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)

The dataset 'day' has the following categorical variables:

'**season**', '**yr**' (an acronym used for the years), **mnth**' (an acronym used for the months), **holiday**', '**weekday', 'workingday', 'weathersit'** (an acronym used for the weather type), **'hum' (**an acronym used for humidity)

- **Season:** (1:'spring', 2:'summer', 3:'fall', 4:'winter'). Demands for Bike is highest in Fall and lowest in Spring.

- **Yr**: (0: 2018, 1: 2019). Two years of data are available. Almost 65% growth from the year 2018 to the year 2019. This business concept was gaining popularity

- **Mnth**: (1 to 12 respectively for January to December). More in the year 2019 (1) than in the year 2018 (2) for each month of the year. Also highest in August and September.

- **Holiday**: We dropped this variable since if it is a 'working day', then it is not a holiday and if it is a holiday then it is not a working day. Hence it was redundant.

- **Workingday**: Demands for Rental Bikes:
  - is high on working days when compared to non-working days for registered users
  - is almost the same for working days and non-working days for casual users
  - Total is high on working days when compared to non-working days due to registered users

- **Weathersit**: (1: 'Clear or Few clouds', 2: 'Mist and Few Clouds', 3: 'Light Snow and Light Rain', 4: 'Heavy Rain + Heavy Snow')
  - Rental count is low when there is rain or snow
  - Rental count is high when there is a clear or few clouds
  - No data available for Heavy Rain + Heavy Snow

- **Weekday**: (0 to 6 respectively for Sunday to Saturday)
  - Total Rentals have an almost similar count on all 7 days of the week
  - On weekends (Saturday and Sunday), Casual count is more than that on weekdays
  - On weekends (Saturday and Sunday), the Registered count is less than that on weekdays

- **Hum**: Scientific fact: Temperature and humidity have an inverse relationship, meaning that as the temperature increases, the relative humidity decreases. Hence this variable dropped as it is redundant.

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)

The dummy variables are made to span the range of values of the categorical variable using one-hot encoding. The values of each dummy variable range from 1 to 0. The corresponding category's presence is represented by a 1 and its absence by a 0. There will be K dummy variables if the category variable has K categories.

To drop the base/reference category while constructing dummy variables, use drop_first = True. This prevents multi-collinearity from entering the model if all dummy variables are present. Multicollinearity occurs when two or more explanatory variables have a high correlation. When 0 appears in a single row for every other dummy variable in each category, it is easy to determine the reference category. K-1 levels can represent the K-levels.

As an Example: The seasons are 'Spring' 'Fall', Summer', and 'Winter'. We could represent these 4 levels by 3 levels:

- 000 = Fall
- 100 = Spring
- 010 = Summer
- 001 = Winter

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)

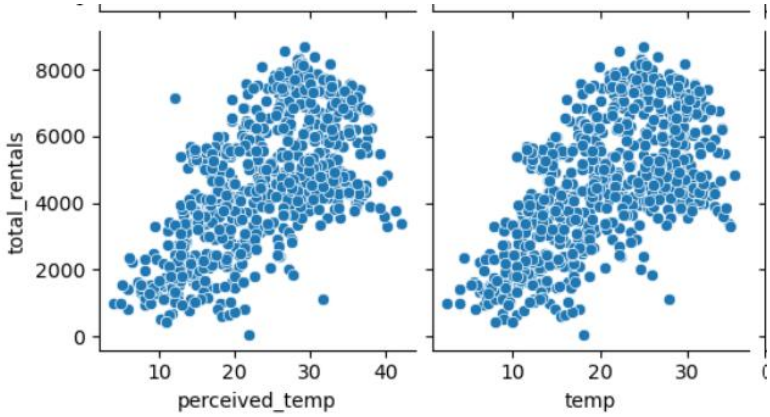
- 'temp' correlates highest with the target variable (approx. 63%).
- It seems that 'registered' has the highest correlation, but 'registered' is part of the target variable. We could decide whether a rental is 'casual' or 'registered' only once we have rented the bike.
- 'atemp' is a derived variable from temp/windspeed/humidity. Hence ignoring its correlation.

