Team 143
Daniel Abravanel
James Drobny
Hunter Hayes
Kevin Lin
Collin Hopkins

Visualizing the Marathon

Introduction - Motivation & Problem

Runners love to analyze data, cross finish lines, and collect bling, but when it comes to visualizing race results, current organizers only show simple data tables. The goal of this project is to reimagine how race data is presented to the participants, give the participant the ability to predict podium qualifying times in different age groups, and enhance race day experience with social media analytics.

In 2018, 7.1 million race results were recorded, down 13% from 2016, *Anderson (2019)*. The past 5 years have seen a steady decline in race participation, unlike the previous 16 years of year-over-year growth. As runners have more ability to socialize workouts via wearables, race organizers have to create new ways to entice runners. From this work, the organizers will be able to gather additional information from the participants through social media to make more meaningful improvements for the following year. The above will be addressed across three topics: 1) innovative ways to visualize race results, 2) prediction of podium qualifying times per age group, and 3) the possible impact of race sentiment analysis on performance.

Survey

Race Result Visualizations

A major focus of this project was to create innovative race visualizations that can be used by race organizers. Disney Marathon 2016-2020 data was collected through APIs and web scraping, which were then transformed and prepared for visualizations. *Hanken's (2016)* focus was using data from multiple sources (weather, runner tracking, & medical tent) to create real-time runner visualizations. Although real-time tracking is out of scope for this project, this work follows a similar approach and combines data sources to create dynamic map visualizations of factors during past races. *Oliveira's (2013)* approach to visualize the tough part of the race course using heart rate data is not applicable, but this work employs the runner's pace at different splits as an alternative approach to mapping the toughest part of the course. *Smyth (2018)* discusses the impact of pacing for fast/slow starters, *Deaner (2016)* compares the marathon pacing in men and women, and *Hanely (2018)* discusses how racers tend to exhibit packing behavior. These papers' scientific analyses were helpful in building our understanding of the type of work that has already been implemented with race data, and inspired us to bring it all together into one analysis. We incorporated similar logic and thought processes into our visualizations of past races so runners could access the information they seek in one place.

Prediction of Qualifying Finish Times based on Age Groups and Environmental Factors

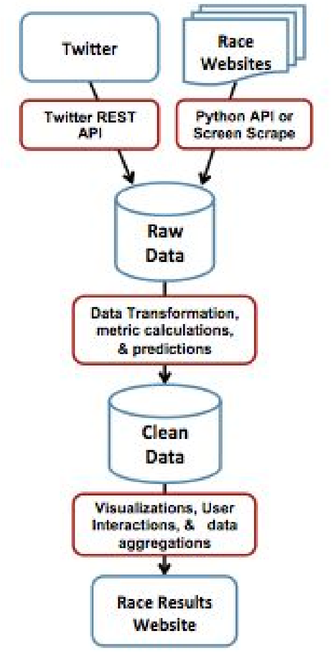
For most participants, the race is just about crossing the finish line; but for some, it is about a top 3 division finish to collect additional accolades. The top finishers' data broken down by age group and gender from 2000 - 2020 were collected using APIs and web scrapes. *Hammerling (2014)* determined the

2013 unfinished Boston Marathon results using local regression using the KNN method. A similar model was created for this work, but rather than predicting the finish times of individual runners based on their split times, we used historical race data to predict podium qualifying times for each age group. *Vihma (2010)* determined that race temperature was the most important factor, *Roach (2012)* analyzed weather, humidity, body fat, blood tests, & aerobic fitness, while *Knechtle (2014)* used multiple line regression to predict half-marathon times in male and female runners. *Nikolaidis (2018)* analyzed the Boston Marathon finishing times since 1897 and determined the importance of factors such as temperature, pressure, precipitation, wind speed, and wind direction. Due to its sensitivity, participant health data is ungatherable, which limited the predicting factors to widely available data such as temperature, humidity, previous race data, age group, division, sex, race, and race distance. All in all, however, we found a moderate correlation between qualifying finish time and a linear combination of sex, age group, temperature and humidity. As such, our visualization allows the user to vary these parameter values to predict podium qualifying times under various conditions.

Race Sentiment

The race sentiment analysis follows the workflow outlined in *Arun et al (2017)*, using Twitter REST API to collect tweets, R-studio to clean the text, and then applying a sentiment classification score to the tweet. *Orland et al (2016)* uses the twitter API to collect tweets from Tour de France cyclists during the race using the individual twitter handles. This work follows a similar approach, except data were collected at the hashtags & race's official twitter account level. *Kabir et al (2018)* classifies the positive and negative words in the tweet using the *J.Breen (2011)* positive/negative word text file. Using these word files as a starter, the team updated said file based on the words specific to race events to improve the classification. To visualize the data, *Venkatesakumar et al (2018)* displays the tweets of the fans relative to the time of the IPL Cricket Final. *Baron (2011)* further discusses the emotional state of participants during the run. Combining the approach in these papers allowed us to tie twitter users to their respective race performances, providing additional insight into sentiment vs. race results (Figure 1).

