**ASSIGNMENT 1**

**AIM:** Compute and display summary statistics for each feature

available in the dataset. (eg. minimum value, maximum

value, mean, range, standard deviation, variance and

percentiles

Data Visualization-Create a histogram for each feature in

the dataset to illustrate the feature distributions. Plot each

histogram.

Create a boxplot for each feature in the dataset. All of the

boxplots should be combined into a single plot. Compare

distributions and identify outliers.

In this assignment, students are expected to study

statistical tools, Data Visualization tools and then

implement point 1-4.

Objective: Students are expected to study

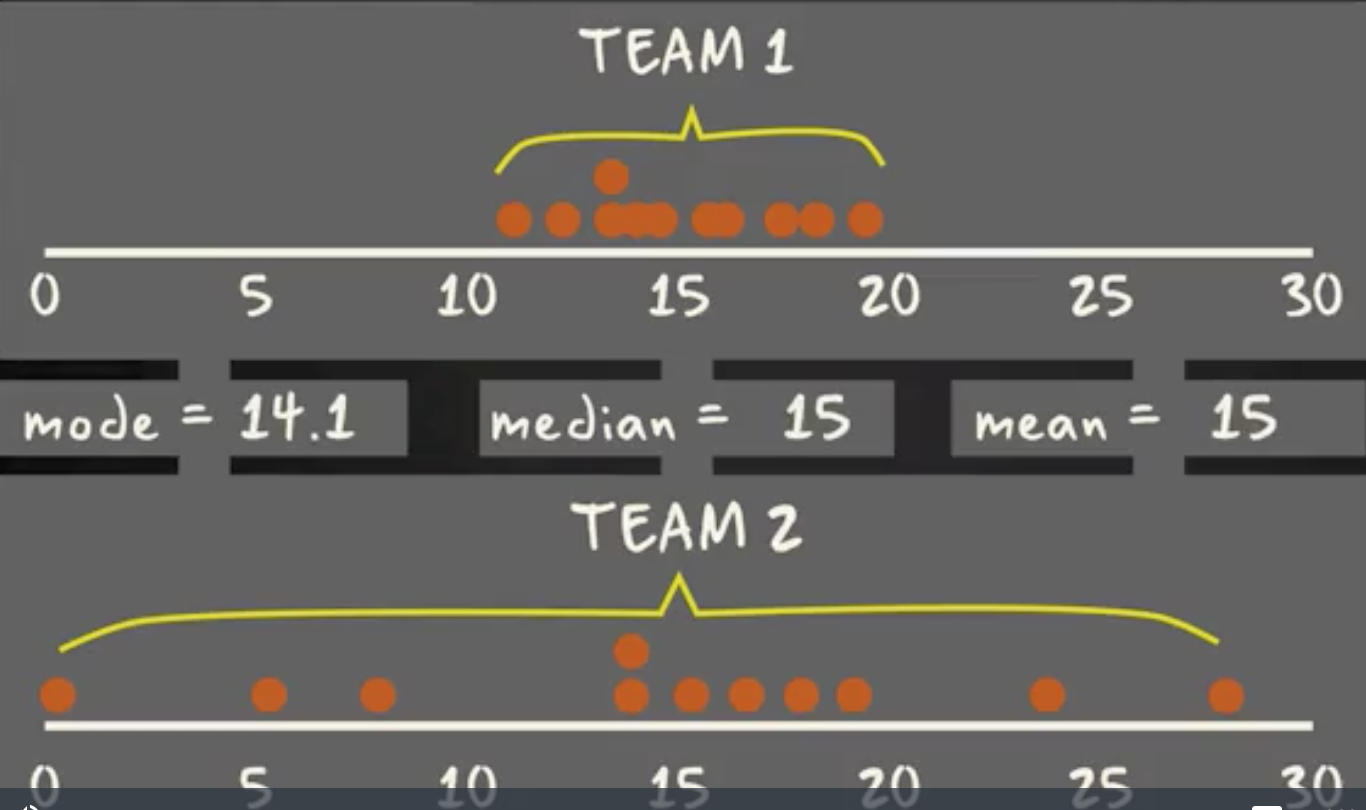
statistical tools, Data Visualization tools and then

implement point 1-4.

**Theory:**

For some distributions/datasets, you will find that you need more information than the measures of central tendency (median, mean, and mode).

