Credit EDA Case Study

For a loan application, the company has to decide for loan approval based on the applicant’s profile, identify patterns which indicate if a client has difficulty paying their instalments

- Bank wants to decide the driving factors behind loan default.
- Two types of decisions bank has to take:
  - ❖ If the applicant is likely to repay the loan, then bank should approve the loan
  - ❖ If the applicant is not likely to repay the loan, then bank should not approve the loan

Prepared By:
Dipali Visapure
Vaibhav Swarnkar

Data Manipulations

- Import the dataset Application.csv
- Inspecting the dataset.
  - ❖ Use functions like , describe, shape, info, columns , dtypes etc.
- Data quality and missing value checks
  - ❖ Finding the percentage of missing values for all columns
  - ❖ Removing columns with >= 50% null values
- Handling missing values for columns with 13% or less than null
  - ❖ Check the columns with missing values
  - ❖ Suggest methods to Impute these missing Values.
  - ❖ Following Fields are considered for Imputation methods(Mean, Median or Mode)

Categorical –

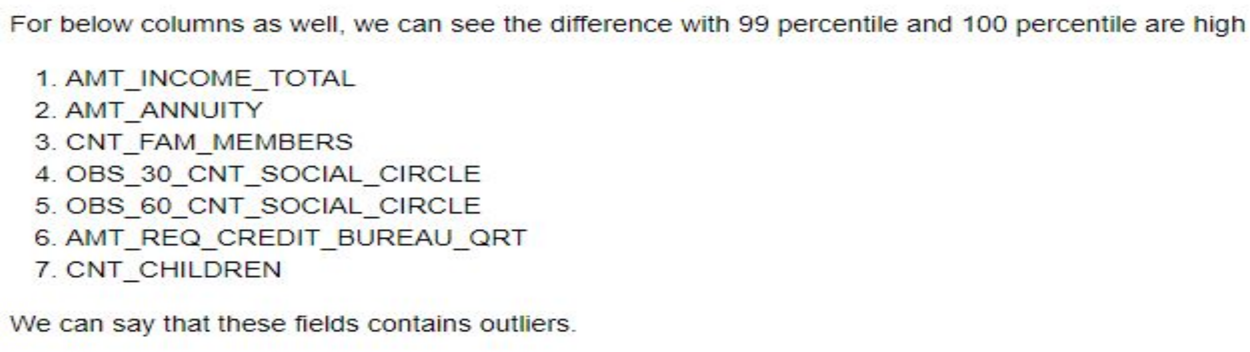
NAME_TYPE_SUITE(Mode)

Continuous –

AMT_REQ_CREDIT_BUREAU_DAY,(Mean)
OBS_30_CNT_SOCIAL_CIRCLE (Median)
EXT_SOURCE_2 (Mean)
AMT_GOODS_PRICE(Mean)

Data Manipulations

- Converting values – Negative to Positive (Field- DAYS_BIRTH)
- Handling Outliers
  - ❖ Use functions like describe, quantile to identify outliers, fields indented for outliers are:

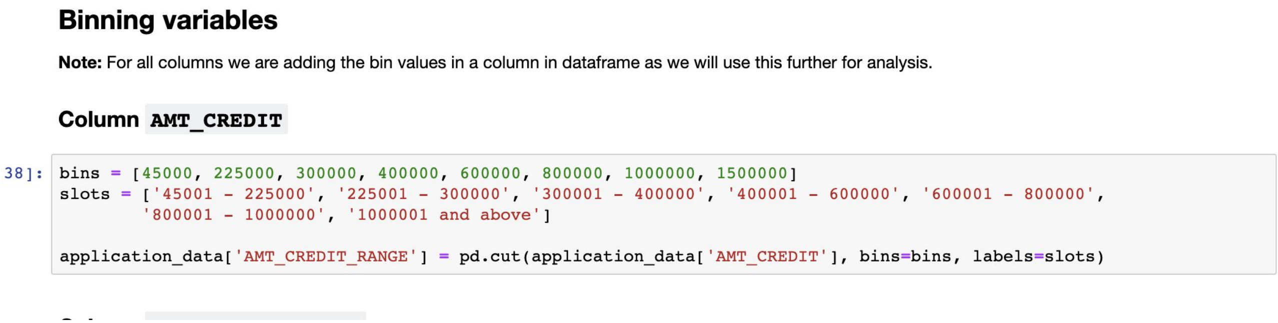


For below columns as well, we can see the difference with 99 percentile and 100 percentile are high

1. AMT_INCOME_TOTAL
2. AMT_ANNUITY
3. CNT_FAM_MEMBERS
4. OBS_30_CNT_SOCIAL_CIRCLE
5. OBS_60_CNT_SOCIAL_CIRCLE
6. AMT_REQ_CREDIT_BUREAU_QRT
7. CNT_CHILDREN

We can say that these fields contains outliers.

- Binning of continuous variables
  - ❖ Use functions like describe, quantile to identify outliers, fields indentified for binning are: AMT_CREDIT, AMT_INOME_TOTAL, CNT_FAM_MEMBERS & DAYS_BIRTH



Binning variables

Note: For all columns we are adding the bin values in a column in dataframe as we will use this further for analysis.

