# Credit EDA Case Study


## Submitted By:

1. **Dibyalok Bhutiarai**
2. **Surya Mangipudi**


# Problem Statement – I & Problem Statement – II

### Data Understanding

This dataset has 3 files as explained below:

*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.

**Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.**

1. First as shown below we downloaded the files and start the analysis
2. We need to study both the files and try to check the columns to select.
3. We need to look for Null Values in the columns and take a decision on how to deal with the Null Values i.e. either remove the fields or put some relevant data if we can. For the Null Values, we checked for the % of Null Values for each entry in the Input file. For the entries, where the % of Null Values were more than 50%, those fields were dropped.
4. Then we merged both the tables and did an analysis.

**Identify the missing data and use the appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)**

- We run the below query to find the Percentage of Missing Values:
    - ld_application_percent = round(100*(old_application.isnull().sum()/len(old_application.index)), 2)
    -

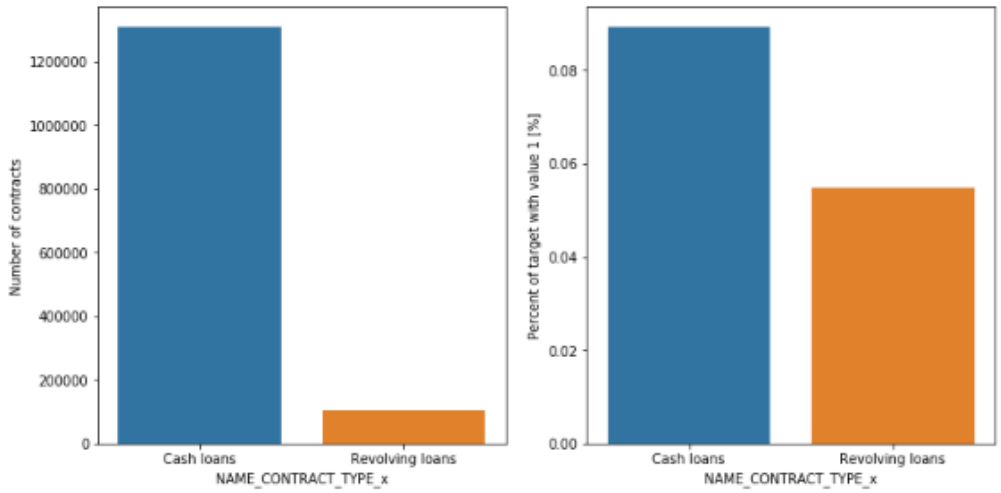
- **<u>The below columns has the more than 50% of missing columns and we don't need this for our analysis and hence we removed them.</u>**

  - 'RATE_INTEREST_PRIMARY'
  - 'RATE_INTEREST_PRIVILEGED'
  - 'AMT_DOWN_PAYMENT'
  - 'RATE_DOWN_PAYMENT'

- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.
- Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.
- Include visualizations and summarize the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.

# ANALYSIS

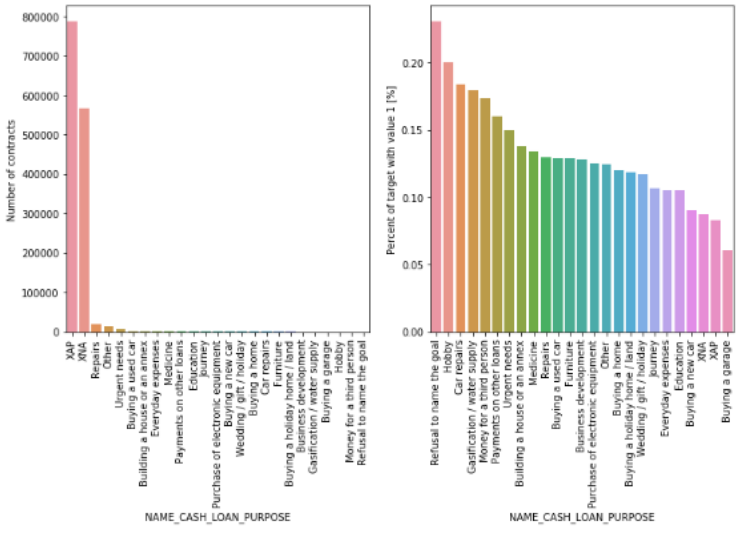
```
plot_p_stats('NAME_CONTRACT_TYPE_x')
```




**Inference**:

It appears from above bar chart that the metrics that relate to revolving loans though very less when compared to cash loans, but the non-payment % stands comparatively relatively high as well

