ISyE 6740 – Spring 2023
Project Report

Team Member Names: Caleb Rauscher crauscher6@gatech.edu

Project Title: Marathon Finish Time Prediction

Problem Statement

There are about 1 million runners per year attempting to complete a full marathon. Running a full marathon requires an individual to run a 26.2 mile / 42.2 km course. To prepare for a full marathon a typical training plan encompasses 3-7 days of running per week for 16 weeks. Due to the physical demands of running a full marathon physical and mental recovery is needed between finishing a marathon and the next training cycle. With the amount of training that is required and the time for recovery that is required it is difficult to fully train for more than 1 or 2 marathons a year. Many runners will participate in numerous marathons per year but to fully prepare to run a personal best marathon an individual needs to focus on a proper training cycle and fully recover before training for the next race. Given this timeframe there are not many opportunities for feedback on what worked well and what did not work well in each training cycle. This is a reason why you can see individuals closer to 40 years old than 20 years old winning major marathons. It takes years of trial and error to improve on race finish times. Shorter distance races are typically dominated by individuals closer to 20 years old, for shorter distance races there is less build up in training and recovery needed so variations in training can be attempted and evaluated frequently. The goal of this report is to evaluate which factors in marathon training impact the finish time the most.

The idea is to determine if high mileage, faster pace, more days running, or some other feature contributes more positively or negatively to the final race time. It would appear obvious that running more and running at a faster pace will improve the race time the most but increasing mileage and pace increases the risk of running related injuries and overtraining. Many amateur marathoners approach race day over trained, which means due to excessive training their body has not fully recovered and they will not be able to perform their best. Another issue with overtraining is the possibility of an injury. Injuries could pop up at any time in the training cycle. Injuries early on could result in lost workouts which could lead to less training time or an attempt to overcome the lost workouts which could lead to more injuries. Injuries that occur late in a training cycle could carry into the race which will reduce the ability to run at full potential.

Data

With the high usage of GPS enabled running watches and the sharing of this data to various websites such as Garmin Connect, Strava, Runkeeper, and MapMyRun there is an increase in available training and racing data. To obtain the marathon training data, publicly reported data on Strava will be used. Existing data has already been collected (Alfonseca, Watanabe, & Duarte, 2022) and will be used to identify marathon training cycles. To identify the training cycle an individual that has completed a run between 41.8 km and 42.5 km will be used as the ending point of the race and then looking back over the previous 16 weeks to obtain the data for the training cycle. Since a marathon is measured as the shortest distance from the starting line to the finish line and it is very difficult or impossible to run that exact line due to other runners being on the course a buffer is used to determine the exact distance of the race.

Typical marathon training plans are 16 weeks with each week slightly increasing in duration and/or effort. This buildup lasts for 2-3 weeks then a recovery week occurs which reduces the mileage and the process repeats itself. The last 3 weeks from the marathon date the taper begins where the mileage reduces to allow the body to recover for the race.

The original data provides:

- **Datetime**: the date of the running activity
- **Athlete**: an integer to identify the runner and provide anonymity
- **Distance**: the total distance of the run, in kilometers
- **Duration:** the total time of the run, in minutes
- **Gender:** “M” or “F”
- **Age Group:** age interval: 18-34, 35-44, 55+
- **Country:** country of origin of the athlete
- **Major:** list of major marathons the runner has run

Since the original data was used for all types of runners the first step was to remove runners that did not complete a run between 41.8 km and 42.5 km. Exploration of the data showed that there were finish times far longer than a typical marathon cutoff time. This could be marathoners that had physical difficulties or an extreme marathon that was off-road with significant elevation gain or other environmental factors. Any marathons finish times greater than 8 hours were removed from the data set. Each data point represents a single running event, since the goal is to evaluate a training plan the runner was grouped by athlete and the preceding 16 weeks from a marathon distance run. This resulted in some runners having multiple training cycles. With the data grouped by athlete and training week the following features were added:

- **Minimum Distance**: Shortest run of the week in kilometers
- **Maximum Distance**: Longest run of the week in kilometers
- **Mean Distance**: Average run of the week in kilometers
- **Total Distance**: Cumulative kilometers of the week
- **Minimum Duration**: Shortest run of the week in minutes
- **Maximum Duration**: Longest run of the week in minutes
- **Mean Duration**: Average run of the week in minutes
- **Slowest Pace**: Slowest run of the week in % of race pace
- **Fastest Pace**: Fastest run of the week in % of race pace
- **Average Pace**: Average run of the week in % of race pace
- **Number of runs**: Total number of runs of the week

With the above features added for each week the data set used for this report resulted in a datapoint for each runner/training cycle with each feature above and the week number. This results in 16 weeks * 11 features = 176 features. Additional categorical features of **Gender**, **Age Group**, and, **Country** were also added for a total of 180 features.

