UNIT 2

Data Science and Descriptive Statistics

Prof. Keyur Prajapati

Understanding data, its type, and data science

Data types:

Structured data: Data that's stored, processed, and manipulated in a traditional relational database management system.

✓ Unstructured data: Data that's commonly generated from human activities and that doesn't fit into a structured database format.

✓ Semi-structured data: Data that doesn't fit into a structured database system, but is nonetheless structured by tags that are useful for creating a form of order and hierarchy in the data.

Continued …

You can use data insights to bring about the following types of changes:

- ✓ Optimize business systems and returns on investment (those crucial ROIs) for any measurable activity.
- ✓ Improve the effectiveness of sales and marketing initiatives —whether that be part of an organizational marketing campaign or simply a personal effort to secure better employment opportunities for yourself.
- ✓ Keep ahead of the pack on the very latest developments in every arena
- ✓ Keep communities safer.
- ✓ Help make the world a better place for those less fortunate

Continued …

- In the world of data space, the era of Big Data emerged when organizations are dealing with petabytes and exabytes of data. It became very tough for industries for the storage of data until 2010. Now when the popular frameworks like **Hadoop** and others solved the problem of storage, the focus is on processing the data. And here **Data Science plays a big role**. Nowadays the growth of data science has been increased in various ways and so on should be ready for the future by learning what data science is and how can we add value to it.

What is Data Science?

- So now the very first question arises is, "**What is Data Science?**" Data science means different things for different people, but at its gist, **data science is using data to answer questions.** This definition is a moderately broad definition, and that's because one must say data science is a moderately broad field!

Continued …

- **Pillar of data science**
- Domain Knowledge:
  - Most people thinking that domain knowledge is not important in data science, but it is essential. The foremost objective of data science is to extract useful insights from that data so that it can be profitable to the company's business. If you are not aware of the business side of the company, how to ask the right questions from the right people so that you can perceive the appropriate information you need to obtain the information you need. There are some visualization tools used on the business end like Tableau that help you display your valuable results or insights in a proper non-technical format such as graphs or pie charts that businesspeople can understand.
- Math Skills:
  - Linear Algebra: Multivariable Calculus & Optimization Technique: These three things are very important as they help us in understanding various machine learning algorithms that play an important role in Data Science.
  - Statistics & Probability Understanding: Statistics is very significant, as it is a part of Data analysis. Probability is also significant to statistics, and it is considered a prerequisite for mastering machine learning.
- Computer Science:
  - Programming Knowledge: One needs to have a good grasp of programming concepts such as Data structures and Algorithms. The programming languages used are Python, R, Java, Scala, C++. It is also useful in some places where performance is very important.
  - Relational Databases: One needs to know databases such as SQL or Oracle so that he/she can retrieve the necessary data from them whenever required.
  - Non-relational databases: There are many types of non-relational databases but mostly used types are Cassandra, HBase, MongoDB, CouchDB, Redis, Dynamo.
  - Machine Learning: It is one of the most vital parts of data science and the hottest subject of research among researchers so each year new advancements are made in this. One at least needs to understand basic algorithms of Supervised and Unsupervised Learning.
  - There are many libraries available for Python and R like matplotlib, Pandas, NumPy, Seaborn, Bokeh, dplyr, ggplot2, etc.
  - Distributed Computing: It is also one of the most important skills to handle a large amount of data because one can't process this much data on a single system. Tools that mostly used are Apache Hadoop and Spark. The two main parts of these tools are HDFS (Hadoop Distributed File System) that is used for collecting data over a distributed file system and another part is map-reduce which we manipulate the data. One can write map-reduce in programs in Java or Python. There are various other tools such as PIG, HIVE, etc.

