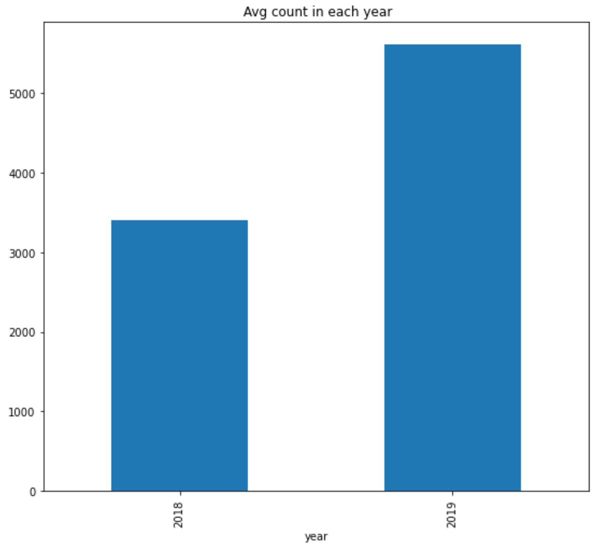
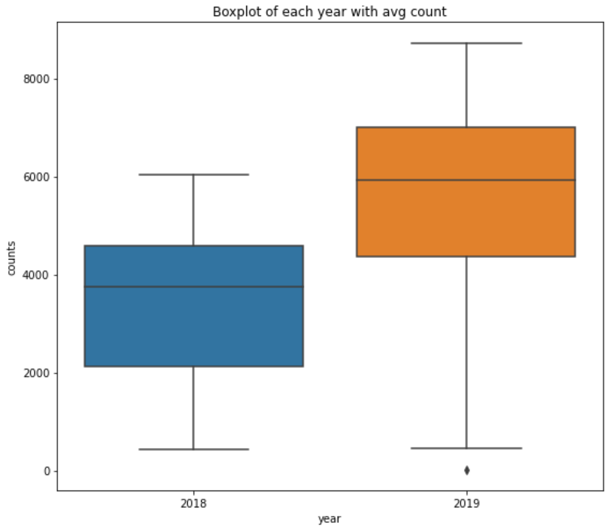
Assignment-based Subjective Questions

