Springboard

CAPSTONE PROJECT 1

REPORT

HOUSE PRICE PREDICTION

SPRINGBOARD

JUNE ,2019

Prepared by:

Mehreteab Kidane

Mentor

Kenneth Gil-Pasquel

OBJECTIVES OF THIS PROJECT

- To determine the causes of increase in housing price
- To predict the price of housing price in the future
- To identify the effect of increase in housing price

DATA OVERVIEW

- Original dataset found from Kaggle website
- Kings County ,Seattle ,Washington
- House sold between May 2014 and May 2015
- 21613 observations and 20 features

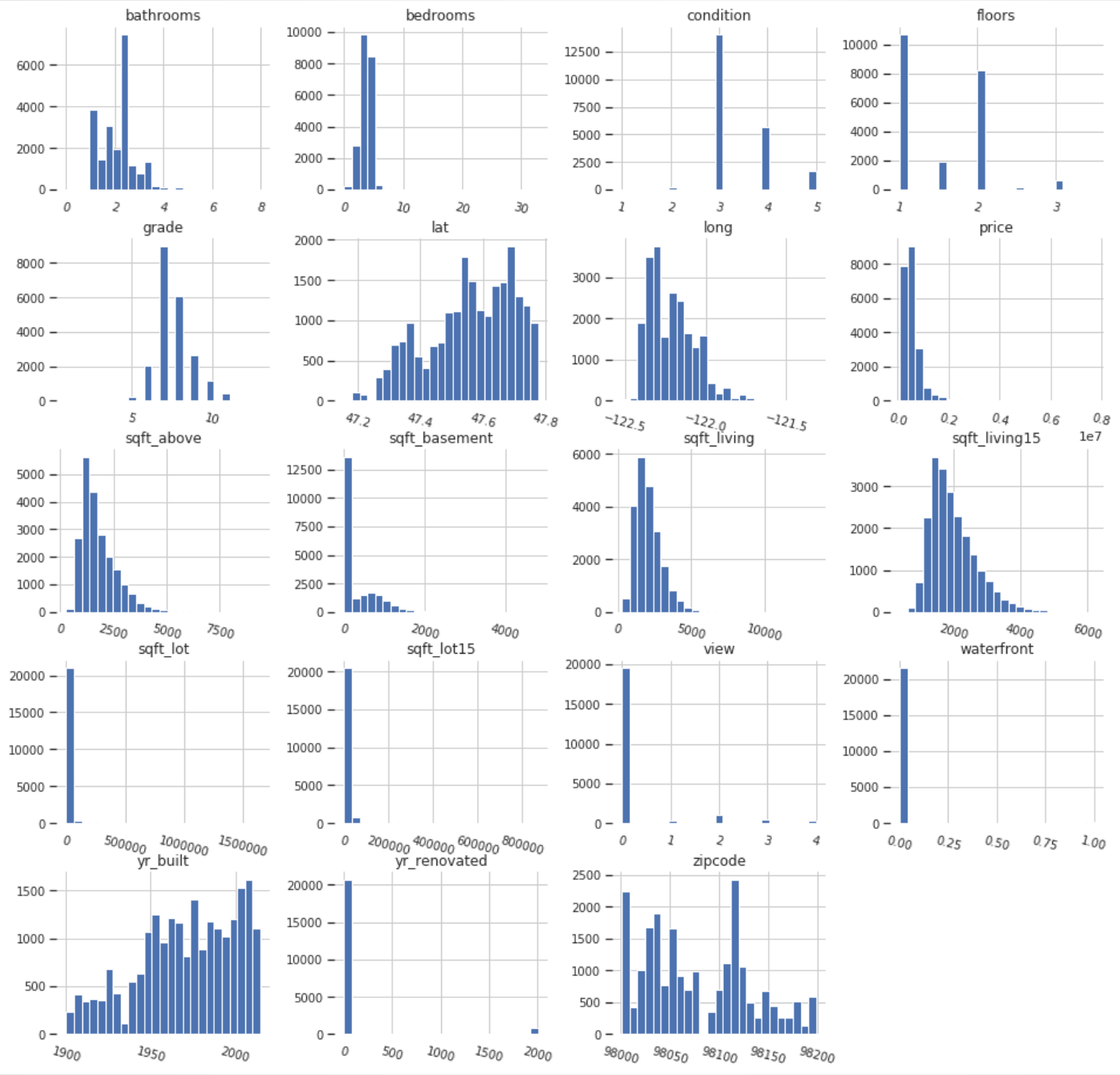
WHO BENEFITS FROM THIS …

- Housing Developers
- Individuals who will purchase house in the future
- Real estate company and brokers
- It can help the government estimate the price of housing in the future

EXPLORATORY DATA ANALYSIS

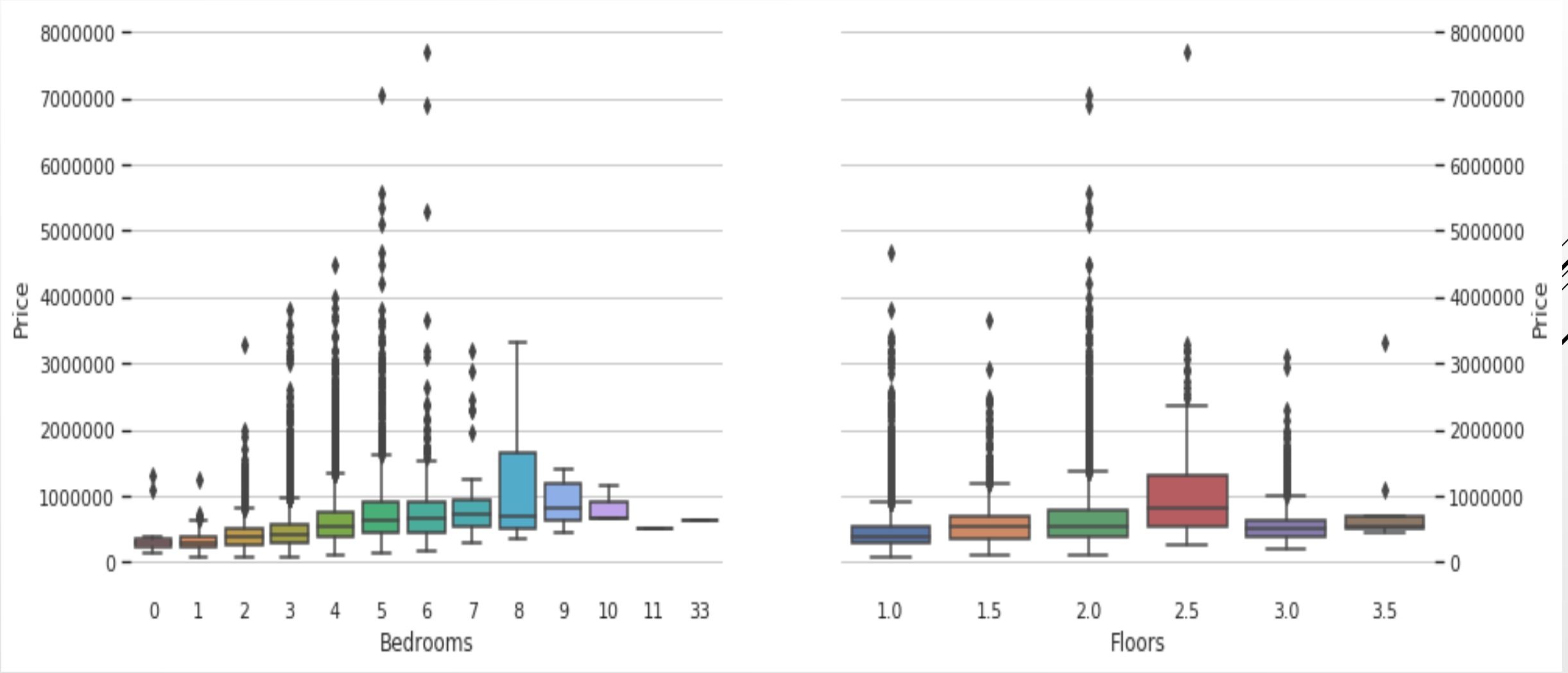
- I tried to use a diverse set of data visualization tools

