Nubank Data Challenge - Exploratory Analysis

by Adriano Freitas

This exploratory analysis aims to better understand the Nubank client dataset to help build machine learning algorithms to predict: This exploratory analysis aims to better understand the Nubank client dataset to help build machine learning algorithms to predict:

- Probability of default
- Probability of fraud
- Probability of spending

In the end, these algorithms will help us build a template to approve or not a client and set an initial limit for your card.

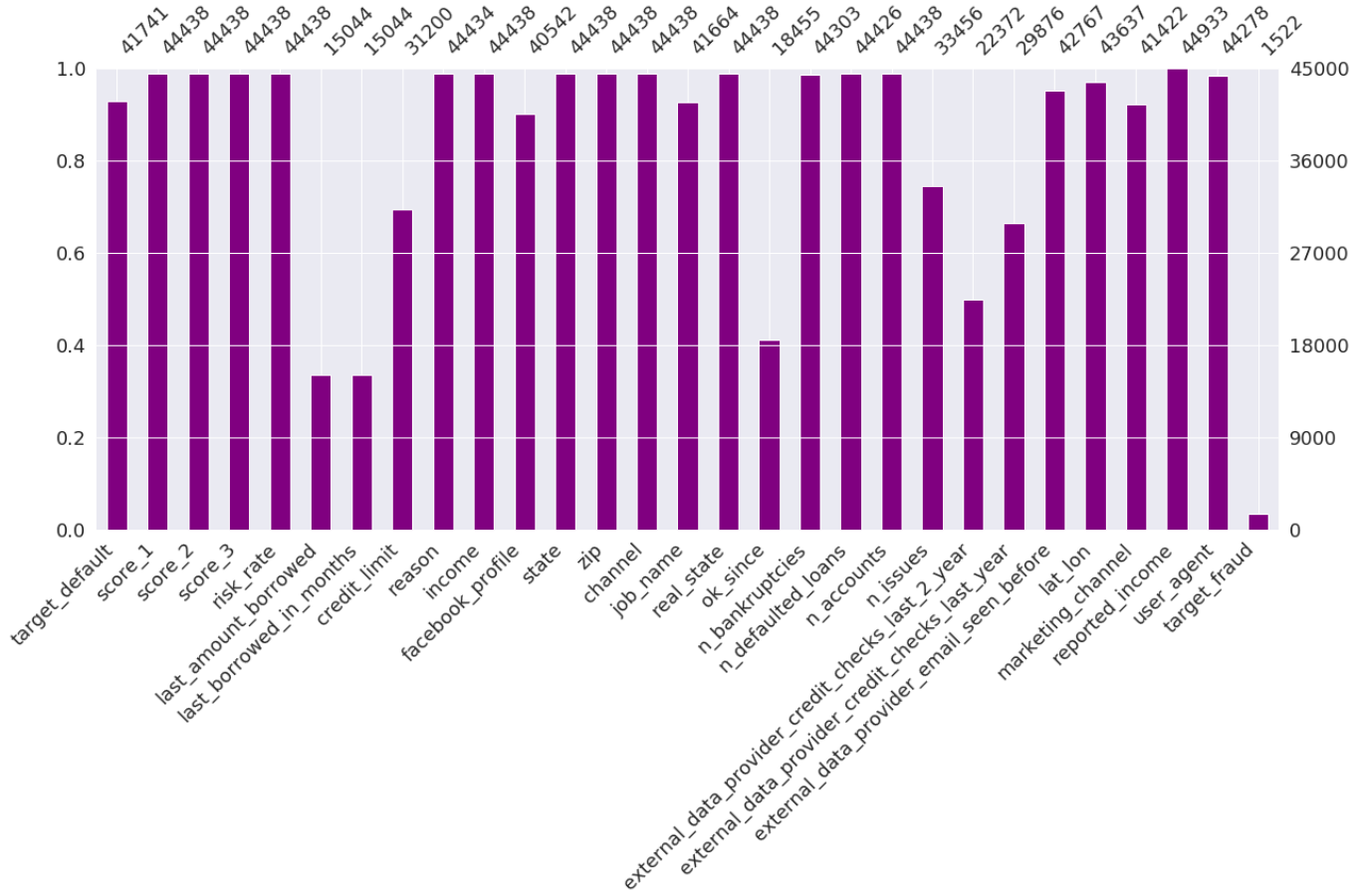
Checking Missing values

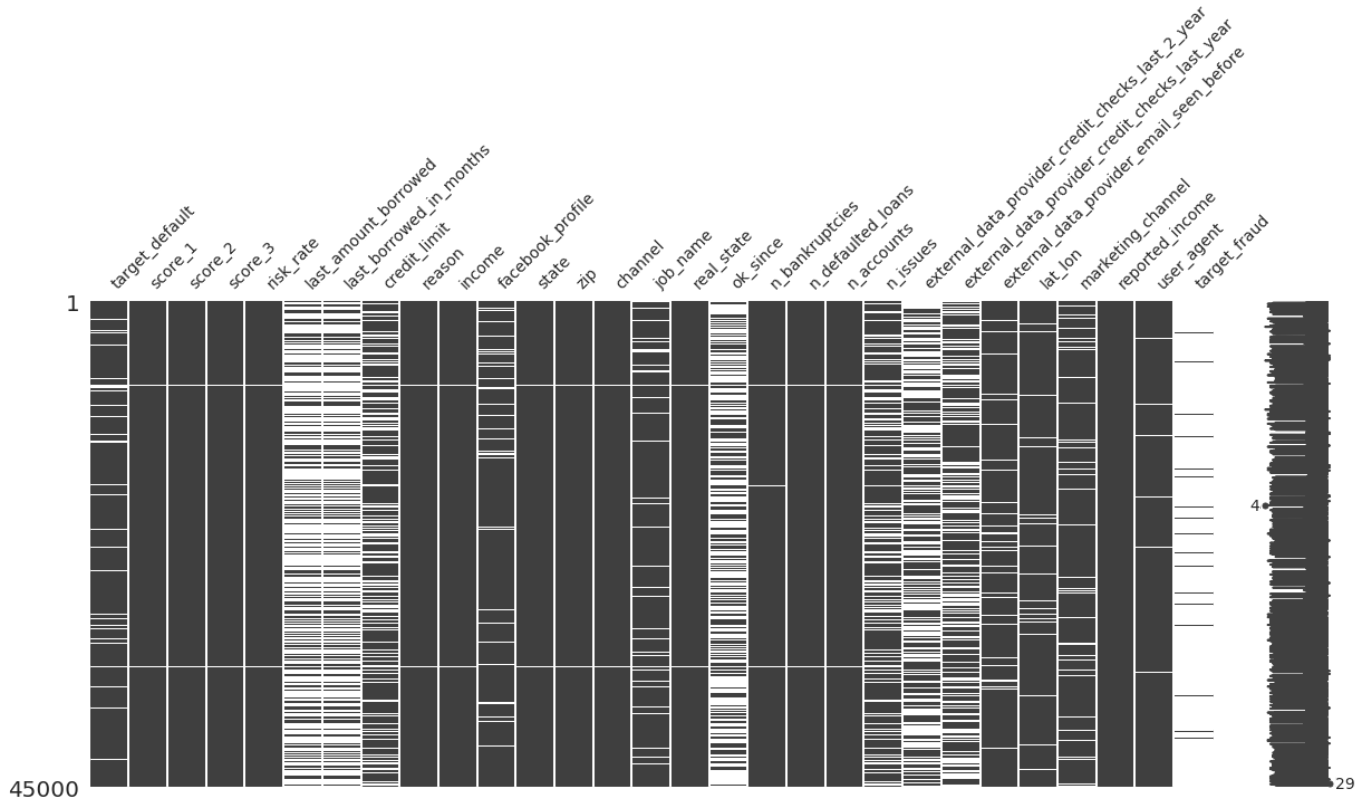
Let's take a look inside the missing values on dataset

