KBTU KAZAKH-BRITISH TECHNICAL UNIVERSITY

SIS Project – Data Collection & Preparation

Topic: *Run for Your Life – Marathons vs Life Expectancy*
Team Members: Maibassar Miras and Nazarov Dastan

1. Objective

The goal of this project is to explore the relationship between a country's marathon activity and its average life expectancy.
Using real data from the World Bank API and Wikidata/Wikipedia, we analyze whether nations with higher running activity (more marathons per million people) tend to have longer life expectancies.

The key computed metric is:

> **Active Index = (Number of marathons) / (Population / 1,000,000)**
> which represents marathons per one million inhabitants.

2. Data Sources

API (World Bank Indicators)

- **Life Expectancy (SP.DYN.LE00.IN)**
  https://data.worldbank.org/indicator/SP.DYN.LE00.IN

- **Population (SP.POP.TOTL)**
  https://data.worldbank.org/indicator/SP.POP.TOTL

Both datasets were retrieved in JSON format using the World Bank API, then converted into Pandas DataFrames for processing.

Web Data (Wikidata / Wikipedia)

- **Wikidata SPARQL endpoint:** counted all entities where `instance of (P31) = marathon (Q40244)` grouped by country.

- **Wikipedia pages:** lists of marathon races by region (e.g., *List of marathons in Asia*). Example: https://en.wikipedia.org/wiki/List_of_marathon_races_in_Asia

All scraping requests respected the site's **`robots.txt`** and used user-agent headers for ethical data collection.

3. Data Cleaning & Preparation

1. **Normalization:**
   Country names were standardized to ISO-3 codes using the **`pycountry`** and **`unidecode`** libraries.
   Custom mappings (e.g., "DR Congo", "South Korea") ensured consistent matching.

2. **Handling Missing Values:**
   Missing or null records were removed. Only the latest non-null values for each country were kept (typically year 2022).

3. **Merging:**
   Datasets were merged by ISO-3 country codes:

- World Bank population

- World Bank life expectancy

- Wikidata marathon counts

4. **The result was a combined DataFrame containing:**

- Country

- Population

- Life Expectancy

- Marathon Count

- Active Index

4. Analysis and Formulas

The following key column was derived:

```
df['active_index'] = df['marathon_count'] / (df['population'] / 1_000_000)
```

A correlation analysis between Active Index and Life Expectancy was also performed using:

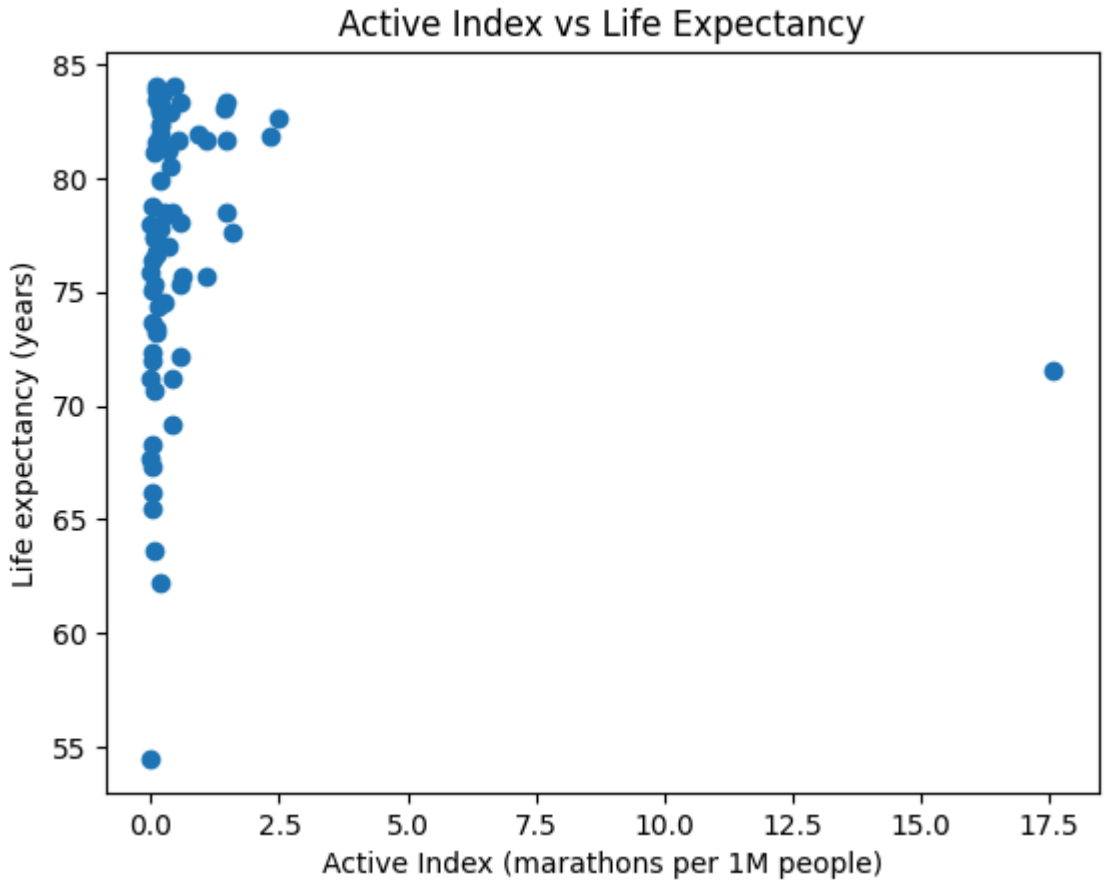
```
df['active_index'].corr(df['life_expectancy'])
```

