**Introduction to Statistics**

To become a successful Data Scientist you must know your basics. Math and Stats are the building blocks of Machine Learning algorithms. It is important to know the techniques behind various Machine Learning algorithms in order to know how and when to use them. Now the question arises, what exactly is Statistics?

"Statistics is a Mathematical Science pertaining to data collection, analysis, interpretation and presentation."

Statistics is used to process complex problems in the real world so that Data Scientists and Analysts can look for meaningful trends and changes in Data. In simple words, Statistics can be used to derive meaningful insights from data by performing mathematical computations on it.

Several Statistical functions, principles and algorithms are implemented to analyse raw data, build a Statistical Model and infer or predict the result.

The field of Statistics has an influence over all domains of life, the Stock market, life sciences, weather, retail, insurance and education are but to name a few.

Moving ahead. Let’s discuss the basic terminologies in Statistics.

**Terminologies in Statistics – Statistics for Data Science**

One should be aware of a few key statistical terminologies while dealing with Statistics for Data Science. I’ve discussed these terminologies below:

* *Statistics Example – Math And Statistics For Data Science*

Now let’s focus our attention on Descriptive Statistics and see how it can be used to solve analytical problems.

**Understanding Descriptive Analysis**

When we try to represent data in the form of graphs, like histograms, line plots, etc. the data is represented based on some kind of central tendency. Central tendency measures like, mean, median, or measures of the spread, etc are used for statistical analysis.

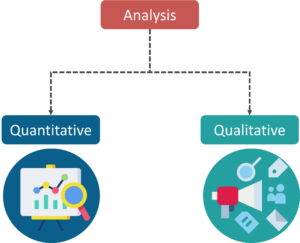
*.*

* *A****Sample****is a subset of the Population*
* *A****Variable****is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item.*
* *Also known as a statistical model, a statistical****Parameter****or population parameter is a quantity that indexes a family of probability distributions.* *For example, the mean, median, etc. of a population.*

Before we move any further and discuss the categories of Statistics, let’s look at the types of analysis.

**Types of Analysis**

An analysis of any event can be done in one of two ways:

****

*Types of Analysis – Math And Statistics For Data Science*

1. **Quantitative Analysis:** Quantitative Analysis or the Statistical Analysis is the science of collecting and interpreting data with numbers and graphs to identify patterns and trends.
2. **Qualitative Analysis:** Qualitative or Non-Statistical Analysis gives generic information and uses text, sound and other forms of media to do so.

For example, if I want a purchase a coffee from Starbucks, it is available in Short, Tall and Grande. This is an example of Qualitative Analysis. But if a store sells 70 regular coffees a week, it is Quantitative Analysis because we have a number representing the coffees sold per week.

Although the purpose of both these analyses is to provide results, Quantitative analysis provides a clearer picture hence making it crucial in analytics.

**Categories in Statistics**

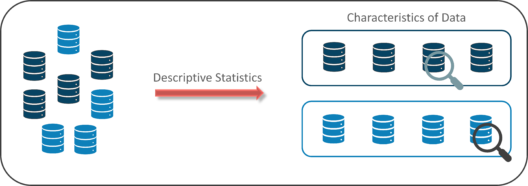
There are two main categories in Statistics, namely:

1. Descriptive Statistics
2. Inferential Statistics

**Descriptive Statistics**

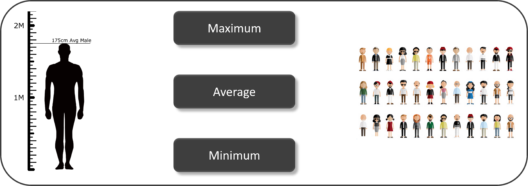
 Descriptive Statistics uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.

Descriptive Statistics helps organize data and focuses on the characteristics of data providing parameters.

