Credit EDA Case Study

Submitted By:
- Aakashnidhi Prasad
- Sumit Kumar

“ Understanding how the bank deal with loans and how consumer attributes and loan attributes influence the tendency of defaults. Finding out various patterns and representing the results to help the bank reduce the credit risk and interest risk.

Flow Of Exploratory Data Analysis

DATA SOURCING → DATA CLEANING → ANALYSIS

DATA SOURCING
- *application_data.csv*
- *previous_application.csv*

DATA CLEANING
- Handling the Nulls
- Checking the Outliers

ANALYSIS
- Univariate
- Bivariate and Multivariate

Application Data Analysis

Data Understanding and Cleaning

- Checked the sample data from this dataset
- Shape of the Dataset: **(307511, 122)**
- No. of columns with more than 30% nulls: **30** (Removed all these columns)
- Imputed the mode values to the null values of categorical columns wherever required

```
app_data.head()
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREI
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

```
app_data.shape
```

```
(307511, 122)
```

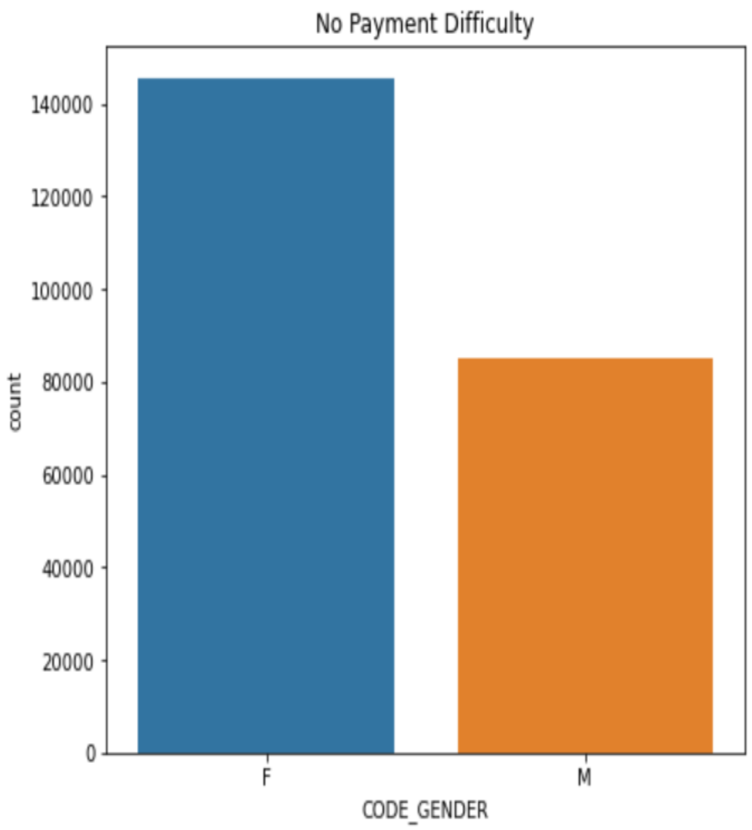
```
# Dropping all the columns having more than 30% null values in app_data
app_data_final = app_data.drop(labels=list(emptycol.index), axis = 1)
```

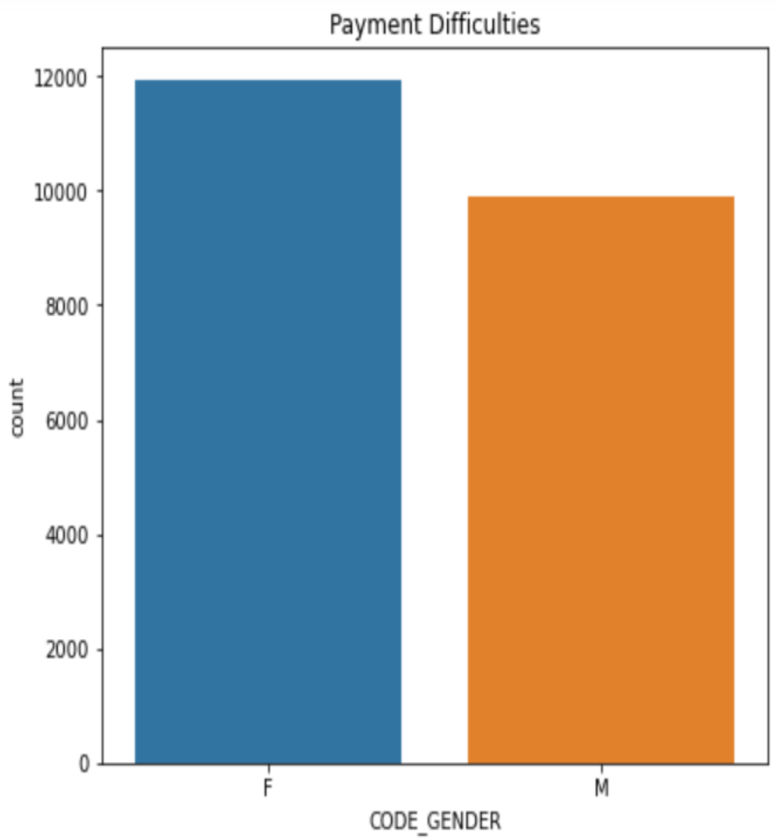
```
# We have XNA in gender which means null value, so we are replacing it with F as mode for f is more and so probablity of it being
app_data_final['CODE_GENDER'] = app_data_final['CODE_GENDER'].apply(lambda x: 'F' if x == 'XNA' else x)
```

