

# Data Analysis with Python

Full tutorial for beginners

# About this tutorial

1. What is Data Analysis
2. [Real example Data Analysis with Python](#)
3. [How to use Jupyter Notebooks](#)
4. [Intro to NumPy \(exercises included\)](#)
5. [Intro to Pandas \(exercises included\)](#)
6. [Data Cleaning](#)
7. Reading Data SQL, CSVs, APIs, etc
8. [Python in Under 10 Minutes](#)

# What is Data Analysis?



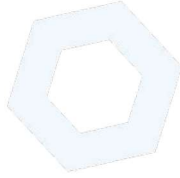


# What is Data Analysis

> *A process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusion and supporting decision-making.*



[Definition by Wikipedia.](#)



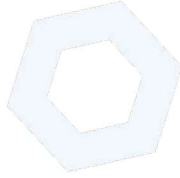


# What is Data Analysis

> A process of *inspecting, cleansing, transforming* and modeling data with the goal of discovering useful information, informing conclusion and supporting decision-making.



[Definition by Wikipedia.](#)





# What is Data Analysis

> A process of inspecting, cleansing, transforming and **modeling data** with the goal of discovering useful information, informing conclusion and supporting decision-making.



[Definition by Wikipedia.](#)



# What is Data Analysis

> A process of inspecting, cleansing, transforming and modeling data with the goal of **discovering useful information**, informing conclusion and supporting decision-making.



[Definition by Wikipedia.](#)



# What is Data Analysis

> A process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, **informing conclusion and supporting decision-making**.

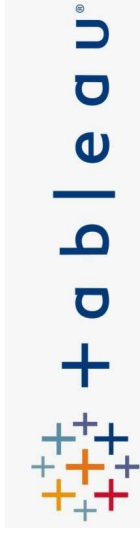


[Definition by Wikipedia.](#)



# Data Analysis Tools

## Auto-managed closed tools



## Programming Languages



## Auto-managed closed tools

👉 Closed Source 🧑

👉 Expensive 💰

👉 Limited 😞

👉 Easy to learn 🧑

## Programming Languages

👉 Open Source 🌟

👉 Free (or very cheap) 🤑

👉 Extremely Powerful 💪

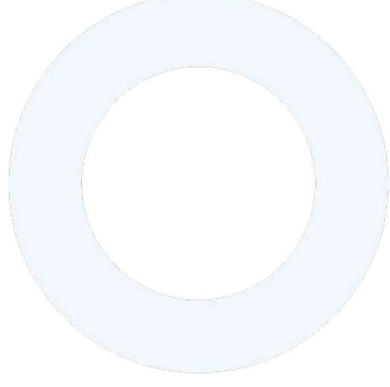
👉 Steep learning curve 🧑

# Why Python for Data Analysis?

# Why Python for Data Analysis?

*Why would we choose Python over R or Julia?*

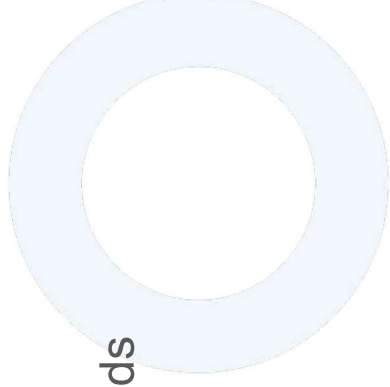
- 👍 very simple and intuitive to learn
- 👍 “correct” language
- 👍 powerful libraries (not just for Data Analysis)
- 👍 free and open source
- 👍 amazing community, docs and conferences



# When to choose R?

*Python, sadly, is not always the answer*

- When R Studio is needed
- When dealing with advanced statistical methods
- When extreme performance is needed



# The Data Analysis Process



## Data Extraction

- SQL
- Scrapping
- File Formats
  - CSV
  - JSON
  - XML
- Consulting APIs
- Buying Data
- Distributed Databases

## Data Cleaning

- Missing values and empty data
- Data imputation
- Incorrect types
- Incorrect or invalid values
- Outliers and non relevant data
- Statistical sanitization

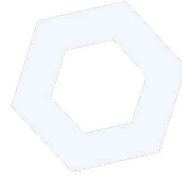
## Data Wrangling

- Hierarchical Data
- Handling categorical data
- Reshaping and transforming structures
- Indexing data for quick access
- Merging, combining and joining data

## Analysis

- Exploration
- Building statistical models
- Visualization and representations
- Correlation vs Causation analysis
- Hypothesis testing
- Statistical analysis reporting
- Reporting

- Building Machine Learning Models
- Feature Engineering
- Moving ML into production
- Building ETL pipelines
- Live dashboard and reporting
- Decision making and real-life tests

