DATA SCIENCE CAREER TRACK CAPSTONE PROJECT-1

King County

House Price Prediction

Gokmen Oran
March 2019

Table of Contents

Problem Statement

"Is it possible to predict the sale price of a house from information about that house provided in the dataset, such as square footage of the home, number of bedrooms, number of bathrooms, number of floors, condition, grade, etc?"

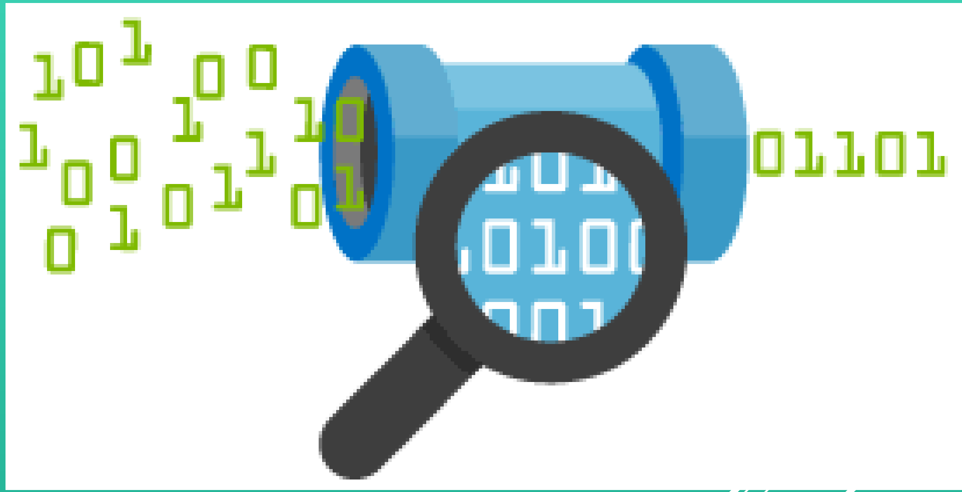
Data Specifications

- Houses sold in King County, Seattle, Washington between May 2014 and May 2015.
- 21613 observations and 21 features
- No missing data
- Uploaded to the Kaggle website by the user harlfoxem (Data Source: https://www.kaggle.com/harlfoxem/housesalesprediction).

King County

Data Wrangling

- “View”, “Sqft_living15”, “Sqft_lot15” features – further investigation from King County official website
- 177 duplicated rows - kept the last entries, dropped the first ones
- “id”, “date”, “lat”, “long” columns ➔ dropped
- Removed 16 observations with zero bathroom or zero bedroom



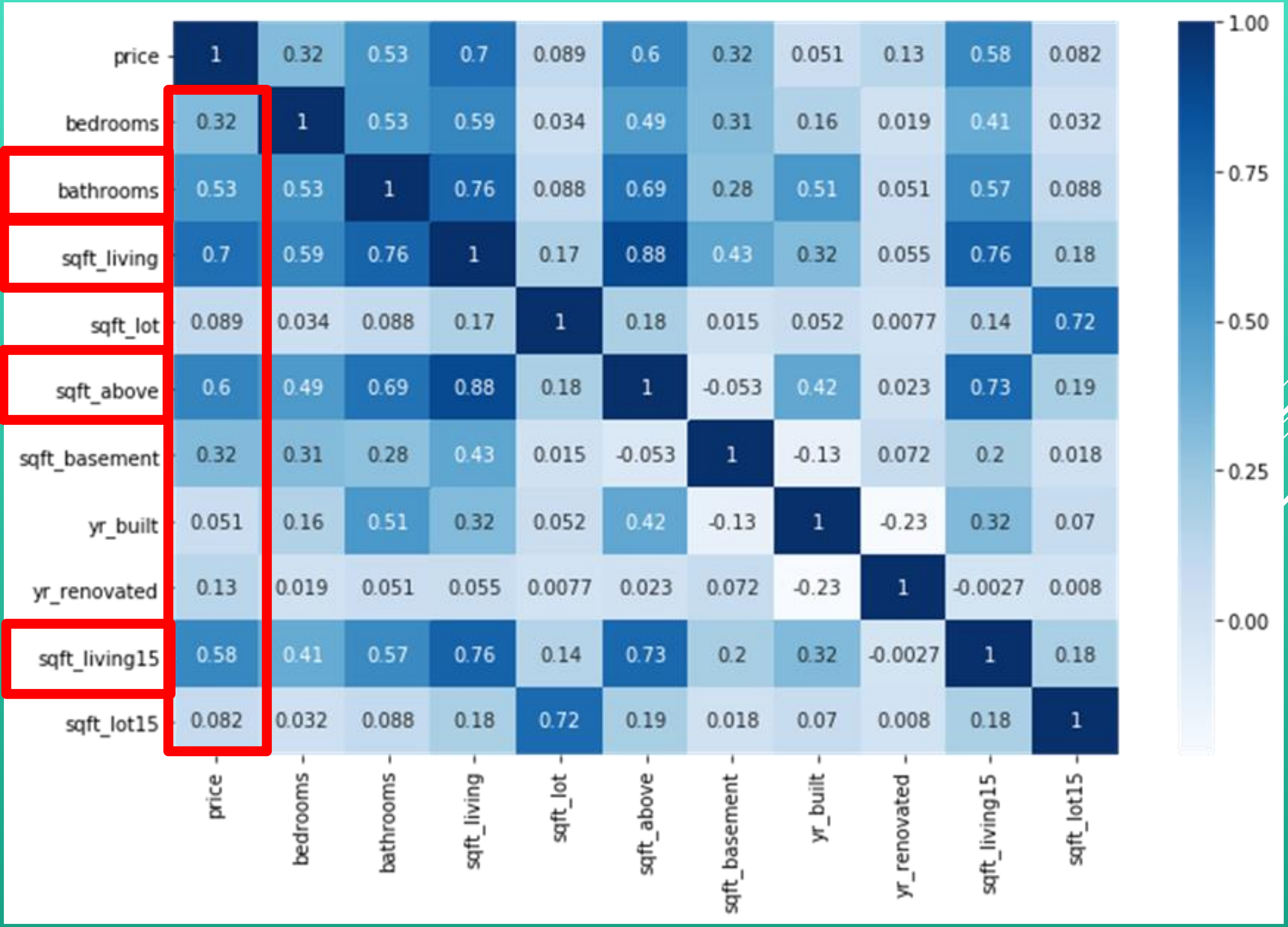
Data Wrangling

- Removed 1 observation with 33 bedrooms
- “floors”, “waterfront”, “view”, “condition”, “grade” columns ➔ **category**
- “zipcode” column ➔ **string**

- 21419 rows and 17 columns left in the dataset
- Datatypes: category(5), float64(2), int64(9), object(1)

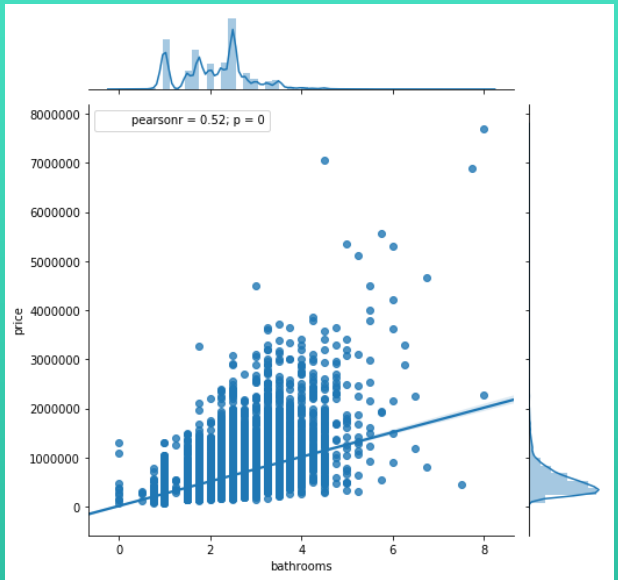
Exploratory Data Analysis

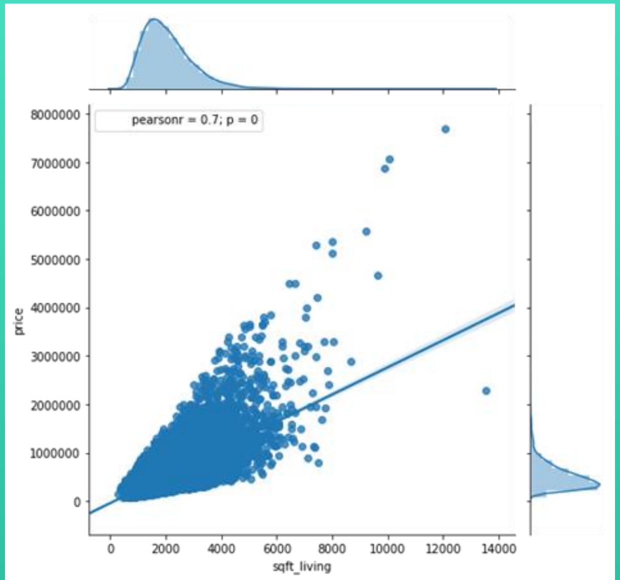
Correlation between the continuous variables

