HOUSE PRICE PREDICTION

- Change categorical values into numerical values

- mainroad (yes/No) : (1/0)
- guestroom (yes/No) : (1/0)
- basement (yes/no) : (1/0)
- hotwaterheating (yes/no) : (1/0)
- A.C. (yes/no) : (1/0)
- pref area (yes/no) : (1/0)
- furnishing status (furnished/semifurnished/unfurnished) : (1/0.5/0

* this will help in correlation analysis as well as at the time of model building

- Descriptive statistics

- mean (avg. of a column)
- median
- mode
- standard deviation
- variance
- Range

* $\text{mean} = \frac{x_1 + x_2 + \dots + x_n}{n}$, x_i = entries, n → total no. of instances

* median = arrange data in increasing order

If n → odd, $\left(\frac{n+1}{2}\right)^{th}$ entry is median

If n → even, $\left(\frac{n}{2}\right)^{th}$ entry is median.

* mode = most frequent entry in the dataset

Descriptive stats

	Mean	Median	Mode	Range	S.D.
price	4766729.24	434000 ~~4606~~	3500000	11550000	1870440
area	5150.54	~~3.0~~ 4600	6000	14550	2170
bedrooms	2.94 ≈ 3	3.0	3.0	N.N.	N.N.
bathrooms	1.28 ≈ 1	1.0	1.0	—	—
stories	1.80 ≈ 2	2.0	2.0	—	—
mainroad	0.858 ≈ 0.86	1.0	1.0	—	—
guestroom	0.177 ≈ 0.18	0	0	—	—
basement	0.345 ≈ 0.35	0	0	—	—
hotwater	0.045 ≈ 0.05	0	0	—	—
A.C.	0.315	0	0	—	—
parking	0.69	0	0	—	—
pref.area	0.23	0	0	—	—
furnishing		0.5	0.5	—	—
status	N.N.				

(N.N.) : does not make sense for the given parameter

Interpretation

- avg. price is 4766729
- most houses have 3 bedrooms
- most houses are 2 stories
- 86% of houses have main road access
- only 5% of house have ~~A.C.~~ hotwater.
- median price is 434000
- most homes are semifurnished

Q.

