Milestone 4 Project Results

Riley Taylor

Objectives

- In this presentation, I will address the following items:
    - Client Information, Initial Hypotheses, and Initial Approach
- Analysis Results
    - Correlations and Metrics
    - Relation to Initial Hypotheses
- Graphics and Visualizations

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

Preliminary Questions

- Is there a geographic pattern that correlates with the events that each country succeeds in?
  - How would climate affect the number of medals won in the Summer vs Winter Games?
- Is there a geographic pattern that correlates with the number of medals received by each country?

Initial Hypothesis

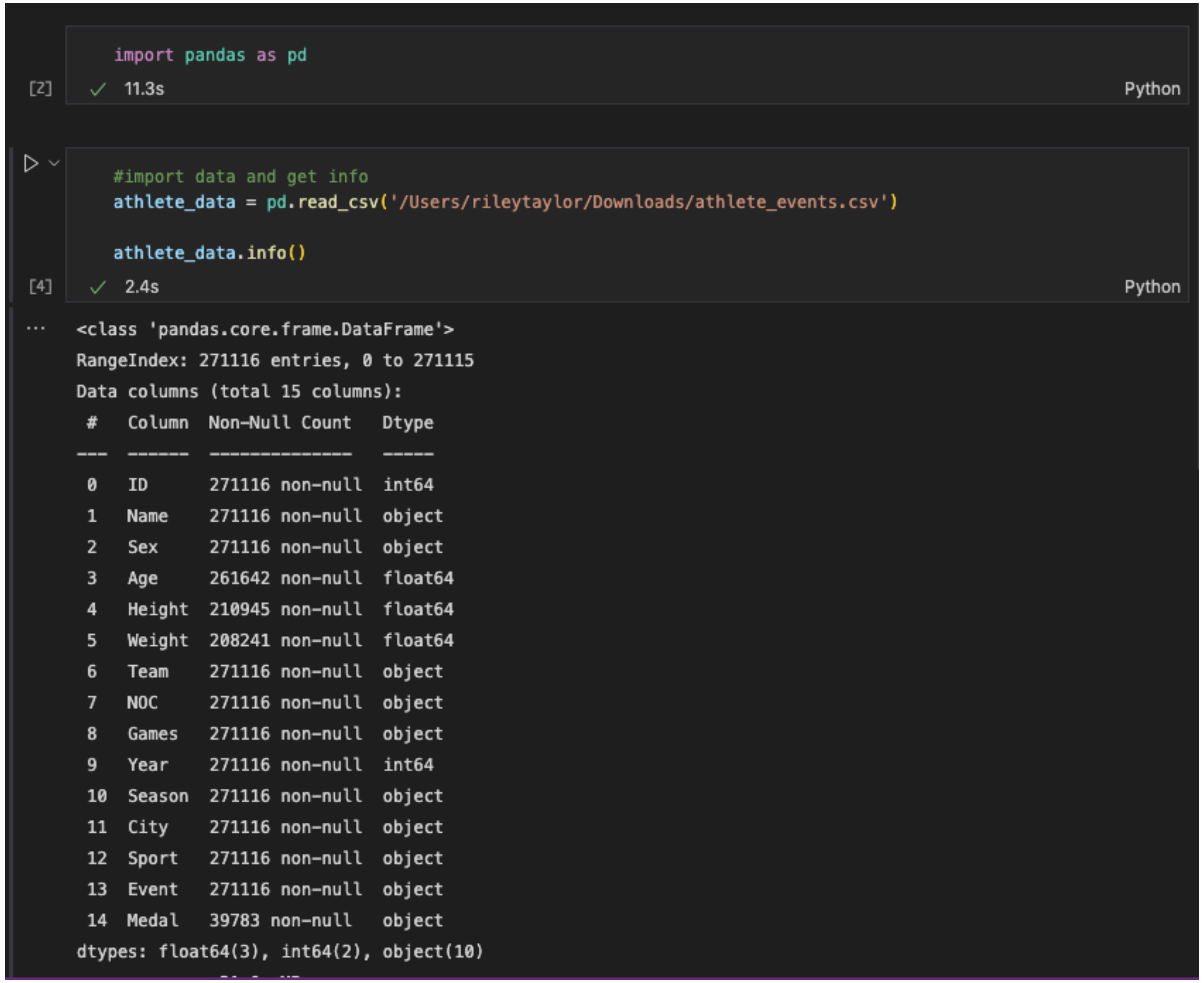
- There will be a correlation between geography and performance
  - Countries with colder climates will perform best in the Winter Games
  - Countries with warmer climates will perform best in the Summer Games
- Countries with higher populations will have higher medal counts
  - A higher population will 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



