King County Housing Market Model

Executive Overview

Usecase:

- Build a predictive model using regression and analyze effect on target feature

Stakeholders:

- King County Housing Administration
- King County Board of Supervisors
- Home Facilities Management Supply Chain

Data

- The data used is a CSV file of all home sales in King County.
- The records span from 2021-2022

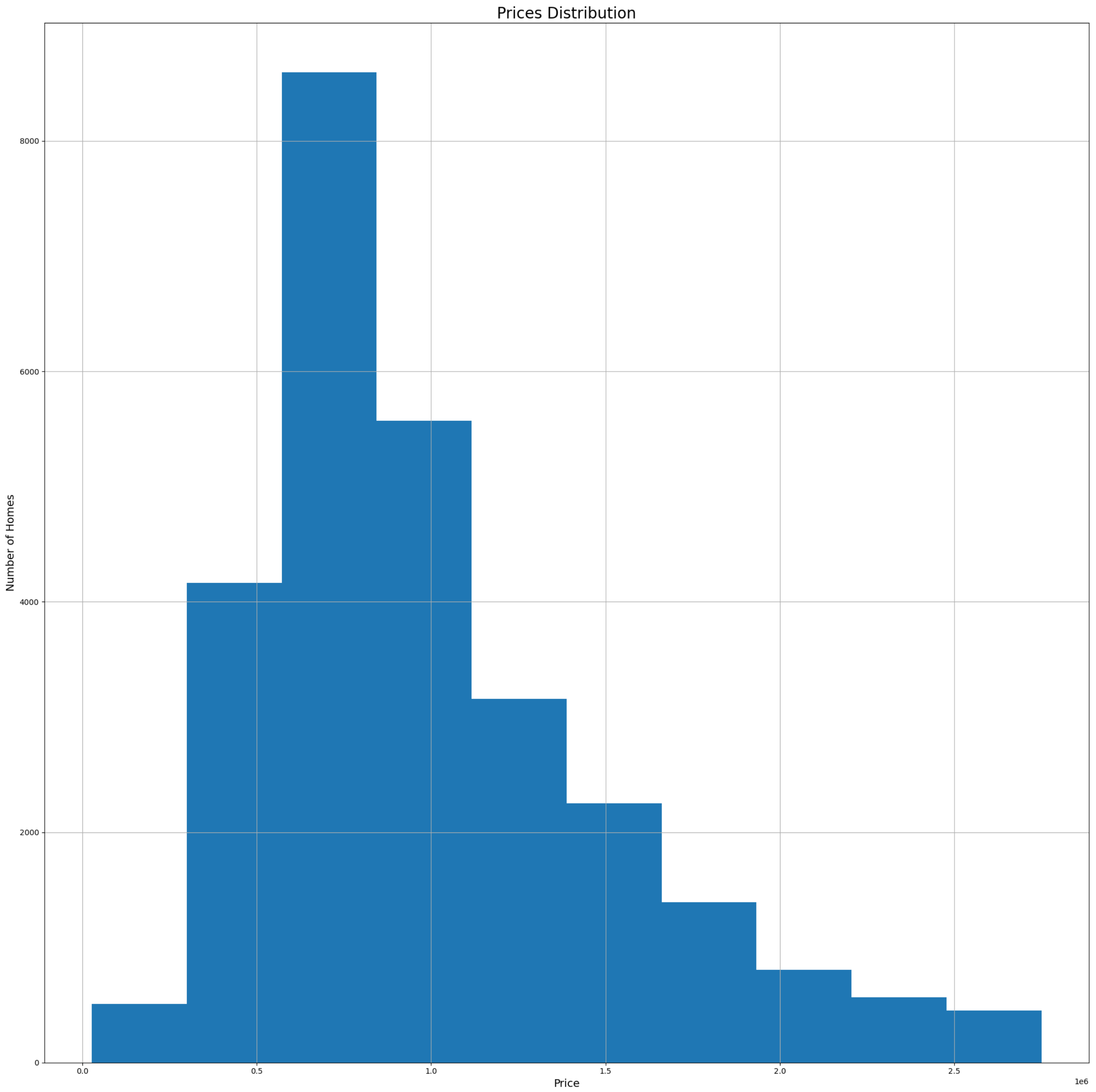
- BeautifulSoup also used to make calls to Washington Hometown Locator, a website containing map data to help us incorporate King County only sales

