




# BOSTON MARATHON FINISHERS ANALYSIS

Ruidi Chen  
Boston University



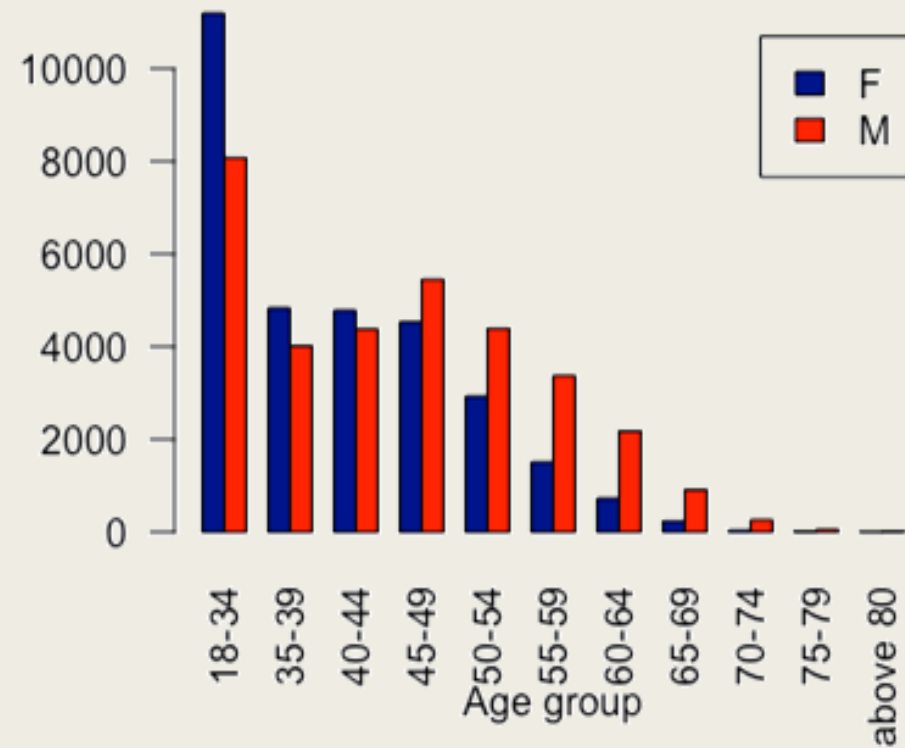
# Overview

- Data Source: <https://www.kaggle.com/rojour/boston-results>
- A list of the finishers of the Boston Marathon of 2015, 2016 & 2017
- Age, Gender, State, Runners' time at 5K, 10K, ..., 40K, Finishing time, Ranking