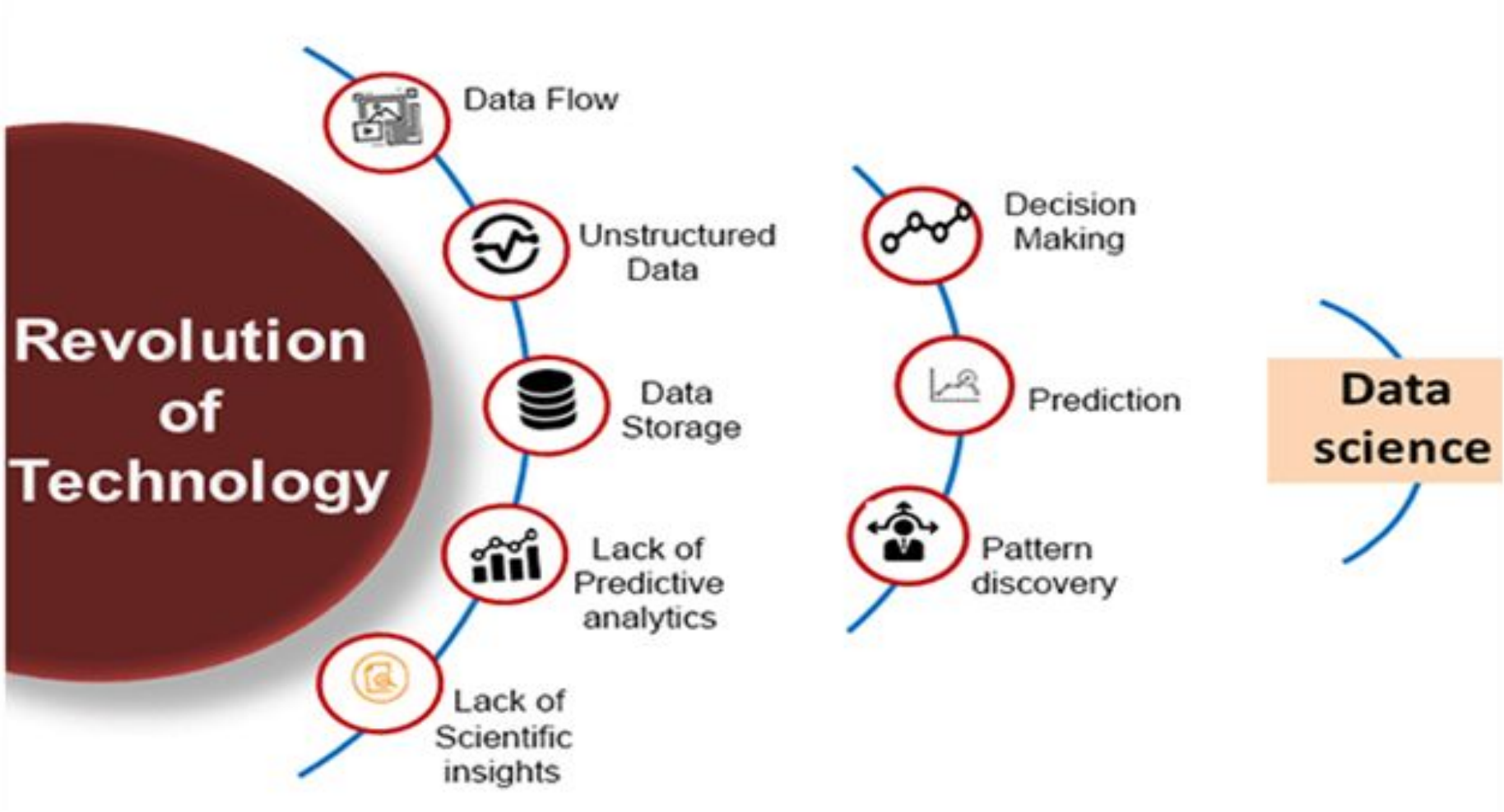
Differentiate…

Data Scientist	Data Analyst	Data Engineer
The focus will be on the futuristic display of data.	The focus of a data analyst is on optimization of scenarios, for example how an employee can enhance the company's product growth.	Data Engineers focus on optimization techniques and the construction of data in a conventional manner. The purpose of a data engineer is continuously advancing data consumption.
Data scientists present both supervised and unsupervised learning of data, say regression and classification of data, Neural networks, etc.	Data formation and cleaning of raw data, interpreting and visualization of data to perform the analysis and to perform the technical summary of data.	Frequently data engineers operate at the back end. Optimized machine learning algorithms were used for keeping data and making data to be prepared most accurately.
Skills required for Data Scientist are Python, R, SQL,	Skills required for Data Analyst are Python, R, SQL, SAS.	Skills required for Data Engineer are MapReduce, Hive, Pig Hadoop, techniques.

