Credit Case Study

Narendra Chinchalapu(EDS21050195)
Shubham Kokate(EDS21050382)

Business Objective

This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

Datasets

1. '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.

2. '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.

3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Application Data

Using pandas library we read csv for application data.

We can observe that there 307511 records and 122 features out of which 65 are float64, 41 are int64 and 16 are Objects.

We can store the numeric and categorical columns separately for further analysis.

```
[6] application_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

```
application_data.shape

(307511, 122)
```

Previous Application Data

For previous applications data we have a huge volume of records at 1670214 with 37 features out of which 15 are float64, 6 are int64 and 16 are objects.

We further need to clean both the datasets if nulls are present.

```
previous_application_data.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
 24  NAME_PRODUCT_TYPE            1670214 non-null  object
 25  CHANNEL_TYPE                 1670214 non-null  object
 26  SELLERPLACE_AREA             1670214 non-null  int64
 27  NAME_SELLER_INDUSTRY         1670214 non-null  object
 28  CNT_PAYMENT                  1297984 non-null  float64
 29  NAME_YIELD_GROUP             1670214 non-null  object
 30  PRODUCT_COMBINATION          1669868 non-null  object
 31  DAYS_FIRST_DRAWING           997149 non-null   float64
 32  DAYS_FIRST_DUE               997149 non-null   float64
 33  DAYS_LAST_DUE_1ST_VERSION    997149 non-null   float64
 34  DAYS_LAST_DUE                997149 non-null   float64
 35  DAYS_TERMINATION             997149 non-null   float64
 36  NFLAG_INSURED_ON_APPROVAL    997149 non-null   float64
dtypes: float64(15), int64(6), object(16)
memory usage: 471.5+ MB
```