- ***Histogram plots***
- ***Boxplots***
- ***Scatterplot***

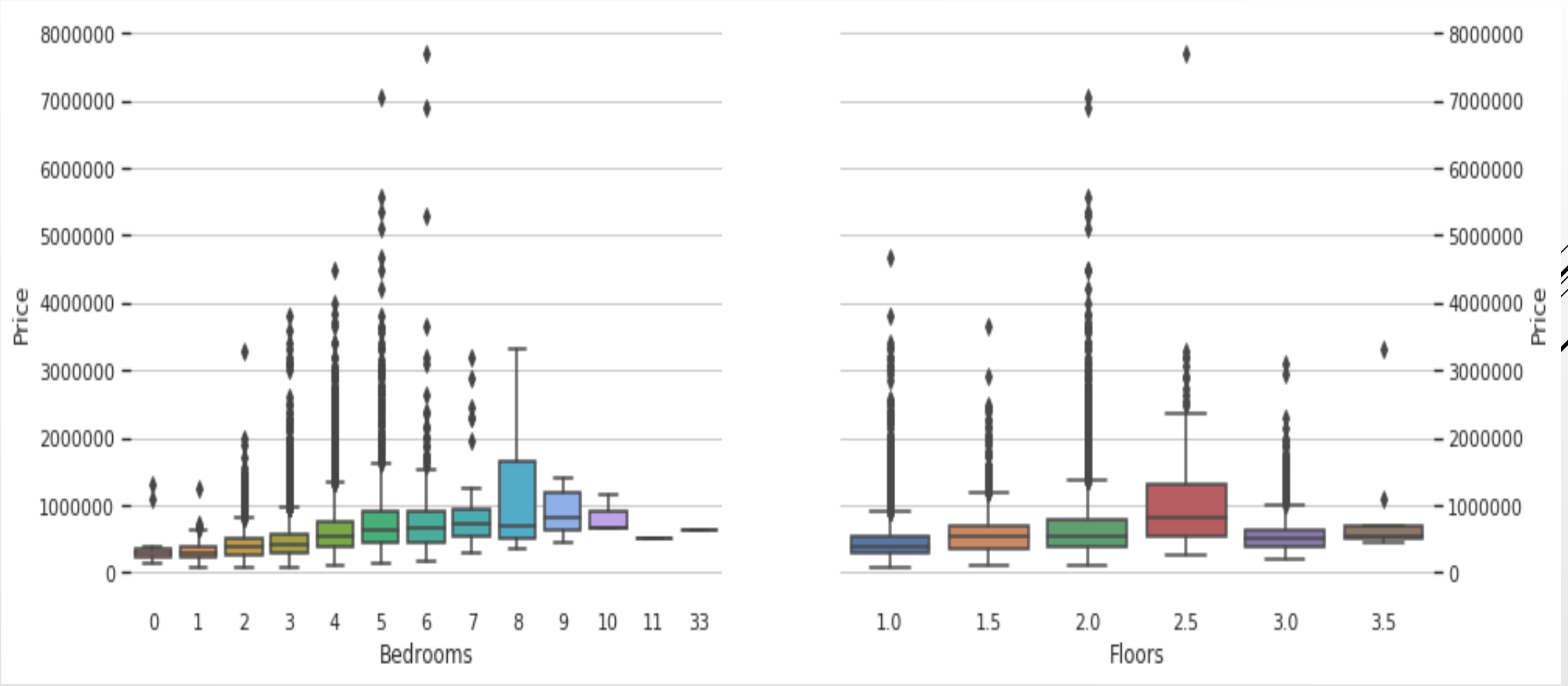


Histogram Plot

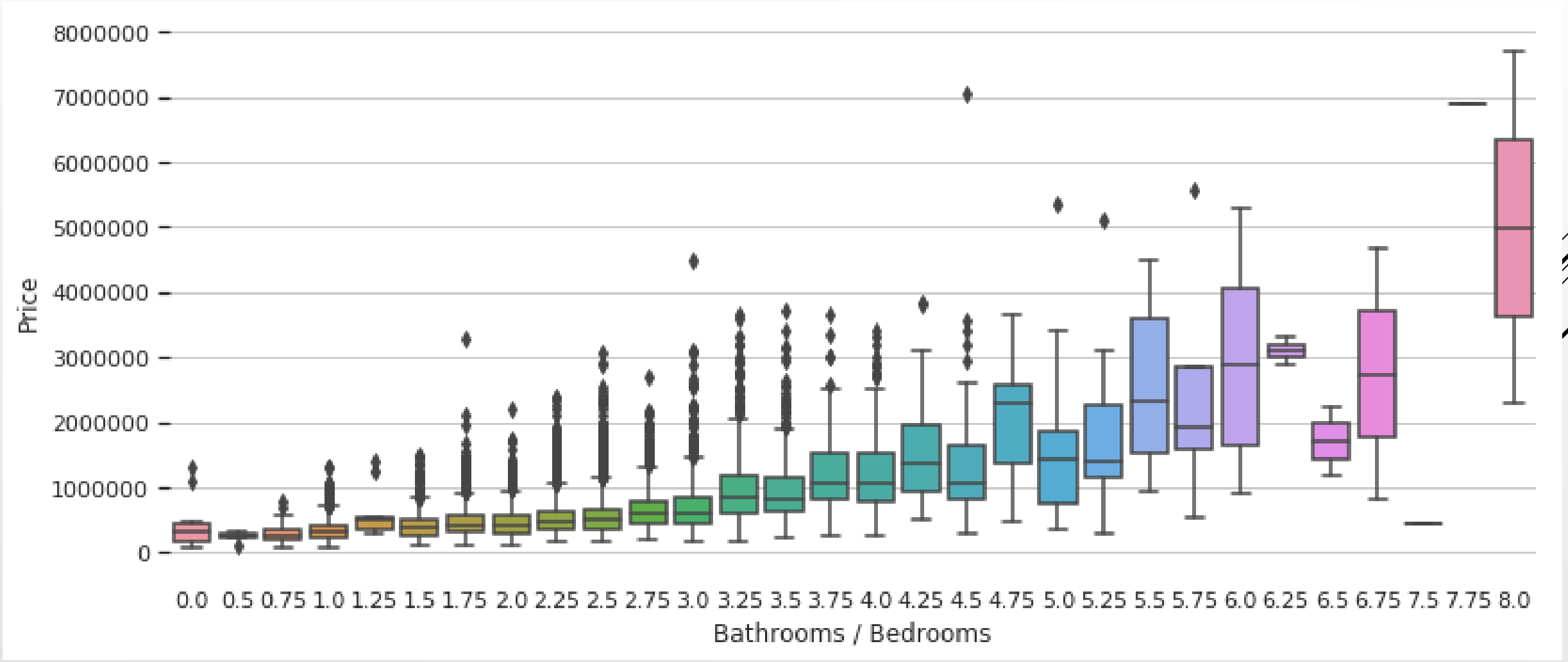
Boxplot 1


8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
Price
0
1
2
3
4
5
6
7
8
9
10
11
33
Bedrooms
1.0
1.5
2.0
2.5
3.0
3.5
Floors

