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? (3 marks)

Ans: I can infer that target variable i.e. cnt is maximum in summer and fall season and in terms of month from may to October. Also for the columns year for 2018 and 2019 cnt almost doubles in 2019, which can be due to business getting popular. Also number of bikes were hired more when the temperature was in the range of 25 to 35, that was also the case when humidity was less than 80 and windspeed is around 10-15.

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

Ans. Drop first is important because it will the drop the extra column created while creating dummy variable for categorical column which well help in reducing the space.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans. Registered

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans. I validated the assumptions of Linear Regression using scatter plot for residuals for constant variance, histogram of residual for normal distribution, vif score for collinearity and scatter plot of predicted vs residual for linear relationship

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans. Weathersit_Rain(-1.009), yr(0.512),mnt_Jan(-0.510)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans. Linear regress algorithm is a supervised machine learning algorithm which performs linear regression to predict the target variable(dependent variable) based on some independent variables.

Linear regression is used to find out a linear relationship between dependent and independent variables to predict the dependent variable.

Basically 4 assumptions assumption are made before linear regression :

1. Linearity: The relationship between X and the mean of Y is linear.
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of X, Y is normally distributed.

A function is assumed with an intercept and coefficient of the independent variable. Then the function is minimized to find out the best fit line for the model. Data is split into 2 parts and one part is used to train the data i.e find out the best fit function by adding features or removing features and then 2nd part is used to test the obtained model.

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans. Anscombe's quartet is a set of 4 data sets with 11 data point with nearly same statistics. Although statistics are same including mean, variance, even the linear regression line but the data is distributed very differently in each set, which implies graphical representation of the data is also very important for analysing.

3. What is Pearson's R? (3 marks)

Ans. Pearson's R or also known as Pearson's Corelation Coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations. Therefor it is a normalized measure of the covariance that means it will always have value between -1 and 1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans. Scaling is normalizing the data to a particular range to speed up the calculations in an algorithm. Most of the times scales in which data is present differ in magnitude, units and range. When an algorithm run on such data it only accounts magnitude and ignore the units which results in incorrect modelling. Normalized scaling brings the values between 0 and 1. It subtracts the minimum value from the data and divide the result by the difference of the maximum and minimum value of the range. In standardized scaling, all the data is brought into normalized standard distribution which has mean zero and standard deviation 1. It is done by subtracting the data by the mean of data and dividing the result by standard deviation of the data.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans. If the independent variable are perfectly correlated then the VIF is infinite.

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

Ans. Quantile-Quantile or Q-Q plot is a plot between two quantiles against each other. It is used to find out weather the 2 sets of data came from the same distribution or not. A 45 degree angle is made if the 2 data sets are from a common distribution. If the 2 data sets are linearly related the points of the Q-Q plot will lie approximately on the line but not necessary on the line y=x.