Math 130 Final Project

Bill Parnell

2/20/2022

Boston Marathon Results 2017

I am looking at results from the 2017 Boston Marathon and summarizing age and finish time.

```
marathon <- read.csv("/Users/billparnell/Documents/chico-classes/math130/data/marathon_results_2017.csv

dim(marathon)
```

```
## [1] 26410    25
```

Univariate Analysis: Age

```
summary(marathon$age)
```

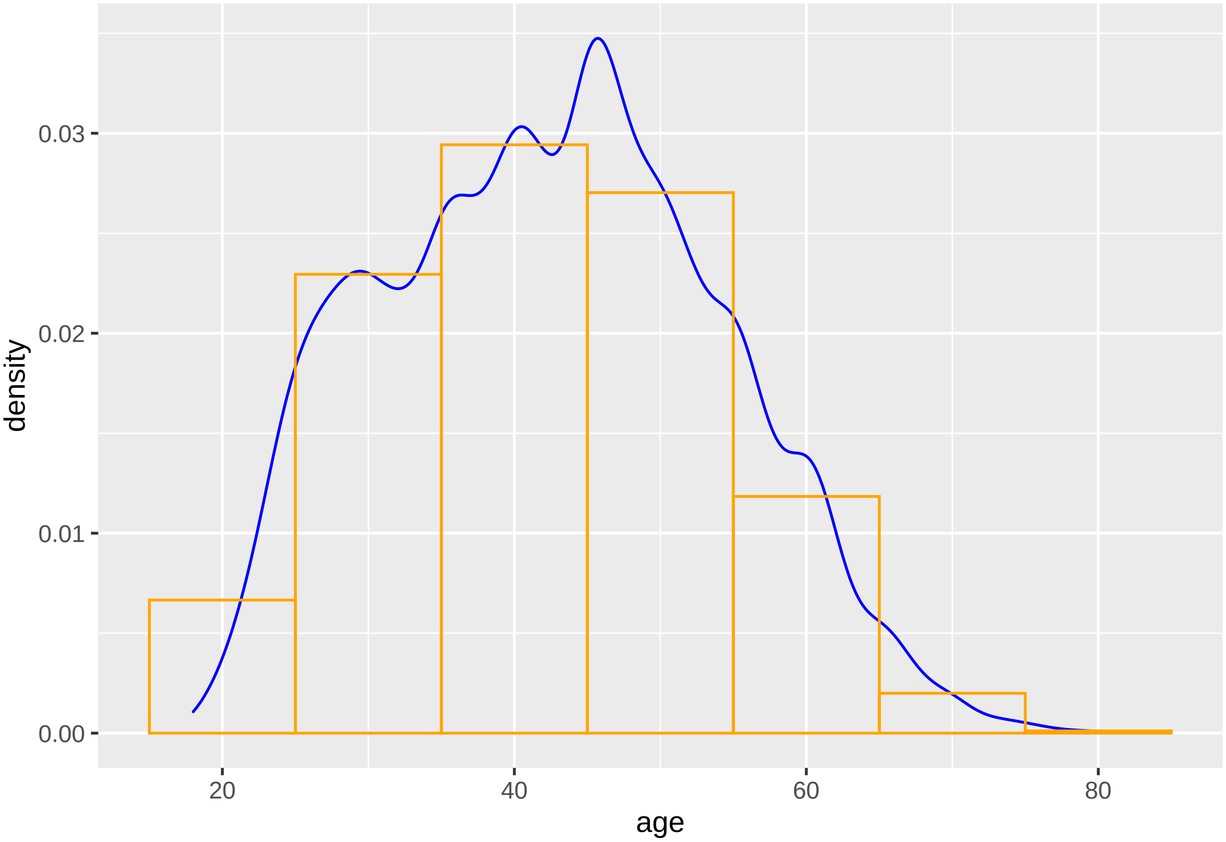
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   34.00   43.00   42.59   51.00   84.00
```

