House Price Prediction

Introduction

‘**Kings County** is located in the U.S. state of California. The population was 152,982 at the 2010 census.[5] The California Department of Finance estimated the county's population was 152,940 as of July 1, 2019.[4] The county seat is Hanford.[6]

Kings County comprises the Hanford-Lemoore, CA metropolitan statistical area, which is also included in the Visalia-Porterville-Hanford, CA combined statistical area. It is in the San Joaquin Valley, a rich agricultural region.

Problem Statement

You are looking to buy a property in Kings County but you are not sure about what are the factors that affect the prices of houses. You have heard from your colleagues that brokers sometimes charge more than the actual cost of the house. Also, Since this is your first house you are not aware of the factors that affect the cost of the house. So in order to make a smart decision you decided to get some insights using the Kings County dataset from Kaggle.

- Can you predict house prices using the features and data available of King County.
- List the top 10 Zip Codes that have the highest house prices and the lowest prices.
- List top five features that contribute to house pricing.
- Is there a correlation between the location and price?

Data Wrangling

Initial Observations

At the first look on the dataset it can be seen that:-

- There are 21 columns and 21613 rows with no null values.
- Date column is of object data type which is changed to Date time type.
- Properties listed in the data set are built between 1900 and 2015, also there are alot of houses that did not go under any kind of renovation(0 value till 75th Percentile).
- Mean of the condition column is 3.4 so we can assume most of the houses are above average in condition.

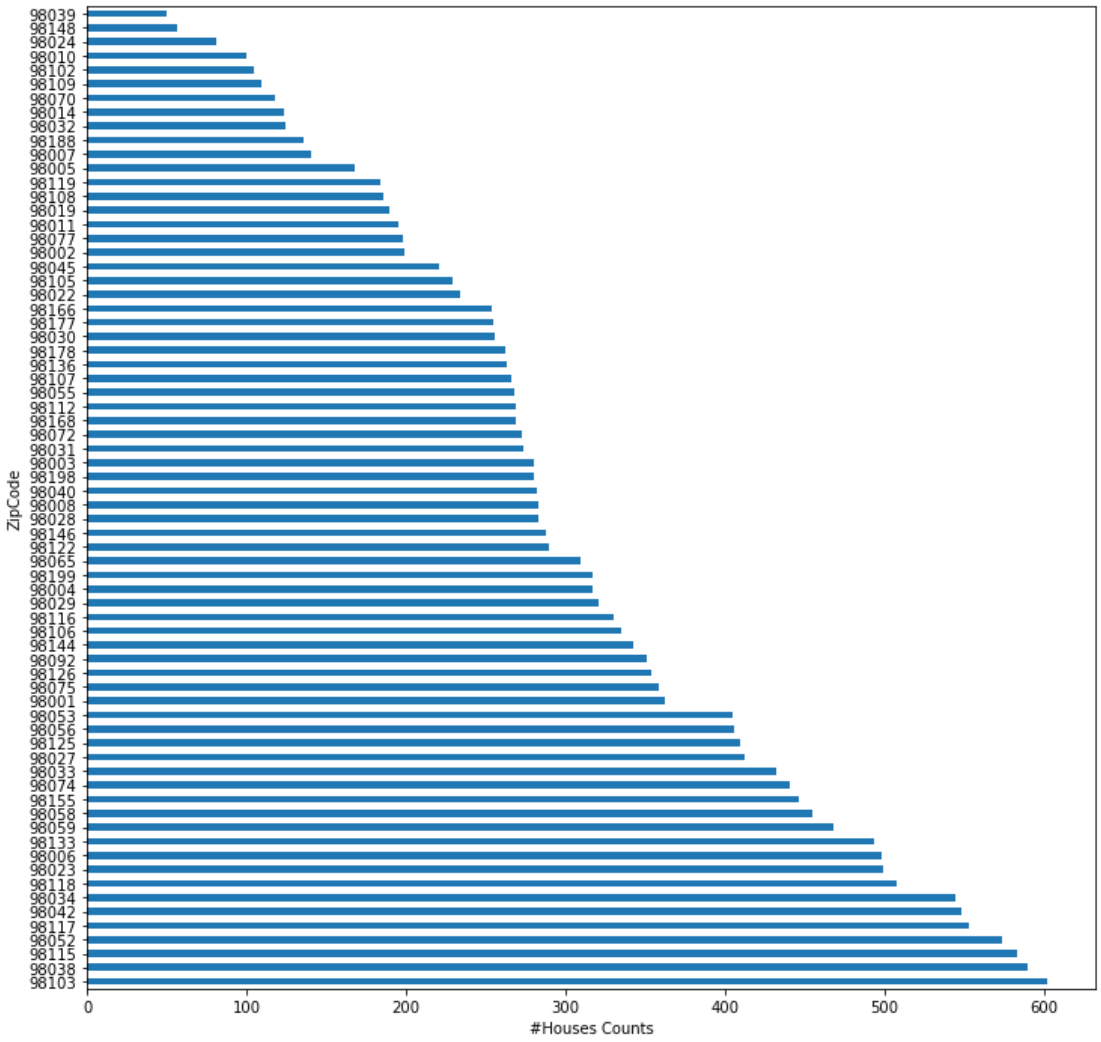
	count	mean	std	min	25%	50%	75%	max
id	21613.0	4.580302e+09	2.876566e+09	1.000102e+06	2.123049e+09	3.904930e+09	7.308900e+09	9.900000e+09
price	21613.0	5.400881e+05	3.671272e+05	7.500000e+04	3.219500e+05	4.500000e+05	6.450000e+05	7.700000e+06
bedrooms	21613.0	3.370842e+00	9.300618e-01	0.000000e+00	3.000000e+00	3.000000e+00	4.000000e+00	3.300000e+01
bathrooms	21613.0	2.114757e+00	7.701632e-01	0.000000e+00	1.750000e+00	2.250000e+00	2.500000e+00	8.000000e+00
sqft_living	21613.0	2.079900e+03	9.184409e+02	2.900000e+02	1.427000e+03	1.910000e+03	2.550000e+03	1.354000e+04
sqft_lot	21613.0	1.510697e+04	4.142051e+04	5.200000e+02	5.040000e+03	7.618000e+03	1.068800e+04	1.651359e+06
floors	21613.0	1.494309e+00	5.399889e-01	1.000000e+00	1.000000e+00	1.500000e+00	2.000000e+00	3.500000e+00
waterfront	21613.0	7.541757e-03	8.651720e-02	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
view	21613.0	2.343034e-01	7.663176e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	4.000000e+00
condition	21613.0	3.409430e+00	6.507430e-01	1.000000e+00	3.000000e+00	3.000000e+00	4.000000e+00	5.000000e+00
grade	21613.0	7.656873e+00	1.175459e+00	1.000000e+00	7.000000e+00	7.000000e+00	8.000000e+00	1.300000e+01
sqft_above	21613.0	1.788391e+03	8.280910e+02	2.900000e+02	1.190000e+03	1.560000e+03	2.210000e+03	9.410000e+03
sqft_basement	21613.0	2.915090e+02	4.425750e+02	0.000000e+00	0.000000e+00	0.000000e+00	5.600000e+02	4.820000e+03
yr_built	21613.0	1.971005e+03	2.937341e+01	1.900000e+03	1.951000e+03	1.975000e+03	1.997000e+03	2.015000e+03
yr_renovated	21613.0	8.440226e+01	4.016792e+02	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.015000e+03
zipcode	21613.0	9.807794e+04	5.350503e+01	9.800100e+04	9.803300e+04	9.806500e+04	9.811800e+04	9.819900e+04
lat	21613.0	4.756005e+01	1.385637e-01	4.715590e+01	4.747100e+01	4.757180e+01	4.767800e+01	4.777760e+01
long	21613.0	-1.222139e+02	1.408283e-01	-1.225190e+02	-1.223280e+02	-1.222300e+02	-1.221250e+02	-1.213150e+02
sqft_living15	21613.0	1.986552e+03	6.853913e+02	3.990000e+02	1.490000e+03	1.840000e+03	2.360000e+03	6.210000e+03
sqft_lot15	21613.0	1.276846e+04	2.730418e+04	6.510000e+02	5.100000e+03	7.620000e+03	1.008300e+04	8.712000e+05

