alpacanonymous / **project2** Public

Flatiron School Project 2: Predicting House Sale Prices in King County, WA

0 stars 0 forks

Star Unwatch

Code Issues Pull requests Actions Projects Wiki Security

main

alpacanonymous pdf folder ... 1 minute ago 2

View code

README.md

Linear Regression Modeling of King County Real Estate Sale Prices

Authors: Aisha Baitemirova-Othman, Angela Kim, Steven Addison, Wahaj Dar
Instructor: David Elliott

Overview

This project analyzes residential real estate sales in King County, Washington, and uses the data to create a model that predicts price based on the parameters given.

Business Problem

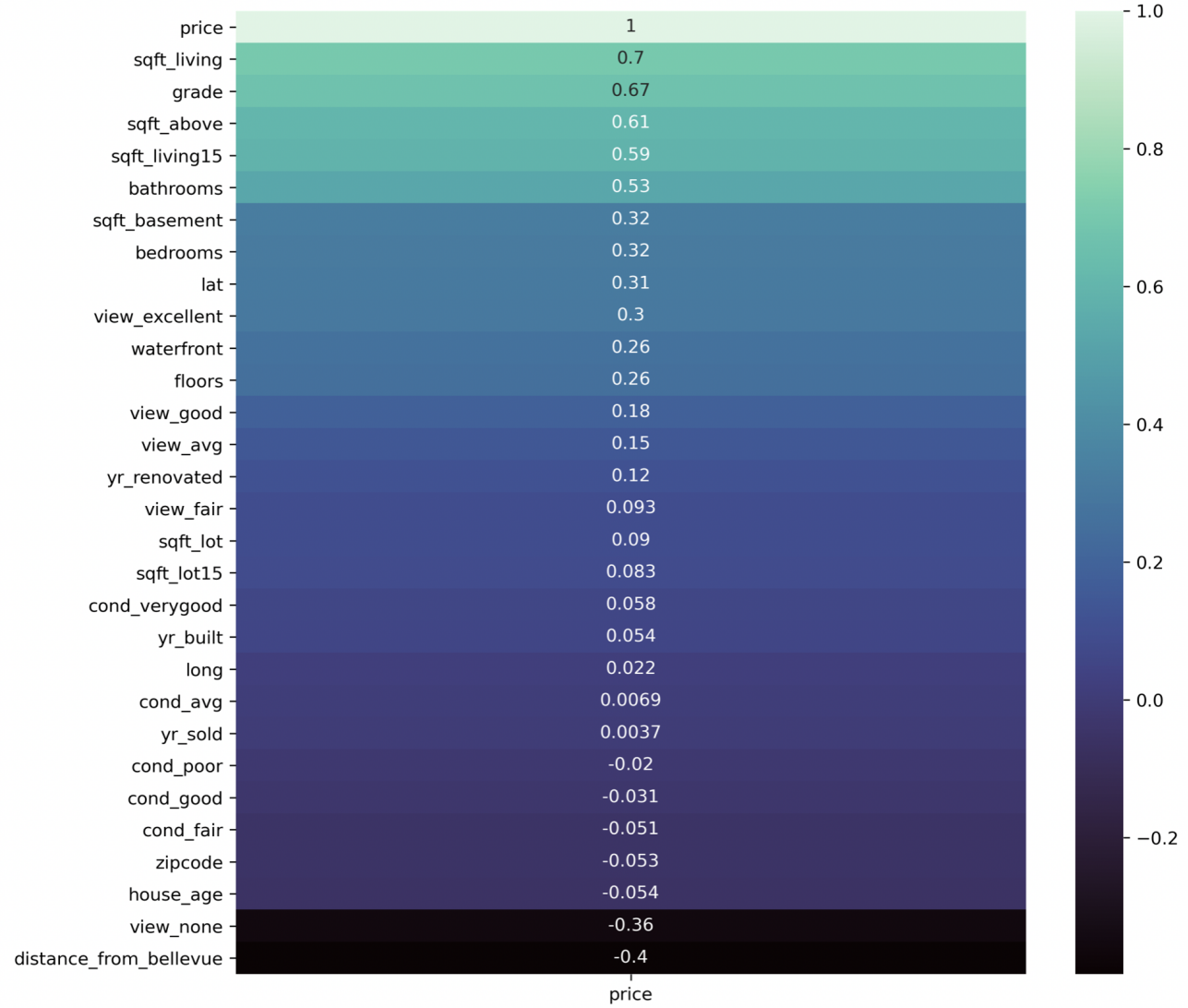
Windermere Real Estate, based in Seattle, Washington, wants to better serve home buyers by being able to accurately present a price point using features of a house (ie. number of bedrooms) that buyers are looking for.

