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 features, including season, month (mnth), year (yr), weekday, working day, and weather conditions (weathersit), have a substantial impact on the target variable 'cnt' (the rental count). These factors reflect various elements that cause fluctuations in bike rental demand.

- **Season**: The demand for bike rentals can be influenced by the season (spring, summer, autumn, winter), as each season brings different weather patterns, holidays, and levels of activity that can either encourage or limit bike usage.
- **Month**: The month variable can highlight seasonal trends, such as increased rentals during holidays, vacations, or when weather conditions change, which can all affect bike rental numbers.
- **Year**: The year variable allows us to track trends over time, capturing factors such as changes in infrastructure, marketing strategies, or evolving societal behaviors that influence bike rentals.
- **Weekday**: The demand for bike rentals can vary throughout the week based on people's work schedules or recreational activities, with more rentals typically occurring on weekdays versus weekends or holidays.
- **Working Day**: Whether a day is a regular working day or a holiday can play a significant role in rental patterns, with lower demand generally observed on holidays.
- **Weather Situation**: Weather conditions have a direct effect on bike rental behavior. Poor weather can lower demand, while favorable conditions can encourage more rentals.

These variables are essential for understanding rental demand and are vital for building accurate predictive models and conducting meaningful analysis.

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)

By using `drop_first=True` when generating dummy variables, we can prevent multicollinearity by removing one category from each categorical variable. This helps eliminate the issue of perfect linear relationships between variables. As a result, the model avoids redundant information, enhancing both its accuracy and interpretability. This is particularly important for algorithms like linear regression, which can be sensitive to 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)

The variables 'temp' and 'atemp' exhibit the strongest correlation 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)

Linear regression models are evaluated by verifying several key assumptions: a linear relationship between the predictors and the target variable, no autocorrelation in the residuals, errors that follow a normal distribution, constant variance of errors (homoscedasticity), and the absence of 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)

The top three features that significantly contribute to explaining the demand for shared bikes are temperature, year, and 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 statistical technique used to model the relationship between a dependent variable and one or more independent variables, assuming a linear connection between them.

The primary objective of linear regression is to identify the line that best fits the data, minimizing the difference between the observed and predicted values. This is typically done using the least squares method.

Linear regression can be either simple or multiple. Simple linear regression uses a single independent variable, while multiple linear regression involves several independent variables.

The performance of a linear regression model is evaluated using metrics such as R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

It’s important to remember that linear regression relies on the assumption of a linear relationship between the variables. Factors like outliers, multicollinearity, and other data issues can affect the accuracy and effectiveness of the model.

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 consists of four distinct datasets that have nearly identical summary statistics, but their graphical representations reveal very different relationships between the variables. Created by statistician Francis Anscombe in 1973, the quartet was designed to emphasize the importance of visualizing data before making conclusions based solely on statistical measures.

Each of the four datasets has the following characteristics:

- Identical mean values for both the independent variable (x) and dependent variable (y),
- Identical variance for both x and y,
- Identical correlation between x and y.

However, the distribution of the data points differs significantly when plotted:

1. **Dataset I**: Displays a strong linear relationship between x and y, with the points closely following a straight line. This dataset would be ideal for linear regression.
2. **Dataset II**: Similar to Dataset I but with more variability around the regression line. The relationship remains linear, but there is greater spread in the data.
3. **Dataset III**: Shows a non-linear relationship. Although there is a high correlation, a linear model would not fit the data well, as the relationship between x and y is curved.
4. **Dataset IV**: Contains an outlier that disrupts the linear relationship. While most of the data points suggest a horizontal line, the outlier skews the regression model.

The key lesson from Anscombe's quartet is that summary statistics alone cannot fully capture the characteristics of the data. Visualizing data is essential, as it can uncover patterns, outliers, and relationships that may not be apparent from statistical summaries alone.

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

- A value of **1** indicates a perfect positive linear relationship, meaning as one variable increases, the other also increases in a perfectly consistent manner.
- A value of **-1** indicates a perfect negative linear relationship, meaning as one variable increases, the other decreases in a perfectly consistent manner.
- A value of **0** means there is no linear relationship between the variables, indicating that changes in one variable do not predict changes in the other.

Values between 0 and 1 (or between 0 and -1) represent varying degrees of correlation. A coefficient close to **1** or **-1**suggests a strong linear relationship, while a coefficient near **0** points to a weak or absent linear relationship.

Pearson's R is computed by dividing the covariance of the two variables by the product of their standard deviations. It specifically measures linear relationships and does not account for non-linear associations between the variables.

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 transforming data to fit within a specific range, which is essential for enhancing the performance of machine learning algorithms. Raw data often contains features with different units, ranges, and magnitudes. If scaling is not applied, algorithms may give undue importance to features with larger values, leading to biased or incorrect outcomes.

Key Differences Between Normalization and Standardization:

1. **Normalization** scales data using the minimum and maximum values of the features, whereas **Standardization** uses the mean and standard deviation of the data.
2. **Normalization** is typically used when the features have varying scales, while **Standardization** is employed when you aim for a zero mean and unit variance.
3. **Normalization** scales the data to a fixed range, often between (0, 1) or (-1, 1), while **Standardization** does not limit the data to a specific range.
4. **Normalization** is sensitive to outliers, which can distort the results, while **Standardization** is more robust to outliers.
5. **Normalization** is useful when the data distribution is unknown, whereas **Standardization** works best for data that follows a normal distribution.
6. **Normalization** is also known as Min-Max Scaling, whereas **Standardization** is 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) is a statistical tool that measures how much the variance of a regression coefficient is inflated due to the correlation among independent variables. It is used to detect multicollinearity in regression models. A VIF greater than 10 indicates a high level of multicollinearity, suggesting potential issues with the model.

Even a VIF above 5 warrants closer scrutiny.

A very high VIF, approaching infinity, indicates near-perfect correlation between two or more predictors, which can cause instability in the model. To address this, removing one of the correlated variables can help reduce multicollinearity and enhance the model's performance.

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 quantiles of a dataset with the quantiles of a theoretical distribution, such as Normal, Exponential, or Uniform. It helps determine whether the data follows a particular distribution. The plot can also be used to compare two datasets to see if they have similar distributions.

When the two distributions being compared are similar, the points on the Q-Q plot will align closely along a straight line. If the points deviate from the line, it suggests a difference between the distributions.

Importance of Q-Q Plot in Linear Regression:

In linear regression, it is crucial to check the assumptions about the distribution of residuals. The Q-Q plot helps verify if the residuals follow a normal distribution, which is an essential assumption in regression analysis. It can also be used to compare the train and test datasets to ensure they come from populations with the same distribution.

Advantages:

- Can be used with sample data.
- Helps detect various distributional characteristics, such as shifts in location or scale, symmetry, and the presence of outliers.

Key Uses of a Q-Q Plot:

- Verifying if two datasets originate from populations with the same distribution.
- Comparing the location and scale of two datasets.
- Assessing if two datasets share a similar distribution shape.
- Analyzing the tail behaviors of distributions.
