My goal for this project was to create a model to predict timeshare resale prices. Though I do not own or plan to buy a timeshare, I was particularly interested in this topic since I commonly here complaints about people trying to get rid of their timeshares. From what I've heard and read online, it seems that the general consensus is that purchasing a timeshare is a giant ripoff unless you intend to visit the same vacation spot and resort every year.

Using Selenium and BeautifulSoup, I scraped data from My Resort Network and TripAdvisor using Google Search results. I removed resorts that did not have TripAdvisor ratings from my dataframe. One thing I found particularly difficult was cleaning up the text scraped from Google.

Once I finished cleaning up the data, I observed the correlations between variables. For baselining, I compared Price with the Bedrooms column. The R^2 was at 0.252 but the p-values were 0. Starting with all of the features, I weeded out features that would lessen my R^2. I noticed collinearity between the PriceMin and PriceMax columns, which made sense because these variables represented the quality and price ranges of the resorts. I kept PriceMin as a feature in my model.

I created a linear regression model using MaxOcc (Maximum Occupancy per Unit), Rating (TripAdvisor Rating), Reviews (Number of TripAdvisor Reviews), PriceMin (Minimum Price for a Standard Room), Bdrs (Number of Bedrooms), and 16 of the dummy variables. This combination of features resulted in the highest R^2 and the p-values were low except for a few of the dummy variables.

Using the same combination of features, I created a polynomial regression model with a degree of 2. Using cross validation, my mean R^2 was 0.5936. After using Lasso in an effort to regularize, I found that my model was better off without regularization. Though polynomial regression may be harder to interpret, I found that this method would better predict timeshare prices than using linear regression.