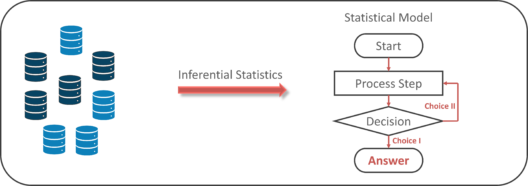


*Descriptive Statistics – Math And Statistics For Data Science*

Suppose you want to study the average height of students in a classroom, in descriptive statistics you would record the heights of all students in the class and then you would find out the maximum, minimum and average height of the class.

****

*Descriptive Statistics Example – Math And Statistics For Data Science*



**Measures Of The Center**

1. **Mean:** Measure of average of all the values in a sample is called Mean.
2. **Median:**Measure of the central value of the sample set is called Median.
3. **Mode:**The value most recurrent in the sample set is known as Mode.

Using descriptive Analysis, you can analyse each of the variables in the sample data set for mean, standard deviation, minimum and maximum.

* If we want to find out the mean or average horsepower of the cars among the population of cars, we will check and calculate the average of all values. In this case, we’ll take the sum of the Horse Power of each car, divided by the total number of cars:

*Mean = (110+110+93+96+90+110+110+110)/8 = 103.625*

* If we want to find out the center value of mpg among the population of cars, we will arrange the mpg values in ascending or descending order and choose the middle value. In this case, we have 8 values which is an even entry. Hence we must take the average of the two middle values.

*The mpg for 8 cars: 21,21, 21.3,22.8,23,23,23,23*  
*Median = (22.8+23 )/2 = 22.9*

* If we want to find out the most common type of cylinder among the population of cars, we will check the value which is repeated most number of times. Here we can see that the cylinders come in two values, 4 and 6. Take a look at the data set, you can see that the most recurring value is 6. Hence 6 is our Mode.

**Measures of the Spread**

Just like the measure of centre, we also have measures of the spread, which comprises of the following measures:

1. **Range:** It is the given measure of how spread apart the values in a data set are.
2. **Inter Quartile Range (IQR):** It is the measure of variability, based on dividing a data set into quartiles.
3. **Variance:** It describes how much a random variable differs from its expected value. It entails computing squares of deviations.
   1. ***Deviation****is the difference between each element from the mean.*
   2. ***Population Variance****is the average of squared deviations*
   3. ***Sample Variance****is the average of squared differences from the mean*
4. **Standard Deviation:** It is the measure of the dispersion of a set of data from its mean.

**CENTRAL LIMIT THEOREM:**

If a random sample of n observations is selected from a population (Any population),

Then when n is sufficiently large, the sampling distribution of x will be approximately normal.

(The larger the sample size, the better will be the normal approximation to the sampling distribution of x.)Cc

# **What It Says**

The central limit theorem (CLT) is simple. It just says that with a large sample size, sample means are normally distributed.

Obviously some of these terms need qualification or explanation. Let’s start at the end and work backwards:

Normal distributed means that a group of numbers follows a bell-shaped curve. Most of the numbers cluster in the middle around the average, and there are fewer numbers at the extremes to the right and left. Looks something like this:

# **A Normal Distribution**

A sample mean is the average of a random subset of a larger group. So if you randomly picked 10 people out of 100 and recorded their heights, the average of those 10 heights would be the sample mean. You could do this many times and, since it is a random selection, the sample mean would be different each time.

What constitutes a large sample is of course subjective, but the frequently cited number here is a sample size should be greater than 30. In practice, your sample might need to be a fair amount bigger, depending on a few factors.

Putting it all together the CLT just says that when you have roughly 30 or more observations in your sample, the average of those numbers is part of a bell-shaped curve. So if you took a bunch of 30+ sized samples and plotted them, they’d look like the normal distribution pictured above — most of the averages would fall toward the center but you’d get a few observations towards the extremes.

The CLT make no assumptions about the distribution of your underlying data. The distribution of people’s heights does not need to be normally distributed in order to know that the sample means of the heights are normally distributed.

# **Why It Matters**

Now you know what the theorem says let’s cover why it matters. Hypothesis testing is the methodology science uses to validate ideas. The framing of a hypothesis test is alway: does the data I have support my idea or could my data just be due to chance?