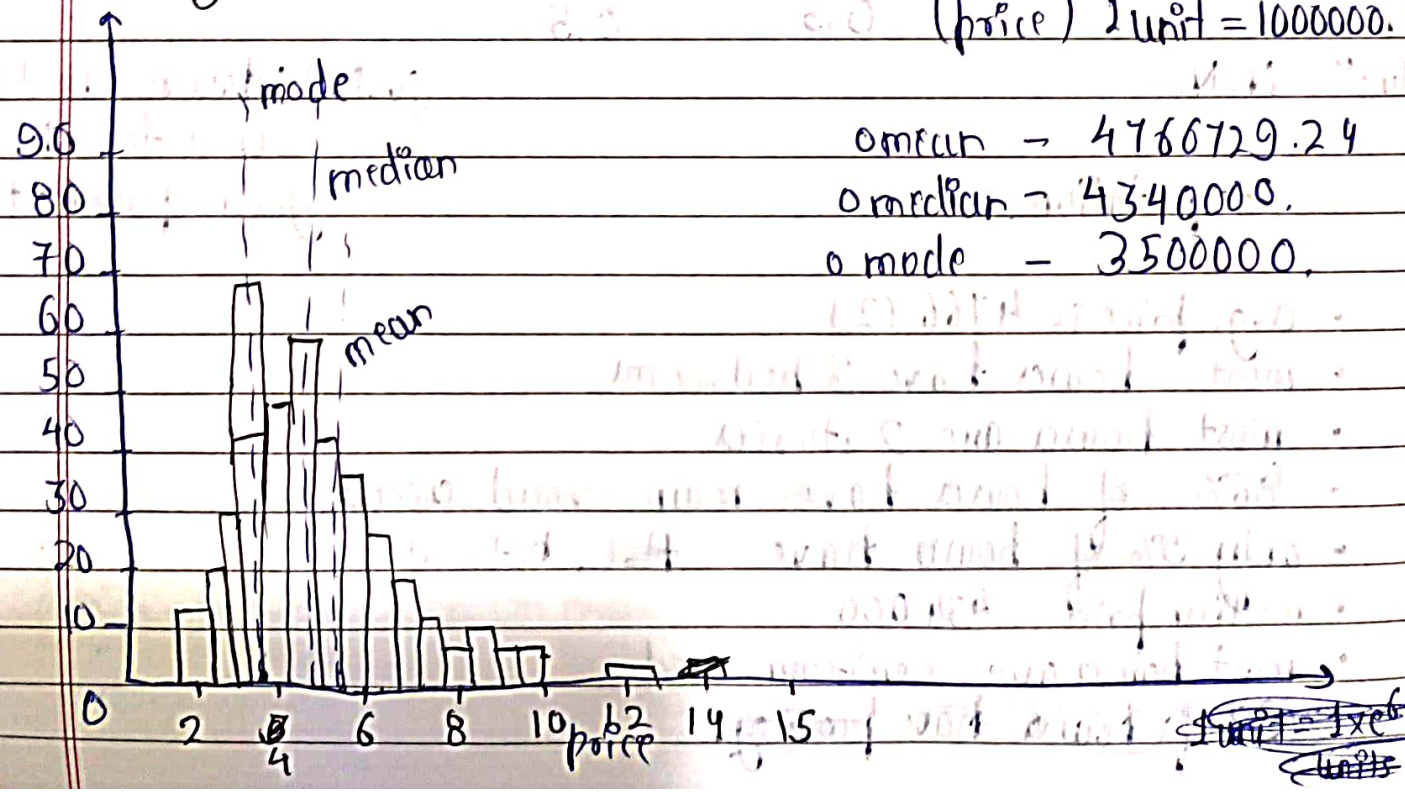
- Max Price of house = ₹ 13300000.
- min Price of house = ₹ 1750000
- max area = 1650 sq.ft
- min. area = 16200 sq.ft
- max bedrooms = 6
- min bedrooms = 1
- range for which price vary is ₹ 11550000.
- range for which area varies is sq.ft (14550)

for Price (mean > median > mode) : suggest positive skewness

Visualization

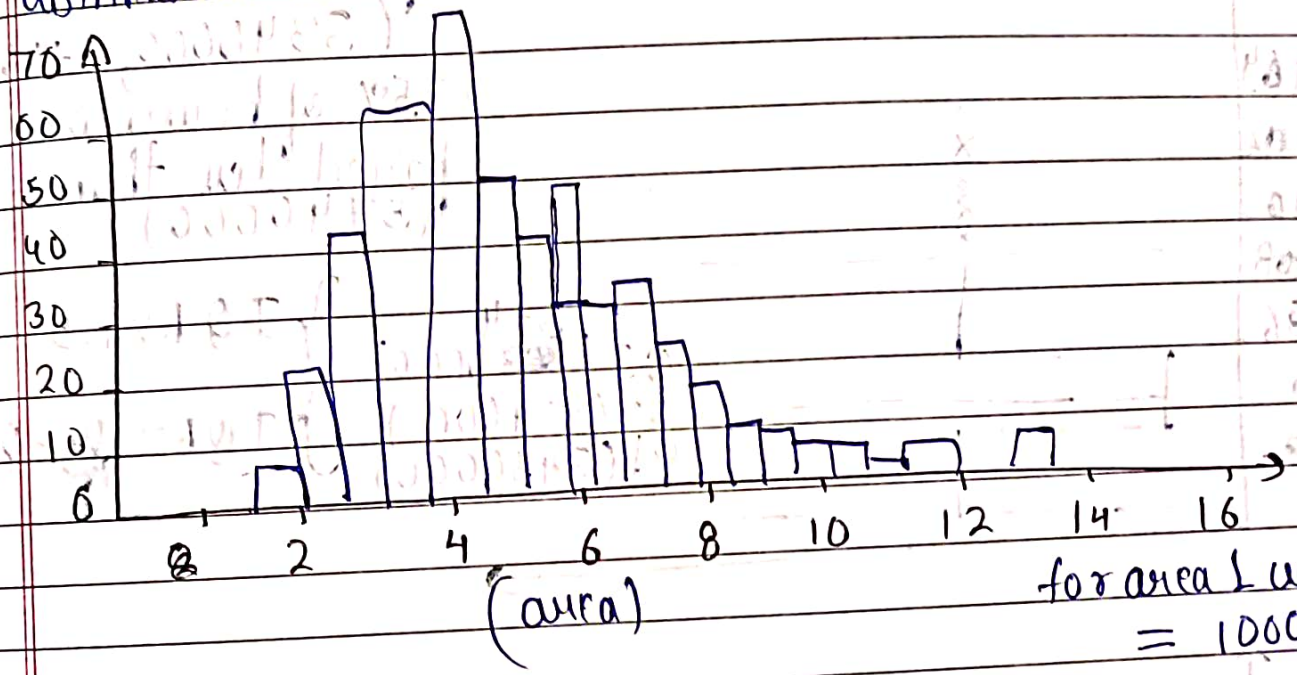
(price) 1 unit = 1000000.

