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: Categorical variable analysis

   - Season: Season 3 has highest demand for rental bikes
   - Year: Bike sharing demand grows next year.
   - Month: Bike demand grows until June, drops in July, august and again increases in September & oct, followed by a drop in last quarter. It could be due to bad weather conditions.
   - Holiday: In holiday period demand increases
   - Weekday & working day: It has almost no impact on bike sharing
   - Weather: The weathersit 1(Good) has highest demand followed by 2 (Moderate) & 3 (Bad)

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

   Answer:

   - In regression model, categorical variables are represented as dummy variables. If a categorical variable has n levels, it is represented as n-1 level i.e. one of the column values can be derived from others.
   - Dropping the first variable helps in reducing the extra column created during dummy variable creation, which further reduces the correlation created among dummy variables.

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

   Answer: "temp" has the highest correlation with the target variable cnt.

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

   Answer: After evaluating the final model following characteristics can be seen:

   - Residual analysis i.e. no relationship between the residual (error) terms.
   - Error terms have constant variance and normally distributed around mean.
   - No multi-collinearity between the independent variables and all variables have variance inflation factor i.e. (VIF) < 5
   - A linear relationship between independent and dependent variables can be seen by plotting the graph.

   Above characteristics points that linear regression assumptions hold true

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

   Answer: Top 3 features contributing towards the final model are:

   1. **Temperature (temp):** With coefficient value of 0.5499, it indicates that a unit increase in temp increases the bike sharing by 0.5499.

   2. **Year (yr):** With coefficient value of 0.2311, it indicates that a unit increase in year increases the bike sharing by 0.2311.

   3. **Weather situation 3:** With coefficient value of (- 0.2880), it indicates that a unit decrease in temperate will increase the bike demand by 0.2880

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer:

2. Explain the Anscombe's quartet in detail.

   Answer:

3. What is Pearson's R?

   Answer:

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

   Answer:

   In multiple linear regression model, a lot of independent variables are Analysed which may be in a different scale which may lead a model with very weird Coefficients that might be very difficult to interpret. Scaling is required for primarily two reasons

   1. Ease of interpretation
   2. Faster convergence for gradient descent methods

   Scaling only affects coefficients and interpretation not the model predictions and parameters like (t-statistics, f-statistics, p-values, r-squared).

   Scaling is performed in two ways:

   1. **Standardized scaling:** The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

2. **Min max scaling/Normalization:** Variables are scaled in such a way that it lies in range of zero and one.

$$X = x - \min(x) / \max(x) - \min(x)$$

Normalization is preferred over standardization because it loses some information in the data, especially about outliers.

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

Answer:

Variance inflation factor (VIF) is a measures the collinearity between predictor variables within a multiple regression. The higher the value, the grater is the correlation between variables. VIF = infinity shows a perfect correlation between two independent variables. In case of a perfect correlation (i.e. R2=1) leads to 1/(1-R2) infinity. To solve the 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.

Answer: