APSSDC

Andhra Pradesh State Skill Development Corporation

Data Science Using Python

ANACONDA

NumPy

pandas

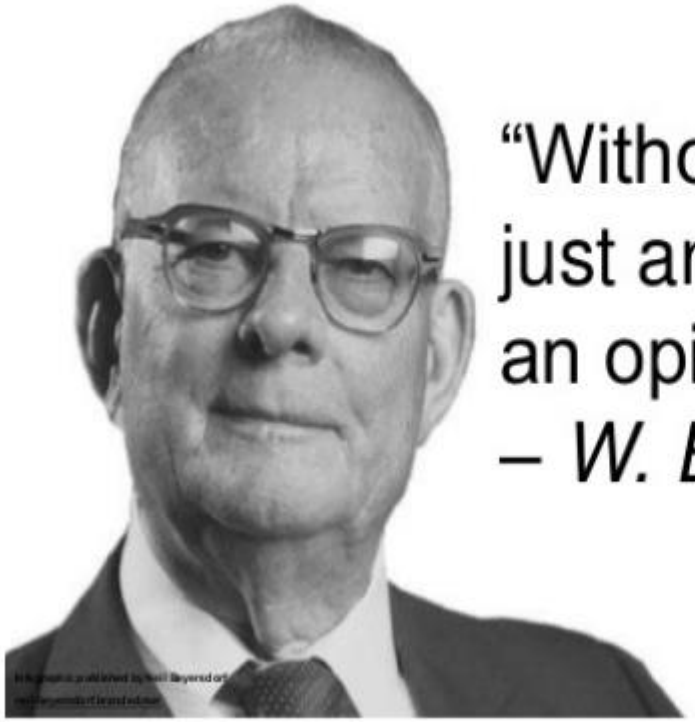
jupyter

matplotlib

scikit learn

What is Data?

Data are facts and statistics collected together for reference or analysis or Prediction.



Interesting insights

Bombardier showcased its C Series jetliner that carries Pratt & Whitney's Geared Turbo Fan (GTF) engine, which is fitted with 5,000 sensors that generate up to 10 GB of data per second. A single twin-engine aircraft with an average 12-hr. flight-time can produce up to 844 TB of data.

Saudi Aramco laid 650km of new pipelines across a mountain range of red sand dunes. How do they monitor that?

