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?
    a. 2019 year essentially was a better year for bikes
    b. Fall followed by Summer were most favorable seasons for riders
    c. Clear weather is preferred by riders; Snow is worst weather for ride rentals
    d. Weekends (Sat, Sun) are not so good for business
    e. Weekdays and working days are favorable
    f. OVERALL – People seem to prefer rental bikes on working day when weather is clear and temperature is good

2. Why is it important to use **drop_first=True** during dummy variable creation?
    a. It helps reduce the number of variables and hence VIF
    b. it reduces the correlations created among dummy variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    a. It is evident from pair-plot Temperature has the highest correlation
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

    There are 4 assumptions we make for linear regression:

    1. Linear relationship between dependent and independent variables:
        a. Scatter plot of each variable as well as the final model should be linear
    2. Residuals should be normally distributed:
        a. The error residuals when plotted should show a normal distribution with mean on 0
    3. Error with constant variance:
        a. Scatter plot between fitted variable and errors/residuals should be varied similarly across the regression line
    4. Residuals should be independent:
        a. Autocorrelation: VIF is low at the end of the model for all important variables confirms that the independent variables are indeed independent

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

    Three most important factors based on Analysis are:

    1. Temprature
    2. 2019
    3. Season/Winter

General Subjective Questions

1. Explain the linear regression algorithm in detail
    a. Linear regression assumes that independent variables have a linear relationship with dependent variable
    b. We first identify linearly aligned and impacting variables
    c. Then we create a model/line equation that explains most of the variance in dataset and hence has the least errors
    d. Use this line equation to explain the dependent variable

2. Explain the Anscombe's quartet in detail.
    a. Anscombe's quartet are the 4 sets of 11 values nearly identical in simple descriptive statistics
    b. However, when plotted on scatterplot, these have different shapes and equations
    c. This is example used to explain the importance of both visuals and mathematics in deciding dependencies among variables

3. What is Pearson's R?
    a. It is correlation coefficient
    b. Describes relation between variables
    c. Varies between -1 and 1 (Negative to positive correlation)
    d. R signifies correlation only and not causation

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    a. Scaling is performed in order to bring all the numeric variables to the same range
    b. This enables variables of very different ranges to be compared efficiently without it's coefficient getting compromised
    c. E.g. House Sales data – number of bedrooms may range from 0-5/6 while carpet area might range from 100 to 2000. Even if these two have equal impact on price of house, the coefficient of carpet area will be lot smaller compared to that of number of bedrooms in order to show equal impact.
    d. Hence, if we scale both to same level – This issue will be resolved.
    e. Normalized Scaling:
        i. Min max scaling takes max value and the min values and compresses all values between 0 and 1 as new min and max
    f. Standardized Scaling:
        i. Scales based on z value with mean at 0 and spread across 1 standard deviation

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    a. If the VIF is infinity, that means VIF = 1/(1-R2) = infinity
    b. It is equivalent to saying R2 is 1
    c. Meaning- there is perfect correlation with at-least 1 other variables

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
    a. It is Quantile Vs Quantile comparison
    b. If two quantiles (on x and y) show roughly straight line, that means they are from the same distribution