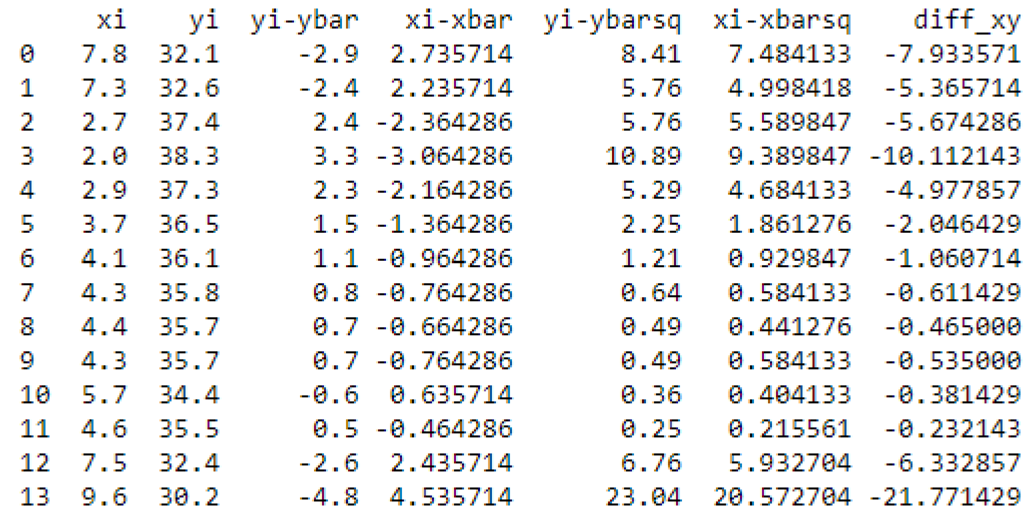
Task 1: Basic statistics

Use the provided data in Table 1 to calculate all statistical results and fill the table

(write and comment your own code)

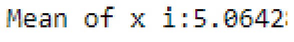


```
      xi    yi  yi-ybar   xi-xbar  yi-ybarsq  xi-xbarsq    diff_xy
0    7.8  32.1     -2.9  2.735714       8.41   7.484133  -7.933571
1    7.3  32.6     -2.4  2.235714       5.76   4.998418  -5.365714
2    2.7  37.4      2.4 -2.364286       5.76   5.589847  -5.674286
3    2.0  38.3      3.3 -3.064286      10.89   9.389847 -10.112143
4    2.9  37.3      2.3 -2.164286       5.29   4.684133  -4.977857
5    3.7  36.5      1.5 -1.364286       2.25   1.861276  -2.046429
6    4.1  36.1      1.1 -0.964286       1.21   0.929847  -1.060714
7    4.3  35.8      0.8 -0.764286       0.64   0.584133  -0.611429
8    4.4  35.7      0.7 -0.664286       0.49   0.441276  -0.465000
9    4.3  35.7      0.7 -0.764286       0.49   0.584133  -0.535000
10   5.7  34.4     -0.6  0.635714       0.36   0.404133  -0.381429
11   4.6  35.5      0.5 -0.464286       0.25   0.215561  -0.232143
12   7.5  32.4     -2.6  2.435714       6.76   5.932704  -6.332857
13   9.6  30.2     -4.8  4.535714      23.04  20.572704 -21.771429
```

Questions:

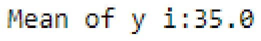
1. Estimate the mean of x and y.

It is calculated by sum of xi /14



```
Mean of x i:5.0642
```

It is calculated by sum of yi /14



```
Mean of y i:35.0
```

2. Calculate the covariance of x and y.

Calculate manually by

Covariance (x,y)=>sigmaxy=>(sum of(xi-xbar)*(yi-ybar))/(1/n) =-5.1

3. Write an implementation with comments to calculate all results.

Yes, I did it.

