House Price Prediction Capstone Project: Milestone 1

Vanika Hans

Introduction

- **Problem Statement:**
  - House prices are an indication of economic growth
  - Important for executives to understand and predict the growth of the market to make future-proof business decisions
- **Need of the Present Study:**
  - House price analysis ***can be very complicated*** due to various features that contribute to the pricing of houses
  - In order to predict the housing prices, we need to understand the features that influence current prices
- **Business/Social Opportunities:**
  - Understanding and predicting the housing market will put the company one step of competitors with a foresight on the economy and business decisions made based on predictions

Data Dictionary

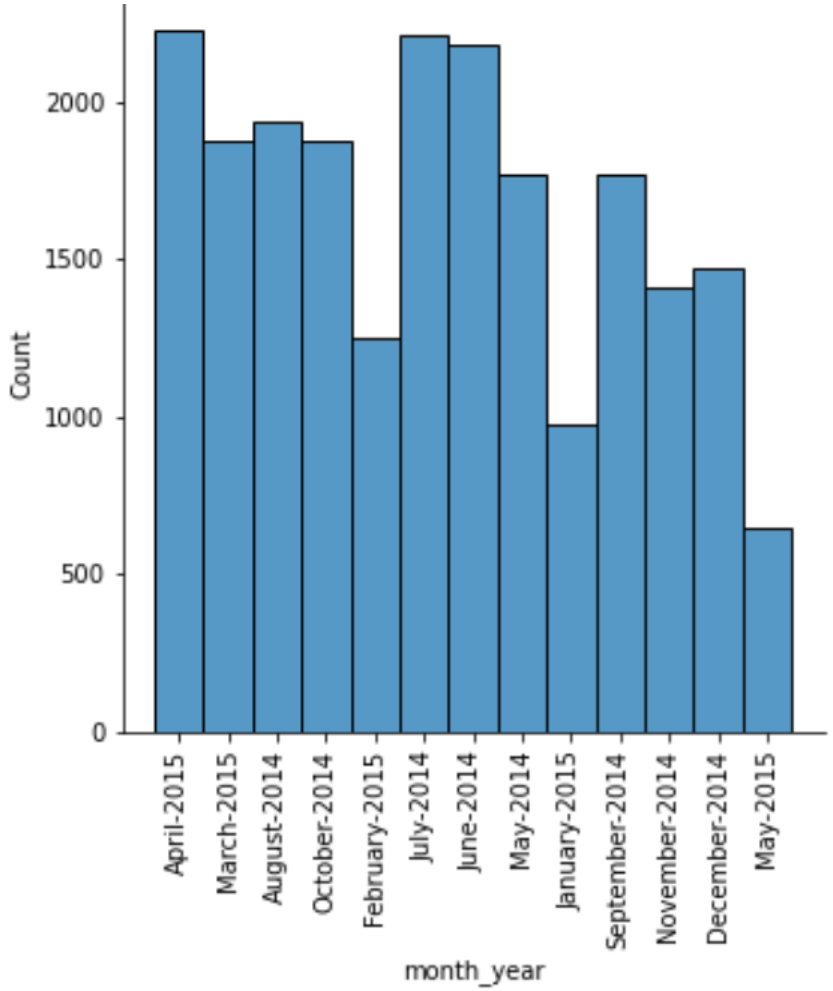
Features that all may or may not contribute to house prices:

Remember when we said it can be complicated?

Index	Attribute	Description	Count	Null	Type
0	cid	a notation for a house	21613	non-null	int64
1	dayhours	Date house was sold	21613	non-null	object
2	price	Price is prediction target (in $)	21613	non-null	int64
3	room_bed	Number of Bedrooms per house	21505	non-null	float64
4	room_bath	Number of bathrooms per bedrooms	21505	non-null	float64
5	living_measure	square footage of the home	21596	non-null	float64
6	lot_measure	square footage of the lot	21571	non-null	float64
7	ceil	Total floors (levels) in house	21571	non-null	object
8	coast	House which has a view to a waterfront (0 - No, 1 - Yes)	21612	non-null	object
9	sight	Has been viewed	21556	non-null	float64
10	condition	How good the condition is (Overall out of 5)	21556	non-null	object
11	quality	Grade given to the housing unit, based on grading system	21612	non-null	float64
12	ceil_measure	square footage of house apart from basement	21612	non-null	float64
13	basement	square footage of the basement	21612	non-null	float64
14	yr_built	Built Year	21612	non-null	object
15	yr_renovated	Year when house was renovated	21613	non-null	int64
16	zipcode	zip code	21613	non-null	int64
17	lat	Latitude coordinate	21613	non-null	float64
18	long	Longitude coordinate	21613	non-null	object
19	living_measure15	Living room area in 2015 (implies-- some renovations) This might or might not have affected the lot size area	21447	non-null	float64
20	lot_measure15	lotSize area in 2015 (implies-- some renovations)	21584	non-null	float64
21	furnished	Based on the quality of room (0 - No, 1 - Yes)	21584	non-null	float64
22	total_area	Measure of both living and lot	21584	non-null	object

Data Report (Univariate): Ceil, Dayhours

- There were multiple entries for the same house (cid) that indicate 176 houses getting listed multiple times.
- House data was collected from May 2014 to May 2015. The least amount of houses in the dataset are in February 2015, January / May 2015

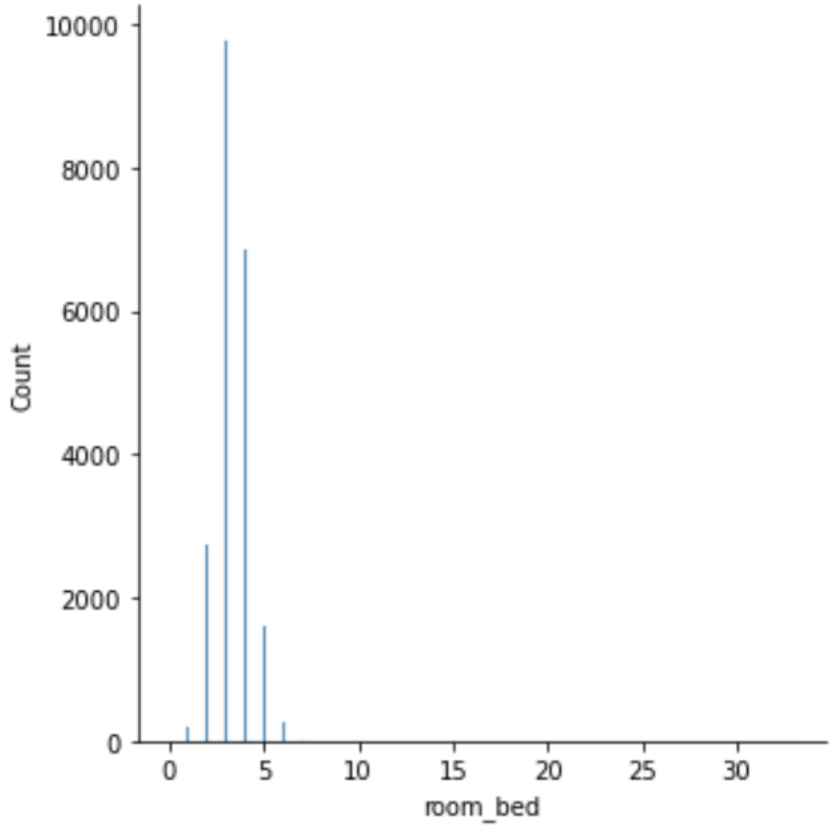


Data Report (Univariate): Room_bed

* Highest count at 3 bedrooms

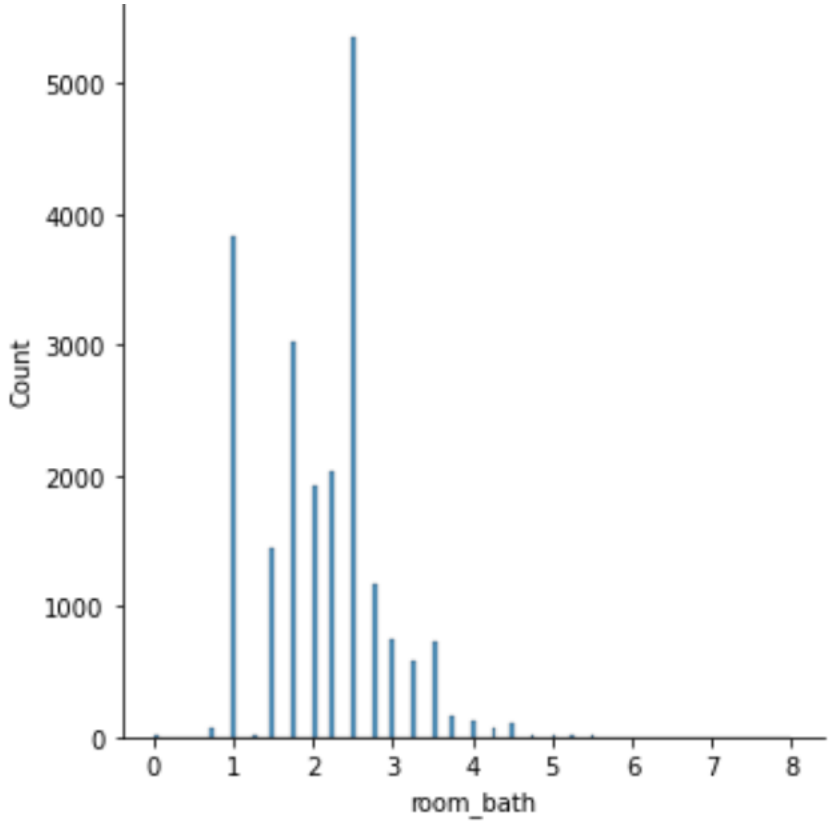
    * Lowest count at 9-33 bedrooooms. These also are definitely the outliers of the dataset

    * Skewed to the right.



