The data needs to go through a cleaning process, so that the analysis can be done. There could be some strange symbols, empty slots, duplicates, and data that is not supposed to be there.

I didn’t know there was a way to make dummies out of variables with a type of object. I started discretization before, so I had to undo and proceed to do the dummies.

Each variable can be analyzed better when a tool of visualization is used. Not every visualization tool is effective with every type of variable. For continuous is better to do a displot or histogram to show the distribution.

Some of the findings in my analysis were:

* 53% of customers are single and 46% are married.
* The average age is 35
* 60% of customers are female
  + 21% of females default
* 40% of customers are male
  + 24% of males default
* 47% of customers have a university education
  + 35% have a graduate school education
  + 16% have a high school education
  + 2% have other type of education
* 78% of customers don’t default
* 22% of customers default
  + 57% are female
  + 43% are male
* There is a correlation of 0.26 between limit balance and high school education. Therefore, there is a positive linear relationship between the limit balance and the high school education.
* There is a strong positive correlation on the status of past payments in September, August, July, June, and May. The variable names are pay\_0, pay\_2, pay\_3, pay\_4, and pay\_5, respectively. Their correlations are 0.32, 0.26, 0.23, 0.21, and 0.20, respectively.
* There is no correlation between the age and whether the customer will default.
* There is also no correlation between default and marital status, gender, education level, and the rest of the variables. Except for the variables mentioned in the point above.

I strongly recommend paying attention to the status of past payments in the 5 most recent months because those variables are the only ones that show a significant linear positive relationship with defaulting. In this case those months are September, August, July, June, and May. If there are more months of delay, the risk of default is higher.