Pricing Homes in King County

March 2023

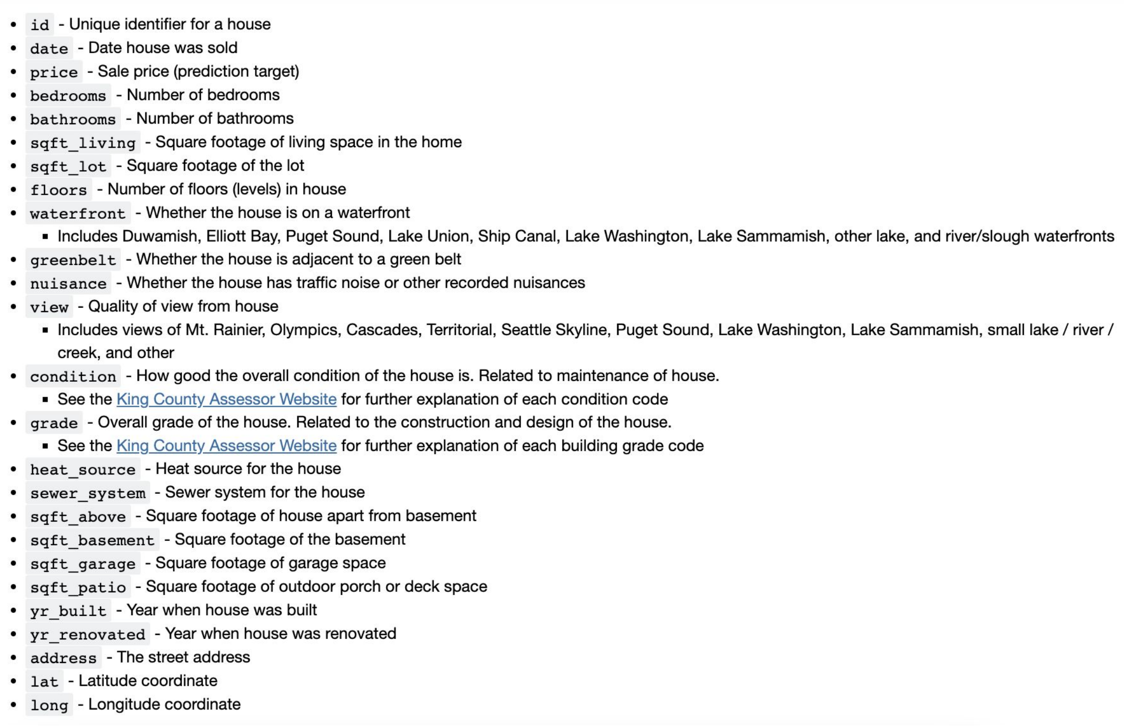
Business Problem and Overview

Information asymmetries is an issue within the housing market. The goal of this analysis is to derive a regression model that aids homebuyers and homesellers in King County by ***pricing the average property*** in the region. Constructing a model will create information transparency and ***encourage competitive, fair prices.***