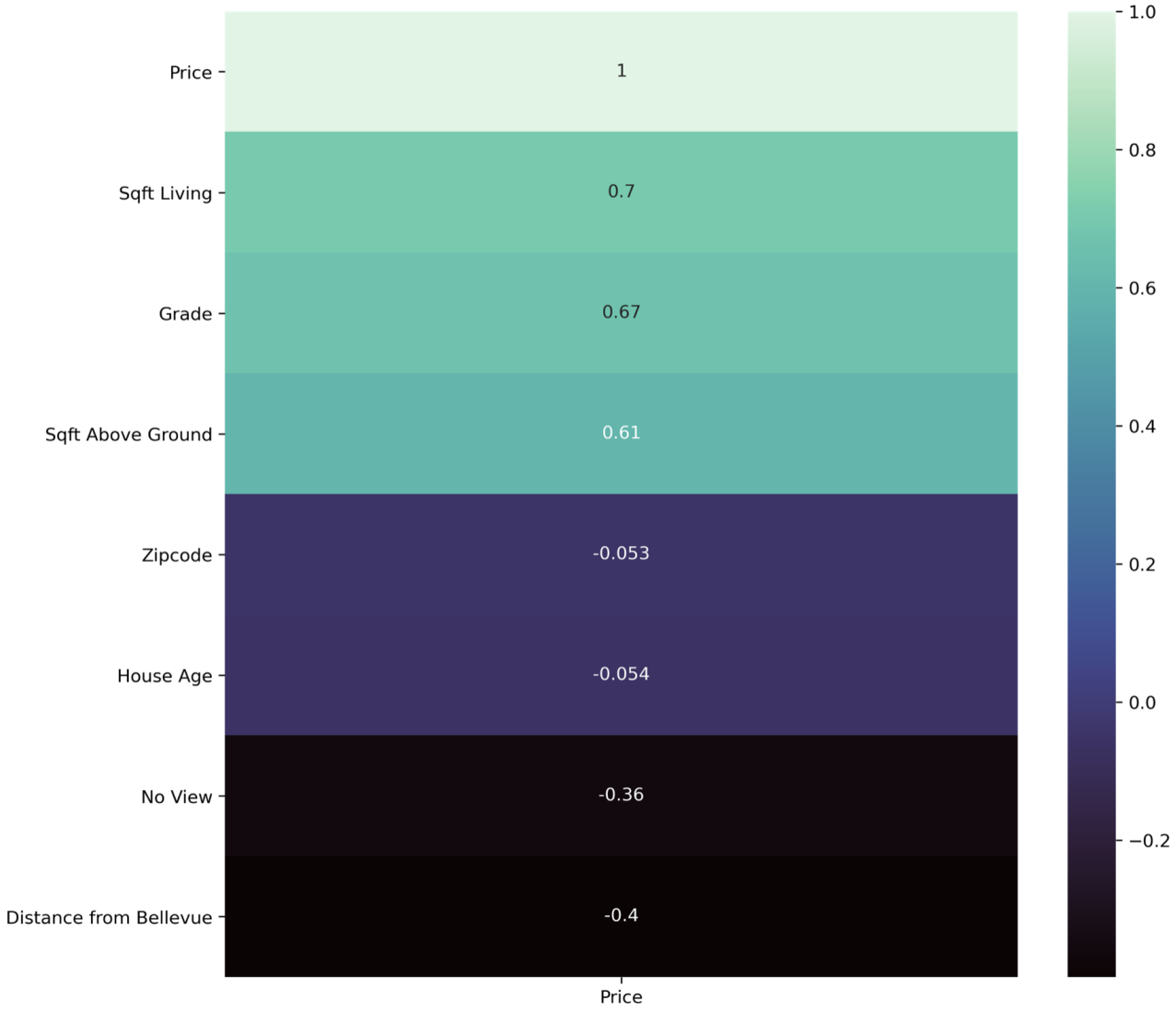
Data Understanding

This dataset contains information about residential real estate sales in King County between May 2014 - May 2015. It includes details such as number of bedrooms and bathrooms, square footage of the home, and various features regarding location.

Data Preparation & Analysis

Correlation between predictor variables and price:



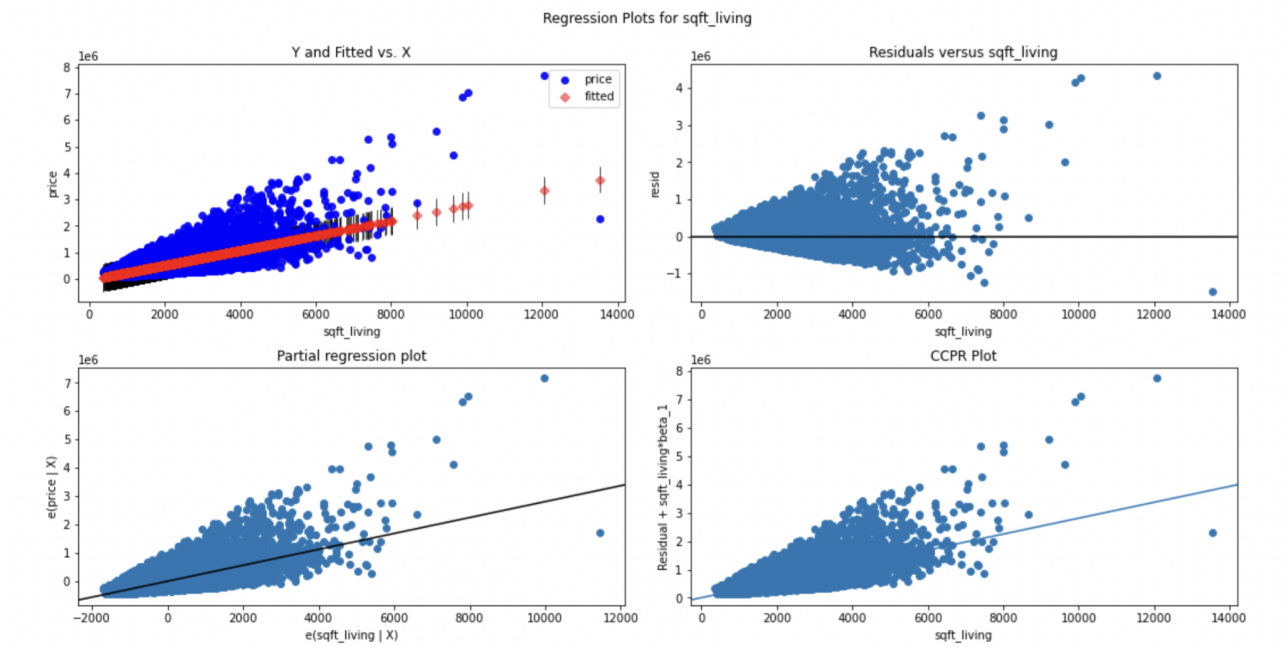


OLS Regression Results

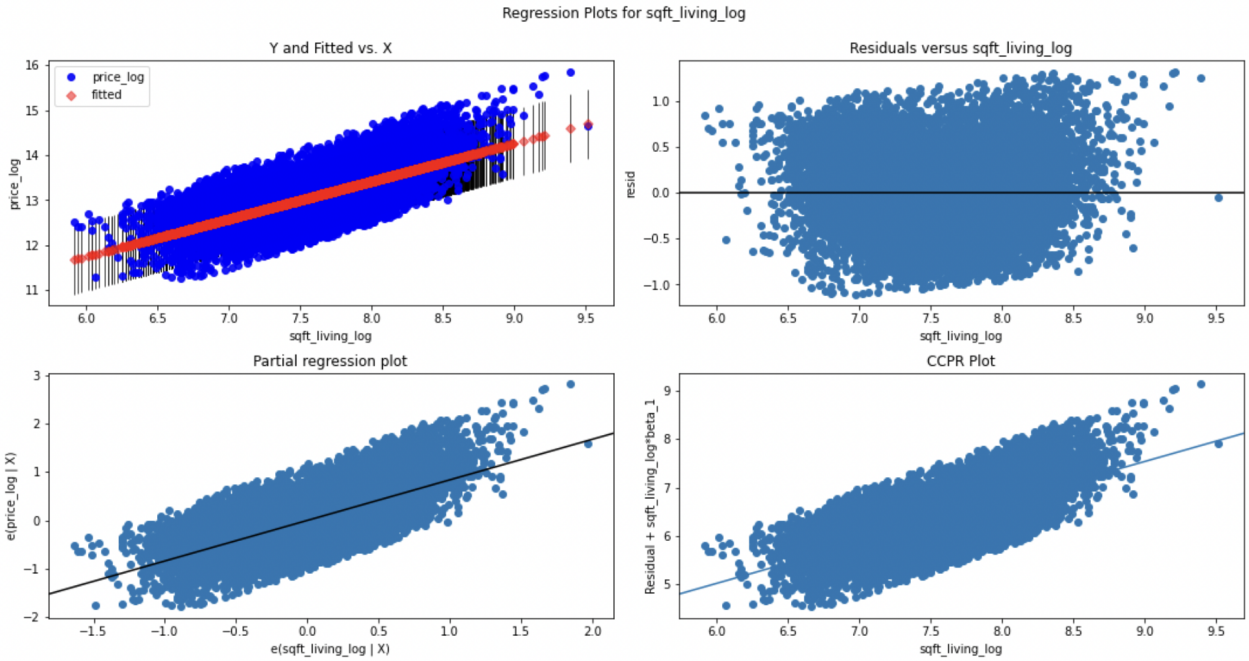
Dep. Variable:	price	**R-squared:**	0.725
Model:	OLS	**Adj. R-squared:**	0.724
Method:	Least Squares	**F-statistic:**	2270.
Date:	Fri, 19 Nov 2021	**Prob (F-statistic):**	0.00
Time:	08:15:18	**Log-Likelihood:**	-2.9347e+05
No. Observations:	21597	**AIC:**	5.870e+05
Df Residuals:	21571	**BIC:**	5.872e+05
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.869e+07	4.53e+06	-6.328	0.000	-3.76e+07	-1.98e+07
bedrooms	-4.093e+04	1899.880	-21.544	0.000	-4.47e+04	-3.72e+04
bathrooms	4.038e+04	3139.238	12.864	0.000	3.42e+04	4.65e+04
sqft_living	111.4975	2.194	50.809	0.000	107.196	115.799
sqft_lot	0.2136	0.046	4.639	0.000	0.123	0.304
floors	2372.2405	3463.560	0.685	0.493	-4416.592	9161.073
waterfront	5.355e+05	1.96e+04	27.378	0.000	4.97e+05	5.74e+05
grade	8.844e+04	2087.403	42.369	0.000	8.43e+04	9.25e+04
sqft_above	78.3972	2.171	36.109	0.000	74.142	82.653
sqft_basement	33.1232	2.548	13.001	0.000	28.129	38.117
yr_built	8639.5763	939.648	9.194	0.000	6797.797	1.05e+04
yr_renovated	24.9298	3.824	6.520	0.000	17.435	32.425
zipcode	-507.6069	31.836	-15.944	0.000	-570.008	-445.206
lat	3.072e+05	1.26e+04	24.379	0.000	2.83e+05	3.32e+05
long	-1.492e+05	1.27e+04	-11.719	0.000	-1.74e+05	-1.24e+05
sqft_living15	8.8410	3.328	2.656	0.008	2.318	15.364
sqft_lot15	-0.1242	0.071	-1.759	0.079	-0.263	0.014
yr_sold	1.954e+04	1877.232	10.410	0.000	1.59e+04	2.32e+04
house_age	1.09e+04	938.899	11.611	0.000	9061.481	1.27e+04
cond_avg	-5.748e+06	9.07e+05	-6.337	0.000	-7.53e+06	-3.97e+06
cond_fair	-5.746e+06	9.07e+05	-6.334	0.000	-7.52e+06	-3.97e+06
cond_good	-5.725e+06	9.07e+05	-6.314	0.000	-7.5e+06	-3.95e+06
cond_poor	-5.785e+06	9.07e+05	-6.376	0.000	-7.56e+06	-4.01e+06
cond_verygood	-5.686e+06	9.07e+05	-6.272	0.000	-7.46e+06	-3.91e+06
view_avg	-5.792e+06	9.07e+05	-6.387	0.000	-7.57e+06	-4.01e+06
view_excellent	-5.552e+06	9.07e+05	-6.121	0.000	-7.33e+06	-3.77e+06
view_fair	-5.765e+06	9.07e+05	-6.356	0.000	-7.54e+06	-3.99e+06
view_good	-5.714e+06	9.07e+05	-6.300	0.000	-7.49e+06	-3.94e+06
view_none	-5.867e+06	9.07e+05	-6.472	0.000	-7.64e+06	-4.09e+06
distance_from_bellevue	-1.331e+04	328.984	-40.459	0.000	-1.4e+04	-1.27e+04

