Assignment Questions

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?

   A -

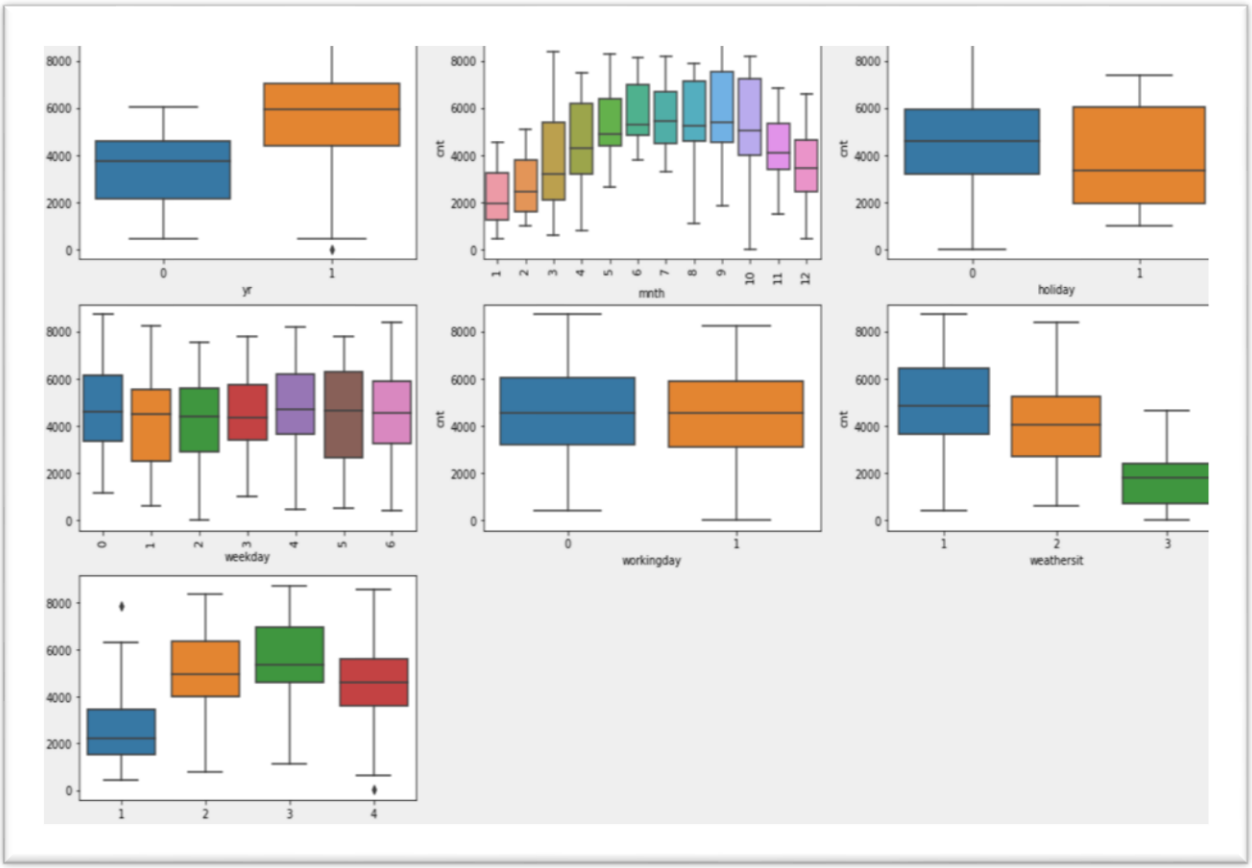

Fig 1: Box plots of categorical variables against the count variable

The variables – year, month, weather situation and season seem to have an effect on the count of riders.

- The year 2018 had a significantly lower number of riders than the year 2019.
- Months June to September have a higher number of riders as compared to the rest.
- The weather situation – clear, partly cloudy has the highest number of riders. Light snowy weather has the least.
- Spring season has the lowest number of riders.

