

#### 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) The year box plot indicates that more bikes are rent during 2019.

The season box plot indicates that more bikes are rent during fall season.

The working day and holiday box plot indicates that more bikes are rent during normal working days than on weekends or holidays.

The month box plots indicates that more bikes are rent during Sept.

The weekday box plots indicates that more bikes are rent during Saturday.

The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, Partly cloudy weather.

2. Why is it important to use drop\_first=True during dummy variable creation?

A) It 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.

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

A) The 'temp' variable 'registered' has the highest correlation with the target variable 'cnt' .

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

A) From R-Squared and adjusted R-Squared value of both train and test dataset we could conclude that the above variables can well explain more than 80% of bike demand. Coefficients of the variables explains the factors effecting the bike demand.

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

A) The Top 3 features contributing significantly towards the demands of share bikes are: weathersit\_Light\_Snow(negative correlation) and yr\_2019(Positive correlation) and temp(Positive correlation).

#### General Subjective Questions

1. Explain the linear regression algorithm in detail.

A) Linear regression is one of the basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. Linear Regression Algorithm is a machine learning algorithm based on supervised learning. It is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Linear regression equation can be written as:  $y = mx + c$  ,  $m$  = Slope of the line,  $c$  = y-intercept of the line,  $x$  = Independent variable from dataset,  $y$  = Dependent variable from dataset.

## 2. Explain the Anscombe's quartet in detail.

- A) 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 analysing it and the effect of outliers on statistical properties. Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points.

Those 4 sets of 11 data-points are given below.

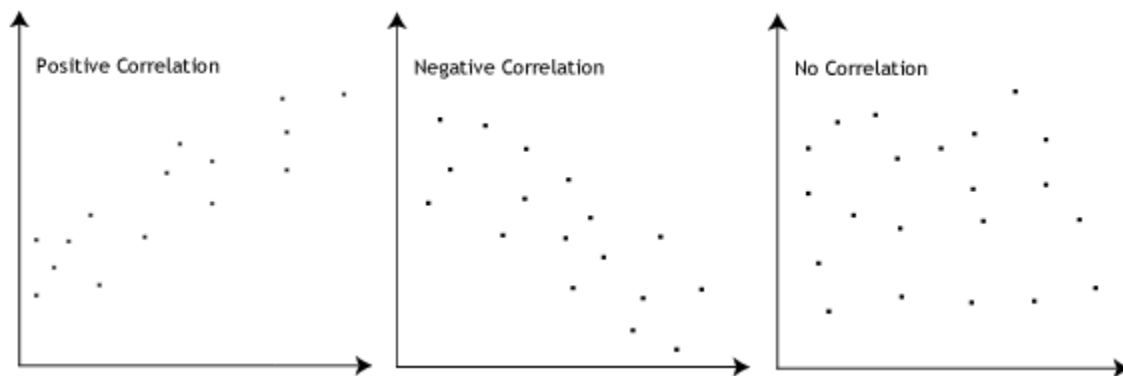
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

## 3. What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's  $r$ , the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between  $-1$  and  $1$ .

The Pearson's correlation coefficient varies between  $-1$  and  $+1$  where:

- $r = 1$  means the data is perfectly linear with a positive slope ( i.e., both variables tend to change in the same direction)
- $r = -1$  means the data is perfectly linear with a negative slope ( i.e., both variables tend to change in different directions)
- $r = 0$  means there is no linear association
- $r > 0 < 5$  means there is a weak association
- $r > 5 < 8$  means there is a moderate association
- $r > 8$  means there is a strong association



Formula

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

Here,

- =correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable
- =values of the y-variable in a sample
- =mean of the values of the y-varia

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

A) It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

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

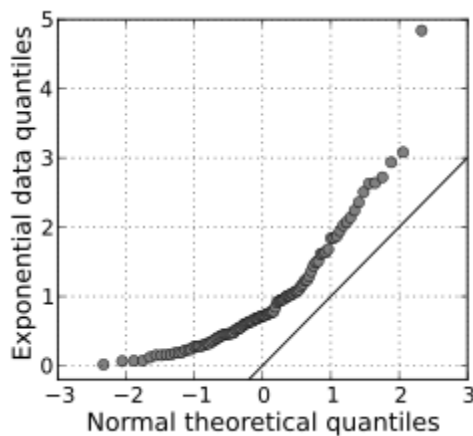
A) If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get  $R^2 = 1$ , which lead to  $1/(1 -$

R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. 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).

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

A) Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q-Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line  $y = x$ . If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line, but not necessarily on the line  $y = x$ . Q-Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

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