2020

CAPSTONE PROJECT

HOUSING DATA ANALYSIS

ABSTRACT

CREATE A MODEL TO PREDICT THE HOUSE PRICE BASED ON THE HISTORICAL DATA

ML Group 1

Table of Contents

Table of Figures

1 Introduction.

1.1 Summary of problem statement

Housing price prediction is a complicated task. For the good prediction of the house price one must consider different parameters like locality, size of the house, aging factor, different amenities available and condition of the house. Market valuations of nearby houses also plays important role in price of a given house.

In a given problem inner-city data set of 21613 houses is given with 23 different features of a house. All the houses are having different prices based on the values of different features given in below section. Price is a target variable and other 22 variables have an impact on the price. Careful study of the different features and their relation to other features as well as target variable is needed for better price prediction.

cid	dayhours	price	room_bed	room_bath	living_measure	lot_measure	ceil	coast	sight	condition	quality	ceil_measure	basement	yr_built	yr_renovated	zipcode	lat	long	living_measure15	lot_measure15	furnished	total_area
3034200666	20141107T000000	808100	4	3.25	3020	13457	1	0	0	5	9	3020	0	1956	0	98133	47.7174	-122.336	2120	7553	1	16477
8731981640	20141204T000000	277500	4	2.5	2550	7500	1	0	0	3	8	1750	800	1976	0	98023	47.3165	-122.386	2260	8800	0	10050
5104530220	20150420T000000	404000	3	2.5	2370	4324	2	0	0	3	8	2370	0	2006	0	98038	47.3515	-121.999	2370	4348	0	6694
6145600285	20140529T000000	300000	2	1	820	3844	1	0	0	4	6	820	0	1916	0	98133	47.7049	-122.349	1520	3844	0	4664
8924100111	20150424T000000	699000	2	1.5	1400	4050	1	0	0	4	8	1400	0	1954	0	98115	47.6768	-122.269	1900	5940	0	5450
5525400430	20140715T000000	585000	3	2.5	2050	11690	2	0	0	4	9	2050	0	1989	0	98059	47.5279	-122.161	2410	10172	1	13740
2419600075	20141201T000000	465000	3	1.75	1480	6360	1	0	0	3	7	1480	0	1954	0	98133	47.7311	-122.353	1480	6360	0	7840
114101161	20140829T000000	480000	3	1.5	2100	67269	1	0	0	4	7	1220	880	1949	0	98028	47.7592	-122.23	1610	15999	0	69369
7011201550	20140707T000000	780000	4	2	2600	4800	1	0	2	3	8	1400	1200	1953	0	98119	47.637	-122.371	2050	3505	0	7400
7203000640	20140918T000000	215000	4	1	1130	7400	1	0	0	4	7	1130	0	1969	0	98003	47.3437	-122.316	1540	7379	0	8530
7518503685	20141009T000000	402000	2	1	710	5100	1	0	0	5	7	710	0	1905	0	98117	47.6765	-122.381	1530	5100	0	5810
7300400150	20141027T000000	299000	4	2.5	2350	6958	2	0	0	3	9	2350	0	1998	0	98092	47.3321	-122.172	2480	6395	1	9308
2215800050	20150415T000000	785000	4	2.5	3440	56192	2	0	0	3	9	3440	0	1994	0	98053	47.6969	-122.046	3150	44431	1	59632
7443000480	20150507T000000	865000	4	2	2750	5527	2	0	0	3	8	2130	620	1901	1987	98119	47.6513	-122.368	1290	1764	0	8277
5072100095	20141117T000000	554000	5	2.5	3440	12900	1	0	2	4	8	1720	1720	1958	0	98166	47.4426	-122.342	2100	10751	0	16340
1387301730	20150202T000000	361000	3	1.5	1200	7236	1	0	0	3	7	1200	0	1975	0	98011	47.739	-122.194	1680	7800	0	8436
1310430130	20141009T000000	459000	4	2.75	2790	6600	2	0	0	3	9	2790	0	2000	0	98058	47.4362	-122.109	2900	6752	1	9390
3352400351	20141121T000000	200000	3	1	1480	5600	1	0	0	4	6	940	540	1947	0	98178	47.5045	-122.27	1350	11100	0	7080
3678900110	20140610T000000	403000	2	1	1100	3598	1	0	0	4	7	1100	0	1926	0	98144	47.5738	-122.313	1240	3598	0	4698
2474400250	20140630T000000	327500	3	2.25	2310	7200	2	0	0	3	8	2310	0	1990	0	98031	47.4051	-122.193	1960	7201	0	9510
8820900029	20140610T000000	700000	5	2.75	3100	9825	2	0	2	4	8	3100	0	1950	1982	98125	47.7188	-122.281	2120	8400	0	12925
263000050	20141031T000000	730000	3	2.5	2160	8809	1	0	0	3	9	1540	620	2014	0	98103	47.6994	-122.349	930	5420	1	10969
9406500350	20141229T000000	207000	2	1.5	1068	1158	2	0	0	3	7	1068	0	1990	0	98028	47.753	-122.244	1078	1278	0	2226
9533100145	20150205T000000	750000	3	1	1120	8549	1	0	0	3	7	1120	0	1952	0	98004	47.6294	-122.205	1440	8640	0	9669
5694500105	20141204T000000	595000	2	2	1510	4000	1	0	0	4	7	1010	500	1900	0	98103	47.6582	-122.345	1920	4000	0	5510
3291800710	20141120T000000	338000	4	3	2090	7500	1	0	0	3	7	1370	720	1986	0	98056	47.4888	-122.182	1810	7650	0	9590
9126100815	20141217T000000	500000	3	2	1560	1156	3	0	0	3	8	1560	0	2014	0	98122	47.605	-122.304	1560	1728	0	2716
3416600800	20150209T000000	834000	4	2.5	2370	4000	1.5	0	2	5	8	1980	390	1928	0	98144	47.601	-122.294	2440	5750	0	6370
7855000460	20141007T000000	1450000	3	2.75	3940	9671	1	0	4	5	9	2140	1800	1967	0	98006	47.5654	-122.158	3390	9360	1	13611
6204410330	20141020T000000	432000	4	1.75	2410	8400	1	0	0	3	7	1600	810	1978	0	98011	47.7341	-122.2	1850	8400	0	10810
9206500250	20140909T000000	1100000	4	4	3770	8899	2	0	0	3	10	2940	830	2006	0	98074	47.6476	-122.079	3300	8308	1	12669
3172600031	20150327T000000	325000	3	1.5	1590	7936	1	0	0	3	7	1590	0	1956	0	98106	47.5201	-122.366	1590	7936	0	9526
2896600020	20150325T000000	460000	3	1.75	1520	7700	1	0	0	3	7	820	700	1969	0	98034	47.7226	-122.219	1420	7674	0	9220
7237500390	20141110T000000	1570000	5	4.5	6070	14731	2	0	0	3	11	6070	0	2004	0	98059	47.5306	-122.134	4750	13404	1	20801
1115600130	20140930T000000	415000	4	2.5	2891	6499	2	0	0	3	9	2891	0	2014	0	98001	47.3359	-122.257	2550	8383	1	9390
9187200275	20150420T000000	905000	4	2.25	2240	5000	2	0	0	3	8	1770	470	1900	2014	98122	47.6027	-122.295	2120	5000	0	7240
224069195	20140616T000000	759950	3	2.5	3100	23790	2	0	0	3	9	3100	0	2002	0	98075	47.5882	-122.011	2250	40854	1	26890
2769600480	20150430T000000	600000	2	2	1270	5000	1	0	0	3	6	1270	0	1944	0	98107	47.6729	-122.363	2190	5000	0	6270
301400240	20140922T000000	282900	4	2.5	1710	3500	2	0	0	3	7	1710	0	2014	0	98002	47.3448	-122.217	1710	3500	0	5210

Figure 1-1 .CSV File of the data.

1.2 Data Set.

cid: *a notation for a house*

dayhours: *Date house was sold*

price: *Price is prediction target*

room_bed: *Number of Bedrooms/House*

room_bath: *Number of bathrooms/bedrooms*

living_measure: *square footage of the home*

lot_measure: *square footage of the lot*

ceil: *Total floors (levels) in house*

coast: *House which has a view to a waterfront*

sight: *Has been viewed*

condition: *How good the condition is (Overall)*

quality: *grade given to the housing unit, based on grading system*

ceil_measure: *square footage of house apart from basement*

basement_measure: *square footage of the basement*

yr_built: *Built Year*

yr_renovated: *Year when house was renovated*

zipcode: *zip*

lat: *Latitude coordinate*

long: *Longitude coordinate*

living_measure15: *Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area*

lot_measure15: *lotSize area in 2015(implies-- some renovations)*

furnished: *Based on the quality of room*

total_area: *Measure of both living and lot*

1.3 Data Insights.

1.3.1 Data Types.

There are 4 features with the decimal values (float64), 18 with integer (int64) and 1 is object type.

```
Data columns (total 23 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   cid                21613 non-null  int64
 1   dayhours           21613 non-null  object
 2   price              21613 non-null  int64
 3   room_bed           21613 non-null  int64
 4   room_bath          21613 non-null  float64
 5   living_measure     21613 non-null  int64
 6   lot_measure        21613 non-null  int64
 7   ceil               21613 non-null  float64
 8   coast              21613 non-null  int64
 9   sight              21613 non-null  int64
 10  condition          21613 non-null  int64
 11  quality            21613 non-null  int64
 12  ceil_measure       21613 non-null  int64
 13  basement           21613 non-null  int64
 14  yr_built           21613 non-null  int64
 15  yr_renovated       21613 non-null  int64
 16  zipcode            21613 non-null  int64
 17  lat                21613 non-null  float64
 18  long               21613 non-null  float64
 19  living_measure15   21613 non-null  int64
 20  lot_measure15      21613 non-null  int64
 21  furnished          21613 non-null  int64
 22  total_area         21613 non-null  int64
dtypes: float64(4), int64(18), object(1)
memory usage: 3.8+ MB
```

1.3.2 Dataset Overview.

1.3.2.1 Variable Types.

There are 20 variables with numerical values, 2 with Boolean type and 1 is categorical type.

1.3.2.2 Dataset Info.

No. of. Variables are 23.

No. Of. Observations are 21613.

No. Of. Missing Cells are 0.

No. Of. Duplicate rows are 0.

1.3.2.3 Dataset Warnings.

`basement` has 13126 (60.7%) zeros — Zeros

`dayhours` has a high cardinality: 372 distinct values — Warning `sight` has 19489 (90.2%) zeros — Zeros

`yr_renovated` has 20699 (95.8%) zeros — Zeros

`total_area` is highly correlated with `lot_measure` — High Correlation

`lot_measure` is highly correlated with `total_area` — High Correlation

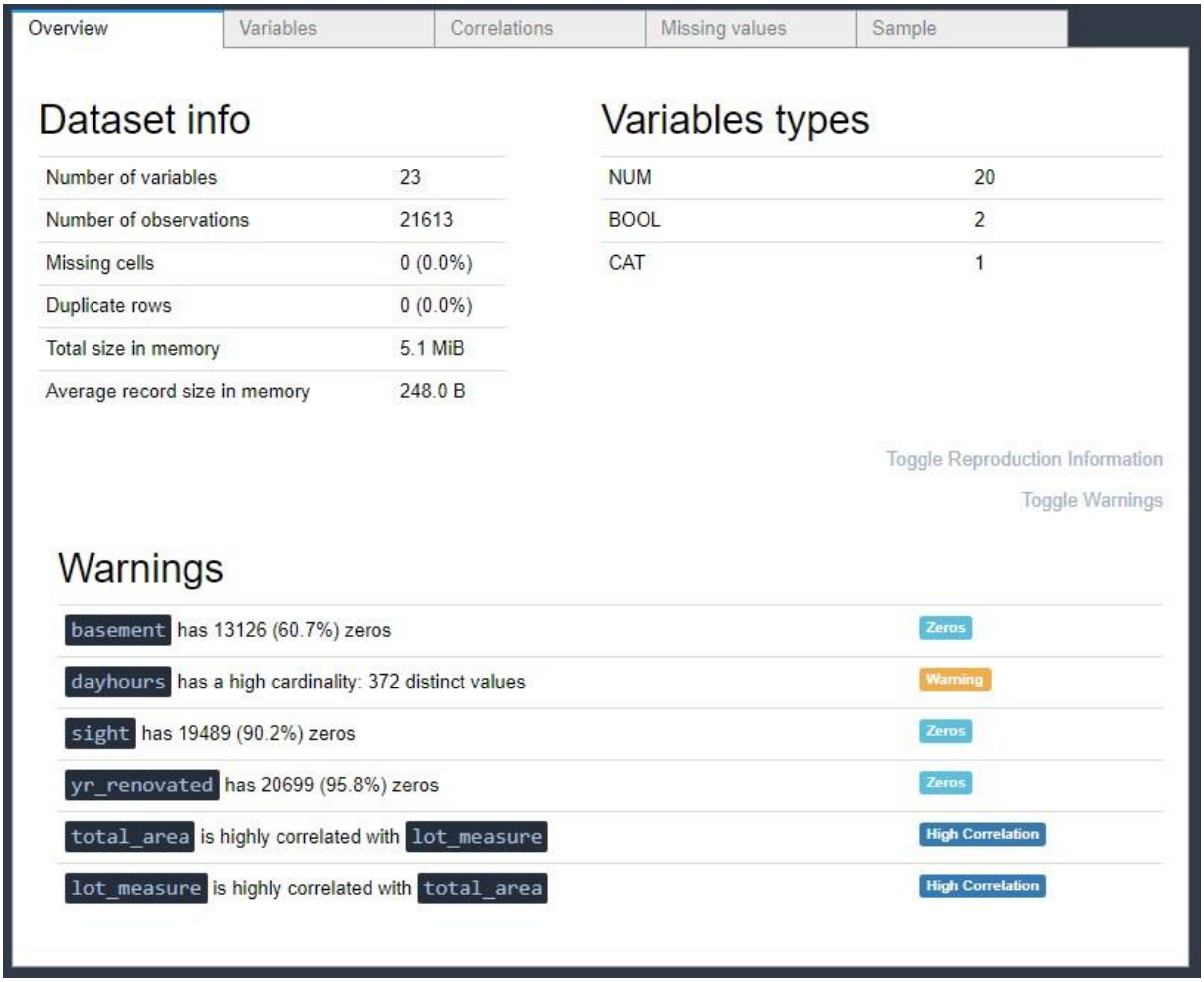


Figure 1-2 Dataset Overview using Pandas Profiling.

1.3.3 Missing Values

There are no missing values present in the dataset.

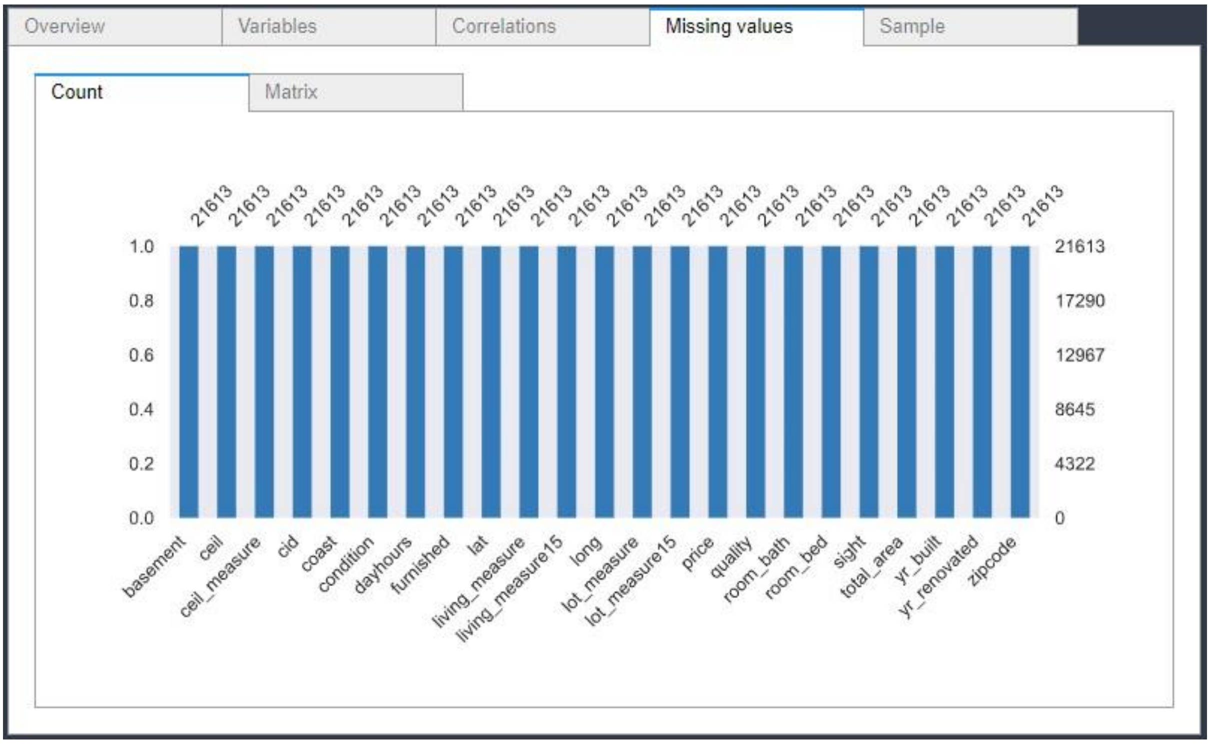


Figure 1-3 Missing Value Count Graph from Pandas Profiling.

Figure 1-4 Missing Value Matrix Graph from Pandas Profiling.

1.3.4 Nan Values

There are no nan values present in the data frame.

```
cid                  0
dayhours             0
price                0
room_bed             0
room_bath            0
living_measure       0
lot_measure          0
ceil                 0
coast                0
sight                0
condition            0
quality              0
ceil_measure         0
basement             0
yr_built             0
yr_renovated         0
zipcode              0
lat                  0
long                 0
living_measure15     0
lot_measure15        0
furnished            0
total_area           0
dtype: int64
```

2 EDA and Pre-processing.

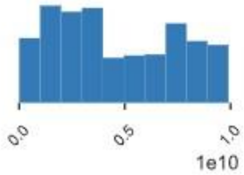
2.1 Univariate Analysis.

