Introduction to Data Science

- In a world of data space where organizations deal with petabytes and exabytes of data, the era of Big Data emerged, the essence of its storage also grew. It was a great challenge and concern for industries for the storage of data until 2010.
- Now when frameworks like Hadoop and others solved the problem of storage, the focus shifted to processing of data. Data Science plays a big role here. All those fancy Sci-fi movies you love to watch around can turn into reality by Data Science.
- Nowadays it’s growth has been increased in multiple ways and thus one should be ready for our future by learning what it is and how can we add value to it. Without any hunches, let’s dive into the world of Data Science.
- After touching to slightest idea, you might have ended up with many questions like What is Data Science? Why we need it? How can I be a Data Scientist?? etc? So let’s clear out ourselves from this baffle.

What is Data Science?

- Data Science is kinda blended with various tools, algorithms, and machine learning principles. Most simply, it involves obtaining meaningful information or insights from structured or unstructured data through a process of analyzing, programming and business skills.
- It is a field containing many elements like mathematics, statistics, computer science, etc. Those who are good at these respective fields with enough knowledge of the domain in which you are willing to work can call themselves as Data Scientist.
- It’s not an easy thing to do but not impossible too. You need to start from data, it’s visualization, programming, formulation, development, and deployment of your model. In the future, there will be great hype for data scientist jobs. Taking in that mind, be ready to prepare yourself to fit in this world.

How Data Science Works?

- Data science is not a one-step process such that you will get to learn it in a short time and call ourselves a Data Scientist. It’s passes from many stages and every element is important. One should always follow the proper steps to reach the ladder. Every step has its value and it counts in your model. Buckle up in your seats and get ready to learn about those steps.

  - **Problem Statement:** No work start without motivation, Data science is no exception though. It’s really important to declare or formulate your problem statement very clearly and precisely. Your whole model and it’s working depend on your statement. Many scientist considers this as the main and much important step of Date Science. So make sure what’s your problem statement and how well can it add value to business or any other organization.

- **Data Collection:** After defining the problem statement, the next obvious step is to go in search of data that you might require for your model. You must do good research, find all that you need. Data can be in any form i.e unstructured or structured. It might be in various forms like videos, spreadsheets, coded forms, etc. You must collect all these kinds of sources.
- **Data Cleaning:** As you have formulated your motive and also you did collect your data, the next step to do is cleaning. Yes, it is! Data cleaning is the most favorite thing for data scientists to do. Data cleaning is all about the removal of missing, redundant, unnecessary and duplicate data from your collection. There are various tools to do so with the help of programming in either R or Python. It's totally on you to choose one of them. Various scientist have their opinion on which to choose. When it comes to the statistical part, R is preferred over Python, as it has the privilege of more than 12,000 packages. While python is used as it is fast, easily accessible and we can perform the same things as we can in R with the help of various packages

- **Data Analysis and Exploration:** It’s one of the prime things in data science to do and time to get inner Holmes out. It’s about analyzing the structure of data, finding hidden patterns in them, studying behaviors, visualizing the effects of one variable over others and then concluding. We can explore the data with the help of various graphs formed with the help of libraries using any programming language. In R, GGplot is one of the most famous models while Matplotlib in Python.
- **Data Modelling:** Once you are done with your study that you have formed from data visualization, you must start building a hypothesis model such that it may yield you a good prediction in future. Here, you must choose a good algorithm that best fit to your model. There different kinds of algorithms from regression to classification, SVM(Support vector machines), Clustering, etc. Your model can be of a Machine Learning algorithm. You train your model with the train data and then test it with test data. There are various methods to do so. One of them is the K-fold method where you split your whole data into two parts, One is Train and the other is test data. On these bases, you train your model.

- **Optimization and Deployment:** You followed each and every step and hence build a model that you feel is the best fit. But how can you decide how well your model is performing? This where optimization comes. You test your data and find how well it is performing by checking its accuracy. In short, you check the efficiency of the data model and thus try to optimize it for better accurate prediction. Deployment deals with the launch of your model and let the people outside there to benefit from that. You can also obtain feedback from organizations and people to know their need and then to work more on your model.

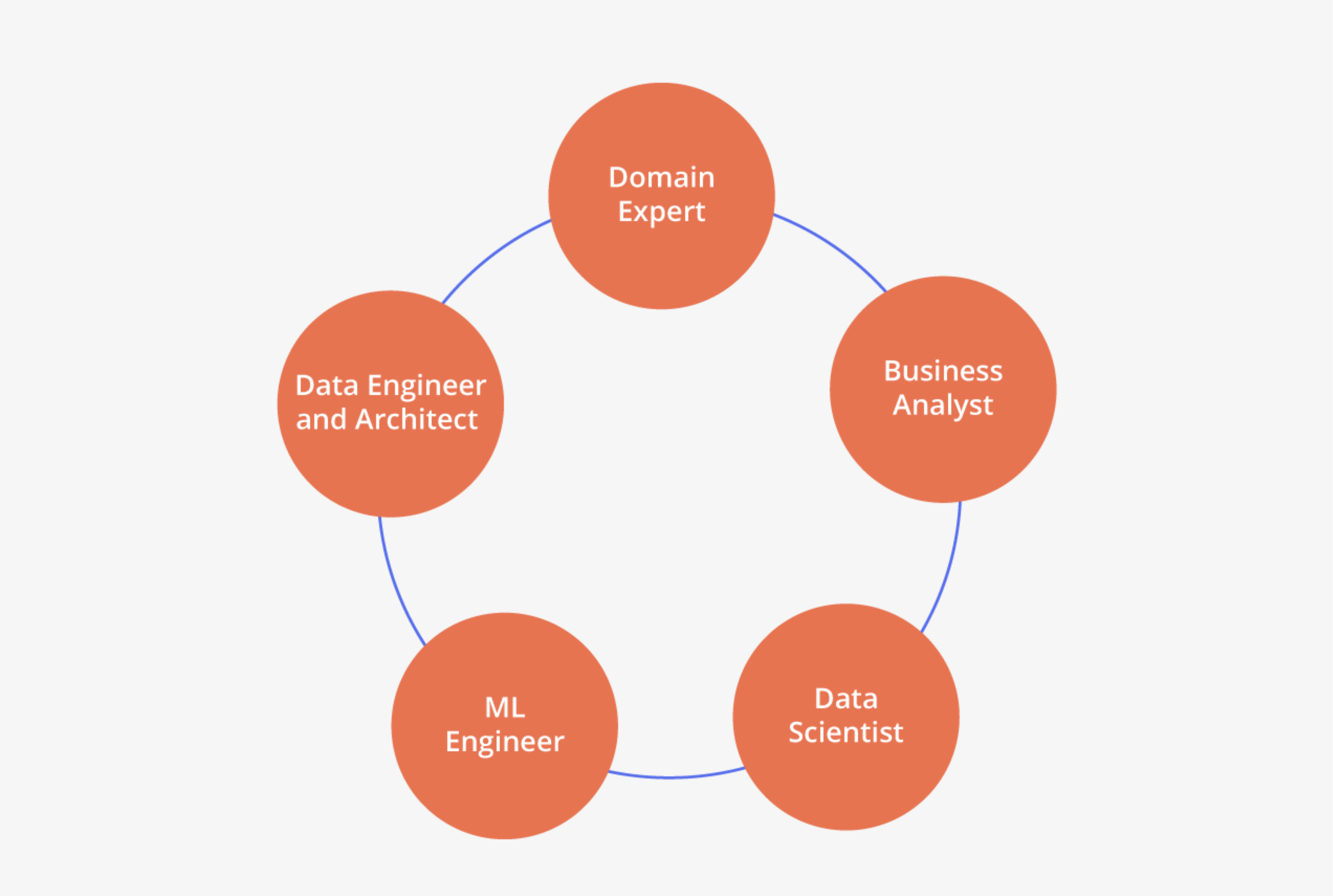
Data Science Life Cycle:

