The ExPatApp

Where to next?

Allison, Ben, Jackie, Nicole & Tom

The B-team

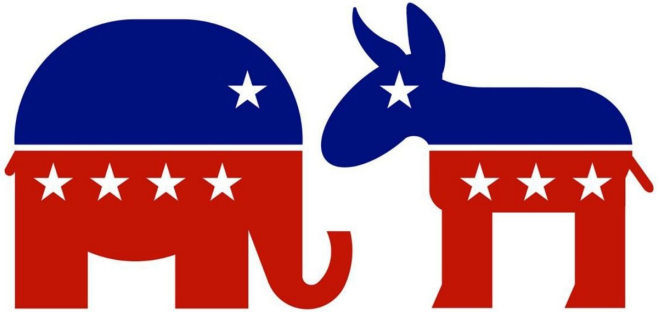
Allison, Ben, Jackie, Nicole, & Thomas

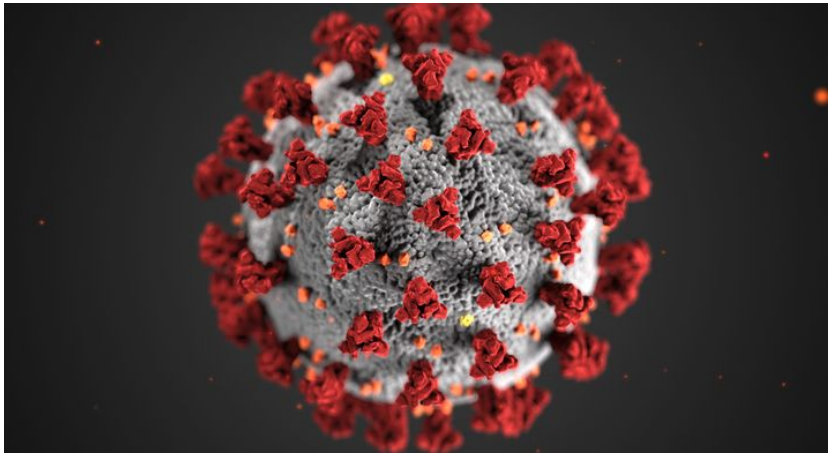
Overview

As we adjust to a post-pandemic world, the American people have changed how and where they work and live. As the American political landscape continues to evolve – and the rise of remote work allows workers to live wherever they find internet access – the idea of emigration from the US has renewed interest.

This analytical project aims to curate an index ranking of countries tailored to Americans interested in trying a new life in a foreign country.

Overview





Questions to consider before moving abroad

1. **Economic parity:** How stable and/or developed is the economy?
2. **Health outcomes:** Are the people who live there healthy?
3. **Political system:** Is it democratic? Are the people ‘free’?
4. **Education system:** What are the average person’s schooling outcomes?
5. **Culture:** Would an American be welcome there? How happy are the people who live there?

Data Exploration for Country-Level data

- ➔ Multiple country-level data sources need to be connected and synchronized to build a comprehensive dataset to answer these questions

- ➔ Explored various proxy indicators/metrics for the five key factors that would be important for Americans considering emigration

Data Exploration for Country-Level data

- **Economy**
  - **Sources**: United Nations, World Bank, Newspapers, Academic studies
  - **Metrics**: United Nations' Human Development Index (HDI), Gross Domestic Product/Gross National Income (GDP/GNI), Income inequality (Gini), Cost of Living (COL), Internet speed, Big Mac Index
- **Health**
  - **Sources**: UN, World Bank, World Health Org., Org. for Economic Cooperation and Development (OECD),, Non-Governmental Organizations (NGOs)
  - **Metrics**: Life expectancy, Happiness index, Quality of life
- **Politics**
  - **Sources**: UN, World Bank, OECD, The Economist, Freedom House, NGOs
  - **Metrics**: Human Freedom Index, Democracy Index, Global Freedom Scores
- **Education**
  - **Sources**: UN, World Bank, OECD, CIA World Factbook
  - **Metrics**: Literacy, average years of schooling, prevalence of advanced degrees
- **Lifestyle**
  - Sources: UN, World Bank, Statista, UNESCO
  - **Metrics**: Religious freedom index, racial/ethnic diversity, climate, % of English-speakers, Climate

Data Extraction

1. **Economic parity:** UN Human Development Index scores, incl. GNI per capita
2. **Health outcomes:** Health Adjusted Life Expectancy (HALE), World Bank/UN
3. **Political system:** *The Economist's* Democracy Index, Regime Type
4. **Education system** : Literacy rates and mean years of schooling (World Bank)
5. **Culture**: Freedom of Religion data (UN)