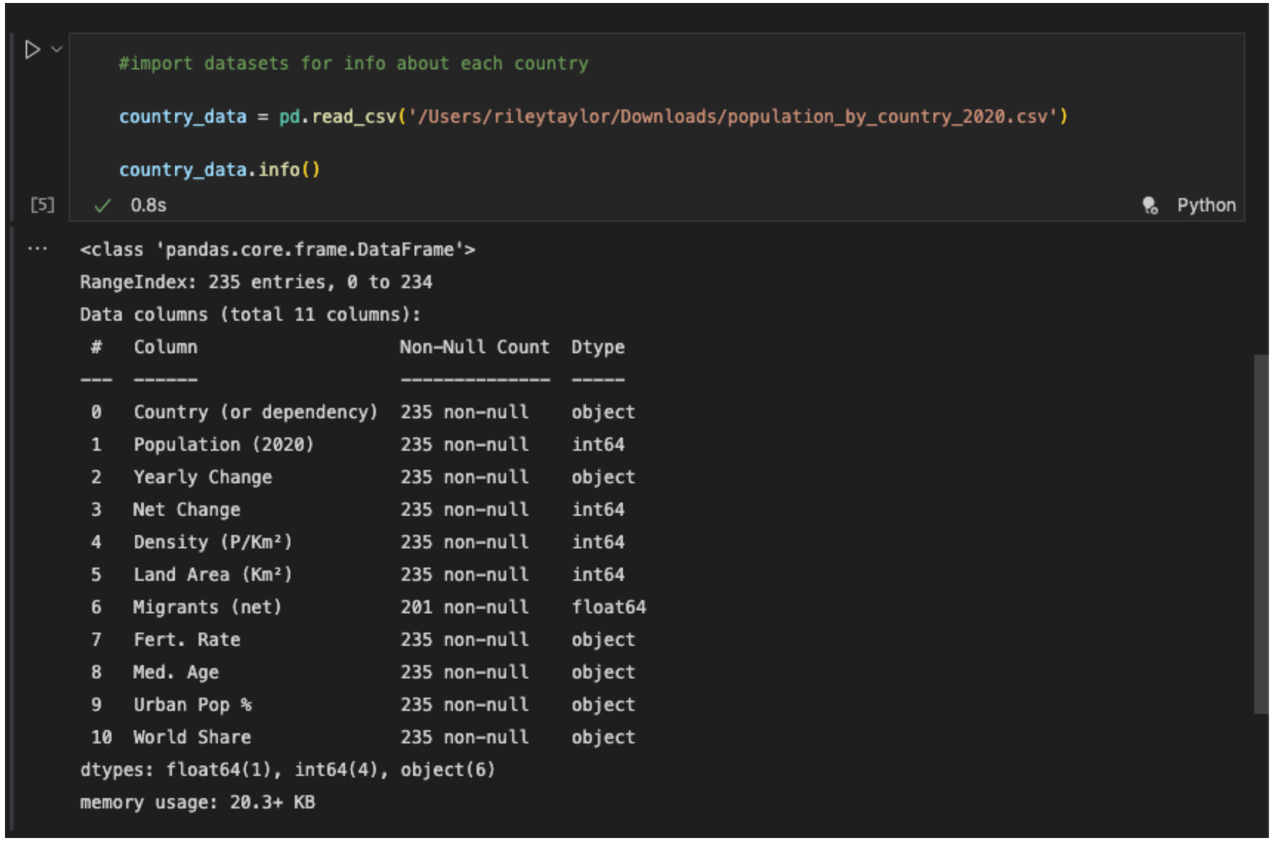
```
import pandas as pd
```

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

athlete_data.info()
```

```
<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)
```

Importing Additional Data

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

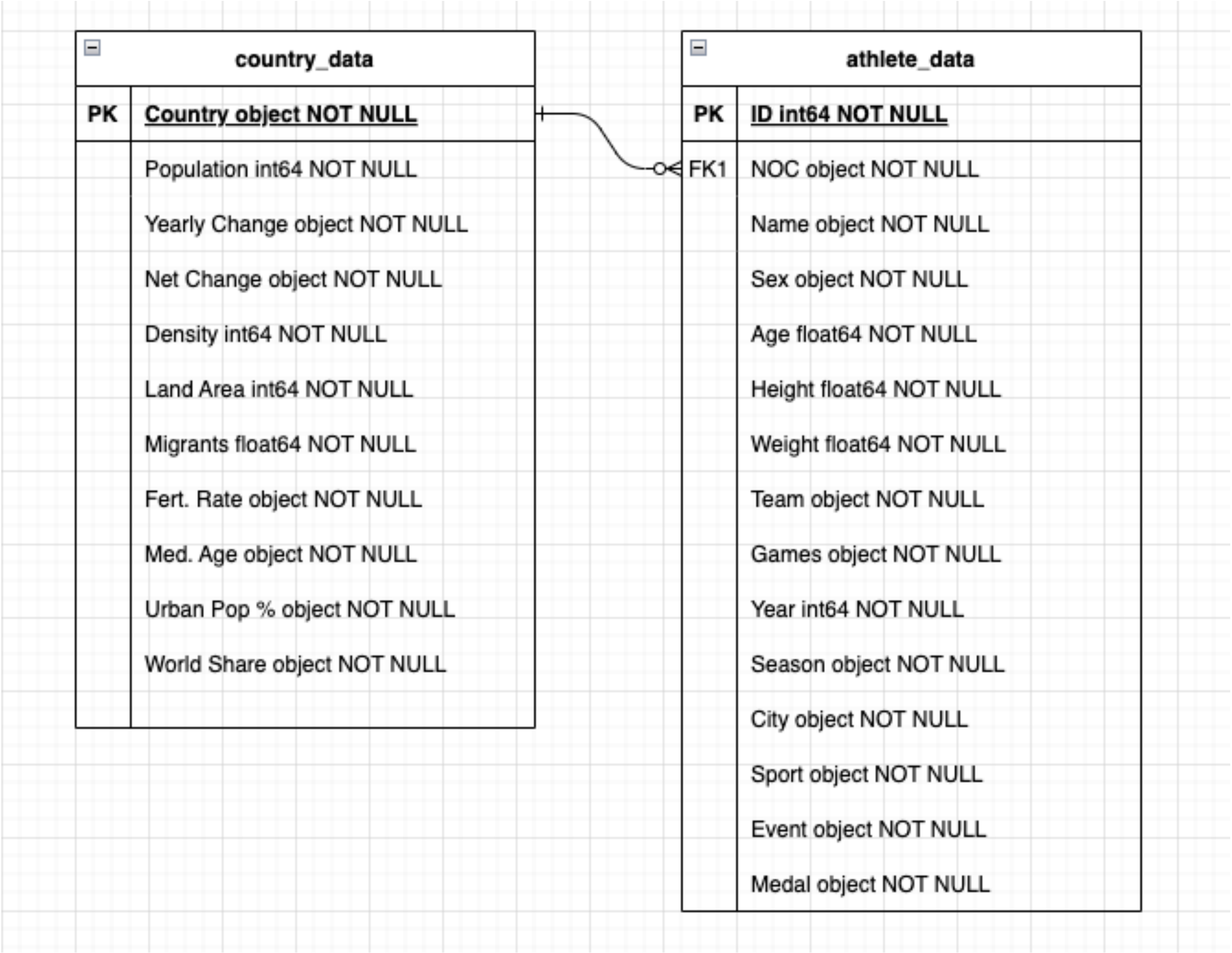


