Predicting House Price with History Data

Tao Chen

2019.04

Agenda

- Background and requirements
- Data acquisition and introduction
- Exploratory Data Analysis
- Predictive Modeling
- Conclusion & Future direction
- Q&A

Background and requirements

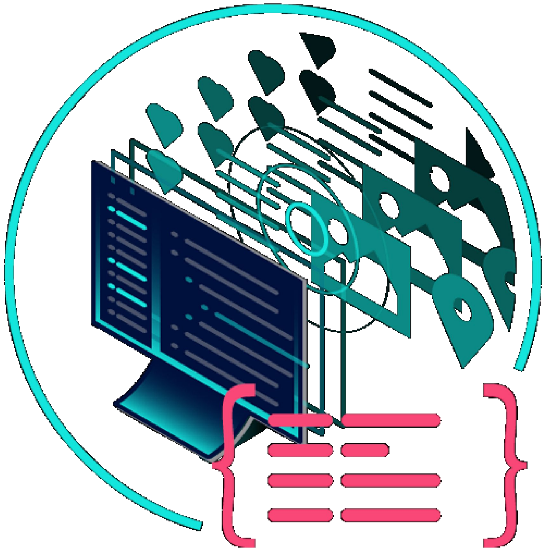
- The real estate market in The United States is very open and vigorous, a scientific way to estimate the house price is needed.
- Data science approach is used to get insights from history house sales data.
- We can build a predictive model to predict the house price, as the target is a continuous variable, a regression approach is selected.

Data acquisition

#	Feature	Description
1	id	It is the unique numeric number assigned to each house being sold.
2	date	It is the date on which the house was sold out.
3	price	It is the price of house which we have to predict so this is our target variable and apart from it are our features.
4	bedrooms	It determines number of bedrooms in a house.
5	bathrooms	It determines number of bathrooms in a bedroom of a house.
6	sqft_living	It is the measurement variable which determines the measurement of house in square foot.
7	sqft_lot	It is also the measurement variable which determines square foot of the lot.
8	floors	It determines total floors means levels of house.
9	waterfront	This feature determines whether a house has a view to waterfront 0 means no 1 means yes.
10	view	This feature determines whether a house has been viewed or not 0 means no 1 means yes.
11	condition	It determines the overall condition of a house on a scale of 1 to 5.
12	grade	It determines the overall grade given to the housing unit, based on King County grading system on a scale of 1 to 11.
13	sqft_above	It determines square footage of house apart from basement.
14	sqft_basement	It determines square footage of the basement of the house.
15	yr_built	It determines the date of building of the house.
16	yr_renovated	It determines year of renovation of house.
17	zipcode	It determines the zipcode of the location of the house.
18	lat	It determines the latitude of the location of the house.
19	long	It determines the longitude of the location of the house.
20	sqft_living15	Living room area in 2015(implies-- some renovations)
21	sqft_lot15	lotSize area in 2015(implies-- some renovations)

- Open dataset of house sales price in King County, Seattle, USA.
- CSV file contains 20 house features plus the price, along with 21613 observations.
- Foursquare API is used to get neighborhood data as a complement for this dataset.

Get Neighborhood Data with Foursquare API



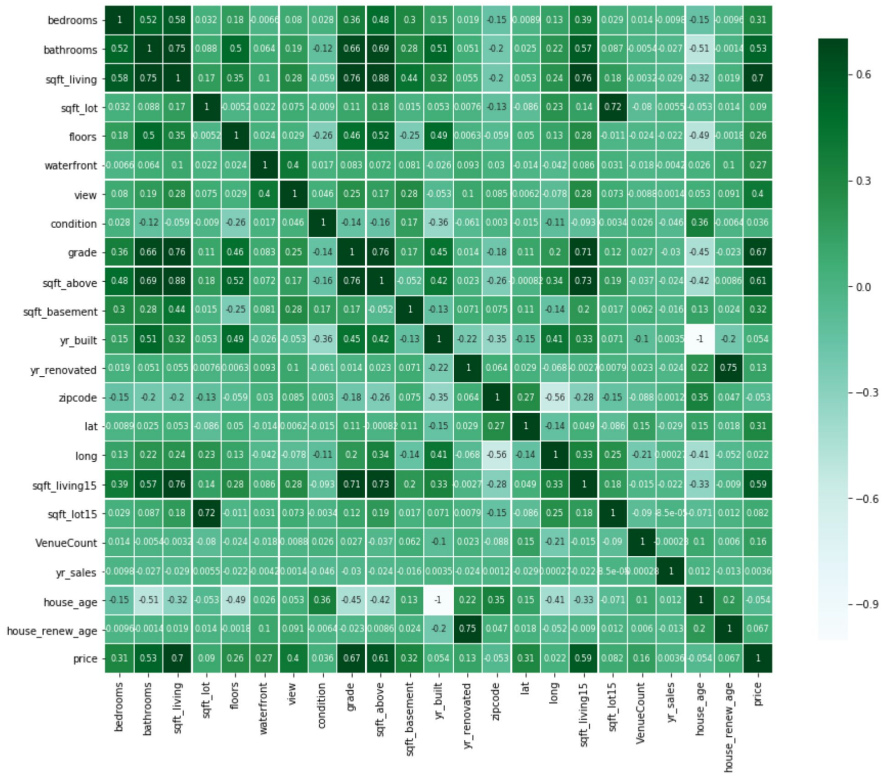
FOURSQUARE DEVELOPERS

- Use “Get Venue Recommendations” API to get the neighborhood data.
- Return a list of recommended venues near the house , include the name,location,categories of the venues.
- Add a new feature VenueCount to the dataset , which indicates the number of venues near the house.

