King County House Sales

Priscilla Suero

Housing Sales

- Predicting house prices from prior house sales
- Provided the King County House Sales Dataset
- Data from 2014-2015
- Over 19,000 house sales

Data Snapshot

- Bedrooms: 1 - 8
- Bathroom: .5 - 4.75
- Square Footage of Lot:
  - 520 sqft - 49,936 sqft
- Built Year: 1900-2015

Features

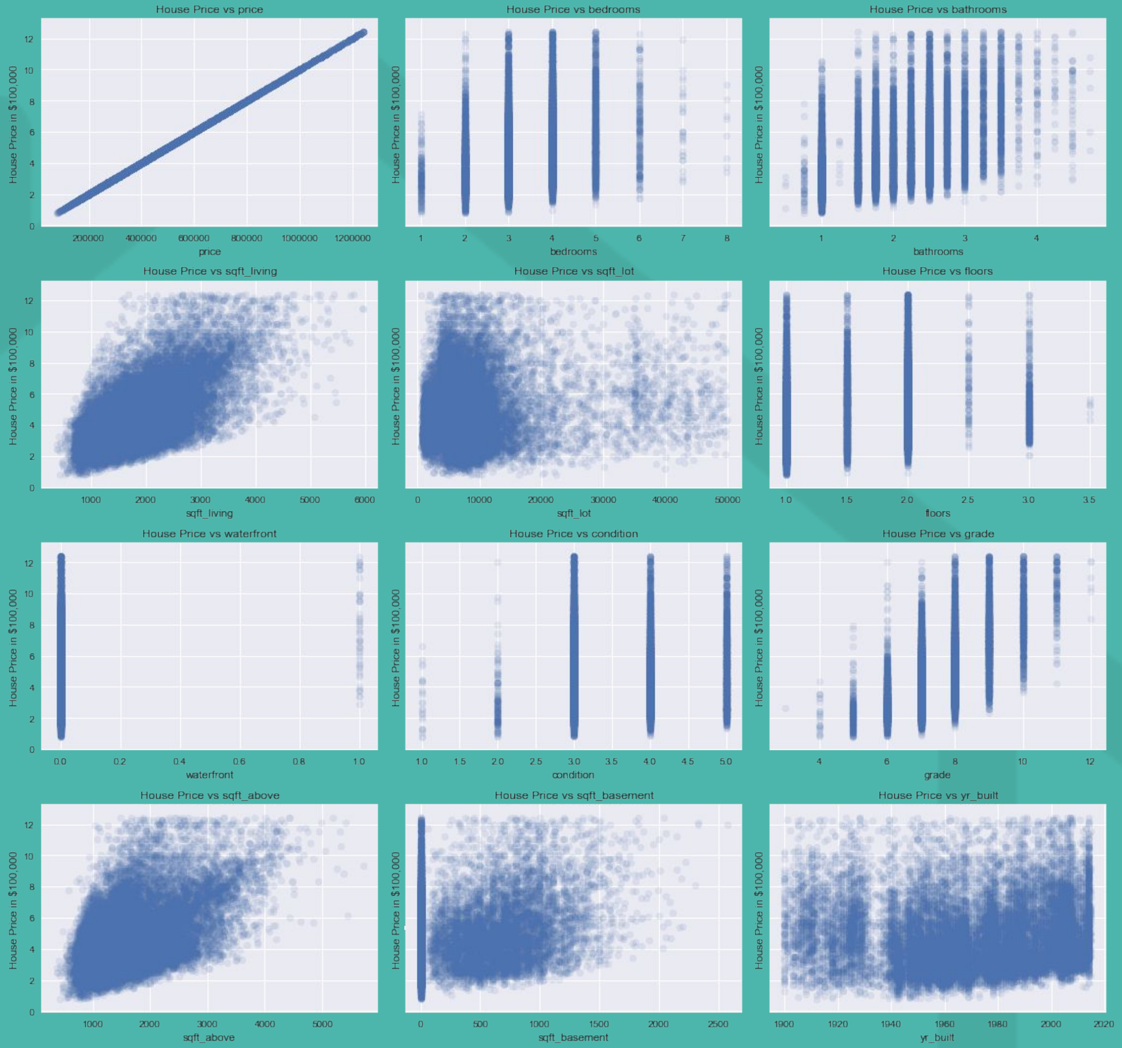
- **id** - unique identified for a house
- **dateDate** - house was sold
- **pricePrice** - is prediction target
- **bedroomsNumber** - of Bedrooms/House
- **bathroomsNumber** - of bathrooms/bedrooms
- **sqft_livingsquare** - footage of the home
- **sqft_lotsquare** - footage of the lot
- **floorsTotal** - 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