Data Report (Univariate): Room_bath

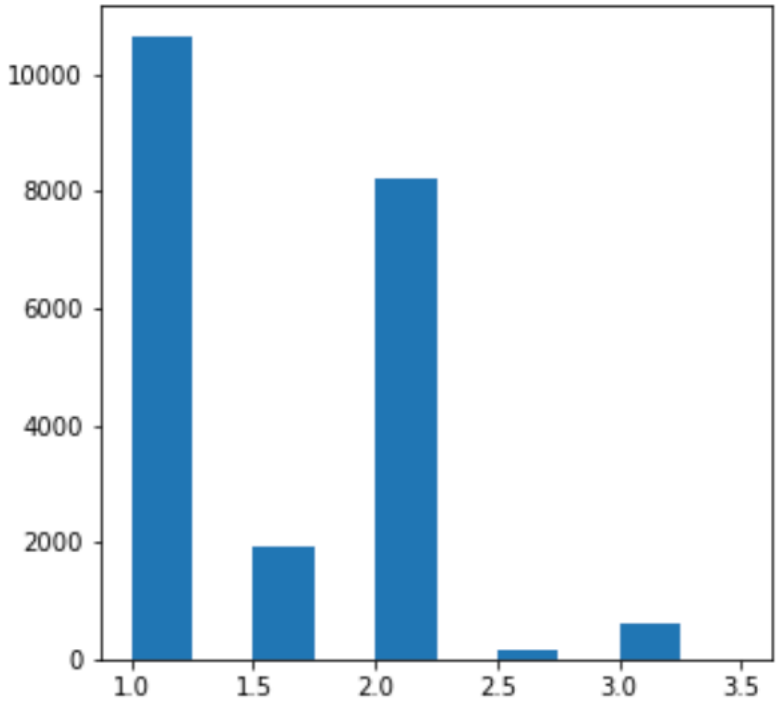
- Highest count at 2.5 bathrooms
- Outliers from 4.25-7.5 bathrooms. Not as common in the dataset.



Data Report (Univariate): Ceil

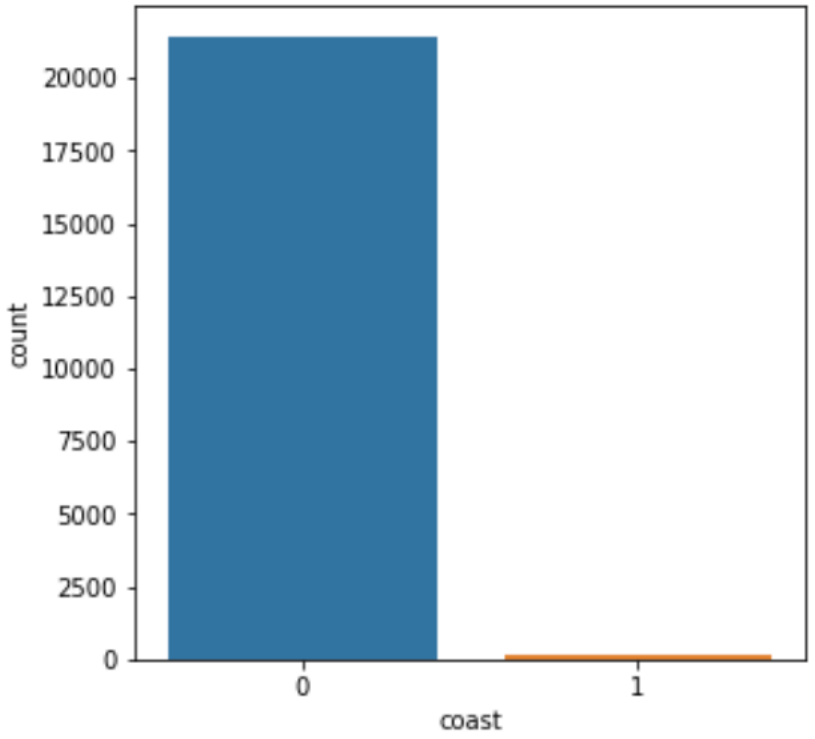
* Observations:

    * Visually confirmed that the total numbers of floors are overwhelmingly 1 and 2 is overwhelmingly high compared to the others.



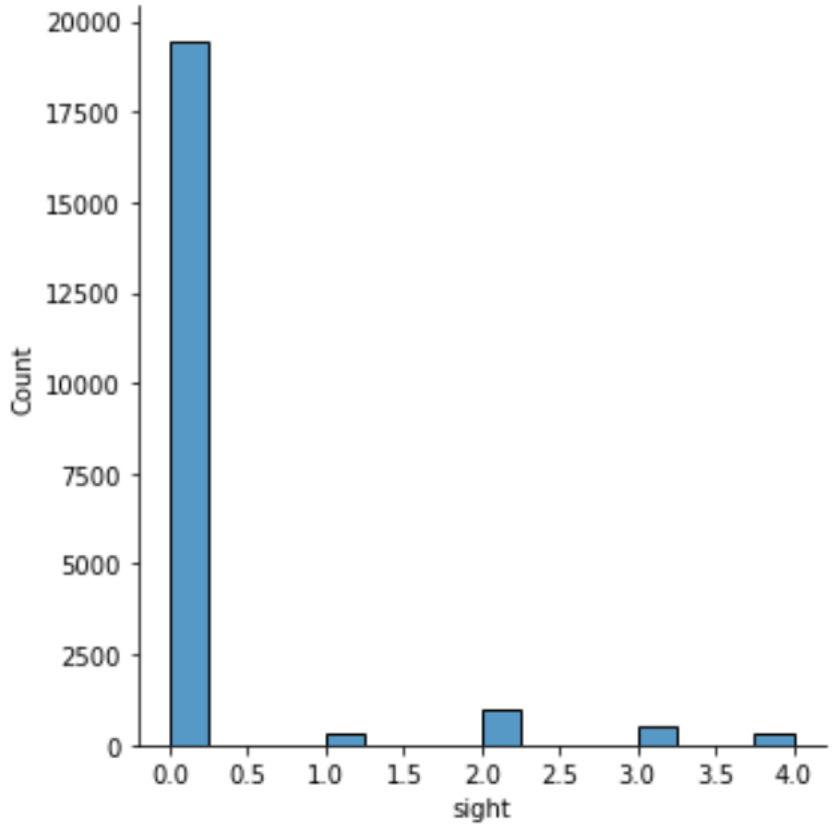
Data Report (Univariate): Coast

* As expected, most properties in the dataset do not have a coast, and a couple do.



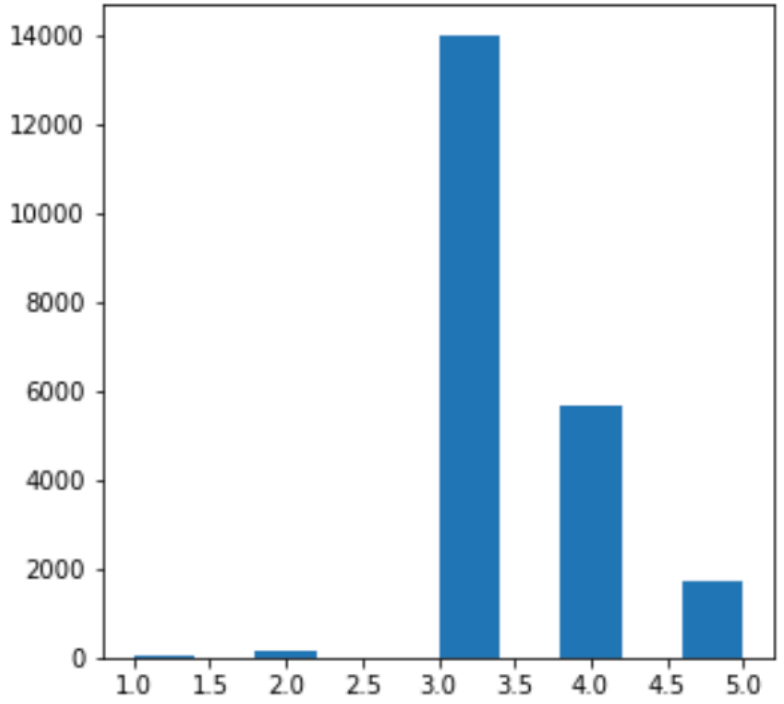
Data Report (Univariate): Coast

* Most of the properties in the dataset have not been viewed, and interestingly more properties have been viewed twice and three times more than once.



Data Report (Univariate): Condition

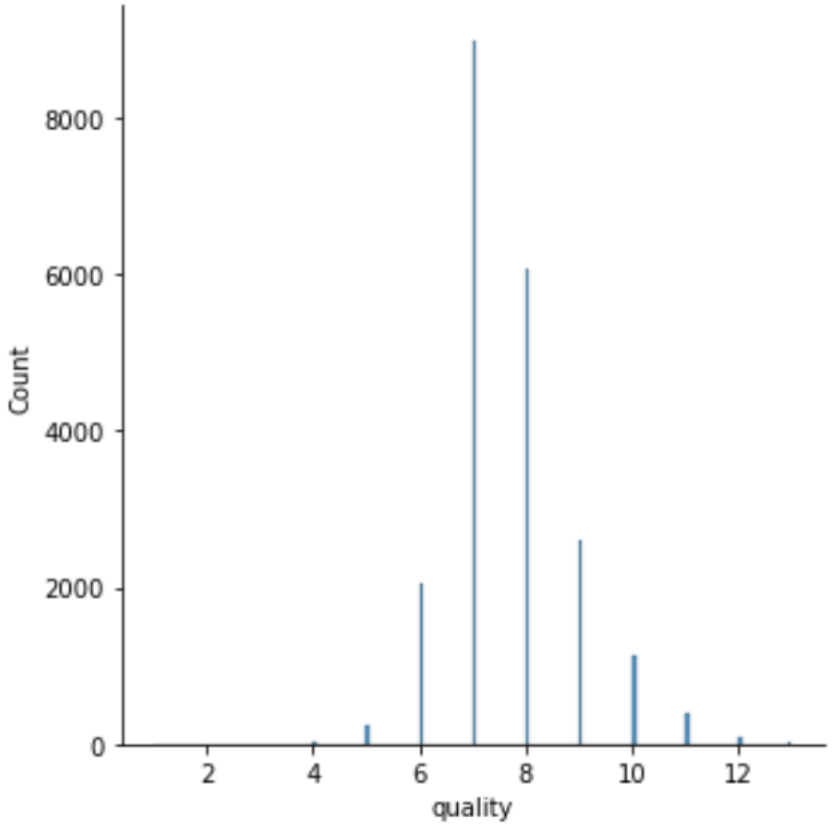
* Most of the houses are rated 3 stars, with some at 4 and 5.



