King County Real Estate Model

Business Case

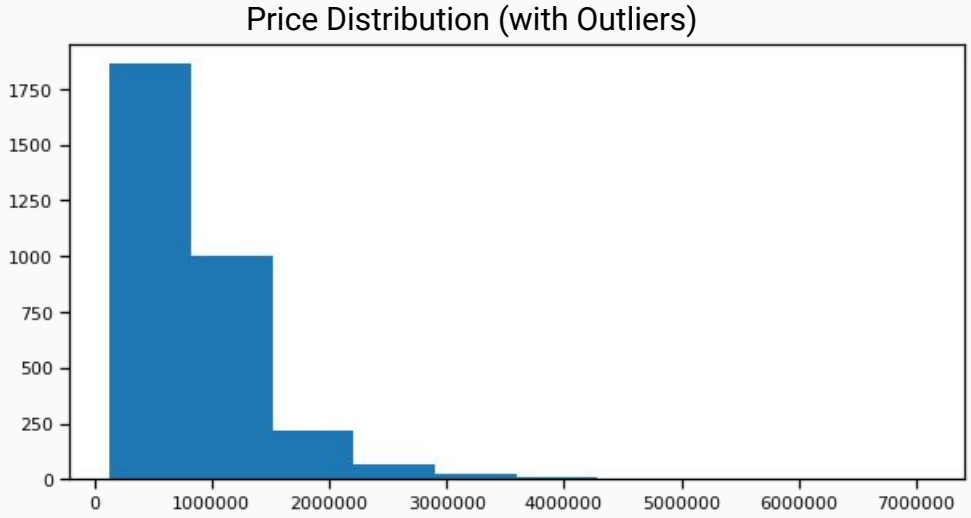
King County Real Estate is a luxury real estate company serving sellers and buyers in the high income earning areas of King County, Seattle. The company wants to understand which features translate to higher housing prices in these areas, as well as develop a model to predict price based on housing features.

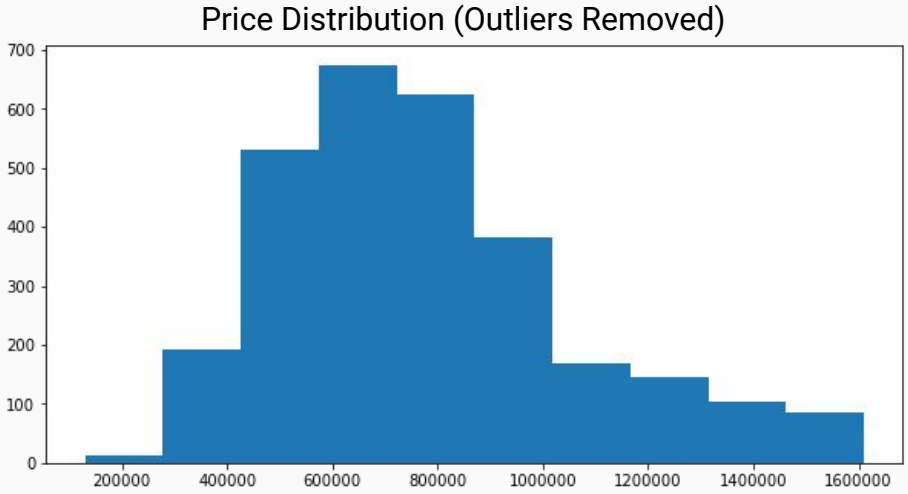
The Data

- Data was sourced from King County Housing Dataset CSV
- Additional census data (AGI) taken from census.gov
- Total data used was from 3192 homes split 80/20 for training and testing
- Variables include price, id, date sold, bedrooms, bathrooms, sqft living, sqft lot, floors, waterfront, condition, grade, sqft above, sqft basement, year built, year renovated, zip code, latitude, longitude, sqft living (houses within 15 blocks), sqft lot (houses within 15 blocks)

