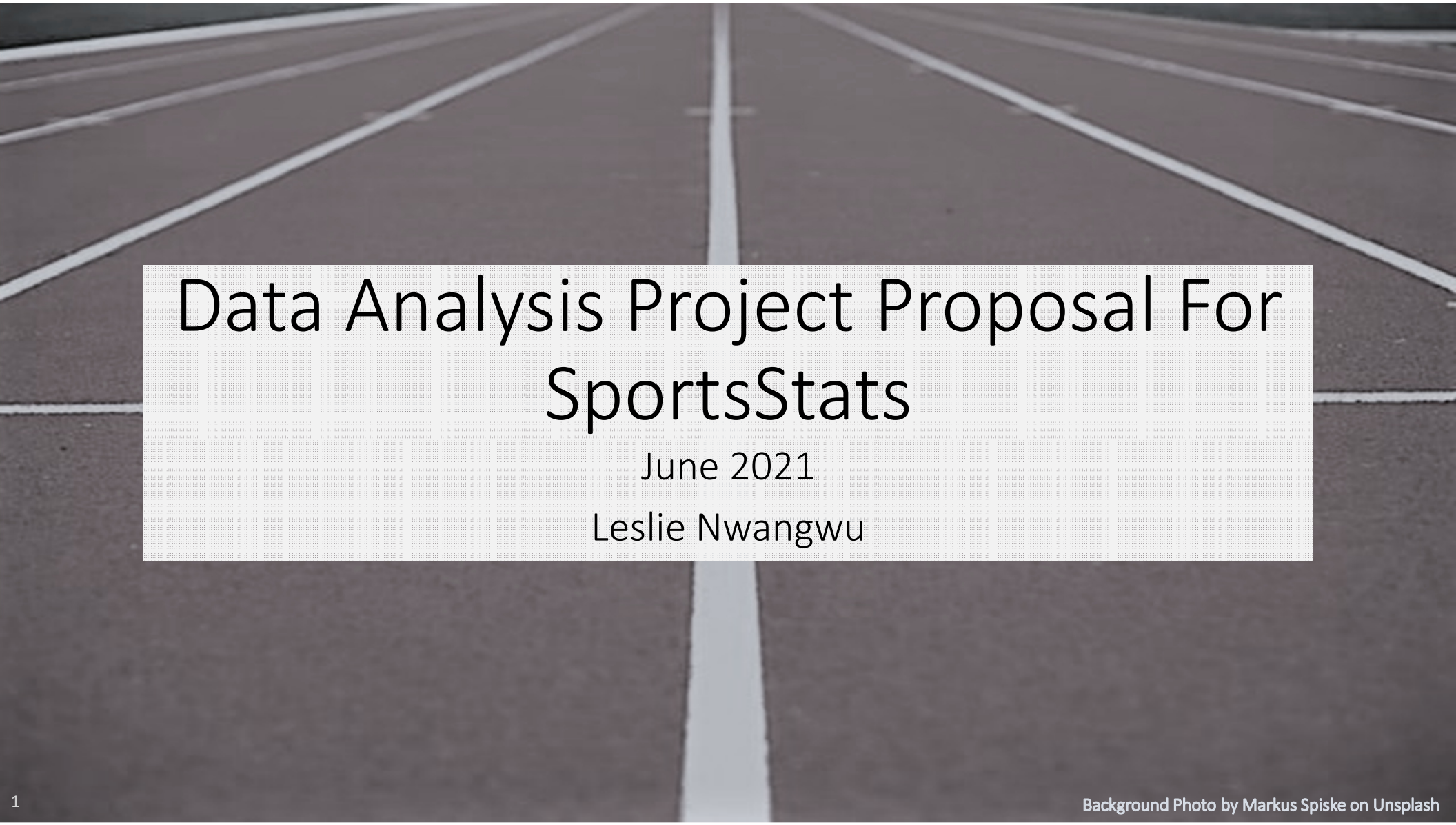
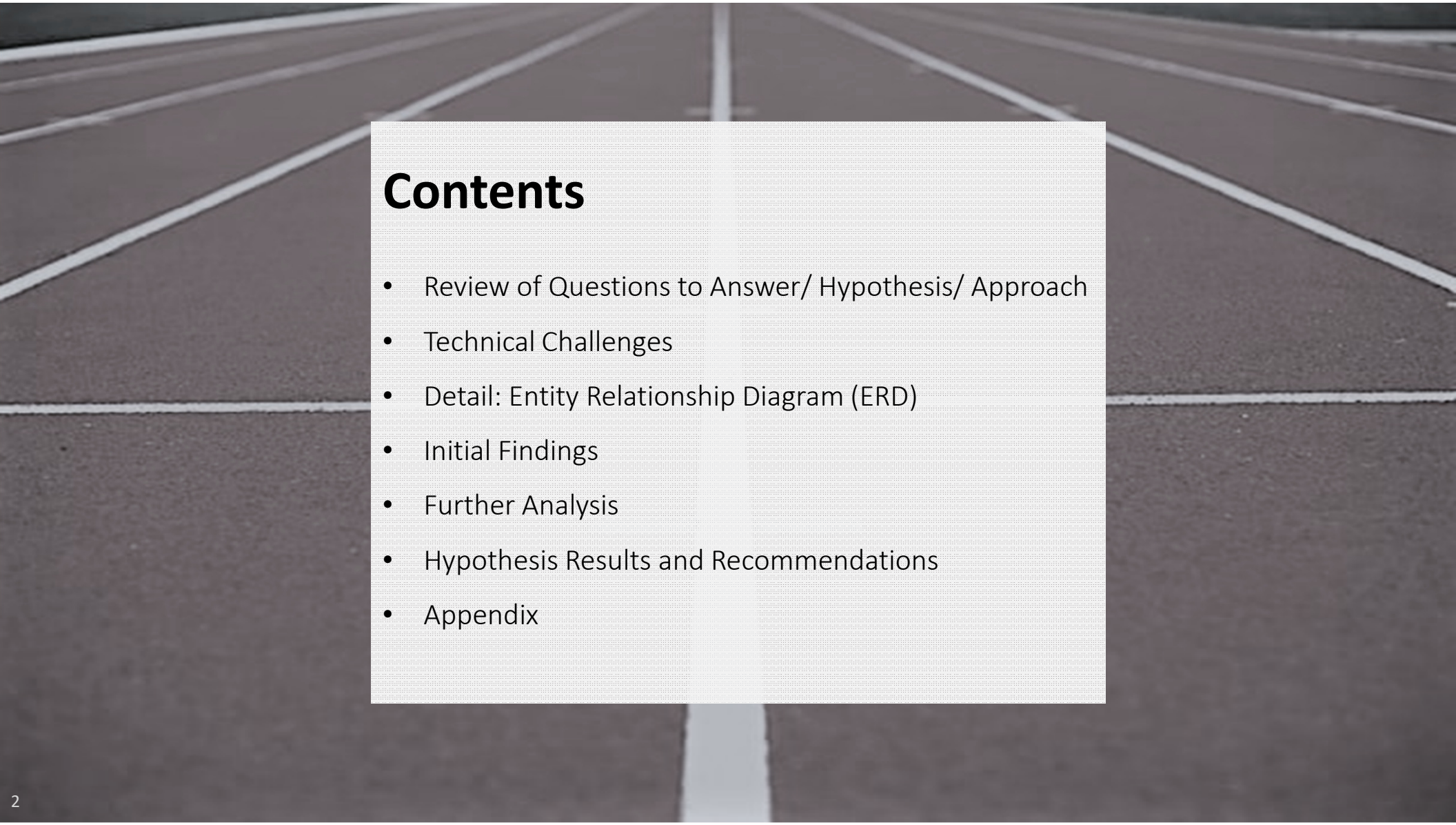


