Credit EDA Assignment

By: Satyendra

INTRODUCTION

For the purpose of this case study, two data sets were provided, namely:

- Application Data
- Previous Application Data

First, we took the Application Data dataset for analysis.

<u>Data cleaning</u> was done before analysis. Following were the steps followed:

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

```python
#method 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 more than **30% NaN** values

```python
# Iterate over columns in DataFrame and delete those with where >30% of the values are null
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. Post these actions, we decided on imputing values on few columns to further make the data set usable.

```python
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	

As you can see from above screenshots, we have decide to impute mean values to the **AMT_GOODS_PRICE** and **EXT_SOURCE_2** columns.

4. We then further decided to impute the mode values to the **NAME_TYPE_SUITE** column

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

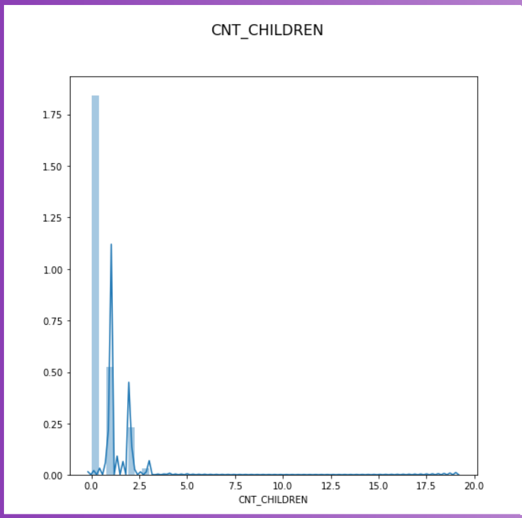
```
Unaccompanied      248526
Family              40149
Spouse, partner     11370
Children             3267
Other_B              1770
Other_A               866
Group of people       271
Name: NAME_TYPE_SUITE, dtype: int64
```

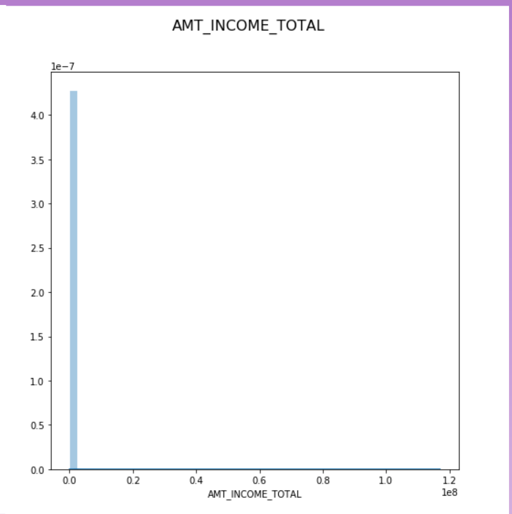
```
#Here "Unaccompanied" data has the highest mode.We can fill missing values  with Unaccompanied

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

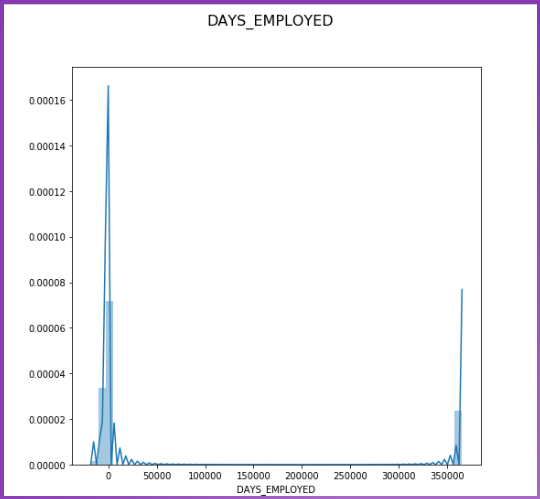
<u>Find outliers in the given data frame</u>

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

Here are some of the values we could spot as outliers with the help of plots:







Above plot for **CNT_CHILDREN** show a large outlier (19). Since a family cannot or very rarely have 19 children.

In the **DAYS_EMPLOYED** there is a value present at 36k range, this won't be possible. This error could have occurred during data entry

In the plot **AMT_INCOME_TOTAL**, we can visually see that the MAX amount is way larger than the other statistical data [Mean, (25,50,75) percentiles]
Now that we have identified the ouliers, we have removed them and plotted them again to observe the difference.

Further more we have done some modification of values in oreder to make the analysis of data easier.

Converted Date of birth to age and also did binning of salaries into High, Medium and Moderate Levels.

ANALYSIS OF APPLICATION DATA

Then we proceeded with the analysis of data.

Divided data into two separate dataframes with defaulter and good clients.

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

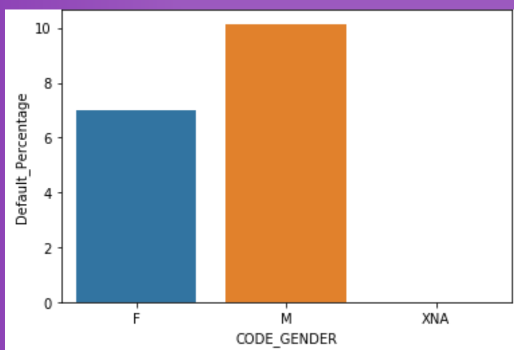
Target value 0 indicates that the client is not a defaulter thus a good client.

Target value 1 indicates client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample.

- **<u>Univariate Analysis of Categorical and Numerical Data</u>**

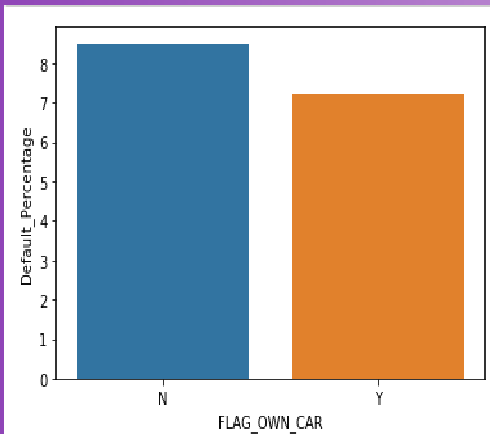
  Checked for clients that are likely to be defaulters/ unlikely to pay back the loan by analysing various columns in the data frame.

  - Based on **CODE_GENDER** (gender of client)



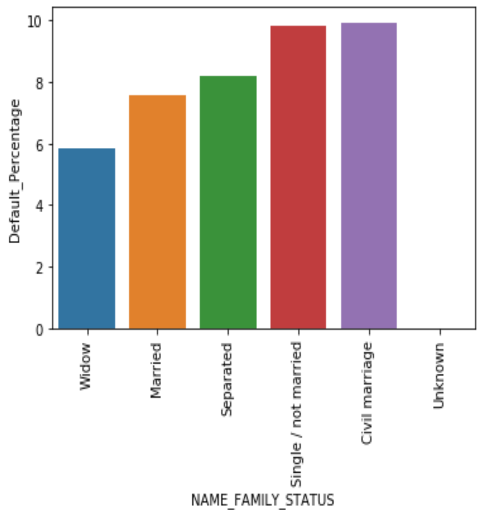
So, from above plots and data we can cleary see that the Female clients are a better TARGET as compared to the Male clients. Observing the percent of defaulted credits, male client have a higher chance of not returning their loans [10.14%], compared to the female clients [7%].

