## Lessons Learned Prepare and Explore the Data, Assn 5 Task 2

## Celeste Hofer

**Recommendations Based on Findings:**

The exploratory data analysis elucidated several interesting points on the data. It is recommended that predictive modeling be used to determine which customers are most likely to default. The current default rate is around 22%, we want to significantly reduce this in the future. The number of customers who defaulted was: 6636 and the number of customers who did not was: 23364. The data we have is of high enough quality that we should be able to build predictive models that have high accuracy. The models will be used to predict the likelihood of default.

**Information Found Providing Business Value:**

* **Customer education:** Frequency ranked in descending order is University, High School, Graduate School, and then Other. Other is not described and has only a few entries.
* **Age:** The most frequent age range of customers is mid-twenties to mid-forties. The average and mean age of customers are 35 and 34 years old, respectively.
* **Balance Limit:** The balance limit ranges from $10,000 to $1,000,000. The most frequent limits are less than $300,000. Customers with Graduate School education receive the highest balance limit; those with a High School education have the lowest balance limit.
* **Loan Default:**
  + Examining gender, 3763 females defaulted, as opposed to 3873 males. For both sexes, those with University educations, then High School, then Graduate School were most likely to default. This is alignment with the total customer population totals.
  + When marital status is considered, for both genders the marital status most likely to default is divorced status, followed by single status, other status, with married status being the least frequent.
  + The majority of customers who default are age 30 and over.
  + The educational level of those who default, in descending order is University, then Graduate School, then High School.
* **Other:** Analysis was not used to examine past payments nor payment history.
  + The data had no missing values. The data also had no quasi-constant features (near zero variance). No duplicate rows of data were found.
  + Correlated features were found and removed. This was done with both the original data frame and data frame with features removed.

**The main lessons I've learned from this experience were:**

Data cleaning should have been done earlier in this exercise. Next time an analysis is conducted I will change the sequence. Adding in the replica of the Titanic analysis approach was not seamless in light of the other analysis that had been done. My solution was to create a second data frame that had all of the extra features added.

Graphing and charting using python packages was extremely helpful in order to provide a visual inspection of the data analysis. Using the seaborn plots was slightly challenging as catplot sometimes had to be used to replace factorplot. At times factor plot was still usable.