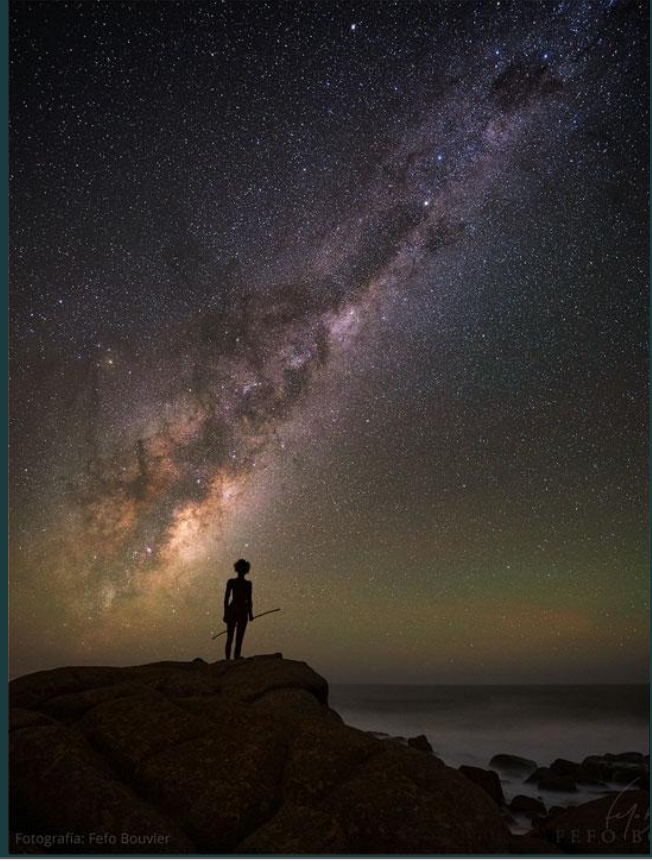
BRICS & IDIA: Data Analytics Training

Statistical Analysis

Narusha Isaacs-Klein

21 May



BRICS ASTRONOMY

Ural Federal University
named after the first President of Russia B.N.Yeltsin

مولانا آزاد نیشنل اُردو یونیورسٹی
मौलाना आज़ाद नेशनल उर्दू यूनिवर्सिटी
MAULANA AZAD NATIONAL URDU UNIVERSITY
(Accredited Grade "A+" by NAAC)

IDIA

Content *overview*

Why Python?

Open-source ecosystem: Collaborative, free, and constantly evolving.

Versatility: From data cleaning to machine learning.

Integration: Libraries work seamlessly together.

Community: Widely adopted in academia and industry.

Key Tools Covered:

Pandas: Data manipulation.

NumPy: Numerical computing.

SciPy: Advanced statistics.

Statsmodels: Statistical modeling.

NumPy – The Foundation

What is NumPy?

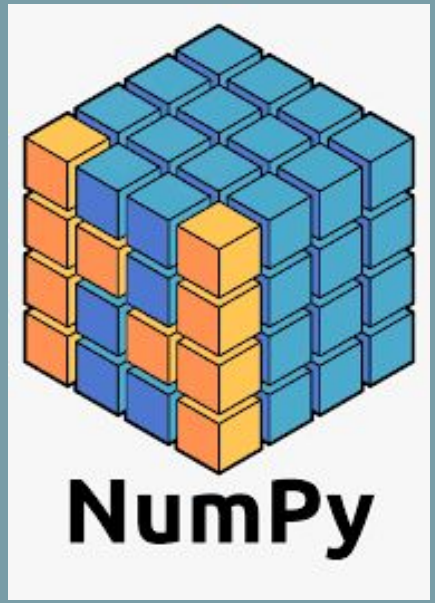
- **Core library** for numerical computing.
- Provides **n-dimensional arrays** (vectors, matrices).

Why It Matters:

- **Efficiency:** Optimized for speed (written in C/Fortran).
- **Universal:** Underpins nearly all scientific Python libraries.
- **Math operations:** Linear algebra, Fourier transforms, random number generation.

Use Cases:

- Handling large datasets (e.g., astronomical measurements).
- Matrix operations for simulations.



Pandas – Data Manipulation

What is Pandas?

- **DataFrame-centric library** for structured data.
- Built on NumPy, optimized for tabular data.

Key Features:

- **Load data:** CSV, Excel, SQL, JSON.
- **Clean data:** Handle missing values, duplicates.
- **Filter/transform:** Subset rows, calculate new columns.
- **Merge datasets:** Combine data from multiple sources.

Why It Matters:

- Simplifies **data wrangling** (80% of data science work).
- Enables reproducible workflows.