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)

Assumptions:

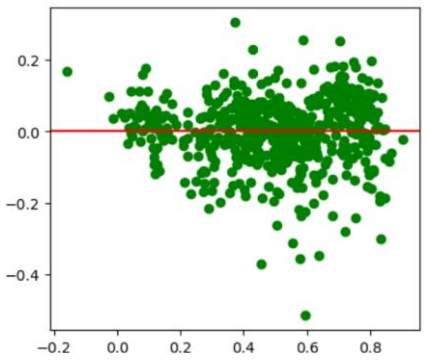
1- **The relationship between independent and dependent variables is linear:**
The most important assumption is the relationship between the explanatory variables and the study variable is linear. This indicates that when the independent variable changes, the dependent variable changes proportionately. Scatter plots and residual plots can be used to visually evaluate this.

Scatter plot: Creating a scatter plot of the independent variables vs. the dependent variable to validate the linear relationship between them. Example: We plotted the Seaborn pairplot and found that 'atemp' and 'temp' has a linear relationship with the target variable 'total_rentals':



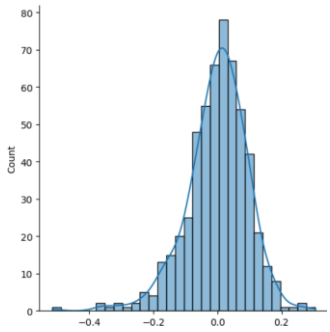
2- **Homoscedasticity:** Plot the residuals against the predicted values. The data is homoscedastic if the points are randomly scattered around zero or in a horizontal band. The data is heteroscedastic if the points form a pattern, like a funnel, curve, or cluster.

The points are randomly scattered around zero or in a horizontal band in the plot below. Hence our data is Homoscedastic.



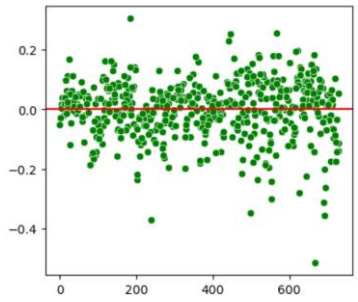
3- **Multivariate Normality – Normal Distribution of the residuals**
The residuals should have a normal distribution with the 'mean' at zero (almost). This assumption ensures that hypothesis tests, confidence intervals, and p-values are valid.

The residual in the plot below has a normal distribution with the 'mean' at zero (almost),



4- **Independence of Errors:** The error values should be independent of each other and should not show any type of correlation between them.

Error terms are independent. We could homogenously fit line Y = 0 in the graph.



5- **Multicollinearity:** VIF (Variance Inflation Factor) is a check for Multicollinearity. Multicollinearity can be dealt with by dropping a few variables (one by one). Multicollinearity should be zero or minimum**.**

Example: In our model, 'atemp' and 'temp' are highly correlated and have very high VIF values and thus there was multicollinearity. As we dropped 'atemp', the VIF value of 'temp' dropped significantly.

6- **Absence of Endogeneity:** The independent variables and the error terms should not be correlated otherwise leads to biased and inconsistent estimates of the regression coefficients.
We see below that the correlation between residual (res) is almost zero with the independent variable 'temp'

	temp	res
temp	1.000000e+00	7.024591e-16
res	7.024591e-16	1.000000e+00

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)

Top 3 features that contribute significantly to explaining the demand for shared bikes:

1- **'temp' (Means Temperature):** This variable has the highest positive coefficient (weight) of 0.428. This means a unit increase in temperature keeping other variables fixed would spur the demand for rental bikes by approx. 42.8%.

2- **'Light Snow and Light Rain' (Our original variable in the dataset was 'weathersit'):** This variable has the highest negative coefficient (weight) of 0.289. This means a unit increase in Snow and Rain attributes keeping other variables fixed would discourage the demand for rental bikes by approx. 28.9%.

3- **'yr' (Means Year):** This variable has the 2nd highest positive coefficient (weight) of 0.245. This means that even if all other factors are taken away, the business will still grow by about 25% year by year by approx. 24.5%.

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:

- It is a "Supervised Machine Learning Algorithm" meaning that the past data with labels is used for building the model.
- Based on the independent input variable, it forecasts the **continuous output variables** by fitting a linear equation to observed data. such as the estimation of home values according to several factors, such as the age of the property, its distance from the main road, its location, its area, etc.
    - Simple Linear Regression (SLR): The number of independent variables is 1
    - Multiple Linear Regression (MLR): Has more than 1 independent variable
    - Univariate Linear Regression: Single dependent variable
    - Multivariate Linear Regression: More than one dependent variable.
- Note: Linear regression guarantees interpolation but not extrapolation.
- The regression line is straight, and Linear regression follows the normal distribution

- Simple Linear Regression – A single independent variable is used.
    - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
- Multiple Linear Regression – Multiple independent variables are used.
    - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$ is the line equation for MLR.
- $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
- $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradients.$

- Cost functions – The cost functions identify the best possible values for the $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ and predict the probability of the target variable(s). The minimization approach helps to reduce the cost functions and provides the best-fitting line to predict the dependent variable.

