Data Science Career Track
Capstone Project-1
Milestone Report

King County
House Price Prediction

Gokmen Oran

February 2019

1. Introduction:

The price of a house is dependent on various factors like size or area, number of bedrooms, bathrooms, floors, location of the house, the price of other houses, and many other factors. Real estate investors would like to find out the actual cost of the house in order to buy and sell real estate properties. They will lose money when they pay more than the current market cost of the house and when they sell for less than current market cost. The banks also want to find the current market price for a house, when they use someone's house as collateral for loans. Sometimes loan applicant overvalues their house to borrow the maximum loan from the bank. Banks and financial institutions also provide mortgage loan to home buyers. Local home buyers want also predict the price of the house to find out if a seller is asking for too much. A local seller also wants to predict their house price and find out how much is a fair market price.

The 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?**"

I will use a dataset of house prices from King County, Seattle, State of Washington. The goal of this analysis is to predict the price of house in King County, based on the variables provided in the dataset. The dataset was uploaded to the Kaggle website by the user harlfoxem (Data Source: https://www.kaggle.com/harlfoxem/housesalesprediction). Unfortunately, the user has not indicated the source of the data but I note that King County has an open data platform where the data may have originated.

1.1. Data Specifications:

The dataset has 21 house features columns, along with 21613 observations. Rows are specifications of houses sold in King County between 05/02/2014 and 05/27/2015.

List of attributes and explanations of features below.

Attribute	Explanation
id	Notation for a house
date	Date house was sold
price	Price is prediction target
bedrooms	Number of Bedrooms/House
bathrooms	Number of bathrooms/House
sqft_living	Square footage of the home
sqft_lot	Square footage of the lot

floors	Floors (levels) in house
waterfront	House which has a view to a waterfront
view	Has been viewed
condition	How good the condition is (Overall)
grade	Overall grade given to the housing unit, based on King County grading system
sqft_above	Square footage of house apart from basement
sqft_basement	Square footage of the basement
yr_built	Built Year
yr_renovated	Year when house was renovated
zipcode	Zip Code
lat	Latitude coordinate
long	Longitude coordinate
sqft_living15	Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area
sqft_lot15	LotSize area in 2015(implies-- some renovations)

1.2. Data Preprocessing:

1.2.1. I loaded the dataset in csv format and read it in the jupyter notebook after importing necessary libraries. I applied **df.head()** and **df.info()** to see some basic information about dataset. There are 21613 entries, and all features look like have no non-null entries. I will check it one more.

1.2.2. I wrote the explanation of column names for better understanding. In the Kaggle website;

"View" feature is explained as "Has been viewed".

"**Sqft_living15**" is explained as "Living room area in 2015(implies-- some renovations) This might or might not have affected the lot size area".

"**Sqft_lot15**" is explained as "LotSize area in 2015(implies-- some renovations)".

The explanations do not make sense. I made a further investigation about these feature from King County official website and I reached out results below:

View: The quality of view grades are as: 0 = Unknown, 1 = Fair, 2 = Average, 3 = Good, 4 = Excellent . "View" feature indicates the quality of view for the house. https://www5.kingcounty.gov/sdc/Metadata.aspx?Layer=parcel_extr

Sqft_living15: The average of square footage of interior housing living space for the nearest 15 neighbors**.**

Sqft_lot15: The average of square footage of the land lots of the nearest 15 neighbors.

1.2.3. I checked both duplicated values and missing values in the dataset. “Id” column is a unique number for each houses. When I checked it (house_data.id.unique()), I saw that There are 177 dublicated rows. In these dublicated rows, everything is the same but price. I kept the last entries, dropped the first ones. After dropping duplicated rows , the dataset has 21436 entries. Also I checked any missing value with (.isnull().any()). There is no missing values.

1.2.4. 4. I reduce the dataset by dropping columns that won't be used during the analysis.I inspected the useless features. “id” column has only one unique value for each observations and that did not impact or change anything in the data. 'date', 'lat', 'long' columns are also has no meaning for analysis. For that reason, I dropped those four columns. (h_data.drop(['id', 'date', 'lat', 'long'], axis = 1, inplace=True)).

1.2.5. I checked the unique values for 'bedrooms', 'bathrooms', 'waterfront', 'view', 'condition', 'grade' columns. I saw that one house has 33 bedrooms and its 'sqft_living' (square footage of the house) is 1620. Most likely, there is a mistake and I dropped this line.

1.2.6. 10 bathroom and 13 bedrooms values are zeros. Most likely, there are mistakes about these observations. I did not want them to have negative effect on calculations, and I dropped them.

1.2.7. I changed datatypes of columns ‘floors’, 'waterfront', 'view', 'condition', 'grade' into category, and column 'zipcode' into str. As a result, total 21419 rows and 17 columns left in the dataset. Datatypes are : category(5), float64(2), int64(9), object(1). At first memory usage was 3.5 MB, and after changing the data types, it became 2.1 MB.

2. Exploratory Data Analysis

2.1. 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).