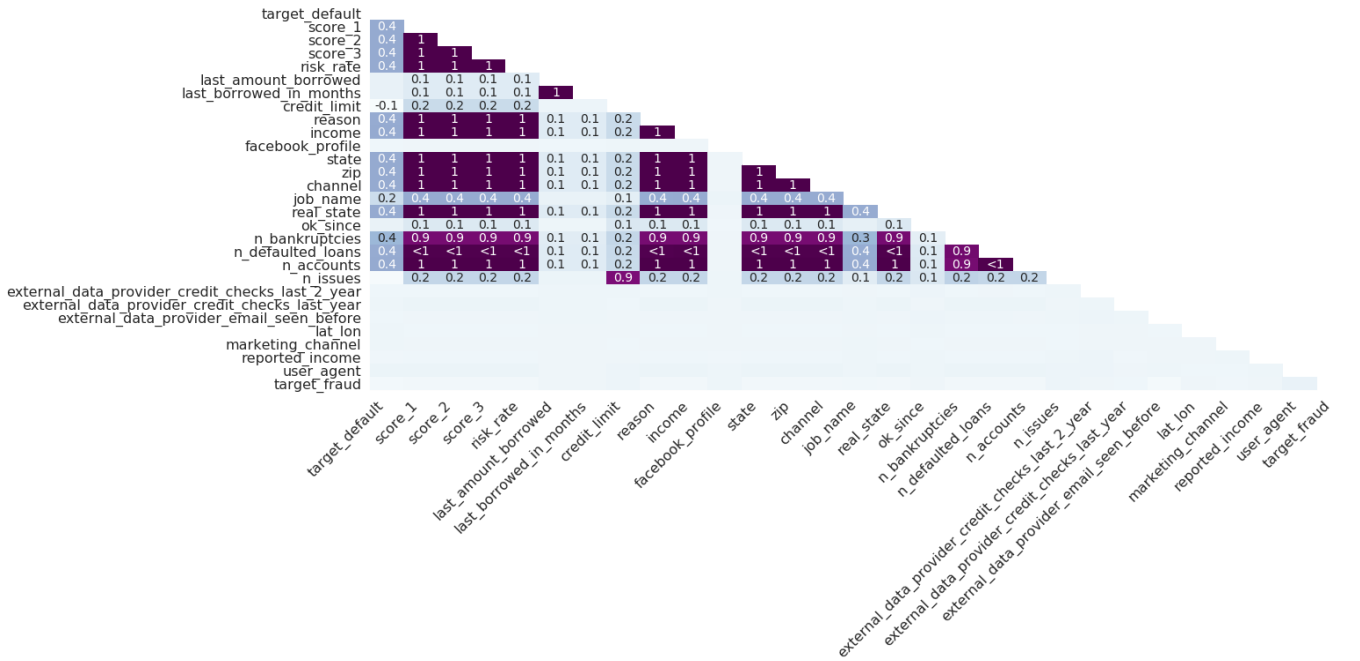
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f202f73f128>
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2037505eb8>
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f203540fc50>
```







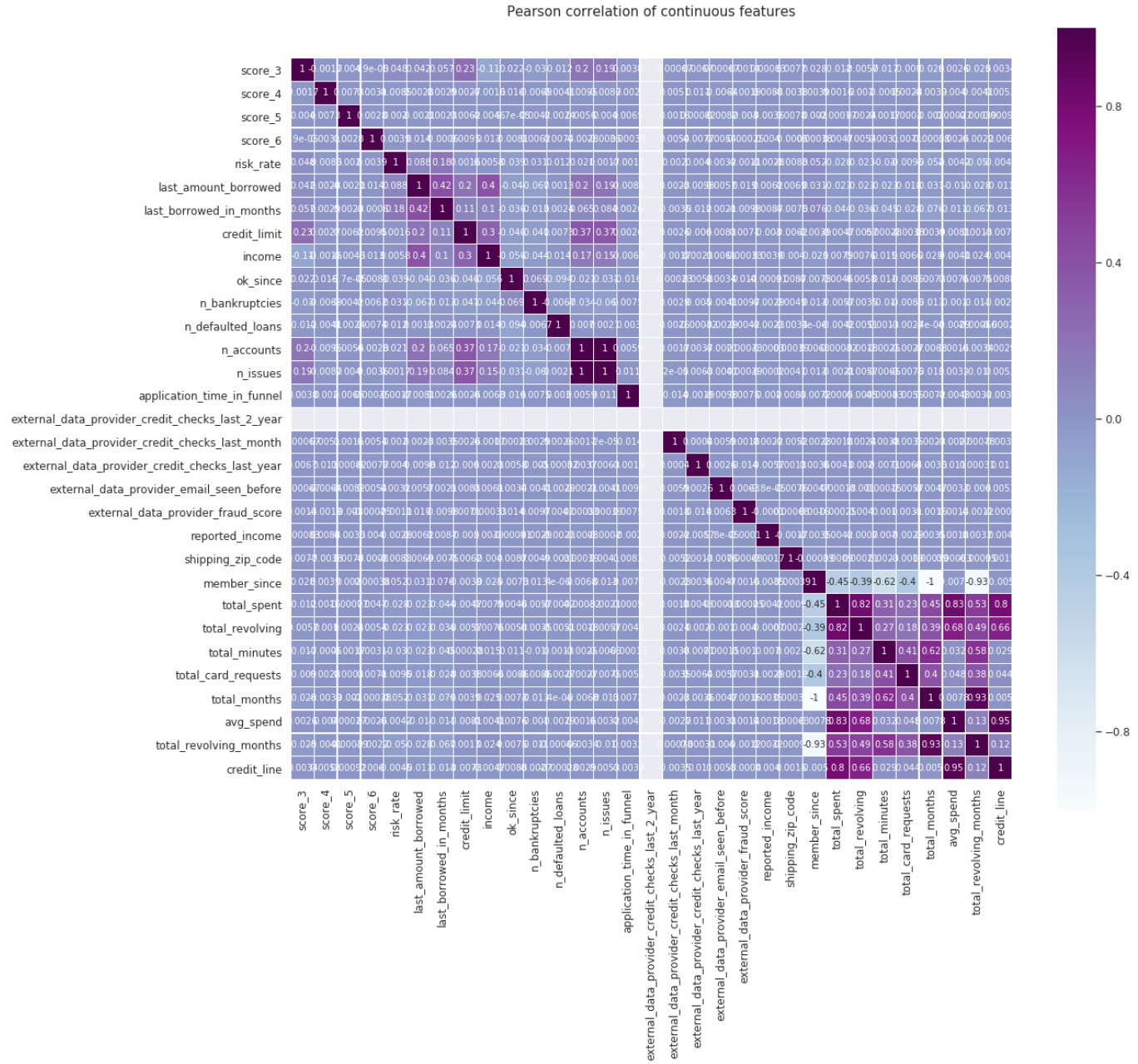
Pearson correlation matrix

As we can see, there is no much linear correlation between continuous features, except for some calculated ones.

