Multiple Regression on House

Presented by: Jiarong Chen

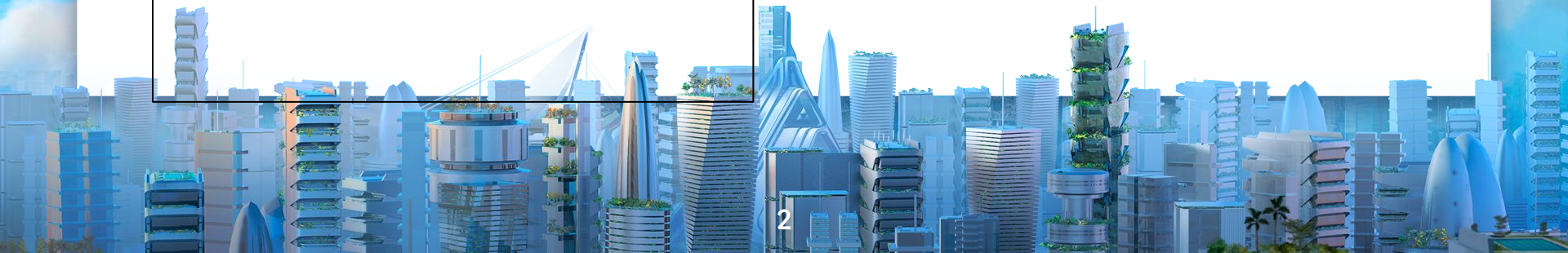
Background

Business problem

A house buyer assigns me a task about the house in King County. He wants to buy a house in this area but doesn't have any ideas about the housing market. And he has some preferred features in his mind, he wants to have a predicted price so that he can prepare for that.

My questions and plan

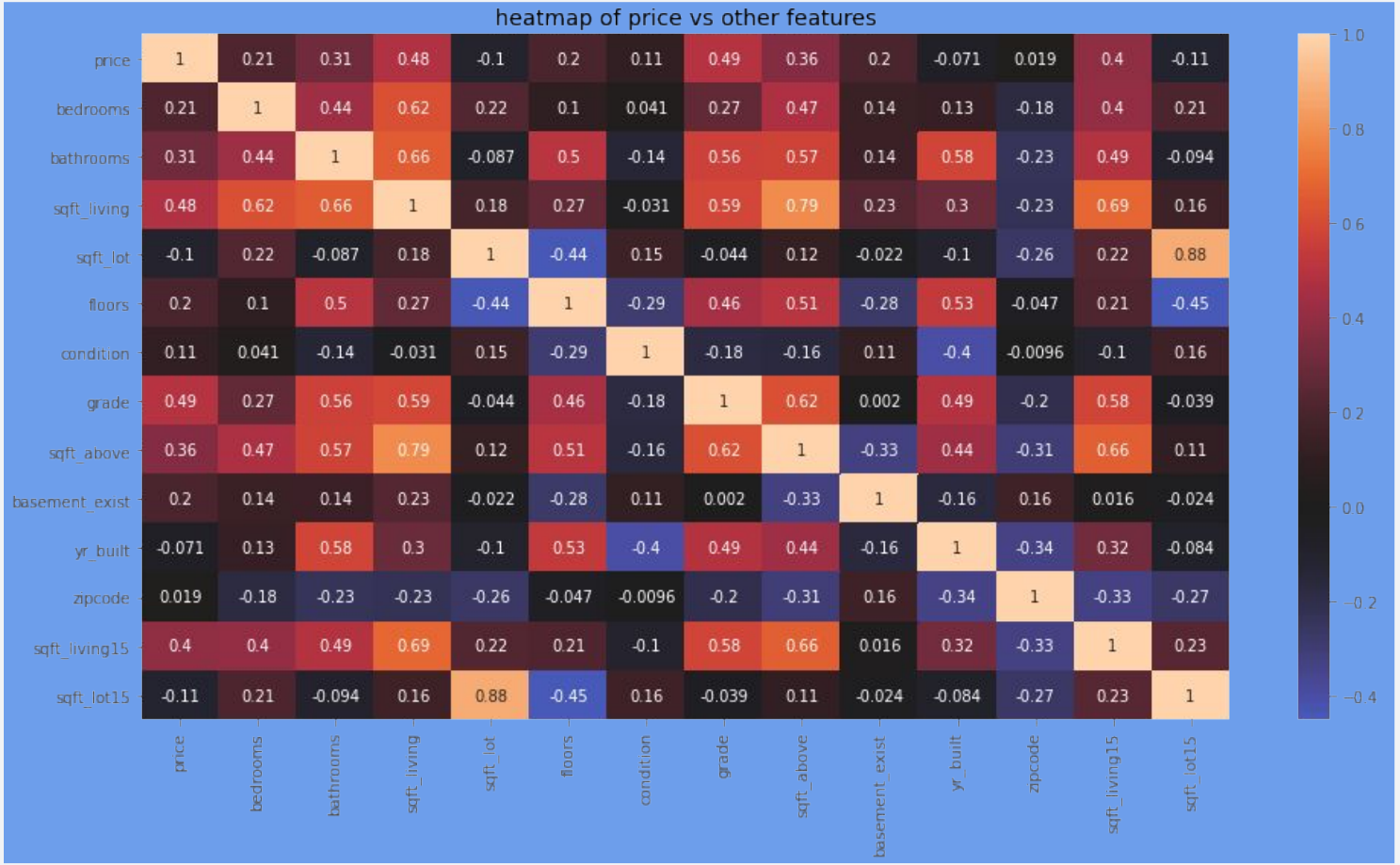
1. **What features does he need to concern about?**
   Find the most related features with the price.
2. **How the footage of the house(sqft_living) affect the price?**
   Find the correlation between them and the regression model.
3. **Which neighborhood is better to invest ?**
   Find the neighborhood by grade, price, year built sorting
4. **How much should he prepare for the dream house?**
   Find the prediction of price with model.