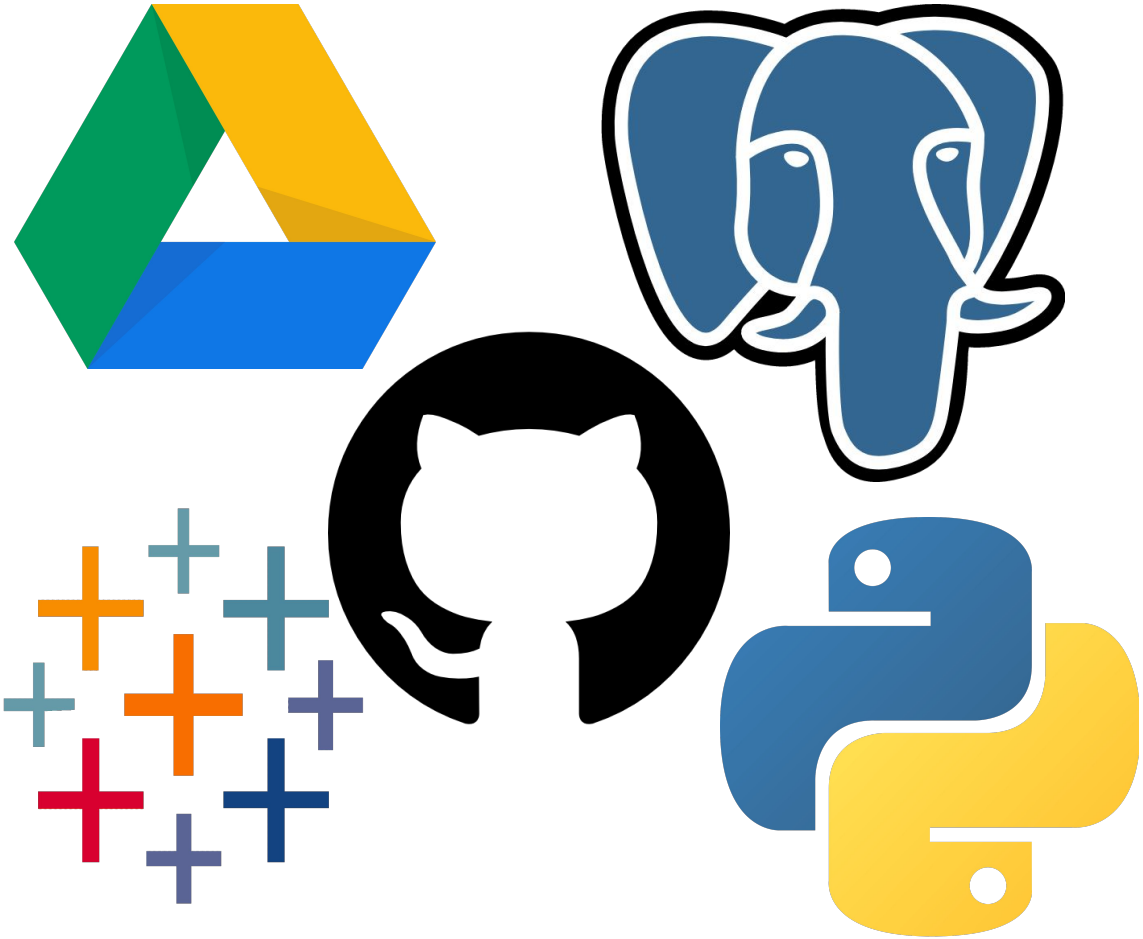
THE WORLD BANK

The Economist

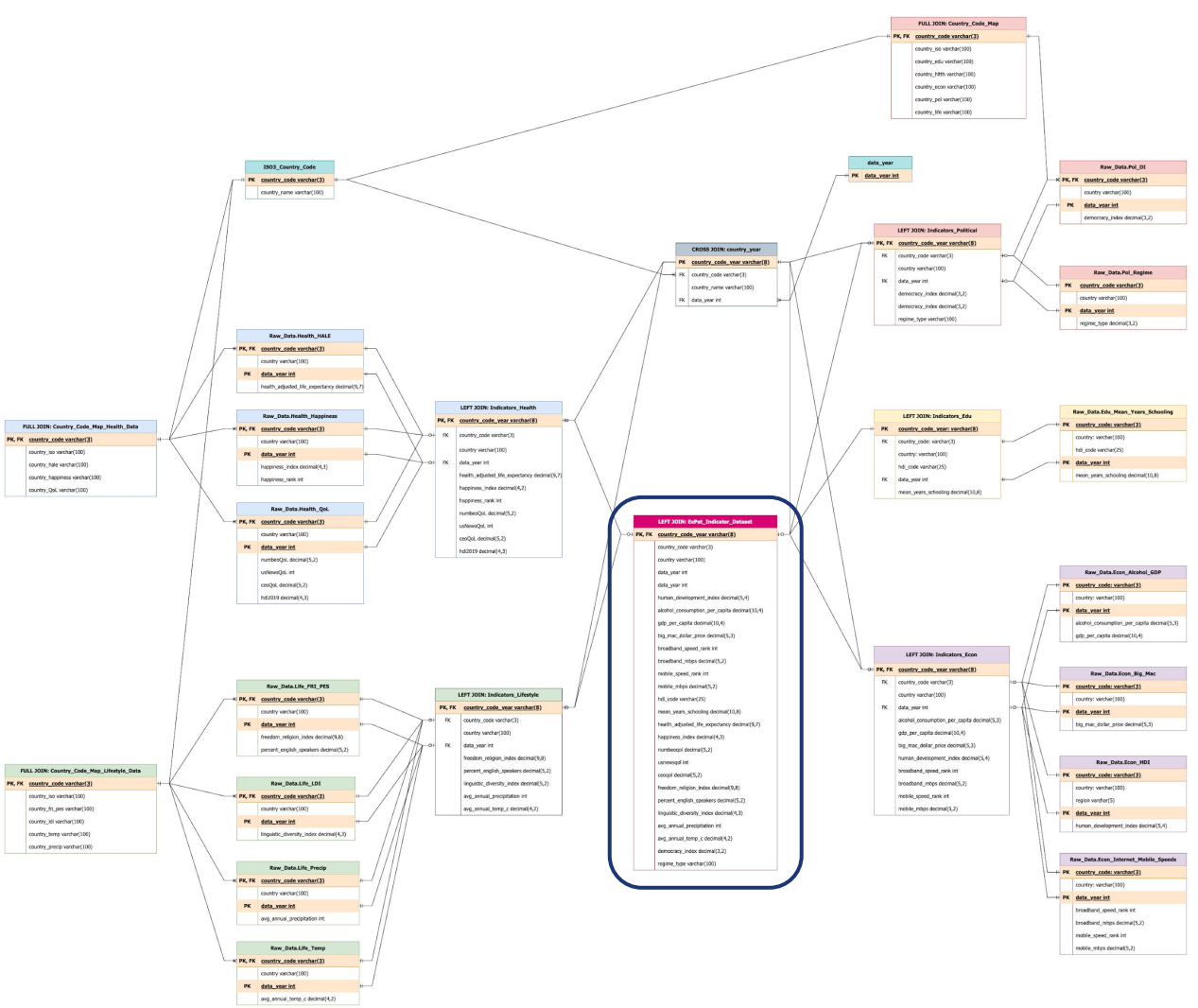
Methodology

Tools

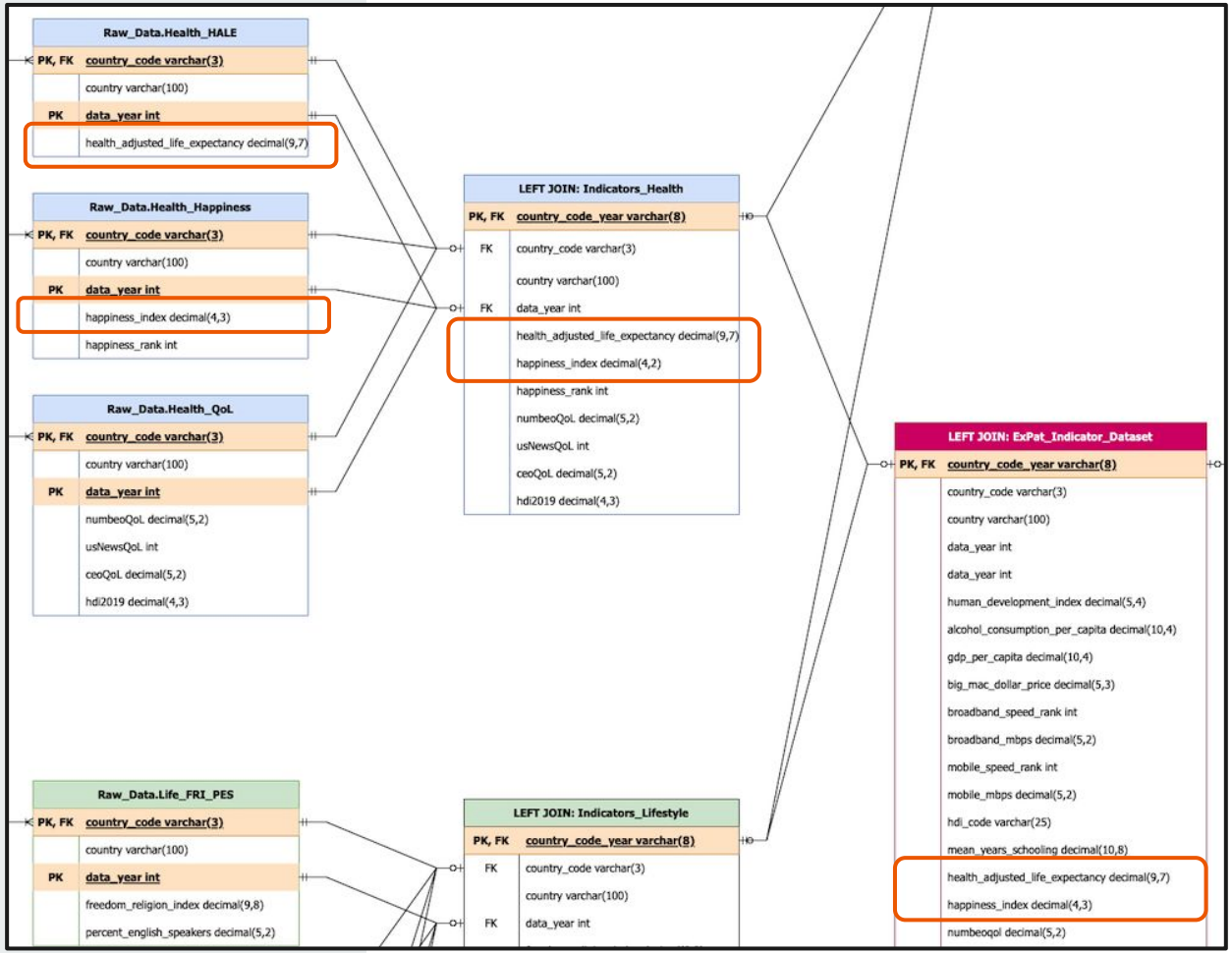
- Google Drive
- GitHub
- Postgres
  - pgAdmin
- Python, Pandas
  - SQLAlchemy
  - scikit-learn
- Tableau Public



