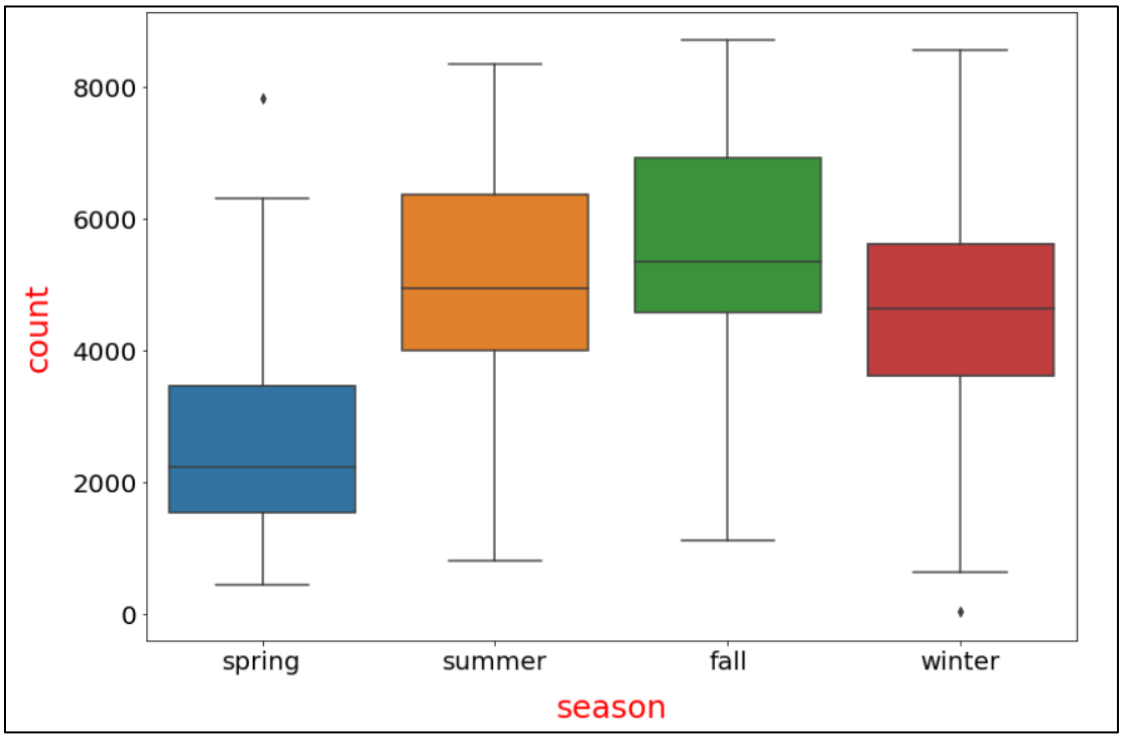
Assignment-based Subjective Questions

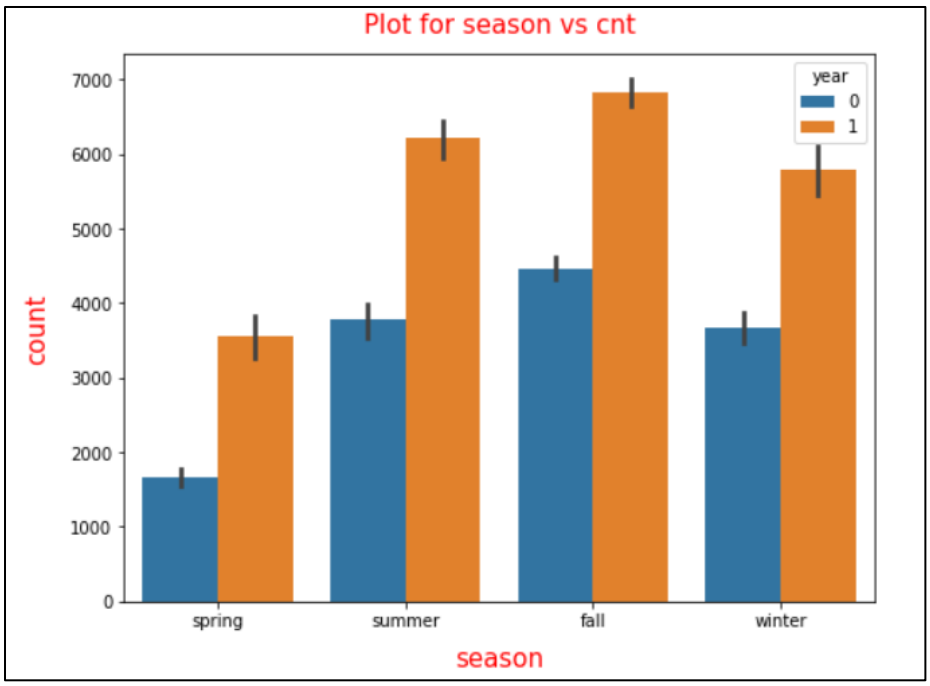
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Below is the list of Categorical variables and their effects on dependent variables that is on count of bikes rental:

Season:



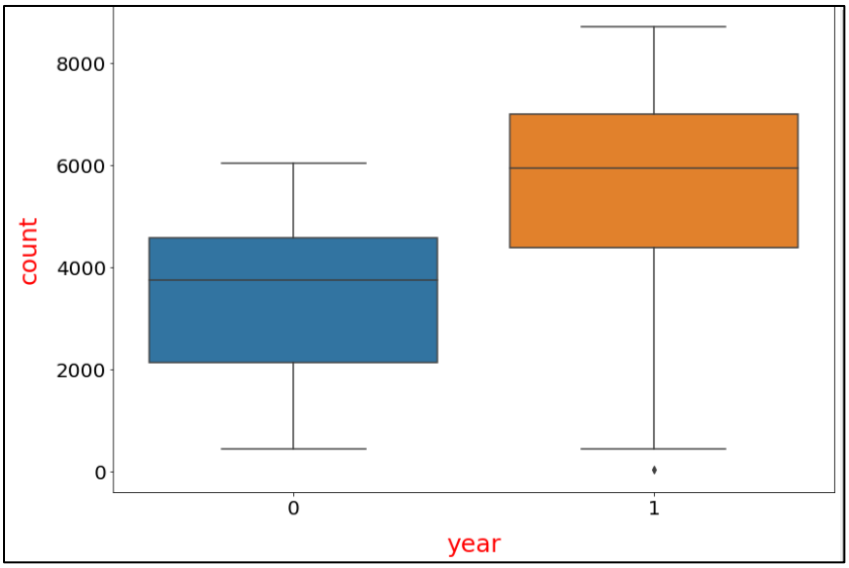
- We can see from the above box plot, `count of total rental bikes (Dependent variable 'cnt') is **high in season "Fall"` (Predictor variable) having 75% of the box lies in around 7000 for the season "Fall".** Then followed by season `"Summer"` then in `"winter"`.
- Whereas season **`"spring" has much less bike rental counts`.**



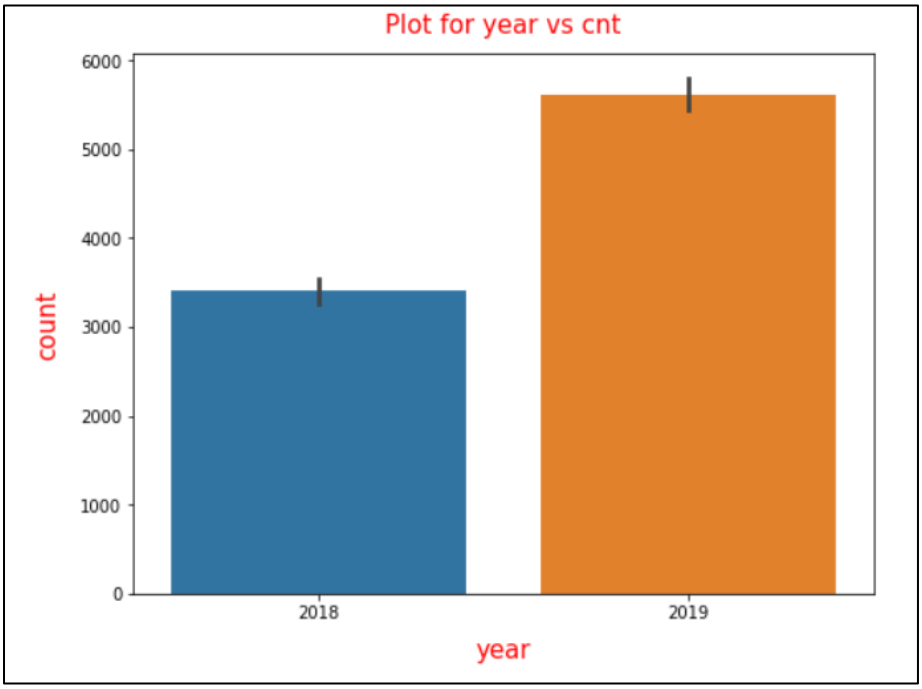
- From the above bar plot we can see that the count of total rental bikes (Dependent variable 'cnt') is **`high in season "Fall" (Predictor variable) and within the year of 2019.**Followed by season summer then in winter`.

- We can see that the count of total rental bikes (Dependent variable 'cnt') is also **`high in season "Fall" (Predictor variable) for the year of 2018`**.Also followed by season summer then in winter.

Year:

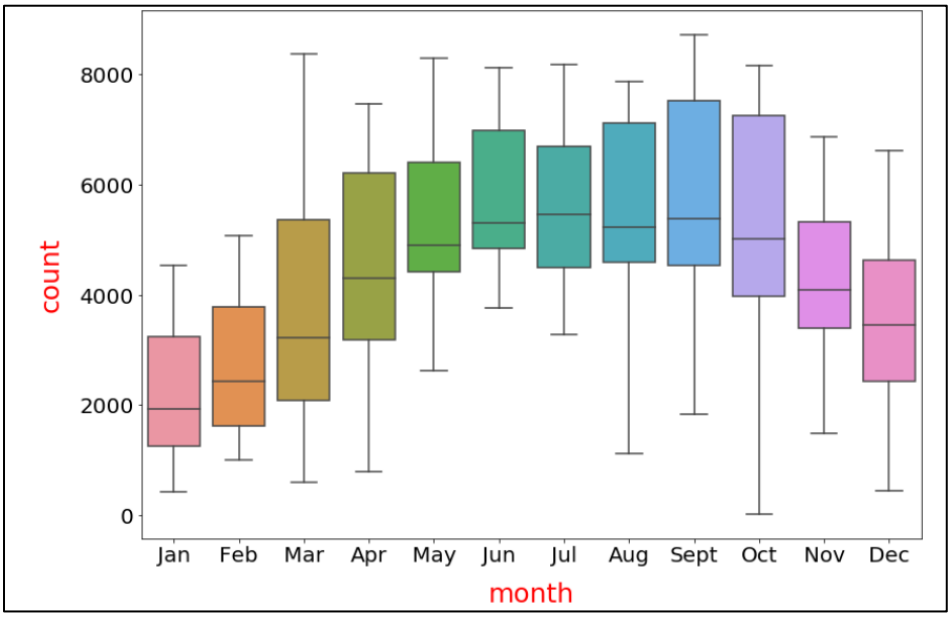


- We can see from the above box plot, count of total rental bikes (Dependent variable 'cnt') **is high in the year (Predictor variable) 2019 than 2018.**