```
# Getting column info of final app dataset
app_data_final.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 72 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   SK_ID_CURR                  307511 non-null  int64
 1   TARGET                      307511 non-null  int64
 2   NAME_CONTRACT_TYPE          307511 non-null  object
 3   CODE_GENDER                 307511 non-null  object
 4   FLAG_OWN_CAR                307511 non-null  object
 5   FLAG_OWN_REALTY             307511 non-null  object
 6   CNT_CHILDREN                307511 non-null  int64
 7   AMT_INCOME_TOTAL            307511 non-null  float64
 8   AMT_CREDIT                  307511 non-null  float64
 9   AMT_ANNUITY                 307499 non-null  float64
 10  AMT_GOODS_PRICE             307233 non-null  float64
 11  NAME_TYPE_SUITE             306219 non-null  object
 12  NAME_INCOME_TYPE            307511 non-null  object
 13  NAME_EDUCATION_TYPE         307511 non-null  object
 14  NAME_FAMILY_STATUS          307511 non-null  object
 15  NAME_HOUSING_TYPE           307511 non-null  object
 16  REGION_POPULATION_RELATIVE  307511 non-null  float64
```

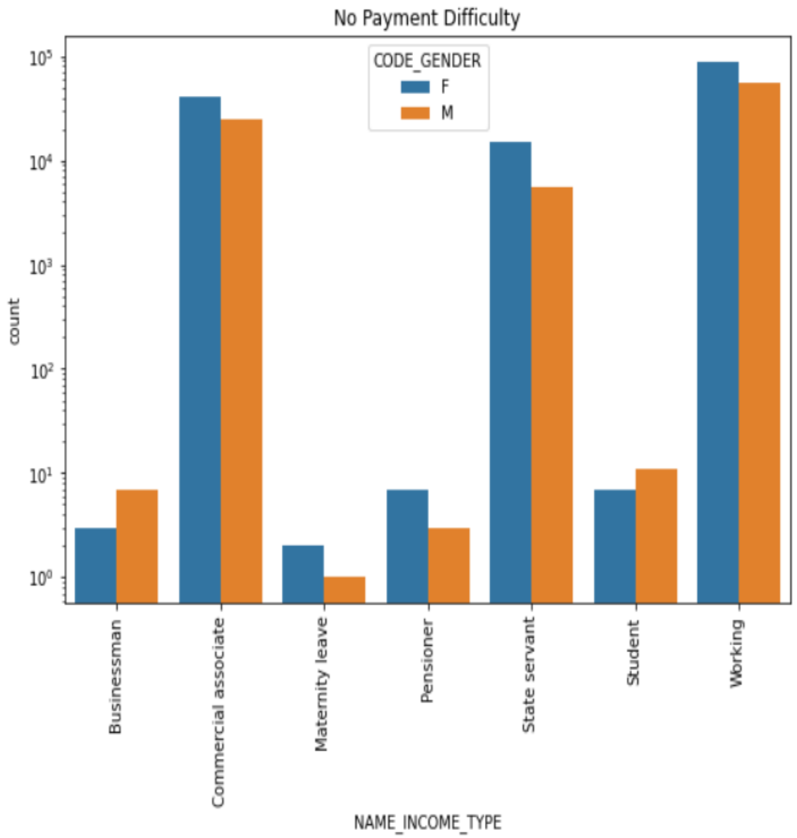
Gender Distribution

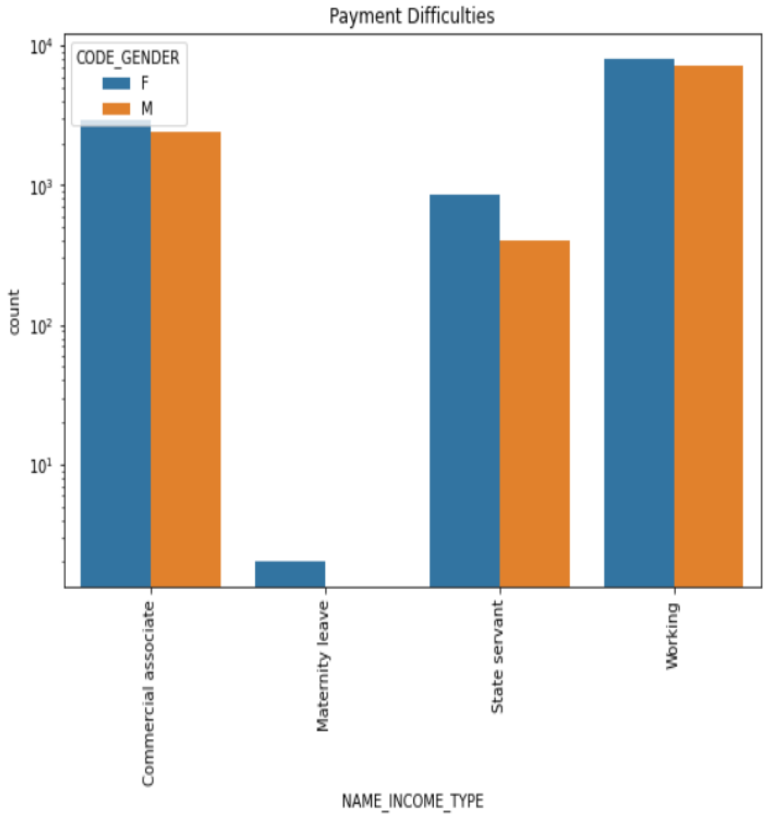




- Females have more credit than Males. Ratio wise Males are facing more difficulties in paying back the loan.

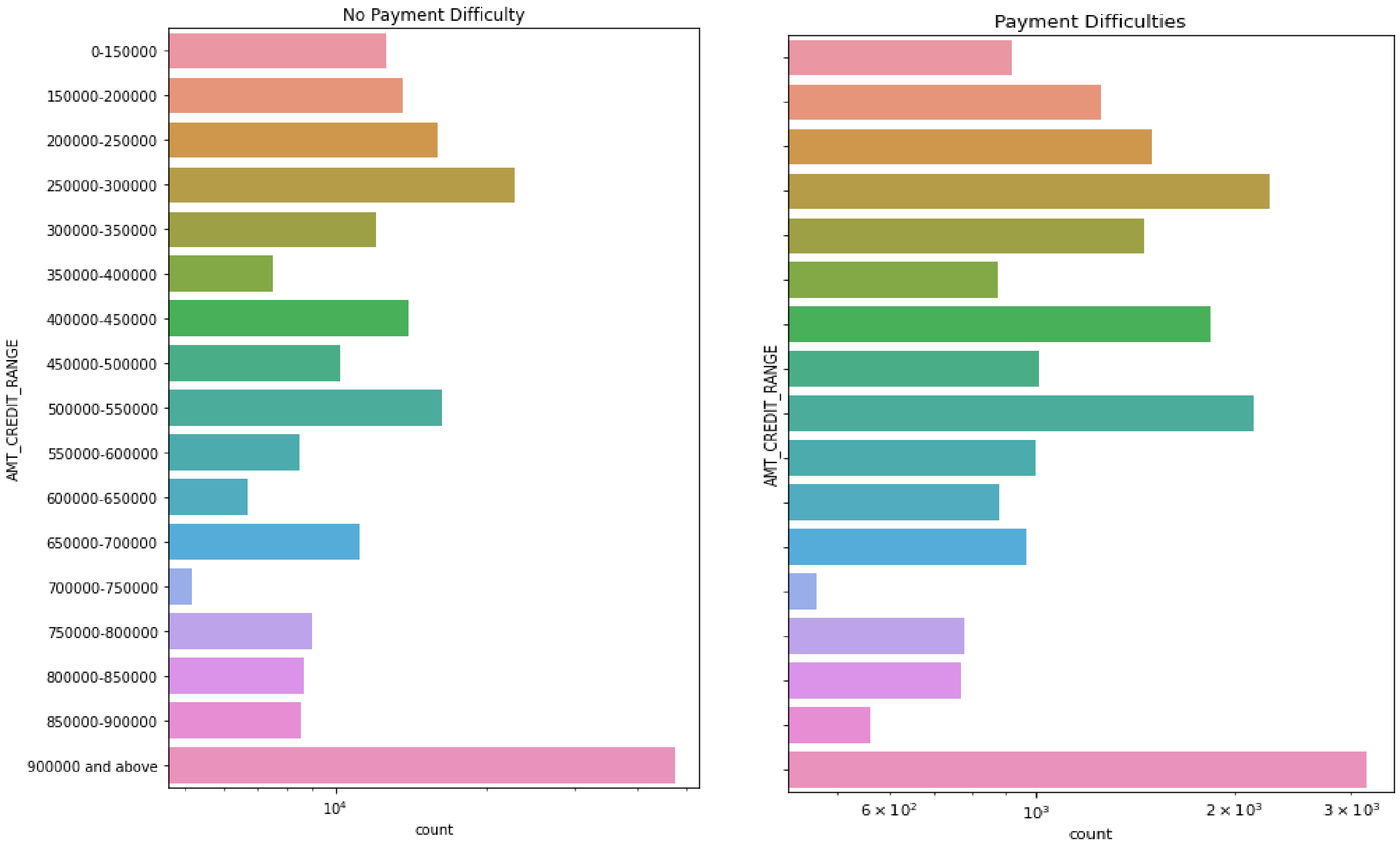
Occupation, Gender and payment Difficulty





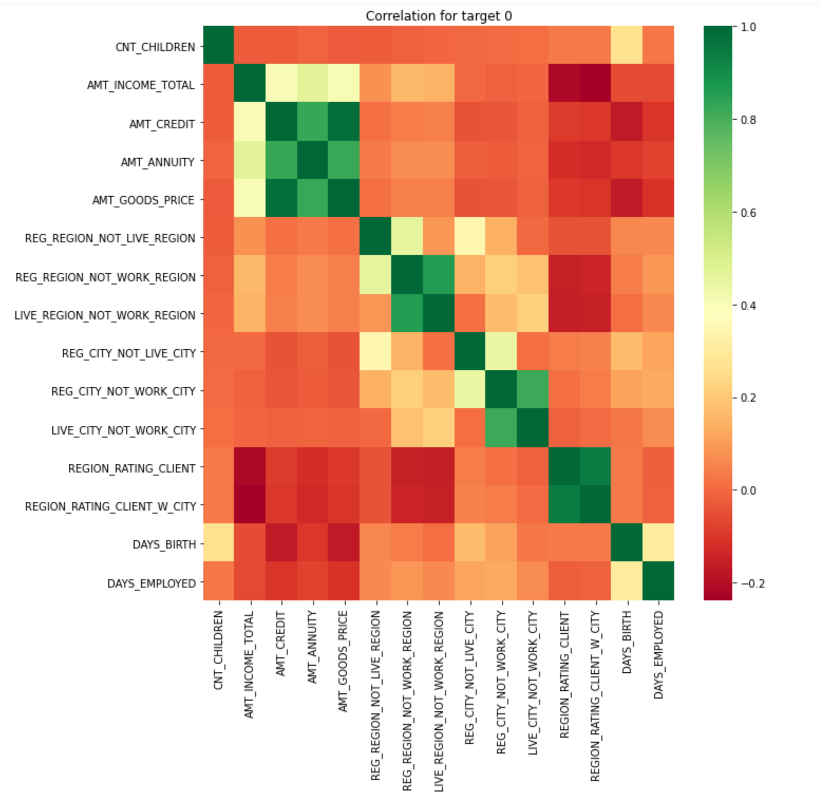
- State servant, Working and Commercial associate have much more credits then others.

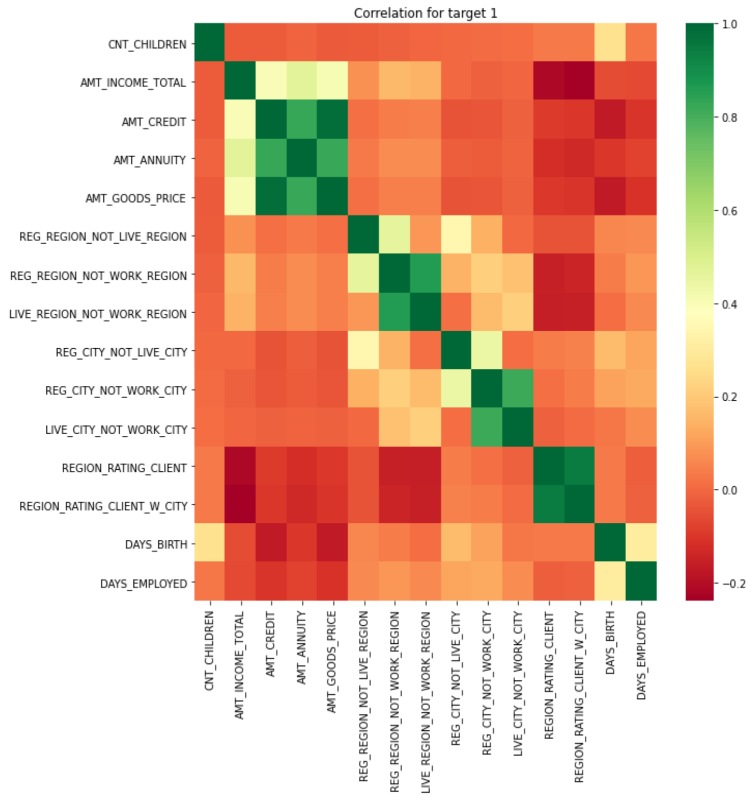
AMOUNT CREDIT AND PAYMENT DIFFICULTY



- Most people take credit in range of 900000 and above. Applicants with higher credit amount has higher default rate.

Correlation between Numerical variables





- Total Income is having high negitive correlation with region_client_rating i.e. if income is high, city rating is on lower side and vise versa.

Previous Application Data Analysis

Data Understanding and Cleaning

- Checked the sample data from this dataset
- Shape of the Dataset: **(1670214, 37)**
- Removed columns with more than 20% nulls
- Imputed the mode values to the null values of categorical columns wherever required