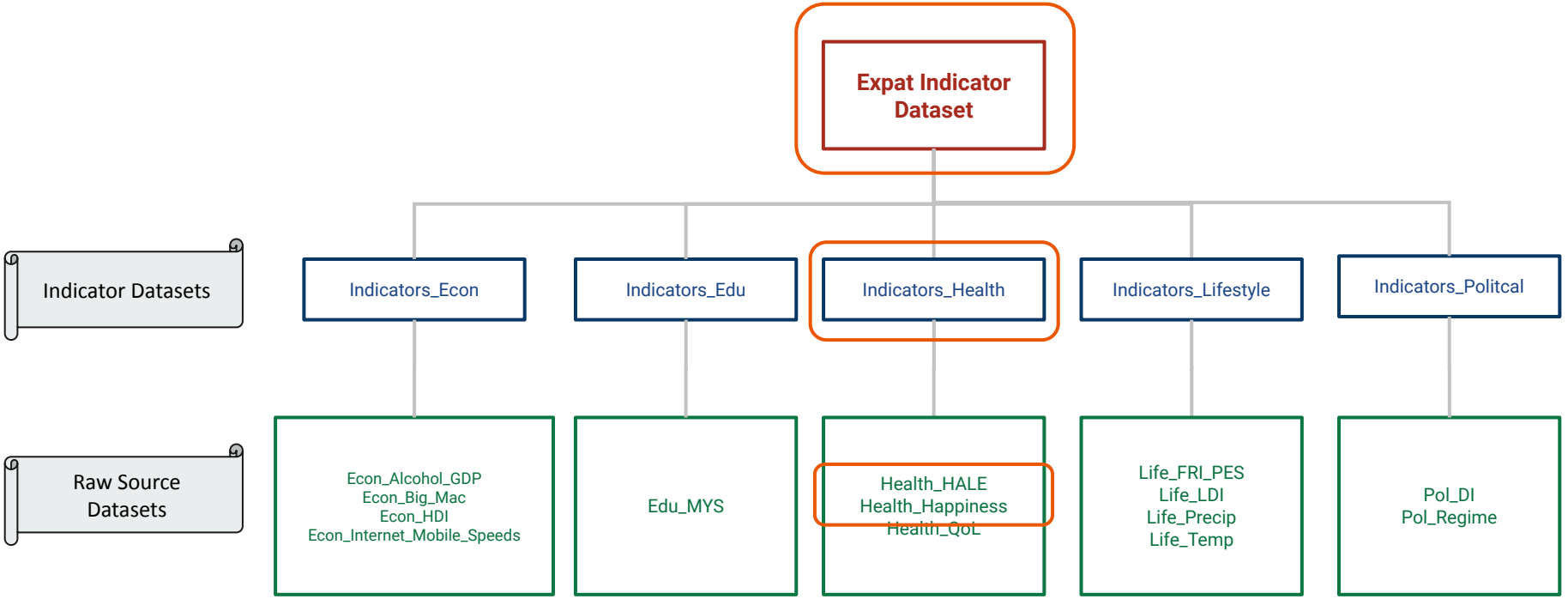
ExPatApp Database ERD



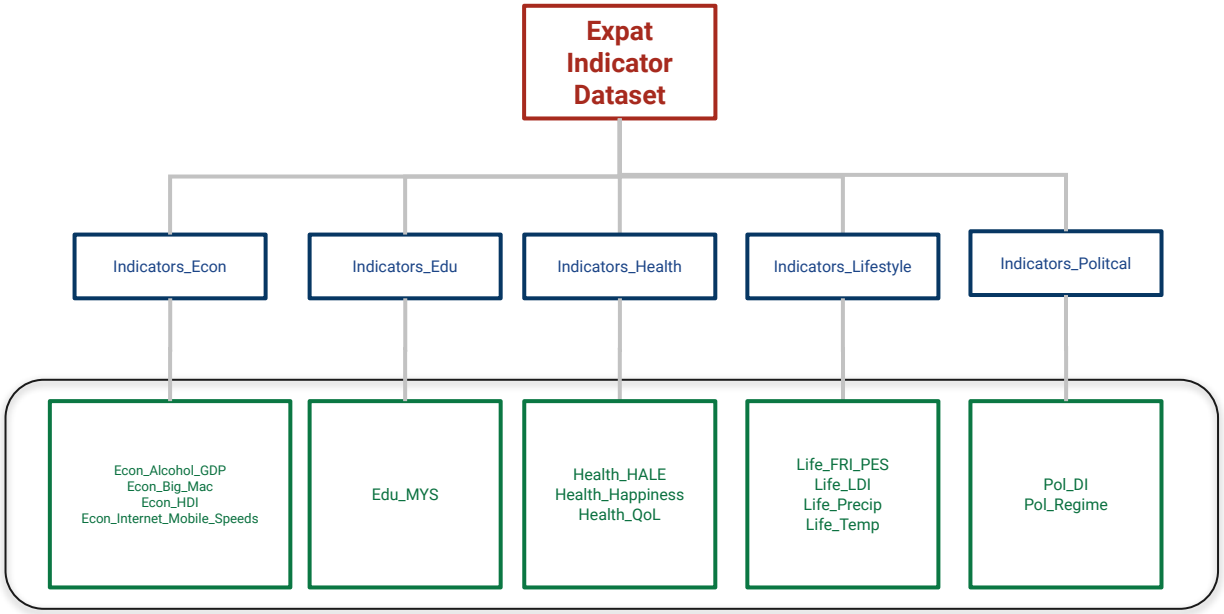
ExPatApp Database ERD



ExPatApp SQL Database Overview



RawData Schema



- ➔ Includes 14 static tables
- ➔ All tables have the following columns:
  - Country Name
  - Data Year (2000-2022)
  - Proxy Indicator/Metric