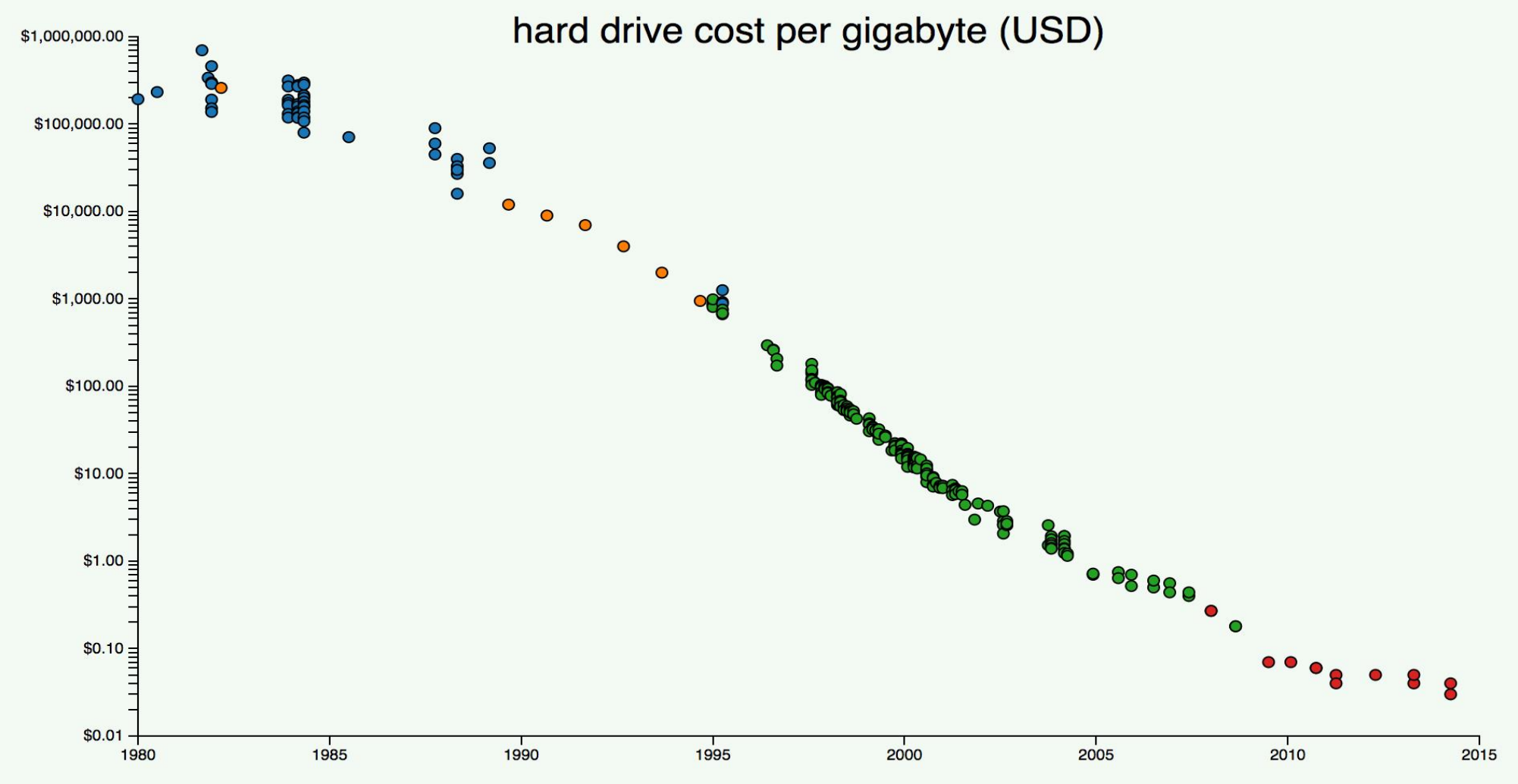
Using 100,000 sensors and data points on wells, pipelines, plants and terminals, it directs every drop of oil and cubic foot of gas that comes out of the kingdom

One study predicts that by 2020, 1.7 MB of data will be created every second for every person on earth.

The average number of AI projects for a business is expected to increase to 35 by 2022 from four this year, according to a Gartner Inc. survey of about 100 organizations of various sizes, many of them with annual revenue of $1 billion to $3 billion. The research and advisory firm also said the number of its clients requesting help in dealing with AI suppliers grew 57% between 2017 and 2018.