Examining the data

An evaluation of the breakdown of the dataset by the categorical features yields the following figures. Figure 1 shows the breakdown of **Male** and **Female** runners. There are significantly more males in this dataset.

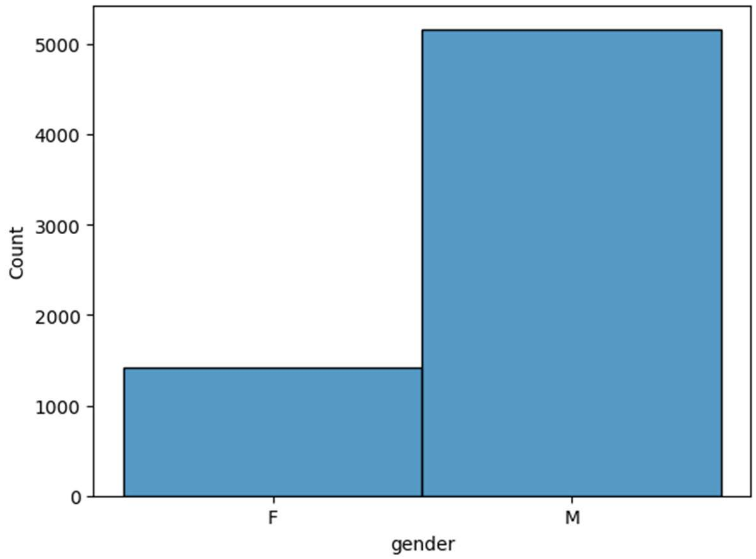

Figure 1: Gender

Figure 2 shows the breakdown by **age_group**, ages 35-54 are the most representative group with significantly more than ages 18-34 which is significantly more than the 55+ age group.

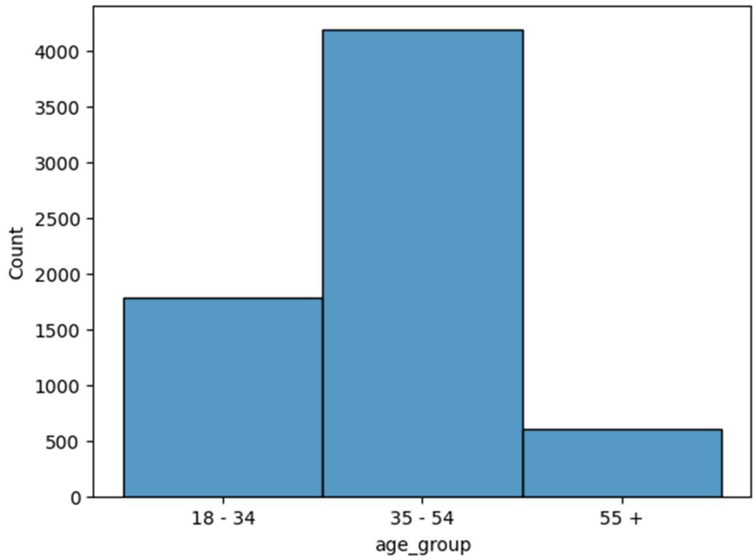

Figure 2: Age Group

Figure 3 shows the representation of each country in the data set. For visibility only the top 20 represented countries are shown in Figure 3. The United States, United Kingdom, and Germany are the majority in this data set. This is likely more of an indication of Strava's adoption in those countries than the number of runners in the countries.