```
sd(marathon$age)
```

```
## [1] 11.41947
```

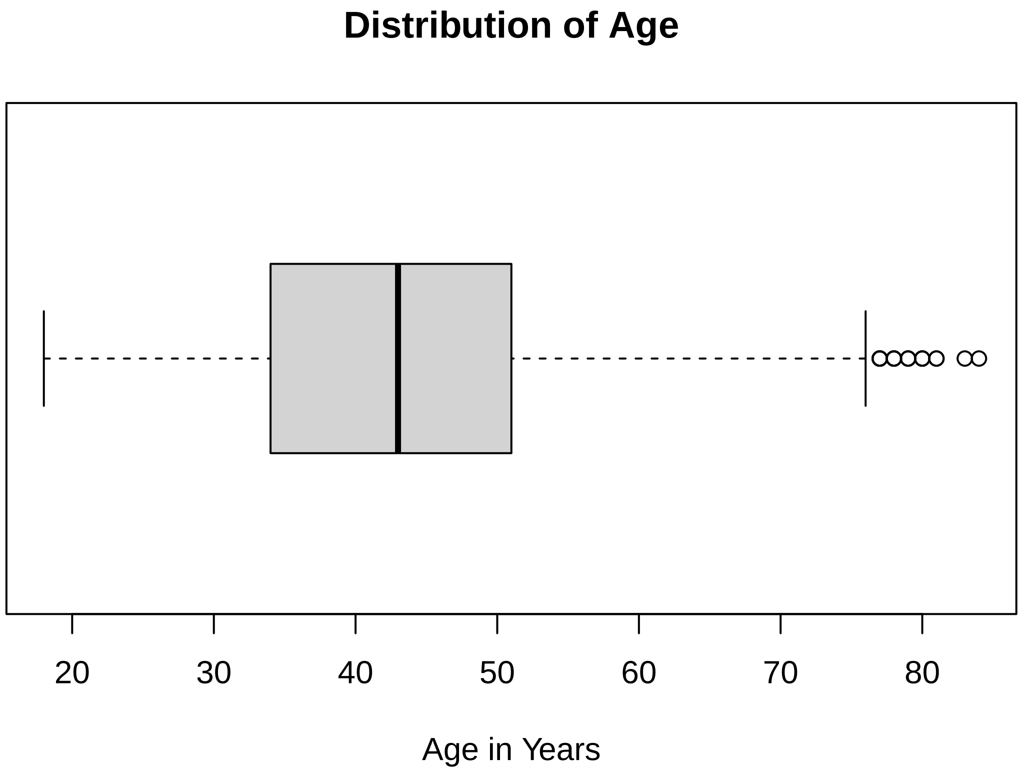
The mean age for runners in the population was 42.59 year with a standard deviation of 11.42 years.

```
ggplot(marathon, aes(x=age)) + geom_density(col="blue") + geom_histogram(aes(y=..density..), color="oran
```



The largest bin of ages was the 10 year bin centered on 40 years old. The curve is skewed a bit, likely due to the lack of runners under a certain age.

```
boxplot(marathon$age, horizontal = TRUE, main = "Distribution of Age", xlab="Age in Years")
```



The main cluster of ages is between 30 and 50, which can be seen clearly here.

Univariate Analysis: Finish Time

