DATA ENGINEERING PROJECT: ANALYSIS OF WORLD POPULATION DISTRIBUTION

Implementing a Simple ETL Pipeline using Python

By Jahval Romiz Septrada

Project Preview

Tujuan project ini yaitu membangun basic pipeline ETL end-to-end. Dimulai dengan mengubah data tabel mentah populasi global yang didapat dari hasil webscrapping dari website **Wikipedia** menjadi data yang siap dianalisis. Dari data tersebut, didapat insight tentang distribusi populasi manusia di seluruh dunia.


https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population

Contents hide

(Top)
Method
Sovereign states and dependencies by population
See also
Explanatory notes
References

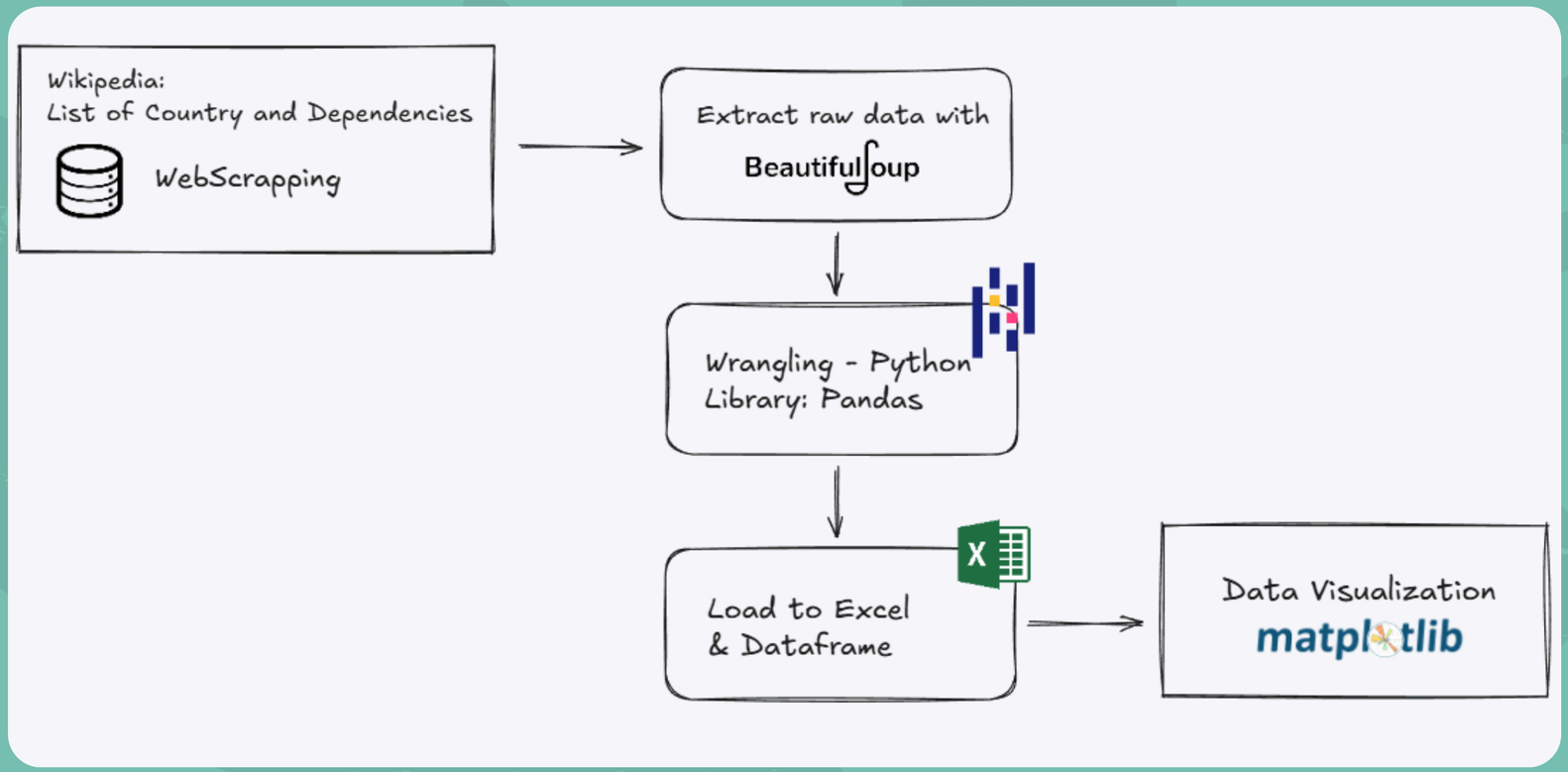
List of countries and territories by total population