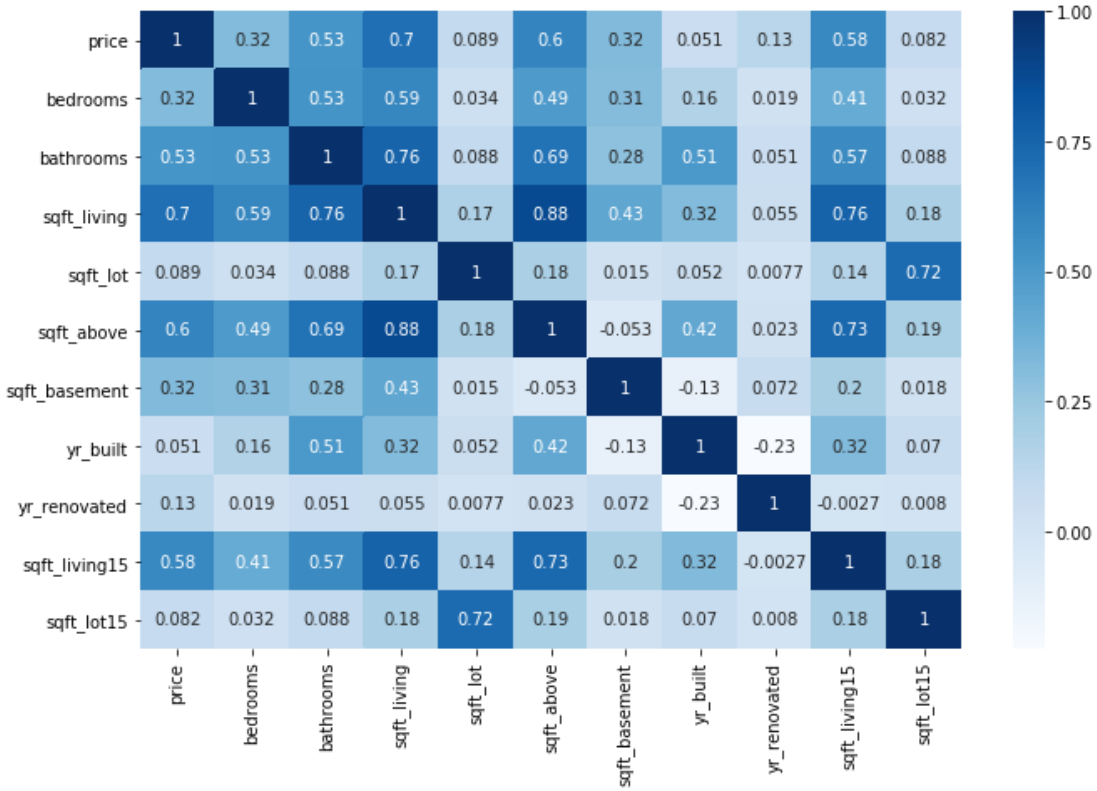
My main question was **"Is it possible to predict the sale price of a house from information about that house provided in the dataset, such as square footage of the home, number of bedrooms, number of bathrooms, number of floors, condition, grade, etc?**"

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.

2.2. Correlations and Associations between Variables

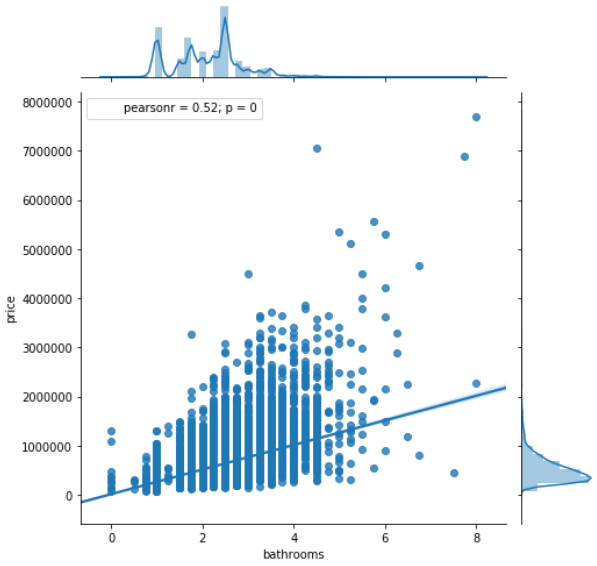
2.2.1. Continuous Variables

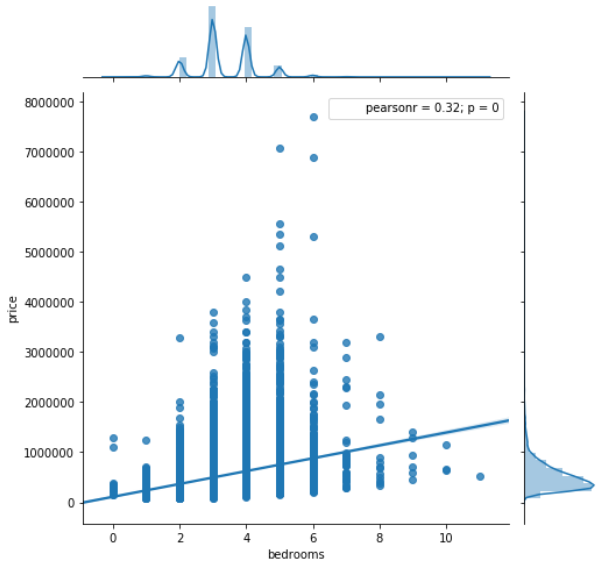
Let us analyze now the relationship between the independent variables available in the dataset and the dependent variable that we are trying to predict (i.e., price). This analysis should provide some interesting insights for our regression models. We will be using heat map, scatterplots and correlations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