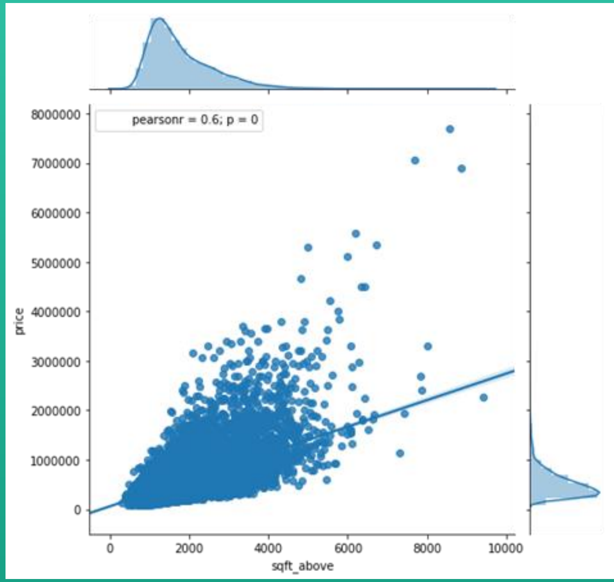


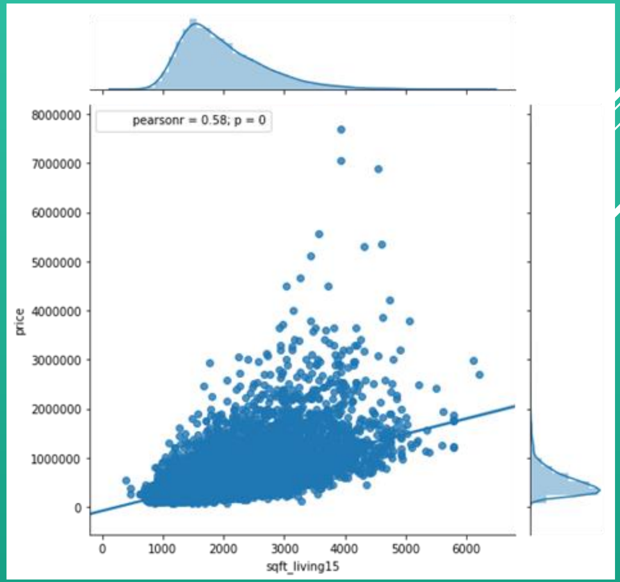
Exploratory Data Analysis

Correlation between the continuous variables









Exploratory Data Analysis

- There are many zeros in the “sqft_basement” and “yr_renovated” variable.

- Created new two columns (**'basement_present', 'renovated'**), and changed their types into category.

- If the house has basement; **'basement_present'** ➔ 1, **otherwise** ➔ 0

- If the house was renovated **'renovated‘ ➔ 1, otherwise ➔ 0**

Exploratory Data Analysis

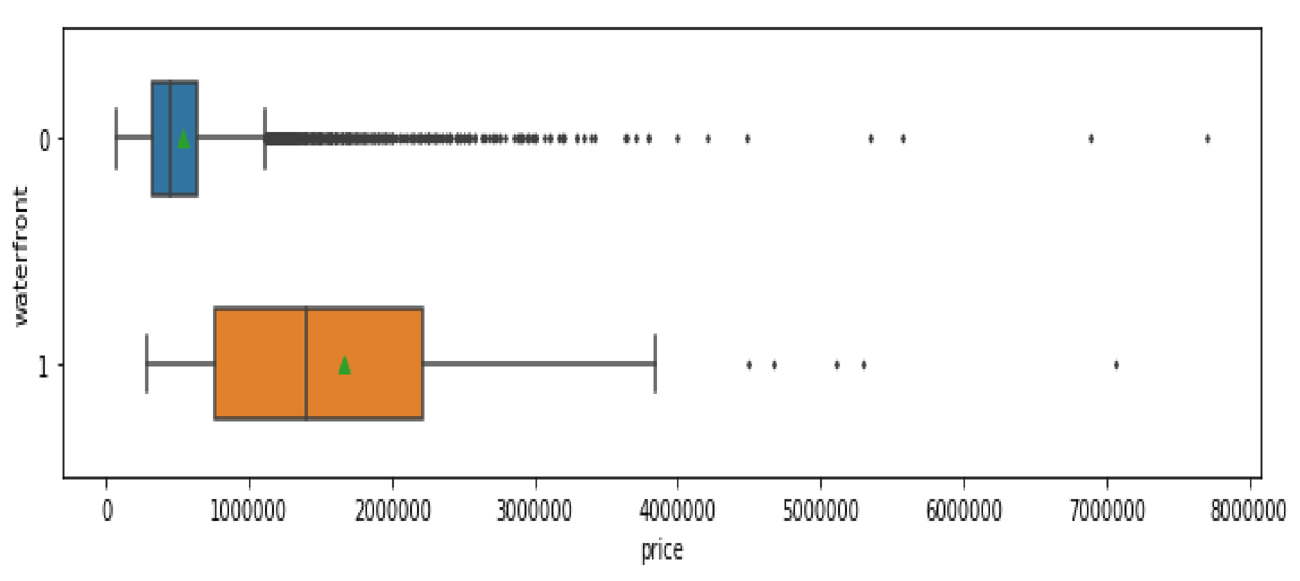
Relationship between ‘house price’ and the categorical variables ('waterfront', 'basement_present', 'renovated',

