House Sales in King County, USA

Predict house price using regression

Presented By Team NHEEEQMT
National Higher Education Entrance Examination Question Making Team

Including ADHC, ALBY, ESYC, ICQY

Contents

“Every good powerpoint has contents”
Bill Gates

01. Backgruond
02. Data Description
03. Data Visualisation
04. Data Preprocessing
05. Model & Evaluation
06. Conclusion

Background

Introduction to our project

- This dataset contains house sale price for King County, which includes Seattle. It includes houses sold between May 2014 and May 2015.
- It's a great dataset for evaluating simple regression models.

Why it is important

- Predicting house price can be helpful to people who want to purchase houses.
- House price are hard to predict in normal life. That’s why Machine Learning is so important.
- Our work lets people understand the variation of house price easily, and lets everyone predict house price.

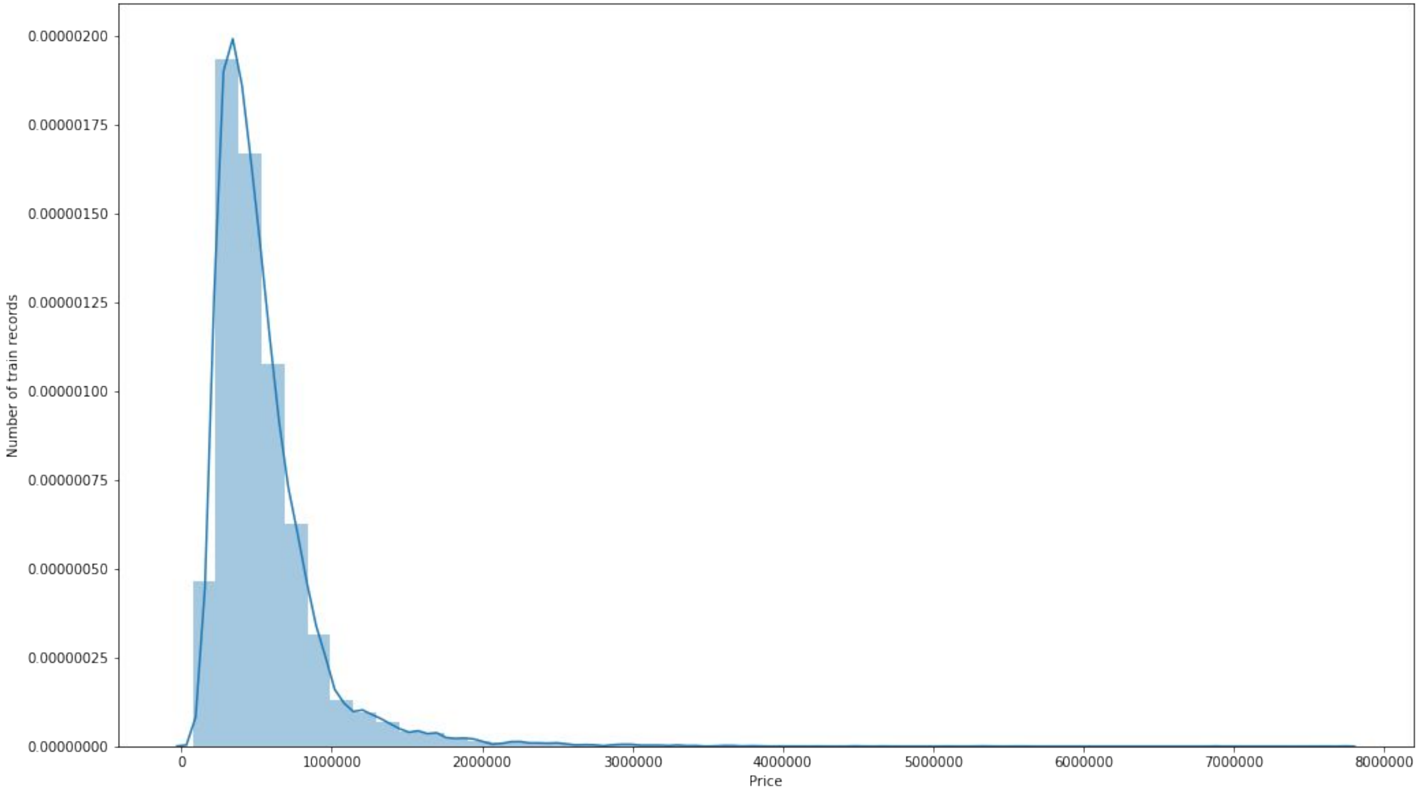
Data Description

The Data We Had

- When we got the data, we had 21 features, and 21613 data entries.

Distribution of Target

- We can see that most of the price is between 0 and 1e6. Most of the price is near 5e5.

