Global Data Analysis API

Group 6

Kenneth Le, Kurai Monica Matiki, Eric Tran, Greg Presneill

Aim and purpose

Chosen Track: Data Engineering

This project aims to develop a reliable, scalable API that allows users to easily access and analyze global data trends, such as world population growth, by providing comprehensive datasets. The API supports analysts in querying and investigating key statistics for research and decision-making.

Why did we choose this data?

We chose world population growth data because it is a critical indicator that affects various aspects of society, such as economic development, resource allocation and policy making.

Motivation #1: How can I use all this global data and extract only the data I need?

Motivation #2: How can our API empower users to answer their questions with data?

Ethical considerations

The project utilizes publicly available, open-source datasets from trusted organizations such as the World Bank. We carefully verify the allowed use of these datasets before incorporating them into our system.

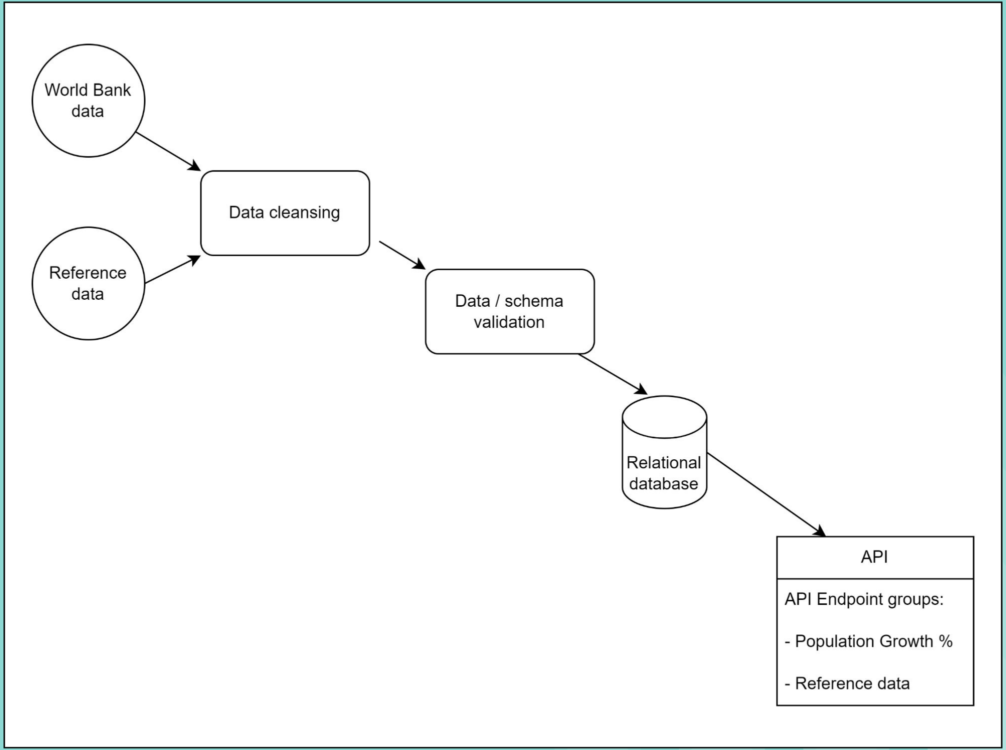
1. **Copyright and Fair Use**: The dataset has been reviewed for copyright protections, and its use complies with the applicable fair use or licensing agreements.
2. **Intended Use Documentation:** the intended usage of the dataset, both present and future, has been documented to ensure compliance with the term of use.
3. **Data Collection Verification**: The dataset's collection process has been investigated to confirm that it was sourced from authorized and legitimate providers.

cc creative commons

CC BY 4.0

ETL Workflow

1. Extract data (World Bank data & reference data)
2. Data cleansing
3. Data/schema validation
4. Transform data as required to match database schema
5. Create blank database, initialise table schemas
6. Load data into database
7. Design API endpoints
8. Implement queries to allow end users to extract data through the API



Data Sourcing

Data Sources and Extraction Process

- World Bank Data
- Web Data (for Country Codes)

THE WORLD BANK

- Downloaded CSV files
- Used Python to read through the csv files and create data frames

Data Cleansing

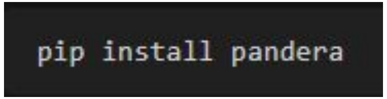
- Converting raw data into a suitable format.

