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 sales are happening around season 3, around 32%.
    - 'mnth' : Sales embark on a growth trajectory from May that stays put until Aug till it starts slowly declining hence after.
    - 'weathersit' : 68.61% of bookings happened in the 'Clear, Few clouds, Partly cloudy, Partly cloudy' weather.
    - 'holiday' : 97% of bookings happened when there wasn't a holiday, this clearly introduces skew in the data.
    - 'weekday' : The sales are evenly distributed among the weekdays, however, with a slight uptick on the 5th day of the week.
    - 'workingday' : 70% of sales happened on a working day. Which denotes a good influence on the independent variable.

2. **Why is it important to use drop_first=True during dummy variable creation?**
    - Drop first is a technique to avoid the dummy variable trap, where we can drop one of the categorical variables created to avoid multicollinearity between the variables which may hinder 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?**
    - checking normality of residuals.
    - checking linearity, which can be done with a scatter plot or residuals plot.
    - checking for multicollinearity among independent variables, which was done with 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 3 features (in order of importance) are**:**
        i. **temp:** The coefficient for this variable is 0.546436. This means that as temperature increases by one unit, the dependent variable increases by 0.546436 units.

ii. **yr**: The coefficient for this variable is 0.233233. This means that as the year increases by one unit, the dependent variable (unit sales of bike) increases by 0.233233 units.

iii. **season_4**: The coefficient for this variable is 0.131612. This means that in season 4, the dependent variable increases by 0.131612 units compared to other seasons, all else equal.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
    - The linear regression algorithm is a statistical method for modelling the relationship between two variables. It involves finding the line of best fit through a scatter plot of data points. The algorithm calculates the slope and y-intercept of this line, which can then be used to predict the value of the dependent variable based on the value of the independent variable. the line of best fit is chosen so as to minimize the sum of the squared differences between the actual data points and the predicted values on the line. This calculation is usually done using the method of least squares, which involves finding the parameters of the line that minimize the sum of the squares of the residuals (i.e., the difference between each data point and the 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.