CREDIT EDA CASE STUDY

Prepared by:

Joel Lukose Johnson

INTRODUCTION

Two data sets were provided for the purpose of this case study namely:

- Application Data
- Previous Application Data

First, I took the Application Data dataset for analysis.

Data cleaning was done before analysis. Following were the steps done to receive cleaned data:

1. Found out the % of **NaN** values in each of the columns to determine which values to remove.

```
#calculating 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(app_data)
```

2. Removed columns with more than **40% NaN** values

```
# Removing columns where null values are greater than 40%
for col, values in app_data.iteritems():
    if get_perc_of_missing_values(app_data[col]) > 0.40:
        app_data.drop(col, axis=1, inplace=True)
app_data
```

3. Post these actions, I decided on imputing values on few columns to further make the data set usable.

```
# Since "AMT_GOODS_PRICE"  & "EXT_SOURCE_2" has very low missing values,
# we can use mean values to fill fill those columns

app_data['AMT_GOODS_PRICE'].fillna((app_data['AMT_GOODS_PRICE'].mean()), inplace=True)
app_data['EXT_SOURCE_2'].fillna((app_data['EXT_SOURCE_2'].mean()), inplace=True)
app_data[["AMT_GOODS_PRICE","EXT_SOURCE_2"]].describe().T
```

	count	mean	std	min	25%	50%	75%	max
AMT_GOODS_PRICE	307511.0	538396.207429	369279.426396	4.050000e+04	238500.000000	450000.000000	679500.000000	4050000.000
EXT_SOURCE_2	307511.0	0.514393	0.190855	8.173617e-08	0.392974	0.565467	0.663422	0.855

As seen from the above images, I need to impute the mean values into the **AMT_GOODS_PRICE** and **EXT_SOURCE_2** columns.

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

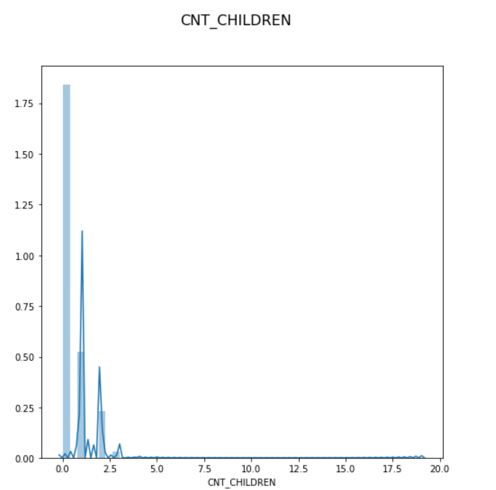
```
app_data.NAME_TYPE_SUITE.value_counts()
```

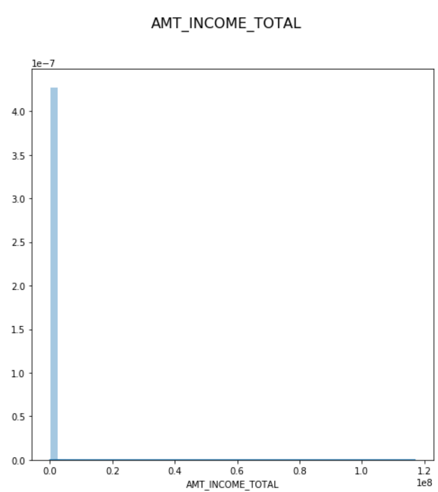
```
Unaccompanied      248526
Family              40149
Spouse, partner     11370
Children             3267
Other_B              1770
Other_A               866
Group of people       271
Name: NAME_TYPE_SUITE, dtype: int64
```

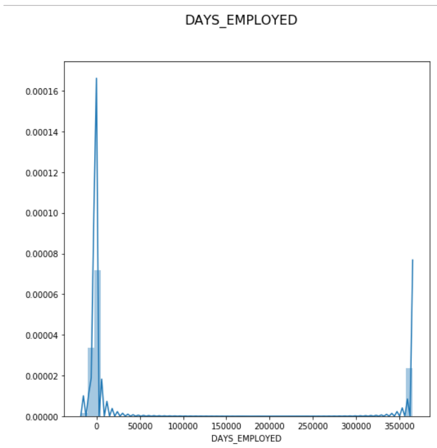
```
# We can fill missing values  with "Unaccompanied" data since it has the highest mode.

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

Finding outliers in the given data frame

Here are some of the columns where I could spot values as outliers with the help of plots:







Above box plot for **CNT_CHILDREN** show 19 as an outlier. Since a family can't or very rarely have 19 children, it is treated as an outlier.

Above plot for **DAYS_EMPLOYED** shows that there is a value present at 36k range which is not possible.

Above plot for **AMT_INCOME_TOTAL** shows the max amount is much larger than other statistical data.

Now that the outlier have been identified, I removed them and plotted them again to observe the difference.

Furthermore I have done some modification of values in order to make the analysis of data easier. These would be converting Date of Birth to age and binned salaries into different levels called High, Medium and Moderate.

ANALYSIS OF APPLICATION DATA

I then proceeded with the analysis of data.

Divided data into separate dataframes called defaulter and good client.

```
good_client = app_data[app_data.TARGET == 0]
defaulter_client = app_data[app_data.TARGET == 1]
good_client.info()
```

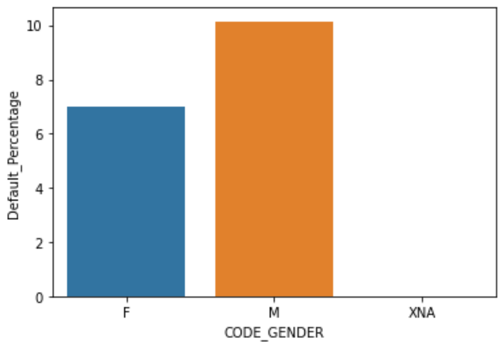
Target value 0 indicates that the client is not a defaulter thus a good client.

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.

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

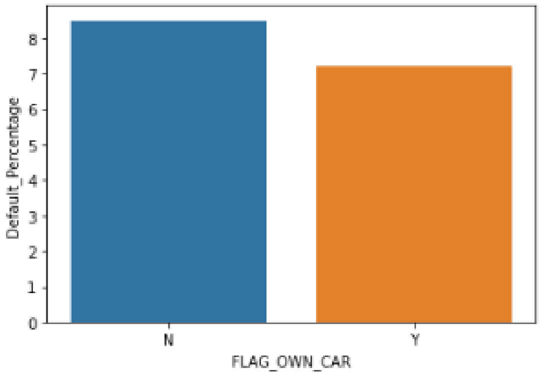
  Checking to see which clients are unlikely to pay back the loan by analysing various columns in the data frame.

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



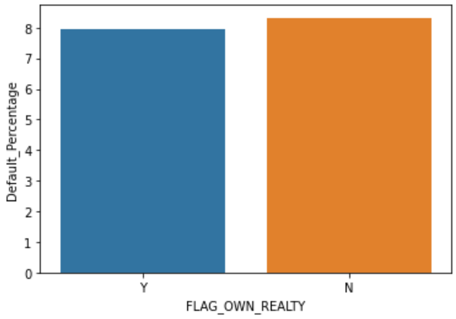
Male client have a higher chance of not returning their loans [10.14%] compared to the female clients [7%]. Therefore we can see that Female clients are a better TARGET compared to the Male clients.

  - Based on **FLAG_OWN**_CAR (clients that own a car or not)



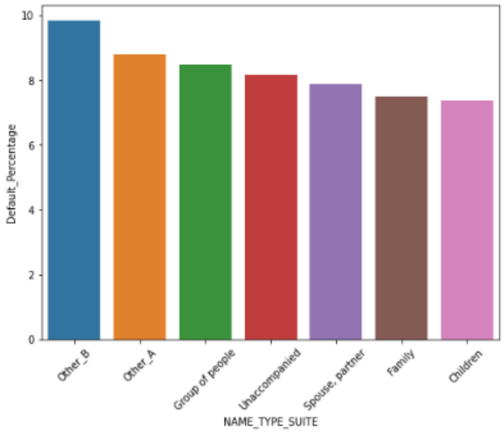
