Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

   All categorical variables that have an effect on dependent variables are temperatures, light rain and the year 2019.

   The effect of the temperature has the highest contribution in explaining the demand for bikes. With every unit increase in temperature, the demand goes up by 0.55 units.

   Light rain as the highest negative contribution to the demand. With every unit of light rain increase, there is decrease in demand by 0.29 units.

   And the last one is, the year 2019. From 2018 to 2019, there was a growth of 0.2331 units and hence if all conditions remains same, the next year growth can take this unit of growth for granted.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

   `drop_first=True` is essential during dummy variable creation:

   - Prevents Multicollinearity: Avoids perfect correlations among dummy variables, ensuring reliable model estimates.
   - Allows Meaningful Intercept Interpretation: Establishes a reference group for the intercept, representing its expected value.
   - Reduces Information Redundancy: Preserves categorical information without unnecessary variables.
   - Enhances Model Efficiency: Improves computational speed and potentially performance by reducing variables.
   - Combats Overfitting: Mitigates the risk of models adapting too closely to training data noise, leading to poor generalization.

In essence, this setting safeguards model integrity, interpretation clarity, and overall performance when working with categorical variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlationwith the target variable? (1 mark)**

   The temp and atemp variables are highly corelated and one of them can be dropped.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

   By Residual Analysis of the Train Data

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

   Top Gears for Bike Rides:

   - Temperature: Every degree hotter fuels a 0.55-unit ride surge, making it the sunshine king. Forget coffee, grab your wheels – warmth is the ultimate boost!
   - Light Rain: Even a sprinkle chills demand by 0.29 units, so watch the forecast and adjust your route. Every drop counts!
   - 2019: This year saw a 0.2331-unit jump compared to 2018, marking a turning point. Buckle up for more exciting climbs ahead!

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

   - It's a method for predicting a target variable based on linear relationships with other variables.
   - It finds the "line of best fit" that minimizes the difference between actual and predicted values.
   - The model produces coefficients (slopes) that indicate how much the target variable changes for each unit change in a predictor.
   - The intercept represents the predicted value when all predictors are zero.

Key steps:

1. Prepare data: Clean and split into training and testing sets.
2. Fit model: Find the best-fitting line using an algorithm like Ordinary Least Squares.
3. Evaluate: Assess performance and check model assumptions.
4. Predict: Use the model to forecast for new data.

Common uses: Predicting prices, grades, sales, stock prices, and more.

Strengths: Simple, efficient, widely applicable.

Limitations: Assumes linearity, sensitive to outliers, affected by multicollinearity.

2. Explain the Anscombe's quartet in detail. (3 marks)

The Anscombe's Quartet: four datasets with identical stats, radically different stories. Plotting reveals:

- Linear Lovers: Perfect straight line, textbook romance.
- Curvy Fling: A parabolic whisper, not linear friends.
- Outlier Intruder: Single point crashes the party, skewing the scene.
- Independent Souls: No connection whatsoever, just points in space.

The lesson? Stats can lie. Visualize, question, embrace complexity!

3. What is Pearson's R? (3 marks)

Pearson's R, a number between -1 and 1, gauges the linear connection between two continuous variables. Positive values mean they rise and fall together, negative means one goes up as the other down, and 0? Just friends, no connection. Think of it as a data love meter!

4. What is scaling? Why is scaling performed? What is the difference between normalized scalingand standardized scaling? (3 marks)

Scaling: The Great Equalizer

- Why Scale? Algorithms often work better when features have similar scales. It prevents bias towards those with larger ranges and helps with outlier control and convergence.
- Normalization vs. Standardization:
    - Normalization squeezes values into a set range (like 0 to 1).
    - Standardization centers values around a mean of 0 with a standard deviation of 1.
- Choose wisely: Normalize for algorithms sensitive to ranges (neural nets, k-NN), standardize for those sensitive to outliers (linear regression, SVM).

Key takeaway: Scale features for fairer, smoother ML journeys!

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Two main culprits cause infinite VIF:

1. Perfect Collinearity: One feature is exactly a linear combination of others, making the model unstable and uninterpretable.
2. Singular Matrix: Features have linear dependencies, leading to unreliable coefficients and predictions.

To tackle these culprits:

- Identify highly correlated features: Check the correlation matrix for near-perfect correlations.

- Drop or combine redundant features: Remove overlapping information or create new features from combinations.
- Apply regularization techniques: Techniques like Ridge or Lasso can manage the effects of multicollinearity.

Remember, infinite VIF flags potential instability. Taking action ensures your model delivers reliable and interpretable results.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

Visual check for normality: Q-Q plots compare data quantiles to a theoretical distribution (often normal). Points on a straight line signify good fit, deviations reveal potential issues with your linear regression model. This helps interpret results and guide improvements.

Key takeaway: A simple but powerful tool for checking a vital assumption and boosting your model's validity.