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)

   Season: The Usage in the season of fall is pretty high.
   Year: Across both the years 2019 has shown a higher demand.
   Months: The demand has been high during the time period of May-Oct and drops later.
   Holiday: The mean during a non holiday is much higher compared to that of a holiday.
   Weekdays: I do not notice much of a variation across the week with respect to mean for cnt.
   Working day: Non working days have a huge variance. Means are almost similar.
   Weather: The number of decreased as the likely hood of rain increased.

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

   Ideally we would try to create the dummy variables since the categorical columns which are not binary will be considered as numerical or else and thus we would convert it to 0's and 1's. But during this conversion ideally we would only need n-1 for n levels. Since all values with a '0' is typically not considered and thus we use drop_first = True.

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

   The temp has the 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)

   The R2 on the test data must be similar to that of the train data. Also, the residual analysis where in you plot the difference between the predicted and the actual value should be a normal distribution.

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

   Temperature, Windspeed and Month (Oct) significantly impact the output and was judged based on the co-efficient.

General Subjective Questions

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

   Linear Regression is a ML model which allows one to predict the outcome of a target variable based on a dependent variables. The goal is to find the best fit line which gives us predicted values that are closest to the output. Ideally this best fit line has an equation of a line like Y= mx+c. and the coefficients here which should be figured out for the best fit line would be that of m, c.

   The model is responsible for coming up with the best coefficients which would minimize the mean square error. We use cost function for this. The algorithm iterates several times with different coefficient values and based on the gradient descent it figures out the best values for the coefficients.

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

   This is a set of four datasets. Each of them include 11 x-y pairs of data. Ideally though the summary statistics are important relying alone on this might not be a good idea because outliers might influence the data which cannot be visualized in this. Though it has identical statistical properties such as R2, mean, variance, correlation it has different representations of them.
   D1: This dataset fits the linear regression pretty well.
   D2: This dataset shows non linear data which cannot fit the regression model.
   D3 & D4: These both show us outliers which cannot fit the regression model.

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

   It is used to measure the strength of the correlation with values in the range of -1 to +1. The closer it is to 1 or -1 the stronger the positive or negative correlation. So this can be used by us to figure out how closely two variables influence the output of each other based on the data that we would have.

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

   While the scaling typically does not effect the p values or the prediction it does have an impact on the coefficients. Ideally the comprehension of the coefficient would not be so effective. So thus since in real time the features vary widely in terms of units, magnitude and range scaling would be important to interpret the features on the same scale.
   The Normalized scaling rescales the values in a way that data has a mean of 0 and a standard deviation of 1(Unit variance) but where as Standardization (Min max scaler) rescales them into the range in between 0 and 1.

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

   Ideally VIF- variance correlation factor is a parameter through which I would be able to filter out features which have multicollinearity with rest of the variables in our system. The formula for this in terms of R would be I/(1-R^2). In terms of a perfect correlation this R2 in denominator would be 1 and thus the overall value would be infinite.

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

(3 marks)

Q-Q plot or quantile-quantile plot allows us to assess if a set of data plausibly came from some theoretical distribution such as exponential or normal. In the case of linear regression we would use Q-Q to check the distribution of residuals whether it is normal or not which is one of our assumption.