Team Name: Hack Pack

Life Expectancy Analysis Report

Introduction

This report presents a comprehensive analysis of global life expectancy statistics using data from the UN, World Bank Group, World Health Organization, and the CIA World Factbook. The analysis highlights life expectancy trends by age, sex, and over time, revealing disparities and improvements among countries. Additionally, the concept of Healthy Life Expectancy (HALE) is discussed.

---

1. Data Sources Overview

- **Table 1:** UN estimates of life expectancy at different ages (2023)
- **Table 2:** Changes in life expectancy from 2019 to 2023
- **Table 3:** World Bank Group historical trends (2014–2022)
- **Table 4:** World Health Organization (2019) life expectancy and HALE
- **Table 5:** CIA World Factbook (2022) by sex and sex gap
- **Table 6:** Life expectancy estimates in the OCED for 2022

---

2. Project Overview

This project focuses on analyzing life expectancy data for various countries, with emphasis on differences between males and females across time and age groups. The goal is to understand:

- Gender-based longevity gaps
- Time-based changes in life expectancy
- Trends across different geographic regions

Data will be scraped and compiled from **Wikipedia**, cleaned using Python and regular expressions, analyzed statistically, visualized using data plots, and stored in a **MongoDB** database. The project concludes with an optional Streamlit dashboard to display interactive insights.

3. Project Objectives

We aim to answer questions like:

- Which countries have the **highest** and **lowest** life expectancy?
- Where is the **male-female life expectancy gap** largest or smallest?
- How has life expectancy **changed from 2019 to 2023** in various countries?
- Are countries improving or declining in health over time?

---

4. Methodology

Data Extraction

- Scrape tables using **BeautifulSoup** and **Pandas** .
- Export and store raw tables in Excel format.

Data Cleaning and Regular Expressions

- Handle nulls, invalid or missing values.
- Normalize country names and formats.
- Use **Regex** to:
    - Parse columns with multiple values into structured forms.

Data Analysis

- Compute:
    - Mean, median, and standard deviation of life expectancy.
    - Gender gap by country and year.
    - Year-over-year growth or decline.
- Compare:
    - Continents (Asia, Europe, etc.)
    - WHO vs World Bank vs CIA reports

Data Visualization

- **Bar charts** for male vs. female life expectancy.
- **Line graphs** for year-over-year trends.
- **Heatmaps** to show global life expectancy distribution.
- **Histograms/Boxplots** for gap analysis.

Data Storage

- Final processed data stored in **MongoDB** for flexible querying and scalability.

Bonus: Streamlit Web App

- A simple dashboard for exploring:
  - Top/bottom countries
  - Male-female gap over time
  - Country-specific graphs

---

1. Cleaning & Data Overview phase

Sheet1:UN: Change of life expectancy from 2019 to 2023

This dataset presents the change in life expectancy across different countries between 2019 and 2023,based on United Nations estimates, highlighting trends of increase or decrease during this period**.**

Sheet2:World Health Organization (2019)

This dataset, published by the World Health Organization in December 2020, provides life expectancy at birth for all populations as of 2019. If countries have equal life expectancy values, they are further sorted by HALE (Healthy Life Expectancy) for the total population.

Sheet3:CIA World Factbook (2022)

This dataset, published by the CIA World Factbook in 2022, provides life expectancy data categorized by total population, male, female, and the gender gap between males and females.

Sheeet4:World Bank Group (2022)

The dataset provides World Bank Group estimates for life expectancy across different countries in 2022, focusing only on nations with populations exceeding 50,000 people. It includes comparative data for the years 2014, 2019, and 2022 to analyze trends, particularly noting that 2014 marked a local peak in life expectancy for some leading countries. Data rounding may cause minor inconsistencies (up to 0.01 year) when compared to the raw calculations. The main columns include the country name, life expectancy for the three reference years, and the population size in 2022

- Here are some of our cleaning process on "**sheet 4"**:

```
import pandas as pd
sh4=pd.read_excel(r"C:\Users\sh138\OneDrive\Documents\sh4.xlsx",header=1)
sh4.columns= ['Countries and territories', 'All', 'Male', 'Female', 'Sex gap','2014', '2014:2019', '2019','2019:2020', '2020', '2020:2021',
```

- ✓ First, we read the data then use 'header' to handle the index and the columns.
- ✓ Secondly, we gave each column its proper name.

```
sh4.head()
```

	Countries and territories	All	Male	Female	Sex gap	2014	2014:2019	2019	2019:2020	2020	2020:2021	2021	2021:2022	2022	recovery from COVID-19: 2019:2022	
0	Macao SAR, China	85.38	82.75	87.98	5.23	83.88	1.09	84.98	0.21	85.18	0.22	85.40	-0.02	85.38	0.40	NaN
1	Japan	84.00	81.05	87.09	6.04	83.59	0.77	84.36	0.20	84.56	-0.11	84.45	-0.45	84.00	-0.36	NaN
2	Hong Kong SAR, China	83.66	80.70	86.77	6.07	83.94	1.22	85.16	0.34	85.50	0.04	85.53	-1.87	83.66	-1.49	NaN
3	French Polynesia	83.55	81.33	85.98	4.65	82.21	0.78	83.00	-0.78	82.22	-2.73	79.49	4.06	83.55	0.55	NaN
4	Switzerland	83.45	81.60	85.40	3.80	83.20	0.71	83.90	-0.90	83.00	0.75	83.75	-0.30	83.45	-0.45	NaN

```
sh4.drop(' ', axis=1, inplace=True)
```

- ✓ Dropping a null column.

```
dtype: object
cols_to_convert = ['All', 'Male', 'Female', 'Sex gap', '2014', '2014:2019', '2019', '2019:2020', '2020', '2020:2021', '2021:2022', '2022', 'recovery from COVID-19: 2019:2022']

for col in cols_to_convert:
    sh4[col] = (sh4[col].astype(str)
                .str.replace('−', '-')
                .str.replace(',', '.')
                .str.replace(r'[^\d.-]', '', regex=True))
    sh4[col] = pd.to_numeric(sh4[col], errors='coerce')
```

```
sh4['All'] = sh4['All'].astype(float)
sh4['Male'] = sh4['Male'].astype(float)
sh4['Female'] = sh4['Female'].astype(float)
sh4['Sex gap'] = sh4['Sex gap'].astype(float)
sh4['2014'] = sh4['2014'].astype(float)
sh4['2014:2019'] = sh4['2014:2019'].astype(float)
sh4['2019'] = sh4['2019'].astype(float)
sh4['2019:2020'] = sh4['2019:2020'].astype(float)
sh4['2020'] = sh4['2020'].astype(float)
sh4['2020:2021'] = sh4['2020:2021'].astype(float)
sh4['2021:2022'] = sh4['2021:2022'].astype(float)
sh4['2022'] = sh4['2022'].astype(float)
sh4['recovery from COVID-19: 2019:2022'] = sh4['recovery from COVID-19: 2019:2022'].astype(float)
```

- ✓ There was some unknown symbols, we handled them then convert the data types to be appropriate

```
columns_to_fill = [
    "All", "Male", "Female", "Sex gap",
    "2014", "2014:2019", "2019", "2019:2020",
    "2020", "2020:2021", "2021:2022", "2022",
    "recovery from COVID-19: 2019:2022"
]

for col in columns_to_fill:
    sh4[col] = sh4[col].fillna(sh4[col].mean())
```

- ✓ Handling nulls values.

Sheet5:UN: Estimate of life expectancy for various ages in 2023

This dataset presents the United Nations' 2023 estimates of life expectancy at various ages for different countries, with location links directing to detailed demographic profiles for each country.

Sheet6:OECD (2022)

This dataset provides life expectancy estimates for OECD countries in 2022, with the default sorting based on life expectancy values from 2019.