Data Analysis Project Proposal For SportsStats

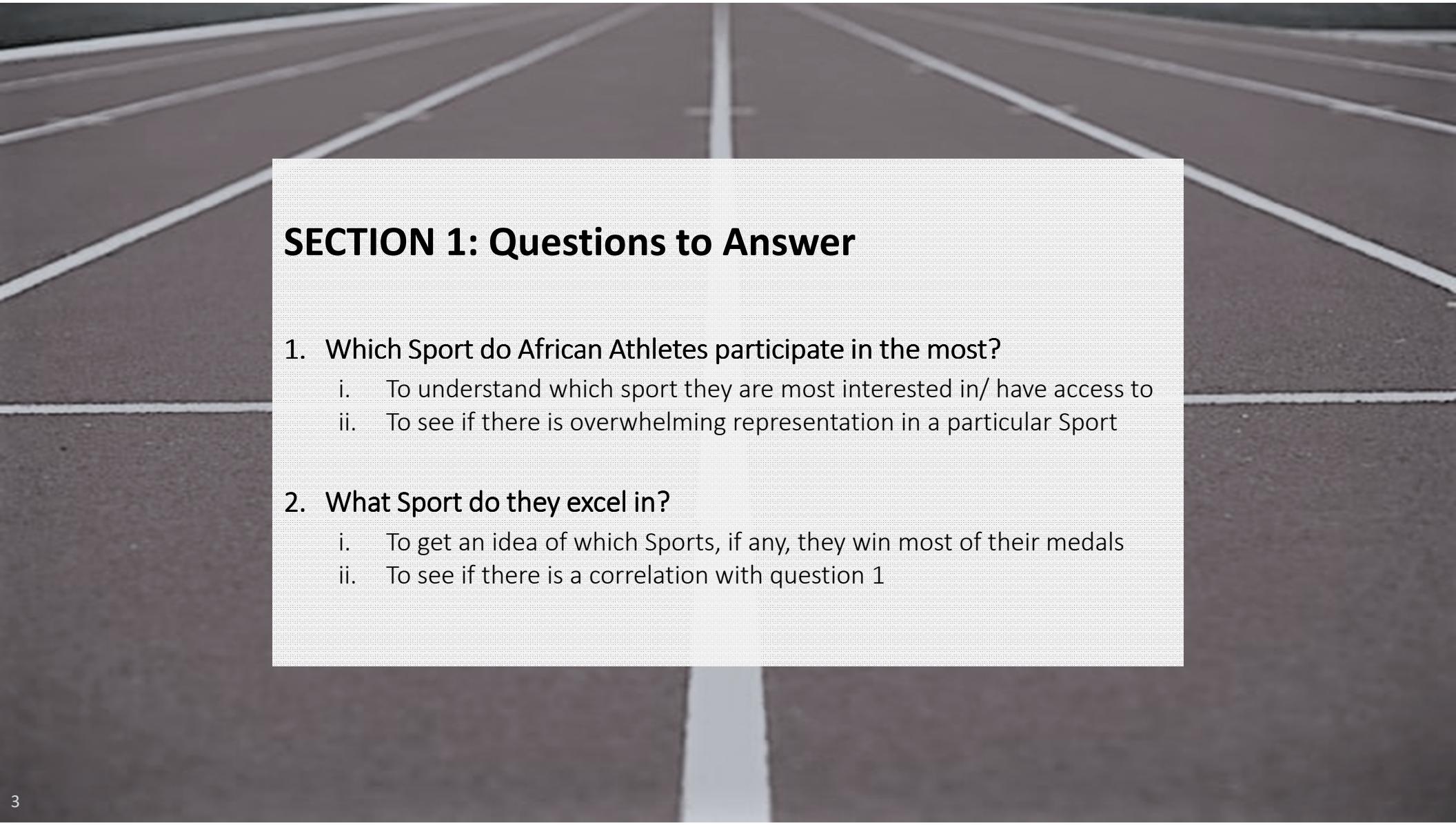
June 2021

Leslie Nwangwu



Contents

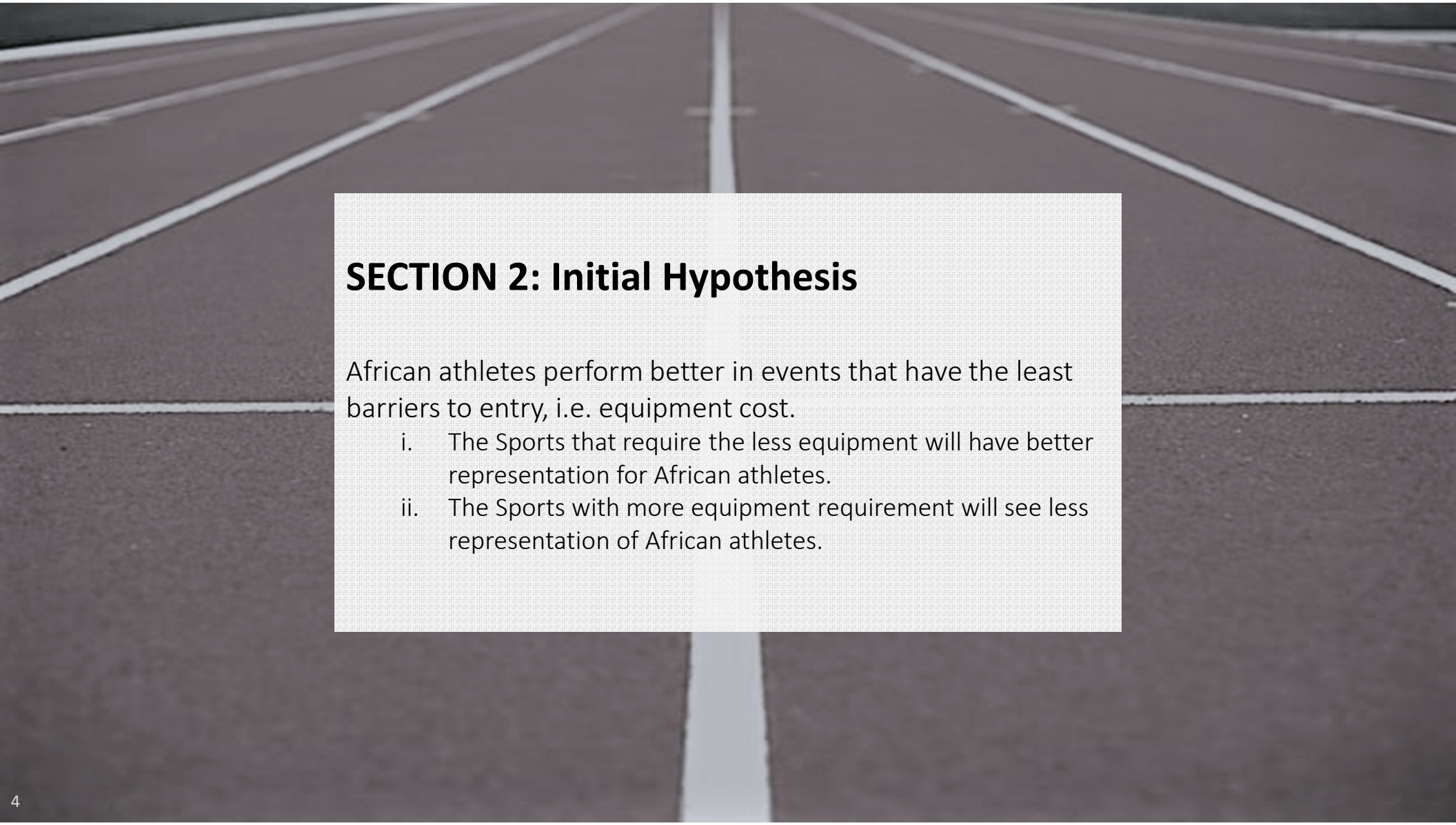
- Review of Questions to Answer/ Hypothesis/ Approach
- Technical Challenges
- Detail: Entity Relationship Diagram (ERD)
- Initial Findings
- Further Analysis
- Hypothesis Results and Recommendations
- Appendix