The way scientists quantify due to chance is by assessing the likelihood of observing their data given that their idea is wrong. So if it’s likely that you’d see the data you collected even if your idea is wrong, then the data doesn’t provide support for the idea. Makes sense, right?

Now comes the tricky part: how do you understand how likely some data is when your hypothesis is wrong? To do this you need to construct the range of values you could see given that your hypothesis is wrong, and then assess the likelihood of your observed value in that context. Luckily this is exactly what the CLT allows you to do.

# **An Example**

Let’s say you’re a data scientist at a software company and you’ve been asked to quantify how engaging your homepage is. You have a week to get an answer to this question. You decide to use average time spent on the homepage as the metric that captures this idea, and think that if your homepage is engaging the true average time spent on it should be over five minutes.

While you could measure the average time spent on your homepage by all users for all time, that wouldn’t exactly be efficient or conducive to answering the question within a week.

So instead you take a sample of 10% of your users over seven days and measure the average time they spend on your homepage. The average time spent on the homepage among this sample is 6.2 minutes. But as with the heights of the random group mentioned above, there will be variation in your estimate, since you’re only measuring a subset of the users you care about.

Since there will be some discrepancy in the average time spent in your sample compared to all your users, you need to assess the likelihood that your result is due to chance. That is, what if the true time spent on your homepage is five minutes or less, but because of random variation the subset of users you measured had an average of 6.2?

Luckily, as long as your sample size is bigger than 30, you can use the central limit theorem to construct what the distribution of time spent on your homepage would look like if your hypothesis is wrong, i.e. when the true average time spent is not greater than five minutes. This is called the distribution under the null hypothesis or the [null distribution](https://en.wikipedia.org/wiki/Null_distribution).

The CLT says the null distribution will be normal (i.e bell-shaped), and it also says you can approximate the values you need to construct the null distribution with values from your sample.

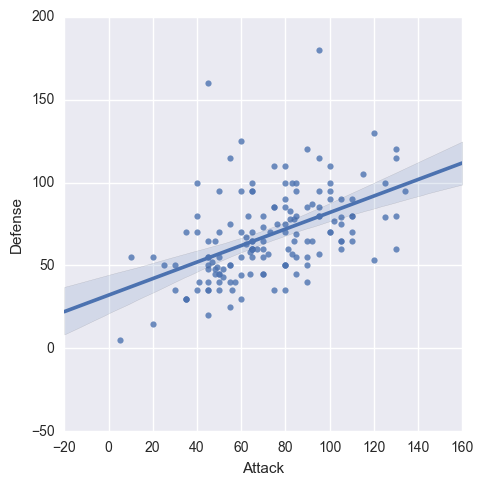
## Data visualization using Seaborn's plotting functions:

One of Seaborn's greatest strengths is its diversity of plotting functions. For instance, making a **scatter plot** is just one line of code using the lmplot() function.

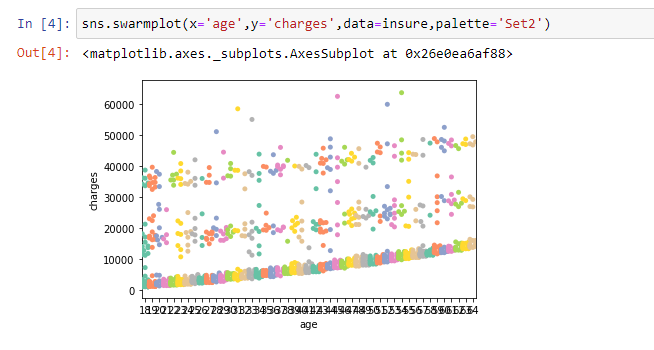
There are two ways you can do so.

* The first way (recommended) is to pass your DataFrame to the data= argument, while passing column names to the axes arguments, x= and y=.
* The second way is to directly pass in Series of data to the axes arguments.

