King County Real Estate Price Focast

•••

November 2020

Andrew Banner and Trevor Mott

Defining The Problem

- Create a accurate pricing model for new listings within King County
    - Use the model to gain additional listings by accurately predicting the sale sale price
    - Use model for negotiations between seller and buyer
    - Provide great visuals to allow the home buyer/seller to gain a greater understanding of market
- Target audience
    - Buyers and sellers looking within the average price range for the market.

Understanding the Raw Data

- Housing data set for King County, Washington
- 21613 Total Houses
- All houses sold in 2014 and 2015
- 70 Different Zip Codes within data
  - Categorical Variable

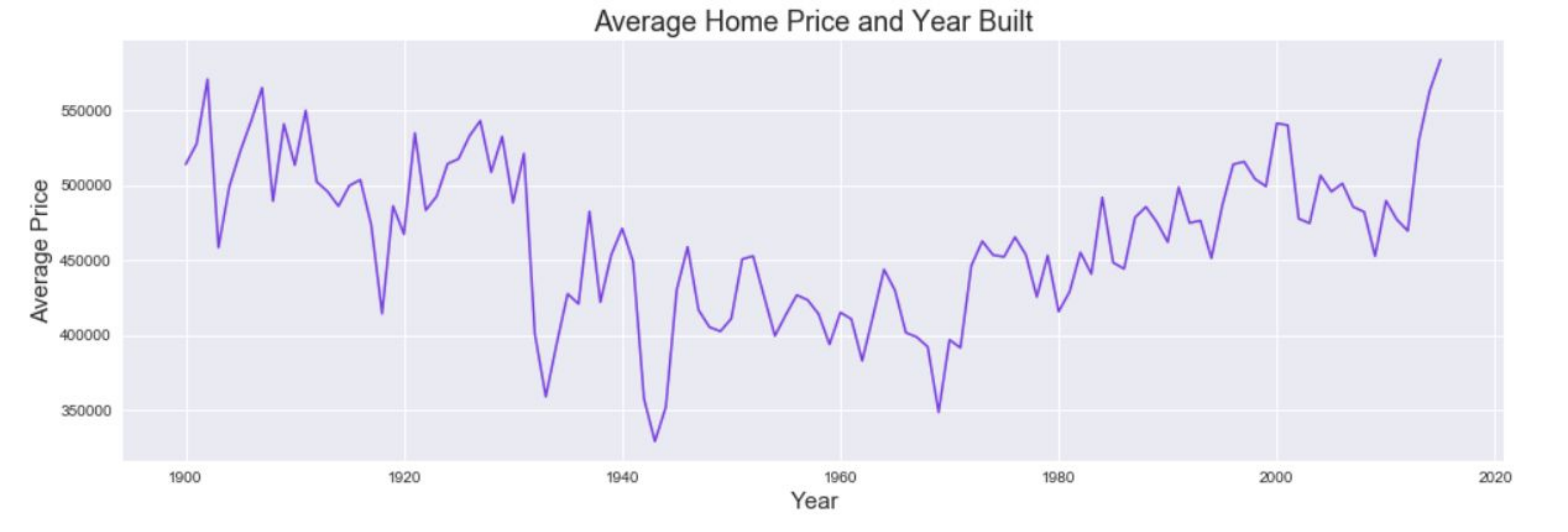
Column Names and descriptions for Kings County Data Set

- **id** - unique identified for a house
- **dateDate** - house was sold
- **pricePrice** - is prediction target
- **bedroomsNumber** - of Bedrooms/House
- **bathroomsNumber** - of bathrooms/bedrooms
- **sqft_livingsquare** - footage of the home
- **sqft_lotsquare** - footage of the lot
- **floorsTotal** - floors (levels) in house
- **waterfront** - House which has a view to a waterfront
- **view** - Has been viewed
- **condition** - How good the condition is (Overall)
- **grade** - overall grade given to the housing unit, based on King County grading system
- **sqft_above** - square footage of house apart from basement
- **sqft_basement** - square footage of the basement
- **yr_built** - Built Year
- **yr_renovated** - Year when house was renovated
- **zipcode** - zip
- **lat** - Latitude coordinate
- **long** - Longitude coordinate
- **sqft_living15** - The square footage of interior housing living space for the nearest 15 neighbors
- **sqft_lot15** - The square footage of the land lots of the nearest 15 neighbors