(year – 0: 2018, 1:2019; Month – 1-12: January to December; weather situation – 1: Clear, 2: Mist, 3: Light Snow, 4: Heavy Rain; Season – 1: Spring, 2: Summer, 3: Fall, 4: Winter)

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

N = number of dummy variables for a particular variable.
N-1 = number of dummy variables are enough to explain all the N variables.

Example – Say, we have to create dummy variables for season. There are 4 seasons – winter, summer, spring and fall.

The following table explains all the 4 seasons –

Winter	Summer	Spring
0	0	0
1	0	0
0	1	0
0	0	1

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

We are able to explain all the 4 seasons using only 3 variables. **Hence, in order to reduce the number of dummy variables in the dataset that will be used for building the model, we use drop_first = True.**

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

   Temperature (temp) and actual temperature (atemp) have the highest correlation with the target variable (cnt). (If we look at the pair-plot, registered and casual have high correlations. However, we are not considering these variables here as their summation makes up the target variable i.e. cnt.)

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

   There are 4 assumptions of linear regression:
   - Target variable and input variable should be linearly dependent.
   - Error terms are normally distributed
   - Error terms are independent of each other.
   - Error terms have a constant variance (Homoscedasticity).

   Checking the assumptions:

   a. Linearity

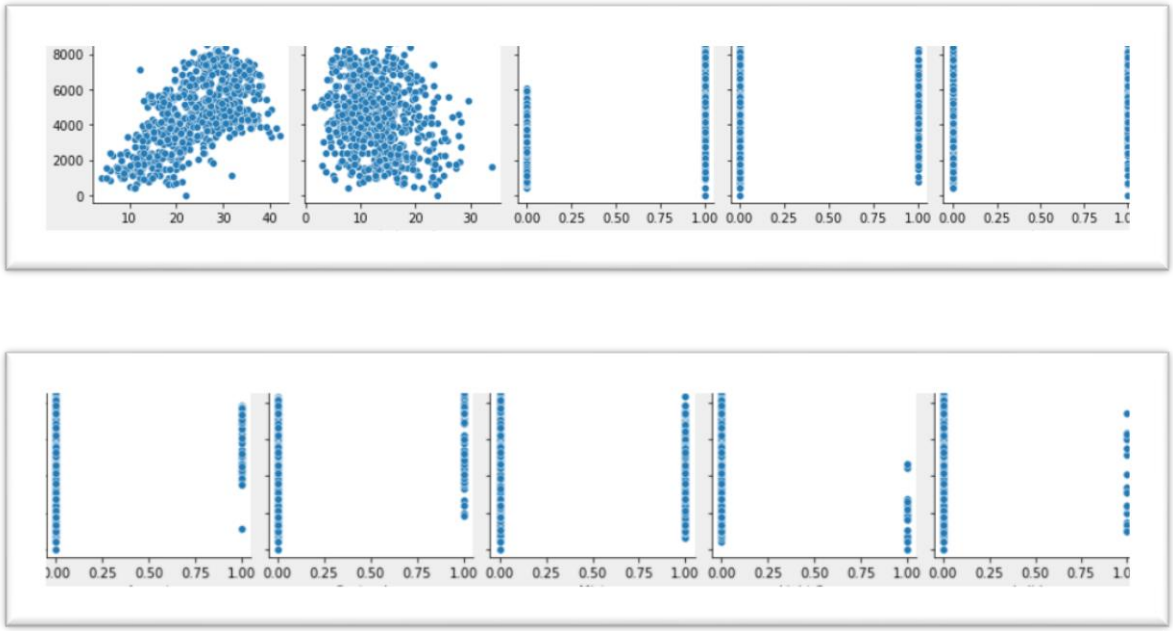

Fig 2: Scatter plots count variable with the top 10 variables

There are only 2 numerical variables in the final set of variables. Only atemp has a linear relationship with cnt and windspeed does not seem to have any relationship with the cnt variable.

b. Checking if the error terms are normally distributed:

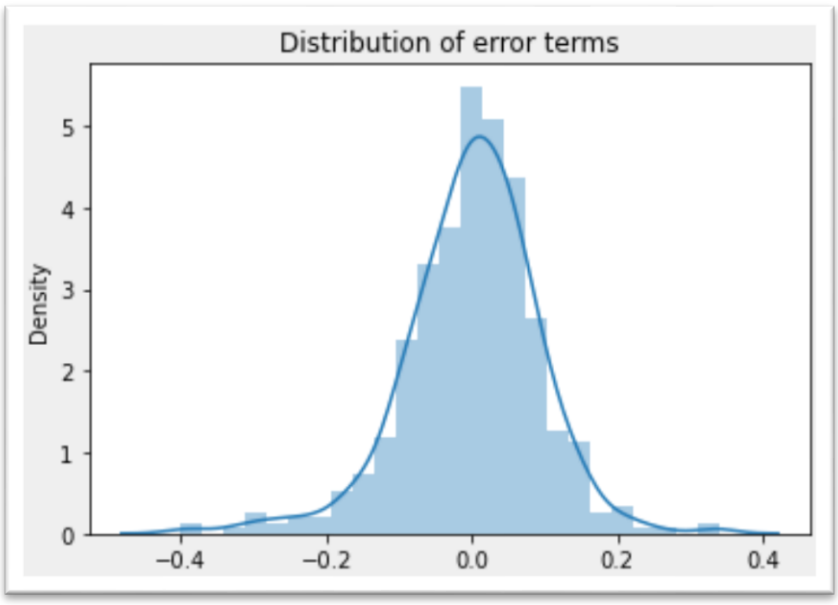

Fig 3: Distribution of error terms

The mean of residuals is 0 and the residuals are distributed normally. This checks out one of the assumptions of linear regression.

c. Checking for independence of error terms:

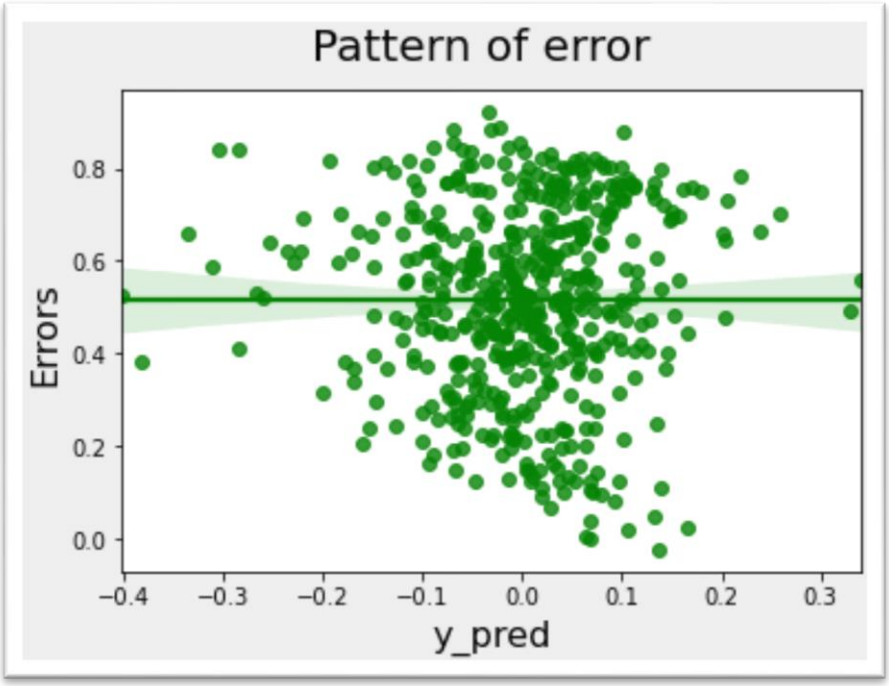

Fig 4: Independence of error terms

The error terms appear to be independent. They don't have a pattern

d. Checking for homoscedasticity:

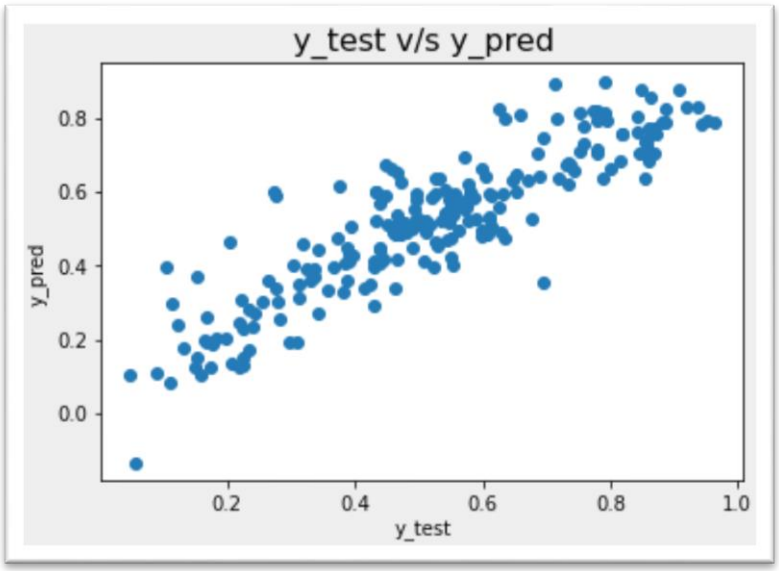

Fig 5: Homoscedasticity of error terms

Errors seem to have a constant variance.

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

   Actual Temperature, Light Snow* and year are the top 3 features. These features were selected due to their coefficients.

   Type of correlation:
   a. atemp is positively correlated to the target variable and hence to the demand of the shared bikes (Coefficient = 0.54).
   b. Light Snow* is the next highest correlation. It is negatively correlated to the demand (Coefficient = -0.28).
   c. Yr (year) is the 3rd highest correlation. It is positively correlated to the demand. (Coefficient = 0.23).

*Light Snow signifies - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   There are 2 categories of machine learning algorithms – supervised and unsupervised. Regression is a type of supervised algorithm. Supervised models are those where the independent variables that are being used to predict the target variable has labels. Example – Season can have 4 labels – summer, winter, spring and fall.

   Linear Regression is a type of regression algorithm. It can be simple or multiple. Simple linear regression is applied when there is only 1 independent variable being used to predict the target. Multiple is applied when more than 1 independent variables are being used to predict the output.

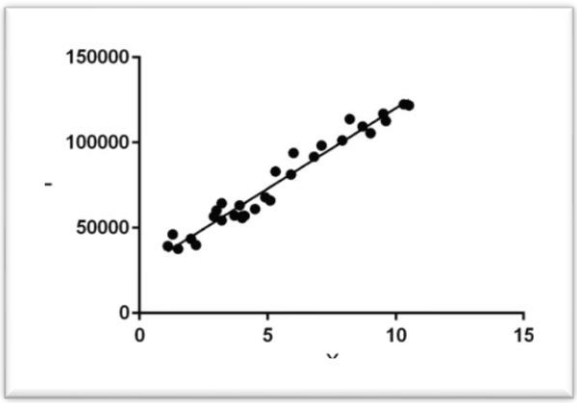

Fig 6: X v/s Y

Here, X is the independent variable which is being used to predict the value of Y (output or dependent or target variable). The line that is drawn through the points is the best fit line or the regression line. The equation of the line is:

$$y = \beta_0 + \beta x$$

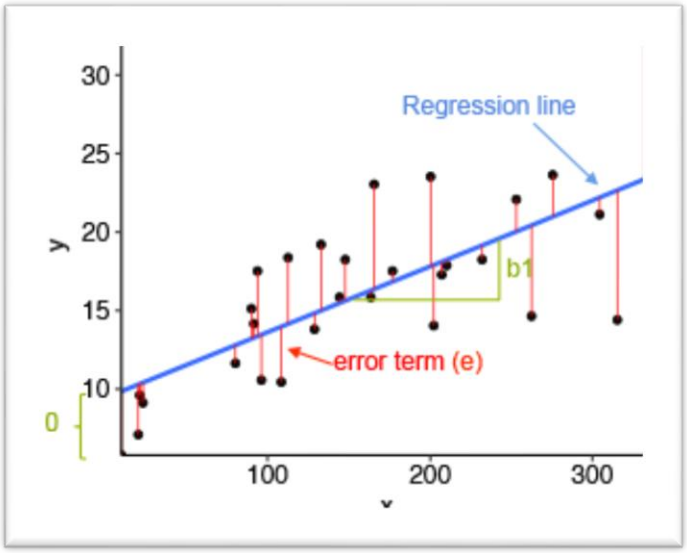

Fig 7: Error terms

This graph is for simple linear regression.

