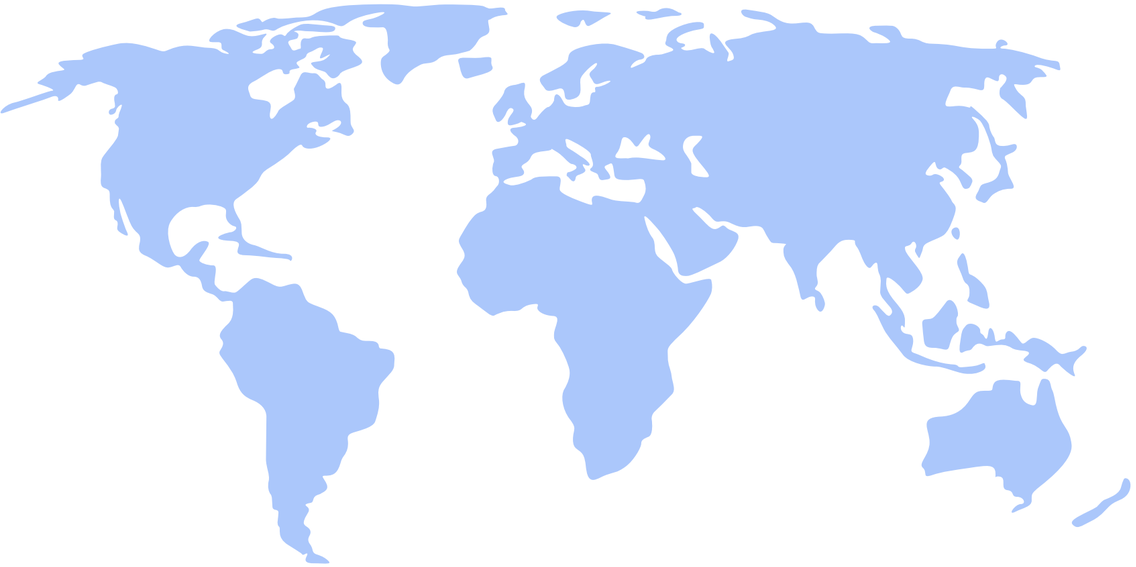


# ETZ - GDP

**Data Science presentation by:**
**Eytan Rahlin**
**Tony Hasson**
**Ziv Simchoni**

# LETS MEET
# OUR TEAM MEMBER

## Ziv simchoni

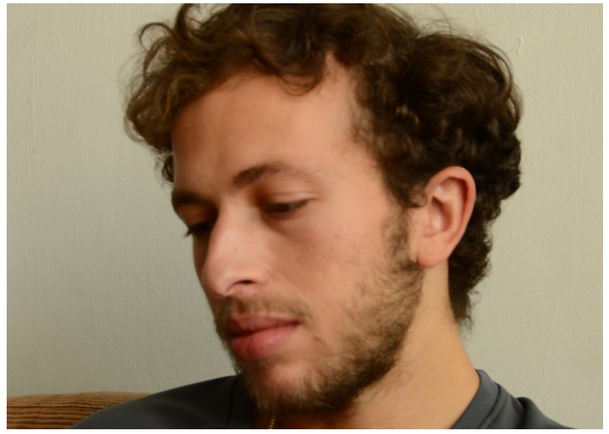
**Leading Developer**

**Technology enthusiast and a passionate learner with a keen interest in Python and Android ROMs.**

## Tony Hasson

**Leading Developer**

**A software dev with the goal of achieving expertise in Python, Data analyzing & management.**



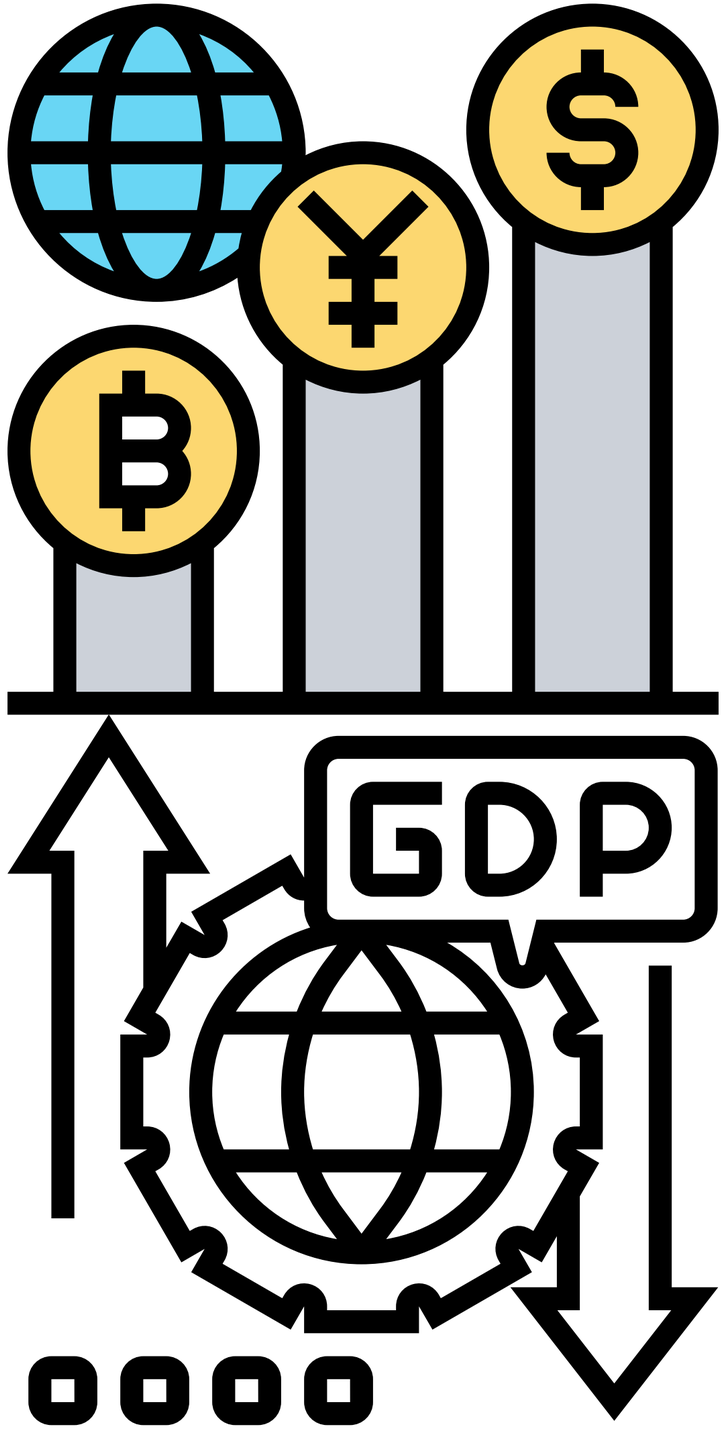
## Eytan rahlin

**Leading Developer**

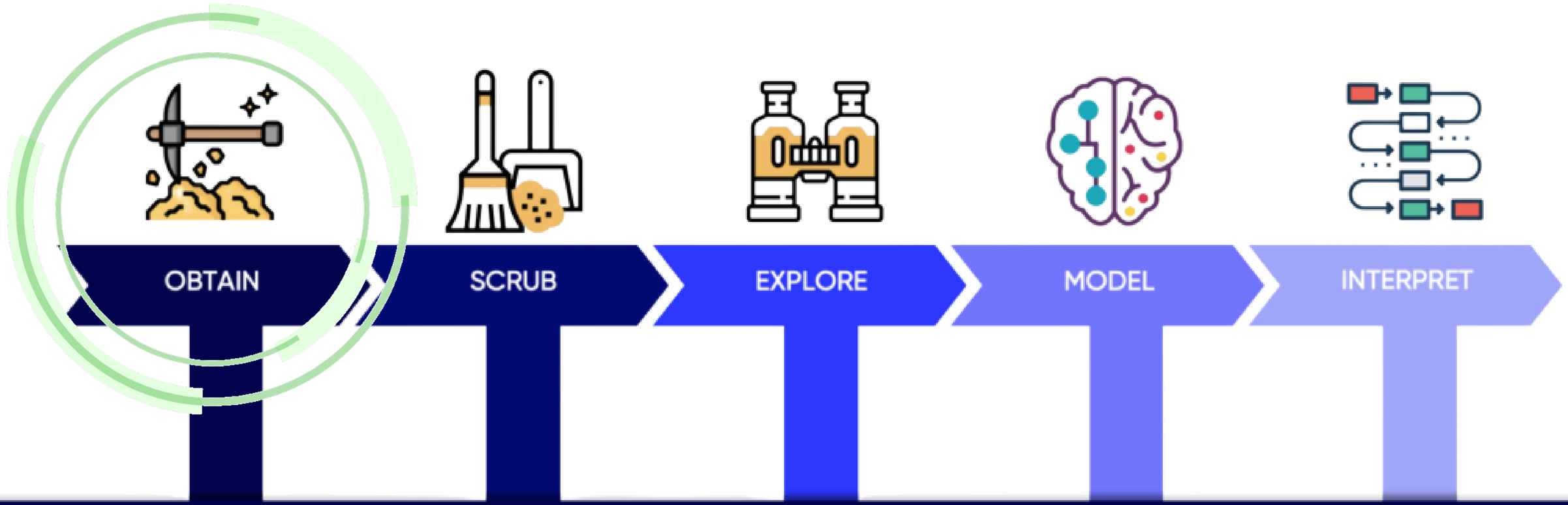
**Outgoing and responsible with a strong passion for programming. Unique problem solving approach.**

# WHAT IS GDP?

- Gross domestic product (GDP) is the monetary value of all finished goods and services made within a country during a specific period.
- GDP provides an economic snapshot of a country, used to estimate the size of an economy and growth rate.
- GDP is a key tool to guide policy-makers, investors, and businesses in strategic decision-making.



# Data Science Process




**O** — **Gather data from relevant** sources

**S** — **Clean data to formats** that machine understands

**E** — Find significant patterns and trends using statistical methods

**M** — Construct models to predict and forecast

**N** — Put the results into good use

# OBTAIN THE DATA

**Scraping:** Wikipedia.org, Numbeo.com, OurWorldInData.org, Burningcompass.com
**CSVs:** data.worldbank.org

**Step**

**01** We've scraped the relevant data and built a CSV from every resource.

**02** We've realized that we cannot combine all the CSVs into one database thats because the combined DB would have too much missing data - due to not overlapping years.

**03** We decided to build a separate CSV for the scraping and for the downloaded data.
It was much easier to use two different CSV's as our main data points.
"df_Full_DataBase" contains all data from 1960 through 2020
"df_Scrape" contains all data from 2009 through 2020

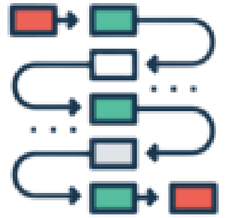

## What data have we gathered?

Various GDP information, Indexes growth rate, Population Data, Cost of living and spending per country.

## Have we encountered data duplications?

Yes, we had a few countries that appeared more than once but with a different name.

# Data Science Process



OBTAIN | SCRUB | EXPLORE | MODEL | INTERPRET

**O**

**Gather data from relevant** sources

**S**

**Clean data to formats** that machine understands

**E**

Find significant patterns and trends using statistical methods

**M**

Construct models to predict and forecast

**N**

Put the results into good use

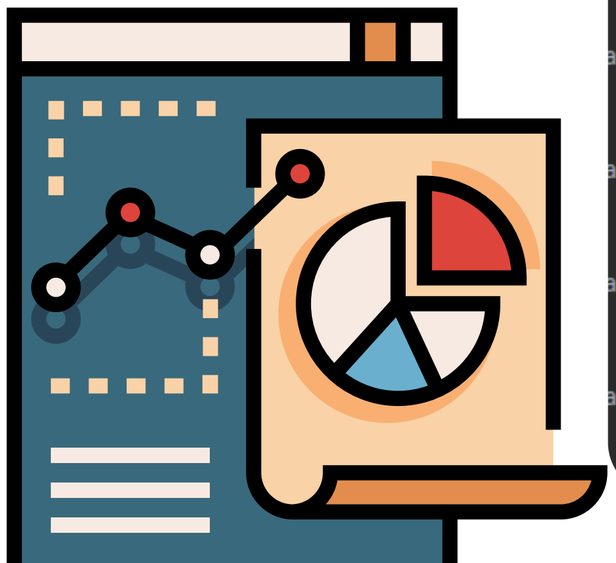
# DATA

## Country names

We noticed some countries have a lot of missing data, although the data is found in the databases.
After EDA we realized that different databases write the country names differently, see Ex.1.

## Missing Values

We've noticed there is too much missing data, hence we decided to use **linear regression** to complete the missing values.



The data is distributed between the 2 CSV's:
**df_scrape**: Cols: 20 X Rows: 1637 = 38,740
**df_Full_DataBase**: Cols: 16 X Rows: 10432 = 166,912
**Total data points: 205,652**

```
rameIterator in arr_df:
aFrameIterator.loc[
 DataFrameIterator["Country"] == r"Cote d'Ivoire", "Country"
 "Ivory Coast"
aFrameIterator.loc[
 DataFrameIterator["Country"] == "Slovak Republic", "Country"
 "Slovakia"
aFrameIterator.loc[
 DataFrameIterator["Country"] == "Yemen, Rep.", "Country"
 "Yemen"
aFrameIterator.loc[
 DataFrameIterator["Country"] == "Egypt, Arab Rep.", "Country
 "Egypt"
aFrameIterator.loc[
 DataFrameIterator["Country"] == "Korea, Dem. Rep.", "Country
 "North Korea"
aFrameIterator.loc[
 DataFrameIterator["Country"] == "Korea, Dem. People's Rep.",
 "North Korea"
aFrameIterator.loc[
 DataFrameIterator["Country"] == "Korea, Rep.", "Country"
"South Korea"
```

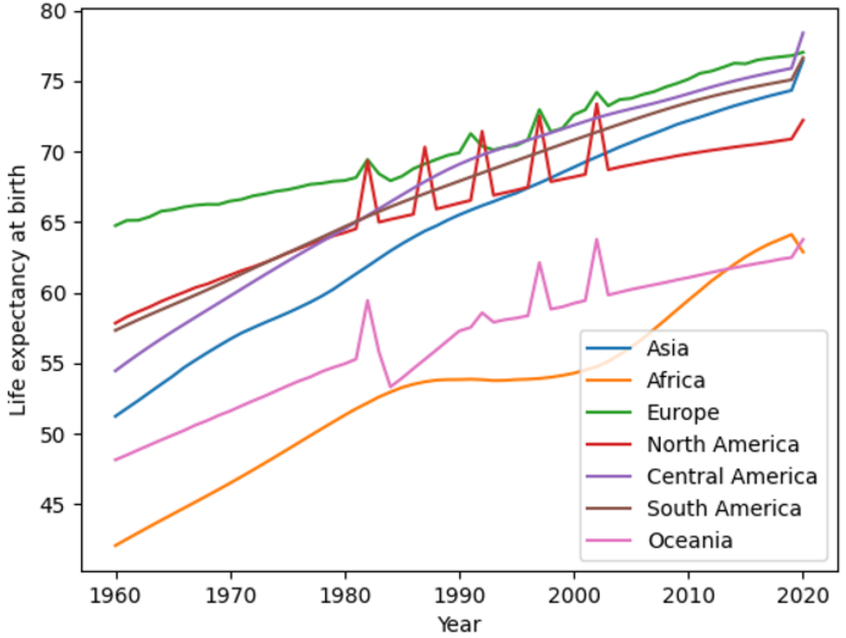
Ex.1

# HOW HAVE WE DEALT WITH OUTLIERS?

We found out about a few outlier countries by viewing anomalies when plotting graphs.

Afterwards, we realized that many countries with population under 100k have missing/not much/false data.

In conclusion, we have decided to delete all the countries with population under 100k, and thus we have dealt with a major amount of data anomalies.


palau

69.06926829
17.58951758
17.58090746
17.57229735
17.56368723
71.84463415
17.546467
17.53785689
17.52924677
17.52063666
70.49365854
17.50341643
17.49480631
17.4861962
17.47758608
69.12926829

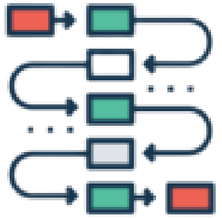
tuvalu

-33.4302439
-31.99547039
-30.56069687
-29.12592335
-27.69114983
-26.25637631
-24.82160279
-23.38682927
-21.95205575
-20.51728223
-19.08250871
-17.6477352
-16.21296168
-14.77818816
-13.34341464
-11.90864112
-10.4738676
-9.039094079
-7.60432056
-6.169547041
-4.734773522
-3.300000002
-1.865226483
-0.430452964

nauru

38.175
16.7824
1.81844
4.39982
6.9812
9.56257
12.1439
14.7253
17.3067
17.523
22.4694
17.7081
17.8007
17.8933
17.9859
37.496
18.171
18.2636
18.3561
18.4487
41.376
18.6339

# Data Science Process



OBTAIN | SCRUB | EXPLORE | MODEL | INTERPRET

**O**

**Gather data from relevant** sources

**S**

**Clean data to formats** that machine understands

**E**

Find significant patterns and trends using statistical methods

**M**

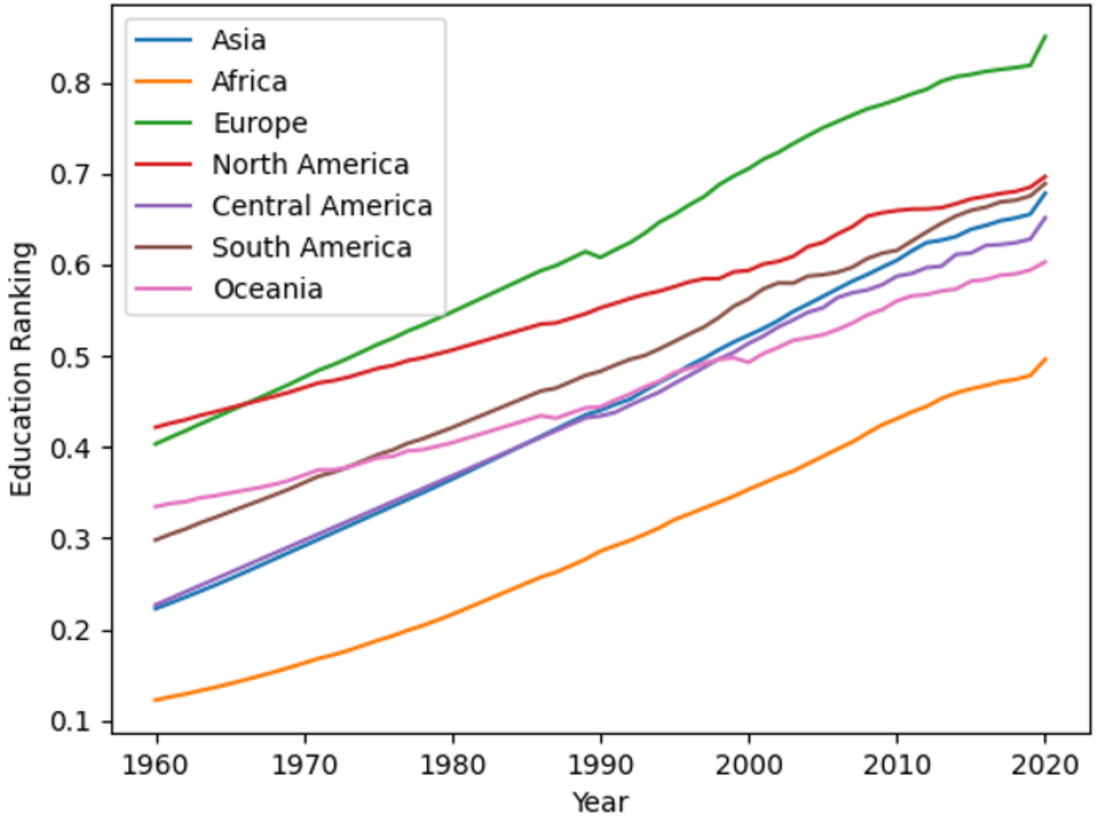
Construct models to predict and forecast

**N**

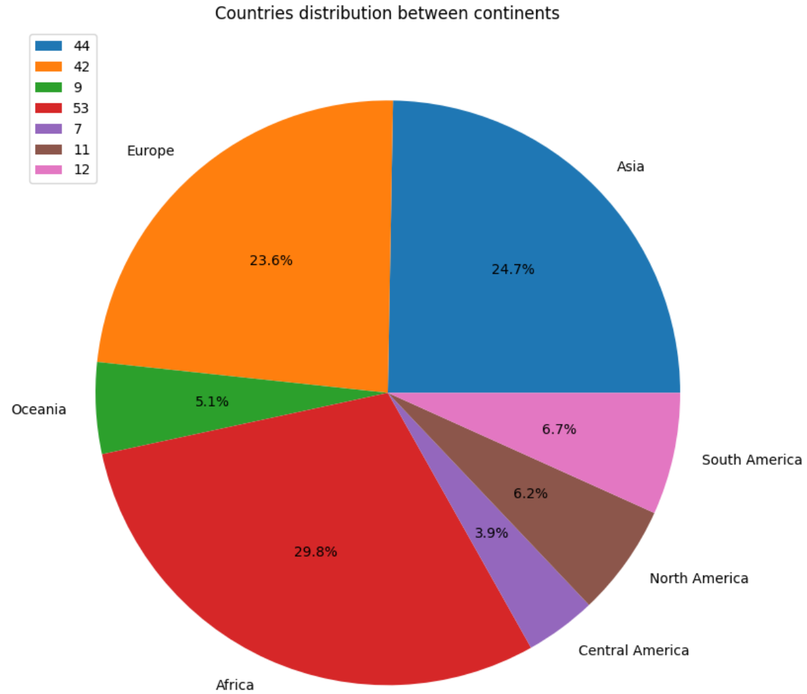
Put the results into good use

# WHAT HAVE WE LEARNED FROM VIEWING OUR DATA?

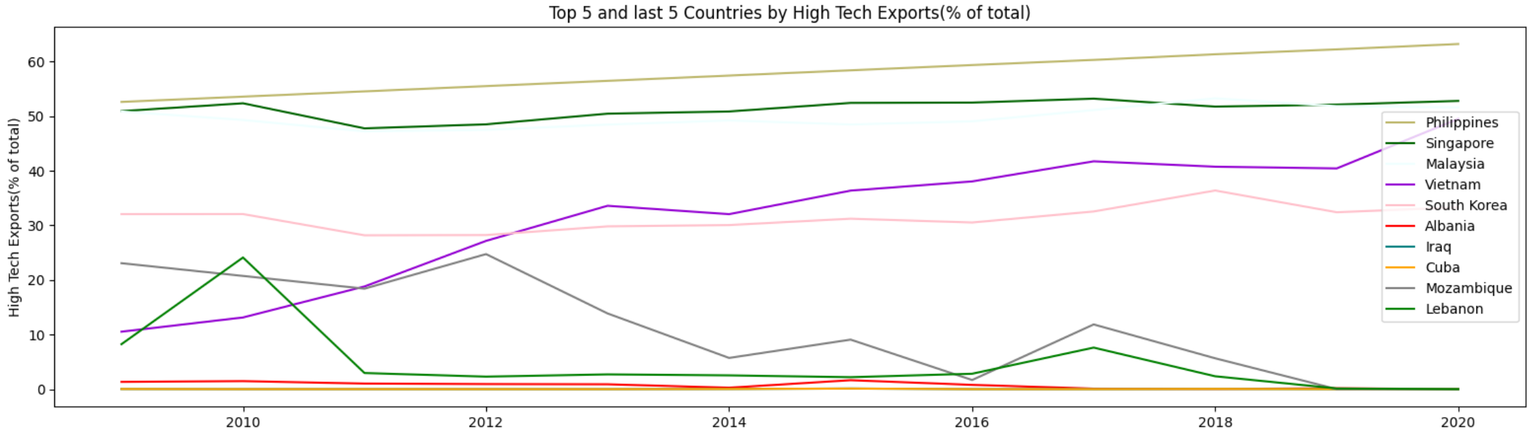
Asia had a large increase and almost surpassed North America.
Overall, education improves as the years go by.