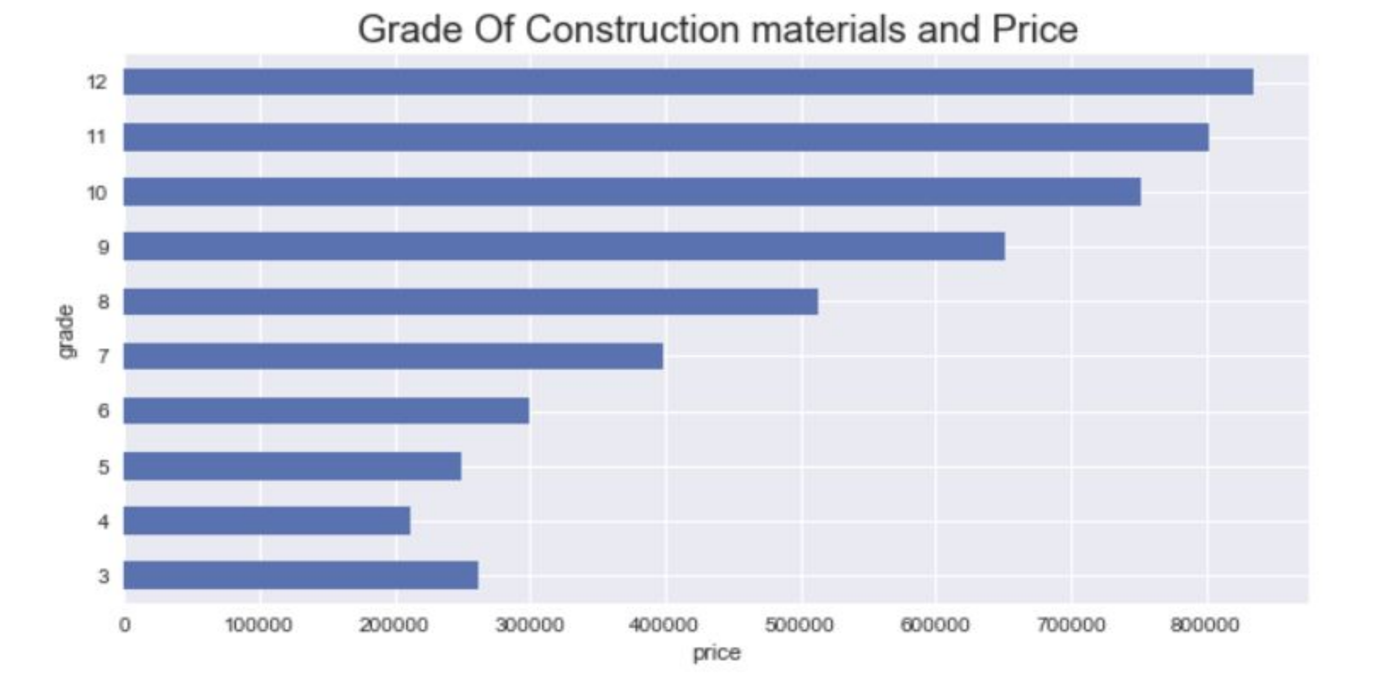
Cleaning Data for target audience

- Created new column of Price per Square Foot
  - Very common real estate metric which was not included in original data
- Removed outliers
  - Price outliers
    - Extremely expensive or inexpensive houses will skew results
  - House size Outliers
    - Number of Bedroom, Bathroom, and Floors
- Removed Unnecessary columns
  - Id
  - Date
  - Year Renovated