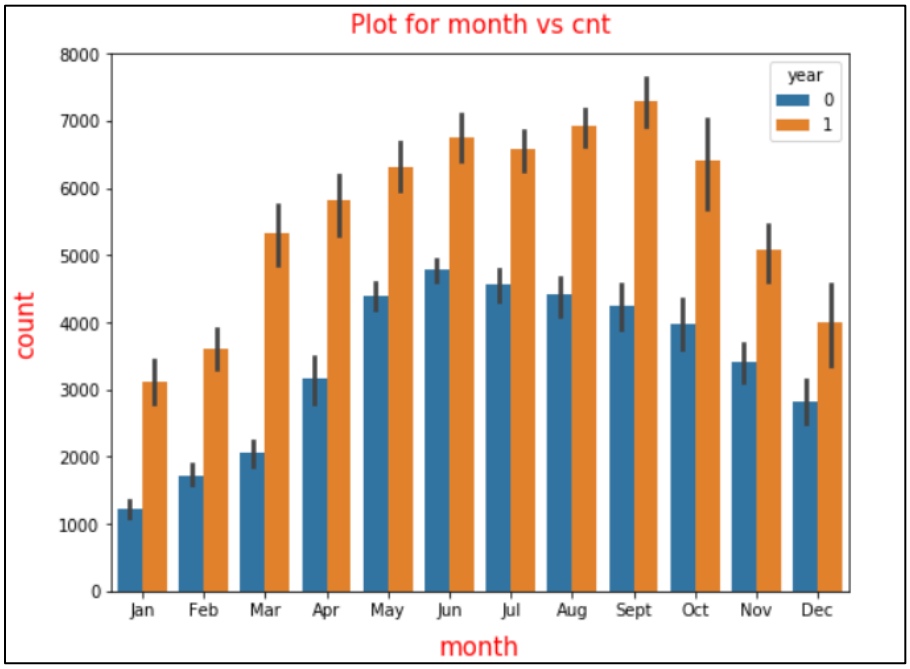
- From the above bar plot we can see that the count of total rental bikes (Dependent variable 'cnt') **is high in the year (Predictor variable) 2019 than 2018.**

Month:



- We can see from the above box plot, count of total rental bikes (Dependent variable 'cnt') **is high is the month (Predictor variable) of "September" followed by "October" then in "August".**

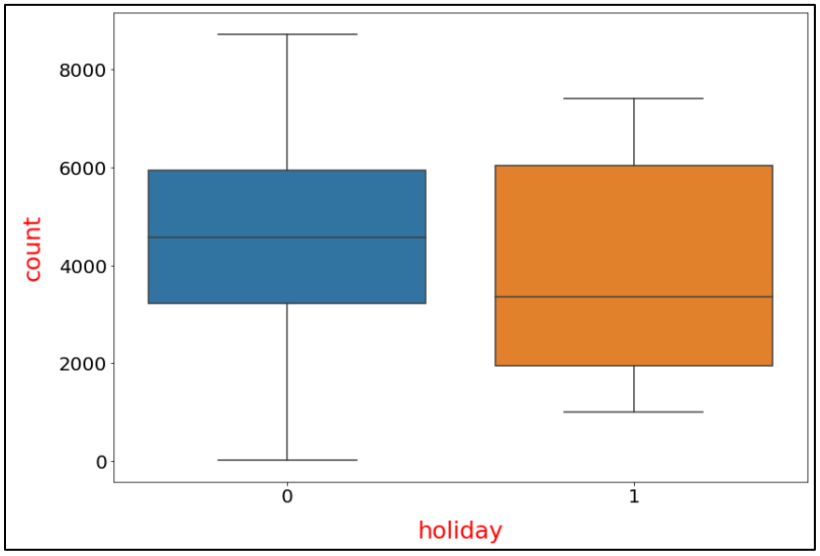
- Whereas **month "January" has much less bike rental counts**.



- If we see both years 2018 and 2019 separately in the bar plot, **then we can say the count of total rental bikes (Dependent variable 'cnt') is high in month of "September" in the year 2019.**

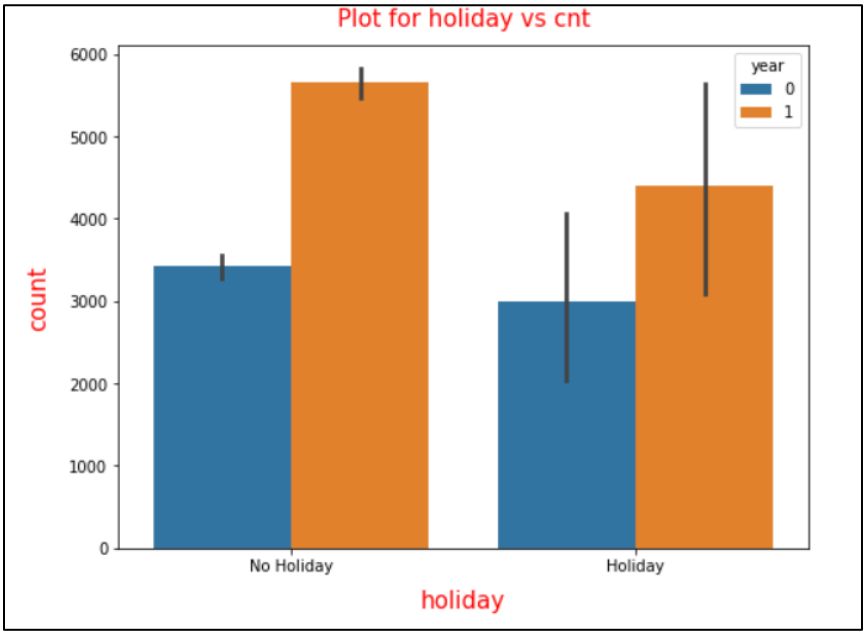
- The demand of rental bikes (Dependent variable 'cnt') **is high in month of "June" in the year 2018.**

Holiday:



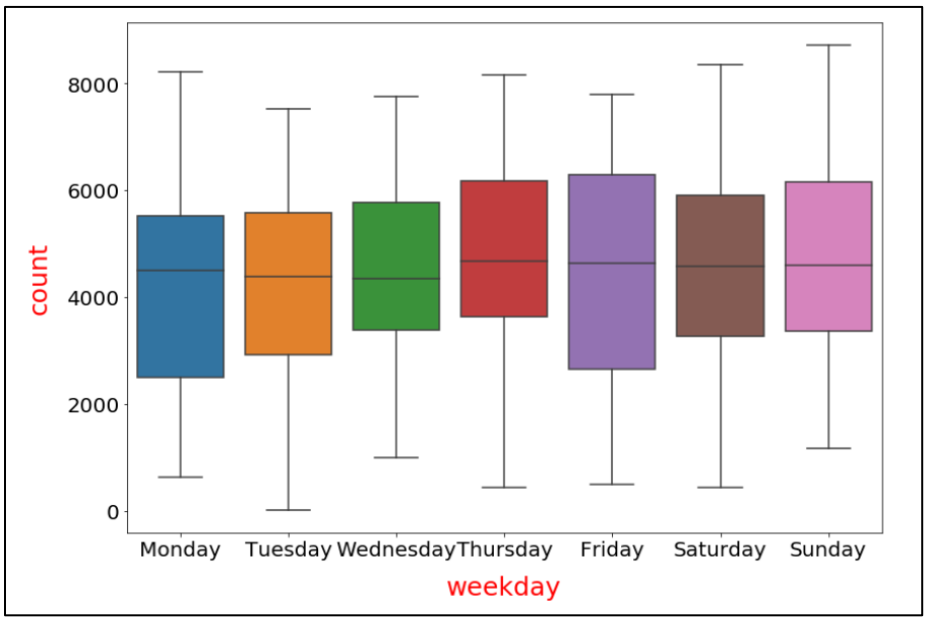
- We can see from the above box plot, count of total rental bikes (Dependent variable 'cnt') **is high when it is not a holiday (Predictor variable) as the median lies more than 4000 and upper fence of the box is more than 8000.**

- Whereas **the median and the upper fence value for the count of total rental bikes (Dependent variable 'cnt') is less when there is a holiday.**



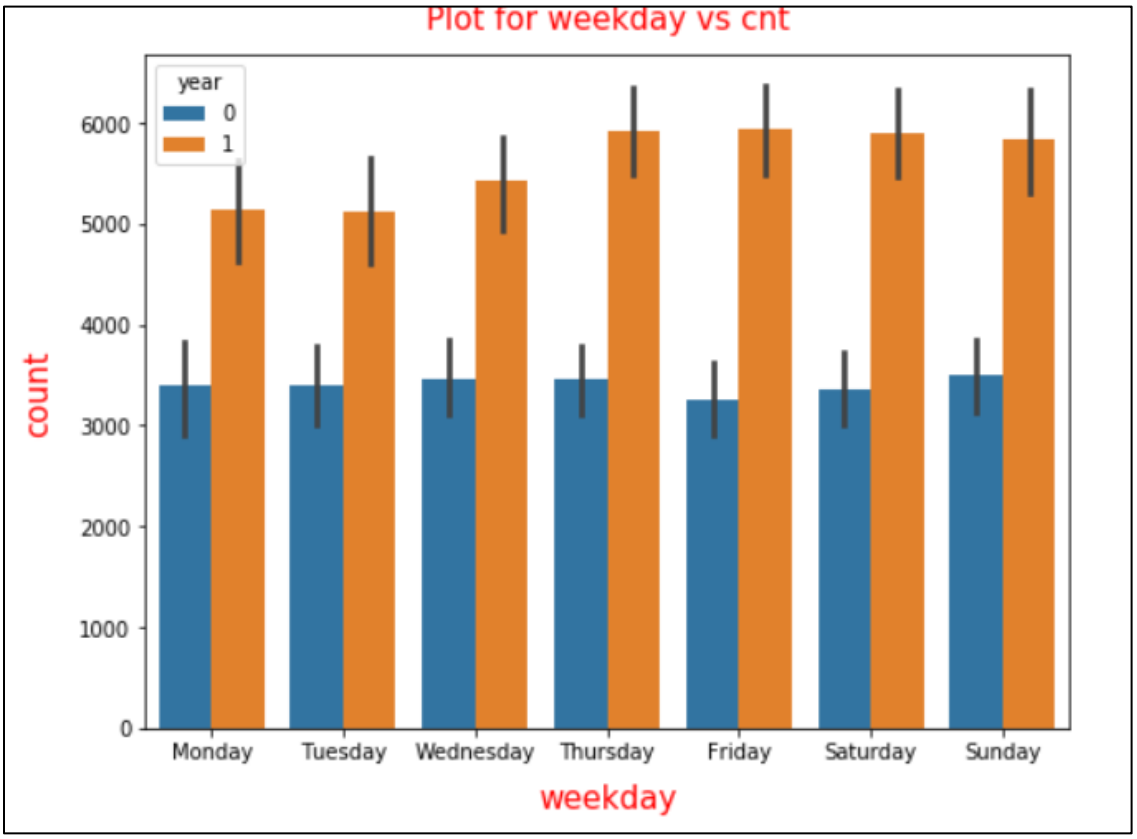
- From the above bar plot we can see that the count of total rental bikes (Dependent variable 'cnt') **is high when there is not holiday for both the years 2018 and 2019.**

Days of the Week:



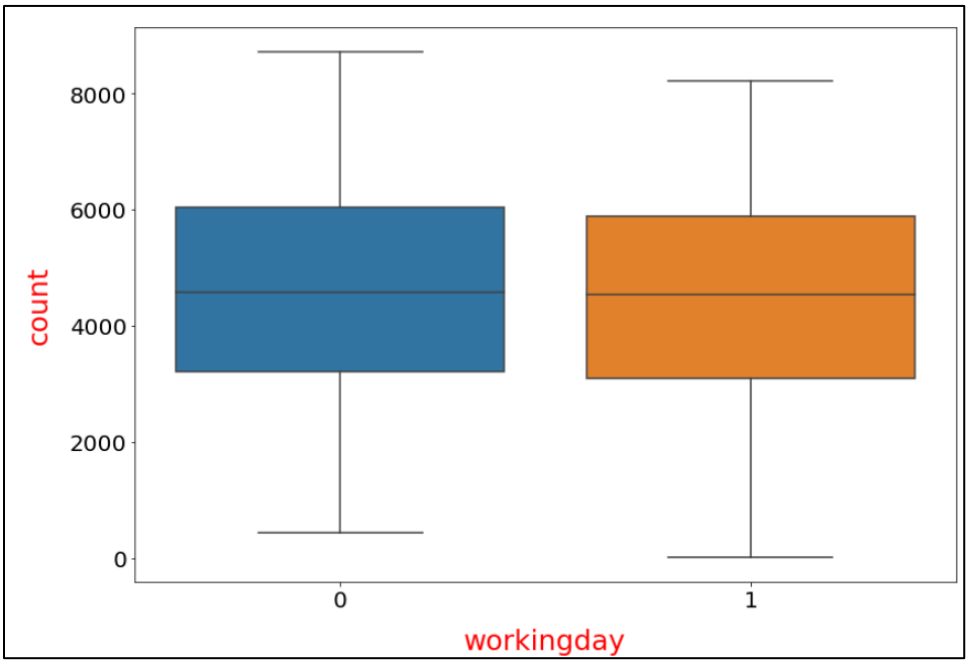
- We can see from the above box plot, count of total rental bikes (Dependent variable 'cnt') **is high in "Friday" followed by "Thursday" in weekdays (Predictor variable) and in weekend, it is high in "Sunday".**

- Whereas the **demand of the rental bikes (Dependent variable 'cnt') is less in other days of the week**.

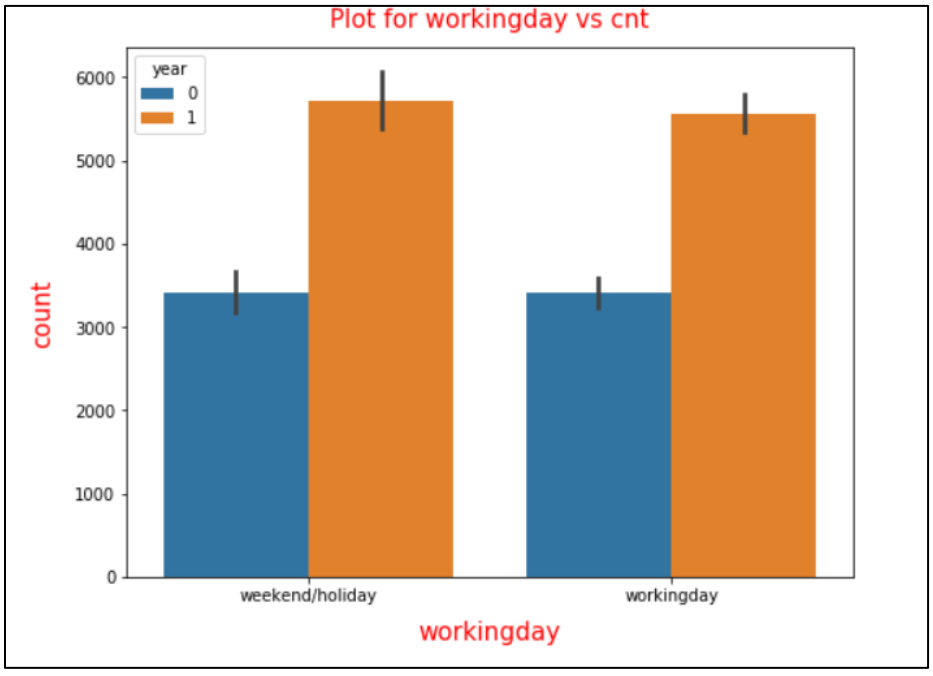


- From the above bar plot we can see that the demand of rental bikes (Dependent variable 'cnt'**) is high in Friday, Thursday and in Saturday respectively in the year 2019.**
- Whereas the **demand of rental bikes (Dependent variable 'cnt') is high in Wednesday, Thursday, Tuesday and in Monday respectively in the year 2018.**

Working Day:

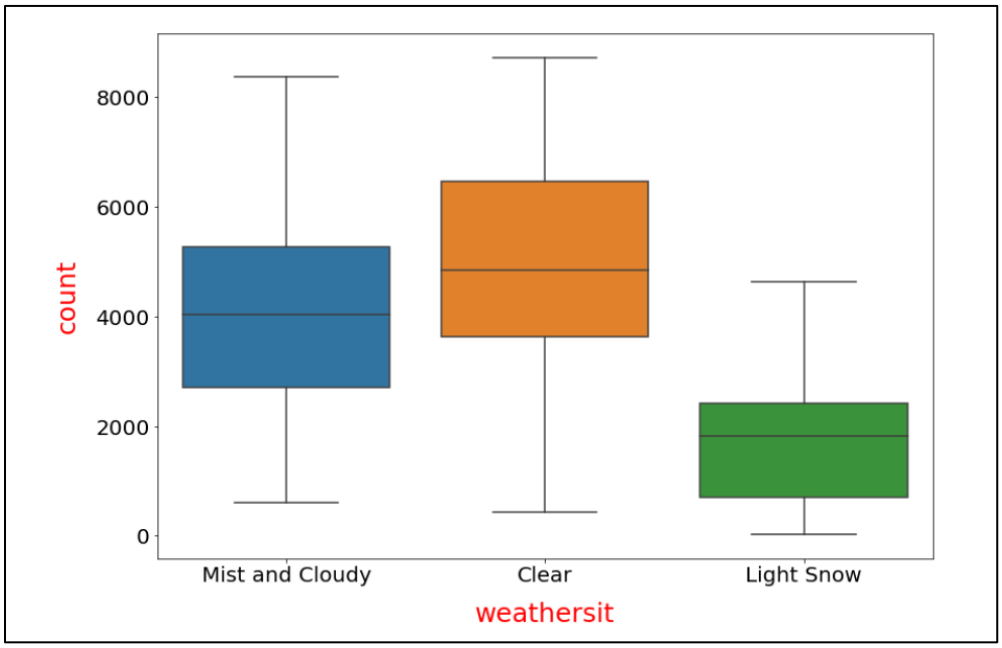


- We can see from the above box plot, count of total rental bikes (Dependent variable 'cnt') **is high in non-working days (Predictor variable) (weekend or in holidays) having 75% of the box lies in around 6000 and upper fence is more than 8000.**
- Whereas the **demand of the rental bikes (Dependent variable 'cnt') is less in working days**.



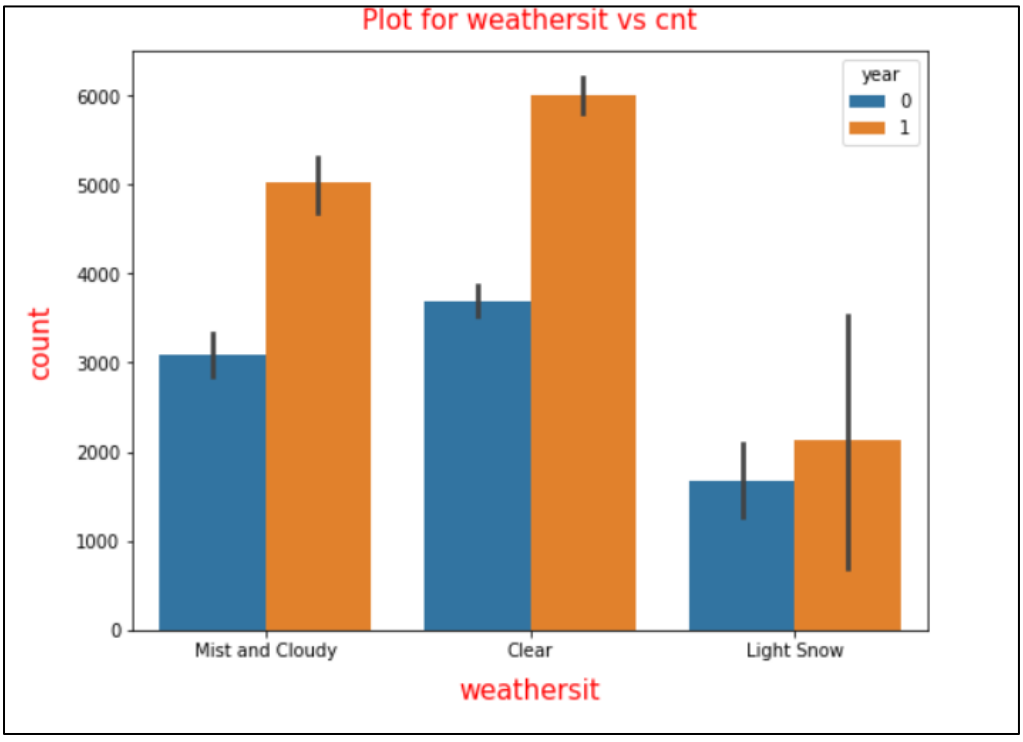
- From the above bar plot we can see that the count of total rental bikes (Dependent variable 'cnt'**) is high when there is weekend or holiday for both the years 2018 and 2019.**

Weather sit:



- We can see from the above box plot, count of total rental bikes (Dependent variable 'cnt') **is high when the sky is clear (Predictor variable), followed by Mist cloudy.**

- Whereas the demand of the rental bikes (Dependent variable 'cnt') **is much less when there is Light snow (Predictor variable).**



- Room the above bar plot we can see that the demand of rental bikes (Dependent variable 'cnt'**) is high when sky is clear for both the years 2018 and in 2019.**

- From the above bar plot we can see that the demand of rental bikes (Dependent variable 'cnt'**) is much less when there is light snow for both the years 2018 and in 2019.**

Linear Regression Model:

Below based on predicted model equation, we have highlighted the categorical variable:

$$cnt=0.1909+0.4777\times temperature+0.0910\times \mathbf{Sept}+0.0621\times \mathbf{summer}+0.0945\times \mathbf{winter}+0.2341\times \mathbf{year}-0.2850\times \mathbf{LightSnow}-0.0787\times \mathbf{MistandCloudy}-0.0554\times \mathbf{spring}-0.0963\times \mathbf{holiday}-0.1481\times windspeed$$

- We can cee the ('cnt' dependent variable) **demand of rental bikes is dependent on the September month, that is on September the count of the rental bikes increased by 0.0910 units keeping all other feature constant.**

- We can cee the ('cnt' dependent variable) demand of rental bikes is dependent on **the Summer season, that is on summer the count of the rental bikes increased by 0.0621 units keeping all other feature constant.**

- We can cee the ('cnt' dependent variable) demand of rental bikes is dependent on the **Winter season as well, that is on winter the count of the rental bikes increased by 0.0945 units keeping all other feature constant.**

- We can cee the ('cnt' dependent variable) demand of rental bikes is **dependent on the Light and Snow weather, that is on Light and Snow whether situation the demand of the rental bikes has decreased by 0.2850 unit keeping all other feature constant.**

- We can cee the ('cnt' dependent variable) demand of rental bikes is **dependent on the Mist and Cloudy weather, that is on Mist and Cloudy whether situation the demand of the rental bikes has decreased by 0.0787 unit keeping all other feature constant.**

- We can cee the ('cnt' dependent variable) demand of rental bikes is **dependent on the Spring season, that is on spring the count of the rental bikes decreased by 0.0554 units keeping all other feature constant.**

- We can cee the ('cnt' dependent variable) demand of **rental bikes is dependent on the holidays, that is on holidays the demand of the rental bikes decreased by 0.0963 units keeping all other feature constant.**

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

We will have non-numeric variables in the data sets. These variables are **also known as categorical variables. Obviously, these variables cannot be used directly in the model, as they are non-numeric**.