```
<Figure size 1296x1152 with 0 Axes>
```

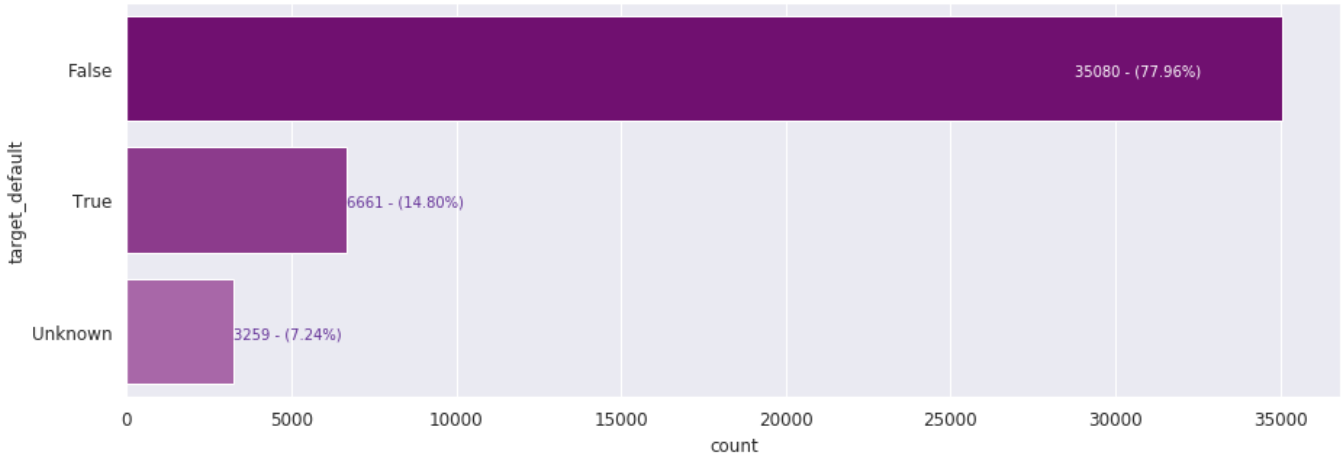
```
Text(0.5,1.05,'Pearson correlation of continuous features')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f20354cb240>
```



Default Analysis

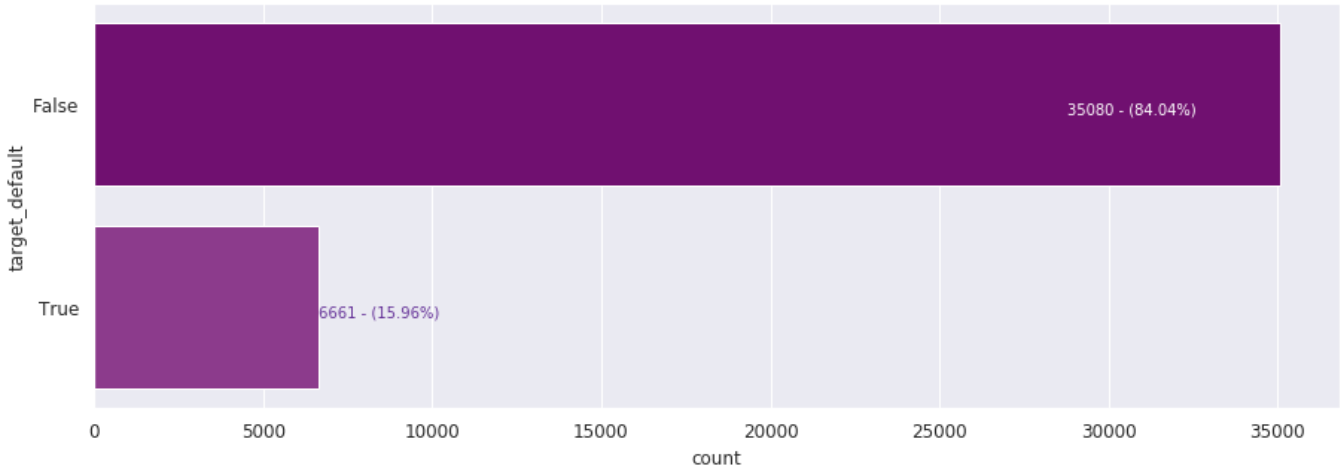
```
AxesSubplot(0.125,0.125;0.775x0.755)
```



Unkown data

There is a small number of observations on our dataset wich we don't know if was defaulted or not, let's exclude them