- Data Science is the amalgamation of two fields – Data and Science. Data is any real or imaginary thing and science is nothing but systematic study of world both physical and natural.
- So Data Science is nothing but systematic study of data and derivation of knowledge using testable methods to do predictions about the Universe.
- In simple words its applying science on data which may be of any size and from any source. Data has become a new oil that is driving businesses today. That's why understanding the data science project life cycle is crucial.
- As a Data Scientist or Machine Learning Engineer or as a Project Manager you must be aware of the important steps.

What is a Data Science Life Cycle?

- A data science lifecycle indicates the iterative steps taken to build, deliver and maintain any data science product.
- All data science projects are not built the same, so their life cycle varies as well. Still, we can picture a general lifecycle that includes some of the most common data science steps.
- A general data science lifecycle process includes the use of machine learning algorithms and statistical practices that result in better prediction models.
- Some of the most common data science steps involved in the entire process are data extraction, preparation, cleansing, modelling, and evaluation etc.
- The world of data science refers this general process as “Cross Industry Standard Process for Data Mining”.

Who Are Involved in The Projects?



Domain Expert:

The data science projects are applied in different domains or industries of real life like Banking, Healthcare, Petroleum industry etc. A domain expert is a person who has experience of working in the particular domain and knows in and out about the domain.

Business analyst:

A business analyst is required to understand the business needs in the domain identified. The person can guide in devising the right solution and timeline for the same.

Data analysis: Univariate, Bivariate and Multivariate data and its analysis

In the field of data, there is nothing more important than understanding the data that you are trying to analyze. In order to understand the data is it important to understand the purpose of the analysis because this will help you save time and dictate how to go about analyzing the data. There are a lots of different tools, techniques and methods that can be used to conduct your analysis. You could use software libraries, visualization tools and statistic testing methods. However, this article will be an Introduction to Univariate, Bivariate and Multivariate analysis.

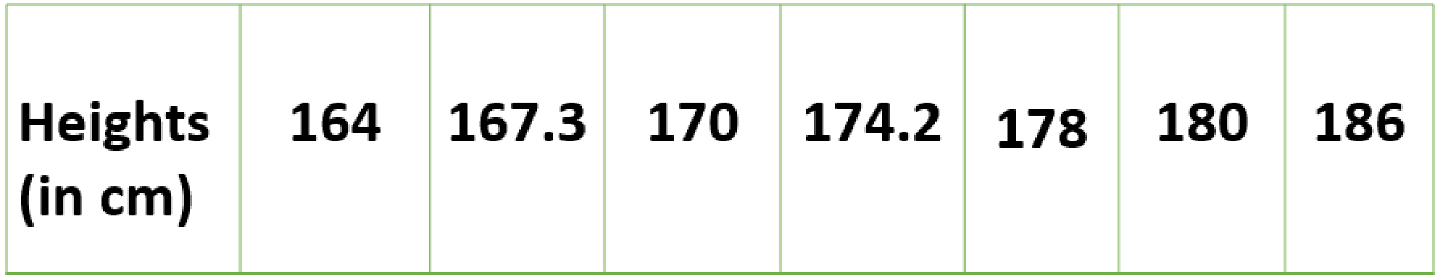
First we must understand the types of variables:

- Categorical variables — variables that have a finite number of categories or distinct groups. Examples: gender, method of payment, horoscope, etc.
- Numerical variables — variables that consist of numbers. There are two main numerical variables.
- Discrete variables — variables that can be counted within a finite time. Examples: the change in your pocket, number of students in a class, numerical grades, etc.
- Continuous variables — variables that are infinite in number often measured on a scale of sort. Examples: weight, height, temperature, date and time of a payment, etc.

However, depending on the type of variable, it can also be changed to another variable for ease of use. For example the date and time could be broken down to year, month and time could be categorized into AM and PM sales. A common technique for continuous variables is “binning” the variables into categories. For example, the weight of a person can be categorized into “below average”/“slim”, “average” and “above average”/“obese” by setting ranges.

