Peyten Boutwell

Prof. John Chandler

November 20, 2020

Dataset Share #2

120 Years of Olympic History

1. Data set source: I downloaded the data from Maven Analytics.
2. Data set description: The dataset contains 120 years of Olympic history, including all the games from Athens 1896 to Rio 2016. The data downloads as a zip file, containing two csv files: athlete\_events.csv and country\_definitions.csv. Each row in the “athlete\_events.csv” file, corresponds to an individual athlete competing in an individual event. It includes information about the athlete (ID, name, sex, age, weight, height, country) and information about the event (games, year, city, sport, event, medal). The “country\_definitions.csv” file contains the abbreviation for the country (NOC), region, and notes.
3. Descriptive statistics:

* The file “athlete\_events.csv” is 35,335 KB, containing 271,117 rows of data, with 15 different rows.
* Tokens: 447,684
* Unique tokens: 22,415
* Lexical diversity: 0.050
* Average token length: 5.94
* Ten most common tokens:

1. Metres (33,219)
2. Skiing (18,899)
3. Country (18,266)
4. De (17,578)
5. Mixed (16,940)
6. X (15,452)
7. Skating (9,445)
8. Kilometres (8,996)
9. Team (7,612)
10. Individual (6,770)

* I opened the “athlete\_events.csv” directly in excel to conceptualize the data a bit further. I created a pivot table and organized it by NOC and count of medals. Here are the top five countries based on medal count over the past 120 years:
  + USA – 18,853 (United States)
  + FRA – 12,758 (France)
  + GBR – 12,256 (UK)
  + ITA – 10,715 (Italy)
  + GER – 9,830 (Germany)

1. Interesting questions: Some questions I would look further into are:

* What countries send the most athletes to the Olympics?
  + I can imagine that some countries send more athletes for the winter games compared to the summer games and vise versa. I think depending on your country’s geographics, some events are easier to specialize in than others.
* What if the percent of female athletes over time?
  + I can imagine that the number of female athletes has gradually increased over time. It would be interesting to see how the percent of female athletes changes before and after women’s rights issues in the 1920s.
* What are the most popular events?
  + I would imagine that there are some events that receive more athletes to compete in them than others (for example, badminton vs. swimming).

1. Approach for a single question: I would look further into the question, “What countries send the most athletes to the Olympics?”

* I would choose this question because I feel that I can look into some more specific things that relate to the question. For example, do those countries send more athletes to the summer or winter games? What percent of the athletes from the specified countries are female versus male?
* Steps I would take to analyze the main question: I would make sure the data is clean. Then I could create a database for the data, using Google Big Query (I theoretically could use DB Browser SQLite). Once I uploaded the data into GBQ, I could create specific queries to answer my question. I feel that this would be the best approach for my question because it is more about creating the right queries to pull a numerical answer to describe my question. I could then create a chart to visualize the data for a reader to understand easily.

GitHub Repo: https://github.com/pboutwell252/Dataset-share-2