```
#head of the data frame.
prev_app.head()
```

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKI
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0	17145.0	
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN	607500.0	
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN	112500.0	
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN	450000.0	
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN	337500.0	

```
# checking the nu
prev_app.shape
```

```
(1670214, 37)
```

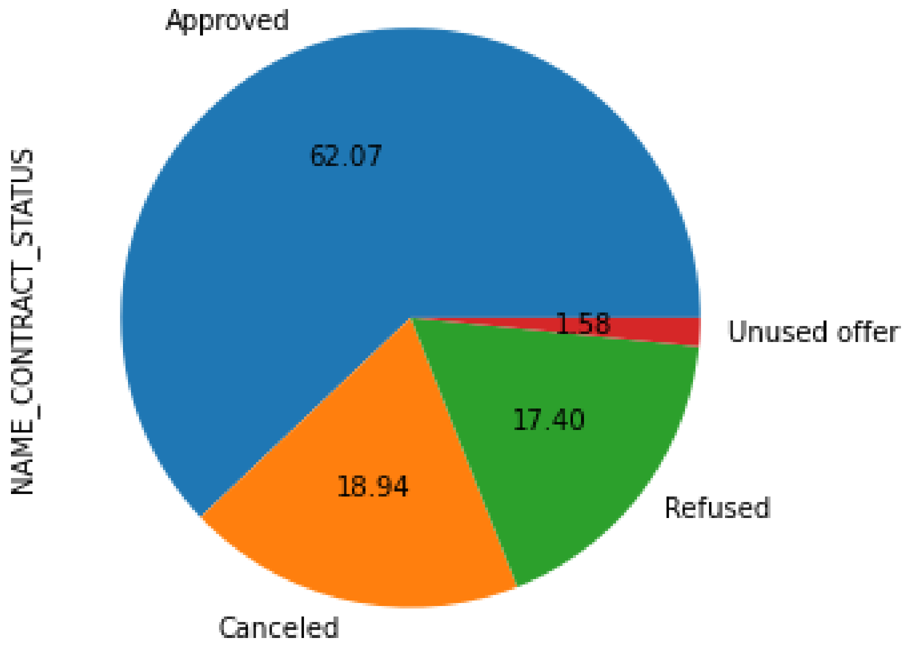
```
# Changing 'XNA' and 'XAP' to NaN
prev_app.loc[prev_app.NAME_CONTRACT_TYPE.isin(['XNA','XAP']),"NAME_CONTRACT_TYPE"]=np.NaN
prev_app.loc[prev_app.NAME_CASH_LOAN_PURPOSE.isin(['XNA','XAP']),"NAME_CASH_LOAN_PURPOSE"]=np.NaN
prev_app.loc[prev_app.NAME_PAYMENT_TYPE.isin(['XNA','XAP']),"NAME_PAYMENT_TYPE"]=np.NaN
prev_app.loc[prev_app.CODE_REJECT_REASON.isin(['XNA','XAP']),"CODE_REJECT_REASON"]=np.NaN
prev_app.loc[prev_app.NAME_CLIENT_TYPE.isin(['XNA','XAP']),"NAME_CLIENT_TYPE"]=np.NaN
prev_app.loc[prev_app.NAME_GOODS_CATEGORY.isin(['XNA','XAP']),"NAME_GOODS_CATEGORY"]=np.NaN
prev_app.loc[prev_app.NAME_PORTFOLIO.isin(['XNA','XAP']),"NAME_PORTFOLIO"]=np.NaN
prev_app.loc[prev_app.NAME_PRODUCT_TYPE.isin(['XNA','XAP']),"NAME_PRODUCT_TYPE"]=np.NaN
prev_app.loc[prev_app.NAME_SELLER_INDUSTRY.isin(['XNA','XAP']),"NAME_SELLER_INDUSTRY"]=np.NaN
prev_app.loc[prev_app.NAME_YIELD_GROUP.isin(['XNA','XAP']),"NAME_YIELD_GROUP"]=np.NaN
```

