King County Real Estate - Data Modelling

Iman Kumarasinghe

Outline

- Business Problem
- Data
- Methods
- Results
- Conclusions

Business Problem

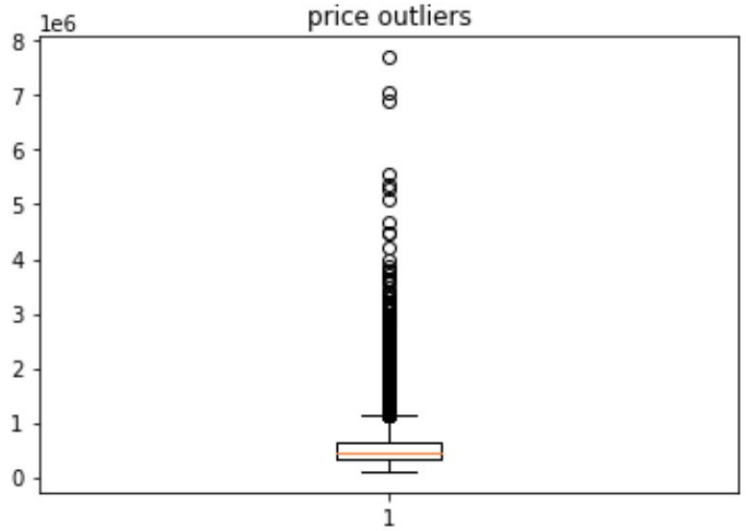
- King County Real Estate is helping home buyers to purchase homes.
- They need help assisting singles and couples in purchasing affordable homes.

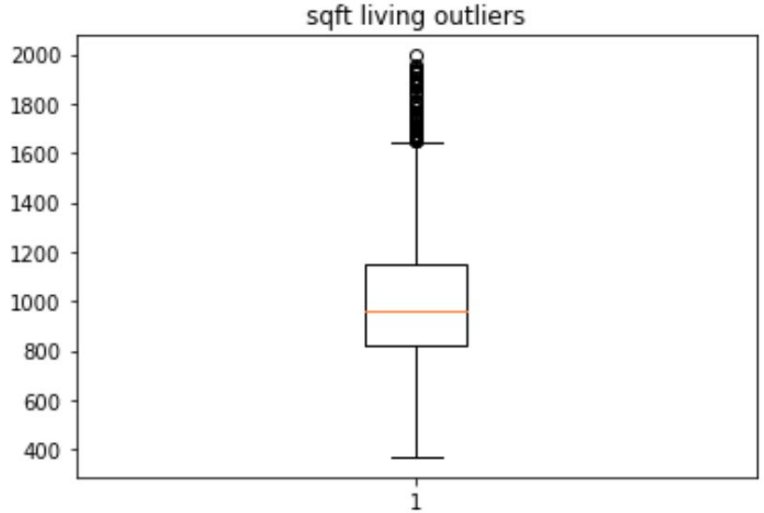
Data - Data Cleaning

- Removed columns that weren’t going to be used for the modelling
- Converting data types - fixing date and sqft_basement columns
- Cleaning the NaNs
- Cleaning the dependent variable - price
- Cleaning independent variables like bedrooms, bathrooms, sqft_living, sqft_above
- Grade being categorical - created dummy variables