Observed Inconsistencies

1. Date Column is of object data type. Changed to DateTime.
2. The Grade Column has values ranging from 1-13 but 2 is missing.

Initial Exploratory Data Analysis

1. 98103 is the most popular zip code with 602 listed selling properties.



2. There were few houses that had no bathrooms and no bedrooms but I am assuming they are really old and are not logged.
3. There was one house listed with a living area greater than lot and one floor and no basement. So this listing was dropped.
4. One listing was dropped with duplicate entries.

List of Possible Categorical Features

1. View
2. Waterfront
3. Condition
4. Grade
5. Zip Code

List of Possible Numerical Features

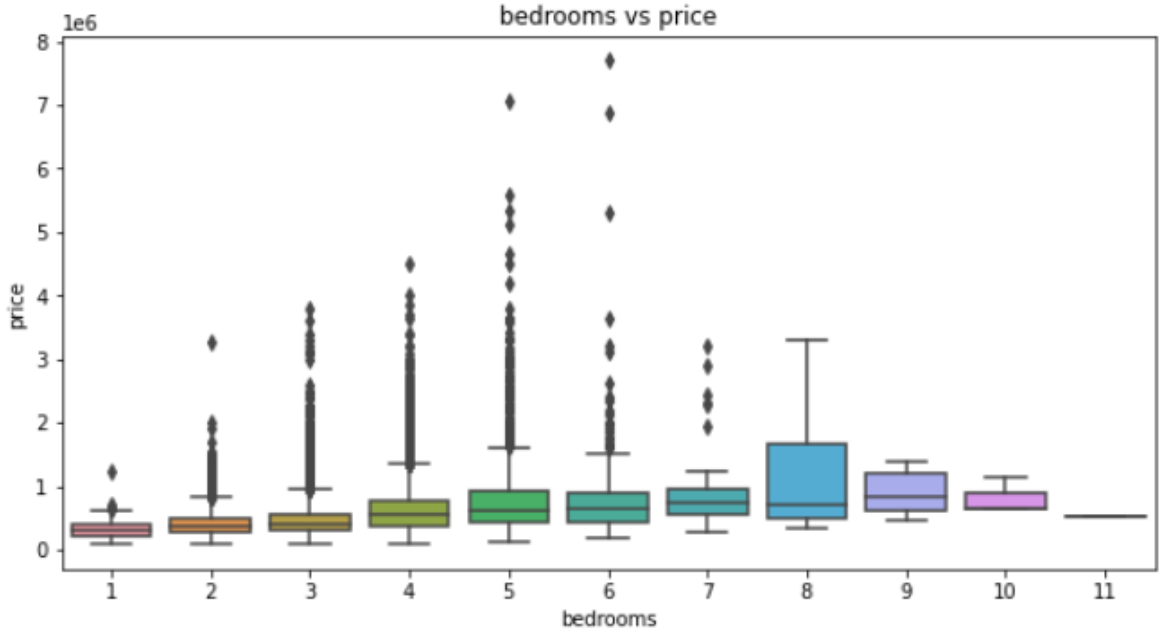
1. Bedrooms
2. Bathrooms
3. Sqft_living
4. Sqft_lot
5. Floors
6. Sqft_above
7. Sqft_basement
8. Yr_built
9. Yr_renovated
10. Lat
11. Long
12. Sqft_living15
13. Sqft_lot15

