CSC5003 - Project

Data Exploration and Analysis

Charles Meyer

TELECOM SudParis

IP PARIS

Table of contents

Project motivation and objective

A topic of your choice! ➡ Running

Lot of data in endurance sports with connected watchs

Objective?

- **Strava** is an American internet service for tracking physical exercise which incorporates social network features.

TELECOM SudParis

IP PARIS

Dataset

	datetime	athlete	distance	duration	gender	age_group	country	major
0	2019-01-01	0	0.00	0.000000	F	18 - 34	United States	CHICAGO 2019
1	2019-01-01	1	5.27	30.200000	M	35 - 54	Germany	BERLIN 2016
2	2019-01-01	2	0.00	0.000000	M	35 - 54	United Kingdom	LONDON 2018,LONDON 2019
3	2019-01-01	3	10.50	43.950000	M	18 - 34	United Kingdom	LONDON 2017
4	2019-01-01	4	9.66	48.650000	M	35 - 54	United States	BOSTON 2017
...	...	...	...	...	...	...	...	...
13290375	2019-12-31	37594	4.39	22.083333	M	18 - 34	United Kingdom	BERLIN 2017
13290376	2019-12-31	37595	23.31	88.266667	M	18 - 34	United States	BERLIN 2019,NEW YORK 2015
13290377	2019-12-31	37596	0.00	0.000000	M	18 - 34	United States	BOSTON 2017
13290378	2019-12-31	37597	6.49	33.733333	F	18 - 34	United States	BOSTON 2015
13290379	2019-12-31	37598	25.67	223.000000	M	35 - 54	China	TOKYO 2012

13290380 rows × 8 columns

- **datetime**: date of the running activity
- **athlete**: a computer-generated ID for the athlete
- **distance**: distance of running (kilometers)
- **duration**: duration of running (minutes)
- **gender**: gender ('M' of 'F');
- **age_group**: age interval ('18 - 34', '35 - 54', or '55 +')
- **country**: country of origin of the athlete
- **major**: marathon(s) and year(s) the athlete ran

TELECOM SudParis

IP PARIS

Dataset

Two other datasets

Running world records Dataset

	event	world_record	athlete	athlete_age	country	venu_location	record_date	distance	duration	gender
0	100 Meters	9.58	Usain Bolt	22	Jamaica	Berlin, Germany	8/16/09	0.1000	0.159667	M
1	200 Meters	19.19	Usain Bolt	22	Jamaica	Berlin, Germany	8/20/09	0.2000	0.319833	M
2	400 Meters	43.03	Wayde Van Niekerk	24	South Africa	Rio de Janeiro, Brazil	8/14/16	0.4000	0.717167	M
3	800 Meters	01:40.9	David Rudisha	23	Kenya	London, Great Britain	8/9/12	0.8000	1.681667	M
4	1000 Meters	02:12.0	Noah Ngeny	20	Kenya	Rieti, Italy	9/5/99	1.0000	2.200000	M
5	1500 Meters	03:26.0	Hicham El Guerrouj	23	Morocco	Rome, Italy	7/14/98	1.5000	3.433333	M
6	2000 Meters	4:43.1	Jakob Ingebrigsten	22	Norway	Brussels, Belgium	8/9/23	2.0000	4.718333	M
7	3000 Meters	07:17.6	Jakob Ingebrigsten	24	Norway	Chorzów, Poland	25/8/24	3.0000	7.293333	M
8	5000 Meters	12:35.4	Joshua Cheptegei	23	Uganda	Monaco, Monaco	8/14/20	5.0000	12.590000	M
9	10 Kilometers	26:11.0	Joshua Cheptegei	24	Uganda	Valencia, Spain	10/7/20	10.0000	26.183333	M
10	Half Marathon	57:31.00	Jacob Kiplimo	21	Uganda	Lisbon, Portugal	11/21/21	21.0975	57.516667	M
11	Marathon	2:00:35	Kelvin Kiptum	24	Kenya	Chicago, IL, USA	8/10/23	42.1950	120.583333	M
12	50 Kilometers	2:38:43	CJ Albertson	29	USA	San Francisco, CA, USA	8/10/22	50.0000	158.716667	M
13	100 Kilometers	6:05:41	Aleksandr Sorokin	40	Lithuania	Bedford, United Kingdom	4/23/22	100.0000	365.683333	M
14	100 Meters	10.49	Florence Griffith-Joyner	28	USA	Indianapolis, IN, USA	7/16/88	0.1000	0.174833	F
15	200 Meters	21.34	Florence Griffith-Joyner	28	USA	Seoul, South Korea	9/29/88	0.2000	0.355667	F
16	400 Meters	47.6	Marita Koch	28	East Germany	Canberra, Australia	10/6/85	0.4000	0.793333	F
17	800 Meters	01:53.3	Jarmila Kratochvílová	32	Czechoslovakia	München, Germany	7/26/83	0.8000	1.888333	F