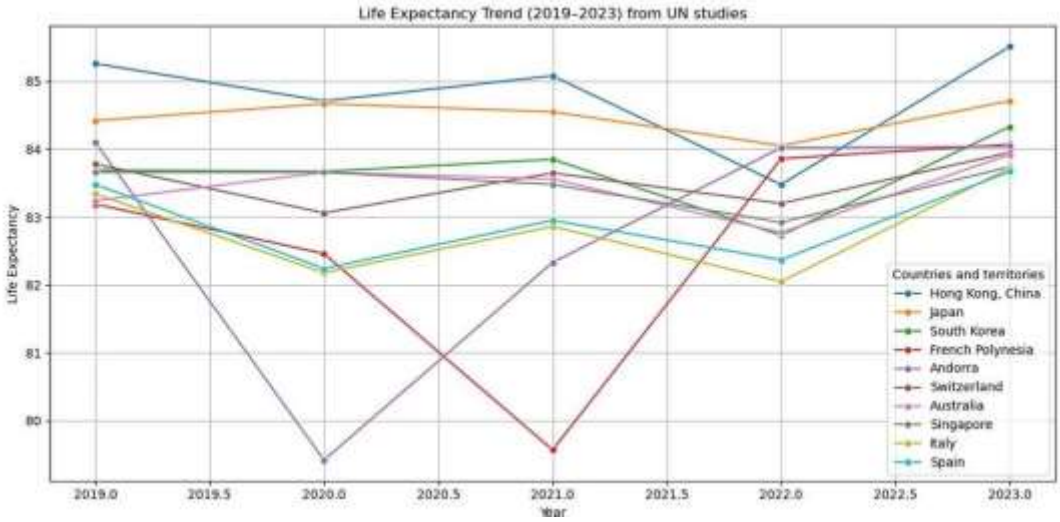
---

2. Analysis phase

i. First table (Estimate of life expectancy for various ages in (2023):

	Countries and territories	2023
0	Hong Kong, China	85.51
1	Japan	84.71
2	South Korea	84.33
3	French Polynesia	84.07
4	Andorra	84.04

- The highest 5 countries estimated life expectancy in 2023



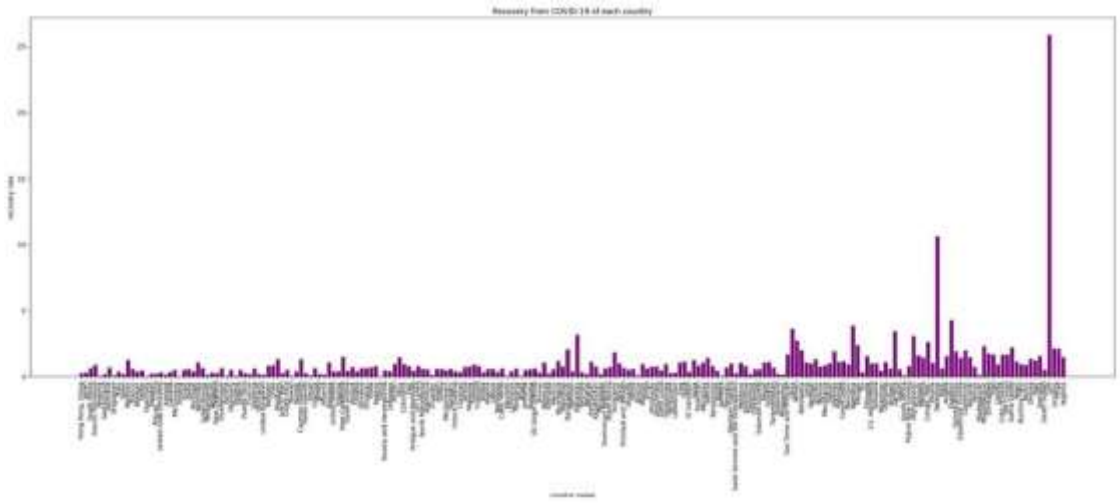
- From this time line graph we found that:
  - In 2019, Hong Kong, China had the highest life expectancy, and French Polynesia had the least.
  - Andorra got the highest drop in 2020, followed by French Polynesia in 2021.
  - Hong Kong, China held the lead from 2019 to 2023, with the exception of 2022.

	Countries and territories	Recovery from COVID-19: 2019:2023
10	Réunion	1.23
25	Bermuda	1.06
42	Maldives	1.33
47	Cayman Islands	1.31
53	Panama	1.08
...	...	...
205	Somalia	1.57
207	CAR	25.88
208	Lesotho	2.13
209	Chad	2.08
210	Nigeria	1.45

  - Now we can see the countries with high recovery after COVID-19 (Above average of life expectancy)

	Countries and territories	Recovery from COVID-19: 2019:2023
0	Hong Kong, China	0.25
1	Japan	0.29
2	South Korea	0.64
3	French Polynesia	0.88
4	Andorra	0.06
...	...	...
192	Madagascar	0.13
196	Liberia	0.93
201	Burkina Faso	0.91
202	Benin	0.89
206	South Sudan	0.51

- And these are the countries with low recovery after COVID-19 (less of life expectancy)



- This bar plot shows the recovery from COVID-19 of each country.

---

ii. 2- The second table (World Bank Group (2022)):

	Countries and territories	2014:2022
139	St. Vincent and the Grenadines	5.50
86	Lebanon	4.55
166	Yemen	3.66
92	Oman	3.51
141	Guatemala	3.29
143	Ukraine	2.60
131	Paraguay	2.41
95	Colombia	2.38
128	Jamaica	2.35
124	Belize	2.35

- The countries with highest wiggle in life expectancy from 2014 to 2019.

	Countries and territories	recovery from COVID-19: 2019:2022
2	Hong Kong SAR, China	1.49
19	Iceland	0.99
37	Greece	1.00
57	USA	1.35
58	Costa Rica	2.11

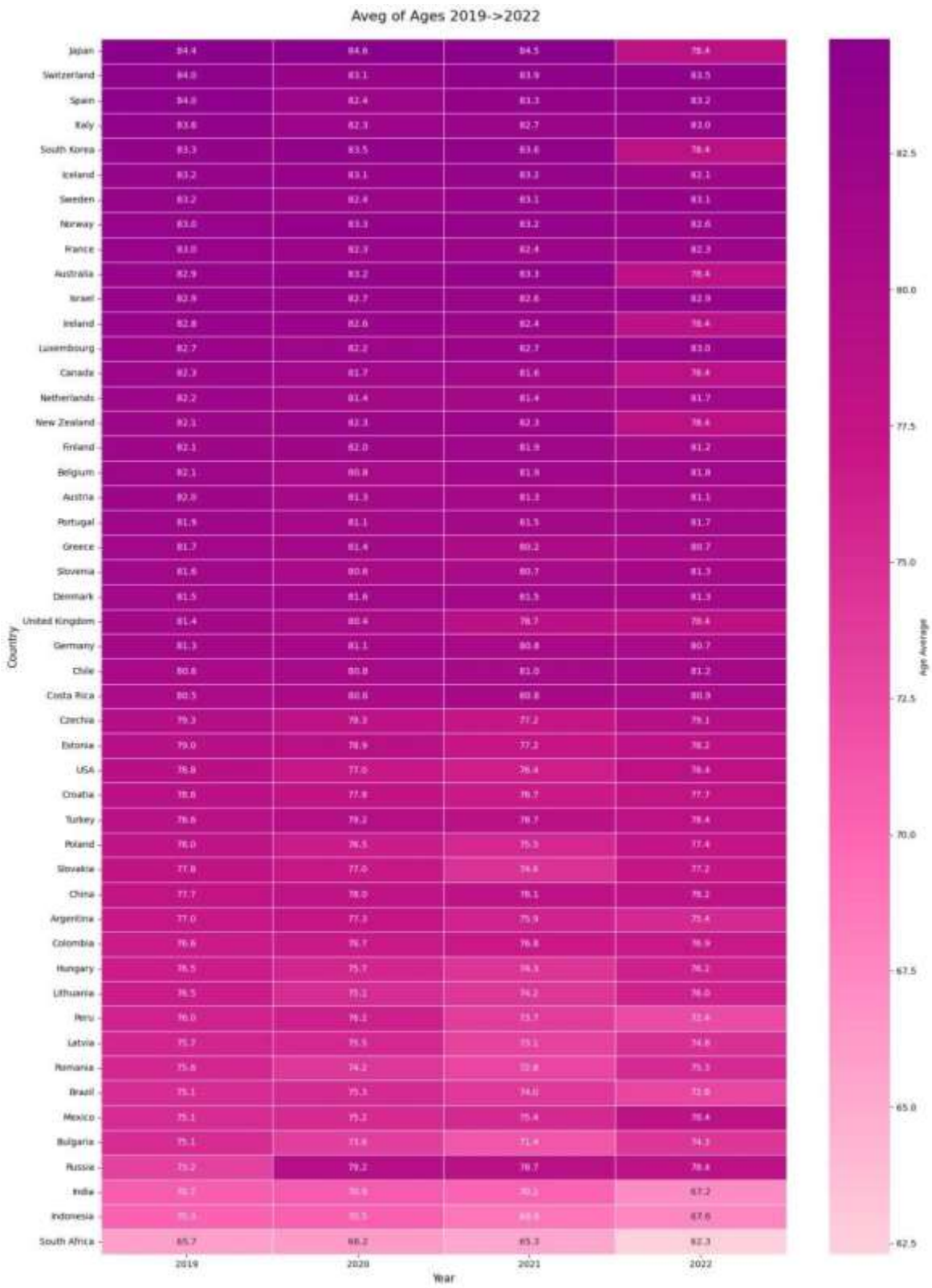
- Countries with the highest recovery from COVID-19.

