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:

Impact of Categorical Variables on the Dependent Variable ('cnt'):

- **Year (yr):**
    - There is a noticeable increase in 'cnt' from 2018 to 2019, reflecting higher and more variable bookings in 2019. This trend aligns with the growing popularity of bike-sharing systems as the demand increases.
- **Month (mnth):**
    - The number of bookings ('cnt') rises from January to a peak around July or August, which corresponds to the highest median and IQR during these months. A decline follows toward December, highlighting the seasonal variation in bike usage.
- **Weekday:**
    - The median counts are relatively consistent across the weekdays, with minor variations in the IQRs. This consistency indicates that bike bookings do not significantly differ from one weekday to another.
- **Season:**
    - **Spring:** Exhibits the lowest median count with a smaller interquartile range (IQR), indicating both lower and less variable bookings.
    - **Summer and Fall:** Both seasons show higher median counts and larger IQRs, suggesting increased and more variable bike bookings.
    - **Winter:** Although the median count in winter is lower than in summer and fall, it is still higher than in spring, with a slightly smaller IQR. These trends are likely influenced by the seasonal weather conditions.

- **Working Day:**
    - There is a slight increase in the median count on working days compared to non-working days. However, the IQR remains similar, suggesting only a minor difference in bookings between working and non-working days.
- **Weather Situation (weathersit):**
    - Bookings are highest in clear weather conditions, with the highest median and IQR, indicating greater and more variable usage. In contrast, 'Light Rain' weather results in significantly lower bookings, with the lowest median and IQR. Unsurprisingly, there are no bookings in 'Heavy Rain' weather.

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

Ans:
By dropping one category (using `drop_first=True`), you reduce redundancy and the number of parameters in the model. This not only improves model efficiency but also reduces the risk of overfitting, particularly when dealing with a large number of categories. Importantly, no information is lost because the absence of all dummy variables indicates the presence of the dropped category.

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

Ans: temp

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

Ans:

- **Linearity:** If the p-values for the predictors are small (typically less than 0.05), it indicates a statistically significant linear relationship between the predictors and the dependent variable. In the final model, all p-values are < 0.05.

- **Normality of Residuals:** To check this assumption, plot a histogram of the residuals. The residuals should approximately follow a normal distribution.

- **No Multicollinearity:** To assess multicollinearity, calculate the Variance Inflation Factor (VIF) for each predictor. A VIF value above 5 or 10 signals high multicollinearity, which should be addressed.

- **Homoscedasticity (Constant Variance of Error Terms):** Examine the Residuals vs. Fitted Plot by plotting the residuals against the predicted values. The plot should display a consistent spread of residuals across the range of predicted values, indicating constant variance.

- **Independence of Error Terms:** Use the Durbin-Watson Test to assess autocorrelation. A value close to 2 suggests that the error terms are independent and there is no autocorrelation.

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: To determine the key features that significantly impact the demand for shared bikes, we should focus on the absolute values of the regression coefficients. Features with higher absolute values are more influential in explaining the demand. Therefore, in our final model, these features are considered the most important, **temp - 0.3425, 2019(yr) – 0.2372, Light Rain(weathersit) – 0.2007**

General Subjective Questions

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

Steps in Linear Regression Analysis

Ans:

1. **Data Collection and Preparation**: Gather data for the dependent and independent variables. Handle missing values, outliers, and perform necessary transformations.
2. **Exploratory Data Analysis (EDA)**: Analyze the data to understand the relationships between variables. Use scatter plots, correlation matrices, and other visualizations to explore linearity and identify potential multicollinearity.
3. **Model Specification**: Choose the form of the regression model (e.g., simple vs. multiple regression). Select the independent variables to include in the model.
4. **Model Fitting (Estimating Coefficients)**: Use the OLS method to estimate the coefficients β0,β1,…,βn\beta_0, \beta_1, \dots, \beta_nβ0,β1,…,βn. This involves solving the minimization problem to find the line of best fit.
5. **Model Evaluation**: Evaluate the model's performance using metrics like R-squared, Adjusted R-squared, Mean Squared Error (MSE), and others. Check for violations of the regression assumptions (e.g., using plots of residuals, VIF for multicollinearity).
6. **Model Interpretation**: Interpret the estimated coefficients to understand the relationship between the independent and dependent variables. For instance, a positive coefficient suggests a positive relationship, while a negative coefficient indicates a negative relationship.
7. **Prediction**: Use the fitted model to make predictions on new data.
8. **Model Validation**: Validate the model using a separate test dataset or through techniques like cross-validation to ensure it generalizes well to unseen data.