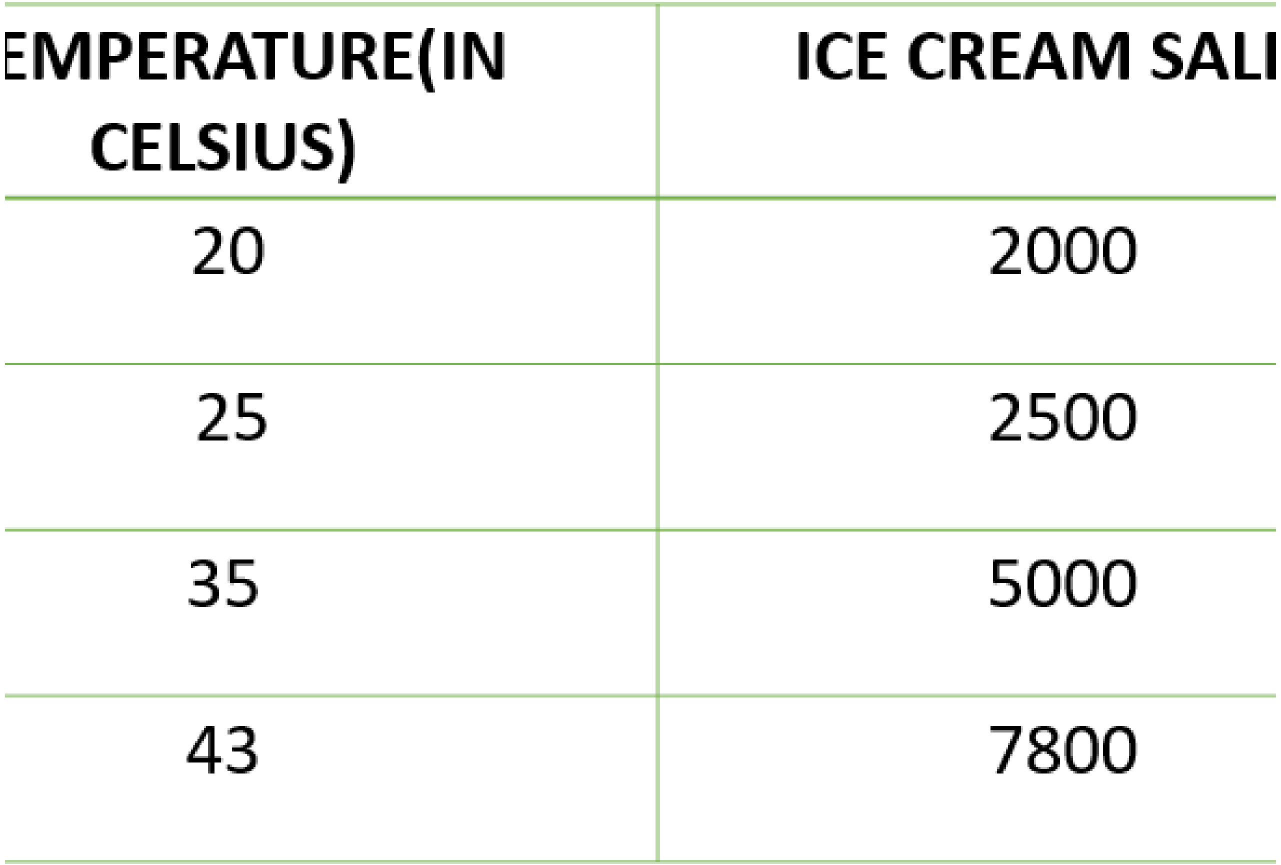
Univariate, Bivariate and Multivariate data and its analysis

1. Univariate data –
This type of data consists of **only one variable**. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a univariate data can be height.



Heights (in cm)	**164**	**167.3**	**170**	**174.2**	**178**	**180**	**186**

Bivariate Analysis



EMPERATURE(IN CELSIUS)	ICE CREAM SALI
20	2000
25	2500
35	5000
43	7800

- Bivariate analysis is where you are comparing two variables to study their relationships. These variables could be dependent or independent to each other. In Bivariate analysis is that there is always a Y-value for each X-value.
- The most common visual technique for bivariate analysis is a scatter plot, where one variable is on the x-axis and the other on the y-axis. In addition to the scatter plot, regression plot and correlation coefficient are also frequently used to study the relationship of the variables. For example, continuing with the iris dataset, you can compare “*sepal_length*” vs “*sepal_width*” or “*sepal_length*” vs the “*petal_length*”to see if there is a relationship.

Multivariate Analysis

- Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables. For three variables, you can create a 3-D model to study the relationship (also known as Trivariate Analysis). However, since we cannot visualize anything above the third dimension, we often rely on other softwares and techniques for us to be able to grasp the relationship in the data.
- In terms of visualization, Seaborn library in Python allows for pairplots where it generates one large chart of selected variables against one another in a series of scatter plots and histograms depending on the type of variable, also known as scatter plot matrix. Again, in the series to come, I will provide the code and examples of this.

Data types and their measurement scale

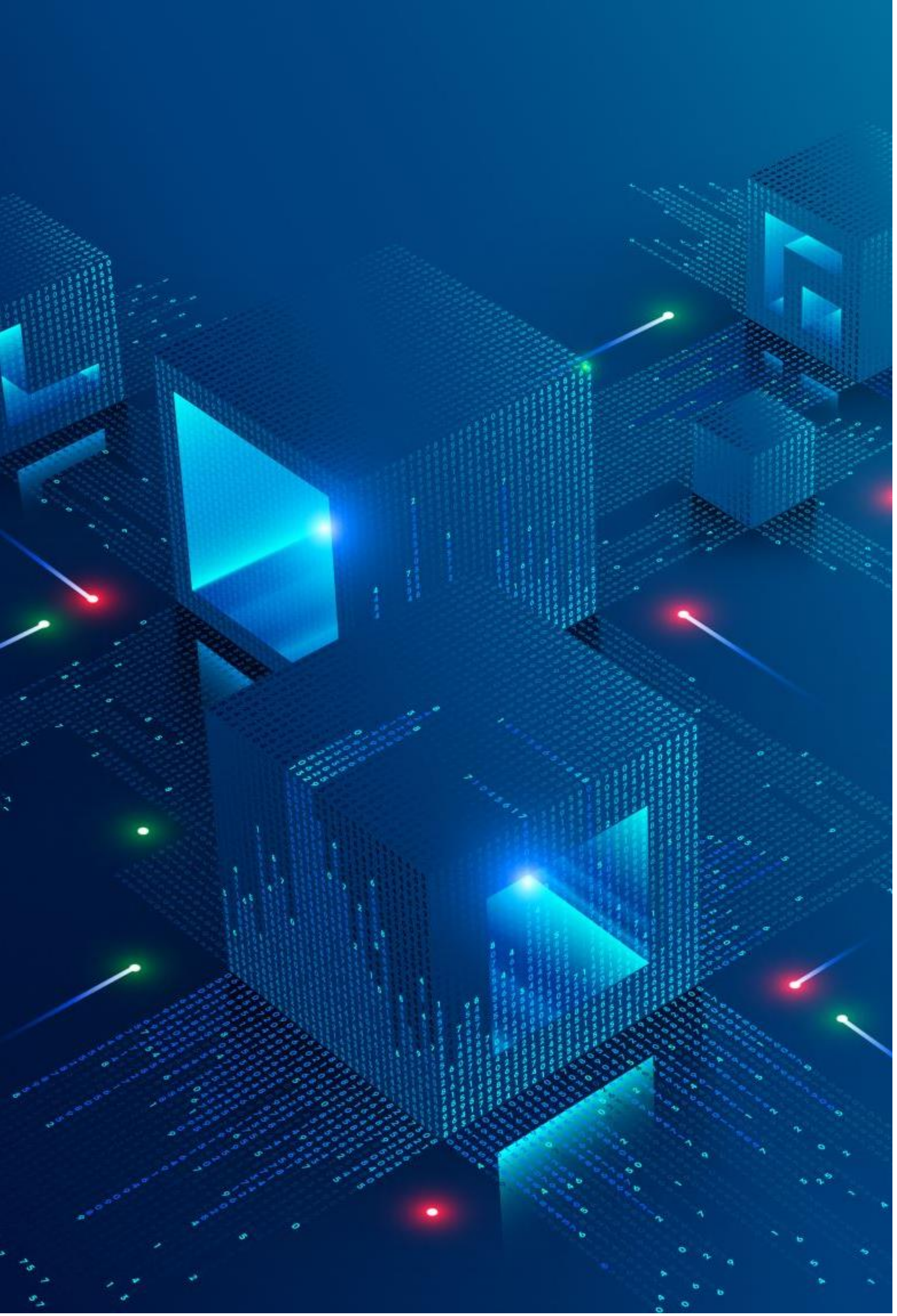
There are different ways to categorize data based on the way it has been collected or its structure.

