HW4-Q3

Andres Potapczynski (ap3635)

11/13/2018

Research questions

For our final project we are analyzing the 2018 NYC Marthon data set. This data consists of over 50 thousand participants from all over the world. Given the popularity of the event we are excited to answer some research questions that we have like:

- **What are best running strategies?** We have the time each person took between the marathon's 5 K splits, thus we can see if the professional athletes' pacing strategy differs from the novice ones. Maybe it is the case that novices start really fast (relative to their overall time) and burn out faster while professional athletes maintain a constant pace.
- **What are the clusters that we observe in the data?** The idea here would be to learn how different given features like: age, gender, country of origin but also made up features like (number of runners in the country of the athlete) separate the data. We would probably get clusters like middle-aged American runners (which could all be also members of a charity teams) or professional athletes. We could start with simple techniques like k-means but also try to find embedding via matrix factorization.
- **What's the country composition of the data?** This is an exploratory plot, but a priori we have no idea from what countries or cities in the world do people come to run the NYC marathon. Nor do we know if the running strategies differ from country or if it is more likely that for certain age groups some countries are more prevalent than other. The good thing about this analysis is that it allows us to show maps, which I think they make really interesting visualizations.
- **On average, what are the most relevant features that predict performance?** Very unlikely but we could predict a persons' performance based on their age, country of origin, sex. Yet, probably the pacing strategy could be a good indicator of performance. For example, if you ran to fast the first 20 K then there is a high change that you would be top 1,000 or something like that.

For this homework I will start the analysis for the first question. Before jumping into the analysis it is worth mentioning that the process of getting the data set has been arduous. We have scraped the data from the web and also generated some new features like the ones that allow me to do this analysis. We are also expecting to expand our set of analysis with data of previous NYC marathons.

Now I load the data

```
file = './DBs/marathon_2018_aug.csv'
marathon <- read_csv(file)
marathon$variable = factor(marathon$variable, levels = c('5k_1',
                                                         '10k_2',
                                                         '15k_3',
                                                         '20k_4',
                                                         '25k_5',
                                                         '30k_6',
                                                         '35k_7',
                                                         '40k_8'))
marathon$bib = factor(marathon$bib)
```

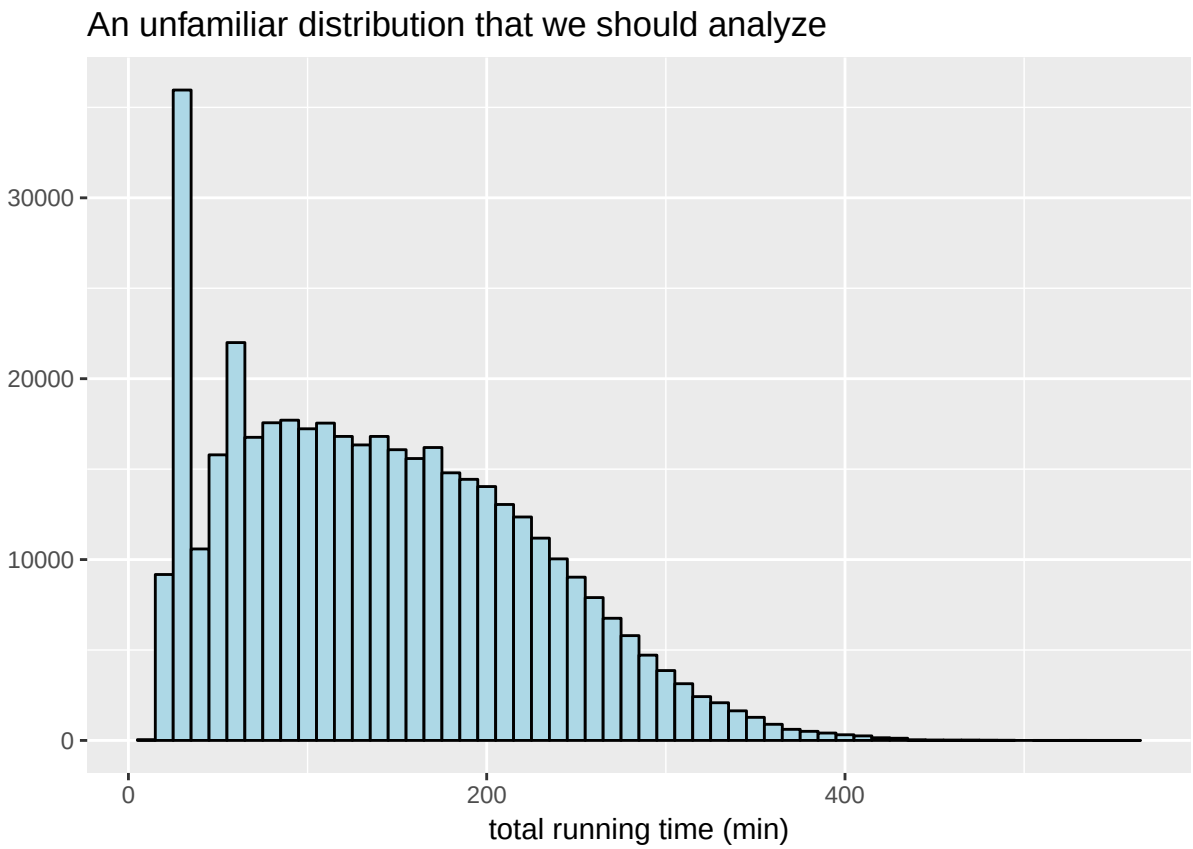
Some relevant columns are:

- `__gender_age__`: encodes the gender and the age of the runner
- `__team__`: the name of the sponsor or the charity team that the athlete belong to.

- __pace__: relative to its overall pace, how did the athlete change its speed during the race.
- __sec__: the number of seconds it takes to run each marathon split.
- __place__: encodes the city or country where the athlete comes from.
- __variable__: factor variable that decouples the data into 5 km marathon splits.

Visualization 1 - Running times distributions

```
ggplot(data = marathon, mapping = aes(x=sec / 60)) +
  geom_histogram(binwidth = 10,
                 fill='lightblue',
                 color='black') +
  ylab('') +
  xlab('total running time (min)') +
  ggtitle('An unfamiliar distribution that we should analyze')
```



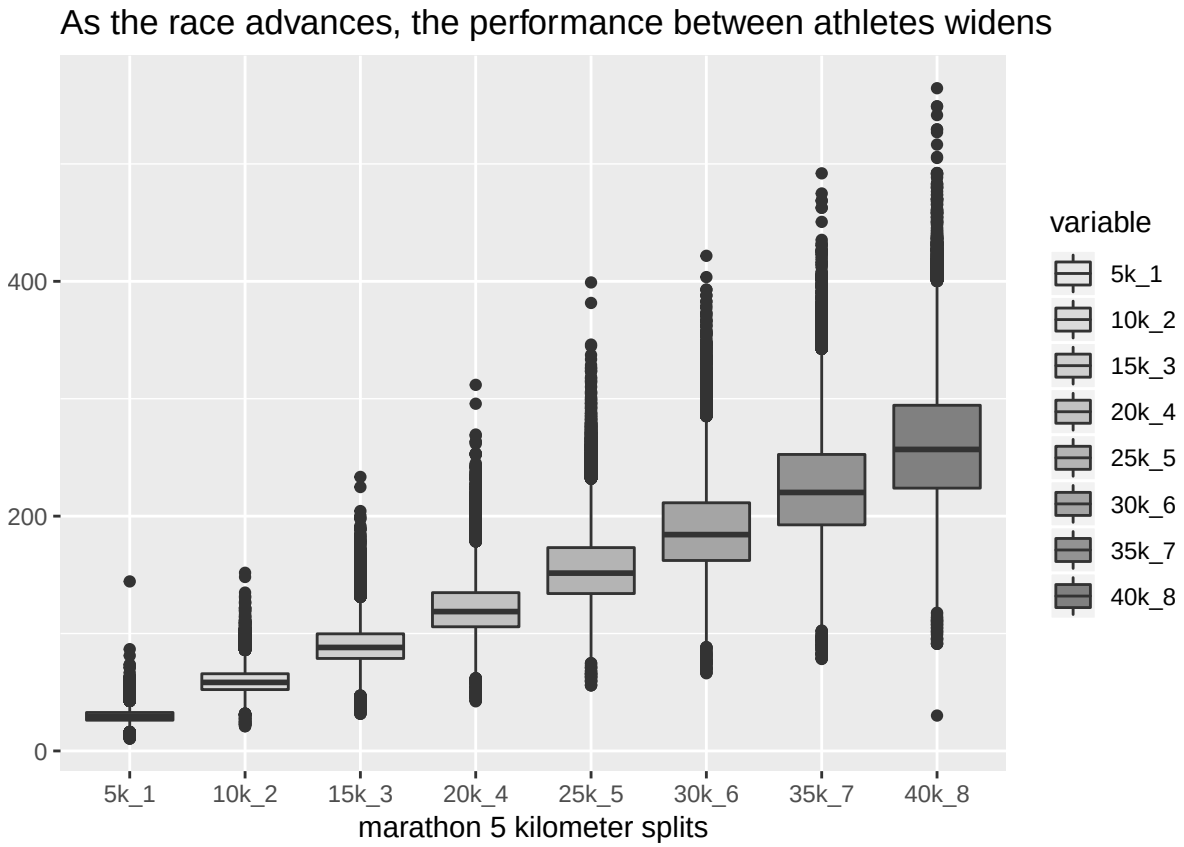
Recalling the statement of the Central Limit Theorem I was expecting a more Normal-shaped distribution. What we observe is completely different. There is a high concentration at the beginning of the distribution probably due to the professional athletes. In contrast, there is a heavy tail which, comes from the rest of the athletes.