- mean - 4766729.24
- median - 4340000.
- mode - 3500000.



- distribution visulization shows positive skewness i.e. (most of houses are priced at lower ends only few houses are very expensive).

- distribution for area.



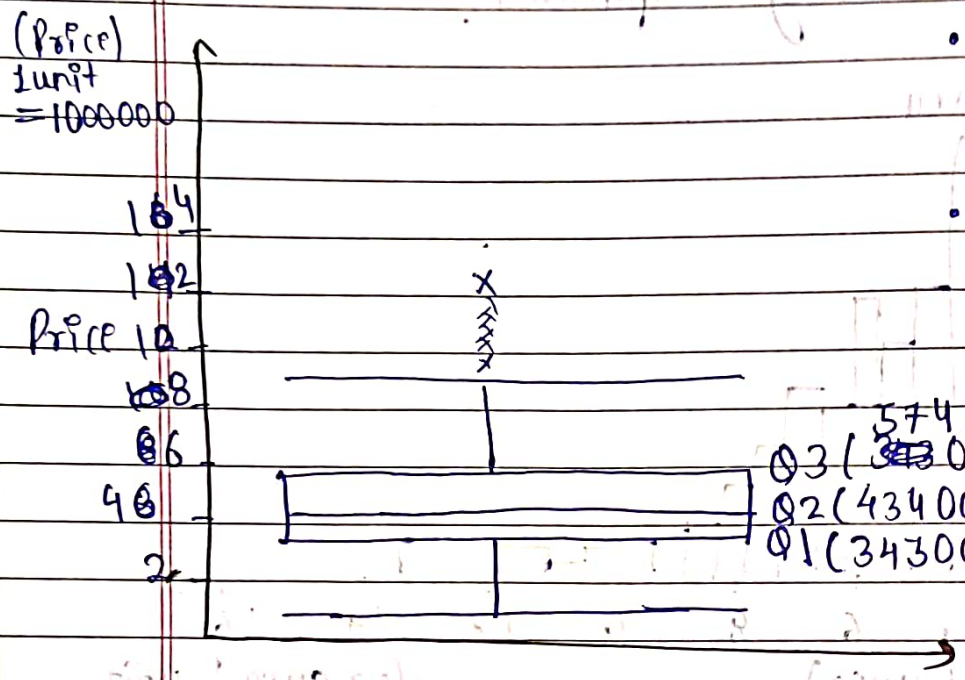
for area 1 unit
= 1000 units

from visualization we can tell it is also +vely skewed.

⇒ most of houses area is around the smaller size.

Visualization for S.D. and IQR.