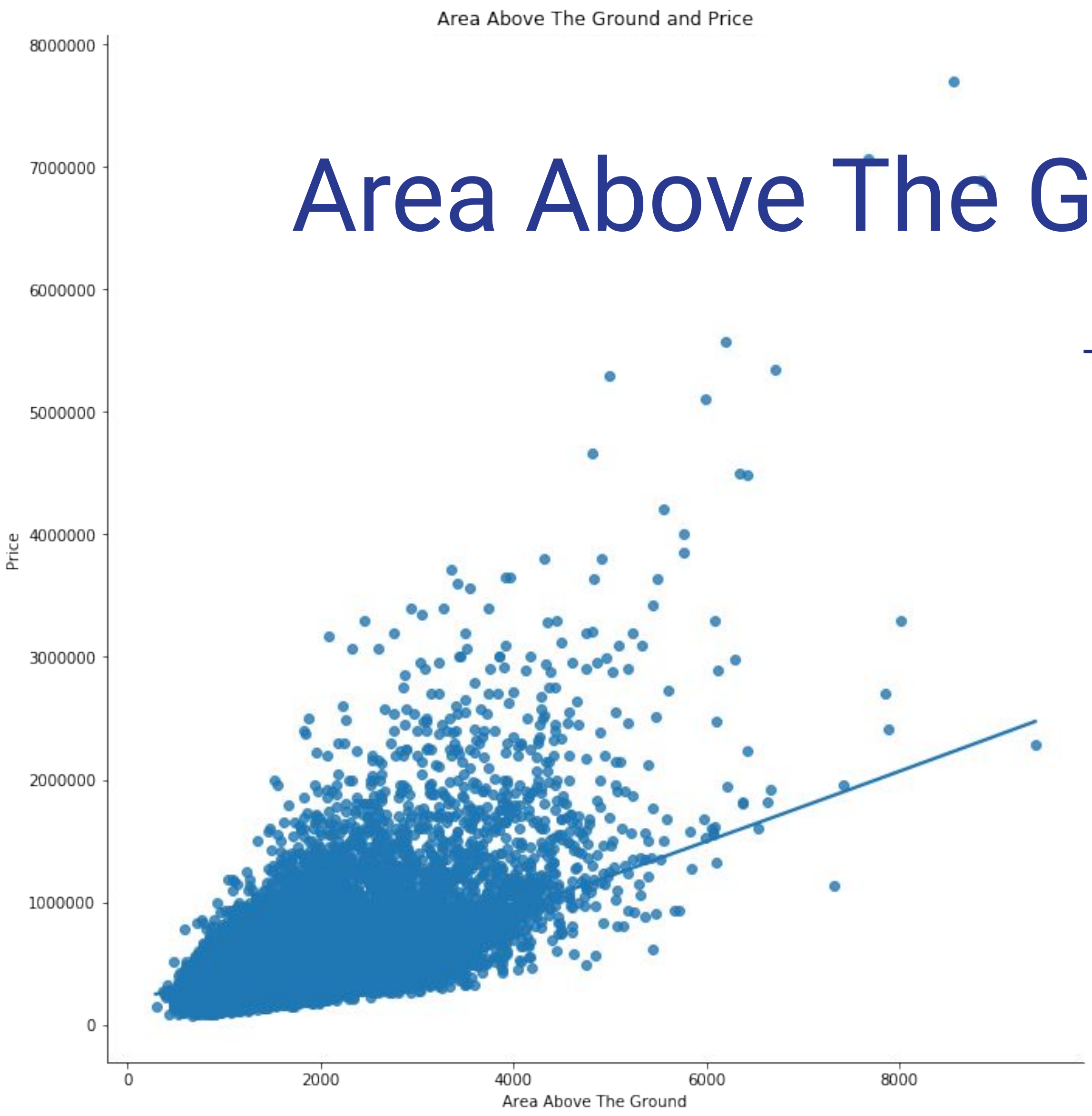


Features and Data Types

Feature	Type	Feature	Type	Feature	Type	Feature	Type	Feature	Type
id	int64	sqft_living	int64	condition	int64	yr_renovated	int64	sqft_lot15	int64
date	object	sqft_lot	int64	grade	int64	zipcode	int64	-	-
price	float64	floors	float64	sqft_above	int64	lat	float64	-	-
bedrooms	int64	waterfront	int64	sqft_basement	int64	long	float64	-	-
bathrooms	float64	view	int64	yr_built	int64	sqft_living15	int64	-	-

