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)

The categorical variables, such as season, weather conditions, and days of the week, had significant effects on bike demand. For example, winter and summer seasons showed a positive impact on demand, likely due to more favorable weather conditions for bikes. However, adverse weather conditions, such as light snow and mist, negatively affected the demand. Similarly, weekends (like Sunday) showed a slight decline in demand, possibly due to fewer commuters using bikes for work-related travel. The regression analysis confirmed that variables like Saturday (coef: 0.0861, $p < 0.05$) and Sunday (coef: 0.0410, $p < 0.05$) significantly influenced 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)
Using drop_first=True helps prevent the issue of multicollinearity in regression models. When creating dummy variables, each categorical variable is represented by multiple binary columns. If all categories are included, one can be perfectly predicted by the others, causing redundancy and numerical instability

---

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 temp (temperature) had the highest correlation with bike demand (cnt).

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, I used **VIF (Variance Inflation Factor) and p-values** as key indicators and followed below approach

1. **Checking for Multicollinearity:** I calculated VIF for all independent variables. If a variable had a **VIF greater than 5**, it indicated high multicollinearity, meaning that the variable was highly correlated with others.

2. **Assessing Statistical Significance:** I checked the **p-values** of each variable. If a feature had a **p-value greater than 0.05**, it meant that it might not have a significant impact on the target

variable (`cnt`).

3. **Final Adjustments:** After dropping highly collinear and statistically insignificant features, I re-ran the model to ensure stability and improved interpretability.

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 were:

1. **yr (Year):** Coefficient = 0.2359, strong positive impact on demand due to the increasing popularity of bike-sharing services.

2. **temp (Temperature):** Coefficient = 0.5690, major positive effect on demand, as favorable temperatures encourage biking.

3. **winter (Winter Season):** Coefficient = 0.1345, significant positive effect, indicating high demand during this season.

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 supervised learning algorithm used for predicting continuous values. It models the relationship between an independent variable (X) and a dependent variable (Y) using the equation: Y = b0 + b1X1 + b2X2 + ... + bnXn + e where:

- b0 is the intercept,

- b1, b2, ..., bn are coefficients,

- e is the error term. The algorithm minimizes the sum of squared residuals (differences between predicted and actual values) using the Ordinary Least Squares (OLS) method. It assumes linearity, normality of residuals, no multicollinearity, and homoscedasticity. The model in this project achieved an Adjusted R-squared of 0.812, indicating strong predictive power.

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 consists of four datasets that have identical statistical properties (mean, variance, correlation, and regression line) but have very different distributions when plotted.

It demonstrates the importance of visualizing data before analysis. The four datasets highlight how outliers, non-linearity, and clustering can distort conclusions drawn purely from summary statistics.

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 called Pearson correlation coefficient, measures the linear relationship between two variables. It ranges from -1 to 1:

- **1:** Perfect positive correlation
- **0:** No correlation
- **-1:** Perfect negative correlation It is calculated as the ratio of the covariance of the two variables to 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 adjusting the range of numerical variables to improve model performance.

- **Normalization (Min-Max Scaling):** Scales values between 0 and 1 using the formula: (X - min(X)) / (max(X) - min(X)) It is useful when features have different ranges.
- **Standardization (Z-score Scaling):** Centers data around zero with unit variance: (X - mean(X)) / std(X) It is useful for algorithms that assume normal distributions.

I have used Min-Max scaling

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 (Variance Inflation Factor) becomes infinite when there is perfect multicollinearity, i.e. one predictor variable is an exact linear combination of others. This can happen due to duplicate columns, inclusion of dummy variables without dropping one category, or highly correlated predictors. In such cases, the regression model cannot compute reliable estimates, leading to infinite VIF values, as seen with holiday, workingday, Sunday and Saturday in my analysis.

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 used to check whether residuals follow a normal distribution. It compares the quantiles of residuals against a theoretical normal distribution. If residuals align closely with the diagonal line, they are normally distributed. A Q-Q plot is essential in linear regression to validate the assumption of normality, which affects inference reliability and model performance. Our model's Q-Q plot suggested some deviation from normality, as confirmed by the Omnibus and Jarque-Bera tests.