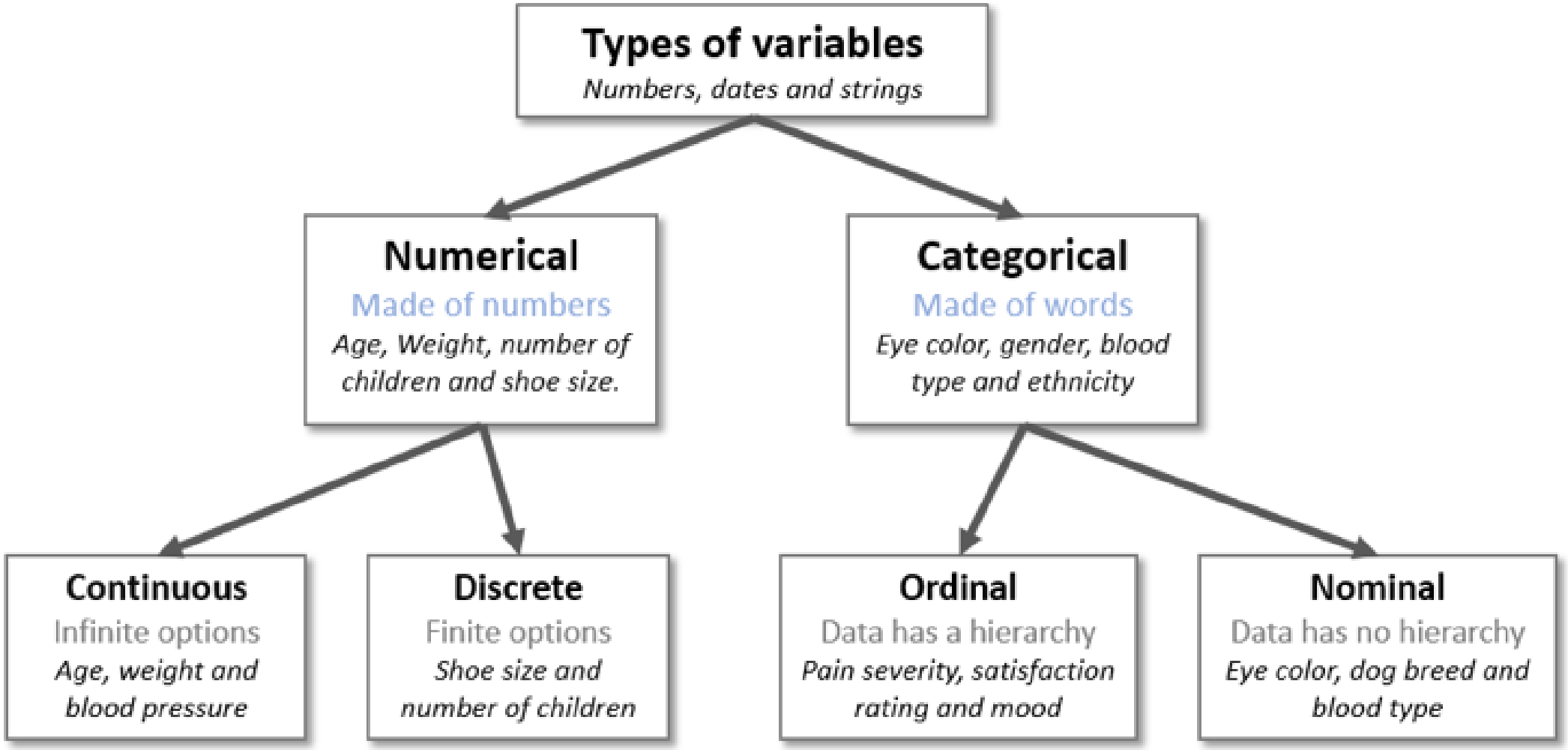
As per the report by NASSCOM and Blueocean, India is reigning big data analytics with a value of $1.2 billion placing it among the top 10 big data analytics markets in the world. They have also anticipated the growth becoming eight-fold by 2025, soaring to $16 billion. With this vision in mind, every sector is now looking forward to Data analytics for its evolution.

Storage capacity, size & cost



Data Generation

DATA TYPES IN STATISTICS



https://en.wikipedia.org/wiki/Statistical_data_type

NUMERICAL DATA

1. DISCRETE DATA

If its values are distinct and separate. In other words: We speak of discrete data if the data can only take on certain values. This type of data can’t be measured but it can be counted. It basically represents information that can be categorized into a classification. An example is the number of heads in 100 coin flips.

You can check by asking the following two questions whether you are dealing with discrete data or not: Can you count it and can it be divided up into smaller and smaller parts?