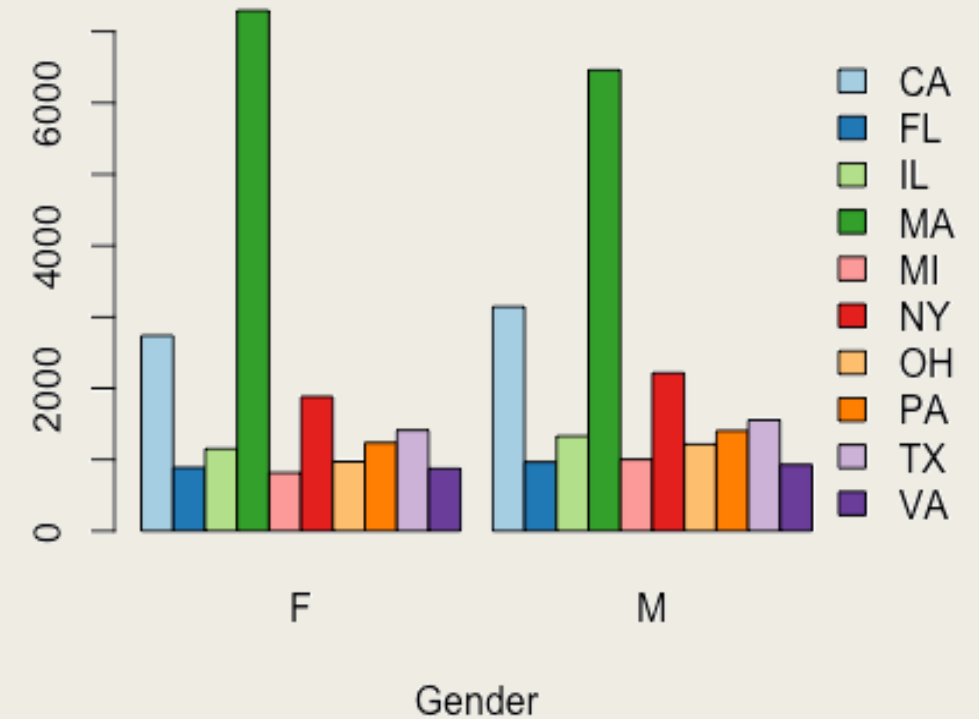


# Exploratory Data Analysis

Finisher Distribution by Age and Gender

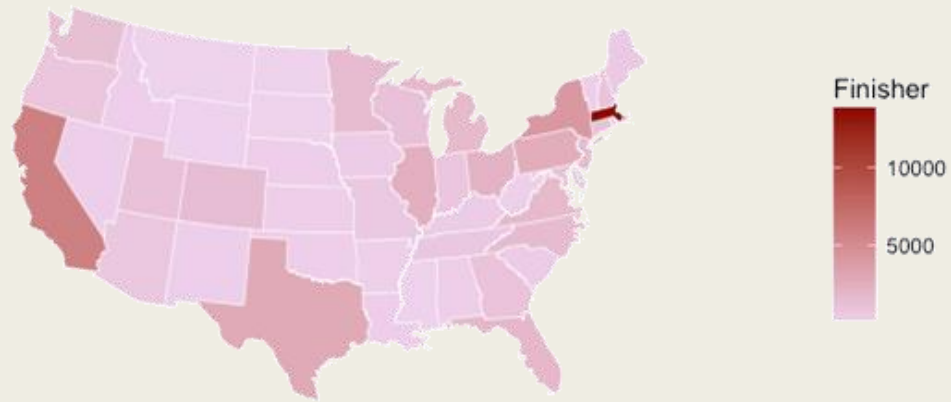


Finisher Distribution by Gender and State

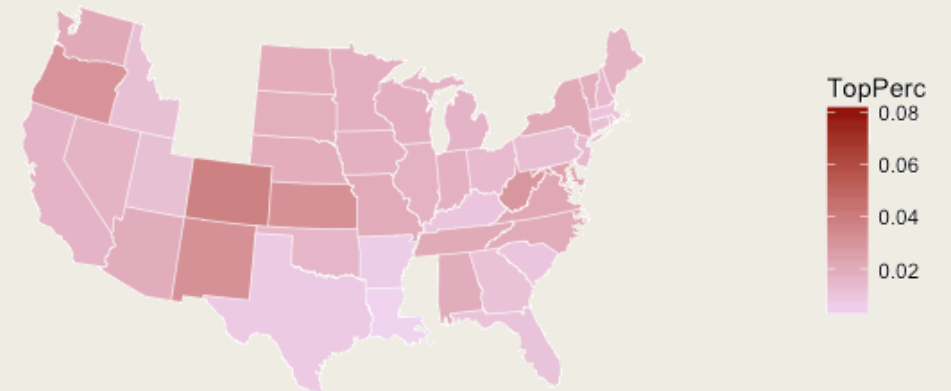


