EDA CASE STUDY

Gaurav Garg
Subrahmanyam Vedula

Problem Statement

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.

The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default

Initial Steps taken before Analysis:

- Imported the “Application_data data set
- Checked the structure of the data
- Checked the percentage of missing values.
- Deleted all the columns which are having missing %age more than 45%
- Imputed other columns with missing values less than 45% with either median or mode
- Converted “Days_Birth” Column into age
- Checked the outliers of the numerical columns (data driven columns)
- Checked the imbalance %age of the Target column
- Divided the Data set into two data set (1 & 0) for EDA Analysis

Univariate Analysis on Application Data:

- Categorical Univariate Analysis:

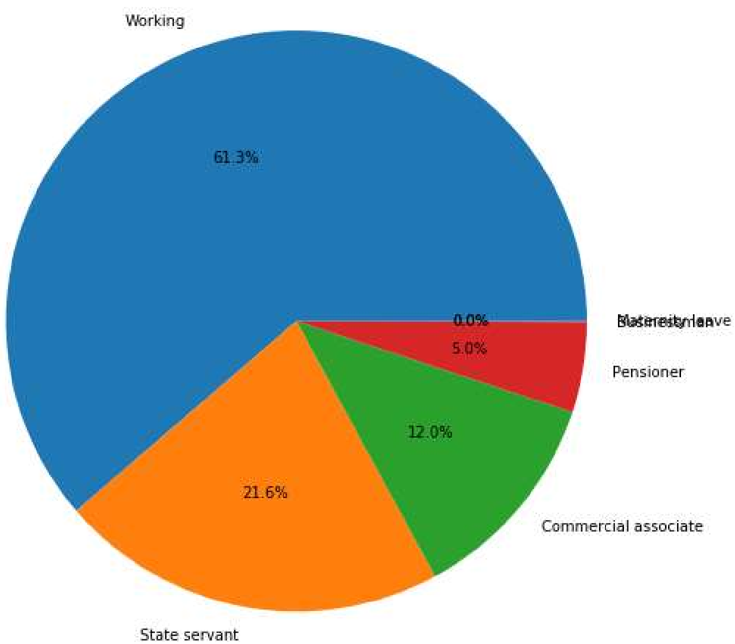
  Columns under consideration are:
  - NAME_INCOME_TYPE
  - OCCUPATION _TYPE
  - ORGANIZATION_TYPE
  - CODE_GENDER
  - NAME_EDUCATION_TYPE

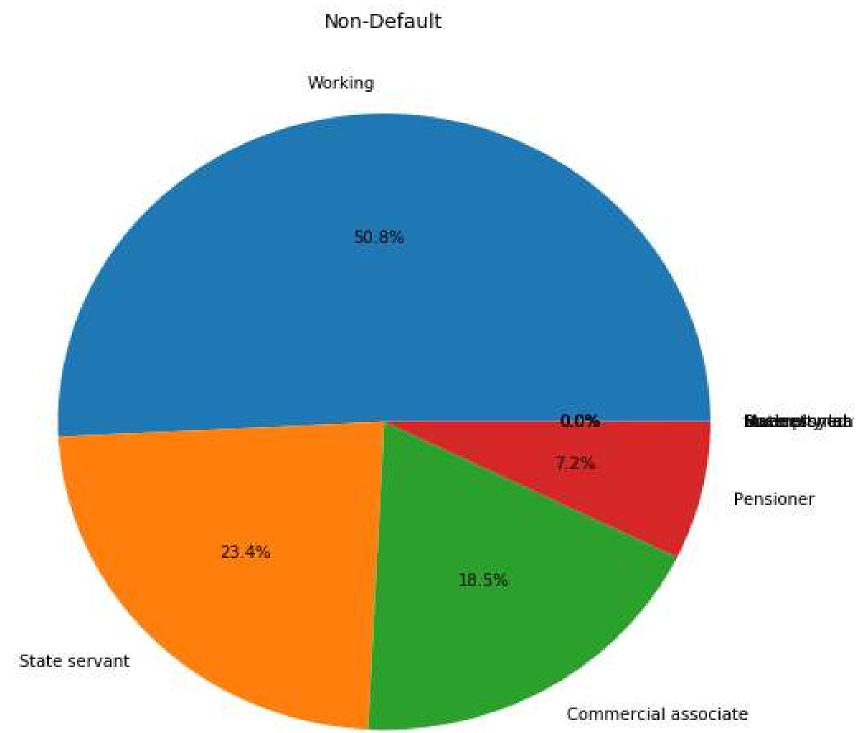
- Continuous Univariate Analysis:

  Columns under consideration are:
  - AMT_INCOME_TOTAL
  - AGE
  - AMT_CREDIT
  - DAYS_EMPLOYED
  - AMT_ANNUITY

Insights from Univariate Analysis:

Column : NAME_INCOME_TYPE

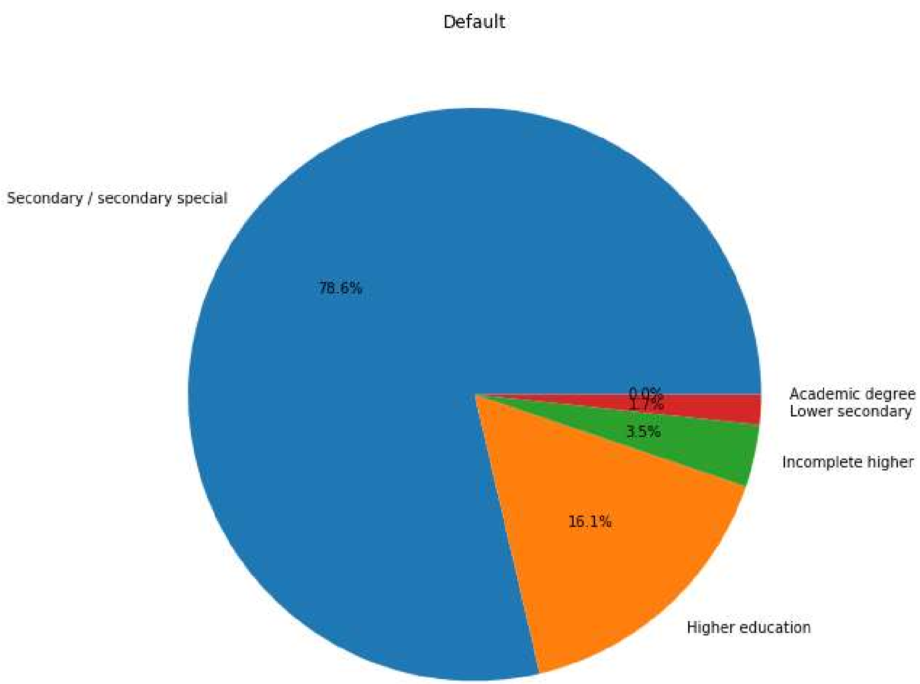


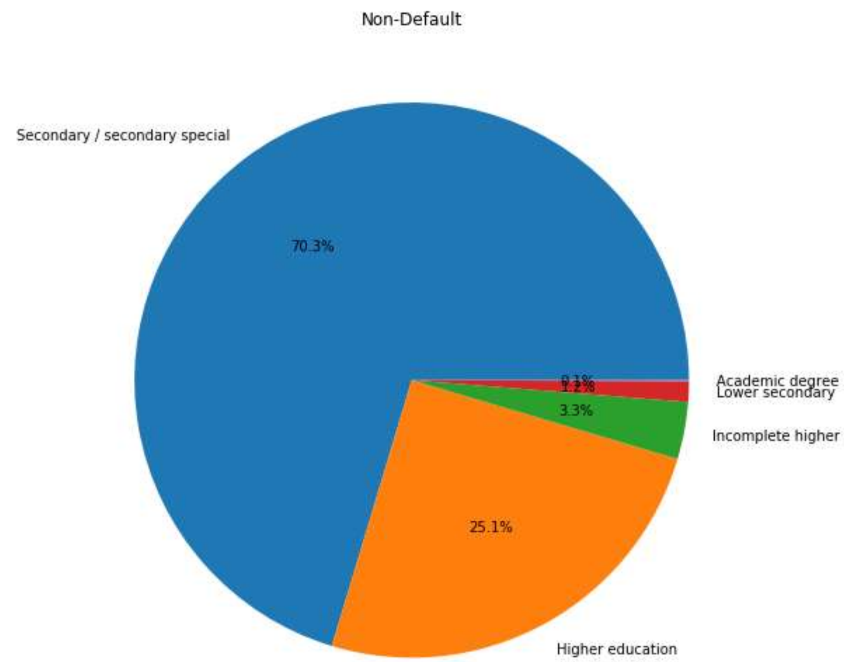


Commercial associates are less likely to default & working class are more likely to Default

Insights from Univariate Analysis:

Column : NAME_EDUCATION_TYPE



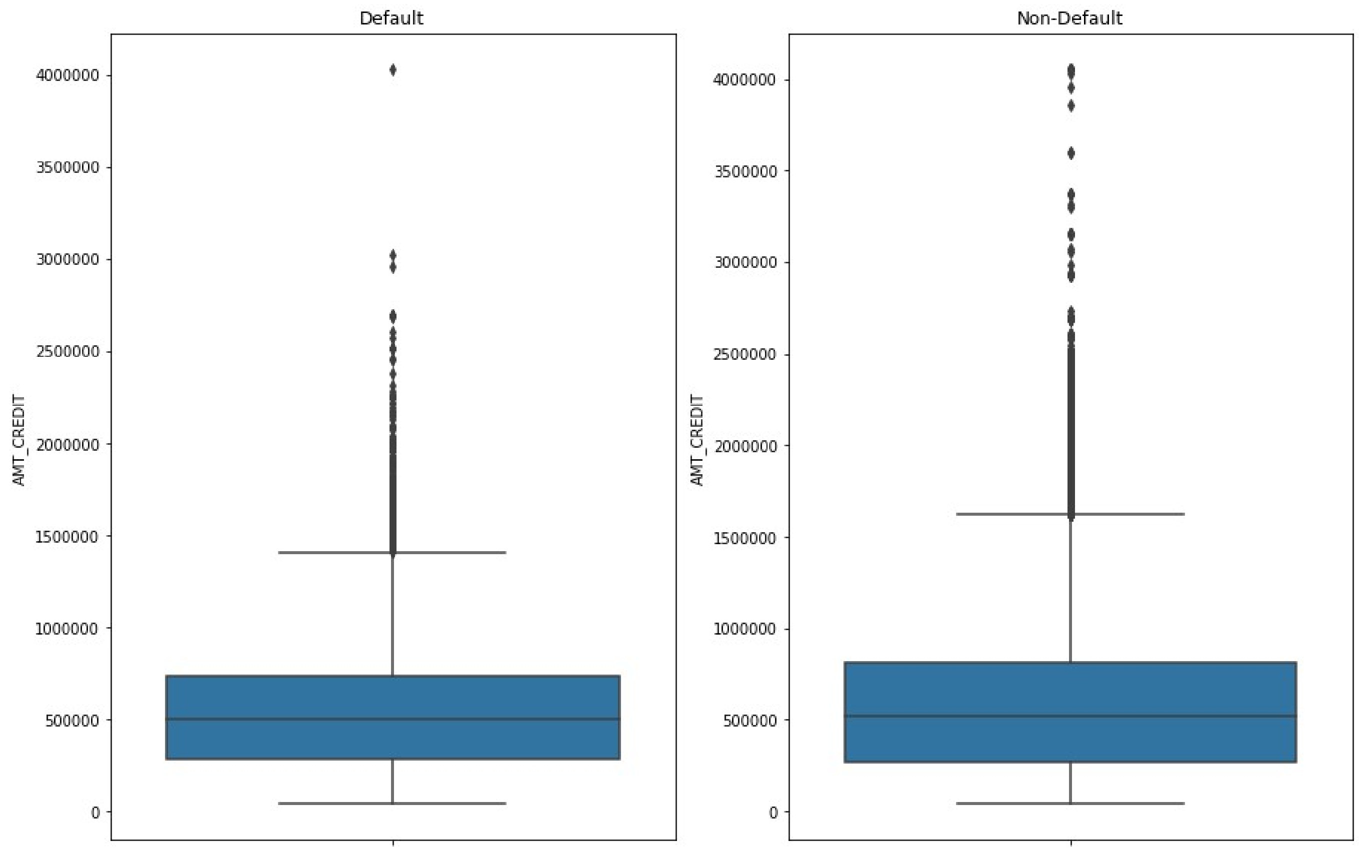


Applicants with Higher Education are 1.5 times more likely to be Non-defaulters

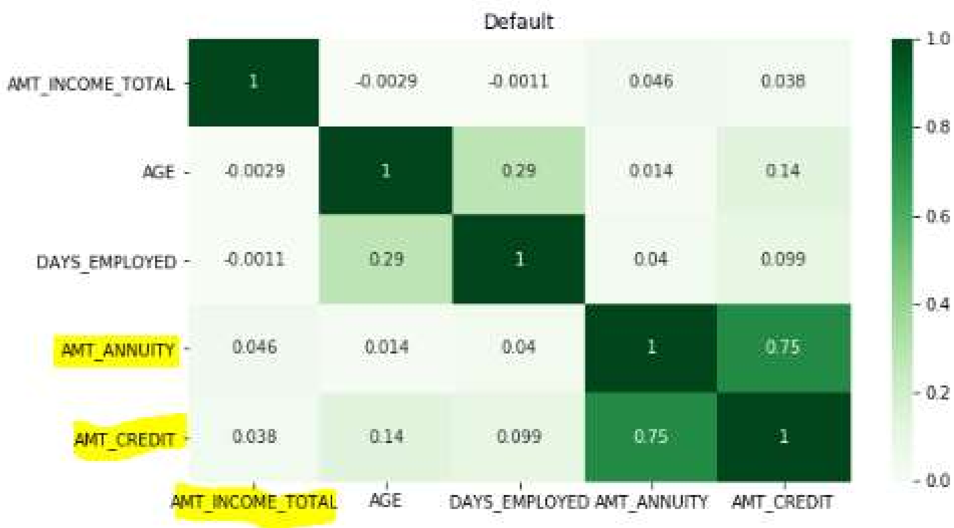
Insights from Univariate Analysis:

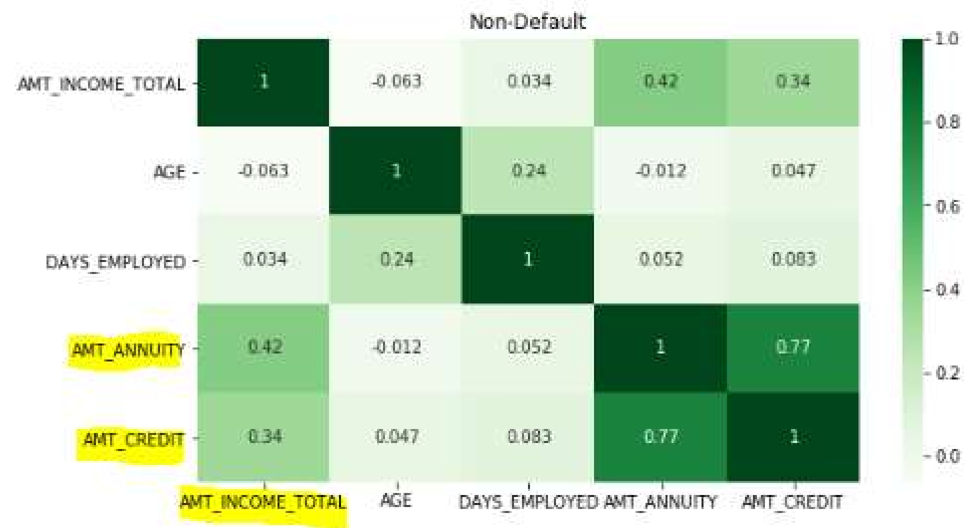
Column : AMT_CREDIT

Credit value above 15 Lakhs are less likely to Default



Correlation Matrix between numerical columns:





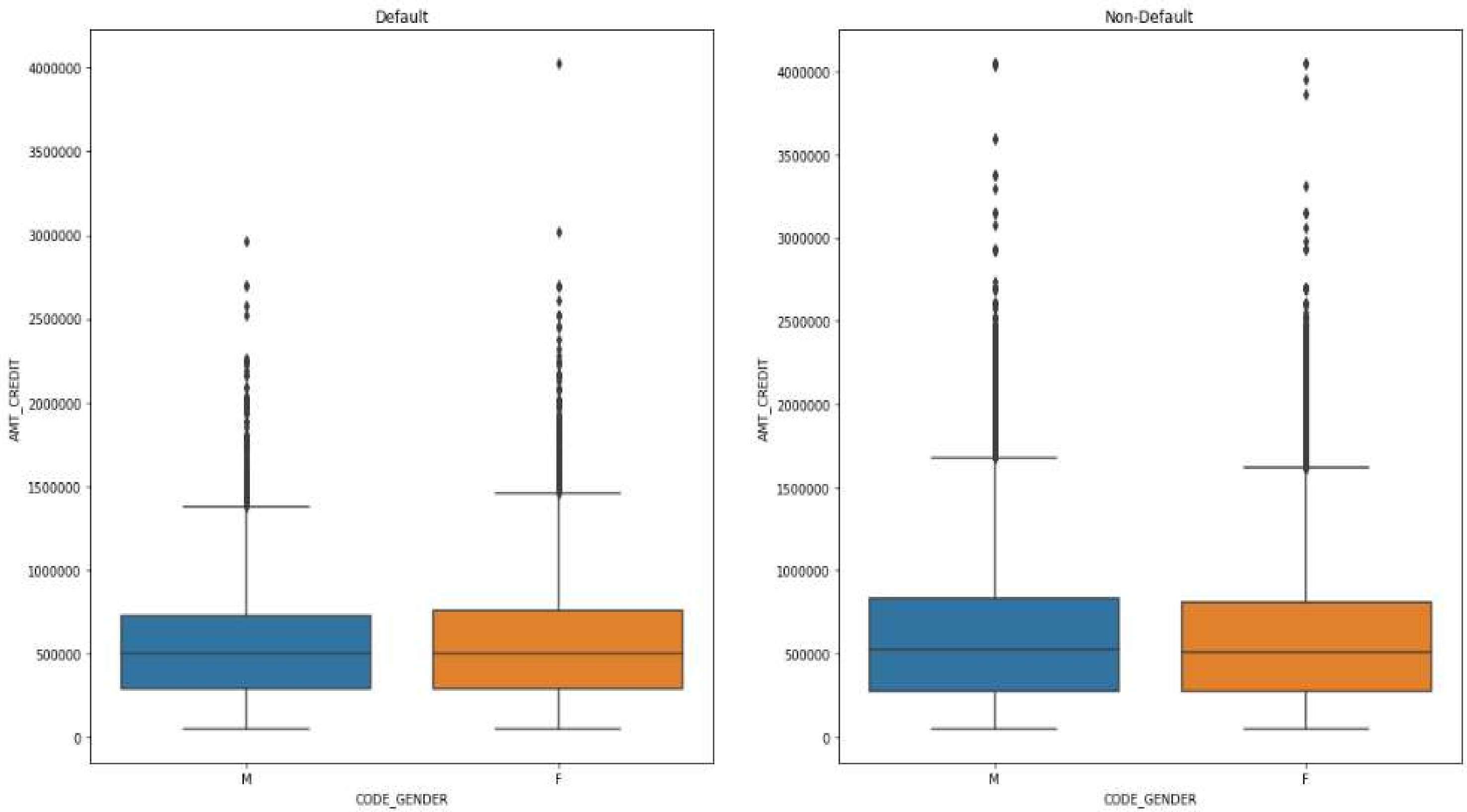
Inferences :