EDA - Price

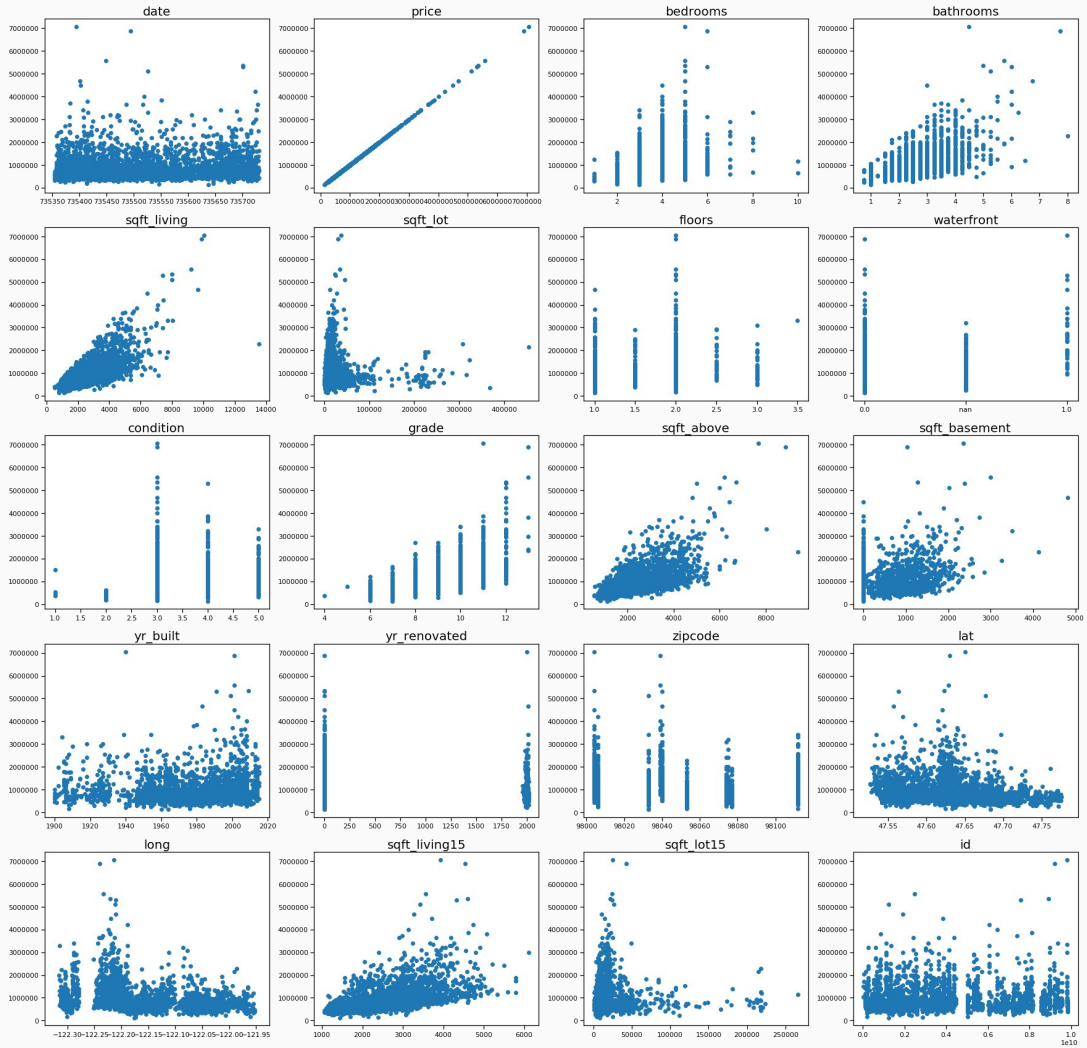
- Range of $130,000 to $1,600,000 after removing outliers
- Fairly normally distributed
- Median price: $730,500
- Std Dev: $283,000





