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)

The categorical variables season, month (mnth), year (yr), weekday, working day, and weather situation (weathersit) have a significant impact on the dependent variable 'cnt' (bike rentals count). These factors contribute to fluctuations in rental demand based on various conditions:

- Season: Different seasons (spring, summer, autumn, winter) influence bike rental demand due to weather changes and seasonal activities.
- Month: Specific months can reflect patterns in rental behavior, possibly due to holidays, school vacations, or weather conditions.
- Year: Changes across years can capture evolving trends or external factors influencing bike rental patterns.
- Weekday: The day of the week affects rentals, with higher demand on workdays or weekends depending on user behavior.
- Working day: Whether it's a working day or holiday plays a role in determining rental activity, with holidays likely seeing lower demand.
- Weather situation: Weather conditions, such as sunny or rainy days, greatly affect whether people opt for bike rentals.

Together, these variables significantly explain variations in bike rentals, making them crucial in understanding and predicting rental patterns.

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 helps prevent multicollinearity by removing one of the categories. This ensures that the dummy variables are not highly correlated with each other, which can lead to model instability and inaccurate coefficient estimates in algorithms like linear regression. Dropping the first category removes redundancy and makes the model more interpretable.

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)

The 'temp' and 'atemp' has the highest 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)
The validation of linear regression models involves ensuring the following assumptions: a linear relationship between independent and dependent variables, no autocorrelation in the residuals, errors are normally distributed, residuals exhibit constant variance (homoscedasticity), and no multicollinearity among the predictors.

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)
Top 3 features that has significant impact towards explaining the demand of the shared bikes are season, temperature and year

---

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 predictive modeling technique that identifies the relationship between a dependent variable (target) and one or more independent variables (predictors). It models how the dependent variable changes in response to the independent variables. When there is only one independent variable, it is called simple linear regression, whereas multiple independent variables lead to multiple linear regression. The result is a straight line that best describes the relationship between the variables.
This relationship can be either a positive or negative linear trend, depending on how the variables are related. The objective of linear regression is to determine the optimal values for the parameters, often denoted as a0 and a1, that define the best-fitting line. The goal is to minimize the error between the observed data and the predicted values. Techniques like Recursive Feature Elimination (RFE) or the Mean Squared Error (MSE) function are used to determine the best values for these parameters, ensuring that the regression line fits the data as accurately as possible.

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 datasets that have nearly identical simple descriptive statistics yet have very different distributions and appear quite differently when graphed. The dataset was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before performing any analysis, especially when interpreting statistical measures.

The four datasets in Anscombe's quartet all have the same:

- Mean of the x-values (average of the independent variable),
- Mean of the y-values (average of the dependent variable),
- Variance of the x-values (spread of the independent variable),
- Variance of the y-values (spread of the dependent variable),
- Correlation between the x and y values.

However, the datasets exhibit very different relationships when plotted. Here's a breakdown of the four datasets:

1. Dataset I: This dataset shows a strong linear relationship between x and y, with a clear upward trend. A linear regression model would fit this data well, as the relationship is approximately linear.
2. Dataset II: This dataset also has a linear relationship between x and y, but the data points are more spread out compared to Dataset I. The relationship is still linear, but there is more variability, and the points deviate more from the line.
3. Dataset III: This dataset has a clear non-linear relationship between x and y. Although the correlation coefficient is still high, the relationship is curved, and a linear regression would not be an appropriate model to describe this data.
4. Dataset IV: This dataset contains an outlier in the form of a single extreme data point, which significantly affects the regression line. The other data points suggest a near-constant value of y, but the outlier creates a misleading relationship if a linear regression model is used.

Anscombe's quartet emphasizes the importance of data visualization. Despite having identical statistical summaries, the relationships between the variables in these datasets are very different. This highlights that relying solely on statistical measures like mean, variance, and correlation can be misleading without visualizing the data to understand its true nature.

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 evaluates the strength and direction of the linear relationship between two continuous variables. The value of Pearson's R ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship, meaning as one variable increases, the other also increases in a perfectly proportional manner.
- -1 indicates a perfect negative linear relationship, meaning as one variable increases, the other decreases in a perfectly proportional manner.
- 0 indicates no linear relationship between the variables, meaning changes in one variable do not predict changes in the other.

Values between 0 and 1 or between 0 and -1 indicate varying degrees of correlation. For example, a value of 0.8 indicates a strong positive correlation, whereas -0.5 indicates a moderate negative

correlation.

Pearson's R is calculated using the covariance of the two variables divided by the product of their standard deviations. It is useful for determining the degree to which two variables are linearly related, but it is not suitable for detecting non-linear relationships.

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 transforming data so that it fits within a specific range or scale. It is an essential pre-processing step that ensures consistent data input, improving the performance of algorithms. Raw data often contains features with different magnitudes, units, and ranges. If scaling is not applied, algorithms may give more weight to features with larger values, leading to biased or incorrect results.

Differences Between Normalization and Standardization:

1. Normalization uses the minimum and maximum values of the features, while Standardization uses the mean and standard deviation to scale the data.
2. Normalization is ideal when features have different scales, whereas Standardization is used to achieve a zero mean and a unit standard deviation.
3. Normalization scales data between a defined range (typically 0 to 1 or -1 to 1), while Standardization does not have a fixed range and can produce values outside of this range.
4. Normalization is sensitive to outliers, as extreme values can distort the scaling. In contrast, Standardization is not as affected by outliers.
5. Normalization is useful when the distribution of data is unknown, while Standardization works best when the data is approximately normally distributed.
6. Normalization is also known as Min-Max Scaling, while Standardization is often referred to as Z-Score Normalization.

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 a regression coefficient is inflated due to the correlation with other independent variables. It quantifies the degree of multicollinearity in a regression model. A VIF value greater than 10 is generally considered high, indicating a significant amount of multicollinearity.

- A VIF above 5 should also be investigated, as it may signal potential issues.

A very high VIF (close to infinity) suggests perfect multicollinearity, where two or more independent variables are highly correlated, resulting in an R2=1R^2 = 1R2=1. This leads to an inflated VIF value, making the model unstable. In such cases, it's important to remove one of the correlated variables to resolve the multicollinearity issue and improve the model's reliability.

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 two probability distributions by plotting their quantiles against each other. It helps assess whether a set of data could have come from a specific theoretical distribution such as Normal, Exponential, or Uniform distribution.

The Q-Q plot is used to check if two distributions are similar. When the distributions are closely aligned, the plot will show a linear trend. This linearity is essential for confirming if data follows the expected distribution. It can also be used to verify assumptions in linear regression, where variables need to follow a multivariate normal distribution. A histogram or Q-Q plot can test this assumption.

Importance of Q-Q Plot in Linear Regression:

- Normality check: Q-Q plots help determine if the residuals (errors) in a linear regression model are normally distributed. This assumption is crucial for reliable statistical inference, such as hypothesis testing and confidence intervals.
- Train and Test Set Validation: Q-Q plots can confirm if both training and testing datasets come from populations with the same distribution.

Advantages:

- Works with any sample size.
- Detects changes in location, scale, symmetry, and outliers in data.

Uses of Q-Q plot:

- Comparing distributions: It helps check if two datasets come from populations with similar distributions.
- Location and scale: Ensures the datasets have the same location (mean) and scale (variance).
- Distribution shape: Confirms if both datasets share the same shape of distribution.
-
