**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? (3 marks)

Ans: Using bar plot I am able to compare how change in one variable affects dependent variable. Here are some inferences.

1. With using BoxPlot it is clear that , there is no much outliers in the given data set.
2. During Fall and Summer seasons, bike rental demand is more. It is also observed that rentals are more in Fall and Summer seasons of both years 2018 and 2019.
3. Demand is increased in 2019 when compare to 2018.
4. Count is more that 5000 during the months from May to October. Also highest bikes are counted in 2019 September.
5. Count in all weekdays is nearly same.
6. More rental bikes if whethersit is clear.

2. Why is it important to use **drop\_first=True** during dummy variable creation? (2 mark)

Ans: drop\_first=True will remove the first dummy variable after creation, it is important to remove as we can represent n levels with n-1 dummy variable there is no need of extra dummy variable to build a model. If we miss to delete there will be a redundancy and multicollinearity get introduce which affects computation of coefficients.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: temp

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: Firstly we can evaluate the model by examining difference between actual observed target values with predicted values. This difference is called as Residuals. Residuals can be analysed and interpret our assumptions with

1. Histogram plot. If histogram plot shows residual points are bell structure it means residuals are normally distributed.
2. Q-Q plot. If residual points are falling approximately on standard normal distribution points, indicates residuals are normally distributed.
3. Residuals vs fitted plot. This assesses if residuals are randomly scattered around 0 (there is no pattern) indicates model is appropriate.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans:

1. Temp
2. Light snow rain
3. Windspeed

**General Subjective Questions**

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression algorithm well fit for continuous variables with linear relationship. Linear relationship means that dependent variables are gets change when there is a change in independent variable.

Algorithm:

Step 1: Understanding and Data visualization.

Step 2: Data Preparation.

Step 3: Data Split into Train data and Test data.

Step 4: Scaling Train data using MinMax scaling for 0/1 values.

Step 5: Building a linear model.

1. Pop train target/dependent variable to y axis (y\_test) and remaining independent variables to (x\_test)
2. Apply Automated approach RFE (Recursive Feature Elimination).
3. Adding constant to X train linear model
4. Applying Ordinary Least Squares.
5. Evaluating P-Value and VIF for Multicollinearity in the model.

Step 6: Evaluating Model - Residual Analysis

1. Error terms are normally distributed.
2. Error terms are centred at zero.
3. Examine difference between actual observed value vs predicted value

2. Explain the Anscombe’s quartet in detail. (3 marks)

3. What is Pearson’s R? (3 marks)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)