Data - Checking

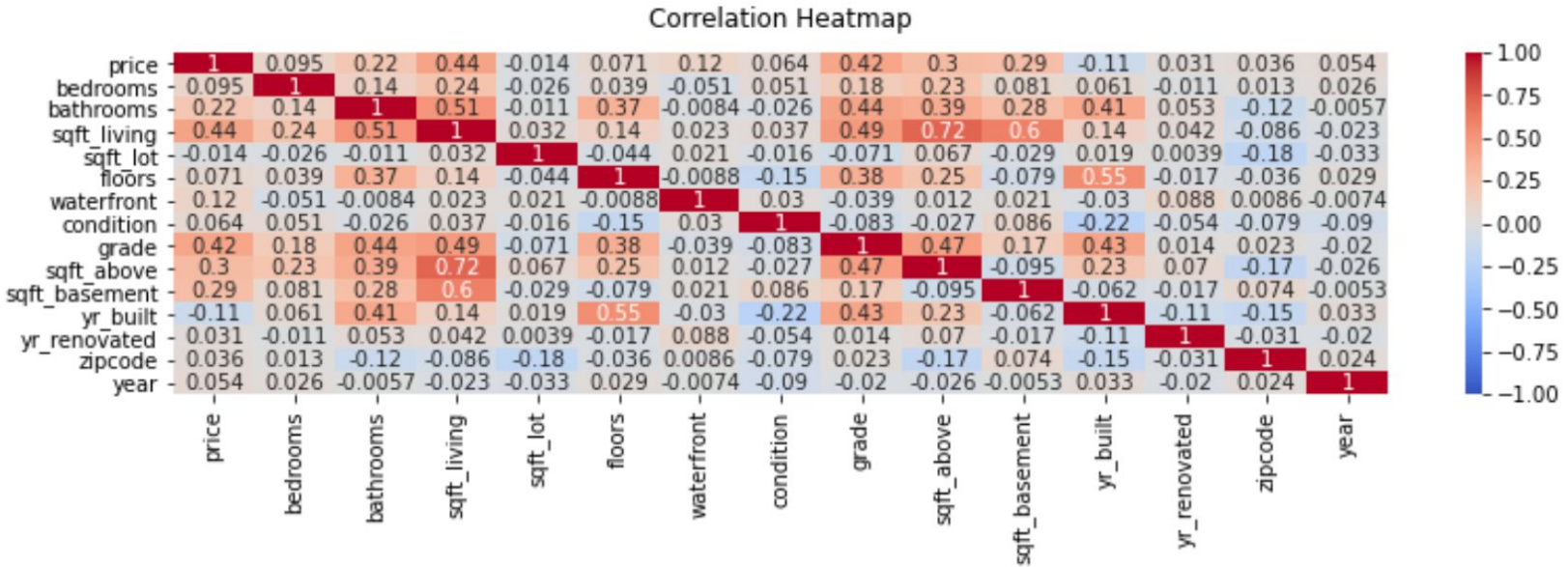
- Removing outliers





Data - Checking

- Correlation Heatmap



Model Creation - First Model

- Test 30%
- Train 70%

Model Creation - First Model

OLS Regression Results

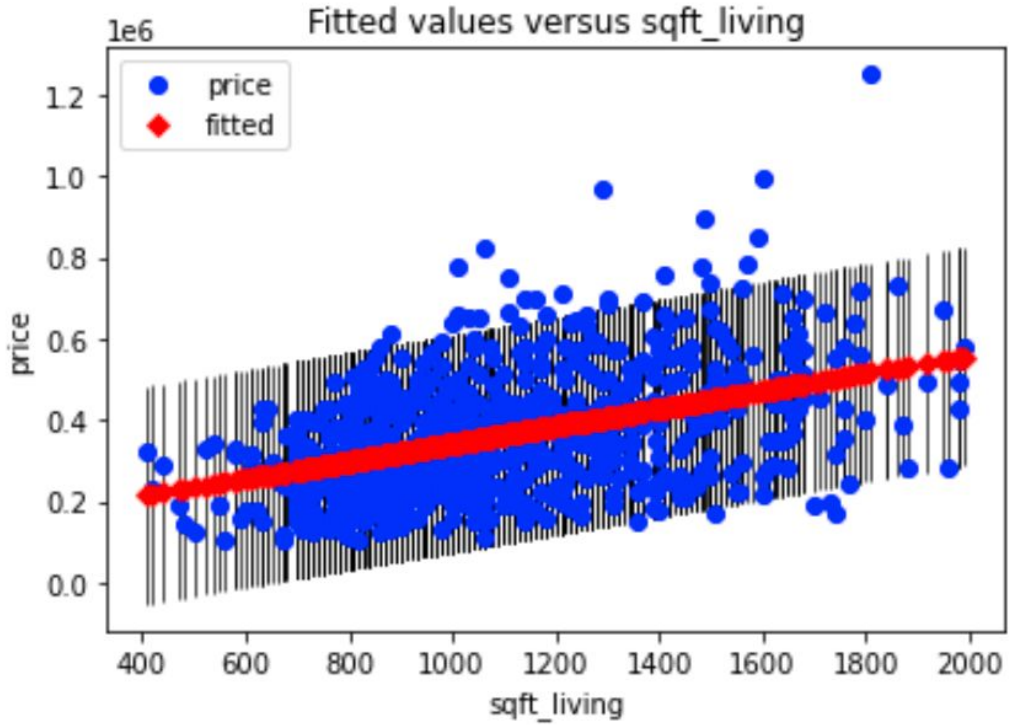
Dep. Variable:	price	**R-squared:**	0.194
Model:	OLS	**Adj. R-squared:**	0.193
Method:	Least Squares	**F-statistic:**	175.1
Date:	Sun, 24 Sep 2023	**Prob (F-statistic):**	5.65e-36
Time:	20:26:30	**Log-Likelihood:**	-9676.7
No. Observations:	731	**AIC:**	1.936e+04
Df Residuals:	729	**BIC:**	1.937e+04
Df Model:	1		
Covariance Type:	nonrobust		

