UPGRAD ASSIGNMENT – EDA CASE STUDY

Problem Statement : To understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default for a Loan lending company.

Submitted By:

- **Aruna Durai**
- **Prerana V**

STEPS FOLLOWED IN EDA CASE STUDY

1. We first read application.csv as it is the primary file.
2. We Inspected the dataset (“dfApp” is the data frame in which application.csv was loaded.
3. Routine Data Check was carried out for dfApp.(application_data.csv)

```
In [ ]: print(dfApp.columns) ##inspecting the columns
        print(dfApp.shape) ##inspecting the shape
        print(dfApp.dtypes) ##inspecting the datatypes of variables
        print(type(dfApp)) ##inspecting the variable on application dataset.

        print(dfApp.info()) ##list down all the columns along with name ,no of non null values,datatype,memory usage
        print(dfApp.describe()) ##describe dataset's mean,std,min,25%,50%,75%,max
```

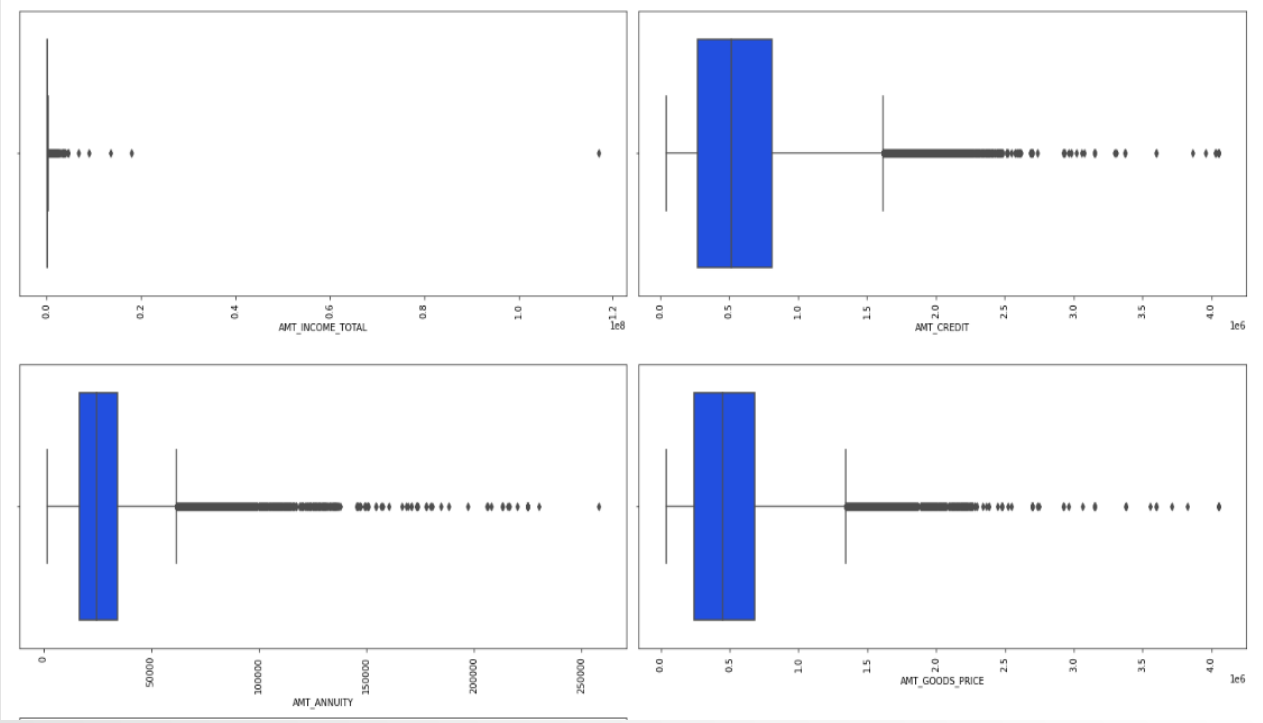
We removed the columns which having null percentage. Greater or equal to 50%. Around 41 columns got removed and totally 81 columns were remaining

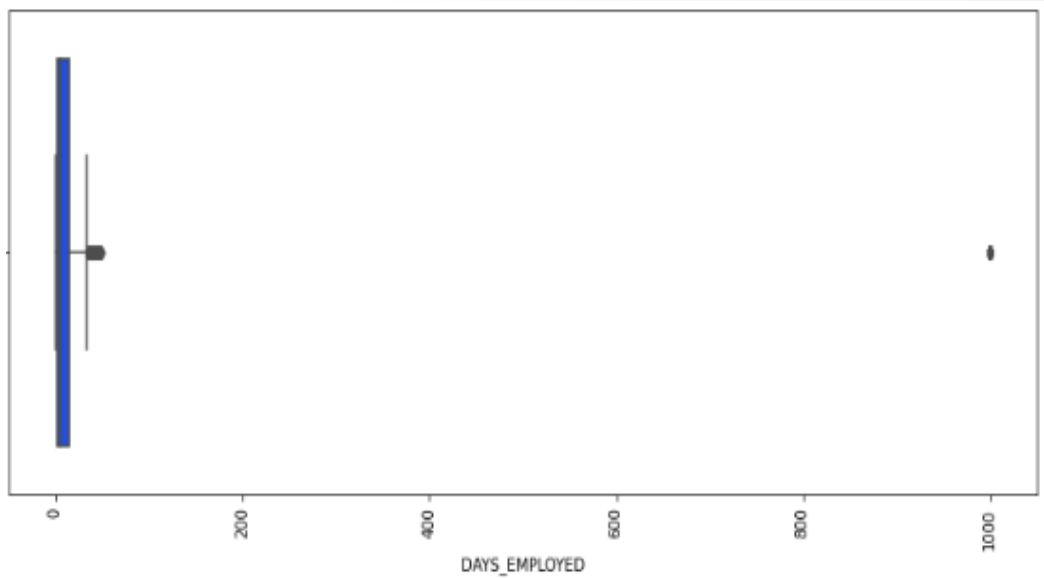
1. We found best metric to impute the columns with less null percent (13% approximately).
2. 'AMT_REQ_CREDIT_BUREAU_DAY ', 'AMT_REQ_CREDIT_BUREAU_DAY', ' ‘AMT_REQ_CREDIT_BUREAU_MON' , 'AMT_REQ_CREDIT_BUREAU_QRT', 'AMT_REQ_CREDIT_BUREAU_YEAR' were imputed with median (since these columns have high outliers , imputing with mean will affect the entire columns)
3. We converted the column’s data type into appropriate data type.

E.g.: dfApp['HOUR_APPR_PROCESS_START'] = pd.to_timedelta (dfApp.HOUR_APPR_PROCESS_START , unit=‘h’)

STEPS FOLLOWED IN EDA CASE STUDY

- We found outliers for 5 numerical variables .'AMT_INCOME_TOTAL' , 'AMT_CREDIT' , 'AMT_ANNUITY' , 'AMT_GOODS_PRICE', 'DAYS_EMPLOYED'





- Binning of continuous variable.
  - DAYS_BIRTH, AMT_INCOME_TOTAL,AMT_CREDIT columns have been binned.