Data Report (Univariate): Quality

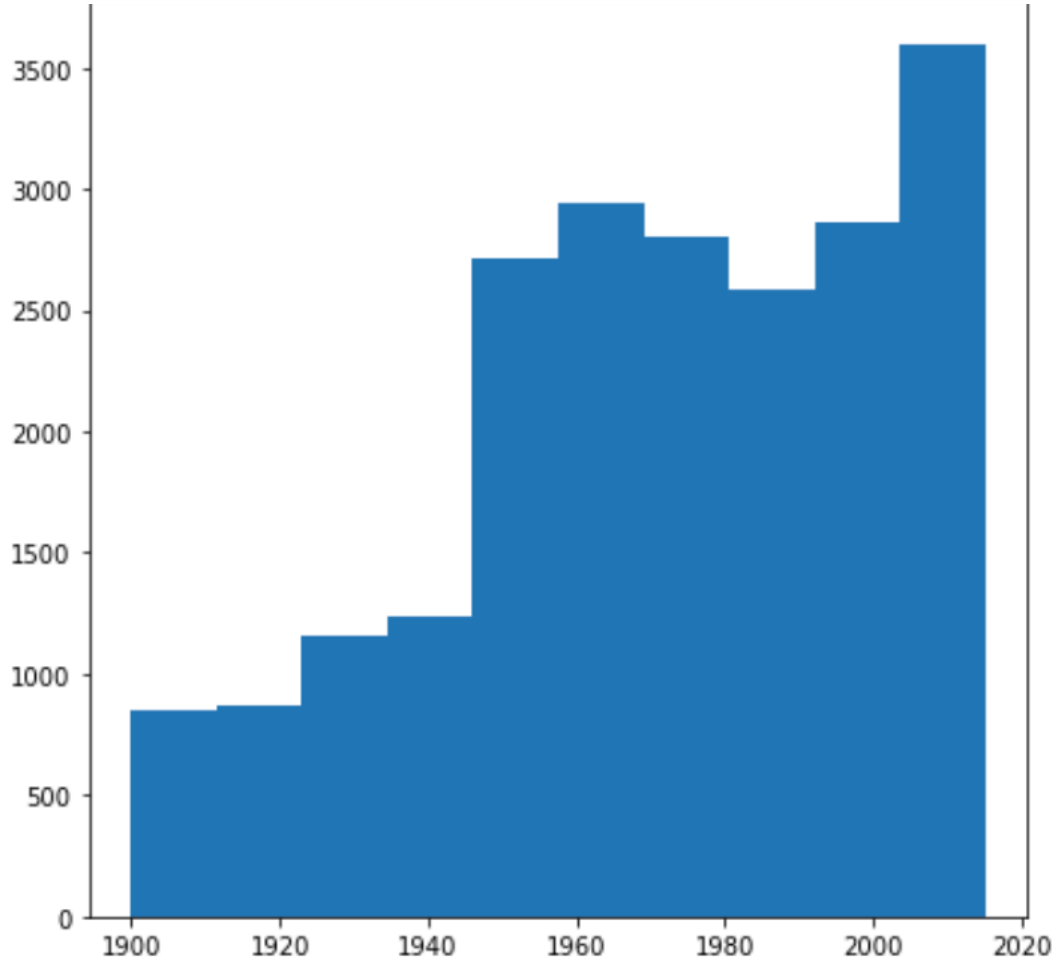
* The quality is based on a grading system that goes up to 13 - need to understand this better.

* But, the most of the houses have a quality of 7 and in the range of 6-10.



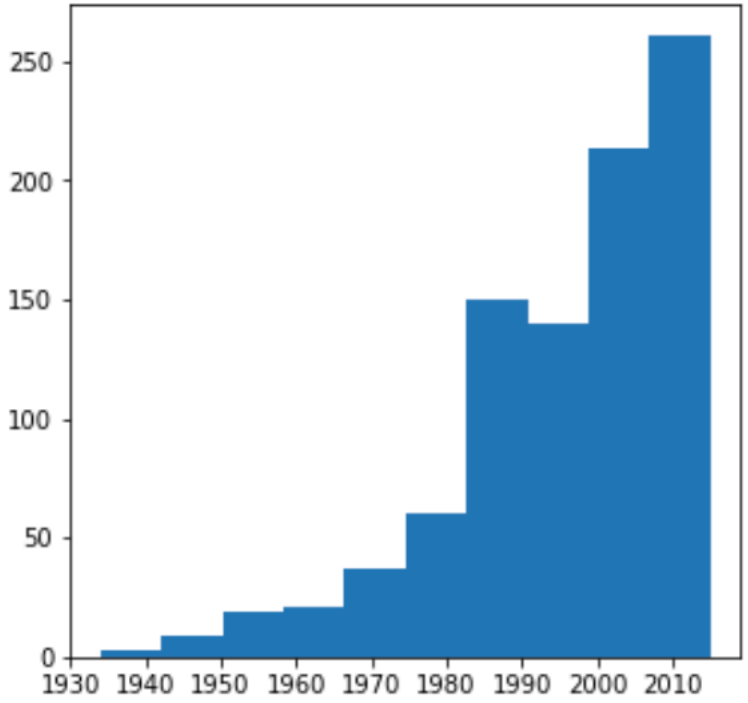
Data Report (Univariate): yr_built

- The yr_built is skewed to the left, indicating that most of the houses in the dataset are trending to be newer.
- The range of years shown is from 1900 to 2015 .



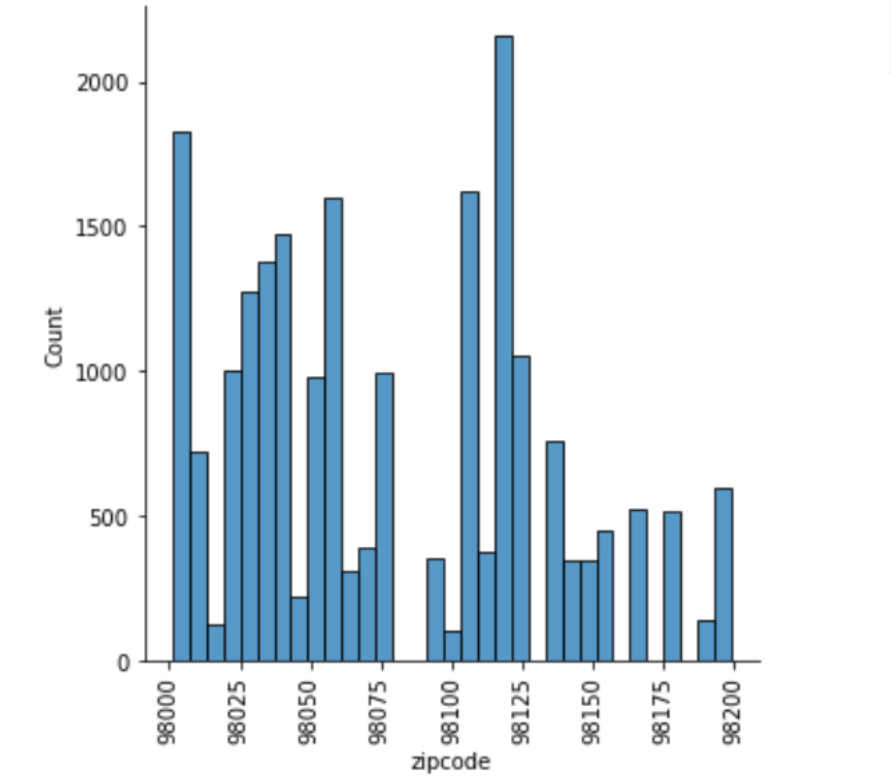
Data Report (Univariate): yr_renovated

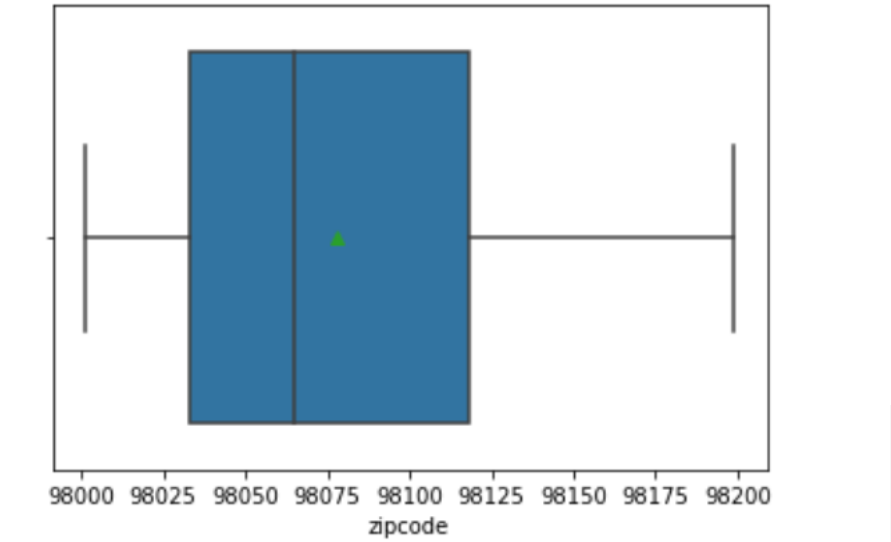
- * The year renovated is very much skewed to the left - which makes sense. Most of the renovations in the dataset are recent.