Features

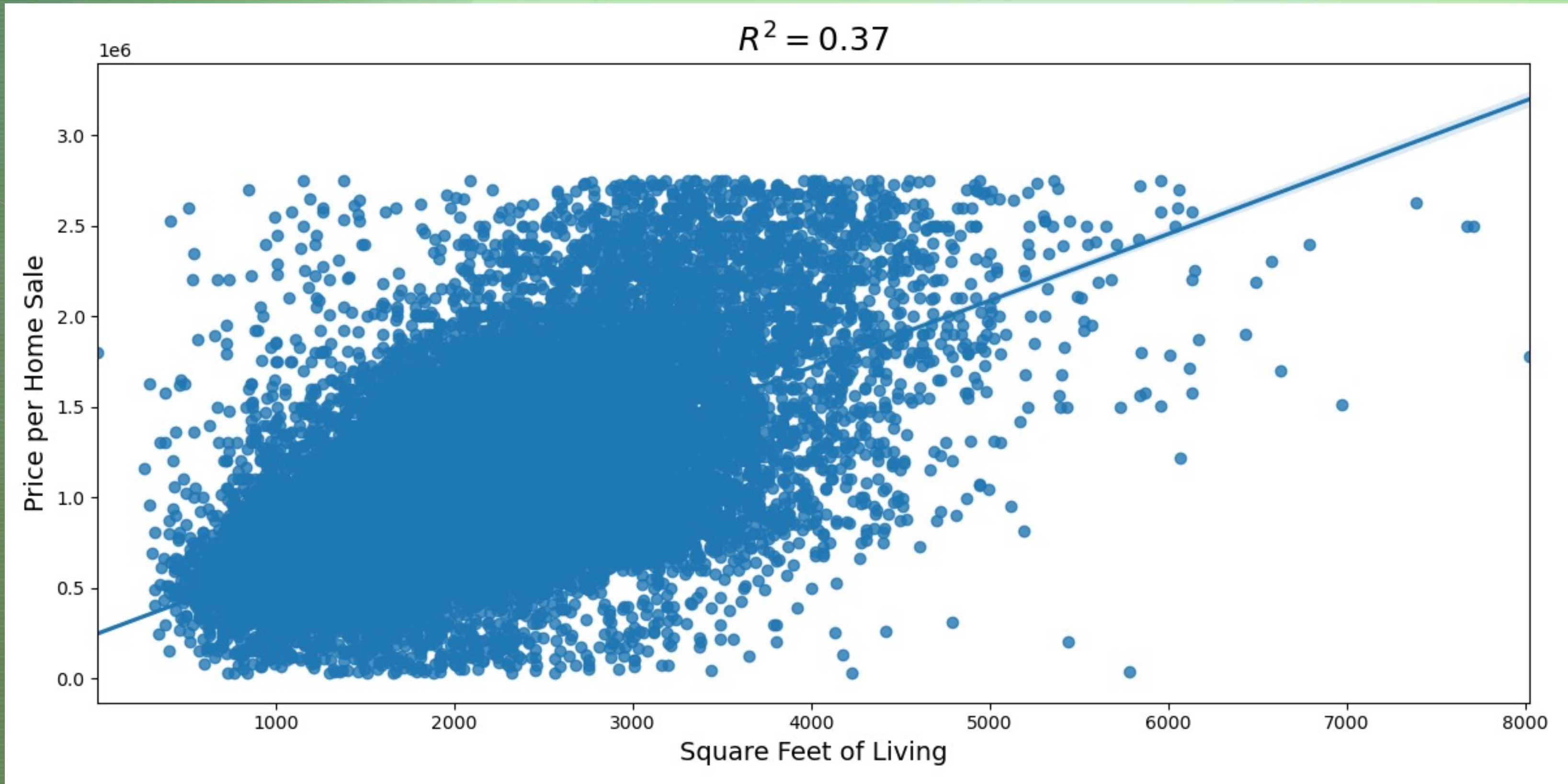
- The response variable shall be price.
- There are 25 predictor variables, all with different kinds of values and classifications

1. Number Bedrooms
2. Number Bathrooms
3. Square Feet Living Space
4. Square Feet Lot
5. Floors
6. Greenbelt
7. Nuisance
8. View
9. Condition
10. Grade
11. Heat Source
12. Square Footage (excl. Basement)
13. Square Footage Basement
14. Square Footage Garage
15. Square Footage Patio
16. Year Built
17. Year Renovated
18. Address
19. Lat
20. Long
21. Age (engineered)
22. Zipcode (engineered)
23. City (engineered)