```
previous_application_data.shape
```

```
(1670214, 37)
```

Data Cleaning for Application Data

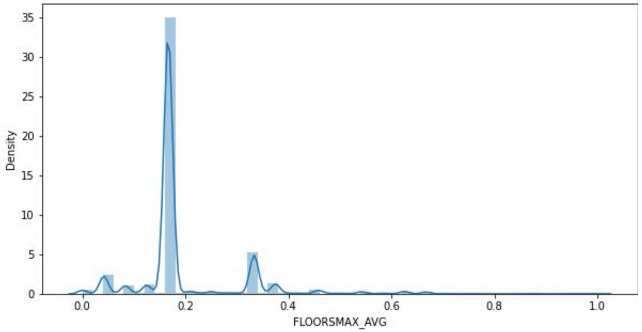
We calculated the percentage of missing values in each feature of application dataset. For handling these missing values we used following approach.

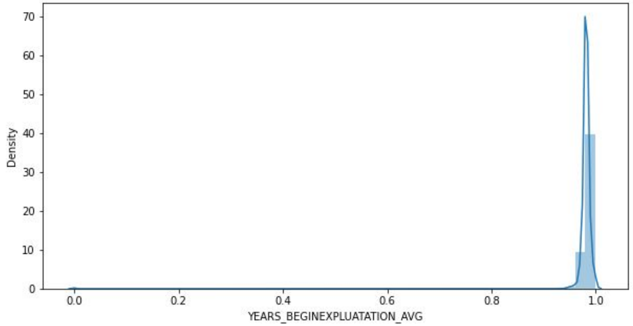
1. **Features with more than 50% missing values:** We can drop these features as most of the data is unavailable.
2. **Features with missing values between 50-10% :** We can impute data for these features with either median or mode depending upon the type of feature.
3. **Features with missing values less than 10% :** We can drop rows for these features as the amount of missing values is less thus is insignificant in analysis.

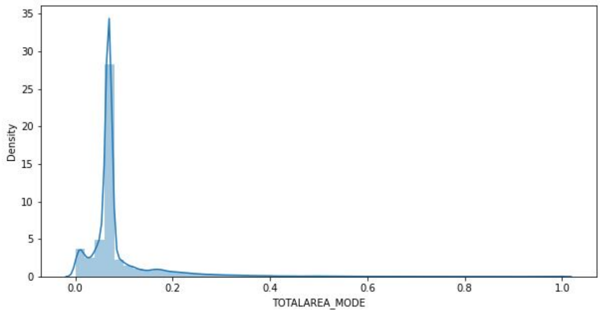
Features with missing values between 10-50%

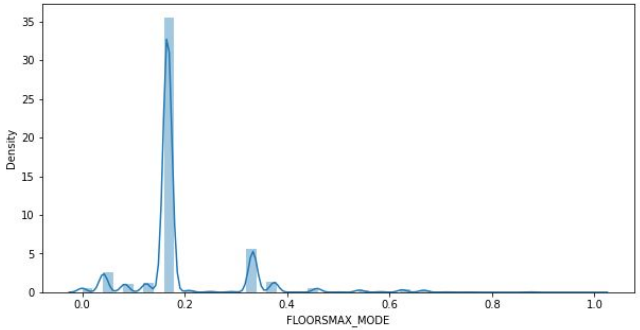
Following features had missing values percentage between 10 to 50%: We imputed the missing values with either Median or Mode depending on whether the feature is categorical or numeric.

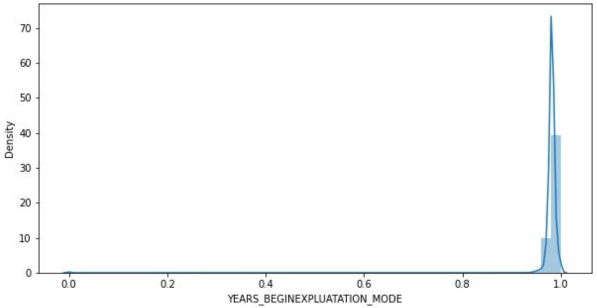
- FLOORSMAX_AVG
- FLOORSMAX_MODE
- FLOORSMAX_MEDI
- YEARS_BEGINEXPLUATATION_MODE
- YEARS_BEGINEXPLUATATION_MEDI
- TOTALAREA_MODE
- EMERGENCYSTATE_MODE
- OCCUPATION_TYPE
- EXT_SOURCE_3
- AMT_REQ_CREDIT_BUREAU_HOUR
- AMT_REQ_CREDIT_BUREAU_QRT
- AMT_REQ_CREDIT_BUREAU_MON
- AMT_REQ_CREDIT_BUREAU_WEEK
- AMT_REQ_CREDIT_BUREAU_DAY
- AMT_REQ_CREDIT_BUREAU_YEAR



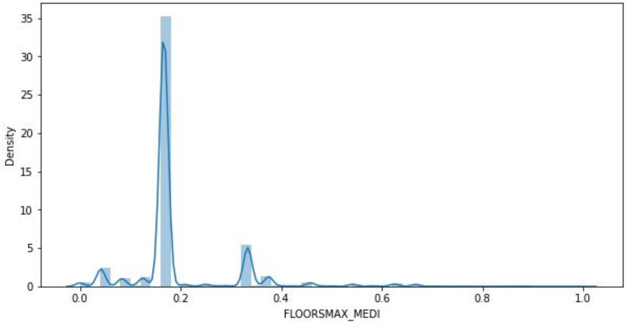


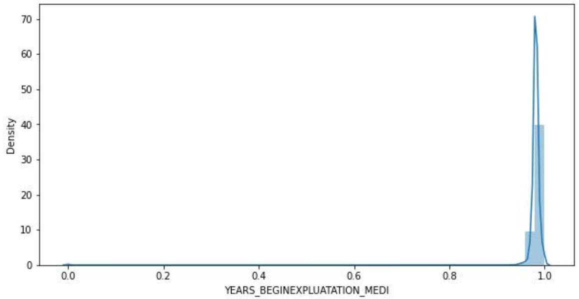