SECTION 1: Questions to Answer

1. Which Sport do African Athletes participate in the most?
    i. To understand which sport they are most interested in/ have access to
    ii. To see if there is overwhelming representation in a particular Sport

2. What Sport do they excel in?
    i. To get an idea of which Sports, if any, they win most of their medals
    ii. To see if there is a correlation with question 1



SECTION 2: Initial Hypothesis

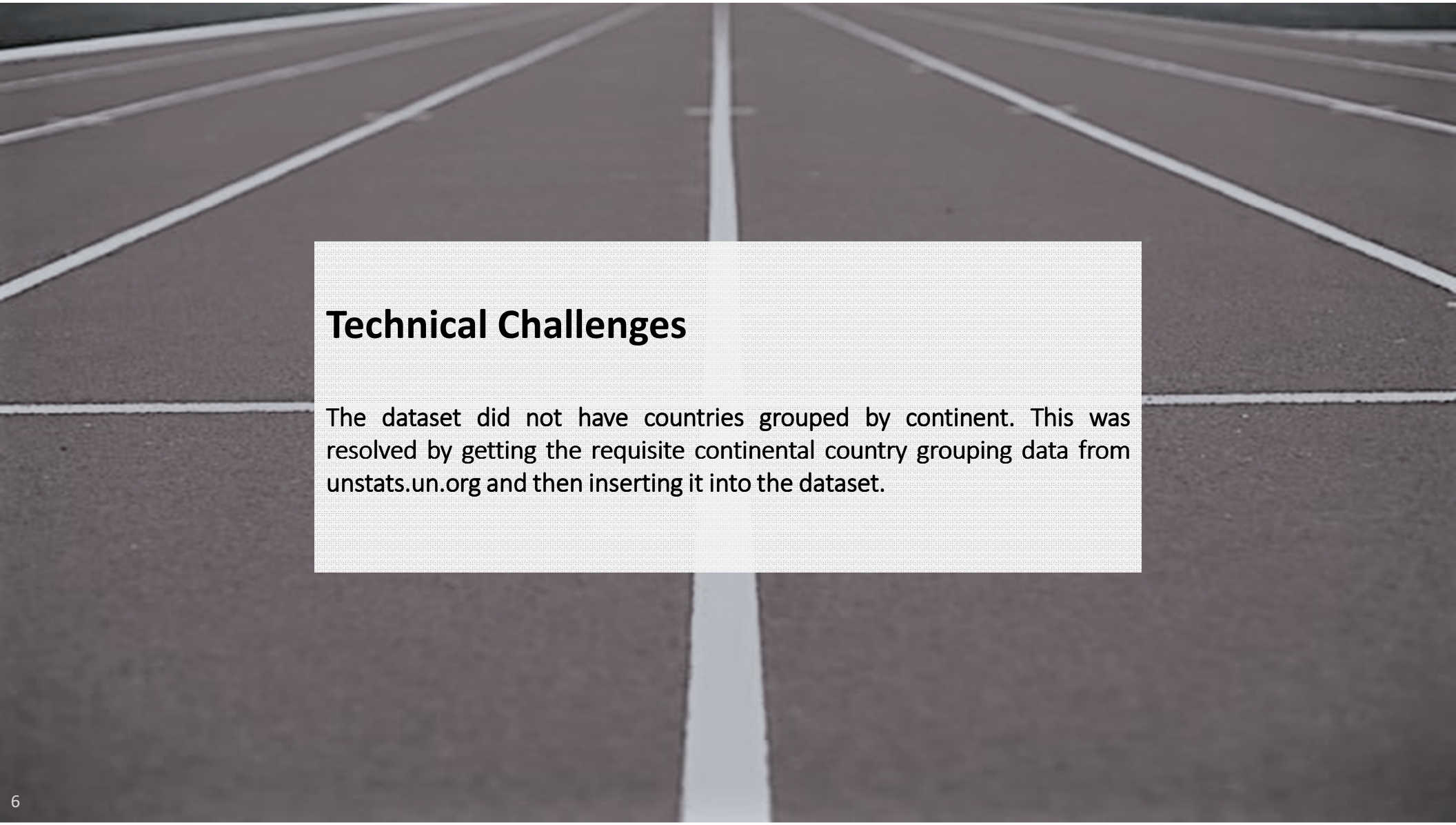
African athletes perform better in events that have the least barriers to entry, i.e. equipment cost.

