

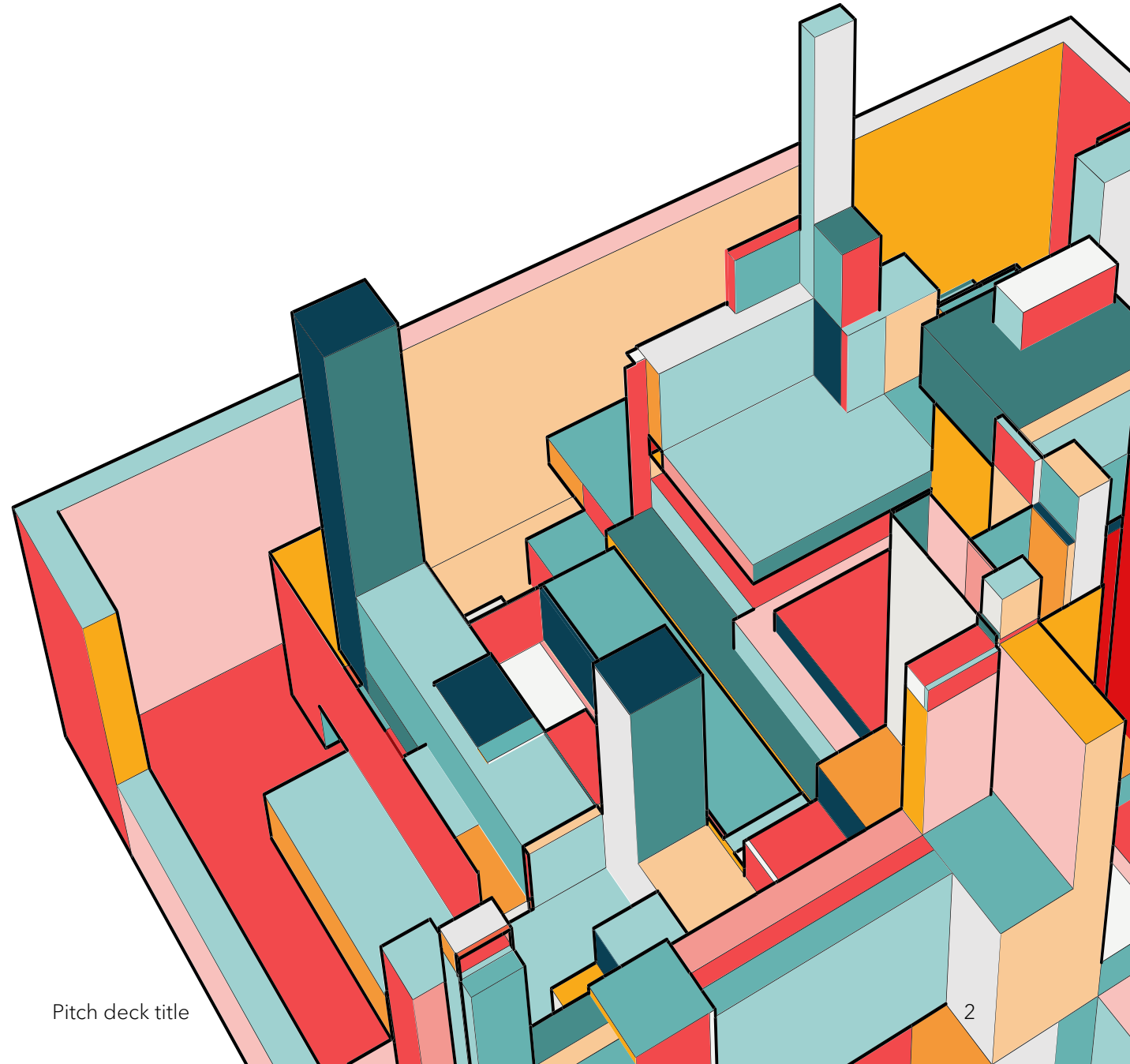
An abstract geometric composition featuring various 3D rectangular blocks and rectangles in shades of teal, orange, red, and pink. The blocks are arranged in a layered, isometric fashion, creating a sense of depth and architectural structure. The background is a solid light teal color.