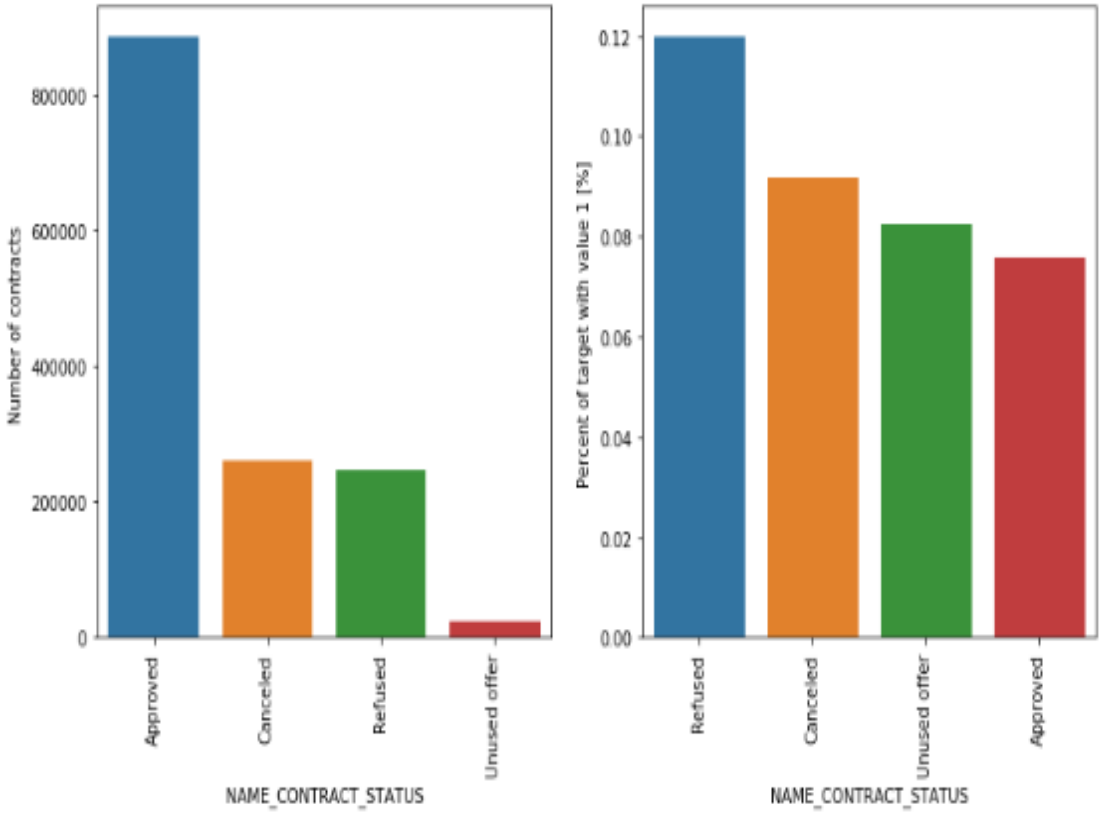
```
In [23]: plot_p_stats('NAME_CASH_LOAN_PURPOSE', True, True)
```



**Inference**:

Even customers who are falling into 1) not revealing the goal for the loan 2) Hobby 3) Car rentals who are all very low / few from contracts metrics appear standing high over the non-payment % metrics.

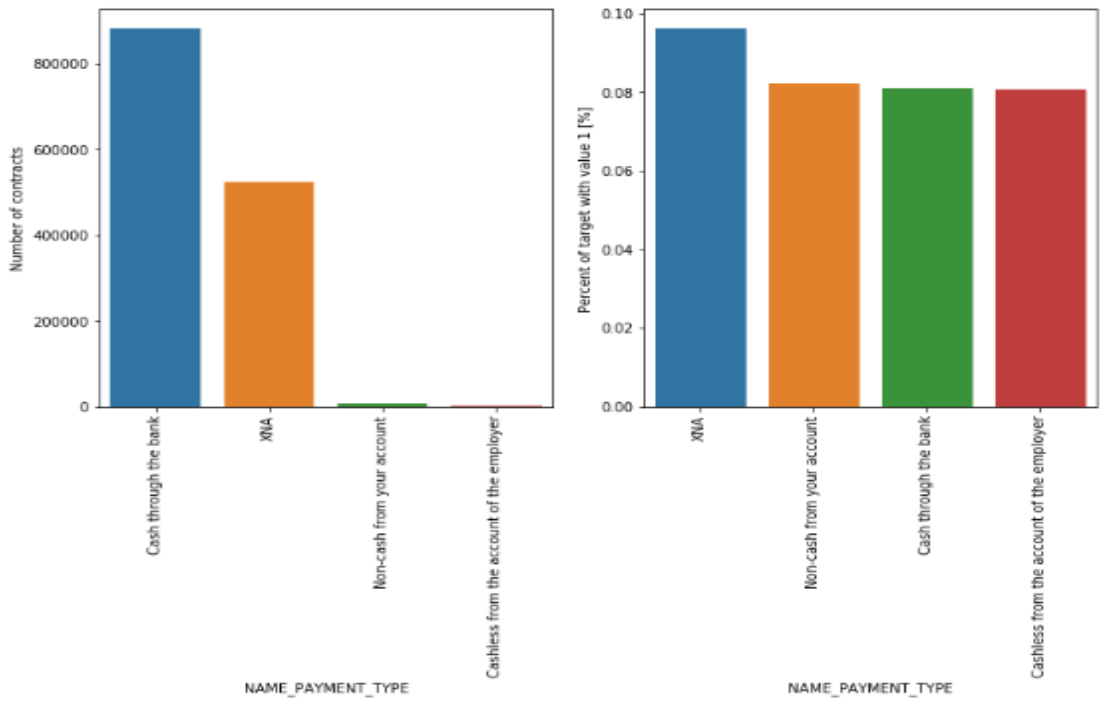
```
plot_p_stats('NAME_CONTRACT_STATUS', True, True)
```




