EDA Case Study

Providing EDA analysis on the current & previous application data for the bank on potential parameters of payment difficulties, loan approval, cancellation, rejection etc.

Problem Statement

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history.
- Use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
  - 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.

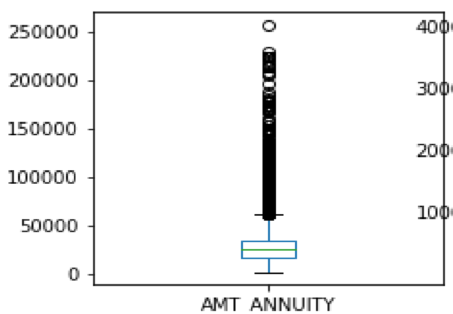
- When a client applies for a loan, there are four types of decisions that could be taken by the client/company):
  - Approved
  - Cancelled
  - Refused
  - Unused offer

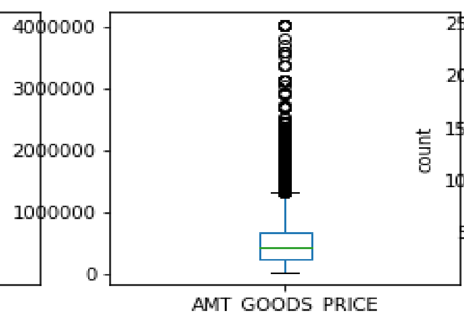
Available Data For Analysis

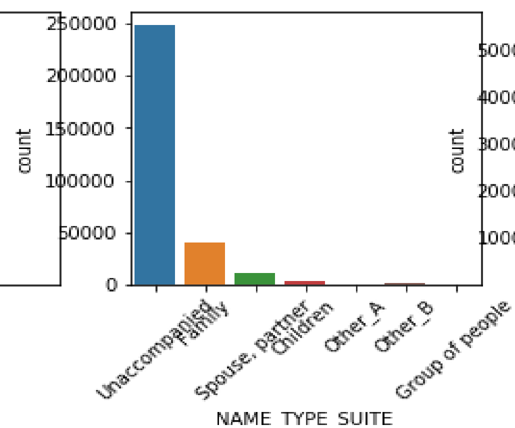
- *'application_data.csv'*
  - contains all the information of the client at the time of application.
  - The data is about whether a **client has payment difficulties.**
- *'previous_application.csv'*
  - contains information about the client's previous loan data.
  - It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**
- Brief Actions Items to be taken
  - Remove unnecessary data
  - Propose the cleaning method of available data
  - Perform univariate & bivariate analysis on the data
  - Fetch the inferences against TARGET column and NAME_CONTRACT_STATUS Columns.

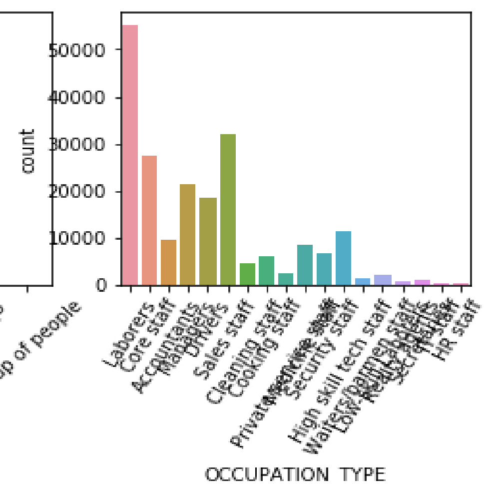
Data Sourcing & Cleansing

- Initial Application data is found with 307K entries across 122 columns
- As some of the columns has missing values from 47% spreading to 70% and more. Dropped all columns with missing values more than 47%.
- About 21 columns with various document availability clubbed to one single column ‘No. Of documents’.
- Identified the data spread in each column and proposed to change the data type of below columns from float to integer as data suggests:
  - DAYS_REGISTRATION
  - CNT_FAM_MEMBERS
  - DAYS_LAST_PHONE_CHANGE
- Later identified some more columns with less missing values and proposed the below approaches for imputation:
  - AMT_ANNUITY – Impute with median due to outliers
  - AMT_GOODS_PRICE – Impute with median due to outliers
  - NAME_TYPE_SUITE – Impute with mode value ‘Unaccompanied’ as it clearly holds majority
  - OCCUPATION_TYPE – Impute with mode value ‘Laborers’ as it holds more values



