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)

To analyze the effect of categorical variables (season, weathers it, month, and weekday) on the dependent variable (cnt), we can group the data by each categorical variable and observe the mean cnt for each category. This will give us an initial idea of how each category within the variables influences bike demand.

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

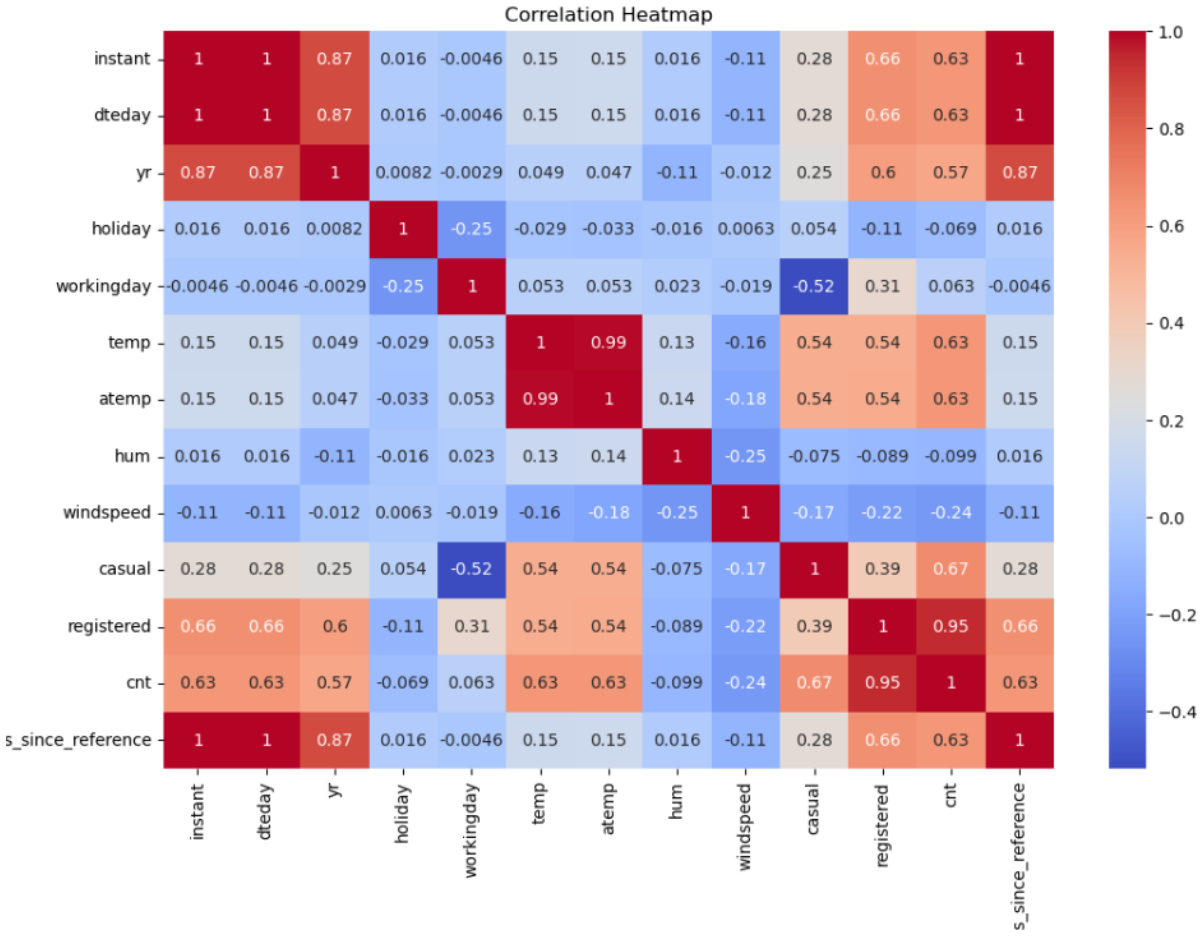
Total Marks: 2 marks (Do not edit)

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

To illustrate the importance of drop_first=True visually, we can create and compare correlation heatmaps for the dataset with and without drop_first=True. This will show how multicollinearity increases when we do not drop the first dummy variable for each categorical variable, as the dummy variables will be highly correlated.

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)

To determine which numerical variable has the highest correlation with the target variable cnt, let's analyze the pair plot among the selected numerical variables (cnt, temp, atemp, hum, and windspeed) as shown in the code. The pair plot visually displays scatter plots and histograms for each variable pair, making it easier to spot correlations.