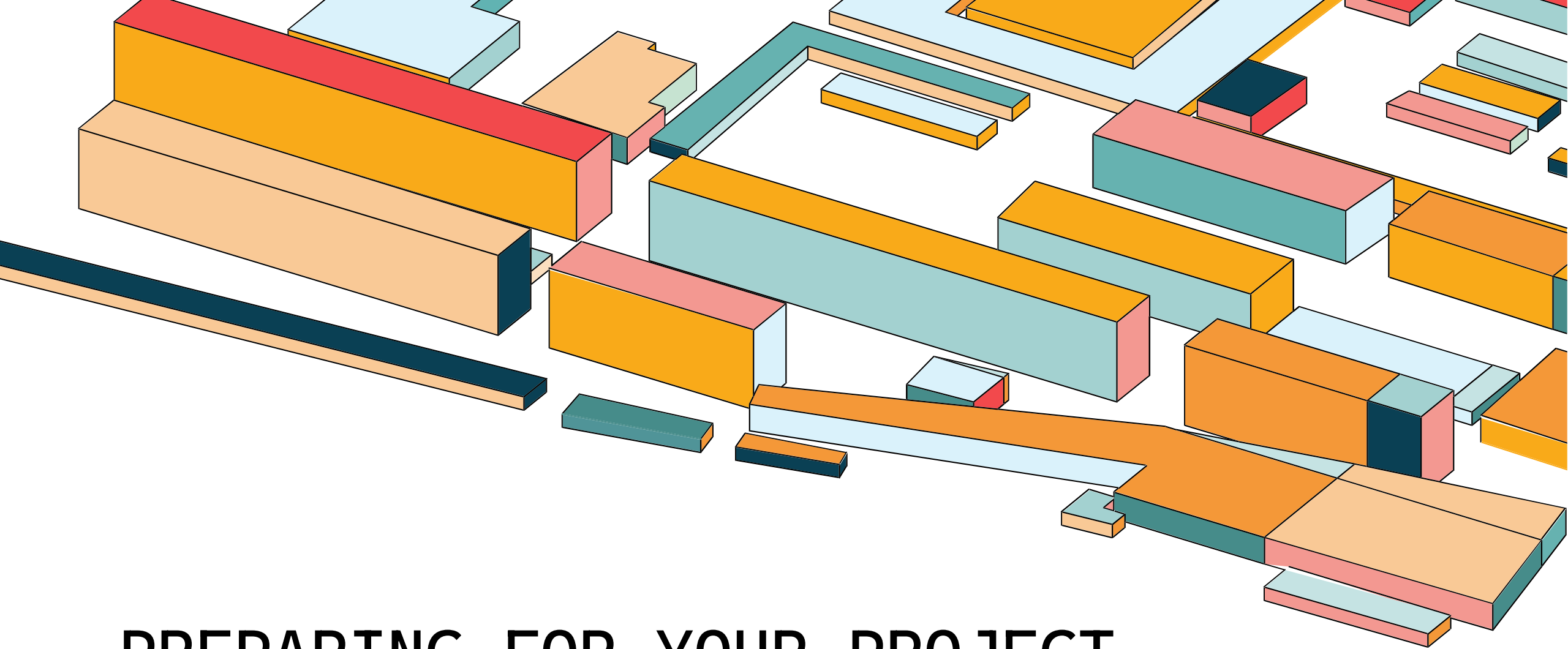
CAPSTONE PROJECT

SABA HAJEBI

TABLE OF CONTENTS

1. PREPARING YOUR PROPOSAL
2. DEVELOPING PROJECT PROPOSAL