```
#print the information of variables to check their data types.
prev_app.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
 #   Column                       Non-Null Count    Dtype
---  ------                       --------------    -----
 0   SK_ID_PREV                   1670214 non-null  int64
 1   SK_ID_CURR                   1670214 non-null  int64
 2   NAME_CONTRACT_TYPE           1670214 non-null  object
 3   AMT_ANNUITY                  1297979 non-null  float64
 4   AMT_APPLICATION              1670214 non-null  float64
 5   AMT_CREDIT                   1670213 non-null  float64
 6   AMT_DOWN_PAYMENT             774370 non-null   float64
 7   AMT_GOODS_PRICE              1284699 non-null  float64
 8   WEEKDAY_APPR_PROCESS_START   1670214 non-null  object
 9   HOUR_APPR_PROCESS_START      1670214 non-null  int64
 10  FLAG_LAST_APPL_PER_CONTRACT  1670214 non-null  object
 11  NFLAG_LAST_APPL_IN_DAY       1670214 non-null  int64
 12  RATE_DOWN_PAYMENT            774370 non-null   float64
 13  RATE_INTEREST_PRIMARY        5951 non-null     float64
 14  RATE_INTEREST_PRIVILEGED     5951 non-null     float64
 15  NAME_CASH_LOAN_PURPOSE       1670214 non-null  object
 16  NAME_CONTRACT_STATUS         1670214 non-null  object
 17  DAYS_DECISION                1670214 non-null  int64
 18  NAME_PAYMENT_TYPE            1670214 non-null  object
 19  CODE_REJECT_REASON           1670214 non-null  object
 20  NAME_TYPE_SUITE              849809 non-null   object
 21  NAME_CLIENT_TYPE             1670214 non-null  object
 22  NAME_GOODS_CATEGORY          1670214 non-null  object
 23  NAME_PORTFOLIO               1670214 non-null  object
```

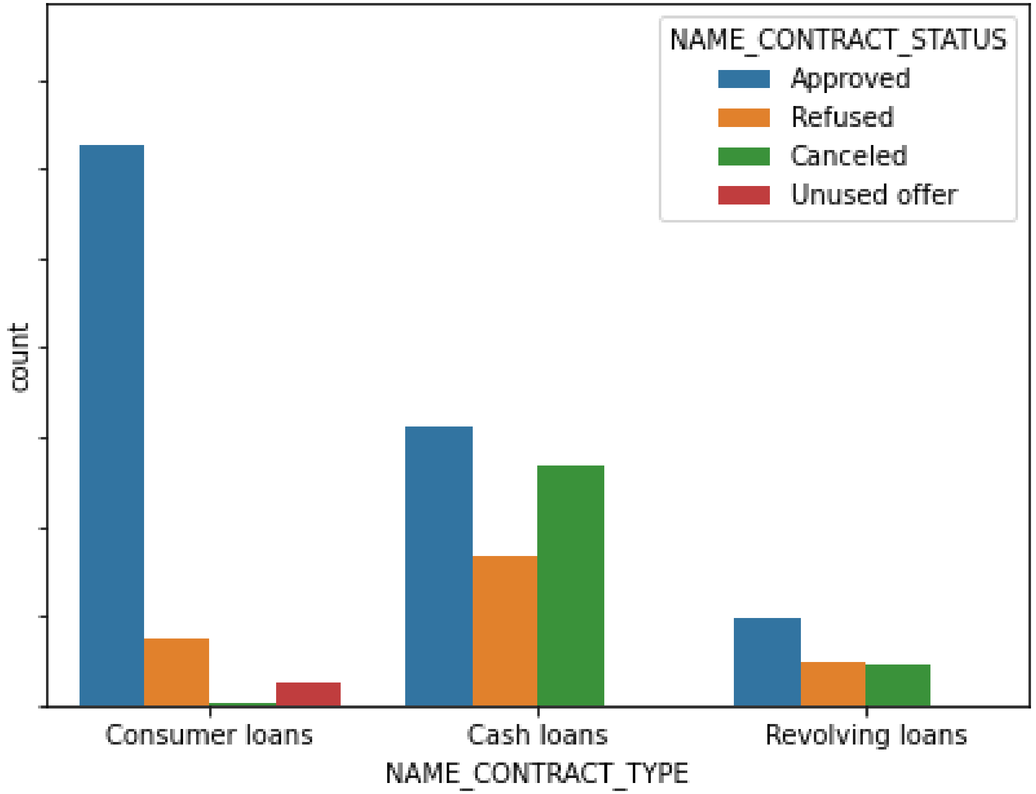
```
# Dropping columns with more than 20% of null values
prev_app = prev_app.loc[:,prev_app.isnull().mean()<=0.2]
prev_app.head()
```

Loan Status