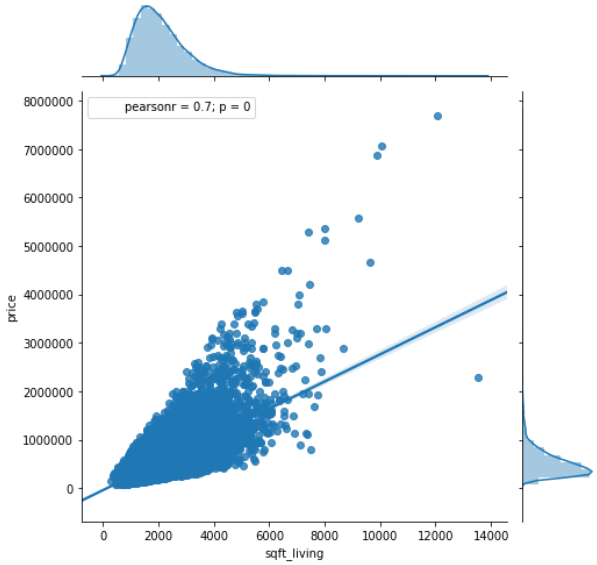


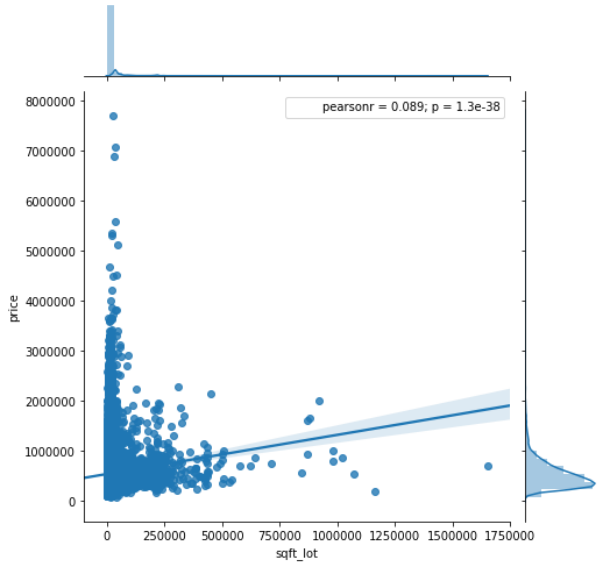
The correlation heat map provides a summary of correlation between the continuous variables in the data. The objective behind analyzing correlation between the continuous variables in the data was to identify variables that have significant linear relationship with price and those who do not. Further, the table helps to identify relationship between potential predictors.

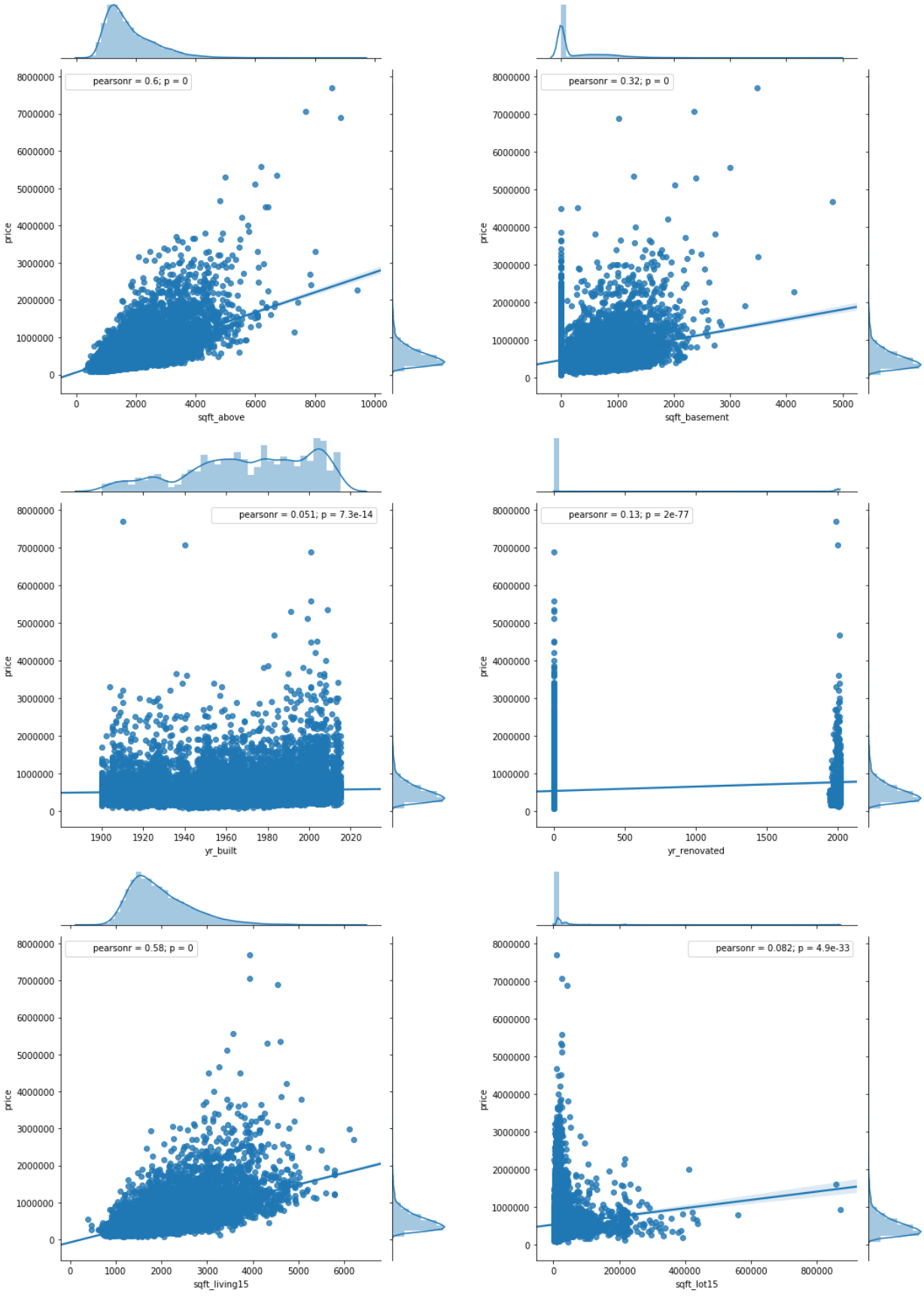
Now let's make a deeper analysis about the relationship between **'bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'sqft_living15', 'sqft_lot15'** and **'price'.** Since variables are measured on a continuous scale, we can use Pearson's coefficient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