Model Creation - First Model

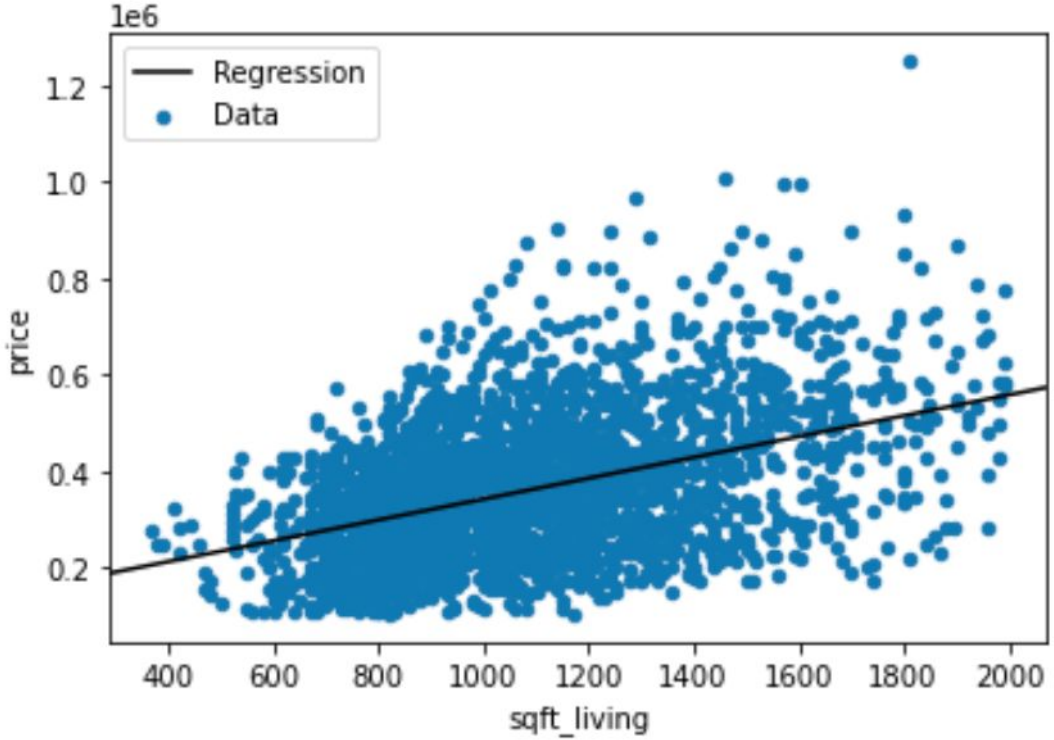
	coef	std err	t	P>|t|	[0.025	0.975]
const	1.255e+05	1.86e+04	6.763	0.000	8.9e+04	1.62e+05
sqft_living	216.3117	16.348	13.232	0.000	184.217	248.407

Omnibus:	82.728	**Durbin-Watson:**	1.942
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	136.488
Skew:	0.748	**Prob(JB):**	2.30e-30
Kurtosis:	4.499	**Cond. No.**	4.19e+03

Fitted values



Regression Line



Model Creation - Second Model

- Variables used were
  - Sqft_living
  - Grade
  - Sqft_basement
  - Bathrooms
  - Bedrooms
  - Waterfront
  - Yr_built

