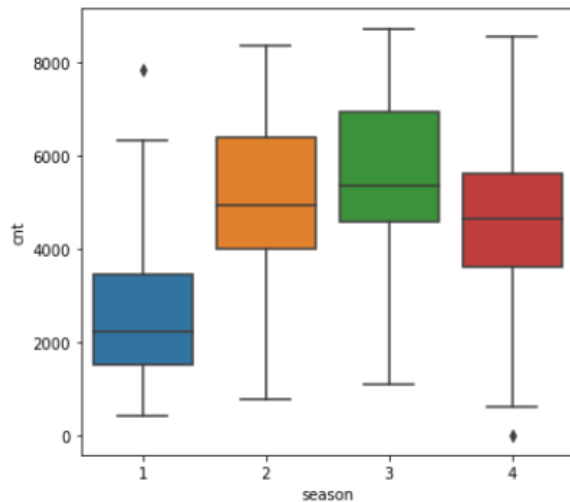


## Assignment-based Subjective Questions

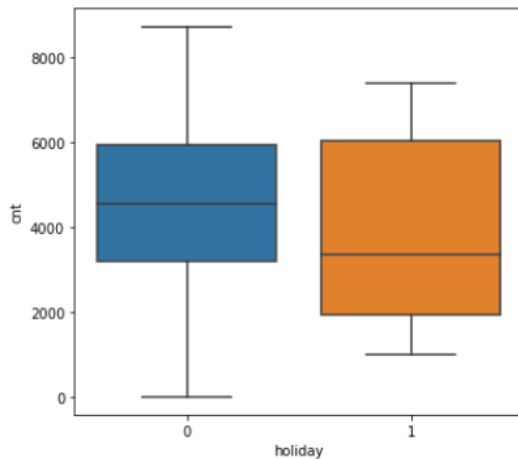
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans:

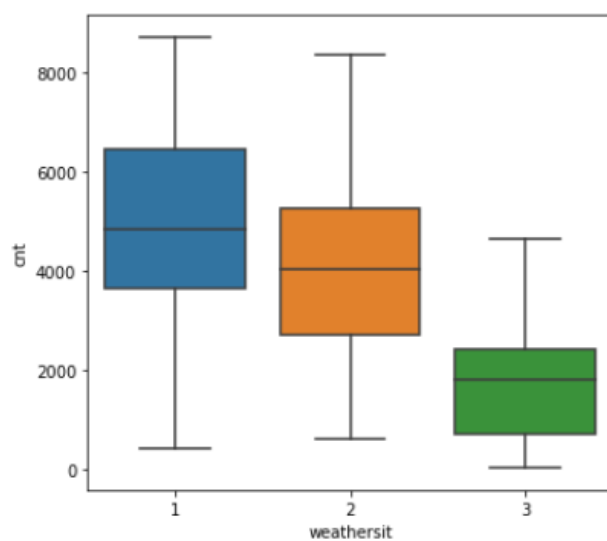
**Seasons:** Count is higher in fall (highest), summer and winter and less in spring



**Holiday:** When its is a non-holiday day, count is higher



**Weather:** when weather is clear count is highest, and in heavy rain there is no bike rented counted



2. Why is it important to use drop\_first=True during dummy variable creation? (2 mark)

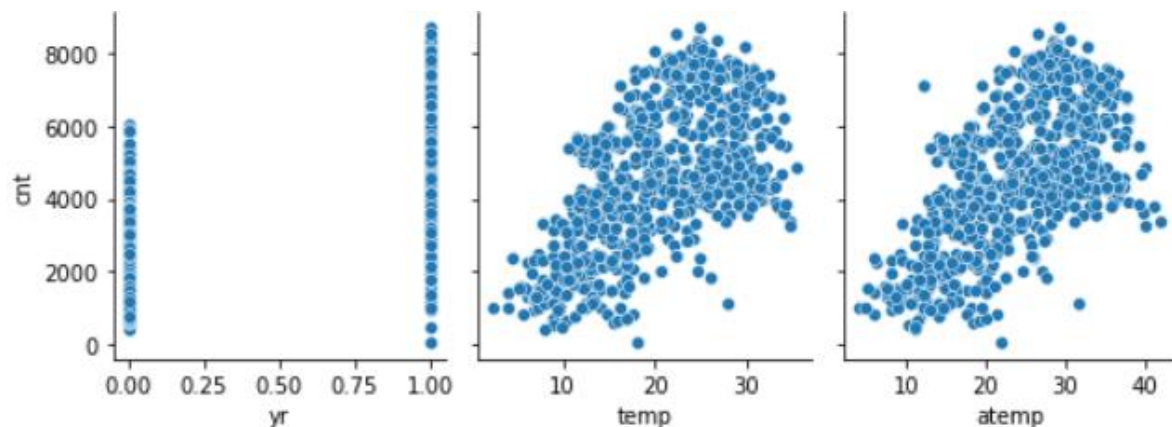
Ans: N number of categorical variables can be explained by N-1 dummy values

