



EXPLORING PITCH DATA IN R

**Did Zack Greinke
pitch differently in July?**

Chapter outline

- Data: Greinke's every pitch from multiple months
- Explore pitch velocity
- July vs. other months
- Graphical skills to compare distributions
- Impact of fastball velocity on hitting outcomes

Data description

```
> names(greinke)
[1] "p_name"           "pitcher_id"           "batter_stand"
[4] "pitch_type"       "pitch_result"         "atbat_result"
[7] "start_speed"      "z0"                   "x0"
[10] "pfx_x"            "pfx_z"                "px"
[13] "pz"               "break_angle"          "break_length"
[16] "spin_rate"        "spin_dir"             "balls"
[19] "strikes"          "outs"                 "game_date"
[22] "inning"           "inning_topbot"        "batted_ball_type"
[25] "batted_ball_velocity" "hc_x"                 "hc_y"
[28] "pitch_id"         "distance_feet"
```



```
> head(greinke[ , 5:6])
      pitch_result atbat_result
1           Ball           Walk
2 Swinging Strike           Single
3   Called Strike       Home Run
...
```

Examining dates: game_date

```
> head(greinke$game_date)
[1] "10/3/2015" "10/3/2015" "10/3/2015" "10/3/2015" "10/3/2015"
[6] "10/3/2015"

> class(greinke$game_date)
[1] "character"

> greinke$game_date <- as.Date(greinke$game_date, "%m/%d/%Y")

> head(greinke$game_date)
[1] "2015-10-03" "2015-10-03" "2015-10-03" "2015-10-03" "2015-10-03"
[6] "2015-10-03"

> class(greinke$game_date)
[1] "Date"
```

Separating dates

```
> library(dplyr)
> library(tidyr)

> greinke <- separate(data = greinke, col = game_date,
                      into = c("year", "month", "day"),
                      sep = "-", remove = FALSE)

> head(greinke[, 21:24])
  game_date year month day
1 2015-10-03 2015    10   03
2 2015-10-03 2015    10   03
3 2015-10-03 2015    10   03
4 2015-10-03 2015    10   03
5 2015-10-03 2015    10   03
6 2015-10-03 2015    10   03
```



EXPLORING PITCH DATA IN R

Let's practice!



EXPLORING PITCH DATA IN R

Subsets and histograms

Pitch velocity: `start_speed`

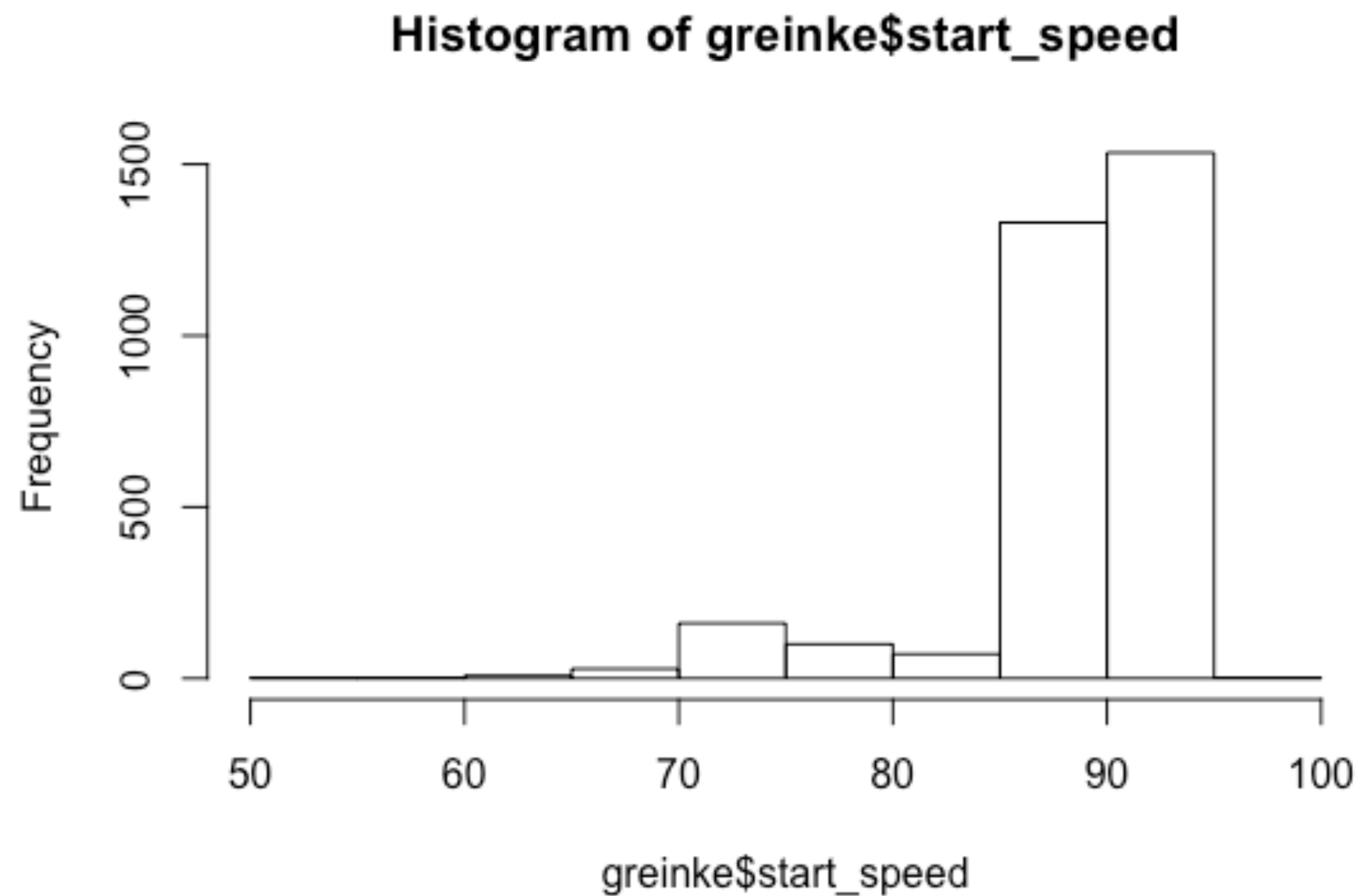
```
> head(greinke$start_speed)
[1] 94.2 92.4 92.7 86.9 92.8 87.8

> class(greinke$start_speed)
[1] "numeric"

> summary(greinke$start_speed)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
 52.20   87.30   89.80   88.44   91.80   95.40     3
```