Based on Data Collection: Data can be categorized into three types based on how data has been collected.

Cross-Sectional Data: Any data points/values captured on multiple variables over one specific time period is termed as cross-sectional data. Ex: attributes of the employee such as age, salary, level, team for the year 2019.

Time-Series Data: Any data points/values captured on a single variable over multiple periods is called time-series data. Ex: sales of smartphones on a monthly, quarterly, yearly basis.

Panel Data: A combination of both the cross-sectional and time-series data is known as Panel data. Ex: GDP of the various country over different periods.



Based on Structure. Another important way to classify data is based on their structure. It can be categorized into two types.

1. **Structured Data**: All the data points which have a specific structure and can be arranged in tabular form (also known as a matrix) with rows and columns are called structured data. Ex: Salary of employees arranged with employee id.
2. **Unstructured Data**: All the data points which are not arranged into any tabular format are unstructured data. Ex: Emails, videos, clickstream data, etc.

- ***70%*** of the available data is unstructured and while analyzing or building any analytics model one has to convert unstructured data to a structured one.
- Another problem which most of the beginners with data analytics domain face is, even the structured data is available what to do with it, how to use it, how it can be measured and how to infer insights from that.
- And for all these, Measurement scales becomes important. One must be aware that if the structured data is available how we can measure them and how those can be differentiated based on measurement.

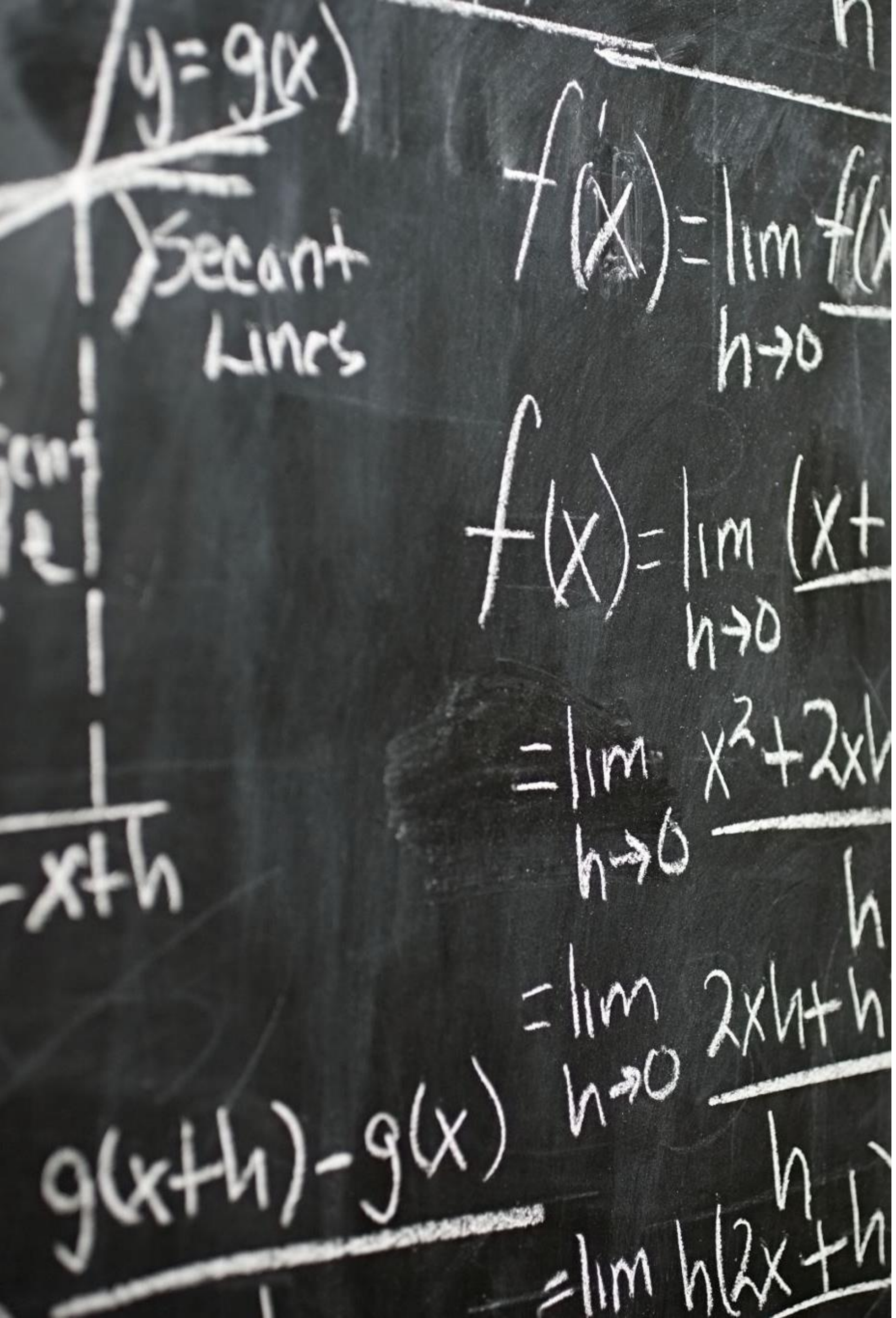
Data can be divided into four parts based on a measurement scale.

