**Assignment-based Subjective Questions**

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

**Answer:** The demad of bike is less in the month of `spring` when compared with other seasons

- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with workign day and non working day.
- The bike demand is high when weather is `clear and Few clouds` however demand is less in case of `Lightsnow and light rainfall`.

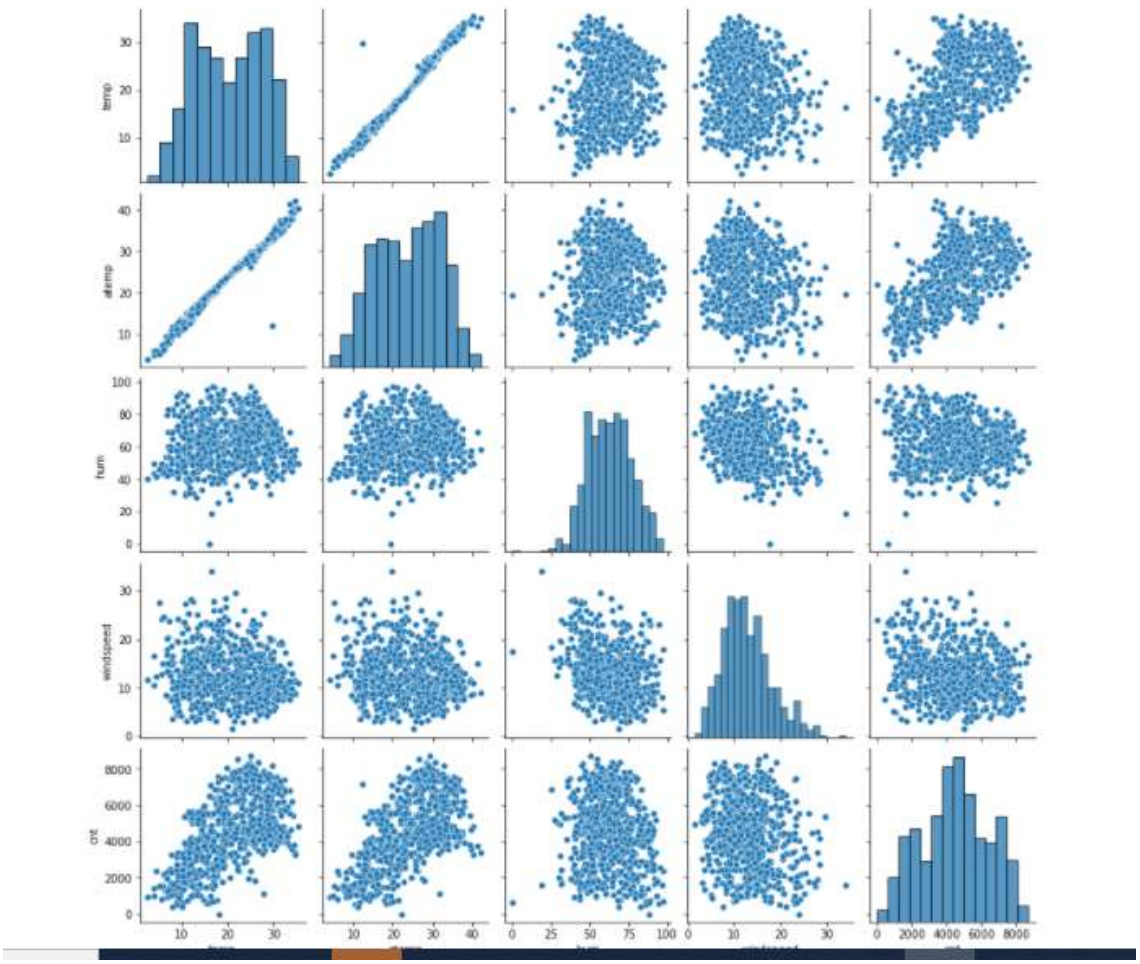
**Question 2: Why is it important to use drop_first=True during dummy variable creation?**

Answer: 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

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

Answer:

```
In [28]: sns.pairplot(df, vars=['temp', 'atemp', 'hum', 'windspeed', 'cnt'])
         plt.show()
```

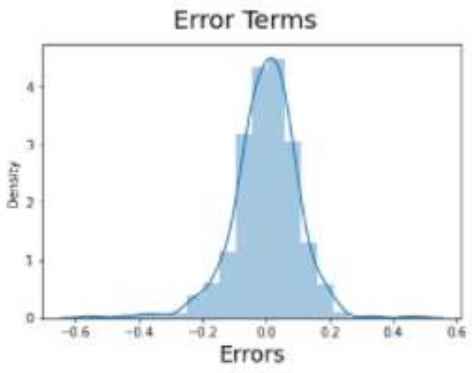



**Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?**

**Answer:**

```
In [133]: #Checking ASSUMPTION OF NORMALITY:
          # Plot the histogram of the error terms
          fig = plt.figure()
          sns.distplot((res), bins = 20)
          fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
          plt.xlabel('Errors', fontsize = 18)                         # X-Label

Out[133]: Text(0.5, 0, 'Errors')
```




**Residuals distribution should follow normal distribution and centred around 0.(mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not.The above diagram shows that the residuals are distributed about mean = 0.**

**Question: 5 Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bike**

**Answer:** we have top 3 features are

1.atemp - coefficient :0.4597

2.yr - coefficient : 0.2320

3. Light rain_Light snow_Thunderstorm- coefficient -0.2359

**General Subjective Questions**

**Question 1: Explain the linear regression algorithm in detail**.

Answer: inear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values.Linear Regression is the most basic form of regression analysis.Regression is the most commonly used predictive analysis model. Linear regression is based on the popular equation “y = mx + c”. It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x).

