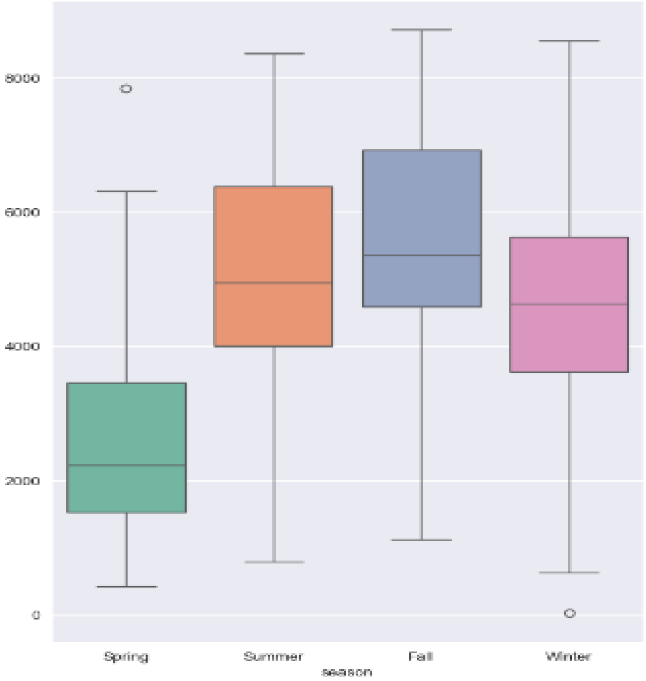
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

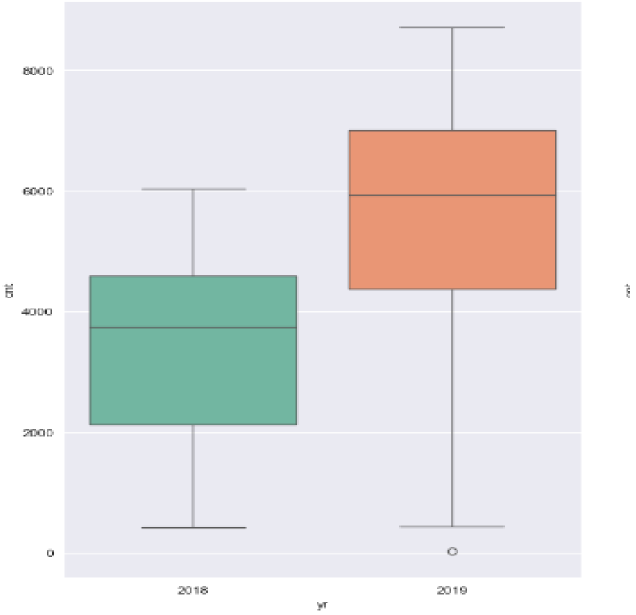
Ans:

Below are the analysis of various categorical variables as below,

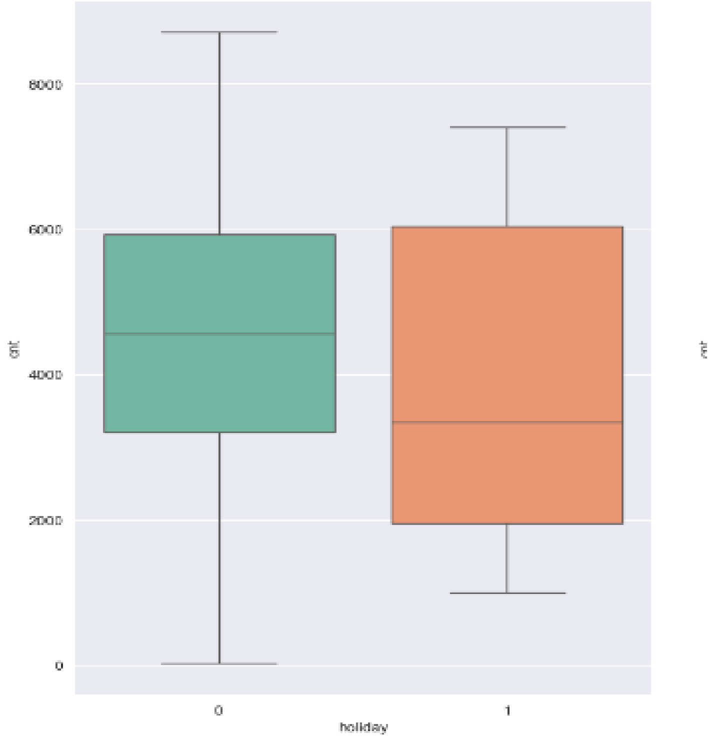
1. Bike rentals are highest in summer and fall, and lowest in spring and winter. Seasonality has a significant impact on the number of bike rentals.



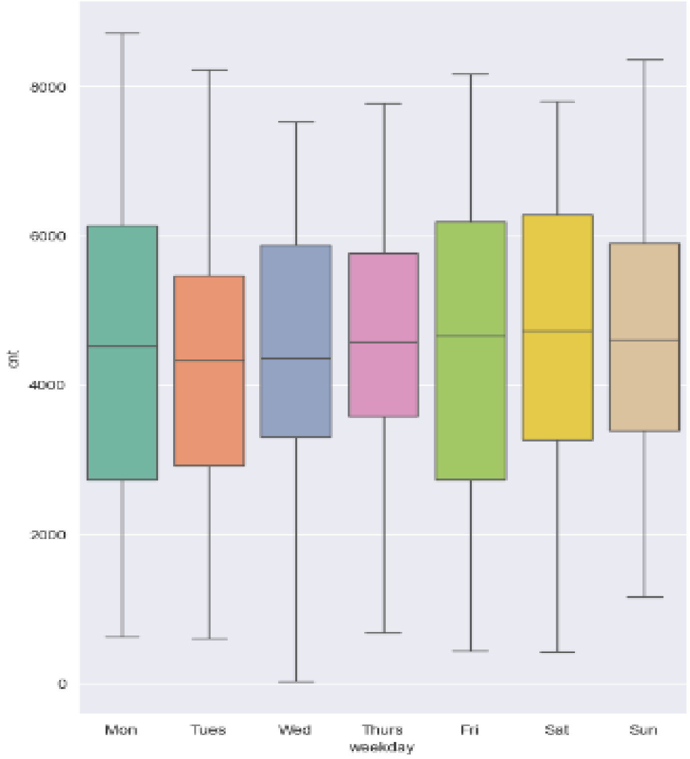
2. There is a noticeable increase in bike rentals in the year 2019, as '1' indicates 2019 from the Data Dictionary, possibly due to the increased popularity of the bike-sharing service.



