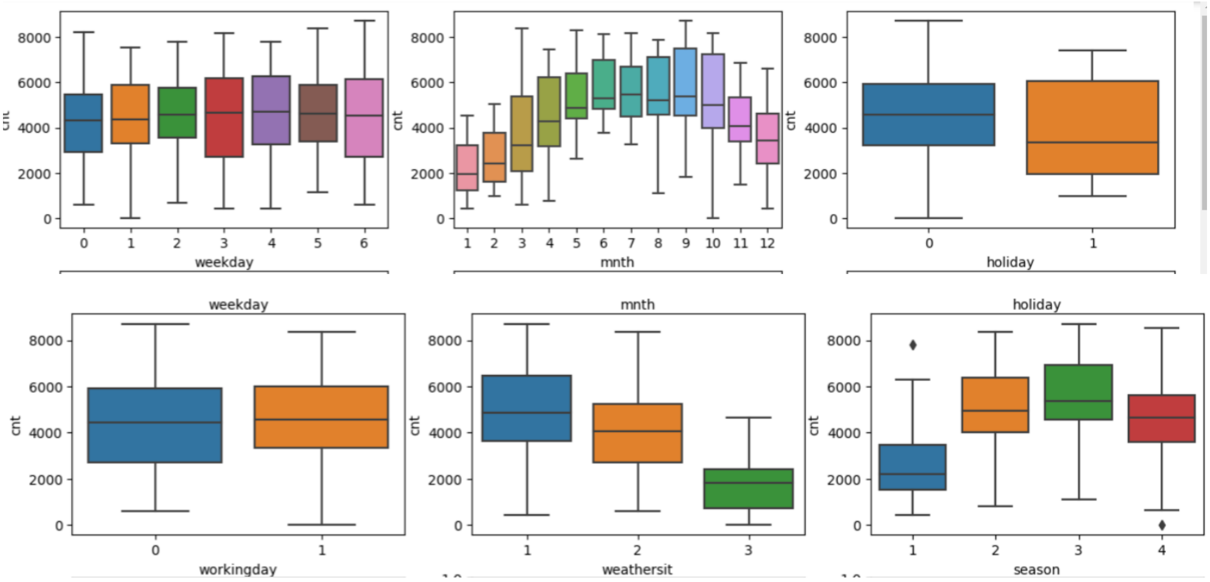
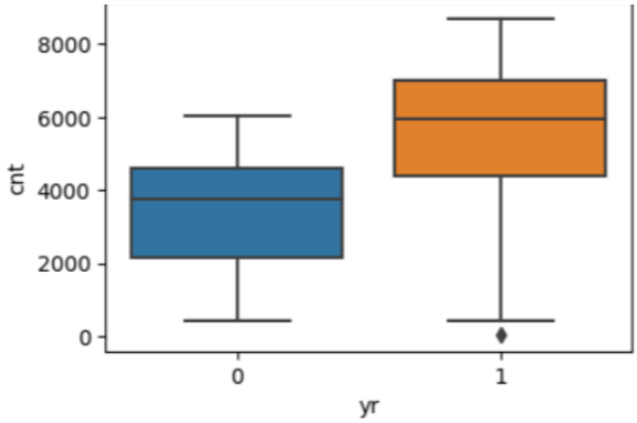
Assignment-based Subjective Questions

Ans 1) According to the dataset , the inference has been recorded in the python notebook. Pasting a screenshot of the same.





1. If we see cnt v/s mnth the count is high for the June and July months and a drop in the latter months.
2. For seasons 2 and 3 which are summer and fall.
3. During holidays less people are renting bike than non-holiday days ...And many other inferences we can draw