EDA - Features

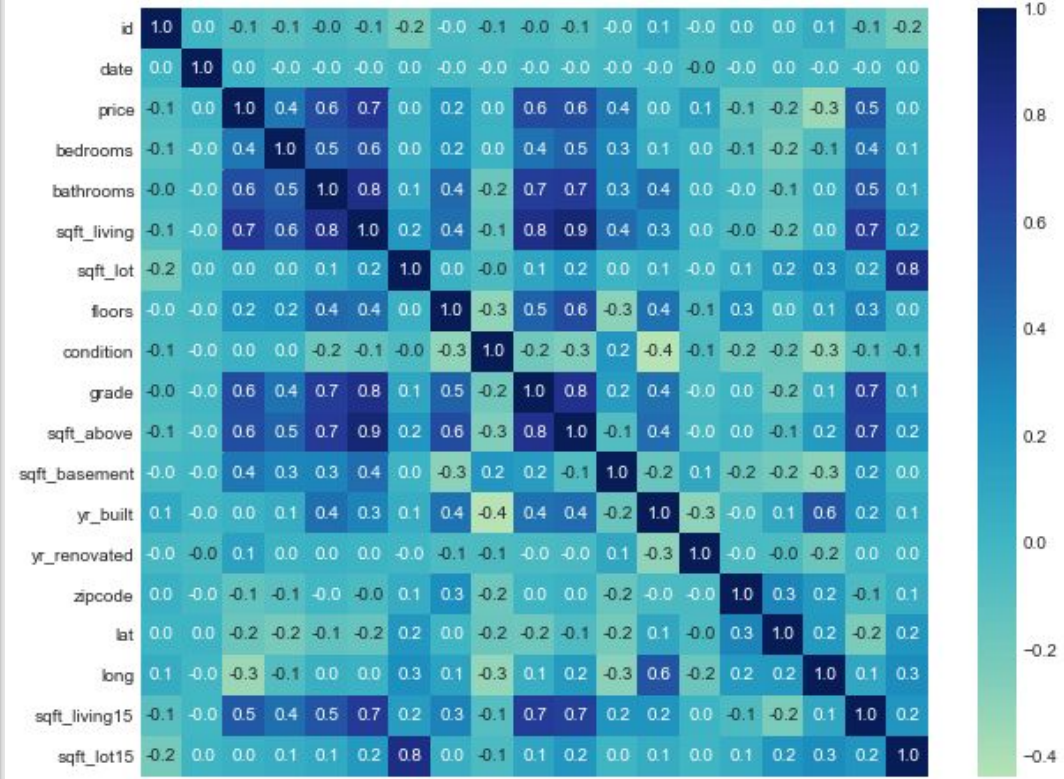
- Certain variables initially appear to have linear relationship with dependent variable (sqft living, bathrooms, bedrooms, sqft above, sqft living 15, sqft basement)



EDA - Assumptions

Strong multicollinearity between certain features in baseline:

- Sqft living, sqft above
- Sqft above, bathrooms
- Sqft lot, sqft lot 15
- Sqft living, sqft living 15
- Sqft above, grade

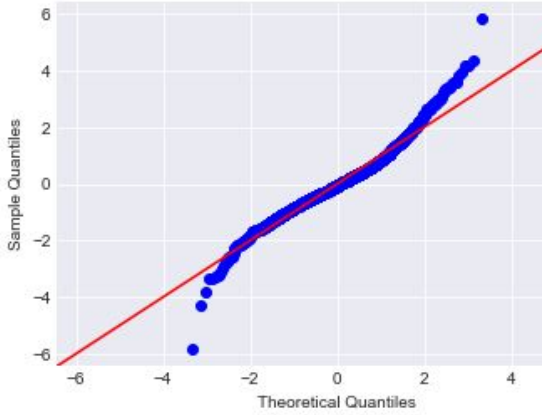


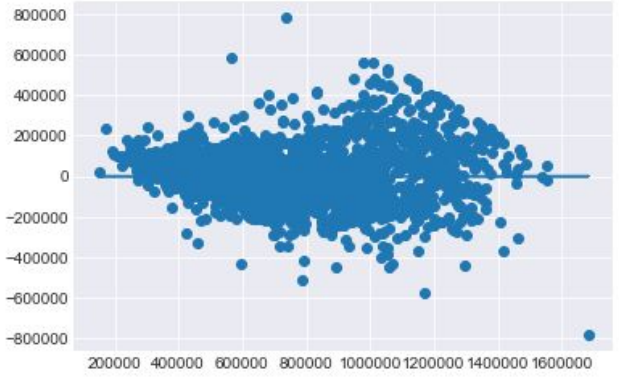
Baseline Model

- 77% of variation in housing prices explained by features
- RMSE: On average, model is $134,000 off in its predictions
- QQ plot still showing heavy tails (JB score: 666)
- Model displays slight heteroscedasticity
- Many features with insignificant p-values, and displays multicollinearity

