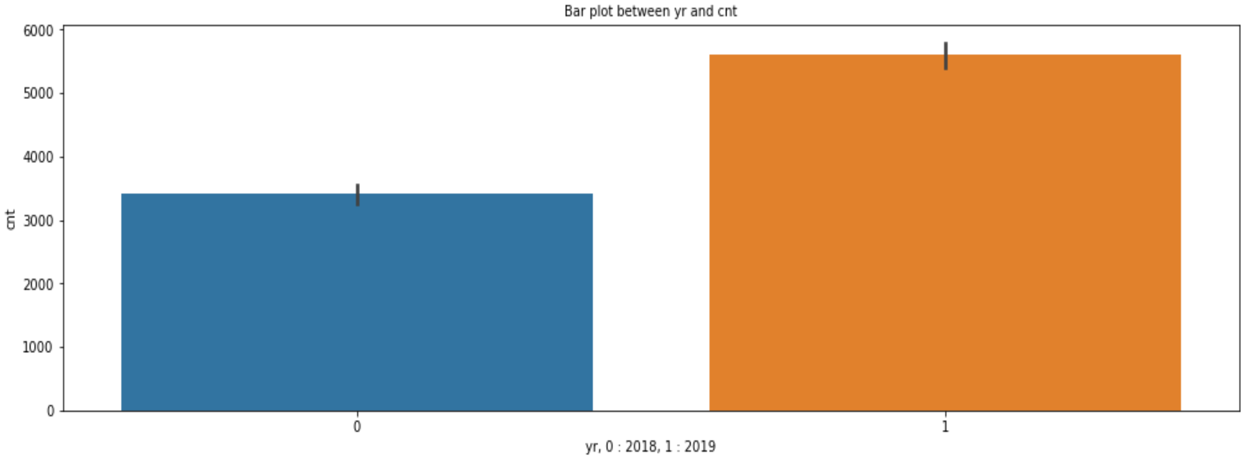
Assignment-based subjective questions:

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

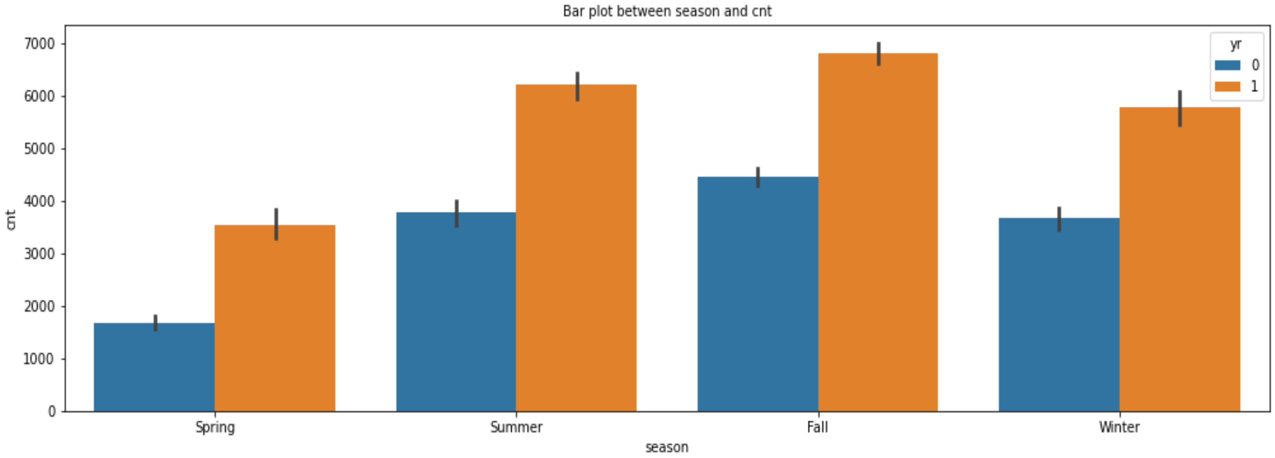
   I conducted bivariate analysis of categorical variables with our dependent variable i.e. cnt. Some of the variables showed some dependencies and some did not. Based on the relevance, I have added my inferences below along with the graphs to support it.

   - Year Vs Count :



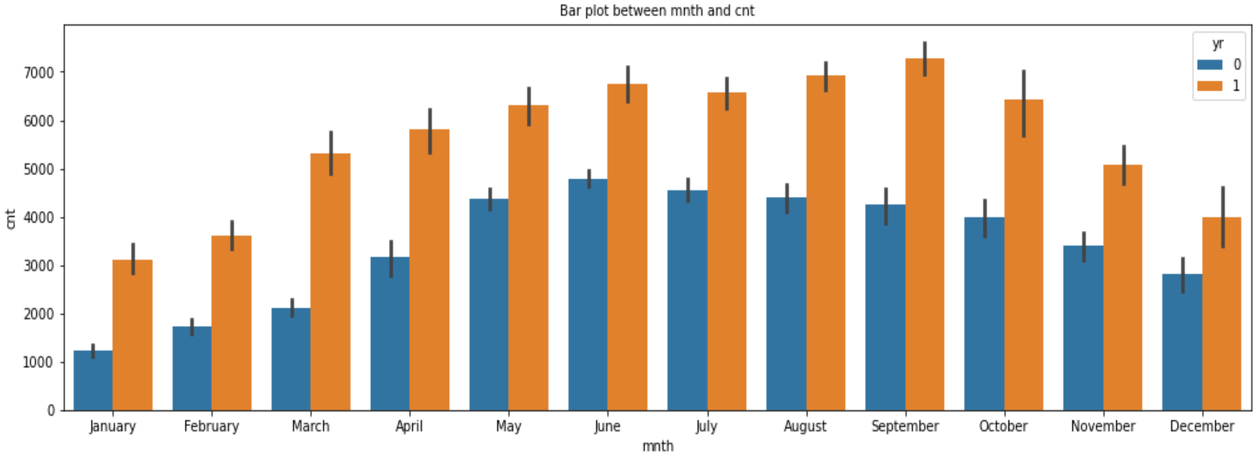
From the plot above we can see that there is an increase in the customers of the company from 2018 to 2019, which is a positive sign. This could mean that the services provided by the company are good and customers are liking it and hence the numbers have increased. But this may be vague to decide just based on the data of two years sales. But on a whole it gives us a good big picture.

- Season Vs Count for each Year :



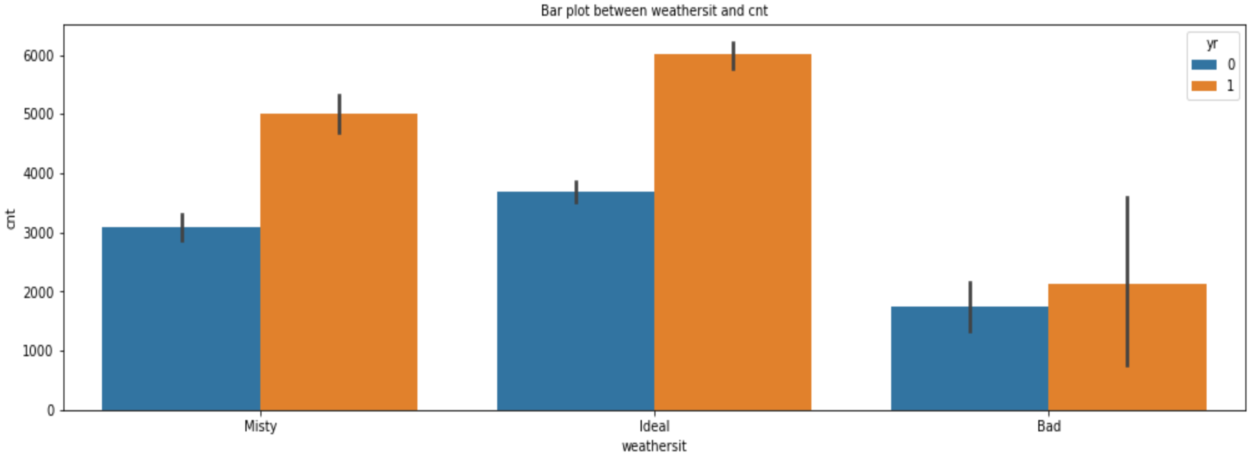
Somewhat similar trend can be seen here, There has been maximum demand in bikes during Summers and Fall for both the years. This can be possibly explained because most of the students have their holidays during this period and those who are working tend to take leaves for holidays and travels at this time of year.

- Month Vs Count for each Year :



This plot confirms our previous inference, the months of summer and fall are having the maximum demand of bikes for both the Years. Thus, maximum bookings.

- Weather conditions Vs Count for each Year :



It is quite obvious that the demands for bike would be very less for Bad weather compared to the Misty or Ideal one. We can also see that the Yearly rise in bookings of Misty and Ideal weather is more as compared to the Bad weather.

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

We use get_dummies() function from the Pandas library to create dummies which creates “n” dummies and the reason we use drop_first = True is to drop one level of category which is beneficial for our model, so it can converge or learn from the data in a better way,there is no harm in keeping that extra category it’s just so for the convenience of our model. This can be understood in a better way by an example;

We have a categorical column of Seasons in our assignment dataset. There are 4 categories, i.e. Spring, Summer, Winter, Fall. While creating dummies if we do it without drop_first=True command, we will create 4 levels of dummies for this column which will explain each category individually. On the other hand if we use drop_first=True command, we will create three levels of dummies, 4th level will be understood by the logic that if level 1 is not there, level 2 is not there and also level 3 is not there then it is obvious that it is level 4. Thus, there wont be any requirement of the 4th level.
Thus, we can say drop_first=True is important.

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

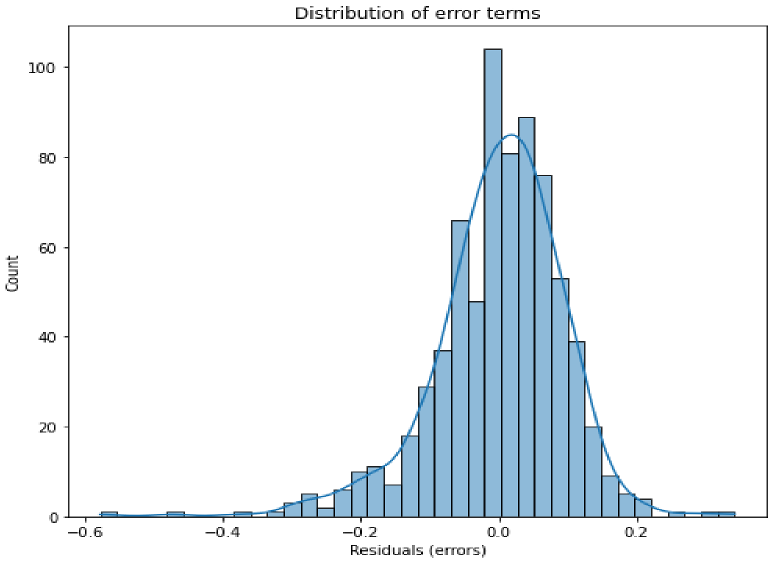
   From the pair-plot we can infer that ‘temp’ variable has the highest correlation with the target variable i.e. Count. The scatter plot of temp vs cnt shows very good positive relation and some sort of linearity is visible.

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

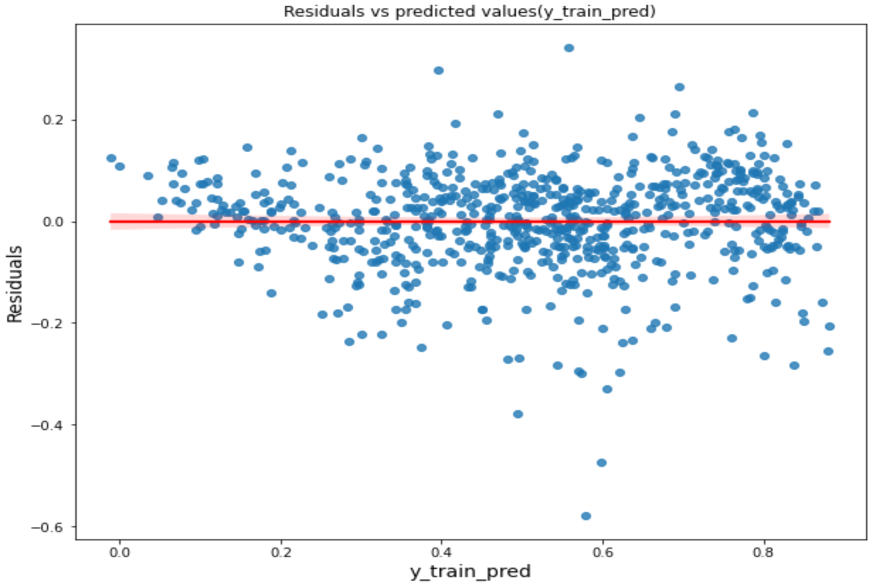
   There are 4 main assumptions when we build any Linear Regression model. We will see each one below:

   a. **There should exist some linear relationship between X and Y :**
   When we made the pairplot of the continuous variables we saw in temp vs cnt and atemp vs cnt that the linear relationship was clearly visible. Some moderate linearity was observed in the Windspeed Vs Cnt. Here count is our target variable.

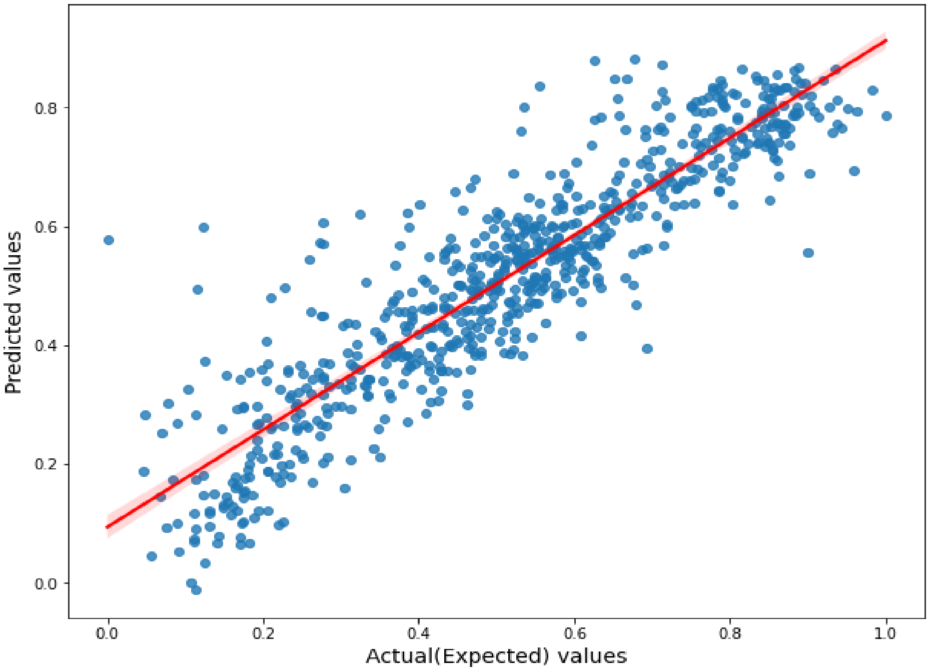
   b. **Error terms should always be normally distributed:**
   I have used the training set here to plot the errors (residuals) which is nothing but the difference between the actual values and the predicted values. We plotted the error terms in the form of histogram, which confirmed the normal distribution around 0.



c. **Error terms should be Independent:**
   We confirmed this assumption by plotting the Residuals vs y_train values which we used to train our model. This graph did not show any pattern or shape and there was large variance in the points.



d. **Error terms should have constant variance:**
   We plotted the residuals with the predicted values and observed a linear trend with the predicted value(line) was close enough to the actual points.



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

   Based on reading the model parameters of the given data, we find that the following three features contribute significantly to the demand:

   - Temp
   - Weather conditions- Ideal
   - Season – Winter

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Linear Regression is a type of Supervised Learning method, where in we predict the values or behavior of some dependent variable based on the relationship with various independent variables. It is used mainly in the case of regression of continuous variables.
   In Linear regression, the data is modelled using straight line equation. In mathematical terms, our model tries to find out the best fit line that passes through the given set of data points with minimum error. We can say that the line is he predicted value and the points are the actual values.

   Linear Regression modelling is of two types based on the number of independent variables:

   a. **Simple Linear Regression (Y = mX+C) :**

It has one dependent variable and one independent variable.

b. **Multiple Linear Regression (Y = C + m1X1 + m2X2 + … + mnXn) :**
It has one dependent variable and multiple independent variables.

Here,
Y = Dependent Variable
X = Independent Variable
C = Y intercept
m,m1,…,mn = Slope

The Linear Regression Algorithm iterates on the slope and intercept values of the Given points and updates it so as to find the line which is best suited for the given Data points, in short it finds the line best suited for the given model.
It uses Residual Sum of Squares and Total sum of squares to give us an idea of the accuracy of the model with 'N' features; it is upto us which features we need to use or remove based on our domain knowledge, in short we focus on minimising the error between the actual values and the predicted values.

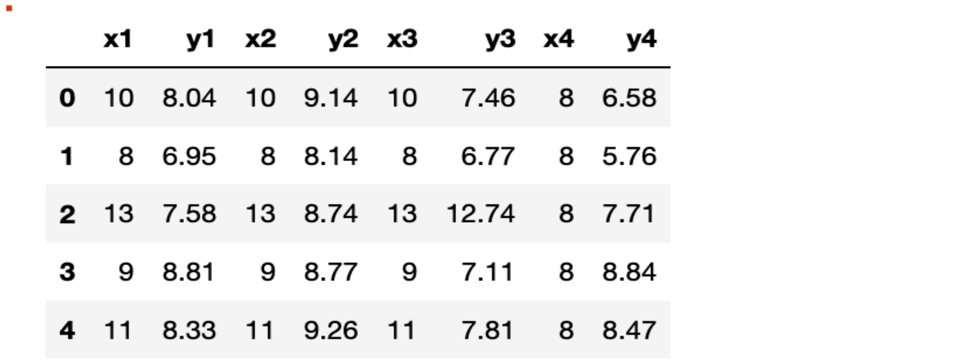
This leads us to the problem of **over fitting** and **under fitting**, **overfitting** is where the model memorizes the data instead of learning it; this is the case where there are too many variables and making the model so complex and it ends up failing to generalize, underfitting is where the model is just too simple and lack of features leads to the model.
We also need to **validate some assumptions of Linear regression** and violating any one of them can lead to serious errors in our model with our model failing to learn or generalizing on the dataset.

There are some assumptions to validate in linear regression on the training set and violating any of these can introduce some serious errors in our model.

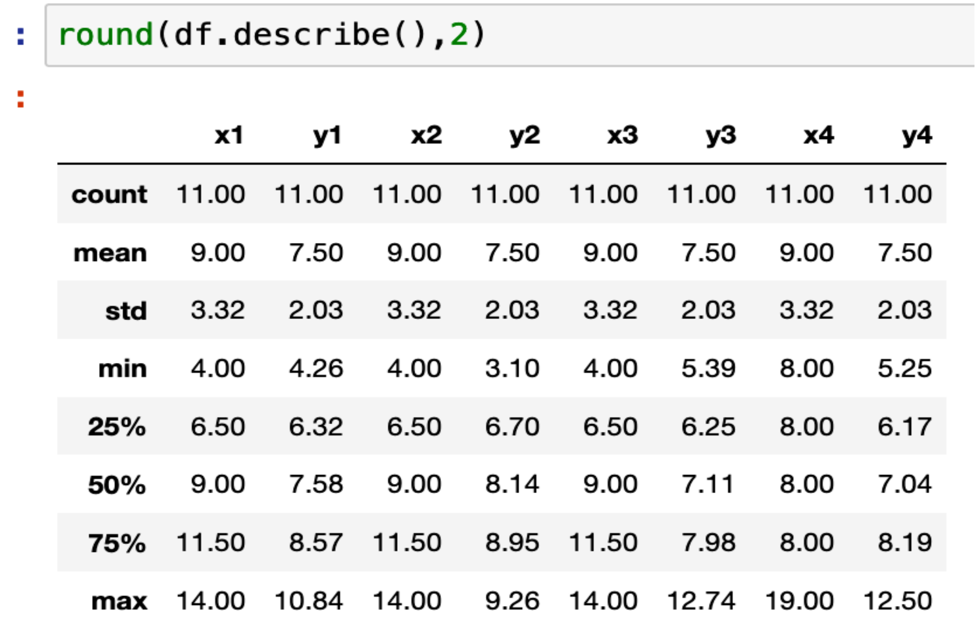
- There should be some linear relationship between X and Y.
- Error terms should be normally distributed around mean zero.
- Error terms should be independent of each other.
- Error terms should have constant variance.

2. Explain the Anscombe’s quartet in detail.

According to Wikipedia **Anscombe’s quartet** is a set of four data sets each of them having extremely similar descriptive statistics (Mean, standard deviation etc). It was devised by a statistician who goes by the name of **France Anscombe** ;wanted to show the need of graphing the data and not relying on the descriptive statistics along to conclude. Following is a dataframe consisting of Anscombe’s quartet and its descriptive statistics;



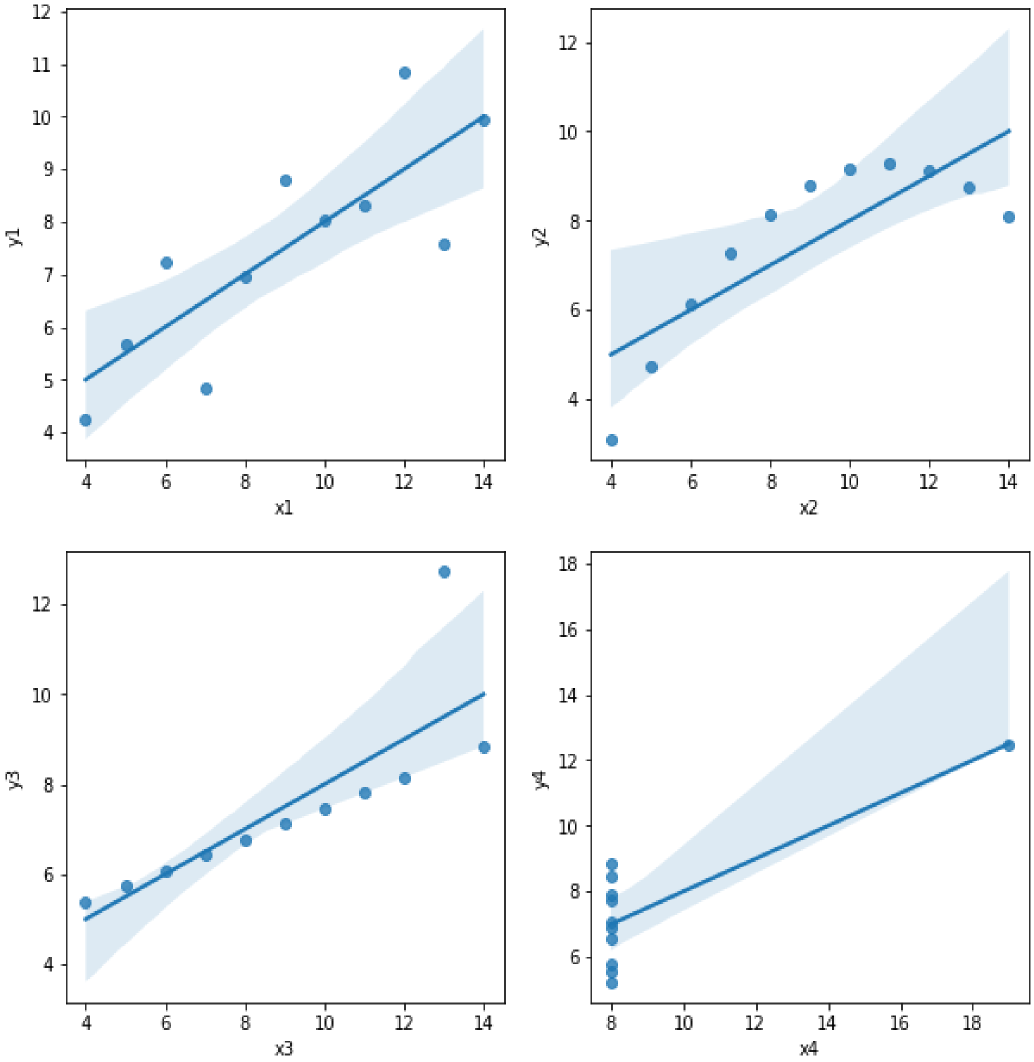
	x1	y1	x2	y2	x3	y3	x4	y4
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47



```
round(df.describe(),2)
```

	x1	y1	x2	y2	x3	y3	x4	y4
count	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00
mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
std	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
min	4.00	4.26	4.00	3.10	4.00	5.39	8.00	5.25
25%	6.50	6.32	6.50	6.70	6.50	6.25	8.00	6.17
50%	9.00	7.58	9.00	8.14	9.00	7.11	8.00	7.04
75%	11.50	8.57	11.50	8.95	11.50	7.98	8.00	8.19
max	14.00	10.84	14.00	9.26	14.00	12.74	19.00	12.50

As one can see the descriptive stats are almost similar across the X and Y, so one would assume that the graphs or a scatterplot of these points will be the same as well; we will plot the graphs and then give our final verdict on it;



In the top the left column we can see that there exists a linear relationship between X and Y.

In the top right corner it is obvious that there is a non linear relationship between X and Y.

In the bottom left corner there exits a perfect Linear relationship except for that one outlier maybe for this a different regression line would fit.

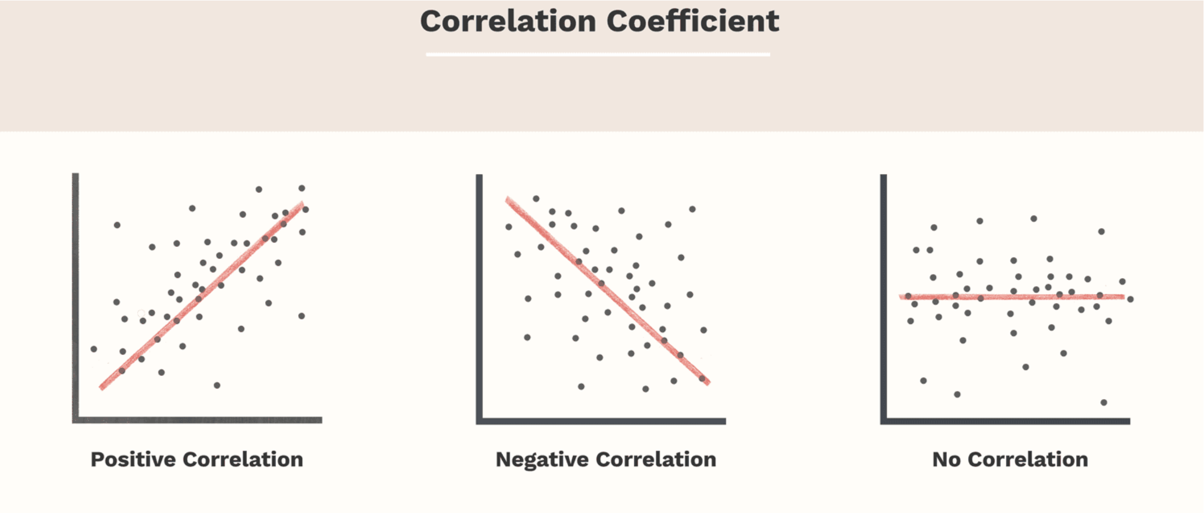
In the last figure we see that the points are sort of stacked as the value of X is constant with the exception of one point, this shows that one outlier is enough to promote high correlation coefficient.

And from the graphs above one can clearly feel the need to visualize the data instead of only relying on basic descriptive statistics.

3. What is Pearson’s R?

Pearson’s R or Pearson’s correlation coefficient gives us the linear correlation between two variables in other words it gives us the summary of strength of correlation between those two variable and it is given sum of products of X and Y subtracted by product of individual sum of X and Y divided by their standard deviation.

Pearson's r can take values ranging from -1 to 1 and can be categorized as Positive Correlation, Negative Correlation and No Correlation.



From the first graph we can see a Positive correlation(greater than zero) where one variable goes the other goes up as well.

From the second graph we can see a Negative correlation(less than zero) where one variable increases the other decreases.

The third graph depicts No correlation(equal to zero) where the points are scattered everywhere and no visible pattern can be identified.

4. **What is Scaling? Why is scaling performed? What is the difference between Normalized and standard scaling?**

Scaling or more commonly known as Feature scaling is done to standardize the data in Machine Learning in the pre-processing phase ; i.e. collapsing the data into a range of similar values this is done so that the larger values are not given more weightage for example The weight measure in pound will always be greater than weight measured in Kgs which is obviously not the case; so to remove this mismatch we do Feature scaling but we don't do it in Linear regression as there is only one independent variable. The difference between Normalized(Minimax) and standardized scaling are as follows:

Normalized scaling	Standardized scaling
Transforms the data in the range of [0,1] or [1,-2]	Transforms the data by removing the mean and scaling the data about mean being 0 and SD 1
It used when we do not known the distribution of the features in our data	It is used when we are certain of the distribution of data(normal distribution)
It is sensitive to outliers	It is not as sensitive to outliers compared to normalized scaling
It is calculated by subtracting the data in individual column by the minimum value and diving it by the range of that column X – Xmin/Xmax - Xmin	It is calculated by subtracting the data in individual column by the mean of that column divided by the standard deviation of that column X – Mean/ SD

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

   The formula for VIF (Variance Inflation Factor) is given by 1/(1-R^2). When VIF is below 5 the Multi collinearity is relatively low; when it is in the range of 5-10 one should investigate the variable and anything beyond 10 should be eliminated at once.
   The reason for VIF being “inf” is as the R^2(R square) approaches 1 i.e the dependent variables are able to explain the variance of independent variables brilliantly; the value in the denominator becomes closer to zero as it indicates extremely high correlation between independent variables and needs to be dropped after careful consideration of P-values as changing or dropping even one variable changes the VIF of other variables.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in Linear Regression.**

   Q-Q plot or Quantile-Quantile plot is a probability plot and is plotted using the quantiles of (x,y) where x represent the quantile of the first dataset and y represents the quantile of the second dataset. We compare two probability distributions (x,y).
   It is used to determine if the two dataset come form a common distribution, we plot the points on the X,y axis and if the distributions are similar on both the datasets, then

the points will (approximately) lie on the line with slope 1 i.e the identity line which has an angle of 45 degrees from the X-axis,
Let’s take an example from our Bike sharing assignment by taking the distribution of error terms or residuals, now according to an assumption in Linear regression residuals should be normally distributed; well approximately because the model which is built cannot be 100% linear; so a line is drawn at a 45 degree angle from the X-axis to fit a normally distributed data points(residuals) and we see if the points are on the line or around the line, if we see such pattern then our residuals are normally distributed.