Clients that own a car are more likely to repay their loans compared to a client that does not own a car according to the above graphs and data

- Based on **FLAG_OWN_Realty** (clients that owns property)



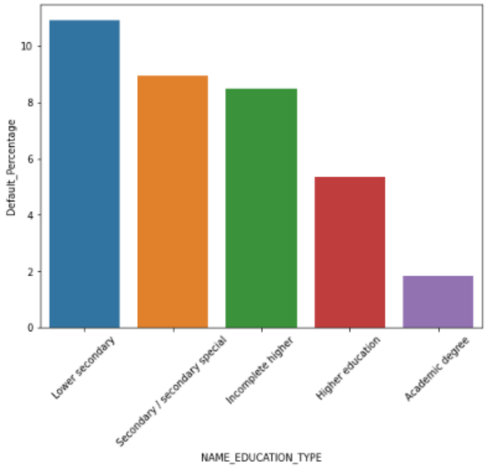
From the above graph and data, repayment of loans of clients that own property are similar or barely below those that do not own property. Therefore, it is difficult to decide a target based on this

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



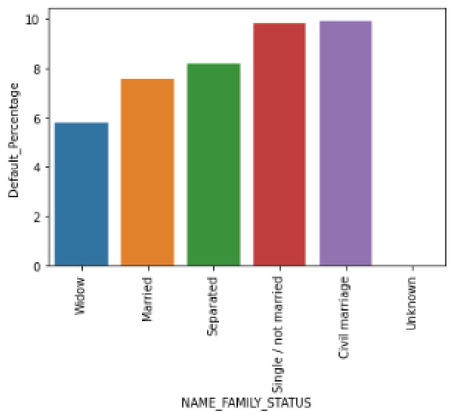
From the above graph and data, repayment of loans are similar across all type suites. Therefore, it is difficult to decide a target based on this.

- Based on **NAME_EDUCATION_TYPE**



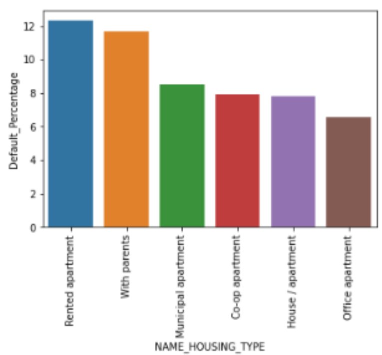
From the above graph and data, it can be seen that more educated clients are more likely to repay loans.

- **Based on NAME_FAMILY_STATUS**