Model Creation - Second Model

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.361
Model:	OLS	**Adj. R-squared:**	0.359
Method:	Least Squares	**F-statistic:**	195.4
Date:	Sun, 24 Sep 2023	**Prob (F-statistic):**	3.22e-230
Time:	20:26:30	**Log-Likelihood:**	-31894.
No. Observations:	2434	**AIC:**	6.380e+04
Df Residuals:	2426	**BIC:**	6.385e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.159e+06	1.84e+05	17.127	0.000	2.8e+06	3.52e+06
sqft_living	110.2521	12.210	9.029	0.000	86.308	134.196
grade	8e+04	3751.286	21.327	0.000	7.26e+04	8.74e+04
sqft_basement	30.3883	14.897	2.040	0.041	1.175	59.601
bathrooms	1.919e+04	8712.856	2.202	0.028	2102.610	3.63e+04
waterfront	2.1e+05	2.9e+04	7.228	0.000	1.53e+05	2.67e+05
yr_built	-1767.9747	99.040	-17.851	0.000	-1962.187	-1573.763
bedrooms	-1.353e+04	9542.929	-1.418	0.156	-3.22e+04	5178.406

Omnibus:	258.390	**Durbin-Watson:**	1.973
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	496.694
Skew:	0.689	**Prob(JB):**	1.39e-108
Kurtosis:	4.732	**Cond. No.**	1.72e+05

Model Creation - Third Model

- Variables used were
  - Sqft_living
  - Grade
  - Sqft_basement
  - Bathrooms
  - Yr_built

Model Creation - Third Model

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.346
Model:	OLS	**Adj. R-squared:**	0.345
Method:	Least Squares	**F-statistic:**	256.8
Date:	Sun, 24 Sep 2023	**Prob (F-statistic):**	9.55e-221
Time:	20:26:31	**Log-Likelihood:**	-31922.
No. Observations:	2434	**AIC:**	6.386e+04
Df Residuals:	2428	**BIC:**	6.389e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.152e+06	1.86e+05	16.979	0.000	2.79e+06	3.52e+06
sqft_living	110.6077	12.136	9.114	0.000	86.809	134.406
grade	7.84e+04	3781.920	20.731	0.000	7.1e+04	8.58e+04
sqft_basement	31.6411	15.019	2.107	0.035	2.190	61.092
bathrooms	1.878e+04	8808.185	2.132	0.033	1508.717	3.61e+04
yr_built	-1771.3892	100.114	-17.694	0.000	-1967.706	-1575.072

Omnibus:	308.070	**Durbin-Watson:**	1.957
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	664.527
Skew:	0.759	**Prob(JB):**	5.01e-145
Kurtosis:	5.061	**Cond. No.**	1.71e+05

Model Creation - Fourth Model

- Variables used were
  - Sqft_living
  - Grade
  - Sqft_basement
  - Bathrooms
  - Yr_built
- Log transformation

Model Creation - Fourth Model

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.346
Model:	OLS	**Adj. R-squared:**	0.345
Method:	Least Squares	**F-statistic:**	256.8
Date:	Sun, 24 Sep 2023	**Prob (F-statistic):**	9.55e-221
Time:	20:26:32	**Log-Likelihood:**	-31922.
No. Observations:	2434	**AIC:**	6.386e+04
Df Residuals:	2428	**BIC:**	6.389e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.152e+06	1.86e+05	16.979	0.000	2.79e+06	3.52e+06
sqft_living	110.6077	12.136	9.114	0.000	86.809	134.406
grade	7.84e+04	3781.920	20.731	0.000	7.1e+04	8.58e+04
sqft_basement	31.6411	15.019	2.107	0.035	2.190	61.092
bathrooms	1.878e+04	8808.185	2.132	0.033	1508.717	3.61e+04
yr_built	-1771.3892	100.114	-17.694	0.000	-1967.706	-1575.072

Omnibus:	308.070	**Durbin-Watson:**	1.957
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	664.527
Skew:	0.759	**Prob(JB):**	5.01e-145
Kurtosis:	5.061	**Cond. No.**	1.71e+05

Final Results

- Little difference between third and fourth models
- R squared explains 35% of the variance in price
- Variables were statistically significant
- An increase in grade = $78000
- An increase in basement space = $31
- An increase in bathroom = $19190
- An increase in sqft living space = $110
- Kurtosis is over 3 - still may be outliers
- Model is statistically significant

Recommendations

1. Square ft of a home, square ft of the homes basement, the grade, number of bathrooms and the year a house was built can explain around 35% of the variance in house prices.
2. Consider the square footage of a home
3. Pay attention to the grade given
4. Consider how many bathrooms needed

Limitations

- Further iterations
- Other variables can be included such as the number of people who have shown interest in the home either through the website or with the real estate.

Thank You!

Email: imannnk@hotmail.com
GitHub: @imannnk