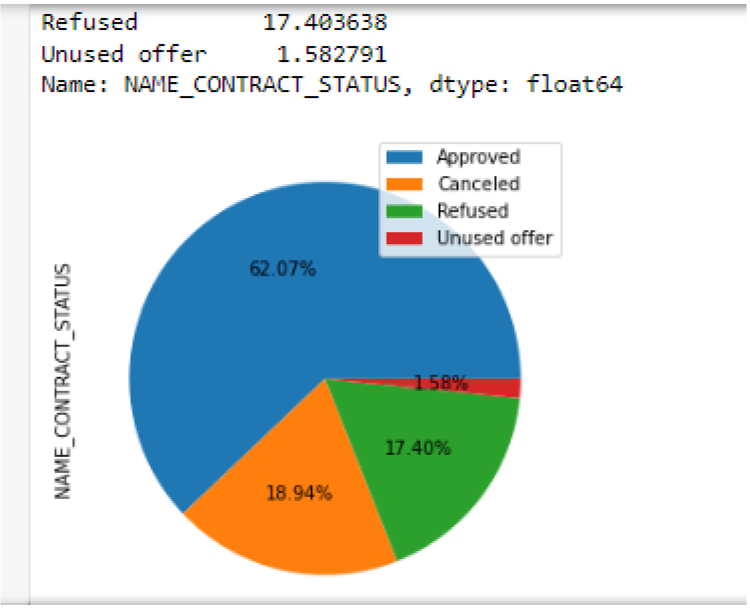
- **Most of the Applicants are have Low Goods Price**

Data Imbalance check



Conclusion >> 62% of the Applicants have the loan approved, 19%, 17% applicants are rejected or canceled and 2% are unused¶

Plot on Categorical columns



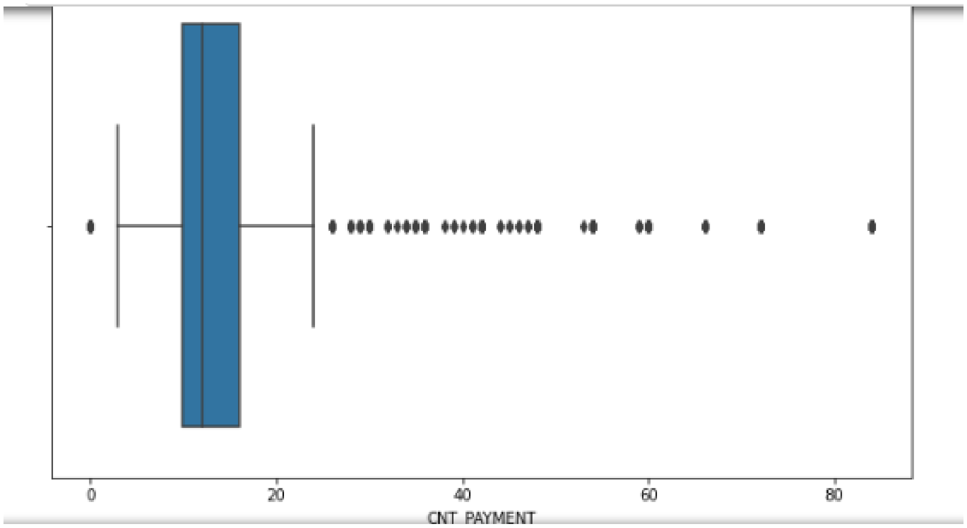
```
Refused          17.403638
Unused offer      1.582791
Name: NAME_CONTRACT_STATUS, dtype: float64
```

Conclusion:

1. NAME_CONTRACT_TYPE - 45% Applicants received Cash loans,44% Applicants received Consumer loans,12% received Revolving during previous application
2. WEEKDAY_APPR_PROCESS_START - All the days have almost equal number of previous loan application
3. NAME_CONTRACT_STATUS - 62% of applications are approved, 19% Cancelled, 17% Refused and 2% unused
4. NAME_PAYMENT_TYPE - 62% of Payment type are Cash through bank, 32% Other modes
5. NAME_CLIENT_TYPE - 74% of Applicants are Repeaters, 18% are New applicants, 8% are Refreshed Appplicants
6. NAME_SELLER_INDUSTRY - 51% are from other Industries, 24%,17% are from Consumer electronics, Connectivity Industry respectively
7. CHANNEL_TYPE - 43% Channel type is Credit and cash offices, 29% are country wide
8. NAME_YIELD_GROUP - Majority of the yield group are others

9. PRODUCT_COMBINATION - Most used PRODUCT COMBINATION is Cash followed by POS household with interest, POS mobile with interest

