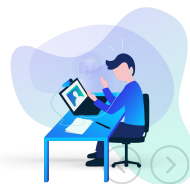


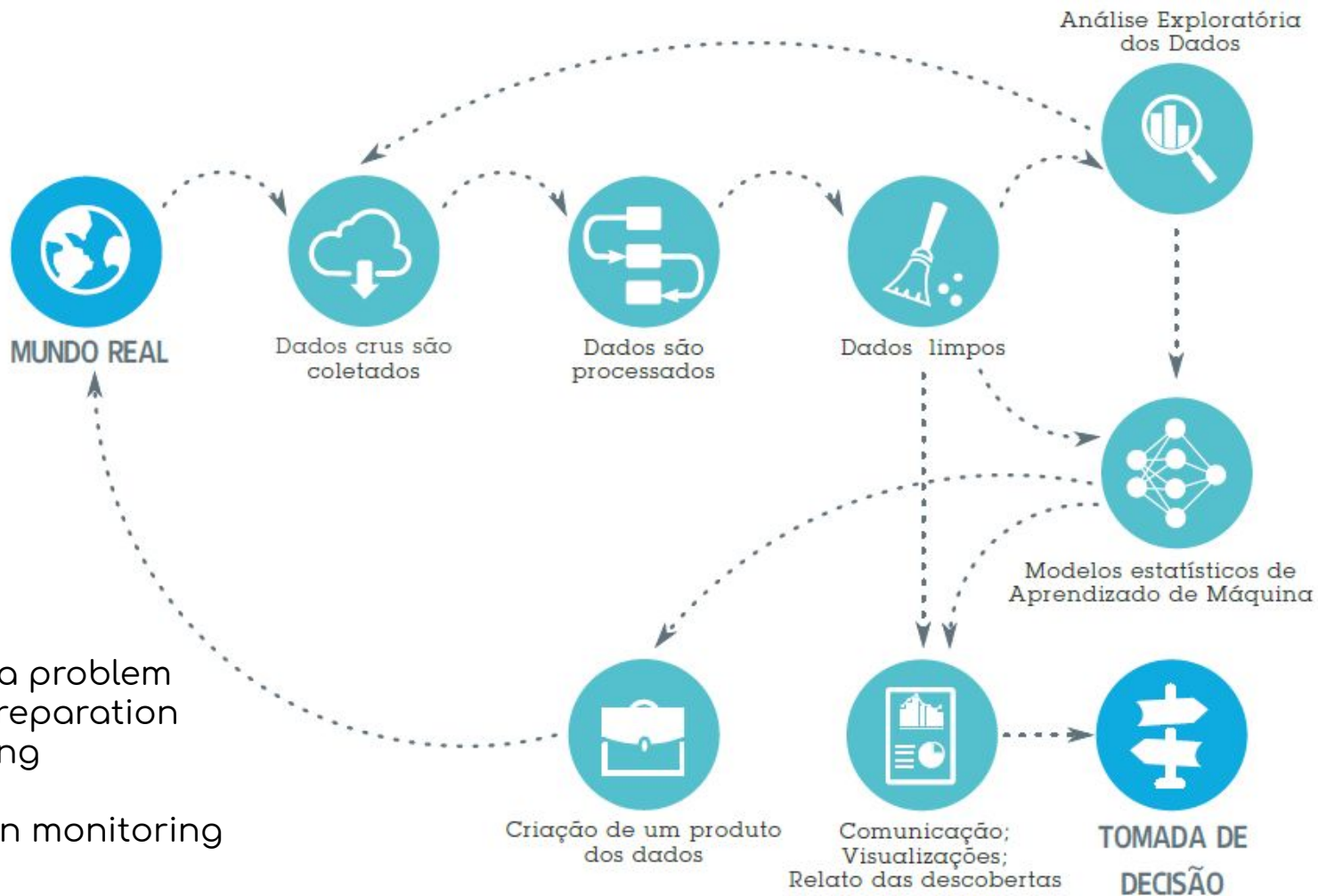
## *Lesson #04*

# *Introduction to Pandas*

*Mar. 2020*  
*Real Python*

1. What do you think makes a good data scientist?
2. Give a few examples of “best practices” in data science.
3. Can you outline the steps in a data science project?
4. What data would you love to acquire if there were no limitations?
5. Tell me a compelling story about data that you have analyzed.