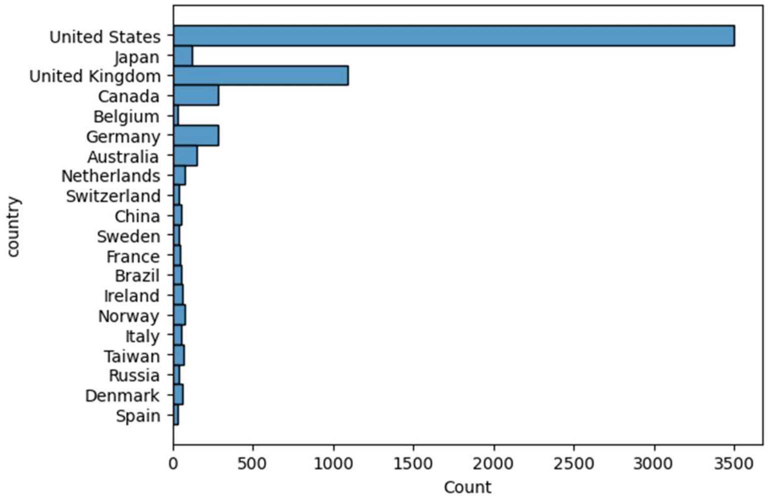

Figure 3: Country

Of the 6,583 collected marathon distanced runs the distance, duration, and pace data is described in Table 1. Since the marathon runs are collected as 41.8 km to 42.5 km it is not surprising that the distances are captured so closely. It is interesting that the minimum duration is 2 hours and 12 minutes which is a very competitive finish time while the 75th percentile run is 4 hours and 2 minutes which is a finish time many casual marathoners would be happy with. The maximum finish time of 7 hours and 51 minutes is likely a finisher that walked the course.

Table 1: Marathon event data

	Distance	**Duration**	**Pace**
Mean	42.28 km	221.13 minutes	5:14 minutes per km
Standard Deviation	0.148 km	47.56 minutes	1:07 minute per km
Minimum	41.8 km	132.0 minutes	3:07 minutes per km
25th percentile	42.2 km	188.0 minutes	4:26 minutes per km
50th percentile	42.29 km	211.57 minutes	5:00 minutes per km
75th percentile	42.4 km	242.0 minutes	5:44 minutes per km
Maximum	42.49 km	471.0 minutes	11:08 minutes per km

The histogram of finish time in Figure 4 indicates close to a normal distribution with a skew to the right that accounts for slower runners.

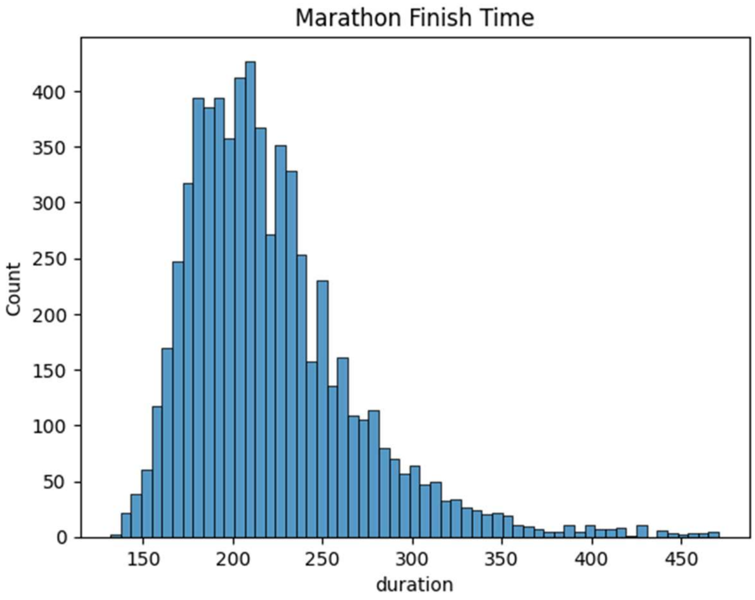

Figure 4: Histogram of Marathon Finish Time (minutes)

Figure 5 shows a box plot of age_group vs finish time. It is interesting to see in the 55+ age_group there are a larger spread of women in the 25th – 75th quartile range. There are numerous runners on the slower side of the lower whisker but this is typical of marathons with a number of elite and competitive runners along with many recreational/social runners.

