Nathaniel Medina
Nicolas Pappas
Ash Mohan
Rick Yang

# Project Milestone 3
# Boston Marathon Run Times

**Brief Reminder:**

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. Also, we all enjoy running and have goals to complete marathons in the future.

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 observe 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 in each year. We will also observe the top performers from each

year and see how their times have changed. Ultimately, our goal is to see whether or not age, gender, and country has an effect on performance.

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. Participants run the race for the experience of participating in a world-known famous event without trying to have the best run time. Due to this, we believe that the average run time among all participants has decreased over the years. (First hypothesis was PROVEN FALSE. Runners are actually becoming faster.)

More participants are running in the Boston Marathon as the race increases in popularity and competitiveness around the world (similar to the World Olympics). The global prestige from competing in this race is encouraging participants from various countries to compete in the Boston Marathon.

**Potential Design:**

- 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.

**Current Status:**

- Completed analysis of various trends, such as :
    - Average run time across all participants
    - Number of participants based on year
    - Number of participants based on age split across years
    - Number of participants based on gender
    - Number of participating countries by year

**(Graphs are below)**

**Tasks Remaining:**

- Finish creating visualizations of trends
- Written analysis for completed visualizations
- Heat map for participating countries

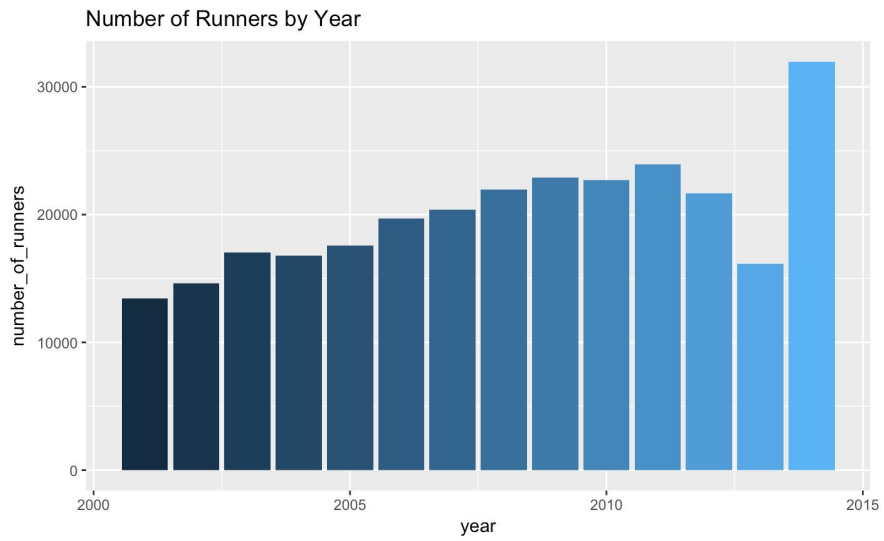
**Task division among group members:**

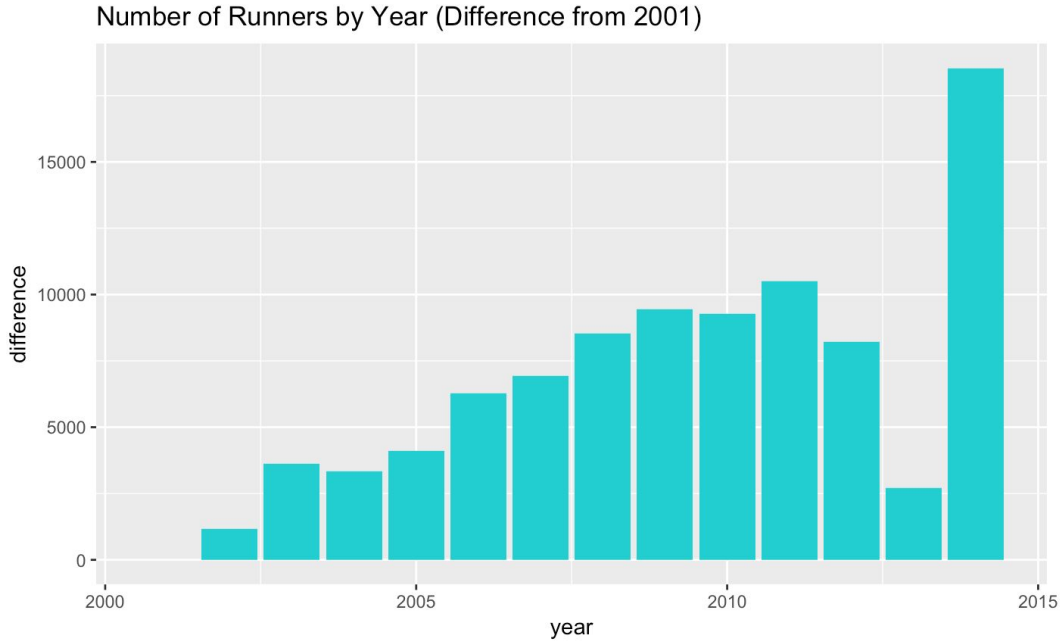
- Clean and group data → Everyone
- High-level summaries → Nathaniel
- Subcategory summaries → Ash
- R/Python Visualizations → Rick
- Tableau Heat Map → Nicolas

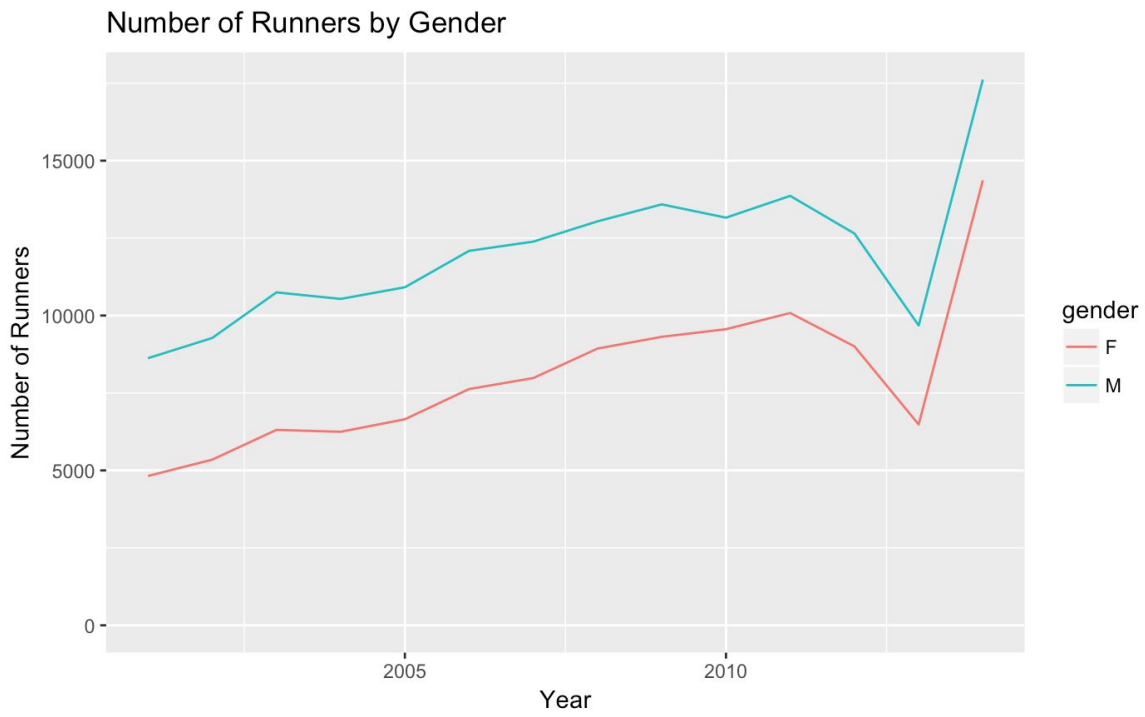
**Proposed Timeline for completing the project:**