Understanding Target Market



Understanding the market



Understanding the market

Model objective:
Make improvements from a baseline model to create a model which accurately predicts house prices for our target audience

Ordinary Least Squares (OLS) Regression Key Terms

- **R-Squared (r^2)**
  - Represents the a goodness-of-fit measure for the entire model
    - For example, a r^2 score of .68 means that 68% of the data fits the regression model
- **Coefficients**
  - Describes mathematical relationship between independent and dependent variable
- **P-Value**
  - Describes weather the relationship between independent and dependent variable is statistically significant

Baseline Model

- Price will be used as dependent variable throughout modeling process
- R-squared value = .699
- Root Mean Square Error(RMSE)
  - Standard deviation of the residuals
    - Residuals are a measure of how far from the regression line the data points lie.
  - Measure of concentration of data points around the line of best fit
- RMES for for baseline model:
  - $198314.69

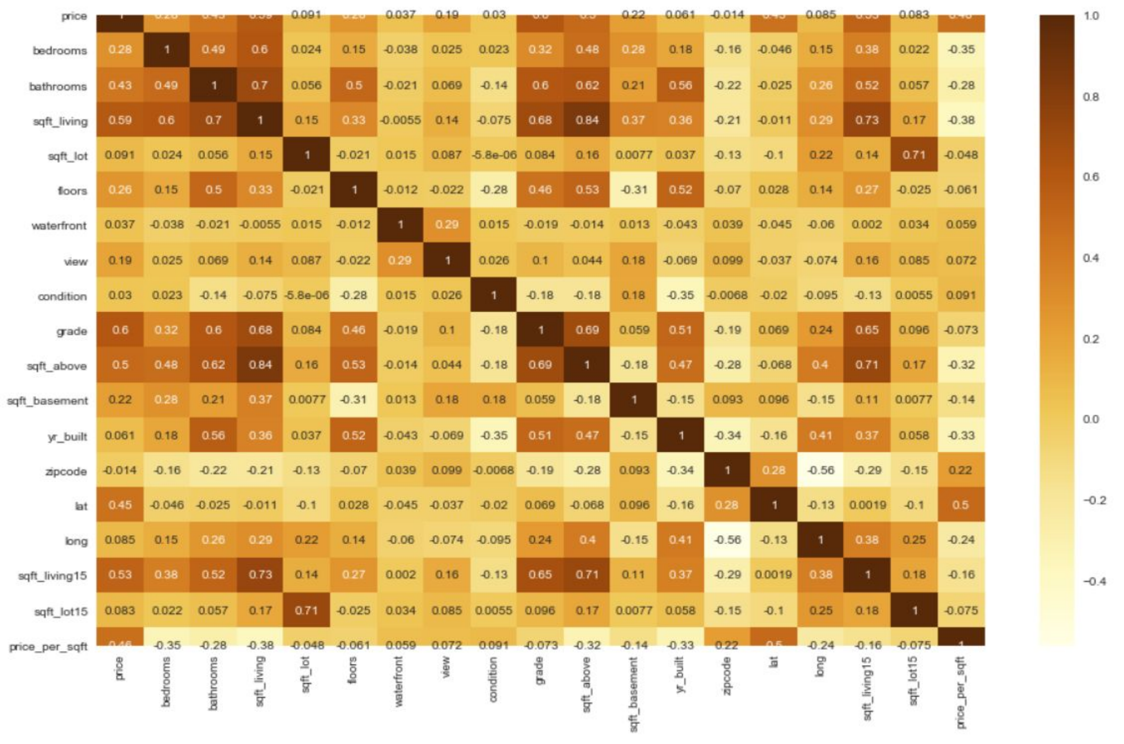
OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.699
Model:	OLS	**Adj. R-squared:**	0.699
Method:	Least Squares	**F-statistic:**	2361.
Date:	Sat, 28 Nov 2020	**Prob (F-statistic):**	0.00
Time:	13:03:34	**Log-Likelihood:**	-2.3543e+05
No. Observations:	17290	**AIC:**	4.709e+05
Df Residuals:	17272	**BIC:**	4.710e+05
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
id	-1.307e-06	5.32e-07	-2.456	0.014	-2.35e-06	-2.64e-07
bedrooms	-3.405e+04	2079.721	-16.371	0.000	-3.81e+04	-3e+04
bathrooms	4.45e+04	3607.657	12.334	0.000	3.74e+04	5.16e+04
sqft_living	108.7783	2.530	43.002	0.000	103.820	113.737
sqft_lot	0.0865	0.058	1.495	0.135	-0.027	0.200
floors	5528.2263	3896.712	1.419	0.156	-2109.725	1.32e+04
waterfront	5.625e+05	1.95e+04	28.831	0.000	5.24e+05	6.01e+05
view	5.326e+04	2346.748	22.694	0.000	4.87e+04	5.79e+04
condition	2.521e+04	2560.508	9.848	0.000	2.02e+04	3.02e+04
grade	9.426e+04	2363.737	39.876	0.000	8.96e+04	9.89e+04
sqft_above	70.5935	2.468	28.600	0.000	65.755	75.432
sqft_basement	38.1848	2.915	13.101	0.000	32.472	43.898
yr_built	-2622.2225	75.143	-34.896	0.000	-2769.510	-2474.934
yr_renovated	20.8861	4.051	5.156	0.000	12.946	28.826
zipcode	-486.9280	19.824	-24.562	0.000	-525.785	-448.071
lat	5.949e+05	1.19e+04	50.135	0.000	5.72e+05	6.18e+05
long	-1.96e+05	1.45e+04	-13.488	0.000	-2.25e+05	-1.68e+05
sqft_living15	22.6219	3.747	6.038	0.000	15.278	29.966
sqft_lot15	-0.3330	0.082	-4.064	0.000	-0.494	-0.172

