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)

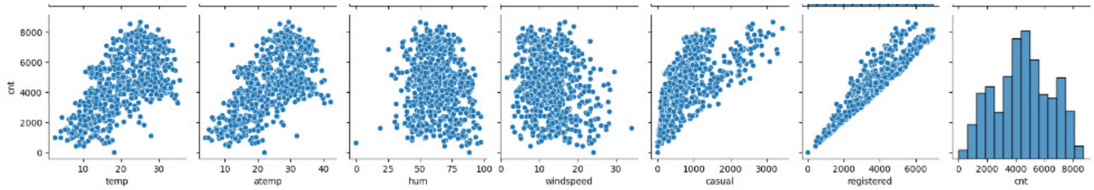
   The categorical variables considered for the test from dataset are workingday, holiday, season, weathersit, month, weekday and year. Some of these are of type object and some have numerical values. Get all the dummy variables for all the categorical variables. For those which are not object, make dummy values for them. And then concatenate both. Dependent variable is Count. As we reduce the number of the columns from 28 to 15 to 7, we see that dependent variable count , has reduced the R squared. There is a significant reduction in R squared using specific dataset.

   .

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

   The categorical variables which are of type object, these need to be removed before converting the existing values to 1s and 0s as in the given dataset. All variables with integer are replaced with dummy variables. And this requires the value to be an integer. Since variable of type object cannot be converted into dummy variables as 1s and 0s, hence first remove the variables of type object, get dummies for the variables which are of integer type, and then merge both object and integer related variables into dataframe. This way we don't corrupt the data.

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



   Registered and casual independent variables have highest correlation with count target variable.

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

First we split the train set as 75% and test set as 25% data. Use the scaling variables as 'temp', 'hum', 'windspeed','casual','registered'. Correlation coefficients were shown to see which variables are highly correlated. Temp is seen to be most correlated after casual and registered as per heatmap. We can do a pairplot for temp and cnt now. Before that, drop casual and registered since they add up to count.

Use statsmodel OLS method to build the training model. And verify the R-squared and Adj R-squared. It shows as follows: The values are promising.

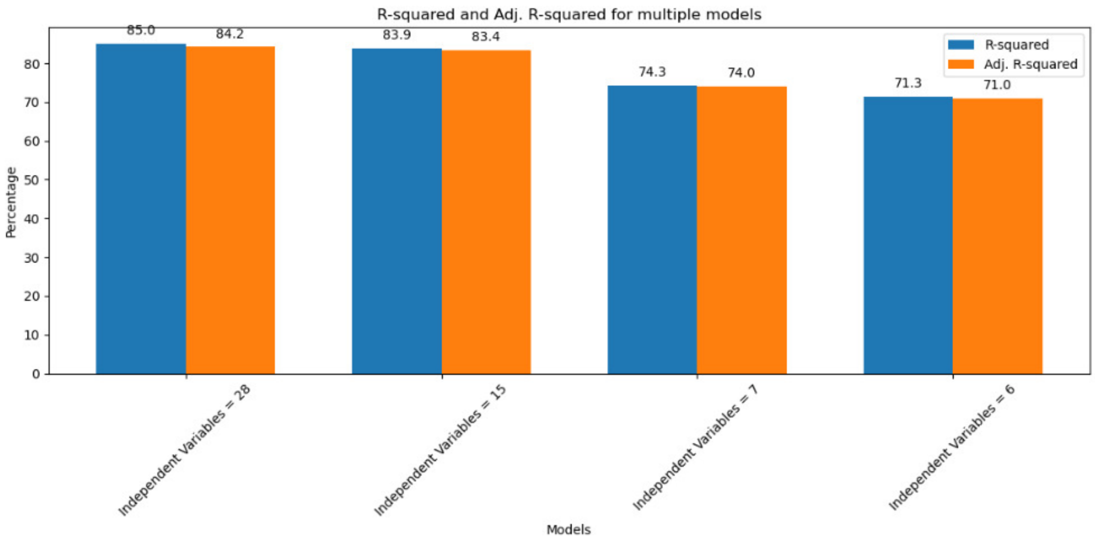
OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.850
Model:	OLS	**Adj. R-squared:**	0.842
Method:	Least Squares	**F-statistic:**	109.1
Date:	Mon, 11 Mar 2024	**Prob (F-statistic):**	1.78e-194
Time:	01:24:07	**Log-Likelihood:**	-4398.3
No. Observations:	547	**AIC:**	8853.
Df Residuals:	519	**BIC:**	8973.
Df Model:	27		
Covariance Type:	nonrobust		