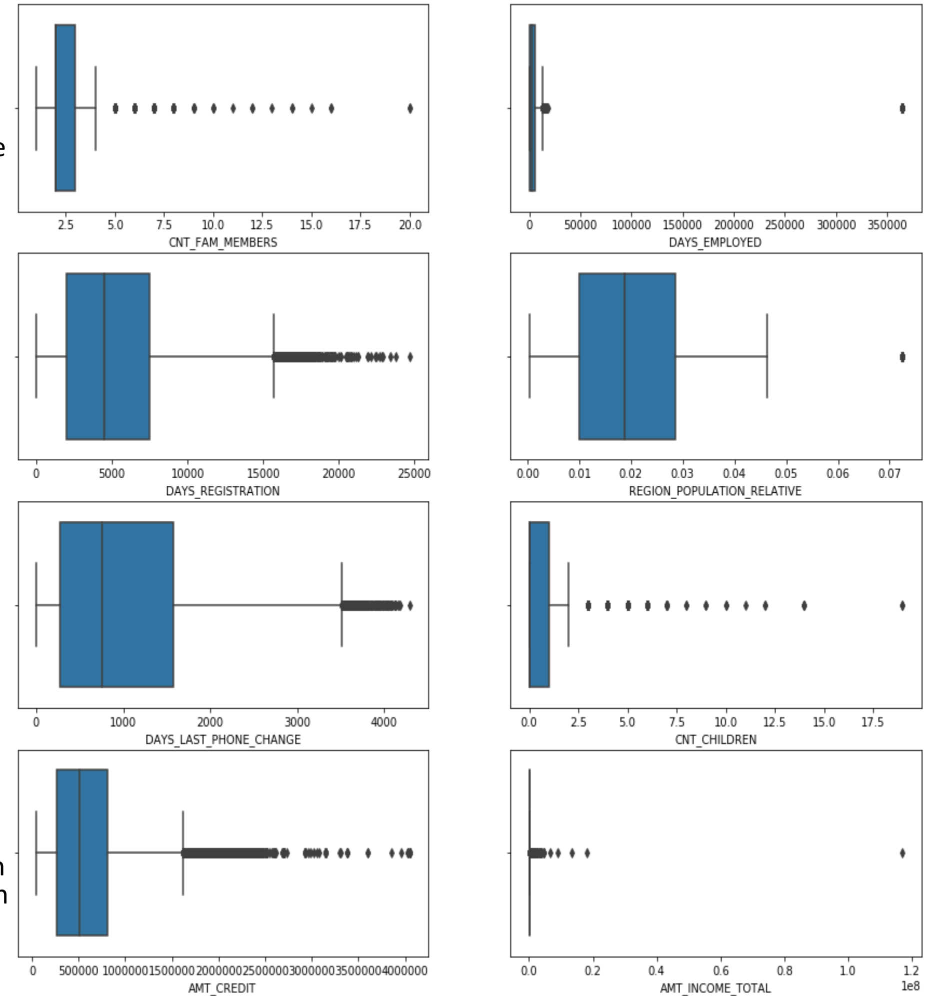






Data Sourcing & Cleansing

- About six various columns around AMT_REQ_CREDIT_BUREAU exist with about 13% missing values.
  - First five columns can be imputed with 0 as majority of the values are distributed around 0.
  - The last column for YEAR needs to be imputed with 1 as it has some data spread away from 0 and the median value is 1.
- There are about 5 different days columns respectively DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE. As all these columns are in negative, converted them to positive number for clear analysis.
- Out of all continuous columns, the following are picked to handle outliers as these columns has too distinct outliers:
  1. CNT_FAM_MEMBERS – Drop off rows with more than 8 family members as they are very less in numbers.
  2. CNT_CHILDREN – Drop off rows with more than 7 children as they are very less in numbers.
  3. REGION_POPULATION_RELATIVE – Drop off all the rows with value more than 0.07 as they are very less in numbers.
  4. AMT_INCOME_TOTAL – Drop off all the rows with greater than 900K value which is the 99.9th percentile.
  5. DAYS_EMPLOYED – There is a garbage value 365243 which needs to be cleared out and then later impute with median value as it still has some outliers.



Data Conversion & Binning

- Convert the DAYS_BIRTH to AGE so that data is more understandable and analysis will be pretty easier.
- Below columns are identified to perform binning as these continuous variables and categorizing these may help in further analysis.
  1) AGE – Youth (25-35), Middle Age(35-50) & Veterans (>50)
  2) AMT_INCOME_TOTAL – Low, Below Avg, Above Avg, High & Very High
  3) AMT_GOODS_PRICE – Low, Average, High & Very High
  4) DAYS_EMPLOYED – Junior, Senior, Middle, Highly Experienced
  5) EXT_SOURCE_2 – Low, Good, Better & Best
  6) REGION_POPULATION_RELATIVE – Low, Average, High & Very High

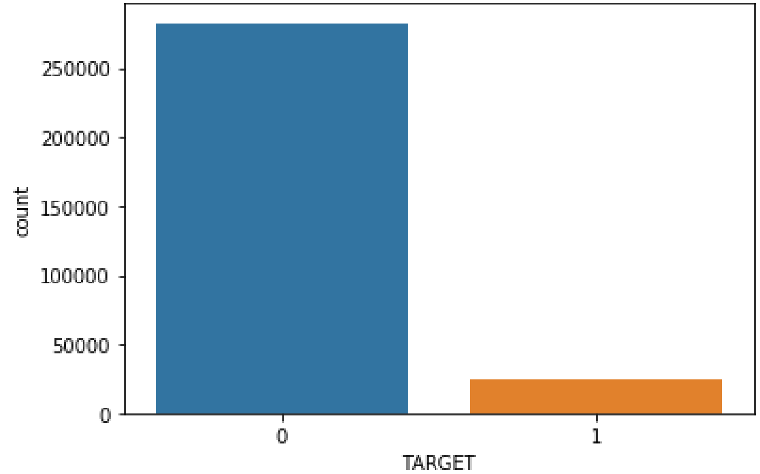
```
a_df.EMPLOYMENT_EXP.value_counts()

Junior        82196
Senior        78150
Middle        62454
Highly_Exp    29335
Name: EMPLOYMENT_EXP, dtype: int64
```

