**Data Exploration**

Data exploration is the first step in data analysis involving the use of [data visualization](https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization) tools and statistical techniques to uncover data set characteristics and initial patterns.

During exploration, raw data is typically reviewed with a combination of manual workflows and automated data-exploration techniques to [visually explore data sets](https://www.techtarget.com/searchbusinessanalytics/tip/Seven-good-data-visualization-practices-for-visual-integrity), look for similarities, patterns and outliers and to identify the relationships between different variables.

This is also sometimes referred to as exploratory data analysis, which is a statistical technique employed to analyze raw data sets in search of their broad characteristics.

**Why is data exploration important?**

Humans are visual learners, able to process visual data much more easily than numerical data. Consequently, it's challenging for data scientists to review thousands of rows of data points and infer meaning without assistance. Data visualization tools and elements like colors, shapes, lines, graphs and angles aid in effective data exploration of [metadata](https://www.techtarget.com/whatis/definition/metadata), enabling relationships or anomalies to be detected.

### What industries use data exploration?

Any business or industry that collects or utilizes data can benefit from data exploration. A few common industries include software development, healthcare and education. The advanced visualization techniques employed by data exploration and business intelligence tools enable businesses and stakeholders to better understand performance metrics by [making raw data more comprehensible and creating a "story"](https://www.techtarget.com/searchbusinessanalytics/feature/4-features-of-great-data-visualization-and-storytelling) around it.By visualizing patterns and finding commonalities in complex data flows, data exploration can help enterprises make data-driven decisions to streamline processes, better target their ideal audience, increase productivity and achieve greater returns.

### What is the best language for data exploration?

The most commonly used statistical methods in data exploration are the [R programming language](https://www.techtarget.com/searchbusinessanalytics/definition/R-programming-language) and Python. Both are open source data analytics languages. While R is best for statistical analysis, [Python](https://www.techtarget.com/whatis/definition/Python) is better suited for machine learning algorithms.

When we learn Data Science as beginners we came across a very common term known as Central Tendency Measure with 3 M’s, Mean, Median and Mode. One should understand what do these terms mean and when they are given priority while analyzing any data set and concluding a decision depending upon the types of data.

## **The measure of Central Tendency**

We should first understand the term Central Tendency. Data tend to accumulate around the average value of the total data under consideration. Measures of central tendency will help us to find the middle, or the average, of a data set. If most of the data is centrally located and there is a very small spread it will form an asymmetric bell curve. In such conditions values of mean, median and mode are equal.

When we learn Data Science as beginners we came across a very common term known as Central Tendency Measure with 3 M’s, Mean, Median and Mode. One should understand what do these terms mean and when they are given priority while analyzing any data set and concluding a decision depending upon the types of data.

## **Types of Data**

The data available for analysis always have two main categories, Quantitative and Qualitative. Quantitative data has numerical values such as time, speed, etc. whereas Qualitative data have non-numerical values such as color, yes or no, etc. There are two types of Quantitative data, Discrete and Continuous. The Discrete data has values that can not be broken down into fractions such as numbers on dice, number of students in a class, etc. while the Continuous data can be available in fractional values such as the height of a person, distance, etc.

The data can be further classified into Nominal, Ordinal, Interval, and Ratio data. Nominal data is categorical data such as gender, religion, etc. Ordinal data is a measure of rank or order such as exam grades, position in the company, etc. Interval data is a measure of equality and interval such as 30 oC is hotter than 15 oC, water is in liquid form when the temperature is in between 0 oC and 100 oC. etc. When in addition to setting up inequalities we can also form quotients such data is known as Ratio data, such as the ratio of height, weight, etc.

## **The measure of Central Tendency**

We should first understand the term Central Tendency. Data tend to accumulate around the average value of the total data under consideration. Measures of central tendency will help us to find the middle, or the average, of a data set. If most of the data is centrally located and there is a very small spread it will form an asymmetric bell curve. In such conditions values of mean, median and mode are equal.