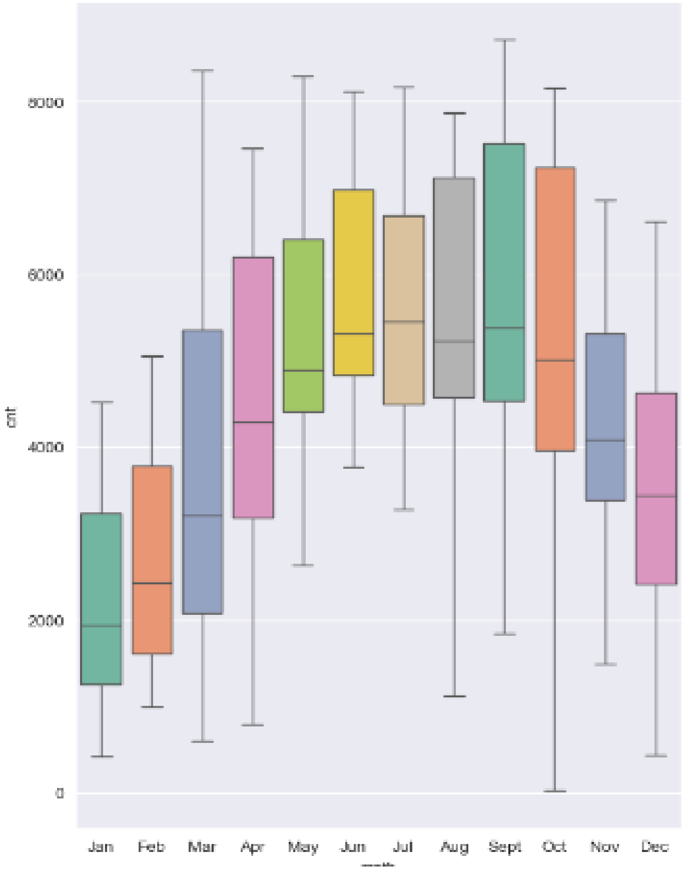
3. Bike rentals are slightly more common on non-holidays, indicating that people may use bikes more for commuting or daily activities rather than on holidays.



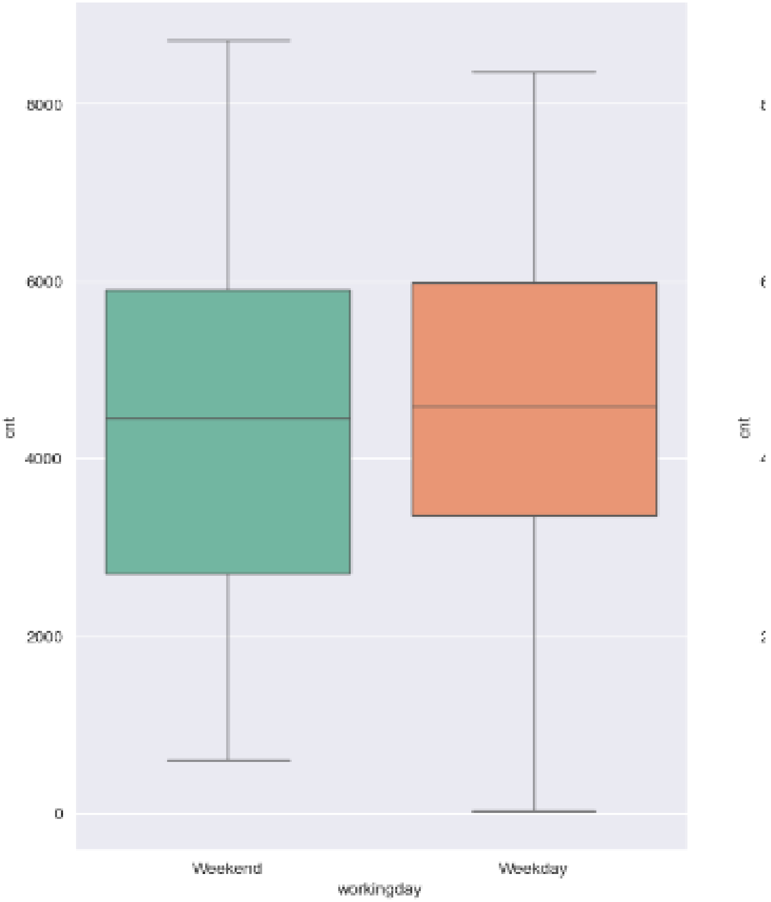
4. Bike rentals appear consistent throughout the week, suggesting that the day of the week is not a strong factor in the rental count.



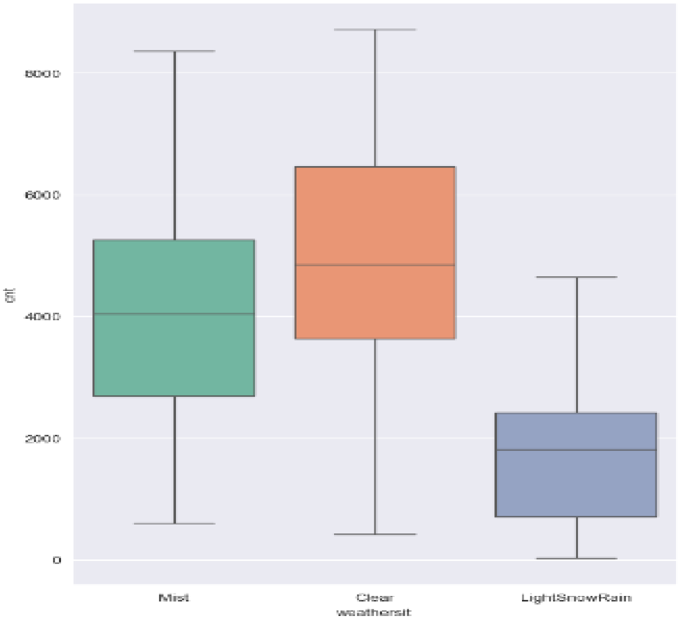
5. Rentals are lower in the winter months (December, January, February) and confirm the seasonal effect, with peak rentals from (March), followed by the summer months (June to September), and start decreasing from October.