Africa has the most countries in the world

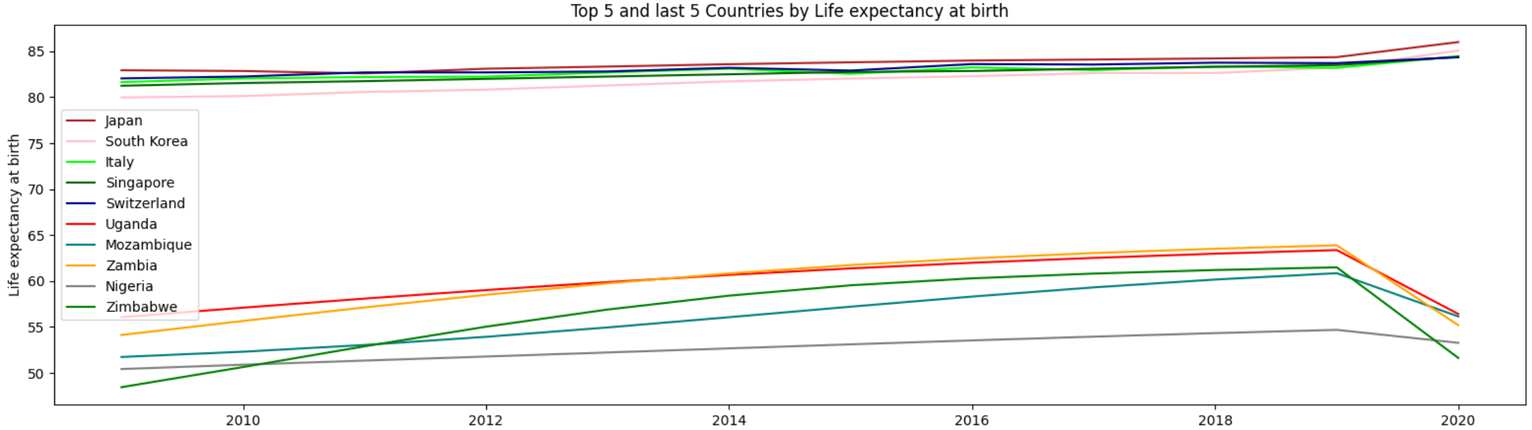


Avg: 22.2%
Israel is above avg with: 28%



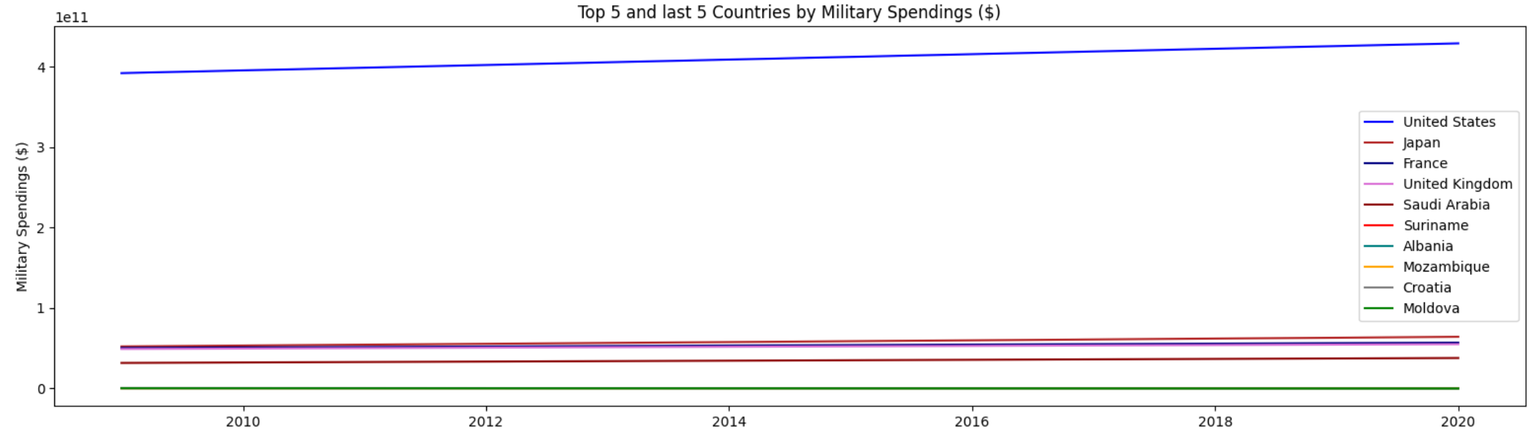

The Philippines' biggest export products by value in 2020 were electronic circuits, computers, computer parts and accessories, insulated wire or cable and printing machinery.

Avg: 72.4 years
Europe with the highest: 78 years old

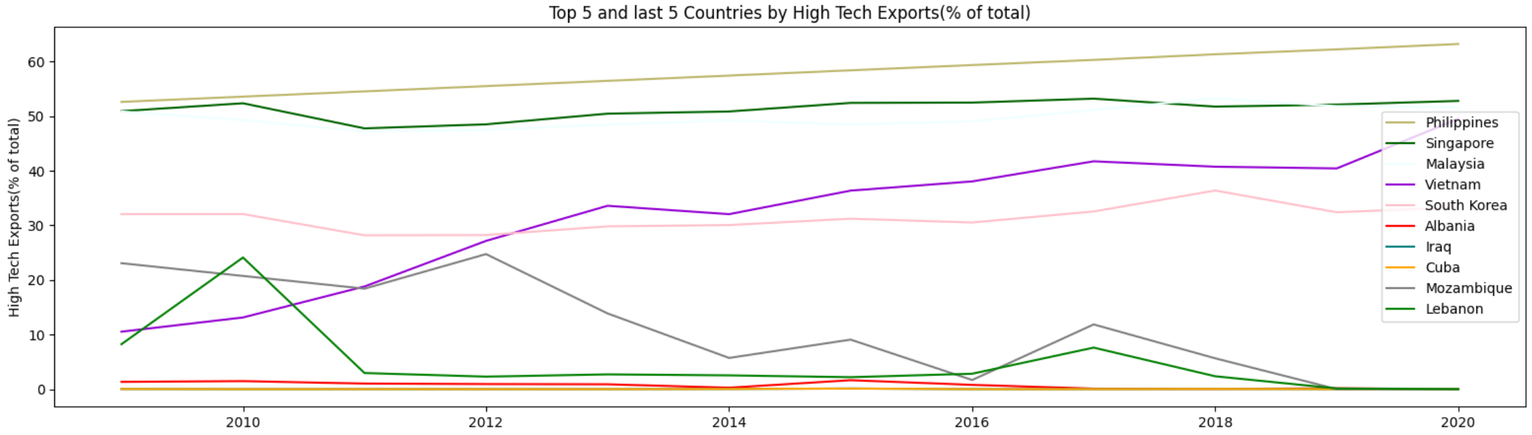



Life expectancy difference between Top 5 and Bottom 5 countries is more than 30 year!

USA spent in 2020 429 Billions$ on military
4 times more the the next big spender: Japan

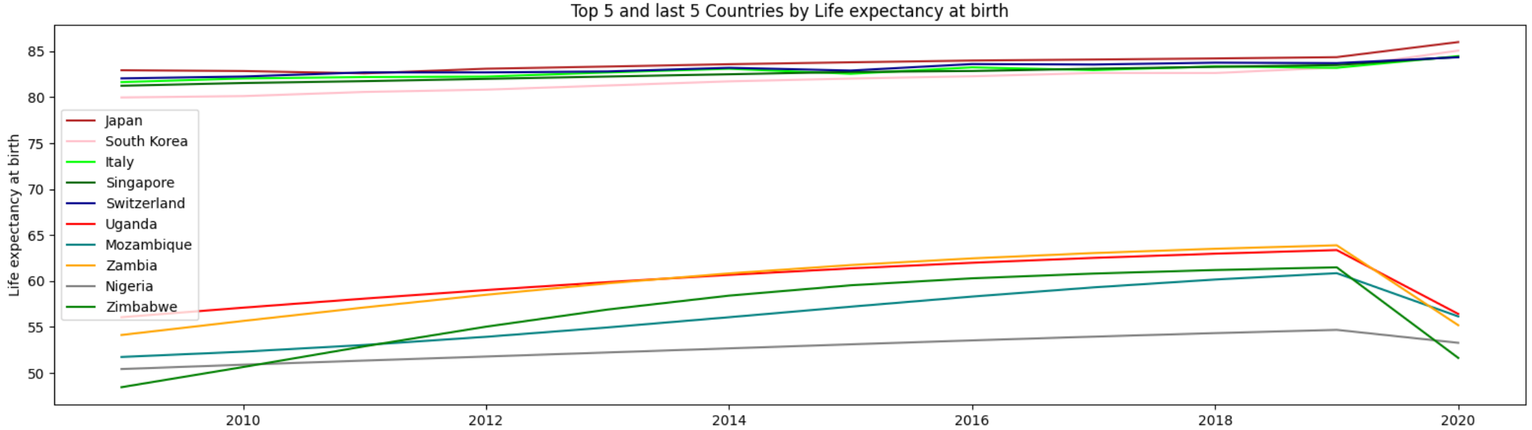


Avg: 22.2%
Israel is above avg with: 28%



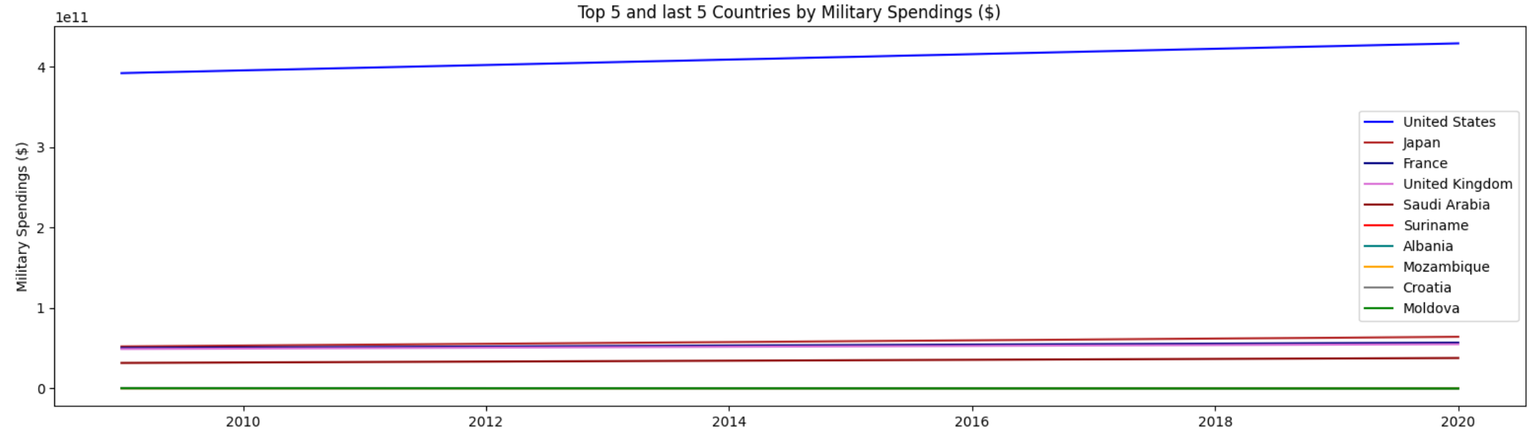

The Philippines' biggest export products by value in 2020 were electronic circuits, computers, computer parts and accessories, insulated wire or cable and printing machinery.

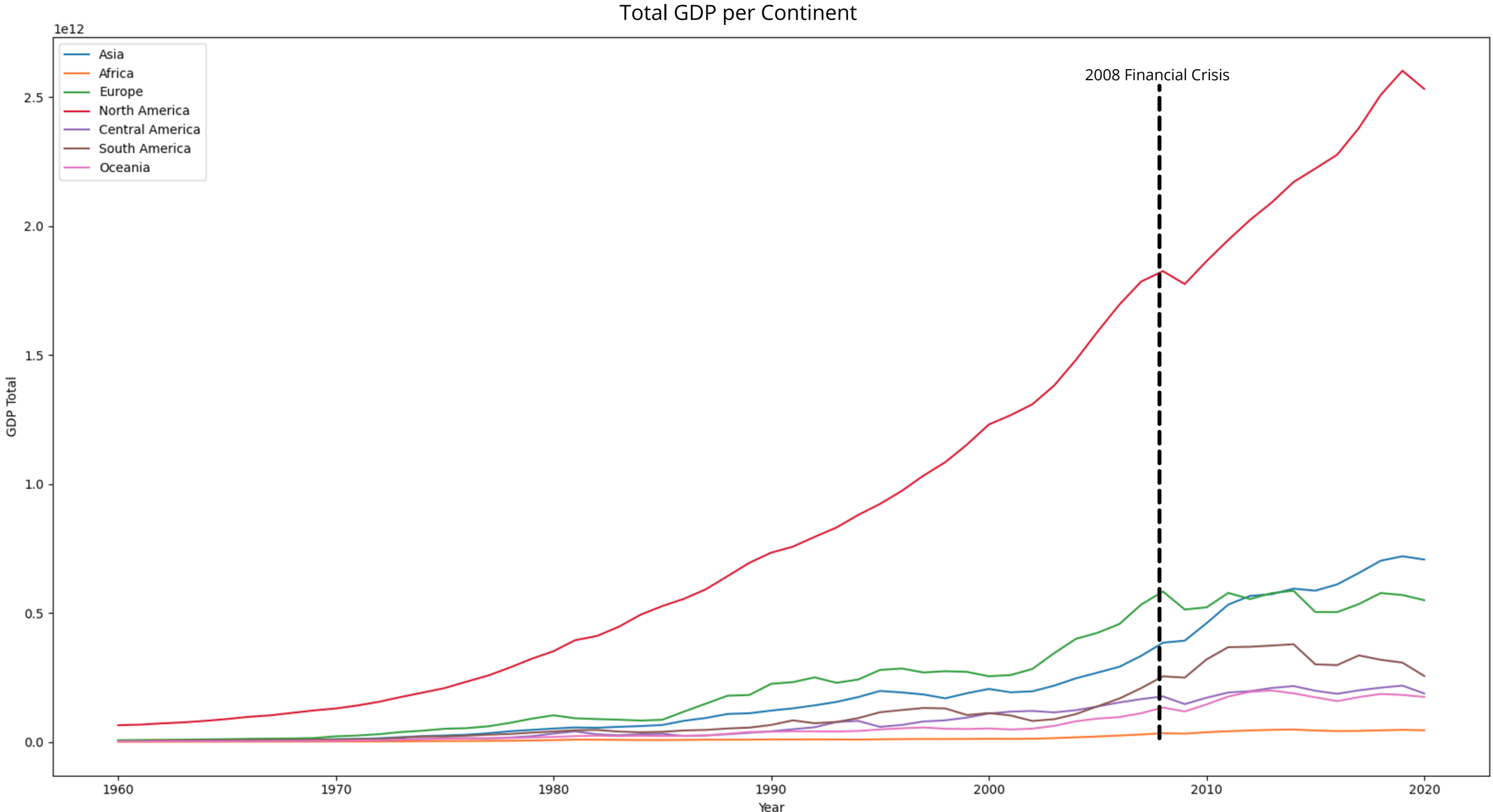
Avg: 72.4 years
Europe with the highest: 78 years old

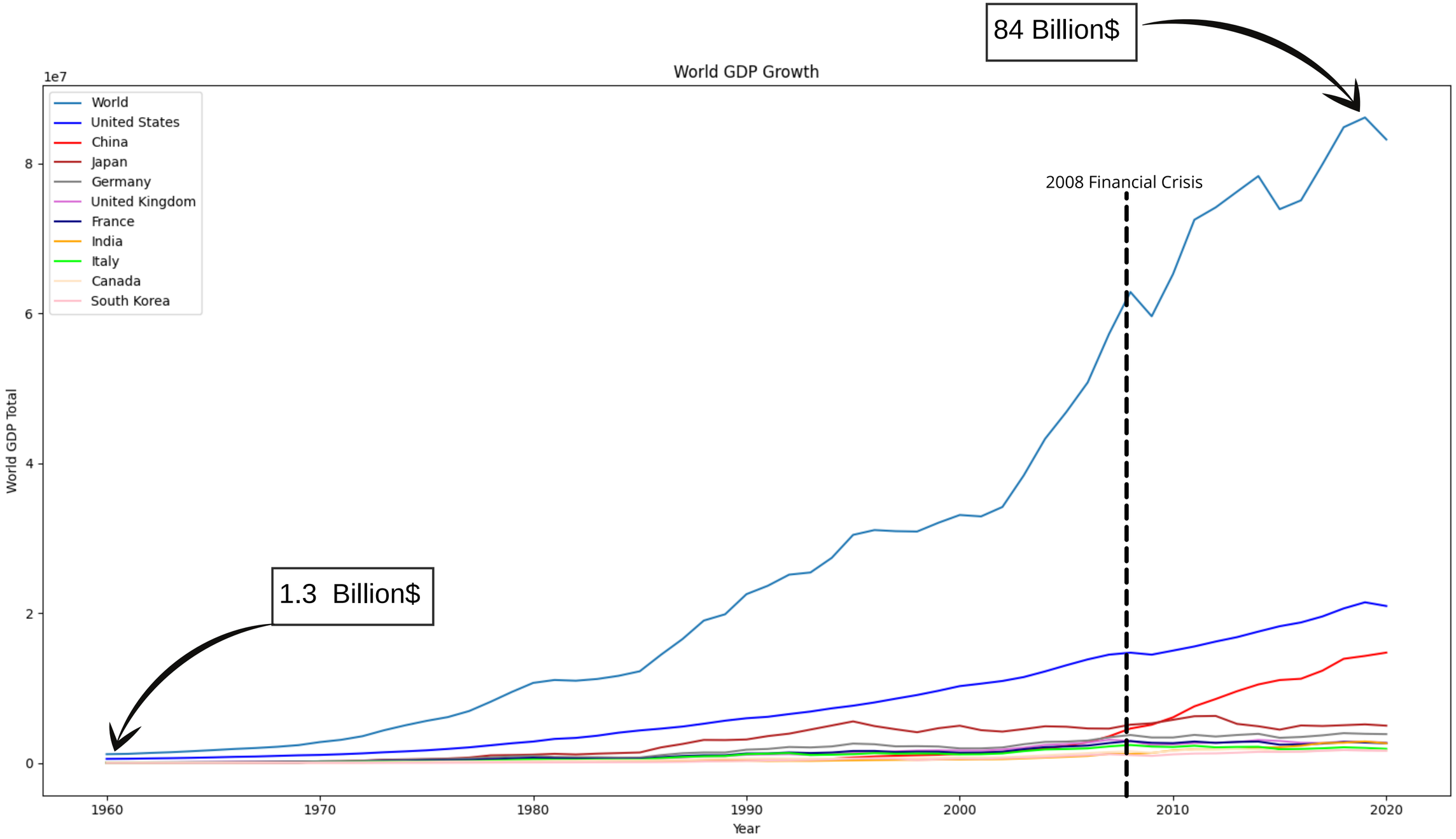



Life expectancy difference between Top 5 and Bottom 5 countries is more than 30 year!

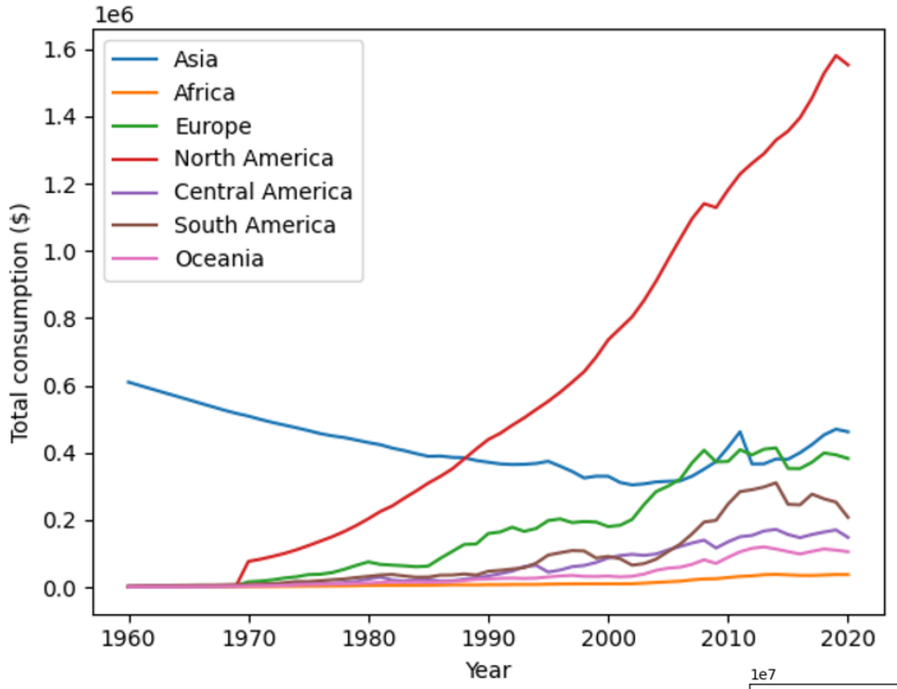
USA spent in 2020 429 Billions$ on military
4 times more the the next big spender: Japan



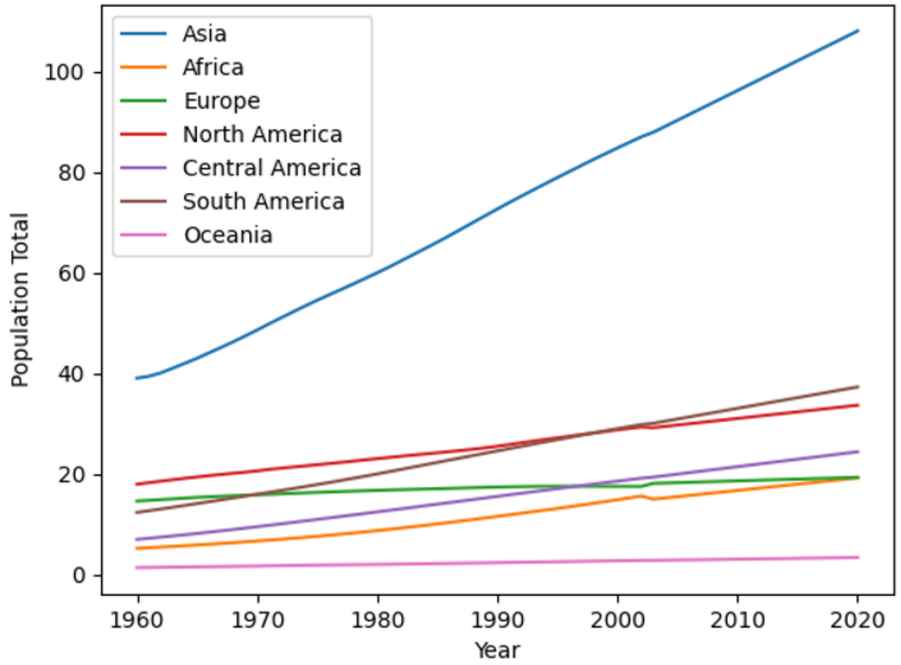

Total GDP per Continent
1e12
Asia
Africa
Europe
North America
Central America
South America
Oceania
2008 Financial Crisis
GDP Total
Year
0.0
0.5
1.0
1.5
2.0
2.5
1960
1970
1980
1990
2000
2010
2020

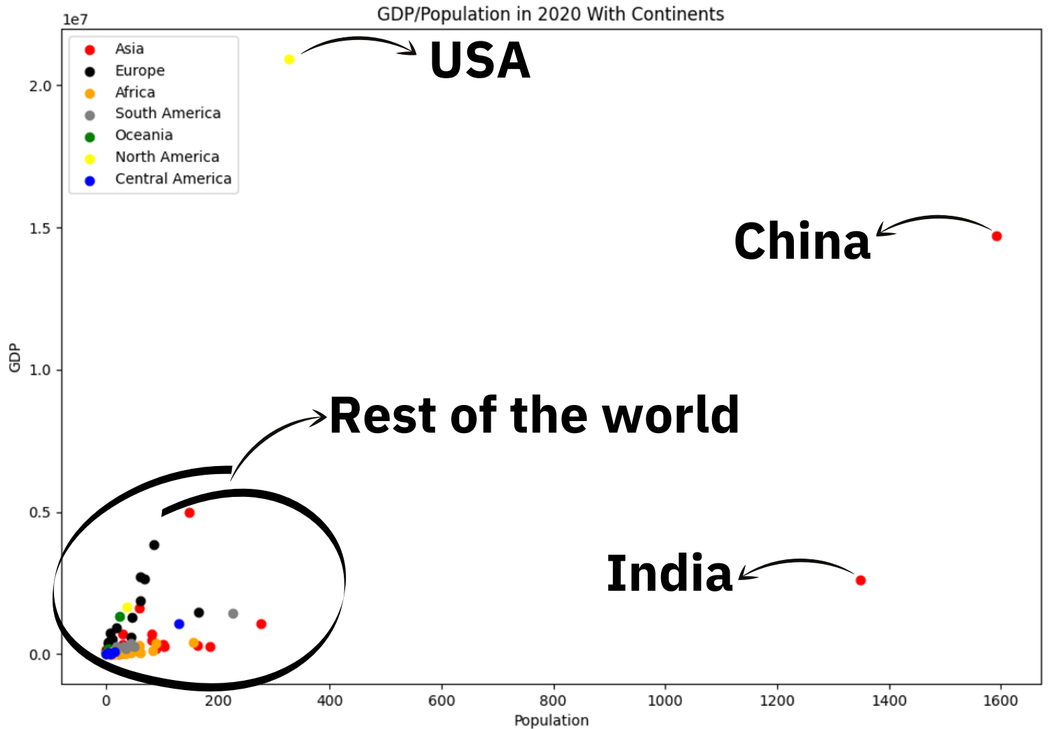

84 Billion$
World GDP Growth
1e7
World
United States
China
Japan
Germany
United Kingdom
France
India
Italy
Canada
South Korea
2008 Financial Crisis
1.3 Billion$
World GDP Total
Year
0
2
4
6
8
1960
1970
1980
1990
2000
2010
2020