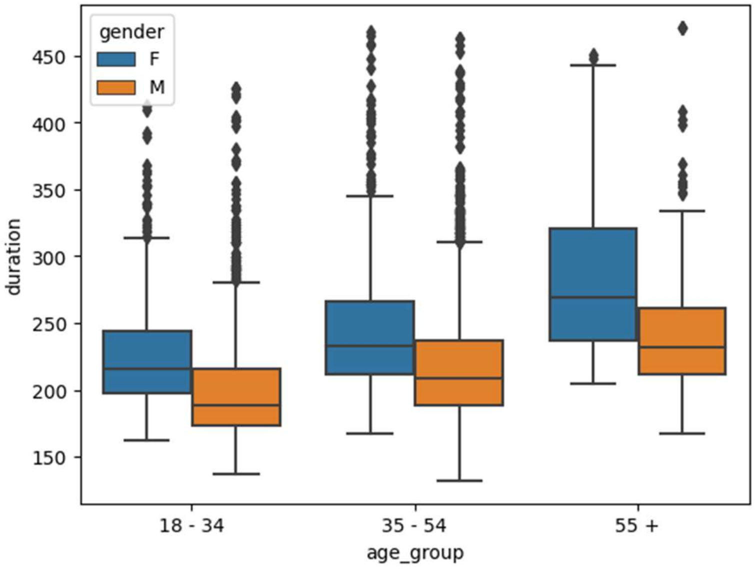

Figure 5: Boxplot of age groups

Methodology

There are many popular marathon training styles that exist which have their main focus on high mileage, faster pace, or an 80/20 mix, which encompasses running most miles at an easier pace with some faster workouts. As a first step a clustering algorithm of the training data will be performed to identify different training styles.

Clustering

The K-Means algorithm is used to identify different training styles. The number of clusters (K) is varied from 1 through 10 to identify the best number of clusters. Both the distortions and the inertia are used in elbow plots to identify the ideal number of clusters to choose. The distortion for each K value is calculated as the average of the squared distances from the cluster center of each center. The inertia is the sum of squared distances of samples to their closest cluster center. Both elbow plots don't have a very clear indication of the K value but comparing both a K value of 5 or 6 looks to be reasonable.

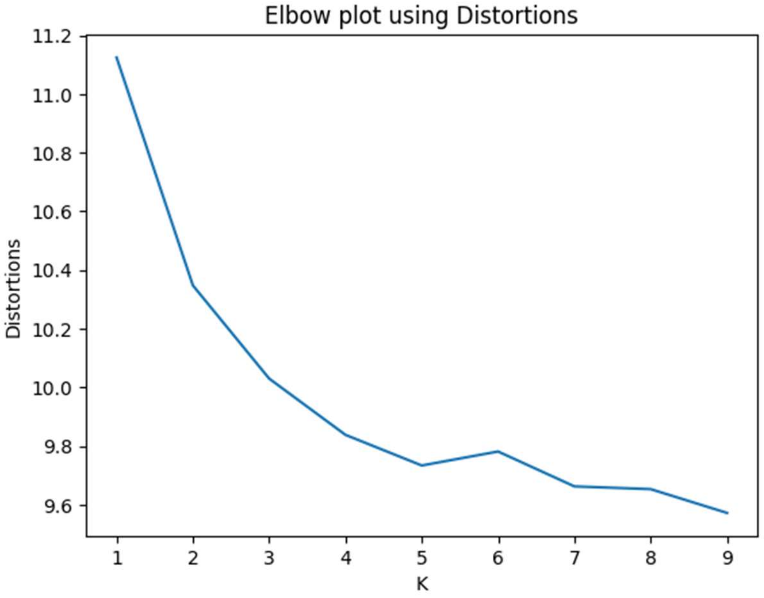

Figure 6: Elbow Plot with Distortion

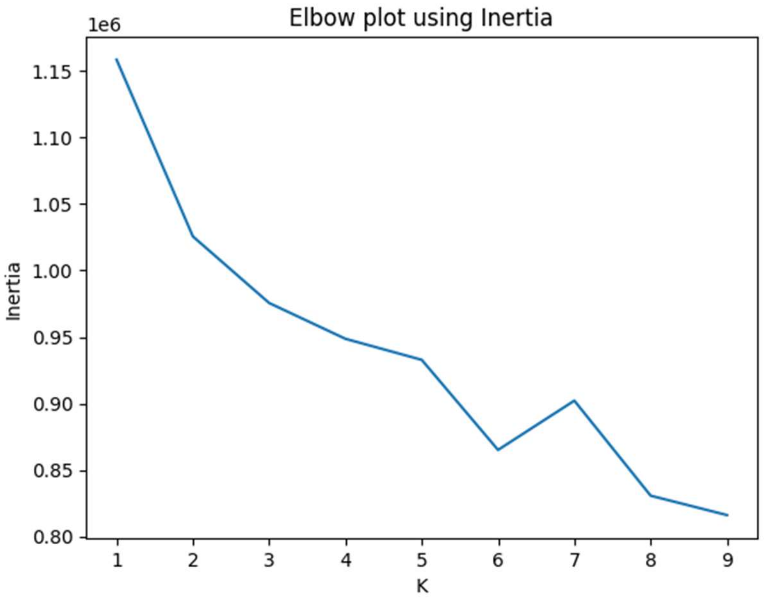

Figure 7: Elbow plot using Inertia

Principal Component Analysis is used to visualize the data with k=5 and k=6. In the k=5 and k=6 plots there are 2 data points that are high on the y axis that are their own cluster. The data points don't indicate anything unusual so it does not appear that they are outliers.