```
summary(marathon$finish_time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.160   3.472   3.861   3.968   4.363   7.971
```

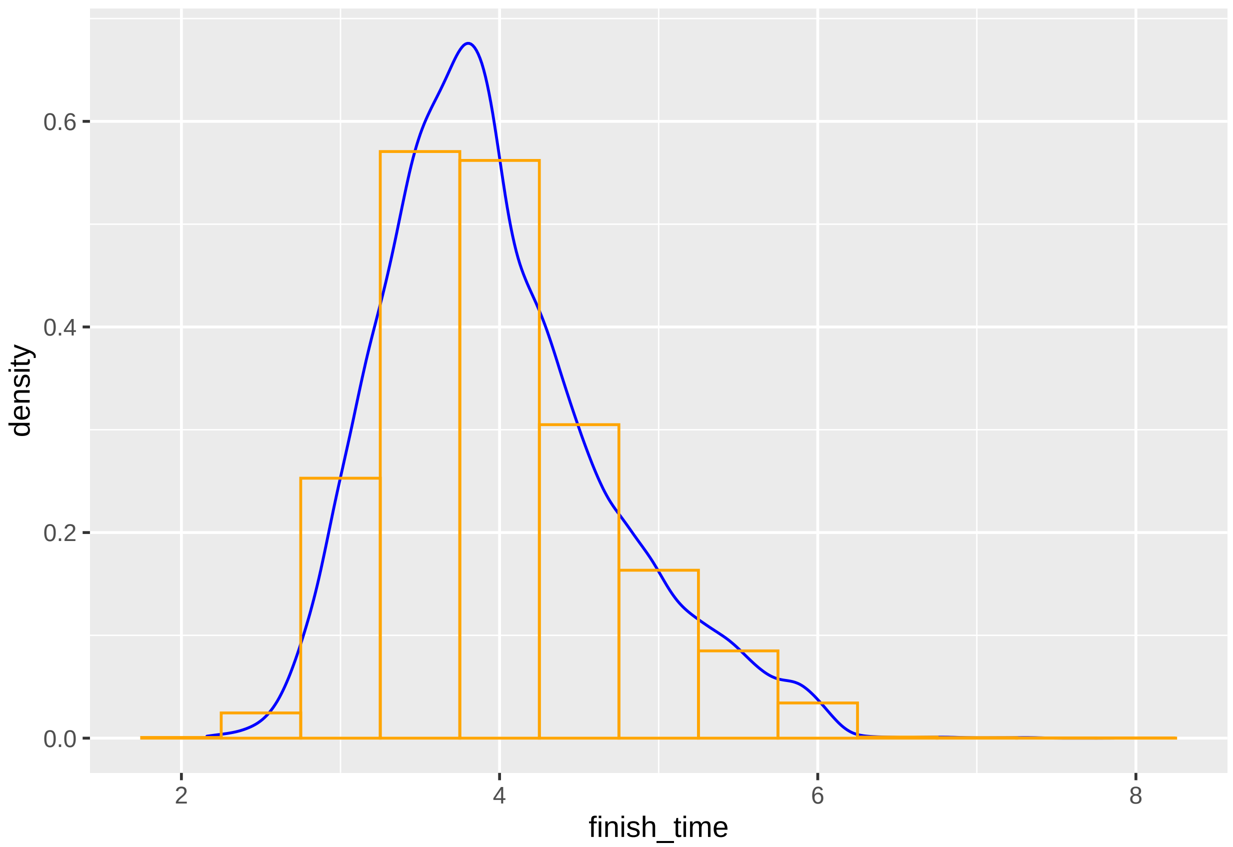
We can see hear a nice summary of min/max, quartiles, and mean. The mean finish time was 3.968 hours (about 3 hours and 58 minutes).

```
sd(marathon$finish_time)
```

```
## [1] 0.7024676
```

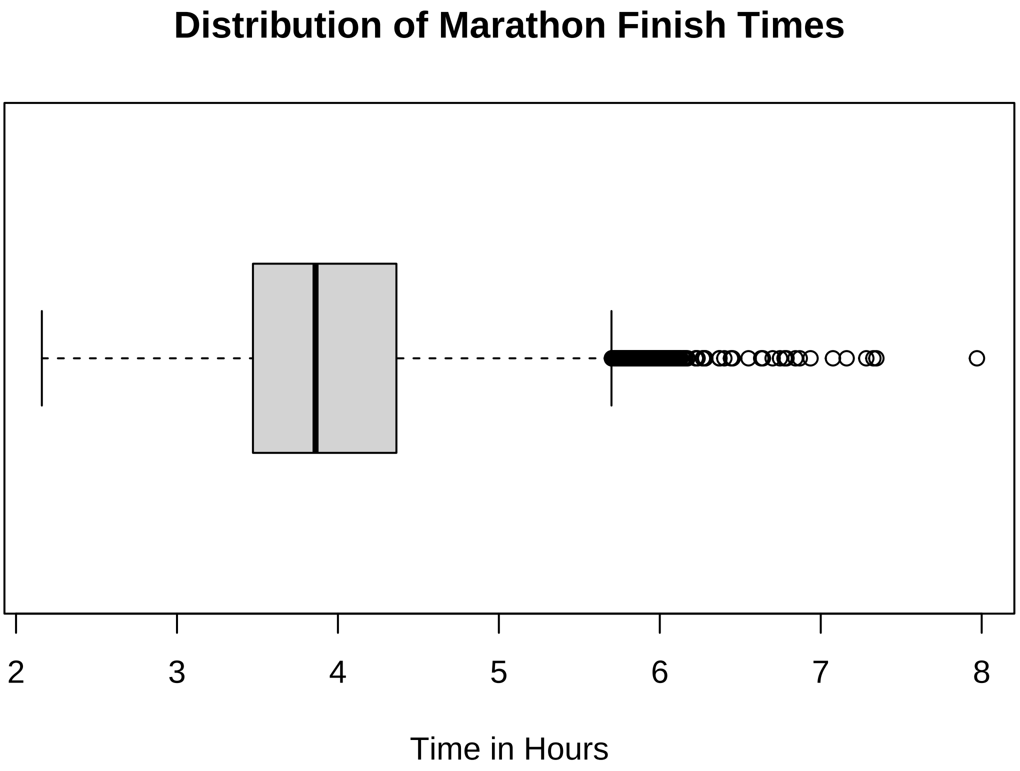
The standard deviation was 0.702 hours (about 42 minutes).