Plot on Numerical columns



Conclusion: Few Columns are with outliers are below

1. HOUR_APPR_PROCESS_START has few outliers and there small difference between mean and median

2. AMT_CREDIT Column has a few outliers and there is a huge difference between the 99th percentile and the max value, also we could see huge variation in mean and median due to outliers

3. AMT_ANNUITY Column has a few outliers and there is a huge difference between the 99th percentile and the max value, also we could see significant variation in mean and median due to outliers

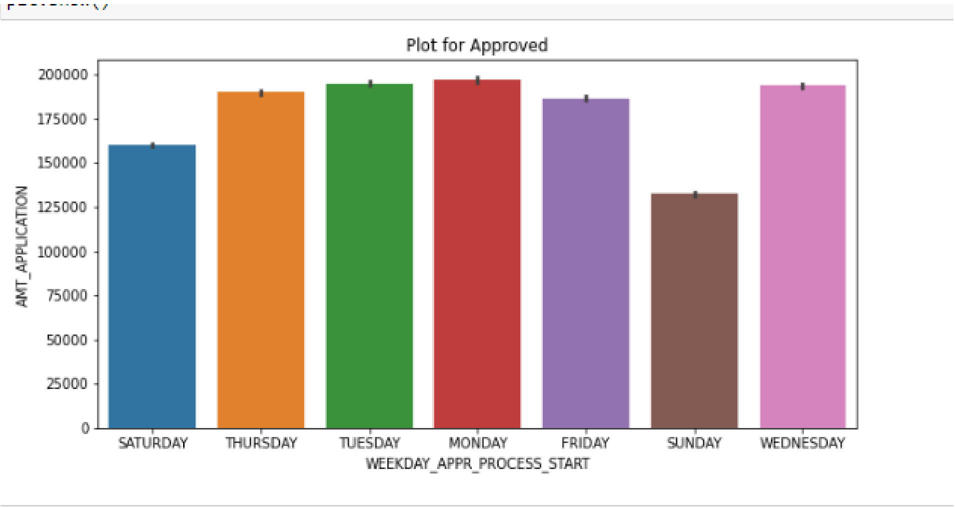
4. AMT_GOODS_PRICE Column has a few outliers and there is a huge difference between the 99th percentile and the max value, also we could see significant variation in mean and median due to outliers

5. AMT_APPLICATION Column has a few outliers and there is a huge difference between the 99th percentile and the max value, also we could see huge variation in mean and median due to outliers

6. CNT_PAYMENT Column has few outliers and there small difference between mean and median

7. DAYS_DECISION has few outliers and there small difference between mean and median

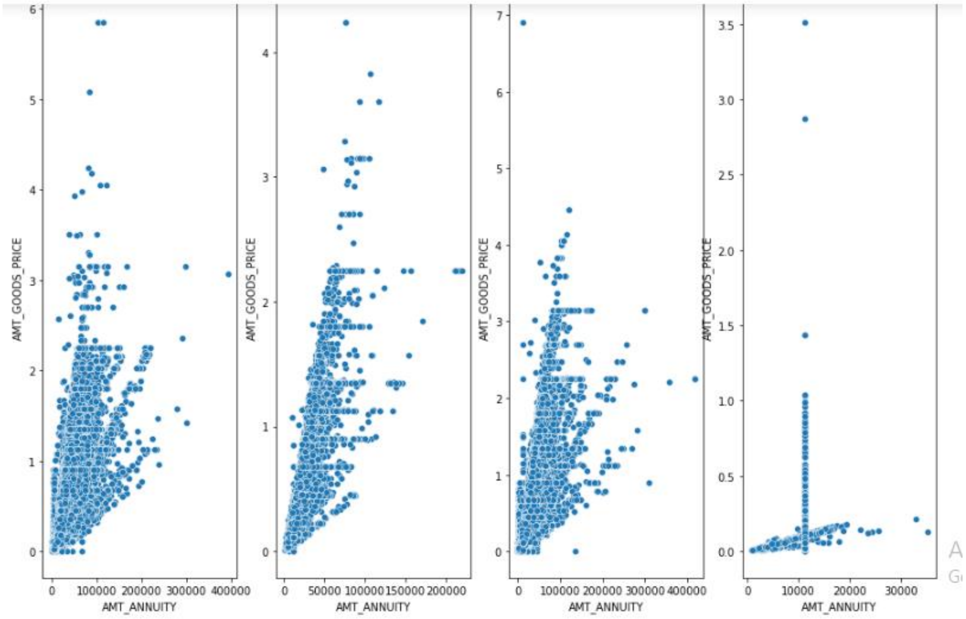
Bivarient Analysis between WEEKDAY_APPR_PROCESS_START VS AMT_APPLICATION



Conclusion:

1. The Credit Amount of applicants with approved status is high on Monday and Wednesday than other days, and least on Sunday
2. The Credit Amount of applicants with cancelled status is high on Sunday and almost equal on other days
3. The Credit Amount of applicants with rejected status is least on Sunday and more on Monday and Wednesday
4. The Credit Amount of applicants with unused offer status is almost equal on all days

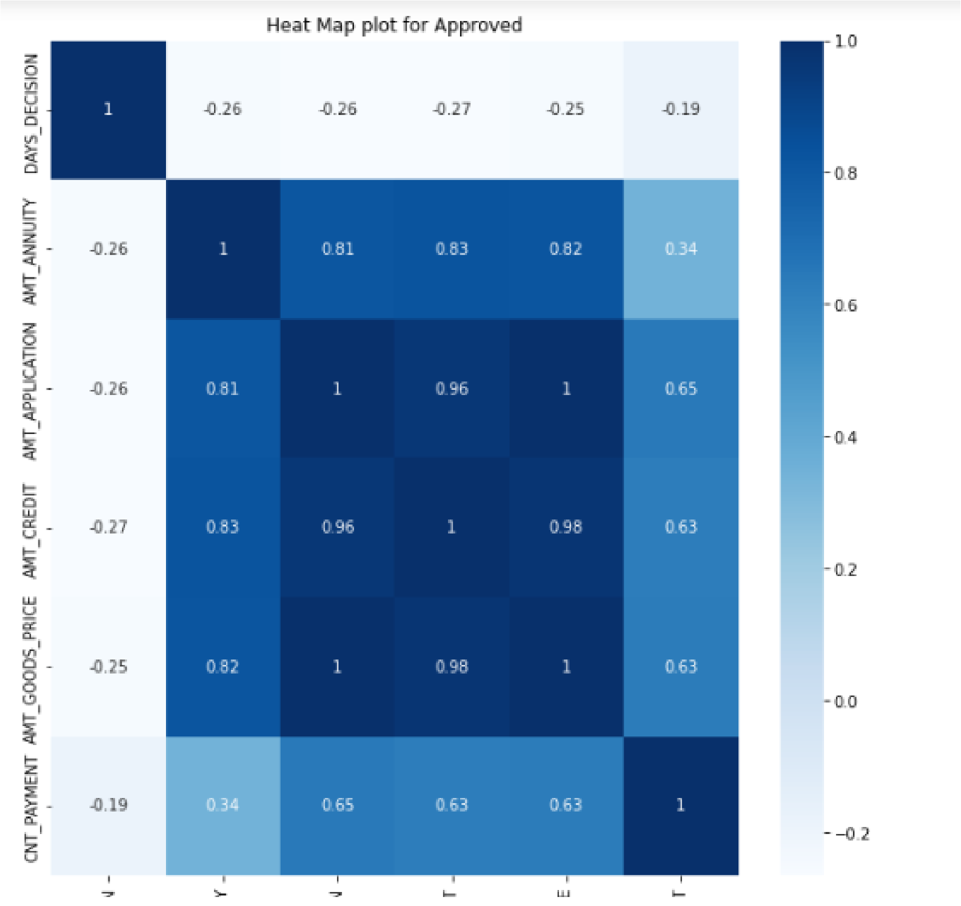
Bivarient Analysis between AMT_ANNUITY VS AMT_GOODS_PRICE



Conclusion:

1. For loan status as Approved,Refused,Cancelled Amount of annuity increases with goods price
2. For loan status as Refused it has no linear relationship

