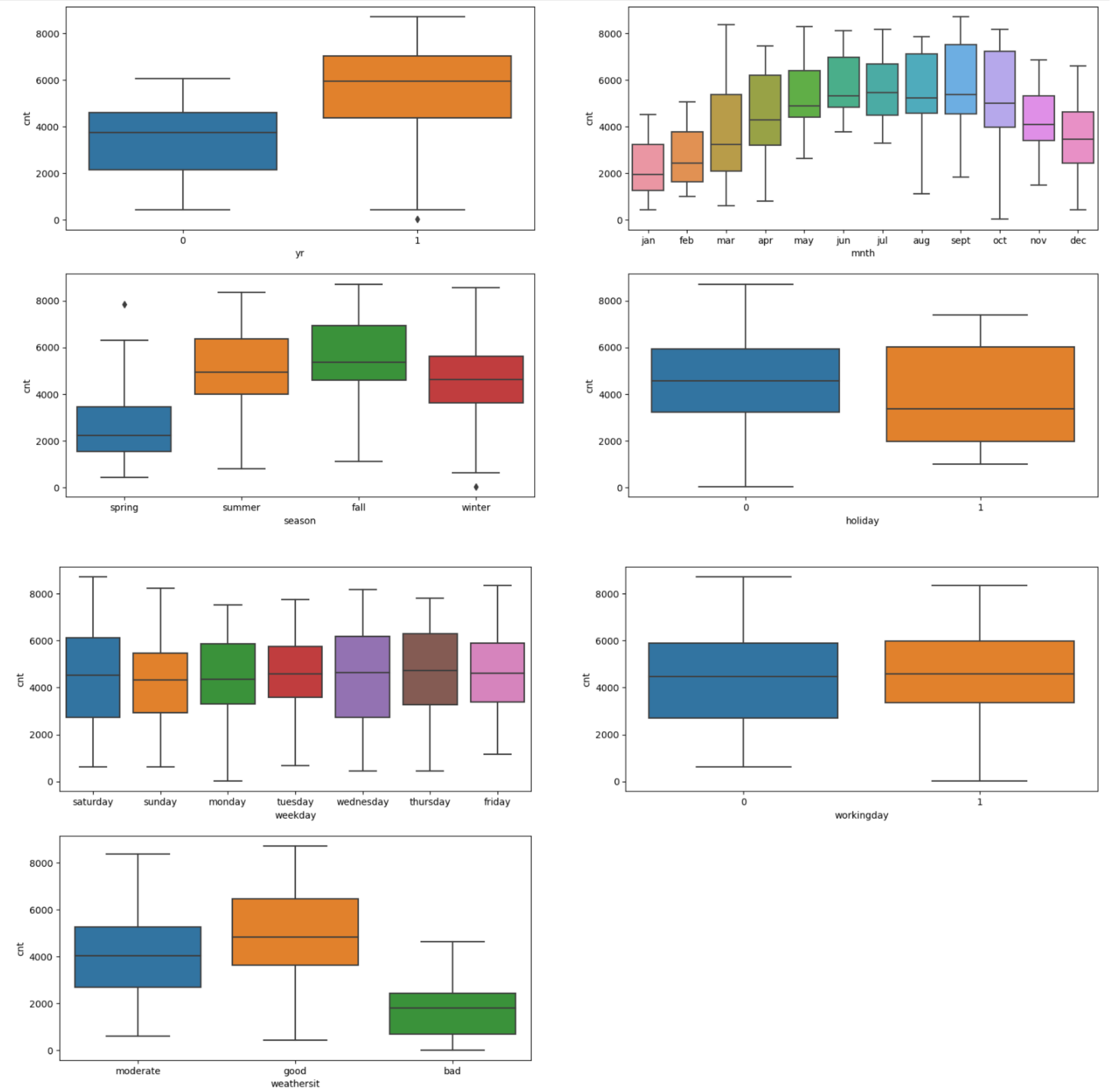
Assignment-based Subjective Questions

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

Answer :

In the given dataset there are some categorical variable like season, yr, holiday, weekday, workingday, and weathersit and mnth. Visualization of these variables is as follows:



These variables had the following effect on our dependant variable:-

- Yr : rentals in the year 2019 is more than 2018
- Mnth : There is incremental growth from Jan till Sept but afterwards it is decreasing. Highest cnt is in Sept and Jan have least cnt

- Season : It is clear that highest demand is in fall and sping season have least value of cnt. Summer and winter have moderate cnt value

- Holiday : During holiday there is less cnt value that means less demand of rental bikes

- Weekday: The cnt of rentals is almost same throughout the week.