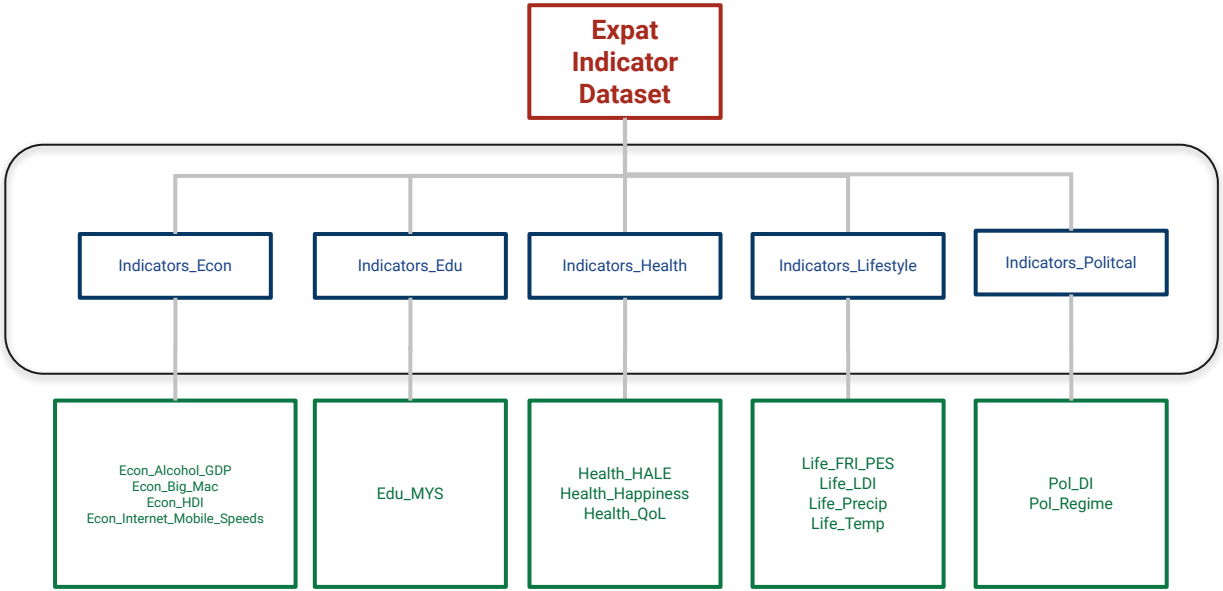
Country Code Mapping

- ➔ In order to combine all the raw datasets together to create the indicator datasets, **all** the **country name** and **data year** columns would need to match each other
- ➔ **ISO 3166-1 alpha-3** (ISO3) codes are standardized three-letter country codes defined and published by the International Organization for Standardization (ISO) → represent countries, dependent territories, and special areas of geographical interest
- ➔ Therefore, we created **country code map tables** by performing *full joins* with the *country name* in the ISO3_codes table and the *country names* of the raw datasets.

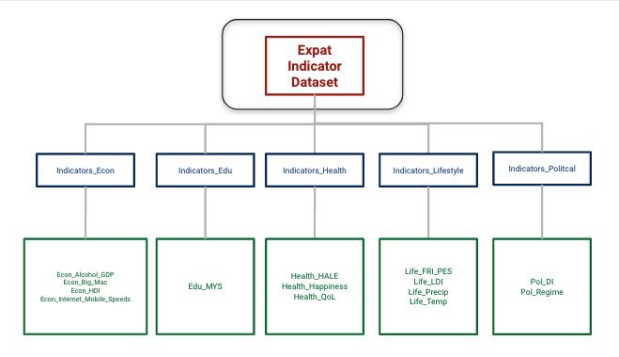
country_code	country_iso	country_fri_pes	country_ldi	country_temp	country_precip
AUS	Australia	Australia	Australia	Australia	Australia
BRA	Brazil	Brazil	Brazil	Brazil	Brazil
CAN	Canada	Canada	Canada	Canada	Canada
FSM	Micronesia (Federated States of)	Federated States of Micronesia	Micronesia, Federated States of	Federated States of Micronesia	Micronesia
HKG	China, Hong Kong SAR	Hong Kong		Hong Kong	Hong Kong
IND	India	India	India	India	India
ITA	Italy	Italy	Italy	Italy	Italy
JPN	Japan	Japan	Japan	Japan	Japan
MCO	Monaco		Monaco	Monaco	Monaco
USA	United States of America	United States	United States	United States	United States

Indicator Datasets

- **Country_year** – this table lists all the distinct combinations of country code and data year (5,750 rows)
- Raw data tables within each indicator category are then joined (left) together by:
  - Country Code
  - Data Year



ExPatApp Indicator Dataset



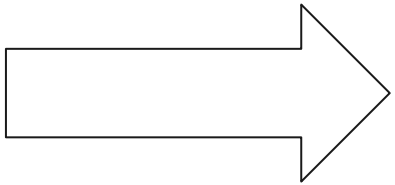
LEFT JOIN: ExPat_Indicator_Dataset	
PK, FK	country_code_year varchar(8)
	country_code varchar(3)
	country varchar(100)
	data_year int
	data_year int
	human_development_index decimal(5,4)
	alcohol_consumption_per_capita decimal(10,4)
	gdp_per_capita decimal(10,4)
	big_mac_dollar_price decimal(5,3)
	broadband_speed_rank int
	broadband_mbps decimal(5,2)
	mobile_speed_rank int
	mobile_mbps decimal(5,2)
	hdi_code varchar(25)
	mean_years_schooling decimal(10,8)
	health_adjusted_life_expectancy decimal(9,7)
	happiness_index decimal(4,3)
	numbeoqol decimal(5,2)
	usnewsqol int
	ceoqol decimal(5,2)
	freedom_religion_index decimal(9,8)
	percent_english_speakers decimal(5,2)
	linguistic_diversity_index decimal(4,3)
	avg_annual_precipitation int
	avg_annual_temp_c decimal(4,2)
	democracy_index decimal(3,2)
	regime_type varchar(100)

From **country_year** table:
ensure all possible combinations of country and year appear in dataset

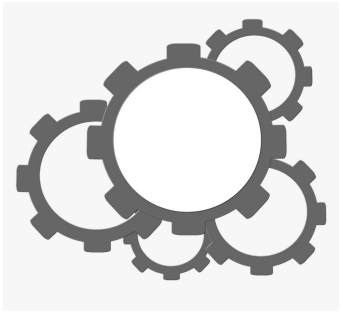
SQL to Python

INPUT

ExPat_Indicator_Dataset



MACHINE LEARNING MODEL



- ➔ A connection string via the **psycopg2-binary** Python package can potentially be used to connect PostgreSQL and Python
- ➔ Currently importing the **CSV version** of the dataset into Python for testing and ease of use

