Exploratory Data Analysis

1. Introduction:

After wrangling and cleaning process, 21419 rows and 17 columns left in the dataset. Datatypes are : category(5), float64(2), int64(9), object(1).

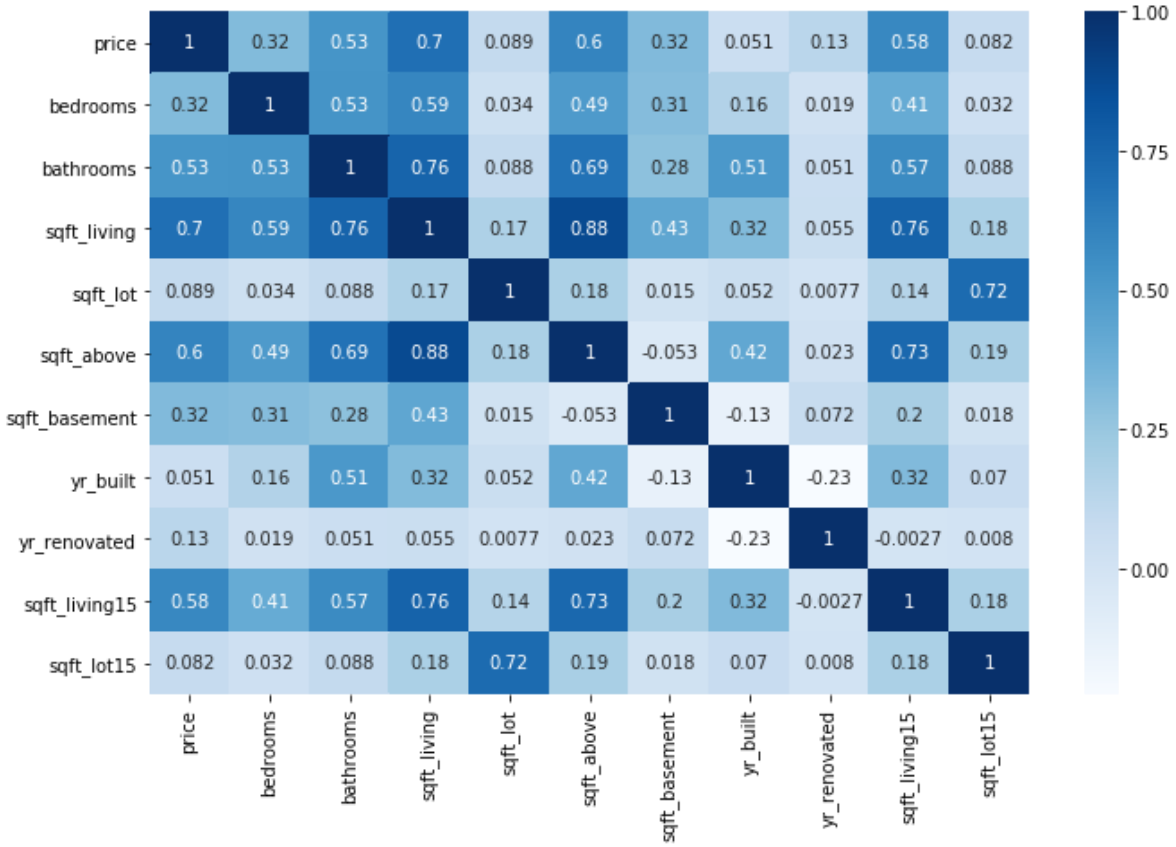
My main question is **"Is it possible to predict the sale price of a house from information about that house provided in the dataset, such as square footage of the home, number of bedrooms, number of bathrooms, number of floors, condition, grade, etc?**"

Now, I will deep dive into the details about features and examine their relationships with price. I will try to visualize my analysis with appropriate plots.

2. Correlations and Associations between Variables

Let us analyze now the relationship between the independent variables available in the dataset and the dependent variable that we are trying to predict (i.e., price). This analysis should provide some interesting insights for our regression models. We will be using heat map, scatterplots and correlations coefficients (e.g., Pearson, Spearman) to explore potential associations between the variables.

