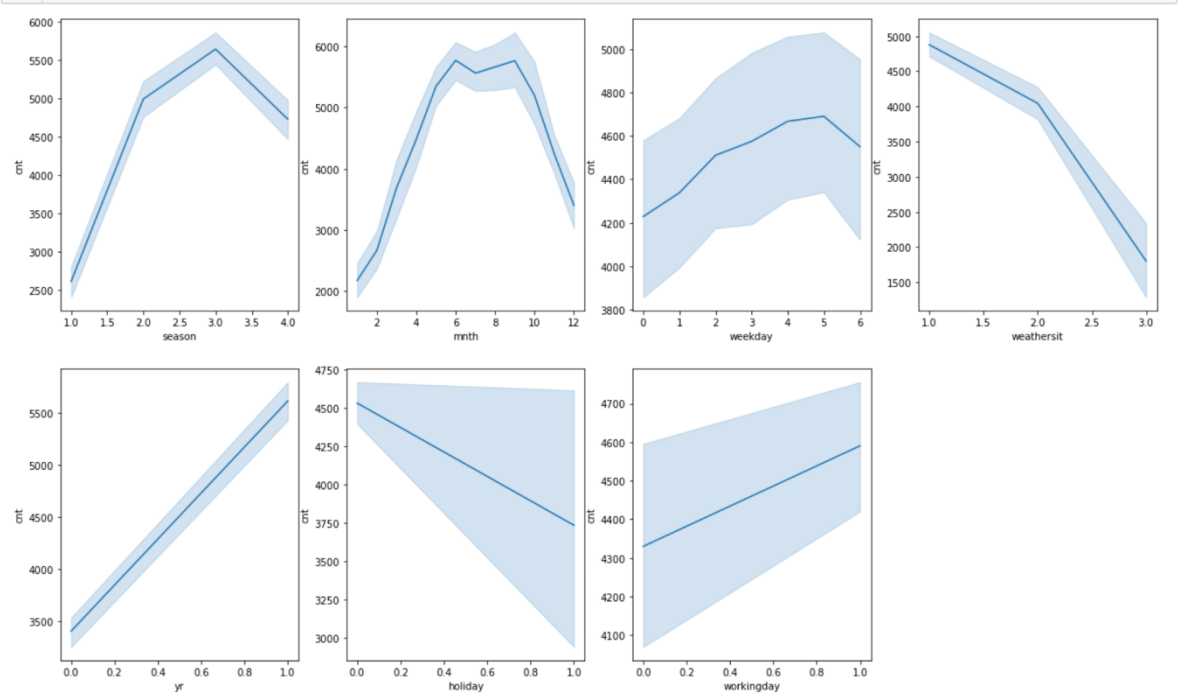
Question: - **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