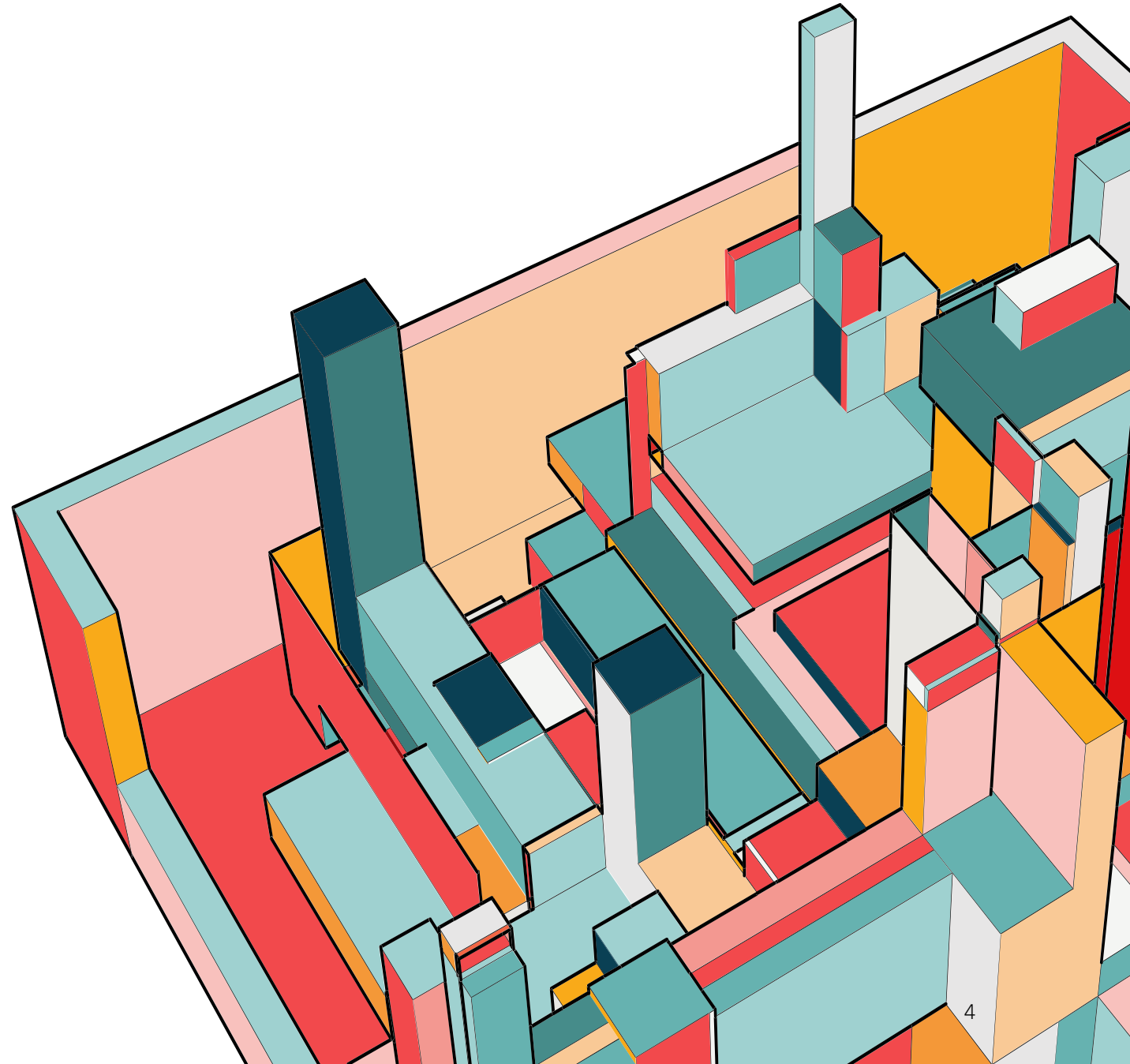
**PREPARING FOR YOUR PROJECT
PROPOSAL**

WHICH CLIENT DID YOU SELECT AND WHY?