Data Report (Univariate): yr_renovated

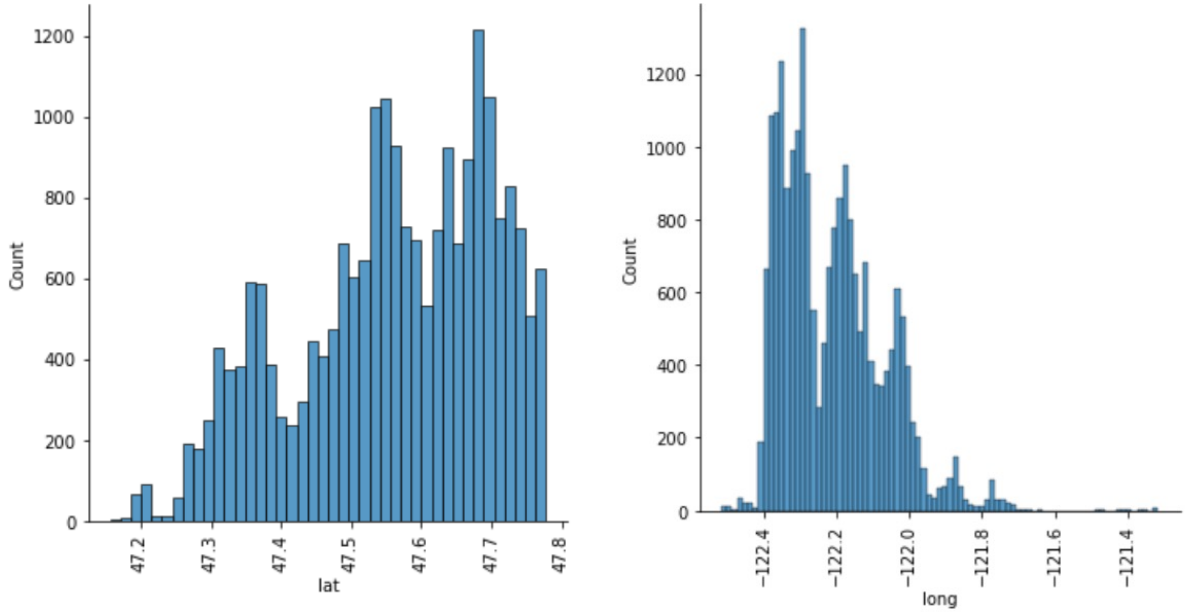
* The zipcodes in the dataset are spread out pretty well - no one area was dominated. So we have a good spread of location in the dataset





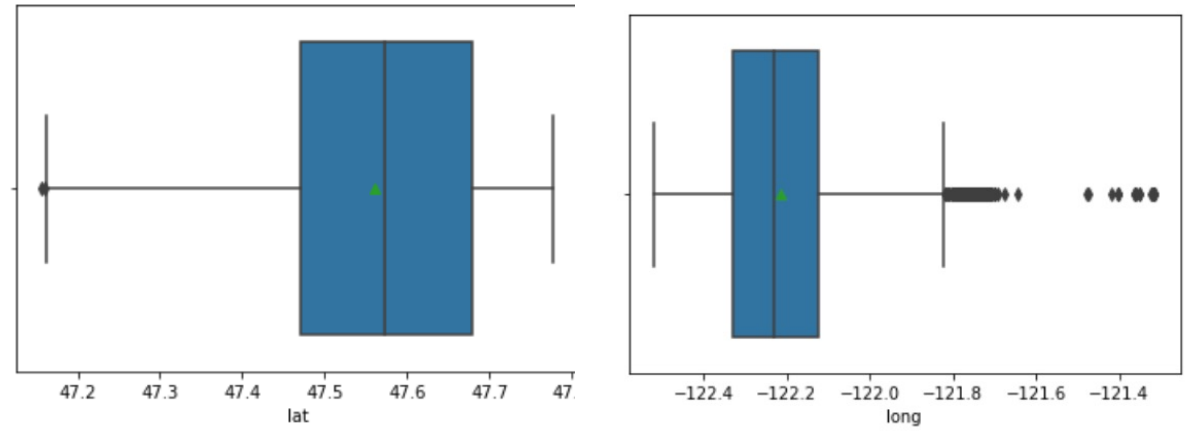
Data Report (Univariate): lat and long

* From latitude values above, we can see that 75% of the data lies within 47.47 to 47.67 and for the longitude values, 75% of the data lies between -122.51 to -122.125. Taking into account the minimum and maximum values as well, we can conclude that the dataset is mostly analyzing houses in the state of Washington.



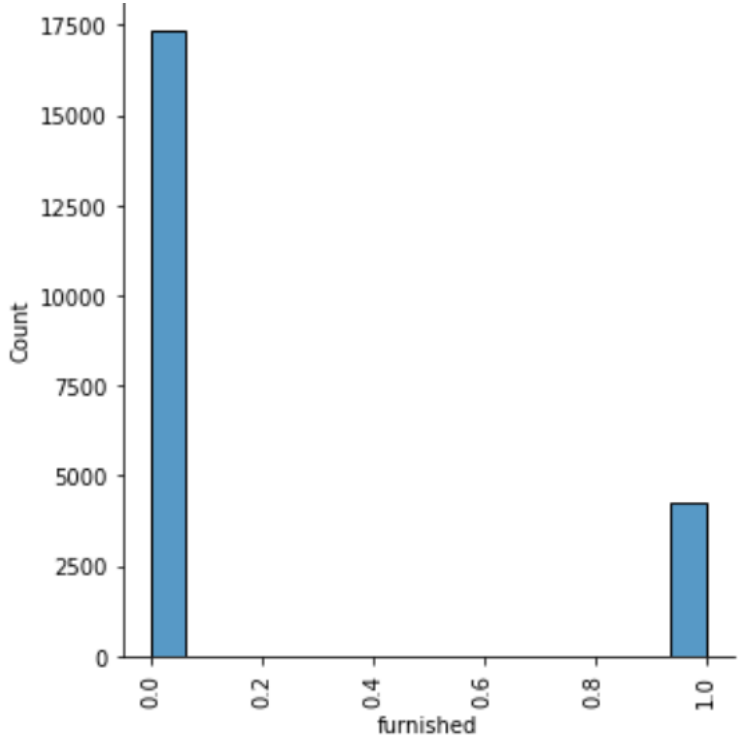
8]: <AxesSubplot:xlabel='lat'>

]: <AxesSubplot:xlabel='long'>



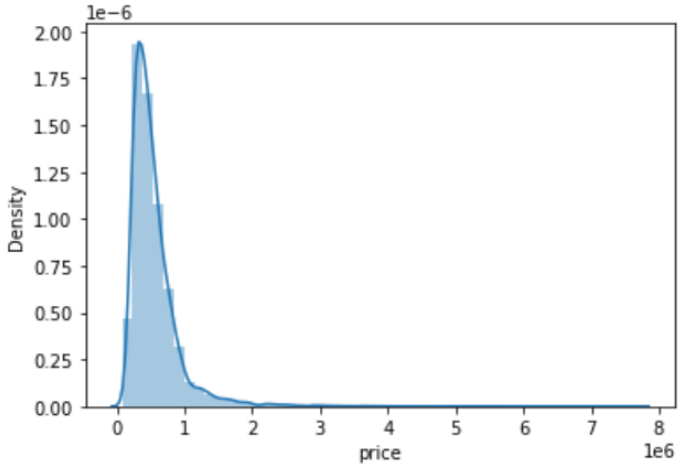
Data Report (Univariate): Furnished

* The majority of the houses are unfurnished with a small but significant amount furnished.

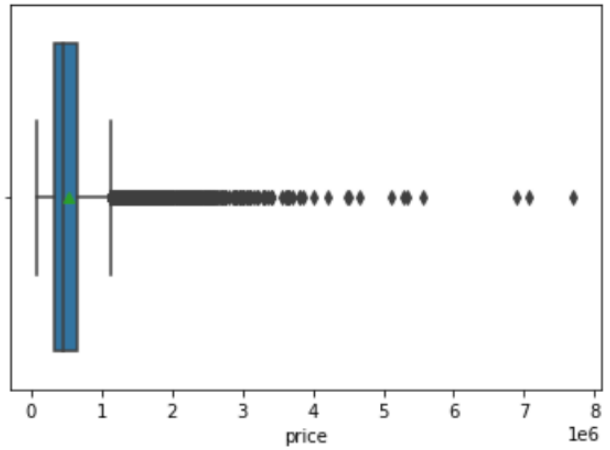


Data Report (Univariate): Price

* House prices are skewed far to the right, with a ton of outliers at a high price

