BOOM-BIKE-SALES-PREDICTION-USING-LINEAR-REGRESSION

Problem Statement:

A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system.
A US bike-sharing provider Boom Bikes has recently suffered considerable dips in their revenues due to the ongoing Corona pandemic. The company is finding it very difficult to sustain in the current market scenario. So, it has decided to come up with a mindful business plan to be able to accelerate its revenue as soon as the ongoing lockdown comes to an end, and the economy restores to a healthy state.
In such an attempt, Boom Bikes aspires to understand the demand for shared bikes among the people after this ongoing quarantine situation ends across the nation due to Covid-19. They have planned this to prepare themselves to cater to the people's needs once the situation gets better all-around and stand out from other service providers and make huge profits.
They have contracted a consulting company to understand the factors on which the demand for these shared bikes depends. Specifically, they want to understand the factors affecting the demand for these shared bikes in the American market. The company wants to know:
1. Which variables are significant in predicting the demand for shared bikes?
2. How well those variables describe the bike demands
Based on various meteorological surveys and people's styles, the service provider firm has gathered a large dataset on daily bike demands across the American market based on some factors.

Business Goal:

You are required to model the demand for shared bikes with the available independent variables. It will be used by the management to understand how exactly the demands vary with different features. They can accordingly manipulate the business strategy to meet the demand levels and meet the customer's expectations. Further, the model will be a good way for management to understand the demand dynamics of a new market.

Author : **GONDESI RUDRA DEEPAK**
Aim : **TO PREDICT THE DEMAND FOR SHARED BIKES AND REQUIRED TO MODEL THE DEMAND FOR SHARED BIKES**
Keywords: **PYTHON, EDA, DATA VISUALIZATION, ML(LINEAR RERESSION)**

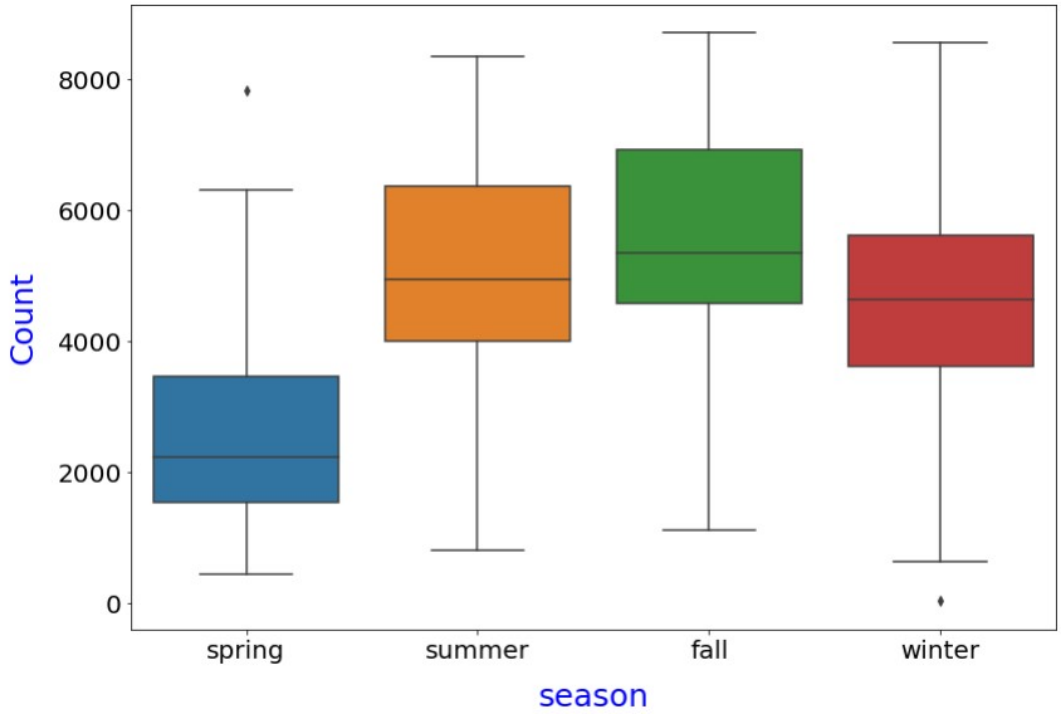
Assignment-based Subjective Questions:

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

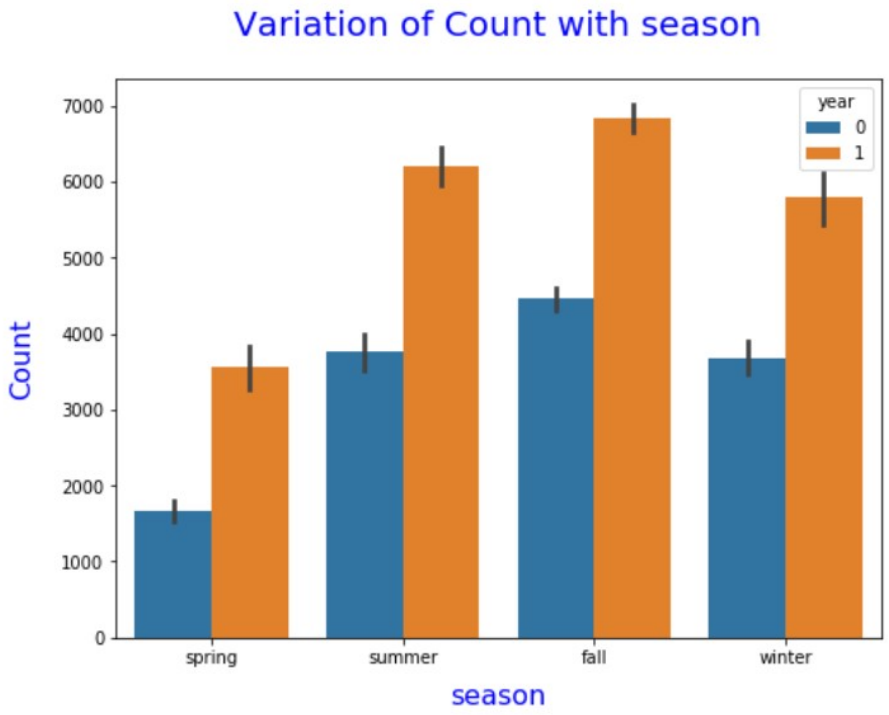
Solution

For analyzing the effect of categorical variables on the dependent variable we have performed Bivariate Analysis on 'season', 'year', 'month', 'holiday', 'weekday', 'workingday', 'weathersit' columns against target variable 'Bike Rental count'.

Season vs. Count:

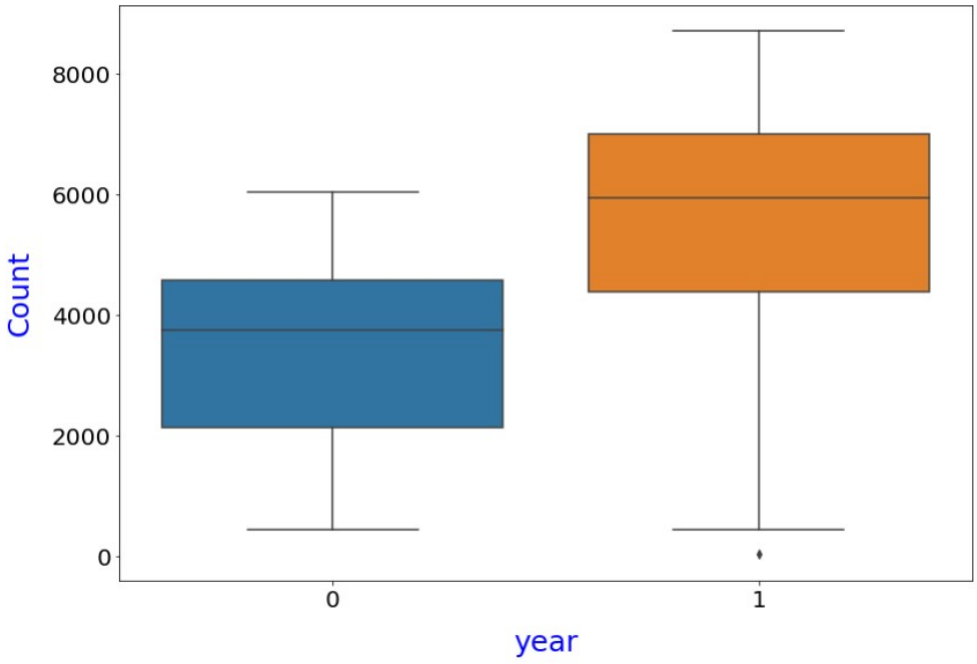


- From the above box plot between season and count, we can see that **count of bike rentals** is **very high** in season **fall**, having its 50% value is more than 5000 (approx.) and 95% is more than 8000.
- Whereas the season **spring** is having **very low bike rental count**, with 75% of box is below 4000.
- From the above box plot between

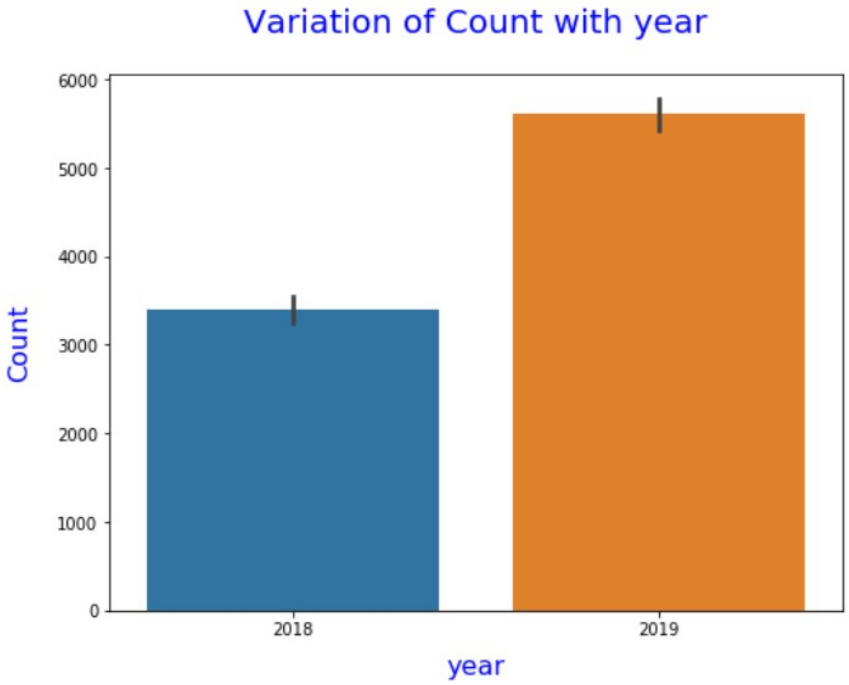


- From above bar plot we can see that, for both the year of **2018** and **2019**, the bike rental count is **high** in season **fall**.
- Between the two years, **2019** is having more rental counts, i.e., more than 6500 (approx.), comparing year **2018**, where the count is more than 4000 (approx.) in **fall**.
- Whereas, the bike rental count is **low** in **spring** for both 2018 and 2019.

year vs. count:

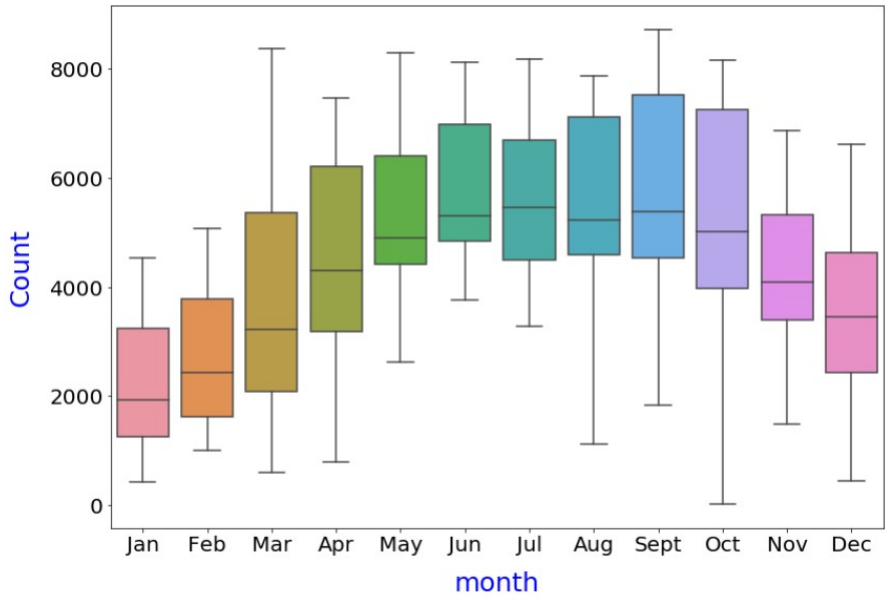


year and count, we can see that **count of bike rentals** is **very high** in the year **2019**, having its 95% value more than 8000 and 25% box value is more than 4000(approx.) •
Where, the 50% box value for 2018 is less than 4000.

