

[bit.ly/pyladies\\_pandas](http://bit.ly/pyladies_pandas)



# Exploring Wikipedia with Pandas

Zareen Farooqui

# What is Pandas?



Open-source Python library for  
data analysis and modeling

**panel data**

# Goals

- High-level overview of Pandas library
- Write code on your own and complete challenge questions
- Lots of ideas, code & data to play around with after tutorial

# Introducing the Data

Pageviews

Clickstream

~150 MB	1.18 GB
~ 5 million rows	25,615,007 rows

Let's explore what's  
happening in the world  
right now....

<a href="#">pageviews-20160921-010000.gz</a>	21-Sep-2016 02:48	38565958
<a href="#">pageviews-20160921-020000.gz</a>	21-Sep-2016 03:31	38032592
<a href="#">pageviews-20160921-030000.gz</a>	21-Sep-2016 04:27	38119794
<a href="#">pageviews-20160921-040000.gz</a>	21-Sep-2016 05:26	37877998
<a href="#">pageviews-20160921-050000.gz</a>	21-Sep-2016 06:22	38632298
<a href="#">pageviews-20160921-060000.gz</a>	21-Sep-2016 07:58	41147282
<a href="#">pageviews-20160921-070000.gz</a>	21-Sep-2016 08:54	45359742
<a href="#">pageviews-20160921-080000.gz</a>	21-Sep-2016 09:17	49001483
<a href="#">pageviews-20160921-090000.gz</a>	21-Sep-2016 10:21	51429348
<a href="#">pageviews-20160921-100000.gz</a>	21-Sep-2016 11:27	52671794
<a href="#">pageviews-20160921-110000.gz</a>	21-Sep-2016 12:20	53004512
<a href="#">projectviews-20160921-000000</a>	01-Sep-2016 02:24	21101
<a href="#">projectviews-20160901-010000</a>	01-Sep-2016 03:37	21316
<a href="#">projectviews-20160901-020000</a>	01-Sep-2016 06:47	20516
<a href="#">projectviews-20160901-030000</a>	01-Sep-2016 07:38	21307
<a href="#">projectviews-20160901-040000</a>	01-Sep-2016 08:33	21520

Download last pageviews file  
(not projectviews)

<https://dumps.wikimedia.org/other/pageviews/2016/2016-09/>

# Pageviews Data

	project	article	requests	bytes_served
1228192	en	Main_Page	306095	0
2315811	en.m	Main_Page	185612	0
465978	de.m	Wikipedia:Hauptseite	66556	0
1555004	en	Special:Search	56479	0
3221294	es.m	Wikipedia:Portada	45600	0