Visualization 2 - Wider variances

```
ggplot(data = marathon) +
  geom_boxplot(aes(variable, sec / 60, fill=variable)) +
```

```
  scale_fill_grey(start=0.9, end=0.5) +
  ylab('') +
  xlab('marathon 5 kilometer splits') +
  ggtitle('As the race advances, the performance between athletes widens')
```

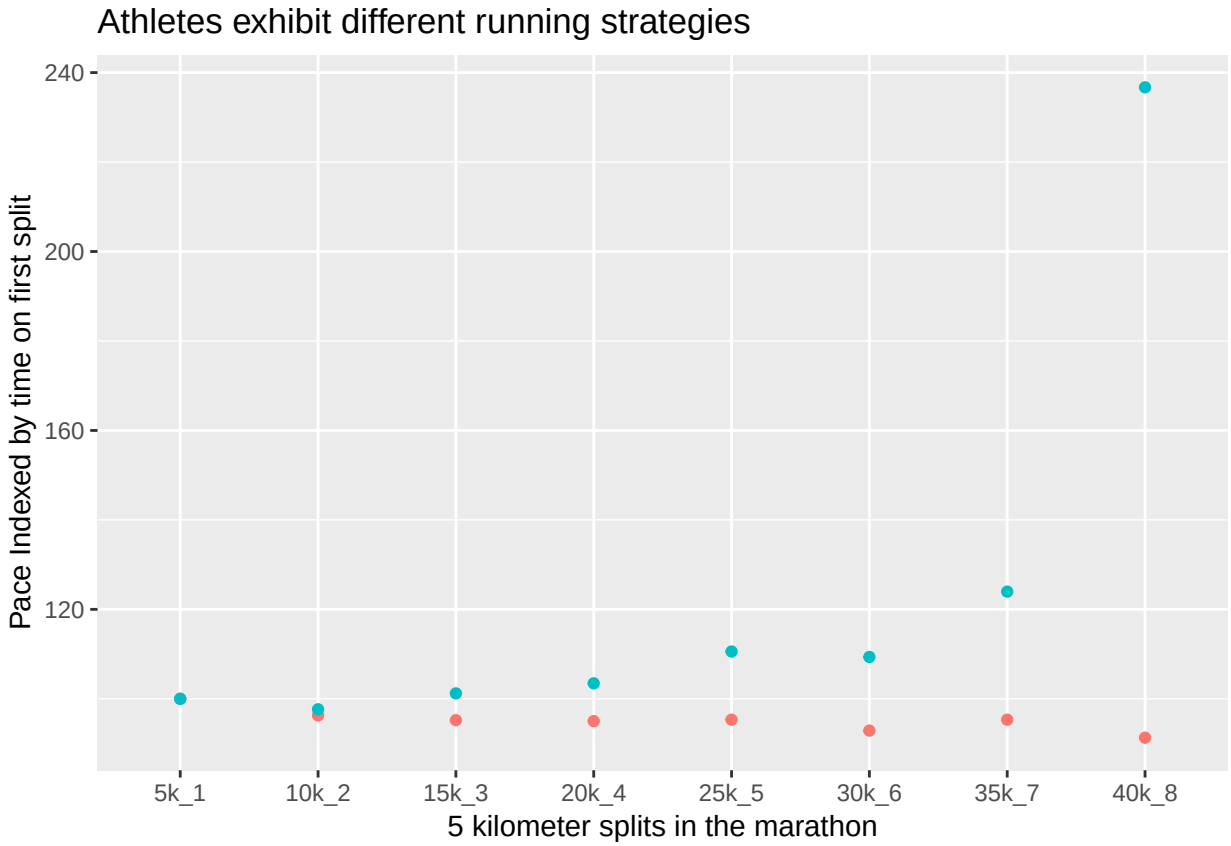
```
## Warning: Removed 1330 rows containing non-finite values (stat_boxplot).
```



as expected, the more we advance in the race the discrepancies in performance become more evident.

Visualization 3 - Runner's stategies

```
data_filtered <- marathon %>% filter(bib == 1 | bib == 600)
ggplot(data = data_filtered, mapping = aes(x = variable, y = pace_index, color=bib)) +
  geom_point() +
  theme(legend.position="none") +
  ylab('Pace Indexed by time on first split') +
  xlab('5 kilometer splits in the marathon') +
  ggtitle('Athletes exhibit different running strategies')
```



Can you guess which is the runner that performed best? The two strategies that we observe above are:

- **Increase the pace between splits and then give it all out**. For one athlete we obser that the first 20k are around the same pase as its first 5 k. Then it starts increasing between 25k to 35k to finally increase its speed 240 % to close the race.
- **Maintain a constant speed throughout**. All the splits appear to be random noise centred around a little less than what the athelte started.

Stay tuned to know the who's who in the above graph and the rest of the upcoming results from the analysis!