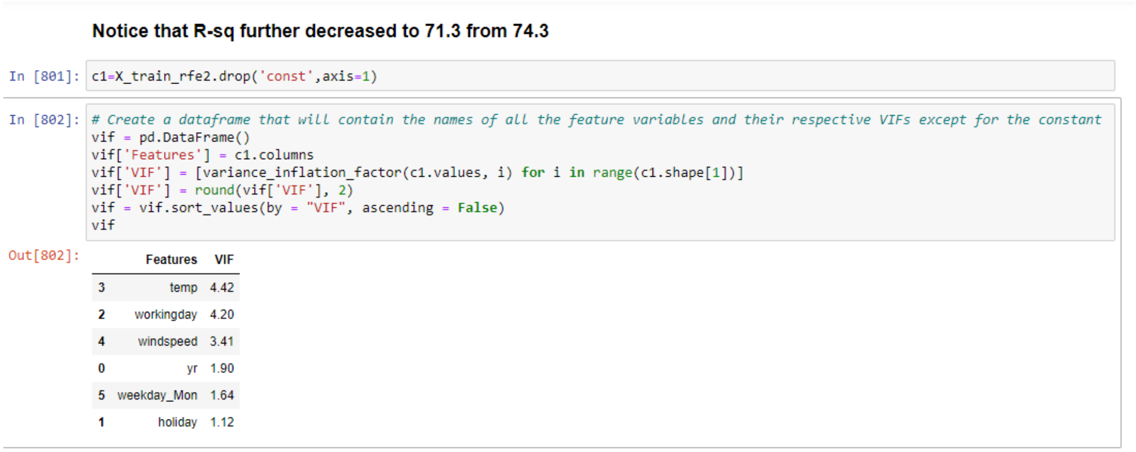
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	2053.7150	322.300	6.372	0.000	1420.543	2686.887
yr	1994.1427	67.490	29.547	0.000	1861.555	2126.730
holiday	49.0697	204.696	0.240	0.811	-353.065	451.204
workingday	744.3853	108.061	6.889	0.000	532.095	956.675
temp	3782.5622	394.198	9.596	0.000	3008.142	4556.982
hum	-1724.0033	329.944	-5.225	0.000	-2372.193	-1075.814
windspeed	-1495.3417	216.923	-6.893	0.000	-1921.497	-1069.186
season_spring	-600.0587	254.645	-2.356	0.019	-1100.321	-99.796

WE can also see temp has highest coeff, and lowest p-value.

This model has an Adjusted R-squared value of 84.2% which is very good. But let's see if we can reduce the number of features and exclude those which are not much relevant in explaining the target variable. The next step is to reduce the number of independent variables to 15 and then to 7. At the same time observe, compare the R-squared & Adj. R-squared values.



