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 IS THE AUDIENCE …

- 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

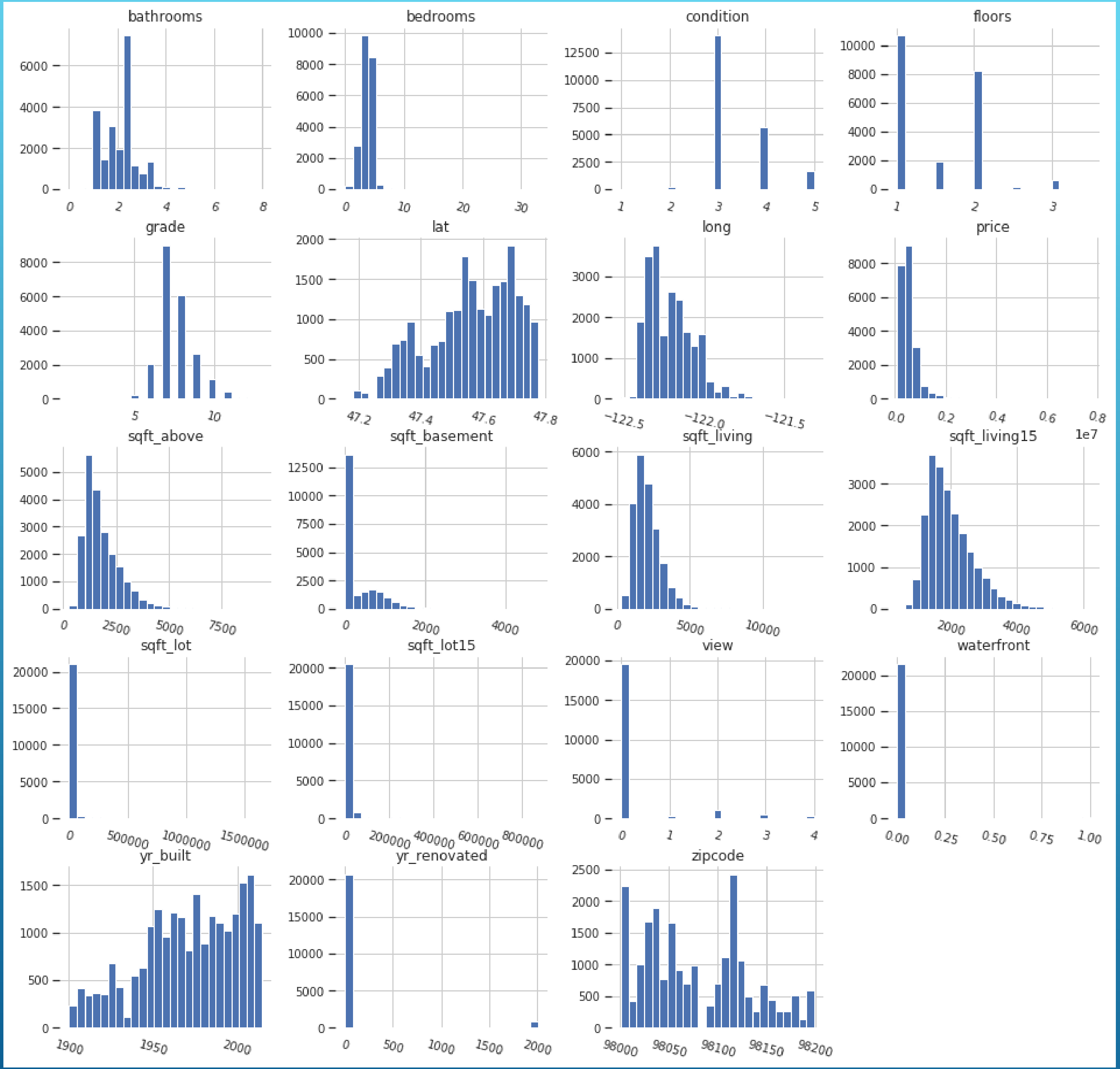
❑ ***Histogram plots***

