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?

   Answer –
   - Demand of bikes is higher during summer and summer seasons.
   - Bike demand is increasing by every year. The demand in 2019 is higher than the demand in 2018.
   - Bike demand is also highly affected by the weather situation. The bookings are more when there is a clear weather and there are no bookings during rainstorm.

2. Why is it important to use **drop_first=True** during dummy variable creation?
   Answer –
   - To avoid duplicate features which implies the same value.
   - To avoid multicollinearity between features which will affect the model performance.
   - To reduce the number of features which will improve

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Answer –
   - By looking at the pairplot, the variable 'temp' and 'atemp' has the highest correlation (0.63) with the target variable 'cnt'

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Answer –
   - By plotting the distribution of residual errors and visualizing if the errors are normally distributed with the standard deviation 1.
   - By plotting y_test and y_pred and verify if there is a linear relationship.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Answer –
   - **Temperature** (temp) - 0.4318
   - **Weather Situation 3** (light_snowrain) – 0.2846
   - **Year** (yr) – 0.2413

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is one of the easiest and most popular machine learning algorithms. It is used for predicting continuous numerical variables using statistical methods. It shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.

Mathematically, linear regression can be represented as:

$$y= a_0+a_1x+ \varepsilon$$

where

y = Dependent or Target Variable
X = Independent variable
a0 = intercept of the line
a1 = Linear regression coefficient
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Types of Linear Regression
Linear regression can be further divided into two types of the algorithm:
Simple Linear Regression:
If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
Multiple Linear regression:
If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

2. Explain the Anscombe's quartet in detail.

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.

3. What is Pearson's R ?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

Scaling 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.
There are two type of scaling techniques –

- Normalization or MinMaxScaling
- Standardization or StandardScaling

Normalization or MinMaxScaling:

Normalization or MinMaxScaling brings the values into the range of 0 and 1 . The formula for min max scaling is

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$x = \frac{x - \text{mean}(x)}{sd(x)}$$

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

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.

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 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.