CREDIT EDA CASE STUDY

INTRODUCTION

There are two datasets are provided for this case studies as follows :

- Application Data
- Previous Application Data

Application Data dataset analysis :

Data Cleaning

1. Found out the % of missing values in each column so as to determine which value to delete.

```
# To calculate percentage of NaN values in DataFrame
def get_perc_of_missing_values(series):
    num = series.isnull().sum()
    den = len(series)
    return round(num/den, 3)
get_perc_of_missing_values(application_data)
```

2. Removed columns with > **30% NaN** values

```
# Iterate over columns in DataFrame and delete the values are null and > 30%

for col, values in application_data.iteritems():
    if get_perc_of_missing_values(application_data[col]) > 0.30:
        application_data.drop(col, axis=1, inplace=True)
application_data
```

3.Imputing values on columns to make the data set usable.

```
application_data['AMT_GOODS_PRICE'].fillna((application_data['AMT_GOODS_PRICE'].mean()), inplace=True)

application_data['EXT_SOURCE_2'].fillna((application_data['EXT_SOURCE_2'].mean()), inplace=True)
```

	count	mean	std	min	25%	50%	75%	
AMT_GOODS_PRICE	307511.0	538396.207429	369279.426396	4.050000e+04	238500.000000	450000.000000	679500.000000	405
EXT_SOURCE_2	307511.0	0.514393	0.190855	8.173617e-08	0.392974	0.565467	0.663422	

Imputed mean values to the **AMT_GOODS_PRICE** and **EXT_SOURCE_2** columns

4. Imputing the mode values to the **NAME_TYPE_SUITE** column

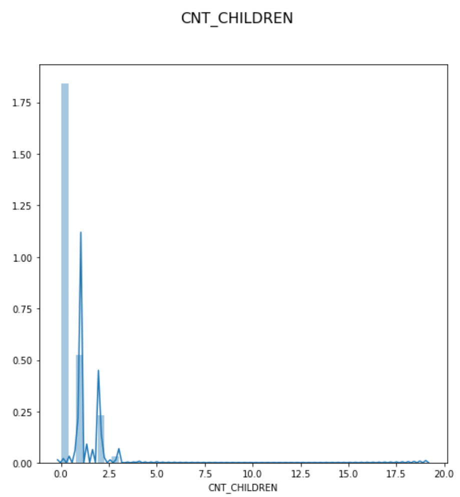
```
application_data.NAME_TYPE_SUITE.value_counts()
```

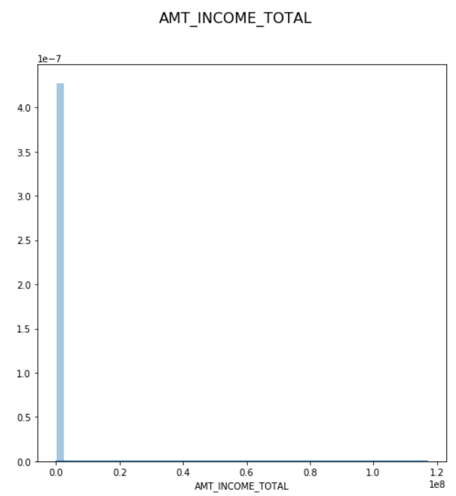
```
Unaccompanied      248526
Family              40149
Spouse, partner     11370
Children             3267
Other_B              1770
Other_A               866
Group of people       271
Name: NAME_TYPE_SUITE, dtype: int64
```

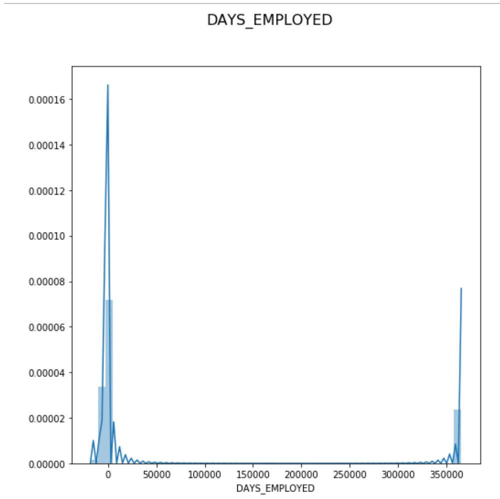
```
# Unaccompanied data has the highest mode, so filling missing values with Unaccompanied

application_data["NAME_TYPE_SUITE"].fillna(application_data["NAME_TYPE_SUITE"].mode()[0],inplace=True)
```

OUTLIERS

Spot outliers in the columns and find reasons for this outlier value presence.







Plot of **CNT_CHILDREN** show a large outlier (19). Since a family rarely have 19 children.

Plot of **AMT_INCOME_TOTAL**, the MAX amount is larger than the other statistical data [Mean, (25,50,75) percentiles]

Plot of **DAYS_EMPLOYED** there is a value present at 36k range, this won't be possible.

Binning of salaries into High, Medium and Moderate Levels and converted date of birth to age.

ANALYSIS OF APPLICATION DATA

Divided data into defaulter and good clients dataframes

```
good_client = application_data[application_data.TARGET == 0]
defaulter_client = application_data[application_data.TARGET == 1]
```

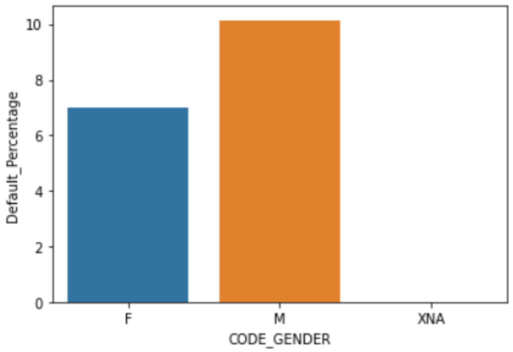
Target == 0 => The client is not a defaulter thus a good client.

Target == 1 => The client with payment difficulties, had late payment more than X days on at least one of the first Y instalments of the loan in sample.

Univariate Analysis of Categorical and Numerical Data

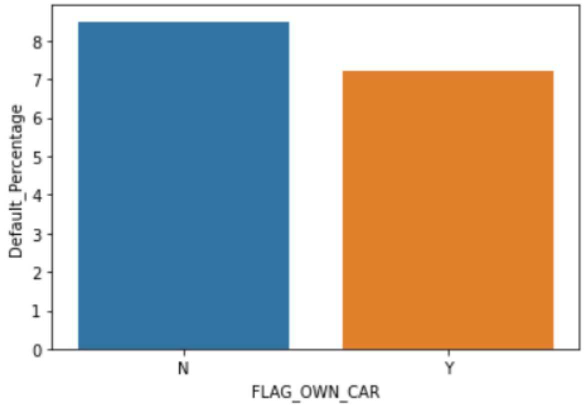
Checking for clients that might to be defaulters/unlikely to pay back the loan by analysing various columns in the data.

- ❑ Based on **CODE_GENDER** (Client's gender)



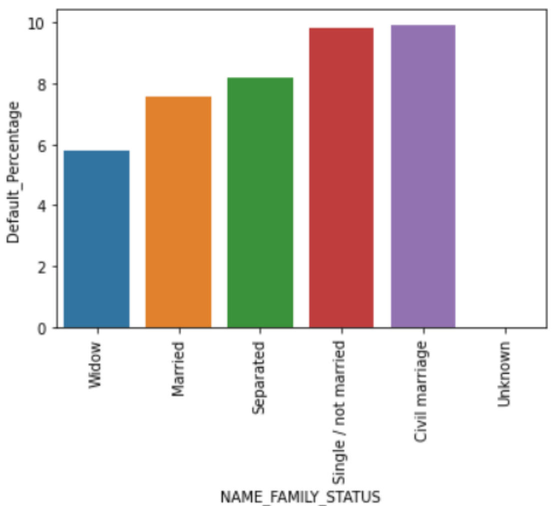
The Female clients are a better TARGET as compared to the Male clients. By observing the percent of defaulted credits, male client have a higher chance of not returning their loans [10.14%], with compared to the female clients [7%].

❑ Based on **FLAG_OWN_CAR** (client owns a car or not)



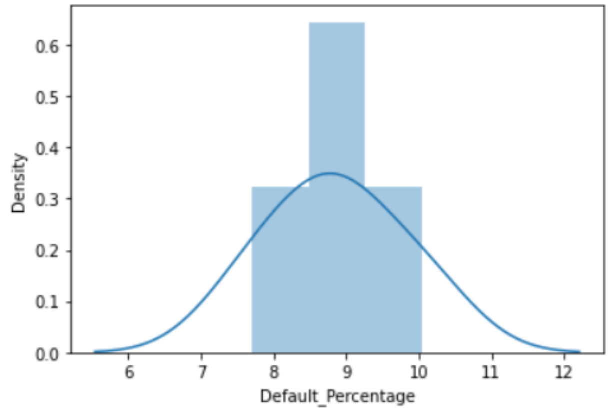
The clients, owns a car are less likely to not repay the loan when compared to the ones that does not own a car. The loan non-repayment rates of both the Car Owners and Non-Car Owners are very close.

❑ Based on **NAME_FAMILY_STATUS**



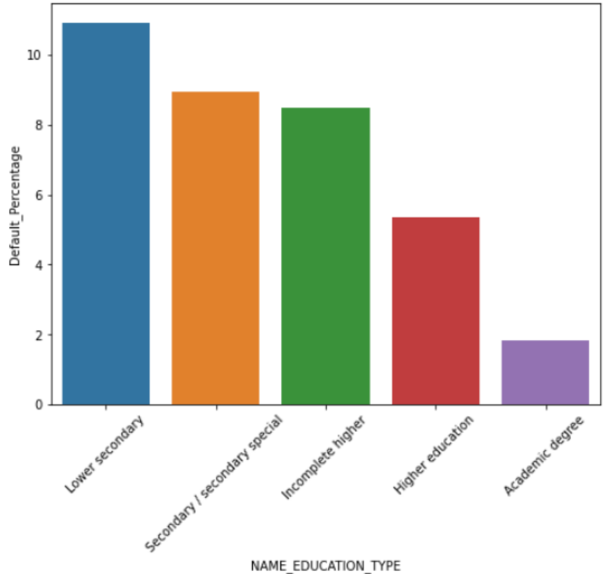
Percentage of non-repayment of loan is at highest for civil mariage and is lowest for widows.

