CAPSTONE PROJECT(NOTE1)

BY- KOSHY MATHEW
February 27, 2022
Batch - March C 21

Table of contents :-

List of Tables

List of Figures

1) Introduction of the business problem:

a) Defining problem statement:-

A house value is simply more than location and square footage. Like the features that make up a person, an educated party would want to know all aspects that give a house its value. For example, you want to sell a house and you don't know the price which you may expect — it can't be too low or too high. To find house price you usually try to find similar properties in your neighbourhood and based on gathered data you will try to assess your house price

Objective – To create multi linear regression model to predict prices of houses in Seattle city and also of neighbouring areas in Washington D.C using given variables which will be described in later .

b) Need of the study/project

In USA the real estate market growing in high pace , if we now specifically speak of Seattle city , The Seattle housing market was red-hot last year. Last year's record-breaking sales occurred despite record-low inventory levels. No month had a supply greater than a month. By and large, industry analysts define a balanced market as having an inventory of four to six months. The Seattle area home prices continue to rise beyond the reach of many buyers. The median home sold for $828,111 in King County, up 14.2 percent from 2020.

Since Feb 2012, the home values in the city of Seattle have appreciated by nearly 153.5% — Zillow Home Value Index. The home values increased consistently, starting in late 2012 and continuing through 2018. After that, it marked the beginning of a sustained downturn in prices which lasted for over a year. In 2018, prices took a steep drop. From July 2018 onward the home values started declining and they continued so until November of 2019. The trajectory has shifted from last Oct 2019 to an upward trend.
(https://www.noradarealestate.com/blog/seattle-real-estate-market/)

The formulation of business problem revolves around four aspects - Objective, Scope, Constraints & Significance of this project. What's the achievement out of this project is concerned with objective & significance of the project.

c. Understanding business/social opportunity

this project will be beneficial for the stake holders in real estate business. Following are the same

1. **buyers** – who looks after best prices

2. **sellers** – who wants to keep the property prices in check or balance based upon the situations

3.**builders** – looking for best features and locations that will fetch best prices

4.**real estate agents** – third party people who are looking for best commission indirectly affected by prices

5**. mortgage lenders** – like banks and loan lenders whose interest or profits indirectly affected by the prices of the houses and its involved factors

Business opportunity – so we are basically going to study the trend of prices of the houses in Seattle and its neighbouring areas during the period 2014-2015

2)Data Report

a) Understanding how data was collected in terms of time, frequency and methodology

data was collected for houses with various feature and studying the trend of the same , so prices or features are taken of period 2014 -2015 of Seattle city area and its neighbouring places

b) Visual inspection of data (rows, columns, descriptive details

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
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          21505 non-null  float64
 4   room_bath         21505 non-null  float64
 5   living_measure    21596 non-null  float64
 6   lot_measure       21571 non-null  float64
 7   ceil              21571 non-null  object
 8   coast             21612 non-null  object
 9   sight             21556 non-null  float64
 10  condition         21556 non-null  object
 11  quality           21612 non-null  float64
 12  ceil_measure      21612 non-null  float64
 13  basement          21612 non-null  float64
 14  yr_built          21612 non-null  object
 15  yr_renovated      21613 non-null  int64
 16  zipcode           21613 non-null  int64
 17  lat               21613 non-null  float64
 18  long              21613 non-null  object
 19  living_measure15  21447 non-null  float64
 20  lot_measure15     21584 non-null  float64
 21  furnished         21584 non-null  float64
 22  total_area        21584 non-null  object
dtypes: float64(12), int64(4), object(7)
memory usage: 3.8+ MB
```

Table 1.1 (info of the data)

1. There are 21613 enteries
2. there are 23 variables out of which 16 are float or integer and 7 are shown as object
3. it can be observed that there are null values as total non values in some variables does Match the total entries

Further we checked for duplicates which was 0

8 point summary :-

	count	mean	std	min	25%	50%	75%	max
cid	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	5.401822e+05	3.673622e+05	7.500000e+04	3.219500e+05	4.500000e+05	6.450000e+05	7.700000e+06
room_bed	21505.0	3.371355e+00	9.302886e-01	0.000000e+00	3.000000e+00	3.000000e+00	4.000000e+00	3.300000e+01
room_bath	21505.0	2.115171e+00	7.702481e-01	0.000000e+00	1.750000e+00	2.250000e+00	2.500000e+00	8.000000e+00
living_measure	21596.0	2.079861e+03	9.184961e+02	2.900000e+02	1.429250e+03	1.910000e+03	2.550000e+03	1.354000e+04
lot_measure	21571.0	1.510458e+04	4.142362e+04	5.200000e+02	5.040000e+03	7.618000e+03	1.068450e+04	1.651359e+06
sight	21556.0	2.343663e-01	7.664376e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	4.000000e+00
quality	21612.0	7.656857e+00	1.175484e+00	1.000000e+00	7.000000e+00	7.000000e+00	8.000000e+00	1.300000e+01
ceil_measure	21612.0	1.788367e+03	8.281025e+02	2.900000e+02	1.190000e+03	1.560000e+03	2.210000e+03	9.410000e+03
basement	21612.0	2.915225e+02	4.425808e+02	0.000000e+00	0.000000e+00	0.000000e+00	5.600000e+02	4.820000e+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
living_measure15	21447.0	1.987066e+03	6.855196e+02	3.990000e+02	1.490000e+03	1.840000e+03	2.360000e+03	6.210000e+03
lot_measure15	21584.0	1.276654e+04	2.728699e+04	6.510000e+02	5.100000e+03	7.620000e+03	1.008700e+04	8.712000e+05
furnished	21584.0	1.967198e-01	3.975279e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00

Table 1.2 (description of data frame)

So we can see firstly scales of variables are totally different which will create problems in modelling so had to be scaled on later stage before modelling

Further as part of 8 point summary - count , mean , median (50%) , standard deviation , min , max , 1st quartile(25%) , 3rd quartile (75%) can be seen for individual variables in Table 1.2

c) Understanding of attributes (variable info, renaming if required)

Let's individually understand the variables

1. cid – its basically unique id of houses
2. day hours – time and date when house was sold
3. price – cost of the house when sold
4. room_bed -Number of Bedrooms/House
5. room_bath- Number of bathrooms/bedrooms
6. iving_measure -square footage of the home
7. lot_measure -square footage of the lot
8. Ceil- Total floors /levels in house
9. Coast -House which has a view to a waterfront
10. Sight -Has been viewed
11. Condition -How good the condition is (Overall)
12. Quality - grade given to the housing unit, based on grading system
13. Ceil_measure - square footage of house apart from basement
14. Basement_measure - square footage of the basement
15. yr_built - Built Year
16. yr_renovated - Year when house was renovated
17. Zipcode - zip
18. Lat- Latitude coordinate
19. Long - Longitude coordinate
20. living_measure15 - Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area
21. lot_measure15 - lotSize area in 2015(implies-- some renovations)
22. Furnished - Based on the quality of room
23. Total_area - Measure of both living and lot area

Cleaning and variable manipulation as per required :-

1. so we checked for any duplicates and it was zero

2. It was observed in the info certain variable which naturally should be of integer or float data type is showing object data type which are as follows :-

a) Ciel – its basically number of floors

b) Coast – its categorical but its labelled with number so should be integer type

c) Condition – again its categorical but labelled with number so should be integer or float type

d) Yr_built – this variable have years so should be float or int

e) Long – its basically longitude I.e geo spatial description with number so should be integer or float type

f) Total area – its living area + lot area so should be in integer or float

So to check the issue done value count on data frame with object data type , and result came out with common anomaly that's " $" sign in columns .

So then replaced those $ sign with 'nan' values for future imputations .now data type automatically changed to int/float after treatment

Following is the info of data frame with treated anomalies and subsequently changed data type :-

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
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          21505 non-null  float64
 4   room_bath         21505 non-null  float64
 5   living_measure    21596 non-null  float64
 6   lot_measure       21571 non-null  float64
 7   ceil              21541 non-null  float64
 8   coast             21582 non-null  float64
 9   sight             21556 non-null  float64
 10  condition         21528 non-null  float64
 11  quality           21612 non-null  float64
 12  ceil_measure      21612 non-null  float64
 13  basement          21612 non-null  float64
 14  yr_built          21598 non-null  float64
 15  yr_renovated      21613 non-null  int64
 16  zipcode           21613 non-null  int64
 17  lat               21613 non-null  float64
 18  long              21579 non-null  float64
 19  living_measure15  21447 non-null  float64
 20  lot_measure15     21584 non-null  float64
 21  furnished         21584 non-null  float64
 22  total_area        21545 non-null  float64
dtypes: float64(18), int64(4), object(1)
memory usage: 3.8+ MB
```