The Data

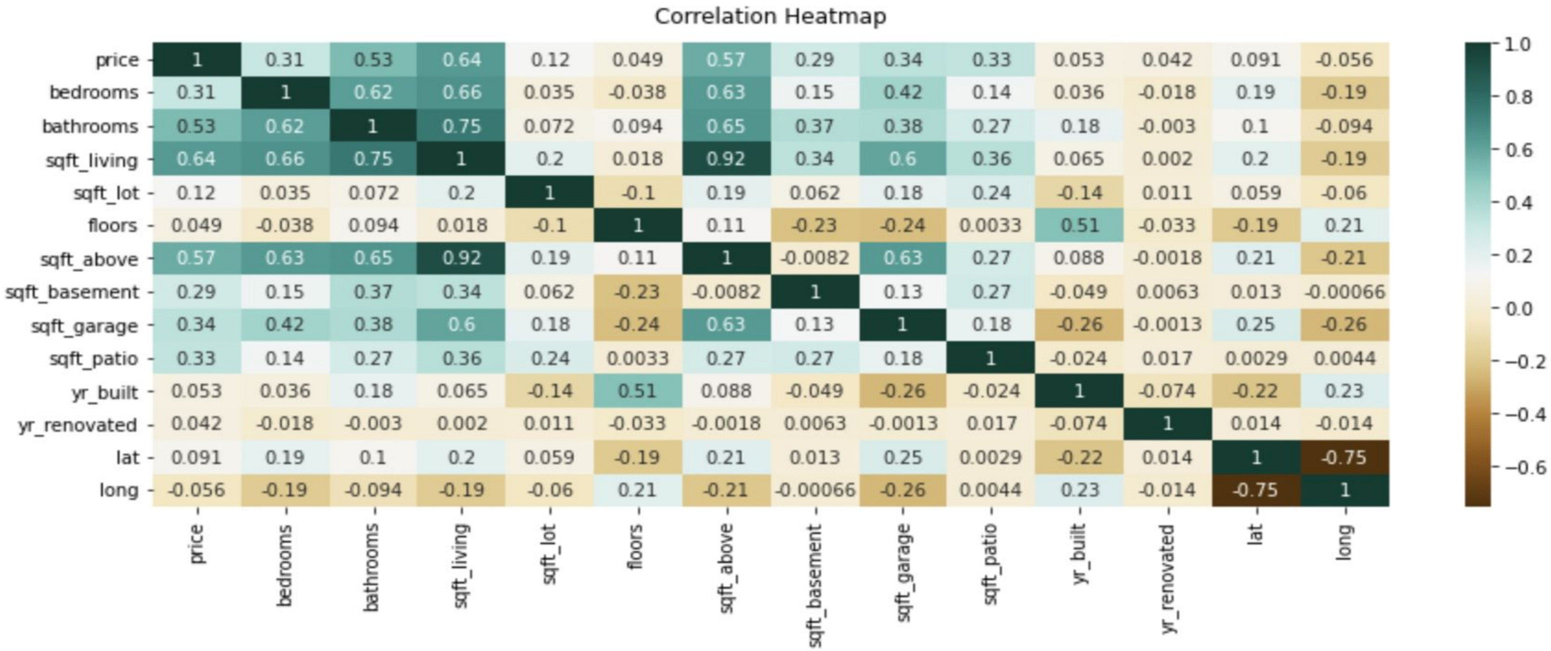
- To create a regression model, public data on homes sold is obtained from King County's government website

- A table showing the dataset's columns and their descriptions is displayed on the right