❑ Based on **CNT_CHILDREN**



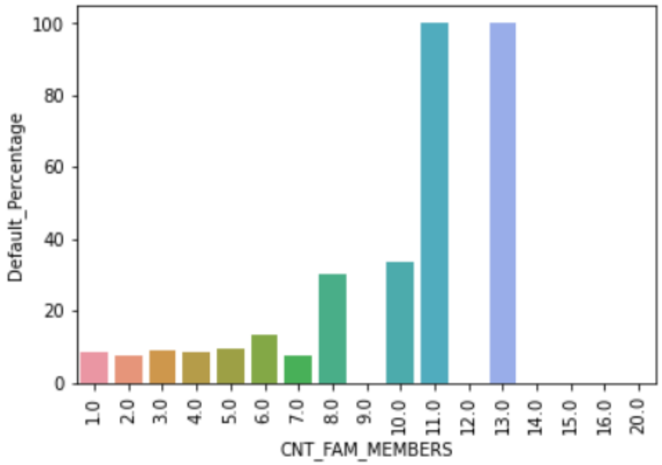
Ther is more chance for a client with more children to not repay the loan back. The more the number of children the more difficult it is for the client to repay the loan due to more personal expenditures.

❑ Based on **NAME_EDUCATION_TYPE**



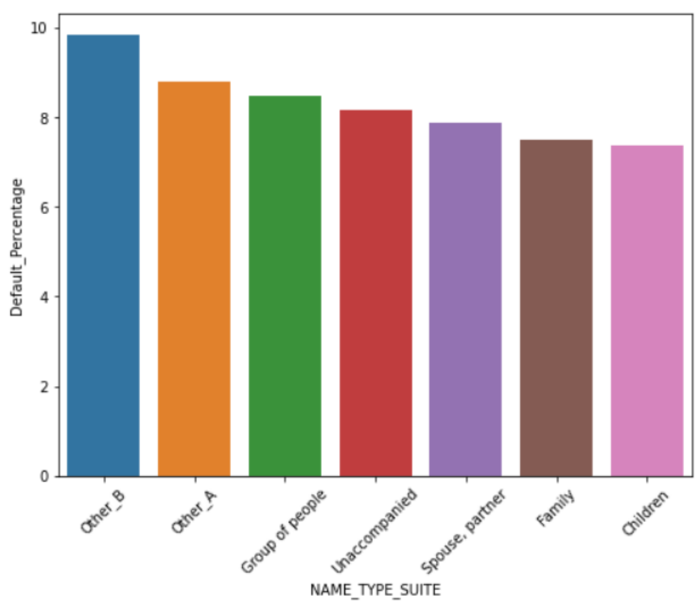
The more educated clients are likely to repay their loans because they will be having more stable jobs with monthly income.

❑ Based on **CNT_FAM_MEMBERS**



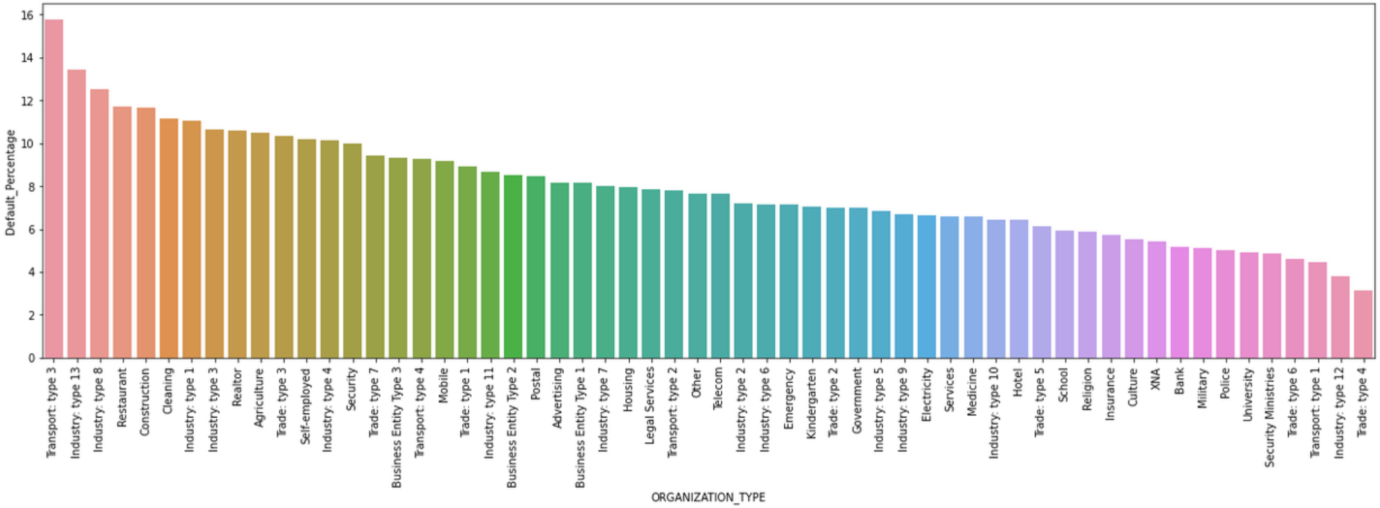
Families with 11,13 members shows highest default rate, but their count is very less [2]

❑ Based on **NAME_TYPE_SUITE**



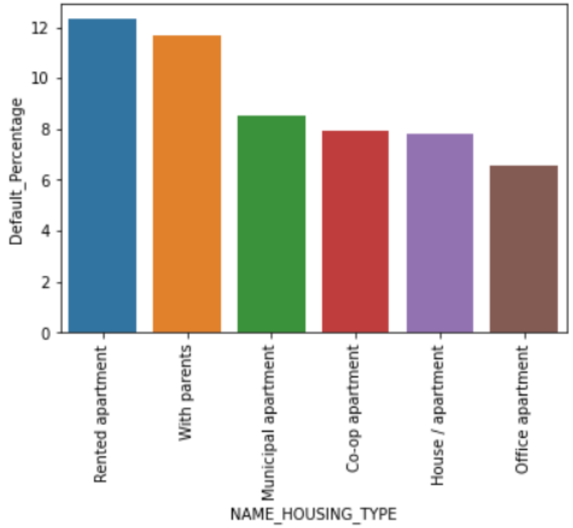
Other_B followed by Other_A are unlikely to pay back their loans.

- Based on **ORGANISATION_TYPE**



From above graph, highest number of non-repayment can be seen in Applicants who work in Transport Type3.

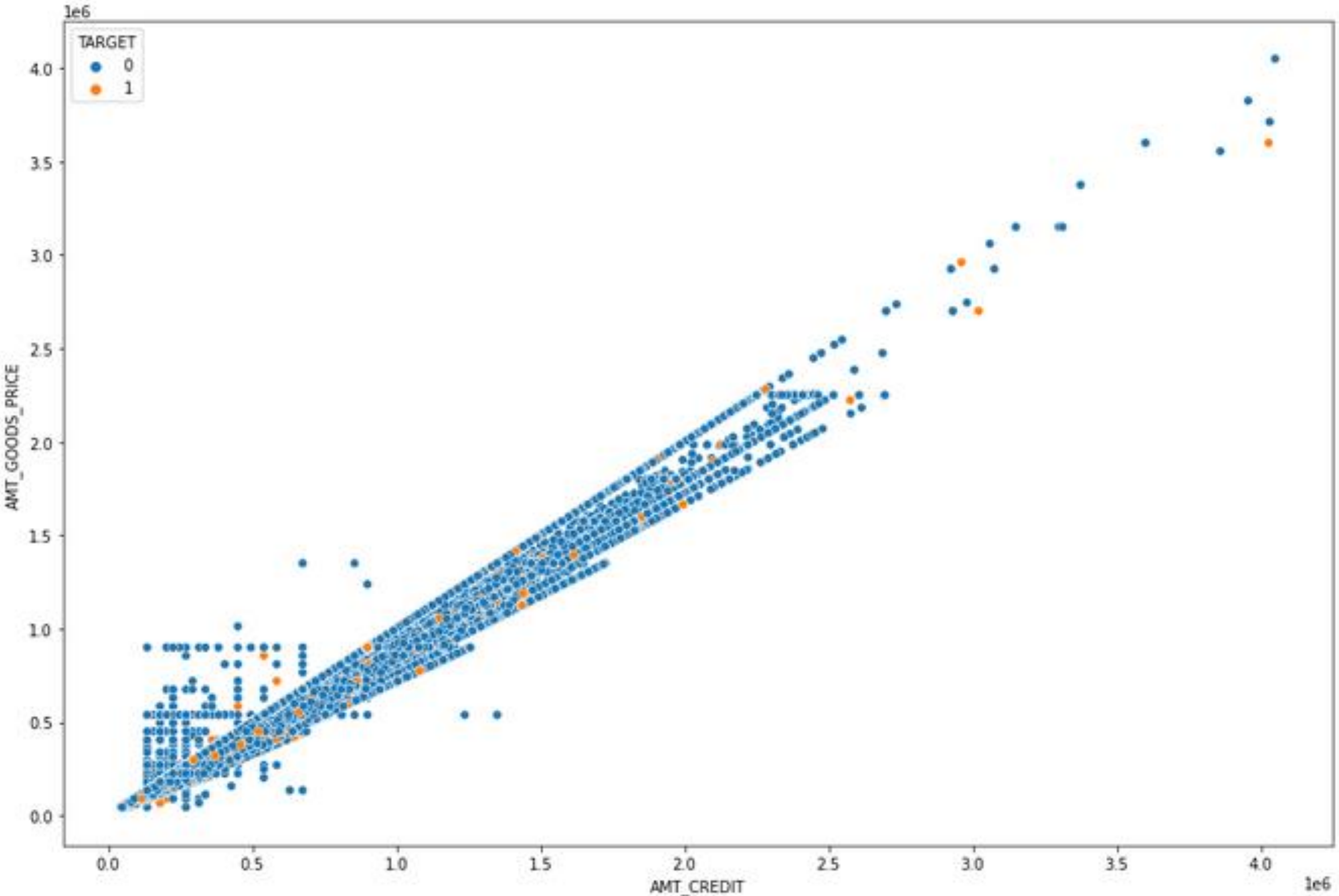
- Based on **NAME_HOUSING_TYPE**



People with rented apartments are less likely to pay back their loans.