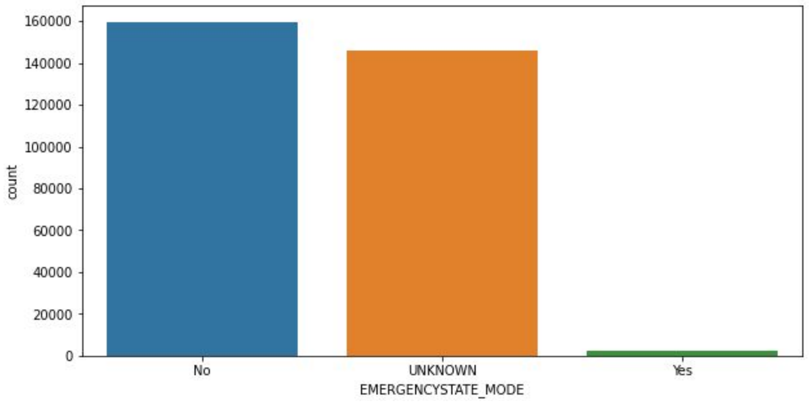




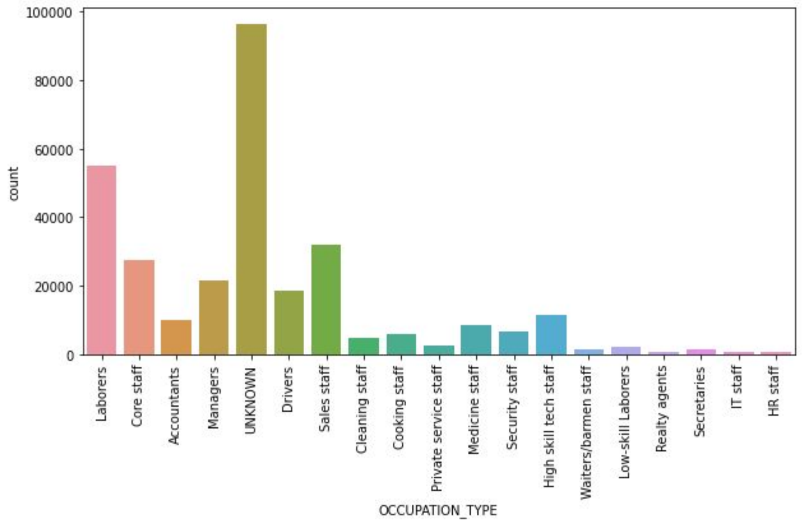
We can observe that these numeric features are normally distributed and some are skewed.
For features with Year information majority of the data is constant.







Before Handling the EMERGENCYSTATE MODE column,it is highly skewed towards **NO.** Instead of Imputing mode of the column, we added **Unknown** as third Category



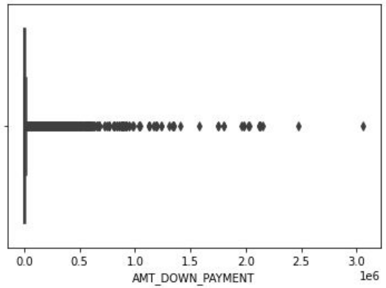
Before Handling the OCCUPATION TYPE column,it is highly skewed towards **LABORERS.** Instead of Imputing mode of the column, we added **Unknown** as a Category

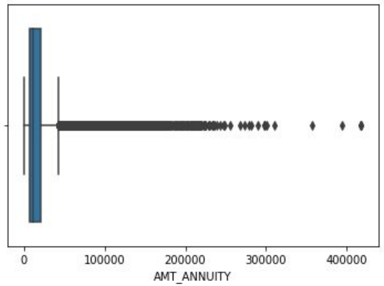
Data cleaning for Previous Application Data

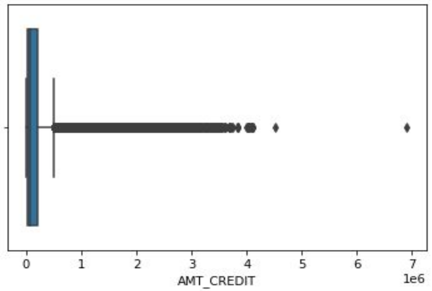
- We can drop columns **RATE_INTEREST_PRIVILEGED**, **RATE_INTEREST_PRIMARY** as they have almost all data missing (99%)
- We can drop rows containing null values for **PRODUCT_COMBINATION**, **AMT_CREDIT** as the percentage of null values is less than 10%
- For rest of the feature we can impute median and mode depending on whether the feature is numeric or categorical.

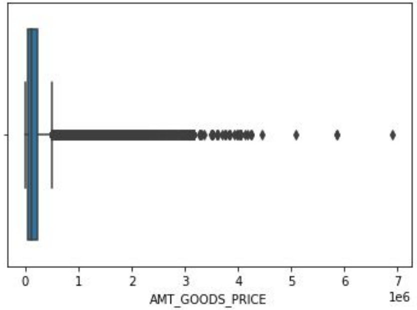
	Column	Null_Percent
13	RATE_INTEREST_PRIMARY	99.643698
14	RATE_INTEREST_PRIVILEGED	99.643698
6	AMT_DOWN_PAYMENT	53.636480
12	RATE_DOWN_PAYMENT	53.636480
20	NAME_TYPE_SUITE	49.119754
31	DAYS_FIRST_DRAWING	40.298129
32	DAYS_FIRST_DUE	40.298129
33	DAYS_LAST_DUE_1ST_VERSION	40.298129
34	DAYS_LAST_DUE	40.298129
35	DAYS_TERMINATION	40.298129
36	NFLAG_INSURED_ON_APPROVAL	40.298129
