# EDA CASE STUDY – Presentation Doc

## Objective:

For this case study, I studied the data and tried to create a profile for a potential defaulter. During this process I went through the whole data and tried exploring the probability of a person being a defaulter based on different characteristics like gender, income group, job type and so on.

## Author:

Tanoy Majumdar

## Date:

6/28/2021

## Data:

I was basically provided with two data sets and a data dictionary.

1. Application data: This dataset contained the information about the current loan applications which contained information like whether the applicants have defaulted till now or not, the gender of the applicants, the loan amount, the price of the goods that was purchased, the job type, the organization type, the area. how many documents were submitted,etc
2. Previous data: These dataset contained the information of the previously applied loans of the current applicants. Among other things this dataset contained the different statuses of the loans i.e accepted, refused, cancelled and unused offer. It also contained information like what was the loan amount, loan type, Goods category, etc.
3. Columns description: This was a data dictionary which told us what the different columns meant in both the previous datasets

## Libraries used:

For this analysis, I used the following 3 python libraries:

1. matplotlib.pyplot
2. numpy
3. pandas

## Exploratory Data Analysis:

We started of first by reading the application data, which contained 307511 rows and 122 columns. Next I went on to delete the columns which contains more than 13 percent missing values. However I didn’t delete the OCCUPATION_TYPE column as I thought this column might provide me some interesting insights. In the whole process I dropped 42 of the original rows

I did the same for previous application dataset where there were 1670214 rows and 37 columns. Here I deleted columns with 40 or more percent missing data. I ended up dropping 9 rows.

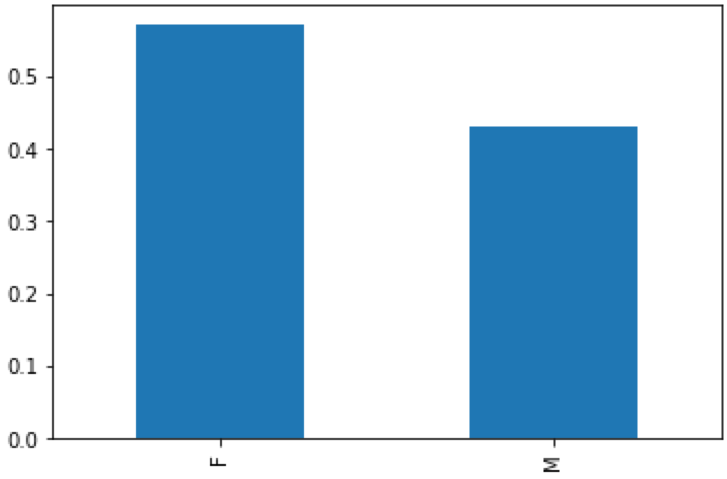
Next for both the datasets I did a preliminary check using the info and describe keywords. After this I started the actual analysis.

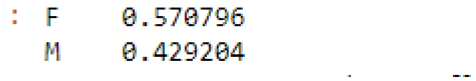
## Checking the percentage of defaulters in the main population (total number of applicants):

Here we saw that of the main population there were around 92 percent non defaulters and around 8 percent defaulters. So I bifurcated the whole dataset into two distinct datasets (One with only defaulters and other with non defaulters). This will help us to create a basic profile of the defaulter.

## Analysis based on the percentage of male and female in the defaulter list:

We created a bar graph to see that 57 percent of the defaulters were female and 43 were male.Below is the bar chart:



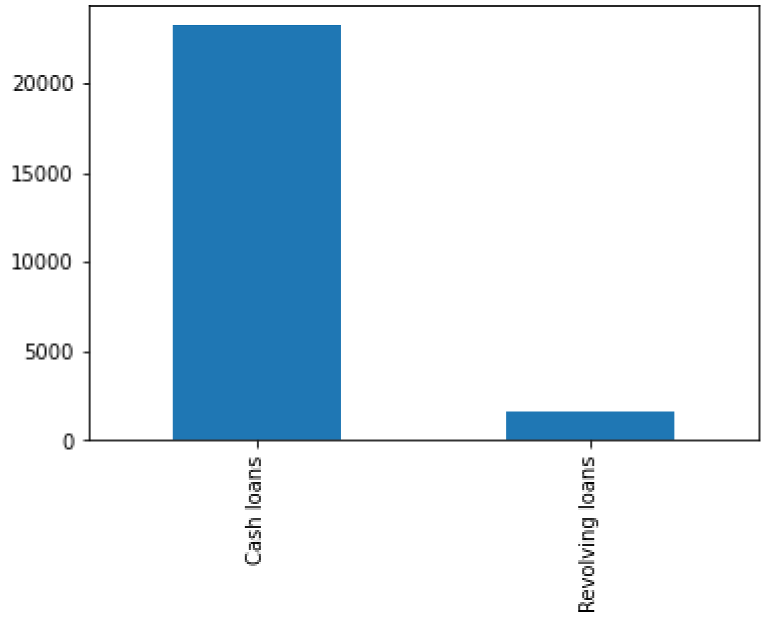


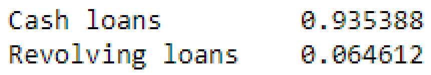
```
: F    0.570796
  M    0.429204
```

## Analysis based on the contract type of the loan of the defaulter:

Here we wanted to find out what was the most defaulted loan type. After that we wanted to see what was the percentage of the defaulters in the main population based on the different loan types.i.e for cash loans what is the percentage of people that has defaulted and so on.

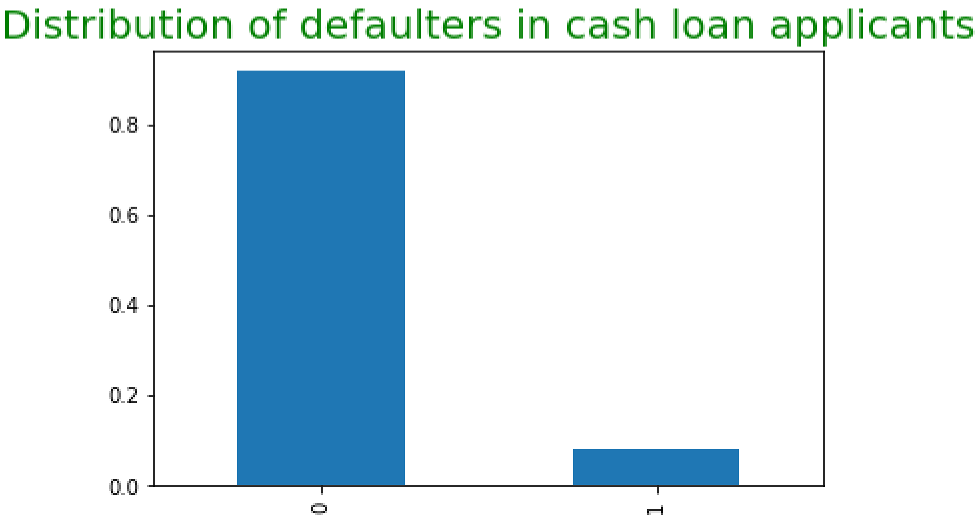
Below was the first graph to see what was the distribution of the loan type among the defaulter:

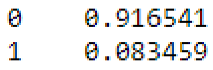




```
Cash loans         0.935388
Revolving loans    0.064612
```

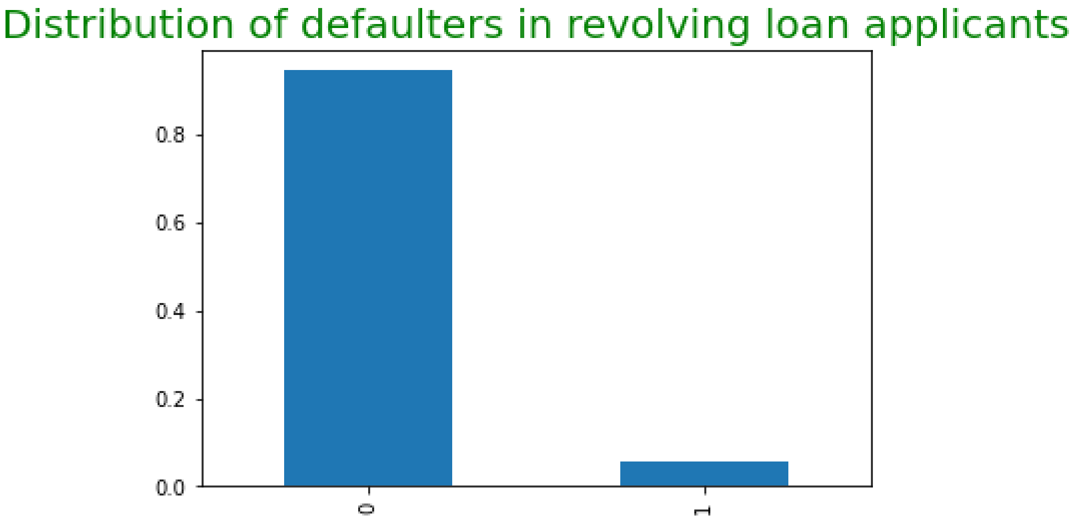
So we see that around 93 percent of the defaulters were applicants of cash loans and around 6.5 percent were revolving loans. In the next graph we will see an analysis based on the cash loan applicants