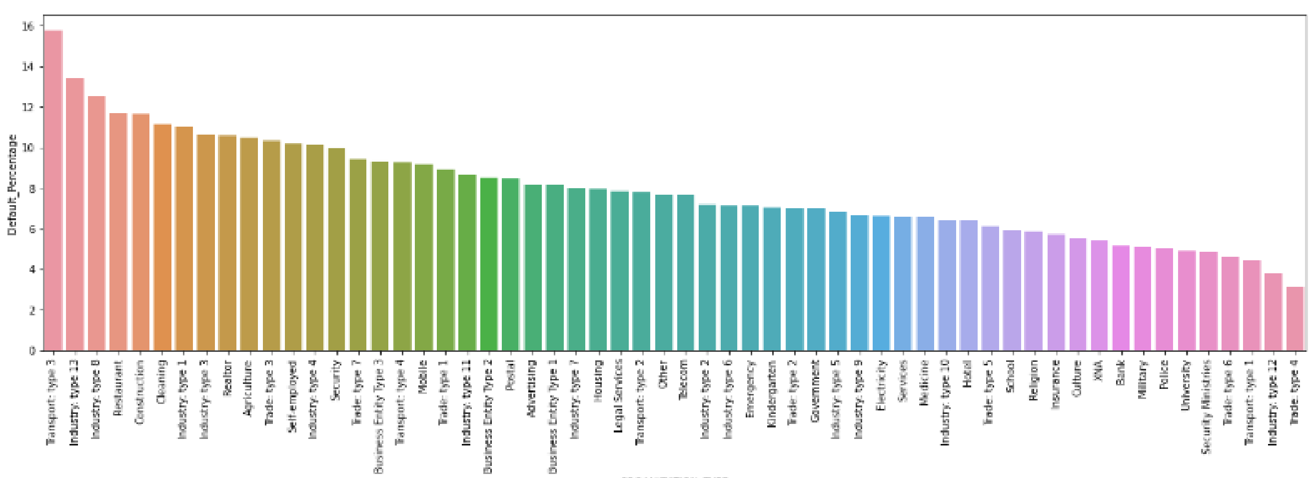
From the above graph and data, it is seen that the percentage of non-repayment of loan is at highest for civil mariage and is lowest for widows

- Based on **NAME_HOUSING_TYPE**



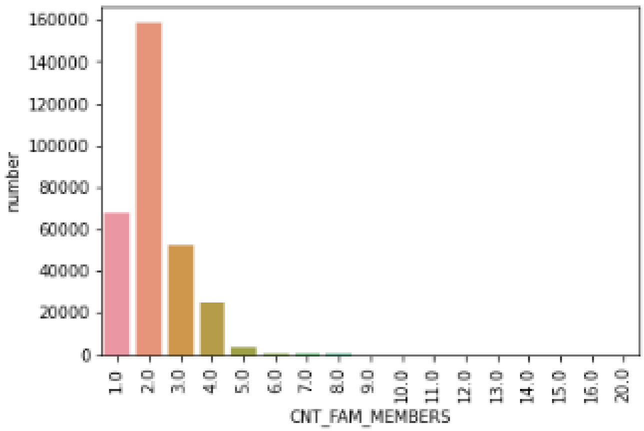
From the above graph and data, it can be seen that people with rented apartments are less likely to pay back their loans

- Based on **ORGANISATION_TYPE**

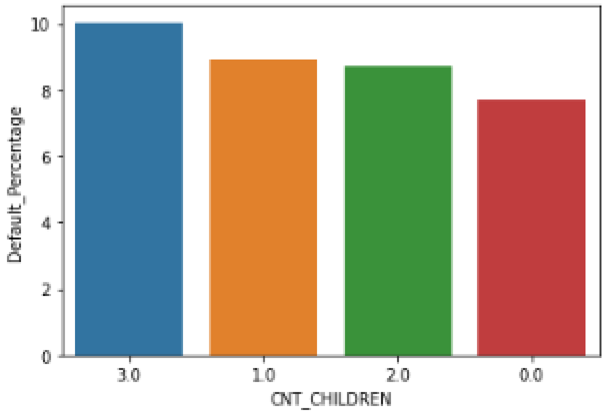


From the above graph, highest number of non-repayment can be seen in Applicants who work in Transport Type3.

- Based on **CNT_FAM_MEMBERS**



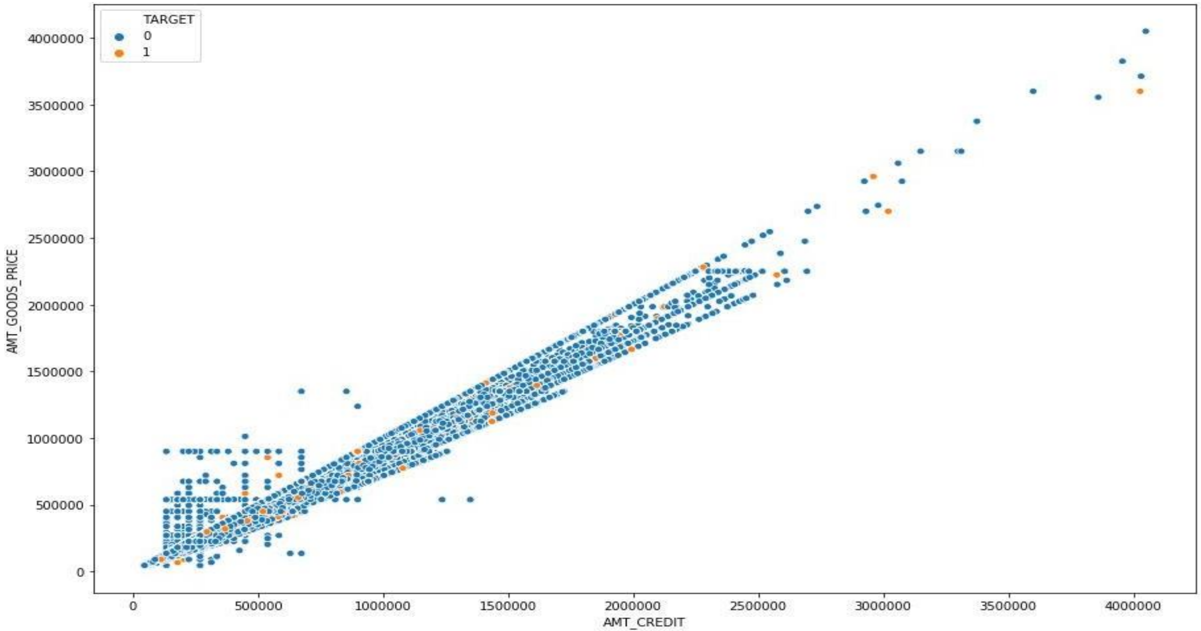
- Based on **CNT_CHILDREN**



From the above graph and data, there is a higher chance that a client with more children is unlikely repay the loan.

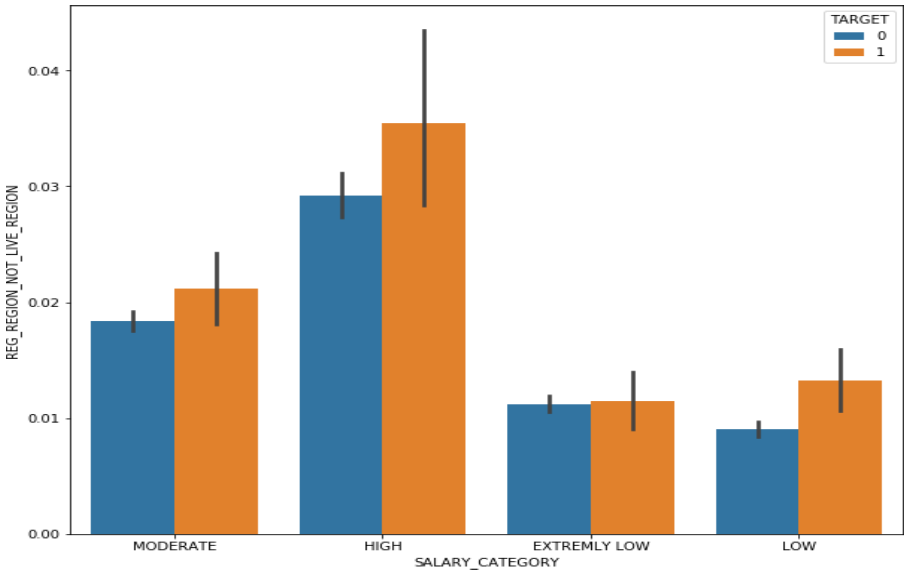
- **BIVARIATE ANALYSIS**

AMT_CREDIT vs AMT_GOODS_PRICE