- Workingday: Median count is almost same throughout the week.

- Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavourable. Highest count was seen when the weathersit was' Clear, Partly Cloudy'.

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

   a. Using **drop_first=True** during dummy variable creation is important to avoid multicollinearity in regression analysis and to simplify the interpretation of the model.
   b. Including dummy variables for all categories of a categorical variable without dropping one can introduce multicollinearity because the sum of the dummy variables for each category will always be constant (equal to 1).
   c. Suppose we have categorical variable "Colour" with three categories: Red, Blue, Green. If we create dummy variables without dropping one, we would have the following representation:

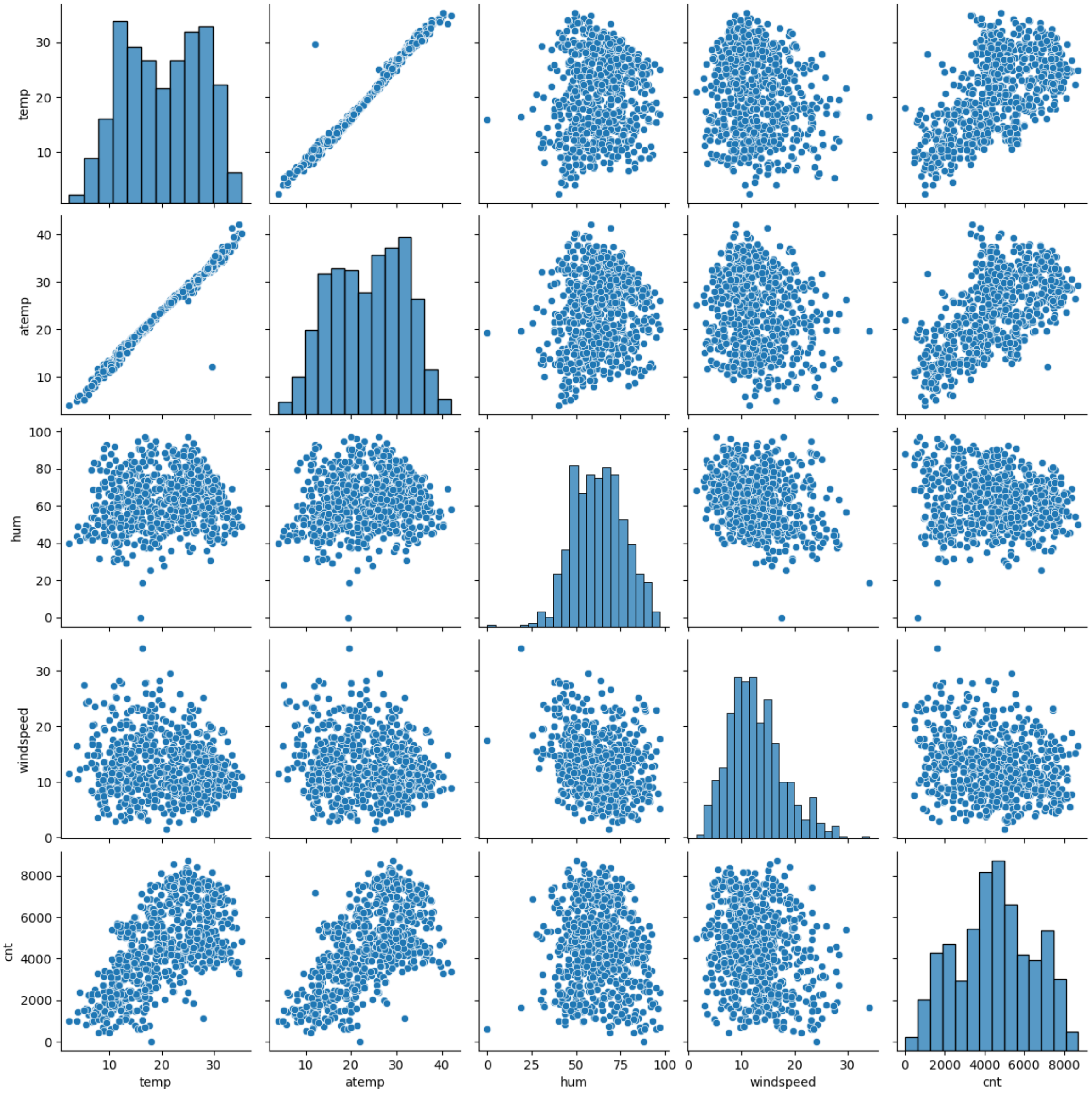
Colour	Red	Blue	Green
Red	1	0	0
Blue	0	1	0
Green	0	0	1