eg: Male is 1000 which can be explained by 000 as well, so we can drop the 1<sup>st</sup> column, reducing extra category column.

Gender	Male	Female	Transgender	Other
Male	1	0	0	0
Female	0	1	0	0
Transgender	0	0	1	0
Other	0	0	0	1

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

Ans: Temp and atemp has the highest correlation with target variable.



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

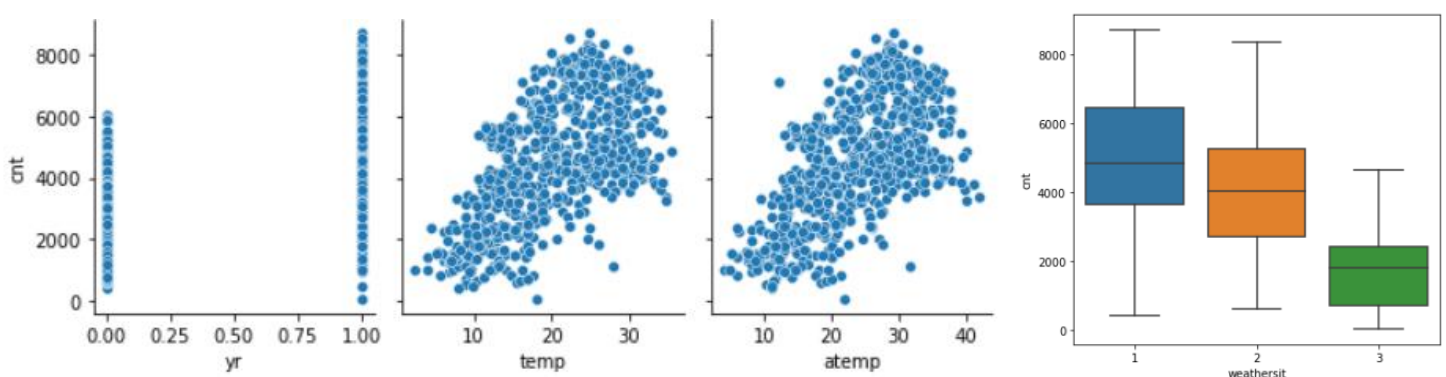
Ans: 1. VIF values to remove multicollinearity

2. Residual errors are homoscedastic

3. Residuals follow normal or almost normal distribution

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

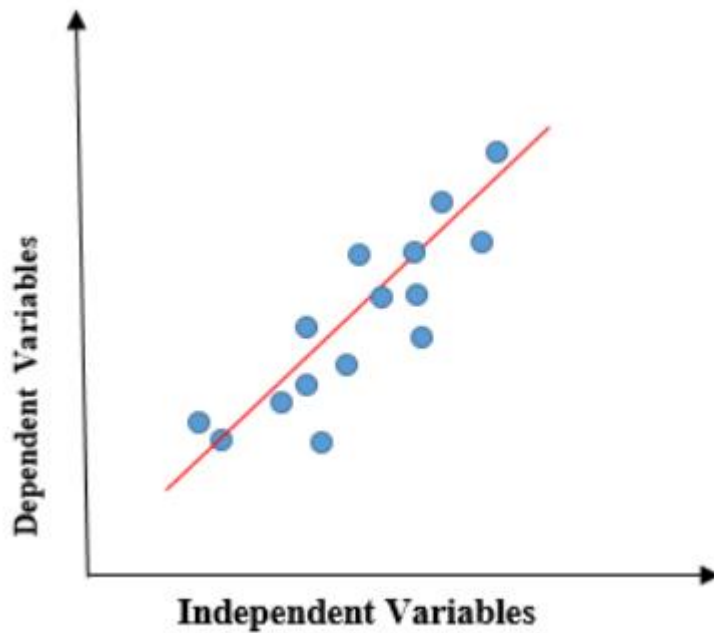
Ans: based on summary and univariate analysis of features, temp, year and weather has the influence on the count of bike sharing.