1.
Most Related Features

a. Visualization

b. Summary


heatmap of price vs other features
price 1 0.21 0.31 0.48 -0.1 0.2 0.11 0.49 0.36 0.2 -0.071 0.019 0.4 -0.11
bedrooms 0.21 1 0.44 0.62 0.22 0.1 0.041 0.27 0.47 0.14 0.13 -0.18 0.4 0.21
bathrooms 0.31 0.44 1 0.66 -0.087 0.5 -0.14 0.56 0.57 0.14 0.58 -0.23 0.49 -0.094
sqft_living 0.48 0.62 0.66 1 0.18 0.27 -0.031 0.59 0.79 0.23 0.3 -0.23 0.69 0.16
sqft_lot -0.1 0.22 -0.087 0.18 1 -0.44 0.15 -0.044 0.12 -0.022 -0.1 -0.26 0.22 0.88
floors 0.2 0.1 0.5 0.27 -0.44 1 -0.29 0.46 0.51 -0.28 0.53 -0.047 0.21 -0.45
condition 0.11 0.041 -0.14 -0.031 0.15 -0.29 1 -0.18 -0.16 0.11 -0.4 -0.0096 -0.1 0.16
grade 0.49 0.27 0.56 0.59 -0.044 0.46 -0.18 1 0.62 0.002 0.49 -0.2 0.58 -0.039
sqft_above 0.36 0.47 0.57 0.79 0.12 0.51 -0.16 0.62 1 -0.33 0.44 -0.31 0.66 0.11
basement_exist 0.2 0.14 0.14 0.23 -0.022 -0.28 0.11 0.002 -0.33 1 -0.16 0.16 0.016 -0.024
yr_built -0.071 0.13 0.58 0.3 -0.1 0.53 -0.4 0.49 0.44 -0.16 1 -0.34 0.32 -0.084
zipcode 0.019 -0.18 -0.23 -0.23 -0.26 -0.047 -0.0096 -0.2 -0.31 0.16 -0.34 1 -0.33 -0.27
sqft_living15 0.4 0.4 0.49 0.69 0.22 0.21 -0.1 0.58 0.66 0.016 0.32 -0.33 1 0.23
sqft_lot15 -0.11 0.21 -0.094 0.16 0.88 -0.45 0.16 -0.039 0.11 -0.024 -0.084 -0.27 0.23 1
price
bedrooms
bathrooms
sqft_living
sqft_lot
floors
condition
grade
sqft_above
basement_exist
yr_built
zipcode
sqft_living15
sqft_lot15
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Summary

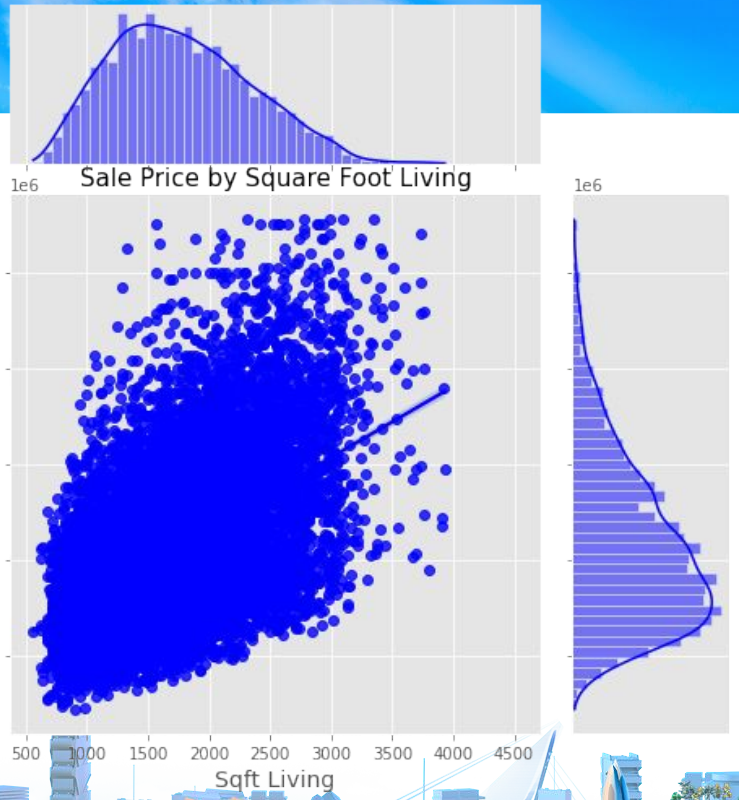
```
[('price', 1.0),
 ('grade', 0.49261027416442066),
 ('sqft_living', 0.48222641077964523),
 ('sqft_living15', 0.3979299346232877),
 ('sqft_above', 0.36084369884935624),
 ('bathrooms', 0.3112049323614764),
 ('bedrooms', 0.21432266736607247),
 ('floors', 0.20297606726310086),
 ('basement_exist', 0.19564092494704588),
 ('condition', 0.10759942018093975),
 ('zipcode', 0.019162250231750035),
 ('yr_built', -0.07066375804244265),
 ('sqft_lot', -0.10333797300264337),
 ('sqft_lot15', -0.1146877232085822)]
```

