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?

   The number of bikes rented depended up on the temperature which in turn is significant on the below categorical variables –

   a. Season - Summer and Winter seasons had higher rentals as compared to other seasons. Spring season had a negative correlation.
   b. Weather situation – Snow and Mist impacted the rentals negatively.
   c. Month of the year – During the summer, winter rentals were higher thus during (Jun – Sep) impacted rentals positively.
   d. Weekday - Day of the week had a slighter impact as compared to other categorical variables.

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

   N levels on a categorical variable can be explained in N - 1 columns. This will also avoid multicollinearity.

   Example : 4 seasons

   100 – Summer, 010 – Winter, 001 – Fall. When value is '000' it explains that it is 'Spring'

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

   The 'temp' column had highest correlation with the target variable 'cnt'.

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

   training set?

   Assumption of Linear regression is that the error terms have a normal distribution with mean as zero. Plotted a histogram to view the error terms distribution between the y_train and y_train_predicted value.

   As the error distribution was normal and mean was 0, the assumption was passed.

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

   1. **Temperature** has the highest co-efficient. So if it increases by 0.5402, count of bikes rented will increase as well. Business can concentrate on temperature
   2. 2019 rentals has improved as compared to 2018. **Year** is the second influencing factor.
   3. **Summer and Winter season** has more rentals as compared to other seasons. Can focus more during these times.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear regression is a supervised algorithm that learns based on the training data provided and predicts how the other independent variables affect the target variable. In other words, it explains the relationship between one or more variables and derive inferences.

2. Explain the Anscombe's quartet in detail.

   Anscombe's quartet is basis to visualize data rather than depend/see only the statistical figures. It has four datasets which will produce identical statistical measures and indicated in graphs for better understanding and visualizing.

   Let's say a kind of pairplot.

3. What is Pearson's R?

   Pearson's R is the most common way to measure the strength of relation between two variables. It lies between -1 and 1.

   When one variable changes, the other changes in the same direction.

   Example – An positive increase in A will increase the value of B in positive scale.

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

   Scaling is a process of data preparation and is applied on the independent variables to normalize the data within a range.

   Scaling is performed to generalise the data points as the gradient descent concept is used in linear regression

   Normalization is also called MinMaxScaling – All values are scaled between value of 0 and 1. It compresses the data between these two values.

   Standardization – Scales the data points between the mean (0) and sigma (1)

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

   When VIF is infinite, it means that the variable is completely predictable by another variable in the dataset. This shows a perfect correlation as R2 = 1 so VIF = 1/1-R2 = inf.

   We can drop one the variables which is causing the multicollinearity.

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

   Quantile – Quantile plot is a probability plot to compare two probability distributions.

It is used to determine if the two distributions are similar and have common behaviour. It can help us to check if the two datasets come from the same sample population.

When training and test datasets are received separately in linear regression, Q-Q plots can help us confirm if the data came from same population with similar distribution.