1) Nominal Scale

2) Ordinal Scale

3) Interval Scale

4) Ratio Scale



1. **Nominal Scale**: All the data points which are qualitative in nature falls in this category. These are also referred to as categorical variables. Ex: Marital Status (Single, married, etc.). No arithmetic operation (addition, subtraction, multiplication or division) can be performed on such variables.
2. **Ordinal Scale**: All the data points from the ordered set falls in this category. Ex: Ratings on a 1–5 scale (5 being highest and 1 being lowest). Here the order of the set is fixed, but no arithmetic operation can be performed such as we know, rating 4 is better than 2, but two 2 ratings cannot be equaled to rating 4.

1. **Interval Scale**: All the data points which have been taken from some fixed interval set. Ex: Temperature (in centigrade), IQ level. In such variables, addition or subtraction can be performed but division doesn’t make sense. As you can say Mumbai has 10 centigrade more than Bangalore, but you saying that Mumbai is twice hotter than Bangalore is not right, thus ratios don’t make sense here.
2. **Ratio Scale**: All the data points which are quantitative in nature falls in this category. Ex: Sales of a product, the salary of an employee, etc. Here all the arithmetic operations can be performed and comparison can be made as such that Ram earns twice of what Shyam earns, thus ratios make sense.

- Thus, by looking at the data, one can infer what kind of data is available like nominal, ordinal, etc. which eventually helps a data analyst/scientist, while building any analytics model for understanding different variables, doing exploratory data analysis, doing data imputation, and performing one-hot encoding.
- And not only it becomes important in predictive analytics, but it also helps in descriptive analytics. You can't do Exploratory data analysis if you don’t have information about the type of data. Once you identify the type of data then lot of univariate and bivariate analysis, visualization and calculation such as mean, mode, median, etc. can be performed to infer insight from data.

What is Statistics?

- Statistics is the science of collecting data and analyzing them to infer proportions (sample) that are representative of the population. In other words, statistics is interpreting data in order to make predictions for the population.

Branches of Statistics:

There are two branches of Statistics.

- DESCRIPTIVE STATISTICS : Descriptive Statistics is a statistics or a measure that describes the data.
- INFERENTIAL STATISTICS : Using a random sample of data taken from a population to describe and make inferences about the population is called Inferential Statistics.

Descriptive Statistics

- In Descriptive statistics, we are describing our data with the help of various representative methods like **by using charts, graphs, tables, excel files etc**. In descriptive statistics, we describe our data in some manner and present it in a meaningful way so that it can be easily understood.
- This means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

Commonly Used Measures

1.Measures of Central Tendency

2.Measures of Dispersion (or Variability)

Measures of Central Tendency

A Measure of Central Tendency is a one number summary of the data that typically describes the center of the data. These one number summary is of three types.

- **Mean :** Mean is defined as the ratio of the sum of all the observations in the data to the total number of observations. This is also known as Average. Thus mean is a number around which the entire data set is spread.

Commonly Used Measures

- **Median :** Median is the point which divides the entire data into two equal halves. One-half of the data is less than the median, and the other half is greater than the same. Median is calculated by first arranging the data in either ascending or descending order.
- If the number of observations are odd, median is given by the middle observation in the sorted form.
- If the number of observations are even, median is given by the mean of the two middle observation in the sorted form.
- An important point to note that the order of the data (ascending or descending) does not effect the median.

Commonly Used Measures

- **3. Mode :** Mode is the number which has the maximum frequency in the entire data set, or in other words,mode is the number that appears the maximum number of times. A data can have one or more than one mode.
- If there is only one number that appears maximum number of times, the data has one mode, and is called **Uni-modal**.
- If there are two numbers that appear maximum number of times, the data has two modes, and is called **Bi-modal**.
- If there are more than two numbers that appear maximum number of times, the data has more than two modes, and is called **Multi-modal**.

Example to compute the Measures of Central Tendency

Consider the following data points.

17, 16, 21, 18, 15, 17, 21, 19, 11, 23

Mean — Mean is calculated as

$$Mean = \frac{17 + 16 + 21 + 18 + 15 + 17 + 21 + 19 + 11 + 23}{10} = \frac{178}{10} = 17.8$$

Median

Median — To calculate Median, lets arrange the data in ascending order.

11, 15, 16, 17, 17, 18, 19, 21, 21, 23

Since the number of observations is even (10), median is given by the average of the two middle observations (5th and 6th here).

$$Median = \frac{5^{th}\ Obs + 6^{th}\ Obs}{2} = \frac{17 + 18}{2} = 17.5$$

Mode

Mode — Mode is given by the number that occurs maximum number of times. Here, 17 and 21 both occur twice. Hence, this is a Bimodal data and the modes are 17 and 21.

Note-

1. Since Median and Mode does not take all the data points for calculations, these are robust to outliers, i.e. these are not effected by outliers.
2. At the same time, Mean shifts towards the outlier as it considers all the data points. This means if the outlier is big, mean overestimates the data and if it is small, the data is underestimated.
3. If the distribution is symmetrical, Mean = Median = Mode. Normal distribution is an example.

Measures of Dispersion (or Variability)

Measures of Dispersion describes the spread of the data around the central value (or the Measures of Central Tendency)

1. **Absolute Deviation from Mean** — The Absolute Deviation from Mean, also called Mean Absolute Deviation (MAD), describe the variation in the data set, in sense that it tells the average absolute distance of each data point in the set. It is calculated as

$$Mean\ \ Absolute\ \ Deviation = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \overline{X}\right|$$

Variance

2. Variance — Variance measures how far are data points spread out from the mean. A high variance indicates that data points are spread widely and a small variance indicates that the data points are closer to the mean of the data set. It is calculated as

$$Variance = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$$

Standard Deviation

3. Standard Deviation — The square root of Variance is called the Standard Deviation. It is calculated as

$$Std\ Deviation = \sqrt{Variance} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}$$