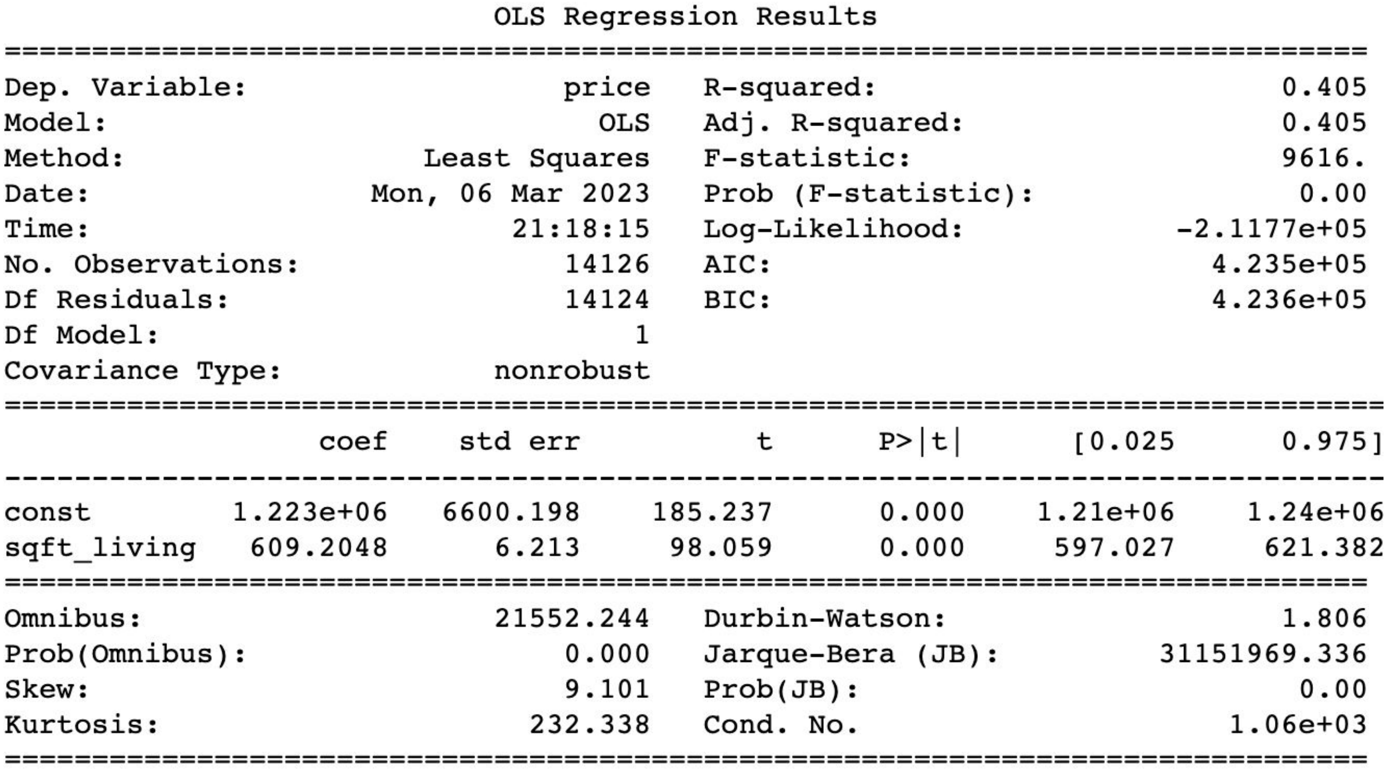
- `id` - Unique identifier for a house
- `date` - Date house was sold
- `price` - Sale price (prediction target)
- `bedrooms` - Number of bedrooms
- `bathrooms` - Number of bathrooms
- `sqft_living` - Square footage of living space in the home
- `sqft_lot` - Square footage of the lot
- `floors` - Number of floors (levels) in house
- `waterfront` - Whether the house is on a waterfront
  - Includes Duwamish, Elliott Bay, Puget Sound, Lake Union, Ship Canal, Lake Washington, Lake Sammamish, other lake, and river/slough waterfronts
- `greenbelt` - Whether the house is adjacent to a green belt
- `nuisance` - Whether the house has traffic noise or other recorded nuisances
- `view` - Quality of view from house
  - Includes views of Mt. Rainier, Olympics, Cascades, Territorial, Seattle Skyline, Puget Sound, Lake Washington, Lake Sammamish, small lake / river / creek, and other
- `condition` - How good the overall condition of the house is. Related to maintenance of house.
  - See the King County Assessor Website for further explanation of each condition code
- `grade` - Overall grade of the house. Related to the construction and design of the house.
  - See the King County Assessor Website for further explanation of each building grade code
- `heat_source` - Heat source for the house
- `sewer_system` - Sewer system for the house
- `sqft_above` - Square footage of house apart from basement
- `sqft_basement` - Square footage of the basement
- `sqft_garage` - Square footage of garage space
- `sqft_patio` - Square footage of outdoor porch or deck space
- `yr_built` - Year when house was built
- `yr_renovated` - Year when house was renovated
- `address` - The street address
- `lat` - Latitude coordinate
- `long` - Longitude coordinate

