Assignment-based Subjective Questions

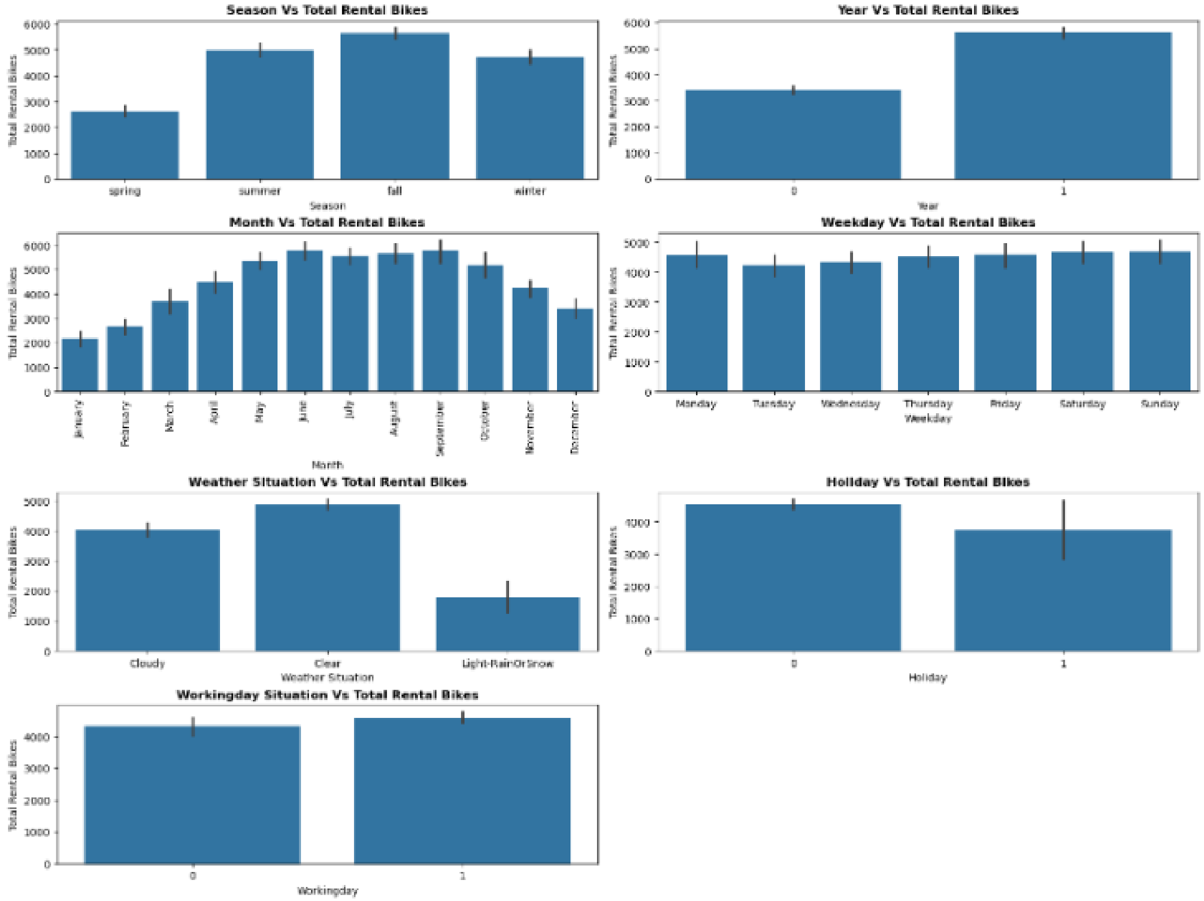
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?   (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Below is the effect of categorical variables season,yr,mnth,holiday,weekday,workingday, weathersit,temp,atemp,hum,windspeed on rental count

- Rentals are more during fall season
- Rentals are more on a day with Clear weather
- Rentals are more on a holiday day
- June and September have more rentals
- Rentals are little more on a weekday which is neither weekend nor holiday
- Rentals are distributed same across all the weekdays except Saturday where there is an increase compared to other days
- Growth is rentals between 2018 and 2019



Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit) drop_first=true will drop the first dummy variable and will give n-1 dummies out of n discrete categorical levels which will reduce correlation among dummy variables.

