Introduction to Data Science

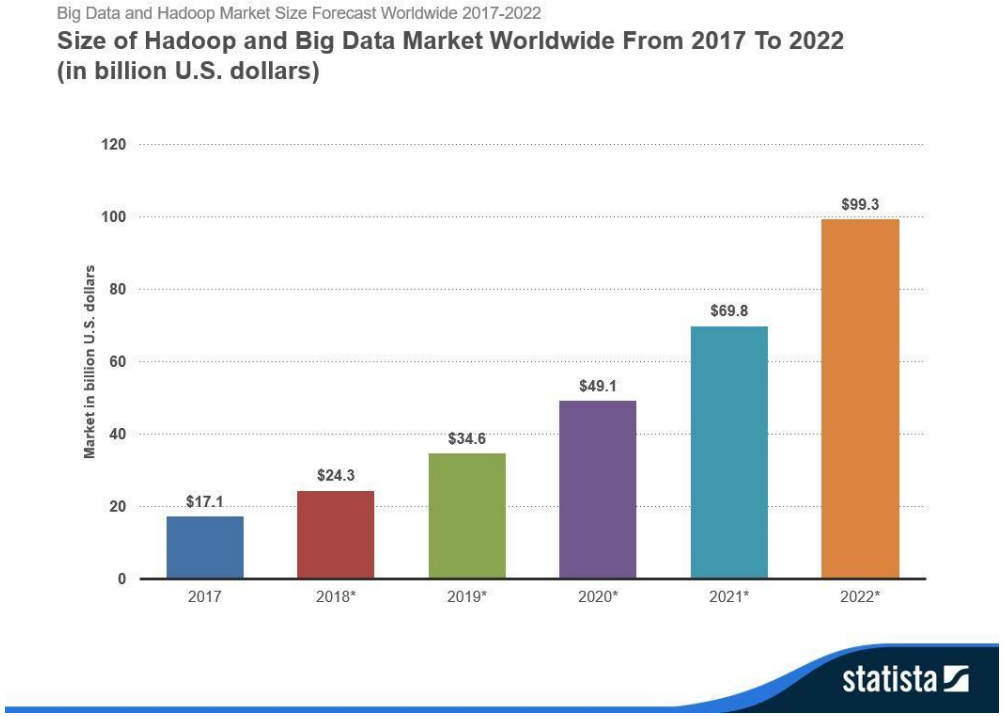
Data Science Academy

Data Science has been one of the biggest tech buzz word for the last 5-10 years!

- Data Science, Artificial Intelligence, Machine Learning, Big Data
- Those terms have been bouncing around every tech site and been heavily glamourized (even vilified) by the media!

The Big Data Industry is Growing Rapidly!

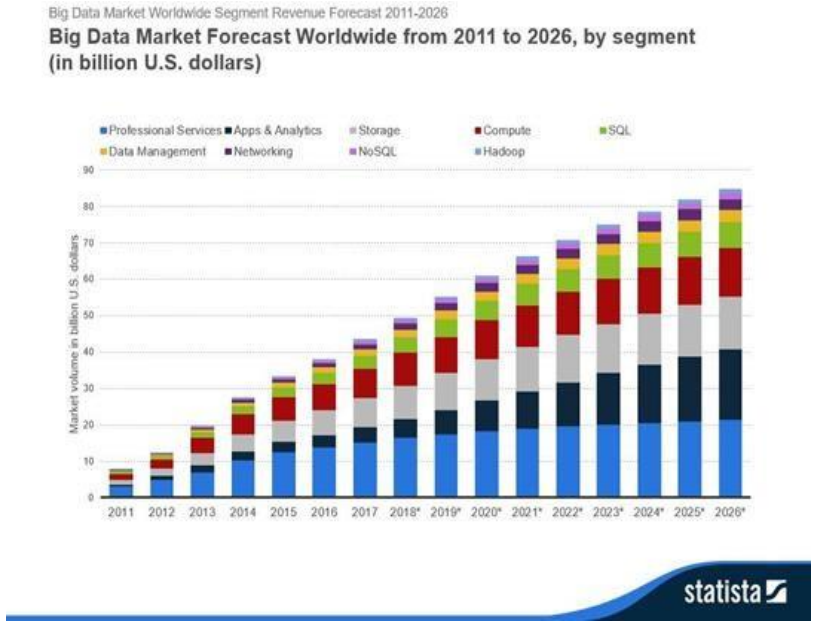
IDC Forecasts Revenues for Big Data and Business Analytics Solutions Will Reach $189.1 Billion This Year with Double-Digit Annual Growth Through 2022



The Demand for Data Scientists is only going up!

Demand for data scientists is booming and will only increase

Fueled by big data and AI, demand for data science skills is growing exponentially, according to job sites. The supply of skilled applicants, however, is growing at a slower pace.



18,002 views | Oct 14, 2019, 07:15am

The Birth Of The Data Science Generation

Forbes Technology Council

Ike Kavas Forbes Councils Member
Forbes Technology Council COUNCIL POST | Paid Program
Innovation

Has this ever happened to you?

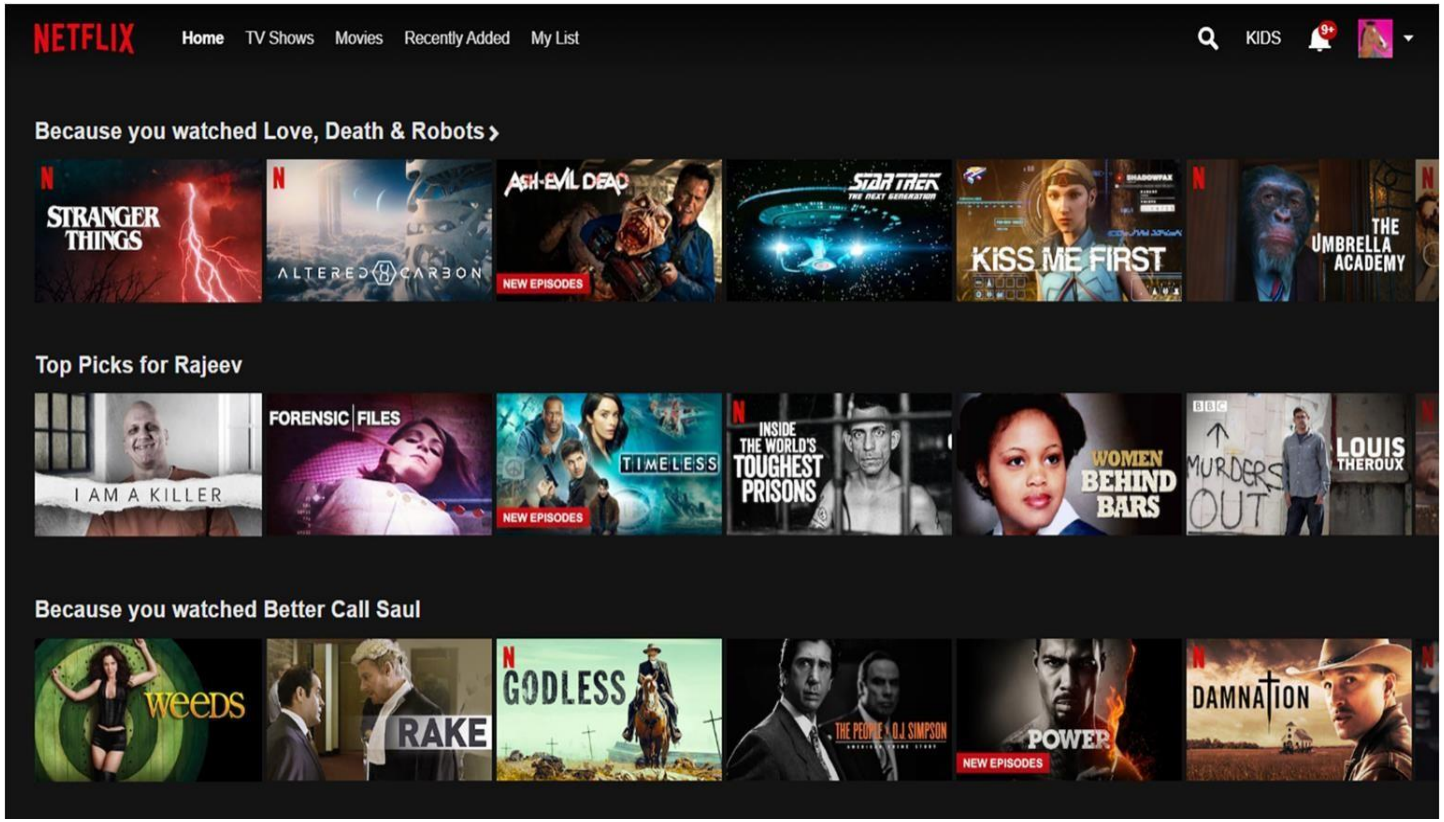
- You're thinking about something, maybe it's the new printer you wanted, or a skiing trip you were planning. Or perhaps thinking about starting a gym.
- Then BAM! You see an online Ad for the exact thing you wanted.

Gmail

How do those online giants like Google and Facebook know you so well?

- Unfortunately, it's not magic
- It's DATA
- And it's everywhere (including yours)
- Let's take a look at 5 super interesting examples of how Data Driven companies are changing the business landscape

Netflix's Recommendations



The Ubiquity of Data Opportunities

- Automated Recruitment of Employees
- Crypto, Forex & Stock Price Predictions, Credit Risk Scoring Models
- Health Analytics – Disease Prediction, finding cures etc.
- Computer Vision – Understanding what is being seen to build things like self driving cars and facial recognition
- Agriculture
- Manufacturing
- Creating Art and Music
- Self Driving cars and Robots
- Chat Bots
- And thousands more! There is no shortage on areas we can apply Data Science

Here's how data science helped Zoomcar capture 75% of Indian market

"Data lake and models built on Kafka helped us achieve enhanced customer experience at the best possible price," Arpit Agarwal, Director, Decision Science, Zoomcar, says.

Nikhar Aggarwal | ETCIO | October 31, 2019, 09:23 IST

Data = Value = Better Decisions = More Profits!