# Asia's population growth is the largest by far




# North America spends the most by far



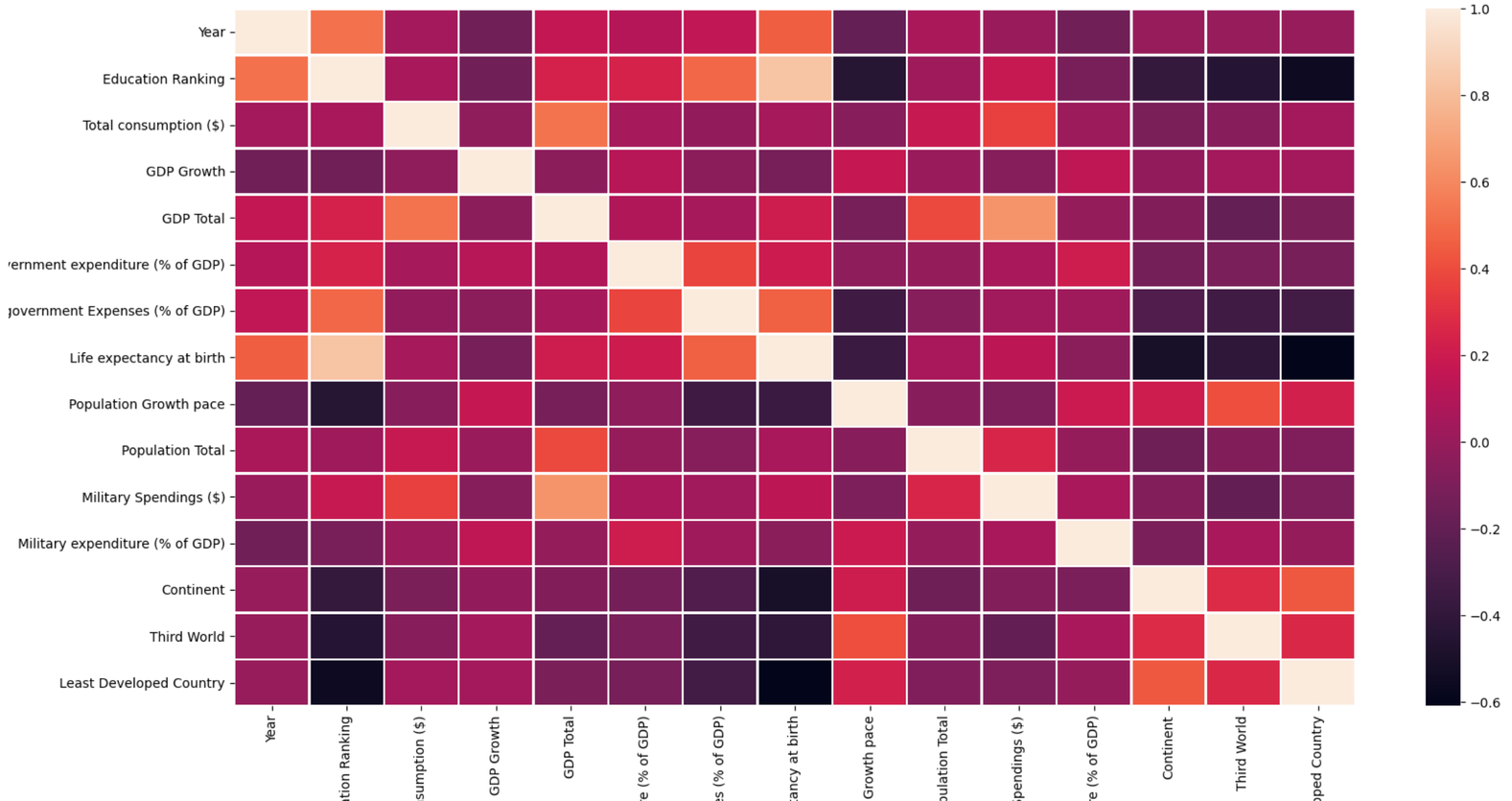



# HEAT-MAP & CORRELATION SCORE

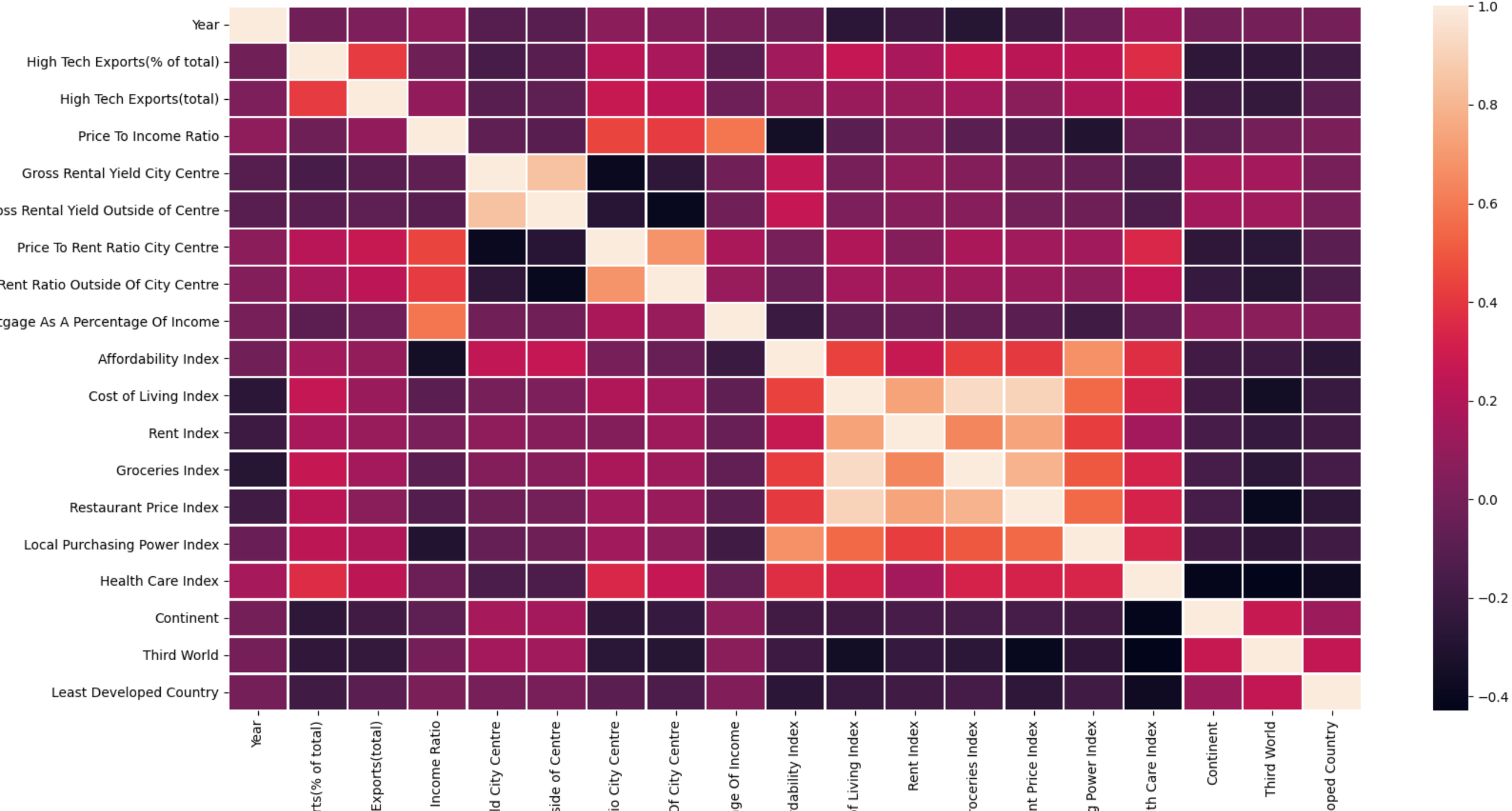
We took each Database's column and built a heatmap with correlation score.

- Iterated on highest correlated columns.
- Plot.
- Saved the best.

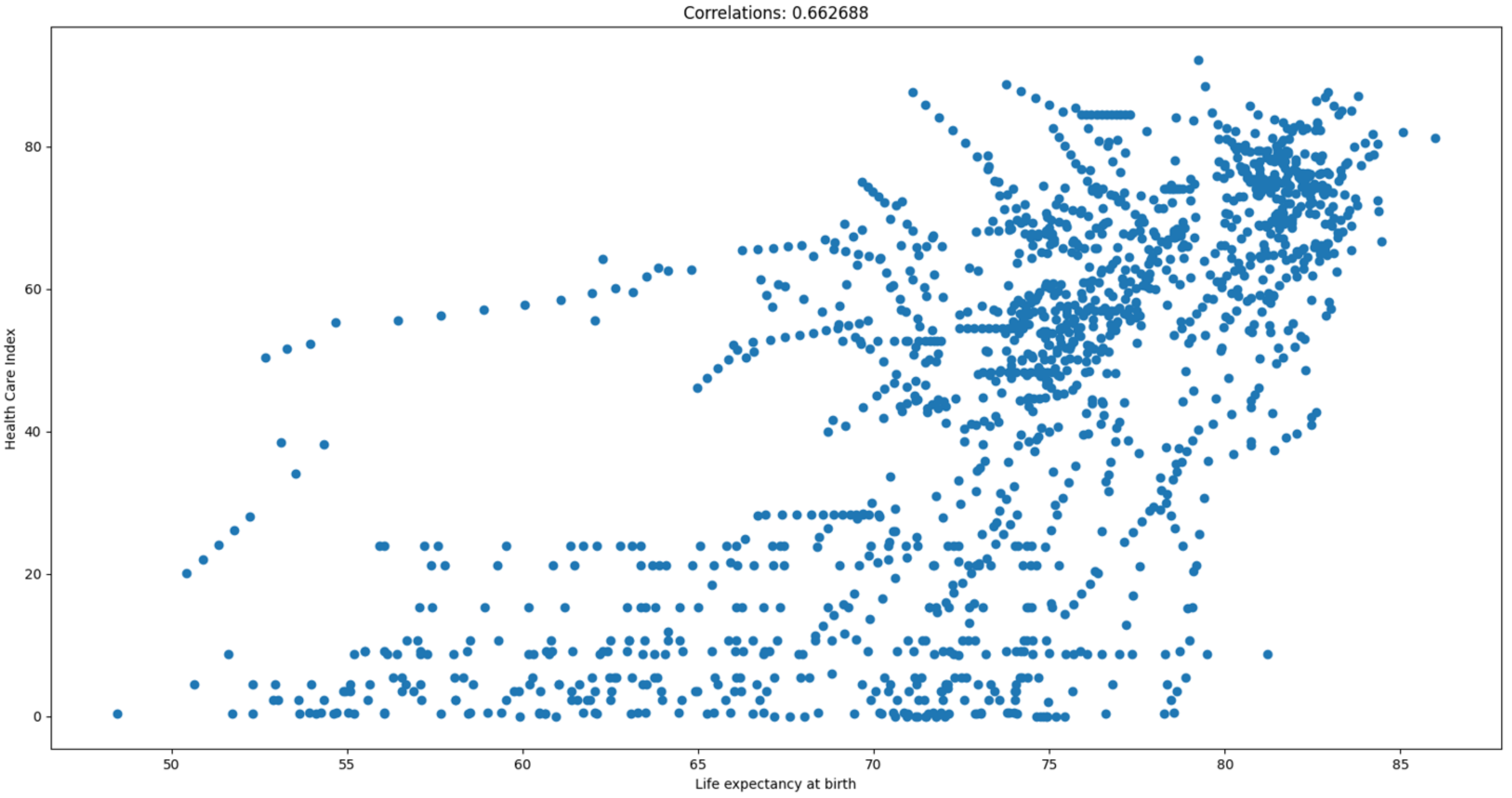
Heatmap df_full_database



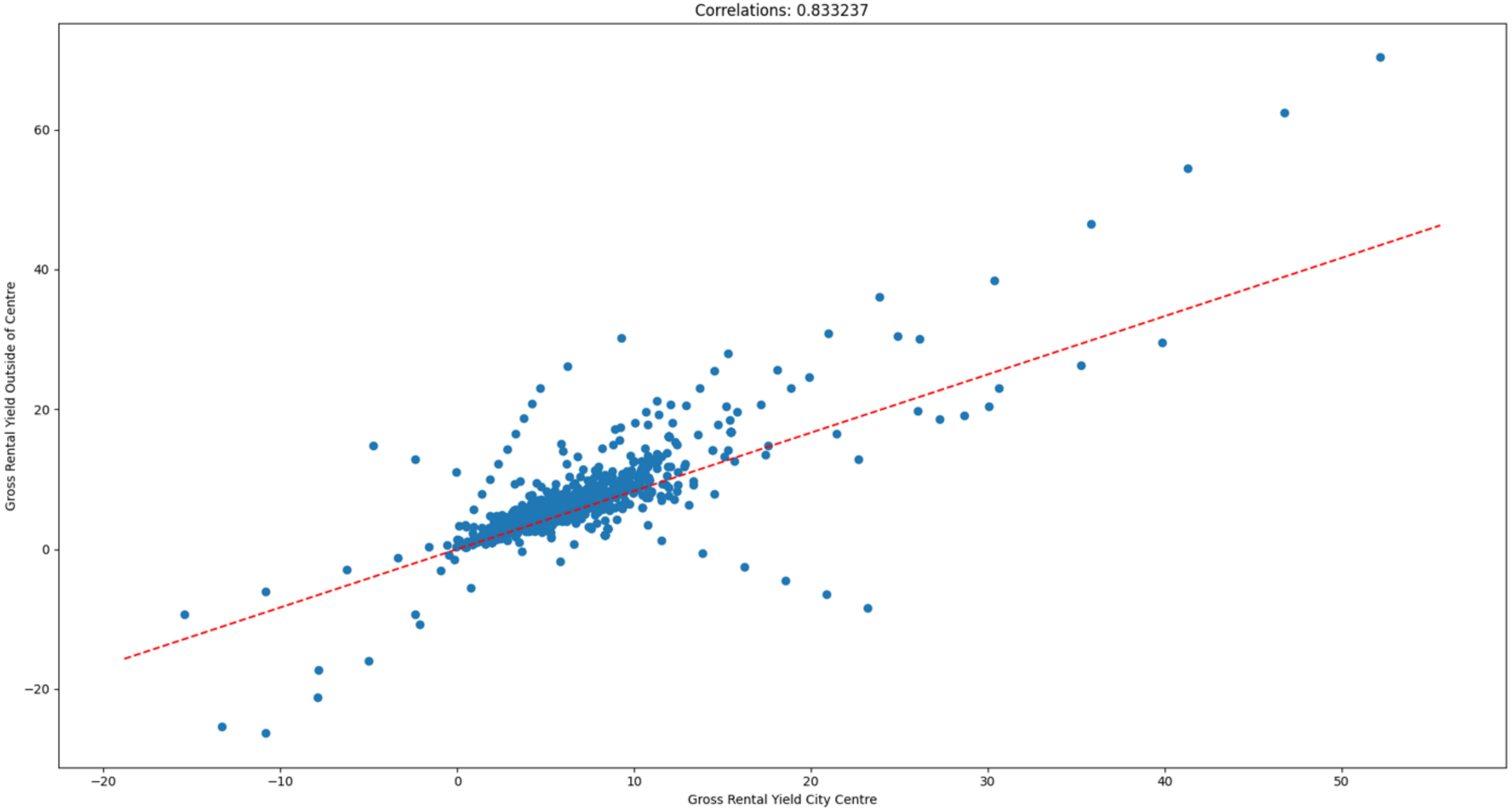
# Heatmap df_scrape



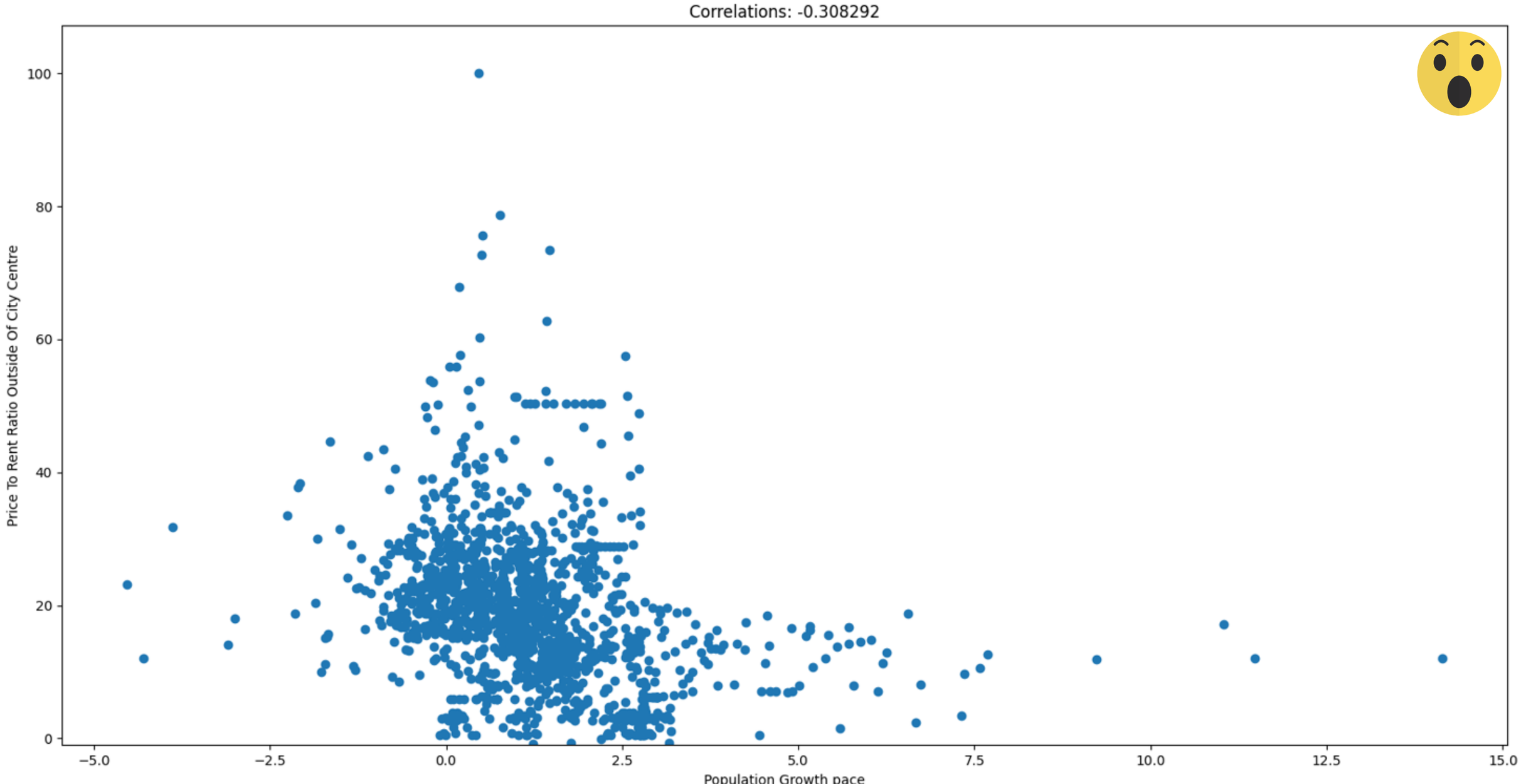
# Countries with higher Healthcare index will likely result in high life expectancy among its residents.



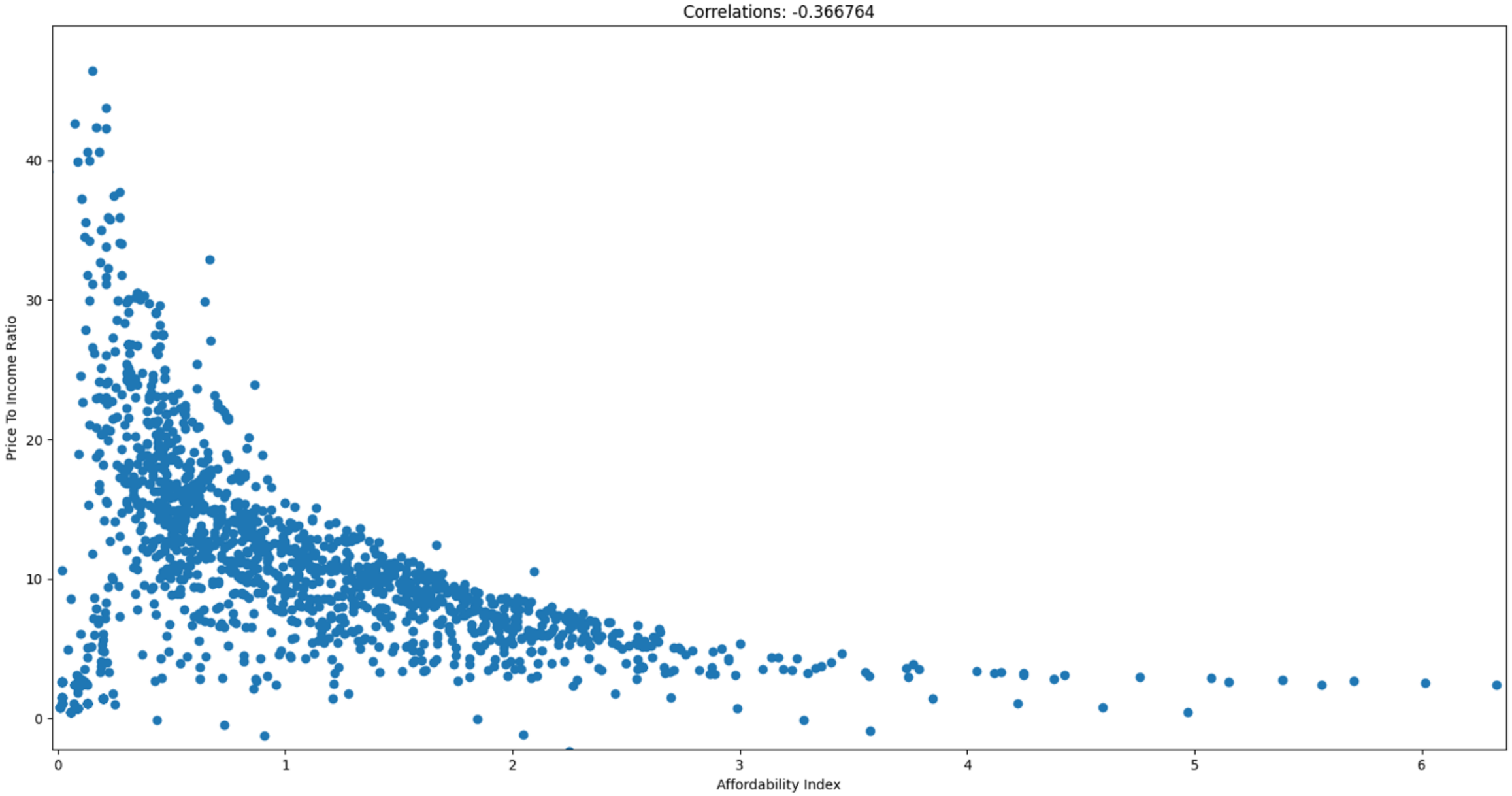
# Rental prices increase in the city center will likely result in price increases outside the center too



