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)

Some Categorical Variables' effect on dependent variable.
Weathersit like Heavy Rain/Snow is no record in dataset, it may not have a significant impact, but when it does occur, it could negatively affect dependent variable.
Workingday: There are significantly more rentals on working days compared to non-working days
Season: The distribution across seasons appears fairly balanced
Year and month: The distribution of data across years is roughly equal

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)
to get k-1 dummies out of k categorical levels by removing the first level. This ensures the model does not have redundant variables to avoid Multicollinearity

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)

temp or atemp has the highest correlation with the target 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)
Checked Plot histogram of residuals vs. fitted values. It's roughly bell-shaped and symmetric
Verified using Variance Inflation Factor (VIF) and p-value that all selected features are low p-value (<0.05) and low VIF (<5) except temp

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)

Top 3 features contributing significantly: temp, year_2019 (yr), weathersit_Light Snow/Rain(weathersit)

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>

It is a supervised learning algorithm used for predicting continuous values based on input features. It establishes relationship between a dependent variable (target) and one or more independent variables (features) by fitting a straight line through the data points. It minimizes the sum of squared differences between actual and predicted values, assuming linearity and no multicollinearity among predictors

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>

It is a set of four datasets that have nearly identical statistical properties but look very different when plotted to demonstrate the importance of data visualization in statistical analysis. Each of the quartet's datasets has identical summary statistics, including mean, variance, and correlation coefficient, as well as the same linear regression line equation when plotted against each other.
Dataset 1: A simple linear relationship between two variables with minor scatter around the line of best fit.
Dataset 2: A non-linear relationship that could be modeled by a quadratic function, not a straight line.
Dataset 3: A perfect linear relationship except for one outlier point.
Dataset 4: A completely random scatter of points with no discernible pattern but still producing the same regression equation.

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>

also called the Pearson correlation coefficient (PCC), is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables.
It is a valuable tool for understanding linear relationships in data, with applications across various fields including business and social sciences.

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 value to ensure that all features contribute equally to a model by transforming them into a common scale. it is performed to ensure that all features contribute equally to a model's performance and to improve numerical stability.

Why is Scaling Performed?

Scaling ensures that all features contribute equally to a model by bringing them to a similar range. It is essential because:

1. Prevents Large Features from Dominating – Large-scale features (e.g., income vs. age) can overshadow smaller ones.
2. Speeds Up Gradient Descent – Helps optimization algorithms converge faster.
3. Improves Distance-Based Models
4. Enhances Model Interpretability – Ensures regression coefficients reflect actual feature importance

The difference between normalized scaling and standardized scaling:

Normalization (Min-Max Scaling)

- Rescales values to a fixed range [0,1]; [-1,1]
- Preserves relative distances but compresses large values
- Sensitive to Outliers
- Use Normalization when data is bounded and needs a fixed range.

Standardization (Z-Score Scaling)

- Centers data around zero mean (0) with unit variance (1)
- Maintains original distribution shape but shifts and scales it.
- Not really sensitive to Outliers
- Use Standardization when data is normally distributed or contains 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>

Because there is perfect multicollinearity, meaning one predictor is an exact linear combination of other predictors. It will be caused if two columns are identical or one feature can be perfectly predicted from others

---

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>

It is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not

Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. It helps to identifying outliers, skewness, or heavy tails. It also supports to comparing distributions of two datasets so it’s importance like a crucial diagnostic method in linear regression to access whether the residuals follow a normal distribution.