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)***

'yr' (2018 or 2019) has high positive contribution in explaining the variations in 'cnt' (total count of rental bikes).
'weathersit' (weather situation) has high negative contribution in explaining the variations in 'cnt'
Other categorical variables (months and days) did not have significant effect on the dependent variable.

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

While creating the dummy variables, one of the variables can be completely described by the remaining dummy variables, i.e., one of the dummy variables will have a high VIF with the other dummy variables.

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

'atemp' (feeling temperature)

4. ***How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)***

Low P-values for predictor coefficients.
Low VIF values among predictors.
Residual analysis: errors are roughly a normal distribution with mean centred at 0.
Similar r-squared values on train data and test data

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

'atemp' (feeling temperature) has high positive contribution
'windspeed' has high negative contribution
'hum' (humudity) has high negative contribution

General Subjective Questions

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

Linear Regression is a method of predicting a continuous target variable from one or more predictor variables. Using the Ordinary Least Squares (OLS) method, we try to find a linear relationship between the predictors and the target such that the residual sum of squares (RSS) is minimized.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$$

where,
y is the target variable,
X_1, X_2, ... are the predictor variables,
β_1, β_2, ... are the coefficients of the predictor variables,
β_0 is the y-intercept

RSS is the sum of squares of the difference between the predicted values and the actual values
TSS is the sum of squares of the difference between the predicted values and the mean value

R^2 and Adjusted-R^2, are used to determine how much variance in the target variable is explained by the predictors. R^2 only considers RSS and TSS where as Adjusted-R^2 considers the number of predictors while penalizing the usage of unnecessary predictors.

Once a model is derived, we check to see if the error terms are normally distributed with mean around 0. Then, a hypothesis testing is conducted to check the significance of the derived coefficients. Depending on the significance, a feature may be dropped or retained in the final model.

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

Anscombe's quartet is a set of 4 datasets having very similar statistic properties but are visually very different when plotted on a graph. This emphasizes the importance of analyzing plots during EDA to gain more insights about the data, apart from the statistical descriptions.

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

Pearson's R or Correlation Coefficient is a measure of linear correlation between two variables. It ranges from -1 to 1.
A value towards 1 signifies a positive correlation (one increases as the other increases)
A value towards -1 signifies a negative correlation (one decreases as the other increases)
A value of 0 signifies that there is no relation between the two variables.

Correlation Coefficient is not the slope of the line.

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

Scaling is a way of transforming the values of numerical features.
It is done so as to get a meaningful interpretation of the contributions of different features to the target variable.
Normalized scaling, a.k.a. minmax scaling fits all the values between 0 and 1
Standardized scaling, are scaled so that the mean is 0 and the standard deviation is 1

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

When a feature can be completely described by one or more features, then the R^2 value between those features becomes 1. Since

$$VIF = \frac{1}{1 - R^2}$$

VIF tends towards ∞ when R^2 becomes 1.

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

A Q-Q plot or a quantile-quantile plot is a plot of quantile-distributions of two variables. It can be used in Linear Regression to check if two variables are scaled or skewed versions of each other.