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)

I have observed the following points in categorical variables.

Yr:

- The count for the year 2018 is lower compared to 2019.
- The median value for 2018 is below the 2019 median.
- More values are spread in the lower quartile in 2018, but they are equally distributed in 2019.

Working day:

- The upper limit for both working days and non-working days is the same.
- The minimum count is higher on working days compared to non-working days.
- Values are distributed equally in the upper and lower quartile for both working days and non-working days.

Season:

- Spring season has a lower count.
- Summer and fall seasons have higher counts.
- Winter season has a higher count than spring season.
- Values are distributed equally in the upper and lower quartile during winter, while more values are distributed in the upper quartile during spring, summer, and fall.

Mnth:

- The count is high from April to October.
- The count is low in January, February, November, and December.
- Values are more distributed in the upper quartile across months.

Holidays:

- The upper limit for both working days and non-working days is the same.
- The minimum count is higher on working days compared to non-working days.
- Values are distributed equally in the upper and lower quartile for both working days and non-working days.

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)

When the parameter `drop = first` is set to true, the initial category of each categorical feature is excluded. This action decreases the number of dummy variables by one for each categorical feature, thereby preventing multicollinearity.

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

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)

The following steps are used to validate the training set:

- Check the R square value after removing each independent variable to observe the change and determine the dependency on specific variables.
- Evaluate the P value and ecoefficiency values.
- Calculate the VIF for independent variables removed from the model.

Assumptions:

- The combination of independent variables results in a high R square value.
- Multicollinearity is checked and addressed with selected independent variables.
- Verify that the VIF value is within the acceptable range.

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)

Year
Temperature
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 widely used machine learning algorithm. It predicts the relationship between dependent and independent variables, illustrating a linear connection between them.

There are two types of linear regression

1. Simple Linear Regression

   Simple linear regression uses one independent variable to predict a numeric dependent variable.

2. Multiple Linear Regression

   When multiple independent variables are employed to predict the value of a numeric

dependent variable, this method is referred to as multiple linear regression.

There are two categories of regression line relationships.

1. Positive linear relationship

A positive linear relationship occurs when the dependent variable increases as the independent variable increases.

2. Negative linear relationship

A negative linear relationship occurs when the dependent variable decreases as the independent variable increases. q

In linear regression, we aim to find the best fit line by minimizing the error between predicted and actual values. This line has the least error.

We need to follow the below steps to create linear regression model

1. Reading, understanding, and visualizing the data.
2. Preparing data for modeling, including train-test split and rescaling.
3. Training the model.
4. Conducting residual analysis.
5. Making predictions and evaluating on the test data.

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 comprises four datasets with identical descriptive statistics—means, variance, R squared, and linear regression lines—but they appear differently when plotted on a graph.

Each dataset has unique x-y relationships and variability patterns, yet they share the same summary statistics, including mean, variance, correlation coefficient, and linear regression line.

Anscombe's quartet shows why exploratory data analysis and data visualization are essential. It highlights the limitations of relying solely on summary statistics to identify trends, outliers, and other details.

Anscombe's quartet datasets reveal the following details:
Dataset 1: Linear x-y relation matching the regression line.
Dataset 2: Strong nonlinear relationship following a curve.
Dataset 3: Linear relationship with one influential outlier.
Dataset 4: Identical data points except for one non-conforming outlier.

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 correlation is a statistical tool used to assess the relationship between two quantitative (continuous) variables, assuming that the relationship is linear. This relationship can be visually depicted using a scatter plot, where a straight line represents the trend. The extent to which data points align with this line is measured by Pearson's correlation coefficient, denoted as r. This coefficient is also known as Pearson's product-moment correlation coefficient or simply the correlation coefficient.
Pearson's r functions both as a descriptive statistic, summarizing the strength and direction of the linear relationship, and as an inferential statistic. It can be tested for statistical significance, enabling researchers to draw inferences and make conclusions based on the data.
The calculation of Pearson's r allows for subsequent analysis:
Interpretation:

- $r > 0$ indicates a positive correlation, meaning an increase in one variable is associated with an increase in the other.
- $r < 0$ indicates a negative correlation, meaning an increase in one variable is associated with a decrease in the other.
- $r \approx 0$ suggests no linear correlation, implying that changes in one variable do not reliably predict changes in the other.

Strength of correlations:

- $|r| < 0.3$: Weak linear relationship
- $0.3 \leq |r| < 0.7$: Moderate linear relationship
- $0.7 \leq |r| < 1.0$: Strong 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)

<Your answer for Question 9 goes here>
Scaling is the process of adjusting the range and distribution of data values to facilitate easier comparison or effective use in various types of analysis. It is essential for ensuring that variables with different units of measurement or ranges are on a comparable scale. This is particularly crucial in machine learning, where variables may vary significantly in scale.

Normalized scaling, also known as Min-Max scaling:

- Transforms values to a range of 0 to 1
- Is used when data needs to be changed within a fixed range
- May not scale outliers correctly

Standardized scaling, also referred to as z-score standardization:

- Is more effective when data is normally distributed
- Transforms data to have a mean of 0 and a standard deviation of 1
- Facilitates easier comparison of variables with different units

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>

VIF (Variance Inflation Factor) is used to assess the degree of multicollinearity in a set of linear regression variables. It quantifies how much the variance of a regression coefficient is increased due to the correlation of that variable with other independent variables in the model. A high VIF indicates that a particular predictor is highly correlated with one or more of the other predictors, which can lead to instability in the estimated regression coefficients.

Perfect multicollinearity occurs when one independent variable is dependent on another independent variable through a perfect linear combination. When high VIF values are identified, it is necessary to check for correlated variables and remove them. This can be identified by performing pairwise correlation. The issue of multicollinearity can be resolved by removing one of the variables causing the multicollinearity.

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 assesses whether a dataset follows a specific theoretical distribution by comparing the quantiles of the observed data to those of another distribution.

- To create a Q-Q plot:
- Sort the data
- Calculate quantiles
- Calculate theoretical quantiles
- Plot the data

The following points can be understood from the plot:

- If the points lie approximately along a straight line, the data follows the theoretical distribution.
- Deviation from the straight line suggests departures from the theoretical distribution.