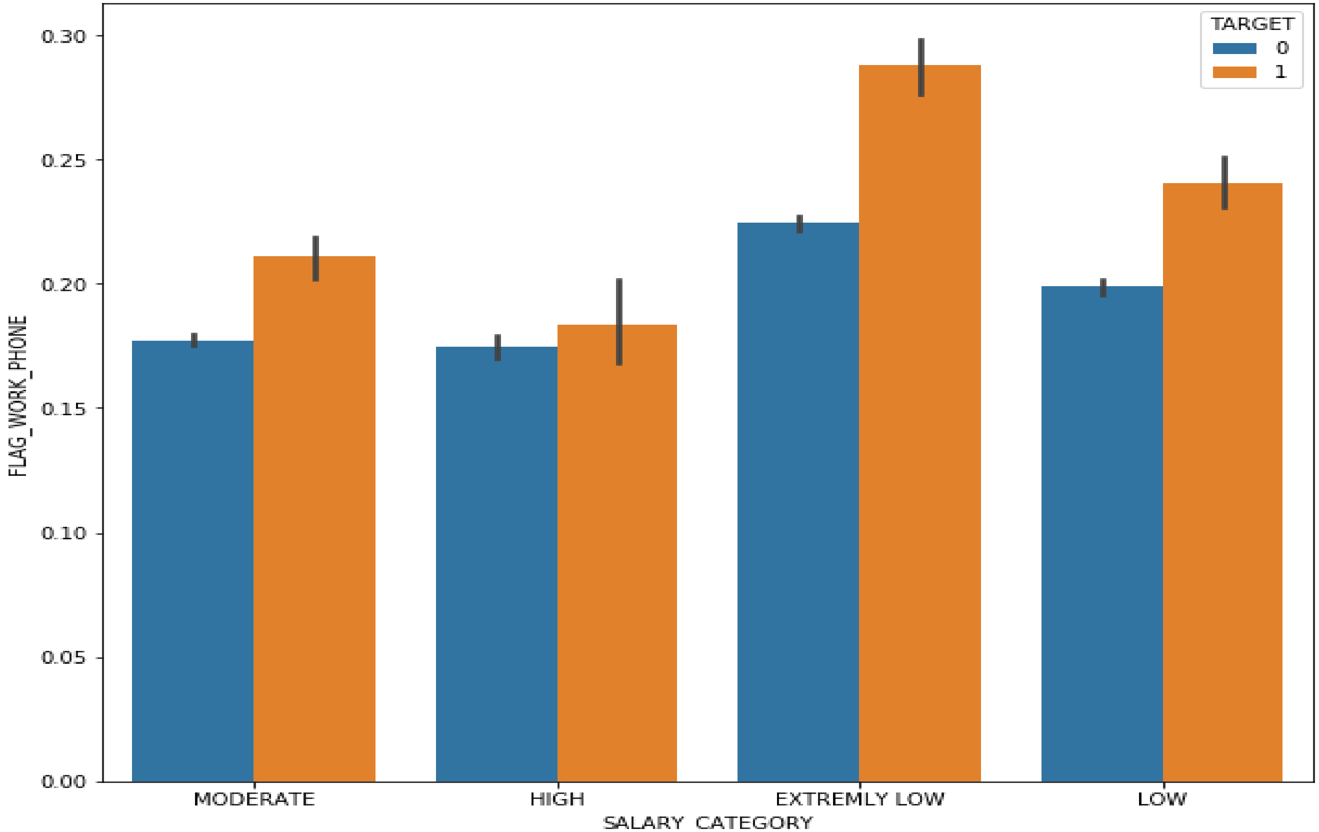
BIVARIATE ANALYSIS

- **AMT_CREDIT vs AMT_GOODS_PRICE**



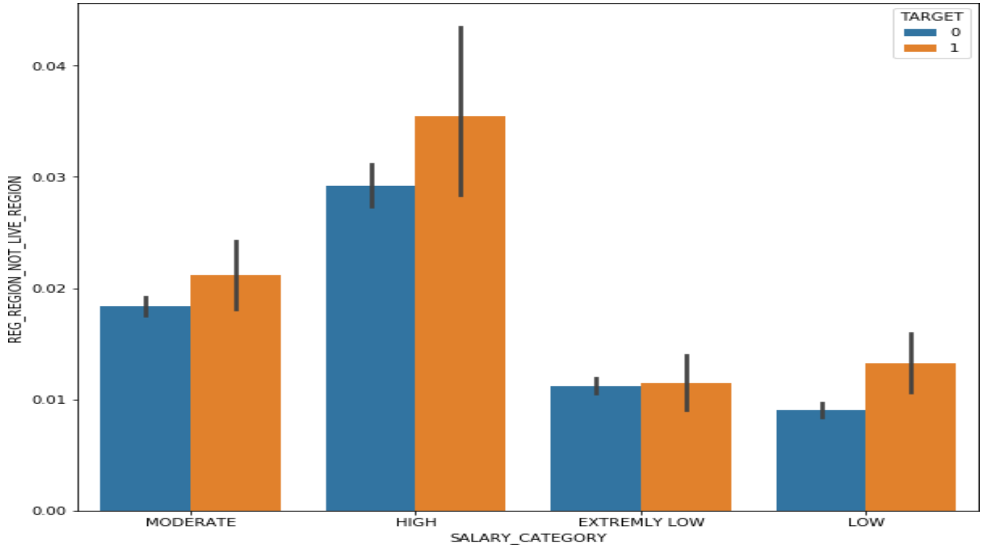
Credit amount and the Amount goods price are correlated with the defaulters and defaulters are linearly increasing as variable increases.

❑ Salary Category vs Client with provided Home Number



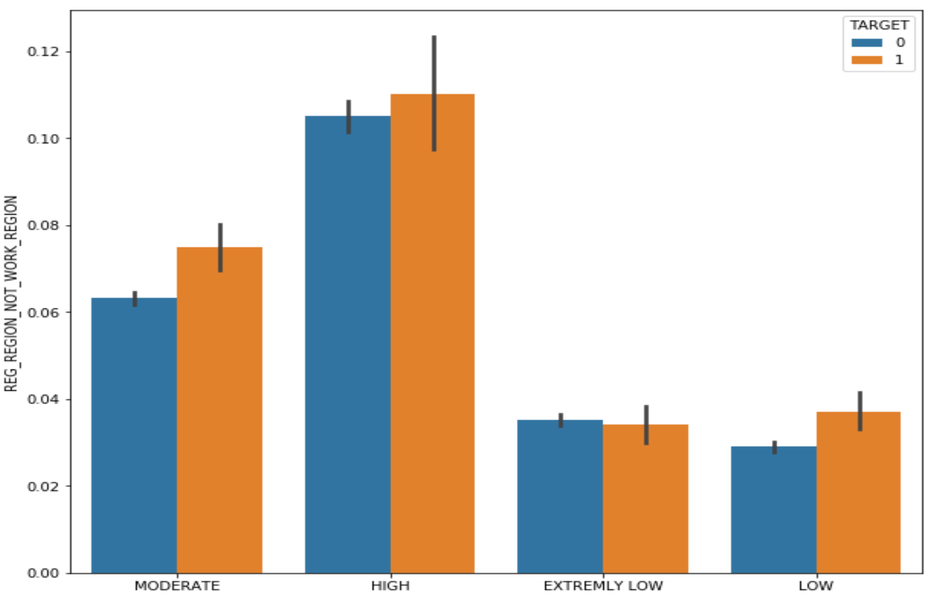
Client with low salary has more chance to be a defaulter, when did not provide the Home phone number. Approx.. 30% people provided the phone number

- **Salary vs Client Whose Permanent Address Not Match With Contact Address at Region Level**