While creating a Linear regression model **we use "get_dummies" or we create dummy variable** basically to convert categorical variable to numerical variable, or ***more precisely to deal with Categorical predictor variable***.

Syntax*:* *pandas.get_dummies(data, prefix=None, prefix_sep='_', dummy_na=False, columns=None, sparse=False**, drop_first=False**, dtype=None)*

Basically, ***drop_first=True*** means whether to get **k-1 dummies out of k categorical levels** by removing the first level.

Let's explain the ***above statement with an example***:

Dataset:

	Gender	Name
0	Male	Rishi
1	Female	Sumana
2	Male	Amitabh
3	Male	Prasanth
4	Female	Devika

After creating dummy variables from **'Gender' categorical column** using '***get_dummies'***, we get below data frame using ***drop_first=False as a parameter of get_dummies'***:

	Female	Male
0	0	1
1	1	0
2	0	1
3	0	1
4	1	0

- Where **0 represent not a Female and 1 represent Female** for 'Female' column
- Where **0 represent not a Male and 1 represent Male** for 'Male' Column

So, for categorical column '**Gender'** , we have 2 (k) level, by using ***drop_first=True*** , we get **k-1 dummies out of k categorical levels** by removing the first level as shown below.

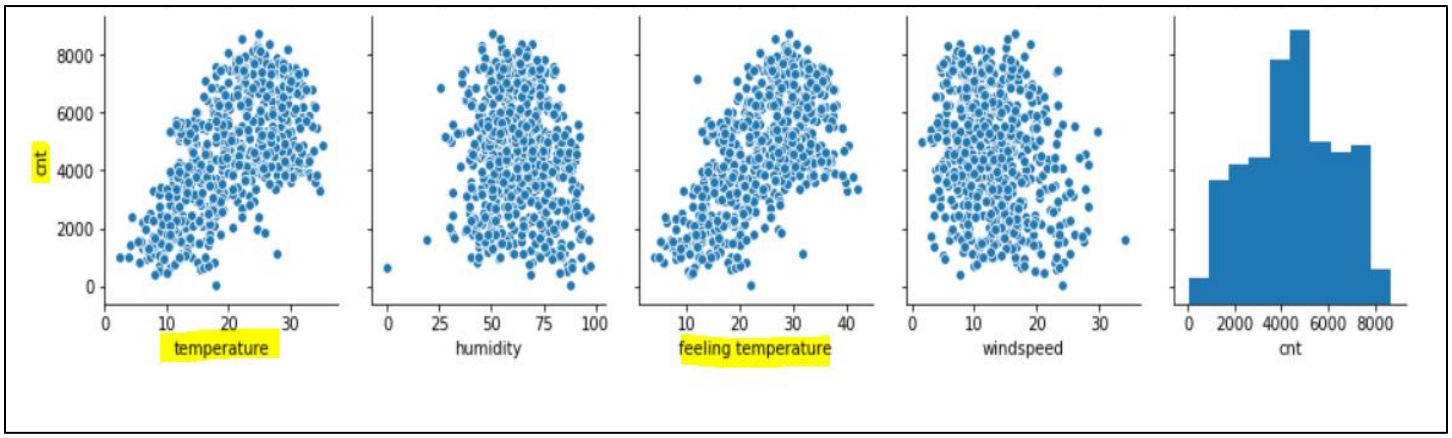
	Male
0	1
1	0
2	1
3	1
4	0

So, without using two separate columns to represent the Gender, we can use only one column 'Male' for the same purpose. In the above data frame, **0 represent Female and 1 represent Male**.

Hence, without increasing the number of variables for representing the same thing which makes the model unnecessary complex, we use ***drop_first=True*** to get **k-1 dummies out of k categorical levels** by removing the first level. We can represent the same thing out of it.

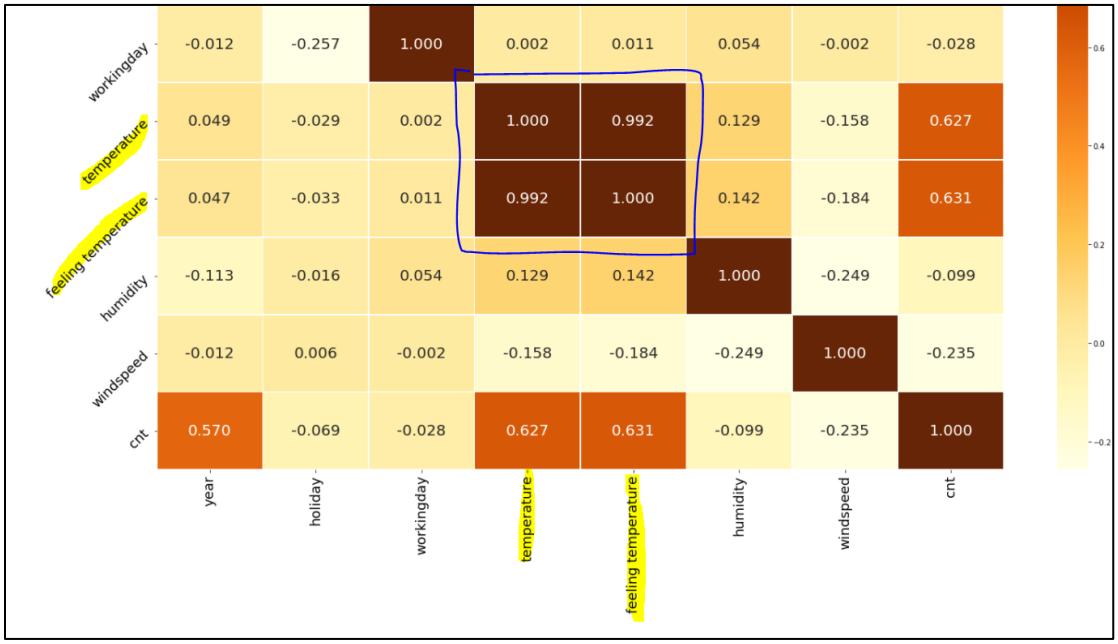
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Looking **at the pair plot** below:



From the above **pair plot Temperature (temp) and Feeling temperature (atemp) has highest correlation with the target variable count of the rental bikes ('cnt').**

But as **Feeling temperature (atemp) is highly correlated with Temperature (temp**) as well which may will introduce **multicollinearity effect in the model,** which we can see from the below heatmap**, hence we have dropped the Feeling temperature (atemp) column**.

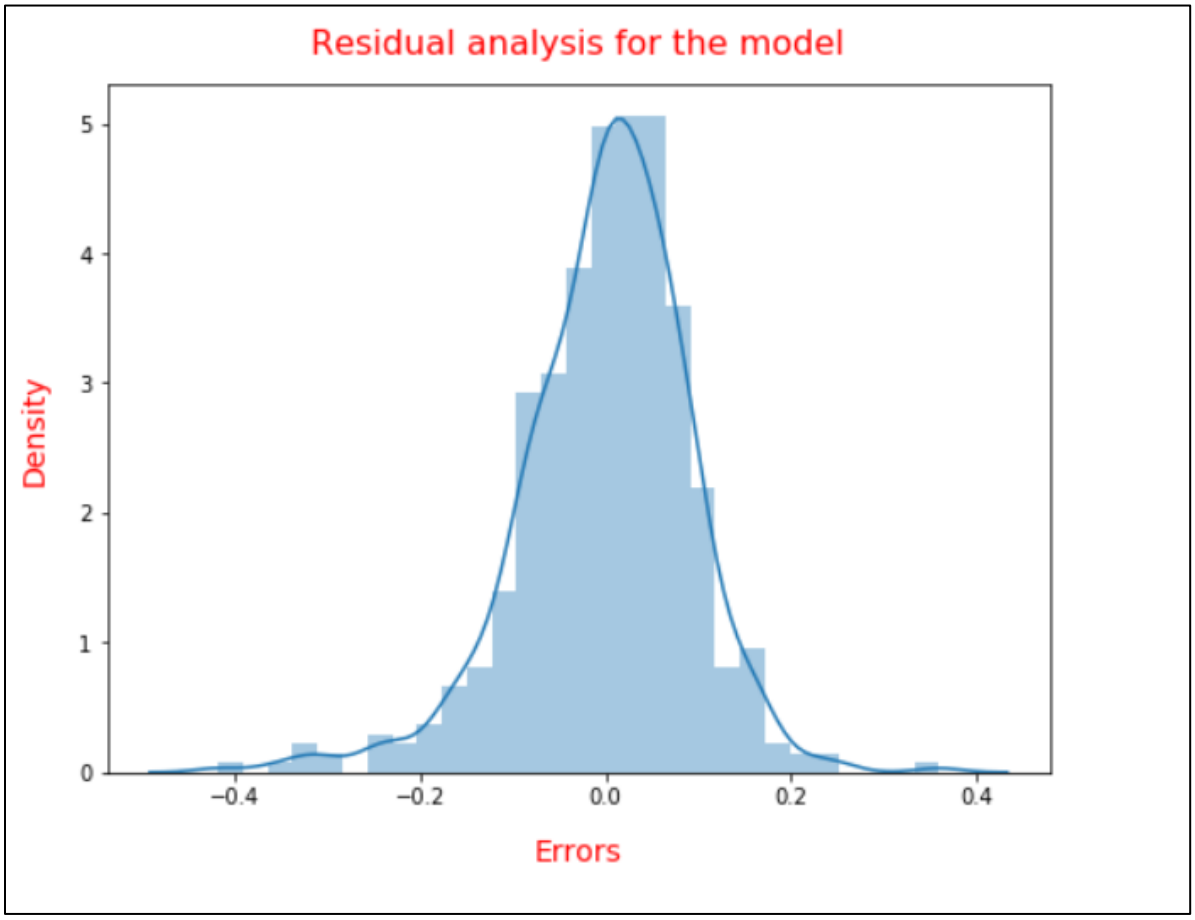


Hence, now we have only **Temperature (temp) which has highest correlation with the target variable count of the rental bikes ('cnt').**

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

Below **assumption of the Linear Regression we have validated** after building the model on training set:

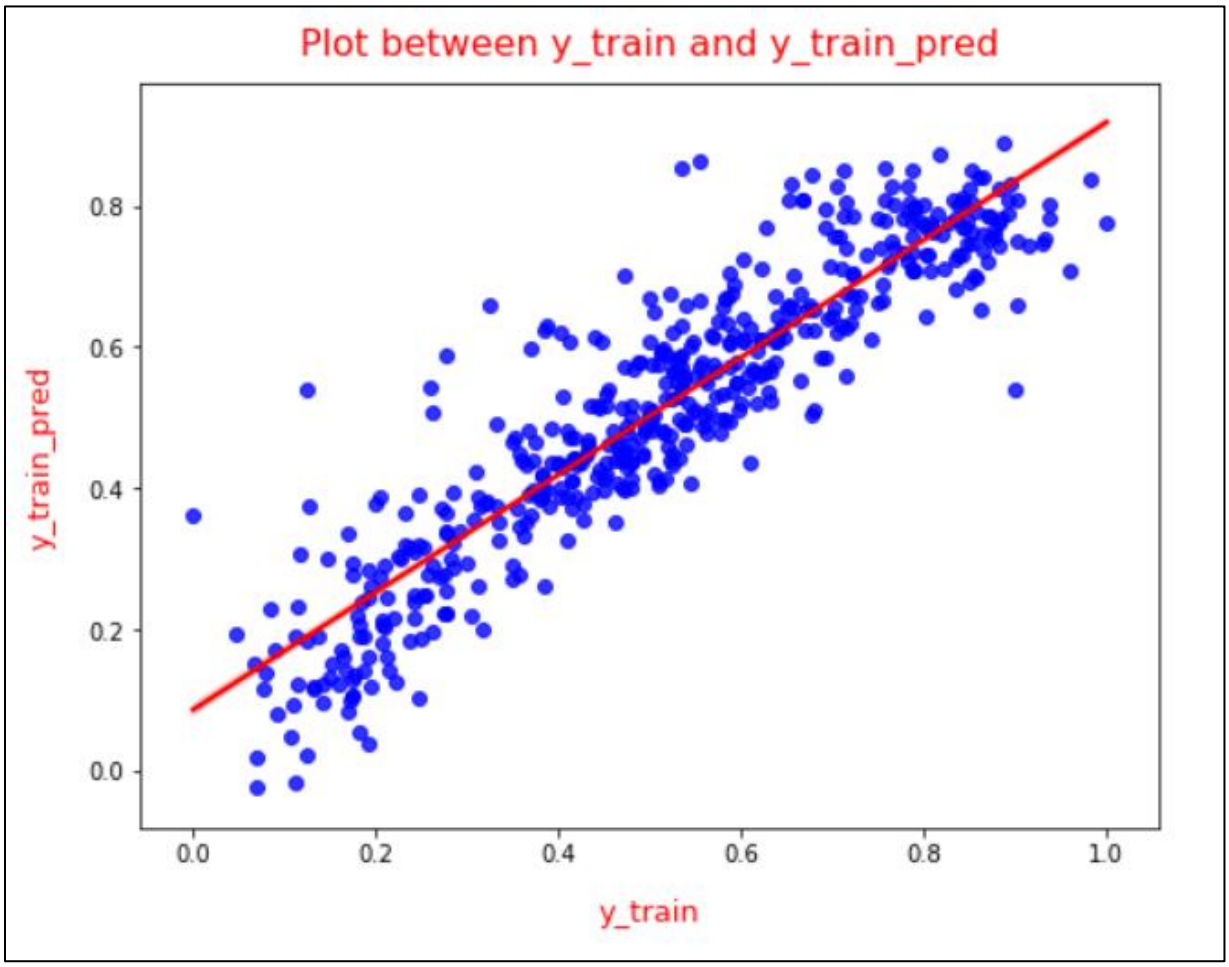
- **Normality:** Error terms should distribute normally with zero mean.



From the **above distribution plot, we can see that the error terms are normally distributed with mean 0**.

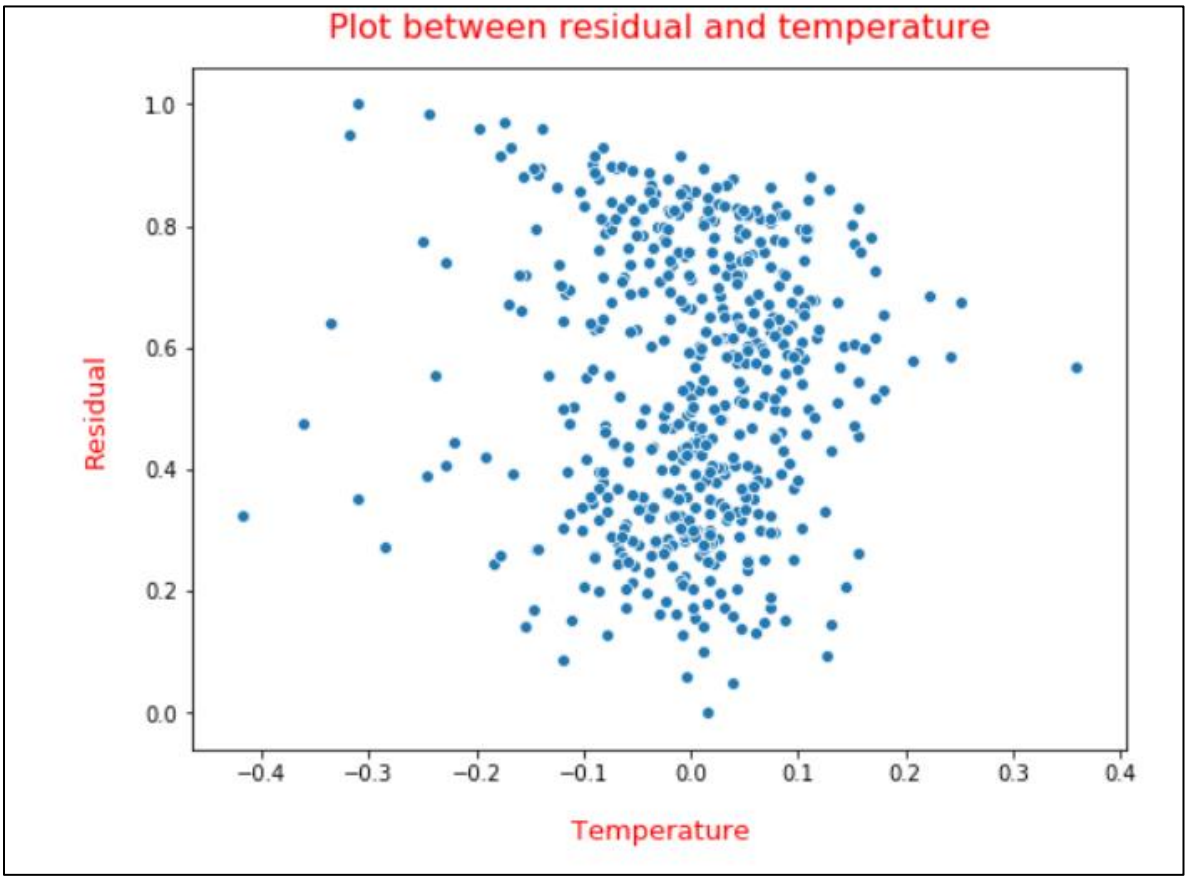
This is one of the assumptions for linear regression, which is true in our case.

- **<u>Homoscedasticity</u>:** The variance of error terms should be constant.



The above **regression plot between y_train and y_train_pred are equally distributed along the regression line. Hence, the variance is constant.** Hence, the assumption of equal variance/ **Homoscedasticity** becomes true for our model.

- **<u>Independence</u>:** Error Terms are independent of each other



From the **above scatter plot we can see that the error terms are independent of each other, no such pattern is there within residuals when we plotted it against a predictor variable temperature**. This is one of the assumptions for linear regression, which is true in our case.

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

Linear Regression Model:

Below based on predicted model equation, we have highlighted the top 3 features contributing significantly towards explaining the demand of the shared bikes**:**

cnt=0.1909**+0.4777×temperature**+0.0910×Sept+0.0621×summer+0.0945×winter**+0.2341×year−0.2850×LightSnow**−0.0787×MistandCloudy−0.0554×spring−0.0963×holiday−0.1481×windspeed

Model Interpretation for Temperature Feature:

We can cee the ('cnt' dependent variable) **demand of rental bikes is very much dependent on the temperature that is one unit rise in temperature, increases the count of the rental bikes by 0.4777 units keeping all other feature constant**.

Business Justification for Temperature Feature:

A US bike-sharing provider **Boom Bikes can focus more on Temperature, increase in temperature will increase the demand of bikes.**

Model Interpretation for Light and Snow weather Feature:

We can cee the ('cnt' dependent variable) demand of rental bikes is dependent on the Light and Snow weather, **that is on Light and Snow weather situation the demand of the rental bikes has decreased by 0.2850 unit keeping all other feature constant.**

Business Justification for Light and Snow weather Feature:

Now seeing to weather sit variable, we have got **negative coefficients** for Light snow weather, **Business can give some offers/Discounts to increase the demand.**

Model Interpretation for Year Feature:

We can cee the ('cnt' dependent variable) demand of rental bikes is dependent on the years, **that is every year the demand of the rental bikes has increased by 0.2341 unit keeping all other feature constant.**

Business Justification for year Feature:

We can see demand for bikes was more in 2019 than 2018, **As there is increase in 2019 and might be facing dips in their revenues due to the ongoing Corona pandemic** and by the time Corona Virus reduces the things will be better.

Business should do a better marketing and **if more people comes to know about boom bikes then every year the demand will increase more and more.**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

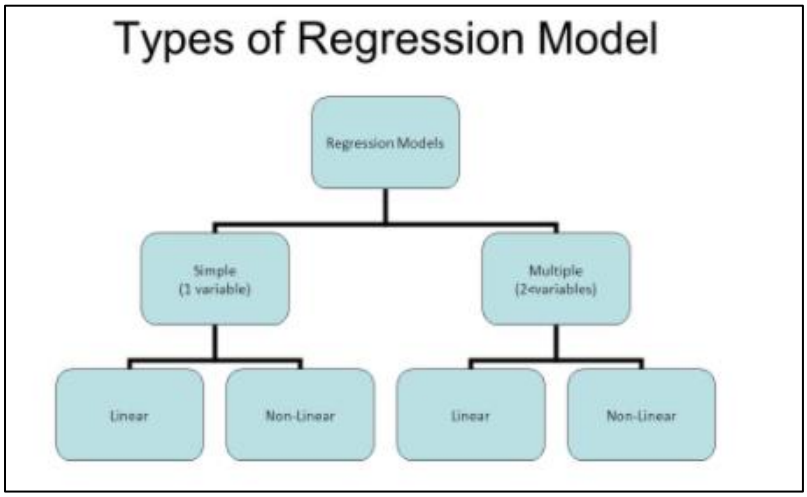
Before we come to, what is Linear Regression Algorithm, let us first understand what Regression Analysis is:

Regression Analysis: Regression analysis consists of a set of machine learning methods, specifically it is one of the ***Supervised Machine learning algorithms*****,** that allow us to predict a continuous outcome variable (y) based on the value of one or multiple predictor variables (x).

The goal of regression model is to build a mathematical equation that defines y as a function of the x variables. This equation can be used to predict the outcome (y) based on new values of the predictor variables (x).

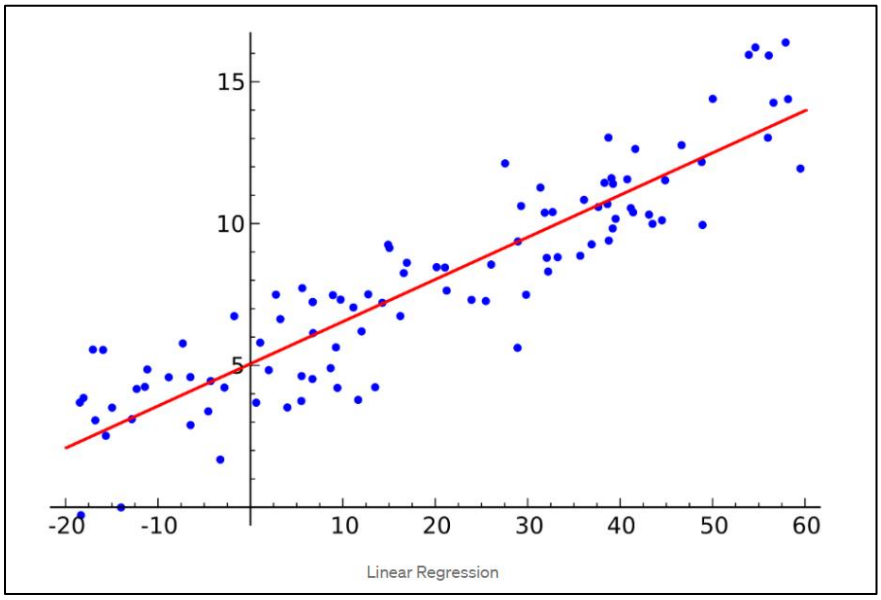
Output Of regression Analysis: The output variable to be predicted is a **continuous variable**, e.g. salary of Employee.