Top 5 Related Features

grade,
sqft_living,
sqft_living15,
sqft_above,
bathrooms

2.
Home Size Effect

Find the correlation between price and sqft_living and make the regression model.



correlation coefficient:
0.482226

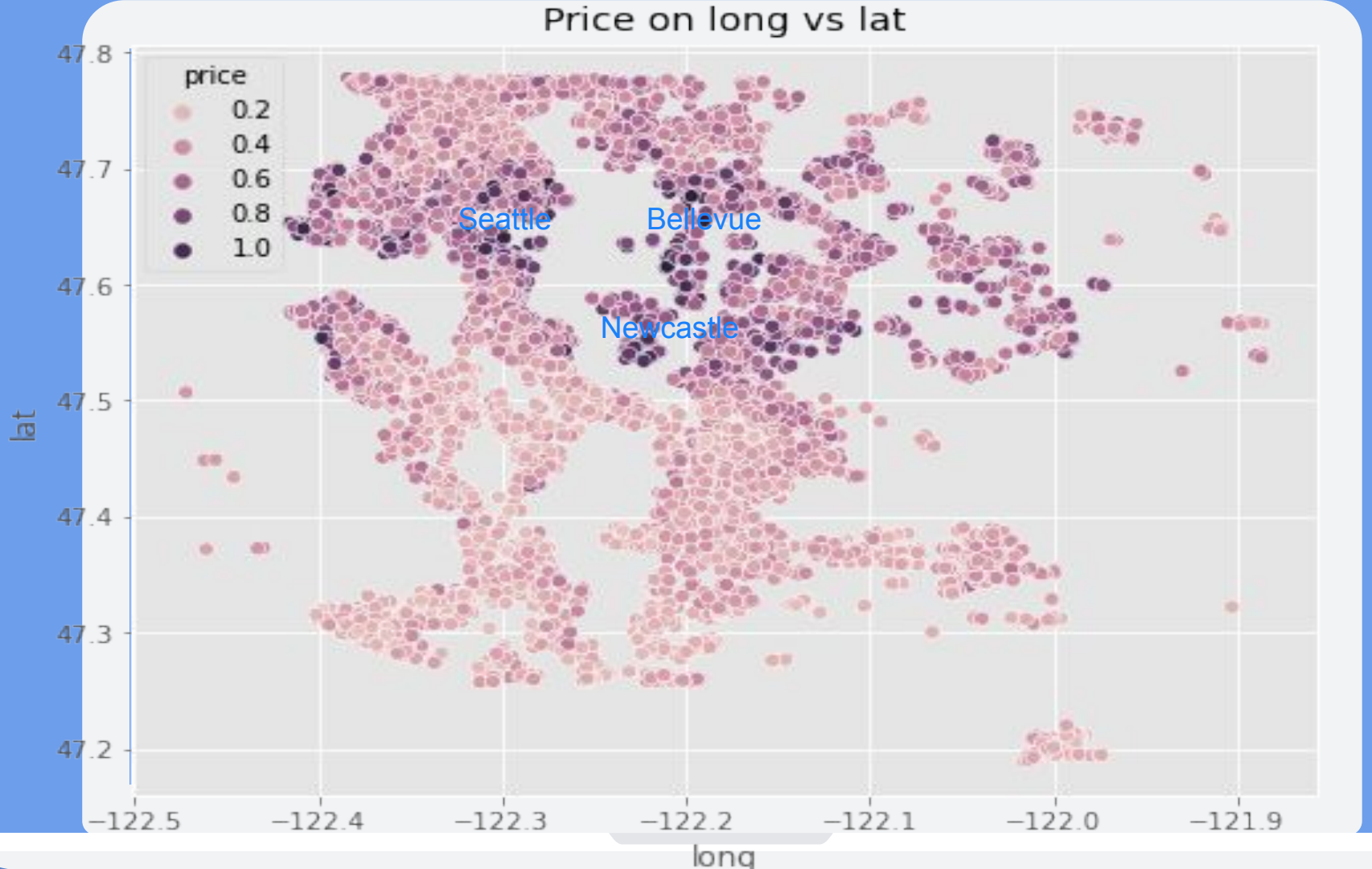
