

Meet Pandas

FinTech
Lesson 3.2



Class Objectives

By the end of today's class, you will be able to:



Describe the benefits of Pandas over spreadsheets to manipulate data on financial use cases.



Explain what a DataFrame is and how it differs from a series.



Create DataFrames from CSV files and use basic commands to manipulate them.



Clean data using built-in commands of DataFrames.



Manipulate data using DataFrame indexes.



Describe the basic theory and calculations of returns using Pandas.



Create basic data visualizations with Pandas' built-in plotting functions.

Why Pandas?

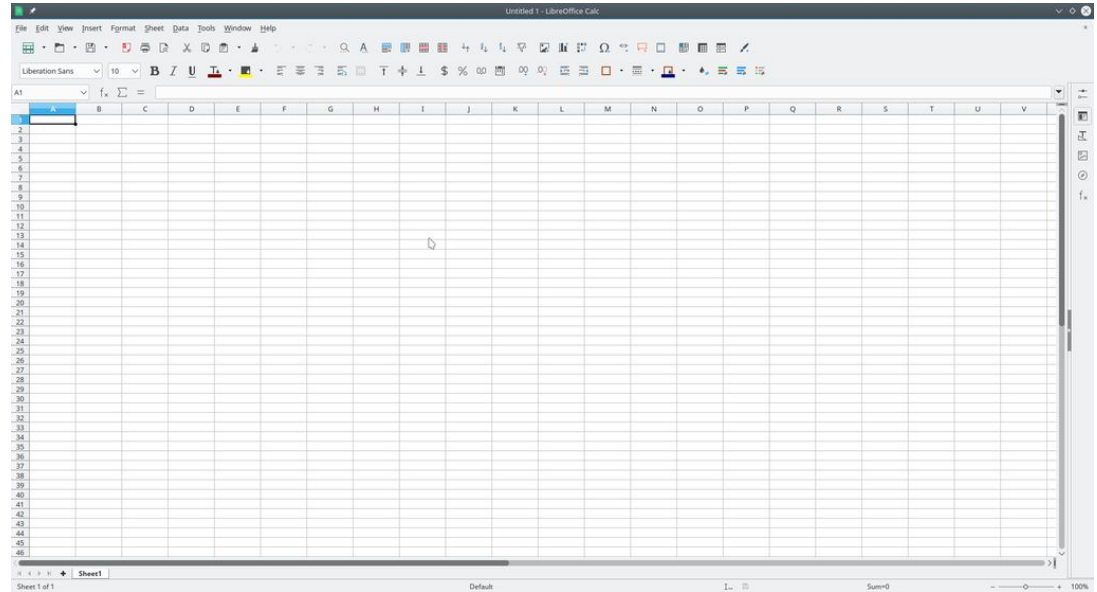
Spreadsheets Are AWESOME.

The Rise of Spreadsheets



A screenshot of a terminal window with a green header bar. The header bar contains the text 'C11 (L) TOTAL' on the left and 'C1 25' on the right. The terminal displays a spreadsheet with the following data:

	A	B	C	D
	ITEM	NO.	UNIT	COST
1	MUCK RAKE	43	12.95	556.85
2	BUZZ CUT	15	6.75	101.25
3	TOE TONER	250	49.95	12487.50
4	EYE SNUFF	2	4.95	9.90
			SUBTOTAL	13155.50
			9.75% TAX	1282.66
			TOTAL	14438.16



The Pain of Using Spreadsheets

Have you ever felt like this?



The Pain of Using Spreadsheets

Spreadsheets are great, but they can become a pain when you are dealing with complex data:

- Calculations are often not reproducible.
- Data can be overwritten in the spreadsheet.
- Data cleaning may overwrite the original data.
- Sharing spreadsheets is difficult.
- Combining data from multiple spreadsheets is difficult.
- Spreadsheets often demonstrate poor performance.
- Large datasets are not handled well.

65421 stop	incidence	minnes	ota	Francis Teller	5 months
464654 ok	paid	Chicag		Mike Michaelson	8 months
7824 ok	paid	minnes	n	James Kowalscky	4 months
		Chicag		James Turner	7 months
		Chicag		Emma Smith	3 months
		Chicag		Bryce Teller	7 months
		minnes	ota	John van Persie	8 months
		wiscon:	in	Jordan Tate	5 months
		Chicag		Mindy Spencer	8 months
		michig:	n	Michael Jones	4 months
		Chicag		Terry Flanagan	5 months
		wiscon:	in	Thomas Tursen	7 months
		minnes	ota	Treapwodd Mint	3 months
		Chicag		Tim Berenger	6 months
		michig:	n	Jonas Stone	8 months
		Chicag		Tobby Rapaport	6 months
		Chicag		Peter Bayega	5 months
		minnes	ota	Javier Ortiz	7 months
		Chicag		James Rodrigues	8 months
		michig:	n	Timmy O'Flanagan	3 months
		minnes	ota	Mike Mcfly	4 months
		Chicag		Jeremiah Tully	6 months
		Chicag		Clemence Sanchez	8 months
		michig:	n	Timmy Richard Lee	6 months



Pandas to the Rescue

Fortunately, we have Pandas to help us mung data on Python.