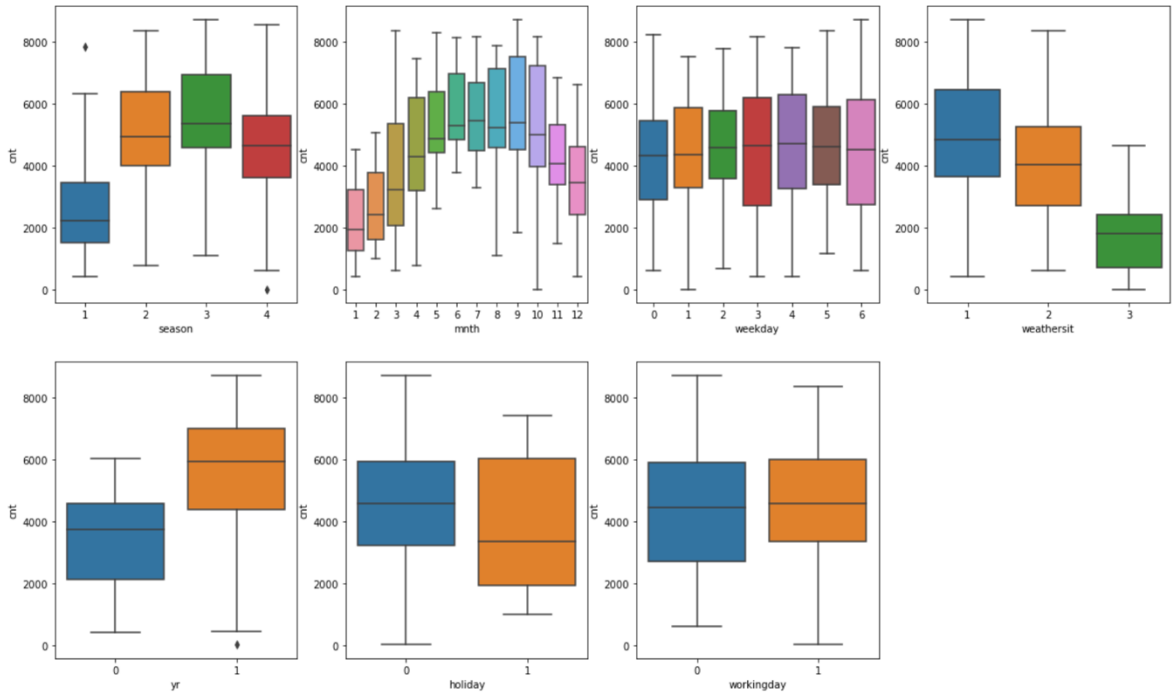
Answer**: -** While performing our analysis we found that there were following categorical variables

- **season, mnth, holiday, weathersit, yr**

we performed two visualization one's box plot and the other as line chart, here are the findings

- **season**:- Bike demand is high in summer and fall, and it lowers down in winter while spring has the lowest of the bike demand
- **weathersit**:- When the weather is bad (Heavy Rain, Ice Pallets, Thunderstorm) then there is no bike demand, while when the weather is clear the bike demand is highest and decreases linearly when weather situation is mist and cloudy and is lowest when weather situation is light snow and light rain
- **yr**:- Bike demand is increasing with a steep curve, with the time passing. Essentially that means more the company is becoming older it is becoming popular and thus increasing the bike demand
- **mnth**:- We see that the demand started to increase from second month, and it is max in **6,7,8** month.
- **holiday**:- Demand is decreasing when there is a holiday





Question:- **Why is it important to use drop_first=True during dummy variable creation?**

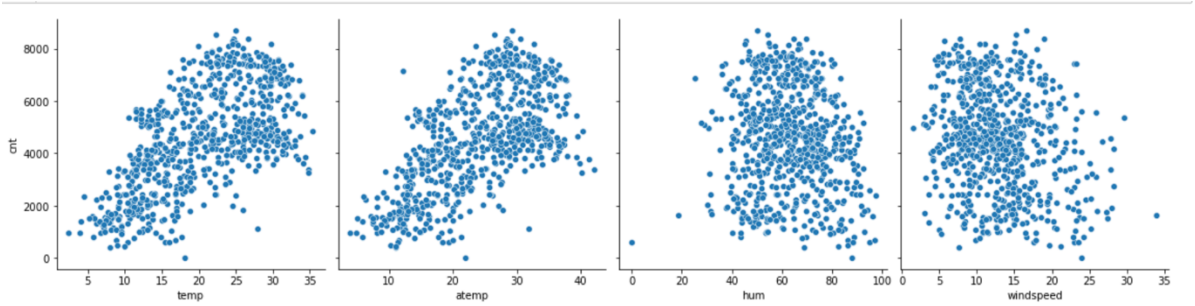
Answer :- If we don't drop the first values while creating dummy variable. we will have columns that are corelated with each other and we have to deal with problems of multicollinearity.

For example :- a column with n categories will create n dummy variables wherein the same information can be conveyed by n-1 dummy variables.

If we don't drop it then we may introduce additional variable and that may affect our model, this affects models adversely where cardinality is small.

