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)

i) yr – a categorical variable has the highest coefficient, so if the yr value is 1 (ie 2019), it will have highest impact on target variable cnt, so on year on year basis cnt will keep increasing.

ii) weathersit_3 – ie categorical dummy variable for "Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds" , has the 2nd highest coefficient tough -tve, so it impacts cnt in a negative way so if the weather is like as mentioned cnt will decrease significantly

iii) season_4 – ie categorical dummy variable for season "winter" has the 3rd highest coefficient , we conclude there will be increase in cnt in winter season

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

Ans: Any categorical variable having 'n' distinct values can be expressed with n-1 dummy variables having values 0 or 1. Like only 1 dummy column can be used to express 2 values of the categorical column ie 0 and 1,two dummy variables can be used to express the categorical column with 3 distinct values (0,0) (0,1) (1,0). Thus instead of creating n dummy columns to express n values of categorical column we can just create n-1 dummy columns.

i) Which leads to less no. of variables and thus less complexity of our model.
ii) It also decreases the multicollinearity in our data set as the rest of the variables can define the state of the dropped variable , otherwise their sate will become dependent on each other.

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

temp , considering registered as non useful which I have dropped later.

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

By plotting dist plot of residuals and it should be close to normal distribution.

residuals = y_test - y_test_predict

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

i) yr – has the highest coefficient , year on year the cnt seem to be increasing

ii) weathersit_3 – has the 2nd highest coefficient though negative, ie categorical dummy variable for "Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds", so it impacts cnt in a negative way so if the weather is like as mentioned cnt will decrease significantly

ii) temp – has the 2nd higest coefficient , so more the temp more the increase in cnt

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear regression equation is defined as:
   y = C0 + C1.x1 + C2.x2 + ……Cn.xn

   where y is the target variable , C0…Cn are coefficients and x1…..xn are the predictor feature variables

   The aim of this linear regression algorithm is to find the values of coefficients which will predict the value of y as close as possible to observed value of target variable say Y.

   Firstly random coefficient values are picked and RSS(Residual Sum of Squares) is calculated for those set of coefficients.

   Then with every iteration those values are tuned by minimizing the RSS(Residual Sum of Squares) using gradient decent.

2. Explain the Anscombe's quartet in detail.

   Anscombe's quartet is used to show the importance of plotting datasets.

   Anscombe's quartet is a collection of 4 datasets , which have same mean , standard deviation, regression line. But their plots are very different.

3. What is Pearson's R?
   Pearsons R is a measure of corelation between variables , its value ranges from -1 to 1, a value close to -1,1 shows more correlation. A value towards zero shows less correlation.

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

   Ans:

   a) scaling is a technique used during model training to bring features to same scale
   b) If scaling not done , the coefficients we will get for features after training show vast difference , and will not show the correct significance of features. And our inference will go wrong.

      Also the convergence for gradient decent will take more time.

   c) standardized scaling – changes the distribution of feature into standard normal distribution ie with mean =0 and std dev of 1. Given by x-mean(x)/stddev(x).

      min max scaler – it normalizes the value of x between 0 and 1. Given by

      Xmax – X/Xmax- Xmin.

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

   VIF is given by 1/1-R squared , the vale of R squared varies from 0 to1 .If 2 variables are highly correlated , R-squared will tend to 1 and value of VIF hill tend to infinity.

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

   Q-Q plots is used to see if

   i) Data set follows a certain type of distribution – theoretical quantiles for distribution against which data set is to be compared are calculated , actual quantiles of data set are calculated and theoretical and actual quantiles are plotted. A straight line means the data set follows the distribution being compared

   ii) 2 data sets follow same kind of distribution – quantiles of the two datasets are calculated and plotted against each other , a straight line means that the datasets have similar kind of distribution

   In linear regression Q-Q can be used to:

   i) To check if error terms follow the normal distribution which is important assumption for linear regression
   ii) Check if the training and test set follows the same distribution