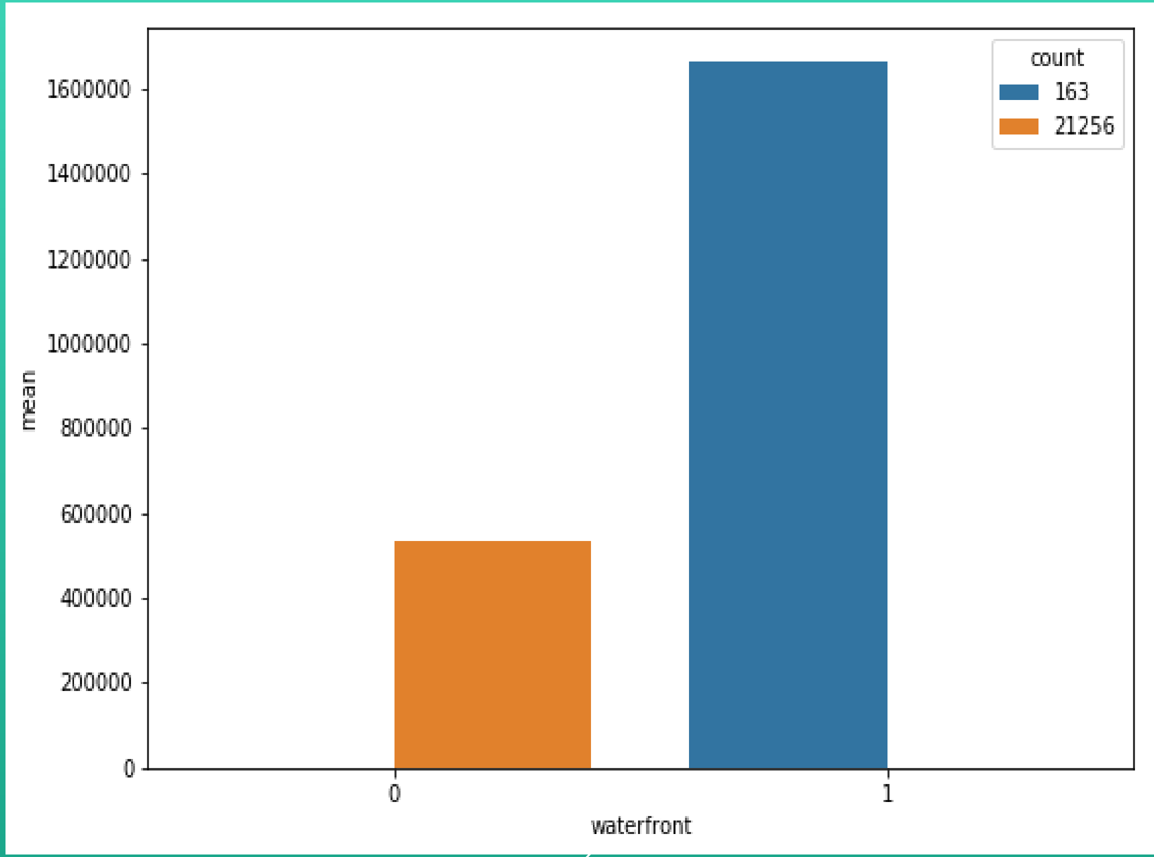
point-biserial correlation

	price
waterfront	0.26
basement_present	0.18
renovated	0.12

Exploratory Data Analysis

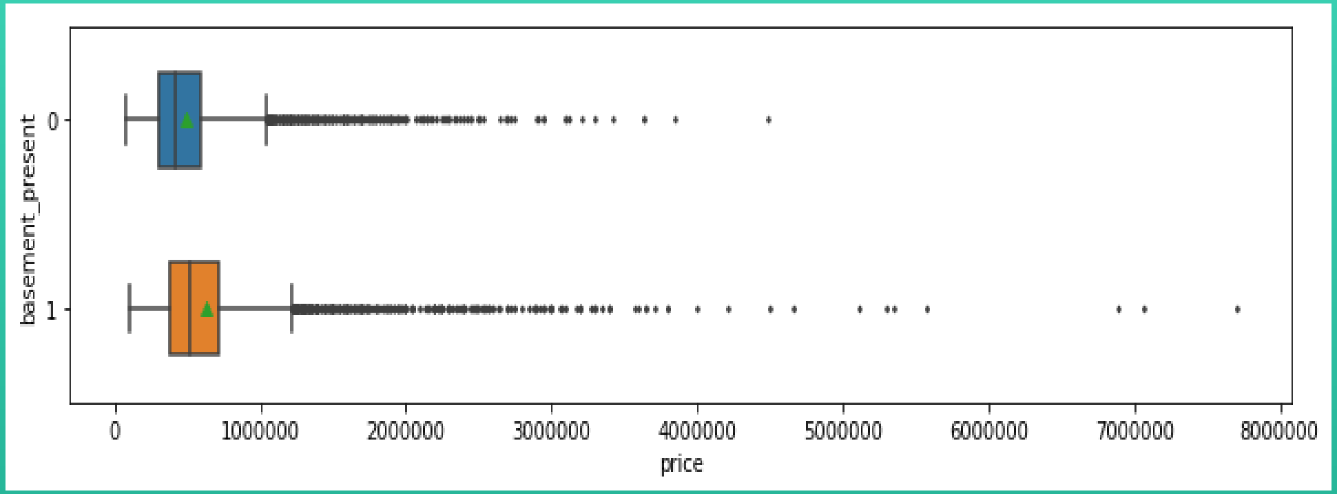
‘house price’ and 'waterfront‘ (r = 0.26)

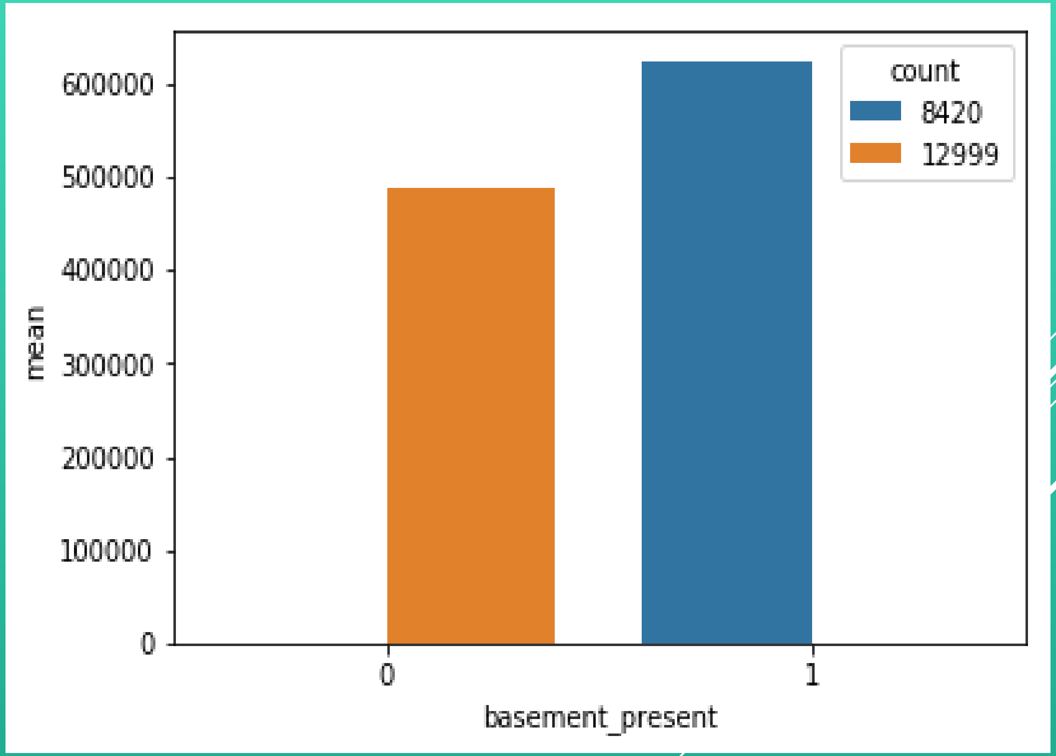




Exploratory Data Analysis

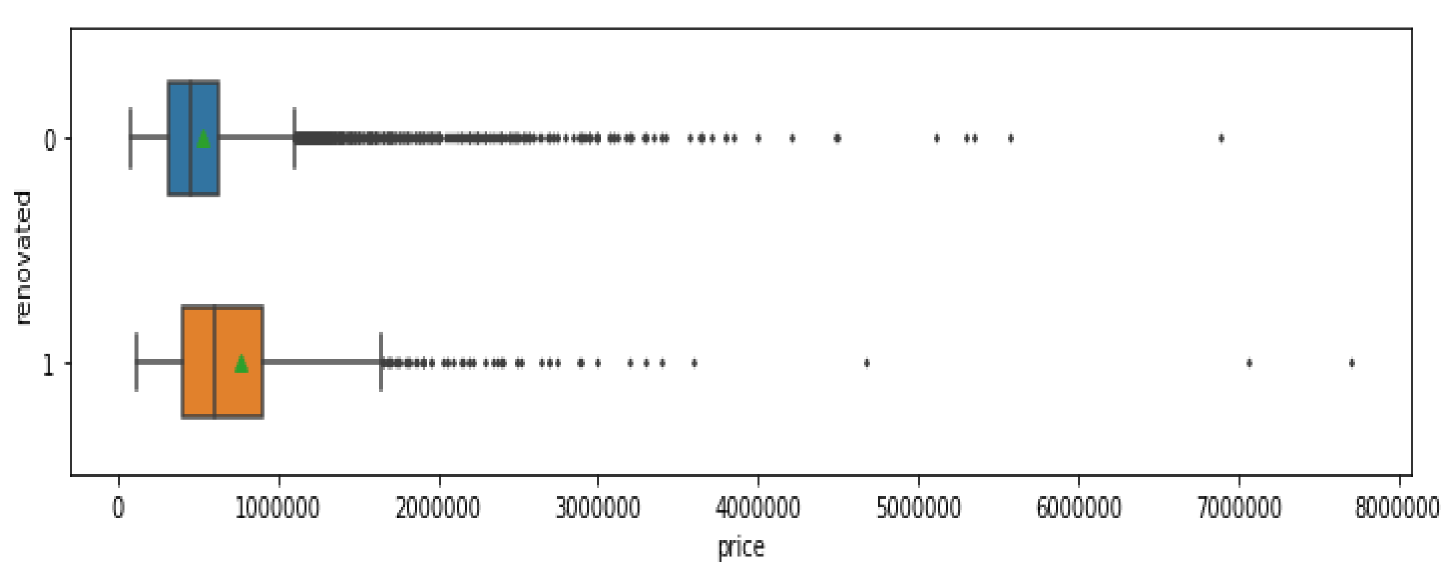
‘house price’ and 'basement_present’ (r = 0.18)