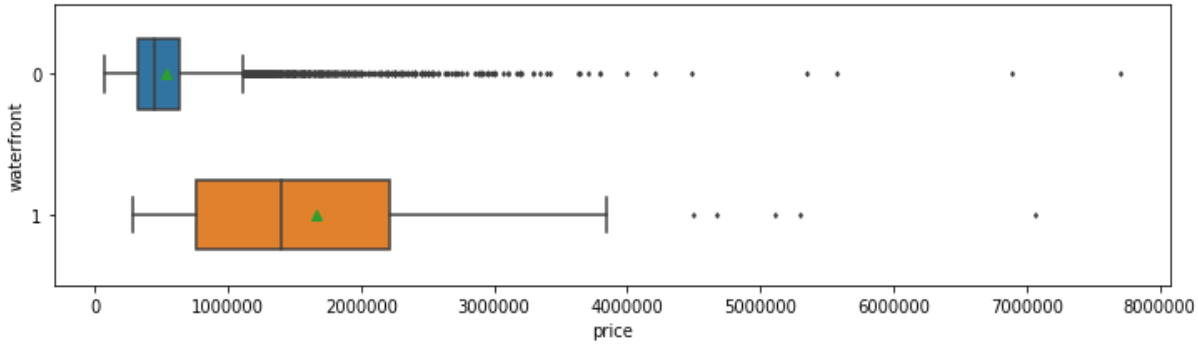
2.2.2. Categorical Variables

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.

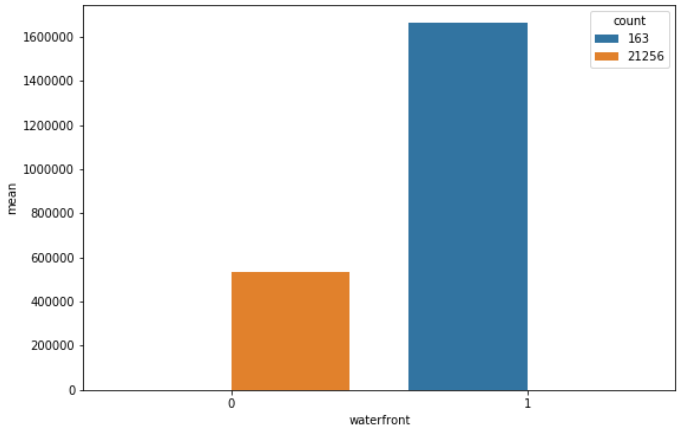
- **Waterfront:**

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.

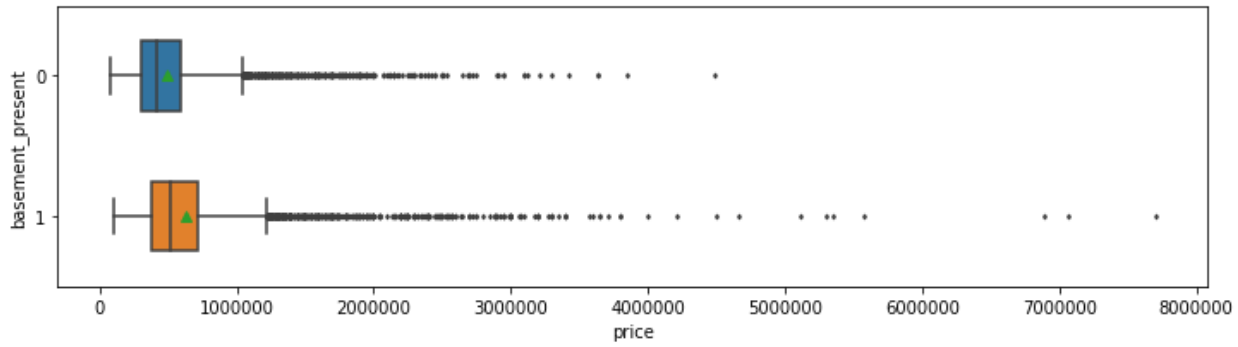
Water front	mean	min	max	count	std
0	5.331728e+05	78000.0	7700000.0	21256	3.416181e+05
1	1.661876e+06	285000.0	7062500.0	163	1.120372e+06



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.

