Introduction to Pandas

Data Analysis with Python

Analyzing 24,000 Irish Traditional Tunes

What is Pandas? 🐼

Pandas is Python's most popular library for data analysis and manipulation

- Built on top of NumPy
- Provides easy-to-use data structures
- Essential for data science workflows
- Used by millions of data professionals

> "Pandas makes working with data feel natural and intuitive"

Why Learn Pandas?

- **Industry Standard** - Used in data science, finance, research, and more
- **Powerful** - Handle datasets with millions of rows
- **Versatile** - Read/write Excel, CSV, SQL, JSON, and more
- **Fast** - Optimized C code under the hood
- **Open Source** - Free and constantly improving

Core Data Structures

DataFrame - The main workhorse

- 2D table with rows and columns
- Like an Excel spreadsheet, but programmable
- Each column can have different data types

Series - A single column

- 1D array with labels
- Think of it as one column from a DataFrame

Our Dataset: Irish Traditional Music 🎵

We'll be working with a real dataset of **24,000 traditional Irish tunes**

Source: The Session (thesession.org) - world's largest collection of Irish traditional music

Format: ABC notation - a text-based music notation system

Dataset Features

What's in our data?

Column	Description	Example
`id`	Unique identifier	1, 2, 3...
`title`	Tune name	"Cooley's"
`tune_type`	Category	reel, jig, hornpipe
`key_sig`	Musical key	Emin, Dmaj, Gmaj
`time_sig`	Time signature	4/4, 6/8, 9/8
`downloaded`	Popularity metric	4432, 2721...
`notation`	ABC music code	Full tune notation

Sample Data

id	title	tune_type	key_sig	downloads
1	Cooley's	reel	Emin	4432
2	Bucks Of Oranmore	reel	Dmaj	2721
3	Boil The Breakfast	reel	Gmaj	812
10	The Butterfly	slip jig	Emin	2767
12	Cliffs Of Moher	jig	Ador	2850

Types of Tunes in Irish Music

Reels (4/4 time)

Fast dance tunes, most common type

Jigs (6/8 time)

Bouncy, lilting rhythm

Hornpipes (4/4 time)

Slower than reels, swung rhythm

Slip Jigs (9/8 time)

Graceful, flowing tunes

Your Learning Journey

Week 1-2: **Fundamentals**

Load data, explore, select, and filter

Week 3-4: **Analysis**

Group, aggregate, sort, and rank data

Week 5-6: **Advanced**

String operations, data cleaning, visualization

Week 7-8: **Project**

Comprehensive analysis and presentation

Essential Pandas Operations

```
import pandas as pd

# Load data
df = pd.read_csv('irish_music.csv')

# Explore
df.head()           # First 5 rows
df.info()           # Column types and info
df.describe()       # Statistical summary
```

Essential Pandas Operations (cont.)

```
# Select columns
df['title']                     # Single column
df[['title', 'tune_type']]      # Multiple columns

# Filter rows
df[df['tune_type'] == 'reel']  # Only reels
df[df['downloaded'] > 1000]    # Popular tunes

# Sort
df.sort_values('downloaded', ascending=False)
```

Essential Pandas Operations (cont.)

```
# Group and aggregate
df.groupby('tune_type')['downloaded'].mean()

# Count values
df['tune_type'].value_counts()

# Handle missing data
df.dropna()             # Remove missing
df.fillna(value)        # Fill missing
```

Questions We'll Answer

1. What's the most popular tune type?
2. Which key signature is most common?
3. Is there a relationship between key and popularity?
4. What makes a tune popular?
5. How do different tune types compare?
6. Are certain time signatures more popular?

Real-World Skills You'll Gain

✅ **Data Cleaning** - Handle messy, real-world data
✅ **Exploratory Analysis** - Find patterns and insights
✅ **Statistical Thinking** - Calculate and interpret metrics
✅ **Visualization** - Create meaningful charts
✅ **Problem Solving** - Answer questions with data
✅ **Communication** - Present findings clearly

Why This Dataset?

Real-World

Actual data from a live website, not sanitized classroom examples

Interesting

Cultural heritage, music, and patterns to discover

Complete

Has everything: numbers, text, categories, missing values

Scalable

24,000 rows - big enough to matter, small enough to explore

Your Final Project

You'll conduct a **comprehensive analysis** answering:

- What patterns exist in Irish traditional music?
- Which tunes and types are most popular?
- How do musical characteristics relate to popularity?
- What insights can inform musicians and researchers?