Column `AMT_CREDIT`

38]:

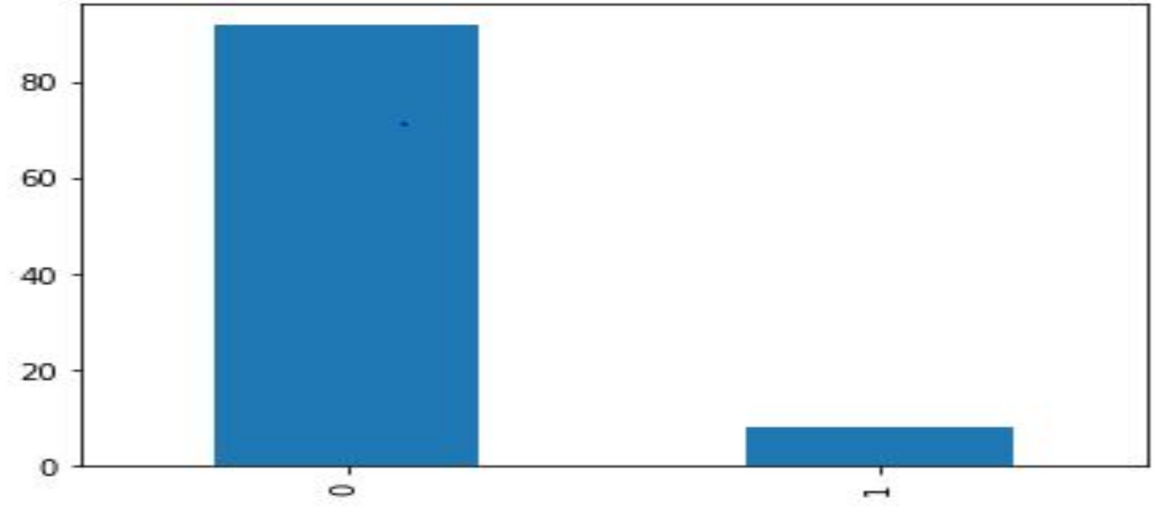
```
bins = [45000, 225000, 300000, 400000, 600000, 800000, 1000000, 1500000]
slots = ['45001 - 225000', '225001 - 300000', '300001 - 400000', '400001 - 600000', '600001 - 800000',
        '800001 - 1000000', '1000001 and above']

application_data['AMT_CREDIT_RANGE'] = pd.cut(application_data['AMT_CREDIT'], bins=bins, labels=slots)
```

Analysis

- Check the Imbalance percentage- Target(0):Target(1)

i.e Defaulters vs Non-Defaulters

Here, our data set is highly imbalanced, with almost 92% customers are paying their loans on time and are not defaulters. So approx. Imbalance percentage 8%



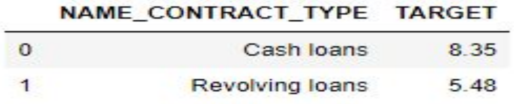
- Split the original dataset - Create two subsets using column TARGET(0/1)

Analysis

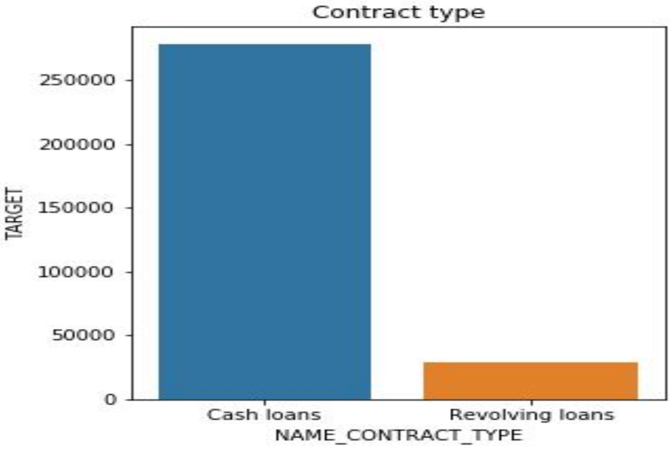
- Analysis wrt Target Variables:

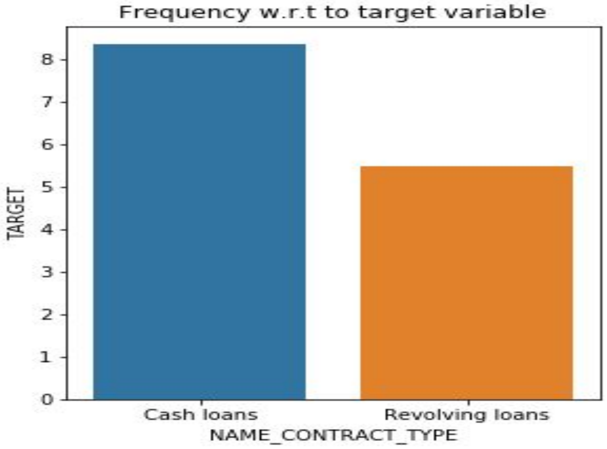
- NAME_CONTRACT_TYPE –

Univariate analysis of this column reveals that in the default list with the Cash loans type have around 8.35% of defaulters, more than the Revolving loans. It is anticipated that Cash loans might have more defaults than Revolving loans.