Types of Regression Algorithm:



Linear Regression: Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a **linear relationship** between x (input) and y(output). Hence, the name is **Linear** Regression.

In the figure below, the regression line is the best fit line for our model. This represent the linear regression.



There are two types of **linear regression** that we have learnt so far:

- Simple linear regression
- Multiple linear regression

Simple linear regression: Simple linear regression, plots one independent variable X against one dependent variable Y. Technically, in regression analysis, the independent variable is usually called the predictor variable and the dependent variable is called the response or Target variable.

Denotes the linear relationship between ***only one Predictor (X) variable and only one Target (Y) variable***.

Equation for simple linear Regression: $$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon$$

yi=dependent variable

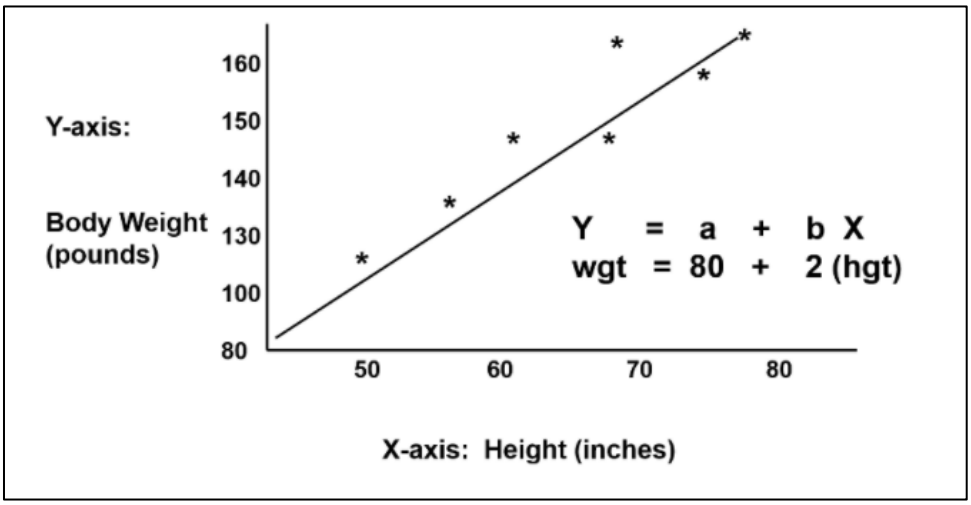
xi=explanatory / Predictor variables

β0=y-intercept (constant term)

β1=slope coefficients for each explanatory variable

ϵ=the model's error term (also known as the residuals)

Below graph represent the Simple linear regression plot where **independent or predictor variable is Height (X)** and **Dependent or target variable is Body weight (Y).**



Multiple linear regression: Multiple linear regression, plots one dependent variable Y against multiple independent variable X Technically, in regression analysis, the independent variable is usually called the predictor variable and the dependent variable is called the response or Target variable.

Denotes the linear relationship between ***multiple Predictor (X) variable and one Target (Y) variable.***

Equation for Multiple linear Regression:

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_px_{ip}+\epsilon$$

where, for i=p observations:

yi=dependent variable

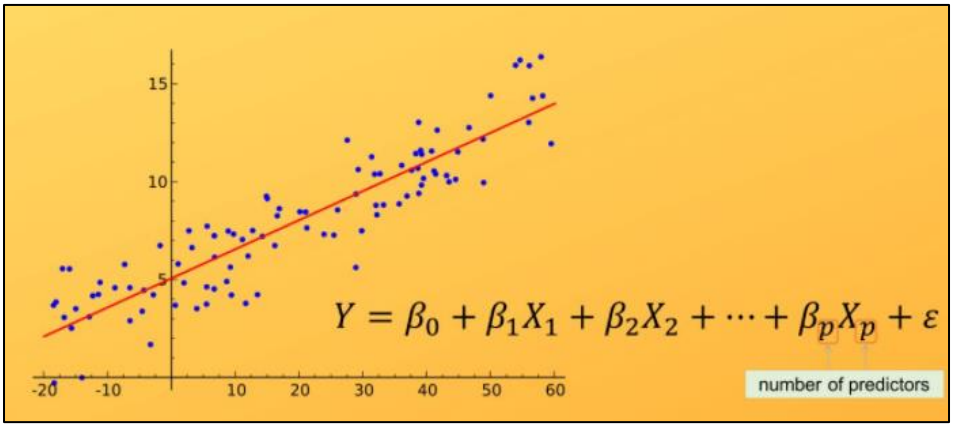
xi=explanatory / Predictor variables

β0=y-intercept (constant term)

βp=slope coefficients for each explanatory variable

ϵ=the model's error term (also known as the residuals)

Below graph represent the **multiple linear regression** plot.



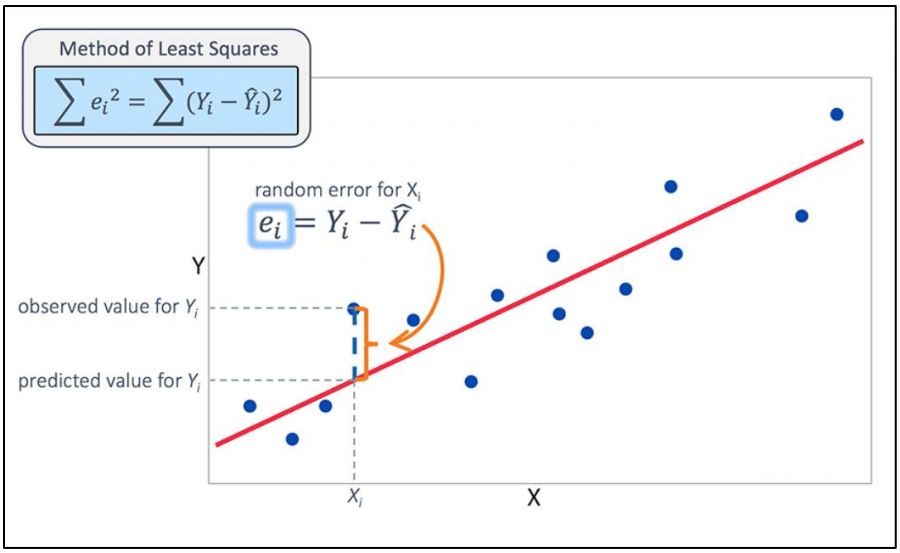
Now we must understand how to plot or how to get the **best fit line in the above simple or in multiple linear** regression model.

Best Fit line: Line of best fit refers to a line through a scatter plot (shown above) of data points that best expresses the relationship between those points. Statisticians typically use the ordinary least squares method to arrive at the geometric equation for the line, either though manual calculations or though programming.

It is found by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable

Ordinary least square or Residual Sum of squares (RSS) — Here the cost function is the (y(i) — y^ (pred)) ² ***which is minimized to find that value*** of β0 and β1, to find that best fit of the predicted line.

Below is the **residual** (y(i) — y^ (pred)) shown graphically:



Residual sum of Squares:	RSS= ∑Ni=1(residual)2
	RSS= ∑Ni=1(ei)2
	RSS= ∑Ni=1(Yi−Yi^)2

where:

Y is the original target score

Y^ is the predicted target score

e is the residual

Below is the RSS equation for Simple and multiple linear regression model:

Simple Regression

$$RSS = \sum_{i=1}^{N}(Y_i - \hat{B}_0 - \hat{B}_1 X_i)^2$$

Multiple Regression

$$RSS = \sum_{i=1}^{N}(Y_i - \hat{B}_0 - \hat{B}_1 X_{i1} - \cdots - \hat{B}_n X_{in})^2$$
$$= \sum_{i=1}^{N}(Y_i - \hat{B}_0 - \sum_{j=1}^{P} \hat{B}_j X_{ij})^2$$

Strength of Linear Regression Model:

The strength of the linear regression model can be assessed using 2 metrics:

1. R2 or Coefficient of Determination
2. Residual Standard Error (RSE)

R2 or Coefficient of Determination: R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

Mathematically, it is represented as: **R2 = 1 - (RSS / TSS)**

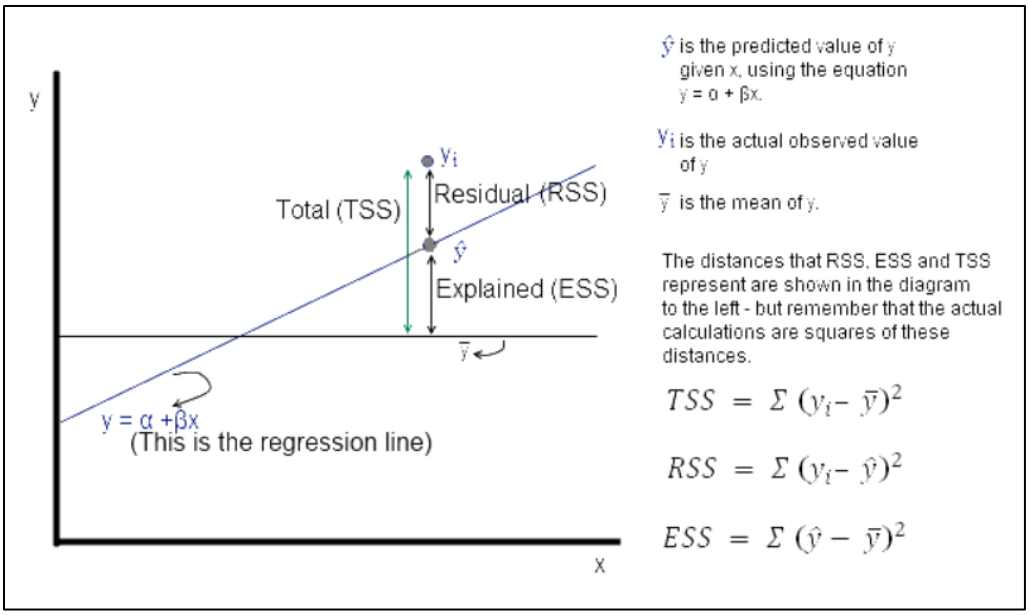
where RSS= Residual sum of square

TSS = Total Sum of square

TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable.