7	AMT_GOODS_PRICE	23.081773
3	AMT_ANNUITY	22.286665
28	CNT_PAYMENT	22.286366
30	PRODUCT_COMBINATION	0.020716
5	AMT_CREDIT	0.000060

Outliers in Previous Application









For investigating outliers we use boxplot. For Previous Application data we have considered amounts features to find outliers if any. It is evident from the plots that there are outliers present in these features.

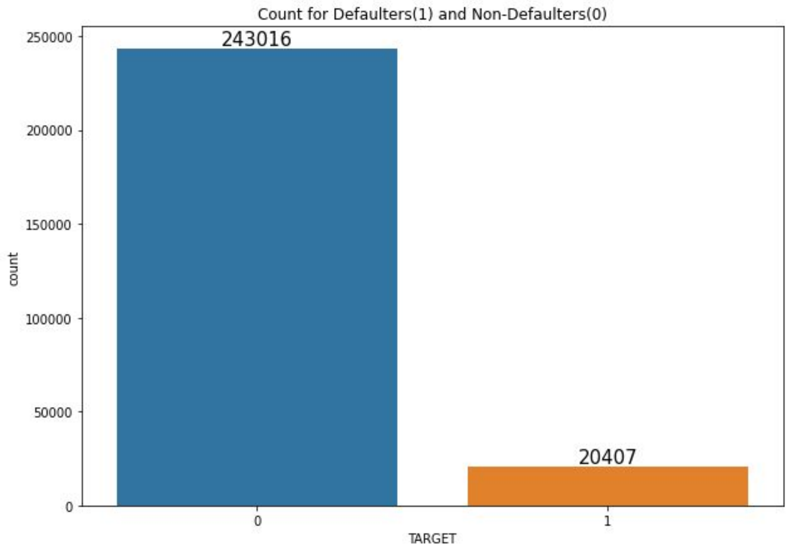
To handle the outliers we can ignore the data after 99 percentile.

Univariate Analysis for Application Data

Analysing the count for target variable 'Target' we have two values 0 for Non defaulters and 1 for Defaulter.

Using seaborn library we can plot a count plot for this feature and we can see that there is a huge data imbalance for this feature.

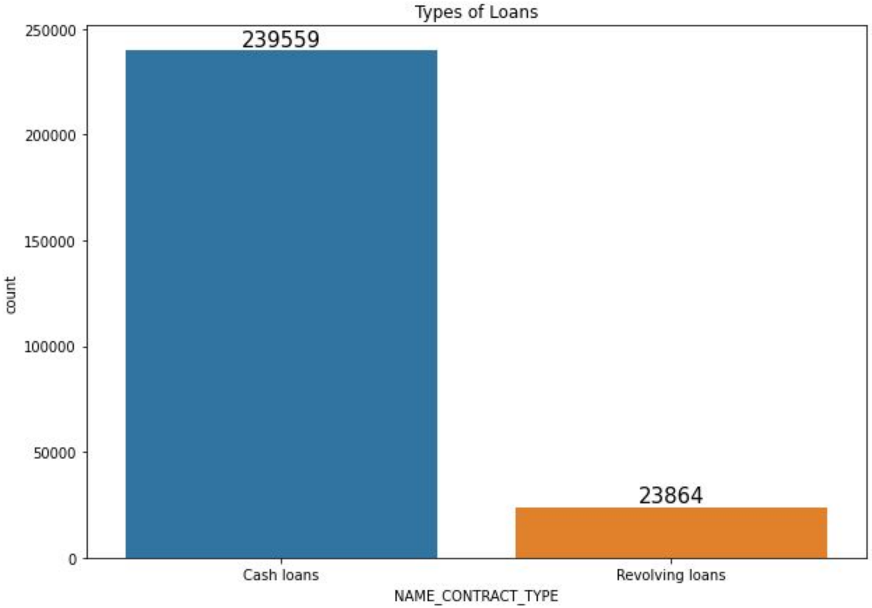
7.74% of total sample data are defaulters.



Type of loans

From the sample data we can conclude that majority of applications are made for Cash loan rather than Revolving loans.

Nearly **90.94%** of applications were made for Cash loans.

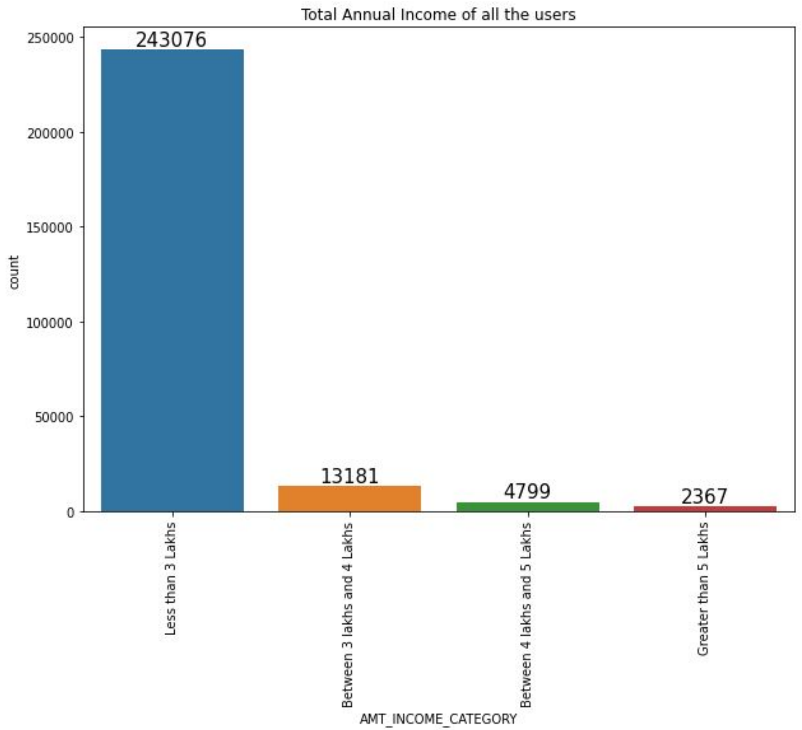


Income Category

For the income category analysis we created bins as:

- Less than 3 lakh
- Between 3 and 4 lakh
- Between 4 and 5 lakh
- Greater than 5 lakh

From the sample data we can state that nearly **92.25%** of the applicants have total annual income less than 3 lakh

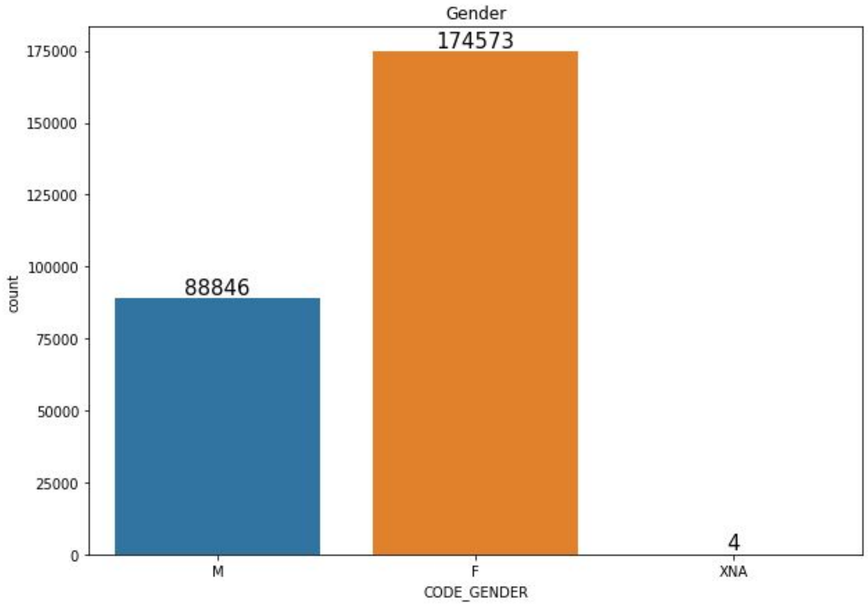


Gender

Most of the applicants are Females.

A very few data points are unknowns(XNA).