	NAME_CONTRACT_TYPE	TARGET
0	Cash loans	8.35
1	Revolving loans	5.48



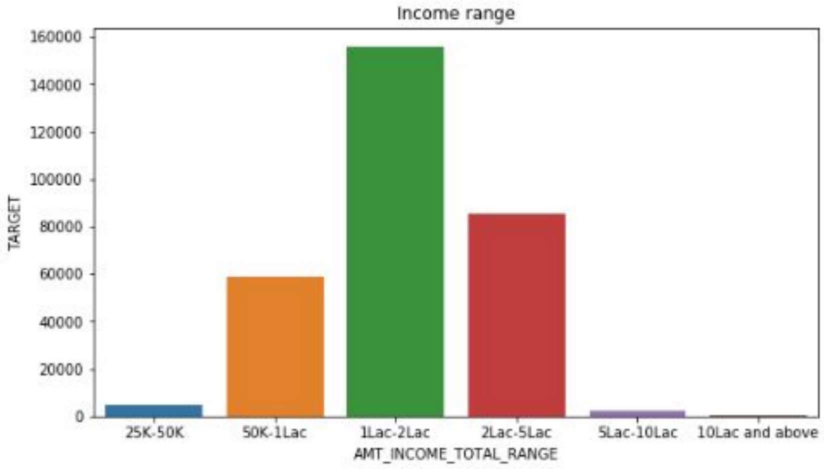


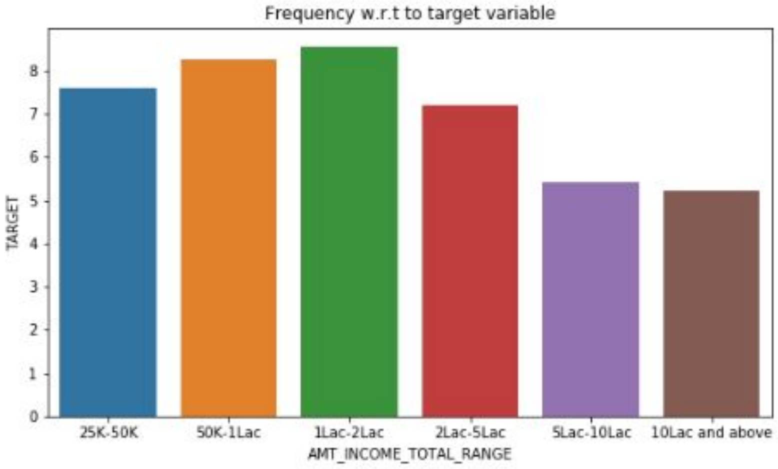
Analysis

- Analysis wrt Target Variables:

- AMT_INCOME_TOTAL

Univariate analysis of this column reveals that in the default list, people with income range between 1Lac - 2Lacs have 8.55% chances of being a default.



