Faculty of Sciences of the
University of Porto

Department of Computer Science

Advanced Topics in Databases

Practical Assignment

Andreia Genro Aníbal Silva Rui Fernandes

June, 2022

Abstract

This report describes the practical assignment of the Advanced Topics in Databases course.

This practical assignment consists in creating a data warehouse and conducting data analysis on it, as well as creating graphical reports using the Python library `matplotlib`.

In this report, we briefly describe our approach to the problem and discuss the decisions we made.

Contents

List of Figures

1 Introduction

The data warehouse contains data from national swimming competitions at the master level (*i.e.* class of competitive swimming for swimmers 25 years and older), namely Troféu Pescada 2021 and the Summer 2021 Championship.

Structure of the Report

The remainder of the report is structured as follows:

- In Section 3, **Data Analysis & Visualization**, we provide some insight into the data
- Finally, Section 4, **Conclusions & Future Work**, concludes the report and suggests remarks for future work.

2 Data Model

In this section, we describe the data contained in the data warehouse.

3 Data Analysis & Visualization

3.1 Number of Athletes by Age

To determine to determine the average age of the athletes, we can run the following SQL query:

```
SELECT AVG(age(birthdate))
FROM annp_final.athlete
```

From this, we can see that the average age of the athletes is 46 years, 6 months and 31 days.

We can also determine who's the youngest athlete by running the following SQL query:

```
SELECT *
FROM annp_final.athlete
ORDER BY age(birthdate) ASC
LIMIT 1;
```

- **Name:** Ana Mónica Eloi
- **Gender:** F
- **Birthdate:** 29/12/1996
- **Age:** 25 years

On the other hand, we can learn information about the oldest athlete by running the following SQL query:

```
SELECT *
FROM annp_final.athlete
ORDER BY age(birthdate) DESC
LIMIT 1;
```

- **Name:** Virgílio Zacarias Costa
- **Gender:** M
- **Birthdate:** 21/07/1931
- **Age:** 90 years

Finally, to determine the number of athletes by age, we can run the following SQL query using the PostgreSQL's built-in `age` function:

```
SELECT COUNT(*), EXTRACT(YEAR FROM age(birthdate)) AS age
FROM annp_final.athlete
GROUP BY age
ORDER BY age ASC;
```

We can then plot the result, as illustrated in Figure 2.

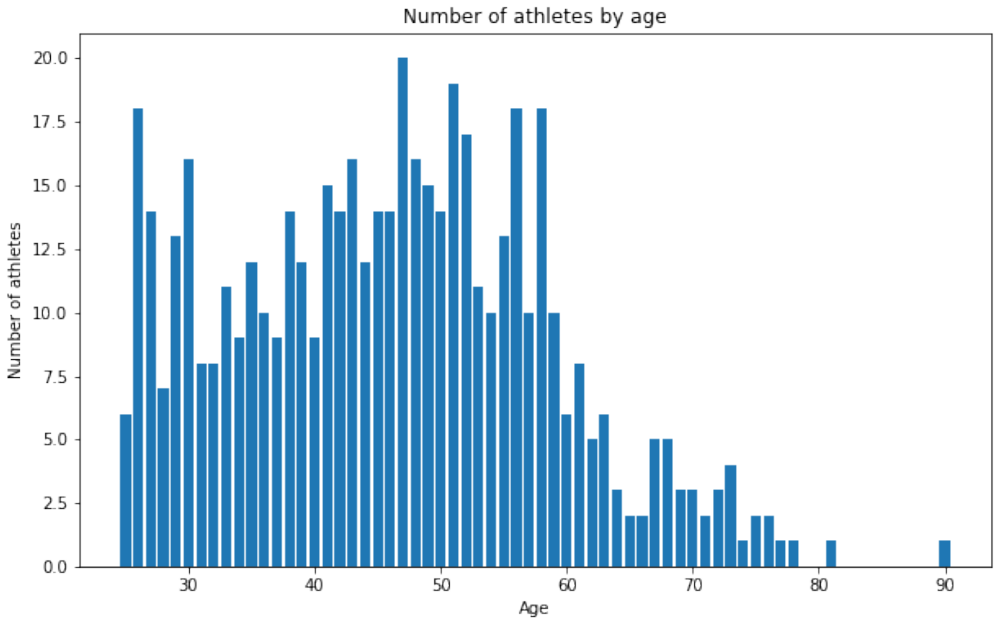

Figure 1: Number of athletes by age

3.2 Number of Athletes by Nation

To determine the number of athletes by nation, we can run the following SQL query:

```
SELECT COUNT(*) nationCount, nation
FROM annp_final.athlete
GROUP BY nation
ORDER BY nationCount ASC;
```

We can then plot the result, as illustrated in Figure 2.