Response

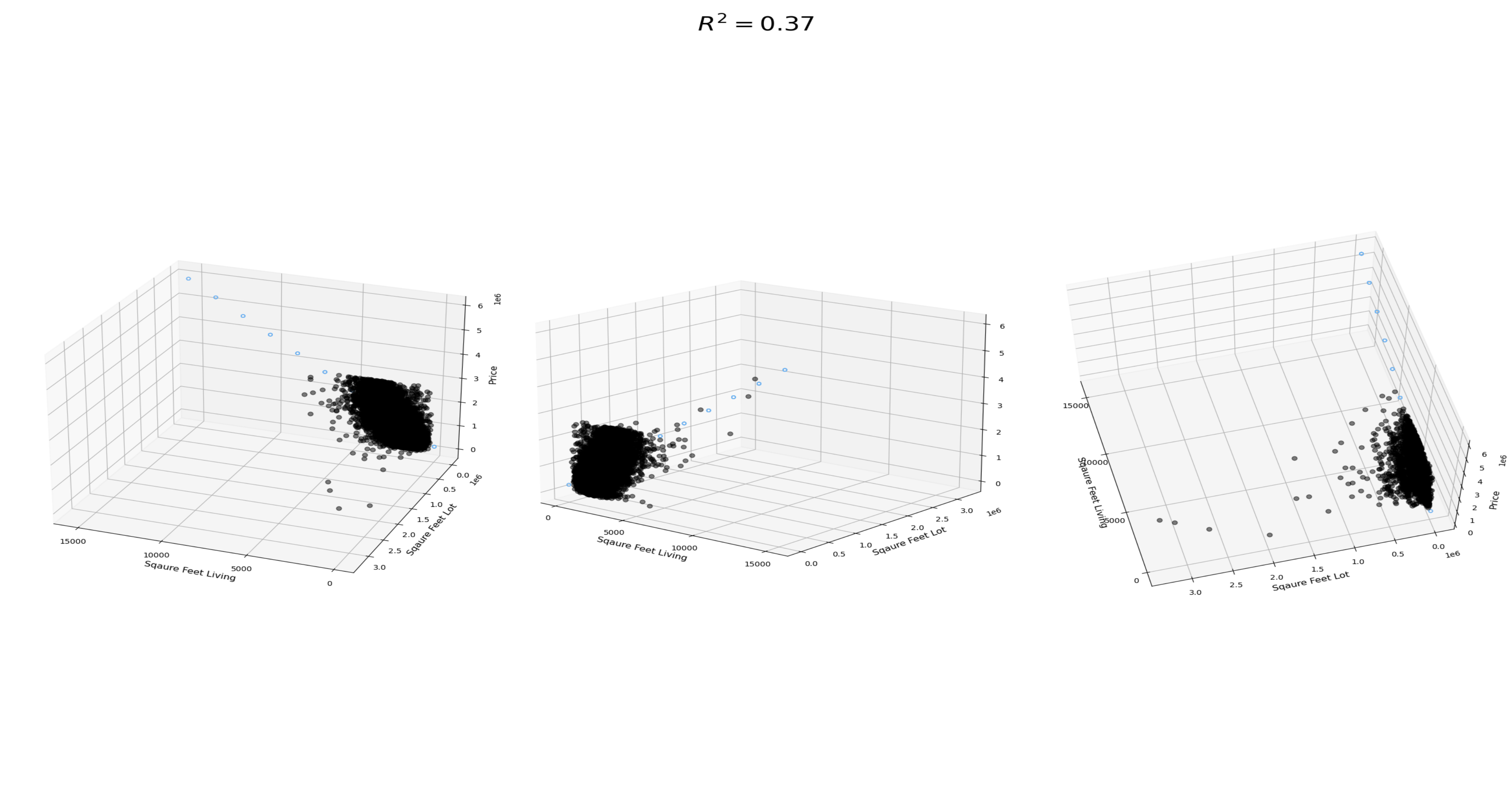

Prices Distribution
Number of Homes
Price
0
2000
4000
6000
8000
0.0
0.5
1.0
1.5
2.0
2.5
1e6

