Data Science

Python is the programming language of choice for data scientists. Although it wasn't the first primary programming language, its popularity has grown throughout the years.

Prerequisites:

Four Basic Concepts:

1. Linear Algebra
2. Statistics
3. Probability
4. Calculus

3 Major Topics for Data Science

1. Statistics
2. Data Analytic
3. Data Visualization

Statistics

– the science of collecting, organizing, analyzing, and interpreting data.

Main Branches of Statistics

1. **Descriptive** – describes a sample.
2. **Inferential** – uses the sample data to make inferences or conclusions about a larger population.

Common tools of Descriptive Statistics

1. **Measure of Central tendency** – tell us where most values fall or the typical value of a data set.
   - **Mode** – the value that shows up the greatest number of times in data set. For example, in the data set {1,1,3,5,7,10} the mode is 1 because it appears more than any other number.
   - **Mean** – or average, is the sum of all the numbers divided by the number of elements in the sample. For example, in the data set {1,1,3,5,7,10} the mean is 4.5

- **Median** - is the middle number of the sorted data set. Sort the data set first before getting the median. For example, in the data set {1,1,3,5,7,10} since we have an even number of elements, the median is the average of two terms in the middle. In this data set the median is 4 (or 3+5/2). Example 2: {1,2,4,5,11} the median is 4.

  Note: In python, it is easy using different libraries.

  Examples:

  Mean

```
1  import numpy
2  speed = [99,86,87,88,111,86,103,87,94,78,77,85,86]
3  x = numpy.mean(speed)
4  print(x)
```

[2] ✓ 1.3s

```
89.76923076923077
```

```
1  import numpy
2  speed = [1,2,3,4,5]
3  x = numpy.mean(speed)
4  print(x)
```

[3] ✓ 0.1s

```
3.0
```

  Median

```
1  import numpy
2  speed = [99,86,87,88,111,86,103,87,94,78,77,85,86]
3  x = numpy.median(speed)
4  print(x)
```

[4] ✓ 0.1s

```
87.0
```

  Mode

```
1  from scipy import stats
2  speed = [99,86,87,88,111,86,103,87,94,78,77,85,86]
3  x = stats.mode(speed)
4  print(x)
```

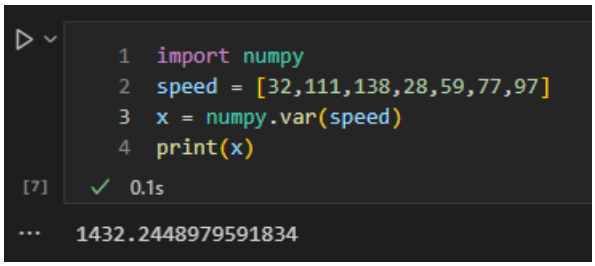
[5] ✓ 17.8s

```
ModeResult(mode=array([86]), count=array([3]))
```

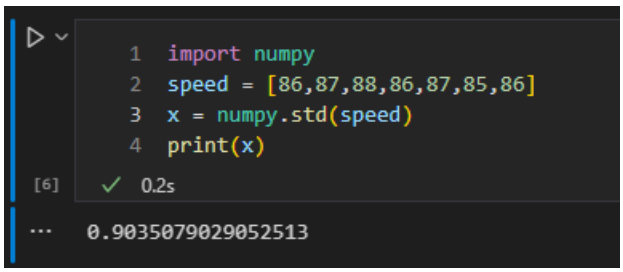
2. **Measures of Dispersion or spread**

Common measures of dispersion

- **range** – is the difference between the high and low values.
    - {1,1,3,5,7,10} the range is 9 (10-1)
- **variance** – measures how far a data set is spread out.



- **standard deviation** – measures how far each observed value is from the mean. Standard deviation is a number that describes how spread out the values are.



Note: If the standard deviation is small, the data values are close to the mean value. If it is high, the data values are widely spread out from the mean value.

Data Analysis

– is the process of collecting, modeling, and analyzing data to extract insights that support decision-making.

Types of Data Analytic:

- **Descriptive**
    - Descriptive Analytics is the examination of data or content, usually manually performed, to answer the question "What happened?" (or What is happening?), characterized by traditional business intelligence (BI) and

visualizations such as pie charts, bar charts, line graphs, tables, or generated narratives.