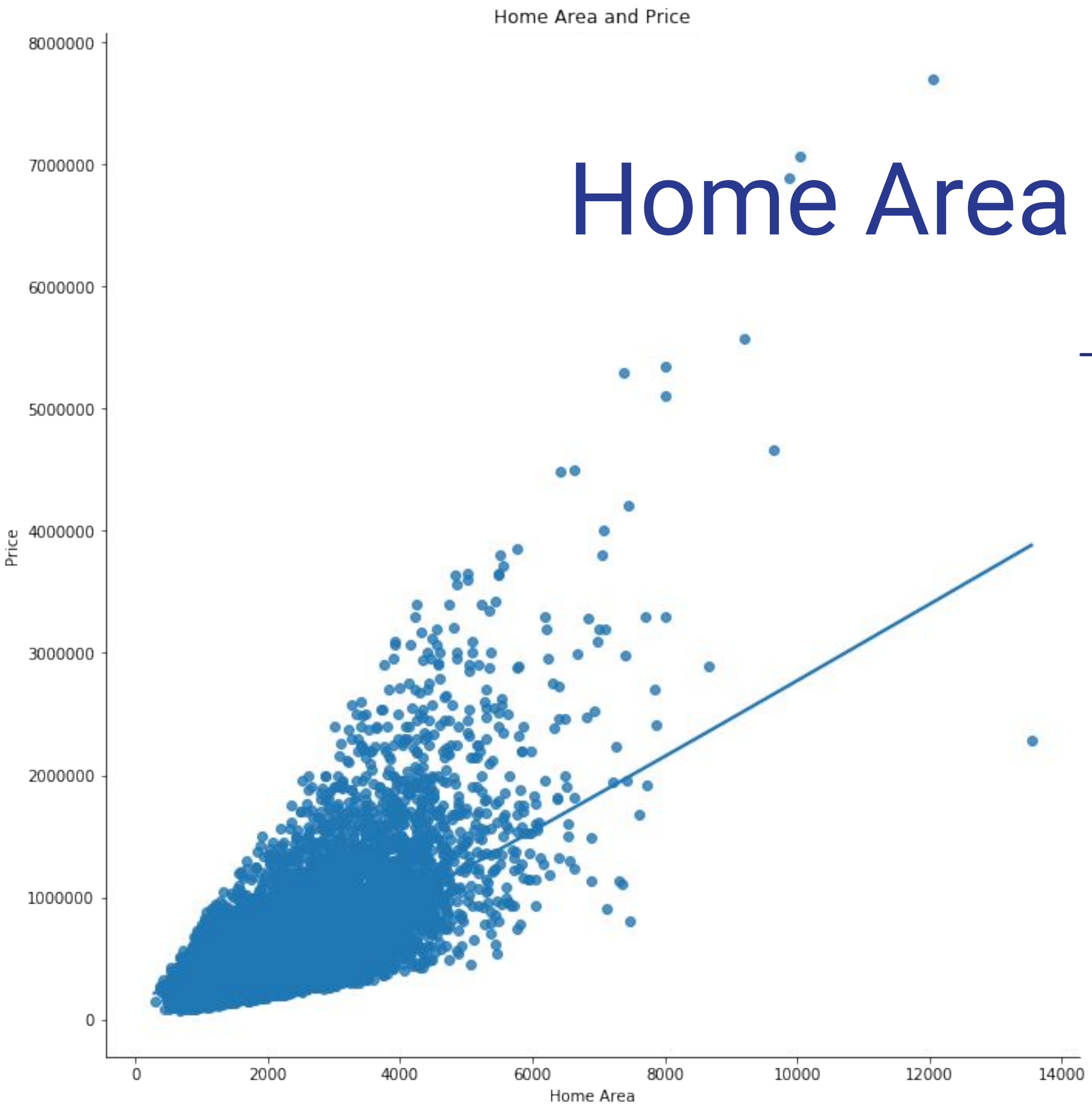
Data Visualisation

Area Above The Ground and Price



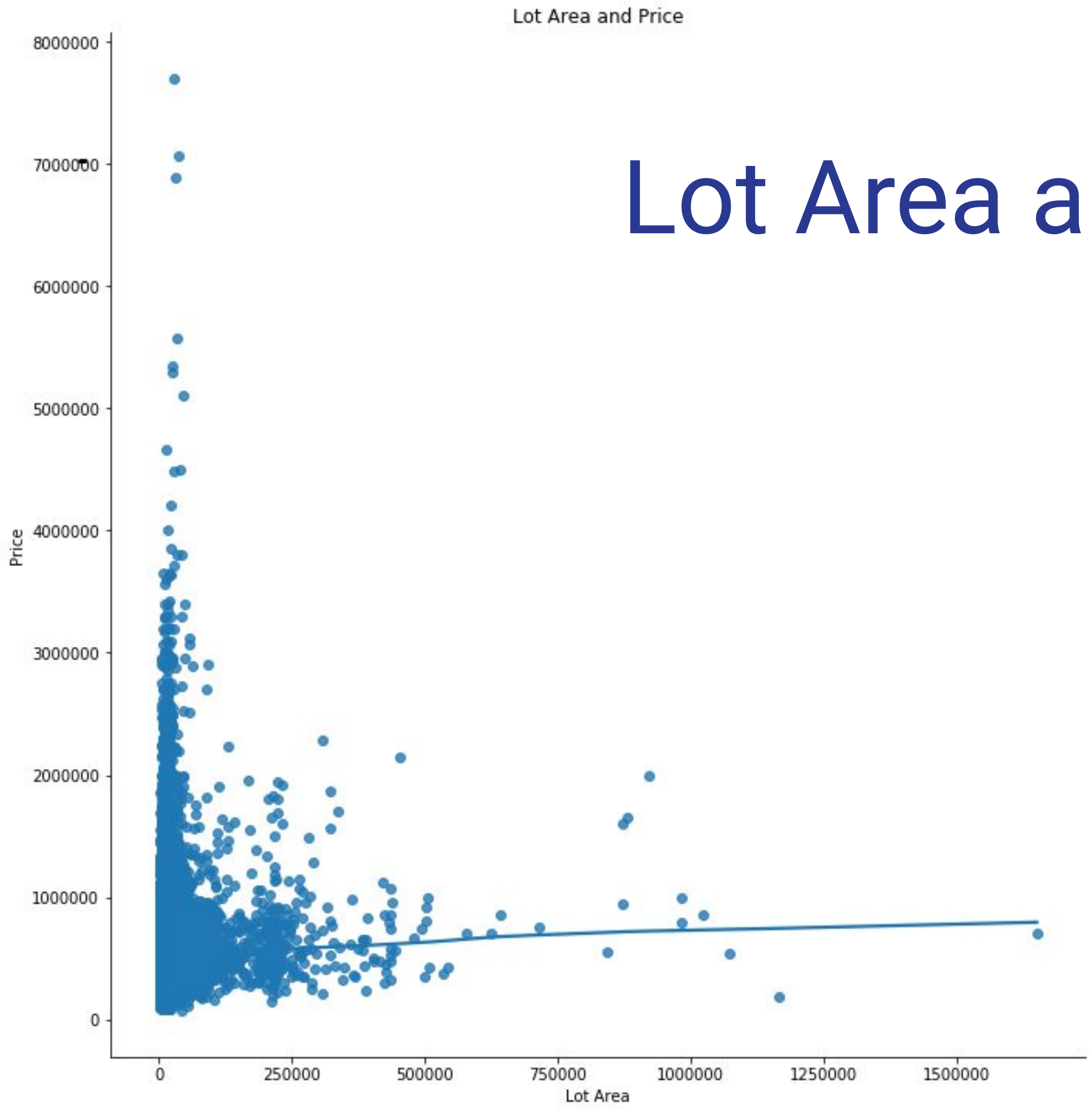
- Clearly, as size of the area above the ground goes up, the house price go up. While the area above the ground is between 2000-4000, price is mainly 1,000,000.

Home Area and Price



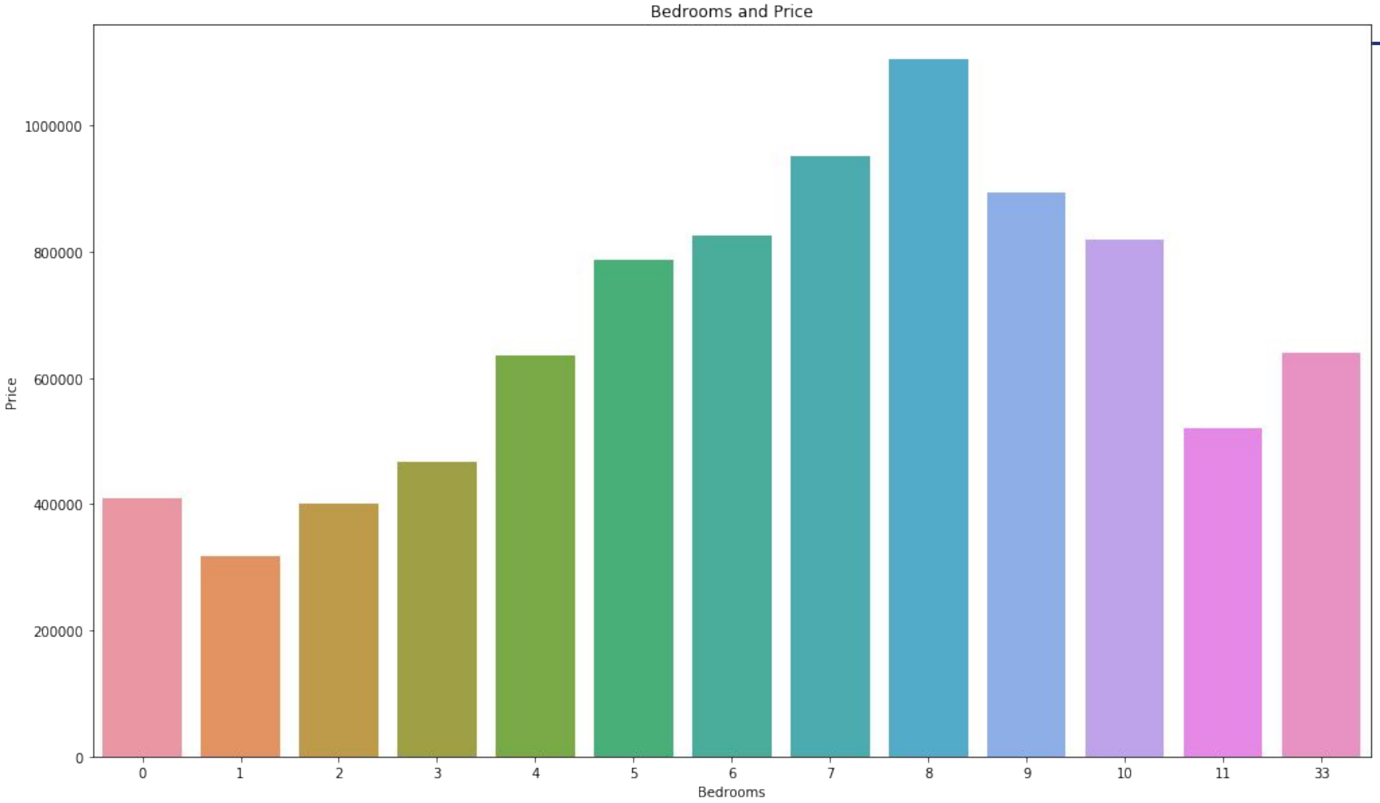
- Already, we can start to form a picture of the relationship between size of home and the price houses were sold for. Clearly, as size of the home goes up, the house price usually go up. While the area of the home is between 2000-4000, price is mainly 1,000,000.

