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)

Season (1-Spring, 2-Summer, 3-Fall, 4-Winter), season likely play major role in bike rental. Rental might be expected higher during summer or fall and lower in winter. Strong positive coefficient for Summer and Fall.
Year (0=2018, 1=2019) - positive coefficient due to yearly demand increase.
Month (1-12): Captures seasonal trends beyond just seasons, warmer months (April - October)
Weekday: (0-Sunday,..6 Saturday): Weekday and Weekends have strong coefficient.
Weather Situation (1=Clear, 2=Mist, 3=Light Snow,4=Heavy Rain/Snow): Bad weather has negative coefficient while good weather have positive coefficient.

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)

Using drop_first=True during dummy creation is important to avoid multicollinearity, or specially for dummy variable trap.
Dummy variable trap - when converting a categorical variable into multiple dummy (one-hot encoded) variables, one of them is redundant.

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)

Registered has the highest correlation with 'cnt' because bike rentals include registered users
Temp and atemp has also some strong correlation, as bike rental increase during comfortable weather
Hum and windspeed have weaker correlations

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 - relationship between independent and dependent variables should be linear, we identify relationship using scatter plot or correlation matrix

Should not have Multicollinearity - independent variables should not be highly correlated with each other, we can find out using VIF (variance inflation factor)

Residual Analysis - The variance of residuals should remain constants across all levels of predicted values, we can use distplot to see the smooth curve

---

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)

weathersit_3 (0.3064): this means that bad weather (light snow, light rain, etc) significantly reduces bike demand.

Yr (0.2504): the year variable (0 for 2018 and 1 for 2019) indicates that bike demand increased in 2019 compared to 2018

Windspeed (0.2502): higher wind speed negatively impact bike demand, making it less likely for people to use shared bikes

---

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 used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find the best-fitting line (or hyperplane) that predicts the target variable from the input features

Simple Linear Regression: y = b0 + b1 X + e
Involves single predictor variable
Y = target (dependent variable), X is feature (independent variable), b0 is intercept, b1 is the coefficient (slope of the line), e is the error term (residual)

Multiple Linear Regression: y = b0 + b1X1 + b2X2 + … + BnXn + e
Involves multiple predictor variables
Y = target (dependent variable), X1,X2,…,Xn are feature (independent variable), b0 is intercept, b1,b2,..bn are the coefficient (slope of the line) for each features, e is the error term (residual)

Find coefficients (b - beta) that minimises the error between predicted and actual values using a loss function, typically Mean Squared Error, MSE

Process:
* Model Representation: write the regression equation
* Estimate Coefficients: Use methods like Ordinary Least Squares (OLS)
* Prediction: Apply the model to predict target values
* Evaluation: Use metrics like R-Squared, MSE etc

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 datasets that have nearly identical simple descriptive statistics (mean, variance, correlation), but when graphed, reveal very different distributions and relationships. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data before making conclusions.

Purpose:

Show that summary statistics (mean, variance, correlation) can be misleading, and data visualization is crucial for proper analysis.

The Four Datasets:

Dataset 1: A simple linear relationship.
Dataset 2: A perfect quadratic relationship (non-linear).
Dataset 3: A linear relationship with one outlier.
Dataset 4: A vertical line with no variance in X, but variance in Y.

Identical Statistics:

All four datasets have:
Same mean of
X and Y,
Same variance of X and Y,
Same correlation between
X and Y.

Key Insight:

Even with identical summary statistics, the datasets have very different patterns and relationships. Visualizing the data is essential to understanding it.

Visualizing:

Scatter Plots of each dataset show the varying relationships (linear, quadratic, outliers, and vertical lines), highlighting how descriptive statistics alone can be insufficient.

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** is a statistical measure of the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where values closer to 1 or -1 indicate a stronger linear relationship, and values closer to 0 suggest little or no linear correlation

1. Definition:
    - Pearson's r measures the linear correlation between two variables.
2. **Formula**: $r = \Sigma[(Xi - \bar{X})(Yi - \bar{Y})] / \sqrt{[\Sigma(Xi - \bar{X})^2 \Sigma(Yi - \bar{Y})^2]}$
   Where:
    - Xi and Yi are the individual data points,
    - $\bar{X}$ and $\bar{Y}$ are the means of the variables X and Y.
3. Range:
    - r ranges from -1 to 1:
        - **r = 1**: Perfect positive linear relationship.
        - **r = -1**: Perfect negative linear relationship.
        - **r = 0**: No linear relationship.
4. Interpretation:
    - **r > 0**: Positive correlation (as one variable increases, the other tends to increase).
    - **r < 0**: Negative correlation (as one variable increases, the other tends to decrease).
    - **r ≈ 0**: Little to no linear relationship.
5. Assumptions:
    - Both variables should be continuous and approximately normally distributed.
    - The relationship between the variables should be linear.
6. Limitations:
    - Only measures linear relationships; does not capture non-linear associations.
    - Sensitive to outliers, which can distort the 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>

What is Scaling?
Scaling is the process of adjusting the range or distribution of data features to make them comparable.

Why is Scaling Performed?

- Improves model performance.
- Prevents features with larger ranges from dominating.
- Speeds up algorithm convergence.

Difference Between Normalized and Standardized Scaling:

1. Normalized Scaling:
    - Rescales data to a specific range (usually [0, 1]).
    - Used when data needs to fit a fixed range.
2. Standardized Scaling:
    - Rescales data to have a mean of 0 and standard deviation of 1.
    - Used when data follows a normal distribution or needs to handle outliers.

Key Difference:

- **Normalization**: Data is within a specific range.
- **Standardization**: Data has a mean of 0 and a standard deviation of 1.

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 value of VIF becomes infinite when there is **perfect multicollinearity** among the predictor variables, meaning one variable can be exactly predicted from others. This leads to a singular matrix in the calculation, making the VIF calculation impossible and resulting in an infinite value.

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 assess if a dataset follows a specific theoretical distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the expected distribution.

Use and Importance of a Q-Q Plot in Linear Regression:

1. Assess Normality of Residuals:
   - Helps check if the residuals (errors) follow a normal distribution, a key assumption in linear regression.
2. Identify Deviations:
   - Outliers or deviations from a straight line in the Q-Q plot indicate non-normality in the residuals.
3. Model Validity:
   - If residuals are normally distributed (as shown by a straight line in the Q-Q plot), the linear regression model's statistical tests and confidence intervals are valid.
4. Detect Skewness or Heavy Tails:
   - The Q-Q plot can reveal skewness or heavy tails in the data, indicating a need for transformation or different modeling approaches.