Baseline



$R^2 = 0.37$
1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Price per Home Sale
1000
2000
3000
4000
5000
6000
7000
8000
Square Feet of Living

Model 1



Grade

Dep. Variable:	price	**R-squared:**	0.356
Model:	OLS	**Adj. R-squared:**	0.355
Method:	Least Squares	**F-statistic:**	1062.
Date:	Sun, 02 Oct 2022	**Prob (F-statistic):**	0.00
Time:	15:25:53	**Log-Likelihood:**	-2.7653e+05
No. Observations:	19268	**AIC:**	5.531e+05
Df Residuals:	19257	**BIC:**	5.532e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.001e+06	3.87e+04	25.890	0.000	9.25e+05	1.08e+06
grade_2 Substandard	-1.2e-09	3.55e-10	-3.384	0.001	-1.9e-09	-5.05e-10
grade_3 Poor	-4.608e+05	1.59e+05	-2.899	0.004	-7.72e+05	-1.49e+05
grade_4 Low	-4.397e+05	7.91e+04	-5.562	0.000	-5.95e+05	-2.85e+05
grade_5 Fair	-4.169e+05	4.58e+04	-9.099	0.000	-5.07e+05	-3.27e+05
grade_6 Low Average	-3.58e+05	3.97e+04	-9.012	0.000	-4.36e+05	-2.8e+05
grade_7 Average	-1.613e+05	3.89e+04	-4.146	0.000	-2.38e+05	-8.5e+04
grade_8 Good	8.218e+04	3.9e+04	2.109	0.035	5805.256	1.59e+05
grade_9 Better	4.886e+05	3.94e+04	12.399	0.000	4.11e+05	5.66e+05
grade_10 Very Good	8.563e+05	4.11e+04	20.834	0.000	7.76e+05	9.37e+05
grade_11 Excellent	1.08e+06	5.08e+04	21.244	0.000	9.8e+05	1.18e+06
grade_12 Luxury	1.292e+06	8.77e+04	14.726	0.000	1.12e+06	1.46e+06
grade_13 Mansion	-9.612e+05	3.8e+05	-2.532	0.011	-1.71e+06	-2.17e+05

Omnibus:	1728.213	**Durbin-Watson:**	2.044
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2594.997
Skew:	0.697	**Prob(JB):**	0.00

Model 9

```python
Y9 = df_modeling['price']
X9 = df_modeling[['sqft_living',
                  'grade_8 Good',
                  'grade_9 Better',
                  'grade_10 Very Good',
                  'grade_11 Excellent',
                  'grade_12 Luxury',
                  'bathrooms',
                  'condition_Poor',
                  'condition_Fair',
                  'condition_Average',
                  'condition_Good',
                  'condition_Very Good',
                  'floors_1.0',
                  'floors_1.5',
                  'floors_2.0',
                  'floors_2.5',
                  'floors_3.0',
                  'floors_3.5',
                  'floors_4.0',
                  'city_Auburn',
                  'city_Bellevue',
                  'city_Black Diamond',
                  'city_Bothell',
                  'city_Enumclaw',
                  'city_Fall City',
                  'city_Preston',
                  'city_Ravensdale',
                  'city_Redmond',
                  'city_Renton',
                  'city_Sammamish',
                  'city_Seattle',
                  'city_Skykomish',
                  'city_Snoqualmie',
                  'city_Woodinville',
                  'sqft_above',
                  'sqft_lot',
                  'yr_built',
                  'sqft_patio',
                  'sqft_garage',
                  'age',
                  ]
                 ]
```

Dep. Variable:	price	**R-squared:**	0.599
Model:	OLS	**Adj. R-squared:**	0.598
Method:	Least Squares	**F-statistic:**	756.3
Date:	Sun, 02 Oct 2022	**Prob (F-statistic):**	0.00
Time:	15:26:00	**Log-Likelihood:**	-2.7195e+05
No. Observations:	19268	**AIC:**	5.440e+05
Df Residuals:	19229	**BIC:**	5.443e+05
Df Model:	38		
Covariance Type:	nonrobust		

