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)

Below are the Inferences

- **Year:** A positive coefficient of 992.37 suggests that bike rentals increased significantly in the newer year (2019 when compared to 2018)
- **Working Day:** A positive coefficient of 173.47 indicates that bike rentals are slightly higher on working days compared to non-working days
- **Temperature:** The large positive coefficient of 1233.74 suggests that warmer temperatures strongly drive bike rentals
- **Humidity:** A negative coefficient of -439.25 implies that higher humidity discourages bike rentals
- **Windspeed:** A negative coefficient of -352.72 indicates that higher wind speeds reduce bike rentals

Categorical variables are encoded as season_summer, month_July, weekday_Sunday), and their coefficients reflect the difference in bike rentals

Season:

- season_summer: A positive coefficient of 303.49 suggests that summer sees higher bike rentals compared to spring
- season_winter: A larger positive coefficient of 508.77 indicates even higher bike rentals in winter relative to other seasons

Month:

- month_July: A negative coefficient of -121.34 suggests that bike rentals decrease in July compared to other months
- month_September: A positive coefficient of 207.04 suggests an increase in bike rentals in September compared to other months

Weekday:

- weekday_Sunday: A positive coefficient of 159.18 indicates that bike rentals are higher on Sundays compared to other weekday

---

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)

It is important to use drop_first as True, when creating dummy variables. Each category is represented as a binary column (1 or 0). If you create dummy variables for all categories, the resulting dataset will include redundant information.

The sum of all dummy columns for each row will always be 1, making one column redundant. This redundancy leads to perfect multicollinearity, which violates the assumptions of regression models and causes issues like: 1. Inflated standard errors. 2. Unreliable coefficient estimates

Spring	Summer	Fall	Winter
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

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)

Temperature

- The variable temperature has the highest correlation with the target variable based on the regression coefficient
- It aligns with expectations, as favourable weather conditions are likely to encourage bike rentals
- Temperature has the highest positive coefficient: 1233.74
- This suggests that temperature has the strongest positive relationship with the bike rentals

Year: A significant positive coefficient: 992.37 shows a strong positive trend in bike rentals over time, but less than temperature

Humidity: A negative coefficient: -439.25 indicates an inverse relationship; higher humidity decreases bike rentals

Windspeed: Another negative coefficient: -352.72 suggests higher wind speeds discourage bike rentals

---

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)

Linearity using Residual plot

- Plot the residuals (errors) against the predicted values

- Look for randomness in the residual distribution. If patterns or trends exist, the linearity assumption may be violated

Homoscedasticity

- Residuals should have constant variance across all levels of predicted values
- Use the residual plot again to verify this. Increasing spread indicates heteroscedasticity

Normality of Residuals

- Residuals should follow a normal distribution
- Use a histogram and Q-Q plot for visual inspection

Multicollinearity

- Check the Variance Inflation Factor (VIF) for all independent variables. VIF > 10 indicates high multicollinearity

Model Performance Check Train vs Test R^2

Train dataset: $R^2 = 0.820$; Adjusted $R^2 = 0.810$

Test dataset: $R^2 = 0.800$; Adjusted $R^2 = 0.790$

The similarity in R^2 values for training and testing indicates the model generalizes well without overfitting

---

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 are Temperature, Year and Season

Temperature: Coefficient: 1233.74. As the temperature increases, bike demand increases significantly. Favourable weather conditions encourage biking

Year: Coefficient: 992.37. Increase in bike-sharing adoption over the years

Season Winter: Coefficient: 508.77. Bike demand in winter is higher compared to spring and other seasons

---

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 to model the relationship between a dependent variable y and independent variables X

Linear Regression is carried out as follows

- Equation 1. Simple Linear Regression 2. Multiple Linear Regression
- Objective: Minimize the sum of squared residuals (errors) to find the best-fit line
- Fitting the Model: The coefficients β is calculated using the Ordinary Least Squares (OLS) method
- Assumptions

  Linearity: Relationship between X and y is linear

  Homoscedasticity: Constant variance of errors

  Normality: Errors are normally distributed

  Independence: Observations are independent

  No Multicollinearity: Independent variables are not highly correlated
- Evaluation**:** Metrics like R^2, Adjusted R^2, root mean squared error (RMSE), and mean squared error (MAE) are used to evaluate model performance

---

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: mean, variance, correlation, and regression line that have nearly identical statistical properties and it looks very different when plotted

It demonstrates the importance of visualizing data before drawing conclusions from statistical measures

All datasets have the same 1. mean and variance for x & y and 2. the correlation coefficient and 3. the linear regression equation

The 4 plots are 1. Linear relationship, 2. Nonlinear relationship, 3. Outlier affects the line fit and 4. Single influential point which determines the regression line

It is important to visualize data to identify patterns, outliers, or non-linear relationships that statistical summaries may miss to identify

---

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 correlation coefficient measures the linear relationship between two variables

It helps assess the strength and direction of a linear relationship between variables

R=Cov(X,Y)/σXσY

- Covariance between X and Y
- σX, σY is Standard deviations of X and Y

It ranges as below

- R∈[−1,1]
- R=1 Perfect positive linear relationship
- R=−1 Perfect negative linear relationship
- R=0 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)

<Your answer for Question 9 goes here>

Scaling transforms features to a specific range or distribution and it is performed for linear regression algorithms which is sensitive to feature magnitudes and prevents features with larger ranges from dominating the model and speeds up optimization algorithms

There are two types of scaling 1. Standardized and 2. Normalized scaling

Standardized scaling

- It centres the data by subtracting the mean and dividing by the standard deviation
- Results in zero mean and unit variance

Normalized scaling

- Scales values to a fixed range, typically [0, 1]
- Useful when feature values need to be bounded

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>

The Variance Inflation Factor (VIF) measures multicollinearity among independent variables

The Infinite VIF occurs when one variable is a perfect linear combination of others. This results in division by zero during VIF calculation where R^2=1

VIF=1/1−R^2

We can fix the infinite VIF by removing one of the collinear variables and use techniques like Principle Component Analysis (PCA) to reduce dimensionality

---

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>

The Quantile-Quantile (Q-Q) plot compares the quantiles of a variable's distribution to the quantiles of a theoretical distribution (for example normal distribution)

The Q-Q plots in Linear Regression validates the assumption of normality of residuals**,** if residuals are normally distributed, the points will lie close to the diagonal line

The Q-Q plots ensures reliable confidence intervals and hypothesis tests and detects heavy tails in residuals

---