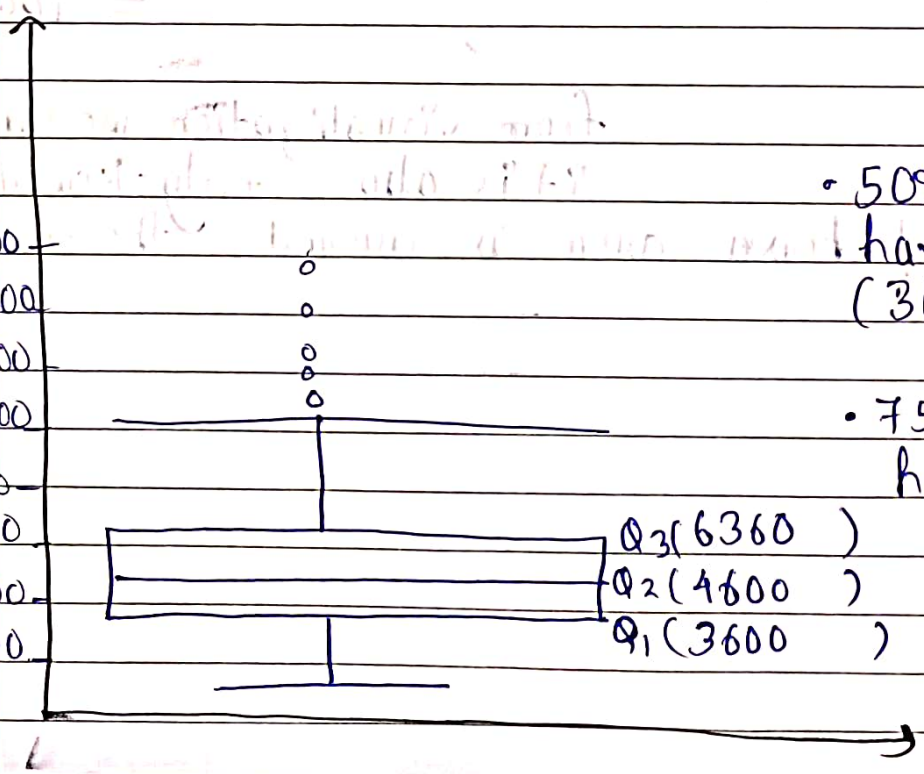
IQR = $Q_3 - Q_1$



- 50% of houses are priced b/w (3340000, 5740000)
- 75% of houses are priced less than (5740000)

} I.Q.R × 1.5

} IQR = $Q_3 - Q_1$



- 50% of houses have area b/w (3600 sq.ft and 6360 ft).
- 75% of houses have area less than (6360).

we see that some points are outside the (1(5xIQB) whiskers

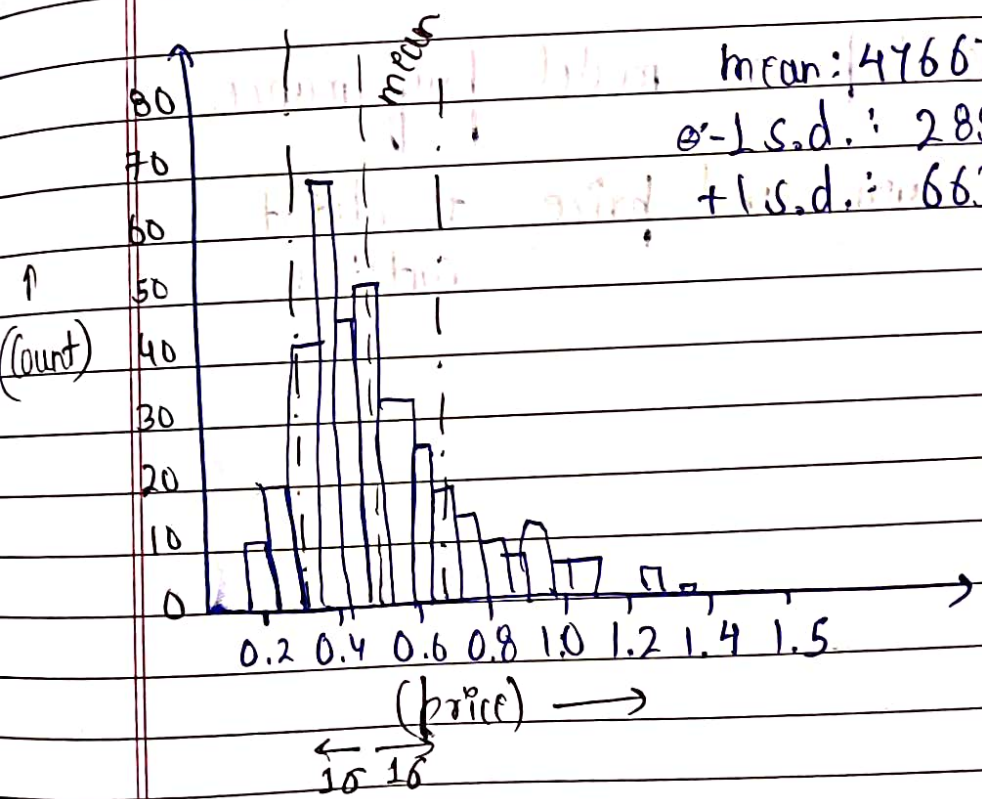
⇒ potential outliers.