In above example all table we can see all colours are highly corelated. We can work after dropping one column also. If we drop Green column, that will be conclude by Red = 0, Blue = 0

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

Answer :

Below is the pair plot generated for the numerical variables :



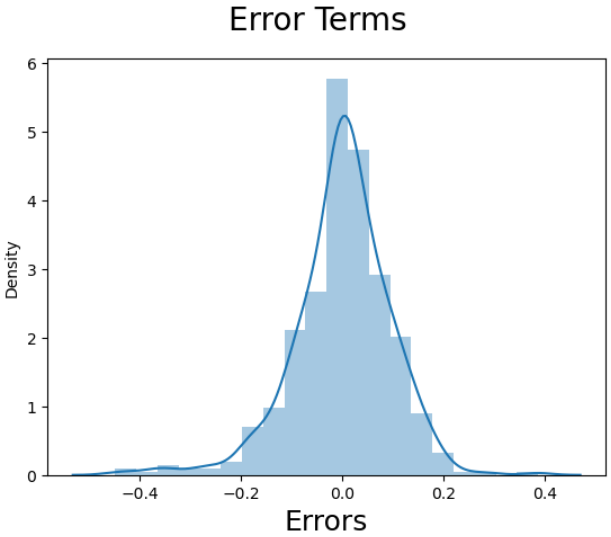
From above pair plot we can say that variables atemp and temp have highest co-relation between them.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

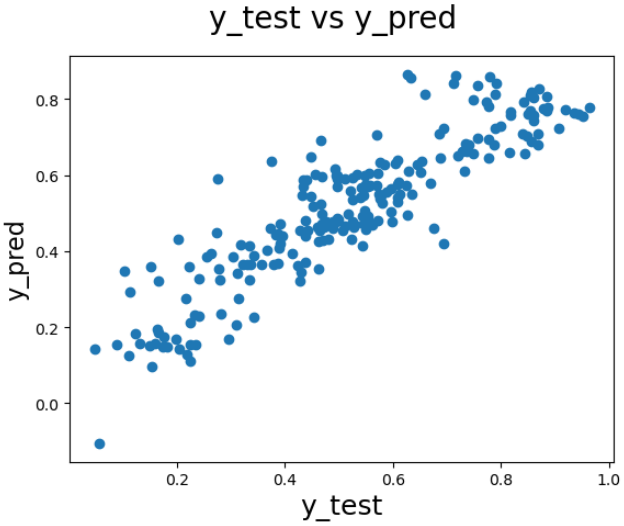
Answer :

The following tests were done to validate the assumptions of linear regression:

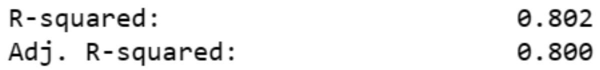
1. In the given plot below we see that residuals are distributes around mean = 0. As we know, residuals distribution should follow normal distribution and centered around 0 (mean = 0). We can validated this assumption about residuals by below distplot of residuals and can say that the residuals are following normal distribution.



2. linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to get the quantitative idea about how much the feature variables are correlated with each other in the new model. Refer to the notebook for more details.