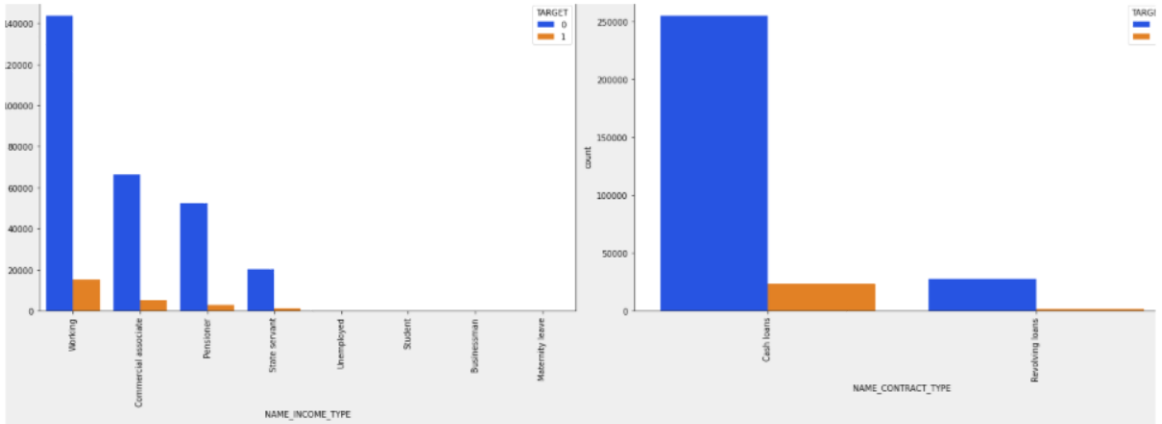
STEPS FOLLOWED IN EDA CASE STUDY

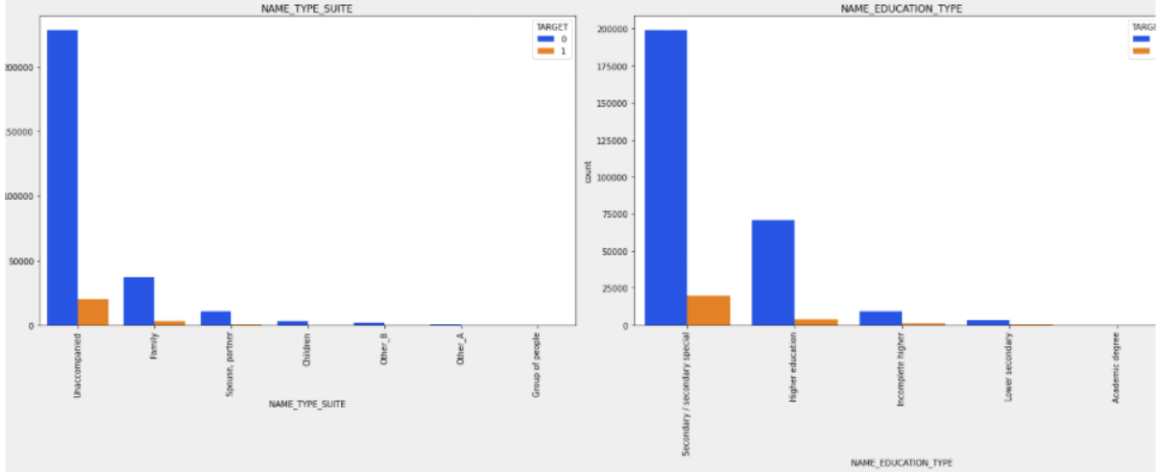
ANALYSIS OF APPLICATION_DATA:

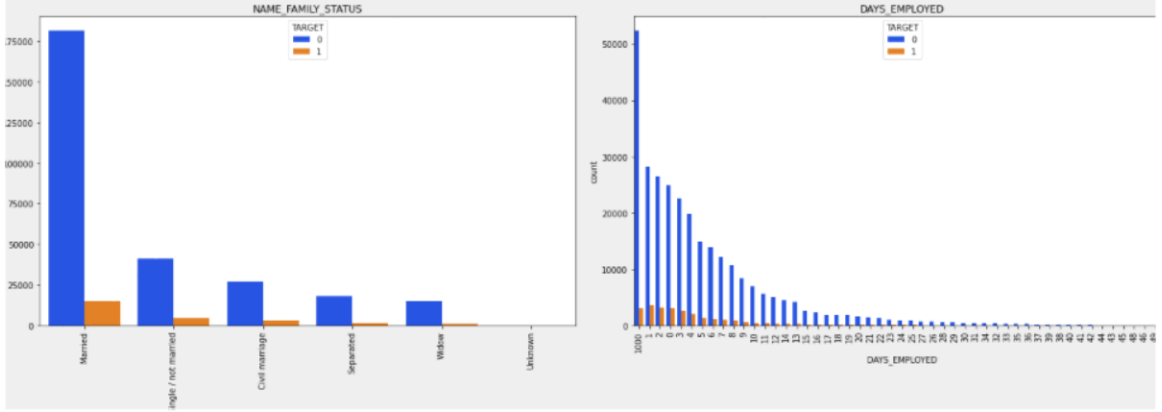
- We found data imbalance for “TARGET” & “NAME_CONTRACT_TYPE” columns.
  - Data Imbalance Ratio for Target = 11.39 Contract Type Data Imbalance=9.5
- Splitting into Target 1 and Target 0 : We have split the entire data frame into Target 1 and Target 0 w.r.t the Target column.
- TARGET - 1
  - We can see that Female have more difficulties in paying loan than male
  - We can see that people are majorly completed secondary education
  - We can see that people mostly reside in apartment.
  - Majority of people have own house but don't own car.
  - We can see that almost they are married out of target with payment difficulties.
  - Mostly target1 don't have child and are working.
  - Most of the people belong to the labor class and reside in region with rating 2(with or without considering city).
- TARGET – 0
  - We can see that target0 also more or less follows the target1 variable’s pattern
- Correlation
  - We found that the correlations variables are same for both target0 and target1 even though the correlation values are different.

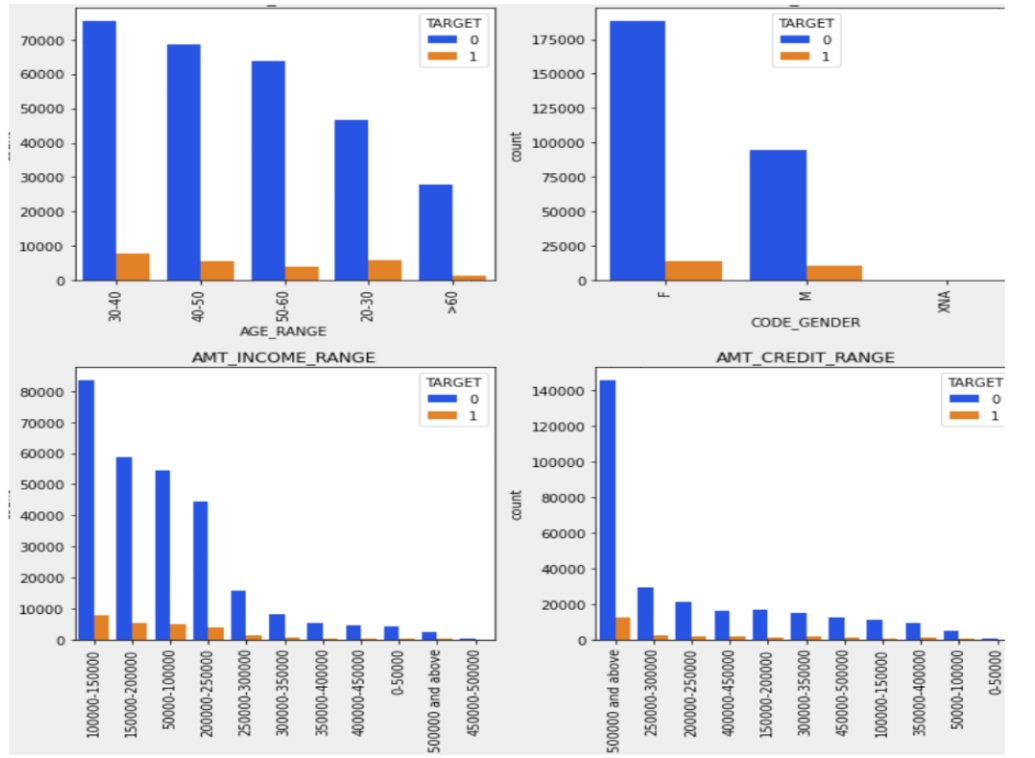
UNIVARIATE ANALYSIS OF APPLICATION_DATA.CSV

