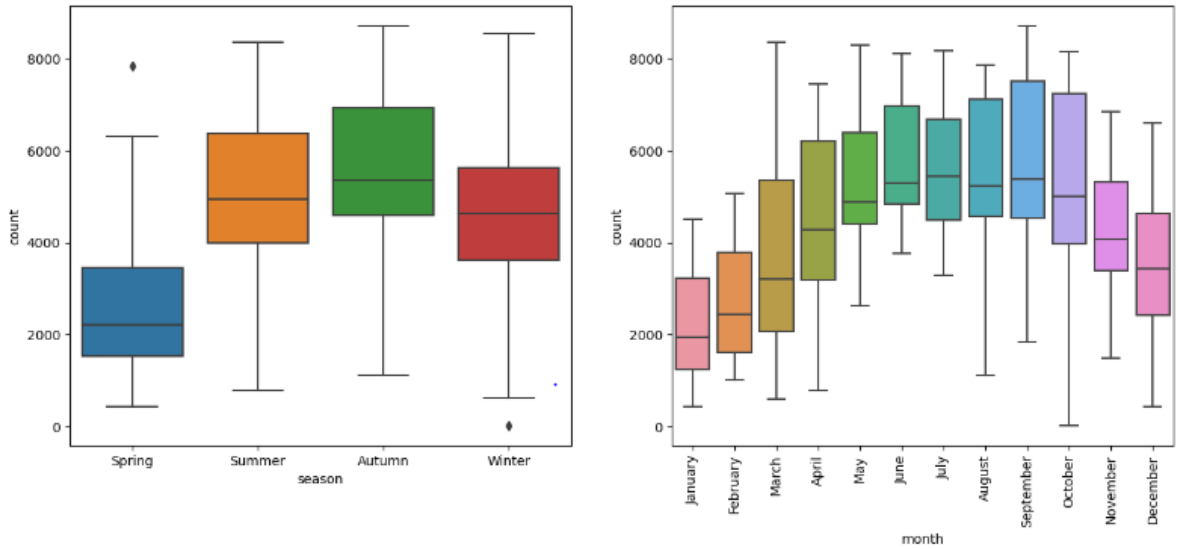
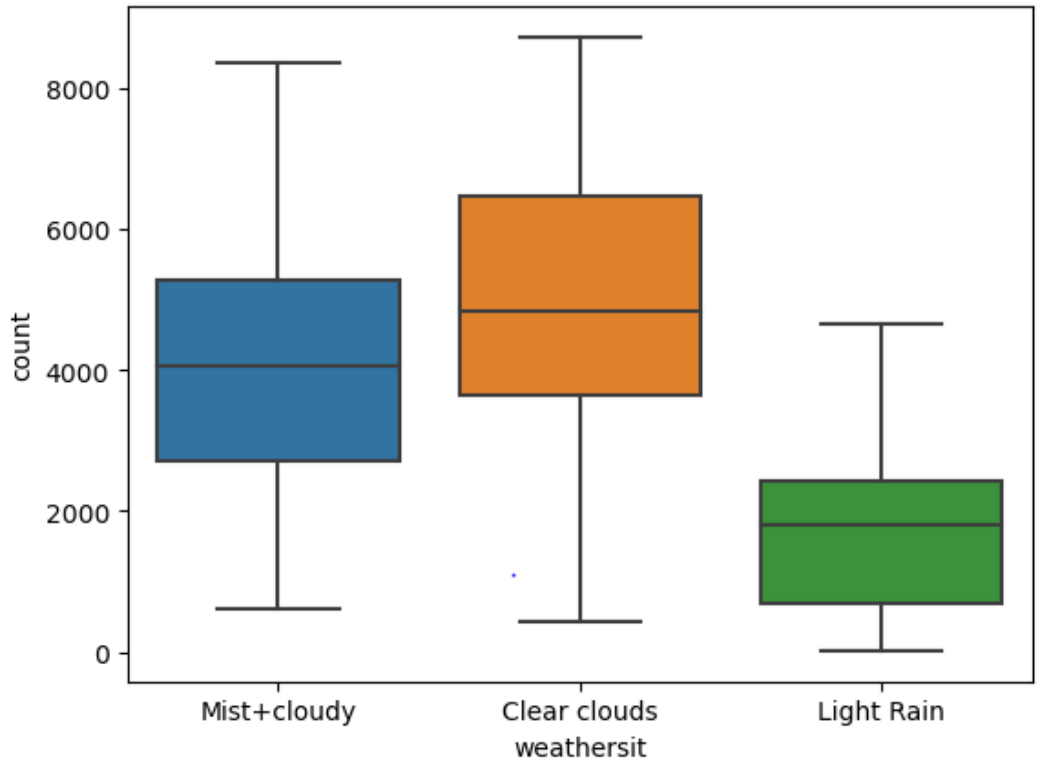
Assignment-based Subjective Questions