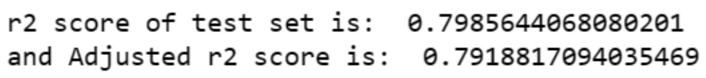


3. R-squared and adjusted r-score of training set are:



```
R-squared:                         0.802
Adj. R-squared:                    0.800
```

And R-squared and adjusted r-score of test set are:



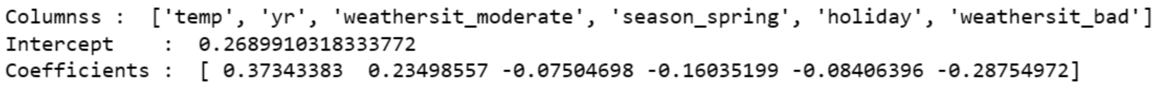
```
r2 score of test set is:  0.7985644068080201
and Adjusted r2 score is:  0.7918817094035469
```

So, the difference is about 0.01 which is accepted. Therefore, we can say that our assumption is correct.

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

   Answer :

   Columns and their coefficient of the final model are as follow:



```
Columnss :  ['temp', 'yr', 'weathersit_moderate', 'season_spring', 'holiday', 'weathersit_bad']
Intercept    :  0.2689910318333772
Coefficients :  [ 0.37343383  0.23498557 -0.07504698 -0.16035199 -0.08406396 -0.28754972]
```

From above data we can conclude top 3 features as follow:

1. temp : Coefficient is – 0.37343383 : we can say that demand will increase with increase in temprature
2. yr - coefficient : 0.23498557 – demand will increase yearly
3. weathersit_Light Snow & Rain - coefficient : -0.28754972 – This concludes, bad weather decreases demand of bikes

General Subjective Questions

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

   Answer:

   - Linear regression is a statistical method used for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors or features) by fitting a linear equation to the observed data.

   - Its primary goal is to establish a linear relationship that can be used to make predictions or understand the association between variables.

   - Here's a step-by-step explanation of the linear regression algorithm:

     a. Data Collection: Gather a dataset that includes both the dependent variable (Y) and one or more independent variables (X). The data should be suitable for regression analysis, meaning there should be a potential linear relationship between the variables.

     b. Data Preprocessing: Clean and preprocess the data. This includes handling missing values, removing outliers, and scaling or standardizing the variables if necessary.

c. Model Representation: In simple linear regression, where you have one independent variable, the model can be represented as:

   Y = b0 + b1*X + ε

   Y represents the dependent variable.

   X represents the independent variable.

   b0 is the intercept (the value of Y when X is 0).

   b1 is the slope (the change in Y for a one-unit change in X).

   ε represents the error term, accounting for unexplained variability.

   In multiple linear regression, where you have more than one independent variable, the equation is extended accordingly.

d. Model Training: The goal is to find the best-fitting line (or hyperplane in multiple linear regression) that minimizes the sum of squared differences between the predicted values and the actual values. This is typically done using a method called "Ordinary Least Squares" (OLS) or other optimization techniques.

e. Model Evaluation: Assess the quality of the model using various statistical metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2), and others. These metrics help determine how well the model fits the data.

f. Inference and Prediction: Once the model is trained and evaluated, you can use it for inference and prediction. You can make predictions on new, unseen data by plugging the values of the independent variables into the regression equation.

g. Assumptions: Linear regression relies on certain assumptions, including linearity (the relationship is linear), independence of errors (residuals are uncorrelated), constant variance of errors (homoscedasticity), and normality of errors (residuals follow a normal distribution). These assumptions should be checked and, if violated, appropriate measures should be taken.

---

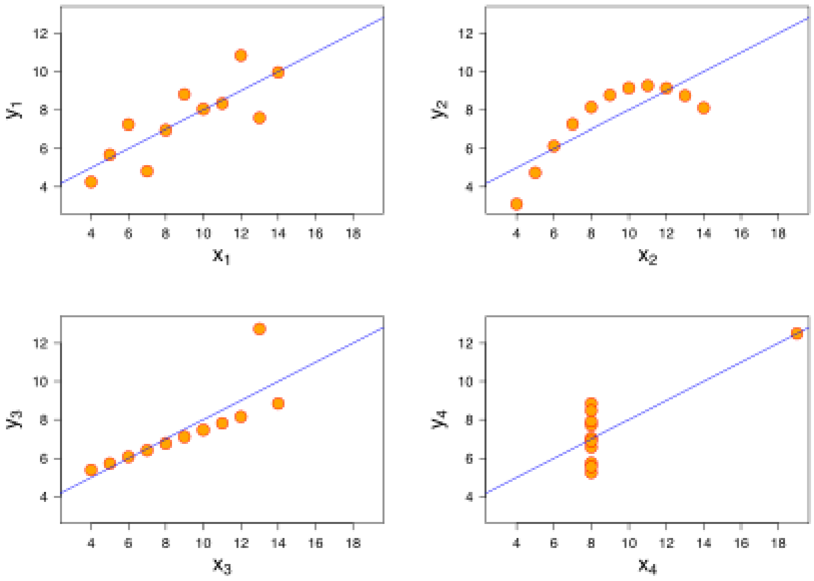
2. Explain the Anscombe's quartet in detail. (3 marks)

