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?**

From my analysis of the categorical variable from the dataset, I inferred that the value of R square decreases as I removed columns from the model.

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

if we don't drop the first column then the dummy variables will be correlated. This may affect some models adversely and the effect is stronger when the cardinality is smaller.

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

Highest correlation is between “atemp” and “temp” I.e, 0.99

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

By plotting scatter plot with the y_test and y_predicted.

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

'yr','spring' and 'Light-snow+rain'

General Subjective Questions

1 - Explain the linear regression algorithm in detail.

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior 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.

Example for that can be let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behavior or patterns from the historical data.

Mathematically, we can write a linear regression equation as:

$$Y = a + bx$$

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.

Anscombe's Quartet. It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven (x,y) pairs

All the summary statistics you'd think to compute are close to identical:

- The average x value is 9 for each dataset
- The average y value is 7.50 for each dataset
- The variance for x is 11 and the variance for y is 4.12
- The correlation between x and y is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

3- What is Pearson's R?

Pearson's R also known as Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

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 the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

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.

The q-q or quantile-quantile is a scatter plot which helps us validate the assumption of normal distribution in a data set. Using this plot we can infer if the data comes from a normal distribution. If yes, the plot would show fairly straight line