HW4_Q3

Antonia Lovjer

11/13/2018

The data set that we are using was taken from the NYRR Marathon website, and includes data on each individual finisher of the NYRR 2018 NY Marathon. The entire data set contains 52670 observations and 33 original variables.

```r
data <- read.csv("/Users/antonialovjer/Documents/Columbia/EDAV/Final Project/marathon_2018.csv")
head(data)
```

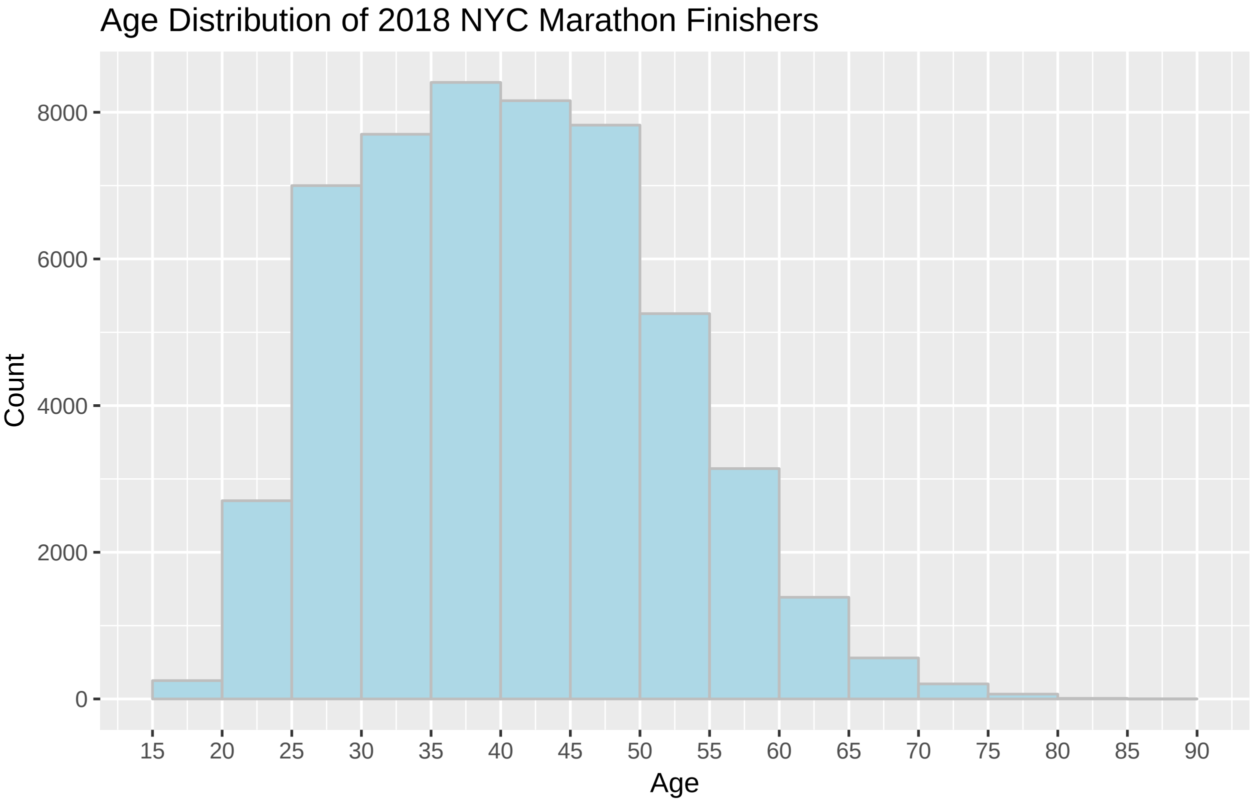
```
##   bib gender_age gun_place gun_time              name official_time
## 1   1        M25         3  2:06:26 Geoffrey Kamworor       2:06:26
## 2   2        M22         2  2:06:01      Shura Kitata       2:06:01
## 3   3        M28         1  2:05:59     Lelisa Desisa       2:05:59
## 4   4        M27         4  2:08:30      Tamirat Tola       2:08:30
## 5   5        M26         5  2:10:21    Daniel Wanjiru       2:10:21
## 6   6        M24         8  2:12:40      Festus Talam       2:12:40
##   pace_per_mile percentile_age.graded              place place_age.graded
## 1         04:50                97.25% Kapchorwa District                3
## 2         04:49                97.57%        Addis Ababa                2
## 3         04:49                 97.6%        Addis Ababa                1
## 4         04:55                95.69%        Addis Ababa                4
## 5         04:59                94.33%               Embu                5
## 6         05:04                92.69%               Iten                9
##   place_age.graded_of place_age.group place_age.group_of place_gender
## 1              30,581               2              2,876            3
## 2              30,581               1                773            2
## 3              30,581               1              2,876            1
## 4              30,581               3              2,876            4
## 5              30,581               4              2,876            5
## 6              30,581               2                773            8
##   place_gender_of place_overall place_overall_of splint_10k splint_15k
## 1          30,581             3           52,697    0:30:51    0:45:49
## 2          30,581             2           52,697    0:30:48    0:45:47
## 3          30,581             1           52,697    0:30:51    0:45:48
## 4          30,581             4           52,697    0:30:48    0:45:48
## 5          30,581             5           52,697    0:30:51    0:45:49
## 6          30,581             8           52,697    0:30:50    0:45:47
##   splint_20k splint_25k splint_30k splint_35k splint_40k splint_5k
## 1    1:00:45    1:15:44    1:30:20    1:45:19    1:59:40   0:15:43
## 2    1:00:39    1:15:45    1:30:20    1:45:19    1:59:50   0:15:43
## 3    1:00:44    1:15:45    1:30:20    1:45:19    1:59:40   0:15:43
## 4    1:00:45    1:15:44    1:30:21    1:45:19    2:01:21   0:15:46
## 5    1:00:49    1:16:18    1:31:32    1:47:16    2:03:22   0:15:46
## 6    1:00:46    1:15:46    1:30:20    1:46:11    2:04:07   0:15:45
##   splint_half   team time_age.graded
## 1     1:03:59   NIKE         2:06:26
## 2     1:03:55   NIKE         2:06:01
## 3     1:03:57   NIKE         2:05:59
## 4     1:03:58 adidas         2:08:30
## 5     1:04:10 adidas         2:10:21
## 6     1:03:59 adidas         2:12:40
```

Histogram of Age

The variable that I will be examining here is **age** of the marathon finishers. First we will look at the distribution of the ages in a histogram.

The distribution of the data appears to be approximately normal centered around 40. Overall, there are more younger people finishing the race in comparison to older people. From this plot, my next questions would be how official finish times differ by age, and the relationship between age and split times for the runners.

```
plot1 <- ggplot(data, aes(x=age)) +
  geom_histogram(color='grey', fill='lightblue', binwidth=5, center=62.5) +
  ggtitle("Age Distribution of 2018 NYC Marathon Finishers") +
  scale_x_continuous(breaks=seq(15,90, by=5)) +
  ylab("Count") +
  xlab("Age")
plot1
```

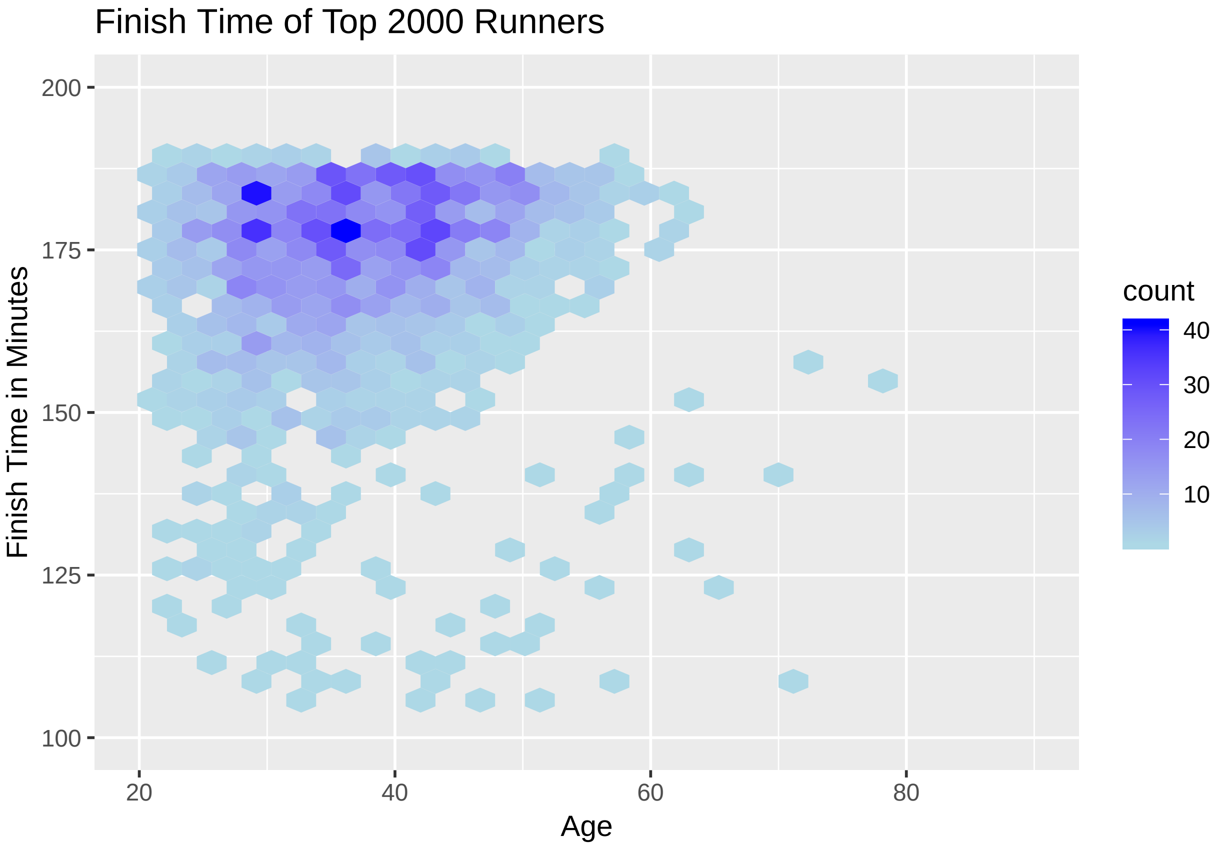


Sample of Finish Times by Age

Next we look at a hexagonal heat map of the age and finish times for the first 2000 marathon finishers.

From this plot we see that overall the top finishers tend to be younger (< 60), with the fastest runner finishing the race in 96 minutes. The majority of the data is centered around 177 minutes, and in the 30-45yr range of the scale. The age observation is consistent with the conlcusion from the histogram above showing that the ages of the runners are centered around 40, and that there are fewer older runners compared to younger runners. Another interesting observation is that there are no old runners with long finish times present in the top 2000. The fact that the finish times are so concentrated around 177 begs the question of what is the distribution of the finish times for the entire data set. It is most likely the case that the majoirty of the runners place in an average marathon range, and that this starts with many of the runners who are also in the top 2000.

