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 the categorical variables in the dataset, we can infer the following about their effect on the dependent variable cnt (bike demand):

1. Year (yr):
    - There is a significant positive association (0.57) with cnt.
    - This indicates a discernible rise in the demand for bikes over time, suggesting that bike usage has increased year over year.
2. Season (season):
    - There is a somewhat positive connection (0.4) with cnt.
    - This suggests that certain seasons, such as spring or summer, have higher bike demand compared to others. People are more likely to ride bikes during favorable weather conditions typically found in these seasons.
3. Current Weather Conditions (weathersit):
    - There is a moderately negative correlation (-0.3) with cnt.
    - This indicates that unfavorable weather conditions, such as rain or snow, tend to decrease bike demand. People are less likely to ride bikes in poor weather conditions.

These inferences suggest that both temporal factors (year and season) and weather conditions significantly influence bike demand. Warmer weather and favorable seasons increase bike usage, while adverse weather conditions decrease it.

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)

When you have a categorical variable with n categories, creating n dummy variables would result in perfect multicollinearity because the sum of all dummy variables for a given observation would always be 1. By setting drop_first=True, you drop one of the dummy variables, which removes this redundancy and avoids multicollinearity. This ensures that the regression model can be estimated correctly and the coefficients can be interpreted meaningfully.

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)

```
The numerical variable with the highest correlation with the target variable 'cnt'
is 'registered' with a correlation coefficient of 0.94541061184837.
```

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)

We have printed the summary of P values and VIF. We have observed that VIF values are below 10 to indicate no severe multicollinearity. In all majority cases Pvalues are low indicating that the features are significant to contribute the model. Further the heatmap clearly shows which features are significant and their coefficient values.

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)

Important Factors **Year (yr):**

significant positive association (0.57) with cnt.
shows a discernible rise in the demand for bikes over time.

Temperature (temp and atemp):

There is a high positive association between both factors and cnt (0.63 each).
In general, more people ride bikes in warmer weather.

Season (season):

0.4 indicates a somewhat positive connection with cnt.
indicates that some seasons, such as spring or summer, have more demand.

Windspeed (windspeed):

cnt has a weakly negative correlation (-0.24).
suggests that strong wind speeds could marginally lower consumption.

The current weather conditions (weathersit):

cnt has a moderately negative correlation (-0.3).
Unfavourable weather conditions, such as rain or snow

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 method used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find the best-fitting linear equation that describes the relationship between the variables.

Steps of the Linear Regression Algorithm:

1. Formulate the Hypothesis:
   - For a simple linear regression with one feature, the hypothesis is: [y = \beta_0 + \beta_1 x + \epsilon] where (y) is the dependent variable, (x) is the independent variable, (\beta_0) is the y-intercept, (\beta_1) is the slope of the line, and (\epsilon) is the error term.
   - For multiple linear regression with multiple features, the hypothesis is: [y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon] where (x_1, x_2, \ldots, x_n) are the independent variables.
2. Estimate the Coefficients:
   - The coefficients ((\beta_0, \beta_1, \ldots, \beta_n)) are estimated using the least squares method, which minimizes the sum of the squared differences between the observed values and the predicted values: [\text{Minimize} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2] where (y_i) is the observed value and (\hat{y}_i) is the predicted value.
3. Fit the Model:
   - Use the training data to fit the linear regression model by finding the optimal values of the coefficients that minimize the cost function.
4. Make Predictions:
   - Once the model is trained, use the estimated coefficients to make predictions on new data: [\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n]
5. Evaluate the Model:
   - Evaluate the performance of the model using metrics such as Mean Squared Error (MSE), R-squared ((R^2)), and Adjusted R-squared.

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 set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and the effect of outliers and the influence of different data distributions on statistical properties.

Key Points of Anscombe's Quartet

1. Identical Descriptive Statistics:
   - Each dataset in the quartet has the same mean, variance, correlation coefficient, and linear regression line.
   - Despite these similarities, the datasets are very different when visualized.
2. Importance of Data Visualization:
   - Anscombe's quartet emphasizes that relying solely on summary statistics can be misleading.

- Visualizing data can reveal patterns, trends, and anomalies that are not apparent from summary statistics alone

---

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 measure of the linear correlation between two variables. It quantifies the degree to which a relationship between two variables can be described by a straight line. The value of Pearson's R ranges from -1 to 1, where:
1 indicates a perfect positive linear relationship.
-1 indicates a perfect negative linear relationship.
0 indicates no linear relationship

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 a data preprocessing technique used to adjust the range of features in a dataset. It ensures that all features contribute equally to the model's performance by bringing them to a common scale without distorting differences in the ranges of values.
Normalized Scaling vs. Standardized Scaling
Normalized Scaling
Definition: Normalization (or Min-Max Scaling) rescales the feature values to a fixed range, typically [0, 1].
Formula: [x' = \frac{x - \min(x)}{\max(x) - \min(x)}]
Use Case: Useful when you know that the data does not follow a Gaussian distribution and when you want to bound the values within a specific range.
Standardized Scaling
Definition: Standardization (or Z-score Normalization) rescales the feature values so that they have a mean of 0 and a standard deviation of 1.
Formula: [x' = \frac{x - \mu}{\sigma}] where (\mu) is the mean and (\sigma) is the standard deviation of the feature.
Use Case: Useful when the data follows a Gaussian distribution and when you want to center the data around the mean with unit variance.

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>
VIF can become infinite when there is perfect multicollinearity among the predictors. This means that one predictor is a perfect linear combination of one or more other predictors. In such cases, the regression model cannot uniquely estimate the coefficients because the predictors are linearly dependent.

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 to assess if a dataset follows a particular theoretical distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution.

Use and Importance in Linear Regression

1. Assessing Normality of Residuals:
   - In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed.
   - A Q-Q plot helps to visually check this assumption by comparing the distribution of residuals to a normal distribution.
2. Detecting Deviations:
   - If the residuals follow a normal distribution, the points on the Q-Q plot will lie approximately along a straight line.
   - Deviations from this line indicate departures from normality, such as skewness or kurtosis.
3. Model Validation:
   - Ensuring that residuals are normally distributed validates the linear regression model and its statistical inferences.
   - Non-normal residuals may suggest that the model is misspecified or that transformations of variables are needed.