i. The Sports that require the less equipment will have better representation for African athletes.
ii. The Sports with more equipment requirement will see less representation of African athletes.

SECTION 3: Data Analysis Approach

1. Aggregate functions.
   i. To determine which sports African athletes won medals in grouped by Country.
   ii. To see if more medals have been won in any particular Sport than others.
   iii. To group the athletes participating in the games by Sport and see if any Sport is heavily represented.

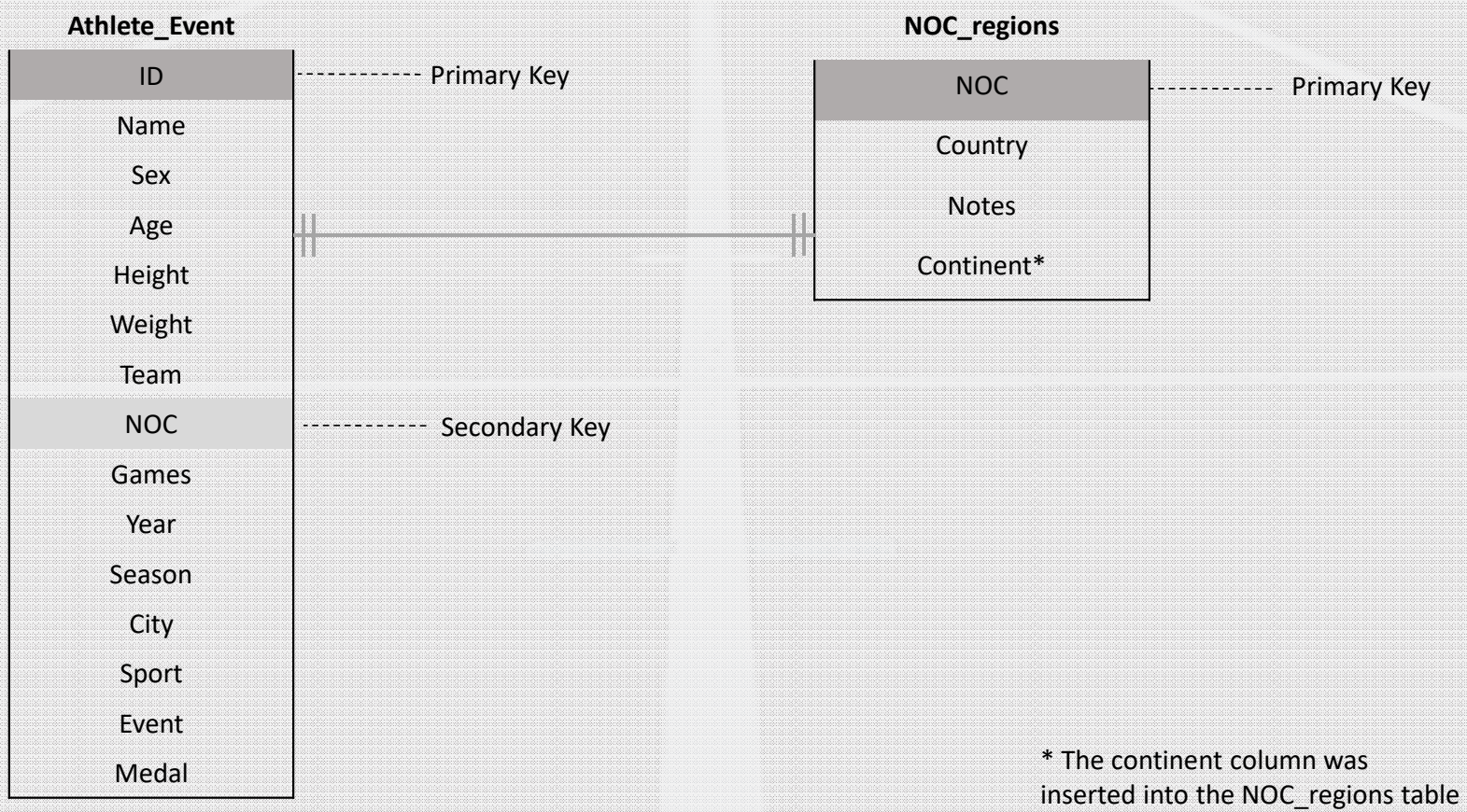
2. Comparative analysis.
   i. To see if the results of the analysis are unique to African contingents.
   ii. Determine if any other pattern emerges.



