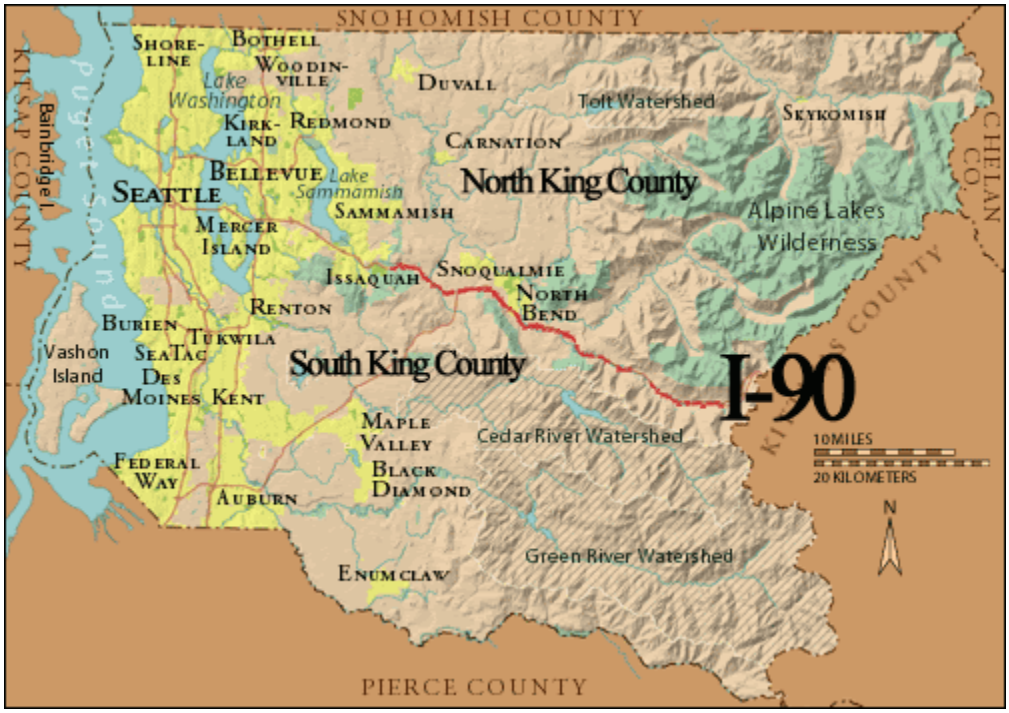
King County Home Prices

Regression Analysis



Main Objective of Analysis

Using various regression models, the goal of this analysis was to forecast prices of homes of King County in Washington state based on various attributes of the homes which included factors such as home size, location, and grading. The primary focus of the analysis was prediction, to estimate home prices after running several regression algorithms and after comparing different statistical error scores, determine which was the best model. The results of this analysis are also useful for interpretation and there is potential for assessments on which factors of a home determine prices.

Description of Data Set

King County is the most populous center in the state of Washington with a population of 1,931,249 people and 789,232 households according to the 2010 US Census. The data set is a sample of 21,597 homes in King County with the following 21 attributes:

- Identification notation for a house (int64)
- Date house was sold (object)

- Price of house (float64)
- Number of bedrooms (int64)
- Number of bathrooms (float64)
- Square footage of house (int64)
- Square footage of lot (int64)
- Total floors of house (float64)
- If house has view to a waterfront (int64)
- If a house has been viewed (int64)
- Condition of house (int64)
- Grade of house (int64)
- Square footage of house apart from basement (int64)
- Square footage of the basement (int64)
- Year built(int64)
- Year renovated (int64)
- Zip code (int64)
- Latitude (float64)
- Longitude (float64)
- Living area in 2015 (int64)
- Lot area in 2015 (int64)