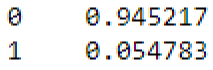




```
0    0.916541
1    0.083459
```

This shows us that of all the cash loan applicants there were around 8 percent defaulters. For revolving loans the percentage of non defaulters were



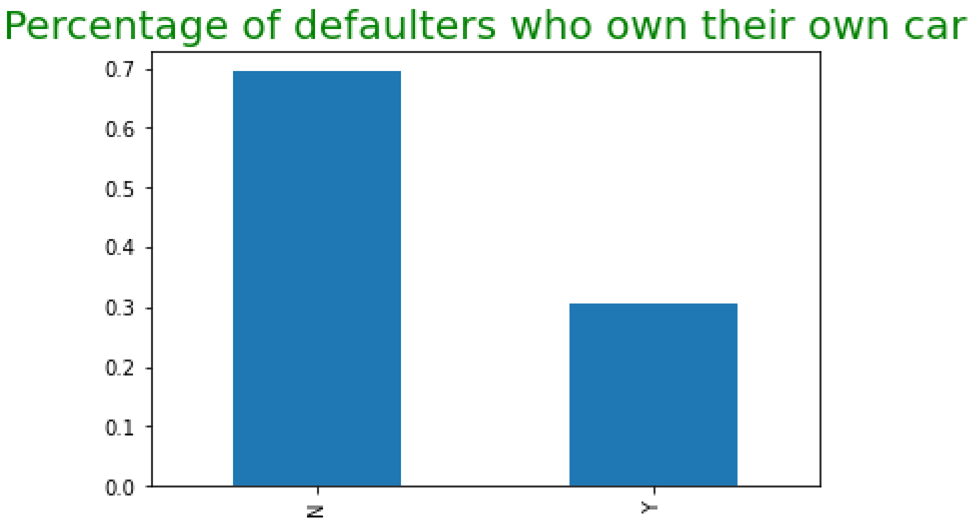


```
0    0.945217
1    0.054783
```

Here we can see that only 5.4 percent of the people  who have taken revolving loans are defaulters

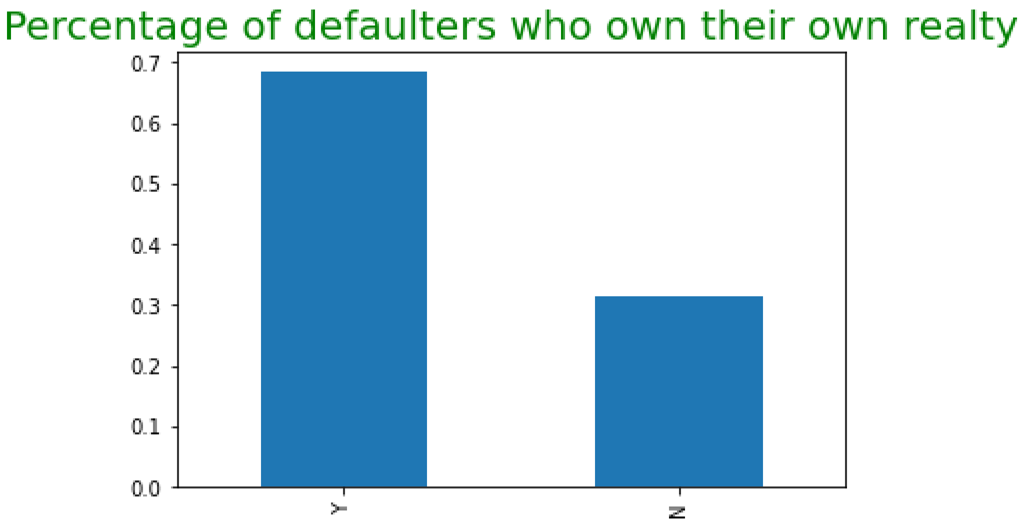
### Analysis based on whether the defaulters own their own car:

By this analysis we see that around 2/3 rd of the defaulters doesn't own their own car. Below is the graph for this:



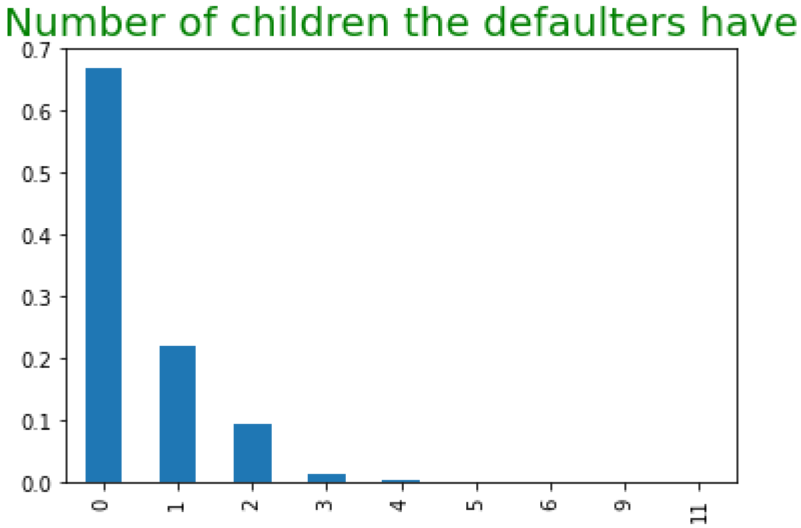
### Analysis based on whether the defaulters own their own realty:

By this analysis we see that around 2/3 rd of the defaulters own their own car. Below is the graph for this:



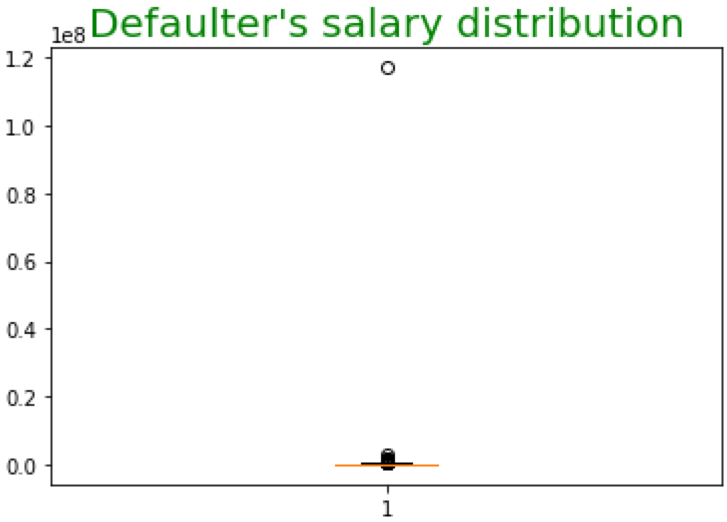
### Analysis on the number of children of the defaulters:

Here we tried to see the number of children the defaulters have. We see that almost 67 percent of the defaulters have no children.