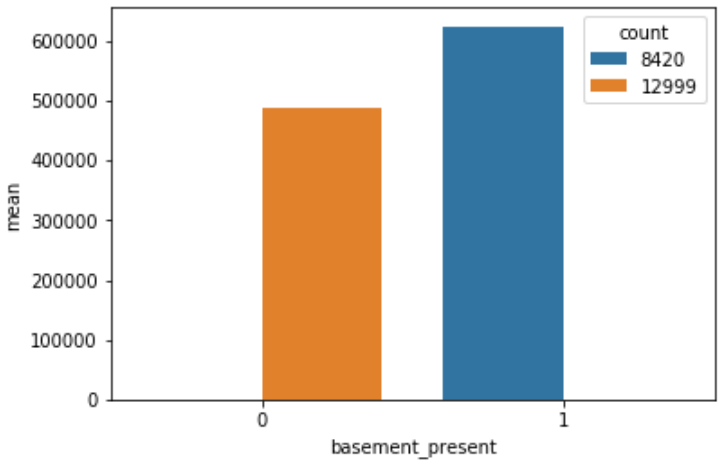
- **Basement:**

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

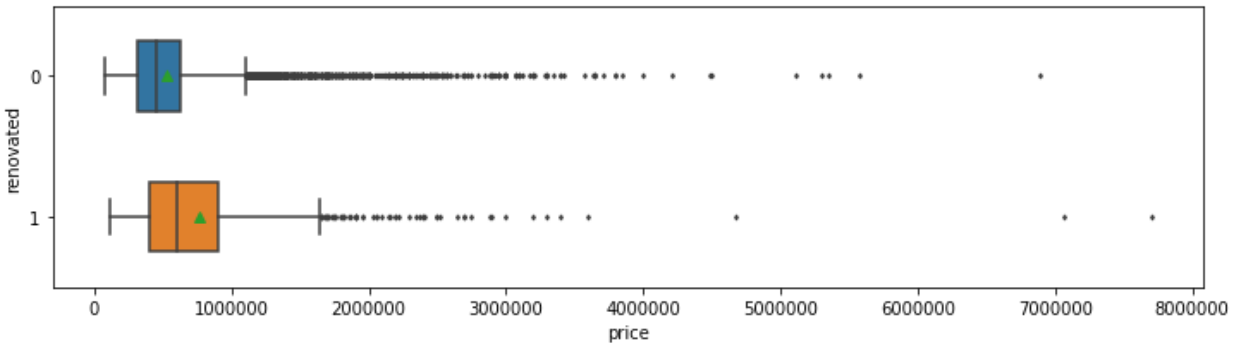
1	623965.418646	100000.0	7700000.0	8420	442262.585805



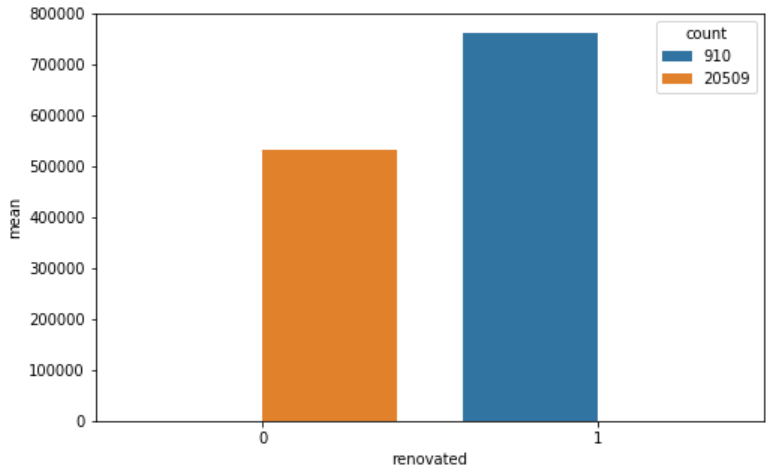
The presence of basement improves the house prices as we observe that most of the houses with basement have higher home prices. This is validated by a positive value of the point-biserial correlation (point biserial correlation r between price and basement_present is 0.18).

- **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

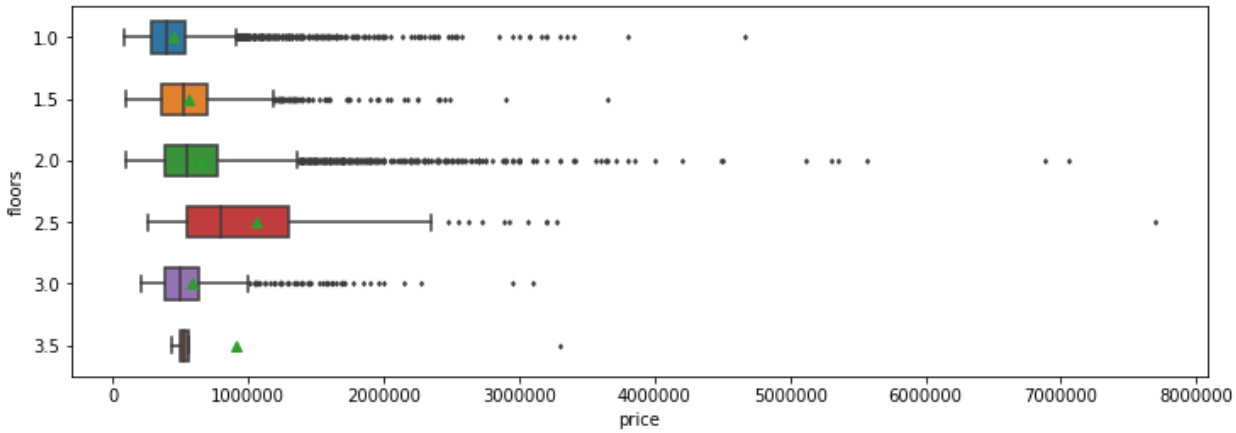


The renovation improves the house prices as we observe that most of the renovated houses have higher home prices. This is validated by a positive value of the point-biserial correlation (point biserial correlation r between price and renovated is 0.12). However, in the dataset we have only 910 renovated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