	Countries and territories	recovery from COVID-19: 2019:2022
0	Macao SAR, China	0.40
1	Japan	0.36
3	French Polynesia	0.55
4	Switzerland	0.45
5	Faroe Islands	0.65
...	...	...
193	Cote d'Ivoire	0.40
197	South Sudan	0.34
198	Central African Republic	0.55
199	Nigeria	0.72
201	Chad	0.26

- Countries with the least recovery from COVID-19.

	Countries and territories	Sex gap
102	Armenia	10.58
143	Ukraine	10.38
103	Belarus	10.29
107	Russia	10.20
80	Latvia	9.80
115	Georgia	9.70
81	Vietnam	9.36
105	Mongolia	9.30
142	Moldova	9.10
117	El Salvador	8.98

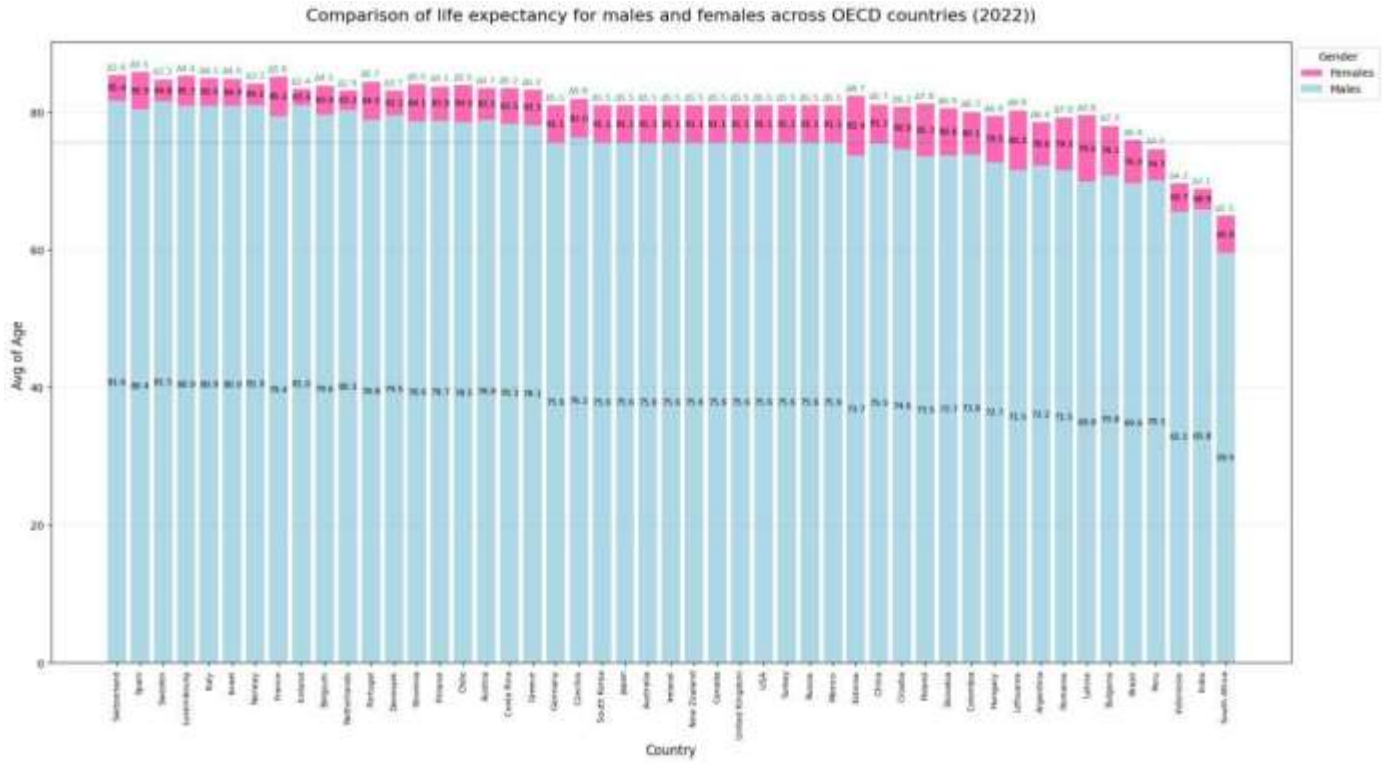
- The countries with the highest gap between male and female life expectancy.



- In this heat map we compare the Average of ages in Countries based on OECD report and we can know that :

1) Japan and Switzerland has the highest value.
2) South Africa and Indonesia has lowest value
3) Some countries have firmness in values over years
4) Color changes mean that there is increase or decrease in values

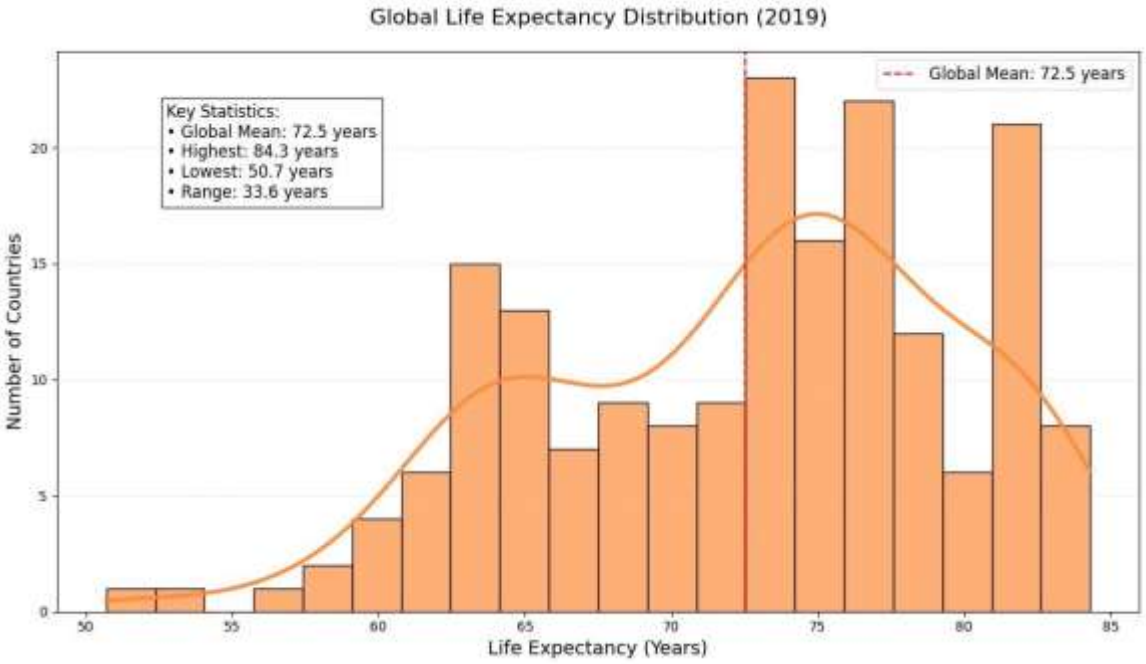
---



- In this plot we compare the ages between males and females in each country and we concluded that :