Mathematically, TSS is: $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Consider the diagram below: Yi is the actual observed value of the dependent variable, y-hat is the value of the dependent variable according to the regression line, as predicted by our regression model. We want to get is a feel for is the variability of actual y around the regression line, i.e. the volatility of ϵ. This is given by the distance yi minus y-hat. Represented in the figure below as RSS. The figure below also shows TSS and ESS



RSE: RSE (Residual square error) is a measure of lack of fit of the model to the data at hand. In simplest terms, if the RSE value is very close to the actual outcome value, then your model fits the data well. If there is a large difference between the values, then the model does not fit the data well.

$$RSE = \sqrt{\frac{RSS}{df}}$$

df = n-2, where n = number of data-points

Assumption of Linear Regression: There are five important assumptions associated with a linear regression model:

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Error terms are independent of each other.
4. **Normality**: Error Terms are normally distributed.
5. **Multicollinearity**: The independent variables are linearly independent of each other.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by **the statistician Francis Anscombe** to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

The essential thing to note about these datasets is that ***they share the same descriptive statistics***. But things change ***completely***, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

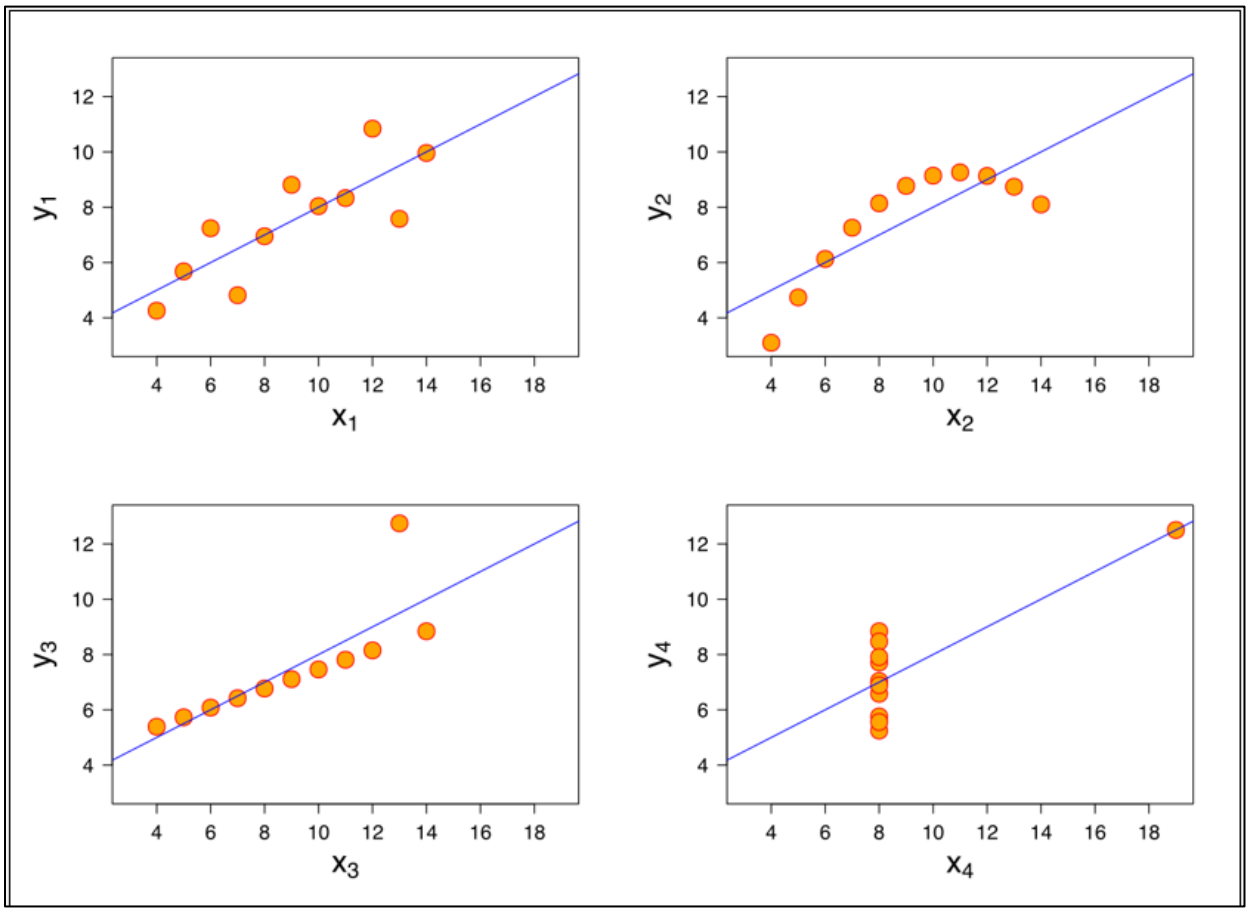
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Quartet's Summary Stats

The summary statistics show that the means and the standard deviation are identical for x and y across the groups:

- Mean of x is 9 and mean of y is **7.50** for each dataset.
- The correlation coefficient between x and y is **0.816** for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well, but each dataset is telling a different story:



Interpretation of the above Graphs:

- Dataset I appear to ***have clean and well-fitting linear models***.
- Dataset II is ***not distributed normally***.
- Dataset III, the distribution is linear, but ***the calculated regression is thrown off by an outlier***.
- Dataset IV shows that ***one outlier is enough to produce a high correlation coefficient***.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

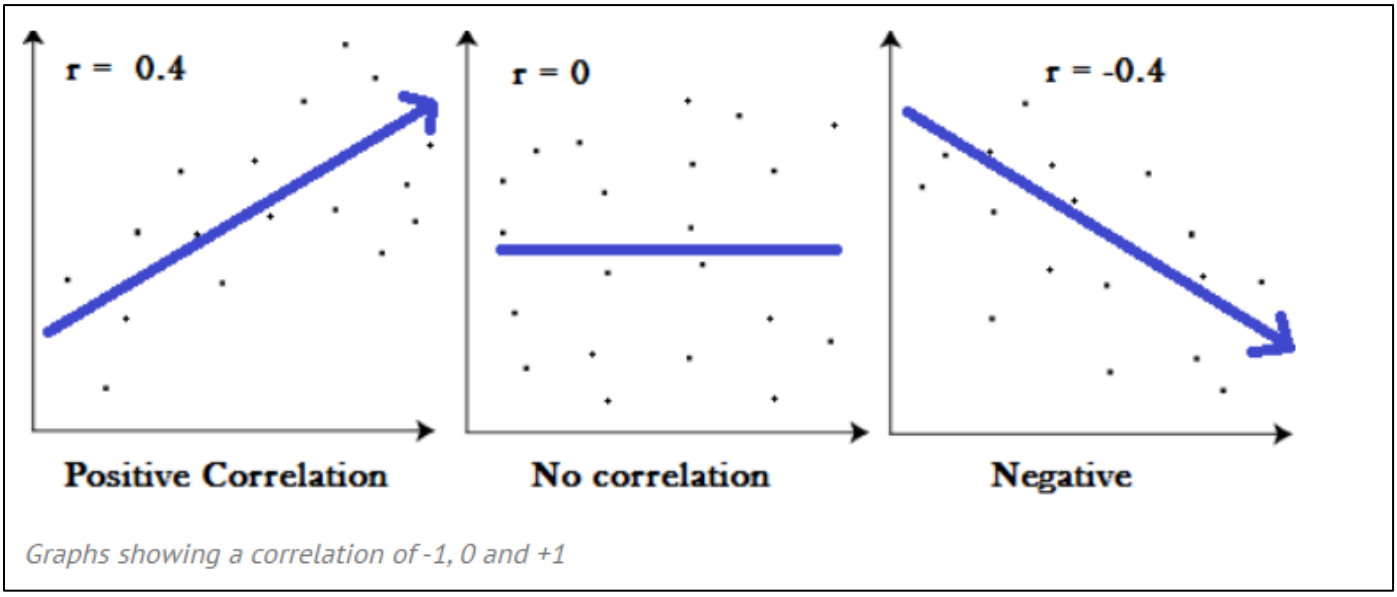
3. What is Pearson's R?

Let us first discuss what is correlation coefficient before we describe Pearson's R

Correlation coefficients: Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is **Pearson's. Pearson's correlation** (also called **Pearson's R**), it is a correlation coefficient commonly used in ***linear regression.***

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas ***return a value between -1 and 1***, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



Graphs showing a correlation of -1, 0 and +1

Interpretation of the above graph:

- ***A correlation coefficient of 1*** means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. **For example**, shoe sizes go up in (almost) perfect correlation with foot length.
- ***A correlation coefficient of -1*** means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. **For example**, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- ***A correlation coefficient 0*** means that for every increase, there isn't a positive or negative increase. The two just aren't related.

Pearson Correlation

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation is ***the Pearson Correlation***. The full name is the **Pearson Product Moment Correlation (PPMC)**. It shows ***the linear relationship between two sets of data***.

Pearson's correlation coefficient formula -

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\, n\sum x^2 - (\sum x)^2 \,][\, n\sum y^2 - (\sum y)^2 \,]}}$$

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

For example: Up till a certain age, a child's height will keep increasing as his/her age increases. Of course, his/her growth depends upon various factors like genes, location, diet, lifestyle, etc. This could be a good example of **linear correlation** and which can ***be measure by Pearson's correlation coefficient***.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling: Most of the times, dataset will contain features highly varying in magnitudes, units and range. Feature ***Scaling is a technique to standardize / Normalize the independent features*** present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

It should be **done specifically after we split the data into train and test set**, and it is important to do so because ***we don't want the model to learn anything from train dataset*** when it is predicting on test dataset.

Example: If an algorithm is not using feature scaling method then it can consider the value 300 **cm** to be greater than 5 **meters** but ***that's not true*** and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes.

Why scaling is performed: Real world dataset contains features that highly vary in magnitudes, units, and range. Scaling should be performed when the scale of a feature is irrelevant or misleading. It basically helps to normalize the data within a range.

Advantage of Scaling:

- It helps to interpret the coefficient properly
- It helps for faster convergence of gradient descent

What is Normalization?

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging **between 0 and 1**. It is also known as **Min-Max scaling**.

Normalization equation -

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, **Xmax and Xmin are the maximum and the minimum values** of the feature respectively.

Above equation Interpretation:

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1.

What is Standardization?

Standardization is another scaling technique where the values are centered around ***the mean (0) with a unit standard deviation (1)***. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Standardization equation -

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature.

Note that in this case, the values are not restricted to a range.

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

Let me first explain what is VIF

VIF: Variance inflation factor (VIF) is a measure of the amount of ***multicollinearity*** in a set of multiple regression variables.

Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model.

Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model, it's presence can adversely affect your regression results. The **VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity** in the model.

VIFs are **calculated by taking a predictor and regressing it against every other predictor** in the model.

VIF Equation -

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

VIF Interpretation:

- VIF starts at 1 and has no upper limit
- VIF = 1, no correlation between the independent variable and the other variables
- VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others

Variance inflation factors range from 1 upwards. The numerical value for VIF tells you what percentage the variance is inflated for each coefficient.

Now let's come back to the point that **when can VIF be infinity**?

If there is perfect correlation ***between each independent variable (predictor) with one another (correlation coefficient is perfectly 1.00), then VIF = infinity***. A large value of VIF indicates that there is a correlation between the variables.

