Bike Sharing Assignment – Questions

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

   Ans: Below are the observations -
   1. The bookings by registered users are high during weekdays while casual users book high during weekdays
   2. Higher number of bookings are done on weekdays than the weekends
   3. Sundays have the least bookings. There is a gradual increase in bookings as the week progresses until Thursday and then it goes down
   4. Casual books have a reverse trend to total bookings – The bookings are high on Sunday and gradually reduce during the weekdays until Thursday and subsequently increase

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

   Ans: Every process of dummy variable creation involves creating n columns where n = number of distinct values of the categorical variable. Since the values are binary in nature where 1 represents presence of the value and 0 the absence, the interpretation can be managed with n-1 dummy columns, since the nth column always represents the absence of values in the remaining n-1 columns. **drop_first=True** drops the nth column and hence reduces the number of dependent variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Ans: Registered and casual users

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Ans: Following steps were carried out
   1. Verified the trend to see if the distribution of residual errors follows normal distribution
   2. Verified the trend to see if there is no co-relation between the residual errors and target variable
   3. No values of variables with High P-value or VIF value was found

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans: The months of August September and October and the season of Summer is when the demands are at the best

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Ans: Linear regression model is a way of associating a scalar response with one or more connected independent variables**.** There are 2 types in this -
   1. Simple Linear Regression - If the model considers only one independent variable to explain the dependent variable
   2. Multiple Linear regression - If the model considers multiple independent variables to explain the dependent variable

   The attempt in Linear regression model is to fit the values of the dependent and independent variables within a linear trend with an associated accuracy.

   In case of Simple linear regression, a straight line is expected to be fit to depict the dependency between the dependent and independent variable in a way that the error values (Difference between the actual value of the dependent variable and its value if depicted on a line in a scatter plot) is minimum.

   The relation between the dependent and independent variable is described in the equation

   y = Beta0 + Beta1 * x

   Where Y is the dependent variable and x is the independent variable

   Beta0 is a constant, that represents the value of dependent variable when x = 0

   Beta1 is the co-efficient that define the proportional increase of dependent variable with the value of the independent variable

   In case of multiple linear regression, a straight line is expected to be fit to depict the dependency between the dependent and independent variables in a way that the error values (Difference between the actual value of the dependent variable and its value if depicted on a line in a scatter plot) is minimum.

   The relation between the dependent and independent variable is described in the equation

   y = Beta0 + Beta1 * x1 + Beta2 * x2 + Beta3 * x3 + Beta4 * x4 ….

   Where **y** is the dependent variable and **x** is the independent variable

   Beta0 is a constant, that represents the value of dependent variable when x = 0

   Beta1 is the co-efficient that defines the proportional increase of dependent variable **x1** with the value of the independent variable

   Beta2 is the co-efficient that defines the proportional increase of dependent variable **x2** with the value of the independent variable

   And so on.

2. Explain the Anscombe's quartet in detail.

   Ans: Anscombe's Quartet is a group of four identical data sets, but have some peculiarities in the dataset that render the linear regression model useless if found. When plotted on a scatter plot, all four datasets display a trend that shows that they are not connected. Hence to make sure that the right model is built and no errors seep in, visualization of the data before problem solving is important

3. What is Pearson's R?

   Ans: Pearson's R is a measure of the linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations. Hence the values of the measure are converted to always lie between -1 and 1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Ans:
   - Scaling is the process of deducing the values of all independent variables that are to be considered for a mathematical modeling to a common scale, thus removing the bias associated with the nature of the value and the different measures used to represent them.
   - Scaling is performed in order to bring all independent variables to a common scale of measurement so that the values are finally interpretable in terms of each other
   - The difference between normalized scaling and standardized scaling is that Normalization tries to represent all values with their proportionate value within a scale of 0 to 1 while standardized scaling tries to represent all values centered around their mean.

   Formulae –
   1. Normalized scaling: X = (x – xmin) / (xmax-xmin)
   2. Standardized scaling = X = (x-mu) / sigma

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans: This happen when there is perfect linear co-relation between the independent variables and the particular variable having the VIF value when it is made dependent on the rest of the independent variables. hence the r2 for the variable becomes 1 rendering the VIF value to be infinity

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Ans: The Q-Q plot, or quantile-quantile plot, is a graphical tool to assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.