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

***Ans.*** *Following are the effects for categorical variables on the dependent variable:*

***a.*** *Company should focus on expanding business during Fall, Summer and Winter*

***b.*** *September month has shown great demand.*

***c.*** *It has been observed that the demand for bike rentals had gone up from 2018 to 2019. So we can say that it will go up once the situation gets normal post Covid*

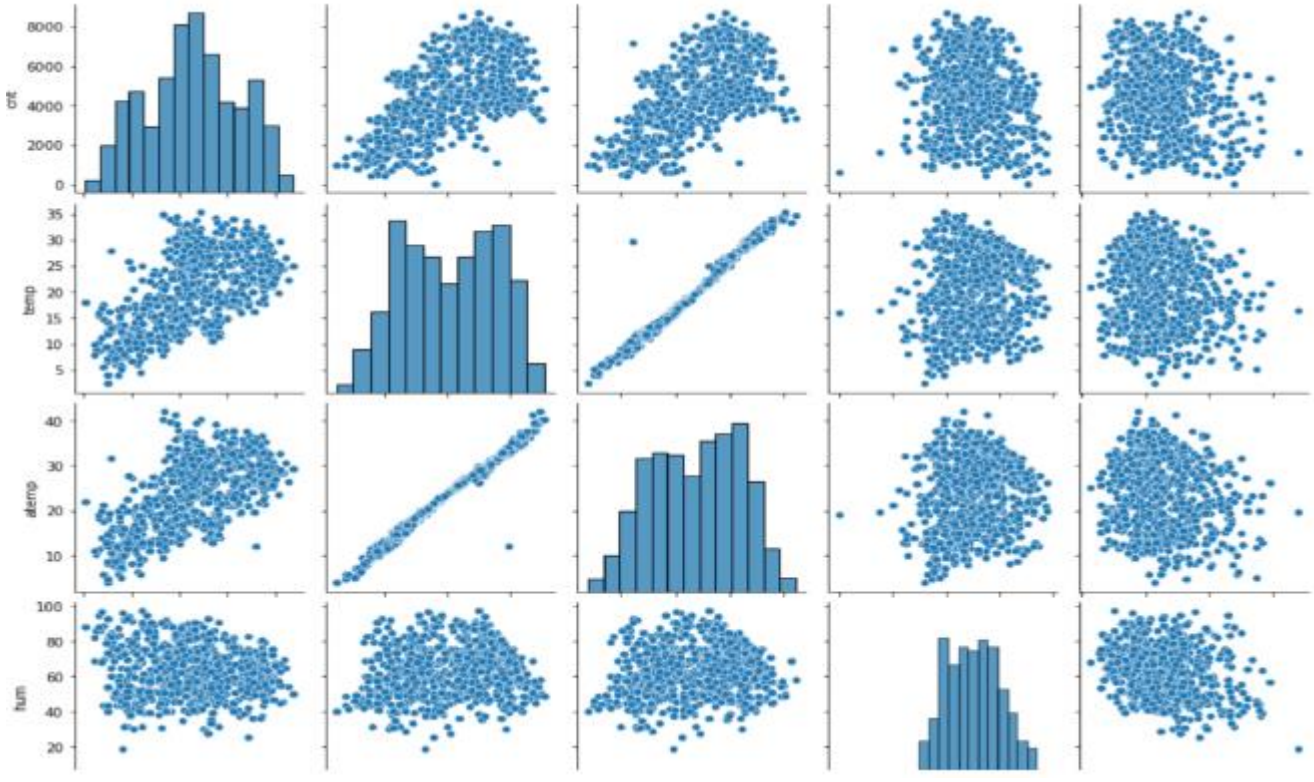
***d.*** *There would be fewer bookings during Bad and no demand in severe weather conditions.*

***e.*** *There is no much demand during the holidays*

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

*Ans: The drop_first =True is important as it helps in reducing the extra column created during creation of dummy variable. Dropping the column is important because the importance or value of that left over variable can be found by remaining variables. So to avoid redundancy we are dropping a column. This helps in the column to become linearly independent.*

