Milestone 1 Project Proposal

Riley Taylor

Objectives

- In this presentation, I will address the following items:
  - Selection of the client
  - Description of the necessary steps to import and clean the data
  - Initial data exploration
  - A proposed ERD model

Client Description

- I’ve decided to work with SportStats to analyze data about previous Olympic medal winners
  - SportsStats is a sports analysis firm that works to provide insights to their partners
- I’m looking to provide an analysis that will develop a news story or discover key health insights based on geography
  - SportStats primarily works with local news and elite personal trainers to identify patterns/trends highlighting certain groups/events/countries, etc
  - These regular partners will benefit from the findings in order to reach a general public audience, as well as the elite athletic community

Preliminary Questions

- Is there a geographic pattern that correlates with the events that each country succeeds in?
- Is there a geographic pattern that correlates with the number of medals received by each country?

Initial Hypothesis

- There will be a correlation between geography and performance
  - Countries with colder climates will perform best in the Winter Games
  - Countries with warmer climates will perform best in the Summer Games
- Countries with higher populations will have higher medal counts
  - A higher population will be correlated with a higher number of competitive athletes to choose for the national team for each event

Approach

- I’ll primarily be looking at the frequency of medal wins and will separate by Summer vs Winter Games
  - From there, I’ll analyze by country, sport, and event
- Columns I expect to primarily analyze:
  - Team, Games, Year, Season, Sport, Event, and Medal
- Target Metric:
  - Count of medals by season and country

Importing the Athlete Dataset

- I imported the Athlete Dataset using Python's Pandas Library
- The data was imported as athlete_data and the info is displayed

```
import pandas as pd
```
[2] ✓ 11.3s Python

```
#import data and get info
athlete_data = pd.read_csv('/Users/rileytaylor/Downloads/athlete_events.csv')

athlete_data.info()
```
[4] ✓ 2.4s Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 15 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   ID      271116 non-null  int64
 1   Name    271116 non-null  object
 2   Sex     271116 non-null  object
 3   Age     261642 non-null  float64
 4   Height  210945 non-null  float64
 5   Weight  208241 non-null  float64
 6   Team    271116 non-null  object
 7   NOC     271116 non-null  object
 8   Games   271116 non-null  object
 9   Year    271116 non-null  int64
 10  Season  271116 non-null  object
 11  City    271116 non-null  object
 12  Sport   271116 non-null  object
 13  Event   271116 non-null  object
 14  Medal   39783 non-null   object
dtypes: float64(3), int64(2), object(10)
```

Import Additional Data

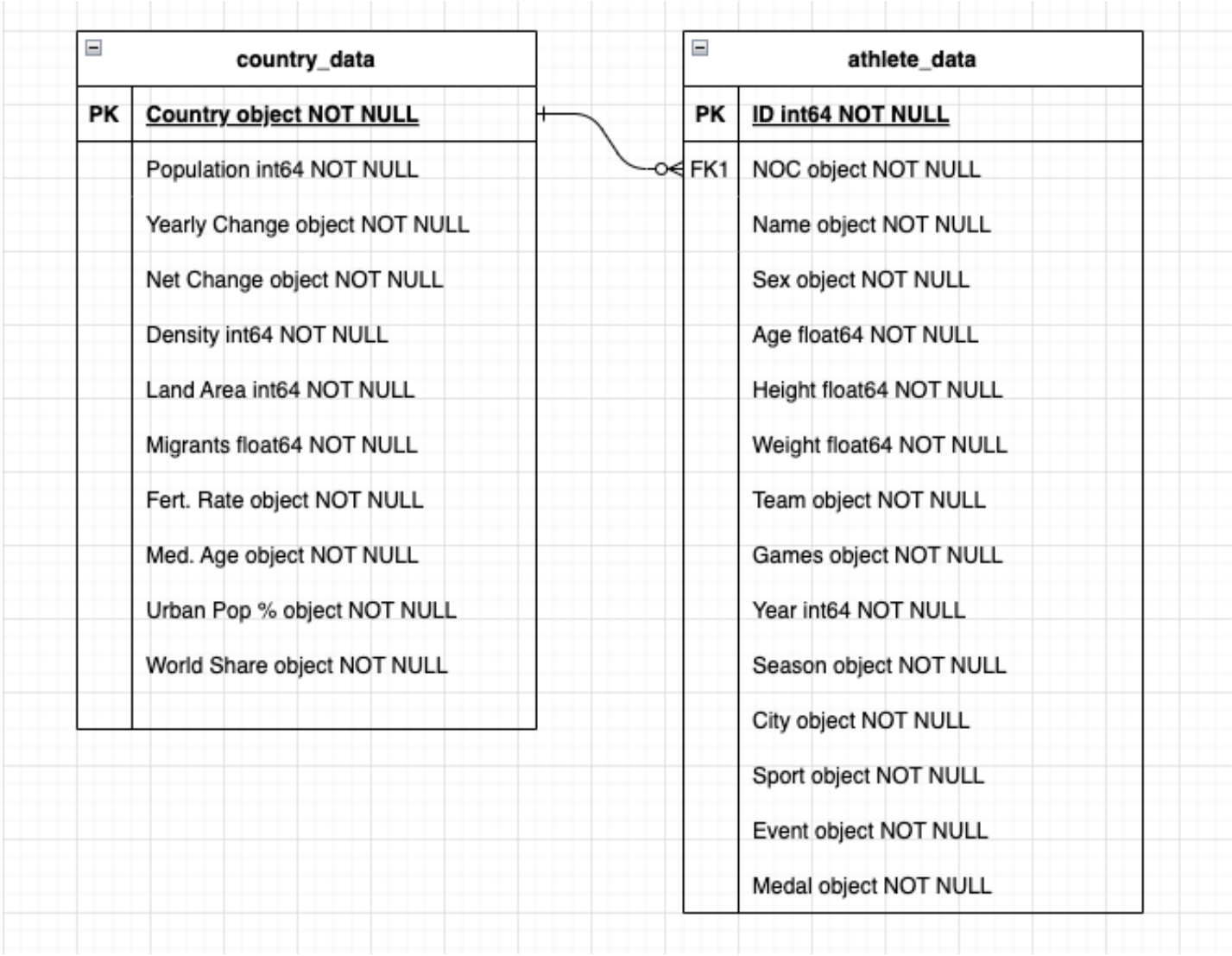
- I decided to import another Dataset from Kaggle to be able to analyze data from each country
- The data was imported as country_data and the info is displayed