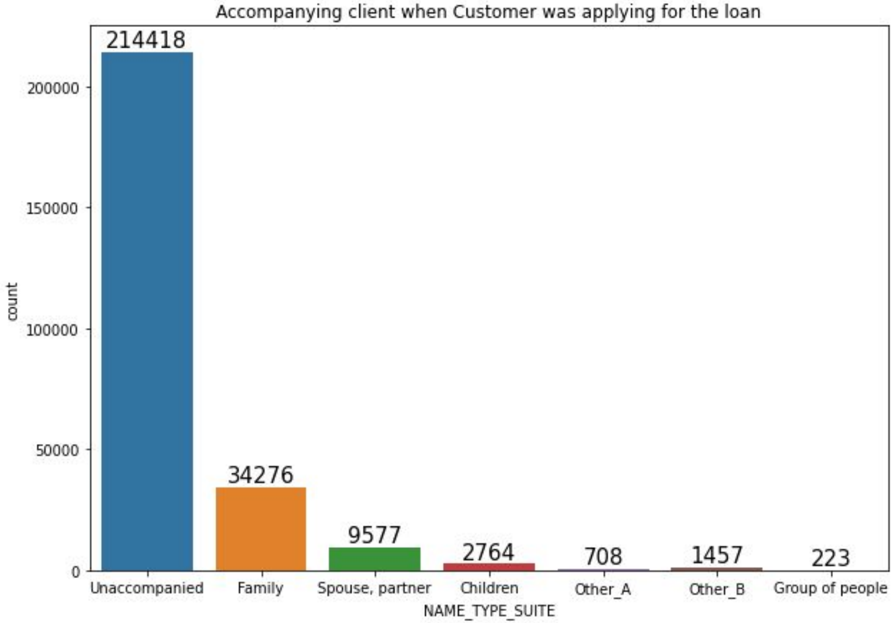
Out of all the applicants Female applicants are nearly **66.27%**



Name type Suite

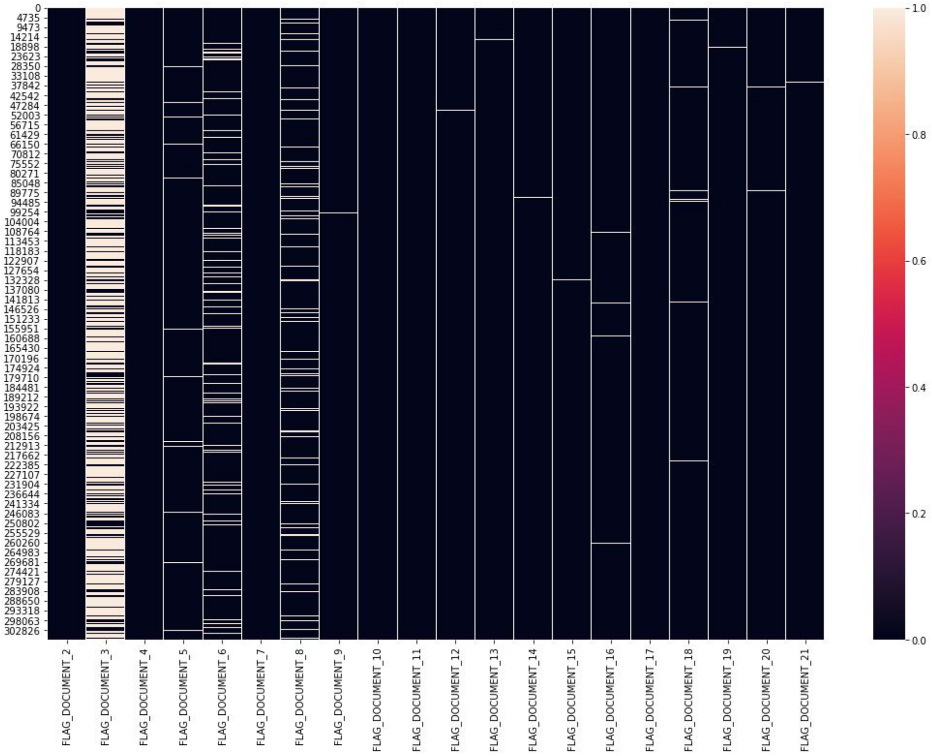
This feature indicates who the applicant was accompanied with.

Almost **81.2%** of the applicants were unaccompanied.



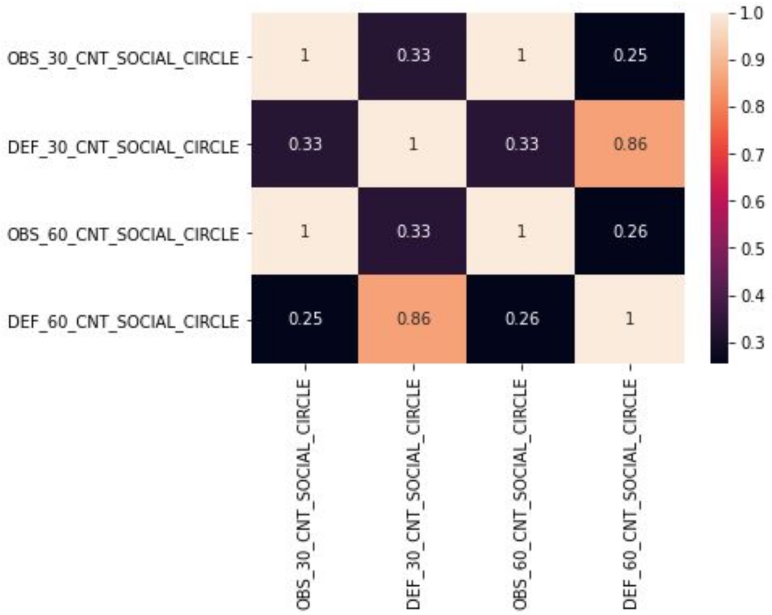
Documents

- There were 20 features regarding documents.
- On plotting a heatmap, it is evident that Document 3 is the required document in most of the cases.
- For a few other cases document 6 and document 8 is required.



Social Info

- DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE are highly correlated
- OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE represent same data



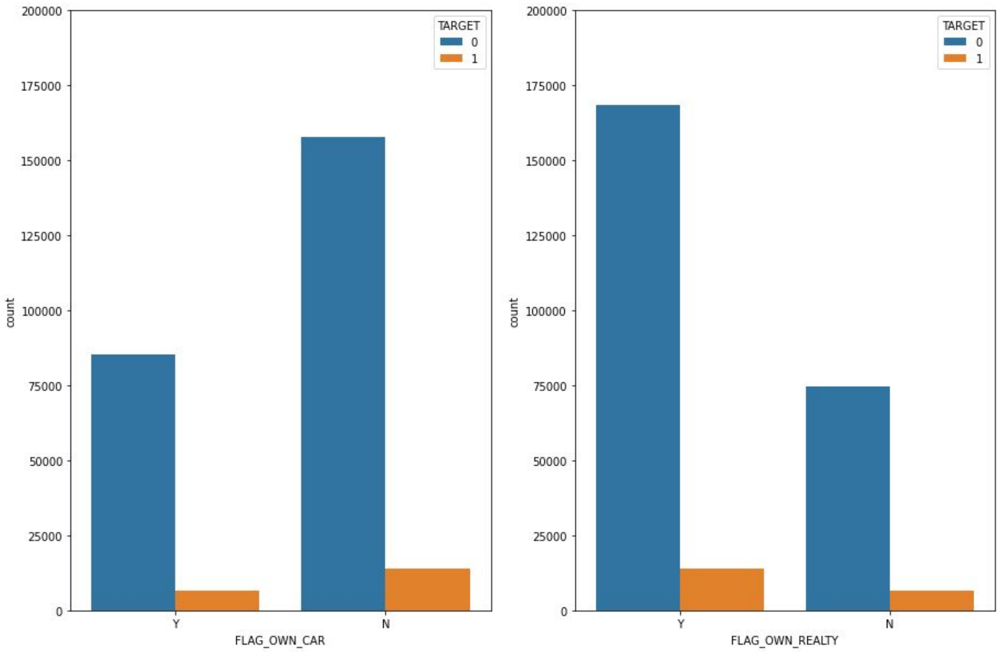
Bivariate Analysis for Current Application Data

We have considered following features for Bivariate Analysis:

- FLAG_OWN_CAR VS TARGET
- FLAG_OWN_REALTY VS TARGET
- NAME_INCOME_TYPE VS TARGET
- NAME_EDUCATION_TYPE VS TARGET
- CODE_GENDER VS TARGET
- NAME_FAMILY_STATUS VS TARGET

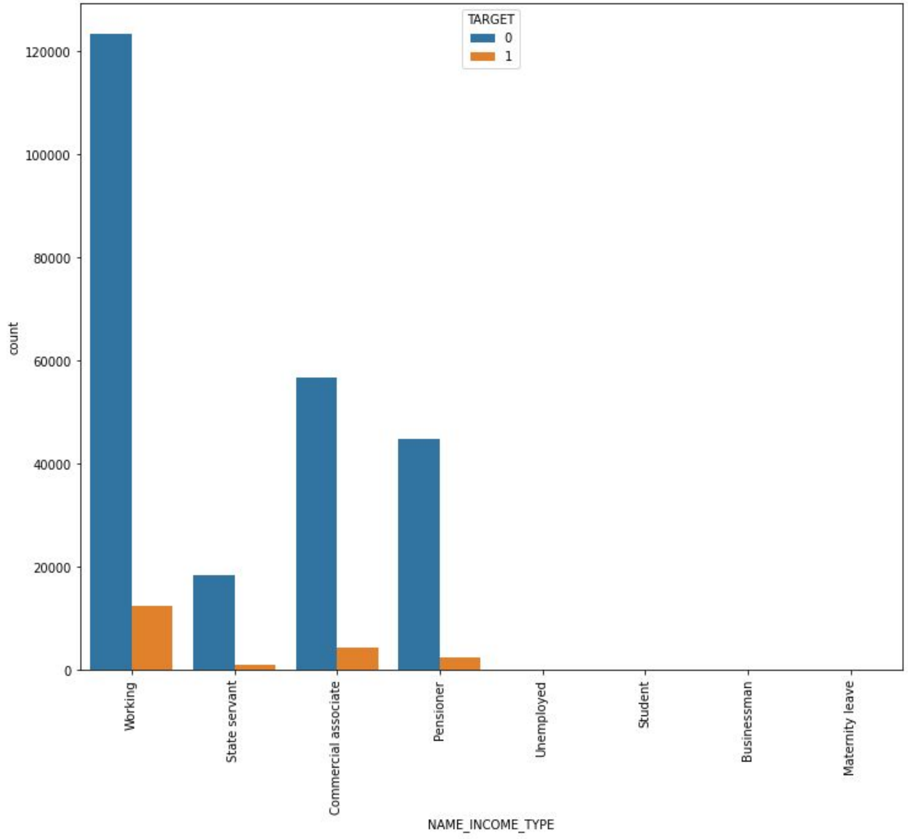
Asset Info

- Most of the applicants own realty
- Most of the applicants do not own cars
- People not owning realty and car and have a slightly higher default rate than the people who own realty and car



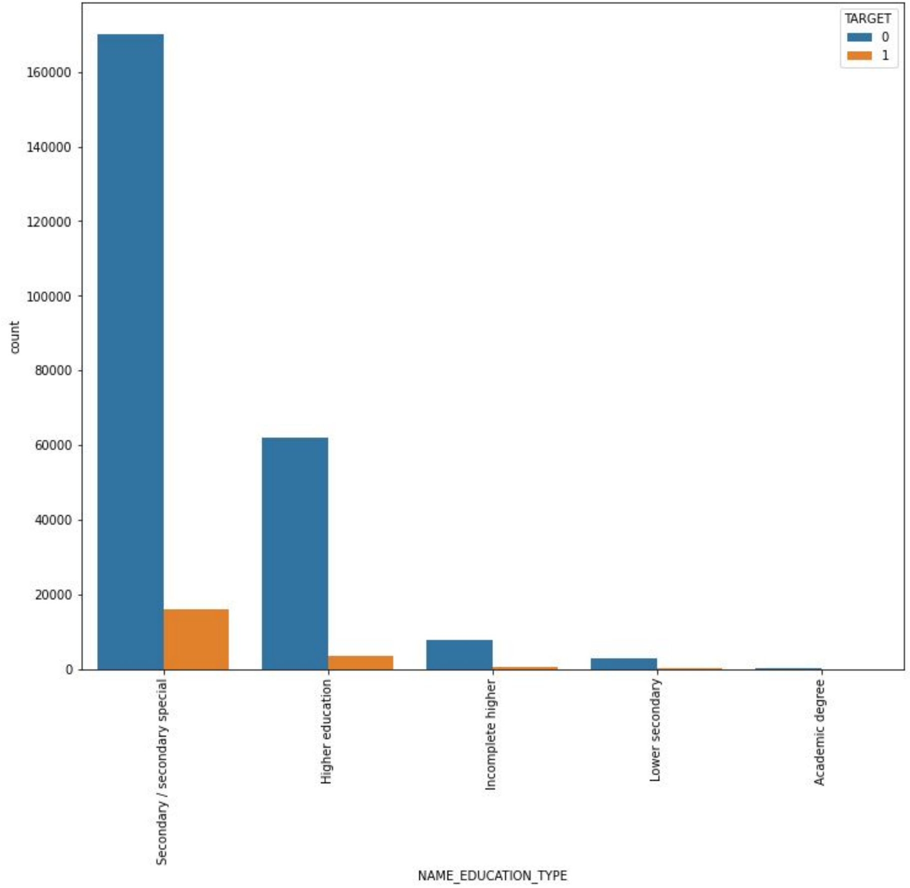
Occupation Info

- Most of the applicants are working.
- 'Unemployed', 'Student', 'Businessman', 'Maternity leave' have very few data in the dataset to contribute in the analysis.



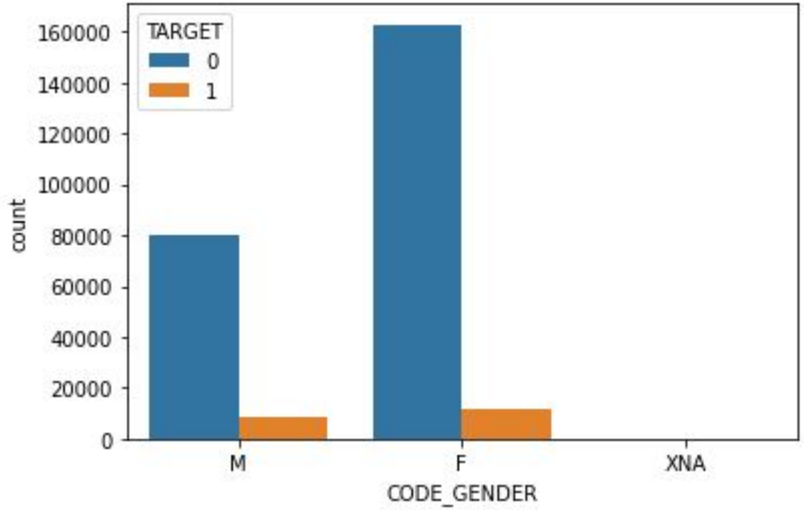