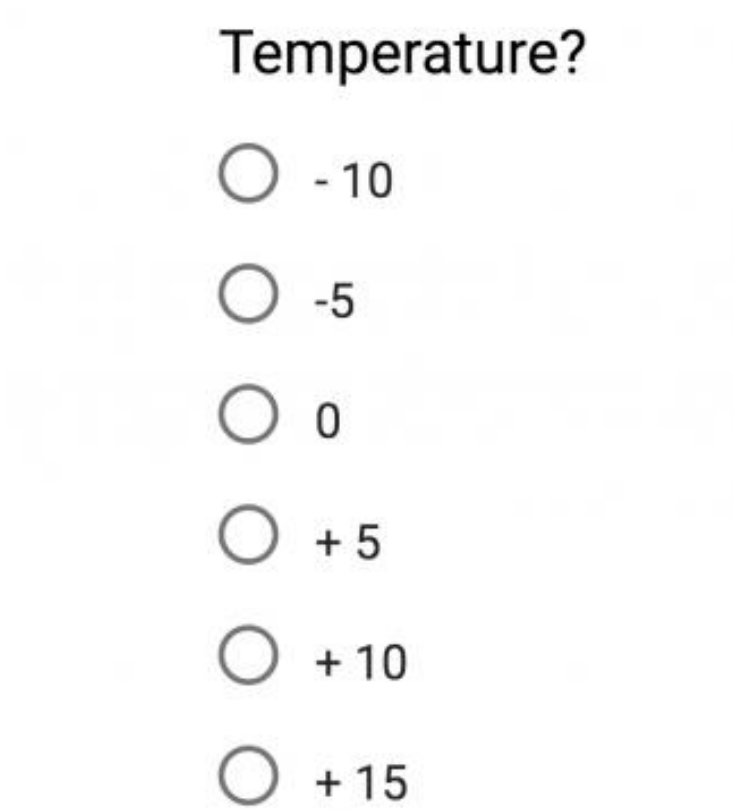
2. CONTINUOUS DATA

Continuous Data represents measurements and therefore their values can’t be counted but they can be measured. An example would be the height of a person, which you can describe by using intervals on the real number line.

Contd..

Interval Data

Interval values represent ordered units that have the same difference. Therefore we speak of interval data when we have a variable that contains numeric values that are ordered and where we know the exact differences between the values. An example would be a feature that contains temperature of a given place like you can see



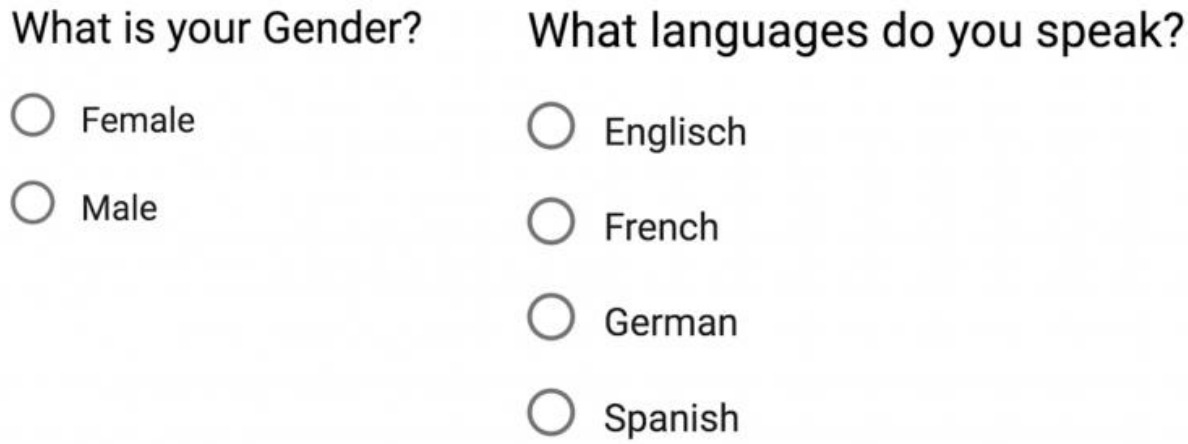
CATEGORICAL DATA

Categorical data represents characteristics. Therefore it can represent things like a person's gender, language etc. Categorical data can also take on numerical values (Example: 1 for female and 0 for male). Note that those numbers don't have mathematical meaning.