- AMT_ANNUITY and AMT_CREDIT are highly co-related columns for bath Target values (1&0).
- Non-default has higher co-relation between AMT_INCOME_TOTAL VS AMT_ANNUITY & AMT_INCOME_TOTAL VS AMT_CREDIT vis-a-vis Defaulters.

Bivariate Analysis on Application Data:

- In Categorical-Categorical, columns considered are:
  - NAME_CONTRACT_TYPE with CODE_GENDER
  - FLAG_OWN_REALTY with CODE_GENDER
  - NAME_EDUCATION_TYPE with FLAG_OWN_REALTY
  - NAME_HOUSING_TYPE with FLAG_OWN_REALTY
  - NAME_CONTRACT_TYPE with NAME_FAMILY_STATUS
- In Categorical-Continuous, columns considered are:
  - CODE_GENDER with AMT_CREDIT
  - AMT_INCOME_TOTAL with CODE_GENDER
  - FLAG_OWN_REALTY with AMT_INCOME_TOTAL
  - FLAG_OWN_REALTY with AGE
  - NAME_CONTRACT_TYPE with AGE
- In Continuous-Continuous, columns considered are:
  - AMT_INCOME_TOTAL with AGE
  - AMT_INCOME_TOTAL with AMT_ANNUITY
  - DAYS_EMPLOYED with AMT_CREDIT
  - AGE with DAYS_EMPLOYED
  - AMT_GOODS_PRICE with AGE

Insights from Bivariate Analysis:



Columns:
AMT_CREDIT vs
CODE_GENDER

Outcome:

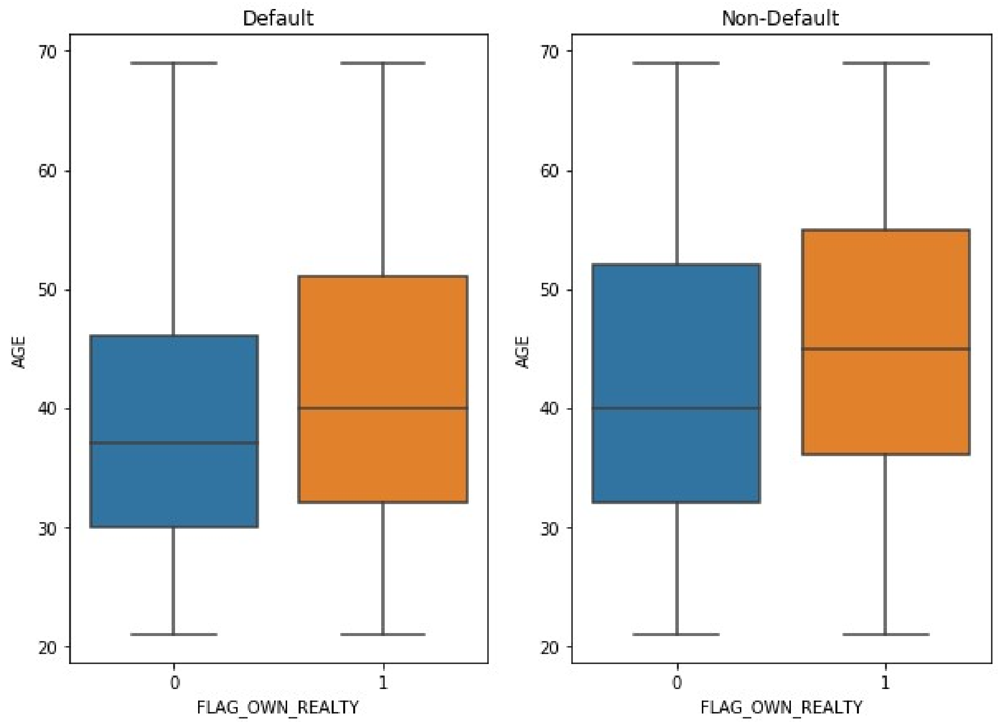
In case of Defaulters, 75% percentile is greater for Females whereas in case of non-defaulters it is greater for Males.