❑ ***Boxplots***

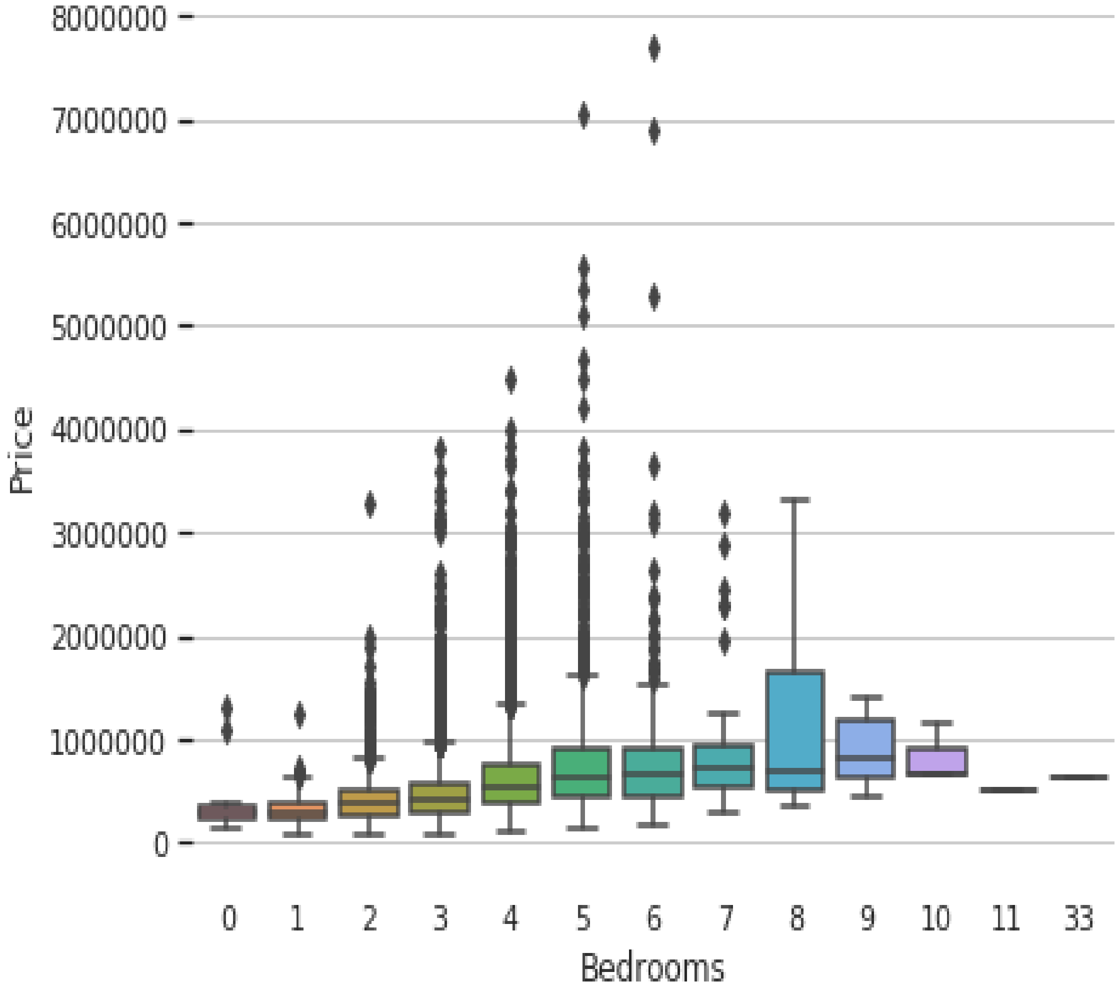
❑ ***Scatterplot***