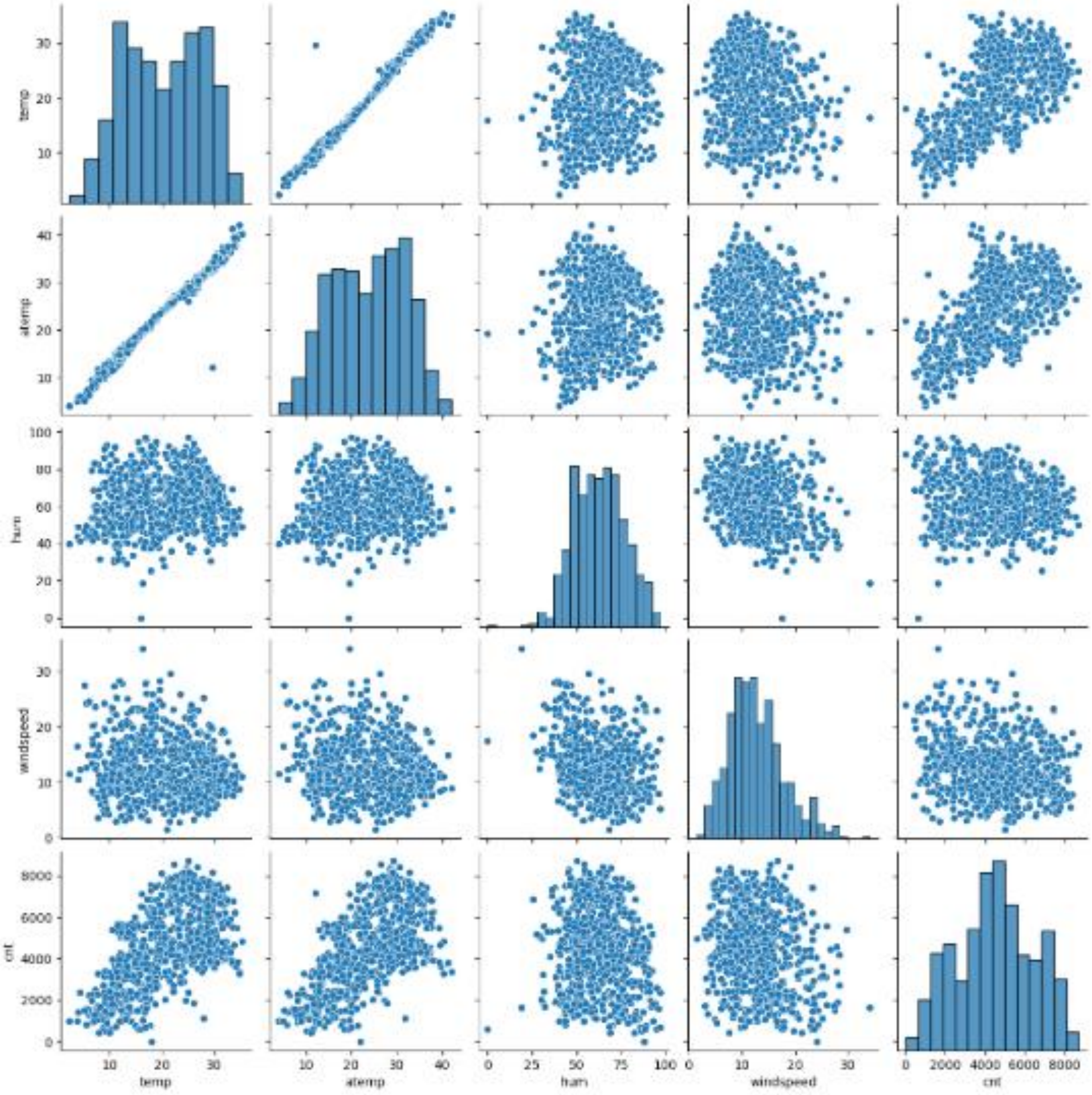
Ex. A Categorical column has 3 values furnished, semi-furnished and unfurnished. If one variable is furnished and other is semi-furnished, then it is obvious the other value is unfurnished. So, we do not need 3rd variable to identify the unfurnished

