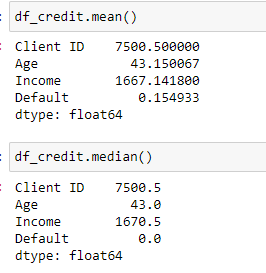
Constructing model:

In order to build a probabilistic model, we ought to identify dependent and independent variables. Since default can be either 0 or 1 and our question is to estimate the risk of clients who are not paying the money back. Therefore, our dependent variable is “Default”, whereas others are independent. To proceed with this assessment, I’m going to use Excel and Python to build the model.

Initially, let us look at how key indexes of the statistics look. To do that, we use Python due to its simplicity. 

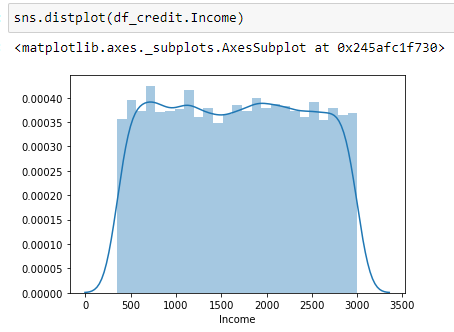
Now we already know the median and mean of all columns and rows.

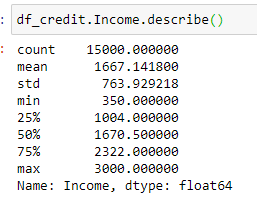
To identify the risks of clients, I’m going to compare different independent variables to our dependent variable – “Default”. I may assume that it is crucial to look at the income column because there is usually a tendency that people with lower income tend to default more rather than customers with income>1000.

|  |  |  |
| --- | --- | --- |
| **average age** | **average income** |  |
| 36.01462995 | 1149.743546 |  |
| **median age** | **median income** |  |
| 33 | 895.5 |  |

There is key provided information above about income and age within customers with default.

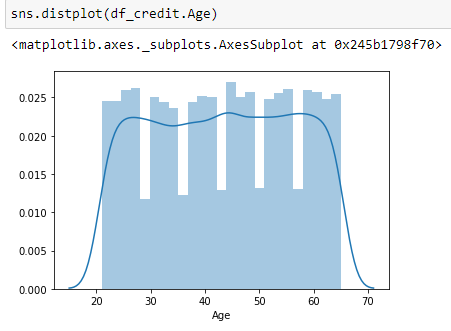
The whole income distribution looks like:





The second row we have to draw attention to is age.

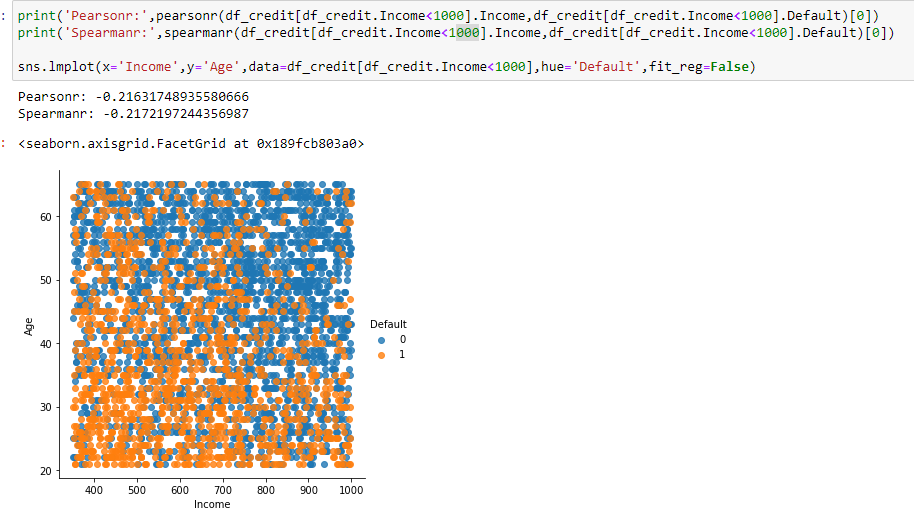
The general distribution appears:



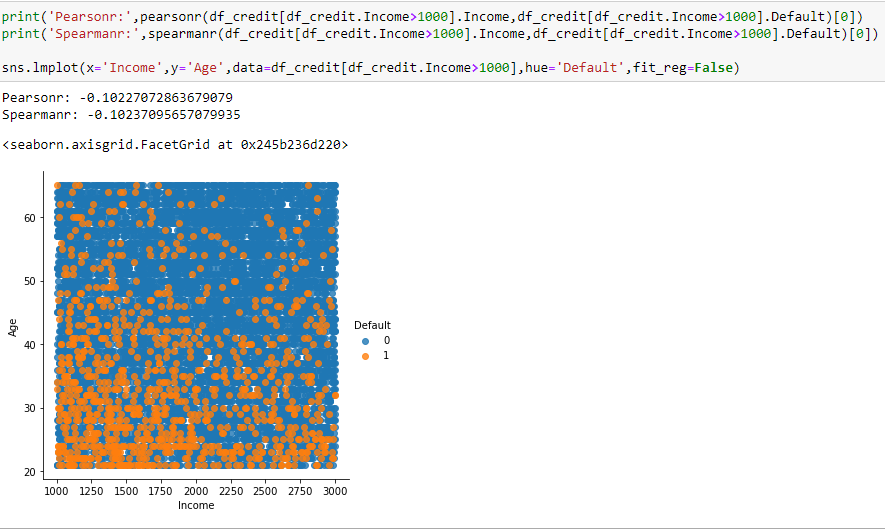
In the end, we also have 2 variables that might brighten our investigation: marital status and client registration date. However, the critical pieces of our analysis are numerical parts. I also may assume that marital status is not important because a general loan may work as a home loan where the income of both partners takes into account. Client registration date might be insufficient as well due to older people can have some other obligations towards another bank.

The model:

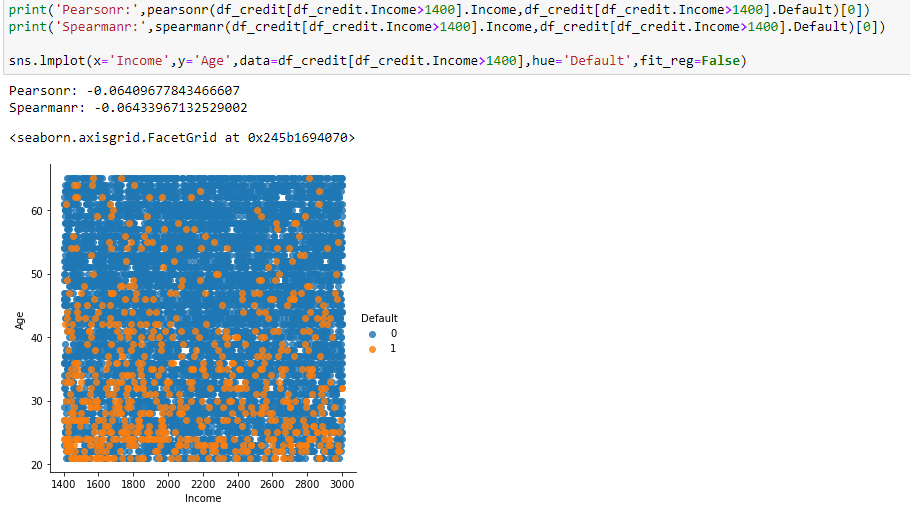
We will use Pearson correlation to estimate the risk of clients. Accordingly, Pearson is relatively easy to understand and evaluate, whereas Pearson’s r is able to examine 2 independent variables (age and income). Initially, we are going to see how age and income are dependent upon default ranges.



Note, that we used the correlation when the income is lower than 1000. Nonetheless, if we change it vice versa, we will get:



If we set our values at income>1400, we have:



Valuation of our analysis

We can conclude that people with income under 1000 have greater number of defaults; yet default often appears with people under around 35 age. The same propensity emerges with people under 35 when their income is greater than 1000 and 1400 (which is approximately median and mean set income). The more person’s income the fewer chances that he can have defaulted.

Indeed, the client risk depends on income and age, hence we have to take it into our consideration in future.

Used tools:

Python’s libraries:

\*pandas - to work with csv file

\*numpy - math library

\*seaborn - graph library to work with numpy

\*matplotlib.pyplot - to plot some parameters in seaborn

\*scipy.stats - statistical tool for Person’s correlation

\*collections - to do counter of some features

\*plotly.tools and plotly.graphs - tools for graphics and

\*warnings - ignore and attention to some issues in code

\*model called df\_credit

Excel:

Filters in data → default 0 and 1/marital status single or married or other

Thank you for your attention,

Sincerely,

Daniil