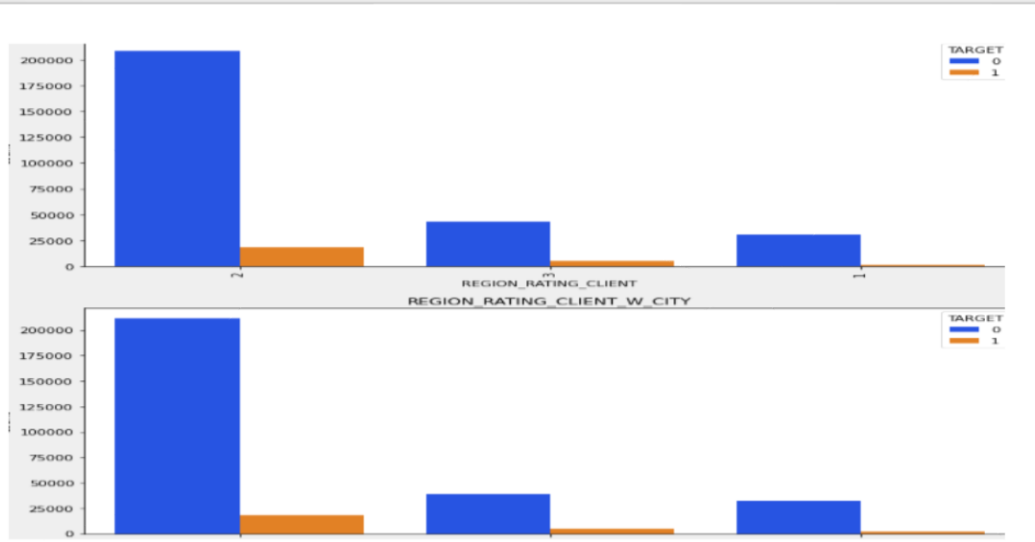
- **When we perform univariate analysis on the complete data frame without splitting into target 1 and target 0 , we get the following insights:**
- **People with less payment difficulties are found to be present in these categories:**
  - **Working Class**
  - **People who apply for Cash loans**
  - **People who are unaccompanied when they visit the bank.**
  - **People with secondary/secondary special education.**
  - **People who are married.**
  - **Work experience with around 2-3 years.**
  - **People owning a Business Entity.**
  - **People who come under Labor class in Occupation type.**
  - **The rating of region with/without city where the client is residing is found to be 2.**
  - **Females**
  - **Age-Range : 30-40**
  - **Amount Income range : 100000 - 150000 and Amt_Credit_Range = 500000 and above.**


TARGET
0
1
Working
Commercial associate
Pensioner
State servant
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE
count
Cash loans
Revolving loans
NAME_CONTRACT_TYPE


NAME_TYPE_SUITE
TARGET
0
1
Unaccompanied
Family
Spouse, partner
Children
Other_B
Other_A
Group of people
NAME_TYPE_SUITE
NAME_EDUCATION_TYPE
count
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE


NAME_FAMILY_STATUS
TARGET
0
1
Married
Civil marriage
Separated
Widow
Unknown
DAYS_EMPLOYED
count
DAYS_EMPLOYED

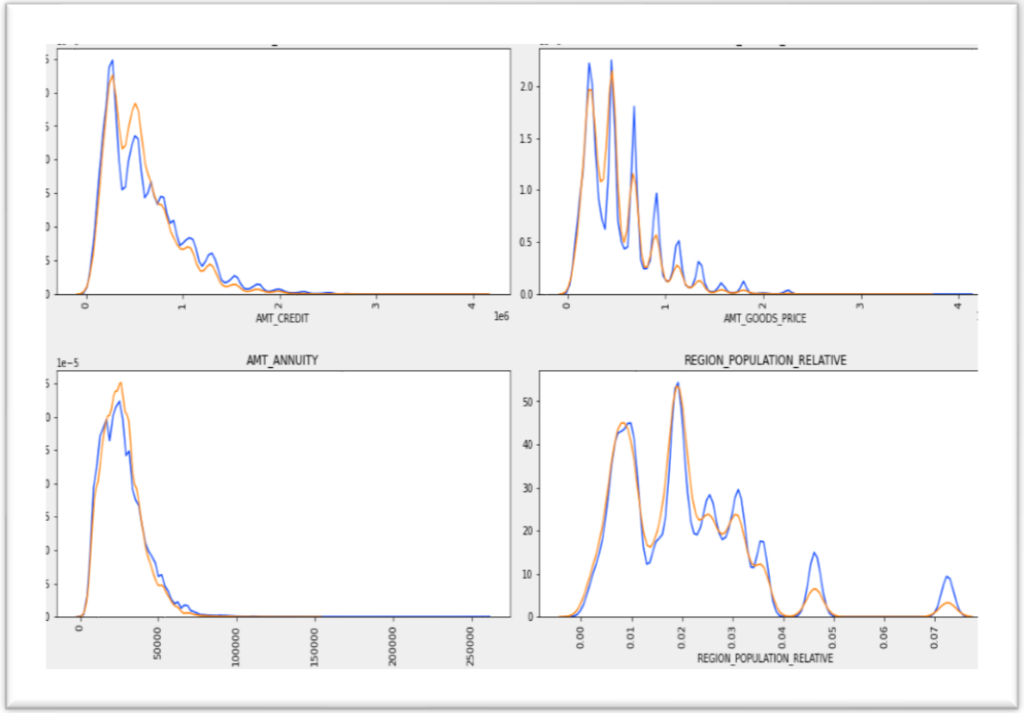

TARGET
0
1
30-40
40-50
50-60
20-30
>60
AGE_RANGE
count
F
M
XNA
CODE_GENDER
AMT_INCOME_RANGE
100000-150000
150000-200000
50000-100000
200000-250000
250000-300000
300000-350000
350000-400000
400000-450000
0-50000
500000 and above
450000-500000
AMT_CREDIT_RANGE
500000 and above
250000-300000
200000-250000
400000-450000
150000-200000
300000-350000
450000-500000
100000-150000
350000-400000
50000-100000
0-50000


TARGET
0
1
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY

UNIVARIATE ANALYSIS – NUMERICAL COLUMNS

- When we perform univariate analysis on the complete data frame w.r.t numerical variables without splitting into target 1 and target 0 , we get the following insights:

- People with less payment difficulties are found to be present in these categories:
  - Mean population density is found to be 0.02 with max being 0.07
  - Mean AMT_GOODS_PRICE is found to be 538396 with max being 4050000
  - Mean AMT_CREDIT is found to be 5999025 with max being 4050000
  - Mean AMT_ANNUITY is found to be 27108 with max being 258025



BIVARIATE ANALYSIS – NUMERICAL COLUMNS

- Outliers in Income category are more among people who have secondary education and are married.
- The Income Mean is higher for academic degree ,higher education in all family status category.
- The credit Mean, Annuity mean is higher for academic degree ,higher education in all family status category.
- For Income variable, majority of the people with family status single and having academic degree occupies 3rd quartile. The credit amount mean for married people is mostly greater than any other family status(even civil marriage is considered as married overall)

MERGING THE DATA FRAMES

- Reading previous_application.csv and Merging both previous_application.csv and application.csv
  - To merge the data , we choose the inner merge as we focus on the data that are common in both the columns. Even if we chooseleft merge, it won't affect the data as the columns that are not common will be filled with NA values and it won't add any value to our analysis.Hence, inner merge is the best option.
  - This data frame is stored as merged_df.
- We carried out Routine Data Check on merged_df.
- To Simplify our analysis, we copied the merged_df to analysis_df and reduced the columns which were not required.
- Finally we considered 'NAME_CONTRACT_TYPE_y', 'NAME_CASH_LOAN_PURPOSE', 'NAME_PAYMENT_TYPE','CODE_REJECT_REASON','NAME_TYPE_SUITE_y','NAME_CLIENT_TYPE','NAME_CONTRACT_STATUS','PRODUCT_COMBINATION','NAME_YIELD_GROUP','NAME_GOODS_CATEGORY','NAME_PORTFOLIO', 'NAME_PRODUCT_TYPE','CHANNEL_TYPE' columns from previous_application.csv for analysis.