Data Cleaning and Transformation

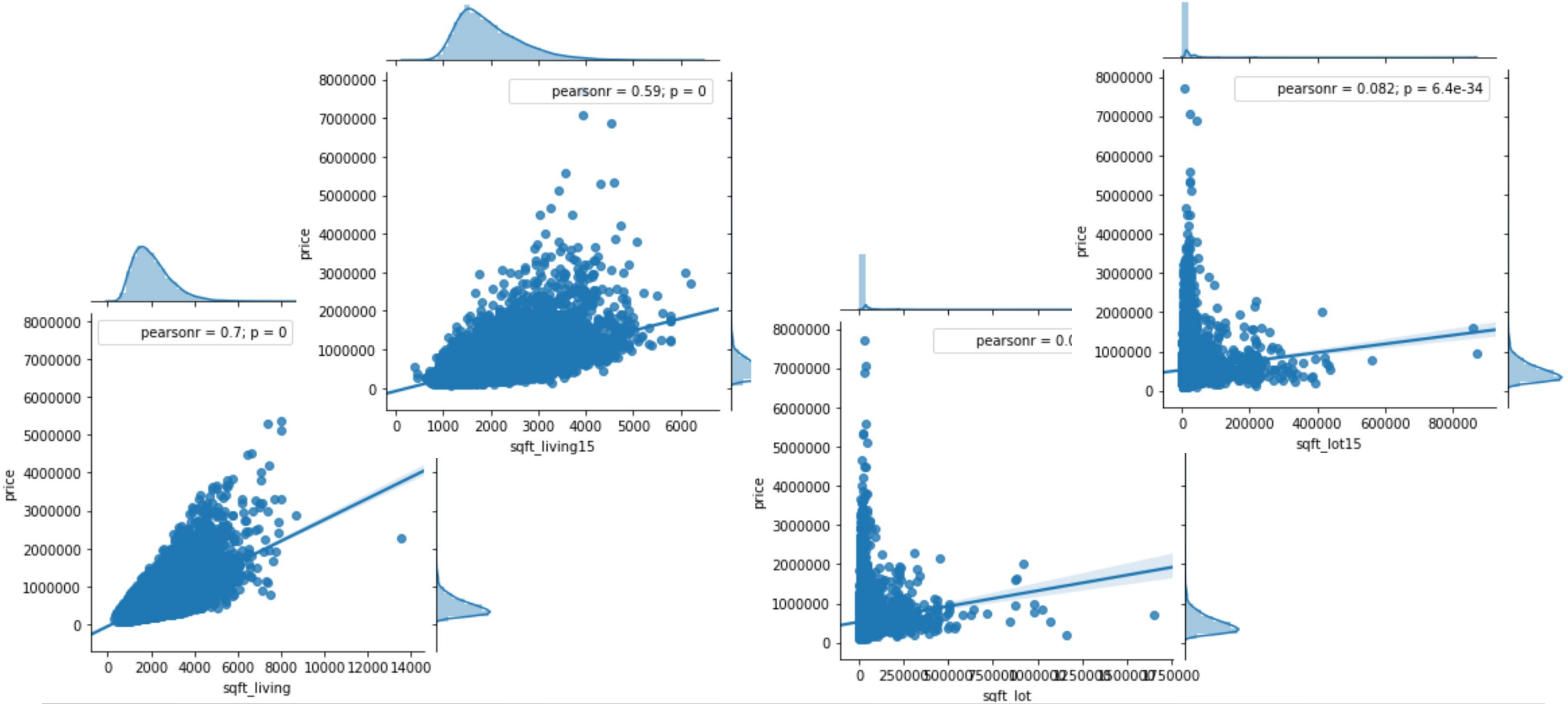
- Drop id and date column.
- Convert yr_built and yr_renovated to house_age and house_renew_age.
- The combination of latitude and longitude has almost the same impact as zipcode on the price prediction.
- Fill NaN values in the derived feature: VenueCount with 0, which means no venue found for that house.