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

Ans. *Let’s talk about each categorical variable one by one-*

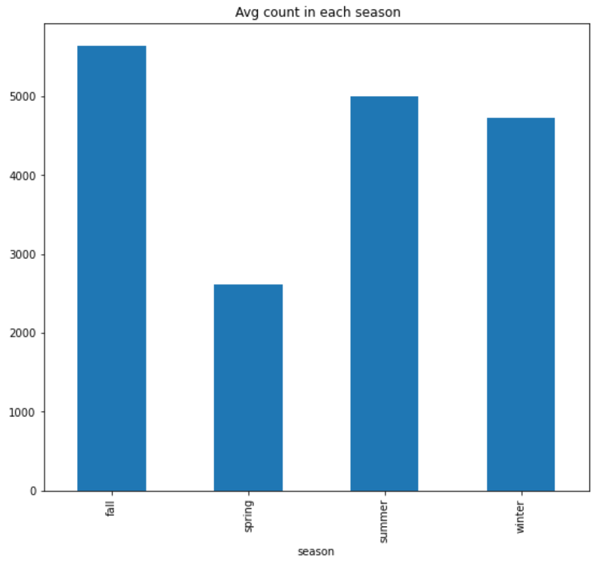
a. ***Year***

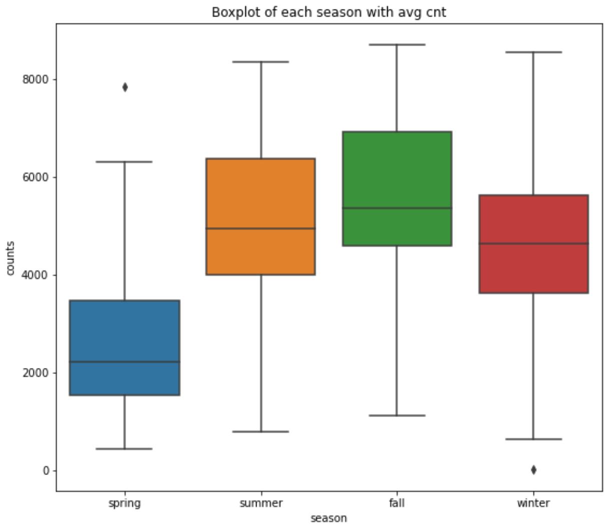

Avg count in each year


Boxplot of each year with avg count

As we can see, the median of the target variable, counts, increases as the year increases. It means it has a linear relationship with our target variable.

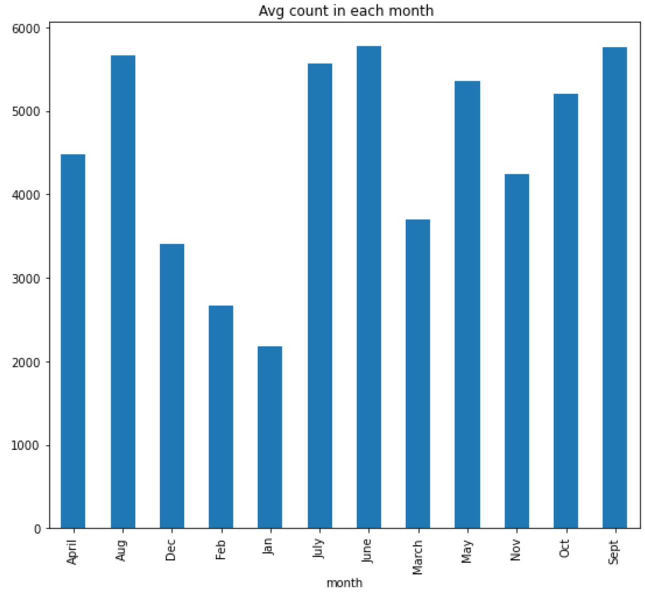
b. Season

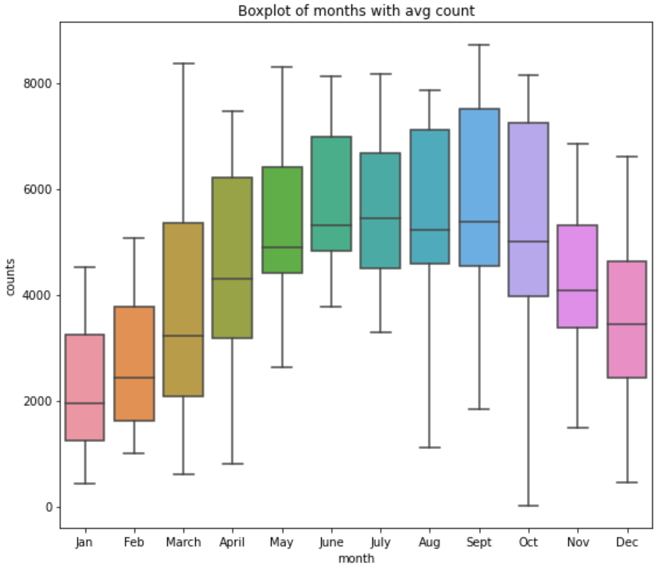

Avg count in each season


Boxplot of each season with avg cnt

As seen from the boxplot, summer and fall have more no of average rental bikes than spring and winter. But after building our model, we found that Spring has a negative impact on the target variable but Summer and Winter have positive beta coefficients.