Lot Area and Price



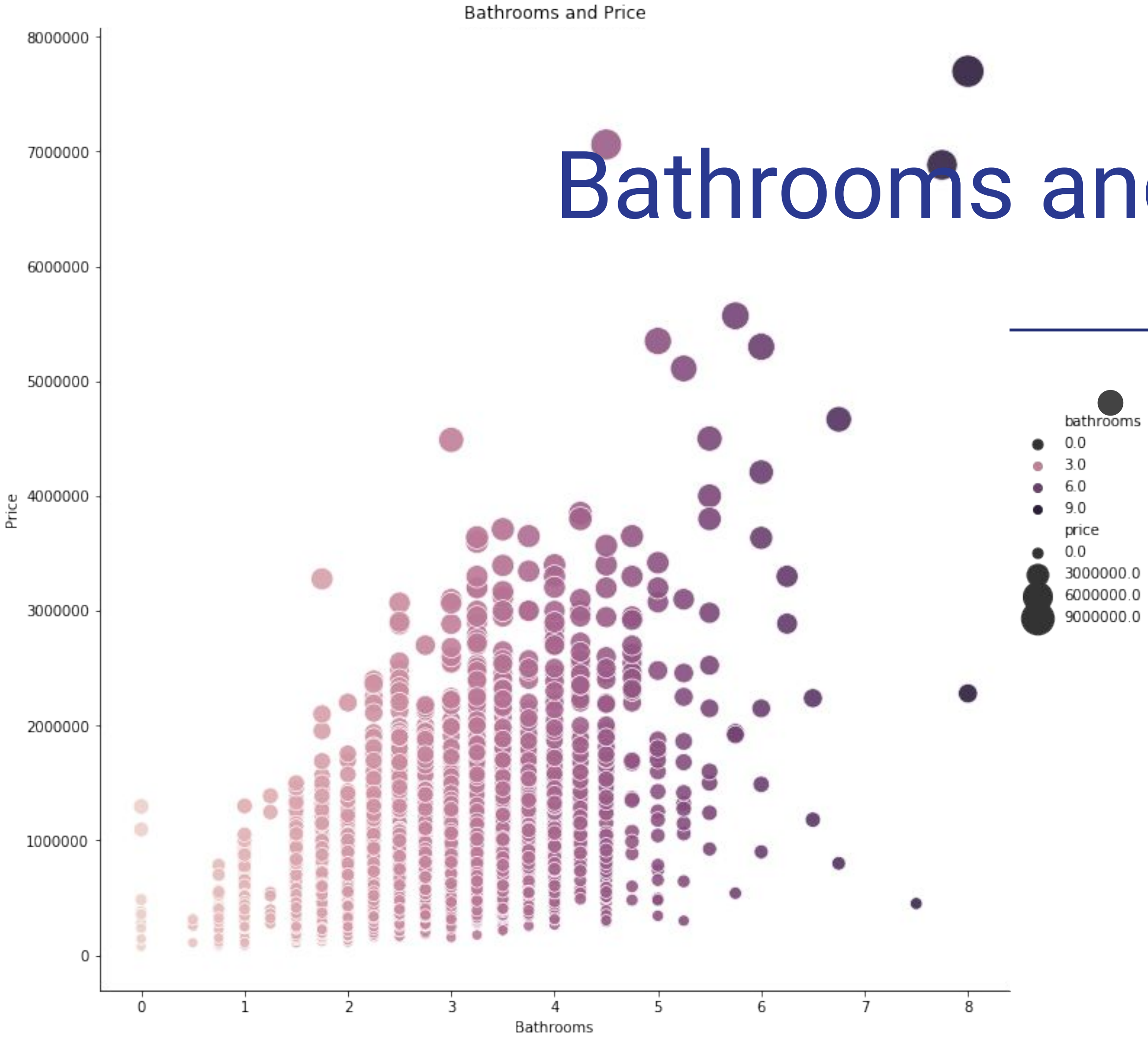
- From this plot we can see that most of the lot areas are 0. Actually, they just have little relation.

Bedrooms and Price



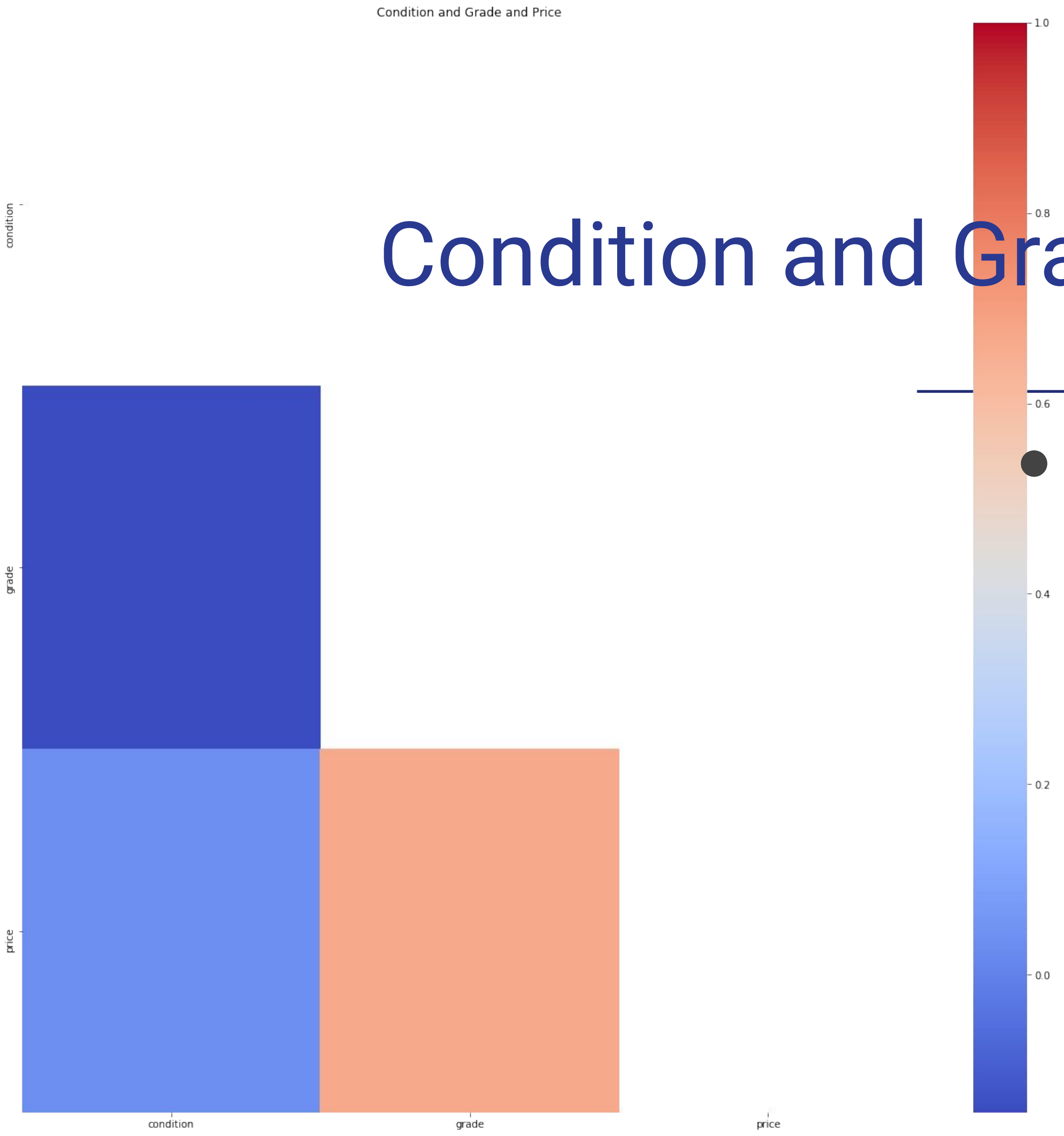
- This picture shows the relationship between the number of bedrooms and the price. From this chart we can see the number of bedrooms goes up but the price sometimes does not go up as the bedrooms.