**Mean:** It is the average of values.

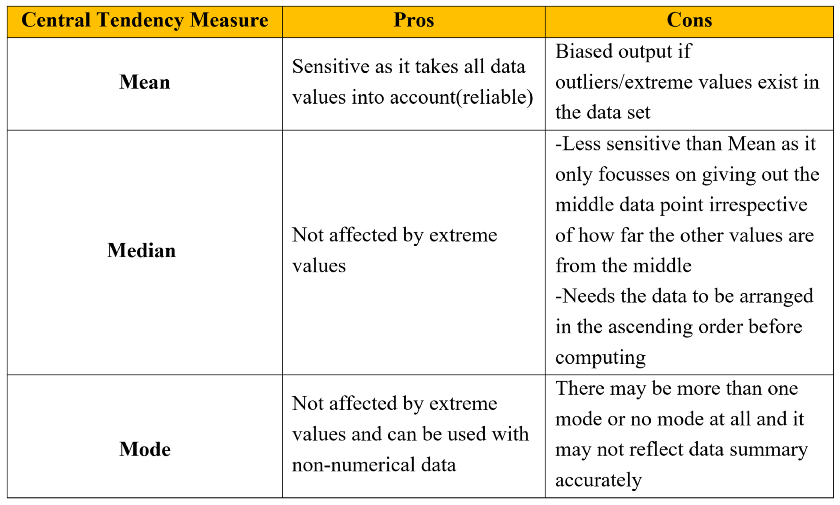
Consider 3 temperature values 30 o C, 40 o C and 50 o C, then the mean is (30+40+50)/3=40 o C.

**Median:** It is the centrally located value of the data set sorted in ascending order.

Consider 11 (ODD) values 1,2,3,7,8,3,2,5,4,15,16.

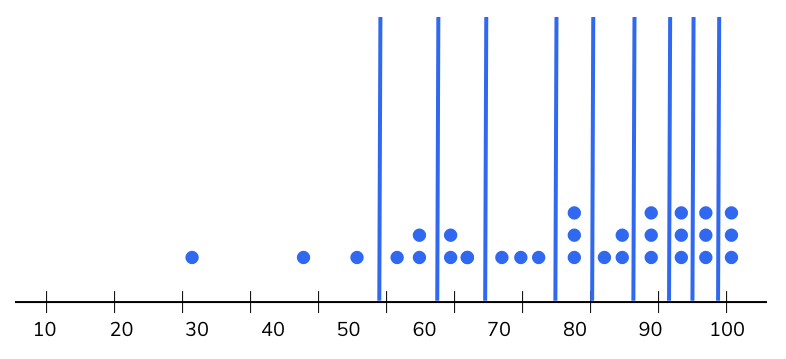
We first sort the values in ascending order 1,2,2,3,3,4,5,7,8,15,16 then the median is 4 which is located at the 6th number and will have 5 numbers on either side.

**Mode:** It is the most frequent value in the data set. We can easily get the mode by counting the frequency of occurrence.



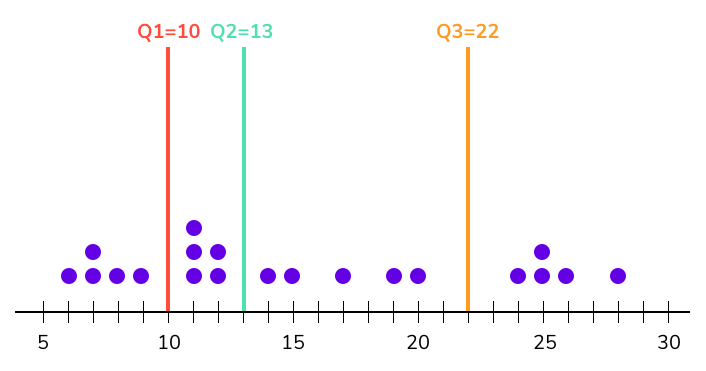
**Quantiles**

Quantiles are the set of values/points that divides the dataset into groups of equal size. For example, in the figure, there are nine values that splits the dataset. Those nine values are quantiles.