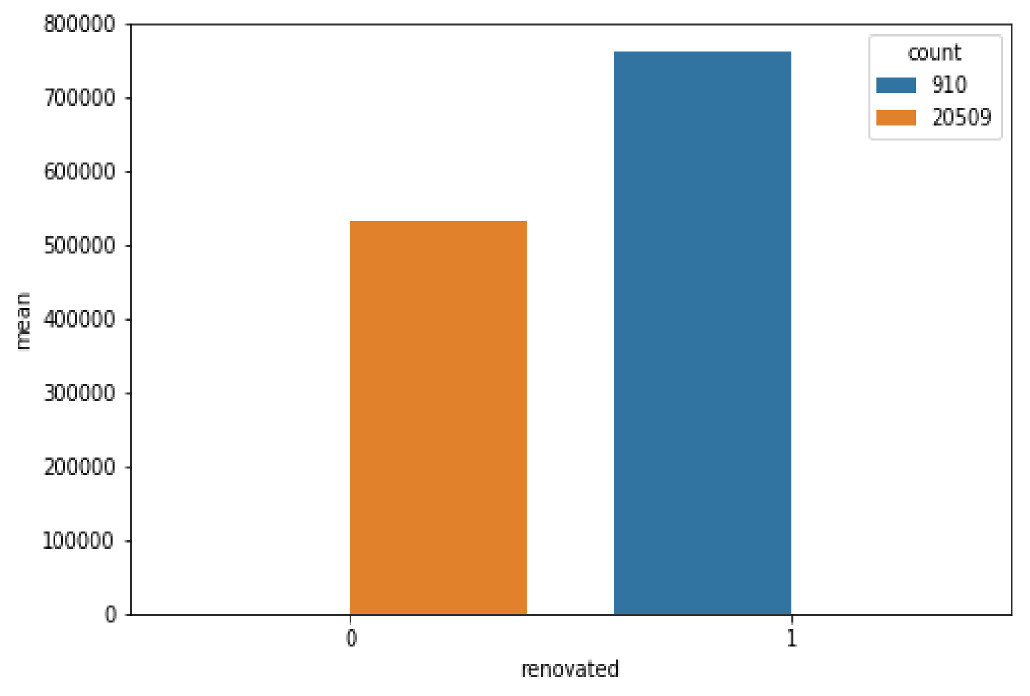




Exploratory Data Analysis

‘house price’ and ‘renovated’ (r = 0.12)





Exploratory Data Analysis

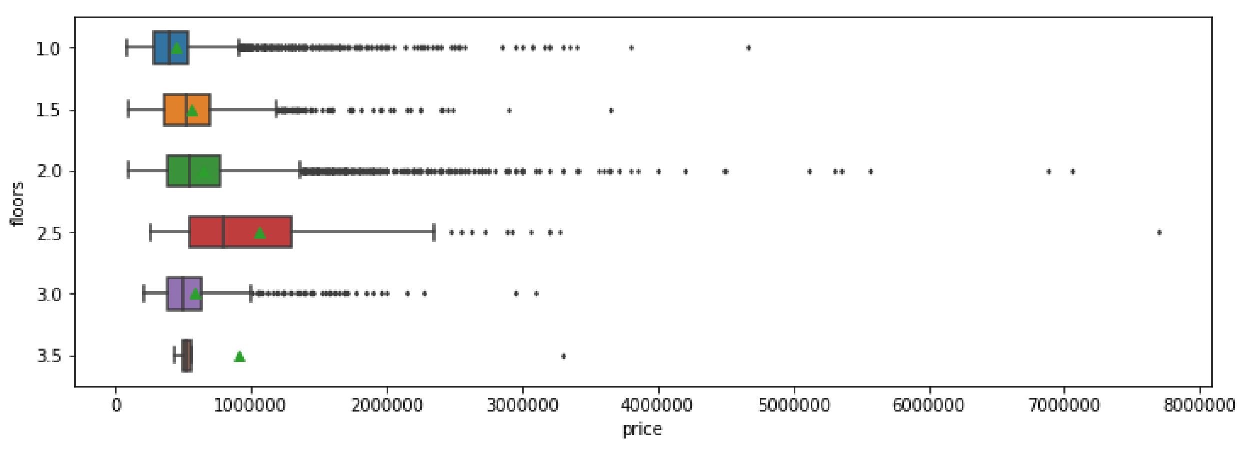
Spearman's rank-order correlation

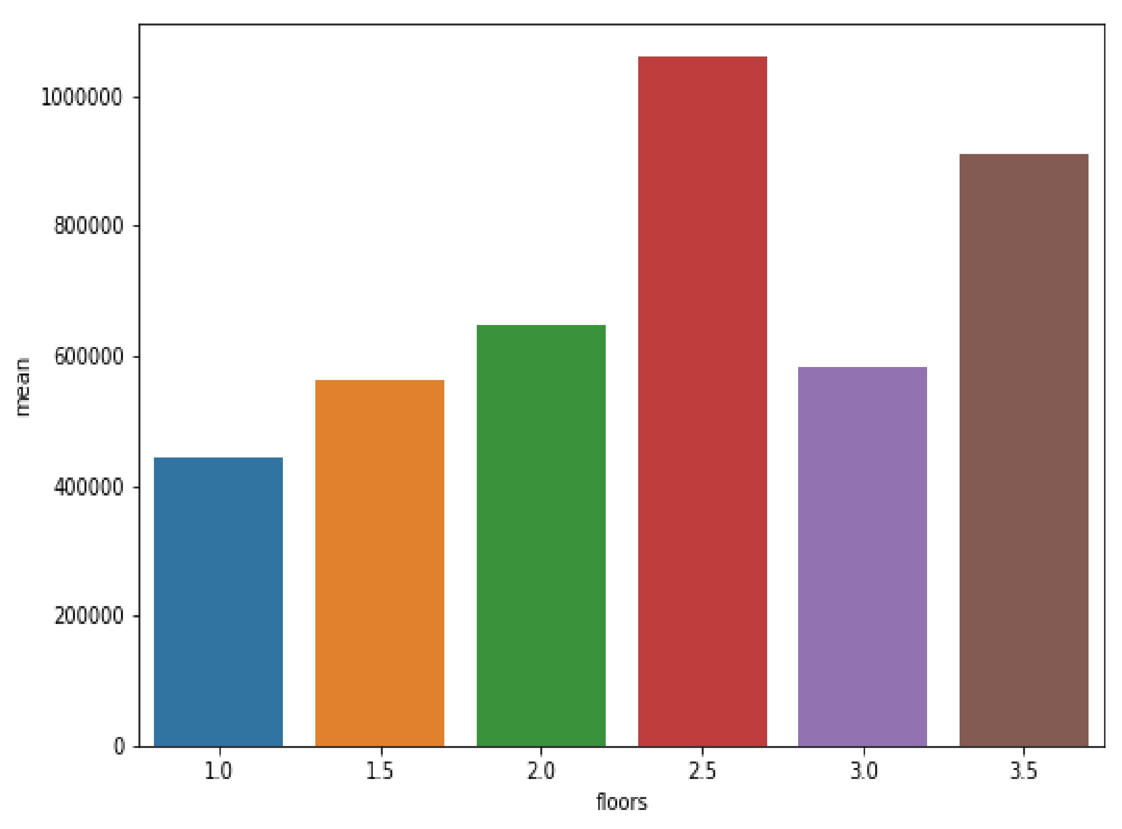
Relationship between 'house price' and the categorical variables ('floors', 'view', 'condition', 'grade')