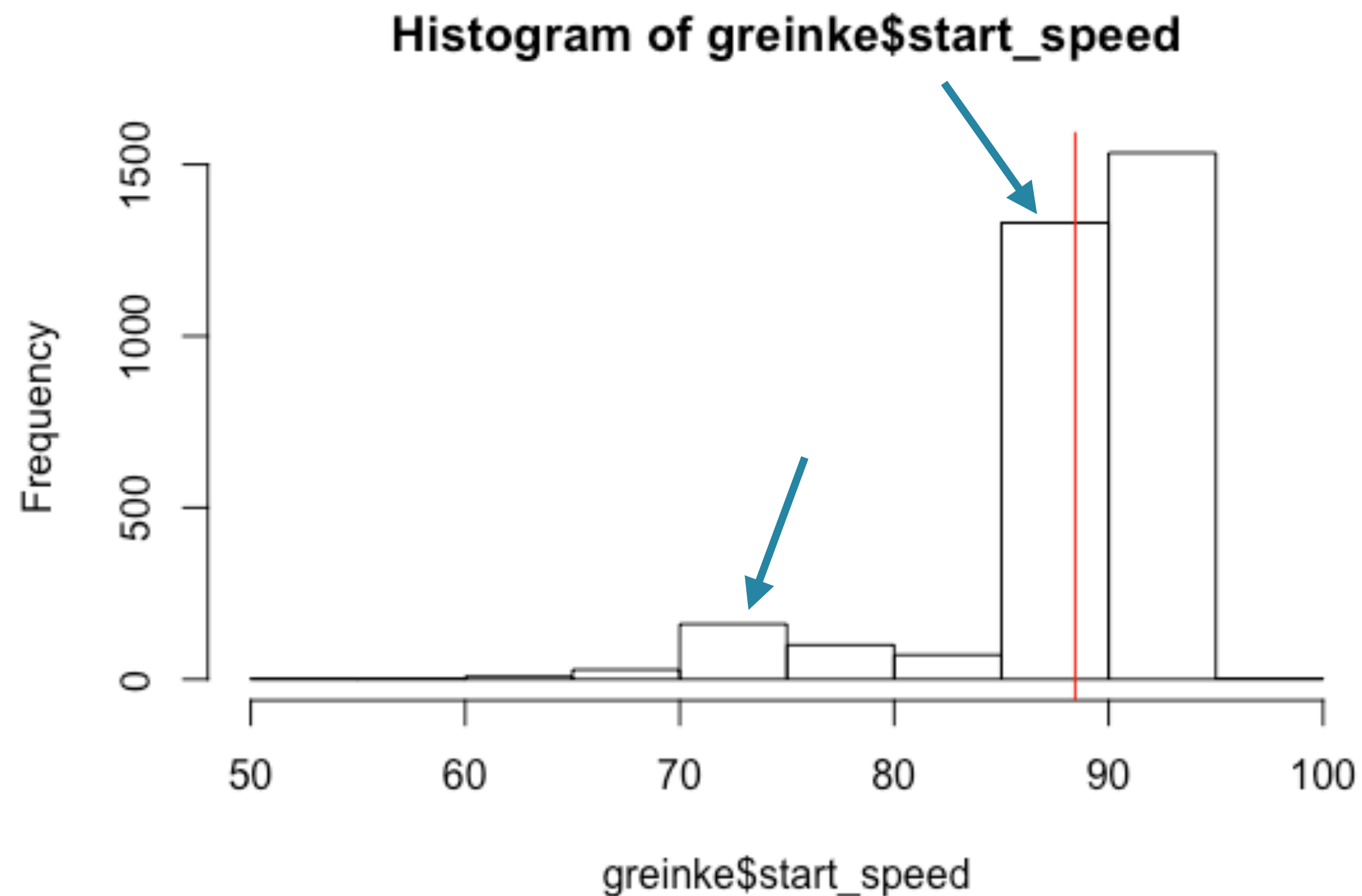

Histograms

```
> hist(greinke$start_speed)
```



Drawing a vertical line with `abline()`

```
> hist(greinke$start_speed)
> abline(v = mean(greinke$start_speed), col = "red")
```



Using `ifelse()`

```
> greinke$slider <- ifelse(greinke$pitch_type == "SL", 1, 0)
> head(greinke[, c(4, length(greinke))])
  pitch_type slider
1         FF      0
2         FF      0
3         FF      0
4         SL      1
5         FF      0
6         SL      1
> greinke$not_slider <- ifelse(greinke$pitch_type != "SL", 1, 0)
> head(greinke[, c(4, length(greinke))])
  pitch_type not_slider
1         FF          1
2         FF          1
3         FF          1
4         SL          0
5         FF          1
6         SL          0
```