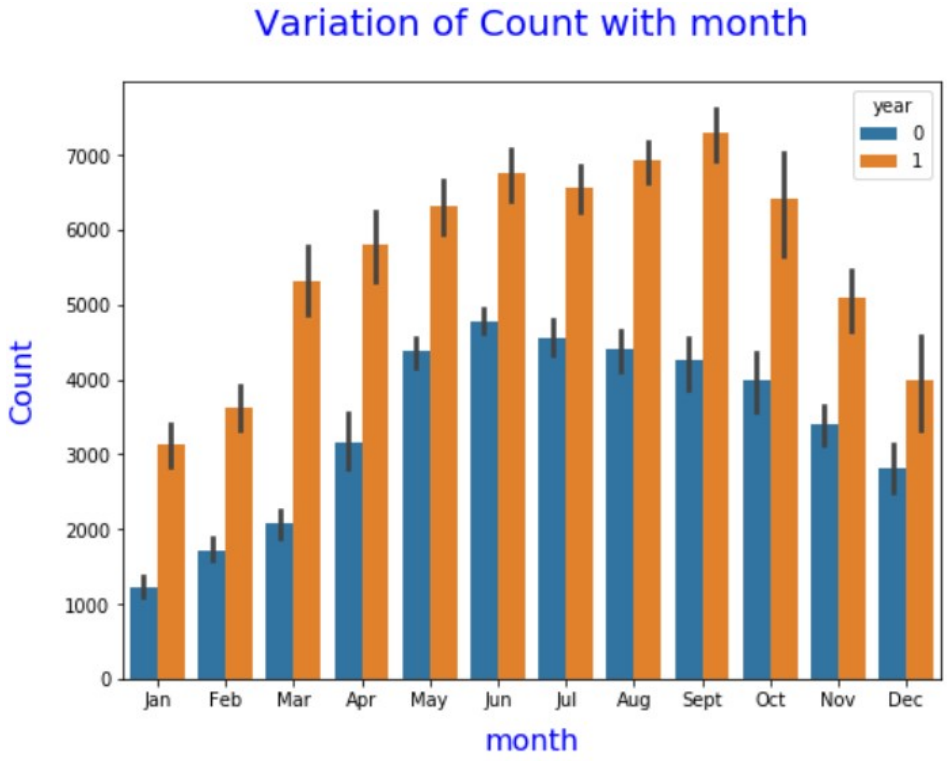


- From the above bar plot, it is clearly seen that year **2019 is having more bike rentals than 2018**.
- Where in **2018** the rental count is around 3500 (approx.), in **2019** the rental count are increases more than 5500 (approx.).

- **<u>month vs. count:</u>**

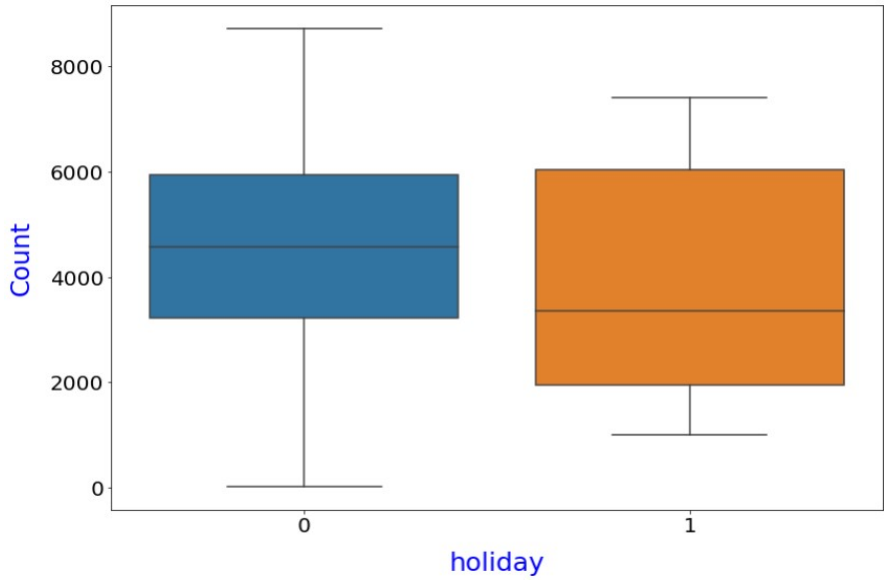


- From the above box plot between **high** month and count, we can see that **count of bike rentals** is **very** for the month **September**, **October** and **August**.
- Whereas, the bike rental count is very low for the month of **January**, **February** and **December**.



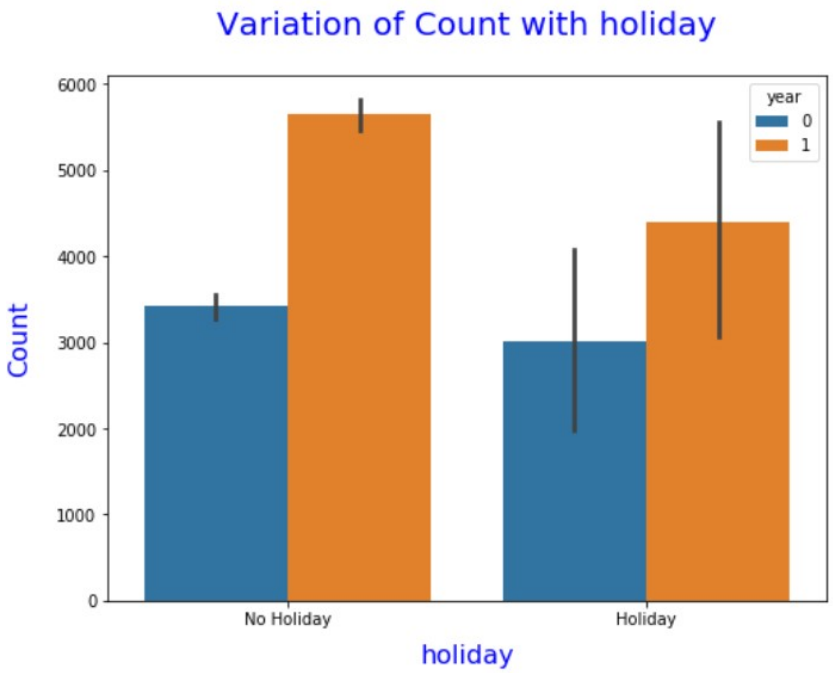
- From the above bar plot, we can see that, for year **2019**, the rental count is **high** in the month of **September**.
- Whereas, for year **2018**, the rental count is **high** in the month of **June**.
- In contrast, the demand of bike rentals is **low** for the month of **January** for both year **2018** and **2019**.

holiday vs. count:



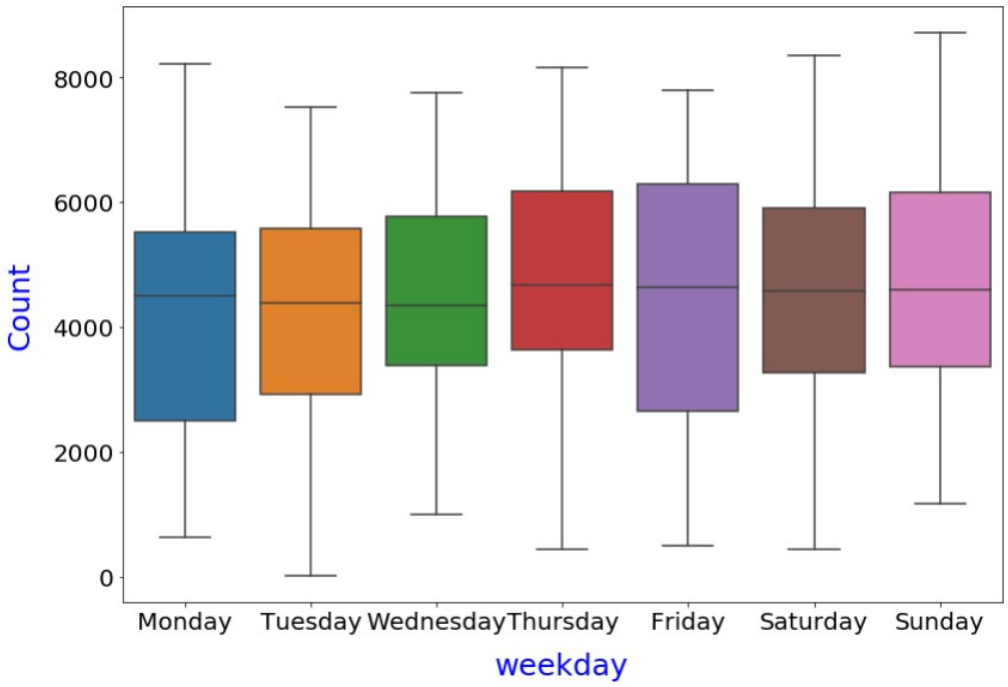
holiday and count, we can see that **count of bike rentals** is **very** when there is **No holiday**, having its 25% value lies between 3000 to 4000 and 95% box value is beyond 8000.

- Whereas, in holidays the 25% of box value lies below 2000.

