Subjective Questions

1. **Inferences about the effect of categorical variables on the dependent variable (bike rentals):**

    - **Season:** Seasonality plays a significant role, with the highest rentals in season 4 (winter) and the lowest in season 1 (spring).
    - **Holiday:** Non-holidays see higher rentals compared to holidays.
    - **Weekday:** Day of the week shows variability in bike rentals, with weekends (especially Saturday) seeing higher rentals compared to weekdays.
    - **Workingday:** Working days have slightly higher rentals compared to non-working days.
    - **Weathersit:** Favorable weather conditions (weathersit 1) significantly increase rentals, while adverse conditions (weathersit 3) drastically decrease rentals.

2. **Importance of using `drop_first=True` during dummy variable creation:**

    - Using `drop_first=True` in `OneHotEncoder` helps in avoiding multicollinearity issues in the dataset. When creating dummy variables for categorical features, if all categories are included, one category becomes a linear combination of the others, causing multicollinearity. By dropping the first category, each category becomes independent, which is important for the model's interpretability and performance.

3. **Highest correlation with the target variable based on the pair-plot among numerical variables:**

    - Based on the pair-plot, the variable 'temp' (temperature) likely has the highest positive correlation with the target variable (bike rentals). This is because higher temperatures are associated with increased bike rentals, as indicated by the strong positive impact mentioned in the key inferences.

4. **Validation of assumptions of Linear Regression after building the model on the training set:**

    - Assumptions of Linear Regression include linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of errors. These assumptions can be validated using diagnostic plots such as residual plots, Q-Q plots, and scatterplots of predicted vs. actual values. Statistical tests like the Durbin-Watson test for autocorrelation can also be used.

5. **Top 3 features contributing significantly towards explaining the demand for shared bikes based on the final model:**

    - Based on the coefficients in the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are likely:

Subjective Questions

1. Temperature (temp)
2. Apparent Temperature (atemp)
3. Weather Situation (weathersit)

Subjective Questions

Linear Regression Algorithm:

- **Overview:** Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent variables (features) and the dependent variable (target).
- **Mathematical Formulation:** In its simplest form, the linear regression model can be represented as: $y=\beta0+\beta1x1+\beta2x2+...+\beta nxn+\epsilon$ y=β0+β1x1+β2x2+...+βnxn+ϵ where:
  - y y is the dependent variable,
  - $x1,x2,...,xn$ x1,x2,...,xn are the independent variables,
  - $\beta0,\beta1,\beta2,...,\beta n$ β0,β1,β2,...,βn are the coefficients,
  - ϵ ϵ is the error term.
- **Objective:** The objective of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the sum of the squared differences between the actual and predicted values.
- **Assumptions:** Linear regression assumes that the relationship between the variables is linear, the errors are normally distributed, and there is no multicollinearity among the independent variables.
- **Prediction:** Once the model is trained, it can be used to predict the dependent variable's values based on new independent variable values.

2. **Anscombe’s Quartet:**

- **Overview:** Anscombe’s quartet is a group of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.), yet appear very different when graphed. It demonstrates the importance of graphing data before analyzing it and highlights the limitations of summary statistics.
- **Purpose:** Anscombe's quartet is often used to illustrate the effect of outliers and the influence of individual data points on statistical properties like correlation and regression lines.
- **Implications:** It emphasizes the importance of visualizing data and not relying solely on summary statistics, as datasets with different underlying patterns can have similar summary statistics.

3. **Pearson’s R:**

- **Definition:** Pearson’s correlation coefficient (Pearson’s R) is a measure of the linear relationship between two continuous variables. It ranges from -1 to 1, where:
  - 1 1 indicates a perfect positive linear relationship,
  - -1 −1 indicates a perfect negative linear relationship, and

Subjective Questions

    - 00 indicates no linear relationship.
- **Formula:** Pearson’s R is calculated as the covariance of the two variables divided by the product of their standard deviations.

4. **Scaling:**

- **Definition:** Scaling is the process of transforming data to a standard scale. It is performed to ensure that the variables are comparable and do not bias the analysis.
- **Importance:** Scaling is important because many machine learning algorithms, such as SVM, KNN, and neural networks, are sensitive to the scale of the input variables. Scaling helps to improve the performance and stability of these algorithms.
- **Normalized Scaling vs. Standardized Scaling:**
    - **Normalized Scaling:** Normalization scales the data to a range of [0, 1] and is useful when the distribution of the data does not follow a Gaussian distribution.
    - **Standardized Scaling:** Standardization scales the data so that it has a mean of 0 and a standard deviation of 1. It is useful when the data follows a Gaussian distribution and is also important for algorithms that assume normally distributed data, such as linear regression.

5. **Infinite VIF Values:**

- **Reason:** The Variance Inflation Factor (VIF) measures the multicollinearity in a regression model. If the VIF for a variable is infinite, it indicates perfect multicollinearity with other variables, meaning that the variable can be perfectly predicted by a linear combination of other variables.
- **Implications:** Infinite VIF values can cause issues in the regression model, such as unstable coefficient estimates and inflated standard errors.

6. **Q-Q Plot (Quantile-Quantile Plot):**

- **Definition:** A Q-Q plot is a graphical technique used to compare the distribution of a dataset to a theoretical distribution, such as a normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution.
- **Use in Linear Regression:** In linear regression, Q-Q plots are used to check the assumption of normality of residuals. If the residuals (the differences between observed and predicted values) are normally distributed, the points on the Q-Q plot will fall approximately along a straight line. Departures from the straight line indicate departures from normality.
- **Importance:** Q-Q plots help to visually assess the goodness-of-fit of a model and identify any deviations from the assumptions of linear regression, such as non-normality of residuals.

Subjective Questions