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

temp and atemp are highly correlated with target variable cnt



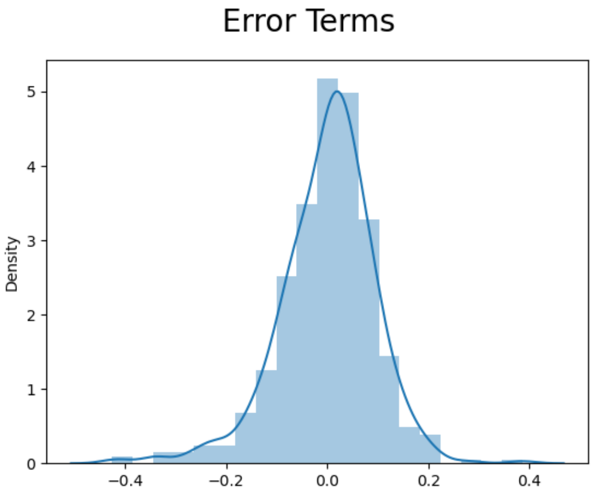
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

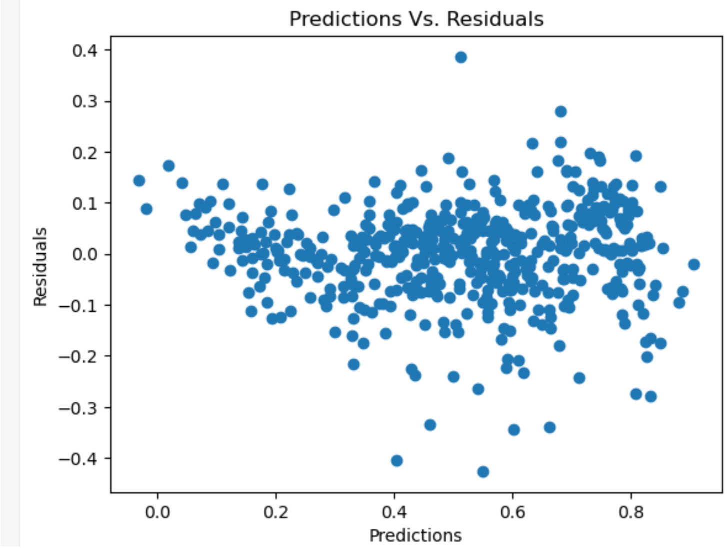
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Distribution of the error terms - The residuals are following the normally distributed with a mean

0.



Looking for patterns in the residuals - the points are plotted in a randomly spread, there is no pattern and points are not based on one side so there is no problem of heteroscedasticity.



Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)
Temp,yr and winter are the top 3 features contributing significantly towards explaining the demand of shared bikes. These features are picked based on the coeff values

---

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is also a type of supervised machine-learning algorithm that learns from the labelled datasets and maps the data points with most optimized linear functions which can be used for prediction on new datasets. It computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation with observed data. It predicts the continuous output variables based on the independent input variable.
For example, if we want to predict house price, we consider various factor such as house age, distance from the main road, location, area and number of rooms, linear regression uses all these parameters to predict house price as it considers a linear relation between all these features and price of house.
Linear regression is of the 2 types:

1. **Simple Linear Regression:** It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

Simple Linear Regression is defined as: Y=β0+β1X1 +ϵ where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope
- ϵ is standard error

2. **Multiple Linear Regression:** It explains the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a 'hyperplane' instead of a straight line. This is an extension of Simple linear regression

