# 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)

# 

**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 in dummy variable creation is important to avoid multicollinearity, which can distort the results of a linear regression model. By dropping the first category, we reduce redundancy in the dataset, ensuring that each piece of information is independent. This helps in making the model’s estimates more 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)

# Among the numerical variables, temp (temperature) and atemp (feeling temperature) showed the highest positive correlation with the target variable 'cnt'. This indicates that warmer temperatures generally increase bike rentals, likely due to more comfortable biking condition.

**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)

We confirmed the assumptions of linear regression by examining the variance inflation factor (VIF), the distribution of error residuals, and the linear relationship between the dependent variable and an independent variable.

1. Verified that the error distribution is normal.
2. Reviewed residuals against predicted values and found no discernible patterns.
3. Observed constant variance in error terms, confirming the assumption of homoscedasticity.

**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)

# Weather Situation (Light Snow/Rain): This feature had the most significant negative impact on bike rentals.

# Month (July): Being in July negatively impacted bike rentals significantly.

# Season (Spring): Spring also showed a significant negative impact on bike rentals.

# 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 an essential statistical technique that models the relationship between a dependent variable and one or more independent variables using a linear approach. This method is characterized by the equation:

y=mx+b

where:

 y represents the dependent variable we aim to predict.

 x is the independent variable used for predictions.

 m denotes the slope, indicating the rate of change in y for each unit change in x.

 b is the y-intercept, the value of y when x=0.

**Steps in Linear Regression:**

1. **Data Collection:** Gather a dataset that includes both independent (features) and dependent (target) variables.
2. **Data Preprocessing:**
   * Address missing values by filling them or removing incomplete rows.
   * Convert categorical variables into numerical format as needed.
   * Partition the dataset into training and testing sets.
3. **Model Training:** Fit the linear regression model to the training data to learn the coefficients (slope and intercept) that minimize prediction errors.
4. **Model Evaluation:** Predict outcomes using the testing set and assess the model's accuracy with metrics like Mean Squared Error (MSE), R-squared, or Root Mean Squared Error (RMSE).

**Model Assumptions:** Linear regression relies on several assumptions:

* Linearity: The relationships between the independent and dependent variables are linear.
* Independence: Observations do not influence each other.
* Homoscedasticity: Errors exhibit constant variance across all levels of the independent variables.
* Normality: The errors are normally distributed.

**Model Evaluation:** Evaluate the model's effectiveness using:

* Mean Squared Error (MSE): The average of squared differences between predicted and actual values.
* R-squared (R²): The proportion of the variance in the dependent variable explained by the independent variables.
* Root Mean Squared Error (RMSE): The square root of MSE, providing an error metric in the same units as the dependent variable.

**Advantages and Disadvantages:**

* **Advantages:**
  + Simplicity and ease of understanding.
  + Interpretable coefficients.
  + Handles both continuous and categorical data effectively.
* **Disadvantages:**
  + Assumes linearity, which might not always hold true.
  + Sensitive to outliers, affecting model accuracy.
  + Performance might degrade if assumptions are violated.

**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 comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset illustrates the importance of visualizing data before analyzing it, as it shows that different data distributions can look similar in statistical terms.

**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, or Pearson correlation coefficient, measures the strength and direction of association between two continuous variables. Values range from -1 to 1, where 1 means a perfect positive linear relationship, -1 means a perfect negative linear relationship, and 0 means no linear relationship.

**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 the process of adjusting the scale of features in data. It is performed to handle disparities in units and to bring all features to a uniform scale, which helps in improving the performance and convergence of machine learning algorithms.

#  Normalized scaling adjusts data varying between [0,1], while standardized scaling transforms data to have a mean of zero and a standard deviation of one.

**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>

# An infinite VIF (Variance Inflation Factor) occurs when perfect multicollinearity exists among the independent variables in a regression model, meaning some variables can be linearly predicted from others with absolute accuracy.

**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 used in linear regression to check the normality assumption of residuals. It compares the quantiles of residuals to the expected normal distribution quantiles. Deviations from the line in a Q-Q plot indicate deviations from normality.

These responses are based on our analysis and general knowledge. If you need any changes or additional details, feel free to ask!