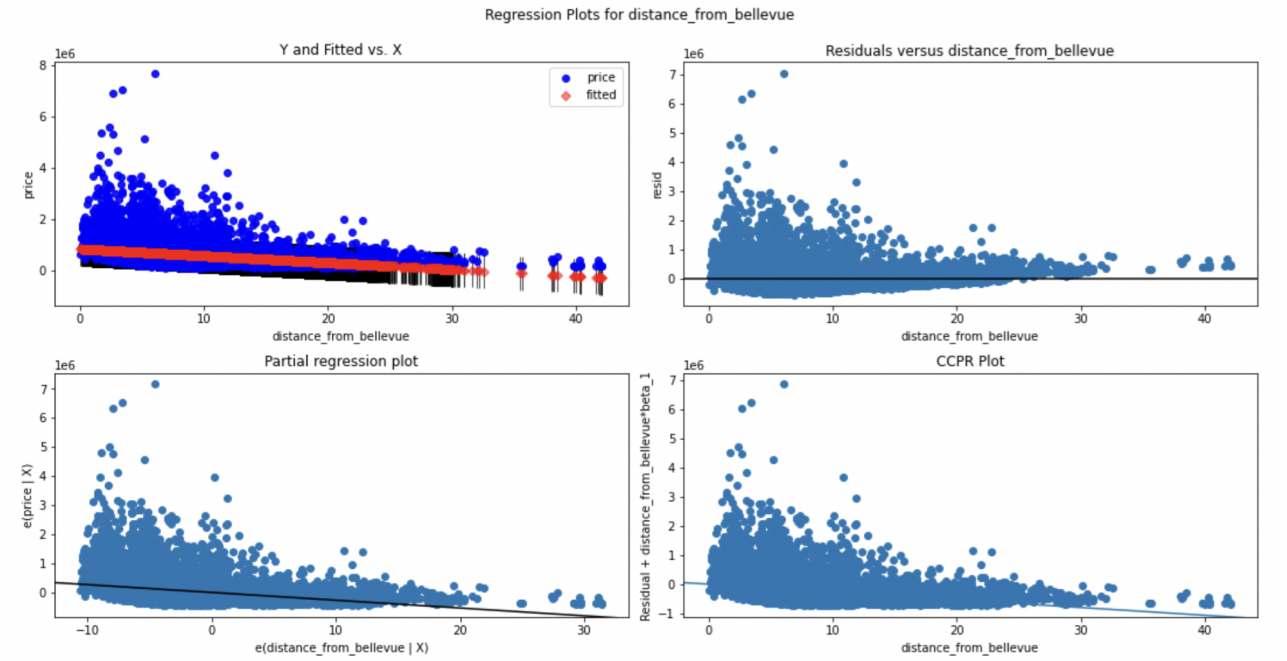
Omnibus:	19227.699	**Durbin-Watson:**	1.989
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2315048.410
Skew:	3.794	**Prob(JB):**	0.00
Kurtosis:	53.150	**Cond. No.**	1.01e+16



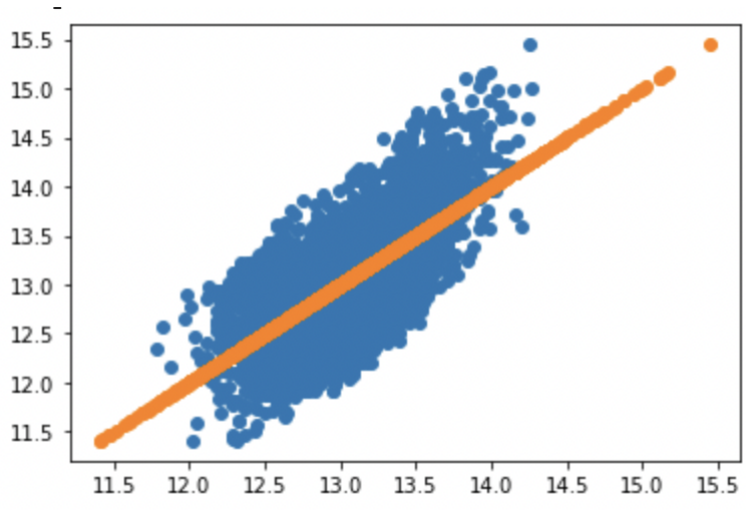
> Plots show heteroscedasticity.