For multiple linear regression,

Say, the independent variables are X_1, X_2, ..., X_n.

The equation of the best fit line will look like this –

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n$$

Finding the best fit line –

y_{pred} - predicted value of y

y_i – actual value of y

A best fit line would be the one that minimises the difference between y_{pred} and y_i. There are several methods to do this –

a. Ordinary least squares method
b. Total sum of squares and R^2 method
c. Residual squared error method

2. Explain the Anscombe's quartet in detail.

A group of 4 datasets that have almost identical statistical properties but appear very different when they are graphed. It shows the importance of doing exploratory data analysis before statistical analysis and the effect of outliers on statistical properties.

Example –

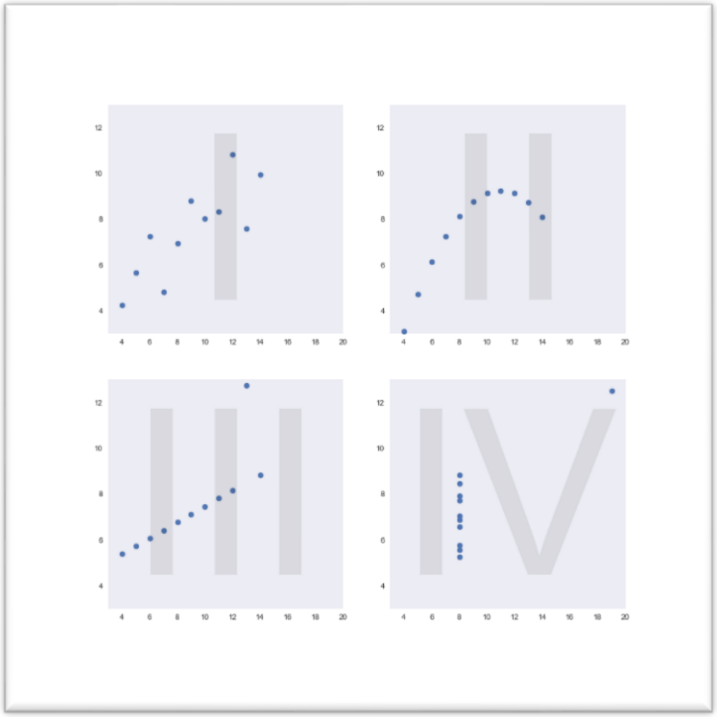

Fig 8: Anscombe's quartet (Graph 1)

The above 4 graphs look very different. But when we find out the best fit line, they appear like this –