Education Info

- Most of the applicants have Secondary/Secondary special education.
- No. of DEFAULTERS and NON DEFAULTERS are highest in Secondary/Secondary Special Category.



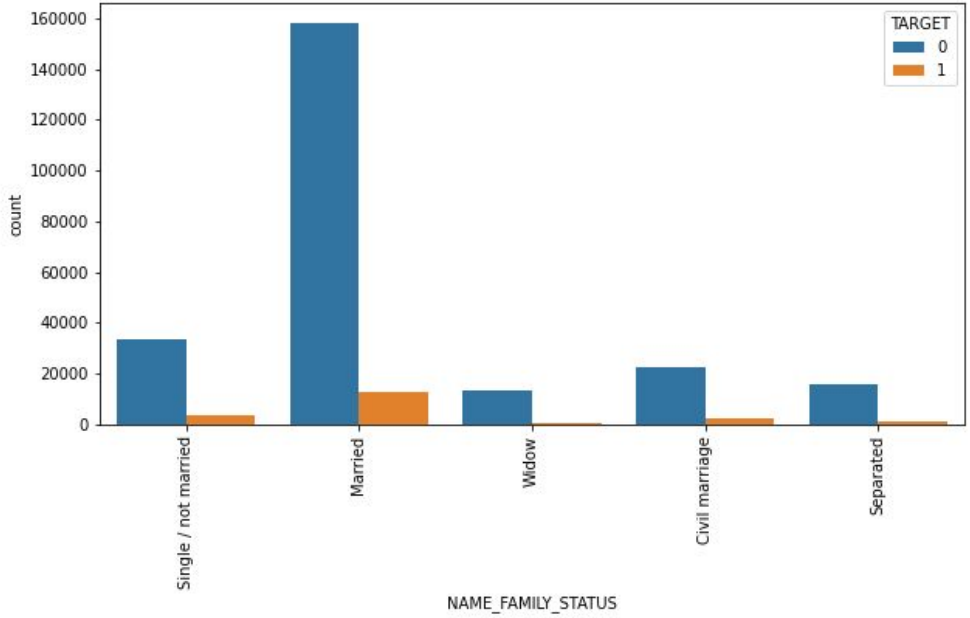
Gender Based Inference

- Female applicants are more than male applicants
- Defaulter percentage is higher for male applicants



Marital Status Info

We can infer that married applicants are relatively safer to sanction loan as the ratio of defaulters to non defaulters is less.



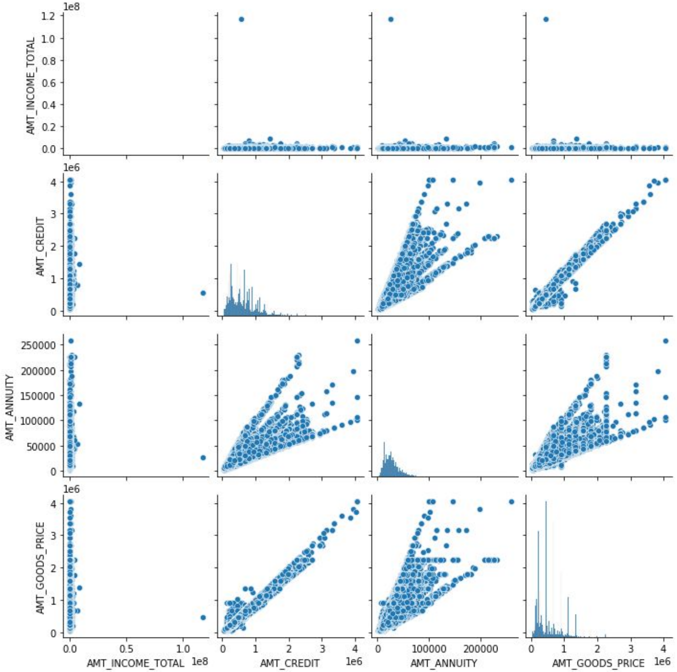
Pairplot for Amounts features

1) The AMT_GOODS_PRICE and AMT_GOODS_PRICE have strong linear correlation.

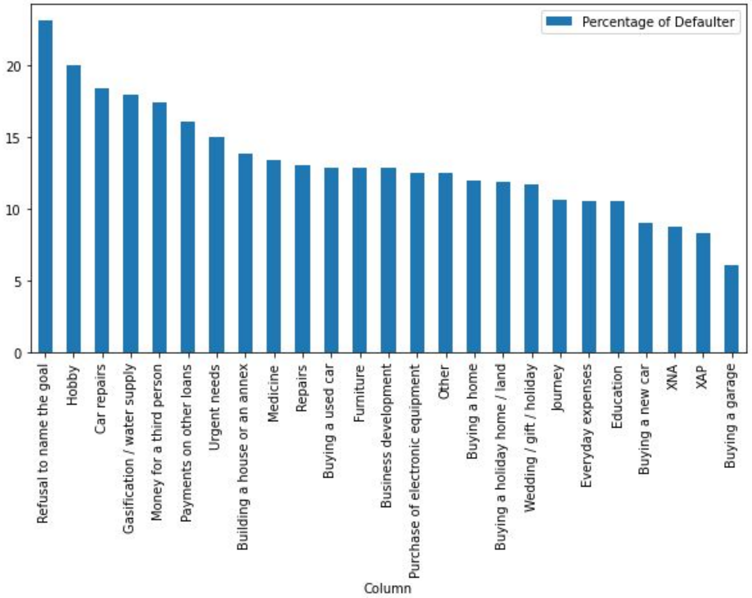
2) There is no linear correlation between AMT_INCOME_TOTAL with other columns

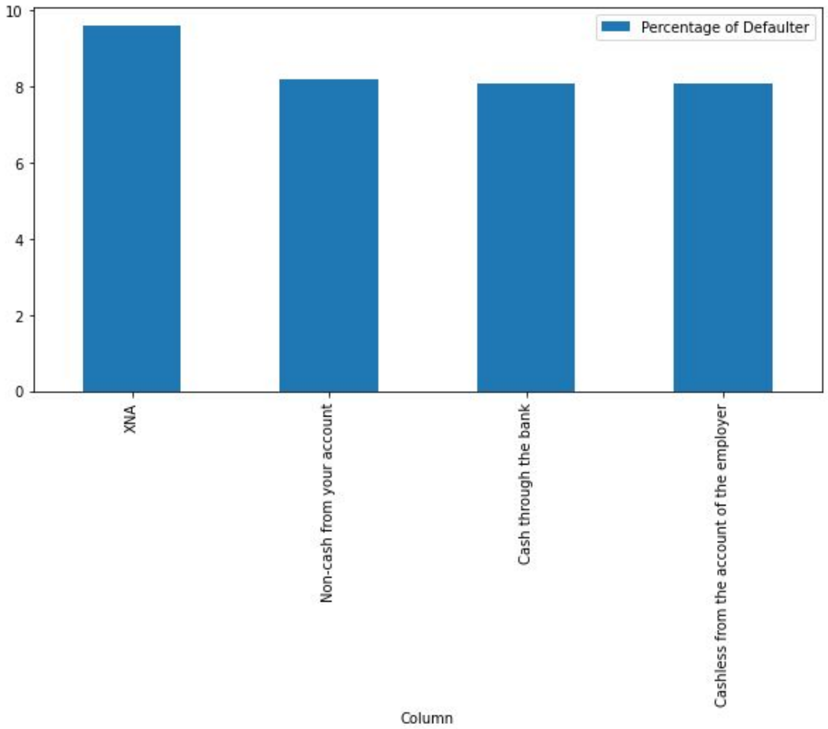
3) AMT_CREDIT and AMT_ANNUITY have weak linear correlation.

4) AMT_GOODS_PRICE and AMT_ANNUITY have weak linear correlation



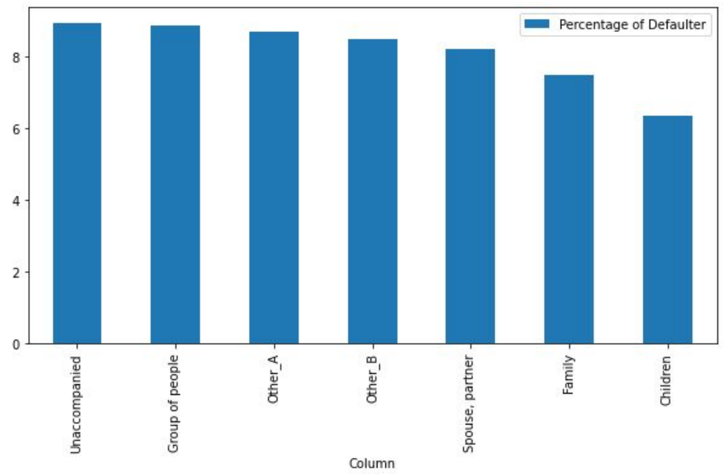
Analysis for Categorical features for Defaulters in Previous Application



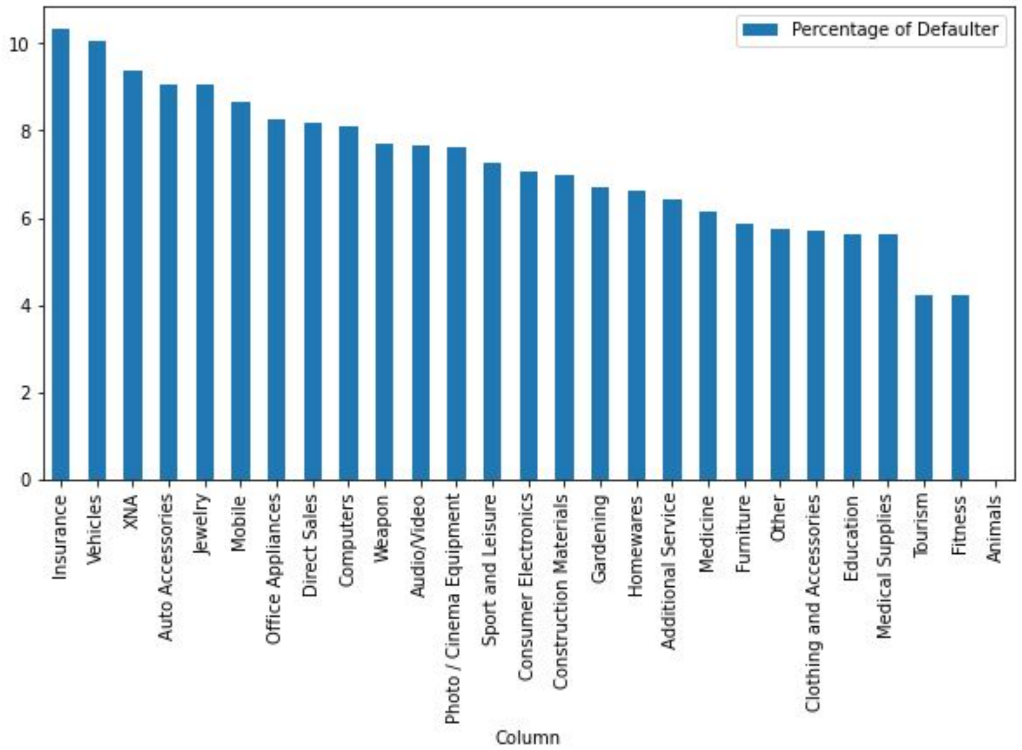


- Applicants refusing to name the goal have highest percentage of defaulters among all other purposes.

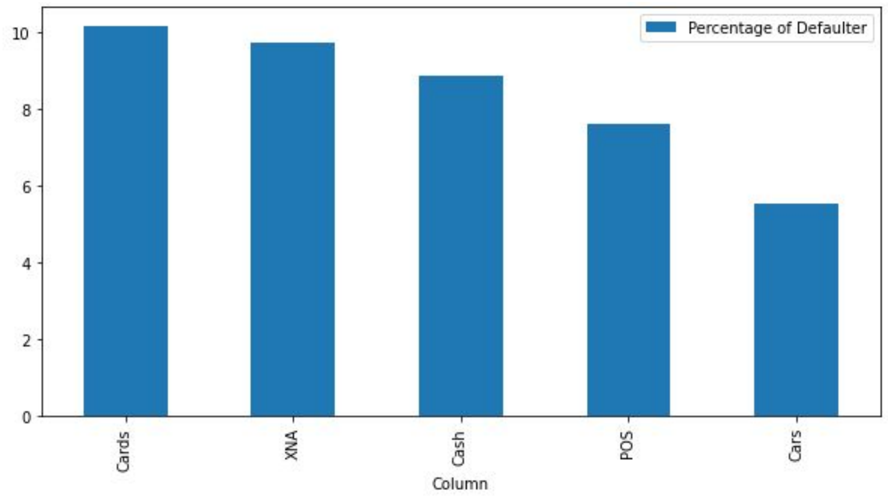
- Most of the customers who defaulted have not provided details for type of Payment.Rest of types have similar defaulter rate.



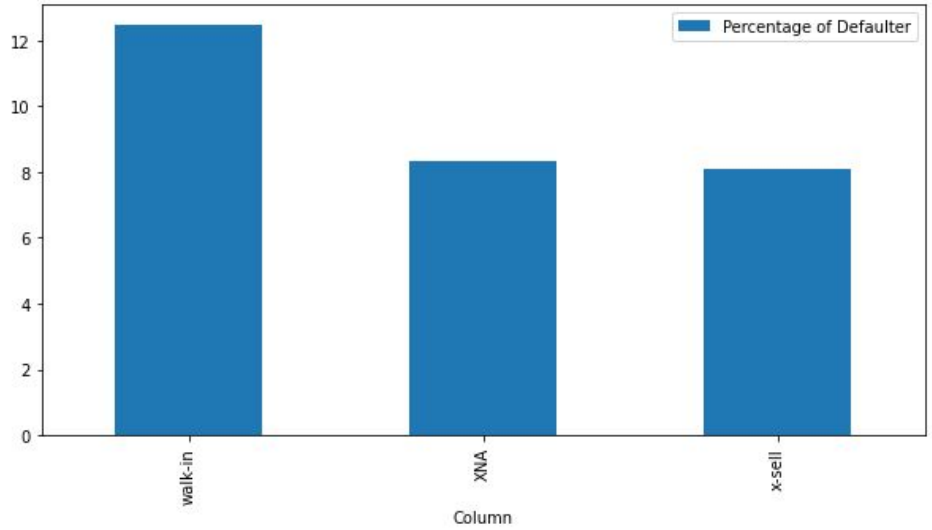
Applicants who were not accompanied during application of loan have more defaulter rate.