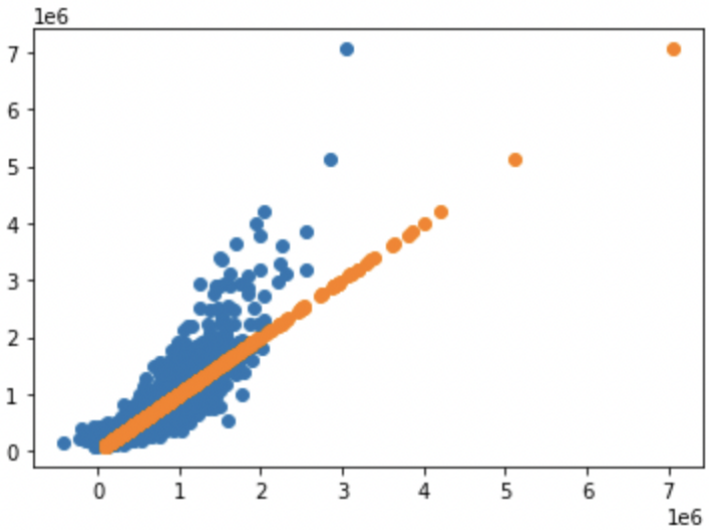


> When 'price' and 'sqft_living' undergo log transformation, they are more normally distributed and more homoscedastic, making them better for modeling.

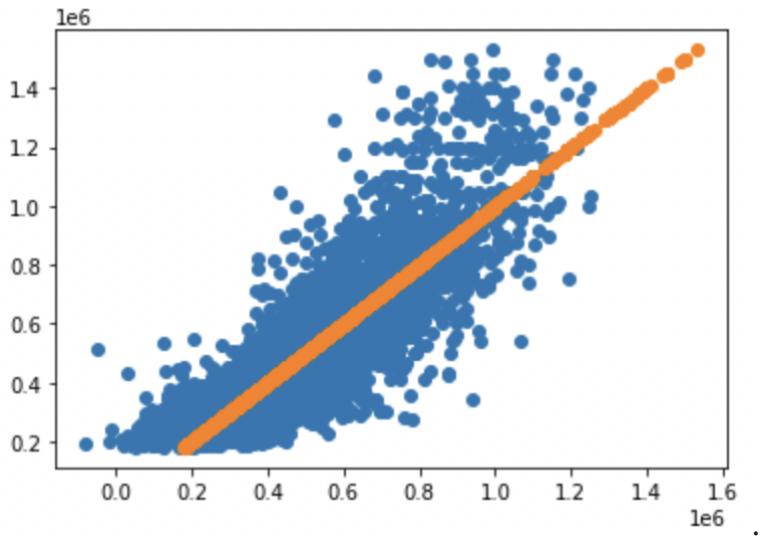


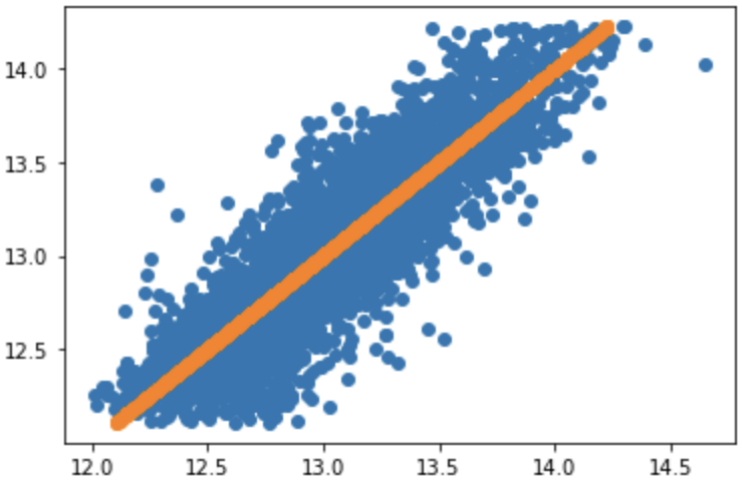
Simple Linear Regression Plot (left) and First Multiple Linear Regression Plot (right)