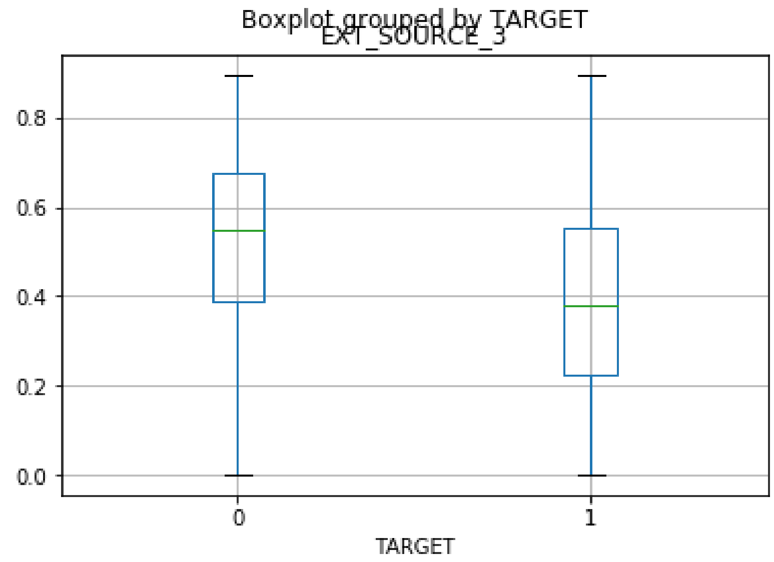
```
a_df.AMT_GOODS_SECTION.value_counts()

Average      97727
High         90496
Low          84891
Very_High    34119
Name: AMT_GOODS_SECTION, dtype: int64
```

```
a_df.EXT_SOURCE_SCORE.value_counts()

Good      98869
Better    82426
Low       78943
Best      46613
Name: EXT_SOURCE_SCORE, dtype: int64
```

Analysis of Application Data

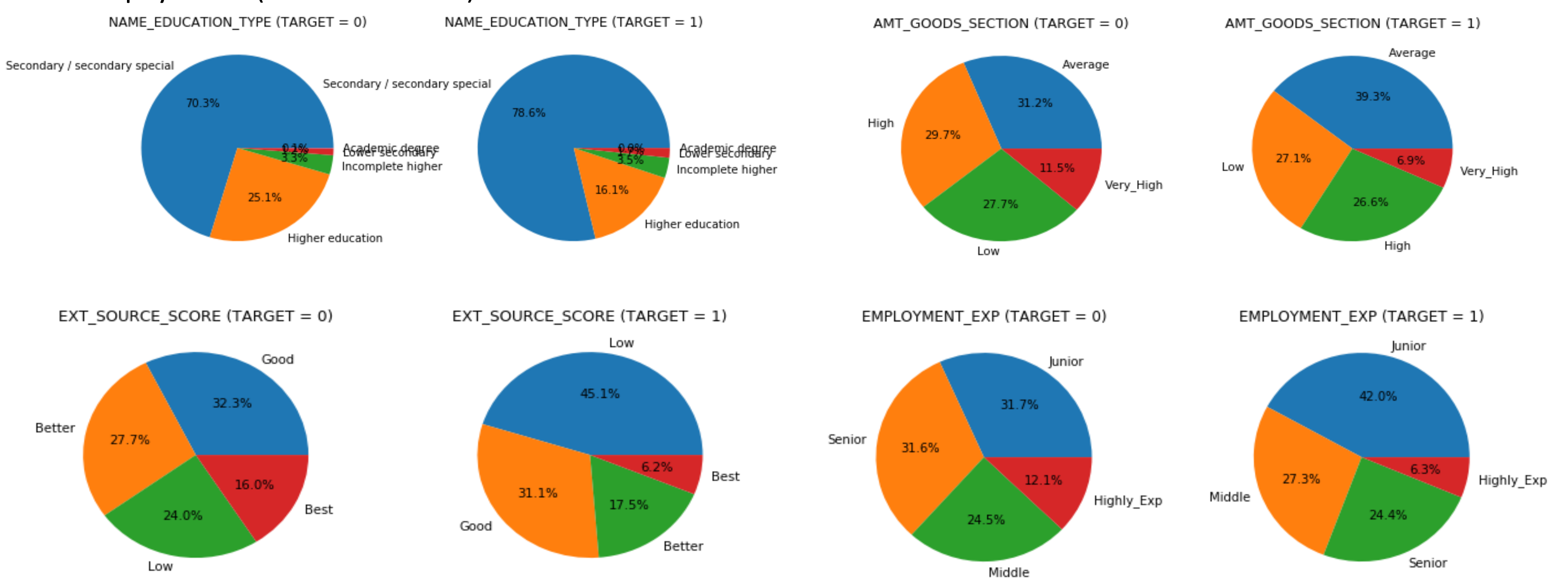
- Data Imbalance: It is observed that almost 92% of the applications are with good history of payments. Whereas 8% are with payment difficulties. The data imbalance is around 11%.
- Identified about 29 columns to further perform analysis by dropping all other less significant columns. Performed a correlation to make sure highly correlated columns are not lost.
- Divided dataframe into two dataframes:
  - One with data corresponds to TARGET value 0. (282K Rows)
  - Another with data corresponds to TARGET value 1. (25K rows)
- Correlation:
  - The correlation in both the dataframes found similar with top three combination being the same.
- Identified the continuous and categorical variables separately to perform analysis.
- The univariate analysis on the continuous variables fetched the below observations:
  - EXT_SOURCE_2 & EXT_SOURCE_3 - The scores are **directly** proportional to application being not getting into payment difficulties (TARGET value 0)
  - AMT_REQ_CREDIT_BUREAU_YEAR - The score is **indirectly** proportional to application being not getting into payment difficulties (TARGET value 0). Higher the score, the application more likely to get into payment difficulties (TARGET value 1).