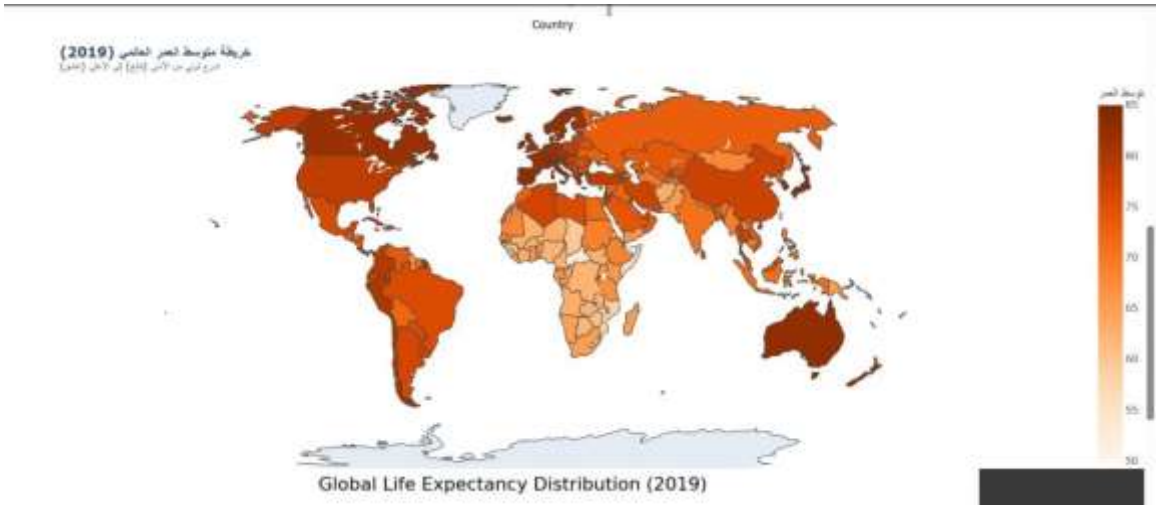
1) In Spain females has the highest value and South Africa has the lowest value
2) In Switzerland men has the highest value South Africa has the lowest value
3) Mostly the difference between genders isn't big and we want to say that this difference may occur because of different levels of health care between them or different cultures between countries
4) The blue line show the averages and white lines show where each country is belong above or below the average line.

iii. Third table (World Health Organization):



- In this plot we understand how life expectancy is distributed among countries and support analysis with statistical data that facilitate the results and we conclude that:

1) Most countries are located in range (65:80).
2) The distribution of ages is asymmetrical.
3) Some countries have low ages (50) that refers to problems of health.
4) Most of countries are close to the average of ages (72.5)
5) There is a clear disparities between countries because the range is a little big.

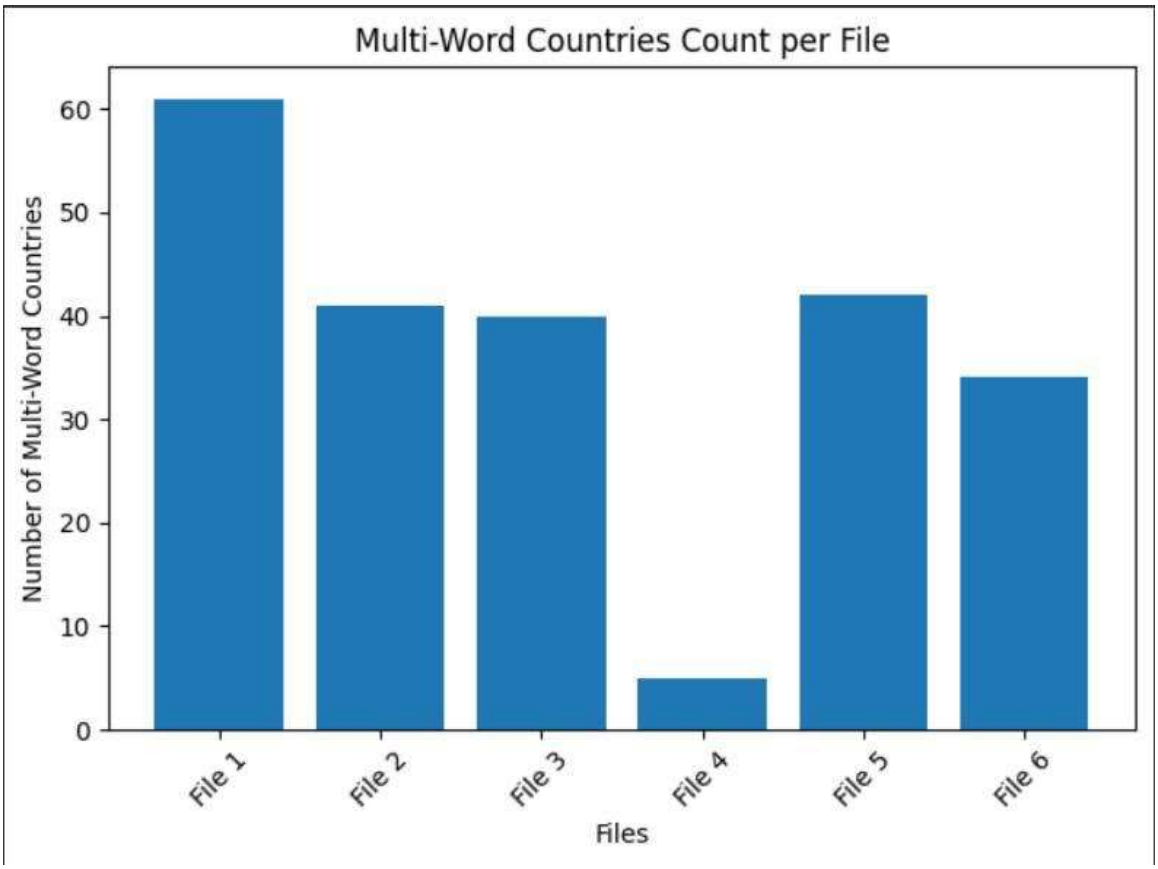


- We Creates an interactive world map (Choropleth Map) showing Life Expectancy per country in 2019 and we concluded that:

1) Western Europe, Australia show darker colors (higher life expectancy).
2) Some African and Asian countries show lighter colors (lower life expectancy).
3) The map reflects health care globally.
4) Some countries have shown significant improvements compared to 2000, indicating progress.