```
#import datasets for info about each country

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

country_data.info()
```

[5] ✓ 0.8s

```
<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



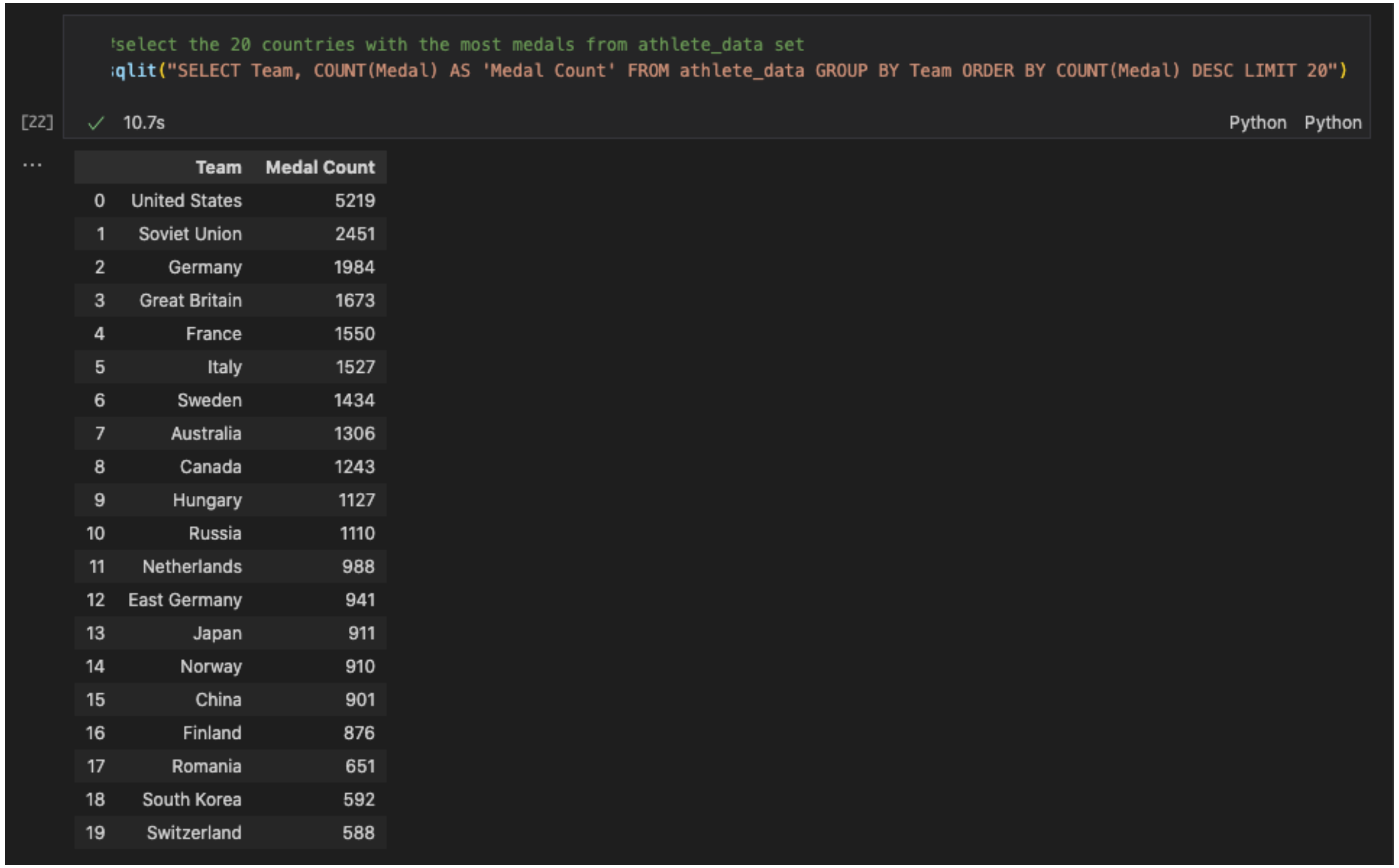
country_data

PK	**Country object NOT NULL**
	Population int64 NOT NULL
	Yearly Change object NOT NULL
	Net Change object NOT NULL
	Density int64 NOT NULL
	Land Area int64 NOT NULL
	Migrants float64 NOT NULL
	Fert. Rate object NOT NULL
	Med. Age object NOT NULL
	Urban Pop % object NOT NULL
	World Share object NOT NULL

athlete_data