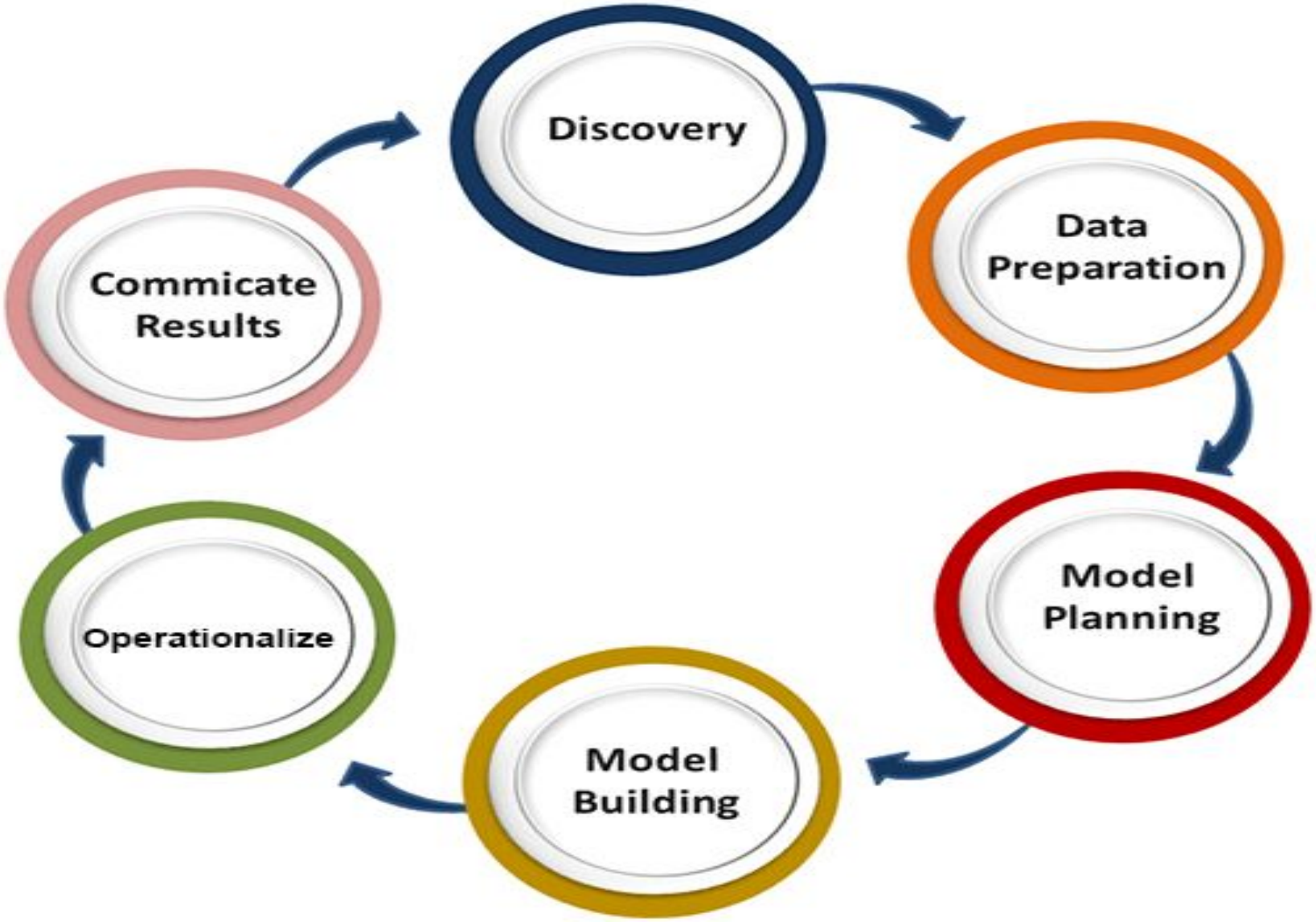
Why do we need data science?

One of the reasons for the acceleration of data science in recent years is the enormous volume of data currently available and being generated. Not only are huge amounts of data being collected about many aspects of the world and our lives, but we concurrently have the rise of inexpensive computing. This has formed the perfect storm in which we have rich data and the tools to analyze it. Advancing computer memory capacities, more enhanced software, more competent processors, and now, more numerous data scientists with the skills to put this to use and solve questions using the data!

Need for Data Science



Data Science Life Cycle



Continued …

•**1. Discovery:** The first phase is discovery, which involves asking the right questions. When you start any data science project, you need to determine what are the basic requirements, priorities, and project budget. In this phase, we need to determine all the requirements of the project such as the number of people, technology, time, data, an end goal, and then we can frame the business problem on first hypothesis level.

•**2. Data preparation:** Data preparation is also known as Data Munging. In this phase, we need to perform the following tasks:

- Data cleaning
- Data Reduction
- Data integration
- Data transformation,

•After performing all the above tasks, we can easily use this data for our further processes.