Results: Initial Analysis and Baseline Model



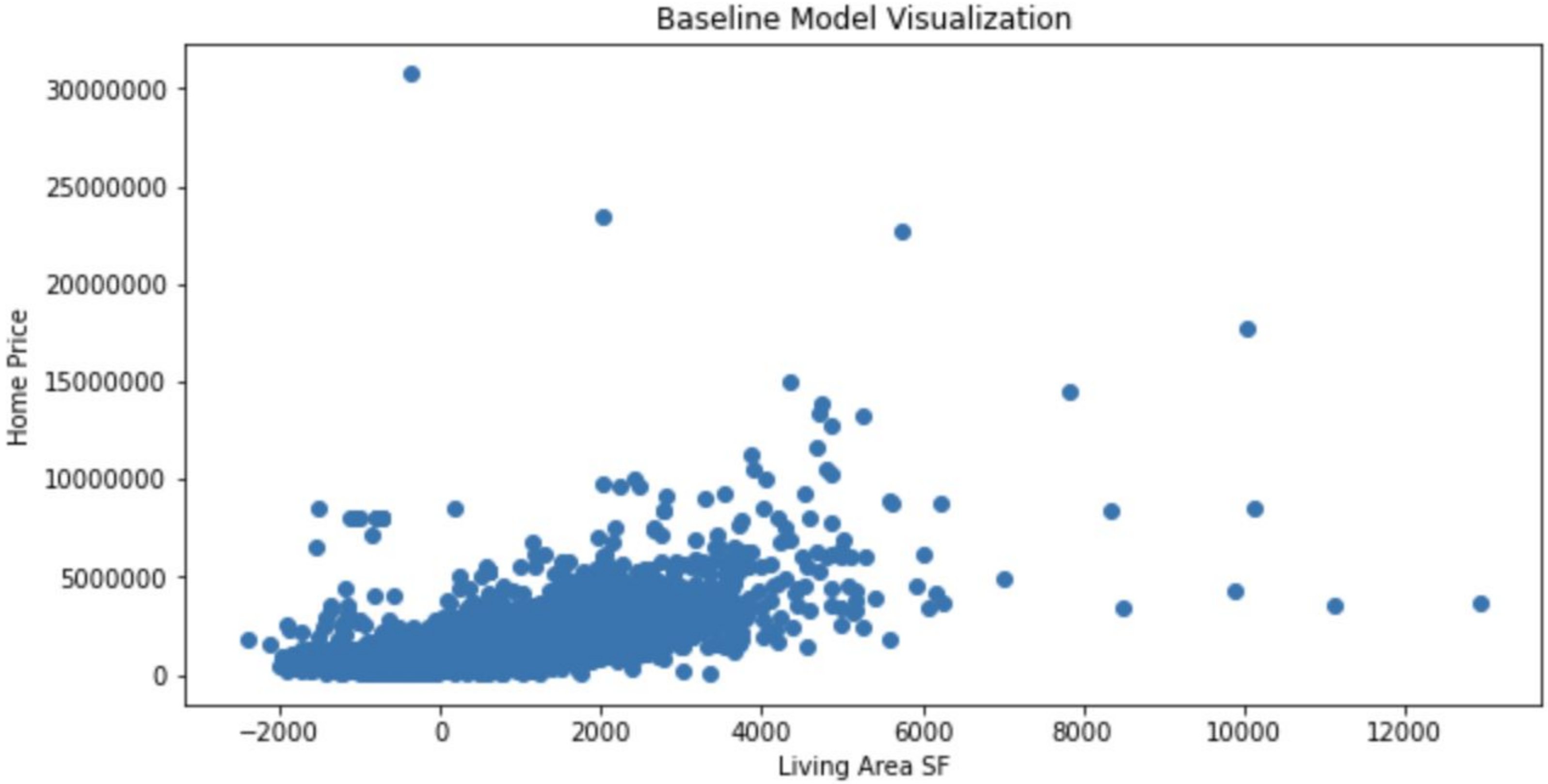
	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	sqft_above	sqft_basement	sqft_garage	sqft_patio	yr_built	yr_renovated	lat	long
price	1	0.31	0.53	0.64	0.12	0.049	0.57	0.29	0.34	0.33	0.053	0.042	0.091	-0.056
bedrooms	0.31	1	0.62	0.66	0.035	-0.038	0.63	0.15	0.42	0.14	0.036	-0.018	0.19	-0.19
bathrooms	0.53	0.62	1	0.75	0.072	0.094	0.65	0.37	0.38	0.27	0.18	-0.003	0.1	-0.094
sqft_living	0.64	0.66	0.75	1	0.2	0.018	0.92	0.34	0.6	0.36	0.065	0.002	0.2	-0.19
sqft_lot	0.12	0.035	0.072	0.2	1	-0.1	0.19	0.062	0.18	0.24	-0.14	0.011	0.059	-0.06
floors	0.049	-0.038	0.094	0.018	-0.1	1	0.11	-0.23	-0.24	0.0033	0.51	-0.033	-0.19	0.21