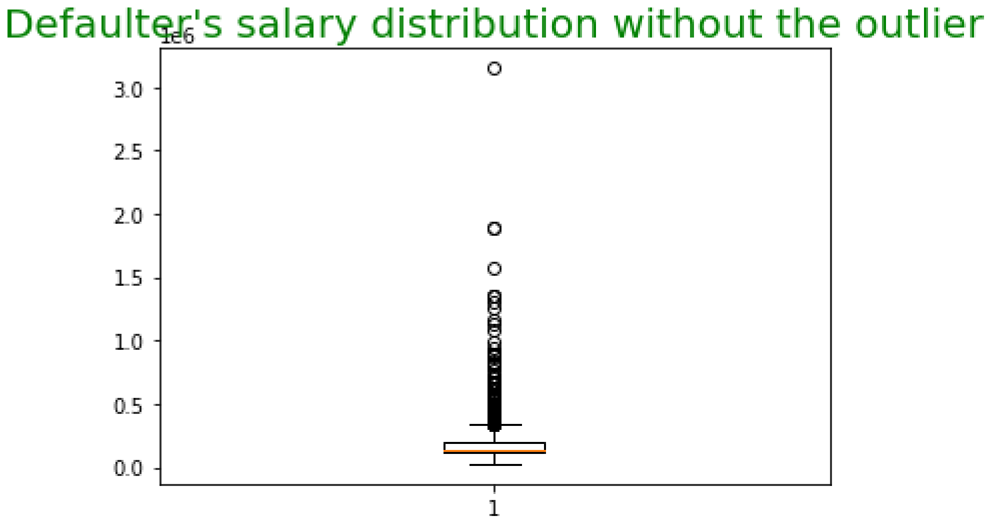


### Analysis based on the salary of the defaulters:

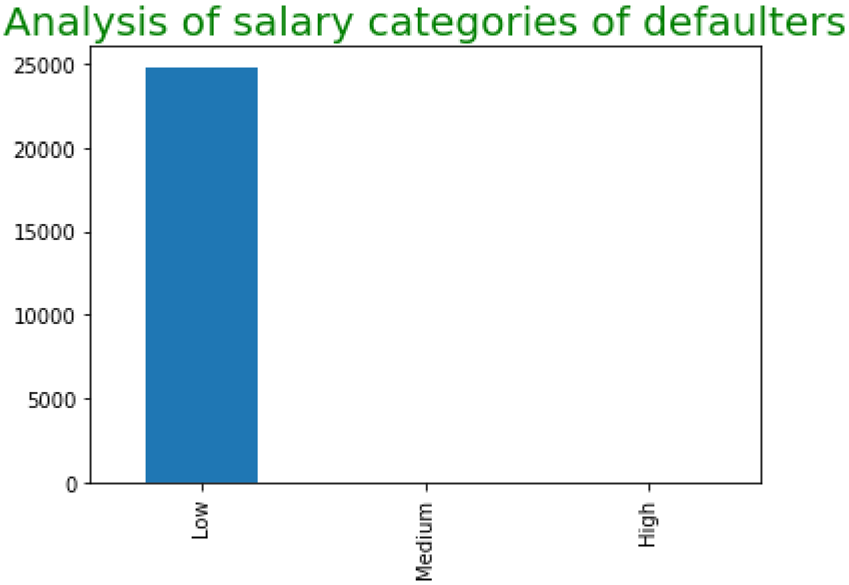
Here we saw the salary distribution of the defaulters. Below is the graph we came upwith:



As we can see that apart from 1 person most of the people are having low salaries. So we tried to get the boxplot without the outlier



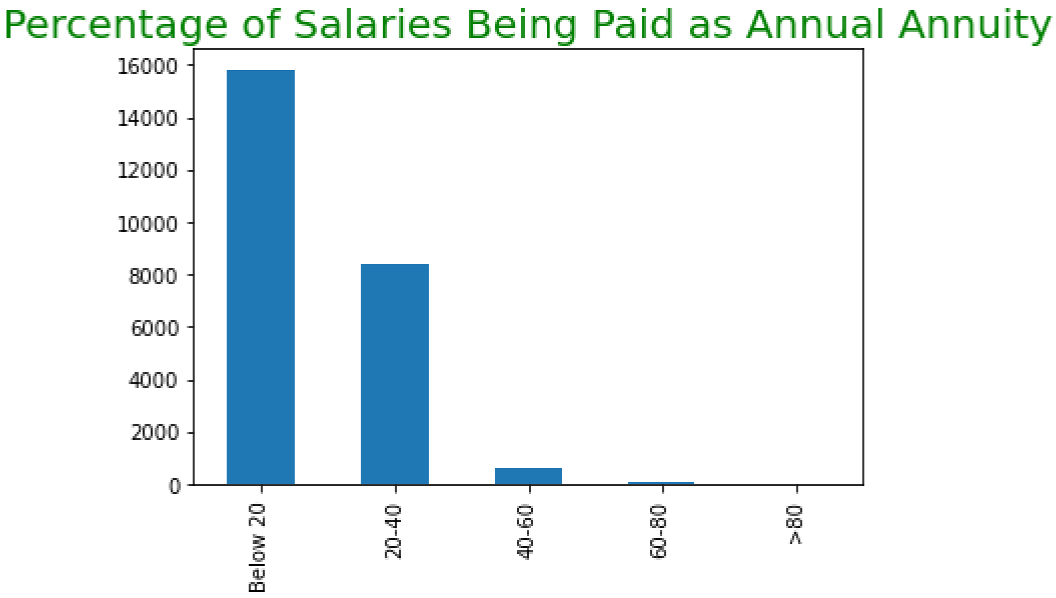
Even this doesn’t help. So we tried binning. We created 3 separate categories _low, medium and high. Below is the distribution:



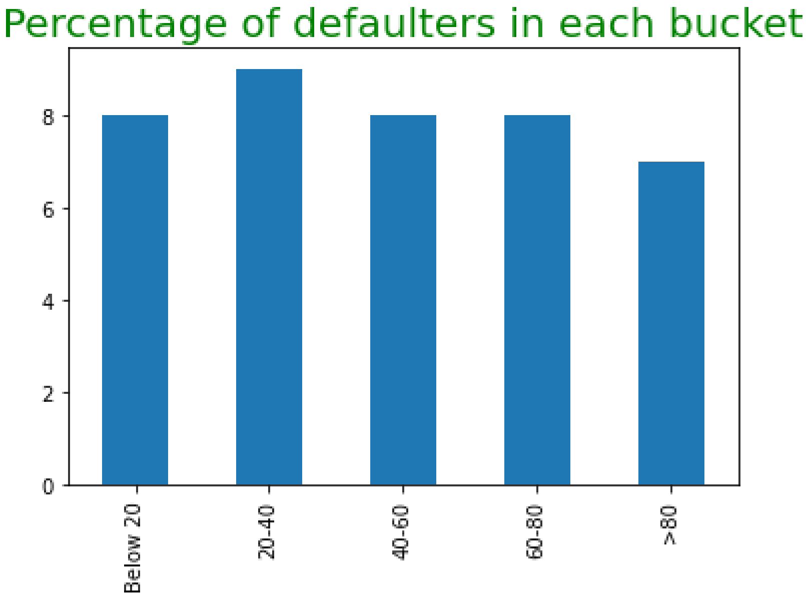
So we see that most of the people are from the low income group.

## Analysis on percentage of salary being paid as Annual Annuity for each defaulters

Here we wanted to see what is the percentage of salary that is paid as the Annual annuity every year.We came up with this graph:


Percentage of Salaries Being Paid as Annual Annuity

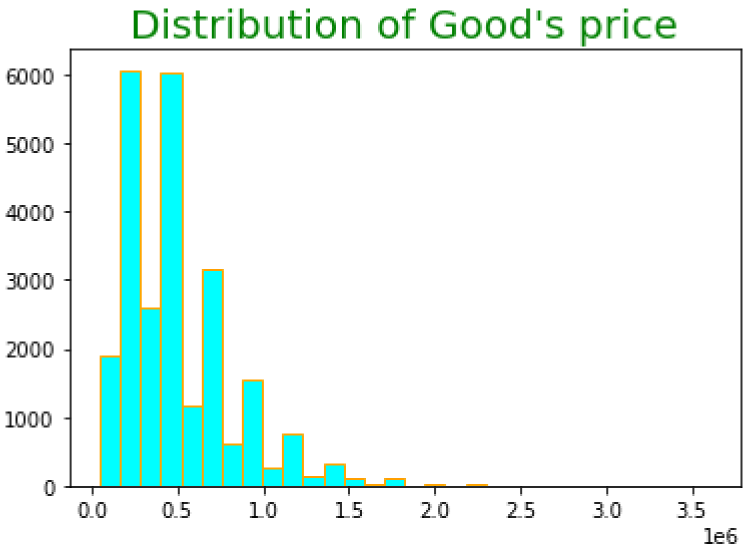
We can see most of the people who defaulted pay below 20 percent of their salary as the annual annuity. I wanted to go one step more in this direction and see what is the percentage of defaulters in each bucket .


Percentage of defaulters in each bucket

Here  we see that people who pay 20 to40 percent of their salary as annual annuity has a slightly more chance of becoming a defaulter.

### Analysis based on the price of the good for which the loan was taken:

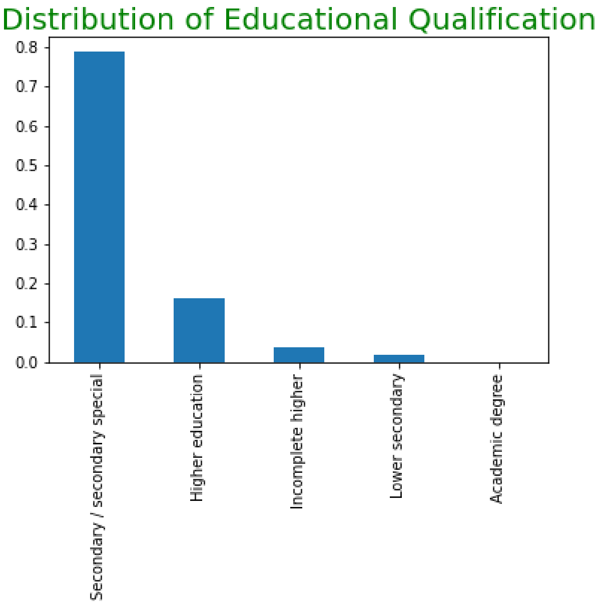
For this we created a histogram, with all the good’s price. Belowis the histogram:



We see that most of the goods costs between 100000 to 750000.