Data Analysts vs. Data Scientists

No Machine Learning
doesn't need to be an ace programmer

Uses Machine Learning
and needs to be good at programming

How Data Scientists Approach Problems

Data Science Academy

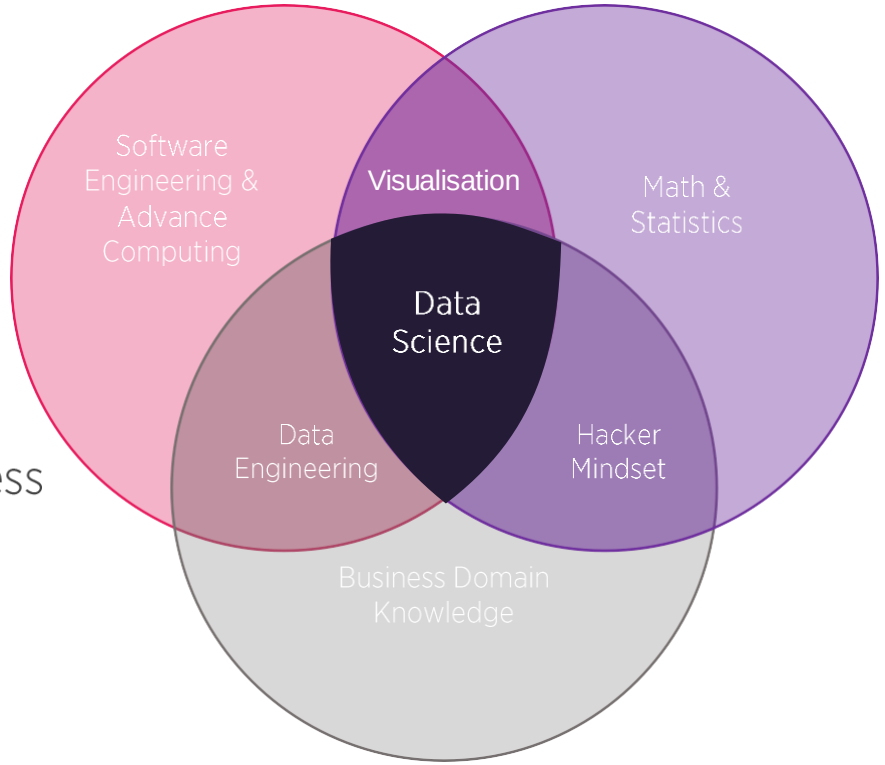
Course outline:

1. Data Science Divisions and Fields
2. What is Data Analytics & Statistics
3. Basic Statistical Concepts (Population, Sample)
4. Descriptive Statistics
5. Types of data
6. Levels of measurements
7. Descriptive statistics (Numerical Methods)
8. Measure of locations
9. Measure of variability
10. Normal Distribution
11. Z-score
12. Empiric Rule
13. Detecting Outliers
14. Why Python for Data Science
15. Companies using Python
16. Why does Python lead the Pack
17. Installing Anaconda
18. Data types - Variables
19. Data Structures
20. Add comments
21. Arithmetic operations
22. Defining a functions
23. Indexing
24. Conditionals
25. For loops
26. While loops and incrementing

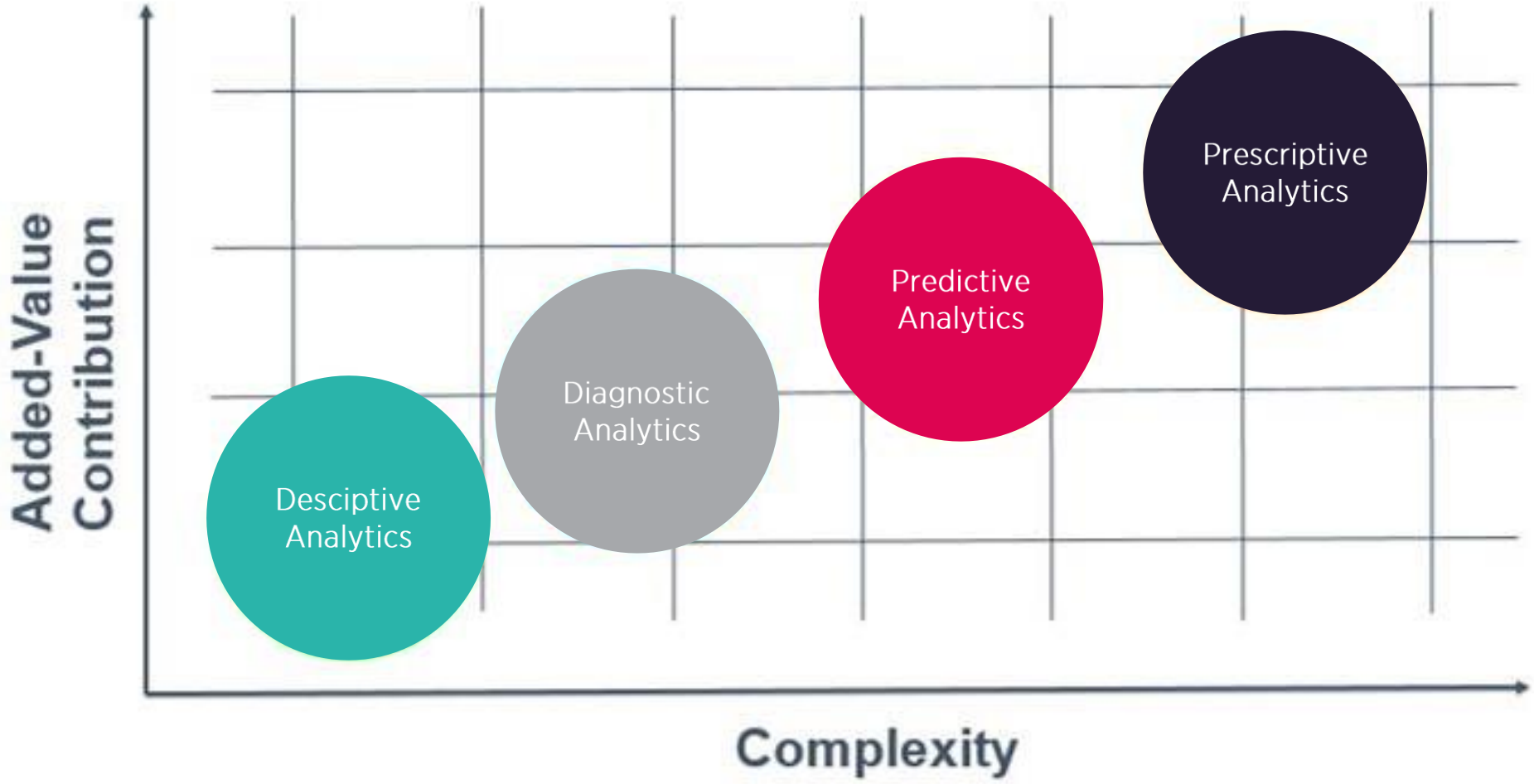
What you'll be able to do after

Understand all aspects that make one a complete Data Scientist

- The software programming chops to mess with Data
- The Analytical and Statistical skills to make sense of Data
- All the Machine Learning Theory needed for Data Science
- Real World Understanding of Data and how to understand the business domain to better solve problems



4 Steps of Data Analytics



Statistics for Data Scientists

Data Science Academy

Statistics – No one likes statistics

- Almost everyone seems to have a dislike their high school or college Statistics classes
- Most found it confusing, difficult and think it's useless real life.
- They couldn't be more wrong…

Statistics – Why is so important?

Everyone deals with statistics!

- Will it rain tomorrow?
- Who's expected to score the most goals in the next World Cup?
- Is Trump going to win the next election?

And in business...

- What month will I have the most sales, or what time of day?
- Should I take out insurance?
- Is the economy doing well?

Everything dealing with forecasting/predicting comes down to Statistics

Companies and Researchers leveraging statistical knowledge know:

- What products of movies you like (think Amazon or Netflix)
- When fraudulent activities are taking place
- Predicting customer demand
- Understanding the cause of certain illnesses
- Understanding what the best advertisements, medications, diets and more!
- Everything dealing with forecasting/predicting comes down to Statistics

“

“Being a Statistician will be the sexiest job over the next decade”

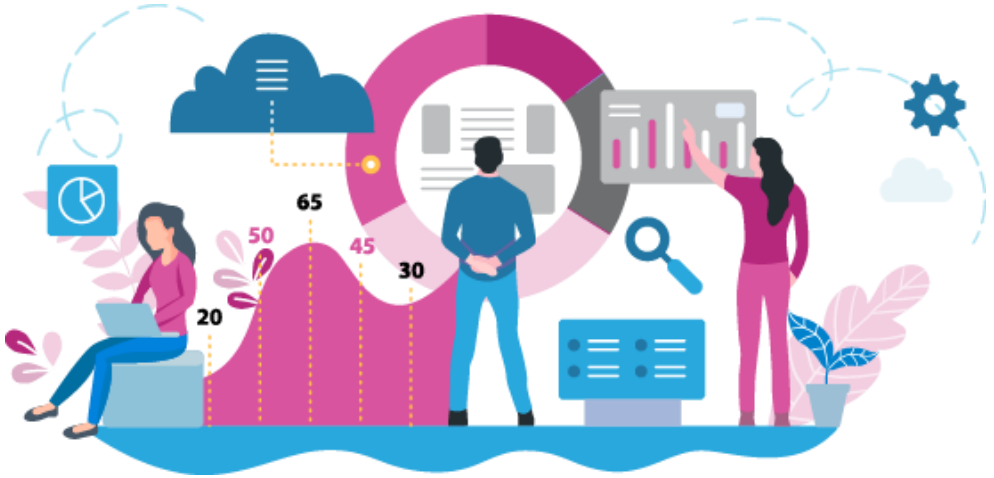
Hal Varian, Chief Economist, Google

Data Science Academy

The Subfields of Statistics

- Descriptive Statistics – Measures or descriptions used to assess some performance or indicator e.g. Batting Averages, GPA

- Inference – Using knowledge from data to make informed inferences, e.g. answering the questions “How often to people get the common cold” or “How many people can afford to a buy a house at the age of 25”



The Subfields of Statistics

- Risk and Probability – What's the likelihood of your rolling a 6 on a dice? Probability is an extensive and important field that is critical for many businesses such as Insurance, Casinos and Finance.

- Correlation and Relationships – How do we know smoking causes cancer? Extensive statistical studies have to be used for Hypothesis testing. This is an area that's often extremely difficult, but extremely useful in making impactful decisions.

The Subfields of Statistics

- Modeling – Many times in movies or documentaries, you'll hear scientists referring to "Their model predicts X". In the real world, especially in data science, modeling is the bread and butter of the job. Building good models that predict some outcome based on the inputs, is critical for many industries. E.g. predicting which customers will purchase Item A.

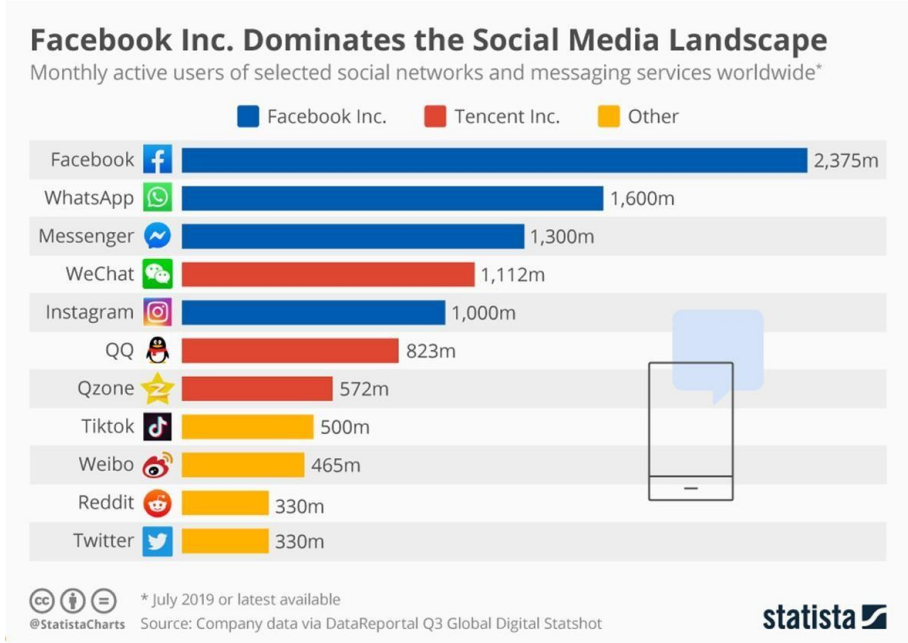
Descriptive Statistics

Descriptive Statistics

- Descriptive statistics are used to describe or summarize data in ways that are meaningful and useful, such that, for example, patterns might emerge from the data.