2.1.1 CID

The Variable CID represents the house notation. It is slightly skewed with the Skewness of 0.243328. Other Statistics and Histogram are as follows.

cid
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	21436	Mean	4580301521
Unique (%)	0.9918104844306668	Minimum	1000102
Missing	0	Maximum	9900000190
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	1000102	Standard deviation	2876565571
5-th percentile	512480335	Coefficient of variation (CV)	0.6280297396
Q1	2123049194	Kurtosis	-1.2605418705143028
median	3904930410	Mean	4580301521
Q3	7308900445	Median Absolute Deviation (MAD)	2543592458
95-th percentile	9297300429	Skewness	0.24332854761624426
Maximum	9900000190	Sum	98994056770455
Range	9899000088	Variance	8.274629486e+18
Interquartile range (IQR)	5185851251		

Figure 2-1 Statistics of CID Feature.

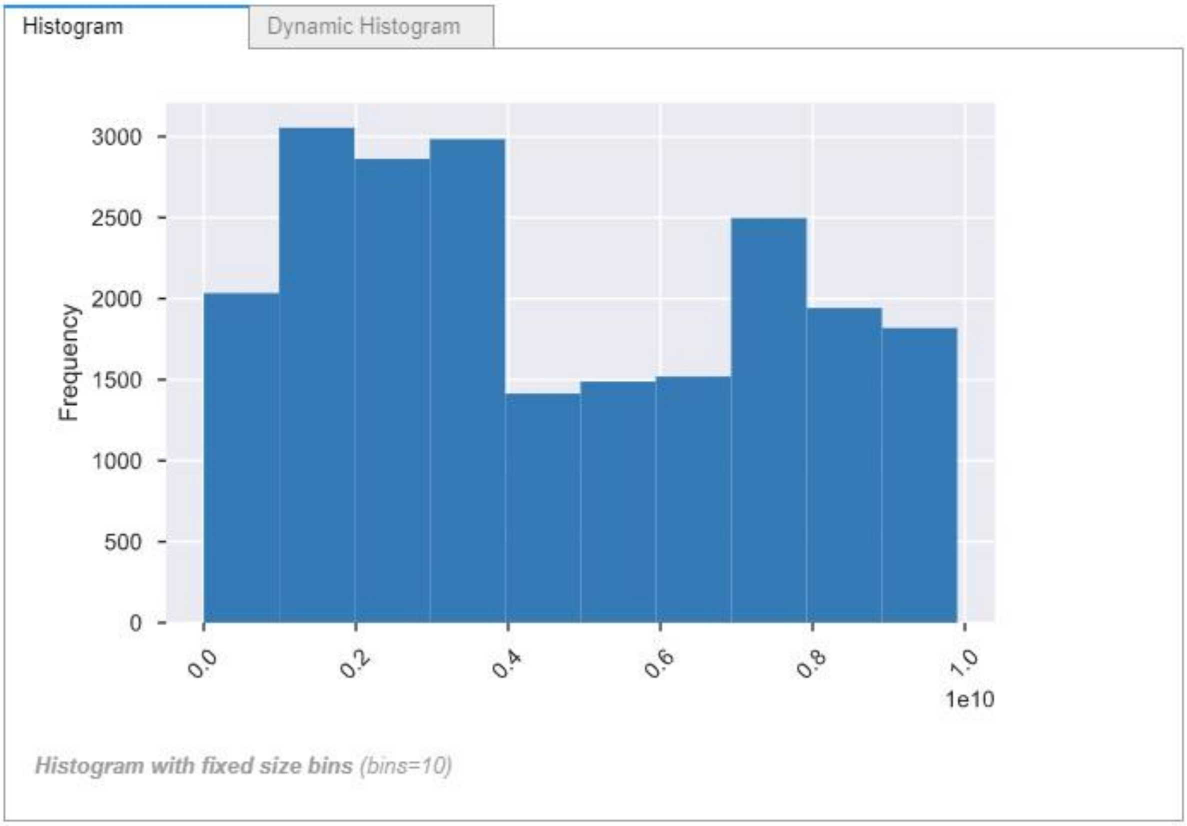


Figure 2-2 CID Distribution Plot.

From the above plot it is clear that the data is skewed.

2.1.2 Dayhours

Dayhours is a categorical variable with high cardinality representing the date and time at which the house was sold.

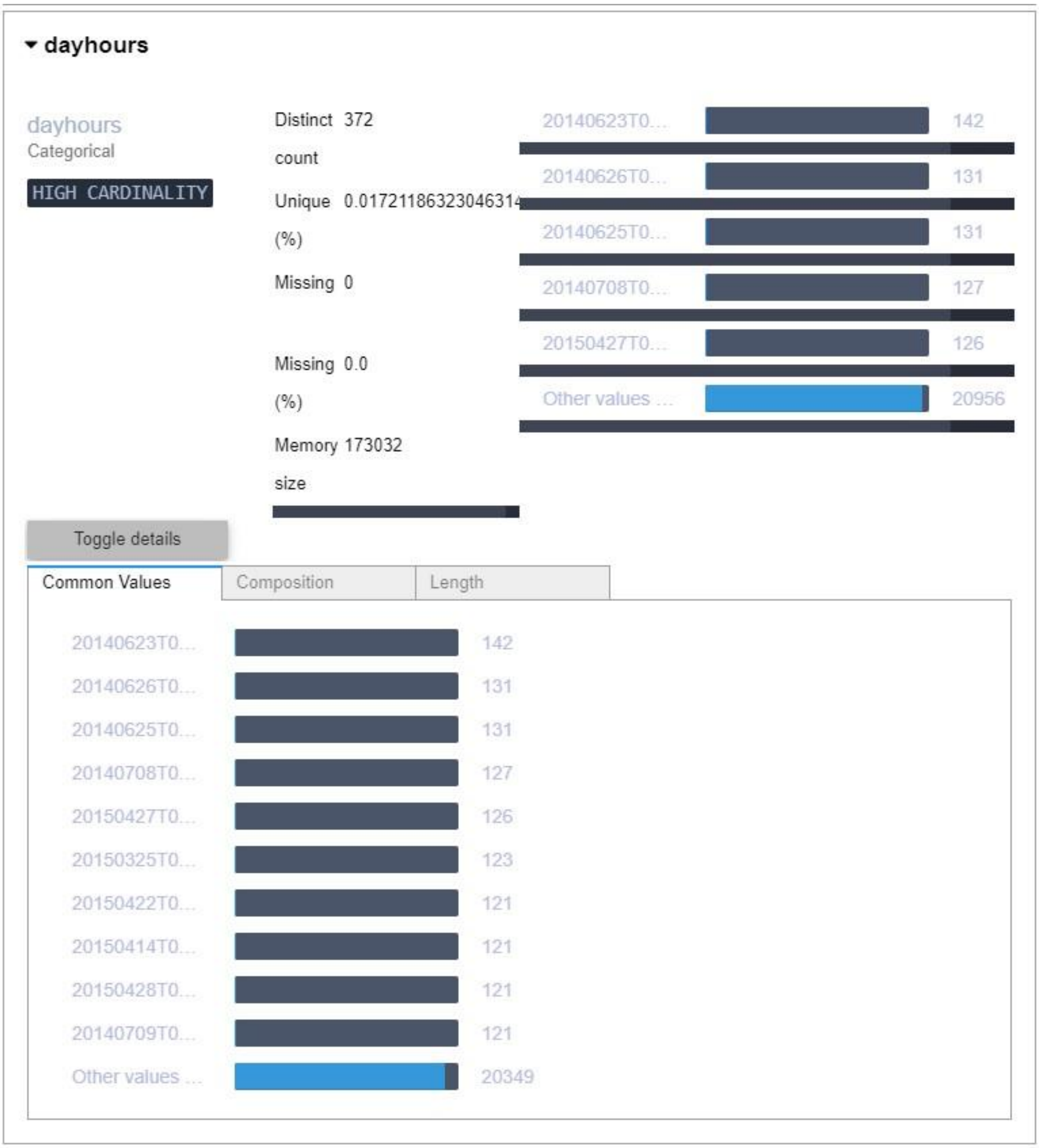


Figure 2-3 Statistics of Dayhours.

2.1.3 Price

It is our target variable. It is highly skewed having skewness of 4.0217155. Other Stats are as follows.

price
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	3625	Mean	540182.1588
Unique (%)	0.16772312959792718	Minimum	75000
Missing	0	Maximum	7700000
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032

Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	75000	Standard deviation	367362.2317
5-th percentile	210000	Coefficient of variation (CV)	0.68007102
Q1	321950	Kurtosis	34.522444410072566
median	450000	Mean	540182.1588
Q3	645000	Median Absolute Deviation (MAD)	234060.8627
95-th percentile	1160000	Skewness	4.021715573048684
Maximum	7700000	Sum	11674956998
Range	7625000	Variance	1.349550093e+11
Interquartile range (IQR)	323050		

Figure 2-4 Statistics of Price.

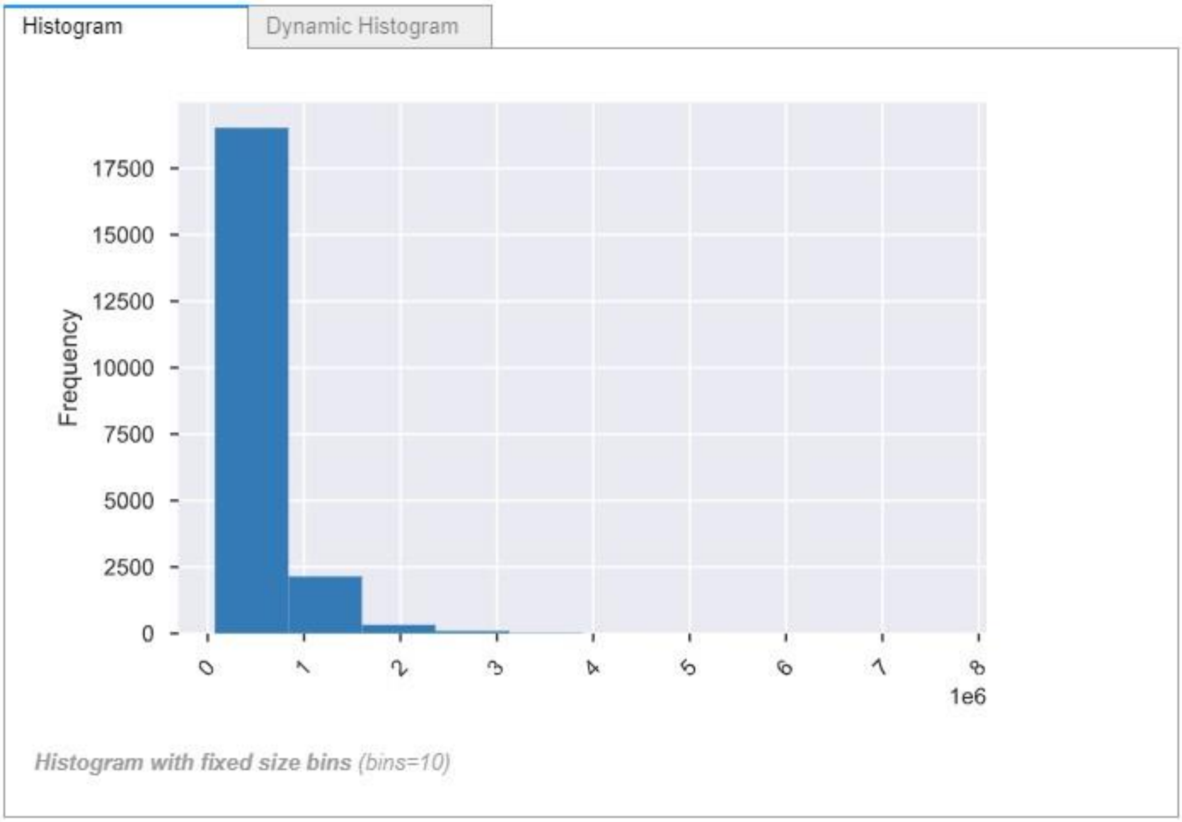
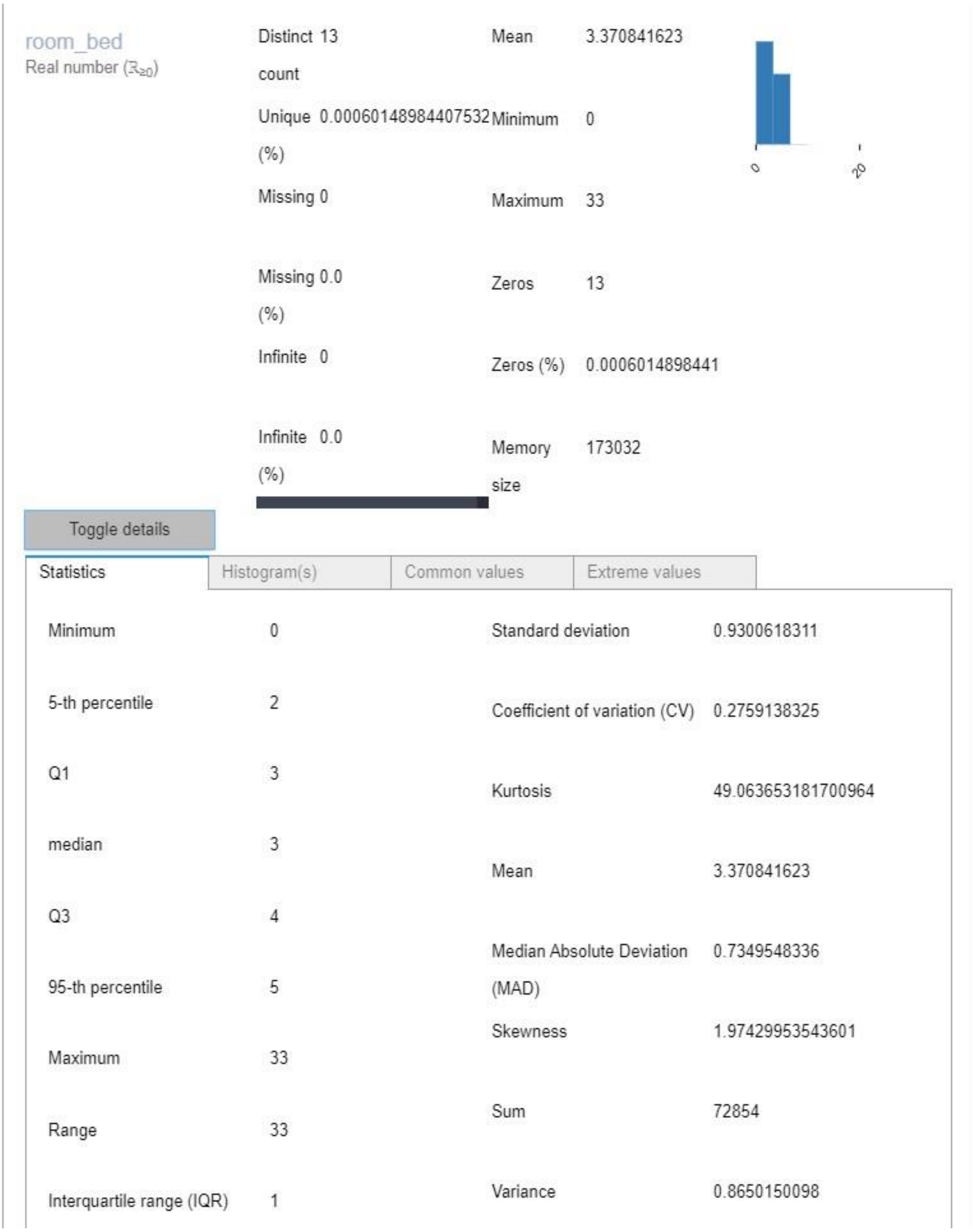


Figure 2-5 Price Distribution plot.

Form the above graph we can clearly see that this feature is highly skewed to left.

2.1.4 Room_Bed

Room_bed is a feature that describes the no of bed rooms in a house. It is having 13 distinct values with a minimum of 0 and a maximum of 33 and a skewness of 1.974299.



room_bed
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	13	Mean	3.370841623
Unique (%)	0.00060148984407532	Minimum	0
Missing	0	Maximum	33
Missing (%)	0.0	Zeros	13
Infinite	0	Zeros (%)	0.0006014898441
Infinite (%)	0.0	Memory size	173032

Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	0	Standard deviation	0.9300618311
5-th percentile	2	Coefficient of variation (CV)	0.2759138325
Q1	3	Kurtosis	49.063653181700964
median	3	Mean	3.370841623
Q3	4	Median Absolute Deviation (MAD)	0.7349548336
95-th percentile	5	Skewness	1.97429953543601
Maximum	33	Sum	72854
Range	33	Variance	0.8650150098
Interquartile range (IQR)	1		

Figure 2-6 Satatistics of room_bed.

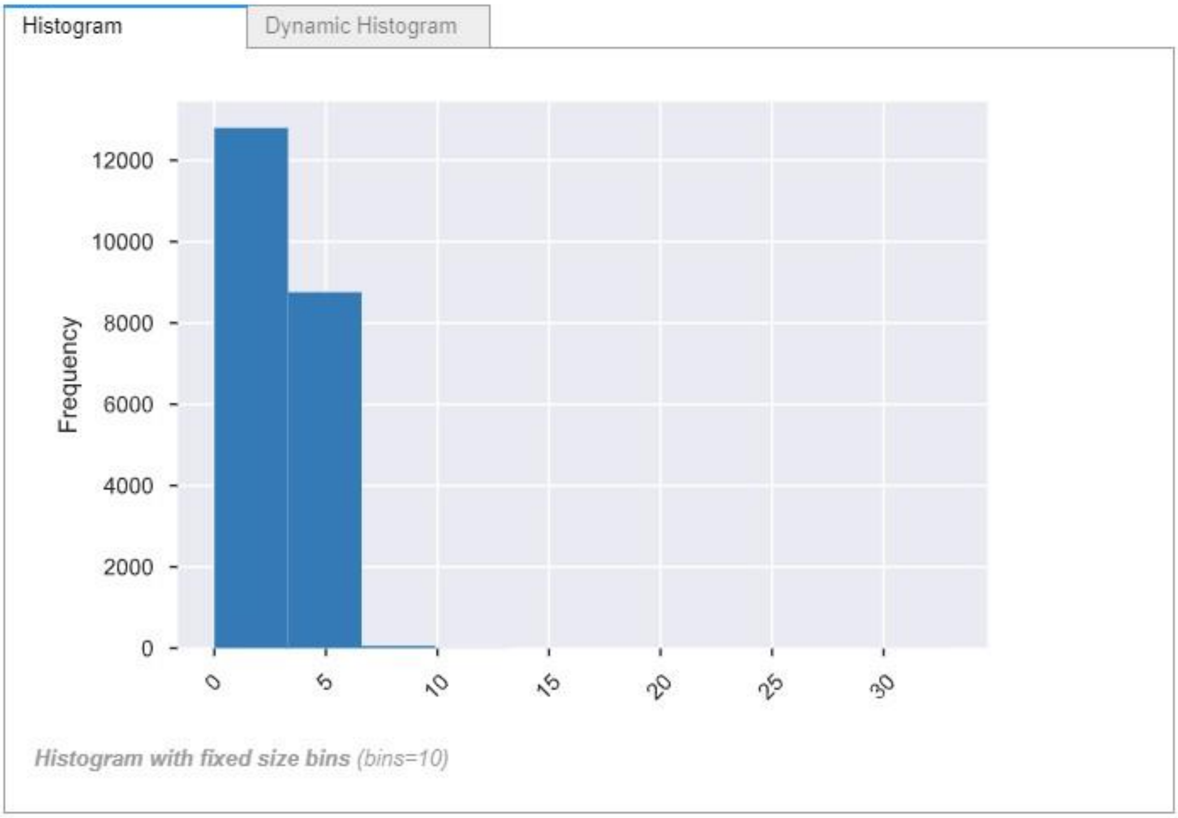


Figure 2-7 Distribution plot of room_bed.