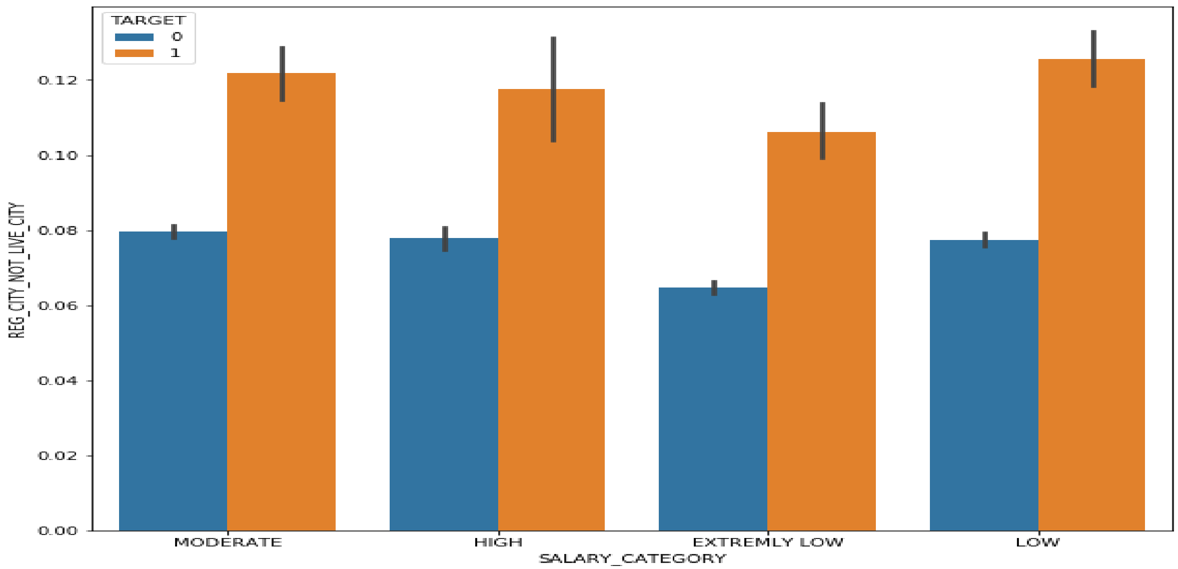
When Client gets lower salary and if his/her Contact address does not match, then there is a Higher chance for him/her to be defaulter

- **Salary vs Client whose Permanent Address not match with Work Address at Region Level**



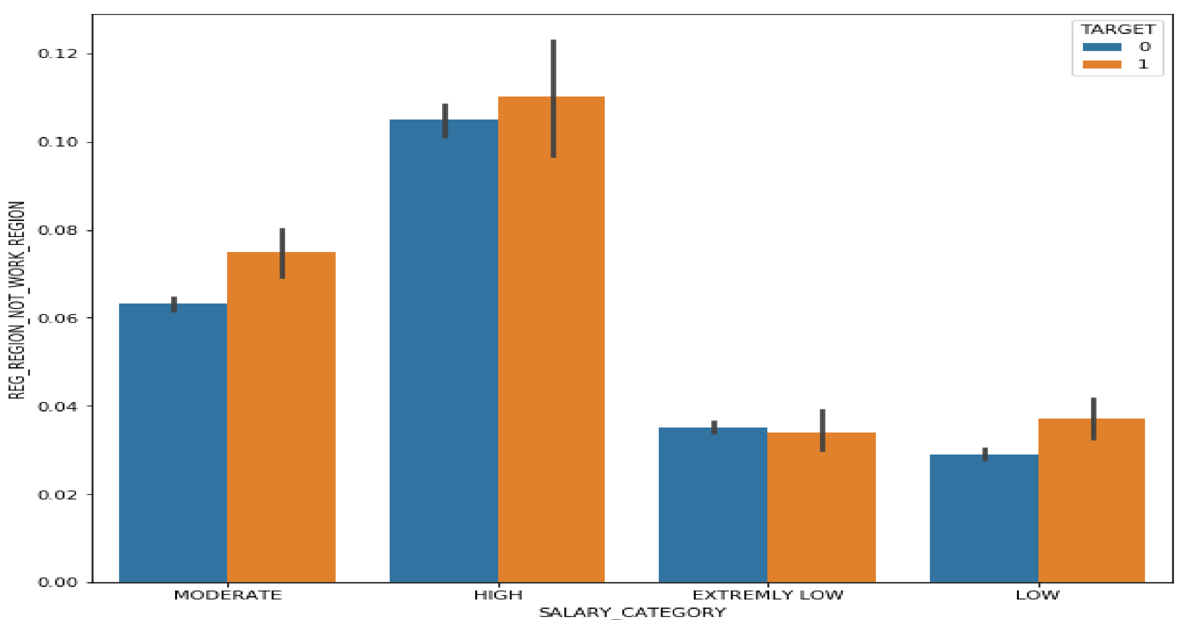
When Client gets lower salary and if his/her Contact address does not match, then there is a Higher chance for him/her to be defaulter

❑ Salary vs Client whose Permanent Address not match with Contact Address at City Level



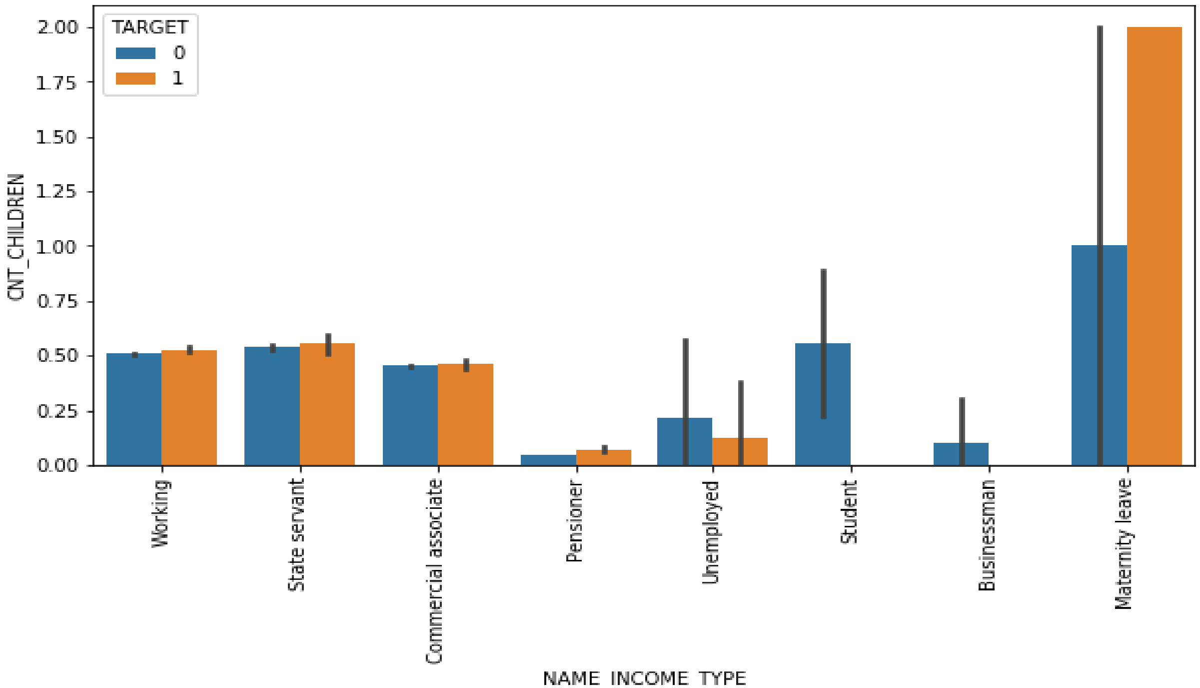
When Client gets LOWER salary and if his/her CONTACT address (CITY-LEVEL) does not match, then there is a Higher chance for him/her to be defaulter.

❑ Salary vs Client whose Permanent Address not match with Work Address at City Level



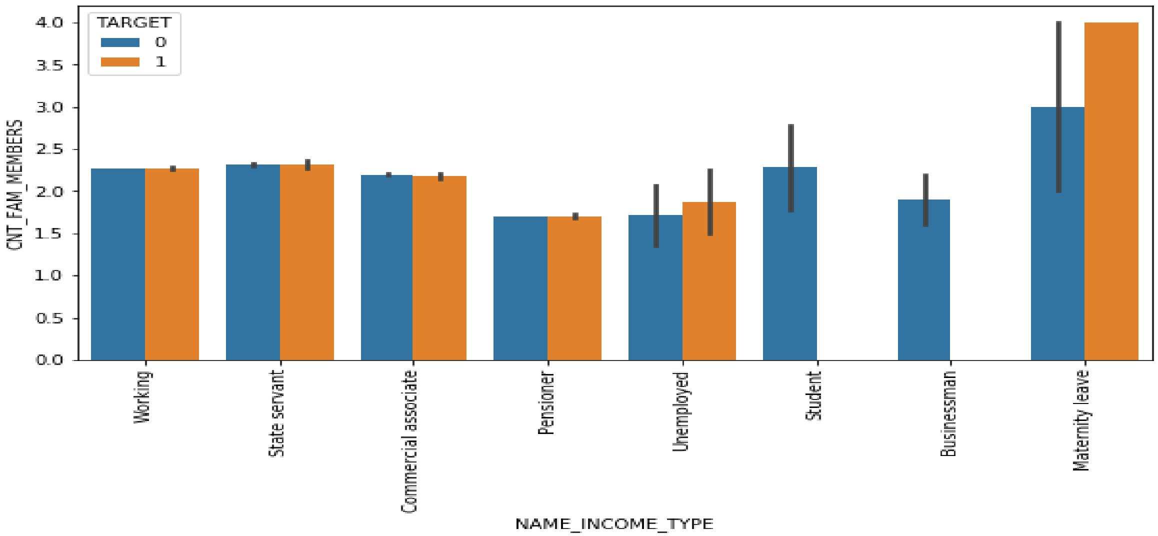
When Client gets HIGH salary and if his/her WORK address (CITY-LEVEL) does not match, then there is a Higher chance for him/her to be defaulter.

