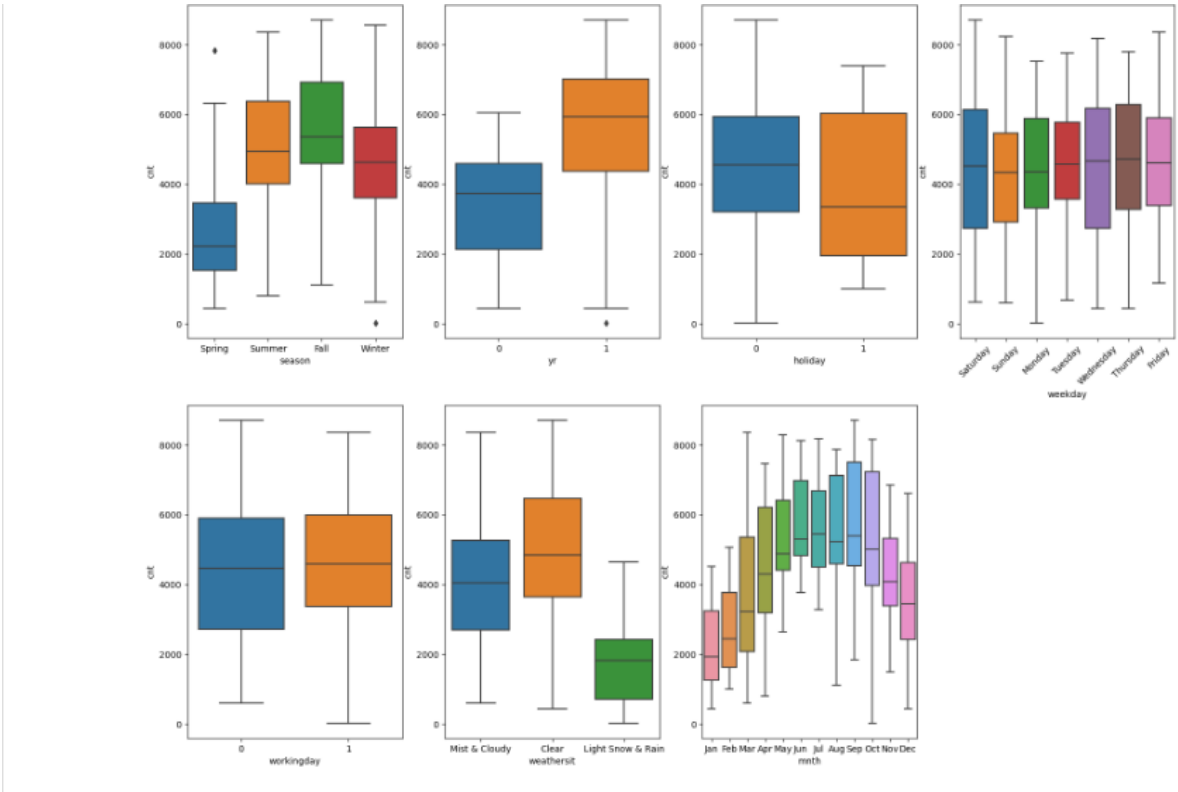
Assignment-based Subjective Questions

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



The categorical variable used in the dataset: season , yr(year) , holiday, weekday ,workingday, and weathersit(weather situation) and mnth(month) . These were visualized using a boxplot.

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

- Season - For the variable season, we can clearly see that the category :Fall, has the highest median, which shows that the demand was high during this season. It is least for: spring.
- Yr - The year 2019 had a higher count of users as compared to the year 2018.
- Holiday - rentals reduced during holiday.
- Weekday - The bike demand is almost constant throughout the week.
- Workingday – From the "Workingday" boxplot we can see those maximum bookings happening between 4000 and 6000, that is the median count of users is constant almost throughout the week. There is not much of difference in booking whether its working day or not.
- Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is quite adverse. Highest count was seen when the weather situation was Clear, Partly Cloudy.
- Mnth - The number of rentals peaked in September, whereas they peaked in December. This observation is consistent with the observations made regarding the weather. As a result of the typical substantial snowfall in December, rentals may have declined