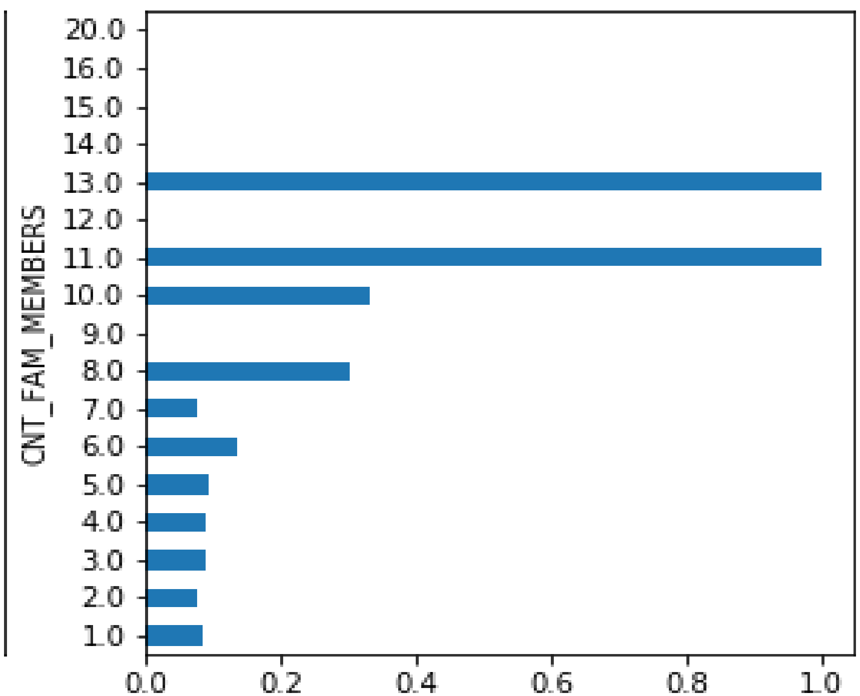
Univariate Analysis of Application Data

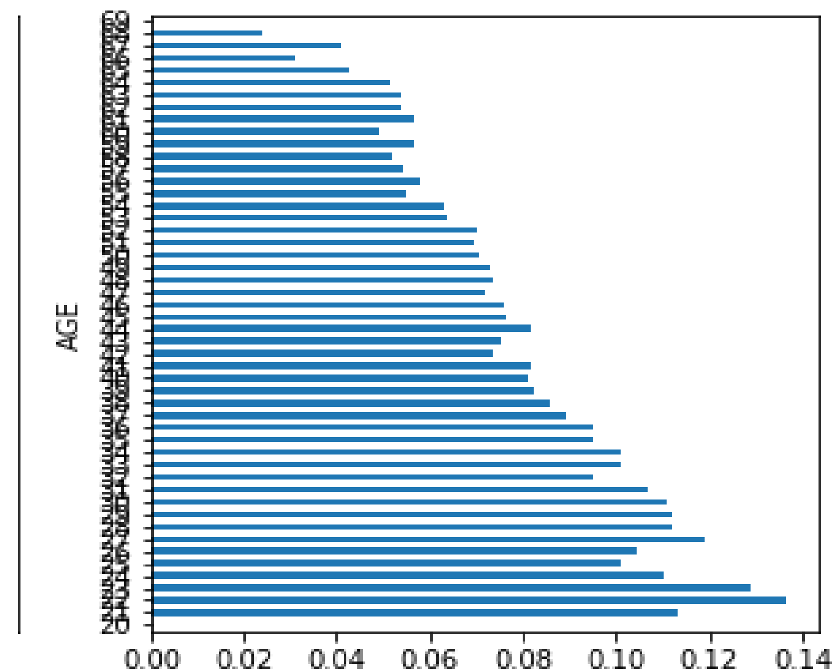
- The univariate analysis on categorical variables fetched the below observations:
  - ➢ NAME_EDUCATION_TYPE - Candidates with Higher Education observed more reliable for payments (TARGET value 0).
  - ➢ OCCUPATION_TYPE - Laborers seems to have more chance of getting into payment difficulties (TARGET value 1).
  - ➢ AMT_GOODS_SECTION - The applications with average goods price (2,50,000 - 5,00,000) seems to tend more towards payment difficulties (TARGET value 1). The application with Very High goods price seems to tend more towards good payments (TARGET value 0).
  - ➢ EMPLOYMENT_EXP - Junior employees (applications with less DAYS_EMPLOYED) tend towards payment difficulties, whereas Highly Experienced employees tend to have no issues with payments.
  - ➢ EXT_SOURCE_SCORE - It clearly shows that as good the score, the applications more likely to have good payment (TARGET value 0).