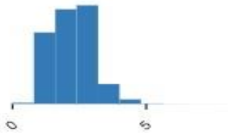
From the above plot the skewness can be clearly seen to left.

2.1.5 Room_Bath

Room_bath is a feature that describes the no of bathrooms a house has.

room_bath
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	30	Mean	2.114757322
Unique (%)	0.00138805348632767	Minimum	0
Missing	0	Maximum	8
Missing (%)	0.0	Zeros	10
Infinite	0	Zeros (%)	0.0004626844954
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	0	Standard deviation	0.7701631572
5-th percentile	1	Coefficient of variation (CV)	0.3641851238
Q1	1.75	Kurtosis	1.2799024441754536
median	2.25	Mean	2.114757322
Q3	2.5	Median Absolute Deviation (MAD)	0.6153609574
95-th percentile	3.5	Skewness	0.511107573347417
Maximum	8	Sum	45706.25
Range	8	Variance	0.5931512887
Interquartile range (IQR)	0.75		

Figure 2-8 Statistics for room_bath

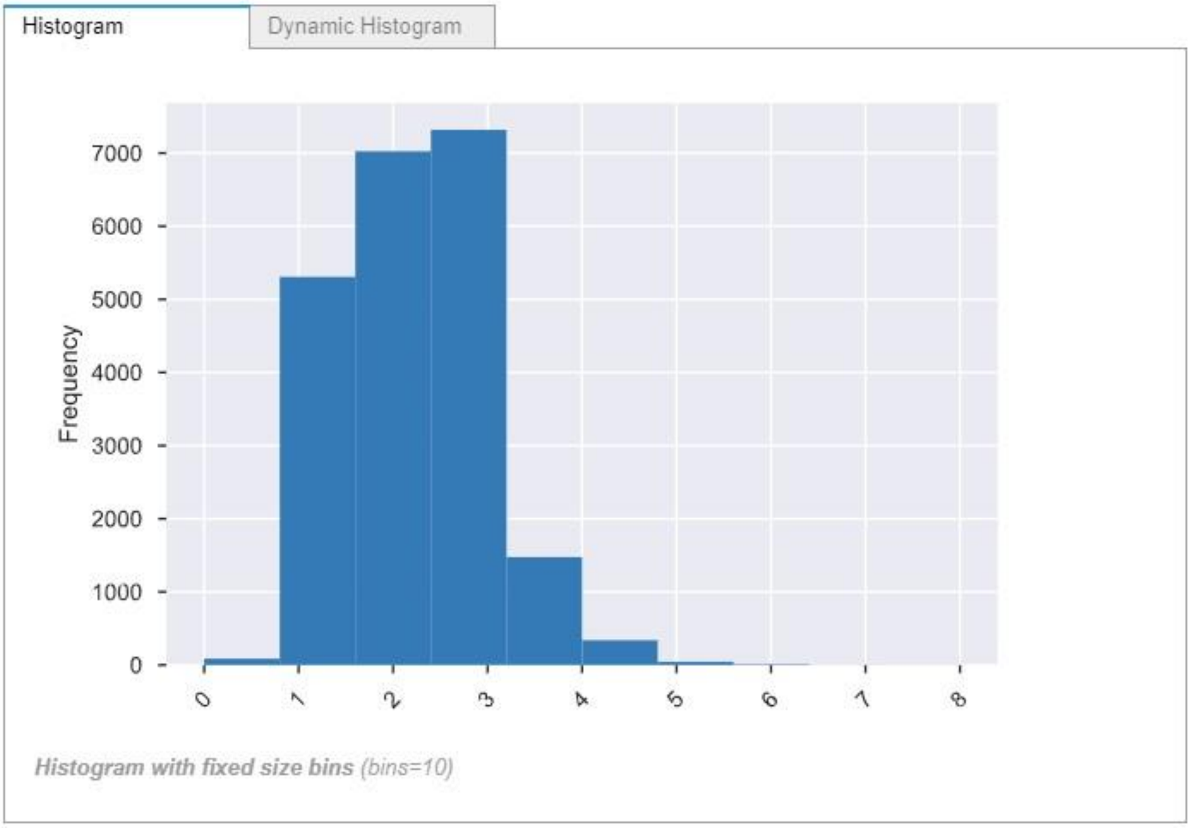


Figure 2-9 Distribution of room_bath.

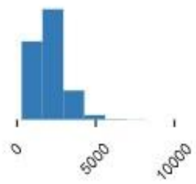
From the above plot it is clear that it is slightly skewed to the left and more than 7000 houses are having no of bathrooms from 2.5 to 3.25.

2.1.6 Living_measure

This feature describes the living area that is available in a particular house. It has 1038 distinct values.

living_measure
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	1038	Mean	2079.899736
Unique (%)	0.04802665062693749	Minimum	290
Missing	0	Maximum	13540
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	290	Standard deviation	918.440897
5-th percentile	940	Coefficient of variation (CV)	0.4415794093
Q1	1427	Kurtosis	5.243092989817112
median	1910	Mean	2079.899736
Q3	2550	Median Absolute Deviation (MAD)	698.3239196
95-th percentile	3760	Skewness	1.471555426802092
Maximum	13540	Sum	44952873
Range	13250	Variance	843533.6814
Interquartile range (IQR)	1123		

Figure 2-10 statistics of living_measure.

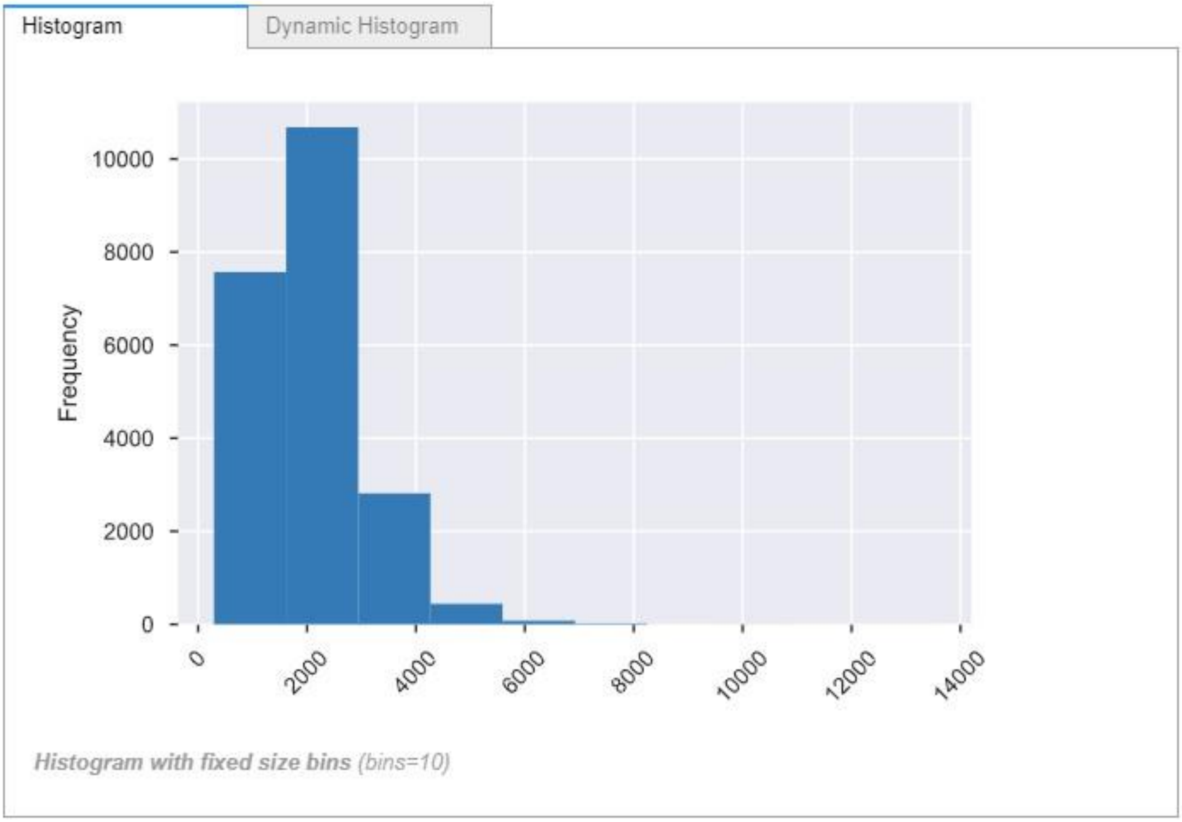
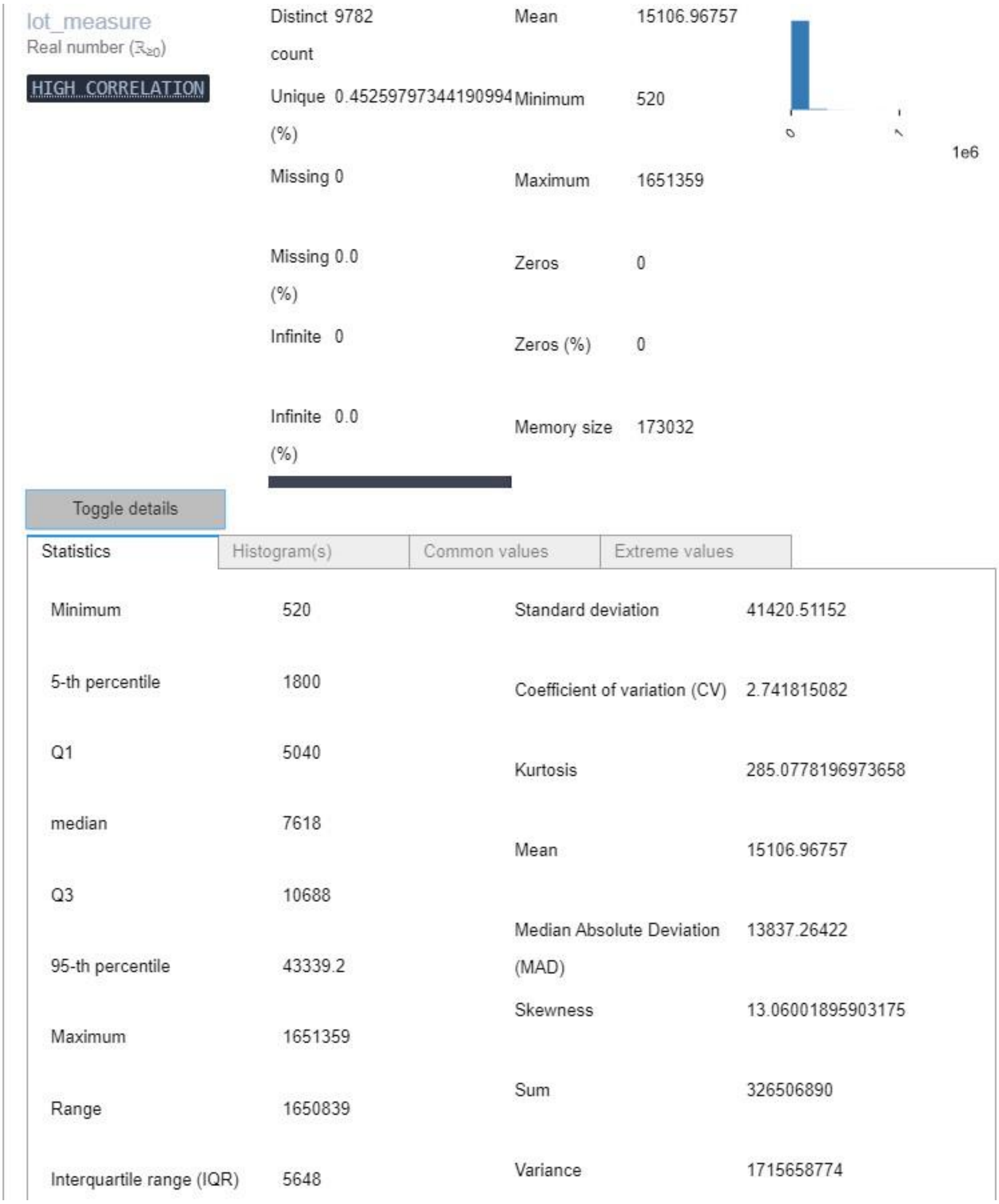


Figure 2-11 Distribution of Living_measure.

It is also highly skewed to the left with a skewness of 1.47.

2.1.7 Lot_Measure.

This feature describes about the square footage of the lot ie the size of the lot in sq foot. It is highly correlated to the feature total_area. It is also highly skewed with a skewness of 13.06.



lot_measure
Real number ($\mathbb{R}_{\geq 0}$)
HIGH CORRELATION

Distinct count	9782	Mean	15106.96757
Unique (%)	0.45259797344190994	Minimum	520
Missing	0	Maximum	1651359
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032

Minimum	520	Standard deviation	41420.51152
5-th percentile	1800	Coefficient of variation (CV)	2.741815082
Q1	5040	Kurtosis	285.0778196973658
median	7618	Mean	15106.96757
Q3	10688	Median Absolute Deviation (MAD)	13837.26422
95-th percentile	43339.2	Skewness	13.06001895903175
Maximum	1651359	Sum	326506890
Range	1650839	Variance	1715658774
Interquartile range (IQR)	5648		

Figure 2-12 Statistics of lot_measure.

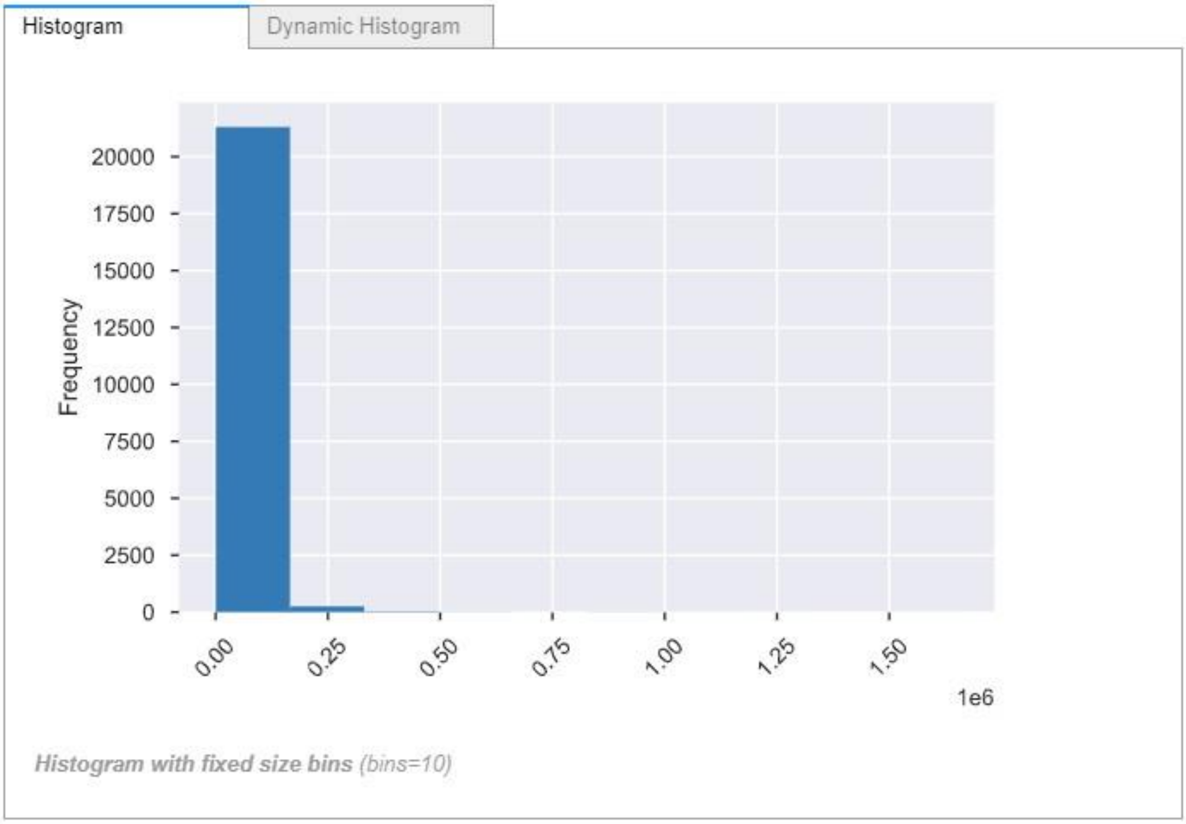


Figure 2-13 Distribution of lot_measure

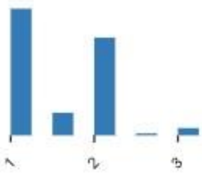
From the above plot it is clear that lot_measure is also skewed to the left.