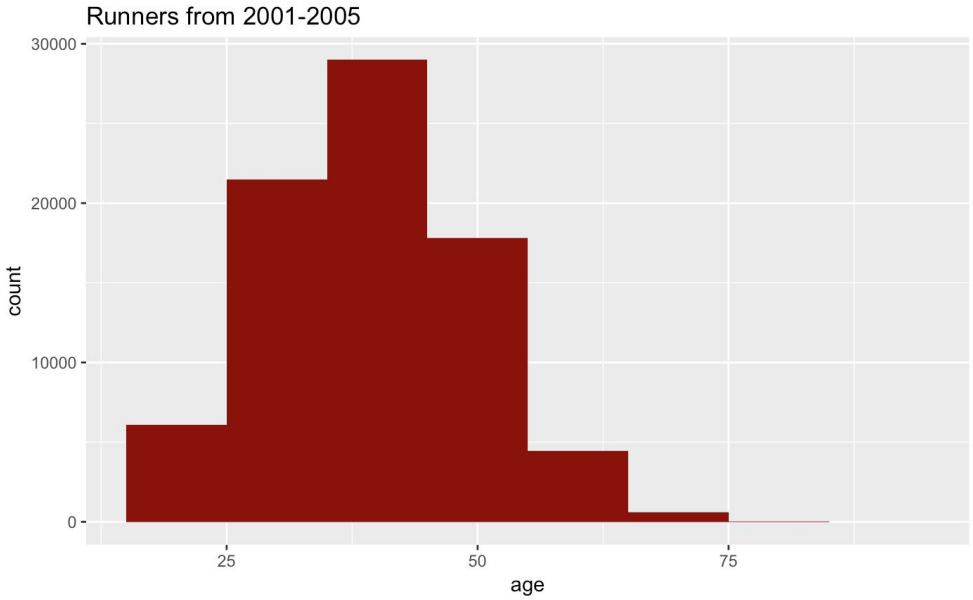
- Next week → finish creating visualizations and heat map
- Thanksgiving break → have the presentation structured and ready to present