**Quartiles**

The three dividing points (or quantiles) that split data into four equally sized groups are called quartiles. For example, in the figure, the three dividing points Q1, Q2, Q3 are quartiles.



A quartile is a statistical term that splits data into quarters or four defined intervals. It basically divides the data points into a data set in 4 quarters on the number line. One thing we need to keep in mind is that data points can be random and we have to put those numbers in line first on the number line in ascending order and then divide them into quartiles. It is basically an extended version of the median. Median divides the data into two equal parts and quartiles divide it into four parts. Once we divide the data, the four quartiles will be:

* 1st quartile or lower quartile basically separates the lowest 25% of data from the highest 75%.
* 2nd quartile or middle quartile is also the same as the median. It divides numbers into 2 equal parts.
* 3rd quartile or the upper quartile separates the highest 25% of data from the lowest 75%.
* Let’s say that we have a data set with N data points:
* X – {X1, X2, X3……….. XN}
* The formula for quartiles is given by:
* **Lower Quartile (Q1) = (N+1) \* 1 / 4**
* **Middle Quartile (Q2) = (N+1) \* 2 / 4**
* **Upper Quartile (Q3 )= (N+1) \* 3 / 4**
* **Interquartile Range = Q3 – Q1**

To understand the data well, only studying measures of central tendency is not enough. One essential measure is how the data is scattered or dispersed. Measures of dispersion indicate how the data is spread or scattered from the measures of central tendency. Measures Of dispersion is also known as “Measures of Variability” because it indicates the variability of the data that how much we still do not know about the data.

four commonly used measures of dispersion.

1. Range
2. Inter-quartile range (IQR)
3. Variance
4. Standard deviation

### ****Range****

The simplest measure of dispersion is Range; it is the difference between the highest value and lowest value in the dataset. It offers a crude insight into the spread of the data, but very susceptible to outliers. The range is helpful when you want to focus on extreme values in the dataset. The formula of Range is:

**Range = Highest value – lowest value**

### ****Inter-Quartile Range****

The interquartile range is a measure of dispersion, as it also measures the variability of the data, IQR indicates how the data in a series is dispersed from the mean. It measures the difference between the third quartile and the first quartile of the data. It means IQR measure the spread of the middle 50% of the dataset. As the IQR goes up the data points are more spread out and if the IQR is small they assumed to be data is spread around the mean. IQR is also very helpful to determine the outlier in the datasets. To calculate IQR first we have to sort the data in ascending order.

**The Formula of IQR is:**

**IQR = Third Quartile – First Quartile**

Let’s understand how to find the interquartile range:

Let’s understand how to find the interquartile range:

Suppose we have a data series 88,89,89,89,90,91,91,91,92

So, to find out the IQR first we have to sort the data on ascending order as the data is already sorted so we don’t need to sort it. Now next find the median (middle value) of the data this is identified as **Q2**, the middle value of the dataset is 90.

88,89,89,89,90,91,91,91,92

As the dataset is divided into two parts, now find the middle value of the first half which is identified as **Q1** is 89 and second half which is identified as **Q3** is 91.

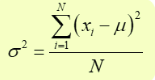
So, the IQR is – = 91-89= 2

We used **IQR**when we are more interested in middle value and less interested in extremes.

**Variance**

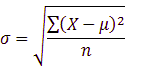
Variance is one of the important measures of dispersion, Variance measure the variability of the data around its mean or average. In other words, variance indicates how the data is deviated or dispersed from its mean or average. High variance means there is more variability or we can say that the data deviates more from its mean whereas low variance means there is less variability. If the variance is zero that means all the values in the data are identical. Variance can never be negative. It is denoted by (sigma square).

**Formula for population variance:**



where **N** is the population size and the **X**are data points and **μ** is the population mean.

**The formula of standard deviation for population:**



where **n**is the sample size and **X**are the data points and μ is the sample mean.