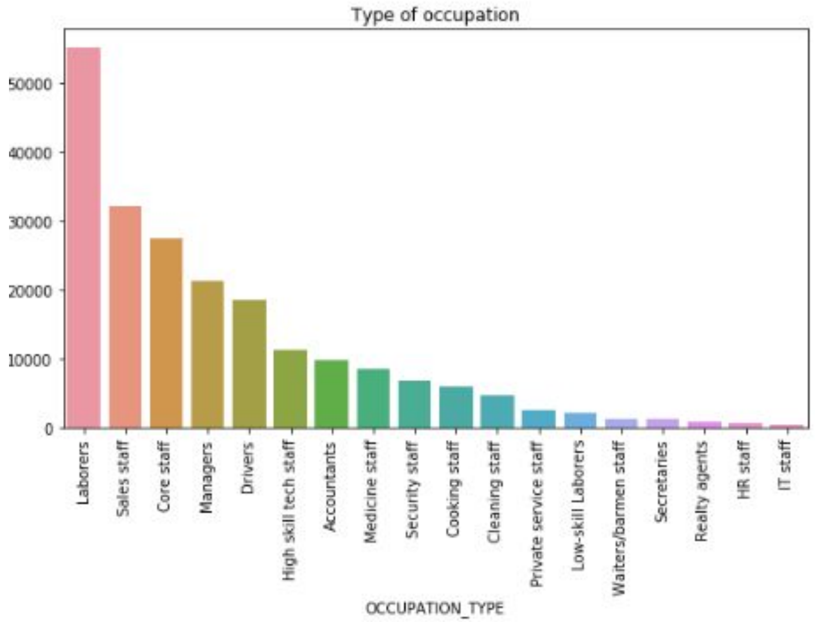


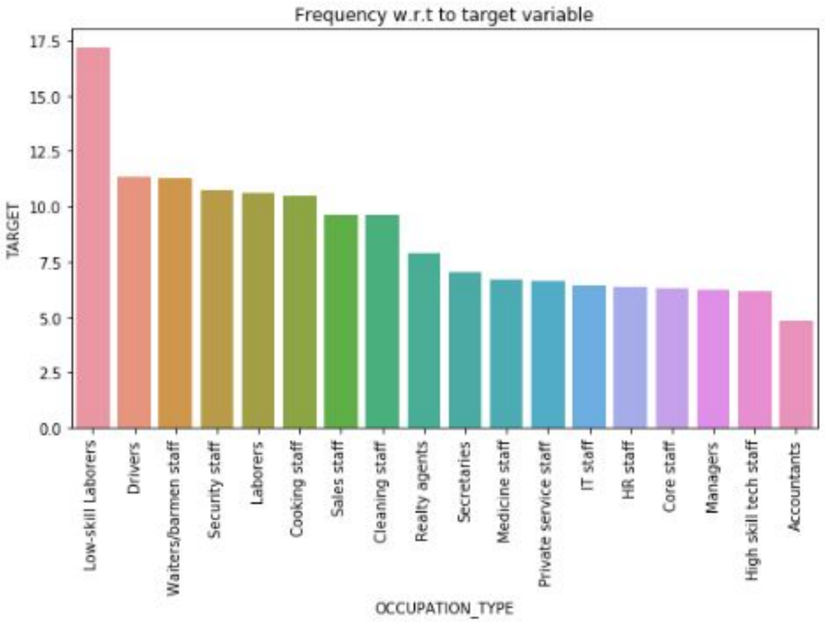
Analysis

- Univariate and Bivariate Analysis wrt Target Variables:

- OCCUPATION_TYPE–

The Occupation type of Low-skill Laborers have the chances of around 17.5% of getting default in loan.





Analysis

- Bivariate Analysis wrt Target Variables:

. CNT_FAM_MEMBERS–

In total, family with 1 or 2 members applied mostly for loan however, family of members 6-10 have chances of 12.61% of being at default and then family of members 11-15.

. NAME_HOUSING_TYPE -

The count of housing type in application_data dataset as House/Apartment highest frequency but when compared W.R.T target variable Rented apartment people 12.31% chances of being at default.

. DAYS_BIRTH -

The age range is between appx 20 and 69, and the loan is taken majorly by age range 54-60. When compared to target variable, it turned out to be that the age range 20-30 seems to be having problem majorly in payments and reduces as the age increases.

. ORGANIZATION_TYPE

When looking at the entire dataset, most of the applicant's type of organization type is Business Entity Type3 and default in payments turns out to be from organization type Transport Type3 with appx rate of ~16%.

CODE_GENDER–

In Input dataset, most of the applicant's gender code is Female , but in Defaulters Male with appx rate of ~10% while Femaile with appx ~7%.

. FLAG_OWN_CAR -

In Input dataset, most of the applicant's FLAG_OWN_CAR flag is 'No', In Defaulters,

Defaulters with FLAG_OWN_CAR Yes :8.5% Defaulters with FLAG_OWN_CAR No : 7.24%.

. FLAG_OWN_REALTY -

In Input dataset, most of the applicant's FLAG_OWN_CAR flag is 'No', In Defaulters,

Defaulters with FLAG_OWN_REALTY is Yes :8.32% Defaulters with FLAG_OWN_REALTY is No : 7.96%.

. NAME_EDUCATION_TYPE –

In Input dataset, most of the applicant's has NAME_EDUCATION_TYPE as econdary / secondary special, But in

Defaulters with Academic degree is : 1.83% Defaulters with Higher education is : 5.36% Defaulters with Incomplete higher is : 8.48% Defaulters with Lower secondary is : 10.93% Defaulters with Secondary / secondary special is : 8.94%

Analysis

- Bivariate Analysis wrt Target Variables:

. CODE_GENDER–

In Input dataset, most of the applicant's gender code is Female , but in Defaulters Male with appx rate of ~10% while Femaile with appx ~7%.

. FLAG_OWN_CAR -

In Input dataset, most of the applicant's FLAG_OWN_CAR flag is 'No', In Defaulters,

Defaulters with FLAG_OWN_CAR Yes :8.5% Defaulters with FLAG_OWN_CAR No : 7.24%.

. FLAG_OWN_REALTY -

In Input dataset, most of the applicant's FLAG_OWN_CAR flag is 'No', In Defaulters,

Defaulters with FLAG_OWN_REALTY is Yes :8.32% Defaulters with FLAG_OWN_REALTY is No : 7.96%.

. NAME_EDUCATION_TYPE –

In Input dataset, most of the applicant's has NAME_EDUCATION_TYPE as econdary / secondary special, But in

Defaulters with Academic degree is : 1.83% Defaulters with Higher education is : 5.36% Defaulters with Incomplete higher is : 8.48% Defaulters with Lower secondary is : 10.93% Defaulters with Secondary / secondary special is : 8.94%

- Bivariate Analysis wrt Target Variables:

. NAME_FAMILY_STATUS–

In Input dataset, most of the applicant's has NAME_FAMILY_STATUS as Married,

In Defaulters with Civil marriage : 9.94 Defaulters with Married : 7.56% Defaulters with Separated : 8.19% Defaulters with Single / not married : 9.81% Defaulters with Widow : 5.82%