Task 2: Linear Regression

1. What is the error when the advertisement for TV is 286.0?

I build a univariate linear regression with x as sales and y as TV, then find the predicted y value for the x value 286.0 ie 20.628 then find out the actual value from the data given ie 15.9 So the error yactual-ypredicted is -4.72

2. Find the total Error between regression line and data points?

I find the total error using Mean Absolute Error: 2.5498 and Root Mean squared Error: 3.2423

3. What is linear regression model for TV sales if the advertisement is considered Radio only?

I build model of radio as independent variable and sales as the dependent variable and find the coefficient and intercept and find the Equation for linear regression model as 9.3116+0.2025*radio

4. What is the error when the advertisement for TV is 13.9?

Using the model, I predicted y (sales)value for the x (radio)value 13.9 since we have 2 different y(sales) values for the same x value. The errors when the advertisement for radio is 13.9: 3.973671 and 3.773671

x	yactual	ypredicted	error
13.9	16.1	12.12	3.9
13.9	15.9	12.12	3.7

Then find out the RMSE ie 3.8

5. Find the total error between regression line and data points.

I find the total error using Root Mean squared Error ie 4.2535

6. What is linear regression model for TV sales if the advertisement is considered newspapers only?

I build model of newspaper independent variable and sales as the dependent variable and find the coefficient and intercept and find the Equation for linear regression model:12.3514+0.05469*newspaper

7. Design Multivariate Regression Model?

I build model of 3(TV, radio, newspaper) independent variable and sales as the dependent variable and find the coefficient and intercept and find the Equation of multivariate linear egression:0.04576*TV+0.1885*radio+-0.00103*newspaper+2.9388

Task 3: Clustering

1. Explore the data (calculate the number of rows and columns, number of missing values, outliers, distribution means, medians, quantiles, visualization, etc.)?

Read the cvs file to the dataframe and fine the head of the dataframe using head() then find the number of missing values by .isna().sum(). Find the rows and columns using shape. Create its heat map and from that RAD and TAX are highly correlated to each other and NOX and DIS are highly negatively correlated. To find the outliers I plot the box plot here I can see that B column has the most number of outliers and NOX,INUX,AGE,RAD,TAX columnS has 0 number of outliers. I find the outliers by implementing → the outliers are when value < Q1 - 1.5 x IQR or Q3 + 1.5 x IQR < value, where IQR = Q3 - Q1 (Interquartile range) an Q1 and Q3 first and second quartiles. I find the mean an median using the functions mean() and median() respectively. I find the quantiles by using the function quantile([.1,.25,.5,.75]) . To Visualize MEDV and AGE (as they have the less correlated data) I plot them using scatter plot.

2. Apply K-means, DBSCAN, Hierarchical Clustering, and BIRCH clustering separately?

- K-means Clustering
  To build the K-mean model I used KMeans() and give 3 as no of clusters and I got the predicted clusters grouped as [0,1,2] by using the 2 columns of the data (MEDV and AGE). Then find out the cluster centers and plot it using scatter plot.
  Then we can see that the clusters I got are not oriented to the centers . So I did the data pre processing using MinMaxScaler and then apply the K-means now we can see the clear cluster.
  Then I find out the sum of squared error for Kmeans, Giving k for 1-10 and plot it to find out the appropriate number for k.
  Here I found it like 3.
- DBSCAN Clustering
  I build model using DBSCAN using the 2 column DIS AND MEDV and give epsilon parameter as a random no 0.7 set the minimum sample as 90 and then plot the scatter plot.
- Brich Clustering
  I build model using Brich using the 2 column DIS AND MEDV and give threshold=0.01 and number of clusters as 3 and then plot the scatter plot.
- Hierarchical Clustering
  Using linkage and dendrogram I plot the graph showing number of clusters build model using AgglomerativeClustering() and plot the scatter plot. Here also I used the 2 column DIS AND MEDV

3. Evaluate and compare the results by calculating the silhouette coefficient and Adjusted Rand Index ?

I find out the silhouette score to evaluate the accuracy of all the models using silhouette_score() comparing the values we can say that DBSCAN model has better accuracy , But since the y actual was not provided in the data Adjusted Rand Index was unable to find out.