❑ INCOME vs CHILDREN Count



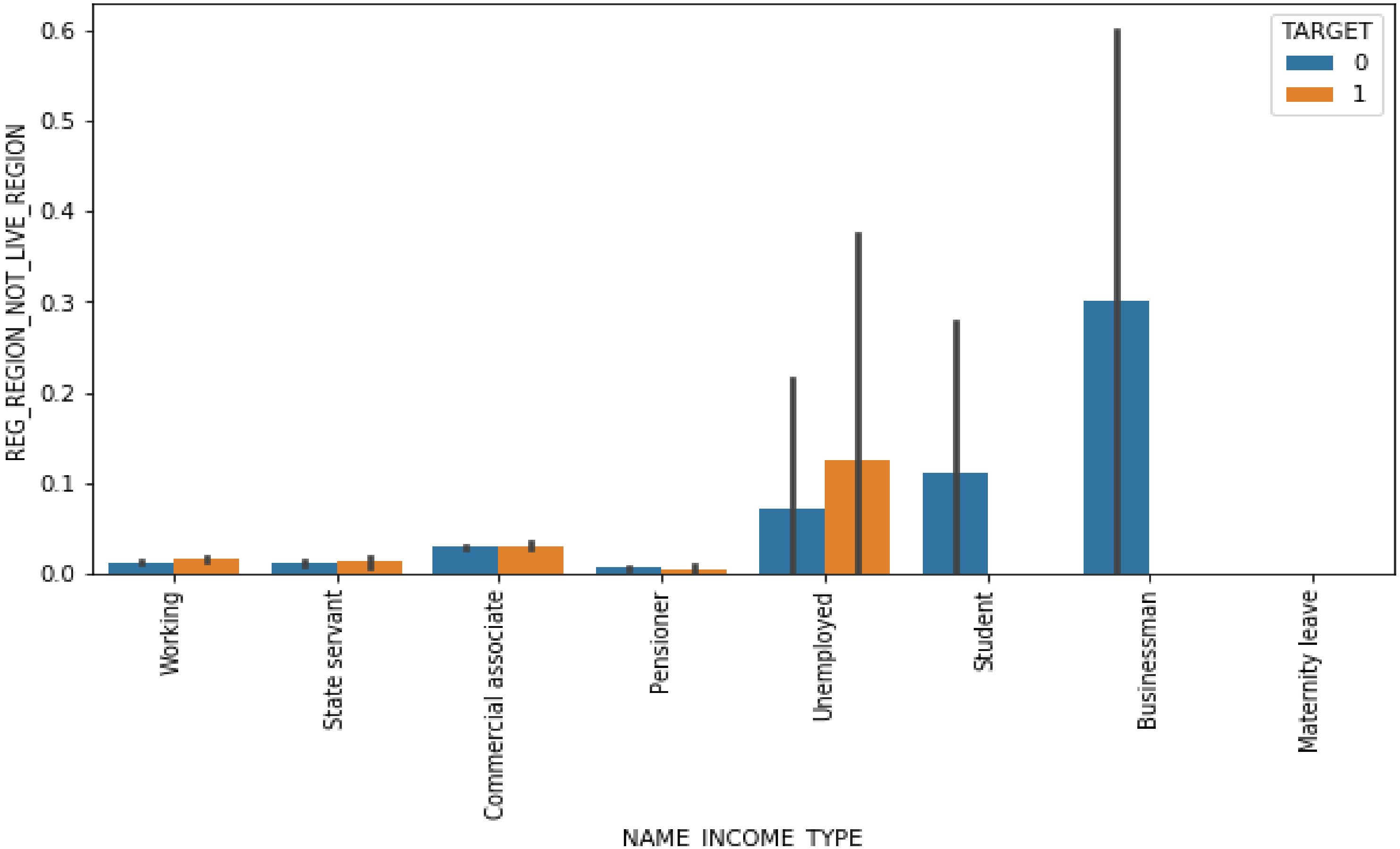
People who getting income via Maternity Leave tends to be more Defaulter when they have more children.

❑ Income vs No. of Family Members



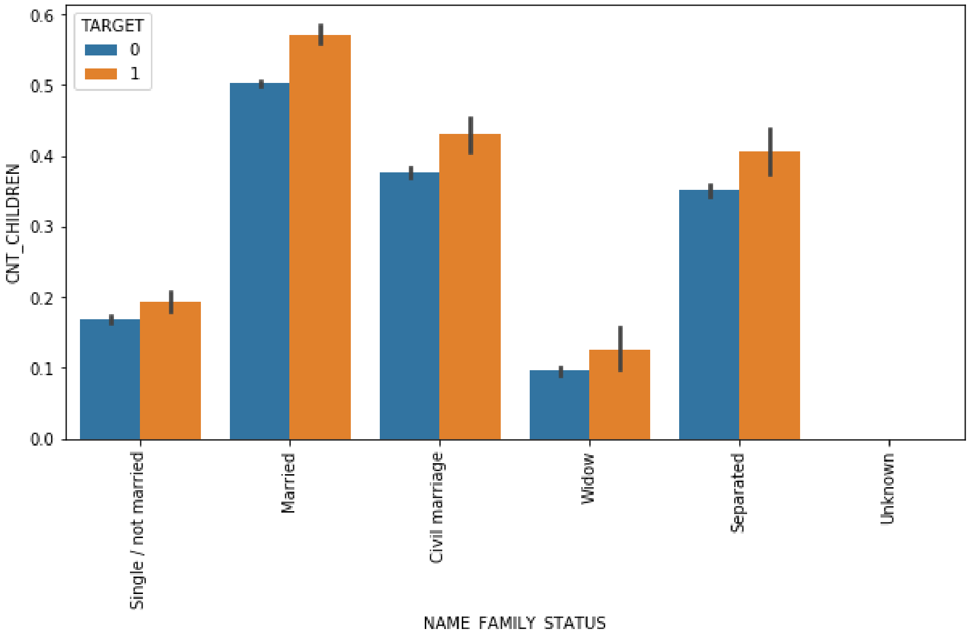
People who getting income via Maternity Leave tends to be more Defaulter when they have more children.

❑ Income Type vs Client whose Permanent Address not match with Contact Address at Region Level



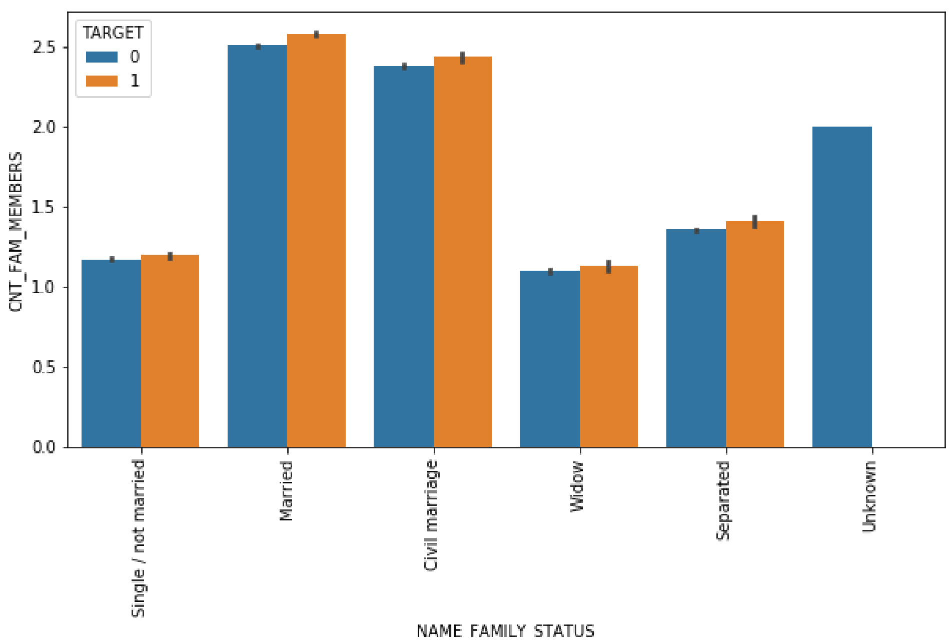
Unemployed Client are has more chance to be a defaulter, when their Permanent Address does not match with the Contact Address in the Regional Level

❑ Family Status vs Count Of Children



Married client and has more children (5+), chances to be a defaulter in High.

❑ Family Status vs Count Of Family Members



Married client and has more children (5+), chances to be a defaulter in High.

Correlation Of Target Variable Vs. Other Variables

```
Correlation.head(6)["TARGET"][1:]
```

```
REGION_RATING_CLIENT_W_CITY    0.060893
REGION_RATING_CLIENT           0.058899
DAYS_LAST_PHONE_CHANGE         0.055218
DAYS_ID_PUBLISH                0.051457
REG_CITY_NOT_WORK_CITY         0.050994
Name: TARGET, dtype: float64
```

```
Correlation.tail(5)["TARGET"]
```

```
AMT_CREDIT                   -0.030369
REGION_POPULATION_RELATIVE   -0.037227
AMT_GOODS_PRICE              -0.039628
AGE                          -0.078263
EXT_SOURCE_2                 -0.160303
Name: TARGET, dtype: float64
```

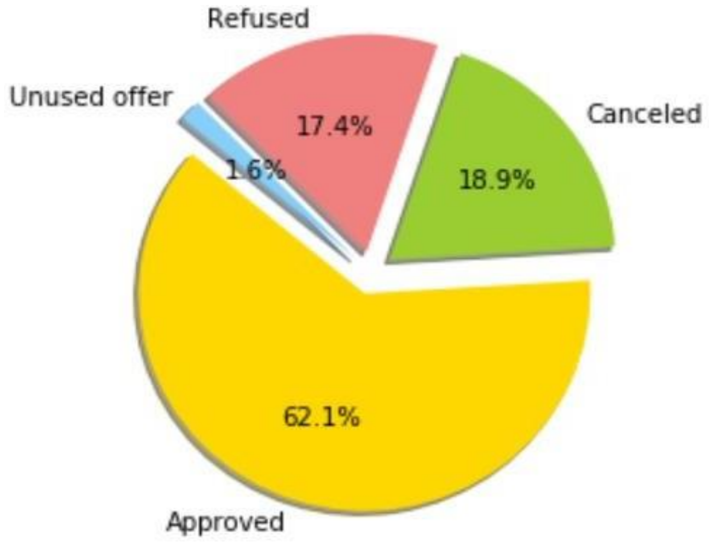
Highly Correlated Variables

1. AMT_CREDIT and AMT_GOODS_PRICE =0.99
2. REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT = 0.95
3. CNT_FAM_MEMBERS and CNT_CHILDREN = 0.87
4. AMT_ANNUITY and AMT_CREDIT = 0.77

Previous Application Analysis

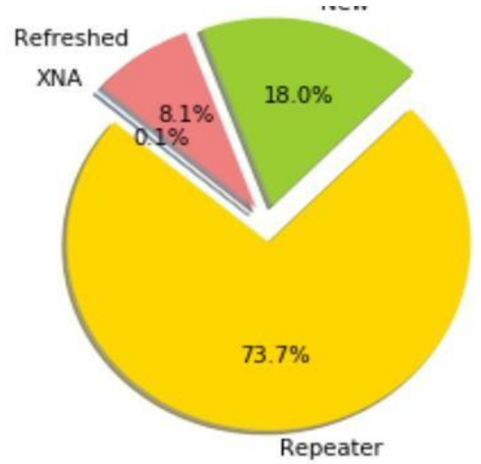
Analysis of the second data set. Data cleaning and analysing the data.