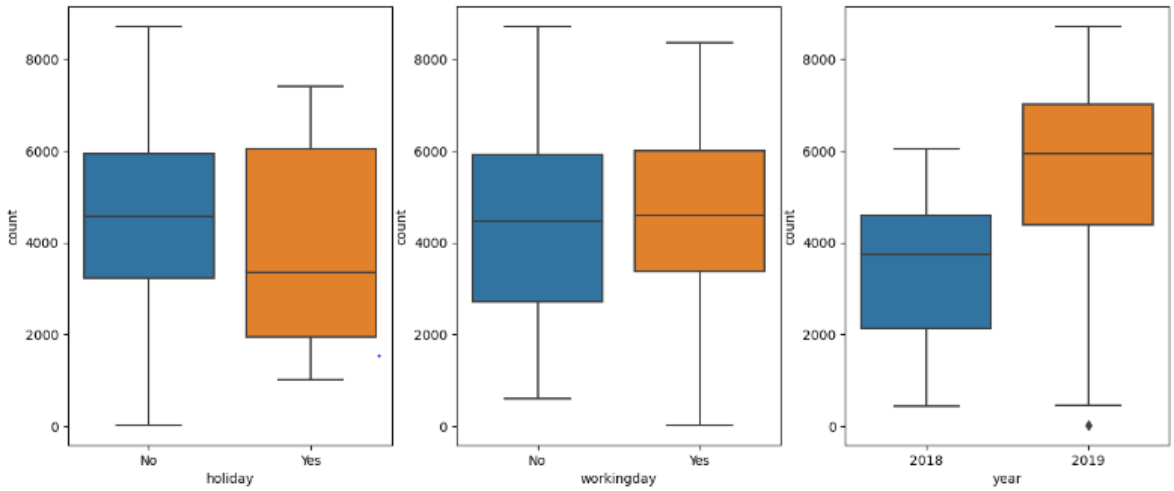
1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



Among seasons, Autumn tend to have high demand, whereas spring has low demand (on considering the Average values) Among months, high demand is from july to September



Among various weather conditions, it is obvious that clear clouds have high demand and Light Rain have low demand



