

Отчёт по Проекту: Беги Ради Жизни - Марафоны и Продолжительность Жизни

Команда: Майбасар Мирас и Назаров Дастан

1. Цель и Как Мы Это Сделали

1.1. Главная Идея

Нам нужно было выяснить, есть ли какая-то связь между тем, насколько страна "спортивна" (по количеству марафонов), и средней продолжительностью жизни её граждан .

Мы ввели **Активный Индекс**. Это количество марафонов, делённое на население страны в миллионах.

1.2. Источники Данных

Использовали два источника, как и требовалось:

- 1. World Bank API2: Взяли самые свежие данные (2023–2024 гг.) по **Населению** и **Продолжительности Жизни**.
- 2. Wikidata SPARQL3: Отсюда взяли **Количество Марафонов**, чтобы посчитать нашу "спортивность".

1.3. Сборка Данных

Самым сложным было собрать все данные вместе:

- 1. **Чистка World Bank:** Убрали всякие агрегаты типа "Низкий доход" или "Восточная Африка", оставили только реальные страны.
- 2. **Нормализация (ISO3):** Все названия стран (из World Bank и Wikidata) перевели в единый трёхбуквенный код (ISO3), чтобы они точно совпали. Некоторые названия пришлось исправлять вручную (например, "United States" на "USA").
- 3. **Слияние:** Объединили три таблицы по этому коду ISO3. В итоге получили чистый набор данных по **76 странам**.

2. Результаты и Главные Выводы

2.1. Статистика (EDA)

Посчитали, насколько наш **Активный Индекс** связан с **Продолжительностью Жизни**:

Корреляция Пирсона (r)	\$1.0000\$	\$-0.0212\$
---------------------------	------------	-------------

Вывод: Коэффициент корреляции $r = -0.0212$ очень близок к нулю. Это значит, что линейной связи между "марафонской активностью" страны и долголетием нет.

2.2. Наблюдения по Графикам

1. **Нет Тренда (Figure 1):** На графике рассеяния (с точками) линия тренда почти плоская. Если бы связь была, она бы шла заметно вверх.
Проблема малых стран (Figure 2, 3): Страны, которые лидируют по Активному Индексу, например, Гренландия (GRL) и Исландия (ISL), просто очень маленькие. Один-два марафона дают им очень высокий индекс, что видно по резким скачкам на столбчатом графике.
2. **Аутлаеры:** Гренландия (GRL) — это яркий пример того, почему наша гипотеза не подтверждается. У неё самый высокий индекс (примерно 17.6), но продолжительность жизни низкая для развитого региона (примерно 71.5 года). На пузырьковом графике (Figure 3), где размер точки показывает население, видно, что богатые страны живут дольше (примерно 80-84 года) независимо от количества марафонов.

Следовательно, на продолжительность жизни, скорее всего, влияют деньги, медицина и уровень жизни, а не только бег. Наш индекс на их фоне не показывает значимого результата.

3. Ограничения Проекта

1. **Корреляция не равно Причинность:** Мы нашли отсутствие связи, но даже если бы нашли сильную связь, это не доказывало бы, что марафоны делают людей здоровыми. Могут быть другие, скрытые факторы (доход, климат).
2. **Временной Пробел:** Данные о марафонах в Wikidata — это общий подсчёт за всё время, а данные о населении и LE — это "снимки" за 2023–2024 гг.
3. **Смещение Данных:** Мы использовали Wikidata, который наполняют пользователи. Это значит, что в странах с меньшим количеством активных пользователей, марафоны могли быть просто не учтены.