Summary

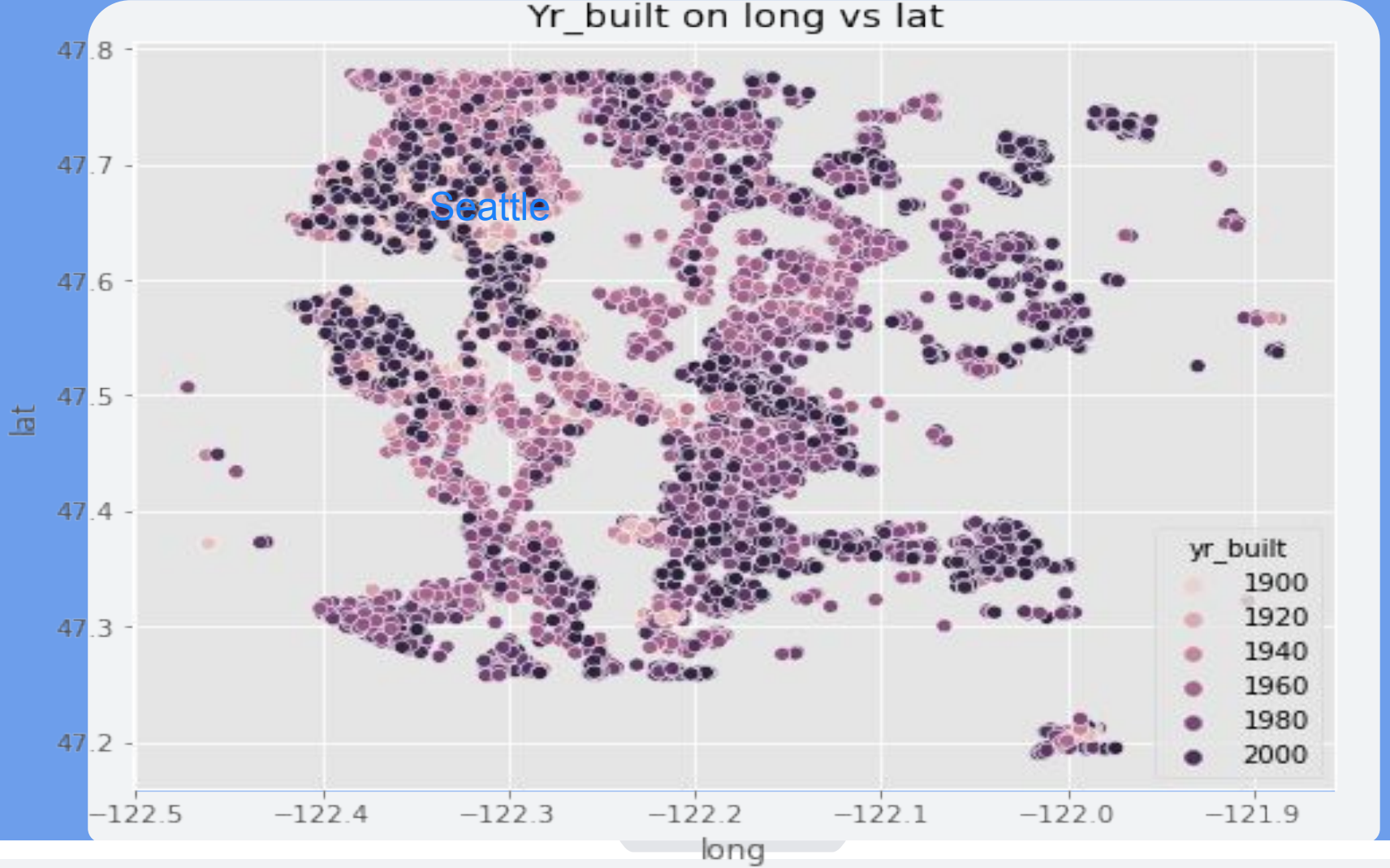
Model : price = 148.0812 * **sqft_living + 170,000**

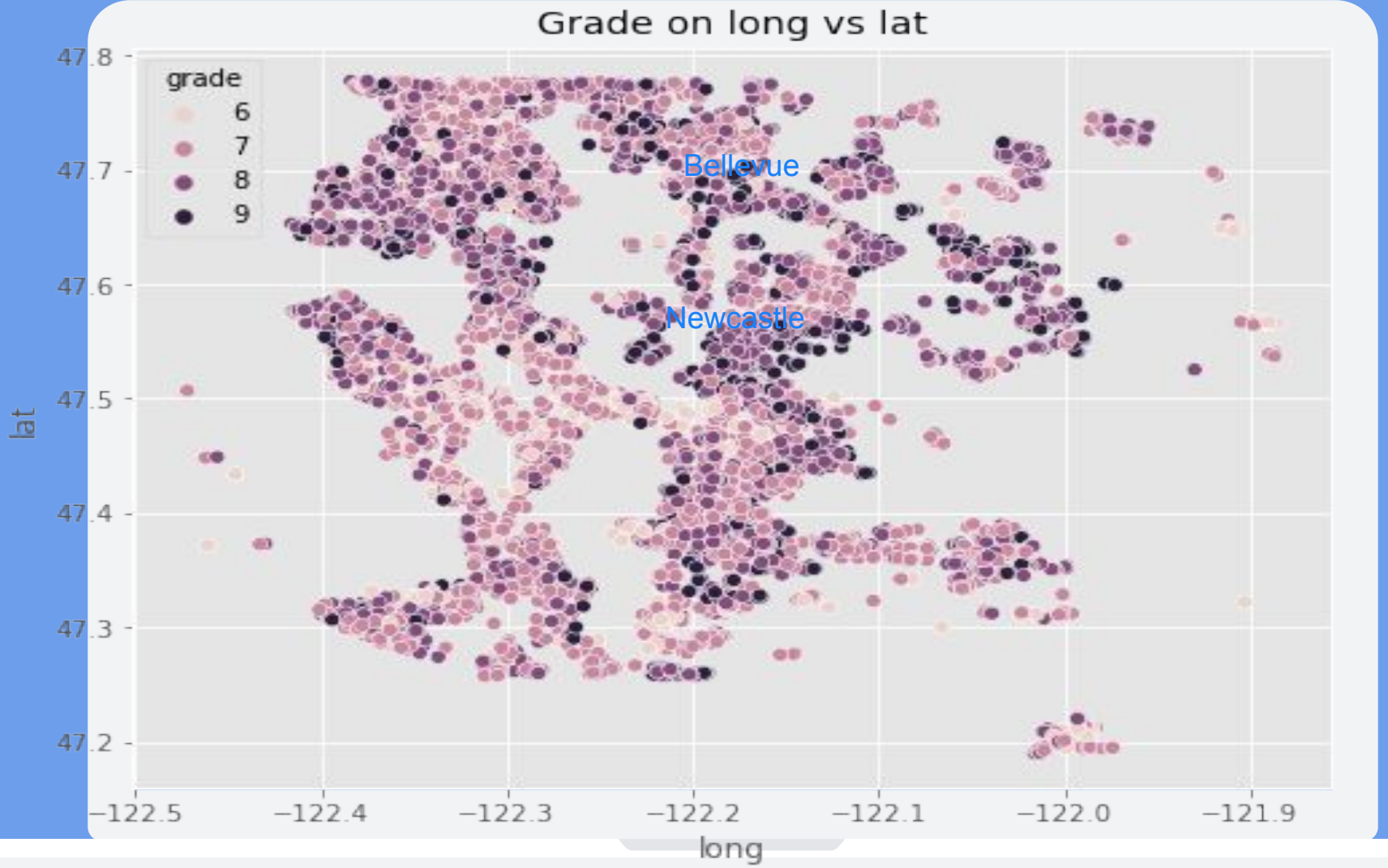
We found that each sqft_living cost $148.08
based on the correlation coefficient of sqft_living.