	price
floors	0.32
view	0.29
condition	0.016
grade	0.656

Exploratory Data Analysis

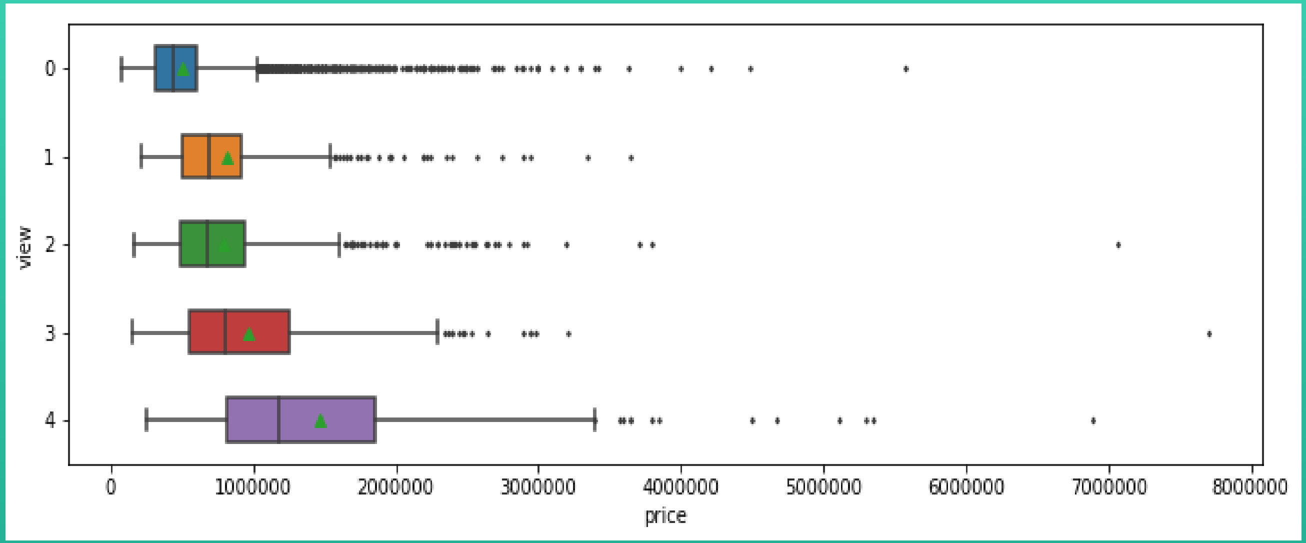
‘house price’ and ‘floors’ (r = 0.32)

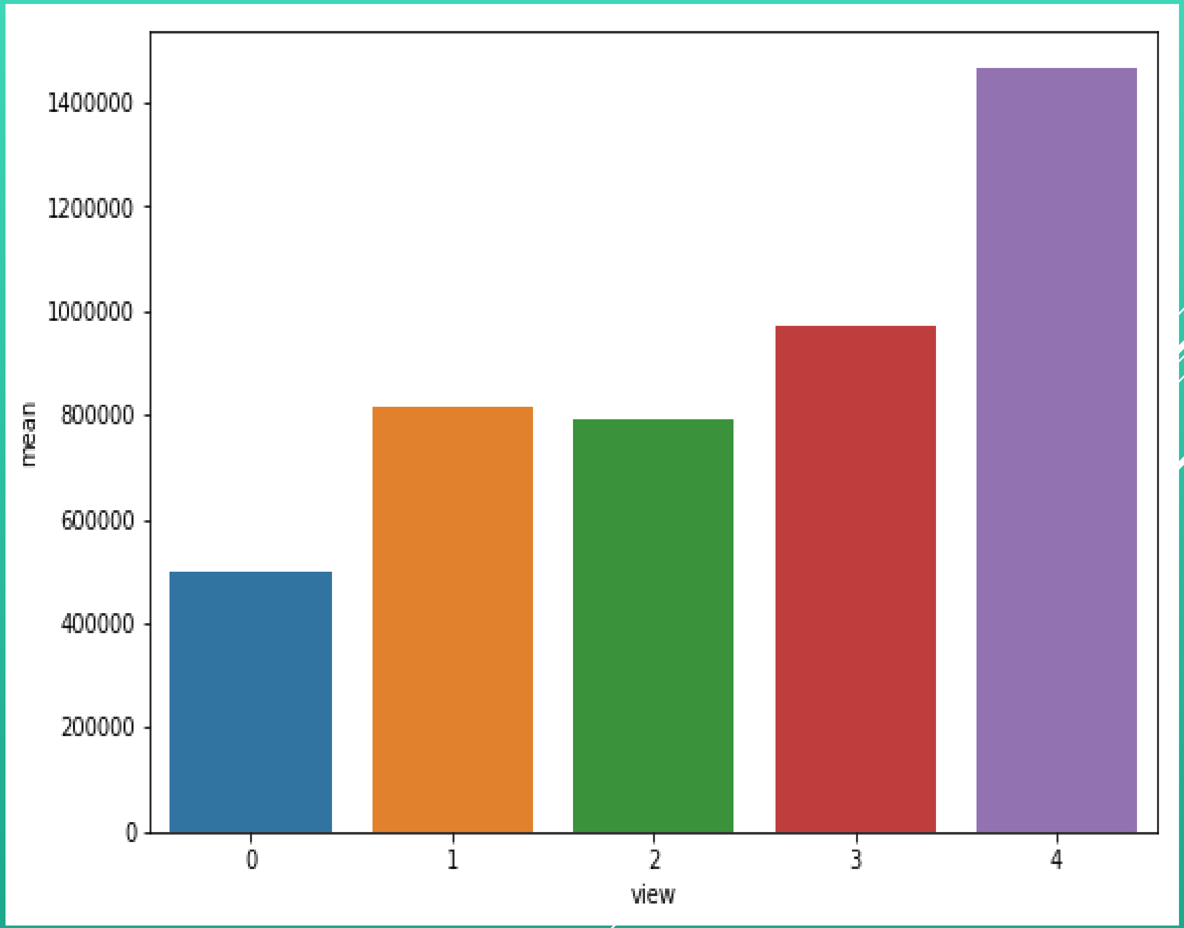




Exploratory Data Analysis

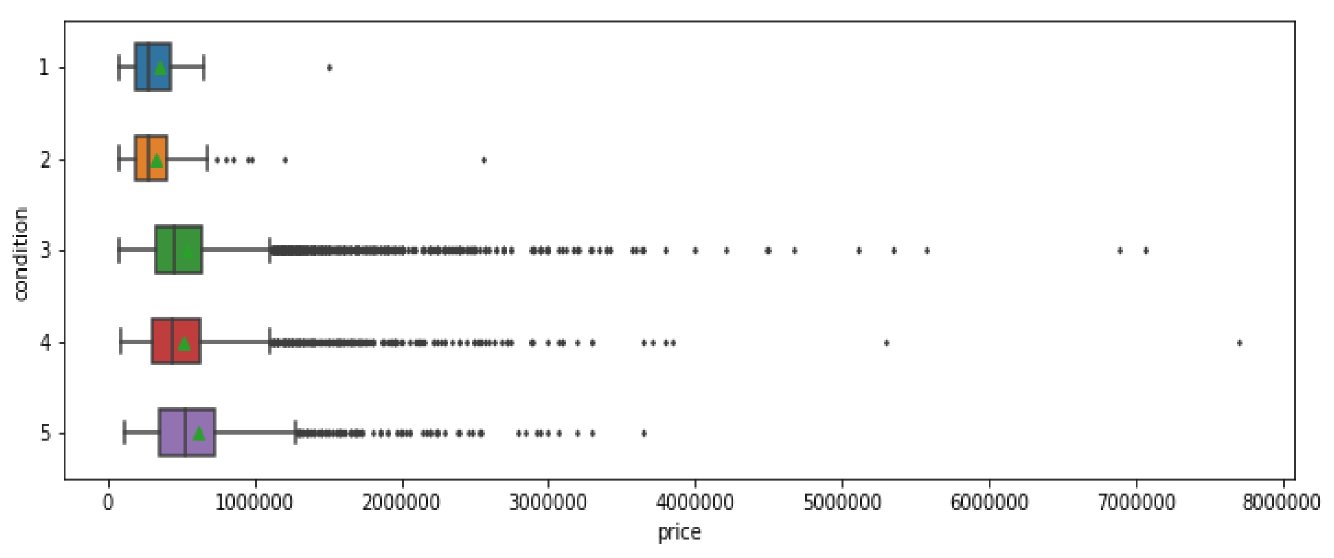
‘house price’ and ‘view’ (r = 0.29)