- Data Cleaning: Handling Missing Values, removing unwanted columns

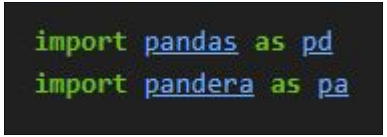
- Data Integration: Combining Data from Different Sources

Data/Schema Validation

Data Validation with...



```
pip install pandera
```

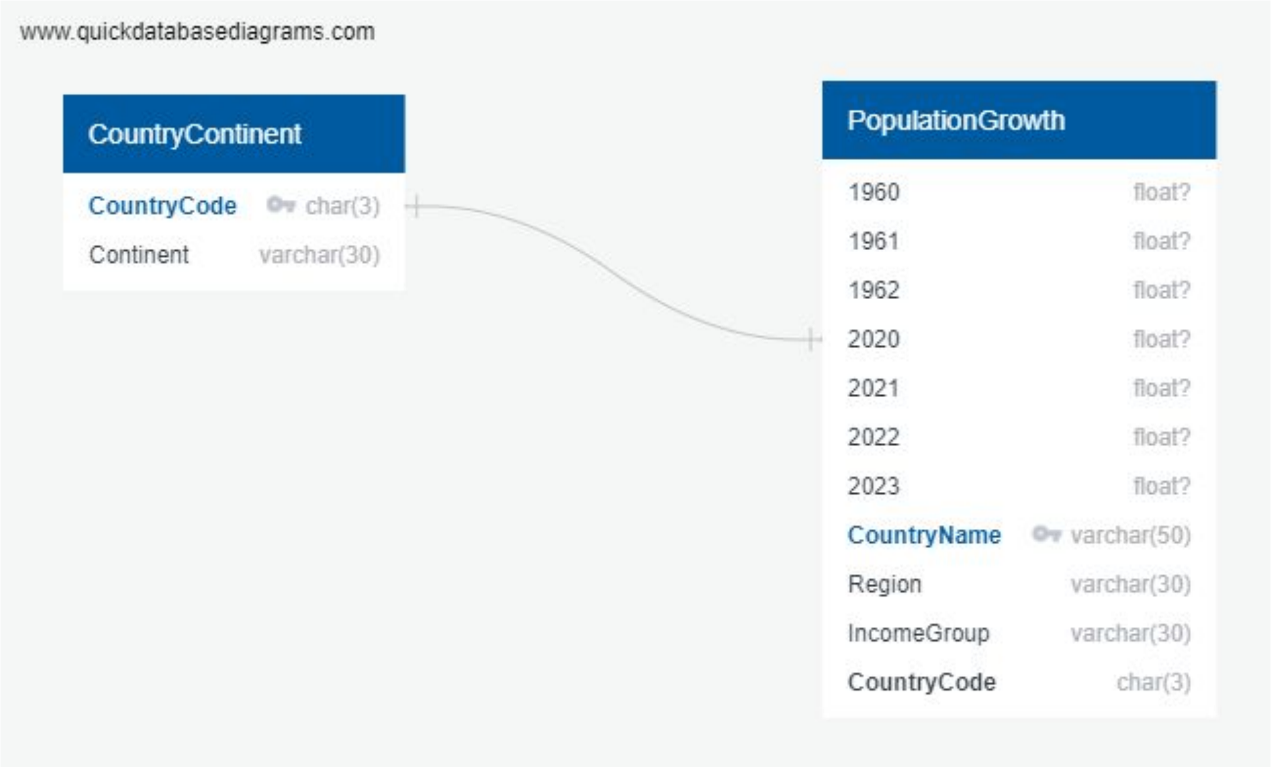


```
import pandas as pd
import pandera as pa
```

UNION

PANDERA

Database/ERD



www.quickdatabasediagrams.com

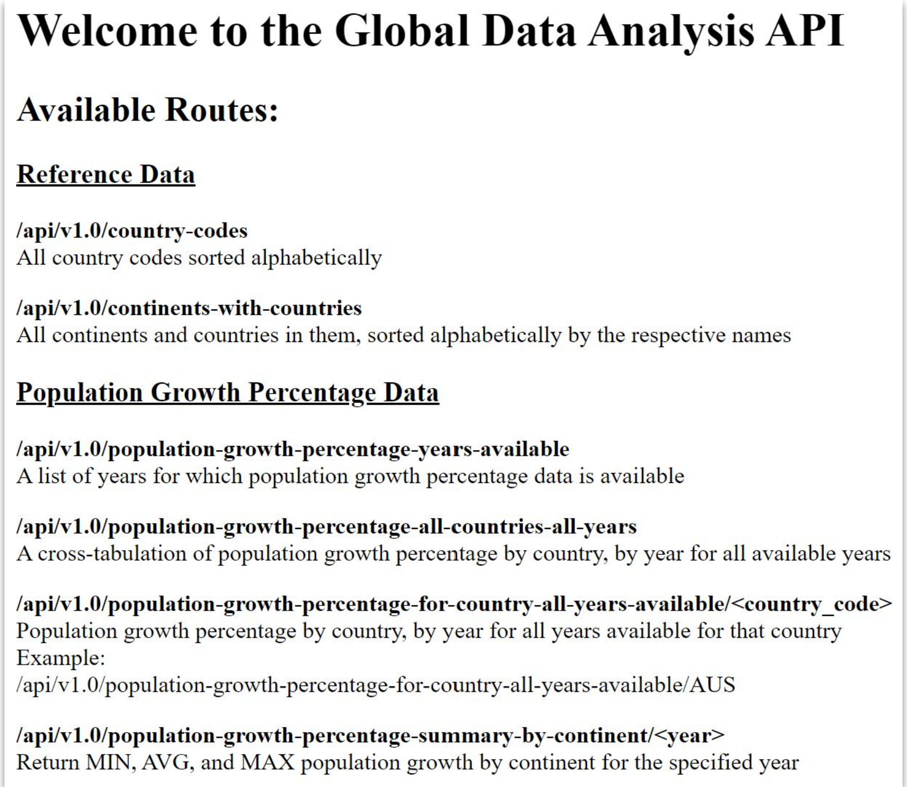
CountryContinent	
CountryCode	char(3)
Continent	varchar(30)

