#### 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....

```
pageviews-20160921-010000.qz
                                                21-Sep-2016 02:48
                                                                           38565958
pageviews-20160921-020000.qz
                                                21-Sep-2016 03:31
                                                                           38032592
pageviews-20160921-030000.gz
                                                21-Sep-2016 04:27
                                                                           38119794
                                                21-Sep-2016 05:26
                                                                           37877998
pageviews-20160921-040000.gz
                                                21-Sep-2016 06:22
                                                                           38632298
pageviews-20160921-050000.gz
                                                21-Sep-2016 07:58
pageviews-20160921-060000.gz
                                                                           41147282
                                                21-Sep-2016 08:54
                                                                           45359742
pageviews-20160921-070000.gz
pageviews-20160921-080000.gz
                                                21-Sep-2016 09:17
                                                                           49001483
                                                21-Sep-2016 10:21
pageviews-20160921-090000.gz
                                                                           51429348
pageviews-20160921-100000.gz
                                                21-Sep-2016 11:27
                                                                           52671794
                                                21-Sep-2016 12:20
                                                                           53004512
pageviews-20160921-110000.gz
                                                01-Sep-2016 02:24
                                                                             21101
projectviews-20160901-010000
                                                01-Sep-2016 03:37
                                                                              21316
projectviews-20160901-020000
                                                01-Sep-2016 06:47
                                                                              20516
projectviews-20160901-030000
                                                01-Sep-2016 07:38
                                                                              21307
                                                01-Sep-2016 08:33
                                                                              21520
projectviews-20160901-040000
Download last pageviews file
              (not projectviews)
```

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

## Pageviews Data

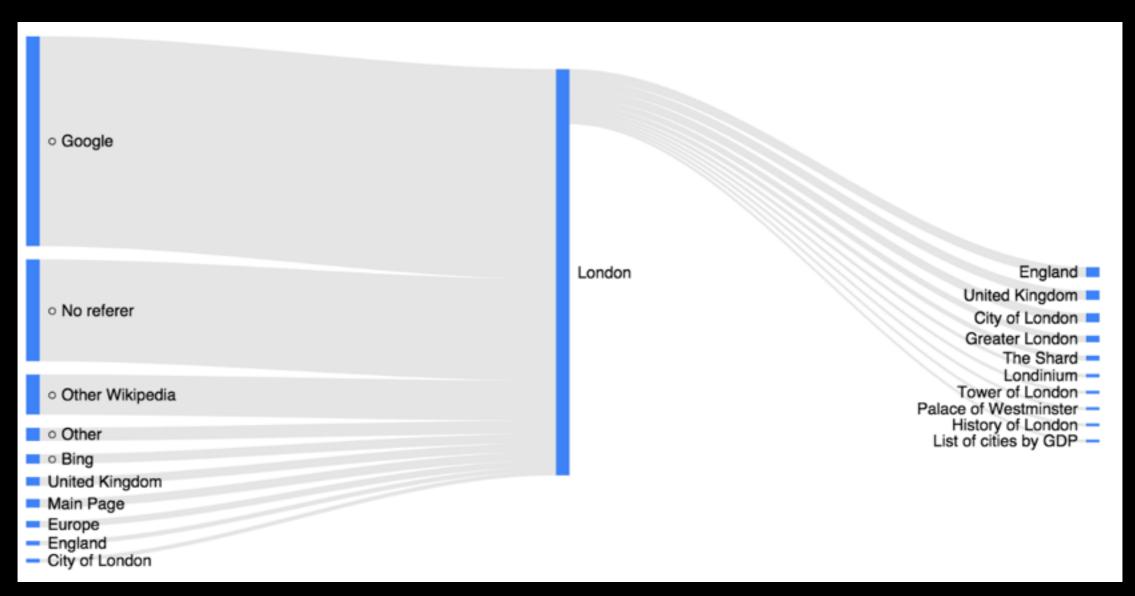
	project	article	requests	bytes_served
1228192	en	Main_Page	306095	0
2315811	en.m	Main_Page	185612	0
465978	de.m	Wikipedia:Hauptseite	66556	0
1555004	en	Special:Search	56479	0
3221294	es.m	Wikipedia:Portada	45600	0

- 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/



ument events endom article nate to Wikipedia Nipedia store

What links here Upload file Permanent Ink Page information Ote this page

Read Edit Viewhistory Search 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

e-tirb is the capital and largest city<sup>(R)</sup> of the

s disbanded most county governments by 2000.74 The

England and the 23rd largest city in the United

Commonwealth of Massachusetts is the United States. Boston also served as the county seat