**Inference:**

The Refused lot though less in count appear the highest as defaulters, but the more surprise fact being the Approved lot, equally getting found ranging close to 60% on the defaulters list

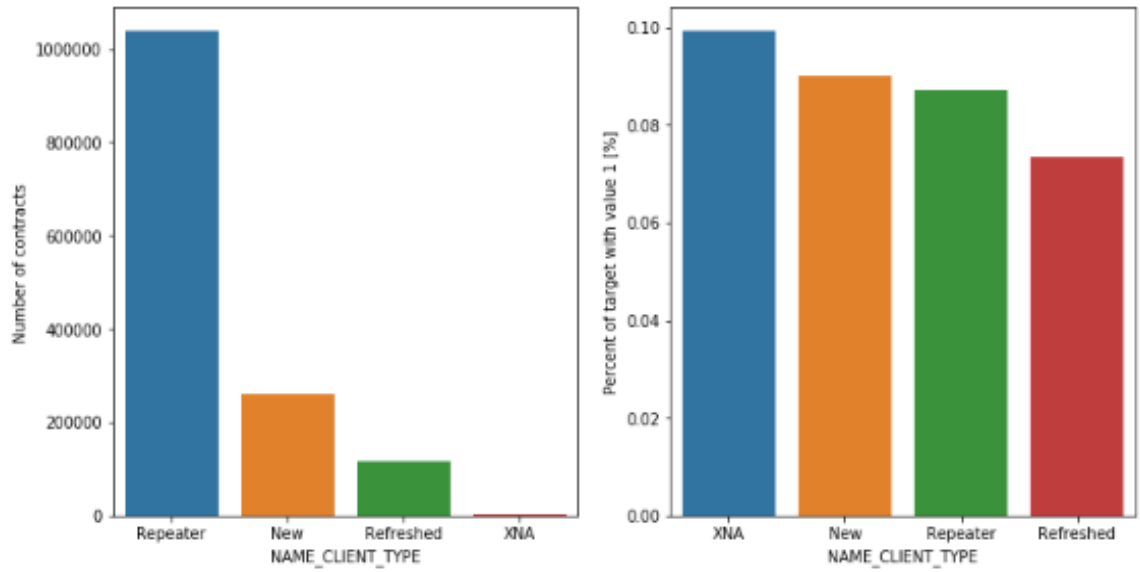
```
plot_p_stats('NAME_PAYMENT_TYPE', True, True)
```




**Inference:**

Irrespective of the mode, all of them stand >80% towards nonpayment % metrics.

