**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?

   We can infer following from the below categorical variables

   - Year – Total count of bike rentals is more in 2019 than in year 2018
   - Month - There is an increasing trend in bike rentals from January and it has peaked in the months of June, July, August, September and from then we can see a decreasing trend till December. Highest rentals are observed in August and lowest is in January
   - Weekdays – Bike rentals are observed more in Thursday and less in Sunday, we can infer that holidays have an impact on bike rentals
   - Working days – Bike rentals are very high in working days which could be due to office commuters than non-working days
   - Weather – Bike rentals are more when weather is very clear, and its very low when it is rainy
   - Windspeed – Bike rentals are less when windspeed is very high. When windspeed is more than 20 we can infer a decrease in bike rentals
   - Temperature – we can infer that when temperatures are in extreme conditions the bike rentals has reduced both when its extreme cold and extreme hot
2. Why is it important to use drop_first=True during dummy variable creation?

   - Drop_first is important to remove extra columns there by reducing the collinearity in variables while framing the model
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   - Variables like Casual, registered are in high correlation with count so we are removing them, also temp and atemp seems to be highly correlated so we can drop atemp

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

   - R square values can be used to validate the linear regression models.We derived R square value of 83 for our LR model

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

   - We can infer that variables like Year, Temperature and Weather and Season has explained the demand of rental bikes well

**Assignment-based Subjective Questions**

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 a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) 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

$Y = a0+a1x+ \varepsilon$

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

Linear regression is again divided into Simple and Multiple linear regression.
When there is a 1-1 relationship between feature and target variable its simple linear regression, when there are multiple features influencing the target variable its called as multiple linear regression

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

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and 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 when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that numerical calculations are exact, but graphs are rough.

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R?

   In statistics, the Pearson correlation coefficient — also known as Pearson's R, the Pearson product-moment correlation coefficient, the bivariate correlation, or colloquially simply as the correlation coefficient — is a measure of linear correlation between two sets of data.

   Formula:

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

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, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.
   - Normalized / Min-Max scaling vs Standardized scaling: Normalized: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

   ```
   X_std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))
   X_scaled = X_std * (max - min) + min
   ```

   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). sklearn.preprocessing.scale helps to implement standardization in Python

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

   Ans: If there is perfect correlation, then VIF (Variance Inflation Factor) = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads 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

**Assignment-based Subjective Questions**

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