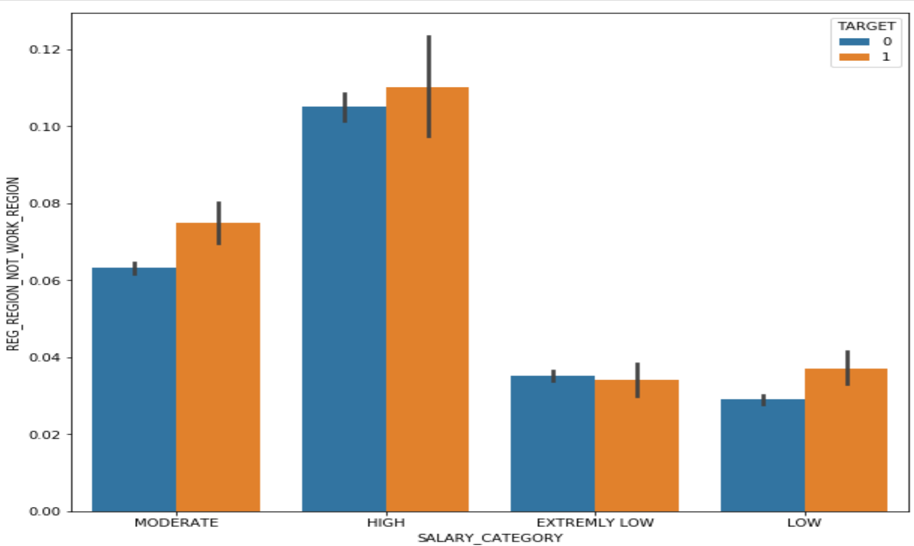
It is found that the Credit amount and the Amount of goods price are more correlated with the Defaulters. As both these variables increase, the Defaulters are linearly increasing as well.

SALARY VS CLIENT WHOSE PERMANENT ADDRESS NOT MATCH WITH CONTACT ADDRESS



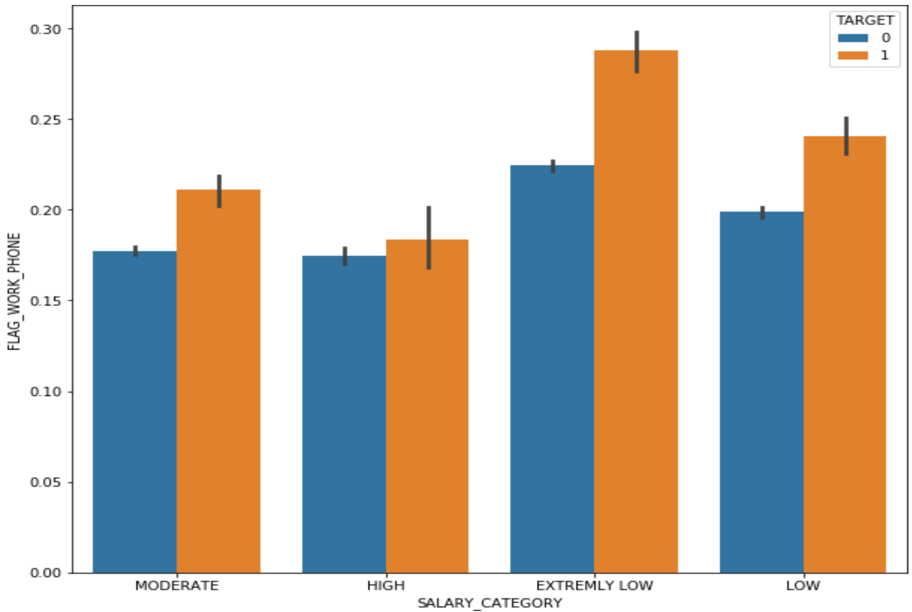
When the Client has a very low salary and the Clients' contact address does not match, there is a high chance the Client is going to be a defaulter.

Salary vs Client whose Permanent Address not match with Work Address



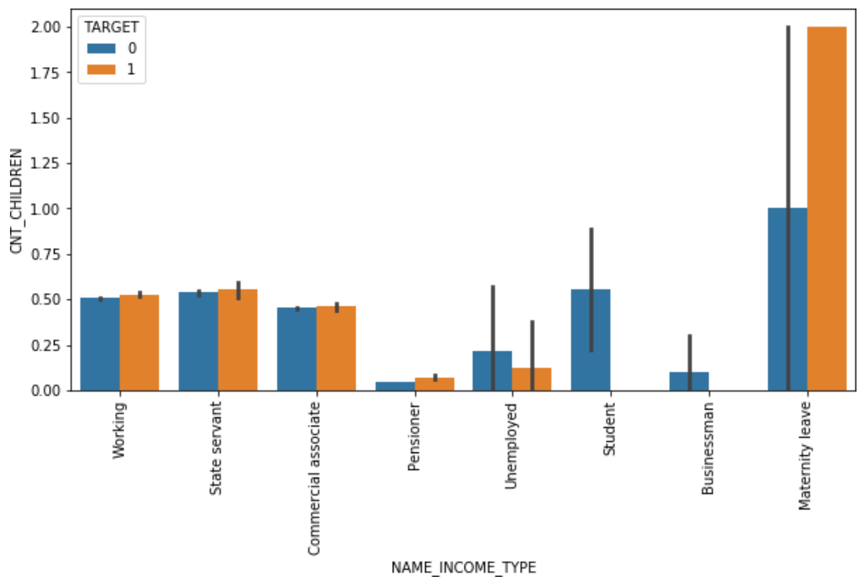
When the Client has a very low salary and if the Clients' work address does not match, there is a high chance for the Client to be a defaulter.

Salary Category vs Client who provided Home Number



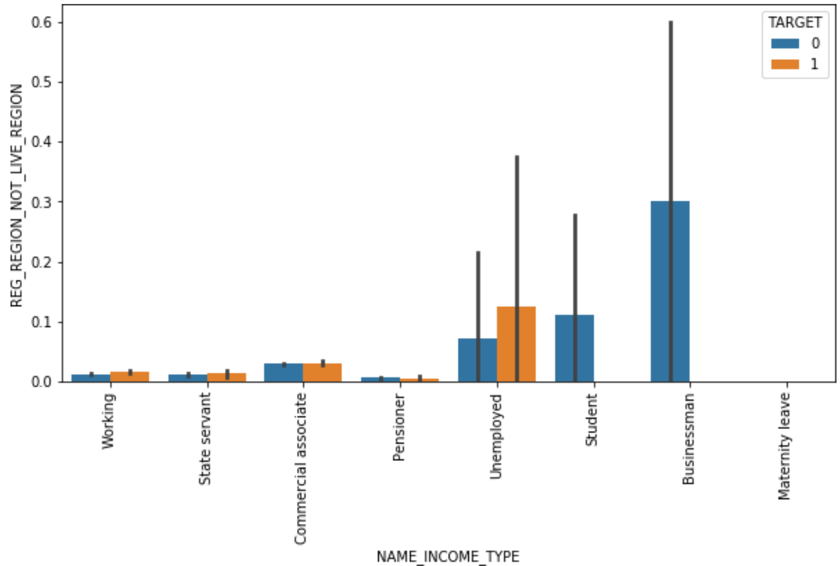