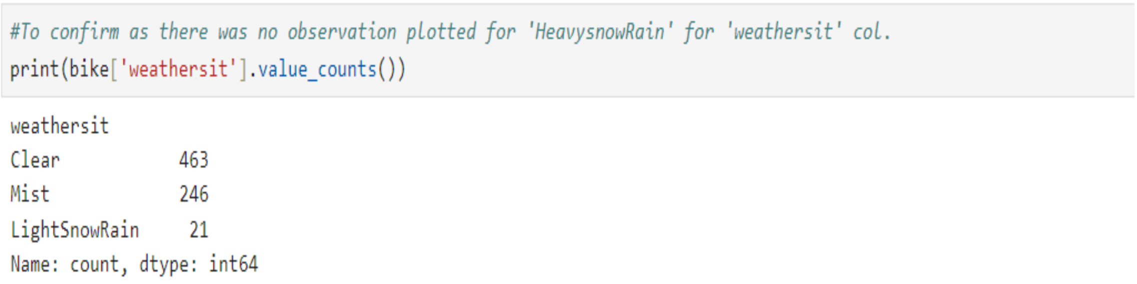
6. We can see maximum bookings happening between 4000 and 6000, the median count of users is almost constant throughout the week. There is no strong difference between working days v/s weekends/holiday, indicating consistent usage for both commuting and leisure.



7. Weather plays a crucial role in bike rentals, with more favourable weather leading to higher rental counts.



8. There were no observations recorded under "HeavySnowRain".



```
#To confirm as there was no observation plotted for 'HeavysnowRain' for 'weathersit' col.
print(bike['weathersit'].value_counts())
```

```
weathersit
Clear           463
Mist            246
LightSnowRain    21
Name: count, dtype: int64
```

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans:

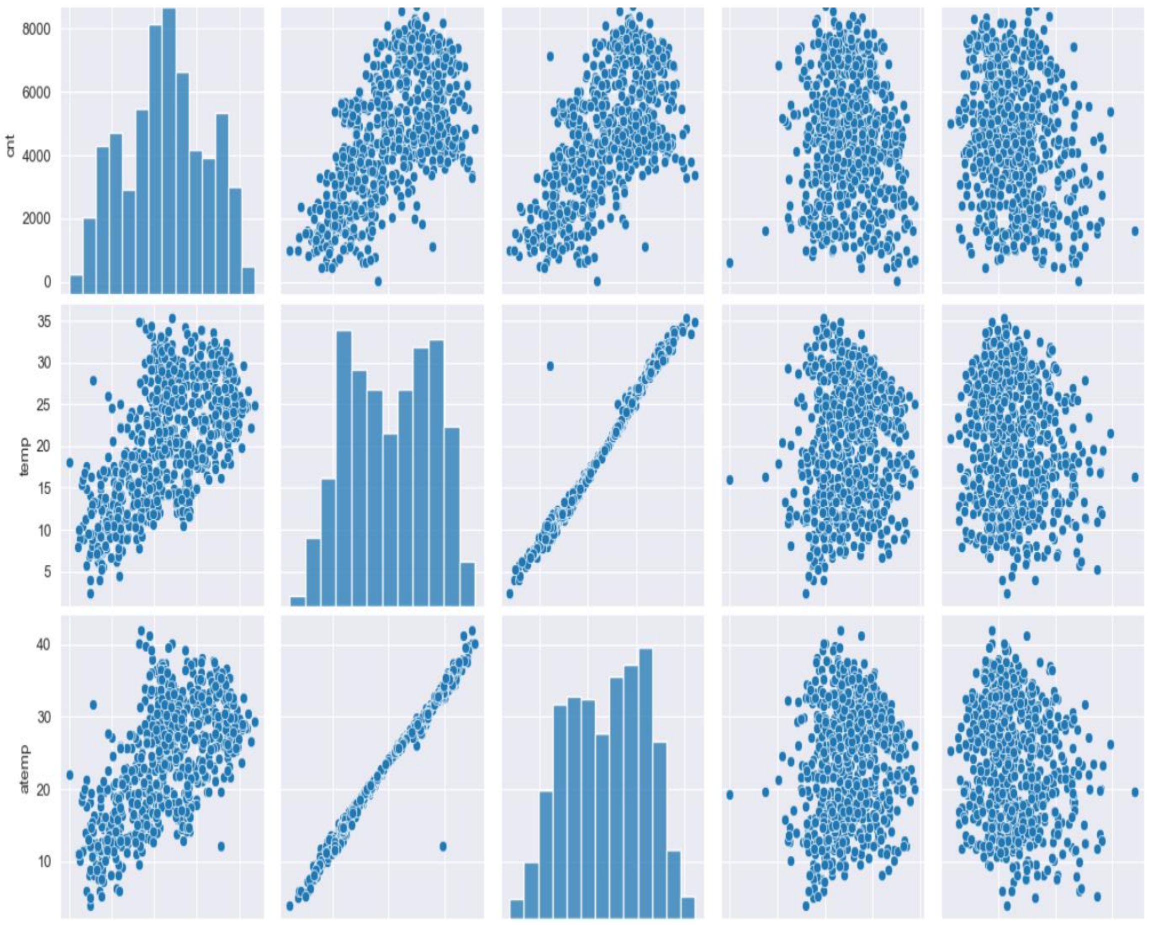
Using drop_first=True during dummy variable creation, prevents multicollinearity by removing one of the dummy variables, which avoids the "dummy variable trap" where the variables become linearly dependent. This ensures the model is well-posed and the coefficients are interpretable.

