Murali b

CAPSTONE PROJECT 1 NOTES SUBMISSION

CONTENTS Page numbers

Tables:   
1.1 Head of the data 3  
1.2 Tail of the data 4

Outputs:   
1.3 Shape of the data 4  
1.4 Describe the data 4  
1.5 Column names 4  
1.6 Renaming column names 5  
1.7 Datatypes details 5  
1.8 Info of the data 6  
1.11 Price variable details 9  
1.13 cid variable counts 9  
1.14 cid unique values 9  
1.16 room bed value counts 10  
1.18 bathroom value counts 12  
1.19 bathroom unique value 12   
1.23 Describe lot\_measure 14  
1.25 cell value counts 15  
1.27 coast variable value counts 16  
1.29 sight variable value counts 17  
1.31 condition variable value counts 18  
1.33 quality analysis value counts 19  
1.35 describe ceil measure 20  
1.37 basement variable count analysis 21  
1.39 year\_built value counts 23  
1.41 year\_renovated value counts 23  
1.43 zipcode unique values 24  
1.45 furnished variable value counts 25  
1.47 living\_measure15 value counts 26  
1.49 lot\_measure15 value counts 27  
1.51 lat value counts 28  
1.53 long variable value counts 29  
1.54 describe total\_area 30

Figures:   
1.9 Histogram 8  
1.10 Variable price 8  
1.12 cid variable analysis 9  
1.15 room bed analysis 10  
1.17 bathroom analysis 11   
1.20 living measure 13  
1.21 describe of living 13  
1.22 lot\_measure variable 14  
1.24 cell variable 15  
1.26 coast variable analysis 16  
1.28 sight analysis 17  
1.30 condition variable analysis 18  
1.32 quality analysis 19  
1.34 ceil\_measure 20  
1.36 variable basement 21   
1.38 year\_built variable 22  
1.40 year\_renovated 23  
1.42 zipcode variable 24  
1.44 furnished variable 25  
1.46 living\_measure15 variable analysis 26  
1.48 lot\_measure15 variable analysis 27  
1.50 lat variable analysis 27  
1.52 long variable analysis 29  
1.54 total\_area variable analysis 30  
1.55 room bed analysis 31  
1.56 room bath analysis 31  
1.58 coast variable analysis 32  
1.59 sight variable analysis 33  
1.60 condition variable analysis 33  
1.61 quality variable analysis 34  
1.62 furnished variable analysis 34  
1.63 cid and price 35  
1.64 dayhours and price 35  
1.65 room\_bed and price 37  
1.66 room\_bath and price 37  
1.67 living\_measure and price 38  
1.68 lot\_measure and price 38  
1. 69 sight and price 39  
1.70 quality and price 39  
1.71 ceil\_measure and price 40  
1.72 basement and price 40  
1.73 year\_renovated and price 41  
1.74 zipcode and price 41  
1.75 lat and price 42  
1.76 living\_measure15 and price 42  
1.77 lot\_measure15 and price 43  
1.78 furnished and price 43  
1.79 Pairplot 44

CAPSTONE PROJECT NOTES 1:

1. **Problem Understanding:**
2. Defining problem statement

Basically, any house is valued based on the area in which it is located and its dimensions which is measured in square feet. (For ex: 30\*40 = 1200sqft). So, any layman or common individual would obviously wish to know all the features related to housing value properties. Hence if in case if we wish to buy any house or sell any house then we will be not knowing the exact amount to which it's being valued whether the cost involved is more or less so in order to address these kind of issues we tend to take samples information which is already existing such as finding same pattern in the same locality and with that data we can try to predict the house value.

1. Need of the study/project

The basic need of the study of this project is to determine the right correct price of the housing value. So, when any person tries to buy or sell a house, usually they won't know the exact price to which it should be sold or bought. Hence the house value might be undervalued or overvalued. Based on the data such as the available information of the house property in that geographical location we can determine the price. Primarily we should ensure to make better understanding of the various important factors and features which helps in determining the right house pricing.

1. Understanding business/social opportunity

Here main ideology is to avoid the loss and time so many people will not be aware of the features which determines the house price. So, to help them further we can provide them a detailed information and guidelines based on house buy/sell services so that people can buy the property with feasible prices so that losses can be minimized and waiting time can be reduced.

1. **Data Report**
2. Understanding how data was collected in terms of time, frequency, and methodology.

