Nathaniel Medina
Nicolas Pappas
Ash Mohan
Rick Yang
Prof.

Boston Marathon Run Times

The Topic/Problem being addressed:

- Ideas:
    - investigating holistic Boston Marathon trends over time, including but not limited to:
        - Number of participants based on:
            - All runners
            - Age
            - Gender
            - Country
            - Returning runners
        - Average run time based on:
            - All runners
            - Age
            - Gender
            - Country
            - Returning runners

With increased worldwide public health awareness and emphasis on healthy lifestyles, the marathon has become a very common activity for people in recent years. The Boston Marathon is notoriously famous for having some of the toughest qualifying standards for a marathon and, as a result, attracts participants from all over the world. Recent studies have claimed that the average run time among participants is getting slower year by year. Is this because humans are getting slower? Or are there other factors at play?

The data set we will be using contains a lot of information about all the runners that participated in the Boston Marathon across 14 years. With such a wide range of data, we will be able to make interesting discoveries about the runners and the Boston Marathon itself. One topic we aim to focus on is the number of participants. We will split this into smaller subcategories such as

age, gender, country of origin, and returning runners. Some questions we hope to answer are: Which countries have the most participants? What is the most common age group? Which gender tends to participate more? How many returning runners are there every year?

Another topic we would like to analyze is the average run time. We will look at the run times of every participant to get a baseline average for each year. After that, we will look at the average run time based on age, gender, and country. (Maybe look at the range of times that were most common for each year and create a visualization of that? We would have to separate times into bins like 2-2.5 hours, etc. For ex. Say out of 30,000 participants a majority of them ran in between 2-2.5 hours one year and 2.5-3 hours the next year)

By analyzing these subcategories, we will be able to address whether or not times have been getting slower over the years and determine if there are any underlying factors that could have caused this.

Our Hypothesis:

An increase in public health awareness and the importance of living a healthy lifestyle has caused the focus of a marathon to lean more towards the social aspect. Due to this, we believe that the average run time among participants has decreased over the years. Participants run the race for the experience of participating in a world-known famous event without trying to have the best run time.

Our Method:

- Group together the data from all 14 years using Python/R
- Observe holistic trends using Python/R
- Create visualizations using Python/R
- Create heat maps using Tableau

To start our data analysis, we will group together the data from 2001-2014 using Python and R. The data will need some cleaning, so once this is done we will be able to investigate trends using both Python and R. After analyzing the data and making visualizations, we will be able to recognize certain trends and make new discoveries. Using Tableau, we can import the data set we create and come up with a way to visualize what we discovered on a global map. One

idea we have right now is creating a heat map of the change in the number of participants and average run times in certain countries across the 14 years.

Our Motivation behind the topic:

- The Boston Marathon is an iconic annual event in the United States. We thought it would be both meaningful and fun to investigate this publicly available Boston Marathon data and find deeper insights and trends about this annual event.

Description of Dataset and sources:

The dataset of Boston marathons from this github link, (https://github.com/llimllib/bostonmarathon/tree/master/results) allows us to collect information from 2001 to 2014. Within each year, multiple detailed data is presented for each individual runner. For each runner, information about his or her name, city of origin, country of origin, gender, age, race time, race ranking, and more is provided.

Simple sketch or mockup of potential design:

- Possible steps include
    - Cleaning and grouping the data
    - Creating high-level summaries of the data
    - Creating summaries of specific subcategories of the data (age, gender, etc)
    - Creating visualizations of these summaries in R/Python
    - Creating a heat map visualization of participating countries using Tableau

List of development tools we plan on using:

- Python for cleaning and grouping the data
- Rstudio for high-level summaries and visualizations
- Tableau for heat map visualizations

Task division among group members:

- Clean and group data → Everyone
- High-level summaries → Nathaniel
- Subcategory summaries → Ash
- R/Python Visualizations → Rick
- Tableau Heat Map → Nicolas

Timeline:

- Mid-Late October → have data cleaned and grouped
- Early-Mid November → have high-level summaries created
- Mid November → have visualizations and heat map created