Case Study
Bank Loan Default Risk Analysis

Exploratory Data Analysis

Content

- Load the data
- Check the data
  - Glimpse the data
  - Check missing data
  - Check data unbalance
- Explore the data
  - Deleting the columns where null values are more than 50%
  - Identifying outliers
  - Univariate, Bivariate and Segmented Univariate analysis

Analysing Application_data.csv file

Load the data

```
In [2]: df_application = pd.read_csv(r'C:\Users\xssadinema\Desktop\Credit EDA\application_data.csv')

In [3]: df_application.head()
```

Out[3]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CRE
0	100002	1	Cash loans	M	N	Y	0	202500.0	40659
1	100003	0	Cash loans	F	N	N	0	270000.0	129350
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	13500
3	100006	0	Cash loans	F	N	Y	0	135000.0	31268
4	100007	0	Cash loans	M	N	Y	0	121500.0	51300

5 rows × 122 columns

Loading the data is the first step of any analysis

Check the data

- Glimpse the data

```
In [3]: df_application.columns

Out[3]: Index(['SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER',
               'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL',
               'AMT_CREDIT', 'AMT_ANNUITY',
               ...
               'FLAG_DOCUMENT_18', 'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20',
               'FLAG_DOCUMENT_21', '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'],
              dtype='object', length=122)
```

```
In [4]: df_application.shape

Out[4]: (307511, 122)
```

Loading data and finding the shape of the data frame which has **307511** rows and **122** columns

- Check missing data

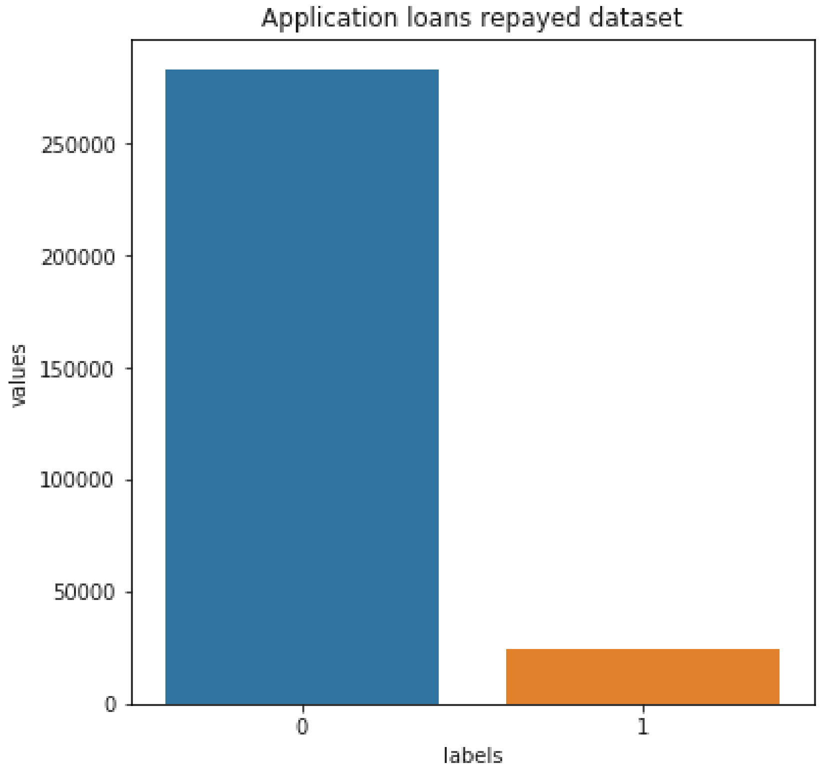
```
In [6]: missing_data(df_application).head(10)
```

Out[6]:

	Total	Percent
COMMONAREA_MEDI	214865	69.872297
COMMONAREA_AVG	214865	69.872297
COMMONAREA_MODE	214865	69.872297
NONLIVINGAPARTMENTS_MODE	213514	69.432963
NONLIVINGAPARTMENTS_MEDI	213514	69.432963
NONLIVINGAPARTMENTS_AVG	213514	69.432963
FONDKAPREMONT_MODE	210295	68.386172
LIVINGAPARTMENTS_MEDI	210199	68.354953
LIVINGAPARTMENTS_MODE	210199	68.354953
LIVINGAPARTMENTS_AVG	210199	68.354953

Checking the total null values and percent of null values column wise

- Check data unbalance



We are checking the data unbalance for the **TARGET** variable.
This is to find the ratio between the clients having difficulty in paying the loan versus all the other cases.
Target 1 means clients having payment difficulties
Target 0 means all the other cases
Ratio between both the cases is **11.39**

Explore the data

- Deleting the columns where null values are more than 50%

  After deleting above columns, the number of columns remained are given as:

```
In [6]: missing_data(df_app).head()
```

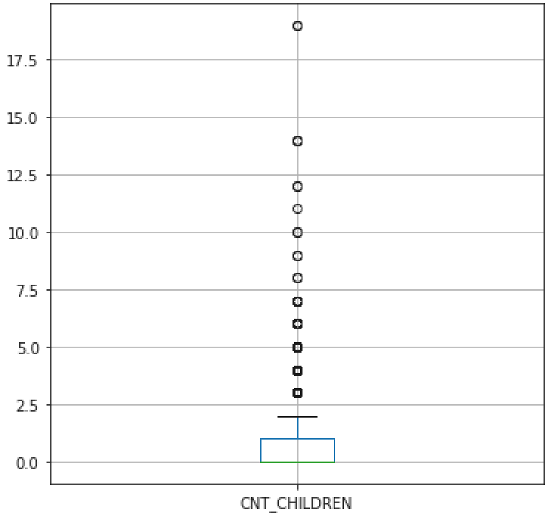
Out[6]:

	Total	Percent
FLOORSMAX_AVG	153020	49.760822
FLOORSMAX_MEDI	153020	49.760822
FLOORSMAX_MODE	153020	49.760822
YEARS_BEGINEXPLUATATION_AVG	150007	48.781019
YEARS_BEGINEXPLUATATION_MEDI	150007	48.781019

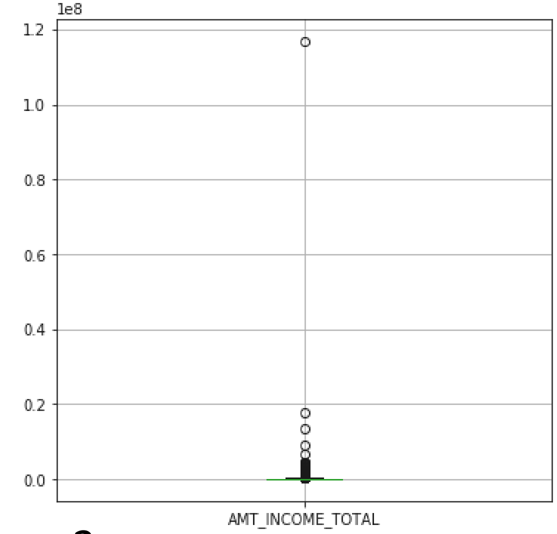
```
In [7]: df_app.shape
```

```
Out[7]: (307511, 81)
```

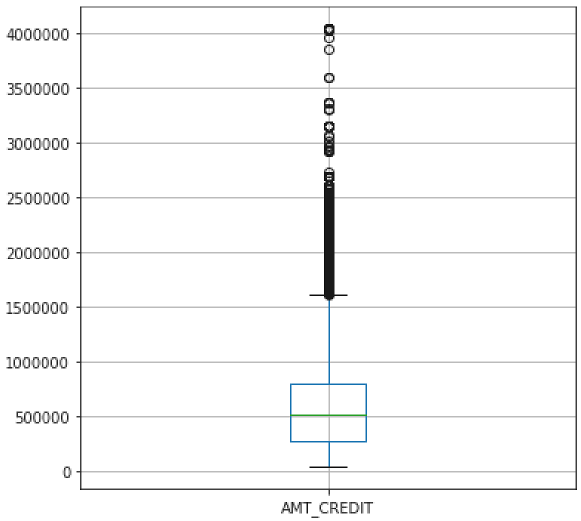
- Identifying Outliers



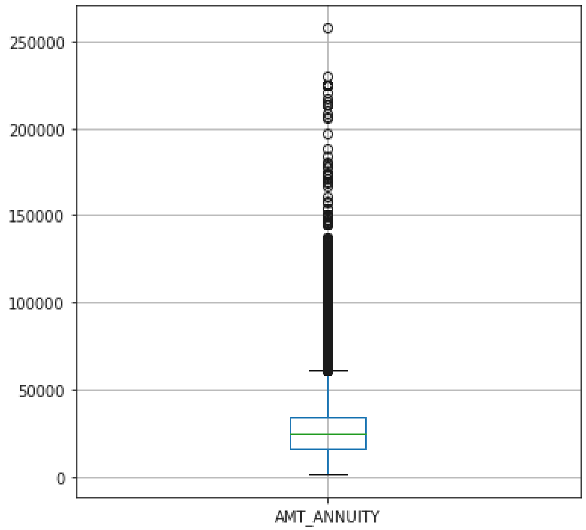
1.



2.



3.



4.

These are the outliers representation of various numerical variables using Box plot.

The various variables shown in the left are:

1. CNT_CHILDREN
2. AMT_INCOME_TOTAL
3. AMT_CREDIT
4. AMT_ANNUITY

- Univariate, Bivariate and Segmented Univariate analysis

Univariate Analysis

- Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; it takes data, summarizes that data and finds patterns in the data.

- Analysing the particular column is called Univariate Analysis. Which either have
  - Categorical variables
  - Quntitative or numerical variables
- We can get summary metrics of the particular column.