**Week 1**

Basic statistical functions for data exploration

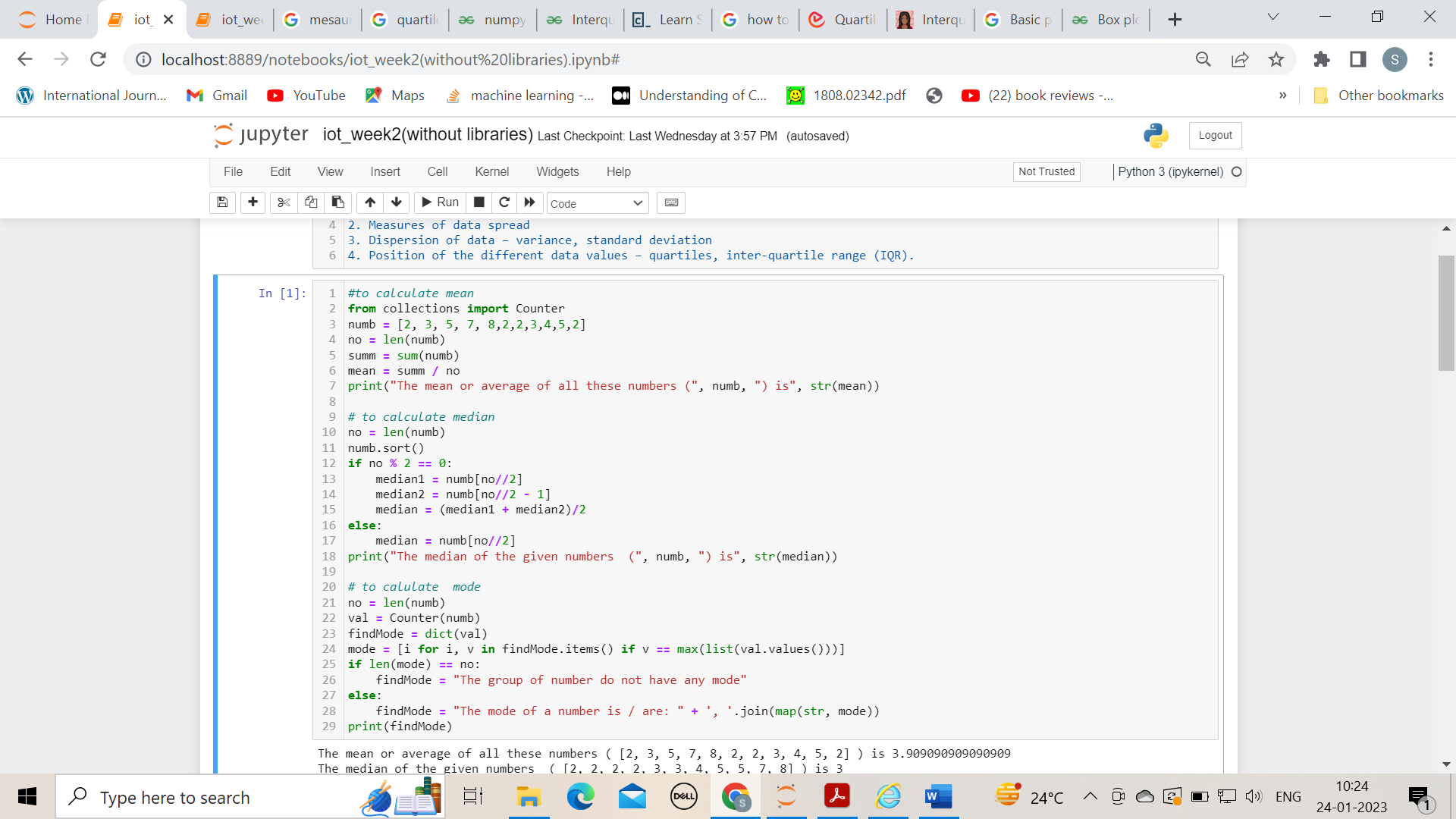
1. Measures of central tendency – mean, median, mode