   Answer:

   a. **Anscombe's quartet** is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

   c. 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.



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them,it's not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R? (3 marks)

Answer:

A. Pearson correlation coefficient, also known as Pearson R statistical test, measures the strength between the different variables and their relationships. Therefore, whenever any statistical test is conducted between the two variables, it is always a good idea for the person analyzing to calculate the value of the correlation coefficient to know how strong the relationship between the two variables is.

B. Pearson's correlation coefficient can range from the value +1 to the value -1, where +1 indicates the perfect positive relationship between the variables considered, -1 indicates the perfect negative relationship between the variables considered, and 0 value indicates that no relationship exists between the variables considered.

C. Below is table that shows correlation:

Pearson correlation coefficient (*r*) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None

Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

D. The Pearson correlation coefficient shows the relationship between the two variables calculated on the same interval or ratio scale. In addition, It estimates the relationship strength between the two continuous variables.

E. The Pearson Correlation Coefficient formula is as follows:

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x}^2 - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y}^2 - (\sum \mathbf{y})^2]}}$$

Where,

r = Pearson Coefficient
n= number of pairs of the stock
∑xy = sum of products of the paired stocks
∑x = sum of the x scores
∑y= sum of the y scores
∑x2 = sum of the squared x scores
∑y2 = sum of the squared y scores

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

Answer:

a. What is scaling?
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

b. Why is it performed?

- Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

- It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

c. The difference between normalized scalingand standardized scaling **:**
Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

---

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

- VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

- For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

  X_1=C+ α_2 X_2+α_3 X_3+⋯

  〖VIF〗 _1=1/(1-R_1^2)

- Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

  X_2=C+ α_1 X_1+α_3 X_3+⋯

  (VIF〗 _2=1/(1-R_2^2)

- If all the independent variables are orthogonal to each other, then VIF = 1.0. *If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.*

VIF	Conclusion
1	No multi collinearity
4-5	Moderate
10 or greater	Sever

---

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

Answer:

a. Q-Q plots are also known as Quantile-Quantile plots. They plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

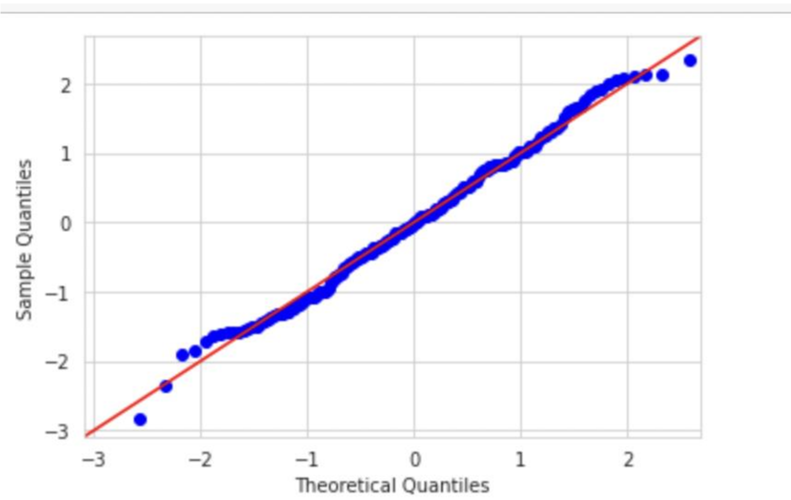
b. QQ plots is very useful to determine:

- If two populations are of the same distribution

- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.

- skewness of distribution

c. Below is the example of Q-Q plot :



d. If we see the left side of the plot deviating from the line, it is left-skewed. When the right side of the plot deviates, it's right-skewed.