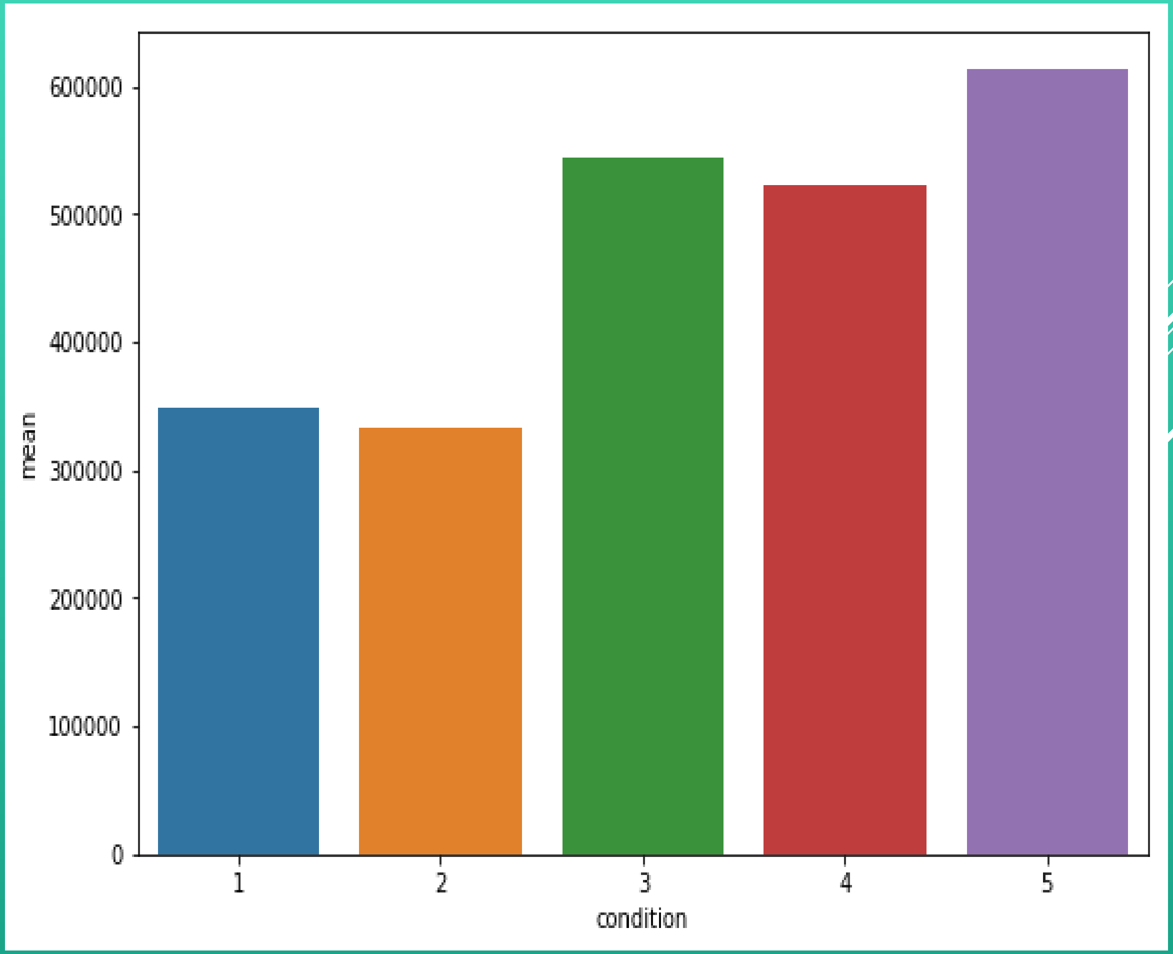




Exploratory Data Analysis

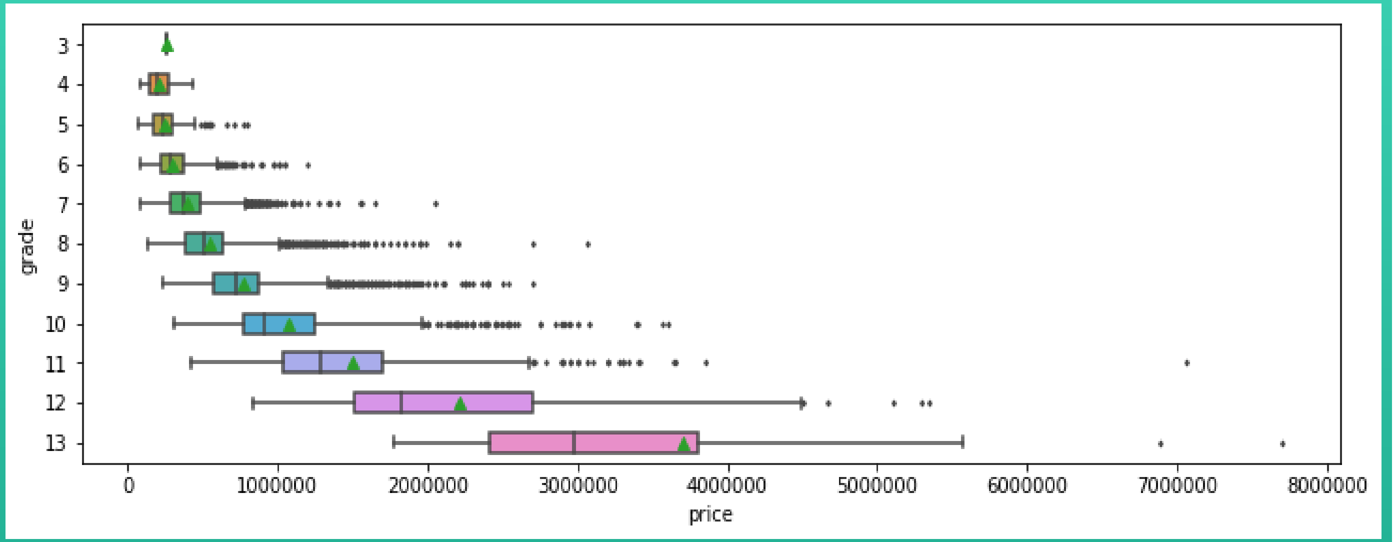
'house price' and 'condition' (r = 0.016)

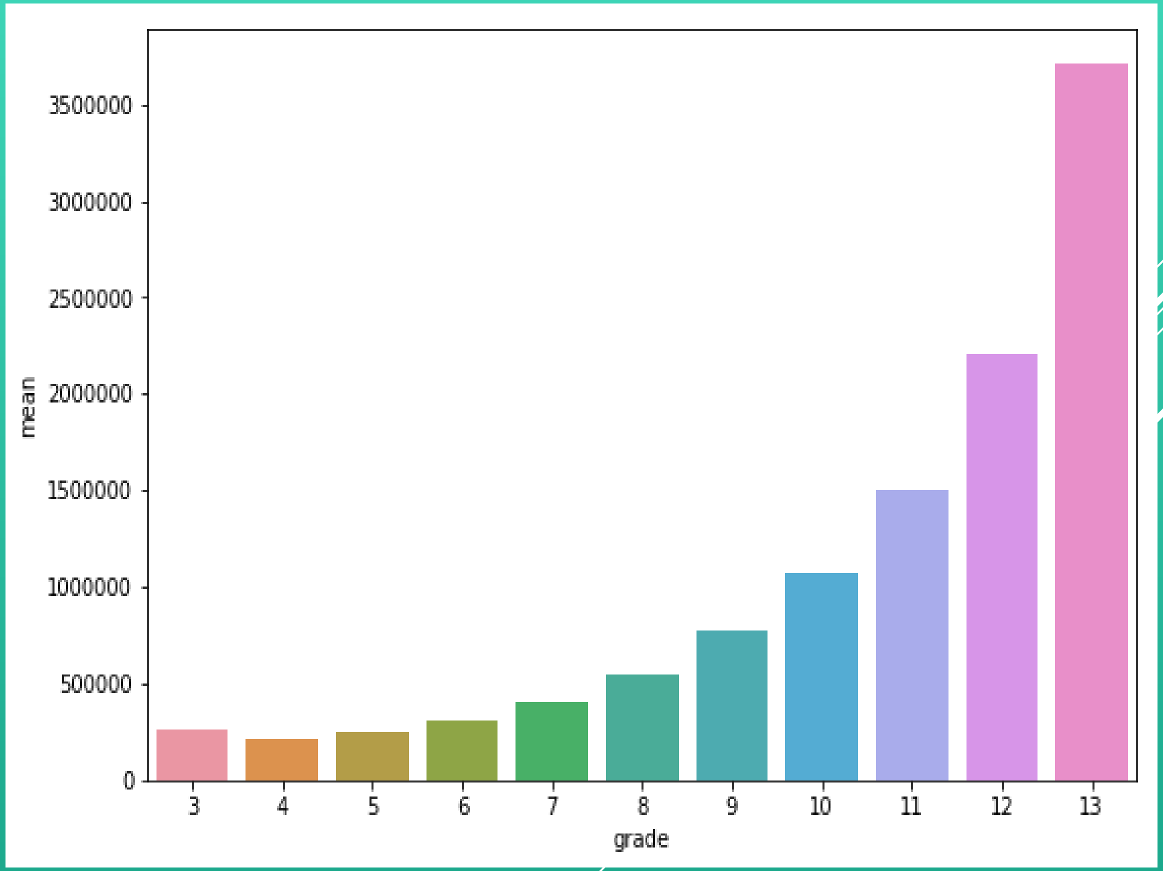




Exploratory Data Analysis

‘house price’ and ‘grade’ (r = 0.656)