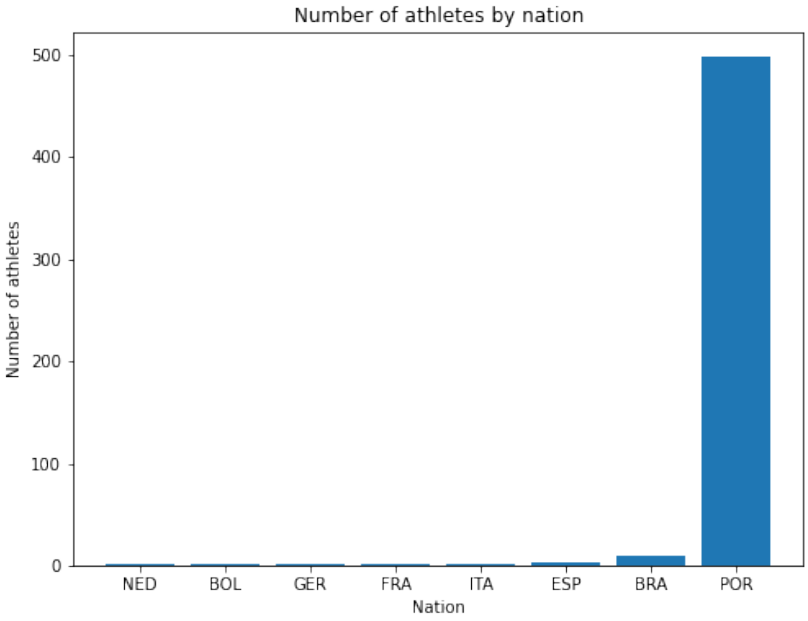

Figure 2: Number of athletes by nation

To have another perspective, we can also plot in a pie chart, as illustrated in Figure 3.

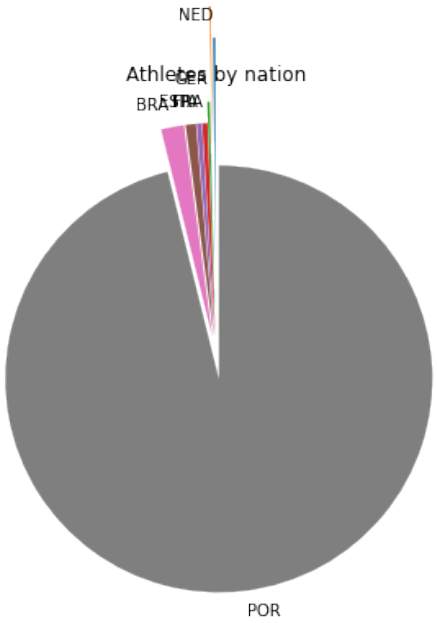

Figure 3: Number of athletes by nation

3.3 Number of Athletes by Gender

To determine the number of athletes by gender, we can run the following SQL query:

```
SELECT COUNT(*), gender
FROM annp_final.athlete
GROUP BY gender;
```

We can then plot the result, as illustrated in Figure 4.

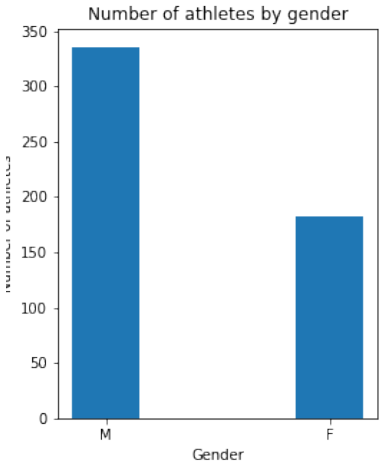

Figure 4: Number of athletes by gender

We can also plot this in a pie chart, as illustrated in Figure 5.

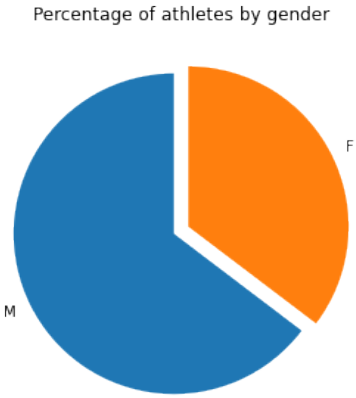

Figure 5: Percentage of athletes by gender

3.4 Number of Events by Gender

To determine the number of events by gender, we can run the following SQL query:

```
SELECT COUNT(*), gender
FROM annp_final.event
GROUP BY gender;
```

We can then plot the result, as illustrated in Figure 6.

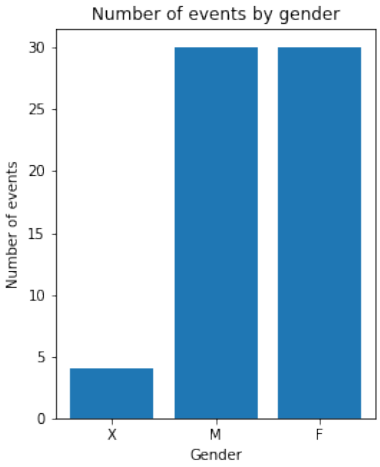

Figure 6: Number of events by gender

Here, the value `X` refers to events that allow athletes from both genders to participate.

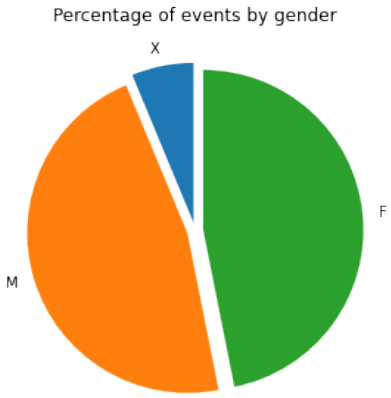

Figure 7: Percentage of events by gender

We can also plot this in a pie chart, as illustrated in **??**.

4 Conclusions & Future Work