Table 1.3 (new table info with treated anomaly)

So we can see that all variables that had anamolies , after treatment has automatically changed to numerical data type

3) Exploratory data analysis:-

1. **Let's first see for meaning full changes in the variables which was observed visually**

   With respect to variables "yr_built" and "dayhours" we can calculate the age of the house by subtracting separated year from "dayhours" and subtracting it with "year_built".

age
49.0
67.0
48.0
5.0
91.0

Table 1.4(first 5 rows of age variable)

Now regarding year of renovations we can convert "yr_renovated" into binary class of whether the house was renovated or not . "0" class if not renovated "1" class if renovated

So if we do value count on this variable we can observe as follows :-

```
0    20699
1      914
Name: renovated, dtype: int64
```

There are 914 houses which are renovated and remaining 20699 are not renovated

2. Now let's see missing values scenario for further treatment

a. For numerical variables:-
Let's observed the percentage of missing values in each variable

```
price              0.000000
room_bed           0.499699
room_bath          0.499699
living_measure     0.078656
lot_measure        0.194327
ceil               0.333133
ceil_measure       0.004627
basement           0.004627
living_measure15   0.768056
lot_measure15      0.134179
total_area         0.314625
age                0.069403
dtype: float64
```

Table 1.5(percentage of missing values in numerical variables)

Now as we can see missing values are less than 1% in each variable , thus let's just simply impute with median value as it has less impact of outliers compared to mean .

So following was the data info after imputation

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   price            21613 non-null  float64
 1   room_bed         21613 non-null  float64
 2   room_bath        21613 non-null  float64
 3   living_measure   21613 non-null  float64
 4   lot_measure      21613 non-null  float64
 5   ceil             21613 non-null  float64
 6   ceil_measure     21613 non-null  float64
 7   basement         21613 non-null  float64
 8   living_measure15 21613 non-null  float64
 9   lot_measure15    21613 non-null  float64
 10  total_area       21613 non-null  float64
 11  age              21613 non-null  float64
dtypes: float64(12)
memory usage: 2.0 MB
```

Table 1.6(data frame info of numerical variable after imputation)

So now we can see there are no null values present and its properly imputed

b. For categorical variables

Let's see percentage of missing value in categorical variable

```
coast         0.143432
sight         0.263730
condition     0.393282
quality       0.004627
furnished     0.134179
renovated     0.000000
zipcode       0.000000
lat           0.000000
long          0.157313
dtype: float64
```

Table 1.7(percentage of missing values in categorical variables)

Again as we can see the percentage of missing value is way less than 1 % , so we can just impute the same with mode or the most frequent in case for categorical variables

so following is info of the categorical variable after imputation

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 9 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   coast      21613 non-null  float64
 1   sight      21613 non-null  float64
 2   condition  21613 non-null  float64
 3   quality    21613 non-null  float64
 4   furnished  21613 non-null  float64
 5   renovated  21613 non-null  float64
 6   zipcode    21613 non-null  float64
 7   lat        21613 non-null  float64
 8   long       21613 non-null  float64
dtypes: float64(9)
memory usage: 1.5 MB
```

Table 1.8(data frame info of categorical variable after imputation)