Range

4. Range — Range is the difference between the Maximum value and the Minimum value in the data set. It is given as

$$Range \quad = Maximum \quad - Minimum$$

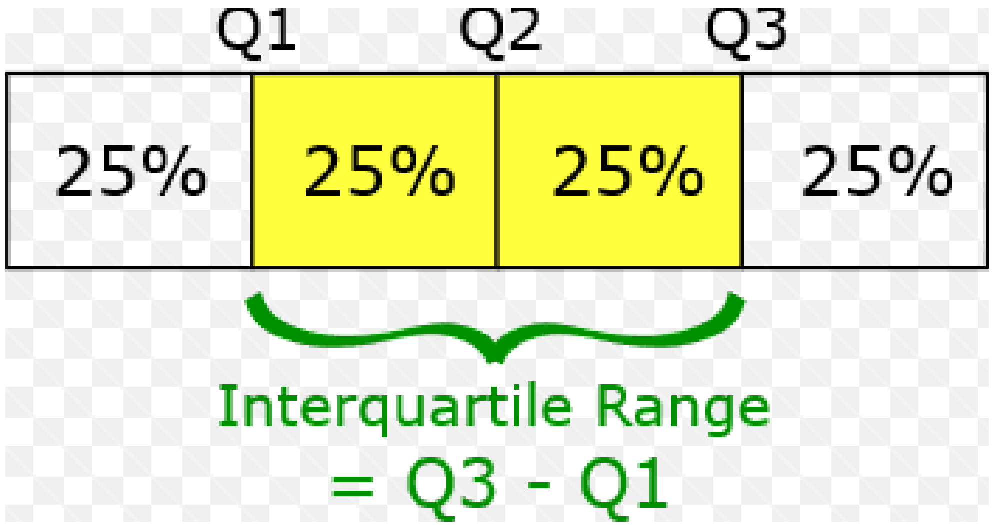
Quartiles

5. Quartiles — Quartiles are the points in the data set that divides the data set into four equal parts. Q1, Q2 and Q3 are the first, second and third quartile of the data set.

25% of the data points lie below Q1 and 75% lie above it.

50% of the data points lie below Q2 and 50% lie above it. Q2 is nothing but Median.

75% of the data points lie below Q3 and 25% lie above it.



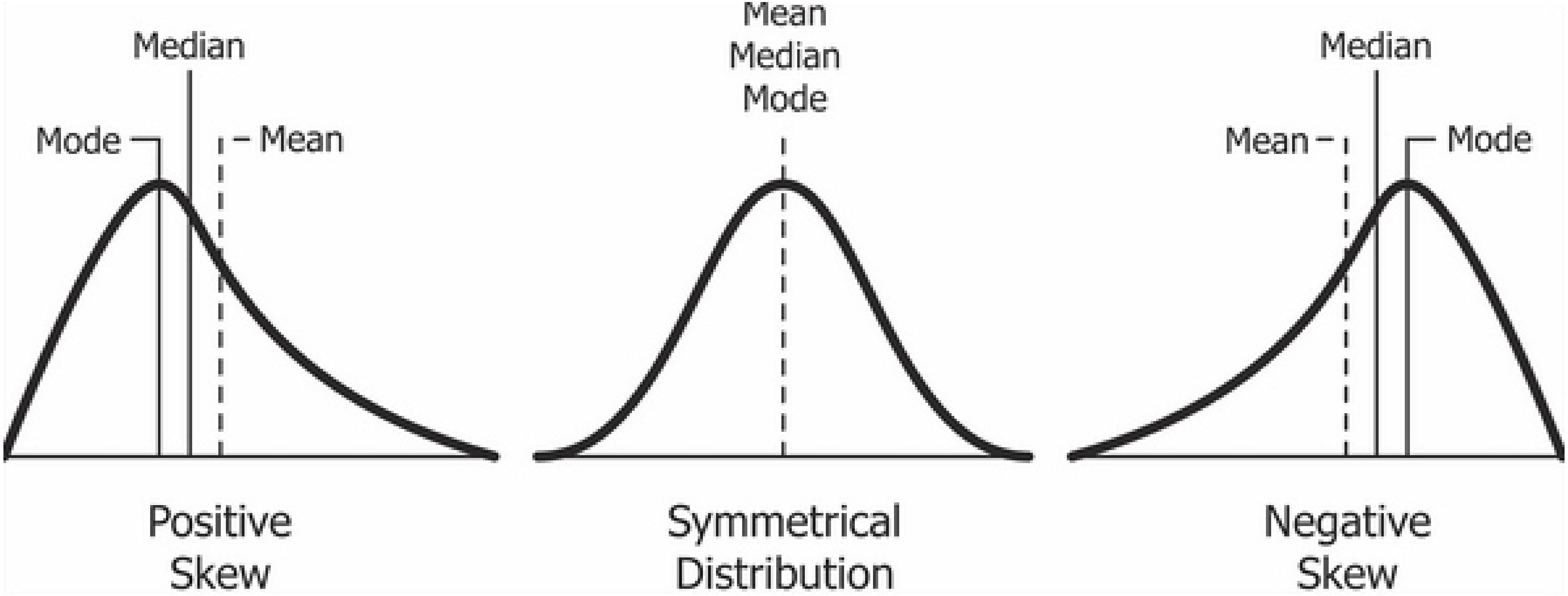
Skewness

6. Skewness — The measure of asymmetry in a probability distribution is defined by Skewness. It can either be positive, negative or undefined.

- Positive Skew — This is the case when the tail on the right side of the curve is bigger than that on the left side. For these distributions, mean is greater than the mode.
- Negative Skew — This is the case when the tail on the left side of the curve is bigger than that on the right side. For these distributions, mean is smaller than the mode.

$$\mathit{Skewness} = \frac{3\,(\mathit{Mean} - \mathit{Median})}{\mathit{Std}\ \mathit{Deviation}}$$

If the skewness is zero, the distribution is symmetrical. If it is negative, the distribution is Negatively Skewed and if it is positive, it is Positively Skewed.

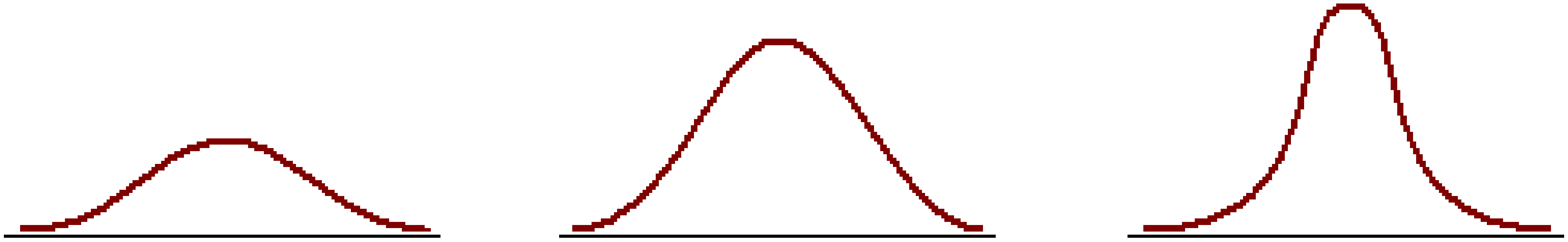


Kurtosis

7. Kurtosis — Kurtosis describes the whether the data is light tailed (lack of outliers) or heavy tailed (outliers present) when compared to a Normal distribution. There are three kinds of Kurtosis:

- Mesokurtic — This is the case when the kurtosis is zero, similar to the normal distributions.
- Leptokurtic — This is when the tail of the distribution is heavy (outlier present) and kurtosis is higher than that of the normal distribution.
- Platykurtic — This is when the tail of the distribution is light(no outlier) and kurtosis is lesser than that of the normal distribution.