When a Client with very low salary does not provide their home phone number at the time of taking loan, there is a higher chance for the Client to be a Defaulter.

INCOME vs CHILDREN Count



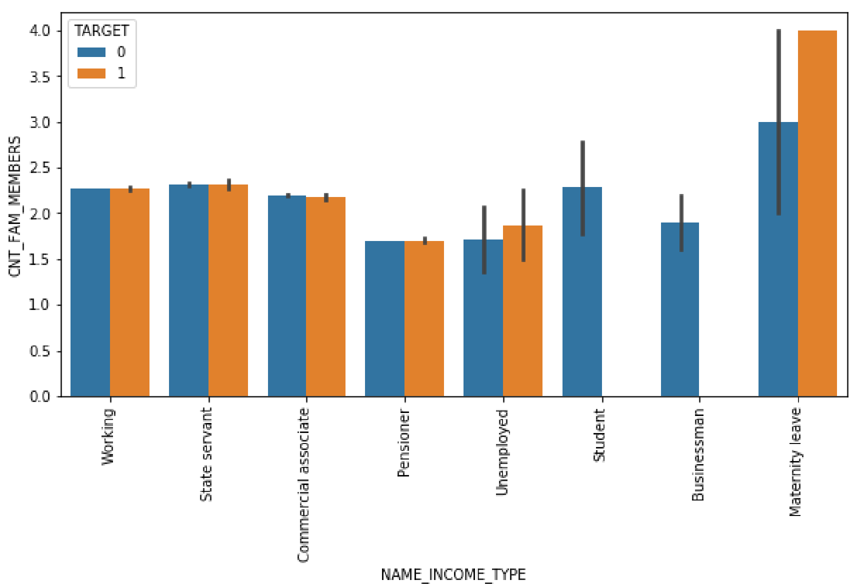
Clients that are getting income via Maternity Leave tend to have a higher chance at being a defaulter when the children count is higher.

Income Type vs Client whose Permanent Address not match with Contact Address



Clients that are Unemployed have a higher chance at being a defaulter when their Permanent Address does not match with the Contact Address.

Income vs No.of.FamilyMembers



Clients that are getting income via Maternity Leave tend to have a higher chance at being a Defaulter when they have more Family Members.

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

CNT_FAM_MEMBERS and CNT_CHILDREN = 0.87

AMT_CREDIT and AMT_GOODS_PRICE =0.99

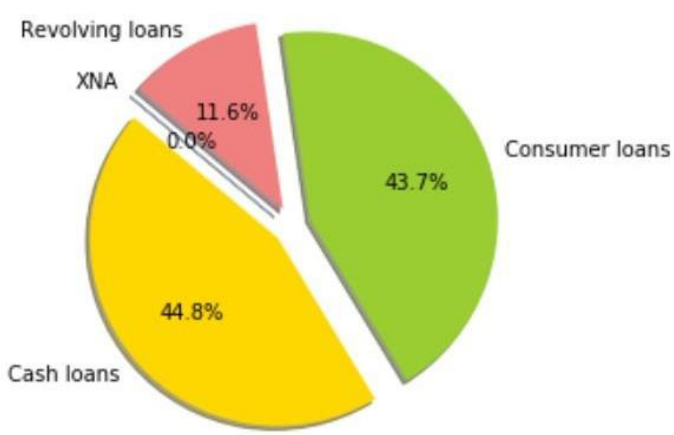
AMT_ANNUITY and AMT_CREDIT = 0.77

REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT = 0.95

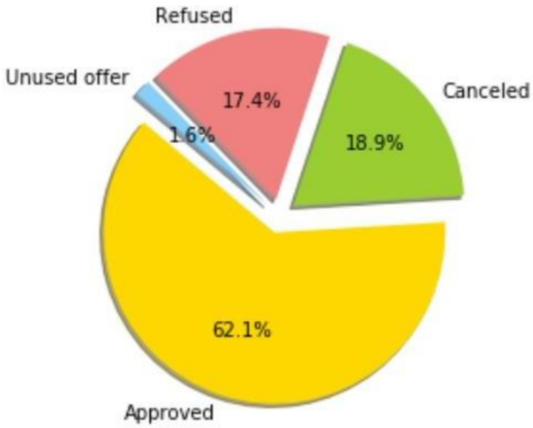
PREVIOUS APPLICATION ANALYSIS

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

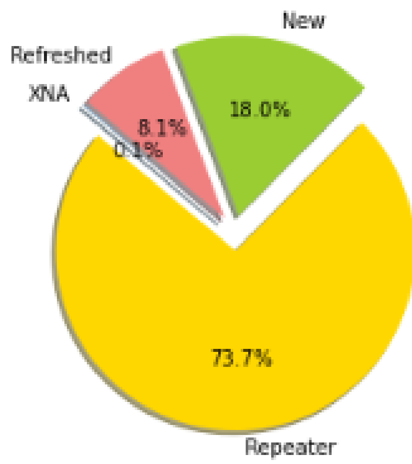
- **Based on Contract Type**



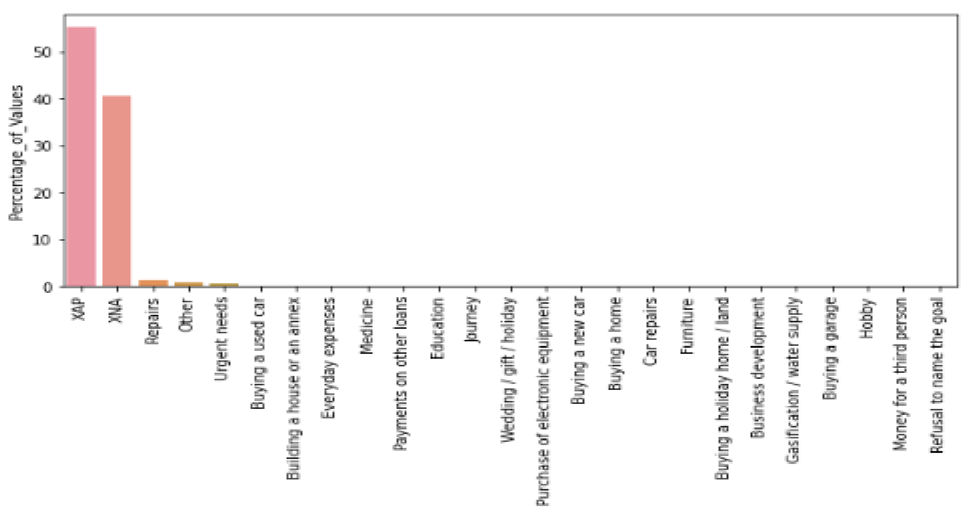
- **Based on Contract Status**



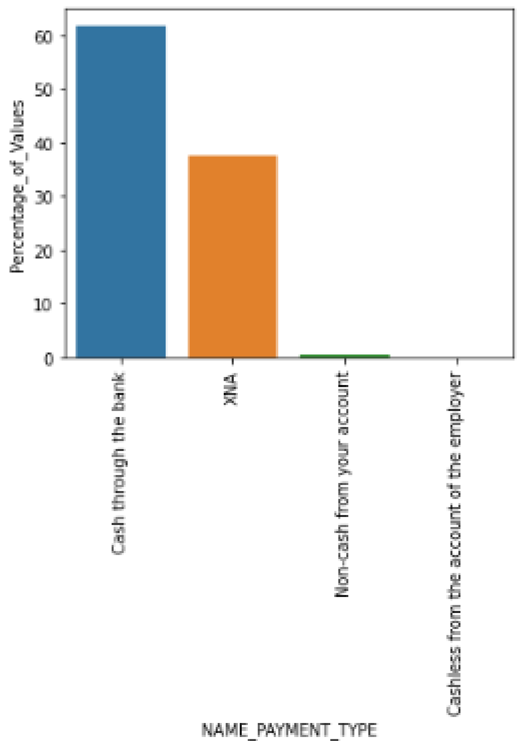
- **Based on Client Type**



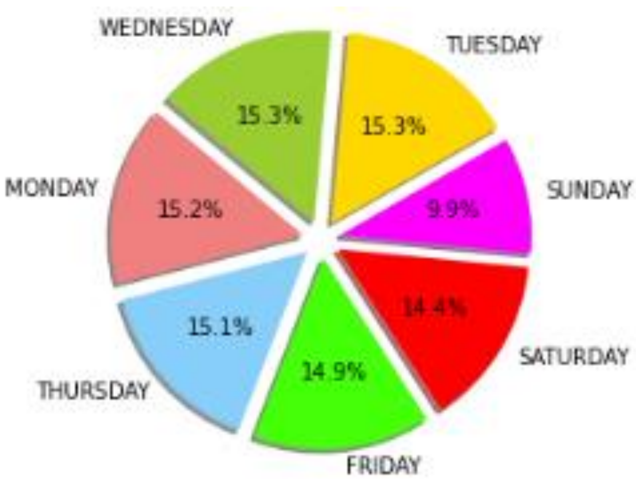
- **Based on Purpose of Loan**



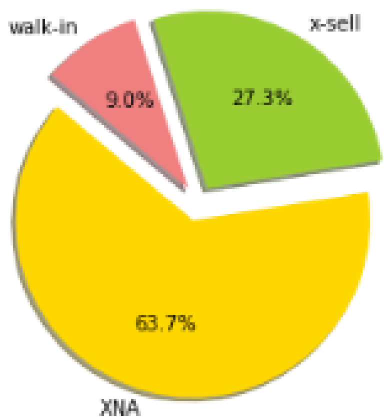
- **Based on Payment Type**



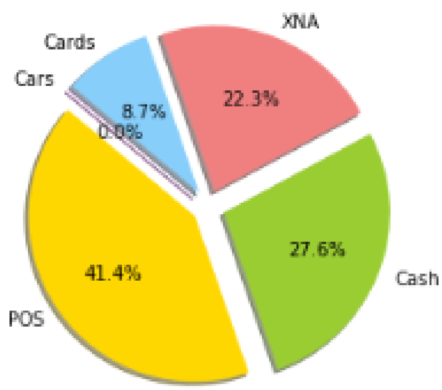
- **Based on Days of Approval**



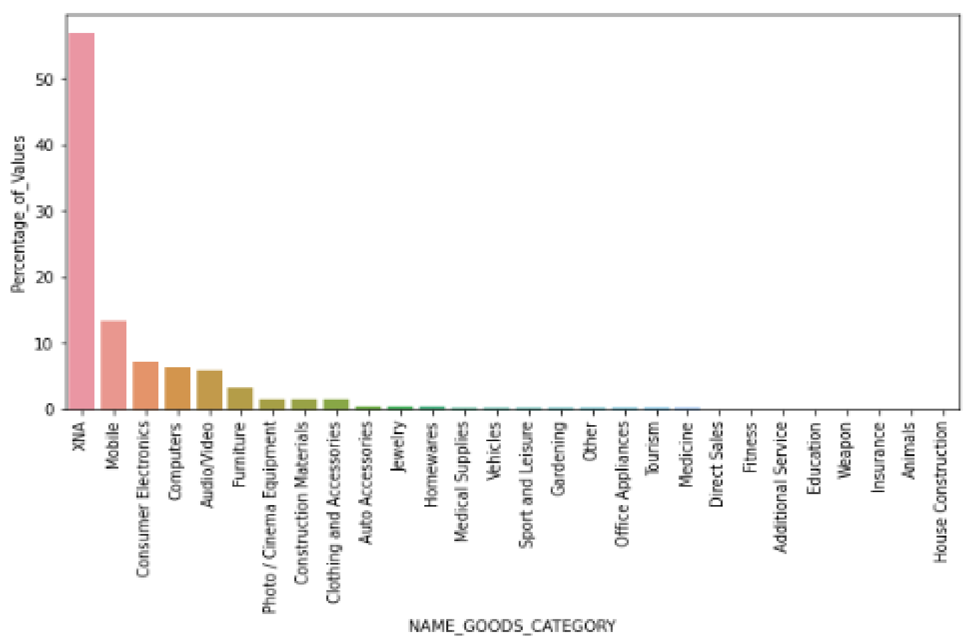
- **Based on NAME_PRODUCT_TYPE**



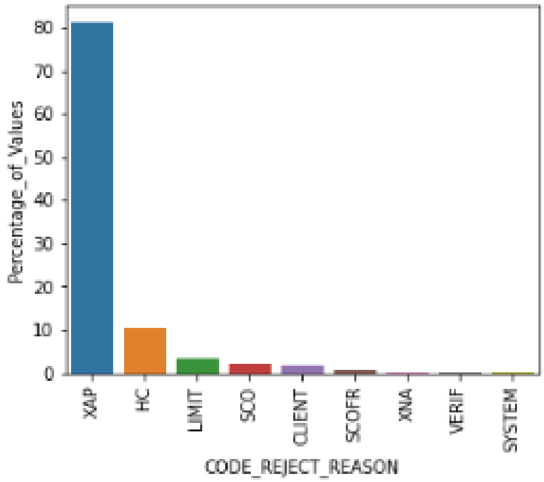
- **Based on NAME_PORTFOLIO**



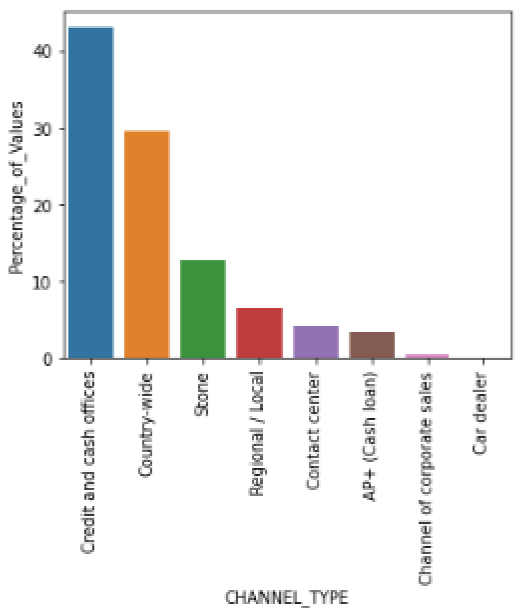
Based on NAME_ GOODS_CATEGORY



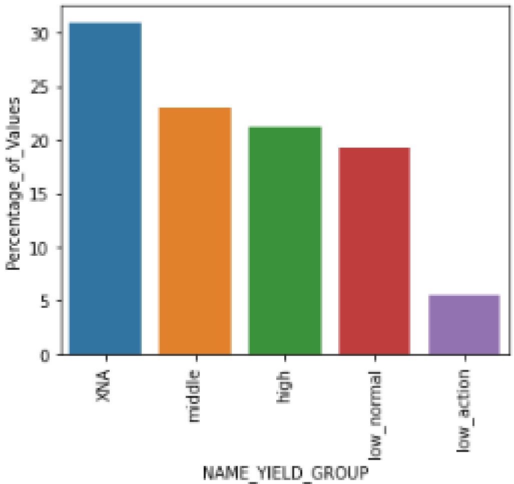
