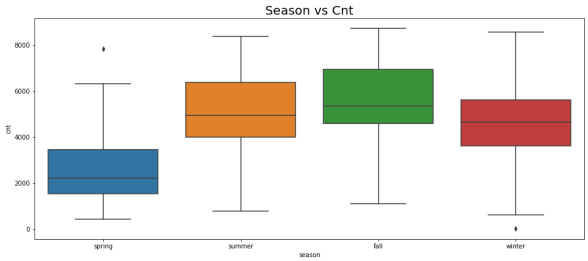
<u>This solution has been attempted by-</u>

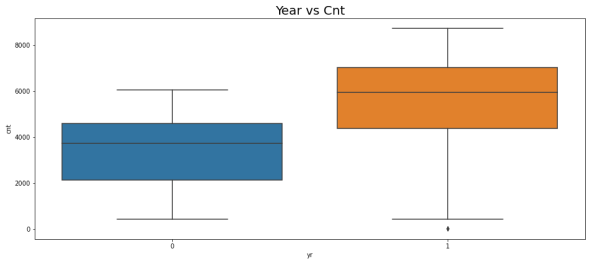
Name: Satvik Praveen
Batch: DS C46 July Batch

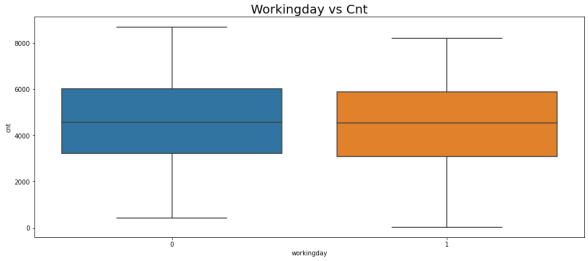
Assignment-based Subjective Questions

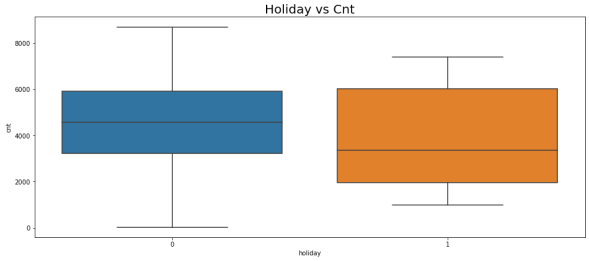
From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

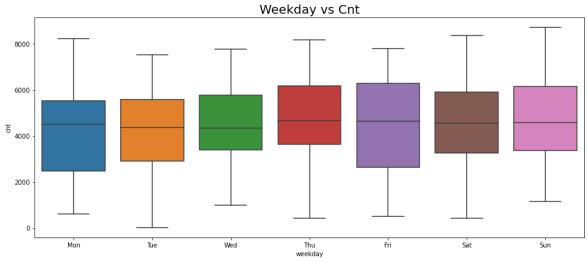
The analysis performed in the bivariate analysis of the EDA let us understand the following about the categorical variables:

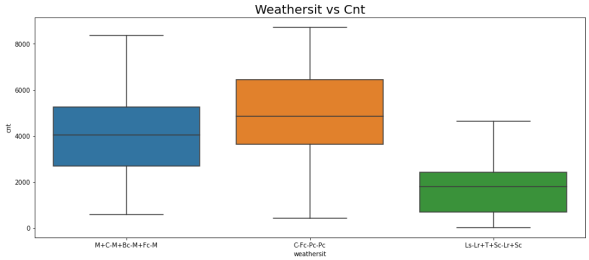


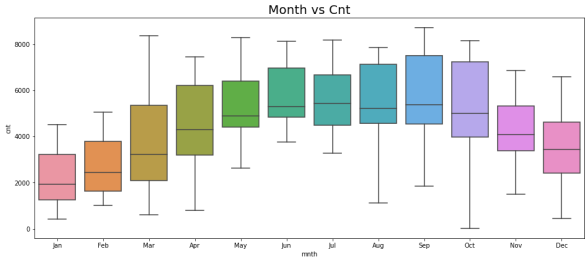












- The plot based on year reveals that more bikes were rented in the year 2019 in comparison to the year 2018.
- The plot based on season reveals that more bikes were rented during fall season followed by summer season.
- The holiday and working day plots show that more bikes are rented during the working days as compared to holidays.
- The sept month is the dominant month in the plot based on month (‘mnth’ vs ‘cnt’) in which more bikes are rented.
- The weekday plot depicts that more bikes are rented on Saturday.
- The plot based on ‘weathersit’ indicates that bikes were rented more when the weather was Clear, Few clouds, Partly cloudy weather..

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

A categorical variable with n levels can be completely described using n-1 dummy variables which indicates that we can drop one of the created variables, Therefore, we use the drop_first = True which helps in reducing the extra column created during the creation of dummy variables. For example: if we have a ‘season’ categorical column which has three levels: summer, winter, rainy. Then, we can describe the categorical column using any two levels as shown:

Before creation of dummy variables:

Season
summer
winter
rainy

After creation of dummy variables:

winter	**rainy**
0	1
1	0
0	0

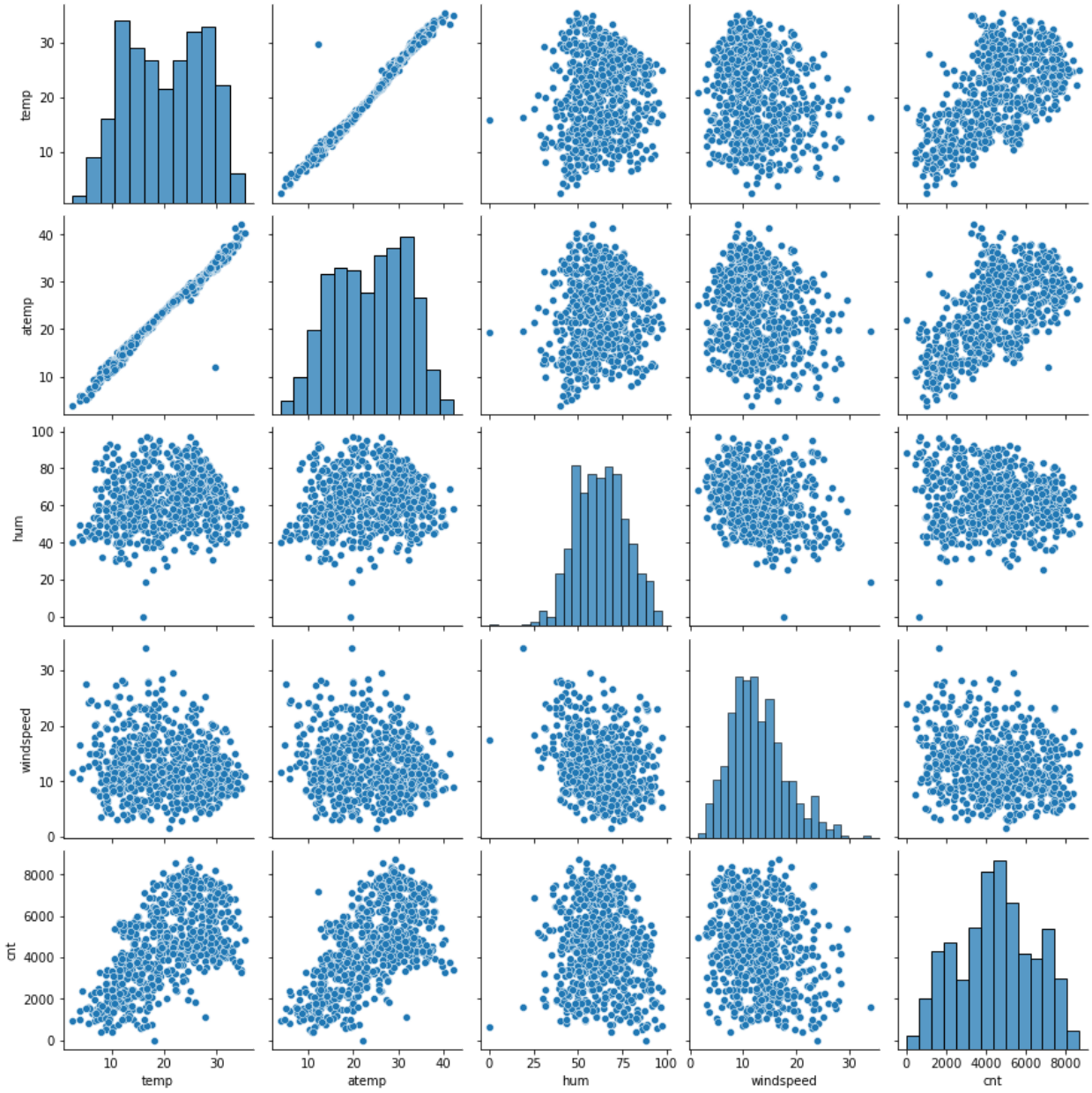
Case-1 01 indicates rainy.
Case-2 10 indicates winter.