3.
Neighborhood Selecting 🏃

a. Price on long vs lat
b. Yr_built on long vs lat
c. Grade on long vs lat
d. Summary


Price on long vs lat
price
0.2
0.4
0.6
0.8
1.0
Seattle
Bellevue
Newcastle
47.8
47.7
47.6
47.5
47.4
47.3
47.2
lat
−122.5
−122.4
−122.3
−122.2
−122.1
−122.0
−121.9
long


Yr_built on long vs lat
Seattle
lat
long
47.8
47.7
47.6
47.5
47.4
47.3
47.2
−122.5
−122.4
−122.3
−122.2
−122.1
−122.0
−121.9
yr_built
1900
1920
1940
1960
1980
2000


Grade on long vs lat
grade
6
7
8
9
Bellevue
Newcastle
47.8
47.7
47.6
47.5
47.4
47.3
47.2
lat
−122.5
−122.4
−122.3
−122.2
−122.1
−122.0
−121.9
long

Summary

Neighborhood Selecting
*High price: Bellevue, Newcastle
Low price: Federal Way, Kent
*Old years: Seattle
*High grade: Bellevue, Newcastle

3.
Prediction Model 🏃

a. Related Features and Model Details

Summary

Related Features:
Number of Bedrooms
Number of Bedrooms
Sqft_lot: footage of the lot
Number of Floors
Condition(1-5, 5 is the best): How good the condition is
Grade(1-9, 9 is the best): overall grade given to the housing unit, based on King County grading system
Yr_built: Built Year
sqft_living15: The square footage of interior housing living space for the nearest 15 neighbors
Basement_exist: have basement or not, 1 is yes, 0 is no
Zip code

Accuracy:
The model can predict with 78.7% accuracy based on these features. Margin of error is around 60,000 USD of target home price.

Recommendation

For buyers: When buying a house, we should more concern about the powerful grade ranking and house size including footage of house, number of bathroom and so on because these features are most related to the house price. If the fund is enough, Bellevue and Newcastle are good neighborhood to invest because the good grade of them. Otherwise, Federal Way and Kent are alternative choices. When buying houses which located at Seattle, we should be careful because there are a lot of old houses located there.

For analysts: Transformation is good thing to improve R-squared and reduce condition number

Future Work

1. Interactions: Find some interactions on the model to see that whether helpful to improve the R-square
2. Kurtosis: Find some ways to reduce kurtosis to make the distribution more normal
3. Detailed Prediction: Give a price prediction to the buyer based on his prefer features using the model
4. More analyses: Try another business case such as helping a house seller

Thank you!

Any questions?

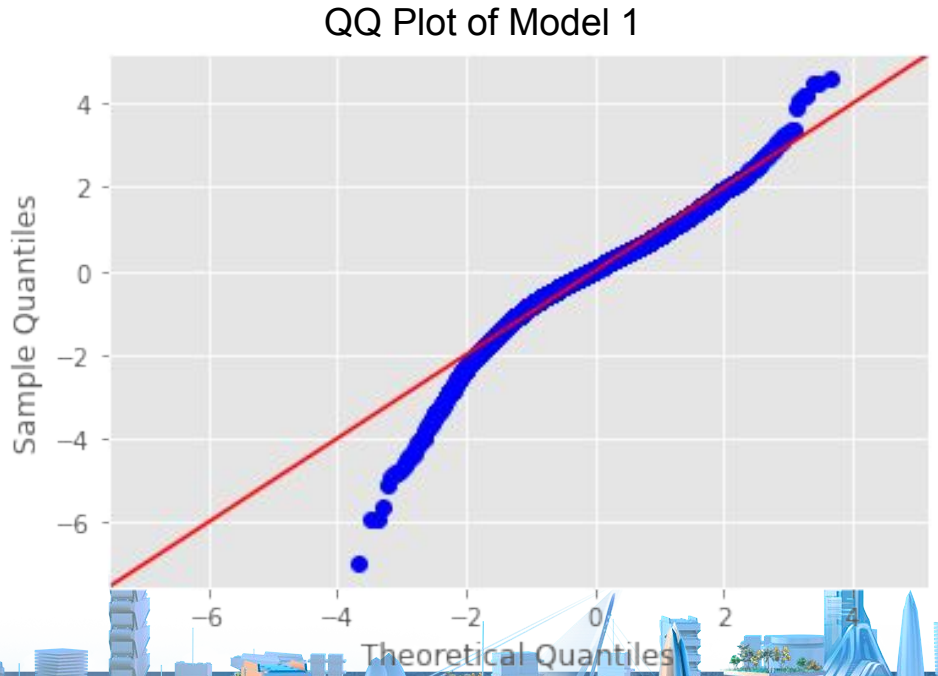
You can find me at:

https://www.linkedin.com/in/jiarong-jr-chen-ba87b214a/
https://github.com/JRRRRRRR

Appendix

a. Model 1: Non-transformation(zip code dummies)
b. Model 2: Log transformations and Standardize(zip code dummies)
c. Model 3: Log transformation and Min-max Scaling(zip code dummies)
d. Comparing
e. Model 1: Non-transformation(distance)
f. Model 2: Log transformations and Standardize(distance)
g. Model 3: Log transformation and Min-max Scaling(distance)
h. Comparing

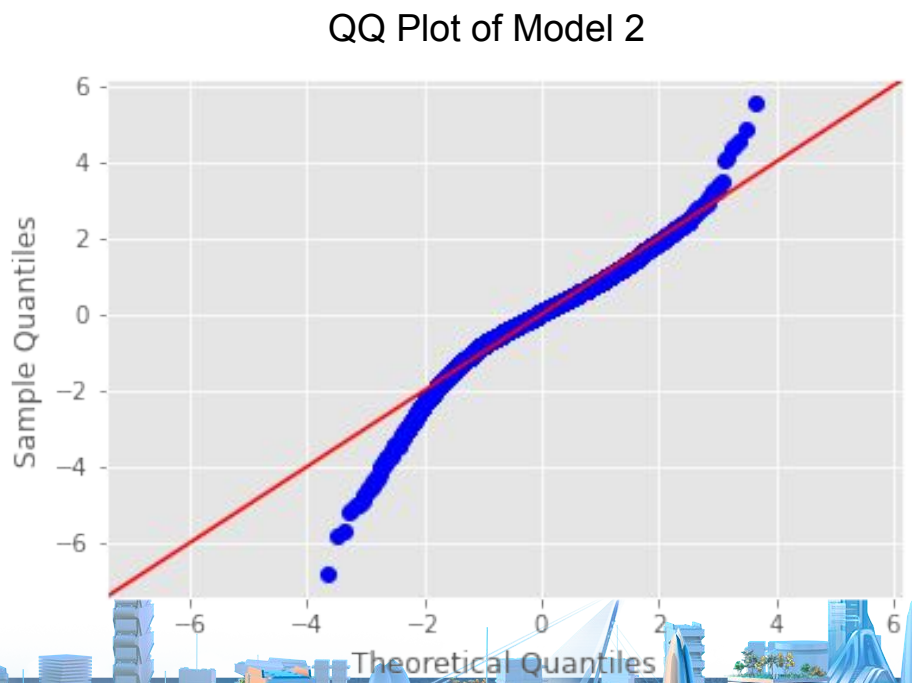
Model 1: Non-transformation(zip code dummies)



Train Mean Squared Error:
6611367342.657117
Test Mean Squared Error:
6726579183.313907
Mean Residuals:
59718.75468026469

R-squared: 0.786
Omnibus/Prob(Omnibus): 0
Skew: 0.627
Kurtosis: 5.761
Durbin-Watson: 2.001
Condition Number: 9.70e+05

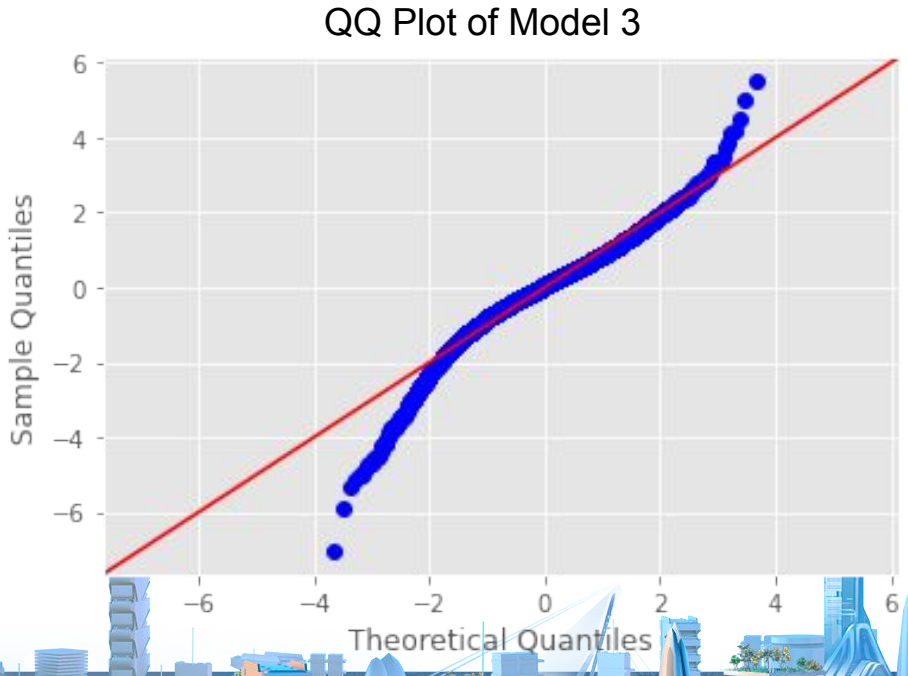
Model 2: Log transformations and Standardize(zip code dummies)