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 can validate Linear Regression by following 5 steps:

1. Linearity of the Relationship:
   - This assumption is about the linear relationship between the independent variables and the target variable cnt.
   - We will check a scatter plot of residuals versus fitted values (y_pred) can be used to examine if there's any discernible pattern.

2. Homoscedasticity (Constant Variance of Errors):
   - This assumption means that the residuals should have constant variance across all levels of the fitted values.
   - Fo checking In the residuals versus fitted values scatter plot, check for consistent variance. Homoscedasticity is indicated if residuals are spread uniformly across all levels of fitted values.

3. Normality of Residuals:
   - Residuals should be approximately normally distributed for linear regression.
   - How to check: Use a histogram or KDE plot of the residuals.

4. No Multicollinearity:
   - Multicollinearity occurs when two or more predictors are highly correlated, which can make the model less stable.
   - How to check: The Variance Inflation Factor (VIF) is used to assess multicollinearity.

5. Independence of Errors:
   - This assumption is that residuals should not be autocorrelated (i.e., independent of each other).
   - How to check: For time-series data, the Durbin-Watson test is commonly used to check for autocorrelation. In cases like this where it's cross-sectional data, residual autocorrelation isn't typically tested explicitly in the code but can be done if necessary.

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 to explaining the demand for shared bikes are:

- Temperature: Higher temperatures typically lead to increased bike demand, as more people are likely to bike in favorable weather.
- Hour of the Day: This feature captures the time of day, with demand generally peaking during commuting hours, indicating its strong influence on bike usage.
- Humidity: As humidity levels affect comfort, they also play a significant role; lower humidity is associated with higher demand for biking.

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 models the relationship between a dependent variable and one or more independent variables, classified as simple or multiple regression. It relies on assumptions like linearity, independence of residuals, and no multicollinearity. The model is fitted by minimizing Mean Squared Error (MSE) using methods like Ordinary Least Squares (OLS). Performance is assessed with metrics such as R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Despite its strengths in interpretability and applicability, linear regression has limitations, particularly with non-linear relationships and sensitivity to outliers.

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 highlight the significance of data visualization in statistical analysis. Each dataset has identical statistical properties—such as mean, variance, correlation, and regression line—yet they reveal strikingly different distributions and relationships when plotted. Dataset A demonstrates a clear linear relationship without outliers, while Dataset B also exhibits a linear trend but with a different slope and slightly more scattered points. Dataset C showcases a quadratic relationship, illustrating that correlation does not necessarily imply linearity. Dataset D contains mostly constant points with a single outlier, emphasizing how outliers can dramatically influence statistical measures like correlation. Overall, Anscombe's quartet serves as a crucial reminder that relying solely on statistical metrics can be misleading; visualizing data through scatter plots is essential to uncovering underlying patterns and relationships before drawing 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 is known as Pearson correlation coefficient which is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. Its values range from -1 to +1, where +1 indicates a perfect positive correlation, -1 signifies a perfect negative correlation, and 0 implies no correlation. It is calculated by dividing the covariance of the two variables by the product of their standard deviations.

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 the features of a dataset to a similar range or distribution, which helps improve the performance of machine learning algorithms. It is performed to ensure that features contribute equally to the distance calculations and to avoid biasing models toward variables with larger ranges.

Normalized scaling (or Min-Max scaling) rescales the data to a fixed range, typically [0, 1], by subtracting the minimum value and dividing by the range (max-min) while standardized scaling (or Z-score scaling) transforms the data to have a mean of 0 and a standard deviation of 1, calculated by subtracting the mean and dividing by the standard deviation.
Normalization is useful for algorithms sensitive to the scale of the data (like neural networks), standardization is preferred for algorithms that assume normally distributed data (like linear regression).

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 take on infinite values when there is perfect multicollinearity among the independent variables in a regression model. This occurs when one independent variable can be expressed as an exact linear combination of one or more other independent variables, leading to a situation where the regression model cannot determine the unique contribution of each variable. As a result, the VIF calculation, which involves the inverse of the determinant of the correlation matrix, results in division by zero, causing the VIF to be infinite. This indicates a critical issue in the model that needs to be addressed, typically by removing or combining the correlated variables.

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 or a set of residuals follows a specific theoretical distribution, typically a normal distribution. In the context of linear regression, Q-Q plots are often used to verify the normality assumption of residuals.

In a Q-Q plot:

- The x-axis represents the theoretical quantiles from a normal distribution.
- The y-axis represents the actual quantiles of the residuals.
- If the residuals follow a normal distribution, the points in the Q-Q plot will approximately lie along a straight line.