Case-3 00 indicates summer.

Clearly, we could describe the column season with three levels using two dummy variables

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

The pair-plot among the numerical variables is as follows:

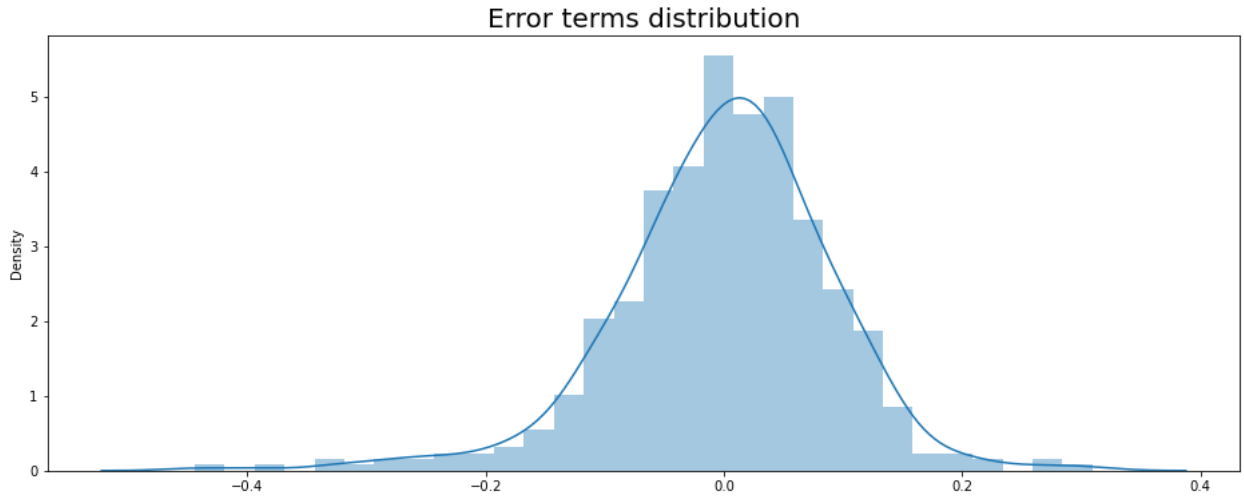


The variables temp and ‘atemp’ seem to have the highest correlation with the ‘cnt’ variable which is the target variable.