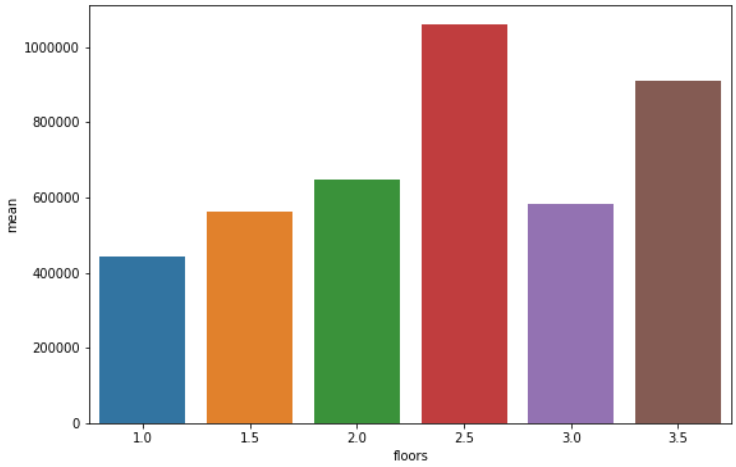
- **Floors:**

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

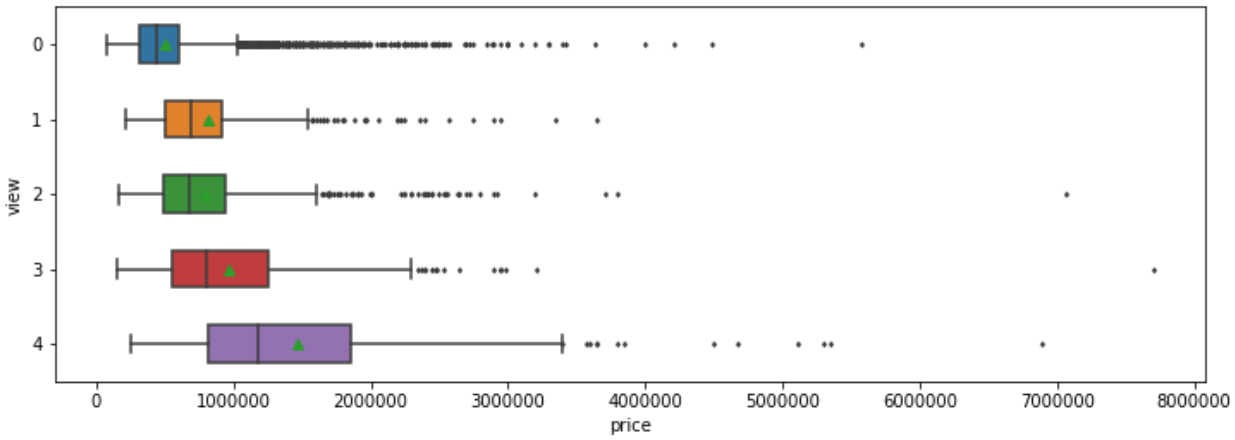
1.5	5.619477e+05	92000.0	3650000.0	1888	3.034928e+05
1.0	4.439819e+05	78000.0	4668000.0	10551	2.639276e+05



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.

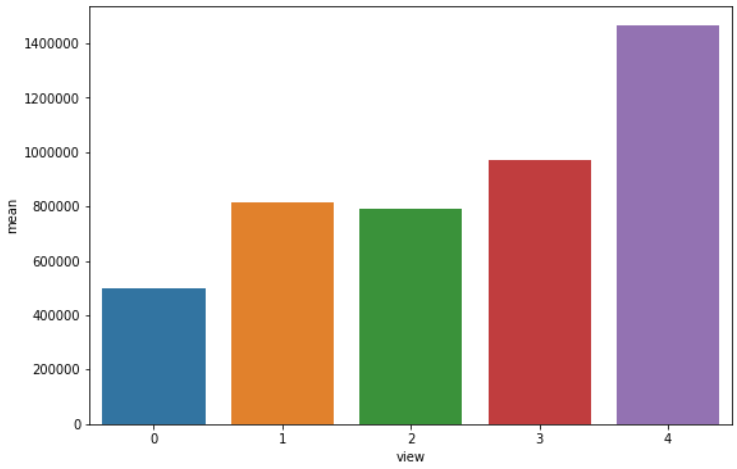
- **View:**

"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

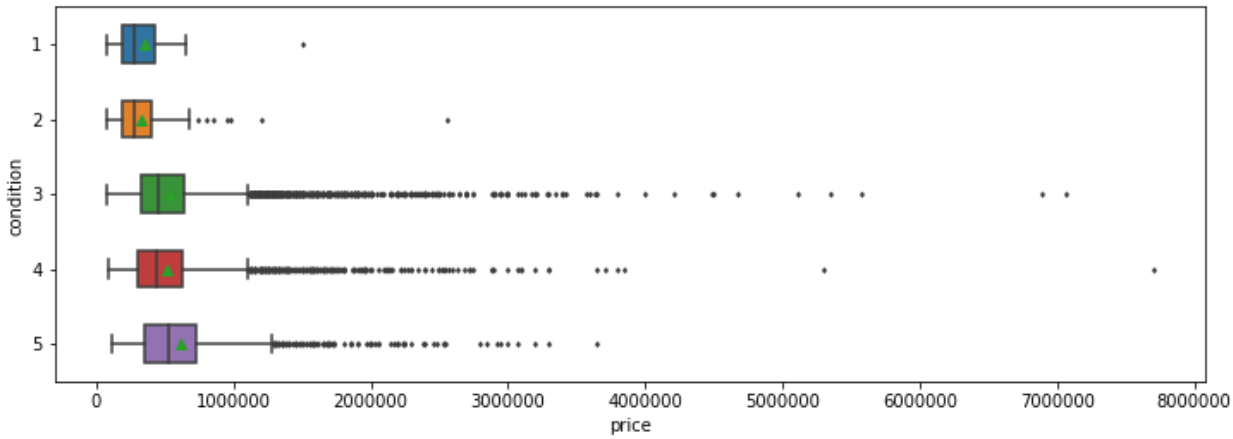
0	4.981983e+05	78000.0	5570000.0	19305	287028.984825



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.