Using `subset()`

```
> greinke$slider <- ifelse(greinke$pitch_type == "SL", 1, 0)

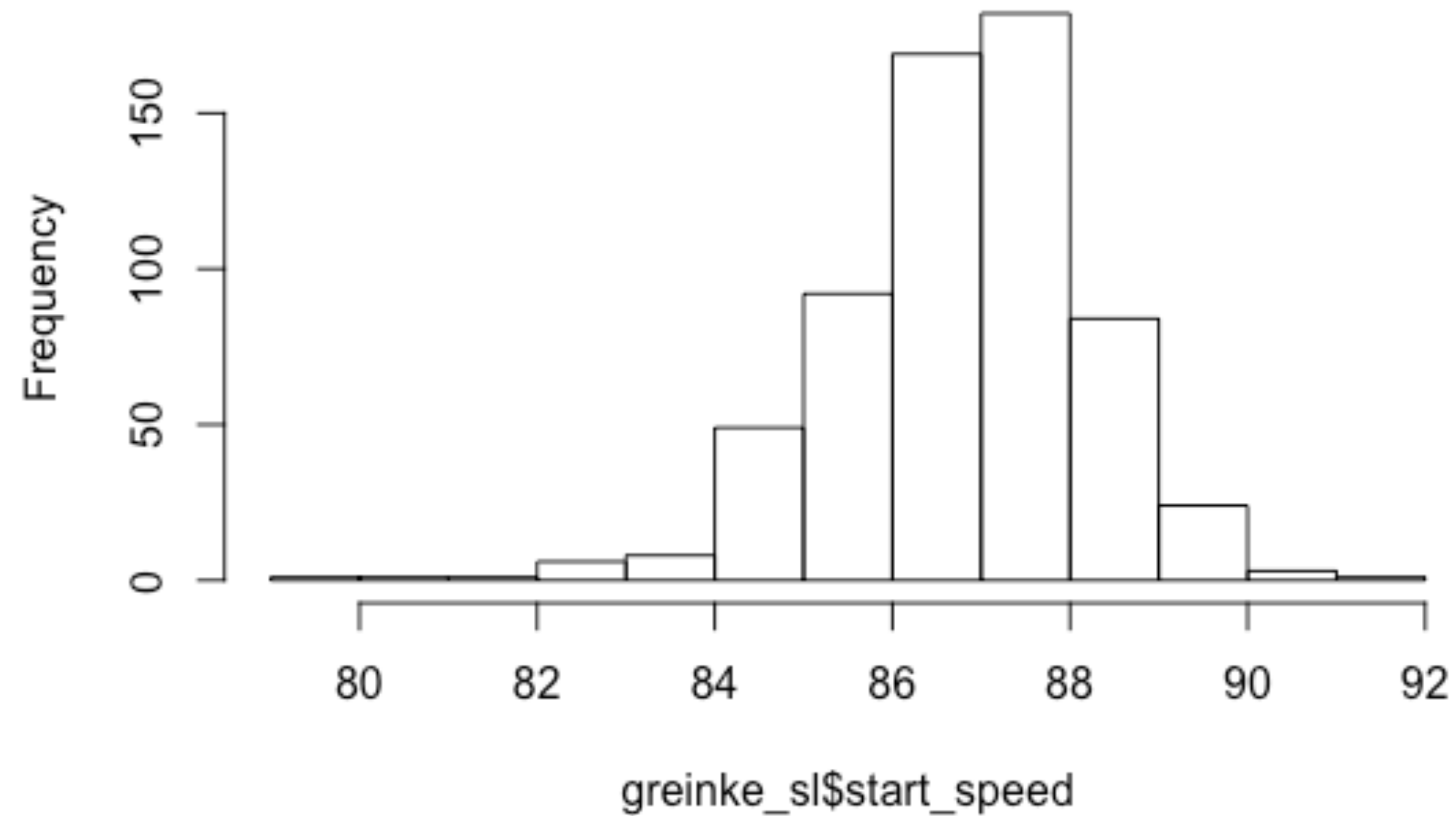
> greinke_sl <- subset(greinke, slider == 1)
> summary(greinke_sl$pitch_type)
CH  CU  EP  FF  FT  IN  SL
 0   0   0   0   0   0 621

> greinke_sl <- subset(greinke, pitch_type == "SL")
> summary(greinke_sl$pitch_type)
CH  CU  EP  FF  FT  IN  SL
 0   0   0   0   0   0 621

> hist(greinke_sl$start_speed)
```

Using `subset()`

Histogram of `greinke_sl$start_speed`





EXPLORING PITCH DATA IN R

Let's practice!