. NAME_INCOME_TYPE -

In Input dataset, most of the applicant's has NAME_INCOME_TYPE as Working, In Defaulters, Defaulters with Income_type Businessman : almost 0%, Defaulters with Commercial associate : 7.48% , Defaulters with Maternity leave : 40.00%, Defaulters with Pensioner : 5.39% , Defaulters with State servant : 5.75% , Defaulters with Student : almost 0% , Defaulters with Unemployed : 36.36% , Defaulters with Working : 9.59% ,

Analysis

- Bivariate Analysis wrt Target Variables:

. NAME_FAMILY_STATUS–

In Input dataset, most of the applicant's has NAME_FAMILY_STATUS as Married,

In Defaulters with Civil marriage : 9.94 Defaulters with Married : 7.56% Defaulters with Separated : 8.19% Defaulters with Single / not married : 9.81% Defaulters with Widow : 5.82%

. NAME_INCOME_TYPE -

In Input dataset, most of the applicant's has NAME_INCOME_TYPE as Working, In Defaulters, Defaulters with Income_type Businessman : almost 0%, Defaulters with Commercial associate : 7.48% , Defaulters with Maternity leave : 40.00%, Defaulters with Pensioner : 5.39% , Defaulters with State servant : 5.75% , Defaulters with Student : almost 0% , Defaulters with Unemployed : 36.36% , Defaulters with Working : 9.59% ,

. CNT_CHILDREN –

In Input dataset, most of the applicant's has CNT_CHILDREN less than 4, But in defaulters people having more children have very high chances to default

. AMT_ANNUITY –

People having less amount of annuity have high chances to default

. CNT_CHILDREN –

In Input dataset, most of the applicant's has CNT_CHILDREN less than 4, But in defaulters people having more children have very high chances to default

. AMT_ANNUITY –

People having less amount of annuity have high chances to default

. NAME_INCOME_TYPE -

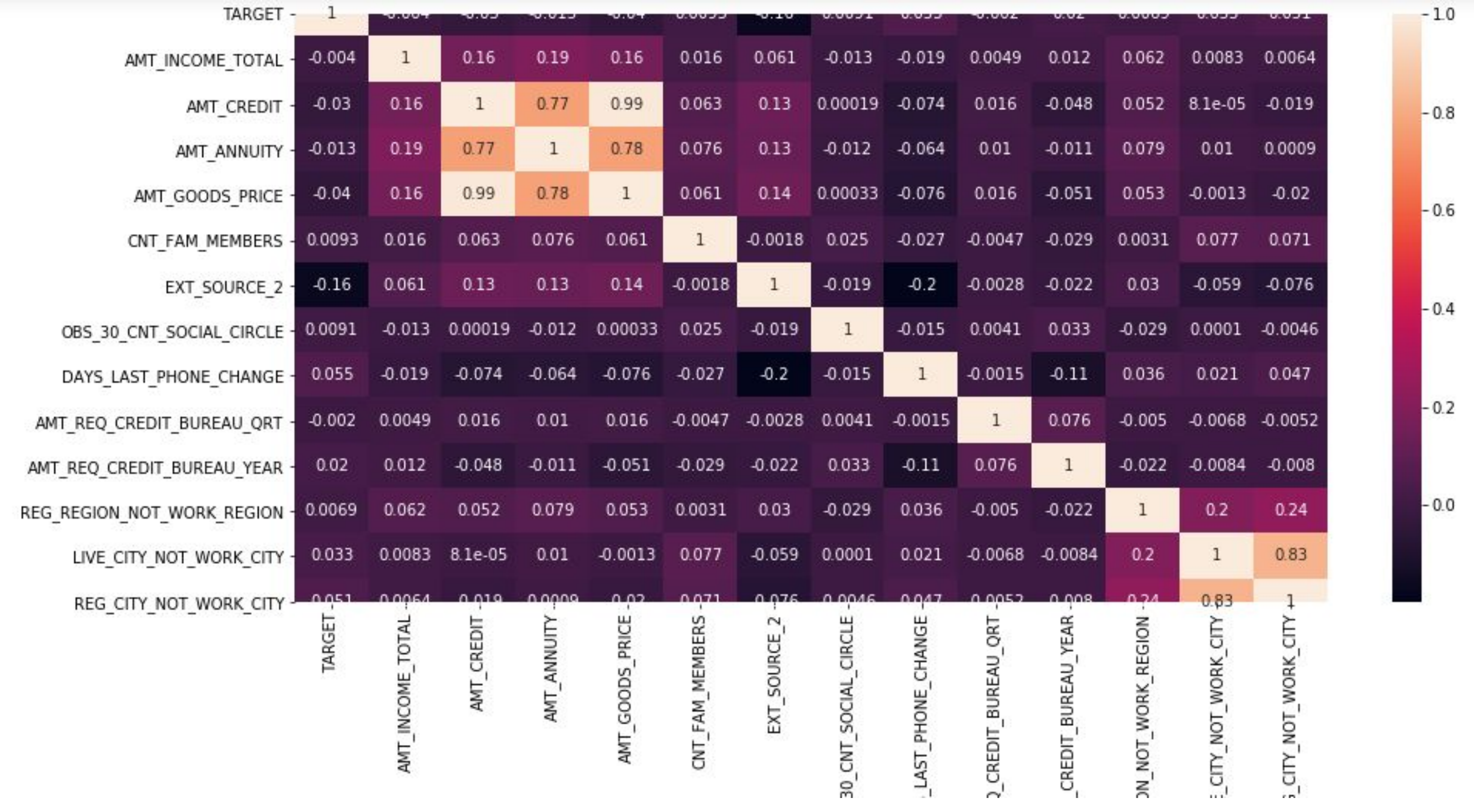
In Input dataset, most of the applicant's has NAME_INCOME_TYPE as Working, In Defaulters, Defaulters with Income_type Businessman : almost 0%, Defaulters with Commercial associate : 7.48% , Defaulters with Maternity leave : 40.00%, Defaulters with Pensioner : 5.39% , Defaulters with State servant : 5.75% , Defaulters with Student : almost 0% , Defaulters with Unemployed : 36.36% , Defaulters with Working : 9.59% ,