❑ **Based on Contract Status**



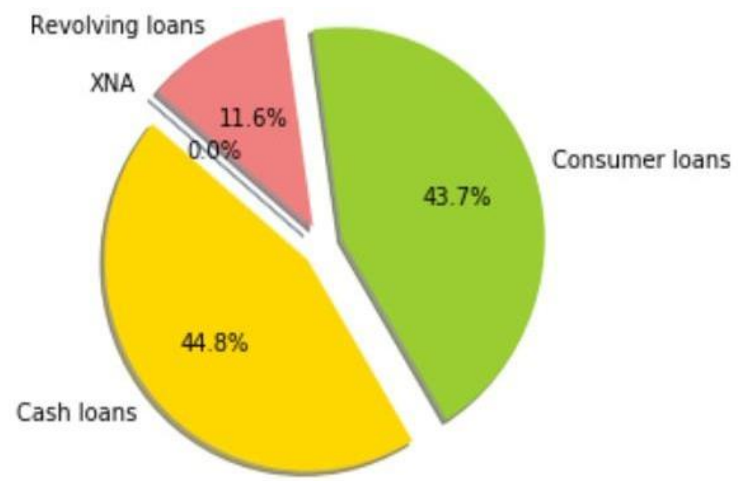
- Approved: 62.1 %
- Cancelled: 18.9 %
- Refused: 17.4 %
- Unused offer: 1.58 %

❑ Based on Client Type

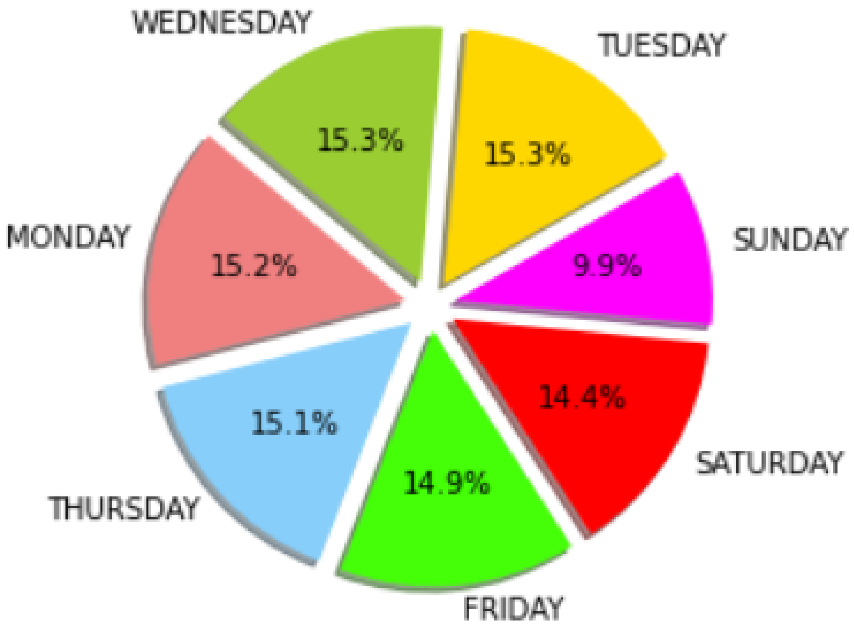


73.4% applicants are repeaters. Only, 18.4% are new clients.

❑ Based on Contract Type

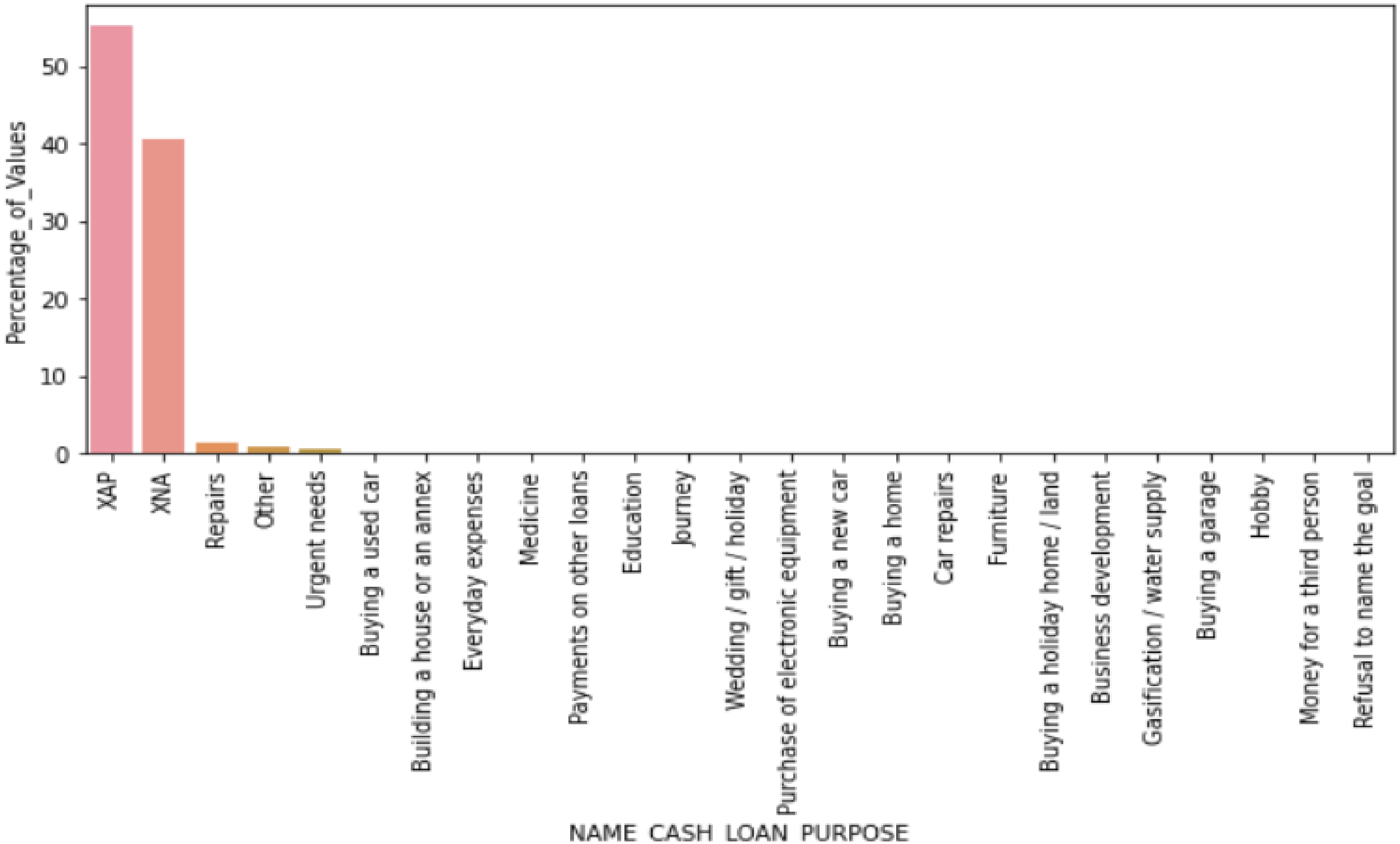


❑ Based on Days of Approval



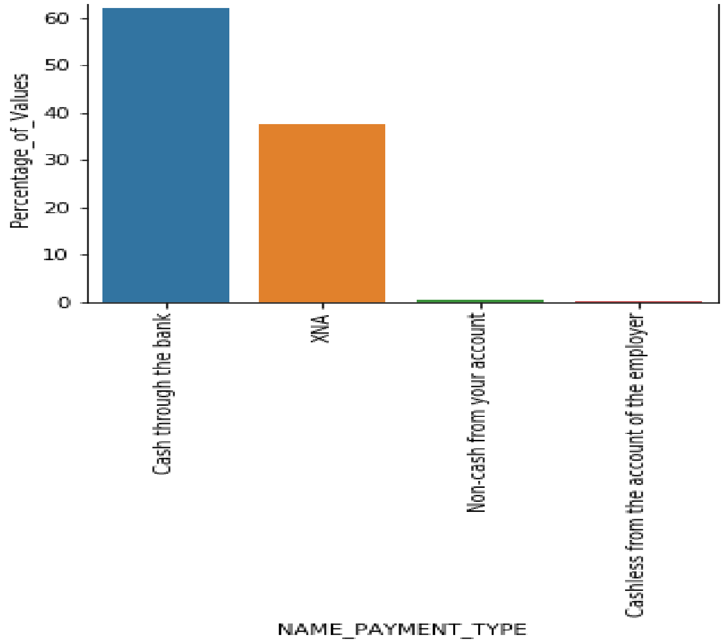
Most of the clients have opted to apply loan on Tuesday. Applicants are very low on weekends.