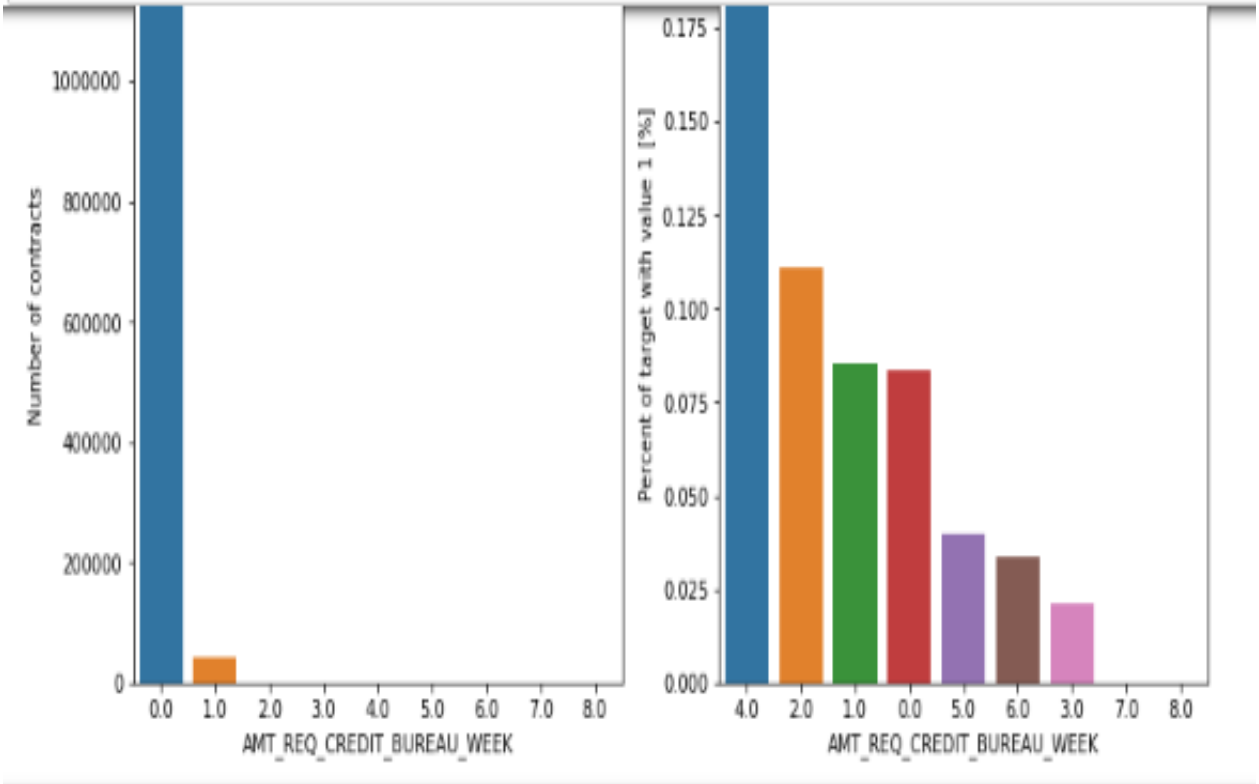
```
plot_p_stats('NAME_CLIENT_TYPE')
```



**Inference:**

The XNA customer type though very small from loan / contract count perspective, stands close to 100% towards nonpayment %. On the converse, the Repeater appears with maximum contracts, but still gets found close to 90% non-payment category. All in All, irrespective of the client type, the non-payment category is always standing more than 75%

```
plot_p_stats('AMT_REQ_CREDIT_BUREAU_WEEK')
```

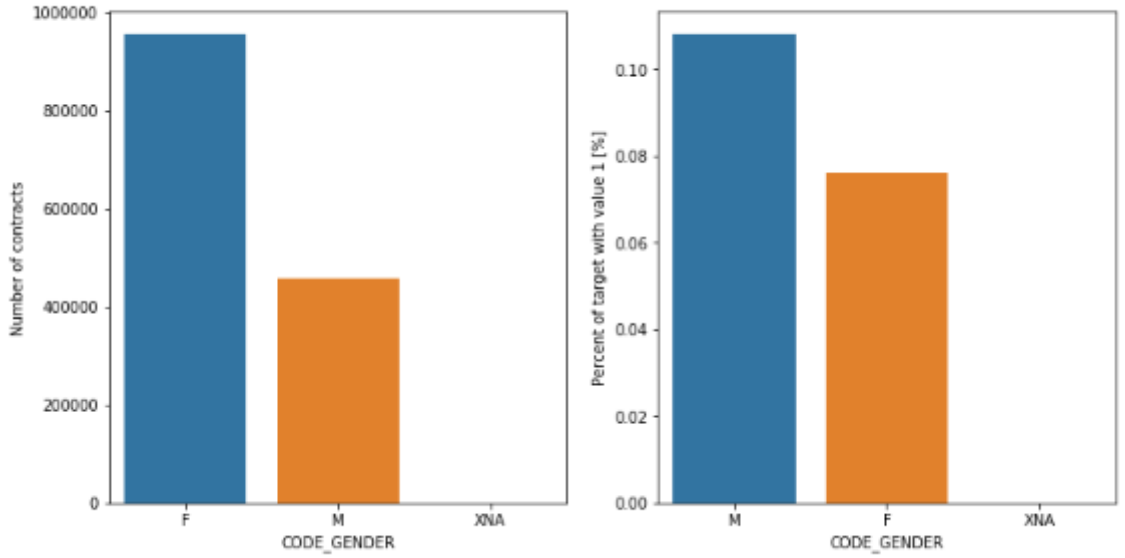



Type *Markdown* and LaTeX: $\alpha^2$

**Inference:**

AMT_REQ_CREDIT_BUREAU_WEEK = 0 stands a whopping 83%, but still the Non Payment % is huge 195% mark.

