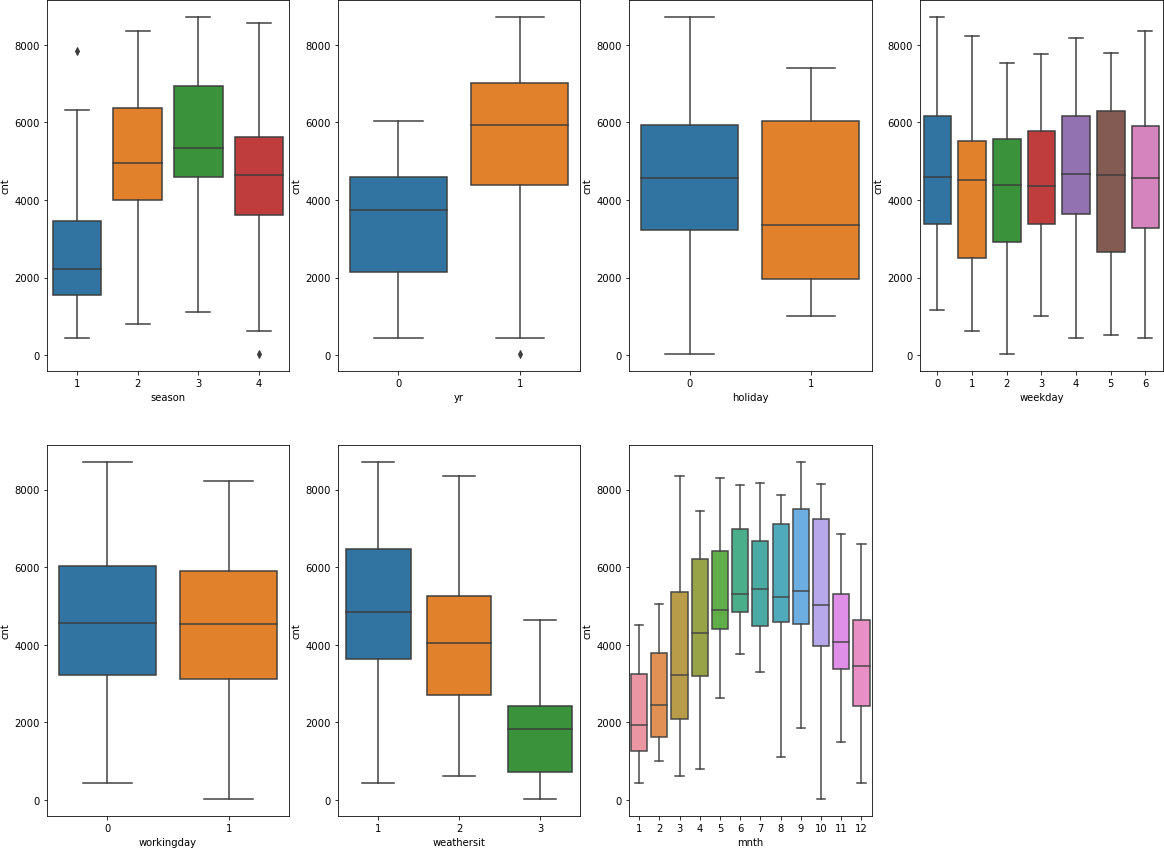
**Assignment-basedSubjectiveQuestions:**

# From your analysis of the categorical variables from the dataset, what could you infer abouttheireffectonthedependentvariable?(3marks)



The categorical variables in the dataset were **season**, **yr**, **holiday**, **weekday**, **workingday**, **weathersit**, and **mnth**. These were visualized using a boxplot (Fig. attached).  
These variables had the following effects on our dependent variable:

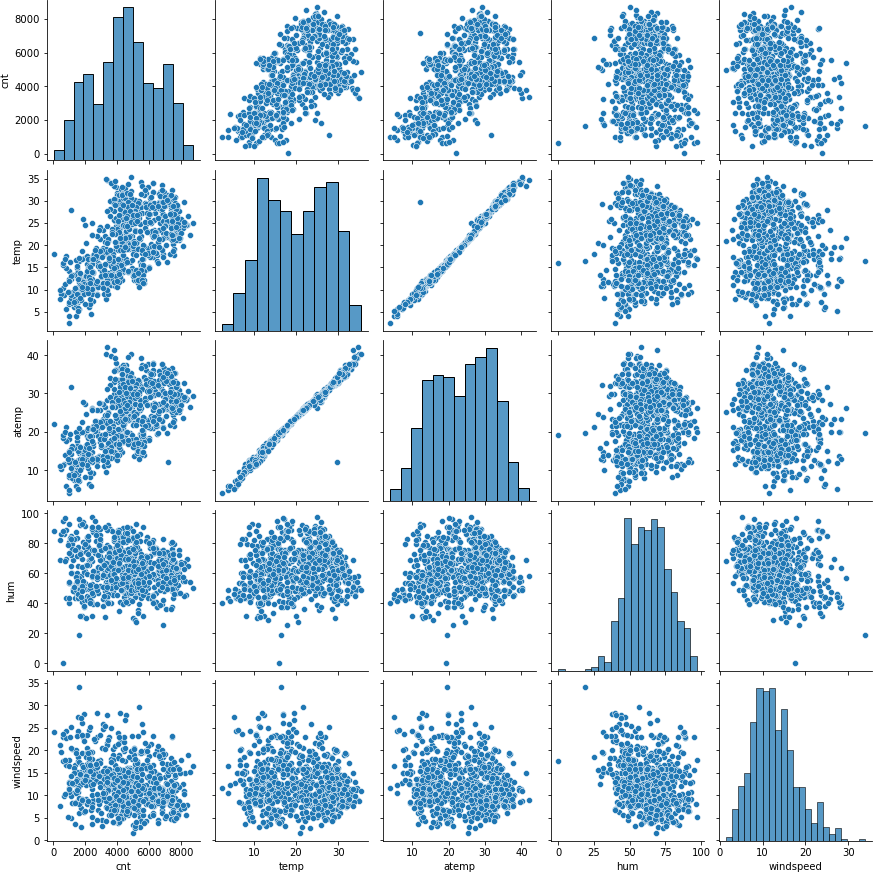
* **Season**: The box plot showed that the spring season had the least value of **cnt**, whereas fall had the maximum value of **cnt**. Summer and winter had intermediate values of **cnt**.
* **Weathersit**: There are no users when there is heavy rain or snow, indicating that this weather is extremely unfavorable. The highest count was seen when the **weathersit** was 'Clear, Partly Cloudy'.
* **Yr**: The number of rentals in 2019 was more than in 2018.
* **Holiday**: Rentals reduced during holidays.
* **Mnth**: September saw the highest number of rentals, while December saw the least. This observation is in accordance with the observation made in **weathersit**. The weather situation in December usually involves heavy snow, which might have caused the rentals to drop.
* **Weekday**: The count of rentals is almost even throughout the week.
* **Workingday**: The median count of users is constant almost throughout the week.

# Whyisitimportanttousedrop\_first=Trueduringdummyvariablecreation?(2mark)

Any variable with ‘n’ levels can be represented with ‘n-1’ dummy variables. If we know that there are 7 known levels and if we label 6 with dummy variables then the remaining one has to be the 7th one and we don’t need it to be encoded into dummy variable. Coming to the importance, it helps in reducing that extra column created while creating dummy variables and therefore results in reduction of the correlations among dummy variables.