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

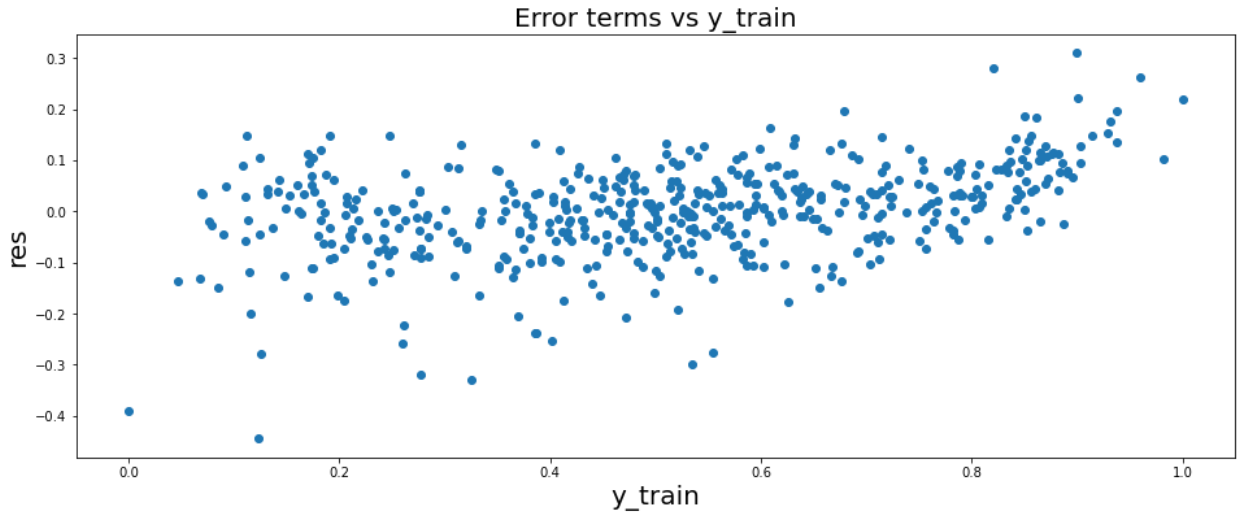
The assumption of Linear Regression were validated using the following:

Plotting the distribution of residuals:



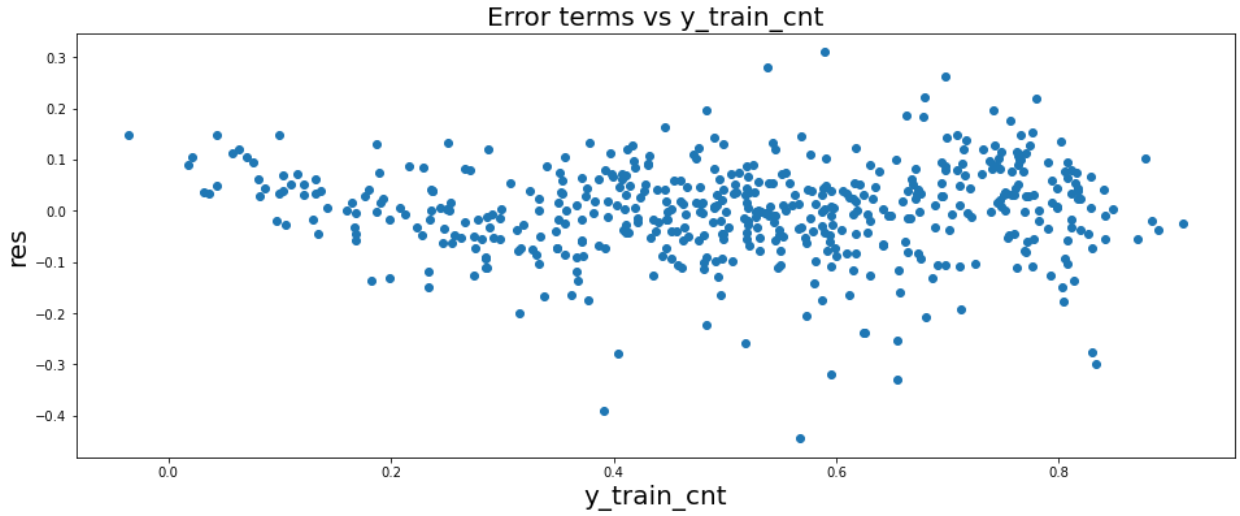
Clearly, the residual distribution is a normal distribution which is centered at 0. Hence, the assumption of normality got verified.