❑ Based on Purpose of Loan



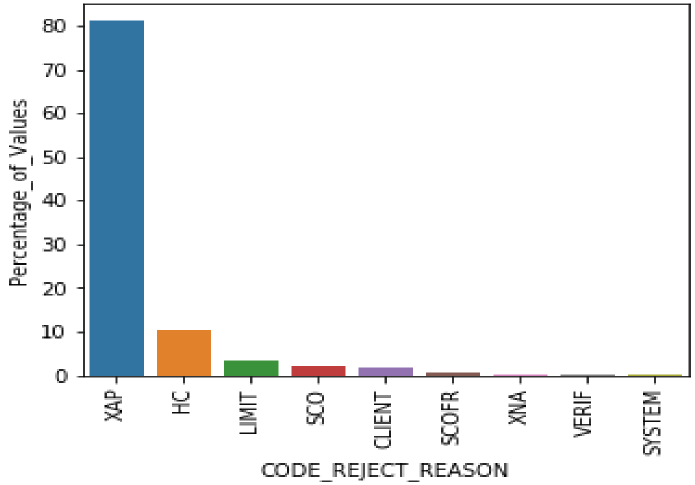
Most Loan purpose was not recorded. **XAP and XNA** values are highest.

❑ Based on Payment Type



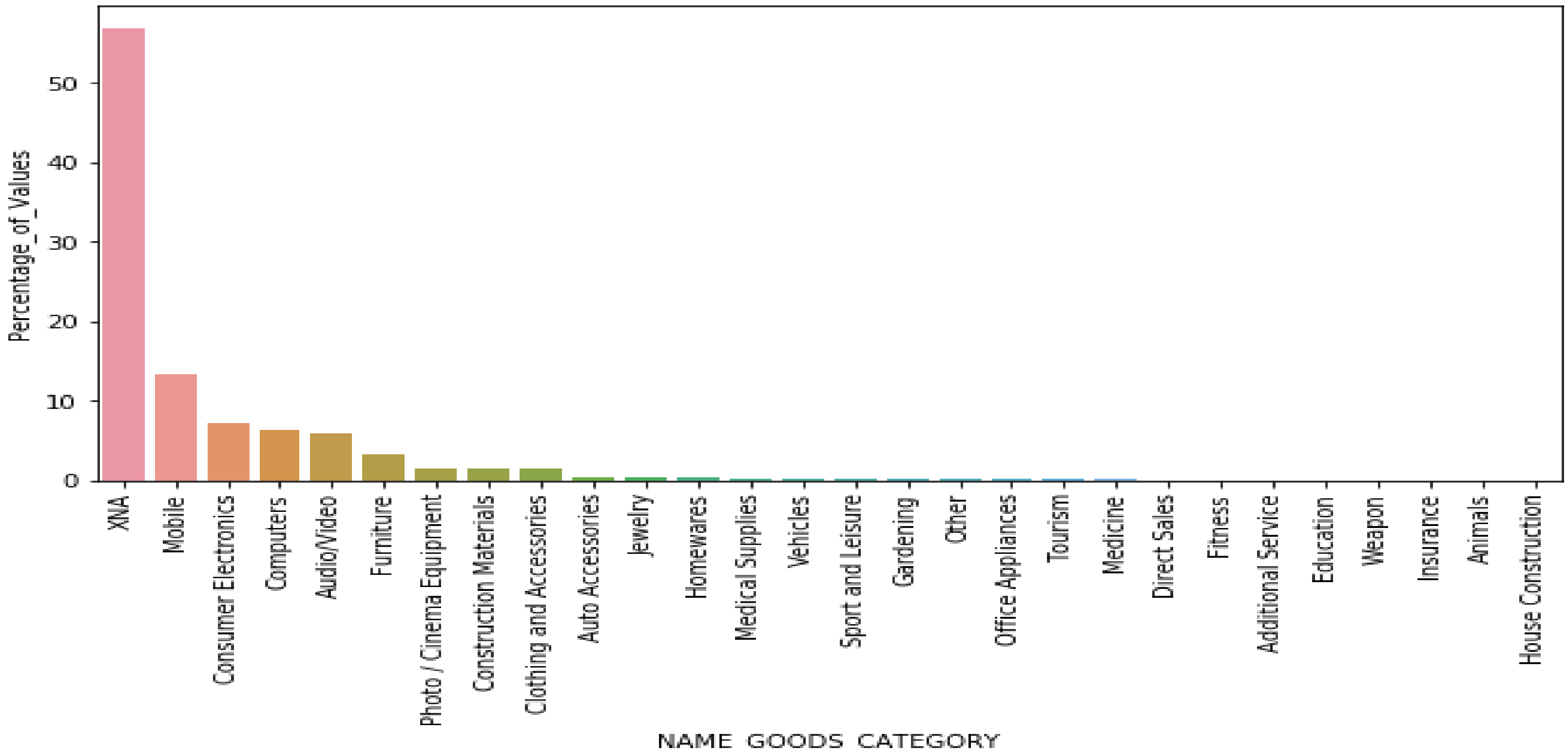
Most people preferred **CASH(62.44%)** as the mode of Payment

❑ Based on Reason of rejection of loan



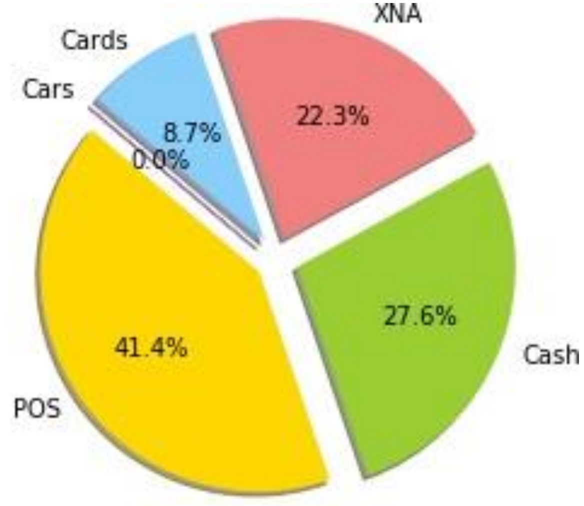
Primary reason for the Loan to get rejected is not recorded(**XAP (81%)**) followed by **HC**.

❑ Based on previous application **NAME_GOODS_CATEGORY**



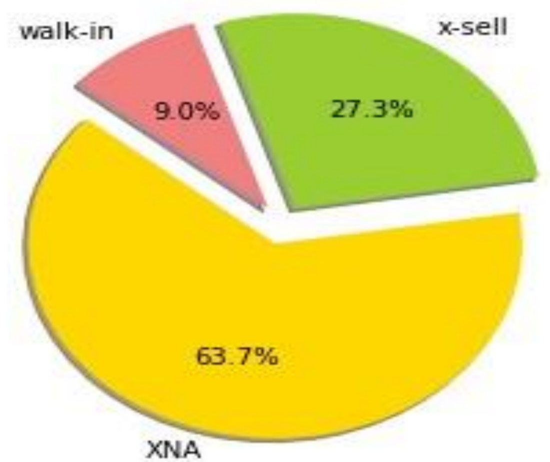
Most clients applied for Mobile and 53.96% of the data is not recorded(XNA).

- Based on previous application **NAME_PORTFOLIO**



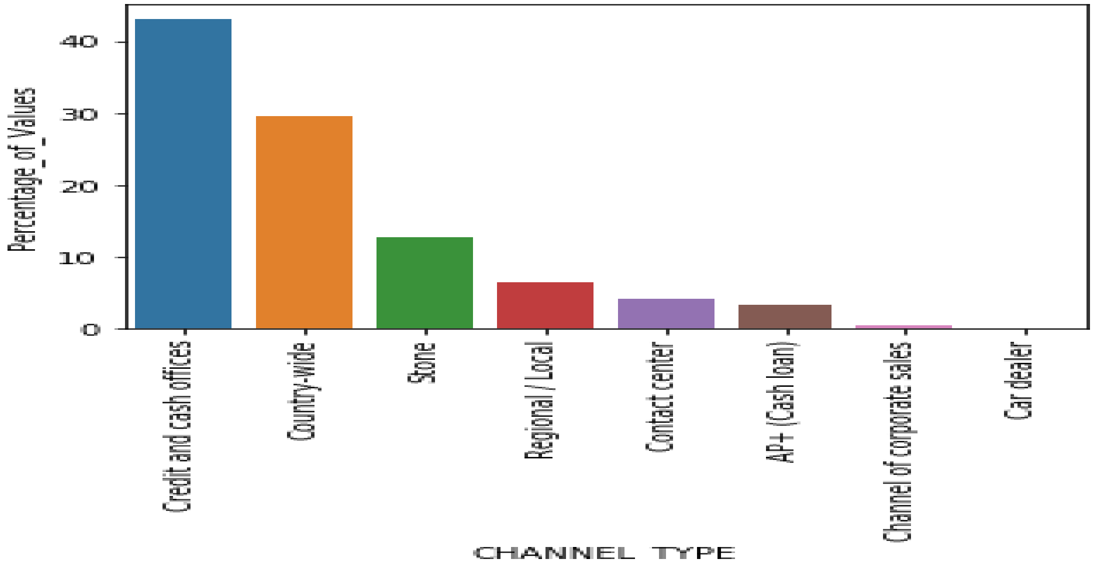
41.4% of the applications were for POS.