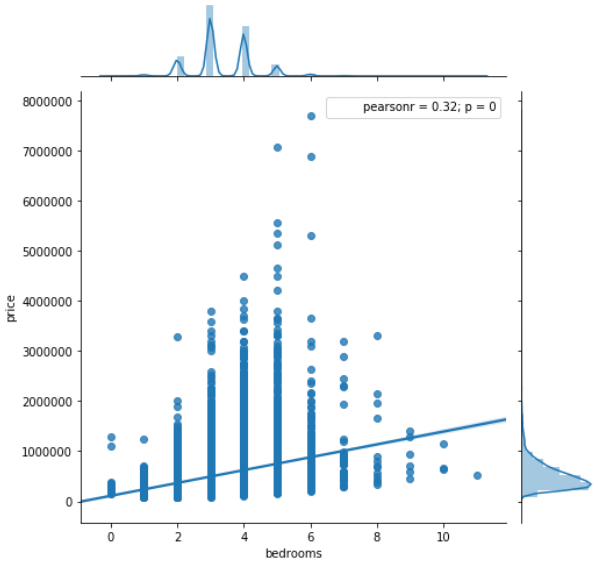
2.a. Continuous Variables

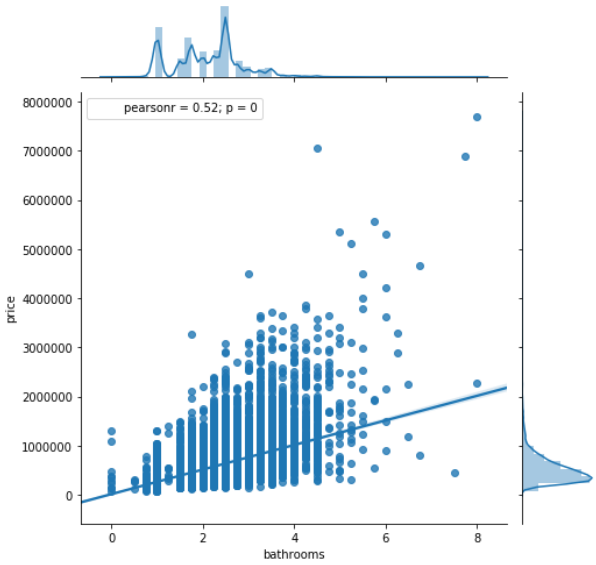


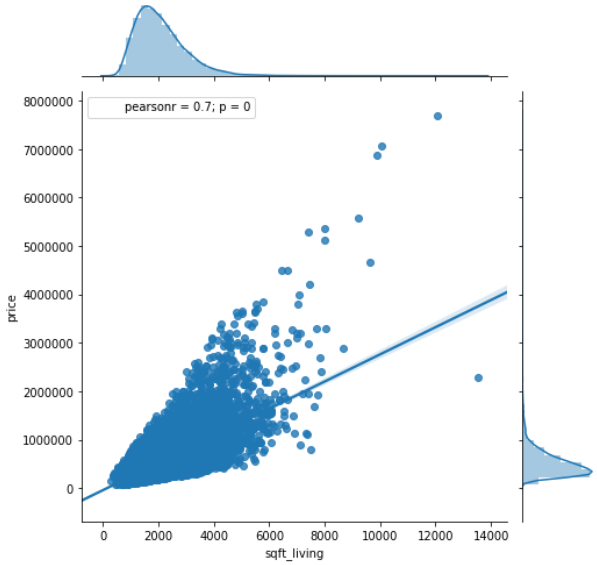
The correlation heat map provides a summary of correlation between the continuous variables in the data. The objective behind analyzing correlation between the

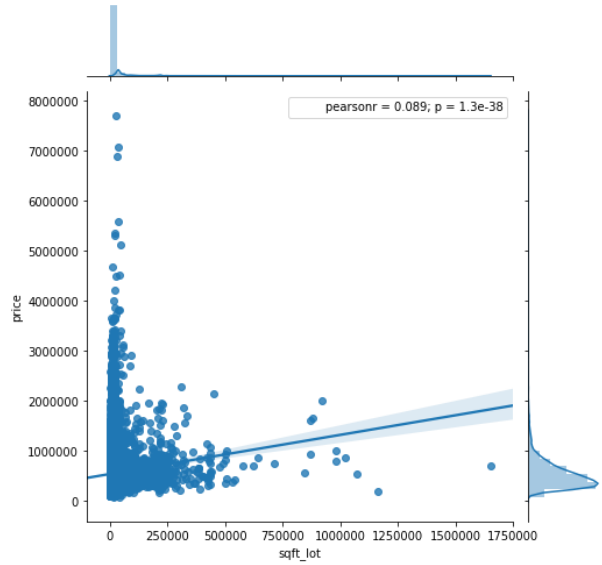
continuous variables in the data was to identify variables that have significant linear relationship with price and those who do not. Further, the table helps to identify relationship between potential predictors.

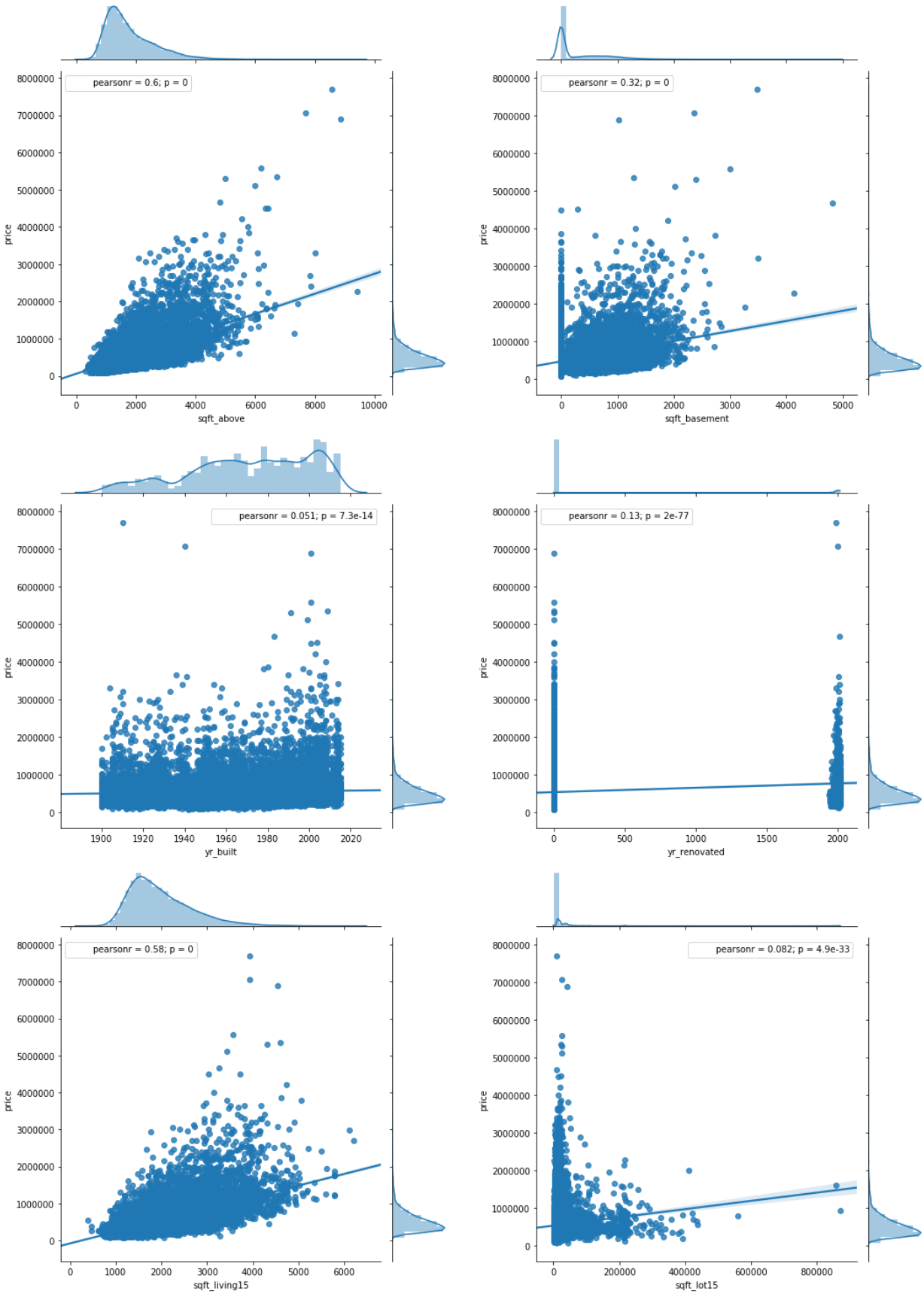
Now let's make a deeper analysis about the relationship between **'bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'sqft_living15', 'sqft_lot15'** and **'price'.** Since variables are measured on a continuous scale, we can use Pearson's coefficient r to measures the strength and direction of the relationship.










