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)

Season: spring season have low cnt ,

year:in 2019 high cnt,

mnth: in mid of the year high cnt(dependent variable)

weathersit: In clear weather high cnt.

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)

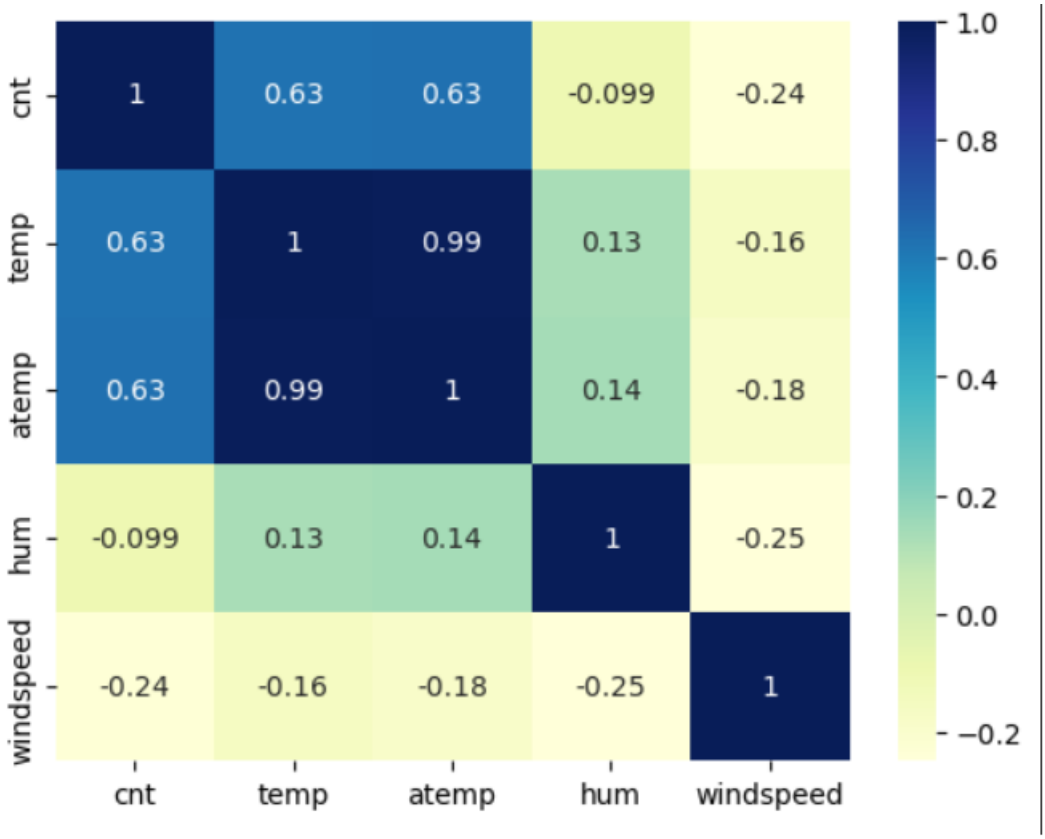
Using drop_first=True during dummy variable creation is important because it prevents multicollinearity by removing one dummy variable, thereby avoiding the "dummy variable trap." Additionally, it helps in reducing the extra column created during dummy variable creation.And we will find great model.

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)

temp-target variable(cnt) and atemp- target variable(cnt) has highest positive correlation which is 0.63.As attached the heatmap of numeric variable.



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)
Plot the residuals (errors) against the predicted values. Residuals have constant variance across all levels of the independent variables.
Calculate Variance Inflation Factors (VIF); values exceeding 5 may indicate problematic multicollinearity.
Plotting y_test and y_pred to understand the spread.

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)
Top 3 features contributing significantly below:
atemp,summer,Dec

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 that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The simplest form, simple linear regression, involves one independent variable and one dependent variable, aiming to find the best-fitting straight line that predicts the dependent variable based on the independent variable.
This method assumes a linear relationship between variables, meaning changes in the independent variable correspond to proportional changes in the dependent variable. The goal is to determine the line that minimizes the sum of squared differences between observed and predicted values, known as the least squares method. Linear regression is widely used for forecasting and determining the strength of predictors.

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 comprises four datasets that share nearly identical summary statistics—such as mean, variance, and correlation—but differ significantly when visualized. Each dataset consists of eleven (x, y) points.
Despite their statistical similarities, plotting these datasets reveals distinct patterns: one shows a linear relationship, another a non-linear curve, the third includes an outlier affecting the regression line, and the fourth has a vertical distribution with a single influential point.

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, or the Pearson correlation coefficient, quantifies the strength and direction of a linear relationship between two variables.
Its value ranges from -1 to 1:

+1 indicates a perfect positive linear relationship.
-1 indicates a perfect negative linear relationship.
0 indicates no linear relationship.
This coefficient helps determine how changes in one variable are associated with changes in another, aiding in understanding and predicting relationships between 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 a data preprocessing technique that adjusts the range of features in your dataset to ensure uniformity.
This process is crucial because many machine learning algorithms perform better when input features have similar
scales, preventing features with larger ranges from dominating the model's learning process.

Normalization (or Min-Max scaling) transforms data to fit within a specific range, typically [0, 1]. This is achieved by subtracting the minimum value of the feature and dividing by the range. Normalization is particularly useful when you want to bound your data within a specific range.

Standardization (or Z-score normalization) adjusts data to have a mean of zero and a standard deviation of one. This is done by subtracting the mean of the feature and dividing by the standard deviation. Standardization is beneficial when the data follows a Gaussian distribution and is essential for algorithms that assume a standard normal distribution.
In summary, scaling enhances model performance by ensuring that each feature contributes equally. Normalization confines data within a specific range, while standardization centers data around the mean with unit variance. The choice between these methods depends on the specific requirements of your analysis and the algorithms employed.

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>
A Variance Inflation Factor (VIF) quantifies how much the variance of a regression coefficient is inflated due to multicollinearity among independent variables. When VIF is infinite, it indicates perfect multicollinearity, meaning one independent variable is an exact linear combination of others. This perfect correlation leads to division by zero in VIF calculations, resulting in an infinite value. Such multicollinearity can destabilize regression models, making coefficient estimates unreliable. To address this, it's essential to identify and remove or combine the perfectly correlated variables to improve model stability.

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 that compares the quantiles of a dataset against the quantiles of a theoretical distribution, such as the normal distribution.
In linear regression, it's essential to check if the residuals (errors) are normally distributed, as this is a key assumption for valid inference.
By plotting the residuals on a Q-Q plot, you can visually assess normality: if the points align closely along a straight line, the residuals are approximately normally distributed. Deviations from this line suggest departures from normality, indicating potential issues with the regression model's assumptions.