Kurtosis



Platykurtic distribution
Low degree of peakedness
Kurtosis <0

Normal distribution
Mesokurtic distribution
Kurtosis = 0

Leptokurtic distribution
High degree of peakedness
Kurtosis > 0

Python Code

to find Mean in python

```python
import numpy as np

# Sample Data
arr = [5, 6, 11]
# Mean
mean = np.mean(arr)

print("Mean = ", mean)
```

to find Mode in python

```python
from scipy import stats

# sample Data
arr =[1, 2, 2, 3]

# Mode
mode = stats.mode(arr)
print("Mode = ", mode)
```

```
Mode = ModeResult(mode=array([2]),
count=array([2]))
```

Python Code

to find Median

```python
import numpy as np

# sample Data
arr =[1, 2, 3, 4]

# Median
median = np.median(arr)

print("Median = ", median)
```

to find Range

```python
import numpy as np

# Sample Data
arr = [1, 2, 3, 4, 5]

#Finding Max
Maximum = max(arr)
# Finding Min
Minimum = min(arr)

# Difference Of Max and Min
Range = Maximum-Minimum
print("Maximum = {}, Minimum = {} and Range = {}".format(
        Maximum, Minimum, Range))
```

```
Maximum = 5, Minimum = 1 and Range = 4
```

Understanding Python's role in data science

- As a new Data Scientist, you know that your path begins with programming languages you need to learn. Among all languages that you can select from Python is the most popular language for all Data Scientists. 7 reasons behind Python's popularity

1. Simplicity

Python is one of the easiest languages to start your journey. Also, its simplicity does not limit your functional possibilities.

What gives Python such flexibility? There are multiple factors:

- Python is a free and open-source language
- This is a high-level programming
- Python is interpreted
- It has an enormous community

Scalability

- Python is a programming language that scales very fast. Among all available languages, Python is a leader in scaling. That means that Python has more and more possibilities.
- Any problem can be decided easily with new updates that are coming. Saying that Python provides the best options for newbies because there are many ways to decide the same issue.
- Even if you have a team of non-Python programmers, who knows C++design patterns, Python will be better for them in terms of time needed to develop and verify code correctness.
- It happens fast because you don`t spend your time to find memory leaks, work for compilation or segmentation faults.

Libraries and Frameworks

- Due to its popularity, Python has hundreds of different libraries and frameworks which is a great addition to your development process. They save a lot of manual time and can easily replace the whole solution.
- As a Data Scientist, you will find that many of these libraries will be focused on Data Analytics and Machine Learning. Also, there is a huge support for Big Data. I suppose there should be a strong pro why you need to learn Python as your first language.

Libraries and Frameworks

Some of these libraries are given below:

- **Pandas**

It is great for data analysis and data handling. Pandas provides data manipulation control.

- **NumPy**

NumPy is a free library for numerical computing. It provides high-level math functions along with data manipulations.

- **SciPy**

This library is related to scientific and technical computing. SciPy can be used for data optimization and modification, algebra, special functions, etc.

4. Web Development

- To make your development process as easy as it is possible only, learn Python. There are a lot of Django and Flask libraries and frameworks that make your coding productive and speed up your work.
- If you compare PHP and Python, you can find that the same task can be created within a few hours of code via PHP. But with Python, it will take only a few minutes. Just take a look at Reddit website — it was created with Python.

Web Development

Here are Pythons Full Stack frameworks for web development:

- Django
- Pyramid
- Web2py
- TurboGears

And here are Pythons micro-frameworks for web development:

- Flask
- Bottle
- CherryPy
- Hug

Also, there is an alternative framework you might want to consider:

- Tornado

Huge Community

- As I have mentioned before, Python has a powerful community. You might think that it shouldn`t be one of the main reasons why you need to select Python. But the truth is vice versa.

Automation

Using Python automation frameworks like PYunit gives you a lot of advantages:

- No additional modules are required to install. They come with the box
- Even if you don`t have Python background you will find work with Unittest very comfortable. It is derivative and its working principle is similar to other xUnit frameworks.
- You can run singular experiments in a more straightforward way. You should simply indicate the names on the terminal. The output is compact too, making the structure adaptable with regards to executing test cases.
- The test reports are generated within milliseconds.