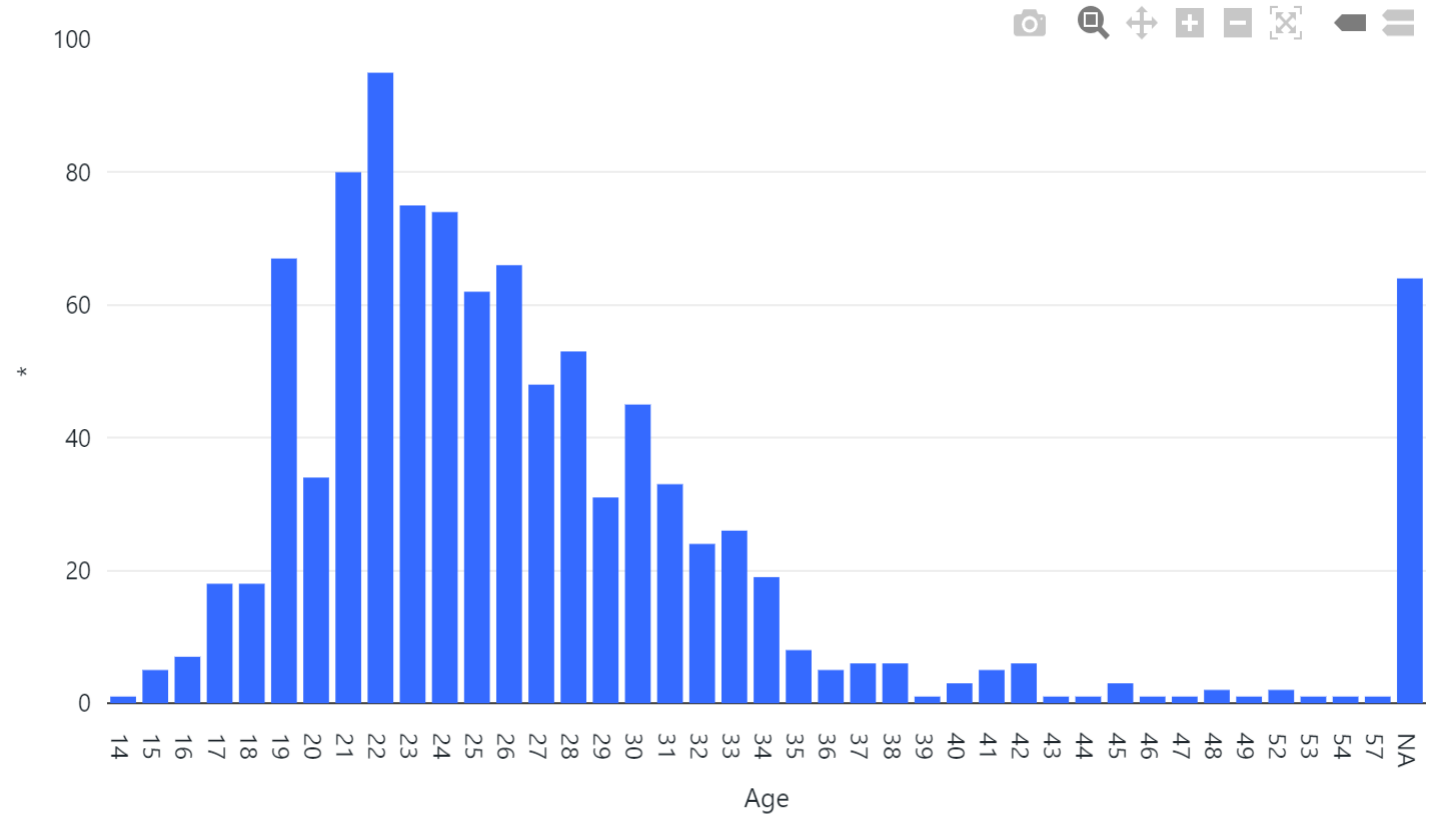
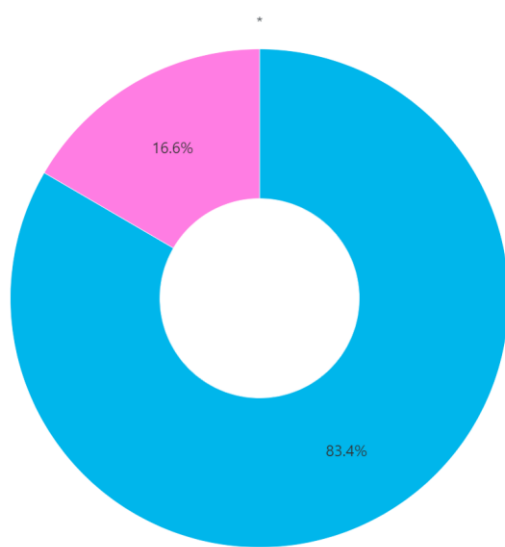
- I CHOSE THE SPORTS STATS CLIENT SINCE I HAVE A STRONG INTEREST IN SPORTS AND FITNESS AND WANTED TO DEVELOP MEANINGFUL INSIGHTS AND PATTERNS ABOUT DIFFERENT SPORTS.