	Location	Population	% of world	Date	Source (official or from the United Nations)	Notes
	World	8,232,000,000	100%	13 Jun 2025	UN projection[1][3]	
1	India	1,417,492,000	17.3%	1 Jul 2025	Official projection[4]	[b]
2	China	1,408,280,000	17.2%	31 Dec 2024	Official estimate[5]	[c]
3	United States	340,110,988	4.1%	1 Jul 2024	Official estimate[6]	[d]
4	Indonesia	284,438,782	3.5%	30 Jun 2025	National annual projection[7]	
5	Pakistan	241,499,431	2.9%	1 Mar 2023	2023 census result[8]	[e]
6	Nigeria	223,800,000	2.7%	1 Jul 2023	Official projection[9]	
7	Brazil	213,421,037	2.6%	1 Jul 2025	Official estimate[10]	
8	Bangladesh	169,828,911	2.1%	14 Jun 2022	2022 census result[11]	[f]
9	Russia	146,028,325	1.8%	1 Jan 2025	Official estimate[13]	[g]
10	Mexico	130,575,786	1.6%	30 Jun 2025	National quarterly estimate[14]	
11	Japan	123,300,000	1.5%	1 Aug 2025	Monthly national estimate[15]	
12	Philippines	114,123,600	1.4%	1 Jul 2025	Official projection[16]	

TOOLS

- **Python** : Programming Language utama untuk scripting & data analysis
- **Colab Notebook** : Sebagai interactive environment untuk menulis script kode & menampilkan visualisasi
- **BeautifulSoup & Request** : Library utama untuk WebScrapping
- **Pandas** : Library untuk manipulasi data seperti pembersihan data, tranformasi, & analisis data
- **Matplotlib** : Library untuk membuat visualisasi dengan grafik analisis data

PIPELINE


Wikipedia:
List of Country and Dependencies
WebScrapping
Extract raw data with
BeautifulSoup
Wrangling - Python
Library: Pandas
Load to Excel
& Dataframe
Data Visualization
matplotlib

DATA COLLECTING

01 Mengirim Request HTTP Get ke URL Artikel Wikipedia

02 Parse string HTML mentah menjadi objek dengan BeautifulSoup

03 Pisahkan Header & Rows hasil pengambilan HTML Table ke list masing-masing

03 Gabungkan Header dan Rows ke Dataframe

RAW DATA

	Location	Population	% ofworld	Date	Source (official or fromthe United Nations)	Notes
0	World	8,232,000,000	100%	13 Jun 2025	UN projection[1][3]	
1	India	1,417,492,000	17.3%	1 Jul 2025	Official projection[4]	[b]
2	China	1,408,280,000	17.2%	31 Dec 2024	Official estimate[5]	[c]
3	United States	340,110,988	4.1%	1 Jul 2024	Official estimate[6]	[d]
4	Indonesia	284,438,782	3.5%	30 Jun 2025	National annual projection[7]	
...	...	...	...	...	...	...
235	Niue (New Zealand)	1,681	0%	11 Nov 2022	2022 Census[252]	
236	Tokelau (New Zealand)	1,647	0%	1 Jan 2019	2019 Census[253]	
237	Vatican City	882	0%	31 Dec 2024	Official figure[254]	[ah]
238	Cocos (Keeling) Islands (Australia)	593	0%	30 Jun 2020	2021 Census[255]	
239	Pitcairn Islands (UK)	35	0%	1 Jul 2023	Official estimate[256]	