Exploratory Data Analysis, Data Cleaning, and Feature Engineering

After performing summary statistics of the original data, the median price for homes was $450,000 while the maximum price was $7.7 million. To examine the spread of the data for prices, a density curve was plotted which indicated that the distribution had low variance and was highly skewed to the right. The most expensive homes in King County are relatively few compared to homes that are more typically priced. To remove outliers, data with z-scores greater than 3 were eliminated and 20,098 observations remained. Plotting the density curve again, the spread of the data was wider and the right skew remained, but wasn't extreme enough to warrant a logarithmic or boxcox transformation and these transformations were not performed for this analysis. The dataset contained no null values. The following attributes were removed: 1. Identification Number, 2. Date of Purchase, 3. Number of views, 4. Year Renovated, 5. Latitude, 6. Longitude. These factors were considered to be relatively unimportant in this analysis. With geographic information such as Latitude and Longitude, negative float values could complicate regression so rather than utilizing those factors, Zip Code was used instead, as location can play a significant role in home prices.

Regression Models

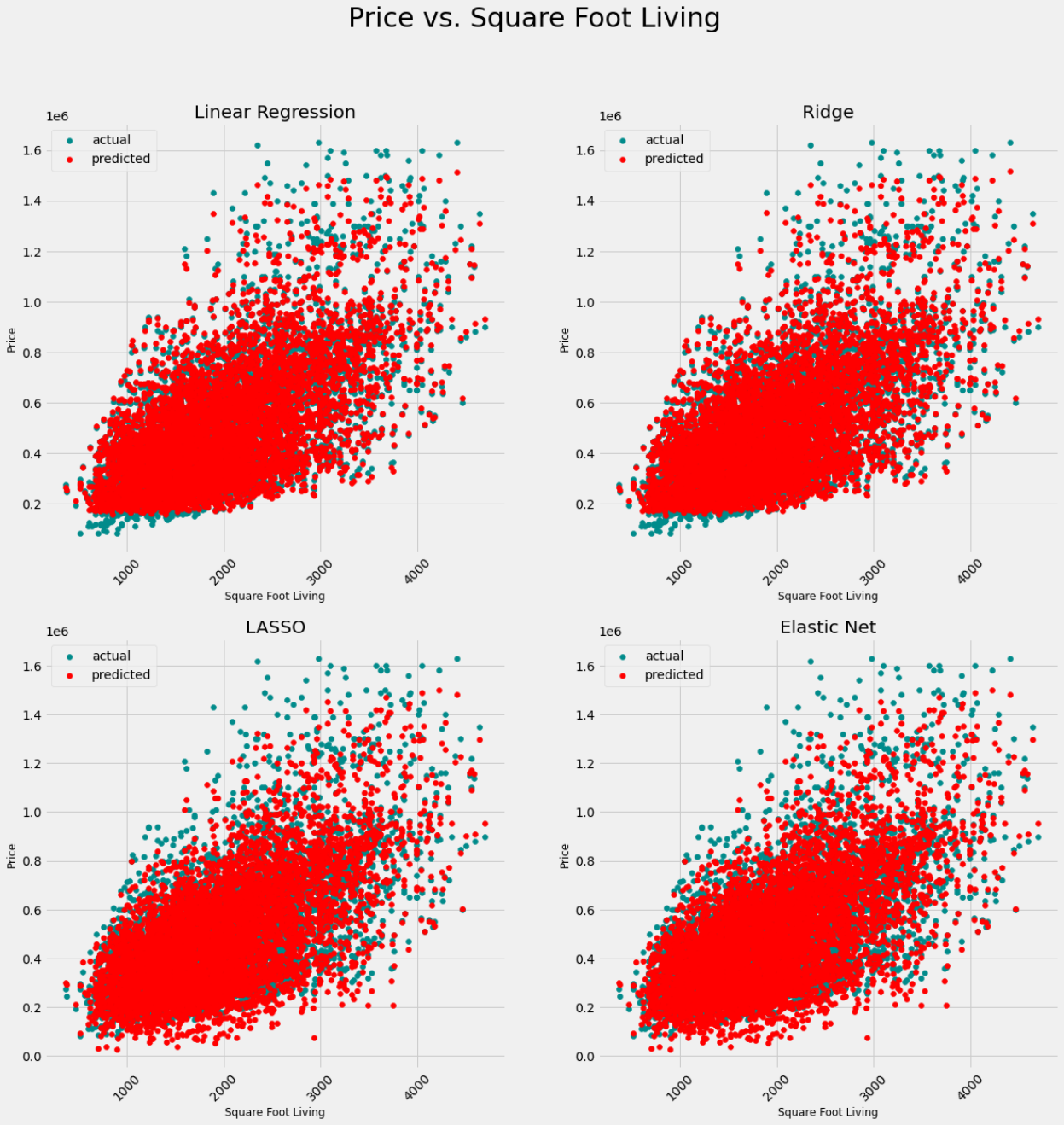
For the purposes of this analysis, four regression models were used, standard linear regression, Ridge, LASSO, and Elastic Net. Three cross validation methods were employed,