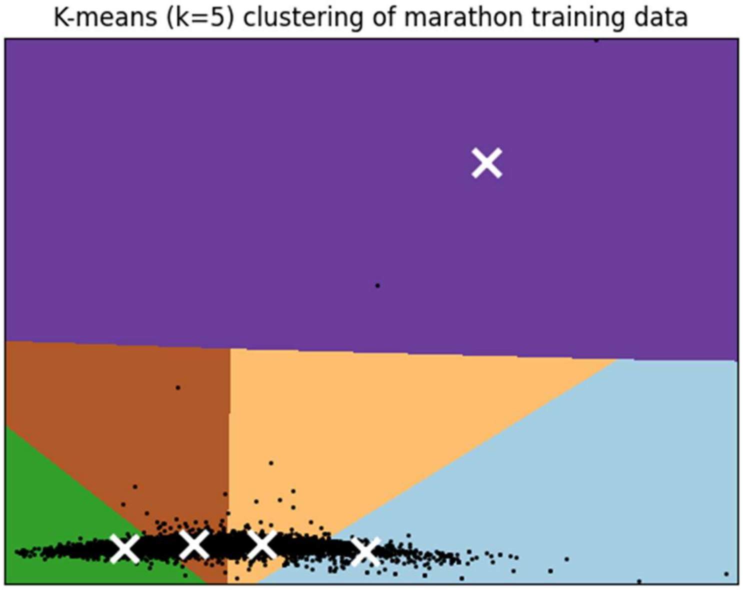

Figure 8: PCA K-means cluster k=5

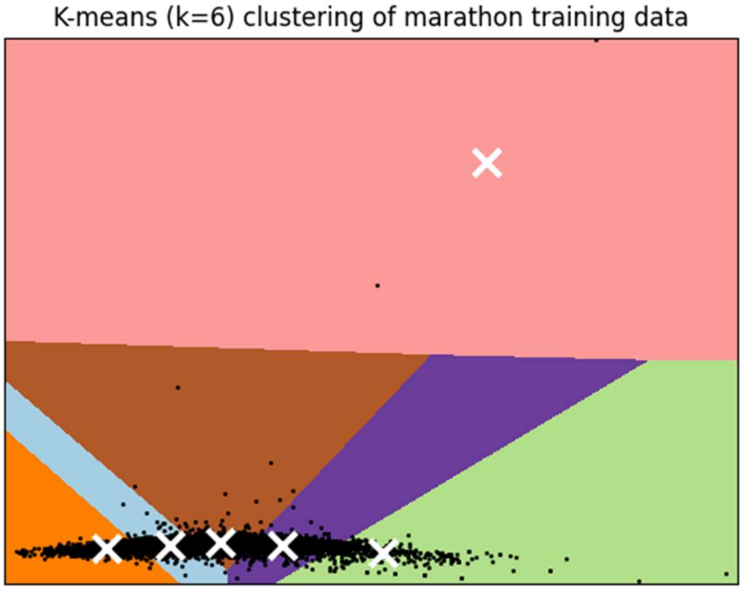

Figure 9: PCA K-means cluster k=6

Feature Selection

Principal Component Analysis (PCA) is used to determine which features are the most important. The loadings of the components in PCA identify which variables explain the variation the most. The most variation is explained in the total distance of the upper middle weeks of the training cycle. Assuming the last 3 weeks are the taper period where mileage is reduced to allow the body to recover for the race then weeks 6 through 13 would be the peak training weeks. The earlier portion of this period is shown to explain the variability the most. Not surprisingly the pace on the first and last week of the training cycle has little to do with explaining the variability.