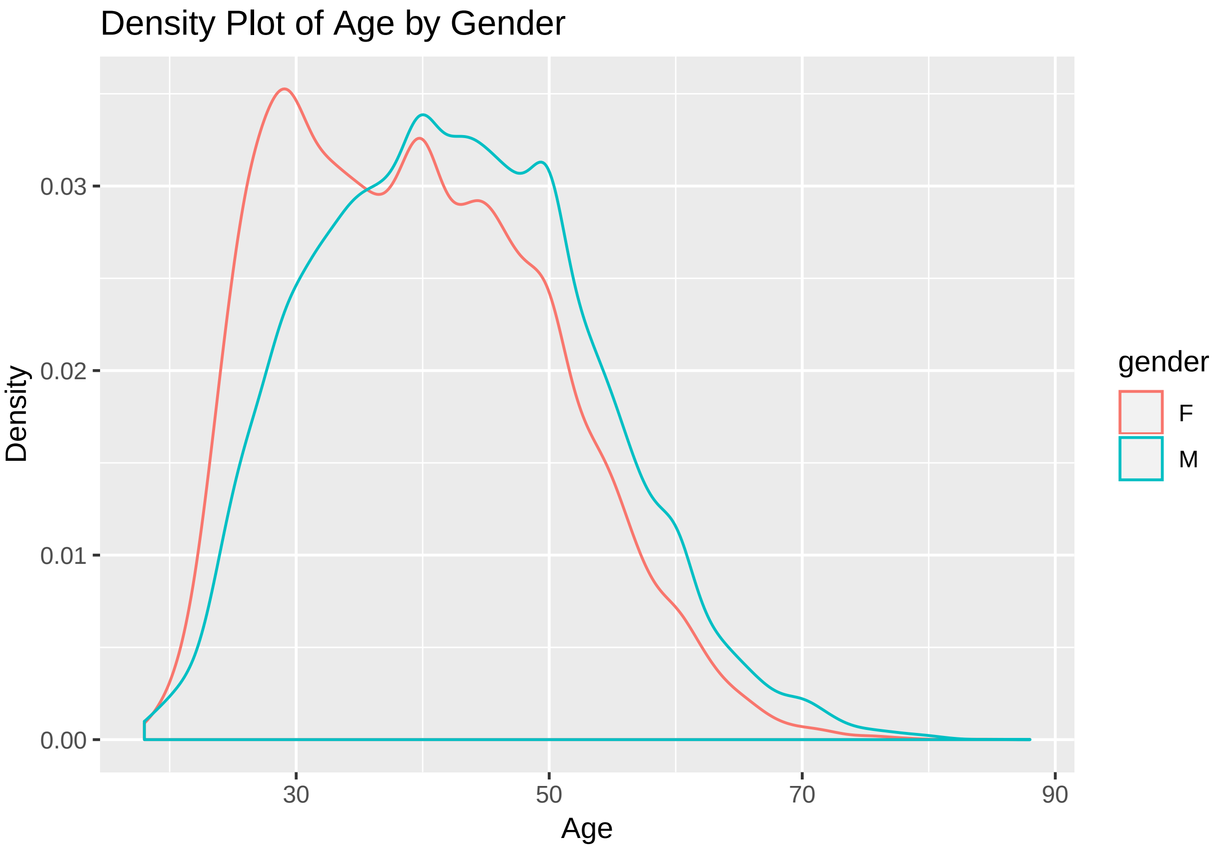
```
data_sample <- data[1:2000, ]

ggplot(data_sample, aes(x=age, y=finish_min)) +
  scale_fill_gradient(low = "lightblue", high = "blue") +
  geom_hex() +
  xlim(20,90) +
  ylim(100,200) +
  ylab("Finish Time in Minutes") +
  xlab("Age") +
  ggtitle("Finish Time of Top 2000 Runners")
```



Density Plot of Age by Gender

Finally we plot density curves of the age distribution seperated by gender. We see that the females tend to be younger in comparison to men in the data set. Further questions to be examined are how the finish times differ by gender, and if gender is connected with any of the other variables such as split times and location of origin.

```
ggplot(data, aes(x=age, color=gender)) +
  geom_density() +
  ggtitle("Density Plot of Age by Gender") +
  ylab("Density") +
  xlab("Age")
```


Density Plot of Age by Gender
Density
0.03
0.02
0.01
0.00
Age
30
50
70
90
gender
F
M