Indian Institute of Technology, Madras
CS6700: Reinforcement Learning
Reinforcement Learning Assignment-I Report

Rajan Kumar Soni - CS18S038

March 2020

Contents

1 Data Visualization and Observations (For extra marks) 3

1.1 Heat Map Visualisation 3

1.2 Box plot visualisation for categorical features 4

1.3 Scatter plot for continuous feature 5

1.4 Distribution of features 7

1.5 TSNE plot 11

1.6 KEY OBSERVATIONS 11

2 Polynomial Regression 12

1 Data Visualization and Observations (For extra marks)

1.1 Heat Map Visualisation

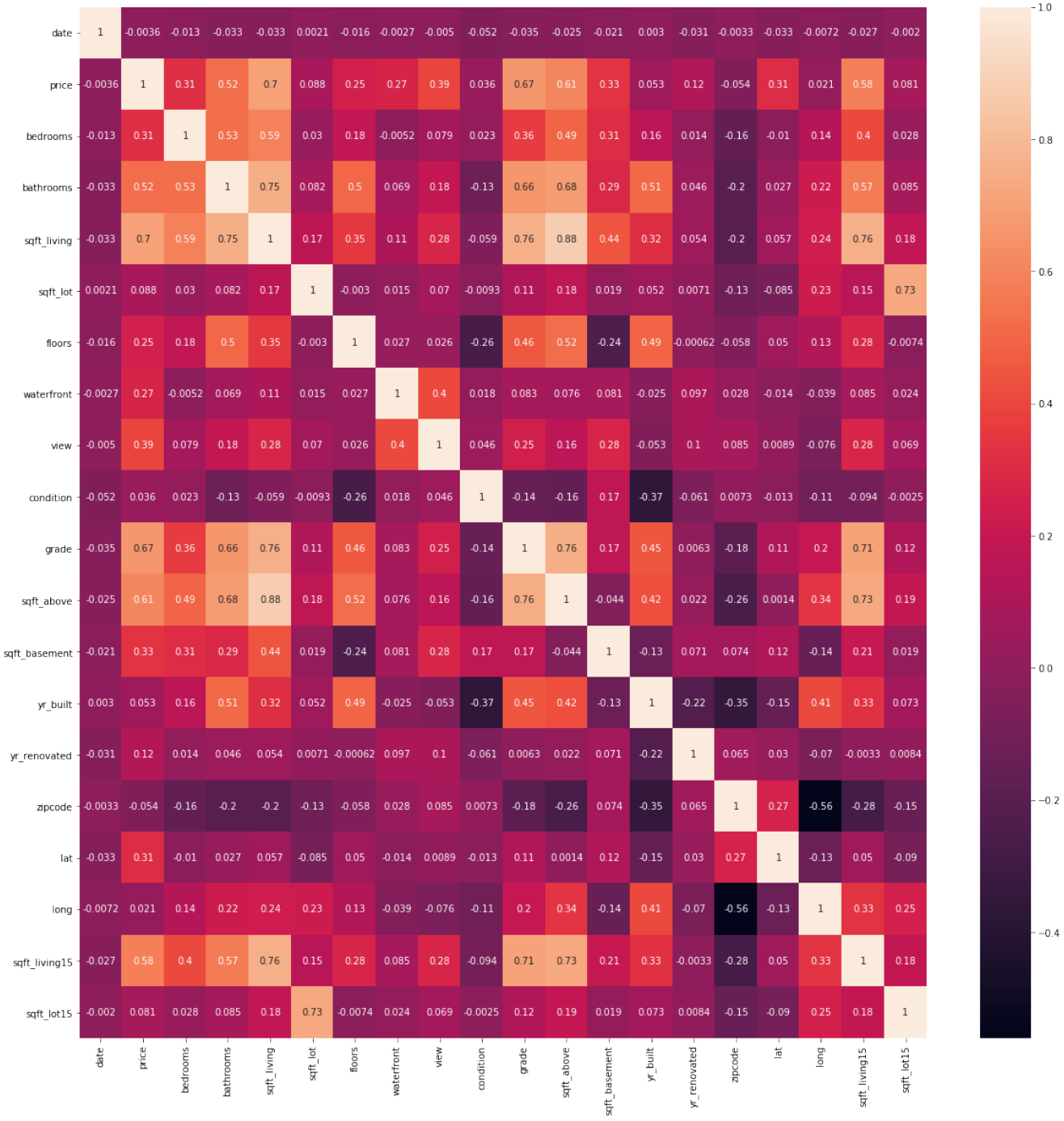
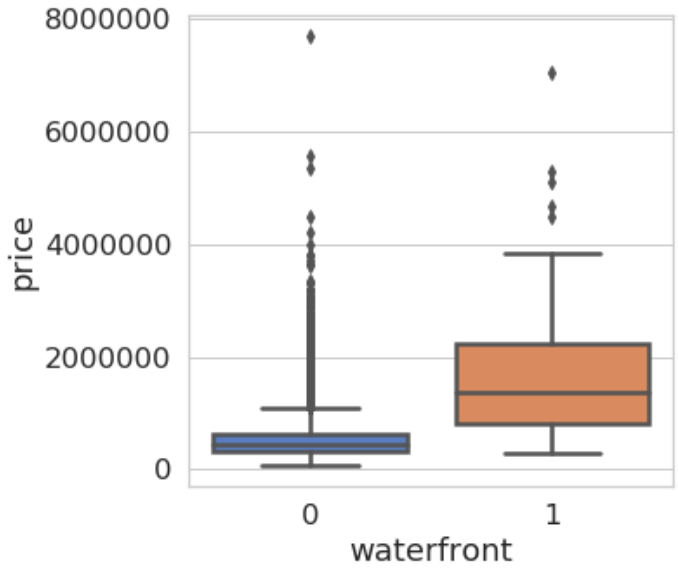
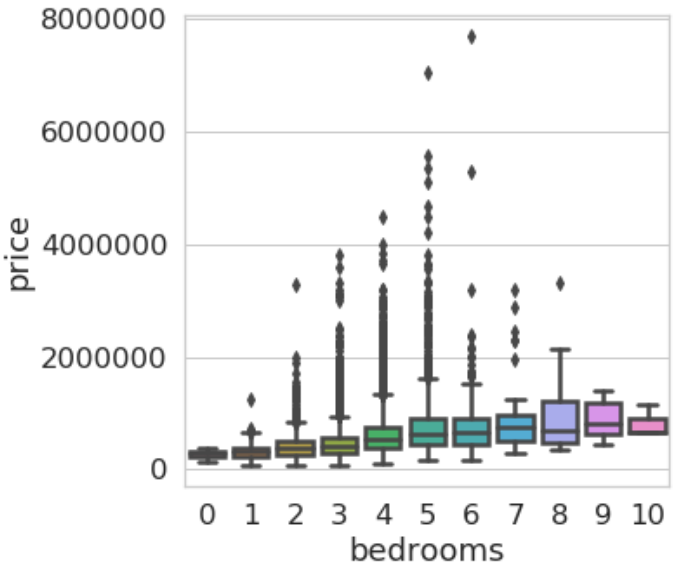
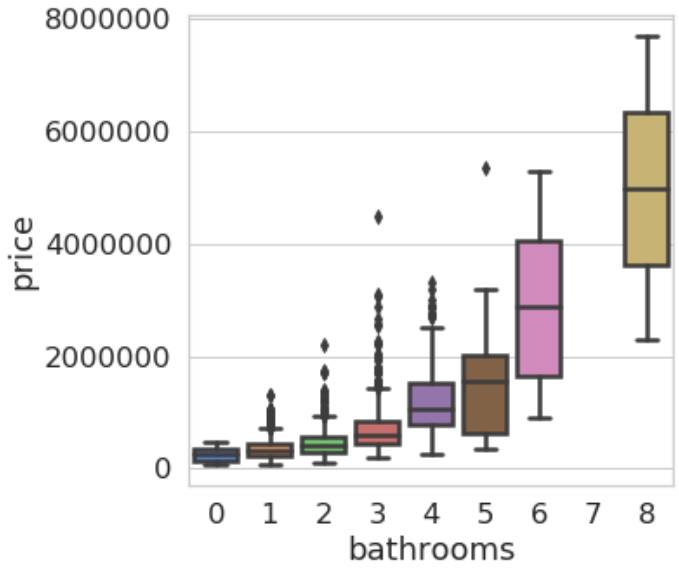