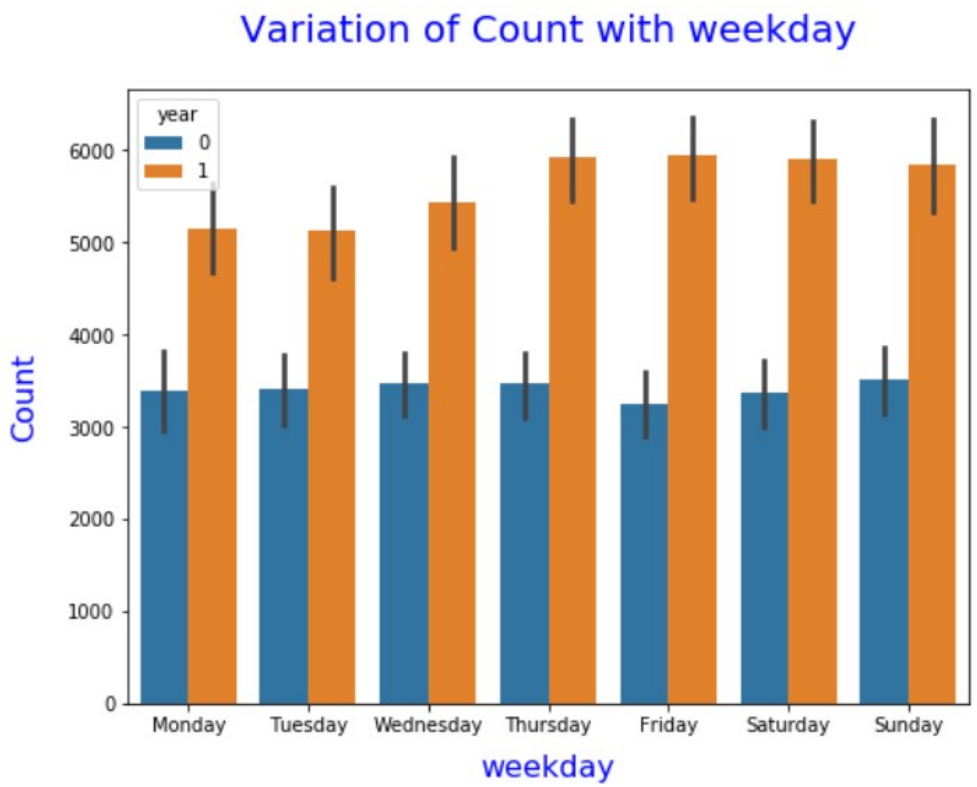


- The above bar plot depicts, that the bike rental count is **high** when there is **no holiday** in both the year **2018** and **2019**.

<u>weekday vs. count:</u>

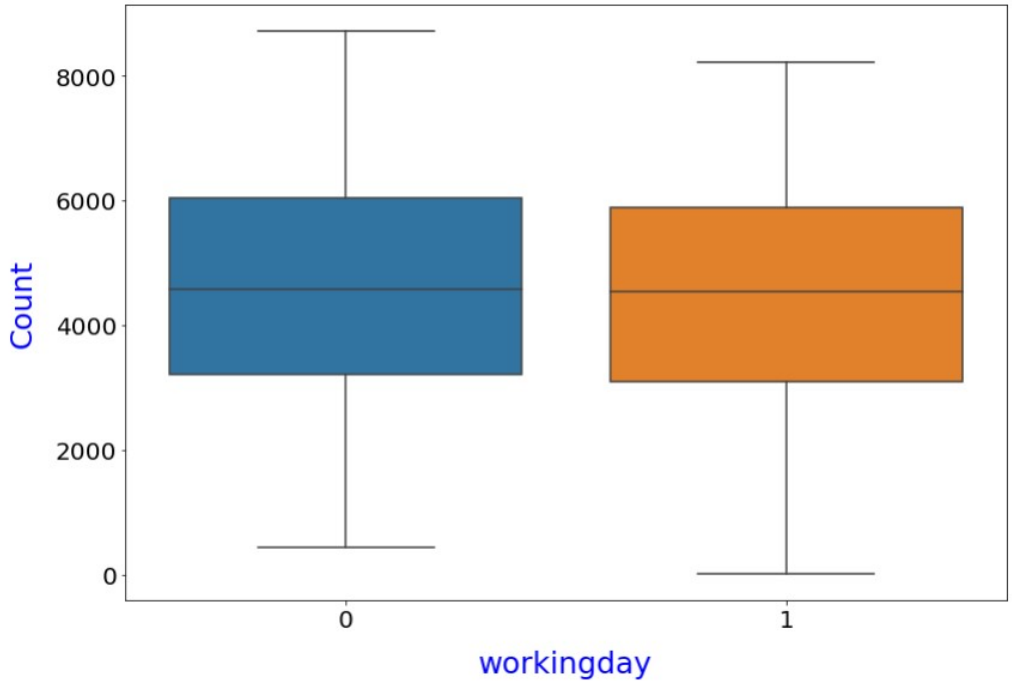


Weekday and count, we can see that **count of bike rentals** is **very** in **Friday**, whereas the rental count is low in **Monday** as its 25% value lies below 3000 (approx.).

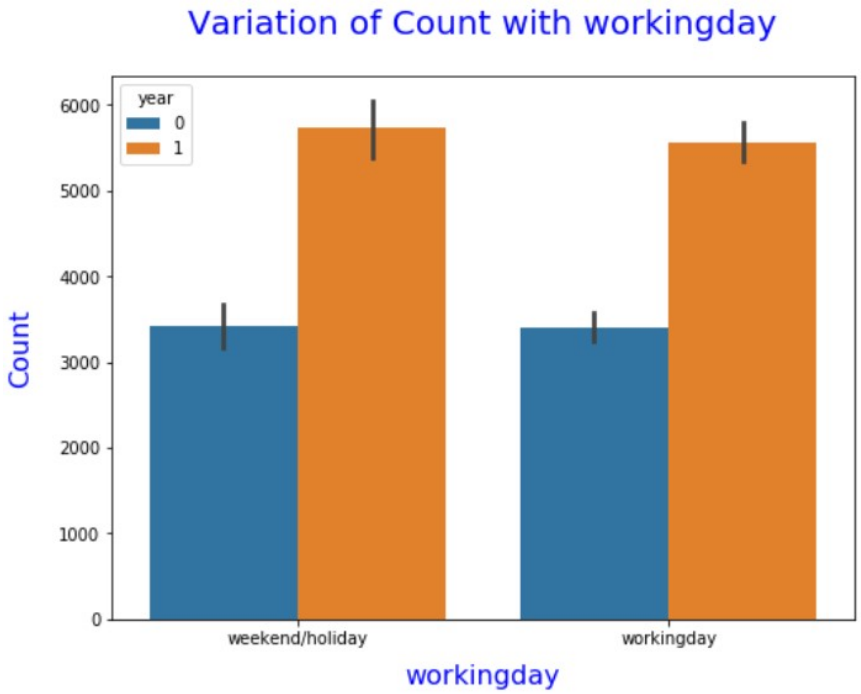


- The above bar plot says that, the demand of bike rental is **high** on **Thursday**, **Friday** and **Saturday** compared to other weekdays in the year of **2019**.
- Whereas, the bike rental demand is **high** on **Monday**, **Tuesday**, **Wednesday** and **Thursday** in the year of **2018**.

working day vs. count:

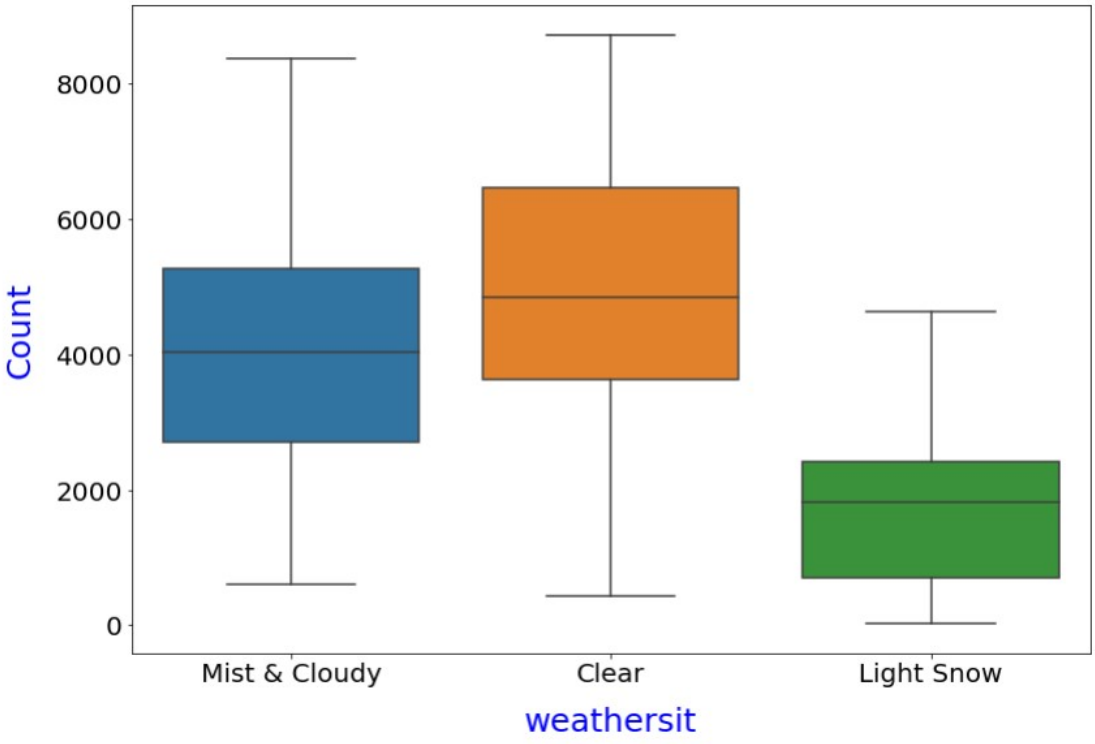


From the above box plot between working day and count, we can see that **count of bike rentals** is **very high** for non-working days, having 75% of box value around 6000 (approx.).



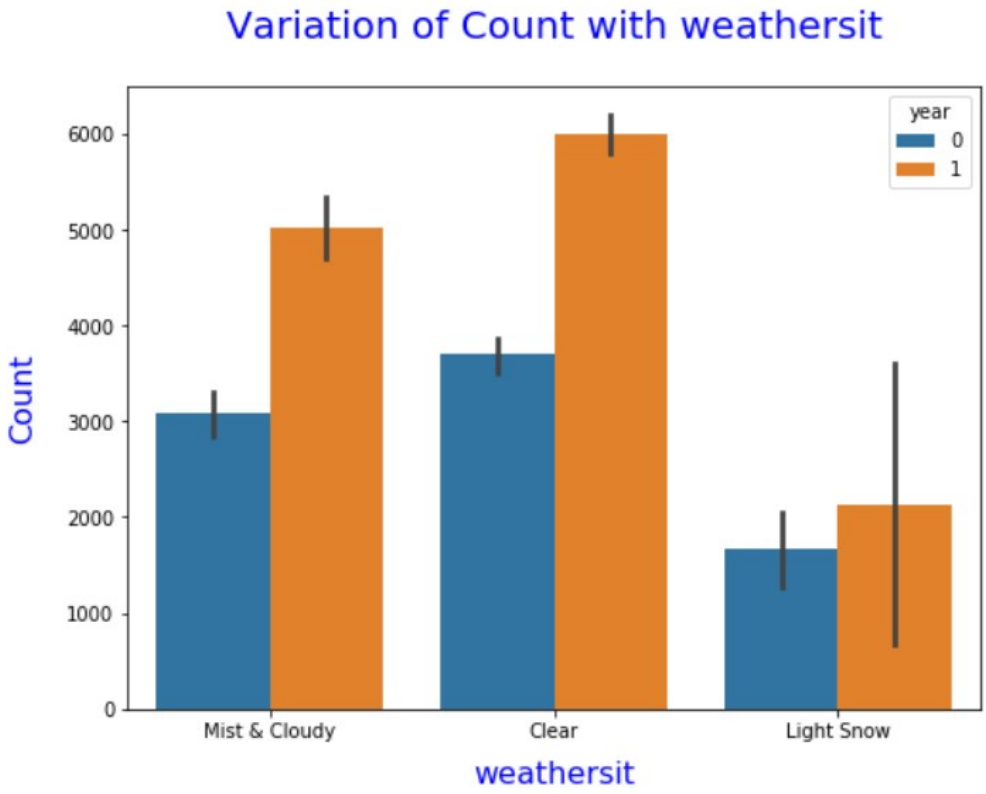
- The above bar plot shows, that the bike rental count is **high** when there is **weekend or holiday** compared to working days in both the year **2018** and **2019**.