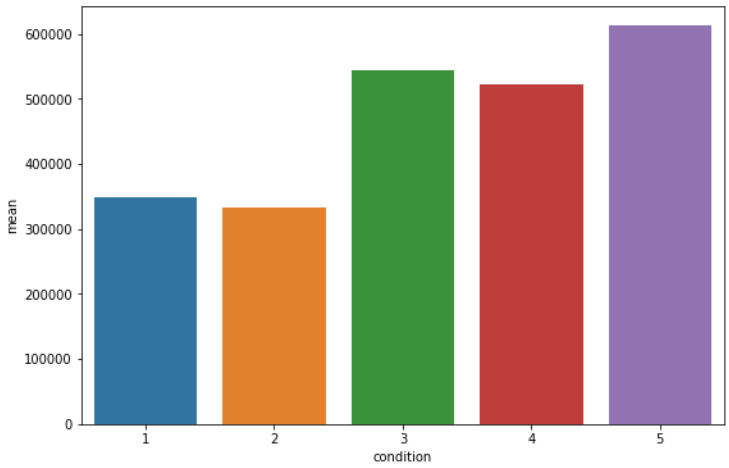
- **Condition:**

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

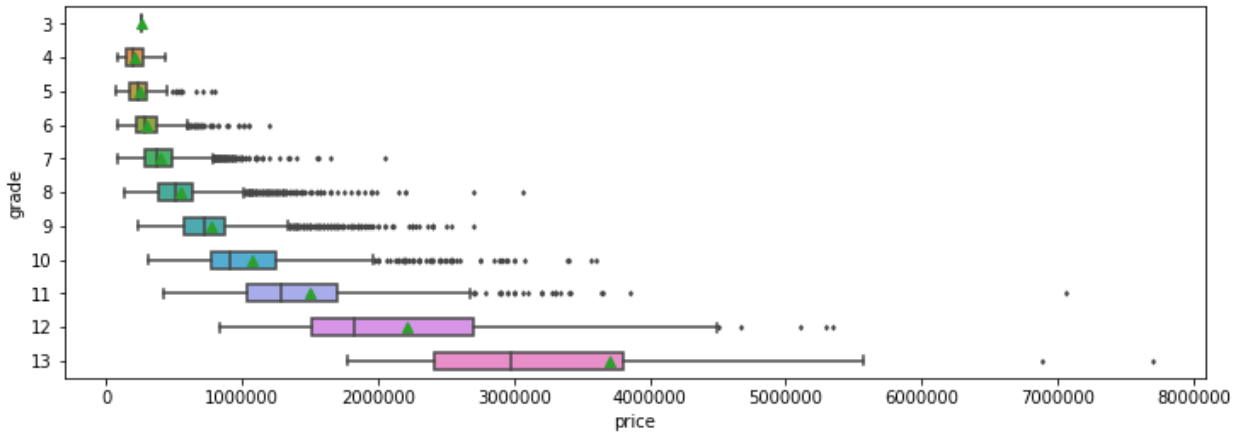
4	522210.459862	89000.0	7700000.0	5643	358171.886173
1	349480.357143	78000.0	1500000.0	28	274653.006112
2	333974.623457	80000.0	2555000.0	162	250749.326239



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}.