Relationship Analysis

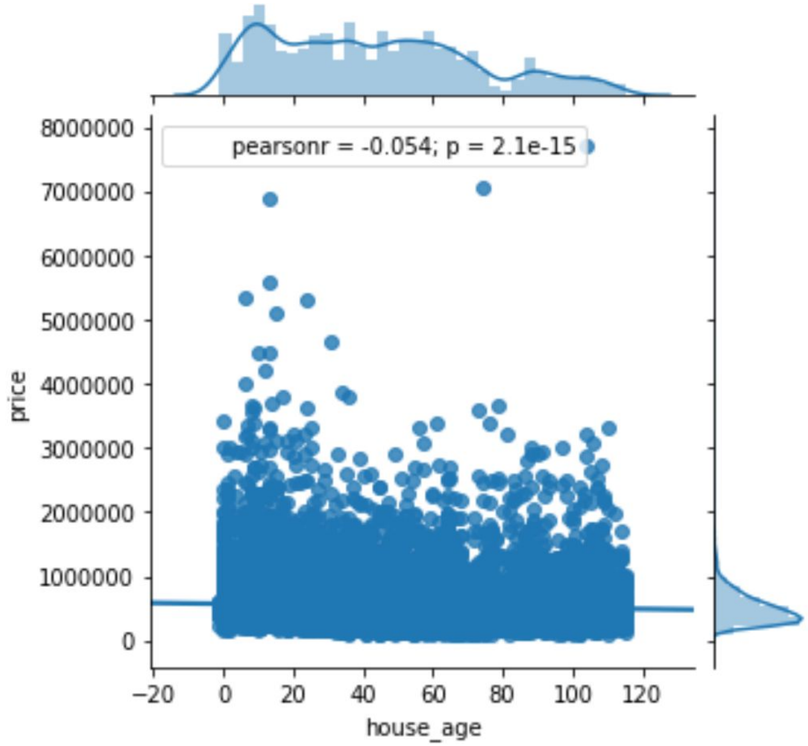


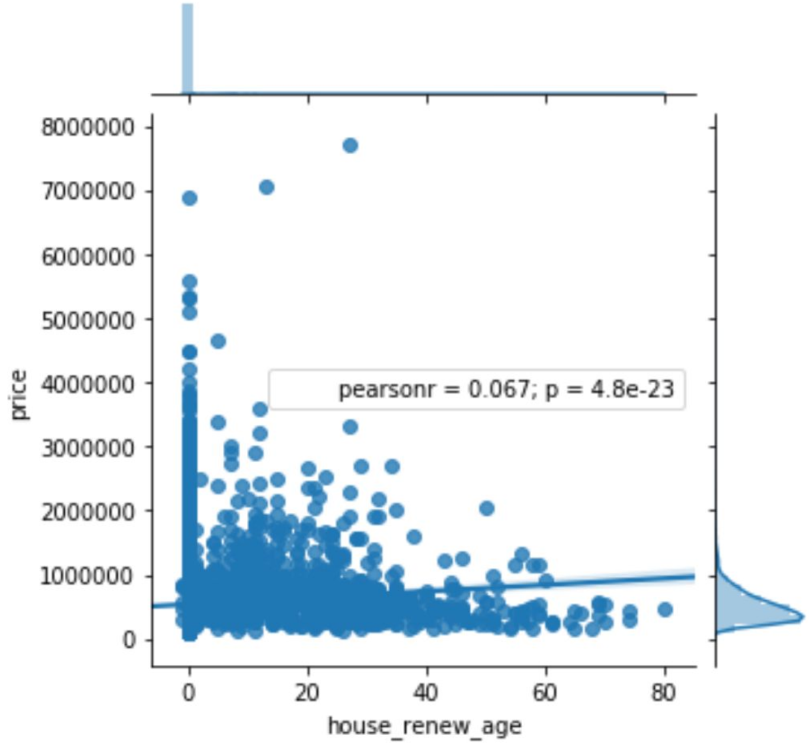
- sqft_living, sqft_above and sqft_basement have strong relationship with price. The 3 variables were also strongly related to each other as sqft_living = sqft_above + sqft_basement.
- sqft_living15 has strong relationship with price.
- sqft_lot, sqft_lot15 and yr_built are poorly related to price.
- Waterfront is slightly associated with price.
- Bedrooms, bathrooms, floors, views, grade have strong connections with price.