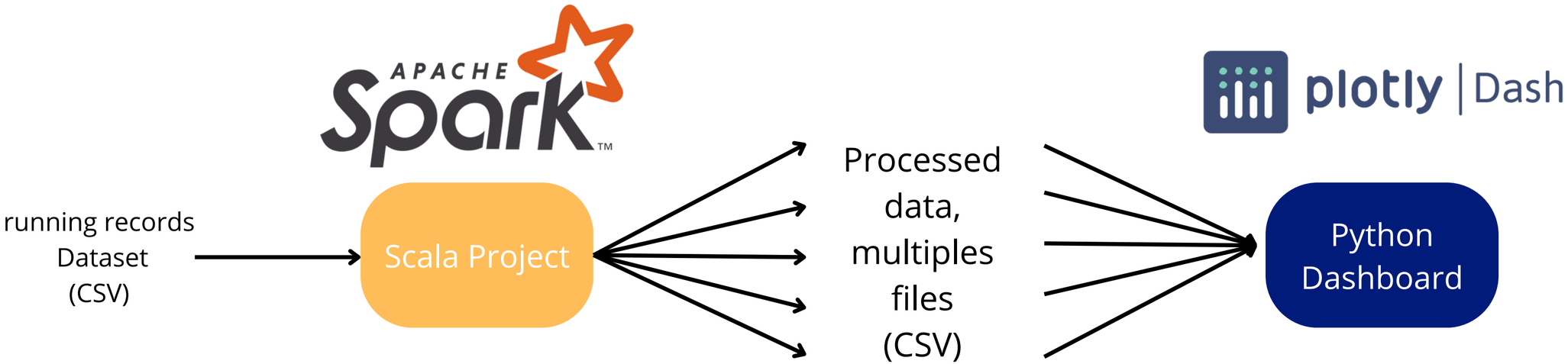
Population Dataset

	country_name	country_code	year	population
0	Aruba	ABW	2019	109203.0
1	Africa Eastern and Southern	AFE	2019	675950189.0
2	Afghanistan	AFG	2019	37856121.0
3	Africa Western and Central	AFW	2019	463365429.0
4	Angola	AGO	2019	32375632.0
...	...	...	...	...
525	Kosovo	XKX	2020	1790152.0
526	Yemen, Rep.	YEM	2020	36134863.0
527	South Africa	ZAF	2020	60562381.0
528	Zambia	ZMB	2020	19059395.0
529	Zimbabwe	ZWE	2020	15526888.0

TELECOM SudParis

IP PARIS

Architecture overview



Data processing

List of functions that perform different data processing

Example

```
def get_per_runner_metrics(df_runner : DataFrame) : DataFrame = {
    val nb_of_weeks = df_runner.select("datetime").distinct().count() / 7.0
    df_runner
    .filter(col("distance") =!= 0)
    .groupBy("athlete")
    .agg(
        count("*").as("nb_of_run"),
        round(sum("distance"), 2).as("total_dist"),
        round(sum("distance") / lit(nb_of_weeks), 2).as("avg_dist_per_week"),
        round(avg("distance"), 2).as("avg_run_dist")
    )
}
```

TELECOM SudParis

IP PARIS

Dashboard

Demo