So now we can see there are no null values present and its properly imputed

3. Now lets see univariate analysis

a. Univariate analysis of numerical variables

i. Distribution plot

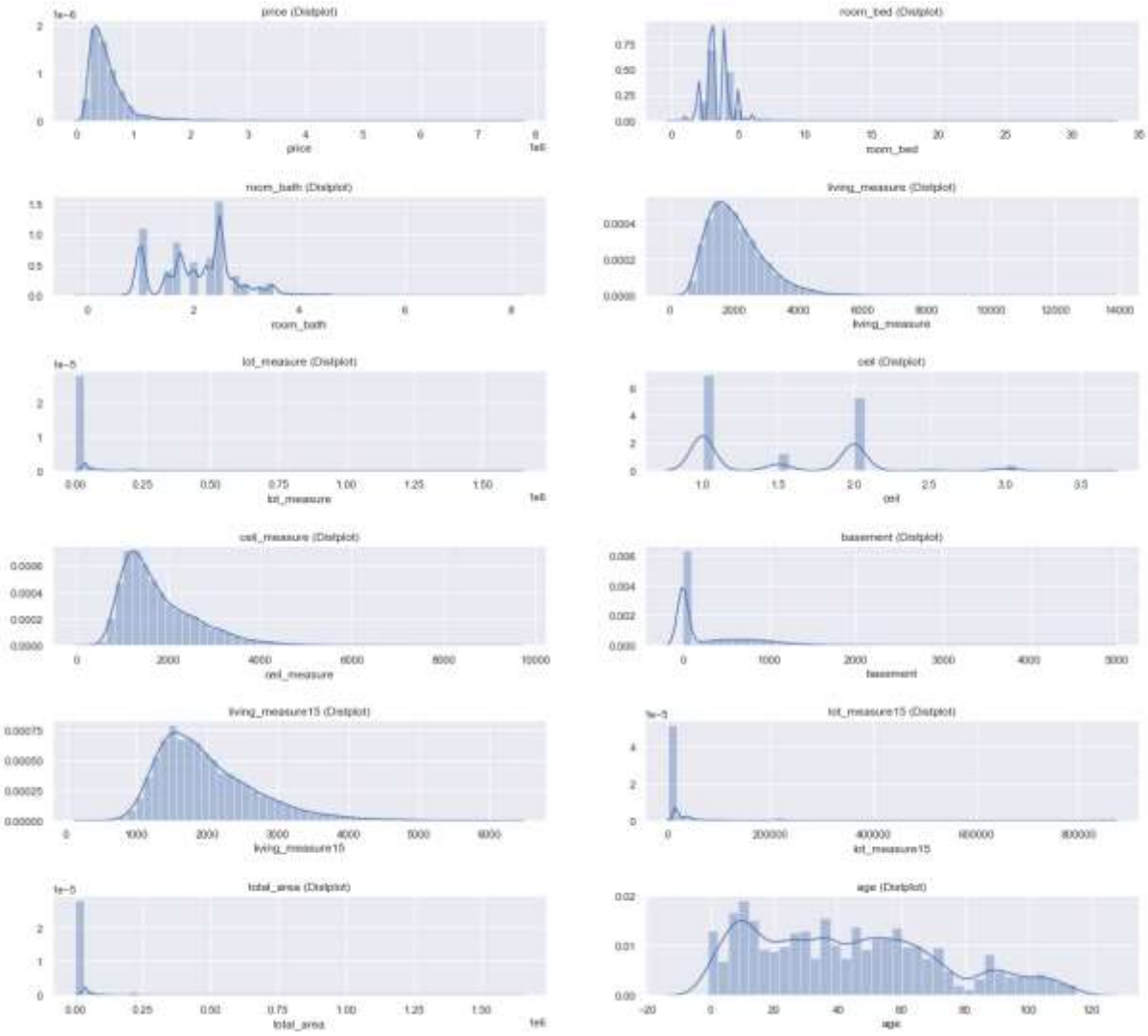

Fig 1.1(distribution plot)

1. we can see variables “price” ,”living measure” , “ciel measure” , “living measure15” , “lot_measure” , “lot_measure15” & “Total_measure” are right skewed so they might have outliers

2. variables “ceil” , “room_bed “ , room_bath” not follows continuous behaviours, they have multiple peaks so they basically are discrete continuous variables

3. we can see long tails for certain variables thus indicating high kurtosis

ii. Kurtosis & skewness

	skewness	kurtosis
price	4.021716	34.522444
room_bed	1.989024	49.448990
room_bath	0.510251	1.304428
living_measure	1.473517	5.253713
lot_measure	13.084880	286.036032
ceil	0.617403	-0.476333
ceil_measure	1.446810	3.402729
basement	1.577965	2.715574
living_measure15	1.116553	1.638682
lot_measure15	9.525543	151.424980
total_area	12.974601	281.866001
age	0.469843	-0.655809

Table 1.9 (skewness and kurtosis)

1. so we can see most of the variables are highly right skewed as values are more than 1 with exception to certain variables like “ceil” , “age” , “room_bath” but these variables showed properties of discreet continuous

2. now if we see kurtosis we can notice variables “Total_area” , ‘lot_measure” , ‘lot_measure15” showing very high positive values which means very long tail and thus high outliers followed by variables ‘price’ & ‘room_bed’ with moderately high values

3. seeing a negative value for ‘ceil’ depicting very lighter tail and thus showing more discreetness in values

iii. Box plot (checking for outliers)

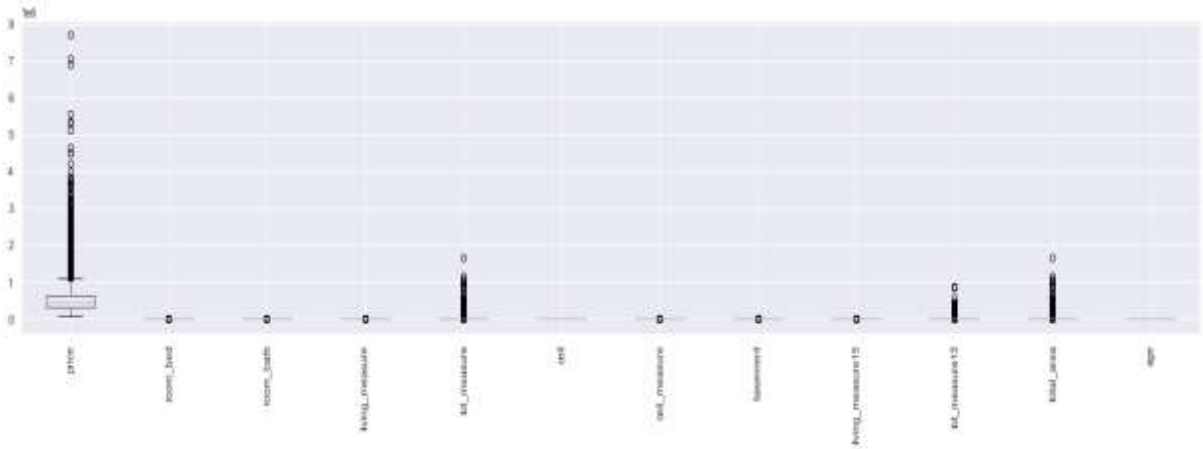

Fig 1.2 (box plot)

	percentage
price	5.362513
room_bed	2.507750
room_bath	2.632675
living_measure	2.646555
lot_measure	11.224726
ceil	0.000000
ceil_measure	2.827002
basement	2.294915
living_measure15	2.498496
lot_measure15	10.142044
total_area	11.187711
age	0.000000

Table 1.10 (outlier percentage)

1. as we assumed from kurtosis value it is proved from table 1.9 that variables ‘total_area’ , ‘lot_measure15 ‘ and ‘lot_measure’ have high outliers followed by ‘price’

2. except ‘age’ and ‘ceil’ all other variables have outliers present

Outlier treatment :-

Now let's treat the outliers as it will create problem during modelling as regression models are very sensitive to model

So we exclude the target variable 'price' from the treatment as during training the independent variable extreme values relativeness causes major issue .