<u>weathersit vs. count:</u>



-

  From the above box plot between weathersit and count, we can see that **count of bike rentals** is **very high** in **clear** weather, having its 75% of value lies above 6000.
- Whereas, in **Light snow**, the **bike rental count** is very low, having its 75 % box value slightly more than 2000 (approx.).



- The above bar plot depicts, that the demand of bike rental is **high** in **clear weather** in both the year **2018** and **2019**.
- Whereas, the demand reduces in **Light Snow** weather in both the year.

Linear Regression Model(*Best Fit Regression Line Equation):*

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

Also, from the linear regression model we can see that the dependent variable Bike rental count is dependent on below categorical variables.

- **September:** The model says that, the dependent variable, i.e., the count of bike rental (cnt) is dependent on September month. **In September month the Bike Rental demand increases by 0.0910 units, keeping all other predictor variables are constant**.
  Summer: The model also says that, the dependent variable, i.e., the count of bike rental (cnt) is dependent on summer season. **In summer season the Bike Rental demand increases by 0.0621 units, keeping all other predictor variables are constant**.

- **Winter:** The model says that, the dependent variable, i.e., the count of bike rental (cnt) is dependent on winter season. **In Winter season the Bike Rental demand increases by 0.0945 units, keeping all other predictor variables are constant**.

- **Light Snow:** We can also see some negative dependencies on predictor variables from the above model equation. The dependent variable, i.e., **the count of bike rental (cnt) is decreases in Light Snow weather situation. In Light Snow weather, bike rental demand decreases by 0.2850 units, keeping all other predictor variables are constant**.

- **Mist & Cloudy:** The dependent variable, i.e., the count of bike rental (cnt) is also decreases in Mist and cloudy weather situation. **In Mist and cloudy weather, bike rental demand decreases by 0.0787 units, keeping all other predictor variables are constant**.

- **Spring:** The dependent variable, i.e., the count of bike rental (cnt) is also dependent on spring season. We can observe that **Bike rental decreases in Spring season. In Spring season, bike rental demand decreases by 0.0554 units, keeping all other predictor variables are constant**.

- **Holiday:** The dependent variable, i.e., the count of bike rental (cnt) is also dependent on holiday. We can observe that Bike rental decreases in holidays. **In holidays, bike rental demand decreases by 0.0963 units, keeping all other predictor variables are constant.**

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

Solution:

First let's see what a Dummy variable is, then we will see the importance to **drop_first = True** during dummy variable creation.

Dummy Variable:

A **Dummy** variable or **Indicator** Variable is an artificial variable created to represent an attribute with two or more distinct categories/levels. But we cannot use these categorical non-numeric variables in linear regression model building as we need numerical variable for the same. In such situation, dummy variable creation is required.

Regression analysis treats **all independent (X) variables in the analysis as numerical**. Numerical variables are interval or ratio scale variables whose values are directly comparable, e.g. '10 is twice as much as 5', or '3 minus 1 equals 2'. Often, however, we might want to include an attribute or nominal scale variable such as 'Education Level or 'Job category' in our study.

For example, suppose the value of Education levels are 1, 2, 3 etc., represents Primary, Secondary, Tertiary education respectively. Here there is no meaning to subtract Education level 2 from Education level 1.

The numbers here are used to indicate or identify the levels of 'Education Type' and do not have intrinsic meaning of their own. Dummy variables are created in this situation to 'trick' the regression algorithm into correctly analyzing attribute variables.

Importance to 'drop_first = True' during Dummy variables Creation:

Dummy variables assign the numbers '0' and '1' to indicate membership in any mutually exclusive and exhaustive category. **The number of dummy variables necessary to represent a single attribute variable is equal to the number of levels (categories) in that variable minus one.**

This is because, for a given attribute variable, none of the dummy variables constructed can be redundant.

For example, say we are having a column Education Level with three unique levels as follows:

Education_Level
Primary
Secondary
Tertiary

Let's see the customer education dataset:

	Cust_ID	Education_Level
0	100897	Primary
1	177965	Secondary
2	154328	Tertiary
3	122789	Secondary
4	298445	Secondary

We cannot use Education Level during Linear Regression Model building as these are non-numeric levels.

That's why we will use pandas **get_dummies** method to convert this categorical column to numerical one.

Syntax:

pandas.get_dummies(data, prefix=None, prefix_sep='_', dummy_na=False, columns=None, sparse=False, drop _first=False, dtype=None)

Where, **Data: D**ata of which to get dummy indicators.

drop_first: Whether to get n-1 dummies out of n categorical levels by removing the first level.

Returns: DataFrame - Dummy-coded data.

After converting categorical column 'Education Level' into Indicator variables we get **3 columns**:

Python Code: *dummies_education_df = pd.get_dummies(df['Education_Level'])*

	Primary	Secondary	Tertiary
0	1	0	0
1	0	1	0
2	0	0	1
3	0	1	0
4	0	1	0

Value 1 in column 'Primary' says the customer having Primary education. Similarly, value 1 in Secondary column indicate customer is having secondary education and so on.

As we can clearly see, there is no need to define three different levels. If you drop a level, say, 'Primary', we will still be able to explain the three levels. Like, whenever there is zero in both Secondary and Tertiary columns, that means the customer is having Primary education.

To achieve that, we will use drop_first = True, so that grt_dummies() always creates n-1 dummy indicator columns for a column having n levels, by dropping the first level.

In our example, if we use **drop_first = True** in get_dummies(), the output will be as follows:

Python Code:

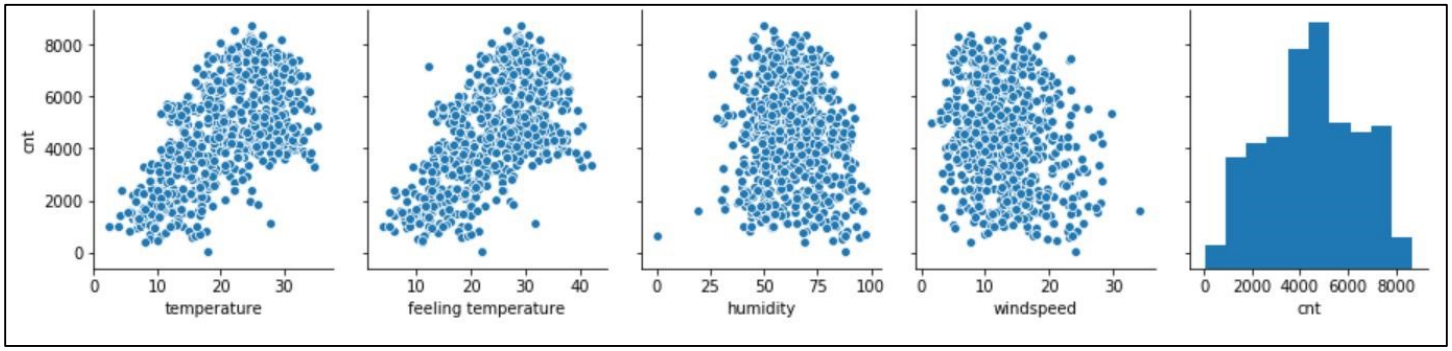
```
dummies_education_df = pd.get_dummies(df['Education_Level'], drop_first = True)
```

	Secondary	Tertiary
0	0	0
1	1	0
2	0	1
3	1	0
4	1	0

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

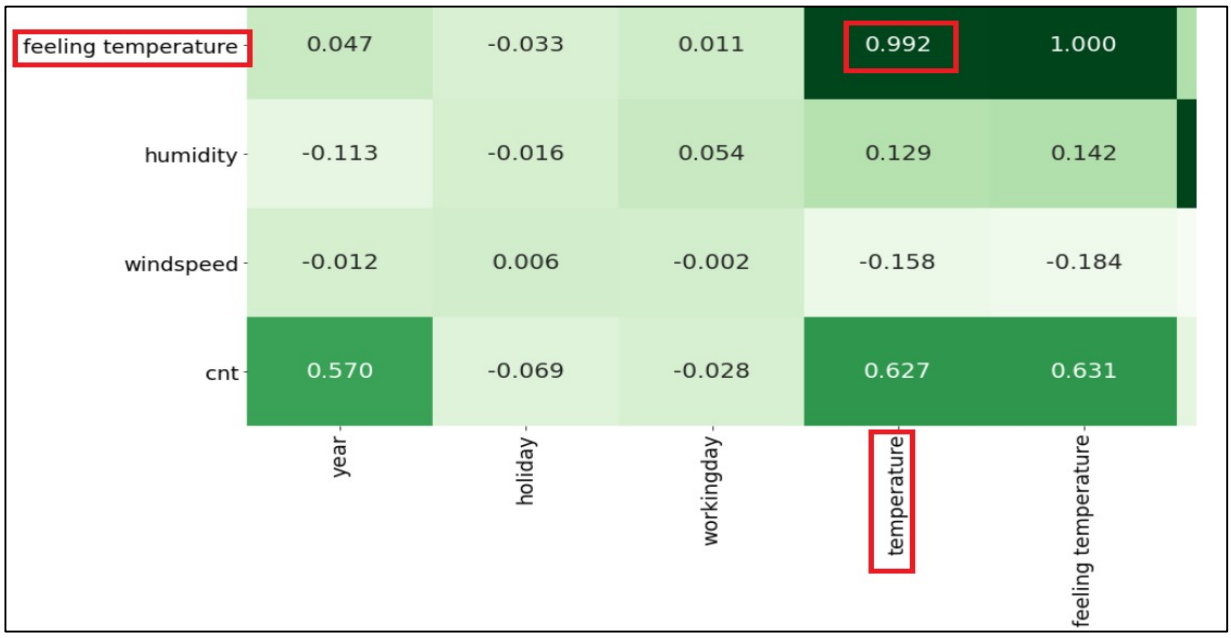
Solution:

Below pair plot shows the correlation among the numerical variables.



From the scatter plot pattern, we can say that **temperature** and **feeling temperature**, both the numerical variables **are strongly correlated** with target variable '**cnt'**.