```
#import datasets for info about each country

country_data = pd.read_csv('/Users/rileytaylor/Downloads/population_by_country_2020.csv')

country_data.info()
```

[5] ✓ 0.8s Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 235 entries, 0 to 234
Data columns (total 11 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Country (or dependency)  235 non-null    object
 1   Population (2020)        235 non-null    int64
 2   Yearly Change            235 non-null    object
 3   Net Change               235 non-null    int64
 4   Density (P/Km²)          235 non-null    int64
 5   Land Area (Km²)          235 non-null    int64
 6   Migrants (net)           201 non-null    float64
 7   Fert. Rate               235 non-null    object
 8   Med. Age                 235 non-null    object
 9   Urban Pop %              235 non-null    object
 10  World Share              235 non-null    object
dtypes: float64(1), int64(4), object(6)
memory usage: 20.3+ KB
```

Create an ERD



Initial Exploration

```
!select the 20 countries with the most medals from athlete_data set
%qlit("SELECT Team, COUNT(Medal) AS 'Medal Count' FROM athlete_data GROUP BY Team ORDER BY COUNT(Medal) DESC LIMIT 20")
```

[22] ✓ 10.7s Python Python

...

	Team	Medal Count
0	United States	5219
1	Soviet Union	2451
2	Germany	1984
3	Great Britain	1673
4	France	1550
5	Italy	1527
6	Sweden	1434
7	Australia	1306
8	Canada	1243
9	Hungary	1127
10	Russia	1110
11	Netherlands	988
12	East Germany	941
13	Japan	911
14	Norway	910
15	China	901
16	Finland	876
17	Romania	651
18	South Korea	592
19	Switzerland	588