Bathrooms and Price



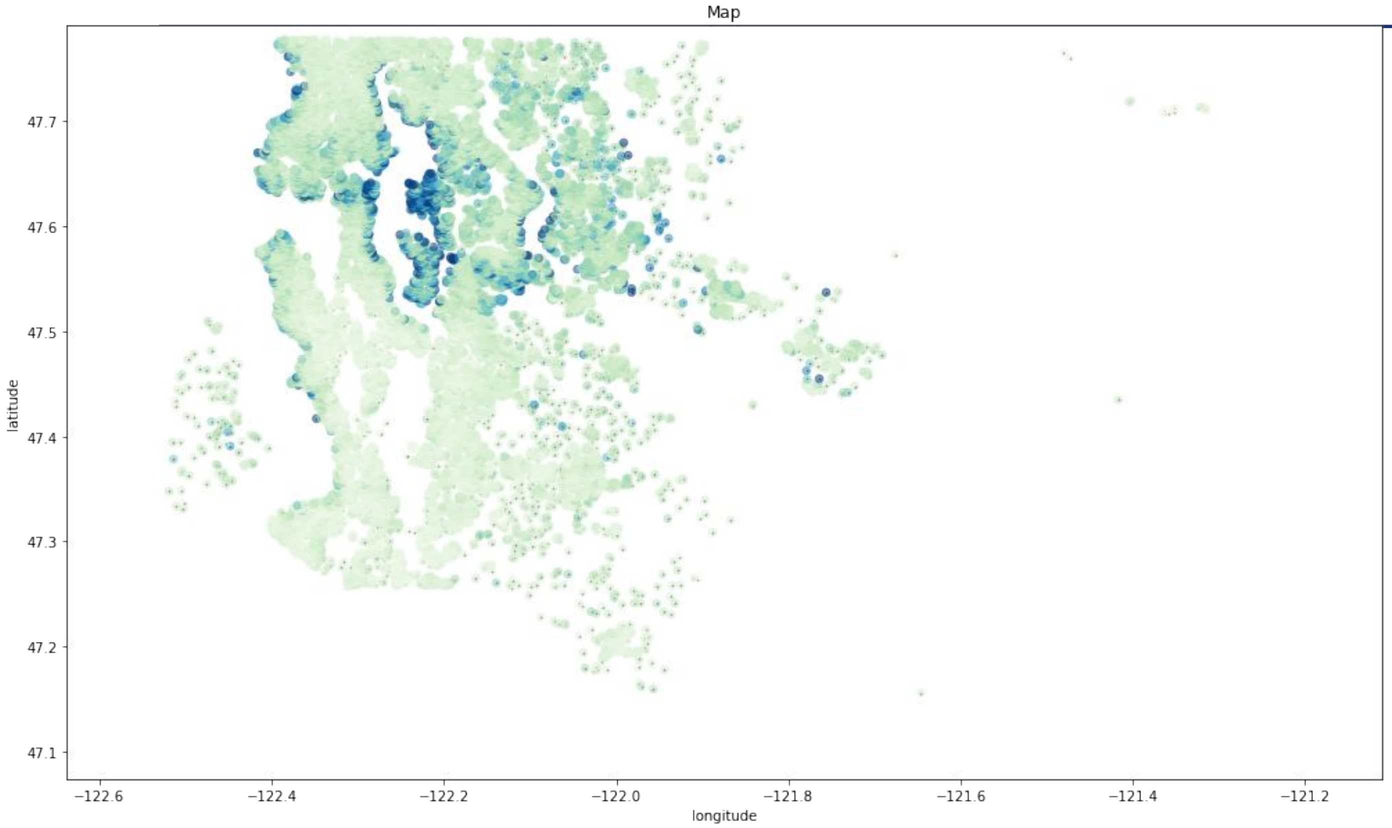
- This plot illustrates the relationship between the price and the bathrooms. At the same time, the size of the point represents the price level, and the higher the point, the higher the house price is.

Condition and Grade and Price



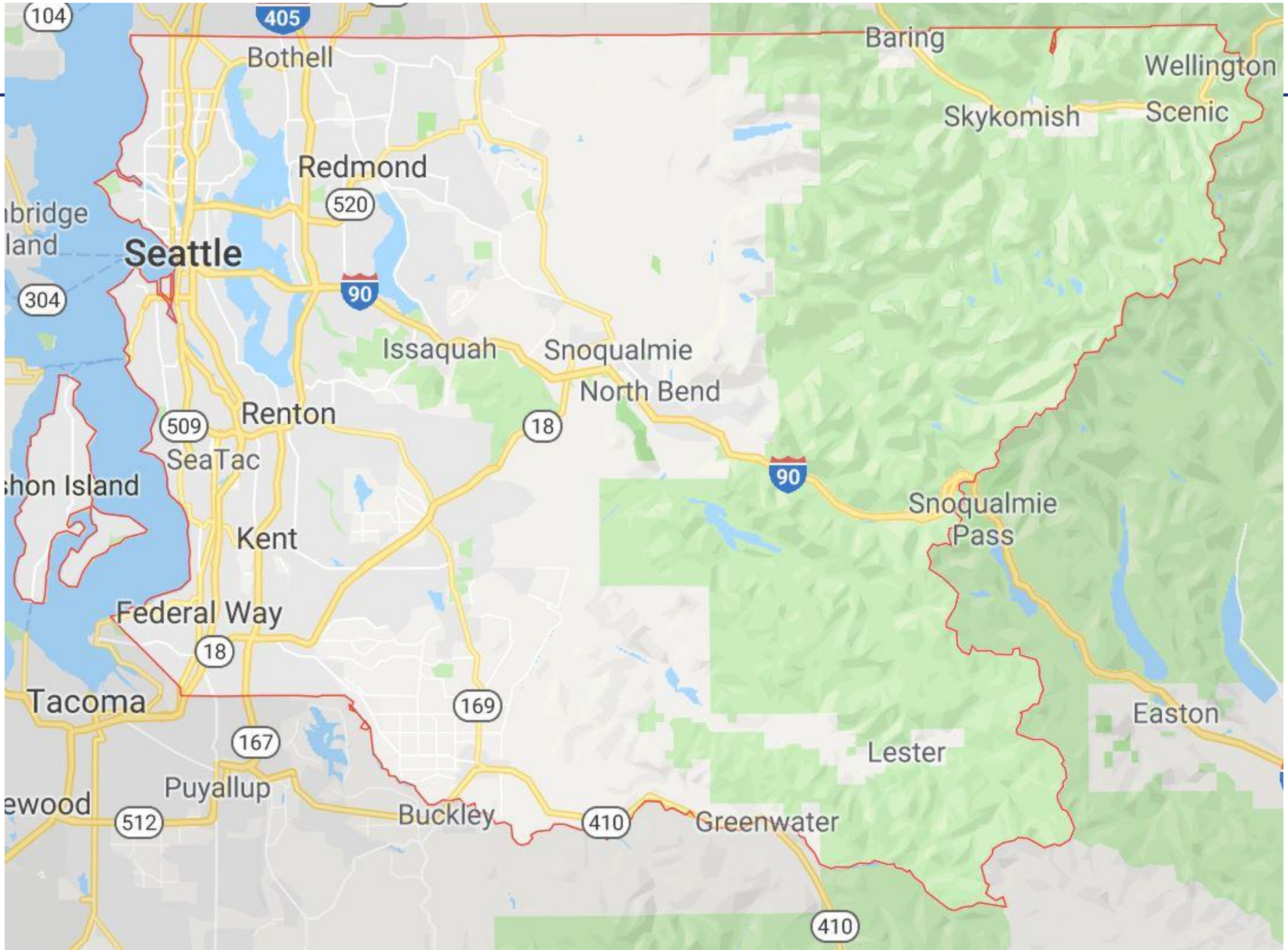
- Price seems to relate to grade more. But grade and condition are both used to judge houses.