So we will be capping the outliers to respective quartiles i.e outliers above upper quartile that is beyond 75% will be capped to 75% and those below lower quartile i.e 25% will be capped to 25%

So after treatment in python , it can be checked in boxplot

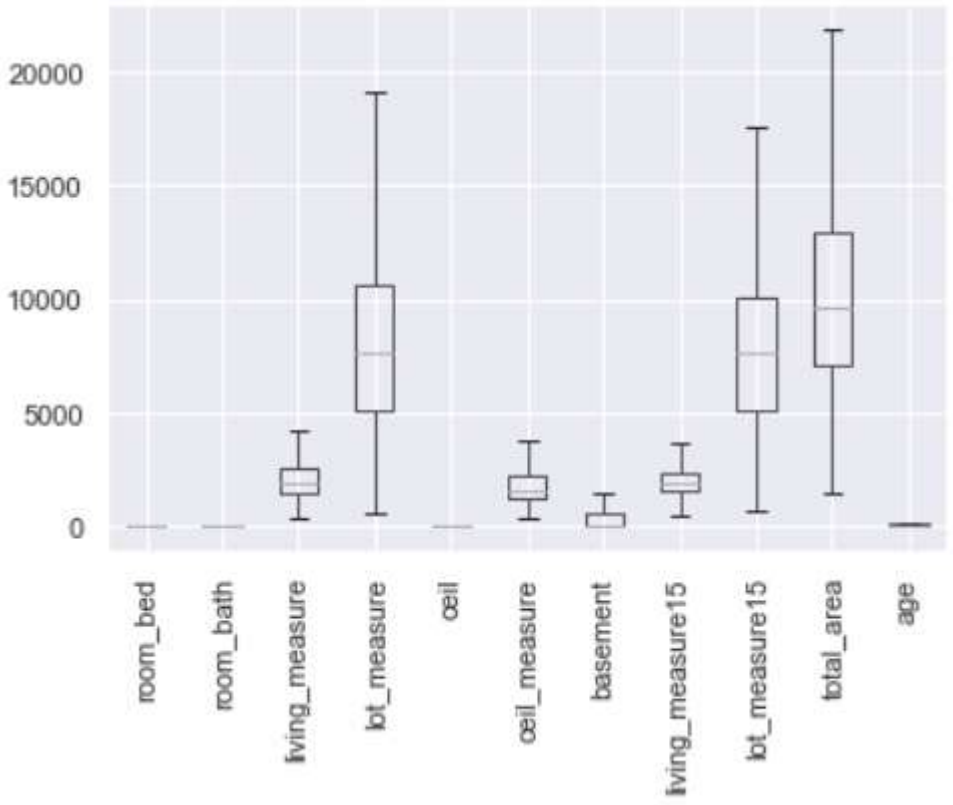

Fig 1.3 (box plot after treatment)

So we can see form boxplot in fig 1.3 after treatment no outliers are present in independent variables

b. Univariate analysis of categorical variables

i. Count plot

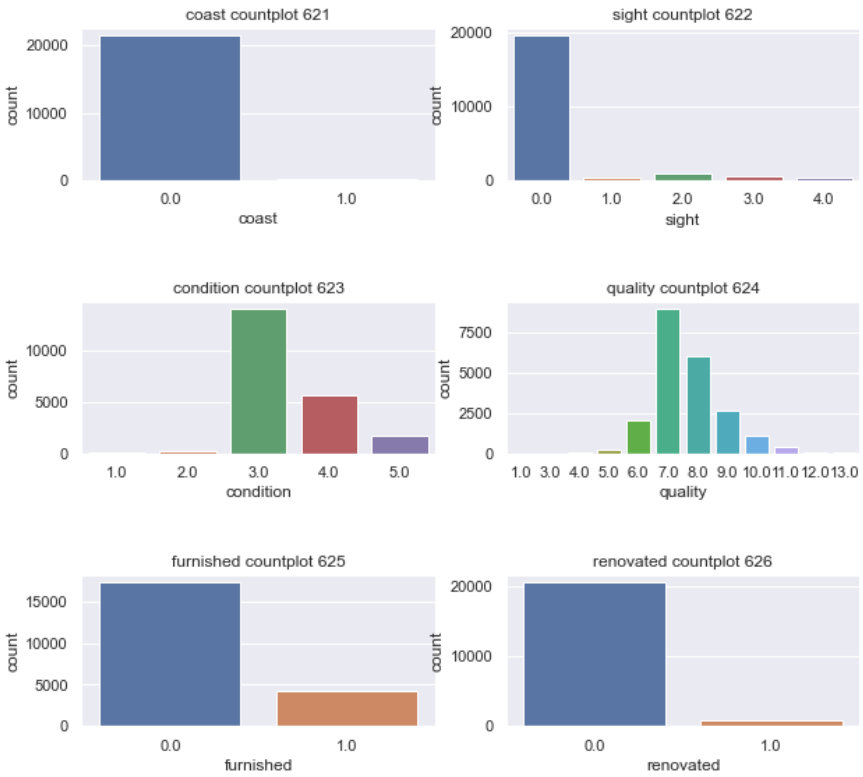

Fig 1.4 (count plot)

So following are the observation :-

1. **coast :-** we can see maximum number of house are in non-coastal i.e non sea facing area depicted by '0' in count plot and in value_counts it was observed only 161 houses are there in coast line ore we can say sea facing

2. **sight :-** we can see maximum number of records in '0' category depicting very less visits before purchase so was in less in demand . so now if see in value counts and say that maximum number of visits meaning high demand of that property then '4' classs has 318 records meaning 318 properties with high demand

3. **condition:-** so we see overall condition of the houses being purchased , its graded 3 and above mostly and its understandable otherwise why customer will buy , but certain property are graded low, in that case we can assume that people bought at low price and renovated and may be further resold at high price

4. **quality :-** so same as we can relate from above , maximum property are given grades 6 and above from 13 that's average and above

5. **furnished :-** so we can see maximum houses were not furnished which was sold depicted by '0' and furnished houses have count as less as number 4246 depicted by '1'

6. **renovated :-** we can see maximum houses not been renovated which was sold depicted by '0' and renovated houses had count as less as number 914 depicted by '1'

4. Now let's see bivariate analysis
    a. Pair plot & correlation heatmap to understand linear relation between variables

