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)

There are multiple factors impacting the bike sharing. I can draw the below inferences from the dataset

1. There is a decrease in bike sharing with increased wind speed.
2. There is decrease in bike sharing if the weather condition is bad.
3. There is decrease in bike sharing on holidays.
4. Bike sharing significantly increases during the fall and sprint seasons.
5. Bike sharing significantly increased in Aug and Sept months.

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, helps to make sure 1) multi-collinearity is prevented 2) make sure the regression equation works correctly 3) Coefficient compared against the dropped category.

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)

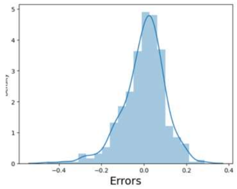
From the pair-plot it looks like the temp and atemp has highest correlation with the target variable.

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)

I conducted a residual analysis to validate the assumptions. When we plot it in a distplot we can see the mean is 0, that means it is 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)

The top three features contributing significantly towards explaining the demand of shared bikes are 1) Temperature 2) Holidays and 3) Season (fall)

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>

A regression model attempts to explain the relationship between a dependent and one or more independent variables. The independent variable is also known as the predictor variable. And the dependent variables are also known as the output variables.

We can represent a linear regression mathematically as
Linear regression
y = B0+B1x1+B2x2+...+BnXn

where,
y-> Dependent variable
x-> independent variable
n-> number of features
Bi - co-efficient values we have to calculate

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 set of four different datasets that have nearly identical statistical properties but appear very different when visualized. It was created by Francis Anscombe in 1973 to highlight the importance of data visualization in statistical analysis.

Each dataset consists of eleven(x,y) pairs and all share the same statistical properties i.e.
Maen of x: 9
Mean of y: 7.50
Variance of x: 11.0
Variance of y: ~4.12
Correlation: 0.816
Linear Regression line: y=3.00+0.50x

Visualization

1. Dataset1 shows a strong linear trend which supports linear regression

2. Dataset 2 x-values are identical to Dataset1 but y-values created curved pattern. This clearly exhibits a non-linear pattern.
3. Dataset 3 - Mostly linear except a single outlier that drastically affects regression line.
4. Dataset 4 – Almost all x-values are same except for one extreme point. The correlation is artificially high because of this single point.

Key Take aways

- Don't rely only on summary statistics
- Outliers and non-linearity impact model
- Correlation does not always imply relationship, data can be structured differently.
- Always check outliers and pattern using scatter plots before making conclusion.

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 (or Pearson correlation coefficient) is a measure of linear correlation between two variables. It quantifies how strongly two variables move together in a linear relationship. The value ranges from -1 to +1

+1 indicates a perfect positive linear relationship.
-1 indicates a perfect negative linear relationship.
0 indicates no linear relationship.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. It is widely used to assess the degree to which changes in one variable predict changes in another.

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 adjusts the range of numerical features in a dataset to ensure they contribute equally to a model. It's crucial for algorithms like gradient descent, k-nearest neighbors, and SVMs, which are sensitive to the magnitude of input features.

- Normalized Scaling (Min-Max Scaling) transforms data to a fixed range, typically [0, 1]. It's useful when you know the minimum and maximum bounds of the data.

- Standardized Scaling (Z-score Scaling) adjusts data to have a mean of 0 and a standard deviation of 1. It's ideal when the data follows a normal distribution or when different features have different units or scales.

Key Difference: Normalization scales data within a specific range, while standardization centers data around zero with unit variance.

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 = 1/ (1-Square of R)

If this results in a zero division, then value of VIF becomes infinite. This states that there is a multi collinearity in the regression model. That means one predictor variable is a perfect linear combination of one or more other predictors. In this case if the "Square of R" becomes 1 then that results in a infinite value forVIF.

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>

Q-Q or Quantile to Quantile plot compares the quantiles of dataset with quantiles of a theoretical distribution such as normal distribution. In linear regression it is used to assess if residuals are normally distributed, a key assumption for valid model inference. Points should align along a diagonal line if the residuals are normally distributed. Deviations from this line can indicate non-normality or outliers, suggesting the need for model adjustments or transformations to improve fit and validity. The Q-Q plot is crucial for ensuring that the assumptions of linear regression are met.