## General Subjective Questions

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

Ans: Linear regression is simple statistical regression method for predictive analysis to show the relationship between dependent and independent variables. It shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis). If there is only one single independent variable to define dependent variable then it is **simple linear regression** if there are more than one independent variable to define target variable then it is called **multiple linear regression**.



To calculate best-fit line linear regression uses a traditional slope-intercept form.

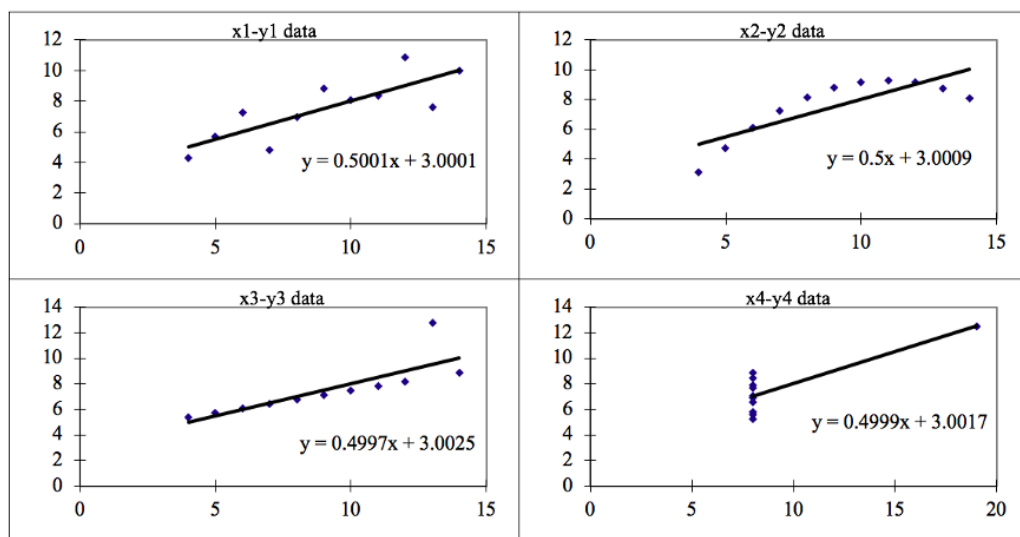
$$y = mx + b \implies y = a_0 + a_1x$$

$y$  = Dependent Variable.  
 $x$  = Independent Variable.  
 $a_0$  = intercept of the line.  
 $a_1$  = Linear regression coefficient.

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

Ans: It comprises four data sets that have almost identical simple descriptive statistics, but due to their peculiarity that fools the regression model if built. When plotted, they have very different distributions.

It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance of plotting the graphs** before analyzing and model building, and the effect of other **observations on statistical properties**.



The four datasets are as follow:

1. **Dataset 1:** this **fits** the linear regression model pretty well.
2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

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

Ans: Pearson's R or the bivariate correlation or Pearson correlation coefficient is a measure of line correlation between two sets of data. It is the ratio of covariance of two variables and product of their standard deviation. Its value lies between -1 to 1. Positive correlation means as one value increases other does too, 0 means values are unaffected and negative correlation means if one increases the other decreases.

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

Ans: scaling is the process to bring all numeric independent variables in same scale, it helps in predicting model better as after scaling coefficient of variable changes. Scaling doesn't affect the prediction in any way.

There are two ways of scaling:

Normalised scaling (min-max scaling) which convert dataset between 0 and 1

Standardised scaling which changes the value to centre them around the mean of the variable in a manner that new mean becomes 0 and standard variation becomes 1

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

Ans: infinite VIF means, very (too much) high value of VIF which shows that variable is totally colinear with other variables. Meaning it can be completely explained by other variables.

In mathematical terms when R-square is 1 then value of VIF is infinite:

$$VIF = 1/(1-R^2)$$

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

Ans: Quantile-Quantile (Q-Q) plot, is a graphical tool to help assess if a set of data comes from some theoretical distribution such as Normal, exponential or Uniform distribution. It helps to determine if two data sets come from populations with a common distribution.

In linear regression it helps to determine if train and test datasets are from same population if both these data sets received separately.