Top 5 rows of the data:

A picture containing text, font, line, number

Description automatically generated  
Table 1.1 Head of the data

Last 5 rows of the data :   
  
A picture containing text, font, number, line

Description automatically generated  
Table 1.2 Tail of the data   
  
  
b) Visual inspection of data (rows, columns, descriptive details)   
  
Shape of the data :   
  
  
Output 1.3 Shape of the data  
  
  
There are 21613 entries with 23 observations.

Describe the data:   
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.4 Describe the data.   
  
We can see the five point summary of the dataset such as count , mean , standard deviation , min and max values and 25% , 50% and 75% percentile values.

c) Understanding of attributes (variable info, renaming if required)   
  
Columns names:   
A picture containing text, screenshot, font, line

Description automatically generated  
Output 1.5 Column names   
We can see 2 column names showing as yr\_built and yr\_renovated so let’s rename it.   
  
Renaming column names: A picture containing text, screenshot, font, line

Description automatically generated  
Output 1.6 Renaming column names

So now we have renamed the column names to year\_built and year\_renovated.

Datatypes details:   
  
A screenshot of a computer program

Description automatically generated with low confidence  
  
Output 1.7 Datatypes details

Info of the data:   
  
A screenshot of a computer

Description automatically generated with medium confidence  
Output 1.8 Info of the data

We can see that there are 16 float datatype variables and 7 object datatype variables.

3) Exploratory data analysis

a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

A picture containing diagram, text, plot, line

Description automatically generatedA picture containing diagram, text, plot, line

Description automatically generated  
Fig 1.9 Histogram

We can see most of the data representation shows as left skewed and the data is uniformly distributed.

Variable: Price   
A picture containing text, diagram, screenshot, line

Description automatically generated  
Fig 1.10 Variable price

A picture containing text, font, screenshot

Description automatically generated  
Output 1.11 Price variable details   
  
We see outliers present in price variable and it’s left skewed distribution.

Variable : cid   
  
A screenshot of a graph

Description automatically generated with low confidence  
Fig 1.12 CID variable analysis   
  
Value counts :   
  
Output 1.13 cid value counts

Unique values:   
  
Output 1.14 cid unique values

We see uniform distribution of data in cid variable analysis and no outliers present.

Variable : room\_bed   
A picture containing text, screenshot, plot, diagram

Description automatically generated  
Fig 1.15 Room bed analysis   
  
We see 9767 houses has 3 bedrooms , 6854 houses has 4 bedrooms and 2747 houses has 2 bedrooms.   
  
Value counts :   
A screenshot of a computer

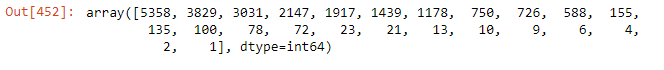
Description automatically generated with medium confidence  
Output 1.16 room bed value counts  
  
Variable : room\_bath   
A picture containing text, screenshot, plot, line

Description automatically generated  
Fig 1.17 Bathroom analysis   
  
We see more than 5358 houses has more than 2 bathroom and 3829 houses has 1 bathroom.

Value counts:   
A picture containing text, screenshot

Description automatically generated  
Output 1.18-bathroom value counts



Unique value:   
  
  
Output 1.19-bathroom unique values.

Variable : living\_measure   
  
A picture containing screenshot, text, diagram, plot

Description automatically generated  
Fig 1.20 living measure

Describe of living\_measure :   
A picture containing text, screenshot, font, number

Description automatically generated  
Fig 1.21 describe of living  
  
We see outliers present in living measure variable and minimum value is 290 and maximum value is 13540.   
Variable : lot\_measure A picture containing text, screenshot, diagram, line

Description automatically generated

Fig 1.22 lot\_measure variable

Describe of lot\_measure :   
  
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.23 describe lot\_measure

We see there is no proper data distribution found in lot\_measure and many outliers present in lot\_measure

Variable: ceil   
A picture containing text, screenshot, diagram, plot

Description automatically generated  
Fig 1.24 cell variable

Value counts :   
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.25 cell value\_counts

We see 10647 values in 1.0 cell variable, 8210 values in 2.0 cell

Variable: coast   
A picture containing text, screenshot, rectangle, display

Description automatically generated  
Fig 1.26 coast variable analysis

Value counts :   
A picture containing text, font, screenshot, white

Description automatically generated  
Output 1.27 coast variable value counts

So we here see that 0.0 has 21421 entries of coast analysis values and 1.0 values with 161 entries.