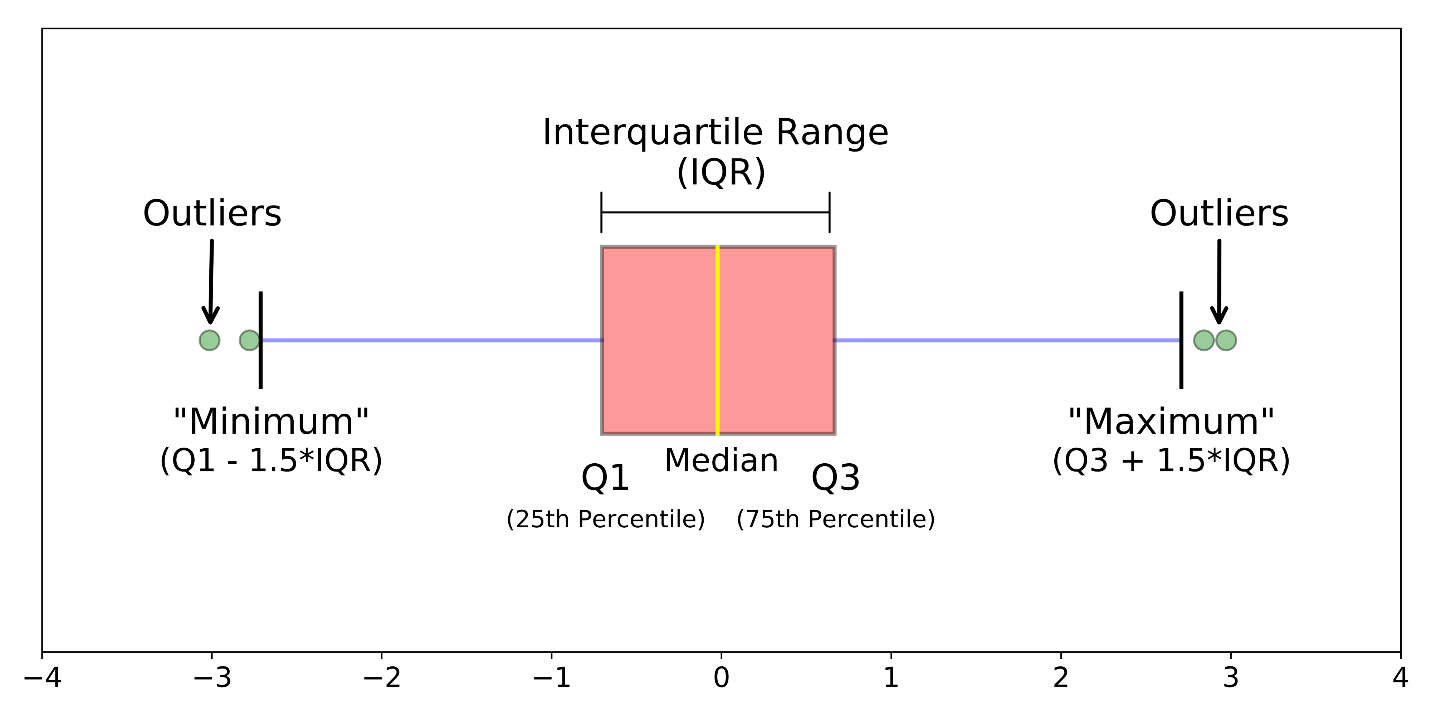




There are times when mean, median, and mode aren’t enough to describe a dataset (taken from [here](https://www.coursera.org/lecture/basic-statistics/1-05-range-interquartile-range-and-box-plot-RbWIZ)).

You need to have information on the variability or dispersion of the data. A boxplot is a graph that gives you a good indication of how the values in the data are spread out. Although box lots may seem primitive in comparison to a [histogram](https://datavizcatalogue.com/methods/histogram.html) or [density plot](https://datavizcatalogue.com/methods/density_plot.html), they have the advantage of taking up less space, which is useful when comparing distributions between many groups or datasets.





Different parts of a boxplot

Boxplots are a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).

**median (Q2/50th Percentile)**: the middle value of the dataset.

**first quartile (Q1/25th Percentile)**: the middle number between the smallest number (not the “minimum”) and the median of the dataset.

**third quartile (Q3/75th Percentile)**: the middle value between the median and the highest value (not the “maximum”) of the dataset.