# Looking at the pair-plot among the numerical variables, which one has the highest correlationwiththetargetvariable?(1mark)

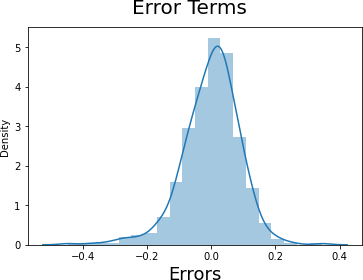
Using the below **pairplot**, it can be seen that **"temp"** and **"atemp"** are the two numerical variables that are highly correlated with the target variable (**cnt**).



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

The following tests were done to validate the assumptions of linear regression:

1. **Linearity**: First, linear regression needs the relationship between the independent and dependent variables to be linear. We visualized the numeric variables using a **pairplot** to check if the variables are linearly related or not. Refer to the notebook for more details.
2. **Normality of Residuals**: Secondly, residuals' distribution should follow a normal distribution and be centered around 0 (mean = 0). We validated this assumption by plotting a **distplot** of the residuals and checking whether they follow a normal distribution. The diagram below shows that the residuals are distributed around mean = 0.



**Multicollinearity**: Thirdly, linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the **VIF** (Variance Inflation Factor) to get a quantitative idea about how much the feature variables are correlated with each other in the model. Refer to the notebook for more details.

# Based on the final model, which are the top 3 features contributing significantly towards explainingthedemandofthesharedbikes?(2marks)

Thetop3featuresare:

1. temp
2. yr
3. weathersit\_Light Rain

**GeneralSubjectiveQuestions**

# Explainthelinearregressionalgorithmindetail.(4marks)

**Linear Regression** is a type of supervised machine learning algorithm that is used for the prediction of numeric values. **Linear Regression** is the most basic form of regression analysis and is the most commonly used predictive analysis model.

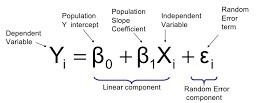
Linear regression is based on the popular equation **y = mx + c**. It assumes that there is a linear relationship between the dependent variable (y) and the predictor(s)/independent variable(s) (x).

In regression, we calculate the best fit line, which describes the relationship between the independent and dependent variables. Regression is performed when the dependent variable is of continuous data type, and predictors or independent variables can be of any data type, such as continuous, nominal/categorical, etc. The regression method tries to find the best fit line that shows the relationship between the dependent variable and predictors with the least error.

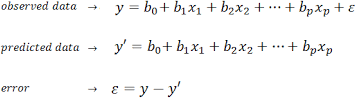
In regression, the output (dependent variable) is a function of an independent variable, the coefficient, and the error term.

Regression is broadly divided into two types: **Simple Linear Regression** and **Multiple Linear Regression**.

1. **Simple Linear Regression (SLR)**: SLR is used when the dependent variable is predicted using only one independent variable.  
   The equation for SLR will be:



**Multiple Linear Regression (MLR)**: MLR is used when the dependent variable is predicted using multiple independent variables.TheequationforMLRwillbe:

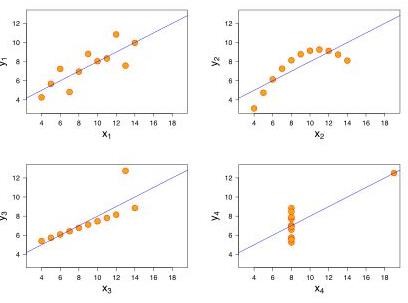


B1=coefficientforX1variable B2=coefficientforX2variable

B3=coefficientforX3variableandsoon… B0 is the intercept (constant term)

# ExplaintheAnscombe’squartetindetail.(3marks)

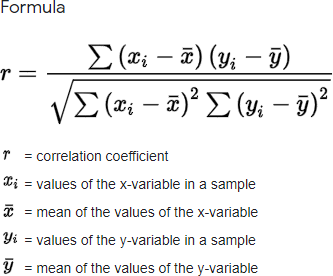
**Anscombe’s Quartet** was developed by statistician **Francis Anscombe**. It includes four datasets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