DESCRIBE THE STEPS YOU TOOK TO IMPORT AND CLEAN THE DATA.

- I IMPORTED THE DATA USING DATABRICKS BY FIRST ATTACHING TO MY CLUSTER, THEN MOUNTING THE DATA AND CREATING A DATABASE AND TABLE .
- SINCE THE DATASET HAS NAN VALUES, I DID NOT TAKE ANY STEPS TO CLEAN THE DATA AS THAT WOULD BE FALSIFICATION OF DATA.



PERFORM INITIAL EXPLORATION OF DATA AND DISPLAY SOME STATS OF DATA



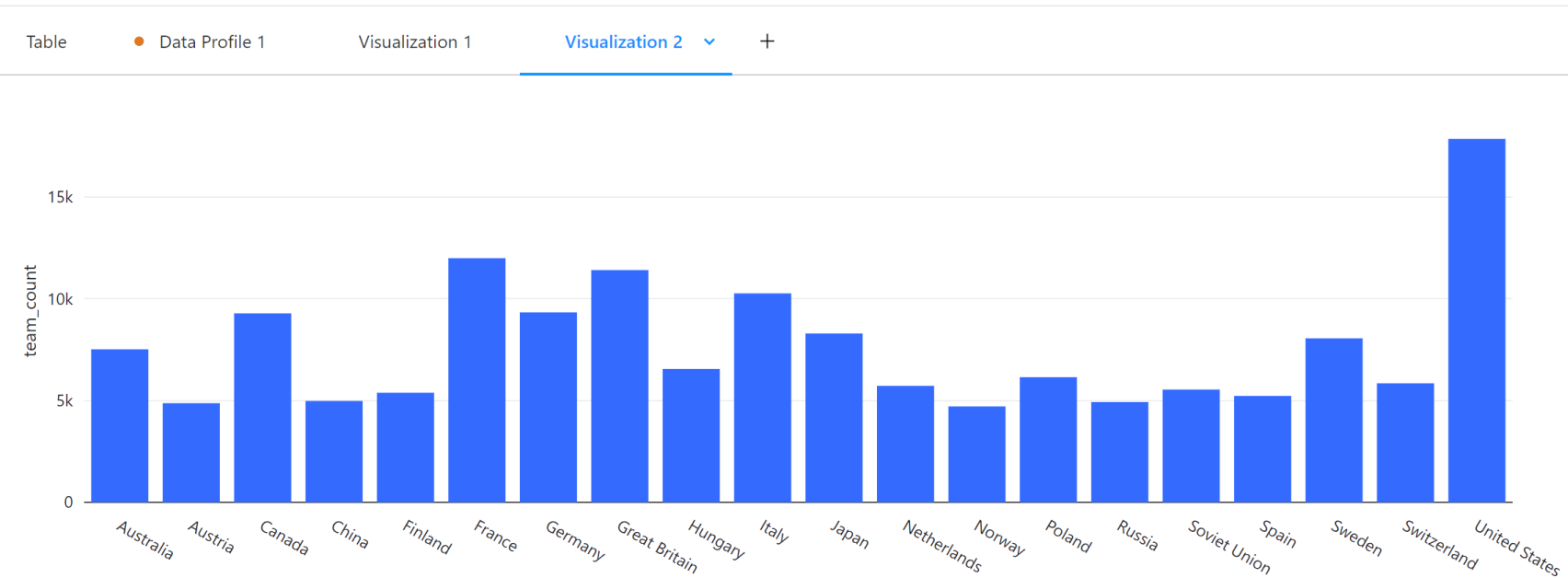
PERFORM INITIAL EXPLORATION OF DATA AND DISPLAY SOME STATS OF DATA

Cmd 1

```
1 SELECT team,  
2     Count(team) AS team_count  