**interquartile range (IQR)**: 25th to the 75th percentile.

**whiskers (shown in blue)**

**outliers (shown as green circles)**

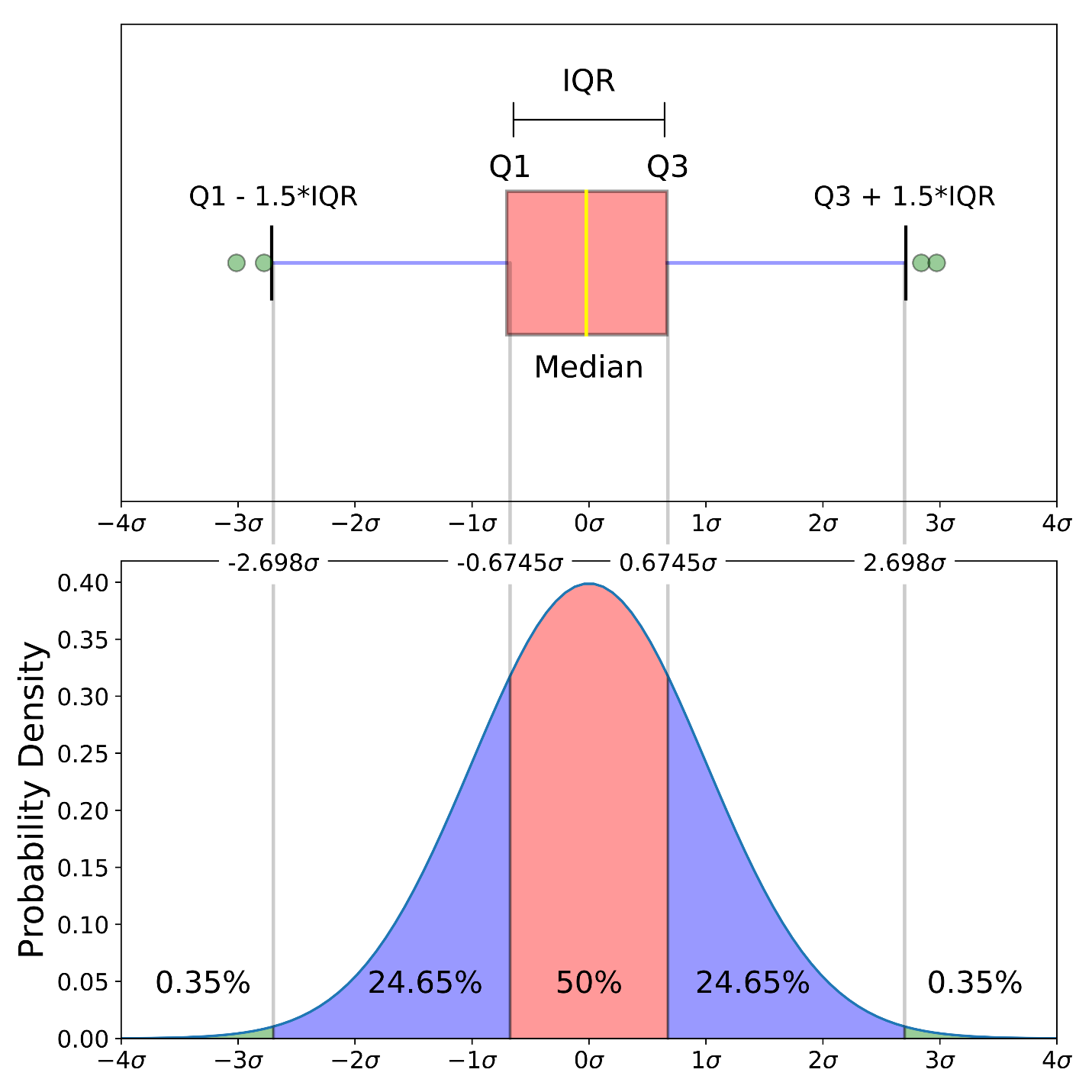
**“maximum”**: Q3 + 1.5\*IQR

**“minimum”**: Q1 -1.5\*IQR

What defines an outlier, “minimum”, or“maximum” may not be clear yet. The next section will try to clear that up for you.

**Boxplot on a Normal Distribution**





Comparison of a boxplot of a nearly normal distribution and a probability density function (pdf) for a normal distribution

The image above is a comparison of a boxplot of a nearly normal distribution and the probability density function (pdf) for a normal distribution. The reason why I am showing you this image is that looking at a statistical distribution is more commonplace than looking at a box plot. In other words, it might help you understand a boxplot.

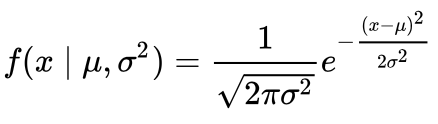
This section will cover many things including:

* How outliers are (for a normal distribution) .7% of the data.
* What a “minimum” and a “maximum” are

**Probability Density Function**

This part of the post is very similar to the [68–95–99.7 rule article](https://towardsdatascience.com/understanding-the-68-95-99-7-rule-for-a-normal-distribution-b7b7cbf760c2), but adapted for a boxplot. To be able to understand where the percentages come from, it is important to know about the probability density function (PDF). A PDF is used to specify the probability of the [random variable](https://en.wikipedia.org/wiki/Random_variable) falling *within a particular range of values*, as opposed to taking on any one value. This probability is given by the [integral](https://en.wikipedia.org/wiki/Integral) of this variable’s PDF over that range — that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. This definition might not make much sense so let’s clear it up by graphing the probability density function for a normal distribution. The equation below is the probability density function for a normal distribution

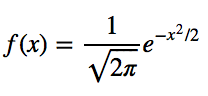




PDF for a Normal Distribution

Let’s simplify it by assuming we have a mean (μ) of 0 and a standard deviation (σ) of 1.



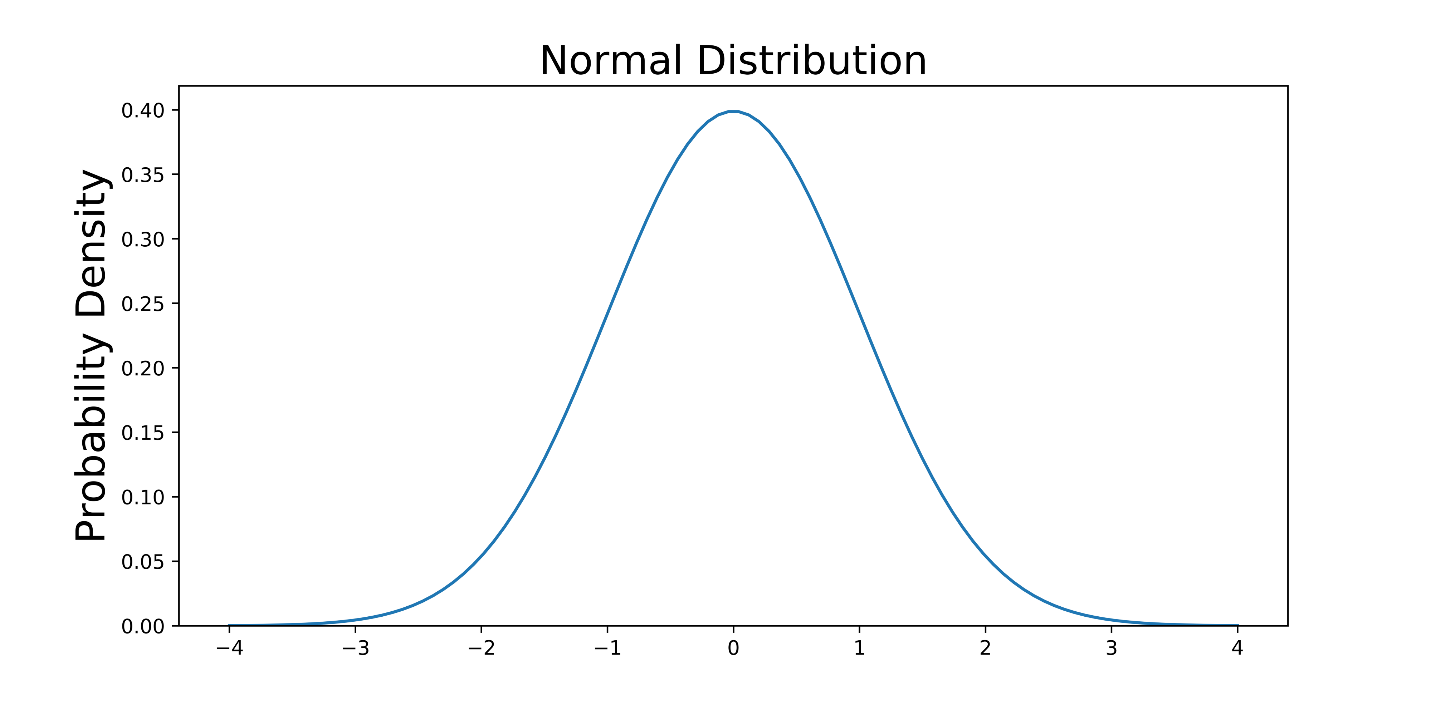


PDF for a Normal Distribution

This can be graphed using anything, but I choose to graph it using Python.

