Project 2 Proposal

Cecily Sun, Justin Hsia, Nobu Yamaguchi

Summary:

Dataset:

- Boston Marathon Raw Data 2001-2016 [name, gender, age, division, country, city]
- Boston Marathon Elevation Data

Supplemental Datasets:

- Hong Kong Marathon Raw Data 2016
- Moscow Marathon Raw Data 2018 (Elevation included)
- Historical Weather Data

Approach:

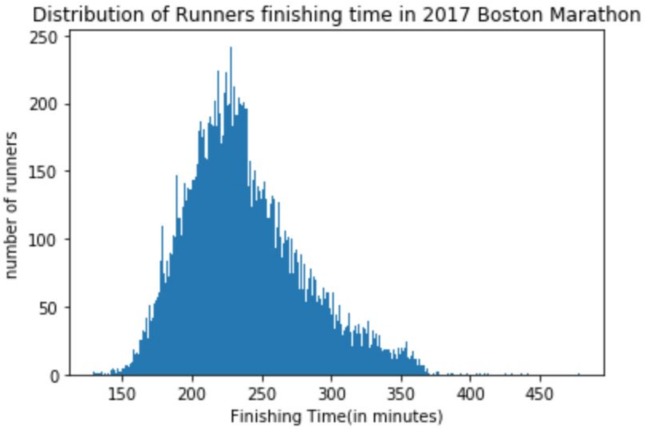
The project will include 3 modules of analysis: Exploratory Data Analysis, Correlation Analysis, and Cluster Analysis. The goal of the project is to generate meaningful insights and actionable recommendations for future Marathon runners in order to help them train well and achieve more.

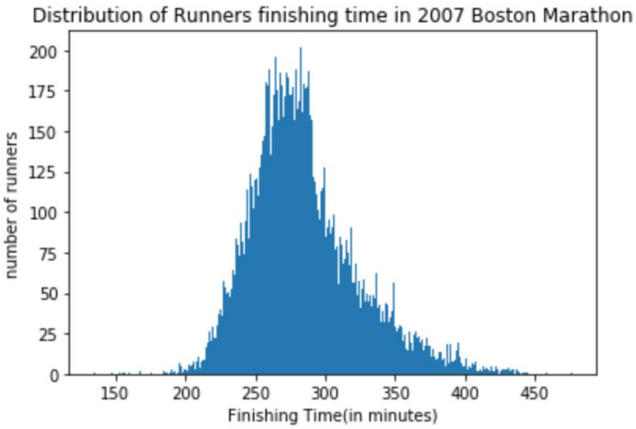
Part 1: Exploratory Data Analysis

In the first part of the project, we will be looking at a 15-year panel data on historical records of Boston Marathon runners. Here are a few questions we want to answer:

- What's the distribution of performance(finishing time) look like among all the runners?
- What's the effect of weather on runner's finishing time?
- What's the ratio between male runners and female runners? Is there a shift towards a certain group over the years? The same question can be applied to different age groups.

In a nutshell, when looking at the distribution of runners finishing time in 2017 vs 2007, there's a very significant trend that the majority of the runners' performance has greatly improved. Both of the distribution looks like a normal distribution, but the mean seems to have moved over to the shorter finishing time.





