Question3

Arthur Herbout

November 14, 2018

First encounter with the Marathon dataset

We have decided for the final project to analyse all the data available for the 2018 New York City Marathon.

In this notebook, I will first show the structure of the race: 4 waves are sequentially made. Since there is a lot of participants, they will not start at the same time. It will be done by groups, I call them waves.

In order to detect those waves, I will use the difference between the official time and the gun time. The official time records the exact time between the runner crosses the start line and the end line. For elite runners, since they start with the gun, there will be no difference. Then we should see some increasing difference between those two values.

I am first starting to separate the gender from the age: they appear in the same variable in the original scrapped dataset.

```
gender <- c()
age <- c()
for (i in 1:nrow(data)){
  gender_ind <-(substr(as.character(data[i,"gender_age"]), 1, 1))
  age_ind <- (substr(as.character(data[i,"gender_age"]), 2,3))
  gender[i] = gender_ind
  age[i] = age_ind
}
data$gender <- gender
data$age <- as.numeric(age)
```

For my analysis to be correct, I have to compute all the differences between the official time and the gun time.

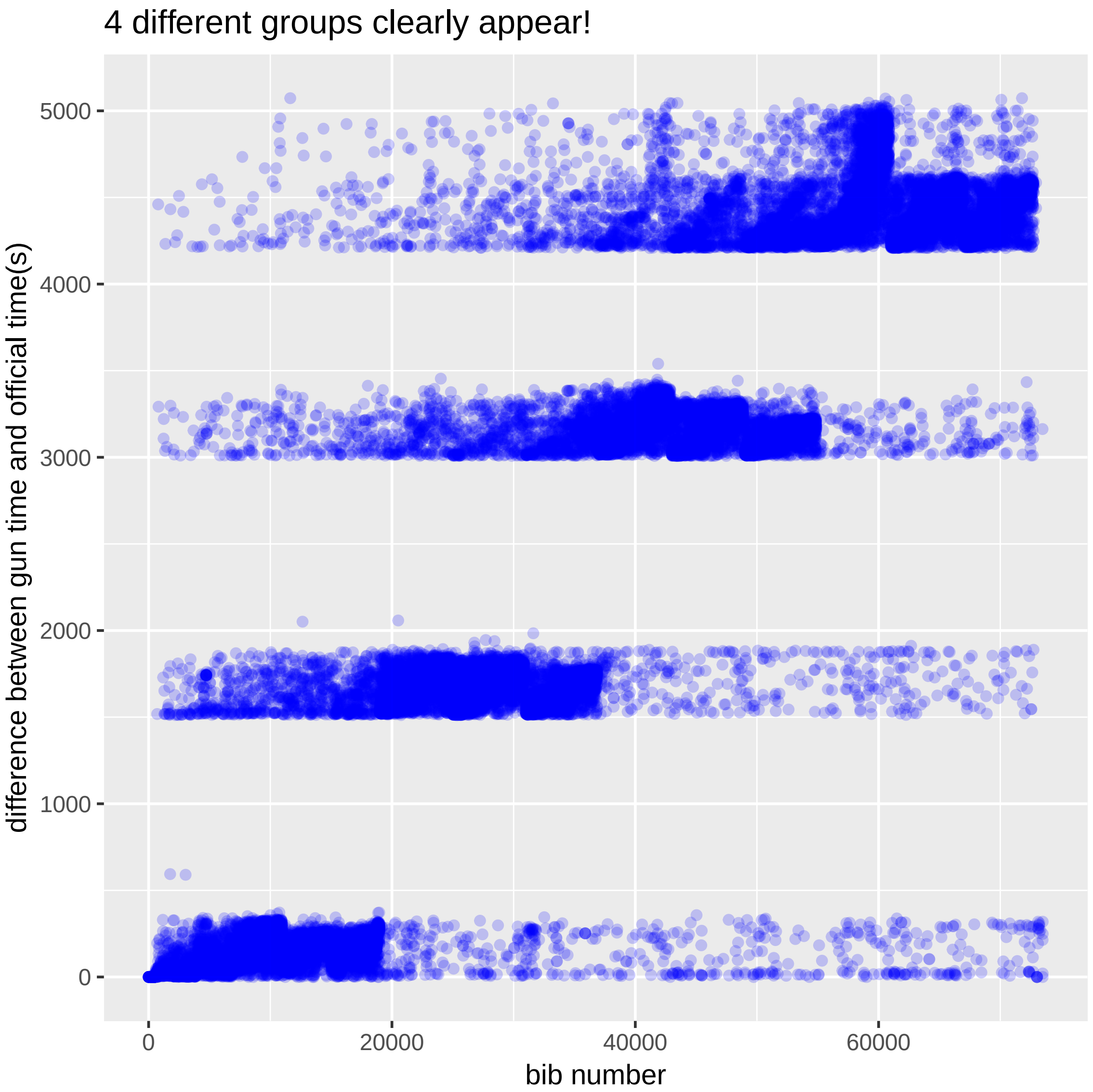
```
diff <- c()
for (i in 1:nrow(data)){
  value <- data[i, "gun_time"] - data[i, "official_time"]
  diff[i] = value
}
data$diff <- as.numeric(diff)
```