```
AxesSubplot(0.125,0.125;0.775x0.755)
```



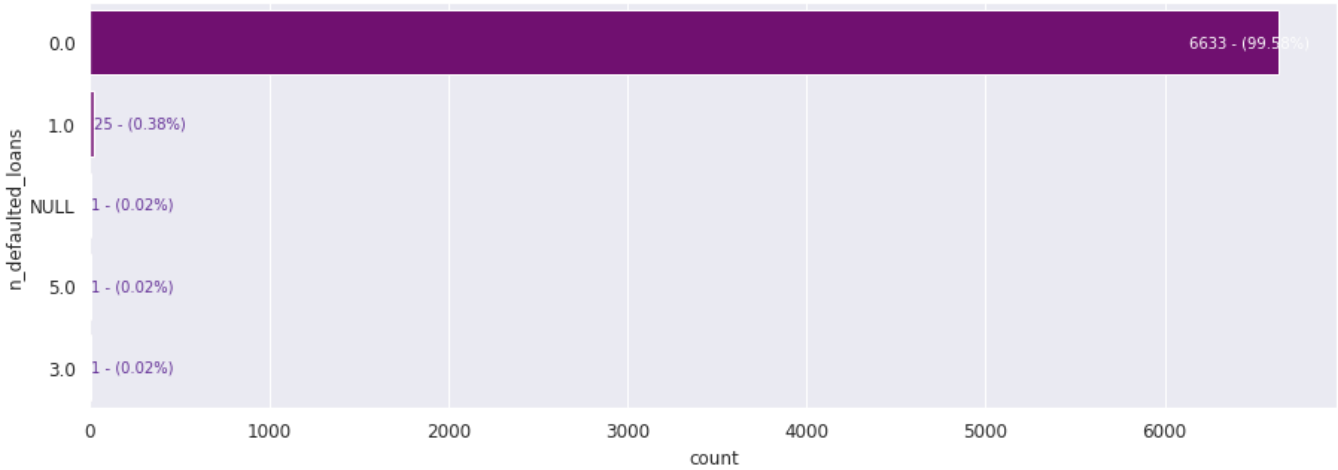
Deeper default analysis

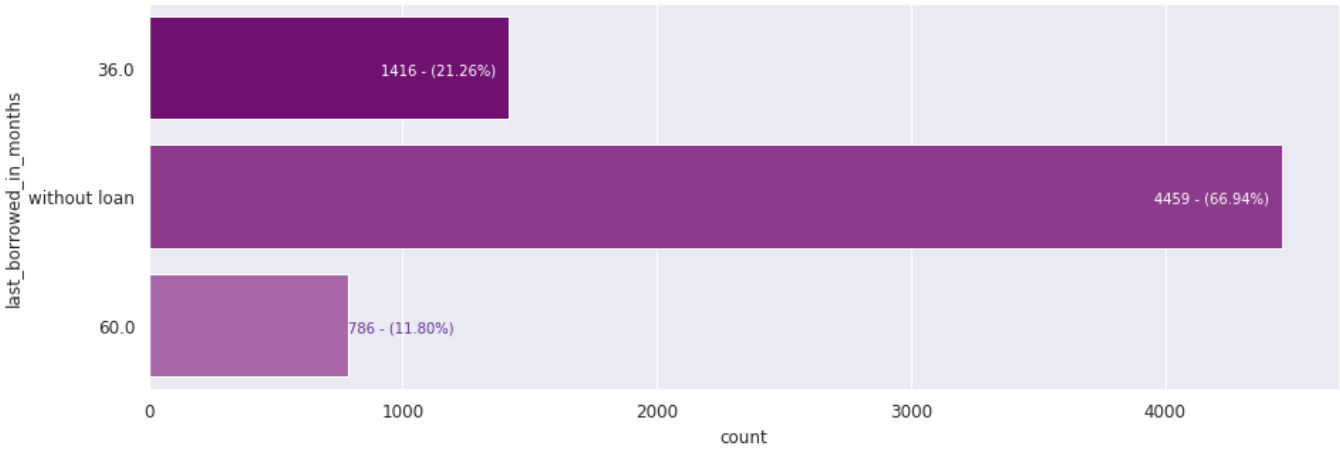
Let's take a close look into the default data, so let's separate only default data into another dataframe.

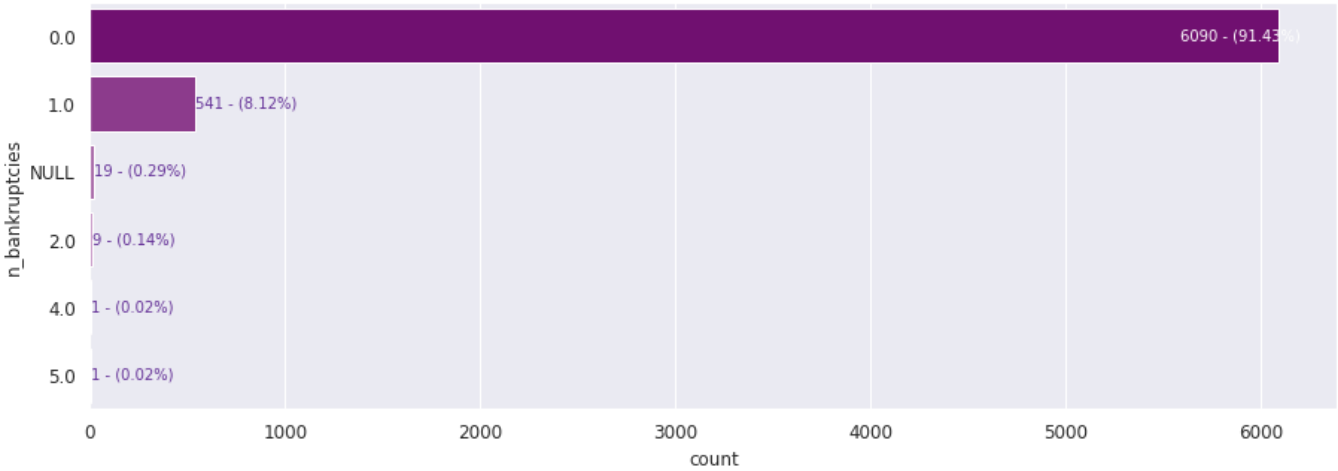
Default history

As we can see, most of the defaulters had not previously been defaulters, nor had they faced bankruptcy, and only a small minority had previously borrowed.