In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

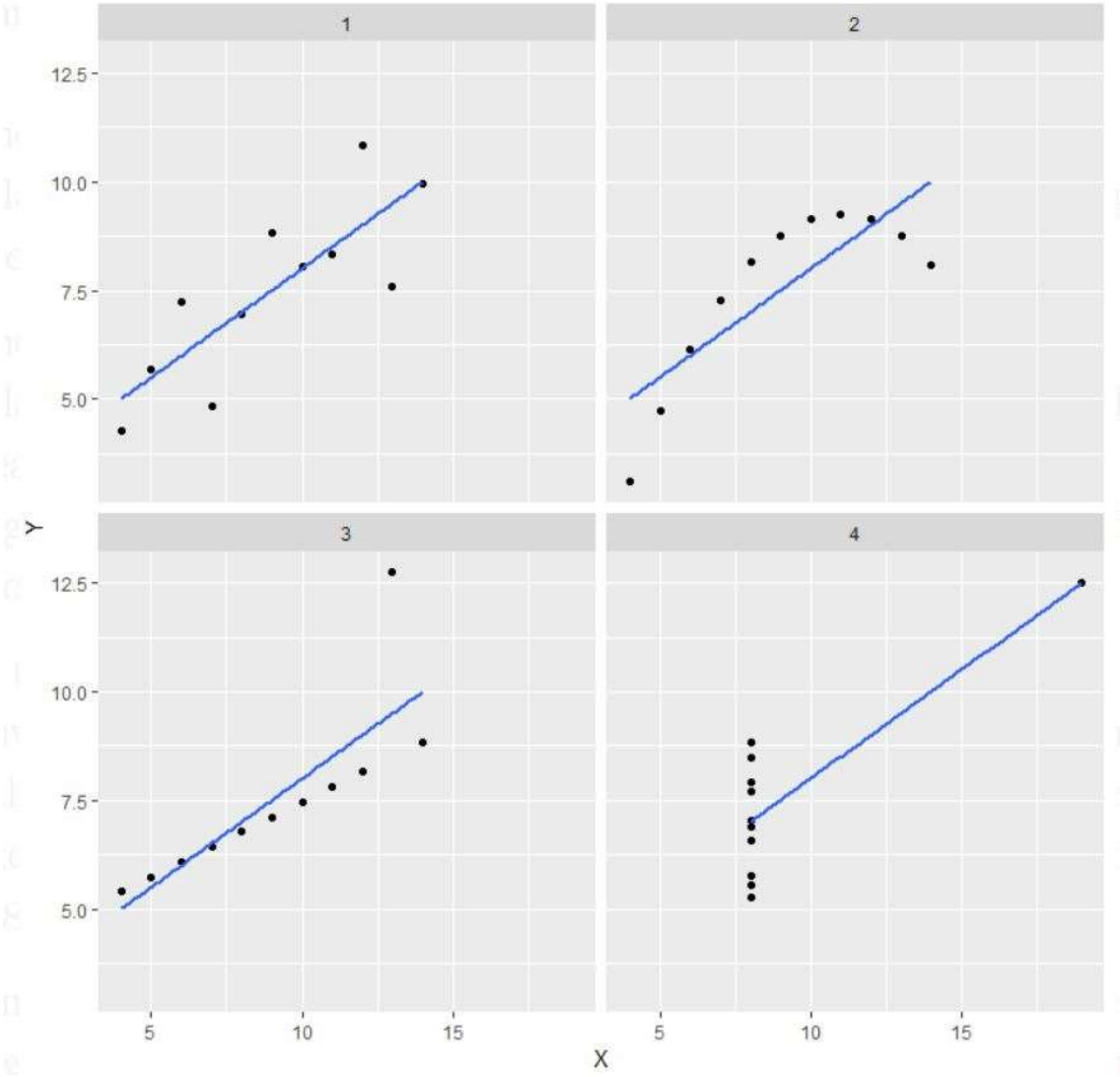
Regression is broadly divided into simple linear regression and multiple linear regression.

1. Simple Linear Regression : SLR is used when the dependent variable is predicted using only one independent variable.

2. Multiple Linear Regression :MLR is used when the dependent variable is predicted using multiple independent variables. The equation for MLR will be: β1 = coefficient for X1 variable β2 = coefficient for X2 variable β3 = coefficient for X3 variable and so on… β0 is the intercept (constant term).

**Question 2: Explain Anscombe's quartet in detail.**

**Answer:** Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph.It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on




**Note:** It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

**Explanation of this output:**

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

**Question 3: What is Pearson's R.**

**Answer:** Pearson's r is a numerical summary of the strength of the linear association between the variables.It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data? r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association

**Question 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

**Answer:** Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

● Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

● Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

**Answer:** VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.(VIF) =1/(1-R_1^2 ). If there is perfect correlation, then VIF = infinity.Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

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

**Answer:** A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.It is used to compare the shapes of distributions.A Q-Q 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.