3 FROM athlete_events_3_csv  
4 GROUP BY team  
5 ORDER BY Count(team) DESC  
6 LIMIT 20
```

SQL ▶ 📊 ⌵ - ✕

▶ (2) Spark Jobs

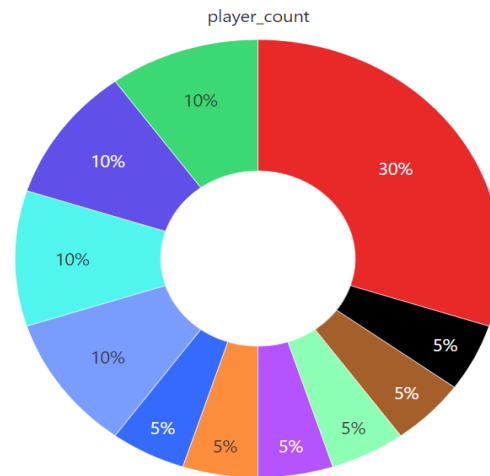


PERFORM INITIAL EXPLORATION OF DATA AND DISPLAY SOME STATS OF DATA

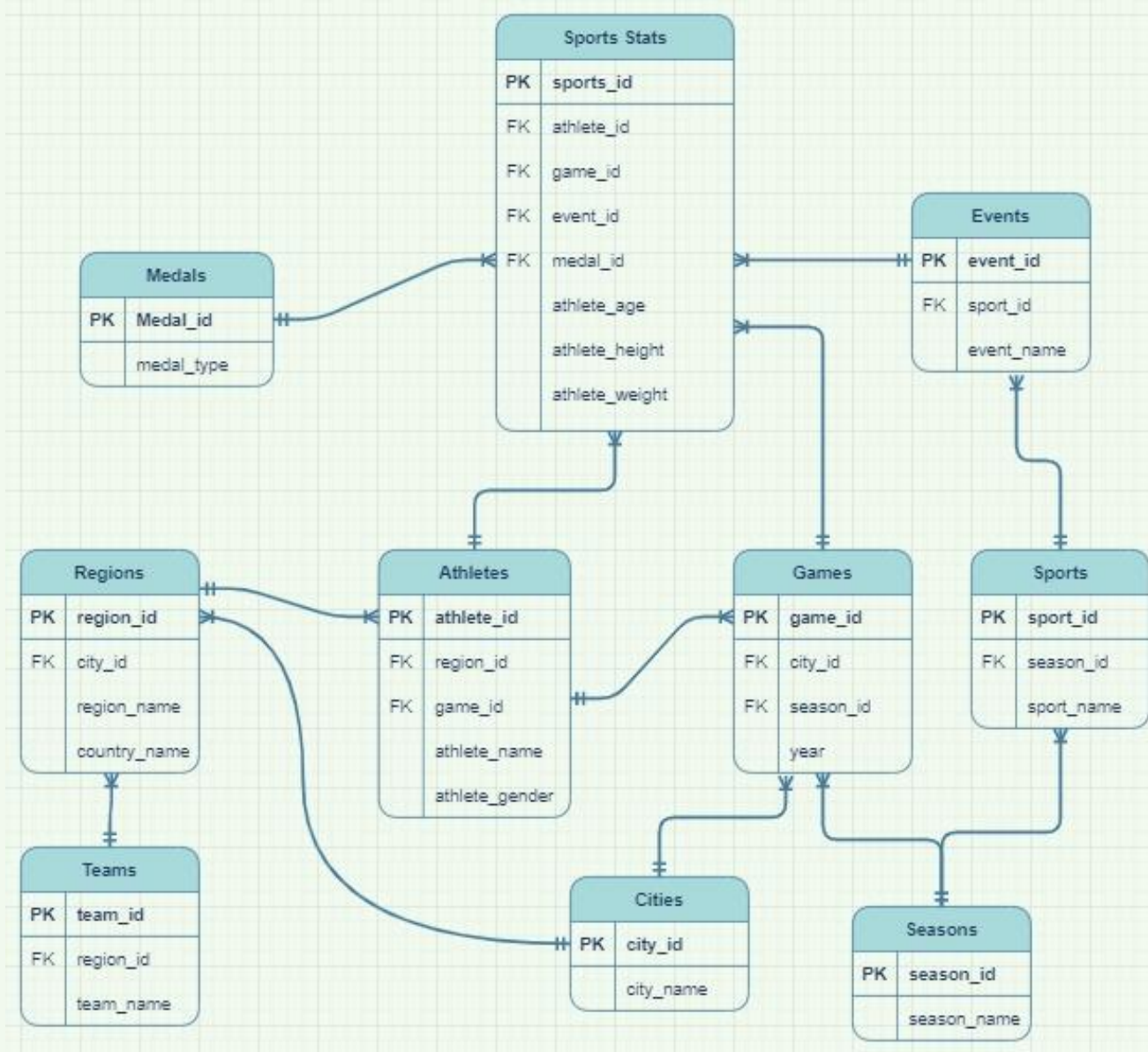
```
SELECT n.region,  
       (SELECT COUNT(id)  
        FROM athlete_events_3_csv) AS player_count  
FROM athlete_events_3_csv AS a  
     INNER JOIN noc_regions_csv AS n  
           ON n.noc = a.noc  
ORDER BY player_count DESC  
LIMIT 20
```