Figure 1. Project Flowchart


Twitter
Race Websites
Twitter REST API
Python API or Screen Scrape
Raw Data
Data Transformation, metric calculations, & predictions
Clean Data
Visualizations, User Interactions, & data aggregations
Race Results Website

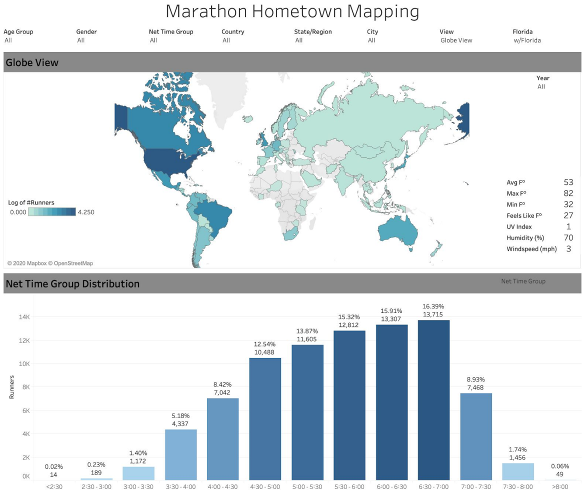
Proposed Method

Visualizations

Distribution

The race results from the Disney Marathon   2016-2020 were collected using python scrapes from PDFs and website screen scrapes. Additional columns (corral assignments, weekend races attempted, multi-year repeat runner) were added via a manually created reference table using bib number lookups and past year comparisons between name and age. Race metrics were calculated (actual start-time, racers that you passed and passed you, overall race pace, pace between race splits, and course location every 30 minutes of race) using R.

Figure 2. Tableau Race Participation Hometown



Marathon Hometown Mapping

Age Group All
Gender All
Net Time Group All
Country All
State/Region All
City All
View Globe View
Florida w/Florida

Globe View

Year All

Avg F° 53
Max F° 82
Min F° 32
Feels Like F° 27
UV Index 1
Humidity (%) 70
Windspeed (mph) 3

Supplemental race weather information (temperature, precipitation, humidity, snow indicator, wind speed, and wind direction) was collected using *worldweatheronline API (2020)* in python. Racer hometown data (altitudes, latitude, longitude, and previous months' weather conditions) were collected using the *Geo-DB Cities API(2020)* in combination with the *worldweatheronline API (2020)*. All these datasets were merged to create an aggregated dataset for visualizations at the runner-race grain.

The resulting Tableau dashboard functions as a powerful exploratory tool into historical race data. The main dashboard combines an interactive hometown map with a modular distribution chart so that the end user can seamlessly explore the data across different attributes and grains. Among the fields offered for analysis are age group, gender, and net finish bucket. Additionally, further analysis went into creatively depicting course location of all participants at a given interval, overall pace, and split pace. The participant scorecard view allows the runner to visualize how many participants they passed, how many passed them, and pace splits at different positions of the course. All of this information is provided in one simple and interactive tableau workbook, and is more insightful than the current running analyses we have seen elsewhere.

Race Results Predictions

The race predictions were determined as a multiple linear regression per age group per gender in the following manner: The data from the past Disney Marathons of the third place finishers were collected along with the supplemental weather data, and were modeled using a multiple-linear regression in R studio. All variables were compared to determine the best predictors using the AIC, BIC and the highest adjusted r-squared. Once the equations were determined, they were added to the Tableau dashboard for end users to determine the time required to finish in the top 3 of their division based on the weather conditions.

Race Sentiment

The Twitter data were collected using a public *Twitter Archive API (2020)* due to constraints with the Twitter REST API. The public archive allowed us to collect historical data based on usernames and keywords for specified time periods. Several popular hashtags related to the Disney Marathon were used to query tweets in the archive for one month around the 2016-2020 marathons. Next, the text of each tweet was pre-processed by removing punctuation, usernames, hyperlinks, and other irrelevant text. Stop words (e.g. "the", "a",etc.) from the NLTK library in Python were removed from the text. Each remaining word of the text was separated for further analysis. With the pre-processed data, *NRC Emotion Lexicon (2020)* was utilized to evaluate the associated emotion around each tweet based on the words used. A dynamic word cloud and accompanying bar chart were built in Tableau for use on each emotion in order to understand which words are most common/essential and to evaluate sentiment trends year over year.

