17/August 2025

Data Science with Python

Araz Shahkarami

Session 1: Introduction and basic concepts of data science

Araz Shahkarami	17/ August	Session 1

Today’s goals (Session duration: 90 minutes)

- Understand the concept of data science
- Familiarity with the data science project cycle
- Identify data types
- Introduction to tools and workspace
- Perform your first data analysis
- Prepare for the next session

Today's

Data Science Definition: Transform data into intelligent decisions

1 Scientific methods

2 Systematic processes

3 Advanced algorithms

4 Computational systems

Project Name

Practical example: Snapp data analysis

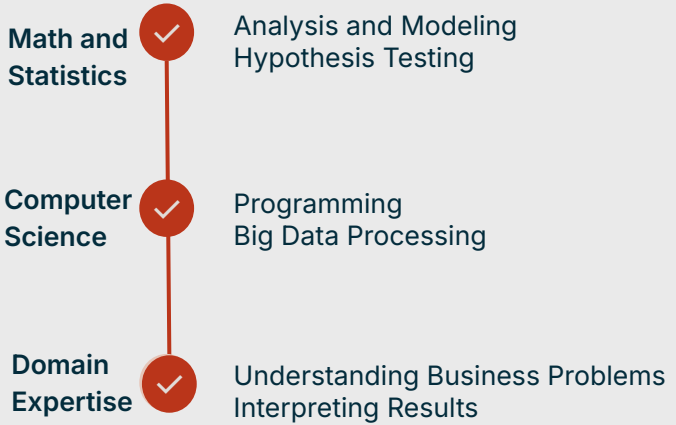
Business Questions:

- ❏ Which areas are in high demand at what times?
- ❏ How to reduce wait times?
- ❏ What is optimal pricing?
- ❏ Which drivers perform best?

Differences from related disciplines(Data Science vs Other Fields)

Feature	Data Science	Traditional Statistics	Business Analytics
Data Volume	Large & Complex	Small Samples	Medium
Goal	Prediction	Hypothesis Testing	Reporting
Tools	Python/R	SPSS/SAS	Excel/BI
Output	Automated Model	Statistical Result	Dashboard

The intersection of three fields:



Data Science Cycle

Exploratory & Preparation Phase	**Problem Definition** Understanding the business/question	**Data Acquisition** Collecting raw data	**Data Preparation** Cleaning, transforming, and structuring data	**Data Analysis** Exploratory analysis, feature engineering
Modeling & Operationalization Phase	**Modeling** Building ML/statistical models	**Evaluation** Testing model performance	**Deployment** Implementing the model in production	**Monitoring** Tracking model performance over time

Exploratory & Preparation Phase

Step one

Problem Definition

Goal:

Translate a business problem into an answerable question

Example:

General problem: "Passengers wait too long"

Specific question: "How can we reduce waiting time from 8 to 5 minutes?"

Tools:

- Stakeholder interviews
- Defining KPIs
- SWOT analysis

Step Two

Data Collection

Data Sources:

Internal: System logs, transactional databases

External: APIs, websites, purchased data

Public: Government, research, open-source

Key Challenges:

- Data quality
- Access and permissions
- Volume and velocity

Step Three

Data Preparation

80% of a project's time is Key Challenges:

Key Activities:

Cleaning: Removing duplicates, handling missing values

Transformation: Changing formats, creating new variables

Integration: Combining data from different sources

Validation: Checking accuracy and logical consistency

Clean data = Reliable results

Step Four

Data Analysis

Extract insights and identify patterns to answer business questions

Key Activities:

Exploratory Analysis (EDA): Summarize data (mean, median, distributions)

Visualize trends (charts, graphs)

Detect outliers and anomalies

Good analysis → Better decisions!

Modeling & Operationalization Phase

Step five

Modeling

Goal:

Develop predictive or prescriptive models to solve the business problem

Key Activities:

Feature Engineering & Selection

Algorithm Selection
Model Training & Tuning
Model Evaluation
Iterative Improvement

Outcome: A deployable model that provides actionable insights or predictions!