Insights from Bivariate Analysis:

Columns:
AGE vs FLAG_OWN_REALTY

Outcome:

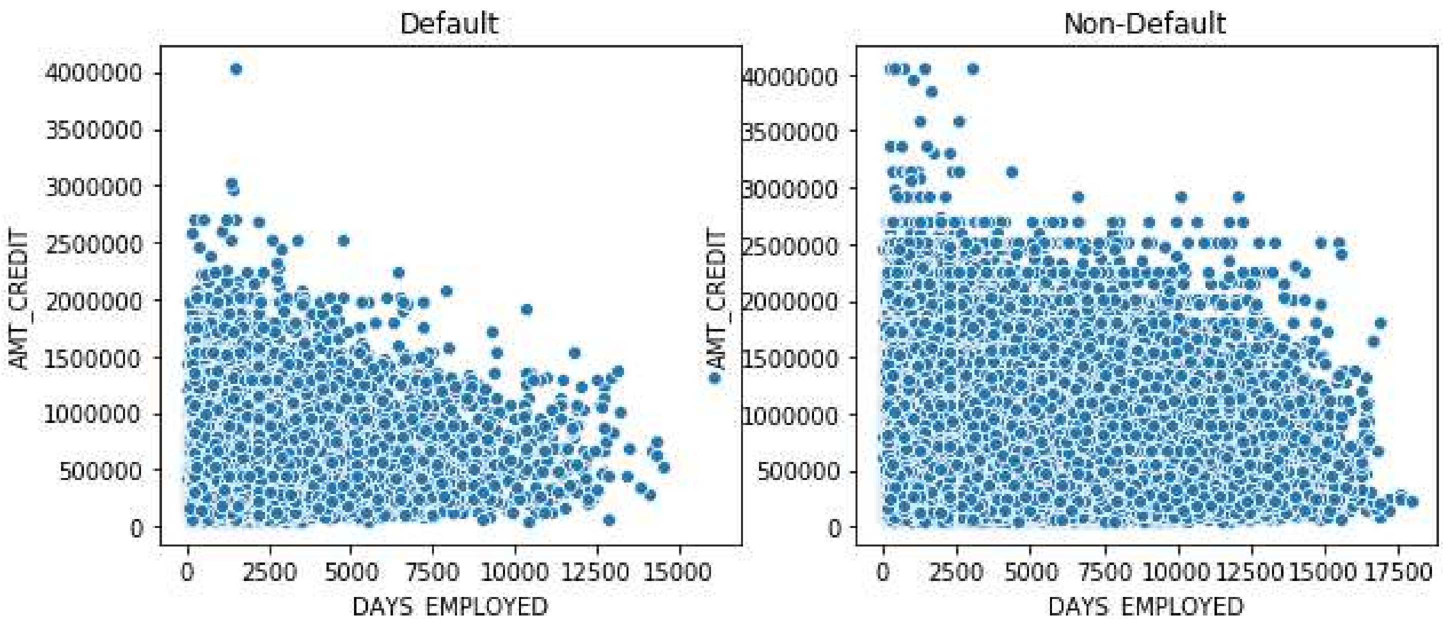
Applicants who are above 45 years of age and do not own a property are less likely to be Defaulters.



Insights from Bivariate Analysis:

Columns: AMT_CREDIT vs DAYS_EMPLOYED

Outcome: **Applicants with more Work Experience and less credit amount are less likely to Default**



Merging with previous application data

- Merged application_data.csv with previous_application.csv using LEFT JOIN.
- Dropped columns with missing values more than 45%.
- Performed Univariate and Bivariate analysis on combined dataset.

```
In [67]: df2 = pd.read_csv("previous_application.csv")
```

Merging of both Application data and Previous application data

```
In [68]: df12 = pd.merge(left=df1,right=df2, how='left', on='SK_ID_CURR')
```

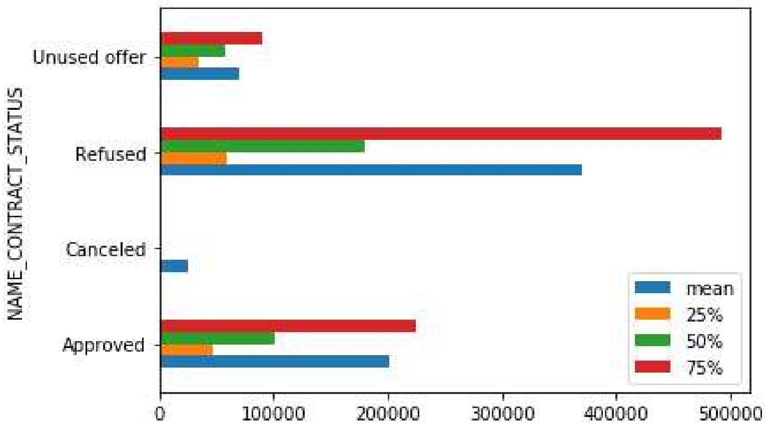
```
In [70]: df12.head()
```

```
: # Shape of the combined data set after dropiing columns with more than 45% missing values
  df12.shape
```

