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?**
   Season – The highest bike booking is done in Fall, followed by Summer & Winter
   Mnth – Most of the booking is done through May and Sep, this means there is a seasonal pattern in the data.
   Weekday – Almost all days are seeing booking for bikes. This might or might not be a good predictor
   Weathersit – 'Clear' weather brings cheer in booking bikes in higher numbers followed by 'Misty' one
   Holiday – More booking is done on 'Non-Holiday' days
   Workingday – More booking is done on 'workingday'

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Multi-collinearity is undesirable, and every time we encode variables with pandas.get_dummies(), we'll encounter this issue. One way to overcome this issue is by dropping one of the generated columns. So, the first transformed variable can be dropped to counter this issue.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   By checking the following things –
   1) Linear relationship
   2) Homoscedasticity
   3) No multicollinearity
   4) Independence of residuals
   5) Normality of error

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes**?
   Temp (+ve corr), yr (+ve corr) and Light_rainsnow (-ve corr)

General Subjective Questions

1. **Explain the linear regression algorithm in detail**.

   Linear regression is an attractive model because the representation is so simple.

   The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

   The linear equation assigns one scale factor to each input value or column, called a coefficient, and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g., moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

2. **Explain the Anscombe's quartet in detail**.
   Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
   It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

3. **What is Pearson's R?**
   Pearson's R is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling**?
   It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

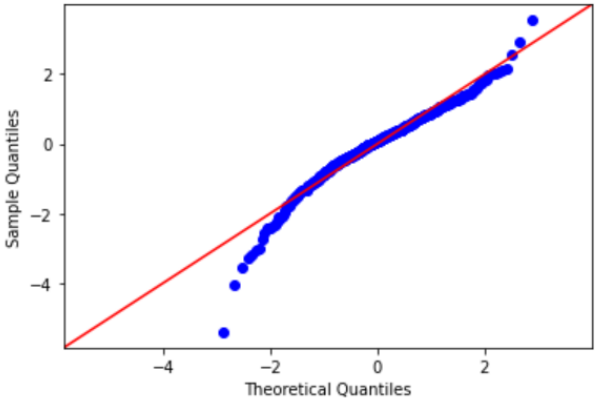
   Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

   Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance)

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen**?
   If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.
   For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
   If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.


Sample Quantiles
Theoretical Quantiles