Even if R2 score is more which means this feature is correlated with other features. ***When R2 reaches to 1.00, the also the VIF reaches infinity.***

Let's see below **example of infinite VIF**:

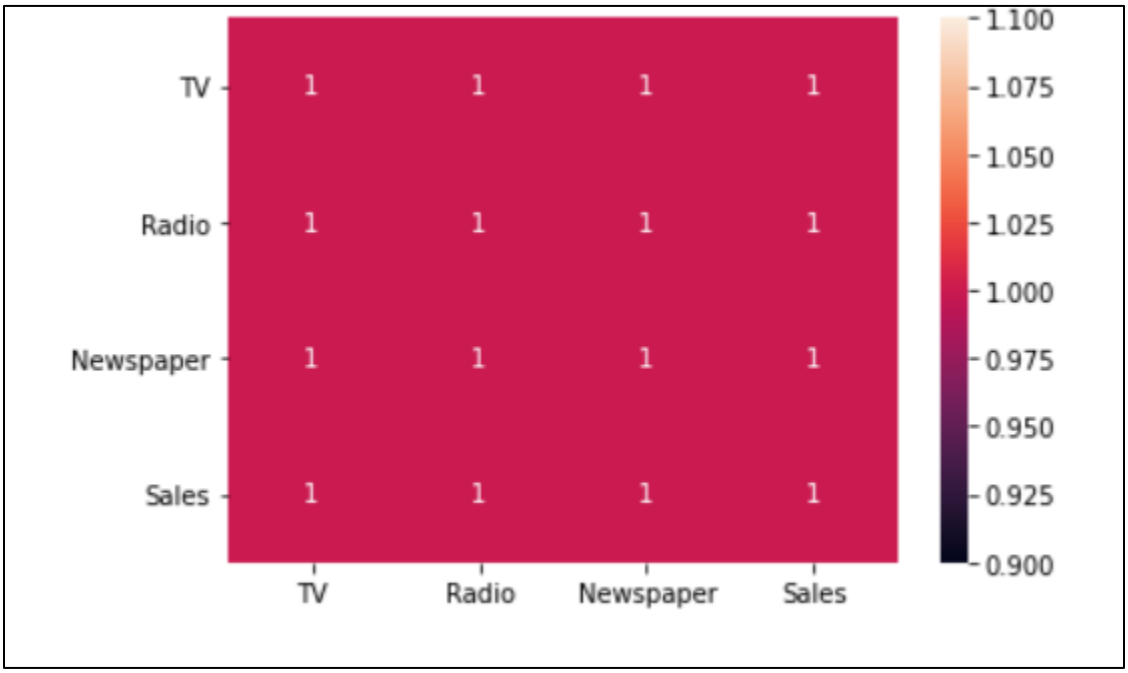
Suppose we have a dataset as below:

	TV	Radio	Newspaper	Sales
0	100	200	300	500
1	200	400	600	1000
2	300	600	900	1500
3	400	800	1200	2000
4	500	1000	1500	2500

The above dataset is basically an advertising dataset where there is **different predictor variable (X) such as TV, Radio and Newspaper. The target variable (y) is Sales.**

The above dataset gives us the how Sales is increasing or decreasing based on TV, Radio and Newspaper budget in an advertising company.

Now let see the correlation between **all predictor variable(X) by a heatmap**.



Above we can see the ***correlation coefficient for all predictor variables(X) (TV, Radio and Newspapers) with each other is perfectly 1***, which means those predictor variables are strongly positively correlated with each other.

Now let's **calculate the VIF** and see the result.

	features	VIF
0	TV	inf
1	Radio	inf
2	Newspaper	inf

Now as per as the explanation here <u>we can see that the VIF is "inf" (Infinity)</u>, because all the predictor variables are perfectly correlated with each other with a correlation value of 1.

Even from the ***above heatmap we can that the independent variable (X) is perfectly correlated with each other (with other independent variable) with a correlation value of 1***.

Hence let's build a model by **taking one independent variable (X) on y-axis against other independent variable (X), basically this is the theoretical procedure how we calculate the VIF (** for checking multicollinearity we check the values of VIF , which is basically calculating the R2 score for one X Variable against other X Variables and then using (**1/1-R2 score**) we get the VIF) , see the model summary on the above dataset.

OLS Regression Results

Dep. Variable:	TV	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	1.079e+32
Date:	Wed, 21 Oct 2020	Prob (F-statistic):	5.16e-64
Time:	13:14:27	Log-Likelihood:	144.81
No. Observations:	5	AIC:	-287.6
Df Residuals:	4	BIC:	-288.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Radio	0.1538	1.48e-17	1.04e+16	0.000	0.154	0.154
Newspaper	0.2308	2.22e-17	1.04e+16	0.000	0.231	0.231

Omnibus:	nan	Durbin-Watson:	0.208
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.474
Skew:	-0.670	Prob(JB):	0.789
Kurtosis:	2.310	Cond. No.	2.12e+16

From the above model summary, we can see ***that the "TV" independent (X) is also perfectly correlated with all the predictor variable such as "radio" and "Newspaper" (X) with a correlation value of 1, hence R2 score value reaches to 1.000 and due to which VIF reaches to "inf" (infinity).***

VIF = 1 / (1 – R2)

When R2 score reaches 1.000 in the above equation, VIF reaches infinity.

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

Q-Q Plots: Quantile-Quantile plots are plots of two quantiles (theoretical and sample) against each other. A quantile is a fraction where certain values fall below that quantile. It is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly follows a Normal Distribution or not.

Procedure of plotting a Q-Q plot:

1. ***Order the items from smallest to largest***.
2. ***Draw a normal distribution curve.*** Divide the curve into n+1 segment.
3. ***Find the z-value (cut-off point) for each segment***. These segments are *areas*, so we need to refer to Z-table to find the corresponding Z- score.
4. Plot data set values (Point 1) against normal distribution cut-off points (Point 3).

Interpretation:

If we get almost straight line on this Q-Q plot indicates the data is **approximately normally distributed**.

Now let's talk about the importance of a Q-Q plot in linear regression.

In the linear regression model, ***we can use this QQ plot in the finding whether our model error terms or residual are normally distributed or not.***

For example:

Suppose we built a linear regression model and based on the that model we got the residual as follows:

Residuals

7.19

6.79

6.31

5.89

5.79

5.19

4.5

4.25

3.77

- Now we are sorting the ***residual in ascending order***:

Residuals

3.77

4.25

4.5

5.19

5.79

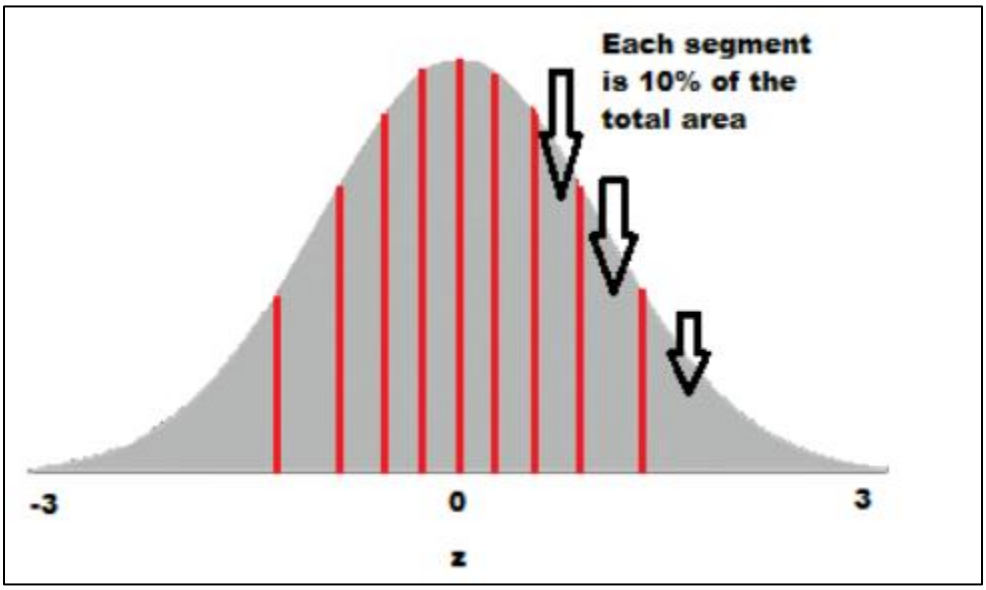
5.89

6.31

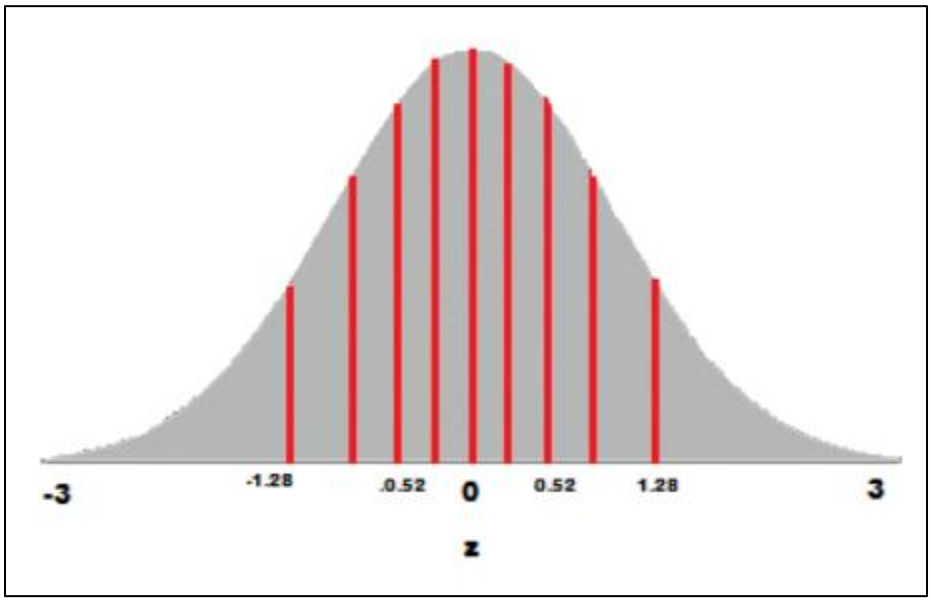
6.79

7.19

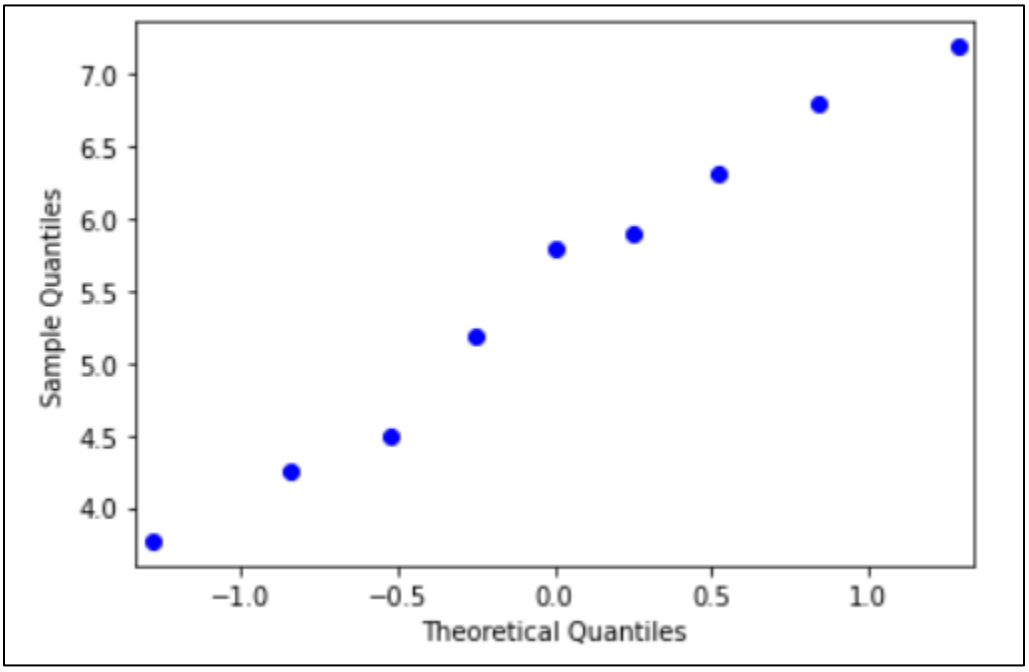
- Draw a normal distribution curve. ***Divide the curve into n+1 segment, here it is total 10 segments***



- ***Find the z-value (cut-off point)*** for each segment, we need to refer the Z table for it.



- *Plot data* ***set values (Point 1) against normal distribution cut-off points***



Interpretation:

From above we can see that residual points in the **Q Qplot falls in almost a Straight line**, ***hence we can tell that the residual or error terms is normally distriuted.***