sqft_above	0.57	0.63	0.65	0.92	0.19	0.11	1	-0.0082	0.63	0.27	0.088	-0.0018	0.21	-0.21
sqft_basement	0.29	0.15	0.37	0.34	0.062	-0.23	-0.0082	1	0.13	0.27	-0.049	0.0063	0.013	-0.00066
sqft_garage	0.34	0.42	0.38	0.6	0.18	-0.24	0.63	0.13	1	0.18	-0.26	-0.0013	0.25	-0.26
sqft_patio	0.33	0.14	0.27	0.36	0.24	0.0033	0.27	0.27	0.18	1	-0.024	0.017	0.0029	0.0044
yr_built	0.053	0.036	0.18	0.065	-0.14	0.51	0.088	-0.049	-0.26	-0.024	1	-0.074	-0.22	0.23
yr_renovated	0.042	-0.018	-0.003	0.002	0.011	-0.033	-0.0018	0.0063	-0.0013	0.017	-0.074	1	0.014	-0.014
lat	0.091	0.19	0.1	0.2	0.059	-0.19	0.21	0.013	0.25	0.0029	-0.22	0.014	1	-0.75
long	-0.056	-0.19	-0.094	-0.19	-0.06	0.21	-0.21	-0.00066	-0.26	0.0044	0.23	-0.014	-0.75	1