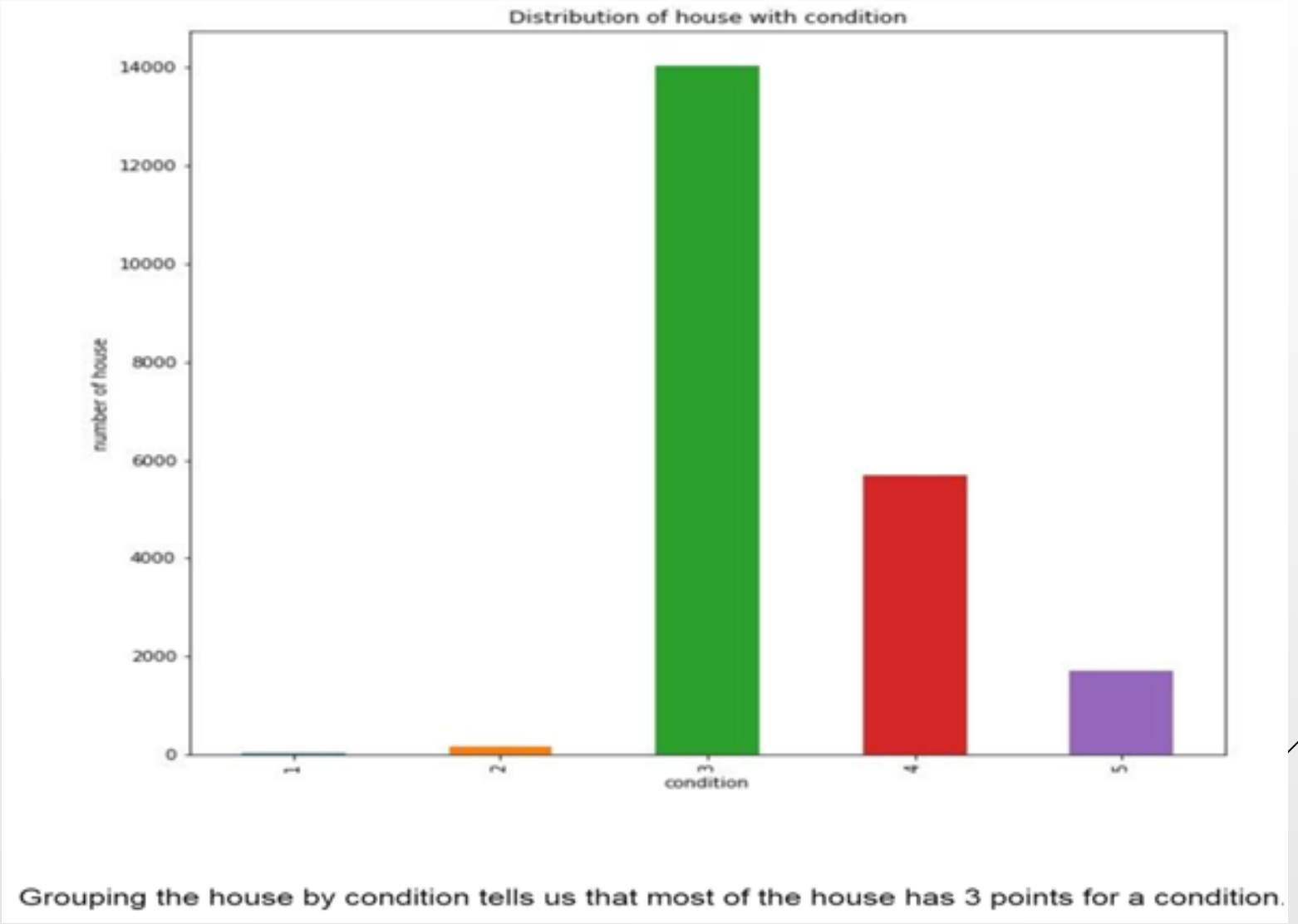
Boxplot 2


8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
Price
0 1 2 3 4 5 6 7 8 9 10 11 33
Bedrooms
1.0 1.5 2.0 2.5 3.0 3.5
Floors