But the **correlation Heatmap shows that both independent variables temperature and feeling temperature are highly positively correlated between each-other, having correlation coefficient as 0.992.**



Hence, we have **dropped** '**feeling temperature**' from our dataset. Thus, there is only one **numerical variable which is highly correlated with target variable, i.e., 'temperature'.**

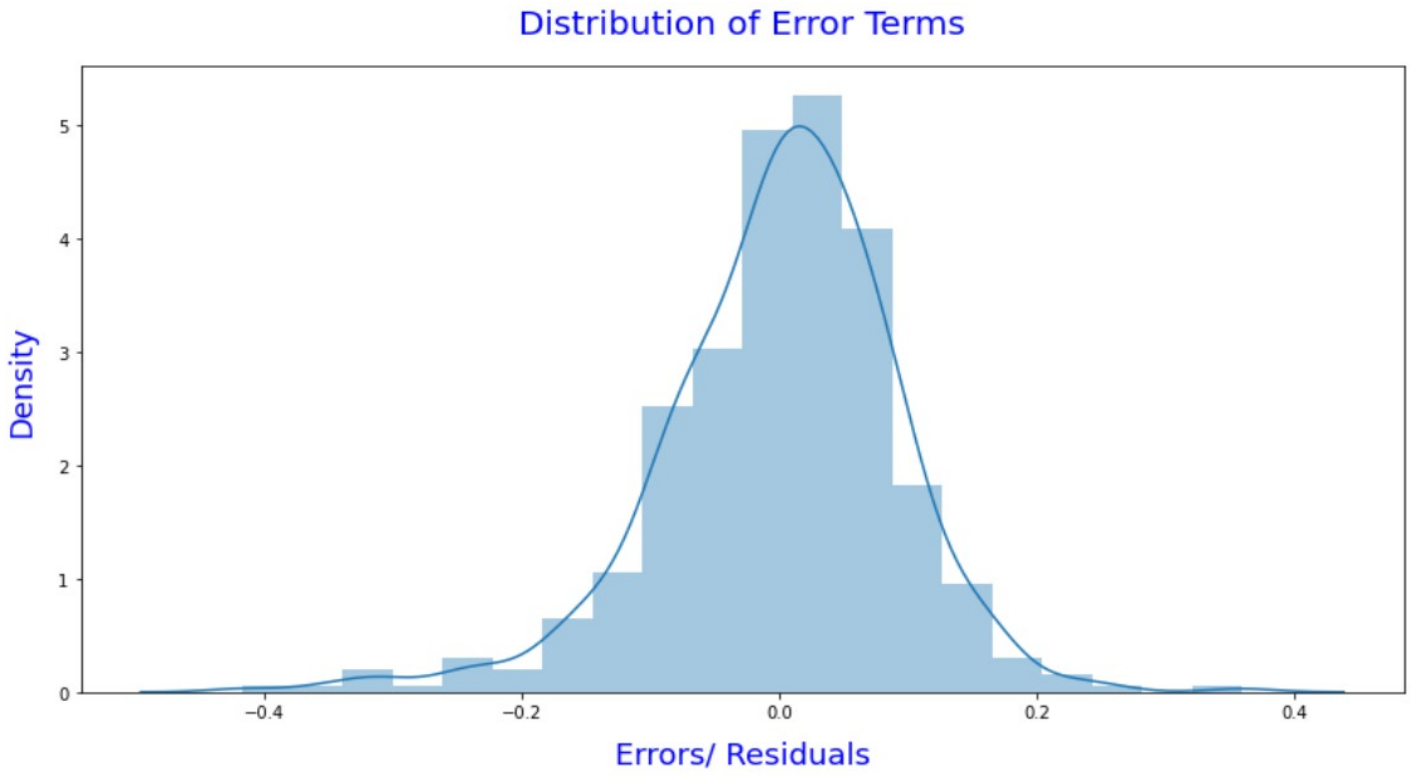
4.How did you validate the assumptions of Linear Regression after building the model on the training set?

Solution:

After building the model, we have validated following assumptions of Linear Regression.

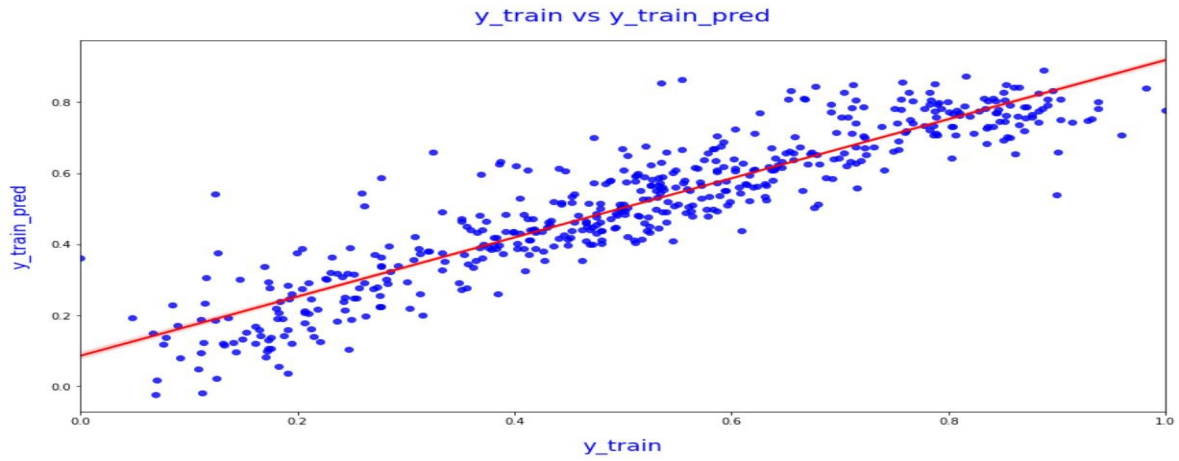
- **Normality**: Error terms are distributed normally with zero mean
- **Homoscedasticity**: The variance of residual is the same for any value of X.
- **Independence**: Observations are independent of each other.

Normality:



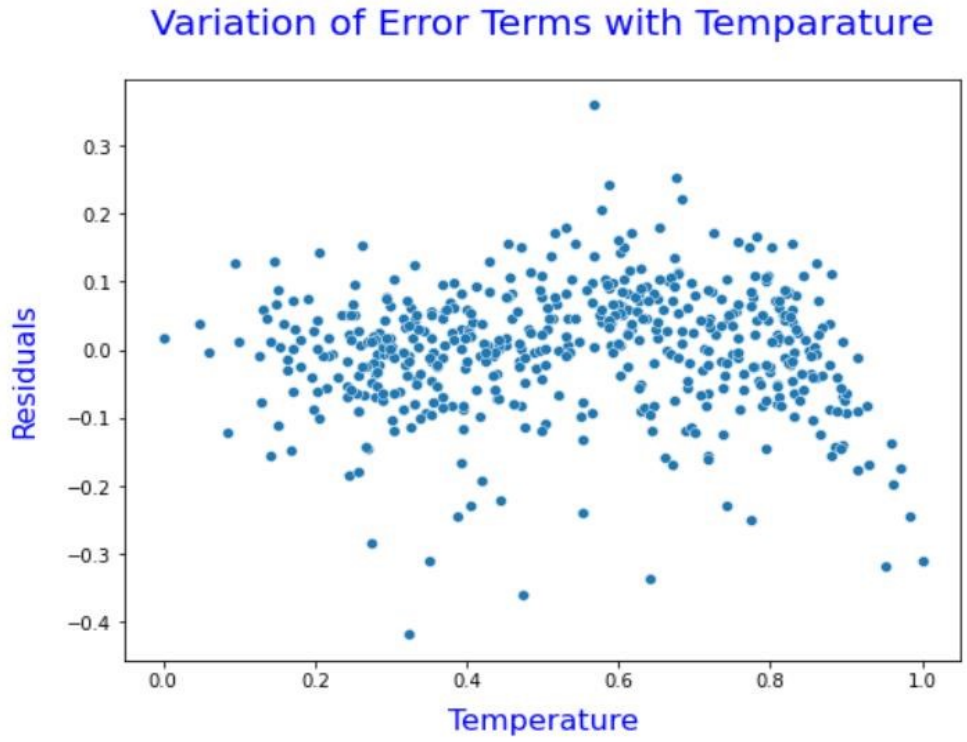
- The above distribution plot shows that the **Residuals/ Error terms** are **normally distributed with zero mean**.
- Hence, our assumption of **Normality becomes true** for this model.

Homoscedasticity:



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

Independence:



- The above scatter plot between the **predictor variable: temperature** and **Residuals** shows that **the Residuals are independent of each-other**. There is no pattern exists among residuals/ error terms.
- Hence, the **assumption of independence becomes true** for our model.

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

Solution

The **correlation coefficients of predictor variables** will help us to identify top 3 features, which are contributing significantly towards explaining the demand of the shared bikes.

Regression Line equation for final model is:

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

Below are the coefficient values (**absolute**) in descending order.

-	temperature	0.477737
-	Light Snow	0.285031
-	year	0.234132
-	windspeed	0.148098
-	holiday	0.096316
-	winter	0.094476
-	Sept	0.090998
-	Mist & Cloudy	0.078741
-	summer	0.062076
-	spring	0.055406

Hence, the **top 3** contributing factors are:

1. **Temperature:**
   We can see from the above model equation, the dependent variable, i.e., the count of bike rental (cnt) is increases with the increase of temperature. **One unit increase of temperature will increase the bike rental demand by 0.4777 units, keeping all other predictor variables are constant**.
   Hence, US bike-sharing provider Boom Bikes, can focus more on temperature, as increase in temperature will increase the demand of bikes.

2. **Light Snow:**
   From the model equation, we can the dependent variable, i.e., **the count of bike rental (cnt) is decreases in Light Snow weather situation**. In Light Snow weather, bike rental demand **decreases by 0.2850 units, keeping all other predictor variables are constant**. As there is less demand in Light snow weather condition, Business can give some **offers/Discounts** or **arrange bike shield** to increase the demand during this weather situation.

3. **Year:**
   We can see **demand for bikes was more in 2019 than 2018**, As there is an increase in demand in 2019 by **0.2341 units, keeping all other predictor variables are constant.** That signifies **the bike rental demand will be increasing in upcoming years as more people come to know about Boom bikesharing provider via different marketing channels**. But business might be facing dips in their revenues due to the ongoing Corona pandemic. By the time Corona Virus reduces the things will be better.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Linear regression is one of the most basic types of regression in machine learning. Before drilling down to **Linear Regression**, first we will discuss about what is **Regression** in Machine Learning.

Regression:
Regression is a classification of Machine Learning Model. It is the most commonly used predictive analysis model. **Regression analysis** consists of a set of *machine learning* methods that allow us to predict a continuous outcome variable (y) based on the value of **one** or **multiple** predictor variables (x).

Briefly, the goal of regression model is to build a mathematical equation that defines y as a function of the x variables. Later, this equation can be used to predict the outcome (y) on the basis of new values of the predictor variables (x). Regression and classification fall under **supervised learning** methods – in which we have the previous years' data with labels which are used to build the model.

Using Regression, the output variable to be predicted is a **continuous variable**, e.g. scores of a student.

Types of Regression:

- Linear regression
- Logistic regression
- Polynomial regression
- Stepwise regression
- Stepwise regression
- Ridge regression
- Lasso regression
- ElasticNet regression

Now, let's move on to Linear Regression.

Linear Regression:
Linear regression is the most simple and popular technique for predicting a continuous variable. 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**.

Types of linear regression:
- Simple linear regression
- Multiple linear regression

Simple Linear Regression:

It is a statistical method that allows us to summarize and study **relationships between two continuous** (quantitative) variables. One variable denoted x is regarded as an independent variable and other one denoted y is regarded as a dependent variable. It is assumed that the two variables are **linearly** related. Hence, we try to find a linear function that predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x).

As there is **only one predictor** variable involved, it is called "**Simple Linear Regression**".