Train Mean Squared Error:
6673004860.434965
Test Mean Squared Error:
6304550879.062024
Mean Residuals:
59705.069439783256

R-squared: 0.787
Omnibus/Prob(Omnibus): 0
Skew: 0.602
Kurtosis: 5.654
Durbin-Watson: 1.997
Condition Number: 119

Model 3: Log transformation and Min-max Scaling(zip code dummies)



Train Mean Squared Error:
6587232865.491896
Test Mean Squared Error:
6655927970.392887
Mean Residuals:
59705.06943978331

R-squared: 0.787
Omnibus/Prob(Omnibus): 0
Skew: 0.602
Kurtosis: 5.654
Durbin-Watson: 1.997
Condition Number: 121

Comparing

Model 1

Train Mean Squared Error:
6611367342.657117
Test Mean Squared Error:
6726579183.313907
Mean Residuals:
59718.75468026469

R-squared: 0.786
Omnibus/Prob(Omnibus): 0
Skew: 0.627
Kurtosis: 5.761
Durbin-Watson: 2.001
Condition Number: 9.70e+05

Model 2

Train Mean Squared Error:
6673004860.434965
Test Mean Squared Error:
6304550879.062024
Mean Residuals:
59705.069439783256

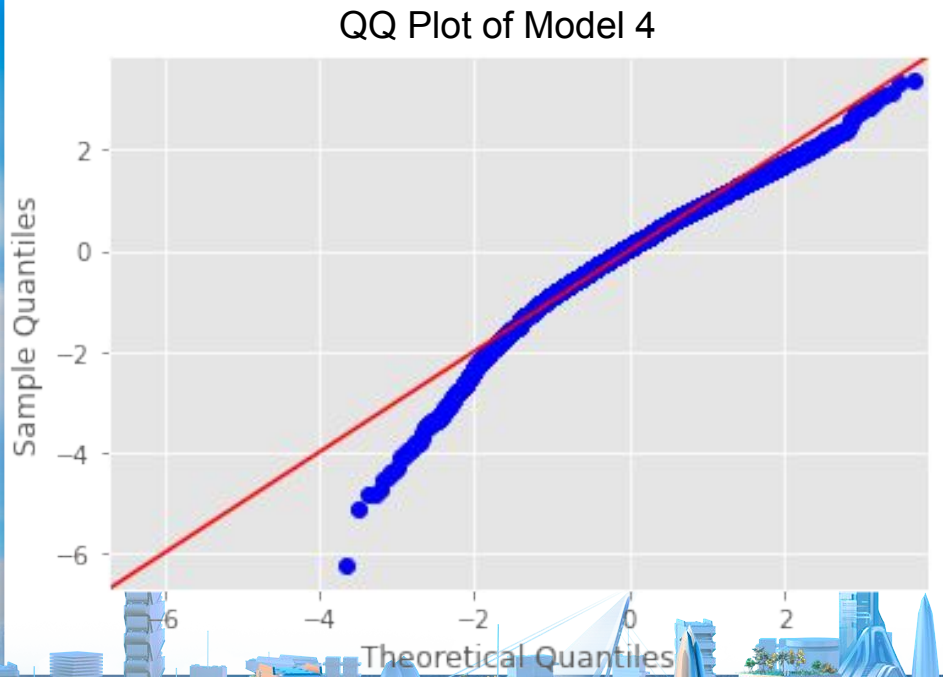
R-squared: 0.787
Omnibus/Prob(Omnibus): 0
Skew: 0.602
Kurtosis: 5.654
Durbin-Watson: 1.997
Condition Number: 119

Model 3

Train Mean Squared Error:
6587232865.491896
Test Mean Squared Error:
6655927970.392887
Mean Residuals:
59705.06943978331

R-squared: 0.787
Omnibus/Prob(Omnibus): 0
Skew: 0.602
Kurtosis: 5.654
Durbin-Watson: 1.997
Condition Number: 121