2.1.8 Ceil

This feature gives us the basic information on the no of floors/levels present in the house. It is having 6 distinct values ranging from 1 to 3.5.

ceil
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	6	Mean	1.494308981
Unique (%)	0.00027761069726553	Minimum	1
Missing	0	Maximum	3.5
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	1	Standard deviation	0.5399888951
5-th percentile	1	Coefficient of variation (CV)	0.361363615
Q1	1	Kurtosis	-0.4847229368298285
median	1.5	Mean	1.494308981
Q3	2	Median Absolute Deviation (MAD)	0.4885226404
95-th percentile	2	Skewness	0.616176721157611
Maximum	3.5	Sum	32296.5
Range	2.5	Variance	0.2915880069
Interquartile range (IQR)	1		

Figure 2-14 Statistics of ceil.

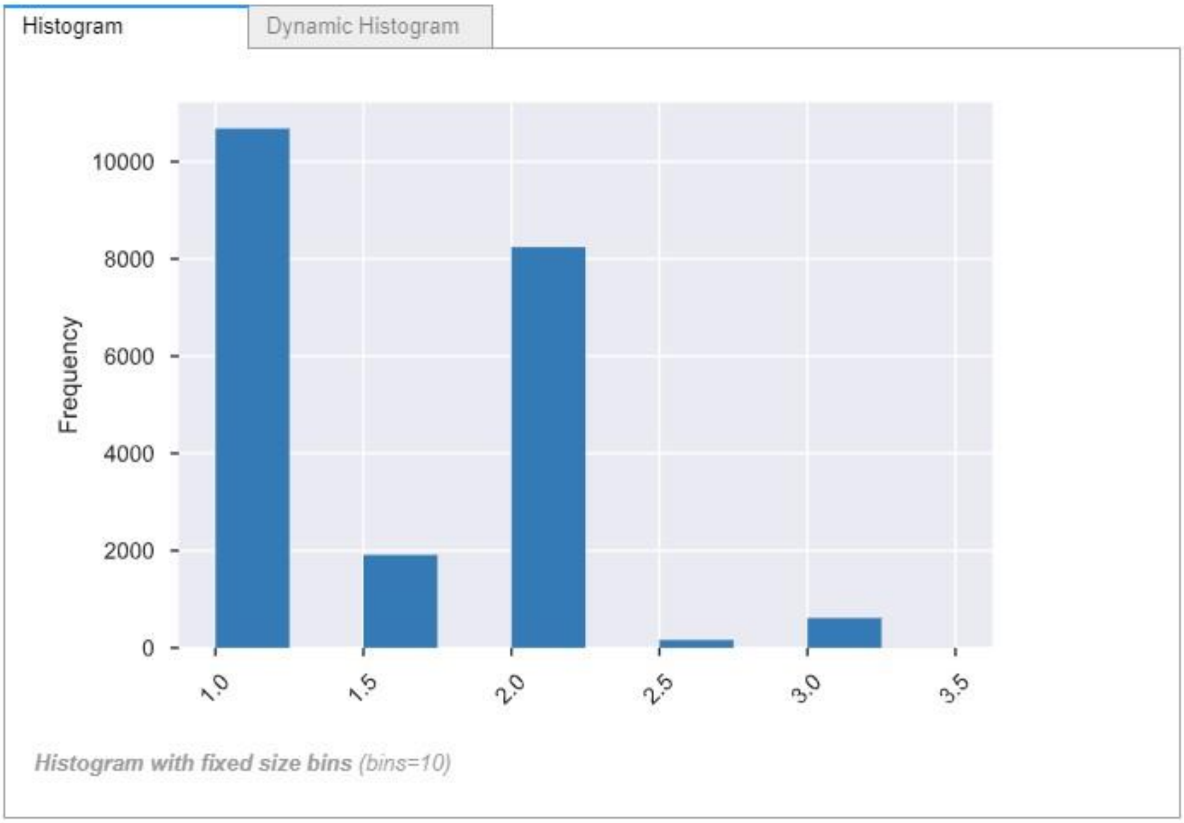


Figure 2-15 Distribution of ceil.

From the above distribution plot, we can see that it is skewed to left and having more than 10,000 houses with 1 ceiling and more than 8,000 houses having 2 ceiling.

2.1.9 Coast

This feature informs us about whether the house is located near a coast or not. The variable is basically of Boolean nature i.e. 0= no and 1= yes. Here the zeros don't indicate null values but tell us about the location of the house is near to the coast or not.

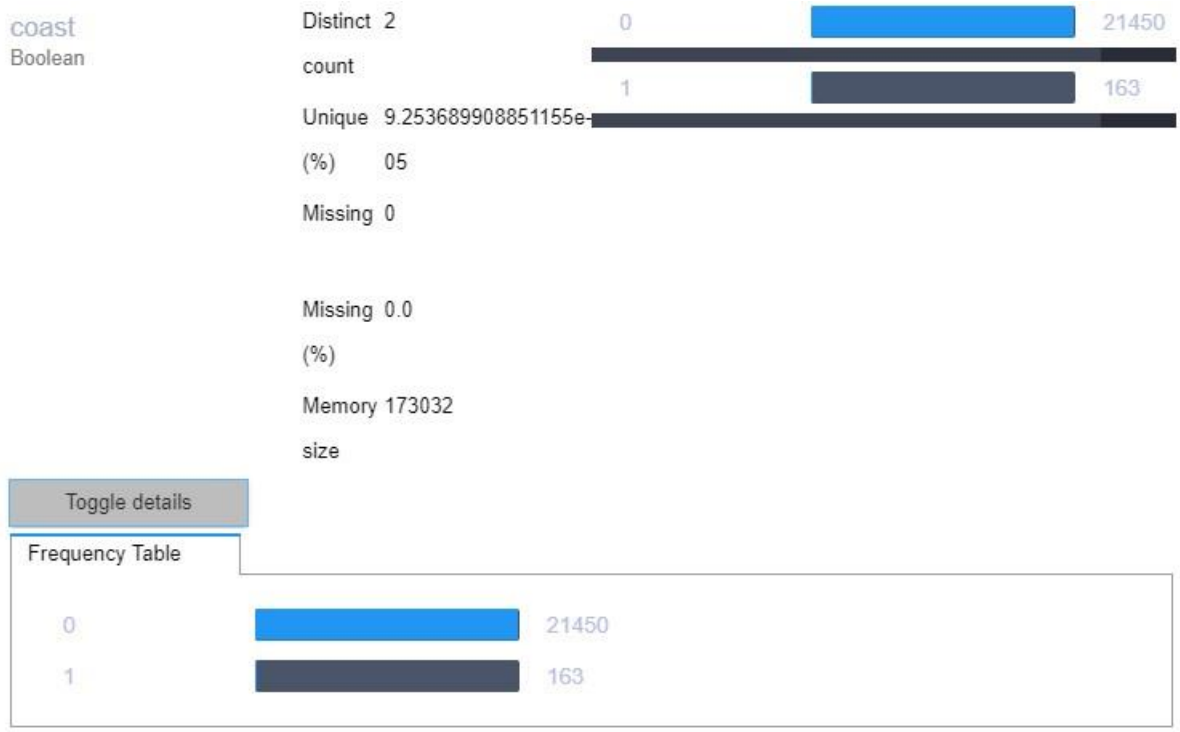


Figure 2-16 Statistics of Coast.

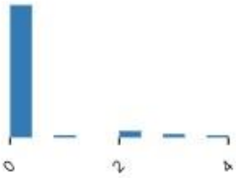
21,450 houses are not located near to a coast where as 163 are located near a coast.

2.1.10 Sight

It gives us information on how many times the sight i.e. house has been viewed.

sight
Real number ($\mathbb{R}_{\geq 0}$)
ZEROS

Distinct count	5	Mean	0.2343034285
Unique (%)	0.00023134224772127	Minimum	0
Missing	0	Maximum	4
Missing (%)	0.0	Zeros	19489
Infinite	0	Zeros (%)	0.9017258132
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	0	Standard deviation	0.7663175693
5-th percentile	0	Coefficient of variation (CV)	3.270620384
Q1	0	Kurtosis	10.893021684601504
median	0	Mean	0.2343034285
Q3	0	Median Absolute Deviation (MAD)	0.4225548992
95-th percentile	2	Skewness	3.3957495932487136
Maximum	4	Sum	5064
Range	4	Variance	0.587242617
Interquartile range (IQR)	0		

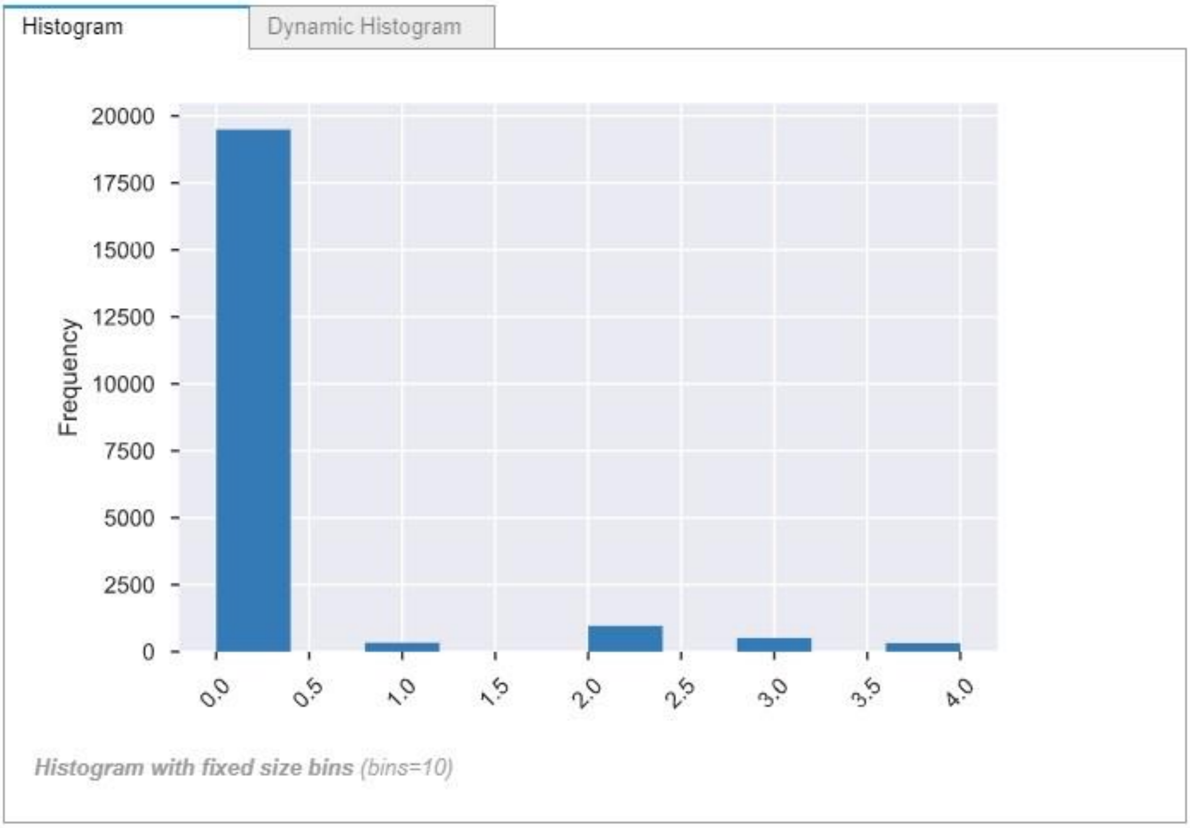
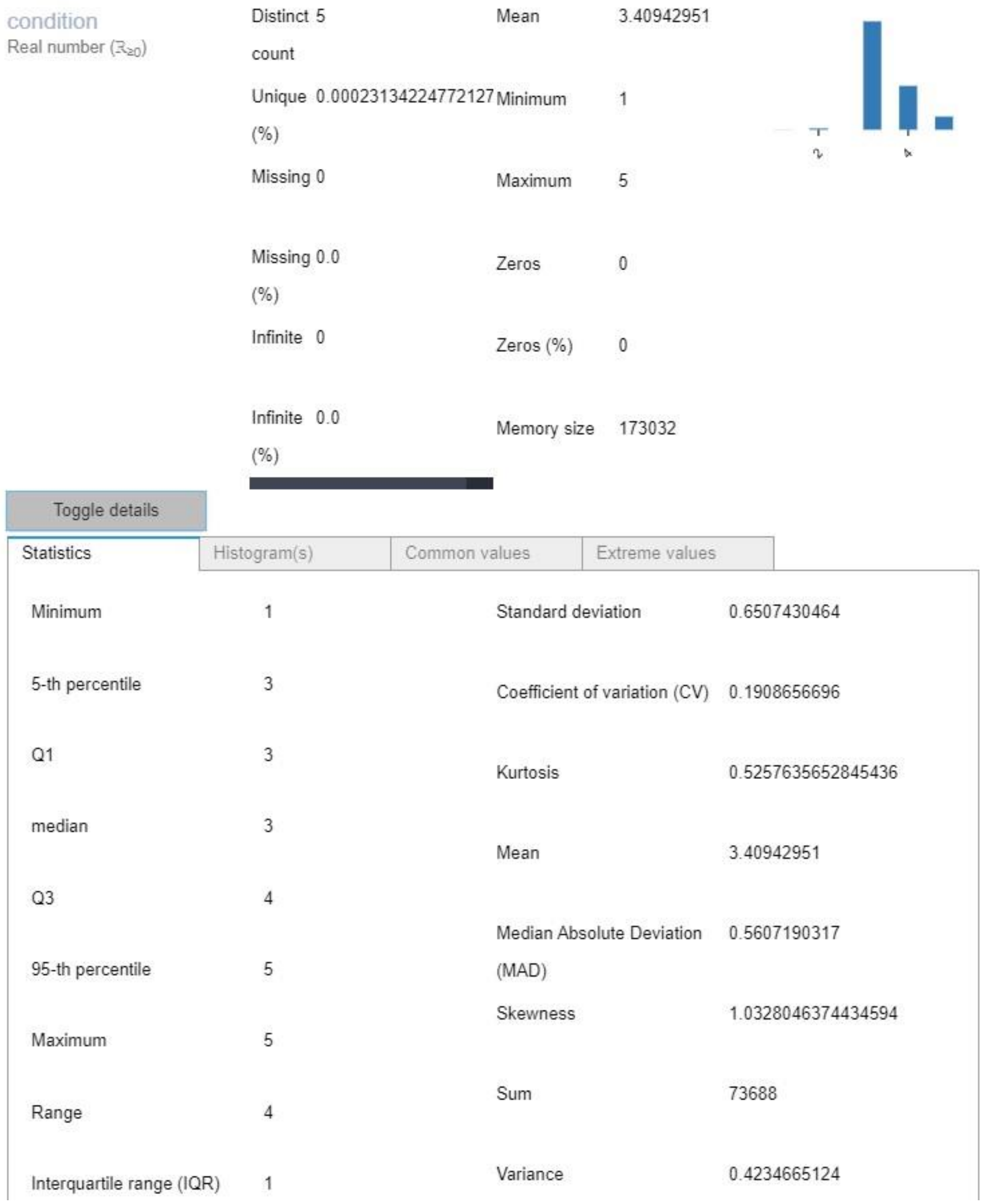


Figure 2-17 Distribution of Sight.

It is skewed to left and has 5 distinct valuers ranging from 0 to 4.

2.1.11 Condition.

It gives information on how good the overall conditions of the house are. The condition of the house is rated on 5 distinct values ranging from 1 to 5. 5 implying the best condition.



condition
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	5	Mean	3.40942951
Unique (%)	0.00023134224772127	Minimum	1
Missing	0	Maximum	5
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032

Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	1	Standard deviation	0.6507430464
5-th percentile	3	Coefficient of variation (CV)	0.1908656696
Q1	3	Kurtosis	0.5257635652845436
median	3	Mean	3.40942951
Q3	4	Median Absolute Deviation (MAD)	0.5607190317
95-th percentile	5	Skewness	1.0328046374434594
Maximum	5	Sum	73688
Range	4	Variance	0.4234665124
Interquartile range (IQR)	1		

Figure 2-18 Statistics of condition.

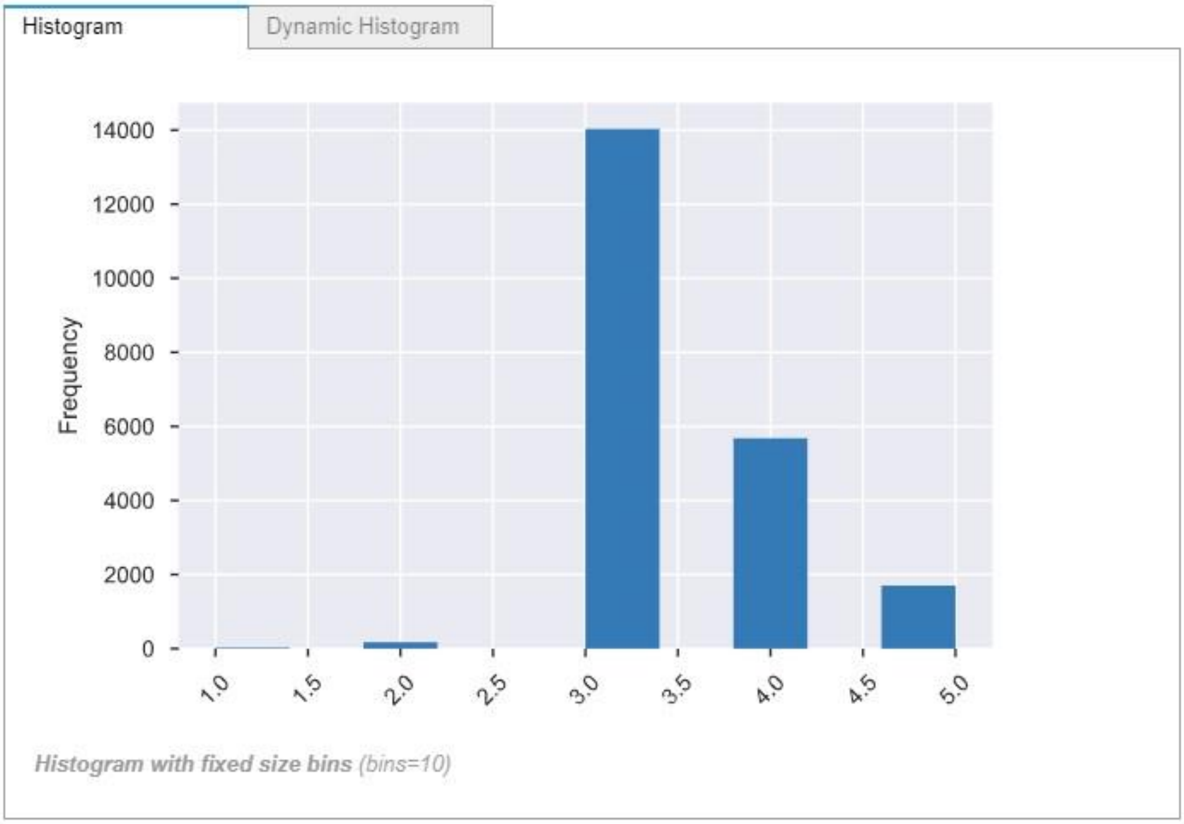


Figure 2-19 Distribution of Condition.