Continuous vs Categorical

- Continouous:
  - Price, Sqft features
- Categorical:
  - Bedrooms, bathrooms, floors, conditions, grade, and waterfront
- These are the variables that can help determine how our model will work

Linearity

- Helpful in determining features
- Linearity in sqft_living and sqft_above shows promise



Initial Modeling

- Initial model is interpreted as having an error of $135,833.92
- So my first model can explain 61.2% of the response data around its mean.

```
Train Mean Squared Error: 18734722588.94
Test Mean Squared Error: 18450854084.6
R Squared: 0.612
Mean Absolute Error: 104553.88
Root Mean Squared Error: 135833.92
Average Predicted Price: 480476.9
Average Actual Price: 480593.35
Difference: -116.45
```

Out[5]:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.613
Model:	OLS	**Adj. R-squared:**	0.612
Method:	Least Squares	**F-statistic:**	613.7
Date:	Wed, 04 Nov 2020	**Prob (F-statistic):**	0.00
Time:	12:58:45	**Log-Likelihood:**	-2.6278e+05
No. Observations:	19842	**AIC:**	5.257e+05
Df Residuals:	19790	**BIC:**	5.261e+05
Df Model:	51		
Covariance Type:	nonrobust		

Final Model

```
Train Mean Squared Error: 18735448148.89
Test Mean Squared Error: 18451939902.3
R Squared: 0.612
Mean Absolute Error: 104551.6
Root Mean Squared Error: 135837.92
Average Predicted Price: 480461.78
Average Actual Price: 480593.35
Difference: -131.57
```

- Final model is interpreted as having an error of $135,837.92
    - A bit larger than the first model

Final model can explain 61.2% of the response data around its mean, similar to first model

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.613
Model:	OLS	**Adj. R-squared:**	0.612
Method:	Least Squares	**F-statistic:**	626.0
Date:	Wed, 04 Nov 2020	**Prob (F-statistic):**	0.00
Time:	14:03:05	**Log-Likelihood:**	-2.6278e+05
No. Observations:	19842	**AIC:**	5.257e+05
Df Residuals:	19791	**BIC:**	5.261e+05
Df Model:	50		
Covariance Type:	nonrobust		

Final Model Interpretation

- Looking at my model, we can see a few items in the OLS on how some specific features can add a difference to housing prices

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.282e+06	1.29e+05	41.017	0.000	5.03e+06	5.53e+06
sqft_lot	-1.782e+04	2474.571	-7.201	0.000	-2.27e+04	-1.3e+04
sqft_above	82.6487	3.132	26.389	0.000	76.510	88.788
sqft_basement	94.2678	3.350	28.141	0.000	87.702	100.834
yr_built	-3015.7895	94.669	-31.856	0.000	-3201.348	-2830.231
yr_renovated	318.0090	96.668	3.290	0.001	128.531	507.487
sqft_living15	4.434e+04	1604.584	27.634	0.000	4.12e+04	4.75e+04
sqft_lot15	-1.636e+04	2375.878	-6.885	0.000	-2.1e+04	-1.17e+04

flr_1.5	9986.2435	4001.973	2.495	0.013	2142.042	1.78e+04
flr_2.0	7131.5363	3617.774	1.971	0.049	40.395	1.42e+04
flr_2.5	3.629e+04	1.38e+04	2.638	0.008	9320.932	6.33e+04
flr_3.0	7.265e+04	7340.368	9.898	0.000	5.83e+04	8.7e+04
flr_3.5	7.129e+04	5.61e+04	1.270	0.204	-3.87e+04	1.81e+05
con_2	3.465e+04	2.97e+04	1.166	0.244	-2.36e+04	9.29e+04
con_3	6.809e+04	2.76e+04	2.464	0.014	1.39e+04	1.22e+05

- For 1 unit increase of the feature below the house price increases by the following:
  - Square foot for house without basement increases $82.64
  - Square footage of basement increases $94.26
  - For having a second floor increases $7,131.53

Next Steps

- Look deeper into the location
- How location affects price
- Zipcodes
    - But also the longitude and latitude that was provided

Thank you!

LinkedIn

Github Repo

sueropriscilla@gmail.com