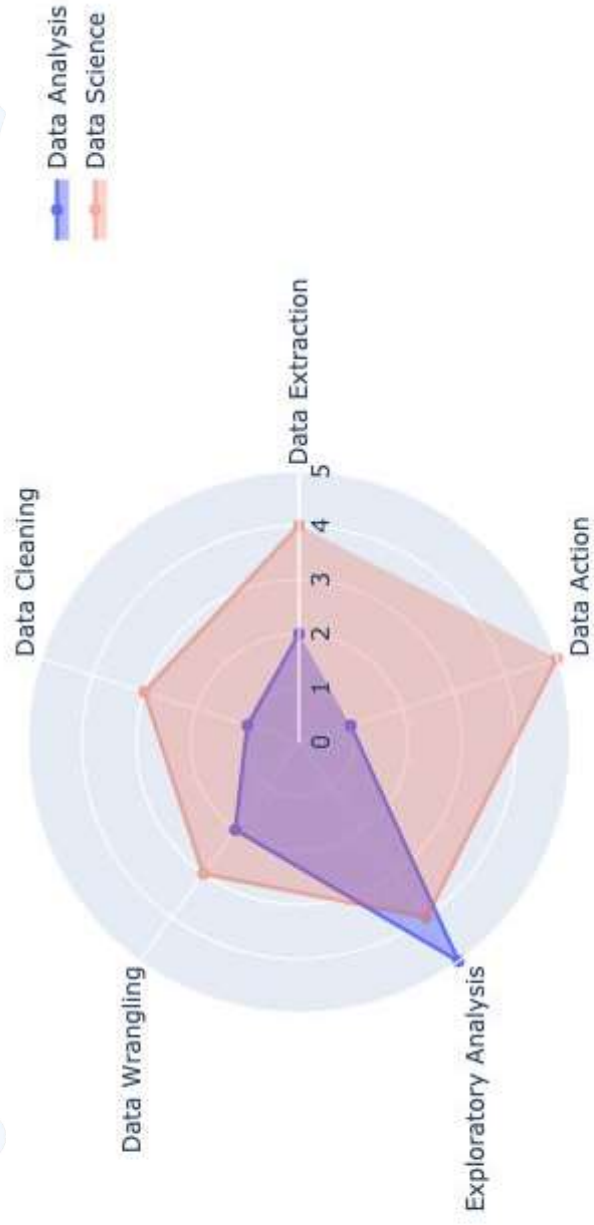




# Data Analysis Vs Data Science

The background of the slide features several faint, light-blue geometric shapes. On the left, there is a solid circle. In the center, there is a large ring or donut shape. On the right, there is a hexagon. These shapes are positioned around the central text.

# The traditional view



# Python & PyData Ecosystem

# The libraries we use...

- [pandas](#): The cornerstone of our Data Analysis job with Python
- [matplotlib](#): The foundational library for visualizations. Other libraries we'll use will be built on top of matplotlib.
- [numpy](#): The numeric library that serves as the foundation of all calculations in Python.
- [seaborn](#): A statistical visualization tool built on top of matplotlib.
- [statsmodels](#): A library with many advanced statistical functions.
- [scipy](#): Advanced scientific computing, including functions for optimization, linear algebra, image processing and much more.
- [scikit-learn](#): The most popular machine learning library for Python (not deep learning)

# How Python Data Analysts Think

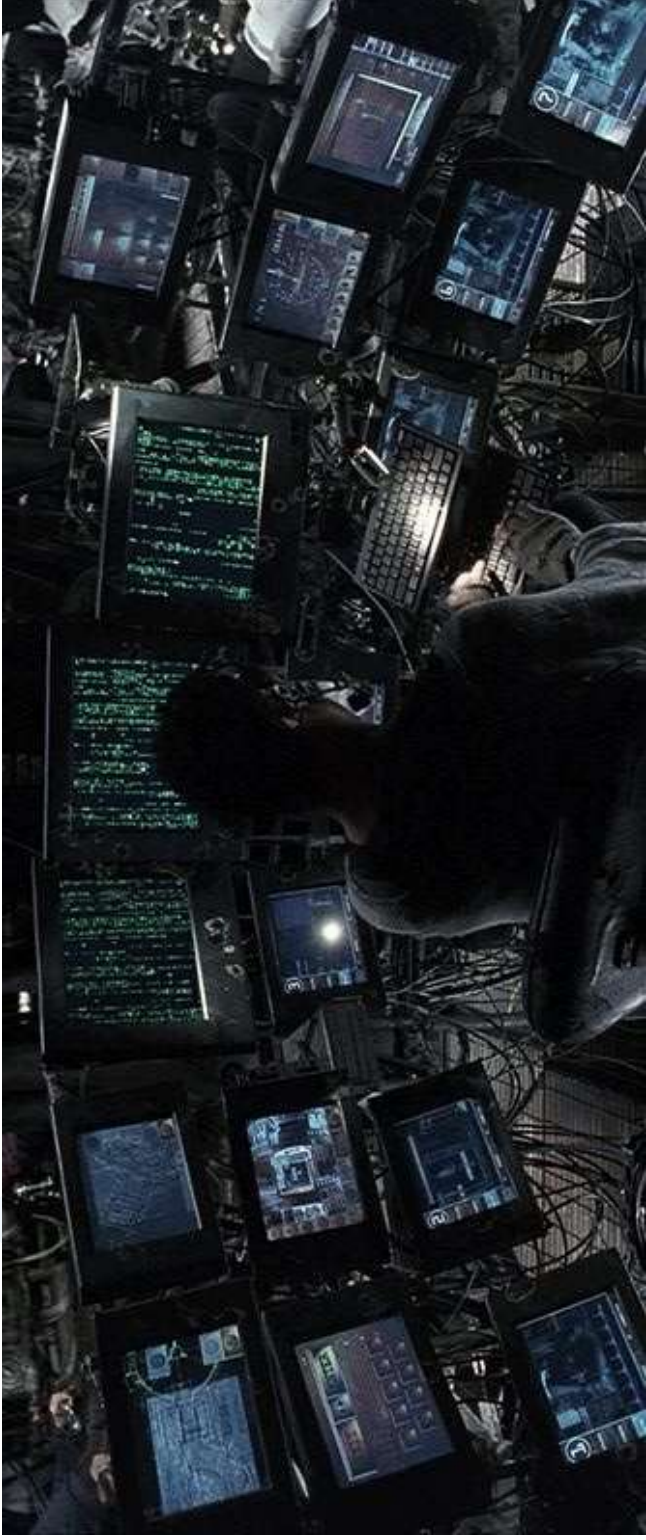
EXCEL, TABLEAU, ETC.