Technical Challenges

The dataset did not have countries grouped by continent. This was resolved by getting the requisite continental country grouping data from unstats.un.org and then inserting it into the dataset.

Entity Relationship Diagram



Initial Findings

Athletics has the highest number of athletes of all sports on all continents

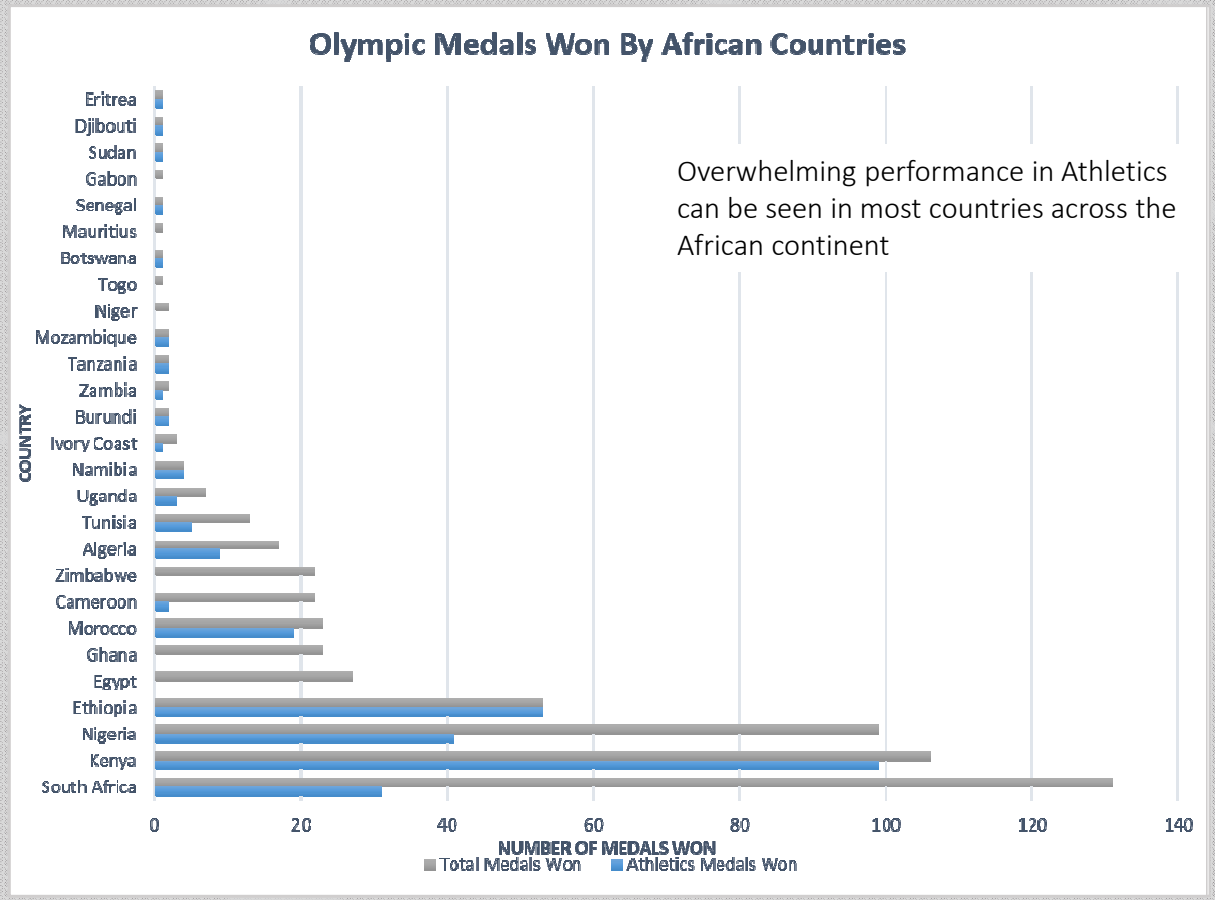
Most expensive Sport	ICE HOCKEY
Least expensive Sport	ATHLETICS