### Analysis based on the educational qualification of the defaulters:

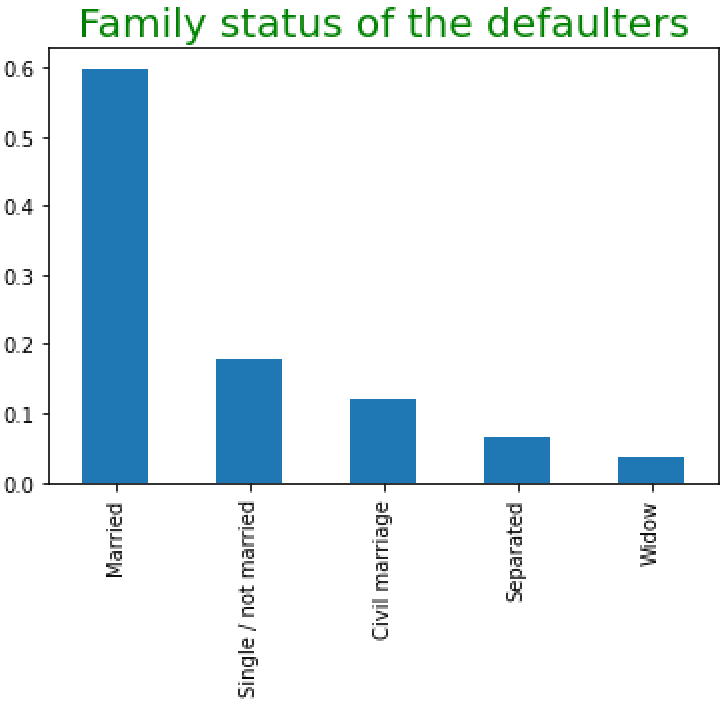
For this we created a graph to see the distribution of different qualifications among the defaulters. Below is the graph:



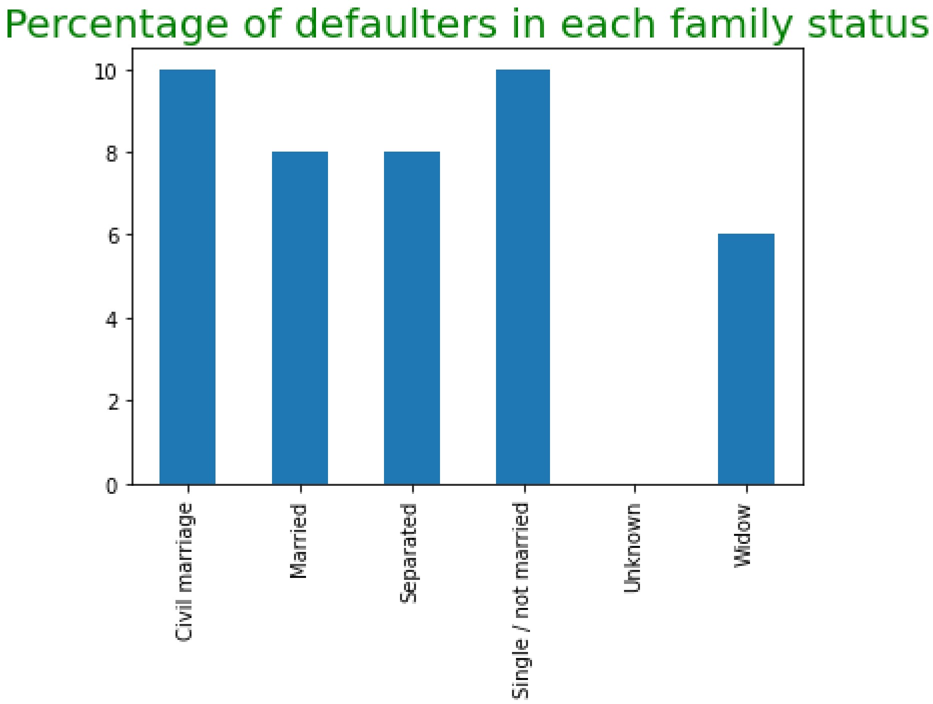
As we can see that almost 80 percent of the defaulters are Secondary or Secondary special qualified

### Analysis based on the family status of the defaulters:

I came up with the below graph:



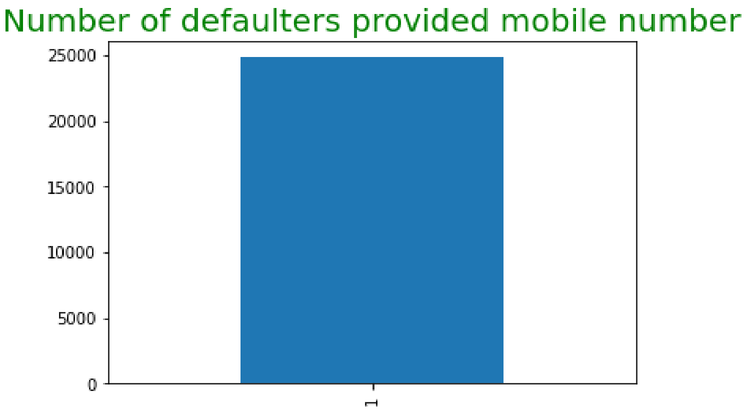
We see that almost 60 percent of the defaulters are married. Lets go one step deeper  and see what is the percentage of defaulters in each falily statuses. We the below graph:



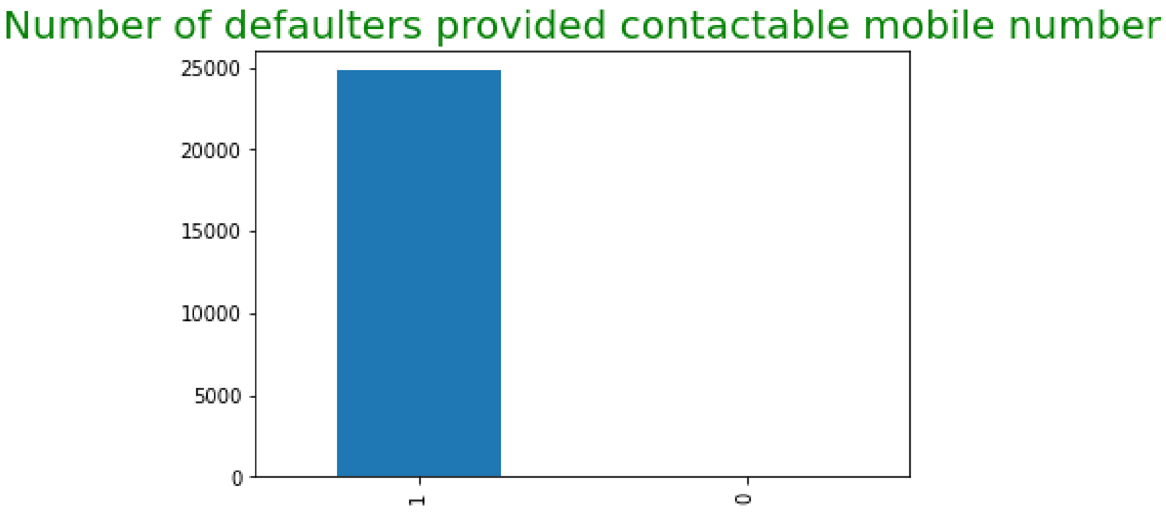
We see a very interesting trend over here that people with civil marriage and people who are single default around 10 percent of the times whereas married people default only around 8 percent of the time. So though the number of married people in defaulters is more but a far more problematic family status is that of civil marriage and single people.

