## 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)

Inferences are as below –

1. Upward trend from 2018 to 2019
2. Maximum bookings in 2018 and 2019 were in the Fall season.
3. Clear or partly cloudy weather attracted more bookings.
4. Most bookings are done between months May through October

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

**drop_first** = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
Let's say we have 3 types of categorical values related to "Furnishing Type" say "Not Furnished", "Semi Furnished" and "Furnished". If we define dummy variables for "Not Furnished" and "Semi_Furnished", then we do not need 3rd variable to identify the "Furnished".

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

"temp" variable has the highest correlation with target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Used the following assumptions –

1. No auto correlation
2. Variables show linearity.
3. Insignificant multicollinearity

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

Top 3 features contributing to the demand of shared bikes are –

1. Temperature i.e. temp
2. Spring season
3. Month of July

## General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression.

There are two major types of Linear Regression –

1. Simple Linear Regression
    a. This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable.
2. Multiple Linear Regression
    a. This involves more than one independent variable and one dependent variable.

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.
In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson's R? (3 marks)

Pearson's R is a numerical summary of how closely the two variables are linearly associated. If the two vary and fluctuate at the same time, then the correlation coefficient will be positive. If when one variable increases the other one decreases and vice versa with respect to high values of one variable and low values of another, then correlation will be negative.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling 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 have to do scaling to bring all the variables to the same level of magnitude

1. Normalized Scaling
    a. Scikit-Learn provides a transformer called MinMaxScaler for Normalization.
    b. It is used when features are of different scales.
    c. Scales values between [0, 1] or [-1, 1]
    d. It is affected by outliers
2. Standardized Scaling
    a. Scikit-Learn provides a transformer called StandardScaler for standardization.
    b. It is used when we want to ensure zero mean and unit standard deviation.
    c. Scales values not bounded to a range
    d. It is less affected by outliers

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

It indicates perfect correlation between two independent variables

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A QQ 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.

Importance of Q-Q plot
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more

insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.

In the case of two data samples, there might be a need to find out whether it is possible to assume that they have common distribution. In this situation, location and scale estimators can be used together to estimate common location and scale for both groups. Also, if the two samples differ, it would be nice to know how they are different.