240 rows × 6 columns

DATA CLEANING

01 Menghapus baris & kolom yang tidak diperlukan (Kolom 'Note' & Baris 1 - World)

02 Menghapus penggunaan '%' pada nilai kolom 'Percentage'

03 Mengubah tipe data yang tidak sesuai ('Population' -> int, 'Date' -> datetime, 'Percentage' -> float)

03 Menghapus nomor sitasi di kolom 'Source'

DATA FINAL

	Location	Population	Percentage	Date	Source
1	India	1417492000	17.3	2025-07-01	Official projection
2	China	1408280000	17.2	2024-12-31	Official estimate
3	United States	340110988	4.1	2024-07-01	Official estimate
4	Indonesia	284438782	3.5	2025-06-30	National annual projection
5	Pakistan	241499431	2.9	2023-03-01	2023 census result
...	...	...	...	...	...
235	Niue (New Zealand)	1681	0.0	2022-11-11	2022 Census
236	Tokelau (New Zealand)	1647	0.0	2019-01-01	2019 Census
237	Vatican City	882	0.0	2024-12-31	Official figure
238	Cocos (Keeling) Islands (Australia)	593	0.0	2020-06-30	2021 Census
239	Pitcairn Islands (UK)	35	0.0	2023-07-01	Official estimate