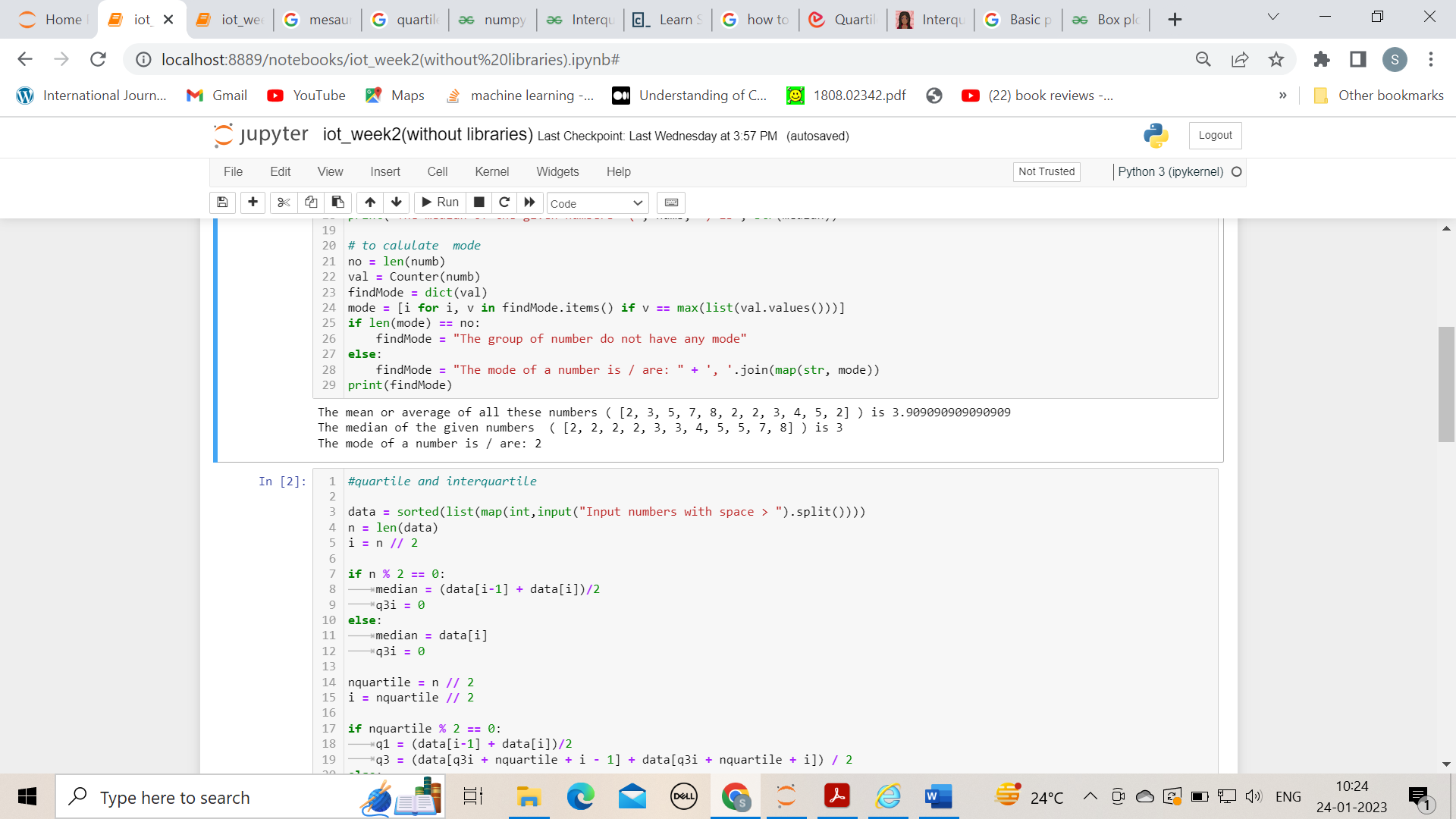
2. Measures of data spread