PK	**ID int64 NOT NULL**
FK1	NOC object NOT NULL
	Name object NOT NULL
	Sex object NOT NULL
	Age float64 NOT NULL
	Height float64 NOT NULL
	Weight float64 NOT NULL
	Team object NOT NULL
	Games object NOT NULL
	Year int64 NOT NULL
	Season object NOT NULL
	City object NOT NULL
	Sport object NOT NULL
	Event object NOT NULL
	Medal object NOT NULL

Initial Exploration



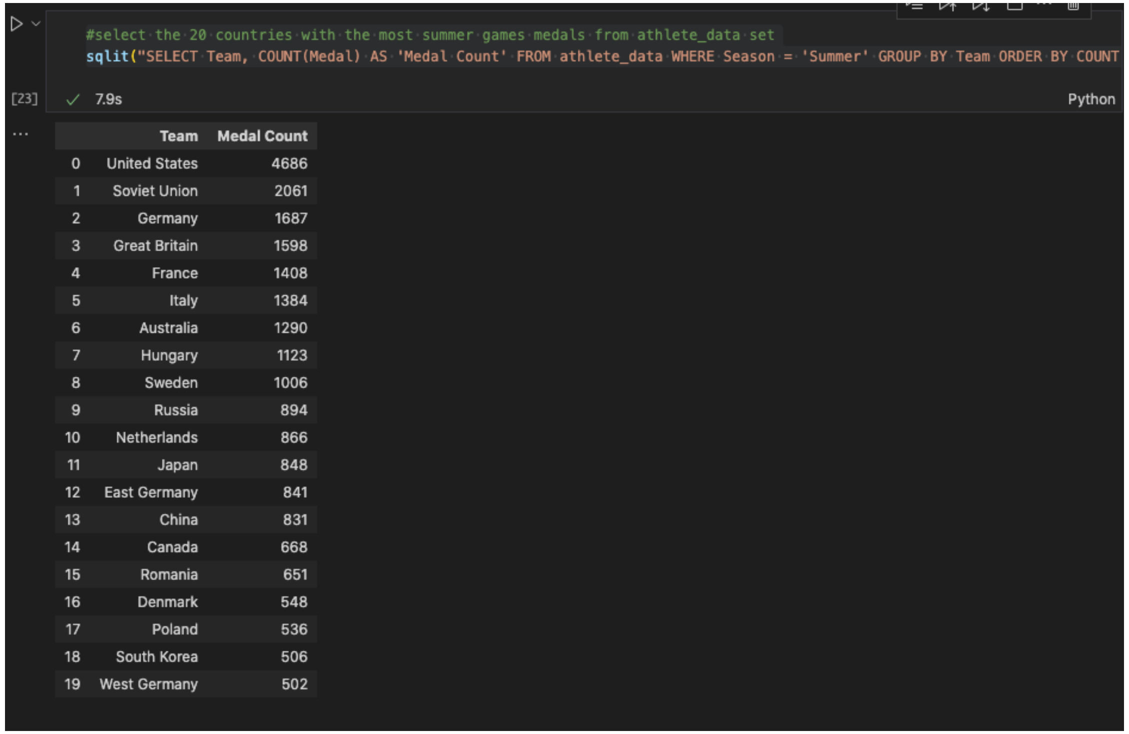
```
select the 20 countries with the most medals from athlete_data set
qlit("SELECT Team, COUNT(Medal) AS 'Medal Count' FROM athlete_data GROUP BY Team ORDER BY COUNT(Medal) DESC LIMIT 20")
```