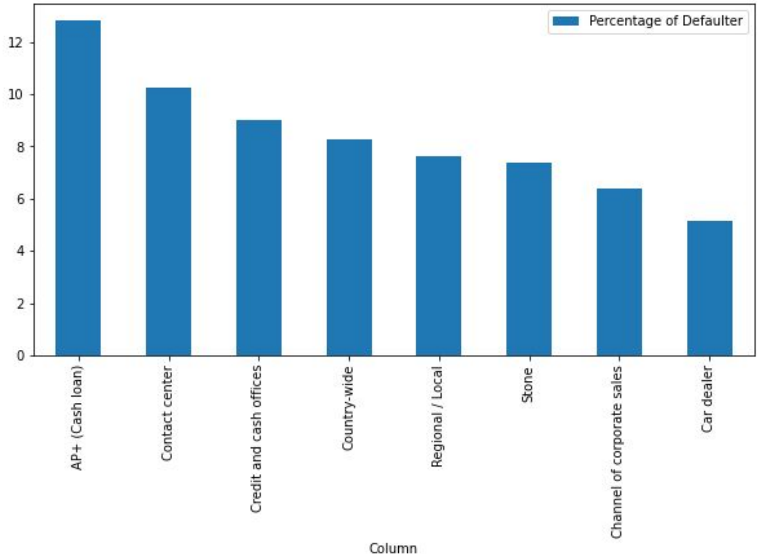
Insurance is the top most category which has a lot of default customers.



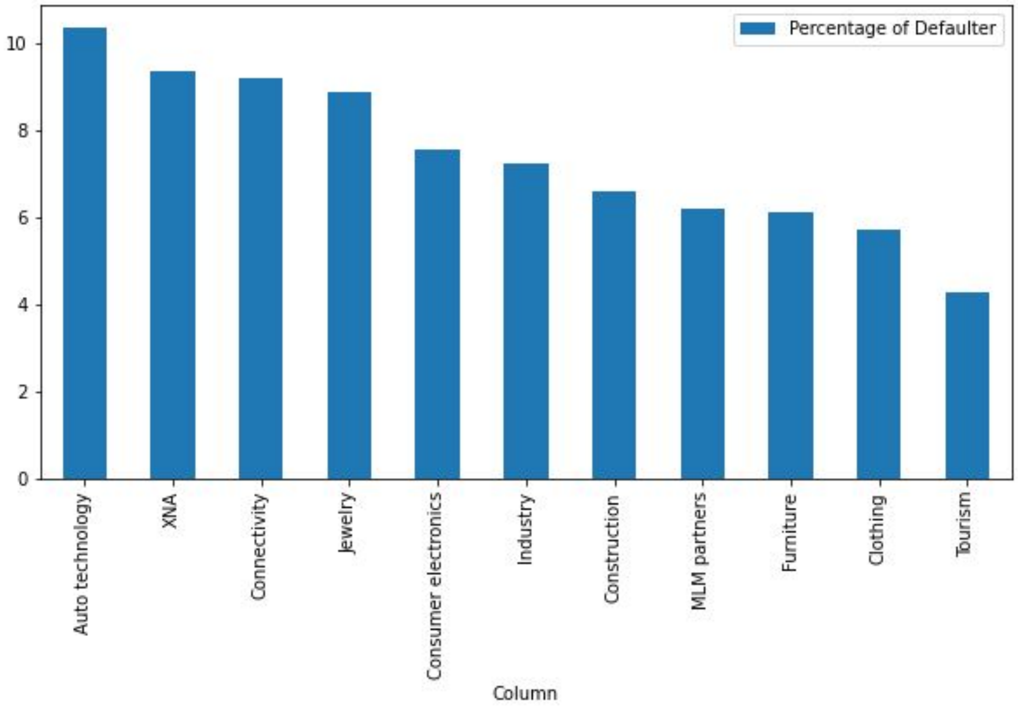
Most of the customers who have cards as name portfolio, have highest default rate.



The people who has previous application as Walk-in, have highest default rate.



Applicants who have AP+(Cash Loan) as channel for acquiring loan have highest default rate.



Applicants having Auto technology as the industry of the seller have highest default rate

Risk Assessments

From the EDA for Defaulters from Previous Application we can infer that for following values in features the risk of the applicant being a defaulter is more:

1. Reason for loan - Refusal to name the goal
2. Type of loan - NA
3. Accompanied Applicant - Unaccompanied
4. Category for loan - Insurance
5. Name portfolio - Cards
6. Channel of Approach - WalkIn
7. Channel of acquiring loan - AP+(Cash loan)
8. Industry of seller - Auto technology

Top 10 Correlation for Defaulters

1.	OBS 30 CNT SOCIAL CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE	0.998270
2.	FLOORSMAX AVG and FLOORSMAX MEDI	0.997295
3.	YEARS BEGINEXPLUATATION MEDI and YEARS_BEGINEXPLUATATION_AVG	0.996139
4.	FLOORSMAX MEDI and FLOORSMAX MODE	0.989472
5.	FLOORSMAX AVG and FLOORSMAX MODE	0.986935
6.	AMT GOODS PRICE and AMT CREDIT	0.982783
7.	YEARS BEGINEXPLUATATION MODE and YEARS BEGINEXPLUATATION AVG	0.980546
8.	YEARS BEGINEXPLUATATION MEDI and YEARS BEGINEXPLUATATION_MODE	0.978163
9.	REGION RATING CLIENT and REGION RATING_CLIENT_W_CITY	0.956637