# Distribution over States

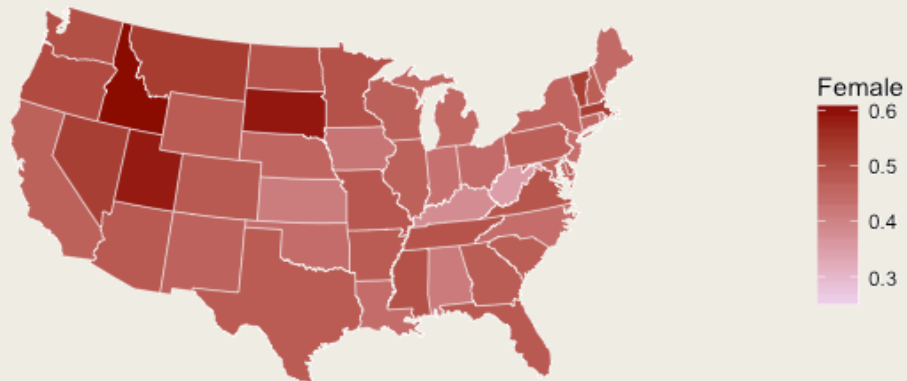
The Number of Finishers over States



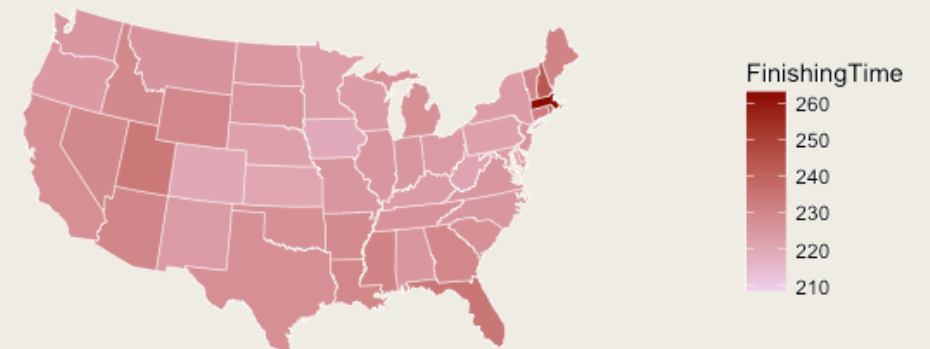
The Proportion of Top-1000 Finishers over States