Machine Learning Model

df_expat

	country_code_year	country_code	country	data_year	human_development_index	alcohol_consumption_per_capita	gdp_per_capita	big_mac_dollar_price
0	ABW_2000	ABW	Aruba	2000	NaN	NaN	41499.6385	NaN
1	ABW_2001	ABW	Aruba	2001	NaN	NaN	39388.3526	NaN
2	ABW_2002	ABW	Aruba	2002	NaN	NaN	37256.6597	NaN
3	ABW_2003	ABW	Aruba	2003	NaN	NaN	37200.0589	NaN
4	ABW_2004	ABW	Aruba	2004	NaN	NaN	39440.6679	NaN
...	...	...	...	...	...	...	...	...
6316	ZWE_2018	ZWE	Zimbabwe	2018	0.569	4.67	3341.6654	NaN
6317	ZWE_2019	ZWE	Zimbabwe	2019	0.571	NaN	3027.6560	NaN
6318	ZWE_2020	ZWE	Zimbabwe	2020	NaN	NaN	2744.6908	NaN
6319	ZWE_2021	ZWE	Zimbabwe	2021	NaN	NaN	NaN	NaN
6320	ZWE_2022	ZWE	Zimbabwe	2022	NaN	NaN	NaN	NaN

6321 rows × 26 columns

Machine Learning Model

➔ Goals of the ML model

- **Cluster** by similarity
- Apples-to-apples comparison
- Use all the data we can

➔ Solutions to meet goals:

- **Unsupervised** learning
- Scale/encode all data
- Use some data as **filters**

```
import pandas as pd
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.decomposition import PCA
from sklearn.cluster import AgglomerativeClustering
import hvplot.pandas
import plotly.figure_factory as ff
```

Machine Learning Model

```
for code in df_expat_latest['country_code']:
  row_total_fudge_factor = 0
  df_placeholder = df_expat_encoded[df_expat_encoded['country_code']==code]
  for column in df_expat_cleaned.columns:
    year_placeholder = 0
    index_value_placeholder = -1
    column_fudge_factor = 0
    for index, row in df_placeholder.iterrows():
      if row['data_year'] > year_placeholder and df_placeholder.notnull().loc[index, column]:
        year_placeholder = row['data_year']
        column_fudge_factor = (2022-year_placeholder)
        index_value_placeholder = row[column]
    if index_value_placeholder > -1:
      df_expat_latest.loc[df_expat_latest[df_expat_latest['country_code']==code].index.values.astype(int)[0], column] = index_value_placeholder
    else:
      df_expat_latest.loc[df_expat_latest[df_expat_latest['country_code']==code].index.values.astype(int)[0], column] = None
    row_total_fudge_factor -= column_fudge_factor
  df_expat_latest.loc[df_expat_latest[df_expat_latest['country_code']==code].index.values.astype(int)[0], 'fudge_factor'] = row_total_fudge_factor
```

- ➔ Build “latest” DataFrame from raw data
- ➔ “Fudge factor” tracks old data
- ➔ Encode/scale & simplify with PCA

country_code	fudge_factor
ABW	-36
AFG	-61
AGO	-66
AIA	-31

```
df_expat_scaled = StandardScaler().fit_transform(df_expat_latest_cleaned)
df_expat_scaled
```

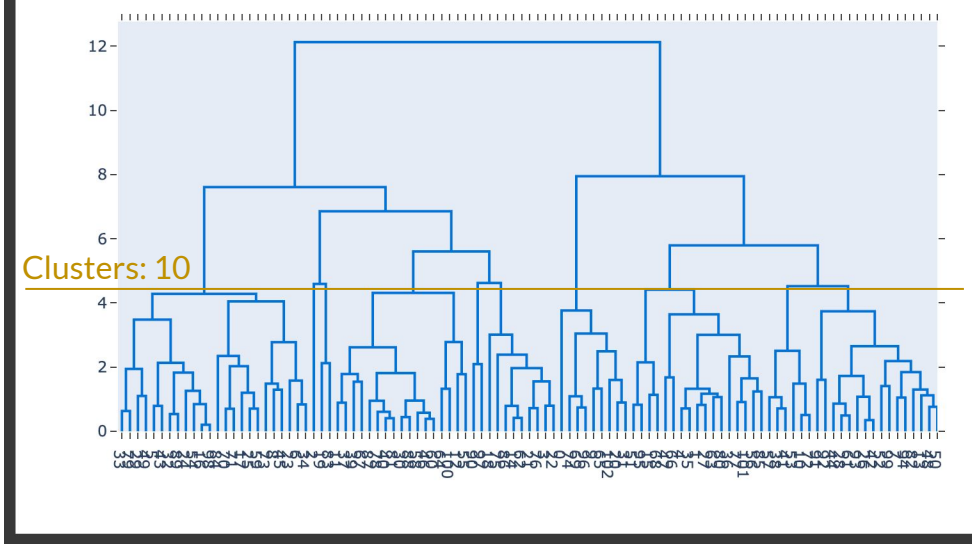
```
# Initialize PCA model for 4 principal components
pca = PCA(n_components=4)
```

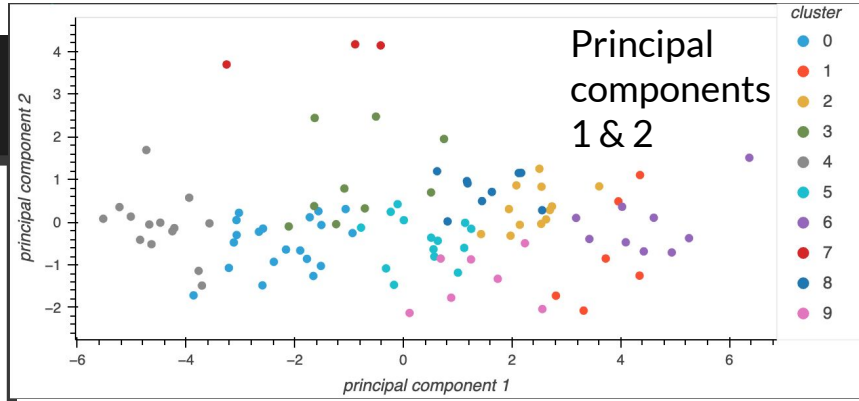
```
pca.explained_variance_ratio_
```

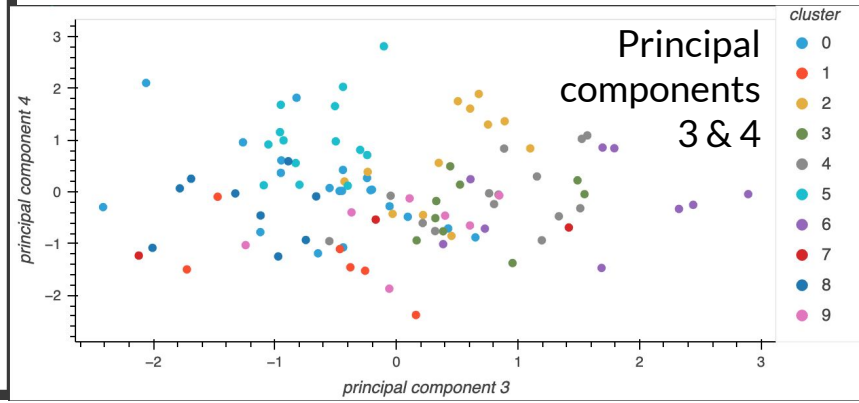
```
array([0.56958387, 0.09201607, 0.07700488, 0.06488941])
```