Target Feature

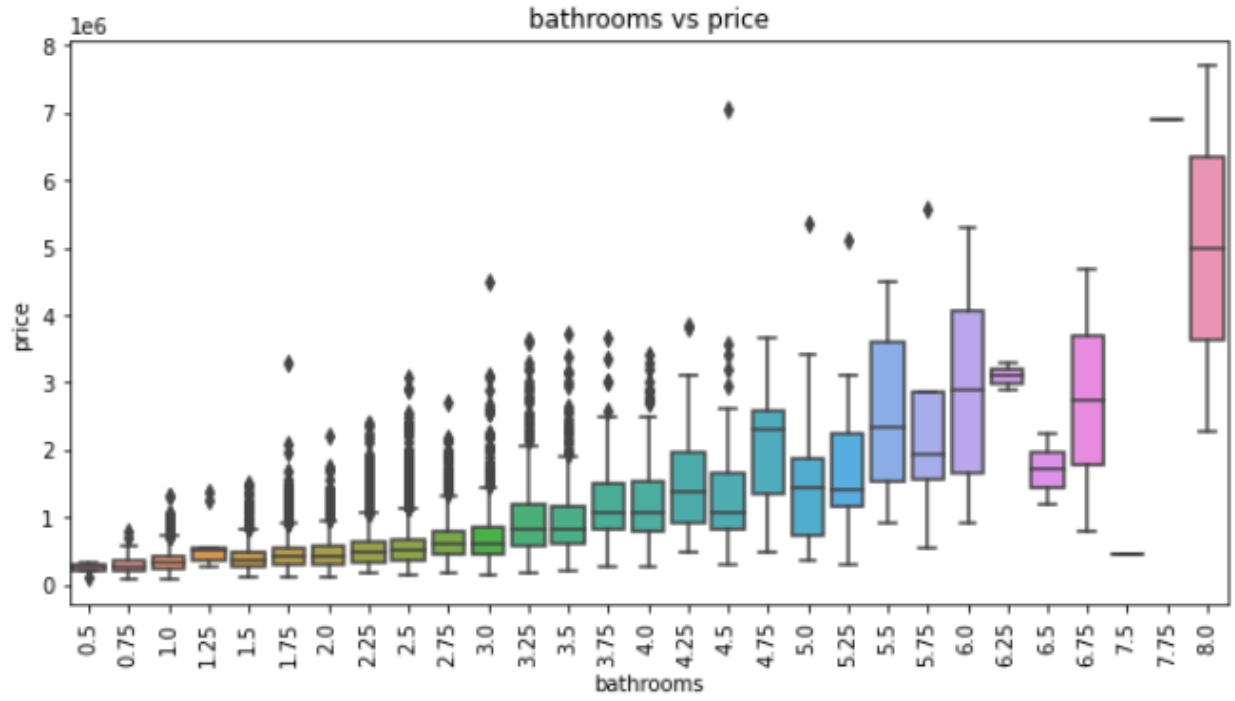
Price

Exploratory Data Analysis

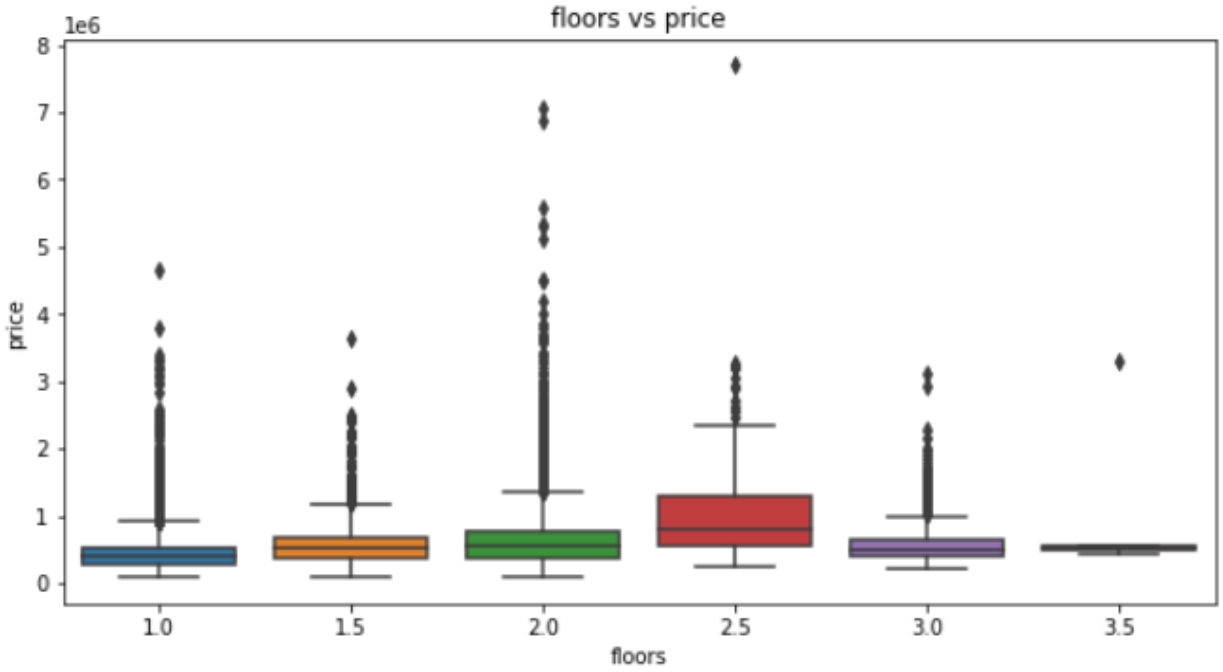
Bedroom vs Price



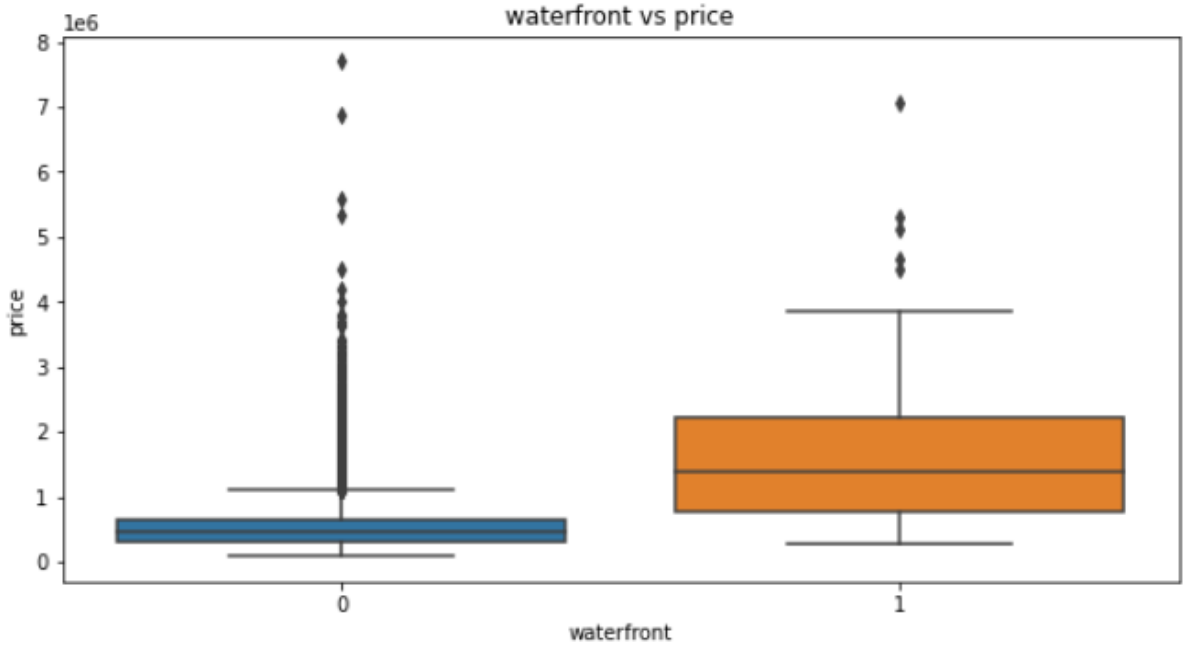
Bathrooms vs Price



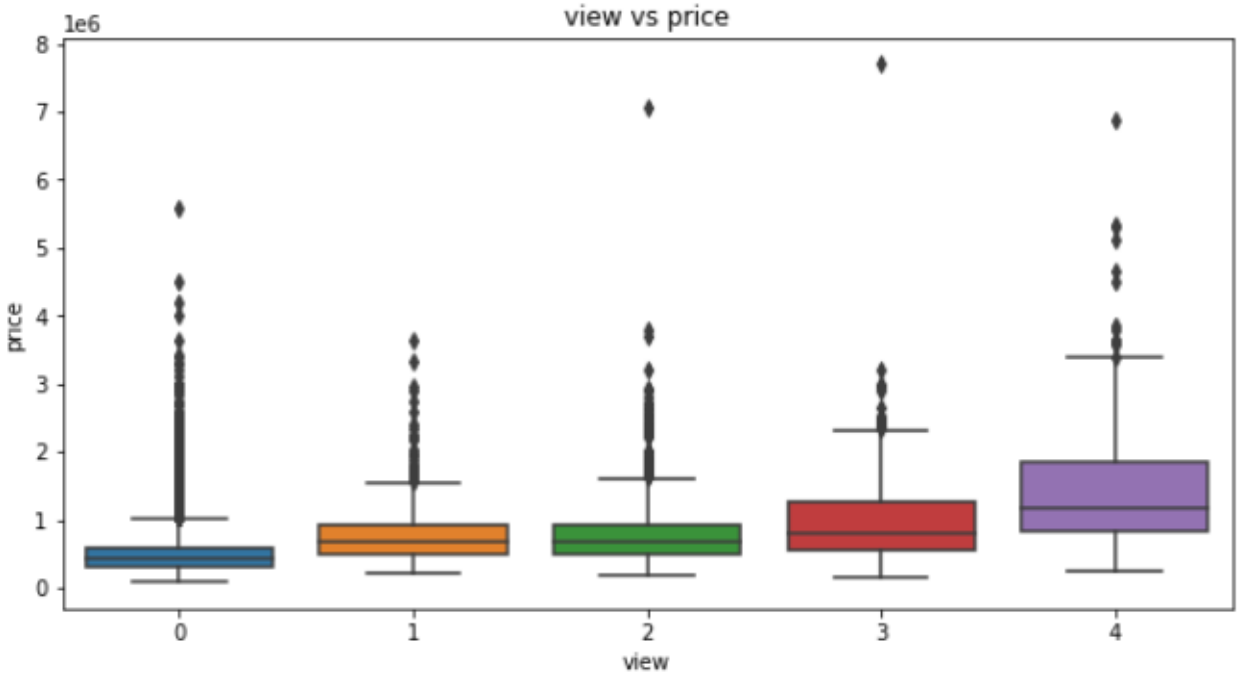
Floors vs Price



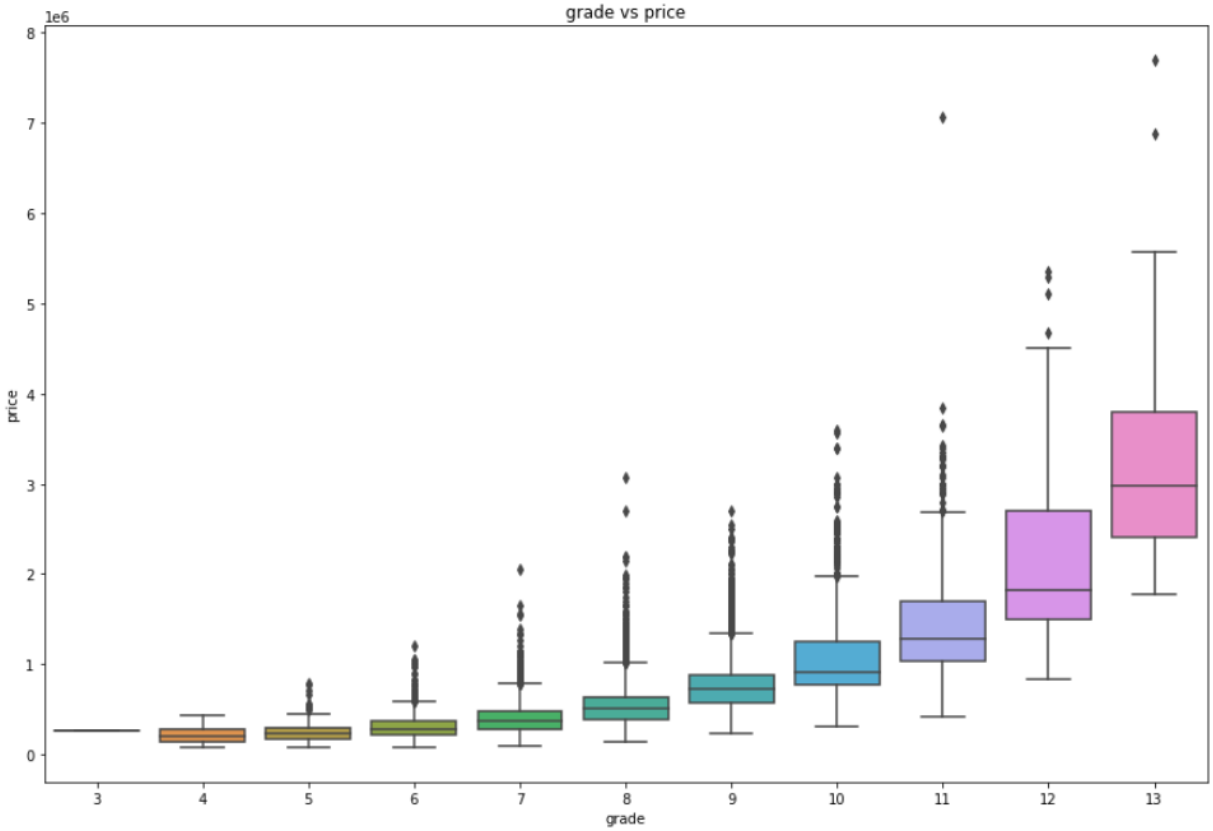
Waterfront vs Price



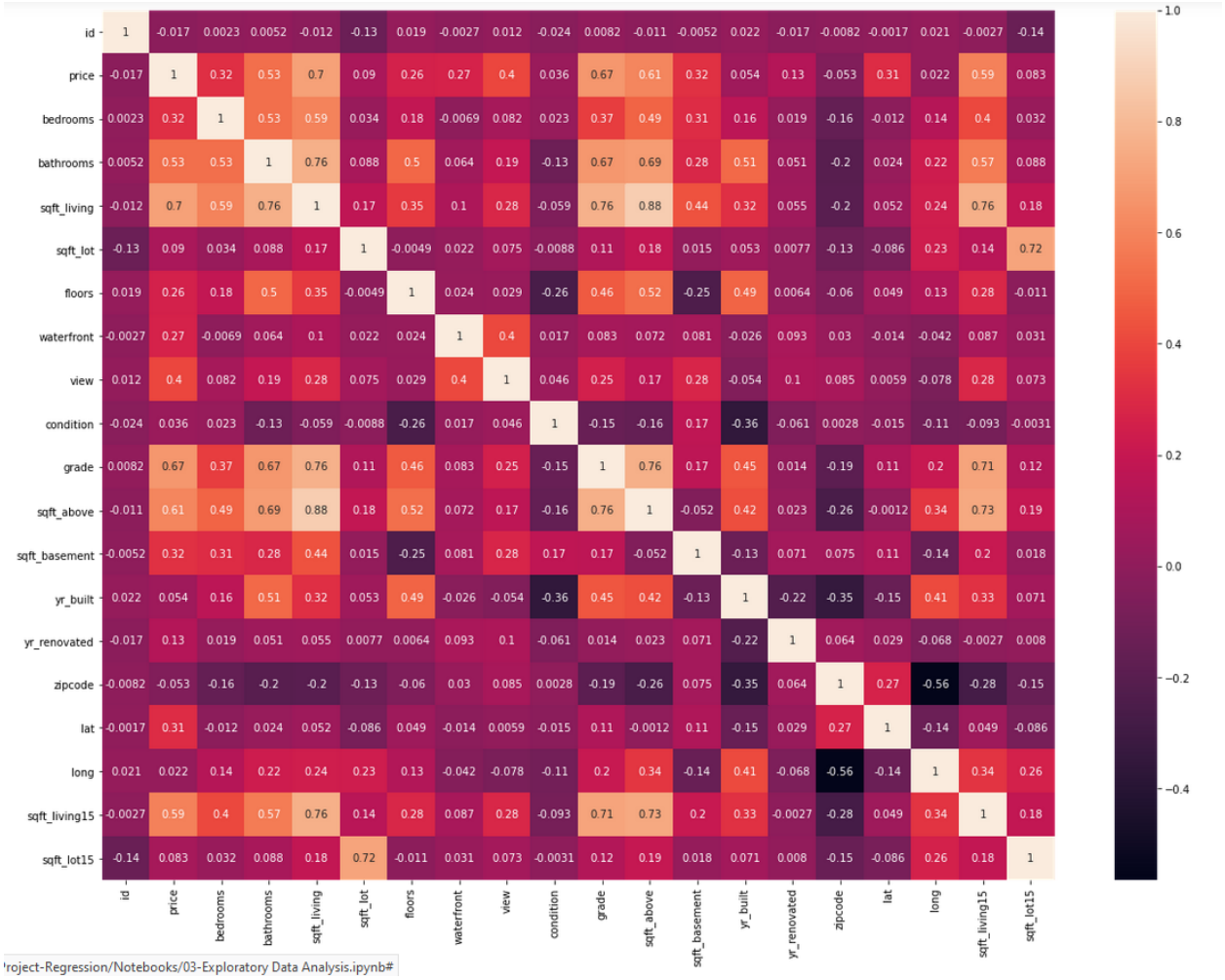