EXPLORING PITCH DATA IN R

Using `tapply()` for comparisons

Using `tapply()`

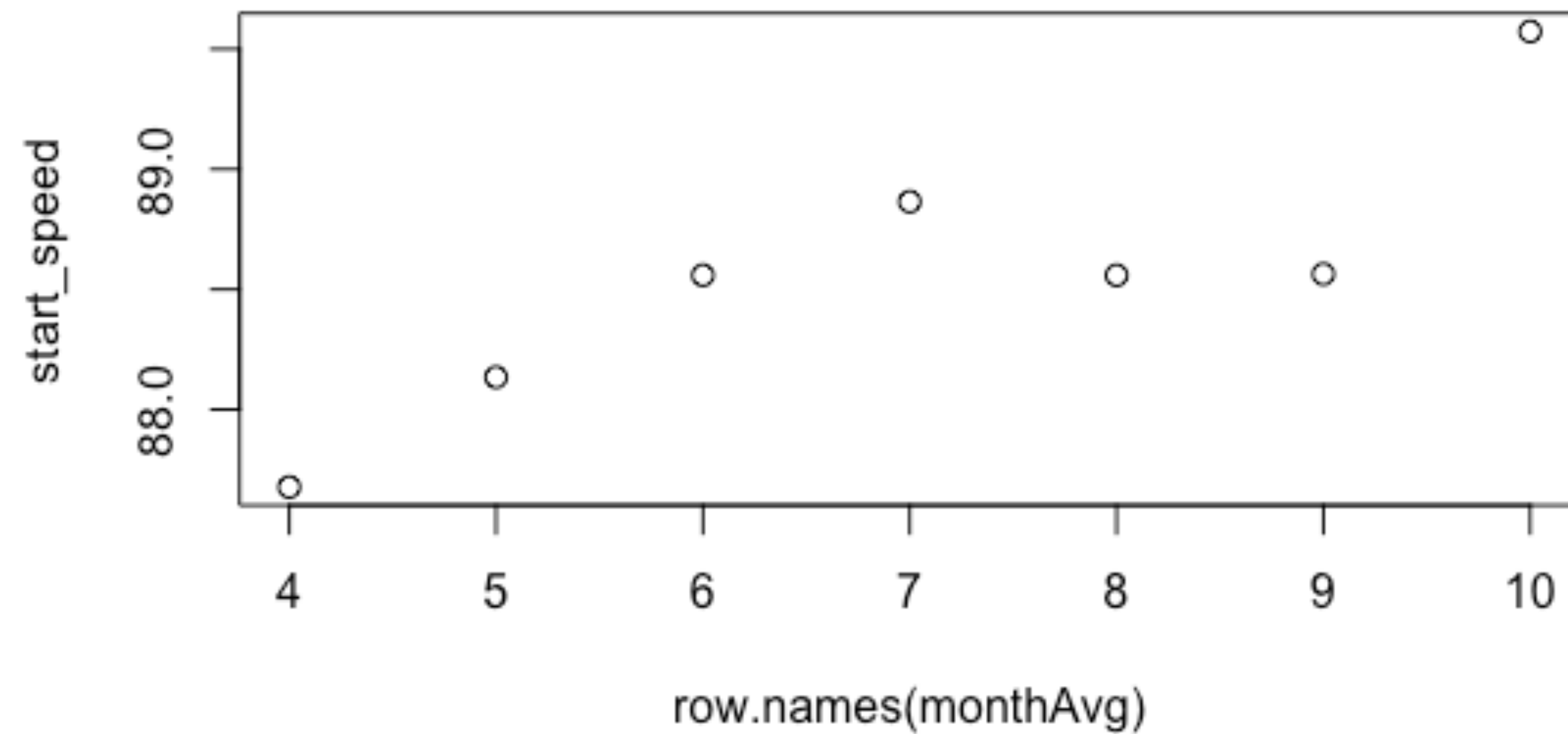
```
> tapply(greinke$start_speed, greinke$month, mean)
      4      5      6      7      8      9     10
87.67758 88.13475 88.55904 88.86489 88.55860 88.56379 89.57315
```



```
> monthAvg <- data.frame(tapply(greinke$start_speed, greinke$month, mean))
> colnames(monthAvg) <- "start_speed"
> monthAvg
  start_speed
4    87.67758
5    88.13475
6    88.55904
7    88.86489
8    88.55860
9    88.56379
10   89.57315
```