Continued …

•**3. Model Planning:** In this phase, we need to determine the various methods and techniques to establish the relation between input variables. We will apply Exploratory data analytics(EDA) by using various statistical formula and visualization tools to understand the relations between variable and to see what data can inform us. Common tools used for model planning are:

- SQL Analysis Services
- R
- SAS
- Python

•**4. Model-building:** In this phase, the process of model building starts. We will create datasets for training and testing purpose. We will apply different techniques such as association, classification, and clustering, to build the model.

Following are some common Model building tools:

- SAS Enterprise Miner
- WEKA
- SPCS Modeler
- MATLAB

•**5. Operationalize:** In this phase, we will deliver the final reports of the project, along with briefings, code, and technical documents. This phase provides you a clear overview of complete project performance and other components on a small scale before the full deployment.

•**6. Communicate results:** In this phase, we will check if we reach the goal, which we have set on the initial phase. We will communicate the findings and final result with the business team.

Applications of Data Science

- **Image recognition and speech recognition:**
  Data science is currently using for Image and speech recognition. When you upload an image on Facebook and start getting the suggestion to tag to your friends. This automatic tagging suggestion uses image recognition algorithm, which is part of data science. When you say something using, "Ok Google, Siri, Cortana", etc., and these devices respond as per voice control, so this is possible with speech recognition algorithm.
- **Gaming world:**
  In the gaming world, the use of Machine learning algorithms is increasing day by day. EA Sports, Sony, Nintendo, are widely using data science for enhancing user experience.
- **Internet search:**
  When we want to search for something on the internet, then we use different types of search engines such as Google, Yahoo, Bing, Ask, etc. All these search engines use the data science technology to make the search experience better, and you can get a search result with a fraction of seconds.
- **Transport:**
  Transport industries also using data science technology to create self-driving cars. With self-driving cars, it will be easy to reduce the number of road accidents.
- **Healthcare:**
  In the healthcare sector, data science is providing lots of benefits. Data science is being used for tumor detection, drug discovery, medical image analysis, virtual medical bots, etc.
- **Recommendation systems:**
  Most of the companies, such as Amazon, Netflix, Google Play, etc., are using data science technology for making a better user experience with personalized recommendations. Such as, when you search for something on Amazon, and you started getting suggestions for similar products, so this is because of data science technology.
- **Risk detection:**
  Finance industries always had an issue of fraud and risk of losses, but with the help of data science, this can be rescued.
  Most of the finance companies are looking for the data scientist to avoid risk and any type of losses with an increase in customer satisfaction.

Data analysis (Univariate, bivariate and Multivariate)

- **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

Continued …

2. Bivariate data –

- This type of data involves **two different variables**. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables. Example of bivariate data can be temperature and ice cream sales in summer season.