# Import all libraries for this portion of the blog post  
from scipy.integrate import quad  
import numpy as np  
import matplotlib.pyplot as plt  
%matplotlib inlinex = np.linspace(-4, 4, num = 100)  
constant = 1.0 / np.sqrt(2\*np.pi)  
pdf\_normal\_distribution = constant \* np.exp((-x\*\*2) / 2.0)  
fig, ax = plt.subplots(figsize=(10, 5));  
ax.plot(x, pdf\_normal\_distribution);  
ax.set\_ylim(0);  
ax.set\_title('Normal Distribution', size = 20);  
ax.set\_ylabel('Probability Density', size = 20);

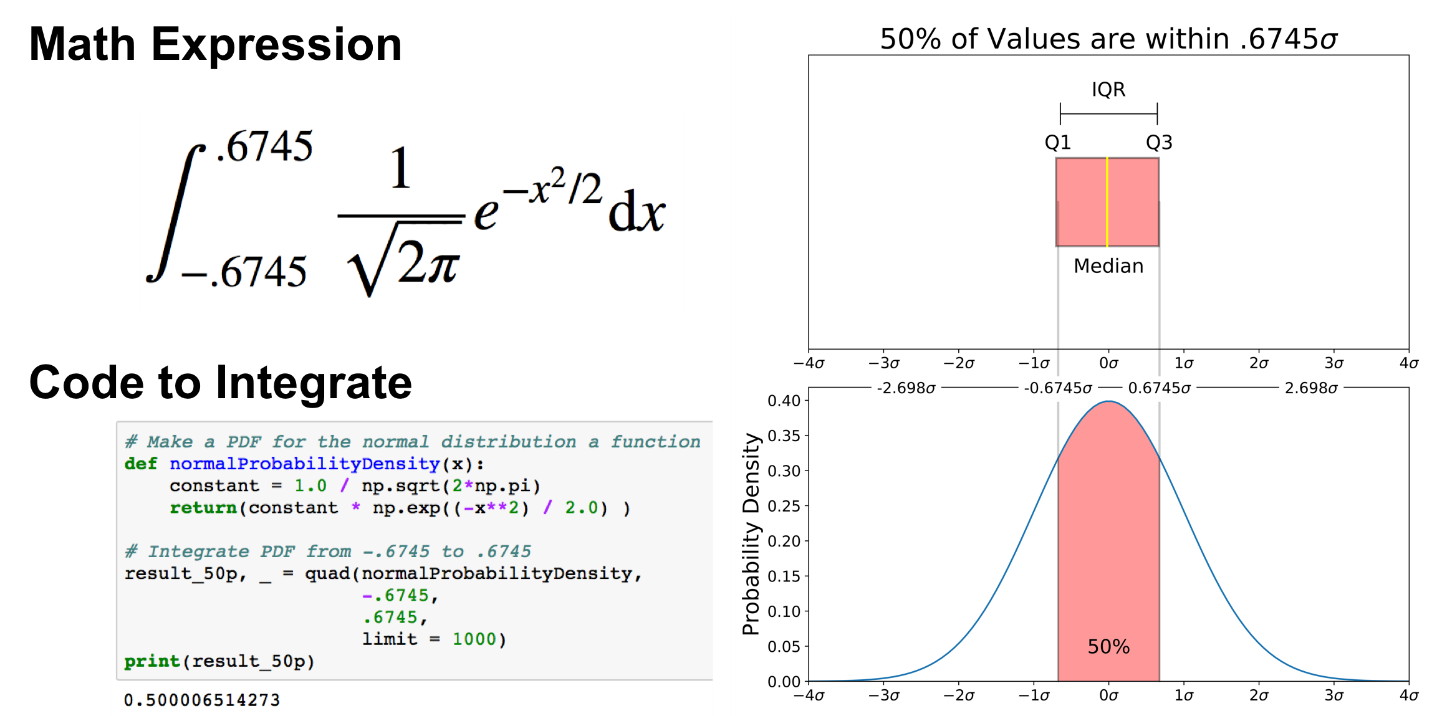




The graph above does not show you the *probability* of events but their *probability density.*To get the probability of an event within a given range we will need to integrate. Suppose we are interested in finding the probability of a random data point landing within the interquartile range .6745 standard deviation of the mean, we need to integrate from -.6745 to .6745. This can be done with SciPy.

# Make PDF for the normal distribution a function  
def normalProbabilityDensity(x):  
 constant = 1.0 / np.sqrt(2\*np.pi)  
 return(constant \* np.exp((-x\*\*2) / 2.0) )# Integrate PDF from -.6745 to .6745  
result\_50p, \_ = quad(normalProbabilityDensity, -.6745, .6745, limit = 1000)  
print(result\_50p)