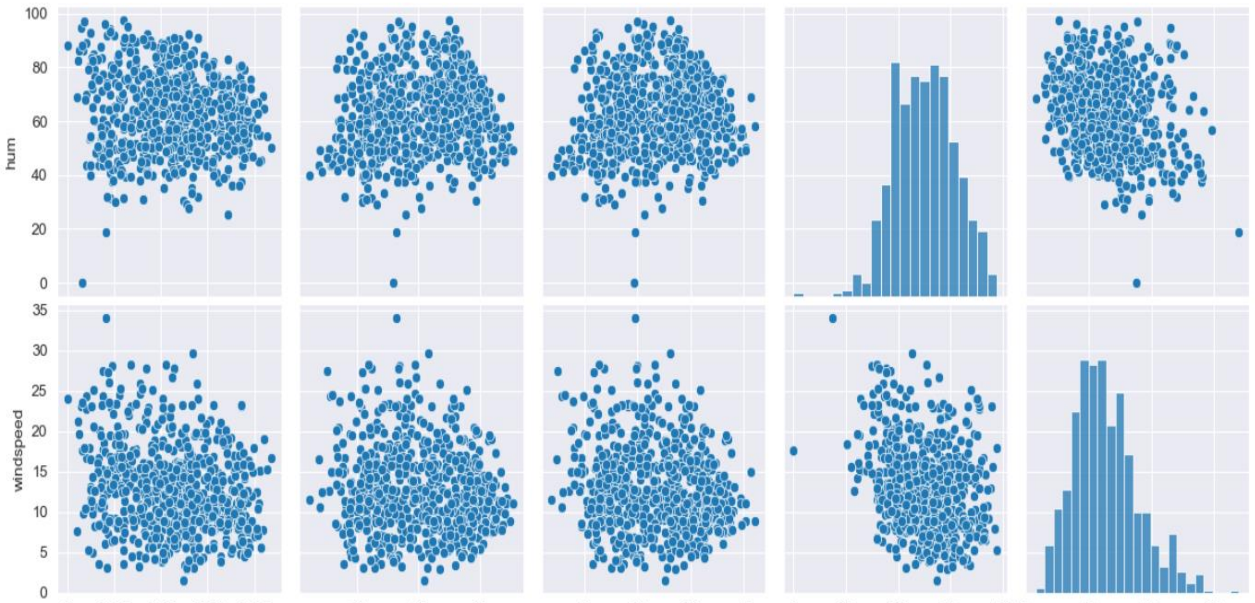
Using one-hot encoding the dummy variables are created to cover the range of values of categorical variable. Each dummy variable have 1 and 0 values. 1 is used to depict the presence and 0 for absence of the respective category. This means if the category variable has 3 categories, there will be 3 dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans:

- The variable "temp" has the highest correlation with the target variable "cnt". This is because the points in the scatterplot are closer to a straight line than any of the other pairs.
- It can also be said that ‘atemp’ and ‘temp’ are linearly correlated and that they move in the same direction.


cnt
8000
6000
4000
2000
0
temp
35
30
25
20
15
10
5
atemp
40
30
20
10

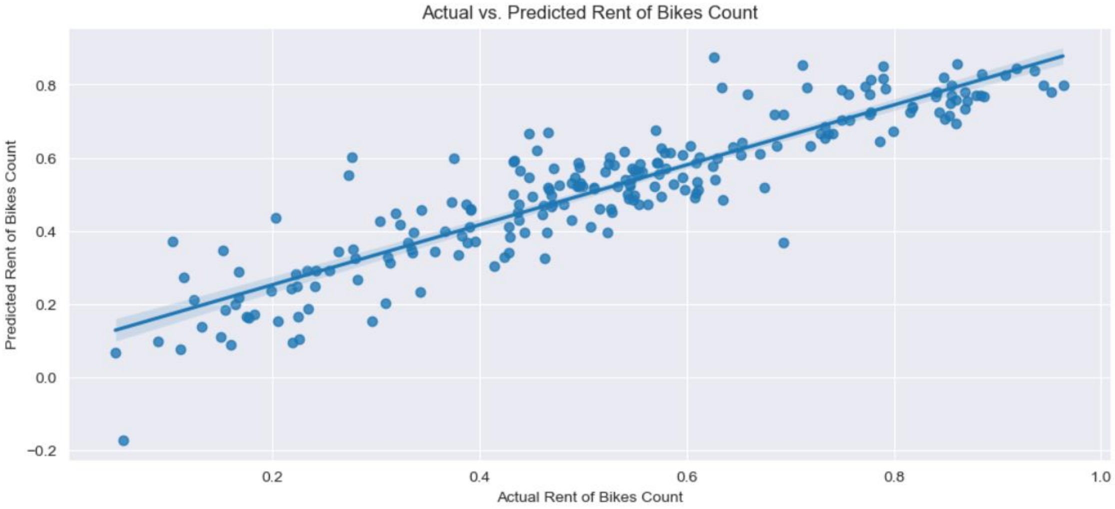

hum
100
80
60
40
20
0
windspeed
35
30
25
20
15
10
5
0
0
2000
4000
6000
8000
cnt
10
20
30
temp
10
20
30
40
atemp
0
25
50
75
100
hum
0
10
20
30
windspeed

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

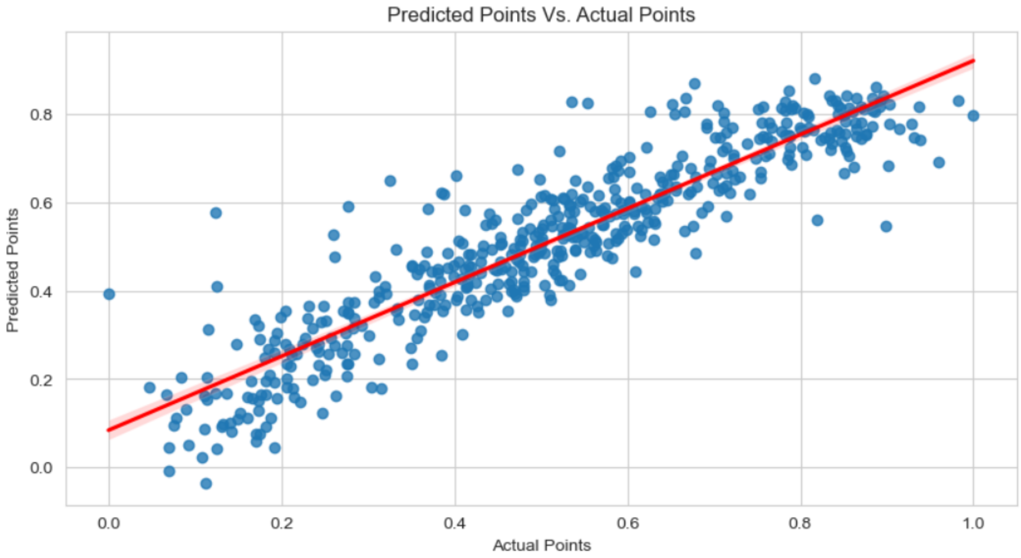
(3 marks)