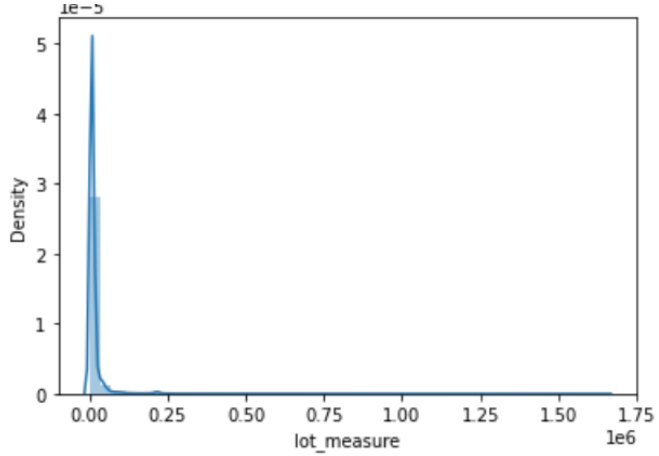


]: <AxesSubplot:xlabel='price'>

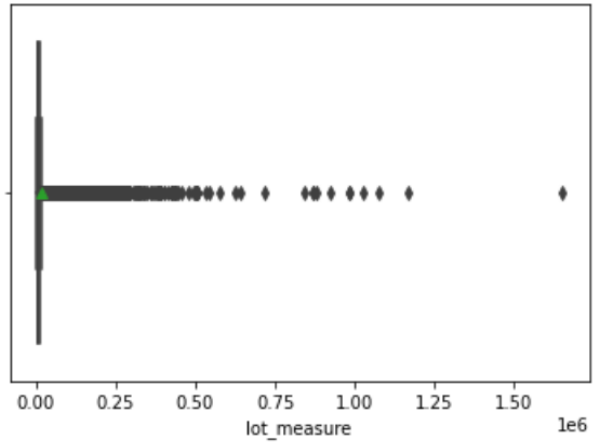


Data Report (Univariate): Price

* Measurements of the house apart from the basement are skewed far to the right, with a ton of outliers of a very large size



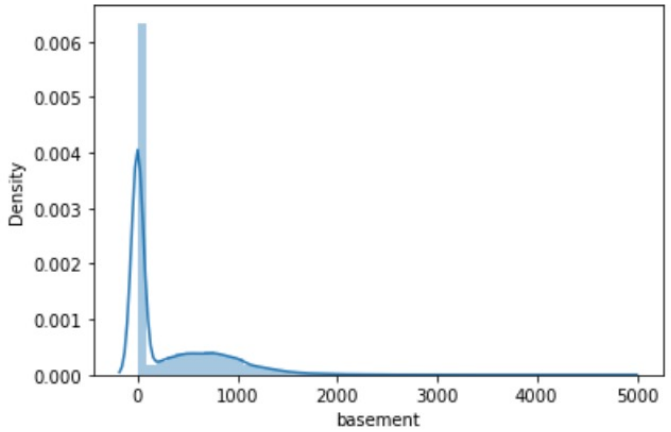
57]: <AxesSubplot:xlabel='lot_measure'>



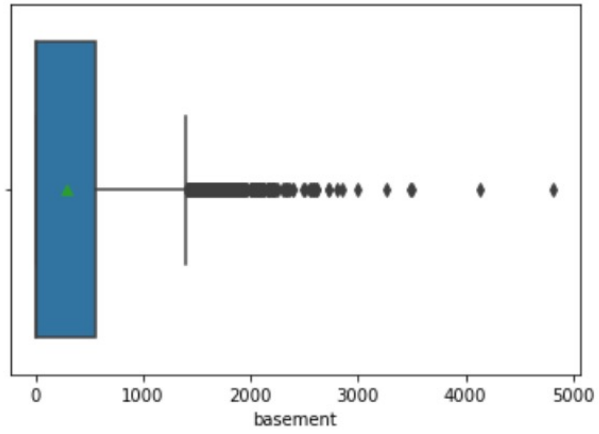
Data Report (Univariate): Basement

* Measurements of lot sizes are skewed far to the right, with a ton of outliers of a very large size.

* Measement sizes are trending to be very small



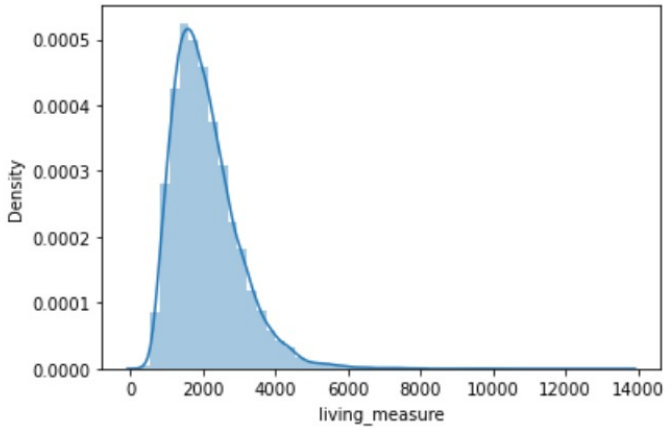
[58]: <AxesSubplot:xlabel='basement'>

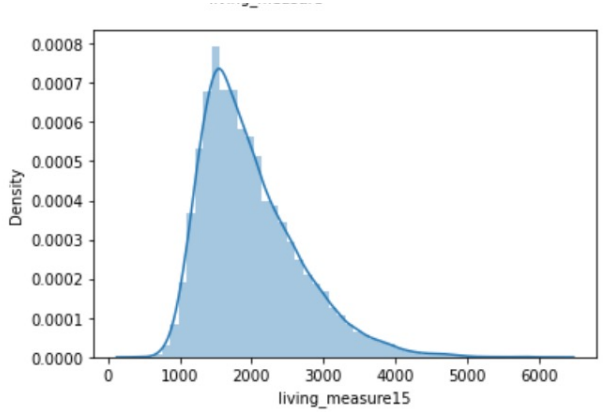


Data Report (Univariate): living_measure and living_measure15

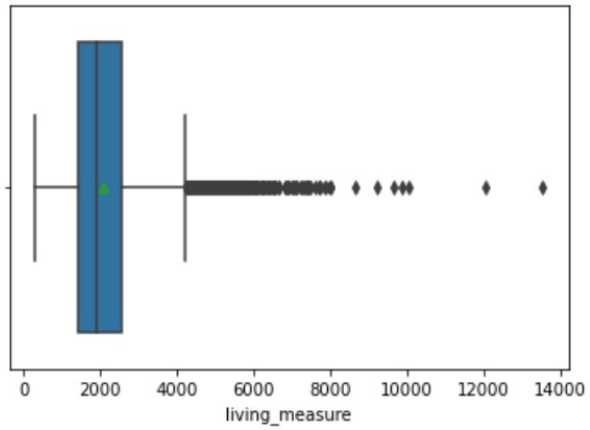
* Measurements of living spaces are skewed far to the right, with a ton of outliers of a very large size above ~4000

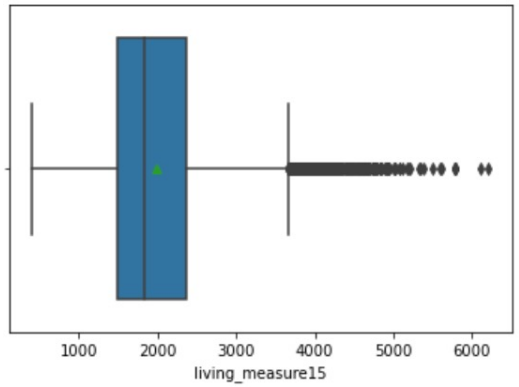
* Measurements of living room area in 2015 - slightly less skewed to the right with some outliers above ~3600.





59]: <AxesSubplot:xlabel='living_measure15'>

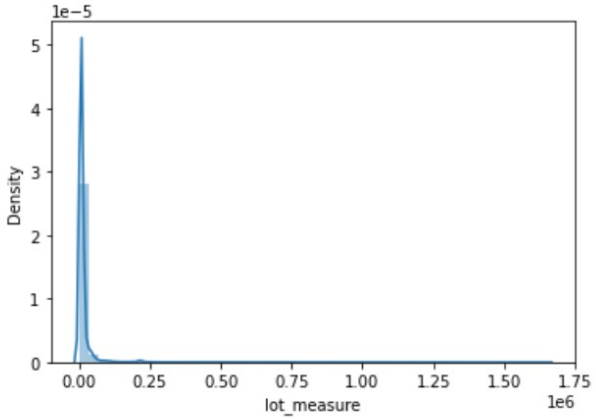


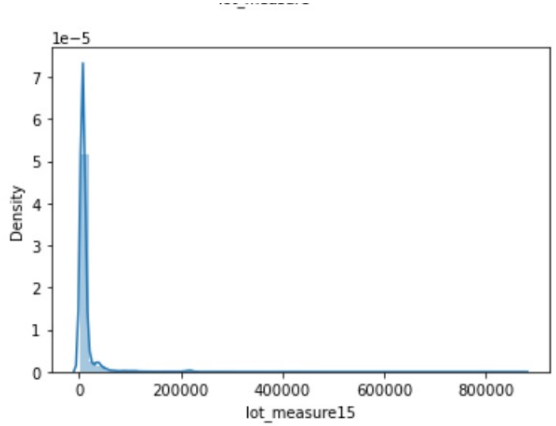


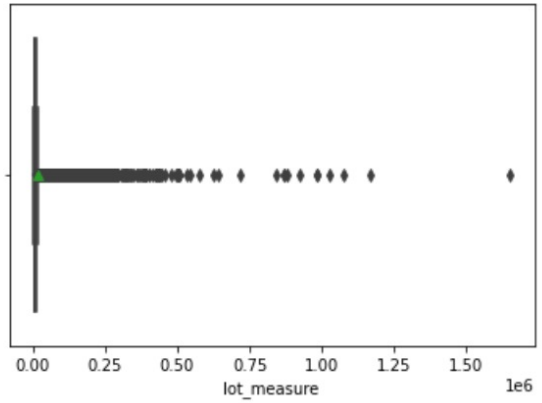
Data Report (Univariate): lot_measure and lot_measure15

* Measurements of lot size and lot size in 2015 - very skewed to the right with some outliers above ~3600.