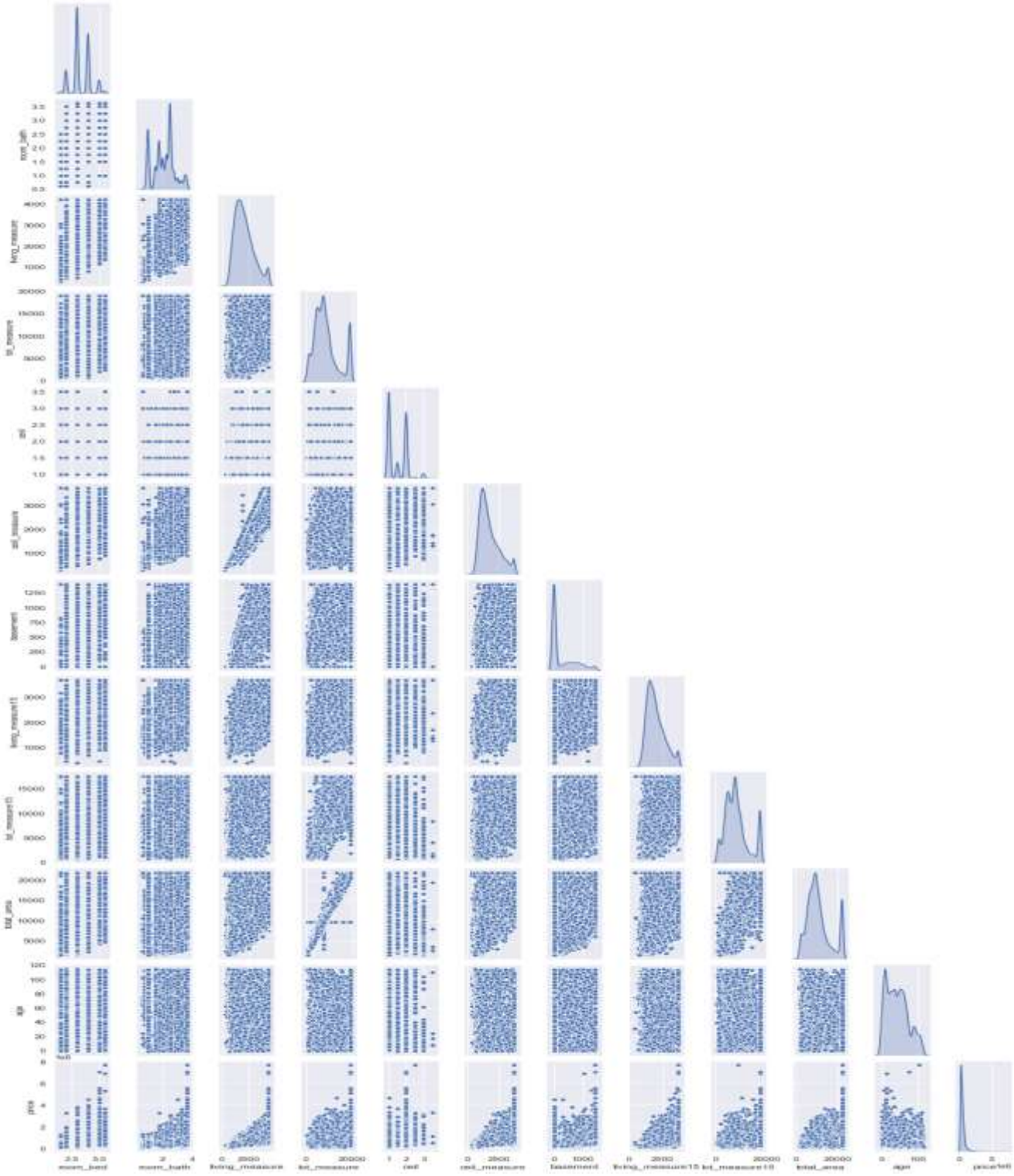

Fig 1.5 (scatter plot)

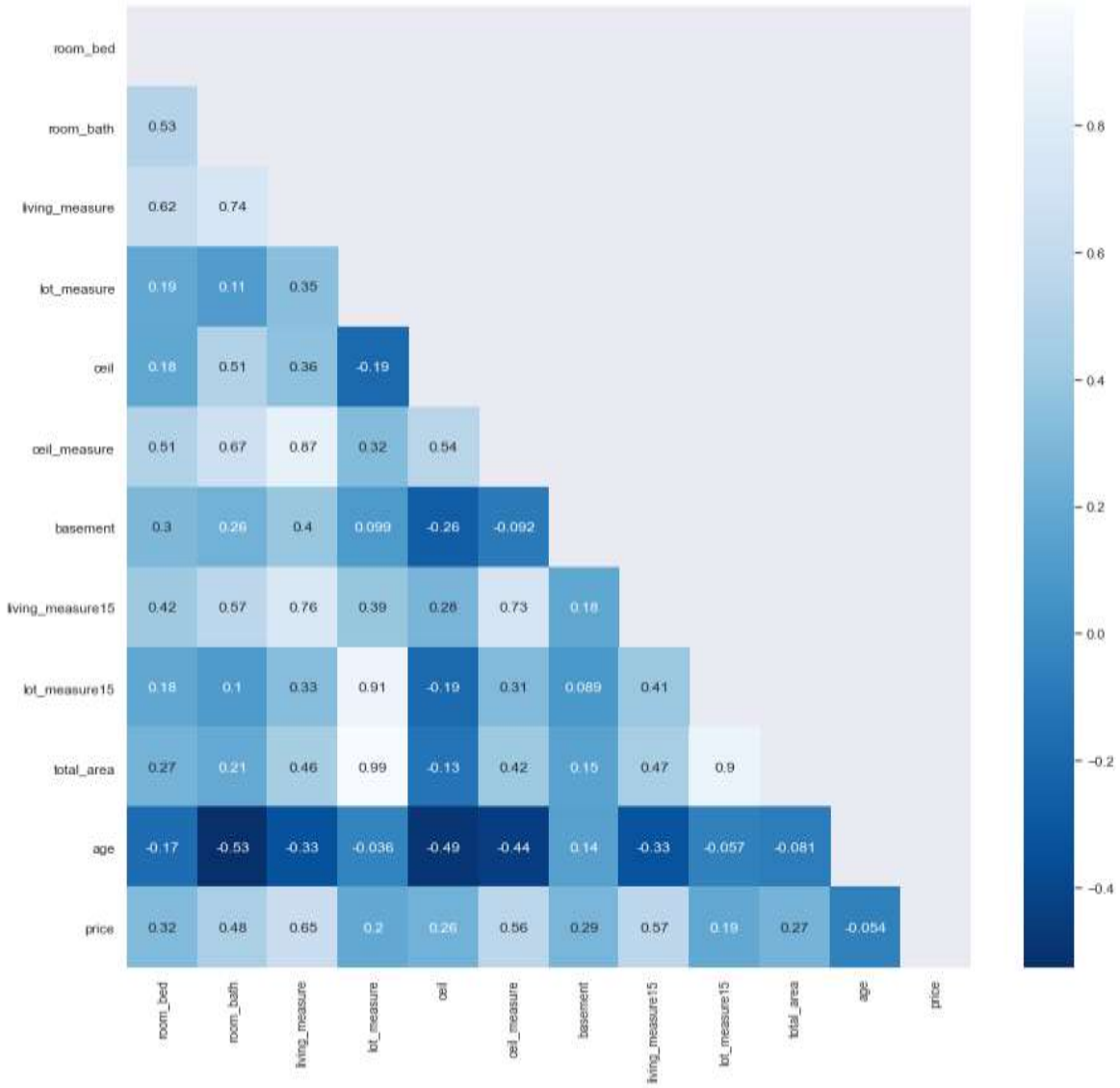

Fig 1.6 (heatmap of correlation)