Bivariate Analysis

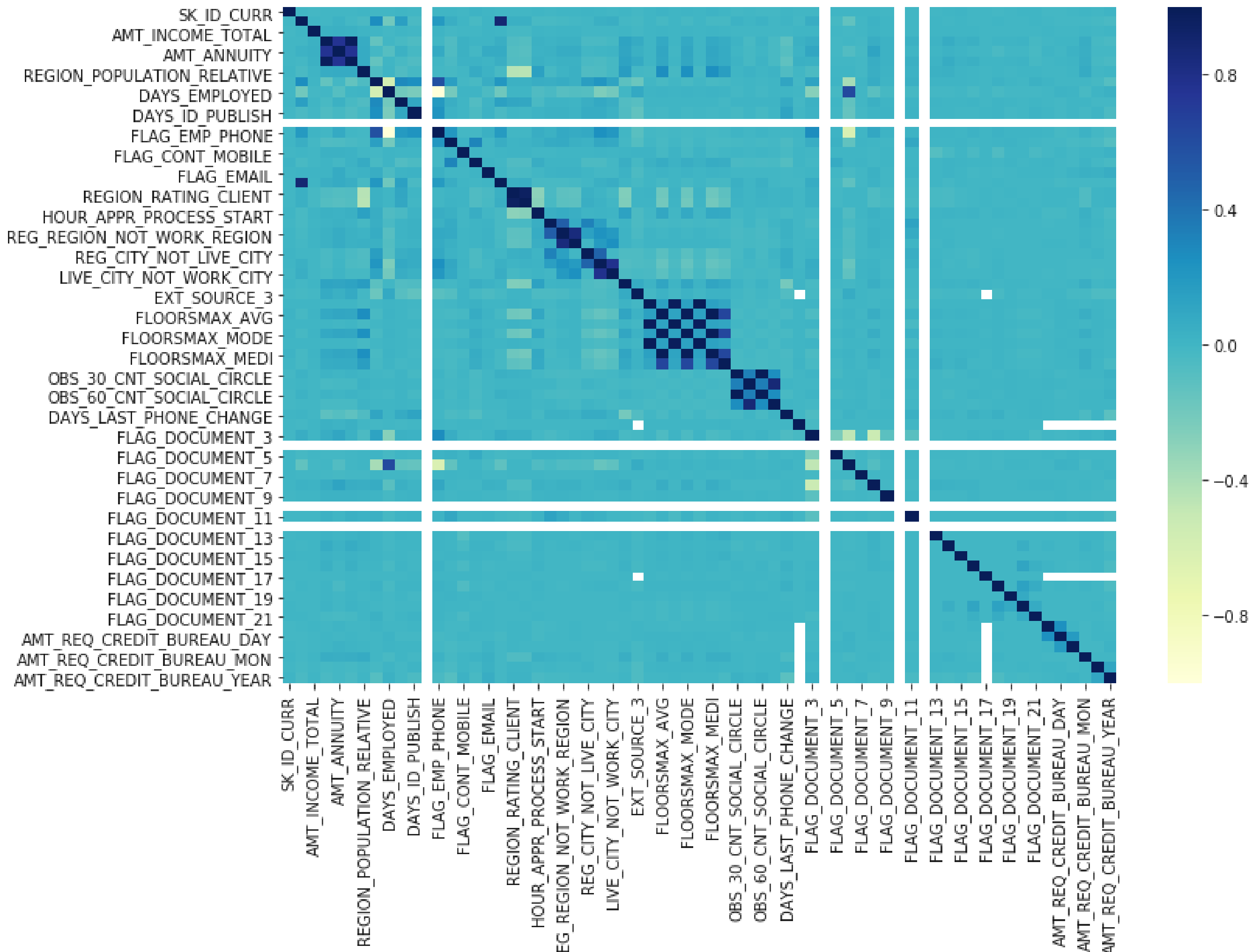
- Bivariate analsis means the relationship between two variables.We need to perform Bivariate Analysis on
  - Continous variables
  - Categorical variables

Segmented Univariate

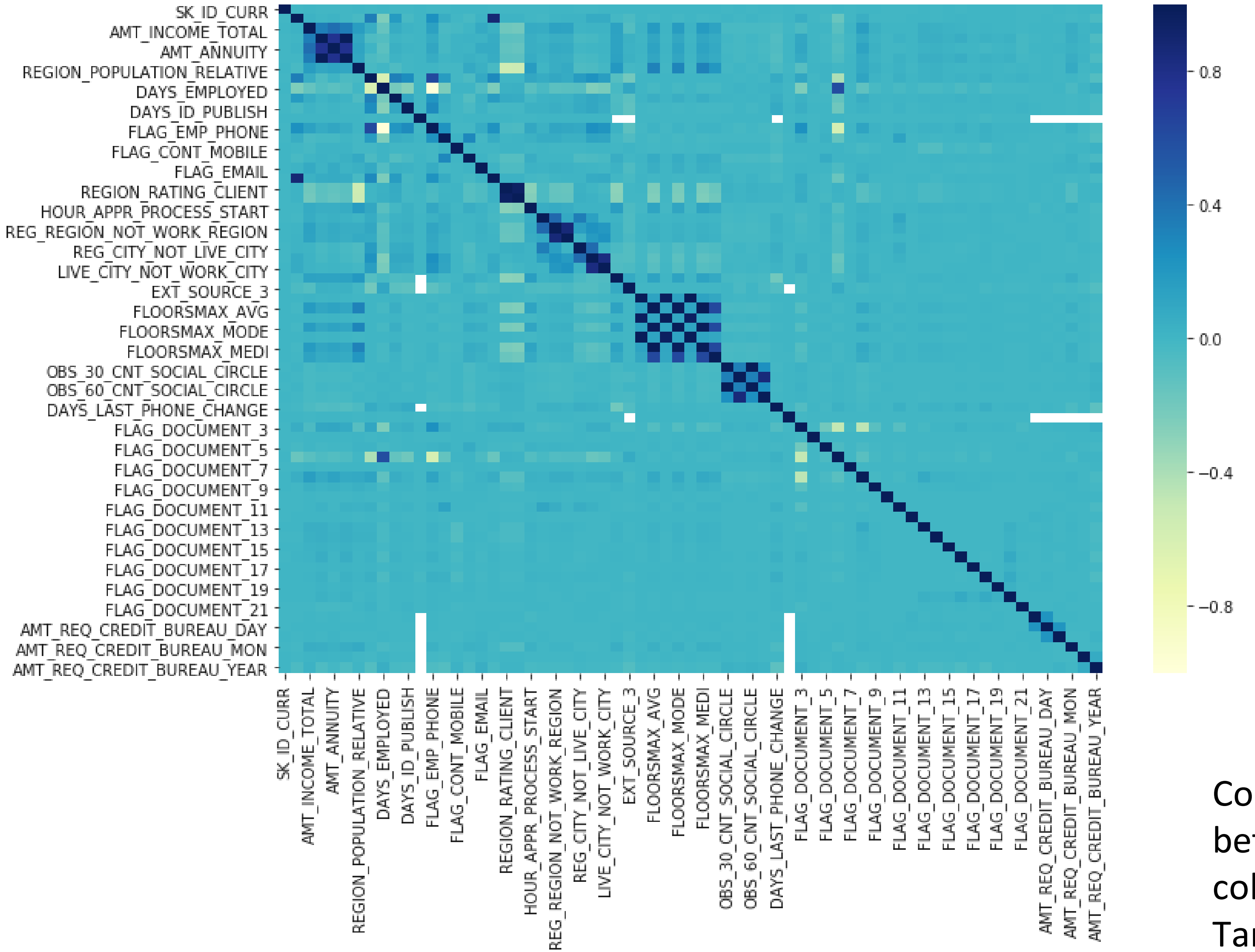
- Segmented univariate analysis allows you to compare subsets of data,it helps you understand how a relevant metric varies across different segments.In the segmented univariate analysis useful insights are extracted by conducting univariate analysis on segments on data.

Dividing into two sets based on Target variable i.e., into Target 0 and Target 1 and doing **bivariate** analysis on continuous(numerical) variables.

Finding correlation between numerical columns of both dataset (Target 0 and Target 1)



Correlation between columns of Target 1



Correlation between columns of Target 0

Top **10 correlation** of Target 1 dataset:

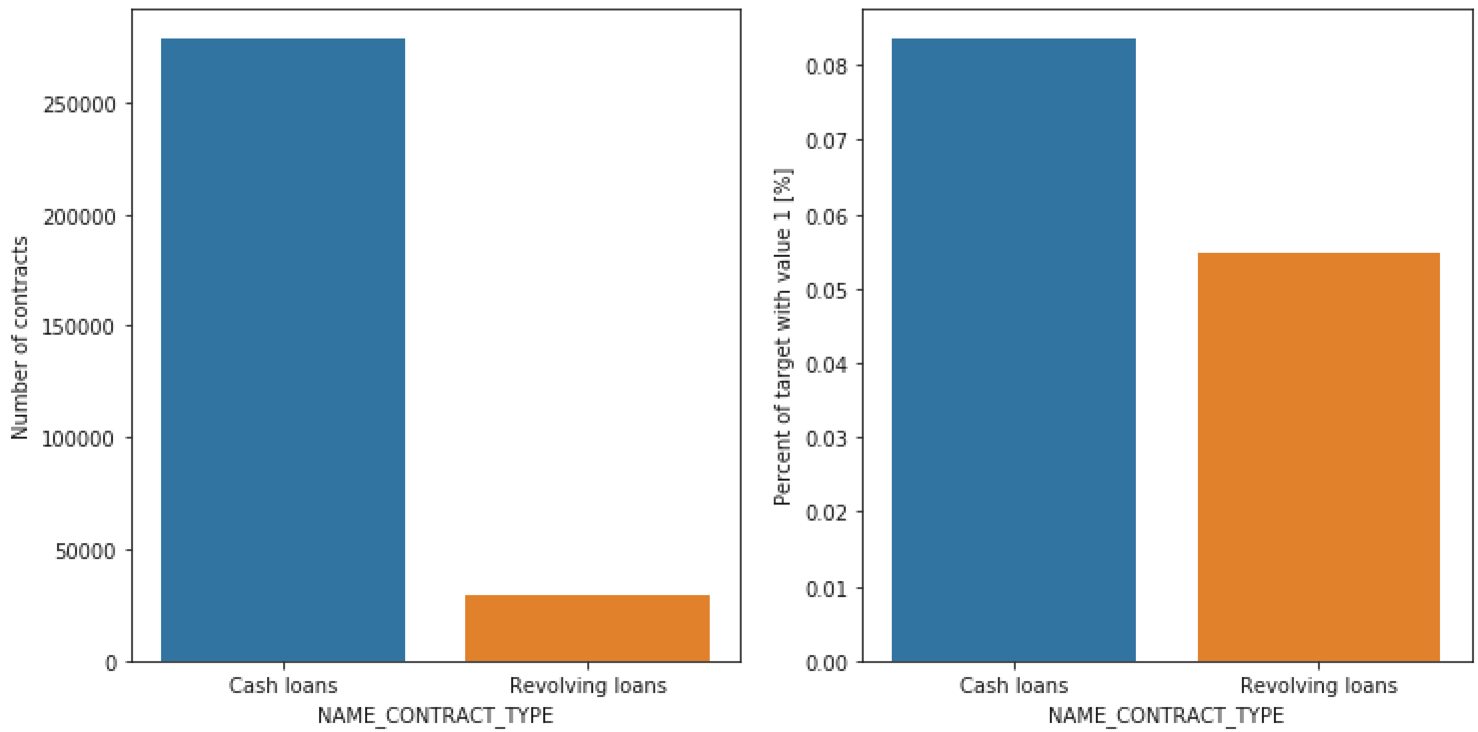
```
Top Absolute Correlations
DAYS_EMPLOYED                  FLAG_EMP_PHONE                   0.999702
OBS_30_CNT_SOCIAL_CIRCLE       OBS_60_CNT_SOCIAL_CIRCLE         0.998269
FLOORSMAX_AVG                  FLOORSMAX_MEDI                   0.997187
YEARS_BEGINEXPLUATATION_AVG    YEARS_BEGINEXPLUATATION_MEDI     0.996124
FLOORSMAX_MODE                 FLOORSMAX_MEDI                   0.989195
FLOORSMAX_AVG                  FLOORSMAX_MODE                   0.986594
AMT_CREDIT                     AMT_GOODS_PRICE                  0.983103
YEARS_BEGINEXPLUATATION_AVG    YEARS_BEGINEXPLUATATION_MODE     0.980466
YEARS_BEGINEXPLUATATION_MODE   YEARS_BEGINEXPLUATATION_MEDI     0.978073
REGION_RATING_CLIENT           REGION_RATING_CLIENT_W_CITY      0.956637
dtype: float64
```