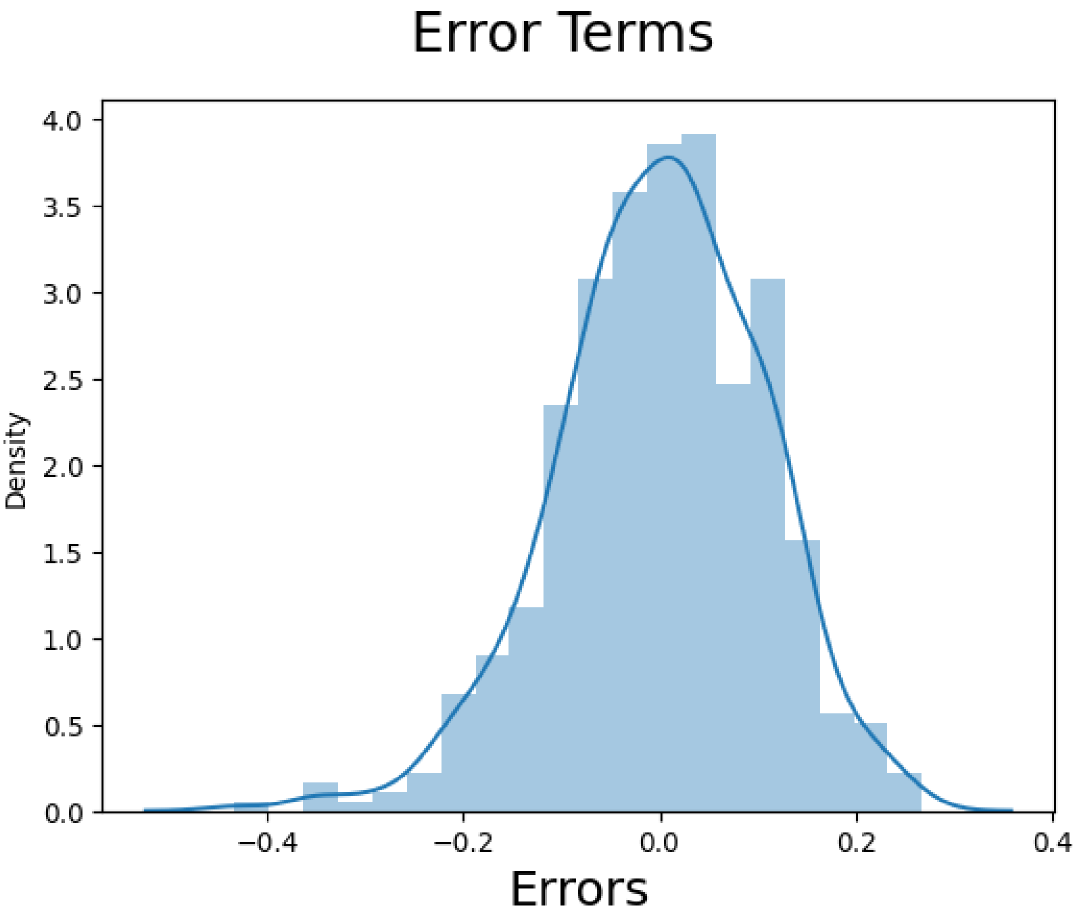
Ans 2) The categorical variables can have n number of categories which we define as levels. According to encoding we require n-1 variables to define these levels.

But using automation it generates n number of levels i.e get_dummies method provides more columns that that is required. Manually you can allocate the variables according to your need, but it is not always possible to use manual. So, we can use both automation and manual work

After generating through this method, we can drop_first= True which drops one of the first columns (from the ones which have been generated) and reduces to n-1 levels.

Ans 3) Looking at the pairplot the highest correlation is between temperature and the target variable i.e. the count.

Ans 4) The assumption can be validated by using the histogram of the residual errors which clearly indicate a normal distribution with error terms centred around 0 i.e. their mean is 0.



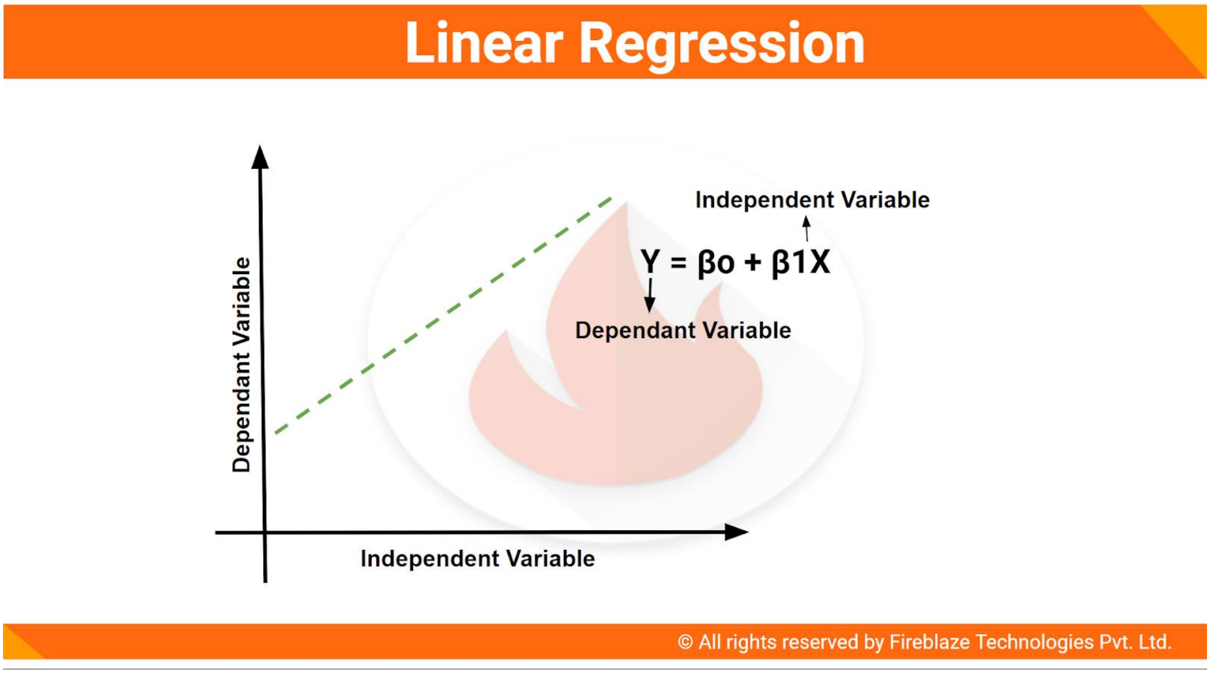
Ans 5) The top 3 features explaining the demand for shared bikes are temperature(temp), year(yr) and humidity(hum). This can be deduced comparing their coefficients in the linear model .