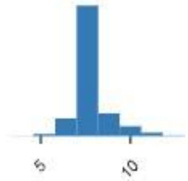
The distribution plot shows that the data is skewed to the right. More than 14,000 houses are rated with 3-star conditions.

2.1.12 Quality.

This feature gives us information om grades provide by a grading system which was given to a house. It has 12 distinct values ranging from 1 to 13 where 13 implies a best quality house.

quality
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	12	Mean	7.656873178
Unique (%)	0.00055522139453106	Minimum	1
Missing	0	Maximum	13
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	1	Standard deviation	1.175458757
5-th percentile	6	Coefficient of variation (CV)	0.1535168116
Q1	7	Kurtosis	1.1909320773987648
median	7	Mean	7.656873178
Q3	8	Median Absolute Deviation (MAD)	0.929600303
95-th percentile	10	Skewness	0.7711032007576065
Maximum	13	Sum	165488
Range	12	Variance	1.381703289
Interquartile range (IQR)	1		

Figure 2-20 Statistics of quality.

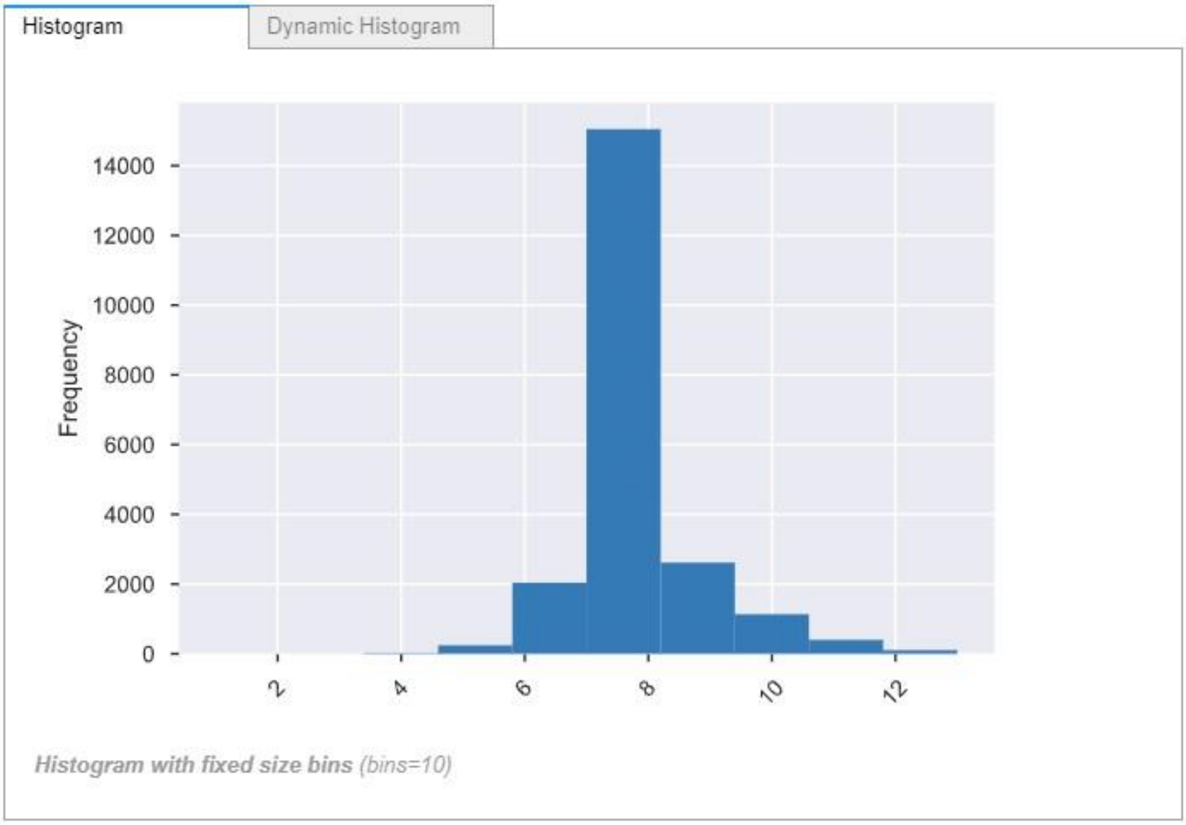


Figure 2-21 Distribution of quality.

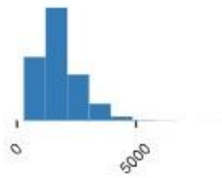
Quality is slightly skewed to right with skewness of 0.771103.

2.1.13 Ceil_Measure

It gives us information on the square foot area of the house excluding basement.

ceil_measure
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	946	Mean	1788.390691
Unique (%)	0.04376995326886596	Minimum	290
Missing	0	Maximum	9410
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	290	Standard deviation	828.0909777
5-th percentile	850	Coefficient of variation (CV)	0.4630369538
Q1	1190	Kurtosis	3.40230362139787
median	1560	Mean	1788.390691
Q3	2210	Median Absolute Deviation (MAD)	640.3860357
95-th percentile	3400	Skewness	1.4466644733818372
Maximum	9410	Sum	38652488
Range	9120	Variance	685734.6673
Interquartile range (IQR)	1020		

Figure 2-22 Statistics of ceil_measure.

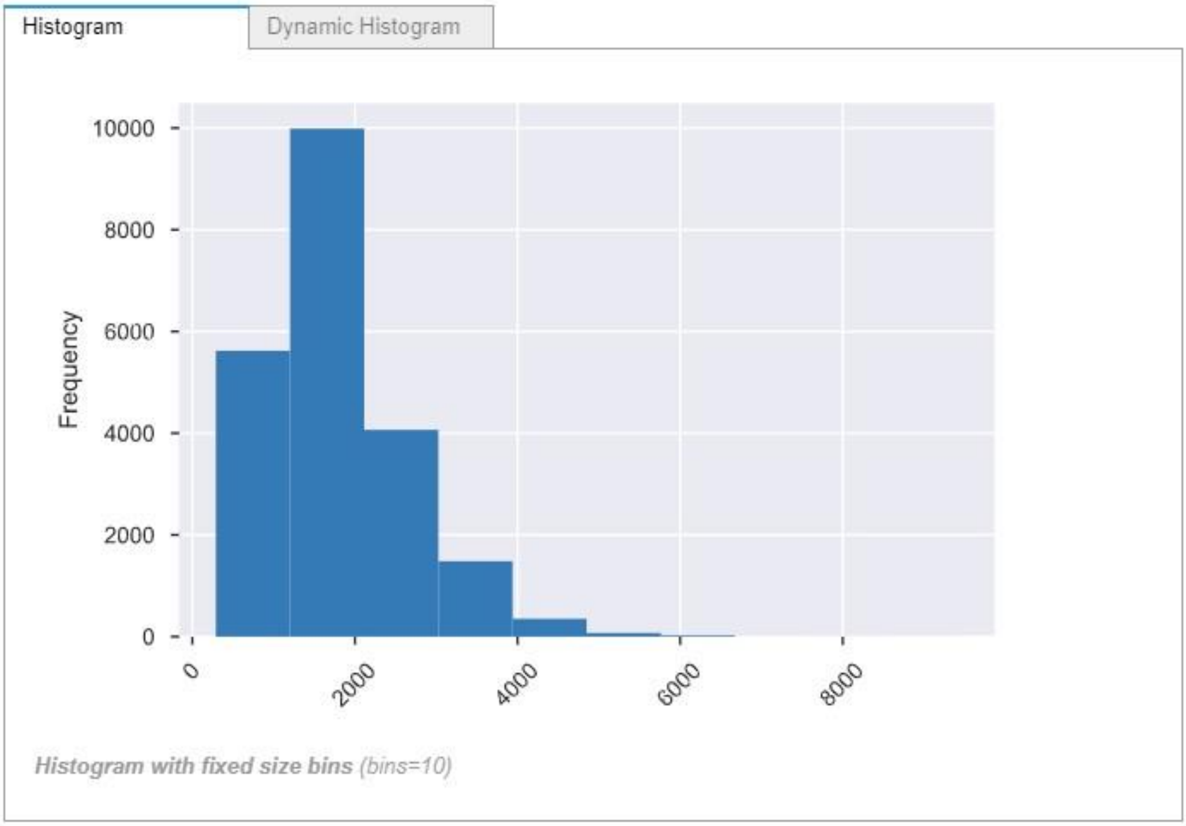


Figure 2-23 Distribution of ceil_measure.

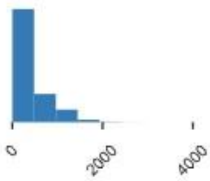
Ceil_measure is also skewed to the left.

2.1.14 Basement.

This feature contains information about the area of the basement. It is having high no of zeros and 306 distinct values. The zeros here indicate that 13,126 houses don't have a basement.

basement
Real number ($\mathbb{R}_{\geq 0}$)
ZEROS

Distinct count	306	Mean	291.5090455
Unique (%)	0.01415814556054226	Minimum	0
Missing	0	Maximum	4820
Missing (%)	0.0	Zeros	13126
Infinite	0	Zeros (%)	0.6073196687
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	0	Standard deviation	442.5750427
5-th percentile	0	Coefficient of variation (CV)	1.518220616
Q1	0	Kurtosis	2.7155742114129353
median	0	Mean	291.5090455
Q3	560	Median Absolute Deviation (MAD)	363.2358668
95-th percentile	1190	Skewness	1.5779650555996247
Maximum	4820	Sum	6300385
Range	4820	Variance	195872.6684
Interquartile range (IQR)	560		

Figure 2-24 Statistics of Basement.

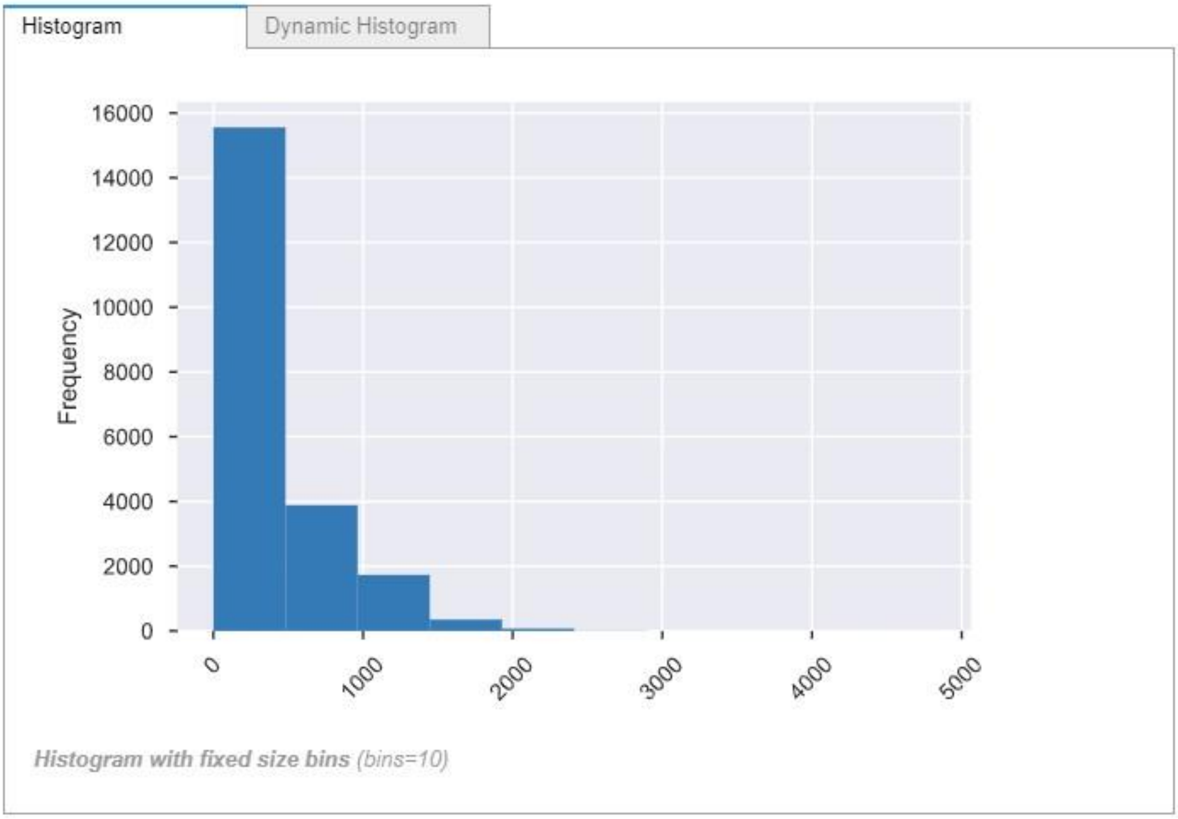


Figure 2-25 Distribution of basement.

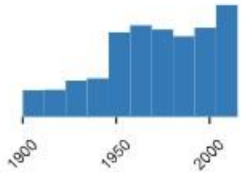
Due to the great no of zeros present in this feature the distribution is highly skewed to the left.

2.1.15 Yr_Built

The year in which the house was built can be seen in this feature. The oldest house being built in 1900 and the newest in 2015. It has 116 distinct values and no zeros or missing columns.

yr_built
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	116	Mean	1971.005136
Unique (%)	0.00536714014713366	Minimum	1900
Missing	0	Maximum	2015
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	1900	Standard deviation	29.3734108
5-th percentile	1915	Coefficient of variation (CV)	0.01490275711
Q1	1951	Kurtosis	-0.657407504733527
median	1975	Mean	1971.005136
Q3	1997	Median Absolute Deviation (MAD)	24.56566156
95-th percentile	2011	Skewness	-0.4698053988143677
Maximum	2015	Sum	42599334
Range	115	Variance	862.7972622
Interquartile range (IQR)	46		

Figure 2-26 Statistics of yr_built.

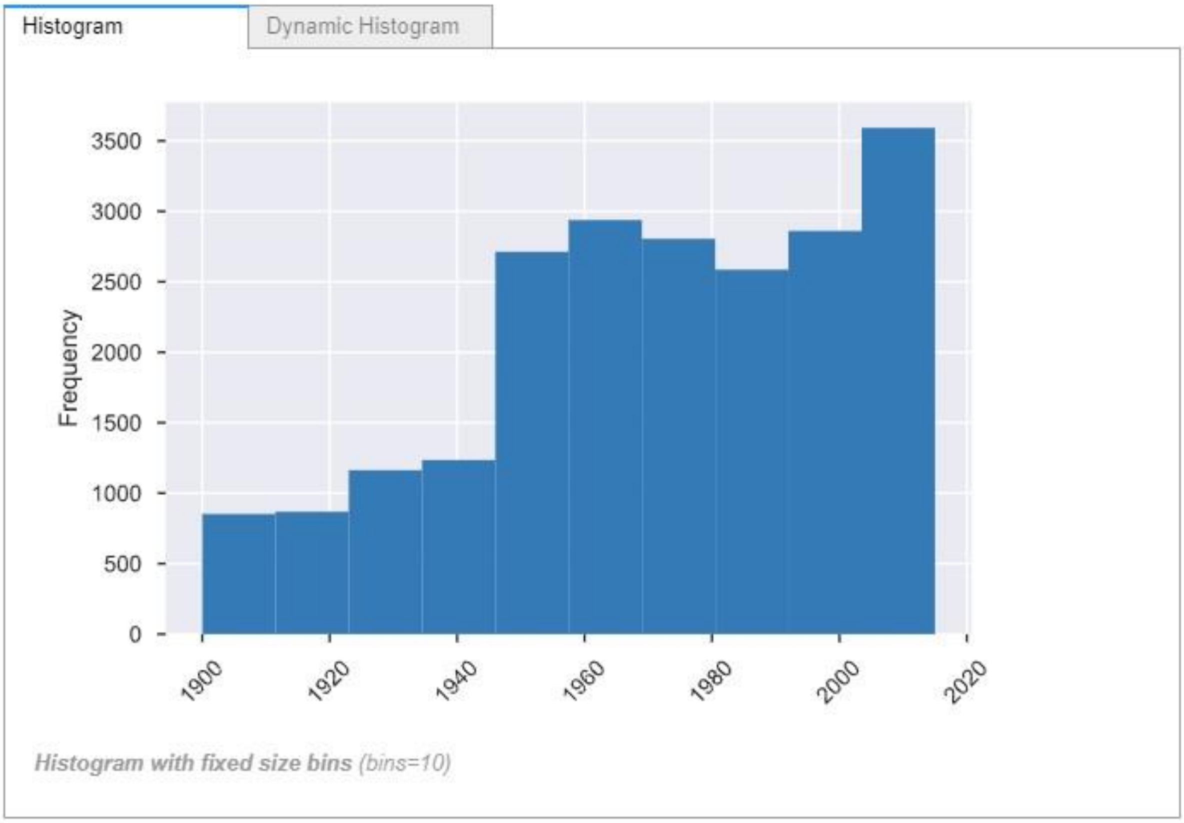


Figure 2-27 Distribution of yr_built.

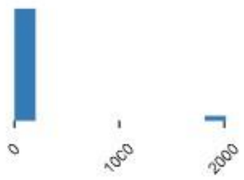
The data is skewed to the left with more no if houses being built between 1945 to 2015.

2.1.16 Yr_Renovated

The year in which the house was renovated. It has 70 distinct values and contains 20699 zeros. The zeros indicate that 95.77% of the houses are not renovated.

yr_renovated
Real number ($\mathbb{R}_{\geq 0}$)
ZEROS

Distinct count	70	Mean	84.4022579
Unique (%)	0.00323879146809790	Minimum	0
Missing	0	Maximum	2015
Missing (%)	0.0	Zeros	20699
Infinite	0	Zeros (%)	0.9577106371
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	0	Standard deviation	401.67924
5-th percentile	0	Coefficient of variation (CV)	4.759105384
Q1	0	Kurtosis	18.701152117599058
median	0	Mean	84.4022579
Q3	0	Median Absolute Deviation (MAD)	161.6658804
95-th percentile	0	Skewness	4.5494933669939295
Maximum	2015	Sum	1824186
Range	2015	Variance	161346.2119
Interquartile range (IQR)	0		

Figure 2-28 Statistics of yr_renovated.