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



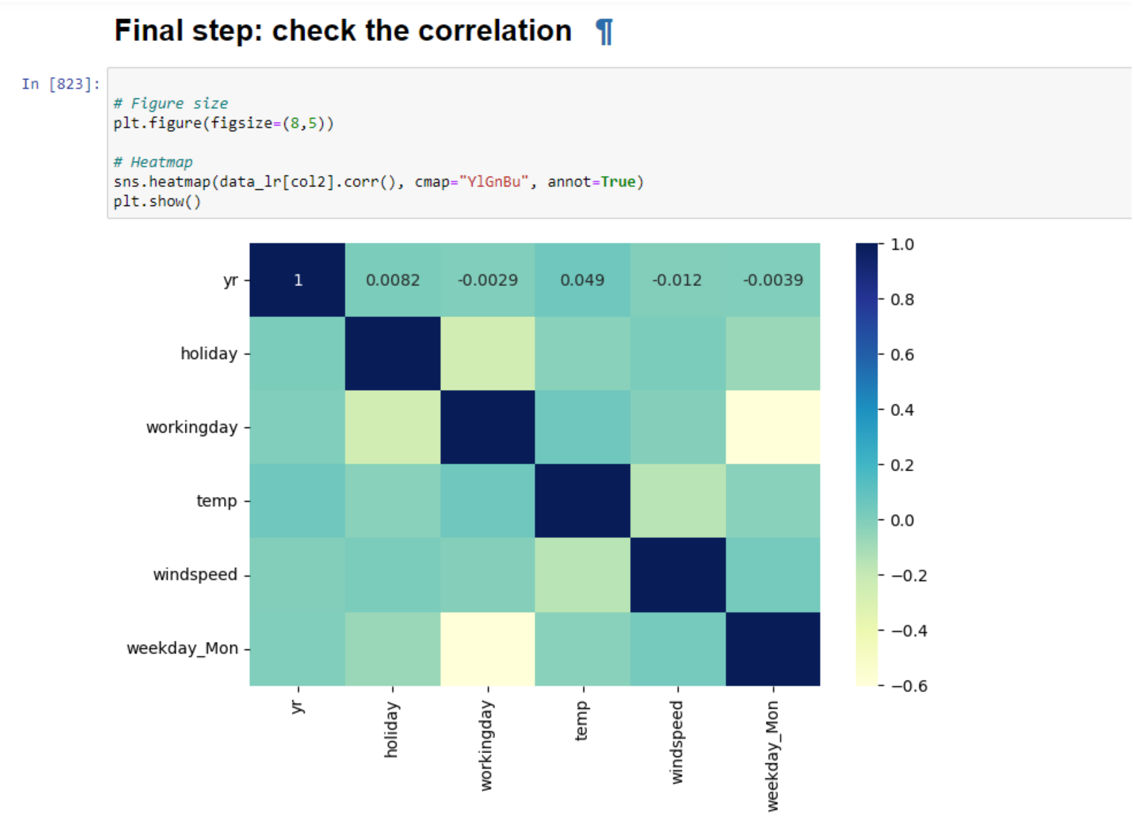
Notice that R-sq further decreased to 71.3 from 74.3

```
In [801]: c1=X_train_rfe2.drop('const',axis=1)
```

```
In [802]: # Create a dataframe that will contain the names of all the feature variables and their respective VIFs except for the constant
          vif = pd.DataFrame()
          vif['Features'] = c1.columns
          vif['VIF'] = [variance_inflation_factor(c1.values, i) for i in range(c1.shape[1])]
          vif['VIF'] = round(vif['VIF'], 2)
          vif = vif.sort_values(by = "VIF", ascending = False)
          vif
```

```
Out[802]:
      Features   VIF
  3       temp  4.42
  2 workingday  4.20
  4  windspeed  3.41
  0         yr  1.90
  5 weekday_Mon 1.64
  1    holiday  1.12
```

Temperature, workingday, windspeed are major 3 factors contributed significantly towards explaining the demand of the shared bikes.



Final step: check the correlation ¶

```
In [823]:
# Figure size
plt.figure(figsize=(8,5))

# Heatmap
sns.heatmap(data_lr[col2].corr(), cmap="YlGnBu", annot=True)
plt.show()
```

General Subjective Questions

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

Linear regression algorithm is a supervised machine learning having model that finds the best fit linear line between independent and dependent variables. Helps in finding the linear relationship between dependent(y) and independent.

There are two types: simple and multiple.

Simple linear regression comes in where only one independent variable is present and model has to find the linear regression of it with dependent variable

Multiple linear regression comes in where there are more than one independent variable for the model, and we need to find relationship.