Map of Price

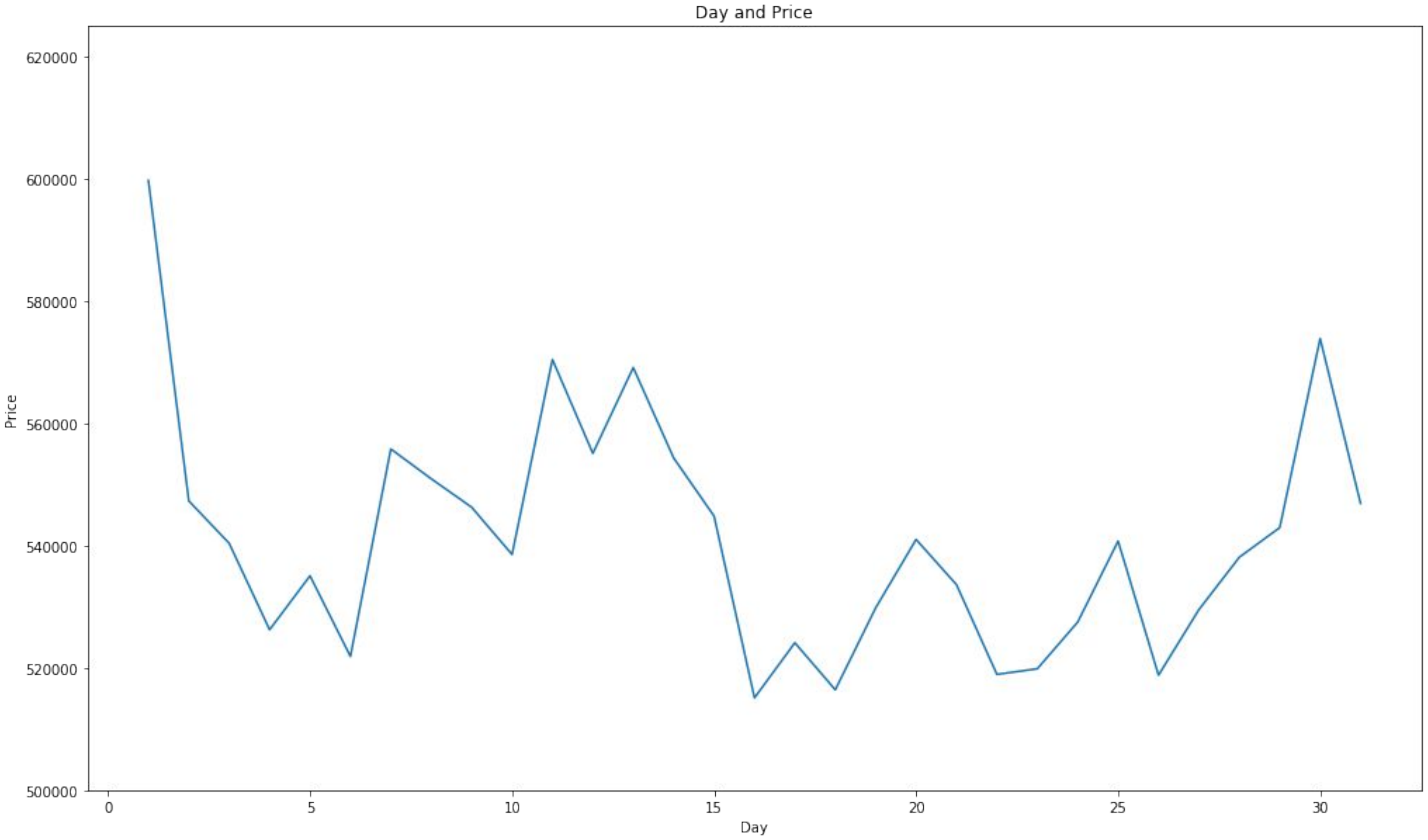


- We can see that the price of the houses near the waters is greater than the price of houses far from the waters.