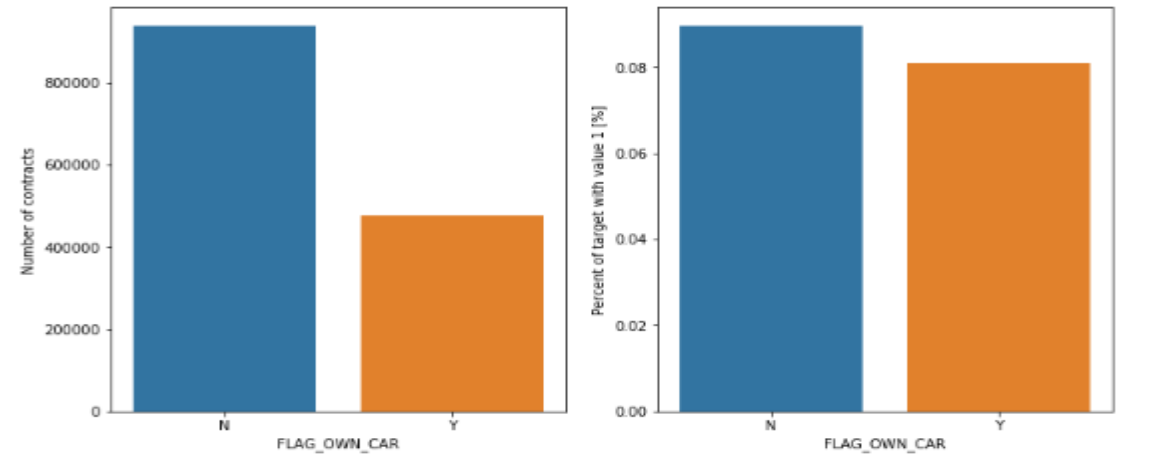
```
plot_p_stats('CODE_GENDER')
```




**Inference:**

Females are close to the million mark on the number of contracts, but get found close to 78% towards the non-payment %. Males seem close to half a million mark, but rather appear more prone on failure metrics towards nonpayment.

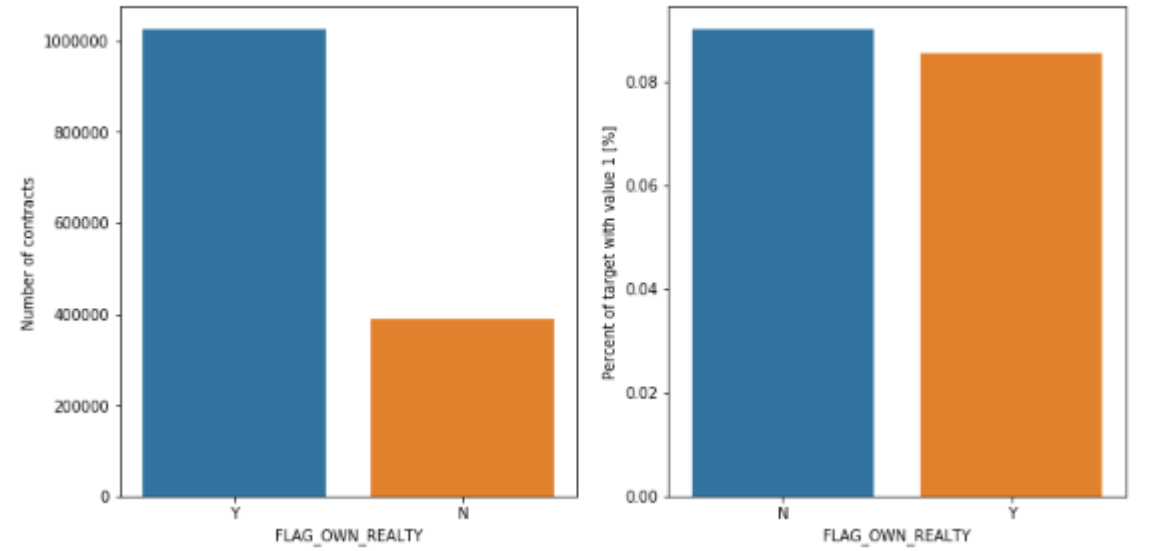
```
plot_p_stats('FLAG_OWN_CAR')
```




**Inference:**

104587 out of 307511 which stands at 34% of customers who own a car, but they stand close to customers who dont own a car, when this comes to non payment % of the taken loan.