TEMPERATURE(IN CELSIUS)	ICE CREAM SALES
20	2000
25	2500
35	5000
43	7800

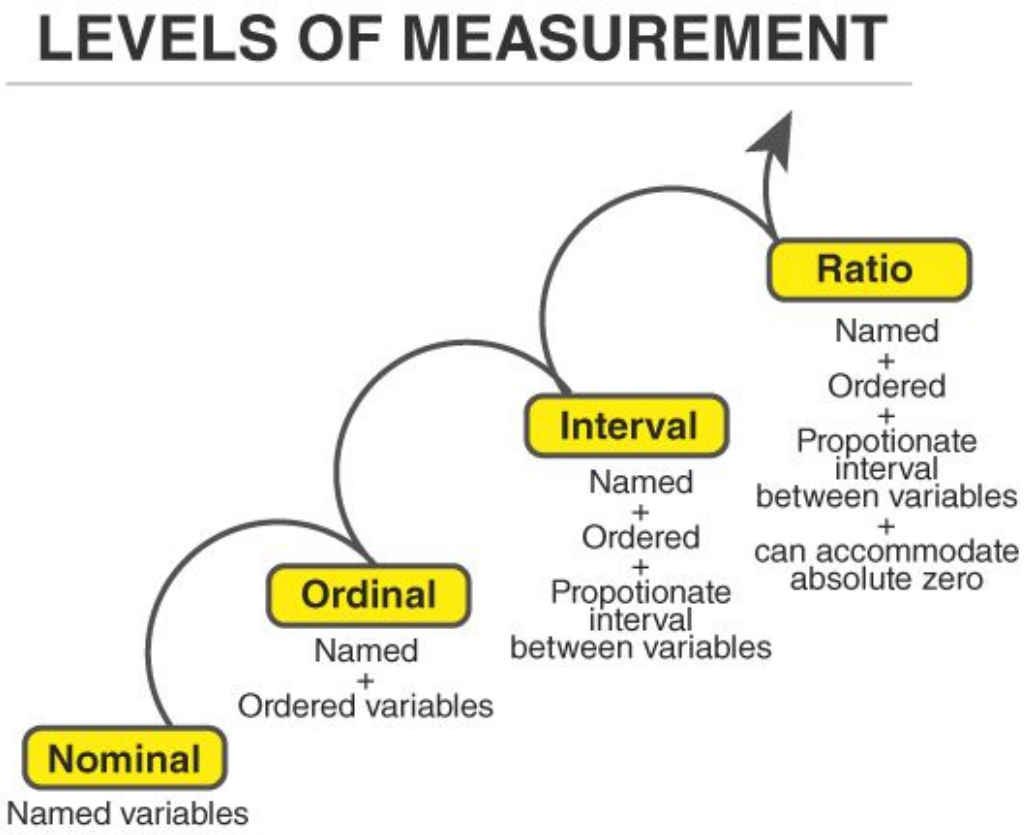
Continued …

- **3. Multivariate data –**

When the data involves **three or more variables**, it is categorized under multivariate. Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

- It is like bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved. Some of the techniques are regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA).

Data measurement scale



Data descriptive statistics (Measures of central tendency, dispersion/variation, measure of location, Shape, and symmetry)

•Descriptive Statistics is the building block of data science. Advanced analytics is often incomplete without analyzing descriptive statistics of the key metrics. In simple terms, descriptive statistics can be defined as the measures that summarize a given data, and these measures can be broken down further into the measures of central tendency and the measures of dispersion.

- Measures of central tendency include mean, median, and the mode, while the measures of variability include standard deviation, variance, and the interquartile range. In this guide, you will learn how to compute these measures of descriptive statistics and use them to interpret the data

Data

•In this guide, we will be using fictitious data of loan applicants containing 600 observations and 10 variables, as described below:

1. Marital status: Whether the applicant is married ("Yes") or not ("No").
2. Dependents: Number of dependents of the applicant.
3. Is graduate: Whether the applicant is a graduate ("Yes") or not ("No").
4. Income: Annual Income of the applicant (in USD).
5. Loan amount: Loan amount (in USD) for which the application was submitted.
6. Term months: Tenure of the loan (in months).
7. Credit score: Whether the applicant's credit score was good ("Satisfactory") or not ("Not satisfactory").
8. Age: The applicant’s age in years.
9. Sex: Whether the applicant is female (F) or male (M).
10. approval status: Whether the loan application was approved ("Yes") or not ("No").

Continued …

- Let's start by loading the required libraries and the data.

```
import pandas as pd
import numpy as np
import statistics as st

# Load the data
df = pd.read_csv("data_desc.csv")
print(df.shape)
print(df.info())
```

```
(600, 10)

 <class 'pandas.core.frame.DataFrame'>
 RangeIndex: 600 entries, 0 to 599
 Data columns (total 10 columns):
 Marital_status      600 non-null object
 Dependents          600 non-null int64
 Is_graduate         600 non-null object
 Income              600 non-null int64
 Loan_amount         600 non-null int64
 Term_months         600 non-null int64
 Credit_score        600 non-null object
 approval_status     600 non-null object
 Age                 600 non-null int64
 Sex                 600 non-null object
 dtypes: int64(5), object(5)
 memory usage: 47.0+ KB
 None
```

Five of the variables are categorical (labelled as 'object') while the remaining five are numerical (labelled as 'int').

Measures of Central Tendency

•Measures of central tendency describe the center of the data, and are often represented by the mean, the median, and the mode.

•**Mean**

•Mean represents the arithmetic average of the data. The line of code below prints the mean of the numerical variables in the data. From the output, we can infer that the average age of the applicant is 49 years, the average annual income is USD 705,541, and the average tenure of loans is 183 months. The command **df.mean(axis = 0)** will also give the same output.

```
df.mean()
```

Output

```
Dependents          0.748333
Income         705541.333333
Loan_amount    323793.666667
Term_months       183.350000
Age                49.450000
dtype: float64
```

Continued ..