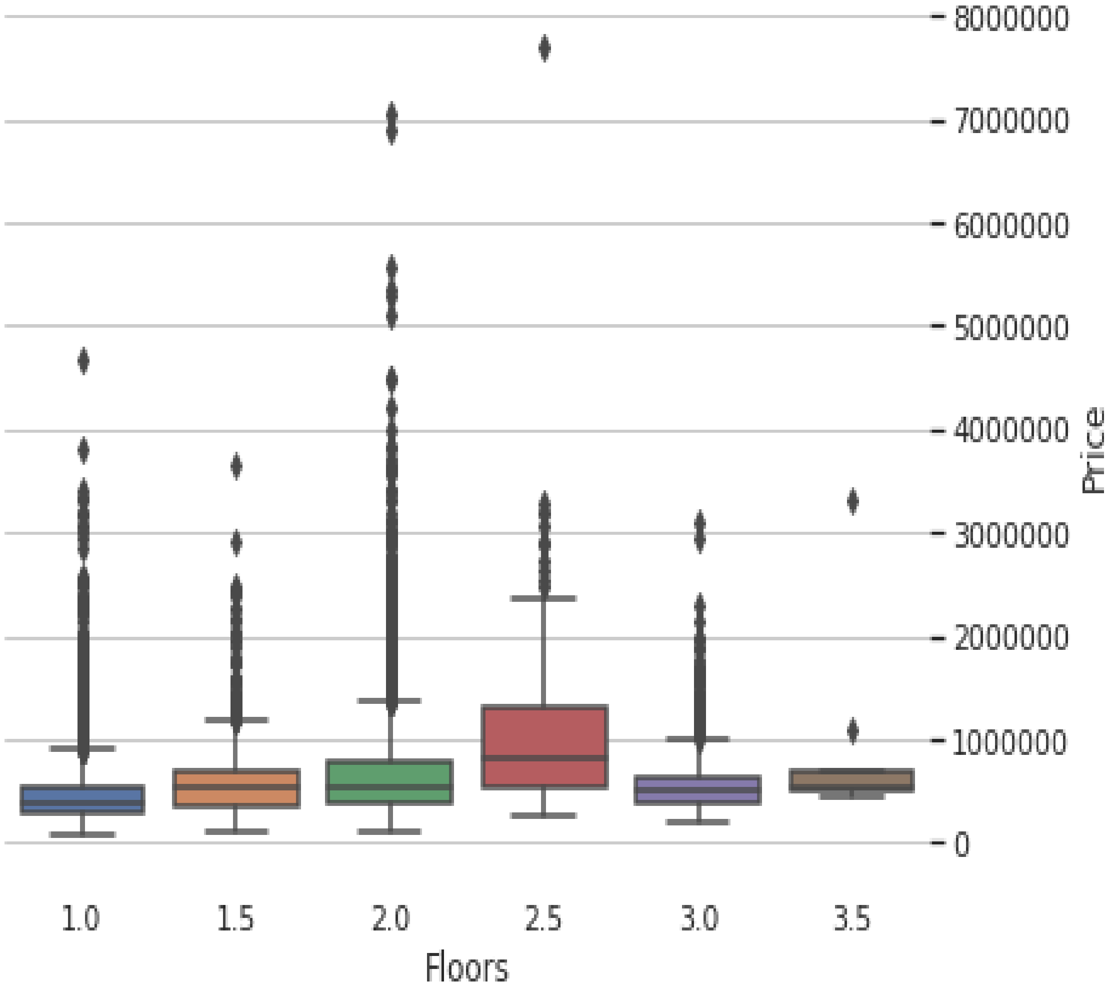
Histogram Plot

Correlations b/n features vs price

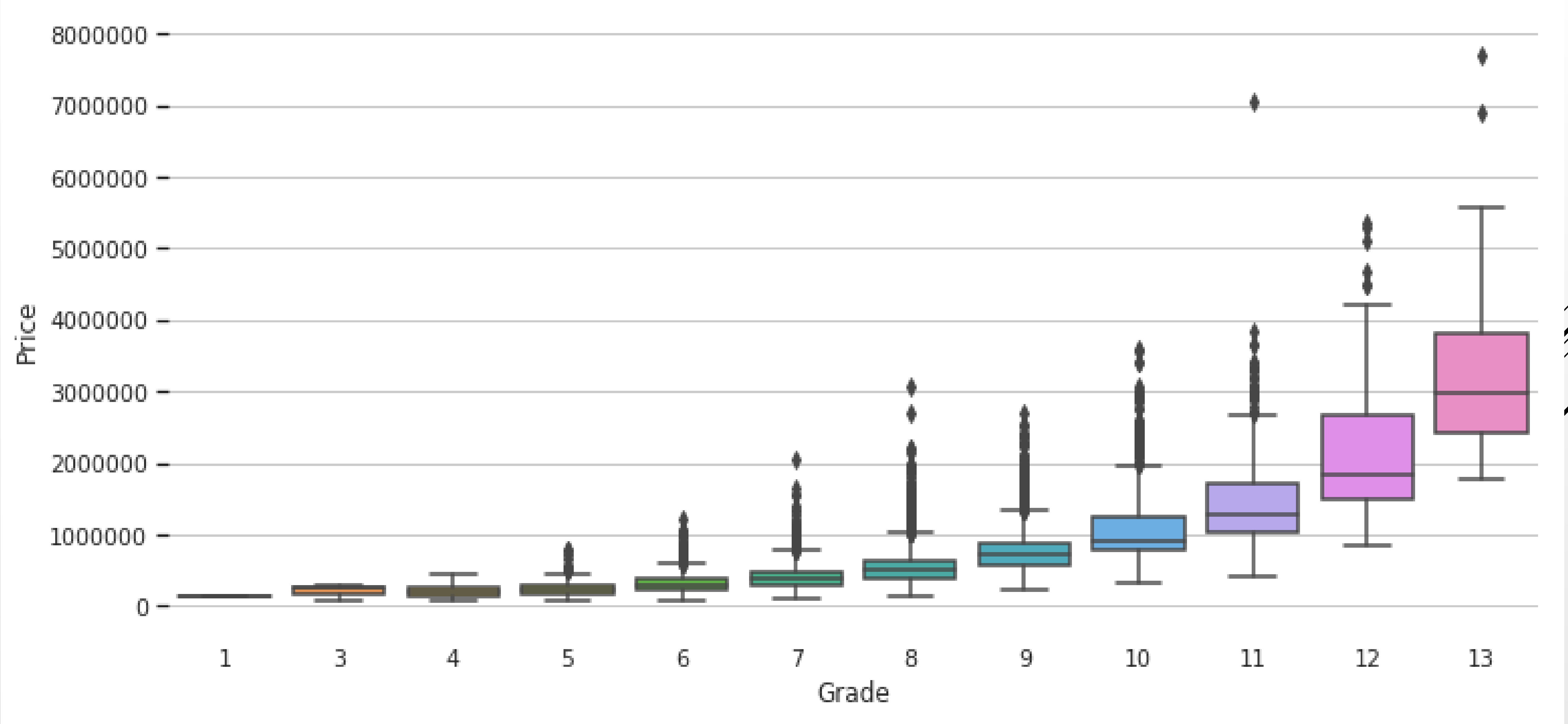