Exploratory Data Analysis

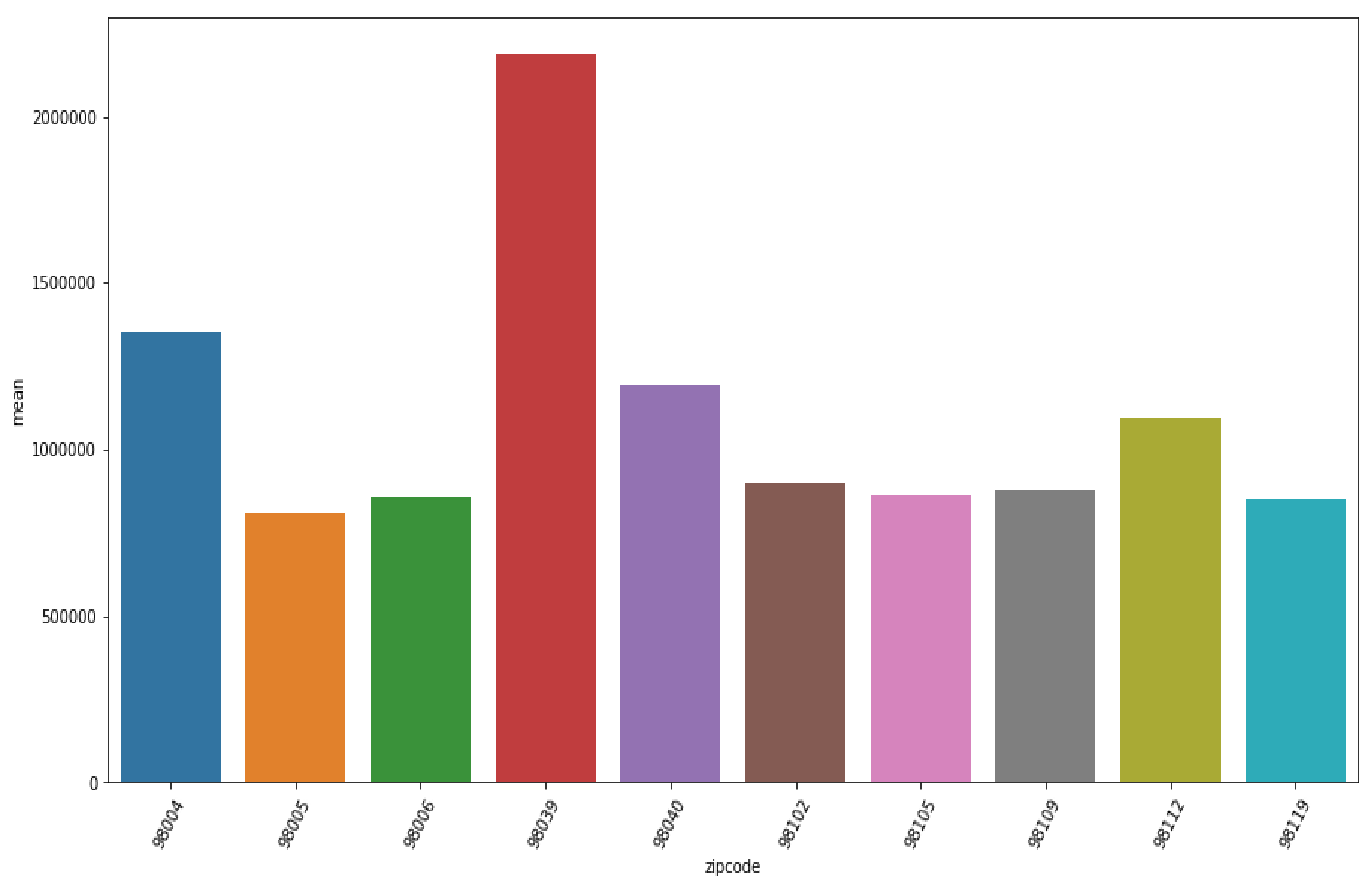
Relationship between 'house price' and the categorical variables ('zipcode')

Zipcode	mean	min	max	count	std
98039	2.186843e+06	787500.0	6885000.0	49	1.163564e+06
98004	1.355387e+06	425000.0	7062500.0	315	7.472826e+05
98040	1.194230e+06	500000.0	5300000.0	282	6.074935e+05
98112	1.096192e+06	169317.0	3400000.0	268	5.947617e+05
98102	8.993954e+05	330000.0	7700000.0	104	7.902389e+05
98109	8.796236e+05	216650.0	3200000.0	109	4.552288e+05

There are 70 unique zip codes in the data, and when we examined the box plot, we can not infer any kind of relationship with price. However, we can say that the house prices in some specific zip code areas are high than the other zip code areas.

Exploratory Data Analysis

Relationship between 'house price' and the categorical variables ('zipcode')



Inferential Statistics

I performed the hypothesis testing to check if the correlation between price and other features (bedrooms, bathrooms, Sqft_living, and Sqft_above) happened by chance.

#H0: There is no significant correlation between **number of bedroom** and **price**.

#Ha: There is a correlation between **number of bedrooms** and **price**.

The p-value is less than level of significance 0.05, so we reject the null hypothesis. There is a correlation between number of bedrooms and price.

I also performed the hypothesis testing between **price** and other features mentioned above.

Machine Learning

Algorithms

- Linear Regression (LR)
- Ridge Regression(RR)
- Lasso Regression(LassoR)
- Support Vector Regression(SVR)
- Decision Tree Regression (DTR)
- Random Forest Regression(RFR)
- Gradient Boosted Regression(GBR).

Performance Evaluation

- Mean Squared Error (MSE)
- Root Mean Square Error (RMSE)
- R2 score
- Mean_Absolute_Error (MAE)

(I evaluated the performance of SVR model only according to R2 score).

Machine Learning

Additional Data Preparation before Applying Models

- I applied one-hot encoding on 'waterfront', 'floors', 'view', 'condition', 'grade', 'basement_present', 'renovated' features.
- I created the copy of ‘h_data’ dataset as ‘h_data_copy’. I applied one-hot encoding on 6 zip codes, which have highest mean of house prices, on ‘h_data’ dataset and I dropped the “zipcode” feature. I used this version on LR, RL, LassoR, and SVR models. I used both version of “zipcode” feature (one-hot encoding applied version and original version) on tree-based models.