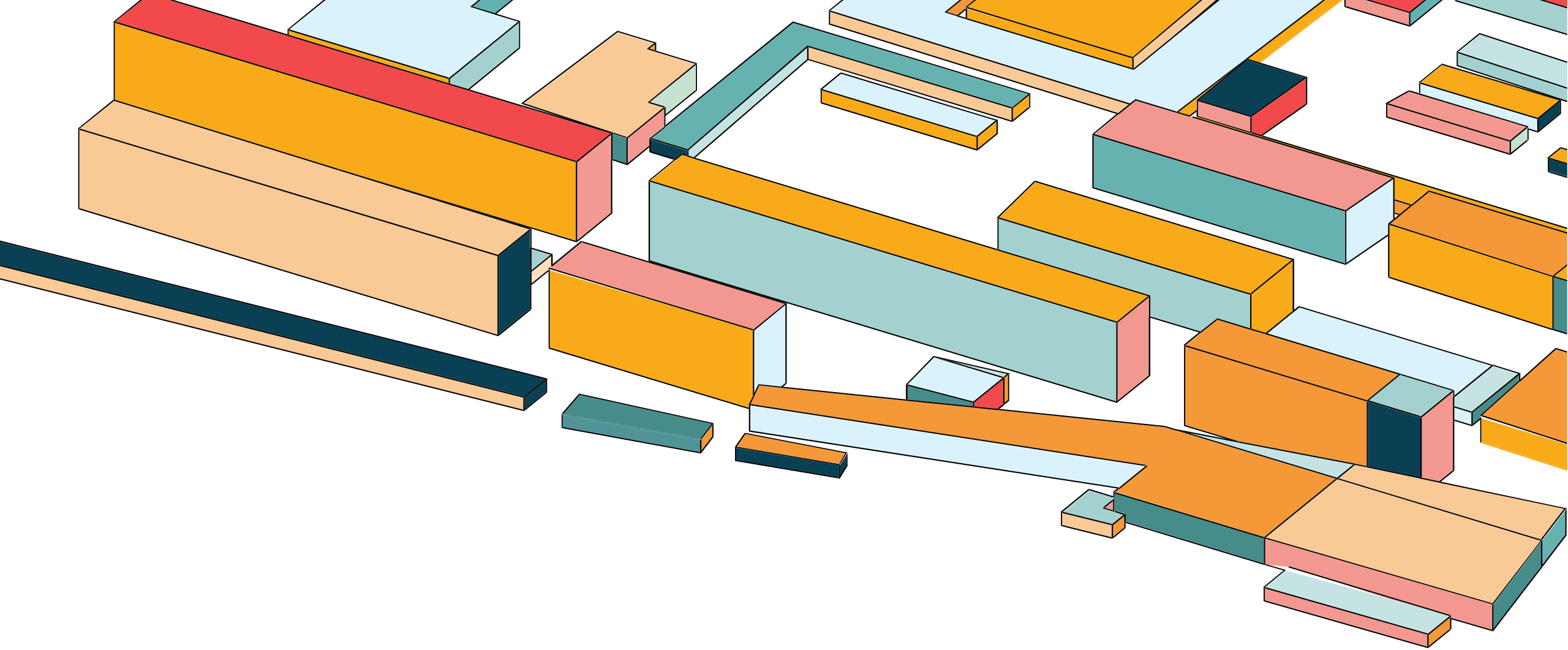
(4) Spark Jobs

Table ● Data Profile 1 Visualization 1 Visualization 2 **Visualization 3** +



- Russia
- China
- Sweden
- Zambia
- Denmark
- USA
- Romania
- Germany
- Greece
- Netherlands
- Indonesia





DEVELOP PROJECT PROPOSAL

DESCRIPTION

MY PROJECT AIMS TO DRAW CONCLUSIONS BASED ON PATTERNS FOUND IN PAST SPORT DATA.

THE ANALYSIS TARGETS FITNESS-ORIENTATED FIRMS IN HOPES OF HELPING THEM UNDERSTAND KEY INSIGHTS ABOUT HOW TO ATTRACT AND MAINTAIN THEIR CLIENT BASE

MY AUDIENCE FOR THIS PROJECT WOULD NOT BE LIMITED TO FITNESS ORGANISATIONS, BUT ALSO SPORTS PLAYERS AND FITNESS COUCHES AS WELL AS ORDINARY PUBLIC USE.


QUESTION

- IS THERE A CORRELATION BETWEEN AGE AND AWARDS WON?
- WHICH COUNTRY HAS WON THE MOST MEDALS?
- IS THERE A CORRELATION BETWEEN GENDER AND SPORT?