1. **Ans:**

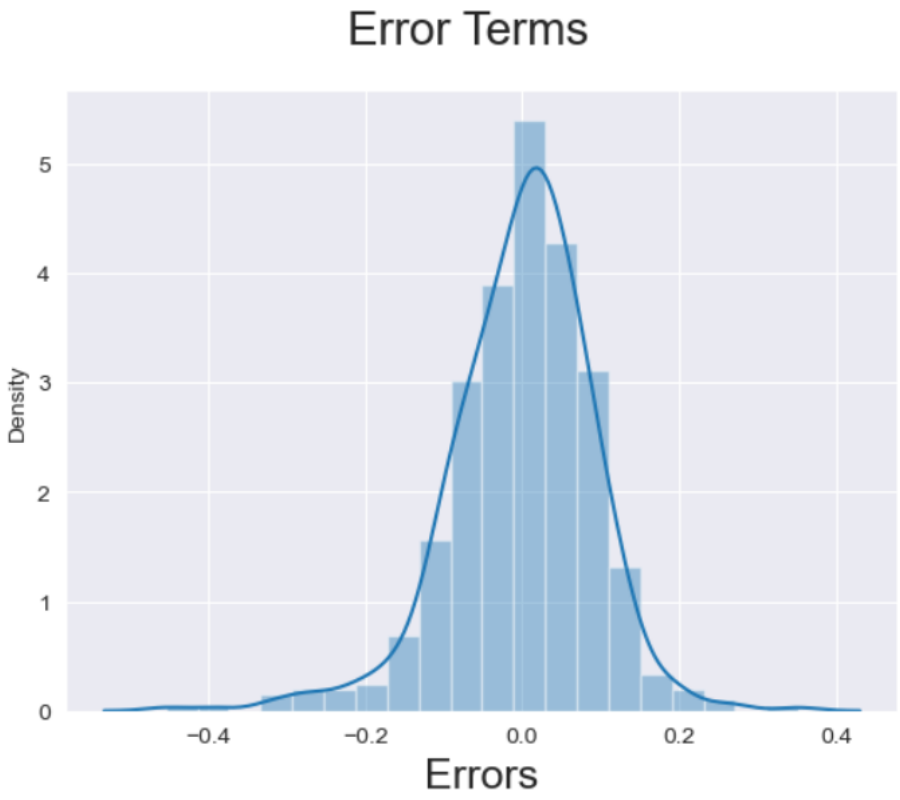
Linear relationship between independent and dependent variables – The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the actual vs predicted plot as shown in the below figure.



2. **Error terms are independent of each other** – We can see there is no specific Pattern observed inthe Error Terms with respect to Prediction, hence we can say Error terms are independent of each other

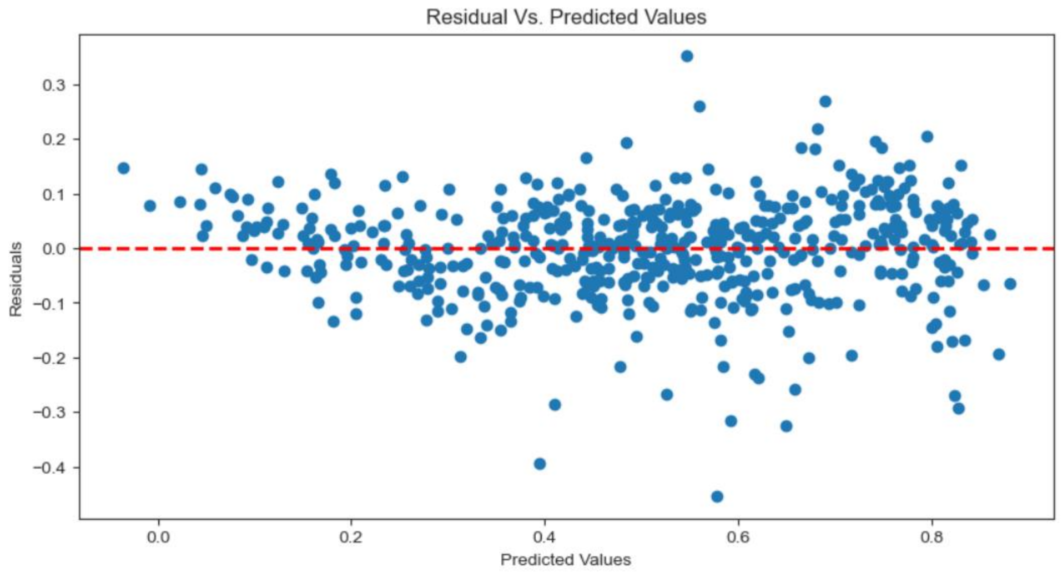


3. Error terms are normally distributed: Histogram and distribution plot helps to understandthe normal distribution of error terms along with the mean of 0. The figure below clearly depicts the same.



4. Error terms have constant variance (homoscedasticity):

We can see Error Terms have approximately a Constant Variance, hence it follows the Assumptionof Homoscedasticity.

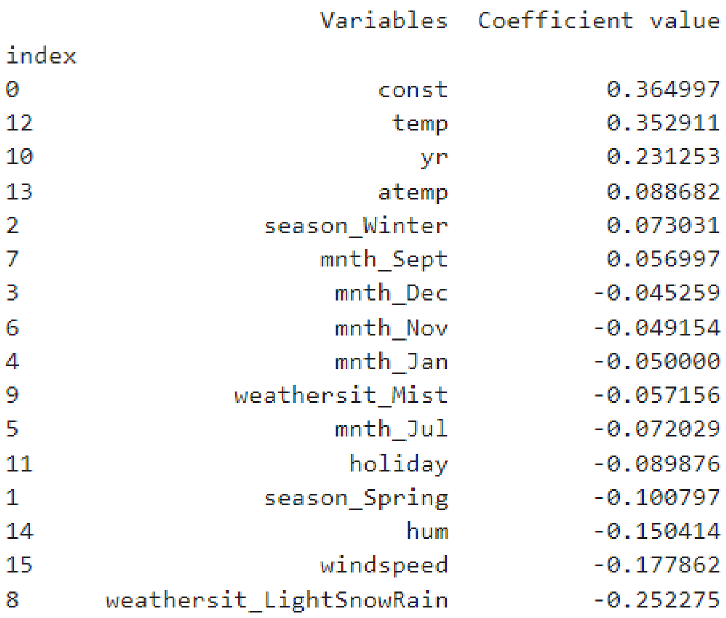


5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: The top 3 variables are:

The significant variables like temperature, year, windspeed, humidity, weather conditions, and seasons provide a strong basis for predicting and understanding the demand for shared bikes.