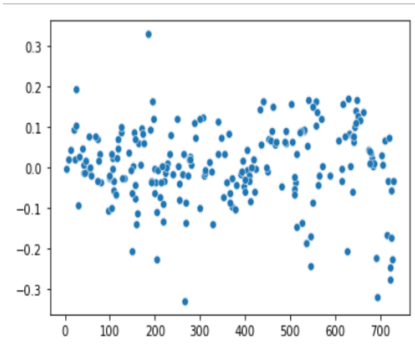
Question:- **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

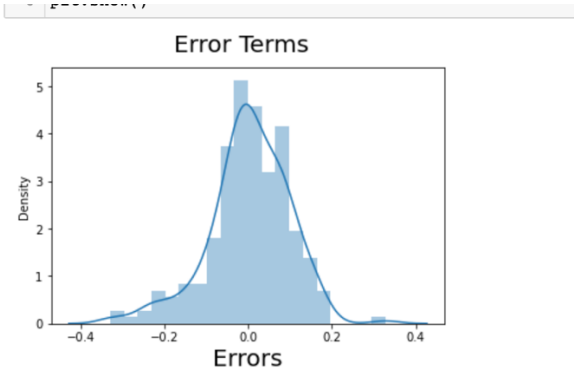
Answer:- Looking at the pair plot we can clearly say that **temp, atemp** are highly corelated with the output variable



Question:- **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Answer:- Once we had our model created we made predictions and calculated our error terms. Then we validated if the error terms are normally distributed by plotting a dist plot of the error terms. We also checked whether the error terms are not showing any pattern by plotting a scatter plot of the error terms. Also, we validated our assumption that error terms have standard variance which is also clearly visible from the dist plot of error terms.





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

Answer:- The top three features are the one that has the highest coefficient which is

- temp 0.491508
- yr 0.233482
- Light_Snow -0.285155 (this is weather situation number 3, according to data dictionary)

Question :- **Explain the linear regression algorithm in detail ?**
Answer :- Linear regression is an algorithm that helps in predicting the continuous variables (numerical data). It's part of supervised learning method in machine learning. Linear regression does not use labels or clustering. It is the most simplest and widely used algorithm in the predictive analytics.

Linear regression is based on the line's equation which is

$Y = mx+c$ (here, c is constant and the m is the coefficient of x)

Linear regression can be applied where the output/dependent variable is a continuous variables and independent variables can be anything.

In linear regression we assume that there is a linear relationship between the predictor and the output variables. And the error terms are normally distributed with no pattern in error terms.

Error terms are essentially the difference between the original value and the predicted value.

Linear regression uses a technique called best fit line, to show a relationship between the predictor and the output variables. In best fit line approach we use least sum of square so that the errors are very minimum when defining the relationship.

In Linear regression the output is essentially an outcome of the coefficient and independent variable, like the line equation (y=mx+c) where y is an outcome of m*x, here m can be treated as a coefficient and x as an independent variable.

Linear regression is further divided in two categories

- **SLR (Simple linear regression)**
  - Simple linear regression is a technique where we just have one independent variable and that is used to find the output variable
  - Formula (Here Beta is the coefficient)
    - Output Variable(y) = Beta * Independent variable (X)+ constant

- **MLR (Multiple linear regression)**
  - MLR is a technique where we use multiple independent variables to predict the output variables
  - Formula (Here Beta_1, Beta_2, Beta_n are coefficient of respective independent variables)
    - Y=Beta_1*X_1+Beta_2*X_2+….+Beta_n*X_n+Constant