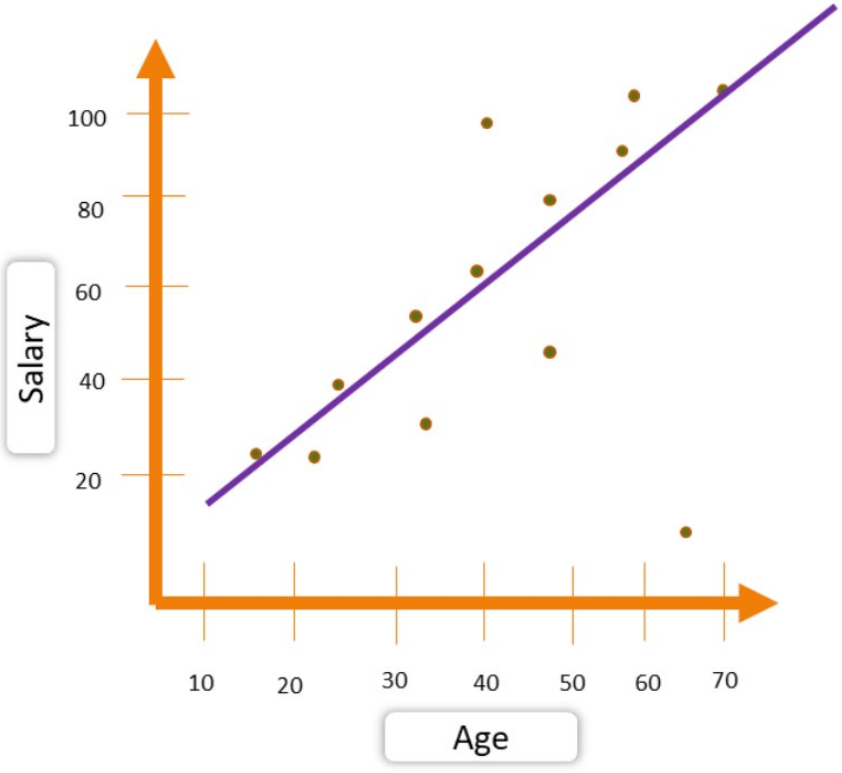
Example, consider predicting the salary of an employee based on his/her age. We can easily identify that there seems to be a correlation between employee's age and salary (more the age more is the salary). Here, employee's age is the independent variable(x) and employee's salary is the dependent/ target variable(y).
The relation ship between two variables x and y are explained using a straight line, called **Regression Line**.

Equation of regression line for simple linear Regression:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where, y = Dependent variable x = explanatory / Predictor variables β_0 = y-intercept (constant term) β_1 = Slope coefficients for each explanatory variable ϵ = The model's error term (also known as the residuals)

In our above example, dependent variable is salary, i.e., Y-axis of below plot denotes Salary and independent / predictor variable is age, i.e., x- axis denotes age.



Multiple Linear Regression:

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the **linear relationship** between the explanatory (independent) variables and response (dependent) variable.

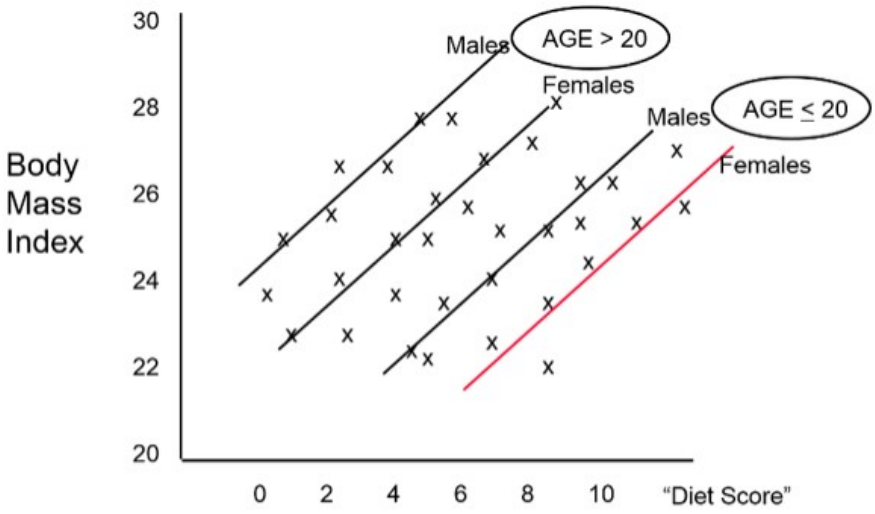
Multiple regression is the extension of simple linear regression that **involves more than one explanatory variable**.

Equation of regression line for Multiple linear Regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_p x_p + \epsilon$

where, for i = p observations:
y = Dependent variable $\mathbf{x_i}$ = Explanatory / Predictor variables $\boldsymbol{\beta_0}$ = y-intercept (constant term) $\boldsymbol{\beta_p}$ = Slope coefficients for each explanatory variable x_i $\boldsymbol{\epsilon}$ = The model's error term (also known as the residuals)

Let's understand this with an example, suppose we want to understand BMI based on diet score, gender and age group, in that case BMI will be our Dependent variable (y) and diet score, gender and age group will be three independent predictor variables, i.e., $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$.

The plot of the multiple linear regression will be look like below.



BMI = 18.0 + 1.5 (diet score) + 1.6 (if male) + 4.2 (if adult)

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3$$

Here, $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, $\boldsymbol{\beta_3}$ are the coefficients, which signifies BMI (y) will be increase by $\boldsymbol{\beta_1}$ terms with one-unit increase of $\mathbf{X_1}$, considering $\mathbf{X_2}$ and $\mathbf{X_3}$ are constant.

In above simple and multiple linear regression plots we saw regression lines, which explains the linearity between dependent and independent variables. Now we will see how to get the best fit regression line.

Best Fit Line:

Best Fit Line is one of the most important outputs of regression analysis. A **line of best fit** is a straight line that is the best approximation of the given set of data. It is used to study the nature of the relation between two variables.

A line of best fit can be roughly determined by drawing a straight line on a scatter plot, so that the number of points above the line and below the line is about equal (and the line passes through as many points as possible).

A more accurate way of finding the line of best fit is the **Ordinary Least Square Method (OLS)**. It is also known as **Residual Sum of Squares (RSS).**

Ordinary Least Square Method:

Using OLS, the best-fit line is found by **minimizing the expression of RSS** (Residual Sum of Squares), also known as **Cost** function of Linear regression model, which is equal to the sum of squares of the residual for each data point in the plot. Residuals/ error terms ($\mathbf{e_i}$) for any data point is found by subtracting predicted value of dependent variable ($\mathbf{Y_{Pred}}$) from actual value of dependent variable ($\mathbf{Y_i}$).

We need to find out $\mathbf{\beta_0}$, $\mathbf{\beta_1}$ coefficients in such a way, that reduces RSS and we can get the best fit predicted line.

Residual sum of Squares:	Residual, $e_i = Y_i - Y_{Pred}$
	$RSS = \sum_{i=1}^{n} (Residual)^2$
	$RSS = \sum_{i=1}^{n} (e_i)^2$
	$RSS = \sum_{i=1}^{n} (Y_i - Y_{Pred})^2$

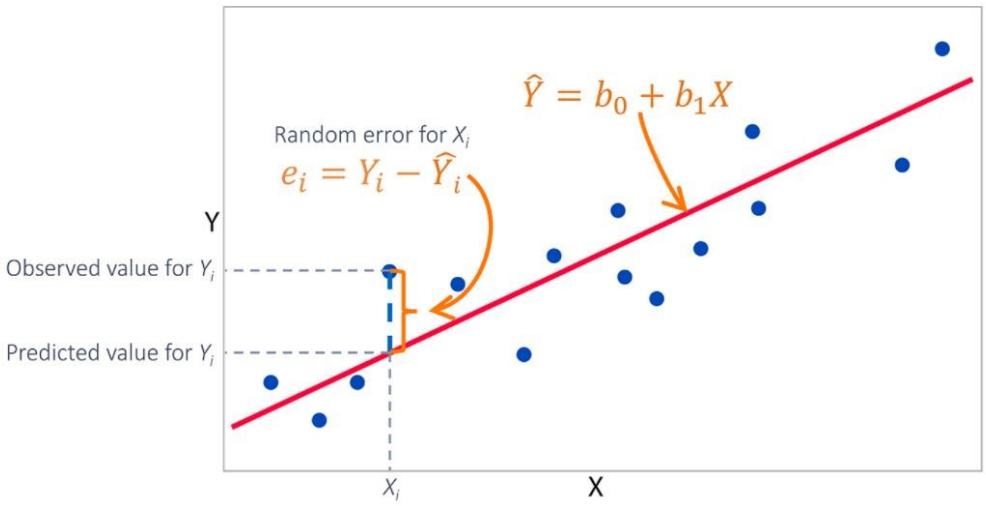
where:

Y_i is the original target score

Y_{Pred} is the predicted target score

e_i is the residual

Below plot depicts the Residuals graphically.



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:

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

R2 or Coefficient of Determination:

R2 is used to check the accuracy of your model, which is R2 statistics. R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.

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

Where,
RSS = Residuals 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:

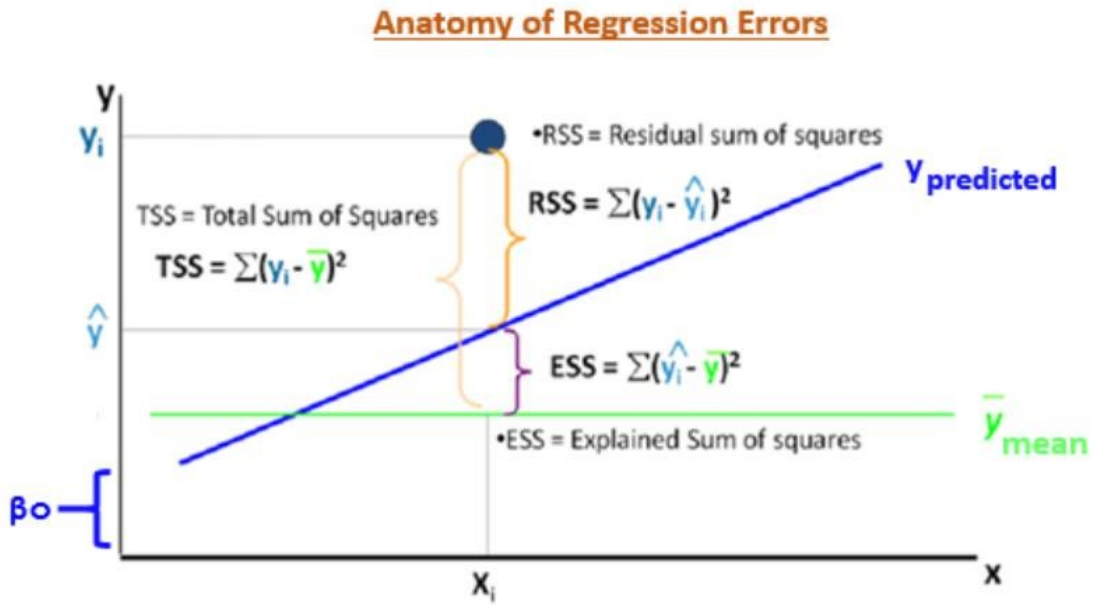
$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

In OLS, the error estimates can be divided into three parts. Following figure represented the error estimates.

Residual Sum of Squares (RSS) – ∑[Actual(y) – Predicted(y)]²

Explained Sum of Squares (ESS) – ∑[Predicted(y) – Mean(ymean)]² **Total**

Sum of Squares (TSS) – ∑[Actual(y) – Mean(ymean)]²



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

The Four Assumptions of Linear Regression:

Before we conduct linear regression, we must first make sure that four assumptions are met:

1. **Linear relationship:** There exists a linear relationship between the independent variable, x, and the dependent variable, y.
2. **Independence:** The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
3. **Homoscedasticity:** The residuals have constant variance at every level of x.
4. **Normality:** The residuals of the model are normally distributed.

For Multiple Linear Regression, there are another assumption, i.e., all the predictor/ independent variables should be linearly independent of each-other. There is no "**Multicollinearity**" exists among independent variable.

2. Explain the Anscombe's quartet in detail.

Solution:

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.

Those 4 sets of 11 data-points are given below.