Temp: 0.549, yr: 0.2296, hum: -0.2286(this has a negative correlation as humidity increases demand decreases)

<u>General Questions</u>

Ans 1) Linear Regression is a machine learning algorithm based on supervised learning.Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables.

The equation in case of one independent variable is given as : Y= (coefficient)X and Y-intercept


Linear Regression

© All rights reserved by Fireblaze Technologies Pvt. Ltd.

It explains that a unit increase in X will increase the Y by coefficient unit. So as we increase the coefficient y will be impacted.

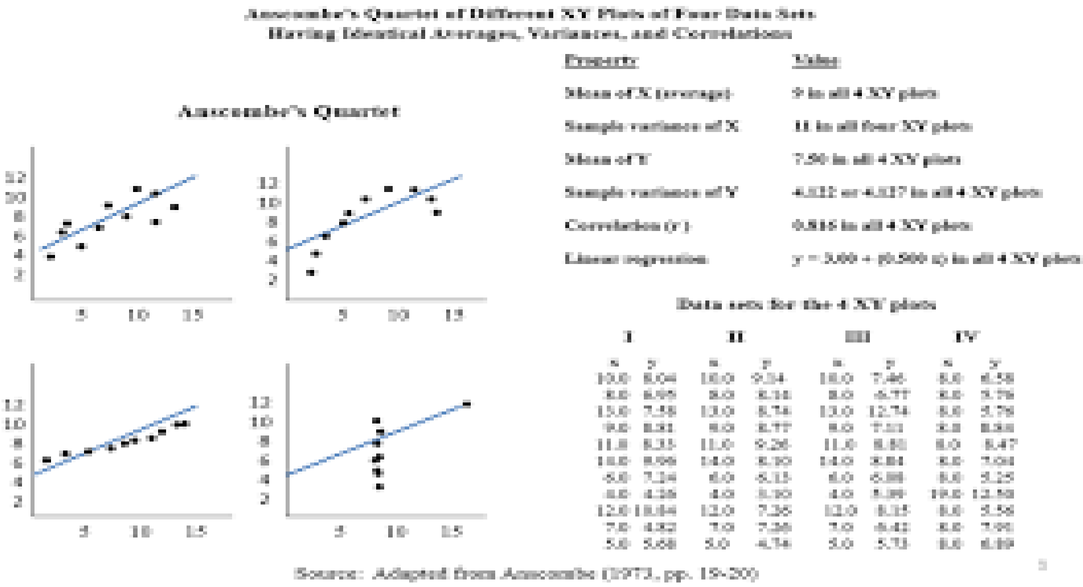
This diagram explains the equation for one independent variable. Here the coefficients are the important term that we need to find. But there can be multiple independent variables on which target is modelled. That is explained as **Multiple Linear Regression.**

Steps that we follow:

- Read the data, analyse it and understand it.
- Split the data into testing and training data set.
- Train the model. In this step you will receive the coefficients of the independent variables and summary of the model.
- Perform residual analysis on the predicted and actual values of the target variable. Also check if the assumptions hold true or not. The distribution should be a normal one. The error terms must be centred around 0 and should be normally distributed.
- Predict the values for the test set using the model and perform evaluations. Fid the r2 value and mean squared error.