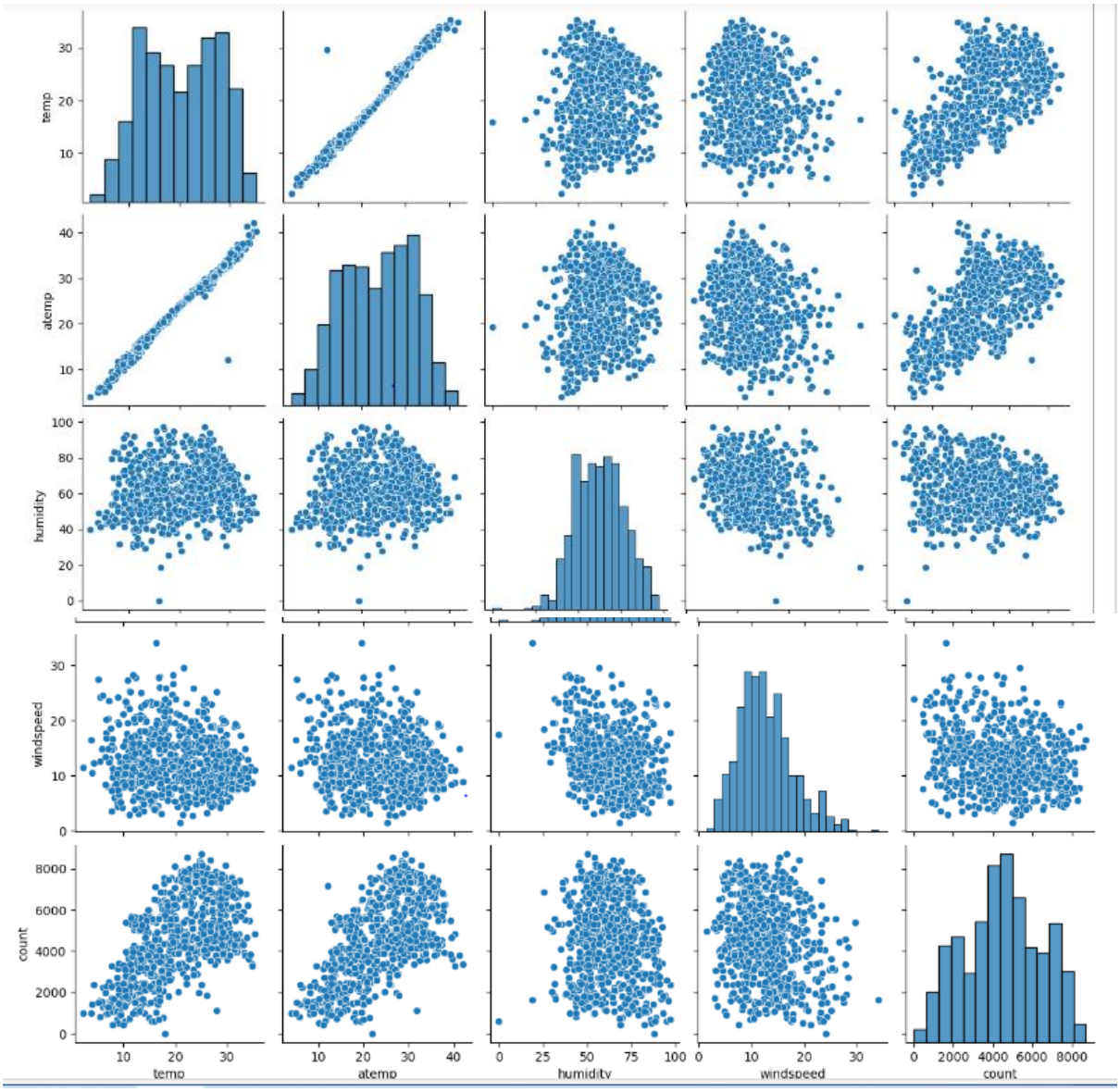
Holidays tend to have less demand than than the working day

Working days have higher demand than than the weekends

Demand in 2019 is higher than 2018.

2) Why is it important to use drop_first=True during dummy variable creation?
Dropping a dummy variables is to avoid redundancy and inflating Coefficients.
   a) Multi collinearity: As the all the dummy variables are included, the inverse of the matrix is not defined. Causing computational complexity for the obtaining the optimized regression coefficients.
   b) Due to high inverse values, the coefficients inflate very high and leading to an inefficient model.

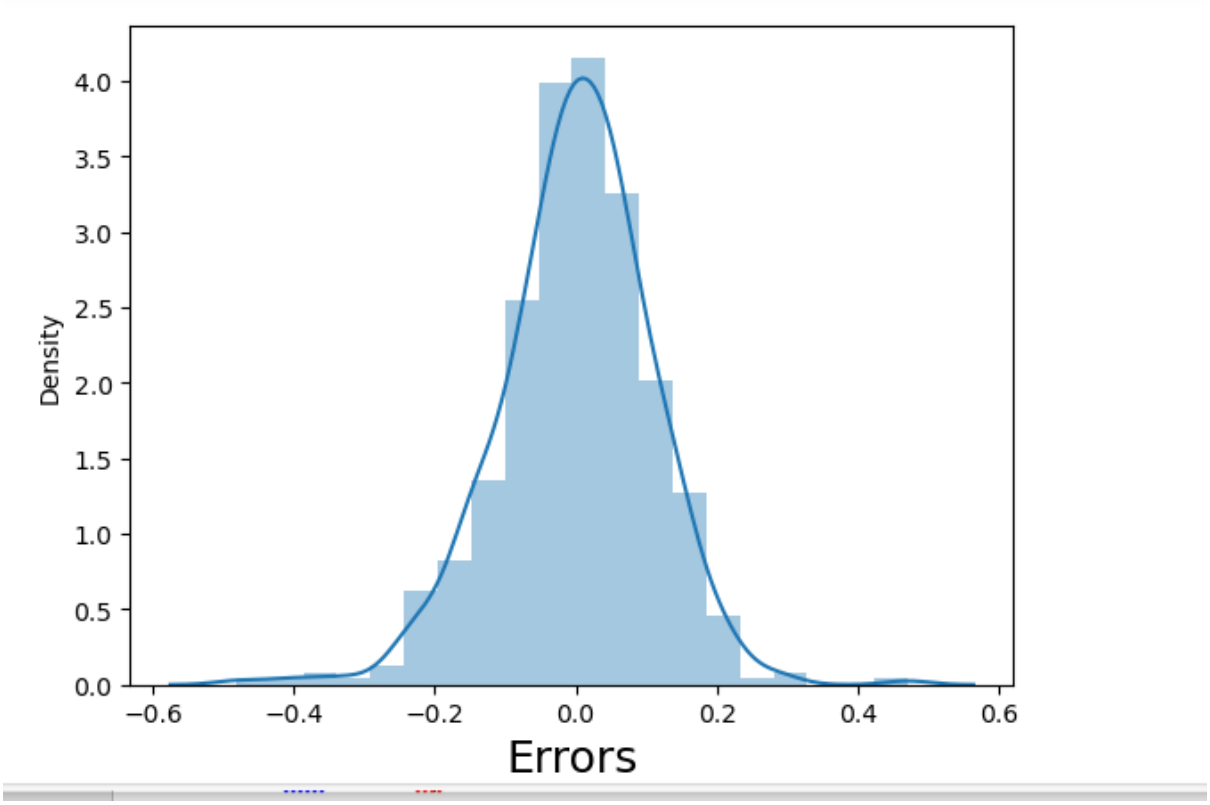
3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