Plotting scatter plot between y_train and residuals:



As we don't see any pattern in the plot. This justifies that the error terms are independent of each other. This justifies the assumption that error terms are independent of each other.

Plotting scatter plot between y_train and residuals:



We can observe that the variance of the error terms remains constant as the value of the error terms changes and hence, the assumption of homoscedasticity is justified here.

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

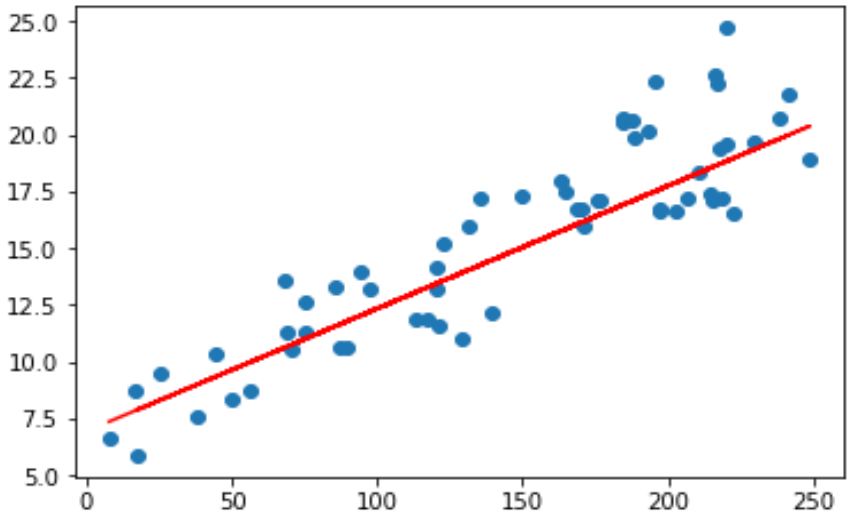
The top 3 features which contribute significantly towards explaining the demand of the shared bikes are as follows:

1. atemp: feeling temperature in Celsius
2. Ls-Lr-+T+Sc-Lr+Sc: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
3. yr: year

General Subjective Questions

Explain the linear regression algorithm in detail.

Linear regression is a machine learning algorithm. It is based on supervised learning in which we have historic data with labels to train the models we intend to design. The regression algorithm is used to modal a target prediction value using independent variables. This can be used to determine the relationship between variables and forecasting. The models differ based on the relationship they design between dependent and independent variables and the number of predictor variables (independent variables, regressors, exogenous variables). The dependent variable is also known by outcome variable, endogenous variable, or regressand.



The task of prediction of the value of a dependent variable is performed by linear regression using the independent variables. The dependent variable is denoted by y and the independent variable is denoted by X. The algorithm establishes a linear relationship between input and output. This relationship given by the hypothesis equation:

$$y = \beta_0 + \beta_1 * X$$

The best - fit line is obtained by minimizing the residual sum of squares (RSS), which is given by:

$$\text{RSS} = \sum_{i=1}^{n} (y_p - y_i)^2$$

where y_p denotes the predicted value and y_i denotes the individual point for a given input. RSS is basically the sum of the squares of residuals at each data point.

Initially, we are given X: input training data, y: historical data with labels. We fit the best line while training the model to predict the value of y for a value of X. The model gets the best fit line by finding best values for betas. Beta_0 is the intercept and Beta_1 is the slope of the line or coefficient of X.
We follow the following method for updating Beta_0 and Beta_1.
We define the cost function J, which is the Root Mean Square Error (RMSE). By minimizing the cost function, we not only reduce the error between the predicted value and the obtained value, but also obtain the best fit line for the model.