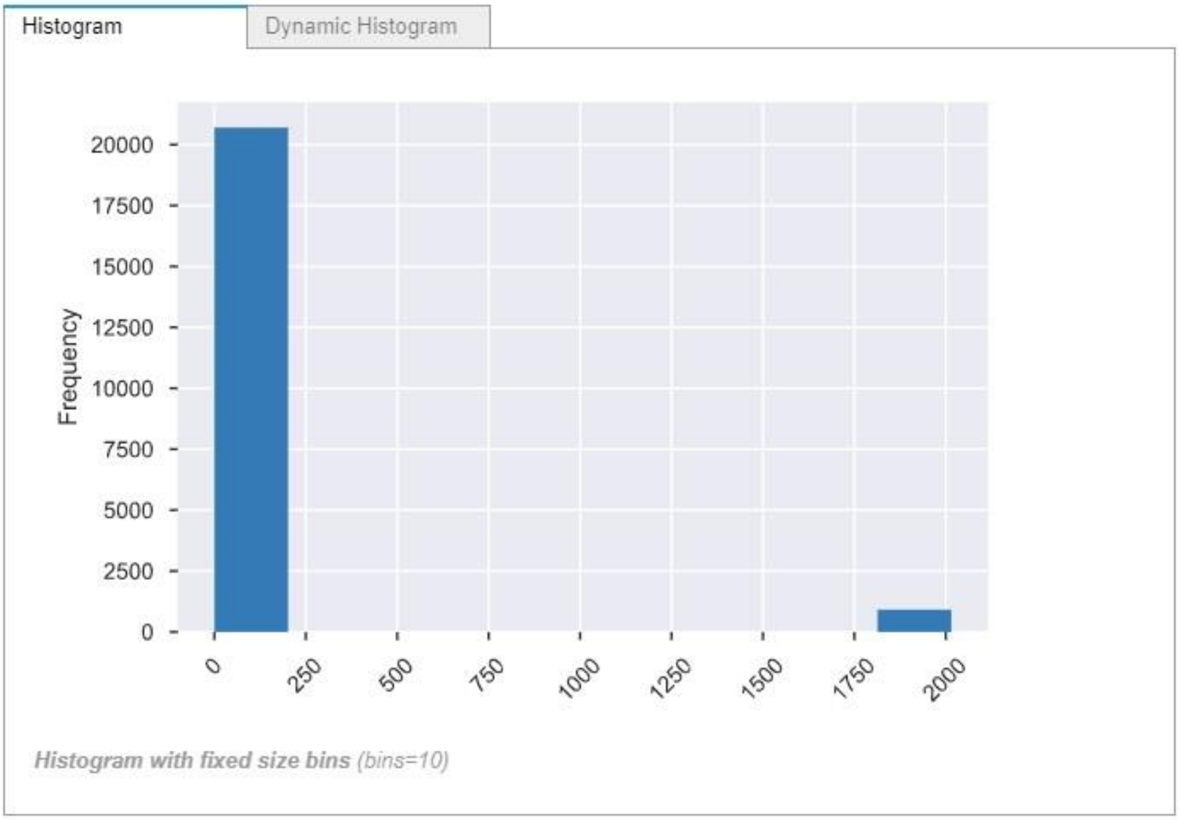


Figure 2-29 Distribution of yr_renovated.

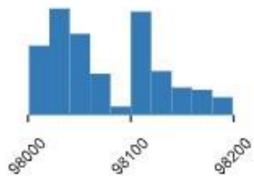
Due to the presence of high no of zeros the data is highly skewed to the left.

2.1.17 Zipcode

This feature contains the zipcode in which the house is located. It has no zeros and missing values.

zipcode
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	70	Mean	98077.9398
Unique (%)	0.0032387914680979	Minimum	98001
Missing	0	Maximum	98199
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	98001	Standard deviation	53.50502626
5-th percentile	98004	Coefficient of variation (CV)	0.0005455357888
Q1	98033	Kurtosis	-0.8534788732101246
median	98065	Mean	98077.9398
Q3	98118	Median Absolute Deviation (MAD)	46.72127898
95-th percentile	98177	Skewness	0.40566120823966473
Maximum	98199	Sum	2119758513
Range	198	Variance	2862.787835
Interquartile range (IQR)	85		

Figure 2-30 Statistics of zipcode.

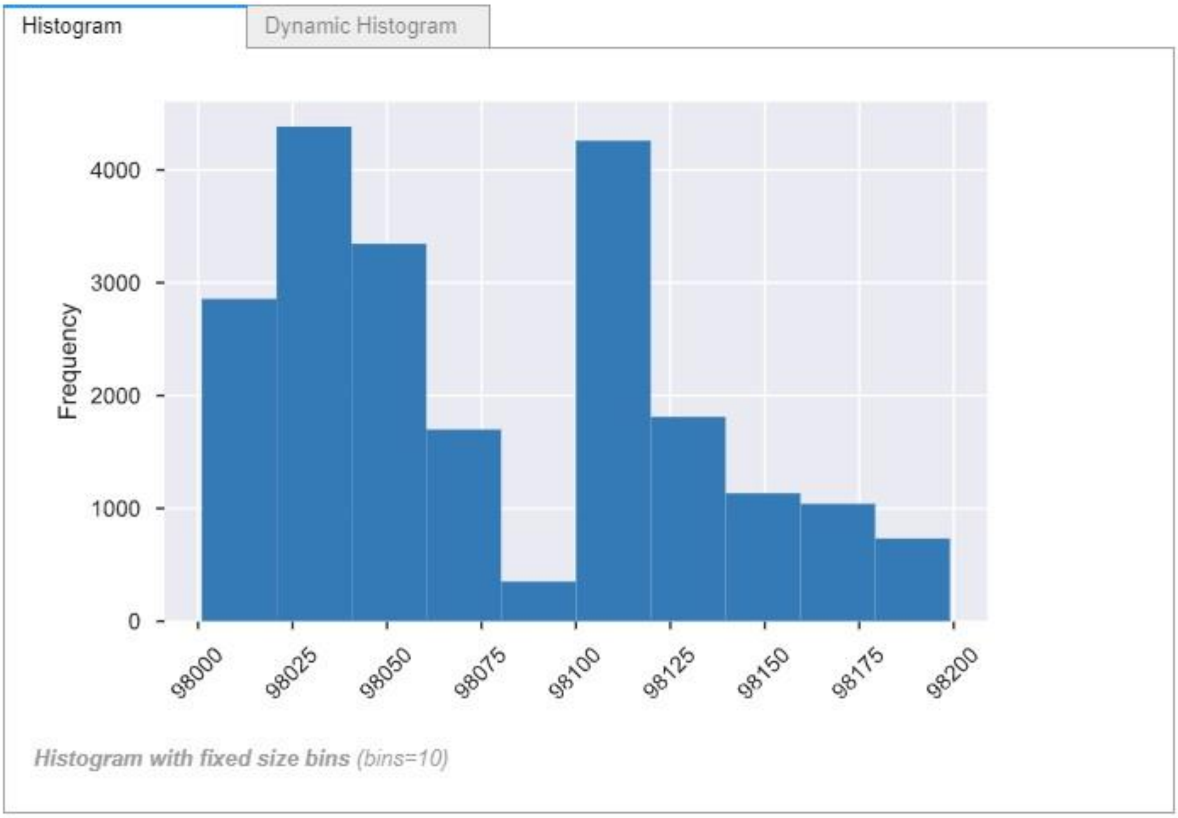


Figure 2-31 Distribution of Zipcode.

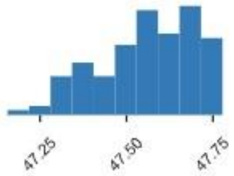
The data is skewed to the left and is having 2 peaks indicating the presence of more no of houses in the area.

2.1.18 Lat

Latitude in which the house is located. It is having 5034 distinct values.

lat
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	5034	Mean	47.56005252
Unique (%)	0.23291537500578355	Minimum	47.1559
Missing	0	Maximum	47.7776
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	47.1559	Standard deviation	0.1385637102
5-th percentile	47.3103	Coefficient of variation (CV)	0.002913447377
Q1	47.471	Kurtosis	-0.6763130016065335
median	47.5718	Mean	47.56005252
Q3	47.678	Median Absolute Deviation (MAD)	0.1148297137
95-th percentile	47.74964	Skewness	-0.48527047653808614
Maximum	47.7776	Sum	1027915.4151000001
Range	0.6217	Variance	0.0191999018
Interquartile range (IQR)	0.207		

Figure 2-32 Statistics of Latitude.

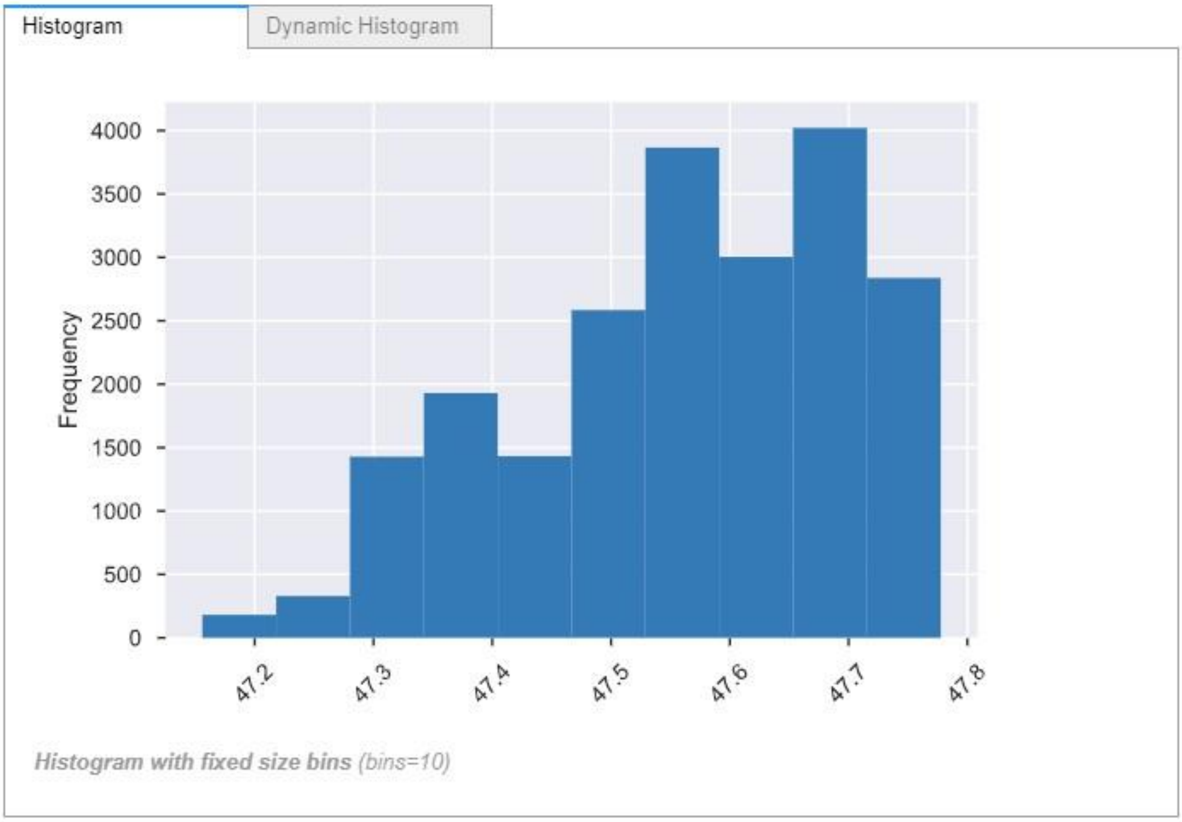


Figure 2-33 Distribution of lat.

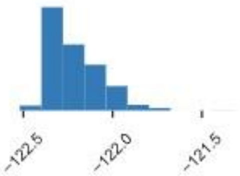
The data is slightly skewed to the right.

2.1.19 Long.

Longitude in which the house is located. It contains negative values.

long
Real number (ℝ)

Distinct count	752	Mean	-122.2138964
Unique (%)	0.03479387405728034	Minimum	-122.519
Missing	0	Maximum	-121.315
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	-122.519	Standard deviation	0.1408283424
5-th percentile	-122.387	Coefficient of variation (CV)	-0.001152310388
Q1	-122.328	Kurtosis	1.0495008872914617
median	-122.23	Mean	-122.2138964
Q3	-122.125	Median Absolute Deviation (MAD)	0.1151608925
95-th percentile	-121.979	Skewness	0.8850529834328087
Maximum	-121.315	Sum	-2641408.943
Range	1.204	Variance	0.01983262202
Interquartile range (IQR)	0.203		

Figure 2-34 Statistics of long.

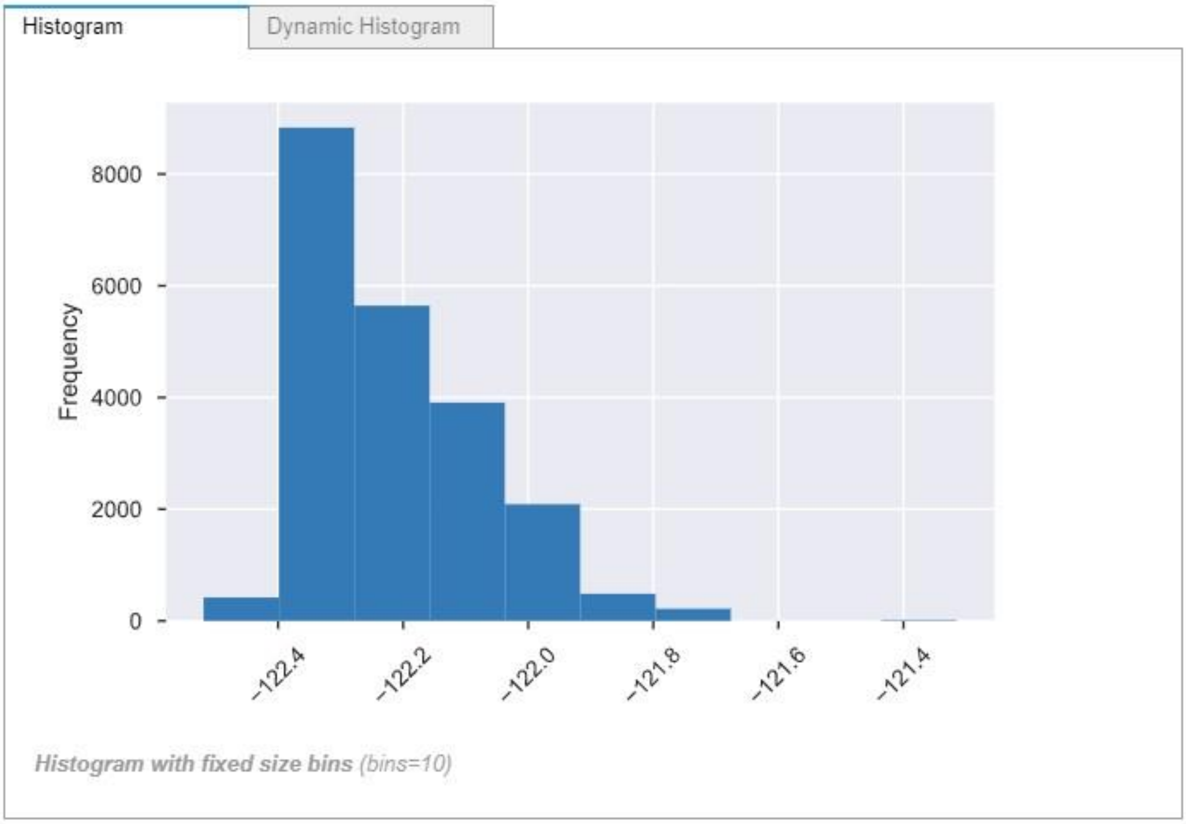


Figure 2-35 Distribution of Long.

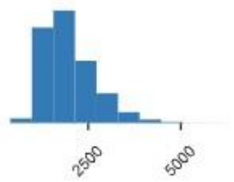
The data is skewed to the left.

2.1.20 Living_measure15

Updated living room area in year 2015(implies-- some renovations) This might or might not have affected the lotsize area.

living_measure15
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	777	Mean	1986.552492
Unique (%)	0.03595058529588673	Minimum	399
Missing	0	Maximum	6210
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	399	Standard deviation	685.3913043
5-th percentile	1140	Coefficient of variation (CV)	0.3450154512
Q1	1490	Kurtosis	1.5970958104616884
median	1840	Mean	1986.552492
Q3	2360	Median Absolute Deviation (MAD)	536.2192073
95-th percentile	3300	Skewness	1.1081812758966965
Maximum	6210	Sum	42935359
Range	5811	Variance	469761.2399
Interquartile range (IQR)	870		

Figure 2-36 Statistics of living_room15.