Bivariate Analysis of Application Data

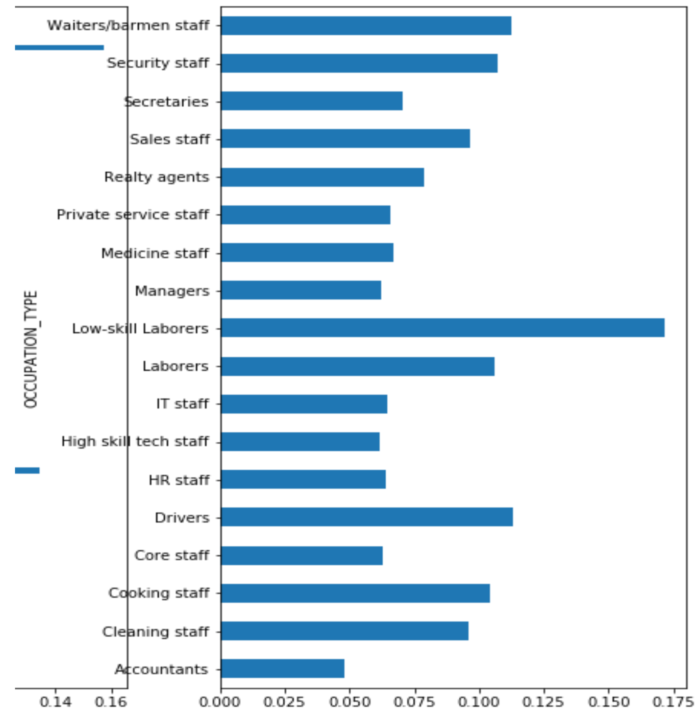
- Bivariate analysis on continuous columns fetched the below observations:
  - Age is directly proportional to the applicant being loyal to payments. Younger people have more chances of delayed payments and older people have very less chance.
  - Count of family members is indirectly proportional to being loyal to payments. As lesser the count of family members, as good to expect safe and on-time payments.

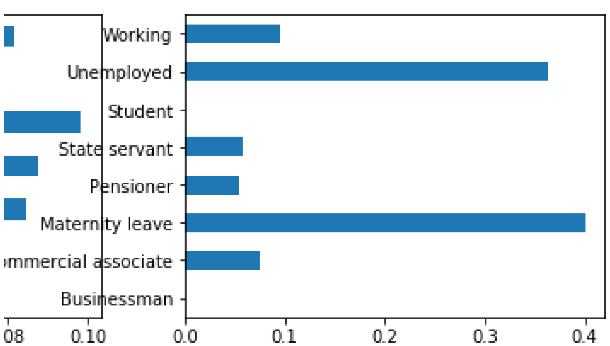


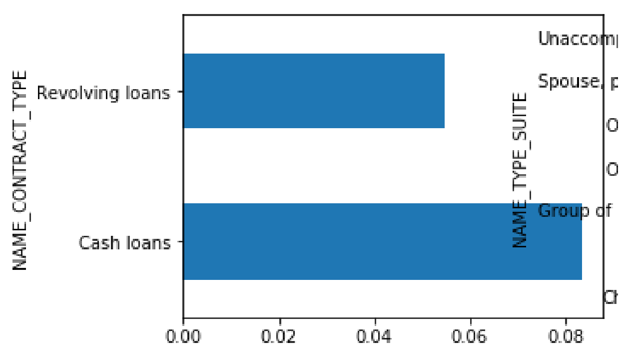


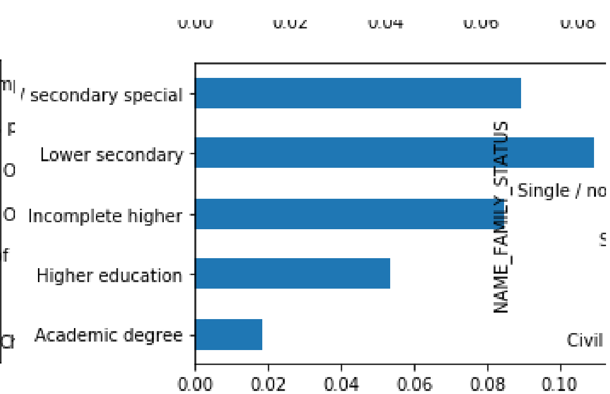
Bivariate Analysis of Application Data

- Bivariate analysis on categorical columns fetched the below observations:
  - ❑ Very clearly Unemployed & Maternity Leave applicants are having very higher chance of falling into payment difficulties.
  - ❑ Cash loans are more prone to payment difficulties compared to Revolving loans.
  - ❑ As higher the applicant is educated there is less chance of falling into payment difficulties.
  - ❑ External Source Score clearly a reliable value. As better the score, there is more not likely to fall into payment difficulties.
  - ❑ Applications from OCCUPATION_TYPE of 'Low-skill Laborers' seems falling more into payment difficulties.