Relationship between house area and price

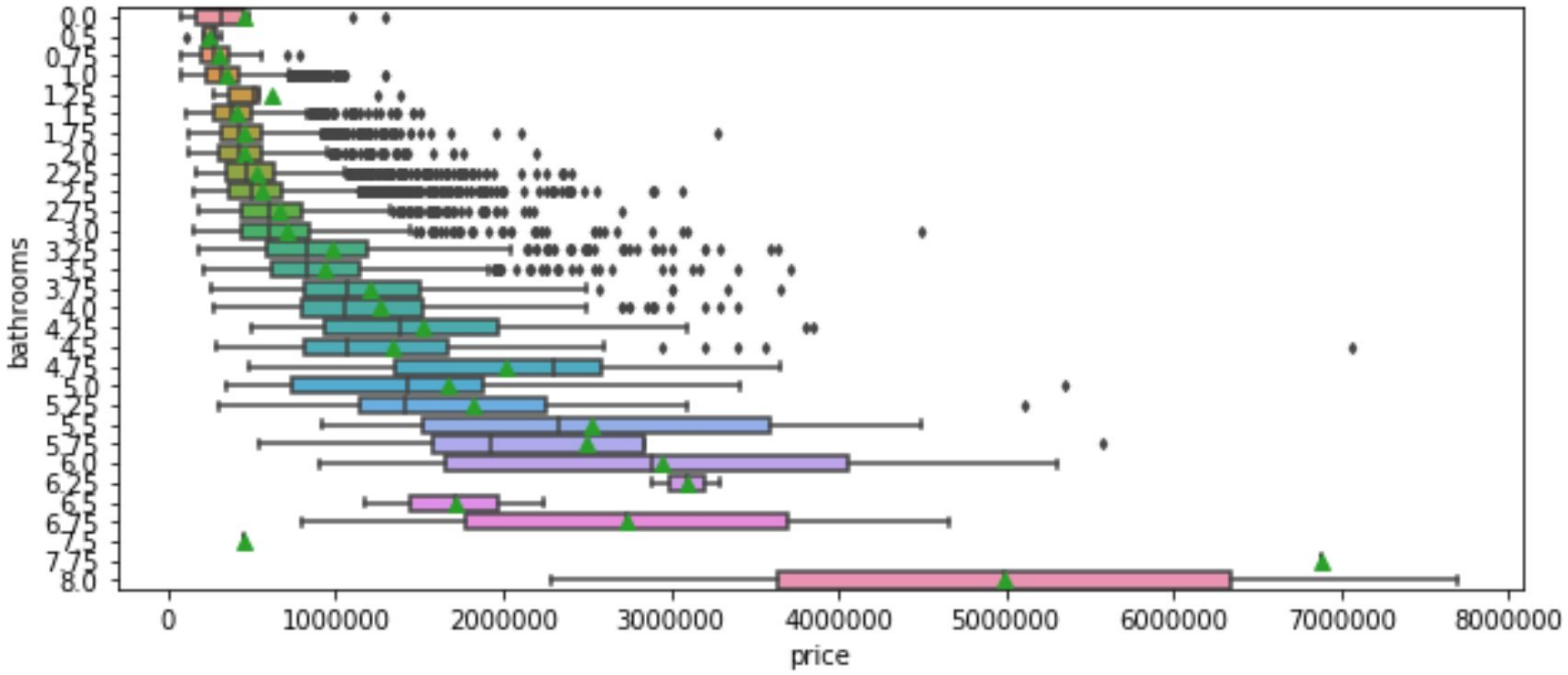


Relationship between house age and price