which can disrupt model performance.

S.D. - tells about the spread of data about the mean.

price = 1870440

area = 2170



mean: 476672.9

-1 s.d.: 289289.63

+1 s.d.: 663716886.

• s.d. tells about how price is spread about mean price.

Outlier detection

Finding Extreme outliers

- we find (> 3 x IQR) and remove them

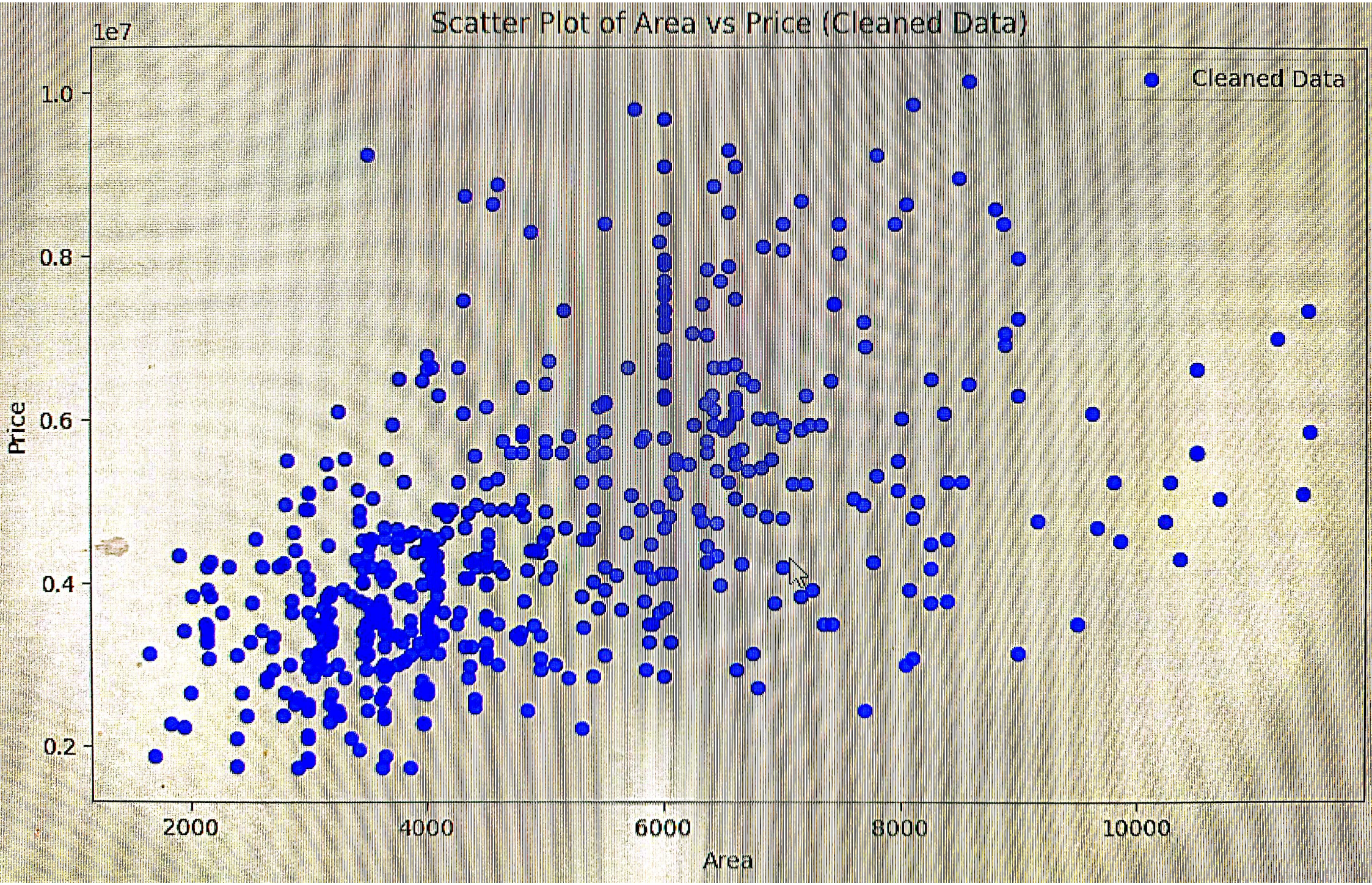
we see that

for Price:

entry with price(₹13300000) is an outlier

for Area is: (16200 sq.ft) and (15600) sq.ft)

we drop them to improve model performance.


Scatter Plot of Area vs Price (Cleaned Data)
1e7
1.0
0.8
0.6
0.4
0.2
Price
Cleaned Data
2000
4000
6000
8000
10000
Area

Using Z-score

we calculate Z-score for potential outliers in price and area

and for $z > 3$ we can say that the points are statistically far enough to consider them outliers.

$$z = \frac{(X - \text{mean})}{\text{standard deviation}}$$

e.g. $z = \frac{13300000 - 4766729.24}{1870440}$

$= 4.56$

$z > 3$, so, it can be considered an outlier

similarly drop these columns

Now, total rows are 532 now.

↳ 13 rows are dropped now.

Correlation matrix

$$f(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

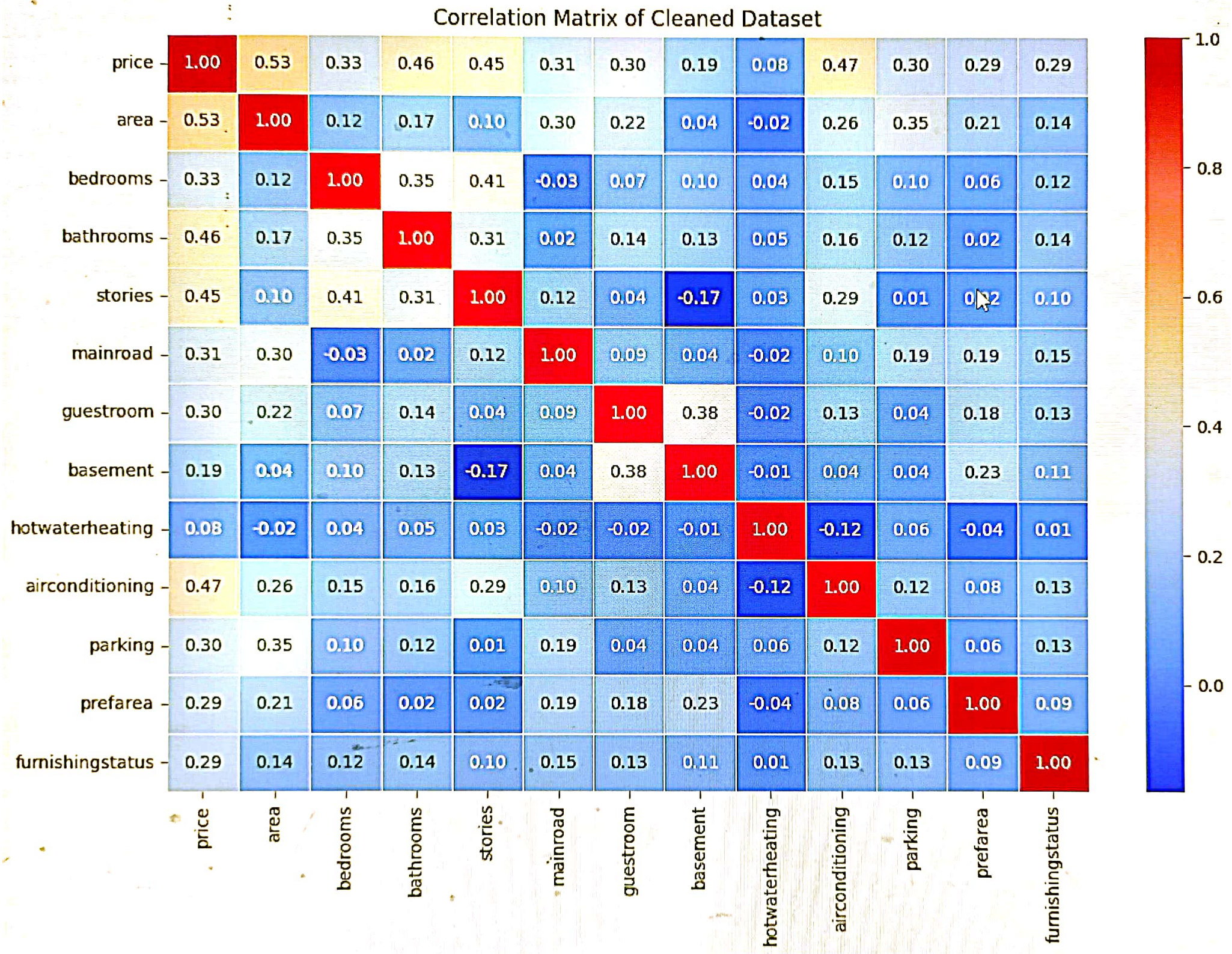
for each pair of features calculate pearson's coefficient.