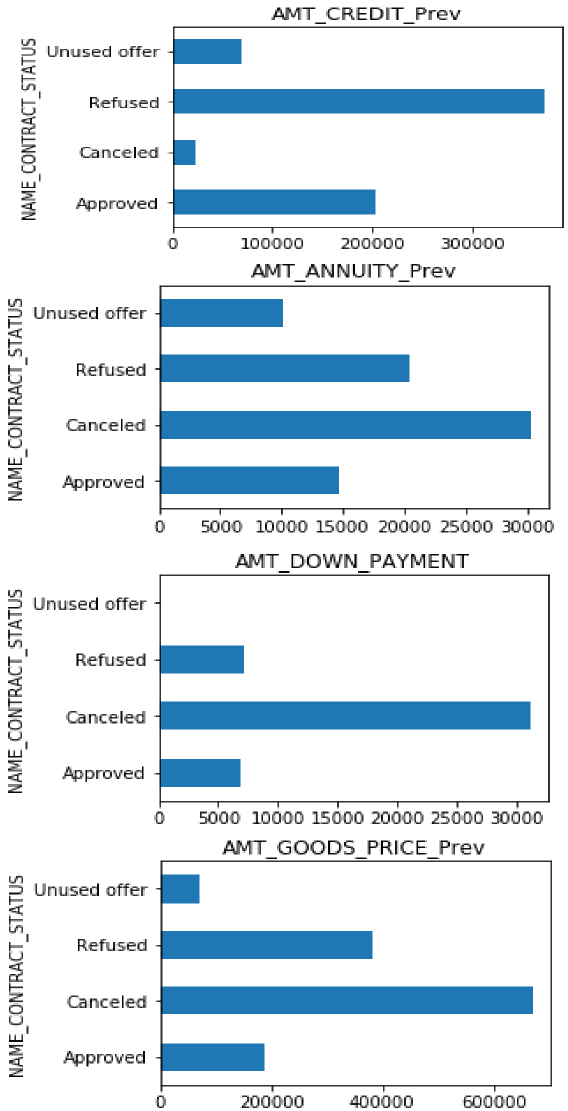
Analysis merging Application & Previous Application Data

Observations out of univariate analysis

- Almost all continuous variables has some outliers
- NFLAG_INSURED_ON_APPROVAL - Comparatively less no. of applications has insurance. Need to determine whether this plays a role in final status
- All the below columns has NA or Garbage values huge in numbers:
  1. NAME_PAYMENT_TYPE
  2. NAME_YIELD_GROUP
  3. NAME_SELLER_INDUSTRY
  4. NAME_GOODS_CATEGORY
  5. CODE_REJECT_REASON

- **Merged the application data which is available after some analysis with previous application data. Dropped some insignificant columns in previous application data for crisp analysis.**

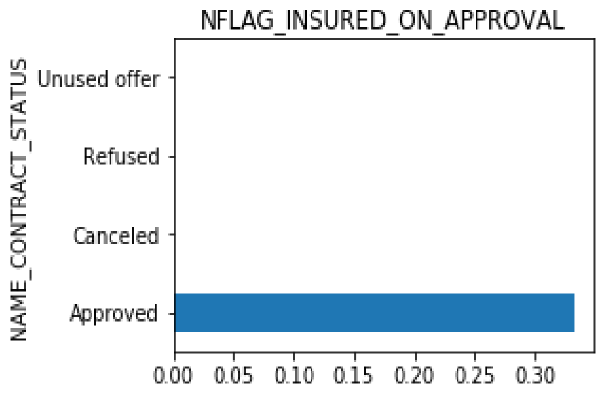
- Below are the observations out of bivariate analysis of continuous variables
  - ❖ AMT_ANNUITY - Higher Annuity amounts leads to cancellation or refusal of applications.
  - ❖ AMT_APPLICATION - Higher amount applications leads to refusal of applications.
  - ❖ AMT_CREDIT - Higher amount credits leads to refusal of applications.
  - ❖ AMT_GOODS_PRICE - Higher goods price leads to cancellation or refusal of applications.