Following are the observation :-

1. we can see very strong linear relation ship of following pairs in terms of linearity visible in scatter plot and correlation coefficients more than 0.80

- total_area &lot_measure - 0.99
- lot_measure15 & lot_measure - 0.91
- lot measure 15& total_area - 0.9
- ceil measure & living area – 0.87

So these variables showcasing high multicollinearity which we will further analysis with specific function “variable inflation factor” or VIF on later stage

2. if we see specifically relation of independent variables with target variable price in here , living measure showcasing highest value with positive correlation and age showcasing lowest correlation

b. Let's see bar plot between target variable "price" and independent categorical variables

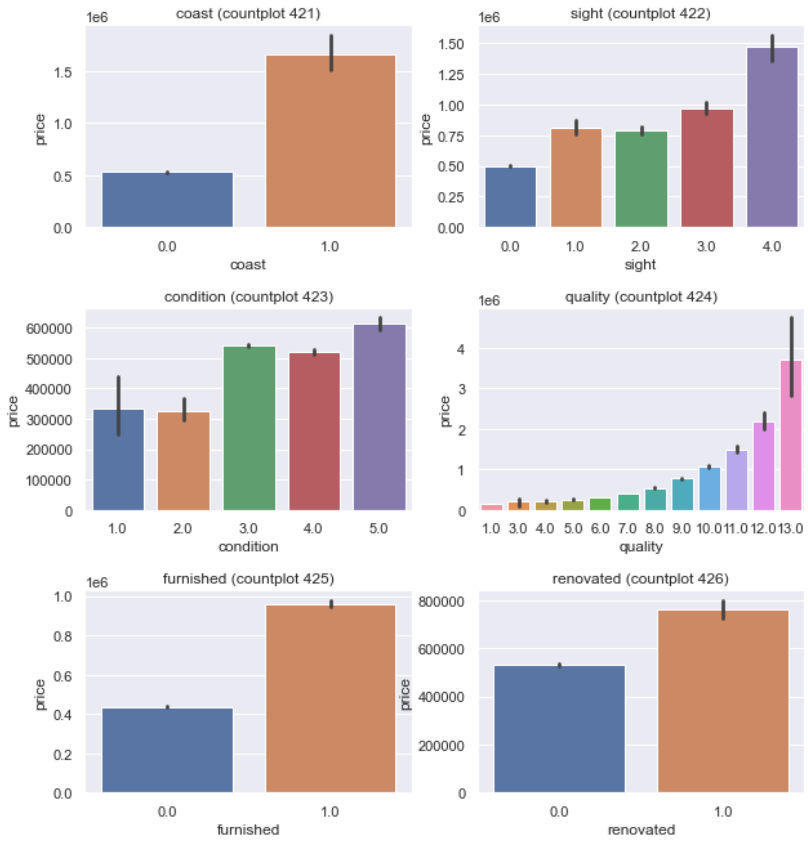

Fig 1.7 (bar plot between price and other independent categorical variables)

Following are the inferences :-

1. **price vs coast :-** we can see mean prices of the houses sea facing or on coastal side are sold on very high prices depicted by '1' as compared to non sea facing houses depicted by '0'

2. price vs sight :- we can see those houses which have been visited maximum number of times fetched the highest prices so in short highest in demand and subsequently least visited houses fetched lowest prices depicted by '0'

3.**price vs condition :-** so again we can see that mean prices for houses with overall condition of the house with higher grades from people fetched higher price but if we see other grades we can observe unusualness for eg. houses graded with 3 and 4 have almost equal mean prices and also same applies for houses with grades 1 and 2 ,this could be because of unbalanced class as it can observed in count plot

4. **price vs quality :-** so its observed again that houses which was graded with highest quality sold at higher price and as grade of quality decreased , the prices also decreased gradually.

5. **price vs furnished :-** here it can be observed furnished houses sold at higher mean prices depicted with '1' as compared to non-furnished houses depicted with '0'

6. **prices vs renovated :-** here we can see renovated house depicted with '1' fetched higher mean prices as compared to non-renovated houses depicted with '0'.