Examples of descriptive Statistics

- Average heights and weights of males (5’9” and 184lbs for the UK)
- Average number of items sold per day at a store



Exploratory Data Analysis

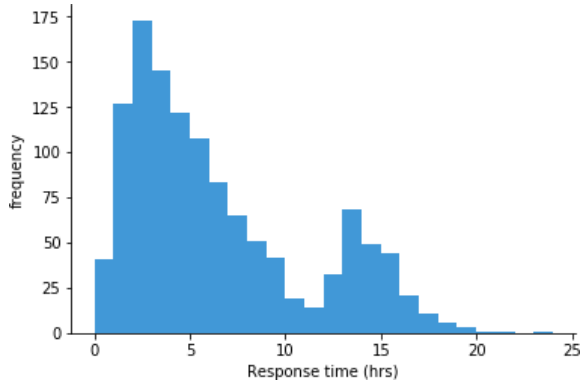
Data Science Academy

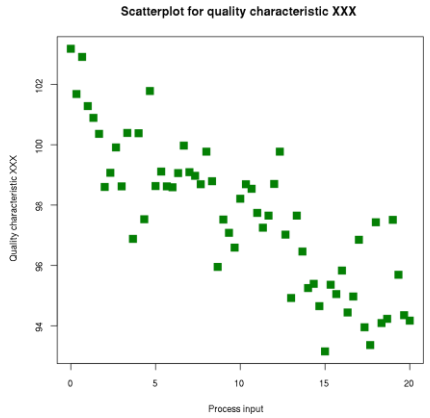
Exploratory Data Analysis (EDA)

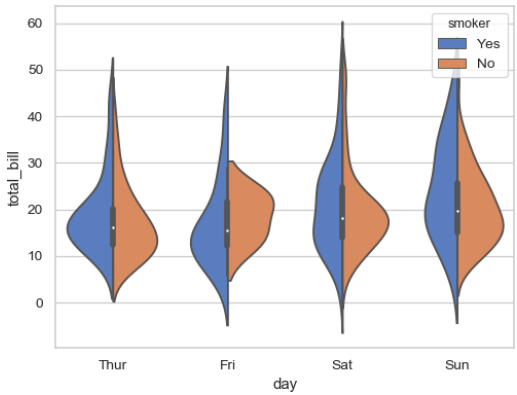
This is the process where we visualize, examine, organize and summarize our data.

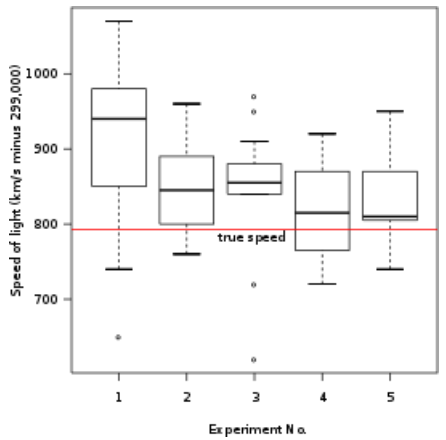
Methods of EDA

- Histogram Plots
- Scatter Plots
- Violin Plots
- Boxplots
- Many more!

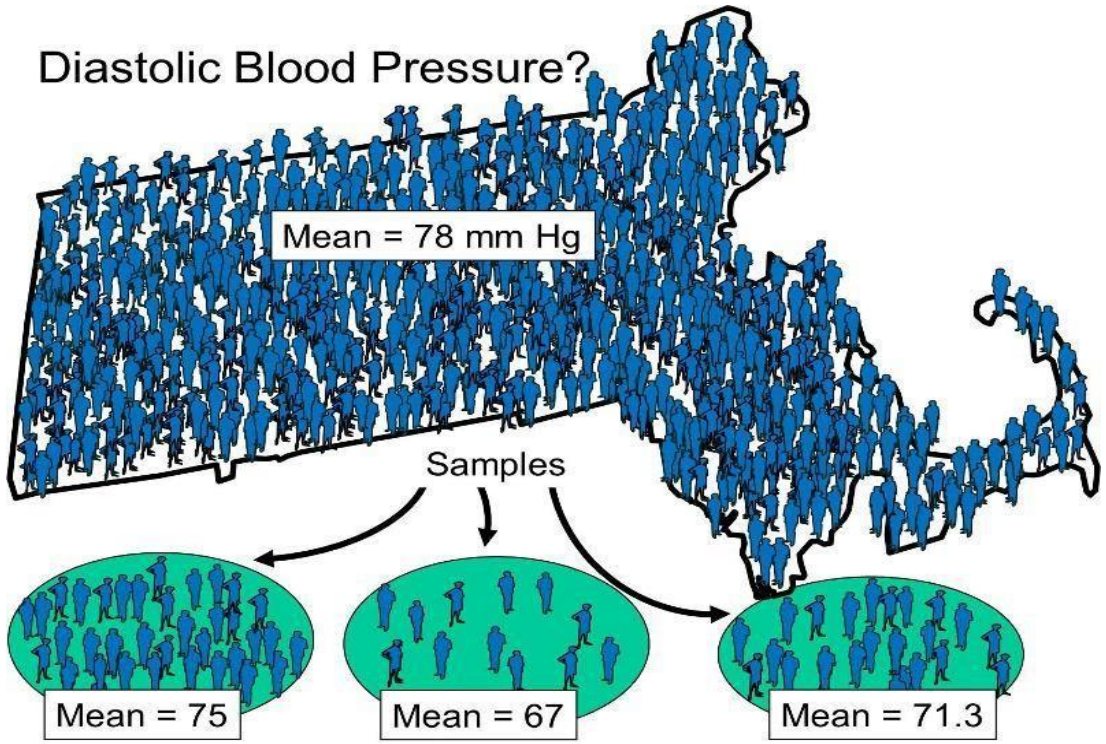








Basic statistical concepts



Sampling Summary

- A subset of a **population** (i.e. the total or all individuals in a set) is called a **sample**
- A good sample aims to be a good **representative** form of that population
- **Sampling Error** is the difference between the parameters or descriptive statistics (i.e. mean etc.) of the population. **Small Sampling Errors** are good

Types of Sampling that we use to create good representations include:

- Random sampling
- Satratified sampling

Variables in Data

Data Science Academy

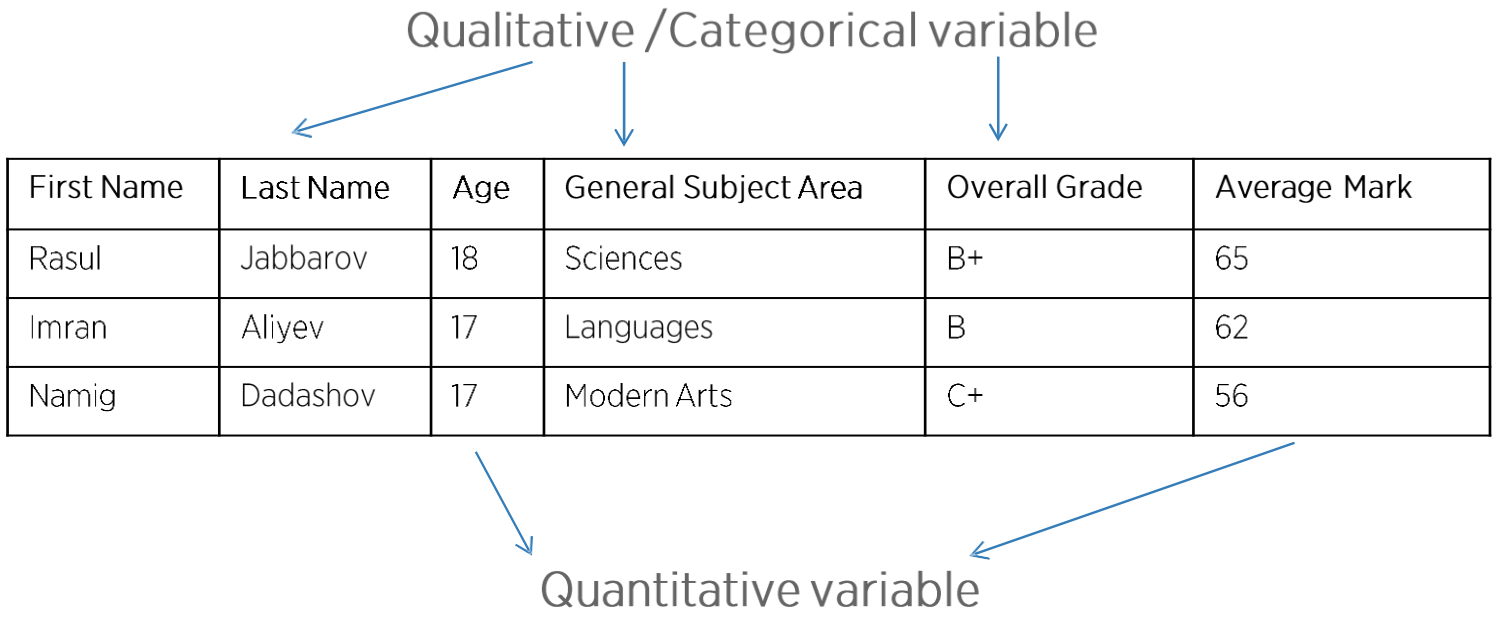
What are the variables?

- Understanding the data we encounter is essential
- Think of data as the information we collect to describe some something

Data can come in two main forms:

- Quantitative variables
- Qualitative or Categorical variables

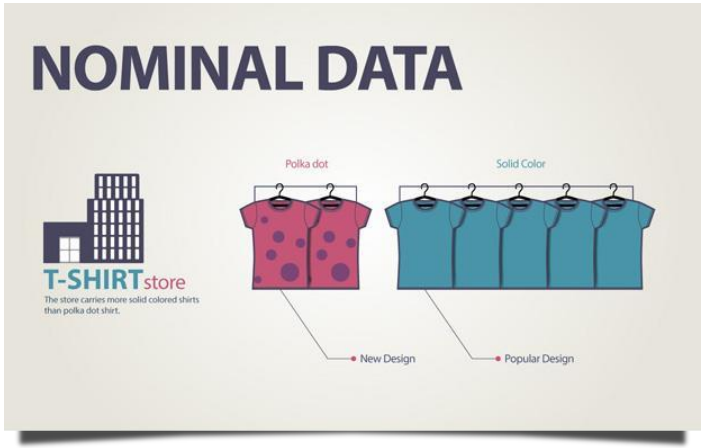
Understanding Variable Types

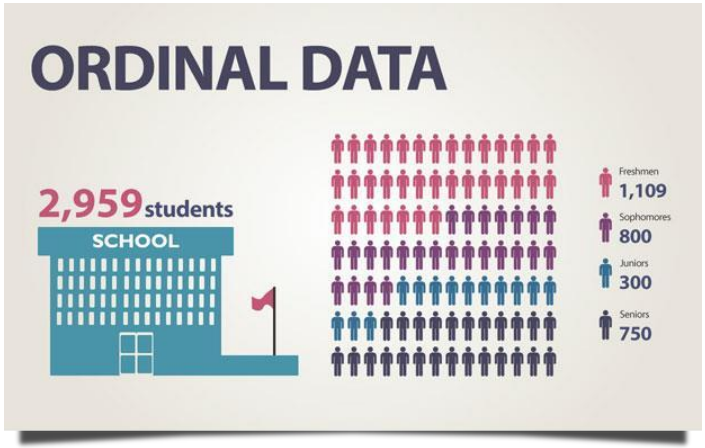


First Name	Last Name	Age	General Subject Area	Overall Grade	Average Mark
Rasul	Jabbarov	18	Sciences	B+	65
Imran	Aliyev	17	Languages	B	62
Namig	Dadashov	17	Modern Arts	C+	56

Nominal and Ordinal

- Nominal scale variables differentiate individual data points e.g. an ID variable or Name
- They say nothing about the value, direction, size or any quantitative measure. They are always qualitative.
- Ordinal scale variables can measure direction, size and values. However, e.g. for a high, low, medium measurements. We don't know exact values i.e. how high or how low.