It is also possible to calculate the mean of a particular variable in a data, as shown below, where we calculate the mean of the variables 'Age' and 'Income'.

```
print(df.loc[:,'Age'].mean())
print(df.loc[:,'Income'].mean())
```

Output:

```
49.45
705541.33
```

In the previous sections, we computed the column-wise mean. It is also possible to calculate the mean of the rows by specifying the **(axis = 1)** argument. The code below calculates the mean of the first five rows.

```
df.mean(axis = 1)[0:5]
```

Output:

```
0     70096.0
1    161274.0
2    125113.4
3    119853.8
4    120653.8
dtype: float64
```

Continued…

- **Median**

- In simple terms, median represents the 50th percentile, or the middle value of the data, that separates the distribution into two halves. The line of code below prints the median of the numerical variables in the data. The command **df.median(axis = 0)** will also give the same output.

```
df.median()
```

Output:

```
Dependents          0.0
Income         508350.0
Loan_amount     76000.0
Term_months       192.0
Age                51.0
dtype: float64
```

Continued …

- **Mode**

- Mode represents the most frequent value of a variable in the data. This is the only central tendency measure that can be used with categorical variables, unlike the mean and the median which can be used only with quantitative data.

- The line of code below prints the mode of all the variables in the data. The **.mode()** function returns the most common value or most repeated value of a variable. The command **df.mode(axis = 0)** will also give the same output.

```
df.mode()
```

Output:

```
|       | Marital_status       | Dependents    | Is_graduate   | Income        |
Loan_amount     | Term_months   | Credit_score  | approval_status       | Age   |
Sex     |
|---    |----------------      |------------   |-------------  |--------
        |-------------  |-------------  |--------------|-----------------
        |-----  |-----  |
| 0     | Yes                  | 0             | Yes           | 333300        |
70000           | 192.0         | Satisfactory  | Yes                   | 55    |
M       |
```

Measures of Dispersion

- **Standard Deviation**

Standard deviation is a measure that is used to quantify the amount of variation of a set of data values from its mean. A low standard deviation for a variable indicates that the data points tend to be close to its mean, and vice versa. The line of code below prints the standard deviation of all the numerical variables in the data.

```
df.std()
```

Output:

```
Dependents          1.026362
Income         711421.814154
Loan_amount    724293.480782
Term_months        31.933949
Age                14.728511
dtype: float64
```

Measures of Shape

- **What are the shapes of a dataset?**

  A distribution of data item values may be symmetrical or asymmetrical. Two common examples of symmetry and asymmetry are the 'normal distribution' and the 'skewed distribution'.

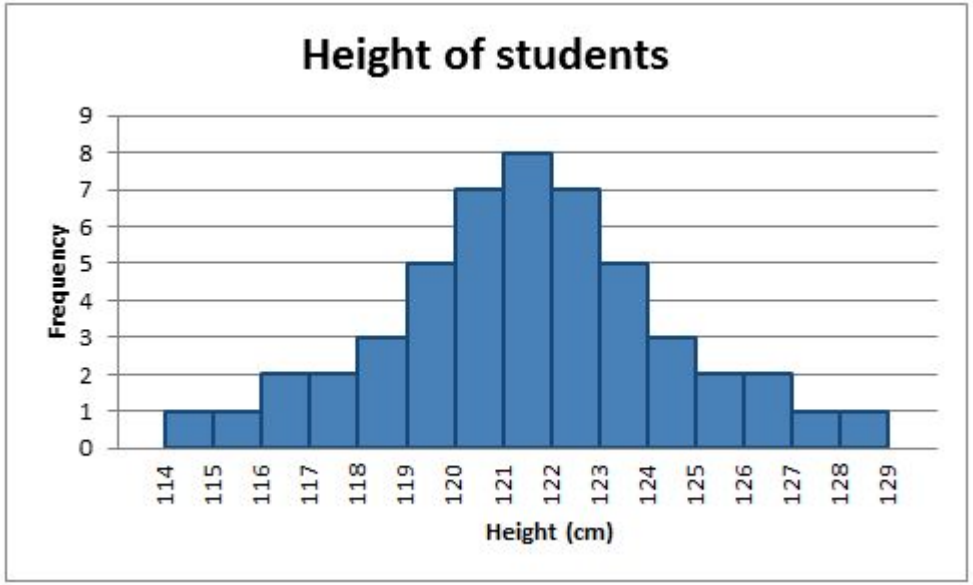
  In a **symmetrical distribution** the two sides of the distribution are a mirror image of each other.

  A normal distribution is a true symmetric distribution of observed values.