Variable : sight   
A picture containing text, screenshot, plot, diagram

Description automatically generated  
Fig 1.28 sight analysis

Value counts :   
A picture containing text, font, screenshot, number

Description automatically generated  
Output 1.29 sight variable value counts

Sight variable has 19437 values of 0.0 sight, 959 values of 2.0 sight, 510 values of 3.0 sight.

Variable: condition   
A picture containing text, screenshot, diagram, plot

Description automatically generated  
Fig 1.30 condition variable analysis   
  
Value counts :   
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.31 condition variable value counts

We see 13978 values of 3.0 condition variable, 5655 values of 4.0 condition variable , 1694 values of 5.0 condition variable.

Variable : quality   
A picture containing text, screenshot, plot, diagram

Description automatically generated  
Fig 1.32 Quality analysis

Value counts :   
A screenshot of a computer

Description automatically generated with medium confidence  
Output 1.33 quality analysis value counts

Quality of the house shows as 8981 entries with 7 quality , 6067 entries with 8 quality , 2615 entries with 9 quality.   
  
  
Variable : ceil\_measure   
  
A picture containing screenshot, diagram, plot, line

Description automatically generated  
Fig 1.34 ceil\_measure variable

Describe of ceil\_measure :   
A screenshot of a computer

Description automatically generated with low confidence  
Output 1.35 describe ceil measure  
  
ceil\_measure variable has minimum value of 290 , maximum value of 9410 , mean value of 1788.

Variable : basement   
A picture containing text, screenshot, diagram, line

Description automatically generated  
Fig 1.36 variable basement

Value counts :   
A screenshot of a computer

Description automatically generated with low confidence  
Output 1.37 basement variable value counts   
  
We can see 13125 houses has no basement , 221 houses has 600 basement , 218 house has 700 basement.

Variable : year\_built   
A picture containing text, screenshot, colorfulness, plot

Description automatically generated

Fig 1.38 year\_built variable

Value counts :   
A screenshot of a computer

Description automatically generated with medium confidence  
Output 1.39 year\_built value counts   
  
Year\_built variable shows as 559 houses built in 2014 year , 454 houses built in 2006 , 450 houses built in 2005 year.

Variable : year\_renovated   
A picture containing screenshot, text, line, plot

Description automatically generated  
Fig 1.40 year\_renovated

Value counts :   
A screenshot of a computer

Description automatically generated with low confidence  
Output 1.41 year\_renovated value counts   
  
year\_renovated variable shows as 20699 houses was not renovated , 91 houses was renovated in year 2014 , 37 houses was renovated in the year 2013.

Variable : zipcode   
A picture containing text, screenshot, colorfulness, line

Description automatically generated  
Fig 1.42 zipcode variable

Unique values :   
A picture containing text, screenshot, font, number

Description automatically generated

Output 1.43 zipcode unique values

Variable : furnished   
A picture containing text, screenshot, diagram, rectangle

Description automatically generated  
Fig 1.44 furnished variable   
  
Value counts :   
A picture containing text, font, screenshot, white

Description automatically generated

Output 1.45 furnished variable value counts   
  
We see 4246 houses are only furnished and 17338 houses are not furnished.

Variable : living\_measure15   
A picture containing text, diagram, plot, screenshot

Description automatically generated  
Fig 1.46 living\_measure15 variable analysis

Value counts :   
A screenshot of a computer code

Description automatically generated with low confidence  
Output 1.47 living\_measure15 value counts   
  
living\_measure15 variable shows as 197 houses with 1540 values, 193 houses with 1440 values , 192 houses with 1560 values.

Variable : lot\_measure15   
A screen shot of a graph

Description automatically generated with medium confidence  
Fig 1.48 lot\_measure15 variable analysis

Value counts :   
A screenshot of a computer

Description automatically generated with medium confidence  
Output 1.49 lot\_measure15 value counts  
  
From the above analysis, we see lot\_measure as 427 houses has values of 5000 , 357 houses has values of 4000 , 287 houses with 6000 values.

Variable : lat   
A picture containing text, screenshot, diagram, plot

Description automatically generated  
Fig 1.50 lat variable analysis

Value counts :   
A screenshot of a computer

Description automatically generated with medium confidence  
Output 1.51 lat value counts

From the above graph, we see that the data is right-skewed and no outliers present in the data.