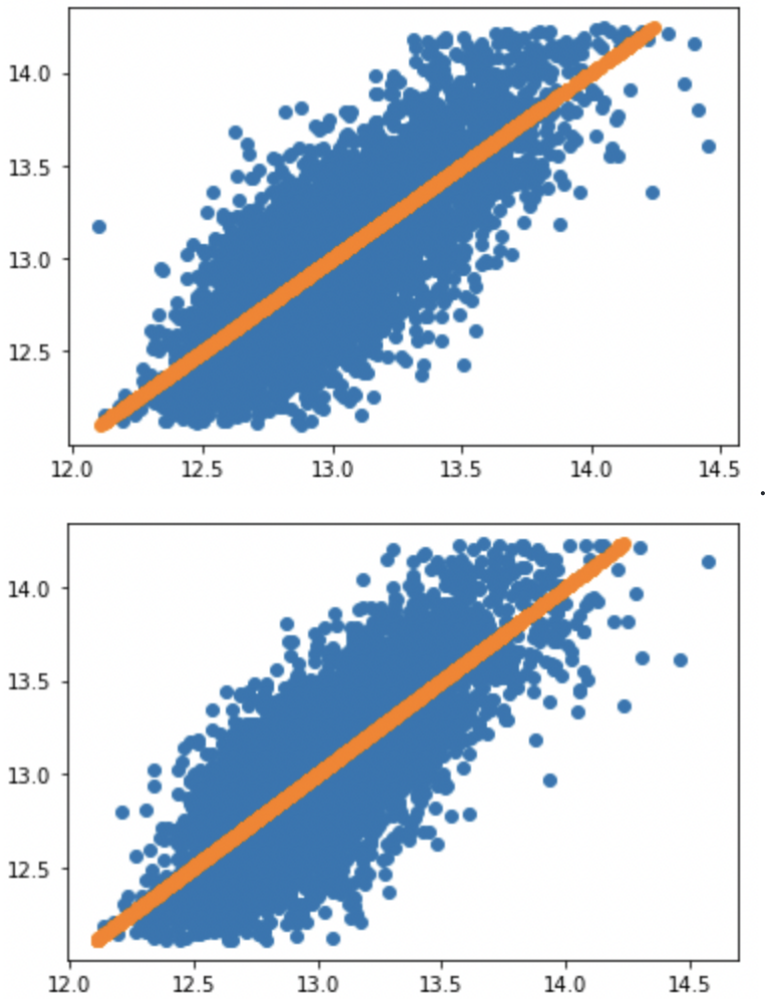


Second Multiple Linear Regression Plot (left) and Third Multiple Linear Regression Plot (right)