239 rows × 5 columns

ANALYSIS RESULT

Rata-Rata Populasi Sedunia

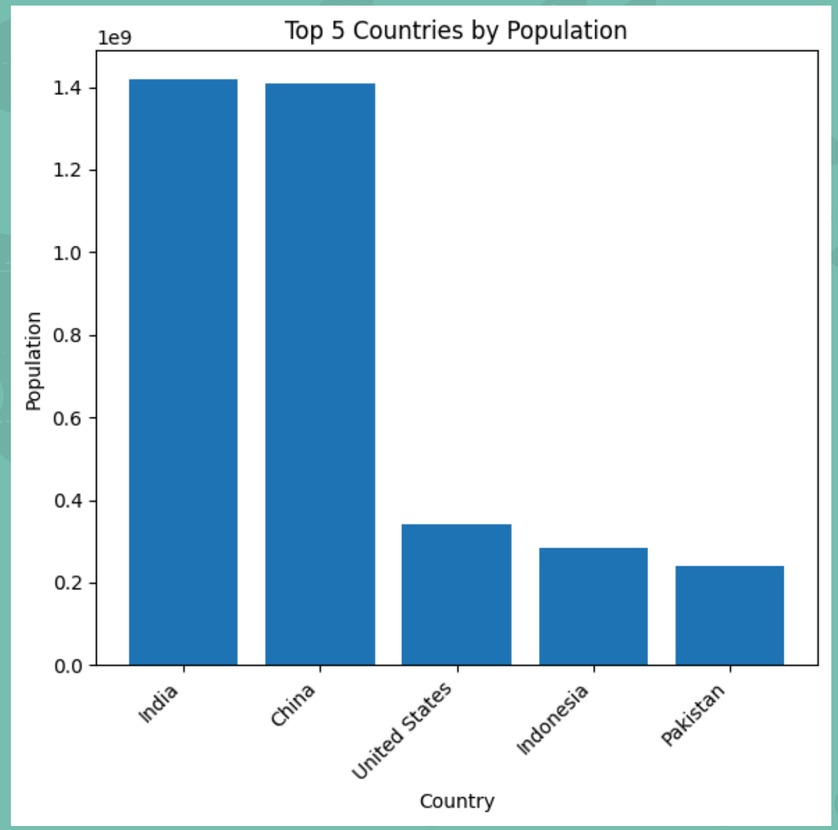
33554601.64

Populasi Negara Terdikit

35

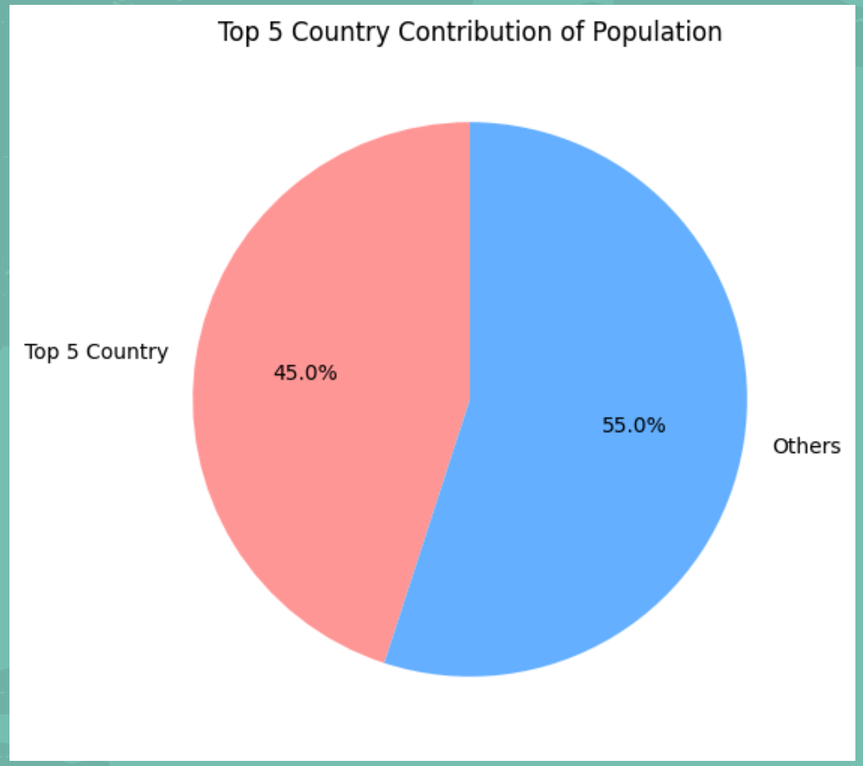
ANALYSIS RESULT

Top 5 Negara dengan Jumlah Penduduk Terbanyak



ANALYSIS RESULT

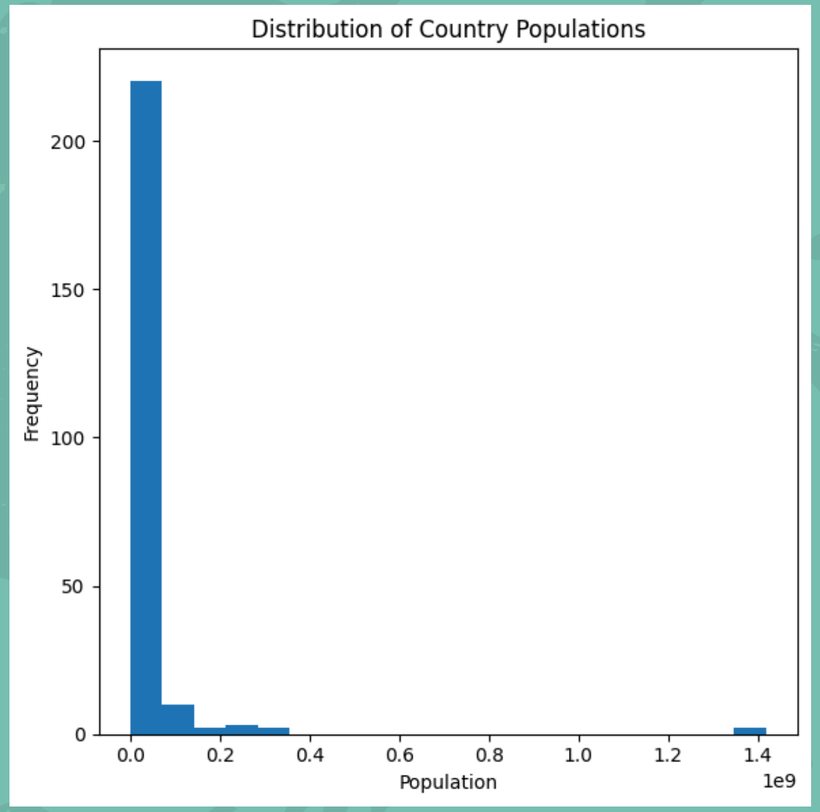
Kontribusi Populasi Top 5 Negara dari Keseluruhan Populasi



Insight:
Dari grafik bisa dilihat bahwa 5 Negara dengan populasi terbanyak memiliki total 45% populasi dari keseluruhan populasi di dunia. Ini menunjukkan bahwa hampir setengah populasi umat manusia tinggal di 5 negara teratas ini saja

ANALYSIS RESULT

Distribusi Populasi Seluruh Dunia



Insight:

Mayoritas negara di dunia memiliki populasi yang relatif kecil. Frekuensi lebih dari 200 menunjukkan ada lebih dari 200 negara dengan populasi kurang lebih di bawah 100 juta orang.

THANK YOU