Formula for Simple linear regression: b_0 is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$Y = b_0 + b_1 x$$

Formula for multiple linear regression: b_0 is the intercept, b_1 , b_2 , b_3, b_4 , is coefficient and $x_1, x_2, x_3, x_4 \dots. x_n$ is the independent variable, and Y is the dependent variable

$$y = b_o + b_1x_1 + b_2x_2 + b_3x_3 \dots. + b_nx_n$$

Motivation for linear regression: to find the best fit linear line and optimal values of intercept and coefficients such that error is minimized

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

   Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

   Key Points of Anscombe’s Quartet

   Statistical Properties: Despite their differences in distribution and appearance when graphed, all four datasets in Anscombe’s quartet share nearly identical statistical properties, including the mean and variance of both x and y variables, the correlation between x and y, and the linear regression line (y = mx + c) that best fits the data.

   Datasets:

   Dataset I typically appears as a simple linear relationship, corresponding closely to the regression line.

   Dataset II is not linear but rather forms a curve. The linear regression line is not a good fit for this data.

   Dataset III appears as a linear relationship similar to Dataset I, but with one outlier affecting both the slope of the regression line and the correlation coefficient.

   Dataset IV shows a relationship where x values are constant for all but one point, demonstrating how a single outlier can heavily influence the regression line.

Importance

Visualizing Data: Anscombe’s quartet underscores the importance of visualizing data before analyzing it. Simply relying on statistical summaries can be misleading and might not reveal the true nature of the data or the relationship between variables.

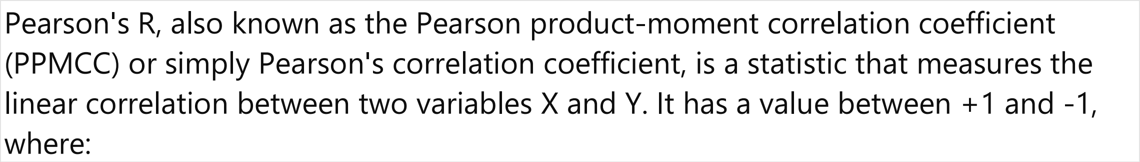
Impact of Outliers: It highlights how outliers can significantly affect the outcome of statistical analyses, such as linear regression, and distort the relationship between variables.

Assumption Checking: It serves as a reminder that the assumptions underlying many statistical methods, such as linearity, homoscedasticity (equal variance), and normality, should always be checked when analyzing data.

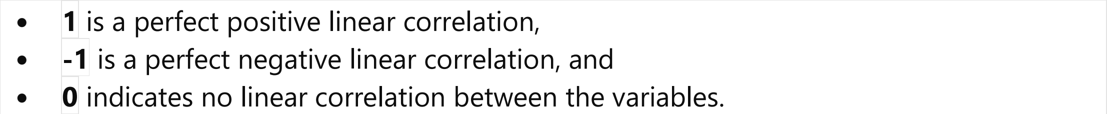
Conclusion

Anscombe's quartet is a crucial demonstration in statistics that emphasizes the need for a thorough exploratory data analysis, including visualizing data, before proceeding with any formal analysis. It showcases that statistical measures can be identical across different datasets and yet represent vastly different realities, highlighting the limitations of summary statistics alone in understanding data complexities.

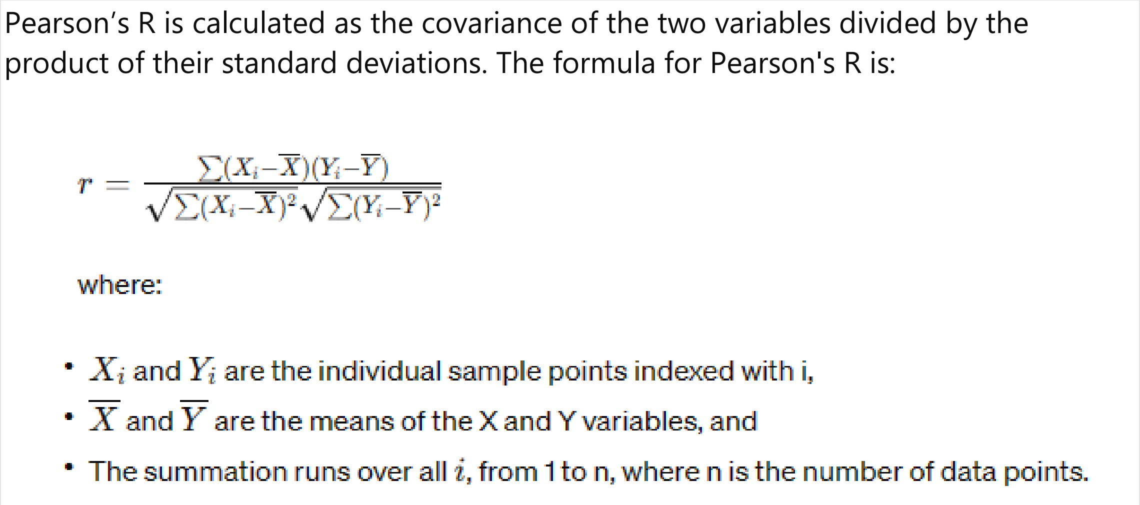
3. What is Pearson’s R? (3 marks)