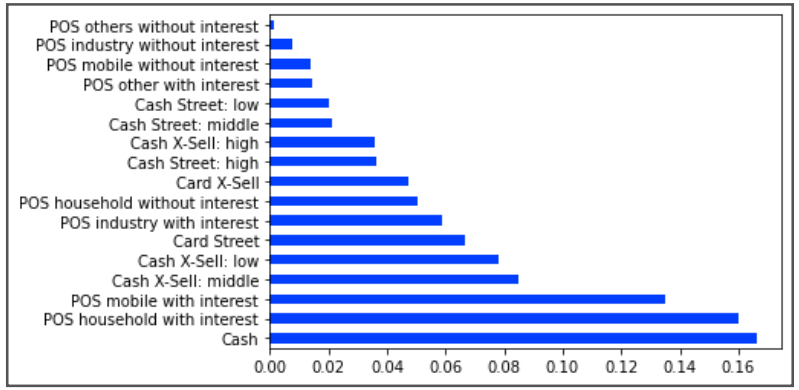
UNIVARIATE ANALYSIS ON MERGED_DATA

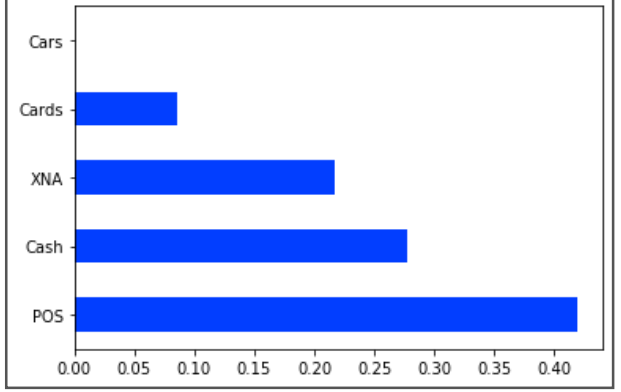
- Note on Univariate analysis:
  - After completing the univariate analysis for all these variables, we only consider the following columns going forward: (We ignore the other columns because majority of values are XNA or XAP.)
    NAME_CONTRACT_TYPE_y
    NAME_PAYMENT_TYPE
    NAME_TYPE_SUITE_y
    NAME_CLIENT_TYPE
    NAME_CONTRACT_STATUS
    PRODUCT_COMBINATION
    NAME_PORTFOLIO
    CHANNEL_TYPE

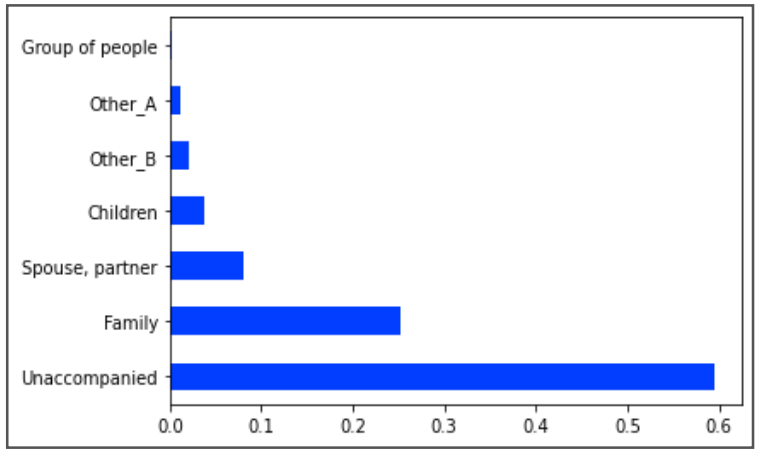
UNIVARIATE ANALYSIS

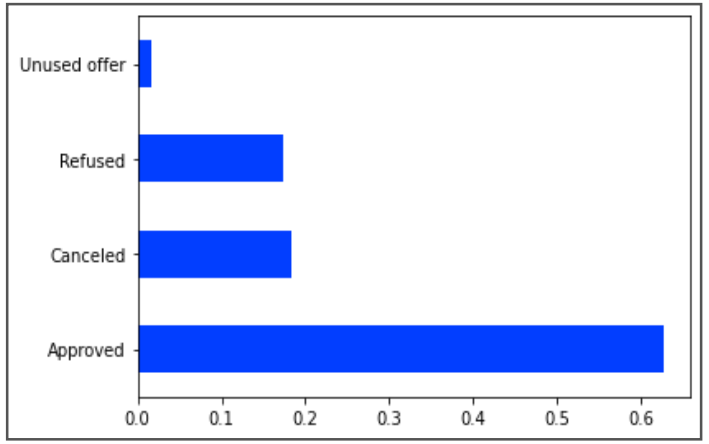
Overall summary of Univariate analysis of prev_application_data.csv variables:

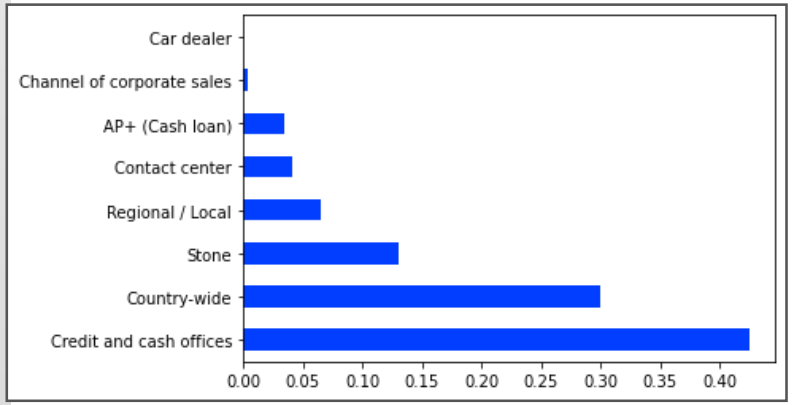
- **Around 43% of applications were acquired through Credit and Cash offices.**
- **Around 45% of the loan applications were found to be under the Category 'POS'.**
- **It is found that ROI (Rate of interest) were found almost equal for high and middle interest values.**
- **Product_Combination were observed at around 16% for both Cash and POS with interest.**
- **It is observed that around 62% of applicants were approved of loan application.**
- **It is observed that around 70% of applicants were repeaters.**
- **It is mostly seen that, in most of the cases , all loan applicants were unaccompanied while applying for loans.**
- **Around 62% of people applied loans for Cash through the bank.**