```
ggplot(marathon, aes(x=finish_time)) + geom_density(col="blue") +
geom_histogram(aes(y=..density..), color="orange", fill=NA, binwidth = .5)
```



Here we can see a nice plot of finish time. The curve is skewed a bit to the left, likely due to the lack of runners below a certain age.

```
boxplot(marathon$finish_time, horizontal = TRUE, main="Distribution of Marathon Finish Times",
xlab="Time in Hours")
```



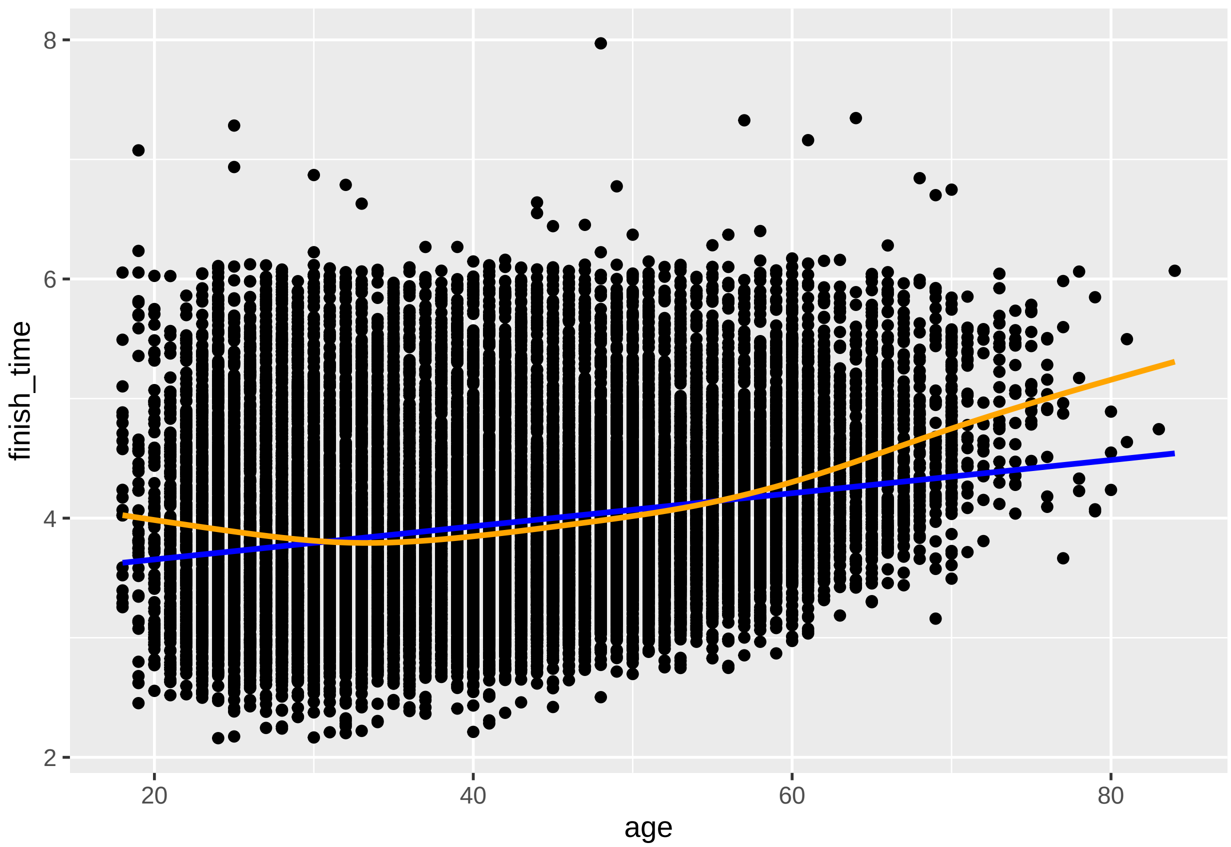
Here we can see the majority of finish times are between 3.5 and 4.5 hours.

Bivarate Description