Additional Data Preparation before Applying Models

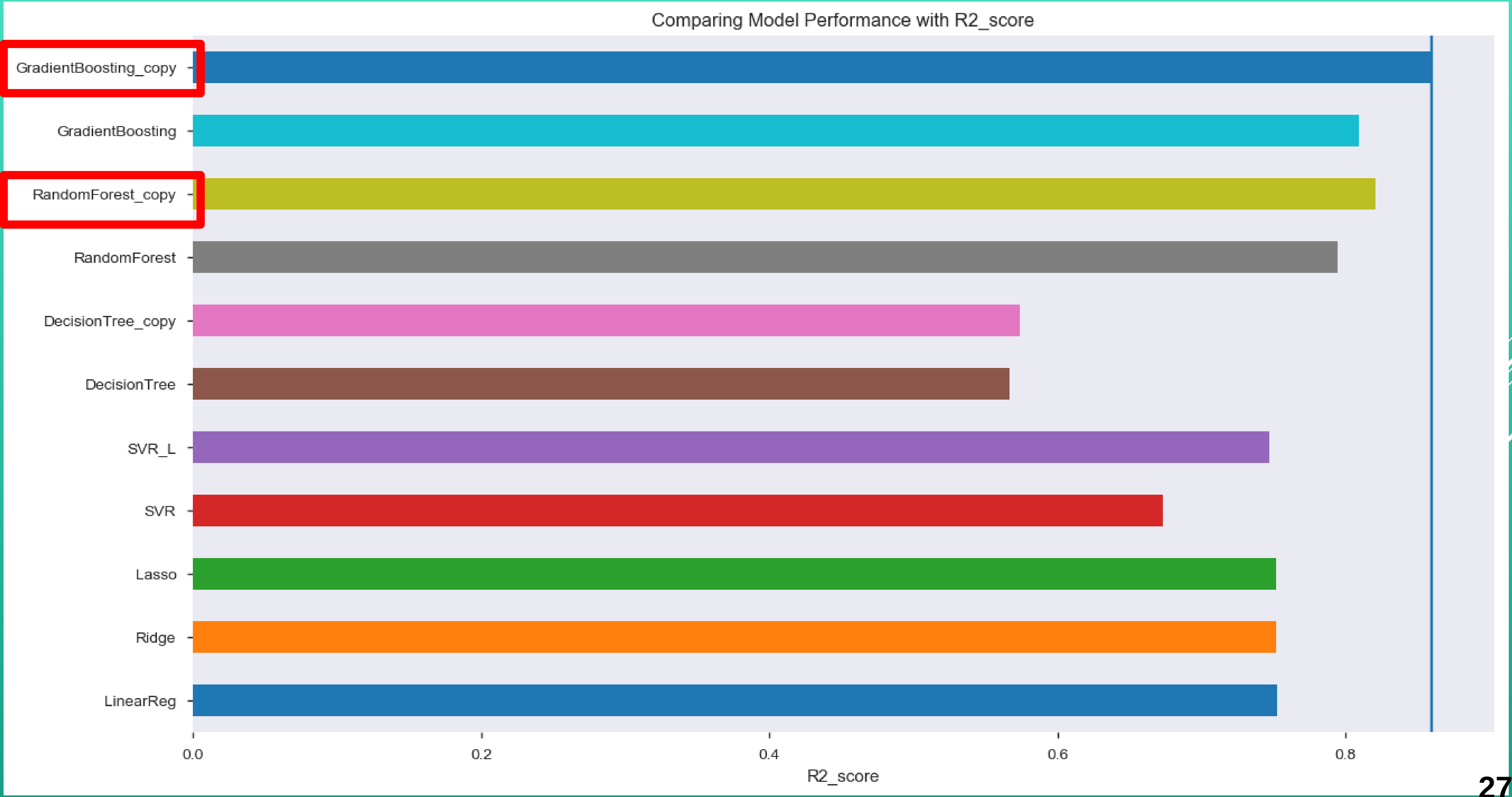
- I applied "pandas.profiling" to see the latest changes and their effects on datasets. 'renovated1' feature is highly correlated with 'yr_renovated' feature (ρ = 0.99997). I decided to drop this feature.
- I divided the data into independent variables "X and X_copy" and target variables "y and y_copy".
- I created a new data frame named "evaluation_matrix" to store the metrics of models.

Machine Learning

	LinearReg	Ridge	Lasso	SVR	SVR_L
Mean_Squared_Error(MSE)	3.37764e+10	3.3788e+10	3.37866e+10	0.326739	0.253243
Root_Mean_Squared_Error(RMSE)	183784	183815	183811	0.571611	0.503232
R2_score	0.751971	0.751886	0.751895	0.673261	0.746757
Mean_Absolute_Error(MAE)	125387	125394	125416	0.328629	0.332053

	Decision Tree	Decision Tree_ copy	Random Forest	Random Forest_ copy	Gradient Boosting	Gradient Boosting_ copy
Mean_Squared_ Error(MSE)	5.89478e+10	5.80211e+10	2.79853e+10	2.44322e+10	2.602865e+10	1.911278e+10
Root_Mean_ Squared_Error(RMSE)	242792	240876	167288	156308	161334	138249
R2_score	0.56713	0.573935	0.794496	0.820588	0.8088645	0.8596496
Mean_Absolute_ Error(MAE)	150482	127169	109315	88127.2	110167	83950.4

Machine Learning



Machine Learning

Hyperparameter Tuning

- Linear Regression, Ridge, Lasso, and Decision Tree Regressor model → GridSearchCV.
- Random Forest Regressor and Gradient Boosting Regressor model → RandomizedSearchCV.
- cv = 5

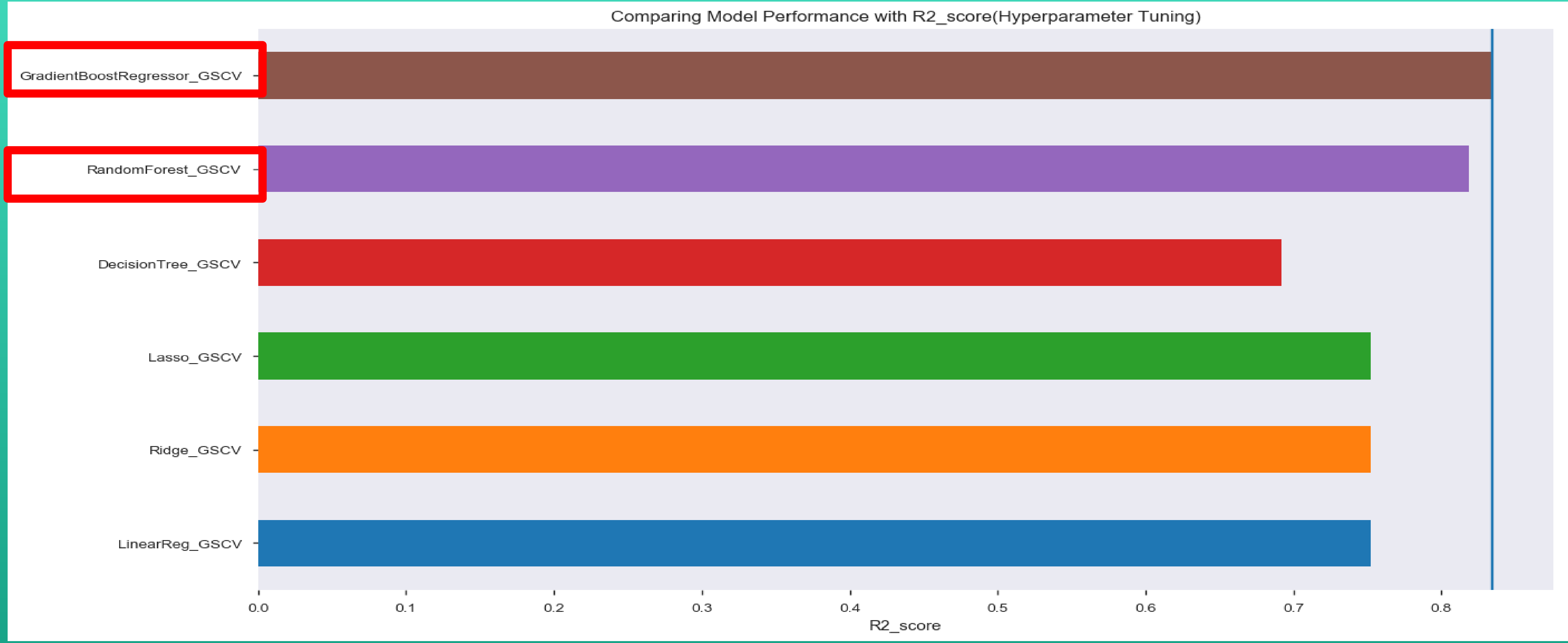
Machine Learning

Hyperparameter Tuning

	LinearReg_GSCV	Ridge_GSCV	Lasso_GSCV	DecisionTree_GSCV	RandomForest_RSCV	GradientBoostRegressor_RSCV
Mean_Squared_Error(MSE)	3.37764e+10	3.3788e+10	3.37639e+10	4.19655e+10	2.47025e+10	2.25726e+10
Root_Mean_Squared_Error (RMSE)	183784	183815	183749	204855	157170	150242
R2_score	0.751971	0.751886	0.752063	0.691836	0.818603	0.834243
Mean_Absolute_Error(MAE)	125387	125394	125386	111556	88326.1	83265.6

Machine Learning

Hyperparameter Tuning



Conclusion

- Gradient Boosted Regression (GBR) is the most effective model with the R2 score around 0.86.

- Random Forest Regression (RFR) is the second better model with the R2 score around 0.82.

- I would recommend using tree-based models, which have higher performance for predicting house prices in King County.