  When a histogram is constructed on values that are normally distributed, the shape of columns form a symmetrical bell shape. This is why this distribution is also known as a 'normal curve' or 'bell curve'.

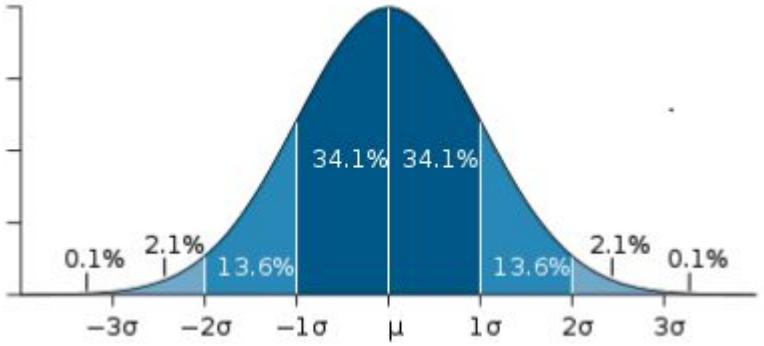
  The following graph is an example of a normal distribution:

Continue


Height of students
Frequency
0
1
2
3
4
5
6
7
8
9
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
Height (cm)

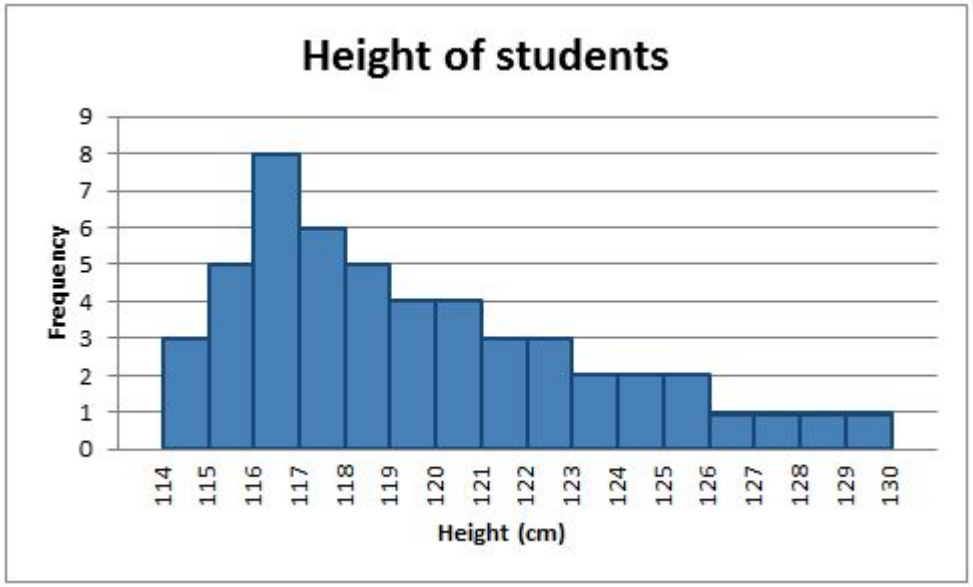
Continued …

- If represented as a 'normal curve' (or bell curve) the graph would take the following shape (where μ = mean, and σ = standard deviation):