Top **10 correlation** of Target 0 dataset:

```
Top Absolute Correlations
DAYS_EMPLOYED                  FLAG_EMP_PHONE                   0.999758
OBS_30_CNT_SOCIAL_CIRCLE       OBS_60_CNT_SOCIAL_CIRCLE         0.998508
FLOORSMAX_AVG                  FLOORSMAX_MEDI                   0.997018
YEARS_BEGINEXPLUATATION_AVG    YEARS_BEGINEXPLUATATION_MEDI     0.993582
FLOORSMAX_MODE                 FLOORSMAX_MEDI                   0.988153
AMT_CREDIT                     AMT_GOODS_PRICE                  0.987250
FLOORSMAX_AVG                  FLOORSMAX_MODE                   0.985603
YEARS_BEGINEXPLUATATION_AVG    YEARS_BEGINEXPLUATATION_MODE     0.971032
YEARS_BEGINEXPLUATATION_MODE   YEARS_BEGINEXPLUATATION_MEDI     0.962064
REGION_RATING_CLIENT           REGION_RATING_CLIENT_W_CITY      0.950149
dtype: float64
```

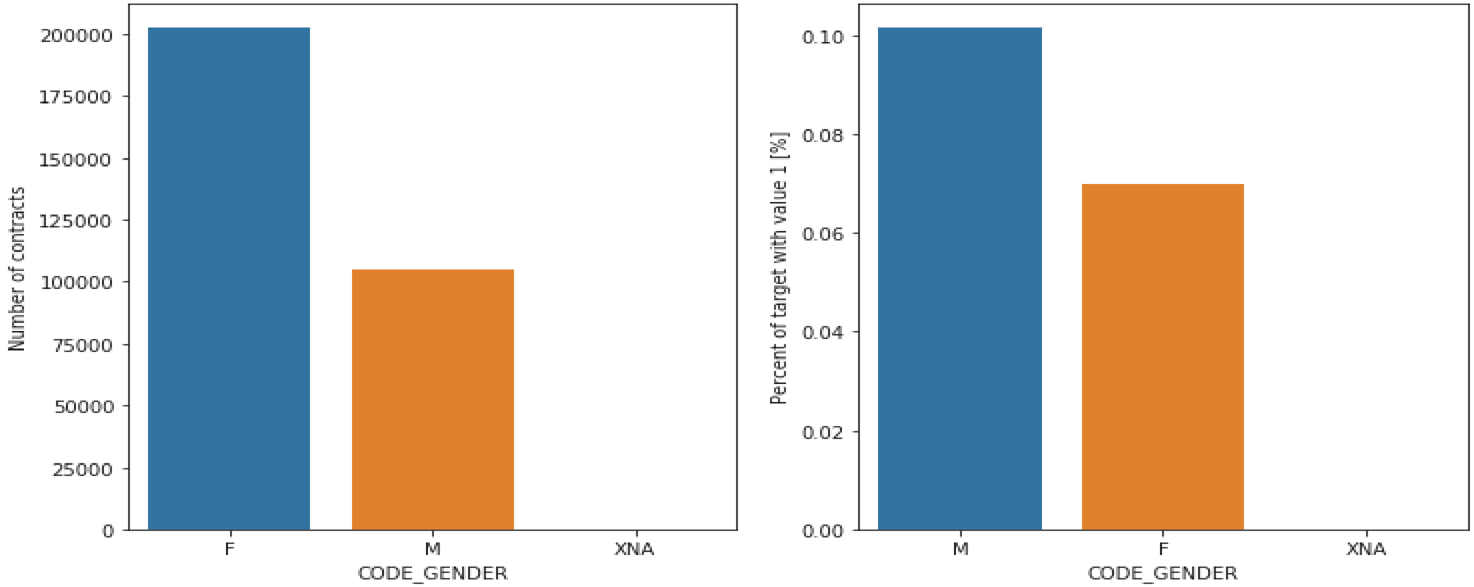
Checking how each column is affecting the client and finding which client is more likely to default and fall under category Target 1 by doing **univariate** and **segmented univariate** analysis.

```
In [53]: plot_stats('NAME_CONTRACT_TYPE')
```

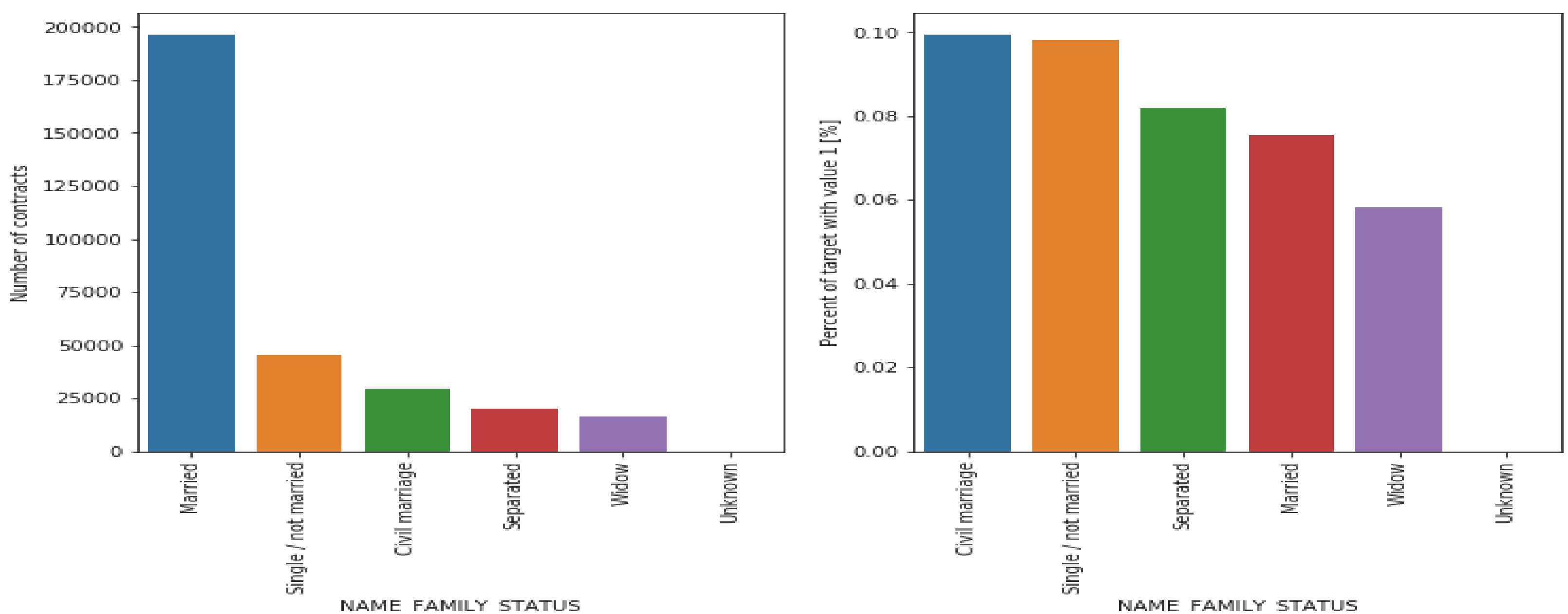


First bar plot shows how many Cash loans and Revolving loans are there in NAME_CONTRACT_TYPE
Second bar plot shows Cash loans and Revolving loans are behaving percentage wise in Target 1 category

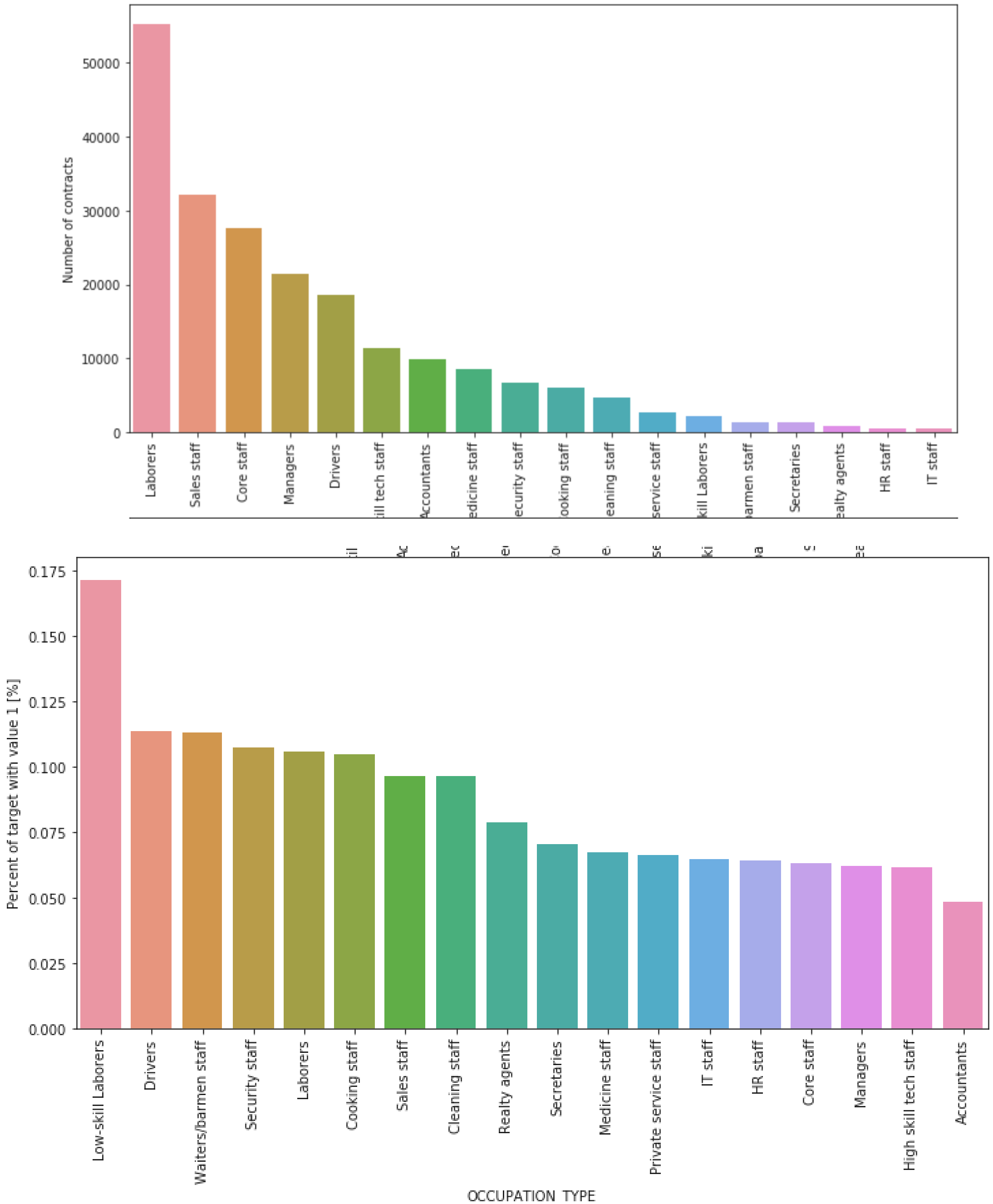
In [54]:
```
plot_stats('CODE_GENDER')
```



In [56]:
```
plot_stats('NAME_FAMILY_STATUS',True, True)
```



In [58]:
```
plot_stats('OCCUPATION_TYPE',True, False)
```



The graphs show the different Occupation Type of the clients.