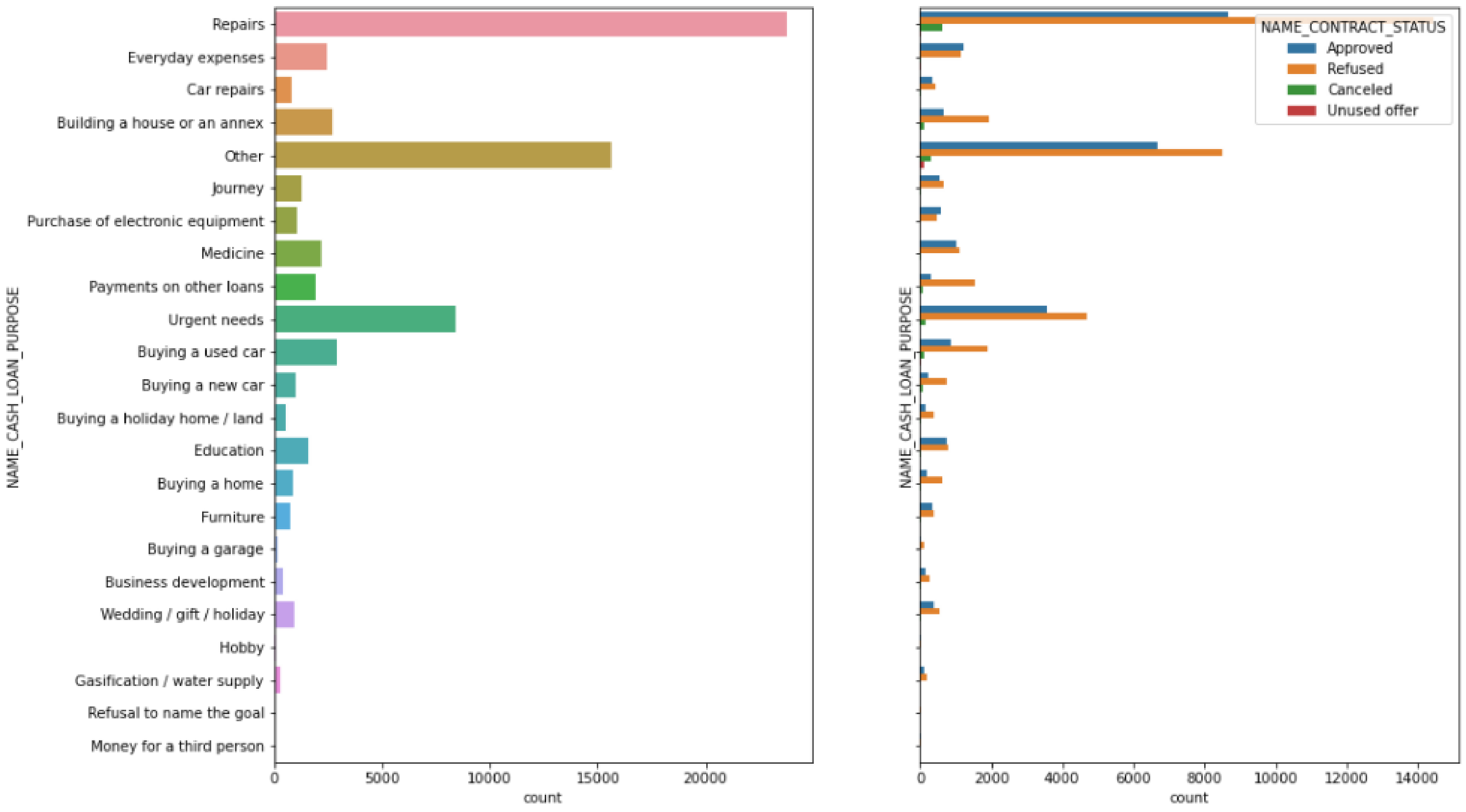
- Approximately 62% of previous loan applications got approved while less than 18% were refused.

Loan type vs loan status



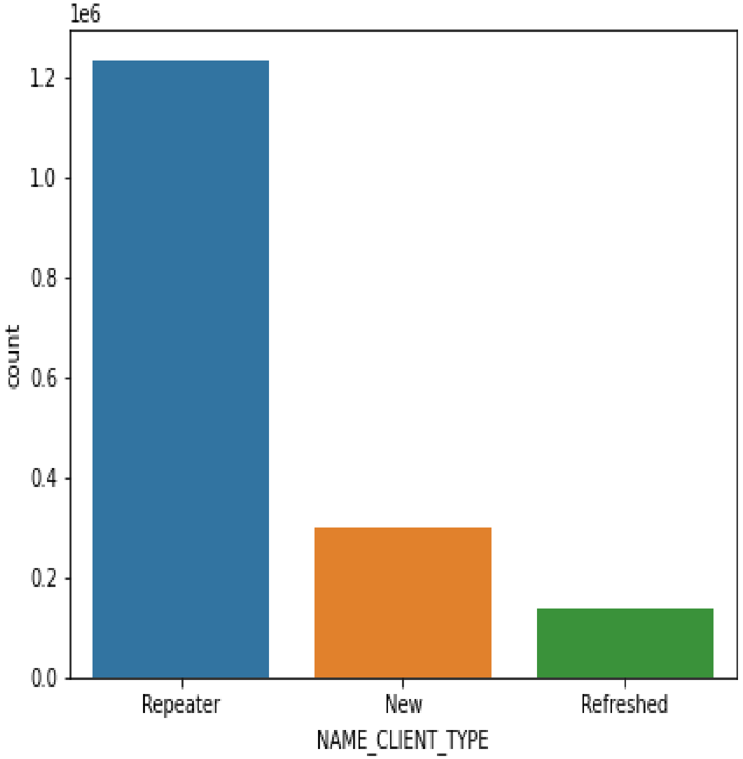
- Majority of the applicants applied for 'Cash Loans' or 'Consumer Loans'. The approval rate was best for 'Consumer Loans'.

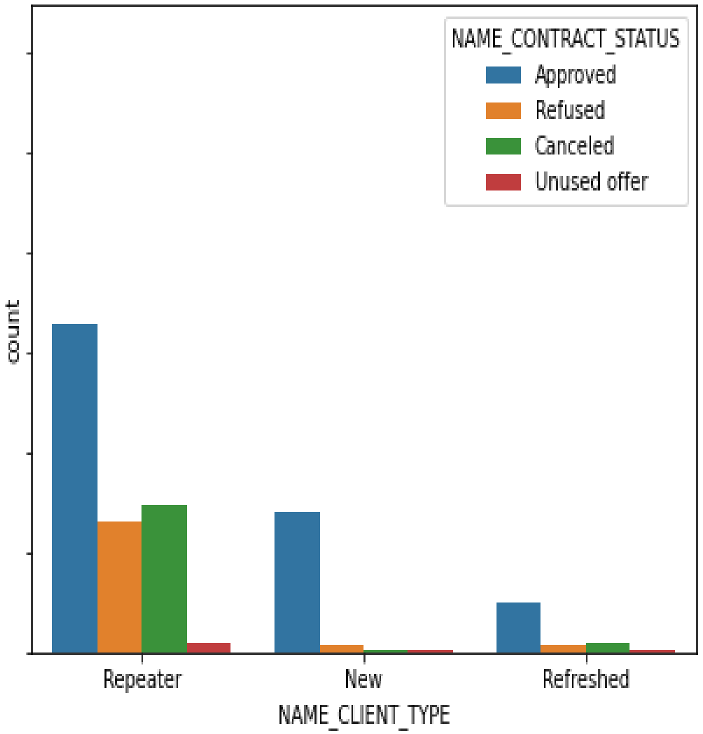
Loan purpose vs status



- Majority of the loan applications were for 'Repair' work.

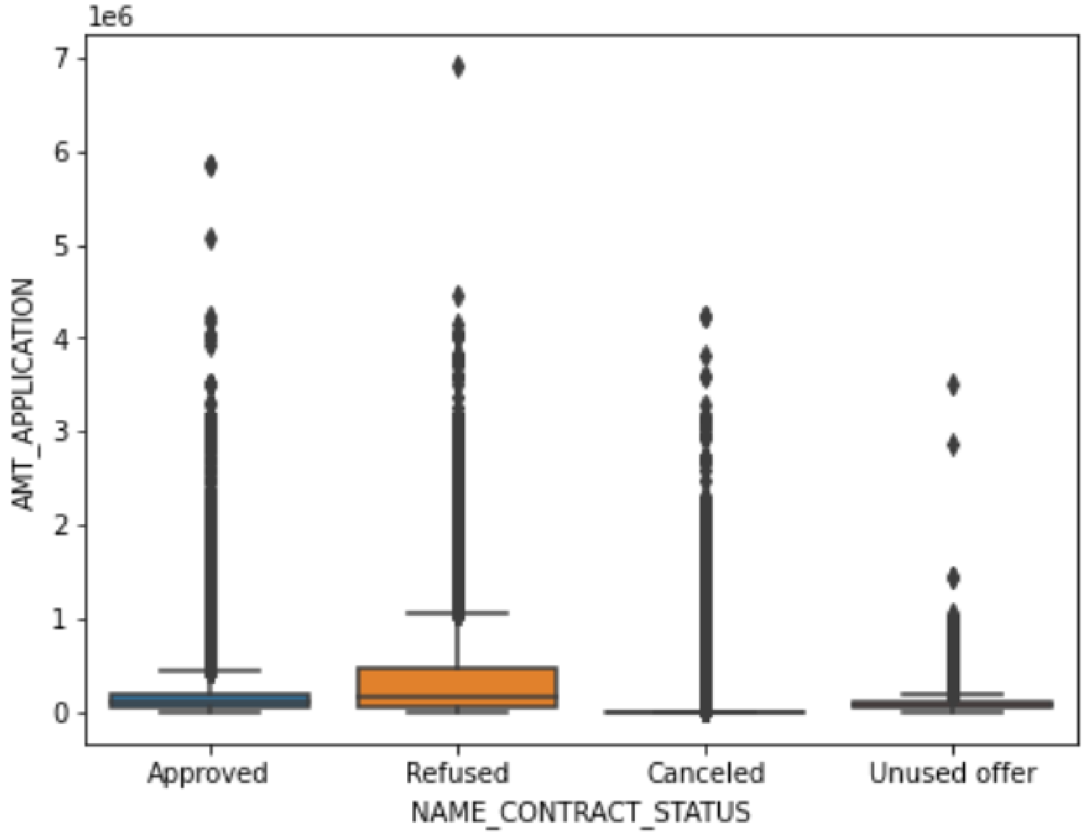
Applicant type vs loan status





- More than 70% of applications were from 'Repeaters'. 'New' applicant group has the best approval rate.

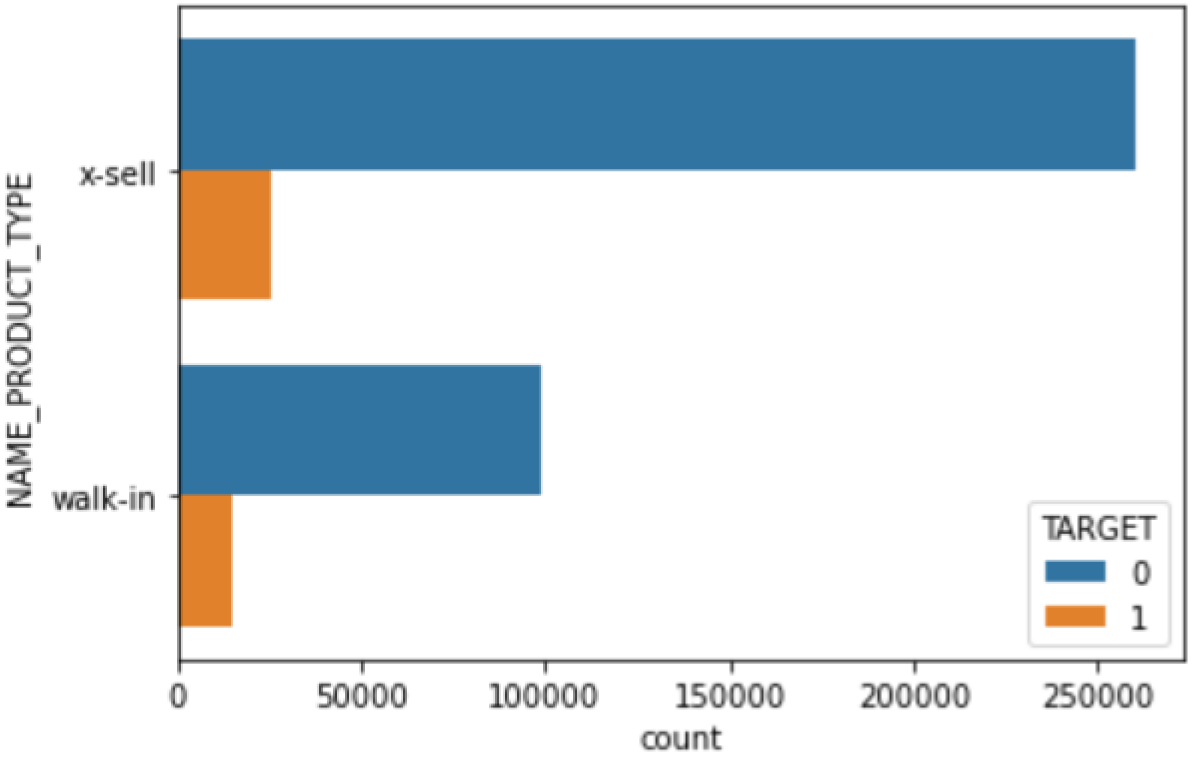
Application amount vs loan status



- Applications with higher loan application amount are likely to be refused. Also, low credit amount are very likely to be cancelled by the applicants.

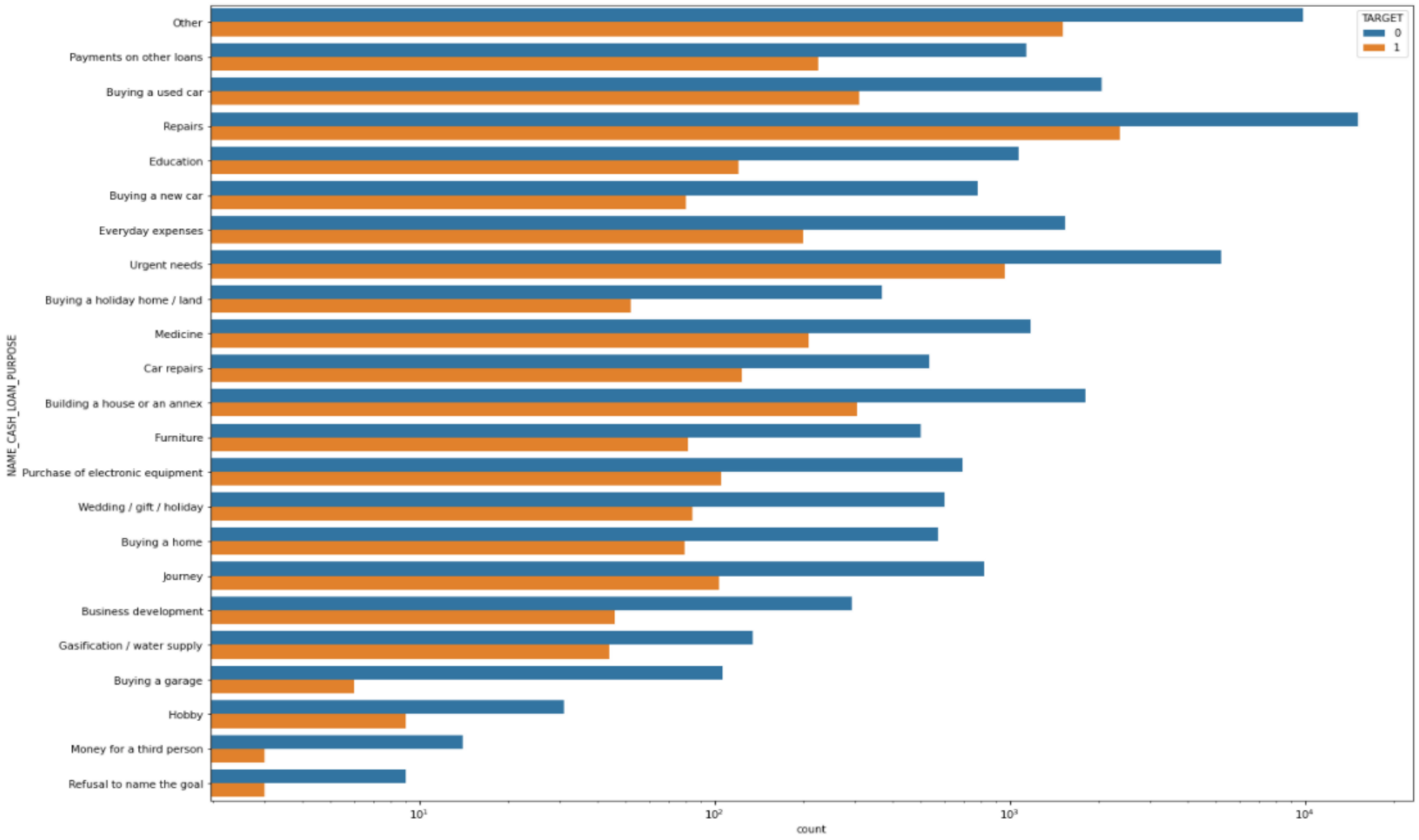
Merged dataset Analysis