c. Zip code vs price

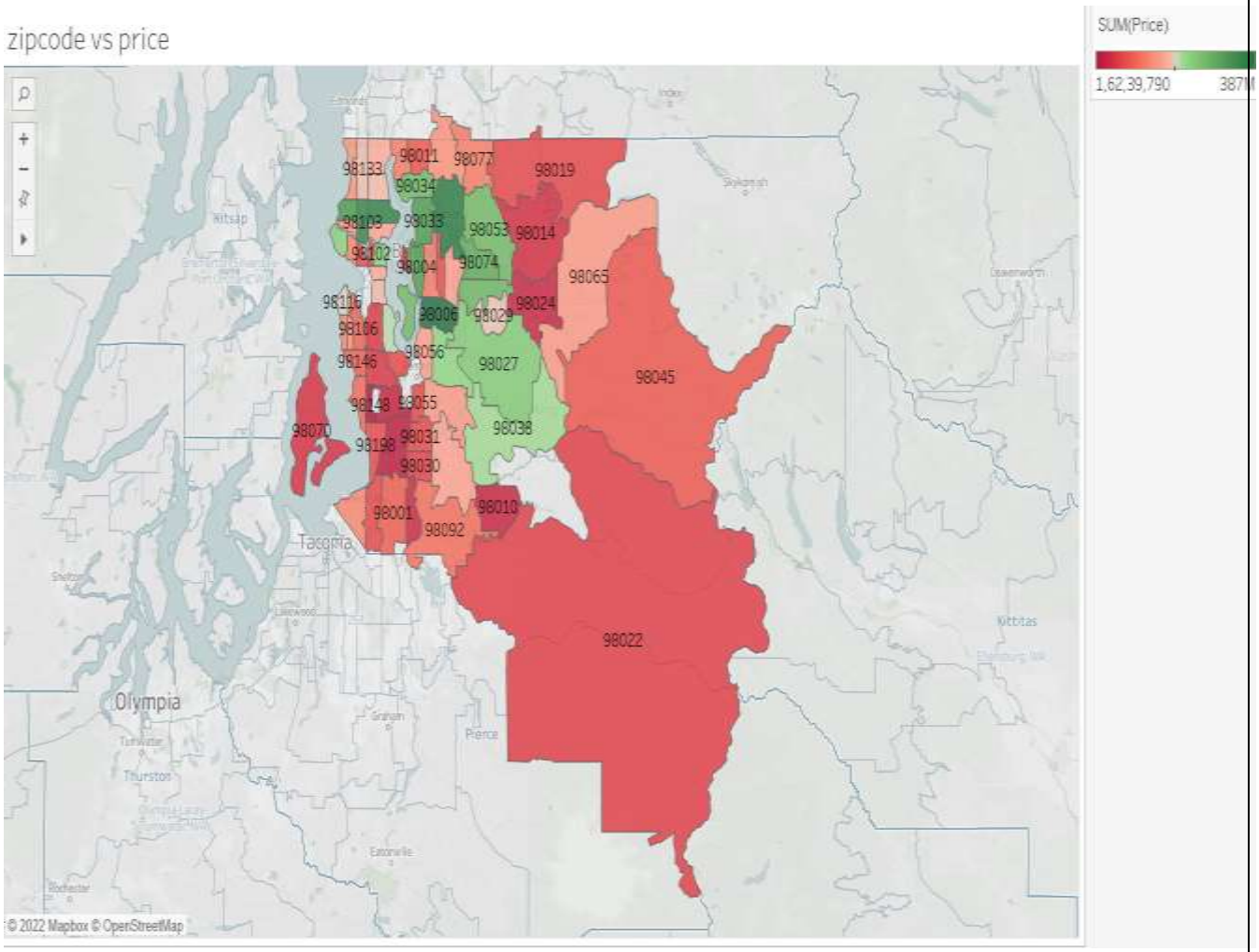

Fig 1.8 (zip code vs price)

A geographical representation of houses prices with respect to the zip code given was plotted in the map of Seattle city , Washington dc using Tableau .

So as we can see from fig 1.8 houses on locations which are in middle of Seattle city are sold on higher prices as compared on the borders .

d. Sight vs quality & sight vs condition

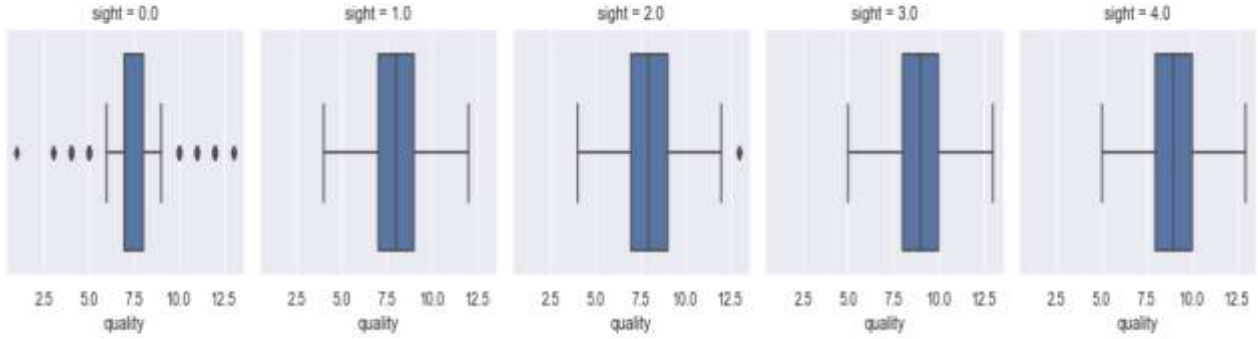

Fig 1.8 (sight vs quality)

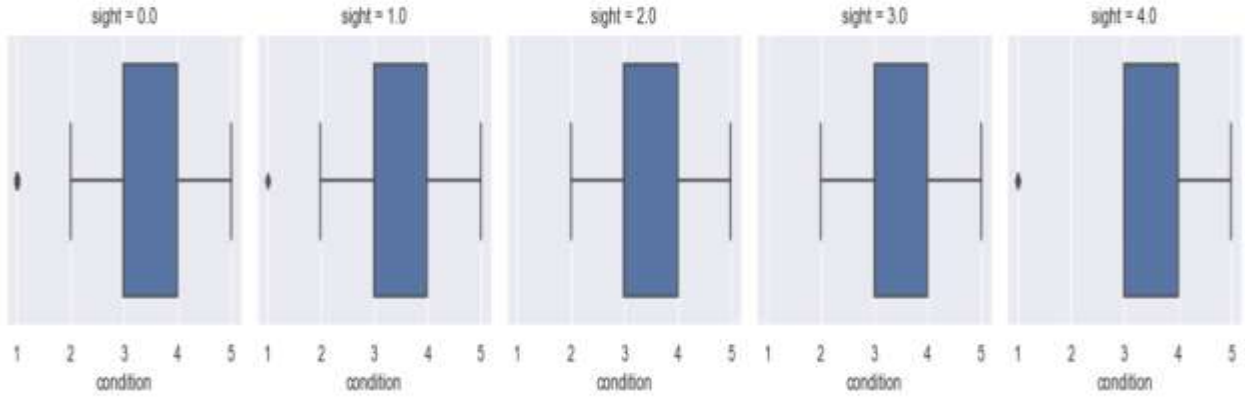

Fig 1.9 (sight vs condition)

So we can see from the two plots that how constant visits of the customer to the property is driven by quality and condition of the property .

From fig 1.8 which tells about relation between sight and quality we can see people visited maximum times on houses which are graded above average

From fig 1.9 we can see people have visited most often in houses with overall condition graded above average

So we can conclude that number of visits directly proportional to quality or condition of the houses or we can also say demand of houses is directly proportional to quality or condition of the houses .

5. **Now let's check the multicollinearity present between independent variables using VIF**

So we will be using variable inflation factor function for this purposes

After applying in first iteration following are the values observed sorted in descending order of value

	variables	VIF
0	total_area	583.415944
1	living_measure	537.626936
2	lot_measure	427.070407
3	ceil_measure	422.106544
4	quality	195.451783
5	zipcode	192.184476
6	condition	35.092891
7	room_bath	32.016409
8	living_measure15	30.699209
9	room_bed	29.437481
10	lot_measure15	28.220340
11	basement	27.756880
12	ceil	20.005786
13	age	6.628110
14	furnished	3.541968
15	sight	1.525734
16	coast	1.217544
17	renovated	1.200138

Table 1.11 (VIF values of each variables)

So as per standards VIF should not be more than 10 but it can be relative ,but here we can notice from table 1.10 certain variables showing very high values thus depicting high case of multicollinearity .