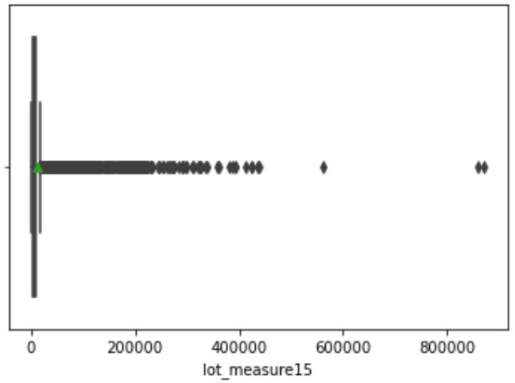
* both look very similar





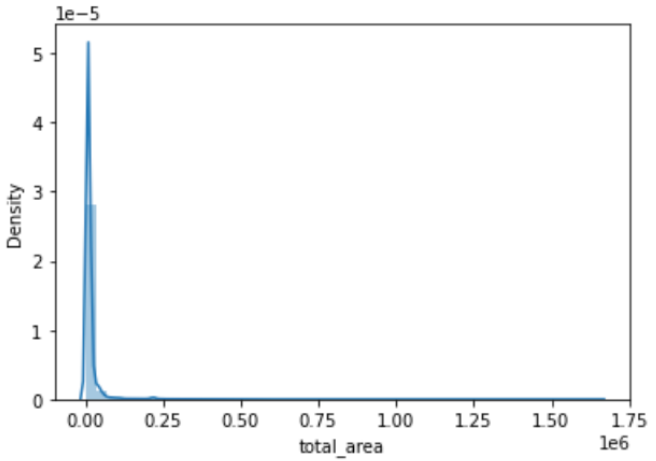


[60]: <AxesSubplot:xlabel='lot_measure15'>

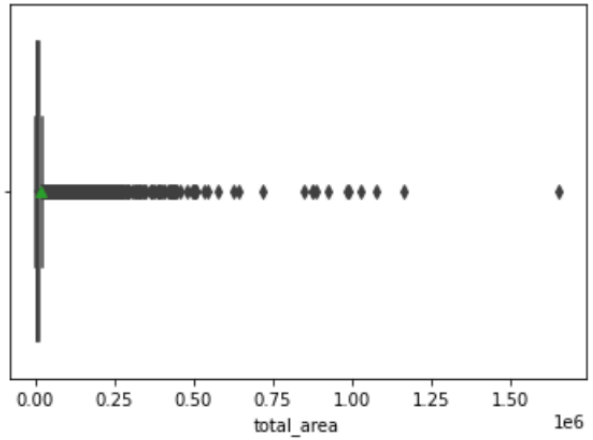


Data Report (Univariate): total_area

* Skewed to the right with a large amount of outliers

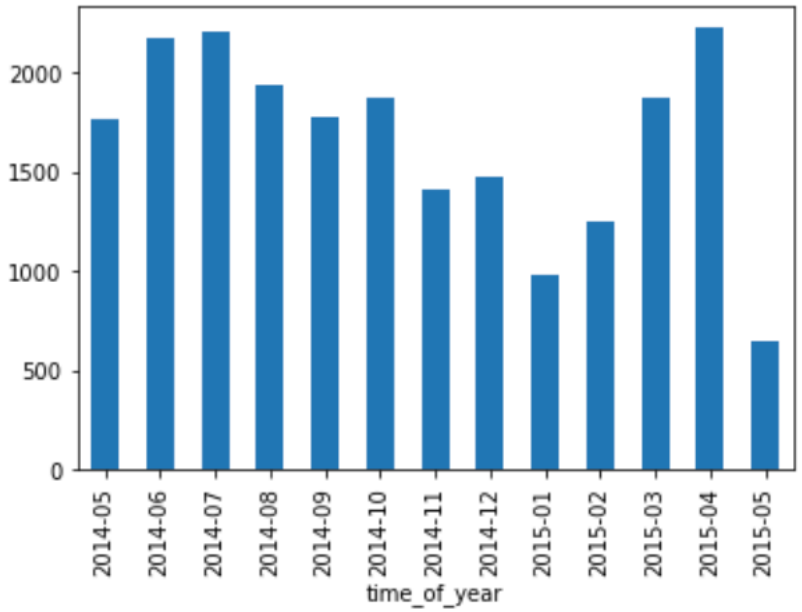


[62]: <AxesSubplot:xlabel='total_area'>



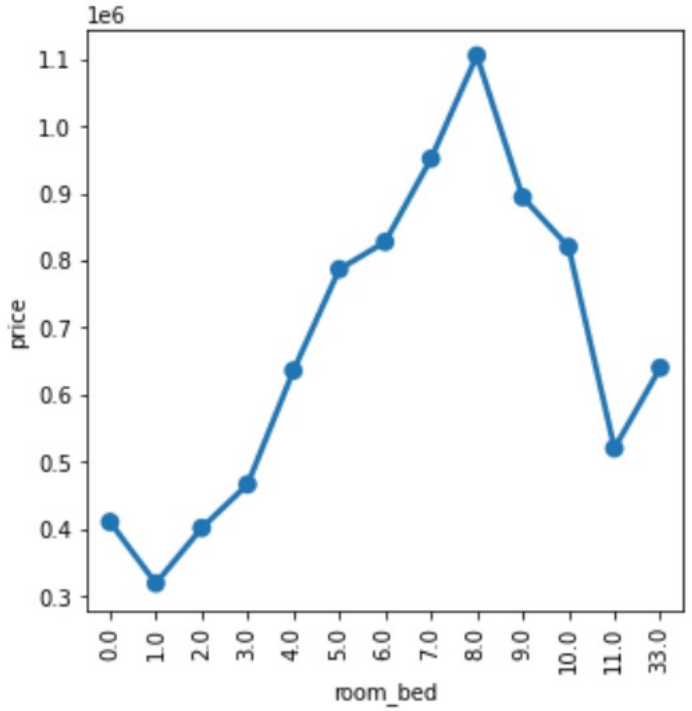
Bivariate Analysis: Price vs Month_Year

- * Months where the house was sold for the highest price were June and July 2014 as well as April 2015
- * Months where the house was sold for the lowest prices were January 2015 and May 2015

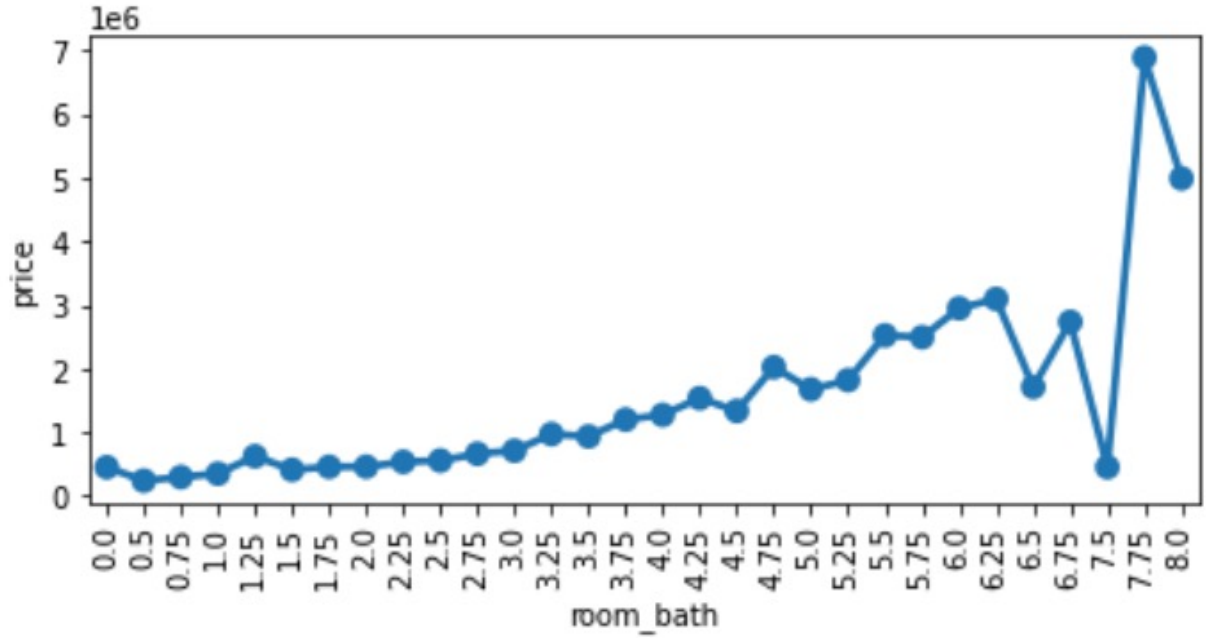


Bivariate Analysis: Price vs Room_bed

* The price becomes higher in value up untl the number of rooms is at 8, for which then the price decreases for rooms with 9-13 rooms

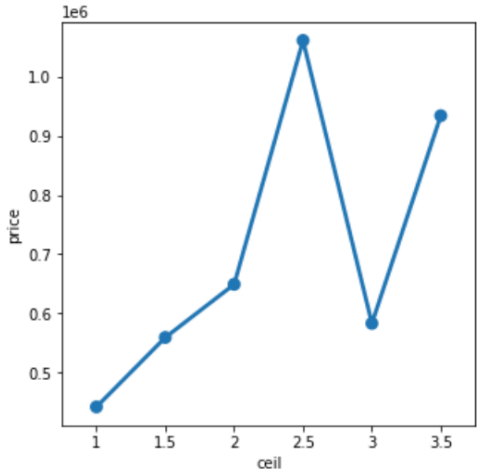


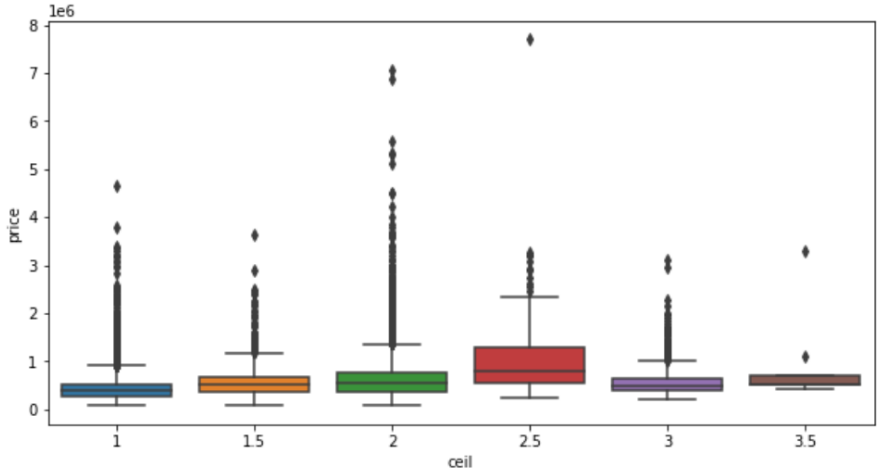
Bivariate Analysis: Price vs Room_bed



- For the most part, the price trends upwards as the number of bathrooms increases