Machine Learning Model

```
# Create a dendrogram to visualize different cluster counts
fig = ff.create_dendrogram(df_expat_pca, color_threshold=0)
fig.update_layout(width=800, height=500)
fig.show()
```







Results

- ML model produces ten country clusters
- Countries most “alike” are grouped together
- Can find “companion countries” easily 🌎

USlike_cluster

```
['Australia',
 'Canada',
 'Switzerland',
 'Germany',
 'Denmark',
 'Finland',
 'Ireland',
 'Republic of Korea',
 'Luxembourg',
 'Netherlands',
 'Norway',
 'Singapore',
 'Sweden',
 'United States of America']
```

Chilelike_cluster

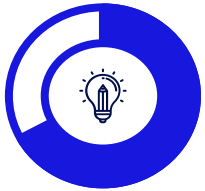
```
['Austria',
 'Belgium',
 'Chile',
 'Spain',
 'Estonia',
 'France',
 'United Kingdom',
 'Hungary',
 'Israel',
 'Italy',
 'Japan',
 'Lithuania',
 'Latvia',
 'New Zealand',
 'Poland',
 'Portugal',
 'Romania',
 'Slovakia',
 'Slovenia',
 'Thailand']
```

Saudilike_cluster

```
['United Arab Emirates', 'China', 'Saudi Arabia']
```

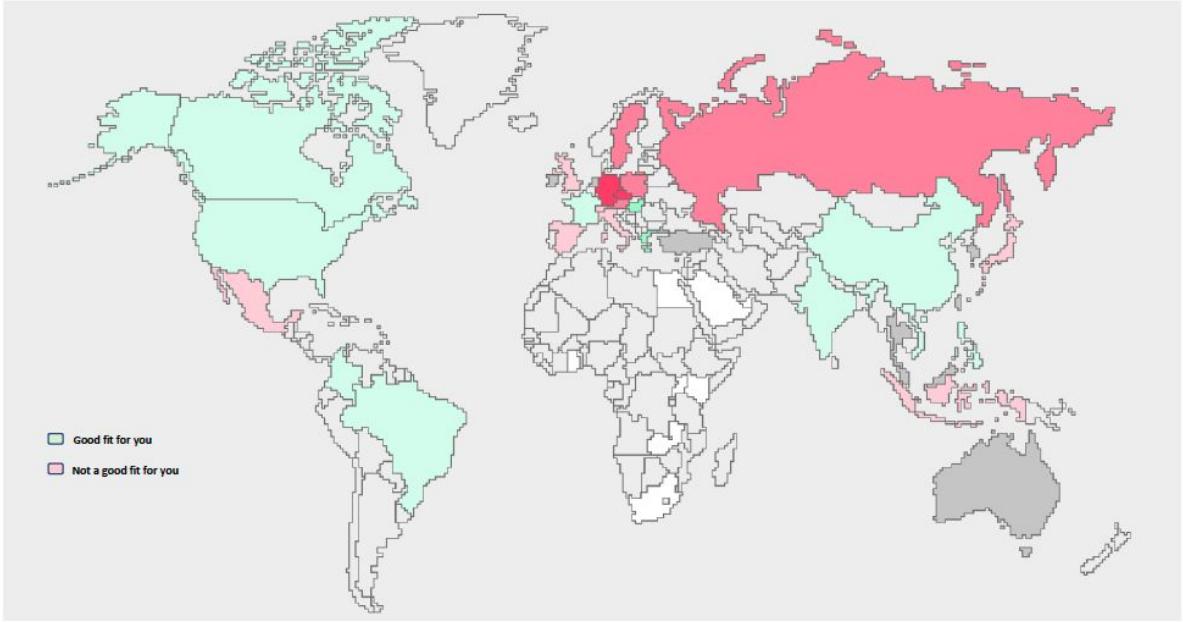

Dashboard
5 4 6 7 8 10 12 14 16
OFF 1 2 3 4 5 DIAL
EMP
9

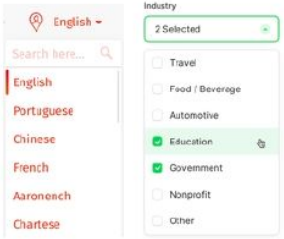
ExPatDash



Economic Development	Education System
70%	50%
High GNI per capita	Mid education outcomes

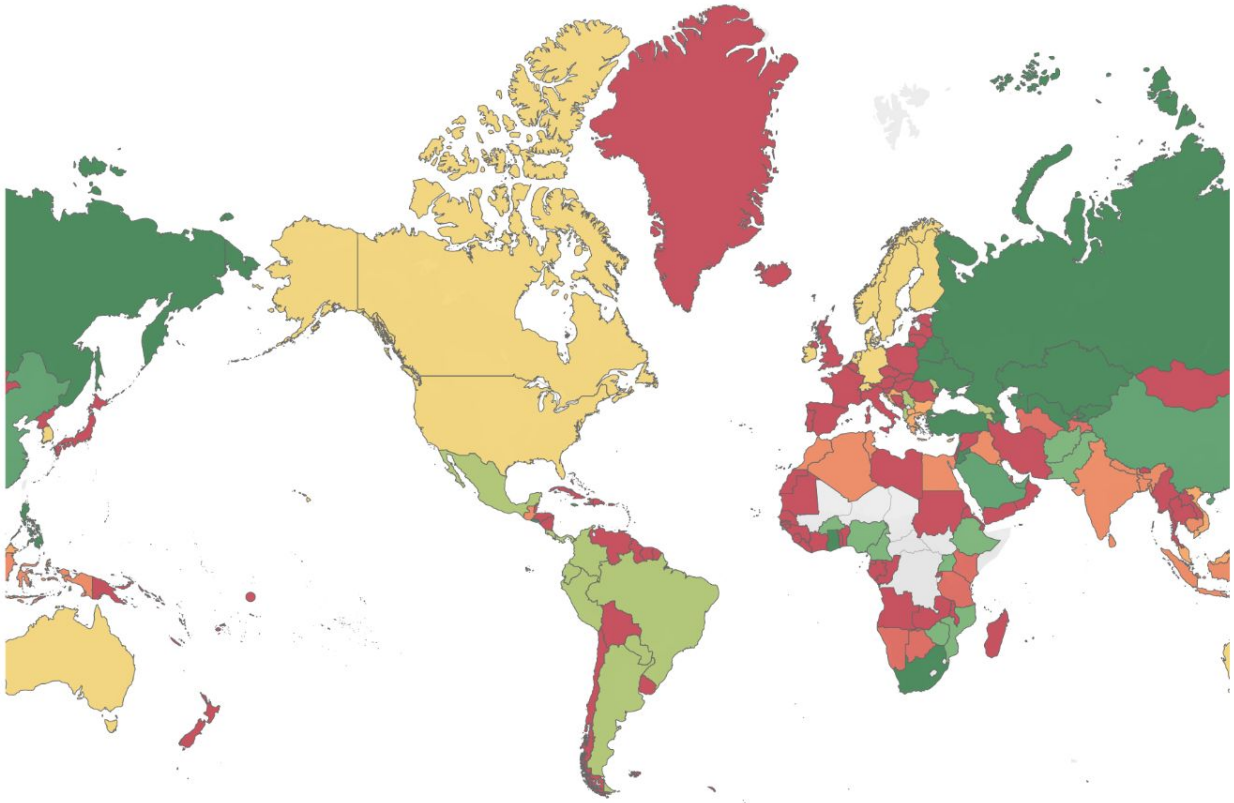
Mock Dashboard from ML Algorithm Rankings