Deliverable: Jupyter notebook with code, visualizations, and insights

Getting Started Today

Lesson 1: Load and Explore

```
import pandas as pd

# Your first pandas code!
df = pd.read_csv('irish_music.csv')

print(f"We have {len(df)} tunes!")
print(f"Columns: {list(df.columns)}")
print(f"\nFirst tune:")
print(df.iloc[0])
```

Installing Pandas

Option 1: Anaconda (Recommended)

```
# Pandas comes pre-installed with Anaconda
conda install pandas
```

Option 2: pip

```
pip install pandas
```

Verify Installation

```
import pandas as pd
print(pd.__version__)
```

Resources for Learning

📚 **Official Documentation:** pandas.pydata.org
📚 **Our Course Materials:** Lessons 1-8 + exercises
📚 **Practice Dataset:** irish_music.csv (provided)
📚 **Community:** Stack Overflow, Reddit r/pandas

💡 **Office Hours:** [Your schedule here]
💡 **Discussion Forum:** [Your platform here]

Course Structure

Lectures (2x per week)

New concepts and live coding demos

Lab Sessions (1x per week)

Hands-on practice with instructor support

Homework

Exercises from each lesson

Final Project

Due Week 8 - comprehensive analysis

Tips for Success

1. **Practice Daily** - Even 30 minutes helps
2. **Type the Code** - Don't just read it
3. **Break Things** - Learn from errors
4. **Ask Questions** - No question is too small
5. **Explore** - Try things not in the exercises
6. **Collaborate** - Discuss with classmates
7. **Have Fun** - Data analysis is creative work!

Let's Look at Real Data

```
import pandas as pd

df = pd.read_csv('irish_music.csv')

# What's the most popular tune?
most_popular = df.nlargest(1, 'downloaded')
print(most_popular[['title', 'tune_type', 'downloaded']])

# Output:
#   title     tune_type  downloaded
# 9 Banish Misfortune  jig    4292
```

Interesting! A jig, not a reel, is most popular 🎵

Common Pandas Patterns

You'll use these constantly:

```
# Load → Filter → Group → Visualize
df = pd.read_csv('data.csv')
reels = df[df['tune_type'] == 'reel']
avg_by_key = reels.groupby('key_sig')['downloaded'].mean()
avg_by_key.plot(kind='bar')
```

This workflow applies to ANY dataset!

Beyond This Course

Pandas is just the beginning:

Next Steps:

- NumPy (numerical computing)
- Matplotlib/Seaborn (advanced visualization)
- Scikit-learn (machine learning)
- SQL (database queries)
- Tableau/Power BI (business intelligence)

All of these build on pandas!

Assessment Breakdown

Component	Weight	Description
Homework	40%	8 weekly assignments
Labs	20%	Participation and exercises
Final Project	30%	Comprehensive analysis
Quizzes	10%	Short knowledge checks

Example: First Analysis

Question: How many tune types are in our dataset?

```
import pandas as pd

df = pd.read_csv('irish_music.csv')
tune_types = df['tune_type'].value_counts()

print(tune_types)
```

Output:

```
reel         12450
jig           7832
hornpipe      2234
slip jig       892
...
```

What Makes This Course Different?

Traditional Approach

- Toy datasets (iris flowers, titanic)
- Artificial examples
- Predictable patterns

Our Approach

- Real cultural data
- Authentic questions
- Surprising discoveries
- Meaningful insights

The Power of Pandas

Before Pandas:

```
# Reading CSV manually - 50+ lines of code
# Filtering data - loops and conditions
# Grouping - complex dictionary logic
# Plotting - connecting to other libraries
```

With Pandas:

```
df = pd.read_csv('data.csv')
df[df['type'] == 'reel'].groupby('key')['downloads'].mean().plot()
```

One line does it all!

Questions to Think About

As we begin, consider:

- What makes some tunes more popular than others?
- Do certain keys sound "better" to listeners?
- How has Irish music evolved over time?
- Can we predict a tune's popularity?
- What cultural factors influence tune types?

Data can help answer these questions!

Your First Assignment

Due: Next Class

1. Install pandas and verify it works
2. Download the `irish_music.csv` file
3. Load the data and run `df.info()`
4. Find and print:
    - Total number of tunes
    - Number of unique tune types
    - The tune with the most downloads
5. Write 3 questions you want to answer with this data

Let's Get Started! 🚀

Today's Lab:

- Install pandas
- Load the dataset
- Run basic exploration commands
- Complete Exercise 1.1 - 1.3

Remember:

- Every expert was once a beginner
- Mistakes are part of learning
- The pandas community is friendly and helpful

Questions?

Ready to dive into data analysis with pandas!

📧 [your.email@university.edu]
💬 Office Hours: [Your schedule]
🔗 Course Website: [Your URL]

Thank You!

Next Lesson: Loading and Exploring Data

See you next class! 🐼📊