Elements of Data Science SCTC 1013

Olympic Mini Project

(revised Feb. 22, 2023)
Project 1: Programming & Data Visualization
Data, Tables (6.1-6.4), Visualization (7.1, 7.2), Cross-classifying (8.3)

In celebration of the Olympic spirit we will analyze trends in a data set which spans the from the 1896 Athens games to Rio in 2016. With this data we will explore trends in medals awarded, sports, and countries, as well as any host country advantage. The dataset is from Kaggle (https://www.kaggle.com), a Data science dataset, coding, and competition site.

We will limit the data to Winter Olympics by using the where method [*.where("Season","Winter")*] which leaves us with 48,564 individual athletes. The dataset encodes missing values for athlete Age as “nan” but in the case of medals athletes who do not win are also encoded “nan”. We can remove extraneous ages some of which are labelled “nan” by restricted the “Age” column to between 0-99 using the .where method and are.between(0,99). For this mini-project this case you will start with a new Python Jupyter notebook on the Temple.2i2c.cloud server based on the Olympic template. Be sure to use comments (lines starting with #) and markdown cells (create a new cell and change it to markdown from the top menu, see: https://www.earthdatascience.org/courses/intro-to-earth-data-science/file-formats/use-text-files/format-text-with-markdown-jupyter-notebook/)

Stage 1:

Starting question:

1. What is the earliest year for a Winter Olympics in this dataset?
2. Examine the distribution of the age of all Olympians with min, max, average, and plot with a histogram.
3. Now examine the age distribution of only gold medal winners again with a histogram. Compare the distributions from question 2 and 3.
   Now look at number of athletes and medals for top 10 countries. Use *.group* method and be sure to exclude “nan” values (“NA” is not available and “nan” is not a number). Examine min, max, and average.
   Example:
   ```
   .where("Medal",are.not_equal_to("nan")).group("Sport
   ").
   sort("count",descending=True).take(np.arange(0,5))
   ```

4. Now look at number of athletes and medals for top 10 countries. Use the *.group* method. What are the top ten countries in number of athletes? The “nan” in the “Medal” column simply means an athletes did not win a medal but still participated. Get the five number summary (min, max, median, mean, and standard deviation) using np.min, np.max, np.median, np.mean, and np.std respectively on the corresponding column array.

5. What are the top ten countries in number of Gold, silver, bronze medals, and total medals? You should have four sets of top ten countries for each of the scenarios. Again, get the five number summary (min, max, median, mean, and standard deviation). Hint: `.where("Medal",are.not_equal_to("nan"))` to get only medal winners. Consider how to create a column for the sum of the three medal categories.
6. What are the top 5 sports in terms of number of athletes?
7. Which sports (top 5) have awarded the most medals?
8. Which sports (top 5) awarded the most medals in Lake Placid, New York (1980, https://www.lakeplacid.com/do/activities/olympic-sites).
9. Remember medals are awarded to each participant on a team, how does this effect the results you found above?

Stage 2: Time trends and comparative results

1. Plot the trend in number of athletes per year.
   Hint: `athletes.group("Year").plot("Year","count")`
2. Plot the number of medals per year.
3. Plot the number of gold medals per year excluding “Ice Hockey”, why hockey?
4. Plot an overlay of gold, silver, and bronze medals as a function of year on the same plot excluding hockey.
5. Compare the US and Norway medal counts as a function of year by overlaying their counts. Hint: You could create separate tables for the US and Norway using an appropriate .where method. Now these tables can be combined using the Table .append method which merges two tables for instance, NORUSA = US.append(Norway). You may also find *.pivot* useful.
6. Now use a scatter plot (*.scatter()*) to look at the number of athletes per year for the US versus that for Norway. What trends do you see?
7. Use a scatter to plot the number of athletes for each country versus the number of medals.

Upload the .html and .ipynb files to Canvas to complete your work.