```
AxesSubplot(0.125,0.125;0.775x0.755)
AxesSubplot(0.125,0.125;0.775x0.755)
AxesSubplot(0.125,0.125;0.775x0.755)
```



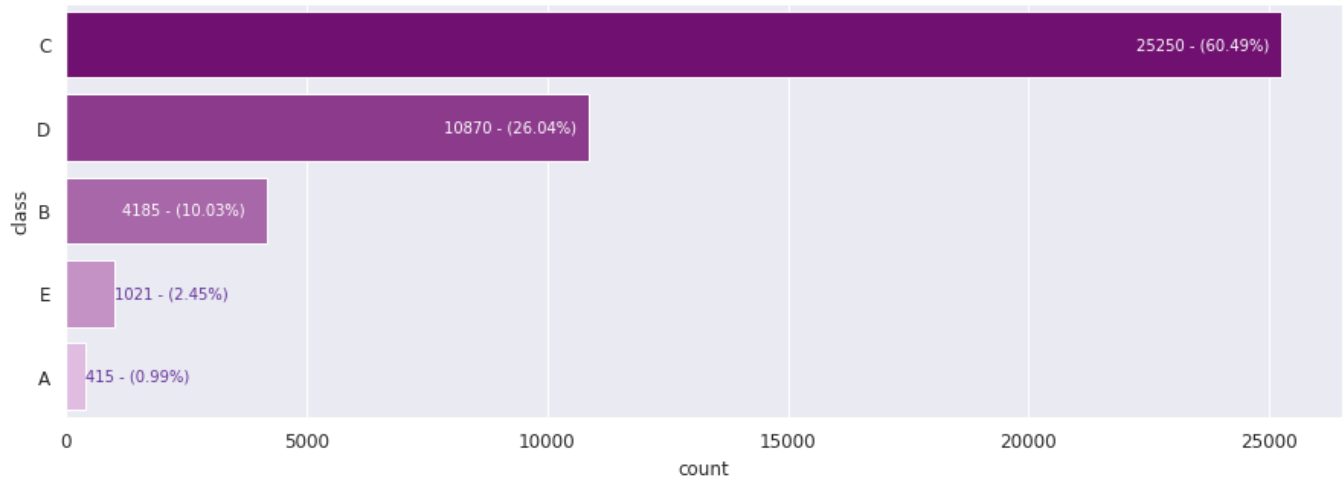




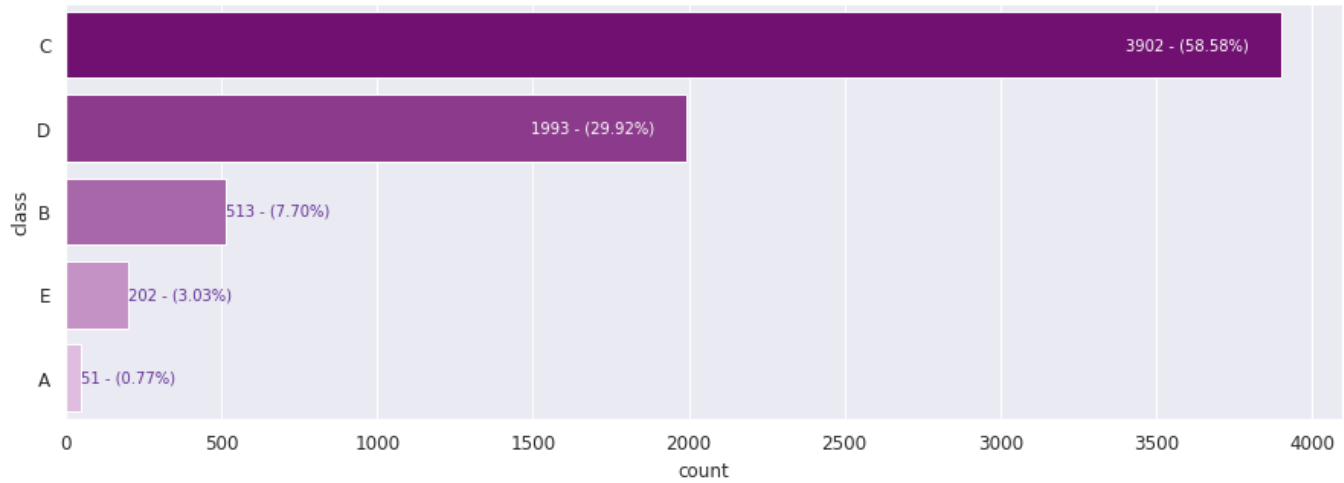
Income

Let's analyse the default against the income

```
AxesSubplot(0.125,0.125;0.775x0.755)
```

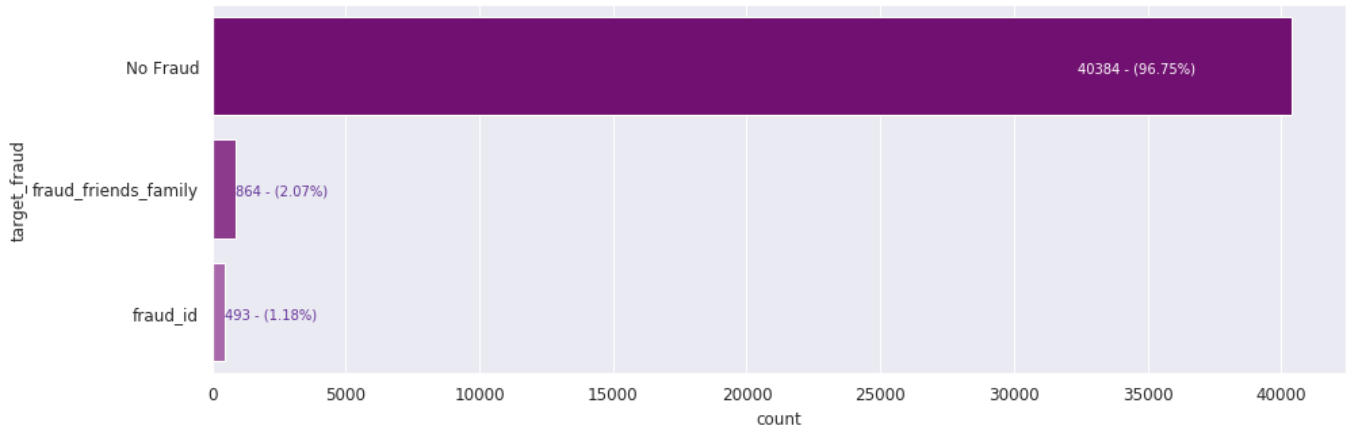