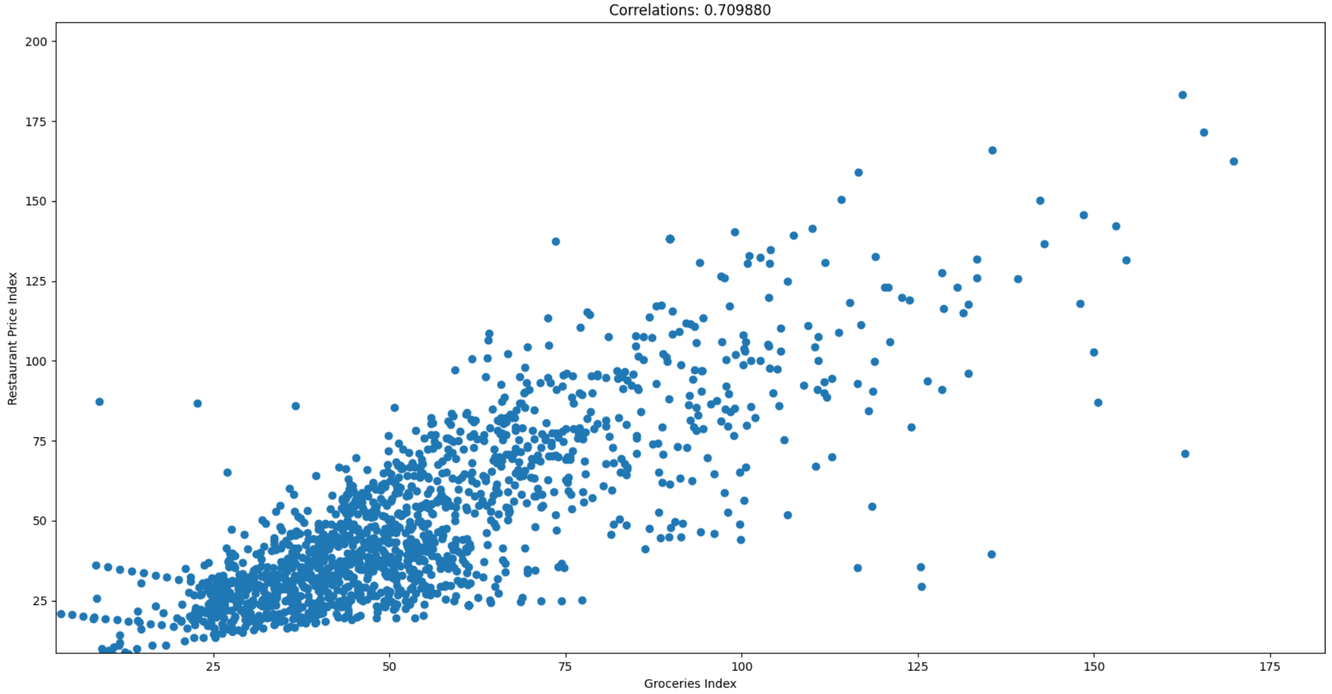
# The population tends to increase, if people buy more houses outside of the city center

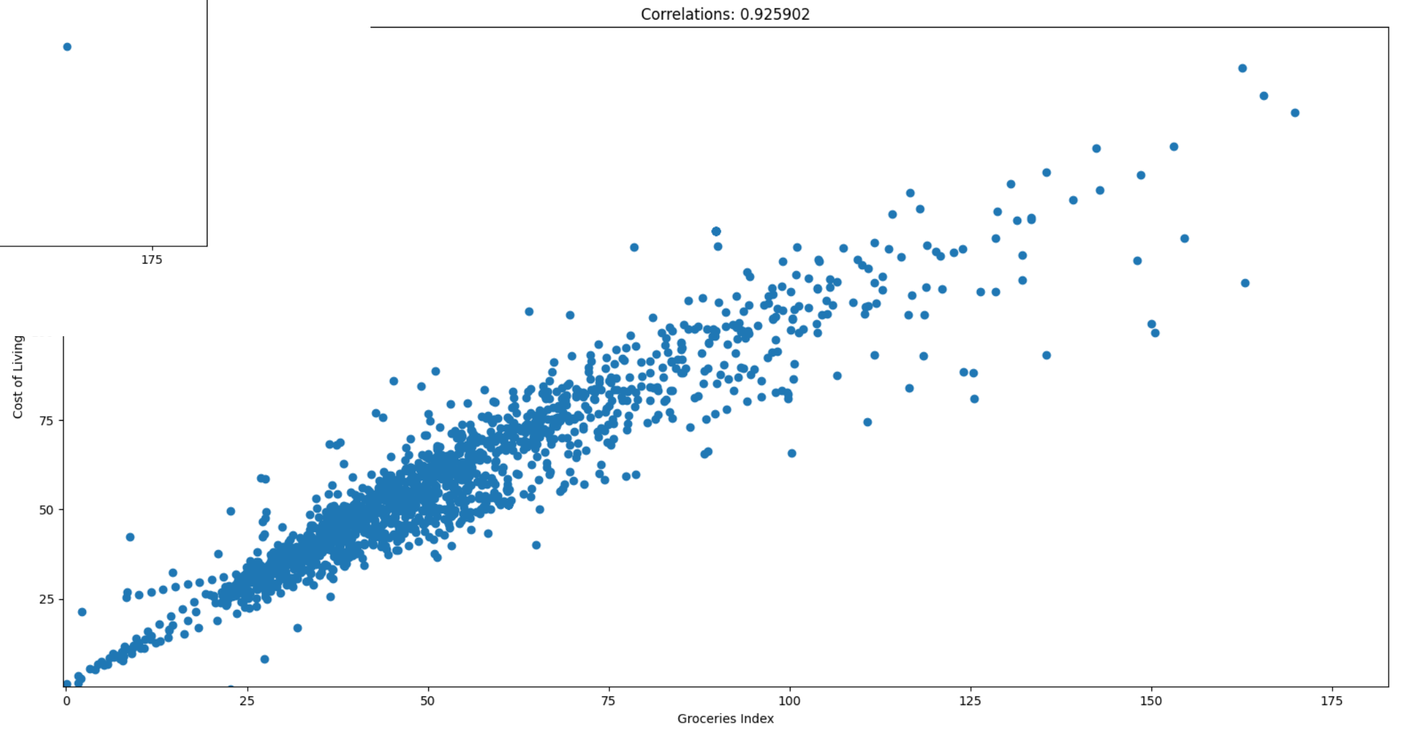


# If you can't afford buying a house, you are more likely to rent a house and vise versa

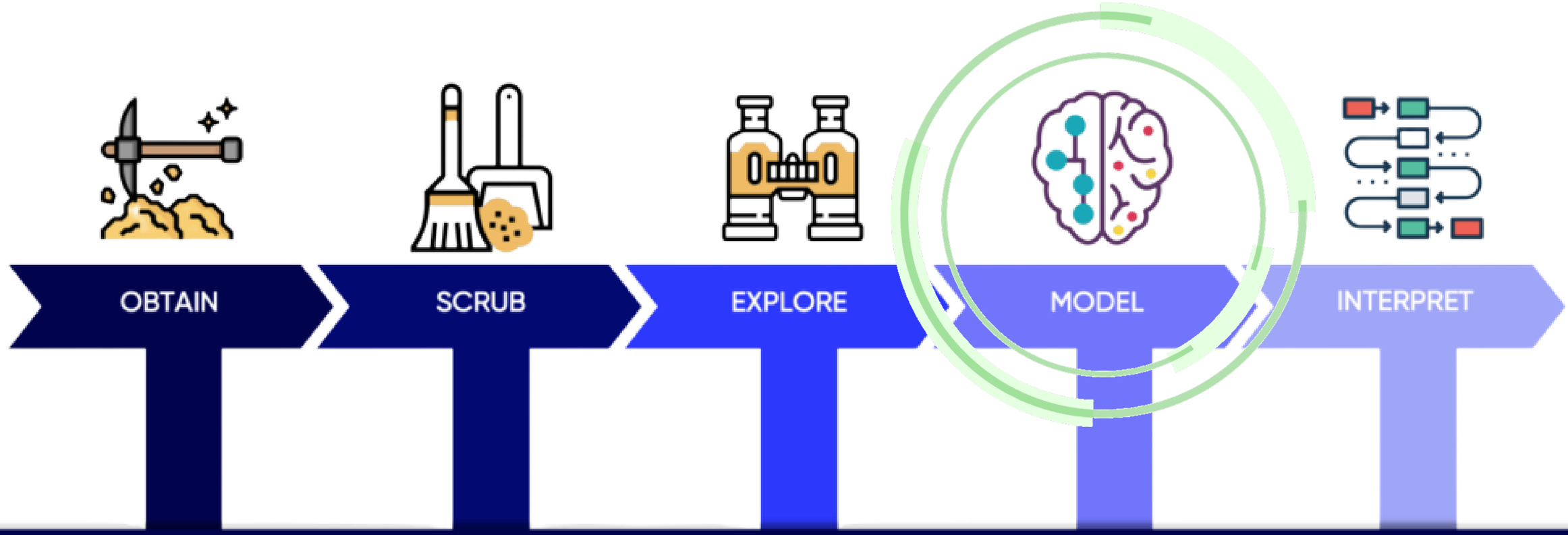


# Restaurants prices and cost of living index tend to increase, correlating with the Groceries prices



# Data Science Process




**O** — **Gather data from relevant** sources

**S** — **Clean data to formats** that machine understands

**E** — Find significant patterns and trends using statistical methods

**M** — Construct models to predict and forecast

**N** — Put the results into good use

# CHOOSING MODELS

- We chose to use **linear regression** in order to predict the GDP of countries in 2030.
- For **clustering**, we decided to use both **KMeans** and **DBSCAN** - to compare the results and choose the better model for us.

# TRAINING & USING OUR LINEAR MODEL

- We trained our model using the GDP data for each country from 1960 to 2020.
- Using the trained model we predicted 2030's data.

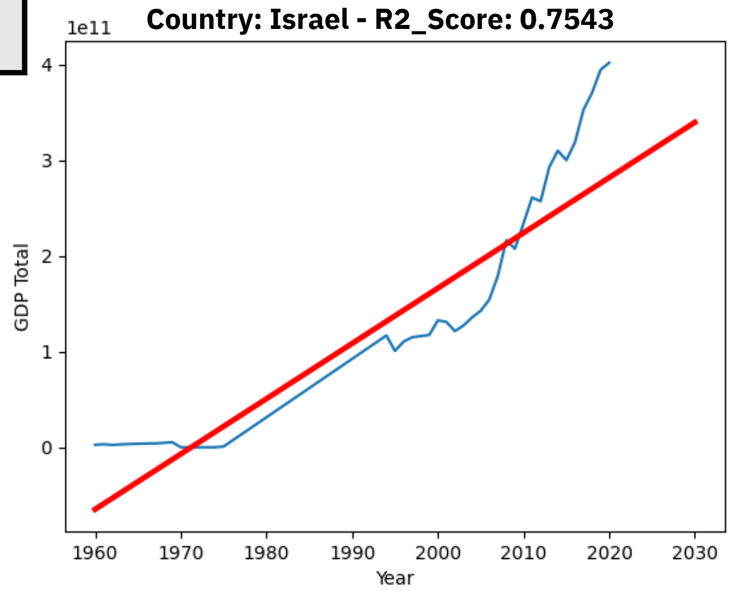
# THE LINEAR REGRESSION MODEL

## COMPARSION R2 SCORE

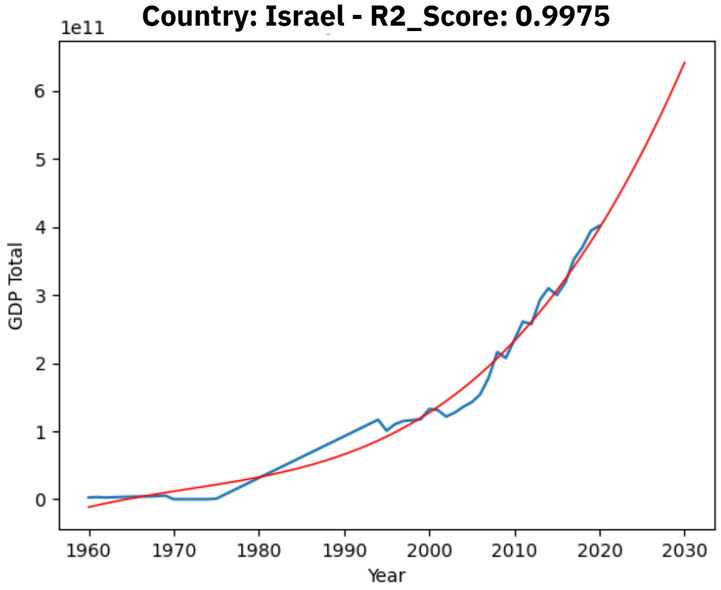
| Subject / Model | World Wide Avg | Israel |
|---|---|---|
| **Linear Regression** | 0.7278 | 0.7543 |
| **Pipeline** | 0.8363 | 0.9975 |

At first we've used linear regression, it was too Linear and not so accurate.
So we made a pipeline with Polynomial (degree 2-5) combined with linear regression for the prediction.
Here is the comparison:

## LINEAR_REGRESSIOON




## PIPELINE



# THE CLUSTERING MODEL

We had too many columns so we decided to downgrade the dimension of the data using **PCA**.
We've done that so we could get 2D graphs.

## FINDING HYPER-PARAMETERS AND MODEL SCORE

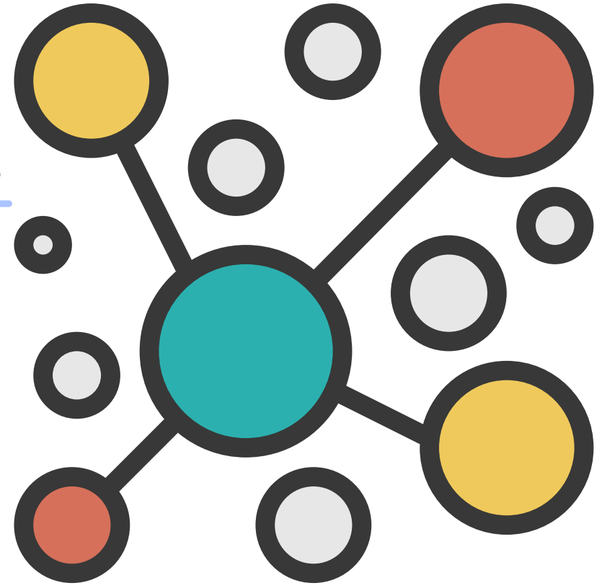
In KMeans and DBSCAN models, choosing the hyper-parameters is crucial for a successful model.

We've created functions for finding the best hyper-parameters for our models.

Afterwards we have determined the models performance using '**Silhouette score algorithm**'.

Overall, the results vary, some graphs had better score with **KMeans** and some with **DBSCAN**.

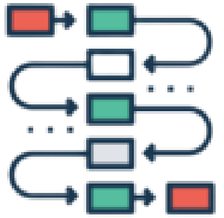
**KMEANS**



**DBSCAN**

Comparison with graphs in slides 30-33

# Data Science Process



OBTAIN | SCRUB | EXPLORE | MODEL | INTERPRET

**O**

**Gather data from relevant** sources

**S**

**Clean data to formats** that machine understands

**E**

Find significant patterns and trends using statistical methods

**M**

Construct models to predict and forecast

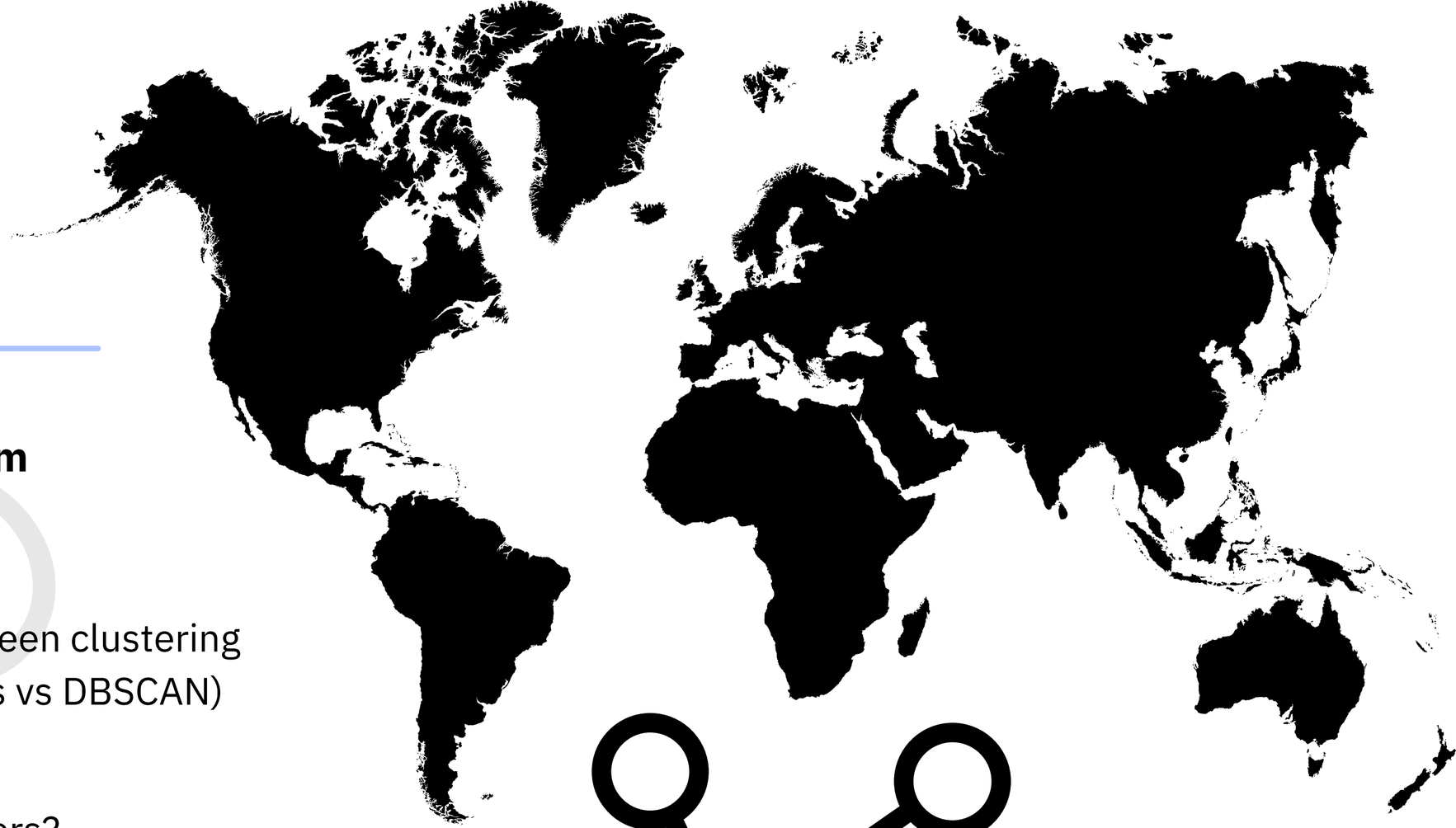
**N**

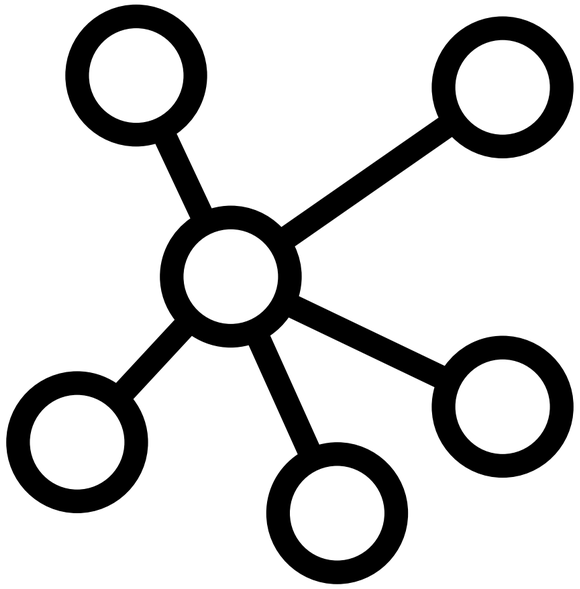
Put the results into good use

# QUESTION 1

**What can we learn from clustering countries?**

- Comparison between clustering methods (KMeans vs DBSCAN)
- Who are the outliers?





# HYPOTHESES

01

**We expected to get 3 clusters**:

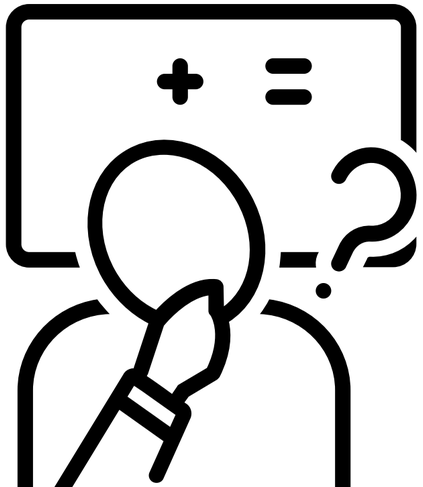
**A.** Top Leaders (USA, China, Russia, etc.)

**B.** Average Countries.

**C.** Below Average.

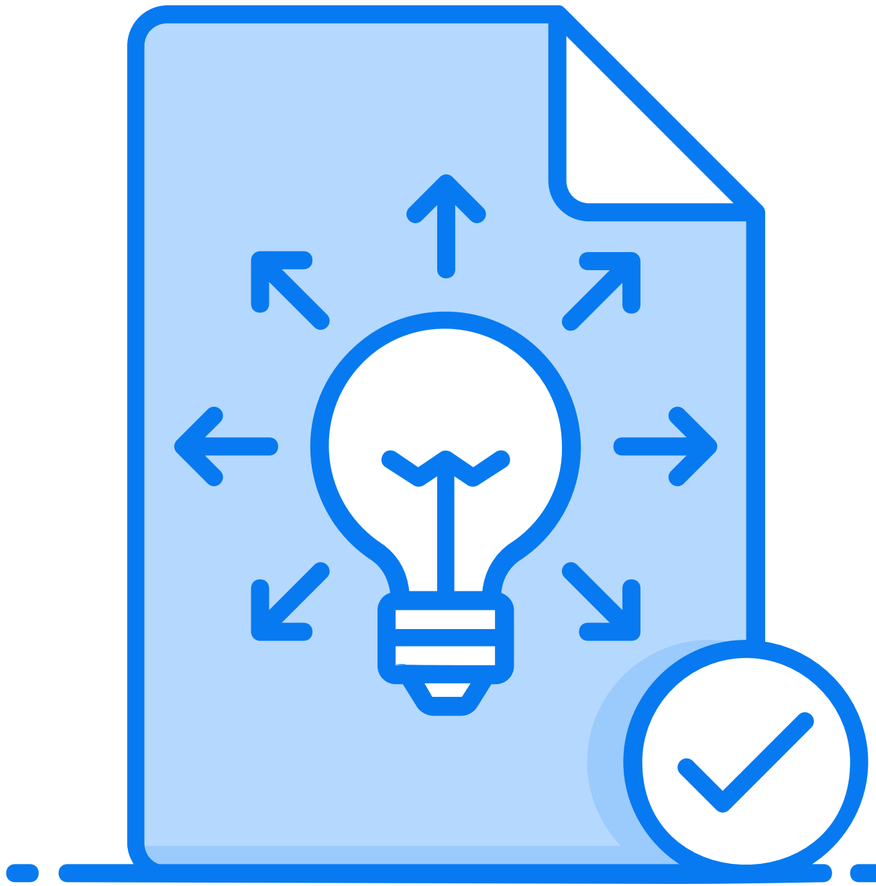
02

Bordering countries will most likely be in the same cluster.