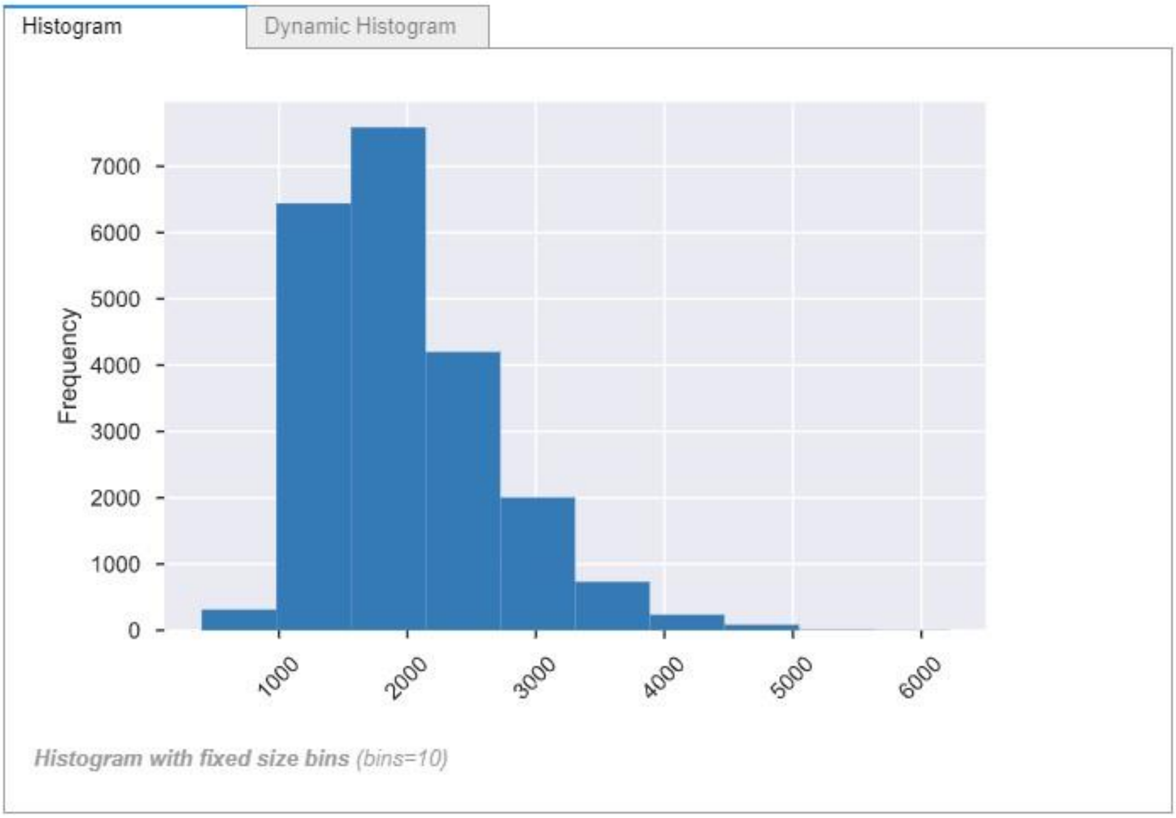


Figure 2-37 Distribution of living_room15.

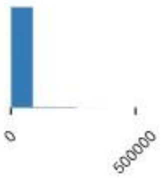
The data is skewed to the left with skewness of 1.108.

2.1.21 Lot_measure.

Updated lot area in year 2015(implies-- some renovations). It has no zeros or missing values.

lot_measure15
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	8689	Mean	12768.45565
Unique (%)	0.4020265580900384	Minimum	651
Missing	0	Maximum	871200
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	651	Standard deviation	27304.17963
5-th percentile	1999.2	Coefficient of variation (CV)	2.138408933
Q1	5100	Kurtosis	150.76311004626973
median	7620	Mean	12768.45565
Q3	10083	Median Absolute Deviation (MAD)	10118.66071
95-th percentile	37062.8	Skewness	9.506743246764398
Maximum	871200	Sum	275964632
Range	870549	Variance	745518225.3
Interquartile range (IQR)	4983		

Figure 2-38 Statistics of lot_measure15.

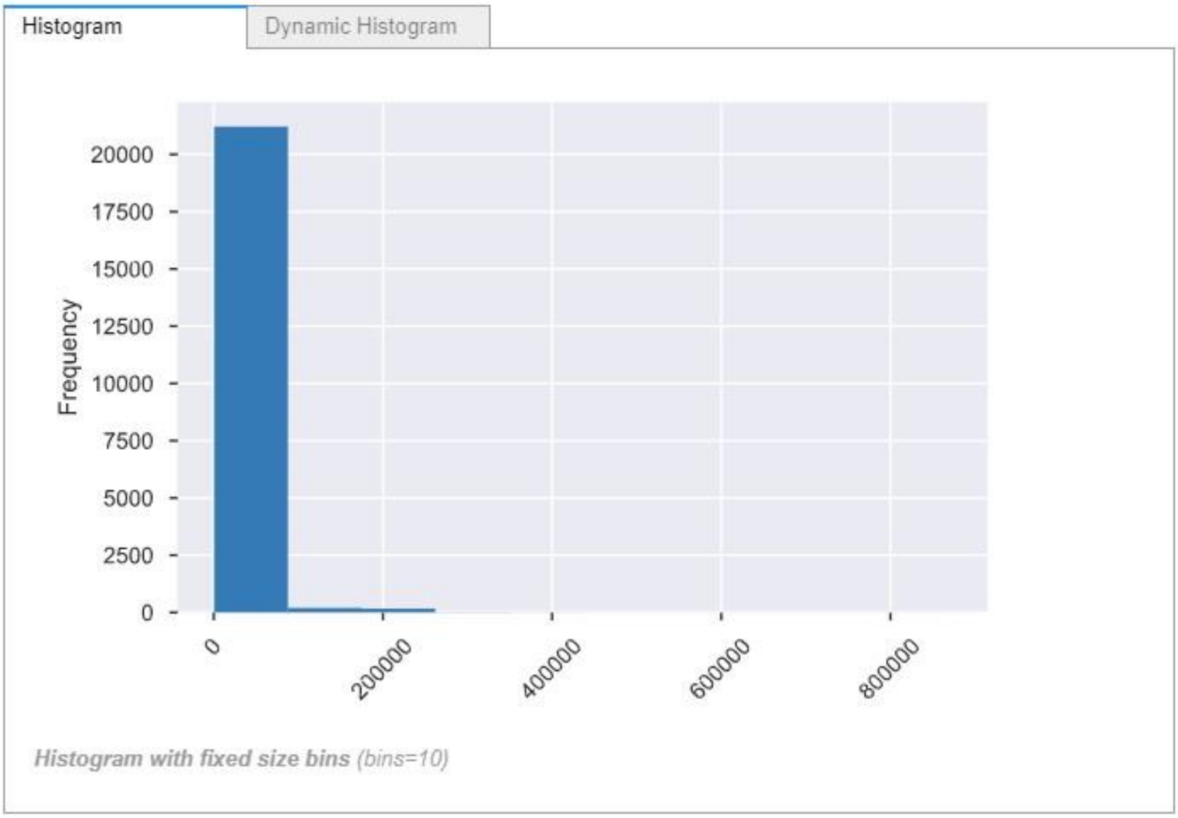


Figure 2-39 Distribution of lot_measure15.

The data is highly skewed to the left.

2.1.22 Furnished

This feature contains information about whether the house is furnished or not. It is having Boolean type data. 0 – indicating not furnished and 1 – indicating furnished. 17362 houses are not furnished and 4251 are furnished.

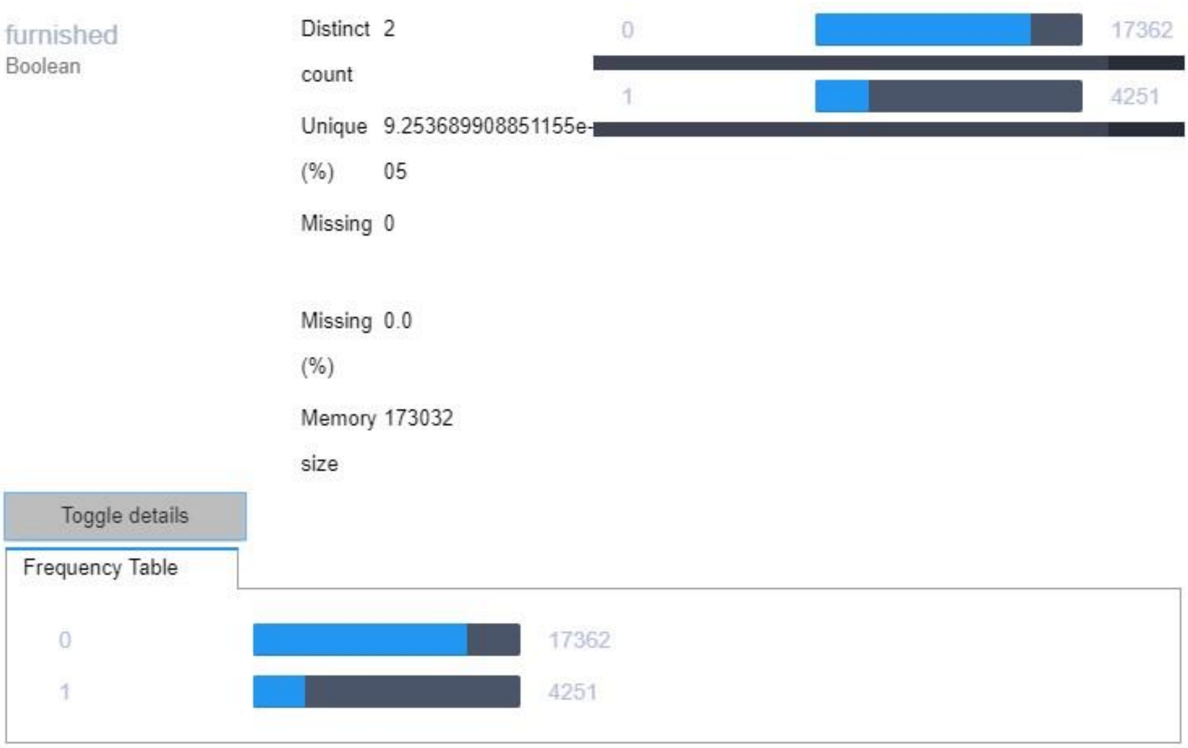


Figure 2-40 Statistics of furnished.

2.1.23 Total_area.

It is the total area in which the house is built and is a sum of living and lot measure. It is highly correlated to lot_measure.

total_area
Real number ($\mathbb{R}_{\geq 0}$)
HIGH CORRELATION

Distinct count	11163	Mean	17186.8673
Unique (%)	0.5164947022625271	Minimum	1423
Missing	0	Maximum	1652659
Missing (%)	0.0	Zeros	0
Infinite	0	Zeros (%)	0
Infinite (%)	0.0	Memory size	173032

Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Minimum	1423	Standard deviation	41589.08122
5-th percentile	3302.2	Coefficient of variation (CV)	2.419817439
Q1	7035	Kurtosis	281.21633093208465
median	9575	Mean	17186.8673
Q3	13000	Median Absolute Deviation (MAD)	14044.49282
95-th percentile	45885.8	Skewness	12.95606682384209
Maximum	1652659	Sum	371459763
Range	1651236	Variance	1729651676
Interquartile range (IQR)	5965		

Figure 2-41 Statistics of total_area.

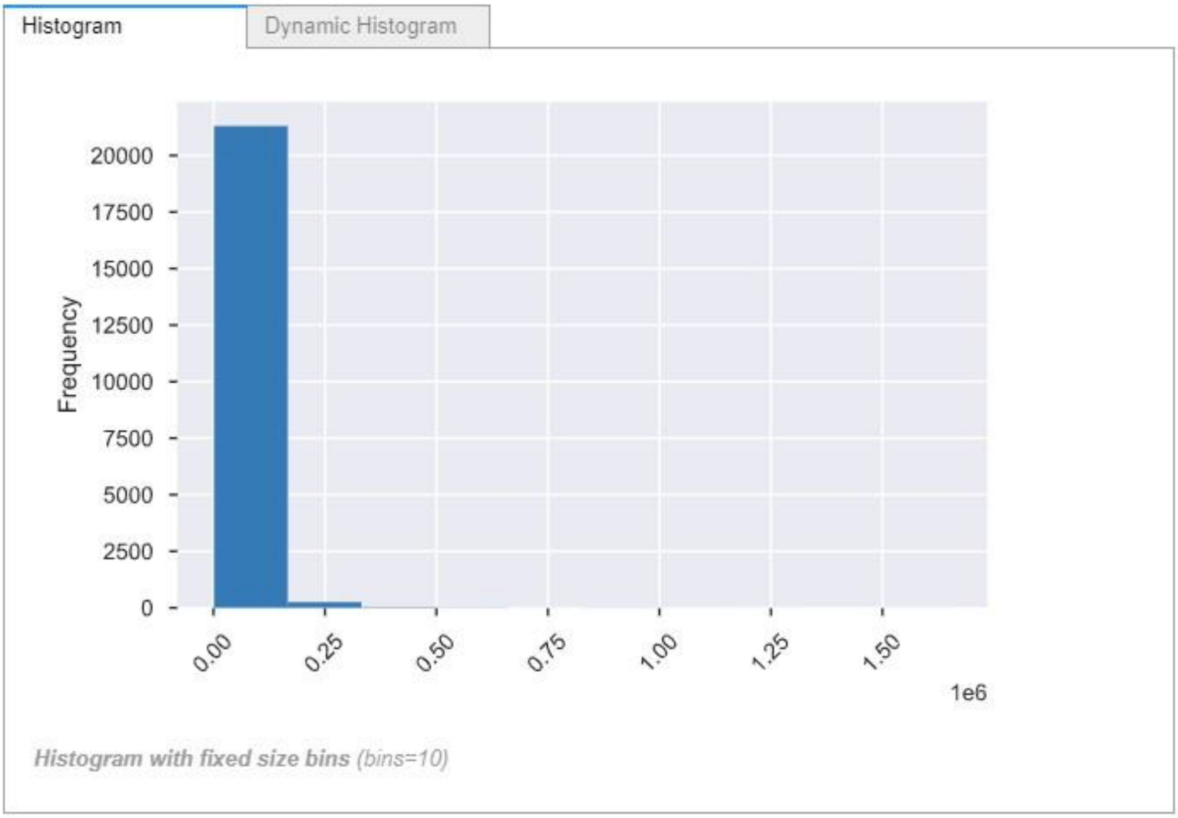


Figure 2-42 Distribution of total_area.

The data is highly skewed to the left.

2.1.24 Summary of Univariate Analysis.

From distribution plot of all the individual features it is observed that features like dayhours, room_bed, room_bath, living_measure,lot_measure, ceil, coast,sight, condition, quality, ceil_measure, basement,yr_built, yr_renovated, furnished, total_area follows skewed distribution with the dispersed values.

Skewness in the features with having certain discrete values like room_bed, room_bath, sight etc … increases bias in the model prediction. This is because values with majority will influence the outcome. For example, number of houses with bedrooms 3 and 5 are much more than number of houses with bedrooms more than 5 and hence 5-bedroom house price will influence the 10-bedroom house price drastically.

2.2 Multivariate Analysis.

2.2.1 Using Pairplot

From the pairplot it is observed that features like ceil_measure, living_measure, room_bath, lot_meausre, total_area are having some linear relationship with the price and among them.

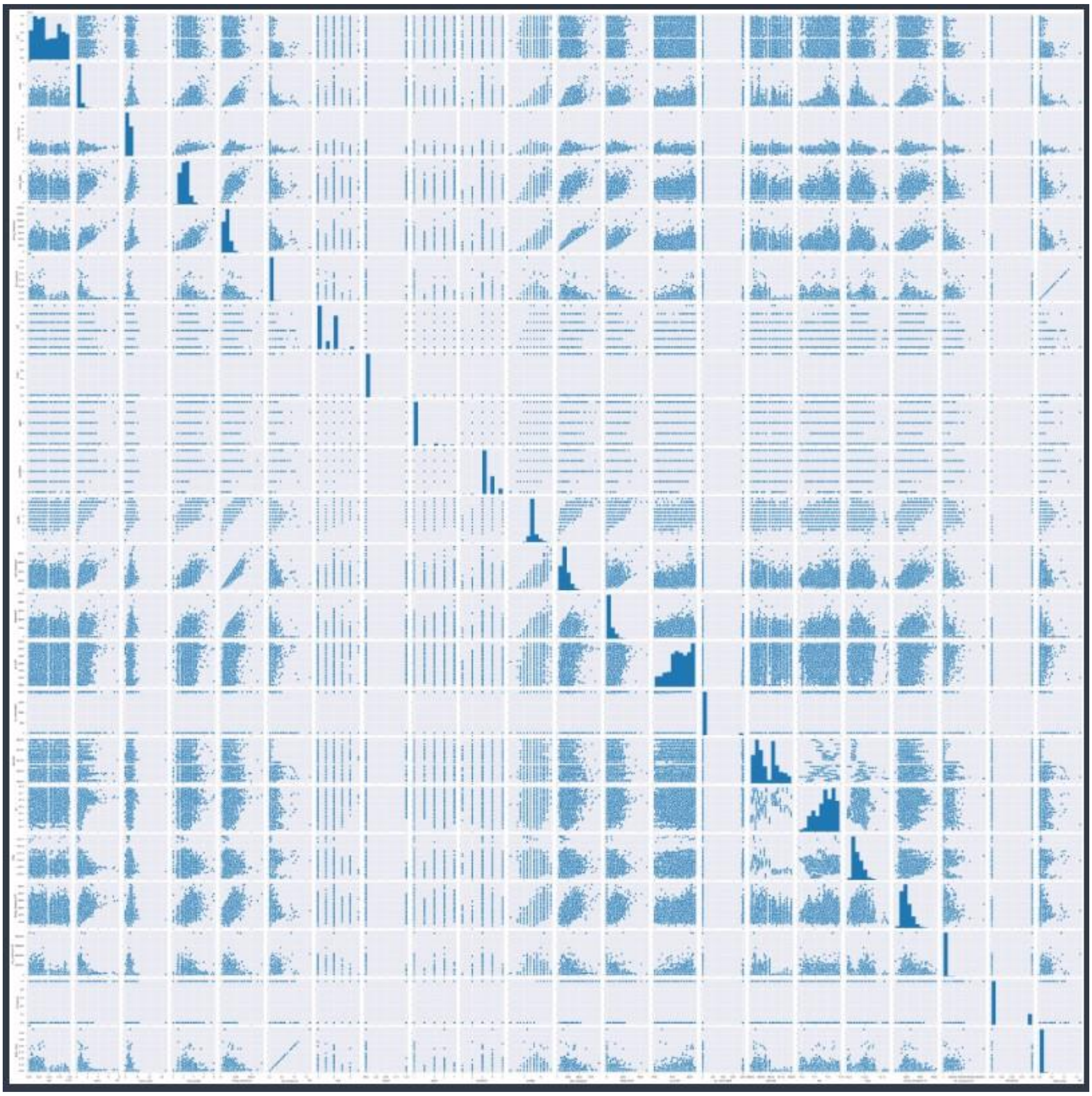


Figure 2-43 Pairplot.

2.2.2 Using Pearson Correlation.

Creating pearson's correlation matrix to find out exact correlation between them and drop/merge features which are highly correlated with other features to avoid multicollinearity issue.

Based on correlation matrix living_measure and ceil_measure highly correlated (.87)

Correlation between lot_measure and total_area is (.99).

