**Term Project**

**Logan Quandt**

**3/1/2022**

I conducted Exploratory Data Analysis on a dataset on housing from California due to my interest in the housing market based on my current housing search. My hypothesis was that house size (bedrooms, square feet of living area and square feet of lot size) are the primary drivers behind housing prices.

After conducting my analysis, I found that bedrooms and square feet of living area were positively correlated with sales price (0.22 and 0.45 respectively) they did not indicate a strong correlation. The Spearman Correlation’s for square feet of living area (0.70) indicated that there may be a non-linear relationship there that it is impacted by outliers.

There are multiple other variables I could’ve used in my analysis that could’ve helped explain some of the pricing differences such as if the house was renovated or not, but I believe that there could have been additional variables such as dollar amounts of repairs needed to reflect if house prices are lowered to account for repairs or sold in pristine condition.

An assumption that I did not agree with was the lack of different geographic areas in the dataset. Most of the data points existed within two zip codes which appeared to be very similar in pricing.

Issues I faced were in my limited experience in coding and Ipython. For issues with coding, I was using packages without fully understanding the scope of functionality within them. For most of the course, I have been using the Think stats package but to solve some graphing errors I ended up using matplotlib which I have limited experience with, and I would’ve like to represent some of my graphs differently. Other issues I had was with the hypothesis testing and knowing exactly what regression model to use.