Multiple Regression on House

Presented by: Jiarong Chen

Background

Business problem

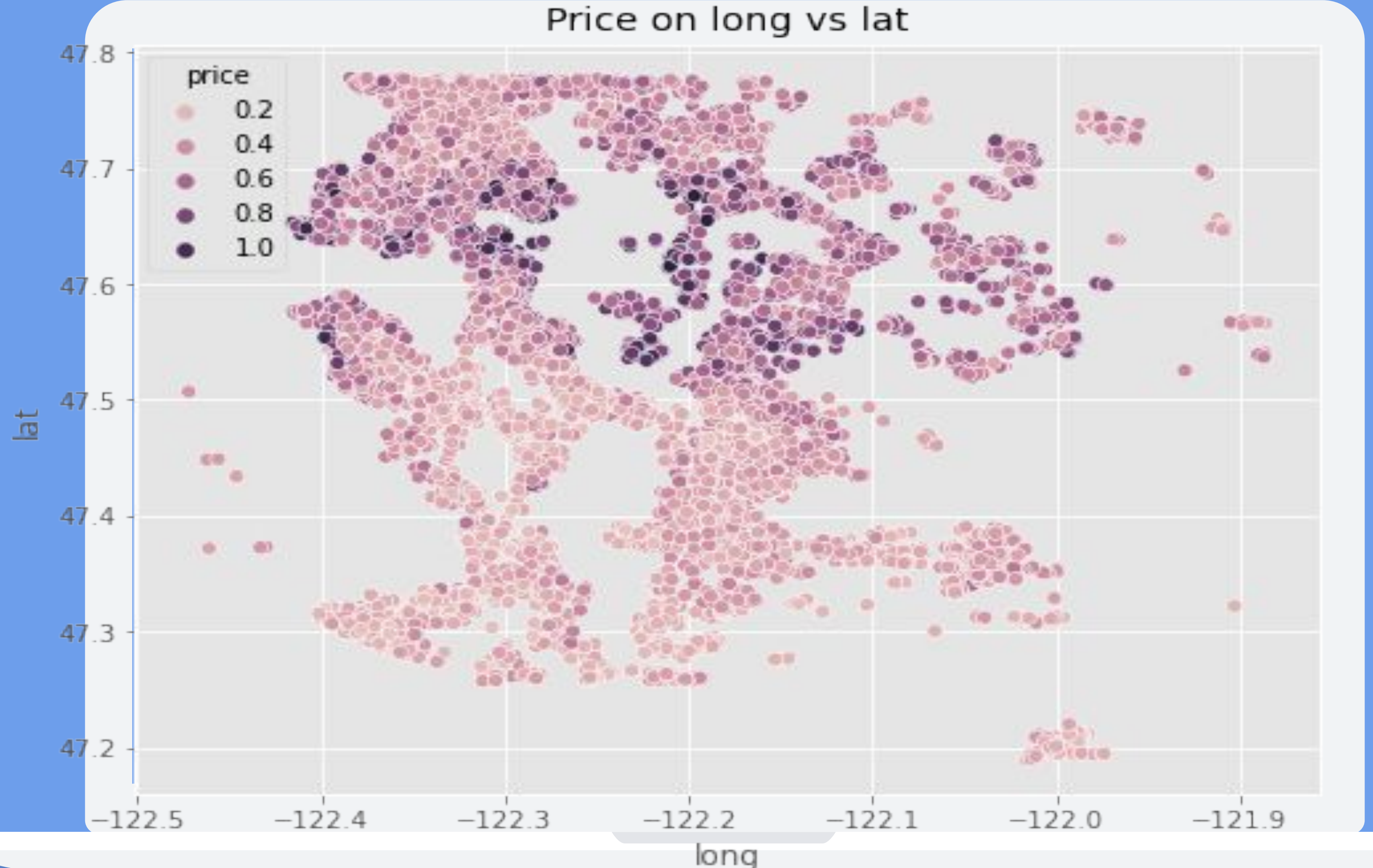
A house buyer assigns me a task about the house in King County. He wants to buy a house in this area but doesn't have any ideas about the housing market. And he has some preferred features in his mind, he wants to have a predicted price so that he can prepare for that.

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. **How much should he prepare for the dream house?**
   Find the prediction of price with model.

1.
Most Related Features

a. Preview
b. Visualization
c. Summary


Price on long vs lat
price
0.2
0.4
0.6
0.8
1.0
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


heatmap of price vs other features
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