we see that ~~coefficients with~~ pair with largest coefficients are

* (price, area) ≡ 0.53
* (price, bathrooms) ≡ 0.46
* (price, storig) ≡ 0.45
* (price, A.C) ≡ 0.47
* (bathrooms, storig) ≡ 0.41

we see that correlation found is not strong

most prominent ones are (price, area, A.C, storig)


Correlation Matrix of Cleaned Dataset
price 1.00 0.53 0.33 0.46 0.45 0.31 0.30 0.19 0.08 0.47 0.30 0.29 0.29
area 0.53 1.00 0.12 0.17 0.10 0.30 0.22 0.04 -0.02 0.26 0.35 0.21 0.14
bedrooms 0.33 0.12 1.00 0.35 0.41 -0.03 0.07 0.10 0.04 0.15 0.10 0.06 0.12
bathrooms 0.46 0.17 0.35 1.00 0.31 0.02 0.14 0.13 0.05 0.16 0.12 0.02 0.14
stories 0.45 0.10 0.41 0.31 1.00 0.12 0.04 -0.17 0.03 0.29 0.01 0.12 0.10
mainroad 0.31 0.30 -0.03 0.02 0.12 1.00 0.09 0.04 -0.02 0.10 0.19 0.19 0.15
guestroom 0.30 0.22 0.07 0.14 0.04 0.09 1.00 0.38 -0.02 0.13 0.04 0.18 0.13
basement 0.19 0.04 0.10 0.13 -0.17 0.04 0.38 1.00 -0.01 0.04 0.04 0.23 0.11
hotwaterheating 0.08 -0.02 0.04 0.05 0.03 -0.02 -0.02 -0.01 1.00 -0.12 0.06 -0.04 0.01
airconditioning 0.47 0.26 0.15 0.16 0.29 0.10 0.13 0.04 -0.12 1.00 0.12 0.08 0.13
parking 0.30 0.35 0.10 0.12 0.01 0.19 0.04 0.04 0.06 0.12 1.00 0.06 0.13
prefarea 0.29 0.21 0.06 0.02 0.02 0.19 0.18 0.23 -0.04 0.08 0.06 1.00 0.09
furnishingstatus 0.29 0.14 0.12 0.14 0.10 0.15 0.13 0.11 0.01 0.13 0.13 0.09 1.00
price area bedrooms bathrooms stories mainroad guestroom basement hotwaterheating airconditioning parking prefarea furnishingstatus
1.0 0.8 0.6 0.4 0.2 0.0

