NYC Data Science Academy

George Goginashvili

Project 2

I scraped two websites of sword manufacturers, Cold Steel and Valiant Armoury. Main goal in my analysis was to understand how the manufacturers position their product / swords regarding price, and if there are specifications / variables that have impact/ statistical relationship with price. In order to achieve this, I extracted the following specifications: blade length, overall length, handle length, price, weigh, name of swords and type of steel. I have used variable name to create another variable katana since I have believed that Japanese swords had statistical relationship with price (I wanted to check this assumption).

In my analysis, I have used box plots for categorical variables such as steel type and Japanese swords, scatter plots with linear approximation for numeric variables, and regression analysis along with correlation matrix to check independent variables on interaction. Box plot for steel shows that type of steel has impact on sword prices. Swords made of steel series 1095, 1084, 1070, 1060, 1050 are priced less than swords made of Damascus steel or 6150 steel. However, these types of steel are hard enough to manufacture "battle ready" swords. Box plot for katana also shows that Japanese swords are priced higher than other swords. Even though box for non-Japanese swords includes more swords and big part of them have higher price than katanas, I believe this happens because Valiant Armoury has higher prices in general and specializes in non-Japanses swords, but all other things held equal because of popular culture Japanese swords are priced more and have stable demand on market. Scatter plots for numeric variables show that blade length, overall length and weight have positive relationship with price. I have created additional variable blade length over overall length, blade_ovarall, to find out if there is a relationship between buyers' desire to buy one handed sword or two handed sword. Even though the relationship turned out to be positive, it has been weaker than that of other numeric variables, which explained negative relationship with handle length and price. People preferred swords with shorter handle over swords with longer handles. In order to further corroborate my findings, I have used two regression analyses, one with only numeric independent variables and another with numeric and categorical independent variables. Regression analysis with numeric independent variable has shown that only relationship between price and blade length has statistical relationship (statistical significance). Regression analysis with numeric and categorical variables has shown that price has only statistical relationship with blade length and categorical variables steel type and katana.

In order to improve analysis, I believe data should be increased ten times and at least twenty more webpages of sword manufacturers must be scraped. I also believe that some of assumptions for regression analysis such as homoscedasticity, outliers, randomness and normality must be further checked. From scatter plots we see that linear models are heteroscedastic. On the other hand, correlation matrices show that independent variables that have statistical relationship

(statistical significance) with price have acceptable level of correlation to exclude interaction among them (however, I would double check this assumption).

In spite of shortcomings in analysis, I may assume that it is commonsensical that steel quality and blade length have a significant impact on prices of swords. I may also assume that because of popular culture Japanese swords maintain a certain niche in swords' market.