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)
   Ans: The categorical variables are season, month, holiday, workingday, weekday, weathersit, yr.
   - **Year vs Count : In the year 2019, bike sharing has become popular**
   - **Month vs Count: September month has been recorded with maximum number of shared bike rides**
   - **Weekday vs Count: Saturday of the week has maximum number of shared bike rides**
   - **Weathersit vs Count: Maximum number of shared bikes are booked during clear weather.**
   - **Maximum number of shared bikes are booked on Weekend or Holiday**

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   Ans: drop_first=True will drop the first columns of the generated dummies. It's good to have n-1 dummies column for n categories, but having n columns makes no harm rather than redundant data. For example,

Gender
Male
Female

⇒

Gender_Male	Gender_Female
0	1

N column for N category ⇒

Gender_Male
0

   If gender_male is 0, then it implicity indicates the customer is female. N-1 columns would be sufficient to prove it.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Ans: The model built on the training set is validated using the R2, Adjusted R2 values and by Residual analysis where the error terms should be normally distributed with mean equals to 0.

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

Ans: Linear regression is a machine learning algorithm that comes under supervised learning methods that uses past data with labels for building models. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (X) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.

Types of linear regression:

- **Simple Linear Regression:** If a single independent variable is used to predict the value of a numerical dependent variable(y), then it is called Simple Linear Regression.

  $$y=\beta0+\beta1X$$

- **Multiple Linear Regression:** If more than one independent variable(X) is used to predict the value of a dependent variable(y), then it is called Multiple Linear Regression.

  $$y=\beta0+\beta1X1+\beta2X2+....+\beta nXn$$

Finding best fit line: Our goal is to find the best fit line which means the error between the predicted values and actual values should be minimized. The best fit line will have least error.

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

Ans: Anscombe's quartet comprises four datasets that have almost 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? (3 marks)

Ans: The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation.

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}}$$

Where r is the correlation coefficient

x is the value of x variable in the sample

y is the value of y variable in the sample

n is the sample size

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

Ans: Scaling is the part of data preparation for building the model. Scaling is the process of normalizing the given data. Because the given data set might have different types of data which is measured using different scale. Say days of the week in numbers, speed of the wind in velocity, temperature in degress or Fahrenheit, humidity, etc. we need to normalize all these values to have all data on same scale for the model to easily interpret the data points.
There are 2 types of Scaling techniques. Normalization or Min-Max scaling and Standardized or Z-score scaling.
Min-Max Scaling: X_new = (X - X_min)/(X_max - X_min)
This transforms the values of the data points into the range of [0,1] or else in the range [-1, 1] if there are negative values in the dataset. It is useful when the there are no outliers as it cannot cope up with them.

Standardized Scaling: X_new = (X - mean)/Std
The standardized data set will have a mean of 0 and standard deviation of 1, but there is no specific upper or lower bound for the maximum and minimum values.
The advantage of Standardisation is that it doesn't compress the data between a particular range as in Min-Max scaling. It is useful, especially if there is are extreme data point (outlier).

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

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.

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.

The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.