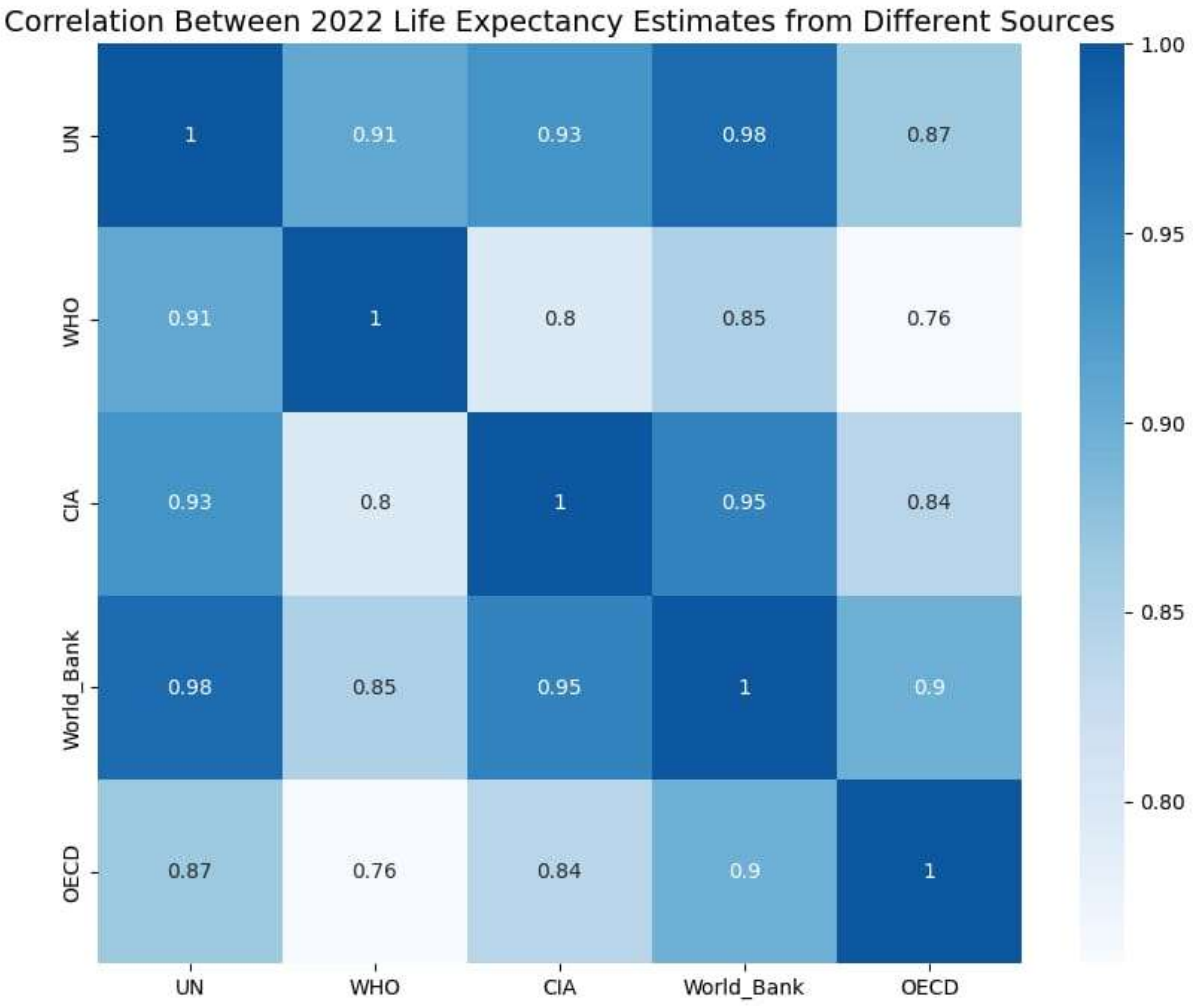
❖ The Regular expression

- These tables have the countries that have more than one word in each file after implement regular expression on columns of countries names in these tables.

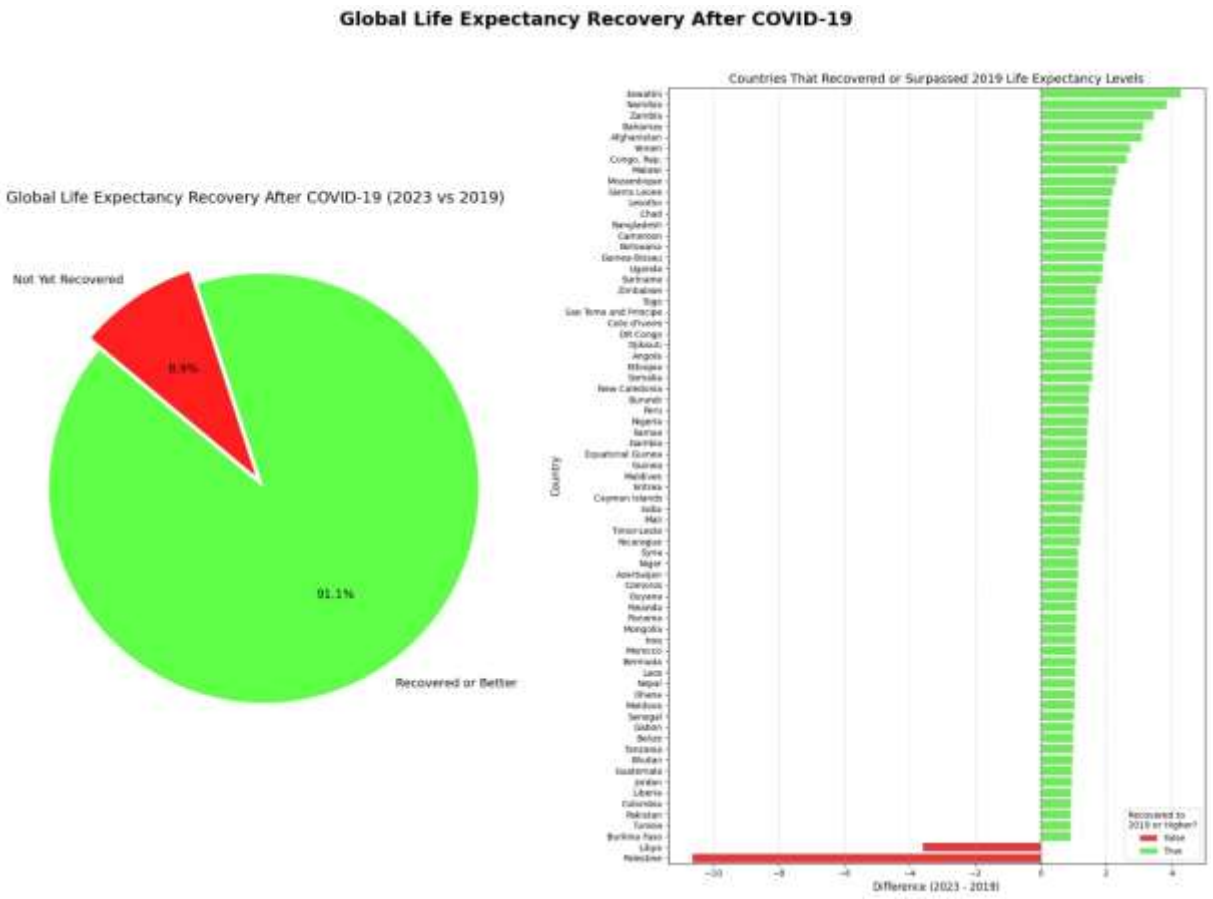


- In this plot we show number of countries that has more than one word in each file after implement Regular Expression on columns that have names of countries.

3. The comparisons and relationships between sources

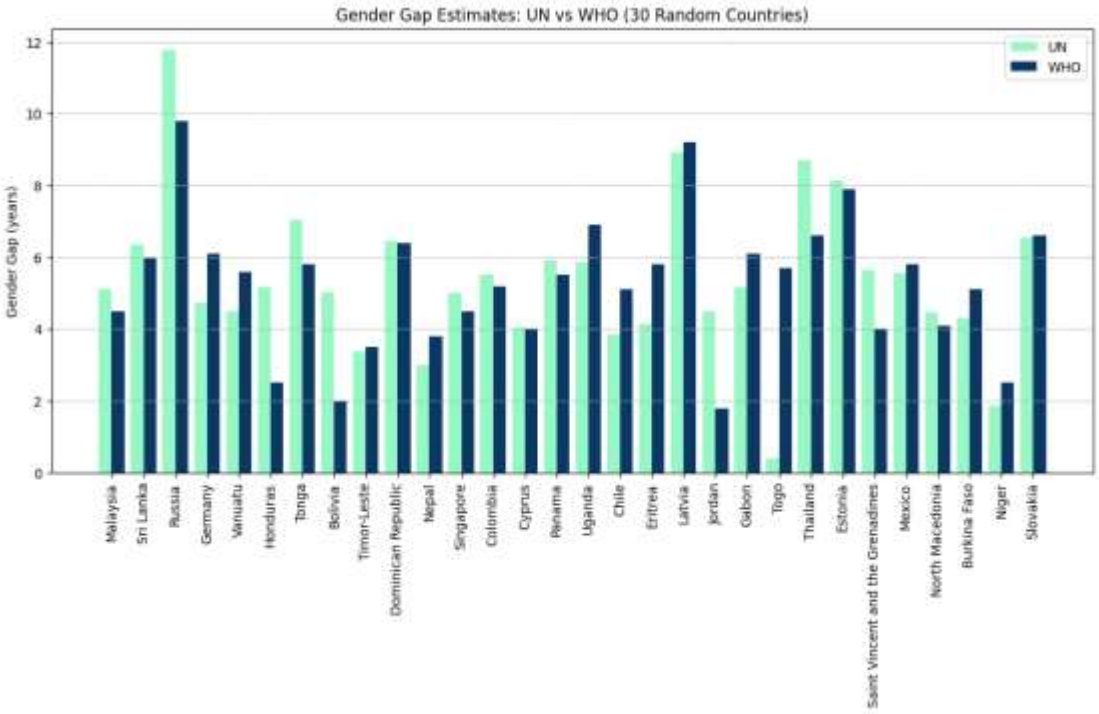


- The heatmap visually represents the degree of consistency among sources of life expectancy estimates for 2022 by examining their correlation coefficients.