Product type vs target



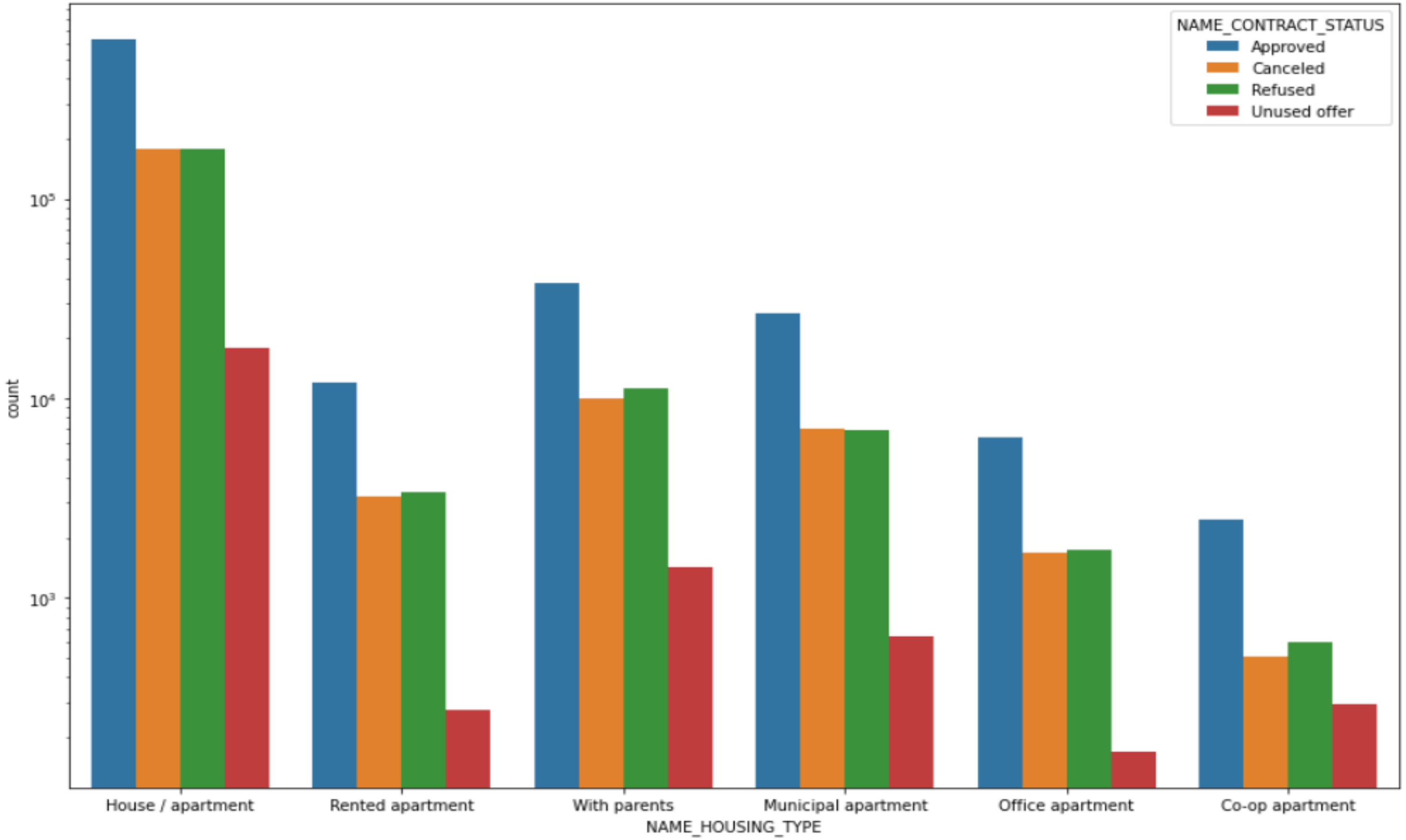
- 'x-sell' product type has significantly lower default rate when compared to 'walk-in'.

Loan purpose vs target



- Loan purposes with 'Repairs' are facing more difficulties in repayment of loan on time.

Housing type vs previous loan status



- People having their own house/apartment have maximum number of applications along with good approval rate

Banks should focus more on Female applicants as they have lower payment difficulty rate than Males.

Banks should focus more on 'Businessmen', 'Students', 'Pensioner' as they don't have any payment difficulties.

Banks should focus less on Business Entity Type 3 and should focus more on Government employees.

Banks should continue focussing on 'x-sell' as it has low default rate.

Also they should focus less on 'Repairs' specific loans as we saw they've faced more difficulty in repayment on time.


THANK YOU