  - Based on **FLAG_OWN**_CAR (client owns a car or not)



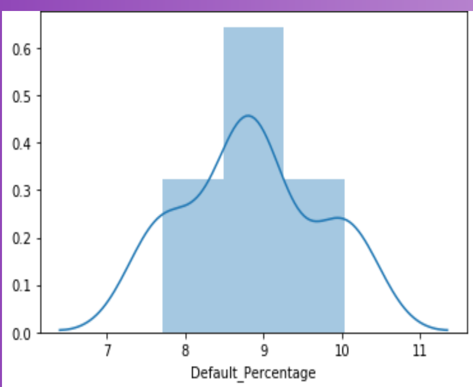
As we can see from above graph,the clients that own a car are less likely to not repay the loan when compared to the ones that do not own a car. The loan non-repayment rates of both the Car Owners and Non-Car Owners are very close. Which is interesting to see and indicates that probably this metric will not be a suitable one when targeting a client.

- **Based on NAME_FAMILY_STATUS**



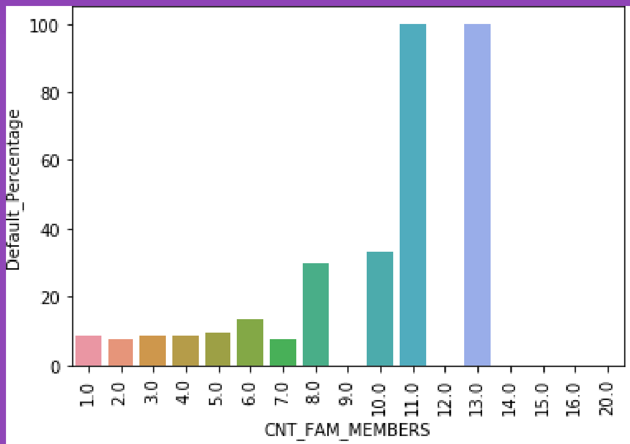
From above graph we can say that the percentage of non-repayment of loan is at highest for civil mariage and is lowest for widows. Which is interesting to see because you expect widows to not payback their loans but it is the opposite here.

- Based on **CNT_CHILDREN**



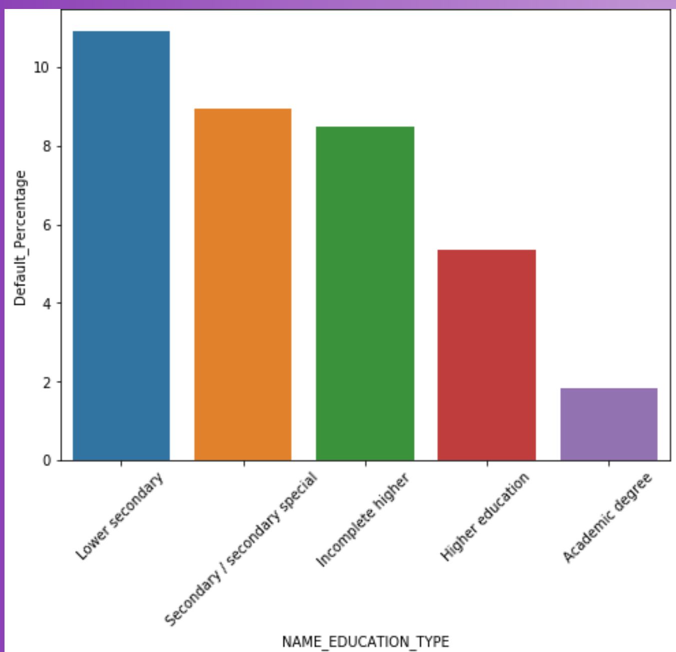
Ther is more chance for a client with more children to not repay the loan back. This can be beacuse of the more liability that is on the client. The more the number of children the more difficult it is for the client to repay the loan due to more personal expenditures.

- Based on **CNT_FAM_MEMBERS**



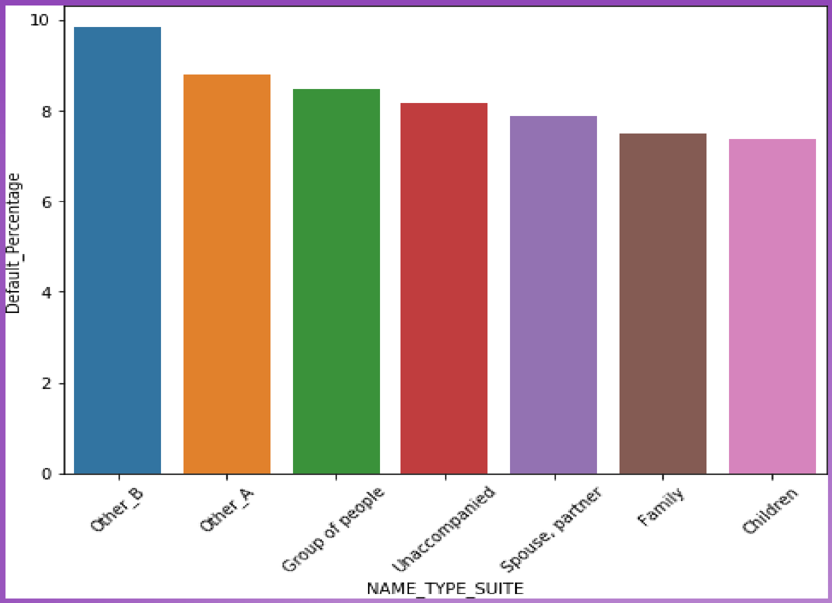
> Though we can see that family with 11,13 members shows highest default rate, but their count is very less[2].

