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?

    - From the season's plot we can see that most of the bookings were done in fall(season 3) followed by summer(season2) and winter(season4).
    - Also we can see that most bookings happened during May, June, July, September, October.
    - Weatherit 1 (Clear, Few clouds, Partly cloudy, Partly cloudy) saw the most number of bookings followed by weathersit 2 (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist).
    - Most of the booking was done during non-holiday.
    - Weekday variables seem to have no clear trend and all the medians are more or less the same indicating.
    - Most of the booking has happened in the working day with median at around 5000.
    - The demand for bikes increased in the year 2019 when compared with 2018.

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

    The **drop_first = True** will drop the first column while creating dummy variables. While creating dummy variables the algorithm creates a number of columns equal to the number of distinct values present in those columns. One column out of those is not necessary, as if all the values are false the last one will be true so we can drop the unnecessary column after creating the dummy variable,

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

    From the correlation matrix we plotted 2 variables temp and atemp have highest correlation(0.63) with target variable cnt.

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

    To validate the assumption of linear regression below are the steps we can follow.

- One can plotting the scatter plot between dependent variable and independent variable and check if they are linearly variable or not
- We can check if there is little to no multicollinearity between the features i.e. the features are not highly correlated.
- Also we need to plot histogram of errors so as to see if they are normally distributed to validate the Linear Regression model.

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

   Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes:

   1. Temperature or temp which has the coefficient of 0.5749.
   2. weathersit_3 : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds with coefficient of -0.3094.
   3. Year or yr which has the coefficient of 0.2304.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear Regression is a method in machine learning where a model is trained on given data with some variable to predict specific behavior related to the data. From the name linear regression, we can derive that the method revolves around 2 linearly correlated variables on the x-axis and y-axis.

Mathematically a linear regression equation is written as

$Y = \beta_0 + \beta_1 x$

where
β_0 - y-intercept of the line.
β_1 - the slope of the line
y - Dependent Variable.
x - Independent Variable.

Further, the above equation can be found by minimizing the cost function using the least square method.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is the collection of four datasets with eleven (x,y) data points which when graphed look very different, despite having similar simple statistical properties.
These datasets were constructed with the purpose to show the effect of outliers on statistical properties.

The quartet is still used to illustrate how important it is to see the distribution of data graphically before starting any analysis of the data..
Using the graphical data one can get to know various anomalies present in the data like outliers, how diverse the data is, etc.
It also helps us understand how easy it is to mislead the regression algorithm using such data.

3. What is Pearson's R?

Pearson's R, also called the Pearson correlation coefficient or the bivariate correlation describes the strength of a linear association between two variables.
Pearson's R shows the distance of the data points from the best fit line drawn through the data of two variables using Pearson product-moment correlation

The value of Pearson's R ranges between +1 to -1.

For two variables Pearson's R-value indicates no relation between them and as the R-value moves to +1 it indicates positive association and if the R-value moves towards -1 it indicates negative association.

The formula to determine the value of Pearson's R is as below:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r=correlation coefficient.
x_i=values of the x-variable in a sample
$\bar{x}$=mean of the values of the x-variable
y_i=values of the y-variable in a sample
$\bar{y}$=mean of the values of the y-variable

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

   Scaling is a step performed while data Pre-Processing is applied to independent variables to normalize that variable's data and bring the same to a specific range.

   Many times collected dataset has highly varying features in magnitudes, units, and range.
   As the algorithm only considers magnitude and not units if scaling isn't done model generated by the algorithm would be incorrect, to solve this issue proper scaling should be done, and all the variables should be brought down to the same level of magnitude.

   The major difference between normalized scaling and standardized scaling is that normalized scaling brings down variables to a range between 0 and 1, whereas standardized scaling scales down variables in such a manner that their mean stays at 0.

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

   The formula to find VIF between 2 variables is as below:

$$\frac{1}{(1-R2)}$$

Now in the case where R2 =1 VIF turns out to be infinite.
The above scenario shows the perfect correlation between the 2 independent variables

An infinite VIF suggests that a corresponding variable may be exhibited by a linear combination of another variable with great precision. In order to solve this perfect multicollinearity issue one needs to drop either of the variables from the model.

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

Q-Q plots also known as Quantile-Quantile plots, plots the quantiles of a sample distribution against the quantiles of a theoretical distribution to determine how many values in a distribution are above or below a certain limit.
This helps us to track down if the set of data we have plausibly matches any theoretical distribution like exponential, normal, or uniform distribution

Q-Q plots help us to summarize any distribution visually.

It also helps one to determine if the 2 datasets are from the same population or not.

Apart from that it also shows how skewed the distribution of data is.

Many aspects like location shift, scale shift, change in symmetry, and the occurrence of outliers can be detected using this method.

While comparing if the distributions turn out to be similar the data point in the Q-Q plot almost lie on the line y = x, and if the distributions that are been compared linearly related, the points in the Q–Q plot will approximately lie on a line which won't be essentially y = x.