Pearson's R, also known as the Pearson product-moment correlation coefficient (PPMCC) or simply Pearson's correlation coefficient, is a statistic that measures the linear correlation between two variables X and Y. It has a value between +1 and -1, where:



- **1** is a perfect positive linear correlation,
- **-1** is a perfect negative linear correlation, and
- **0** indicates no linear correlation between the variables.



Pearson’s R is calculated as the covariance of the two variables divided by the product of their standard deviations. The formula for Pearson's R is:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2}\sqrt{\sum(Y_i - \overline{Y})^2}}$$

where:

- X_i and Y_i are the individual sample points indexed with i,
- $\overline{X}$ and $\overline{Y}$ are the means of the X and Y variables, and
- The summation runs over all i, from 1 to n, where n is the number of data points.

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

Scaling is a technique used in data preprocessing, particularly in the fields of machine learning and data analytics. It involves adjusting the range of the data values to meet specific criteria, often to ensure that different features contribute equally to the analysis or to improve the convergence speed of gradient-based optimization algorithms. Scaling is crucial for models that rely on the distance between data points, such as k-nearest neighbors (KNN) and support vector machines (SVM), and for techniques that involve gradient descent optimization, like neural networks.

Scaling is a technique used in data preprocessing, particularly in the fields of machine learning and data analytics. It involves adjusting the range of the data values to meet specific criteria, often to ensure that different features contribute equally to the analysis or to improve the convergence speed of gradient-based optimization algorithms. Scaling is crucial for models that rely on the distance between data points, such as k-nearest neighbors (KNN) and support vector machines (SVM), and for techniques that involve gradient descent optimization, like neural networks.

Why is Scaling Performed?
Scaling is performed for several reasons:

- Equal Contribution: To ensure that no single feature dominates the model due to its scale. For example, if one feature ranges between 0 and 1, while another feature ranges between 1,000 and 10,000, models might unduly weight the influence of the larger range feature.
- Numerical Stability: Some algorithms, especially those involving distance calculations, can become numerically unstable if the features are on vastly different scales.
- Convergence Speed: Gradient descent and similar optimization algorithms converge faster when features are on a similar scale, as it helps in preventing the optimization path from becoming skewed.
- Algorithm Requirements: Certain algorithms, like SVM and KNN, explicitly require scaling for correct operation.

Normalized Scaling vs. Standardized Scaling

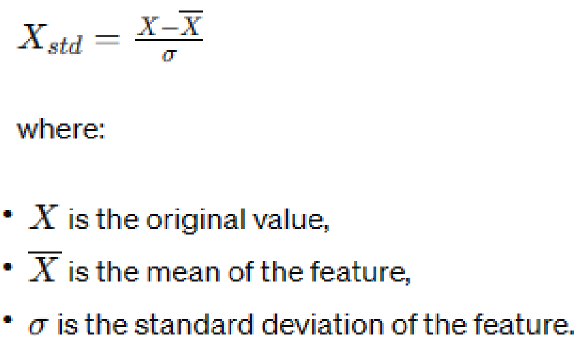
Normalized Scaling:

Normalization, also known as min-max scaling, is a scaling technique that linearly transforms the range of feature values to a common scale, typically [0, 1] or [-1, 1]. The formula for min-max scaling to the range [0, 1] is:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardized Scaling:

Standardization, on the other hand, transforms the data to have a mean of 0 and a standard deviation of 1. It is achieved by subtracting the mean and then dividing by the standard deviation for each feature:



$$X_{std} = \frac{X - \overline{X}}{\sigma}$$

where:

- X is the original value,
- $\overline{X}$ is the mean of the feature,
- σ is the standard deviation of the feature.

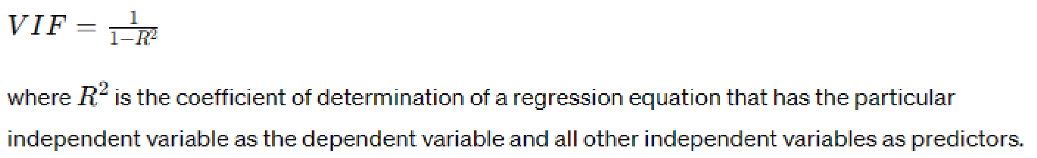
6. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

The variables should be independent of each other, hence no correlation should be there between the independent variables. When we see the correlation matrix or VIF score. If the VIF score is more than 5, then the variables are highly correlated. While doing the modelling fitting in training, while we reduce the number of variables, the VIF value kept reducing as we reduced the number of

variables. Then removing highly correlated values, we will narrow down on variables slightly above 5, which might be the once influencing our analysis or influencing the dependent variable.
The Variance Inflation Factor (VIF) is a measure used primarily in regression analysis to detect the presence and severity of multicollinearity, which is when two or more independent variables in a regression model are highly correlated. The VIF quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated. If no factors are correlated, the VIFs will all be 1.

Why VIF might be infinite

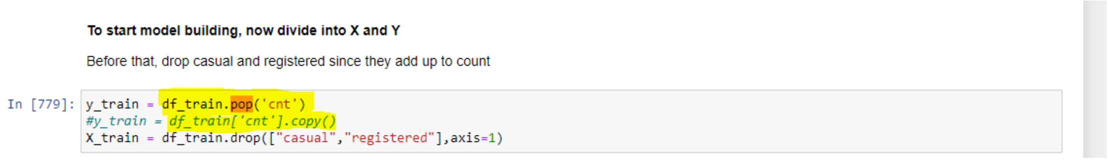
VIF is given by:



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

where R^2 is the coefficient of determination of a regression equation that has the particular independent variable as the dependent variable and all other independent variables as predictors.

VIF might be infinite when R2 is 1. And this can happen when the independent variable is perfectly correlated with another independent variable. For example, we have count as one of the variable in the independent variables. And this one was popped out and added into Y_train. The reason we do this was to remove it from data frame, and put that into Y set. If not, then the same variable exists in both Y and X, and hence the R^2 and Adj.R^2 would be both 1 and shows perfect correlation. Therefore the VIF will be infinte in that case. IT shows something is wrong with our calculation. As shown below instead of *pop()* if we do *copy()* , then VIF would become *Infinite*



To start model building, now divide into X and Y

Before that, drop casual and registered since they add up to count

```
In [779]: y_train = df_train.pop('cnt')
          #y_train = df_train['cnt'].copy()
          X_train = df_train.drop(["casual","registered"],axis=1)
```

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

A Q-Q (Quantile-Quantile) plot is a graphical tool to assess if a set of data plausibly came from some theoretical distribution such as a Normal, Exponential, or Uniform distribution.

It’s a graphical tool to assess if a set of data came from normal, exponential or uniform distribution.

It compares the dataset against the quantiles of theoretical distribution, gives a visual means to evaluate how well the data matches the given distribution

Importance:

Used to verify assumption of normality of residuals. For linear regression, it is important that residuals must be normal. The residuals of a regression model are the differences between the observed values and the values predicted by the model.

Importance of this plot in Linear regression: Used as diagnostic tool to check validity of linear equation model assumptions. Useful to assess the reliability of the hypothesis tests and confidence intervals for regression coefficients.

Improving model accuracy by identifying non-normality in the residuals, this plot can show when transformations of variables might be necessary to improve model fit and accuracy.