NOMINAL DATA

Nominal values represent discrete units and are used to label variables, that have no quantitative value. Just think of them as labels. Note that nominal data that has no order. Therefore if you would change the order of its values, the meaning would not change. You can see two examples of nominal features in the right.

The left feature that describes a persons gender would be called „dichotomous", which is a type of nominal scales that contains only two categories.



Contd..

ORDINAL DATA

Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters. You can see an example below:

Note that the difference between Elementary and High School is different than the difference between High School and College. This is the main limitation of ordinal data, the differences between the values is not really known.

Classification of Data in Real-World

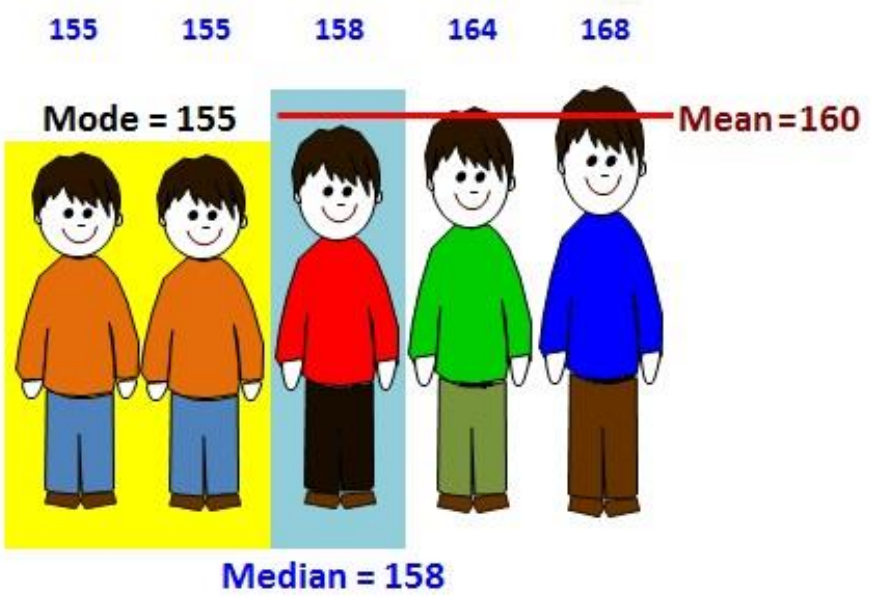
1. Structured Data → Data which is some structure is structured Data → Excel, CSV, DB, tsv, |, -
2. Un-Structured Data → Data which is not having any structure is structured Data → audio, video, image, txt, doc, → Computer Vision, NLP,
3. Semi-Structured Data → follow some structures and store un-structured data
    1. Excel file → Images
    2. NoSql →
    3. Json → {‘aadhar’ :’image’, “fingerprint”:}
    4. XML →
4. Product is good → |product |is |good |

What is Statistics?

A branch of mathematics that takes and transform the data into some useful information which in turn is used to make some decisions.

Statistics is concerned with

- Processing and analyzing data
- Collecting, presenting and transforming data to assist decision maker



Measures of Dispersion

Range: It is the difference between highest value and the lowest value in the data set.

For a given list of numbers: 10, 20, 40, 10, 70 the range is 70 - 10 = 60.

Variance: The average of the squared differences from the mean.

Steps to calculate variance:

- Calculate mean (mean is nothing but average)
- Find difference of each data from mean
- Square all the differences
- Take the average of the squares.

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Standard Deviation: It shows you how much your data is spread out around the mean. Its symbol is **σ** (the Greek letter sigma). It is the square root of the **variance.**