### Analysis based on whether the defaulters provided their mobile number :

With this analysis we saw that everyone of the defaulters provided a mobile number. Below is the graph:



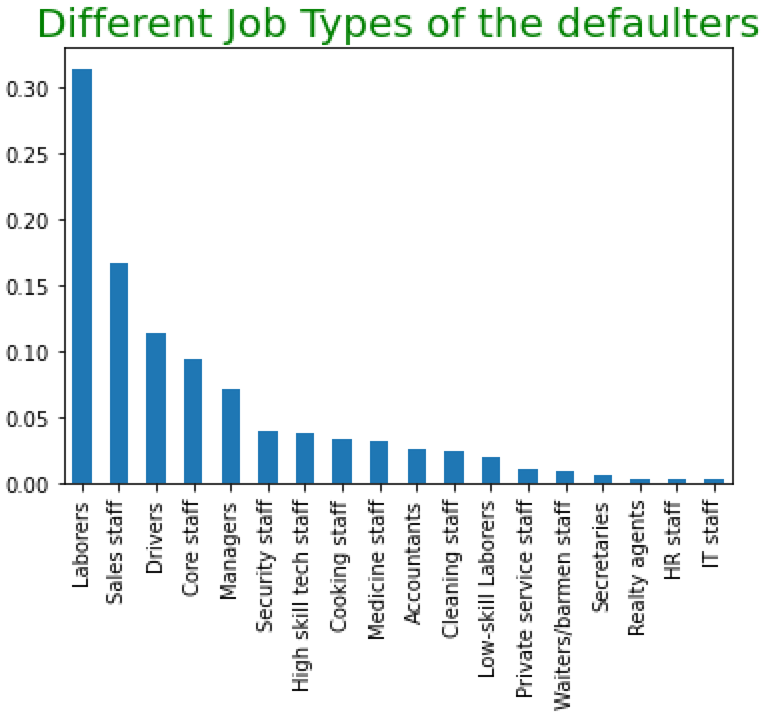
### Analysis based on whether the defaulters provided contactable mobile number :



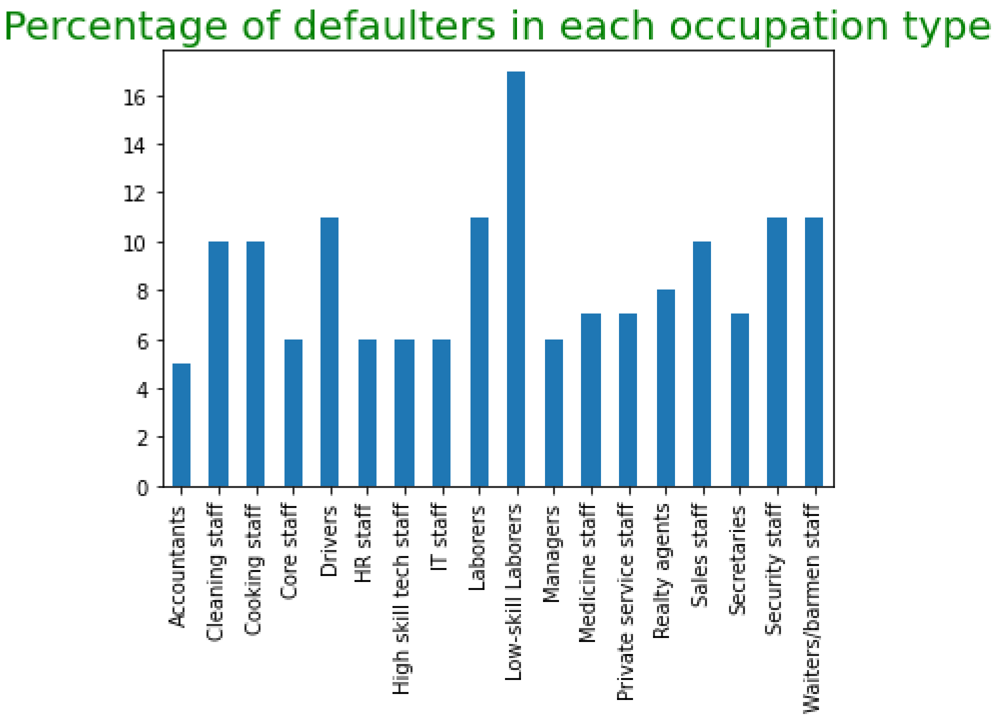
Again we see that almost everyone provided a contactable mobile number barring a negligible few.

### Analysis based on the several job types of the defaulters:

We tried looking for the different  job types of the defaulters:

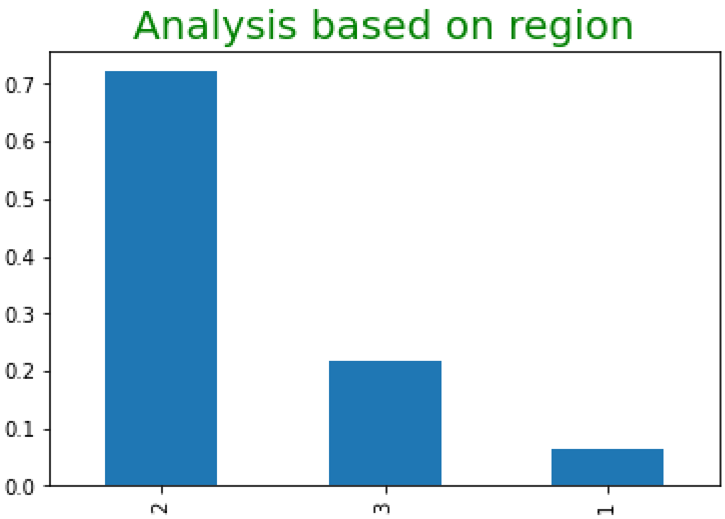


We see that around 30 percent of the defaulters are labourers. Lets try to find out what is the percentage of defaulters in each occupation type:

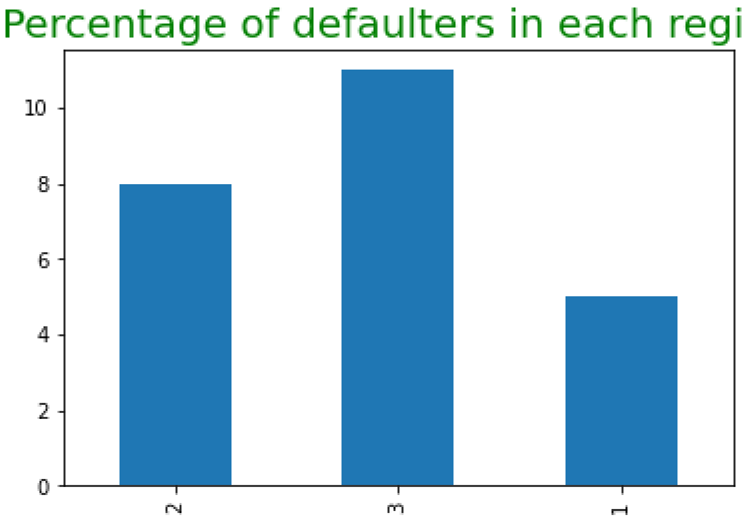


Here we see a slightly different story. We see the people from Low skill laborers  are more probable to default.  Largely we see that most of the low income jobs have very high default rate. These jobs are low skill laborers(16 percent),Labourers,Waiters/Barmen staff,Drivers(10 percent each).

### Analysis based on region:

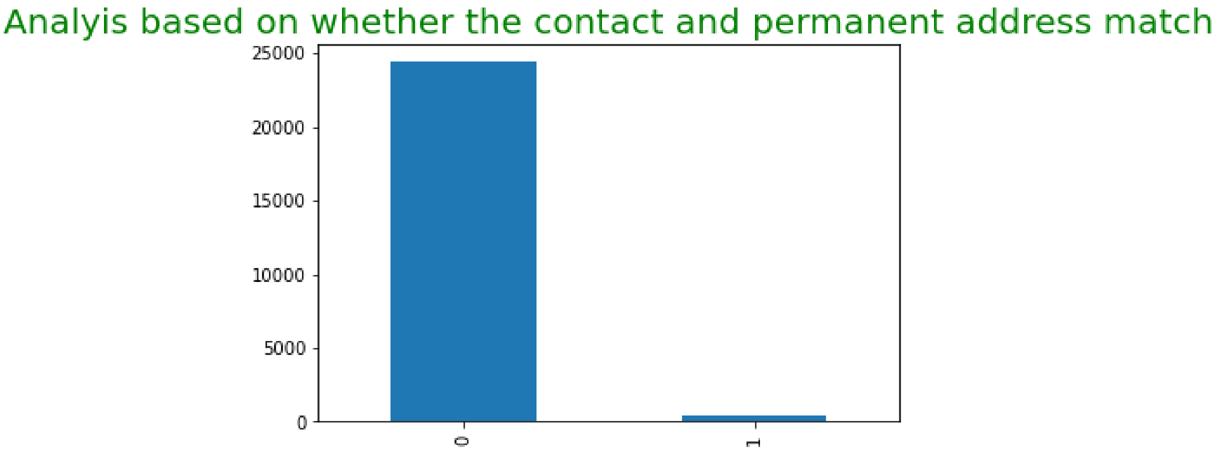


We see that most of the defaulters (around 70 percent of the defaulters) come from region 2. Now let's see what the percentage of defaulters from the regions is. We came up with this graph.



We see that region 3 is most probable to default and region 1 is the safest bet with only around 5 percent defaulters. Region 3 has more than 10.