- Based on previous application **NAME_PRODUCT_TYPE**



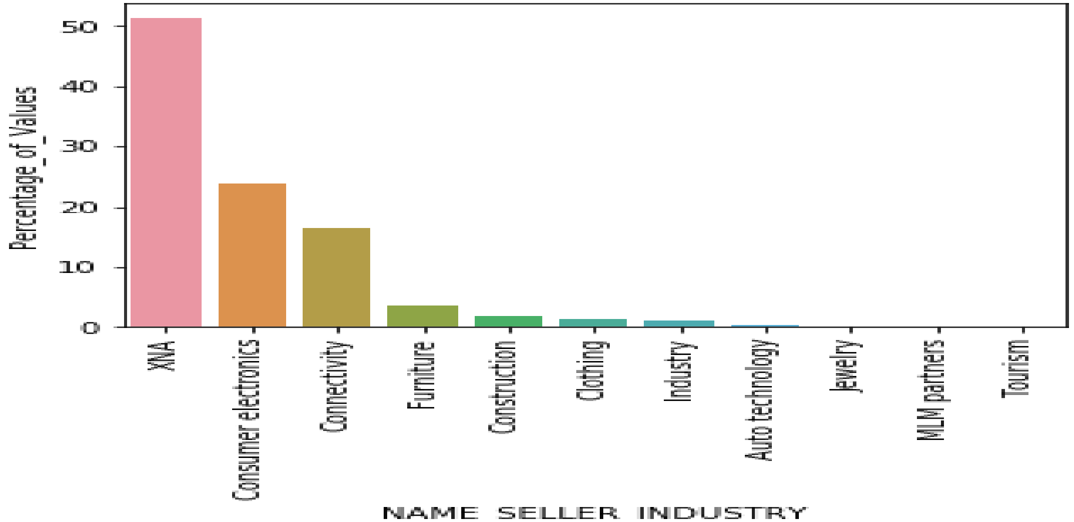
X-sell applications were more than walk-in

- Based on previous application **CHANNEL_TYPE**



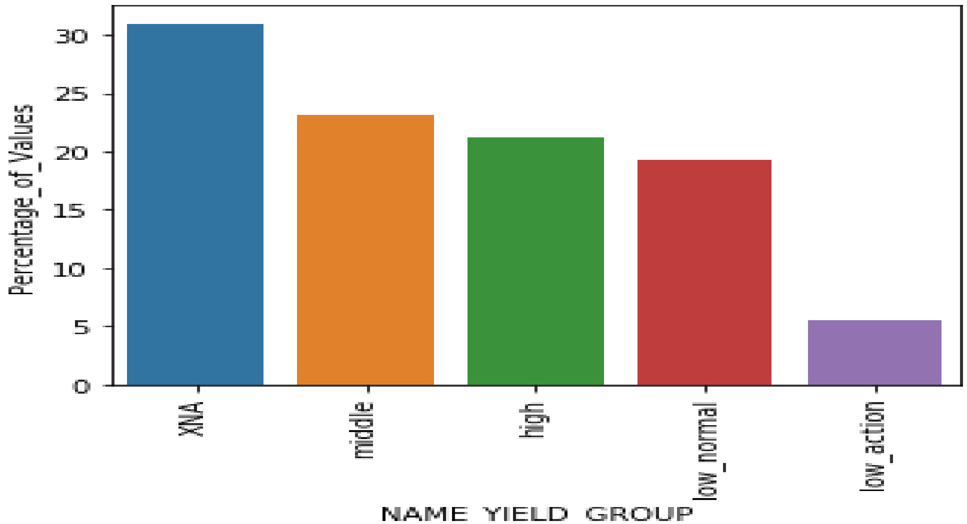
Most clients were asquired from **Credit and Cash Offices**

- Based on N**AME_SELLER_INDUSTRY**



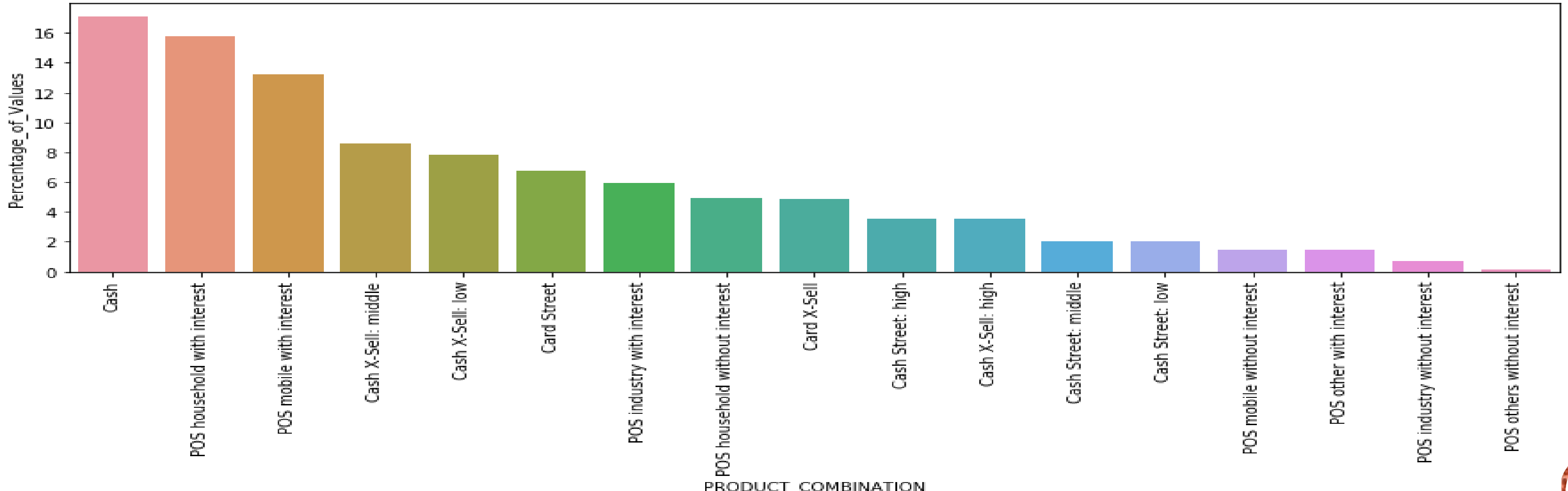
Most Sellers are from **Consumer electronics**

❑ Based on previous application **NAME_YIELD_GROUP**



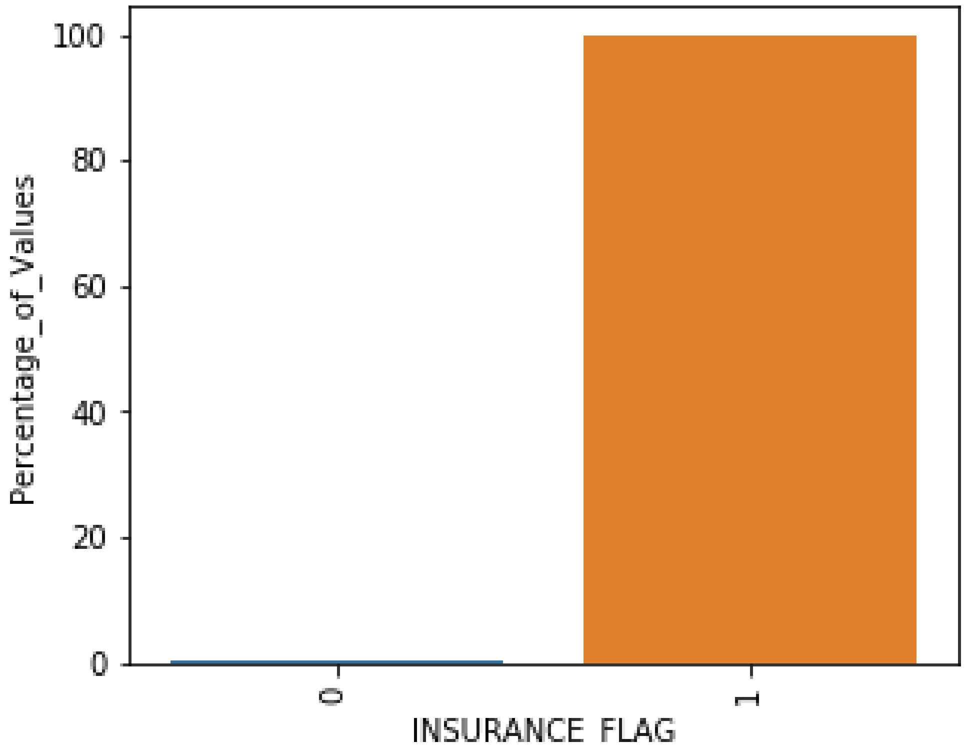
Most group interest rates lie in middle.

❑ Based on **PRODUCT_COMBINATION**



Highest product combination is **Cash** followed by **POS household with interest**

❑ Based on **NFLAG_LAST_APPL_IN_DAY**



For most clients it was the last application of the day.

Merging Application Data and Previous Application

After analysing all the previous and applications data, the correlation of the variable with respect to the Target variable is as following

TOP COORELATION VARIABLES

DAYS_LAST_PHONE_CHANGE	0.059721
REGION_RATING_CLIENT_W_CITY	0.059700
REGION_RATING_CLIENT	0.056932
DAYS_ID_PUBLISH	0.051037
REG_CITY_NOT_WORK_CITY	0.049353

LOW COORELATED VARAIBLES

HOUR_APPR_PROCESS_START	-0.027809
AMT_GOODS_PRICE	-0.032550
REGION_POPULATION_RELATIVE	-0.035028
AGE	-0.074927
EXT_SOURCE_2	-0.154919

THANK YOU !