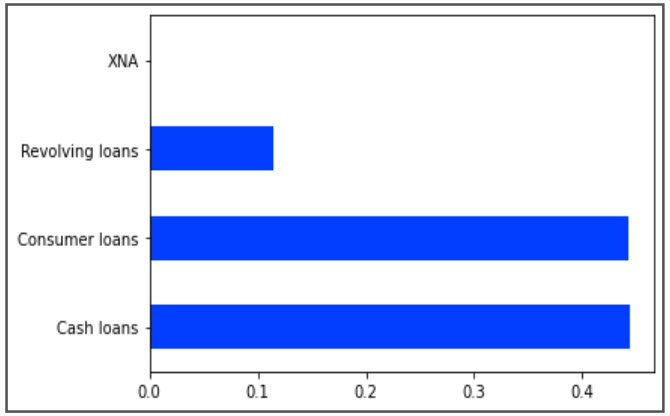






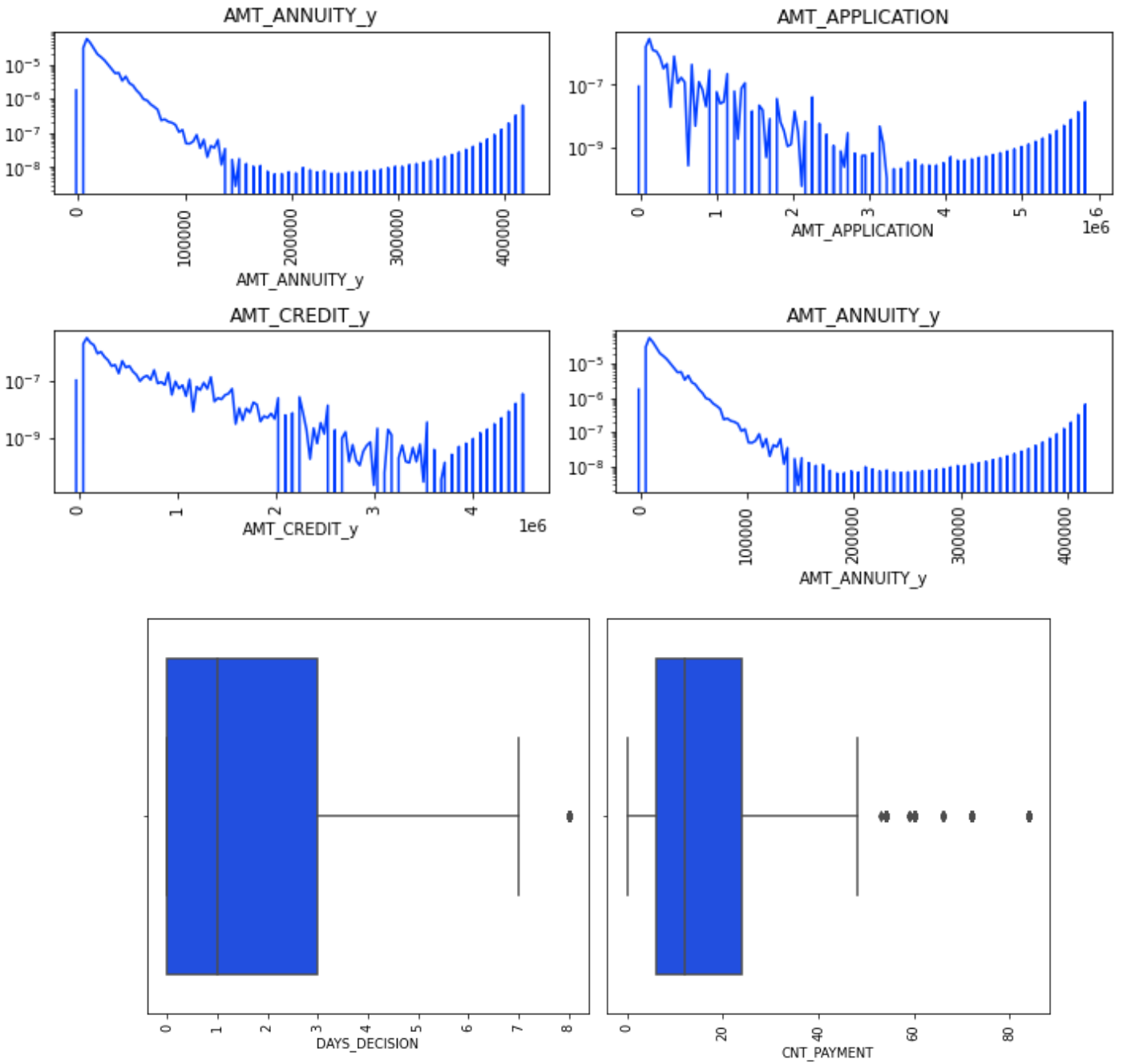






UNIVARIATE ANALYSIS

- Average number of days taken by the bank to convey the decision = 2.
- Mean amount of annuity = 15837 And highest value of annuity = 418058
- Mean amount of application = 175243 And highest value of loan application = 5850000
- Mean amount of annuity = 196354 And highest value of annuity = 4509688
- Average term for all loans is around 16 (months/years as per the unit taken in the dataset)



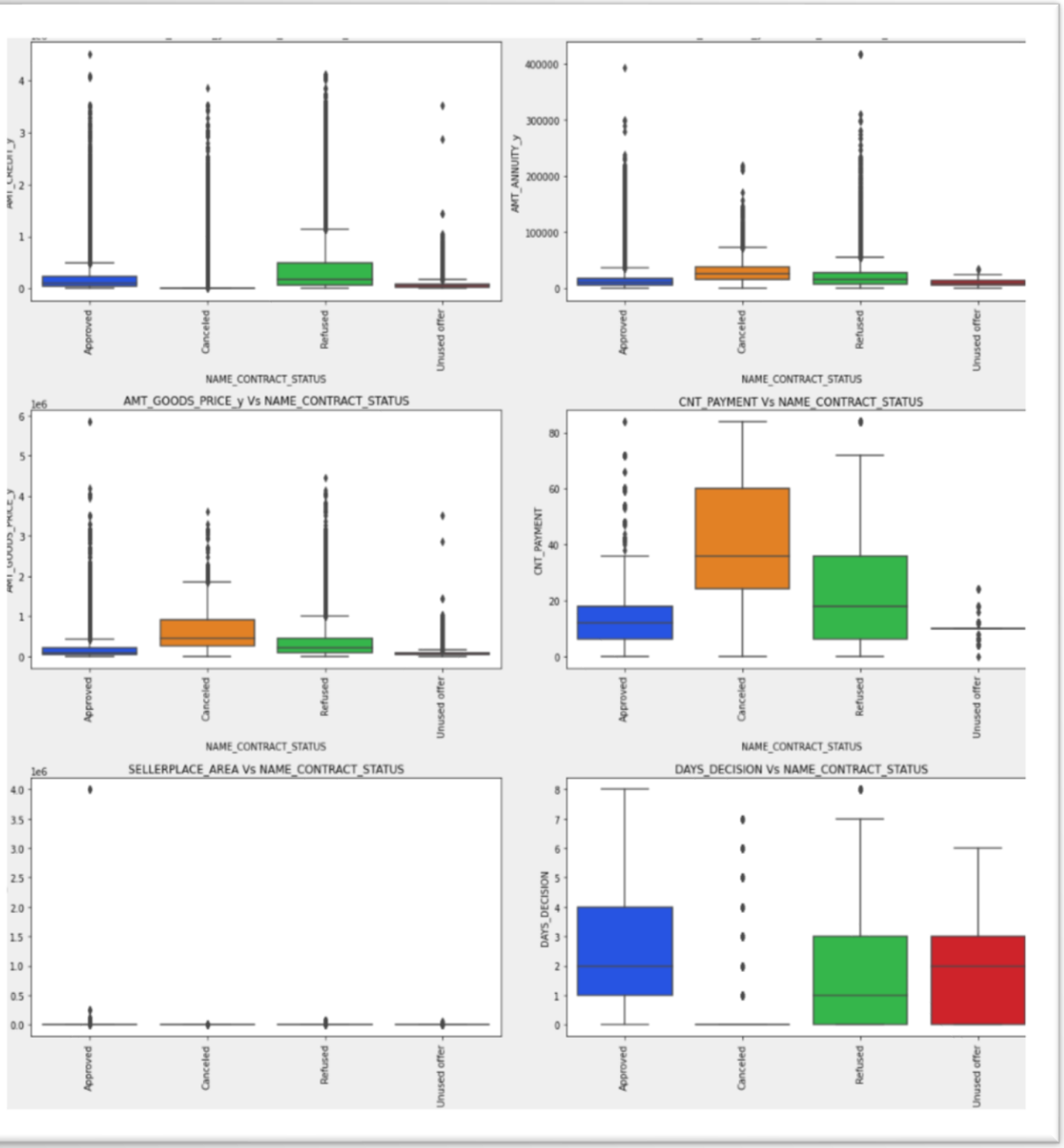
BIVARIATE ANALYSIS ON MERGED DATA.

