Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

(3 marks)

Answer:
From the analysis, we found that categorical variables such as season, yr (year), and weathersit (weather situation) significantly influence the dependent variable (bike rental demand). Specifically, bike rentals were higher during warmer seasons (Summer and Fall compared to Spring), indicating a seasonal effect on demand. The year showed a positive effect, with an increase in rentals in 2019 compared to 2018, reflecting growing popularity or expanded service. Weather situations also had a clear impact, with clear to partly cloudy weather conditions favoring higher rental numbers, while adverse weather conditions (like light snow or heavy rain) significantly reduced demand.

2. Why is it important to use drop_first=True during dummy variable creation?

(2 mark)

Answer:
Using drop_first=True in dummy variable creation is important to avoid the "dummy variable trap," which refers to multicollinearity problems in a linear regression model. It ensures that one category is dropped and not encoded, which helps in reducing the number of features and avoiding redundancy since the dropped category can be inferred from the others, ensuring the model's coefficients remain interpretable and the model itself remains efficient.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

(1 mark)

Answer:
Based on the analysis, temp (temperature) showed the highest correlation with the target variable (cnt), indicating that warmer temperatures are associated with increased bike rental demand. This was observed through both visual analysis and the model's coefficients.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

(3 marks)

Answer:

The assumptions were validated by analyzing residuals, which involved:

Homoscedasticity: Checked by plotting residuals vs. fitted values. A pattern less distribution suggests constant variance.

Normality of Residuals: Assessed using a histogram and Q-Q plot of residuals. A bell-shaped histogram and residuals closely following the line in the Q-Q plot indicate normality.

Linearity: Assumed based on the linear formulation of predictors in the model and can be visually supported by scatter plots of individual predictors against the target.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

(2 marks)

Answer:

Based on the final model, the top 3 features contributing significantly to explaining bike rental demand were yr (indicating yearly growth in rentals), temp (temperature), and weather conditions (weathersit_3, representing adverse weather like light snow/rain).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

(4 marks)

Answer:

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The equation of a linear regression line is y = β0 + β1X1 + ... + βnXn, where y is the dependent variable, X1 to Xn are independent variables, and β0 to βn are the coefficients that linear regression tries to estimate. The goal is to find the line that best fits the data, typically using the least squares method, which minimizes the sum of the squared differences between observed and predicted values.

2. Explain the Anscombe’s quartet in detail.

(3 marks)

Answer:

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset shows a different relationship between the variables, highlighting the importance of visualizing data before analyzing it and the limitations of relying solely on basic statistics for data analysis.

3. What is Pearson’s R?

(3 marks)

Answer:

Pearson's R, or Pearson correlation coefficient, measures the linear correlation between two variables, giving a value between -1 and 1. A value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear correlation. It's used to quantify the degree of linear relationship between variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

(3 marks)

Answer:

Scaling is the process of normalizing the range of features in a dataset. It's important for algorithms that compute distances between data, as it ensures that features contribute equally to the result. Normalized scaling brings values between 0 and 1, while standardized scaling transforms data to have a mean of 0 and a standard deviation of 1, making it easier to compare coefficients as predictors on the same scale.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)

Answer:

An infinite VIF (Variance Inflation Factor) occurs when there's perfect multicollinearity among the predictors, meaning one variable can be perfectly predicted from the others. This usually happens when there's an exact linear relationship between some of the independent variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q

(3 marks)

Answer:

A Q-Q (Quantile-Quantile) plot is a graphical tool to assess if a dataset follows a certain distribution, typically the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution. The closer the points lie to the reference line, the more likely it is that the data follows the distribution. It's useful for checking the normality assumption in linear regression models.