A seaborn heatmap shows that “**sqft_living**” is the **most correlated variable with price**



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.405
Model:                            OLS   Adj. R-squared:                  0.405
Method:                 Least Squares   F-statistic:                     9616.
Date:                Mon, 06 Mar 2023   Prob (F-statistic):               0.00
Time:                        21:18:15   Log-Likelihood:            -2.1177e+05
No. Observations:               14126   AIC:                         4.235e+05
Df Residuals:                   14124   BIC:                         4.236e+05
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const        1.223e+06   6600.198    185.237      0.000    1.21e+06    1.24e+06
sqft_living   609.2048      6.213     98.059      0.000     597.027     621.382
===============================================================================
Omnibus:                    21552.244   Durbin-Watson:                   1.806
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         31151969.336
Skew:                           9.101   Prob(JB):                         0.00
Kurtosis:                     232.338   Cond. No.                     1.06e+03
==============================================================================
```

The baseline model is statistically significant and accounts for **40.5% of price variance**. With a **unit increase in living area**, we can expect price to rise by **$609**.



The baseline model is statistically significant and accounts for **40.5% of price variance**. With a **unit increase in living area**, we can expect price to rise by **$609**.

Results: Overview of Data Transformations

Model 2

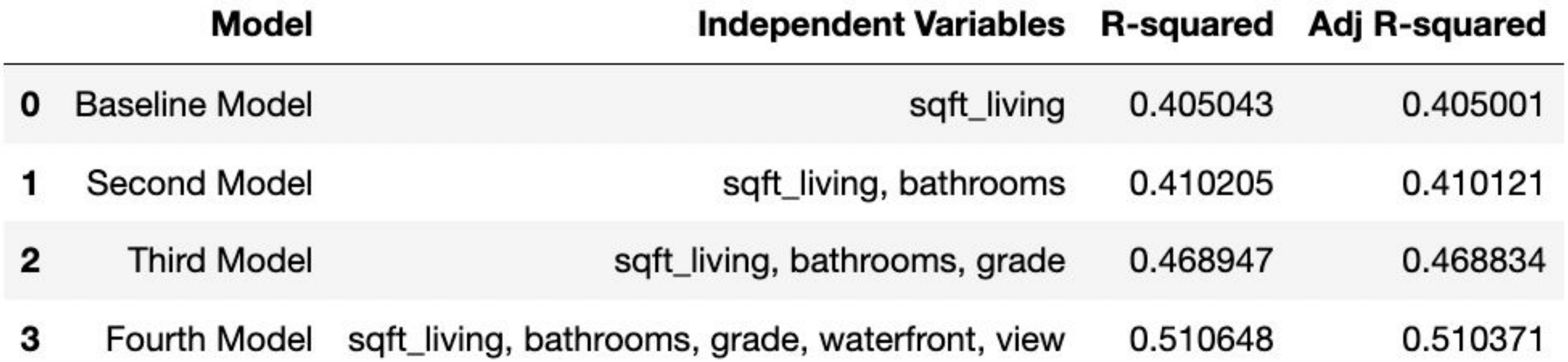
- Adding and centering the discrete variable (“bathrooms”) to create a multiple regression model

Model 3

- Re-mapping and centering the discrete column “grade” to transform it from a string into model-friendly numbers

Model 4

- Incorporating categorical variables (“view” and “waterfront”) and repeated one-hot-encoding to assess the impact that amenities have on average properties

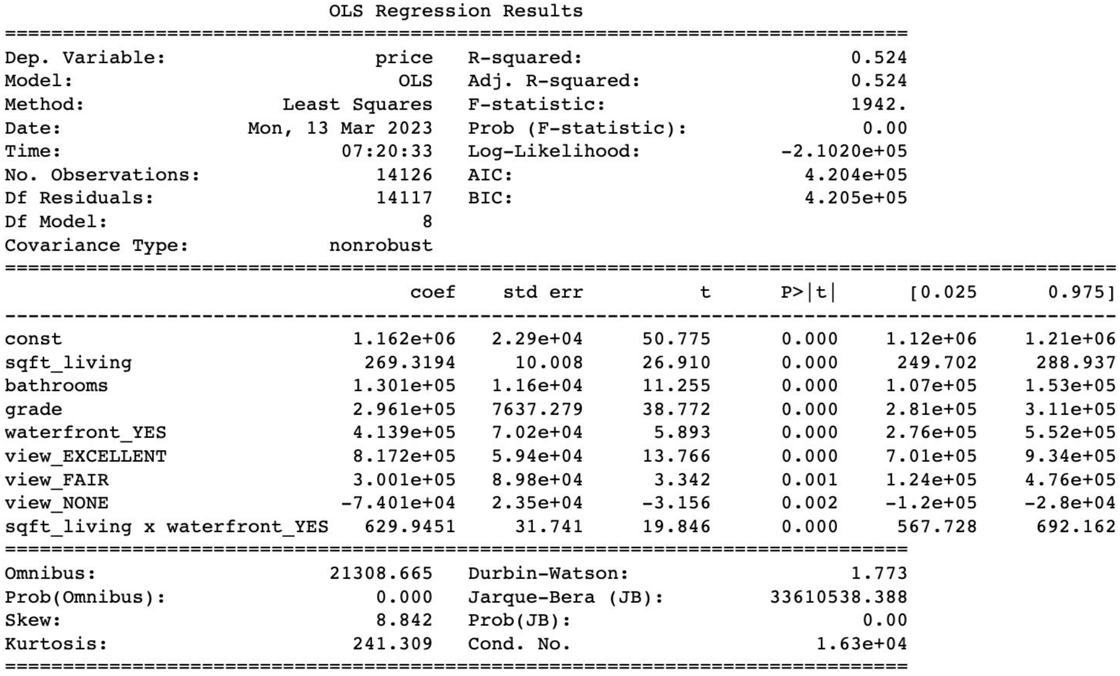


	Model	Independent Variables	R-squared	Adj R-squared
0	Baseline Model	sqft_living	0.405043	0.405001
1	Second Model	sqft_living, bathrooms	0.410205	0.410121
2	Third Model	sqft_living, bathrooms, grade	0.468947	0.468834
3	Fourth Model	sqft_living, bathrooms, grade, waterfront, view	0.510648	0.510371

Results: Summary of Models 2-4

Results: Final Regression Model

The final model builds on the fourth by creating an interaction term from the relevant predictors



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.524
Model:                            OLS   Adj. R-squared:                  0.524