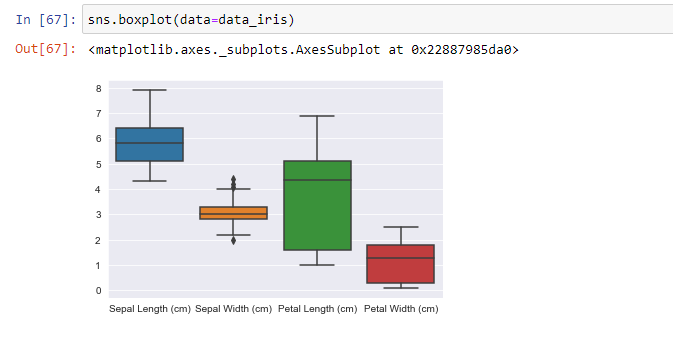
**sns.lmplot(x='Attack', y='Defense', data=df)**



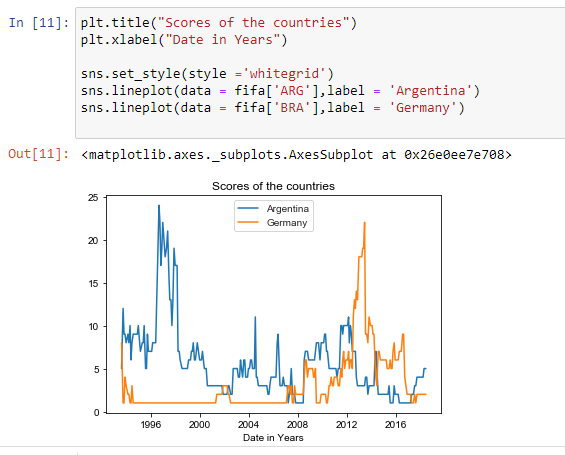
**SWARMPLOT:**

****

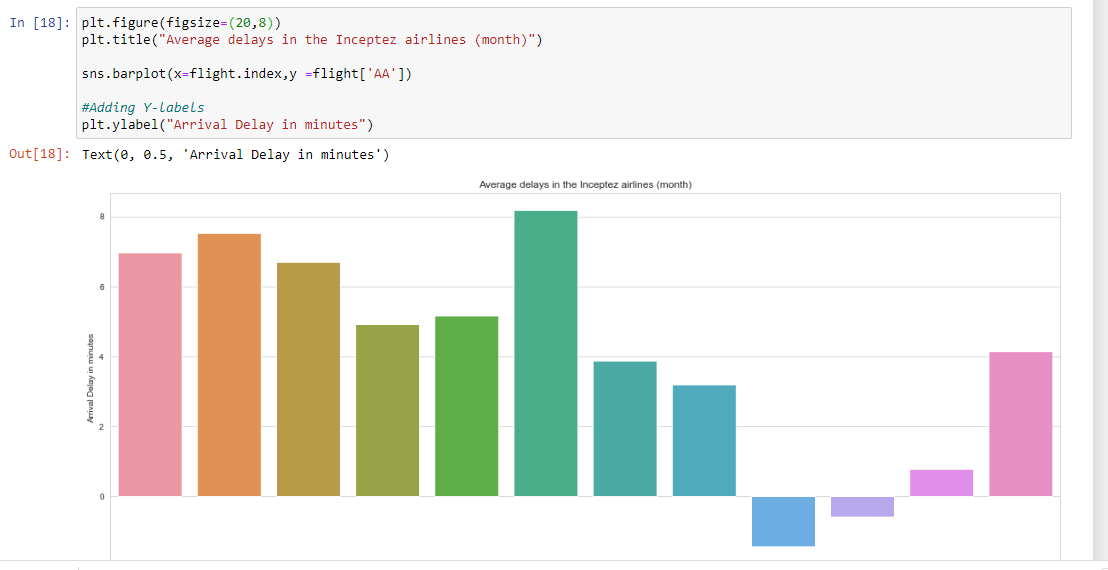
**BOXPLOT:**



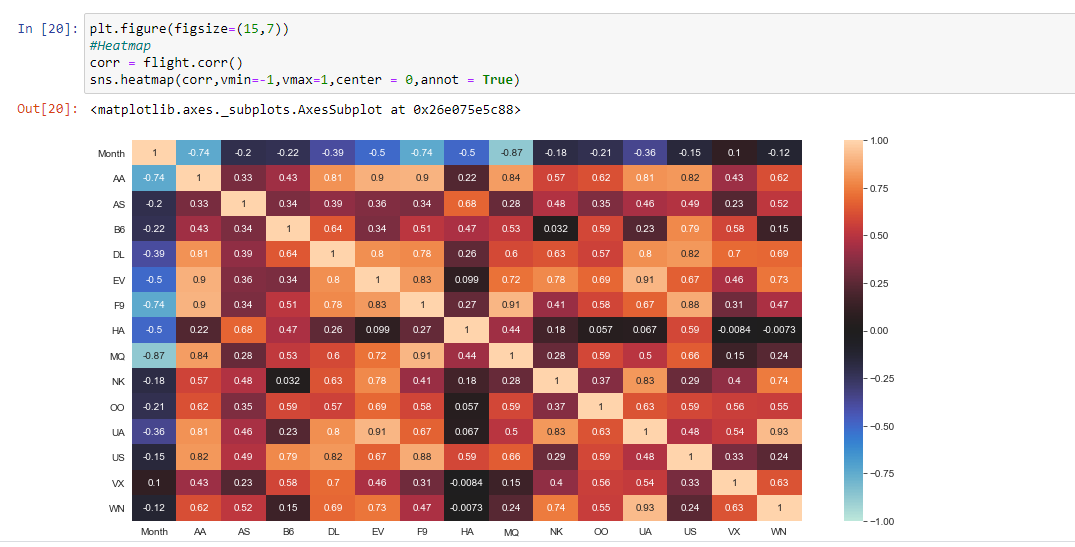