Data is cleaned and outliers were dropped

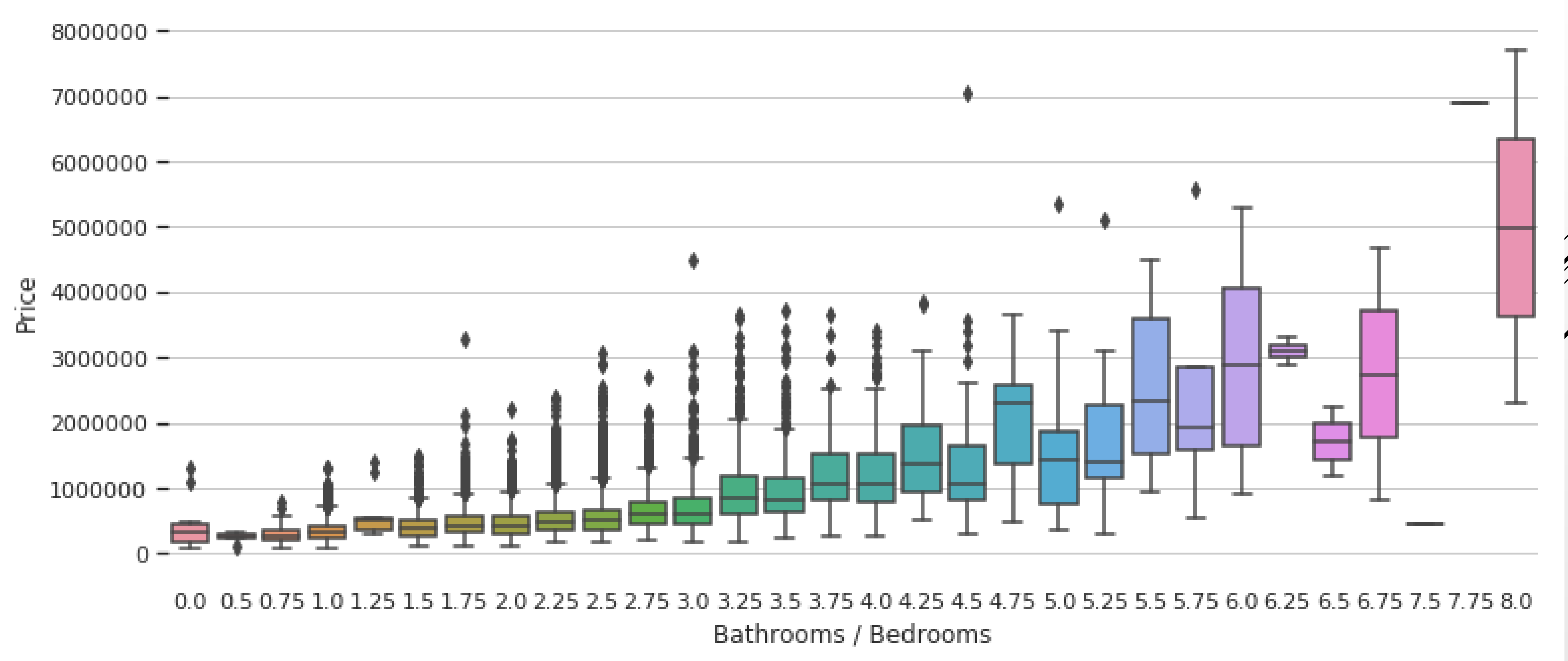