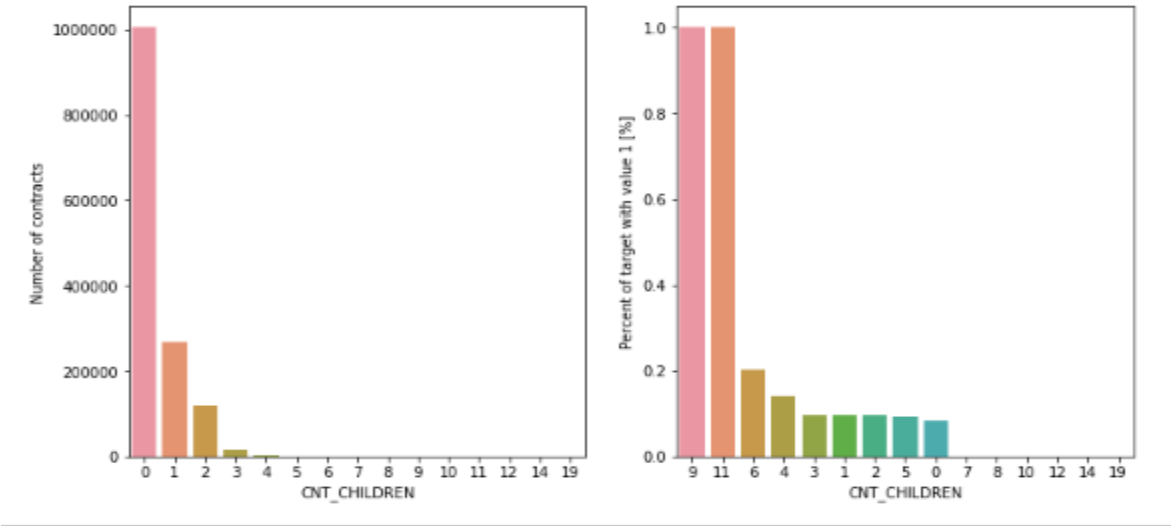
```
plot_p_stats('FLAG_OWN_REALTY')
```




**Inference:**

213312 out of 307511 which stands at ~70% of customers who own a flat/aptcar, but they stand close to customers who dont own the same, when this comes to non payment % of the taken loan.

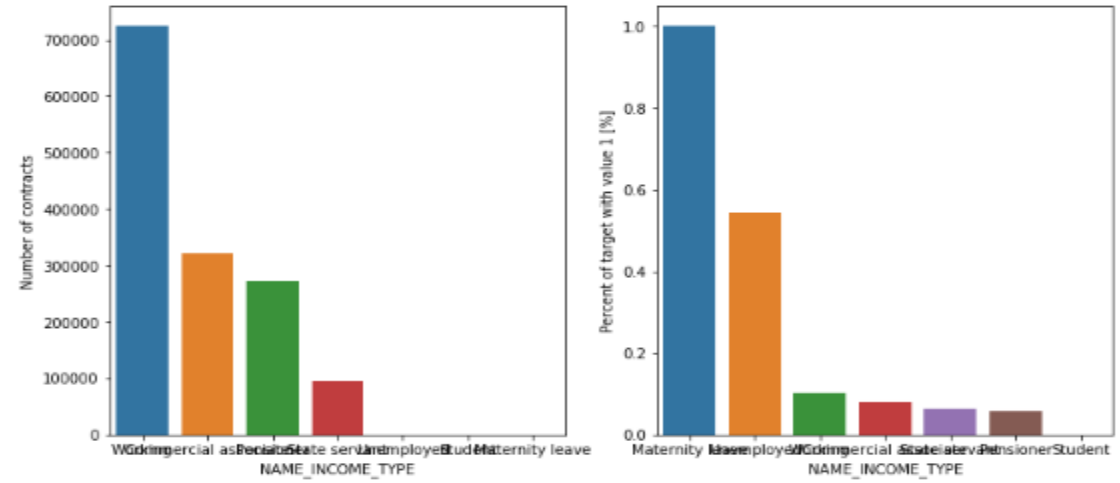
```
plot_p_stats('CNT_CHILDREN')
```




**Inference:**

Clients with no children stands at 10% failure of repaying loan amounts, but clients with child counts as 9 and 11 are failing 100% towards the same loan repayment.

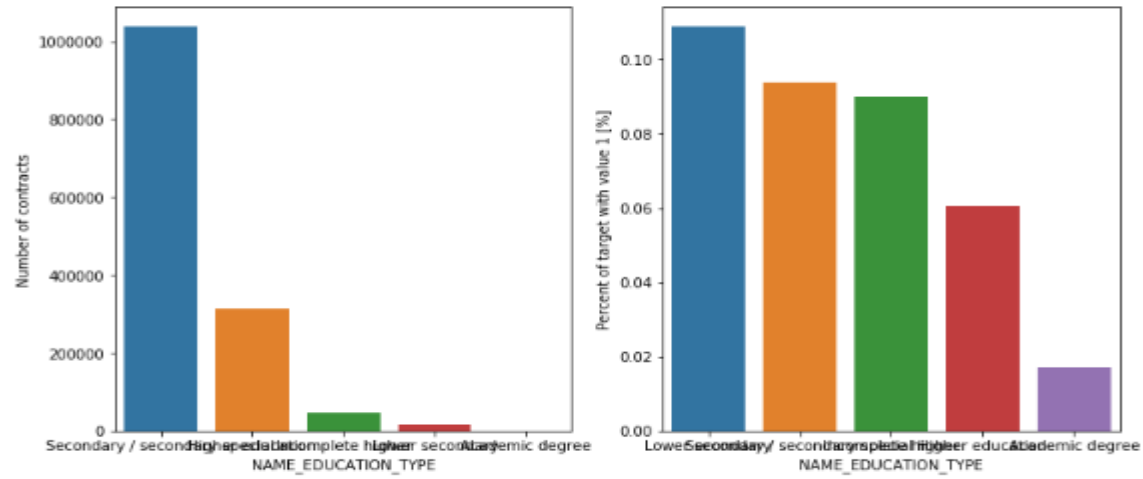
```
plot_p_stats('NAME_INCOME_TYPE')
```