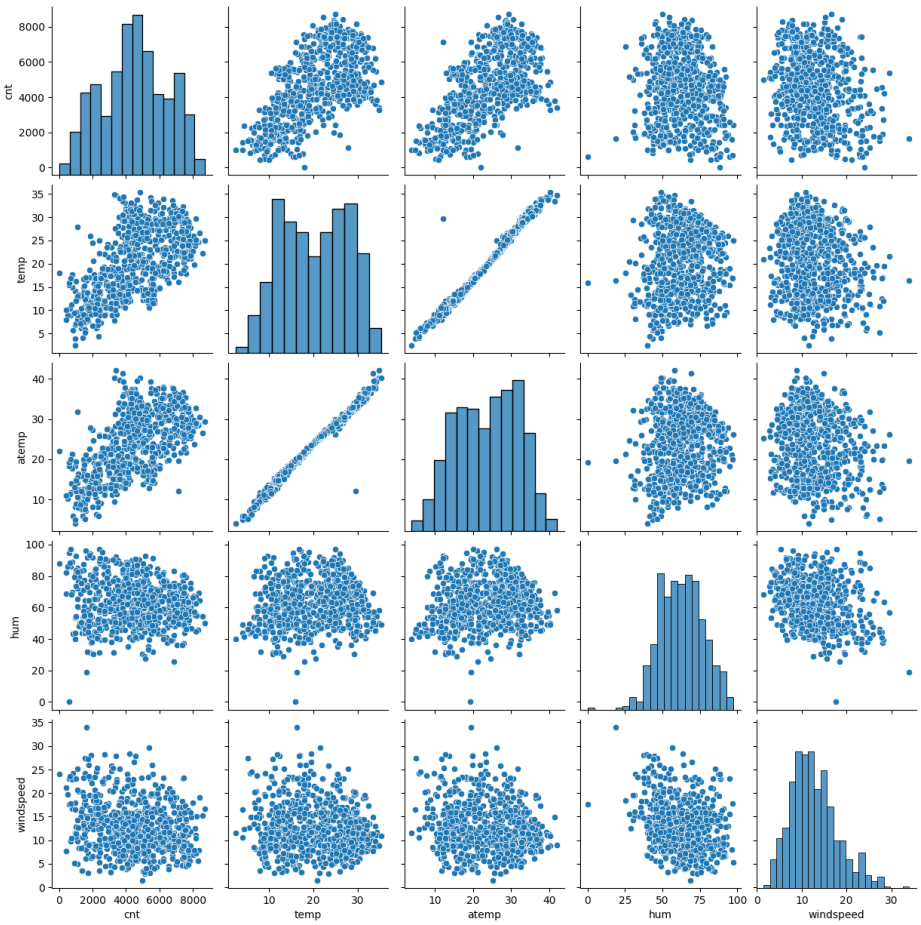
2. **Why is it important to use `drop_first=True` during dummy variable creation?**

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

Consider a Categorical column with 3 types of values, we want to create dummy variable for that column. If one variable is neither furnished nor semi_furnished, then It is unfurnished. So we do not need 3rd variable to identify the unfurnished.

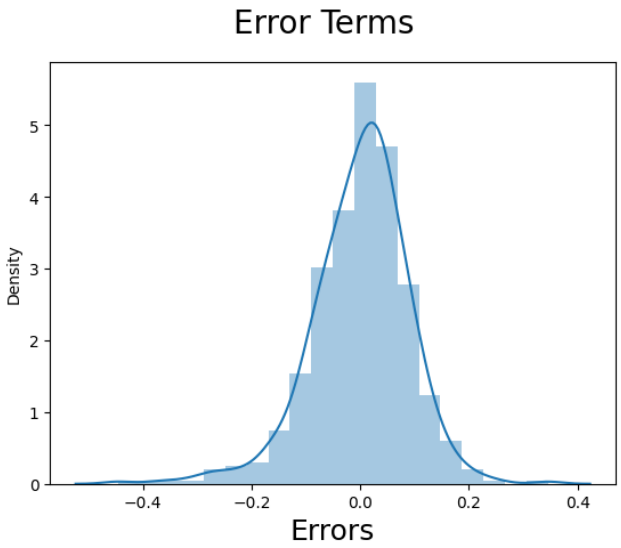
Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
"temp" and "atemp" are the two numerical variables which are highly correlated with the target variable (cnt).



4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   We have done following tests to validate assumptions of Linear Regression:

   a. There should be linear relationship between independent and dependent variables. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not.
   b. Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not.