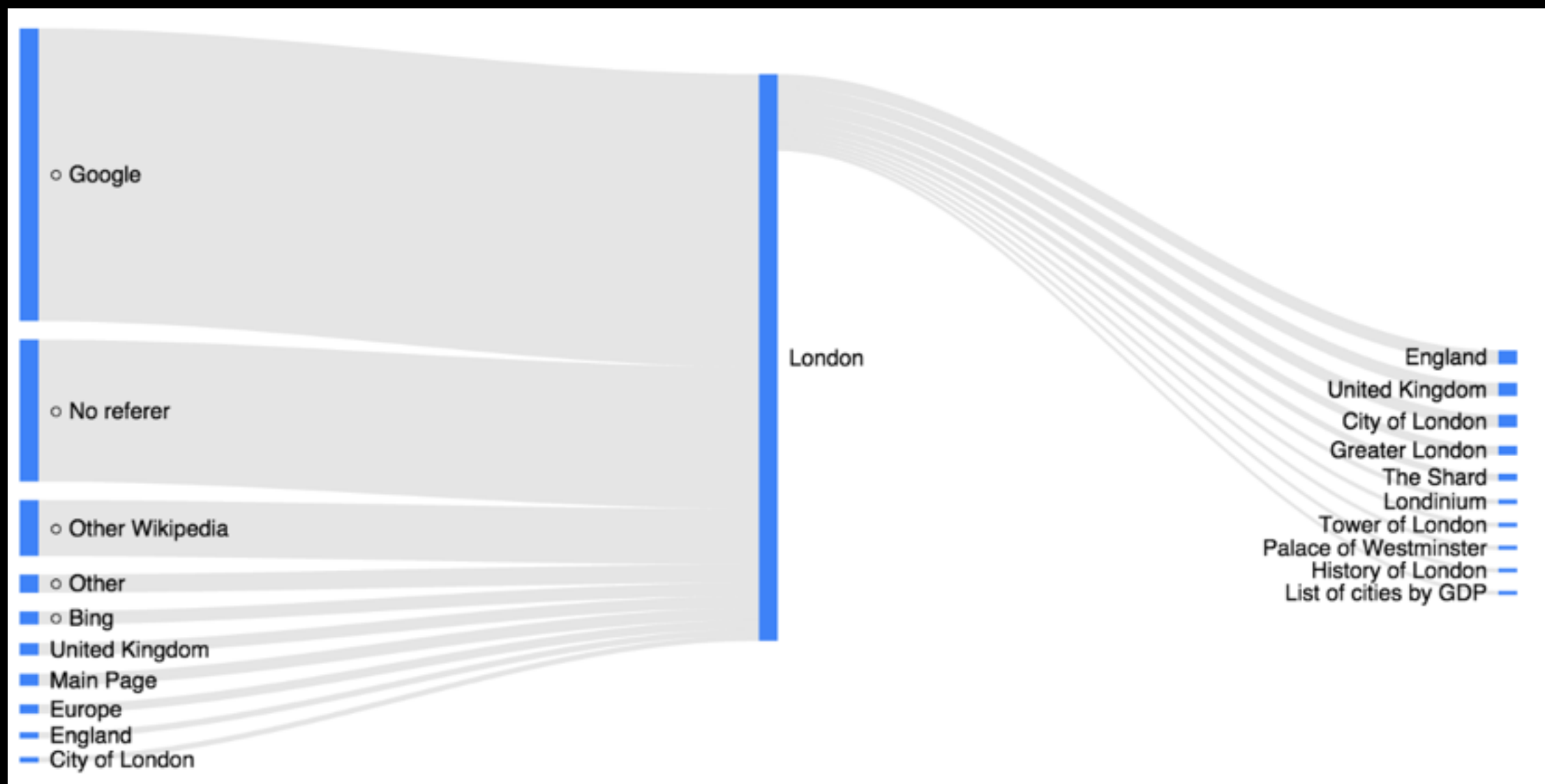
- **project**: Wikimedia project (290+ languages, many different project like wikibooks, wikinews, wikiquotes)
- **article**: title of current article
- **requests**: number of times that link was visited
- **bytes\_served**: total size of content returned




Let's explore what was  
happening in the world  
March 2016....

# Clickstream

how people get to a Wikipedia article and what links they click on



Source: [https://ewulczyn.github.io/Wikipedia\\_Clickstream\\_Getting\\_Started/](https://ewulczyn.github.io/Wikipedia_Clickstream_Getting_Started/)



WIKIPEDIA  
The Free Encyclopedia

Article **Talk**

Read **Edit** View history


## Boston

From Wikipedia, the free encyclopedia

*This article is about the capital of Massachusetts. For the English town it was named after, see **Boston, Lincolnshire**. For other uses, see **Boston (disambiguation)**.*

**Boston** (pronounced ˈbɒstən listen) is the capital and largest city<sup>[R]</sup> of the Commonwealth of Massachusetts<sup>[R]</sup> in the United States. Boston also served as the county seat of Suffolk County until Massachusetts disbanded most county governments by 2000.<sup>[R]</sup> The city proper covers 48 square miles (124 km<sup>2</sup>) with an estimated population of 667,137 in 2015,<sup>[R]</sup> making it the largest city in New England and the 23rd largest city in the United States.<sup>[R]</sup> The city is the economic and cultural anchor of a substantially larger metropolitan area called Greater Boston, home to 4.7 million people and the tenth-largest metropolitan statistical area in the country.<sup>[R]</sup> Greater Boston as a commuting region is home to 8.1 million people, making it the sixth-largest in the United States.<sup>[R]</sup>

One of the oldest cities in the United States, it was founded in 1630 by Puritan settlers from England during the American Revolution, such as the Battle of Bunker Hill, and the Siege of Boston continued to be an important port and culture.<sup>[R]</sup><sup>[R]</sup> Through land reclamation beyond the original peninsula, its area is now drawing over 20 million visitors per year, including the public school, Boston Latin School, and the Boston Public Garden (1634).