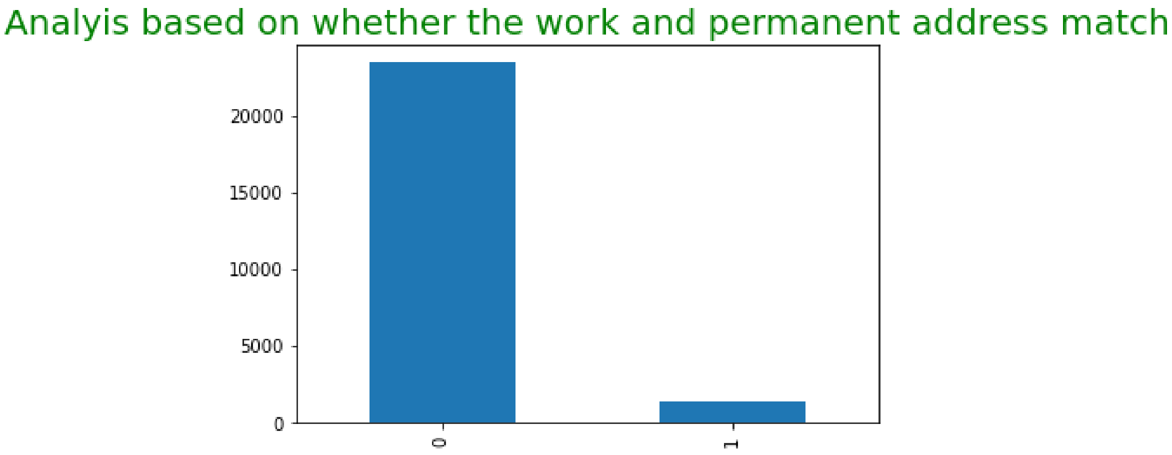
### Analysis based on whether the contact and the permanent address match:



We see that most of the contact and permanent address match

### Analysis based on whether the work and the permanent address match:

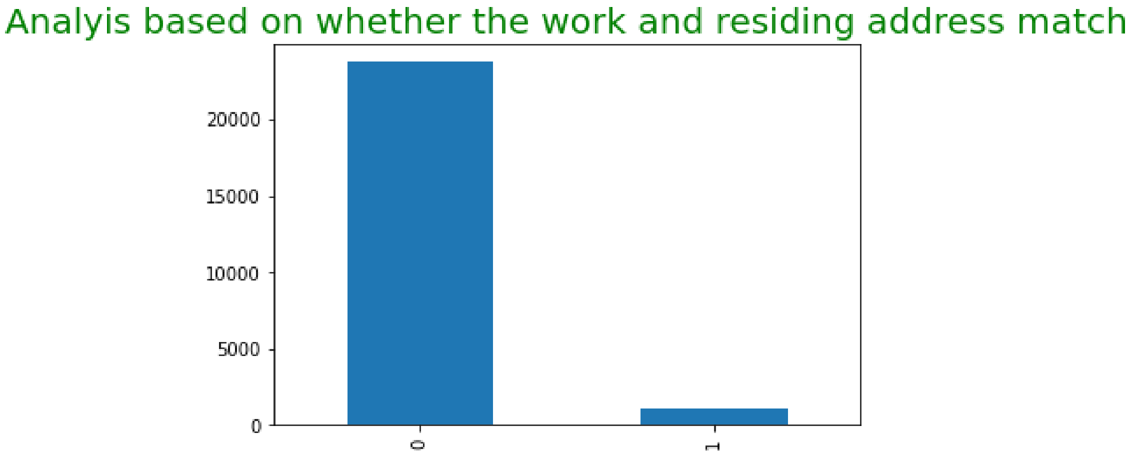
Below is the graph that we got:



Again we see that barring a negligible few the work and the permanent address match

### Analysis based on whether the work and the residing address match:

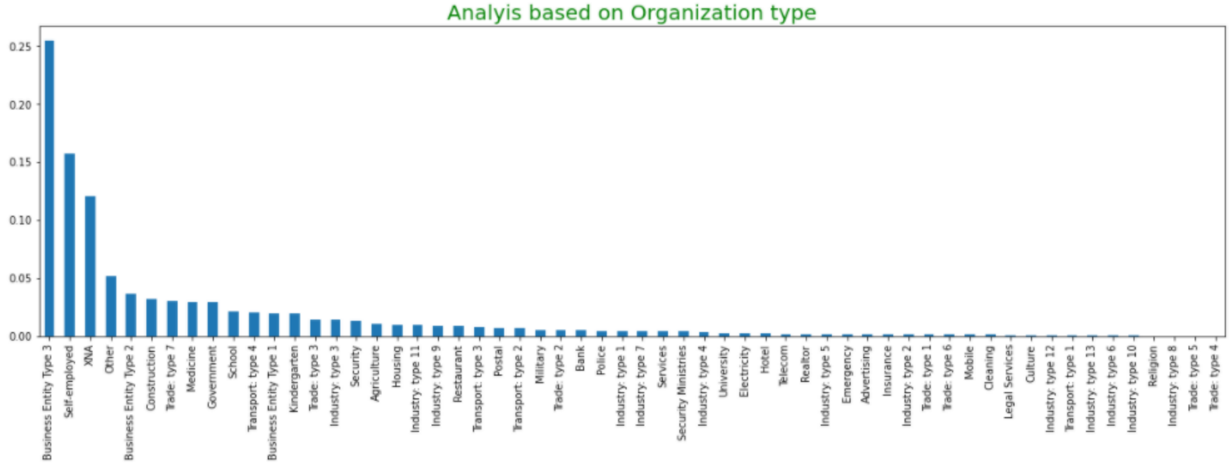
Below is the graph we got:



Again we see that almost for everyone the work and the residing region match

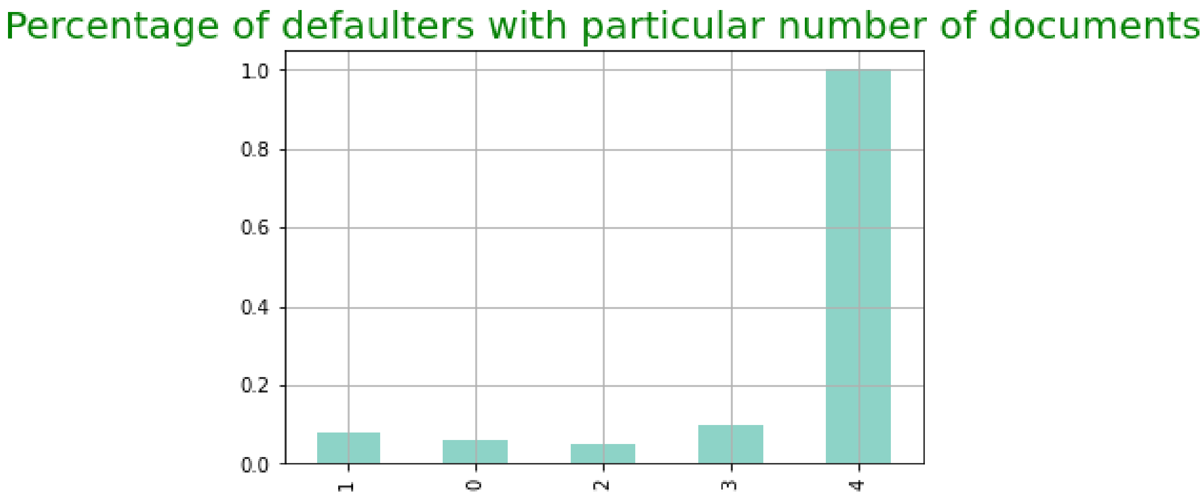
## Analysis based on the organization type:

Below is the distribution of organization type among the defaulters:



We see that the most common organization type is business entity type 3. Which constitutes around 25 percent of all the defaulters

## Analysis based on the number of documents provided:



We see here that people who submitted 4 documents all of them defaulted, people with 3 documents have defaulted around 10 percent of the times.

### Analysis of the previous application dataset:

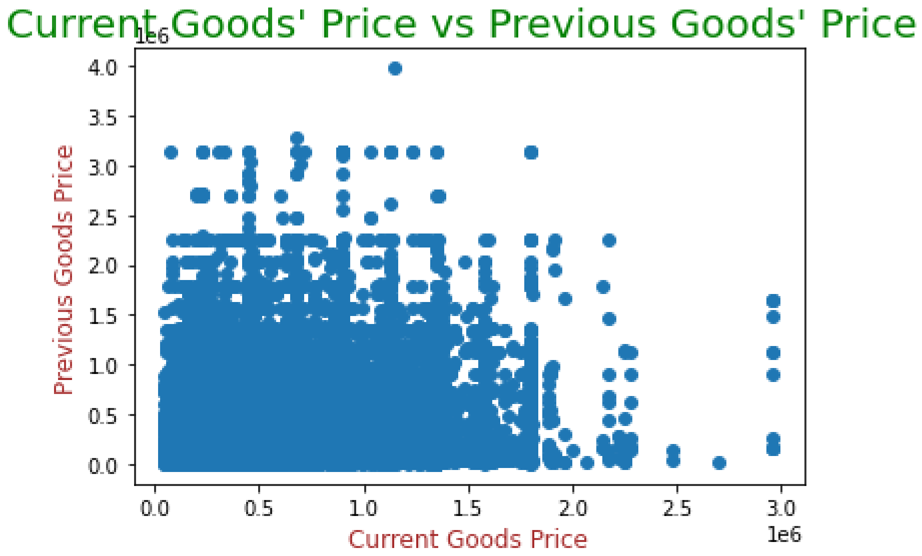
Here in this data frame we have the data which is the previous loan information of the current applicants. So I decided to join the two data frames: application data and previous data. We made this merge using the column “SK_ID_CURR”.