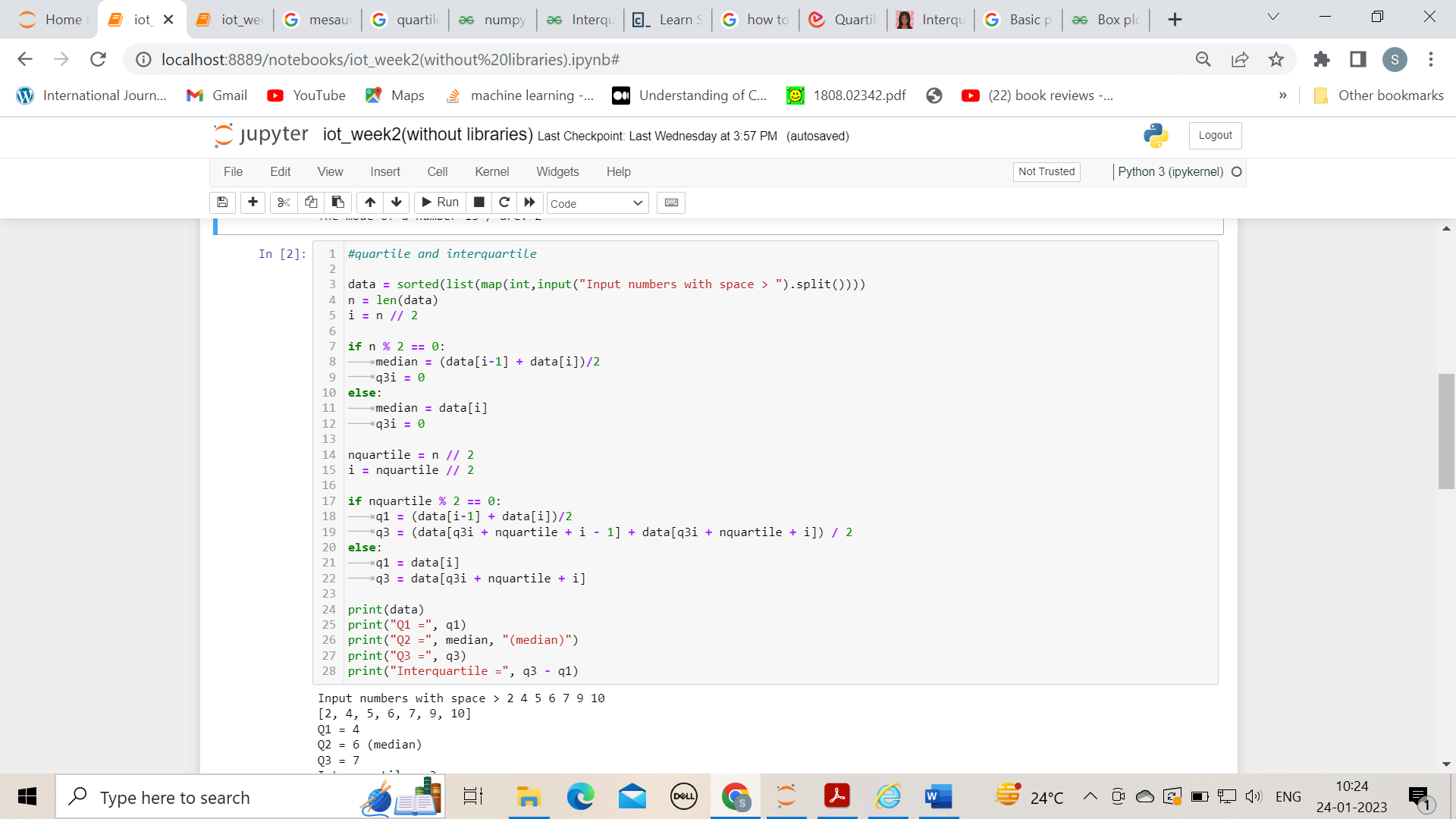
3. Dispersion of data – variance, standard deviation

