Simple Linear Regression Assignment

### Author: Nitesh Baranwal Date: 12th July, 2017

# Predicting Credit Amount using Simple Linear Regression

## Purpose:

The purpose of this analysis is to review provided data set and identify correlation between various metrics. We would also like to analyze a few key categorical factors. In particular, we want to understand the relation between Amount and Duration.

## Dataset:

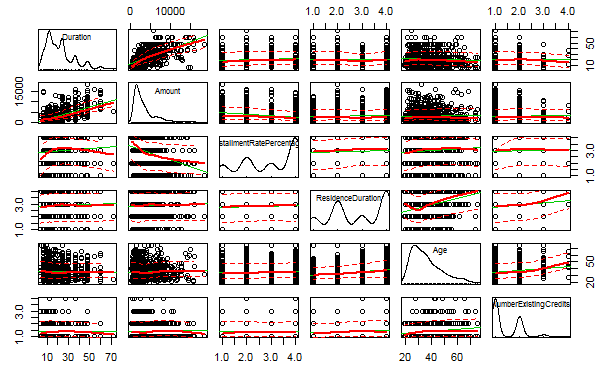
German Credit dataset

## Dataset Description:

# of records: 1000

# of columns: 62

1. There are 6 numeric columns with more than 2 values which can be considered for regression.
2. Scatter plot for these shows that there is positive correlation between Amount and Duration. On further inspection we can see that **correlation coefficient between Amount and Duration is 0.624 which reflects strong positive correlation**.
3. There is weak negative correlation between Installment rate percentage and Amount
4. There is weak positive correlation between Age and Residence duration
5. Class can be considered a categorical factor with 30% Bad and 70% Good values spread
6. Most of the population, about 94%, has some form of employment
7. Most of the population, more than 96%, is foreign



Scatter plot matrix for metrics in German credit dataset

## Final Model:

**Regression model to predict amount using duration: Amount = 213.216 + 146.3 x Duration**

Intercept is not significant as p-value is 0.127 but correlation coefficient is significant based on it value p-value ≤ 0.05.

39% of variation in Amount is explained by Duration (i.e. r-squared = 0.39)

**Correlation coefficient between Amount and Duration is 0.624 which reflects strong positive correlation**

Overall, the model is significant based on p-value <=0.05

## Conclusion:

There is strong positive correlation between Amount and Duration and generated model is significant to predict amount using duration.