Tom Ferko

DSC 530 week 10 Final

The outcome of my EDA of the realtor dataset was very successful overall. I was primarily looking at the measurable variables that had an effect on price, so I ended up looking at number of bedrooms, number of bathrooms, lot size, square feet, and most importantly price. One major concern that I ran into was outliers. The dataset was massive, around 500k observations. And even with this many observations, there were still major missing pieces of the data. I orginally tried to compare east to west only to realize that much of the west data was missing, so I had to pivot and change to house vs non-house which still turned out to be very interesting. I faced a major issue trying to get the analytical model to line up with the data, and couldn’t end up getting it to work in the end, and that could be on me for trying to make a model fit that wasn’t designed to. I felt that the prices of homes fell under a pareto distribution with a few very expensive houses skewing data towards larger numbers. But when I put the model in place and tried to plot it, it just didn’t work. I think it will end up just taking more time to get right, and fully understand it.