**Lessons Learned Report for Class 5 – Task 2**

The EDA exercise was a good practice to become familiar with the python syntax. It is a bit difficult to move to python after months of R work. I found data visualization easier in R and Tableau so I don’t think I’d rely on python for visualization. I also did not like Jupyter Notebook as much as I liked RStudio.

Regarding the dataset, I created a new subset of data to explore the behavior of customers who defaulted their loans. I found a statistically significant relationship between age, sex, marital status and education with defaulting loans. For example:

* Early to mid-career customers are more likely to default
* Customers with higher education levels (Education levels 4 to 6) can be considered low risk.
* Education level 2 (university level) is high risk in both genders.
* Singles are more likely to default their loans. Divorced customers are less likely to default.
* Overall, the age range of divorced customers and those with higher education levels is higher.
* Number of female defaulters are higher than males.

Based on this study, I found that solely relying on age, education and marital status can be misleading, as in general, higher education achievements and divorce usually occur in older ages. I think it would be a good start to combine these into a new variable and build a detailed classification model that predicts the exact demographics of customers who default their loans.