The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize **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
10.0	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

Summary

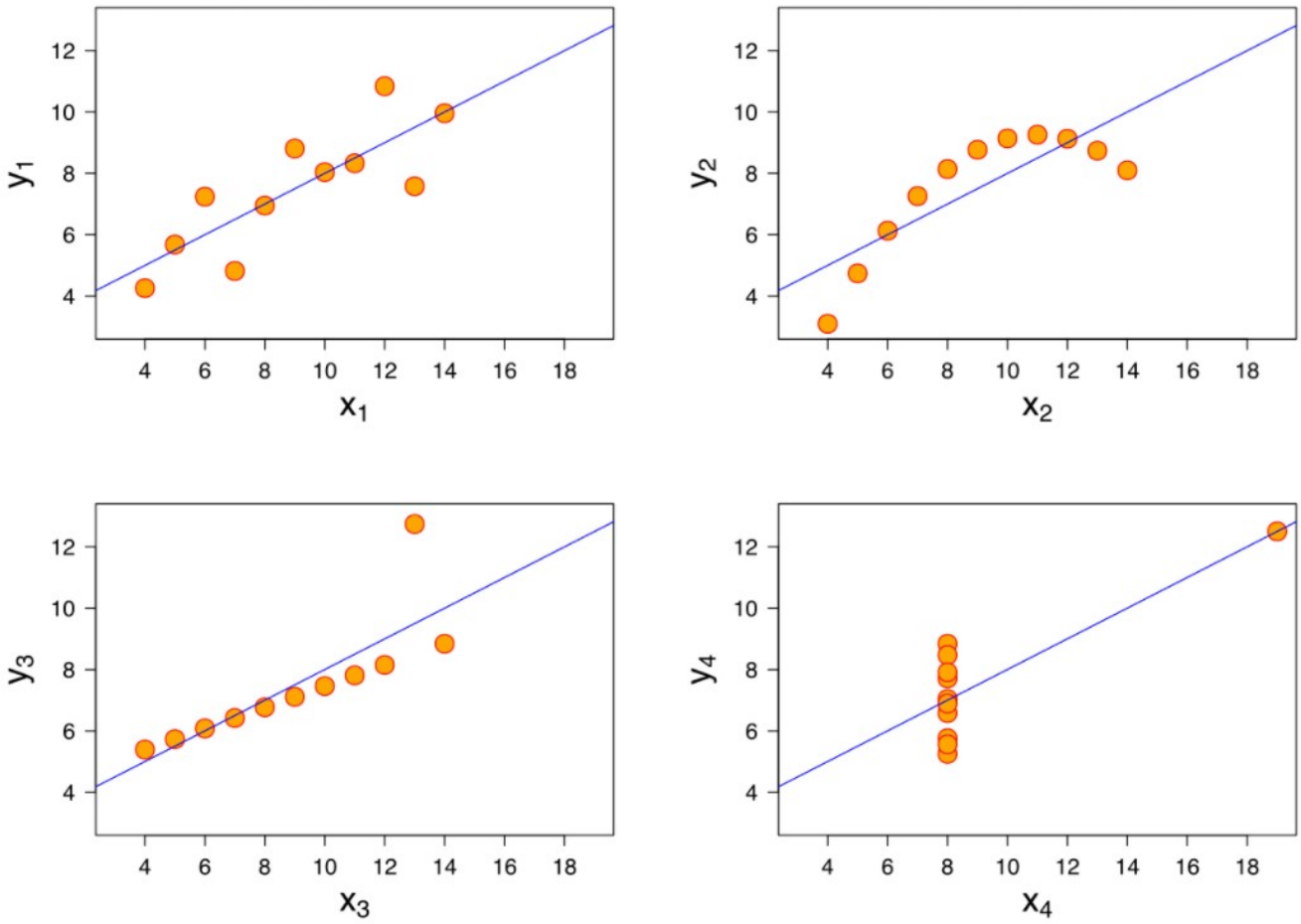
Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

Summary Statistics:

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is
  0.816 for each dataset

All four data sets are identical when examined using **simple summary statistics**, but vary considerably when graphed.



Explanation of above plots:

- **Plot I** (top left) - The scatter plot shows a **linear** relationship between x and y.
- **Plot II** (top right) - The figure concludes that there is a non-linear relationship between x and y.
- **Plot III** (bottom left) – Figure shows a perfect linear relationship for all the data points except one which seems to be an **outlier** which is indicated be far away from that line.
- **Plot IV** (bottom right) – It shows an example when one high-leverage point is enough to produce a high correlation coefficient.

3. What is Pearson's R?

Solution:

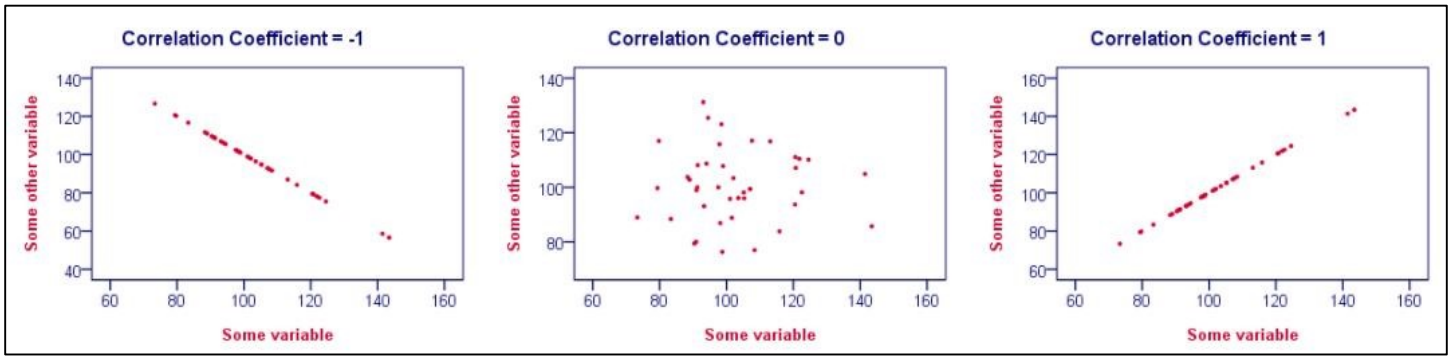
In statistics, the **Pearson correlation coefficient** (PCC), also referred to as **Pearson's R**, the Pearson productmoment correlation coefficient (PPMCC), or the bivariate correlation, is a statistic that measures **linear correlation** between two variables X and Y.
Now, the question comes to our mind is what is **Correlation / Correlation coefficient**?

Correlation: In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data.

Correlation Coefficient: Correlation coefficients are used in statistics to measure how strong a relationship/ correlation 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*) 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.



Interpretation of above graph:

- **Correlations are never lower than -1.** A correlation of -1 indicates that the data points in a scatter plot lie exactly on a straight descending line; the two variables are perfectly negatively linearly related.

Example, Petrol in fuel tank decreases in (almost) perfect correlation with distance travel in kilometers.

- A **correlation of 0** means that two variables don't have any linear relation whatsoever. However, some nonlinear relation may exist between the two variables.

- **Correlation coefficients are never higher than 1.** A correlation coefficient of 1 means that two variables are perfectly positively linearly related; the dots in a scatter plot lie exactly on a straight ascending line.
  Example, human height goes up in (almost) perfect correlation with age.

Pearson Correlation:

The **Pearson product-moment correlation coefficient** (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit.

The Pearson correlation coefficient, *r*, **can take a range of values from +1 to -1**.

Pearson's correlation coefficient formula -

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

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

Σx^2 = the sum of squared x scores

Σy^2 = the sum of squared y scores

For example: Shoe size of a person increases with his/her foot size. Of course, his/her body growth depends upon various factors like genes, 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?

Solution:

Scaling:

It is also known as feature scaling, it is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

It is a step of Data Pre-Processing which is applied to independent variables or features of data. It basically helps to normalize the data within a range.

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 30 inch to be greater than 1 meter 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. Normalization should be performed when the scale of a feature is irrelevant or misleading and not should Normalize when the scale is meaningful.

Formally, if a feature in the dataset is big in scale compared to others then this big scaled feature becomes dominating and needs to be scaled. Here feature scaling helps to weigh all the features equally.

Advantage of Scaling:

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

What is Normalization?

Normalization: Also known as **Min-Max scaling** this technique re-scales a feature or observation value with distribution value between 0 and 1.

Normalization equation - $$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Here, **max(X) and min(X) are the maximum and the minimum values** of the feature respectively.

Interpretation of the above equation:

- When the value of X is the minimum value in the column, the numerator will be 0, and hence $\mathbf{X_{new}}$ 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 $\mathbf{X_{new}}$ is 1
- If the value of X is between the minimum and the maximum value, then the value of $\mathbf{X_{new}}$ is between 0 and 1.

What is Standardization?

Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

Standardization equation - $$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

$\mathbf{X_{mean}}$ is the mean of the feature values.

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

Solution:

Before discussing in what scenario, the value of VIF is infinite, lets first describe what is Variance Inflation Factor (VIF) is.

Definition of VIF:

- The **variance inflation factor** *(VIF)* quantifies the extent of correlation/ relationship between one predictor and the other predictors in a model. It is used for diagnosing **collinearity**/**multicollinearity**.

- The VIF estimates how much the variance of a regression coefficient is **inflated** due to **multicollinearity** in the model.

- The extent to which a predictor is correlated with the other predictor variables in a linear regression can be quantified as the ***R-squared*** **statistic** of the regression where the predictor of interest is predicted by all the other predictor variables (X_i).

- The variance inflation factor for the estimated regression coefficient β_j — denoted **VIFj** — is just the factor by which the variance of β_j is "**inflated**" by the existence of correlation among the predictor variables in the model.

In particular, the variance inflation factor for the j^{th} predictor is:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Where, R_j^2 is the R^2 - value obtained by regressing the j^{th} predictor on the remaining predictors.

How do we interpret the variance inflation factors for a regression model?

- A VIF of 1 means that there is no correlation among the j^{th} predictor and the remaining predictor variables, and hence the variance of b_j is not inflated at all.
- VIFs exceeding 5 warrant further investigation.
- VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

Why the value of VIF becomes infinite?

An **infinite VIF** value indicates that the corresponding predictor variable may be expressed exactly by a linear combination of other predictor variables. A large value of VIF indicates that there is a correlation between the variables.
On the other hands, the formula of the VIF stats that, if the value of R^2 increases ($R^2 \rightarrow 1$), which means one independent variable X*j* can be explained by all other independent or predictor variables, the denominator (i.e., 1 - R^2_j) will decrease, which makes overall VIF infinite.

$$VIF_j = \frac{1}{1 - R_j^2}$$

- If $R^2 \rightarrow 1$, then (1 - R^2_j) becomes **zero**, and VIF_j will become **infinity**.
- Lets, proof the above explanation with an example dataset.
- Taking a dataset of patients' Blood Pressure dependency factor from a hospital.