What links here

Special pages

Permanent link

Page information

Whidele liers

City this page

Create a book

city proper covers 48 square miles (N4 km²) with an estimated population of 667,137 in

Boston -> Greek mythology



Featured conten Current events

About Wikipedia

Community portal

Recent changes Contact page

Upload file

Special pages

Permanent link

Page informatio

Boston, Massachusetts

City of Boston

Plymouth was the site of the first color of the Mayfower, in 1692, the town of most infamous cases of mass hysteria founded the Springfield Armory, which technological advances, including into revolt led by disaffected Revolutionary Convertion.[48] In the 18th century, the world, originated from the pulpit of Nor century. Boston became known as the

Atlantic Ocean WIKIPEDIA From Wikipedia, the free encyclopedia "Atlantic", "North Atlantic", "South Atlantic", "Atlantic Basin", and "Atlantic coast" redirect here. For other uses, see Atlantic (disambiguation).

North Atlantic (disambiguation), South Atlantic (disambiguation), Atlantic Basin (disambiguation), and Atlantic Coast (disambiguation). This article needs additional citations for verification. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. (March 2013) (Learn how and when to remove this template message)

WIKIPEDIA

Randon article

Donate to Wikinedia Wkipedia store

About Wikipedia

What links here

Special pages

Permanent Int

One this page

The Atlantic Ocean is the second largest of the world's oceanic divisions, following the Pacific Ocean. With a total area of about 106,400,000 square kilometres (41,100,000 sq mi),[1] it covers-approximately 20 percent of the bout 29 percent of its water surface area. Its name refers to Afias of Greek mythology, m

The oldest known mention of "Atlantic" is in The Histories of Hi 1.202.4): Atlantis thalassa (Greek: Ankgynic Bákozoo; Englis) Ethiopic Ocean, derived from Ethiopia, was applied to the sout 19th century. Before Europeans discovered other oceans, the synonymous with the waters beyond the Strait of Gibraltar that The early Greeks believed this ocean to be a gigantic river end

The Atlantic Ocean occupies an elongated, S-shaped basin ex Contents Eurasia and Africa to the east, and the Americas to the west. A Featured con interconnected global ocean, it is connected in the north to the Ocean in the southwest, the Indian Ocean in the southeast, an

& Not logged in Talk Contributions Create account Log in

Greek mythology

Greek mythology is the body of myths and teachings that belong to the ancient Greeks, concerning their gods and heroes, the nature of the world, and the origins and significance of their own cult and ritual practices. It was a part of the religion in ancient Greece. Modern scholars refer to and study the myths in an attempt to shed light on the religious and political institutions of Ancient Greece and its civilization, and to gain understanding of the nature of

Wikipeda Clickstream source Viewhistory Search

Greek mythology is explicitly embodied in a large collection of narratives, and implicitly in Greek representational arts, such as vase-paintings and votive gifts. Greek myth attempts to explain the origins of the world, and details the lives and adventures of a wide variety of gods, goddesses, heroes, heroines and mythological creatures. These accounts initially were disseminated in an oral poetic tradition; today the Greek myths are known primarily from Greek Herature. The oldest known Greek Herary sources, Homer's epic poems Had and Odyssey, focus on the Trojan War and its aftermath. Two poems by Homer's near contemporary Hesiod, the Theogony and the Works and Days, contain accounts of the genesis of the world, the succession of divine rulers, the succession of human ages, the origin of human woes, and the origin of sacrificial practices. Myths are also preserved in the Homeric Hymns, in fragments of epic poems of the Epic Cycle, in lyric poems, in the works of the tragedians and comedians of the fifth century BC, in writings of scholars and poets of the Hellenistic Age, and in texts from the time of the Roman Empire by writers such as Plutarch and Pausarias.

Archaeological findings provide a principal source of detail about Greek mythology, with gods and heroes featured prominently in the decoration of many artifacts. Geometric designs on pottery of the eighthcentury BC depict scenes from the Troian cycle as well as the adventures of Heracles. In the succeeding Archaic, Classical, and Hellenistic periods, Homeric and various other mythological scenes appear,



★ 量 側

Rotonda, Museo Pio-

myth-making itself. [1]

of Sulfolk Count

2015,<sup>[10]</sup> making it the largest city in Ne

About Wikipedia

Contact page

#### 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

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

## Thanks!

- @iamzareenf
- @zareenf
- zareenf@gmail.com