5 Python Frameworks For Test Automation:

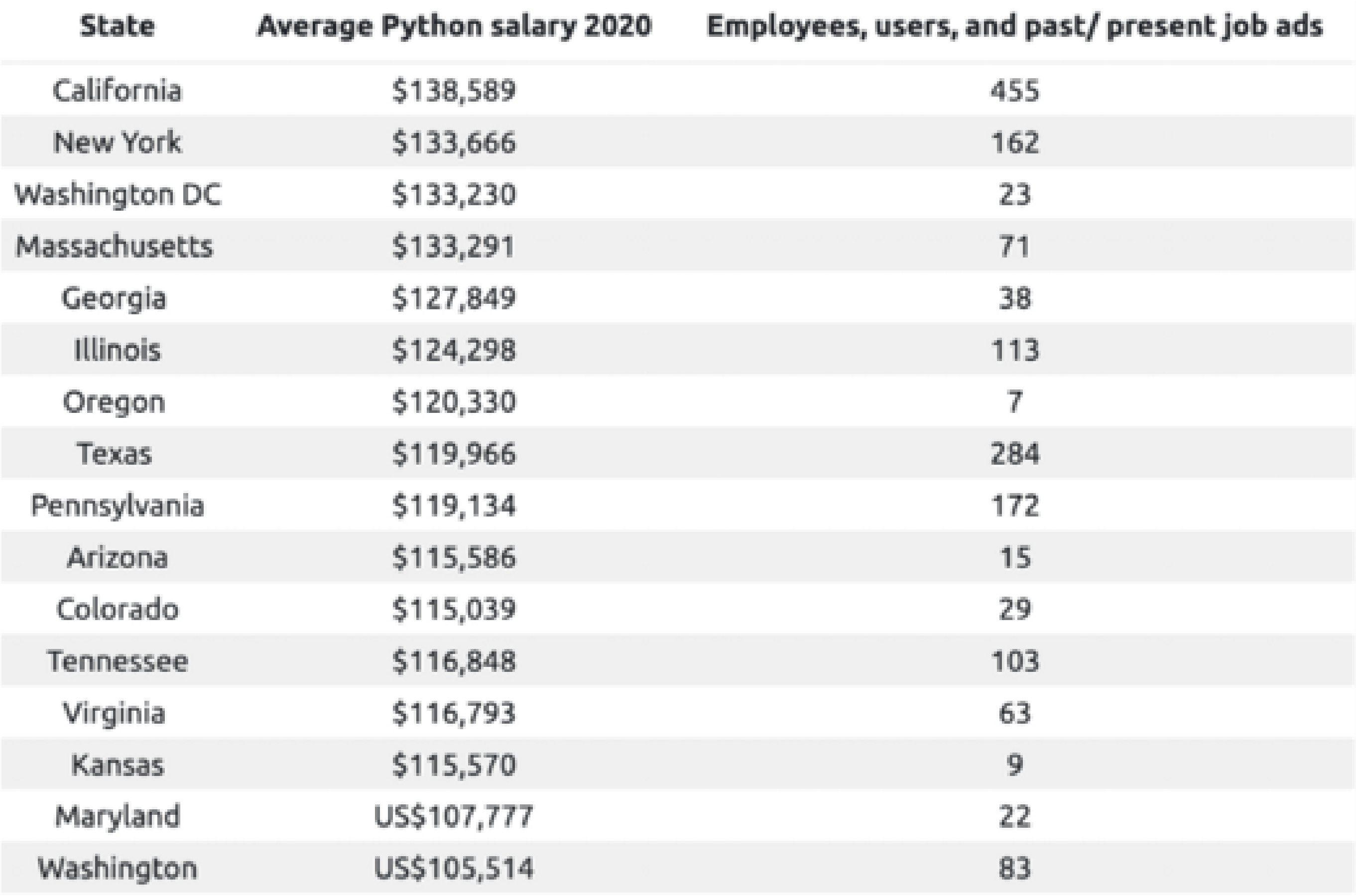
1. Robot Framework
2. UnitTest
3. Pytest
4. Behave
5. Lettuce

Jobs and Growth

Python is a unique language that has powerful growth and opens multiple career opportunities for Data Scientists. If you learn Python you can consider multiple additional jobs you might want to make the switch to in the future:

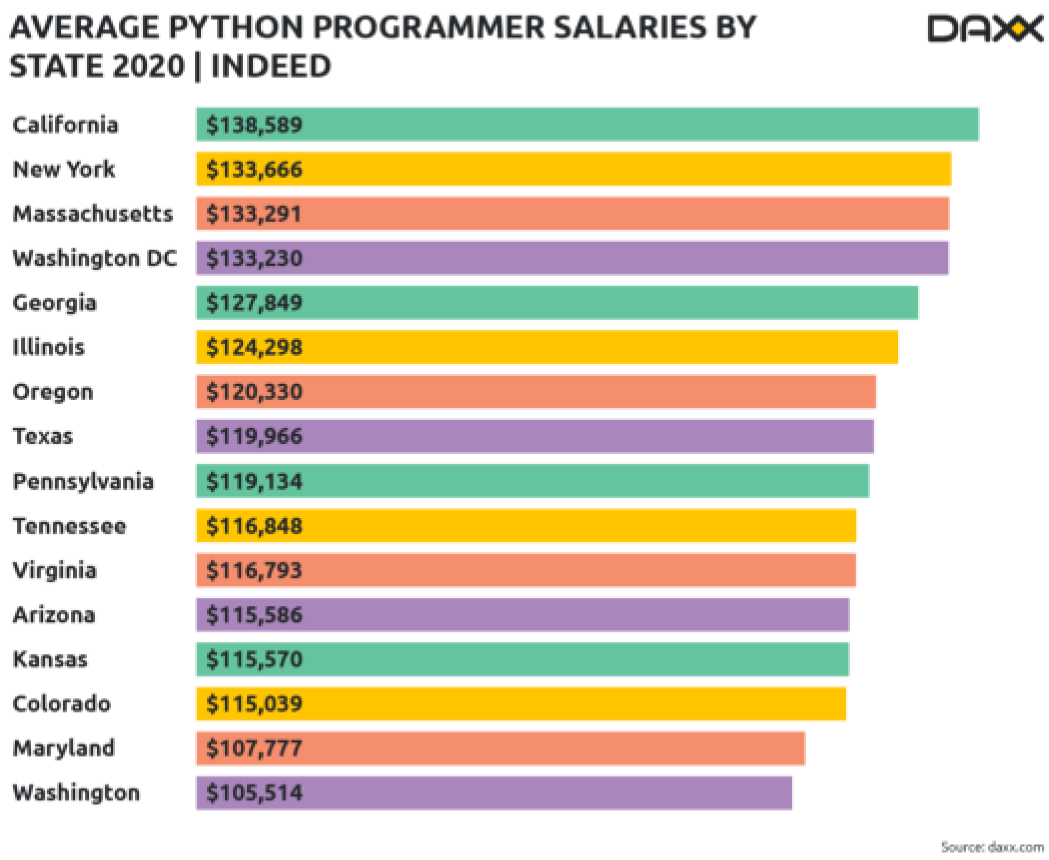
1. Python Developer
2. Product Manager
3. Educator
4. Financial Advisors
5. Data Journalist

Salary



State	Average Python salary 2020	Employees, users, and past/ present job ads
California	$138,589	455
New York	$133,666	162
Washington DC	$133,230	23
Massachusetts	$133,291	71
Georgia	$127,849	38
Illinois	$124,298	113
Oregon	$120,330	7
Texas	$119,966	284
Pennsylvania	$119,134	172
Arizona	$115,586	15
Colorado	$115,039	29
Tennessee	$116,848	103
Virginia	$116,793	63
Kansas	$115,570	9
Maryland	US$107,777	22
Washington	US$105,514	83

Salary



AVERAGE PYTHON PROGRAMMER SALARIES BY STATE 2020 | INDEED

Source: daxx.com

Conclusion

Python is a base for any Data Scientist. There are many reasons to select this powerful programming language, so it’s up to you which reason will be main. You should definitely consider Python due to its possibilities and ongoing improvement, which will help you to build amazing products and help businesses.