Use Cases:

- Cleaning messy survey data.
- Aggregating experimental results.

SciPy – Scientific Computing

What is SciPy?

- **Advanced toolbox** for science and engineering.
- Built on NumPy, with specialized submodules.

Key Features:

- Statistics: Distributions, hypothesis testing (t-tests, ANOVA).
- Optimization: Minimize/maximize functions.
- Signal processing: Filtering, Fourier analysis.
- Integration: Solve differential equations.

Why It Matters:

- Provides peer-reviewed algorithms (no reinventing the wheel)
- Bridges gap between theory and application.

Statsmodels – Statistical Modeling

What is Statsmodels?

- **Library for estimating and testing statistical models.**

Key Features:

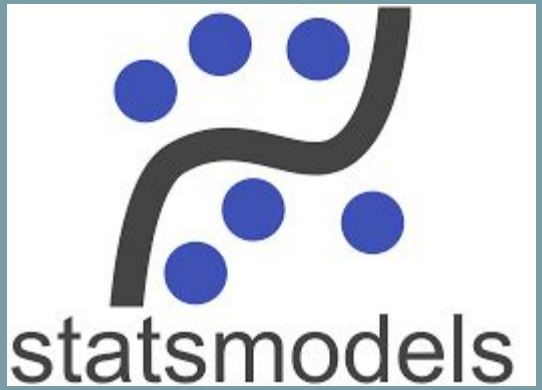
- **Regression:** Linear, logistic, time-series.
- **Hypothesis tests:** T-tests, chi-square, ANOVA.
- **Diagnostics:** Model accuracy, residual analysis.

Why It Matters:

- Provides **interpretable results** (p-values, confidence intervals).
- Essential for **quantitative research.**

Use Cases:

- Predicting galaxy brightness from distance.
- Validating experimental hypotheses.



Workflow Integration

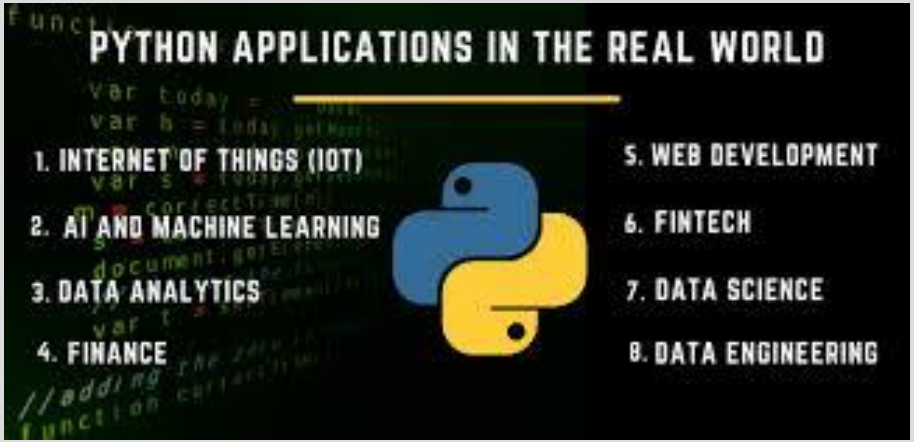
Typical Data Analysis Pipeline:

1. **Load & Clean:** Pandas.
2. **Transform:** NumPy/Pandas.
3. **Analyze:** SciPy/Statsmodels.
4. **Model:** Statsmodels.

Why Integration Matters:

- **End-to-end analysis** in one ecosystem.
- Reduces errors from switching tools.

Real-World Applications



Astronomy

- Filtering Gaia star catalogs (Pandas).
- Testing if bright stars move faster (SciPy t-test).
- Modeling brightness vs. distance (Statsmodels).

Biology

- Cleaning genomic data (Pandas).
- Comparing gene expression (SciPy ANOVA).

Finance

- Predicting stock trends (Statsmodels time-series).

Conclusion

Python's Scientific Ecosystem:

- **NumPy:** Fast arrays.
- **Pandas:** Clean data.
- **SciPy:** Advanced stats.
- **Statsmodels:** Modeling.

Why Learn These Tools?

- **Industry standard** in data-driven fields.
- **Empowers researchers** to focus on science, not coding.

Ural Federal University
named after the first President of Russia B.N.Yeltsin

مولانا آزاد نیشنل اُردو یونیورسٹی
मौलाना आज़ाद नेशनल उर्दू यूनिवर्सिटी
MAULANA AZAD NATIONAL URDU UNIVERSITY
(Accredited Grade "A+" by NAAC)

Thank you

Q & A