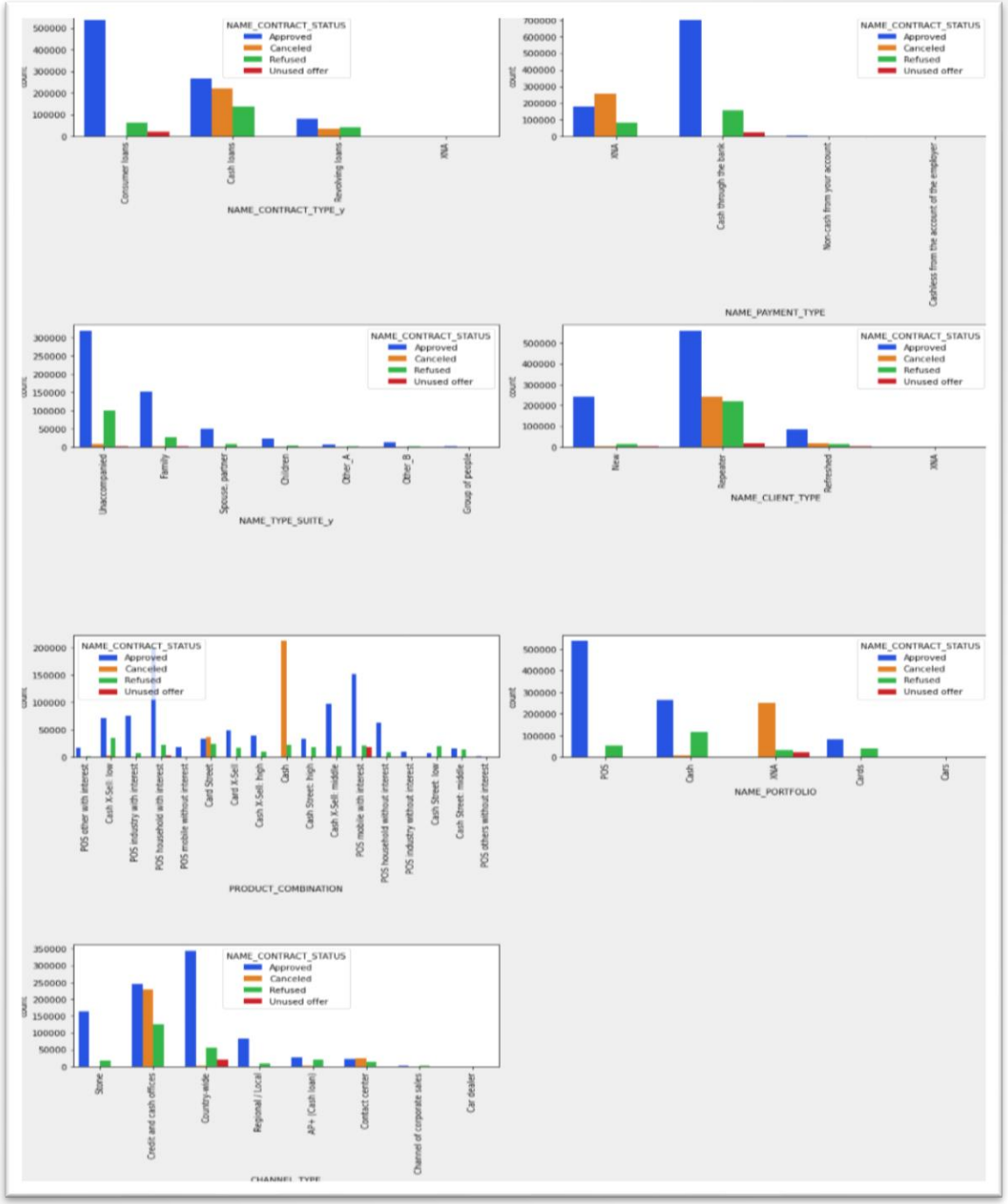
Insights from Bi-Variate analysis w.r.t current application data :

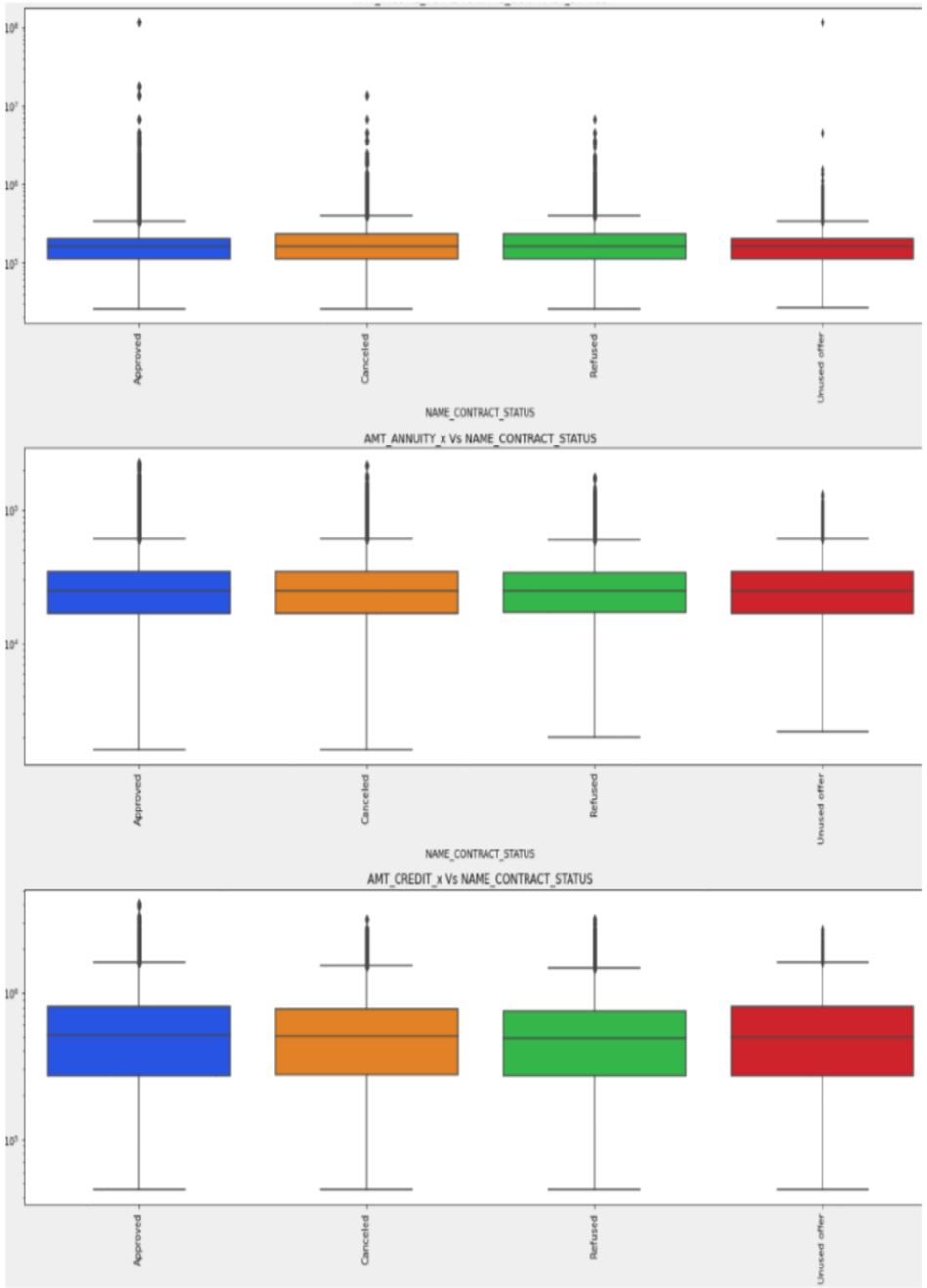
- **People who have a age range between 30-40 and have 2-3 years of work experience, with an income range of 100000 - 150000 and a credit range of 500000 and above have greater chances of approval**
- **It is found that Business Entity and Labor class loans are approved the highest.**
- **It is found that working class with secondary education , who are married and live in their house / apartment have better chances of loan approval.**
- **It is found that Females who own a house and do not own a car have a greater probability of their loan being approved.**
- **When we analyze for numerical columns , mean value for all amount columns remains to be the same.**