Method:                 Least Squares   F-statistic:                     1942.
Date:                Mon, 13 Mar 2023   Prob (F-statistic):               0.00
Time:                        07:20:33   Log-Likelihood:            -2.1020e+05
No. Observations:               14126   AIC:                         4.204e+05
Df Residuals:                   14117   BIC:                         4.205e+05
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                      1.162e+06   2.29e+04     50.775      0.000    1.12e+06    1.21e+06
sqft_living                 269.3194     10.008     26.910      0.000     249.702     288.937
bathrooms                  1.301e+05   1.16e+04     11.255      0.000    1.07e+05    1.53e+05
grade                      2.961e+05   7637.279     38.772      0.000    2.81e+05    3.11e+05
waterfront_YES             4.139e+05   7.02e+04      5.893      0.000    2.76e+05    5.52e+05
view_EXCELLENT             8.172e+05   5.94e+04     13.766      0.000    7.01e+05    9.34e+05
view_FAIR                  3.001e+05   8.98e+04      3.342      0.001    1.24e+05    4.76e+05
view_NONE                 -7.401e+04   2.35e+04     -3.156      0.002    -1.2e+05    -2.8e+04
sqft_living x waterfront_YES  629.9451     31.741     19.846      0.000     567.728     692.162
==============================================================================
Omnibus:                    21308.665   Durbin-Watson:                   1.773
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         33610538.388
Skew:                           8.842   Prob(JB):                         0.00
Kurtosis:                     241.309   Cond. No.                     1.63e+04
==============================================================================
```

Results: Final Regression Model (Data Interpretation)

- The model accounts for **52.4% of the variance in sale price** and models against a reference home with:
  - Average living area
  - Average number of bathrooms
  - Average grade
  - Average views
  - No waterfront
- The model prices the **typical home** with the aforementioned features at **~$1.2 Million**
- We expect a unit increase in “sqft_living” to raise the value of an average home by $269
- We expect a unit increase in “grade” to raise the value of an average home by ~$300K
- Adjusting variables within a home with "nice-to-have" amenities, such as a waterfront, has a greater impact on value
  - A unit increase in "sqft_living" for an average-sized home with a waterfront adds ~$629 instead of $269

Future Considerations

- ❖ Map geographic distribution of homes with “lat” and “long”
- ❖ Leverage other public data from King County’s website such as population and socioeconomic information

Email: kezia.setokusumo@stern.nyu.edu
Github: @keziasetokusumo
LinkedIn: https://www.linkedin.com/in/keziasetokusumo/