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?
    - Bikes usage is more in Fall season
    - Bikes usage is lower in Spring season
    - Bikes usages is more when you have Clear Clouds
    - Bikes usages is less when you have Light Snow
    - Bikes usage increased from 2018 to 2019
2. Why is it important to use drop_first=True during dummy variable creation?
    - drop_first=True should be used to reduce the creation of more columns so that it will reduce the correlation between the variables.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    - 2019 (or the yr column)
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    - Residual Analysis has been done which resulted in normal distribution
    - Verified the VIF and p-Values of the variables
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    - Temp (Temperature)
    - Season
    - WorkingDay

General Subjective Questions

1. Explain the linear regression algorithm in detail.
    - Data Cleaning -
        - Drop rows/columns which are not required post analysis.
        - Data substitution can be done for the columns with a lesser number of missing values.
        - Data Extraction by creating derived variables when required.
        - These can be performed by NumPy and Pandas libraries
    - EDA - Exploratory Data Analysis
        - Identify the impact of the variables (predictors) on target variable
        - Generate the charts and graphs with the data available to generate insights.
        - Create a correlation heatmap between the predictor variables.
        - This can be done using matplotlib.pyplot and seaborn libraries
    - Handling Categorical Variables
        - As we know that Model building needs numerical data. We need to convert the categorical variables to dummy variables
        - This can be done by pd.get_dummies() function
    - Split Train and Test Data
        - Now split the data by 70% - Train Data, 30%- Test Data
    - Once you have Train data :
        - Scale the data using MinMaxScaler to fit and transform data between 0 to 1
    - Model Building
        - You can choose any one of the options
            - Progressive way of including one variable at a time to learn the performance of the model using p-values and VIF. This will be a tedious task
            - Bottom-up way where include all the variables in the first model and remove one by one to learn the performance of the model. This is much better than the previous one.
            - Else we can use the Automated RFE options to finalise the variable list automatically
    - Residual Analysis
        - Predict y_train_pred using final model fixed with x_train data
        - Generate the residuals using y_train-y_train_pred and plot using `sns.distplot(res)`
        - Which should show the normal distribution
    - Finally Evaluation of Model using Test Data
        - Here we only transform and not fit the data using the MinMaxScaler
        - Generate the r2_score using the test predicted values
2. Explain the Anscombe’s quartet in detail - (not well versed on this topic)
3. What is Pearson’s R?
    a. This is called Pearson's correlation coefficient.
    b. This is used for measure the correlation between two continuous variables

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    a. Scaling is a step performed on the predictors to normalize the data between a range.
    b. This increases the speed of the calculations performed while generating the statistics summary of the model.
    c. Normalizing means bringing the continuous variable value between 0 to 1. Where as standardized means replace the values with Z scores.
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    a. This might happen when R2=1 as VIF=1/(1-R2)
    b. So this says that both variables are perfectly correlated
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(not read this concept)