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? (3 marks)

Answer:-

a) Days:- Doesn't have any correlation with Target Variable
b) Seasons:- the Target variables has a very good relevence with Seasons
c) Weekdays:- even though there is gradual increase in Response variable from Monday to Friday and sight decrease from Since Friday to state constant over weekends, but the impact is not very severe.
d) Months:- Months follow the same trend as with Seasons and some of them(after creating dumy variable) could be colinear.

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

Answer:-

a) drop_first=True, reduces a Feature in predict variable dataset, and Reducing a Feature will decrease the penality factors like VIF and make the model more sigficant and make sure that they is no wise swing in Cooefficient.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:-

a) temp and atemp have highest correlation with target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:-

a) Validation steps of Linear Regressions are as follows:-
    1. Predict the Target variable using train data and Calculate Residuals(res_train) for Training Data
    2. Visualise the Residual(res_train) on a histogram and check the whether they are normally distirbuted over Mean value "0"
    3. Predict the Target variable using test data and Calculate Residuals(res_train) for test Data
    4. Visualise the Residual(res_test) on a histogram and check the whether they are normally distirbuted over Mean value "0"
    5. Check the R2 Score for actual Test response and predicted test response, the calculated R2 Score should be within 5% of R2 Score obtained by model on Train data.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:-

a. Year(yr), actual temperature(atemp) and Wind Speed(windspeed) are top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:-

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. 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.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:-

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 analyzing it and the effect of outliers on statistical properties.

Simple understanding:

Once Francis John "Frank" 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

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

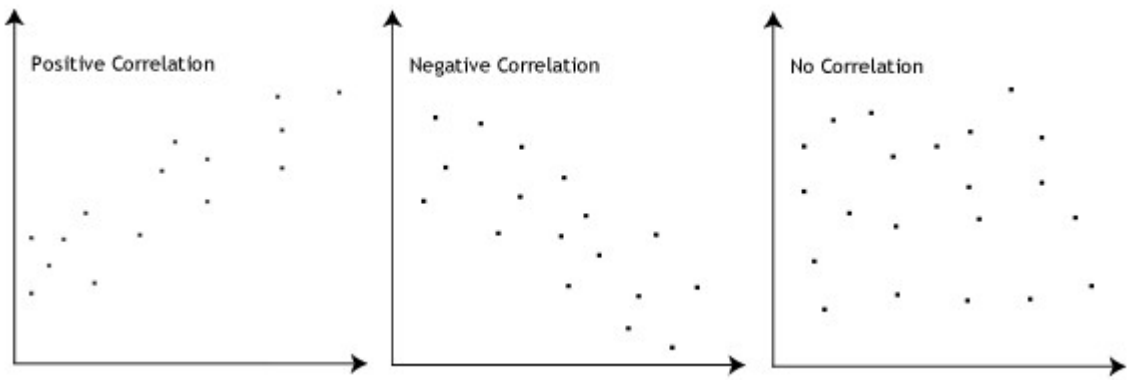
3. What is Pearson's R? (3 marks)

Answer:-

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



Pearson r Formula:-

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

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Answer:-

   - Scaling:- Scaling is a geometric change that linearly enlarges or reduces things. A property of objects or rules known as scale invariance is that they remain unchanged when scales of length, energy, or other variables are multiplied by a common factor. Scaling law, a law that explains how many natural phenomena exhibit scale invariance.
   - scaling performed because:- It is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables. Additionally, it aids in accelerating algorithmic calculations. The majority of the time, the obtained data set includes characteristics that vary greatly in magnitudes, units, and range.
   - The difference between normalized scaling and standardized scaling:-
     The values of a normalized dataset will always fall between 0 and 1. A standardized dataset will have a mean of 0 and a standard deviation of 1, but the maximum and minimum values are not constrained by any specified upper or lower bounds.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 Marks)

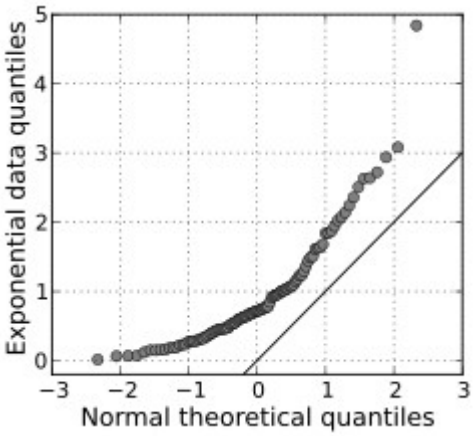
   Answer:-
   If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is 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. (3 marks)

   Answer:-
   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.