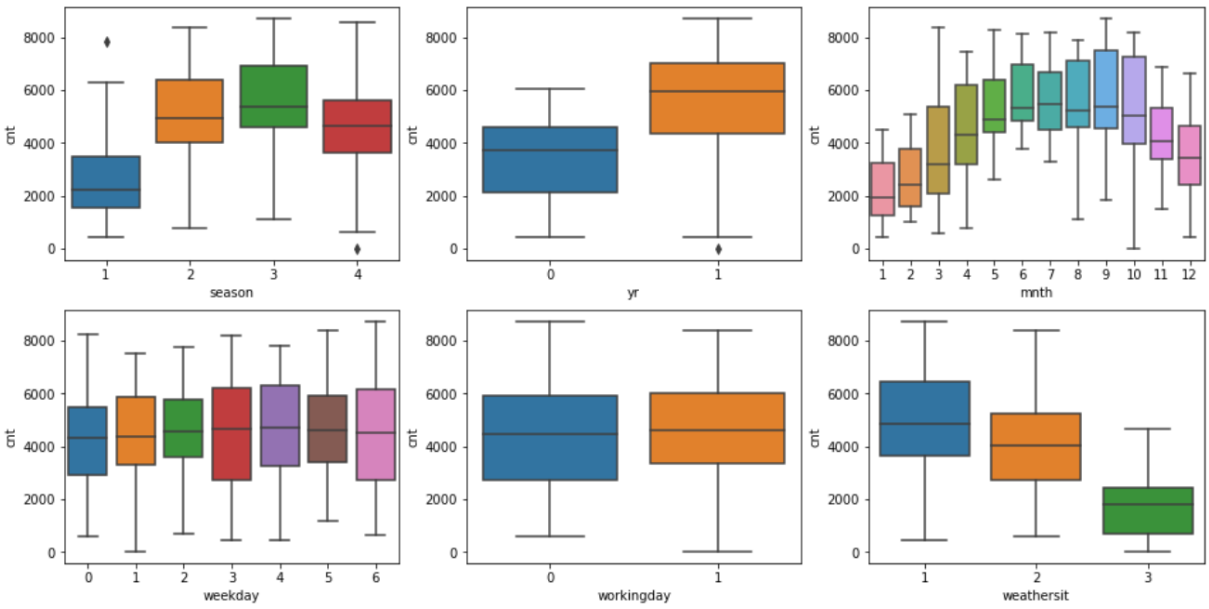
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

[Ans]:



1. Fall season rental is highest among other season
2. Year 2019 has higher bike rental than Year 2018
3. September Month Rental is higher that other months
4. Saturday has highest rental
5. Working day rental is very little high compared to weekend and holiday
6. Bike demand is highest in when weather is Clear

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

[Ans]:

When you have a categorical variable with say 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels. For a variable say, 'Relationship' with three levels namely, 'Single', 'In a relationship', and 'Married', you would create a dummy table like the following:

Relationship Status Single In a relationship Married Single 1 0 0 In a relationship 0 1 0 Married 0 0 1

But you can clearly see that there is no need of defining three different levels. If you drop a level, say 'Single', you would still be able to explain the three levels.

Let's drop the dummy variable 'Single' from the columns and see what the table looks like:

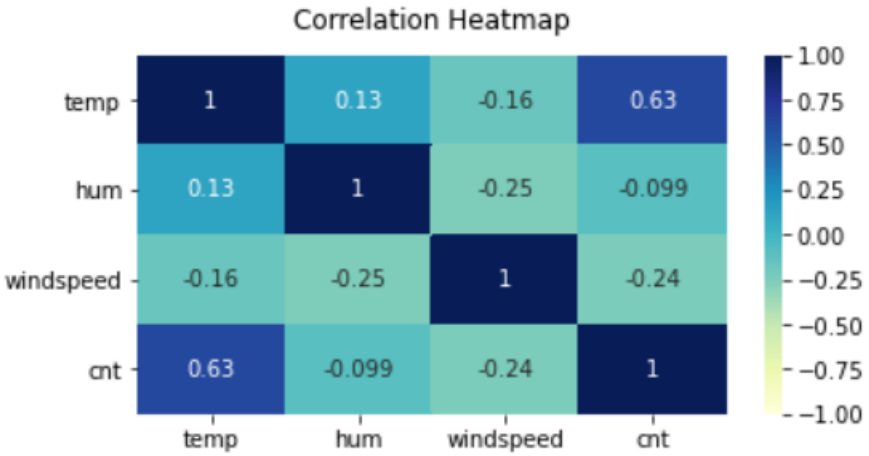
Relationship Status In a relationship Married Single 0 0 In a relationship 1 0 Married 0 1

If both the dummy variables namely 'In a relationship' and 'Married' are equal to zero, that means that the person is single. If 'In a relationship' is one and 'Married' is zero, that means that the person

is in a relationship and finally, if 'In a relationship' is zero and 'Married' is 1, that means that the person is married.