Before going for plots, let us first compute some statistics on this new variable:

```
summary(data$diff)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     305    3011    2368    4236    5073
```

```
ggplot(data, mapping = aes(x = bib, y = diff)) +
  geom_point(alpha=1/5, col= "blue") +
  xlab("bib number") +
  ylab("difference between gun time and official time(s)") +
  ggtitle("4 different groups clearly appear!")
```



This first very simple plot gives us some information about the organization. First, there are clearly 4 different starts: - as expected, the first one starts at 0. It is the one that coincid with the gun start, - A second start occurs 25 minutes after the gun start, - A third start occurs 50 minutes after the gun start, - The fourt start occurs 1 hour 10 minutes after the gun start

Alpha blending also make us realize that each group is fairly homogenous: the higher your bid is, the later you might start your race.

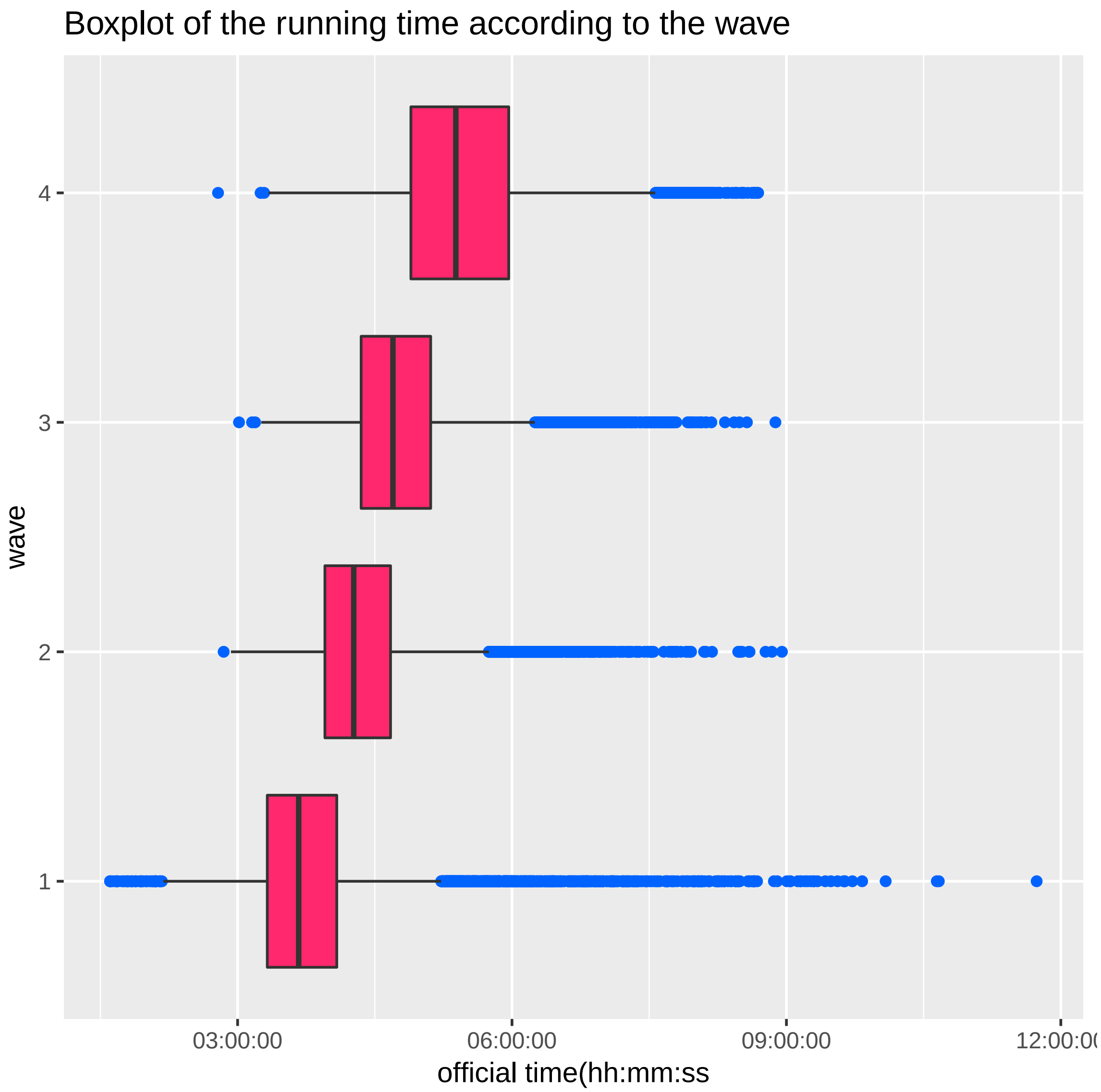
Now I will try to see to what extend the waves are homogenous: is wave 1 significantly better than wave 2? That is the kind of question I want to answer now.

```
wave <- c()
for (i in 1:nrow(data)){
  value <- data[i, "diff"]
  if (value < 1500){
    wave[i] = 1
  } else if (value < 3000){
    wave[i] = 2
  }else if (value < 4000){
    wave[i] = 3
  }else {
    wave[i] = 4
  }
}
data$wave <- as.numeric(wave)
```

```
summary(data$wave)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   3.000   2.524   4.000   4.000
```