  The 2 types of cost function minimization approaches – **Unconstrained and constrained**.
    - The 'Sum of Squared" function is used as a cost function to identify the best-fit line (since the squared function is differentiable). The cost functions are generally represented as
        - The straight-line equation is $Y = \beta_0 + \beta_1 X$
        - The prediction line equation would be $Y_{pred} = \beta_0 + \beta_1 x_i$ and the actual Y is Y_i.
        - $The\ cost\ function\ will\ be\ J(\beta_1, \beta_0) = \sum (y_i - \beta_1 x_i - \beta_0)^2$
    - The unconstrained minimization is solved using 2 methods
        - Closed form
        - Gradient descent
- In the best-fit line approach there are errors while mapping the actual values to the line. These errors are called residuals. To minimize the error squares **OLS (Ordinary Least Square)** is used.
    - $e_i = y_i - y_{pred}$ is provides the error for each of the data points.
    - OLS is used to minimize the total e^2 called the "Residual Sum of Squares" (RSS)
    - $\text{RSS} = = \sum_{i=1}^{n} (y_i - y_{pred})^2$.
    - $\text{The } \beta(s)\ given\ by:$ $(X'X)^{-1}X'Y$
- The Ordinary Least Squared method minimizes the Residual Sum of Squares and estimates beta coefficients.
- The following assumptions must be satisfied in Linear Regression:
    - The relationship between independent and dependent variables is linear
    - Normal Distribution of residuals
    - Homoscedasticity: The data is homoscedastic if the points are randomly scattered around zero or in a horizontal band.
    - Multicollinearity should be zero or minimum**.**
    - Absence of Endogeneity: The independent variables and the error terms should not be correlated

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>

Numerical Statistics like correlation, R^2, mean, Quantiles, variance, and standard deviation are usually considered good enough parameters to understand the variation of some data without looking at every data point. The statistics are great for describing the general trends and aspects of the data.
However, in 1973, Francis Anscombe realized that only statistical measures are not good enough to represent the data sets. He created different data sets with identical statistical properties to illustrate the facts.

Anscombe's Quartet emphasizes the importance of data visualization in detecting nuances, outliers, and diverse relationships in datasets. It also helps to identify trends and anomalies, highlighting that numerical summaries can be deceptive.

Link for Anscombe's Quartet dataset: https://query.data.world/s/6p2ntncvkzj5mnvbpkaswfilryvnrk

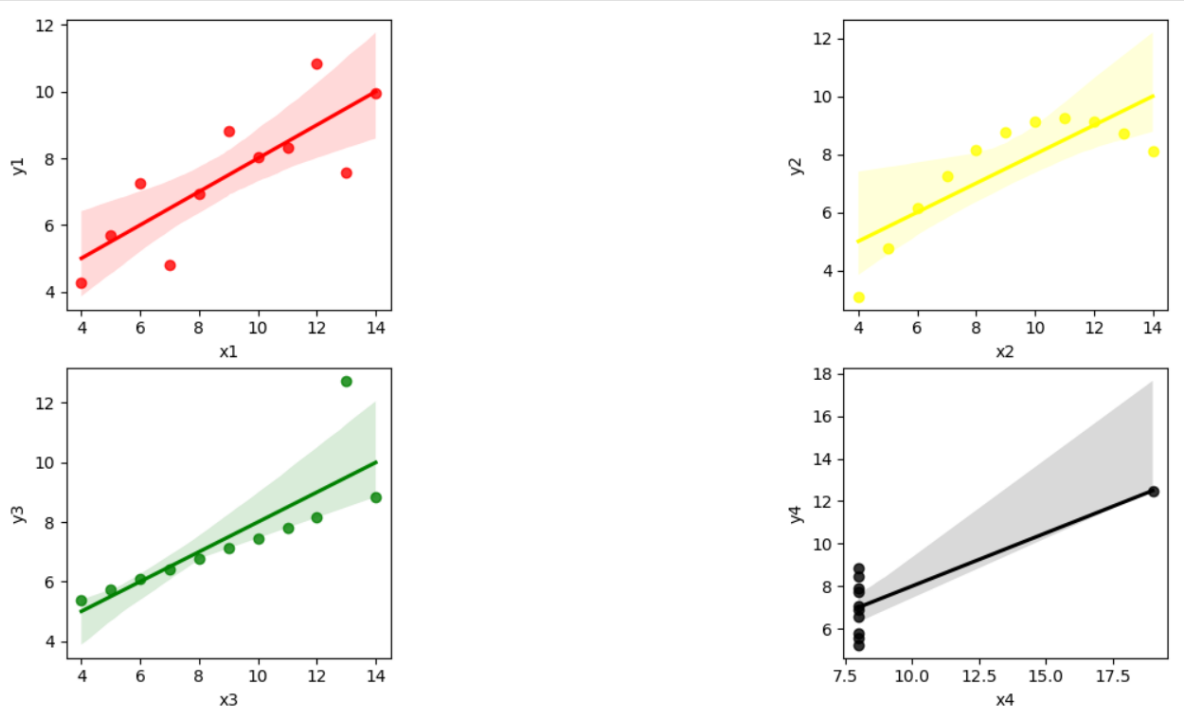
1st 5 Rows of the dataset are below:

	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

The numerical statistic of the dataset is almost the same:

	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 plots of Anscombe's Quartet show a variety of patterns, highlighting the value of data visualization for insights that go beyond numerical summaries.



- Summary points
  - Plotting the data is very important.
  - Outliers should be properly dealt with while analysing the data.
  - Descriptive statistics do not fully describe the dataset in its entirety.

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's R or The **Pearson correlation coefficient** is a descriptive statistic that measures a linear correlation (strength and direction of linear relationship) between two variables.
It is also an inferential statistic. It can be used in hypothesis testing to check the significance of a variable.

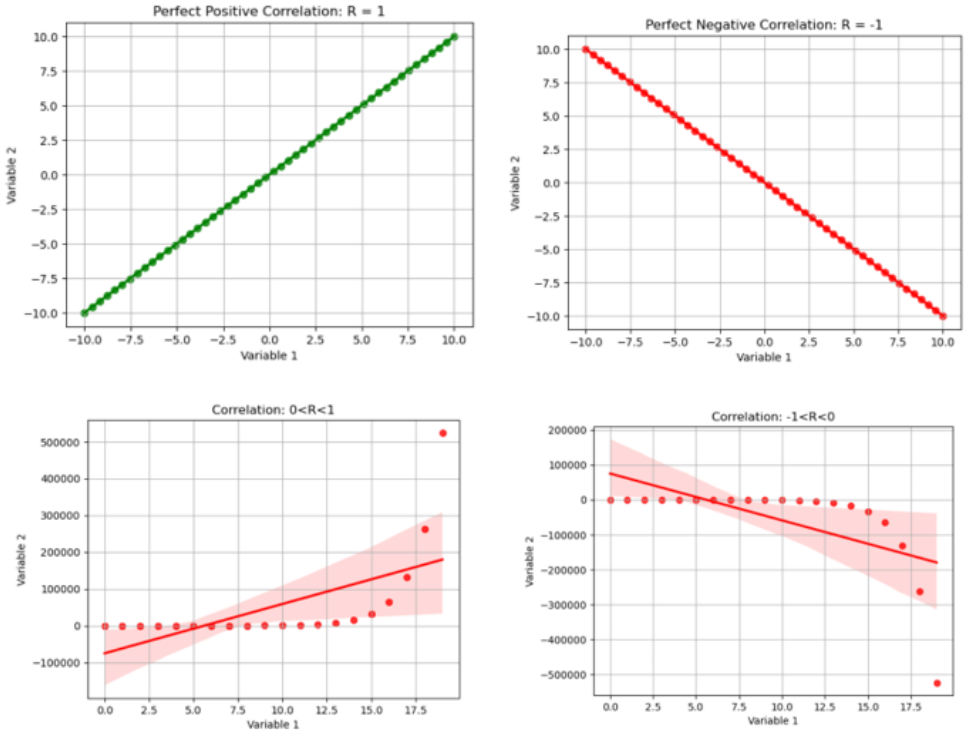
The Pearson's R returns values between -1 and 1. The interpretation of the coefficients is as follows:

- *-1 indicates a strong inversely proportional relationship.*
- *0 indicates no relationship.*
- *1 indicates a strong directly proportional relationship.*

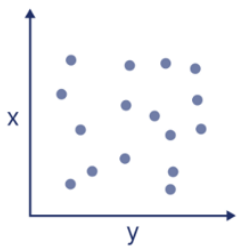
$$r = \frac{n(\Sigma x * y) - (\Sigma x) * (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}}$$

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx^2 = the sum of squared x scores
Σy^2 = the sum of squared y scores

Pearson correlation coefficient (R)	Correlation type	Interpretation	Example
Between 0 and 1	(directly proportional) Positive correlation	The direction of change for both the variables are same	Diesel Automobiles and Pollution level As the count of Diesel Automobiles increases, pollution level too increases
0	No correlation	The variables are unrelated	House prices and planet's gravity House prices have no relation with the gravity of the planet.
Between -1 and 0	(inversely proportional) Negative correlation	The direction of change for the variables is different (opposite)	The power of the Radio waves and distance of the propagation The power of the Radio waves decreases as the distance of propagation increases
1	Perfect Positive correlation	The direction of change for both the variables are exactly the same	Show size and the foot length As foot length increases, so does shoe size.
-1	Perfect Negative correlation	The direction of change for both the variables are exactly the opposite	Speed vs. Time if distance is fixed If distance is fixed, time taken will be less if speed is more



No Correlation: R = 0



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 preprocessing technique that converts feature values to a comparable scale that guarantees that every feature contributes equally to the model. For datasets that contain features with different ranges, units, or magnitudes, it is crucial that bias from features with higher values is avoided and model performance and convergence are enhanced. Common scaling techniques are standardization and normalization (min-max scaling).

Why Scaling is performed:

- Most feature data is collected in public domains where the interpretation of variables and their units is kept as open as possible. As a result, the units and ranges of data are highly variable. If these data sets are not scaled, there is a high chance that the data will be processed without the proper unit conversion. Also, the higher the range, the more likely it is that the coefficients are impaired when comparing the dependent variable variance. The coefficients are the only ones affected by the scaling.
- Scaling does not affect prediction accuracy and parameters like t-statistics, F-statistics, p-values, $R^{2,}$ etc.
- Also, some machine learning algorithms such as gradient descent are sensitive to feature scaling. So, while employing these algorithms, we need scaling.

Normalized (or MinMax Scaling): MinMax scaling normalizes data between 0 and 1. The Min max scaling also helps to normalize outliers.

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

Standardized scaling: Standardization transforms all data points into a standard normal distribution with a mean of zero and a standard deviation of one.

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

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>

The Variance Inflation Factor (VIF) can be infinite when there is perfect multicollinearity in a model, which means that one independent variable can be predicted by another.

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

From the VIF formula, if the R^2 is 1 then the VIF is infinite. $R^2 = 1$ indicates that there is a perfect correlation between 2 independent variables.
Example: Y = X, Y = -X.

This can happen when:

- There are several identical columns in the input dataset, or when the independent variables have a high correlation (correlation coefficients close to 1 or -1).
- Other variables essentially derive the target variable.
    - Example: Profit = Selling price – cost price. If all three variables are there in the dataset and we want to predict profit, then we must drop the cost price and selling price.

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 is a graphical tool for determining whether two sets of data have the same distribution. Q-Q plots are useful in linear regression because they can help you check whether your model's residuals are normally distributed, which is an assumption for many parametric tests and confidence intervals.

We could determine via Q-Q plot:

- Distribution of the data is normal or some other distribution like exponential, uniform, Poisson, etc.
- Whether there are outliers is the data
- Presence of Homoscedasticity
- Skewness of the data

Interpretation of Q-Q plot:

- A similar distribution occurs when all quantile data points align with a straight line at a 45-degree angle to the x-axis
- Y values less than X values: If the y-value quantiles are less than the x-value quantiles.
- X values < Y values: When x-values are lower than y-values quantiles.
- Different distributions occur when all data points deviate from a straight line.

Advantages of Q-Q Plot:

- Allows for non-equal sample sizes.
- Several distributional aspects can be tested concurrently. This plot can detect shifts in location, scale, and symmetry, as well as the presence of outliers.