pearsonr = 0.6; p = 0
price
sqft_above
pearsonr = 0.32; p = 0
price
sqft_basement
pearsonr = 0.051; p = 7.3e-14
price
yr_built
pearsonr = 0.13; p = 2e-77
price
yr_renovated
pearsonr = 0.58; p = 0
price
sqft_living15
pearsonr = 0.082; p = 4.9e-33
price
sqft_lot15

There is a clear linear association between the variable 'sqft_living' (pearsonr = 0.7), 'sqft_above' (pearsonr = 0.6), 'sqft_living15' (pearsonr = 0.58), and 'bathrooms'(r = 0.52) and 'price', indicating a strong positive relationship. The house prices increase with increase in these features of the houses.They should be a good predicator of house price.

The variables 'bedrooms' (pearsonr = 0.32), 'sqft_basement' (pearsonr = 0.32), 'yr_renovated' (pearsonr = 0.13) has low positive correlation with 'price'.

The variables sqft_lot(pearsonr = 0.089), sqft_lot15 (pearsonr = 0.082) and yr_built(pearsonr = 0.051) seem to be poorly related to price.

We can see that there are many zeros in the sqft_basement distribution (i.e., no basement). Similarly, there are many zeros in the yr_renovated variable. Let us create new two columns (**'basement_present', 'renovated'**), and change their types into category. If the columns 'sqft_basement' and 'yr_renovated' have a value, the new columns (**'basement_present', 'renovated')** will take “1” value, otherwise will take “0” value. After that, I will evaluate these two features as categorical data.

2.b. Categorical Variables

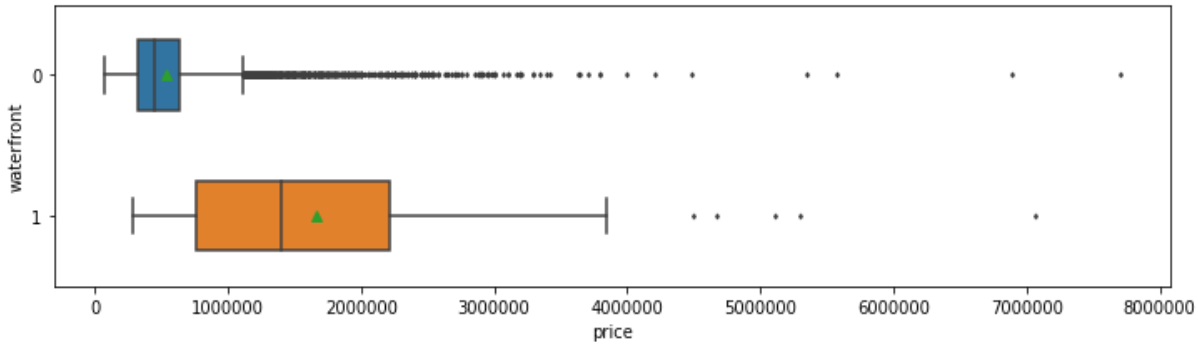
Now let's analyze the relationship between house price and the categorical variables ('floors', 'waterfront', 'view', 'condition', 'grade', ''basement_present', 'renovated', 'zipcode').

Firstly, we will try to assess if having a **waterfront, basement, and renovation** is related to a higher house value. We can use boxplots and a point-biserial correlation coefficient to highlight the relationship between the two variables.

Waterfront:

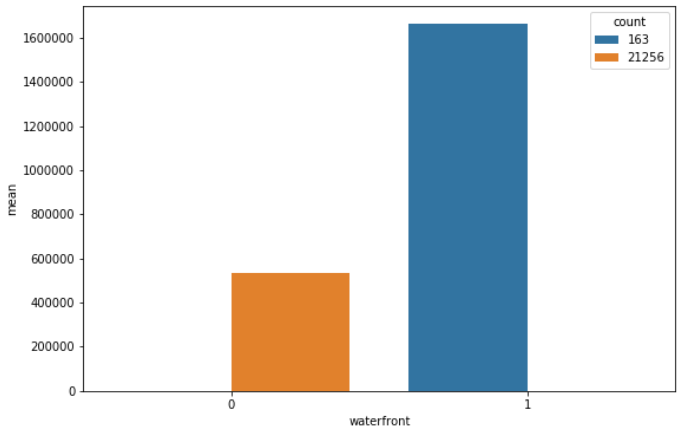
The waterfront feature is 1 if the property has a waterfront view and 0 if

it does not.



The no waterfront box plot is comparatively short. This suggests that overall, house prices in this group are very close to each other. The waterfront box plot is comparatively tall. This suggests that house prices differ greatly in this group.