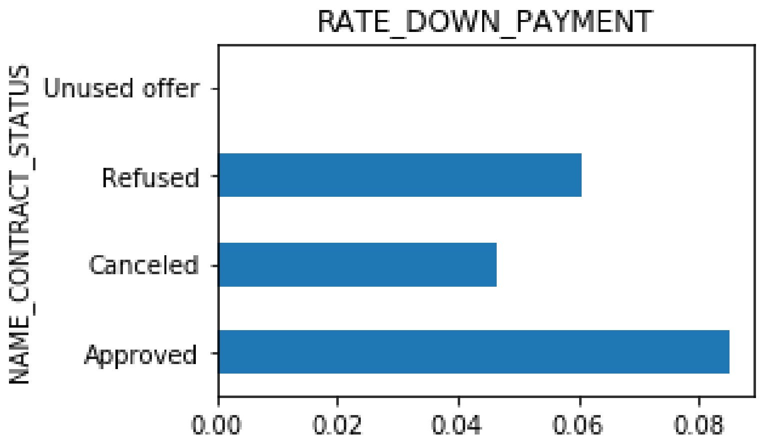


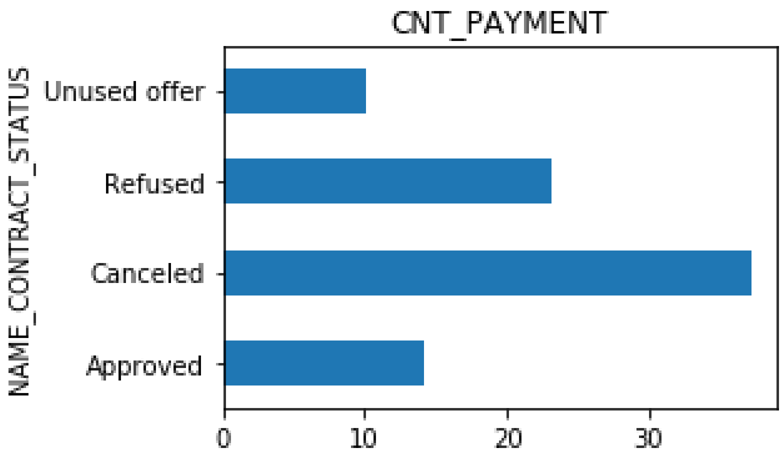
Analysis merging Application & Previous Application Data

Below are the observations out of some more bivariate analysis of continuous variables

- AMT_DOWN_PAYMENT - It is observed that applications with huge down payments are more to get cancelled.
- RATE_DOWN_PAYMENT - Higher down payment rate leads to application approval.
- DAYS_DECISION - As high the number it may lead to approval and as low the number it may lead to cancellation.
- CNT_PAYMENT - As minimum the term of previous application, it is high likely to get approval. As high it is, it may lead to cancellation.
- NFLAG_INSURED_ON_APPROVAL - Applications without insurance are high likely not to get approved.



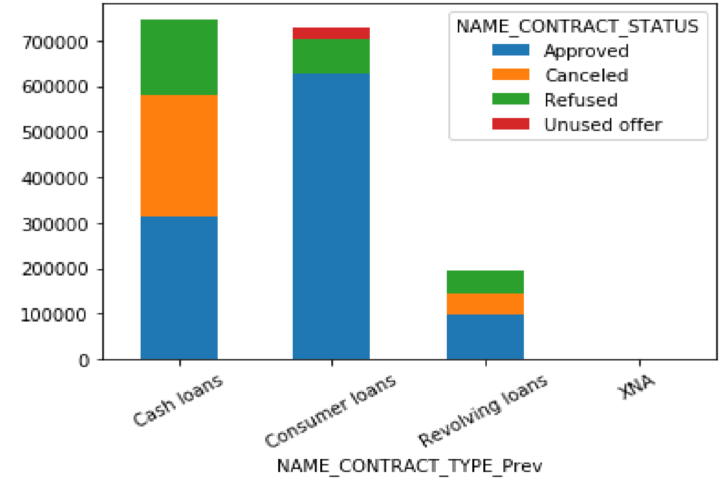


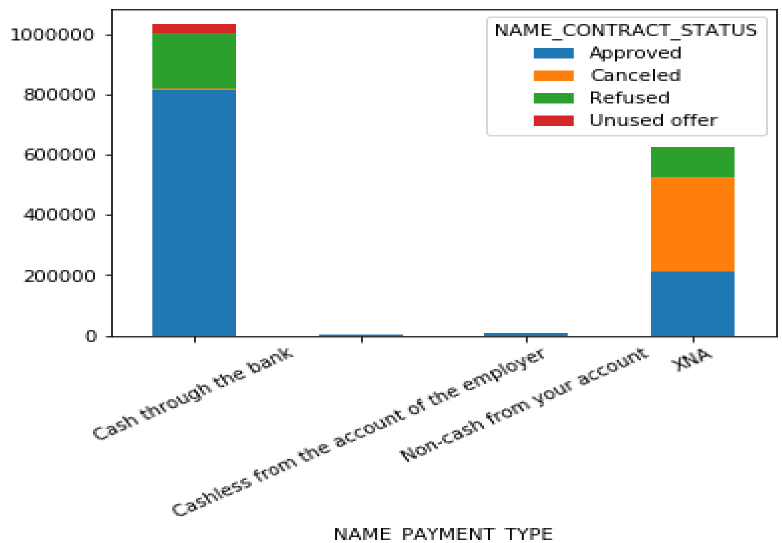


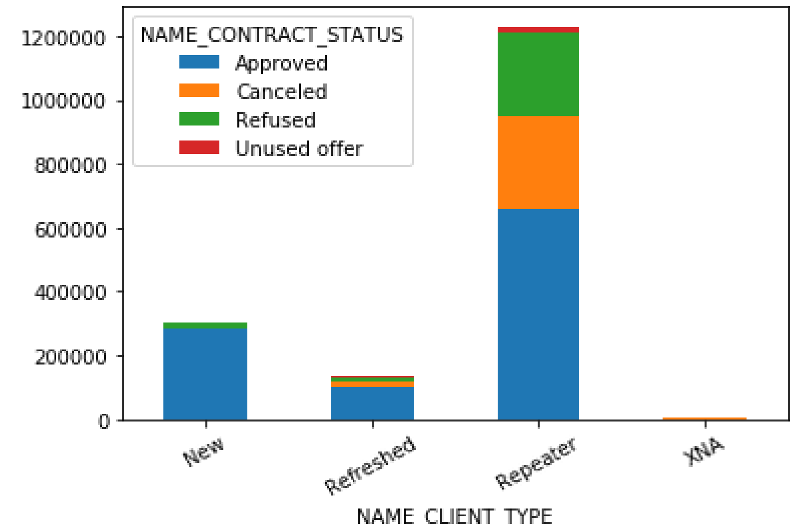
Analysis merging Application & Previous Application Data

Below are the observations of bivariate analysis of categorical variables

- ❖ NAME_CONTRACT_TYPE - Cash loans & revolving loans are more set to get cancelled or refused compared to the consumer loans which has higher chance of getting approved.
- ❖ NAME_PAYMENT_TYPE - High chance of getting cancelled where there is no record of payment type.
- ❖ NAME_TYPE_SUITE - Good chance of getting application cancelled when the applicant is unaccompanied.
- ❖ NAME_CLIENT_TYPE - Repeaters are more likely to get their applications cancelled or refused compared to New & Refreshed.