Table 3: PCA Loadings

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
week_8_total_distance	0.13	0.00	-0.07	-0.01	-0.06	0.05	-0.09	-0.02	-0.09	-0.05
week_7_total_distance	0.13	0.00	-0.08	0.02	-0.02	0.06	0.00	0.00	-0.07	-0.04
week_4_total_distance	0.13	0.00	-0.09	0.06	0.05	0.00	0.02	0.00	0.06	0.06
week_9_total_distance	0.13	0.00	-0.07	-0.04	-0.06	0.05	-0.02	0.07	0.02	-0.02
week_6_total_distance	0.13	0.01	-0.09	0.03	-0.03	0.07	0.10	-0.02	-0.02	0.01
…	…	…	…	…	…	…	…	…	…	…
week_16_fastest_pace	0.01	0.05	0.02	0.02	-0.06	-0.10	-0.02	-0.08	0.05	-0.01
week_16_average_pace	0.01	0.16	0.10	-0.02	0.06	0.06	0.05	0.05	-0.08	0.07
week_7_average_pace	0.01	0.02	0.01	0.00	-0.01	0.02	-0.04	-0.04	-0.01	0.07
week_7_fastest_pace	0.00	0.01	0.00	0.00	-0.01	0.01	-0.05	-0.04	-0.01	0.06
week_1_fastest_pace	0.00	0.04	0.03	0.00	-0.02	-0.04	0.00	-0.02	0.00	0.02

The explained variance principal component 1 is 20% then starts to decrease exponentially. The cumulative explained variance begins to plateau around principal component 75 to 100.

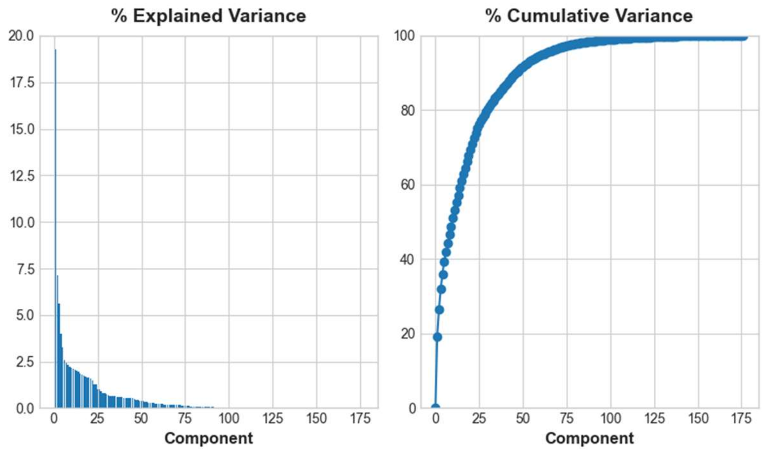

Figure 10: Principal Component explained variance

Cross Validation of a Lasso Regression model was used to find which features predict the finish time the best. The features with a coefficient of 0 will be removed from the prediction model. Cross Validation yielded a best alpha value of 0.05 with an R^2 value of 0.59. Features that Lasso Regression removed are grayed out in Table 4 and the remaining features are left as is. Key observations are in the first half of the training cycle average duration and fastest pace are the main features, in the second half of the training cycle the main features are total distance, longest time, and fastest pace. This aligns with most training plans that focus on speed in the beginning while the accumulated distance increases and the later weeks were longer runs are key to building up endurance.

Table 4: Features

Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
total distance	total distance	total distance	total distance	total distance	total distance	total distance	total distance
longest distance	longest distance	longest distance	longest distance	longest distance	longest distance	longest distance	longest distance
average distance	average distance	average distance	average distance	average distance	average distance	average distance	average distance
shortest distance	shortest distance	shortest distance	shortest distance	shortest distance	shortest distance	shortest distance	shortest distance
shortest time	shortest time	shortest time	shortest time	shortest time	shortest time	shortest time	shortest time
longest time	longest time	longest time	longest time	longest time	longest time	longest time	longest time
average duration	average duration	average duration	average duration	average duration	average duration	average duration	average duration
fastest pace	fastest pace	fastest pace	fastest pace	fastest pace	fastest pace	fastest pace	fastest pace
slowest pace	slowest pace	slowest pace	slowest pace	slowest pace	slowest pace	slowest pace	slowest pace
average pace	average pace	average pace	average pace	average pace	average pace	average pace	average pace
total runs	total runs	total runs	total runs	total runs	total runs	total runs	total runs