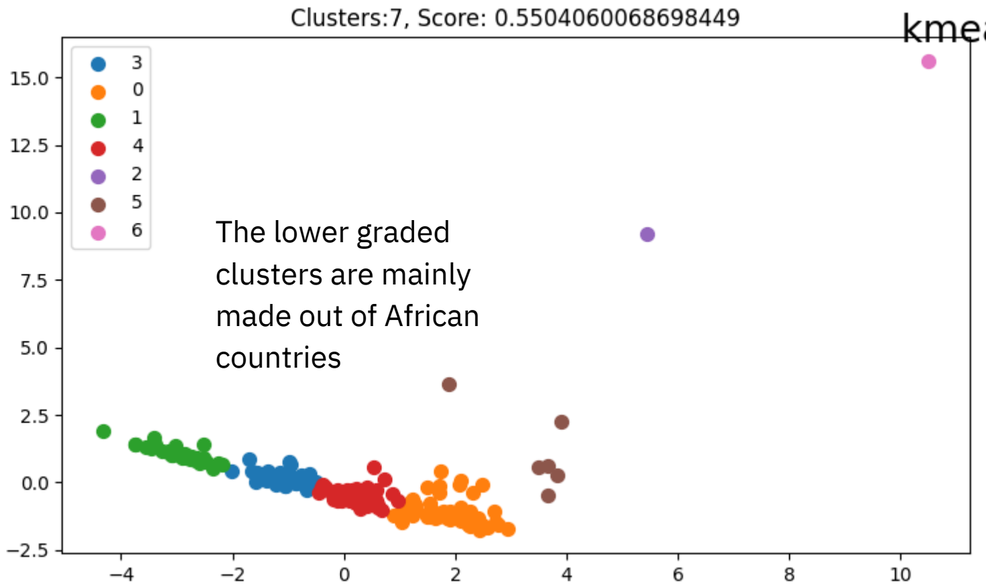
03

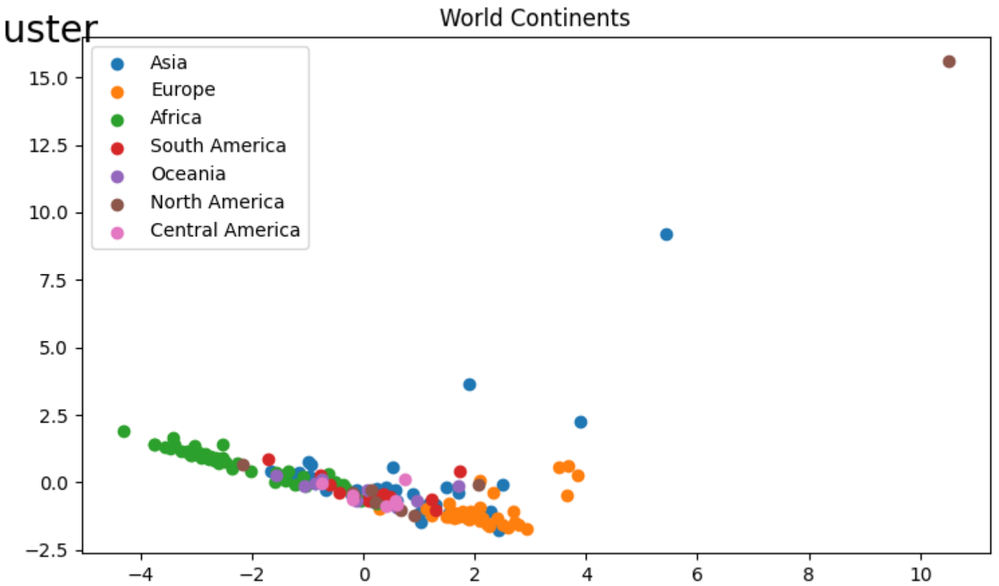
In **DBSCAN -** Top leaders and least developed countries will be considered as outliers.



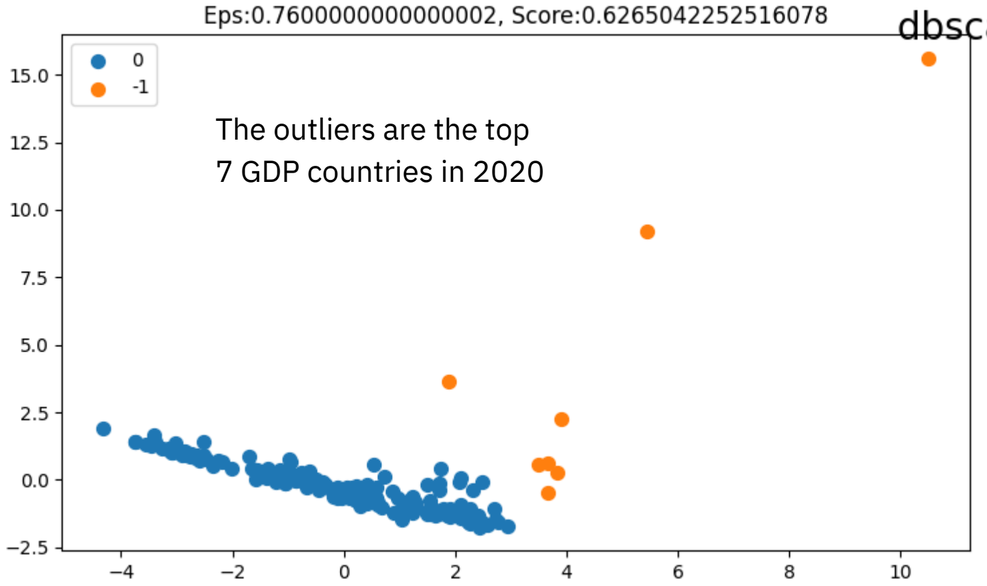
# FULL DATA CLUSTERING

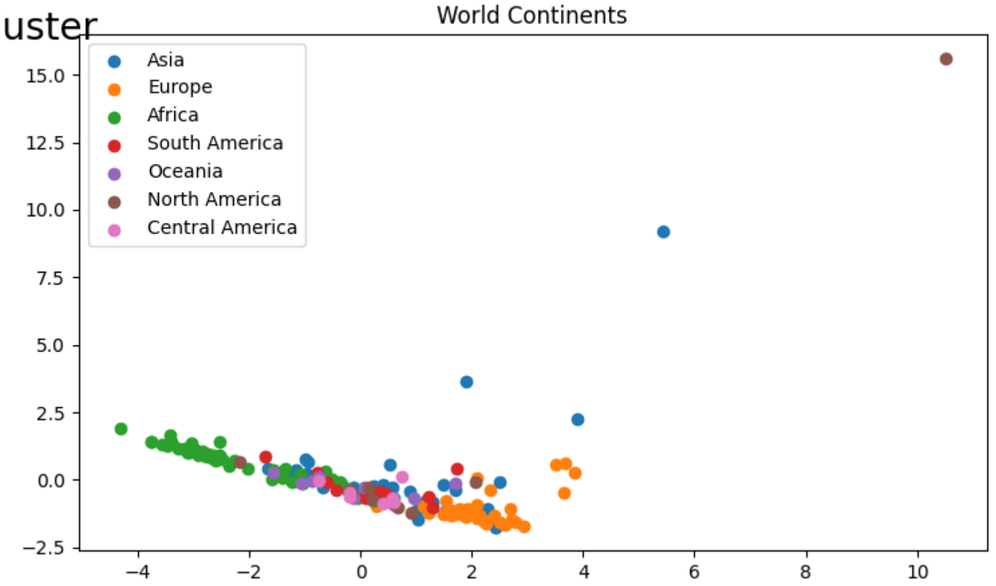
df
kmean-cluster



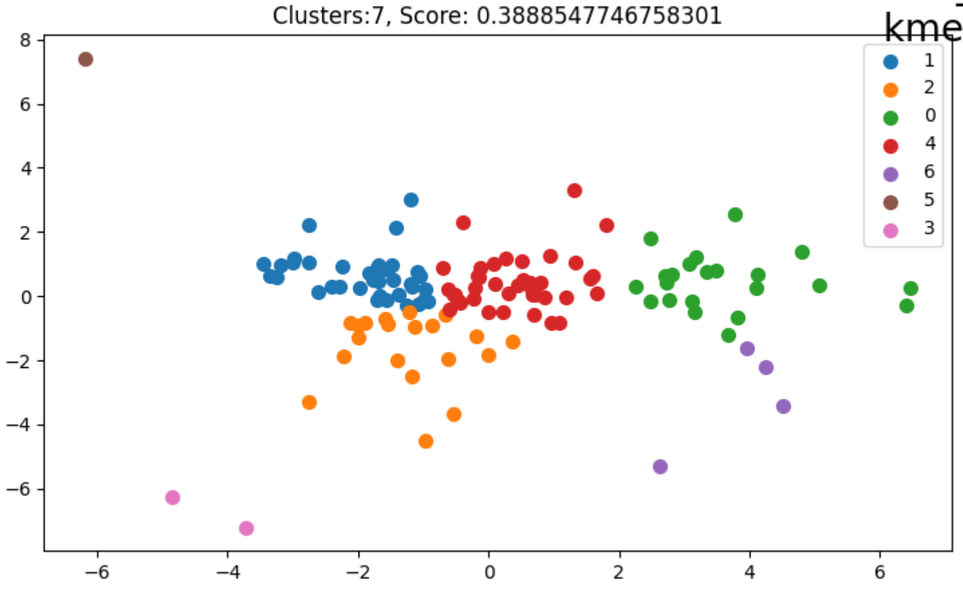

df
dbscan-cluster

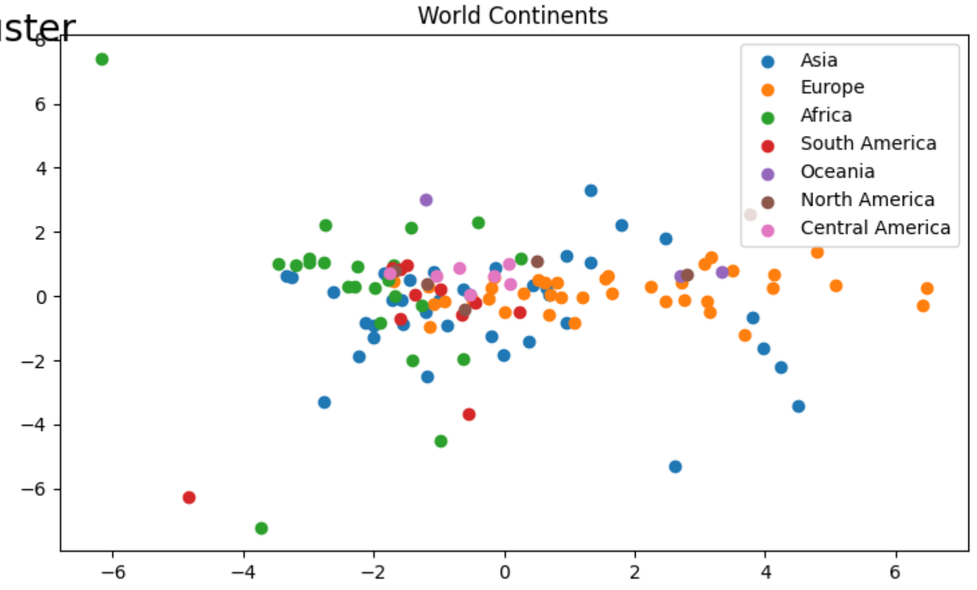


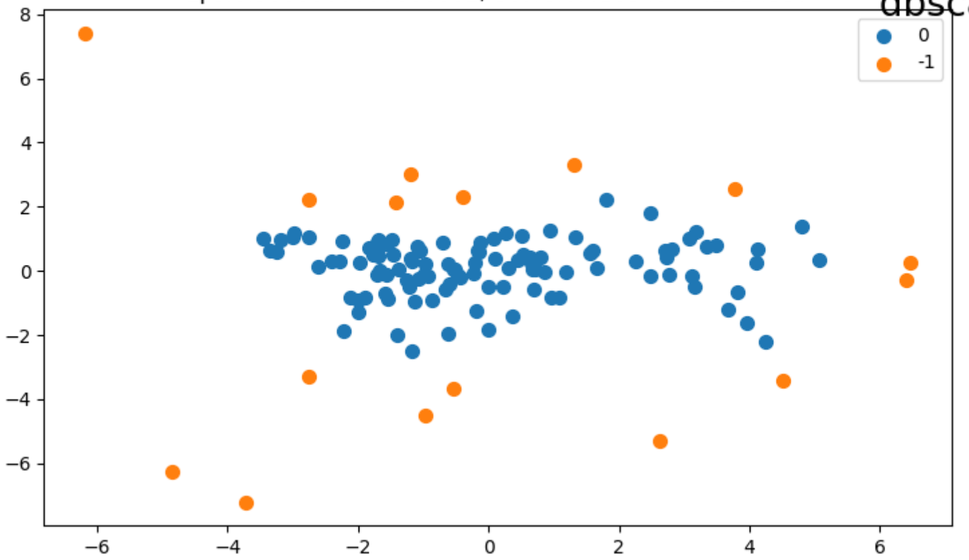
# SCRAPED DATA CLUSTERING

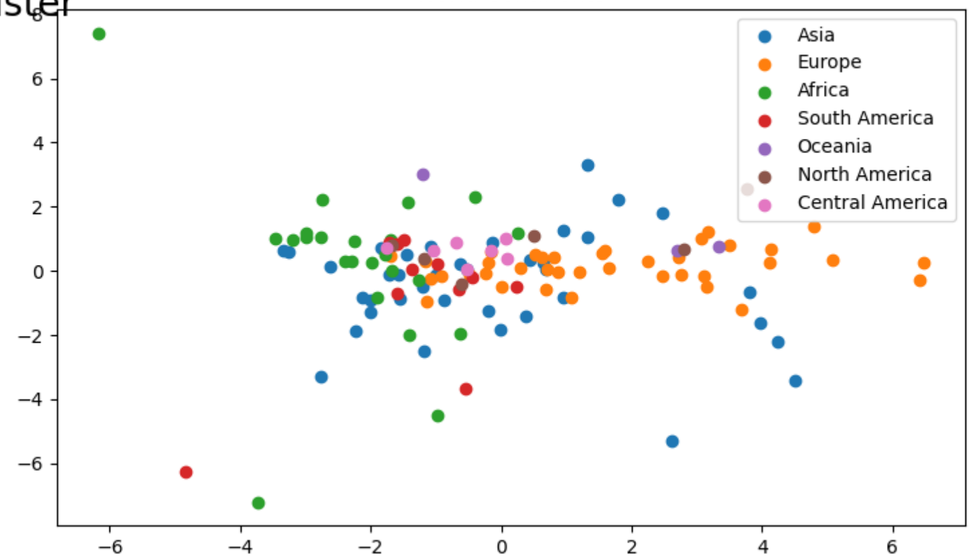
df_scrape
kmean-cluster



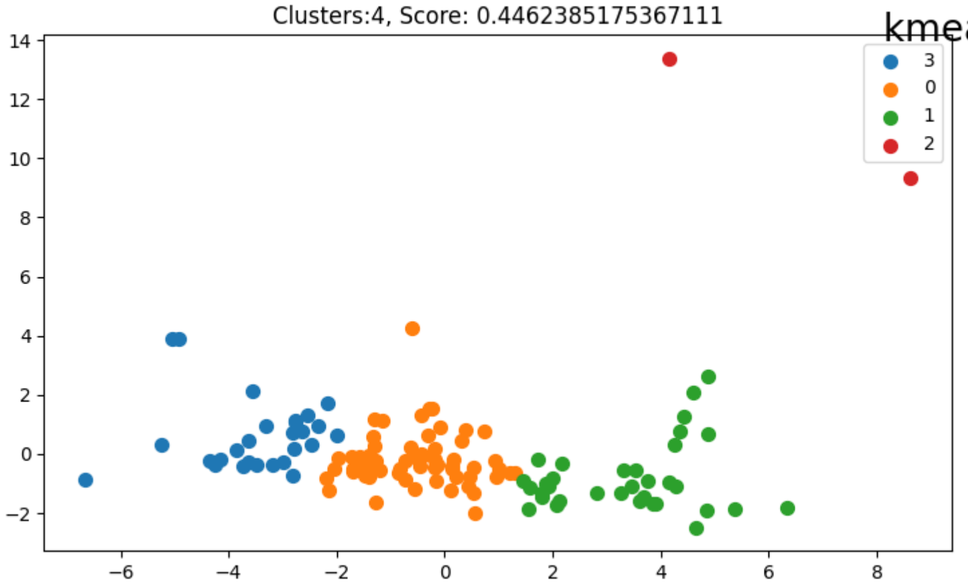

df_scrape
dbscan-cluster

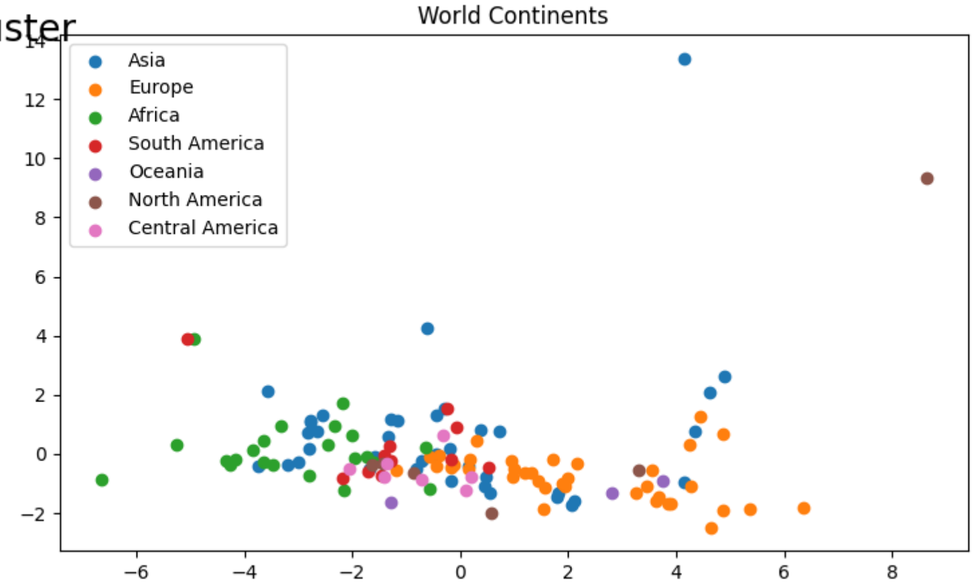


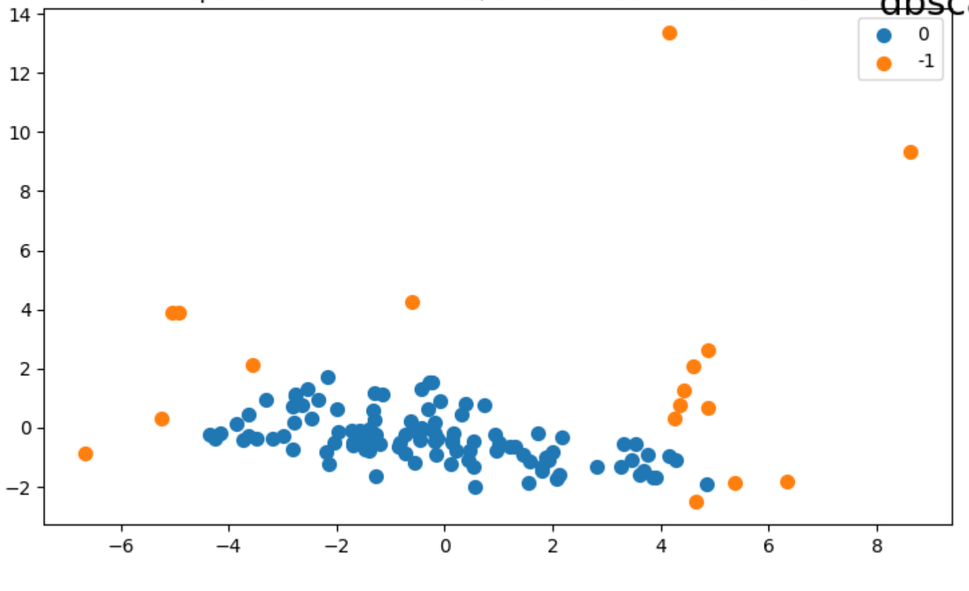
# COMBINATION

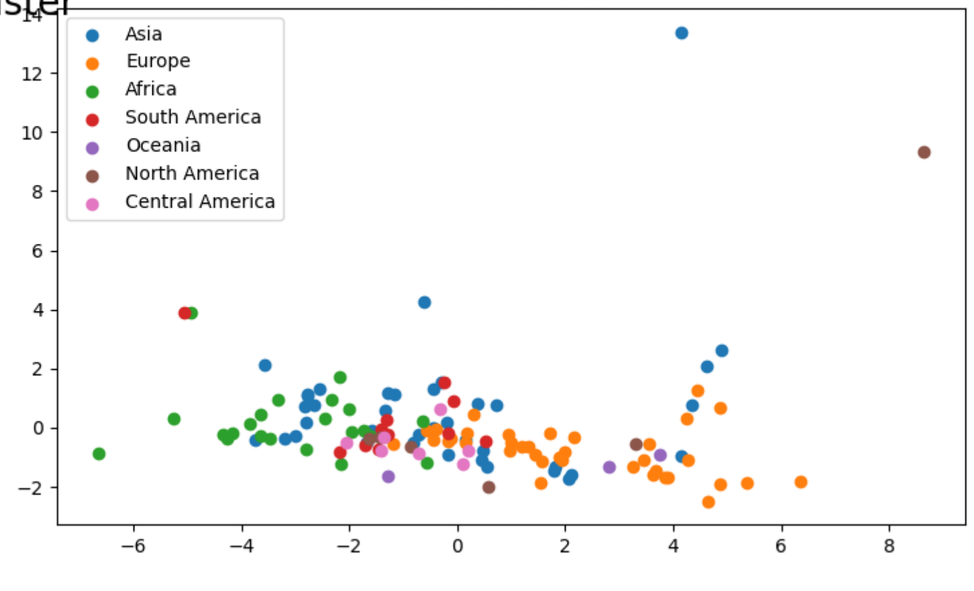
df_total
kmean-cluster




df_total
dbscan-cluster

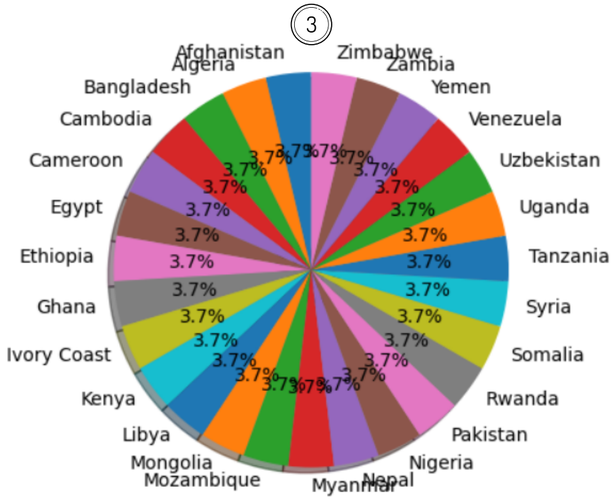


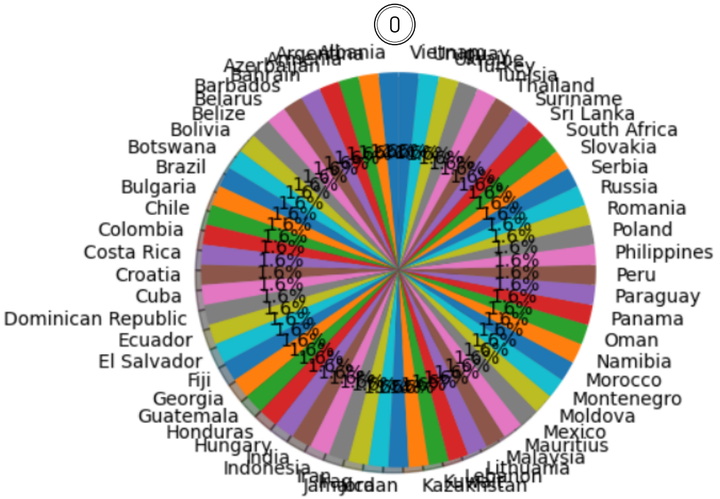
# KMEANS CLUSTERING OF ALL DATA COMBINED WITH THE COUNTRIES IN EACH CLUSTER

| 3 | 0 | 1 | 2 |
|---|---|---|---|
| Afghanistan | Albania, Lithuania | Australia | China |
| Algeria | Argentina, Malaysia | Austria | United States |
| Bangladesh | Armenia, Mauritius | Belgium | |
| Cambodia | Azerbaijan, Mexico | Canada | |
| Cameroon | Bahrain, Moldova | Cyprus | |
| Egypt | Barbados, Montenegro | Czech Republic | |
| Ethiopia | Belarus, Morocco | Denmark | |
| Ghana | Belize, Namibia | Estonia | |
| Ivory Coast | Bolivia, Oman | Finland | |
| Kenya | Botswana, Panama | France | |
| Libya | Brazil, Paraguay | Germany | |
| Mongolia | Bugaria, Peru | Greece | |
| Mozambique | Chile, Philippines | Iceland | |
| Myanmar | Colombia, Poland | Ireland | |
| Nepal | Costa Rica, Romania | Israel | |
| Nigeria | Croatia, Russia | Italy | |
| Pakistan | Cuba, Serbia | Japan | |
| Rwanda | Dominican Republic, Slovakia | Latvia | |
| Somalia | Ecuador, South Africa | Luxembourg | |
| Tanzania | El Salvador, Sri Lanka | Malta | |
| Uganda | Fiji, Suriname | Netherlands | |
| Uzbekistan | Georgia, Thailand | New Zealand | |
| Venezuela | Guatemala, Tunisia | Norway | |
| Yemen | Honduras, Turkey | Portugal | |
| Zambia | Hungary, Ukraine | Qatar | |
| Zimbabwe | India, Uruguay | Saudi Arabia | |
| | Indonesia, Vietnam | Singapore | |
| | Iran, | Slovenia | |
| | Iraq, | South Korea | |
| | Jamaica, | Sweden | |
| | Jordan, | Spain | |
| | Kazakhstan, | Switzerland | |
| | Kuwait, | United Arab Emirates | |
| | Lebanon, | United Kingdom | |