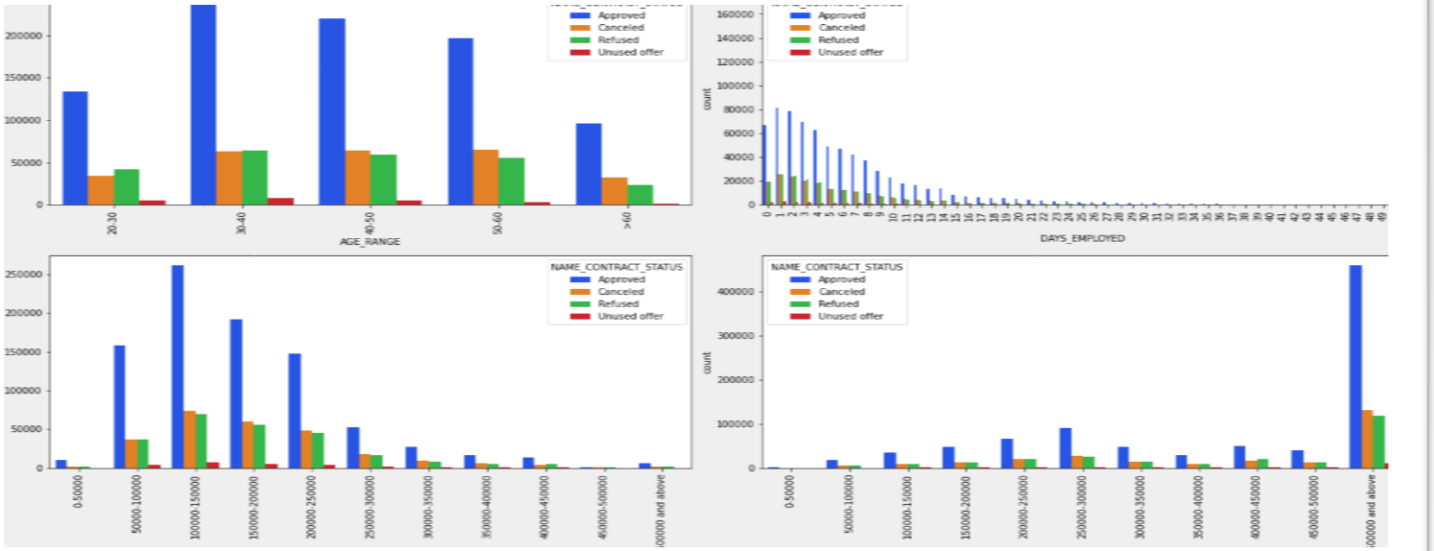
Insights from Bi-Variate analysis w.r.t previous application data:

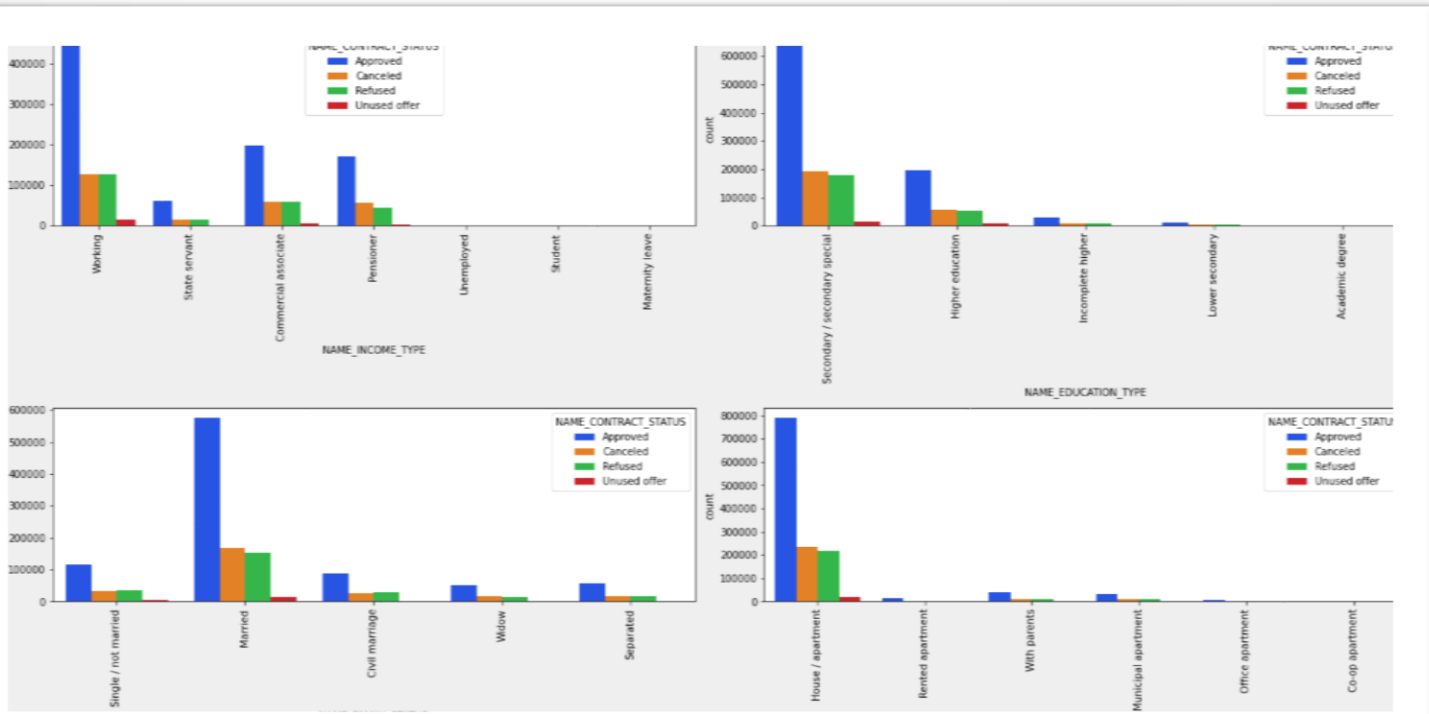
- **People who are unaccompanied have a greater rate of loan approval**
- **People with consumer and cash loans through the bank have a greater rate of approval**
- **Repeaters have a greater rate of loan approval**
- **People having a NAME_PORTFOLIO = POS have a greater rate of loan approval.**
- **People with Product_Combination POS with Household_interest have a greater rate of approval but cancel rate is greater for Product_combination having cash**
- **Country-Wide offices have a greater approval rate.**
- **Mean for all AMT_GOODS_PRICE is found to be higher for cancelled applications.**
- **The term (CNT_PAYMENT) is found to be higher in case of Cancelled applications.**
- **Mean for AMT_ANNUITY AND AMT_CREDIT is found to be almost equal and lies in the range of 100000**