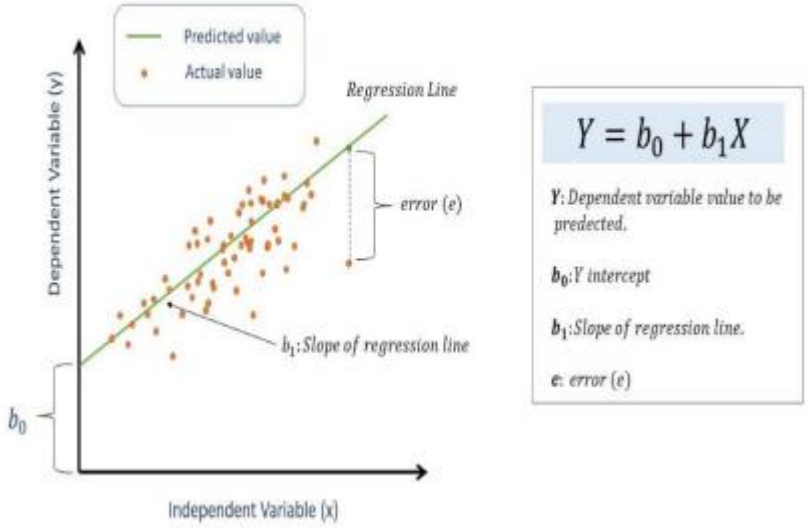
   c. Linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to quantify how strongly the feature variables in the new model are associated with one another.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   The top 3 significant features are:
   1. temp - coefficient : 0.472115
   2. yr - coefficient : 0.234283
   3. weathersit_Light Snow & Rain - coefficient : -0.285425

General Subjective Questions

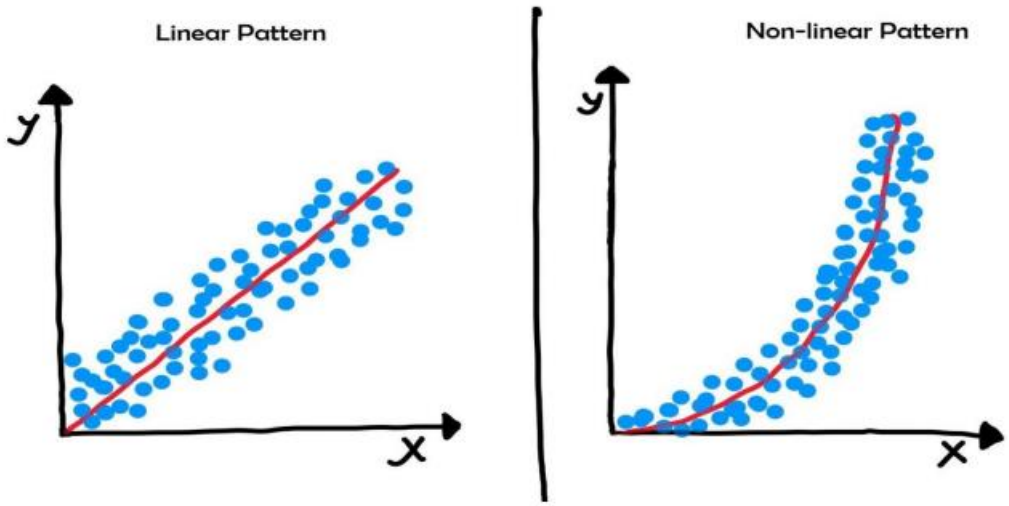
1.Explain the linear regression algorithm in detail.

In simple terms, linear regression is a method of finding the best straight line fitting to the given data, i.e., finding the best linear relationship between the independent and dependent variables. In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Residual Sum of Squares Method. Linear regression models can be classified into two types depending upon the number of independent variables:
Simple linear regression: This is used when the number of independent variables is 1.
Multiple linear regression: This is used when the number of independent variables is more than 1.