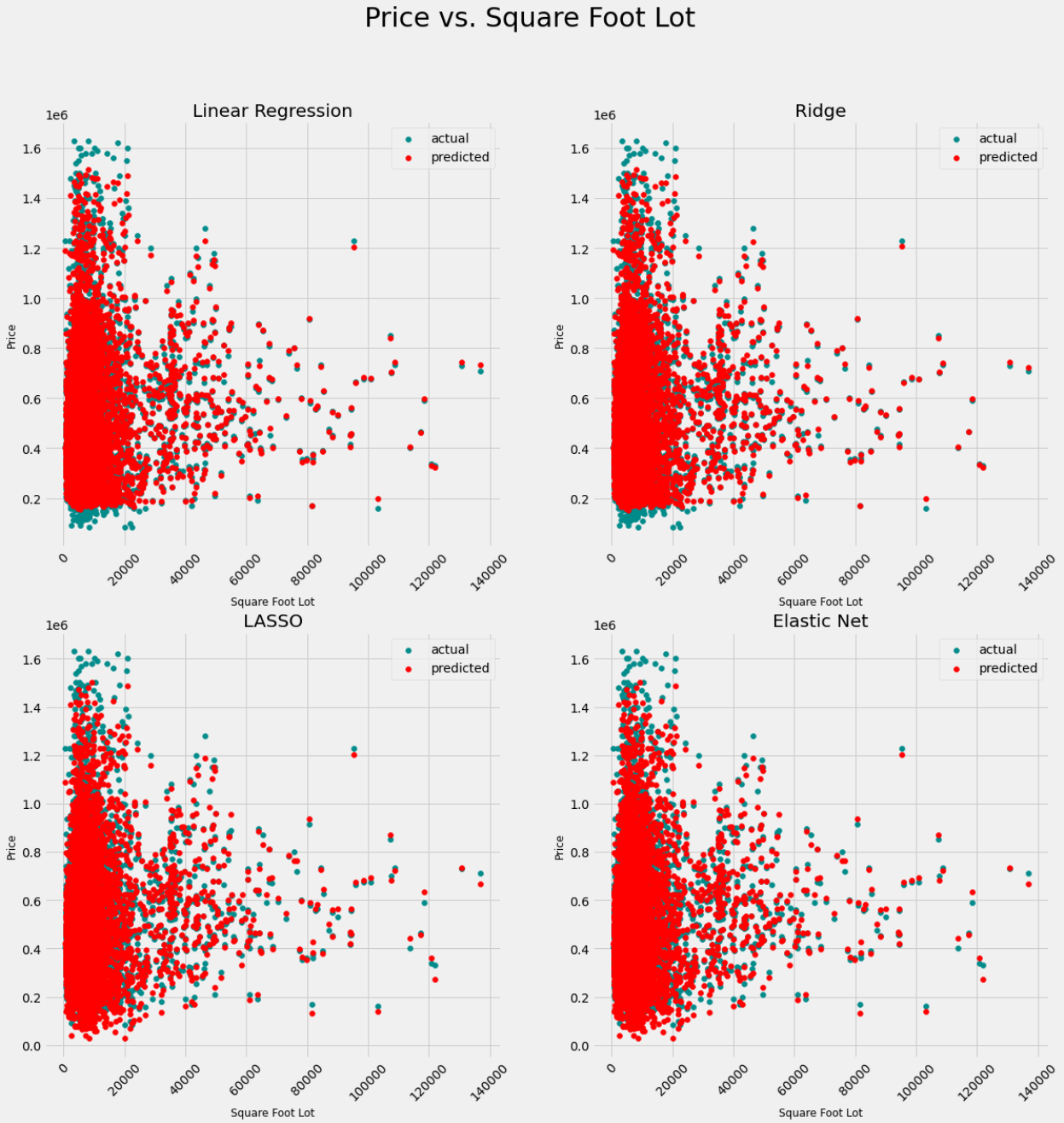
including a 70/30 train/test split, K Fold and Stratified K Fold and all three were used in each model. These were compared to determine the best R^2 and root mean square error scores for the 70/30 train test split and average scores over 10 folds for K Fold and Stratified K Fold to assess which model performed the best. 2nd degree polynomial features were added, and the data was scaled using `StandardScaler`. After running `GridSearchCV`, the optimal hyperparameters were: $\lambda=0.001$ for Ridge, $\lambda=10^{-11}$ for LASSO, $\lambda=10^{-11}$ and $\beta=0.9$ for Elastic Net. Running the models using these hyperparameters and various cross validation approaches resulted in the following scores:

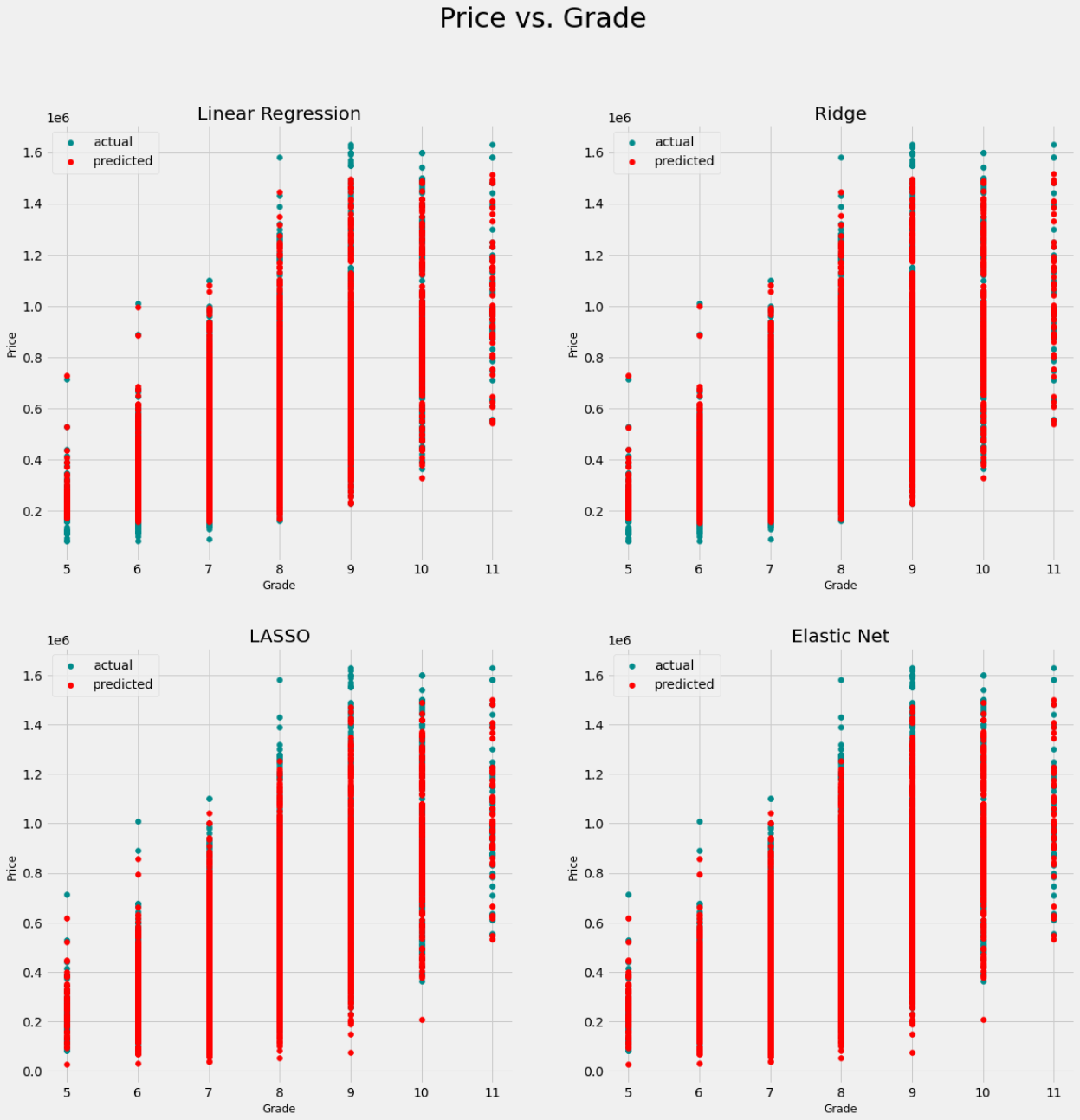
	model	r2	rmse	kf	stratified kf
0	Linear Regression	0.995572	15960.768028	0.995441	0.995422
1	Ridge	0.995571	15961.331626	0.995441	0.995412
2	LASSO	0.980083	33546.850900	0.980342	0.980292
3	Elastic Net	0.980083	33546.850900	0.980342	0.980356