. CNT_CHILDREN –

In Input dataset, most of the applicant's has CNT_CHILDREN less than 4, But in defaulters people having more children have very high chances to default

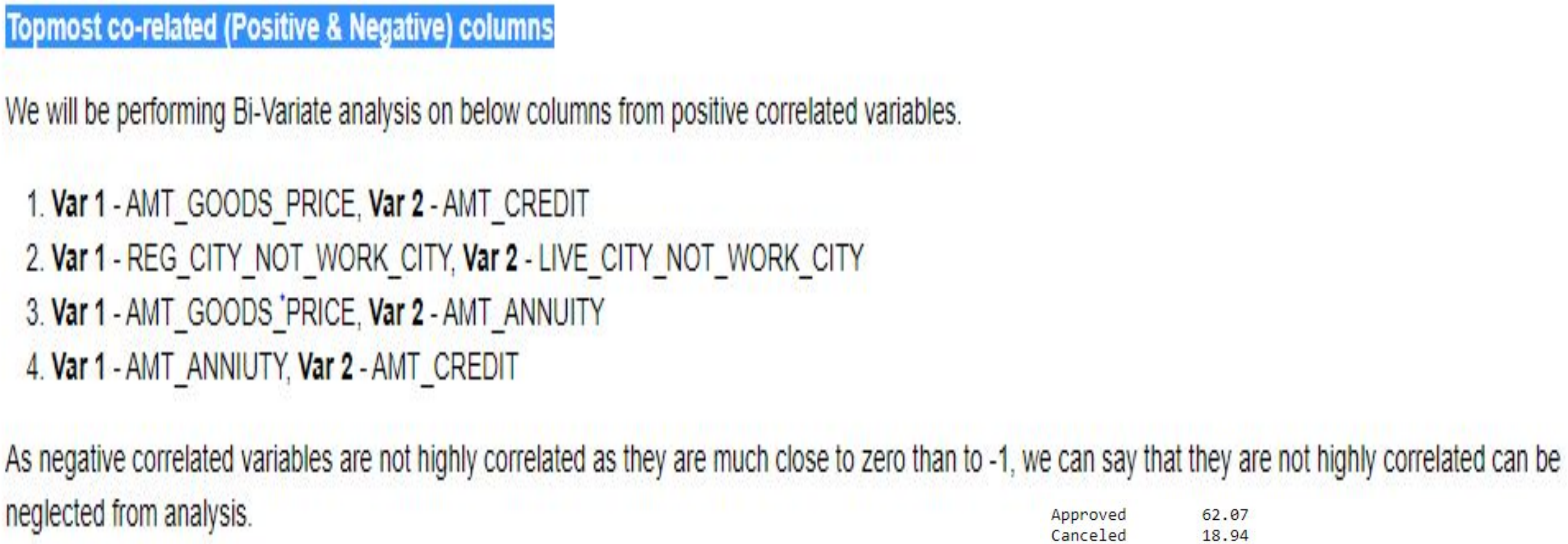
Analysis

Create corr_matrix for Selected numerical varibales & Plot



Analysis

- Identify Topmost co-related (Positive & Negative) columns



Topmost co-related (Positive & Negative) columns

We will be performing Bi-Variate analysis on below columns from positive correlated variables.

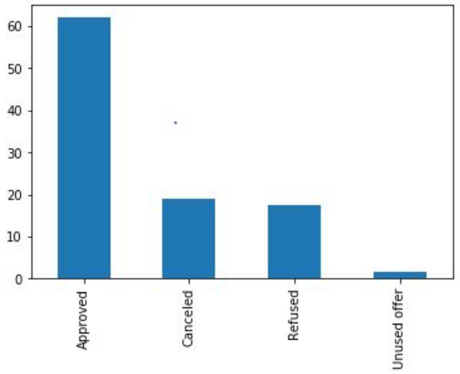
1. **Var 1** - AMT_GOODS_PRICE, **Var 2** - AMT_CREDIT
2. **Var 1** - REG_CITY_NOT_WORK_CITY, **Var 2** - LIVE_CITY_NOT_WORK_CITY
3. **Var 1** - AMT_GOODS_PRICE, **Var 2** - AMT_ANNUITY
4. **Var 1** - AMT_ANNIUTY, **Var 2** - AMT_CREDIT

As negative correlated variables are not highly correlated as they are much close to zero than to -1, we can say that they are not highly correlated can be neglected from analysis.