Boxplot 3

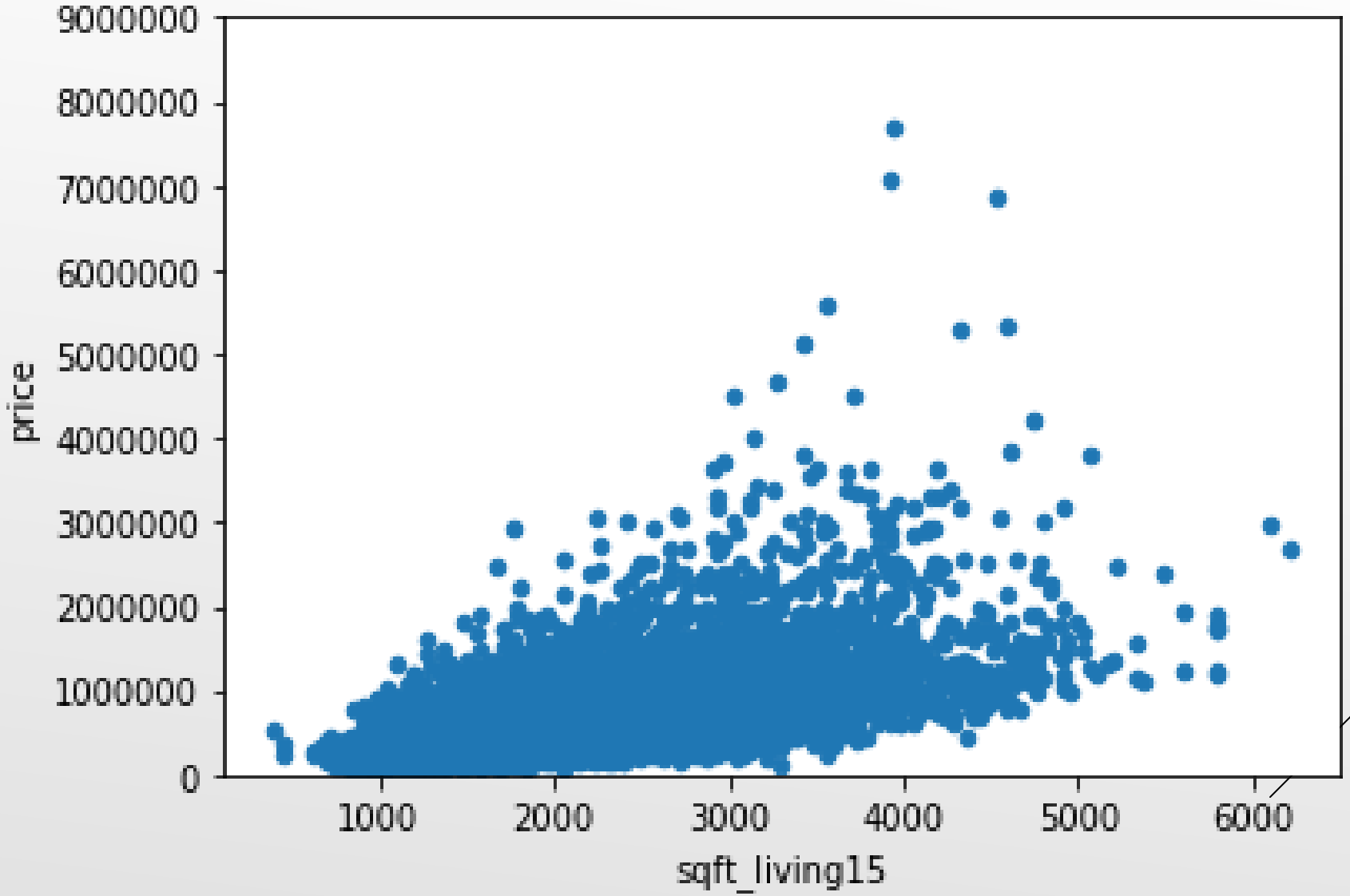

8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
Price
0.0 0.5 0.75 1.0 1.25 1.5 1.75 2.0 2.25 2.5 2.75 3.0 3.25 3.5 3.75 4.0 4.25 4.5 4.75 5.0 5.25 5.5 5.75 6.0 6.25 6.5 6.75 7.5 7.75 8.0
Bathrooms / Bedrooms

Boxplot 4



Grouping the house by condition tells us that most of the house has 3 points for a condition.