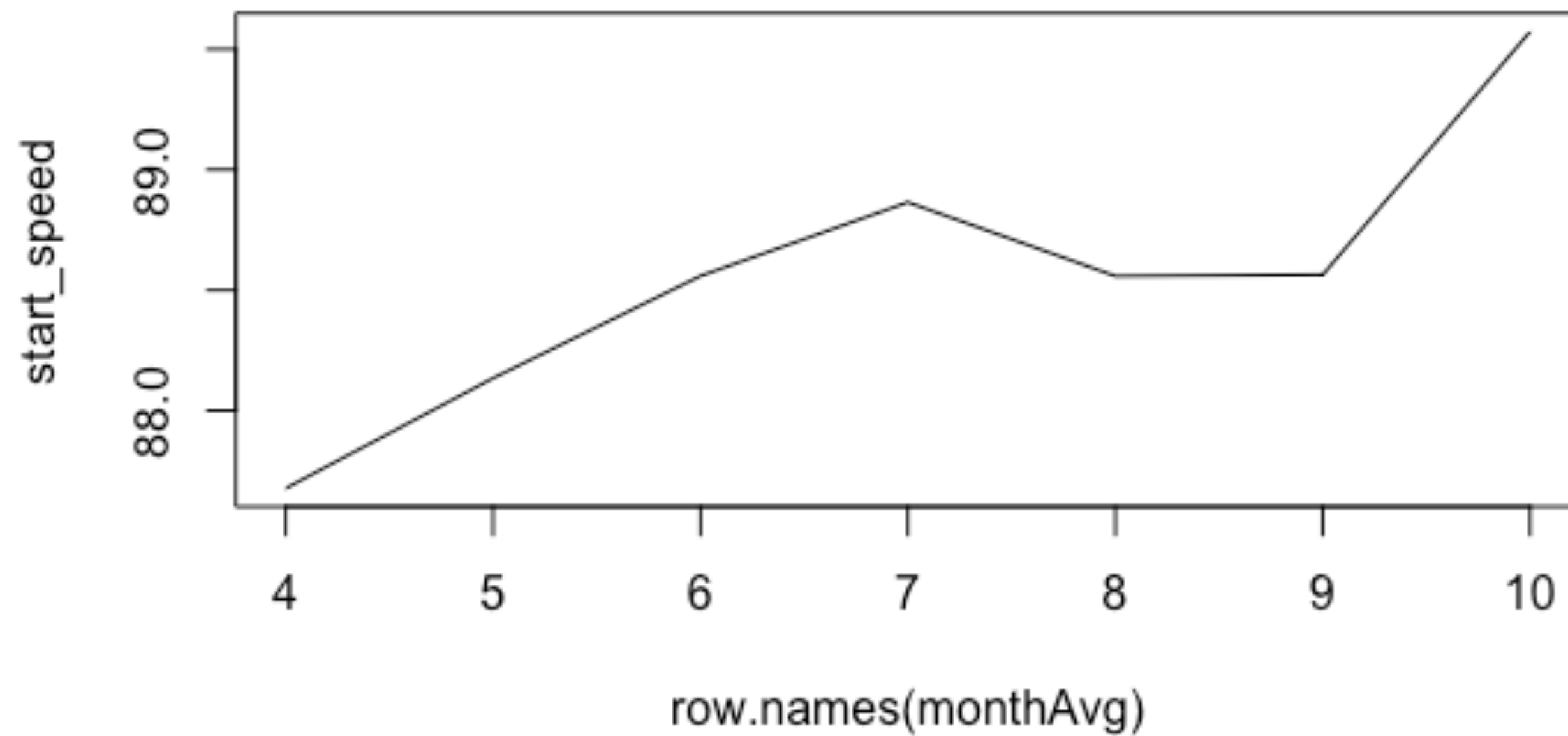

`tapply()` and `plot()` for time series

```
> plot(start_speed ~ row.names(monthAvg), data = monthAvg)
```



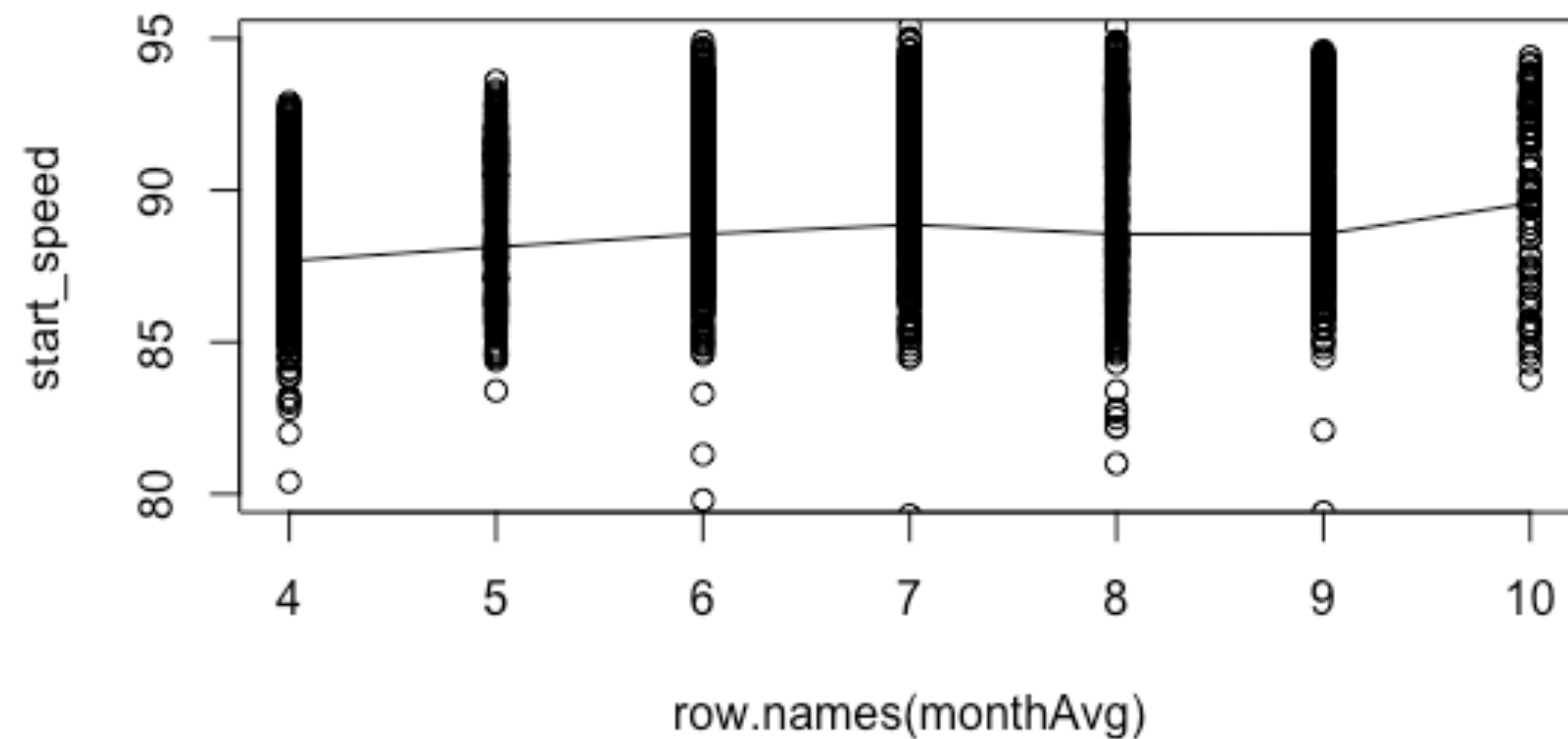
`tapply()` and `plot()` for time series

```
> plot(start_speed ~ row.names(monthAvg), data = monthAvg, type = "l")
```



