Project Proposal

Made by: Akash Chovatiya

Preparing for Proposal

1. Which client/dataset did I select and why?
   - The SportsStats Olympics Dataset has been chosen to work with. As I like to play sports, I decided to go with a sports analysis. The analysis may be helpful to news media to develop their story, and to personal trainers to develop their strategies to give necessary health insights to their customers.

2. The steps I took to import and clean the data.
   - I downloaded the SportsStats dataset.
   - To work on the dataset, I first tried to work with databricks environment. But as I have earlier worked with jupyter notebook and pandas library, I decided to go with jupyter notebook.
   - I imported the dataset in jupyter notebook environment. I used the pandas library, imported it as ‘pd’ and used ‘pd.read_csv(file path)’ command to import the data.
   - While looking at datatypes of the columns, I found that ‘Age’ column was given ‘float64’ datatype. To make it proper, I changed its datatype to ‘int64’.

Preparing for Proposal

3. Initial exploration of data and some screenshots or some stats of the data I am looking at.

- I explored the data and found that the dataset is the list of the participants who had participated in the Olympics from 1896 to 2016.
- The oldest participant was found the age of 97, the resident of United States, who participated the Olympics of 1928 in Art Competitions category.

```
In [18]: print(sports.Year.max())
         print(sports.Year.min())

         2016
         1896
```

```
In [8]: sports.loc[sports['Age']==sports.Age.max()]
```

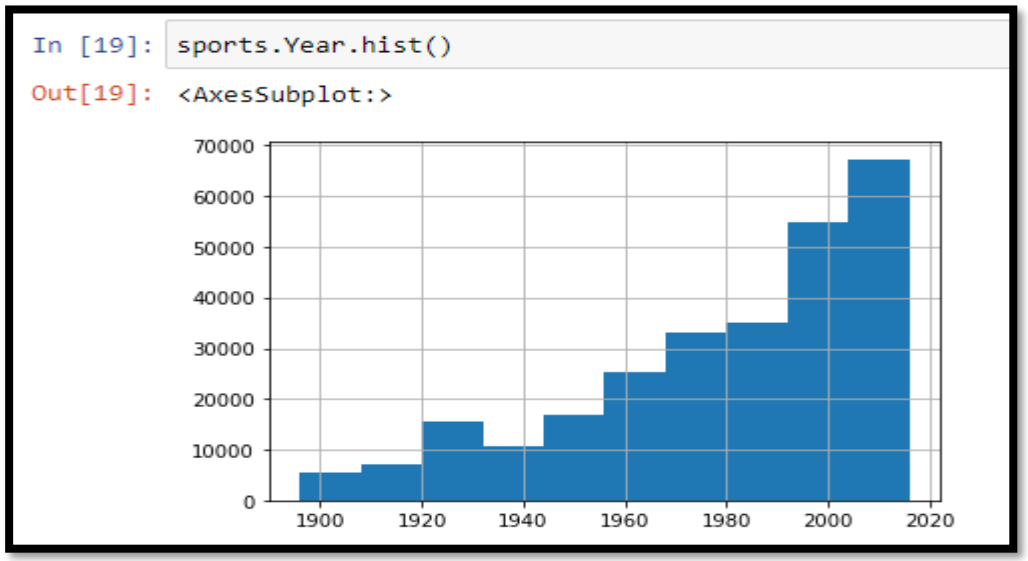
Out[8]:

