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)

1.'cnt' is increased in year 2019 compared to 2018
2.'cnt' is max in season3 compared to all seasons
3.'cnt' is max in weathersit1(clear weather) compared to all weathersit2,3

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)

1. This redundancy can lead to issues in regression models, where multicollinearity increases standard errors and can make the model's coefficients unstable.
2.With drop_first=True, the model can interpret coefficients more reliably, and each dummy represents the effect of that category relative to the dropped (baseline) category.
3. This leads to more stable coefficient estimates and improved interpretability, as each dummy coefficient will show the effect of being in that category compared to the baseline.

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)

'Registered' having more correlation with target variable 'cnt'

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)

1. The relationship between each predictor and the target variable should be linear.
2. The residuals (errors) should be independent of each other, meaning there's no autocorrelation.
3. The residuals should have constant variance (homoscedasticity) across all levels of the independent variables.
4. Residuals should follow a normal distribution, particularly important if you need to make inference statements (like confidence intervals or hypothesis tests).
5. Predictors should not be highly correlated with each other, as this can lead to unreliable coefficient estimates.
6. Outliers and points with high leverage can disproportionately affect the model.

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)

---

1. Casual-
   Coefficient: 0.7158
   p-value: < 0.001 (highly significant)
   Interpretation: This is the largest positive coefficient, indicating that the number of casual users has a substantial, direct impact on total bike rentals. For each unit increase in casual users, total rentals (cnt) are expected to increase by around 0.7158 units.
2. Weekday_6
   - **Coefficient**: -0.1476
   - **p-value**: < 0.001 (highly significant)
   - **Interpretation**: This negative coefficient suggests that bike rentals decrease on this particular day (day 6, likely Saturday or Sunday). It may reflect lower demand due to weekend patterns or lower commuting needs.
3. season_3
   - **Coefficient**: 0.0969
   - **p-value**: < 0.001 (highly significant)
   - **Interpretation**: Being in season 3 (likely fall) is associated with a noticeable increase in demand, possibly due to more favorable biking conditions. This coefficient shows a strong seasonal effect on bike rental demand.

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>

---

1. **Objective of Linear Regression**
   Linear regression aims to model the relationship between one or more independent variables (predictors) and a continuous dependent variable (target) by fitting a linear equation to the data. The goal is to find the line (or hyperplane in multi-dimensional data) that best represents this relationship, minimizing the difference between predicted and actual values.
2. For simple linear regression with one predictor, the model can be written as:

   $Y =\beta o+\beta_1 x +e$

   where y is the target variable, x is the predictor, β_0 is the intercept, β_1 is the slope, and ϵ\epsilonϵ is the error term. In multiple linear regression, with multiple predictors, the equation generalizes to:

$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots \beta_n x_n + e$

3. **Optimization using Ordinary Least Squares (OLS)**
   Linear regression uses the **Ordinary Least Squares (OLS)** method to estimate the coefficients (β\betaβ values). OLS minimizes the sum of the squared differences between the observed values and the predicted values. where yi is the actual value, and y^i is the predicted value. By minimizing this cost function, the algorithm finds the best-fitting line or plane.
4. Linear regression relies on several key assumptions:

   Linearity: The relationship between predictors and the target variable is linear.
   Independence: Observations are independent of each other.
   Homoscedasticity: Constant variance of errors across all levels of predictors.
   Normality: Residuals (errors) are normally distributed.
   Checking these assumptions is crucial for reliable predictions and statistical inferences from the model.

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 is a group of four datasets that demonstrate the importance of visualizing data before analysis, as they all have nearly identical summary statistics but differ greatly in their distributions and relationships.
When plotted, the four datasets reveal different relationships:

- Dataset 1 shows a simple linear relationship.
- Dataset 2 shows a non-linear relationship (curved pattern).
- Dataset 3 contains a strong linear relationship, but an outlier heavily influences it.
- Dataset 4 has a single outlier that defines the linear pattern, with otherwise constant y-values.
  These differences demonstrate that data visualizations can uncover nuances, such as outliers or non-linearities, that numerical summaries alone can't reveal, underscoring the value of graphical analysis in data science.

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, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It's widely used to assess how changes in one variable are associated with changes in another.

Pearson’s R ranges from -1 to 1:

- **R = 1** indicates a perfect positive linear relationship (as one variable increases, the other does too).
- **R = -1** indicates a perfect negative linear relationship (as one variable increases, the other decreases).
- **R = 0** indicates no linear relationship between the 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 in Data Preprocessing
Scaling is a data preprocessing technique used to adjust the range of features so they can be compared on a similar scale. This is crucial in machine learning algorithms that are sensitive to the magnitude of features, such as k-nearest neighbors, support vector machines, and gradient descent-based models, where large differences in scale can bias the results.
Purpose of Scaling
Scaling ensures that each feature contributes equally to the model’s learning process, which helps the model converge faster and improves accuracy. Without scaling, features with larger numerical ranges might dominate and distort the learning process.

Normalization (Min-Max Scaling): This scaling technique rescales features to a specific range, typically [0, 1] or [-1, 1].
Standardization (Z-score Scaling): Standardization rescales features to have a mean of 0 and a standard deviation of 1

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>

An infinite value of **Variance Inflation Factor (VIF)** typically indicates a severe issue with **multicollinearity** among the predictor variables. This occurs when one predictor variable is a perfect or near-perfect linear combination of other variables in the model, leading to the following key issues:

1. **Perfect Multicollinearity**: If a predictor variable is an exact linear combination of other predictors, the regression model cannot estimate the variable’s unique contribution to the dependent variable. As a result, the VIF calculation—which relies on the R-squared value of the variable regressed against others—leads to an undefined (or "infinite") result since the R-squared is 1.

2. **Mathematical Breakdown in Calculation**:
   VIF is calculated using the formula:
   $VIF=1/(1-R^2)$
   where R^2 is the coefficient of determination for the predictor variable as a function of all other predictors. When R^2 approaches 1 (indicating near-perfect multicollinearity), the denominator approaches zero, causing the VIF to spike to infinity.

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 (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 the theoretical distribution.

Uses of a Q-Q Plot:

1. **Normality Check**: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps visualize this assumption by plotting the residuals against a normal distribution.
2. **Detection of Outliers**: Q-Q plots can help identify outliers in the dataset. Points that deviate significantly from the reference line may indicate outliers or influential observations that could affect the model's performance.
3. **Model Diagnostics**: By assessing the distribution of residuals, Q-Q plots help validate the linear regression model. If the residuals deviate from normality, it may suggest that the model is misspecified or that transformations of the dependent variable may be necessary.

Importance in Linear Regression:

- **Assumption Validation**: The validity of many statistical tests and methods in linear regression relies on the normality of residuals. A Q-Q plot provides a visual way to check this assumption.
- **Model Improvement**: Identifying non-normality can prompt model refinements, such as adding polynomial terms, applying transformations, or exploring alternative modeling approaches.
- **Interpretation and Reporting**: Q-Q plots offer a clear and intuitive visualization that can be easily communicated in reports and presentations to demonstrate the fit of the model and the adherence to underlying assumptions.