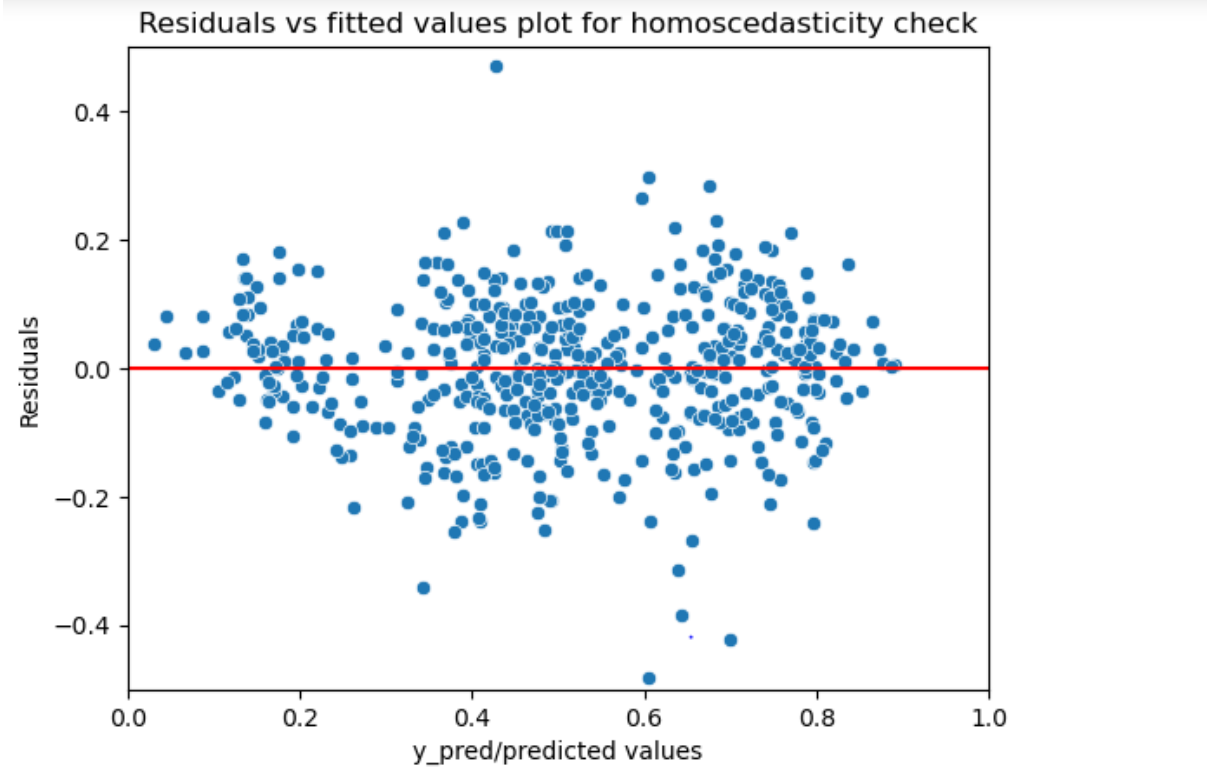
From observation, Count is highly correlated with temp and atemp.