RFE

- Post Model 9, we use Recursive Feature Engineering (RFE) using Decision Tree Algorithm: Random Forest.
- We called a list of the ranked best to worst predictive features based on backward elimination of a training and test set that will be modeled on:

```
1 sqft_living
2 lat
3 long
4 sqft_lot
5 sqft_above
6 age
7 grade_7 Average
8 sell_year
9 sqft_patio
10 grade_8 Good
11 sqft_garage
12 yr_built
13 sqft_basement
14 view_EXCELLENT
15 bathrooms
```

```
15 bathrooms
16 grade_9 Better
17 condition_Average
18 view_NONE
19 bedrooms
20 heat_source_Gas
21 city_Bellevue
22 condition_Very Good
23 grade_6 Low Average
24 yr_renovated
25 grade_10 Very Good
26 heat_source_Oil
27 floors
28 grade_13 Mansion
29 view_GOOD
30 view_AVERAGE
```

Final Model

```
X_final = df_modeling[
                      [
                       'sqft_living',
                       'sqft_lot',
                       'sqft_above',
                       'sqft_patio',
                       'lat',
                       'long',
                       'sell_year',
                       'age',
                       'grade_7 Average',
                       'grade_8 Good',
                       'grade_9 Better',
                       'grade_10 Very Good',
                       'sqft_garage',
                       'yr_built',
                       'view_EXCELLENT',
                       'sqft_basement',
                       'condition_Average',
                       'condition_Good',
                       'condition_Very Good',
                       'view_NONE',
                       'heat_source_Oil',
                       'heat_source_Gas',
                       'city_Bellevue',
                       'city_Seattle'
                       ]
                      ]
```

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.652
Model:	OLS	**Adj. R-squared:**	0.651
Method:	Least Squares	**F-statistic:**	1566.
Date:	Sun, 02 Oct 2022	**Prob (F-statistic):**	0.00
Time:	15:32:43	**Log-Likelihood:**	-2.7060e+05
No. Observations:	19268	**AIC:**	5.412e+05
Df Residuals:	19244	**BIC:**	5.414e+05
Df Model:	23		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-3.607e+08	9.74e+06	-37.021	0.000	-3.8e+08	-3.42e+08
sqft_living	172.2158	9.935	17.334	0.000	152.742	191.690
sqft_lot	0.3488	0.036	9.600	0.000	0.278	0.420
sqft_above	116.8426	10.561	11.063	0.000	96.141	137.544
sqft_patio	46.4693	10.323	4.502	0.000	26.236	66.703
lat	1.472e+06	1.75e+04	84.057	0.000	1.44e+06	1.51e+06
long	1.924e+04	2.1e+04	0.916	0.360	-2.19e+04	6.04e+04
sell_year	9.715e+04	3063.401	31.712	0.000	9.11e+04	1.03e+05
age	4.921e+04	1530.886	32.142	0.000	4.62e+04	5.22e+04
grade_7 Average	-1.535e+04	7823.145	-1.962	0.050	-3.07e+04	-14.237
grade_8 Good	9.19e+04	9025.565	10.182	0.000	7.42e+04	1.1e+05
grade_9 Better	2.858e+05	1.14e+04	25.123	0.000	2.64e+05	3.08e+05
grade_10 Very Good	4.305e+05	1.58e+04	27.261	0.000	4e+05	4.61e+05

Final Model (cont.)

sqft_garage	18.0695	11.248	1.607	0.108	-3.977	40.116
yr_built	4.794e+04	1534.180	31.249	0.000	4.49e+04	5.09e+04
view_EXCELLENT	2.734e+05	2.18e+04	12.559	0.000	2.31e+05	3.16e+05
sqft_basement	27.8161	8.125	3.424	0.001	11.891	43.742
condition_Average	6.955e+04	2.31e+04	3.011	0.003	2.43e+04	1.15e+05
condition_Good	1.242e+05	2.32e+04	5.355	0.000	7.87e+04	1.7e+05
condition_Very Good	2.021e+05	2.38e+04	8.500	0.000	1.56e+05	2.49e+05
view_NONE	-1.274e+05	7738.175	-16.465	0.000	-1.43e+05	-1.12e+05
heat_source_Oil	-5445.3391	9168.220	-0.594	0.553	-2.34e+04	1.25e+04
heat_source_Gas	4.198e+04	5812.279	7.222	0.000	3.06e+04	5.34e+04
city_Bellevue	3.324e+05	1.05e+04	31.654	0.000	3.12e+05	3.53e+05
city_Seattle	-1.694e+04	7631.152	-2.219	0.026	-3.19e+04	-1977.957