The team also attempted to connect specific runners to their tweets by using the FuzzyWuzzy package in Python and joining the twitter usernames to runner names. The likelihood of a username matching a runner name was determined, and if it met or exceeded a Levenshtein distance score of 80, the previous 6 months of tweets for that username were used to evaluate sentiment of the runner leading up to the race. This data was then combined with the race results to examine sentiment in closer detail by runner and helped determine whether sentiment is correlated with race results. In sum, no correlation between sentiment and final runtime was found, as twitter sentiment could not be properly gauged for all participants. In the future, we propose collecting social media information on the runners either by offering incentives to do so or gathering that data from elsewhere.

Experiments

Questions to be Answered via Analysis

- Can we accurately predict race qualifying finish times based on the information we gathered?
- Do runners from different countries/states perform statistically better as a group?
- Do runners who train in harsher conditions (i.e. higher elevation, more extreme temperatures) perform better in the actual marathon?
- Do pre or mid-race sentiment significantly affect final runtime?
- What is the overall sentiment of each race?

Race Results Visualizations

The first visualization of the race results was attempted in D3. After a two week trial period with this large data set, it was determined that in the time frame required for the project, it was best to use a visualization tool with which the team was more familiar. The team agreed to switch to Tableau.

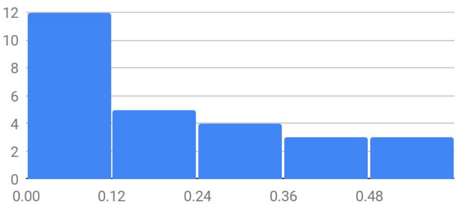
Through our use of Tableau, we found that most of the participants' hometowns were in the United States, and most of the US runners were from Florida. Whether this is realistic or includes some user-input error was not questioned -- rather a parameter to include/exclude Florida from the results was added to the dashboard. It was interesting also to see that the next countries in terms of representation were Canada, Brazil and Australia, but that Russia and China were heavily underrepresented when considering their population sizes.

Our visualization of the age category distribution also portrayed a heavier participation rate for the 35-39 age group, and a gradual drop-off to each side with a right skew. Interestingly, female runners exceeded male runners in all years with the exception of 2019.

Moreover, we built the dashboard in a linear fashion so that it can be followed like a story -- The first tab covers summary statistics for the race population, and through a combination of filter and parameter actions, the end user can progress through the dashboard into the more granular runner data. For example, if a user were interested in the 35-39 age group data, they could click on that bar in the distribution chart and menu-navigate to the runner details specific to those runners. From there, the user can once again click on a specific runner and menu-navigate to that runner's scorecard. On said scorecard, a user can see all races in which they competed, their relevant metrics, as well as their average pace, passing metrics, and visualized split times throughout the race.

Race Results Predictions

Figure 3: Adj. R^2 Distribution for the 28 models

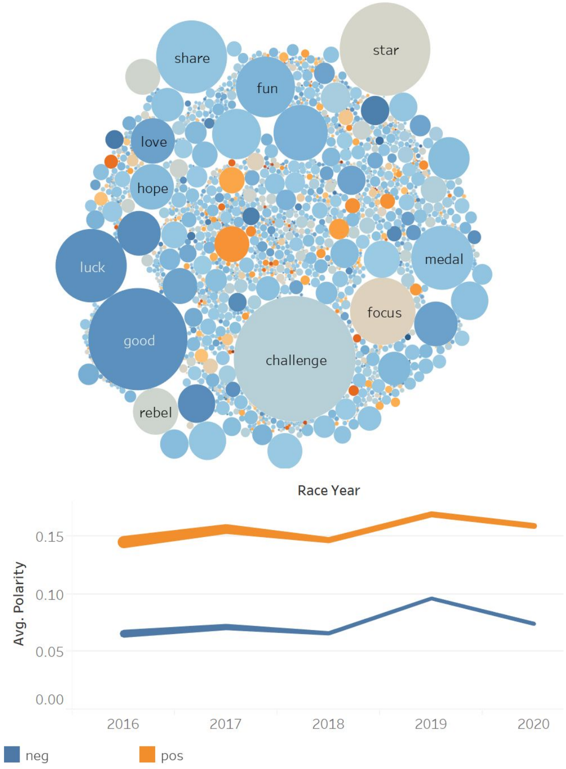


The goal was to accurately predict the 3rd place finisher's time in next year's race. Using R-Studio, 20 years of third place finishers' times were run using a stepwise regression. The first attempt included all the race divisions and genders in one model. This model achieved an adjusted-r squared of 0.94 which may have been artificially inflated due to age group and gender being included as attributes. To focus the predicting variables on environmental factors, a different model was created for each division (age group and gender). A stepwise regression was run on the data using 10 different historical race weather values (multiple temperatures, windspeed, humidity). The adjusted R-squared values were between 0 - 0.59 with Division Female 35-39 producing the best results.