Correlation

price	1.000000
grade	0.600028
sqft_living	0.587663
sqft_living15	0.533609
sqft_above	0.495720
price_per_sqft	0.458125
lat	0.451340
bathrooms	0.426600
bedrooms	0.283191
floors	0.262949
sqft_basement	0.219279
view	0.192805
sqft_lot	0.091171
long	0.084751
sqft_lot15	0.083089
yr_built	0.061039
waterfront	0.036707
condition	0.029584
zipcode	-0.014223



Training/Improving Model

- **Categorical Variables**
  - Assigned to non continuous or obvious categorical data
  - Zip Codes were given the categorical designation.
    - Assigned dummy variables to represent each zip code within model.
    - Allows for the significance of zip code to have its individual influence on model

- **Collinear Variables**
  - Variables that are highly correlated with other variables can skew results.
  - Dropping collinear variables will improve the predictive power and results for model.
  - Sqft_above was chosen to be dropped

pairs	
(sqft_above, sqft_living)	0.844410
(sqft_living15, sqft_living)	0.728793
(sqft_lot, sqft_lot15)	0.708516
(sqft_living15, sqft_above)	0.708036
(bathrooms, sqft_living)	0.703585

Additional Feature Selection Through Stepwise

- **Many of the remaining variables had high P-values**
  - Use stepwise function to iterate through model and remove all variables deemed statistically insignificant
    - All values with a P-value higher than .05
- **Test-Train-Split**
  - Creates two, randomly chosen subsets to compare
  - One set of data us unchanged and one set is used to train model
- **Cross Validation**
  - Technique used for assessing the statistical analysis on a model
    - Used to estimate accuracy not improve accuracy

Final Model

- Final R-squared value = .919
  - Over 90% of the data can be explained within the model
- Root Mean Square Error(RMSE) for both test and train set:

```
RMSE for train set: 52809.0694374566
RMSE for test set: 54877.40702905297
RMSE difference: -2068.3375915963697
```

