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. More number of customer in clear weather
    b. More number of customer in fall season
    c. More number of customer in september month

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

drop_first=True is used to reduce the extra column created during dummy variable creation. It helps in reducing the correlations created among dummy variables.

If you have three level Categorical column, (BAT, BOWL, FIELD), you can create n-1 columns for dummy variable, as BAT can be 10, BOWL can be 01 and FIELD can be 00

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

temp has the highest correlation and **atemp** is similar **temp**

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

To assess/validate the assumption of Linear Regression after building the model on the training set.

- **Linear relationship**: Using Scatterplot, it should be easy to visualize the linear relationship plot
- **Absence of Multicollinearity**: Having no related predictors in the multiple linear regression model
- **Independence of residuals/Normality of Errors**: If residual is not normally distributed then it means there is a presence of outlier.

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

Top three features are:

- Weather Situation
- Temperature

- Wind Speed

General Subjective Questions

6. Explain the linear regression algorithm in detail.

Linear regression is a predictive analysis model, it is used to relate the dependent variable/ predictor variables to one independent variable/an outcome variable.

For example, in bike sharing assignment, number of customers, “cnt”, an outcome variable, it is predicted using predictor variables such as season, temperature, wind speed etc.

7. Explain the Anscombe’s quartet in detail.

It is a group of four dataset where statistical analysis looks the same. But when we visualize them they look different. It helps us in understanding the importance of data visualization and how it is easy to fool the regression model.

8. What is Pearson’s R?

It is a way to measure a linear correlation. This number lies between -1 and 1, where it measures strength and direction two variables.

9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling means putting the feature in the same range, for example we scaled temp to be in the same range with others.

So we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

In normalization, we map the minimum feature value to 0 and the maximum to 1. Hence, the feature values are mapped into the [0, 1] range whereas In standardization, we don’t enforce the data into a definite range. Instead, we transform to have a mean of 0 and a standard deviation of 1

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

If we have a perfect correlation in all variable then we can see Infinite VIF, which also means that we are getting r-squared as 1.

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

Q-Q plots are also known as Quantile-Quantile plots. This plots the quantiles of a sample distribution against quantiles of a theoretical distribution.

Q_Q plot is important as it ensures the ML Model is based on the right distribution. It is used for determining the following:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behavior.