1. Define a problem
2. Data preparation
3. Modeling
4. Deploy
5. Solution monitoring

1. **Plan the experiment:** this involves answering questions like:
  - a. What is the objective of the project
  - b. What kind of data
  - c. What is the desired accuracy ...
2. **Gather raw data:** make sure that we have the right kind of data and large enough data set for the desired accuracy.
3. **Gather tools:** use standard libraries, and aims for reproducibility
4. **Execute experiments**
5. **Analyze**
6. **Comunicate:** this involves preparing summaries of Experimental results, plots, publication, ...



# Project Inspiration

REUTERS GRAPHICS



## Notes from the 2008 financial crisis

# 10 YEARS ON

Five Reuters reporters who covered the collapse of Lehman Brothers and its aftermath examine what has changed — and what has not — in the decade since.

SEPTEMBER 10, 2018

Enter >>

up of the world in the mid-2000s. Tech has mushroomed. Tech worth nearly \$4 trillion. Apple's iPhones, Amazon's relentless expansion beyond online retailing, Google's dominance of internet search, Facebook's sway over social media. Microsoft's re-birth as a cloud computing giant.

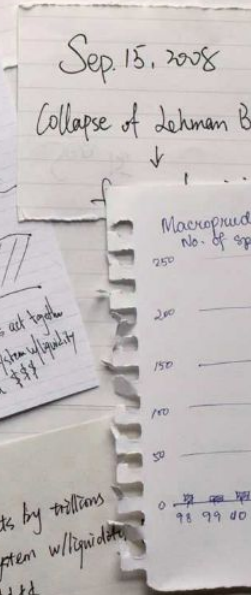


A. DEPRESSION.  
RICH ↔ POOR  
widening the gap

member of the EU.  
Poverty rates are higher in BULGARIA & ROMANIA  
Greece Place.

emerging markets, particularly across U.S. stocks winners after crisis. Europe stocks are losers. Emerging markets so so.

Before the crisis, U.S. stocks dominated global equity markets. But emerging markets, particularly across Asia, steadily gained share. U.S. stocks winners after crisis. Europe stocks are losers. Emerging markets so so.



- Decline of realist parties  
- Backlash from the voters  
- New anti-establishment parties in Europe  
- US voter identity  
- Independent

BAZOOKA  
expanded balance sheets by trillions  
Flooded banking system w/liquidity.  
WISH WISH \$\$\$\$.