Variable : long A picture containing text, screenshot, diagram, plot

Description automatically generated  
Fig 1.52 long variable analysis

Value counts :   
A screenshot of a computer

Description automatically generated with medium confidence  
Output 1.53 long variable value counts  
  
Here, we can observe that there are many outliers present in the long variable analysis and data is almost skewed distribution.

Variable : total\_area

A screenshot of a graph

Description automatically generated with medium confidence  
Fig 1.54 total\_area variable analysis

Describe of total\_area :

A picture containing text, font, screenshot, line

Description automatically generated  
Output 1.54 describe total\_area   
  
No proper data distribution found in total\_area variable and outliers present in total\_area variable.

Analysis of room\_bed :   
A picture containing text, screenshot, diagram, plot

Description automatically generated  
Fig 1.55 room\_bed analysis

We see 10000 houses has 3 bedrooms , 7000 houses has 4 bedrooms , 3000 houses has 2 bedrooms.

Analysis of room\_bath :   
A picture containing text, screenshot, diagram, plot

Description automatically generated  
Fig 1.56 room\_bath analysis   
  
Here, we can see 6000 houses has more than 2.5 bathrooms , 4000 houses has 1 bathroom , 3000 houses has 1.75 bathrooms.

Analysis of Ceil : A picture containing screenshot, rectangle, diagram, plot

Description automatically generated  
Fig 1.57 cell variable analysis  
  
We see 10000 houses has only 1 ceil , 8000 houses with 2 ceil , 2000 houses with 1.5 ceil values.

Analysis of Coast : A picture containing screenshot, text, rectangle, plot

Description automatically generated  
Fig 1.58 coast variable analysis   
  
22500 houses are situated near coast side area and remaining houses are not located near coast side.

Analysis of Sight :   
A picture containing plot, rectangle, line, software

Description automatically generated  
We see 18500 values with 0 sight values, 500 values with 2.0 sight values.   
  
Fig 1.59 sight variable analysis

Analysis of Condition :   
A picture containing text, screenshot, diagram, rectangle

Description automatically generated  
Fig 1.60 condition variable analysis   
  
Here, we see 14000 houses are in good condition with 3 rating , 6000 houses with 4 rating.

Analysis of Quality : A picture containing diagram, plot, screenshot, line

Description automatically generated  
From the above graph, we see 9000 houses are in good quality of 7 , 6000 houses with 8.0 quality rating.   
  
  
Fig 1.61 quality variable analysis

Analysis of Furnished :   
A picture containing text, screenshot, rectangle, diagram

Description automatically generated  
Fig 1.62 furnished variable analysis   
  
  
Here, we see 3500 houses are only furnished but 17500 houses are not furnished.   
  
  
b) Bivariate analysis (relationship between different variables , correlations)

Variable used : cid and price   
A picture containing text, screenshot, diagram

Description automatically generated  
Fig 1.63 cid and price   
  
We see all the points are scattered at the bottom of the graph so the value keeps decreasing.

Variable used : dayhours and price   
A screen shot of a graph

Description automatically generated with low confidence  
Fig 1.64 dayhours and price   
  
Here, we see there is more scattered relationship between dayhours and price

Variable used : room\_bed and price   
A picture containing text, screenshot, plot, line

Description automatically generated  
Fig 1.65 room\_bed and price   
  
There is an upward direction in the trend of datapoints.

Variable used : room\_bath and price   
A picture containing text, screenshot, plot, line

Description automatically generated  
Fig 1.66 room\_bath and price   
  
As the value keeps increasing the data points is getting scattered on the right side of the plot.

Variable used : living\_measure and price   
A picture containing text, screenshot, diagram, font

Description automatically generated  
Fig 1.67 living\_measure and price   
  
We see there is a deep steep downward trend in the values as the values keeps decreasing.

Variable used : lot\_measure and price   
A picture containing text, screenshot, diagram, plot

Description automatically generated  
Fig 1.68 lot\_measure and price  
  
When lot\_measure and price are plotted against each other we see there is no proper data representation between them.

Variable used : sight and price   
A picture containing text, screenshot, line, diagram

Description automatically generated  
Fig 1.69 sight and price   
  
We see linear data representation of datapoints between sight and price.