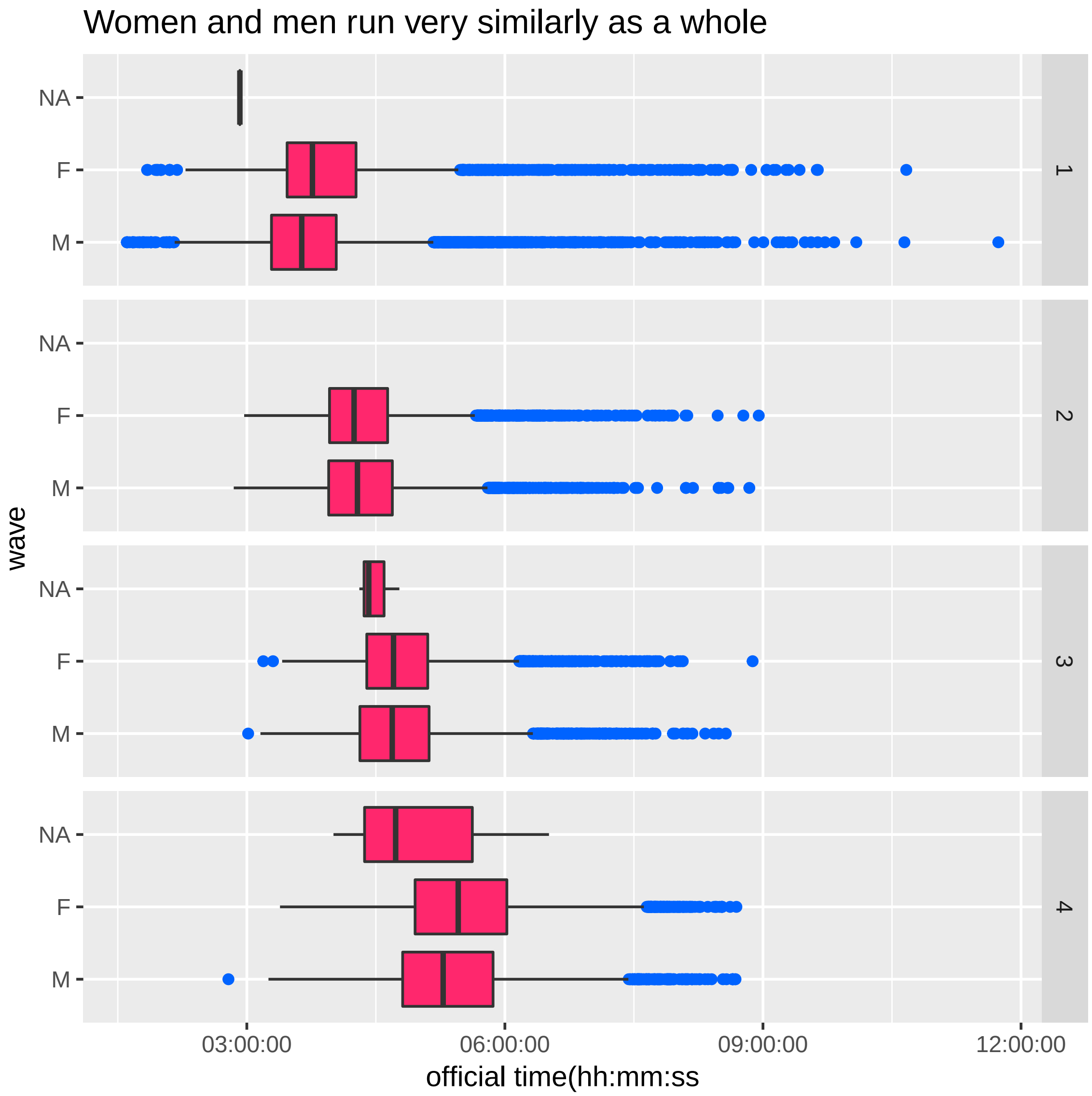
```
ggplot(data, aes(x = factor(wave, levels = c('1','2','3','4')) , y = official_time)) +
  geom_boxplot(fill = "#FF276D", outlier.colour = "#0063FF") +
  coord_flip() +
  ylab("official time(hh:mm:ss") +
  xlab("wave") +
  ggtitle("Boxplot of the running time according to the wave")
```



What we didn't see previously was the actual official time given the wave. Here we see that the groups are well contructed: they clearly show different patterns. We should also mention here that the outliers for the fastest group are in fact the wheelchair participants.

Is there any particular gender repartition among the waves?

```r
ggplot(data, aes(x = factor(gender, levels = unique(gender)) , y = official_time)) +
  geom_boxplot(fill = "#FF276D", outlier.colour = "#0063FF") +
  facet_grid(wave~.) +
  coord_flip() +
  ylab("official time(hh:mm:ss") +
  xlab("wave") +
  ggtitle("Women and men run very similarly as a whole")
```



Some interesting comments can be made at this point: - The distribution for the second wave is almost similar between men and women. This group is composed, I guess, by people who run regularly but are not athletes. They do not push, by training, their capabilities as far as their body can. Therefore, the physical differences are not visible for this group. - For the first wave, the elite runners, the gender distinction is clearly visible. The plot shows that the physical differences show up when the training is close to perfect: the innate differences appears.

Now that I have the different new features, I can save this cleaner dataset.

```
write_csv(data, './clean_marathon.csv')
```

Many other questions now come to mind: - what happen during the race? Do men and women behave the same? Are their splits different? - what is the best running tactics? - what is the age factor?