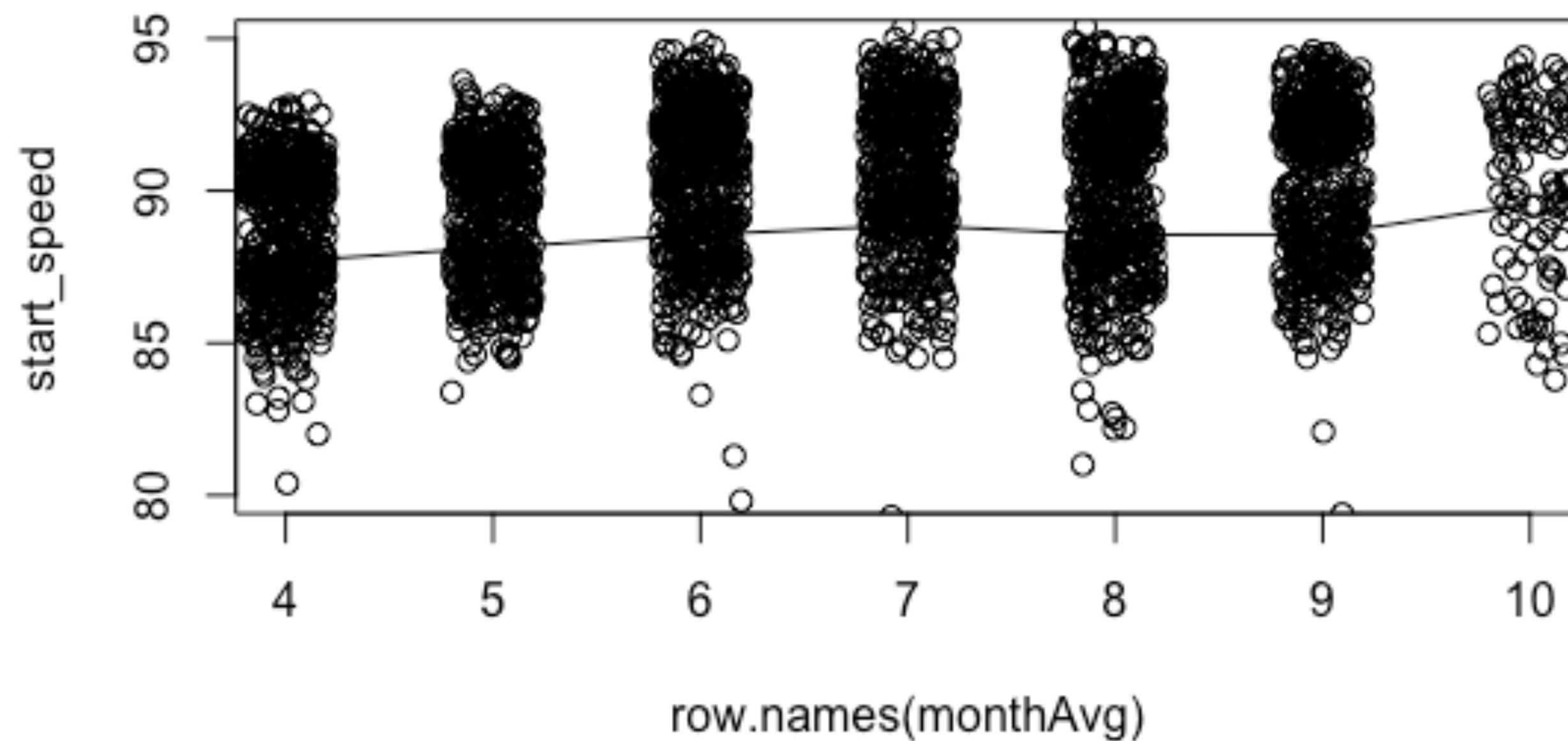
Too much overlap

```
> plot(start_speed ~ month, data = monthAvg)  
> points(greinke$start_speed ~ greinke$month)
```



Jittering points with `jitter()`

```
> plot(start_speed ~ row.names(monthAvg), data = monthAvg,  
       type = "l", ylim = c(80, 95))  
> points(jitter(greinke$start_speed) ~ jitter(greinke$month))
```