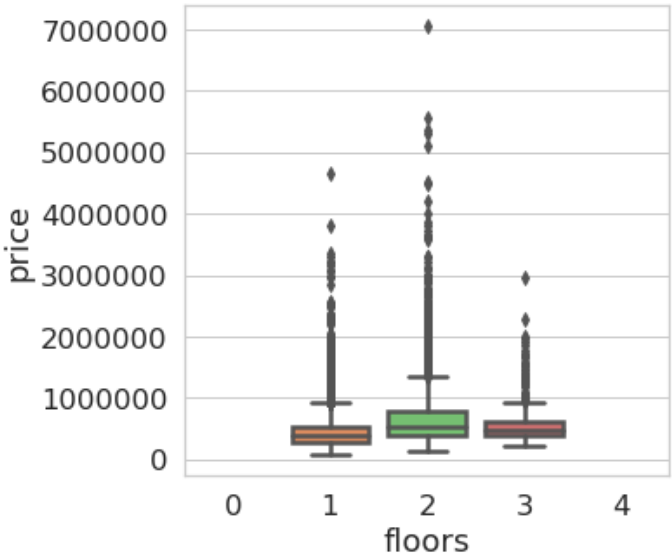
Figure 1: Heat Map

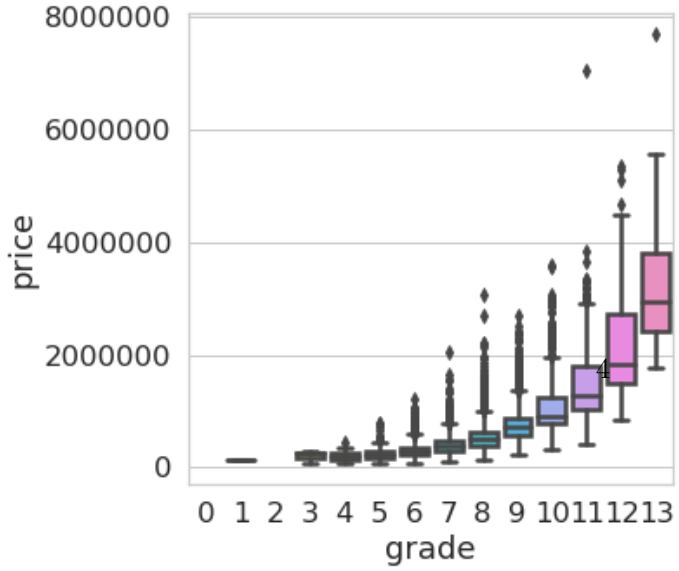
1.2 Box plot visualisation for categorical features

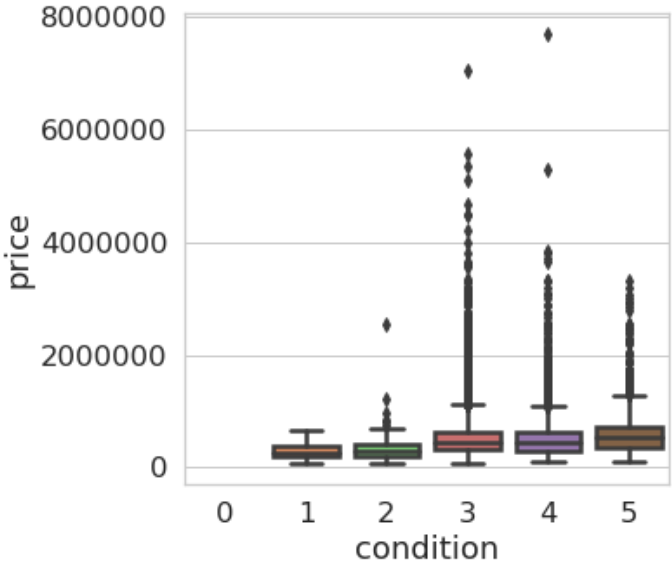




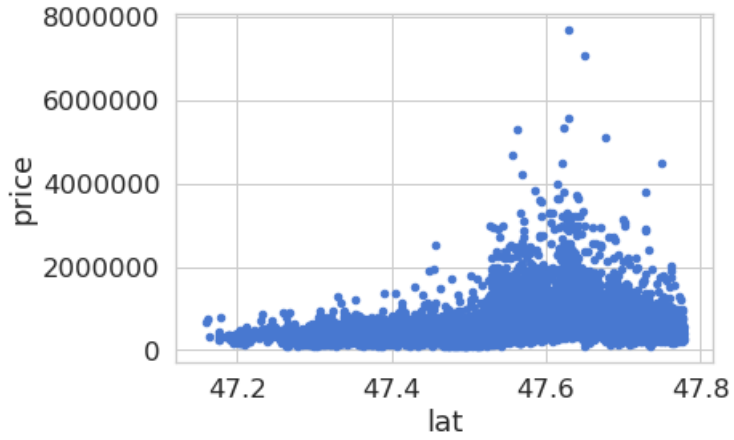


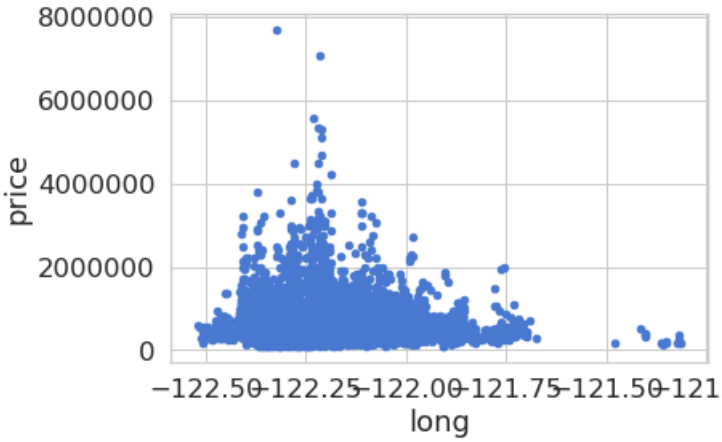


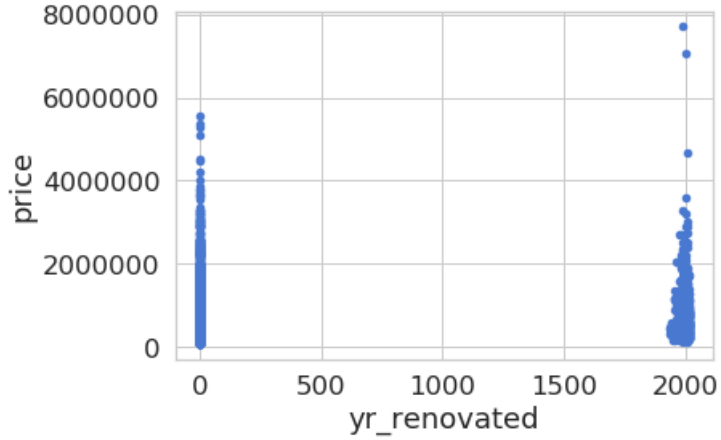


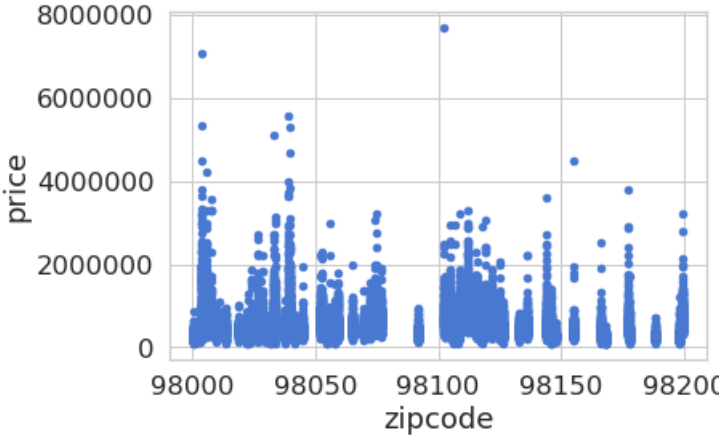


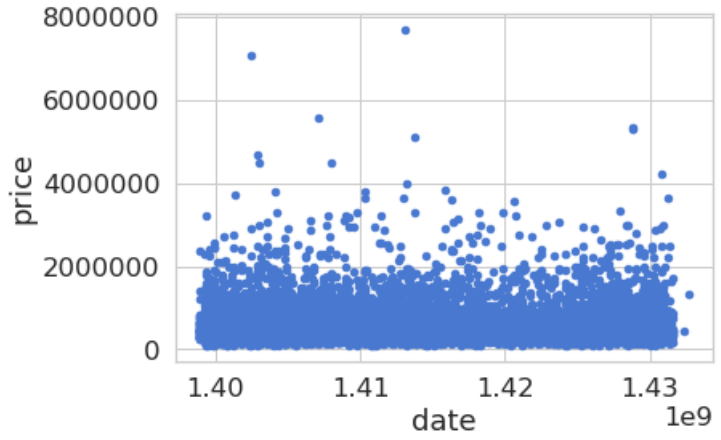
1.3 Scatter plot for continuous feature

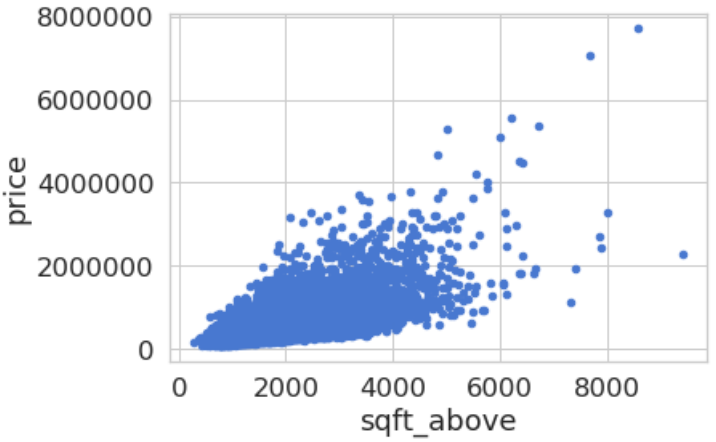


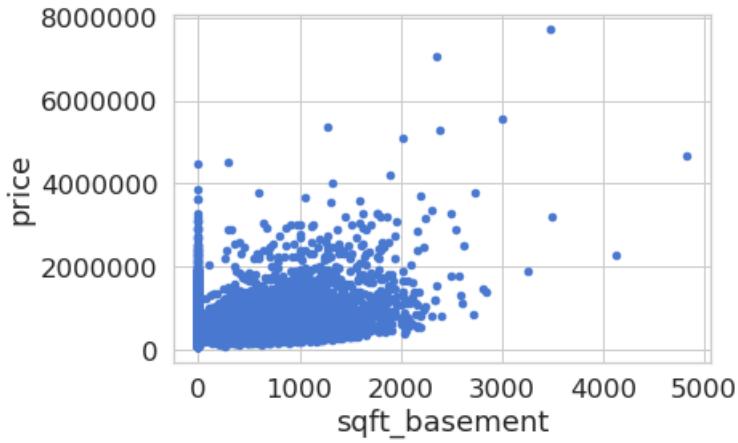


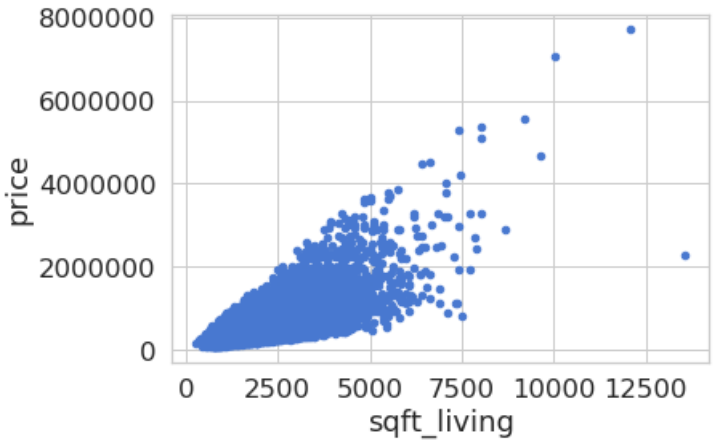


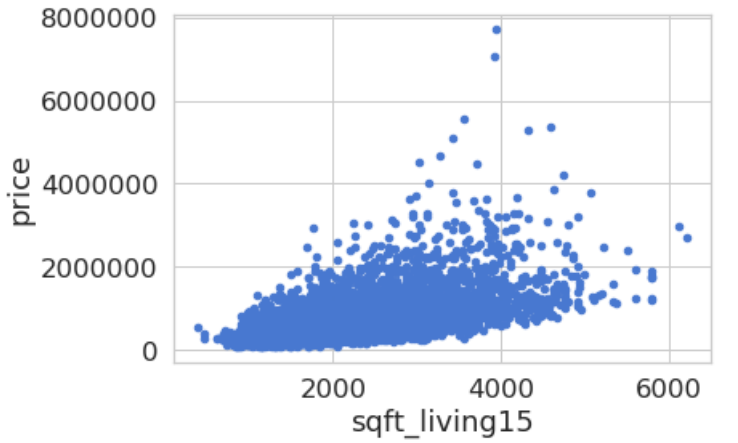


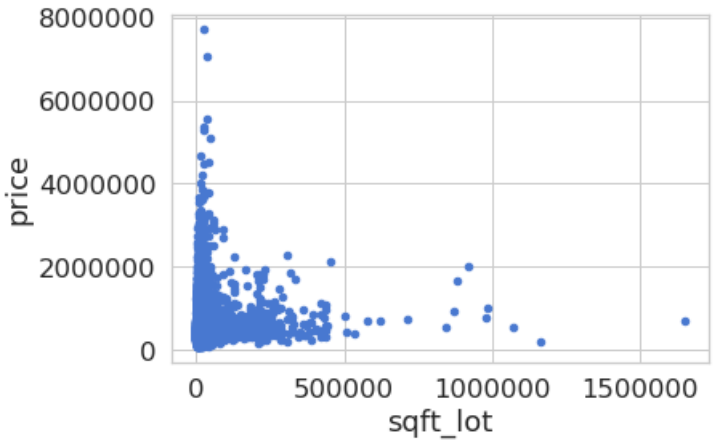


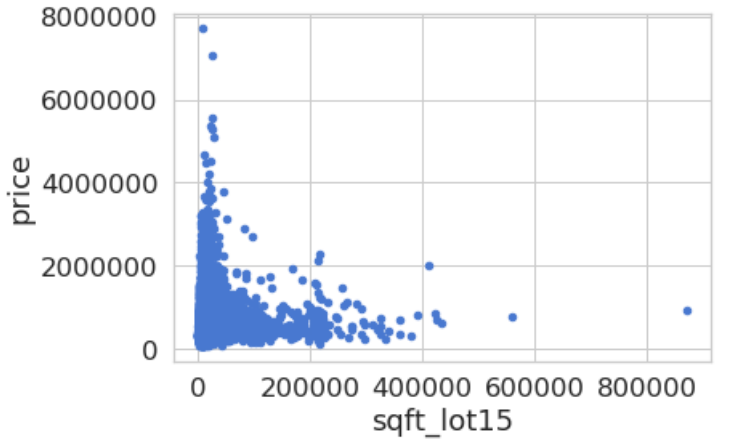


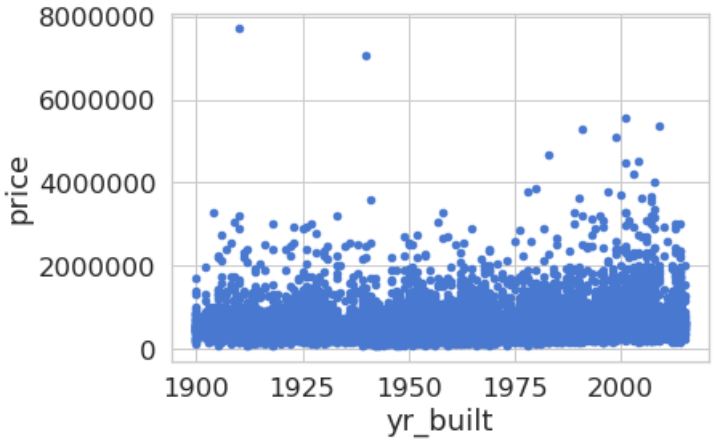

price
8000000
6000000
4000000
2000000
0
0
1000
2000
3000
4000
5000
sqft_basement


price
8000000
6000000
4000000
2000000
0
0
2500
5000
7500
10000
12500
sqft_living

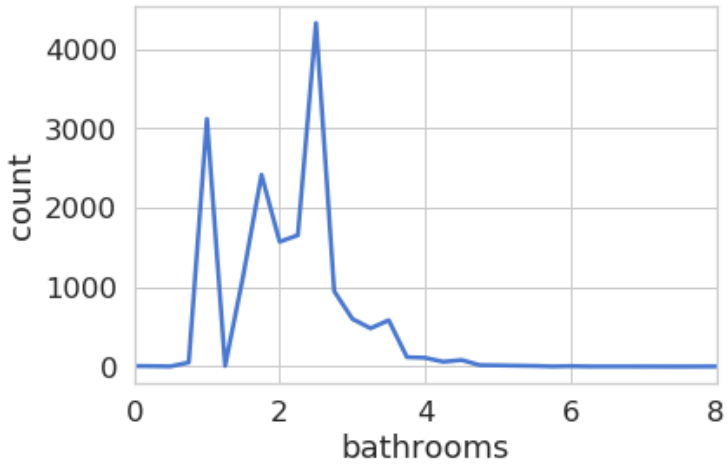

price
8000000
6000000
4000000
2000000
0
2000
4000
6000
sqft_living15

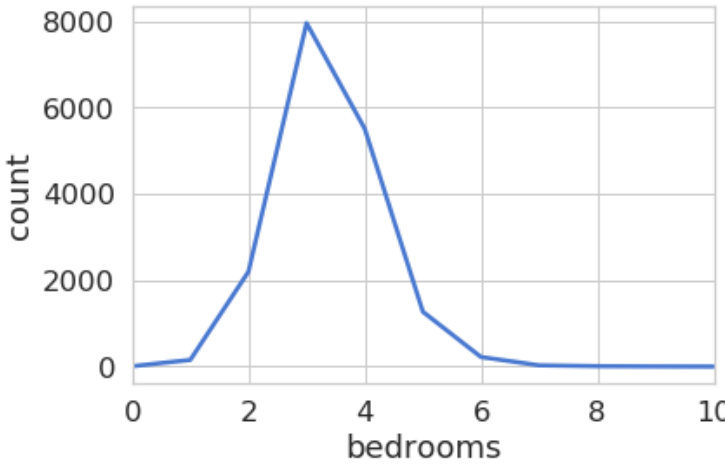

price
8000000
6000000
4000000
2000000
0
0
500000
1000000
1500000
sqft_lot

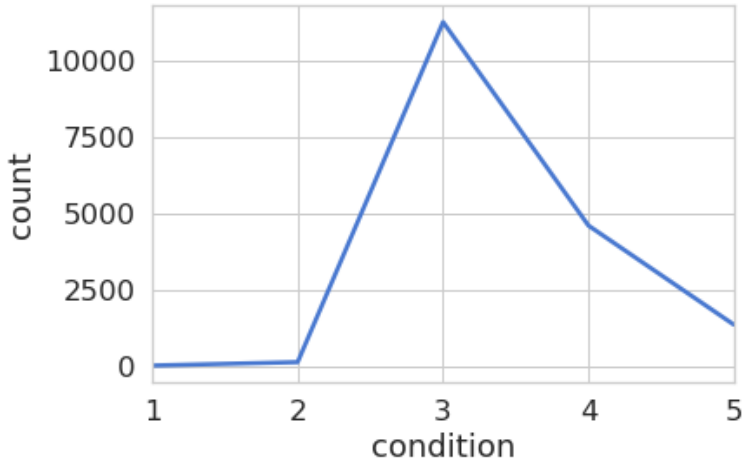

price
8000000
6000000
4000000
2000000
0
0
200000
400000
600000
800000
sqft_lot15

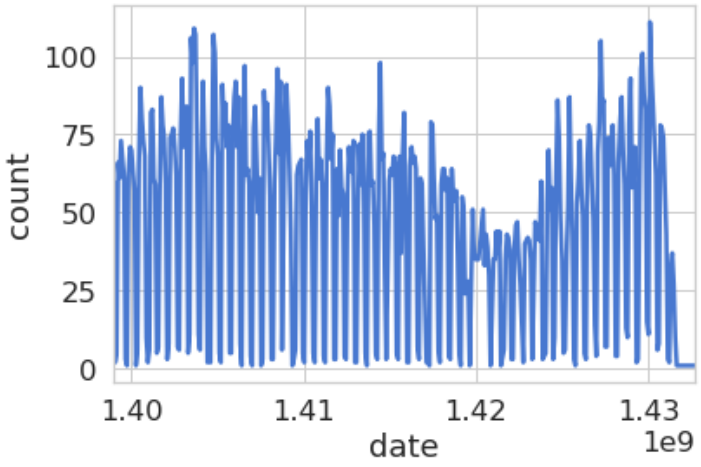

price
8000000
6000000
4000000
2000000
0
1900
1925
1950
1975
2000
yr_built

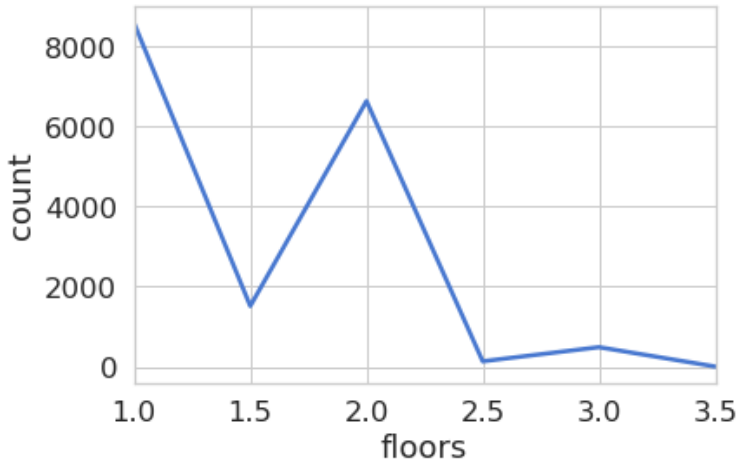
1.4 Distribution of features

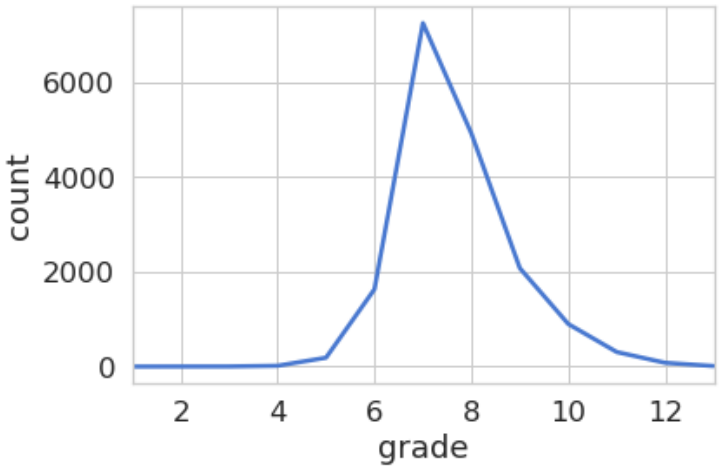


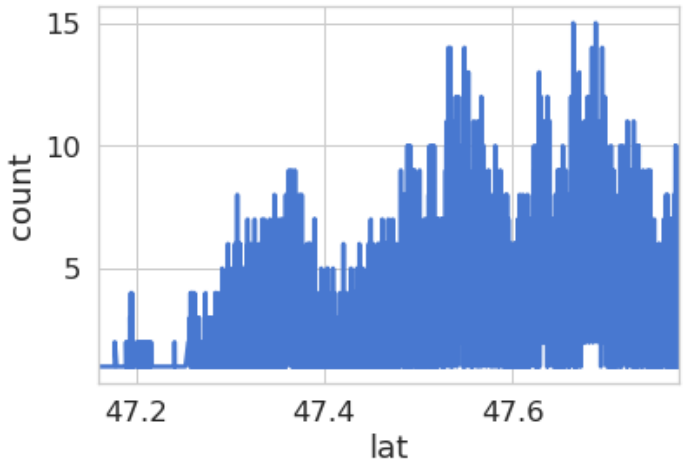


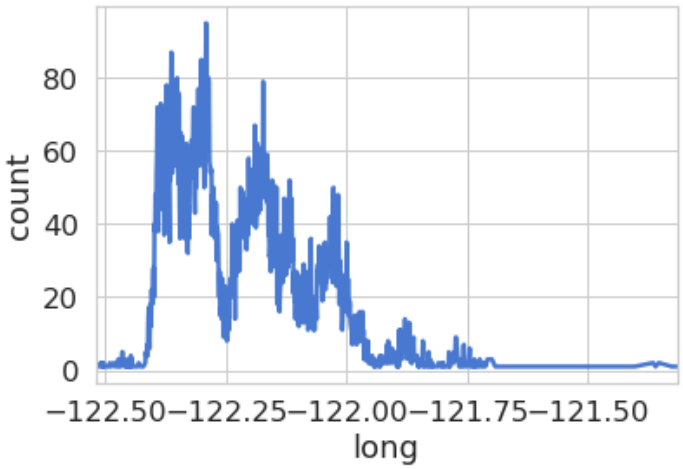


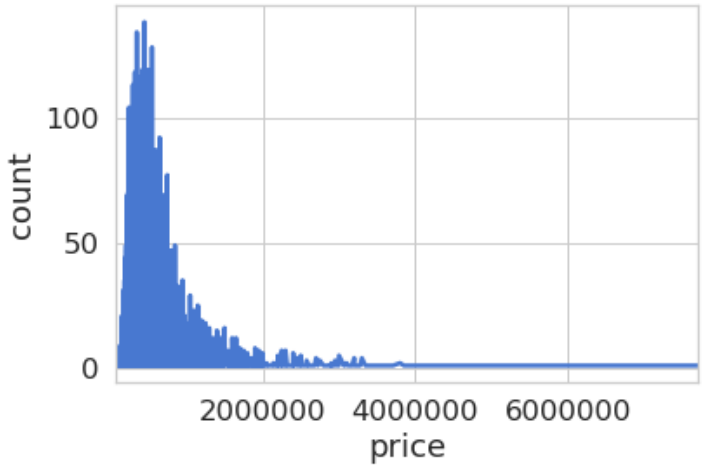


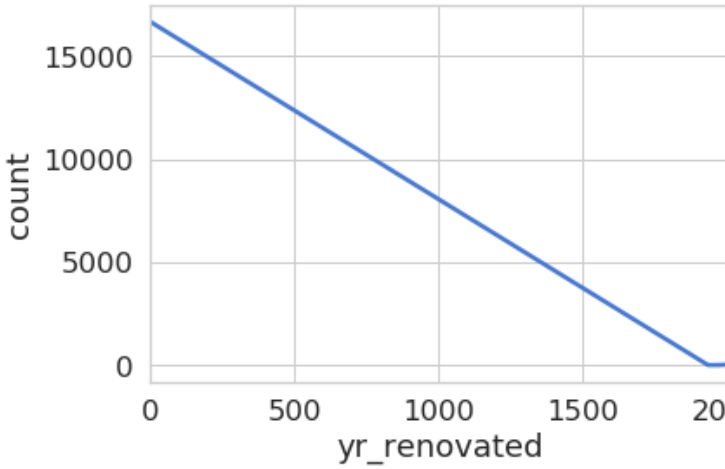


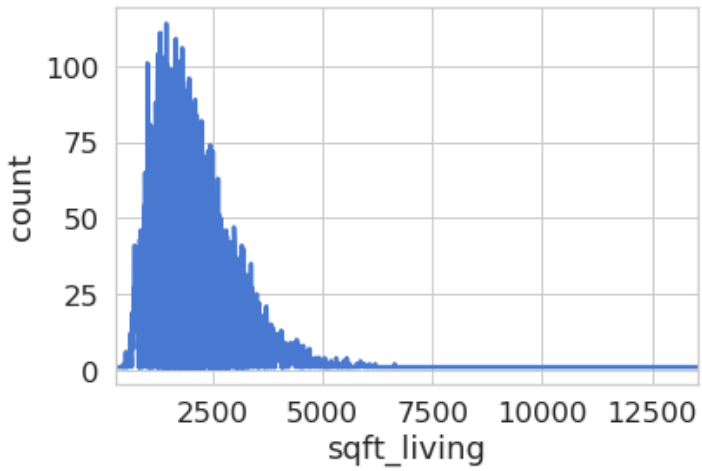


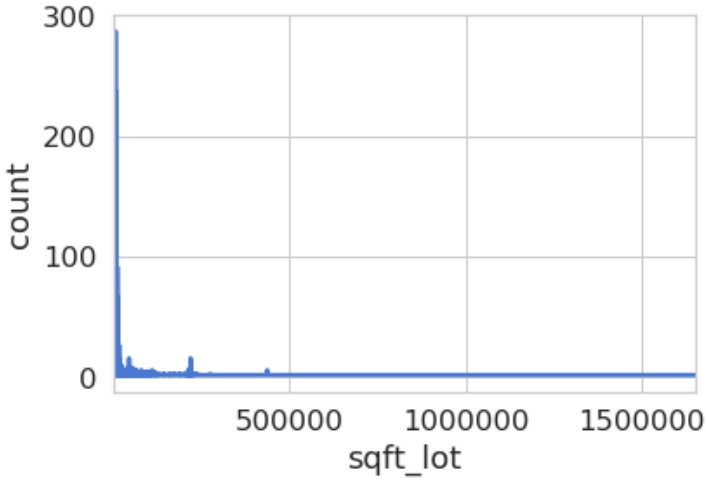

15
10
5
count
47.2
47.4
47.6
lat

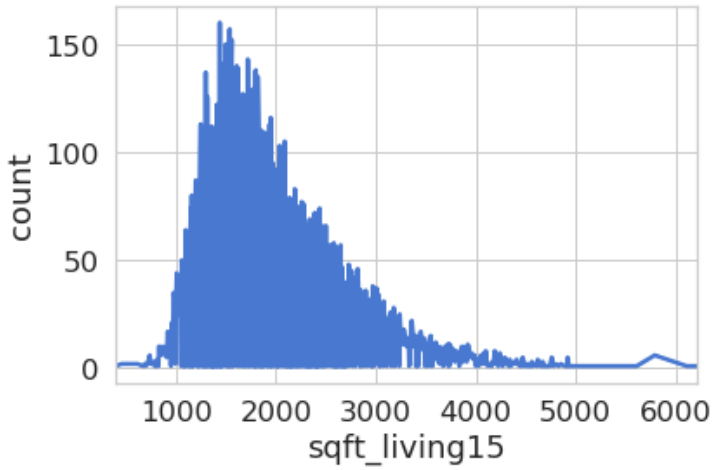

80
60
40
20
0
count
−122.50
−122.25
−122.00
−121.75
−121.50
long