AMT_GOODS_PRICE_y Vs NAME_CONTRACT_STATUS
CNT_PAYMENT Vs NAME_CONTRACT_STATUS
SELLERPLACE_AREA Vs NAME_CONTRACT_STATUS
DAYS_DECISION Vs NAME_CONTRACT_STATUS
AMT_ANNUITY_y
CNT_PAYMENT
DAYS_DECISION
NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer


NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_TYPE_y
Consumer loans
Cash loans
Revolving loans
XNA
NAME_PAYMENT_TYPE
Cash through the bank
Non-cash from your account
Cashless from the account of the employer
NAME_TYPE_SUITE_y
Unaccompanied
Family
Spouse, partner
Children
Other_A
Other_B
Group of people
NAME_CLIENT_TYPE
New
Repeater
Refreshed
PRODUCT_COMBINATION
POS other with interest
Cash X-Sell: low
POS industry with interest
POS household with interest
POS mobile without interest
Card Street
Card X-Sell
Cash X-Sell: high
Cash
Cash Street: high
Cash X-Sell: middle
POS mobile with interest
POS household without interest
POS industry without interest
Cash Street: low
Cash Street: middle
POS others without interest
NAME_PORTFOLIO
POS
Cash
Cards
Cars
CHANNEL_TYPE
Stone
Credit and cash offices
Country-wide
Regional / Local
AP+ (Cash loan)
Contact center
Channel of corporate sales
Car dealer

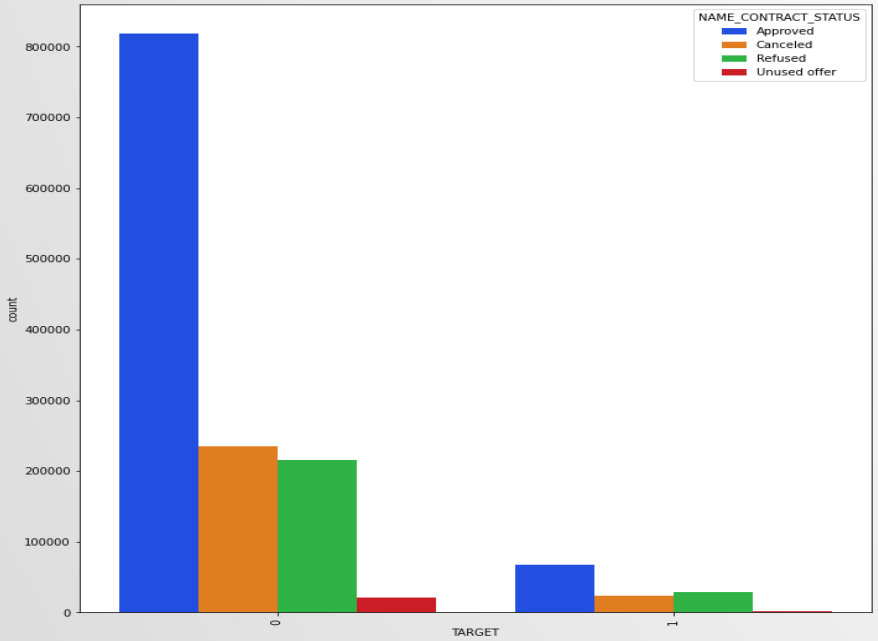

Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS
AMT_ANNUITY_x Vs NAME_CONTRACT_STATUS
AMT_CREDIT_x Vs NAME_CONTRACT_STATUS


Approved
Canceled
Refused
Unused offer
AGE_RANGE
20-30
30-40
40-50
50-60
>60
count
DAYS_EMPLOYED
NAME_CONTRACT_STATUS
0-50000
50000-100000
100000-150000
150000-200000
200000-250000
250000-300000
300000-350000
350000-400000
400000-450000
450000-500000
500000 and above


Approved
Canceled
Refused
Unused offer
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Maternity leave
NAME_INCOME_TYPE
count
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE
Single / not married
Married
Civil marriage
Widow
Separated
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment

INFERENCE FROM THE EDA

- The target column in the current application data has an imbalance percentage of 91% (People with no payment difficulties) and 9% (Defaulters).
- When we perform univariate and bivariate analysis on different variables w.r.t Target 0 and Target 1, we infer that the trend for both these variables remain the same.
- With the variables obtained from the first data we perform analysis on Prev_Application_Data.csv and extract variables which have a great impact on loan approvals.
- We find that the trends for loan approval and trends on target variable mostly remains the same.
- We extract the columns that have a high influence on the Target variable. These variables have a high ratio of people with no payment difficulties.
- On performing analysis on merged data , we find that loan approval rate for such columns are fairly high when compared to other variables.



On plotting, Target v/s Name_Contract_Status , we can infer the number of defaulters were high for variables on which the loan was approved.

For target = 0 - Loan was approved.

Hence, we need to find all the categories where loan was approved.

The categories are listed in the next slide.

INFERENCE FROM THE EDA

The loan approval rate was found to be very high for these cols and hence we must focus on these variables while providing the loan.

- INCOME_TYPE : WORKING
- NAME_CONTRACT_TYPE : CASH LOANS & CONSUMER LOANS
- NAME_SUITE_TYPE: UNACCOMPANIED & FAMILY
- EDU : SECONDARY
- FAMILY_STATUS: MARRIED (INCLUDING CIVIL MARRAIAGE)
- DAYS EMPLOYED : < 5 YEARS
- OCC_TYPE: LABORERS AND SALES_STAFF
- ORG_TYPE : BUSINESS_ENTITY_3 & SELF-EMPLOYED
- REGION_RATING
- AMT_INCOME_RANGE = 100000 - 150000
- AMT_CREDIT_RANGE = 500000 AND ABOVE
- AGE = 30-40
- THEY OWN A HOUSE BUT DO NOT A OWN A CAR

WE CAN CONCLUDE THAT WHILE LENDING LOANS WE MUST ENSURE THAT THESE PEOPLE WHO FALL UNDER THE ABOVE MENTIONED CATEGORIES HAVE A GREATER DEFAULTER VALUE. HENCE THE BANK MUST TAKE PRECAUTIONARY MEASURE WHILE APPROVING THE LOANS FOR SUCH VARIABLES.

OTHER INTERESTING INSIGHTS:

- Loan approval was confirmed within 2 days in case of approved ot refused loan.
- The average term of loan was 14 months.
- Most of the labor class were males.
- Repeaters had a greater probability of loan approvals.
- Married people have high mean for credit amount.

THANK YOU!