**INFERENCE:**

158774/30751 = ~52% from "Working" class appear the most at taking contracts but they appear to be failing only by 10% at paying back the loan amounts. On contrary, the 5 customers on Maternity Leave appear failing by by full 100% on the repay loan amount. It helps the bank to share the loan amounts to Working Class

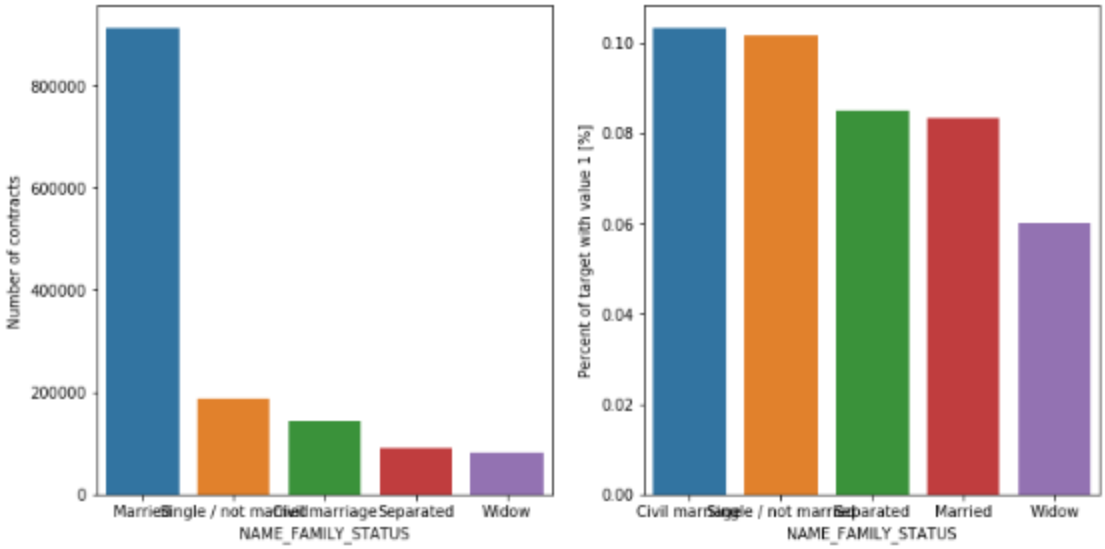
```
plot_p_stats('NAME_EDUCATION_TYPE')
```




**INFERENCE:**

Education definetely has  a role to play when it comes to loan repayment. Lower Secondary and Secondary education categories seem the highest at loan defaulters and Academic Degree and Higher Education seem lower on this loan defaulter types.

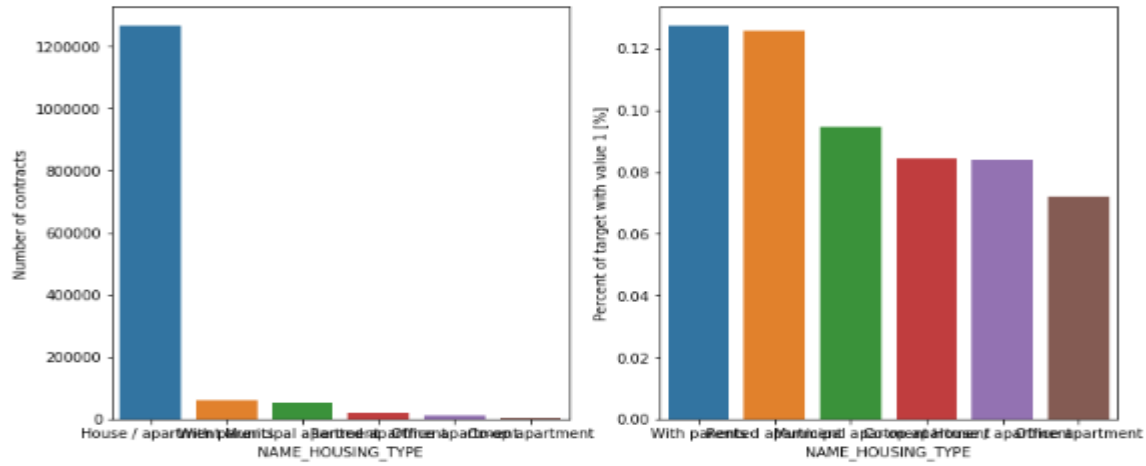
```
plot_p_stats('NAME_FAMILY_STATUS')
```




**INFERENCE:**

Customers falling under Civil Marriage, though comprising <10% of the loan/contract takers, seem the highest at not paying back. The risk appears very high here. This appears the same case with Married lot. On contrary, Widows who appear <5% on taking loans, appear by far safe when it's coming to repaying the loan amount back.