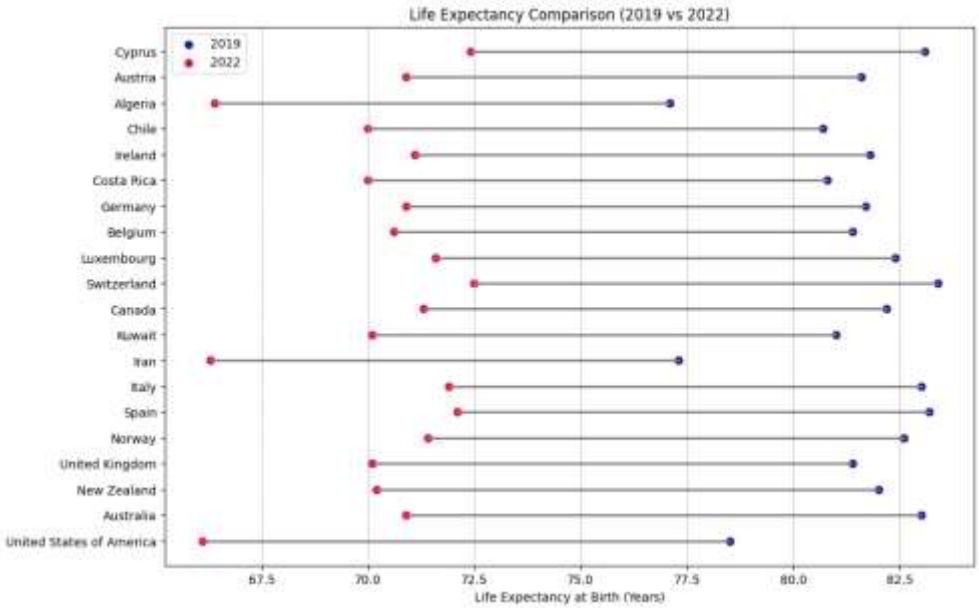


- The pie chart illustrates the percentage of countries that have recovered or surpassed their pre-pandemic life expectancy compared to those that have not yet recovered.

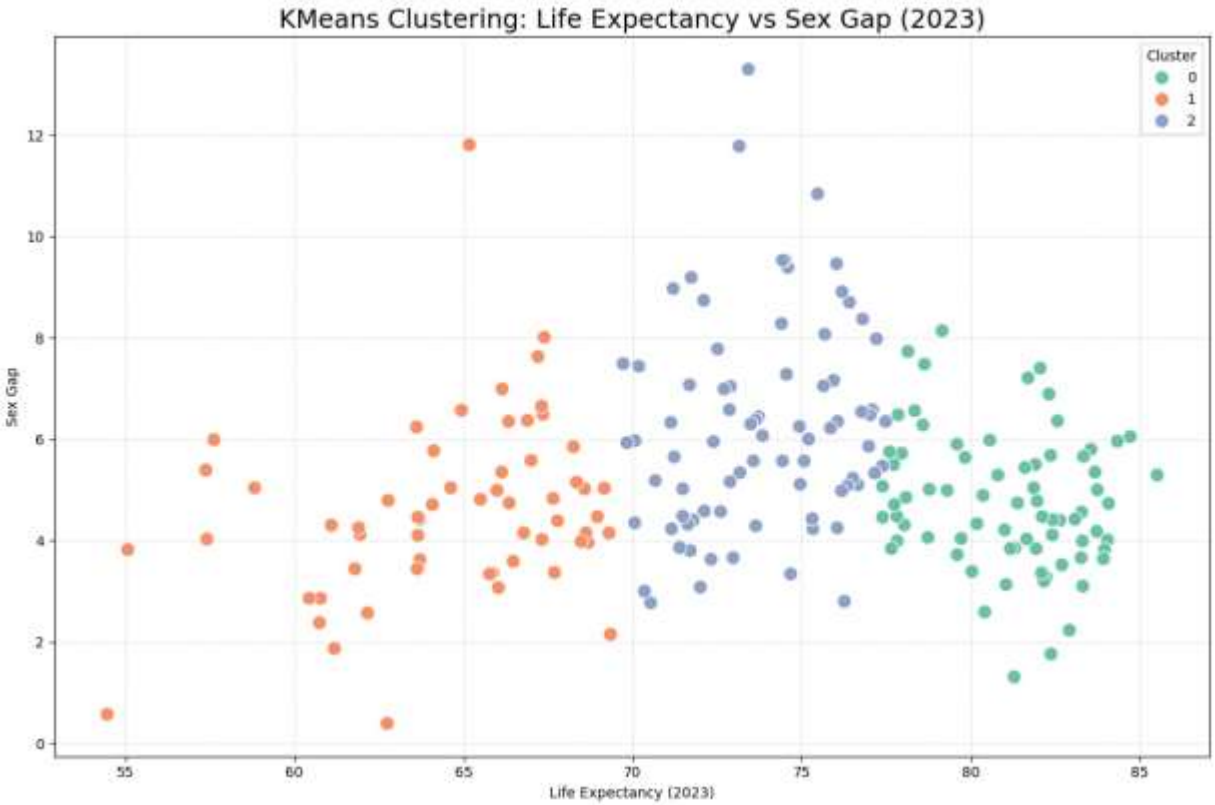
- The bar chart highlights the change in life expectancy for each country between 2019 and 2023, distinguishing between countries that have recovered and those that have not.



- The bar chart compares gender gap estimates in life expectancy between United Nations and World Health Organization data for 30 randomly selected countries.



- This chart displays the top 20 countries most affected by the COVID-19 pandemic in terms of decline in life expectancy at birth, based on a comparison between 2019 and 2022 data



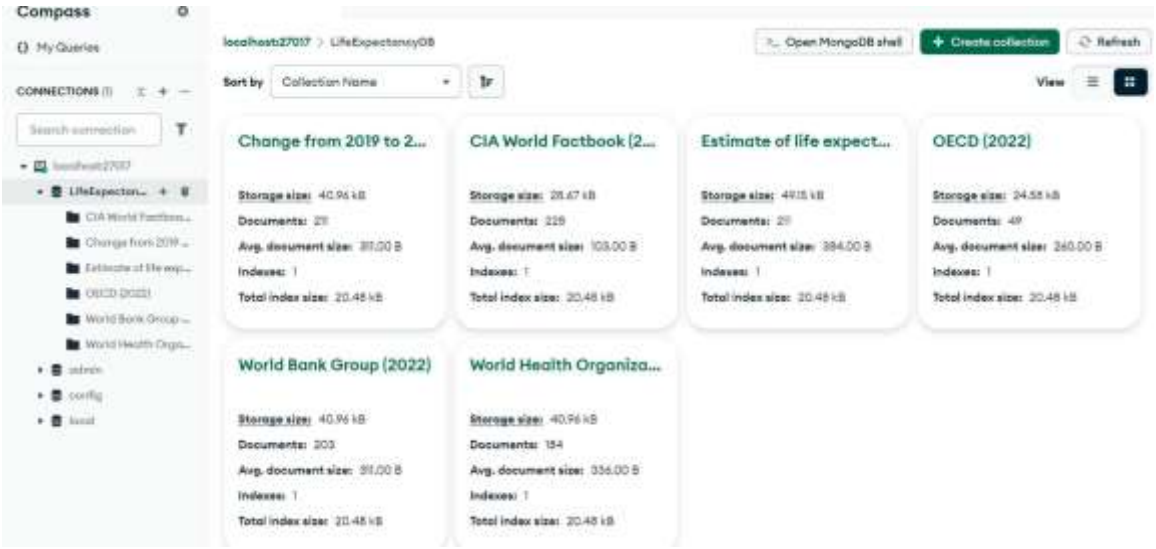
	Cluster 0	Cluster 1	Cluster 2
0	Antigua and Barbuda	Libya	Hong Kong, China
1	Sri Lanka	Yemen	Japan
2	Argentina	Botswana	South Korea
3	Ecuador	Laos	French Polynesia
4	Guam	Senegal	Andorra
...	...	...	...
75	Guyana	NaN	NaN
76	Turkmenistan	NaN	NaN
77	Greenland	NaN	NaN
78	Philippines	NaN	NaN
79	Sao Tome and Principe	NaN	NaN