- Conclusion

- Contract type Revolving loans are less when compared to the Revolving loans
- The number of female clients are more, almost double the number of male clients. The percentage of males have a higher chance of not returning their loans '10%' when compared with women '7%'.
- The number of clients having cars were half than the number of clients dont have cars. Both of them having 8% chance of not returning the lone amount.
- The clients that owns real estate are more than double of the ones that doesn't own real estate. Both categories have not-repayment rates less than '8%'.
- Among all the income types Pensioner, State Servent '6%' having more chance of not repaying the lone, when compared to other income types
- Majority of the clients have Secondary/secondary special education, followed by Higher education clients. The people with Academic degree have less than '2%' not-repayment rate. The Lower secondary category have the largest rate of not returning the loan '11%'.
- Most of clients are married, followed by Single/not married and civil marriage. Civil marriage has the highest percent of not repayment '10%', with Widow having the lowest percent of not repaying the lone '6%'.
- The people who are staying in Rented apartment and With parents has higher chance '10%' of not-repaying the lones.
- Laborers taken more number of lones, followed by Sales staff. The category with highest percent of not repaid loans are Low-skill Laborers '17%', followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.
- Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%) are the organisations which does not pay the lone amount.
- Clients with family size of 11 and 13 have not payed the lones at all. Families having 10 or 8 members having the percentage of not repaying of loans is over 30%. Families with 6 or less members have repayment rates close to the 10%.

Analysing Previous_application.csv file

Load the data

```
previous_data.head()
```

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_APPR_PROCESS_START	HOUR_APPR_PROCESS_START	...
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0	17145.0	SATURDAY	15	..
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN	607500.0	THURSDAY	11	..
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN	112500.0	TUESDAY	11	..
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN	450000.0	MONDAY	7	..
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN	337500.0	THURSDAY	9	..

5 rows × 37 columns

Merging two data sets for analysis :

```
[ ] data=application_data.merge(previous_data, left_on='SK_ID_CURR', right_on='SK_ID_CURR', how='inner')
```

```
[ ] data.head()
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE_x	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT_x	AMT_ANNUITY_x	...	NAME_SELLER_INDUSTRY	CNT
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.5	24700.5	...	Auto technology	
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	...	XNA	
2	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	...	Furniture	
3	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6750.0	...	Connectivity	
4	100006	0	Cash loans	F	N	Y	0	135000.0	312682.5	29686.5	...	XNA	

5 rows × 158 columns

Loading the data is the first step of any analysis and merging it with application_data.csv file

Check the data

- Glimpse the data and deleting unnecessary columns

```
[ ] data.head()
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE_x	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT_x	AMT_ANNUITY_x	...	NAME_SELLER_INDUSTRY	CNT
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.5	24700.5	...	Auto technology	
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	...	XNA	
2	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	...	Furniture	
3	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6750.0	...	Connectivity	
4	100006	0	Cash loans	F	N	Y	0	135000.0	312682.5	29686.5	...	XNA	

5 rows × 158 columns

Dropping unnecessary columns :

```
[ ] data.drop(['FLAG_DOCUMENT_2','FLAG_DOCUMENT_3','FLAG_DOCUMENT_4','FLAG_DOCUMENT_5',
            'FLAG_DOCUMENT_6','FLAG_DOCUMENT_7','FLAG_DOCUMENT_8','FLAG_DOCUMENT_9','FLAG_DOCUMENT_10',
            'FLAG_DOCUMENT_11','FLAG_DOCUMENT_12','FLAG_DOCUMENT_13','FLAG_DOCUMENT_14','FLAG_DOCUMENT_15',
            'FLAG_DOCUMENT_16','FLAG_DOCUMENT_17','FLAG_DOCUMENT_18','FLAG_DOCUMENT_19','FLAG_DOCUMENT_20',
            'FLAG_DOCUMENT_21','FLAG_EMP_PHONE','OBS_30_CNT_SOCIAL_CIRCLE','OBS_60_CNT_SOCIAL_CIRCLE',
             'DEF_30_CNT_SOCIAL_CIRCLE','DEF_60_CNT_SOCIAL_CIRCLE','YEARS_BEGINEXPLUATATION_MODE',
             'YEARS_BEGINEXPLUATATION_MEDI','EXT_SOURCE_1','EXT_SOURCE_2','EXT_SOURCE_3','APARTMENTS_AVG',
             'BASEMENTAREA_AVG','YEARS_BEGINEXPLUATATION_AVG','YEARS_BUILD_AVG','COMMONAREA_AVG','ELEVATORS_AVG',
             'ENTRANCES_AVG','FLOORSMAX_AVG','FLOORSMIN_AVG','LANDAREA_AVG',
             'LIVINGAPARTMENTS_AVG','LIVINGAREA_AVG','NONLIVINGAPARTMENTS_AVG','NONLIVINGAREA_AVG',
             'APARTMENTS_MODE','BASEMENTAREA_MODE','YEARS_BUILD_MODE','COMMONAREA_MODE',
             'ELEVATORS_MODE','ENTRANCES_MODE','FLOORSMAX_MODE','FLOORSMIN_MODE',
             'LANDAREA_MODE','LIVINGAPARTMENTS_MODE','LIVINGAREA_MODE','NONLIVINGAPARTMENTS_MODE',
             'NONLIVINGAREA_MODE','APARTMENTS_MEDI','BASEMENTAREA_MEDI','YEARS_BUILD_MEDI','COMMONAREA_MEDI','ELEVATORS_MEDI','ENTRANCES_MEDI'
             'FLOORSMAX_MEDI','FLOORSMIN_MEDI','LANDAREA_MEDI','LIVINGAPARTMENTS_MEDI','LIVINGAREA_MEDI','NONLIVINGAPARTMENTS_MEDI',
             'NONLIVINGAREA_MEDI','FONDKAPREMONT_MODE','HOUSETYPE_MODE','TOTALAREA_MODE','WALLSMATERIAL_MODE',
             'EMERGENCYSTATE_MODE','FLAG_MOBIL', 'FLAG_WORK_PHONE',
             'FLAG_CONT_MOBILE', 'FLAG_PHONE', 'FLAG_EMAIL','REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY',],axis=1,inplace=True)
```

- Check missing data

Finding out columns which are having null values more than 50% and dropping :

```
[ ]  total = data.isnull().sum().sort_values(ascending = False)
     percent = (data.isnull().sum()/data.isnull().count()*100).sort_values(ascending = False)
     missing_data=pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
     missing_data
```

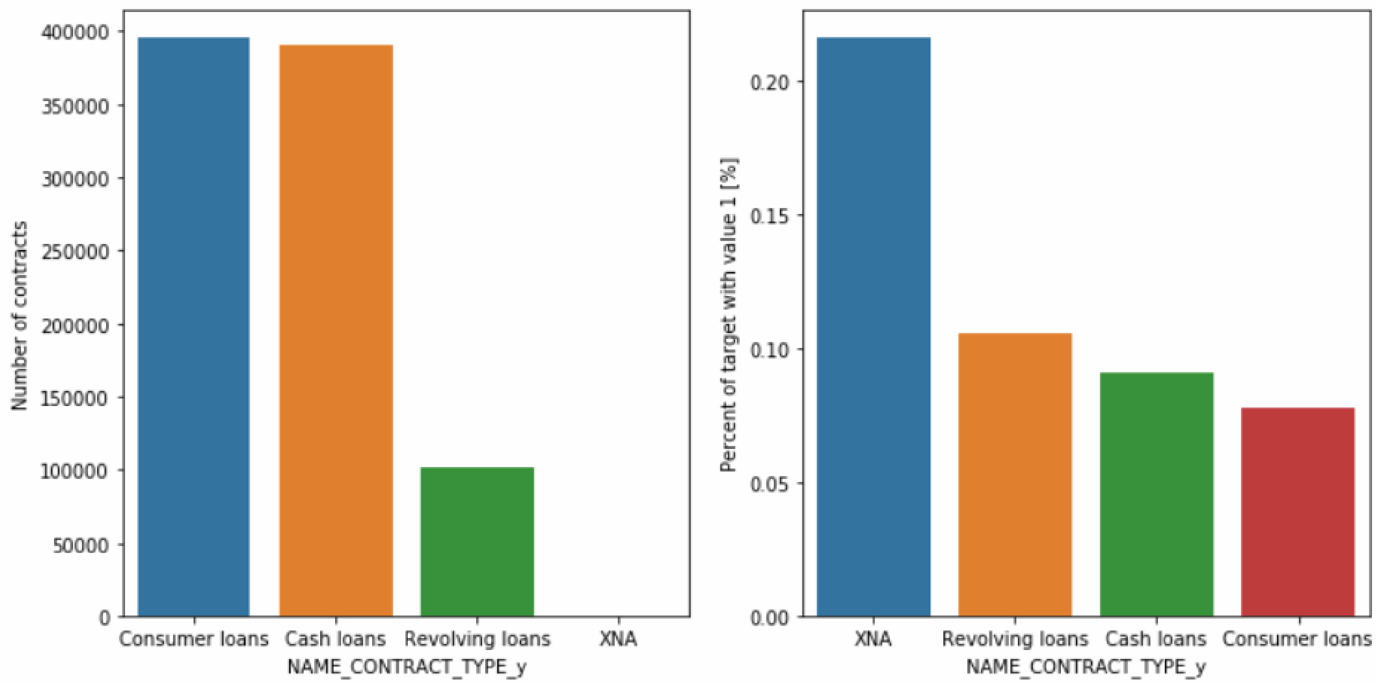
```
[ ]  df=data[missing_data[missing_data['Percent']>=50].index]
     data.drop(df,axis=1,inplace=True)
```