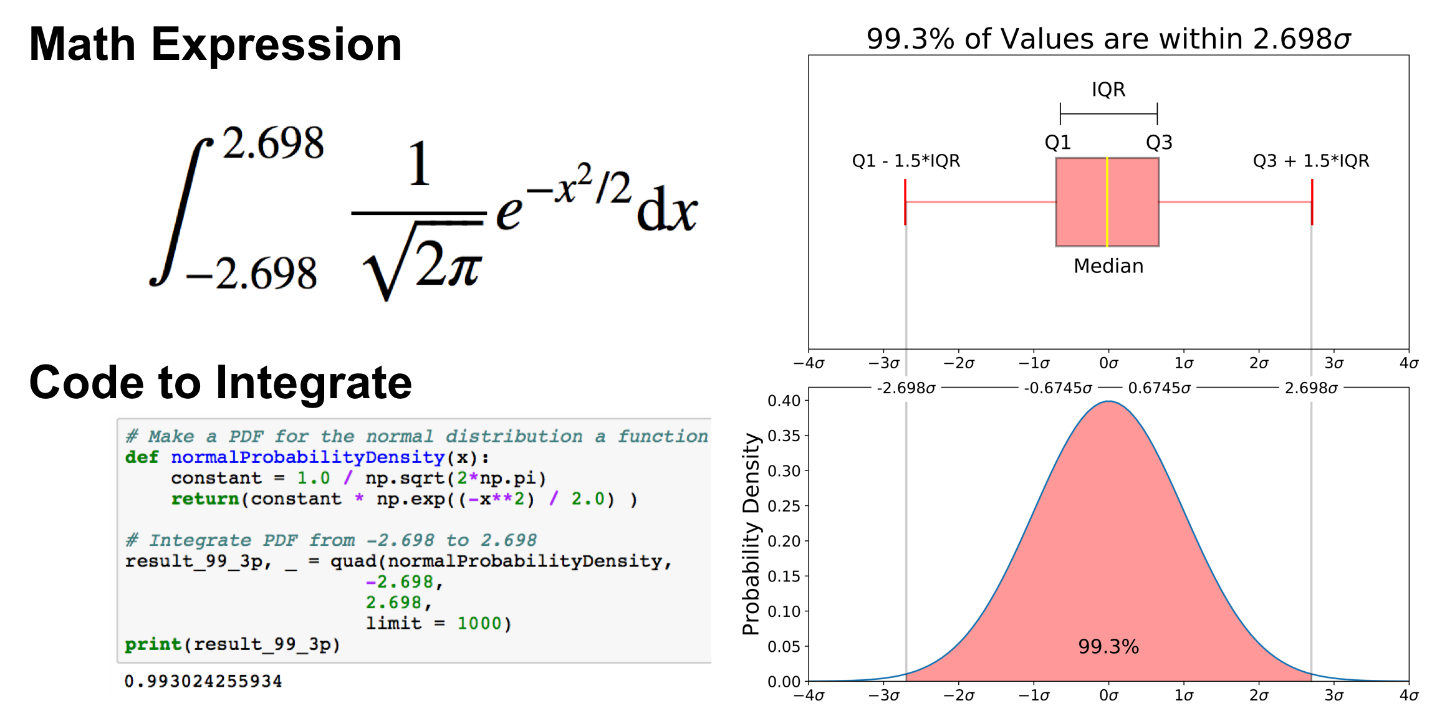




The same can be done for “minimum” and “maximum”.

# Make a PDF for the normal distribution a function  
def normalProbabilityDensity(x):  
 constant = 1.0 / np.sqrt(2\*np.pi)  
 return(constant \* np.exp((-x\*\*2) / 2.0) )# Integrate PDF from -2.698 to 2.698  
result\_99\_3p, \_ = quad(normalProbabilityDensity,  
 -2.698,  
 2.698,  
 limit = 1000)  
print(result\_99\_3p)

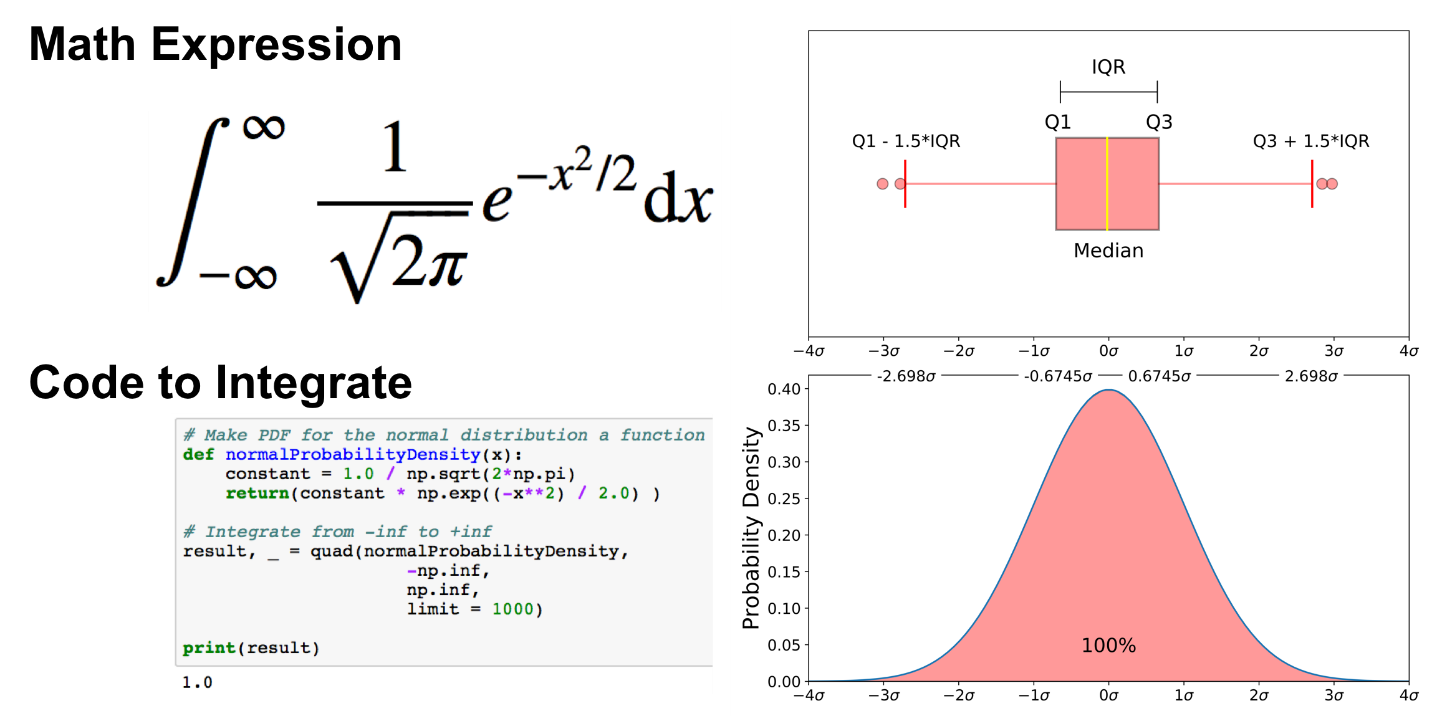




As mentioned earlier, outliers are the remaining .7% percent of the data.