$$J = \frac{1}{n} \sum_{i=1}^{n} (y_p - y_i)^2$$

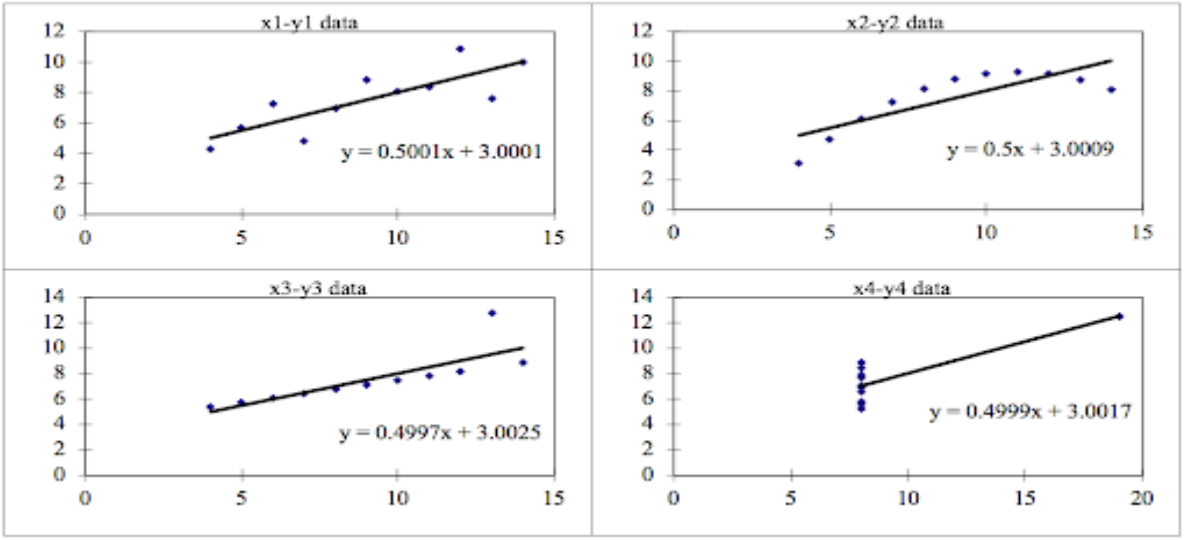
We minimize the cost function, J, using the gradient descent algorithm. The idea behind the gradient descent algorithm is that we start with random values of Beta_0 and Beta_1. Then, we iteratively update their values in order to minimize the cost.

Explain the Anscombe's quartet in detail.

Consider that we have a dataset which has a particular set of summary statistics such as mean, standard deviation, correlation coefficient and line of best fit. Now, consider three other datasets which produce the similar summary statistics. We might think that all these datasets have similar kinds of distribution. However, if we plot the scatter plots for all the datasets, we might realize that the distribution of data points are quite different. When these plots are collated in a single frame, they are called Anscombe's quartet.

Consider the case below:

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	



Clearly, this tells us the importance of visualizing the data before applying various algorithms to build the models out of them. Plotting the data also helps us identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Further, linear regression can be considered a fit for the data with linear relationships and is incapable of handling any other kinds of datasets.

What is Pearson's R?

Pearson's R is also known by the following:

- Pearson correlation coefficient
- Bivariate correlation

- Pearson product-moment correlation coefficient
- The correlation coefficient

It is a descriptive statistical measure that is used to summarize the characteristics of a dataset. The magnitude of Pearson's R describes the strength of the linear relationship between two quantitative variables. The sign depicts the direction, i.e, +ve sign denotes a direct relationship whereas -ve sign denotes an inverse relationship.
The range of values generally lie in between -1 and 1.

The following table is generally considered for reference in order to make an inference of the relationship strength (which is also known as effect size) between two quantitative variables:

Pearson's R	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between -.3 and 0	Weak	Negative
Between -.5 and -.3	Moderate	Negative
Lesser than -.5	Strong	Negative

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

Scaling is a technique to standardize the independent features present in the dataset in a fixed range.
For instance,

- In the case of standardized scaling, the scaling is done such that the entire features gets transformed to have a mean of 0 and standard deviation of 1.
- In case of min-max scaling also known as normalization, we scale the features so that all the data points are between 0 and 1.

Scaling is performed for the following reasons:

1. It speeds up the process of gradient descent algorithm
2. The features that can have different ranges of magnitude of data points, are scaled to have similar characteristics so that comparisons can be made between the coefficients of the features while modeling using linear regression.