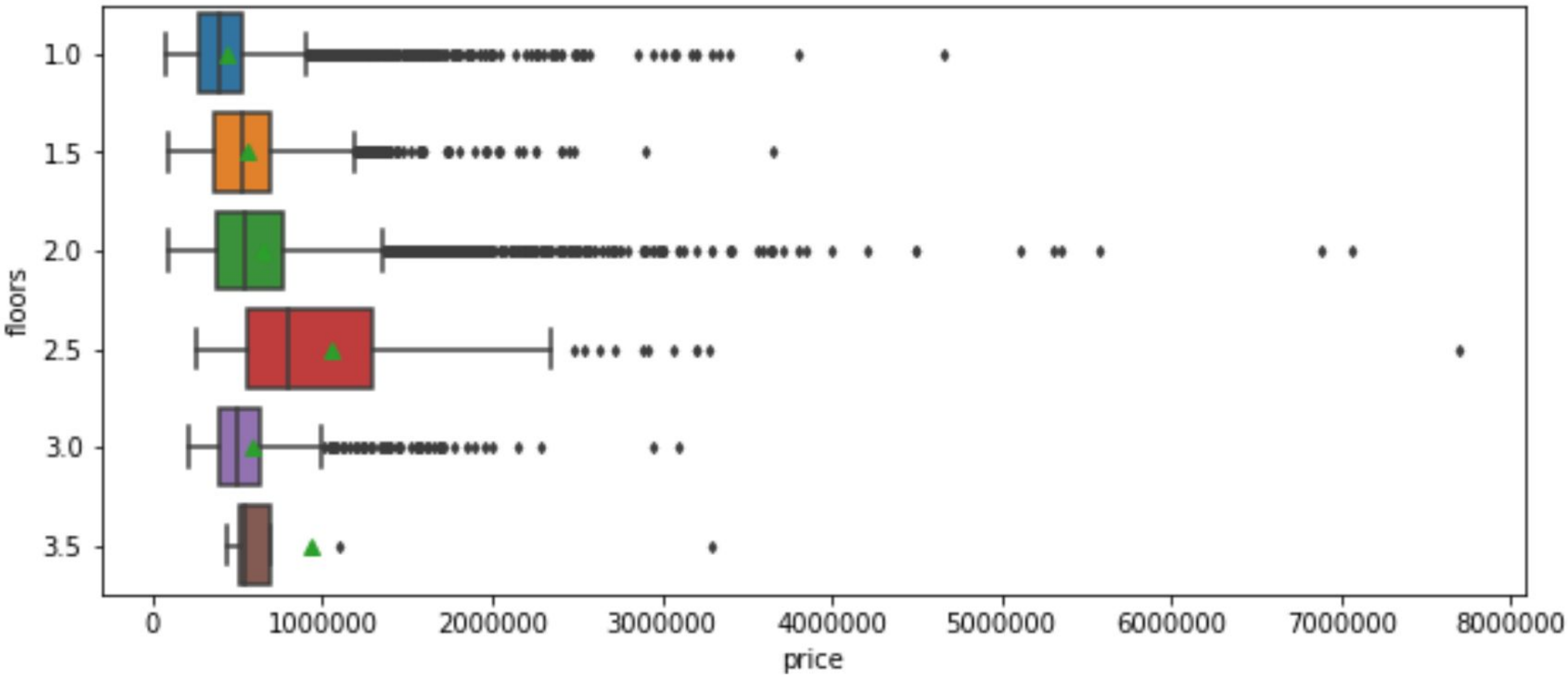




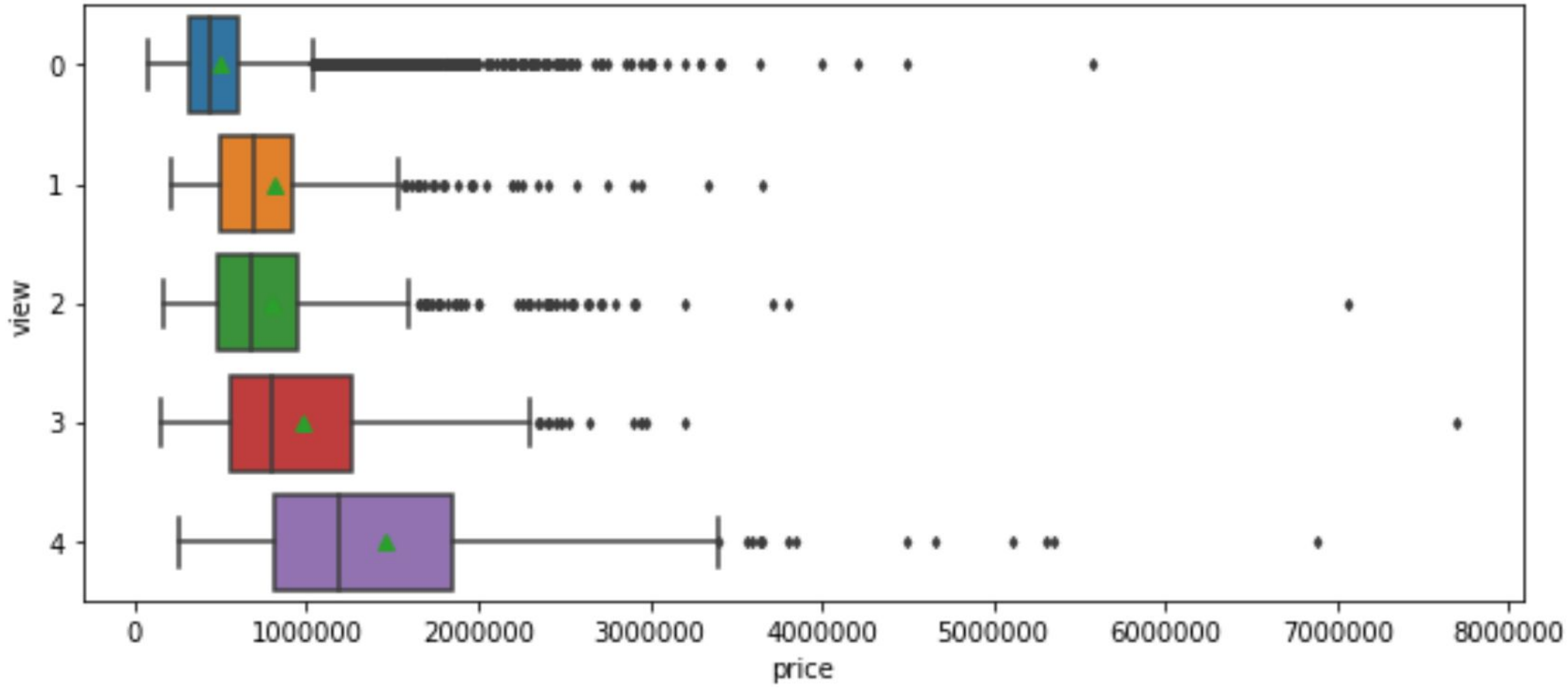
Relationship between bathrooms and price