The ability to gather more health related data (height, weight, body fat percentage) along with more training data (miles run per week or past race results per runner) would allow for a better prediction model.

Race Sentiment

Figure 4: Twitter Word Cloud & Sentiment Scores



The team used the Twitter data associated with Disney Marathon keywords to explore sentiment, emotion association, and relationship to race finish time. The tweets were pre-processed to remove punctuation, stop words, and other invaluable information. A Naive Bayes classifier was then used to estimate polarity and subjectivity of each tweet. Initial examination of positive and negative classifications of tweets showed that most tweets used similar words, regardless of sentiment. In order to further explore the data, each word was matched to an emotional association to get an overall emotional association of the tweets. Hashtags were also separated from the tweets to examine most frequently used hashtags based on emotion or sentiment classification.

The analysis shows that the majority of words with emotional association are positive, with the top emotions being anticipation, joy and trust. The least common emotions are disgust, sadness and anger. Word clouds were used to provide visibility to most commonly used words for the sentiment and emotional association of runners.

Tying specific tweets to actual runners proved to be difficult. The team attempted to match runners based on runner name and twitter handle of a tweet, but there was not enough information available to verify the match between twitter accounts and runners. Accomplishing this task requires finding additional criteria to match on which are not publicly available due to Twitter privacy restrictions. With the matched data, the team combined sentiment of users with race data to determine whether there was a relationship between sentiment and race finish time. The team used random forest regression and linear regression models for the prediction. The models showed that there was an improvement in overall r-squared value when using runner sentiment, but because the matches to runners were not verifiable, the team did not use these variables in the final model.

The team also explored predicting sentiment for runners based on race data. Due to the difficulty of matching twitter usernames to runner names, this prediction was not successful. Lastly, the team attempted to predict overall race sentiment based on data around the race, but due to low sample sizes of races, the model was not effective.

Conclusion

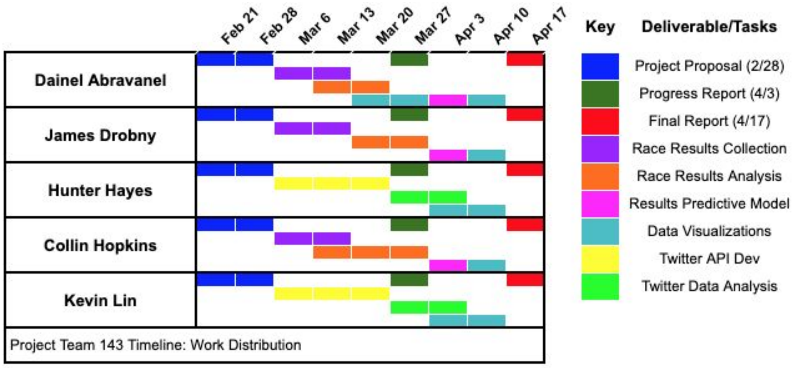
Our work delves into both a summary visualization of the race data as well as a deeper dive into the disparate variables that could help predict runner performance. From our research, we did not find a visualization project that effectively analyzes marathon data over multiple years with return runners, and also incorporates weather, geography and sentiment. Thus, we hope this work will provide a more holistic perspective on any marathon data. In fact, while we did run this analysis on marathon data for the Walt Disney World races from 2016-2020, this analysis can be recreated for any marathon or list of marathons to be compared. Moreover, the visualization introduced in the project is in and of itself innovative as it provides interactive views of different grains by which the end user can dig into race-year and runner-year grains seamlessly.

The twitter sentiment analysis attempted to explain runner performance with pre or mid-race sentiment. Our attempt to match twitter handles to runner names by similarity score [FuzzyWuzzy package in python] did not yield the best results, but did manage to correctly match to approximately 1-2% of runners per year. Race time prediction was done using a random forest regression model with scikit-learn and did not yield a meaningful correlation with an R squared value of 0.67. If we were to run this analysis for other races in the future, it would be imperative to ask for runners' twitter handles if they would like to see the predictive component of the analysis on their run.

Project Work Distribution

The team was split into 2 separate teams: the twitter sentiment team and the race data team. All the individuals did the work required during the project.

