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)

The categorical variables in the dataset likely include features such as **season, yr, mnth, holiday, weekday, and workingday.** Their effect on the dependent variable (**cnt**) can be analyzed as follows:

1. **Season**: Different seasons may show varying bike usage trends. For example, summer might see higher ridership than winter due to favorable weather.
2. **Year**: If the dataset spans multiple years, we might observe a general upward or downward trend in bike rentals over time. But in the dataset only 2 years.
3. **Month**(**mnth**): Monthly variations may indicate seasonal effects or changes in commuting behavior.
4. **Holiday**: Holidays might show lower ridership if people commute less or higher if biking is used for leisure.
5. **Weekday** : Weekdays might exhibit higher rentals during workdays compared to weekends.
6. **Workingday**: This could differentiate between work-related commuting patterns and leisure-based biking trends.

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 avoid **multicollinearity**, which occurs when one dummy variable can be perfectly predicted from the others. By dropping the first category (drop_first=True), we prevent this redundancy, ensuring the model remains stable and 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)

temp variable has the highest correlation.

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**: Checked using scatter plots between predicted values and residuals to ensure no clear patterns exist.
- **Homoscedasticity**: Examined residual plots to verify that residual variance remains constant across all fitted values.
- **Multicollinearity**: Used the Variance Inflation Factor (VIF) to detect highly correlated independent variables (VIF > 10 indicates multicollinearity issues).
- **Normality of Residuals**

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 are:

1. temp
2. weathersit_3(weather situation)
3. yr (year)

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 for predicting a continuous dependent variable (Y) based on one or more independent variables (X). It assumes a **linear relationship** between the dependent and independent variables.

The equation of a simple linear regression model is:

$$Y=b0+b1X+\epsilon$$

Assumptions of Linear Regression

For Linear Regression to work well, it assumes:

1. **Linearity** ⟶ The relationship between X and Y is linear.
2. **Independence** ⟶ The observations are independent of each other.
3. **Homoscedasticity** ⟶ The variance of errors is constant across all values of X.

4. **No Multicollinearity** → Independent variables should not be highly correlated.
5. **Normality of Residuals** → The errors should be normally distributed.

How Linear Regression Works

Step 1: Fit the Model (Find Best-Fit Line)

The goal is to find the values of b_0, b_1, ..., b_n that minimize the error between the actual and predicted values.

The most common approach is **Ordinary Least Squares (OLS)**, which minimizes the **Sum of Squared Errors (SSE)**:

$$SSE=\sum(Yi-Y\hat{}i)2$$

Where:

- Yi = Actual value
- Y^i = Predicted value

OLS finds the best line that minimizes this error.

Model Evaluation Metrics

After training, we evaluate the model using following metrics:

1. **R-squared (r^2)**
    - Measures how well the independent variables explain the variability in Y.
    - Value ranges from **0 to 1** (higher = better).
2. **Mean Squared Error (MSE)**
    - Measures the average squared difference between actual and predicted values.
    - Lower MSE = better model.
3. **Root Mean Squared Error (RMSE)**
    - Square root of MSE, easier to interpret.
4. **Mean Absolute Error (MAE)**
    - Measures the average absolute difference between actual and predicted values.

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)

1. It demonstrates how **statistical summaries (like mean, variance, and correlation) can be misleading** without visualizing the data.Each dataset in the quartet has nearly **identical statistical properties**, but when plotted, they reveal **very different distributions and relationships** between the variables.
2. **The Four Datasets**

   Each dataset consists of **11 (x, y) pairs**, and they share the following similar statistical properties:

   - **Mean of X**: ≈ 9
   - **Mean of Y**: ≈ 7.5
   - **Variance of X**: ≈ 11
   - **Variance of Y**: ≈ 4.12
   - **Correlation between X and Y**: ≈ 0.816
   - **Regression equation**: y=3+0.5xy = 3 + 0.5xy=3+0.5x

   Despite these similarities, the scatter plots reveal **drastically different patterns**.

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, also known as the **Pearson correlation coefficient (PCC)**, measures the strength and direction of the linear relationship between two continuous variables. It is represented by the symbol rrr and has a value between **-1 and 1.**

1. **Positive Correlation (r>0):** As one variable increases, the other also increases. A positive slope in a scatter plot indicates a positive correlation.
2. **Negative Correlation (r<0):** As one variable increases, the other decreases. A negative slope in a scatter plot indicates a negative correlation.
3. **No Correlation (r=0):** No relationship between the variables.

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 **transforming numerical features** so they have a **similar range**, preventing some variables from dominating others due to their magnitude.For example, in a dataset with **height (cm)** and **salary (USD)**, salary values are much larger, which could negatively impact models like linear regression.

Why is Scaling Performed?

1. **Improves Model Performance** : Many ML algorithms perform better with scaled features.
2. **Prevents Dominance** : Large-magnitude features don't overpower smaller ones.
3. **Speeds Up Convergence** : Gradient-based models like Logistic Regression and Neural Networks train faster.
4. **Required for Distance-Based Models** : k-NN, SVM, and PCA are highly sensitive to feature scales.
5. **Better Interpretation** : Helps in understanding patterns in data.

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)

The **Variance Inflation Factor (VIF)** measures multicollinearity in regression models. It quantifies how much a predictor variable is correlated with other predictors

Why Does VIF Become Infinite?

VIF becomes **infinite** when **r^2=1**, meaning **perfect multicollinearity** exists.

1. One predictor is a perfect linear combination of others
2. Dummy variable trap (if one-hot encoding is done without `drop_first=True`)
3. Duplicate or highly correlated columns

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 is a graphical tool used to check if a dataset follows a particular distribution, typically the normal distribution. It compares the quantiles of the sample data with the quantiles of a theoretical distribution. If the data is normally distributed, points in the Q-Q plot will align closely along a 45-degree diagonal line.

Use of Q-Q Plot in Linear Regression:

In linear regression, we assume that residuals (errors) are normally distributed. A Q-Q plot is used to validate this assumption:

1. **Check Normality of Residuals** : Ensures residuals are normally distributed for valid hypothesis testing.
2. **Detect Skewness** : If points deviate from the line on one side, residuals may be skewed.
3. **Identify Heavy Tails (Kurtosis Issues)** : If residuals have long tails, they may follow a non-normal distribution.
4. **Detect Outliers** : Extreme deviations from the Q-Q line indicate possible **outliers**.