```
AxesSubplot(0.125,0.125;0.775x0.755)
```



Fraud Analysis

```
AxesSubplot(0.125,0.125;0.775x0.755)
```



Fraud loss over the time

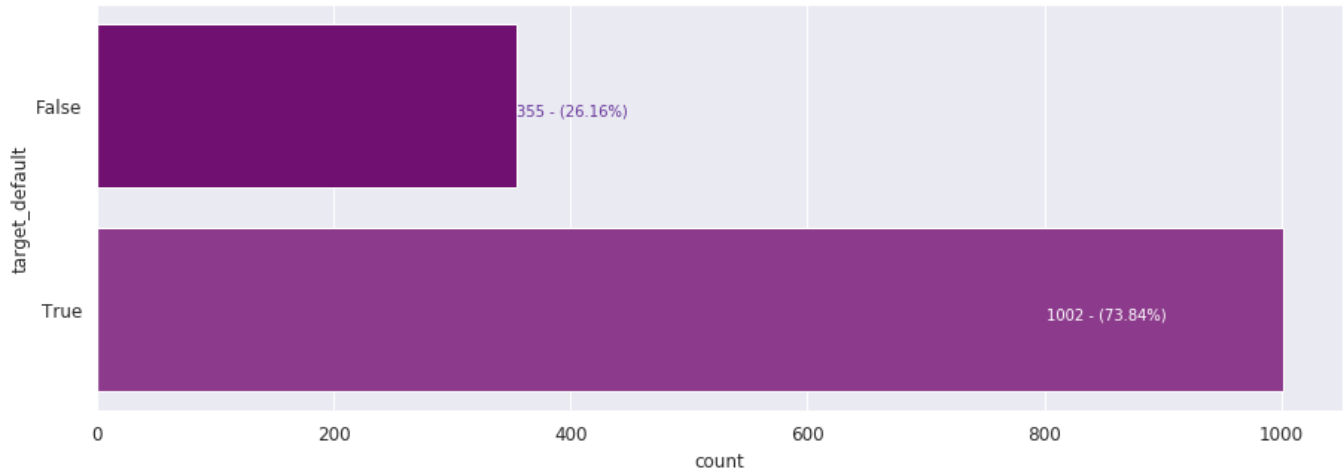
```
(1357, 53)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f202dbe04a8>
```

Fraud vs Default

Let's take a look into only the subset of our data that refers to fraud. How the default is present in this piece of data?

```
AxesSubplot(0.125,0.125;0.775x0.755)
```



Default without fraud

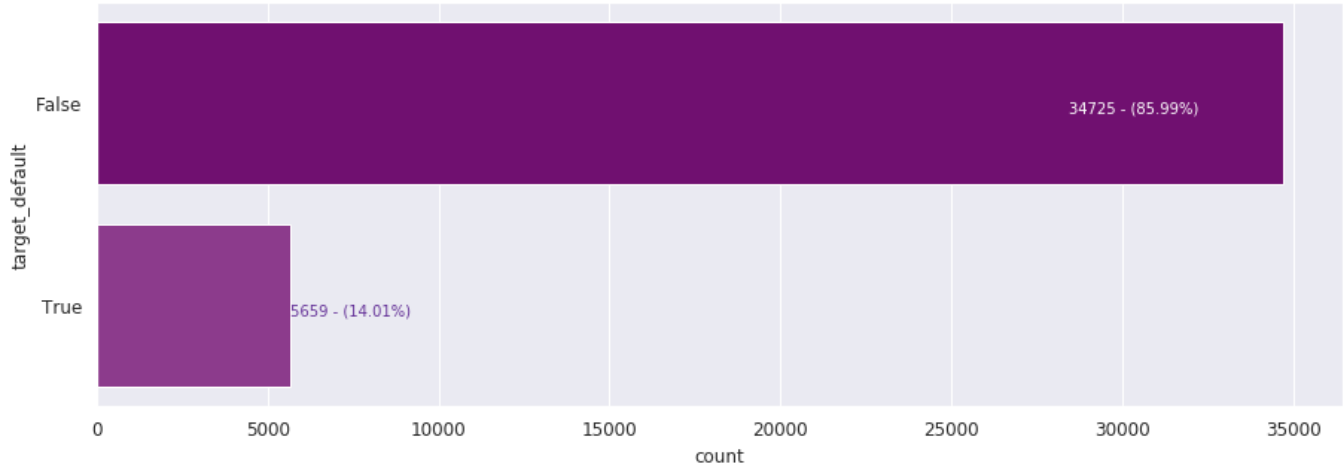
As we can see, almost 74% of our frauds end in default. What is an expected behavior.

We need to take care of removing fraud cases from our analysis when talking about default. Makes no sense trying to predict the probability of default (PD) when dealing with a fraud case, once we already know most of the frauds are made with intention of default.

Ideally, the fraud analysis will run before the PD analysis, so a default caused by a fraud is, first of all, a mistake in fraud analysis.

From this point on, when analyzing PD, we will always look for a view of the dataset without fraud.

