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)

Weather Situation: The 'weathersit' variable has a noticeable impact on bike rentals. Clear or partly cloudy weather (weather situation 1) shows the highest median and overall distribution of bike rentals. As the weather deteriorates (weather situations 2, 3, and 4), the median rental count decreases, and the variability in rentals also changes. Heavier rain, snow, or thunderstorm significantly reduces rentals.

Season: Seasons 2 and 3 (likely summer and fall based on the provided mapping) show moderate effects on bike rentals, suggesting higher rental counts compared to other seasons. The box plots likely show differences in the distribution of rentals across seasons, with some seasons exhibiting higher median rental counts and potentially less variability.

Month (mnth): The month of the year influences rentals. There's likely a seasonal trend with lower rentals in winter months and higher rentals in warmer months, especially during summer (June-August).

Year (yr): The box plots for the year variable would likely show an overall upward trend in bike rentals between the two years, indicating that the number of rentals is increasing over time.

Weekday (weekday): Weekday likely exhibits a weekly pattern, potentially showing lower rentals on weekends (higher on weekdays).

Working Day (workingday): Working days may show a different rental pattern compared to non-working days (weekends and holidays).

Holiday (holiday): Holidays might have lower rental counts compared to non-holiday days.

In summary: Weather and season are the most prominent factors influencing bike rentals. The other variables (month, year, weekday, working day, holiday) likely contribute to smaller variations or more subtle patterns in the rental counts. The visualizations (boxplots) give the distribution of rentals for each category, providing additional insights beyond just the mean counts for each category.

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 helps to remove multicollinearity (also known as the dummy variable trap) in regression models. Multicollinearity occurs when one dummy

variable can be perfectly predicted by others, which affects the interpretability and stability of regression coefficients. Dropping the first category ensures the model remains identifiable, as one category serves as the baseline for comparison, reducing redundancy and making the model more efficient.

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)

Temperature and Apparent Temperature: Both seem to be significant predictors of 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)
We have done the following checks

1. linear Relation ship check between Independent and Dependant variables.
2. Check the Normal distribution between the residuals such that they have constant variance.
3. there is no visible patterns in residuals and are Indpendant among them.
4. Also checked Insignificant MultiCollinearity among variables using VIF.

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)
Following Features with positive coefficients increase the target variable

1. year
2. atemp (feeling temperature in Celsius)
3. winter season

while following negative coefficients decrease it

1. season_spring
2. month_nov
3. weathersit_Light (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

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)

Linear regression is a method used to understand the relationship between a target variable (Y) and one or more independent variables (X).
It fits a straight line through the data to predict Y based on the values of X.
For example, in the equation $Y = m\,x + c$, m represents how much Y changes when X increases by 1.

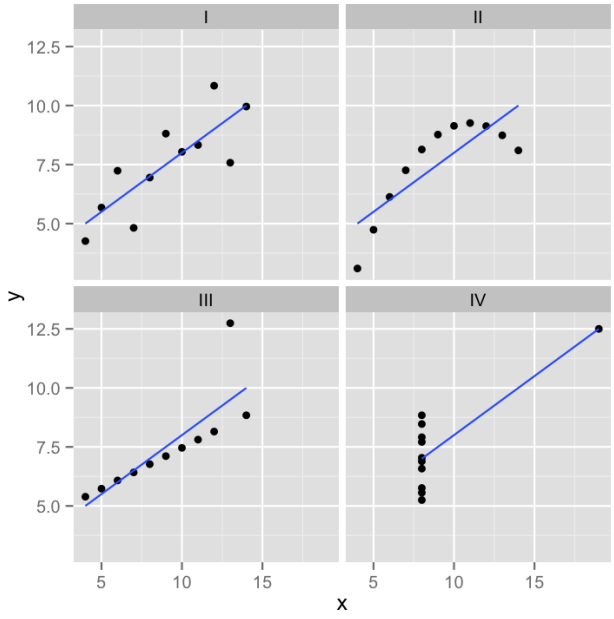
In multiple linear regression, this concept is extended to include multiple independent variables (X 1 , X 2 , . . . , $X\,n$), helping us understand how these factors together explain the variation in Y. The better the variables explain Y , the more accurate the predictions will be. This approach also allows us to use the relationship to estimate future outcomes.
$\hat{y}$ is the respective estimate of the *y*-value. This means that for each *x*-value the corresponding *y*-value is estimated.
If all points (measured values) were exactly on one straight line, the estimate would be perfect. However, this is almost never the case and therefore, in most cases a straight line must be found, which is as close as possible to the individual data points. The attempt is thus made to keep the error in the estimation as small as possible so that the distance between the estimated value and the true value is as small as possible.