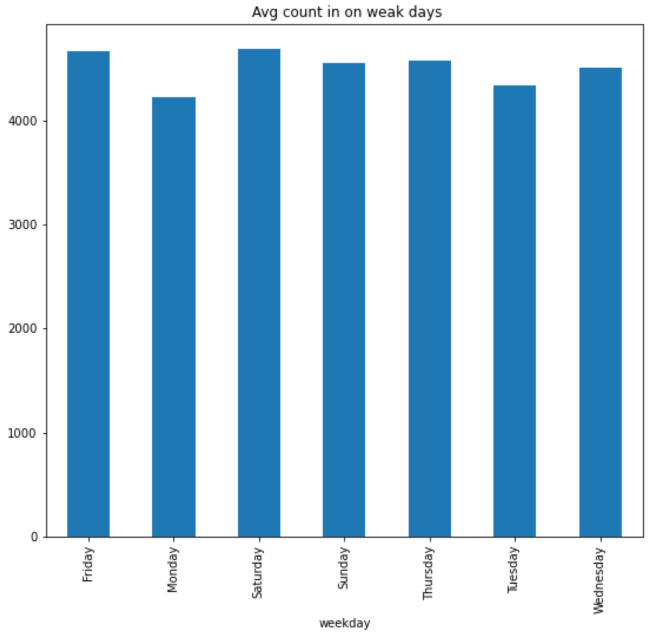
c. Month

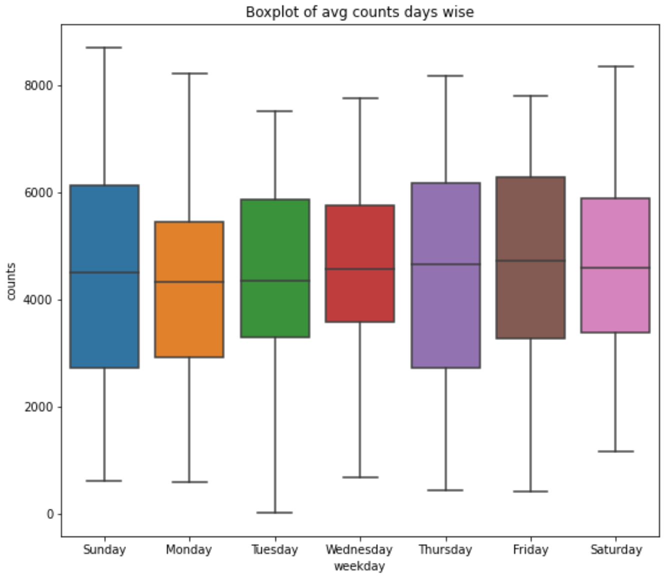




It can be seen that the count of rental bikes is increasing from January to July and then slowly it starts decreasing. After building our model, we found that only September has a good effect on our target variable and its beta coefficient is positive.

