Use of Pandas Profiler is a nice tool to do an Initial analysis of data. It provides an excellent Univariate analysis of each variable in the data set; however, besides the correlation heat map, it does not provide any information regarding relationships between the variables (bi-or multivariate analysis). Also, the heat map is small and hard to review. I found it more useful to execute a correlation matrix with numbers to get a deeper understanding of correlations.

The data do not have many strong correlations. The strongest is marriage and age (0.41). Beyond this, there are several groups of variables with small correlations (0.2-0.3 range) with Limit Balance having the strongest of these with Billing Amounts, Education and Pay Status. Next was Default with Pay Status, then Pay Status with Billing Amounts and Education. Billing Amounts and Pay Amounts also have relatively strong correlations (0.3), but that is expected.

When expanding Bi-Variate analysis beyond the correlation analysis, start by looking at the Dependent Variables.

Expanding on Multi-Variate analysis through graphs and tables, the following were found:

* Default rate of the entire data population is 22%, significantly high considering that all of these customers were deemed “credible”.
* Married Status: defaults rates highest to lowest are: Divorced (25%), Married (23%), Single (21%), and then Other( 9%)
* Gender: Females make up 60% of the population and have a default rate of 21% while Males have a default rate of 24%.
* Education: default rates highest to lowest are: High School (25%), University (24%), Graduates (19%), Other (7%).
* Age: In General, younger people have lower default rates than older. The trend starts out with very young individuals reaching defaults rates around 20%, then rates decline to about 18% for individuals in their early 30’s, then rates gradually increase and become more volatile to an average of about 25% for people 55 and above.

Regarding Limit Balance vs. Age, Limits increase with age and Max out in the early 30’s, then slowly decrease to the early sixties, then become more volatile and increase again until max age of 79. This is in line with what we would expect from what we learned with Age default rates, except the increase in Limit Balance after the early 60’s.

Next steps

1. Conduct Machine Learning Analysis using categorical regression as Defaults being the dependent variable. Primary independent variables being Pay Status, Limit Balance, Billing Amounts and then Education.
2. Conduct Machine Learning Analysis using linear regression as Limit Balance being the dependent variable. Primary independent variables being Billing Amounts, Pay Status, Education, and Pay Amounts. Will also explore how Defaults affects the model.