- Import dataset previous_application.csv
- Plot the Decision frequency wrt column NAME_CONTRACT_STATUS

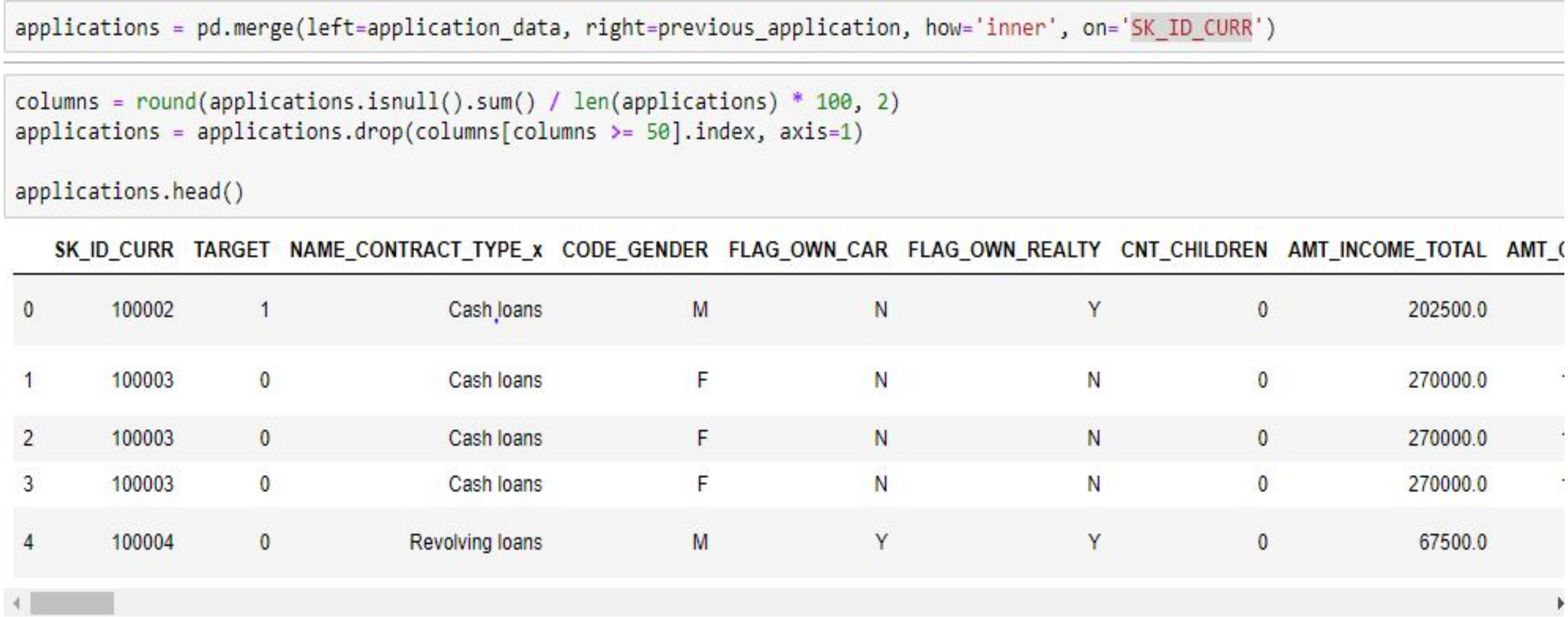
From the previous year's total applied application almost 62% were approved, while ~18% were cancelled, ~17% were refused and ~1% were Unused offer.

```
Approved        62.07
Canceled        18.94
Refused         17.40
Unused offer     1.58
Name: NAME_CONTRACT_STATUS, dtype: float64
```



Merge and manipulate 2 datasets

- Merge two input datasets, application_data.csv and previous_application.csv on field SK_ID_CURR with Inner Join.
- Drop the columns with more than 50 % Null values.