Interval & Ratio

- **Interval scales** can tell us the size difference between categories, however the still can't tell us the exact value e.g. Date of Birth

- **Ration scales** are similar to interval, but they have the added property of having an inherent zero. E.g. height or weight.

Continuous and Discrete Variables

- Goals are **discrete**, there is no way a player can score a fraction of a goal. Discrete variables measure quantity and value, but have no interval measurement between adjacent values.

- Height however, is a continuous. Just because we give whole number integer values, that doesn't mean Player 1 and Player 3 are exactly the same height. Player 1 can be 177.3cm while Player 3 can be 176.9cm. You can perhaps never have exactly the same value in height of two people as there would be nanometer differences in height

Player	Goals	Height
Player 1	43	177cm
Player 2	25	180cm
Player 3	3	177cm

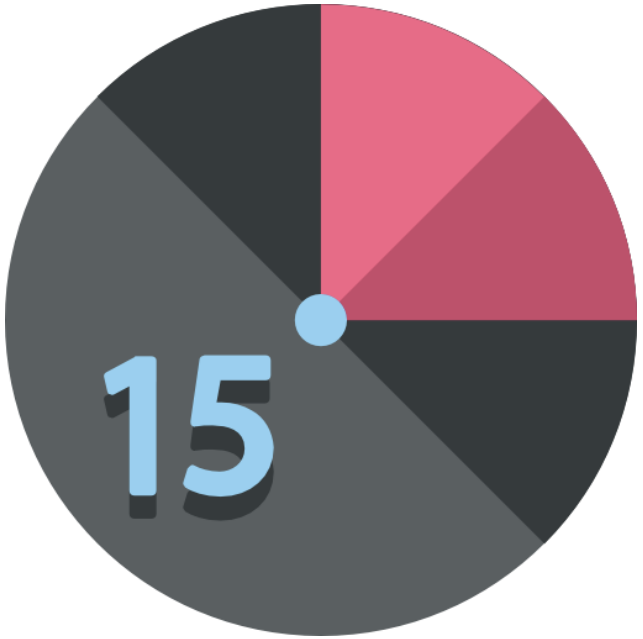
Descriptive Statistics: Numerical Methods

Measures of location

Measures of variability

Measures of Location

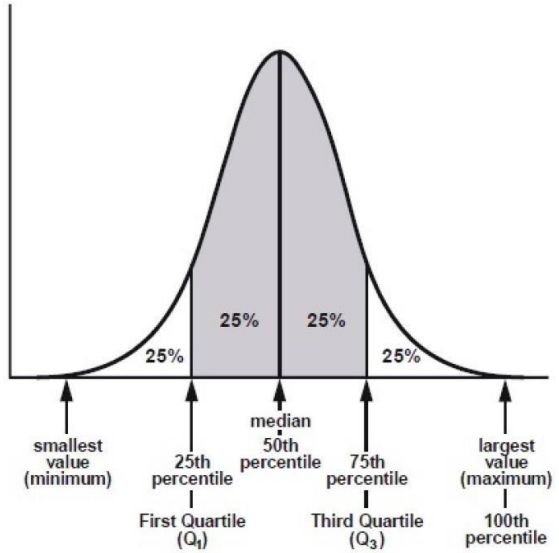
Percentiles



Quartiles

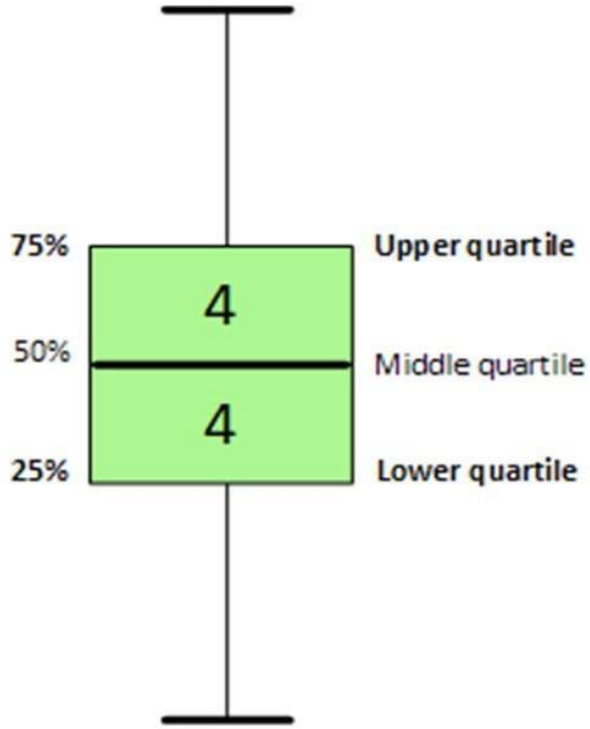
Percentiles

- A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.
- Admission test scores for colleges and universities are frequently reported in terms of percentiles.



Outliers Rule of Thumb

- A value can be considered an outlier if it exceeds 1.5X the difference between the upper quartile and the lower quartile (the inter Quartile range).



Inter Quartile Range = Upper quartile - Lower quartile

Measures of location

Mean

The mean is the most widely spread measure of central tendency. It is the simple average of the dataset. (Easily affected by outliers)
The formula to calculate the mean is:

- $\frac{\sum_{i=0}^{N} Xi}{N}$ or $\frac{X1+X2+X3+\cdots+XN-1+XN}{N}$

In Excel, the mean is calculated by:
=AVERAGE()

Median

The median is the midpoint of the ordered dataset. It is not as popular as the mean, but is often used in academia and data science. That is since it is not affected by outliers.
In an ordered dataset, the median

is the number at position $\frac{n+1}{2}$

If this position is not a whole

number, it, the median is the simple average of the two numbers at positions closest to the calculated value.

In Excel, the mean is calculated by:
=MEDIAN()

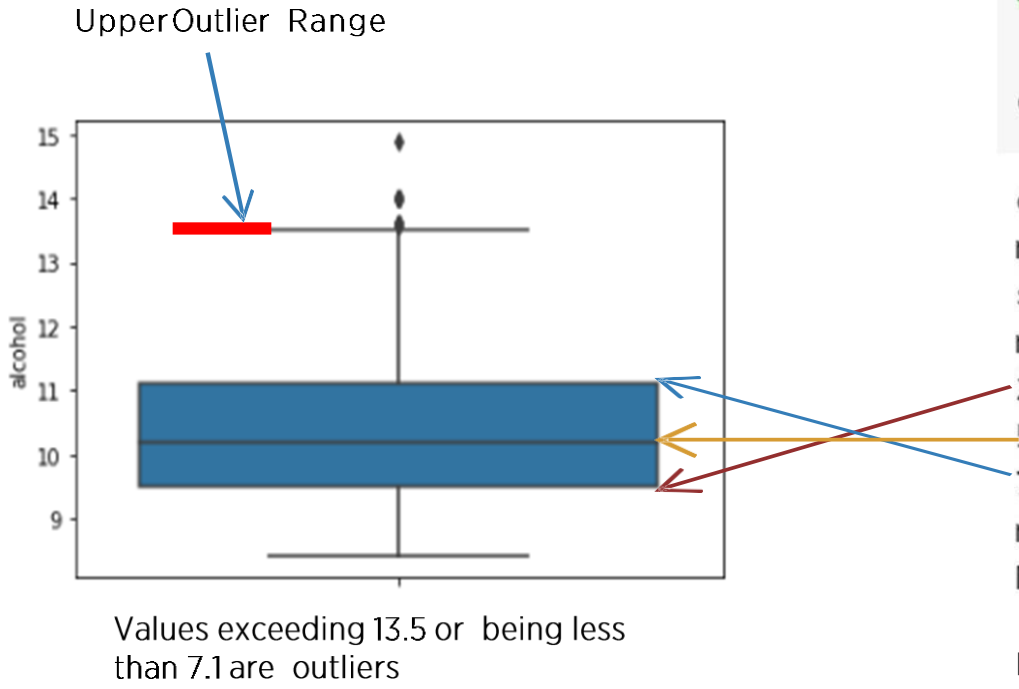
Mode

The mode is the value that occurs most often. A dataset can have 0 modes, 1 mode or multiple modes. The mode is calculated simply by finding the value with the highest frequency.

In Excel, the mean is calculated by:
=MODE.SNGL() -> returns one mode

=MODE.MULT() -> returns an array with the modes. It is used when we have more than 1 mode.

Inter Quartile Ranges for



Values exceeding 13.5 or being less than 7.1 are outliers

Red Wines

```
# Showing the quartile ranges of alcohol content

df[df['type'] == 'red']['alcohol'].describe()
```

```
count    1599.000000
mean       10.422983
std         1.065668
min         8.400000
25%         9.500000
50%        10.200000
75%        11.100000
max        14.900000
Name: alcohol, dtype: float64
```

Upper Quartile = 11.1
Lower Quartile = 9.5

Inter Quartile Range = 11.1 – 9.5 = 1.6

Outlier Ranges = 1.6 * 1.5 = 2.4

Upper Outlier Range = 11.1 + 2.4 = 13.5

Lower Outlier Range = 9.5 - 2.4 = 7.1

The Mean

- The mean is quite simple and we will have discussed it before so you intuitively know that mean is the same as average.