All of the models performed very well, with high R^2 scores, close to 1. Running standard linear regression and Ridge resulted in the best scores, with standard linear regression only slightly better than Ridge. This indicates that all of the attributes in this data set bear important weight and penalizing parameters does not lead to better results. The models that zero some of the parameters such as LASSO and Elastic Net had much higher root mean square error scores. So standard linear regression stands out as the best model in this analysis. And out of the cross validation approaches, a 70/30 split proves to be the best method in this case.

The following graphs plot predictions and true data together for Price versus Square Foot Living, Square Foot Lot, and Grade. Note that while standard linear regression was chosen as the best model, all four performed well and the plots look very similar, with a few minor variations with predictions closely matching true values.




Price vs. Square Foot Lot
Linear Regression
Ridge
LASSO
Elastic Net
actual
predicted
Price
Square Foot Lot
1e6
0
20000
40000
60000
80000
100000
120000
140000


Price vs. Grade
Linear Regression
Ridge
LASSO
Elastic Net
actual
predicted
1e6
Price
Grade

Key Findings and Insights

All of the values in the cleaned dataset were numerical, either integers or floats. Categorical data was completely absent which facilitated good regression modeling. Error scoring was very good for all four models. Because the standard linear regression model performed the best, at least the majority of the parameters were important in carrying out predictions. This present analysis is predictive rather than interpretive, so the focus is on how accurate the forecasts were. Standard linear regression performed the best, with Ridge a close second and LASSO and Elastic Net, which penalize the coefficients the most performed less well. But the R^2 scores for the four models were high and they would all do well at estimating home prices.

Next Steps

Keeping the longitude and latitude attributes, a geographic heat map can be created to visualize what parts of King County have higher or lower home prices. A hypothesis can be made that urban areas have more expensive homes than rural areas. This kind of information can facilitate a more interpretive approach to analysis. Using regression to predict house prices in this dataset was fairly straightforward. A potential future analysis would involve classification algorithms to determine the grade of a house based on various attributes in the data set, as the grades are integers in a small fixed range and can be converted into categories.