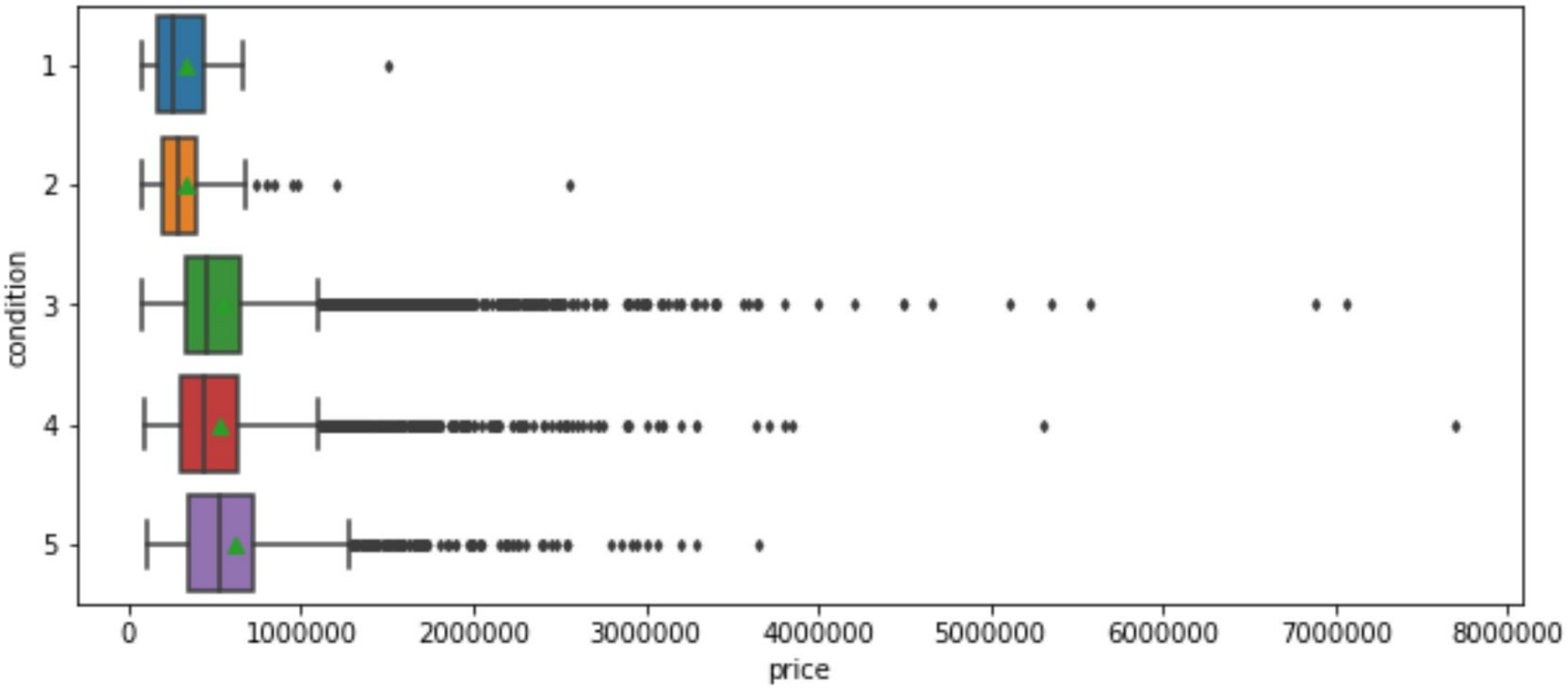
Relationship between floors and price



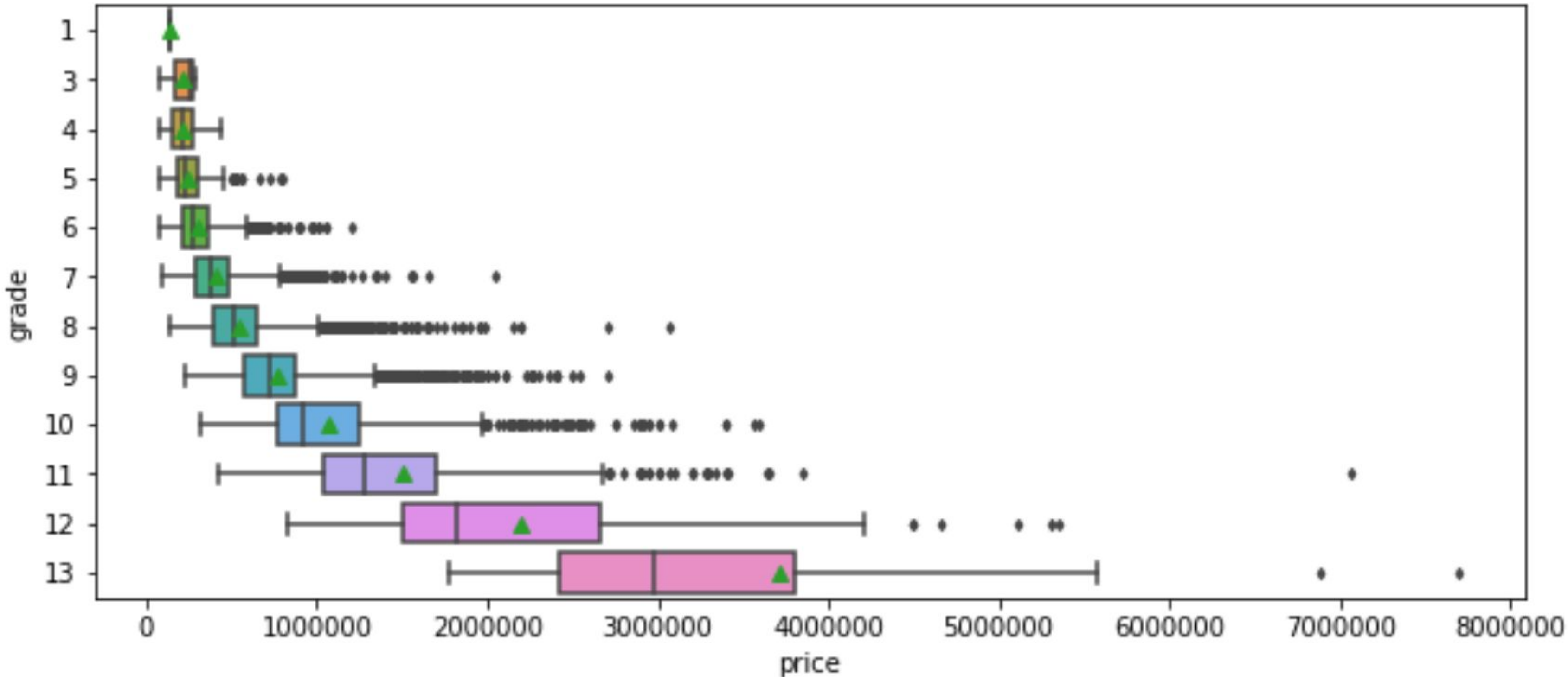
