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)

   Answer: From the basic analysis of the categorical variables we can clearly see that the numbers of total users are higher for 'Fall' season, and there is a significant increase in number of users in the year 2012 that that of 2011. We can also infer that the demand of shared bikes are higher during the Holidays. There is a much higher demand of bikes in a clear or a cloudy day than that of a rainy day.

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

   Answer: This actually helps us reduce Redundancy caused by the dummy variables, if we do not drop a column from the dummy variable then there is a chance of high correlation between some of those variables which will cause higher P value while building the model and will be making a particular dummy variable redundant.

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

   Answer: The variable 'registered' 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? (3 marks)

   Answer: After building the final model I have checked if the residuals have any relation with the predicted Y values using a scatter plot and from the plot it was clear that there was no relationship between them. I have also checked if the residuals are normally distributed or not using a distribution plot with a bin size of 20, and the data points came out to be normally distributed.

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

   Answer: Based on the model the most important 3 factors are , 'season', 'yr' and 'holiday'.

General Subjective Questions

1. Explain the linear regression algorithm in detail

   Answer: The structure of linear regression algorithm is,
   $Y = b0 + b1*x1+b2*x2+…+bn*xn$

Using this algorithm we can predict the change of a variable (Y) using the another variable(X) if they are linearly related with each other, the algorithm learns form the given dataset and finds out the b0 and corresponding bn's for all available xn in the dataset, the cost function that is used here is,
cost = 1/n(squared root ((y actual – y predicted)squared). And it used gradient descent to optimise this cost function to get the optimal b0 and all corresponding bn.

2. Explain the Anscombe's quartet in detail.

   Answer: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

3. What is Pearson's R?

   Answer: Pearson's correlation coefficient (R) is a measure of the strength of the association between the two variables

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

   Answer: Scaling is the method used to transform data which are belonging to different scales in the same scale.

   Distance algorithms like KNN, K-means, and SVM are most affected by the range of features. This is because behind the scenes they are using distances between data points to determine their similarity. This is the main reason of data scaling.

   The presence of feature value X in the formula will affect the step size of the gradient descent. The difference in ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

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

   Answer: An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. It actually happens when R squared becomes 1 for an independent variable.

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

   Answer: A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

   It is mostly used to check the normality of the distribution of a given numerical continuous data set, in linear regression it is used to plot the residuals.