costs measured:
registration, equipment, travel, camps, other

Source: money.com

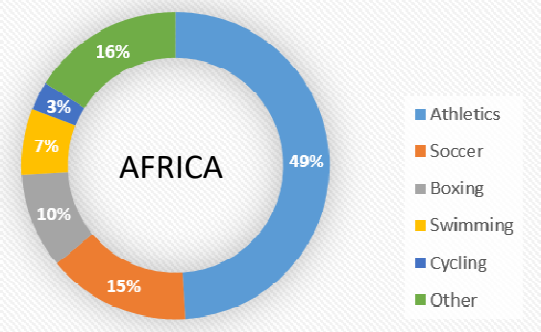
Athletics contingents as a percentage of total athletes for each Continent

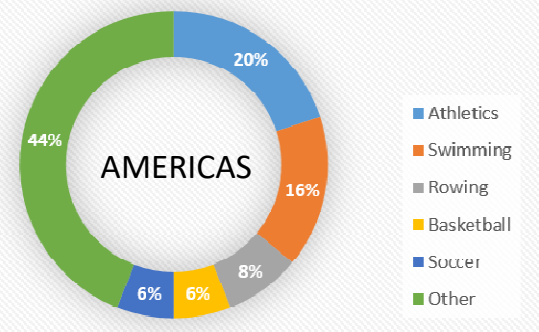
Africa	33%
Americas	21%
Europe	17%
Asia	16%
Oceania	15%

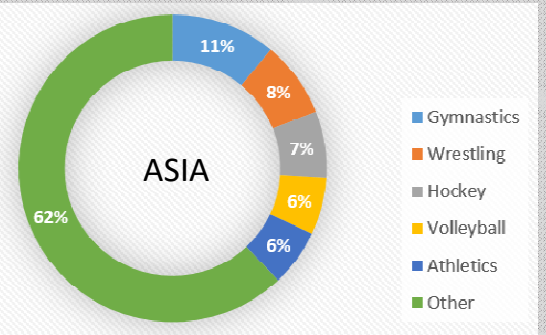


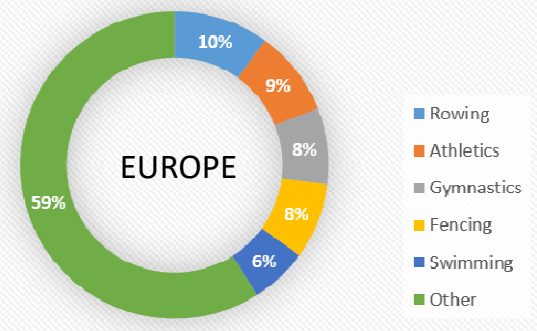
Further Analysis

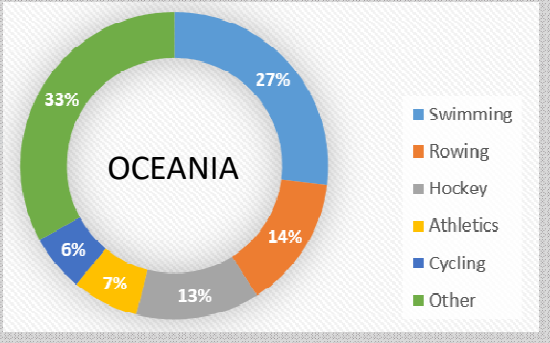
Medals Won by each Continent Grouped by Sport











Athletics accounts for half of Africa's medals.
Such inordinate performance in a single sport is clearly unique to Africa.

Hypothesis Results

- ✓ African athletes do have a heavy representation in one sport, Athletics.
- ✓ Athletics is the sport that is most represented on all the continents, however, this is particularly emphasized in Africa where athletics accounts for one-third of all African Olympic athletes.
- ✓ Athletics, Soccer, and Boxing account for 76% of all medals won by African athletes, with Athletics alone accounting for 49% of total medals won.
- ✓ The findings clearly indicate that African athletes do perform better in events that have the least barriers to entry, i.e. equipment cost.

Recommendations

- Athletics should be the immediate focus of any Sports brand looking to enter the African market.
- There is enormous opportunity for the development of facilities/equipment in other sports on the continent.