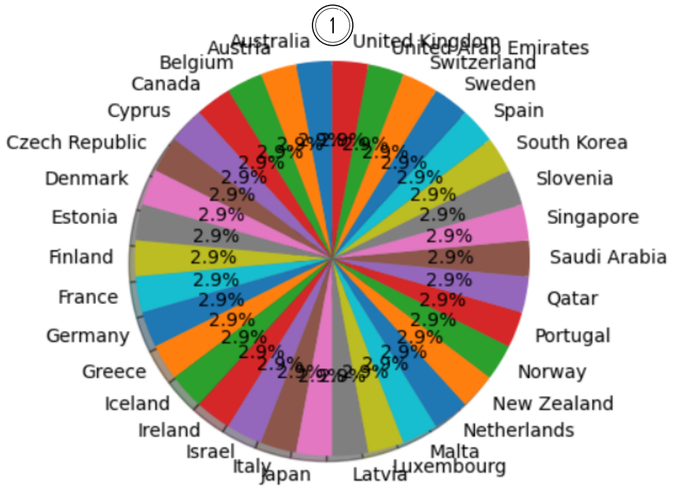
### Way below Average Countries



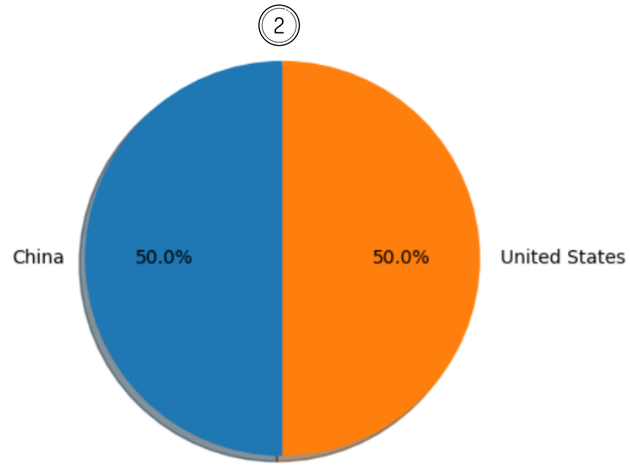

### Average Countries




### Above Average Countries




### World Biggest Leaders



# CONCLUSIONS

01

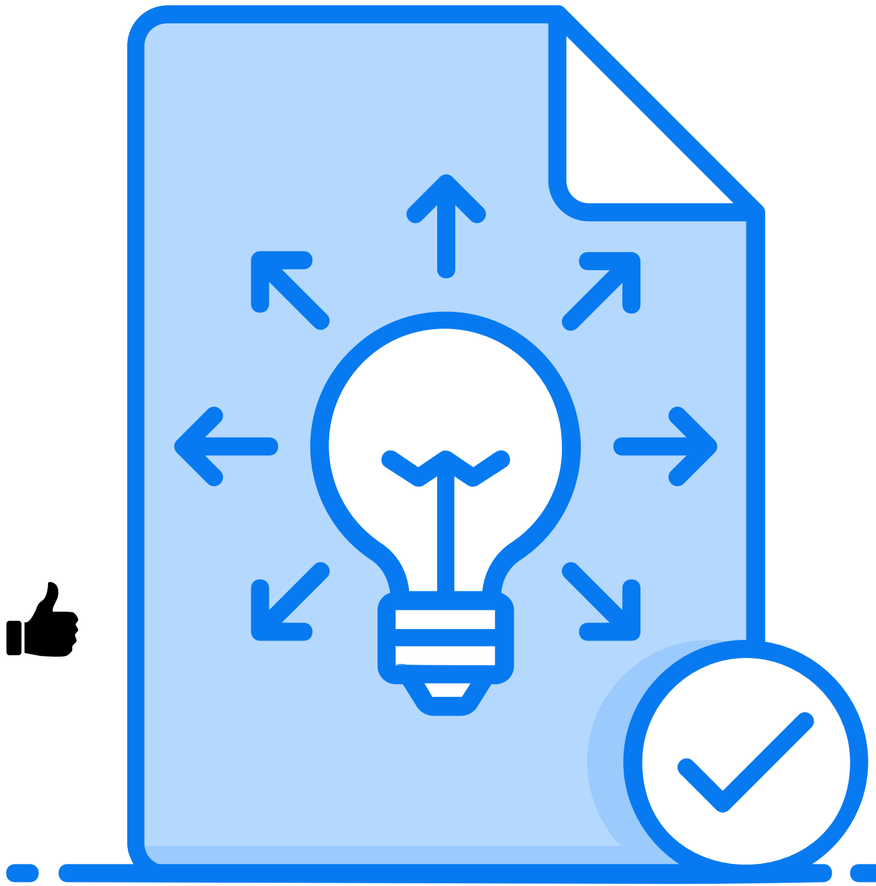
After running the KMeans cluster algorithm we got more than 3 clusters as seen in the previous slide.

02

Bordering countries are likely to be in the same cluster.
Most African countries are in cluster 3.
Most European countries are in cluster 1.
Most American countries are in cluster 0.

03

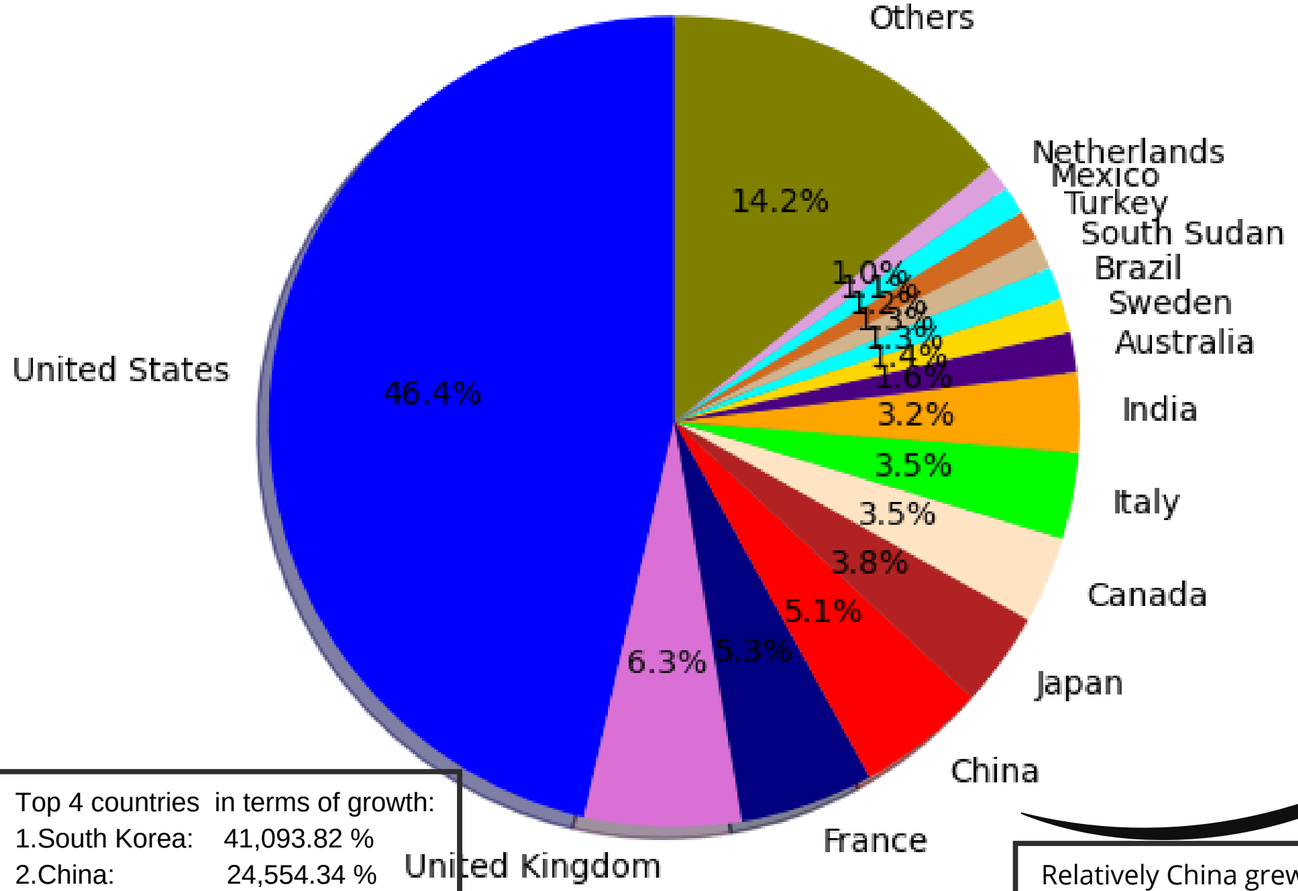
After running the DBscan we were amazed to see that he clustered the outliers which some world leaders and some least developed countries.

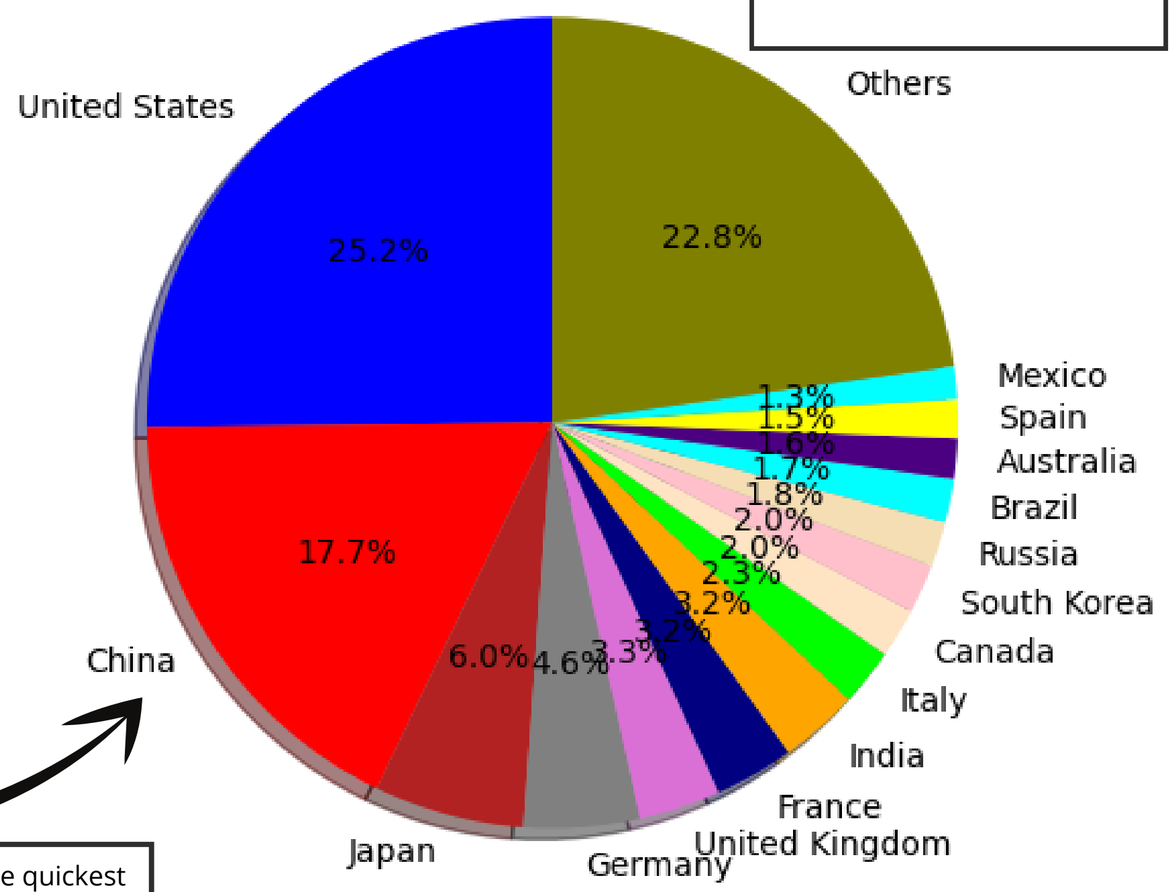


# QUESTION 2

**How would the World GDP look like in 2030?**

- Who will be the world leading countries?
- Will the current world leading countries keep their place in the future?
- Comparison between 1960 & 2020.



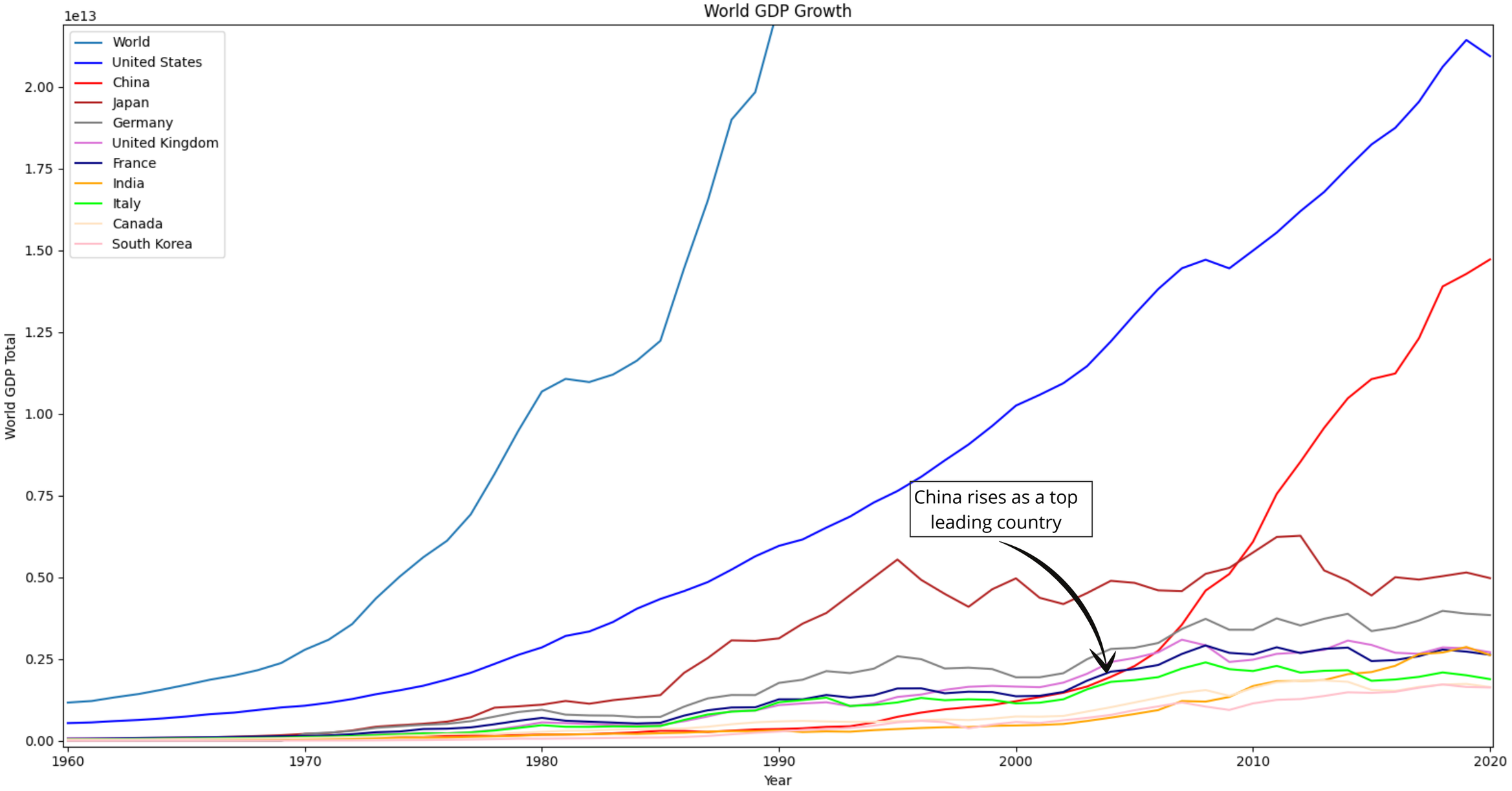




All other countries are getting stronger
Africa's GDP is still the lowest in terms of growth and total.

Top 4 countries in terms of growth:
1.South Korea: 41,093.82 %
2.China: 24,554.34 %
3.Japan: 11,129.35 %
4.United States: 3,753.60 %

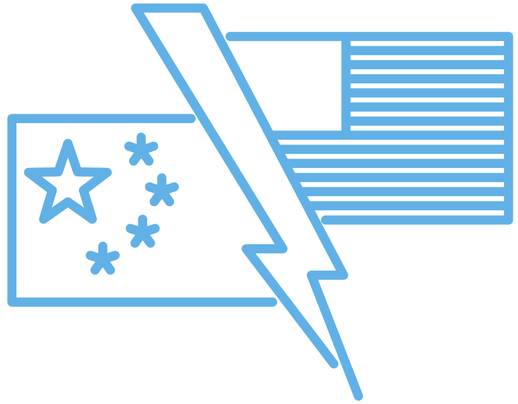
Relatively China grew the quickest by 340%

# Slide 14 Zoomed in

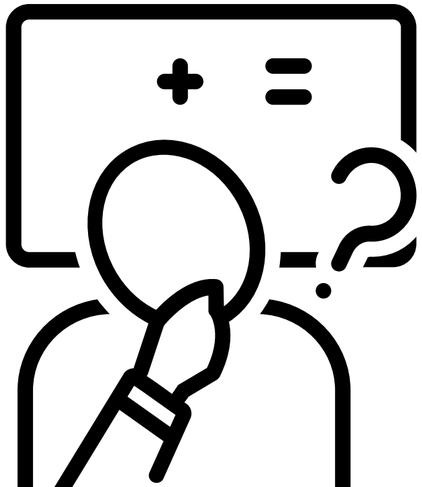


# HYPOTHESES

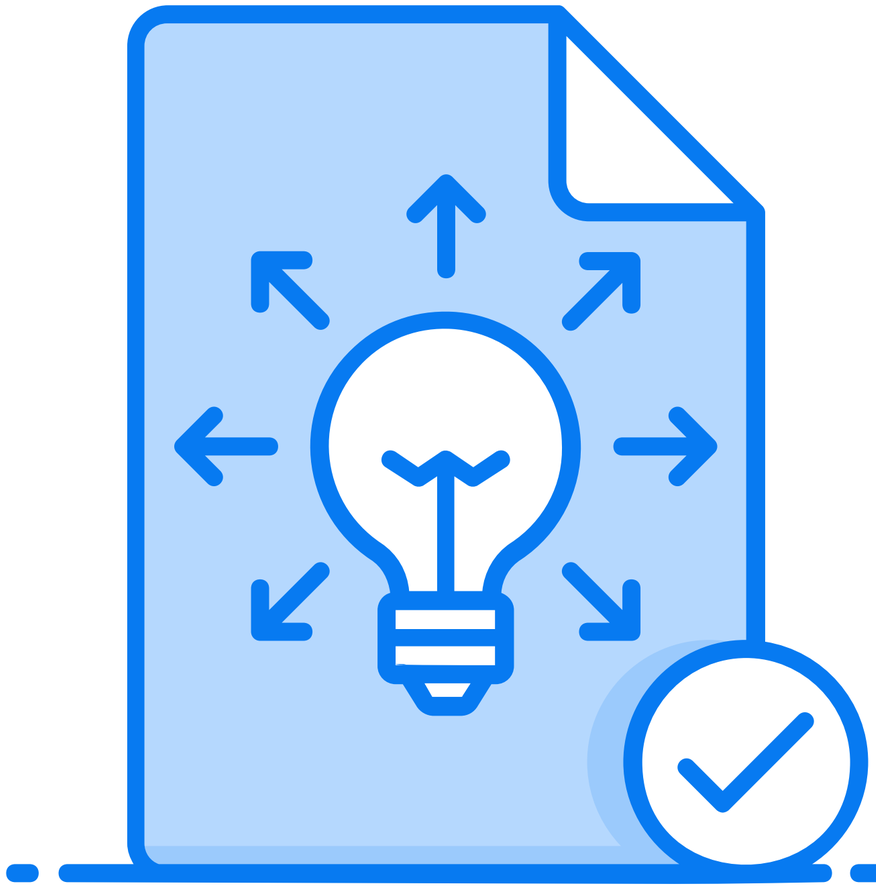
01 China will outgrow USA in the near future.



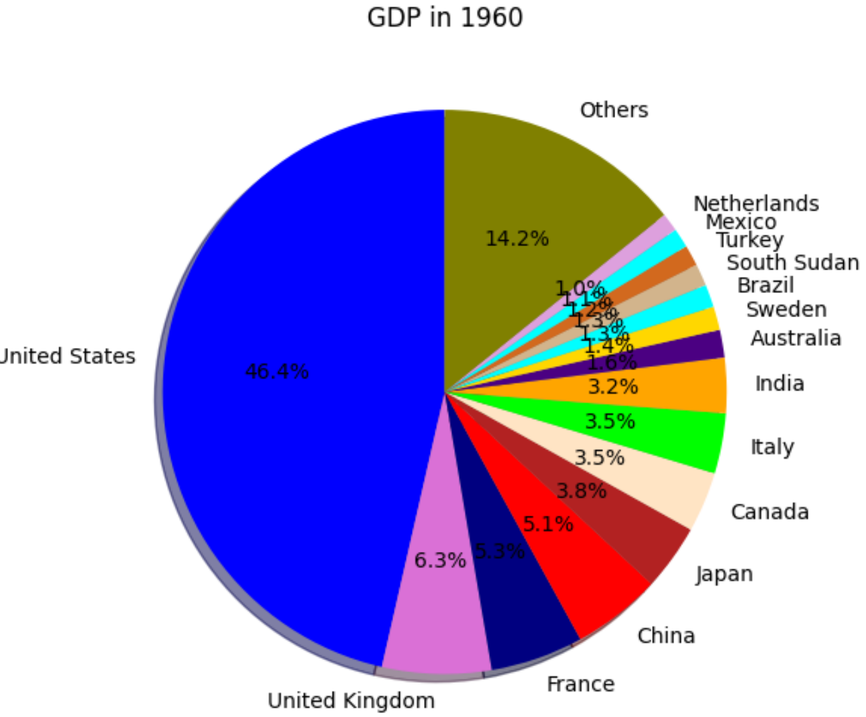
02 "Others" countries GDP will grow relatively to total GDP.

