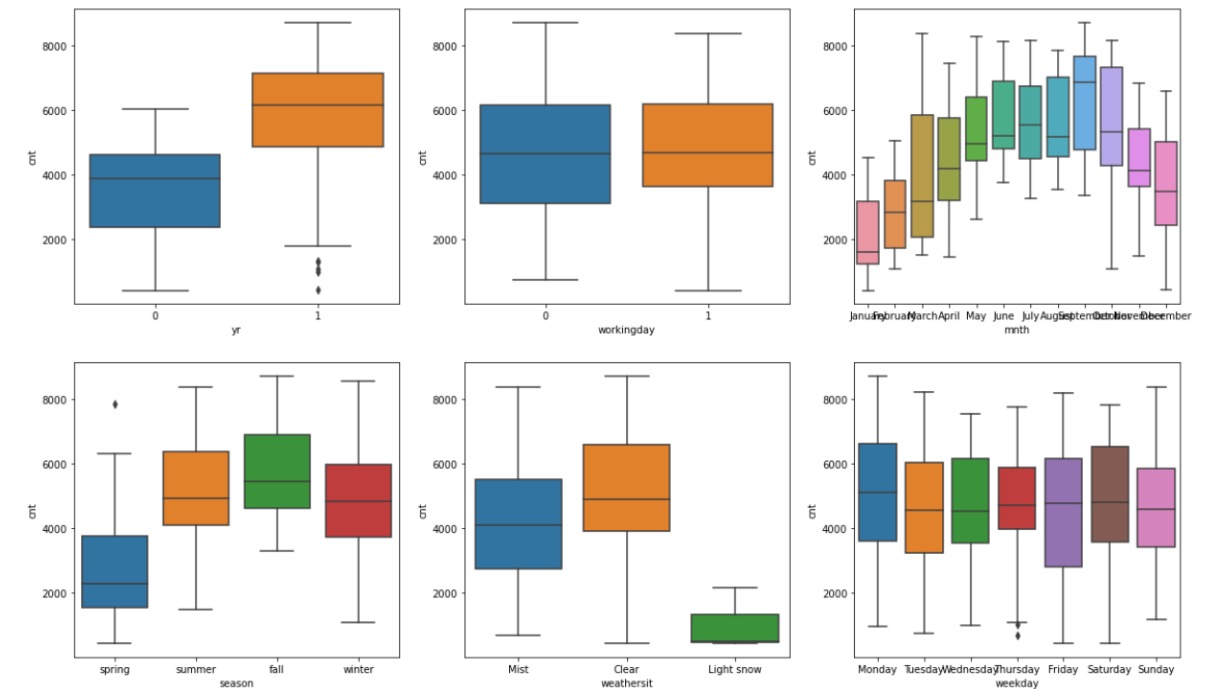


## I. Assignment-based Subjective Questions

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

Answer :



From these graph , we can see that

- ➔ Year 2019 has large counts
- ➔ No much difference in for working days and holidays
- ➔ June , July has maximum count records
- ➔ “Fall” season has large records for count
- ➔ No significant difference for weekdays. But weekends and adjacent days (Monday , Friday) has large records.

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

Answer : The reason we use `drop first = true` is, we need  $n-1$  dummy variables for a feature with  $n$  levels . Excluding this one we will have  $n$  dummies which is not the desired case

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

Answer : “temp” and “atemp” has the highest correlation with the target variable “cnt”

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

Answer : I have validated the below assumption we made for ideal linear regression model

- I. Target and predictors have a linear relationship with them
- II. There is no multi collinearity in our data frame
- III. Our residuals are homoscedastic. They are standard normally distributed with equal variances

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

Answer : “yr”, “Month\_September”, “Spring\_”