Relationship between view and price



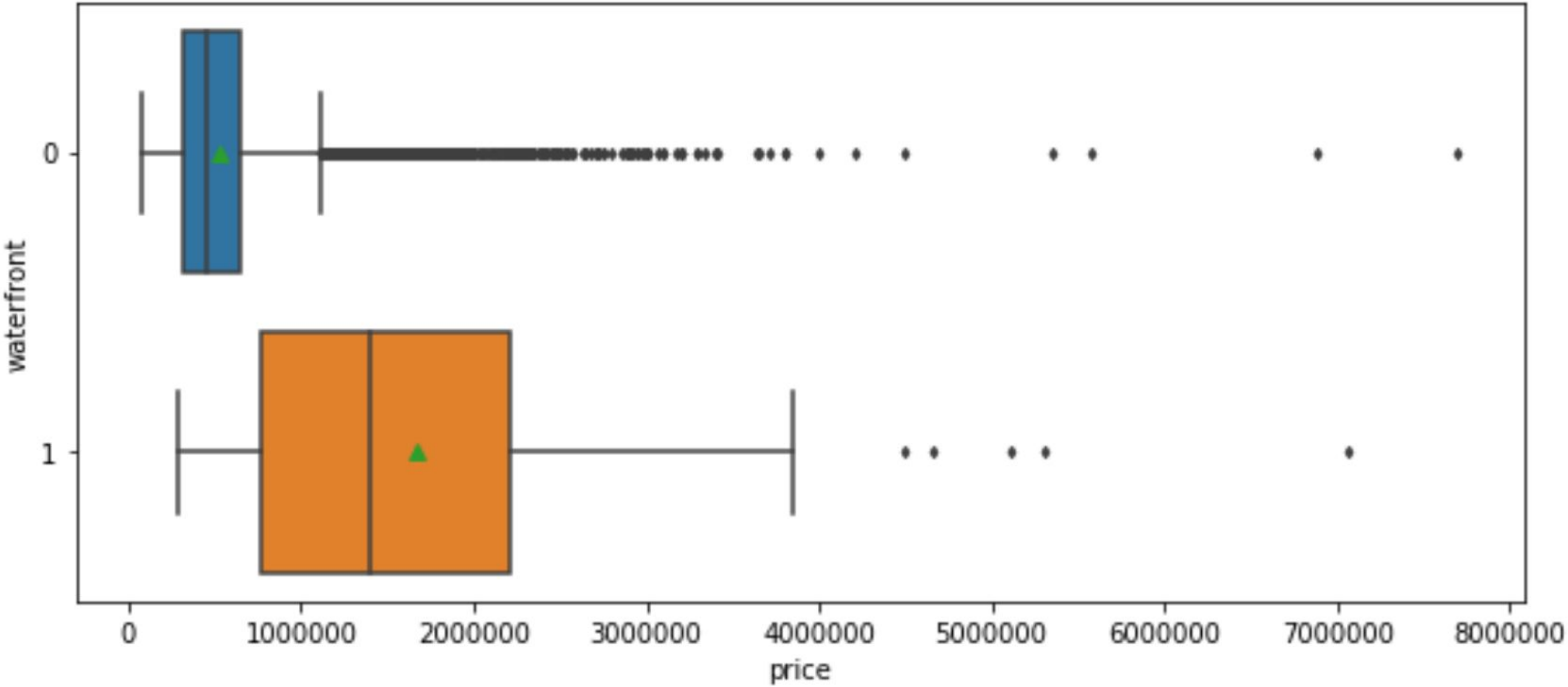
Relationship between condition and price