PopulationGrowth	
1960	float?
1961	float?
1962	float?
2020	float?
2021	float?
2022	float?
2023	float?
CountryName	varchar(50)
Region	varchar(30)
IncomeGroup	varchar(30)
CountryCode	char(3)

How would users use the API?

The Flask-based API is designed to run as a web server, that will return JSON data for queries such as:

- List the Countries and Years for which data is available
- Population growth % for all countries by year
- Individual country population growth % by year
- Minimum, Average, and Maximum Population growth % by continent for a specified year



Welcome to the Global Data Analysis API

Available Routes:

Reference Data

/api/v1.0/country-codes
All country codes sorted alphabetically

/api/v1.0/continents-with-countries
All continents and countries in them, sorted alphabetically by the respective names

Population Growth Percentage Data

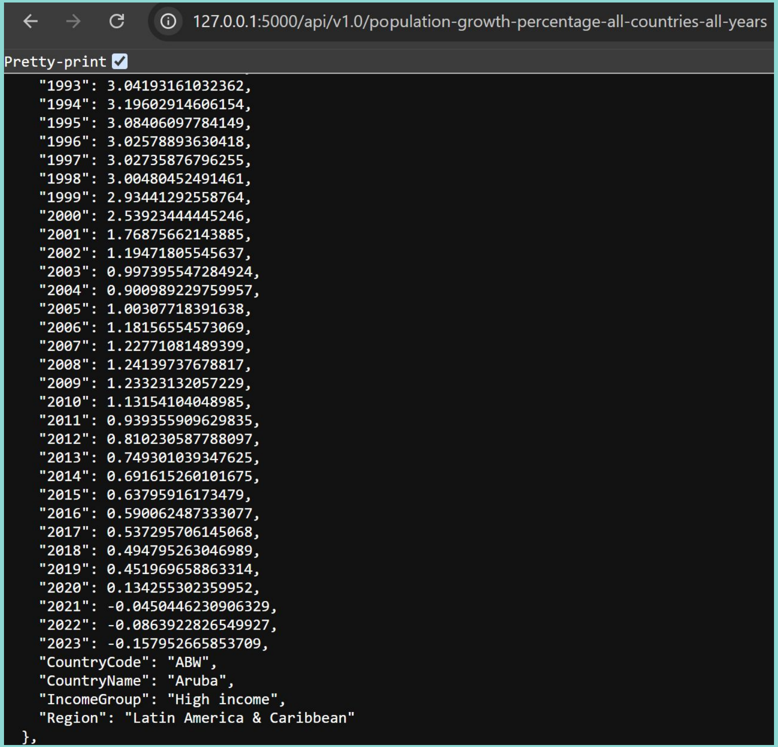
/api/v1.0/population-growth-percentage-years-available
A list of years for which population growth percentage data is available

/api/v1.0/population-growth-percentage-all-countries-all-years
A cross-tabulation of population growth percentage by country, by year for all available years

/api/v1.0/population-growth-percentage-for-country-all-years-available/<country_code>
Population growth percentage by country, by year for all years available for that country
Example:
/api/v1.0/population-growth-percentage-for-country-all-years-available/AUS

/api/v1.0/population-growth-percentage-summary-by-continent/<year>
Return MIN, AVG, and MAX population growth by continent for the specified year