They're all visual tools...

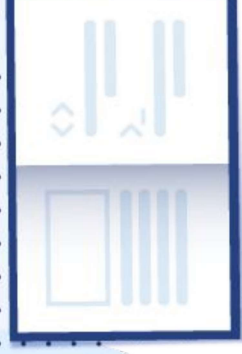
	A	B	C	D	E	F	G	H	I	J	K	L	M
19	http://s-ee9d05-1.sj	Not started	Active	431747565	Draft	Ballinmore							
20	http://s-ee9d05-1.sj	Not started	Active	577174756	Draft	Bay Area							
21	http://s-ee9d05-1.sj	Not started	Active	804502480	Published	Bay Area							
22	http://s-ee9d05-1.sj	Not started	Active	177912527	Draft	Bay Area							
23	http://s-ee9d05-1.sj	Not started	Active	765135561	Final	Bay Area							
24	http://s-ee9d05-1.sj	Not started	Active	28006725	Final	Bay Area							
25	http://s-ee9d05-1.sj	Not started	Active	73320333	Final	Bay Area							
26	http://s-ee9d05-1.sj	Not started	Active	560148690	Final	Bay Area							
27	http://s-ee9d05-1.sj	Not started	Active	820565535	Published	Bay Area							
28	http://s-ee9d05-1.sj	Not started	Active	389715560	Published	Bay Area							
29	http://s-ee9d05-1.sj	Not started	Active	490412451	Published	Bay Area							
30	http://s-ee9d05-1.sj	Not started	Active	847905585	AP-Induction	Bay Area							
31	http://s-ee9d05-1.sj	Not started	Active	102806725	AP-Induction	Bay Area							
32	http://s-ee9d05-1.sj	Not started	Active	609358647	AP-Induction	Bay Area							
33	http://s-ee9d05-1.sj	Not started	Active	670576662	AP-Induction	Bay Area							
34	http://s-ee9d05-1.sj	Not started	Active	644355600	AP-Induction	Bay Area							
35	http://s-ee9d05-1.sj	Not started	Active	435565457	AP-Induction	Bay Area							
36	http://s-ee9d05-1.sj	Not started	Active	54514559	AP-Induction	Bay Area							
37	http://s-ee9d05-1.sj	Not started	Active	838056334	AP-Induction	Bay Area							
38	http://s-ee9d05-1.sj	Not started	Active	398138506	AP-Induction	Bay Area							
39	http://s-ee9d05-1.sj	Not started	Active	888162325	AP-Induction	Bay Area							
40	http://s-ee9d05-1.sj	Not started	Active	494417243	AP-Induction	Bay Area							
41	http://s-ee9d05-1.sj	Not started	Active	304525414	AP-Induction	Bay Area							
42	http://s-ee9d05-1.sj	Not started	Active	353203666	AP-Induction	Bay Area							
43	http://s-ee9d05-1.sj	Not started	Active	296045545	AP-Induction	Bay Area							

RMOTR  
BY LINE

# Thinking like a Python Data Analyst



**And finally,  
why Python?**







>20%

Salary increase for a Data Analyst  
that knows Python and SQL.

# About this tutorial

1. What is Data Analysis
2. **Real Example Data Analysis with Python**
3. How to use Jupyter Notebooks
4. Intro to NumPy ([exercises included](#))
5. Intro to Pandas ([exercises included](#))
6. Data Cleaning
7. Reading Data SQL, CSVs, APIs, etc
8. Python in Under 10 Minutes