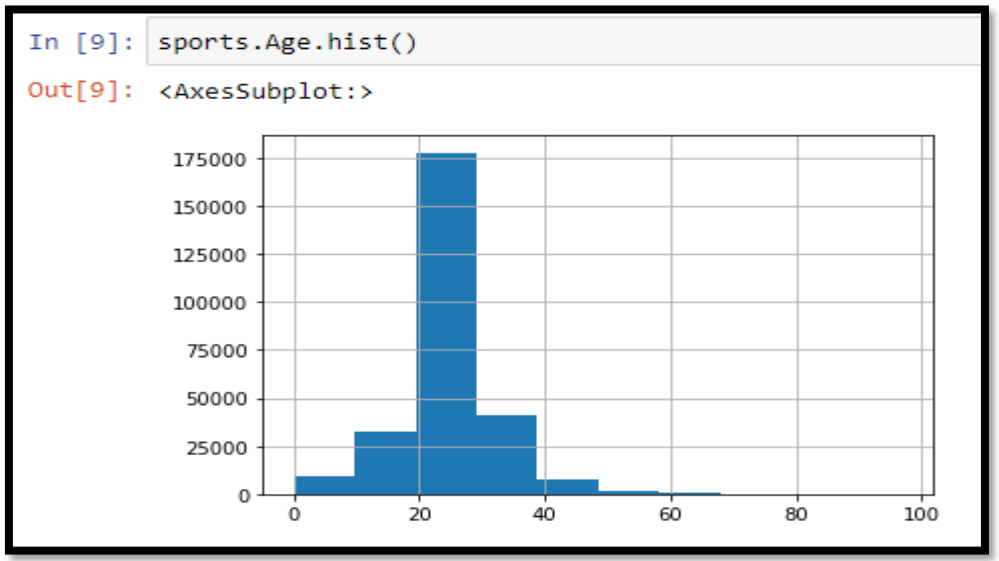
	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
257054	128719	John Quincy Adams Ward	M	97	NaN	NaN	United States	USA	1928 Summer	1928	Summer	Amsterdam	Art Competitions	Art Competitions Mixed Sculpturing, Statues	NaN

Preparing for Proposal

3. Initial exploration of data and some screenshots or some stats of the data I am looking at.

- The highest number of participants come from the age group of 20 to 24 (around 176000 participants).
