## 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)

   Ans. a) cnt values are decrearising with weathersit. Heavy rain does not have any booking. b) Booking are more on working day.c) Booking are less on holiday.

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

   Ans. If we have n dummy (mutually exclusive) variables then we can represent them with n-1 variables, because we will have $2^{(n-1)}$ combination with n-1 variables which is pretty enough for n variables.

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

   Ans. Registered variable.

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

   Ans. By comparing the y_test values & y_predicted values with the help of r2_score(y_test, y_predicted) and plotting the distribution curve for error values.

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

   Ans.weathersit, atemp & regi_casual

## General Subjective Questions

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

   Ans. Linear Regression algorithm is based on linear equation (straight line) model. In this one variable say dependent variable will have linear relation with independent variable. This relation can be positive or negative or constant. It means the dependent variable will increase or decrease or remain constant respectively with increase in independent variable.

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

   Ans. A statistician Francis Anscombe have analysed four data sets in 1973. The graphs for these data sets were looking very different, but he found that statistical characteristics were same four all four graphs. The mean of x, mean of y, linear regression line, sample variance of x & y values etc were same for all four graphs.

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

   Ans. Pearson's R (also known as Pearson's correlation) used to measure how strong a relationship is between two variables. It is commonly used for linear regression. The Pearson's correlation coefficient is denoted with the symbol "R". The correlation coefficient formula returns a value between 1 and -1. Here,

   -1 indicates a strong negative relationship

   1 indicates strong positive relationships

   And a result of zero indicates no relationship at all

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

   Ans. Scaling is just the conversion of numeric values of variable in different scale. For example, if we have one variable say distance and the values are in meter than we can scale the values in kilometre. Some time we have variables with large difference in values for example one variable have values between 1 to 10 while other variable has values between 1-10000. If this will be the scenario, then one variable will diminish the effect of another variable. For this reason, we will do the scaling. In this example, either we will divide the values of second variable with 1000 to drag them in range of 1-10 or we will multiply the values of first variable with 1000 to drag them in range of 1-10000. In the normalized scaling, the minimum value of that feature gets transformed to 0 while the maximum value will transform to 1 and all other values are normalized between 0 and 1. Formula for it is $x_{new}=(x-x_{min})/(x_{max}-x_{min})$. While in standardized scaling, it has a mapping of mean value as 0 and variance equals 1. Formula for it is $x_{new}=(x-\mu)/\sigma$.

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

   Ans. If a variable has perfectly correlated with other then we will get infinite VIF. If the correlation will be 1 then we will have VIF infinity because $VIF=1/(1-R^2)$.

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

   Ans. Q-Q plot (quantile-quantile plot) is used to checked that the data is normally distributed or not. For this, we plot the sample distribution values against the normal distribution values. For example, we plot the 75th percentile of sample value against 75th percentile of standard normal distribution value as so on for other percentiles. If we get a straight line on the plot, it means the data is distributed perfectly. In linear regression, many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.