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

Residuals tend to have zero mean



Homoscedasticity Check



Residuals tend to be around zero and very close to each other.

Multi collinearity Checks:

	Features	VIF
0	windspeed	4.00
7	workingday_Yes	3.29
1	season_Spring	2.00
2	season_Summer	2.00
4	year_2019	1.88
3	season_Winter	1.73
6	weekday_Monday	1.56
9	weathersit_Mist+cloudy	1.56
5	month_September	1.18
8	weathersit_Light Rain	1.08

VIF tends to have value less than 5, so there are no multi collinear effects.

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

Top 3 features are:

1. Year_2019 : 0.2475
2. Light Rain: -0.3
3. Season_spring : -0.2969

General Subjective Question

1) Explain the linear regression algorithm in detail

Linear regression identifies the relationship between the dependent variable and a single or multiple variables.

Linear Regression Equation:

$Y= C_0 + C_1X_1 + C_2X_2+....C_pX_P + \varepsilon$

Where Y is the dependent or target variable

X_1 , X_2 ,….,X_p are independent variables.

C_0 –Intercept (constant)

C_1, C_2 , C_3 ….are coefficients of the independent variables.

Steps followed in linear Regression model:

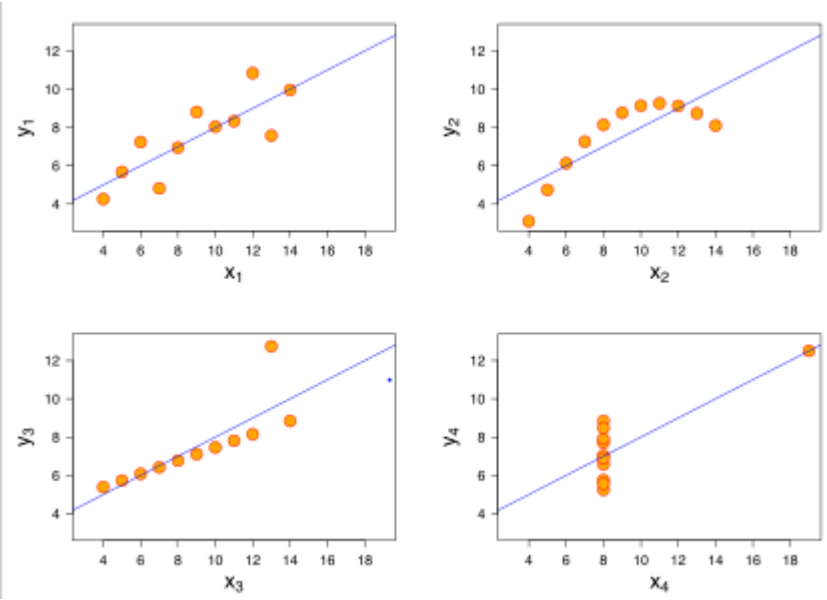
a) Collection of Data Points: Collecting the data from various sources and combining it for the analysis
b) Data Pre-Processing: Checking the dataset for any abnormalities or outliers. Treatment of missing values/outliers.
c) Data Visualization: To get an initial idea on the variables and its relationship with other variables. Graphical plots are utilized.
d) Data Processing : Scaling the numerical variables for the final analysis.
e) Splitting the data set in to train and test data
f) Training the regression model on the train dataset to find the relationship between the dependent and the independent variables. Finding the Coefficients using gradient algorithm.
   Gradient descent method minimizes the cost function (square od residuals) for the coefficients.
g) Model Evaluation: Check the trained model on the test data on key parameers such as Adjusted R-squared, MSE etc..
h) Check for linearity , normality and Homosdescadicity of errors.

2) Explain the Anscombe's quartet in detail.

Quartet comprises of 4 datasets in which statistical properties are identical, but when plotted on a graph they showcase a different distribution.

Developed to empower the importance of visualization prior to analysis.

Also studies the importance of outliers and other significant statistical properties.



i) Dataset -1 (Top Right) : Linear Relationship and well fitted by linear regression

ii) Data set-2 (Top left) : Nonlinear Relationship .Have similar statistical properties to dataset -1 but here it is not well suited for linear regression.

iii) Data set-3 (Bottom left) : Linear Relationship. The fitted line is influenced by an outlier. If the outlier has been eliminated, then the fitting would be more perfect. Shows the influence of outliers on the final regression model.

iv) Data set-4 (Bottom left): one outlier has tilted the regression line enormously. If the outlier is eliminated, the developed regression would be completely different.