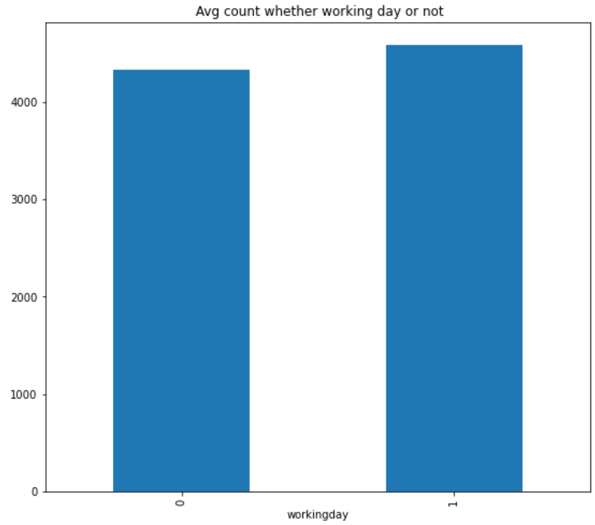
d. Weekday

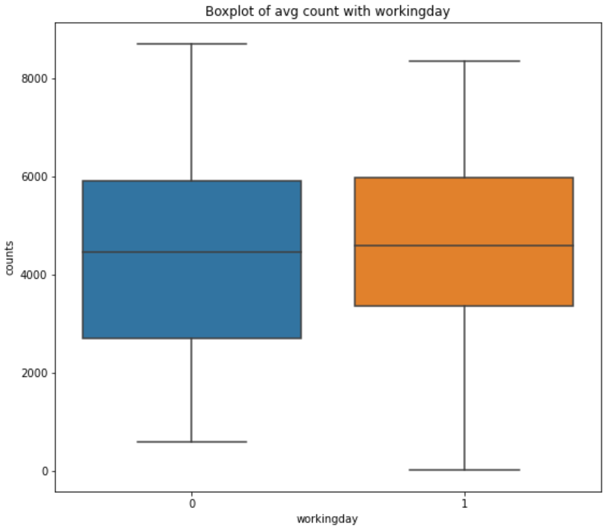




Weekday variable does not have much impact on the total count of rental bikes. The median and mean remains almost similar each day of the week. That’s why in our final model, we don’t have any weekdays. It is insignificant to our model.

e. Working day





Workingday also does not have any impact on our target variable, counts. Whether it is a working day or non working day, the number of rental bikes is not affected by it. That’s why we do not have this variable after building our model.

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

   Ans: *While creating the dummy variables, there is a creation of 1 extra column which we don't need actually. So we drop that extra column by dropping the first column with the help of drop_first = True. All the columns can be represented by the 1 less column. Ex. if we have n levels in a categorical column, we just need n-1 dummy columns to represent all the n levels in that column. So to avoid any extra column while building the model, we drop the first column by default.*

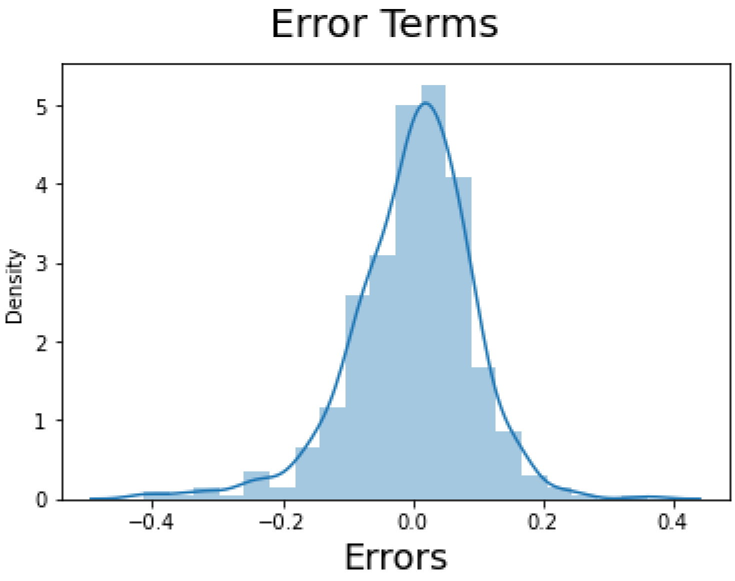
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   Ans: *Temperature variable (temp : 0.63)*

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

   Ans: *To validate the assumptions of Linear Regression after model building, I did the following things-*

   a. *Checked if error terms are normally distributed. Yes they are.*



b. *Checked if the sum of error terms must be zero.*

sum((y_train - y_train_pred)) = -5.103695244201845e-13 (very very close to zero)

c. *Error terms should be independent of each other.*

AFter plotting the error terms, there should be no pattern between error terms.

d. *No correlation among independent variables.*

Checked the VIF after building the model. All the independent variables have the value of less than 5. So we can say that there is no correlation among the independent variables.

e. *Error terms have constant variation (Homoscedasticity).*

By plotting the error terms, we can easily visualise if they have constant variance or not. In this case, yes they are having.