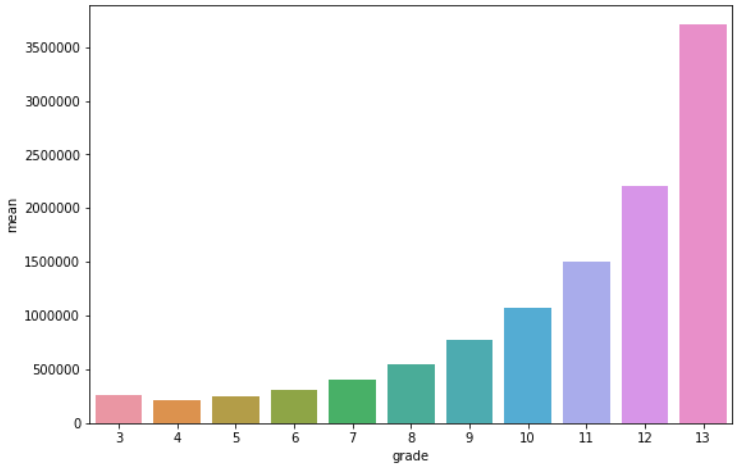
- **Grade:**

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

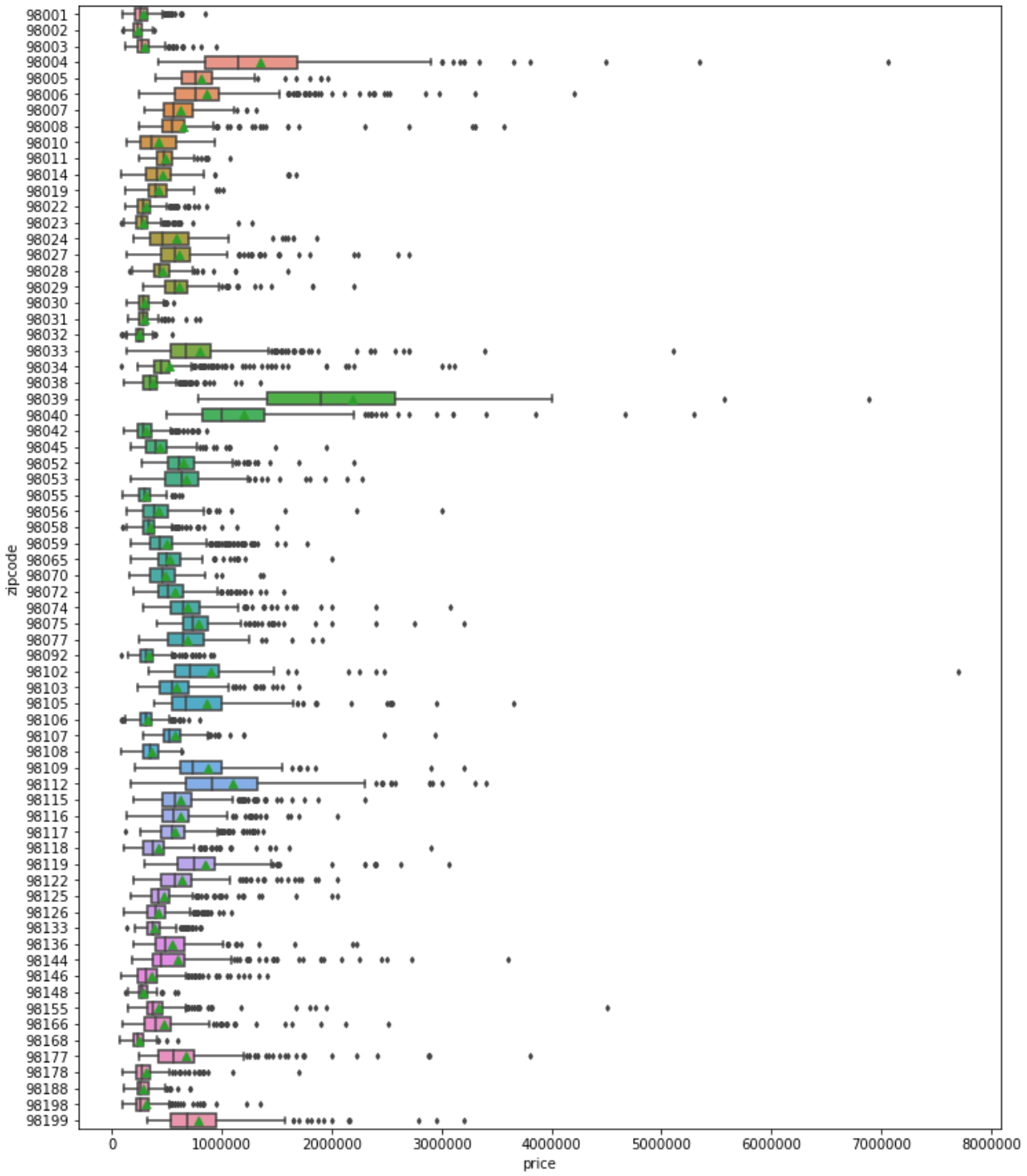


There is strong positive correlation between grade and price of the house. We can easily understand that from the graphs and the table. There is only 1 house with the grade 3. If we exclude this house, prices gradually increase with grade. This is validated by a value of the spearman correlation (r =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

2.2.3. Zip code Feature:

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

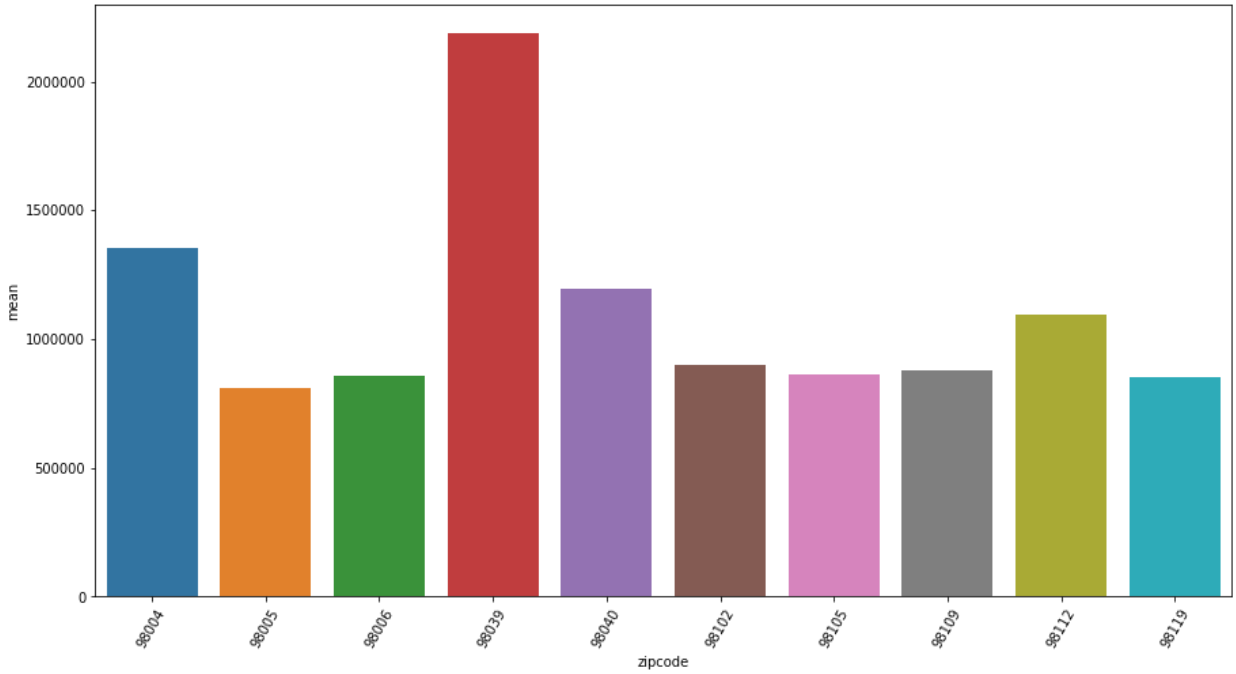
There are 70 unique zip codes in the data, and when we examined the box plot below, we cannot infer any kind of relationship with price. However, we can say the house prices in some specific zip code areas high than the other zip code areas.



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.

2.2.4 Inferential Statistics

- **Bedroom and Price**

I conducted a hypothesis test to check if there is no significant correlation between a number of bedroom and price. The p-value for the hypothesis test is less than the level of significance 0.05, so we reject the null hypothesis. So I support that there is a correlation between a number of bedrooms and price.

- **Bathrooms and Price**

I also conducted a hypothesis test to check the correlation between a number of bathrooms and price. The p-value for the hypothesis test is less than the level of significance 0.05, so we reject the null hypothesis and suggest that there is a correlation between a number of bathrooms and price.

- **Sqft_living and Price**

Similarly, I conducted a hypothesis test to check the correlation between sqft_living and price. The p-value for the hypothesis test is less than the level of significance 0.05, so we reject the null hypothesis and suggest that there is a correlation between sqft_living and price.

- **Sqft_above and Price**

Lastly, I conducted a hypothesis test to check if there is a correlation between sqft_above and price. The test suggests that there is a correlation between sqft_above and price.