View vs Price



Grade vs Price

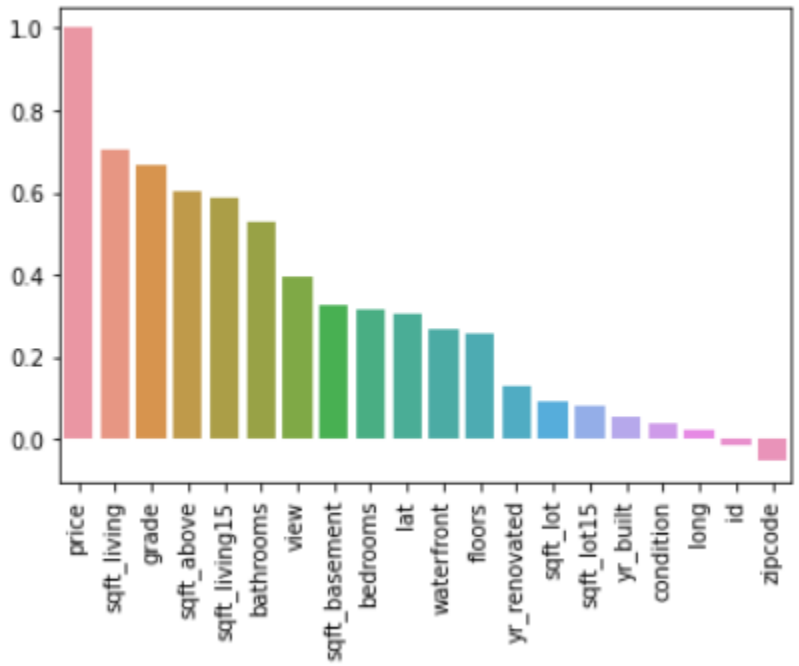


Correlation Matrix



Price is highly correlated with sqft_living, grade, bedrooms, sqft_above, sqft_living15. This makes sense as the size increase price goes up.

Plotting Features with maximum correlation with price



Zip Code wise summary

	0	1	2	3	4
zipcode	9.800100e+04	9.800200e+04	9.800300e+04	9.800400e+04	98005.000
Listing_per_Zip_Code	3.610000e+02	1.990000e+02	2.800000e+02	3.170000e+02	168.000
ZipCode_Total_sqft_living	6.872660e+05	3.239210e+05	5.400870e+05	9.221600e+05	446343.000
ZipCode_Total_sqft_lot	5.403088e+06	1.496009e+06	2.968867e+06	4.154038e+06	3348036.000
ZipCode_Total_sqft_above	6.222760e+05	3.029850e+05	4.657940e+05	7.670900e+05	362913.000
ZipCode_Total_sqft_basement	6.499000e+04	2.093600e+04	7.429300e+04	1.550700e+05	83430.000
ZipCode_Total_sqft_living15	6.606660e+05	2.943270e+05	5.253350e+05	8.478800e+05	431401.000
ZipCode_Total_sqft_lot15	4.050014e+06	1.509511e+06	2.728930e+06	4.059192e+06	3085786.000
Zipcode_Mean_housePrice	2.811949e+05	2.342840e+05	2.941113e+05	1.355927e+06	810164.875