	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

Technical Challenges

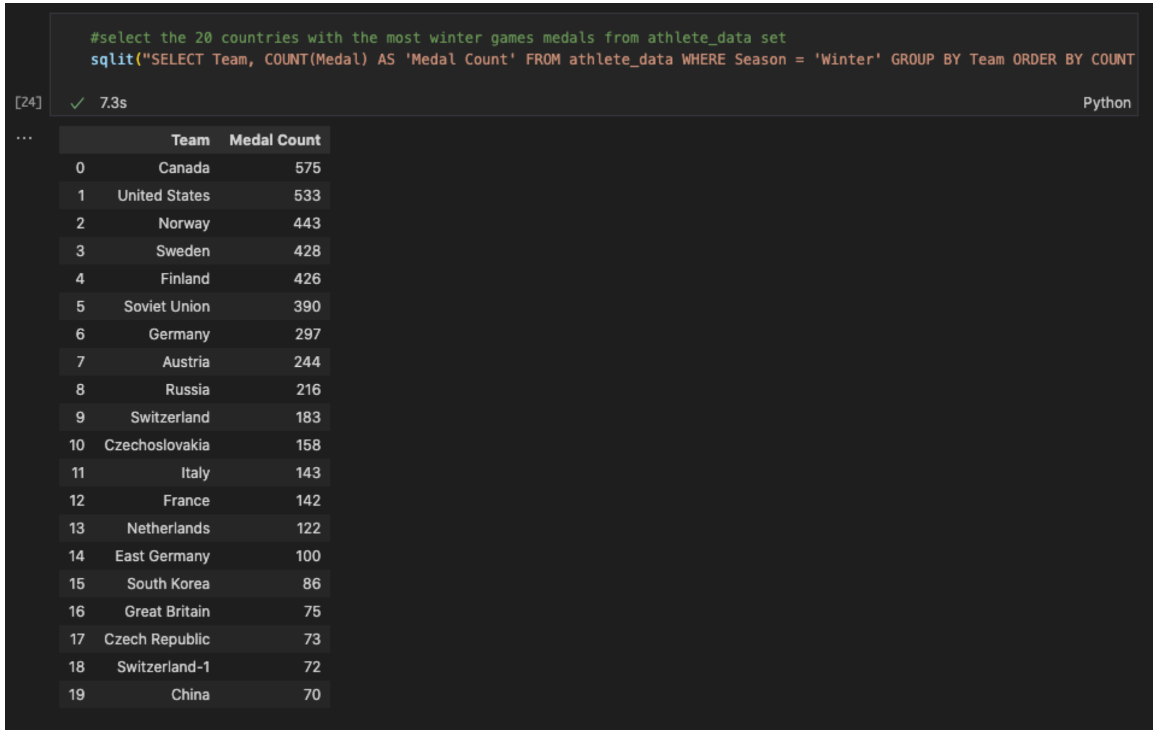
- Since SQL queries were run within Python using SQLite, there were a few limitations
  - However, these were later resolved once the data was cleaned accordingly, and the formatting issues were addressed within my queries

Initial Findings



```
#select the 20 countries with the most summer games medals from athlete_data set
sqlit("SELECT Team, COUNT(Medal) AS 'Medal Count' FROM athlete_data WHERE Season = 'Summer' GROUP BY Team ORDER BY COUNT
```