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)

Anscombe's Quartet is a set of four datasets that have nearly identical statistical properties but differ significantly in their distributions and relationships when visualized. It was introduced by statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it statistically.

Key Characteristics of Anscombe's Quartet
Each dataset in the quartet has nearly the same:

1. mean of the x and y values
2. variances of x and y
3. Correlation between x and y
4. Linear regression line.



Above is a visual representation of Anscombe's Quartet, showing the distinct patterns in each dataset while their statistical properties remain similar

---

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)

Pearson's r is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It ranges from −1 to +1 and is widely used in statistics and data analysis. The values of both the sample and population Pearson correlation coefficients are on or between −1 and 1.
The Pearson correlation coefficient is symmetric: corr(X,Y) = corr(Y,X).
The further the coefficient is from zero, whether it is positive or negative, the better the fit and the greater the correlation.
The values of -1 (for a negative correlation) and 1 (for a positive one) describe perfect fits in which all data points align in a straight line, indicating that the variables are perfectly correlated.

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)

Scaling can be used for

1. Minimizing Computation in Gradient-Descent-Based Models:
   Scaling helps reduce computation requirements for models like regression and neural networks that rely on gradient descent. When features are on vastly different scales,

gradient updates can become unstable, slowing down convergence or even causing failure to converge.

2. Comparing Beta Coefficients for Interpretation:
   After building regression models, scaling ensures that beta coefficients (model weights) are directly comparable. When all features are on a common scale, it becomes easier to interpret their relative importance for inference purposes.
3. Improving Algorithm Performance:
   The machine learning algorithms perform better when features are scaled, as it ensures faster and more efficient optimization.

Standardization transforms the data so that it has a mean of 0 and a standard deviation of 1. This method does not bound the data to a specific range, but rather centers the data around 0.

Characteristics:

- The transformed data has a **mean of 0** and **standard deviation of 1**.
- It is **not bound** to a fixed range, so the values can be positive or negative.
- **Not sensitive to outliers** as much as normalization, because the mean and standard deviation are used instead of the min and max.

Normalization (or Min-Max scaling) transforms features to a fixed range, typically [0, 1], by using the minimum and maximum values of the data.

Characteristics:

- The scaled data will always be between **0 and 1** (or any other specified range, but typically [0, 1]).
- Sensitive to outliers. If there is an outlier in the dataset, it can compress the range of the other values, as it takes the **min** and **max** values into account.
- This is particularly useful when you have data that doesn't follow a normal distribution and you want to scale it to a fixed range.

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)

As VIF measures Multicollinearity between variables/columns . if there are other columns which are highly correlated with the dependent variable then VIF becomes infinity .

When using dummy variables to encode categorical variables, if all levels of the categorical variable are included without excluding one (e.g., omitting a reference level), perfect multicollinearity arises. This results in infinite or very high VIF values.

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)

A Q-Q plot is a graphical method for determining how well a set of data follows a known probability distribution. It compares the distribution of the data to a theoretical distribution, typically a normal distribution, by plotting the quantiles of the data against the quantiles of the theoretical distribution.
If the data follows the theoretical distribution, the points on the plot will lie along a straight line.

Also in **linear regression** it helps to check if the Error terms are normally distributed or not ,which will help you check the efficiency of model you have created.

Common uses :
It is used for checking if your dataset is following the Normal distribution or not.
Assessing Normality of Residuals:
Linear regression assumes that the residuals are normally distributed.
If the Q-Q plot shows large deviations from the straight line, the normality assumption is violated ,so which indicates we need to rebuild the model by refitting variables.
Identifying Skewness:
Upward curve (right tail above line): Indicates positive skewness.
Downward curve (right tail below line): Indicates negative skewness.