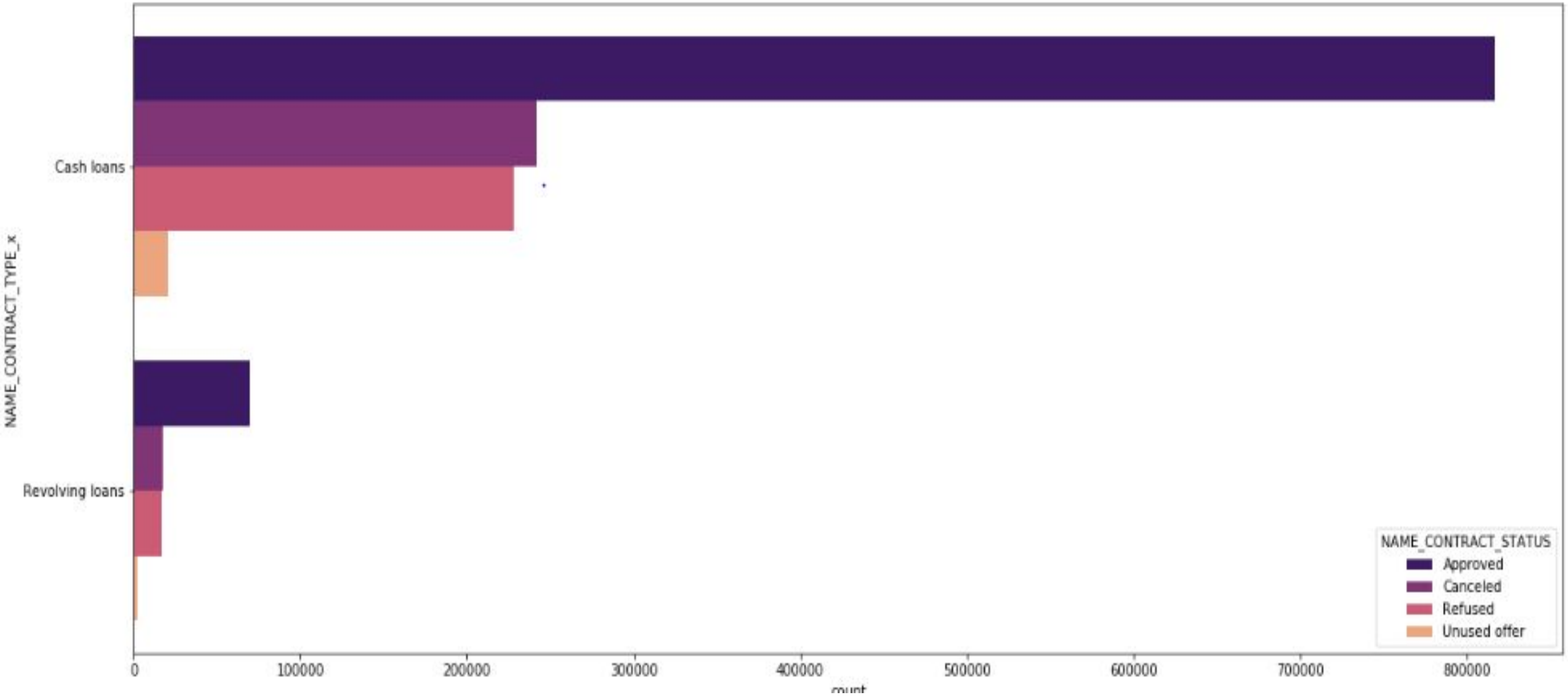
```
applications = pd.merge(left=application_data, right=previous_application, how='inner', on='SK_ID_CURR')
```

```
columns = round(applications.isnull().sum() / len(applications) * 100, 2)
applications = applications.drop(columns[columns >= 50].index, axis=1)

applications.head()
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE_x	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_(
0	100002	1	Cash loans	M	N	Y	0	202500.0	
1	100003	0	Cash loans	F	N	N	0	270000.0	
2	100003	0	Cash loans	F	N	N	0	270000.0	
3	100003	0	Cash loans	F	N	N	0	270000.0	
4	100004	0	Revolving loans	M	Y	Y	0	67500.0	

Analysis of Merged Dataset

- Analyze the merged dataset wrt column NAME_CONTRACT_STATUS
- Field NAME_CONTRACT_TYPE :
- In the analysis, it is uncovered, that Cash loans were approved at the highest rate than compared to Revolving loans



Analysis of Merged Dataset

- Field AMT_INCOME_TOTAL:

Loan was approved mostly for clients with income range of 1Lac to 2Lacs, followed by 2Lacs to 5Lacs, while surprisingly, loan application was also cancelled for the 1Lac to 2Lacs is also highest followed by 2Lacs to 5Lacs income range.

- Field ORGANIZATION_TYPE:

Most of the loans were approved with client organization's type Business Entity Type 3 and then followed by clients whom organization type is unknown. Loan was also cancelled by client's whom organization type is un

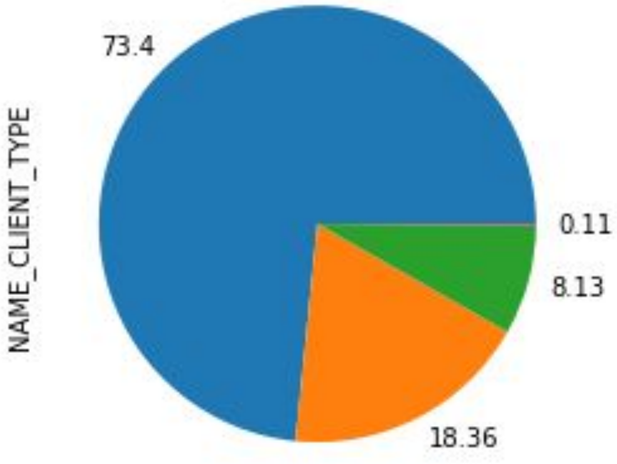
known.

- AGE_BIRTH:

Loan was approved majorly for client's with age range 54-60 follwed by range 42-47 the 31-36. Loan is cancelled by clients highly in range 54-60 followed by age range 61-above.

- Field NAME_CLIENT_TYPE:

Almost 73% of the clients are repeating clients.


NAME_CLIENT_TYPE