- Cross Validation Scores:

```
10 Cross Validation R^2 score for train: 0.917662066078903
10 Cross Validation R^2 score for test: 0.9118442675351133
```

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.919
Model:	OLS	**Adj. R-squared:**	0.918
Method:	Least Squares	**F-statistic:**	2814.
Date:	Wed, 25 Nov 2020	**Prob (F-statistic):**	0.00
Time:	16:59:54	**Log-Likelihood:**	-1.9383e+05
No. Observations:	15768	**AIC:**	3.878e+05
Df Residuals:	15704	**BIC:**	3.883e+05
Df Model:	63		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
sqft_living15	5.7552	1.220	4.717	0.000	3.364	8.147
price_per_sqft	1100.3271	7.670	143.460	0.000	1085.293	1115.361
grade	2.331e+04	732.405	31.825	0.000	2.19e+04	2.47e+04
const	-1.368e+07	3.4e+05	-40.196	0.000	-1.43e+07	-1.3e+07
bedrooms	6370.9119	646.033	9.862	0.000	5104.613	7637.210
sqft_living	168.0949	1.386	121.259	0.000	165.378	170.812
lat	2.907e+05	7095.332	40.975	0.000	2.77e+05	3.05e+05
condition	1.168e+04	735.279	15.891	0.000	1.02e+04	1.31e+04
zip_98040	9.595e+04	5700.928	16.830	0.000	8.48e+04	1.07e+05
zip_98006	4.252e+04	3408.461	12.476	0.000	3.58e+04	4.92e+04
view	1.193e+04	764.024	15.611	0.000	1.04e+04	1.34e+04
zip_98155	-7.891e+04	3402.740	-23.191	0.000	-8.56e+04	-7.22e+04
zip_98133	-7.174e+04	3216.353	-22.305	0.000	-7.8e+04	-6.54e+04
zip_98028	-7.747e+04	3983.178	-19.449	0.000	-8.53e+04	-6.97e+04
zip_98019	-8.632e+04	4692.968	-18.393	0.000	-9.55e+04	-7.71e+04
zip_98005	5.07e+04	5155.315	9.834	0.000	4.06e+04	6.08e+04
zip_98004	7.269e+04	5762.365	12.615	0.000	6.14e+04	8.4e+04
zip_98011	-7.327e+04	4706.360	-15.569	0.000	-8.25e+04	-6.4e+04
zip_98125	-5.298e+04	3381.274	-15.670	0.000	-5.96e+04	-4.64e+04
zip_98034	-4.802e+04	3073.738	-15.622	0.000	-5.4e+04	-4.2e+04

Final Model Conclusions

- **Dramatic increase in R-squared score**
  - **Goodness of fit along regression line was greatly improved**
  - **About 30% increase in strength of the relationship between price and the various independent variables**
- **Dramatic decrease in Root Mean Squared Error Metric**
  - Allows for predictions to be far more accurate.
    - Can estimate the price of a house with and error of about $52809
    - Substantially lower the the $198314 we saw in the baseline model
- **Both Test-Train-Split and cross validation confirmed results**
  - A small Test-Train-Split difference conforms model can be applied to entire data set
  - Cross validation scores confirm the models accuracy throughout entire data set

Future Considerations

- **Dramatic increase in R-squared score**
  - **Goodness of fit along regression line was greatly improved**
  - **About 30% increase in strength of the relationship between price and the various independent variables**
- **Dramatic decrease in Root Mean Squared Error Metric**
  - Allows for predictions to be far more accurate.
    - Can estimate the price of a house with and error of about $52809
    - Substantially lower the the $198314 we saw in the baseline model
- **Both Test-Train-Split and cross validation confirmed results**
  - A small Test-Train-Split difference conforms model can be applied to entire data set
  - Cross validation scores confirm the models accuracy throughout entire data set