```
AxesSubplot(0.125,0.125;0.775x0.755)
```

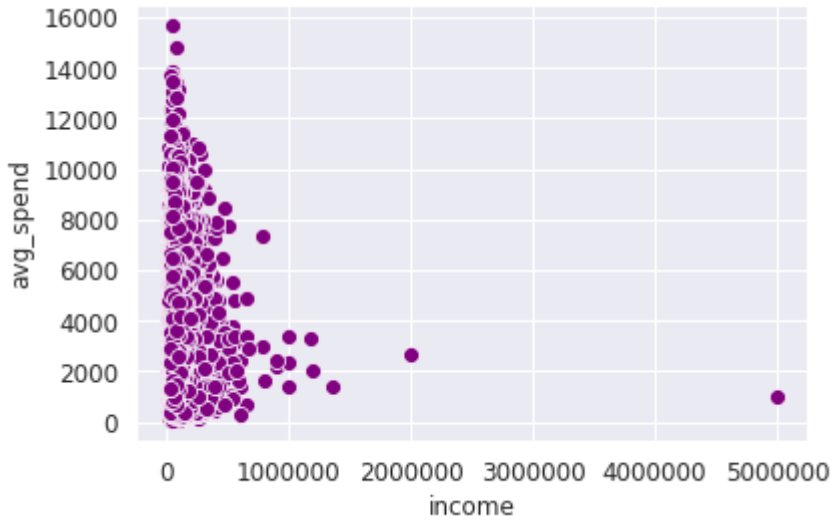


Value Spend vs Income

These graphs will show us how the amount spent is inversely proportional to the person's reported income.

Average spend vs Income

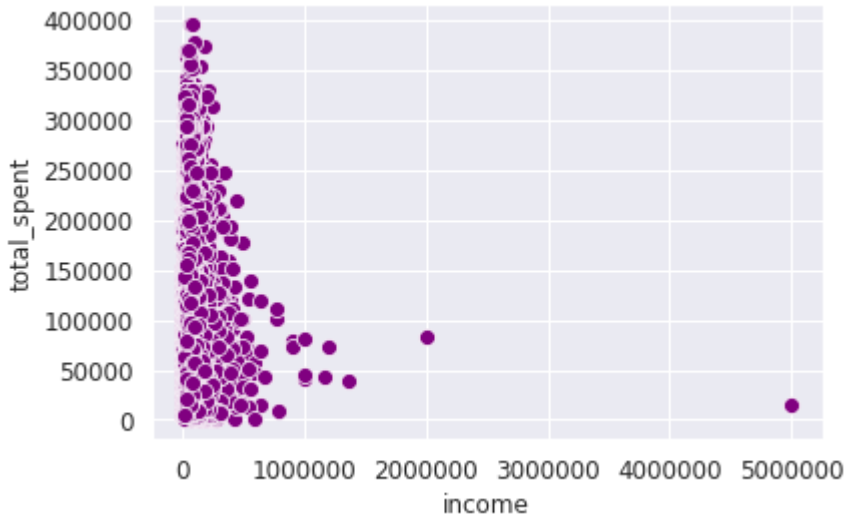
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f202c9f72e8>
```



Total spend vs Income

Here we have to be careful because some people have more or less time as clients.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2034fe2160>
```

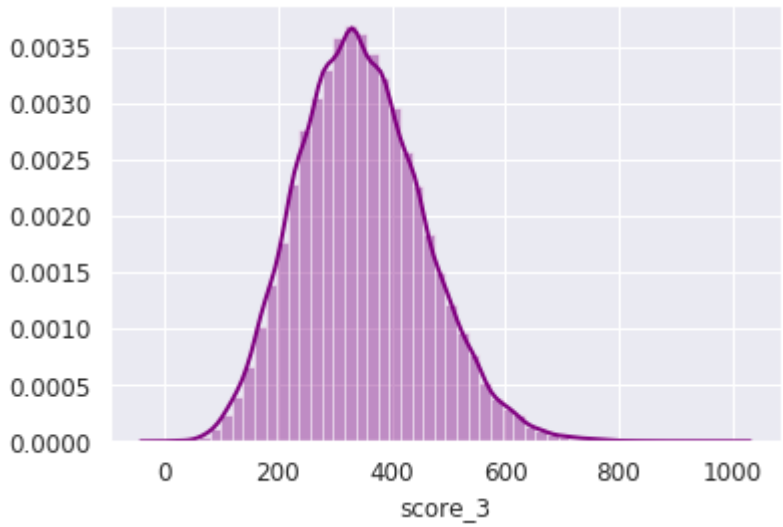


Distribution of scores

Here we can see that the distribution of scores is normalized, as decribed in https://www.nubank.com.br/perguntas (https://www.nubank.com.br/perguntas), section: "Tenho o nome "limpo" e não fui aprovado na análise do Nubank. Por que?".

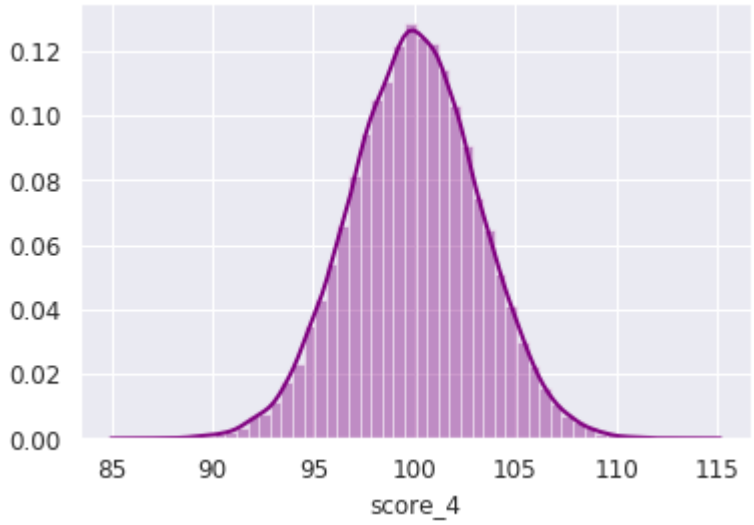
Score 3

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f202bd85ac8>
```