Top 5 Zip Codes with Maximum Listing

zipcode	
98103	601
98038	589
98115	583
98052	574
98117	553

Top 5 Zip Codes with Maximum total sqft_living

zipcode	
98006	1438371
98052	1356738
98038	1265342
98074	1163189
98059	1124910

Top 5 Zip Codes with Maximum total sqft_lot

zipcode	
98022	17296502
98038	14971867
98053	14345424
98027	13537080
98014	12015955

Top 5 Zip Codes with Maximum total sqft_above

zipcode	
98038	1210592
98052	1206728
98006	1145591
98074	1070779
98059	1062240

Top 5 Zip Codes with Maximum total sqft_basement

zipcode	
98038	1210592
98052	1206728
98006	1145591
98074	1070779
98059	1062240

Top 5 Zip Codes with Maximum total sqft_living15

zipcode	
98006	1387235
98052	1339559
98038	1248059
98074	1137171
98059	1099213

Top 5 Zip Codes with Maximum total sqft_lot15

zipcode	
98022	11996802
98038	11826628
98053	11725647
98027	11715451
98092	10604072

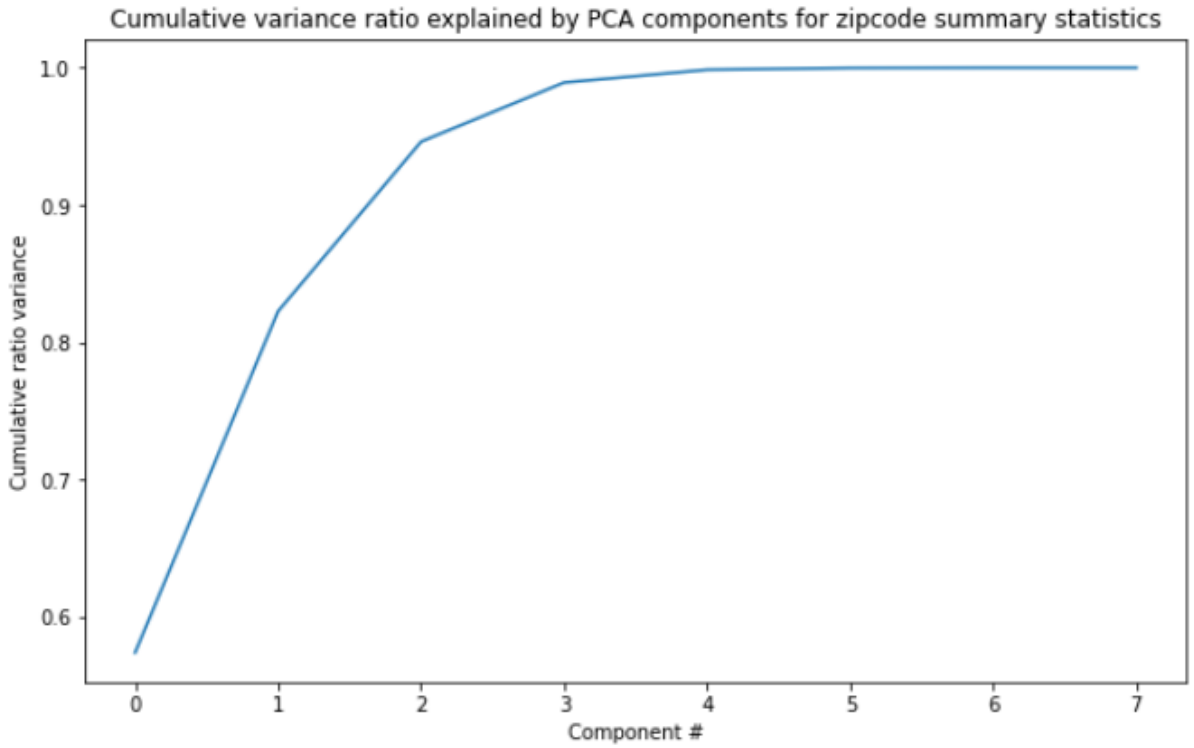
Top 15 Zip Code with highest mean price

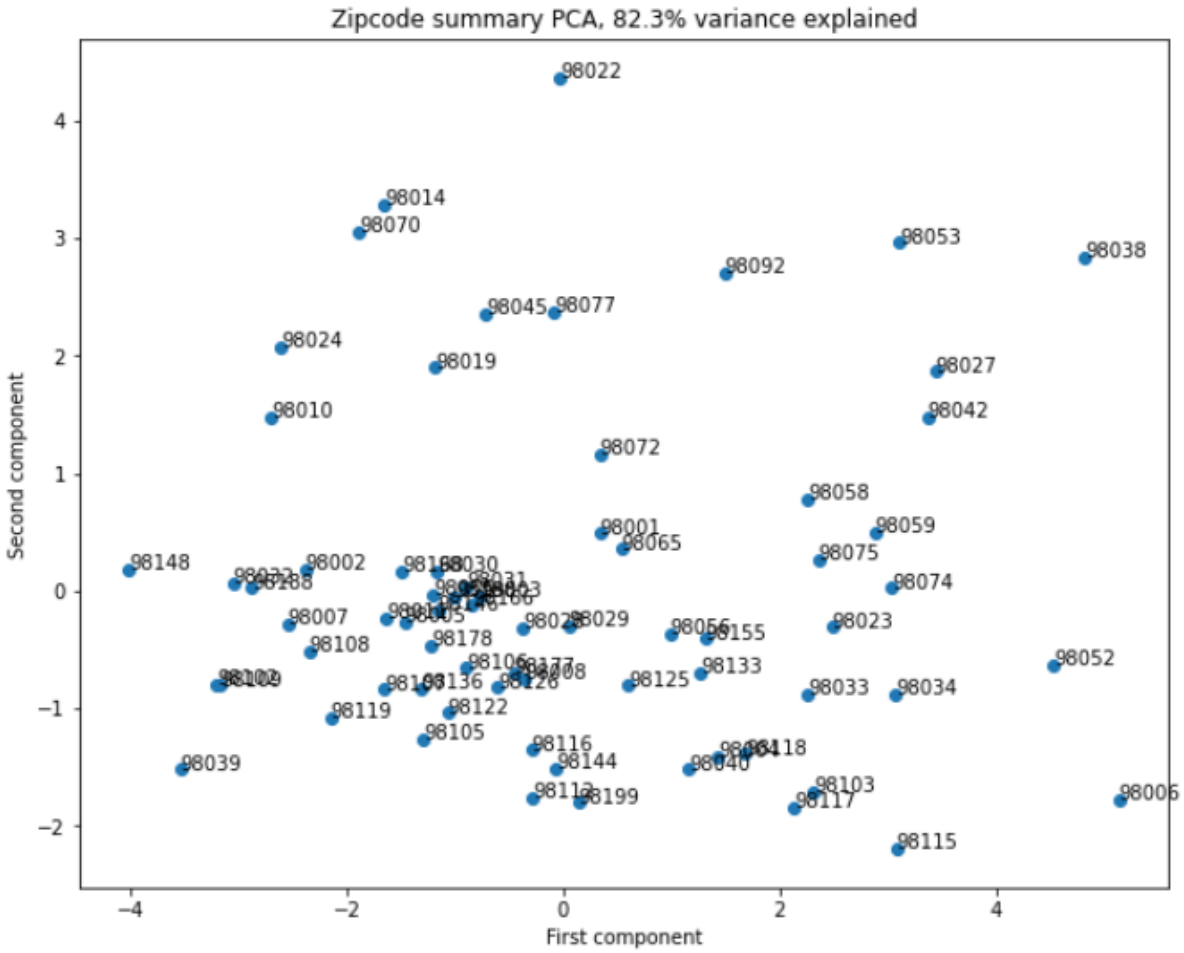
zipcode	
98039	2.160607e+06
98004	1.355927e+06
98040	1.194230e+06
98112	1.095499e+06
98102	8.993954e+05
98109	8.796236e+05
98105	8.628252e+05
98006	8.596848e+05
98119	8.494480e+05
98005	8.101649e+05
98033	8.037195e+05
98199	7.918208e+05
98075	7.905767e+05
98074	6.850617e+05
98077	6.827749e+05
98053	6.771126e+05
98177	6.761854e+05
98008	6.455074e+05
98052	6.452315e+05
98122	6.343602e+05

Please find the observations below:-

- 98038 looks like the most famous zipcode as it made top list for every feature but price.
- 98052 also made it to the list for almost all the feature except lot and price.
- None of the Zip codes that made it to the list of highest mean price has made it to any other features list.

Principal Component Analysis





Data Modeling and Results.

Different versions of Linear Regression and Random Forest were used as my modelling Techniques.
The Results for Each of them are was follows.

