**Group:**

* Kundyz Smith
* Fawn Zou
* Kevin Beygi
* Ugochi Akaluso

**Topic: World Development Indicators**

**What?**: Providing a database including several indicators that can be applied to analyze the objective of minimizing poverty globally as set by the World Bank.

**Why?**: These indicators are so diverse that a single database from one single webpage (database) cannot provide the required indications of movement toward the objective. So, we need to gather the database from several web pages each including some of the indicators.

**How?**: In order to track this objective, it is required to define several themes (as listed below) and within each category some indicators are set and measured. The indicators, however, can be shared within more than one category, for example, the global pattern in CO2 emission, or people access to primary infrastructures in each country. Then, the goal(s) of each category is (are) related to the primary objective of this project, i.e. measuring the global poverty line.

**Introduction**

The World Development Indicators (WDI) is a compilation of relevant, high-quality, and internationally comparable statistics about global development and the fight against poverty. We explored several online resources to gain the required measurements pertaining to more than 1600 indicators for almost 220 economies, with some data series extending back more than 50 years. One of the import parameters that severely impacts the global indicators, is anthological carbon emission. Carbon dioxide emission is causing numerous problems around the world. With the issue of climate change far from resolved, understanding the determinants behind emission levels is necessary to improve forecasting and guide policy making. There is an evidence that emission factor depends on per capita income. Carbon dioxide table provides carbon dioxide emission data from 1960 till 2018 for every country around the world. This table can be used to see the correlation between carbon dioxide emission and income group levels in every country.

In the following, we highlight the six primary indicator categories together with their own (or shared) sub-indicators. Again, all these indicators serve to address both categorical and global objectives as described above.

**Indicators subjects to use by each category:**

* **Poverty and Shared Prosperity:**
  + Population below national poverty lines
  + Poverty gap at national poverty lines
  + International poverty lines: Population below $1.90 a day, Population below $3.20 a day, Population below $5.50 a day
  + Percentage share of income or consumption
  + Annualized growth in mean consumption or income per capita
  + Mean consumption or income per capita
* **People**:
  + Poverty eradication
  + Health
  + Education opportunities
  + Gender equality
  + Water and sanitation
  + Decent work and economic growth
* **Environment**:
  + promote sustainable agriculture
  + Availability of and access to water
  + Challenges that urbanization creates
* **Economy**:
  + Growth of consumption, investment and trade
  + Structure of economy demands
* **States and Market**:
  + Private investment and performance,
  + The public sector’s role in nurturing investment and growth, and
  + The quality and availability of infrastructure essential for growth
* **Global links**
  + Size and direction of trades in goods and services,
  + Movement of people, and
  + Impact of policy interventions

**Output database**

We generate a non-relational database (World\_Development\_Indicator) consisting of three collections (WDI\_general and CO2):

1. WDI\_general collection was extracted from the relational database (database.sqlite extracted from the Kaggle project) consisting of several tables followed by a data munging process. This collection relate 1600 indicators from more than 180 countries with more than 5.6 M rows.
2. CO2 collection was created by extracting two data sets one from Kaggle and the second one from The World Bank website. Data were separately cleaned, merged and converted into JSON format to be added to the database.
3. Unemployment rate-education input collection extracted from web-scraping the World Bank website. These two tables were gone through a data munging process as described earlier, followed by adding the merged table as a collection in the Mongodb database.

**Data references:**

**World Bank**:  h[ttp://datatopics.worldbank.org/world-development-indicators/themes/people.html](http://datatopics.worldbank.org/world-development-indicators/themes/people.html)

**Kaggle**: <https://www.kaggle.com/worldbank/world-development-indicators#Indicators.csv>

**Google public data**: <https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_>

**What useful investigation could be done with the final database?**

As discussed in WWH

**Whether final DB will be relational or non relational, and why?**

Since the dataset is extracted from several resources with different formats, i.e., a heterogenous data set, we prefer the non-relational DB. Also, the generated database will be integrated at the end but from the ACID perspective, we are not worried about the data integrity during the transition step. This database is primarily generated once and occasionally will be updated and the isolation, integrity, or the consistency of data is not the main concern of this local project. After all, we trust each other :) Finally, we are storing records in the same collection that have different fields or attributes. For this purpose, also, we prefer the non-relational DB.

**STEPS:**

1. **Export**: Started with a relational database (Sqlite) from a Kaggle project and the World Bank website
2. **Transform**:
   1. Data cleaning:
      1. *dropped some unnecessary columns and excluded some of the imported CSV files from the sqlite database followed by renaming some of the column headers.*
   2. Data transformation:
      1. *exported sqlite database into three Pandas dataframe where we could easily merge the tables. It was followed by exporting the dataframe into the Mongodb environment.*[*¶*](http://localhost:8888/notebooks/HomeWorks/13-ETL_Project/Notebooks/Kevin.ipynb#exported-sqlite-database-into-three-Pandas-dataframe-where-we-could-easily-merge-the-tables.-It-was-followed-by-exporting-the-dataframe-into-Mongodb.)

**Note**: we opt to apply Pandas DataFrame as a bridge between Sqlite and MongoDB to manipulate data. The other, and actually more professional approach, could be applied by the tools that are intended for this purpose, such as T3 Studio GUI.

**Load**: export data to a non-relational database (MongoDB)