The equation of the best fit regression line $Y = \beta_o + \beta_1 X$

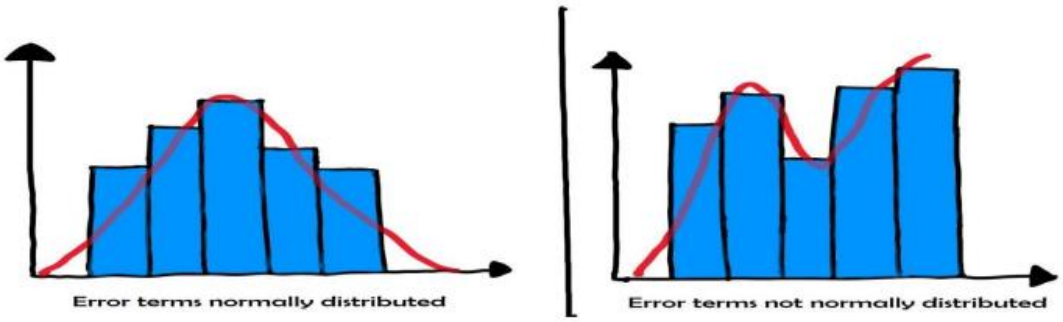
The assumptions of linear regression are:
Assumption about the form of the model: It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the 'linearity assumption'. Assumptions about the residuals:
• Normality assumption: It is assumed that the error terms, ε (i) , are normally distributed.
• Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
• Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ 2 . This assumption is also known as the assumption of homogeneity or homoscedasticity.
• Independent error assumption: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.
Let's understand the importance of each assumption one by one:
1.**There is a linear relationship between X and Y:** X and Y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.



2.Error terms are normally distributed with mean zero(not X, Y):

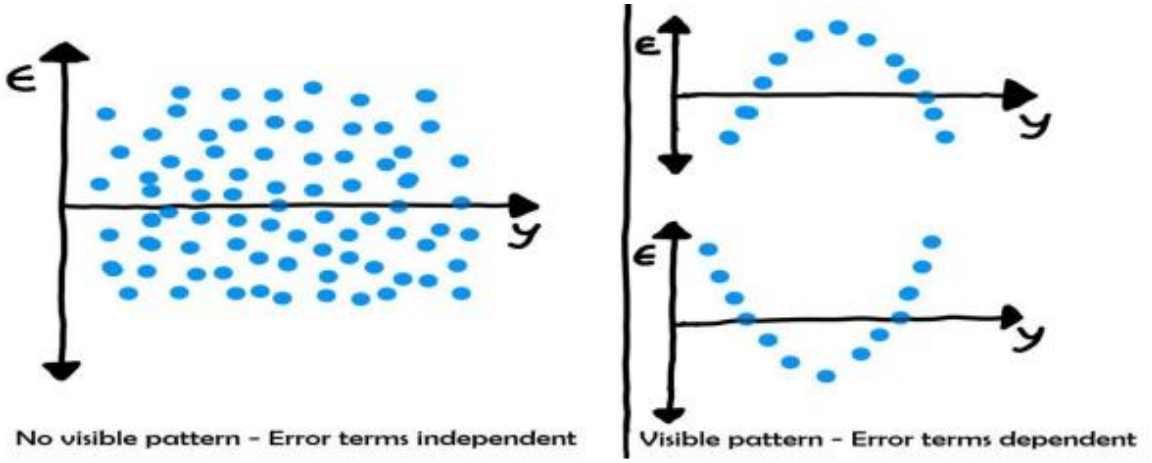
There is no problem if the error terms are not normally distributed if you just wish to fit a line and not make any further interpretations.

But if you are willing to make some inferences on the model that you have built , you need to have a notion of the distribution of the error terms. One particular repercussion of the error terms not being normally distributed is that the p-values obtained during the hypothesis test to determine the significance of the coefficients become unreliable. (You'll see this in a later segment) The assumption of normality is made, as it has been observed that the error terms generally follow a normal distribution with mean equal to zero in most cases.