If scaling is not performed before modeling then the algorithm shall only take the magnitude into account and not the units. This shall result in incorrect modeling.
incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

The difference between normalized scaling and standardized scaling are as follows:

Normalization	**Standardization**
Uses minimum and maximum values to scale the features.	Uses mean and standard deviation to scale the features.
Scales between [0,1]	The scales are not bounded in a certain range.
Used when features are of different scales.	Used when features are to be set with 0 mean and unit standard deviation.
Affected by outliers.	Much less affected by outliers.
MinMaxScaler from Scikit-Learn is used to perform min-max scaling or Normalization.	StandardScaler from Scikit-Learn is used for performing Standardization.
Useful when we do not know about the distribution.	Useful when we know the feature distribution is Normal or Gaussian.
Often called Scaling Normalization.	Often called Z-Score Normalization.

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

The expression for VIF for a feature is:

$$VIF_i = \frac{1}{1-(R_i)^2}$$

Clearly, the VIF of a feature can be infinite if its R-squared value is 1, which means the variance in the feature can be completely explained by another feature of the dataset. That is, there is a perfect correlation between two independent variables. In other words, the corresponding variable can be completely expressed as a linear combination of other variables (which shows

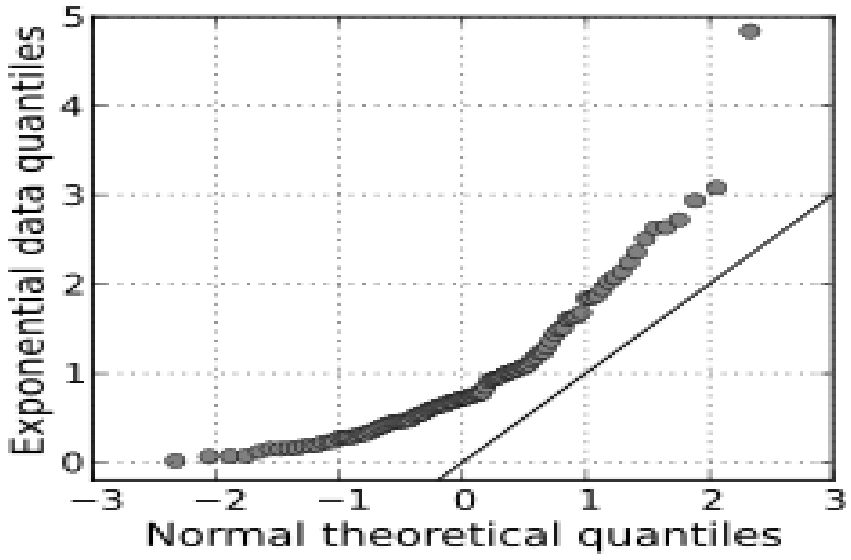
infinite VIF as well) Thus, the given feature does not add any value to our model as it brings a high degree of multicollinearity with itself. Hence, we prefer to drop it.

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

A plot of quantile against quantile for two distributions is called Q-Q plot or Quantile-Quantile plot. A quantile is a fraction which signifies certain values that fall below that quantile.
For instance,
median is a quantile where 50% of the data fall below that point and 50% lie above.
30% quantile is the point where 30% of the data fall below it and 70% lie above it.

The use of Q-Q plots is to find out if two sets of data come from similar distribution. A 45 degree angle is plotted on the Q-Q plot to examine this. If the two data sets come from a common distribution, the points will fall on that reference line. Further, the sample sizes need not be equal and many distributional aspects can be simultaneously tested.

A Q-Q plot showing the 45 degree reference line is as shown:



If the two distributions being compared are similar, the points in the plot shall approximately fall on the line y=x(slope of 45 degree), but not necessarily on the line itself. The greater the departure from the line, the better the evidence that the two datasets have come from populations with different distributions. This can also be utilized as graphical estimation of parameters in a location-scale family of distributions.

The Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how the properties such as location, scale, and skewness are similar or different in two distributions.