Dashboard Main Page

Overview of Results by Cluster and ML main indicators



Democracy Index
0.320 9.750

Freedom Religion Index
0.0000 0.8319

Gdp Per Capita
1,229 110,261

Human Development Index
0.4520 0.9570

Cluster
-1.00 9.00

Health Adjusted Life Expe...
50.80 74.48

Cluster
0 9

Persona “A”

Outputs: Chile, Hungary, Romania, and South Korea

Age: 20s-30s

Relationship status: Single

Employment: Full-time WFH

Drinks alcohol: Yes

Foreign language: None

Financing: Little to none

Reason for leaving USA: Looking for a different political/socio-economic system

Persona “B”

Outputs: France, Italy, Saudi Arabia, and Australia

Age: 55+

Relationship status: Married

Employment: Retired professor

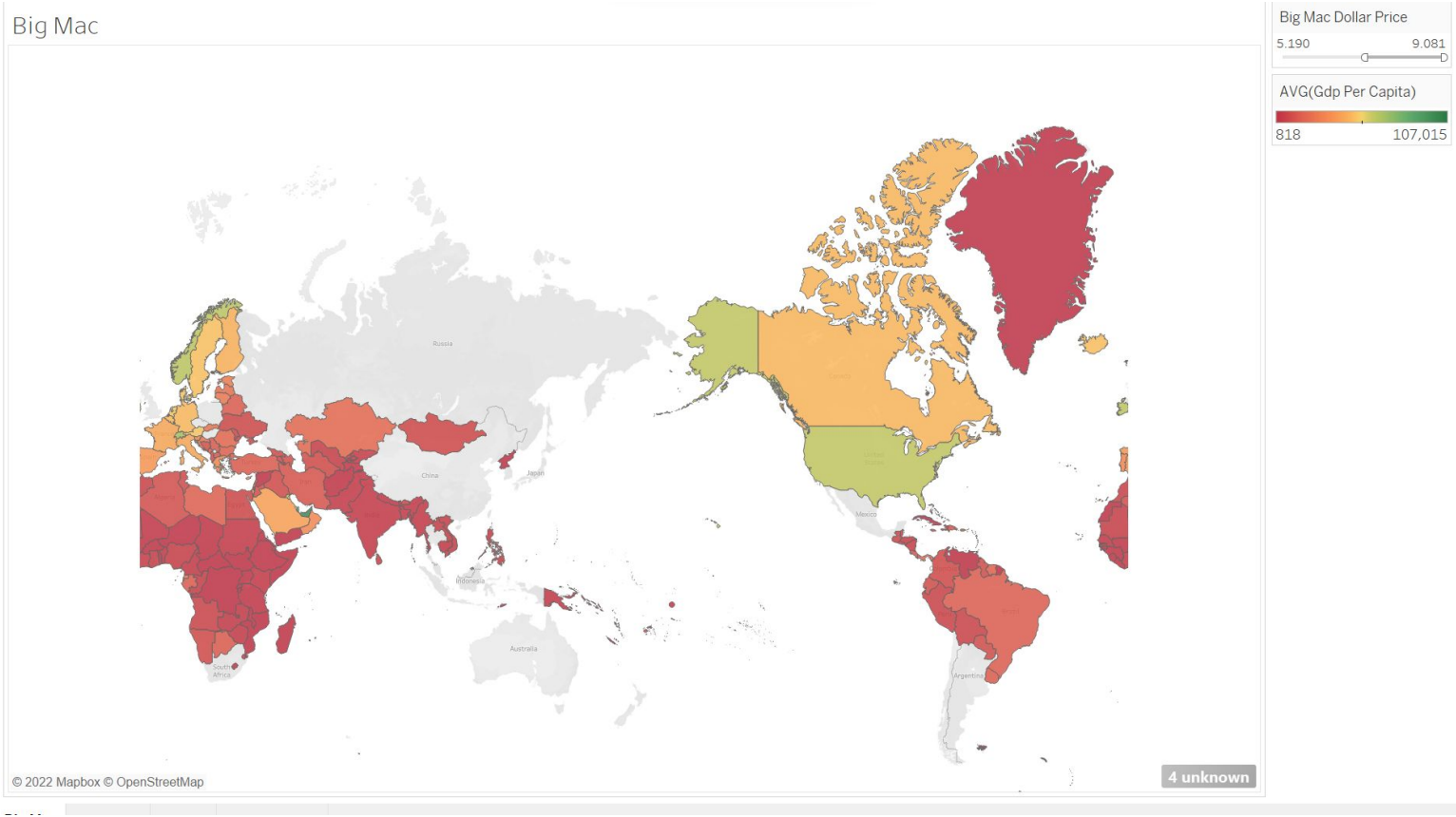