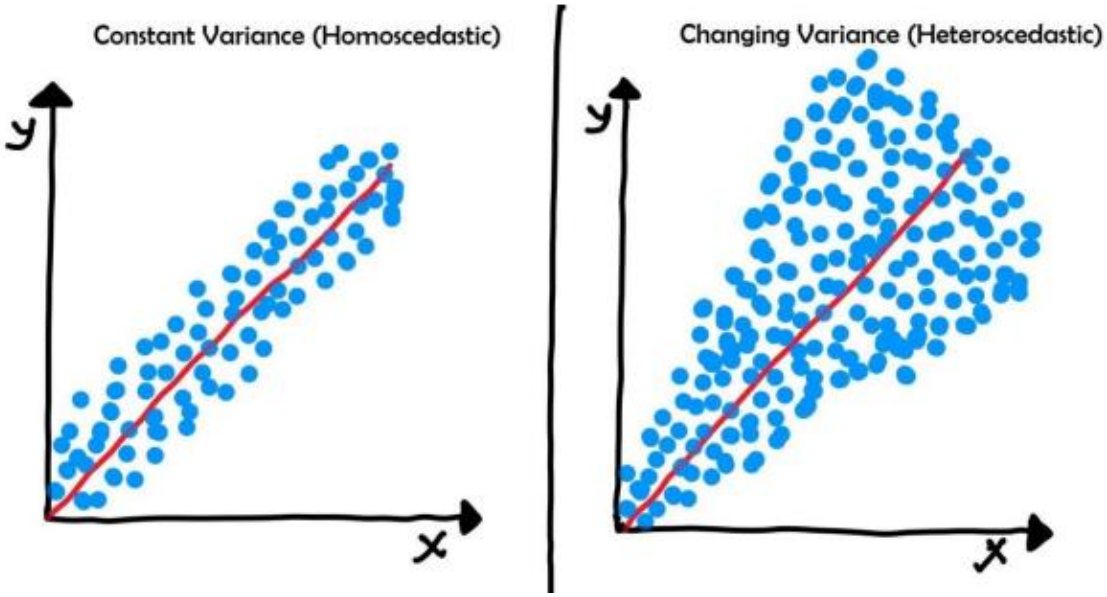


3.Error terms are independent of each other:

The error terms should not be dependent on one another (like in a time-series data wherein the next value is dependent on the previous one).



4.Error terms have constant variance (homoscedasticity):
The variance should not increase (or decrease) as the error values change. Also, the variance should not follow any pattern as the error terms change. Q



2.Explain the Anscombe's quartet in detail.

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 datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

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.

Purpose of Anscombe's Quartet

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.

3.What is Pearson's R?

In Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

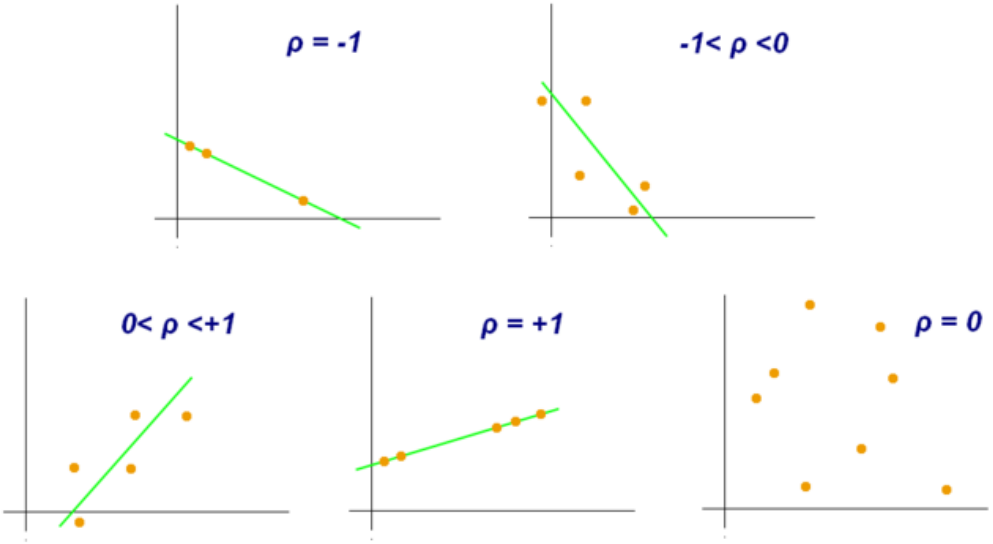
Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Formula

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

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

As can be seen from the graph below, r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association



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

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

● Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.

● Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

Variance inflation factor (VIF) is used to check the presence of multicollinearity in a data set. It is calculated as: Here, VIFi is the value of VIF for the ith variable, Ri2 is the R2 value of the model when that variable is regressed against all the other independent variables. If the value of VIF is high for a variable, it implies that the R2 value of the corresponding model is high, i.e., other independent variables are able to explain that variable. In simple terms, the variable is linearly dependent on some other variables.

The common heuristic we follow for the VIF values is:

> 10: VIF value is definitely high, and the variable should be eliminated.

> 5: Can be okay, but it is worth inspecting.

< 5: Good VIF value. No need to eliminate this variable.

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

The user has to select the variables to be included by ticking off the corresponding check boxes. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well)

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

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions .A Q-Q 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. The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?