- The lowest number of participants are found before 1920 (around 6000), and the highest number of participants are found after 2000 (around 67000). Within just 20 years duration (from 1995 to 2016), the number of participants has increased two time.





Preparing for Proposal

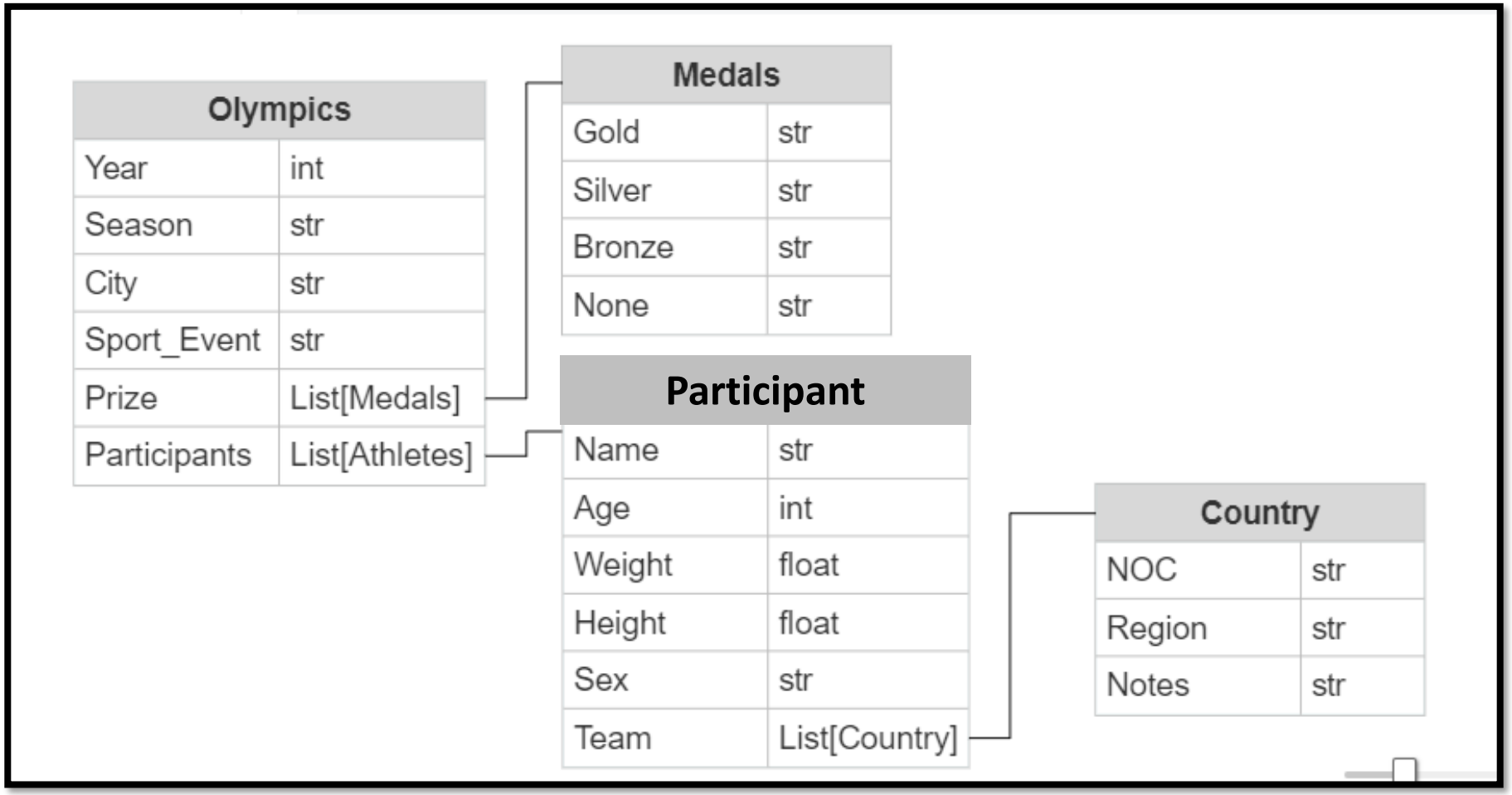
3. Initial exploration of data and some screenshots or some stats of the data I am looking at.

