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)

Ans-1) From the analysis of categorical variables, we can infer that:

- **Season**: Demand is higher in Spring, with a noticeable peak in this season.
- **Month**: September shows a significant surge in demand, likely due to favorable weather conditions.
- **Holiday**: Demand varies on holidays, with fluctuations in rentals during festive periods.
- **Weekday**: Weekends, particularly Sundays, tend to have higher demand.
- **Weather Situation**: Light Snow and Mist + Cloudy weather reduce demand significantly.

These categorical factors significantly influence the demand for bike-sharing services, as inferred from the data analysis.

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)

Ans-2) Using drop_first=True during dummy variable creation is important because it prevents multicollinearity, a situation where one predictor variable can be perfectly predicted from the others.

When categorical variables are converted into dummy variables, each category becomes a separate binary column. If all categories are kept, one column will be a linear combination of the others, which can distort the results of regression models by causing redundant information. By dropping the first category, we avoid this problem, ensuring that the dummy variables remain independent of each other and the model interprets them correctly.

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)

Ans-3) From the pair-plot among the numerical variables, temperature (temp) has the highest correlation with the target variable cnt. This suggests that as the temperature increases, the demand for bike-sharing also tends to rise, indicating a positive relationship.

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)

Ans-4) After building the linear regression model, I validated the assumptions as follows:

- **Linearity**: Checked using scatter plots and the correlation matrix.
- **Normality of Residuals**: Verified with a histogram and Q-Q plot.
- **Homoscedasticity**: Assessed using the residuals vs. fitted values plot, showing no patterns.
- **No Multicollinearity**: Ensured by calculating the Variance Inflation Factor (VIF), which showed no high values.

These checks confirmed that the linear regression assumptions were met.

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)

Ans-5) Based on the final model, the top 3 features contributing significantly to explaining the demand for shared bikes are:

1. **Temperature (temp)** - Higher temperatures correlate with increased bike-sharing demand.
2. **Holiday** - Demand tends to be higher during holidays.
3. **Season** - Spring shows the highest demand, indicating a strong seasonal effect.

Additionally, **September** also plays a significant role due to favorable weather conditions, making it another important factor in the model.

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)

Ans-6) Linear regression is a statistical method that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation. The goal is to find the best-fitting line by minimizing the sum of squared residuals (errors).

Key steps in linear regression:

1. **Fitting the model**: The model finds coefficients for each feature to minimize the error between predicted and actual values using the least squares method.
2. **Assumptions**:
   - The relationship between variables is linear.
   - Residuals (errors) are independent and normally distributed.
   - Homoscedasticity: constant variance of errors.
3. **Evaluation**: The model's performance is assessed using **R-squared** (goodness of fit) and **p-values** (significance of coefficients).

Linear regression helps in understanding relationships between variables and making predictions when assumptions are met.

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)

Ans-7 Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics, such as mean, variance, correlation, and regression line, but they differ in their distributions and visual patterns. The purpose of Anscombe's quartet is to highlight the importance of visualizing data before drawing conclusions based solely on statistical metrics. The four datasets have:

1. **Identical means** for the x and y variables.
2. **Similar variances** for both x and y variables.
3. **Similar correlation** between x and y.
4. **Identical linear regression lines** when analyzed statistically.

However, when plotted, the datasets show very different patterns:

- **Dataset 1** shows a linear relationship with no outliers.
- **Dataset 2** has a curved pattern despite appearing linear from the summary statistics.
- **Dataset 3** contains a strong outlier that distorts the data.
- **Dataset 4** has a vertical pattern with one large outlier affecting the regression.

Anscombe’s quartet emphasizes the necessity of data visualization for proper analysis and avoiding misleading conclusions from summary statistics alone.

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)

Ans-8) Pearson's R (also known as the Pearson correlation coefficient) is a measure of the linear relationship between two variables. It ranges from -1 to 1, where:

- **+1** indicates a perfect positive linear relationship (as one variable increases, the other increases in a perfectly linear manner).
- **-1** indicates a perfect negative linear relationship (as one variable increases, the other decreases in a perfectly linear manner).
- **0** indicates no linear relationship between the variables.

Pearson's R is calculated using the covariance of the variables divided by the product of their standard deviations. It helps to quantify the strength and direction of the linear relationship, making it useful for assessing how closely two variables move together. A value closer to 1 or -1 suggests a strong relationship, while a value near 0 indicates a weak or 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)

Ans-9) Scaling is the process of transforming features to a common scale, making sure that one feature does not dominate others due to its larger magnitude. It is important because many machine learning algorithms, such as linear regression, distance-based algorithms like KNN, and neural networks, perform better or converge faster when the features are on the same scale.

There are two common types of scaling:

- **Normalized Scaling**: This technique scales the data to a fixed range, typically 0 to 1, by subtracting the minimum value and dividing by the range of the data.
- **Standardized Scaling**: In this method, data is rescaled to have a mean of 0 and a standard deviation of 1. It's achieved by subtracting the mean and dividing by the standard deviation.

The key difference is that normalized scaling works on adjusting the data within a specific range, while standardized scaling adjusts data based on the mean and 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)

Ans-10) A VIF (Variance Inflation Factor) value becomes infinite when there is perfect multicollinearity between two or more independent variables in the dataset. This means that one of the independent variables is a perfect linear function of another variable, leading to a situation where the variable's variance cannot be estimated reliably.
In such cases, the model cannot distinguish the individual contributions of these highly correlated variables, resulting in infinite VIF values. This indicates that the independent variables should be reconsidered, and redundant variables should be removed to ensure that multicollinearity does not affect the model's performance.

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)

Ans-11) A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a particular distribution, typically the normal distribution. It compares the quantiles of the data with the quantiles of a theoretical distribution, such as the normal distribution. If the data points lie along a straight line, it indicates that the data follows the expected distribution.

In the context of **linear regression**, the Q-Q plot is essential for validating the **normality of residuals**. For linear regression to produce reliable results, the residuals (the differences between observed and predicted values) should be normally distributed. By examining the Q-Q plot, we can visually check if the residuals deviate significantly from a normal distribution. If the residuals deviate from the straight line, it may suggest violations of the linear regression assumption of normality, potentially affecting the model's accuracy.