otherRegressions

The nonsupervised cluster error means were widely dispered. We pursued to subset the data by agegroups and gender, rather than using the nonsupervised clustering mechanisms and compare the mean error for the subset groups obtained by using crossvalidation. Below is the output from R code for different combinations of subset groups. If we subset the data by agegroup only, kmeans clusters were better for modeling boston marathon predictive finish times.
If we subset the data by agegroup and gender, we were getting on average much less mean errors for females groups compared to kmeans clusters but not the same can be said for male groups.

```
## In age group i=[ 15 , 25 ] the mean error is  342.0702  and number of rows = 4330
## In age group i=[ 25 , 35 ] the mean error is  259.4245  and number of rows = 15465
## In age group i=[ 35 , 45 ] the mean error is  237.2754  and number of rows = 21037
## In age group i=[ 45 , 55 ] the mean error is  249.1888  and number of rows = 17247
## In age group i=[ 55 , 65 ] the mean error is  315.3395  and number of rows = 5386
## In age group i=[ 65 , 75 ] the mean error is  368.3841  and number of rows = 675
## In age group i=[ 75 , 85 ] the mean error is  469.0303  and number of rows = 27

## ####################################################################################################

## For females in age group i=[ 15 , 25 ] the mean error is  306.8224  and number of rows = 2484
## For females in age group i=[ 25 , 35 ] the mean error is  249.5365  and number of rows = 8043
## For females in age group i=[ 35 , 45 ] the mean error is  214.838  and number of rows = 9248
## For females in age group i=[ 45 , 55 ] the mean error is  253.942  and number of rows = 5676
## For females in age group i=[ 55 , 65 ] the mean error is  303.8146  and number of rows = 1032
## For females in age group i=[ 65 , 75 ] the mean error is  235.5316  and number of rows = 82
## For females in age group i=[ 75 , 85 ] the mean error is  69.97445  and number of rows = 3

## ####################################################################################################

## For males in age group i=[ 15 , 25 ] the mean error is  386.6893  and number of rows = 1846
## For males in age group i=[ 25 , 35 ] the mean error is  264.006  and number of rows = 7422
## For males in age group i=[ 35 , 45 ] the mean error is  250.3243  and number of rows = 11789
## For males in age group i=[ 45 , 55 ] the mean error is  244.4575  and number of rows = 11571
## For males in age group i=[ 55 , 65 ] the mean error is  319.0253  and number of rows = 4354
## For males in age group i=[ 65 , 75 ] the mean error is  382.6745  and number of rows = 593
## For males in age group i=[ 75 , 85 ] the mean error is  464.7551  and number of rows = 24
```

We proceeded next using the half marathon time as a predictor, rather than the first split time, to find a better predictor.

```
## Using half marathon time the mean error is  118.2262
```

Using the half marathon time, we have greatly reduced the mean error as, 118.23 using cross-validation. This was an expectedd result as closer we are to finish line, better our predictive analytics will be, hence half marathon time was a better predictor than first split time. It will be more interesting to find the breaking point in the split times that if the runner is lagging behind it then she or he will not be able to finish the boston marathon in time.