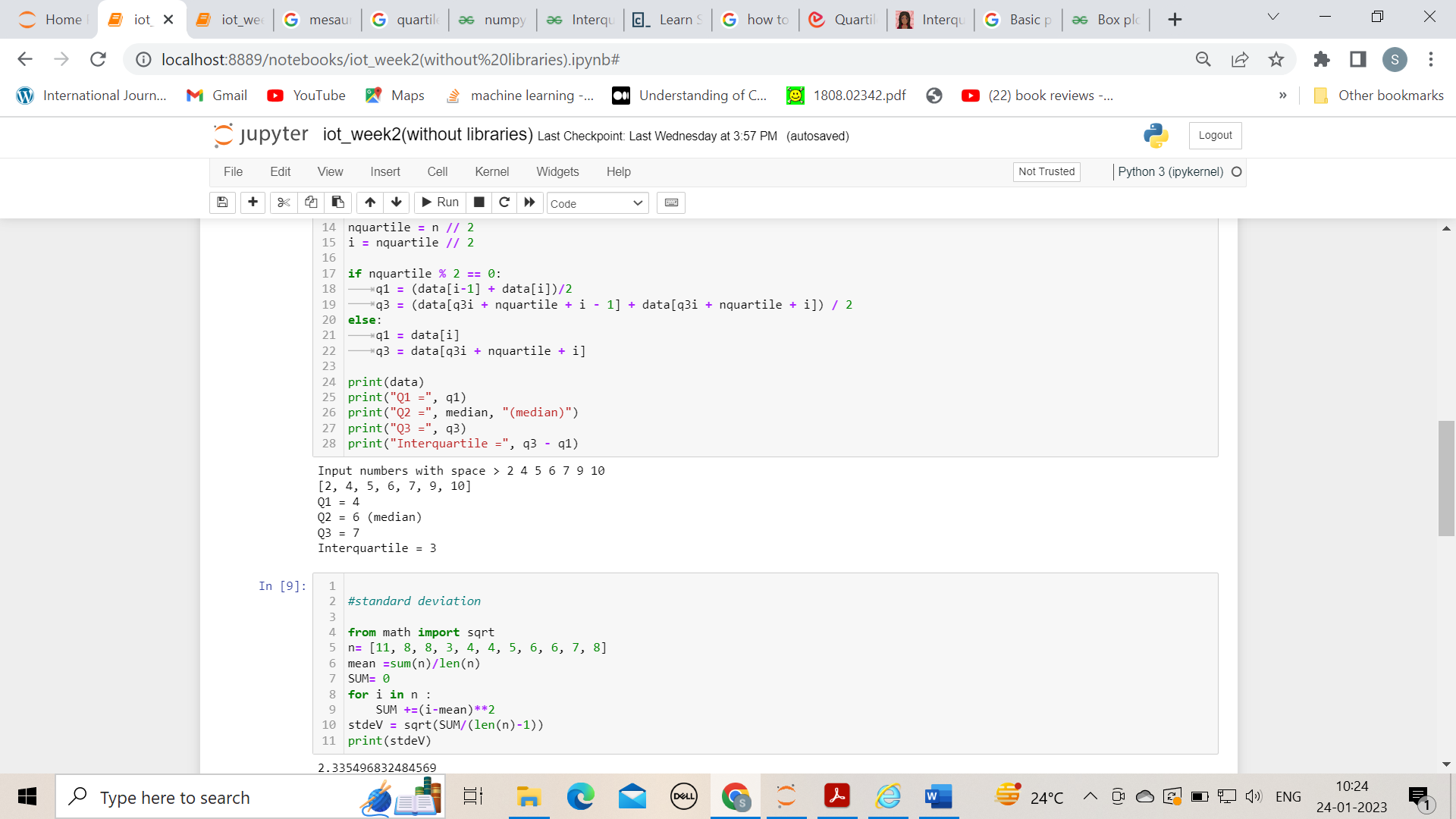
4. Position of the different data values – quartiles, inter-quartile range (IQR).

