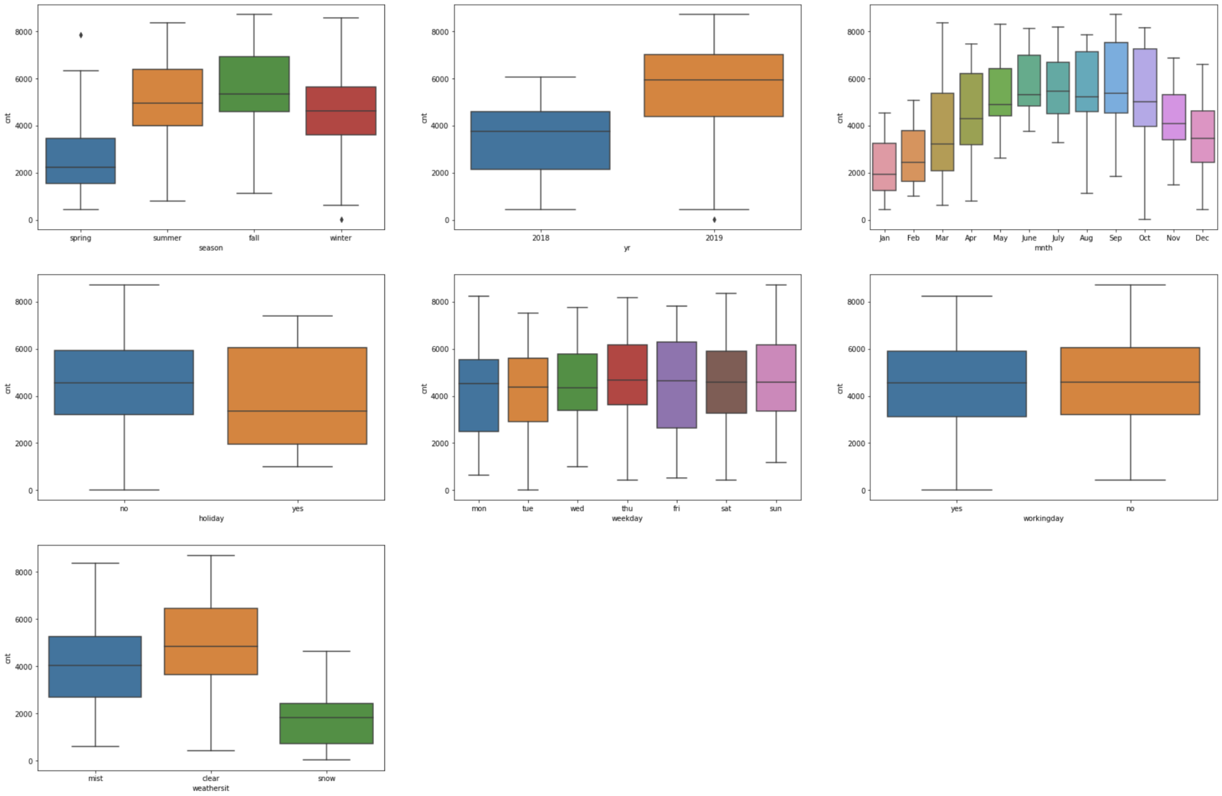
Assignment-based Subjective Questions:

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

The Categorical variables in the dataset are - season, yr, mnth, holiday, weekday, workingday, weathersit.

1. Significant increase in the demand for rentals is noted in 2019, compared to 2018.
2. Weekday and Working Day: The median demand for rentals is observed to be the same during weekdays and working days.
3. Significant decrease in the demand observed during holidays.
4. Significant decrease is observed during snow season, indicating the weather is uncomfortable.
5. During fall, we could notice the increase in demand. Summer and Winter have intermediate demand. Spring is seen to have less demand for rentals.



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

- It helps in reducing the extra column created during dummy variable creation.
- It helps in Reducing Multicollinearity.

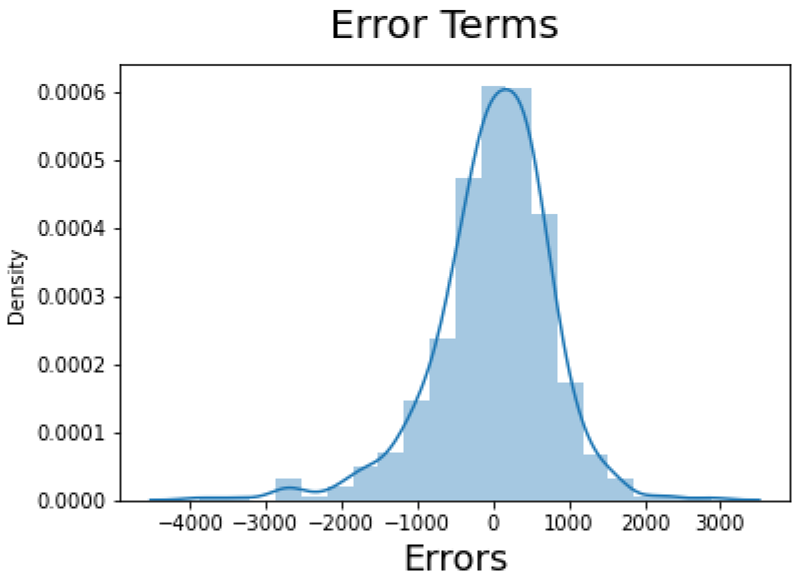
*3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?***