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

   Ans:

   a. *Temperature with beta coefficient of 0.477619*

      It has the strongest contribution among all the independent variables. Ast the temperature increases, the demand for rental bikes is increasing.

   b. *Year with beta coefficient of 0.234459*

      As the year increases, we can say that the demand is increasing.

   c. *Snow/Rain weather with beta coefficient of -0.280241.*

      If the weather is bad - snow or rain is there, from our analysis, we can see that the demand is decreasing.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: *Linear Regression is a type of machine learning algorithm where we try to find out the linear relationship between independent and dependent variables. These are of 2 types-*

a. *Simple Linear Regression*

   In this, we have only 1 independent variable, x and 1 dependent variable, y. We find the linear relationship between these two only. The equation of the line an be given as-

   $y = m * x + c$

Where m is the slope, c is the constant.

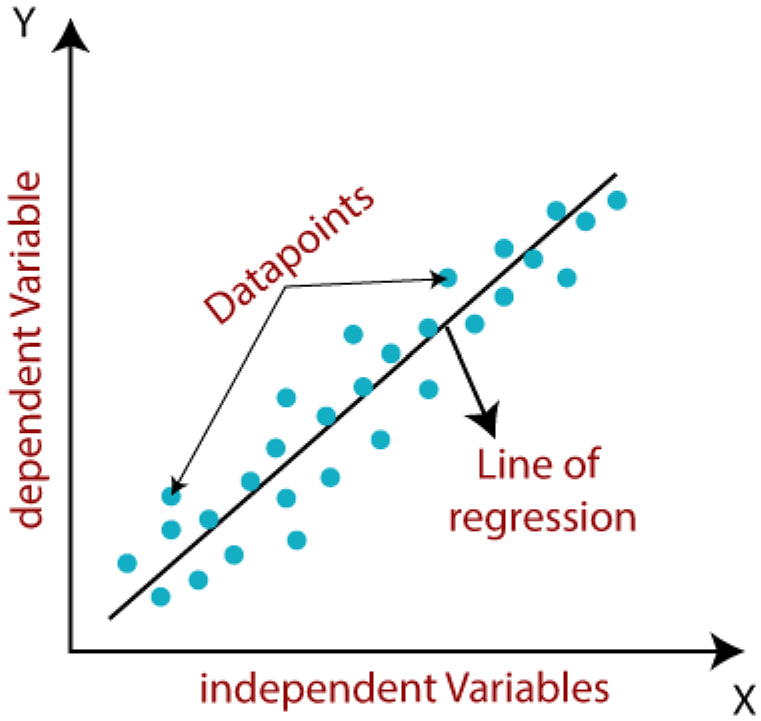
b. *Multiple Linear Regression*

In this, we can have more than 2 independent variables, x1, x2, x3 and so on and 1 dependent variable, y. The equation can be given as-

$y = c + m1 * x1 + m2 * x2 + m3 * x3 + \ldots..$

Where c is the constant and m1, m2, m3 .. are the coefficients of x1, x2, x3, … respectively.

We try to best fit the line into our dataset.



How do we fit the line?

*We first start with some m and c and then we calculate the y = m*x + c*

First we divide our dataset into 2 parts- training set and test set with the ratio of 80:20 or 70:30. We train our model on the training set and then we predict the model that we have built on the test set.

After this, we calculate the squared sum of residuals.

*Ypred = m*x + c is our predicting value.*

residual , r = Ypred - y

Then we calculate the RSS (Residual Squared Sum of error)

$RSS = (Y1pred - y1)^2 + (Y2pred - y2)^2 + (Y3pred - y3)^2$ ……

*where Yipred = mi*x+c*

This RSS is our cost function in the case of Linear Regression and then we try to minimize this cost function using the method called Gradient Descent. We find the optimal values for m and c.

We can minimize it with two ways - Using differentiation and using an iterative method. Software tools use the iterative way to minimize this cost function.

After fitting the best possible lines, we do the residual analysis. In this, we validate the assumptions of linear regression. We perform these checks-

a. *Check if the independent variables are not having any multicollinearity among themselves.*
b. *Check if the error terms are normally distributed.*
c. *Check if the sum of all the error terms is zero.*
d. *Check if the error terms have the constant variance.*

After validating the assumptions, we go for the prediction.

We perform the prediction on the test data set. We use the model that we have built on our training data set and we predict the output using the test dataset.

At last, we see the r2 score of the training dataset and test dataset. There should be much difference between these two.

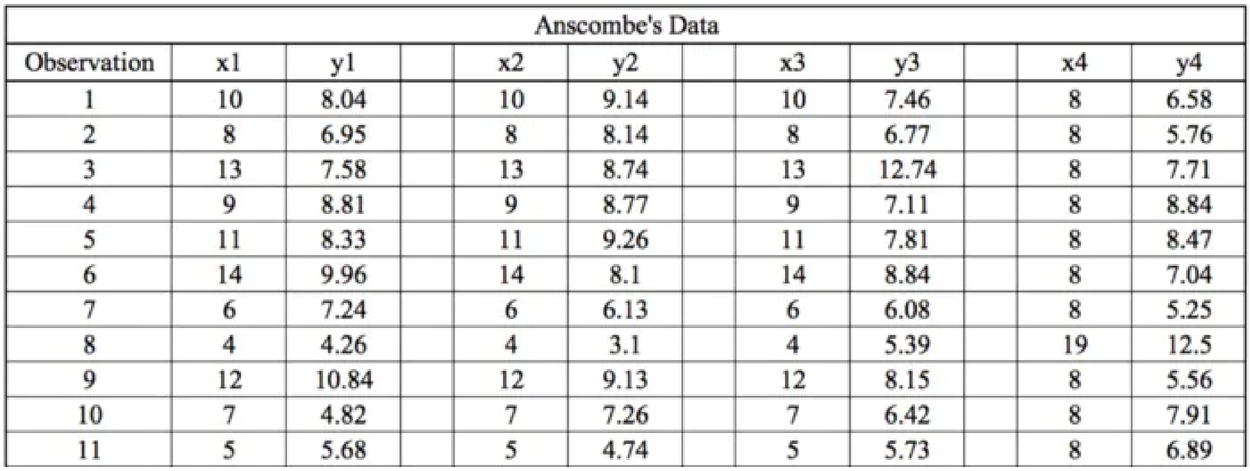
To say we have a good fit model, we also check for F probability. If it is very low, we can say that our model is a good fit.

2. Explain the Anscombe's quartet in detail.

   Ans: *Anscombe's quartet explains the importance of visualising the data before applying any analysis or any algorithm. It says the basic statistics can tell the* *same story for all the dataset but in actuality, there may be a completely different story based on the data when we visualise it.*

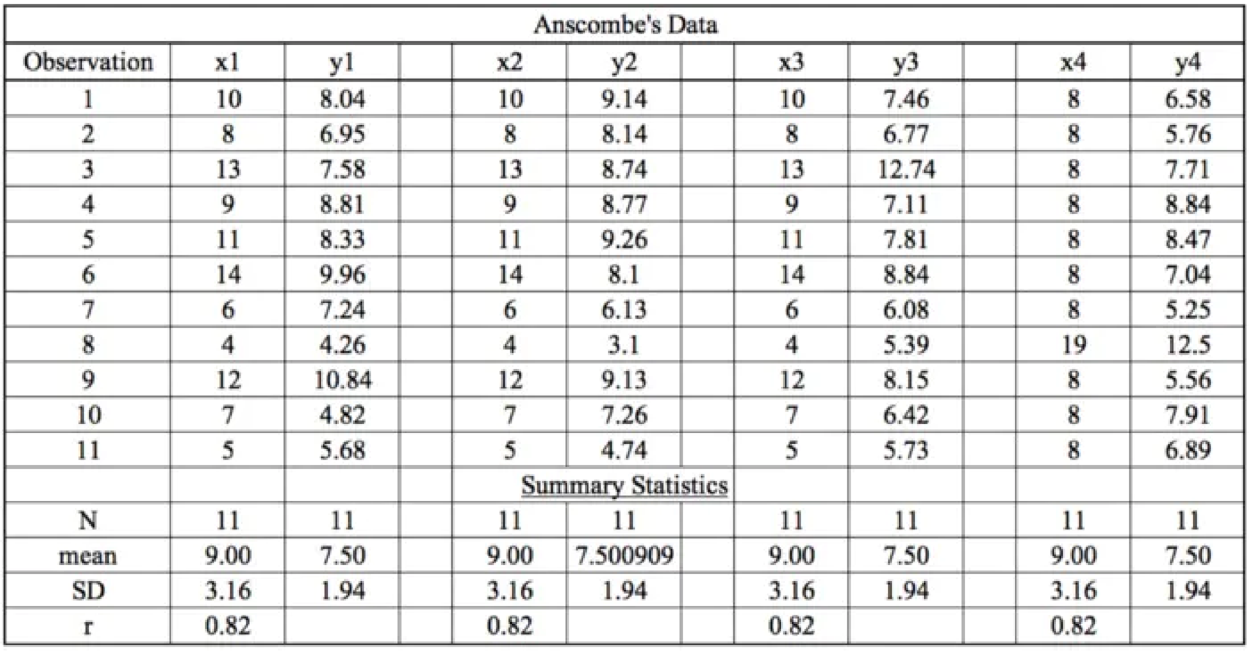
Anscombe’s Quartet comprises 4 datasets which are nearly the same in simple descriptive statistics.

Ex.



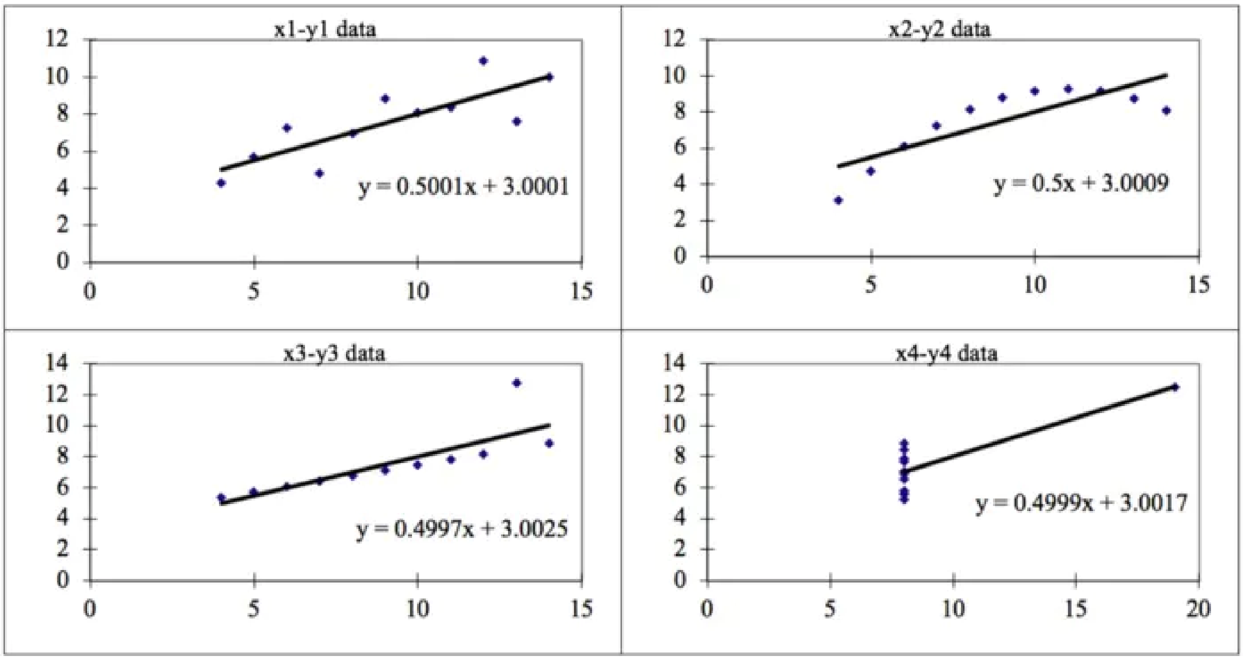
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for all these four datasets are approximately similar and can be computed as follows:



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

It can be seen that the Mean, Median and standard deviation of all the 4 dataset are identical. But what we find when we plot these dataset-



All the 4 dataset can be described in this way-

- *Dataset 1: this fits the linear regression model pretty well.*
- *Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.*
- *Dataset 3: shows the outliers involved in the dataset which cannot be handled by a linear regression model.*
- *Dataset 4: shows the outliers involved in the dataset which cannot be handled by a linear regression model.*

That's why visualising the data is very important before applying any machine learning algorithm to make our model best fit.
(image source: hackernoon.com)

3. What is Pearson’s R?

   Ans: Pearson's r is a numerical summary of the strength of the linear association between the variables. 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.

   In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

   Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change).

We can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. There are certain requirements for Pearson's Correlation Coefficient:

- *Scale of measurement should be interval or ratio*
- *Variables should be approximately normally distributed*
- *The association should be linear*
- *There should be no outliers in the data*

The formula is given by -

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^{2}-(\Sigma x)^{2}][N\Sigma y^{2}-(\Sigma y)^{2}]}}$$

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Ans: *Scaling is a type of preprocessing which we apply to our variables in order to bring them into a particular range. It helps us to get all the variables in some fixed particular range where we can say that the maximum and minimum value of each variable is fixed.*

   We perform scaling because when we get the data, the variables can vary from magnitude and units. There may be a major difference between the magnitude of var x1 and var x2. The beta coefficient that we will get after the model building will also vary and we may ignore the importance of the lower beta values variable. In order to avoid these things, we perform the scaling.

   Also, scaling helps the machine learning algorithm to run faster.

   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.

   Standardized Scaling vs Normal Scaling-

   Standardized Scaling:

   This replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean zero and standard deviation one.

$Scaling = (x - mean(x))/std$

Normalisation:

It brings all the values of the variables between 0 and 1.

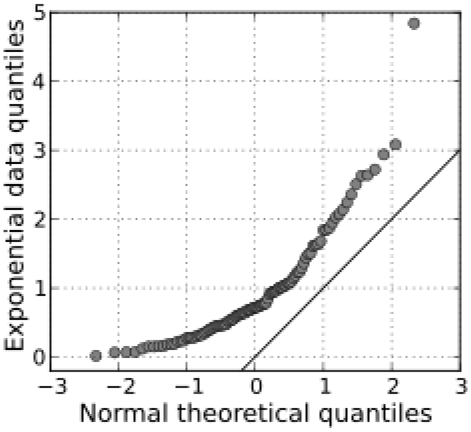
$Scaling = (x - min(x))/(max(x) - min(x))$

It also takes care of outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans: *The large VIF means that the variables are highly correlated. In the case of infinite VIF, it means that the 2 variables are having the perfect correlation between them. When correlation is perfect, its R2 = 1 and then from formula, 1/(1-R2) , it goes to infinite.*

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

Ans: *Q-Q plots are the graphs of two quantiles against each other. It's purpose is to find out if the two datasets are coming from the same distribution or not.*



There is a reference line of 45 degrees in this plot. 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.

In linear regression, the use of this plot is when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with the same distributions.