Variable used : quality and price   
A picture containing text, screenshot, line, plot

Description automatically generated  
Fig 1.70 quality and price  
  
As quality gets increased, we can see the cost of the house prices also getting increased.

Variable used : ceil\_measure and price  
A picture containing text, screenshot, diagram, font

Description automatically generated  
Fig 1.71 ceil\_measure and price

There is again a downward trend followed between ceil\_measure and price

Variable used : basement and price   
  
A picture containing text, screenshot, diagram

Description automatically generated  
Fig 1.72 basement and price

All the datapoints are scattered on the left side when plotted against basement and price.

Variable used : year\_renovated and price   
A picture containing text, screenshot, line, plot

Description automatically generated

Fig 1.73 year\_renovated and price

There is no relationship and only few data points. Being found between year\_renovated and price

Variable used : zipcode and price   
  
A screen shot of a graph

Description automatically generated with low confidence  
Fig 1.74 zipcode and price  
  
As the zipcode increases eventually houses are situated in a are with increasing zipcode values.

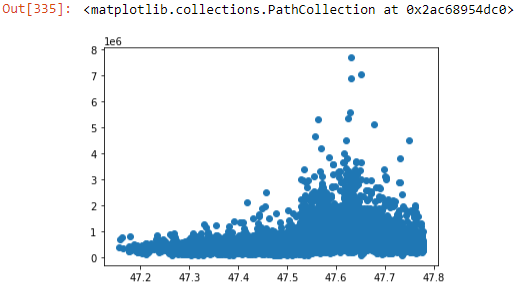
Variable used : lat and price   
  


Fig 1.75 lat and price

We see as the latitude increases even the price gets increased.

Variable used : living\_measure15 and price   
  
A picture containing text, screenshot, diagram, plot

Description automatically generated  
Fig 1.76 living\_measure15 and price

So from the above plot, we can conclude that as living measure gets increased then price gets increased.

Variable used : lot\_measure15 and price   
A picture containing text, screenshot, plot, line

Description automatically generated

Fig 1.77 lot\_measure15 and price   
  
lot\_measure having lesser values has more higher cost of houses.

Variable used : furnished and price   
A screenshot of a graph

Description automatically generated with low confidence

Fig 1.78 furnished and price   
  
Furnished houses has higher price value than houses which is not furnished.

Pairplot :   
A picture containing text, screenshot, handwriting, rectangle

Description automatically generated

Fig 1.79 Pairplot   
  
We see all the variable data representation through pairplot which shows unique distribution of data points when they are related to each other.

Correlation :   
A picture containing text, screenshot, number, font

Description automatically generated  
Output 1.80 Correlation

Heatmap :   
A picture containing screenshot, rectangle, square, colorfulness

Description automatically generated  
Output 1.81 Heatmap

From the heatmap, we can see that there is dense correlation which exists between most of the variables.

b) Missing Value treatment (if applicable)   
  
Duplicate values :   
  
Output 1.82 duplicate values   
  
We can see that there is no duplicate values present in the dataset.

Checking for Null values :   
A screenshot of a computer

Description automatically generated with low confidence  
Output 1.83 checking for null values

We see that there are 108 null values in room\_bed , room\_bath , 17 in living\_measure , 42 in lot\_measure , 1 in coast & quality & ceil\_measure & basement & year\_built, 57 in sight and condition , 166 in living\_measure15 , 29 in lot\_measure15 & furnished & total\_area so let’s fix all this Null values.

After fixing Null values:   
  
We treated Null values with mode and median so now we can see that Null values has been fixed.   
A picture containing text, screenshot, font, menu

Description automatically generated  
Output 1.84 fixing Null values   
  
  
Now we can see that there is no Null values present in the dataset.

d) Outlier treatment (if required)   
A picture containing diagram, line, plan, plot

Description automatically generated  
Fig 1.85 Before fixing outliers

We see many outliers present in the dataset except for lot\_measure , zipcode and lat so let’s fix the outliers.

