state-of-states

Finding correlations between **Wikipedia** data and a country's **economic success**

by **Harrison Pincket** & **Indira Pranabudi**

1.

HYPOTHESIS: The amount of **attention** a country receives on **Wikipedia** is positively correlated to its **economic success**.

A country's **economic success** is defined as the country's **GDP**. Meanwhile, a country's W**ikipedia attention** is computed using:

File size
Number of citations
Number of forward links
Number of edits

2.

DATASET INFORMATION: We scraped Wikipedia pages using a Python library called **BeautifulSoup**. We looked at all the pages associated with a particular country, and ran a Python script to download the files. in total, we downloaded **1,634,173 pages**, which take up **116GB** and took approximately **1 week** to download.

3.

METHODOLOGY:

- Ran a Python script to **extract** the four metrics above for all countries, over all their associated pages.
- **Plot** the changes in the four metrics for each country, compared to its GDP over the years. no positive correlation.
- **Visualize** the number of edits on a world map. countries with either a high GDP or a big population have more edits.
- Calculate the **Pearson correlation** between the number of edits and different metrics (GDP, land area, urban population, electricity usage, etc.)

4.

CHALLENGES

Data collection:

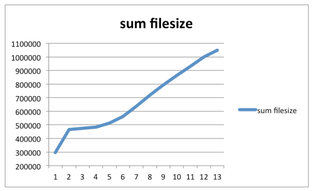
- Determining which pages were associated with a country. Options: title of the page, neighboring pages, community portals, 'list of __ in country' pages, and 'categories'. We settled on 'categories'. Categories and sub-categories are overlapping and broad, leading to unrelated pages. We regulated our **search depth** to reduce volume.
- Handle instances of 404 and 500 **Errors**

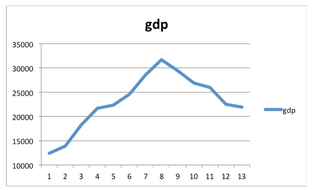
github: **indirap.github.io/state-of-states**

FINDINGS

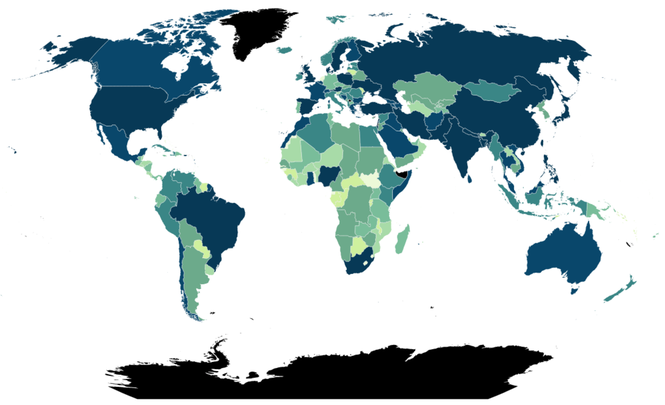
Most Edits:
United Kingdom, United States, India, Israel, Japan

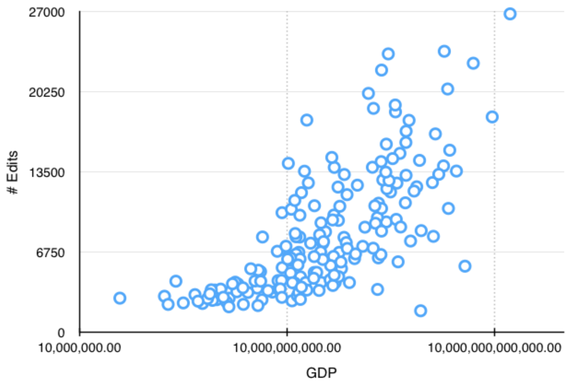
Fewest Edits:
Western Sahara, Netherlands, Seychelles, Swaziland, Lesotho





World Map of Edits





Highest Pearson Correlations:
Citations with GDP (2005 US$): 0.724
Citations with GNI, Atlas method (current US$): 0.707
Citations with GNI (current US$): 0.7066
Citations with GDP (current US$): 0.7061
Citations with # of secure Internet servers: 0.681

Lowest Pearson Correlations:
File size with freshwater withdrawal: -0.000997
File size with ratio of women in ministry: 0.00225
Links with freshwater withdrawal: -0.00522
Links with pre-primary entrance age: 0.00753
Citations with freshwater withdrawal: 0.00891

SAMPLE PAGES

Austria
Grand Duchy of Tuscany
Vienna Summer of Logic
Praetorian Prefecture of Illyricum
Sierra Leone
Armed Forces Revolutionary Council
Parliament of Sierra Leone
Islam in Sierra Leone

CHALLENGES

Entity Resolution:

- **Data from different sources**: some countries that are recognized by the World Bank (our source for GDP information) were not listed as countries on Wikipedia.
- **Reconcile names**: some countries' official names are different from what we commonly know them as (Macedonia is Former Yugoslav Republic of Macedonia; North Korea is Democratic Republic of Korea)