The Origins of Pandas

- [Pandas](#) is one of the most powerful open source libraries in Python for analyzing and manipulating data.
- This library was born on 2008 at [AQR Capital](#) when [Wes McKinney](#) was looking for a solution to offer a high-performance and flexible tool to perform quantitative analysis on financial data.
- Etymology: panel data structures

Why Pandas is Great

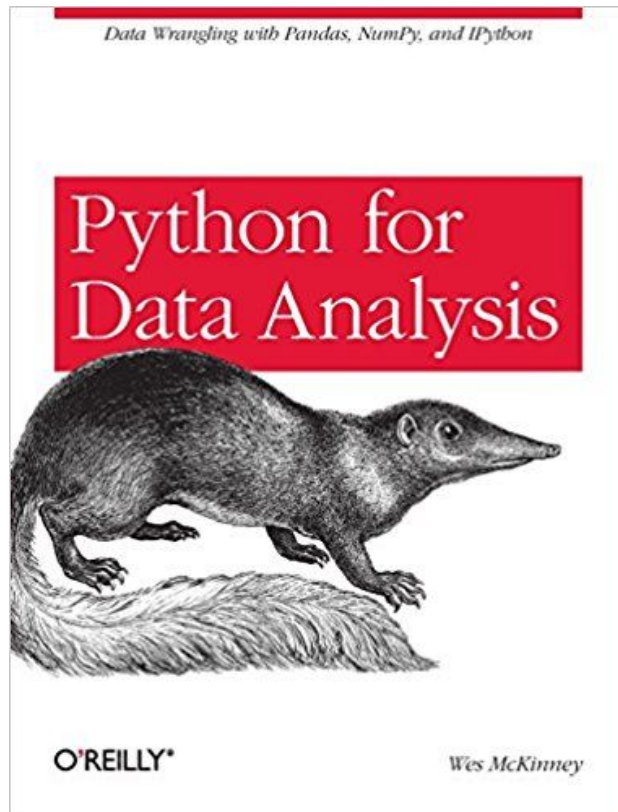
- Python + Pandas = the perfect combination for small experiments or for implementing large-scale production systems to analyze data and make smarter decisions.
- High-performance data structures:
 - Series (1D labeled vectors)
 - DataFrame (2D structures similar to spreadsheets)
 - Panel (Collection of DataFrames as 3D labeled arrays)
- Built-in time series functionality, which is a must for financial and quants analysis



Resources for Learning More About Pandas

- Official website: <https://pandas.pydata.org/>
- Pandas on GitHub: <http://github.com/pydata/pandas>
- *Python for Data Analysis* by Wes McKinney

Python for Data Analysis
by Wes McKinney
(O'Reilly Media, 2017)





There is life beyond Excel
to analyze data. Let's find
the path!



Activity: Reading Stock Data from a CSV File

In this activity, you will get hands-on experience reading CSV files into Pandas. You will use the `read_csv` function, sample data with the `head` function, and create DataFrames with specified column names.

(Instructions sent via Slack.)

Suggested Time:
10 Minutes





Time's Up! Let's Review.



Activity: Spring Cleaning

In this activity, you will be given Harold's stock data and are asked to perform a series of data quality checks to ensure the data is ready for analytical use. The objective of the assignment is for you to learn how to cleanse data using Pandas native functions (`count`, `value_counts`, `isnull`, `sum`, `mean`, `contains`, and `replace`).

(Instructions sent via Slack.)

Suggested Time:
15 Minutes





Time's Up! Let's Review.



Activity: Three-Year Loans

This activity will test your DataFrame indexing skills. You will slice and dice the `loans.csv` data to generate insightful answers regarding three-year loan customers.

(Instructions sent via Slack.)

Suggested Time:
15 Minutes





Time's Up! Let's Review.



Activity: Market Analysis

In this activity, you will create three different charts using Pandas: pie chart, bar chart, and scatter plot.

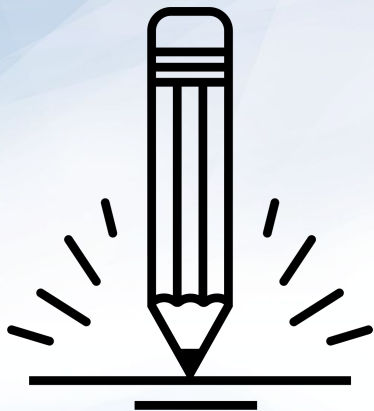
(Instructions sent via Slack.)

Suggested Time:
15 Minutes





Time's Up! Let's Review.



Activity:

Returns Over Date Ranges

In this activity, you will work analyze the last 10 years of historical price data for AMD and plot the daily returns over the last 1-, 3-, 5-, and 10-year time periods.

(Instructions sent via Slack.)

Suggested Time:
15 Minutes





Time's Up! Let's Review.



Decompress