	Team	Medal Count
0	United States	4686
1	Soviet Union	2061
2	Germany	1687
3	Great Britain	1598
4	France	1408
5	Italy	1384
6	Australia	1290
7	Hungary	1123
8	Sweden	1006
9	Russia	894
10	Netherlands	866
11	Japan	848
12	East Germany	841
13	China	831
14	Canada	668
15	Romania	651
16	Denmark	548
17	Poland	536
18	South Korea	506
19	West Germany	502



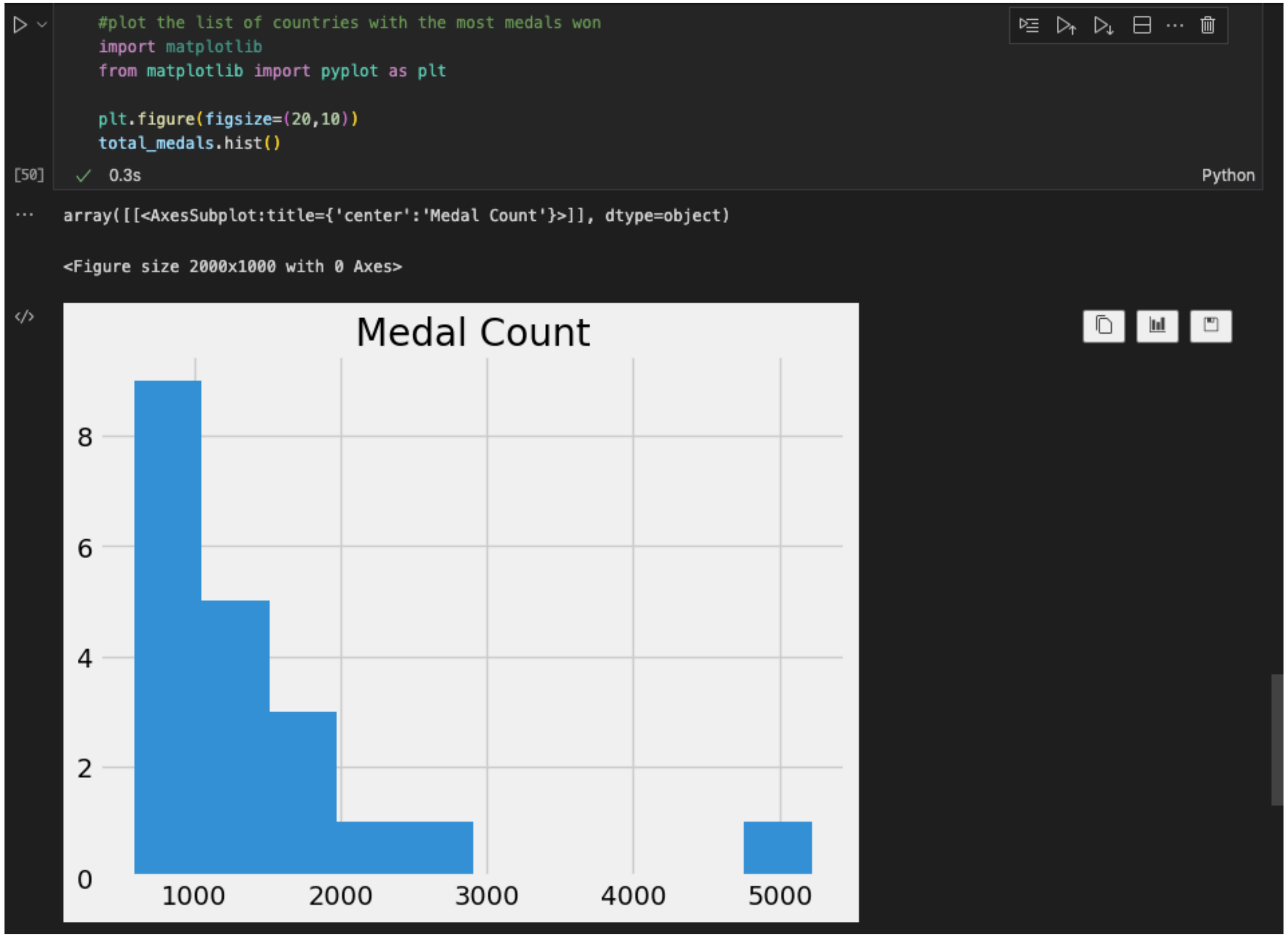
```
#select the 20 countries with the most winter games medals from athlete_data set
sqlit("SELECT Team, COUNT(Medal) AS 'Medal Count' FROM athlete_data WHERE Season = 'Winter' GROUP BY Team ORDER BY COUNT
```

	Team	Medal Count
0	Canada	575
1	United States	533
2	Norway	443
3	Sweden	428
4	Finland	426
5	Soviet Union	390
6	Germany	297
7	Austria	244
8	Russia	216
9	Switzerland	183
10	Czechoslovakia	158
11	Italy	143
12	France	142
13	Netherlands	122
14	East Germany	100
15	South Korea	86
16	Great Britain	75
17	Czech Republic	73
18	Switzerland-1	72
19	China	70

Initial Findings (cont.)