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

[Ans]:

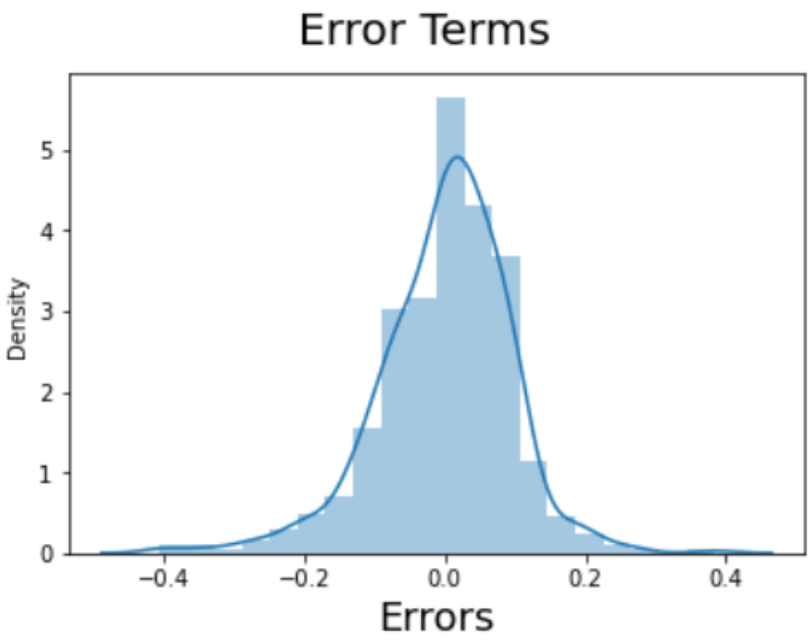


correlation between cnt and temp is high **(0.63).**

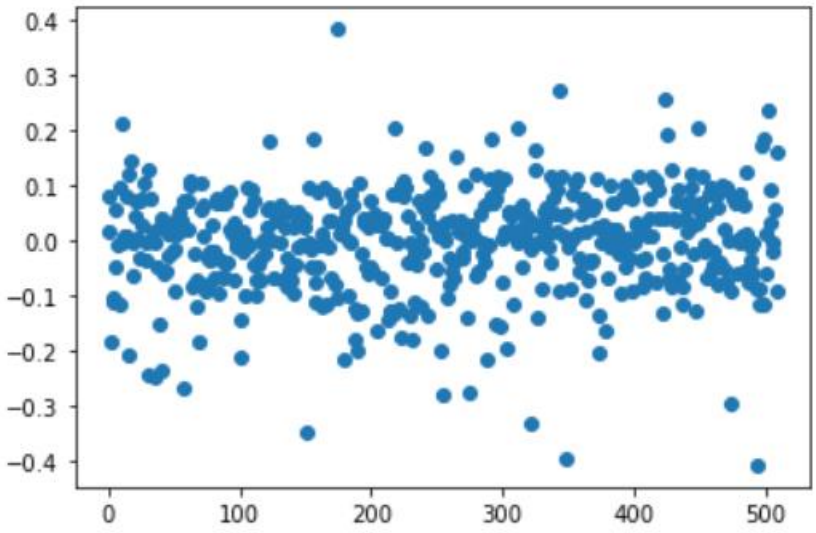
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

[Ans]:

1. Correlation between dependent and independent variable by scatter plot to see liner relation hip x and y. That's shows this problem statement is eligible for Linear regression.
2. Errors terms are normally distributed



3. Residual does not have any pattern.



4. Below VIF data shows there is no on multi collinearity.

	Features	VIF
0	const	20.68
1	temp	2.02
3	spring	1.64
5	July	1.28
6	Sep	1.10
2	windspeed	1.06
7	Light_Snow	1.04
8	Mist	1.04
4	2019	1.02

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

[Ans]:

1. Temperature in Celsius
2. year (2019)
3. month(September)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

[Ans]: Linear regression algorithm explain the relationship between to numerical data set if they have correlation at all. By using linear regression algorithm, you can predict value of y axis data from x axis data. This algorithm helps you to find best fit line into date set which can help to determine value of data in y axis.

