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)

- In the box plot we see that 3 categorical variables Seasons, Month and Week
- Seasons, we see that bikes are used more in the Summer and Fall season
- Months, we see that bikes are used more during summer and fall months (i.e. April to Sept)
- Day of the week, more bikes are used on working day

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

- When we create a dummy variable, we use drop_first = True to delete the first column hence not enabling it to create any extra columns

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

- With the pair plot, we see that the numerical variable with highest correlation with target variable is temp
- It is because, people will ride bikes when the temp is high hence it explains why summer and fall season

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

- The relationship between independent mean and mean of independent is linear
- Observations are independent of each other
- For any fixed value of dependent and independent is normally distributed

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

- Temp
- Sept
- winter

General Subjective Questions

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

- Linear Regression is supervised type of machine learning model
- The output is continuous and has a constant slope
- This method is mainly beneficial for predicting values in a range rather than any categories
- Mainly two types of Linear Regression:
    - SImple
    - Multivariable
- Simple Regression:
    - Can be explained by a line equation y=mx+c
    - Here,
        - Y - dependent variable
        - X - independent variable
        - C - intercept
        - M - slope of the line
- When we have one variable affecting the outcome it comes under the Simple Regression type
- Multivariable Regression
    - When we have more than 1 variable that affect the outcome of the model, it comes under multivariable regression

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

- Anscombe's quartet helps us understand the importance of data visualization and effects of the outliers and other data that influence on statistics
- It has 4 sets of data, the data will have very similar descriptive stats
- If these similar 4 sets are plotted, they appear very different and also have different distributions as well
- As the plotting helps us understand the data, this tells us how important visualization is
- The plotting before building the model can tell us the exact features that can affect the outcome exactly

3. What is Pearson's R? (3 marks)

- Also called as Pearson's Corelation Coefficient
- The value ranges from -1 to 1
- This value will tell us how the variable affects the outcome/target value
- Only a linear correlation can be studied by the pearson's r value

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

- Scaling helps normalizing the range of the independent variables
- When we have multiple features, and we want to have the variable in the same number range. We perform scaling
- **Normalized Scaling:** Scaling is done in between the range of 0-1

- **Standardized Scaling**: The scaling is done such that the mean is 0 and the std deviation is 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

- VIF value gives us a measure that how much variance of regression and the correlation goes up to, this happens due to collinearity
- If the variables are in right angle to each other the VIF = 1
- If not and the relation is perfect VIF=infinity
- Large value of VIF tells us that the variables are correlated

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

- Quantile - Quantile (Q-Q) plot, is a method of graphical representation
- This helps us to check if the data is from theoretical distribution
- Theoretical distribution: Normal, Exponential or Uniform
- We can know if the data originates from the same distribution
- If the data sets come from the same distribution, the points will fall on the line, the graph is plotted on a 45 deg line