Drinks alcohol: A little

Foreign language: Conversational Spanish and French

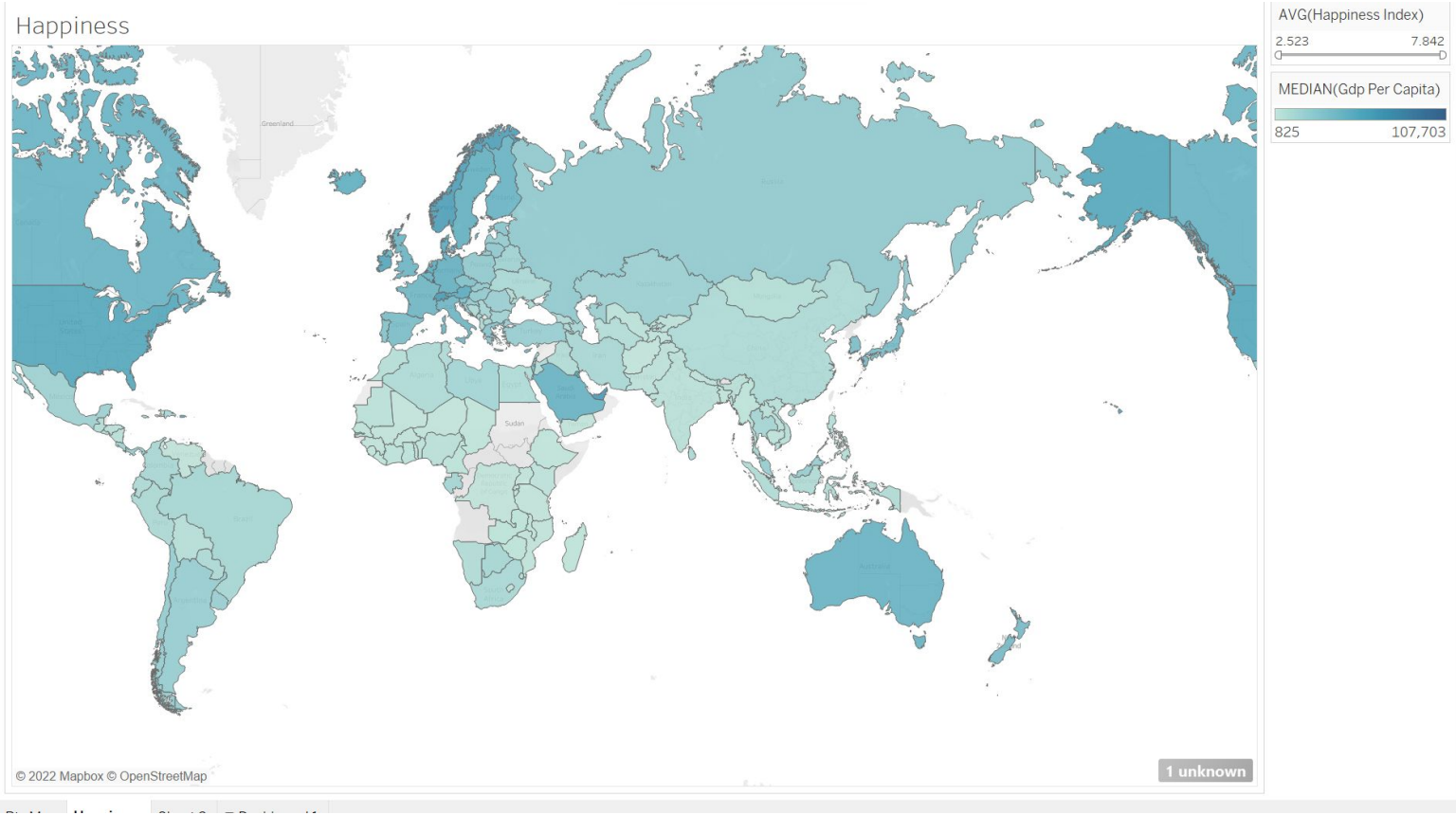
Financing: Full pension (lot$)

Reason for leaving USA: Looking for a second home country to spend their retirement

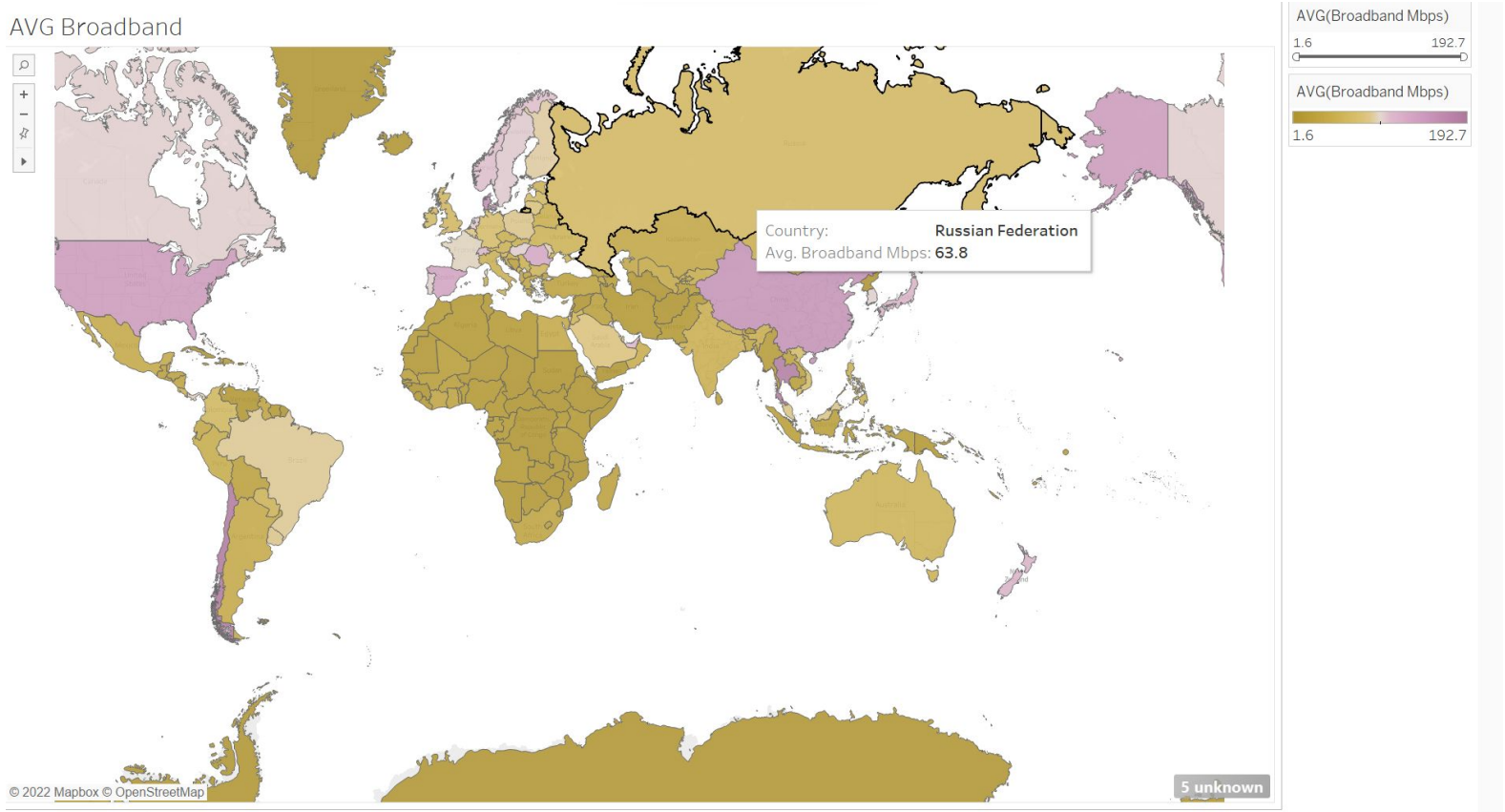
Economic Dashboard Heatmap (GDP per capita and Big Mac Index (USD)



Happiness Filter with Economy Heatmap (GDP per capita, USD)



Internet Speed (Broadband Mbps) Filter and Heatmap



Moving Forward

Hindsight is 20/20

- ➔ Data limitations
- ➔ ETL
- ➔ Dashboard vs. Web App
- ➔ UX Design

Future Analysis

- ➔ Expand range of data
  - Socio-economics
  - Safety
  - Ease
  - Culture
  - Expatriates; Emigration vs. Immigration
- ➔ Upgrade UX
- ➔ Enhance resources

Fin.