1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Season : More than 5000 bookings are there in season 3 and is followed by season 2 and 4. This means that season can be one of the predictors.
Months : The bookings are over 4000 in the months of 6,7,8,9,10. So it has some control over the analysis.
Weathersit : The weathersit has almost 5000 bookings in weather1 and followed by weather2. So weather has some trends for consideration.
Holiday : The variable 'holiday' has nothing much impact on the prediction.
Weekday : It is either weekday or weekend there is no significant change in the trend and is not having much impact.
Workingday : It is close to 5000 booking on working day and is having some impact on the prediction.

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

During dummy variable creation new columns would be created and this drop_first helps to reduce the nomber of columns.

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

The variable 'temp' having the highest co-relation with 0.63 co-eff.

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

The errors are normally distributed with the mean 0. Also from the histogram we could see the normal distribution, hence the assumption is right with linear regression.

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

The top 3 features are ,

'temp', 'weathersit_3', 'yr' which contributes more as per the co-efficients.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

It is used for finding out relationship between variables and forecasting. It performs the task to predict the dependent varaible value (y) based on the independent varaiable (x).
So this regression technique is used to find out the linear relationship between x

and y.

Formula : y = mx + c

2. Explain the Anscombe’s quartet in detail.

It has four quarters nearly simple descriptive statistics with different distributions. Each of the datasets consists of 11 points.
It describes the effect of outliers and other observations on statistical properties.The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson’s R?

It is called as bivariate co-relation or the co-relation coefficient. It is a measure of linear corelation between two sets of data. The result always have a value between -1 and 1.
Correlations equal to +1 or −1 correspond to data points lying exactly on a line (in the case of the sample correlation), or to a bivariate distribution entirely supported on a line (in the case of the population correlation). The Pearson correlation coefficient is symmetric: corr(X,Y) = corr(Y,X).

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

It is a preprocessing step which is applied to independent variables to normalize the distribution with a range. It is required to bring all the variables to same level for calculation.
It doesn't affect p,F,t,R^2.
Normalization brings the data to in the range of 0 and 1 whereas the standardization brings the data to a standard normal distribution which has mean 0 and standard deviation as 1.

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

It shows that there is a perfect corelation between two independent

variables. Incase of perfect corelatation R2 = 1 and the VIF value becomes infinity. So we are dropping one of the columns which is helping for this multi colinearity.

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

These are two qunatiles plotted against each other. It is a fraction that certain value falls below that quantile. If median becomes quantile then the points would be plotted 50% above and below the line.It is used to find out two sets of data from the same distribution. It is used to compare the shapes of the distribution using graphic view.