Week 9	Week 10	Week 11	Week 12	Week 13	Week 14	Week 15	Week 16
total distance	total distance	total distance	total distance	total distance	total distance	total distance	total distance
longest distance	longest distance	longest distance	longest distance	longest distance	longest distance	longest distance	longest distance
average distance	average distance	average distance	average distance	average distance	average distance	average distance	average distance
shortest distance	shortest distance	shortest distance	shortest distance	shortest distance	shortest distance	shortest distance	shortest distance
shortest time	shortest time	shortest time	shortest time	shortest time	shortest time	shortest time	shortest time
longest time	longest time	longest time	longest time	longest time	longest time	longest time	longest time
average duration	average duration	average duration	average duration	average duration	average duration	average duration	average duration
fastest pace	fastest pace	fastest pace	fastest pace	fastest pace	fastest pace	fastest pace	fastest pace
slowest pace	slowest pace	slowest pace	slowest pace	slowest pace	slowest pace	slowest pace	slowest pace
average pace	average pace	average pace	average pace	average pace	average pace	average pace	average pace
total runs	total runs	total runs	total runs	total runs	total runs	total runs	total runs

With these training styles identified each cluster will use supervised learning algorithms to predict the marathon finish time per cluster. The first model used is a Lasso Regression model using the features defined in Table 4 with cross validation of 10 folds. The best model selected was with an alpha of 0.06, and R^2 of 0.55 and the coefficients as:

```
-0.04 * week_5_total_distance
-0.0 * week_6_total_distance
-0.01 * week_9_total_distance
-0.0 * week_10_total_distance
-0.02 * week_11_total_distance
-0.04 * week_13_total_distance
-0.03 * week_14_total_distance
-0.05 * week_15_total_distance
-0.11 * week_1_average_distance
-0.01 * week_7_average_distance
-0.02 * week_10_average_distance
-0.0 * week_11_average_distance
-0.02 * week_12_average_distance
-0.0 * week_13_average_distance
-0.1 * week_14_average_distance
-0.07 * week_16_average_distance
-0.0 * week_1_shortest_distance
0.01 * week_2_longest_time
0.01 * week_3_longest_time
0.03 * week_6_longest_time
0.03 * week_7_longest_time
0.0 * week_9_longest_time
0.03 * week_10_longest_time
0.03 * week_11_longest_time
0.03 * week_12_longest_time
0.03 * week_13_longest_time
0.02 * week_14_longest_time
0.07 * week_15_longest_time
0.06 * week_1_average_duration
0.0 * week_2_average_duration
0.01 * week_3_average_duration
0.0 * week_4_average_duration
0.05 * week_5_average_duration
0.03 * week_8_average_duration
0.01 * week_9_average_duration
0.0 * week_11_average_duration
0.03 * week_13_average_duration
0.14 * week_14_average_duration
0.01 * week_15_average_duration
0.1 * week_16_average_duration
-0.02 * week_1_fastest_pace
-0.07 * week_2_fastest_pace
-0.03 * week_3_fastest_pace
-0.05 * week_4_fastest_pace
-0.08 * week_5_fastest_pace
-0.03 * week_6_fastest_pace
-0.06 * week_8_fastest_pace
-0.01 * week_10_fastest_pace
-0.02 * week_11_fastest_pace
-0.01 * week_12_fastest_pace
-0.04 * week_13_fastest_pace
-0.08 * week_14_fastest_pace
-0.14 * week_15_fastest_pace
-0.2 * week_16_fastest_pace
```

The Lasso model removed additional features and identified the features that result in an increased duration and those that result in a decreased duration. Negative coefficients lead to a decrease in finish time while positive values lead to an increase in the finish time. Total Distance, Average Distance, and Fastest Pace are the key features for running a faster marathon. Longest Time and Average Duration predict a longer finish time. This would imply that increasing weekly distance and running some of the weekly distance at a pace faster than the marathon finish time will improve the marathon finish time. Running slow and long is less likely to lead to a faster marathon finish time.

The dataset was also evaluated on a Ridge Regression, Random Forest, and AdaBoost Regression model to obtain the following metrics.

Table 5: Evaluation of models

Model	**R^2**	**Mean Squared Error**	**Root Mean Squared Error**	**Mean Absolute Error**	**Mean Absolute Percentage Error**
Lasso Regression	0.55	923.94	30.4	22.76	10%
Ridge Regression	0.71	576.46	24.01	16.85	7%
Random Forest	0.95	551.1	23.48	17.32	8%
AdaBoost Regression	0.6	1068.92	32.69	26.97	13%

On the overall dataset the Random Forest Regressor and Ridge Regression performed the best but the accuracy of each model is not very accurate compared to conventional prediction calculators and tests. Since there are many types of runners with a wide range of fitness levels and goals it is not realistic for a one size fits all model. Using K-means clustering the dataset will be divided into clusters then each model will be fit and tested within the cluster.

