## Assignment-based Subjective Questions

**Question 1**. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

**Total Marks**: 3 marks (Do not edit)

**Answer:** <Your answer for Question 1 goes below this line> (Do not edit)

From the analysis of categorical variables such as season, mnth, weekday, and weathersit, we observed that these variables significantly impact the demand for shared bikes.

Specifically:

- Season: Demand for shared bikes was higher during seasons associated with favorable weather (season_3 and season_4).
- Month: Some months showed significant effects on bike rentals, indicating seasonality trends, with increased rentals during warm months.
- Weathersit: Bad weather (weathersit_3) had a negative effect on bike rental demand, as expected, since unfavorable conditions deter usage.

These findings suggest that people are more likely to use bike-sharing services in pleasant weather and during specific months of the year, which should be factored into any strategic decisions regarding bike availability and marketing campaigns.

---

**Question 2.** Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

**Answer:** <Your answer for Question 2 goes below this line> (Do not edit)

Using drop_first=True during dummy variable creation helps to avoid multicollinearity by removing one category for each categorical variable. This avoids the dummy variable trap, where perfect multicollinearity could occur due to redundant variables. Dropping the first dummy variable ensures that the remaining variables are independent and can uniquely describe the relationship with the target variable.

---

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

**Total Marks:** 1 mark (Do not edit)

**Answer:** <Your answer for Question 3 goes below this line> (Do not edit)

Based on the pair-plot analysis, the variable temp (temperature) had the highest correlation with the target variable (cnt). This indicates that bike demand increases with higher temperatures, which is consistent with the assumption that more people opt for bike-sharing services when the weather is warmer.

---

**Question 4.** How did you validate the assumptions of Linear Regression after building the model

on the training set? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** <Your answer for Question 4 goes below this line> (Do not edit)

The assumptions of Linear Regression were validated using the following methods:

1. Linearity: We assessed the residuals vs. fitted values plot. The residuals were randomly scattered, indicating a linear relationship between features and the target variable.
2. Normality of Residuals: A histogram of residuals showed a roughly bell-shaped distribution, suggesting that residuals followed a normal distribution.
3. Homoscedasticity: The residuals vs. fitted values plot did not show a funnel shape, indicating that the variance of residuals was consistent (homoscedasticity).

---

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
**Total Marks:** 2 marks (Do not edit)
**Answer:** <Your answer for Question 5 goes below this line> (Do not edit)

The top three features contributing significantly towards explaining bike demand are:

1. yr: The yr variable had a strong positive coefficient, indicating a significant increase in demand in the second year of data collection.
2. temp: Temperature had a positive effect on bike rental counts, with higher temperatures leading to increased demand.
3. season_4: The winter season (season_4) had a significant positive impact on demand, likely reflecting favorable conditions compared to other seasons.

---

## General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)
**Total Marks:** 4 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is a statistical algorithm used to model the relationship between a dependent variable and one or more independent variables. It aims to find a linear relationship between the features and the target. The linear regression equation is given as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Where:

y is the dependent variable (target).
x_i are the independent variables (features).

β_0 is the intercept.
β_i are the coefficients for the independent variables.
ε is the error term.
The goal is to find the values of β (coefficients) that minimize the sum of squared residuals between the predicted and actual values. Linear regression uses techniques like Ordinary Least Squares (OLS) to estimate the coefficients.

Linear regression assumptions include linearity, normality of residuals, homoscedasticity, and no multicollinearity among features.

---

**Question 7.** Explain the Anscombe's quartet in detail. (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet consists of four datasets that have nearly identical statistical properties (mean, variance, correlation, etc.) but very different distributions and visual appearances when plotted. The purpose of Anscombe's quartet is to illustrate the importance of visualizing data rather than relying solely on summary statistics.

Despite having the same descriptive statistics, the data in Anscombe's quartet exhibit different trends:

1. A linear relationship.
2. A non-linear relationship.
3. A relationship affected by an outlier.
4. A vertical pattern with high variance.

The key takeaway is that visualization is crucial to correctly interpreting data and avoiding misleading conclusions.

---

**Question 8.** What is Pearson's R? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R, also known as the Pearson correlation coefficient, measures the linear correlation between two variables. It ranges from -1 to +1:

+1 indicates a perfect positive linear relationship.
-1 indicates a perfect negative linear relationship.
0 indicates no linear correlation.
Pearson's R helps to understand the strength and direction of a linear relationship between two continuous variables.

---

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is the process of transforming features so that they lie within a similar range. It is performed to:

1. Improve model convergence for algorithms sensitive to the scale of data (e.g., gradient descent).
2. Ensure that all features contribute equally to the model.

The two common types of scaling are:

Normalization: Also known as Min-Max scaling, it rescales features to a range of 0 to 1 using:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization: Transforms features to have zero mean and unit variance using:

$$X_{std} = \frac{X - \mu}{\sigma}$$

Normalization is useful for data that doesn't follow a normal distribution, while standardization is suitable when data is normally distributed.

---

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity between independent variables. Perfect multicollinearity occurs when one feature is a perfect linear combination of other features, meaning that the information provided by one variable is redundant. To resolve this, one of the correlated features should be removed from the model.

---

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether the residuals of a linear regression model are normally distributed. It plots the quantiles of the residuals against the expected quantiles of a normal distribution.

1. Normality Check: If the residuals follow a normal distribution, the points on the Q-Q plot will fall roughly along a straight line.
2. Importance: Checking the normality of residuals is essential because linear regression assumes that residuals are normally distributed. Deviations from normality can indicate that the model may not be appropriate, or the data might need transformation.

---