- **Diagnostics**
  - Diagnostic analytics is a branch of analytics that aims to answer the question, “Why did this happen?” By using diagnostic analytics, companies can gain insights into the causes of patterns they've observed in their data. Diagnostic analytics can involve a variety of techniques, including data drilling and data mining.
- **Predictive**
  - Predictive analytics is the use of data to predict future trends and events. It uses historical data to forecast potential scenarios that can help drive strategic decisions.
  - Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modeling, and machine learning that analyze current and historical facts to make predictions about future or otherwise unknown events.
- **Prescriptive**
  - Prescriptive analytics is the process of using data to determine an optimal course of action. By considering all relevant factors, this type of analysis yields recommendations for next steps. Because of this, prescriptive analytics is a valuable tool for data-driven decision-making.
  - Prescriptive analytics is the third and final phase of business analytics, which also includes descriptive and predictive analytics.

Statistical Data Types:

- **Quantitative Data** – numerical and countable (example: no. of children, price, weight)
  - **Discrete** – can't be divided (e.g., No. of children, because no 2.5 children in a family)
  - **Continuous** – can be divided (e.g., price, weight)

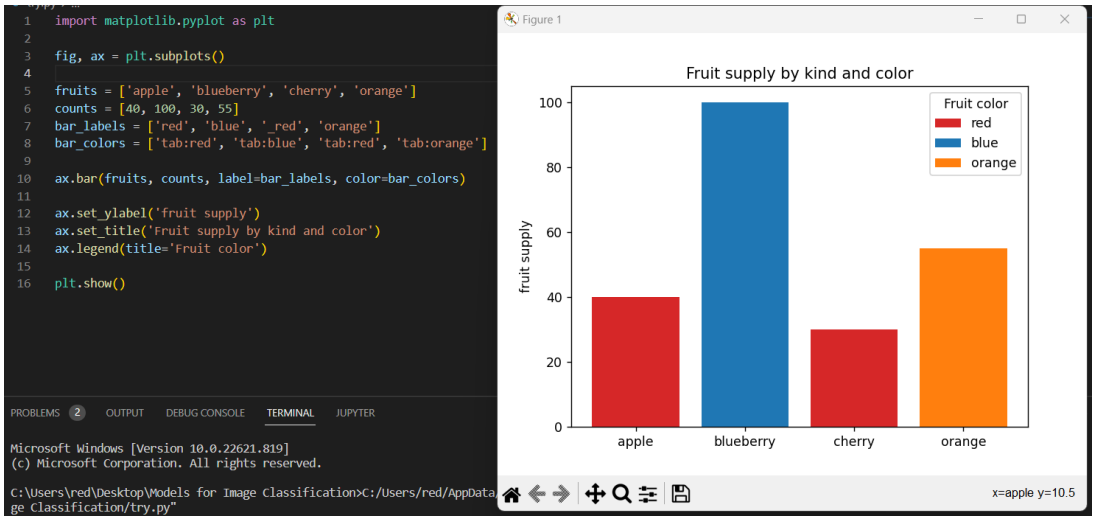
- **Qualitative Data** – categorical and non-countable (example: nationality, Intermediate).
  - **Nominal** (labels e.g., nationality)
  - **Ordinal** (ranking e.g., Intermediate)

Data Visualization

- is the presentation of data in a *graphical* or *pictorial* format. It can be in graph, chart, or other visual format that provides an easy and accessible way of understanding the presented data.

Python graphic libraries for visualizing data

- ***Plotting Libraries:***
  - **Matplotlib** - Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine, designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.



  - **Seaborn** – is a Python data visualization library based on matplotlib. It provides

a high-level interface for drawing attractive and informative statistical graphics. Seaborn is a library mostly used for statistical plotting in Python. It can be considered as an extension of another library called Matplotlib as it is built on top of that. At last, we can say that Seaborn is an extended version of matplotlib which tries to make a well-defined set of hard things easy.

- ***IDE***
  - Jupyter Notebook
- ***Other Libraries***
  - **Numpy** - NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors. NumPy is a NumFOCUS fiscally sponsored project.
  - **Pandas –** pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language. pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals. Its name is a play on the phrase "Python data analysis" itself. Wes McKinney started building what would become pandas at AQR Capital while he was a researcher there from 2007 to 2010.