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?
   - Following are the categorical variables from the dataset and their influence are as follows:
     - Season
       - Demand for the bike is greater in the seasons of Fall, Summer and Winter. However during the spring season the demand for the bike is less.
     - Month
       - Month data also complement the season data. Months corresponding the Fall, summer and winters like from April to November there is a significance high usage of the bike and it is peaking in the month of September
     - Weekday
       - Weekday has statistically very less significance on the overall count which is a good thing as the demand is distributed over the entire week
     - Weather situation
       - Weather situation plays a crucial role, usage is high when the weather is clear or partially cloud. However during light rain or heavy rain usage drops significantly
     - Year
       - 2019 has higher usage compared to 2018 which signifies the business is growing
     - Working day
       - Working day has very less impact. This also complement weekday. So usage is constant over the period of time irrespective of whether it is weekday or weekend.

2. Why is it important to use drop_first=True during dummy variable creation?
   - Drop_First ensures that only limited set of columns is created. For example, if I have column with data 1 and 0. I do not need 2 columns to store the data. I can have only 1 column which stores the value as either 1 or 0.
   - So representing data with limited set of columns will create a compact data set and have less columns to be processed by the algorithm.
   - If I create two columns one for 1 and another for 0 then data set will be having multiple Null values. for example for all the records which has 1 the 0 column will become null and this is difficult to handle for further analysis.
   - In general if I have n values, I need to create n-1 value.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   - Temperature field has highest correlation with target variable.
   - Also to be noted here that casual and registered user have better correlation to target variable, but it cannot be considered as the casual and registered user are parts of “Cnt” which is our target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   - By comparing the R squared value of both training and testing data set.
     - R-squared value for training data set – 84.9%
     - R-Squared value for test data set is 85.4%
   - This indicates that the assumptions which we made during the model training is not overfitting or memorising and it is a good fit.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   - Temperature, 2019, Winter as these are having the high coefficient

General Subjective Questions

1. Explain the linear regression algorithm in detail?
   - Linear regression is a supervised learning algorithm
   - This is used to model relationship between:
     - Dependant variable
       - Target variable. This is the value that will be predicted
     - Independent variable
       - Predictors which are used to determine the outcome of dependant variable
   - Linear regression is represented with formula $y = B0 + B1Xi +ei$

        - Where B1 is called the slope
        - B0 is called the intercept
    - There are two types of linear regression
        - Simple linear regression
            - One predictor is used to determine the target
        - Multiple linear regression
            - Multiple predictor is used to determine the target.
            - There is a problem of multicollinearity

2. Explain the Anscombe's quartet in detail.
    - Anscombe quartet is a very famous dataset created by statistician Franscis Anscombe.
    - It explains the significance of data visualization in statistics.
    - It consists of 4 quartet and each containing 11 data points.
    - Even though data looks different they share summary statistics.
    - In his example, he had 4 dataset which had nearly identical summary statistics like Mean, Variance, Correlation and R-squared . However, visualization of data showed different result the underlying relationships, pattern and even outliers are anomalies started appearing in the visualization. So this quartet teaches us significance of using visualization before doing any further analysis on the same.

3. What is Pearson's R?
    - It is a statistic co-efficient that quantify the strength and direction of linear relationship between two continuos variable.
    - Its value ranges from -1 to +1
        - If the value is -1 then it means negative linear correlation.
        - If the value is 0 then it means there is no linear correlation.
        - If the value is +1 then it means there is a positive linear correlation.
    - For a good linear regression model the value for the Pearson R shouls be higher and closer to 1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    - Scaling is performed to bring the value to single scale.
    - For example, if I am analyzing the house rent data and I have two independent variables like carpet area and number of bedrooms. In general carpet areas will be a 3 digit or 4 digit number like 1200 or 870. However, the number of bedrooms will generally be a single digit number like 3 or 4. When we calculate the co-efficient of this fields the co-efficient values will have a huge difference and this may impact the outcome. So we generally bring everything to a similar scaler to ensure co-efficients are in same scale.
    - Normalized scaling
        - It generally scales the data between 0 and 1
        - It is also called as Min-Max scaling.
        - Formula is : (x – xmin)/(xmax – xmin)
        - They are good to handle outliers.
    - Standardized scaling
        - It generally scales with mean as 0 and standard deviation as 1.
        - It is also called Z score scaling.
        - Formula is : (x- u)/sigma

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    - VIF is calculated as 1/ (1- R2).
    - VIF can only be infinite when 1 -R2 =0. It means R2 value is 1.
    - This can only happen if there is a high collinearity between two variables.
    - One example, if I have data with predictors in different unit. For example, carpet area in feet and in meters. Even though the numbers are different the values are same in different unit of measurement and hence this will return R2 as 1.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
    - Quantile – Quantile plot is used to compare distribution of a dataset to theoretical distribution like normal deviation or standard deviation.
    - It is used to ensure the validity of the tests in regression.
    - This also helps in confirming whether the algorithm is correct or needs further improvement.