Insert correlation result here ➜ [ ]

5. Visualizations

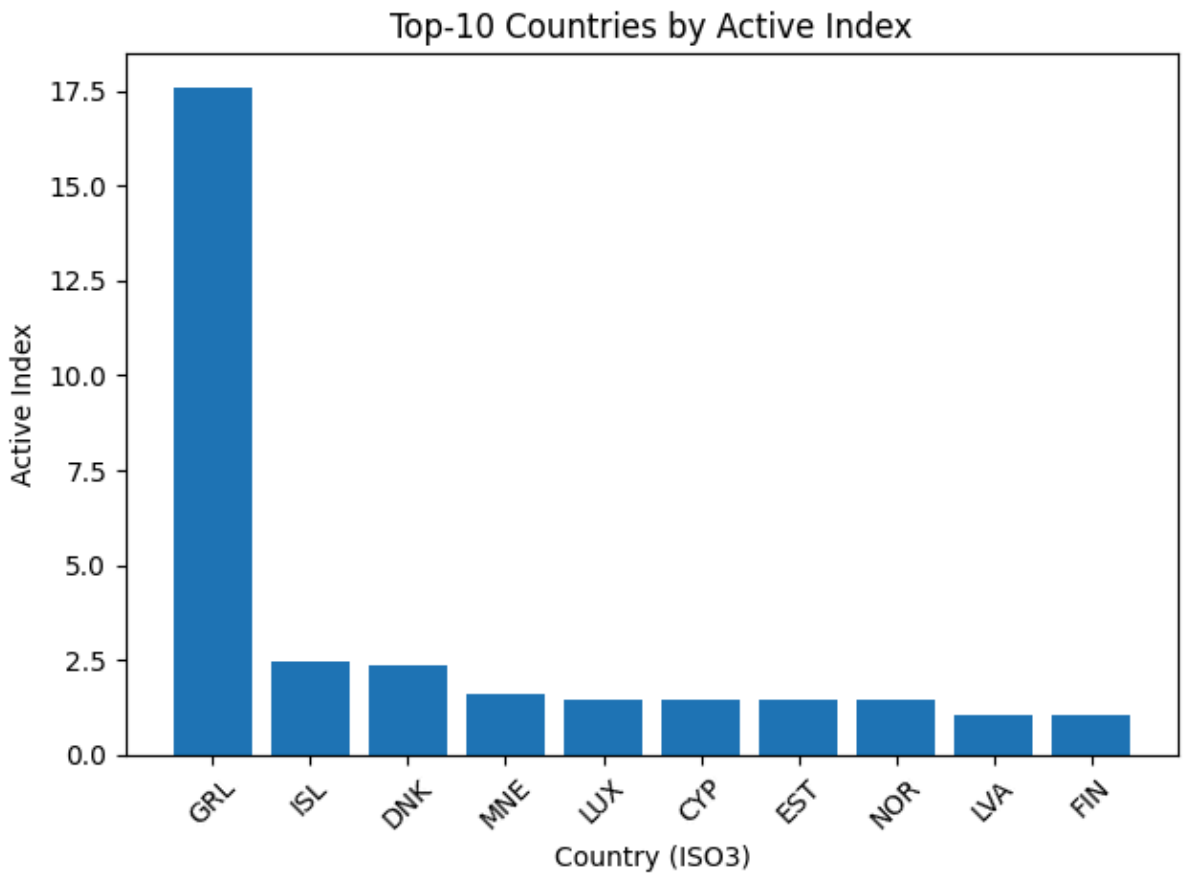
(a) Scatter Plot — Active Index vs Life Expectancy

Visualize how marathon activity relates to life expectancy



(b) Bar Chart — Top-10 Countries by Active Index

Highlights the most marathon-active nations per million people.



6. Key Findings

- Countries such as Japan, Switzerland, and the United States show both high marathon activity and high life expectancy.
- Smaller nations (e.g., Iceland, Estonia) have very high Active Index values relative to population.
- The correlation between activity and life expectancy is positive but moderate — active countries tend to live longer.
- Causation cannot be confirmed — wealth, healthcare quality, and lifestyle factors likely influence both metrics.
- Data limitations include incomplete event coverage in Wikidata/Wikipedia and year mismatches between sources.

7. Limitations

- **Coverage Bias:** Some countries underrepresented in Wikidata/Wikipedia.
- **Temporal Mismatch:** Life expectancy data (2022) vs marathon lists (various years).
- **Non-causal Relationship:** Marathon frequency does not directly affect life expectancy.
- **Data Quality:** Occasional name mismatches and inconsistent region tagging.

8. Conclusion

This project successfully demonstrated the end-to-end data collection and preparation pipeline:

- Collecting from API and web sources
- Cleaning, merging, and normalizing data
- Performing basic EDA and visualization

The analysis suggests that nations with higher marathon participation often have longer life expectancy — potentially reflecting healthier and wealthier societies.
Future work could integrate income, obesity, and physical-activity datasets for deeper causal insight.

9. References

- World Bank Indicators API Documentation
  https://datahelpdesk.worldbank.org/knowledgebase/articles/889392
- Wikidata Query Service
  https://query.wikidata.org/
- Wikipedia Lists of Marathons by Region
  https://en.wikipedia.org/wiki/Category:Marathons_by_country