- The lists of the countries with the most medals by Winter vs Summer Games were incredibly similar
  - The Winter Games list suggests a slight correlation of climate vs number of medals won, but wasn’t statistically significant upon further analysis
- These results led me toward further analysis based on population and the percentage of the population of each country living in urban areas

Visualizations

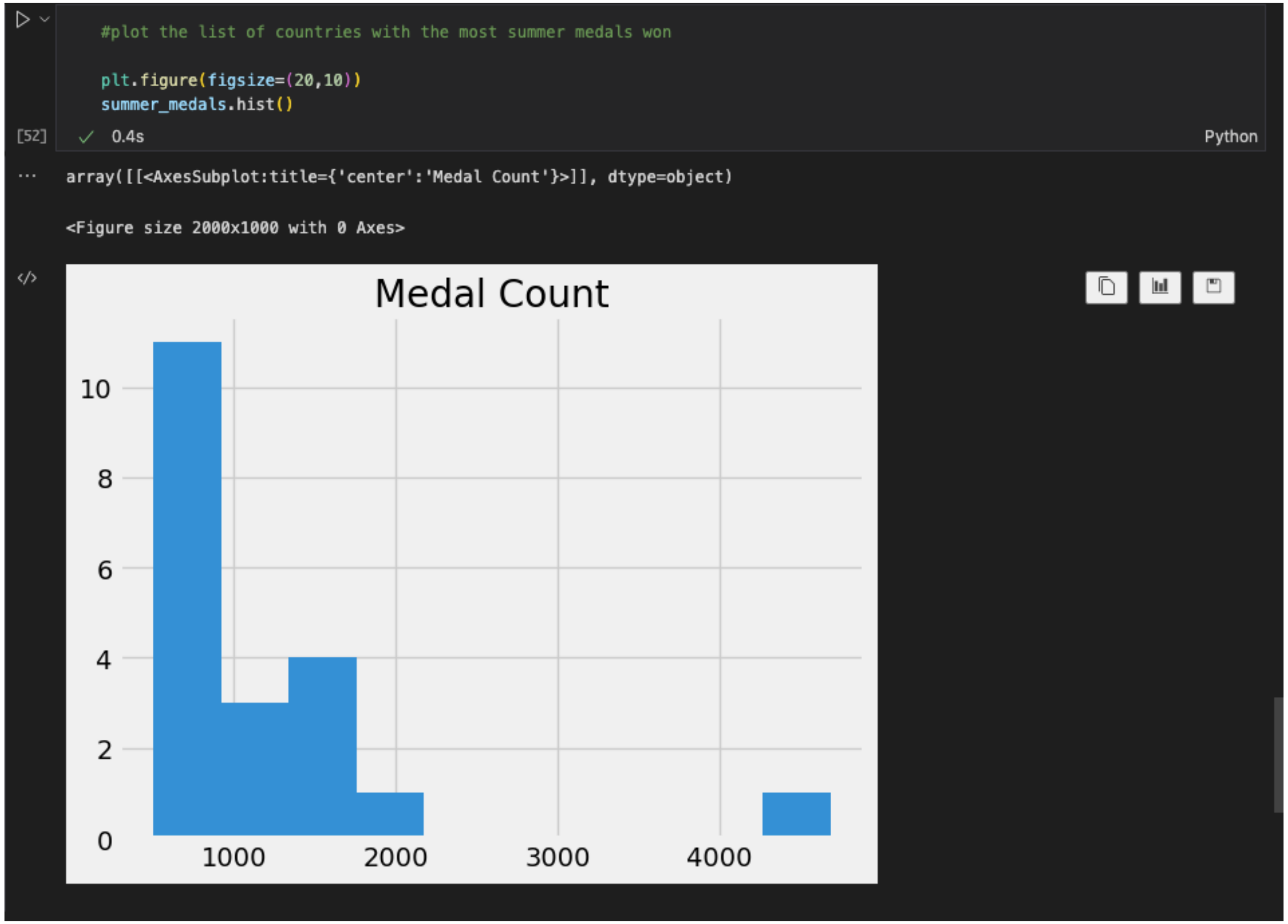


```
#plot the list of countries with the most medals won
import matplotlib
from matplotlib import pyplot as plt

plt.figure(figsize=(20,10))
total_medals.hist()
```

[50] ✓ 0.3s

```
array([[<AxesSubplot:title={'center':'Medal Count'}>]], dtype=object)

<Figure size 2000x1000 with 0 Axes>
```