1. **1.** 'weathersit' : Temperature is the Most Significant Feature which affects the Business positively, Whereas the other Environmental condition such as Raining, Humidity, Windspeed and Cloudy affects the Business negatively.
2. **2.** 'Yr': The growth over the years seems organic given the geological attributes.
3. **3.** season': Winter season is playing the crucial role in the demand of shared bikes.



```
                         Variables  Coefficient value
index
0                            const           0.364997
12                            temp           0.352911
10                              yr           0.231253
13                           atemp           0.088682
2                    season_Winter           0.073031
7                        mnth_Sept           0.056997
3                         mnth_Dec          -0.045259
6                         mnth_Nov          -0.049154
4                         mnth_Jan          -0.050000
9                  weathersit_Mist          -0.057156
5                         mnth_Jul          -0.072029
11                         holiday          -0.089876
1                    season_Spring          -0.100797
14                             hum          -0.150414
15                       windspeed          -0.177862
8          weathersit_LightSnowRain          -0.252275
```

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

A) Linear Regression: A Method for Finding the Best Linear Relationship
Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

- The algorithm uses the best-fitting line to map the association between independent variables with dependent variables.
- There are 2 types of linear regression algorithms,

A) **Simple Linear Regression – A single independent variable is used**.

y=a0+a1x+ε is the line equation for SLR.

Where,

1. a0= It is the intercept of the Regression line (can be obtained by putting x=0)

2. a1= It is the slope of the regression line, which tells whether the line is increasing or decreasing.
3. ε = The error term. (For a good model it will be negligible)

Since MLR is an enhancement of Simple Linear Regression, so the same is applied to the multiple linear regression equation, the equation becomes:

B) Multiple Linear Regression – Multiple independent variables are used.

$Y = \beta 0 + \beta 1X1 + \cdots + \beta pXp + \in$ is the line equation for MLR.

Where,

1. $\beta 0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
2. $\beta 1, \beta 2, \ldots, \beta p = Slope\ or\ the\ gradient.$
3. x1, x2, x3, x4,...= Various Independent/feature variable.
4. ε = The error term.

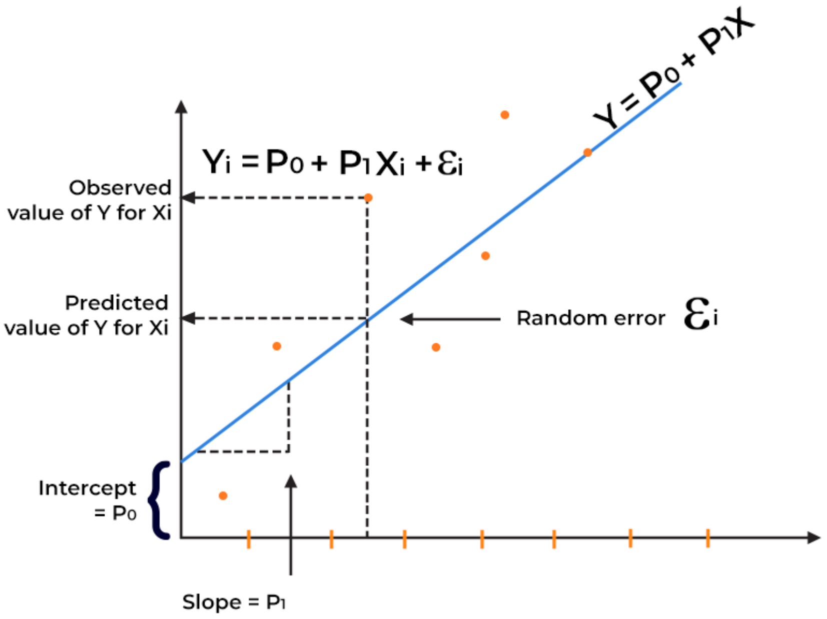


FIG: Below given figure is a classic representation of the Linear Regression Model.

1. **Cost functions**: The cost functions help to identify the best possible values for the $\beta 0, \beta 1, \beta 2, \ldots, \beta p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best-fitting line to predict the dependent variable.
   There are 2 types of cost-function minimization approaches

Unconstrained and Constrained.

1. The sum of squared function is used as a cost function to identify the best

fit line. The cost functions are usually represented as

- The straight-line equation is $Y = \beta 0 + \beta 1X$
- The prediction line equation would be $Ypred = \beta 0 + \beta 1xi$ and the actual Y is as Yi.
- $Now\ the\ cost\ function\ will\ be\ J(\beta 1, \beta 0) = \sum(yi - \beta 1xi - \beta 0)2$

2. The unconstrained minimization are solved using 2 methods
   - Closed form
   - Gradient descent

- While finding the best fit line, we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals.
  To minimize the error squares OLS (Ordinary least square) is used.
  A) $ei = yi - ypred$ is provides the error for each of the data point.
  B) OLS is used to minimize the total e2 which is called as Residual sum of squares.
  RSS=∑i=1n(yi−y^i)2

  where,
  where yi is the observed value, y^i is the predicted value, and n is the number of observations.

Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

2. Explain the Anscombe’s quartet in detail. (3 marks)

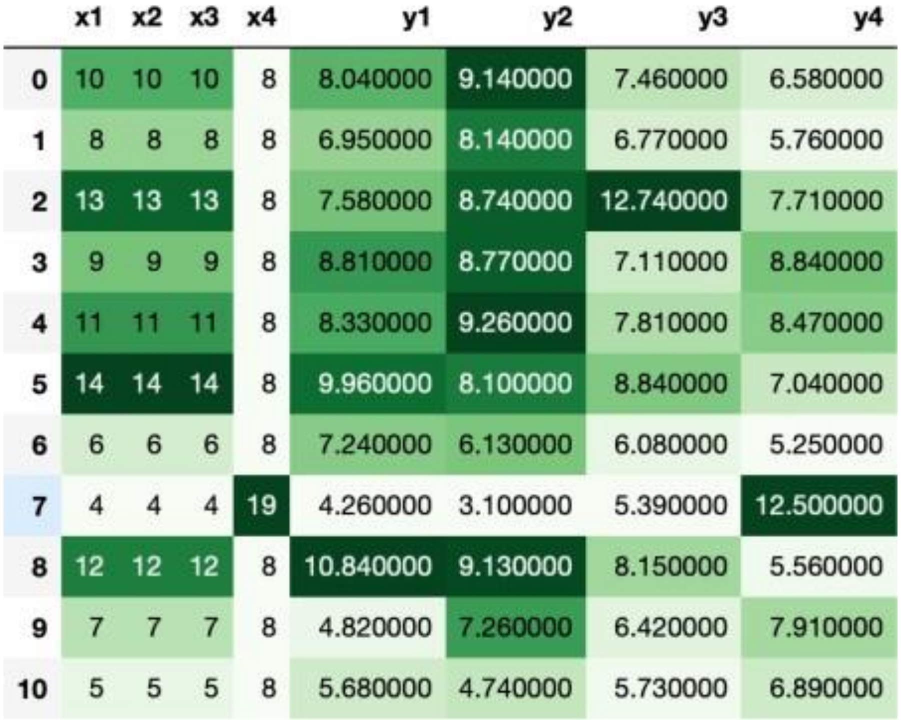
Ans:

Statistics like variance and standard deviation are usually considered good enough parameters to understand the variation of some data without actually looking at every data point. The statistics are great for describing the general trends and aspects of the data.

