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?**
   - 'season': The majority of bike sales, accounting for approximately 32%, occur during season 3.
   - 'mnth': The sales exhibit a growth pattern starting in May and maintaining stability until August, after which they gradually decline.
   - 'weathersit': The weather condition described as 'Clear, Few clouds, Partly cloudy, Partly cloudy' accounts for 68.61% of the bookings.
   - 'holiday': A significant portion, approximately 97%, of the bookings took place on non-holiday dates, which introduces data skewness.
   - 'weekday': Bike sales are evenly distributed across the weekdays, although there is a slight increase in sales on the fifth day of the week.
   - 'workingday': Approximately 70% of the sales occurred on working days, indicating a positive influence on the independent variable.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   - The "Drop first" technique is employed to prevent the dummy variable trap. It involves excluding one of the categorical variables created to avoid multicollinearity, which can potentially undermine the accuracy and efficiency of our model.

3. **Looking at the pairplot among the numerical variables, which one has the highest correlation with the target variable?**
   - cnt is found linearly related to atemp and temp column fields

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    - Assessing the normality of residuals.
    - Examining linearity, which was accomplished through the use of scatter plots or residual plots.
    - Investigating potential multicollinearity among the independent variables, utilizing variance inflation factor (VIF) scores.
    - Checking for homoscedasticity using The White Test.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
    - Based on the final model, the top three features, ranked in order of importance, are as follows:
        - Temperature (temp): With a coefficient of 0.546436, an increase of one unit in temperature leads to a corresponding increase of 0.546436 units in the dependent variable.
        - Year (yr): The coefficient for this variable is 0.233233, indicating that a one-unit increase in the year results in a corresponding increase of 0.233233 units in the dependent variable (unit sales of bikes).
        - Season 4 (season_4): With a coefficient of 0.131612, the dependent variable experiences a 0.131612 unit increase during season 4 compared to other seasons, assuming all other factors remain constant.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   - The linear regression algorithm is a statistical approach used to model the relationship between two variables. It entails determining the line of best fit by analyzing a scatter plot of data points. By calculating the slope and y-intercept of this line, we can make predictions about the dependent variable based on the independent variable's value. The line of best fit is chosen to minimize the sum of squared differences between the actual data points and the predicted values on the line. This optimization process is typically accomplished using the method of least squares, which involves finding the line's parameters that minimize the sum of the squared residuals, representing the differences between each data point and its corresponding point on the line.

2. **Explain the Anscombe's quartet in detail.**
   - Anscombe's quartet is a set of 4 datasets which have identical statistical properties, yet look very different when graphed. Each dataset has the same mean, standard deviation, correlation coefficient, and regression line, but they differ widely in the way that the individual data points are distributed.
     This is meant to illustrate the danger of relying solely on numerical statistics, without also considering the underlying graphical representation of the data.

3. **What is Pearson's R?**
   - Pearson's R is a measure of the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where -1 represents a perfect negative correlation, 1 represents a perfect positive correlation, and 0 represents no correlation.
   - This measure is commonly used in statistics to analyze the relationship between two variables, and it can be calculated using the formula R = (n * sum(x*y) - sum(x) * sum(y)) / sqrt((n*sum(x^2) - sum(x)^2) * (n*sum(y^2) - sum(y)^2))

where n is the number of observations, x and y are the variables of interest, and sum and sqrt represent summation and square root, respectively.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
    - Scaling is the process of transforming the values of a dataset so that they fall within a certain range or have a certain distribution. Scaling is often performed to help standardize the data and avoid issues with numerical overflow or underflow. Normalized scaling - involves

- transforming the data so that the values fall between 0 and 1. In contrast, standardized scaling involves transforming the data so that the values have a mean of 0 and a standard deviation of 1.
- Normalized scaling is useful when the distribution of the data is not Gaussian, since it preserves the relative ordering of the original values. Standardized scaling is useful when the data has a Gaussian distribution, since it converts the values to z-scores, allowing for easier comparison across different datasets.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
    - When the VIF is infinite, it usually means that there is a perfect linear relationship among some of the predictor variables, which makes it impossible to estimate the regression coefficients. This can happen, for example, when two variables are defined by exactly the same data, or when one variable is a linear combination of other variables. In this case, we observe VIF as ‘inf’ for categories.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
    - It is a graphical technique for comparing the distribution of a dataset to a theoretical distribution, such as the normal distribution.
    - Q-Q plots or quantile-quantile plots are useful in linear regression because they allow us to check the assumptions of normality and constant variance for the residuals of the model. If the residuals are normally distributed and have constant variance, the points on the Q-Q plot should fall close to the straight line. If the residuals deviate from these assumptions, it may indicate that the model is not an appropriate fit for the data.