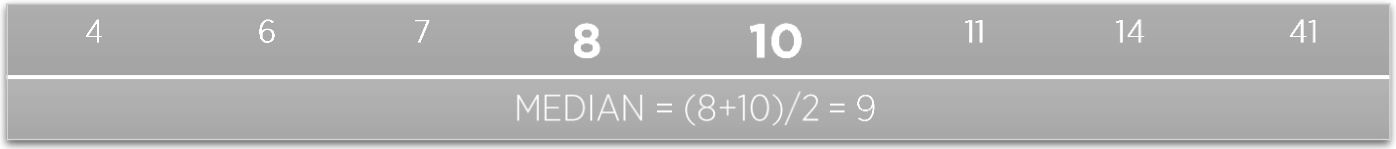
$$\text{Mean} = \frac{10 + 16 + 7}{3} = 11$$

Person	ARPU
Naila	10
Afag	16
Sanan	7
Mean	**11**

Median

- Many people confuse Means and Medians.
- Remember while means are the average of all values, Median is average of the two middle values or the actual middle value itself (depending on if the quantity of data is odd or even)

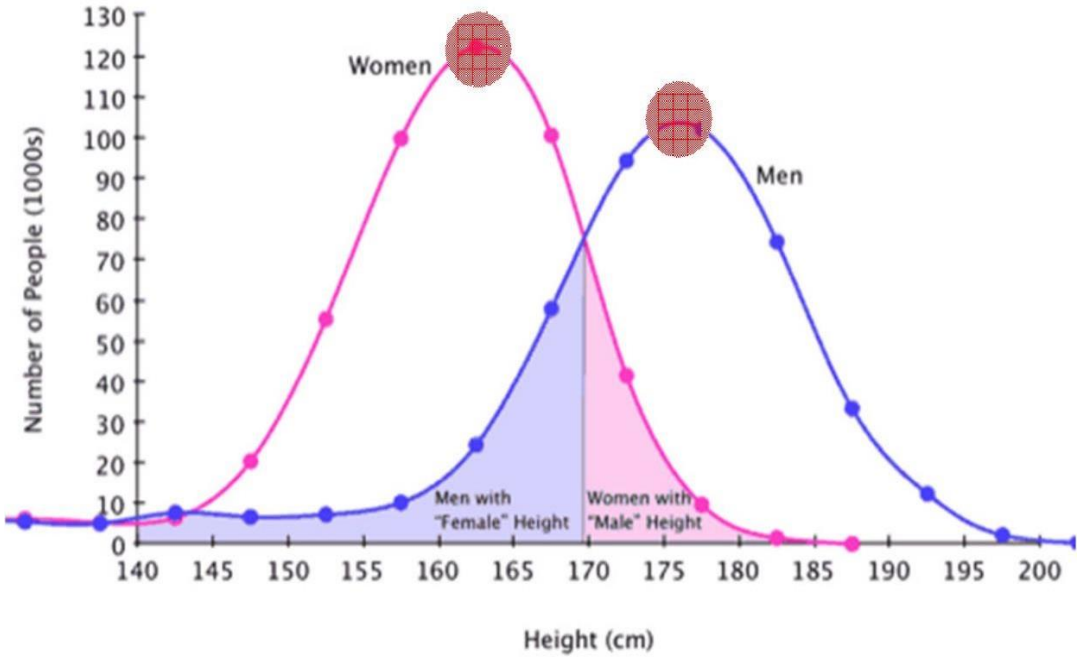
Mean= 10.86



Mean= 12.63

Mode

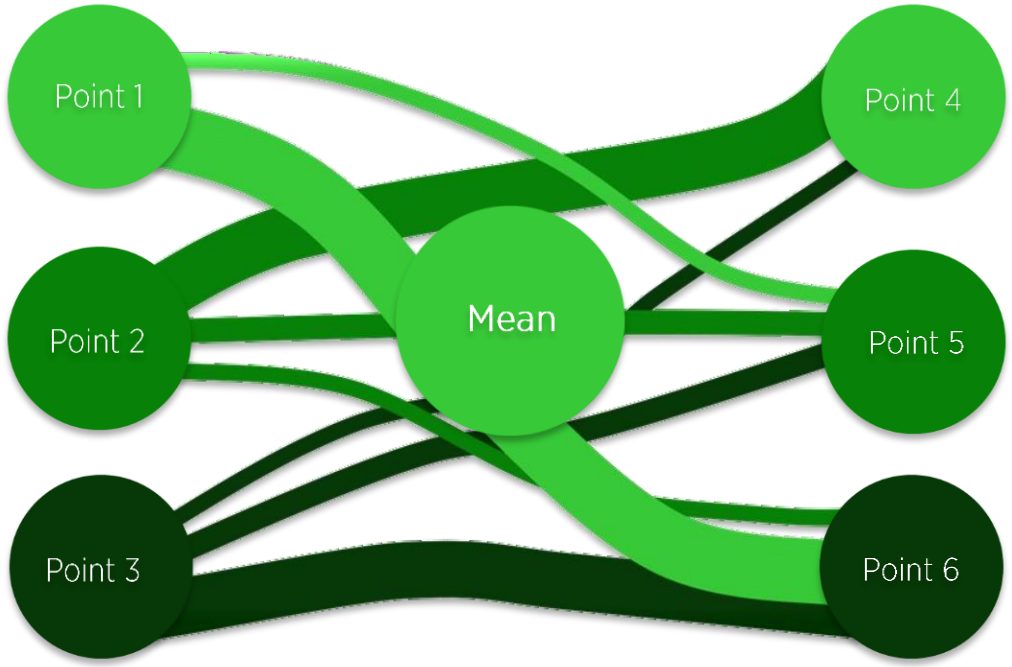
- The Mode is simply the most frequent item in distribution.
- In any Kernel Density plot (KDE in Seaborn), the mode is always the peak



Measures of Variability

- It is often desirable to consider measures of variability (dispersion), as well as measures of location.
- For example, in choosing supplier A or supplier B, we might consider not only the average delivery time for each, but also the variability in delivery time for each.

Measures of Variability



Calculating variance in Excel:

Simple variance: =VAR.S()

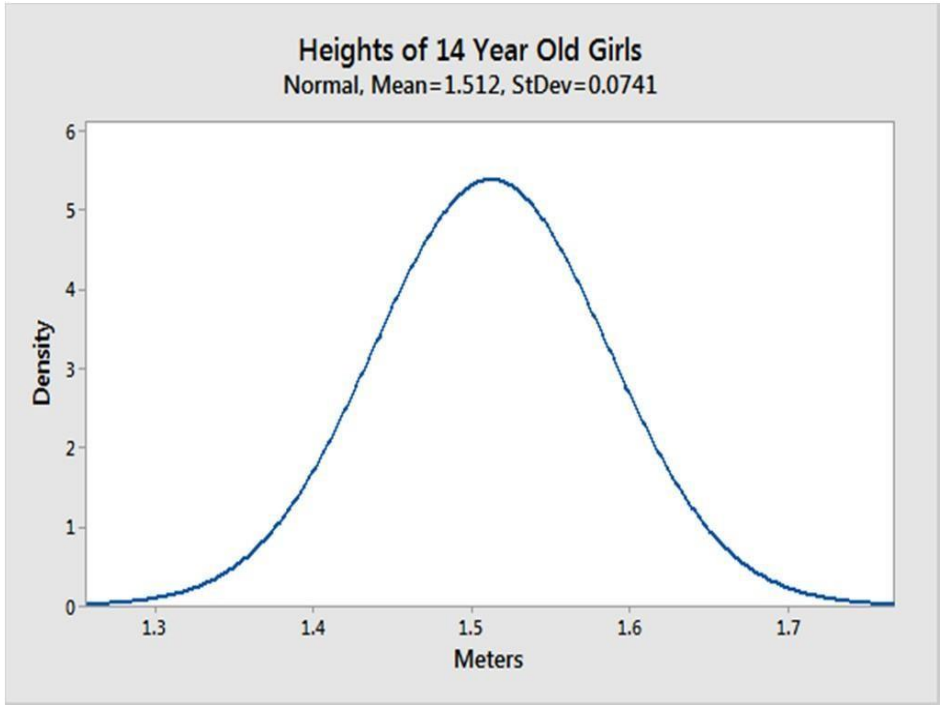
Population variance: =VAR.P()

Simple standard deviation: =STDEV.S()

Population standard deviation: =STDEV.P()

- Variance and standard deviation measure the dispersion of a set of data points around ist mean value.
- There are different formulas for population and sample variance & standard deviation. This is due to the fact that the sample formulas are the unbiased estimators of the population formulas.
- Sample variance formula $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
- Population variance formula $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$
- Sample standard deviation formula $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$
- Population standard deviation formula $\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$