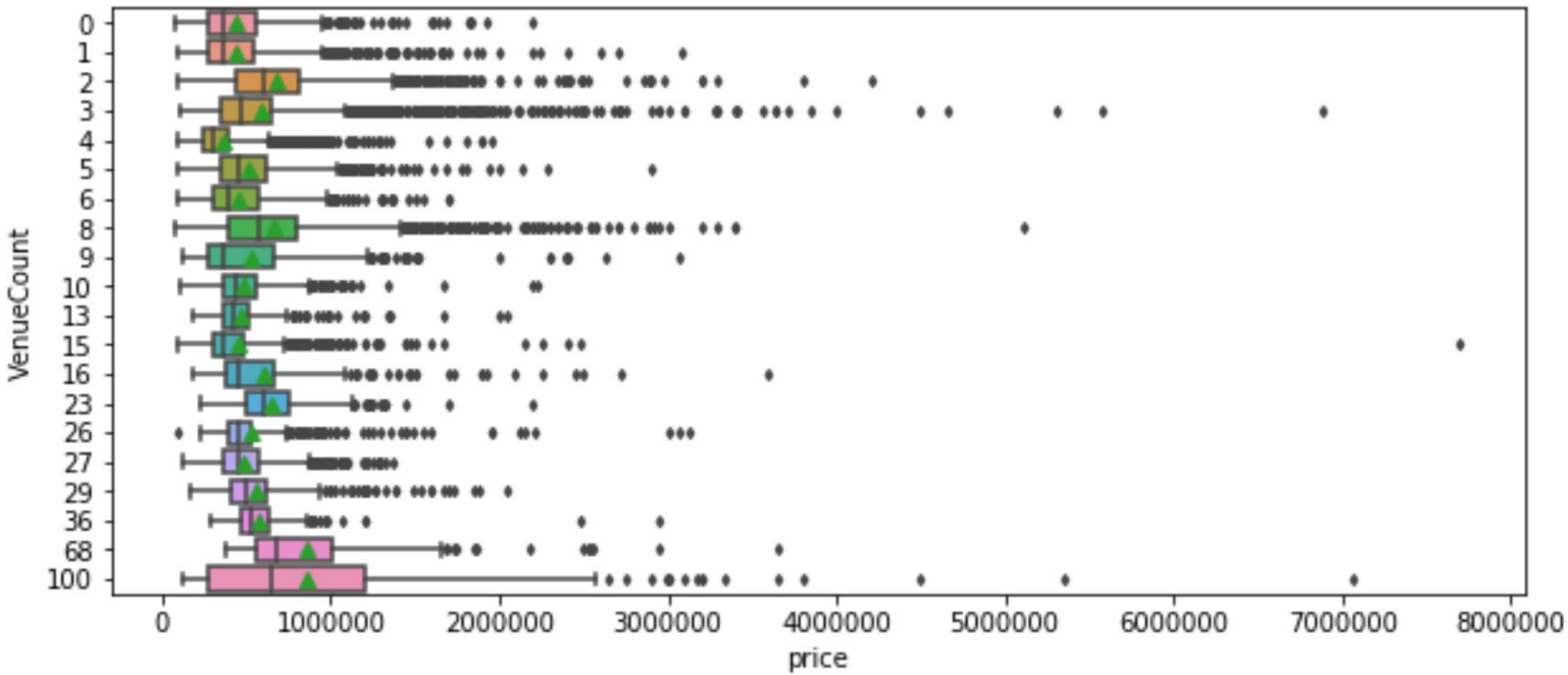
Relationship between grade and price



Relationship between waterfront and price



Relationship between nearby venues count and price

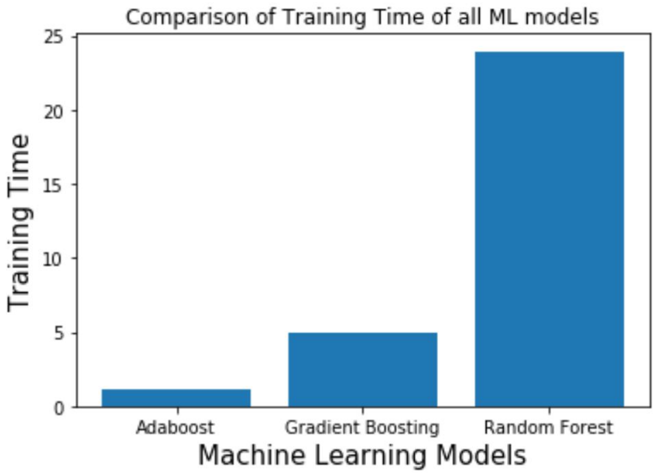


Feature selection

#	Feature	Selected(Y/N)	Reason
1	bedrooms	Y	Contribute to price.
2	bathrooms	Y	Contribute to price.
3	sqft_living	Y	Contribute to price.
4	sqft_lot	N	Similar feature selected and no strong correlation.
5	floors	N	No strong relationship.
6	waterfront	Y	Contribute to price.
7	view	Y	Contribute to price.
8	condition	Y	Contribute to price.
9	grade	Y	Contribute to price.
10	sqft_above	N	Similar feature selected
11	sqft_basement	N	Similar feature selected and no strong correlation.
12	yr_built	N	Transformed to another feature.
13	yr_renovated	N	Transformed to another feature.
14	zipcode	Y	Contribute to price.
15	lat	N	Similar feature selected
16	long	N	Similar feature selected
17	sqft_living15	N	Similar feature selected
18	sqft_lot15	N	Similar feature selected and no strong correlation.
19	house_age	Y	Contribute to price.
20	house_renew_age	Y	Contribute to price.
21	VenueCount	Y	Contribute to price.

Predictive Modeling

	Model	Accuracy Score	Variance Score	R2 Score
0	AdaBoost	0.525	0.401	0.379
1	Random Forest	0.780	0.746	0.746
2	Gradient Boosting	0.820	0.804	0.804



- AdaBoostRegressor, RandomForestRegressor and GradientBoostingRegressor were used.
- Split dataset into 20% of test data and remaining 80% will be used for training the model.
- Evaluation Metrics: R2-score , Accuracy Score and Explained Variance Score.

Conclusion & Future direction

- Analyzed the house sales dataset and relationship between the house price and the independent variables.
- Identified sqft_living, grade, view, bathrooms , bedrooms among the most important features that affect a house's sale price.
- Three regression models were built, GradientBoostingRegressor model had the best performance.
- Intuitively, mature neighborhood with consummate supportive commercial and residential facilities will help a house get higher price, one of the further directions is to get more neighborhood data to enhance the model.

Q&A

Thanks!