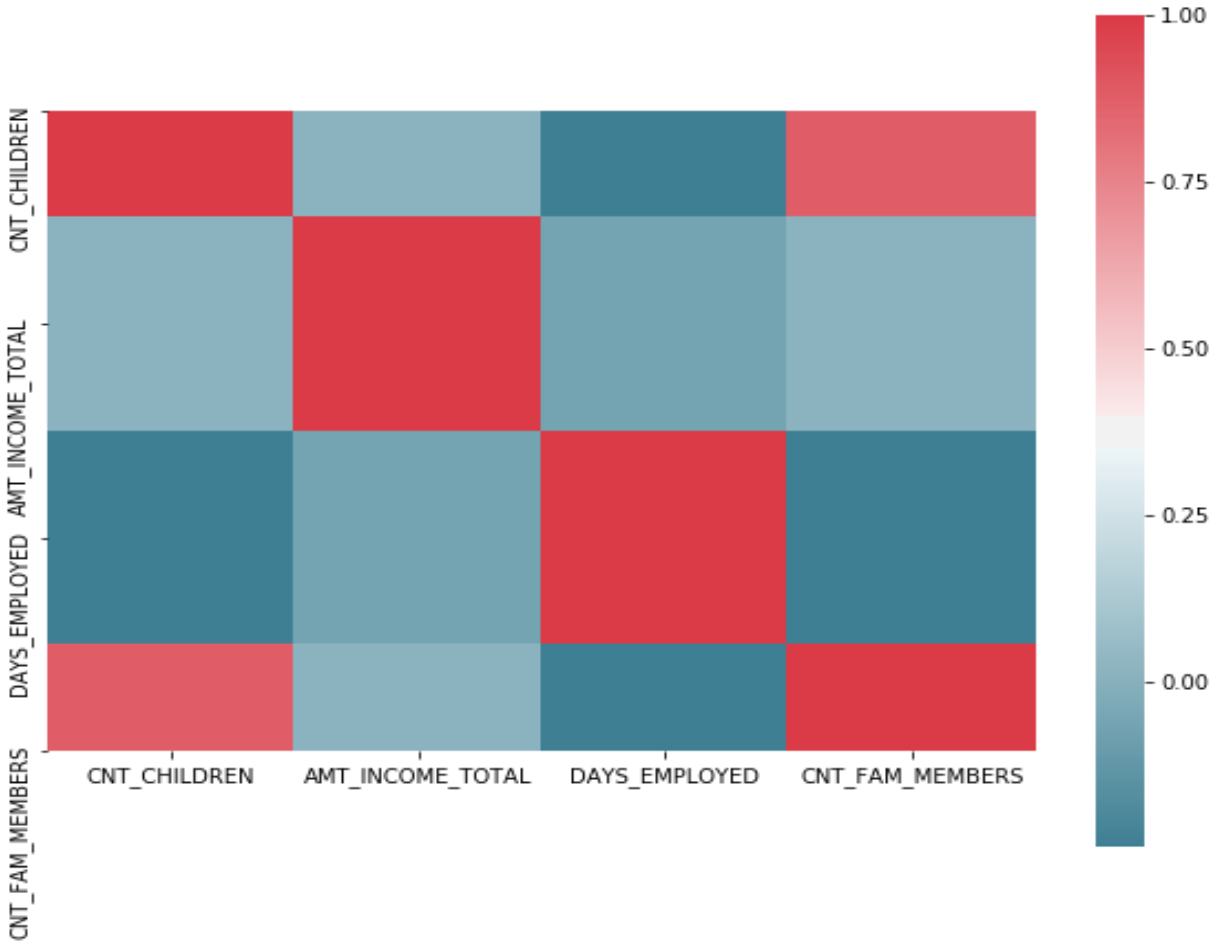
```
plot_p_stats('NAME_HOUSING_TYPE')
```




**INFERENCE:**

272868 out of 307511 = ~89%, who are House / Apartment owners seem the most taking loans, and are ranged around 80%+ at not able to pay back the loan amount on time. On the reverse, people staying with parents who comprise only about 4% appear the worst case scenario at loan repayment ranging more than 130% defaulters

To understand the correlation between the datasets we need to use a heatmap for visualization.




**INFERENCE**

The heat map gives the relation of correlation among the different columns or variables. The variables having correlation are

CNT_CHILDREN,CNT_INCOME_TOTAL,DAYS_EMPLOYED,CNT_FAM_MEMBERS.