EXPLORING PITCH DATA IN R

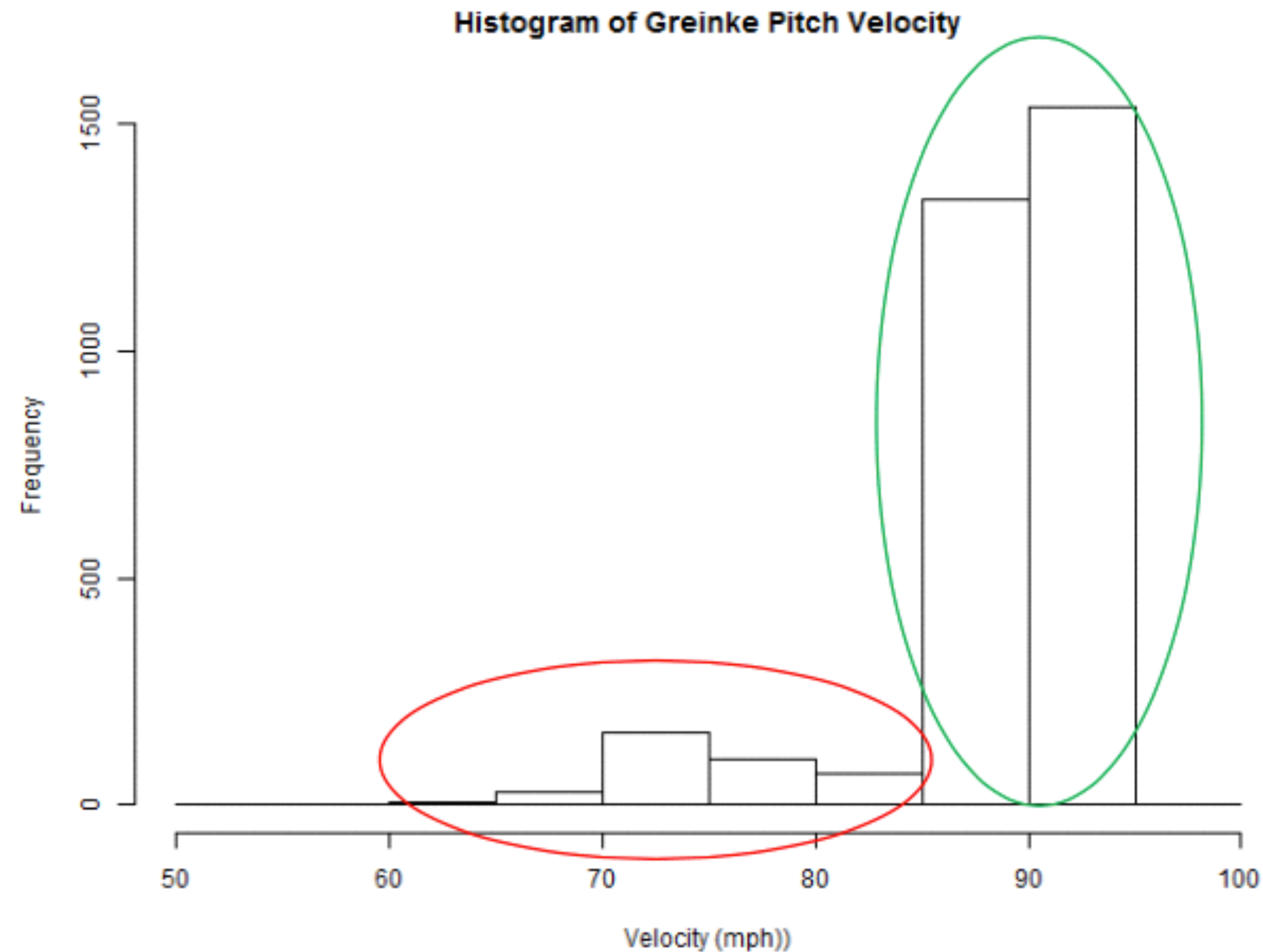
Let's practice!