Normal Distributions Example



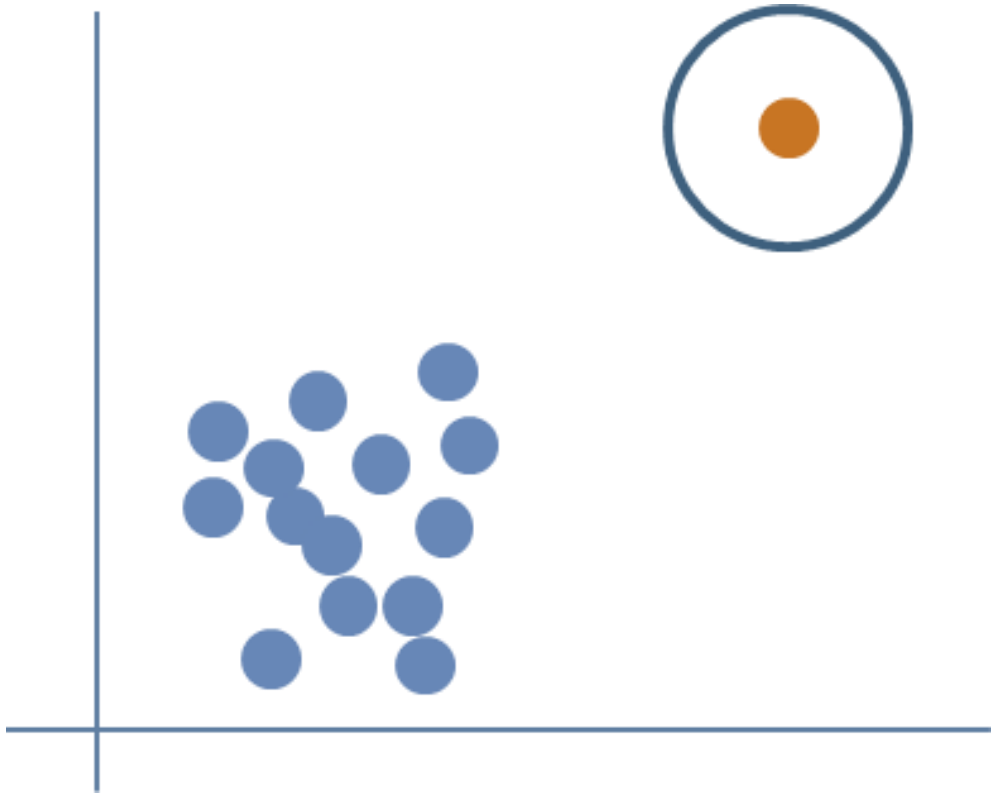
Data Science Academy

Detecting Outliers

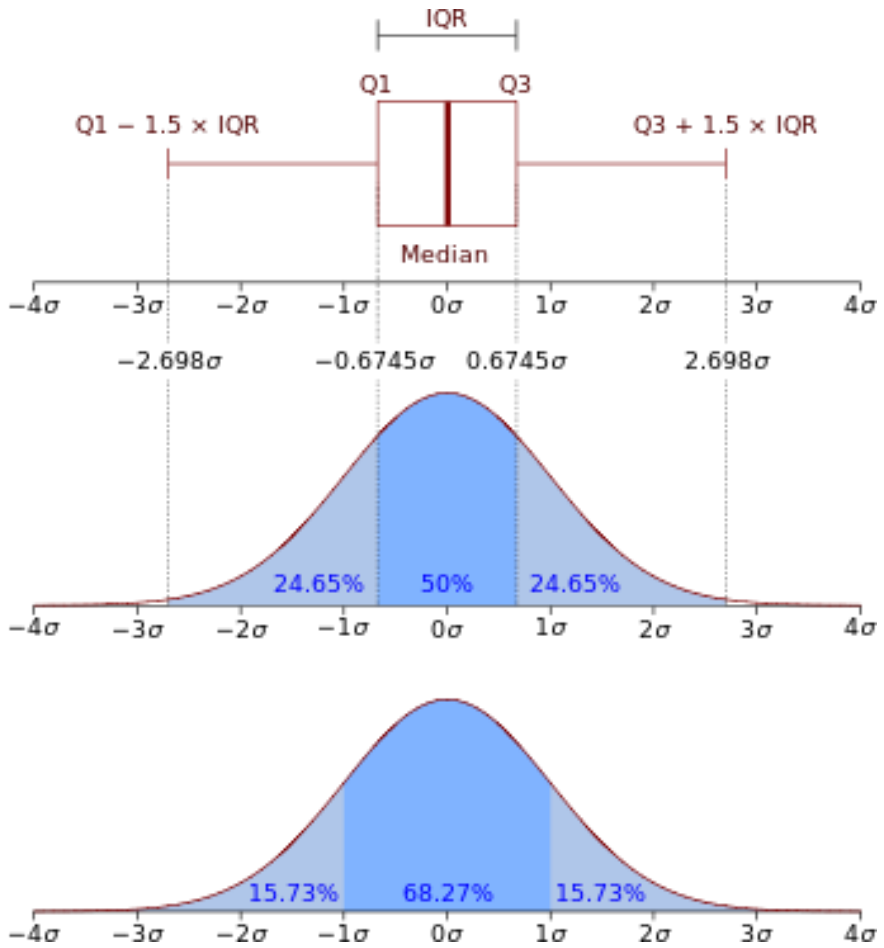
Data Science Academy

Detecting Outliers

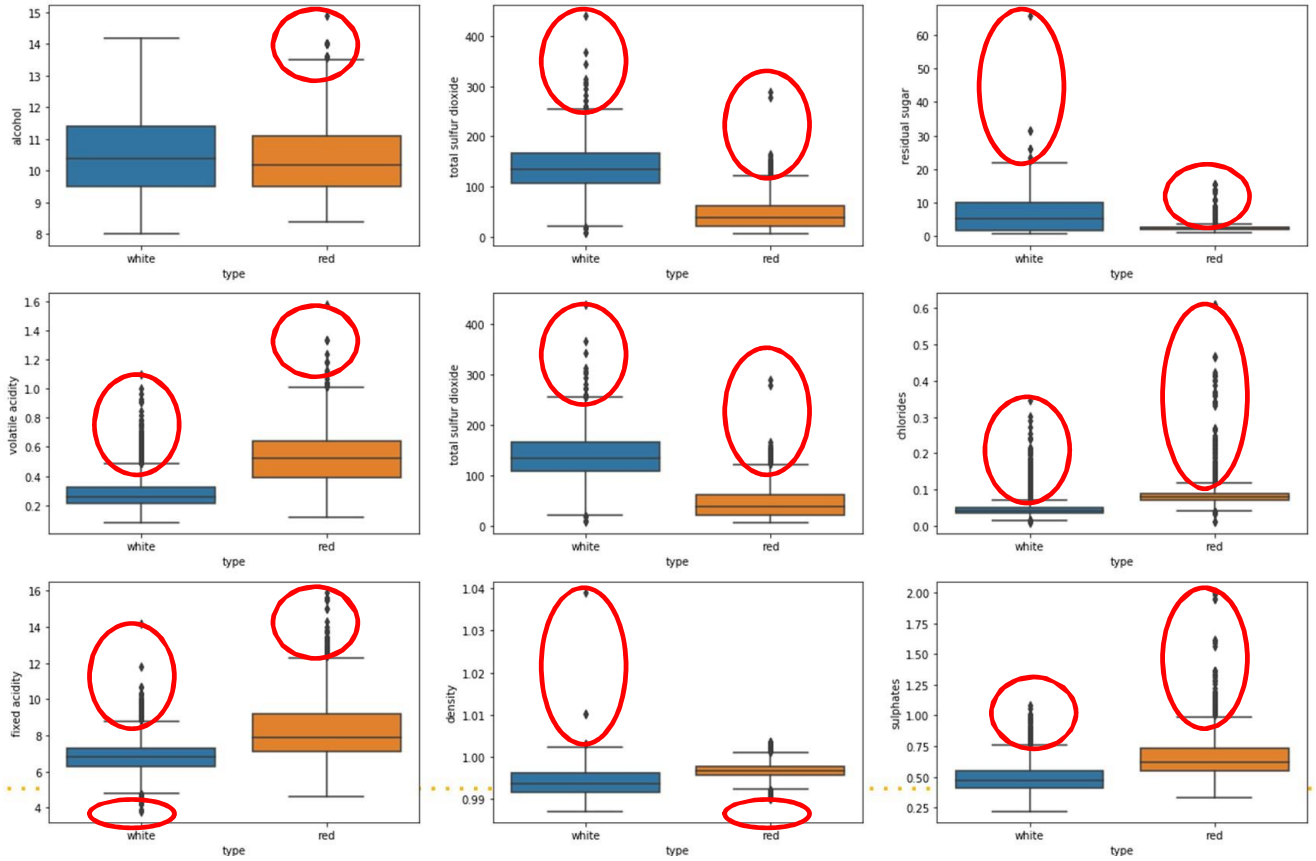
- An outlier is an unusually small or unusually large value in a data set.
- A data value with a z-score less than -3 or greater than +3 might be considered an outlier.
- It might be an incorrectly recorded data value.
- It might be a data value that was incorrectly included in the data set.
- It might be a correctly recorded data value that belongs in the data set !



Detecting Outliers



Analyze Frequency Distributions to Find Outliers



Data Science with Python

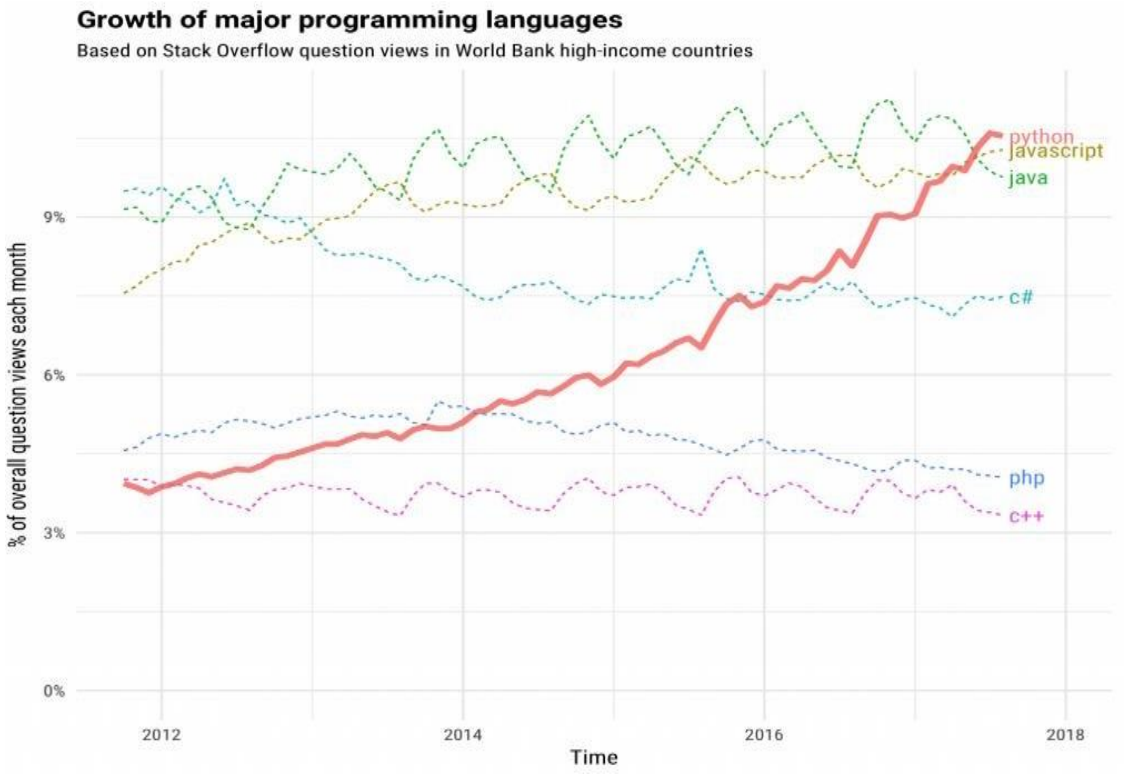
Data Science Academy

Python

- Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes **code readability** with its notable use of significant whitespace
- Python is an interpreted, high-level, general-purpose programming language.
  - **Interpreted** – Instructions are execute directly without being compiled into machine-language instructions. Compiled languages unlike interpreted languages, are faster and give the developed more control over memory management and hardware recourses.
  - **High-level** – allowing us to perform complex tasks easily and efficiently

Why Python for Data Science

- Python competes with many languages in the Data Science world, most notably R and to a much lesser degree MATLAB, JAVA and C++.



Companies using python

Why does Python Lead the Pack?

- It is the only general-purpose programming language that comes with a solid ecosystem of scientific computing libraries.
- Supports a number of popular Machine Learning, Statistical and Numerical Packages (Pandas, Numpy, Scikit-learn, TensorFlow, Cython)
- Supports easy to use iPython Notebooks, especially handy for view Data Science work.
- Quite easy to get started
- As a general purpose language it allows more flexibility such as building webservers, APIs and a plethora of other useful programming libraries.

7

What is Anaconda?

- Free and open-source distribution of the python and R
- Predominantly used for Data Science, Machine Learning and large scale data processing
- Over 12 million user
- 1400 packages..