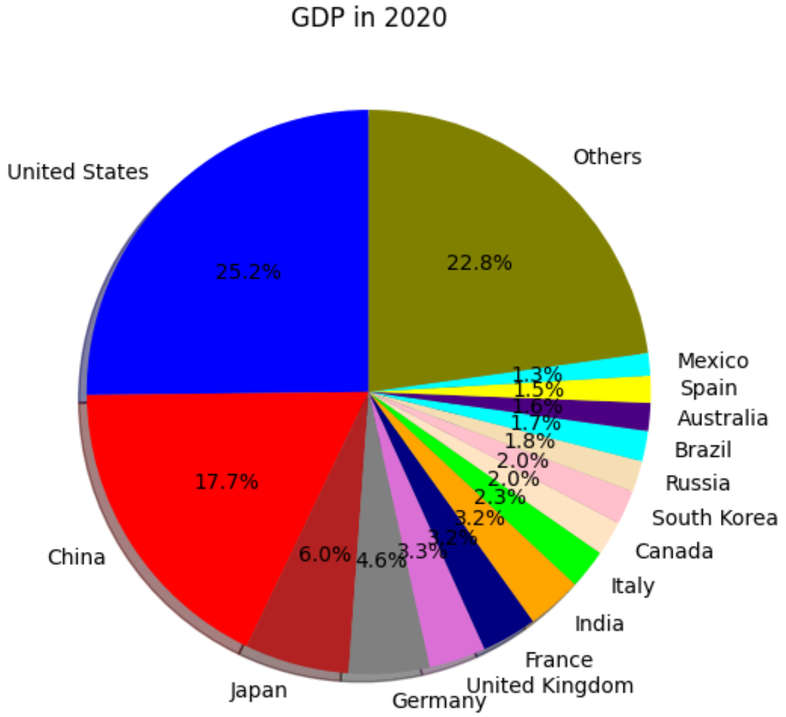


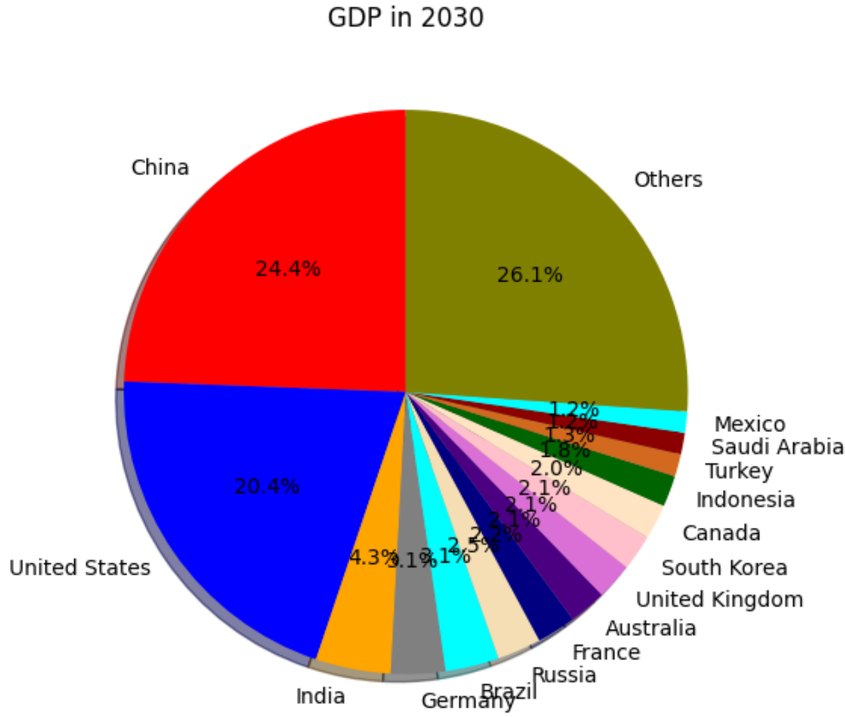
03 New countries will rise to the top.



# Comparision between years

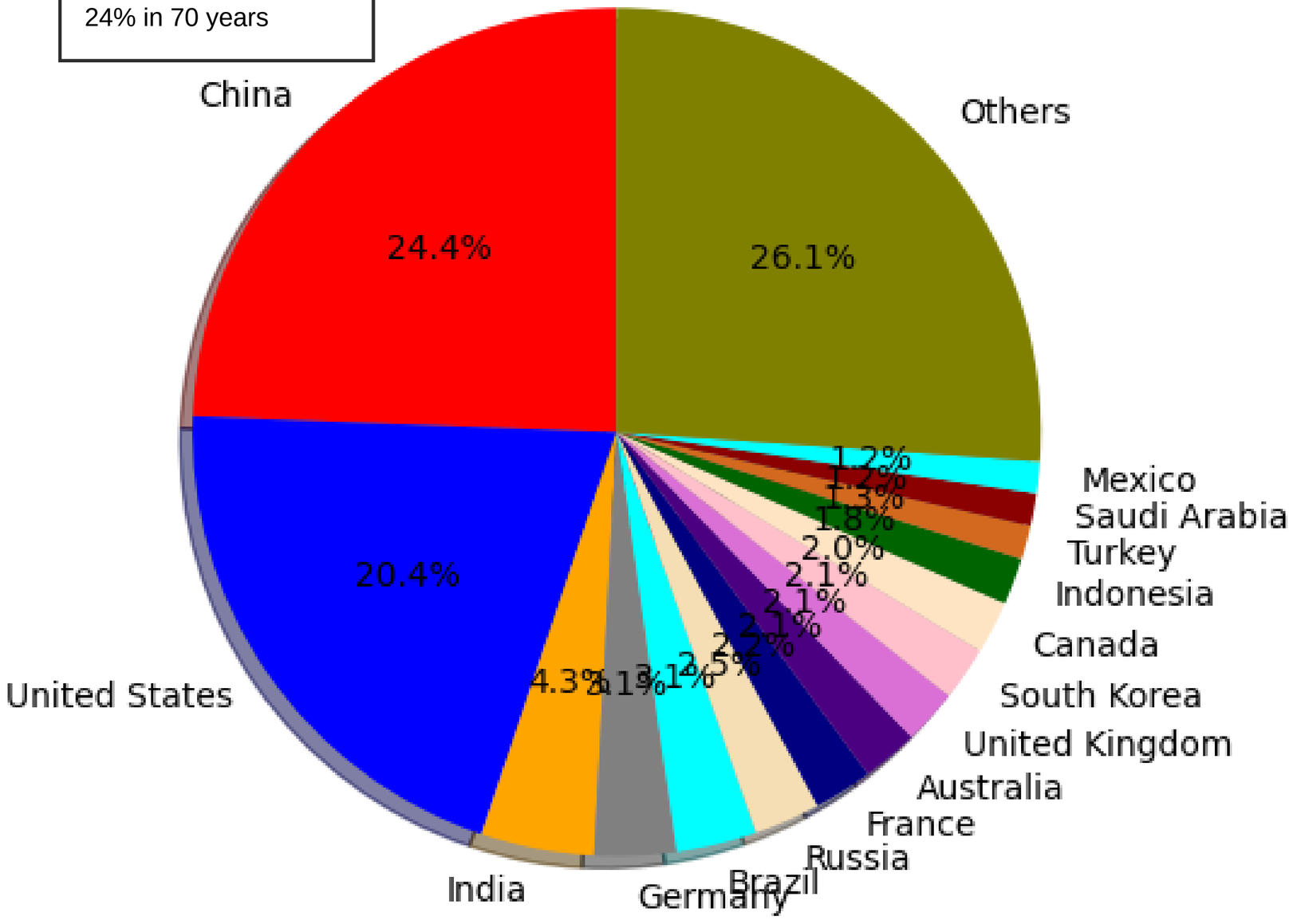


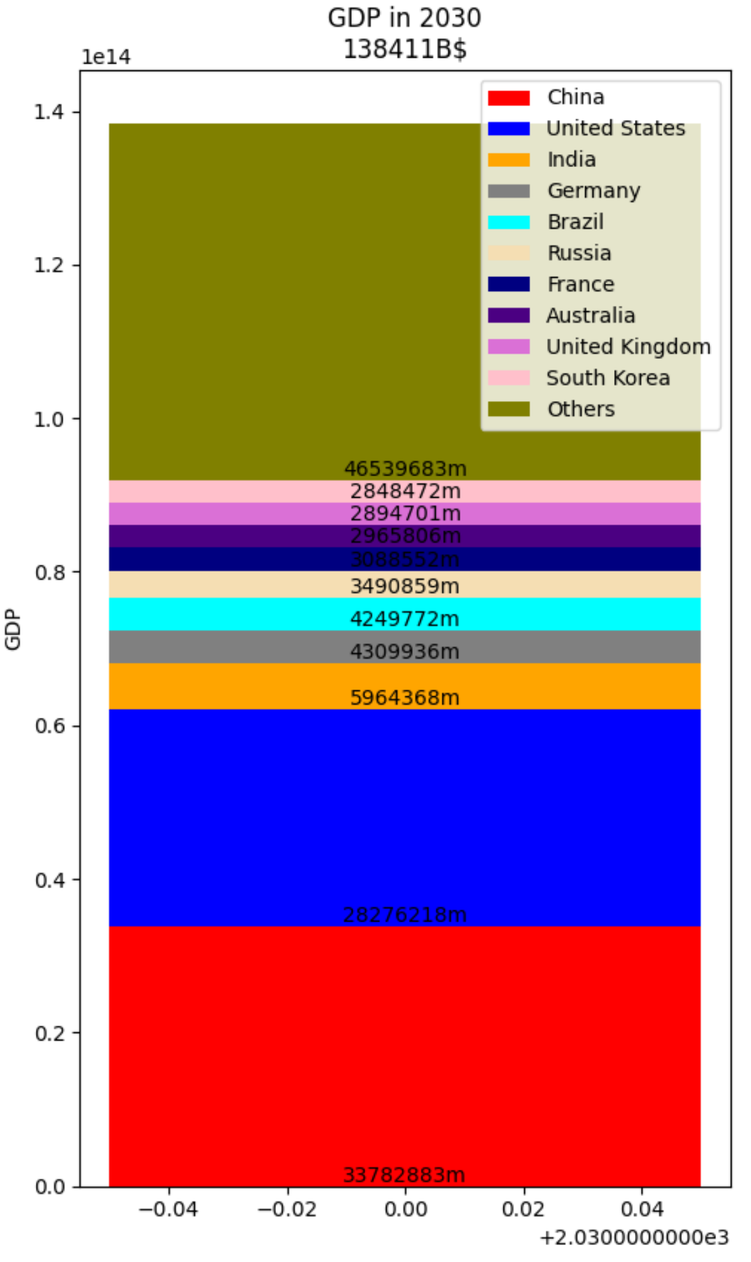



# GDP in 2030

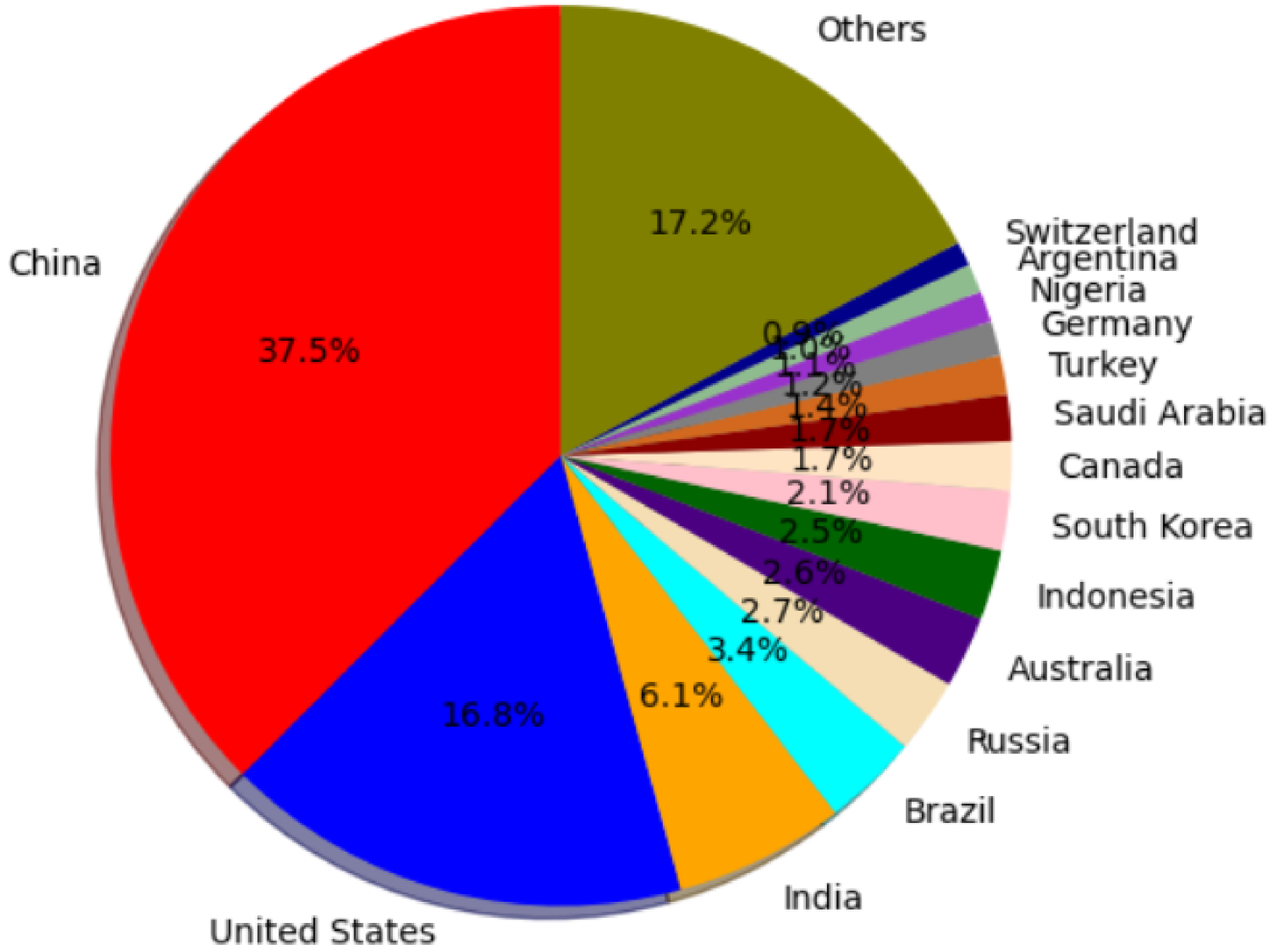
China grew from 5% to 24% in 70 years



# BONUS PREDICTION 2050

- China outgrows USA by far.
- India, Brazil and Russia are in the top 5.
- Nigeria and Indonesia are in the top 15.



# CONCLUSIONS

01

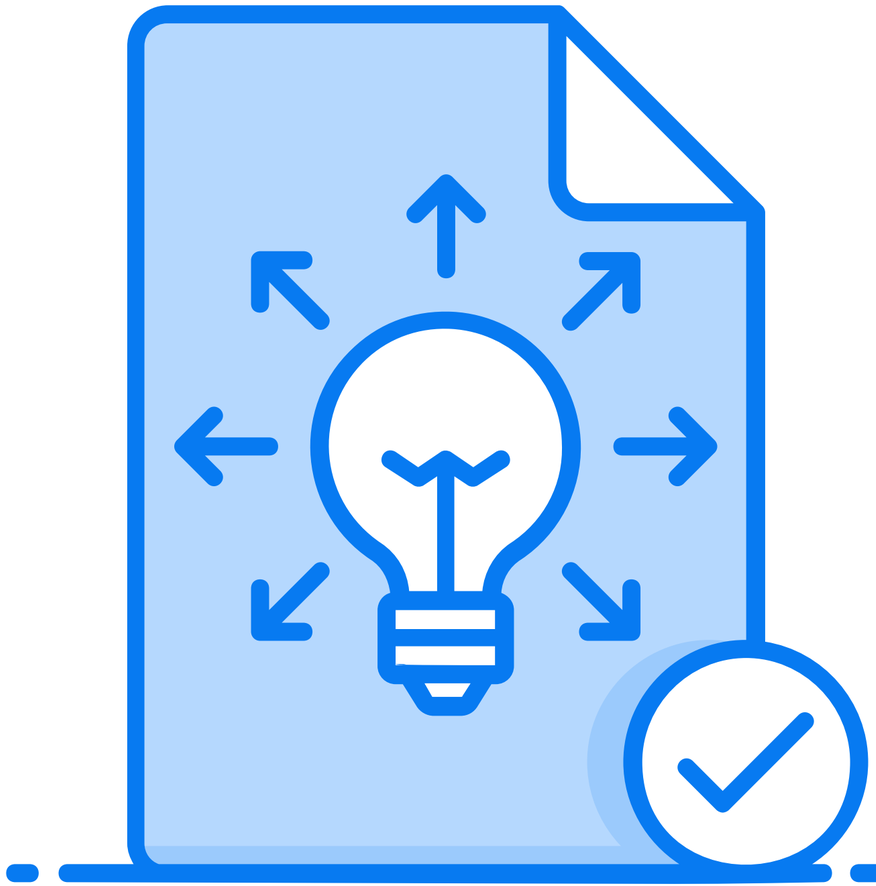
Based on the data, we predict that China will outgrow USA in the near future, to be precise - by year 2025.
China will be leading USA by 5,506,665m

02

Countries in the "Others" section GDP will grow relatively to the total GDP.
In 2020 they we're 22.8%
In 2030 they grew to 26.1%

03

New countries emerged to the top.
e.g Saudi Arabia, Indonesia, Turkey.



# QUESTION 3

**What can we learn from the data and graphs about history?**

## THE DATA EXPLAINS OUR HISTORY

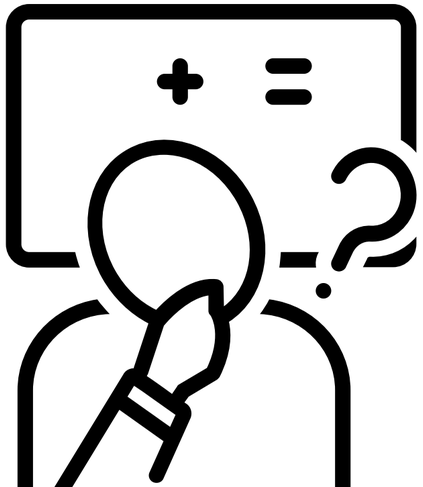
> Throughout history, humans killed each other while trying to gain control and power.

This fact, is represented in our data and reveals "dark history" for some.

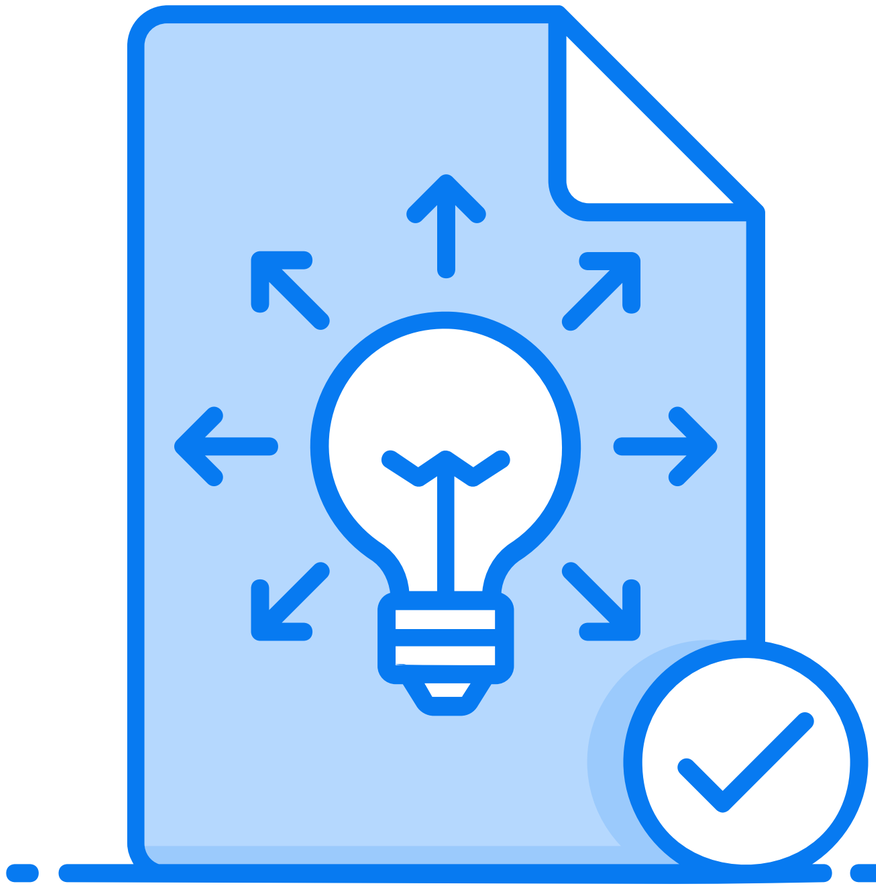
# HYPOTHESES

01 Significant events will stand out in the graphs.

02 Least developed/Smaller countries are more sensitive to historical events.



03 Revolution / Collapsion can impact the country's  GDP.



# GENOCIDE, WARS & USSR LISTS

## Genocide & Wars

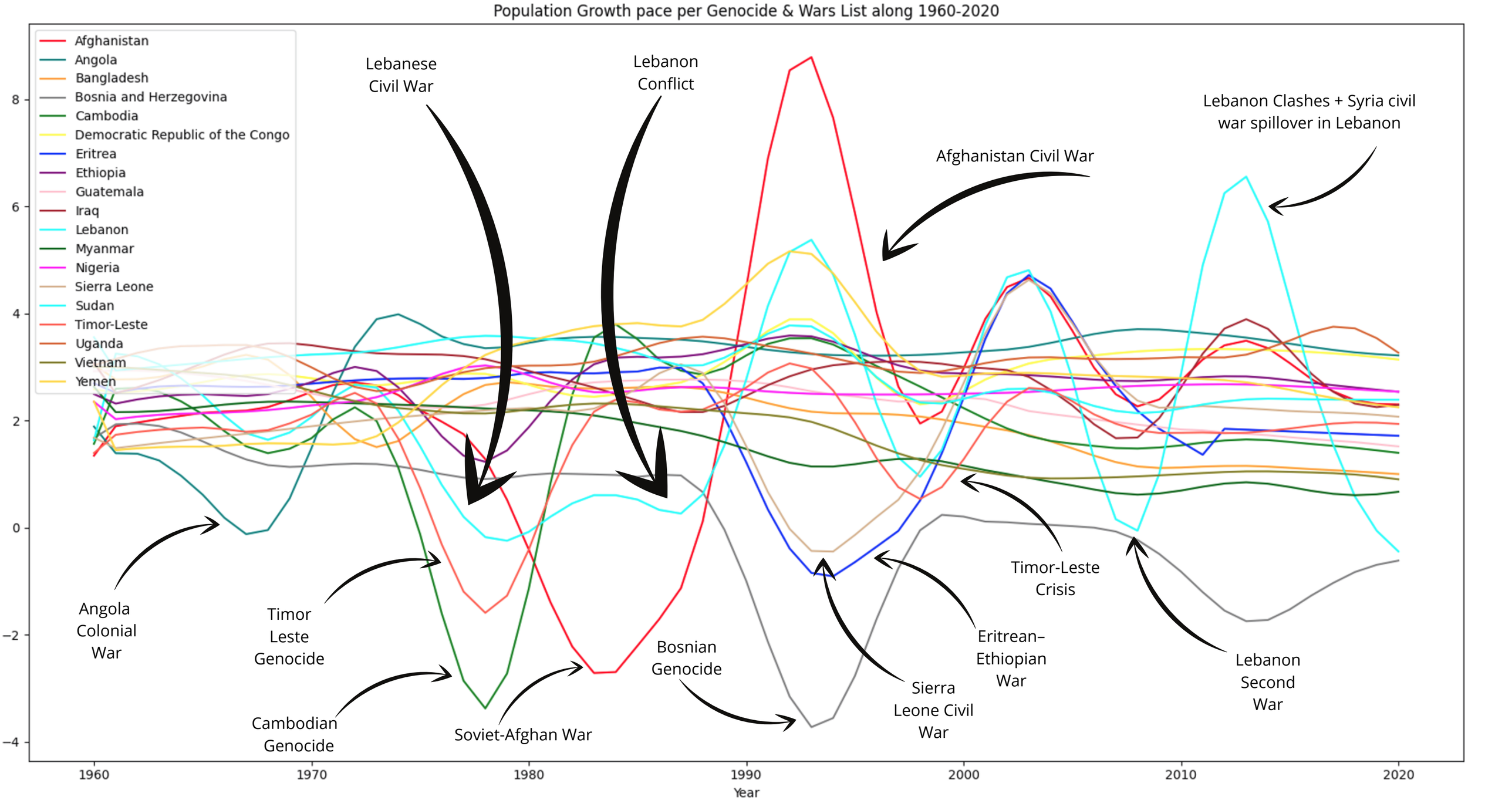
Based on the outliers and inconsistent data in the graphs we've concluded those countries have been through war/s, Civil war/s and genocide.