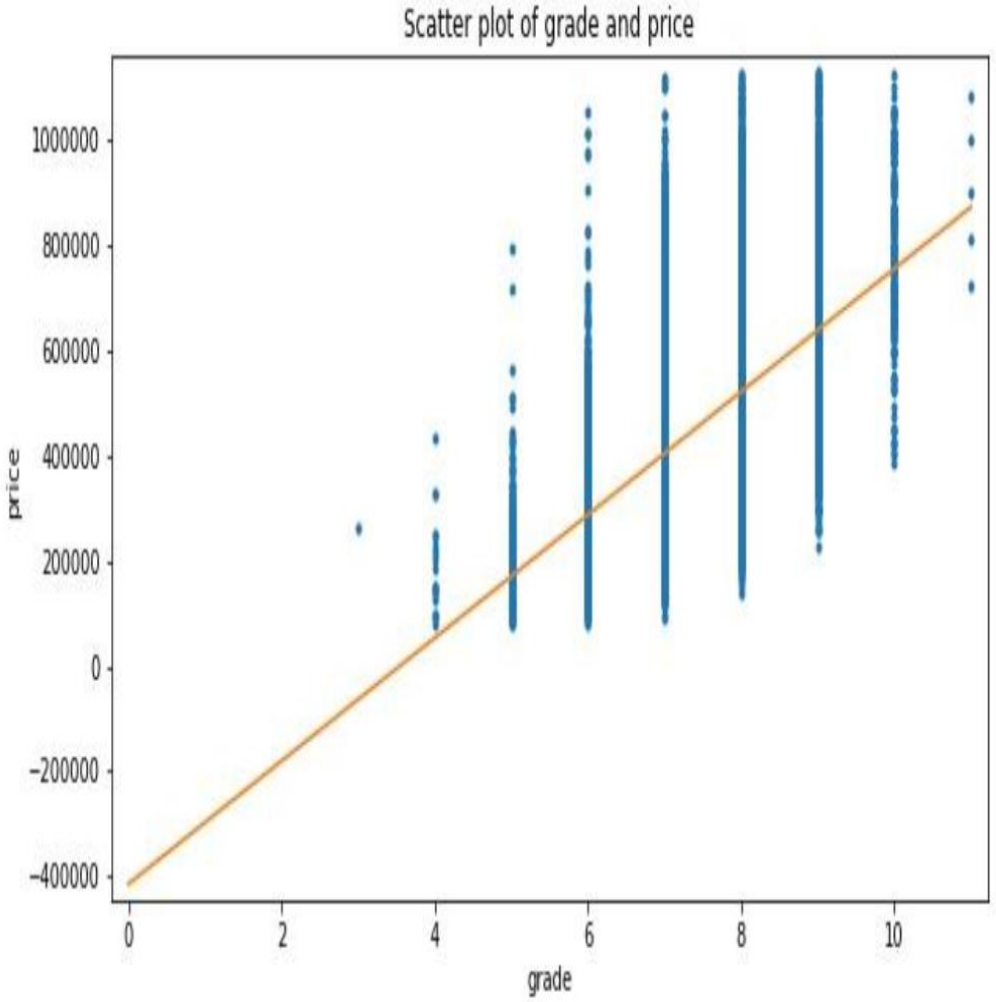
Scatter plot 4



Correlation b/n features and price

I calculated the correlation coefficient for all the variables to find the best predictors of the house price.

Features	CORRCOEFF	
sqft_lot15	-0.107535	Weak negative
sqft_lot	-0.089069	Weak negative
waterfront	0.055702	Very Weak positive
condition	0.078840	Very Weak positive
view	0.218874	Weak positive
bedrooms	0.235083	Weak positive
floors	0.238493	Weak positive
sqft_basement	0.239227	Weak positive
bathrooms	0.360725	Strong positive
sqft_above	0.403418	Strong positive
sqft_living15	0.439548	Strong positive
sqft_living	0.524052	Strong positive
grade	0.546210	Strong positive



MACHINE LEARNING MODELS

- Linear Regression
- Random Forest Regression

METRICS USED TO EVALUATE

- Root Mean Squared Error (RMSE)
- Mean squared error (*MSE*)
- Mean absolute error (*MAE*)

CONCLUSION

From the output, it is visible that the random forest algorithm is better at predicting house prices for the Kings County housing dataset, since the values of MAE, RMSE, and MSE for random forest algorithm are far less compared to the linear regression algorithm