**LINEPLOT:**



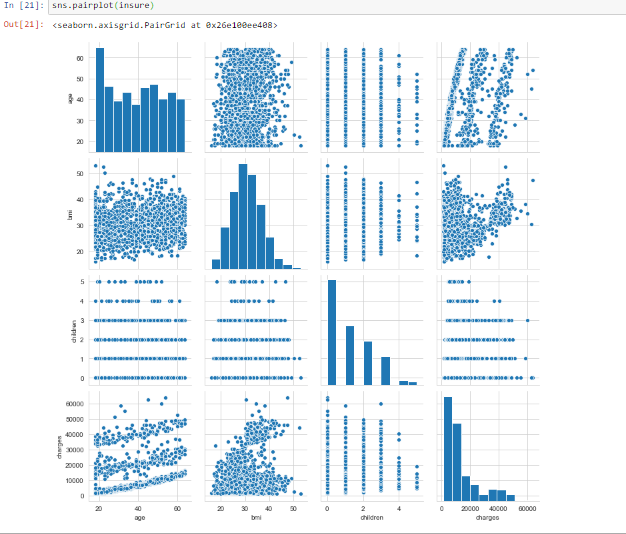
**BARPLOT:**



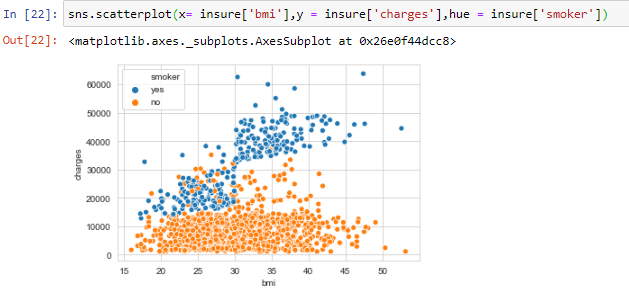
**HEATMAP WITH CORR VALUE:**



**PAIRPLOT**:



**SCATTERPLOT:**



**TO BE CONTINUED…….**