HYPOTHESIS

- THE OLDER THEY PLAYER, THE HIGHER THE PROBABILITY THAT THEY HAVE A MEDAL.
- THE COUNTRY WITH THE MOST PLAYERS HAS THE MOST MEDALS.
- THERE ARE MORE MEN THAN WOMEN WHO PARTICIPATED IN SPORTS.

APPROACH

- IS THERE A CORRELATION BETWEEN AGE AND AWARDS WON?
 - I WILL PLOT THE DISTRUBUTION OF AGE AGAINST THE DISTRUBTION OF MEDALS WON TO SEE IF THERE IS A CORRELATION
 - IS THERE A CORRELATION BETWEEN THE COUNTRY WITH THE MOST PLAYERS AND THE COUNTRY THATS HAS WON THE MOST MEDALS?
 - I WILL CATEGORISE THE COUNTRIES IN ORDER OF MOST MEDALS WON AND THE COMPARE THAT AGAINST A PLOT OF WHICH CONTRIES HAVE THE MOST PLAYERS
 - IS THERE A CORRELATION BETWEEN GENDER AND SPORT?
 - I WILL PLOT THE DISTRIBUTION OF GENDER AGAINST THE DSITRUBTION OF SPORT TO DRAW INSIGHTS ABOUT THE RELATIONSHIP
- 

THANK YOU