Figure 5: Deliverables and Work Distribution



Works Cited

1. Andersen, Jens Jakob. “The State of Running 2019.” *RunRepeat*, 15 Jan. 2020, runrepeat.com/state-of-running.

2. Arun, K., Srinagesh, A., & Ramesh, M. (2017). Twitter sentiment analysis on demonetization tweets in India using R language. *International Journal of Computer Engineering and Information Technology, 9*(6), 119-124.

3. Baron B. , Moullan F. , Deruelle F. and Noakes T.D. , 2011, The role of emotions on pacing strategies and performance in middle and long duration sport events, British Journal of Sports Medicine 45(6), www.ncbi.nlm.nih.gov/pubmed/19553226.

4. Deaner R.O. , Carter R.E. , Joyner M.J. and Hunter S.K. , 2014, Men are more likely than women to slow in the marathon, Medicine and Science in Sports and Exercise 4(3), 607–616.

5. G. Oliveira, J. Comba, R. Torchelsen, M. Padilha and C. Silva, "Visualizing Running Races through the Multivariate Time-Series of Multiple Runners," 2013 XXVI Conference on Graphics, Patterns and Images, Arequipa, 2013, pp. 99-106.

6. Hammerling, D., Cefalu, M., Cisewski, J., Dominici, F., Parmigiani, G., Paulson, C., & Smith, R. L. (2014). Completing the results of the 2013 Boston marathon. *PLoS One, 9*(4)

7. Hanken T , Young S , Smilowitz K , Chiampas G , Waskowski D . Developing a Data Visualization System for the Bank of America Chicago Marathon (Chicago, Illinois USA). *Prehosp Disaster Med*. 2016;31(5):572–577

8. Hanley, B. (2016). Pacing, packing and sex-based differences in olympic and IAAF world championship marathons. *Journal of Sports Sciences, 34*(17), 1675-1681.

9. J. Breen. positive-words.txt [Online]. Available: https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial201107/blob/master/data/opinion-lexicon-English/positive-words.txt

10. J. Breen. negative-words.txt [Online]. Available: https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial201107/blob/master/data/opinion-lexicon-English/negative-words.txt

11. Kabir, A. I., Karim, R., Newaz, S., & Hossain, M. I. (2018). The power of social media analytics: Text analytics based on sentiment analysis and word clouds on R. *Informatica Economica, 22*(1), 25-38.

12. Knechtle, B., Barandun, U., Knechtle, P., Zingg, M. A., Rosemann, T., & Rüst, C.,A. (2014). Prediction of half-marathon race time in recreational female and male runners. *SpringerPlus, 3*(1), 1-8.

13. Nikolaidis, P. T., Gangi, S. D., Chtourou, H., Rüst, C. A., Rosemann, T., & Knechtle, B. (2019). The role of environmental conditions on marathon running performance in men competing in the Boston Marathon from 1897 to 2018. *International Journal of Environmental Research and Public Health, 16*(4)

14. Orland Hoeber, Larena Hoeber, Maha El Meseery, Kenneth Odoh, Radhika Gopi, (2016) "Visual Twitter Analytics (Vista): Temporally changing sentiment and the discovery of emergent themes within sport event tweets", Online Information Review, Vol. 40 Issue: 1, pp.25-41, https://doi.org/10.1108/OIR-02-2015-0067

15. Roach, L. E. (2012). *Thermoregulation and marathon performance: Relationships of predictability of marathon performance, ambient weather conditions, BSAM:MT, BSA:ML, percent body fat, and aerobic fitness (Order No. 1532041).*

16. Smyth, Barry. “Fast Starters and Slow Finishers: A Large-Scale Data Analysis of Pacing at the Beginning and End of the Marathon for Recreational Runners.” *Journal of Sports Analytics*, IOS Press, 22 Aug. 2018, content.iospress.com/articles/journal-of-sports-analytics/jsa205.

17. Venkatesakumar, R., & Pachayappan, M. (2018). Sentiment analysis of tweets- IPL 2018 final. *SCMS Journal of Indian Management, 15*(3), 71-80.

18. Vihma, T. (2010). Effects of weather on the performance of marathon runners. *International Journal of Biometeorology, 54*(3), 297-306.

19. Geo-DB Cities API. wirefreethought. https://rapidapi.com/wirefreethought/api/geodb-cities.

20. Weather API. worldweatheronline. http://api.worldweatheronline.com/premium/v1/past-weather.

21. Twitter Archive API. https://github.com/marquisvictor/Optimized-Modified-GetOldTweets3-OMGOT.

22. NRC Emotion Lexicon. http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.