A: The variable registered has the highest correlation with the cnt variable. However, this variable is not known in advance.
Hence - If we consider the list of all independent variables, we could say <u>temp is highly correlated with cnt.</u>

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

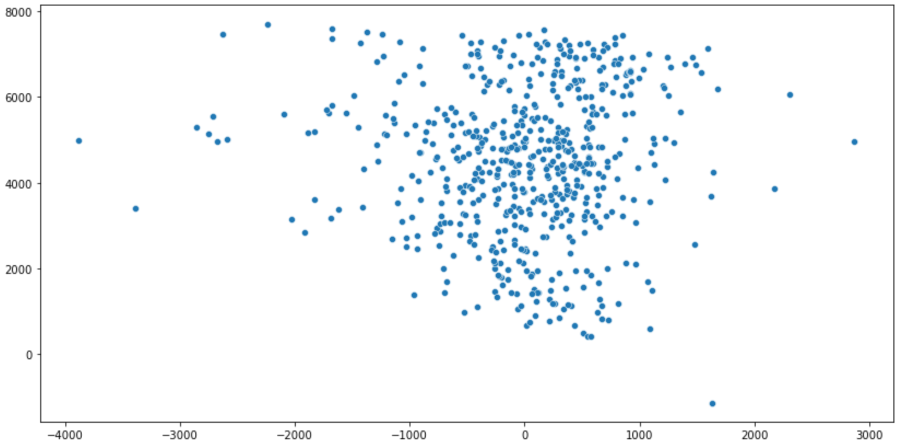
1. Linear relationship: By drawing a pairwise scatter plot / heatmap between all the numeric variables.

2. Verified that the error terms are normally distributed.



    a.

3. Multicollinearity -
    a. Since temp and atemp variables are highly correlated, removed atemp.
    b. Also verified the VIF at every iteration of RFE to ensure the variables are not correlated.
4. Verified that the error terms have constant variance

a. 

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

The top 3 features in the model are :

1. Temp
2. Weathersit_snow
3. Year

General Subjective Questions

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

A: Linear regression is a type of algorithm in supervised learning. The target variable is a continuous variable.

Here, we assume there is a linear relationship between independent and dependent variables.

We are trying to predict the value of the target variable by fitting the **best possible line** using a set of already known independent variables.

For fitting the best possible line, we try to reduce the sum of squared errors. There are two types of linear regression:

1. Simple linear regression - Here we have a single dependent variable and a single independent variable. We predict the value of a dependent variable using only one dependent variable.
2. Multiple linear regression - Here we have a single dependent variable and multiple independent variables.

The algorithms used widely are:

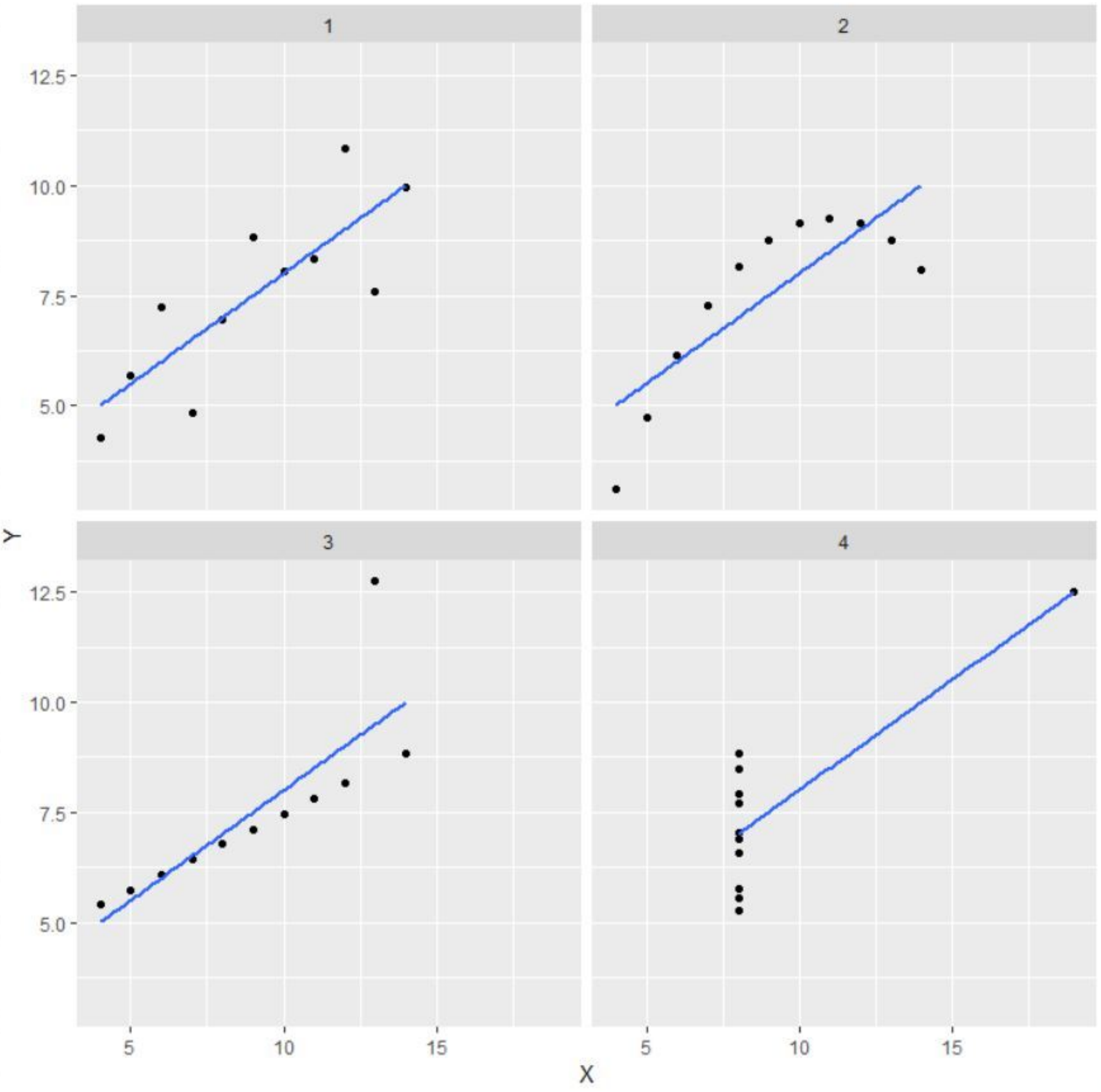
1. Ordinary Least Squares - This is most commonly used in the context of simple linear regression.

2. Gradient Descent -
    a. Here, we start at a random coefficient in the graph.
    b. Then we calculate the derivative of the cost function w.r.t beta parameters.
    c. Then compute the coefficient using the formula: coefficient = coefficient - alpha * delta.
    d. Alpha is the learning rate of the algorithm.
    e. This process is repeated until the cost function converges to minimum.

2. Explain the Anscombe's quartet in detail.

It consists of four different datasets, which have the same statistical properties. But they have a different distribution when plotted on a graph.

This is used to show the importance of graphing data, and to see the effect of outliers.



Graph -1 : it indicates a simple linear relationship between the variables x and y.
Graph -2: The variables x and y are having a non- linear relationship.
Graph -3: It shows the points are having a perfectly linear relationship. However, we could see that the outlier has impacted the fit of the line.
Graph -4: when one high-leverage point is enough to produce a high correlation coefficient.

3. What is Pearson's R?

A: A Pearson R is used to indicate if two variables are having a linear relationship. The values of pearson coefficients are between -1 and 1.

If pearson_r = close to 1, variables are positively correlated.

If pearson_r = close to -1, variables are negatively correlated.
If pearson_r = 0, variables do not have a relationship.

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

A: Scaling is a part of data pre-processing step applied to the independent variables to get the values in a specific range.

If the values are too large, it would take a large amount of time for the beta values to reach the optimum range.

Also, By scaling the parameters, it only affects the coefficients but not the evaluation metrics like R Squared, F statistics etc.

Normalised Scaling: Here, we replace the values with their z-scores. It brings the data to a standard normal distribution with mean 0 and standard deviation 1.

Min Max Scaling: It brings the data in the range 0 to 1. It preserves the shape of the original distribution.

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

VIF (Variance inflation factor) tells us how the independent variables are related to each other.
If VIF > 10, then the variable is considered to be collinear with others, and we drop the variable.

The formula for VIF = 1/ (1- R2).
If VIF = infinity, then 1-R2 = 0 => R2 = 1

Hence, This indicates that there is a multicollinearity within the independent variables.

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

A Q-Q plot tells us if two data sets are having the same distribution. This is a plot of quantiles of the first data set against the other.

Here, we draw the quantiles of the two data sets, and draw a line at 45 degrees angle. If both the distributions are similar, points will fall on that line.

In the context of linear regression, it can be used to see if test and train are having the same distribution, same location and scale.