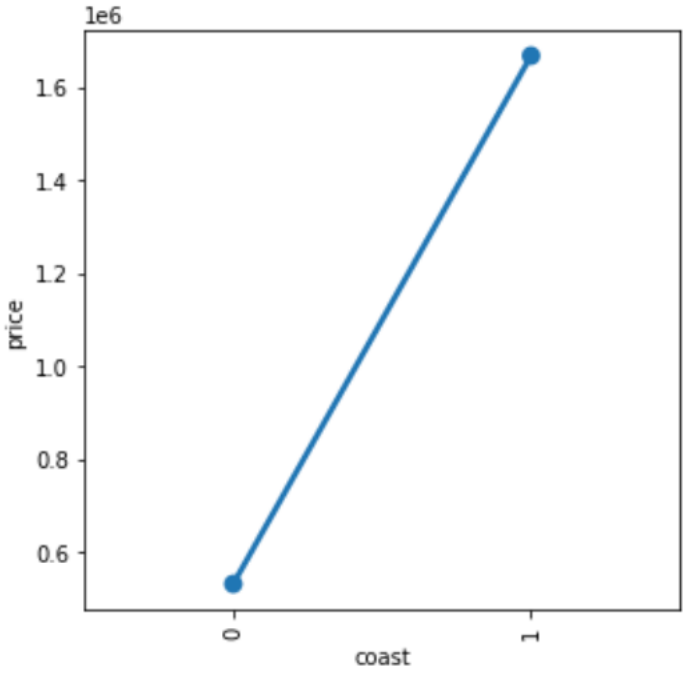
Bivariate Analysis: Price vs Ceil





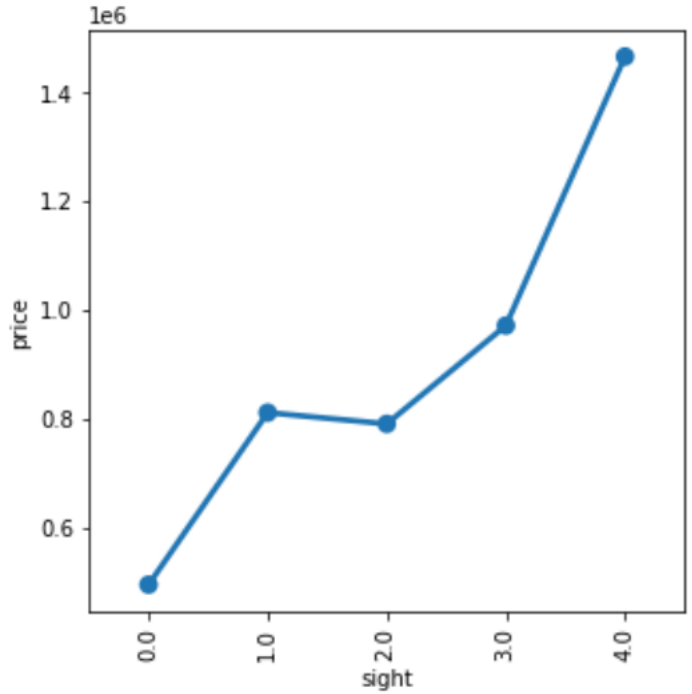
- The price increases as the number of floors increases, but briefly drops at 3 floors

Bivariate Analysis: Price vs Coast



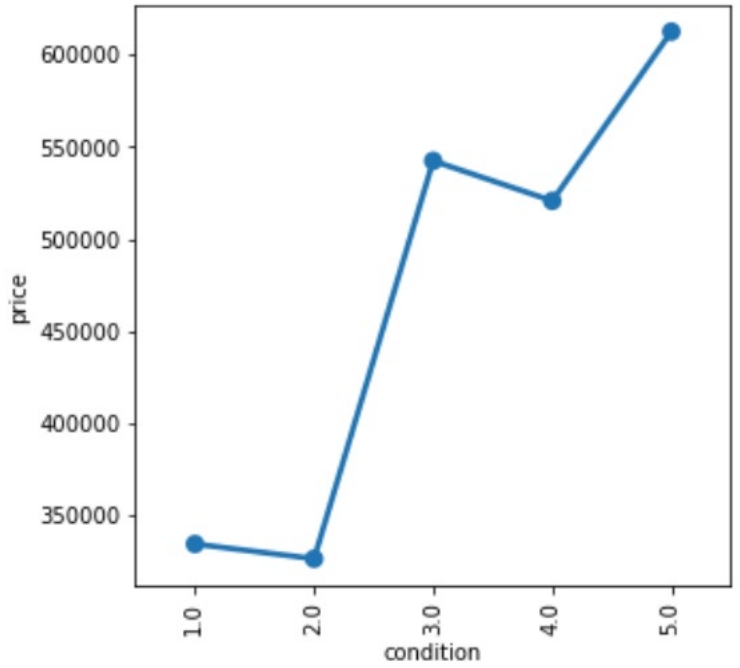
- Undubiously, having a coast view increases the price

Bivariate Analysis: Price vs Sight



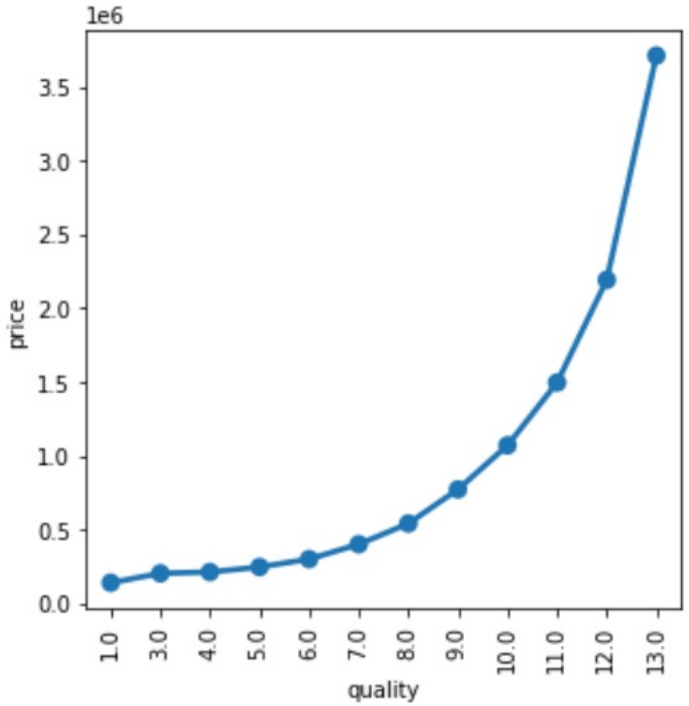
- having the house viewed increases the price

Bivariate Analysis: Price vs Condition



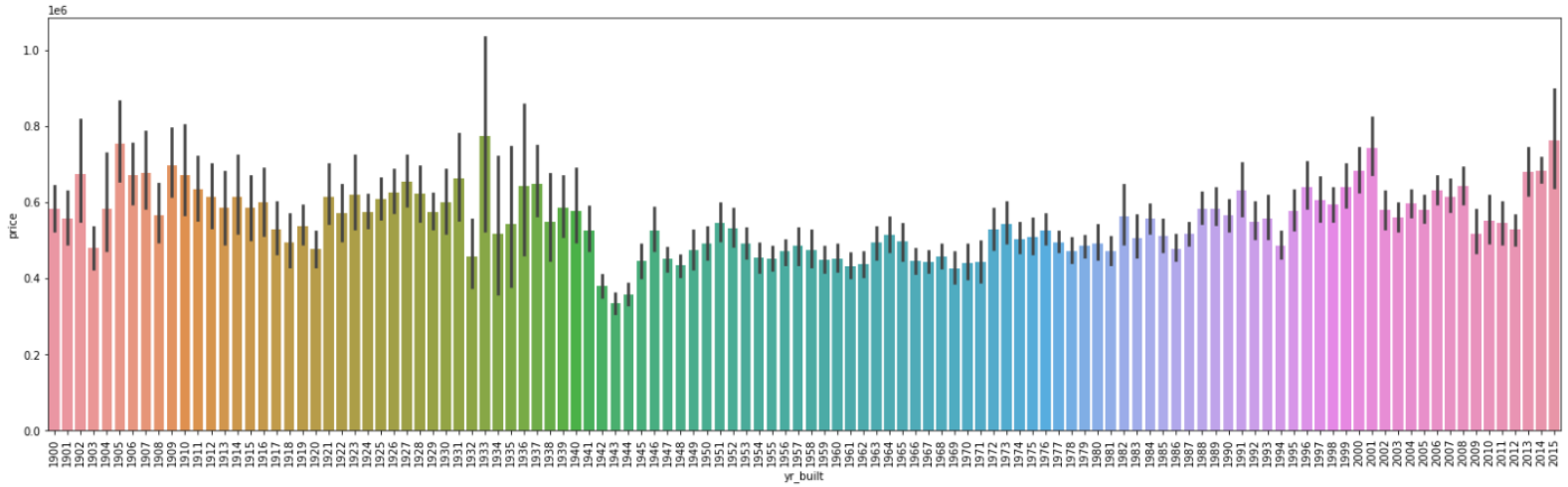
- the higher the condtiion, the higher the price