- **Based on Reason of rejection of loan**



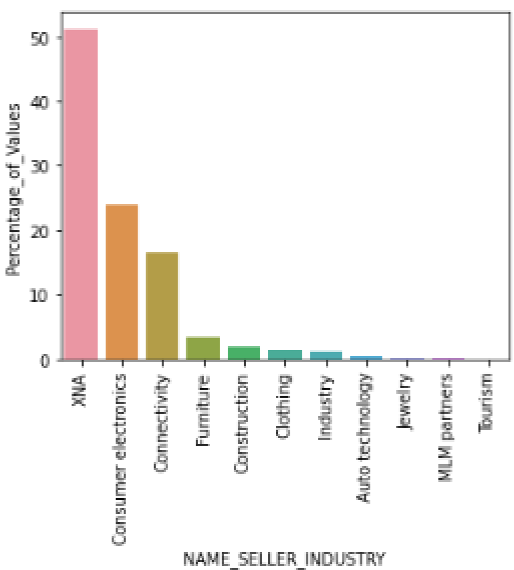
- **Based on CHANNEL_TYPE**



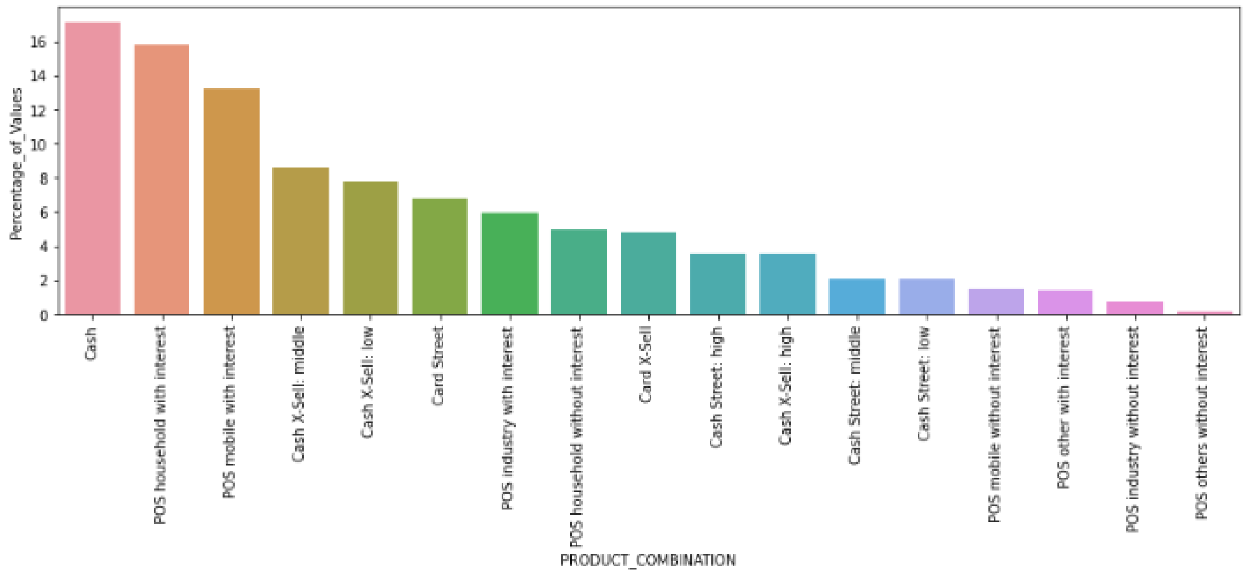
- **Based on NAME_YIELD_GROUP**



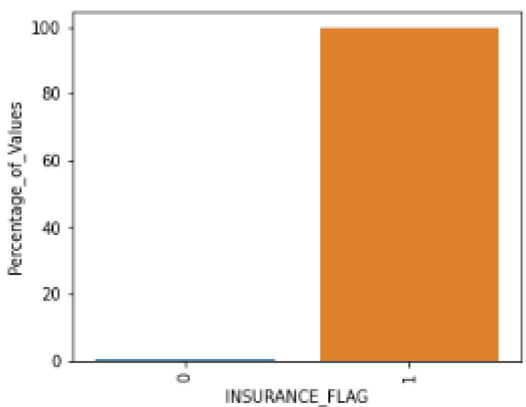
- **Based on NAME_SELLER_INDUSTRY**



- **Based on PRODUCT_COMBINATION**



- **Based on NFLAG_LAST_APPL_IN_DAY**



MERGING APPLICATION DATA AND PREVIOUS APPLICATION

After analyzing all the previous and current applications, I once again checked the correlation of the variable with respect to the Target variable. I 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

```
AMT_GOODS_PRICE              -0.032550
REGION_POPULATION_RELATIVE   -0.035028
AGE                          -0.074927
EXT_SOURCE_2                 -0.154919
```

As seen in the application Data, mostly the variables are more or less familiar which has been contributing more to the **DEFAULTERS** prediction.