After fixing outliers :   
A screenshot of a computer

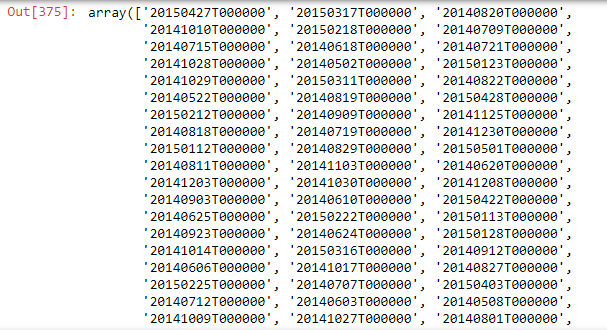
Description automatically generated with low confidence  
Fig 1.86 After fixing outliers

Now we can that there is no outliers present in the dataset and it has been take care already.

e) Variable transformation (if applicable)

Cid variable unique values:   
  
Output 1.87 cid variable unique values

Cid variable shape:   
  
Output 1.88 Cid variable shape

Dayhours variable unique values :   
  
  
Output 1.89 Dayhours variable unique values

Dayhours variable shape :   
   
Output 1.90 Dayhours variable shape  
  
Price variable unique values :   
  
Output 1.91 Price variable unique values  
  
Price variable shape :   
  
Output 1.92 Price variable shape  
  
Room\_bed variable unique values:   
  
Output 1.93 Room\_bed variable unique values  
  
Room\_bed variable shape :   
  
Output 1.94 Room\_bed variable shape

Room\_bath variable unique values :   
A picture containing text, font, line, screenshot

Description automatically generated   
Output 1.95 Room\_bath variable unique values  
  
Room\_bath variable shape :   
  
Output 1.96 Room\_bath variable shape  
  
living\_measure variable unique values :   
  
Output 1.97 living\_measure variable unique values  
  
living\_measure variable shape :   
  
Output 1.98 living\_measure variable shape  
  
lot\_measure variable unique values :   
   
Output 1.99 lot\_measure variable unique values  
  
lot\_measure variable shape :   
  
Output 2.1 lot\_measure variable shape  
  
ceil variable unique values:   
  
Output 2.2 ceil variable unique values  
  
coast variable shape :   
  
Output 2.3 coast variable shape  
  
sight variable unique values :   
  
Output 2.4 sight variable unique values  
  
sight variable shape :   
  
Output 2.5 sight variable shape

condition variable unique values :   
  
Output 2.6 condition variable unique values  
  
quality variable unique values :   
   
Output 2.7 quality variable unique values  
  
quality variable shape :   
  
Output 2.8 quality variable shape  
  
ceil\_measure variable unique values :   
A picture containing text, screenshot, number, font

Description automatically generated  
Output 2.9 ceil\_measure variable unique values

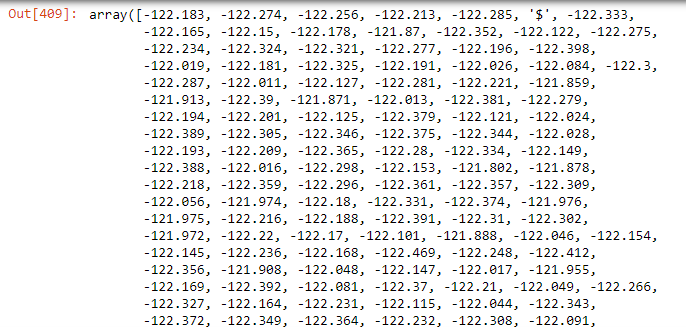
Ceil\_measure variable shape :   
  
  
Output 2.10 Ceil\_measure variable shape

basement variable unique values:   
A picture containing text, screenshot, font, number

Description automatically generated  
Output 2.10 basement variable unique values   
  
basement variable shape :   
  
Output 2.11 basement variable shape   
  
year\_built variable unique values :   
A picture containing text, screenshot, font, number

Description automatically generated  
Output 2.12 year\_built variable unique values  
  
year\_renovated variable unique values :   
A picture containing text, screenshot, font, number

Description automatically generated  
Output 2.13 year\_renovated variable unique values

year\_renovated variable shape :   
  
Output 2.14 year\_renovated variable shape  
  
zipcode variable shape:   
  
Output 2.15 zipcode variable shape  
  
lat variable unique values :   
  
Output 2.16 lat variable unique values  
  
lat variable shape :   
  
Output 2.17 lat variable shape  
  
long variable unique values :   
  
Output 2.18 long variable unique values

living\_measure15 variable unique values:   
A picture containing text, screenshot, font, number

Description automatically generated  
Output 2.18 living\_measure15 variable unique values  
  
living\_measure15 variable shape:   
  