1.3 Feature Engineering

feature scaling:

Target: "price" (y)

$$z = \frac{x - \mu}{\sigma}, \quad \mu \rightarrow \text{mean}, \quad \sigma \rightarrow \text{S.D.}$$

normalize the features

feature selection:

* Price (Y): target variable
* area (X_1)
* bedrooms (X_2)
* bathrooms (X_3)
* stories (X_4)
* parking (X_5)

to make the model simple we can combine some features

we combine bedrooms and bathrooms as total rooms.

total_rooms = bedrooms + bathrooms

Final selected features

- X_1 (area)
- X_2 (total rooms)

Linear Regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

β_0: Intercept
β_1: coefficient of area
β_2: coefficient of total_rooms

x_1: area
x_2: total rooms
ε: error term
y: target variable.

Analytical method:

$$\beta = (X^T X)^{-1} X^T y$$

y: target
X: Design matrix

Gradient descent method:

Cost function:

$$J(\beta) = \frac{1}{2m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$J(\beta)$: cost function
y_i: actual input
$\hat{y}_i$: predicted output

Model Training

~~β_0, β~~

Initialize parameters

$$\beta_0 = 0, \quad \beta_1 = 0, \quad \beta_2 = 0$$

Compute gradients

$$\frac{\partial L}{\partial \beta_0} = \frac{-2}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$$

$$\frac{\partial L}{\partial \beta_1} = \frac{-2}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})) x_{1i}$$

$$\frac{\partial L}{\partial \beta_2} = \frac{-2}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})) x_{2i}$$

Iteration one

$$\frac{\partial L}{\partial \beta_0} = 0$$

$$\frac{\partial L}{\partial \beta_1} = -1.0671$$

$$\frac{\partial L}{\partial \beta_2} = -0.9173$$

Update Coefficients $\alpha = 0.01$

$$\beta_0^{(new)} = \beta_0 - (0.01)\frac{\partial L}{\partial \beta_0}$$

$$= 0 - (0.01) \times 0$$
$$= 0$$

$$\beta_1^{(new)} = \beta_1^{(old)} - (0.01) \times (-1.0671)$$

$$= 0.0107$$

$$\beta_2^{(new)} = \beta_2^{(old)} - (0.01) \times (-0.9173)$$

$$= 0.0092$$

Iteration (2)

$$\frac{\partial L}{\partial \beta_0} = 0$$

$$\frac{\partial L}{\partial \beta_1} = -1.0426$$

$$\frac{\partial L}{\partial \beta_2} = -0.8954$$

Update coefficients:

$$\beta_0^{(new)} = \beta_0^{old} - (0.01)\frac{\partial L}{\partial \beta_0}$$

$$= 0 - (0.01) \times 0$$

$$= 0$$

$$\beta_1^{(new)} = \beta_1^{old} - (0.01)(-1.0426)$$

$$= 0.0211$$

$$\beta_2^{(new)} = \beta_2^{(old)} - (0.01)(-0.8954)$$

$$= 0.0181$$

Model evaluation (After convergence using numpy)

mean absolute error: $\frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$

RMSE: $\sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$

model after 2 situations

$$y = 0 + 0.0211(x_1) + 0.0181(x_2)$$

we will calculate for 3 samples.

x_1 (area)
x_2 (total rooms)

(Area) (rooms)
- 7420, 6
- 8960, 8
- 9960, 5

$$\hat{y}_1 = 0 + 0.0211(7420) + 0.0181(6)$$
$$= 156.6706$$

$$\hat{y}_2 = 0 + 0.0211(8960) + 0.0181(8)$$
$$= 189.2008$$

$$\hat{y}_3 = 202.11$$

now,

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$= \frac{1}{3}\left[(13300000 - 156.6706) + (12250000 - 189.2008) + (12250000 - 202.11)\right]$$

= very large value

∴ it is done for very few iterations the value is extremely large after convergence it will decrease.

Final results (using numpy)

$\beta_1 = 0.448$
$\beta_2 = 0.364$

$\beta_0 = -0.0247561 \approx 0$

MAE = 0.604