Checking the total null values and percent of null values column wise

Explore the data

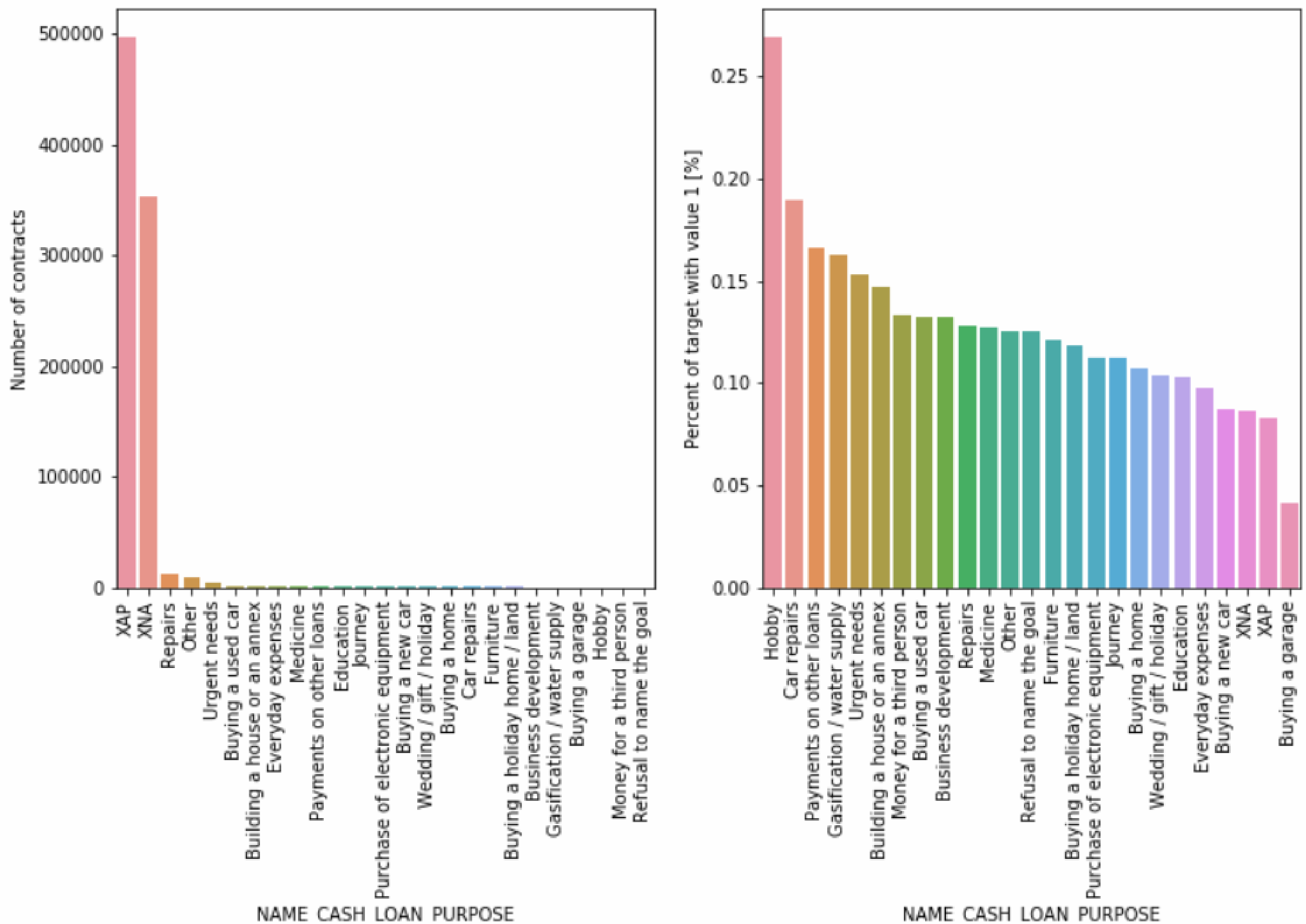
- Analysing the data column wise

```
graph('NAME_CONTRACT_TYPE_y')
```



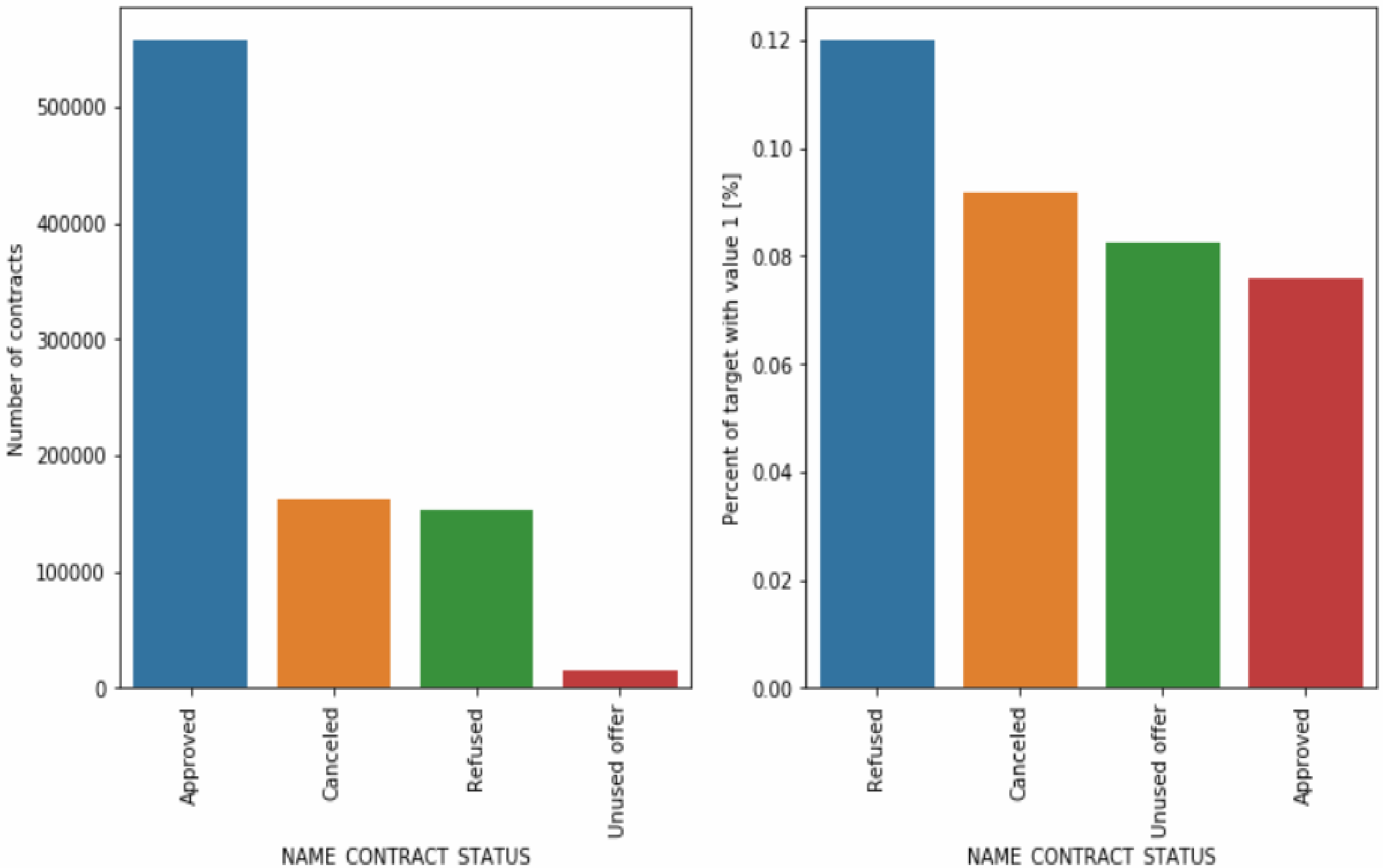
- Among the three types of contracts i.e., Cash loans, Consumer loans, Revolving loans. Cash loans and Consumer loans are almost the same '600K' and Revolving loans are almost equal to '150K'.
- The percent of defaults in Revolving loans is 10% and then 9.5% for Cash loans and 8% for Consumer loans.

```
graph('NAME_CASH_LOAN_PURPOSE', True, True)
```



- The percent of defaults for cash loan are more by the name of Hobby which is 2*% followed by Car repairs 19% and Payments on other loans 17%

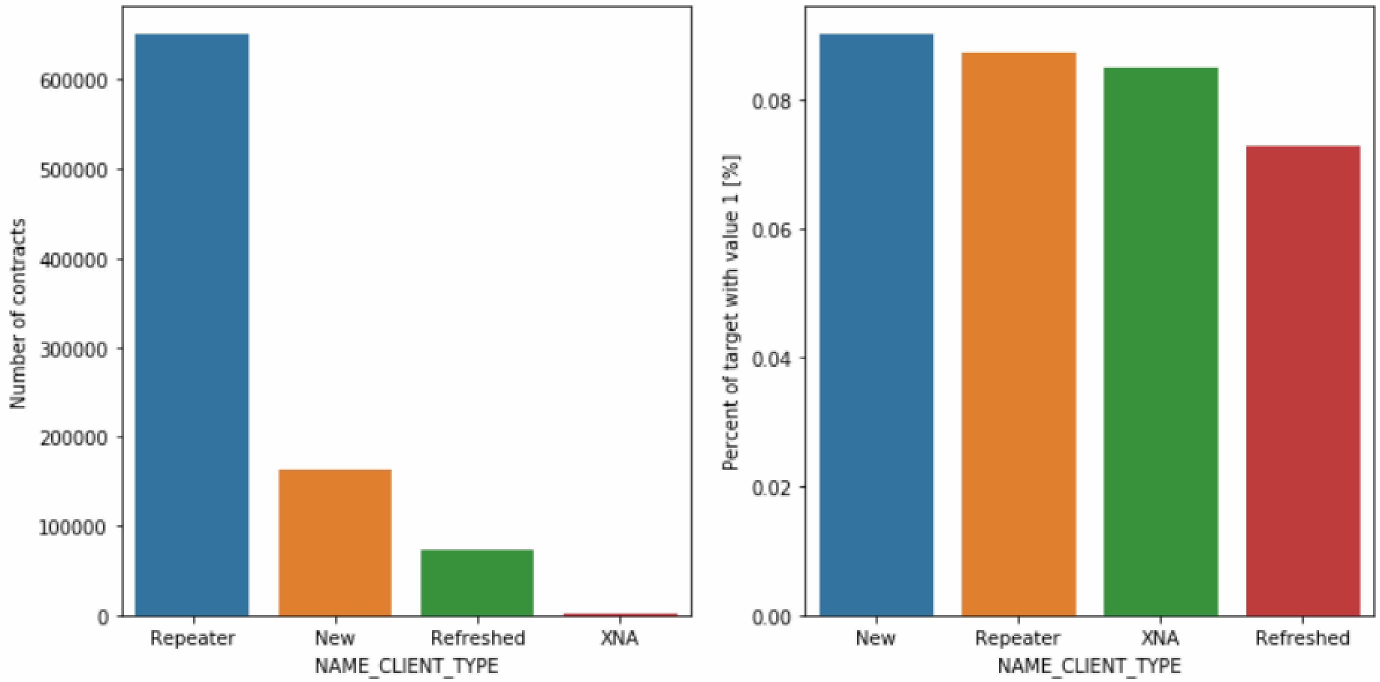
```
graph('NAME_CONTRACT_STATUS', True, True)
```



Most previous applications contract status are Approved 850K, Canceled and Refused status is 240K. There are only 20K in Unused offer.