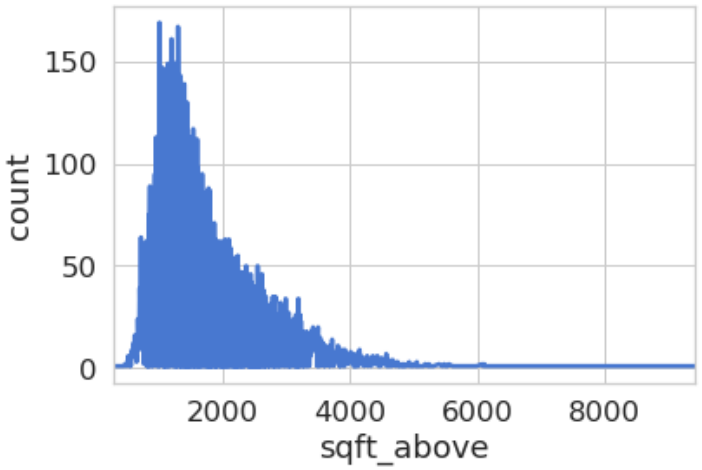

100
50
0
count
2000000
4000000
6000000
price

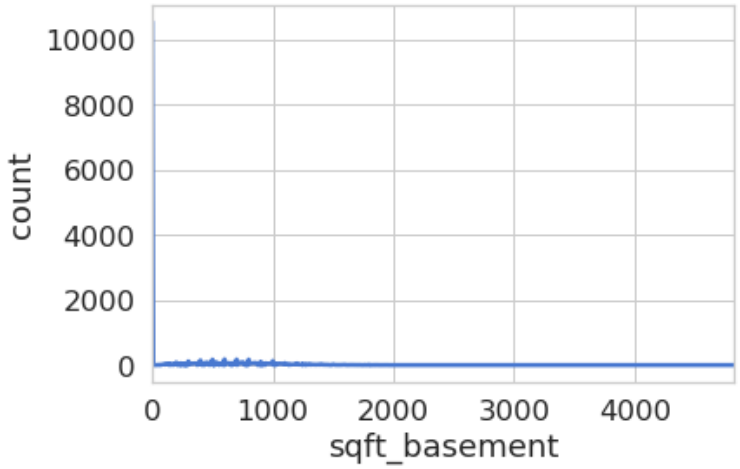

15000
10000
5000
0
count
0
500
1000
1500
yr_renovated

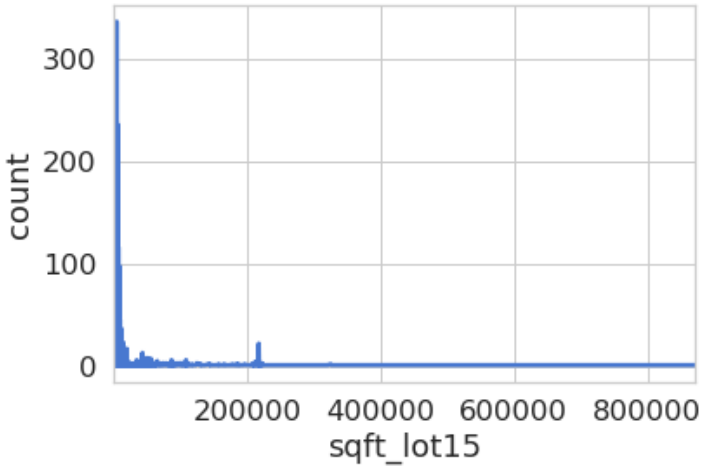

100
75
50
25
0
count
2500
5000
7500
10000
12500
sqft_living

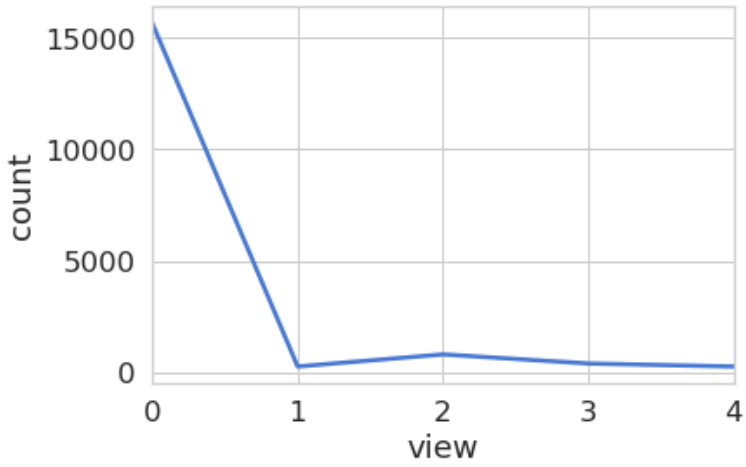

300
200
100
0
count
500000
1000000
1500000
sqft_lot

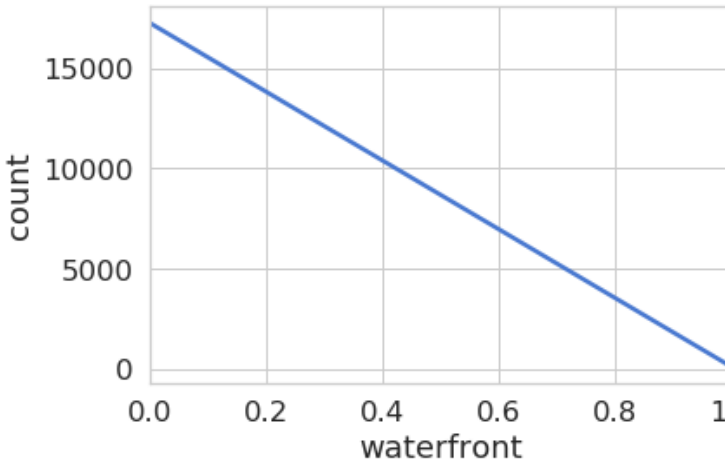

count
150
100
50
0
1000
2000
3000
4000
5000
6000
sqft_living15

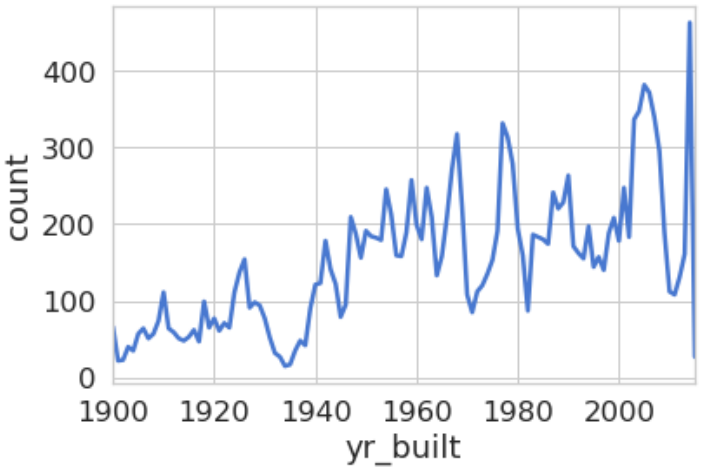

count
150
100
50
0
2000
4000
6000
8000
sqft_above

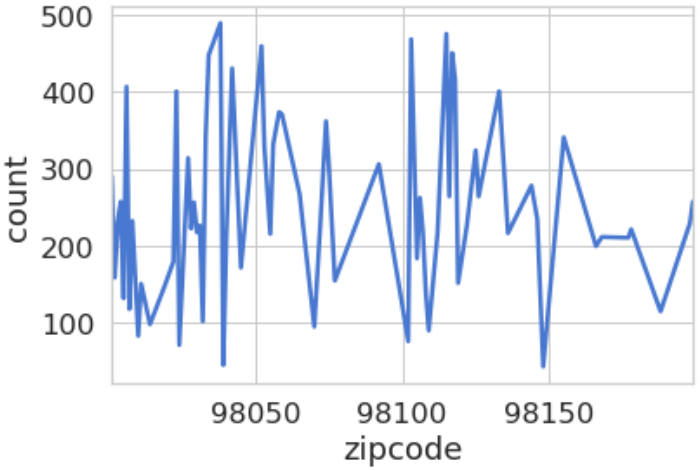

count
10000
8000
6000
4000
2000
0
0
1000
2000
3000
4000
sqft_basement


count
300
200
100
0
200000
400000
600000
800000
sqft_lot15


count
15000
10000
5000
0
0
1
2
3
4
view


count
15000
10000
5000
0
0.0
0.2
0.4
0.6
0.8
waterfront


count
0
100
200
300
400
1900
1920
1940
