Assignment Based Subjective Questions

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- June to September bike demand is high whereas January is probably the lowest
- No matter what weekday demand is steady
- Bike demand is higher in better weather such as clear sky or fewer clouds
- Spring is the worst season for bike sharing
- 2019 was the better year of the two

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

- Helps to reduce an extra column which can be created when making dummy variables

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

- We can see the following correlation with the target variable and temp @0.63

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

- We validate the assumptions made by looking at a few things, 1) Errors have a mean of 0, 2) Linearity of Data, 3) Errors are of a constant variability,

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

- Temp
- Weather
- Yr

Assignment Based Subjective Questions

Explain the linear regression algorithm in detail

- Is a machine learning algorithm of which is supervised learning and predicts based on independently specified variables. Primarily for finding the relationship and correlation between on variables and forecasting. In simple terms it finds the linear relation between an input and output (for example money and happiness)

Explain the Anscombe's quartet in detail.

- Anscombes quartet is a group of dataset in fours that are almost identical in a statistical manner, but that have some strange differences which can trick a regression model if put into such, thus will appear differently when plotted onto commonly used scatter plots.

What is Pearson's R?

- Pearson's R is a summary of the strength of a relationship between different variables, for instance if said variables go up and down together the correlation coefficient (Pearsons R) will be positive whereas if the variables go opposite to each other the coefficient (Pearsons R) will be negative accordingly. Pearsons R is between -1 and +1

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

- Scaling is used to bring down the values of independent values/features on a dataset so they're all on the same scale to help carry out further calculations in algorithm quickly and with ease.
- Normalized scaling is where values are rescaled between 0 and 1 where as standardization is where the mean will be equals to 0 and the standard deviation equals to 1

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

- If the correlation between variables is perfect, then the VIF can be infinite. To solve you would need to remove the variable which is causing said perfect correlation.

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

- A QQ Plot is two quantiles against each other to compare the shape distribution, providing a graphical view of properties between the two distributions and Its purpose is to find out possibly if the two sets from the same distribution.