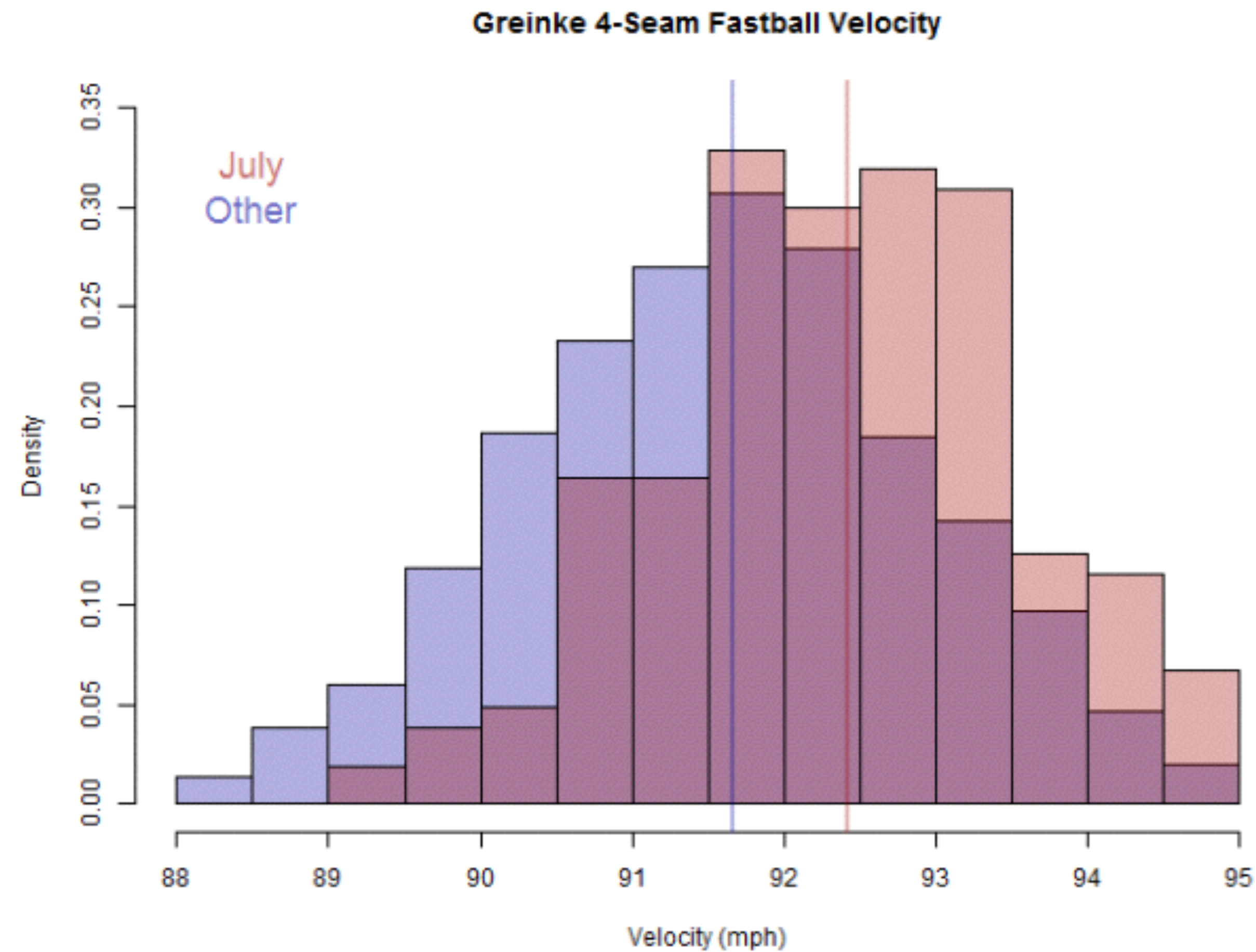
EXPLORING PITCH DATA IN R

Wrap-up

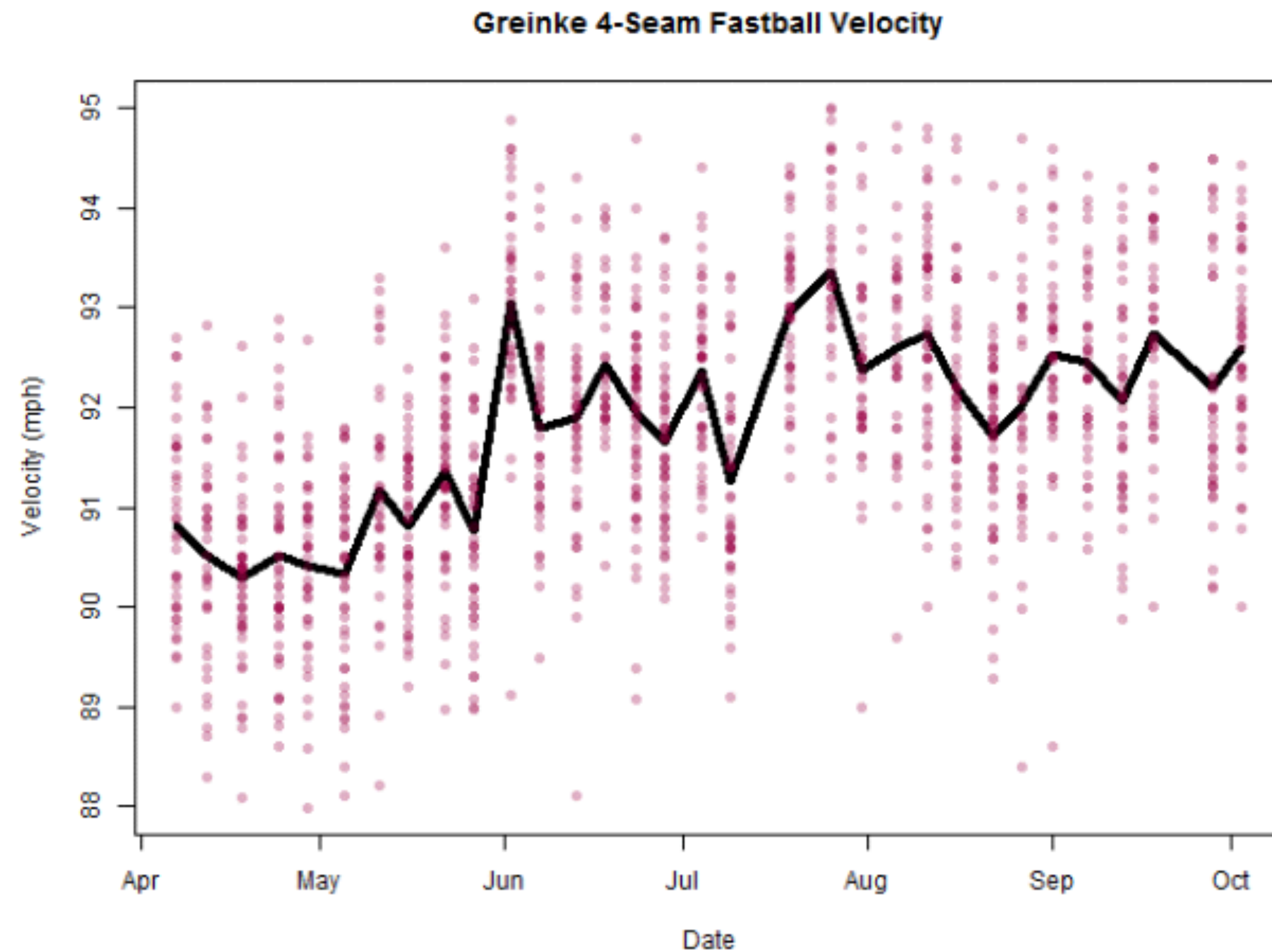
Multimodal velocity distribution



Fastball velocity differences in July



Game-level velocity changes across the year





EXPLORING PITCH DATA IN R

Let's practice!