Ans 2) Anscombe's quartet comprises four data sets that have nearly identical statistical data which includes their variance, mean but when plotted on a graph all 4 are quite different from each other.

This is one of the best examples of emphasizing on the importance of graphical representation, effect of outliers.



Anscombe's Quartet of Different XY Plots of Four Data Sets
Having Identical Averages, Variances, and Correlations

Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

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

Source: Adapted from Anscombe (1973, pp. 19-20)

5

In each panel, the Pearson correlation between the x and y values is the same, r = 0. 816. In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values. Here, the graph 2 doesn't even follow a proper linear regression model. Graph 4 is also completely different from all 3.

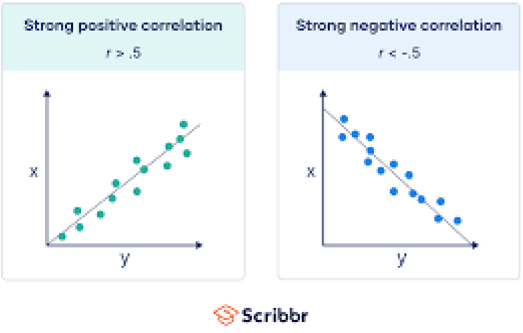
Ans 3) The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between **–1 and 1** that measures the strength and direction of the relationship between two variables.

If range lies between 0 and 1 – This means, there is a positive correlation between the variables, and they move in the same direction. In other words, if one increases the other also increases.

If range lies between -1 and 0 – This means that there is a negative correlation between the variables and the move in the opposite direction. In other words, if one increases the other decreases.

If value is 0 – Then there is no correlation.

High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation. Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation. Low degree: When the value lies below +-0. 29, then it is said to be a small correlation.



Ans 3) It 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 speeding up the calculations in an algorithm.

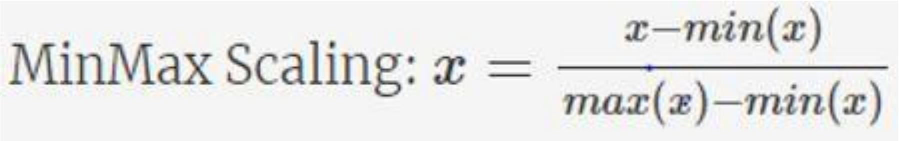
Many a times, the dataset we have has columns varying in units and magnitude. If scaling is not done, then the algorithm will keep note of magnitude and not units thus leading to incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

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.

Two types of scaling are there: -

Normalization/Min-Max Scaling

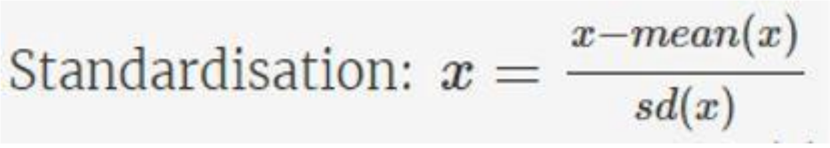
It brings all the data in the range of 0-1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling

Standardization replaces the values by their z scores. It brings all the data in a normal distribution which has mean 0 and standard deviation 1. sklearn.preprocessing.scale helps to implement standardization.

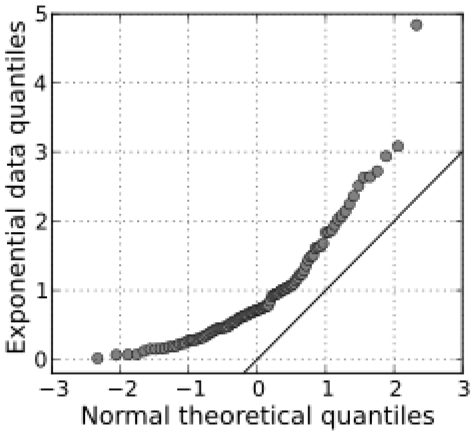


$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

Ans 4) VIF is referred as the correlation between the independent variables. If there is a high correlation, then VIF = infinity. In this case r2=1 which lead to 1/(1-r2) infinity. It means that the variable can be expressed as combination of other variables. To solve this issue, we need to remove or drop the variable having such VIF to avoid multicollinearity. In the dataset, the temp and atemp variables are almost same and thus have perfect correlation leading to infinite VIF.

Ans 5) Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any probability distribution like normal, uniform, exponential.

A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line 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.