- Based on **NAME_EDUCATION_TYPE**



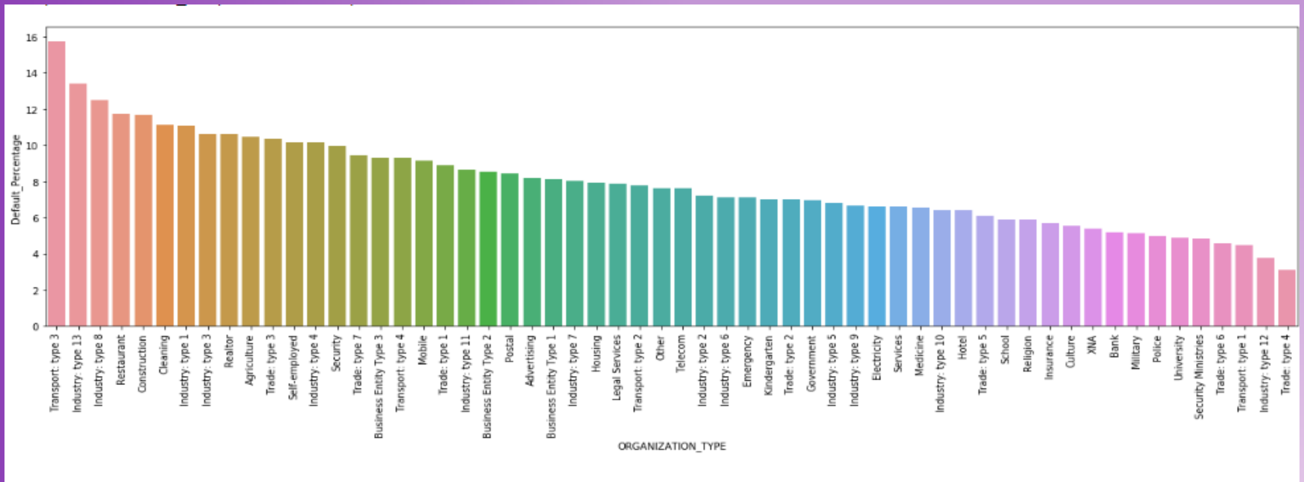
It can be seen from above graph that the more educated clients are likely to repay their loans because they will be having more stable jobs with monthly income.

- Based on **NAME_TYPE_SUITE** (Who was accompanying client when he was applying for the loan)



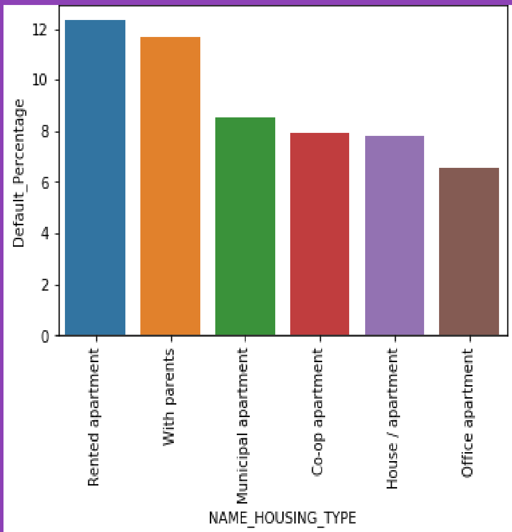
Most clients who were occupied by Other_B followed by Other_A are unlikely to pay back their loans.

- Based on **ORGANISATION_TYPE**



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

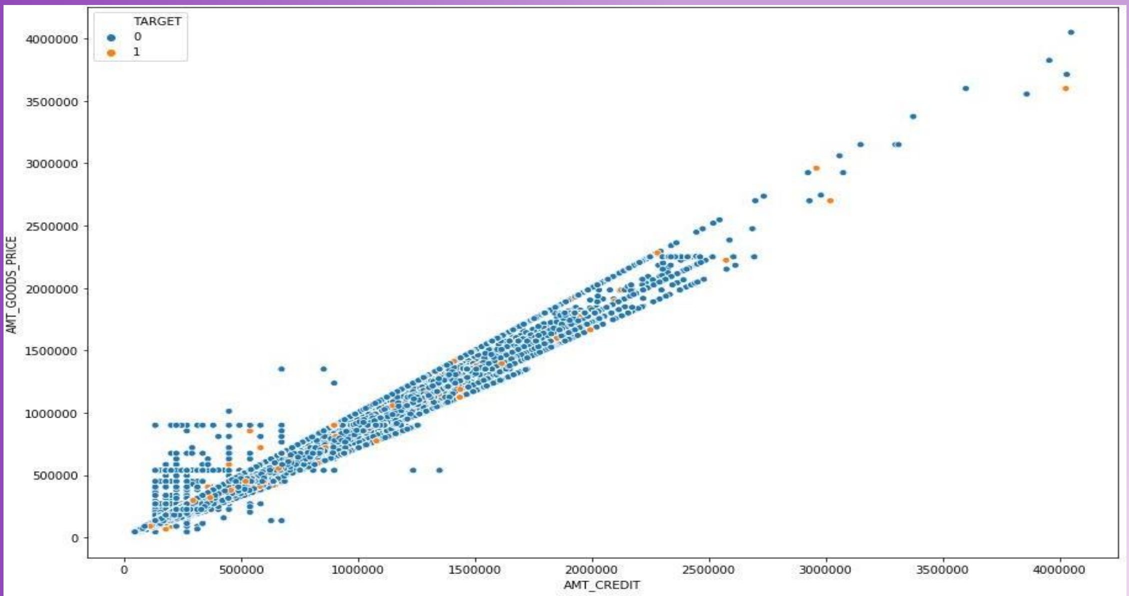
- Based on **NAME_HOUSING_TYPE**



From above graph it can be seen clearly that people with rented apartments are less likely to pay back their loans. This can be because they already have more liabilities compared to other type of people who do not have thia liability.