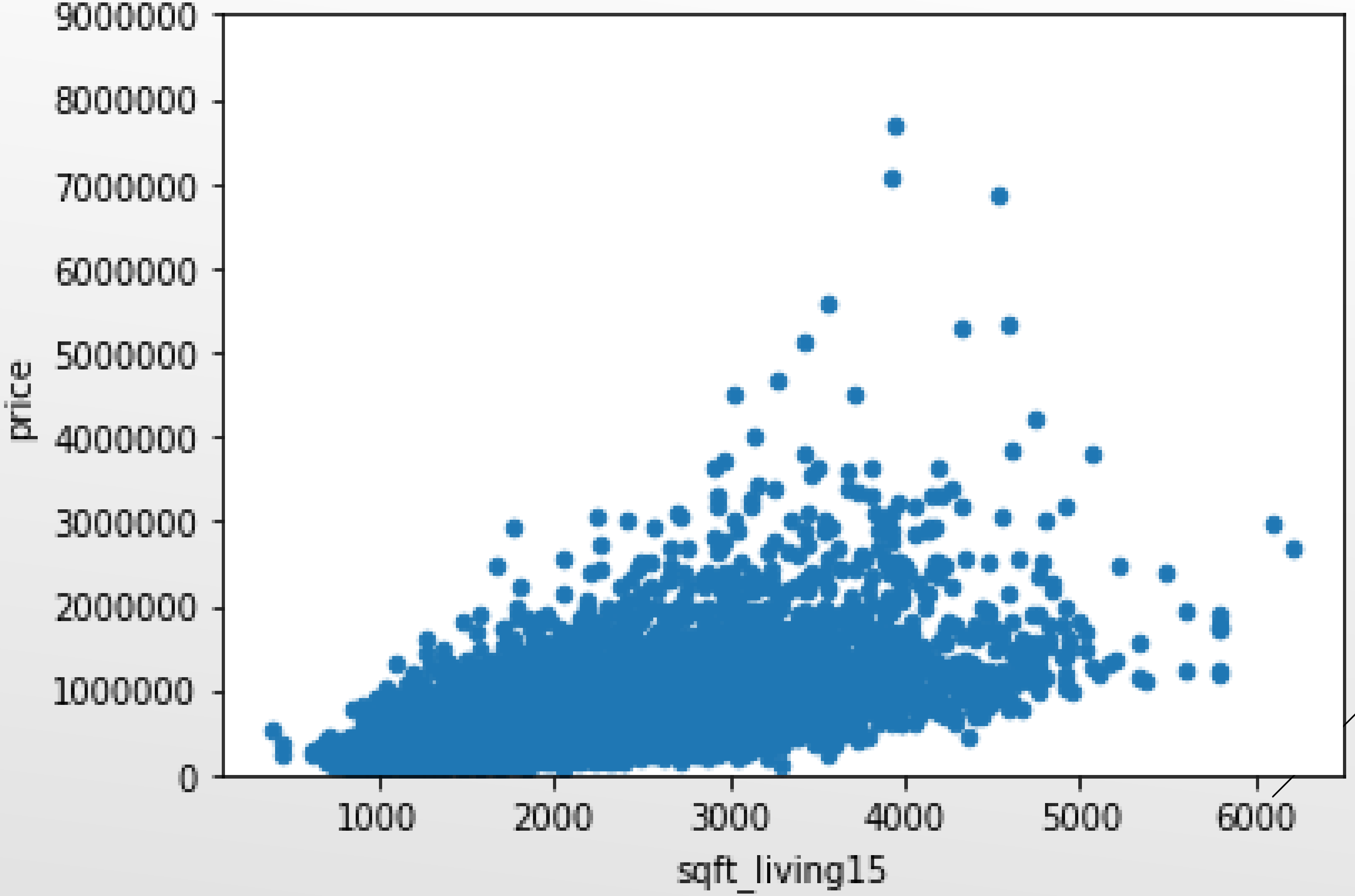
Boxplot : correlation Bedrooms and Floors