Map Of King County

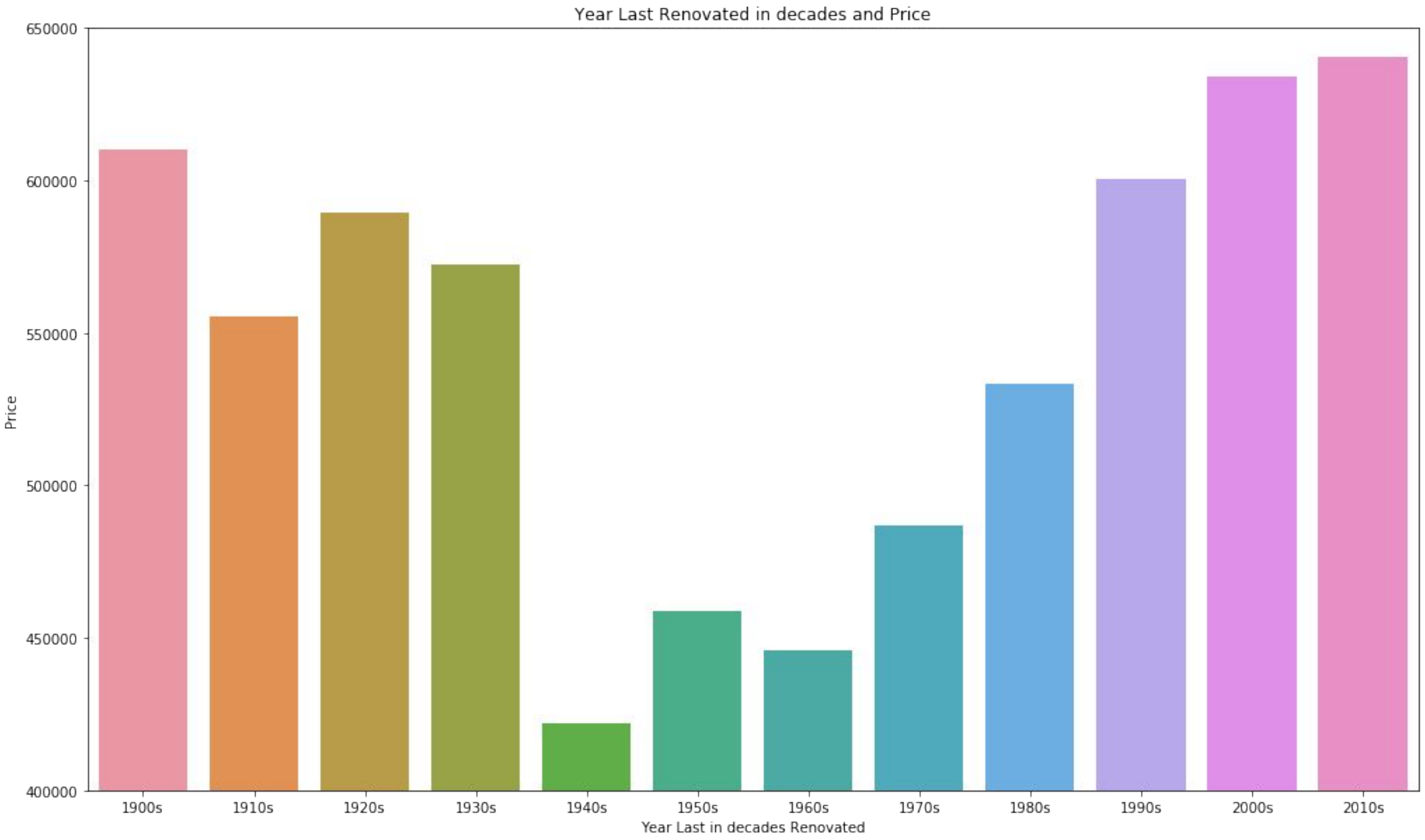


Day Sold and Price



- As we can see, there is not big relationship between the day and the price.

Year Last Renovated in Decades and Price



- We find that **very** old houses worth a lot of money, but old houses do not worth a lot. New houses worth a lot.

Data Preprocessing

Data Preprocessing

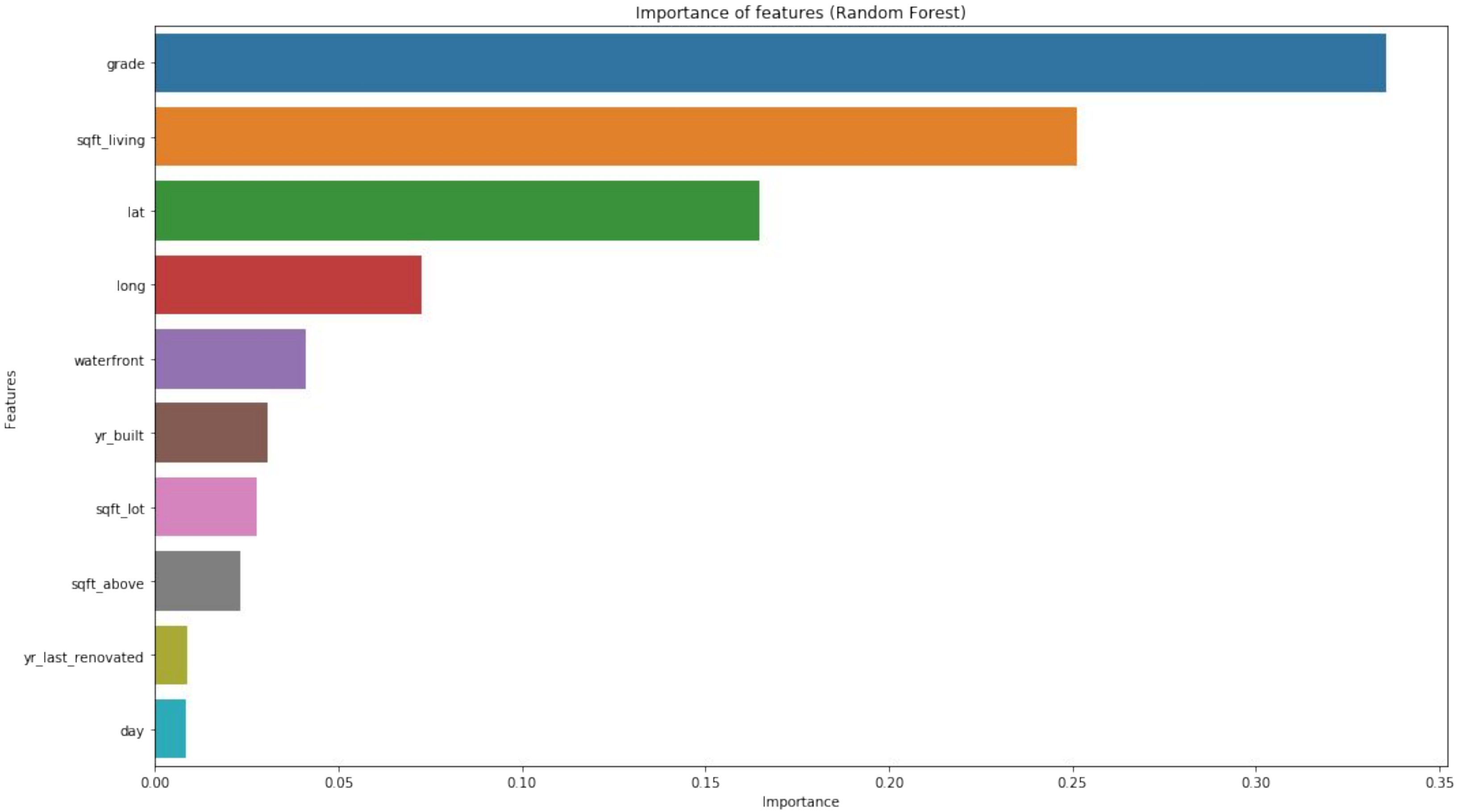
- This time, we only used Min-Max Data Normalization because some of the prices are too big.

