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)

1) Seasonal variations and weather conditions play a key role in bike rentals. "Clear" weather has most number of booking with lowest in "Light rainy". "Heavy rainy day" does not have any booking

2) For 'yr' feature, during year 2018, company is just establishing so has less bookings compared to bookings in 2019

3) Derived categorical features (temp_category, windspeed_category) help in better model interpretation by grouping continuous data into meaningful categories.

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)

1) Keeping all categories can lead to a situation where one category can be predicted from others, which can cause issues in linear regression.

⍰2) Reduces Feature Count: Eliminating one category simplifies the model and improves efficiency

---

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)

In continuous variables, "Temp" and "windspeed", Temp is showing higher correlation with "cnt" variable

---

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)

Independent variables should not be highly correlated. Computed VIF and High VIF (>10) indicates multicollinearity.

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)

From stats model summary

1) Year (yr) → Coef: +2195.46 (p = 0.000) (Bike demand significantly increases over time.)

2) Spring Season (season_spring) → Coef: -2217.55 (p = 0.000): Demand decreases drastically in spring compared to fall
3) High Wind Speed (windspeed_category_High) → Coef: -940.47 (p = 0.000) : High wind speeds negatively impact bike demand.

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>
The Linear Regression Equation
The equation of a simple linear regression (with one variable) looks like:
Y= mX+c
Where:

- Y = Predicted output (dependent variable)
- X = Input feature (independent variable)
- m = Slope (coefficient, weight)
- c = Intercept (bias, the value of Y when X = 0)

For multiple features, the equation becomes:
Y=w1X1+w2X2+...+wnXn+b
Where:

- w1,w2,...,wn are the weights (coefficients) of the features
- b is the intercept

Our goal is to find the best values for the weights wi and b so that the predictions are as close as possible to the actual values.

2 Cost Function – We use the Mean Squared Error (MSE) as the cost function, which measures the difference between predicted and actual values:
Our goal is to minimize MSE, meaning we want the predicted values to be as close as possible to the actual values.

3 Finding the Best Model – Using Partial Derivatives & Gradient Descent
To minimize the cost function (MSE), we use gradient descent, an optimization algorithm that iteratively updates the model parameters.

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 set of four different datasets that have nearly identical statistical properties, but when visualized, they appear very different.

Why is Anscombe's Quartet Important?

- ➔ Shows that summary statistics can be misleading
- ➔ Highlights the need for data visualization
- ➔ Demonstrates the impact of outliers and patterns in data

The Four Datasets in Anscombe's Quartet

Each dataset consists of 11 (x, y) pairs, and all four datasets have the following nearly identical statistical properties:

1 Mean of X: ~9

2 Mean of Y: ~7.50

3 Variance of X: ~11

4 Variance of Y: ~4.12

5 Correlation between X and Y: ~0.816

6 Linear Regression Equation: y=3+0.5xy = 3 + 0.5xy=3+0.5x

Despite having the same summary statistics, they have very different distributions when plotted!

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 correlation coefficient (denoted as r) measures the strength and direction of a linear relationship between two continuous variables.

It gives a value between -1 and 1:

- ➔ r = 1 → Perfect positive correlation
- ➔ r = -1 → Perfect negative correlation
- ➔ r = 0 → No correlation

Important Note:

Pearson's R only detects **linear** relationships, not non-linear patterns.

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 transforming numerical features so that they have a comparable range, preventing one feature from dominating others due to large numerical differences.

Why is Scaling Performed?

1. Improves Model Performance: Many machine learning algorithms (e.g., linear regression, k-NN, SVM) perform better when features are on a similar scale.
2. Speeds Up Gradient Descent: In optimization-based models (like logistic regression, neural

networks), scaling helps convergence by ensuring weights update uniformly.

3. Prevents Features from Dominating: Without scaling, features with larger values (e.g., income in lakhs vs. age in years) will dominate distance-based models like k-NN or K-Means.

Difference between normalized scaling and standardized scaling:

- ➔ Normalization ensures data falls within a specific range, while standardization ensures data has a mean of 0 and a standard deviation of 1.
- ➔ Normalization is best for bounded data, whereas standardization is better for normally distributed data with potential outliers.

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 represents the level of multicollinearity in a dataset. It becomes infinite when there is perfect multicollinearity, meaning one or more features are highly correlated or exactly dependent on another feature.

Reasons for Infinite or Extremely High VIF:

1. Exact Linear Dependency
   - If one feature is an exact linear combination of other features, VIF becomes infinite.
   - Example: If Feature A = 2 × Feature B, then VIF for both will be ∞.
2. Dummy Variable Trap
   - If you create dummy variables for a categorical feature without dropping one category, one variable becomes a linear combination of others.
3. High Correlation Between Features
   - If two or more predictors are highly correlated (e.g., temp and atemp), VIF will be very high.
4. Small Sample Size with Many Features
   - When the number of features is close to the number of samples, VIF can become large due to overfitting.

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 a graphical tool used to check if a dataset follows a normal distribution. It compares the quantiles of the actual data against the quantiles of a theoretical normal distribution.

If the points lie roughly along a 45-degree straight line, the data is normally distributed.

Q-Q Plot Importance in Linear Regression:

In linear regression, one of the key assumptions is that the residuals (errors) should be normally distributed. A Q-Q plot helps verify this assumption.

Q-Q Plot interpretation:

-> Points lie on a straight line → Residuals are normally distributed which is a sign of a good model

-> Other curves or patterns means data is skewed which should note be the case