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)**
   Answer:

I have analyzed the categorical columns using box plots and bar plots. Here are some key insights from the visualizations:

- The fall season has attracted more bookings, with a significant increase in booking counts from 2018 to 2019 in each season.
- Most bookings were made during May, June, July, August, September, and October. The trend shows an increase in bookings from the beginning of the year until mid-year, followed by a decrease towards the end of the year.
- Clear weather conditions resulted in higher booking counts, which is expected.
- Thursdays, Fridays, Saturdays, and Sundays have more bookings compared to the start of the week.
- There are fewer bookings on holidays, which is reasonable as people may prefer to spend time at home with family.
- Bookings are almost equal on working days and non-working days.
- 2019 saw a significant increase in bookings compared to the previous year, indicating good progress in business.

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

Using drop_first=True during dummy variable creation is important for the following reasons:

1. **Avoiding Multicollinearity:** By dropping the first category, we prevent multicollinearity, which occurs when one dummy variable can be perfectly predicted from the others. This ensures that the dummy variables are not highly correlated with each other.
2. **Reducing Redundancy:** It reduces redundancy in the dataset. For a categorical variable with k categories, only k-1 dummy variables are needed. The dropped category serves as a baseline, and the presence of the remaining categories can be inferred.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   Answer:
   ‘temp’ variable 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)**
   Answer:
   I have validated the assumption of Linear Regression Model based on below 5 assumptions -

   - ✓ Normality of error terms
     - o Error terms should be normally distributed

- ✓ Multicollinearity check
    - There should be insignificant multicollinearity among variables.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   Answer:
   Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –
   - Holiday
   - Windspeed
   - month_aug
   - month_feb

General Subjective Questions

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

Linear regression is a statistical algorithm used to model the relationship between a dependent variable and one or more independent variables. The primary objective is to find the linear equation that best predicts the dependent variable. The equation of a simple linear regression model is y=β0+β1xy = \beta_0 + \beta_1 xy=β0+β1x, where yyy is the dependent variable, xxx is the independent variable, β0\beta_0β0 is the intercept, and β1\beta_1β1 is the slope. In multiple linear regression, the model extends to y=β0+β1x1+β2x2+...+βnxny = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_ny=β0+β1x1+β2x2+...+βnxn. The coefficients (β\betaβ) are estimated by minimizing the sum of the squared differences between observed and predicted values. The goodness of fit is evaluated using metrics like R-squared and mean squared error (MSE).

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

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression line, yet appear very different when graphed. The quartet demonstrates the importance of graphing data before analyzing it, as relying solely on summary statistics can be misleading. Each dataset in the quartet shows distinct patterns: a linear relationship, a non-linear pattern, a clustered pattern, and an outlier-influenced pattern. This emphasizes that data visualization is crucial for identifying underlying patterns and anomalies.

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

Pearson's R, or Pearson's correlation coefficient, is a measure of the linear correlation between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. It helps in understanding the strength and direction of the relationship between the variables.

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

Scaling is the process of adjusting the range of features in the dataset so that they can be compared on similar scales. Scaling is performed to improve the performance and convergence speed of many machine learning algorithms, particularly those that use distance measurements or gradient descent.

- **Normalized Scaling:** Adjusts the values to a range between 0 and 1, using the formula x−min⁡(x)max⁡(x)−min⁡(x)\frac{x - \min(x)}{\max(x) - \min(x)}max(x)−min(x)x−min(x). It is useful when the data does not follow a Gaussian distribution.
- **Standardized Scaling:** Adjusts the values to have a mean of 0 and a standard deviation of 1, using the formula x−μσ\frac{x - \mu}{\sigma}σx−μ. It is useful when the data follows a Gaussian distribution.

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

The value of Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity among the predictor variables, meaning that one predictor variable is a perfect linear combination of one or more other predictor variables. This results in division by zero in the VIF formula, as the R-squared value becomes 1, leading to an undefined or infinite value. Perfect multicollinearity makes it impossible to isolate the individual effect of each predictor on the dependent variable.

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 (quantile-quantile) plot is a graphical tool to compare the distribution of a dataset with a theoretical distribution, usually the normal distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points will lie approximately along a straight line. In linear regression, Q-Q plots are used to check the assumption of normality of the residuals. Normally distributed residuals indicate that the model's predictions are accurate and reliable. Deviations from the straight line suggest departures from normality, indicating potential issues with the model.