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

Ans:

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics (such as mean, variance, and correlation) but exhibit very different distributions and relationships when graphed. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it, rather than relying solely on statistical measures. The quartet demonstrates that different datasets can have the same statistical properties, yet differ significantly in structure and behavior.

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

Ans:

Pearson's R, also known as the **Pearson correlation coefficient**, is a measure of the strength and direction of the linear relationship between two continuous variables. It is one of the most commonly used correlation coefficients in statistics.

The Pearson correlation coefficient is calculated using the following formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- x_i and y_i are the individual sample points.
- $\bar{x}$ and $\bar{y}$ are the means of the x and y values, respectively.

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

Ans:

Scaling is a data preprocessing technique used to adjust the range and distribution of features (variables) in a dataset. The goal of scaling is to ensure that the features contribute equally to the analysis and that algorithms that rely on distances or gradients work more effectively. Without scaling, features with larger ranges can dominate the analysis, leading to biased or incorrect results.

Scaling is performed for –

- Algorithm Efficiency
- Equal Contribution
- Improved Performance
- Normalization of Feature Weights

Difference Between Normalized Scaling and Standardized Scaling --

- **Normalization** is typically used when:
  - You need the data to be within a specific range.

- The model assumptions require features to be bounded within a range (e.g., neural networks).
- The dataset contains outliers that are not of interest or need to be kept within a specific range.

- **Standardization** is typically used when:

  - The data follows a normal distribution or needs to be centered around zero.
  - The algorithm assumes that the data is normally distributed or works best with data having a mean of 0 and a standard deviation of 1.
  - The dataset contains outliers, but they are not extreme enough to distort the scaling significantly.

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

   VIF Becomes Infinite because of -

   - **Perfect Multicollinearity**: Perfect multicollinearity occurs when Ri2=1R^2_i = 1Ri2 =1 for a predictor variable. This means that the predictor can be expressed as an exact linear combination of the other predictors in the model.
   - **Mathematical Consequence**: When $R_i^2 = 1$, the denominator in the VIF formula becomes zero:

     $$VIF_i = \frac{1}{1-1} = \frac{1}{0} = \infty$$

     Therefore, the VIF value becomes infinite.

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

   (3 marks)

   Ans:

   A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, typically the normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution to determine if the data conforms to that distribution.

   Importance of Q-Q Plots in Linear Regression :

1. Checking Normality Assumption - A Q-Q plot can help assess if normality assumption holds. If the residuals are not normally distributed, the estimates and inferences made from the model may not be reliable.
2. Identifying Outliers - Q-Q plots can help identify outliers or extreme values that do not fit the expected distribution. Outliers can significantly affect the model, including the estimates of coefficients and the overall fit.
3. Assessing Model Fit - By examining the residuals through a Q-Q plot, you can determine if there are any systematic deviations from normality that might suggest a poor fit.
4. Guiding Model Adjustments - If the Q-Q plot indicates non-normality, you might consider transforming the response variable, applying a different model, or adding additional predictors to better capture the data's structure.

In short, Q-Q plots are a valuable diagnostic tool in linear regression for checking the normality assumption of residuals, identifying outliers, and assessing overall model fit. They provide a visual means to ensure that the underlying assumptions of linear regression are reasonably met, thereby helping to ensure the validity and reliability of the model's inferences.