The results for each model in each cluster are shown in Table 6.

Table 6: Cluster Models

Label	0				
Training Samples	876				
Testing Samples	220				
Slowest Finisher	471				
Fastest Finisher	137				
Std Dev	49.9				
Model	**R^2**	**MSE**	**RMSE**	**MAE**	**MAPE**
Lasso Regression	0.91	288.12	16.97	11.62	5.52%
Ridge Regression	0.91	288.28	17.0	11.56	5.51%
Random Forest	0.97	502.78	22.42	16.26	7.77%
AdaBoost	0.81	781.17	27.95	21.74	10.95%
Label	1				
Training Samples	199				
Testing Samples	50				

Slowest Finisher	459				
Fastest Finisher	160				
Std Dev	51.5				
Model	**R^2**	**MSE**	**RMSE**	**MAE**	**MAPE**
Lasso Regression	0.84	444.52	21.1	15.45	15.45%
Ridge Regression	0.86	558.94	23.64	17.76	7.15%
Random Forest	0.93	727.77	26.98	16.58	6.77%
AdaBoost	0.86	777.36	27.88	20.43	8.4%
Label	2				
Training Samples	1477				
Testing Samples	370				
Slowest Finisher	450				
Fastest Finisher	132				
Std Dev	38.22				
Model	**R^2**	**MSE**	**RMSE**	**MAE**	**MAPE**
Lasso Regression	0.86	232.08	15.23	10.28	4.57%
Ridge Regression	0.86	229.1	15.14	10.39	4.62%
Random Forest	0.92	589.57	24.28	18.16	8.09%
AdaBoost	0.54	794.49	28.19	22.76	10.51%
Label	3				
Training Samples	323				
Testing Samples	81				
Slowest Finisher	463				
Fastest Finisher	142				
Std Dev	49.79				
Model	**R^2**	**MSE**	**RMSE**	**MAE**	**MAPE**
Lasso Regression	0.83	1398.99	37.4	22.14	9.19%
Ridge Regression	0.84	1527.7	39.09	23.76	9.86%
Random Forest	0.94	1017.65	31.9	22.88	9.96%

AdaBoost	0.82	1085.49	32.95	24.09	10.81%
Label	4				
Training Samples	620				
Testing Samples	155				
Slowest Finisher	468				
Fastest Finisher	158				
Std Dev	60.93				
Model	**R^2**	**MSE**	**RMSE**	**MAE**	**MAPE**
Lasso Regression	0.82	850.85	29.17	20.18	7.48%
Ridge Regression	0.82	873.17	29.55	20.18	7.49%
Random Forest	0.95	1034.2	32.16	24.7	9.45%
AdaBoost	0.77	1219.1	34.92	27.53	10.84%
Label	5				
Training Samples	1769				
Testing Samples	443				
Slowest Finisher	471				
Fastest Finisher	138				
Std Dev	37.6				
Model	**R^2**	**MSE**	**RMSE**	**MAE**	**MAPE**
Lasso Regression	0.67	596.16	24.4	17.12	7.76%
Ridge Regression	0.81	314.65	17.74	11.82	5.4%
Random Forest	0.95	515.24	22.7	15.63	7.08%
AdaBoost	0.69	719.26	26.82	19.97	9.31%

There are some improvements by training and testing the models on the clustered data but the range of finishing times is still fairly large within each cluster.

Evaluation and Results

The clustering of the dataset based on marathon training data provided some insight into different training styles but overall it was not able to determine if a particular cluster was better than the others. Principal Component Analysis and Lasso Regression were able to identify some of the key features that defined the marathon training cycle but based on the results there are likely more features that could have been added to the dataset. This methodology should be performed on a dataset that is pulled from known participants that could provide additional features like heart rate, weight, training location, previous marathon finish times, and current marathon finish time goals. Since this dataset was accumulated from publicly provided data from a wide range of participants and it is not known their precise goals with the training cycle there is a lot of variability in the data.

While in some of the clusters the models perform better than on the overall dataset it is still far less accurate than many online race predictor calculators and running tests like Yasso 800s that can be used to predict marathon finish times. Again, with the dataset ranging from a little over a 2-hour finish time and close to around 8 hour finish time there is too much variability in the data to make precise predictions.

In conclusion, the evaluation provided some insights but not a strong enough case to confirm which training style is best or if it can predict a finish time from the training data.