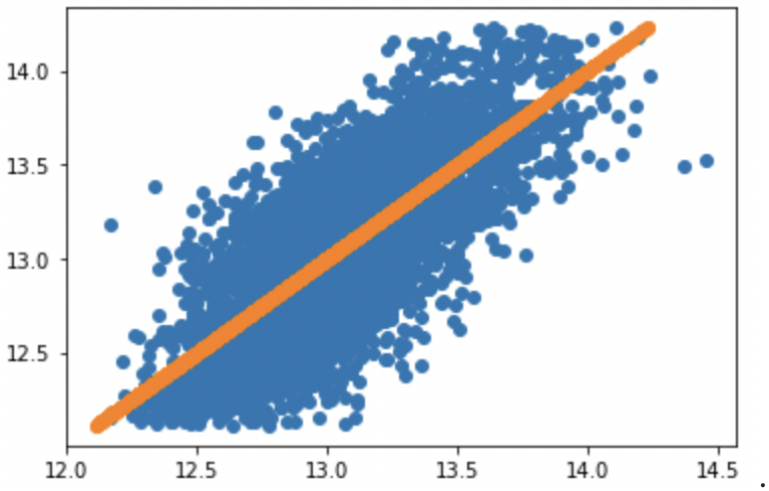


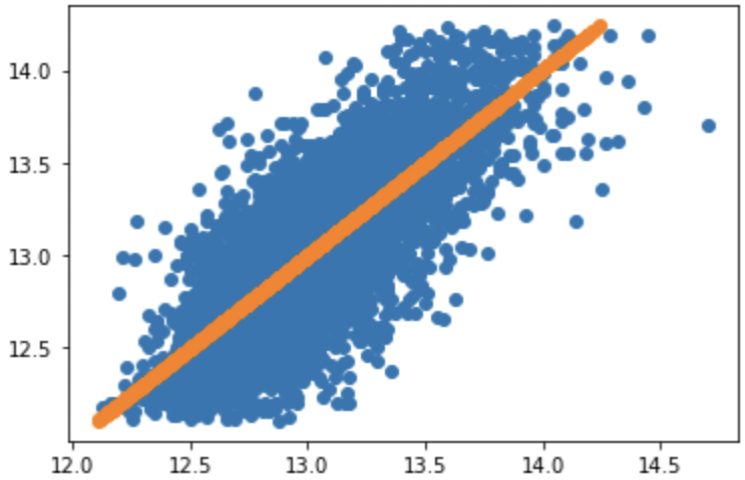


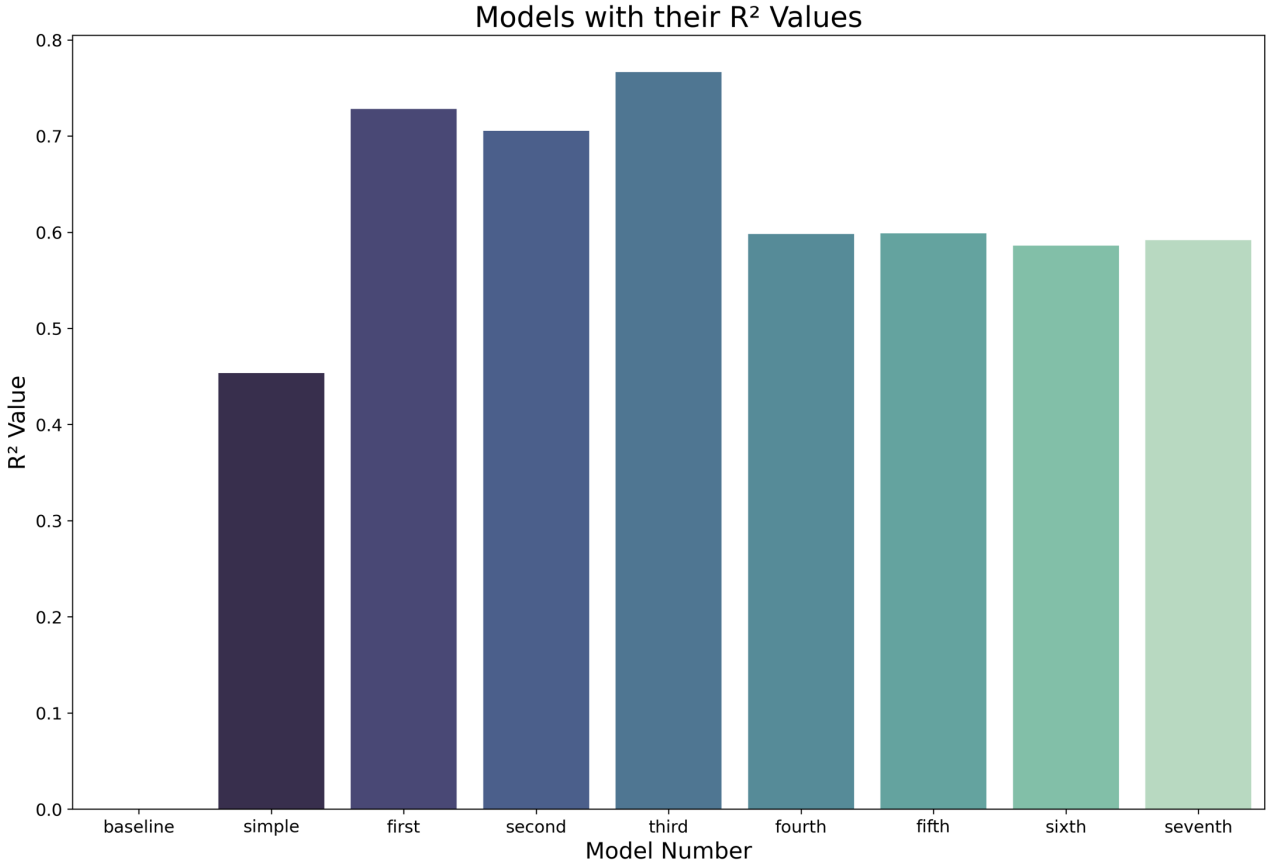
Fourth Multiple Linear Regression Plot (left) and Fifth Multiple Linear Regression Plot (right)