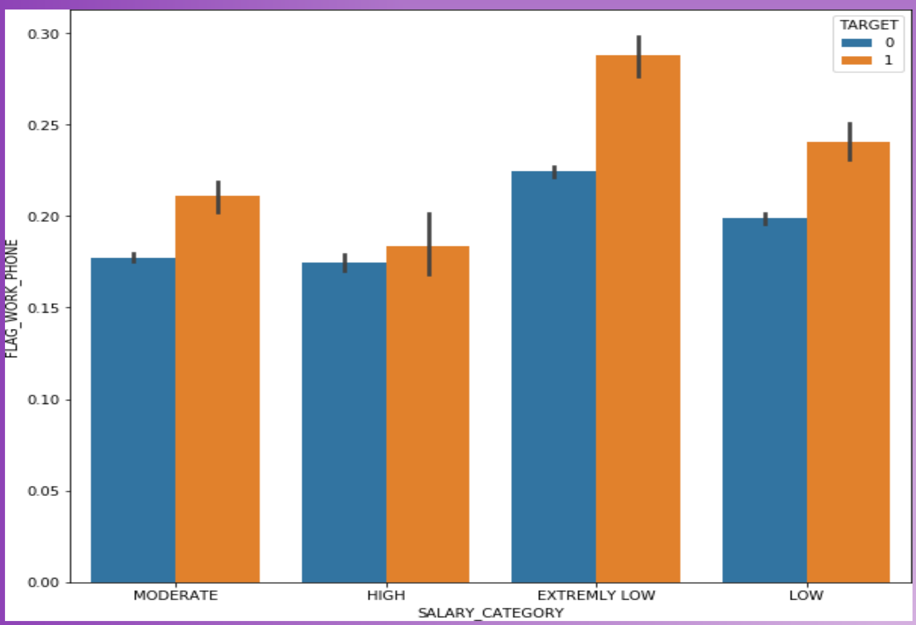
- **<u>BIVARIATE ANALYSIS</u>**

<u>AMT_CREDIT vs AMT_GOODS_PRICE</u>



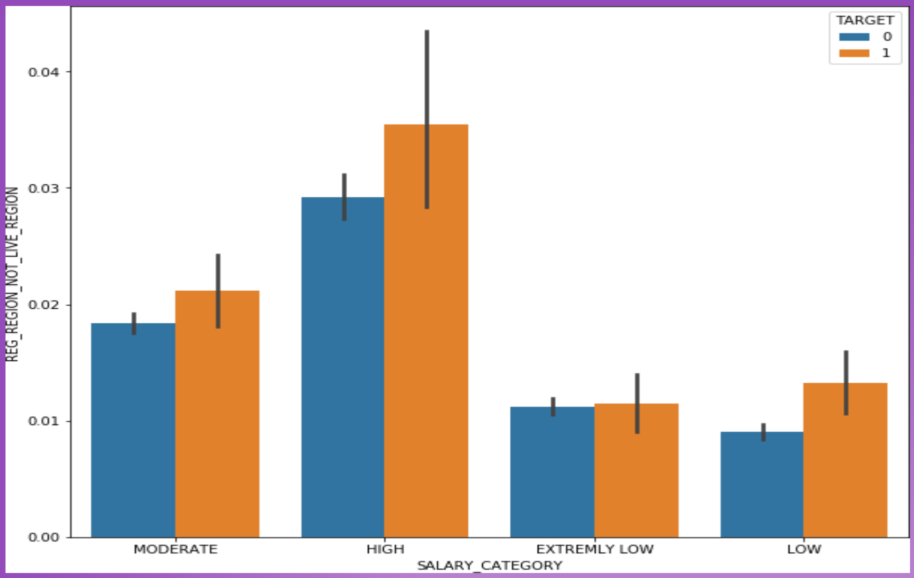
We found that Credit amount and the Amount goods price are more correlated with the Defaulters. The Defaulters are linearly increasing as these both variable increases.

Salary Category vs Clinet who provided Home Number



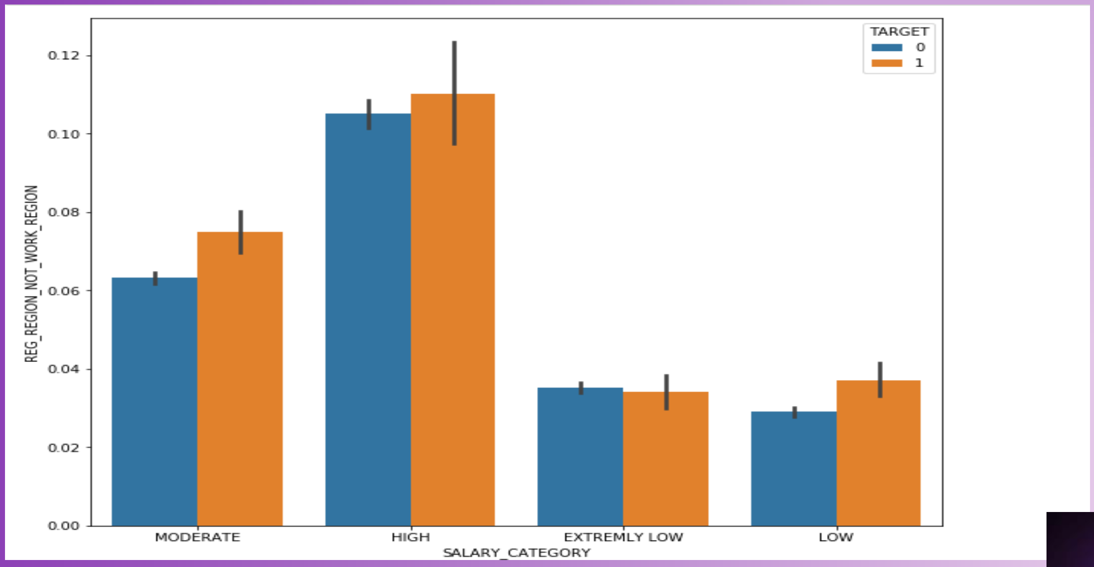
Client with Extremly low salary has more chance to be a Defaulter, when he did not provide the Home phone number. Here approximately 30% people only produced the phone number

SALARY VS CLIENT WHOSE PERMANENT ADDRESS NOT MATCH WITH CONTACT ADDRESS -REGION LEVEL.



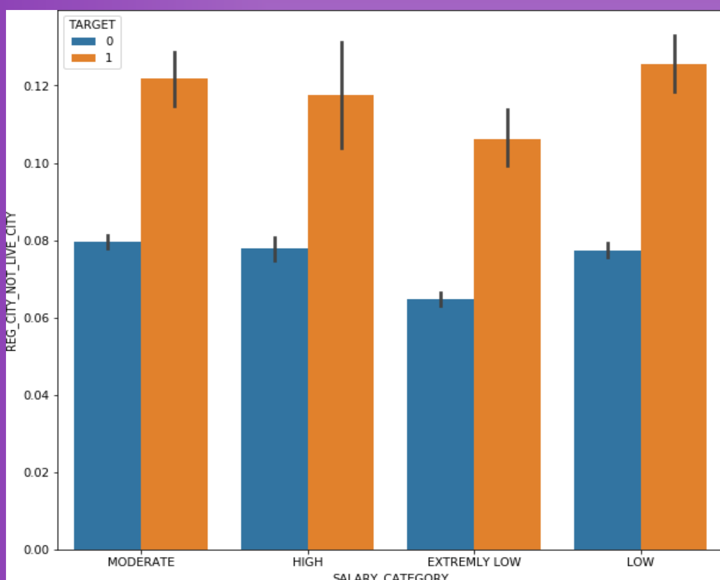
When Client gets Extremly lower salary and if his/her Contact address doest match, then there is a Higher chance for him/her to be defaulter

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



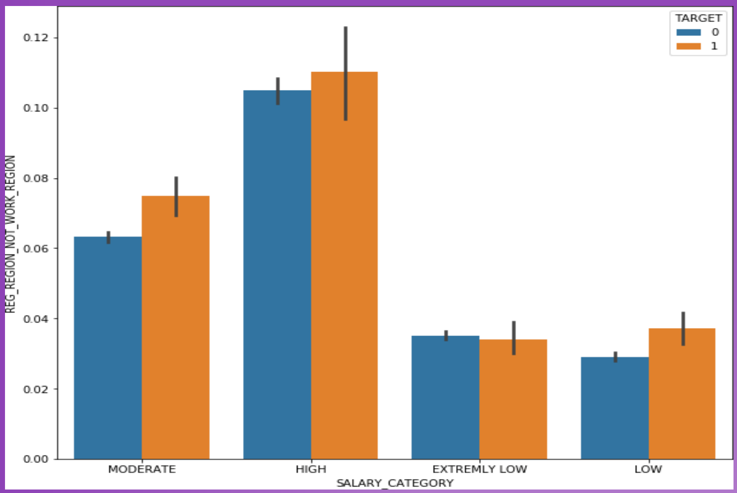
When Client gets Extremply lower salary and if his/her Work address doest match, then there is a Higher chance for him/her to be defaulter.