10.	CNT_CHILDREN and CNT_FAM_MEMBERS	0.885484

Top 10 Correlation for Non Defaulters

1.	OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE	0.998510
2.	FLOORSMAX_AVG and FLOORSMAX_MEDI	0.997253
3.	YEARS_BEGINEXPLUATATION_MEDI and YEARS_BEGINEXPLUATATION_AVG	0.993594
4.	FLOORSMAX_MODE and FLOORSMAX_MEDI	0.988955
5.	AMT_CREDIT and AMT_GOODS_PRICE	0.987022
6.	FLOORSMAX_AVG and FLOORSMAX_MODE	0.986569
7.	YEARS_BEGINEXPLUATATION_AVG and YEARS_BEGINEXPLUATATION_MODE	0.971086
8.	YEARS_BEGINEXPLUATATION_MEDI and YEARS_BEGINEXPLUATATION_MODE	0.962133
9.	REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT	0.950149
10.	CNT_FAM_MEMBERS and CNT_CHILDREN	0.878571

Summary for Current Application Data

- This data is highly imbalanced as number of defaulter is very less in the sample.
- Documents : Most of the applicants did not submit any documents apart from DOCUMENT_3.
- Housing:
  - Most of the applicants live in House/Apartment.
  - Applicants living with their parents or in rented apartment have higher rate of default.
- Social Circle Info: The features show similar trend for defaulters and non defaulters, can be dropped.
- Asset Info :
  - Most of the applicants own realty.
  - Most of the applicants do not own cars.
  - People not owning reality and car and have a slightly higher default rate than the people who own reality and car
- Gender Info :
  - Female applicants are more than male applicants
  - Defaulter percentage is higher for male applicants
  - XNA values can be replaced with "Unknown"

Thank You