Assignement-based Subjective Questions

1. These are the inferences for categorical variables:
   - Season - Spring season has the least average number of rental bikes taken while Fall has the highest average
   - Yr - 2019 has significantly more number of people using rental bikes
   - Mnth - May-October seems to be the peak season for rental bikes which might be due to the pleasant weather at that time of the month
   - Weekday - Most of the days has similar number. Day 0,1 is slightly lesser than the rest of the days
   - Weathersit - Clear weather has more number of people renting bikes while if there is light snow or rain, lesser people prefer rental bikes
   - Holiday - People might be commuting to office during non-holidays and therefore more people rent bikes on non-holidays
   - Workingday - Slightly more number of people use rental bikes on working days, might be to commute to offices

2. Drop-First = True is needed to drop the original columns from the dataset and keep the newly created dummy columns
3. The Variable registered has the highest correlation of 0.94 but it is linearly dependent with 'cnt' variable and can't be considered as part of independent variables. The column 'atemp' has the highest correlation of 0.63. with 'cnt'.
4. By testing the model on the remaining 20% test set that was created from the dataset using train-test split method.
5. The top 3 would be 'atemp', 'season' and 'yr' as they have significantly greater changes for different values.

General Subjective Questions

1. Linear Regression is an algorithm where we plot a linear graph by finding the best fit line using the provided data points (X – independent variables, y – dependent variable) and by reducing the loss function (cost) using Differentiation or Gradient Descent methods. We use the linear graph to predict the outcomes for new values of X.

   The equation of the line would be:
   $y = mX + c$
   where,
   m → slope of the line
   c → y-intercept

   Values m and c can be found using the Gradient Descent method or by differentiation.

2. Anscombe's Quartet comprises of four datasets containg 11 data points (X,Y) each with same descriptive statistics such as same mean of X and Y respectively, standard deviation of X and Y respectively and Correlation. But they appear very different

when plotted in a graph. One of the graph looks linear, the second is non-linear and the third although linear an outlier, and the fourth has all the values parallel to the Y-axis except one which is an outlier.

3. The Pearson's Coefficient (r) is a way to measure linear correlation between two variables. Its value lies between 1 and -1 and defines the strength and the direction of the relation between the two variables. It is calculated using the formula:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

   Where n is the sample size, and x,y are the sample points.

4. Scaling is a Mathematical technique to standardize certain independent features during data pre-processing steps. It is of two types – Standardized and Normalized Scaling. Standardization converts a standard normal distribution to a normal distribution with 0 mean and unit standard deviation. It is only used on Normal Distributions and is also called Z-Score Normalization.
   Formula for Standardization is,
   x' = (x-mean)/standard_deviation

   Normalization is used to transform independent features to a similar scale to reduce the bias of Machine Learning algorithms on features with large magnitudes.
   Formula for Normalization is,
   x' = (x-xmin)/(xmax-xmin)

5. VIF formula is 1/(1-RSquare). Therefore if RSquare is equal to 1, then VIF is infinite. It happens since two variables might have a perfect correlation with each other, equal to 1. To make VIF finite, we need to drop either of the two perfectly correlated variables.

6. Q-Q plot is a graphical method where we plot the quantiles of two datasets to check if the two datasets follow the same distribution type or not. In Linear Regression, we plot the quantiles of the training and test dataset to check if the two datasets have same probability distribution or not, which ensures if the two datasets have been taken from a population having same probability distribution.