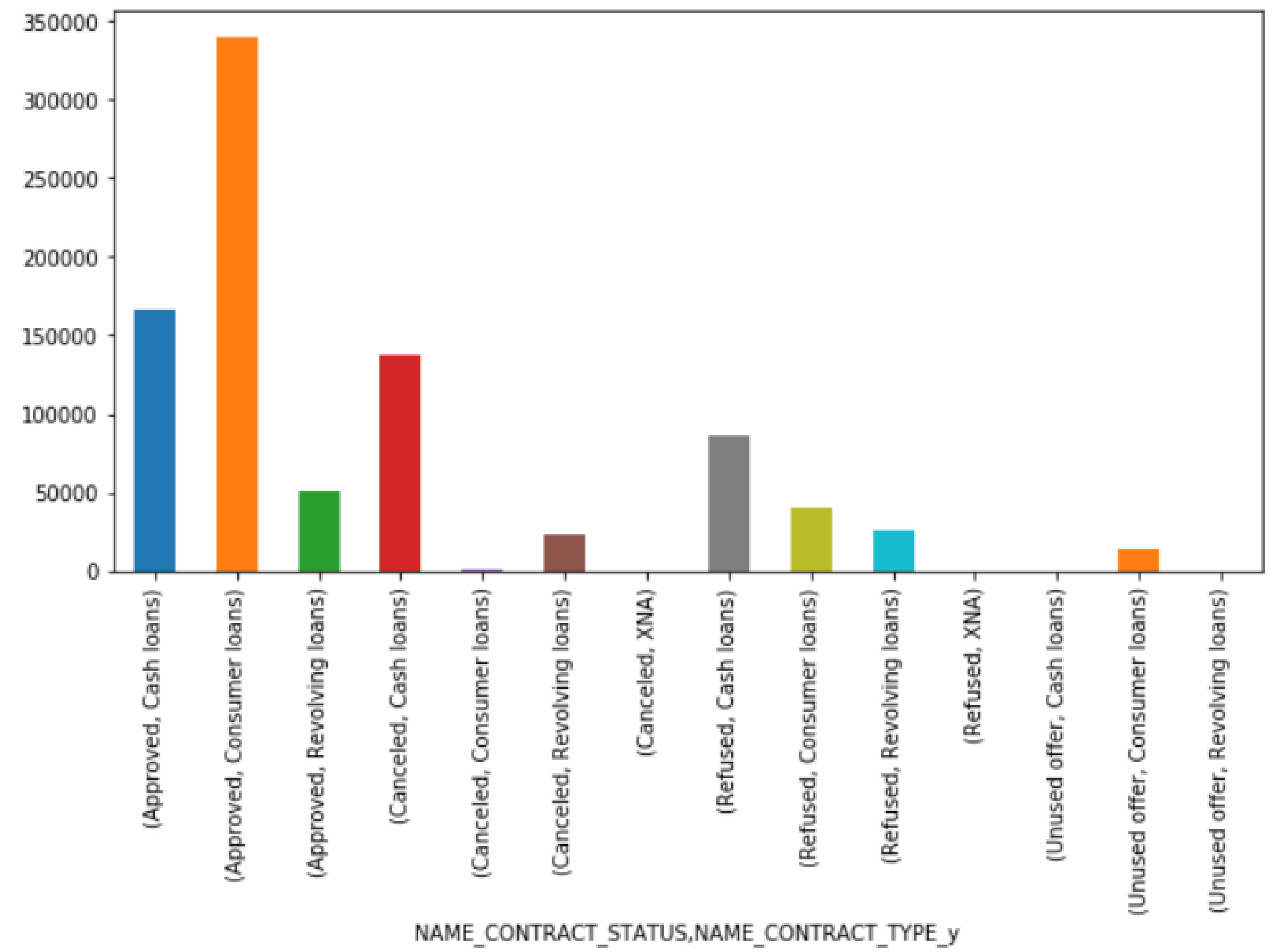
In terms of percent of defaults for current applications, clients with history of previous applications have largest percents of defaults when in their history contract statuses are Refused 12%, followed by Canceled 9%, Unused offer 8% and then Approved which is less than 8%.

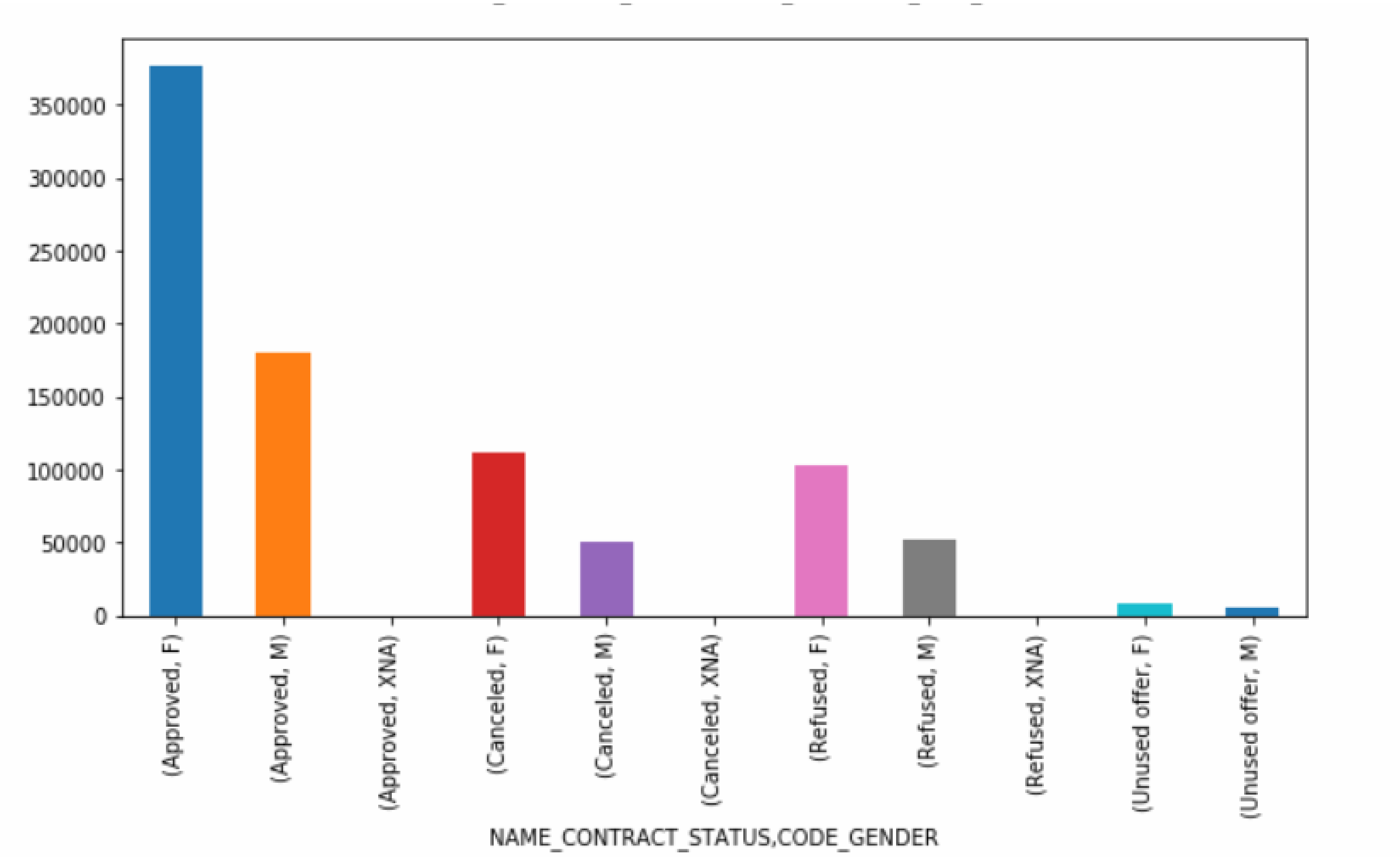
```
graph('NAME_CLIENT_TYPE')
```



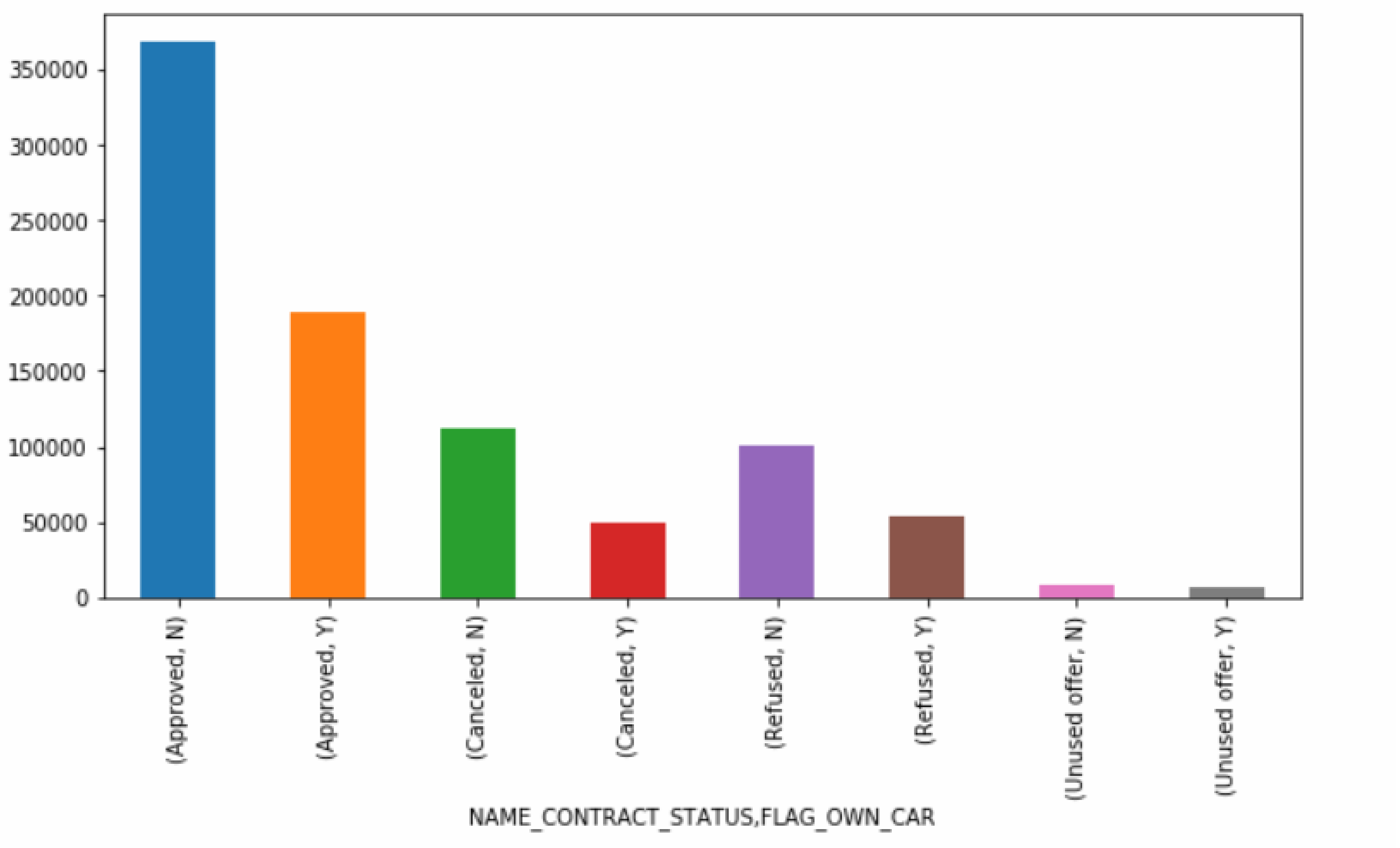
- Most of the previous application data have client type of Repeater(1M), 200K are New and nearly 100K are Refreshed.
- In terms of default the percentage of current applications of clients having defaults ranging from from 8.5%, 8.25% and 7% for New, Repeater and Refreshed respectivelly.