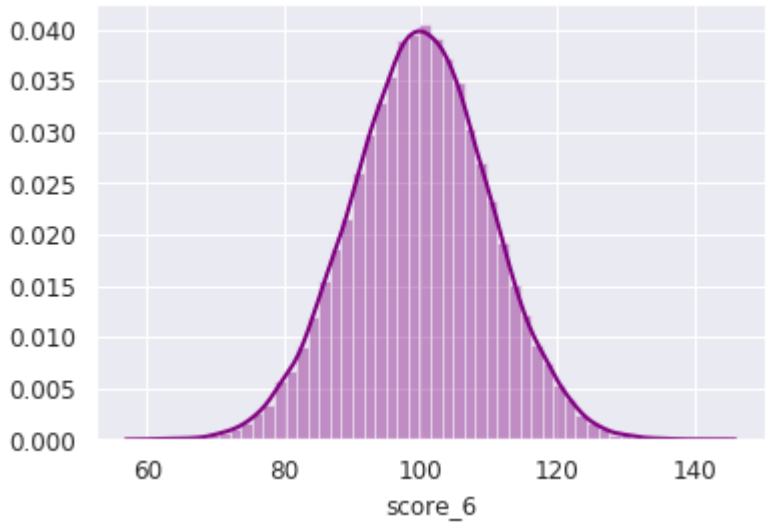
Score 4

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f202d3bc240>
```



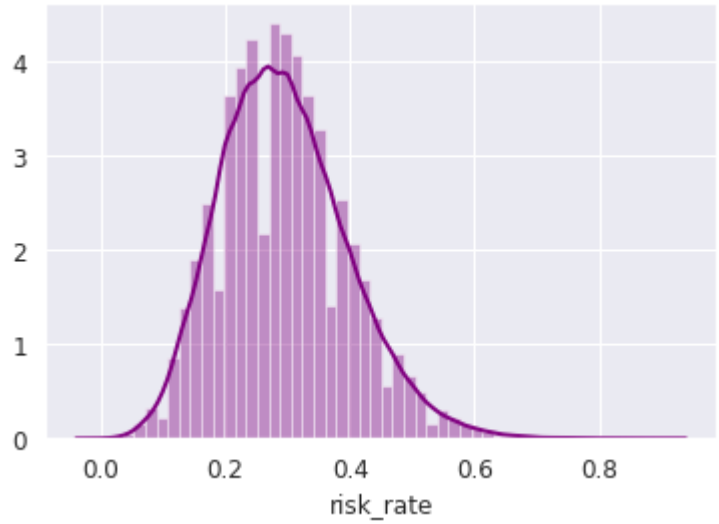
Score 6

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2034fa1780>
```



Risk Rate

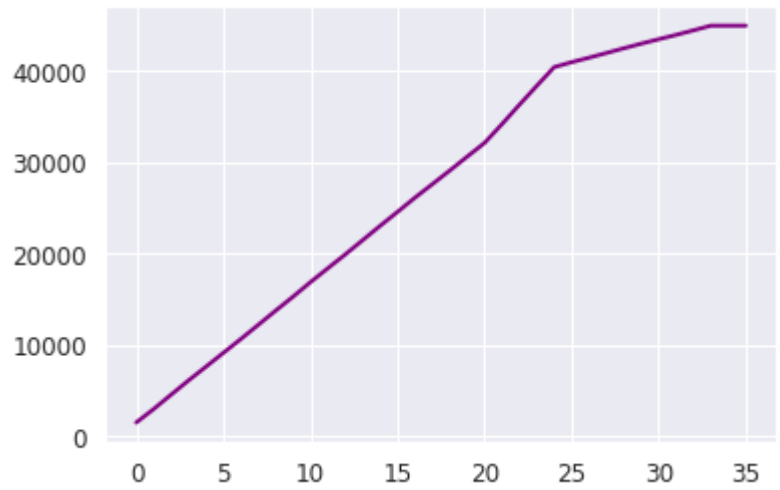
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f202daa8518>
```



Customers

Evolution of customers in time

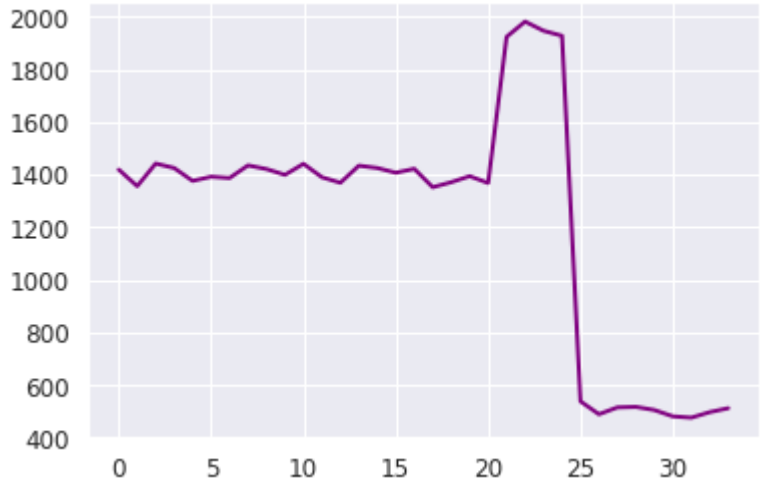
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f20371f8a20>
```



How many customers were approved by month

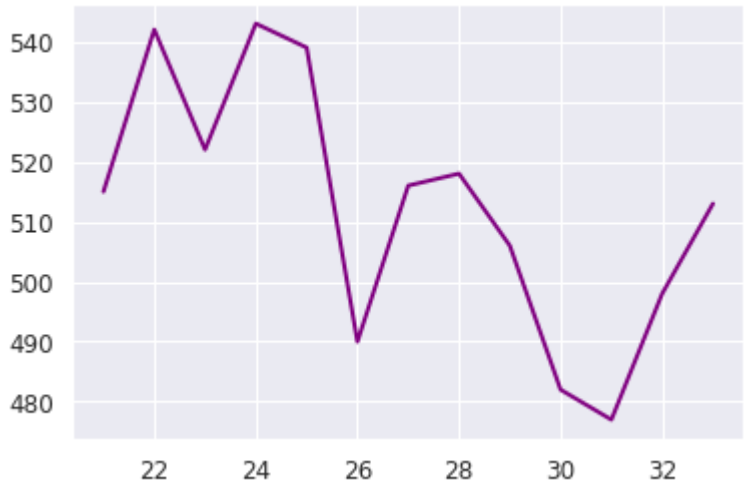
It appears that a mistake was made on month 20th and many customers have been approved, and now to compensate customers are being approved less than normal

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f202da857f0>
```



How many customers approved became defaulters

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f202dc388d0>
```



Reflection

As we can see, it will be a challenging task to predict what has been proposed, because there are few samples of fraud and default in the data set. This reflects the great work that Nubank is doing to prevent this.