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



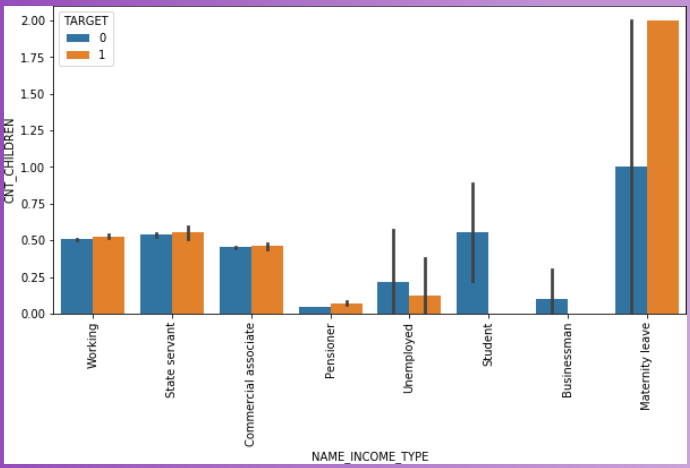
When Client gets LOWER salary and if his/her CONTACT address(CITY-LEVEL)doest match, then there is a Higher chance for him/her to be defaulter.

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



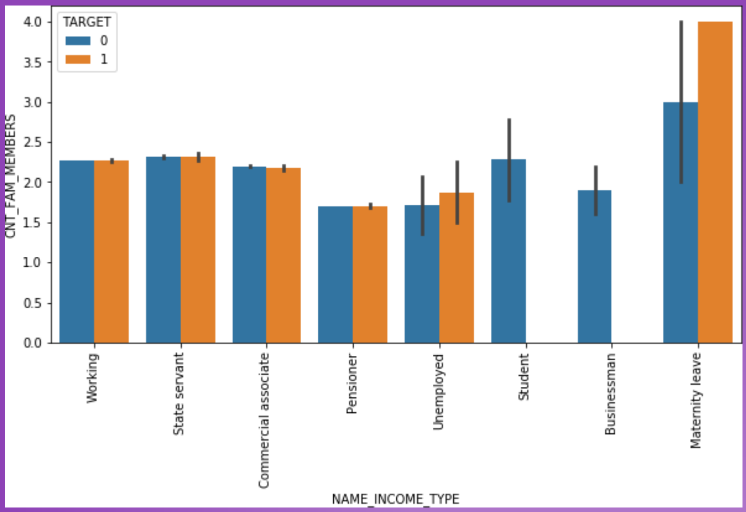
When Client gets HIGH salary and if his/her WORK address(CITY-LEVEL)doesn't match, then there is a Higher chance for him/her to be defaulter.

INCOME vs CHILDREN Count



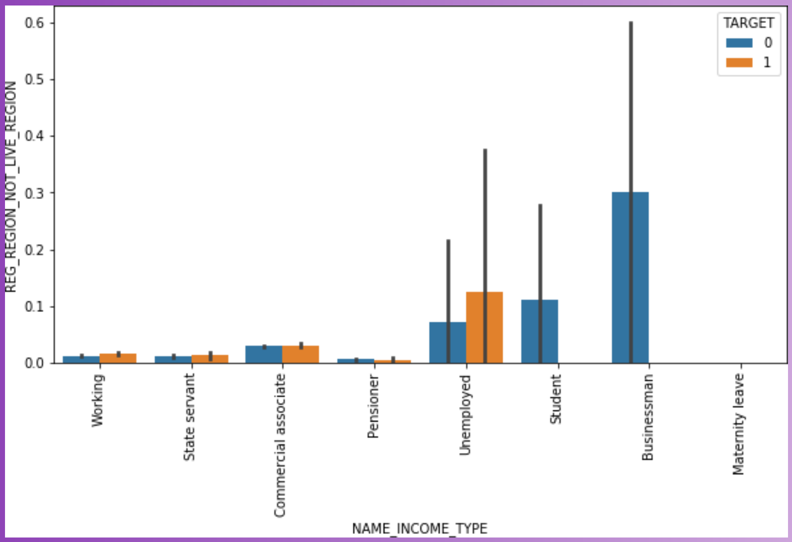
People who getting income via Maternity Leave tends to be more Defaulter when they have more children.

Income vs No.of.FamilyMembers



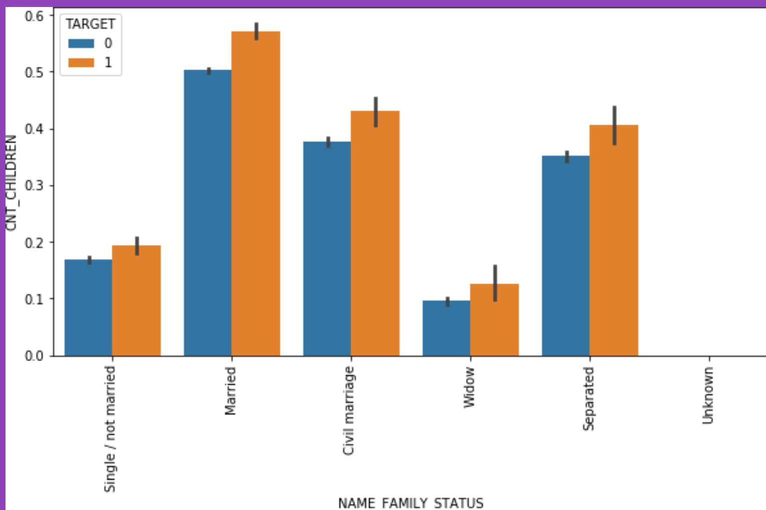
People who getting income via Maternity Leave tends to be more Defaulter when they have more Family Members.

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



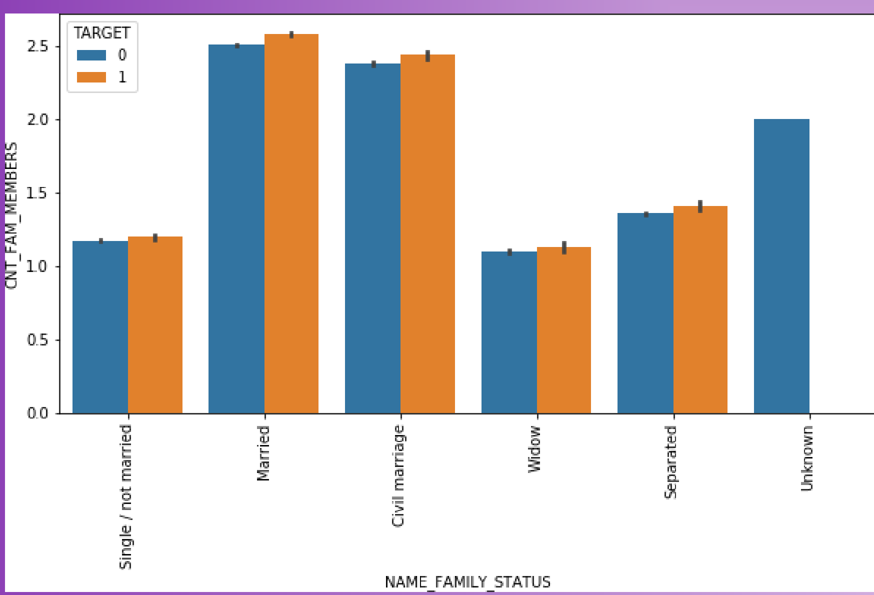
Client who are Unemployed has more chance to be a defaulter , when their Permanent Address does not match with the Contact Address in the Regional Level

Family Status vs Count Of Children



Client who are married and has more children (5+), chances to be a defaulter in High. This may be due to the Economic situation of their family, because of more children.