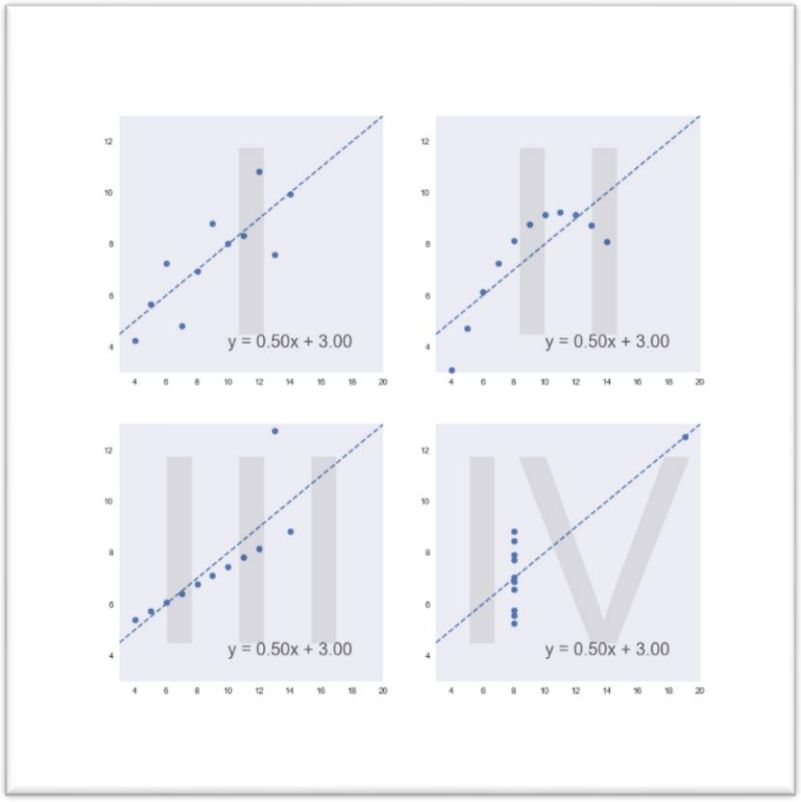

Fig 9: Anscombe's quartet (Graph 2)

So, the best fit line is the same for all of the graphs. Following statistical properties are identical for these graphs –

a. Correlation coefficient
b. Mean of x and mean of y
c. Variance of x and variance of y

3. What is Pearson's R?

Pearson's R is a measure of the strength of linear association between 2 continuous variables – x and y. It is given by -

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

Fig 10: Formula: Pearson's R

The value of r can range from -1 to +1. -1 means that the two variables are highly negatively correlated. One increases with the decrease in another. +1 means that the two variables are highly positively correlated. One increases with the increase in another.

Type of correlation can be depicted by the following graphs -

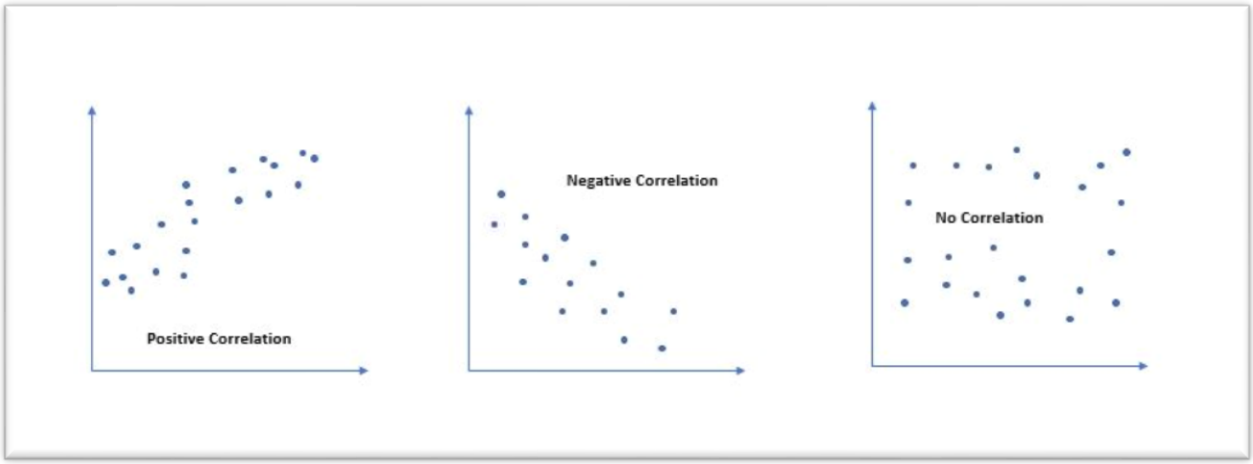

Fig 11: Types of correlation

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

We may come across features in a dataset that have very different magnitudes and units. In order to compare these features to each other and to the output variable, we need to make them of similar magnitudes. Making the values of all the features of similar magnitude is called scaling.