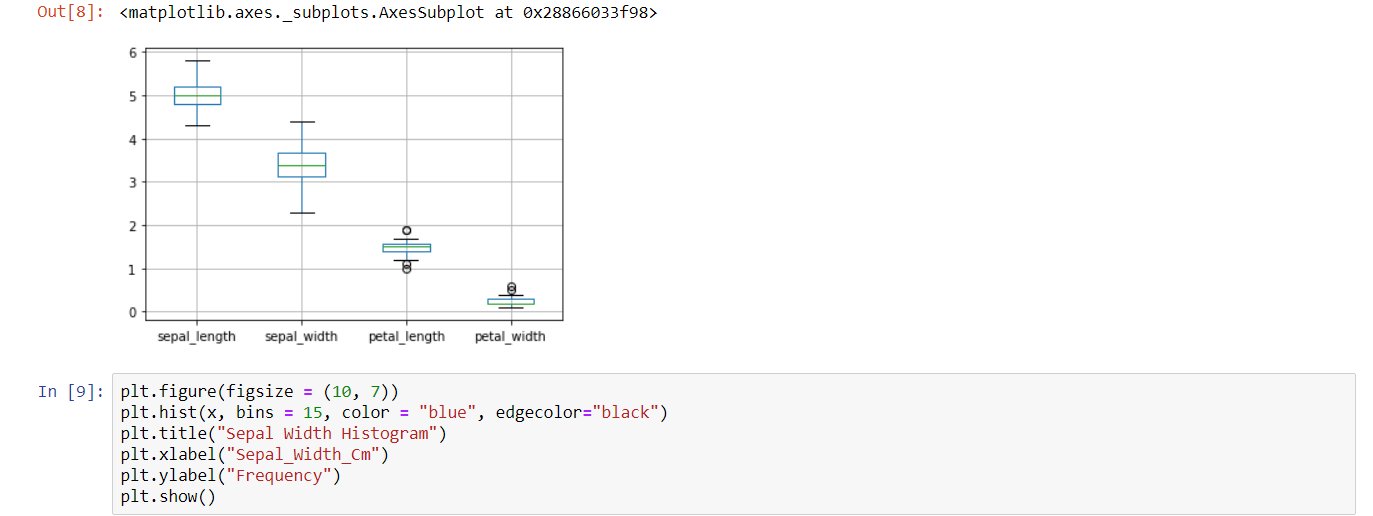
It is important to note that for any PDF, the area under the curve must be 1 (the probability of drawing any number from the function’s range is always 1).