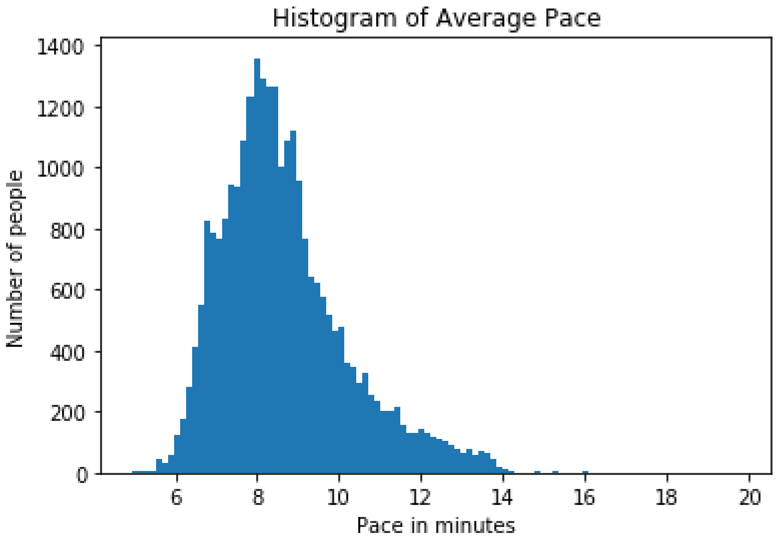
The results from the first part of the analysis are the mainly sanity check and descriptive data. We want to see if there are any outliers in the data sets that we should be aware of before delving into the deeper analysis. A descriptive data analysis also helps the team to come up with more data questions that will, later on, develop into meaningful research questions.

Part 2: Correlation Analysis

In the second part of the project, we will analyze data in closer detail from selected years. We will correlate the analysis with other types of data such as course elevation profile. In addition, we will also look into groupings runner base on other attributes such as state of origin. The questions we want to answer are:

- How does the elevation change affect runners splits and to what extent?
- Does the state origin play into each runner's finishing time? The underlying assumption here is each runner trains at his/her home state.
- Lastly, we will look at the number of participants from each state and attempts to understand if there is a correlation between the two.

Below is a sample graph showing the average mile pace of all runners from the 2016 Boston Marathon. Similar graphs will be generated for runners from each state.



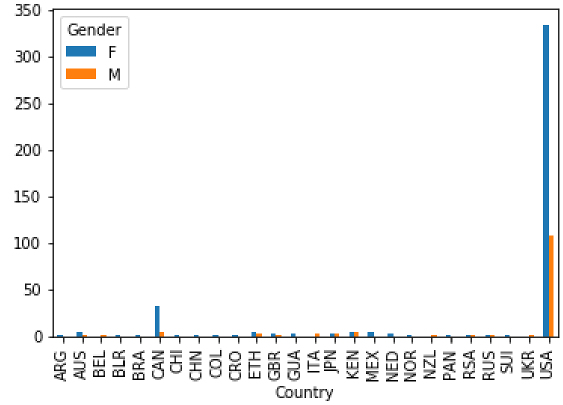
Part 3: Cluster Analysis

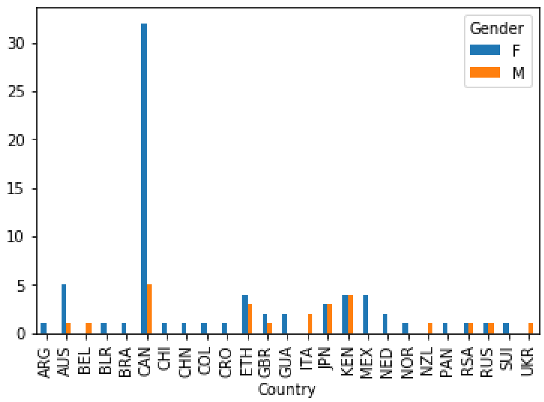
In the third part of the project, we will be clustering the data by time, country, and age. By using the cluster analysis, we want to answer the following questions.

- Are there any differences among countries and ages of elite runners (Finish time < 2:35 for male, Finish time < 3:10 for female)?
- Is the prime runner age truly be early male(30 - 36) and female(30 - 38)?
- How do top finishers running strategy/pattern differ from others? Do top finishers' running patterns evolve year over year?

We may add some other features in the analysis so that we can get more useful insights for marathon runners.

Below is the example of elite runners by country and gender in 2015. Since the number of US runners are so large, we excluded US runners in the second chart. Since the number of elite runners outside the US is small, we will combine data from multiple years for further analysis.





Littlest Bear Initiative

#ISchool5k