- **Finding out to reduce the Defaulters**

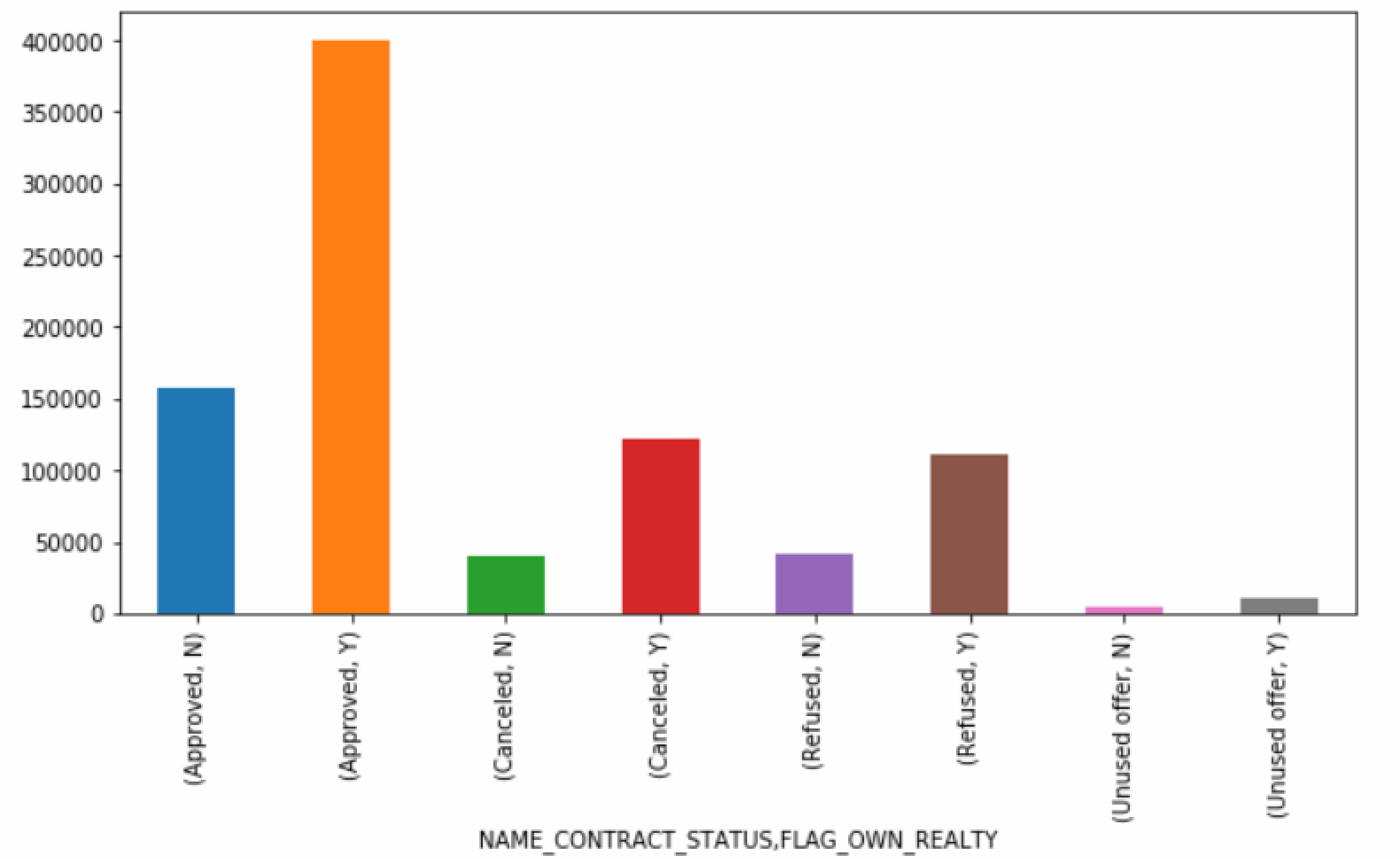


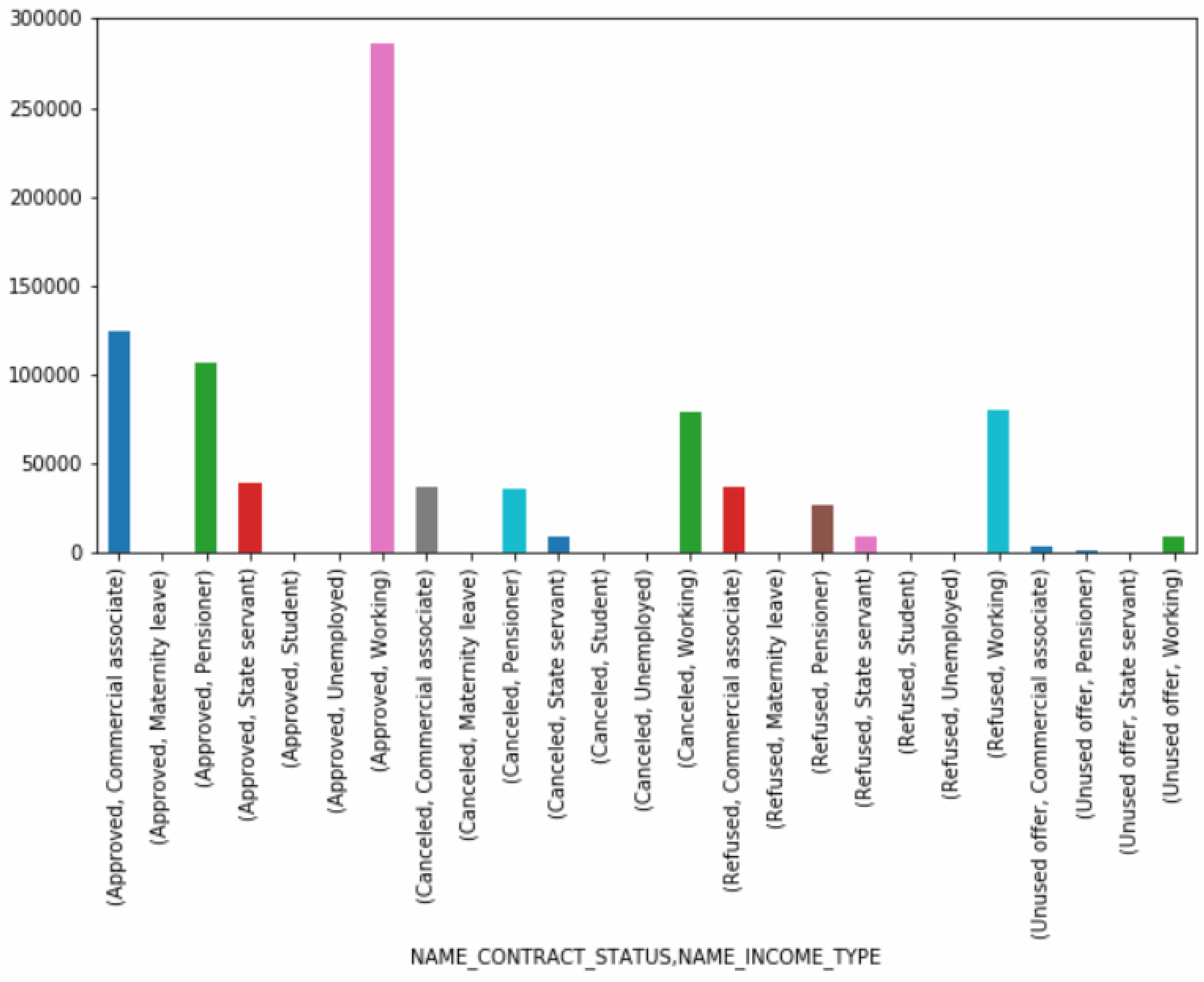


350000
300000
250000
200000
150000
100000
50000
0
(Approved, F)
(Approved, M)
(Approved, XNA)
(Canceled, F)
(Canceled, M)
(Canceled, XNA)
(Refused, F)
(Refused, M)
(Refused, XNA)
(Unused offer, F)
(Unused offer, M)
NAME_CONTRACT_STATUS,CODE_GENDER

