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)

---

1. When the weather is clear more people using bike sharing service.
2. Year 2019 shows more count compared to previous year.
3. Summer & Fall season see people using the service more.
4. Weekdays shows more usage compared to weekends.
5. More people use the service when the windspeed between 8-15.

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 is important for several reasons, primarily related to avoiding multicollinearity and improving the interpretability of your models. It also reduce redundancy and improve efficiency.

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)

Temp & aTemp is having the strongest correlation.

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)

Validating the assumptions of linear regression is crucial to ensure that your model is reliable and that the results can be interpreted correctly. There is a linear regression between dependent variable(cnt) and the independent variables. The error terms are also normally distributed. We are not able to find any visible pattern in residual values.

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)

1. Temperature
2. Year

3. Winter season

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 fundamental and widely used algorithm in machine learning for predicting a continuous dependent variable based on one or more independent variables. It assumes a linear relationship between the 1 variable, meaning the relationship can be modeled by a straight line (in simple linear regression with one independent variable) or a hyperplane (in multiple linear regression with multiple independent variables).

There are 2 types of linear regression:

1. Simple linear regression
2. Multiple linear regression.

The algorithm is relatively straightforward to implement and interpret. The algorithm is relatively straightforward to implement and interpret. The coefficients can be used to understand the relative importance of different independent variables.

Linear regression relies on several key assumptions:

1. Linearity: The relationship between the independent and dependent variables is linear.
2. Independence: The errors are independent to each other.
3. Homoscedasticity: The variance of the errors is constant across all levels of the independent variables.
4. Normality: The errors are normally distributed.

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 are specifically designed to demonstrate the importance of visualizing data before relying solely on summary statistics. These datasets, created by statistician Francis Anscombe in 1973, have nearly identical descriptive statistics (like mean, variance, and correlation) but are drastically different when plotted graphically.

Each dataset in the quartet consists of 11 pairs of (x, y) values. While the exact numbers are not crucial to the understanding, it's important to know that they were carefully constructed to share certain statistical properties.

All four datasets have the following (or very nearly the following) properties:

1. Mean of X
2. Mean of Y
3. Variance of X
4. Variance of Y
5. Correlation coefficient.

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 also known as Pearsons the Pearson correlation coefficient. It is a statistical measure that quantifies the strength and direction of a linear relationship between two quantitative variables. It's a widely used tool in statistics and data analysis to understand how two variables move together.
Pearson's R ranges from -1 to +1:

- +1: Perfect positive correlation. The variables move together in perfect alignment.
- 0: No linear correlation. There's no linear relationship between the variables.
- -1: Perfect negative correlation. The variables move in opposite directions in perfect alignment.

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 is the process of transforming numerical features to a similar range of values. it's a data preprocessing step where all features are brought to a comparable scale, usually by normalization or standardization. It helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardized scaling: It replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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)

If the VIF value is infinite, it means that perfect multicollinearity among your predictor variables. Perfect multicollinearity means that one predictor variable can be perfectly predicted by a linear combination of the other predictor variables.

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 (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It's a powerful visual method for comparing the quantiles of your data to the quantiles of the theoretical distribution you're interested in. Quantiles divide a dataset into equal parts. For example, quartiles divide the data into four equal parts (25%, 50%, 75%, and 100%). A Q-Q plot is created by plotting the quantiles of your data against the corresponding quantiles of the theoretical distribution. If your data follows the theoretical distribution, the points on the Q-Q plot will fall approximately along a straight diagonal line.Deviations from this line indicate that your data does not perfectly match the theoretical distribution