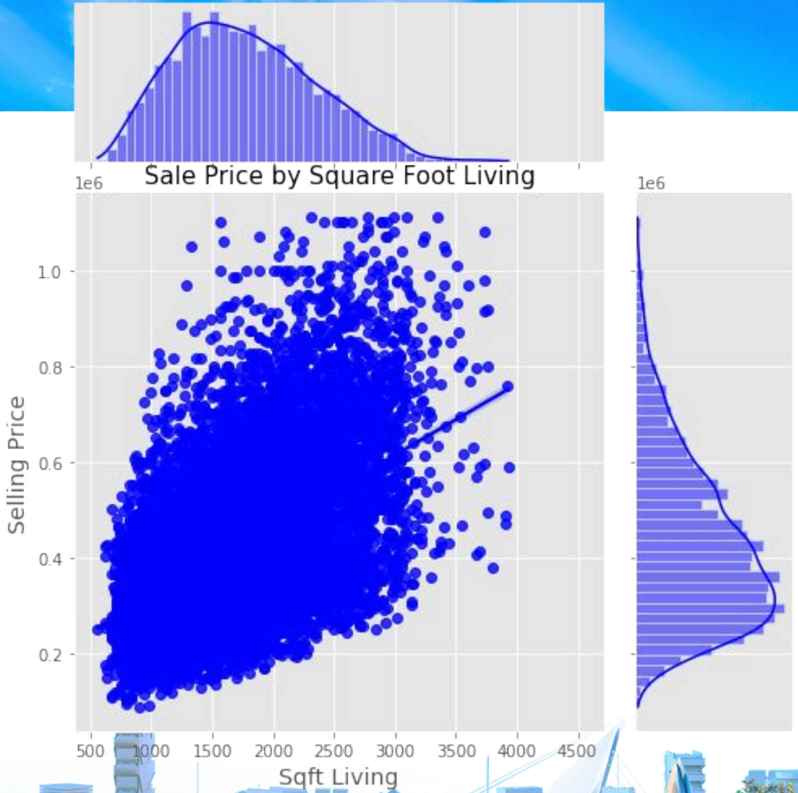
```
[('price', 1.0),
 ('grade', 0.49261027416442066),
 ('sqft_living', 0.48222641077964523),
 ('sqft_living15', 0.3979299346232877),
 ('sqft_above', 0.36084369884935624),
 ('bathrooms', 0.3112049323614764),
 ('bedrooms', 0.21432266736607247),
 ('floors', 0.20297606726310086),
 ('basement_exist', 0.19564092494704588),
 ('condition', 0.10759942018093975),
 ('zipcode', 0.019162250231750035),
 ('yr_built', -0.07066375804244265),
 ('sqft_lot', -0.10333797300264337),
 ('sqft_lot15', -0.11468772320858822)]
```

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.



correction coefficient:
0.482226

Summary

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

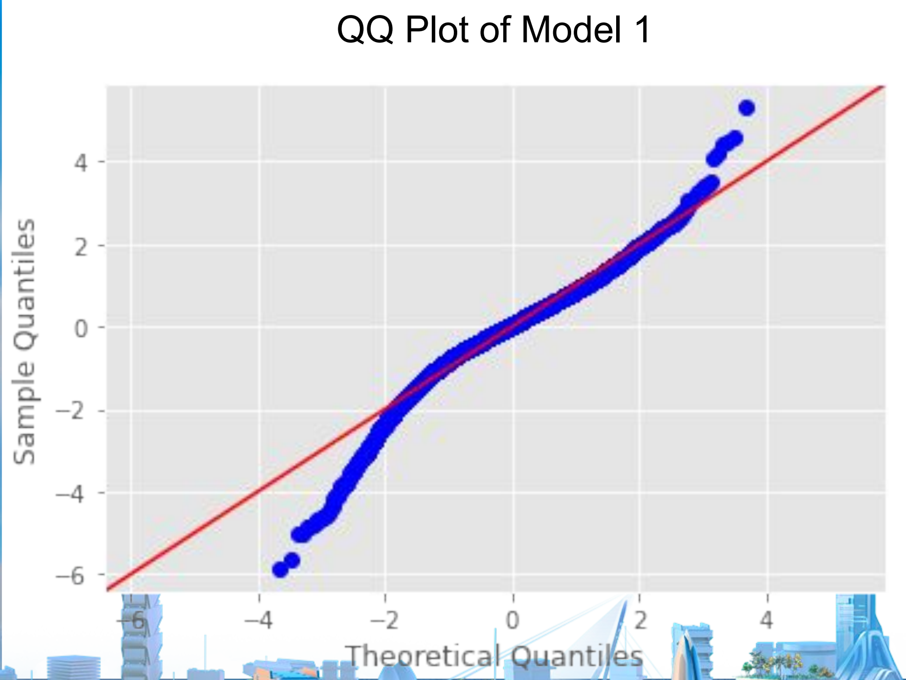
We found that each sqft_living cost $148.08
based on the correlation coefficient of sqft_living.

3.