Predictor	Dummy Regressor Mean			
Metrics	R-Squared	Mean Absolute	Mean Squared Error	Root Mean Squared Error
Train	0.000	234147.780	135997343219.453	368778.176
Test	-1.533	233334.908	125711289643.160	354557.879

Predictor	Linear Regression			
Metrics	R-Squared	Mean Absolute	Mean Squared Error	Root Mean Squared Error
Train	0.912	56659.450	11960805934.041	109365.470
Test	0.928	54830.773	9042055024.967	95089.721

Predictor	Linear Regression with f-regression score function			
Metrics	R-Squared	Mean Absolute	Mean Squared Error	Root Mean Squared Error
Train	0.908	56150.223	12453092666.586	111593.426
Test	0.928	54359.240	9090860132.751	95346.002

	Cross Validation Result				
Test Scores	0.923	0.901	0.907	0.912	0.89407628
Mean	0.907				
STD	0.010				

Predictor	Linear Regression(Median, f-regression) with k = 21			
Metrics	R-Squared	Mean Absolute	Mean Squared Error	Root Mean Squared Error
Train	0.899	56150.223	12453092666.586	111593.426
Test	0.926	54359.240	9090860133	95346.002

	Cross Validation Result				
Test Scores	0.923	0.905	0.909	0.913	0.8990167
Mean	0.910				
STD	0.008				

Predictor	Random Forest			
Metrics	Model Score	Mean Absolute	Mean Squared Error	Root Mean Squared Error
Train				0.000
Test	0.990	6108.744		0.000

Predictor	Random Forest with StandScalar, f_regression			
Metrics	R-Squared	Mean Absolute	Mean Squared Error	Root Mean Squared Error
Train	0.998	2121.615	291432168.950	17071.385
Test	0.991	5355.824	1116839782	33419.153

	Cross Validation Result				
Test Scores	0.988	0.965	0.991	0.975	0.98611789
Mean	0.981				
STD	0.010				

Predictor	Random Forest with Standard Scalar, f_regression, random state =29			
Metrics	Model Score	Mean Absolute	Mean Squared Error	Root Mean Squared Error
Train	0.998	2105.481	297979988.109	17262.097
Test	0.990	5479.396	1130229322.302	33618.883

	Cross Validation Result				
Test Scores	0.988	0.969	0.992	0.978	0.98520669
Mean	0.982				
STD	0.008				