Co-relation between Numerical Columns

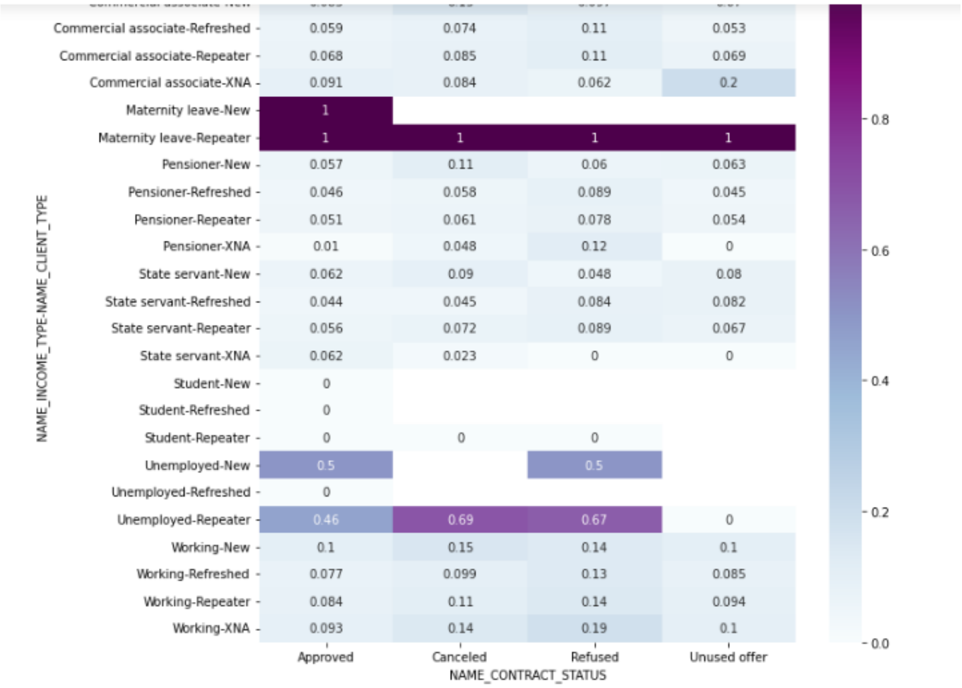


Conclusion:

1. AMT_APPLICATION has higher Corelation with AMT_CREDIT and AMT_GOODS_PRICE,AMT_ANNUITY
2. DAYS_DECISION has negative Corelation with AMT_GOODS_PRICE,AMT_CREDIT, AMT_APPLICATION,CNT_PAYMENT,AMT_ANNUITY

Merge the Application and Previous Application Dataframes

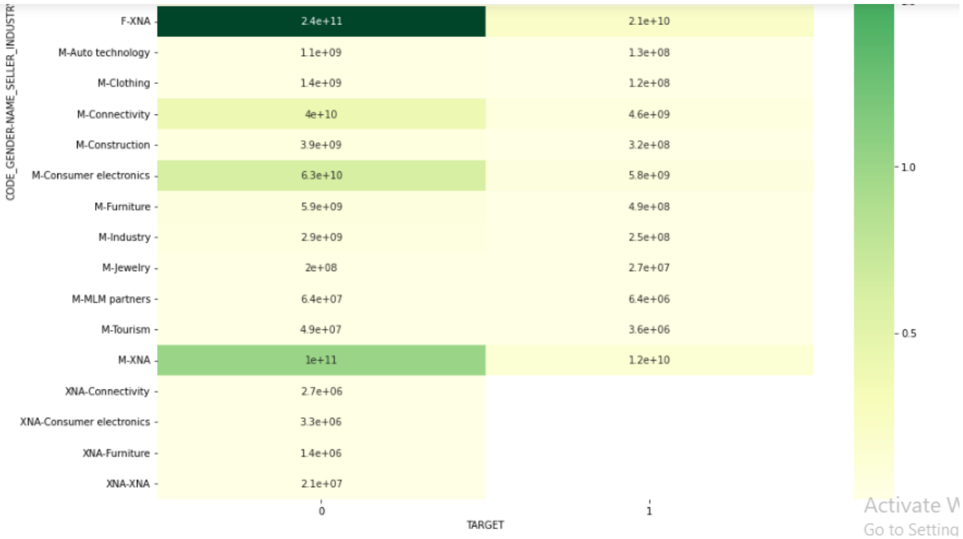
- **Plot and Analysis between NAME_INCOME_TYPE, NAME_CLIENT_TYPE, NAME_CONTRACT_STATUS**



Conclusion:

1. Applicants with income type Maternity leave and client type New are having more chances of getting the loan approved
2. Applicants with income type Maternity leave, Unemployed and client type Repeater are having getting the loan cancelled
3. Applicants with income type Maternity leave, Unemployed and client type Repeater are having getting the loan Refused
4. Applicants with income type Maternity leave and client type Repeater, Working and client type New are not able to utilize the Bank's offer

- **Plot and Analysis between TARGET, CODE_GENDER, NAME_SELLER_INDUSTRY**



Conclusion:

1. Female Applicants from Other Seller industry are more likely to repay the loan on time
2. Male Applicants from furniture industry are less likely to repay the loan on time

OVERALL ANALYSIS AND CONCLUSION:

- Female Applicants from Other Seller industry are more likely to repay the loan on time
- Male Applicants from furniture industry are less likely to repay the loan on time
- Applicants with income type Maternity leave and client type New are having more chances of getting the loan approved
- Applicants with income type Maternity leave, Unemployed and client type Repeater are having getting the loan cancelled

- Applicants with income type Maternity leave, Unemployed and client type Repeater are having getting the loan Refused
- Applicants with income type Maternity leave and client type Repeater, Working and client type New are not able to utilize the Bank's offer

---

Thank You

Submitted by,

Hanumanth A