Victoria Hall

DSC530 – T301

Final Project

Can We Predict the Price of Shoes?

The purpose of this research was to see if we could accurately predict the price of shoes based on several explanatory variables. My hypothesis was that color and brand would help explain and predict the price of shoes. I used a dataset of 10,000 woman’s shoe prices scraped from multiple websites.

My EDA showed a distribution of prices that appeared relatively normal. I did some work to limit outliers like only including colors with more than 50 instances in the dataset and brands with more than 10 instances. This significantly shrank the size of my dataset and allowed for a clearer review. It did appear that certain brands like Lifestride were overrepresented as was the color black.

Because of the relatively normal distribution I was able to utilize testing based on the normal distribution. However, I did eliminate outliers and something I was unable to figure out how to account for is that there were some shoes that had multiple cases. For example, in the dataset the same shoe might have shown up with different prices recorded at multiple instances so there was duplication of average prices. Both of these things could mean I may have misrepresented the distributions of the total population compared to the sample.

Overall, I was able to create an OLS regression model that accounted for around 50% of the variation in pricing. This means that I wasn’t able to explain the other half of the variation. The dataset I used did offer more features for selection that I could have built in but the cleaning I would need to do extract those features was beyond my skill set at this time. They were contained in long strings and would require pulling bits out to change into categorical data. My data was also limited and overrepresented by certain colors and brands. Finally, I struggled a bit with using the combination of categorical and numeric data. I did my best to find ways to perform the analysis but the examples in the textbook mostly focused on numeric data which may mean some of the assumptions I made and tests I performed were not perfect.