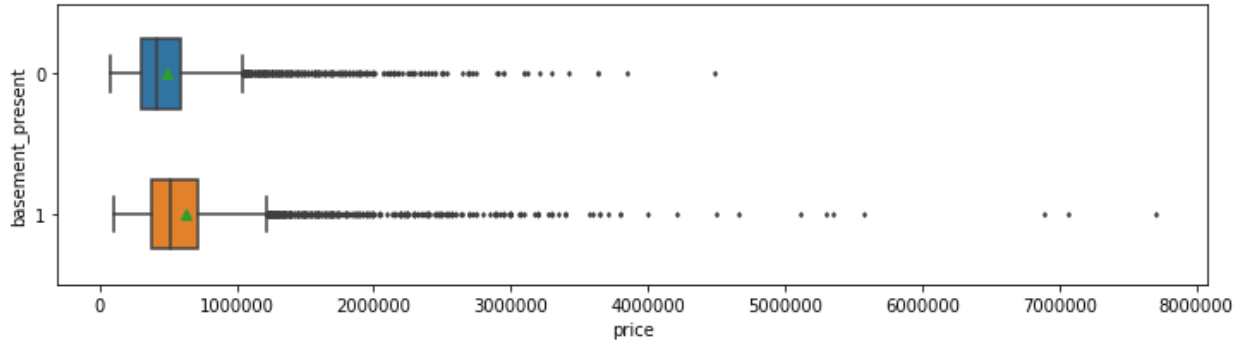
Waterfront	mean	min	max	count	std
0	5.331728e+05	78000.0	7700000.0	21256	3.416181e+05
1	1.661876e+06	285000.0	7062500.0	163	1.120372e+06



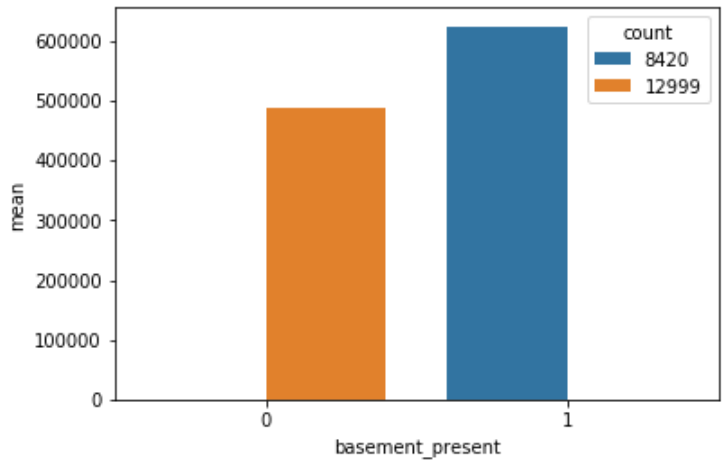
The presence of waterfront improves the house prices as we observe that most of the houses with waterfront have higher home prices. This is validated by a positive value of the point-biserial correlation (point biserial correlation r between price and waterfront is 0.26). However, in the dataset we have only 163 houses which have a waterfront view.

Basement:

The basement_present feature is 1 if the property has a basement and 0 if it does not.



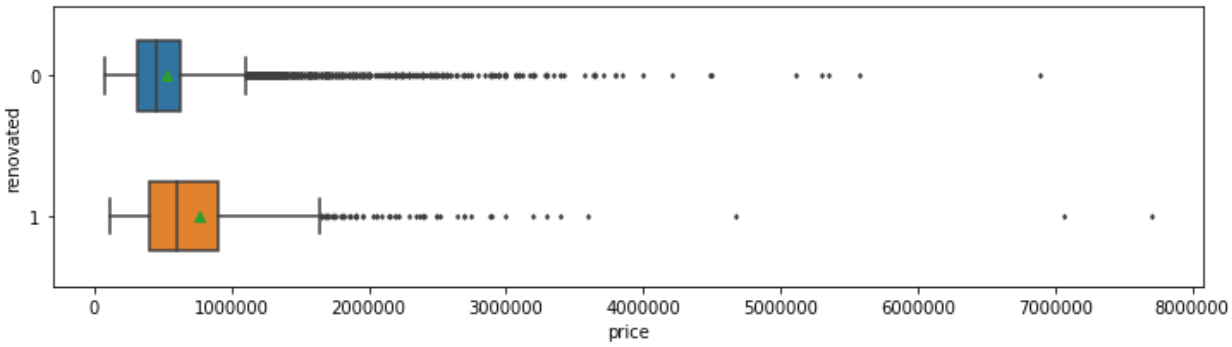
basement_present	mean	min	max	count	std
0	488515.826294	78000.0	4489000.0	12999	297379.112362
1	623965.418646	100000.0	7700000.0	8420	442262.585805



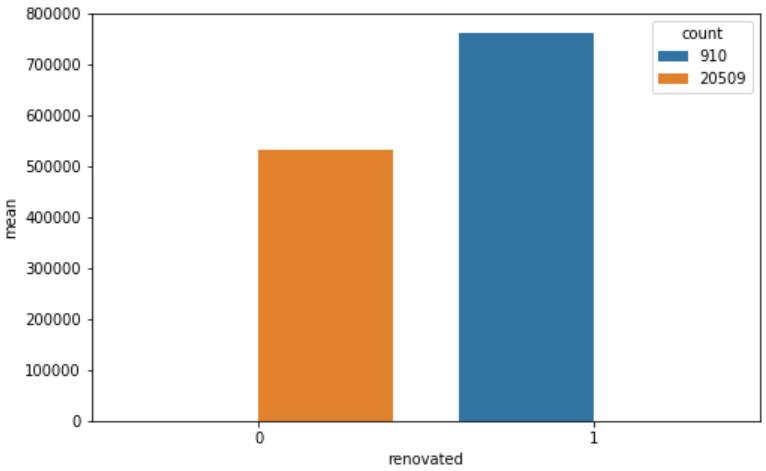
The presence of basement improves the house prices as we observe that most of the houses with basement have higher home prices. This is validated by a positive value of the point-biserial correlation (point biserial correlation r between price and basement_present is 0.18).

