# 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 "Season vs. Rentals per Day" plot highlights that fall records the highest average rentals, closely followed by summer.

Analyzing year-on-year trends, 2019 exhibits a median increase of 2,000 rentals compared to 2018.

The "Month-wise Rentals" plot indicates that September consistently records the highest rentals, with neighboring months also showing elevated numbers, reflecting a clear seasonal pattern.

On holidays, rental counts are generally lower than on working days, accompanied by a greater variability in demand.

Weekday rental patterns reveal no significant differences, except for Thursdays and Sundays, which display higher variability compared to other days.

Upon analyzing the correlations between selected continuous variables and the count of bike rentals (cnt), the following insights emerge:

atemp (Adjusted Temperature): This variable demonstrates the strongest positive correlation with cnt, indicating that an increase in adjusted temperature is closely associated with a rise in bike rentals.

temp (Temperature): While positively correlated with cnt, the strength of this relationship is slightly lower than that of atemp.

hum (Humidity): Among the variables considered, hum exhibits the weakest correlation with cnt, suggesting that variations in humidity have a relatively limited influence on bike rental counts.

temp and atemp: The correlation between these two variables is nearly perfect, underscoring a strong linear relationship. Given this redundancy, it may be prudent to exclude one of them in subsequent analyses to mitigate potential multicollinearity.

# 

**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 prevents duplication, avoids confusion in regression, and ensures predictors are independent.

**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)

# Atemp variable demonstrates the strongest positive correlation with cnt, indicating that an increase in adjusted temperature is closely associated with a rise in bike rentals.

**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)

# I validated Linear Regression assumptions by checking residual normality, constant variance, linearity with predictors, and independence of errors using residual plots, histograms, and statistical tests.

**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)

Temperature (temp): Temperature emerges as the most influential factor positively impacting bike rental demand. Warmer temperatures significantly boost rental activity.

Year (yr): The upward trend in bike rentals over the years highlights a consistent positive impact of time progression on demand.

Winter (winter): Demand appears to increase during the winter season, potentially driven by holiday activities and seasonal recreation.

# 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 (Y) and one or more independent variables (X). It tries to fit a straight line that best represents the relationship between these variables.

**Equation of the Line**: In simple linear regression, the relationship is represented by the equation:  
Y=β0+β1X+ϵY = \beta\_0 + \beta\_1X + \epsilonY=β0​+β1​X+ϵ  
where:

* + YYY is the dependent variable (what we want to predict).
  + XXX is the independent variable (input or feature).
  + β0\beta\_0β0​ is the intercept (value of Y when X is 0).
  + β1\beta\_1β1​ is the slope (rate of change of Y for a unit change in X).
  + ϵ\epsilonϵ is the error term (difference between the predicted and actual values).

**Objective**: The goal of linear regression is to find the best values for β0\beta\_0β0​ and β1\beta\_1β1​ so that the line fits the data well. This is done by minimizing the difference between the predicted values (from the line) and the actual values in the dataset.

**Cost Function**: To measure how well the model fits the data, we use a cost function called Mean Squared Error (MSE), which calculates the average of the squared differences between the predicted and actual values. The model aims to minimize this error.

**Optimization**: The most common way to find the best line is through a method called **Ordinary Least Squares (OLS)**. This method adjusts the values of β0\beta\_0β0​ and β1\beta\_1β1​ iteratively, minimizing the MSE.

**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 consists of four datasets with identical summary statistics (mean, variance, correlation), but different visual patterns when plotted. It shows that summary statistics can be misleading, and it’s important to plot the data to understand its true nature. The key lesson is that visualizing data helps reveal patterns and outliers that numbers alone may miss.

**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>**for 4 marks use simple direct word onlhy

###### ChatGPT said:

ChatGPT

Pearson’s R (or Pearson correlation coefficient) is a statistical measure that shows the strength and direction of the linear relationship between two variables.

**Range**: It ranges from -1 to +1.

* + **+1** means a perfect positive relationship (as one variable increases, the other increases).
  + **-1** means a perfect negative relationship (as one variable increases, the other decreases).
  + **0** means no linear relationship.

**Formula**: It is calculated by dividing the covariance of the variables by the product of their standard deviations.

**Interpretation**:

* + A high positive or negative value (close to +1 or -1) indicates a strong relationship.
  + A value near 0 indicates a weak or no linear relationship.

**Importance**: Pearson’s R is used to understand how strongly two variables are related, helping in prediction and analysis of trends. However, it only measures linear relationships and doesn’t capture non-linear patterns.

**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 range or distribution of features in a dataset to make them comparable. It is done to improve the performance of machine learning models.

**Scaling is Performed**:

* Different features may have different units or scales (e.g., age in years, salary in thousands), which can affect the model’s performance.
* It helps algorithms like distance-based models (e.g., KNN, SVM) perform better.

**Normalized Scaling**:

* Rescales the data to a range between 0 and 1.
* Formula: X\_norm=X−min(X)max(X)−min(X)\text{X\\_norm} = \frac{X - \text{min}(X)}{\text{max}(X) - \text{min}(X)}X\_norm=max(X)−min(X)X−min(X)​

**Standardized Scaling**:

* Rescales the data to have a mean of 0 and a standard deviation of 1.
* Formula: X\_std=X−μσ\text{X\\_std} = \frac{X - \mu}{\sigma}X\_std=σX−μ​, where μ\muμ is the mean and σ\sigmaσ is the standard deviation.

**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>

A **VIF (Variance Inflation Factor)** becomes infinite when two or more independent variables in a regression model are **perfectly correlated**. This means one variable can be exactly predicted using the other(s).

When this happens:

* The model can’t separate their effects properly.
* The VIF formula divides by a very small number, causing it to become infinite.

**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 (Quantile-Quantile plot) is a graphical tool that compares the distribution of a dataset with a theoretical distribution, such as the normal distribution. In linear regression, it is used to check if the residuals (errors) follow a normal distribution. This is important because many regression assumptions, like the validity of hypothesis tests and confidence intervals, rely on normally distributed residuals. If the points in the Q-Q plot form a straight line, it suggests the residuals are normally distributed, confirming the model’s reliability