Model & Evaluation

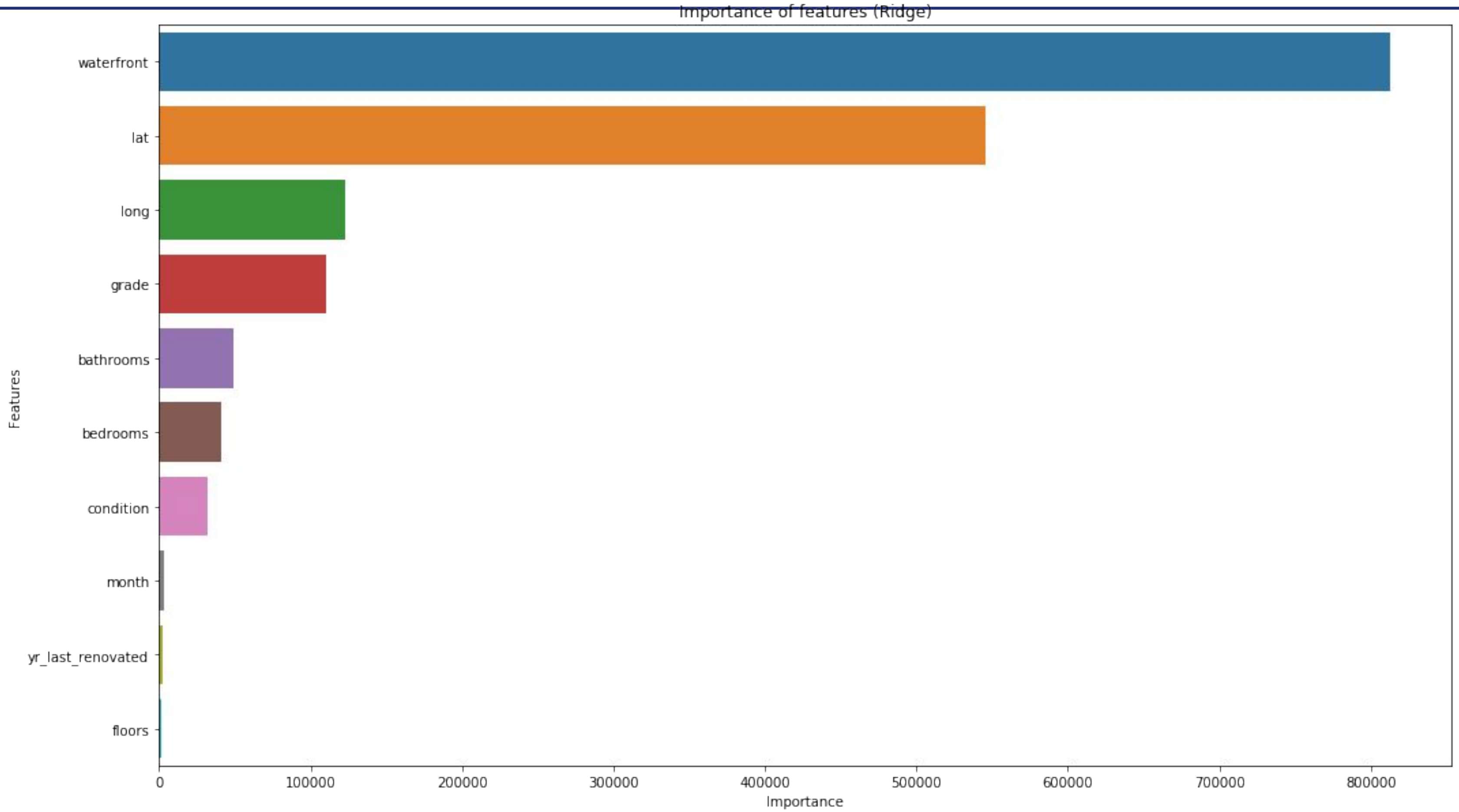
Model

- We’ve used Ridge (Linear Regression) and Random Forest Regression.
- Random Forest performs better because it combines lots of weak regressions into a strong regression. Ridge is only a weak regression.

Evaluation - Random Forest



Evaluation - Ridge

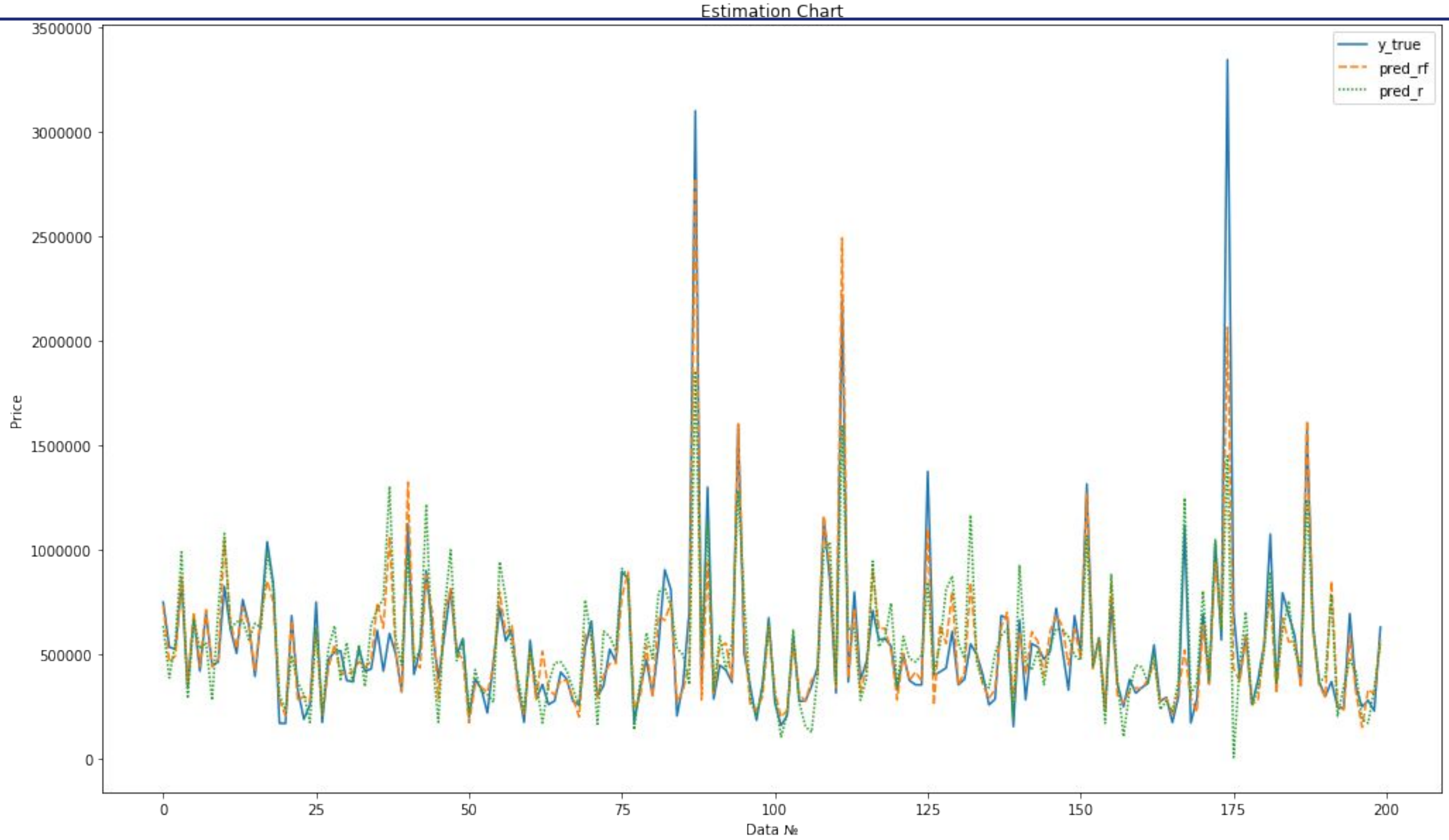


Conclusion

Results

Type	Square Root of MSE	MAE
RF	145564.93137247238	78939.5736967117
R	205446.90222072878	129104.63779307903

Test Results



Value

- We can predict house prices in King County now!

What have we learnt from the camp

- ADHC: Commodity price is really expensive.
- ALBY: Cooperating with team members, data analysis and machine learning.
- ESYC: How to do data analysis and data visualization.
- ICQY: Lets me be interested in knowledge in this field.

Thank you for listening!

NHEEEQMT