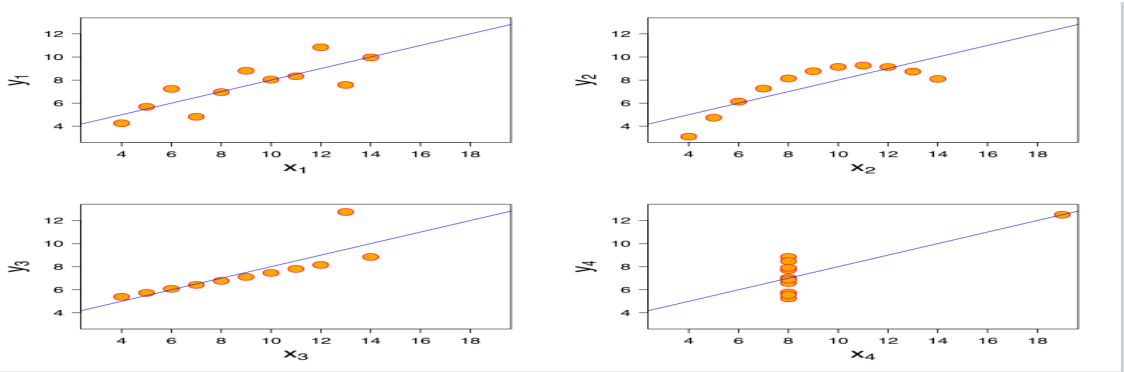
Question :- **Explain the Anscombe's quartet in detail.**
Answer :- Anscombe's quartet proves that statistics itself is not able to explain the data prope rly, and we need visualization to gain insights about the data. This is further explained by taki ng 4 dataset (hence quartet) which are statistically identical.

When these 4 dataset are plotted on a graph we can see how different these are from one another. This was done to demonstrate both the importance of graphing data before analyzing it, and the effect of outlier and other influential observation on statistical properties

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The above data has the same mean (9) for x and std_dev as (3.32) and same mean (7.5) and std_dev (2.03) for y for all the 4 dataset but when plotted on a scatter plot shows different characteristic as shown below.



- First plot shows linear relationship
- Second plot shows non-liner relationship
- Third plot shows linear relationship, except for one datapoint which seems to be an outlier
- Fourth plot shows that even one datapoint (high enough) can produce a high corelation coefficient

Question :- **What is Pearson's R?**
Answer :- Pearson's r also known as Pearson correlation coefficient (PCC). It is a measure of linear corelation between two data sets. It is the variance of two variables, divided by product of standard deviations. It's value is always in between -1 and 1. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

Formula for Pearson R is given as :- where

x_{i}	=	values of the x-variable in a sample
{x}_bar	=	mean of the values of the x-variable
y_{i}	=	values of the y-variable in a sample
{y}_bar	=	mean of the values of the y-variable

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

Question :- **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Answer :- Since most of the data that we work with while creating the model may not be on the same scale, and this may affect the coefficient calculation during our model building. Hence a need is there to bring all data on the same scale. In order to achieve this scaling is performed.

Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range, It also helps in faster convergence of gradient descent.

There are two types of scaling

- normalized scaling
- standardized scaling

normalized scaling is also known as min max scaling where the values are scaled in between 0 and 1. Since all the values are in between 0 and 1, this scaling strategy takes care of outliers

Formula :- X_new = (X - X_min)/(X_max - X_min)

Standardized scaling on the other hand is the transformation of features by subtracting from mean and dividing by standard deviation. the values are centred around the mean with a unit standard deviation. This means mean becomes zero and the resultant distribution has a unit standard deviation.

Standardization can be helpful in cases where the data follows a Gaussian distribution. In Standardized scaling the range can be anything hence outliers are not handled here

Formula :- X_new = (X - mean)/Std_Dev

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

Answer :- VIF or variance inflation factor essentially gives us a measure of how much collinearity is present among the variables that we are measuring.

Formula for VIF = 1 / (1 – R Square)

In case there is a perfect multi-colinearity such that the variable that we are measuring can be perfectly explained by other predictor variables, then we will get R square equal to 1.

Substituting this value in the formula will result VIF in infinite value.

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

Answer :- A q-q plot is is a probability plot, it's a plot of the quantiles of the first data set against the quantiles of the second data set. It helps us assess if the data has normal, uniform or exponential distribution.

If the two datasets being compared are similar, the points in the Q–Q plot will lie on the same line i.e y = x

If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line but not necessarily on the line y = x.

Q-Q plot can be used to find if the two datasets has similar distribution or similar tail behaviour. It also helps to identify whether two datasets come from populations with a common distribution.