Next we deleted the unimportant and already analyzed columns. This resulted in dropping 39 columns. We also extracted only the defaulter records using the column TARGET. We finally had a dataframe with 68 columns and 122360 rows.

### Analysis of Current Goods price and Previous Goods price:

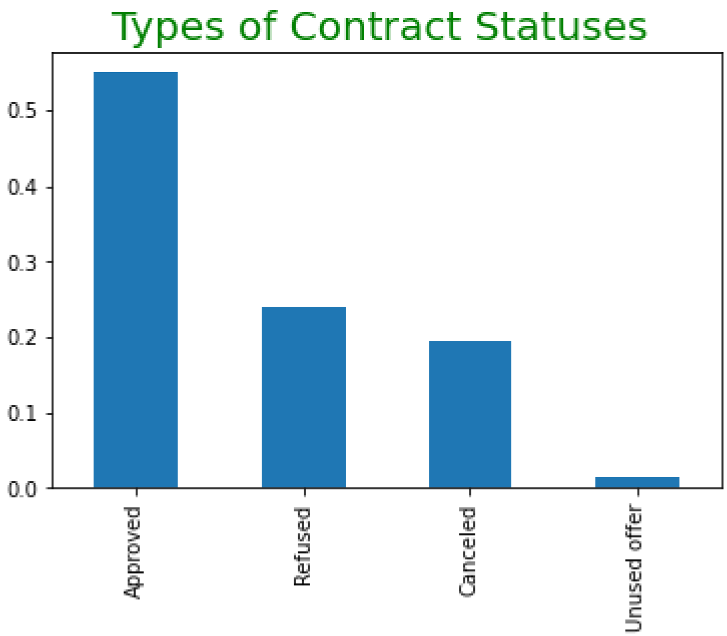
Here we compared the good’s price for which the loan was taken and the good’s price for which the loan was taken previously. We see that most of the previous and the current good’s prices are low.

We  saw the below graph:



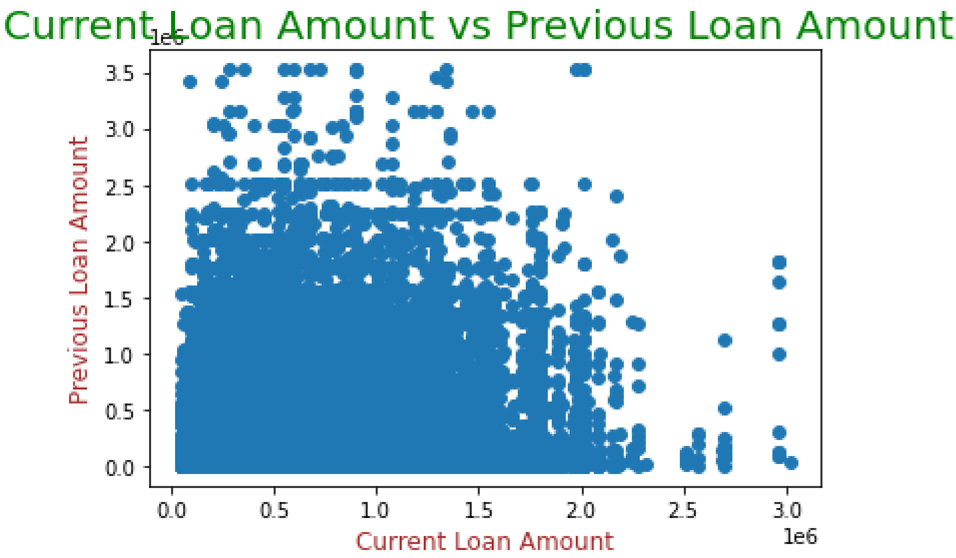
### Analysis of based on previous contract statuses:

We see that more than 50 percent of the previous loan applications were approved. Below is the graph:

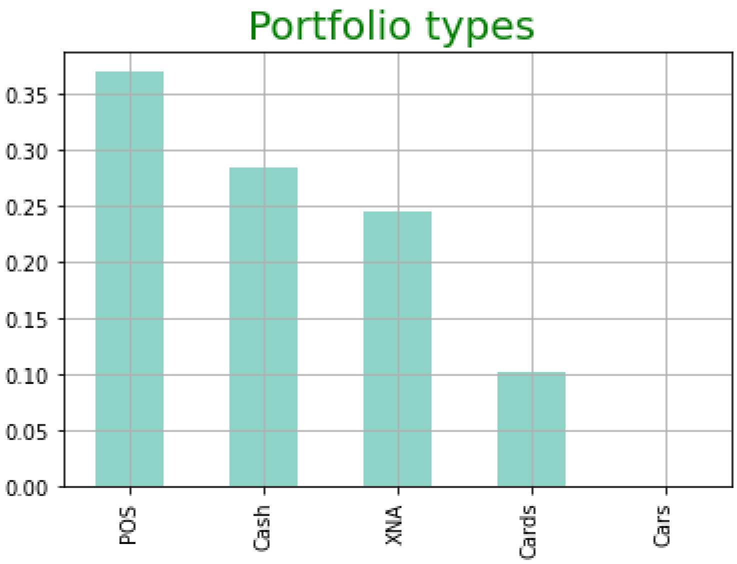


### Analysis of Current Loan Amount and Previous Loan Amount:

Again we see that the current and previous loans were small loans in amount apart from the few ouliers

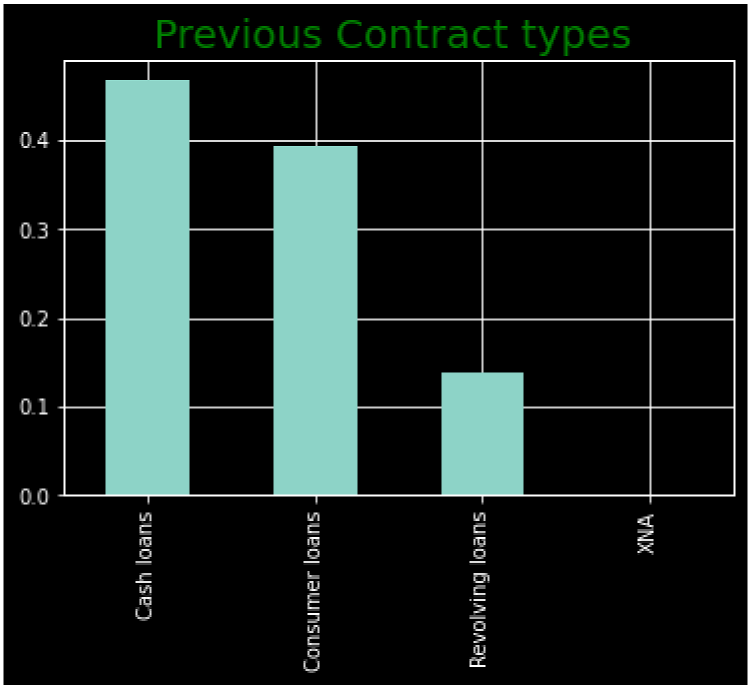


### Analysis based on the portfolio of the previous applications:

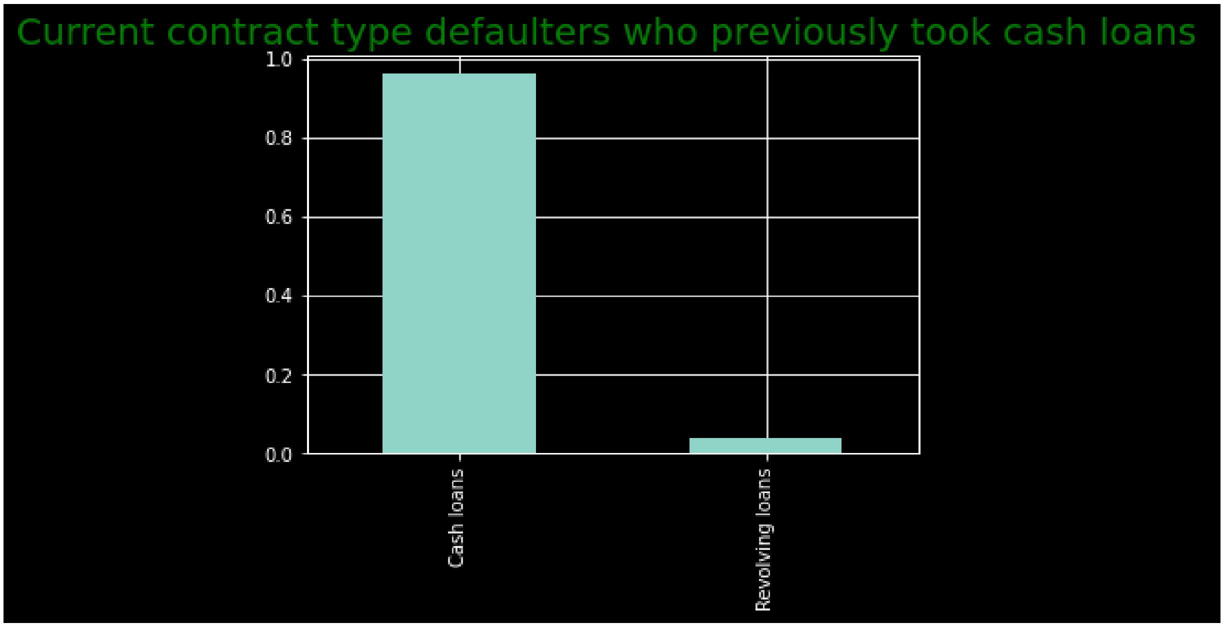


We see here that most of the portfolio typesof the previous loans were POS.