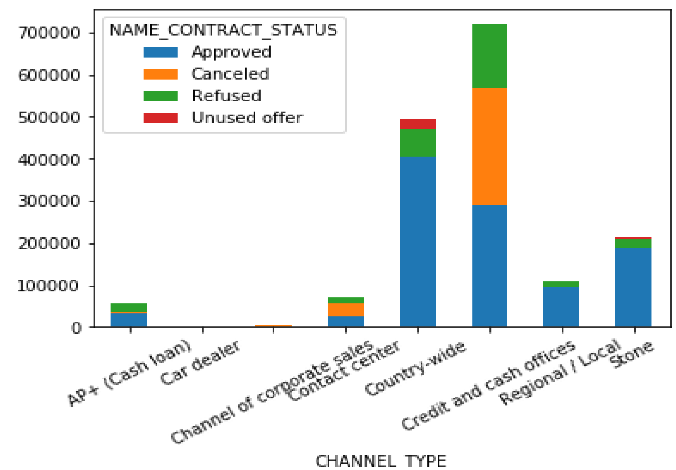


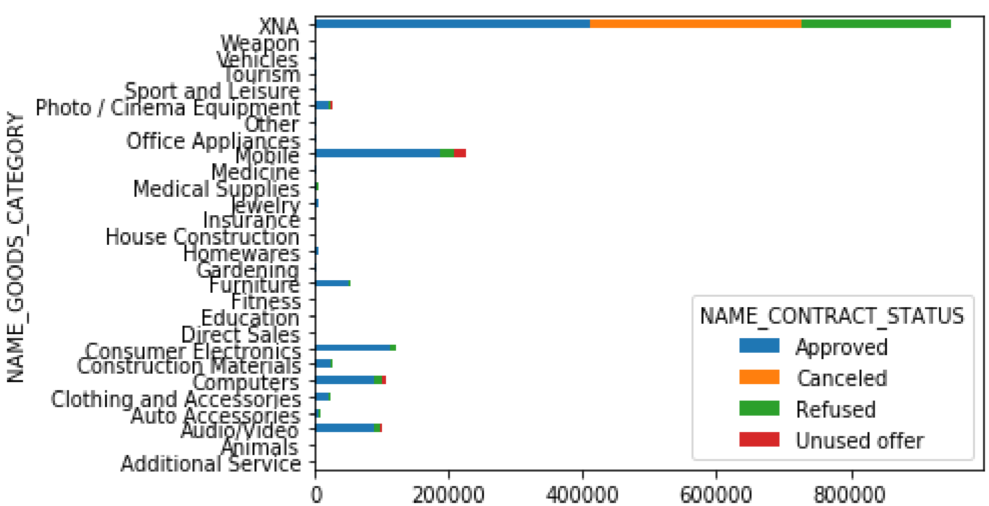


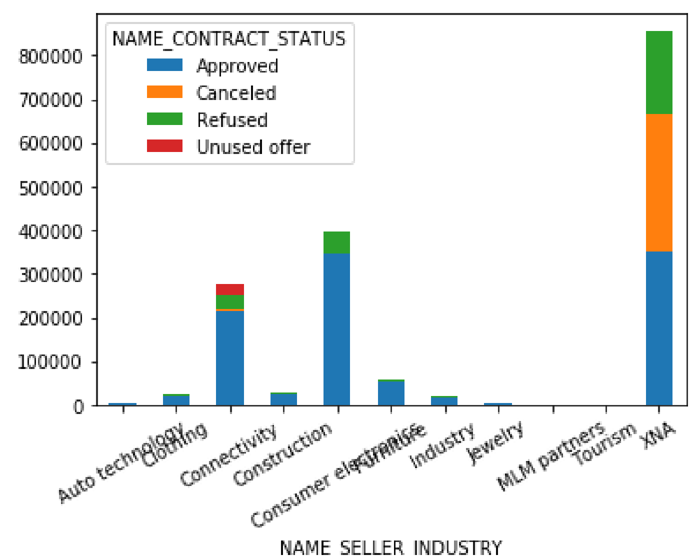
Analysis merging Application & Previous Application Data

Below are the observations out of bivariate analysis of remaining categorical variables

- NAME_GOODS_CATEGORY - More likely to get cancelled / refused in case the goods type is not mentioned.
- CHANNEL_TYPE - Country-wide channel has more cancellation and refusal applications compared to any other channel.
- NAME_SELLER_INDUSTRY - More likely to get cancelled / refused in case of seller industry name is not available.
- NAME_YIELD_GROUP - High likely to get cancelled if the grouped rate interest amount is not available.







THANK YOU!