THANK YOU

Appendix: Samples of SQL used for this Project

African Olympic medalists

```
SELECT ID, Name, country, Sport, Medal, continent
 FROM athlete_events_4_csv
   JOIN noc_country_continent_4_csv
   ON athlete_events_4_csv.NOC = noc_country_continent_4_csv.NOC
 WHERE athlete_events_4_csv.Season = 'Summer'
   AND noc_country_continent_4_csv.continent = 'Africa'
   AND athlete_events_4_csv.Medal <> 'NA'
   AND athlete_events_4_csv.Sport = 'Athletics'
   ORDER BY athlete_events_4_csv.Year DESC
```

▸ (1) Spark Jobs

	ID	Name	country
1	86595	Francine Niyonsaba	Burundi
2	6204	Almaz Ayana Eba	Ethiopia
3	28311	Genzebe Dibaba Keneni	Ethiopia
4	87792	Hellen Onsando Obiri	Kenya
5	20525	Vivian Jepkemoi Cheruiyot	Kenya

Athletics athletes as a % of total African athletes grouped by country

```
SELECT tableAll.country, tableAll.num_of_athletes AS total_athletes, tableAth.athletics_athletes AS athletics_athletes, CONCAT
       (CAST((tableAth.athletics_athletes/tableAll.num_of_athletes) * 100 AS INT), "%") AS athletics_per_total
FROM (
      SELECT noc_country_continent_csv.Country AS country, COUNT (DISTINCT athlete_events_csv.ID) AS num_of_athletes,
             noc_country_continent_csv.Continent
      FROM athlete_events_csv
      JOIN noc_country_continent_csv
           ON athlete_events_csv.NOC = noc_country_continent_csv.NOC
      WHERE athlete_events_csv.Season = 'Summer'
            AND noc_country_continent_csv.continent = 'Africa'
      GROUP BY noc_country_continent_csv.Country, noc_country_continent_csv.Continent ) tableAll
 JOIN (
       SELECT noc_country_continent_csv.Country AS country, COUNT (DISTINCT athlete_events_csv.ID) AS athletics_athletes
       FROM athlete_events_csv
       JOIN noc_country_continent_csv
            ON athlete_events_csv.NOC = noc_country_continent_csv.NOC
       WHERE athlete_events_csv.Season = 'Summer'
            AND noc_country_continent_csv.continent = 'Africa'
            AND athlete_events_csv.Sport = 'Athletics'
       GROUP BY noc_country_continent_csv.Country ) tableAth
      ON tableAll.country = tableAth.country
ORDER BY athletics_per_total DESC
```

▸ (5) Spark Jobs

Athletics	Silver	Africa
Athletics	Gold	Africa

Athletics medals won as a % of total medals won by Africans grouped by country

```
SELECT tableA.Country, tableA.total_medals, tableB.athletics_medals, CONCAT (CAST((tableB.athletics_medals/tableA.total_medals) * 100
       AS INT), "%") AS athletics_percent_of_total_medals
FROM (
      SELECT Country, COUNT(Medal) AS total_medals
      FROM (
            SELECT *
            FROM athlete_events_4_csv
              JOIN noc_country_continent_4_csv
              ON athlete_events_4_csv.NOC = noc_country_continent_4_csv.NOC
            WHERE athlete_events_4_csv.Season = 'Summer'
              AND noc_country_continent_4_csv.continent = 'Africa'
              AND athlete_events_4_csv.Medal <> 'NA'
            ) GROUP BY Country ) tableA
LEFT JOIN (
      SELECT Country, COUNT(Medal) AS athletics_medals
      FROM (
            SELECT *
            FROM athlete_events_4_csv
              JOIN noc_country_continent_4_csv
              ON athlete_events_4_csv.NOC = noc_country_continent_4_csv.NOC
            WHERE athlete_events_4_csv.Season = 'Summer'
              AND noc_country_continent_4_csv.continent = 'Africa'
              AND athlete_events_4_csv.Medal <> 'NA'
              AND athlete_events_4_csv.Sport = 'Athletics'
            ) GROUP BY Country ) tableB
ON tableA.Country = tableB.Country
GROUP BY tableA.Country, tableA.total_medals, tableB.athletics_medals
ORDER BY tableA.total_medals DESC
```

Athletics medals won grouped by continent

```
SELECT continent, COUNT(Medal) AS athletics_medals, Sport
FROM
  (
  SELECT *
  FROM athlete_events_4_csv
    JOIN noc_country_continent_4_csv
    ON athlete_events_4_csv.NOC = noc_country_continent_4_csv.NOC
  WHERE athlete_events_4_csv.Season = 'Summer'
    AND athlete_events_4_csv.Medal <> 'NA'
     AND athlete_events_4_csv.Sport = 'Athletics'
  ) table1
GROUP BY continent, Sport
ORDER BY athletics_medals DESC
```

▸ (2) Spark Jobs

	continent	athletics_medals	Sport
1	Europe	1722	Athletics
2	Americas	1524	Athletics
3	Asia	330	Athletics
4	Africa	279	Athletics
5	Oceania	114	Athletics