$R^2 = 0.652$

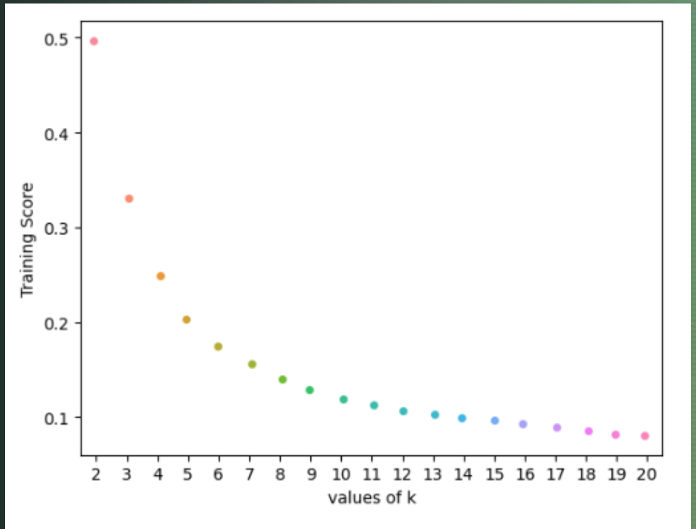
p-value: all less than threshold. (Heat_source_Oil observed at Threshold 0.05)

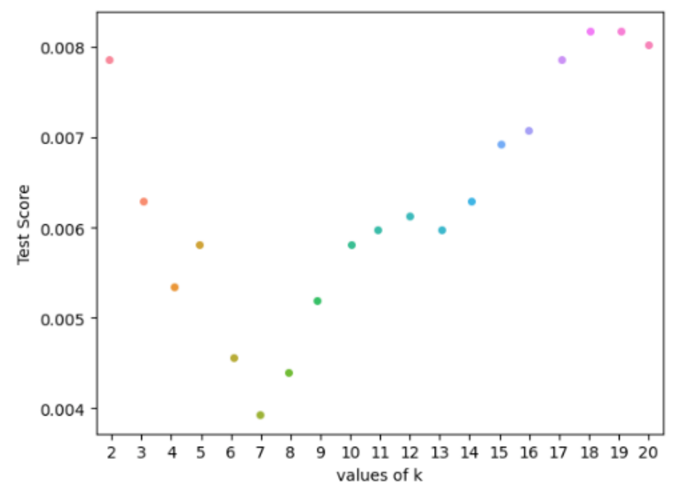
RMSE = 304211

Determinants

- Final Model 9 – RFE Results Deemed Recursive or otherwise unnecessary by feature ranking

k-NN





- We recall the R^2 and RMSE for the **final model** observed was 0.652 & 304211.0600187805.

- The **k-NN algorithm** predictions produce a higher R^2 of 0.685 & a lower RMSE of 282204.3965664784.

We can conclude: the **k-NN** algorithm performs slightly better than our final model when the number of nearest neighbors is set to 20.

Findings

The the R^2 and
RMSE for the final model observed was

- R^2: 0.652
- RMSE: 304211.0600187805

<u>For our usecase, this model seems to be sufficient.</u>

The final model was then used against a trained and tested set using Machine Learning algorithm, **k-NN**; which is Nearest Neighbors algorithms.

We found that:
•the k-NN algorithm performs slightly better than our final model when the
•number of nearest neighbors is set to **20**.

- R^2: 0.685
- RMSE: 304211.0600187805

Recommendations

1. With the final model, we can estimate existing home sales records to form the basis for our classifications for tax revenue

2. In addition, our Home Facilities Stakeholder has a reliable model they can use to help them understand home sale price

3. The k-NN model can help predict the trajectory of future home sale prices

4. This model achieves the objectives of King County:
   - Build a model that provides data on home sale prices
   - and recorded dimensions of homes in the sale.
   - Filter out and clean a DataFrame that included homes
   - located in counties other than KC.
   - Prepare and execute an iterative modeling process with explanations of coefficients.
   - Through EDA and research, information on dimensions about homes in KC.
   - Investigates 'grade' for Home Facilities Management stakeholder.

Further Investigation

- We recommend maybe incorporating webscraped topographical data to engineer new features
- There are many powerful libraries that handle geolocational data structures
- Time-Series Forecasting can be used to help US Census Bureau estimate population growth