   - The United States has won the highest number of Gold Medal (i.e. 2474)

```
In [28]: print(len(sports[(sports.Medal=='Gold') & (sports.Team == 'United States')]))
         2474
```

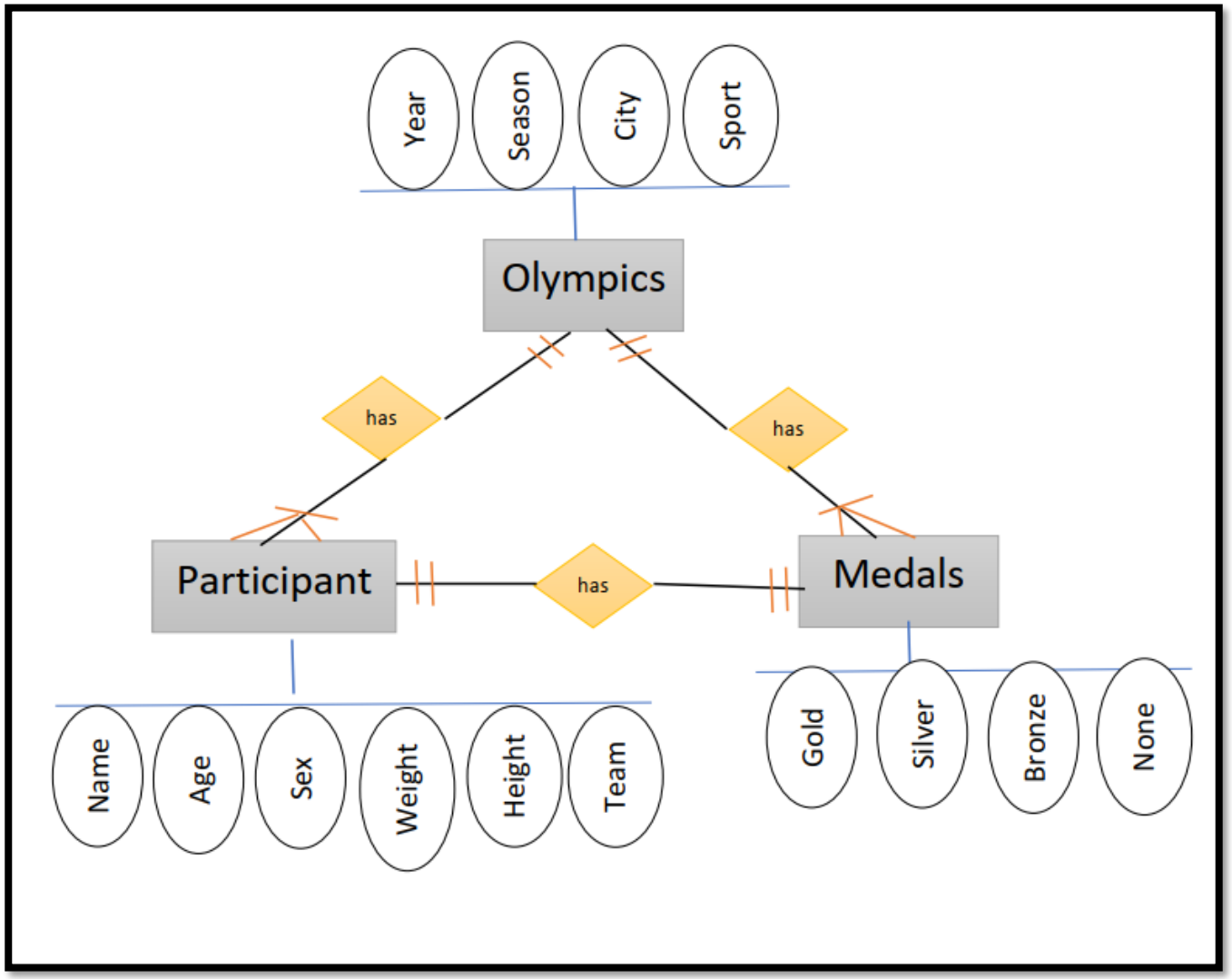
Preparing for Proposal

4. An ERD or proposed ERD to show the relationships of the data I am exploring.



Preparing for Proposal

4. An ERD or proposed ERD to show the relationships of the data I am exploring.



Develop Project Proposal

Description: A 5-6 sentence paragraph describing my project; the details of who might be interested to learn about my findings. Who might be my audience?

- My analysis provides necessary information to news media and elite personal trainers. The analysis is supposed to find the relationship between medals won in the sport events and the age group which won medals. This might help elite personal trainers to target the public of specific group. Moreover, it helps personal trainers to find their strategies by discovering key health insights and focus on that while giving training. On the other hand, news media can develop a news story considering past history of Olympics.
- The potential audience are elite personal trainers, news media and desired Olympics participants.

Develop Project Proposal

Questions: 2-3 questions that I want to answer with the data.

- Which top 5 countries have the most number of participants and/or medals won in Olympics? Is there any relation between number of participants from one country and medals won by that country ?
- Is there any relationship between age group and sport? For example, participants from the specific age group predominantly win in a specific sport?
- The number of participants in last 20 years have been increased more than ever in Olympics history? What might be the reason?
- Is there any trend associated with the season and medal won by the country ?

Develop Project Proposal

2-3 Assumptions - Hypotheses: What are my initial hypotheses about the data? 2-3 assumptions about the data that I'll want to go back to prove or disprove.

- It is common for all countries that majority of their medals are concentrated in a certain number of sports.
- The country/countries win more number of medals because it/they send higher number of participants.
- At most of the times, a player wins a medal in different events of the same sport.

Develop Project Proposal

Approach: 5-6 sentences about what approach I am going to take in order to prove (or disprove) my hypotheses.

- For the first hypothesis, I will look into the list of top 5 sports that have the most medals. I will compute percentage of these sports in relation to the total medals won by particular country. Based on this value, I will come to the conclusion for the first hypothesis. If these sports contain 40-50% of value then I will consider the first hypothesis as true.
- For the second one, I will study the total number of participants of specific country in the Olympics. I will look into the percentage of the participants who won medals. If the more number of participants means more medals then my hypothesis will be considered as true.
- For the third one, I will study the participants who won the medals in different events. I will group by it according to sports. If the same person will be found in many events, then the hypothesis is true.