# **Introduction to Pandas**

**Exploring Data with Pandas**

**Data Cleaning Basics**

**Data Aggregation**

**<<Project 01>>**

**Combining Data with Pandas**

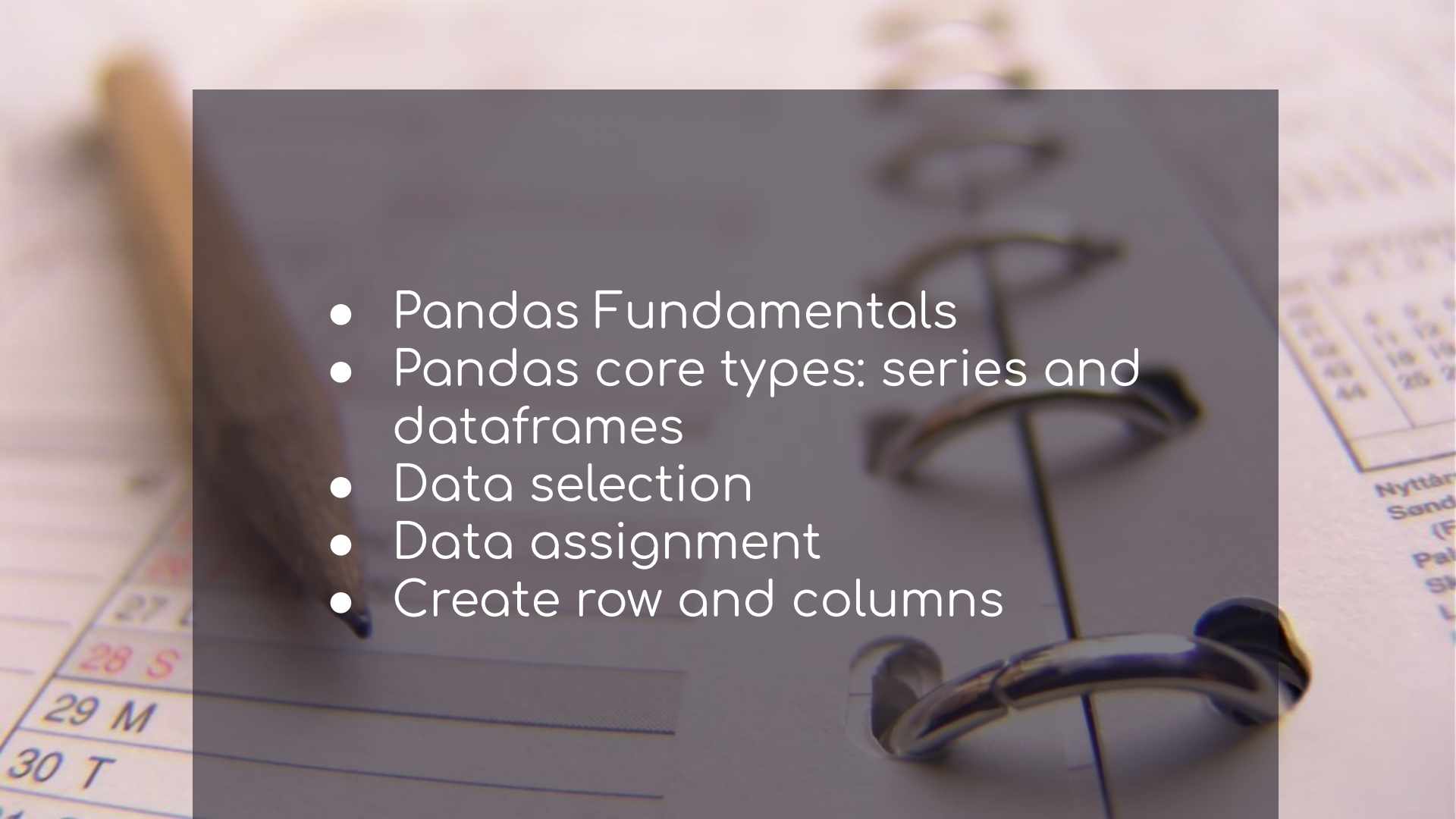
**Transforming Data with Pandas**

**Working with String in Pandas**

**Regular Expression**

**Working with missing and duplicate data**

**<<Project 02>>**

- 
- A spiral-bound notebook is shown from a top-down perspective. A dark-colored pen lies diagonally across the left side of the notebook. On the far left, a portion of a calendar strip is visible, showing dates 27, 28 (marked with a red 'S' for Sunday), 29 (marked with a blue 'M' for Monday), and 30 (marked with a blue 'T' for Tuesday). The notebook's pages are white with faint horizontal lines. The metal spiral binding is visible on the right side of the notebook. The entire scene is overlaid with a semi-transparent dark grey rectangle containing a bulleted list.
- Pandas Fundamentals
  - Pandas core types: series and dataframes
  - Data selection
  - Data assignment
  - Create row and columns

# Update from repository

---

```
git clone https://github.com/ivanovitchm/datascience2020
```

Or ....

```
git pull
```





S&P 500 Index - 1M - TVC O2974.3 H3096.5 L2945.2 C3066.8 +112.6 (+3.81%)

<https://www.tradingview.com/>



Vol (20): No volume is provided by the data vendor.

WMA 24 close



# Global 500

The world's 500 largest companies generated \$32.7 trillion in revenues and \$2.15 trillion in profits in 2018. Together, this year's Fortune Global 500 companies employ 69.3 million people worldwide and are represented by 34 countries. [Click here to view](#)

# Fortune 500

This year's *Fortune* 500 marks the 65th running of the list. In total, *Fortune* 500 companies represent two-thirds of the U.S. GDP with \$13.7 trillion in revenues, \$1.1 trillion in profits, \$22.6 trillion in market value, and employ 28.7 million people worldwide.