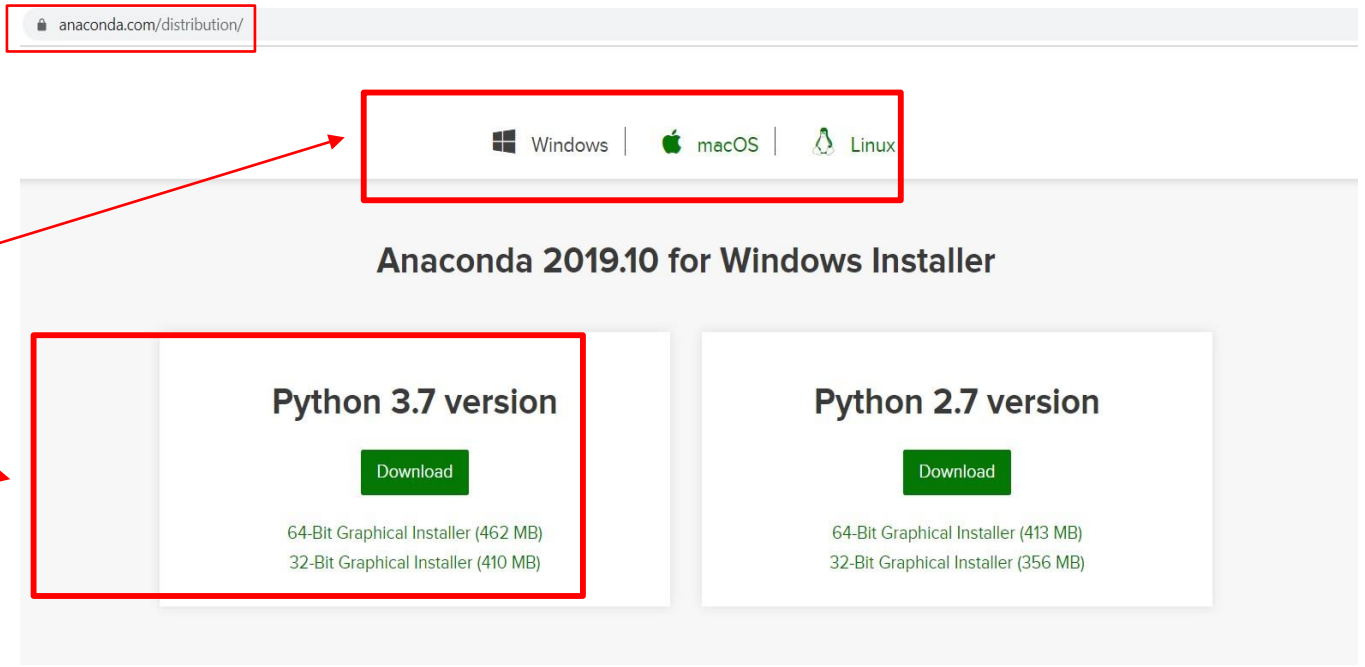
System Requirements

- Operating system: Windows 7 or newer, 64-bit macOS 10.10+, or Linux, including Ubuntu, RedHat, CentOS 6+, and others.
- Older versions of Anaconda are available in archive
- System architecture: Windows- 64-bit x86, 32-bit x86; MacOS- 64-bit x86; Linux- 64-bit x86, 64-bit Power8/Power9.
- Minimum 5 GB disk space to download and install.

Install Anaconda

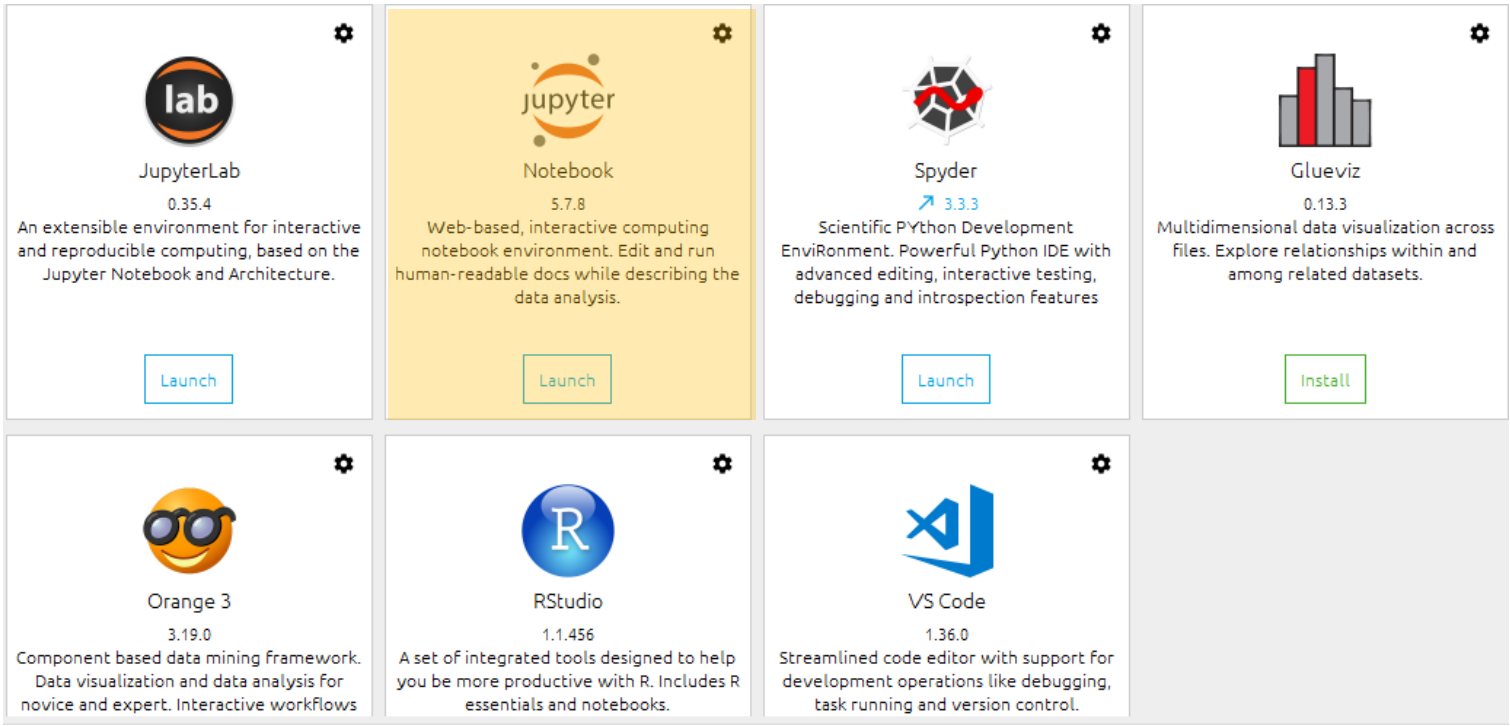
Check the given link and download "Anaconda"

- Choose operating system of your computer
- Download Python 3.7 version

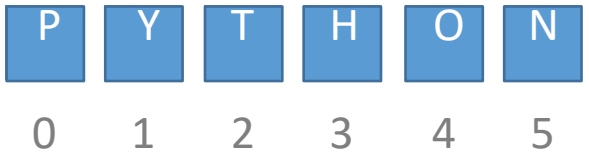


Install Jupyter Notebook

- Run Anaconda
- Open “jupyter Notebook”



Data types in Python - Variables



- Reserved memory locations or space
- Standard Data Types of variables
  - Numbers
  - String
  - Boolean
  - List
  - Tuple
  - Dictionary

True
False

34
56
7383

Integers

3.0
4.56

Float

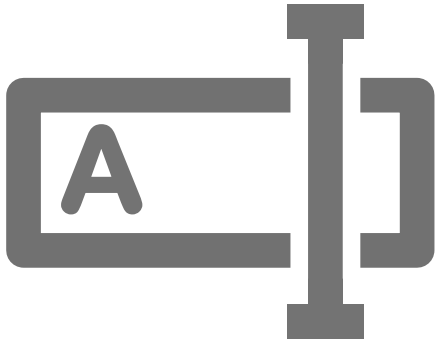
Integer

In Python 3, there is effectively no limit to how long an integer value can be. Of course, it is constrained by the amount of memory your system has, as are all things, but beyond that an integer can be as long as you need it to be.

123

String

- Strings are sequences of character data. The string type in Python is called **str**.
- A string in Python can contain as many characters as you wish. The only limit is your machine's memory resources. A string can also be empty.
- Strings are immutable that means once defined they cannot be changed



Boolean

- Boolean values are the two constant objects *False* and *True*.
- In numeric contexts (for example, when used as the argument to an arithmetic operator), they behave like the integers 0 and 1, respectively.
- The bool() function allows you to evaluate any value, and give you True or False in return.

List

- Sequence of elements
- Similar to array in other programming languages
- Elements are indexed

List1 = [“Ali”, “Namiq”, “Leyla”, “Gunay”]

0 1 2 3

Tuple

- Sequence of elements
- Similar to array in other programming languages
- Tuples are immutable
- Faster to process
- Elements are indexed

Tuple1 = (“Ali”, “Namiq”, “Leyla”, “Gunay”)

0 1 2 3

Dictionary

- Collection of key-value pairs
- Values can be accessed using the Key
- Used for JSON format conversion

Address

Key	Value
Street	Ashiq Ali 2
City	Baku
Region	Absheron
Country	Azerbaijan

Address = { 'Street': Ashiq Ali 2', 'City': 'Baku', 'Region' : 'Absheron', 'Country': 'Azerbaijan'}

Address['Street'] 'Ashiq Ali 2' Address['Region']
'Absheron'

Add Comments

- Comments are sentences not executed by the computer; it doesn't read them as instructions. The trick is to put a *hash sign* at the beginning of each line you would like to insert as a comment.

- If you would like to leave a comment on two lines, don't forget to place the hash sign at the beginning of each line.

- Or another thing you can do is use multiline strings by wrapping your comment inside a set of triple quotes, that's called documentation string

```
In [1]: #This is just a comment and not code!
        print 7,2

        7 2

In [2]: #Comment 1
        #Comment 2
        print 1

        1
```

```
"""
If I really hate pressing `enter` and
typing all those hash marks, I could
just do this instead
"""
```

Arithmetic Operators

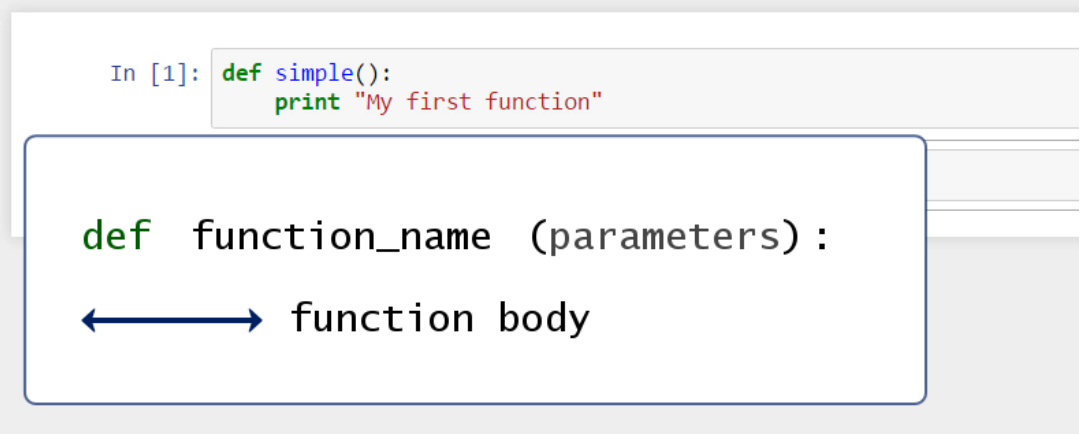
Operator	Description
= =	Verifies the left and right side of an equality are equal
!=	Verifies the left and right side of an equality are not equal
>	Greater than
<	Less than
>=	Greater than or equal to
<=	Less than or equal to
+	Sum (plus)
-	Subtraction (minus)
/	Division
*	Multiplication
**	Superpower

Defining a Function in Python

- Write def at the beginning of the line. Def is neither a command nor a function. It is a **keyword**. To indicate this, Jupyter will automatically change its font color to green.

1. Type **name of the function.**
2. Add a pair of **parentheses**.
3. Place **parameters** of the function if it requires you to have any. It is possible to have a function with zero parameters.
4. Put a **colon** after the name of the function.

- Since it is inconvenient to continue on the same line when the function becomes longer, it is much better to build the habit of laying the instructions on a new line, with an **indent** again.



Indexing

Is it possible to extract the letter “d”?
Yes, *by using square brackets*. Specify the position of the letter we would like to be extracted.