Variables that are showcasing very high values are total area , living_measure , lot_measure , ceiling_measure , quality , and zipcode

Treatment:-

Now lets bring down the VIF value by deleting each variable with highest value in each iteration till the variables left are normalized at values below 10 or close to 10 .

So after each iteration the final variables left with values less than or close to 10 are as below

	variables	VIF
0	total_area	4.198551
1	ceil	3.567320
2	age	2.780900
3	furnished	1.688513
4	basement	1.635204
5	sight	1.461009
6	coast	1.208566
7	renovated	1.110218

Table 1.12(VIF values of variables after treatment)

Now as we can see we are left with 10 independent variables after treatment for multicollinearity which suitable for modelling

6. **Now let's understand the significance of categorical variables on target variable 'Price '**

for this we will conduct ANOVA test to understand the impact of categorical variable on continuous target variable

further we need to convert the independent variable data type to category type

we will conduct ANOVA test using stats library in python so after formulation of relation with our continuous target variable and independent variable

so here hypothesis is as follows

H0 = independent variable does not have any significance impact on dependent variable

H1 = independent variables have significance impact on dependent variable

So here default significance value is 0.05 , so in order to estabilish significance of independent categorical variable on the continuous target variable p value should be less than 0.05 thus rejecting null hypothesis

so following is the ANOVA table output :-

	df	sum_sq	mean_sq	F	PR(>F)
coast	1.0	2.064347e+14	2.064347e+14	6865.814635	0.000000e+00
sight	4.0	3.055843e+14	7.639608e+13	2540.858017	0.000000e+00
condition	4.0	1.684603e+13	4.211508e+12	140.070570	2.141903e-118
quality	11.0	1.204626e+15	1.095115e+14	3642.242983	0.000000e+00
furnished	1.0	3.490175e+10	3.490175e+10	1.160798	2.813123e-01
renovated	1.0	2.627845e+13	2.627845e+13	873.995143	2.547740e-188
zipcode	69.0	5.097703e+14	7.387976e+12	245.716756	0.000000e+00
Residual	21521.0	6.470728e+14	3.006704e+10	NaN	NaN

Table 1.13 (ANOVA table output)

So we can see only variable furnished is having p value more than 0.05 thus failing to reject the null hypothesis and thus proving it does not have significant impact on dependent variable price

So we can conclude that except variable furnished all other categorical variable have significant impact on target variable 'price'.

4) Business insights from EDA

a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business

so we see problem statement its basically regression problem thus target variable is continuous so unbalanced data concept does not apply to target variable

but if we analyse our independent variables there are categorical variables with classes which are unbalanced for example :

variable 'coast' which has depicts significantly higher houses in non coastal area than coastal area

variable 'quality' & condition have class unbalance with certain classes with very less records

variable 'renovated' depicting very less houses which are renovated

b) Any business insights using clustering

let's do clustering using the variables , its an unsupervised technique so there is no concept of dependent and independent variable here

we will be using k-means clustering technique which is distance based algorithm . so before applying the model its highly required that all variables are in same scales , so scaling should be done

so we will be doing z transformation using standard scaler

then after that we need to find the right k value to divide into clusters , for that we will check with elbow method by wss plot

so following is the wss plot trying with k in range 1-11

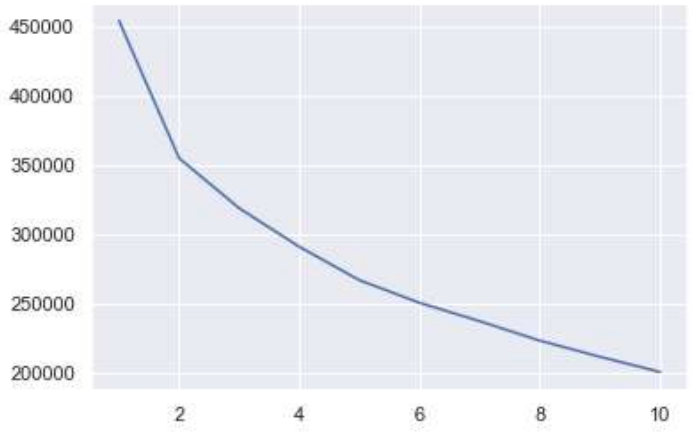

Fig 1.10 (wss plot)

So we can notice from fig 1.10 elbow is formed at 2 thus establishing value of k to be 2 as best . but 2 clusters seems meaningless , at least 3 should be there , so we will consider value of k as 3

So considering k value 3 , we pass in k-means algorithm to form 3 clusters by predicting on the data frame and creating a variable clusters in data frame

So now we see cluster profile by applying group by on cluster column with aggregation median and adding column of frequency of clusters to the data frame of group by

So following is cluster profiling:-

clusters	0	1	2
price	810000.0	374000.00	420000.00
room_bed	4.0	3.00	3.00
room_bath	2.5	2.25	1.75
living_measure	3170.0	1880.00	1480.00
lot_measure	10362.0	8800.00	5250.00
ceil	2.0	1.00	1.00
ceil_measure	2830.0	1590.00	1210.00
basement	0.0	0.00	0.00
living_measure15	2830.0	1890.00	1520.00
lot_measure15	10025.0	8537.00	5320.00
total_area	13726.0	10675.00	6982.00
age	20.0	36.00	65.00
coast	0.0	0.00	0.00
sight	0.0	0.00	0.00
condition	3.0	3.00	3.00
quality	9.0	7.00	7.00
furnished	1.0	0.00	0.00
renovated	0.0	0.00	0.00
freq	4595.0	8852.00	8166.00

Table 1.14 (cluster profiling)

1. first we can see frequency of clusters that is highest for 1 followed by 2 then 0

2. now if see the price median we can see highest median price for cluster 0 followed by 2 then 1

3. now if we see total area 0 again has highest median followed by 1 then then2

4. in living measure 0 has highest median but here we can see 2 has higher median value than 1 so we

So as we can observe from above points with respect to points of price , area and rooms also we can conclude as following

Cluster 0 – luxury houses- with more area , quality , price

Cluster 2 - medium range house – with medium area , quality price

Cluster 1 – low range house – with respect to area , quality & price

c) Any other business insights :-

following are the insights or summary

1. We have observed high demand of properties with higher grade in quality and overall condition leading to higher price
2. Houses located at coastal area is sold at higher price than non coastal rea
3. Furnished houses are sold at very high prices than non-furnished but it came out to be insignificant in ANOVA test
4. Renovated houses are sold at higher prices than non-renovated
5. Houses which are located at centre of Seattle city fetched higher price as compared to the border areas of Seattle city
6. Further age of the houses is not showing that much relation or significance to target variable price, that means age is not affecting the price as such