Normalized Scaling	Standardized Scaling
Values are rescaled such that they end up ranging between 0 and 1. It is also called min-max scaling.	Values are scaled such that they are centred around the mean with a unit standard deviation.
Formula: $X` = (X-X_{min})/(X_{max} - X_{min})$ X_{min} and X_{max} are the minimum and maximum values of the feature respectively.	Formula: $X` = (X - \mu)/S$ μ: mean of the feature values S: Standard deviation of the feature values
Good to use when the distribution of data does not follow a Gaussian distribution.	Helpful where the data follows a Gaussian distribution.

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

   VIF is given by,

   $$VIF = 1/(1-R^2)$$

   When R^2 is 1, the value of VIF will be infinite. $R^2 = 1$ means that the model has completely fit the dataset, i.e., it is overfit. All of the variability in y is explained by the model. It also means that there is perfect multicollinearity. In order to solve for this, we need to drop the variable with an infinite VIF value.

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

   Q-Q plot/Quantile-Quantile plot – They plot the quantiles of a given distribution against quantiles of a theoretical distribution. This helps in understanding if a sample distribution follows any particular type of probability distribution like Gaussian, uniform or exponential.

   Types of graphs for normal, uniform and exponential distribution:

   a. Normal distribution: If the sample dataset is normally distributed, we will get almost a straight line. In this graph, the theoretical and sample distribution, both are normal.

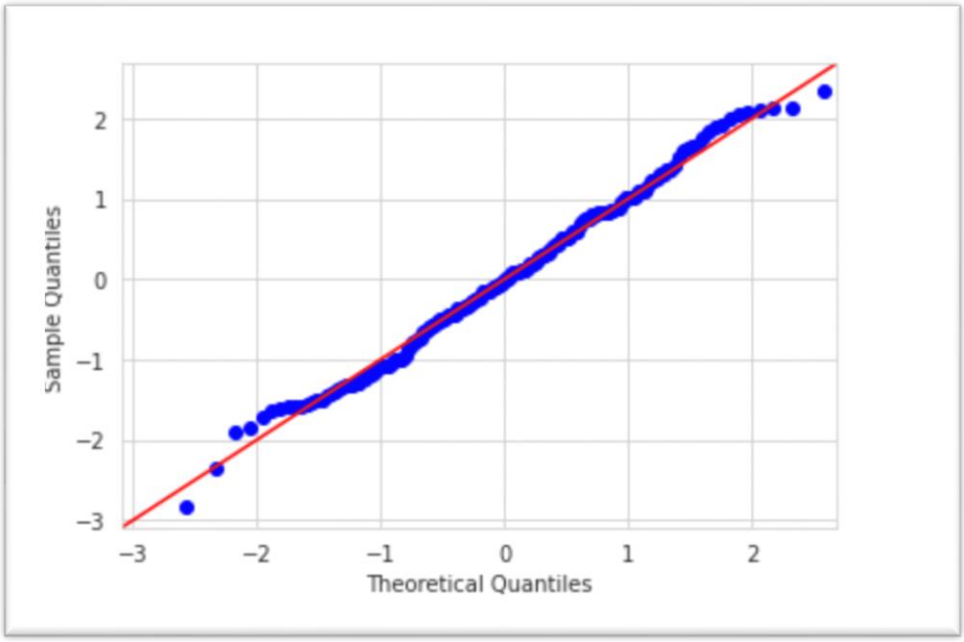

Fig 12: Normal distribution v/s theoretical normal distribution

b. Uniform distribution: If a sample dataset is uniformly distributed and the theoretical one is normally distributed, we get the following graph:

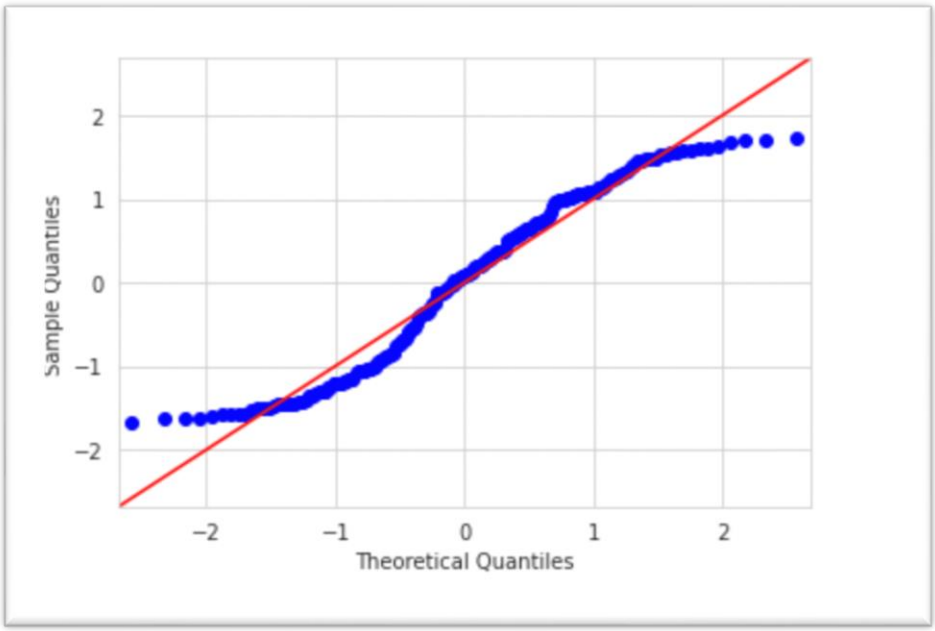

Fig 13: Uniform distribution v/s theoretical normal distribution

c. Exponential distribution: If a variable with exponential distribution is plotted against a theoretical normal distribution, we will get the following kind of graph:

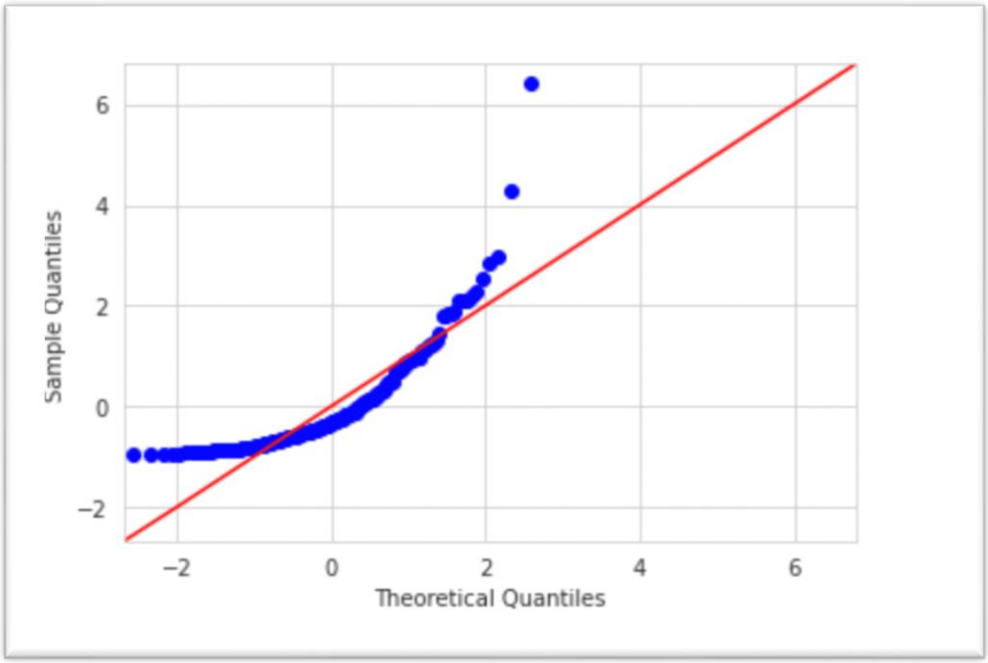

Fig 14: Exponential distribution v/s theoretical normal distribution

Use of Q-Q plots:

a. Determining if 2 populations are of the same distribution
b. Determining if residuals follow a normal distribution
c. Determining the skewness of distribution