Multiple Linear Regression is defined as Y=β0+β1X1+β2X2+…+βnXn+ϵ where:

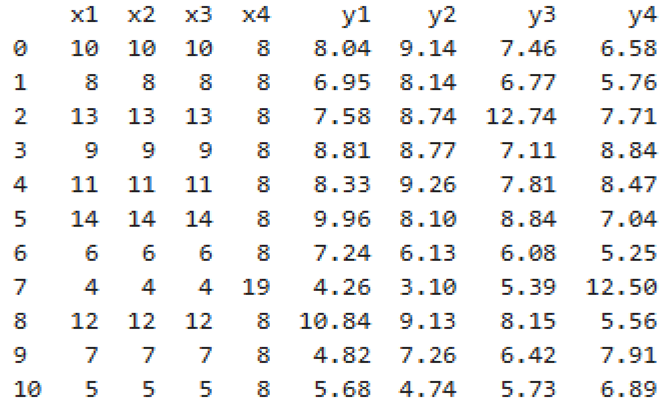
- Y is the dependent variable
- X1, X2, …, Xn are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the slopes
- ϵ is standard error

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

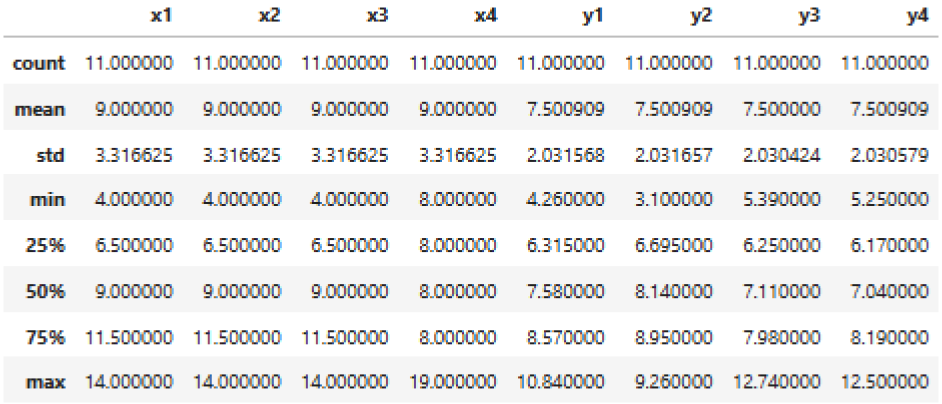
Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.