Step six

Evaluation

Goal:

Rigorously assess model performance to ensure reliability and business readiness

Step seven

Deployment

Goal: Transition from a prototype to a live system that delivers business value

Step eight

Monitoring

Goal: Ensure sustained accuracy, reliability, and business relevance of deployed models

Data Acquisition Pipeline

ETL/ELT Process

Extract Phase

Source	Common Use Cases	Extraction Challenges
Databases	Transactional data, user records	Schema changes, query performance
Flat Files	Legacy systems, external reports	Format inconsistencies, encoding
APIs	SaaS platforms, microservices	Rate limits, authentication
Web Scraping	Competitive intelligence, leads	Legal/robots.txt, HTML volatility

Transform Phase

Category	Operations	Performance Tip
Cleaning	Null handling, outlier clipping	Use vectorized operations
Normalization	Min-max scaling, z-score	Cache intermediate results
Encoding	One-hot, label, embeddings	Avoid one-hot for high-cardinality
Temporal	Date parts, rolling windows	Pre-compute common aggregations
Text	Tokenization, TF-IDF	Pipeline caching

Data Acquisition Pipeline

ETL/ELT Process

Load Phase

Storage Type	Best For	Performance
Data Warehouses	Structured analytics, BI reporting	High
Data Lakes	Raw/unstructured data, ML pipelines	Medium
NoSQL Databases	High-velocity unstructured data	Varies
Streaming Platforms	Real-time processing	Ultra-low latency

Three Primary Data Types

Structured Data	• Tabular format with explicit rows/columns •	**CSV**	Analytics exports - Pandas, Excel
		SQL	Transactional systems - PostgreSQL, MySQL
Semi-Structured Data	• Self-describing but on-tabular • Flexible schema nested/hierarchical)	**JSON**	Key-value pairs - Pandas
		XML	Tag-based- XPath, BeautifulSoup
Unstructured Data	• No inherent organizational model • Requires preprocessing for analysis	**Text/ Images / Audio**	

Thanks for participating

Why Python is the Best Choice for Data Projects

1

Simple & Readable

Natural Syntax → Closer to human language than most programming languages

2

Rich Ecosystem

Pandas-Data manipulation
NumPy-Numerical computing
Matplotlib-Visualization
Scikit-learn-Machine Learning
TensorFlow-Deep Learning

3

Massive Community Support

GitHub: 500K+ Python repositories

Real-World Impact: Used by 90% of Fortune 500 companies

4

Open Source & Free

Zero Licensing Costs
Cross-Platform

Perform your first data analysis

📁 ساختار مخزن

```
data-science-python-course/
│
├── sessions/                    # مطالب هر جلسه
│   ├── session-01/              # جلسه اول
│   │   ├── slides/              # اسلایدها
│   │   ├── notebooks/           # Jupyter Notebooks
│   │   ├── data/                # داده‌های نمونه
│   │   └── exercises/           # تمرین‌ها
│   ├── session-02/
│   └── ...
│
├── projects/                    # پروژه‌های دوره
│   ├── mini-projects/           # پروژه‌های کوچک
│   └── final-project/           # پروژه نهایی
│
├── datasets/                    # مجموعه داده‌های دوره
│   ├── raw/                     # داده‌های خام
│   ├── processed/               # داده‌های پردازش شده
│   └── external/                # داده‌های خارجی
│
├── resources/                   # منابع و مراجع
│   ├── books/                   # کتاب‌ها و مقالات
│   ├── cheatsheets/             # برگه‌های کمک سریع
│   └── links.md                 # لینک‌های مفید
│
├── setup/                       # راه‌اندازی محیط
│   ├── installation.md          # راهنمای نصب
│   └── environment.yml          # محیط conda
│
```

7 Reasons to Become a Data Scientist

1. High Demand & Salary
2. Future-Proof Career
3. Impactful Work
4. Intellectual Stimulation
5. Remote Flexibility
6. Low Barrier to Entry
7. Creativity Meets Logic

Thank you

SCAN ME

If you have any questions, contact:
araz.shah@gmail.com
https://araz.me