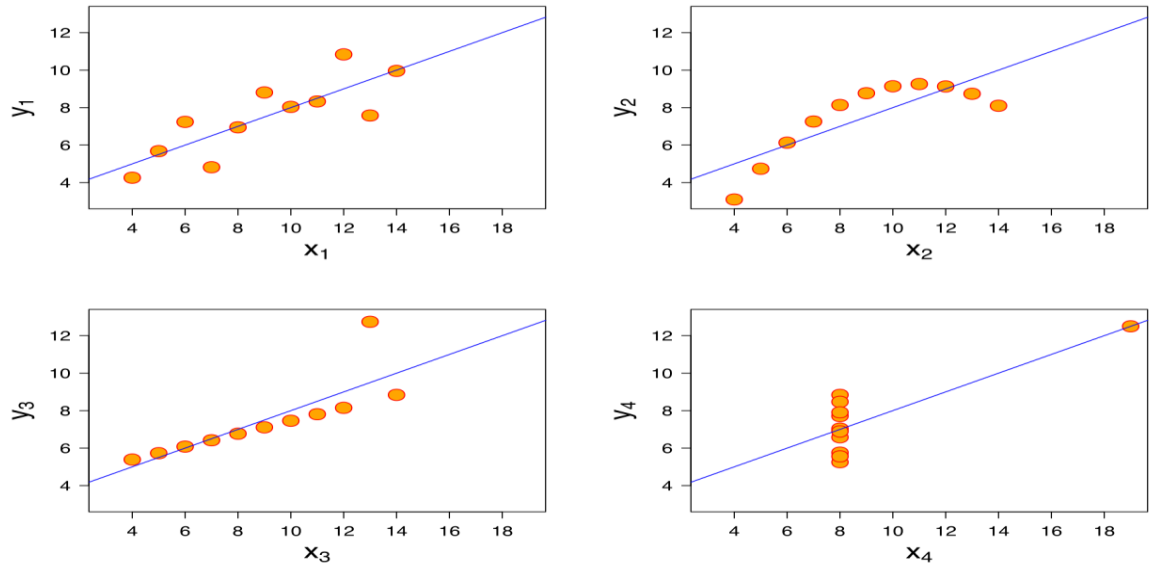
## **II. General Subjective Questions :**

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

Linear regression is a process of predicting the target variable by assuming a linear relationship between the target and predictor variable.

- a. First, we will perform data cleaning and EDA
- b. Once this is done, then the data will be split into train and test (70: 30 ratio)
- c. Then features are eliminated by the constraint( p value  $< 0.05$  , VIF  $< 5$  )
- d. These features are eliminated manually or automatic(Recursive feature elimination)
- e. After then a model is built and trained using training data
- f. We have our linear equation by this time
- g. We'll test the accuracy and efficiency of the model by its performance in test data

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



Answer : These four plots are identical when we interpret using mean , median , mode.

However we can entirely different distributions and behavior in the graphs.

If we observe them they have,

1. The first x1,y1 plot could be predicted using linear regression
2. The first x2,y2 plot is less likely to be predicted by linear models as it has very less linearity
3. The x3,y3 might look comforting , but the outlier effects the whole model. Removing it can suffice to good.
4. The x4,y4 doesn't have any linear relationship with the axes, but it can fool us with good correlation coefficient.

So we should be more careful in visualizing the data before processing it for the model.

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

Answer : Pearson's correlation coefficient (or) Pearson's R is a number which lies between -1 and 1 .

This is an index which uses to measure the correlation between two variables.

$$r = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

$r$  = Pearson Correlation Coefficient

$x_i$  = x variable samples

$y_i$  = y variable sample

$\bar{x}$  = mean of values in x variable

$\bar{y}$  = mean of values in y variable

For eg : If  $R=1$ , This indicates that X,Y are positively correlated , if X is increased , then Y is more likely to increase too.




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

Answer :

The goal of scaling is to have the entire numerical features under one scale. This helps us to analyze them using common plots within them . If scaling is not done, then the features can be incorrectly prioritized for difference in magnitudes within them.

1. Normal scaling is a process of rescaling the data so that they can fit between 0 to 1

**Normalization Formula**


$$X_{\text{new}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$


2. Standardized scaling does not have any boundaries. The main goal of this one is to fit the mean of the data to zero
3. Standardized scaling is more likely opted if we know that the distribution is normal.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

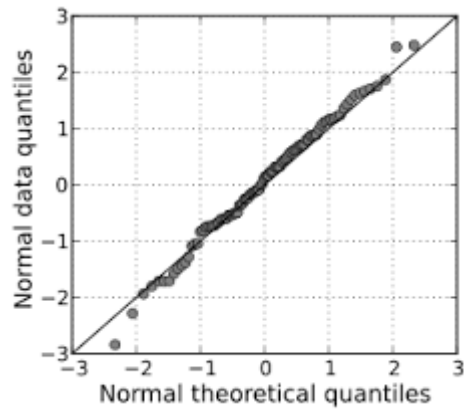
**Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

Answer :  $VIF = 1 / (1 - R_{\text{squared}})$

If VIF is infinity , it means that R square took value as “1” , that leads the denominator to zero! .This indicated ideal positive correlation between the features. Hence, we need to drop that feature to avoid multicollinearity.

**Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

Quantile – Quantile plot or Q-Q plots help us in determining the distribution of a variable using its data points.



Step 1 : Take the data points , split them into 10 (arbitrary) quantiles.

Step 2 : Choose a reference distribution (Normal distributed bell curve) . Feel free to choose the range

Step 3 : Divide the bell curve to exact number of quantiles

Step 4 : Plot all the data points of the variable in y axis (data points of the quantiles) , and the points of bell curve in X axis(data points of the quantiles)

Step 5 : If you see a linear behavior, it means the data points are normally distributed.

This is how the Q-Q plot helps us in determining the nature of distribution.