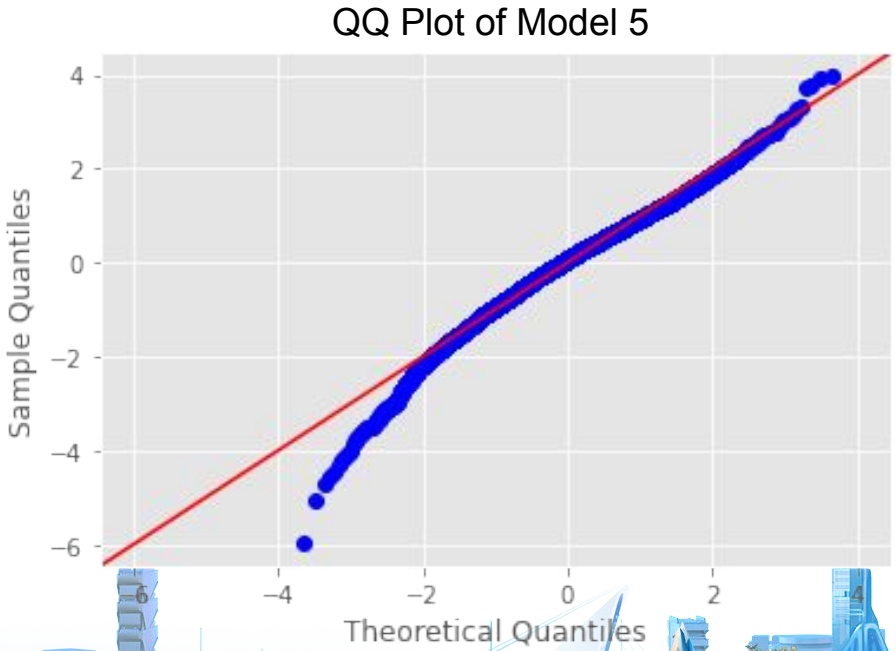
Model 1: Non-transformation(distance)



Train Mean Squared Error:
11110067996.111124
Test Mean Squared Error:
10658766992.5509
Mean Residuals:
80840.24297000362

R-squared: 0.644
Omnibus/Prob(Omnibus): 0
Skew: 0.649
Kurtosis: 4.342
Durbin-Watson: 1.994
Condition Number: 8.15e+05

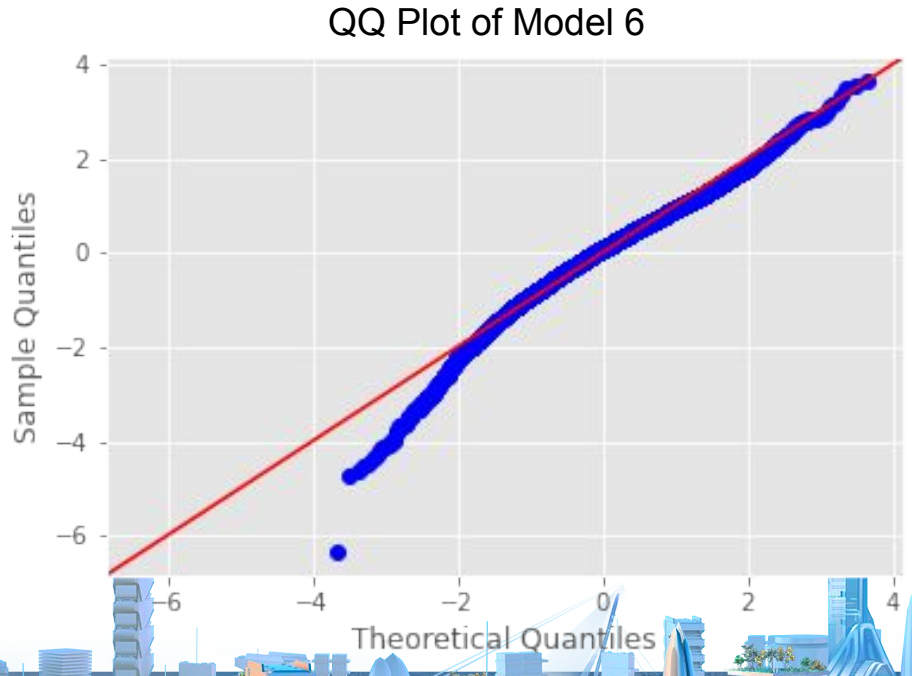
Model 2: Log transformations and Standardize(distance)



Train Mean Squared Error:
9534402926.103527
Test Mean Squared Error:
9531588843.911503
Mean Residuals:
75706.54494344277

R-squared: 0.692
Omnibus/Prob(Omnibus): 0
Skew: 0.408
Kurtosis: 4.037
Durbin-Watson: 1.984
Condition Number: 6.69

Model 3: Log transformation and Min-max Scaling(diatance)



Train Mean Squared Error:
10386628299.195896
Test Mean Squared Error:
10742839512.297537
Mean Residuals:
78490.80390278941

R-squared: 0.662
Omnibus/Prob(Omnibus): 0
Skew: 0.546
Kurtosis: 4.307
Durbin-Watson: 1.987
Condition Number: 20.0

Comparing

Model 4

Train Mean Squared Error:
11110067996.111124
Test Mean Squared Error:
10658766992.5509
Mean Residuals:
80840.24297000362

R-squared: 0.644
Omnibus/Prob(Omnibus): 0
Skew: 0.649
Kurtosis: 4.342
Durbin-Watson: 1.994
Condition Number: 8.15e+05

Model 5

Train Mean Squared Error:
9534402926.103527
Test Mean Squared Error:
9531588843.911503
Mean Residuals:
75706.54494344277

R-squared: 0.692
Omnibus/Prob(Omnibus): 0
Skew: 0.408
Kurtosis: 4.037
Durbin-Watson: 1.984
Condition Number: 6.69

Model 6

Train Mean Squared Error:
10386628299.195896
Test Mean Squared Error:
10742839512.297537
Mean Residuals:
78490.80390278941

R-squared: 0.662
Omnibus/Prob(Omnibus): 0
Skew: 0.546
Kurtosis: 4.307
Durbin-Watson: 1.987
Condition Number: 20.0