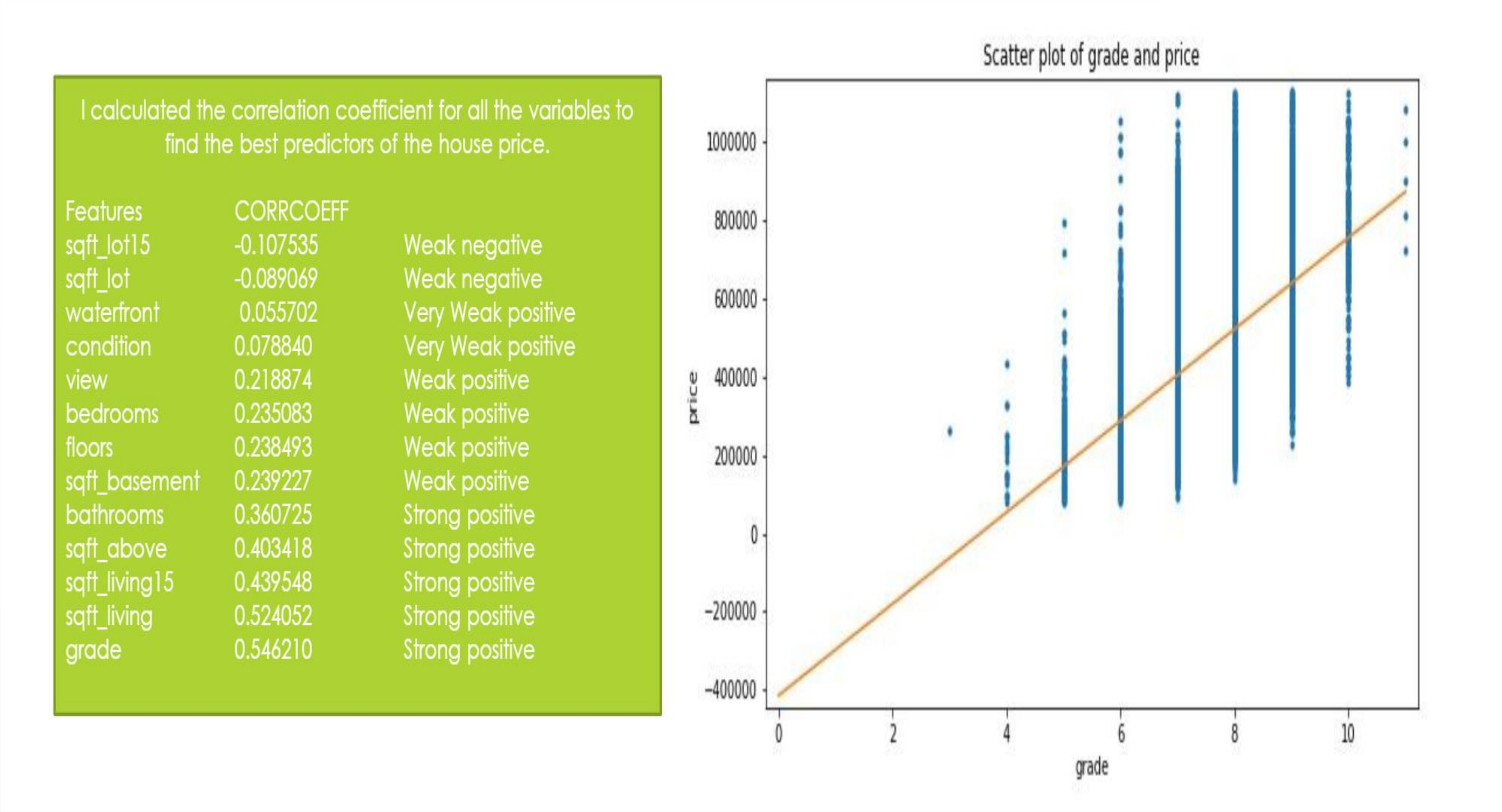
Boxplot : correlation b/n Bathrooms & Bedrooms



Scatter plot for Sqft_living15



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

Scatter plot of grade and price

MACHINE LEARNING MODELS

- Linear Regression
- Random Forest Regression
- Gradient Boosting Regressor
- Decision TreeRegressor

METRICS USED TO EVALUATE

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

METRICS ...CONTINUED

- Accuracy score
- Variance score

CONCLUSION

- **Gradient Boosting model better model**
- **Accuracy score of 88.84%**
- **R-squared of 0.828**

And all the metrics suggests that Gradient Boosting model has better performance. Therefore, it is inferred that **Gradient Boosting** is the suitable model for this dataset.