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?

Answer:

1. **Year:**
    - Higher count of total bike rentals (cnt) in 2019.
2. **Holiday:**
    - Lower count of total bike rentals (cnt) on holidays.
3. **Working Day:**
    - No significant effect on total bike rentals.
4. **Season:**
    - Highest cnt in fall, followed by summer and winter.
    - Lowest in spring.
5. **Month:**
    - Higher cnt from June to September.
    - Lower from January to March.
6. **Weekday:**
    - Higher cnt on Sundays and Thursdays.
    - Lower on Mondays and Tuesdays.
7. **Weathersit_relabelled:**
    - Higher cnt on clear days and days with few clouds.
    - Followed by misty or broken cloud days.
    - Lower cnt on days with light rain, thunderstorms, and scattered clouds.

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

Answer:

When creating dummy variables, using `drop_first=True` helps avoid the creation of unnecessary independent variables. This practice prevents an unnecessary increase in the number of predictors in a prediction model, keeping the model simpler and more efficient.

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

Answer: temp

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

Answer:

Linearity: Validate by plotting a scatter plot between X and Y. A clear positive relationship without turning backward supports the assumption of a linear relationship.

Normality of Residuals: Validate by plotting a distplot for the residuals. A distribution resembling a normal or Gaussian distribution (bell-shaped curve) indicates that the assumption of normally distributed errors is met.

Homoscedasticity: Validate by plotting a scatter plot of residuals vs. predicted values (y hat). If the scatter plot shows no discernible trend and has constant variance, then the assumption of homoscedasticity is validated.

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

Answer: year, temp, weathersit

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

- **Simple Linear Regression:**
  - Models' relationship between one independent (X) and one dependent variable (Y).
  - Equation: (Y = b_0 + b_1. X + ε).
- **Multiple Linear Regression:**
  - Includes multiple independent variables.
  - Equation: (Y = b_0 + b_1 . X_1 + b_2 . X_2 + ... + b_n . X_n + ε).
- **Objective:**
  - Minimize sum of squared differences between observed and predicted values.
- **Training:**
  - Estimate coefficients ((b_0, b_1, ..., b_n)) using methods like OLS or gradient descent.
- **Key Concepts:**
  - Cost function, gradient descent, assumptions, residuals.
- **Steps:**
  - Data collection, preprocessing, training, evaluation, prediction.
- **Assumptions:**
  - Linear relationship, independence of errors, homoscedasticity, normality of errors.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet, comprising four datasets, highlights the importance of visualizing data before applying models. This practice aids in detecting anomalies such as outliers, diversity, and linearity. Linear regression suits linearly related data, emphasizing the need to assess data distribution. Scatter plots of these datasets showcase diverse patterns, underlining the challenge of interpreting them with regression algorithms. Visual inspection is crucial for understanding data characteristics prior to model building.

So Anscombe's quartet is quite useful to understand the data visualization. So before attempting to interpret and model the data or any algorithms. We must first visualize the data set to build a good fit model.

3. What is Pearson’s R?

Answer: The Pearson correlation coefficient also referred as Pearson’s R or bivariate correlation. It is a measure of linear correlation between two data sets. It is a numerical summary of the strength of the linear association between two variables. If its go up and down together its called positive correlation and if it go up and down opposite sites called negative correlation.

The Pearson’s correlation coefficient varies between – 1 and +1 .

Pearson r formula =

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where r = correlation coefficient

xi = value of x variables

xbar = mean of x variables value

yi = value of y variables

ybar = mean of y variables value

What we can understand from r values

r = 1 (Linearly associated with high positive correlation)

r = -1 (Linearly associated with high negative correlation)

r = 0 (No Linearly associated)

r >0>5 = (Weakly Linearly associated)

r >0 5 < 8 = (Moderate Linearly associated)

r > 8 = (Strong Linearly associated)

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

Answer:

- **Scaling in Data Pre-processing:**
  - Normalizes data to a specific range, overcoming issues with high and low values in different units.
- **Importance of Scaling:**
  - Ensures algorithms interpret values uniformly, avoiding misleading results due to varied magnitudes.
- **Normalization Effects:**

    - Impacts coefficients but not other result parameters (R-Square, Adj-R-Square, T-statistic, F-statistics, P-values).
- **Min-Max Scaling (Normalized Scaling):**
    - Uses minimum and maximum values for scaling.
    - Scale values between (0 and 1) or (-1, 1).
    - Highly affected by outliers.
    - Utilizes MinMaxScaler from sklearn.
- **Standardized Scaling:**
    - Uses mean and standard deviation for scaling.
    - Aims for zero mean and 1 standard deviation.
    - Less affected by outliers.
    - Utilizes StandardScaler from sklearn.

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

Answer: VIF provides a measure of multicollinearity among independent variables. If VIF is infinite, it signifies perfect correlation between two independent variables, indicating a high level of multicollinearity. This suggests that both variables have an identical impact, allowing for the consideration of dropping one without significantly affecting the results. An infinite VIF specifically indicates that one variable is precisely expressed as a linear combination of others, emphasizing the redundancy and potential for simplification in the model.

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

Answer: A Quantile-Quantile (Q-Q) plot is a graphical tool used in linear regression to assess whether a dataset follows a theoretical distribution, typically the normal distribution. It compares the quantiles of the observed data with the quantiles of a theoretical distribution. If points on the Q-Q plot fall along a straight line, it suggests that the data follows the expected distribution. Deviations from linearity indicate departures from the assumed distribution. In linear regression, Q-Q plots are vital for diagnosing normality assumptions of residuals, helping identify potential issues like skewness or heavy-tailed distributions, which can impact the reliability of statistical inferences.