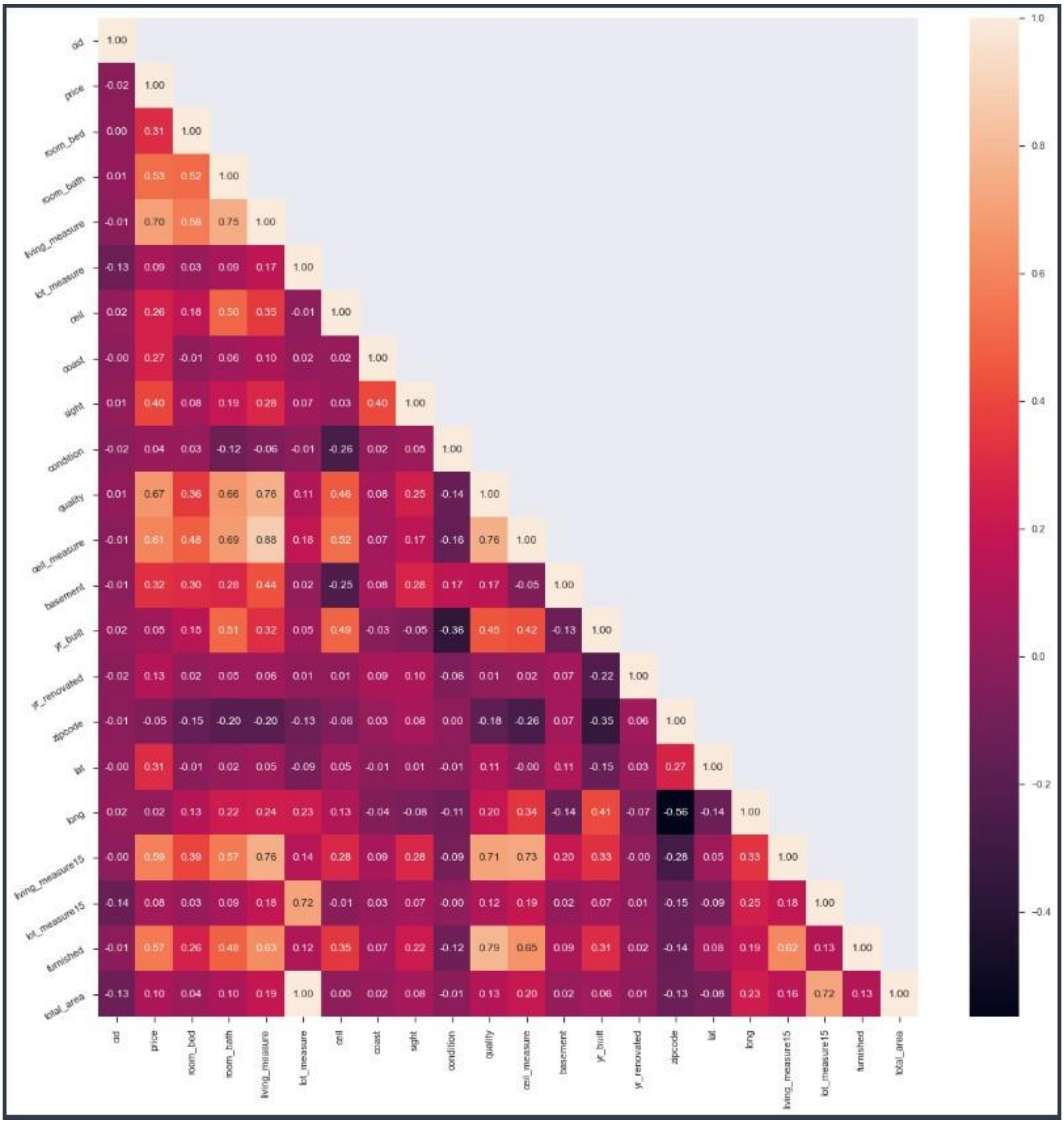


Figure 2-44 Correlation Heatmap.

As the correlation heatmap is quite dense and the no of variables are more it is difficult to isolate the most important features. Therefore, further creating an isolated heatmap of the most important features.

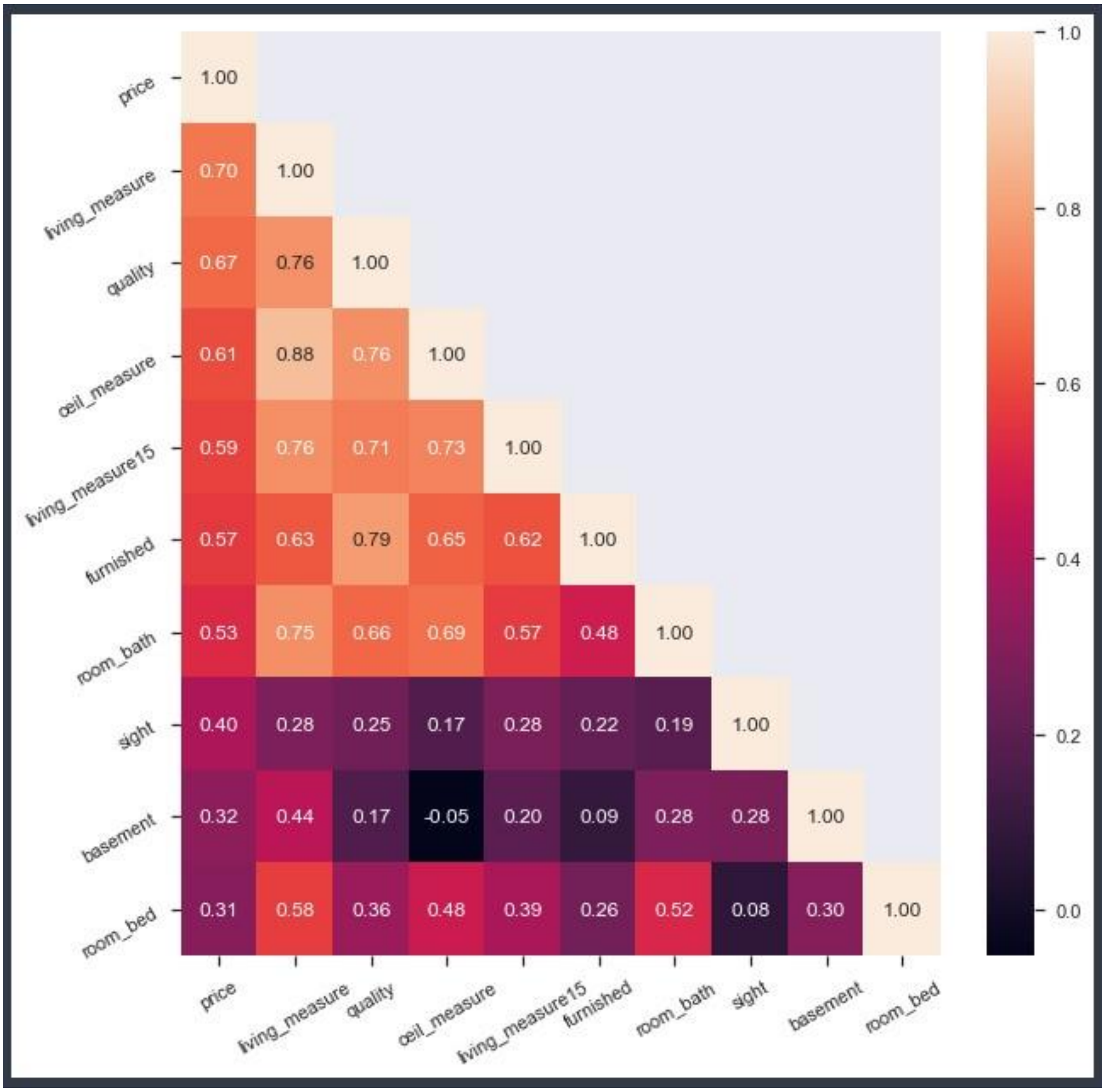


Figure 2-45 Correlation Heatmap of the 10 most important features.

From the above plot the 10 most important features are: -

```
Most Correlated Features
0        price
1        living_measure
2        quality
3        ceil_measure
4        living_measure15
5        furnished
6        room_bath
7        sight
8        basement
9        room_bed
```

2.3 Feature Selection.

From the pairplot it is observed that cid follows mostly uniform distribution and it does not help much on the prediction of the price. So, dropping cid.

Based on correlation heatmap dropping ceil_measure and lot_measure.

2.4 Feature Transformation.

As indicated in univariate analysis lot of features are having skewed dispersed distribution.

Log transformation helps to handle such features while maintaining fair distance between different values of the features.

```
df_log = np.log(df.drop(['price','lat','long'],axis=1)+1)
features = pd.concat([df_log,df[['lat','long']]],axis=1)
Adding +1 in each feature to have non-zero values since ln0 is not defined
```

Zipcode does not have any significance as a numerical value so treating it as a categorical variable and doing label encoding for the same

```
df_zip = pd.get_dummies(df['zipcode'])
#Dropping the zip code and appending corresponding one hot data
df.drop('zipcode',axis=1,inplace=True)
```

As non-linear relationships seen among the different features and between some features and target variable doing polynomial transformation

```
#Adding polynomial features with degree 3

from sklearn.preprocessing import PolynomialFeatures

poly = PolynomialFeatures(3)

features_t = poly.fit_transform(features)
```

From different experiments it is observed that polynomial transform with degree 3 gives best model performance

Finally doing Power transformation (yono-johson) to make the features more Gaussian Like.

```
#Applying tranformation to the features, to make the distribution Gaussian like

from sklearn.preprocessing import PowerTransformer

pt = PowerTransformer()

df_t = pt.fit_transform(features_t)
```

2.5 Feature Scaling.

Scaling is important to make the features unit independent and to avoid the influence of higher values because of asymmetry in the units of different features .

Using MinMax scaling in this case because StandardScaler will not give good performance since the data is already transformed using log.

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

df_m = scaler.fit_transform(df_m)
```

2.6 Impact of Outliers.

Since the features are log transformed followed by MixMaxScaler effect of outliers is almost nullified and there are no significant outliers in the data.

2.7 Principal Component Analysis

After polynomial transformation with degree 3 and label encoding of the feature zipcode total number of features have increased to 1400. To avoid curse of dimensionality performing PCA to only select principal components.

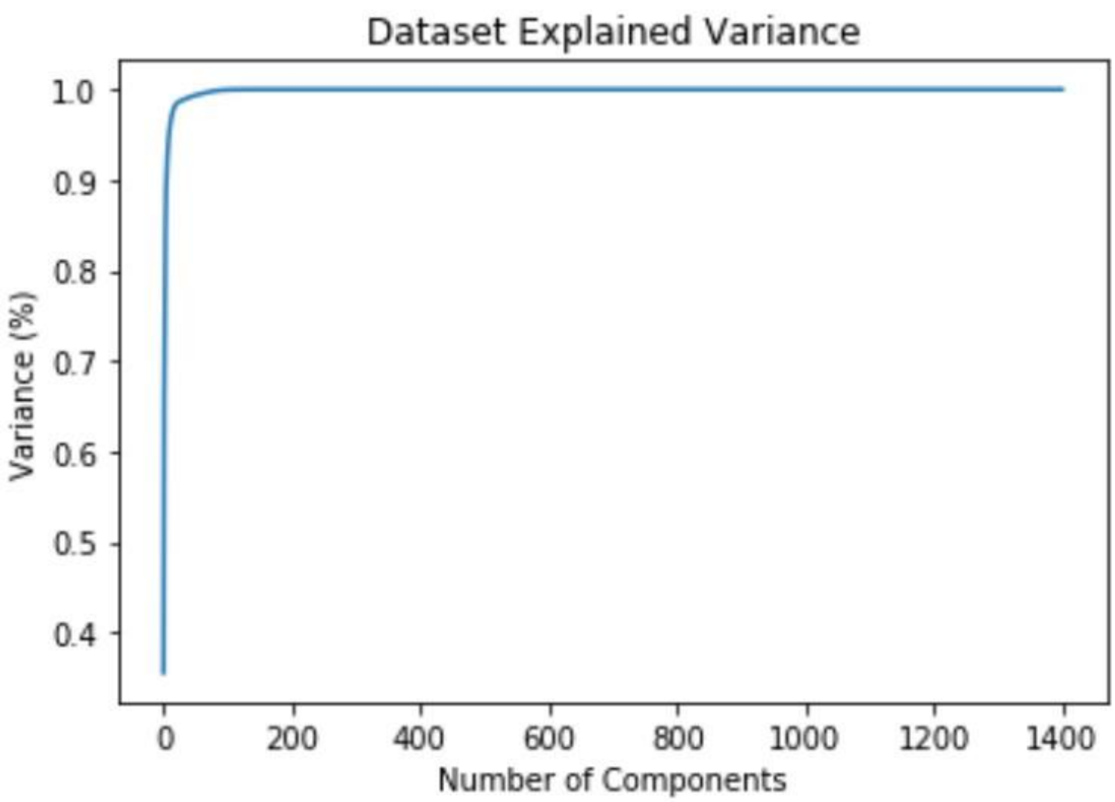


Figure 2-46 Principal component analysis graph with explained variance.

```
print(np.sum(pca.explained_variance_ratio_[0:300]))

0.9999999820403399
```

From the graph it is seen that elbow is around 70 and 99.99% of the variance is explained by 300 features.

3 Model Explore

3.1 Gradient Boosting Regression

Step 1 : After data insight find correlation between the attributes. Correlation functions fall within the range [-1, 1].

```
df1.corr()
```

Step 2 : Visualize the data Set using different descriptive statistics such Box Plot for univariate variable and Scatter plot is used to understand relationship between two different attributes in the dataset. We have compared PRICE (target) vs each of the attribute in the dataset.

Step 3 : Training Regression Model, we tried out different Regression models available in scikitlearn with a k-fold cross validation method. standardize the dataset using StandardScaler function in scikit-learn. This is a useful technique where the attributes are transformed to a standard gaussian distribution with a mean of 0 and a standard deviation of 1.

```
from sklearn.preprocessing import StandardScaler, MinMaxScaler
scaler = MinMaxScaler().fit(X)
scaled_X = scaler.transform(X)
```

Step 4 : Regression Model Comparison

```
# hold different regression models in a single dictionary
models = {}
models["Linear"]        = LinearRegression()
models["Lasso"]         = Lasso()
models["Ridge"]         = Ridge()
models["ElasticNet"]    = ElasticNet()
models["DecisionTree"]  = DecisionTreeRegressor()
models["KNN"]           = KNeighborsRegressor()
models["RandomForest"]  = RandomForestRegressor()
models["AdaBoost"]      = AdaBoostRegressor()
models["GradientBoost"] = GradientBoostingRegressor()
models["XGBoost"] = XGBRegressor()
```

Result

```
Linear: -142129.093, 17540.19
Lasso: -142353.936, 17588.117
Ridge: -132353.057, 14841.722
ElasticNet: -192229.726, 27398.837
DecisionTree: -133850.991, 26396.536
KNN: -125846.724, 24907.204
RandomForest: -94514.224, 20327.081
AdaBoost: -164224.782, 15037.949
GradientBoost: -89958.789, 17725.236

XGBoost: -96634.708, 25801.697
```

Step 4 : Based on the above comparison, we choose Gradient Boosting Regression model outperforms all the other regression models. So, we will choose it as the best Regression Model for this problem.

MAPE Calculation

MAPE - Mean Absolute Percentage Error (TRAIN DATA): 10.190201396000507

MAPE - Mean Absolute Percentage Error (TEST DATA): 16.831630389981463

3.2 Summary

After exploring different algorithms, after exploring relation between attributes we find out Grid Search is one of the algorithms which is giving higher accuracy among all the technique which we tried. In Detailed description is below.

4 Model Selection

Housing price prediction is a regression problem. Since features are transformed using polynomial transformation to avoid the high variance in the output, model with the regularization provides better performance.

Using the Ridge regression with default regularization

4.1 Ridge

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(df_m,df['price'], random_state =
1,test_size= 0.2)

sklearn.linear_model import Ridge

lr = Ridge()

lr.fit(X_train,y_train) lr.score(X_test,y_test)
```

With this R2 score is 83%

4.2 Lasso

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(df_m,df['price'], random_state =
1,test_size= 0.2)

from sklearn.linear_model import Ridge

lr = Lasso()

lr.fit(X_train,y_train)

lr.score(X_test,y_test)
```

With this R2 score is 83.94%

Below chapter contains model tuning for best performance.

5 Model Tuning and Evaluation.

After doing hyper parameter tuning using GridSearch it is observed that alpha = 1e-5 is giving best performance which is ~87%

Below are the results of ridge and lasso regressions with principal components

```
In [125]: from sklearn.model_selection import train_test_split
          X_train, X_test, y_train, y_test = train_test_split(PCA(300).fit_transform(df_m),df['price'], random_state = 1,test_size= 0.3)

In [111]: lasso_params = {'alpha':list(np.logspace(-10,0,11))}
          ridge_params = {'alpha':list(np.logspace(-10,0,11))}

          from sklearn.linear_model import Ridge
          from sklearn.linear_model import Lasso

In [126]: from sklearn.model_selection import GridSearchCV

          gs_lasso = GridSearchCV(Lasso(),param_grid=lasso_params)

          gs_lasso.fit(X_train,y_train)
          gs_lasso.score(X_test,y_test)

Out[126]: 0.8657814445247067

In [127]: from sklearn.model_selection import GridSearchCV

          gs_ridge = GridSearchCV(Ridge(),param_grid=ridge_params)

          gs_ridge.fit(X_train,y_train)
          gs_ridge.score(X_test,y_test)

Out[127]: 0.8672161974720665
```

6 Implications, Limitation and Closing Reflections.

Now, we use these results to discuss whether the constructed model should or should not be used in a real-world setting. Some questions that are worth to answer are:

- *How relevant today is data that was collected from 1900? How important is inflation?*

Data collected from 1900 is not of much value in today's world. Society and economics have changed so much and inflation has made a great impact on the prices.

- *Are the features present in the data sufficient to describe a home? Do you think factors like quality of appliances in the home, square feet of the plot area, presence of basement or not etc should factor in?*

The dataset considered is quite limited, there are a lot of features, like the size of the house in square feet, the presence of basement or not, and others, that are very relevant when considering a house price.

- *Is the model robust enough to make consistent predictions?*

Given the high variance on the price range, we can assure that it is not a robust model and, therefore, not appropriate for making predictions.

- *Is it fair to judge the price of an individual home based on the characteristics of the entire area/location?*

In general, it is not fair to estimate or predict the price of an induvial home based on the features of the entire area like zip, lat and long. In the same area there can be huge differences in prices.