1960
1980
2000
yr_built


count
100
200
300
400
500
98050
98100
98150
zipcode

1.5 TSNE plot

Figure 2: T-SNE Plot

1.6 KEY OBSERVATIONS

Refering to heatmap:-We have plotted here heatmap of correlation between all the pairs of feature(including price)

- As we can see here *sqft_living* and grade shows good correlation with price. But both are

also correlated to each other so it would be better to drop one of them as they are redundant precdictor.

- Date is least correlated to price. So, it can not be considered good feature for price prediction.
- Since bad feature selection may lead to bad result. So many algorithms have developed to tackle this.

Refering to boxplot:- We have plotted boxplot for catogoraial features. Here we have six catogorial feature. For each of the feature we have plotted tis against price.

- For grade and bathroom distribution shifts towards high price as the grade and bathroom increase which shows there is a relation between price and grade, price and bathrooms and can be used as feature for prediction.
- waterfront and bedrooms also show little amount of shift in distribution as the waterfront and bedrooms increases. which shows that higher the number of bedrooms and waterfront higher the price.

Refering to scatter plot:- For non-categorical features we have plotted scatter plot to check how individual feature is related to price.

- features like lat, $sqft_above$, $sqft_basement$, $sqft_living$, $sqft_living15$ shows postive relation with price because as the feature value increases price also increase.
- features like $sqft_lot$, $sqft_lot15$, shows negative relation with price because as the feature value decreases price also decreases.
- features like long, $yr_renovated$, zipcode, date, yr_built does not show significant relation with price

Refering to distribution plot:- It shows how the feature is distributed. For example data points having 3 bathrooms are more in number in trainingset. Like wise for price we can see costlier house are very less in training set campred to cheap houses.Like this we can get idea of all features.

Refering to T-sne plot:- Here we have plotted the training data. We have run unsupervised clustering(t-sne)method on data. As we can see data is clustered or grouped. It shows that houses are divided in groups and each group will have some average price. For each group we can calculate average value for each feature. In this way we can also answer reverse query. Like some one ask-I have this much amount of money what type of house I can buy?,I have this much amount of money can I buy house with four bedrooms etc.

2 Polynomial Regression