Assignment-based Subjective Questions

1. Visualization with seaborn pairplot to understand the pattern of each feature, that understand strong association with the dependent variable based on visual patterns in the pairplot. Also I'm use the sklearn module RFE to check the ranking of variables that understand important in predicting the dependent variable based on RFE ranking.
2. To avoid multicollinearity of features, using "drop_first=True" ensures consistency with this approach.
3. atemp
4. To validate the model, I have use R2 score from skearn to evaluate the performance of this regression model. In this assignment, I have using seaborn.histplot module to plot a histogram of the errors. And find those errors are normally distributed. And I have using a scatter plot of actual vs. predicted values is a visual quickly gauge the performance and validity of the model.
5. Base on RFE module, atemp, weathersit and holiday are most important top 3 features

General Subjective Questions

1. Linear regression models the linear relationship between a continuous target y and one or more predictor features X. It fits a straight line through the data points that minimizes the sum of squared residuals (differences between observed and predicted y). Linear regression estimates the slope coefficients that define the. By examining the bis, we can assess the direction, strength, and significance of the linear association between each X and y. Once fit, the model can predict future values of y based on X. Overall, linear regression is a fundamental tool for modeling and predicting linear relationships.

2. Anscombe's quartet demonstrates that simply looking at summary statistics that regression can be misleading. This dataset contains four very different sets of x and y datas. Yet they have nearly identical descriptive statistics - the same mean and variance of x and y, the same regression line, and correlation coefficient. Actual relationships between x and y are quite different in each set so illustrates the importance of visualizing data before analysis, not relying solely on summary statistics to characterize relationships.

3. Pearson's R is measures the strength and direction of the linear relationship between two continuous variables. It quantifies how well a straight line can describe the association between two variables. Pearson's R ranges from -1 to 1. A value of 0 means no linear correlation, while -1/+1 indicates a perfect negative/positive linear relationship. The sign reflects the direction of association, while the absolute value indicates the strength. Pearson's R is widely used in statistics to summarize linear dependence.

4. When a dataset has features measured at very different scales, rescaling can be useful before analysis. Normalized scaling bounds all features to the same fixed range, like 0 to 1, compressing different scales into a common range. This removes magnitude differences but also distributional aspects. Standardized scaling rescales features to have a mean of 0 and standard deviation of 1, retaining differences in distributions and variation while removing scale differences. Standardization is preferred when distributional aspects are informative, while normalization is useful when only the relationships between feature values matter.

5. An infinite VIF arises when a predictor is perfectly collinear with other predictors which meant those features are perfect collinearity.

6. Q-Q plot is a graphical tool used to assess the validity of distributional assumptions in linear regression. It plots the quantiles of the regression errors against the theoretic quantiles from a normal distribution. If the errors are normally distributed as required for linear regression, the points will approximately lie on the reference line which is a useful diagnostic for assessing a key linear regression assumption.