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



*The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.*

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

1. *Linearity: The relationship between the independent variables and the dependent variable is assumed to be linear. This means that the effect of the independent variables on the dependent variable is a straight-line relationship.*
2. *No autocorrelation: The error terms (residuals) in the regression model should not be correlated with each other. This assumption ensures that there is no systematic pattern or relationship between the residuals.*
3. *Normality of error: The error terms should follow a normal distribution. This assumption allows for valid statistical inference and hypothesis testing.*
4. *Homoscedasticity: The variance of the error terms should be constant across all levels of the independent variables. In other words, the spread of the residuals should be consistent throughout the range of predicted values.*
5. *Multicollinearity: The independent variables should not be highly correlated with each other. Multicollinearity can make it difficult to interpret the individual effects of the independent variables on the dependent variable.*

*Validating these assumptions is crucial for ensuring the reliability and accuracy of the linear regression model's results. If any of these assumptions are violated, it may affect the model's performance and the validity of the inferences drawn from it.*

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

*Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature, year and season.*

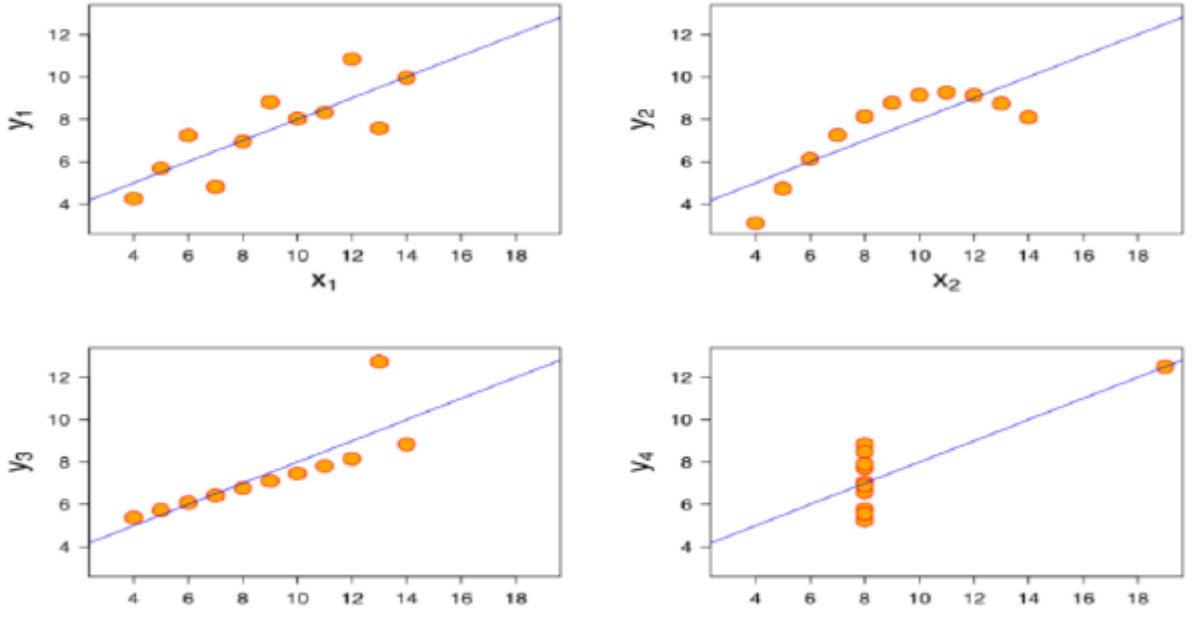
### 1. Explain the linear regression algorithm in detail ?

*Linear regression is a form of predictive modeling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the*

*independent variable. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.*
*A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line and the best fit line should have the least error.*
*In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line*
*for the data points.*

## 2. Explain the Anscombe's quartet in detail?

*Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It was constructed to illustrate the importance of*
*plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.*



● *1 st data set fits linear regression model as it seems to be linear relationship between X and y*
● *2 nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.*
● *3 rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.*
● *4 th data set has a high leverage point means it produces a high correlation coeff.*
*Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning model.*

### 3. What is Pearson's R?

*In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.*

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

*Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.*
*Difference between Normalizing Scaling and Standardize Scaling:*
*1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.*
*2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.*
*3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.*

*4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.*

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

*VIF(VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:*
*A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.*
*A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.*

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

*Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.*
*Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.*
*QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.*
*Importance of QQ Plot in Linear Regression :*
*In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.*

*Advantages:*
- *It can be used with sample size also*
- *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*