Bivariate Analysis: Price vs Quality



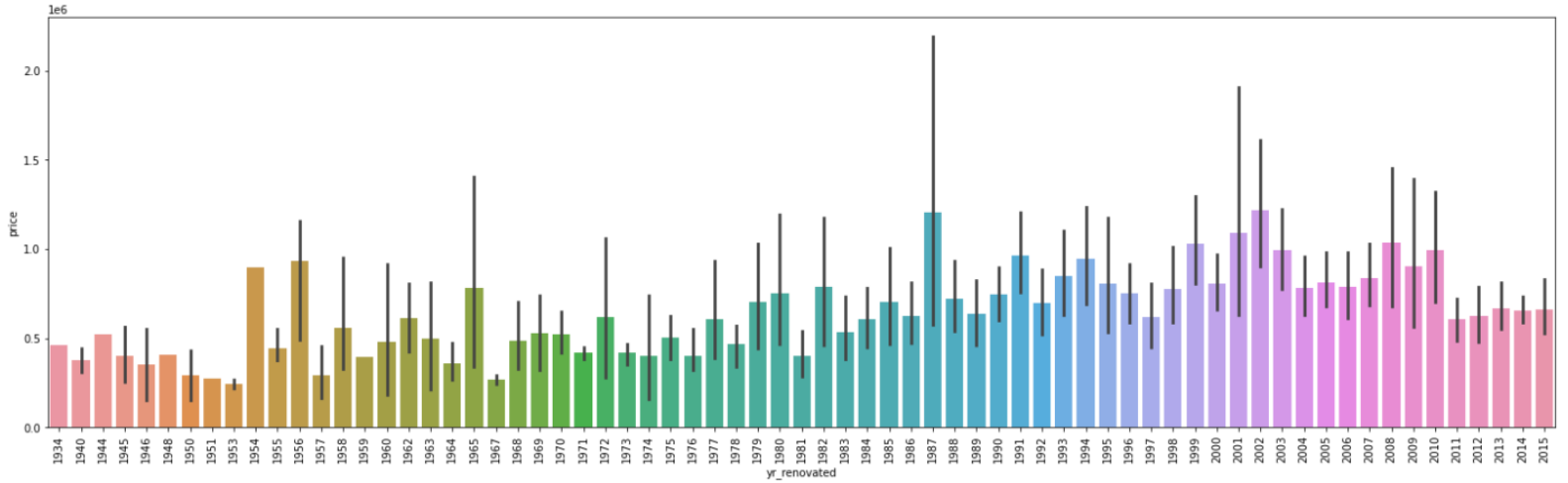
- the higher the quality, the higher the price

Bivariate Analysis: Price vs Year Built



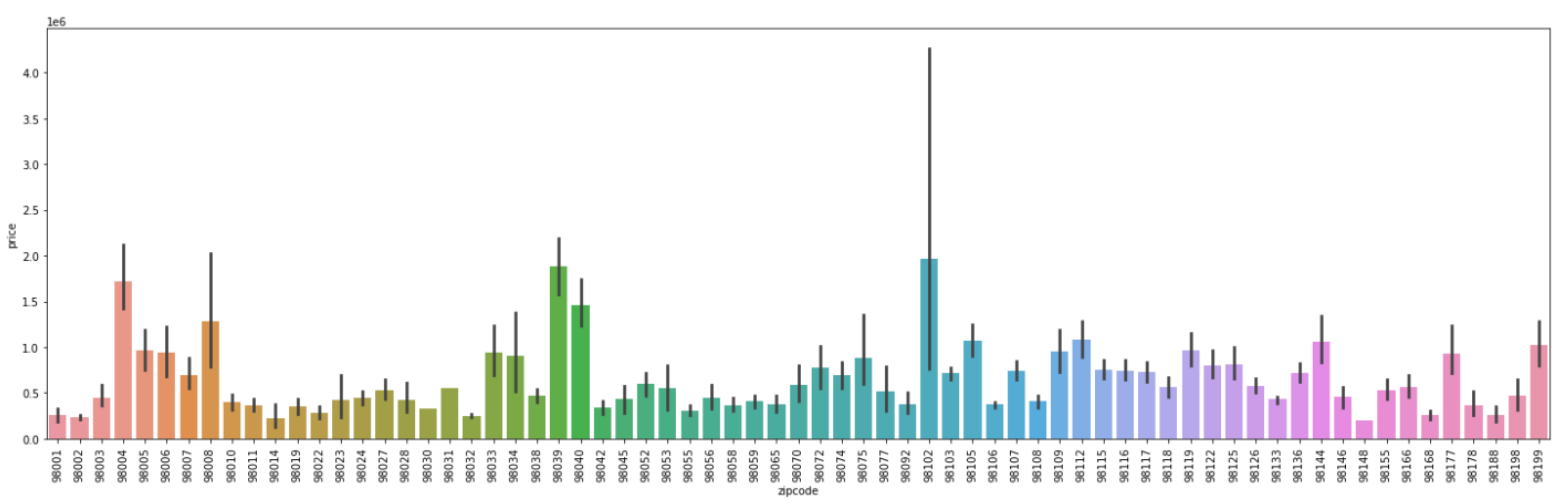
- no specific pattern

Bivariate Analysis: Price vs Yr_renovated



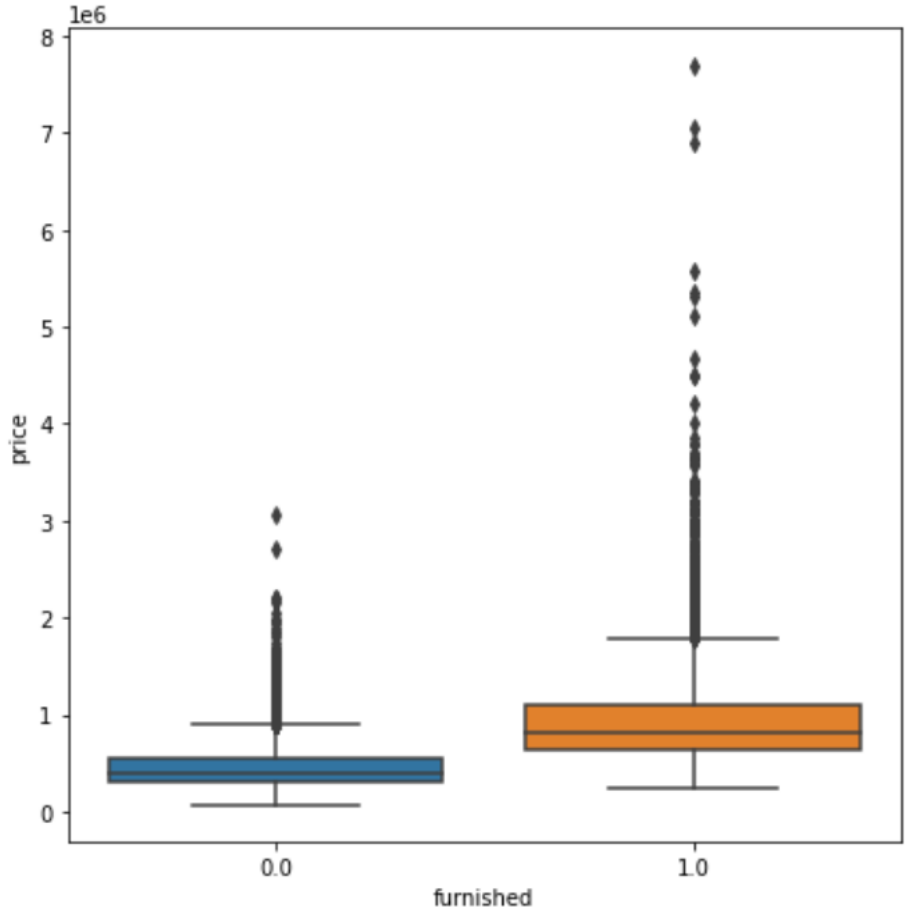
- it does appear overall that houses renovated in more recent years have higher prices than those in years less than ~1978.

Bivariate Analysis: Price vs zipcode



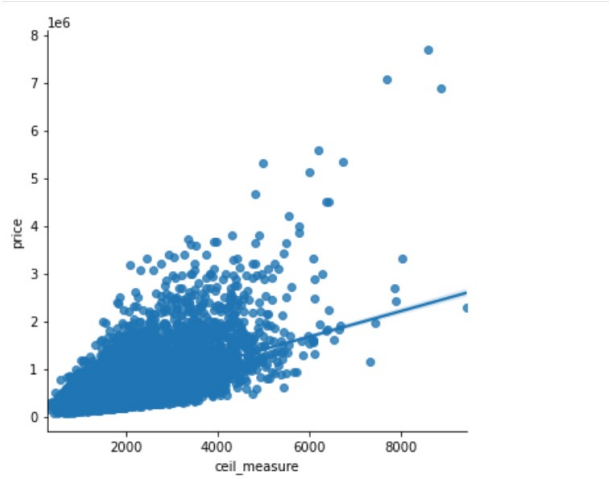
- specific zip codes seem to shine: 99102 (Albion, WA), 98004 (Bellevue,WA), 98008 (King County, WA), 99039 (King County, WA), 99040 (King County, WA)

Bivariate Analysis: Price vs furnished



- not surprisingly, furnished houses have higher prices

Bivariate Analysis: Numerical



- visually, the strongest linear relationships are for ceil_measure (square footage not including basement) and living_m (living room area in 2015).