* Afghanistan
* Angola
* Bangladesh
* Bosnia and Herzegovina
* Cambodia
* Democratic Republic of the Congo
* Eritrea
* Ethiopia
* Guatemala
* Iraq
* Lebanon
* Myanmar
* Nigeria
* Sierra Leone
* Sudan
* Timor-Leste
* Uganda
* Vietnam
* Yemen

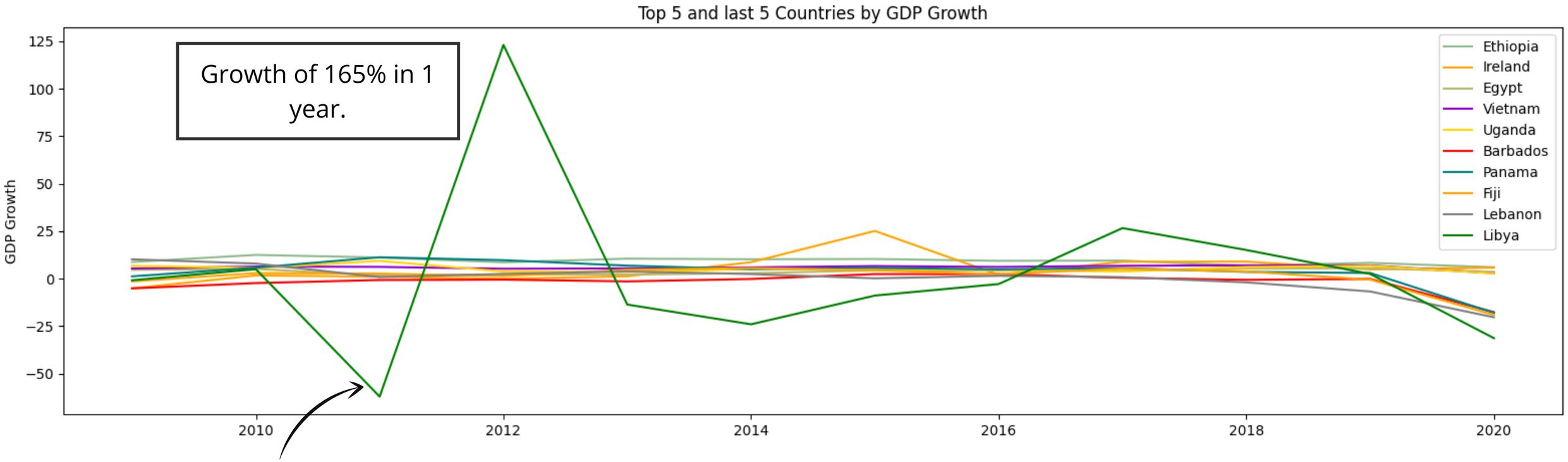
## USSR Countries

The fall of Soviet Union affected many countries especially Russia.


* Armenia
* Azerbaijan
* Belarus
* Estonia
* Georgia
* Kazakhstan
* Kyrgyzstan
* Latvia
* Lithuania
* Moldova
* Russia
* Tajikistan
* Turkmenistan
* Ukraine
* Uzbekistan

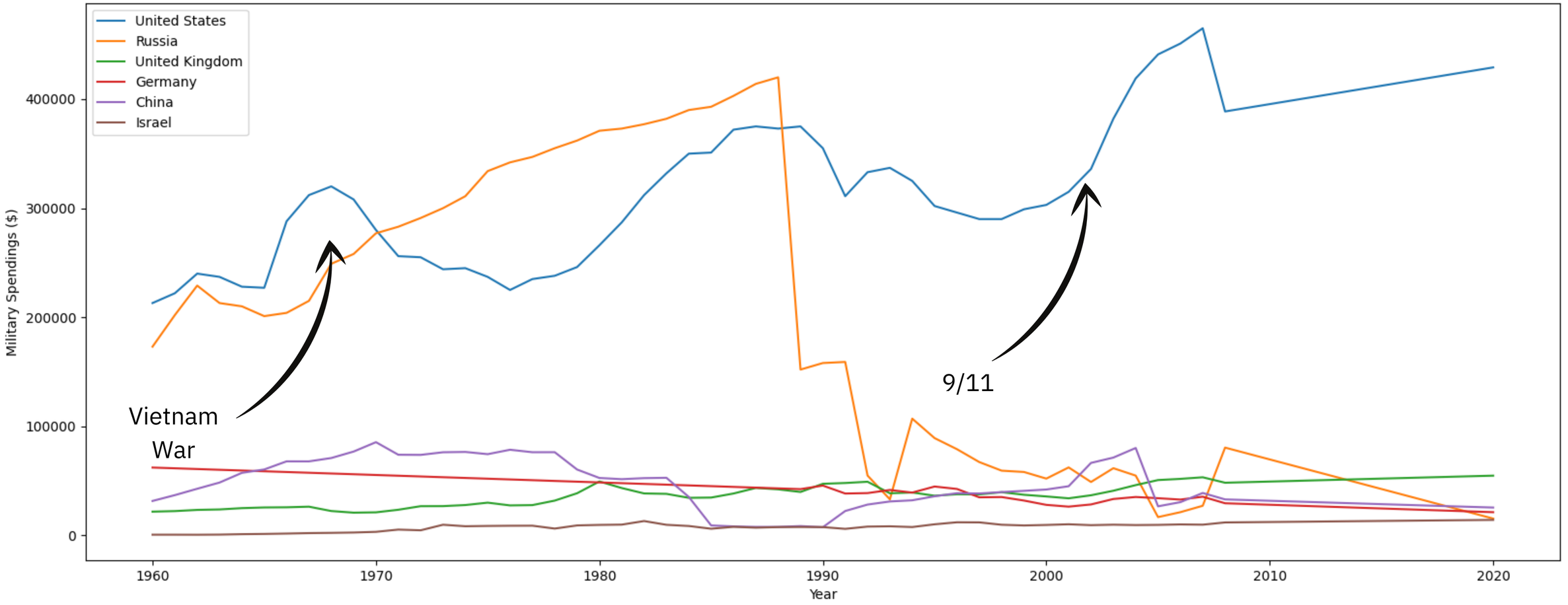

Population Growth pace per Genocide & Wars List along 1960-2020
Afghanistan
Angola
Bangladesh
Bosnia and Herzegovina
Cambodia
Democratic Republic of the Congo
Eritrea
Ethiopia
Guatemala
Iraq
Lebanon
Myanmar
Nigeria
Sierra Leone
Sudan
Timor-Leste
Uganda
Vietnam
Yemen
Lebanese Civil War
Lebanon Conflict
Lebanon Clashes + Syria civil war spillover in Lebanon
Afghanistan Civil War
Angola Colonial War
Timor Leste Genocide
Cambodian Genocide
Soviet-Afghan War
Bosnian Genocide
Sierra Leone Civil War
Eritrean–Ethiopian War
Timor-Leste Crisis
Lebanon Second War
8
6
4
2
0
−2
−4
1960
1970
1980
1990
2000
2010
2020
Year

# Muammar Gaddafi Libya Overtake 2010

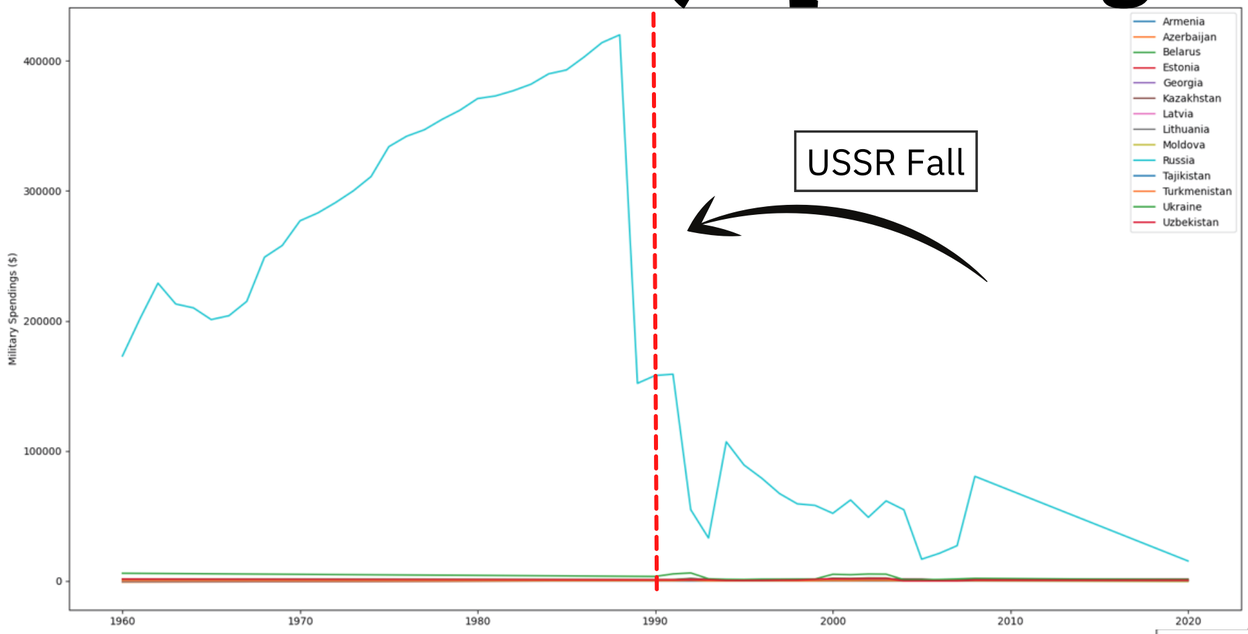



The people overtook corrupted leader Gaddafi

# USA 9/11 & Vietnam War



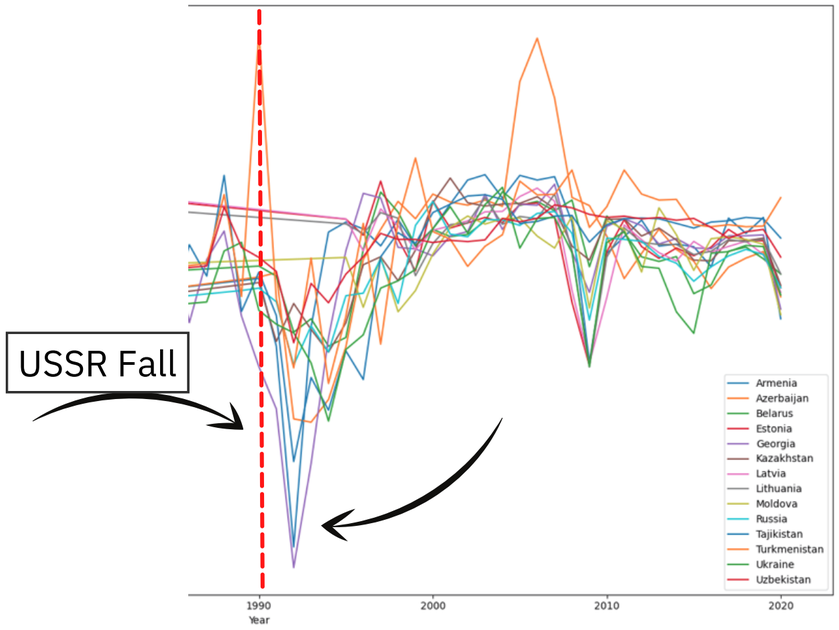
# USSR Military Spending's



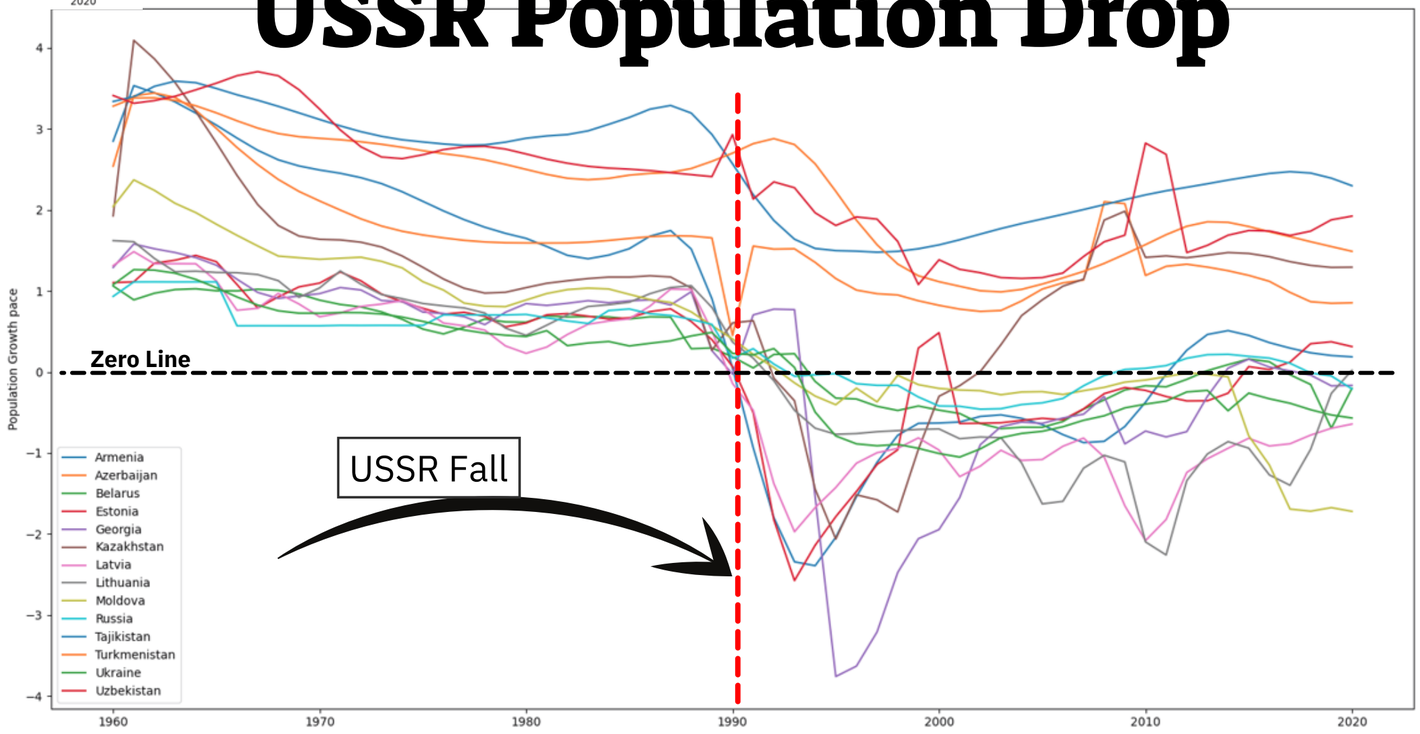

USSR collapsed in 1991,
we can see the effect on the Union members in the following graphs.

- Effect on the population
- Effect on the GDP
- Russia was affected the most regarding military spending's.

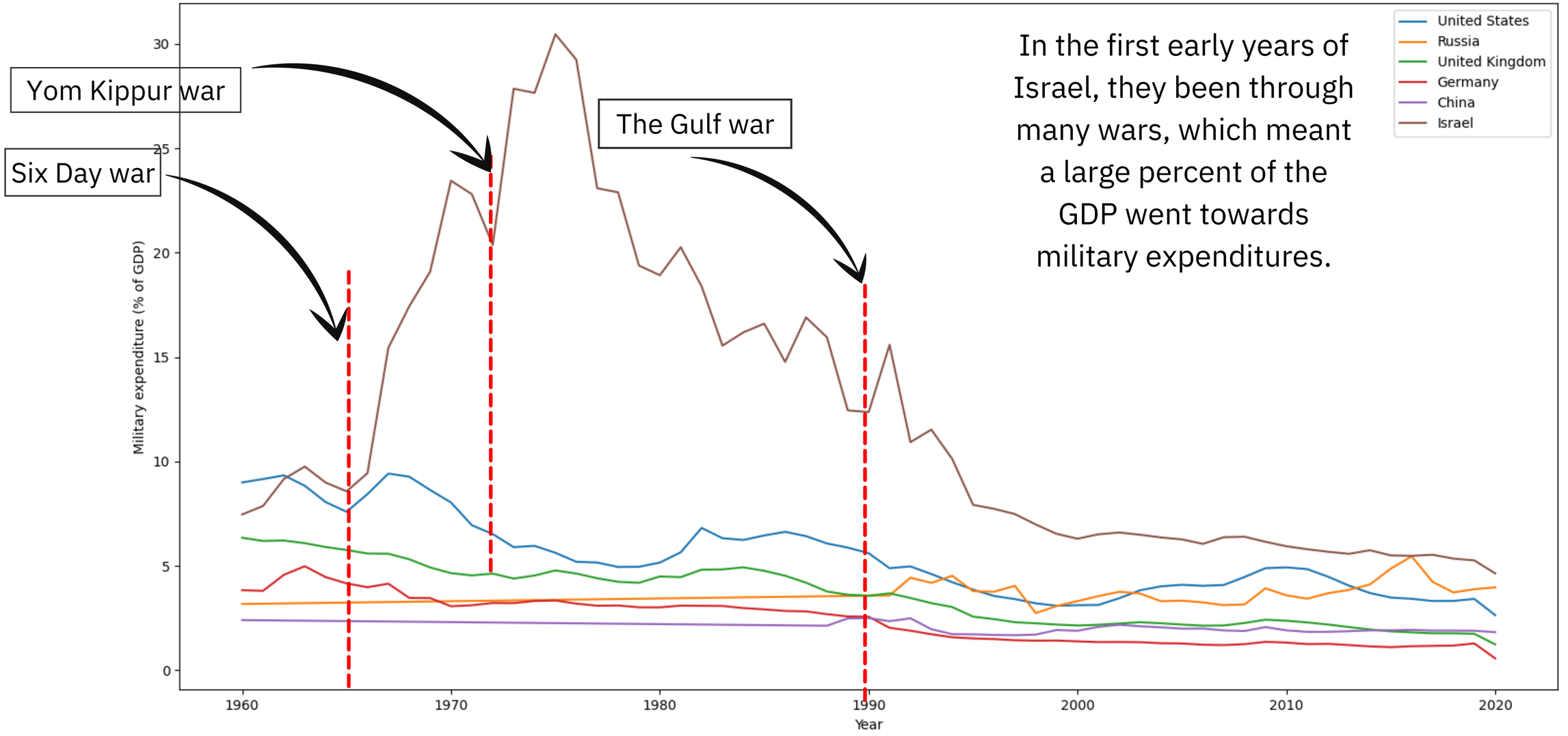
## USSR Total GDP



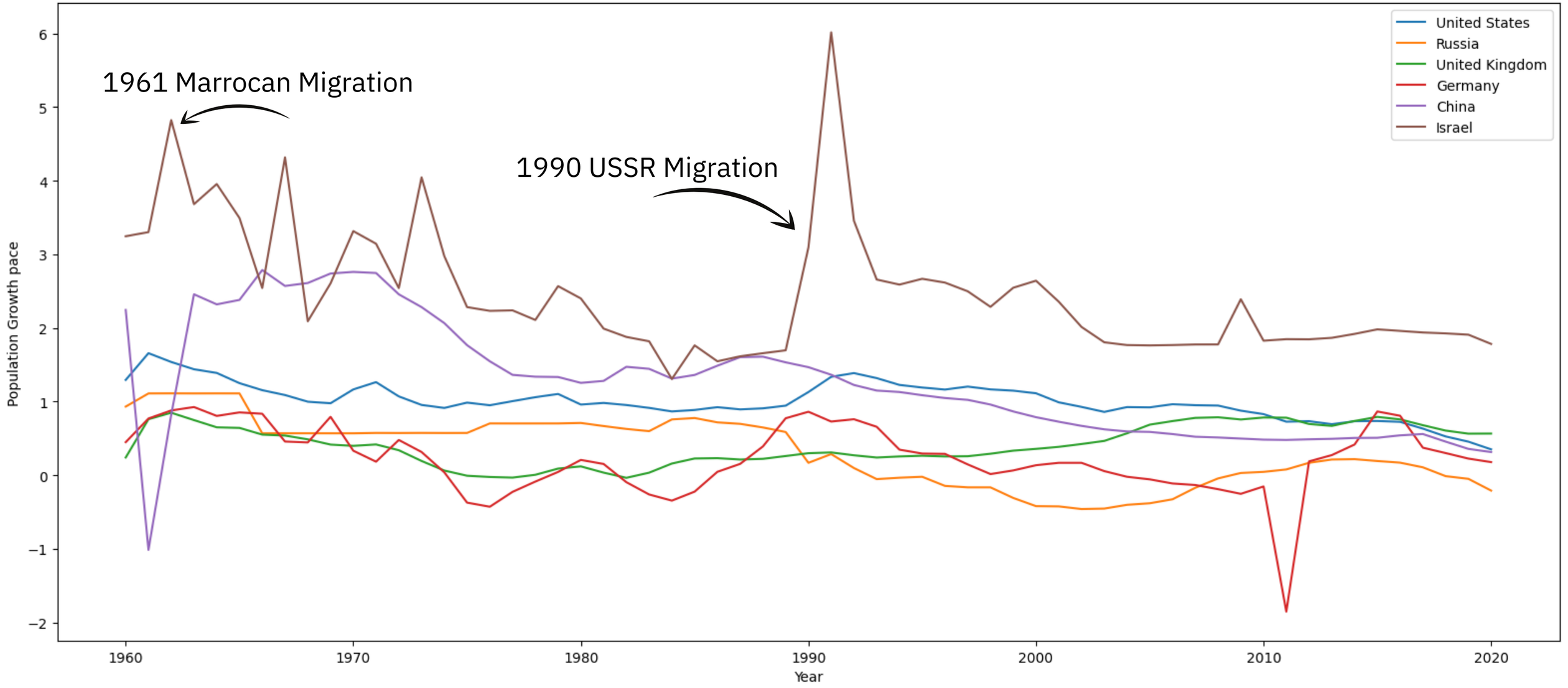

# USSR Population Drop



# Israel Military Spending



# Bonus Data: Population growth ISR



# CONCLUSIONS

01

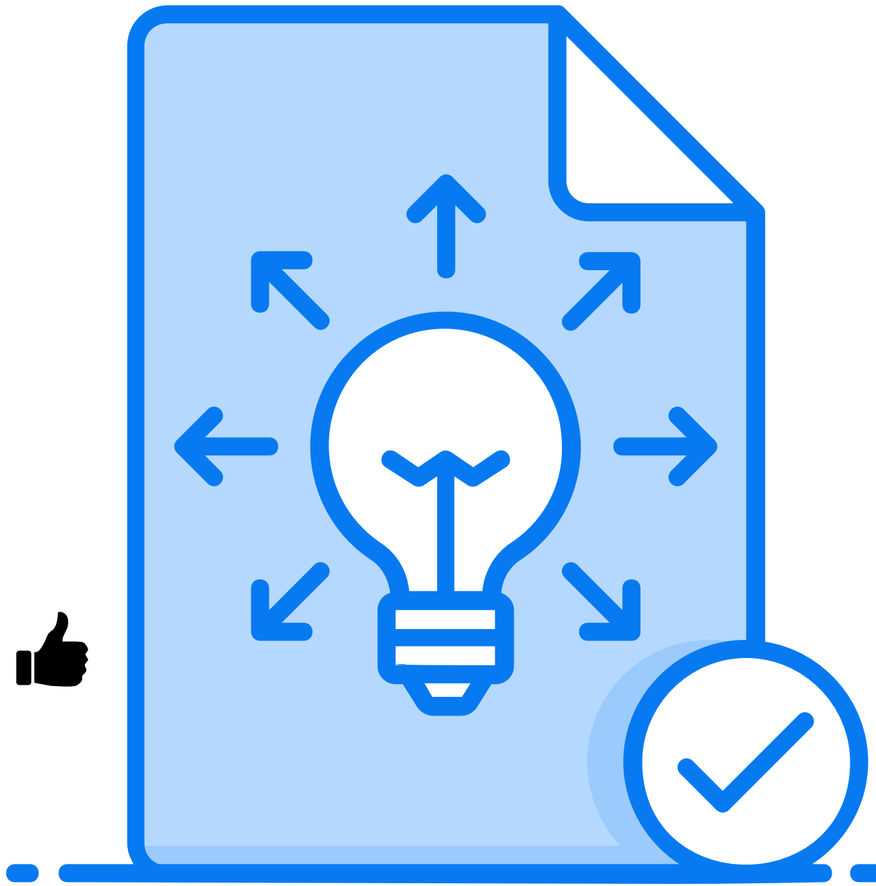
Most significant events do stand out in the graphs, and tell about the rise and fall of countries in term of finance and population.

02

Sadly, least developed/smaller countries are more sensitive to changes.
As shown in the graphs, this is highly accurate  when talking about Africa and Asia.
We've discovered many wars and genocides that are less familiar.

03

We've seen that revolution / collapsion does in fact impact a country's GDP.
i.e USSR collapse, Lybia post Muammar Gaddafi



# Thanks for Watching

## Any Questions?