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**: Rentals are higher in spring and summer and drop in winter, driven by favorable weather.
- **Year** (year): Rentals increased in 2012, indicating growing adoption.
- **Month** (month): Higher rentals occur from May to September; lower in winter months.
- **Weekday**: Rentals are consistent throughout the week, with minor drops on weekends.
- **Holiday**: Rentals decrease on holidays, reflecting reduced commuting.
- **Working Day**: Higher rentals on working days, driven by commuting demand.
- **Weather Situation**: Bad weather reduces rentals, while clear weather increases demand.

Conclusion: Seasonality, holidays, workdays, and weather significantly impact rentals. BoomBikes should focus on peak months, good weather, and workdays, while adjusting resources during holidays and bad weather.

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 variable creation is important to avoid the dummy variable trap and prevent multicollinearity in linear regression models.

Reasons

- Dummy Variable Trap:
    - When all categories of a categorical variable are represented as dummy variables, one category becomes redundant because it can be inferred from the others.
    - For example, if a variable season has four categories (1, 2, 3, 4), creating four dummies will make one of them predictable (e.g., if all other dummies are 0, the value must correspond to the dropped category).
- Multicollinearity Issue:
    - Including all dummies leads to perfect multicollinearity, meaning one predictor is a linear combination of others, which can make the regression coefficients unstable and hard to interpret.
- Solution with drop_first=True:
    - Dropping the first dummy (one category) solves this by removing redundancy. It keeps only (n-1) dummies, where n is the total number of categories, ensuring the model is well-specified.

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 (temperature) or atemp (apparent temperature) has the strongest positive correlation with

bike rentals, as people prefer biking in pleasant weather.

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)

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)

- Linearity
  - The relationship between predictors and the target variable should be linear. If a plot of actual vs. predicted values shows a linear pattern, the assumption holds.
- Normality of Residuals
  - The residuals (errors) should follow a normal distribution. This can be checked using a histogram or Q-Q plot—if the residuals align with a normal curve, the assumption is satisfied.
- Homoscedasticity
  - The variance of residuals should be constant across all predicted values. In a residual vs. predicted plot, random scatter without patterns confirms this assumption.
- Multicollinearity
  - Independent variables should not be highly correlated. A VIF (Variance Inflation Factor) score above 5 suggests multicollinearity, which can destabilize the model.
- Independence of Residuals
  - Residuals should not show patterns over time. Or no auto-correlation

---

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 and one or more independent variables. The goal is to find the best-fitting line (or hyperplane in higher dimensions) that describes this relationship.

The Model

In its simplest form, a linear regression model for one independent variable can be expressed as:
$y = mx + b$

Where:
y is the dependent variable (the value we want to predict)
x is the independent variable (the feature or predictor)
m is the slope of the line, representing how much y changes for a unit change in x
b is the y-intercept, representing the value of y when x is 0
Multiple Linear Regression

For multiple independent variables, the model becomes:

$y = b0 + b1*x1 + b2*x2 + ... + bn*xn$
Where:
y is the dependent variable
x1, x2, ..., xn are the independent variables
b0, b1, b2, ..., bn are the coefficients to be estimated
The Learning Process

The primary goal of linear regression is to find the optimal values for the coefficients (m and b or b0, b1, ...) that minimize the difference between the predicted values and the actual values. This is often done using a method called Ordinary Least Squares (OLS).

Ordinary Least Squares (OLS)

OLS aims to minimize the sum of the squared residuals, where a residual is the difference between the predicted value and the actual value. Mathematically, this can be expressed as:

minimize $\Sigma(y_i - \hat{y}_i)^2$
Where:
y_i is the actual value of the i-th data point
$\hat{y}_i$ is the predicted value of the i-th data point
Model Evaluation

Once the model is trained, its performance can be evaluated using various metrics:

Mean Squared Error (MSE): The average squared difference between the predicted and actual values.
Root Mean Squared Error (RMSE): The square root of the MSE, providing an error measure in the same units as the dependent variable.
Mean Absolute Error (MAE): The average absolute difference between the predicted and actual values.
R-squared: A statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variable(s).
Applications of Linear Regression

Linear regression is widely used in various fields, including:

Finance: Predicting stock prices, forecasting economic trends
Marketing: Analyzing customer behavior, predicting sales
Healthcare: Modeling disease progression, predicting patient outcomes
Social Sciences: Studying social phenomena, predicting election results

Linear regression is a fundamental statistical technique with numerous applications. By understanding its principles and limitations, you can effectively use it to model and predict relationships between variables.

