Assignment-based Subjective Questions

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

Ans: If we don't do so i.e use drop_first = True then the dummy variables will be corelated. This can affect the model adversely. Like there will be trouble converging the iterative models and the list of variable importance can be unfavourably distorted.

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

Ans: temp and atemp are highly correlated

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

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans: Windspeed, Working Day and the Year are mostly contributing significantly towards the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

AnsL Linear regression is a statical model which is used to analze the linear relationship of a dependent variable with set of independent variables.

The relationship can be denoted by

y=mX+b for one independent variable and for more independent variables we can add further m1X1, m2X2 etc…

There can be positive linear relationship when both the dependent and independent variable increases

There can be a negative linear relationship when the independent variable increases and the dependent variable decreases.

There are two types of Linear regression

a. Simple linear regression: it is a linear regression model with single independent variable.
b. Multiple linear regression: It is a extension of the simple linear regression. It has multiple features.

Q2. Explain the Anscombe's quartet in detail.

Ans: It has four datasets with simple statistical properties. They have very different distribution and when plotted on graph seems very different.

Q3. What is Pearson's R?

Ans: It is a measure of correlation between two data sets. It is calculated as the ration between the covariance of two variables and the product of standard deviation.
Its result is always between -1 and 1.

Q4, What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is a process to normalize the rand of independent variables of the data
It is performed when we have multiple independent variables like age, weight, Glucose. All these are measured in different ranges. Scaling would help them all to be in the same range,

Normalized scaling rescales the values into range 0 to 1

It is calculated as

X' = (x- min(x))/(max(x)- min(x))

Where 'x' is the original value and 'x`' is the normalized value.

standardized scaling: In this technique the values are centered around the mean with a unit standard deviation. It rescales the data to have a mean of 0 and a standard deviation of 1

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: VIF is infite in the case where there is a perfect correlation. In such case R2 is one, which implies 1/(1-R2) equals to infinity.
In these cases the corresponding variable me be expressed by a linear combination of other variable.

Q6 . What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

It is the Quantile=Quantile plot which is graphically a method of comparing two probability distributions by plotting their quantities against each other.

It helps two determine if two datasets belongs to population with a common distribution.

If the two distibutions being compared in the Q-Q plot are similar then q-q plot will lie on the line y=x.

If the distibutions are linearly related the q-q points will will lie on a line but not always on line y = x.