### Analysis based on the portfolio of the previous contract types:

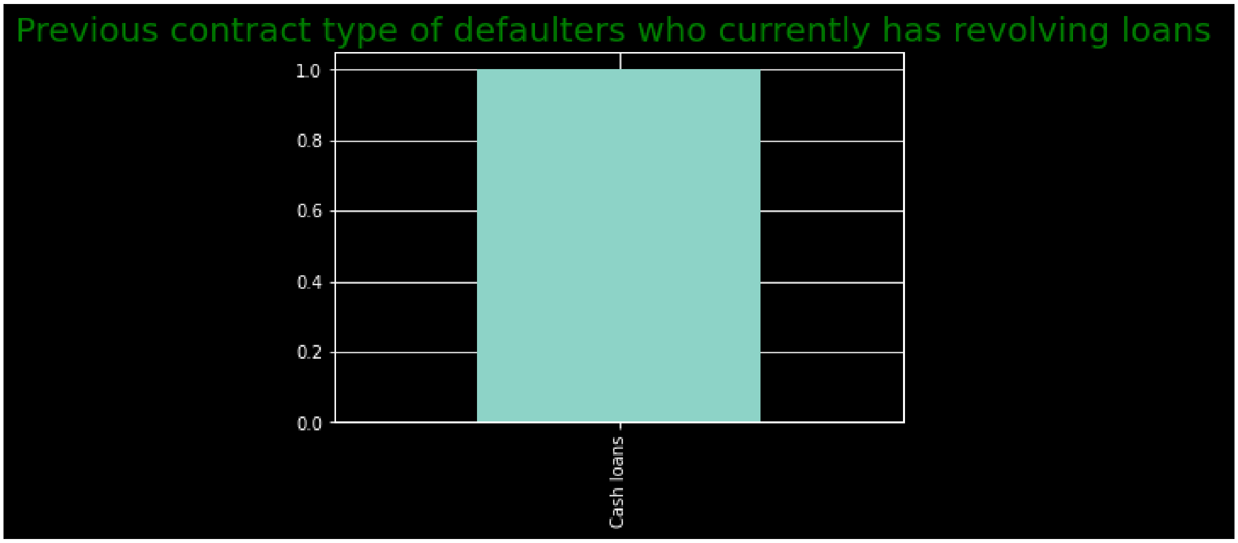


We see here that the previous contract types of the defaulters were mostly cash loans(around 47 percent) and consumer loans(39 percent). So I checked what is the current  loan type of people whose previous loan type was cash. We came up with the below graph:



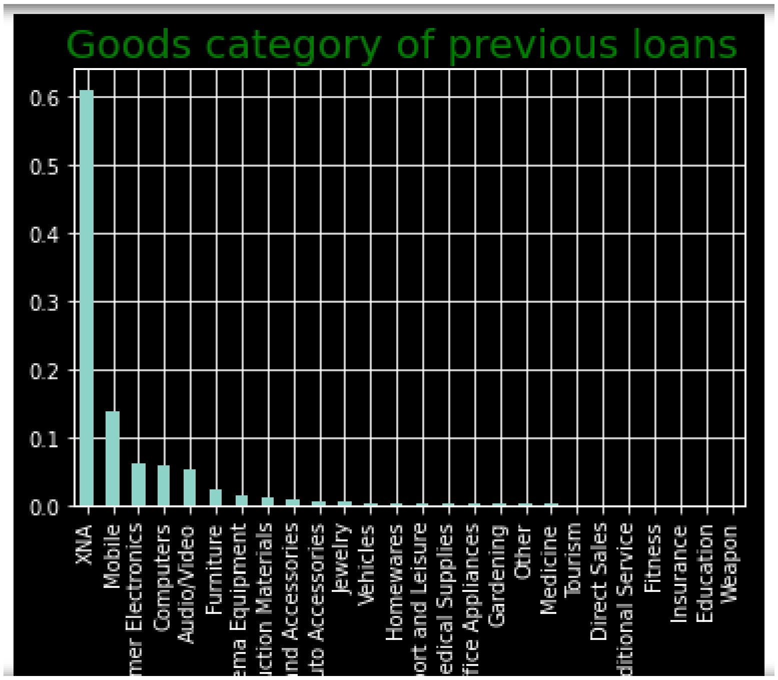
We see most of the defaulters who took a cash loan previously mostly take a cash loan again.

Let us now see what is the what was the previous contract typeof the defaulterswho have currently defaulted using revolving loans. We saw the below graph:



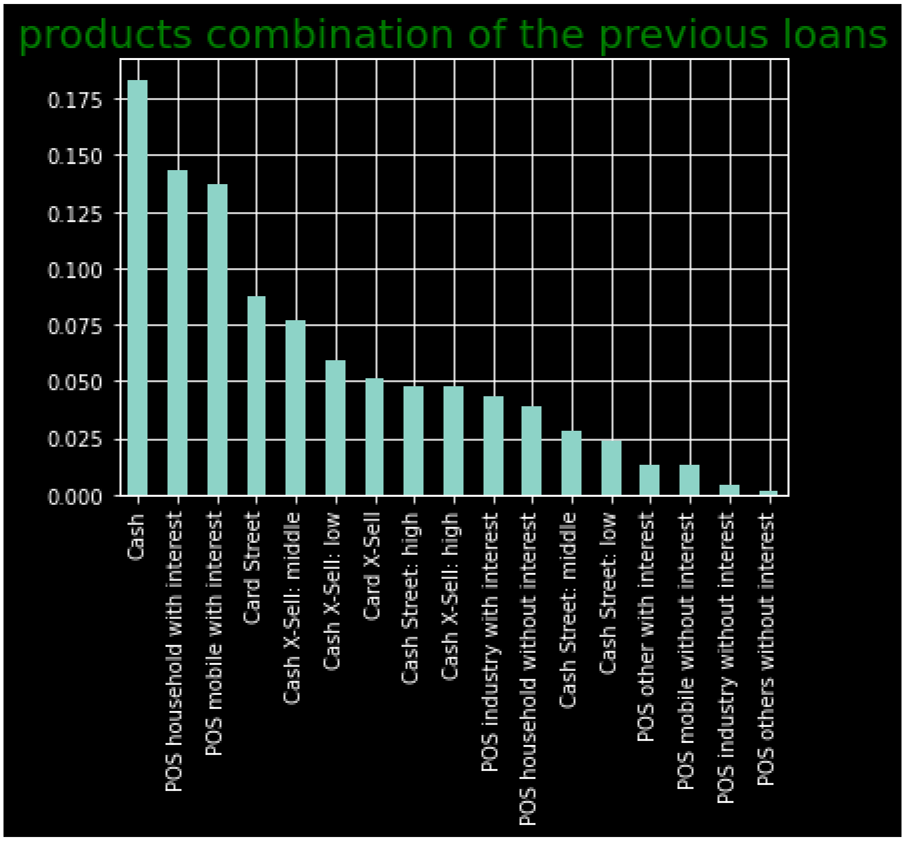
We see here that all the people who has taken cash loans previously and then turned to revolving loans later on has defaulted.

### Analysis based on previous loans' good's category:



We see most of the good's category was not mentioned(XNA). The next category is mobile (15 percent)

### Analysis based on previous loans' good's category:



We see that most of the people(17.5 percent) went for the products combination cash.

## Conclusion:

1. Most of the people who have defaulted are the ones who come from a low income jobs(watchmen,barman,etc)
2. These people most of the times pay around 20 to 40 percent of their salaries as annual annuity.
3. We also saw that most of the defaulters were Female and with a secondary and secondary plus education
4. The defaulters are probable to come from region 3
5. The defaulters mostly end up taking multiple small loans
6. People with a previous cash loan, when they turn to revolving loan they are almost certain to default.