Renovated:

The renovated feature is 1 if the property is renovated and 0 if it is not.



renovated	mean	min	max	count	std
0	531984.914818	78000.0	6885000.0	20509	349656.223164
1	762118.058242	110000.0	7700000.0	910	608430.783572

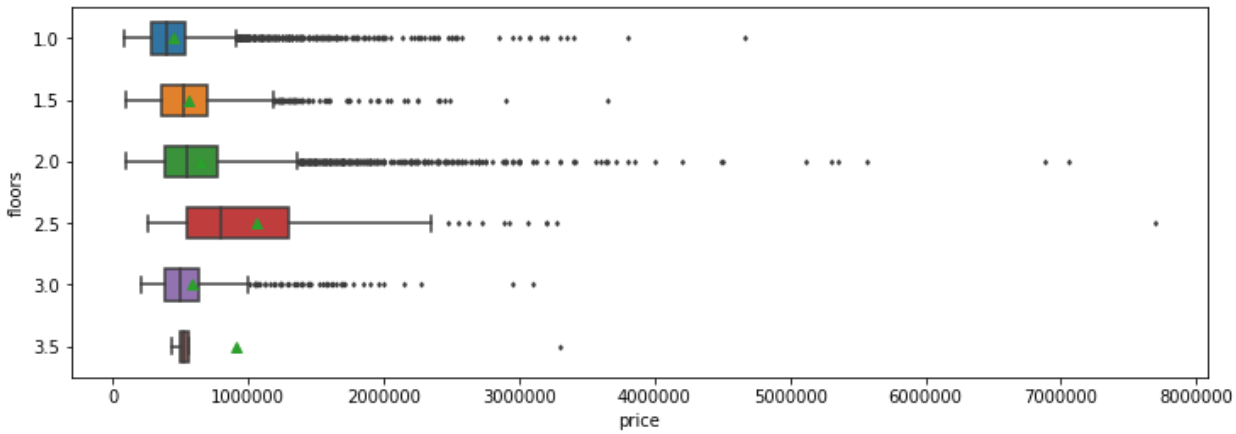


The renovation improves the house prices as we observe that most of the renovated houses have higher home prices. This is validated by a positive value of the point-biserial correlation (point biserial correlation r between price and renovated is 0.12). However, in the dataset we have only 910 renovated houses.

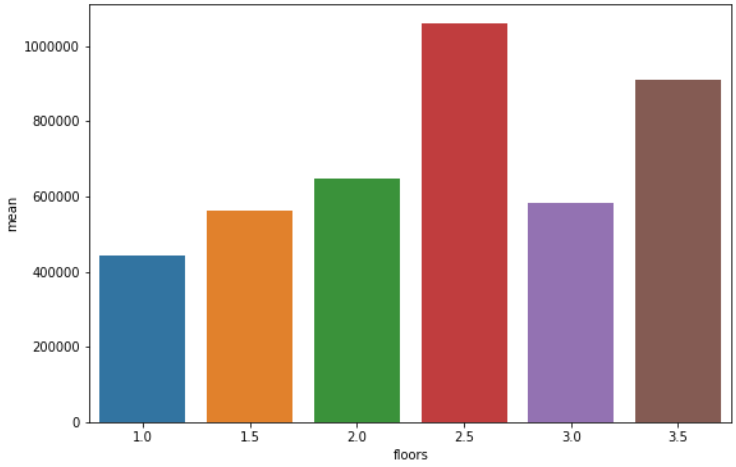
Secondly, we will try to analyze **'floors', 'view', 'condition', and 'grade' features**, and try to figure out whether they are related to a higher house value. We can use the Spearman's rank-order correlation to measure the strength and direction of the relationships between house price and these variables.

Floors:

Floors features shows the number of floors in properties. It can be half values (e.g. 1.5, 2.5 etc) which is most likely because of Mezzanine floors.



floors	mean	min	max	count	std
2.5	1.060346e+06	255000.0	7700000.0	161	8.582595e+05
3.5	9.102143e+05	435000.0	3300000.0	7	1.054669e+06
2.0	6.493117e+05	90000.0	7062500.0	8203	4.340177e+05
3.0	5.832602e+05	205000.0	3100000.0	609	3.389904e+05
1.5	5.619477e+05	92000.0	3650000.0	1888	3.034928e+05
1.0	4.439819e+05	78000.0	4668000.0	10551	2.639276e+05



10,551 of the properties are on asingle floor, 8,203 on 2 floors and 1,888 on 1.5 floors. The number with different values (up to a maximum of 3.5) are much smaller. The number of floors have a positive correlation with the price. This is validated by a positive value of the spearman correlation r (r = 0.32) between price and floors. This relationship is strong for the houses with fewer than 3 floors, and weak for the houses with 3 or 3.5 floors.

View:

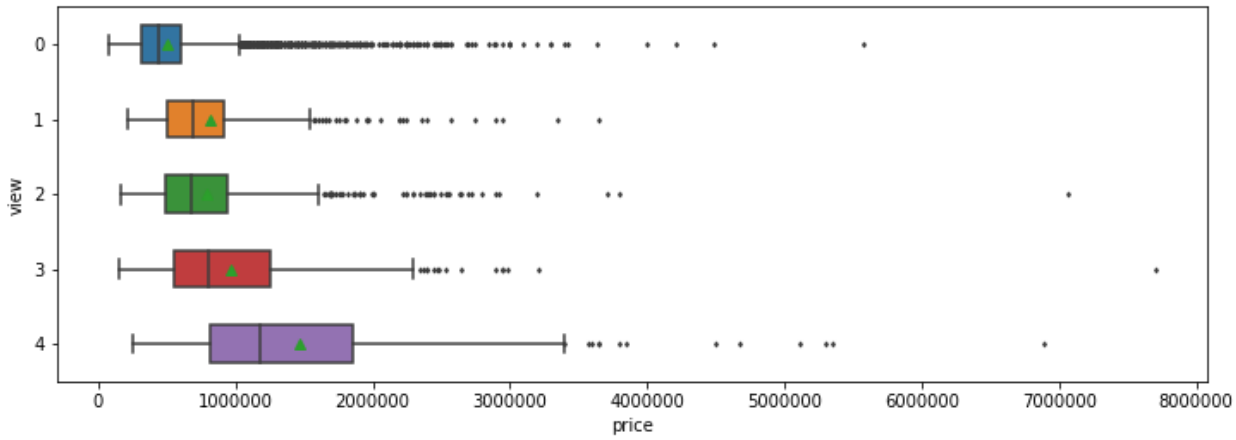
This feature is explained as **"view: Has been viewed"** in the Kaggle website. I made a further investigation about this feature from King County official website and I reached out these results below:

TOTAL VIEW QUALITY: This is the sum of all view's quality. The view's quality can vary from 0 to 4, in 5 different categories; Puget Sound, City/Territorial, Lake Washington/Sammamish, Mountain, and Small Lake/River.
https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=k

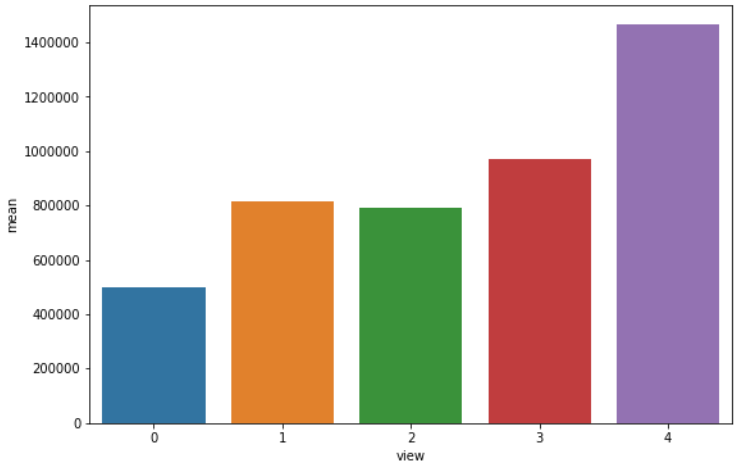
VIEW: For each classification will display blank for no view or "Fair", "Average", "Good" or "Excellent" to reflect the quality of view for that unit.
https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=k

Later they changed the quality of view grades as: 0 = Unknown, 1 = Fair, 2 = Average, 3 = Good, 4 = Excellent
https://www5.kingcounty.gov/sdc/Metadata.aspx?Layer=parcel_extr

In our dataset “view” feature indicates the quality of view for the house.



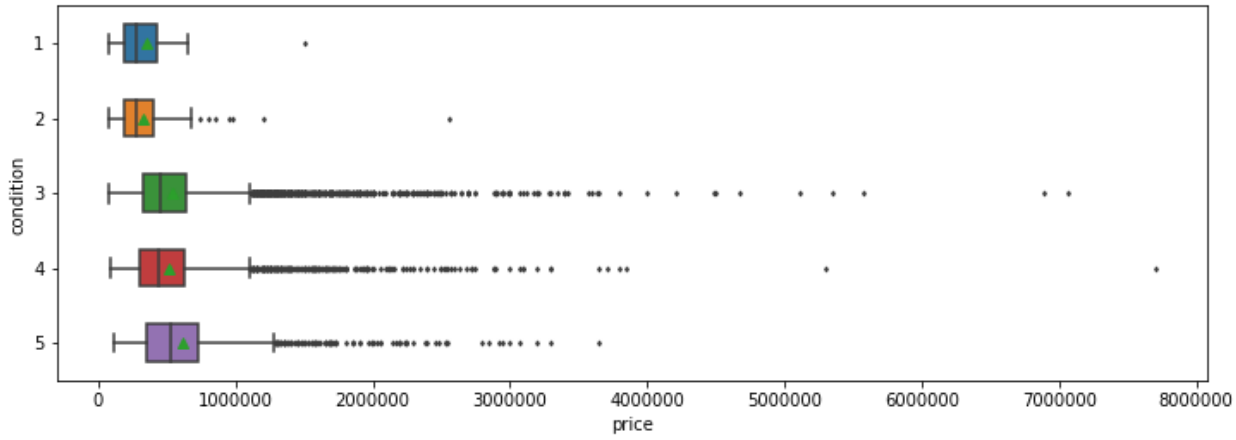
view	mean	min	max	count	std
4	1.465751e+06	252000.0	6885000.0	316	955755.394284
3	9.715104e+05	154000.0	7700000.0	507	613852.966176
1	8.132847e+05	217000.0	3650000.0	331	511395.259933
2	7.930803e+05	169317.0	7062500.0	960	510334.295352
0	4.981983e+05	78000.0	5570000.0	19305	287028.984825



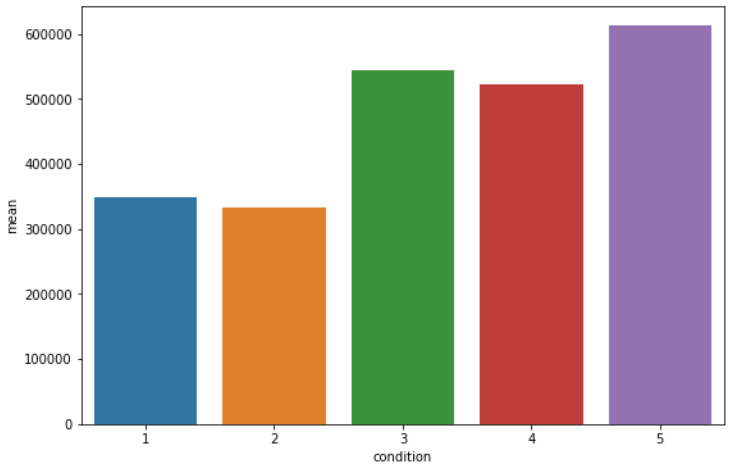
There is a positive correlation between view and price. This is validated by a positive value of the spearman correlation r (r = 0.29). It can be observed that the average house price associated with View 3 and View 4 is much higher than the average house price associated with View 1 and 2, while the average house price associated with View 1 is the lowest. However, 19305 houses’ view quality is 0. Only 823 houses’ view quality is 3 and more.

Condition:

The condition feature indicates the overall condition of the house (1 = Poor, 2 = Fair, 3 = Average, 4 = Good, 5 = Very Good).