The main idea behind linear regression is:

1. Use least squares to fit a line to the data
2. Calculate R2(r square)
3. Calculate P-value of R2.

Finding least squares:

First draw a line in data set and find **residuals** from data point and square them and add them. That call **sum of square of residual**. Keep repeating this step until you find the point which is having least sum of squares. Once you find least sum of squares fit the line and get linear equation(y=mx+c).

It will give two parameters

1. Y axis intercept
2. Slope

If the slope is **not zero** that indicates that you can predict value of some data on Y-axis from x-axis data. But how good this predication. For that you need to have R2.

Calculate R2

1. First take average of all data points in Y-axis and measure distance from mean to the data points and square it then add those square together. This call **ss(mean)**

   ss(mean)= square of (data-mean)

   Variation around the mean= square of (data-mean)/n (n is sample size)

2. Now go back to original plot and then Sum up squared residuals around the least-square fit

   This is called **ss(fit).**

   ss(fit)= square of (data-line)

   variation around fit=var(fit)= square of (data-line)/n

3. Once you get var(mean), var(fit) and var(mean) for x and y axis data lets calculate R2

   R2=var(mean)-var(fit)/var(mean)

   R2 tell you how much of variation in y axis data can be explain by x axis data.

   Example: var(mean)=11.1 and var(fit)=4.4

   So R2= 11.1-4.4/11.1=0.60….

   R2 =0.60 is nothing but 60%

   We can say that data in x axis explain 60% of variance in y axis data.

Calculate P-value of R2

1. P-value of R2 comes from something call “F”.

F=Variation in Y axis data explained by x axis data/The variation in y axis data not explained by x axis data

2. Now get the p-value from F.

Basically R2 squared needs be large and p-value needs to be small to have an interesting output

[This is created from StatQuest you tube Video]

2. Explain the Anscombe's quartet in detail.

[Ans]: Anscombe's quartet basically says that the importance of data visualization when datasets are statistically very close means their mean, median and correlation value etc.

Anscombe who was a statistician did an experiment of 4 data sets which is very close statistically when they visualize those data in x and y coordinates he observed

1. First data set is a perfect linear relation between x and y
2. The Second one shows curved line means non linear relation ship.
3. The Third one shows a perfect linear relation between x and y and with an outlier
4. fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Anscombe's theory basically tells you if a statical summary of datasets are the same that does not mean they are the same data set. Every dataset has its own property. So before taking any decision based on a statical summary, please visualize those data and take a decision are they same or not.

3. What is Pearson's R?

[Ans]: Pearson correlation coefficient is known as Pearson's R. it describes the relationship between two variables. It indicates the relationship between the two variables. Pearson's R value is between -1 and 1.

1.Pearson's R value as 1, indicates that when one variable(x) increases then another variable (y) also increases .

2. Pearson's R value as -1, indicates that when one variable(x) increases then other variable(y) also decreases.

3. Pearson's R value as 0, indicates that there is no relationship between two variable. Basically, it is random

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

[Ans]: Scaling is nothing but transform variable values. scaling does not impact your model.

Why we need a Scale of variables:

1. Ease of interpretation. if you have many variables with different scales then the interpretation is difficult. if you have the same scale then easily you can explain the coefficient of one variable over the coefficient of another variable.
2. Computation is time-consuming because background Gradient descent algorithm is running. if the variable has a different scale then Algo will take high time.

Which factor will impact by Scaling?

It is important to note that scaling just affects the **coefficients** and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc. Even Model accuracy will not change

How to do scaling?
There are two major methods to scale the variables, i.e.

1. standardization
2. Min-Max scaling.

Standardization basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. Min-Max scaling, on the other hand, brings all of the data in the range of 0 and 1.

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

[Ans]: The Variance Inflation Factor (VIF) is a measure of collinearity among predictor variables within a multiple regression.

VIF (VIF= 1/1-R squared)

High VIF means multicollinearity is high

If R squared is 1 then VIF is infinite. R squared will be one

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

[Ans]:
Quantile-quantile plot (Q-Q) plot says that the data set sample came from a population where data is normally distributed. It is a graphical tool basically not very concrete evidence. It is map between sample quantiles vs quantile. The expected value is all the data points in Q-Q plot form moreover a straight line(45 degree plot)

Use cases for Q-Q plot:

1. If the samples data come from the same population or not
2. If sample data are normally distributed or not

The intercept and slope of a linear regression between the quantiles gives a measure of the relative location and relative scale of the samples.

This Q-Q plot indicates that if datapoints closer to the line that means they are from same population