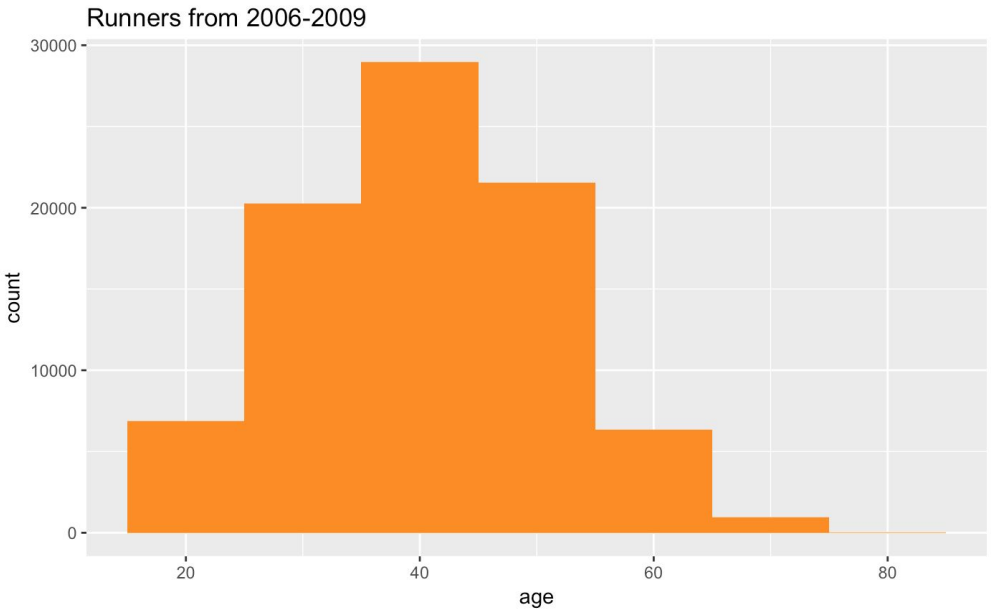
**Below are some created graphs:**

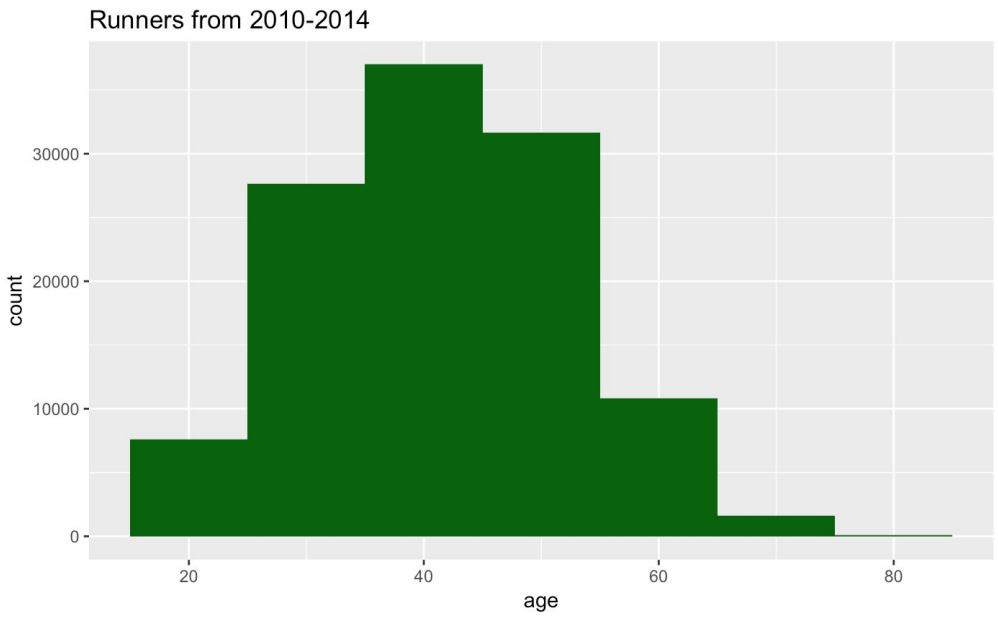

Number of Runners by Year
number_of_runners
30000
20000
10000
0
2000
2005
2010
2015
year