The Proportion of Female Finishers over States



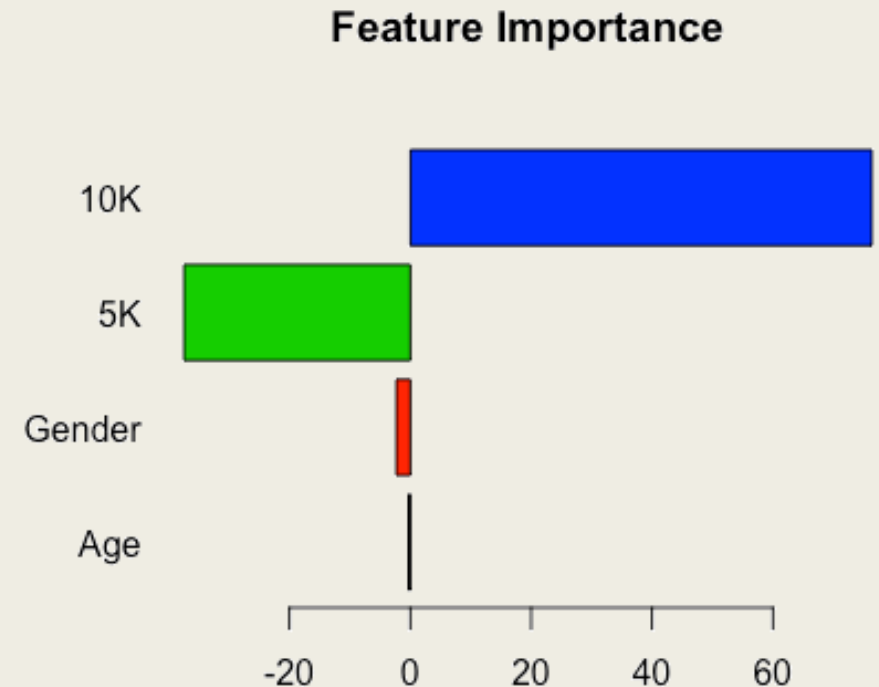
The Average Finishing Time over States



# Predict the Finishing Time

- Linear Regression, LASSO, Ridge
- Age, Gender, Time at 5K and 10K

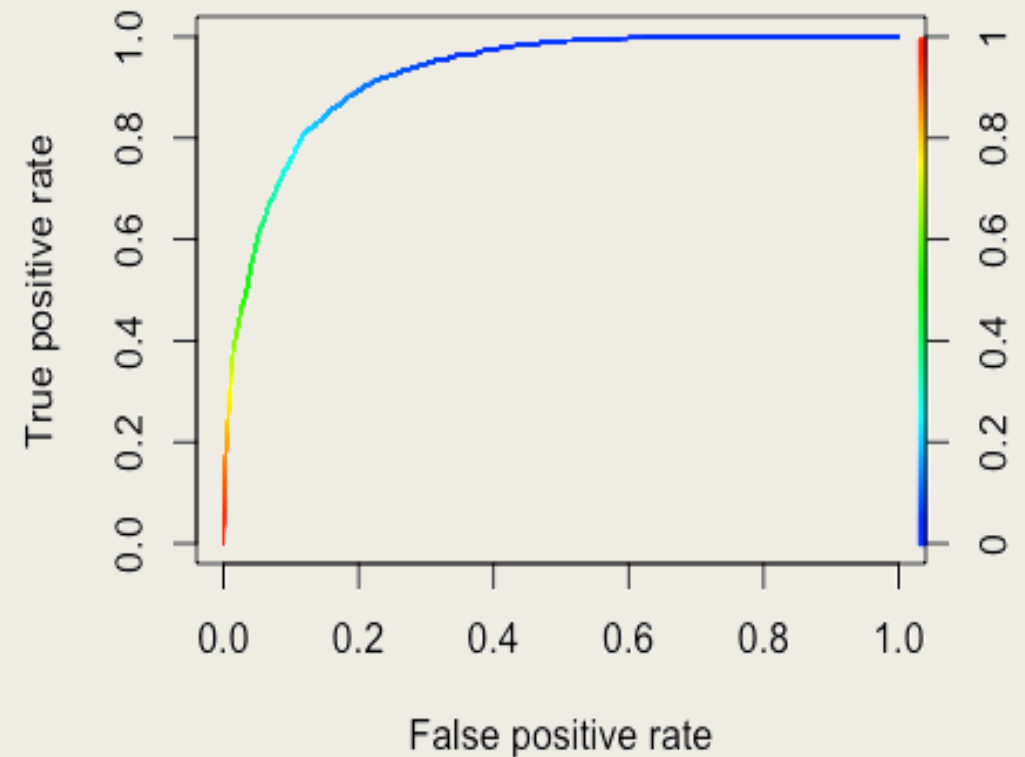
	MSE	R Square
Linear Regression	248.49	0.86
LASSO	248.76	0.86
Ridge	309.24	0.82



# Predict Ranking in the Age Division

- Predict whether in the top 10% of their age division - a **classification** problem
- Predictors: Age, Gender, Time at 5K and 10K
- Method: Logistic Regression
- Out-of-sample Accuracy: 0.89
- Out-of-sample AUC: 0.93
- Confusion matrix

Prediction \ True	FALSE	TRUE
FALSE	10331	392
TRUE	1013	1065



# Fun Facts

- More female participants in young age groups, and more male participants in older groups.
- Excluding MA, California has the most participants.
- Idaho, Utah, and Iowa have the highest percentage of female participants.
- People from Alaska, Colorado and New Mexico run fast!
- People from Massachusetts run the slowest!
- Github link: <https://github.com/Ruidi/BostonMarathon>