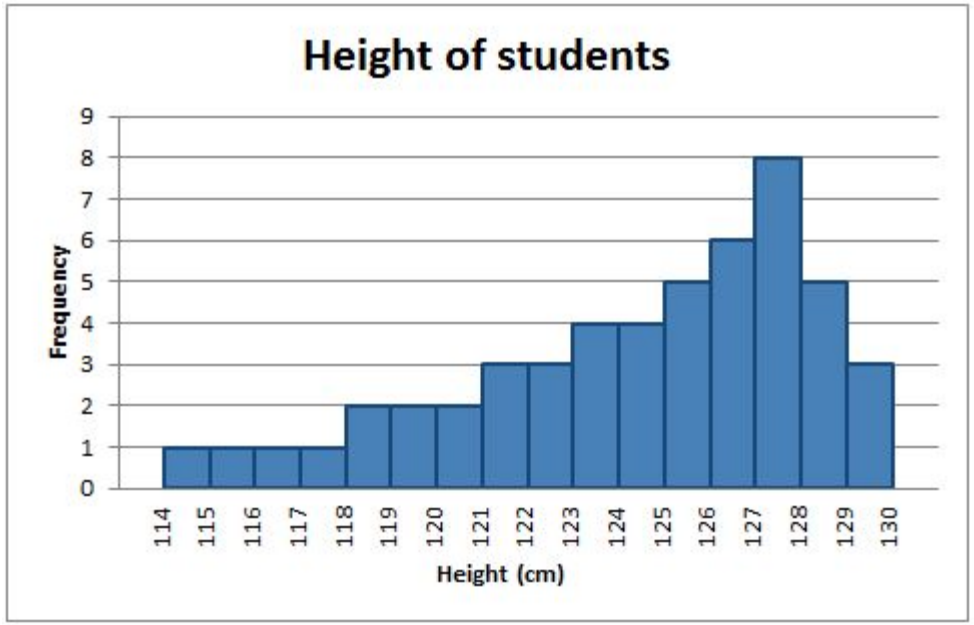
Continued …

- A distribution is said to be **positively skewed** when the tail on the right side of the histogram is longer than the left side. Most of the values tend to cluster toward the left side of the x-axis (i.e. the smaller values) with increasingly fewer values at the right side of the x-axis (i.e. the larger values).



Continued …

- A distribution is said to be **negatively skewed** when the tail on the left side of the histogram is longer than the right side. Most of the values tend to cluster toward the right side of the x-axis (i.e. the larger values), with increasingly less values on the left side of the x-axis (i.e. the smaller values).



Calculations (MEAN, VARIANCE, STANDARD DEVIATION)

Mean ($\bar{x} = \Sigma fx/\Sigma f$)

= *(600 + 470 + 170 + 430 + 300) / **5***

= *1970 / **5***

= 394

Variance (σ^2)

$$= \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

= $(206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2) / \mathbf{5}$

= *42436 + 5776 + 50176 + 1296 + 8836 / **5***

= *108520 / **5***

= 21704

Standard Deviation (σ) ($\sqrt{}$ Variance)

= $\sqrt{21704}$

= 147.32...

= **147**

Range and Interquartile Range

- Range: Min to Max (e.g: 43, 12, 3, 79,54,120,67,93,68,35)

Range = 3 to 120

- Interquartile Range

1.51, 1.53. 1.55, 1.55, 1.79. 1.81, 2.10, 2.15, 2.18,
2.22, 2.35, 2.37, 2.40, 2.40, 2.45, 2.78. 2.81, 2.85

IR(Interquartile Range is : 1.79 to 2.40)

Understanding Python's Role in Data Science

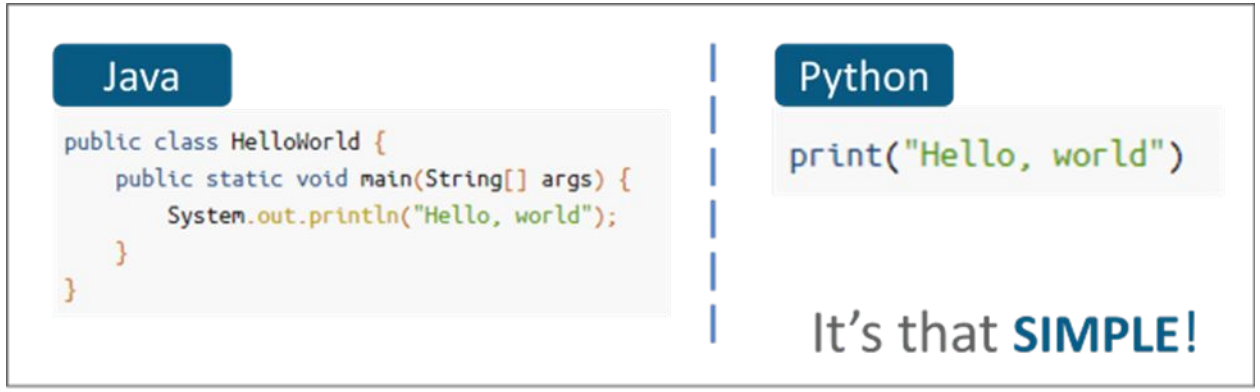
•Simplicity

•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

• In addition, Python is fast in writing. Just compare these 2 examples written in Java and Python:



Continued …

•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.

•Python flexibility is super useful for any problem in-app development

•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.

Continued …

- 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.

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.

Continued …

•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.

•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
- Tornado

Continued …

•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.

•If you don`t get support from other specialists, your learning path can be difficult. That`s why you should know that this won`t happen with your Python learning journey.

•Here is a list of some Python communities:

•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.

THANK YOU