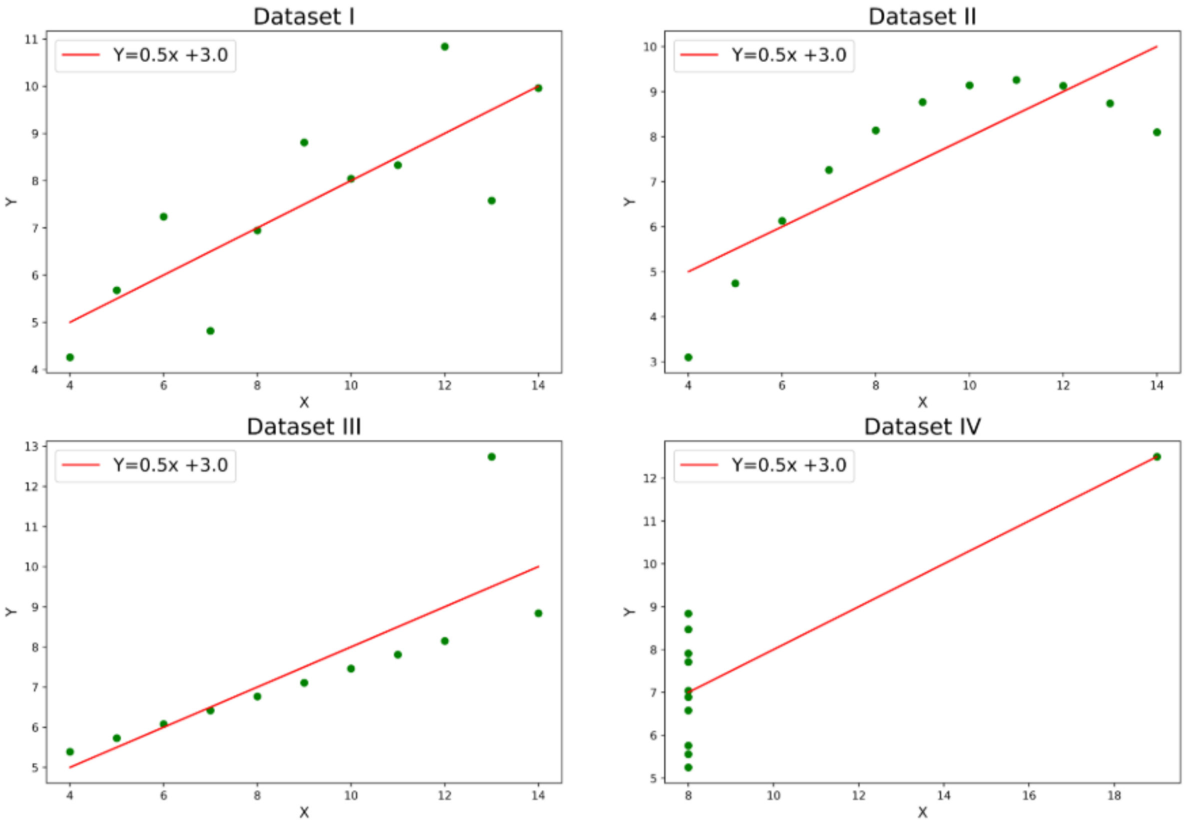
```
    x1  x2  x3  x4     y1    y2     y3     y4
0   10  10  10   8   8.04  9.14   7.46   6.58
1    8   8   8   8   6.95  8.14   6.77   5.76
2   13  13  13   8   7.58  8.74  12.74   7.71
3    9   9   9   8   8.81  8.77   7.11   8.84
4   11  11  11   8   8.33  9.26   7.81   8.47
5   14  14  14   8   9.96  8.10   8.84   7.04
6    6   6   6   8   7.24  6.13   6.08   5.25
7    4   4   4  19   4.26  3.10   5.39  12.50
8   12  12  12   8  10.84  9.13   8.15   5.56
9    7   7   7   8   4.82  7.26   6.42   7.91
10   5   5   5   8   5.68  4.74   5.73   6.89
```

Statistical summary.



	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

Scatter plot and linear regression line for each dataset



Conclusion:

- Anscombe's Quartet exhibits diverse patterns in scatter plots, illustrating the importance of visualizing data for meaningful insights beyond numerical summaries.
- Reveals limitations of summary statistics, emphasizing the need for visual exploration to detect nuances, outliers, and diverse relationships in datasets.
- Underscores that numerical summaries alone can be misleading, emphasizing the crucial role of data visualization in uncovering patterns and outliers.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

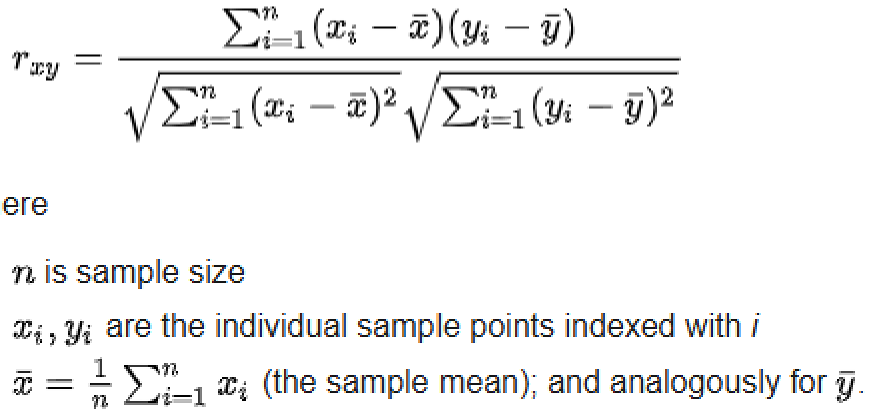
<Your answer for Question 8 goes here>

Pearson Correlation Coefficient (r), often denoted as r, measures the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

