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?

1. The month has the highest median, which means people love bike rides depending on the month
2. From the pair plot we can see that Season 3 and Season 4 registrations were more, 2019 people registered more than in 2018,
3. Month Aug, Sept and October people registered more bikes.
4. Clear, Few clouds, Partly cloudy, Partly cloudy has more registration
5. People register bikes more during weekdays rather than at the weekend
6. People report more during the working days, rather than holidays and weekends

2. Why is it essential to use drop_first=True during dummy variable creation?
-> it helps in reducing the extra column created during dummy variable creation.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
-> I calculated the p-value and VIF after removing the one-by-one features till all the VIFs were below 5 and then tested the model with the test dataset that we split at the beginning of the model creation and also calculated the adjusted r-squared while removing each feature from the dataset.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
-> yr, weather sit and holiday are the three features for explaining the demand for the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   -> Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as we can see the name suggests linear which means the two variables which are on the x-axis and y-axis should be linearly correlated.

   Mathematically, we can write a linear regression equation as:

   $y = a + bx$

   Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

2. **Explain Anscombe's quartet in detail.**
   -> Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.
3. What is Pearson's R?
   -> In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

   The Pearson's correlation coefficient varies between -1 and +1 where:

   r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
   r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
   r = 0 means there is no linear association
   r > 0 < 5 means there is a weak association
   r > 5 < 8 means there is a moderate association
   r > 8 means there is a strong association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   -> **Scaling** means that you transform your data to fit into a specific scale, like 0-100 or 0-1. You want to scale the data when you use methods based on measurements of the distance between data points, such as supporting vector machines and the k nearest neighbours. With these algorithms, a change of "1" in any numeric characteristic has the same importance. Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude into account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

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

   -> If there is perfect correlation, then VIF = Infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

   -> Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.