## The Top 10

Revenues (\$M)

1	Walmart	\$514,405
2	Sinopec Group	\$414,649
3	Royal Dutch Shell	\$396,556
4	China National Petroleum	\$392,976
5	State Grid	\$387,056
6	Saudi Aramco	\$355,905
7	BP	\$303,738
8	Exxon Mobil	\$290,212
9	Volkswagen	\$278,341
10	Toyota Motor	\$272,612

<https://fortune.com/global500/>

[Pull requests](#) [Issues](#) [Marketplace](#) [Explore](#)<https://github.com/cmusam/fortune500>[cmusam](#) / [fortune500](#)[Watch](#) 2[★ Unstar](#) 20[Fork](#) 23[Code](#)[Issues](#) 2[Pull requests](#) 0[Actions](#)[Projects](#) 0[Wiki](#)[Security](#)[Insights](#)

Branch: master ▾

[fortune500](#) / [csv](#) /[Create new file](#)[Upload files](#)[Find file](#)[History](#)

cmusam add 2019 data

Latest commit 618f7ac on 4 Dec 2019

..

<a href="#">fortune500-1955.csv</a>	upload HTML sources and csv files	14 months ago
<a href="#">fortune500-1956.csv</a>	upload HTML sources and csv files	14 months ago
<a href="#">fortune500-1957.csv</a>	upload HTML sources and csv files	14 months ago
<a href="#">fortune500-1958.csv</a>	upload HTML sources and csv files	14 months ago
<a href="#">fortune500-1959.csv</a>	upload HTML sources and csv files	14 months ago
<a href="#">fortune500-1960.csv</a>	upload HTML sources and csv files	14 months ago
<a href="#">fortune500-1961.csv</a>	upload HTML sources and csv files	14 months ago
<a href="#">fortune500-1962.csv</a>	upload HTML sources and csv files	14 months ago
<a href="#">fortune500-1963.csv</a>	upload HTML sources and csv files	14 months ago
<a href="#">fortune500-1964.csv</a>	upload HTML sources and csv files	14 months ago
<a href="#">fortune500-1965.csv</a>	upload HTML sources and csv files	14 months ago
<a href="#">fortune500-1966.csv</a>	upload HTML sources and csv files	14 months ago
<a href="#">fortune500-1967.csv</a>	upload HTML sources and csv files	14 months ago





# f500.csv

company	rank	revenues	revenue_change	profits	assets	profit_change	ceo	industry
Walmart	1	485873	0.8	13643.0	198825	-7.2	C. Douglas McMillon	General Merchandisers
State Grid	2	315199	-4.4	9571.3	489838	-6.2	Kou Wei	Utilities
Sinopec Group	3	267518	-9.1	1257.9	310726	-65.0	Wang Yupu	Petroleum Refining
China National Petroleum	4	262573	-12.3	1867.5	585619	-73.7	Zhang Jianhua	Petroleum Refining
Toyota Motor	5	254694	7.7	16899.3	437575	-12.3	Akio Toyoda	Motor Vehicles and Parts





# PANDAS PROFILING

<https://github.com/pandas-profiling/pandas-profiling>

## Navigation

[Plotly Express](#)

[A single import with built-in datasets](#)

[Scatter and Line plots](#)

[Visualize Distributions](#)

[Ternary Coordinates](#)

[Images](#)

[3D Coordinates](#)

[Polar Coordinates](#)

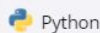
[Maps](#)

[← Back To Python](#)



# Plotly Express in Python

Plotly Express is a terse, consistent, high-level API for rapid data exploration and figure generation.



Python

► [New to Plotly?](#)

## Plotly Express

Plotly Express is the easy-to-use, high-level interface to Plotly, which [operates on "tidy" data](#) and produces [easy-to-style figures](#). Every Plotly Express function returns a `graph_objects.Figure` object whose data and layout has been pre-populated according to the provided arguments.

**Note:** Plotly Express was previously its own separately-installed `plotly_express` package but is now part of `plotly` and importable via `import plotly.express as px`.

This notebook demonstrates various `plotly.express` features. [Reference documentation](#) is also available, as well as a [tutorial on input argument types](#) and one on how to [style figures made with Plotly Express](#).

You can also read our original [Medium announcement article](#) for more information on this library.

# The Data Science Interview Study Guide

Top highlight

121 resources to help you land your data science dream job



SeattleDataGuy

Follow

May 20, 2019 · 7 min read ★