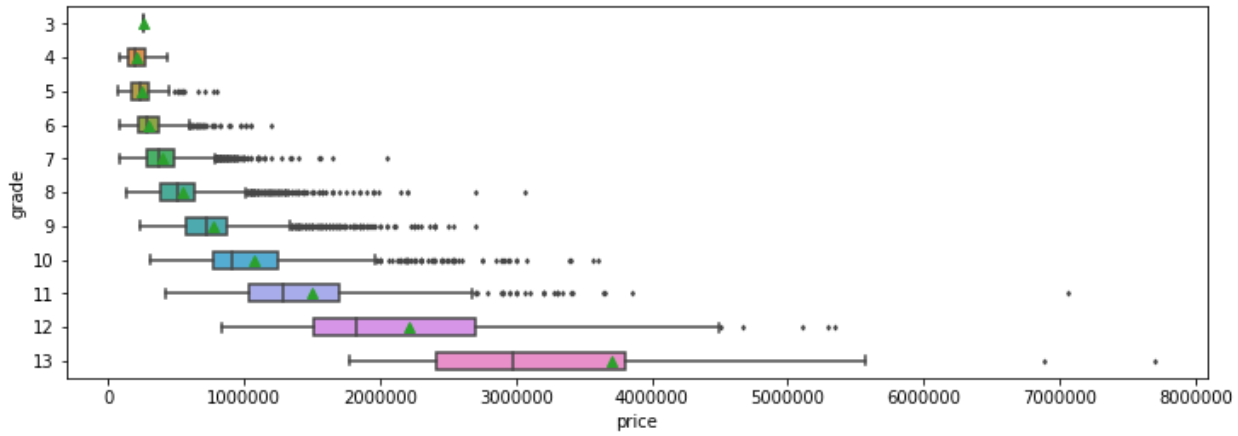
condition	mean	min	max	count	std
5	613111.545670	110000.0	3650000.0	1686	410884.734867
3	543854.461942	83000.0	7062500.0	13900	364900.711906
4	522210.459862	89000.0	7700000.0	5643	358171.886173
1	349480.357143	78000.0	1500000.0	28	274653.006112
2	333974.623457	80000.0	2555000.0	162	250749.326239



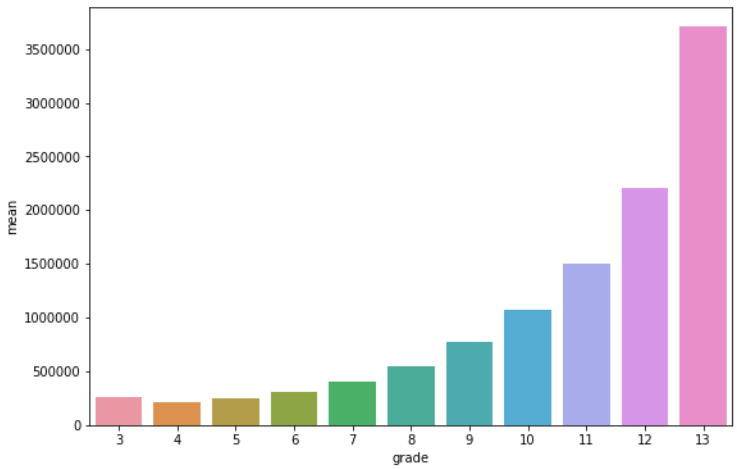
There is very low positive correlation between condition and price. This is validated by a value of the spearman correlation r (r = 0.016). But we can say that there is an increase in price as the condition of the house increases from {1,2} to {3,4,5}.

Grade:

Grade is overall grade given to the housing unit, based on King County grading system. It represents the construction quality of improvements. Grades run from grade 1 (low) to 13 (high).



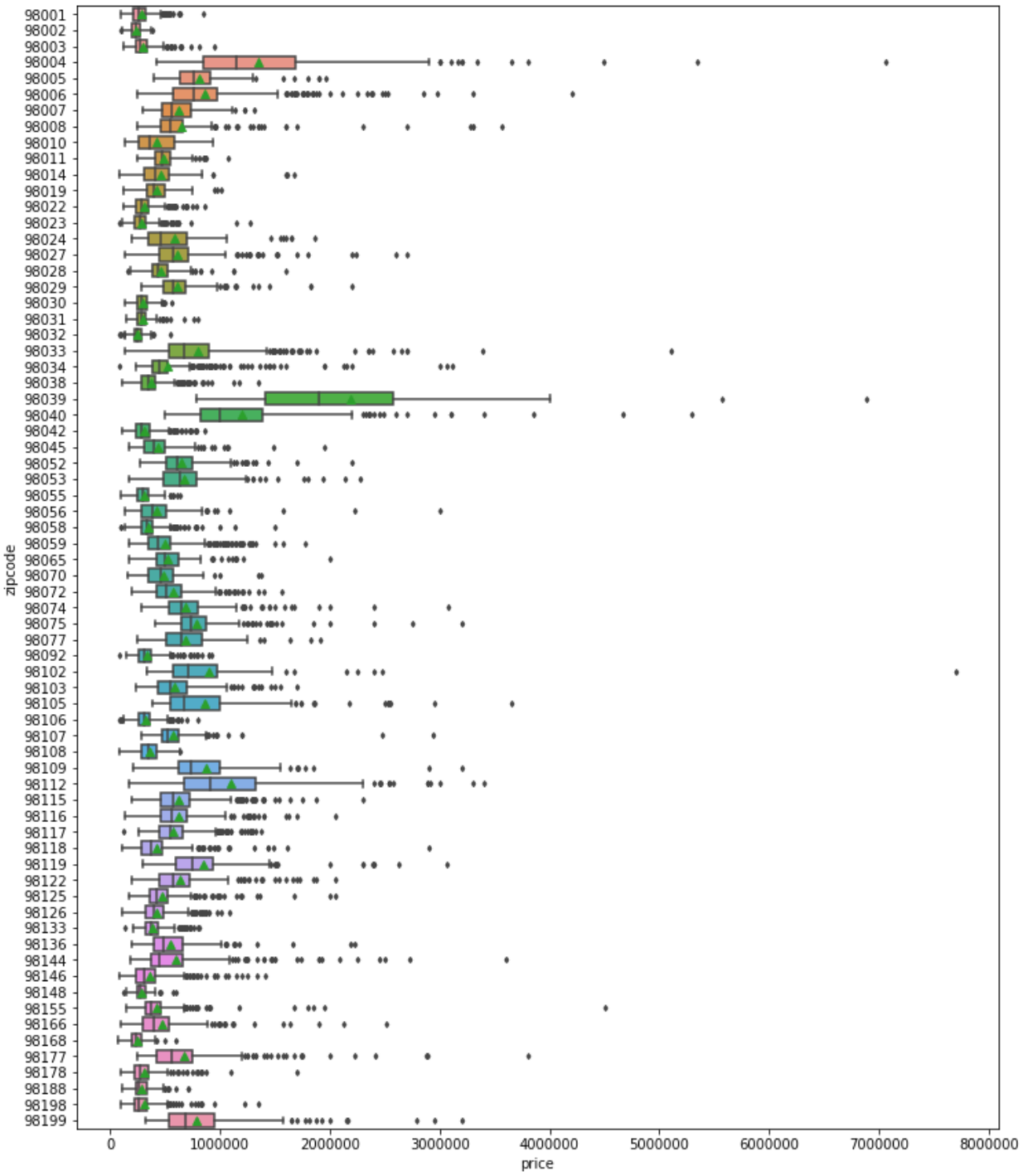
grade	mean	min	max	count	std
13	3.709615e+06	1780000.0	7700000.0	13	1.859450e+06
12	2.212521e+06	835000.0	5350000.0	88	1.029567e+06
11	1.498242e+06	420000.0	7062500.0	396	7.069597e+05
10	1.071612e+06	316000.0	3600000.0	1130	4.837897e+05
9	7.734201e+05	230000.0	2700000.0	2606	3.157671e+05
8	5.433418e+05	140000.0	3070000.0	6041	2.175537e+05
7	4.036257e+05	90000.0	2050000.0	8888	1.555405e+05
6	3.042484e+05	84000.0	1200000.0	1995	1.226688e+05
3	2.620000e+05	262000.0	262000.0	1	NaN
5	2.504543e+05	78000.0	795000.0	234	1.172046e+05
4	2.120019e+05	80000.0	435000.0	27	9.729450e+04