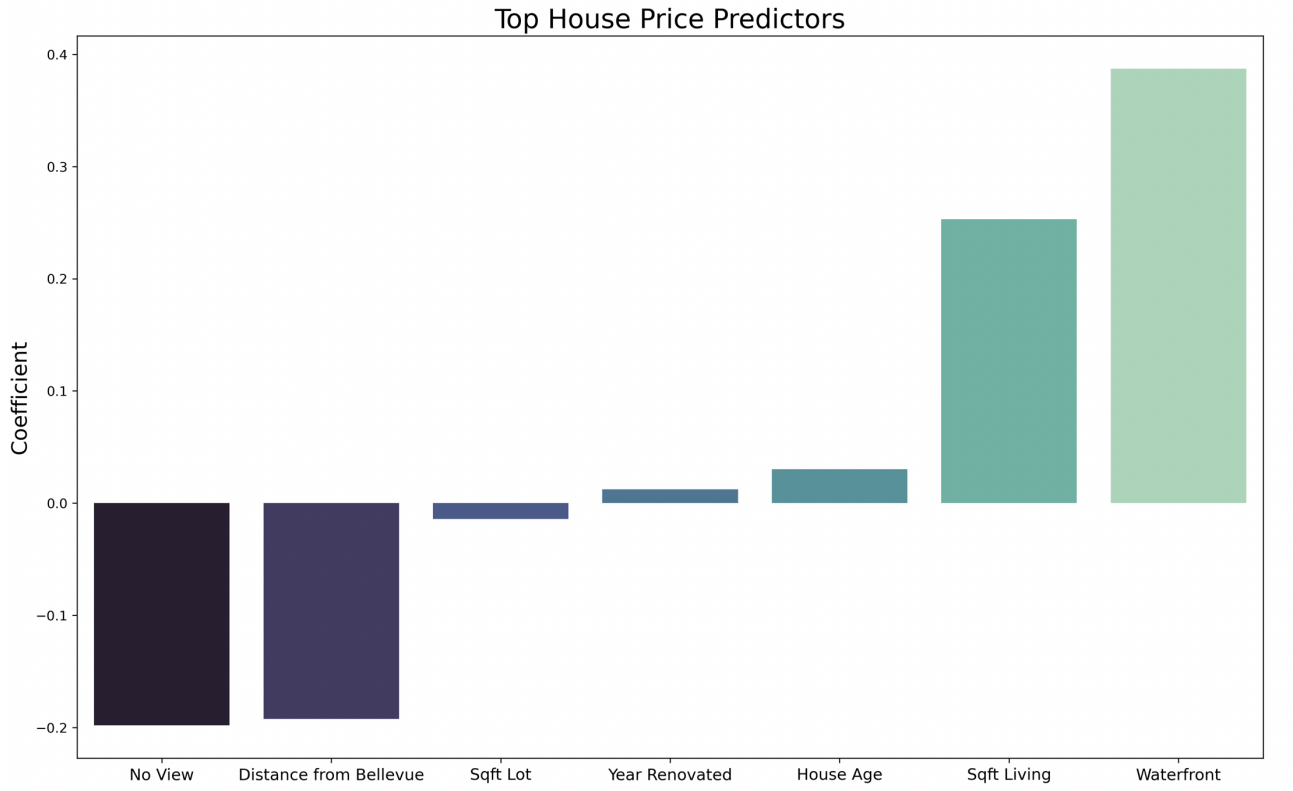


Sixth Multiple Linear Regression Plot (left) and Seventh Multiple Linear Regression Plot (right)


14.0
13.5
13.0
12.5
12.0
12.5
13.0
13.5
14.0
14.5


14.0
13.5
13.0
12.5
12.0
12.5
13.0
13.5
14.0
14.5


Models with their R² Values
R² Value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
baseline
simple
first
second
third
fourth
fifth
sixth
seventh
Model Number



Modeling

We start off with a baseline model using the highest correlated variable with price, which is sqft_living. We log transform price and sqft_living to get them normally distributed.

Baseline Model (DummyRegressor)

- Baseline Train R^2: 0.0
- Baseline Test R^2: -7.611977848931417e-06

Our next model uses simple linear regression with the price_log and sqft_living_log variables.

Simple Linear Regression

- Simple LR Train R^2: 0.4559935622464675
- Simple LR Test R^2: 0.4533592790543598
- Simple LR Train RMSE: 0.38932939001222455
- Simple LR Test RMSE: 0.3863726040140355
- Simple Condition Number: 136.8975981292544

The following models use multiple linear regression with different features to improve our initial models.

Multiple Linear Regression Model 1

- all untouched predictor variables without normalization or scaling
- LR1 Train R^2: 0.7227415083845596
- LR1 Test R^2: 0.7280296563595856
- LR1 Train RMSE: 191512.90263985636
- LR1 Test RMSE: 197177.0082445334
- LR1 Condition Number: 1.006084430986517e+16

Multiple Linear Regression Model 2

- price, sqft_living, and distance_from_bellevue outliers removed.
- LR2 Train R^2: 0.7262127515210657
- LR2 Test R^2: 0.7262127515210657
- LR2 Train RMSE: 125672.09779613405
- LR2 Test RMSE: 126388.69024537751
- LR2 Condition Number: 1.0035752027694244e+16

Multiple Linear Regression Model 3

- maintaining model 2 with price_log, sqft_living_log, and distance_from_bellevue_log
- LR3 Train R^2: 0.7710735650508587
- LR3 Test R^2: 0.7664798818166713
- LR3 Train RMSE: 0.21193655370980347
- LR3 Test RMSE: 0.20882679492683884
- LR3 Condition Number: 1.001320905302204e+16

Multiple Linear Regression Model 4

- LR4 Train R^2: 0.588175511728797
- LR4 Test R^2: 0.5981068463670351
- LR4 Train RMSE: 0.28170020490184783
- LR4 Test RMSE: 0.2815580293362851
- LR4 Condition Number: 9165.499808642742

Multiple Linear Regression Model 5

- maintaining fourth model with several predictor variables scaled
- LR5 Train R^2: 0.5879392642546251
- LR5 Test R^2: 0.5987762963115495
- LR5 Train RMSE: 0.2825149340905738
- LR5 Test RMSE: 0.27915393037643454
- LR5 Condition Number: 24.921890162559567

Multiple Linear Regression Model 6

- maintaining fifth model with only keeping selected columns provided by sklearn.feature_selection.RFE
- LR6 Train R^2: 0.5854459587429532
- LR6 Test R^2: 0.5858729777718898
- LR6 Train RMSE: 0.28363978204734774
- LR6 Test RMSE: 0.28282305780280415
- LR6 Condition Number: 24.091127817685116

Multiple Linear Regression Model 7

- using variables chosen by stepwise regression method
- LR7 Train R^2: 0.5903916264446546
- LR7 Test R^2: 0.5916737150909092
- LR7 Train RMSE: 0.2817047366292947
- LR7 Test RMSE: 0.281559353755675
- LR7 Condition Number: 23.730072313671045

Conclusions

After preparing the data, we made seven multiple linear regression models. Our final model was our best performing model with an R^2 value of 0.592, RMSE of 0.282, and Condition Number of 0.730. Our strongest predictor variables that will increase house prices are square footage of the house and whether the home is located on a waterfront. The strongest predictors that will decrease cost are homes with no view and being located farther from Bellevue.

Through multiple iterations of our model, we came to the conclusion that linear regression is not the best method to make a predictive model with this dataset. Linear regression is ill suited for a dataset with many categorical variables, as is the case with this dataset.

Our next steps would include gathering more data such as more recent home sales and expanding beyond single family homes into condos and apartments. We would also explore more complex modeling algorithms.

```
├── [data]
│    ├── column_names.md
│    └── kc_house_data.csv
├── [images]
├── [pdfs]
│    ├── github.pdf
│    ├── notebook.pdf
│    └── presentation.pdf
├── .gitignore
├── README.md
├── notebook.ipynb
└── presentation.pdf
```

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%