Prediction 🏃

a. Model 1: Non-transformation
b. Model 2: Log transformations and Standardize
c. Model 3: Log transformation and Min-max Scaling
d. Comparing and Summary

Model 1: Non-transformation

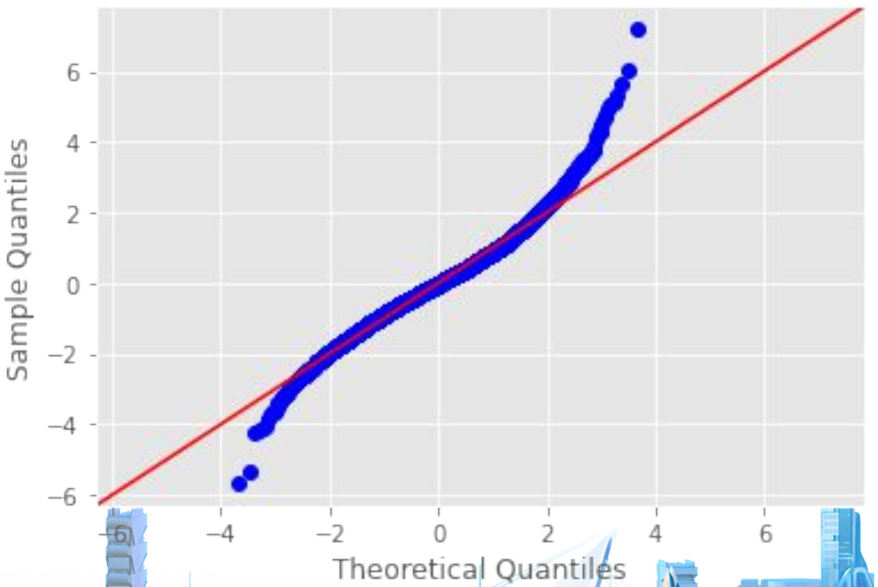


Train Mean Squared Error：6533509266.241655
Test Mean Squared Error：7046160771.013472

R-squared: 0.786
Omnibus/Prob(Omnibus): 0
Skew: 0.627
Kurtosis: 5.761
Durbin-Watson: 2.001
Condition Number: 970,000

Model 2: Log transformations and Standardize

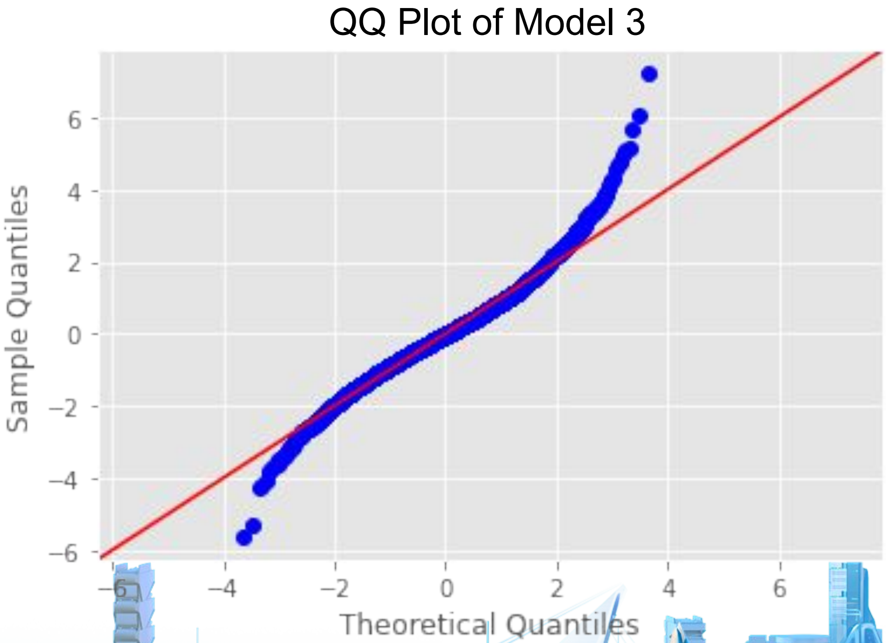
QQ Plot of Model 2



Train Mean Squared Error: 0.18127323976869805
Test Mean Squared Error: 0.19451224725079647

R-squared: 0.816
Omnibus/Prob(Omnibus): 0
Skew: -0.304
Kurtosis: 4.974
Durbin-Watson: 1.999
Condition Number: 119

Model 3: Log transformation and Min-max Scaling



Train Mean Squared Error: 0.004715869788110689
Test Mean Squared Error: 0.004739938713457318

R-squared: 0.816
Omnibus/Prob(Omnibus): 0
Skew: -0.304
Kurtosis: 4.974
Durbin-Watson: 1.999
Condition Number: 130

Comparing and Summary

Model 1	Model 2	Model 3
Train Mean Squared Error: 6533509266.241655	Train Mean Squared Error: 0.18127323976869805	Train Mean Squared Error: 0.004715869788110689
Test Mean Squared Error: 7046160771.013472	Test Mean Squared Error: 0.19451224725079647	Test Mean Squared Error: 0.004739938713457318
R-squared: 0.786	**R-squared: 0.816**	**R-squared: 0.816**
Omnibus/Prob(Omnibus): 0	**Omnibus/Prob(Omnibus): 0**	**Omnibus/Prob(Omnibus): 0**
Skew: 0.627	**Skew: -0.304**	**Skew: -0.304**
Kurtosis: 5.761	**Kurtosis: 4.974**	**Kurtosis: 4.974**
Durbin-Watson: 2.001	**Durbin-Watson: 1.999**	**Durbin-Watson: 1.999**
Condition Number: 970,000	**Condition Number: 119**	**Condition Number: 130**

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

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