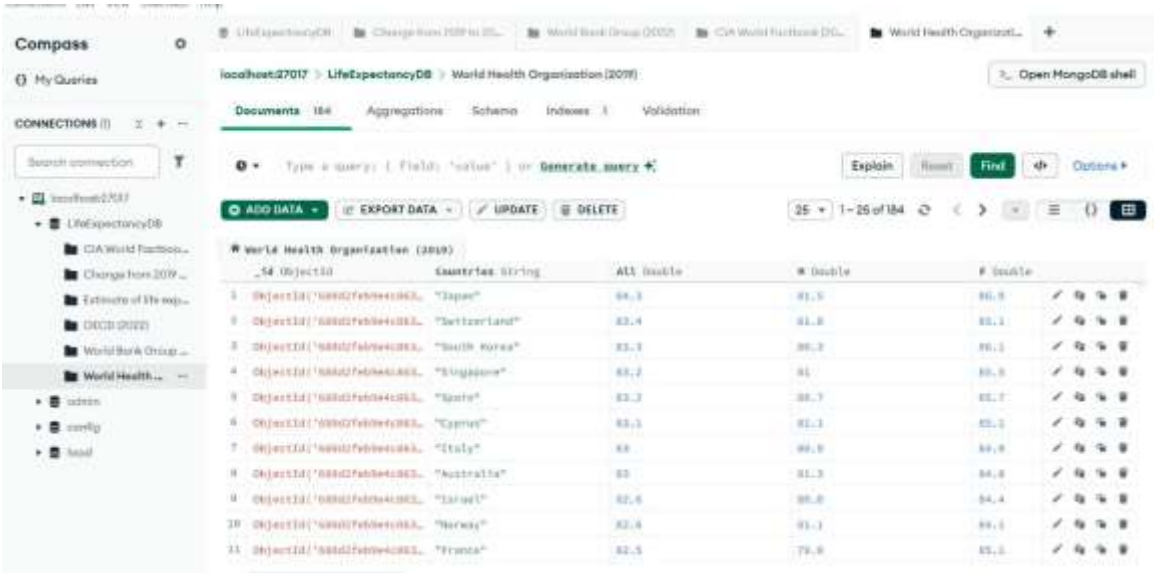
80 rows × 3 columns

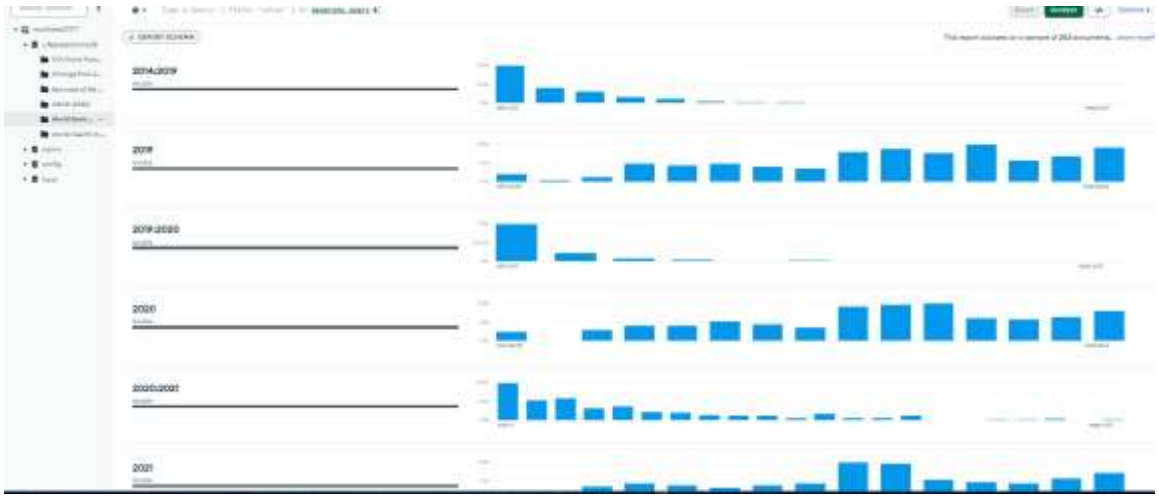
- The chart is used to classify countries into three groups based on their 2023 life expectancy, recovery from COVID-19, and gender gap, understanding the relationship between life expectancy, the gender gap, and COVID-19 recovery levels.

4. Data Storage (MONGODB)

- We stored data from various sources about life expectancy information in a MongoDB database for easier access and analysis.







- This schema provides an overview of how life expectancy data is distributed for each year or period. It reveals any irregularities or missing values across the dataset.

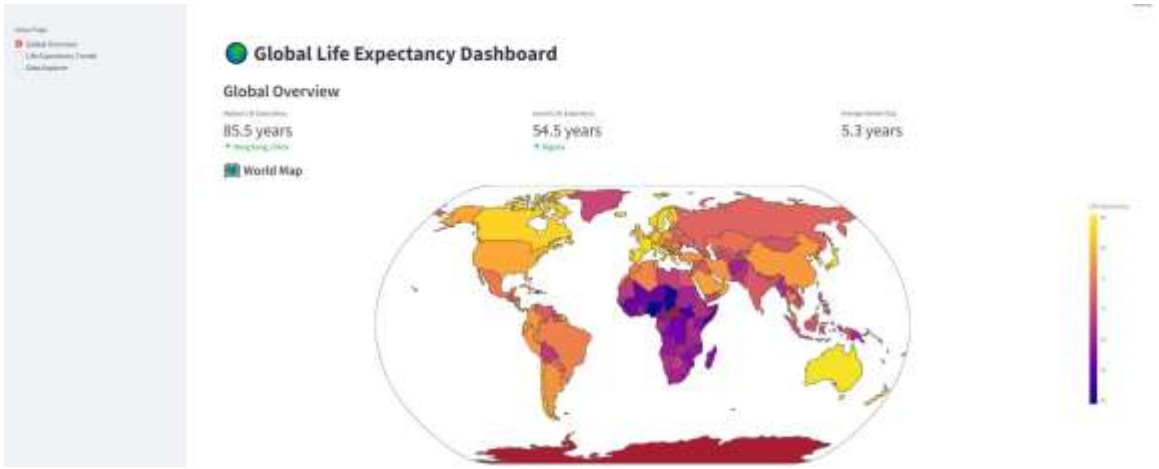
---

5. The Streamlit Phase

- The dashboard contains three interfaces: Global Overview, Gender Analysis, and Data Explorer. Each interface offers a unique perspective on life expectancy.

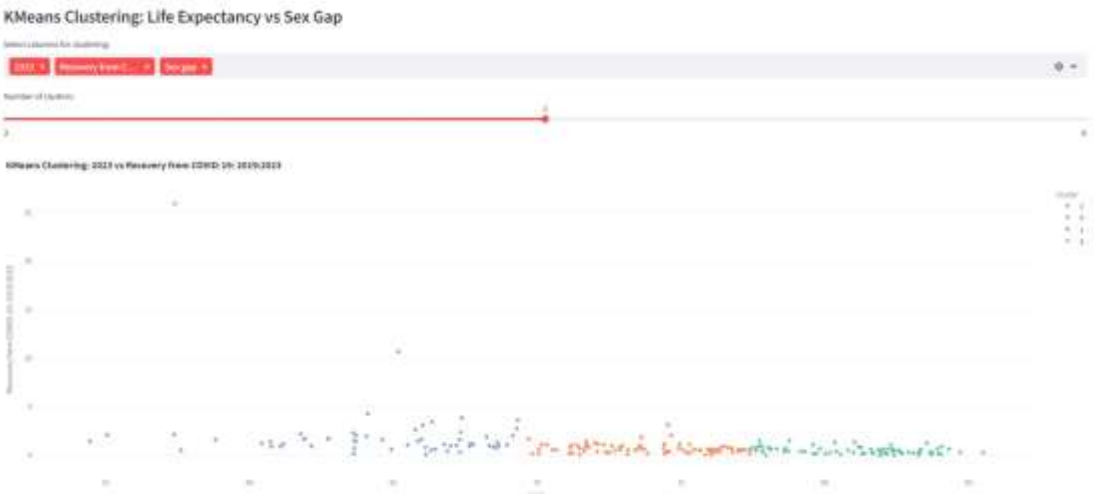
1. Global Overview It displays :

a) Summary Statistics about life expectancy and gender gap.
b) Interactive world map showing life expectancy by country.