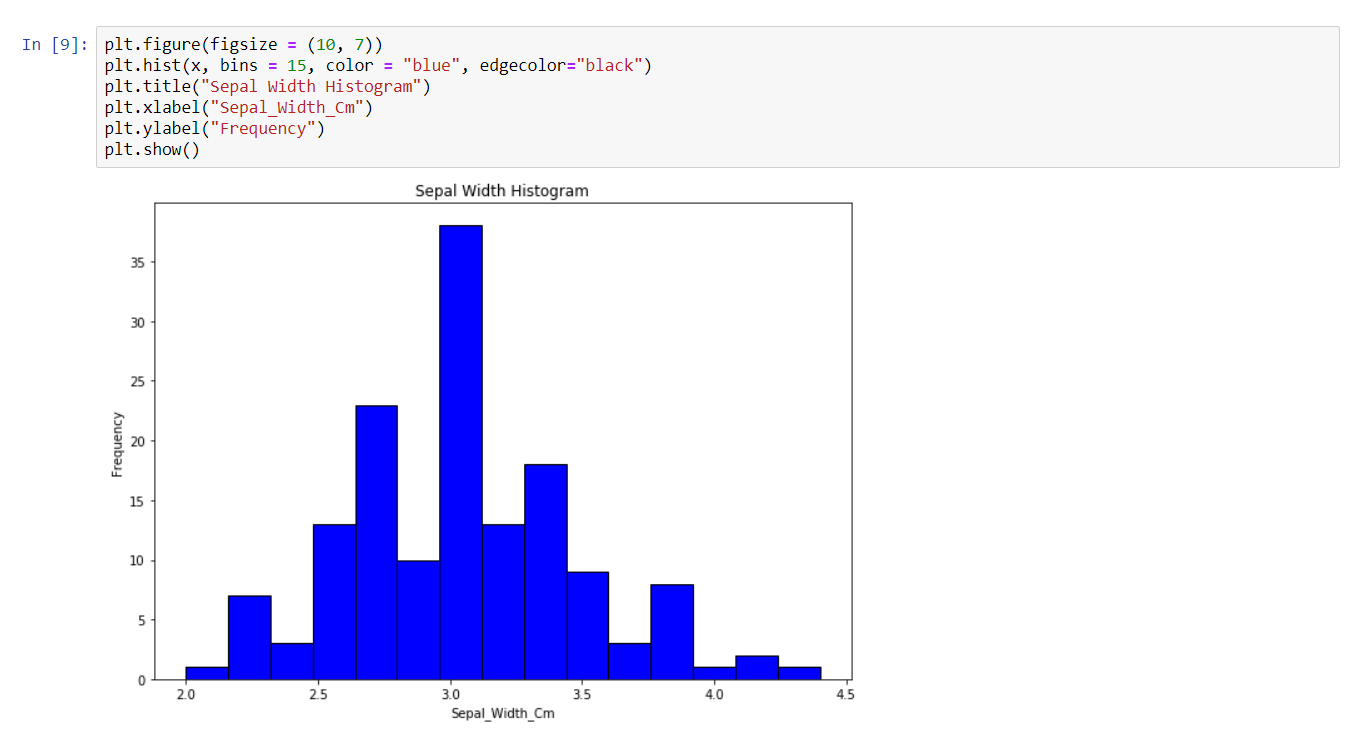
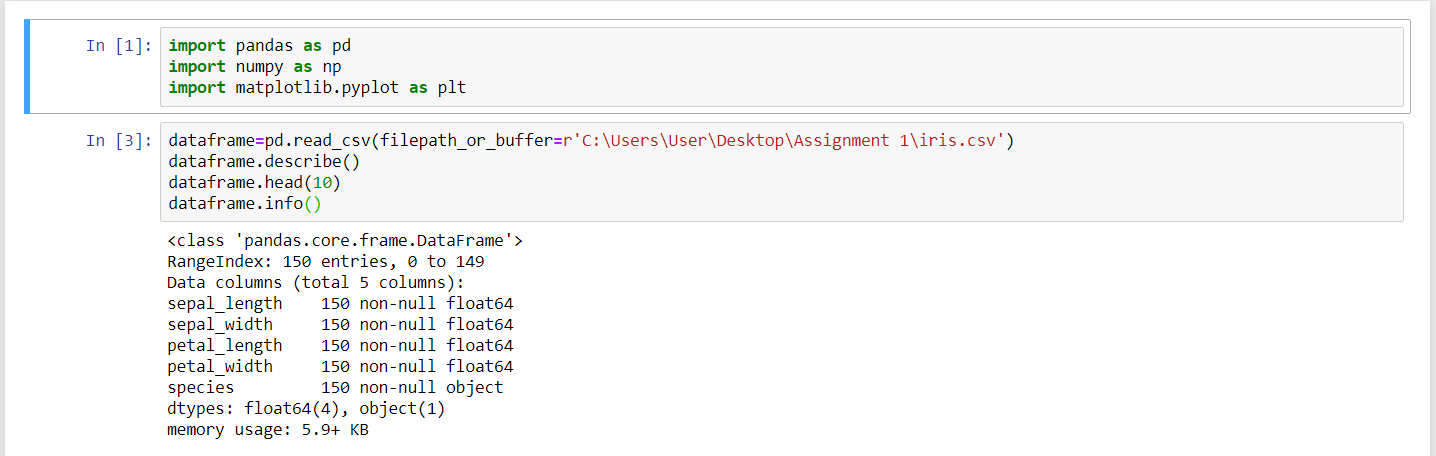
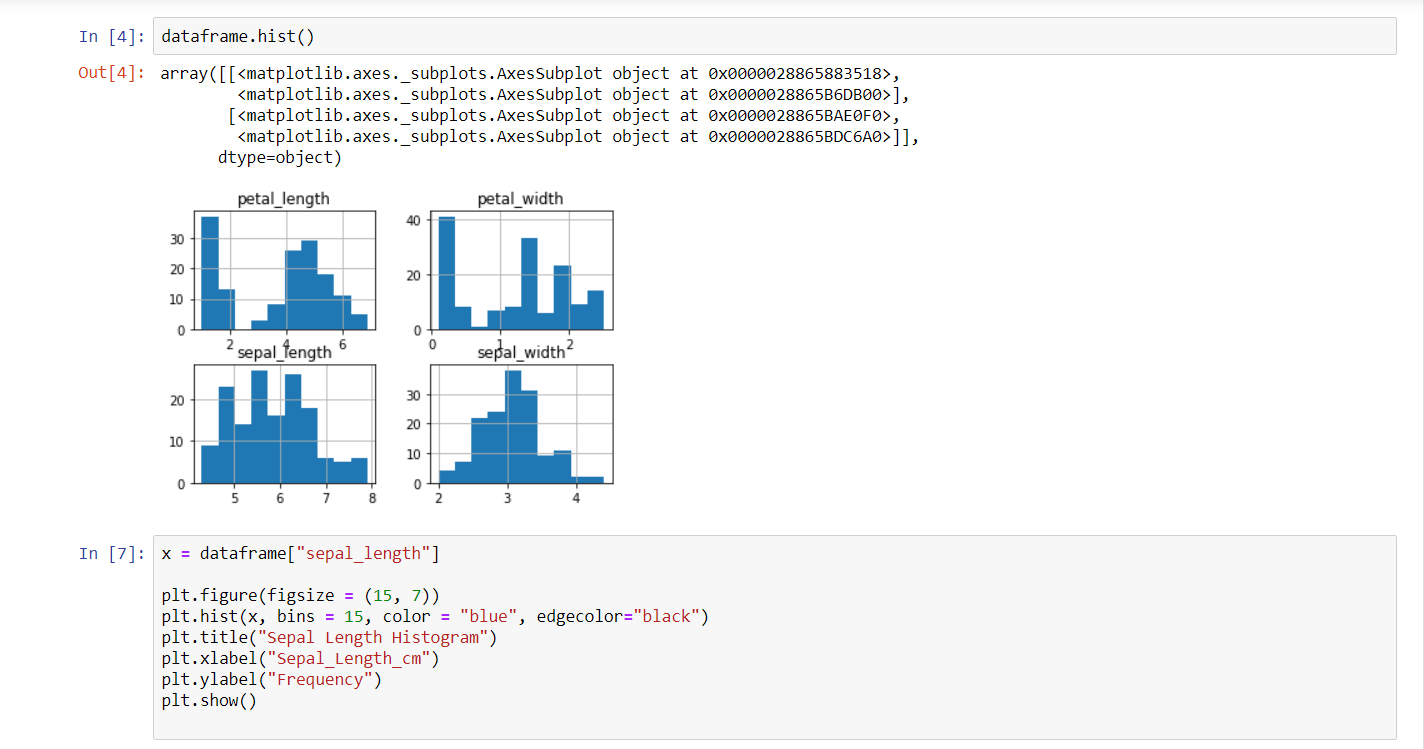
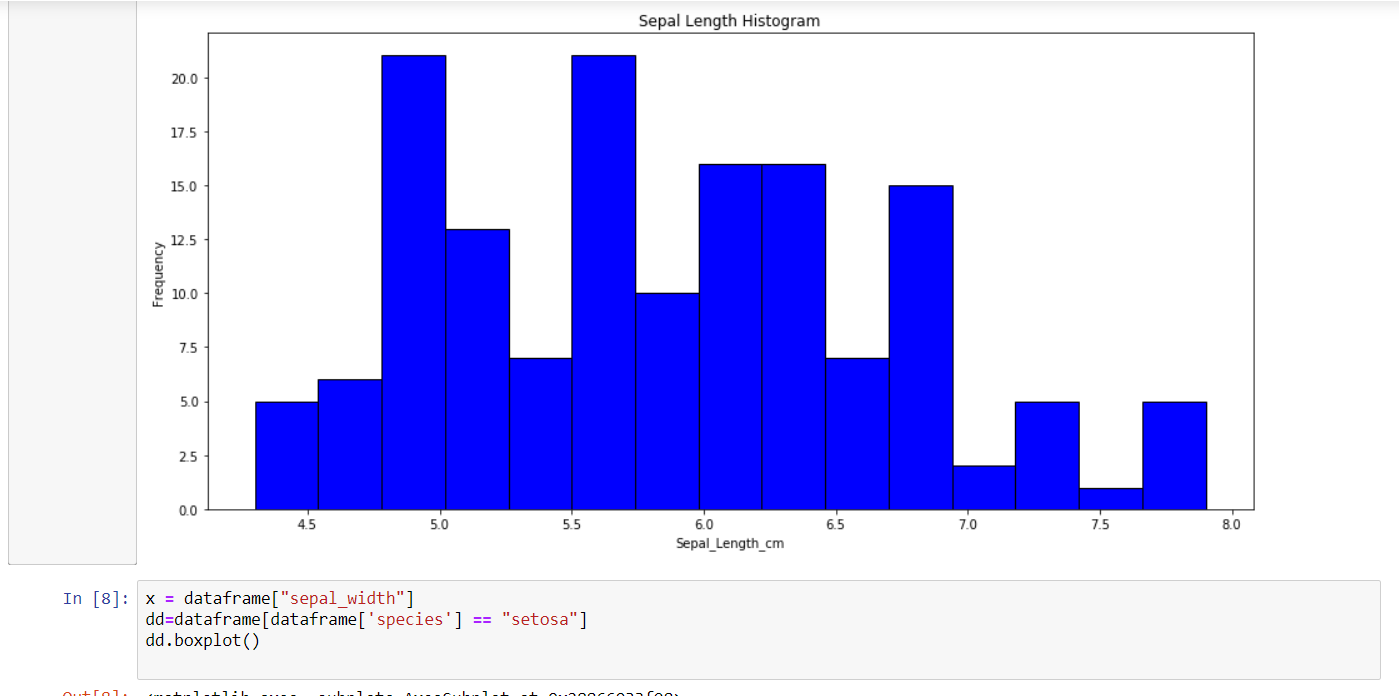
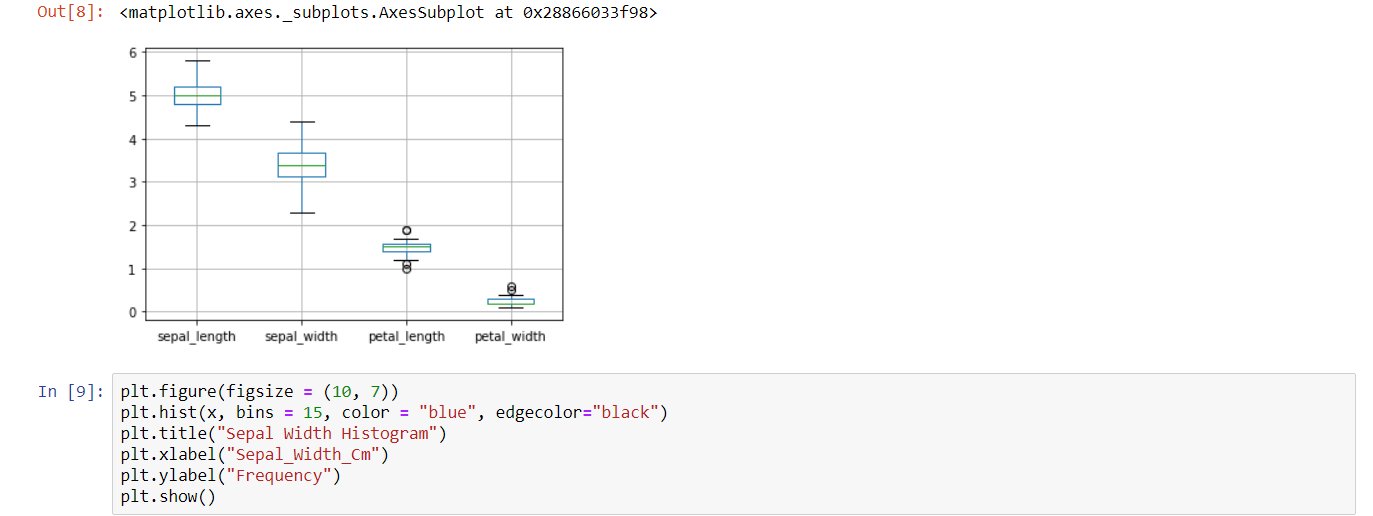




Data Sets used: Iris Data Set. Also, petal width, petal length and all such data sets have been used.

**Expected Output:**



**Conclusion:** We have thus displayed the statistics for each feature in the dataset.