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)

In order to infer the effect of categorical variables on the dependent variable (in this case, the bike demand represented by the cnt column), we need to understand how the categorical variables relate to the bike demand.the dependent variable that we need to analyze are Season, Weather situation, Year.

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)

It is important to use drop_first=True during dummy variable creation to avoid the dummy variable trap and prevent multicollinearity in the regression model.

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)

---

pair plot helps in visualizing the relationships between all numerical variables. It will show scatter plots between pairs of numerical features, and along the diagonal, histograms of individual variables.By observing the scatter plots in the pair-plot, you can visually inspect which variables show the strongest linear relationship with the target variable (cnt).

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 a crucial step after building a model to ensure that the model is reliable and that the results are valid. The assumptions of linear regression are:

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

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)

Look at the coefficients: Features with larger absolute coefficients have a more significant impact on the target variable (bike demand).

Check p-values: Ensure that the p-values for these features are less than the significance level (commonly 0.05). If a feature has a high p-value, it is not statistically significant.

Rank the features: After considering both the coefficients and p-values, rank the features by the magnitude of their coefficients.

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 one of the simplest and most commonly used algorithms in statistical modeling and machine learning. It is used for predicting a continuous target variable based on one or more predictor variables (also known as independent variables or features).

In the context of linear regression, the assumption is that there is a linear relationship between the dependent variable (target) and the independent variables (predictors). This means that we aim to model the relationship as a straight line (or a hyperplane in the case of multiple predictors) that best fits the data.

1. Simple Linear Regression (One Predictor)

In the case of simple linear regression, the relationship between the dependent variable yyy and the independent variable xxx is modeled as:

y=β0+β1x+ϵy = \beta_0 + \beta_1 x + \epsilony=β0+β1x+ϵ

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 is a collection of four data sets that were created by the statistician **Francis Anscombe** in 1973 to demonstrate the importance of visualizing data and the potential pitfalls of relying solely on summary statistics, such as the mean, variance, or correlation. Despite the fact that the four data sets have identical summary statistics, their underlying distributions and relationships are very different, highlighting the importance of graphical analysis in understanding data.

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, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is widely used in data analysis to assess the degree to which two variables are linearly related.

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 refers to the process of transforming features of a dataset into a specific range or distribution. It is done to adjust the features so that they are on a comparable scale or scale of measurement. In datasets, features may have different units or ranges, and scaling ensures that no particular feature dominates or disproportionately influences the model's performance due to differences in scale.

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>

The Variance Inflation Factor (VIF) measures how much the variance of the estimated regression coefficients is inflated due to multicollinearity among the independent variables. It quantifies how strongly a variable is correlated with the other variables in the model.

A VIF value greater than 1 indicates that the variable's variance is inflated due to correlation with other predictor variables. If the VIF value is very high (usually above 5 or 10), it indicates a strong multicollinearity problem, meaning that the feature is highly correlated with other features, which can lead to unreliable estimates of the regression coefficients.

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 used to compare the distribution of a dataset with a theoretical distribution, such as the normal distribution. The plot is used to assess whether the data follows a particular distribution, typically the normal distribution, which is a key assumption in many statistical techniques, including linear regression.