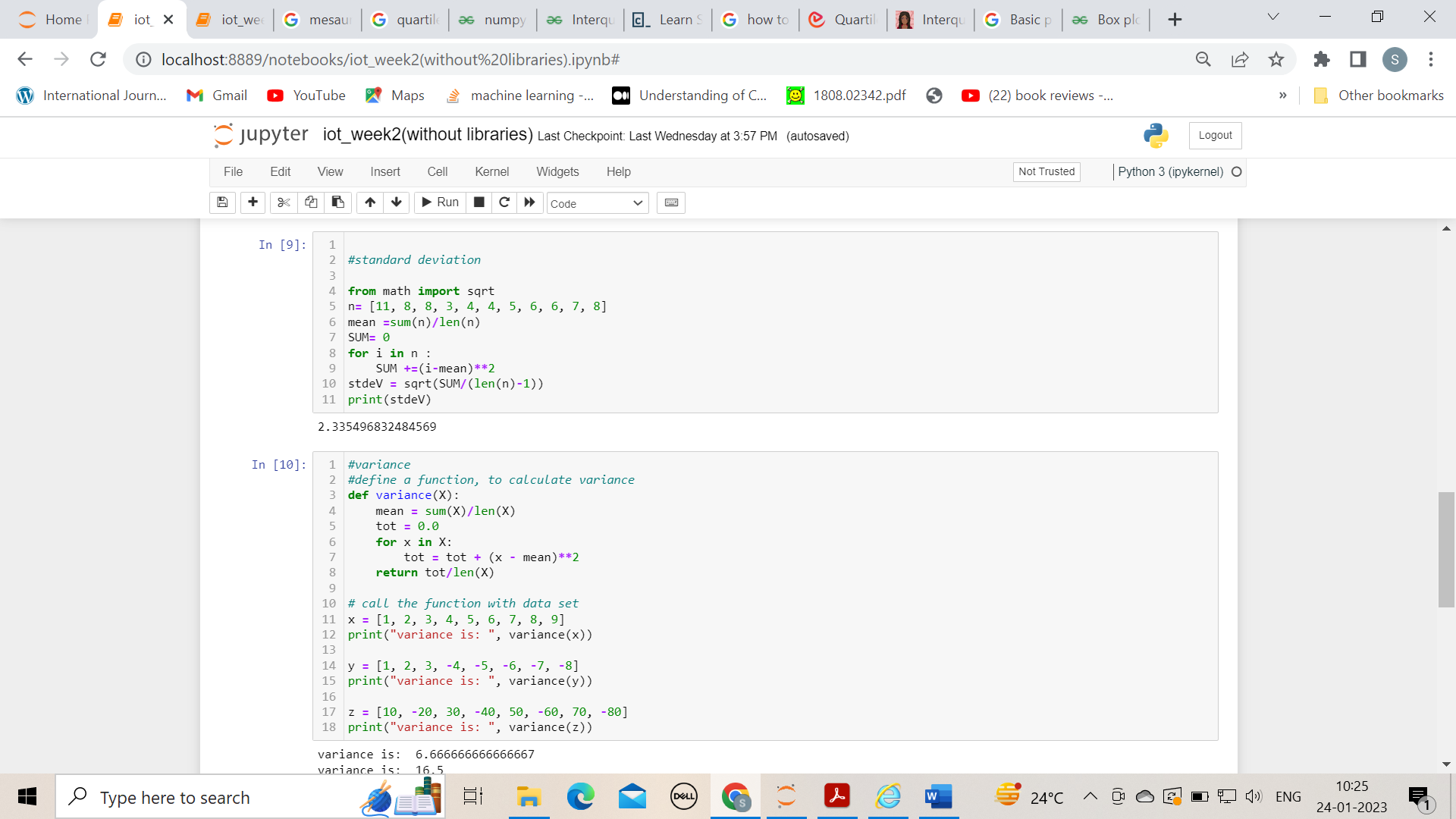
**Without using the libraries**











**Using libraries**