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 group of four datasets that have the same basic statistics (like average, variance, and correlation) but look very different when you plot them. It was created by Francis Anscombe in 1973 to show that numbers alone don't tell the whole story.

What the Four Datasets Show:

Dataset 1:
A clear straight-line relationship. Perfect for linear regression.
Dataset 2:
A curved pattern that can't be captured by a straight line.
Dataset 3:
An outlier messes up the linear relationship.
Dataset 4:
Almost all points are on a vertical line except one, making the linear fit useless.

Why Anscombe's Quartet Matters:

Statistics alone can be misleading—different patterns can give the same numbers.
Visualizing data helps you spot things like outliers or curves that numbers might hide.
It's a reminder to always plot your data to make better decisions.
In short, seeing your data is as important as measuring it!

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 that tells us how strongly two variables are related and the direction of their relationship. It is used to determine if an increase in one variable corresponds to an increase or decrease in another.
Formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- x_i and y_i: Data points for the two variables
- $\bar{x}$ and $\bar{y}$: Mean of each variable

Interpreting Pearson's R:

- **Range**:
  Pearson's R can take values between **-1** and **1**.
  - **+1**: Perfect **positive correlation** (as one variable increases, so does the other).
  - **-1**: Perfect **negative correlation** (as one variable increases, the other decreases).
  - **0**: **No correlation** (no linear relationship between the variables).

In summary, Pearson's R helps you understand how two variables move together, whether in the same or opposite direction, and how strong that relationship is.

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 adjusts feature values to be on the same scale to prevent certain variables from dominating due to larger ranges or units.

Why Perform Scaling?

- Prevents bias in algorithms sensitive to magnitude (e.g., k-NN).
- Speeds up convergence in models like linear regression.
- Ensures all features contribute equally to the model.
- Improves interpretability and performance.

Types of Scaling:

1. **Normalization (Min-Max Scaling)**
   - **Scales data between [0, 1]**.
   - **Good for** data that doesn't follow a normal distribution or neural networks.
   - **Sensitive to outliers**.
2. **Standardization (Z-Score Scaling)**
   - **Centers data at mean = 0** and **standard deviation = 1**.
   - **Useful for** normally distributed data and models like PCA or SVM.
   - **Less affected by outliers**.

Key Difference:

- **Normalization** scales within a specific range (e.g., [0, 1]).
- **Standardization** centers data with mean 0 and unit variance.

In summary, **scaling** ensures that different features contribute equally to the model and improves the performance of algorithms. **Normalization** fits data into a specific range, while **standardization** makes data follow a mean-centered, unit-variance distribution. Choosing the right scaling technique depends on the type of data and the algorithm used.

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 becomes infinite when there is perfect multicollinearity, meaning one feature is fully predictable from others. This can happen due to:

- Duplicate or linearly dependent variables
- Dummy variable trap (all categories included)
- Highly correlated features (correlation close to ±1)

Solution:

- Remove redundant features
- Use drop_first=True for dummy variables
- Apply feature selection or PCA to reduce multicollinearity

In summary, infinite VIF occurs due to perfect multicollinearity, indicating that one feature can be perfectly explained by others, making the regression model unstable.

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 plot compares the distribution of residuals to a normal distribution. In linear regression, it helps check if the residuals are normally distributed, which is important for reliable predictions.
How to Interpret:
Points align along the diagonal: Residuals are normally distributed.
Points deviate from the line: Residuals show outliers or non-normality (e.g., skewness).
Importance:
Ensures the normality assumption for valid regression results.
Identifies issues like outliers or need for data transformations.