3) What is Pearson's R?

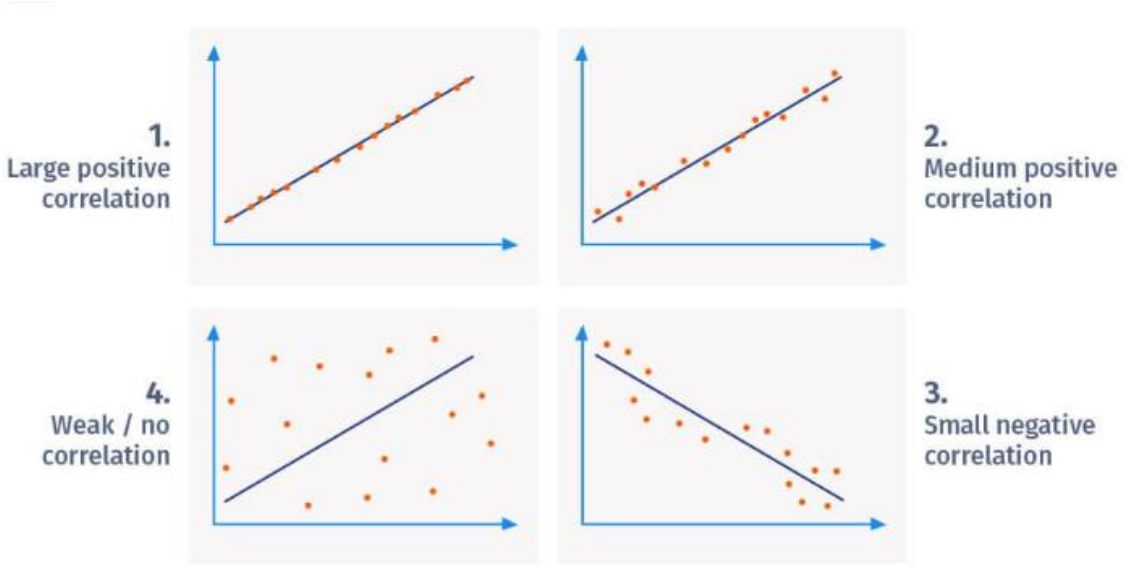
Ans: It means the variability of one variability with respect to other variable. It varies between -1 and +1.

It is calculated using the formula below:

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

X_i and Y_i are the values of X and Y values.

$\bar{x}$ and $\bar{y}$ are the mean values of X and Y.



Scattered plots tend to have low r values.

If both the values tend to rise or fall simultaneously, the R>0

If one of the tend to drop when the other increases or vice versa, then R<0

If there is no variation in a variable, when the other changes, then R=0.

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

Scaling and why it is performed: When multiple variables are involved in the analysis, then all these numerical would be at different scales. When these variables are used in analysis, larger numerical tend to dominate the lower denominations. To avoid this error, all the variables are scaled to bring at the same level.

Normalization Scaling : Varies between 0 and 1.

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

Values ranges between o and 1.

It is sensitive to outliers. It is followed when the data does not follow Gaussian distribution.

Standarization Scaling: Used when the data follows Gaussian disrbution.

Standardization (or Z-Score Scaling) transforms the data to have a mean of 0 and a standard deviation of 1. Use when the data is normally distributed.

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

Variance Inflation factor:

Measure used to identify the presence of multicollinearity in regression model. It occurs when two or more independent variables are highly correlated to each other. In such cases the regression coefficient are inflated.

The VIF of a predictor X:

$1/(1-R^2)$, R^2 is the regression coefficient when X is regressed with other predictors.

VIF is infinite:

It happens only when R^2 is 1 in the above equation, it conveys that a variable is in absolute correlated with one or more variables(predictors)

6) What

A Q-Q plot is a tool used to identify whether the data set follows a certain distribution. It plots the quantiles of the sample data against the quartiles of any theoretical distribution.

Sometimes two datasets can be comparing by plotting the quantiles of the two data sets.

Mainly to check:

a) The distribution of a sample.
b) Do the two datasets belong to the same distribution.
c) Do they have similar shapes.