Francis Anscombe realized in 1973 that only statistical measures are not good enough to depict the data sets. He created several data sets all with several identical statistical properties to illustrate the fact.

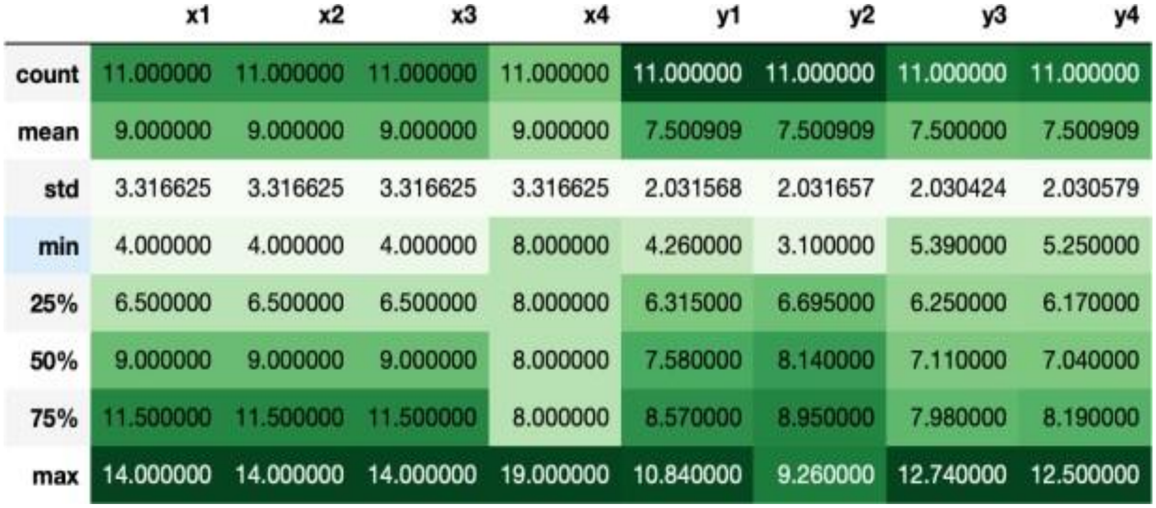
Illustrations:

1) Illustrating one of the examples for the data set as below,



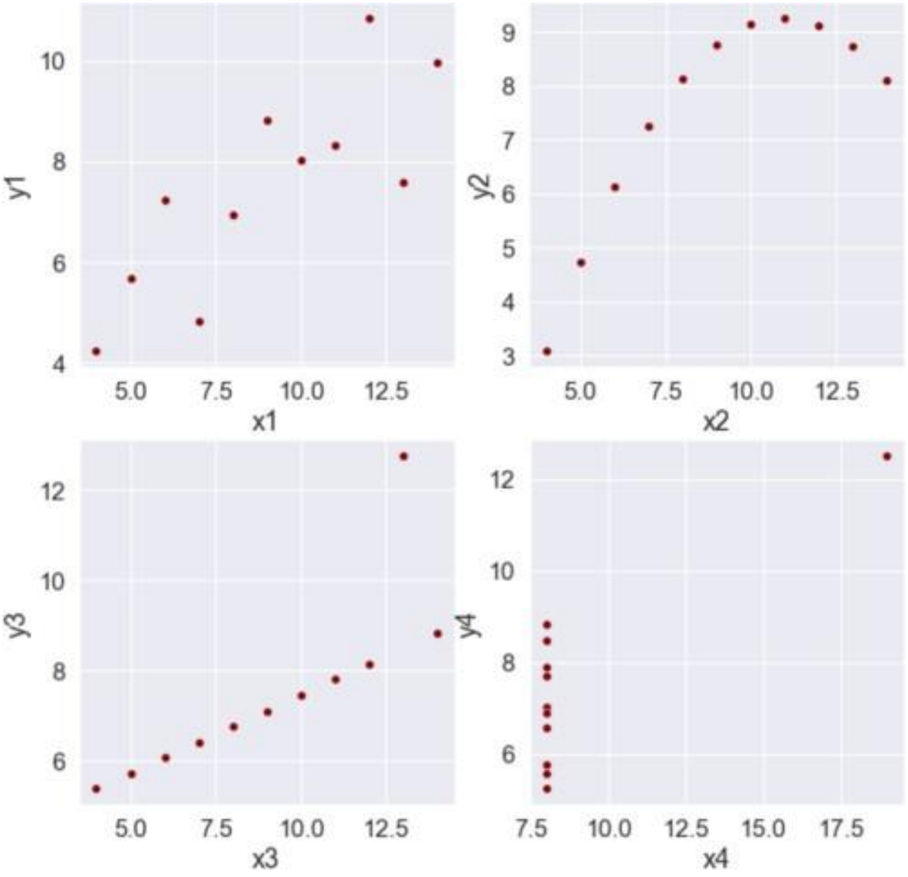
	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.040000	9.140000	7.460000	6.580000
1	8	8	8	8	6.950000	8.140000	6.770000	5.760000
2	13	13	13	8	7.580000	8.740000	12.740000	7.710000
3	9	9	9	8	8.810000	8.770000	7.110000	8.840000
4	11	11	11	8	8.330000	9.260000	7.810000	8.470000
5	14	14	14	8	9.960000	8.100000	8.840000	7.040000
6	6	6	6	8	7.240000	6.130000	6.080000	5.250000
7	4	4	4	19	4.260000	3.100000	5.390000	12.500000
8	12	12	12	8	10.840000	9.130000	8.150000	5.560000
9	7	7	7	8	4.820000	7.260000	6.420000	7.910000
10	5	5	5	8	5.680000	4.740000	5.730000	6.890000

2) If the descriptive statistics are checked for above data set then they all look similar:



	x1	x2	x3	x4	y1	y2	y3	y4
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	9.000000	9.000000	9.000000	7.500909	7.500909	7.500000	7.500909
std	3.316625	3.316625	3.316625	3.316625	2.031568	2.031657	2.030424	2.030579
min	4.000000	4.000000	4.000000	8.000000	4.260000	3.100000	5.390000	5.250000
25%	6.500000	6.500000	6.500000	8.000000	6.315000	6.695000	6.250000	6.170000
50%	9.000000	9.000000	9.000000	8.000000	7.580000	8.140000	7.110000	7.040000
75%	11.500000	11.500000	11.500000	8.000000	8.570000	8.950000	7.980000	8.190000
max	14.000000	14.000000	14.000000	19.000000	10.840000	9.260000	12.740000	12.500000

3) However, when plotting these points, the relation looks completely different as shown below.