* + - Thefirstscatterplot(topleft)appearstobeasimplelinear relationship.
    - Thesecondgraph(topright)isnotdistributednormally;whilethereisarelationbetween them,it’s not linear.
    - In the third graph (bottom left), the distribution is linear, but should have a different regressionlineThecalculatedregressionisoffsetbytheoneoutlierwhichexertsenough influence to lower the correlation coefficient from 1 to 0.816.
    - Finally,thefourthgraph(bottomright)showsanexamplewhenonehigh-leveragepointis enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

# WhatisPearson’sR?(3marks)

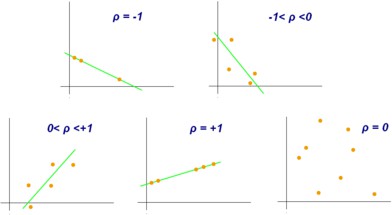
**Pearson's r** is a numerical summary of the strength of the linear association between two variables. Its value ranges from -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us: **“Can we draw a line graph to represent the data?”**



Ascanbeseenfromthegraph below,

r=1meansthedataisperfectlylinearwithapositiveslope

r=-1meansthedataisperfectlylinearwithanegativeslope r = 0 means there is no linear association



# What is scaling? Why is scaling performed? What is the difference between normalized scalingandstandardizedscaling?(3marks)

**Feature scaling** is a method used to normalize or standardize the range of independent variables (or features) in the data. It is performed during the data preprocessing stage to handle varying values in the dataset. If feature scaling is not done, machine learning algorithms tend to weigh larger values higher and consider smaller values as lower, irrespective of the units of the values.

* **Normalization** is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any specific distribution of the data, such as **K-Nearest Neighbors** and **Neural Networks**.

**Standardization**, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

### ****VIF - Variance Inflation Factor****

The **VIF** gives an idea of how much the variance of the coefficient estimate is being inflated due to collinearity. If there is perfect correlation, then **VIF = infinity**. It provides a basic quantitative measure of how much the feature variables are correlated with each other, making it an extremely important parameter for testing our linear model.

Where **R²** is the R-squared value of the independent variable that we want to check. This value shows how well the independent variable is explained by other independent variables. If the independent variable can be perfectly explained by other independent variables, it will have perfect correlation, and its R-squared value will be equal to 1. In this case, **VIF = 1 / (1 - 1)**, which gives **VIF = 1 / 0**, resulting in **infinity**.

The numerical value for **VIF** tells you (in decimal form) what percentage the variance (i.e., the standard error squared) is inflated for each coefficient. For example, a **VIF of 1.9** means that the variance of a particular coefficient is 90% larger than what you would expect if there was no multicollinearity (no correlation with other predictors).

### ****A rule of thumb for interpreting the variance inflation factor:****

* **1** = Not correlated.
* Between **1 and 5** = Moderately correlated.
* Greater than **5** = Highly correlated.

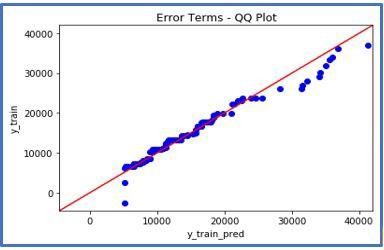
# What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A **Q-Q plot** (Quantile-Quantile plot) is a plot of the quantiles of the first dataset against the quantiles of the second dataset. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. The Q-Q plot is used to answer the following questions:

* **Do two datasets come from populations with a common distribution?**
* **Do two datasets have common location and scale?**
* **Do two datasets have similar distributional shapes?**
* **Do two datasets have similar tail behavior?**

BelowarethepossibleinterpretationsfortwodatasetsusingaQ-Qplot:

1. Similardistribution:Ifallpointofquantilesliesonorcloseto straightlineatanangleof45degree from x -axis
2. Y-values<X-values:Ify-quantilesarelowerthanthex-quantiles.



1. X-values<Y-values:Ifx-quantilesarelowerthanthey-quantiles.

