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)

Categorical variables such as season, weathersit, and workingday significantly impact bike rental demand. The seasonality factor shows that demand is higher in summer and fall compared to winter and spring. Weather conditions also play a crucial role—clear days tend to have higher rentals, whereas heavy rain or snow decreases demand. workingday affects rentals, with weekdays having more registered users, while casual users rent more on weekends.

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 helps to avoid the dummy variable trap, which occurs when there is multicollinearity in the dataset due to redundant categorical representations. By dropping the first category, we ensure that the model does not suffer from perfect multicollinearity, leading to more stable and interpretable regression coefficients.

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 variable temp (temperature) has the highest correlation with the target variable cnt (total bike rentals). This suggests that warmer temperatures are associated with increased bike usage, likely due to favorable riding conditions.

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)

The following validations were performed:

1. **Linearity**: Checked through scatter plots between independent variables and cnt.
2. **Normality of Residuals**: Verified using histogram and Q-Q plot of residuals.
3. **Homoscedasticity**: Examined residual vs. fitted values plot.
4. **Multicollinearity**: Checked using Variance Inflation Factor (VIF) to ensure no severe correlation among predictors.

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. **Temperature (temp)** – A significant predictor as demand increases with favorable temperatures.
2. **Season (season_fall)** – Fall season has a notable positive impact on bike rentals.
3. **Year (yr)** – The year 2019 (yr=1) shows an increasing trend in bike-sharing demand compared to 2018.

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 and one or more independent variables.

The equation is: y=β0+β1X1+β2X2+...+βnXn+εy = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_nX_n + \varepsilon where:

- yy is the dependent variable,
- X1,X2,...,XnX_1, X_2, ..., X_n are independent variables,
- β0\beta_0 is the intercept,
- β1,...,βn\beta_1, ..., \beta_n are coefficients,
- ε\varepsilon is the error term.

Linear regression minimizes the sum of squared residuals using Ordinary Least Squares (OLS) to estimate coefficients. It assumes linearity, independence, homoscedasticity, and normality of residuals for reliable predictions.

.

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)

Anscombe’s quartet consists of four datasets with nearly identical statistical properties (mean, variance, correlation), yet they exhibit different distributions when visualized. It highlights the importance of data visualization in statistical analysis, demonstrating that relying solely on summary statistics can be misleading

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)

Pearson’s R, or Pearson correlation coefficient, measures the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

- **+1** indicates a perfect positive correlation,
- **0** indicates no correlation,
- **-1** indicates a perfect negative correlation. It is calculated as: r=∑(Xi−X¯)(Yi−Y¯)∑(Xi−X¯)2⋅∑(Yi−Y¯)2r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \cdot \sqrt{\sum (Y_i - \bar{Y})^2}}

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)

Scaling transforms data to a specific range or distribution. It is performed to improve model convergence and performance by standardizing numerical features.

- **Normalization (Min-Max Scaling)**: Rescales values to a range of [0,1]. Formula: Xscaled=X−XminXmax−XminX_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}.
- **Standardization (Z-score Scaling)**: Centers data to have mean 0 and standard deviation 1. Formula: Xscaled=X−μσX_{scaled} = \frac{X - \mu}{\sigma}.

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)

VIF (Variance Inflation Factor) measures multicollinearity among independent variables. An infinite VIF occurs when there is perfect multicollinearity, meaning one predictor is a linear combination of others. This leads to singularity in the regression matrix, causing computation issues

---

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)

A Q-Q (Quantile-Quantile) plot compares the distribution of residuals to a normal distribution. It helps validate the assumption of normality in linear regression. If points lie along the 45-degree reference line, residuals follow a normal distribution, ensuring valid p-values and confidence intervals in hypothesis testing.
