# 6.2 Project Milestone

Ali Malenchik

So far in this project I have completed data import and cleaning, as well as exploratory data analysis. After importing the data I found that the data types such as region code and policy sales channel were interpreted as float, so I converted them to string since they are truly categorical variables. I also updated binary values (1 and 0) to Yes and No in the driving license and previously insured variables, in order to make them easier to work with during EDA. I also checked for null values and found there were none. Once all cleaning was complete I split the dataset into 67% train and 33% test.

In exploratory data analysis, I created count plots for the categorical variables and incorporated the policyholder response in order to compare the response proportion for different values. I also created histograms, box plots, and step plots for the numeric values in order to observe the distribution of variables such as age, annual premium, and vintage (age of the car). I also created scatterplots to demonstrate any relationships between the numeric variables (age vs. annual premium, age vs. vintage, annual premium vs. vintage). One interesting observation I had is that despite policyholders being skewed toward the 20's age, very few of these younger policyholders were responsive to cross-selling. Additionally, policyholders with relatively new cars (<1 year in age) were not very responsive to cross-selling, with the lowest proportion of policyholders interested in vehicle performance. Policyholder with older cars (>2 years old) were almost 50% responsive to cross-selling.

After exploratory analysis, I encoded categorical variables in preparation of modeling. Using the encoded variables I created a correlation heatmap. One issue I have run into already is the number of encoded variables, especially due to the region code. During feature selection I will determine if these region codes are valuable or necessary to the model.