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.779
Model:	OLS	**Adj. R-squared:**	0.774
Method:	Least Squares	**F-statistic:**	160.6
Date:	Thu, 19 Nov 2020	**Prob (F-statistic):**	0.00
Time:	17:35:28	**Log-Likelihood:**	-30867.
No. Observations:	2334	**AIC:**	6.184e+04
Df Residuals:	2283	**BIC:**	6.213e+04
Df Model:	50		
Covariance Type:	nonrobust		



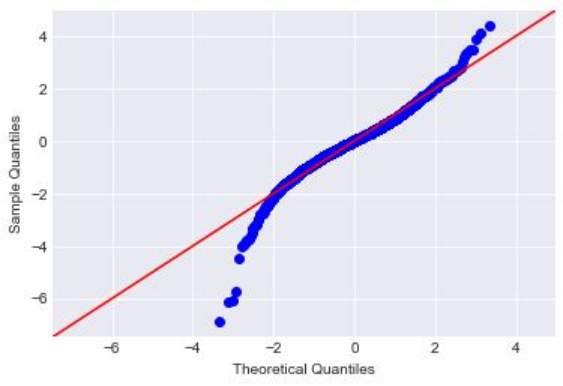


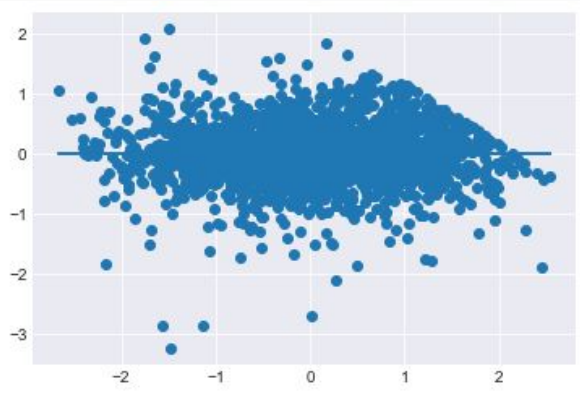
Final Model

- 78% of variation in housing prices explained by features
- RMSE: On average, model is $138,708 off in its predictions
- QQ plot still showing heavy tails (JB score: 1467)
- Model displays less heteroscedasticity
- Multicollinear features removed, as well as insignificant p-values
- **Bottom Line:** Best scoring model that met regression assumptions

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.780
Model:	OLS	**Adj. R-squared:**	0.778
Method:	Least Squares	**F-statistic:**	327.5
Date:	Sun, 22 Nov 2020	**Prob (F-statistic):**	0.00
Time:	18:36:25	**Log-Likelihood:**	-1543.7
No. Observations:	2334	**AIC:**	3139.
Df Residuals:	2308	**BIC:**	3289.
Df Model:	25		
Covariance Type:	nonrobust		





Formula

Housing Price =

Y-int	Date log	# Bath log	Sqft above log	Year built log	Lat log	Sqft living 15 log	Sqft basement	Long	Never renovated	On water	Cond 5	Grade 11	Grade 12	2 floors
-443.00	0.05	0.08	0.58	0.08	0.07	0.25	0.00	-3.63	-0.07	1.66	0.25	0.31	0.42	-0.19

3 floors	7 beds	98006	98033	98039	98040	98053	98074	98075	98077	98112
-0.28	-1.39	-0.86	-0.88	0.32	-0.47	-0.92	-0.90	-0.73	-1.29	-0.39

Conclusions

- Significant features in luxury homes include waterfront property, location (zip codes, longitude), and square foot above ground
- Having more floors or bedrooms does not necessarily imply higher sale price
- **Bottom Line:** location and square footage are the most important features in determining sale price

Next Steps

1. Refine dataset (expand and cut certain zip codes)
2. Subset model for different price ranges
3. Investigate polynomial relationships and interactions between variables in greater detail


Thank You

Modeling

present your final model (or an intermediate one if you think it adds to your presentation)

briefly talk about any feature selection/feature engineering that you did

eliminated variables based on p-values, removed collinear variables

interaction terms, scaling, transforming, etc.

evaluate your model and interpret some of the more important coefficients