Family Status vs Count Of Family Members



Client who are married and has more children (5+), chances to be a defaulter in High. This may be due to the Economic situation of their family, because of more children

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
```

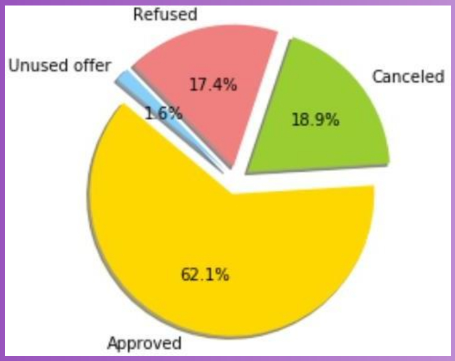
Highly Correlated Variables

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

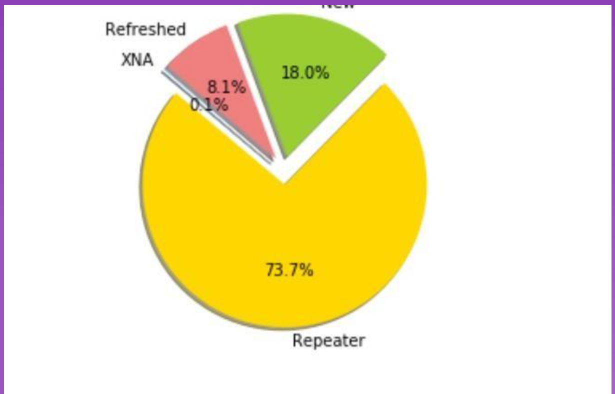
Then we moved on to analysis of the second data set. We performed few data cleaning steps and then moved on to analyzing the data.

- **Based on Contract Status**



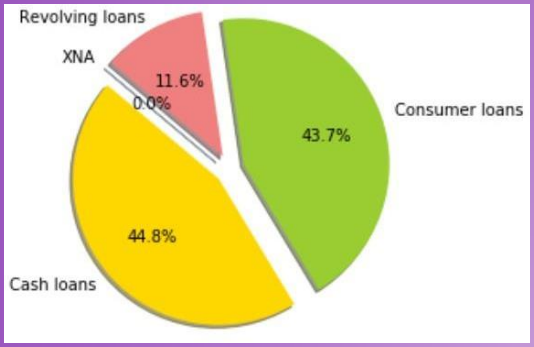
- Approved: 62.1 %
- Cancelled: 18.9 %
- Refused: 17.4 %
- Unused offer: 1.58 %

- **Based on Client Type**

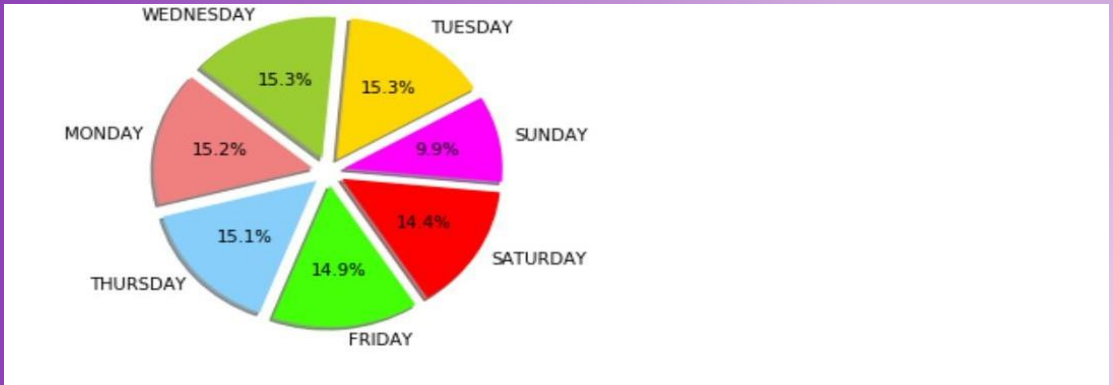


73.4% applicants are repeaters. Only, 18.4% are new clients.

- **Based on Contract Type**

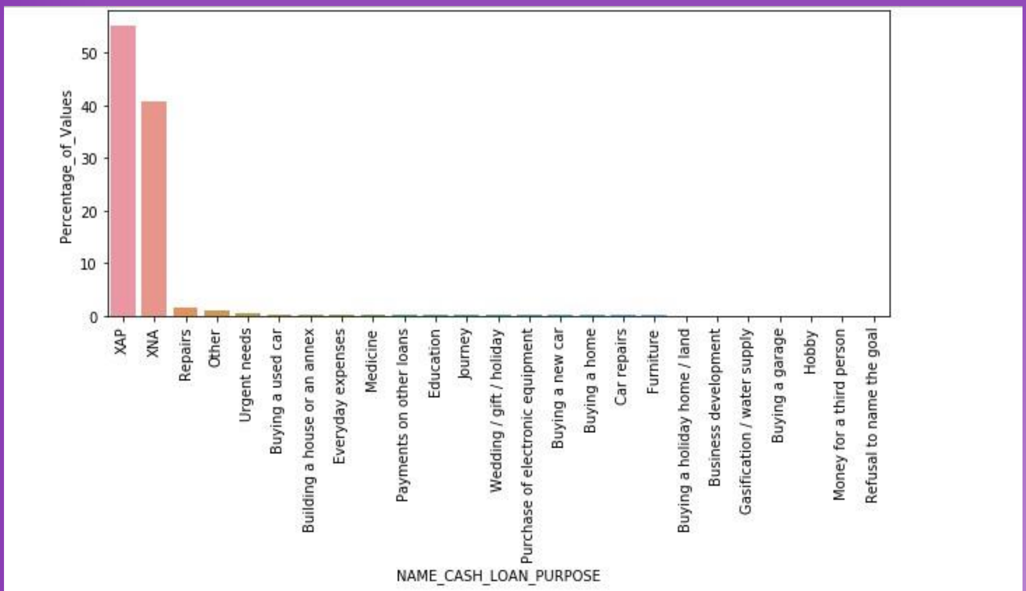


- **Based on Days of Approval**



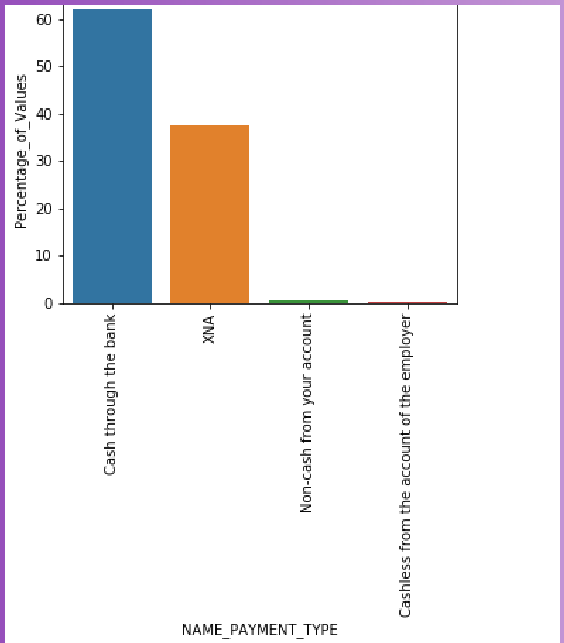
Most of the clients have opted to apply loan on Tuesday. It is very interesting to see that applicants are very low on weekends. We would otherwise assume that the applicants would prefer weekends to apply.

