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)

From the analysis of categorical variables, it can be inferred that they have a significant impact on the dependent variable (bike demand). Variables such as season**,** year, and weather situation show clear variation in demand levels. For example, bike demand is generally higher during favorable seasons and weather conditions and increases noticeably in later years, indicating growth in adoption. Categories like holiday and working day also influence demand, with working days typically showing higher 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)

Using drop_first=True helps avoid the dummy variable trap, which causes perfect multicollinearity when all dummy variables are included. By dropping one category, it becomes the reference category, ensuring that the regression model remains stable and the coefficients are interpretable.

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)

From the pair-plot**,** temperature (temp) shows the highest correlation with the target variable -bike demand.

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 assumptions of Linear Regression were validated using the following methods:

- Linearity & Homoscedasticity**:** Checked using residuals vs. fitted values plot.
- Normality of residuals**:** Verified using a Q-Q plot.
- Multicollinearity**:** Assessed using Variance Inflation Factor (VIF).
- Independence of errors**:** Checked using the Durbin–Watson statistic.

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 3 features contributing significantly to explaining shared bike demand are:

1. Temperature (temp)
2. Year
3. 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)
Linear Regression is a supervised learning algorithm used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship and fits a best-fit line by minimizing the sum of squared errors between actual and predicted values. where coefficients are estimated using Ordinary Least Squares (OLS). Linear regression is widely used due to its simplicity, interpretability, and effectiveness when assumptions are met.

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 that have nearly identical statistical properties such as mean, variance, correlation, and regression line but differ significantly when visualized. It demonstrates the importance of data visualization in statistical analysis, as relying solely on summary statistics can be misleading and hide underlying patterns or outliers.

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 is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. Its value ranges from –1 to +1, where +1 indicates perfect positive correlation, –1 indicates perfect negative correlation, and 0 indicates no linear relationship.

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 is the process of bringing numerical features to a similar range to improve model

performance and convergence.

- Normalization**:** Rescales data to a range of 0 to 1.
- Standardization**:** Centers data around mean 0 with standard deviation 1.
  Scaling is important for algorithms sensitive to feature magnitude, such as linear regression and gradient-based models.

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)
An infinite VIF occurs when there is perfect multicollinearity, meaning one independent variable can be exactly predicted from others. This makes the regression coefficients unstable and unreliable, and it typically happens when duplicate or highly correlated variables are included in the 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)
A Q-Q (Quantile-Quantile) plot compares the distribution of residuals with a theoretical normal distribution. In linear regression, it is used to verify the **normality assumption of residuals**. If points lie close to the reference line, the assumption holds; deviations indicate skewness or outliers, which can affect model validity.