```
ggplot(marathon, aes(x=age, y=finish_time)) + geom_point() +
geom_smooth(se=FALSE, method = "lm", color="blue") +
geom_smooth(se=FALSE, color="orange")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

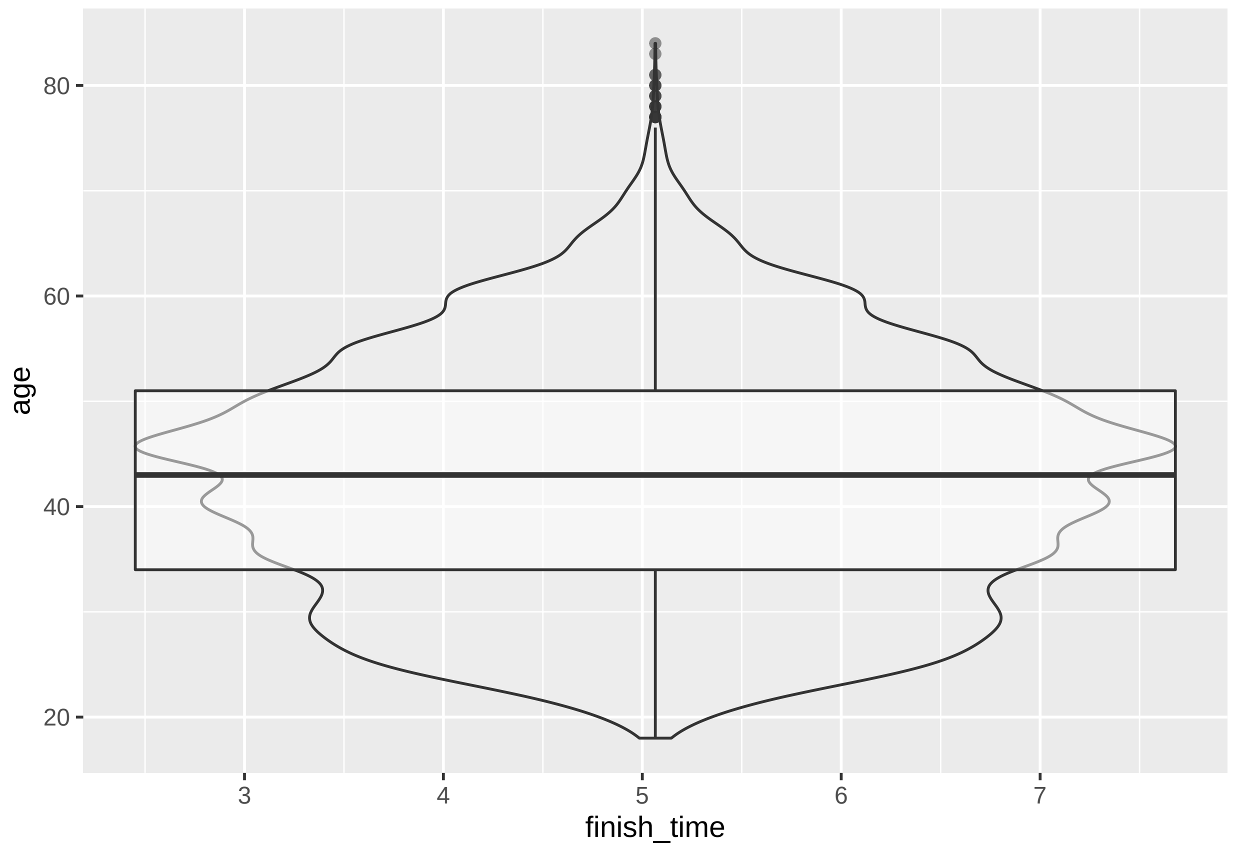
```
## ‘geom_smooth()‘ using method = ’gam’ and formula ’y ~ s(x, bs = "cs")’
```



Here we can see that the best fit line suggests that finish time increases with age. The lowess line shows that there is a subtle effect that with an increase in age from 20's to 30's there is a decrease in finish time, but finish time increases with age thereafter.

```
ggplot(marathon, aes(x=finish_time, y=age, fill=finish_time)) + geom_violin(alpha=.1) +
geom_boxplot(alpha=.5, width=.2)
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```


80
60
40
20
age
3
4
5
6
7
finish_time