c) The tables present the top 50 and bottom 50 countries ranked by life expectancy in 2023, along with the gender gap in these countries

---

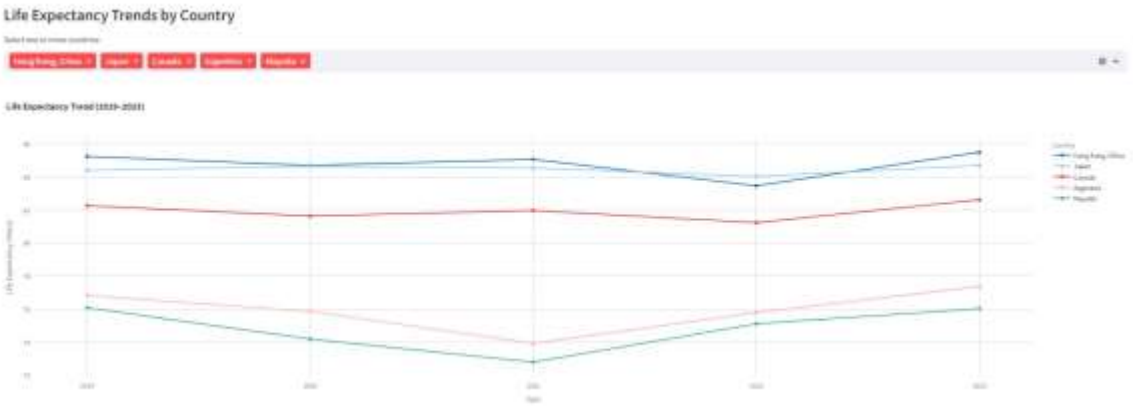


d) This part applies KMeans clustering to group countries based on life expectancy in 2023 and the gender gap and display it in split tables.

2- Life Expectancy Trends It displays :

a) Scatter plot: displays the relationship between life expectancy in 2023 and the gender gap , highlighting specific countries and a trend line to show the general pattern of the relationship between life expectancy and the gender gap.

b) The tables display the top 30 countries with the largest gender gaps in life expectancy, and the 30 countries with the smallest gender gaps.



c) This plot displays the trend of life expectancy for selected countries over the years 2019 to 2023.

d) This chart compares the changes in life expectancy at birth between 2019 and 2022 for different countries, showing how COVID-19 may have impacted life expectancy.

3- Data Explorer : it explore raw datasets from various sources. It desplay :

a) A selection menu to choose a dataset.

b) Basic information like number of rows & columns , column names and some statstics.

c) Finnaly , represent full data with interactive way

Select Page:

- Global Overview
- Life Expectancy Trends
- Data Explorer

Global Life Expectancy Dashboard

Data Explorer

Select a table to display:

Estimate of life expectancy for various ages in 2019

Displaying Estimate of life expectancy for various ages in 2019

Basic info

Number of rows: 211

Number of columns: 33

Column Names:


```
" ]
  0 : "Location"
  1 : "at birth"
  2 : "female(5)"
  3 : "at 1"
  4 : "female(1to4)"
  5 : "at 15"
  6 : "female(5to64)"
  7 : "at 60"
  8 : "at birth.1"
  9 : "at 1.1"
 10 : "at 15.1"
 11 : "60.1"
 12 : "at birth.2"
 13 : "at 1.2"
 14 : "at 15.2"
 15 : "at 60.2"
 16 : "at birth.3"
 17 : "at 1.3"
 18 : "at 15.3"
 19 : "at 60.3"
]
```


Data overview:

	Location	at birth	female(5)	at 15	Female(1to4)	at 60	female(5to64)	at 60	at birth.1	at 1.1	at 15.1	60.1	at birth.2	at 1.2	at 15.2	at 60.2	at birth.3	at 1.3
count	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211
unique	211	Varies	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None
top	Hong Kong	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None
freq	1	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None
mean	None	73.8893	2.8333	59.8133	5.9371	16.7291	5.8008	7.09	71.1991	56.3982	21.3661	6.6078	70.4473	60.1638	18.0196	8.1391	5.2587	5.044
std	None	7.0985	1.9374	6.3389	3.7533	3.8516	1.3437	2.6002	7.0139	6.3616	3.5275	1.5663	7.1835	6.4866	3.3728	3.9366	1.8158	1.9859
min	None	54.46	0.16	45.4	1.21	11.13	3.85	4.31	53.2	44.77	16.79	4.52	54.74	48.79	12.26	4.76	0.39	0.23
25%	None	68.6	0.57	55.745	3.03	14.183	4.613	6.233	66.27	54.185	22.07	5.68	71.13	59.108	15.225	5.62	4.035	3.6
50%	None	74.29	3.21	61.27	5.97	16.4	5.94	7.32	73.31	57.73	24.73	6.41	76.18	64.22	17.91	7.9	5.32	4.75
75%	None	79.985	5.1	64.3	7.97	18.97	7.008	8.641	77.95	62.925	23.605	7.999	81.9	67.695	20.38	9.965	6.29	6.055

	Location	at birth	female(5)	at 15	female(5-64)	at 60	female(65+60)	at 60	at birth.1	at 1.1	at 15.1	60.1	at birth.2	at 1.2	at 15.2	at 60.2	at birth.3	at 1.3	at
0	Hong Kong	85.51	0.21	70.71	2.47	13.19	3.27	11.45	82.84	68.08	21.21	13.26	88.13	73.33	20.21	12.04	5.29	5.27	4
1	Japan	84.72	0.24	69.98	2.54	13.46	3.45	10.95	81.85	66.55	19.58	9.27	87.74	72.93	24.85	12.1	6.25	6.05	4.87
2	South Korea	84.33	0.3	69.63	2.69	13.32	4.14	10.96	81.35	66.91	18.63	8.4	87.14	72.44	21.59	11.68	5.96	5.53	4.93
3	French Polynesia	84.07	0.66	69.65	3.46	13.3	3.72	10.23	81.78	67.27	19.58	8.88	86.6	72.07	23.48	12.19	4.72	4.7	3.9
4	Andorra	84.04	0.26	69.52	2.85	13.47	3.72	10.13	82.1	67.68	19.81	9.91	86.11	71.7	22.18	11.41	4.81	4.55	3.46
5	Switzerland	83.85	0.29	69.31	2.2	12.51	3.48	9.97	82.61	67.38	20.07	9.68	85.92	71.17	22.82	13.62	3.82	3.6	2.79
6	Australia	83.82	0.33	69.27	2.38	13.09	3.28	10	82.1	67.19	20.32	9.14	85.74	71.09	22.91	14.79	3.64	3.62	2.82
7	Singapore	83.74	0.27	69	2.87	12.87	3.87	10.74	81.24	66.61	20.63	9.66	86.24	71.3	23.52	13.3	5	4.98	3.5
8	Italy	83.51	0.29	68.98	2.59	13.51	3.98	9.77	81.37	66.84	19.87	8.77	85.75	70.95	22.76	14.5	4.14	4.16	2.99
9	Spain	83.47	0.33	68.96	2.98	13.24	3.57	10.11	83.05	68.26	19.84	8.96	86.31	71.95	22.4	13.04	5.35	5.32	3.91
10	Réunion	83.35	0.6	68.24	3.1	12.54	3.67	10.36	83.03	68.16	20.37	9.85	86.33	71.89	23.83	13.6	5.8	5.73	3.46
11	France	83.33	0.39	68.71	2.81	12.93	3.98	10.01	83.61	68.95	20	9.41	86.68	71.44	23.19	12.95	5.86	5.59	3.78
12	Norway	83.21	0.23	68.56	2.18	13.75	3.63	9.48	81.73	67.56	18.68	8.37	84.96	70.97	22.84	14.02	3.1	3.05	2.37
13	Malta	83.2	0.79	68.88	2.35	13.16	3.98	10.31	81.27	66.86	18.26	9.83	85.26	71.26	23.35	13.61	3.99	4.4	2.86
14	Guernsey	83.17	0.64	68.91	3	14.01	3.91	9.82	81.83	66.93	19.63	9.35	85.38	71.1	21.77	13.68	4.36	4.58	3.79
15	Sweden	83.16	0.24	68.3	2.43	13.31	3.71	9.53	81.44	66.66	19.39	8.65	85.1	70.55	22.21	14.4	3.66	3.64	3.09