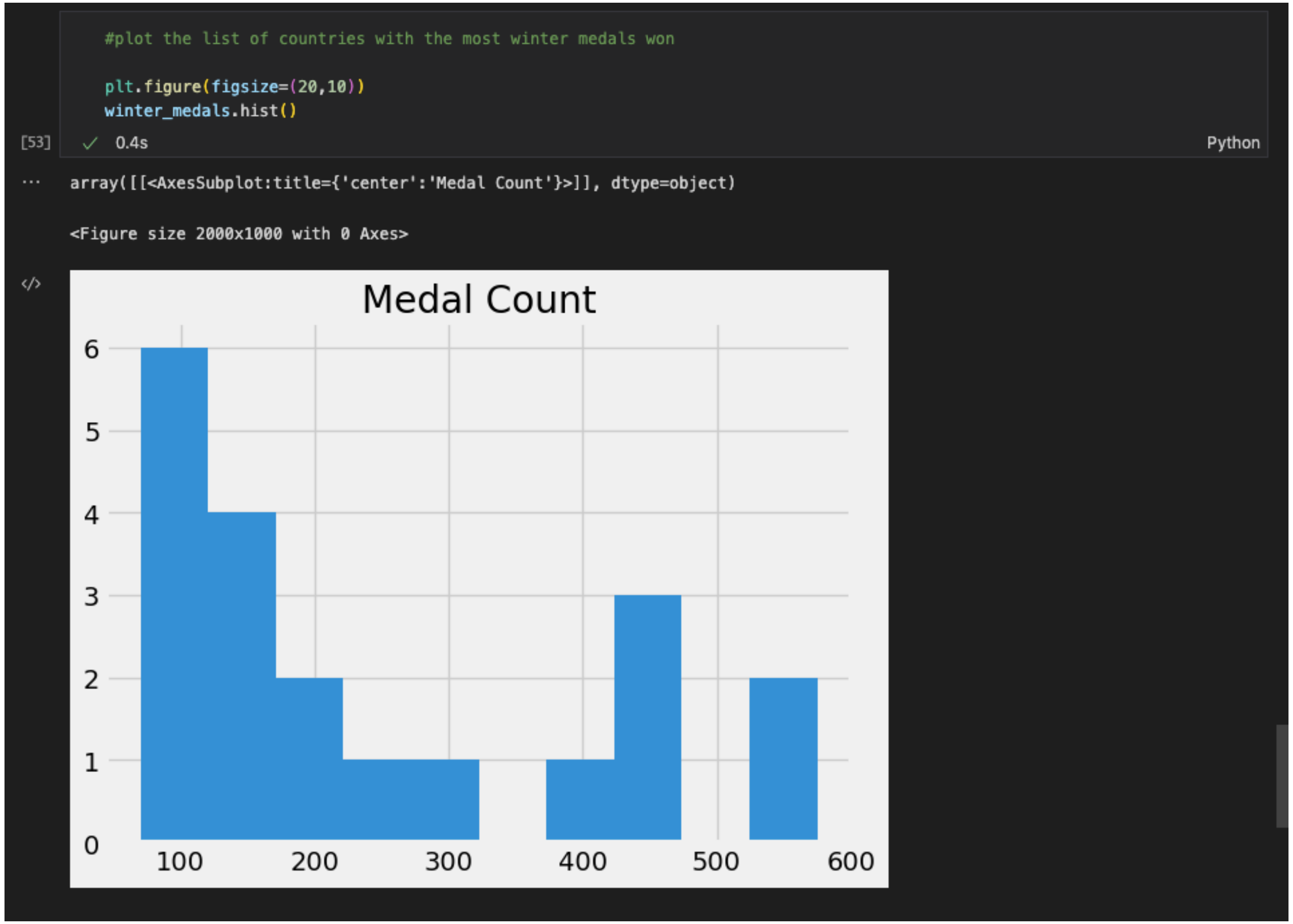
Visualizations (cont.)



```
#plot the list of countries with the most summer medals won

plt.figure(figsize=(20,10))
summer_medals.hist()
```

```
array([[<AxesSubplot:title={'center':'Medal Count'}>]], dtype=object)

<Figure size 2000x1000 with 0 Axes>
```

Visualizations (cont.)



```
#plot the list of countries with the most winter medals won

plt.figure(figsize=(20,10))
winter_medals.hist()
```

[53] ✓ 0.4s Python

```
array([[<AxesSubplot:title={'center':'Medal Count'}>]], dtype=object)

<Figure size 2000x1000 with 0 Axes>
```

Deeper Analysis and Final Findings

- Deeper analysis showed a more significant correlation between population size and the percentage of the population living in urban areas
  - Additional analysis may be needed, but I suspect that pulling in additional data would provide more insight
  - For example, a more urbanized population may suggest a higher GDP, which I suspect may correlate with an increase in the number of medals won