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)

Categories features will have demands as following

- Seasonality: Bike demand is heavily seasonal, with peaks during warm months.
- Weather Sensitivity: Clear weather boosts rentals, while rain and snow decrease them.
- Holidays and Working Days: These indicators affect user type rather than total rentals, as both categories may compensate for each other.

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)

drop_first=True ensures :

- No perfect multicollinearity in regression models.
- Clear interpretation of coefficients relative to a baseline category.
- Decreases the number of features

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(warm temperature) feature had highest correlation with 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 I used r2_score and also tried the model on test dataset which gave a promising result

R2_score on train dataset - **0.7886**

R2_score on test dataset - **0.7794**

I also used residual analysis which showed error distribution at zero

---

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 three features are

- Temp
- Yr
- spring

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 – Simplified Explanation

Linear regression models the **relationship between a dependent variable** (target) and **one or more independent variables** (features) by fitting a **straight line** through the data.

The general equation is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

- y: Target
- x_i: Independent variables
- β_0: Intercept
- β_i: Coefficients (effect of each feature)
- ϵ: Error term

How it Works:**
1. Find the best-fit line by minimizing the cost function (Mean Squared Error - MSE):

$$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$

2. Gradient Descent or analytical methods (like Normal Equation) are used to find the optimal coefficients.

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>

The quartet consists of four datasets, each with:

- **Identical mean** for the x and y variables.
- **Identical variance** of both x and y.
- **Identical linear regression line** (similar slope and intercept).
- **Identical correlation coefficient** (usually around 0.82).

Despite these similarities, the datasets are drastically different in structure, which becomes clear only through visualization.

For data sets

- A typical linear relationship
- A non-linear relationship
- Linear relationship with an outlier
- Vertical outlier affecting correlation

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 known as the Pearson correlation coefficient (PCC), is a statistical measure that indicates the strength and direction of a linear relationship between two variables. It ranges from -1 to +1, where:

+1: Perfect positive linear relationship (as one variable increases, the other increases proportionally).
-1: Perfect negative linear relationship (as one variable increases, the other decreases proportionally).
0: No linear relationship (the variables are uncorrelated).
It is named after the statistician Karl Pearson and is widely used in data analysis, especially in fields like machine learning, statistics, and data science, to explore the correlation between features.

Formula

Formula of Pearson's R

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \cdot \sum(Y_i - \bar{Y})^2}}$$

How to understand Pearson's R

- **|r| < 0.3: Weak correlation**
- **0.3 ≤ |r| < 0.7: Moderate correlation**
- **|r| ≥ 0.7: Strong correlation**
- **r = ±1: Perfect correlation**

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 used to adjust the range of features's values so that they fit within as specific scale. This insures the there is not a very bigg difference between two variable

Eg: Price for a product can be 20,000, where number of quantity can be only 2 so there is very big difference between both the features

- Normalization (Min-Max Scaling)

It Rescales the values to fit within a specific range, typically 0 to 1 but this range can be custom also. Iin this method outliers can be easily removed because all them will be grouped to 0 or 1
Formula

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

- **Standardization (Z-Score Scaling)**

Transforms features so they have zero mean and unit variance. In this method values follows a standard normal deviation (mean = 0 or std = 1). This method is less sensitive to outliers but still it can affect mean and standard deviation.

Formula

$$Z = \frac{X - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation.

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>
Some time VIF becomes infinite when one variable is perfectly acchives linear combination then that R square becomes 0

Scenario for this to occur

$$VIF(X_i) = \frac{1}{1-1} = \frac{1}{0} = \infty$$

Above scenarios can also occour when two variable are identical for example in Boom Bikes problem there was two features like temp and atemp.

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 compare the distribution of a dataset with a theoretical distribution (usually a normal distribution). It plots the quantiles of the sample data against the quantiles of the theoretical distribution. If the data follows the theoretical distribution closely, the points in the Q-Q plot will lie approximately on a straight 45-degree line.

How to understand Q-Q plot

- Points on or near the line: The data follows the theoretical distribution well.
- Curved deviations from the line: The data deviates from normality (e.g., skewed or heavy-tailed distributions).
- S-shaped curve: Indicates light tails (fewer extreme values than a normal distribution).
- Inverse S-shaped curve: Indicates heavy tails (more extreme values than expected under normality).

Use of Q-Q plot

- Assessing Normality of Residuals:
    - Linear regression assumes that residuals are normally distributed. A Q-Q plot allows us to visually confirm this assumption.
    - If residuals deviate significantly from the 45-degree line, it suggests that the

residuals are not normally distributed, which could indicate potential problems with the model.

- Identifying Skewness or Outliers:
  - Deviations from the line (especially at the tails) could indicate skewness or the presence of outliers.
  - This helps in deciding whether to transform variables or handle outliers.
- Ensuring Reliable Confidence Intervals::
  - The normality of residuals is necessary for valid confidence intervals and hypothesis tests on regression coefficients. If the assumption is violated, the results may be unreliable.
- Model Diagnostics:
  - If the Q-Q plot shows significant deviations from normality, it may suggest that a non-linear model or a different transformation of the data would provide better results.
