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)

Analyzing categorical variables in the bike-sharing dataset can provide valuable insights into their effects on bike rental counts. Here's a breakdown of some key categorical variables and their potential influence on the dependent variable (bike rental counts):

Key Categorical Variables

1. Season
   - Categories: Winter, Spring, Summer, Fall
   - Effect: Seasonal variations significantly impact bike rentals. Typically, summer months see the highest rental counts due to favorable weather conditions, while winter tends to have the lowest rentals due to cold temperatures and adverse weather.
2. Day of the Week
   - Categories: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
   - Effect: Weekday and weekend patterns often differ. Rentals may peak on weekends when people engage in leisure activities, while weekdays may show higher usage among commuters. Analysis may reveal that Fridays and Sundays are particularly high-traffic days.
3. Holiday
   - Categories: Yes or No
   - Effect: Holidays can lead to increased bike rentals as people take time off work and engage in recreational activities. An analysis might show spikes in rentals during public holidays compared to regular weekdays.
4. Working Day
   - Categories: Yes or No
   - Effect: This variable distinguishes between working days and non-working days (weekends/holidays). Rentals are generally higher on working days due to commuter usage, but this can vary based on the season and weather conditions.
5. Weather Conditions
   - Categories: Clear, Misty, Light Rain, Heavy Rain, Snow
   - Effect: Weather has a profound impact on bike rentals. Clear weather typically correlates with higher rental counts, while adverse weather conditions (like heavy rain or snow) can lead to significant drops in usage.

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)

By dropping one category (the first one, in this case), you eliminate redundancy. For example, if you have a categorical variable with three categories (A, B, C), including all three would mean that knowing A and B allows you to infer C. Dropping one category helps maintain the integrity of the model and ensures that it can be interpreted correctly without inflated variance in coefficient estimates

---

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 most correlated with the dependent variable cnt (the total number of bike rentals) is temperature.

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)

To validate the assumptions of linear regression after building the model on the bike-sharing dataset, several diagnostic techniques have been employed. Here’s a structured approach to checking these assumptions:

1. Linearity
   - Check: Use scatter plots to visualize the relationship between independent variables and the dependent variable (cnt).
   - Purpose: Ensure that there is a linear relationship; if not, transformations may be needed.
2. Independence of Errors
   - Check: The Durbin-Watson test can be utilized to detect autocorrelation in residuals.
   - Purpose: Values close to 2 indicate no autocorrelation, while values significantly less than or greater than 2 suggest potential issues.
3. Homoscedasticity
   - Check: Create residual plots by plotting residuals against fitted values.
   - Purpose: The spread of residuals should remain constant across all levels of the fitted values. If a pattern is observed (e.g., funnel shape), it indicates heteroscedasticity.
4. Normality of Residuals
   - Check: Use Q-Q plots or histograms of the residuals.
   - Purpose: Residuals should follow a normal distribution; deviations from the straight line in a Q-Q plot indicate non-normality.
5. Multicollinearity
   - Check: Calculate Variance Inflation Factor (VIF) for each independent variable.
   - Purpose: A VIF value greater than 5 or 10 suggests significant multicollinearity, which can distort coefficient estimates.
6. Model Fit Evaluation
   - Metrics: Assess model performance using R-squared, adjusted R-squared, RMSE (Root Mean Square Error), AIC (Akaike Information Criterion), and BIC (Bayesian Information Criterion).
   - Purpose: These metrics help evaluate how well the model explains the variability in cnt and compare it against other models.

By systematically checking these assumptions, you ensure that the linear regression model is valid and reliable for predicting bike rentals in the dataset.

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)

As per my analysis I got temperature, Year and Weather as most important

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>

- Overview of Linear Regression Algorithm
    - Linear regression is a fundamental statistical method used in machine learning and data analysis to model the relationship between a dependent variable and one or more independent variables. The primary goal of linear regression is to find the best-fitting straight line through the data points, which can be used for prediction and inference.
- Key Concepts
    - Dependent and Independent Variables:
        - The dependent variable (often denoted as yy) is the outcome or response variable that you aim to predict.
        - The independent variable (or predictor, denoted as xx) is the variable that influences the dependent variable.
    - Linear Relationship:
        - Linear regression assumes a linear relationship between the dependent and independent variables. This means that changes in the independent variable will produce proportional changes in the dependent variable.
    - Equation of the Regression Line:
        - The relationship is typically expressed using the equation:
            - y=b0+b1x1+b2x2+...+bnxny=b0+b1x1+b2x2+...+bnxn
            - where: b0b0 is the intercept, b1,b2,...,bnb1,b2,...,bn are the coefficients of the independent variables x1,x2,...,xnx1,x2,...,xn respectively

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 collection of four datasets created by statistician Francis Anscombe in 1973. Each dataset consists of eleven (x, y) points and has nearly identical simple descriptive statistics, yet they exhibit very different distributions and visual patterns when graphed. This quartet serves as a powerful illustration of the importance of data visualization in statistical analysis.

Purpose and Significance

The primary purpose of Anscombe's quartet is to demonstrate that relying solely on summary statistics can be misleading. Despite the datasets having similar means, variances, and correlation coefficients, their graphical representations reveal distinct patterns that could lead to different interpretations and conclusions. Anscombe emphasized that "numerical calculations are exact, but graphs are rough," highlighting the necessity of visualizing data before performing analysis or modeling.

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 (PCC), is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. Developed by Karl Pearson in the late 19th century, it is a foundational concept in statistics and data analysis.

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 refers to the process of adjusting the range of independent variables or features in a dataset. The goal is to standardize the features so that they contribute equally to the analysis and modeling processes. Without scaling, features with larger ranges can disproportionately influence the results of machine learning algorithms, especially those that rely on distance calculations, like k-nearest neighbors (KNN) or gradient descent optimization methods.

Why is Scaling Performed?

Scaling is performed for several reasons:

- Equal Contribution: It ensures that all features contribute equally to the distance calculations and model training, preventing features with larger scales from dominating the results
- Improved Convergence: Algorithms like gradient descent converge faster when the data is scaled, as it reduces the likelihood of oscillations and helps reach the optimal solution more quickly
- Enhanced Performance: Many machine learning algorithms perform better with scaled data, leading to improved accuracy and efficiency in model predictions
- Outlier Detection: Scaling can also aid in identifying outliers by adjusting the feature distributions

Key Differences Between Normalized Scaling and Standardized Scaling

Feature	Normalization	Standardization
Range	[0, 1] or [-1, 1]	Typically [-2, 2]

Feature	Normalization	Standardization
Mean	Not necessarily zero	Mean = 0
Standard Deviation	Not necessarily one	Standard deviation = 1
Use Case	When features have different ranges	When data follows a normal distribution
Effect on Outliers	Sensitive to outliers	Less sensitive; can help identify them

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>

Infinite values of the Variance Inflation Factor (VIF) typically occur due to perfect multicollinearity among the independent variables in a regression 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, or quantile-quantile plot, is a graphical tool used in statistics to compare the quantiles of two probability distributions. It provides a visual assessment of how closely a dataset follows a specific theoretical distribution, such as the normal distribution.