- **Based on Purpose of Loan**



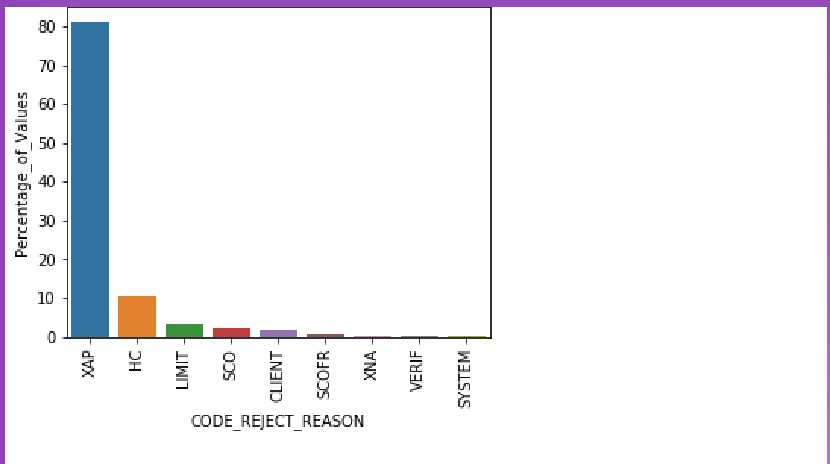
Most Loan purpose was not recorded. **XAP and XNA** values are highest.

- **Based on Payment Type**



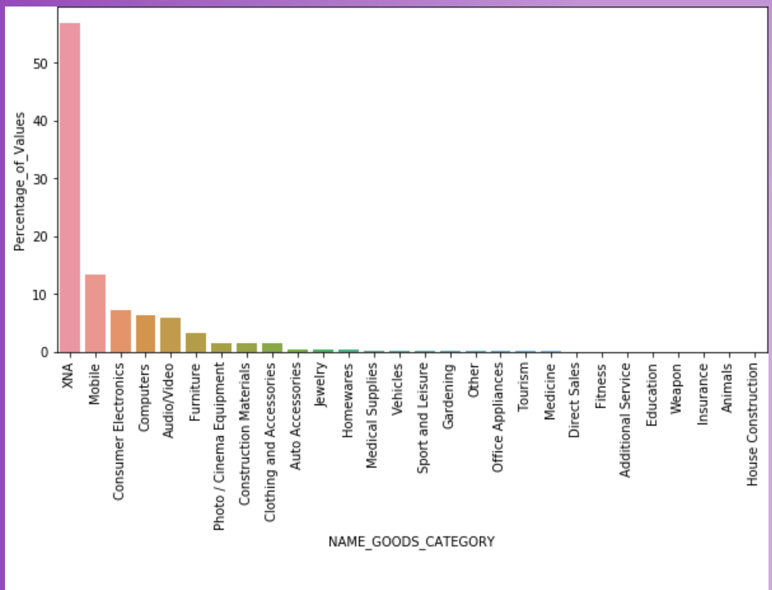
Most people preferred **CASH(62.44%)** as the mode of Payment

- **Based on Reason of rejection of loan**



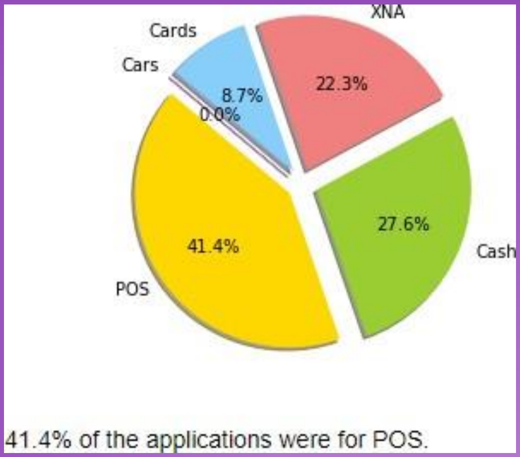
Primary reason for the Loan to get rejected is not recorded(**XAP (81%)**) followed by **HC**.

- Based on What kind of goods did the client apply for in the previous application - **NAME_GOODS_CATEGORY**



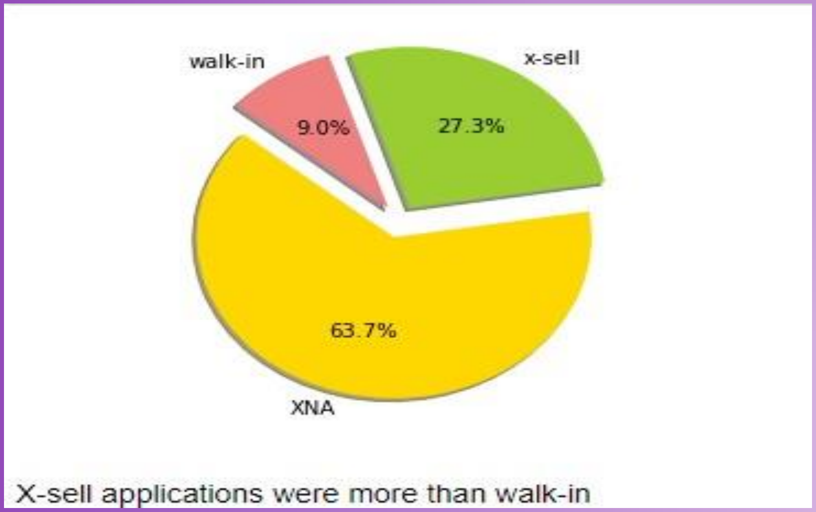
Most clients applied for Mobile and 53.96% of the data is not recorded(XNA).

- Based on Was the previous application for CASH, POS, CAR, … - **NAME_PORTFOLIO**



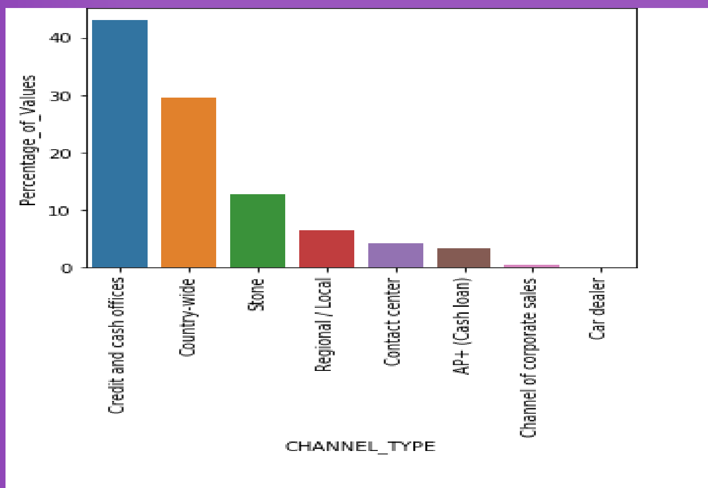
41.4% of the applications were for POS.

- Based on Was the previous application x-sell or walk-in - **NAME_PRODUCT_TYPE**



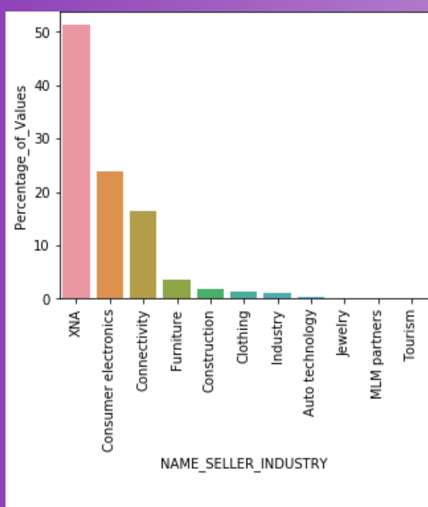
X-sell applications were more than walk-in

- Based on Through which channel we acquired the client on the previous application - **CHANNEL_TYPE**



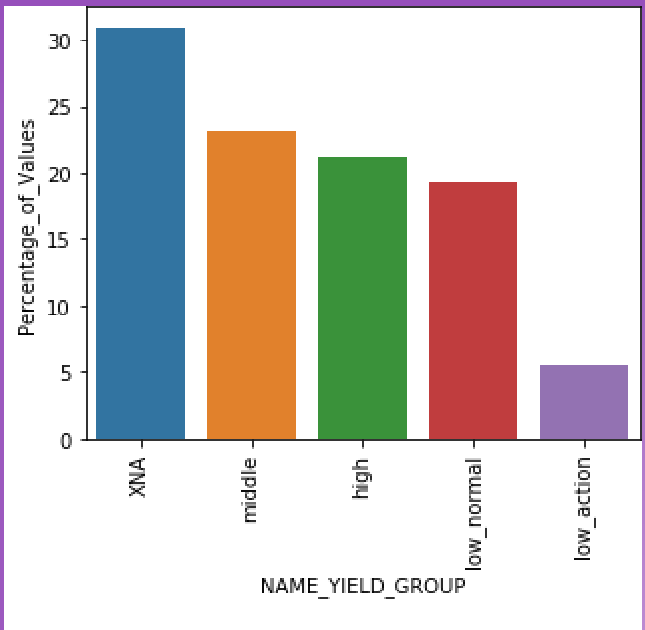
Most clients were asquired from **Credit and Cash Offices**

- Based on The industry of the seller - N**AME_SELLER_INDUSTRY**



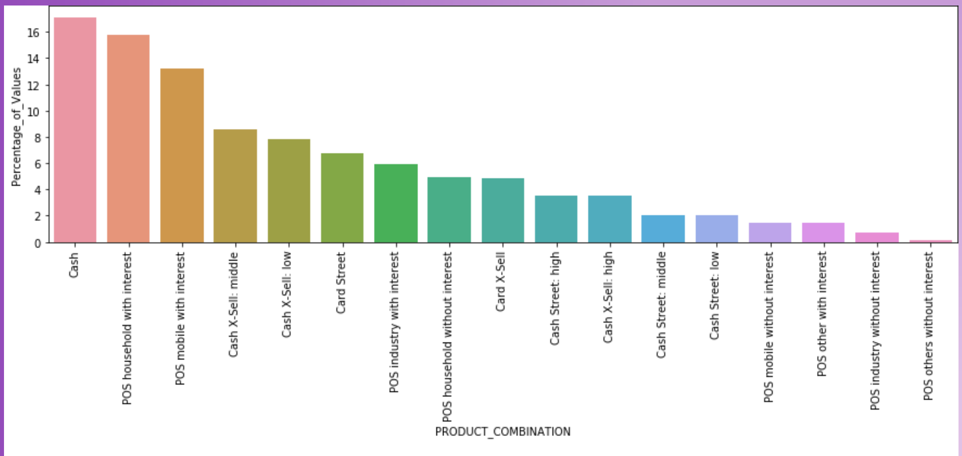
Most Sellers are from **Consumer electronics**

- Based on Grouped interest rate into small medium and high of the previous application - **NAME_YIELD_GROUP**



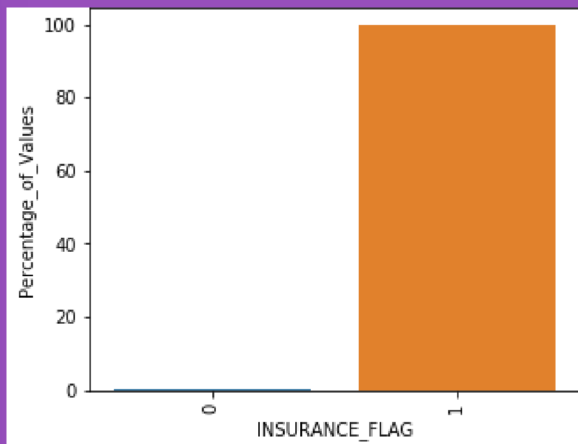
Most group interest rates lie in middle.

- Based on **PRODUCT_COMBINATION**



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

- Based on Flag if the application was the last application per day of the client - **NFLAG_LAST_APPL_IN_DAY**



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

MERGING APPLICATION DATA AND PREVIOUS APPLICATION

After analyzing all the previous and current applications, we once again checked the correlation of the variable with respect to the Target variable. We got the following results.

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
```

Mostly the variables are more or less familiar, as we seen in our Application data, that has been contributing more to the **DEFAULTERS** prediction.