# Credit One EDA – Lessons Learned

**EDA Overview**

The columns of the dataset are identified, and a summary of their meaning was provided. Therefore, both Pandas profiling and manual data wrangling, statistical analysis and visualization methods were used during EDA. Pandas profiling was used to understand the data from a high level, but since the features and target variable meanings are known, wrangling the data and exploring the data from a human perspective of the data provided better insights then Pandas profiling alone. So, filtering and sub-setting the data in ways that related to the goal were a key step of the manual EDA. For example, the data was broken down in those clients that defaulted, and the demographics and credit limits were explored and compared to those clients that did not default. Finally, a major part of this EDA was binning the customers’ ages in to generations and binning limit balances into four categories, something that Pandas profiling did not have the power to do.

**Initial EDA Insights**

The limit balances were grouped into four bins called Bronze, Silver, Gold and Diamond representing the lowest limit balance quartile to the highest limit balance quartile respectively. Although analytics have not been performed yet that demonstrate what tends to drive defaults, the statistics show that default rates are highest in the Bronze credit limit range and default rates decrease as credit limit balances increase. Additionally, grouping the data by different demographics does not change this stat, regardless of the sex or education level of the customer default rate trends are the same in each credit limit balance range.

**Recommendations**

The recommendation is to build regression models that focus on the limit balance as the target continuous variable and explain how likely a customer is likely to default based on the credit they are extended.

**­­**