WIKIPEDIA  
The Free Encyclopedia

Article **Talk**

Read **Edit** View history

## Massachusetts

From Wikipedia, the free encyclopedia

*This article is about the U.S. state. For other uses, see **Massachusetts (disambiguation)**.*

**Massachusetts** (listen help ˌmæsəˈtʃuːts / ˌmæsəˈtʃuːsɪ; officially the **Commonwealth of Massachusetts**), is the most populous state in New England, and a part of the Northeast megalopolis. It is bordered by the Atlantic Ocean to the east, the state of New Hampshire to the north, the state of Rhode Island to the south, and the state of Connecticut to the southwest. The state capital is Boston, and the largest city is Springfield.

The screenshot shows the Wikipedia article for 'Atlantic Ocean'. The title 'Atlantic Ocean' is highlighted with a red box. A red arrow points from this box to the 'Talk' tab in the top right corner of the article content area. The left sidebar contains various navigation links such as 'Random article', 'Donate to Wikipedia', 'Interaction', 'Tools', and 'Print/export'. The main content area displays the introductory text of the article, mentioning its location between North America and Europe, and its status as the second-largest ocean.

<span></span>	<div> <div><span>Featured content</span></div> <div><span>Current events</span></div> <div><span>Random article</span></div> <div><span>Donate to Wikipedia</span></div> <div><span>Wikipedia store</span></div> </div>
<span></span>	<div> <div><span>Interaction</span></div> <div><span>Help</span></div> <div><span>About Wikipedia</span></div> <div><span>Community portal</span></div> <div><span>Recent changes</span></div> <div><span>Contact page</span></div> </div>
<span></span>	<div> <div><span>Tools</span></div> <div><span>What links here</span></div> <div><span>Related changes</span></div> <div><span>Upload file</span></div> <div><span>Special pages</span></div> <div><span>Permanent link</span></div> <div><span>Page information</span></div> <div><span>Wikidata item</span></div> <div><span>Cite this page</span></div> </div>
<span></span>	<div> <div><span>Print/export</span></div> <div><span>Create a book</span></div> </div>

Figure 1. The effect of the number of trials on the number of correct responses. The number of correct responses was significantly higher for the 10 trials condition than for the 5 trials condition. Error bars represent the standard error of the mean.

# Clickstream Data

	prev	curr	type	n
0	Wild_Bill_Hickok	Deadwood,_South_Dakota	link	18
1	(Ghost)_Riders_in_the_Sky:_A_Cowboy_Legend	Stan_Jones_(songwriter)	link	410
2	(Hey_You)_The_Rock_Steady_Crew	Rock_Steady_Crew	link	75
3	(Miss)understood	Heaven_(Ayumi_Hamasaki_song)	link	13
4	(Theme_from)_Valley_of_the_Dolls	Do_You_Know_the_Way_to_San_Jose	link	12

- **prev**: where the hit came from
  - another Wikipedia articles
  - external sources (Google, Facebook, Twitter)
- **curr**: title of current article
- **type**:
  - link: prev and curr are both Wikipedia articles,
  - external: prev is a non-Wikipedia source (Google, Facebook, etc.),
  - other: prev and curr are both Wikipedia articles but the curr article was searched for or the prev article was spoofed
- **n**: number of times that link was visited that month

# Data Structures

- **series:** one-dimensional object similar to a labeled array, list or column in a table

```
s = pd.Series([22, 'Boston', 3.14, -87264034, 'Hello there!'])
s
0      22
1    Boston
2     3.14
3 -87264034
4  Hello there!
dtype: object
```

- **dataframe:** 2 dimensional labeled data structure with rows and columns (spreadsheet, SQL table, dict of Series)

# Thanks!

 @iamzareenf

 @zareenf

 [zareenf@gmail.com](mailto:zareenf@gmail.com)