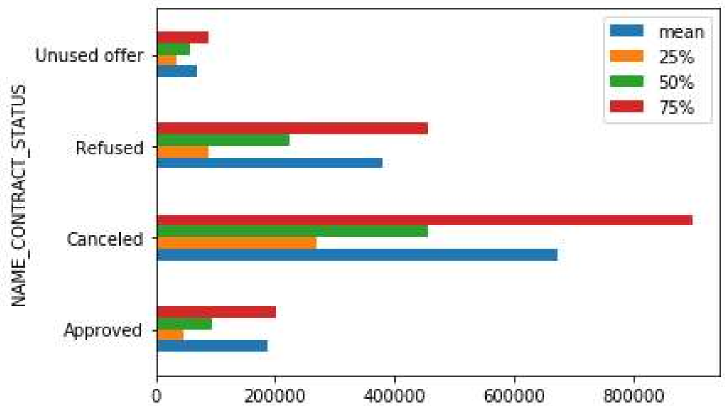
```
: (1430155, 107)
```

Insights from Univariate Analysis on Combined Dataset:

AMT_CREDIT



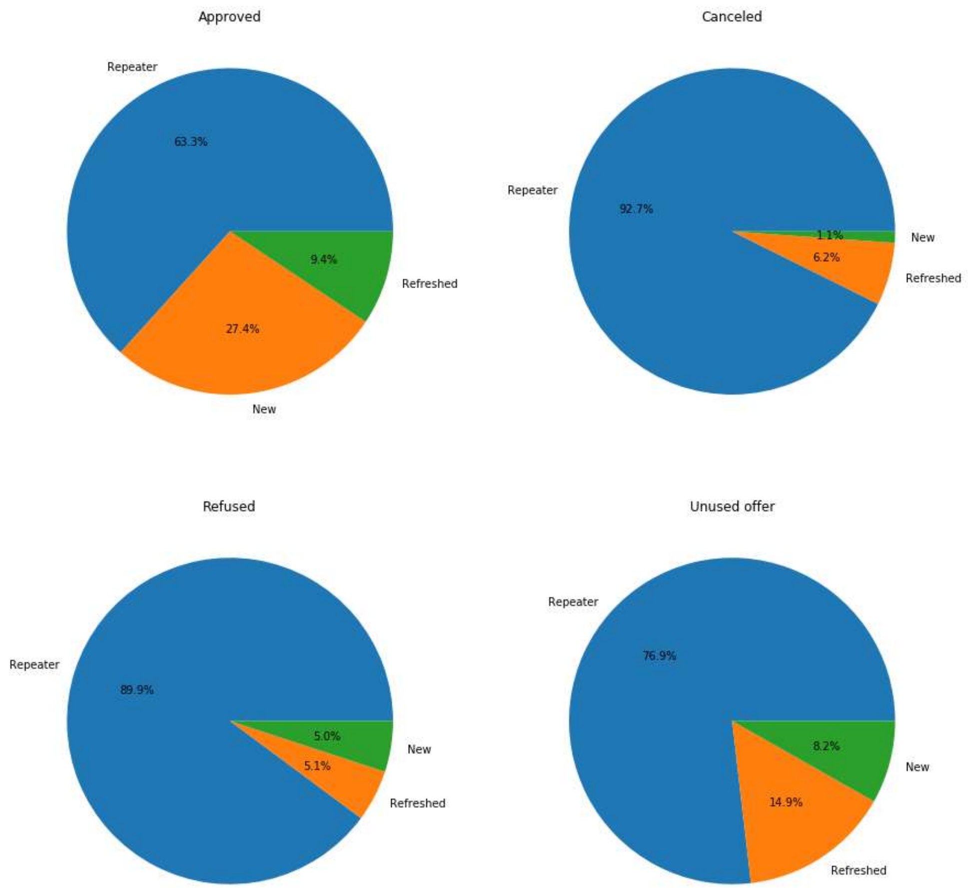
AMT_GOODS_PRICE



Outcomes:

- **Chances of Application being Refused is more if the Credit Amount is more than 2 Lakhs**
- **Chances of Application being Refused or getting cancelled is more if the Amount Goods price is more than 2 Lakhs**

Insights from Univariate Analysis on Combined Dataset:

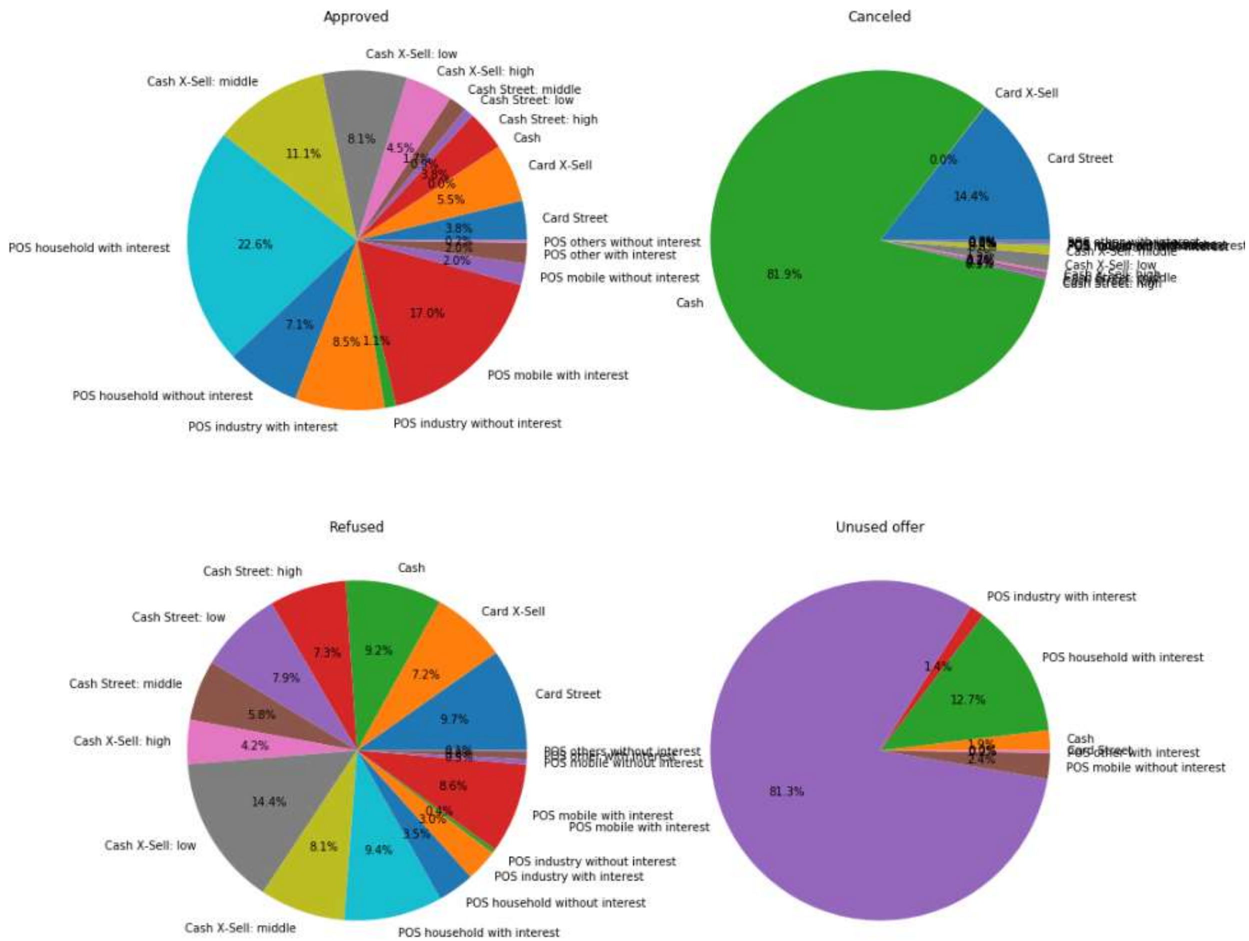


Column : NAME_CLIENT_TYPE

Outcomes:

- **Chances of an Application getting Approved is more for a New Applicant**
- **Chances of an applicant Un-use the offer is more for a Refreshed Applicant**
- **Chances of an Application getting Cancelled or Refused is very high for a Repeater**

Insights from Univariate Analysis on Combined Dataset:



Column: PRODUCT_COMBINATION

Outcomes:

- **Chances of an Application getting approved is more for POS household with Interest and overall POS products (>50%)**
- **Chances of an Application getting Either Cancelled or Refused is more for all Cash products (>50%)**

Conclusion:

- These are driving factors behind load default:
  - AMT_CREDIT
  - NAME_EDUCATION_TYPE
  - NAME_INCOME_TYPE
  - NAME_CLIENT_TYPE
  - PRODUCT_COMBINATION
  - AGE
  - DAYS_EMPLOYED
  - AMT_INCOME_TOTAL
  - CODE_GENDER

Thank you!