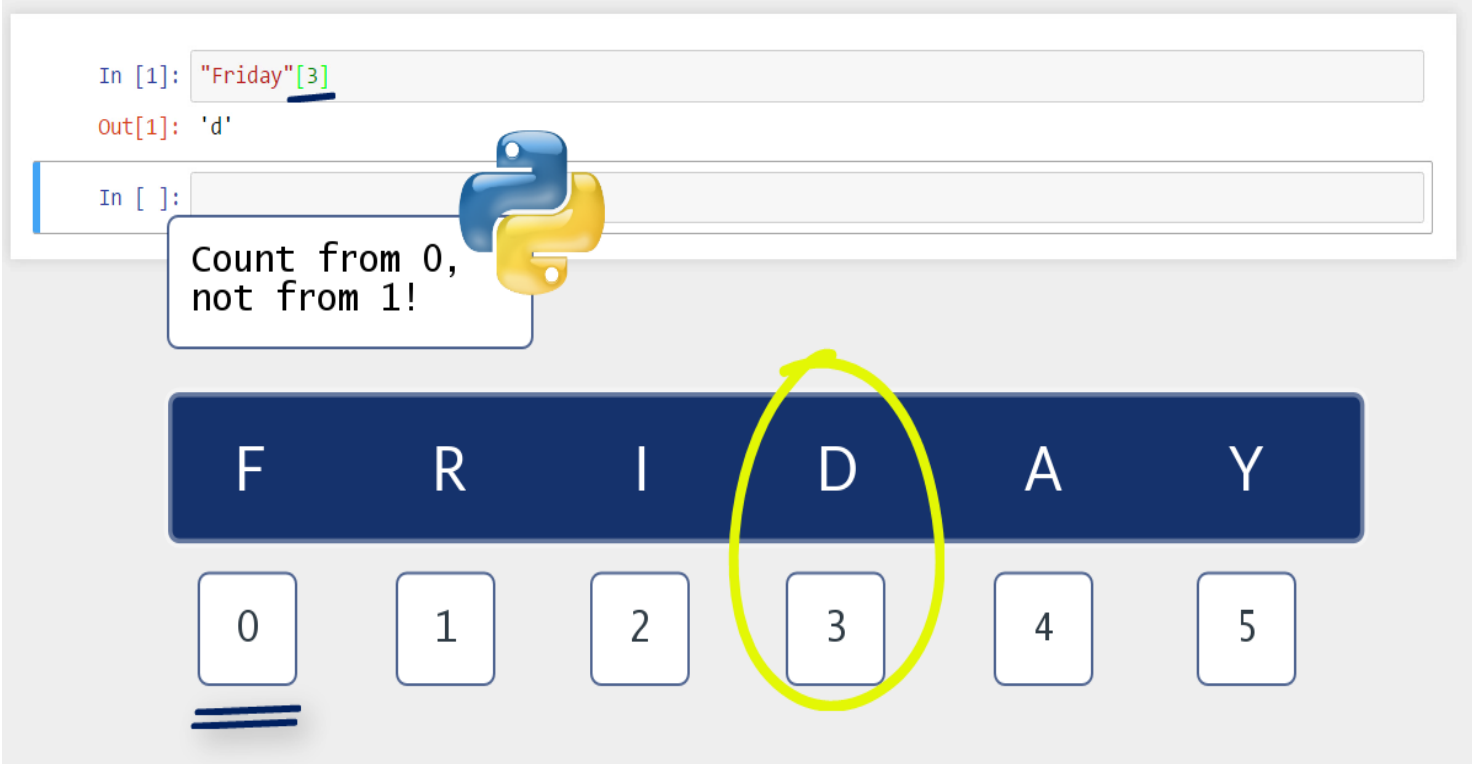
Note:
Make sure you don’t mistake brackets for parentheses or braces:
parentheses – () brackets – [] braces– {}

```
In [ ]: "Friday"[]
```

```
"Name_of_variable"[index_of_element]
```

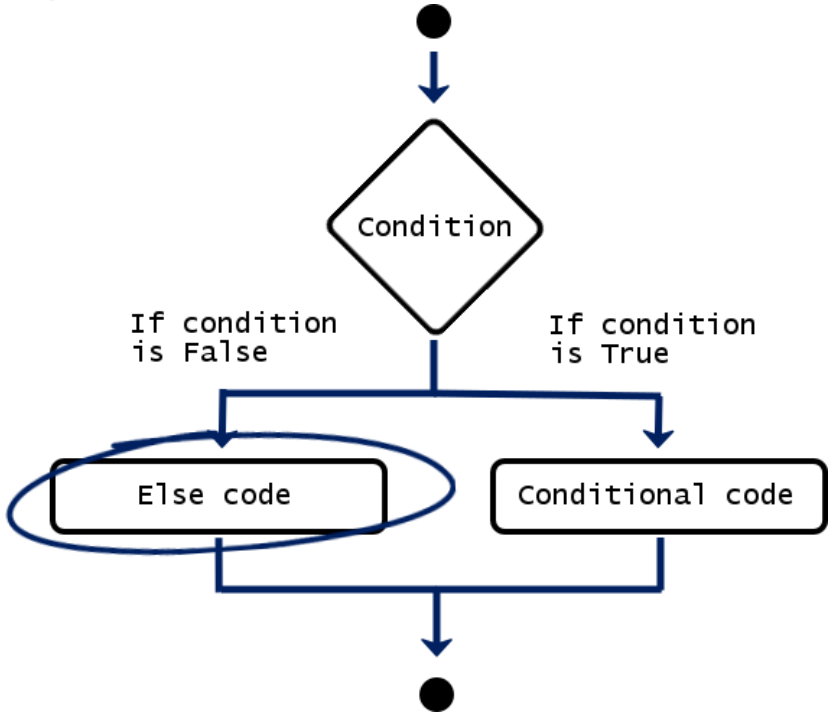
Indexing

- *Note:* in Python, we count from 0, not from 1!
- 0, 1, 2, 3, 4, and so on.
- That's why I'll ask for the 4th letter, 'd', by writing 3 here.



Add an ELSE statement

- if the condition is false, result is *else code*.
- Whether the initial condition is satisfied, we will get to the end point, so the computer has concluded the entire operation and is ready to execute a new one.



Else if, for Brief - ELIF

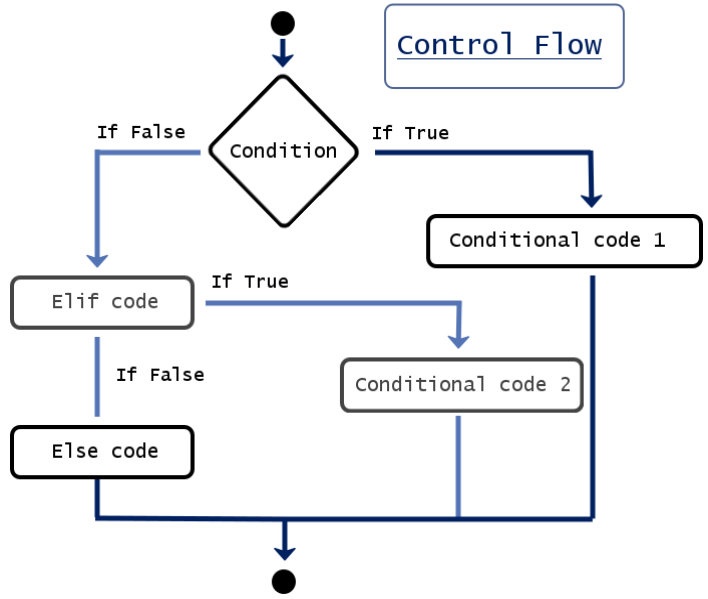
If y is not greater than 5, the computer will think: “else if y is less than 5”, written “elif y is less than 5”, then I will print out “Less”.

```
In [1]: def compare_to_five(y):
            if y > 5:
                return "Greater"
            elif y < 5:
                return "Less"
            else:
                return "Equal"
```

Else if, for Brief - ELIF

- Computer always reads your commands from top to bottom. This is something like the flow of the logical thought of the computer, the way the computer thinks – step by step, executing the steps in a rigid order.

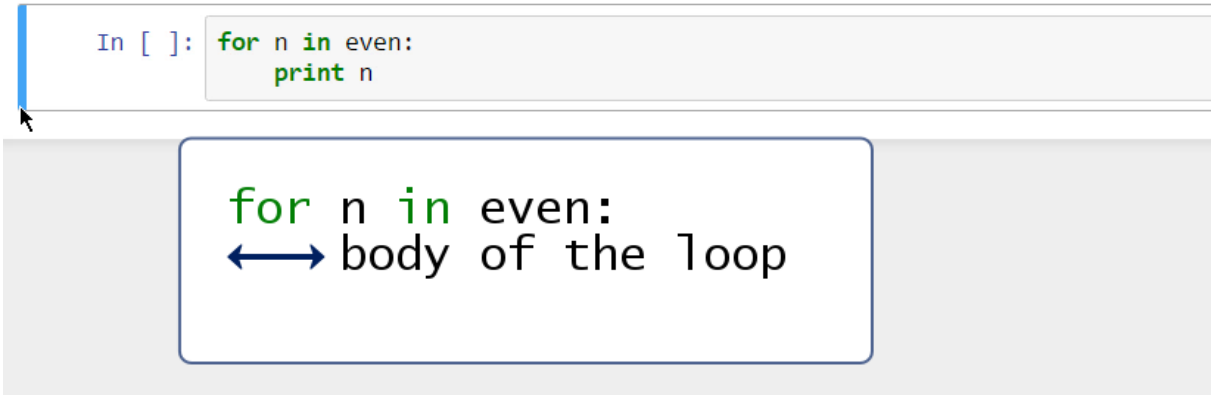
- When it works with a conditional statement, the computer's task will be to execute a specific command once a certain condition has been satisfied. It will read commands from the if- statement at the top, through the elif-statements in the middle, to the else- statement at the end. The first moment the machine finds a satisfied condition, it will print the respective output and will execute no other part of the code from this conditional.



For Loops

- **Iteration** is a fundamental building block of all programs. It is the ability to execute a certain code repeatedly.
- The list "even" contains all the even numbers from 0 to 20. "for n in even", colon, which would mean *for every* element n in the
- list "*even*", do the following: print that element.

- The command in the loop body is performed once *for each* element in the *even* list.



While Loops and Incrementing

The while loop in Python is used to iterate over a block of code as long as the test expression (condition) is true

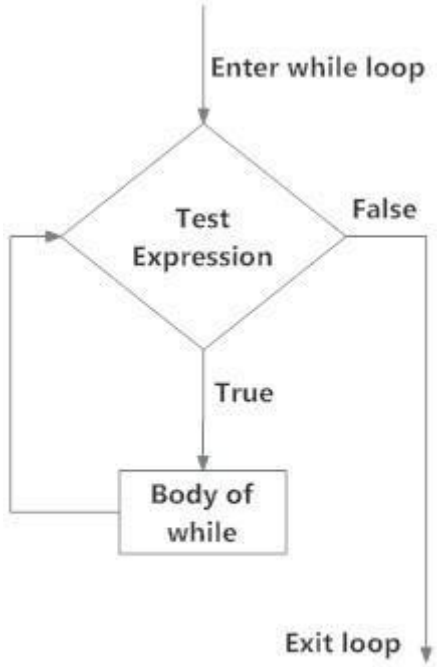


Fig: operation of while loop

Thank you!

Data Science
Academy