Output 2.19 living\_measure15 variable shape  
  
lot\_measure15 variable unique values :   
  
Output 2.20 lot\_measure15 variable unique values  
  
lot\_measure15 variable shape :   
  
Output 2.21 lot\_measure15 variable shape  
  
furnished variable unique values :   
  
Output 2.22 furnished variable unique values  
  
furnished variable shape :   
  
Output 2.23 furnished variable shape   
  
total\_area variable unique values :   
  
  
Output 2.24 total\_area variable unique values

So we have converted the variable type: ceil , coast , condition , year\_built , total\_area and long to string type. We have replaced all “$” values with nan values and then imputes all nan values with KNNImputer.

f) Addition of new variables (if required) :   
  
We have added new variable called house\_land\_ratio by taking the round values of living\_measure divided by total\_area multiplied by 100. The output shows as :   
  
A picture containing text, screenshot, font, line

Description automatically generated  
Output 2.25 new variable details

4) Business insights from EDA   
  
a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business   
  
Yes, the data seems to be unbalanced from the data analysis, hence we so we try to balance the data. We see that more people are interested in buying houses which has more bedrooms and more bathrooms so we should build more houses which has more bathrooms and bedrooms. We can see that houses with basement have higher prices than compared to houses which are very low in number. Therefore, we must build houses which has basement in them and, we can go publicity like posting ads in the sites having basement so that many people will be aware whether any house is for sale or not. People likely to tend to buy houses which has 2 floors so we can focus more on building houses with 2 floors. We see many houses are not furnished so if we do fully furnish or semi furnished then we can attract more people. As per the quality wise, we see many houses are not with good quality so if we improve the quality of the houses, we can demand more prices for the houses.

b) Any business insights using clustering (if applicable)   
  
A screenshot of a computer

Description automatically generated with low confidence  
Output 2.26 clustering

Hierarchical clustering :   
  
A picture containing text, line, font, number

Description automatically generated  
Output 2.27 Hierarchical clustering  
  
Dendogram :   
  
A diagram of a city

Description automatically generated with low confidence  
Fig 2.28 Dendogram

Dendogram with p value as 10:   
  
A picture containing diagram, rectangle, plan, line

Description automatically generated  
Fig 2.29 Dendogram with p value as 10

Hierarchical Clustering Dendrogram :   
  
A picture containing diagram, line, rectangle, parallel

Description automatically generated  
Fig 2.30 Hierarchical Clustering Dendrogram

Cluster values upto 3:   
  
  
Output 2.31 Cluster values upto 3

Hierarchical Clustering Top head values:   
  
A picture containing text, screenshot, number, font

Description automatically generated  
Output 2.32 Hierarchical Clustering Top head values

Hierarchical Clusters value counts :   
A picture containing text, font, white, line

Description automatically generated  
Output 2.33 Hierarchical Clusters value counts  
  
KMeans Clustering:   
A screenshot of a computer

Description automatically generated with medium confidence  
Output 2.34 KMeans Clustering

Elbow Method :   
A graph with a red line

Description automatically generated with medium confidence  
Fig 2.35 Elbow Method

K\_means with clusters = 3 :   
  
Output 2.36 K\_means with clusters = 3  
  
  
Sil\_width :   
  
A picture containing text, screenshot, font, number

Description automatically generated  
Output 2.37 Sil\_width

K\_means1 labels :   
  
A picture containing text, screenshot, font, number

Description automatically generated  
Output 2.38 K\_means1 labels  
  
  
K\_means1 value\_counts :   
  
A picture containing text, font, white, line

Description automatically generated  
Output 2.39 K\_means1 value\_counts

Insights:

When compared to other variables, so we can apply a new strategy to educate the people about the information which most of them are not aware of it. Since many people do not have interest to know about all things so we make sure that they understand the details related to their houses and guiding them which in turn can improve the business. We can create a user-friendly customized website related to all information or a mobile app so that they can check the details about the details of their houses so based on their choice of houses we can make the sales.   
  
  
c) Any other business insights   
  
We can see that properties with higher price value have a greater number of sites when compared to houses with lesser price. So once the data is cleaned then it’s stable for model building. We can see that there is a greater increase in the upward trend in price values with respect to ceil\_measure. We can see that smaller house are in a better condition but with more higher prices.