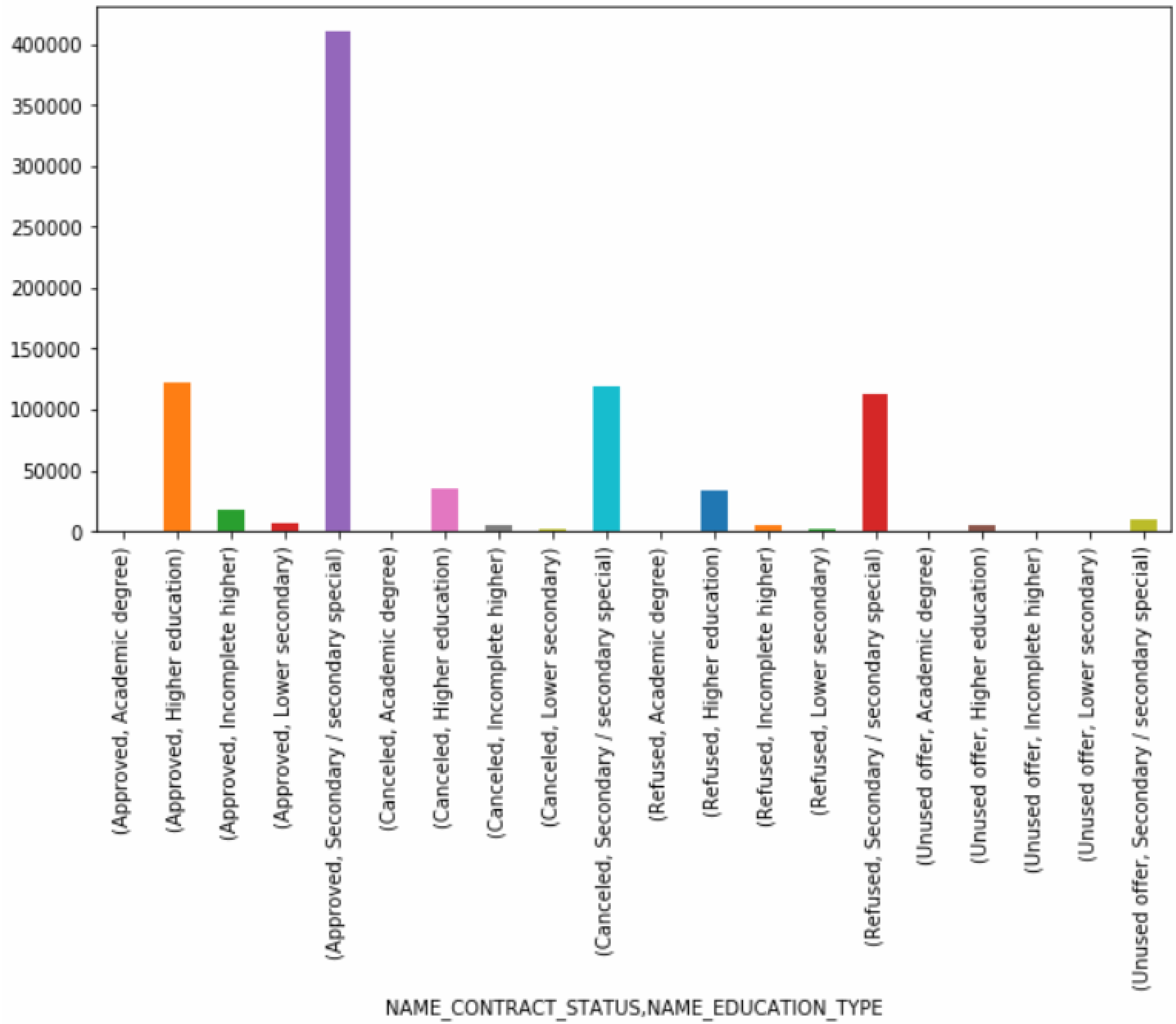


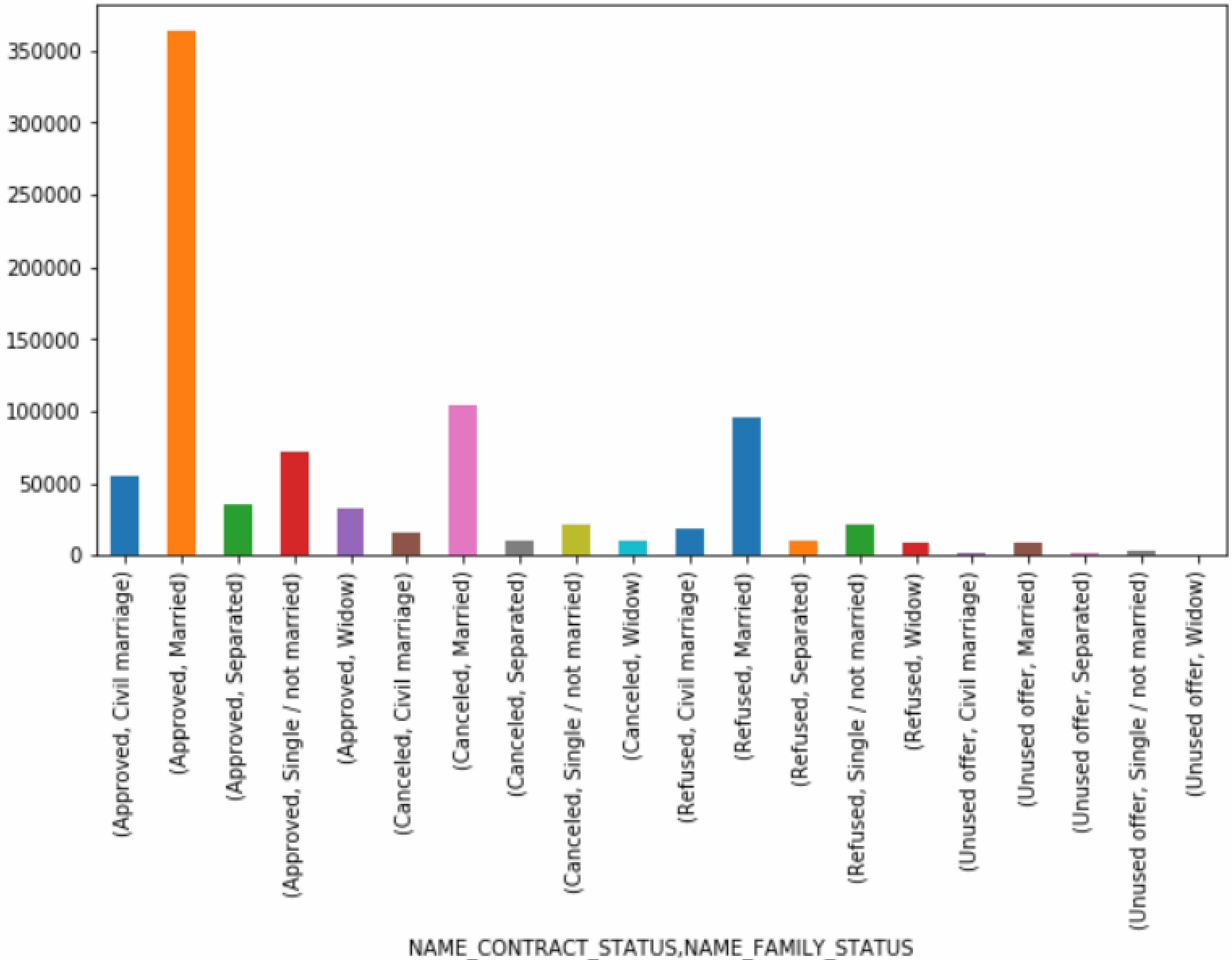
350000
300000
250000
200000
150000
100000
50000
0
(Approved, N)
(Approved, Y)
(Canceled, N)
(Canceled, Y)
(Refused, N)
(Refused, Y)
(Unused offer, N)
(Unused offer, Y)
NAME_CONTRACT_STATUS,FLAG_OWN_CAR


400000
350000
300000
250000
200000
150000
100000
50000
0
(Approved, N)
(Approved, Y)
(Canceled, N)
(Canceled, Y)
(Refused, N)
(Refused, Y)
(Unused offer, N)
(Unused offer, Y)
NAME_CONTRACT_STATUS,FLAG_OWN_REALTY

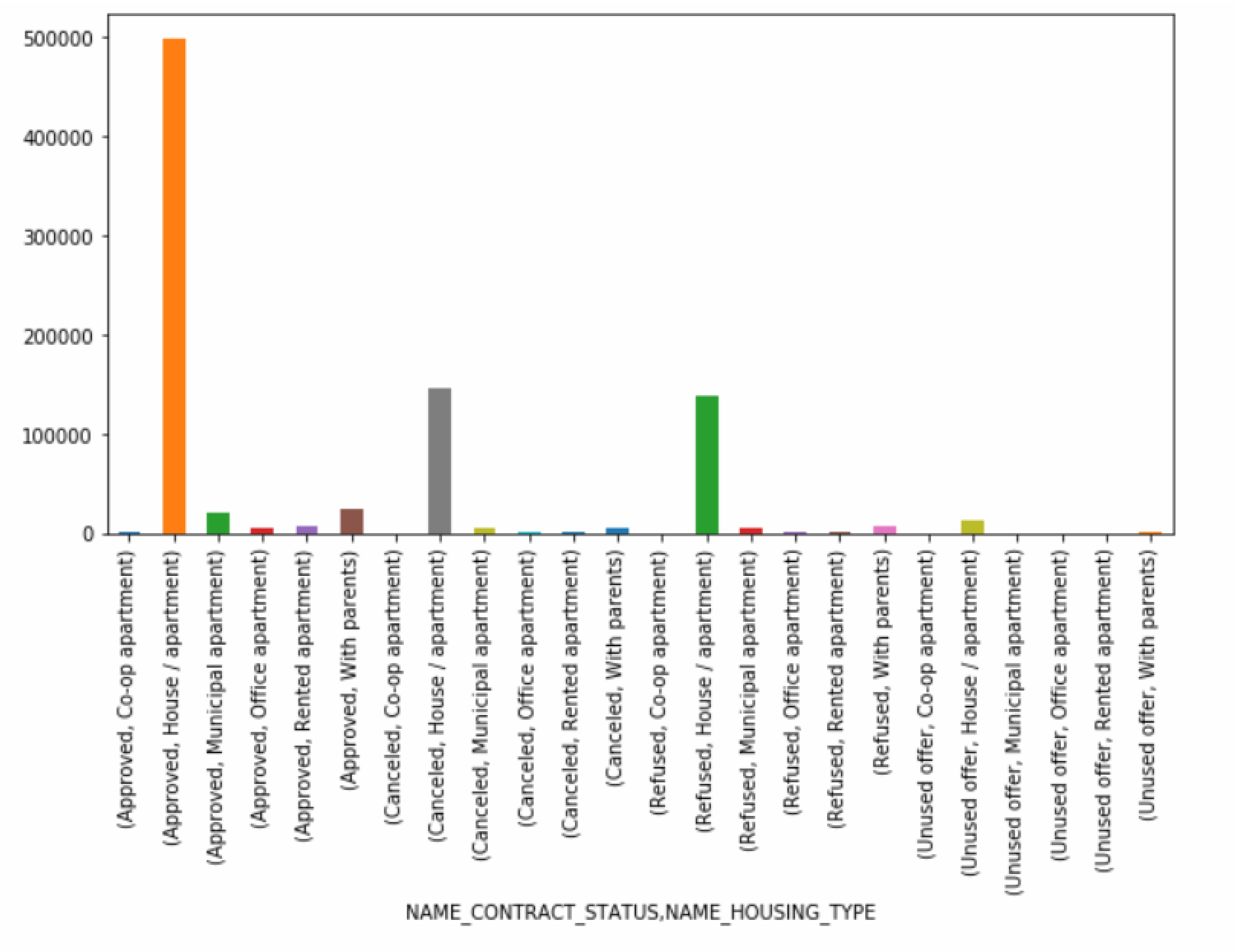


300000
250000
200000
150000
100000
50000
0
(Approved, Commercial associate)
(Approved, Maternity leave)
(Approved, Pensioner)
(Approved, State servant)
(Approved, Student)
(Approved, Unemployed)
(Approved, Working)
(Canceled, Commercial associate)
(Canceled, Maternity leave)
(Canceled, Pensioner)
(Canceled, State servant)
(Canceled, Student)
(Canceled, Unemployed)
(Canceled, Working)
(Refused, Commercial associate)
(Refused, Maternity leave)
(Refused, Pensioner)
(Refused, State servant)
(Refused, Student)
(Refused, Unemployed)
(Refused, Working)
(Unused offer, Commercial associate)
(Unused offer, Pensioner)
(Unused offer, State servant)
(Unused offer, Working)
NAME_CONTRACT_STATUS,NAME_INCOME_TYPE


400000
350000
300000
250000
200000
150000
100000
50000
0
(Approved, Academic degree)
(Approved, Higher education)
(Approved, Incomplete higher)
(Approved, Lower secondary)
(Approved, Secondary / secondary special)
(Canceled, Academic degree)
(Canceled, Higher education)
(Canceled, Incomplete higher)
(Canceled, Lower secondary)
(Canceled, Secondary / secondary special)
(Refused, Academic degree)
(Refused, Higher education)
(Refused, Incomplete higher)
(Refused, Lower secondary)
(Refused, Secondary / secondary special)
(Unused offer, Academic degree)
(Unused offer, Higher education)
(Unused offer, Incomplete higher)
(Unused offer, Lower secondary)
(Unused offer, Secondary / secondary special)
NAME_CONTRACT_STATUS,NAME_EDUCATION_TYPE



350000
300000
250000
200000
150000
100000
50000
0
(Approved, Civil marriage)
(Approved, Married)
(Approved, Separated)
(Approved, Single / not married)
(Approved, Widow)
(Canceled, Civil marriage)
(Canceled, Married)
(Canceled, Separated)
(Canceled, Single / not married)
(Canceled, Widow)
(Refused, Civil marriage)
(Refused, Married)
(Refused, Separated)
(Refused, Single / not married)
(Refused, Widow)
(Unused offer, Civil marriage)
(Unused offer, Married)
(Unused offer, Separated)
(Unused offer, Single / not married)
(Unused offer, Widow)
NAME_CONTRACT_STATUS,NAME_FAMILY_STATUS


500000
400000
300000
200000
100000
0
(Approved, Co-op apartment)
(Approved, House / apartment)
(Approved, Municipal apartment)
(Approved, Office apartment)
(Approved, Rented apartment)
(Approved, With parents)
(Canceled, Co-op apartment)
(Canceled, House / apartment)
(Canceled, Municipal apartment)
(Canceled, Office apartment)
(Canceled, Rented apartment)
(Canceled, With parents)
(Refused, Co-op apartment)
(Refused, House / apartment)
(Refused, Municipal apartment)
(Refused, Office apartment)
(Refused, Rented apartment)
(Refused, With parents)
(Unused offer, Co-op apartment)
(Unused offer, House / apartment)
(Unused offer, Municipal apartment)
(Unused offer, Office apartment)
(Unused offer, Rented apartment)
(Unused offer, With parents)
NAME_CONTRACT_STATUS,NAME_HOUSING_TYPE

- Conclusion

The following are the cases where there are High approval of loans having more defaulters

- The client having Secondary special, Higher education as education
- The client having Family Status as Married
- Whose Occupation is Labour, Core staff and Sales staff

- By reducing the loans approved for the above category of people there might be a change of having less defaulters which helps the organisation to gain profit