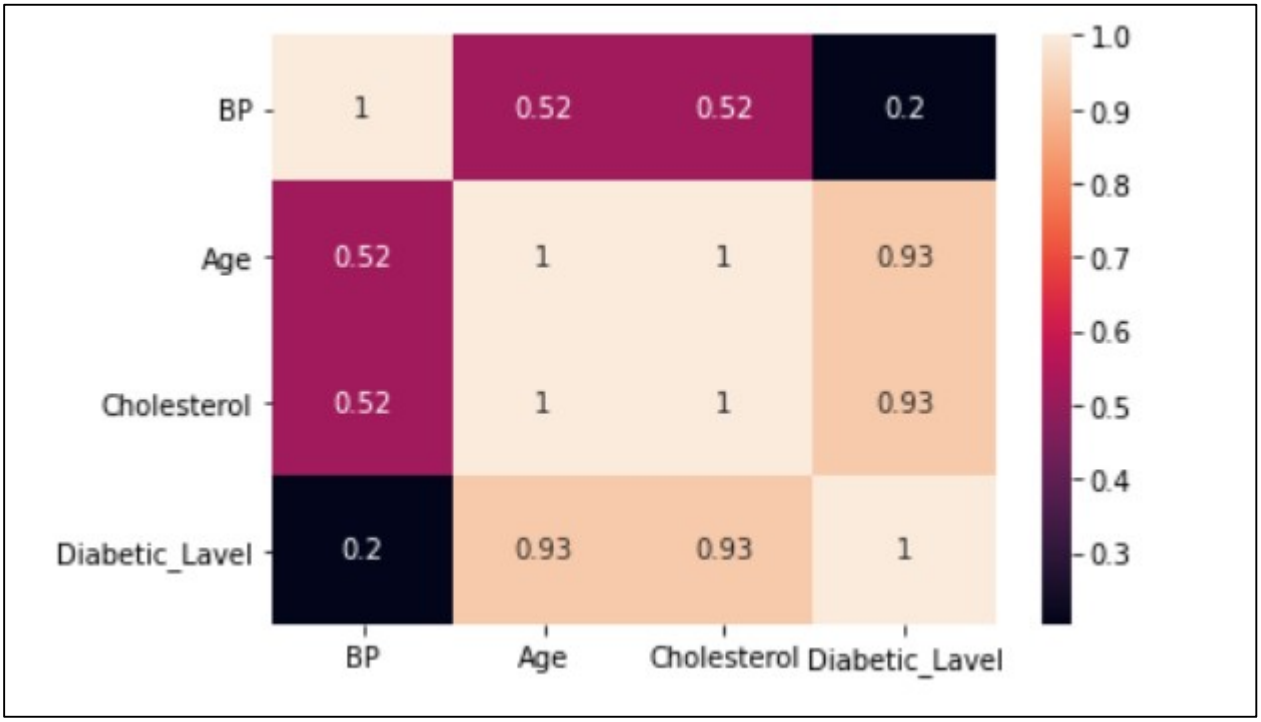
It looks like below:

	BP	Age	Cholesterol	Diabetic_Lavel
0	105	10	100	80
1	115	20	200	90
2	116	30	300	110
3	117	40	400	200
4	112	50	500	300

Here, the column '**BP**' represents blood pressure level and its dependent on patients' Age, Cholesterol, Diabetic level.
So, in this dataset, the Predictor variables are as follows;

- Age
- Cholesterol
- Diabetic level

Let's check the correlation between predictor variables using a seaborn heatmap plot:



From the above plot, we can see that the **Correlation Coefficient** between variable **Age** and variable **Cholesterol** is **perfectly 1**, which means predictor variable Age is strongly positively correlated with predictor variable **Cholesterol**.

Verifying the collinearity using VIF:

After calculating VIF using the **variance_inflation_factor** method of **statsmodels.stats.outlier_influence** library, we got below result.

	Features	VIF
1	Age	inf
2	Cholesterol	inf
3	Diabetic_Lavel	34.33
0	BP	6.21

We can see, the calculated **VIF of feature Age is showing 'infinite'**, which means variable 'Age' can be 100% explained by a linear relationship of other predictor variables. From the heatmap we saw that **correlation coefficient of variable Age and Cholesterol in perfectly 1**, that means **Age is linearly dependent of Cholesterol level and therefore can be explained by Cholesterol.**

Calculating R-squared(j) of model, taking **Xj = Age** as Target and all other predictor variables as independent dataset.

Below summary shows that R^2_j becomes 1. Now, if we replace this R^2_j in our VIF_j formula, we will see that denominator, i.e., $(1 - R^2_j)$ becomes zero, which means **VIF_j = Infinite**

OLS Regression Results

Dep. Variable:	Age	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	2.805e+30
Date:	Sat, 24 Oct 2020	Prob (F-statistic):	3.91e-46
Time:	20:36:43	Log-Likelihood:	149.66
No. Observations:	5	AIC:	-295.3
Df Residuals:	3	BIC:	-296.1
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Cholesterol	0.1000	2.43e-16	4.12e+14	0.000	0.100	0.100
Diabetic_Lavel	0	4.55e-16	0	1.000	-1.45e-15	1.45e-15

Omnibus:	nan	Durbin-Watson:	0.077
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.406
Skew:	-0.202	Prob(JB):	0.816
Kurtosis:	1.664	Cond. No.	13.8

Warnings:

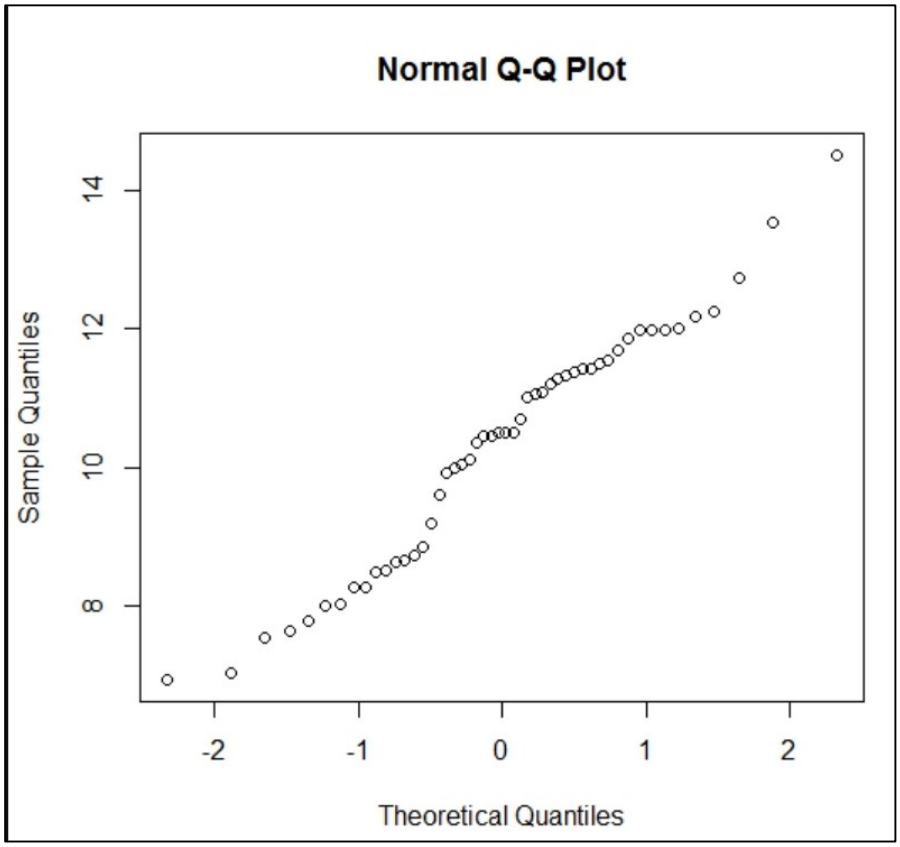
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

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

Solution:

A **Q-Q plot is a scatterplot** created by plotting the **theoretical quantiles** or basically known as the standard normal variate (a normal distribution with mean=0 and standard deviation=1) on the **x-axis** and the **ordered values for the random variable** which we want to find whether it is Gaussian distributed or not, on **the yaxis**. If both sets of quantiles came from the same distribution, we should see the points forming a **line that's roughly straight**.

Here's an example of a Normal Q-Q plot when both sets of quantiles **truly come from Normal distributions**.



Plotting of a Q-Q plot:

- **Order the random variables from smallest to largest**.
- **Draw a normal distribution curve.** Divide the curve into n+1 segment.

- **Find the z-value (cut-off point) for each segment**. These segments are areas. The sum of area of all segment should be 1. So, we need to refer to Z-table to find the corresponding Z-score for each segment partition.
- Plot data set values (step 1) against normal distribution cut-off points (step 3).

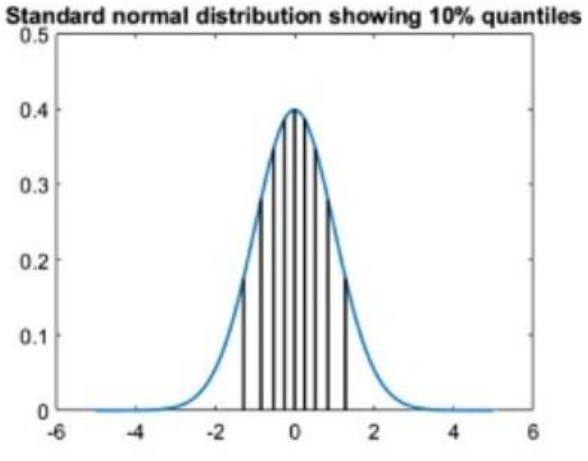
Example of plotting procedure Q-Q plot:

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

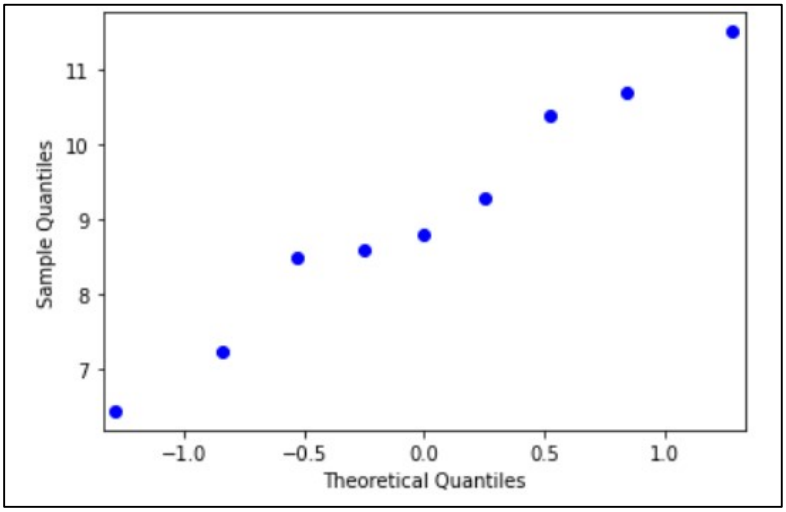
Residuals: 11.5, 10.69, 10.37, 9.29, 8.79, 8.6, 8.5, 7.25, 6.45

Step 1: Now we are sorting the **residual in ascending order**.
Sorted Residuals: 6.45, 7.25, 8.50, 8.60, 8.79, 9.29, 10.37, 10.69, 11.50

Step 2: Draw a normal distribution curve. Divide the curve into **n+1** segment. In our example n = 9. So, there will be total 10 segments in distribution curve and each segment will cover **10%** of area.



Step 3: As these segments are areas, we need to refer the z- table to find z-value (cut-off point) for each segment.
Step 4: Plot data point values (step 1 values) against normal distribution cut-off points (z-values)



Interpretation:

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

Importance of QQ Plot:

We can check normality with distribution plot or histogram also. But the QQ plot is useful in many

distributional aspects like **shifts in location**, **shifts in scale**, **changes in symmetry**, and **the presence of outliers can all be detected** from this plot.

Use of QQ plot in Linear Regression:

If we get almost straight line on this Q-Q plot indicates the data is **approximately normally distributed**. In the linear regression model, **we can use this QQ plot in the finding whether our model error terms (y_train – y_pred) or residual are normally distributed or not.**