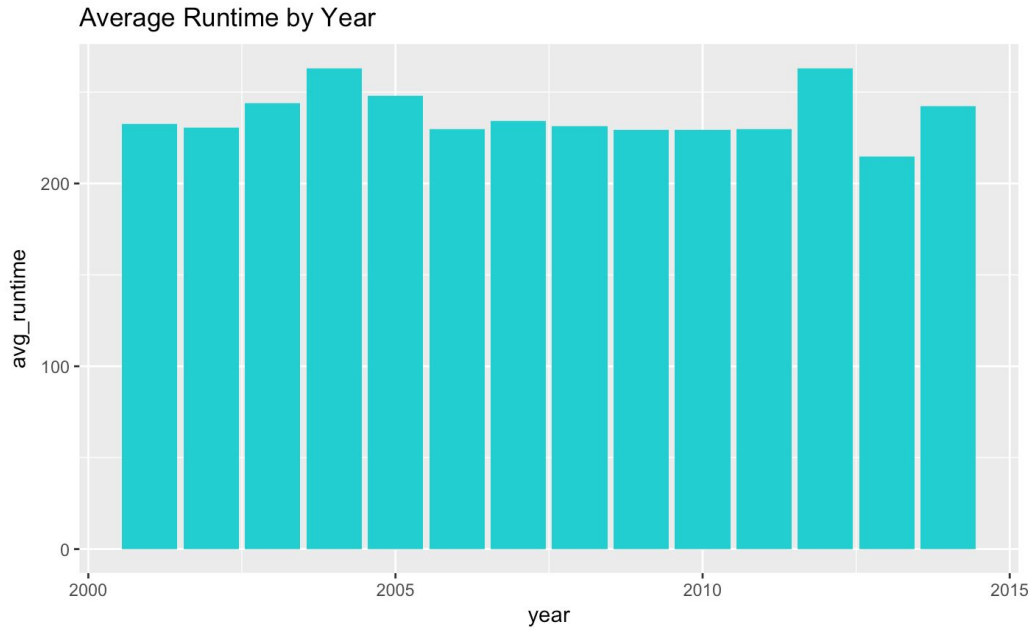

Number of Runners by Year (Difference from 2001)
difference
15000
10000
5000
0
2000
2005
2010
2015
year

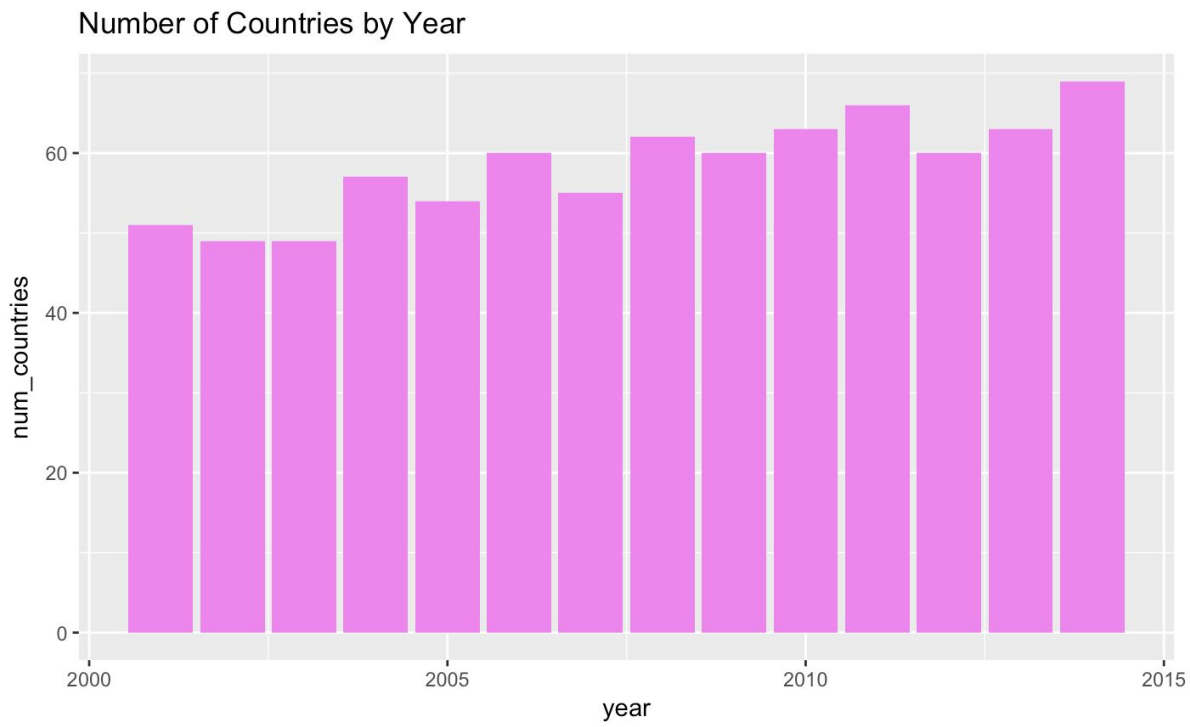

Number of Runners by Gender
15000
10000
5000
0
Number of Runners
2005
2010
Year
gender
F
M


Runners from 2001-2005
30000
20000
10000
0
count
25
50
75
age


Runners from 2006-2009
30000
20000
10000
0
count
20
40
60
80
age


Runners from 2010-2014
30000
20000
10000
0
count
20
40
60
80
age


Average Runtime by Year
200
100
0
avg_runtime
2000
2005
2010
2015
year


Number of Countries by Year
num_countries
60
40
20
0
2000
2005
2010
2015
year