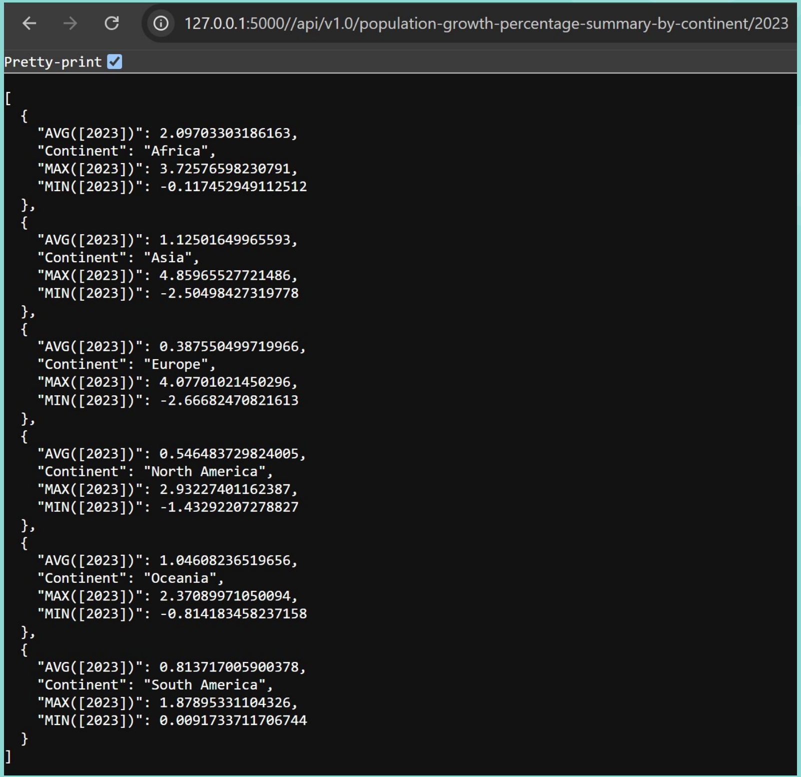
API in Action



127.0.0.1:5000/api/v1.0/population-growth-percentage-all-countries-all-years

Pretty-print ☑

```
    "1993": 3.04193161032362,
    "1994": 3.19602914606154,
    "1995": 3.08406097784149,
    "1996": 3.02578893630418,
    "1997": 3.02735876796255,
    "1998": 3.00480452491461,
    "1999": 2.93441292558764,
    "2000": 2.53923444445246,
    "2001": 1.76875662143885,
    "2002": 1.19471805545637,
    "2003": 0.997395547284924,
    "2004": 0.900989229759957,
    "2005": 1.00307718391638,
    "2006": 1.18156554573069,
    "2007": 1.22771081489399,
    "2008": 1.24139737678817,
    "2009": 1.23323132057229,
    "2010": 1.13154104048985,
    "2011": 0.939355909629835,
    "2012": 0.810230587788097,
    "2013": 0.749301039347625,
    "2014": 0.691615260101675,
    "2015": 0.63795916173479,
    "2016": 0.590062487333077,
    "2017": 0.537295706145068,
    "2018": 0.494795263046989,
    "2019": 0.451969658863314,
    "2020": 0.134255302359952,
    "2021": -0.0450446230906329,
    "2022": -0.0863922826549927,
    "2023": -0.157952665853709,
    "CountryCode": "ABW",
    "CountryName": "Aruba",
    "IncomeGroup": "High income",
    "Region": "Latin America & Caribbean"
  },
```



127.0.0.1:5000//api/v1.0/population-growth-percentage-summary-by-continent/2023

Pretty-print ☑

```
[
  {
    "AVG([2023])": 2.09703303186163,
    "Continent": "Africa",
    "MAX([2023])": 3.72576598230791,
    "MIN([2023])": -0.117452949112512
  },
  {
    "AVG([2023])": 1.12501649965593,
    "Continent": "Asia",
    "MAX([2023])": 4.85965527721486,
    "MIN([2023])": -2.50498427319778
  },
  {
    "AVG([2023])": 0.387550499719966,
    "Continent": "Europe",
    "MAX([2023])": 4.07701021450296,
    "MIN([2023])": -2.66682470821613
  },
  {
    "AVG([2023])": 0.546483729824005,
    "Continent": "North America",
    "MAX([2023])": 2.93227401162387,
    "MIN([2023])": -1.43292207278827
  },
  {
    "AVG([2023])": 1.04608236519656,
    "Continent": "Oceania",
    "MAX([2023])": 2.37089971050094,
    "MIN([2023])": -0.814183458237158
  },
  {
    "AVG([2023])": 0.813717005900378,
    "Continent": "South America",
    "MAX([2023])": 1.87895331104326,
    "MIN([2023])": 0.00917337117067444
  }
]
```

Conclusion

What could this API be used for in the future?

Questions