- Anscombe's Quartet signifies that multiple data sets with many similar statistical properties could still be different from one another when plotted.
- The dangers of outliers in data sets are warned by the quartet.

  Please refer to the bottom 2 graphs.

  If those outliers would have not been there the descriptive stats would have been completely different in that case.
- Important points
  A) Plotting the data is very important and a good practice before analysing the data.
  B) Outliers should be removed while analysing the data.
  C) Descriptive statistics do not fully depict the data set in its entirety.

3. What is Pearson’s R?

(3 marks)

Ans:

Definition:

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that calculates the linear correlation between two continuous variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation.

The interpretation of the coefficients is as follows:

a. -1 coefficient indicates a strong inversely proportional relationship.
b. 0 coefficient indicates no relationship.
c. 1 coefficient indicates a strong proportional relationship.

$$r = \frac{n(\Sigma x * y) - (\Sigma x) * (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

1) N = the number of pairs of scores
2) Σxy = the sum of the products of

   paired scores.

3) Σx = the sum of x scores
4) Σy = the sum of y scores
5) Σx^2 = the sum of squared x scores
6) Σy^2 = the sum of squared y scores

Example: Suppose we have the following data:

x	y
41	3.2
42	3.3
43	3.4
44	3.5
45	3.6

Calculation: To calculate Pearson's R, we first need to calculate the following values:

- n = 5
- Σx = 215
- Σy = 17
- Σxy = 732
- Σx^2 = 9255
- Σy^2 = 57.9

Then, we can plug these values into the formula:

r = (5(732) - (215)(17)) / sqrt((5(9255) - (215)^2) * (5(57.9) - (17)^2)) r = 1

In this example, the Pearson correlation coefficient is 1, indicating a perfect positive correlation between x and y.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:

Definition:

Scaling is the data preparation step involved in building a regression model. Scaling in linear regression adjusts predictor variables to a common scale, improving model performance and interpretation.

Why: Scaling is performed to ensure that predictor variables have a similar range and variance, which improves the stability and convergence of the regression model, and makes it easier to compare the relative importance of different predictors.
If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also the higher the range the higher the possibility, that the coefficients are impaired to compare the dependent variable variance.

The scaling only affects the coefficients. The prediction and precision of prediction stay unaffected after scaling.

There two ways to perform scaling, below are as follows,

- Normalization/Min-Max scaling – The Min-max scaling normalizes the data within the range of 0 and 1. The Min-max scaling helps to normalize the outliers as well.

$$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Standardization converges all the data points into a standard normal distribution where the mean is 0 and the standard deviation is 1.

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)

Ans:

Definition:

The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors.

A VIF value becomes infinite when there is perfect multicollinearity among the predictors, meaning one predictor is a perfect linear combination of others.

$$VIF = \frac{1}{1 - R^2}$$

The VIF formula, signifies when the VIF will be infinite. If the R^2 is 1, then the VIF is infinite. The reason for R2 to be 1 is that there is a perfect correlation between 2 independent variables.

Example: Consider two predictors X1, X2 and X3 in a regression model , where X1 + X2 is exactly X3.

Then, the VIF for X3 will be infinite, because it cannot be estimated uniquely without X1 and X2.

Infinite VIF occurs due to perfect multicollinearity among predictor variables.

.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

Ans:

Definition: Q-Q plots, or quantile-quantile plots, are a graphical representation used to compare the distribution of two data sets(train and test), which can be normal, exponential, or uniform.

A Q-Q plot is a scatter plot that compares the quantiles of two datasets. The x-axis represents the quantiles of a theoretical distribution (usually the standard normal distribution), and the y-axis represents the quantiles of the observed data. The points on the plot are the paired quantiles from both distributions.

A. Use :

1. **Linear regression**: Q-Q plots are essential in linear regression to check the normality of residuals and validate the model.
2. **Data exploration:** Q-Q plots can be used in data exploration to understand the distribution of a variable and identify potential issues.
3. **Advantages:**
   a. They reveal aspects of the distribution, such as location, scale shifts, symmetry changes, and outliers, in a single plot.
   b. The plot can also indicate the sample size.

B. Importance of a Q-Q plot in linear regression:

A Q-Q plot is important in linear regression because it helps to:

1. **Validate model assumptions**: By checking the normality of residuals, a Q-Q plot helps to ensure that the linear regression model is valid and reliable.
2. **Improve model performance**: Identifying and addressing non-normality or outliers in the residuals can lead to a better-fitting model and more accurate predictions.
3. **Prevent misinterpretation of results**: A Q-Q plot can help prevent misinterpretation of results by highlighting potential issues with the residuals, which can affect the validity of the regression model.

C. Examples: A company produces two types of batteries, A and B. The company claims that the lifetimes of both batteries are normally distributed with the same mean and standard deviation. However, some customers have complained that battery A has a shorter lifetime compared to battery B.

Battery A	**Battery B**
50, 60, 70, 80, 90, and so on,	55, 65, 75, 85, 95, and so on.

-First, we need to sort the data in ascending order for Battery A: (data points) and Battery B: (data points)
-Next, we create a Q-Q plot using the sorted data
(This was just an example of a Q-Q plot)

D. Interpretations:

- **Similar distribution:** Data points lie around a straight line at a 45-degree angle from the x-axis.
- **Y values < X values:** Y-values are lower than X-values.
- **X values < Y values:** X-values are lower than Y-values.
- **Different distributions:** Data points deviate from the straight line.

By using Q-Q plots, we can gain a deeper understanding of our data, validate our models, and make more accurate predictions.

*******************THANKYOU*******************