There is strong positive correlation between grade and price of the house. We can easily understand that from the graphs and the table. There is only 1 house with the grade 3. If we exclude this house, prices gradually increase with grade. This is validated by a value of the spearman correlation (r = 0.656).

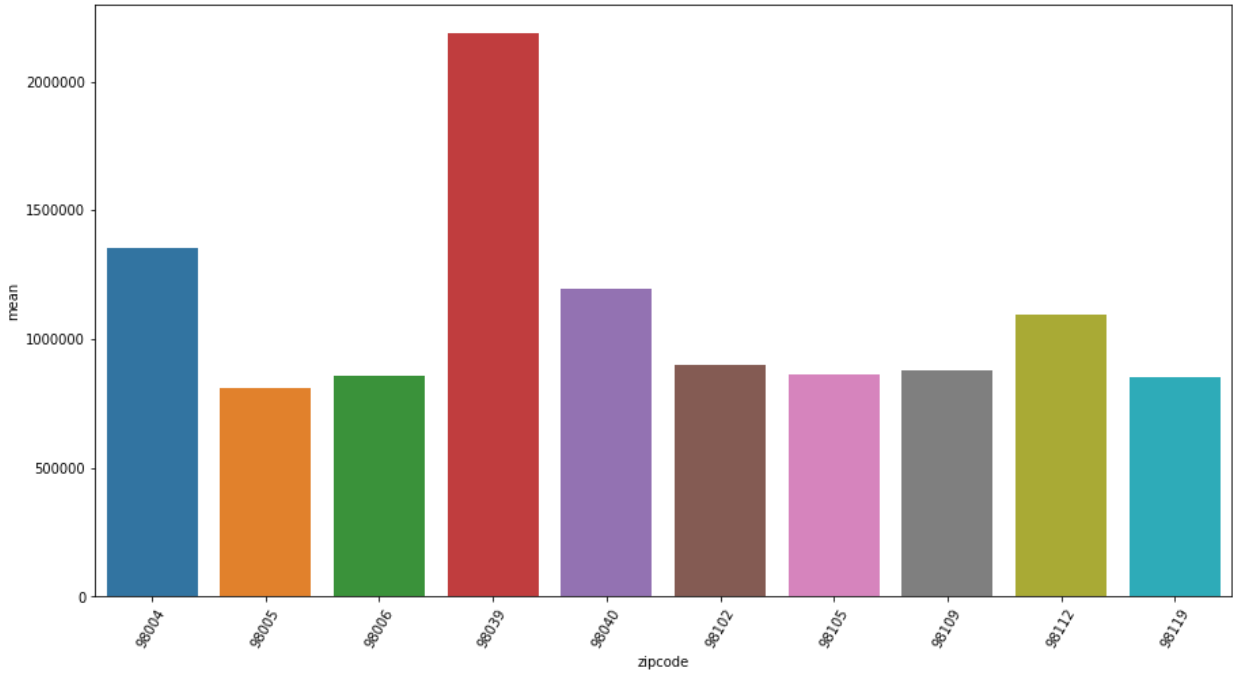
Thirdly, we will try to analyze **'zip code' feature**, and try to figure out whether it is related to a higher house value.

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



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
98105	8.628252e+05	380000.0	3650000.0	229	4.772876e+05
98006	8.578753e+05	247500.0	4208000.0	490	4.455127e+05
98119	8.494480e+05	300523.0	3065000.0	184	4.337225e+05
98005	8.101649e+05	400000.0	1960000.0	168	2.687537e+05

I have sorted the highest average house prices by zip codes and made a table with first 10 above. There are total 2198 houses in these 10 zip code areas.



We can clearly see that average house prices in the first four zip code areas (98039, 98004, 98040, and 98112) are higher than the other areas. There are total 914 houses in these four zip code areas.