- r = 1: Perfect positive linear relationship
- r = -1: Perfect negative linear relationship
- r = 0: No linear relationship between the variables.

Pearson correlation coefficient, when applied to a sample, is commonly represented by $r_{xy}$and may be referred to as sample Pearson correlation coefficient.

Given paired data {(x1,y1),…,(xn,yn)} consisting of n pairs, r_{xy} is defined as below



$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$.

Pearson correlation draws a line of best fit through two variables, indicating the distance of data points from this line. A 'r' value near +1 or -1 implies all data points are close to the line. An 'r' value close to '0' suggests data points are scattered around the line.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

What is scaling?

- Feature Scaling is a technique to standardize the independent features present in the data.
- It is performed during the data pre-processing to handle highly varying values.

Why is scaling performed?

- Scaling is performed during the data pre-processing to handle highly varying values.
- If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling.
- Scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.

What is the difference between normalized scaling and standardized scaling?

- 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 mean is 0 and standard deviation is 1.

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

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

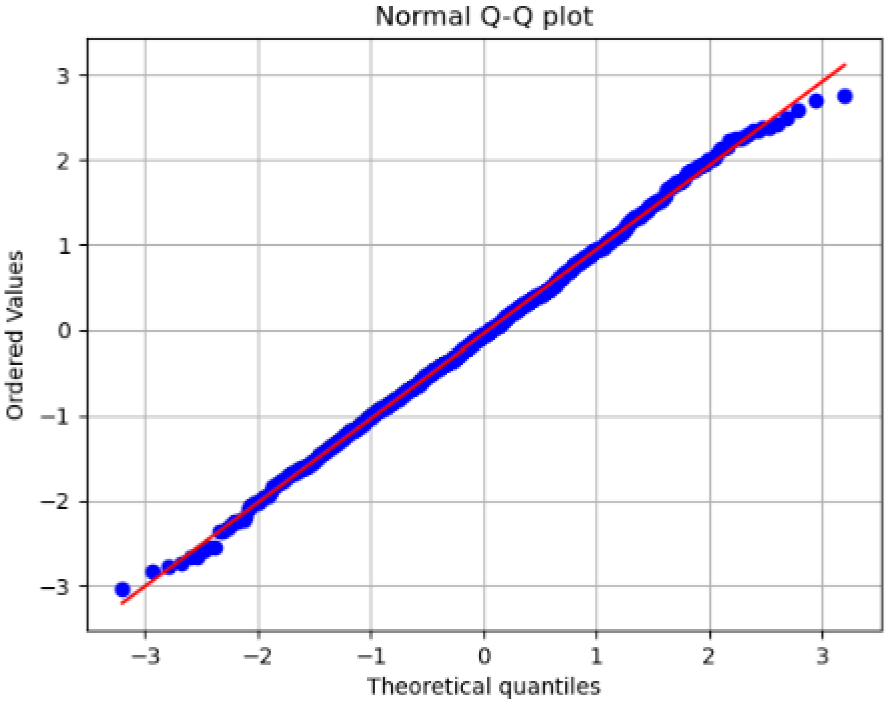
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>
A VIF value of infinity occurs when there is a perfect correlation between two or more independent variables in a regression model. When the correlation between two variables is exactly 1 or -1, the VIF becomes infinite because the denominator in the VIF calculation becomes zero.

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

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>
The quantile-quantile (q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.
A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.



QQ plot is used to check following scenarios:
If two data sets -

i. come from populations with a common distribution
ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior