Analyzing children's relay times for fun and... gold medals

Daniel Kirsch

@kirel

Python Programmer

Daniel Kirsch

@kirel

Haskell Programmer

Daniel Kirsch

@kirel

Ruby Programmer

Daniel Kirsch

@kirel

R Programmer

Daniel Kirsch

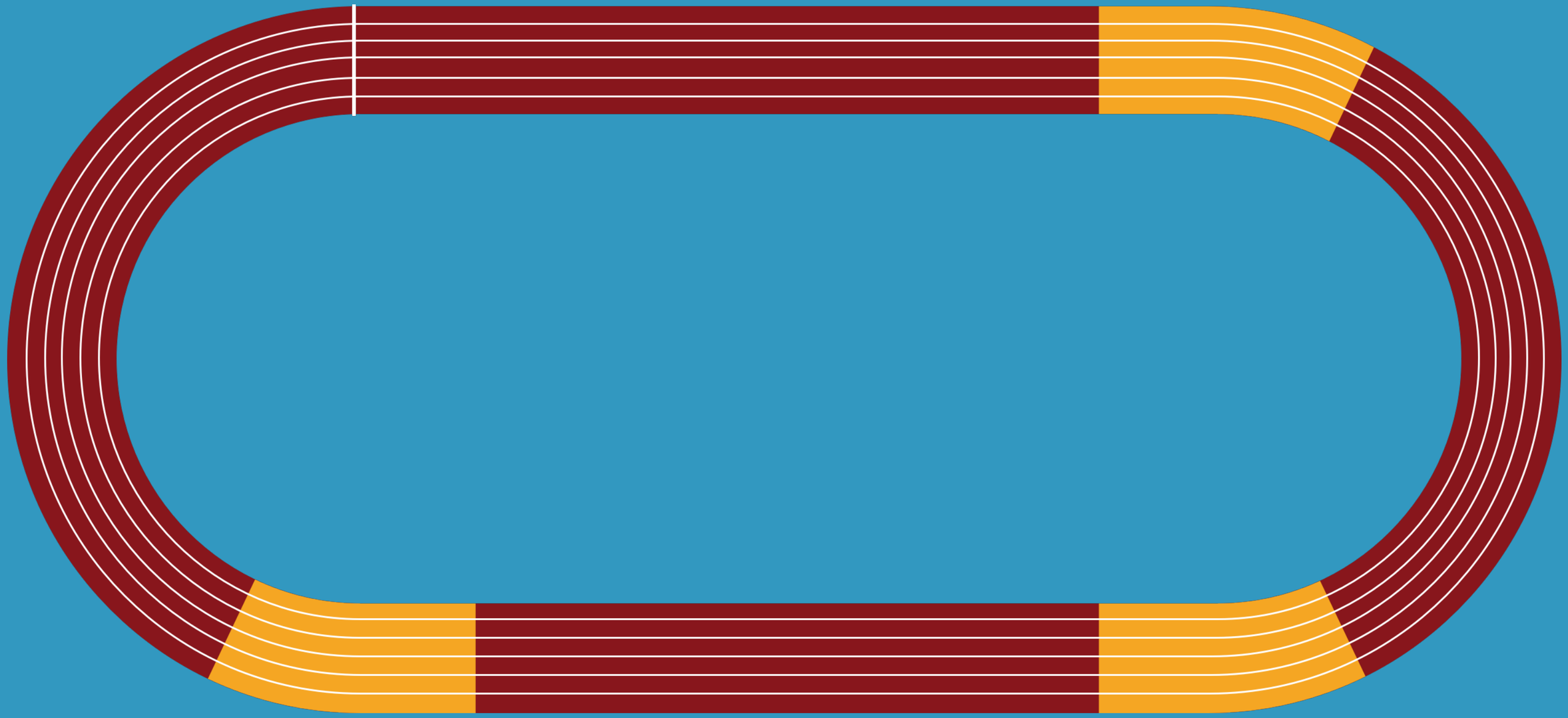
Track & Field Coach


22


844
1

Relay wisdom

- Fast sprinters should run the straight parts
- Let small sprinters do the curves
- Bad hand-off technique goes first or last
- Best starter goes first



Relay wisdom

- **Fast sprinters should run the straight parts**
- Let small sprinters do the curves
- Bad hand-off technique goes first or last
- Best starter goes first

$$\hat{S} = m_1 t_1 + m_2 t_2 + m_3 t_3 + m_4 t_4 + b + \varepsilon$$

Ergebnisliste

DJMM MJU16 Gruppe 2, männliche Jugend U16

Rk.	Name	Punkte
1.	LC Attendorn	6.738,0

Rk.	Name								Punkte
	Bewerb	StNr.	Name	Jg.	Nat.	Leistung	W.	Punkte	
1.	LC Attendorn							6.738,0	
	100m (DM2)	39	Kosch, Jonas	2000		13,55	+0,0	449,0	
	100m (DM2)	41	Kremer, Lukas	2001		14,01	+0,0	413,0	
	800m (DM2)	45	Springob, Maximilian	2000		2:45,56		389,0	
	800m (DM2)	42	Damm, Roman	2000		3:08,13		299,0	
	80m Hürden (DM2)	37	Rennebaum, Antonius	1999		14,96	+0,0	417,0	
	80m Hürden (DM2)	46	Höffer, Sebastian	2000		16,19	+0,0	374,0	
	80m Hürden (DM2)	38	Schneider, Timo	2000		16,24	+0,0	(373,0)	
	Hochsprung (DM2)	45	Springob, Maximilian	2000		1,50 m		479,0	
	Hochsprung (DM2)	36	Krämer, Jason	1999		1,50 m		479,0	
	Weitsprung (DM2)	39	Kosch, Jonas	2000		4,93 m	+0,0	488,0	
	Weitsprung (DM2)	41	Kremer, Lukas	2001		4,00 m	+0,0	388,0	
	Kugelstoßen (DM2)	46	Höffer, Sebastian	2000		8,31 m		393,0	
	Kugelstoßen (DM2)	37	Rennebaum, Antonius	1999		8,10 m		384,0	
	4x100m (DM2)		LC Attendorn 1			55,28		856,0	
	37, Rennebaum Antonius (1999) - 36, Krämer Jason (1999) 38, Schneider Timo (2000) - 39, Kosch Jonas (2000)								
	4x100m (DM2)		LC Attendorn 2			58,12		(751,0)	
	41, Kremer Lukas (2001) - 42, Damm Roman (2000) 43, Springob Martin (2001) - 46, Höffer Sebastian (2000)								

Ergebnisliste

DJMM MJU16 Gruppe 2, männliche Jugend U16

Rk.	Name	Punkte
1.	LC Attendorn	6.738,0

Rk.	Name / Bewerb	StNr.	Name	Jg.	Nat.	Leistung	W.	Punkte
1.	LC Attendorn							6.738,0
	100m (DM2)	39	Kosch, Jonas		[illegible]	13,55	+0,0	449,0
	100m (DM2)	41	Kremer, Lukas		[illegible]	14,01	+0,0	413,0
	800m (DM2)	45	Springob, Maximilian		2000	2:45,56		389,0
	800m (DM2)	42	Damm, Roman		2000	3:08,13		299,0
	80m Hürden (DM2)	37	Rennebaum, Antonius		1999	14,96	+0,0	417,0
	80m Hürden (DM2)	46	Höffer, Sebastian		2000	16,19	+0,0	374,0
	80m Hürden (DM2)	38	Schneider, Timo		2000	16,24	+0,0	(373,0)
	Hochsprung (DM2)	45	Springob, Maximilian		2000	1,50 m		479,0
	Hochsprung (DM2)	36	Krämer, Jason		1999	1,50 m		479,0
	Weitsprung (DM2)	39	Kosch, Jonas		2000	4,93 m	+0,0	488,0
	Weitsprung (DM2)	41	Kremer, Lukas		2001	4,00 m	+0,0	388,0
	Kugelstoßen (DM2)	46	Höffer, Sebastian		2000	8,31 m		393,0
	Kugelstoßen (DM2)	37	Rennebaum, Antonius		1999	8,10 m		384,0
	4x100m (DM2)		LC Attendorn 1			55,28		856,0
	37, Rennebaum Antonius (1999) - 36, Krämer Jason (1999) 38, Schneider Timo (2000) - 39, Kosch Jonas (2000)							
	4x100m (DM2)		LC Attendorn 2			58,12		(751,0)
	41, Kremer Lukas (2001) - 42, Damm Roman (2000) 43, Springob Martin (2001) - 46, Höffer Sebastian (2000)							

```ruby
require 'nokogiri'; require 'open-uri'; require 'pry'; require 'csv'; require 'awesome_print'

urls = %w(...)

REGEX_RELAY = /\dx\d|Staffel/

CSV.open("relays.csv", "wb") do |csv_relay|
CSV.open("performances.csv", "wb") do |csv_perf|
  csv_perf << %w[url tournament tournament_year gender club event nr name year performance pts]
  csv_relay << %w[url tournament tournament_year gender club event performance nr1 name1 year1 nr2 name2 year2 nr3 name3 year3 nr4 name4 year4]
  urls.each do |url|
    txt = open(url)
    doc = Nokogiri::HTML(txt)

    d,m,y = doc.css('td.h1').text.match(/(\d+)\.(\d+)\.(\d+)/).captures
    tournament_year = y.to_i

    tournaments = doc.css('p.ev1').map do |ev1|
      tournament = ev1.css('> a').text
    end.reject { |t| t.nil? || t.empty? }

    genders = tournaments.map { |t|  t =~ /(weiblich|innen)/i ? 'w' : 'm' }

    tables = doc.css('table').select do |table|
      table.css('.tdTEvent, .tdTEvent2').any?
    end.each_with_index.map do |table, i|
      tournament = tournaments[i]
      gender = genders[i]
      table.css('tr').slice_before do |tr|
        tr.css('.hdClub, .hdClub2').any?
      end.map do |slice|
        headerRow = slice.shift
        club = headerRow.css('.hdClub, .hdClub2').text
        slice.select do |tr|
          tr.css('.tdTEvent, .tdTEvent2').any? &&
          # tr.css('.tdTEvent, .tdTEvent2').text =~ REGEX_SPRINT &&
          tr.css('.tdTEvent, .tdTEvent2').text !~ REGEX_RELAY # no relay
...
```

```
performances = read.csv('performances.csv')
numAthletes = performances[c('name', 'year')] %>% table %>% `!=`(0) %>% colSums %>% sum
```

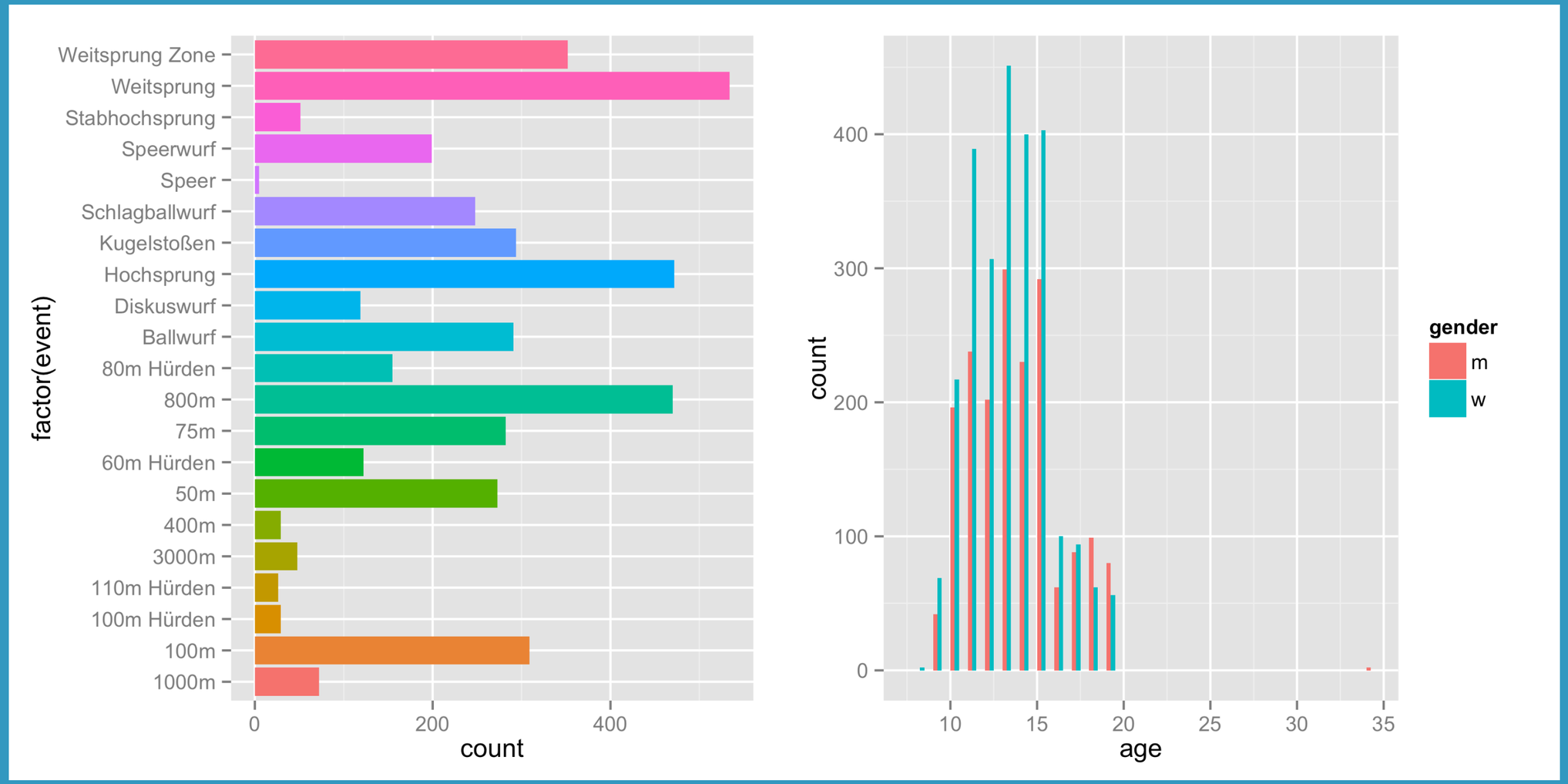
4380 results from 2112 athletes


Volksbank
Rheiner Anzeiger
BARMER GEK
678

```
performances %>% head
```

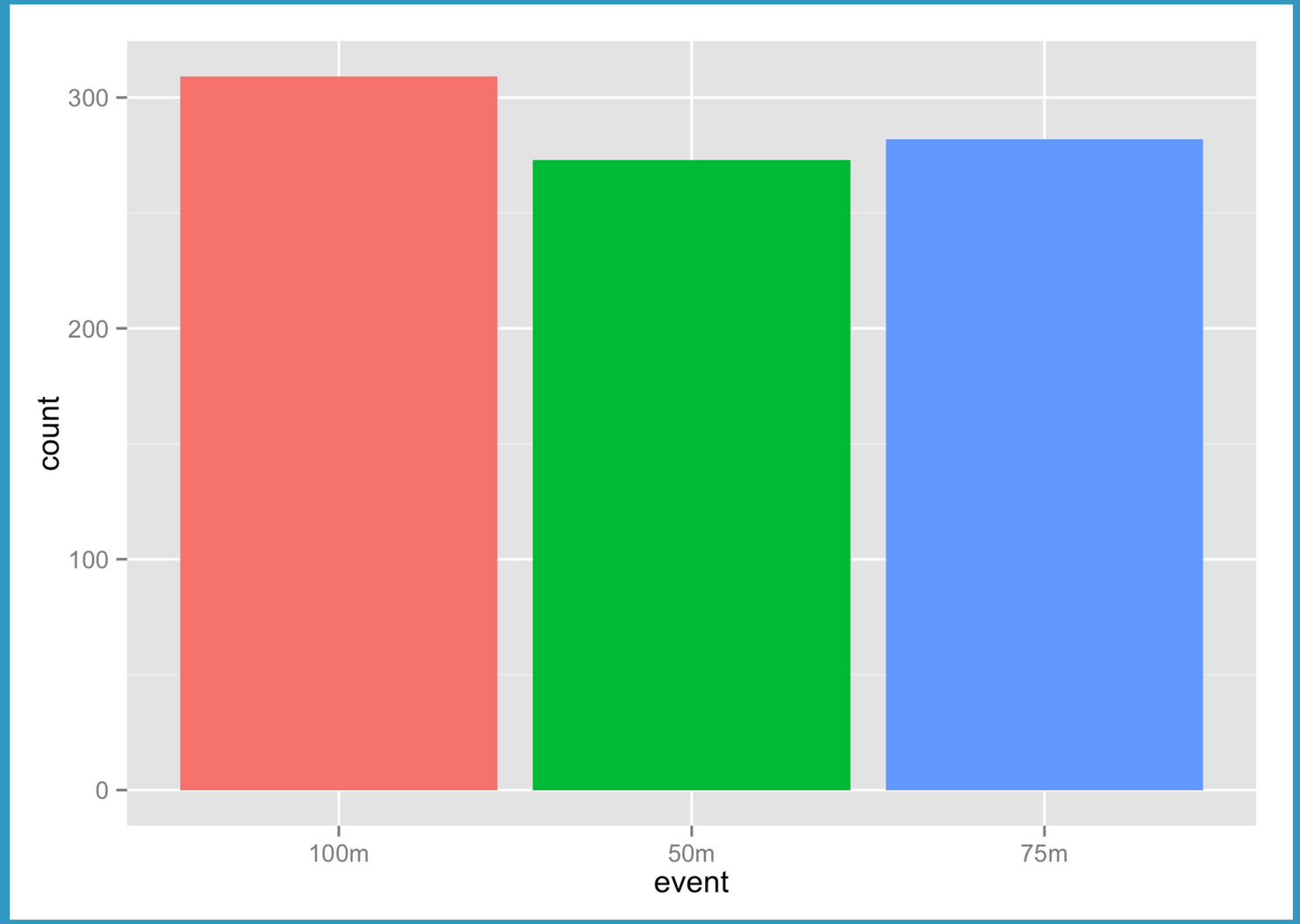
```
##                                                                 url
## 1 https://www.flvwdialog.de/daten/2014/mannschaftswertung9485.html
## 2 https://www.flvwdialog.de/daten/2014/mannschaftswertung9485.html
## 3 https://www.flvwdialog.de/daten/2014/mannschaftswertung9485.html
## 4 https://www.flvwdialog.de/daten/2014/mannschaftswertung9485.html
## 5 https://www.flvwdialog.de/daten/2014/mannschaftswertung9485.html
## 6 https://www.flvwdialog.de/daten/2014/mannschaftswertung9485.html
##                                  tournament tournament_year gender
## 1 DJMM MJU16 Gruppe 2, männliche Jugend U16            2014      m
## 2 DJMM MJU16 Gruppe 2, männliche Jugend U16            2014      m
## 3 DJMM MJU16 Gruppe 2, männliche Jugend U16            2014      m
## 4 DJMM MJU16 Gruppe 2, männliche Jugend U16            2014      m
## 5 DJMM MJU16 Gruppe 2, männliche Jugend U16            2014      m
## 6 DJMM MJU16 Gruppe 2, männliche Jugend U16            2014      m
##           club      event nr                  name year performance   pts
## 1 LC Attendorn       100m 39         Kosch, Jonas 2000       13.55 449,0
## 2 LC Attendorn       100m 41        Kremer, Lukas 2001       14.01 413,0
## 3 LC Attendorn       800m 45 Springob, Maximilian 2000        2.00 389,0
## 4 LC Attendorn       800m 42          Damm, Roman 2000        3.00 299,0
## 5 LC Attendorn 80m Hürden 37  Rennebaum, Antonius 1999       14.96 417,0
## 6 LC Attendorn 80m Hürden 46    Höffer, Sebastian 2000       16.19 374,0
```

```
p1 = ggplot(performances, aes(x=factor(event), fill=factor(event))) + geom_histogram() + coord_flip() + guides(fill=FALSE)
p2 = ggplot(performances, aes(x=age, fill=gender)) + geom_histogram(binwidth=.5, position="dodge")
grid.arrange(p1, p2, nrow=1)
```



```
performances[performances$event=='Speer',]$event = 'Speerwurf'
performances = performances %>% filter(age < 21)
```

```
sprints = performances %>% filter(event == "100m" | event == "75m" | event == "50m")# %>% filter(age > 12)
ggplot(sprints, aes(x=event, fill=event)) + geom_histogram(binwidth=.5, position="dodge") + guides(fill=FALSE)
```



```
sprints100m = sprints %>% filter(event == '100m' & age > 12)
```

308 100m sprint results

```
relays = read.csv('relays.csv')
relays$age1 = relays$tournament_year - relays$year1
relays$age2 = relays$tournament_year - relays$year2
relays$age3 = relays$tournament_year - relays$year3
relays$age4 = relays$tournament_year - relays$year4
relays4x100m = relays %>% filter(event == "4x100m")
```

147 4x100m results

```
relays4x100m %>% head
```

```
##                                                                 url
## 1 https://www.flvwdialog.de/daten/2014/mannschaftswertung9485.html
## 2 https://www.flvwdialog.de/daten/2014/mannschaftswertung9485.html
## 3 https://www.flvwdialog.de/daten/2014/mannschaftswertung9485.html
## 4       https://www.flvwdialog.de/daten/2014/ergebnisliste9770.htm
## 5       https://www.flvwdialog.de/daten/2014/ergebnisliste9770.htm
## 6       https://www.flvwdialog.de/daten/2014/ergebnisliste9770.htm
##                                tournament tournament_year gender
## 1 DJMM MJU16 Gruppe 2, männliche Jugend U16            2014      m
## 2 DJMM MJU16 Gruppe 2, männliche Jugend U16            2014      m
## 3 DJMM WJU16 Gruppe 2, weibliche Jugend U16            2014      w
## 4        DMM Gruppe 2, männliche Jugend U16            2014      m
## 5        DMM Gruppe 2, weibliche Jugend U18            2014      w
## 6        DMM Gruppe 2, weibliche Jugend U18            2014      w
##            club  event performance nr1              name1 year1 nr2
## 1   LC Attendorn 4x100m       55.28  37 Rennebaum Antonius  1999  36
## 2   LC Attendorn 4x100m       58.12  41       Kremer Lukas  2001  42
## 3   LC Attendorn 4x100m       57.94  48      Dürwald Luisa  2000  49
## 4 SuS Schalke 96 4x100m       52.20  93         Laufer Jan  1999  92
## 5 TV Jahn Siegen 4x100m       52.15  25     Wetter Pauline  1998  23
## 6 TV Jahn Siegen 4x100m       58.35  18          Keil Emma  2000  19
##             name2 year2 nr3             name3 year3 nr4            name4
## 1    Krämer Jason  1999  38     Schneider Timo  2000  39      Kosch Jonas
## 2      Damm Roman  2000  43    Springob Martin  2001  46 Höffer Sebastian
## 3    Gieseler Kim  1999  56        Stuff Luisa  2000  50    Elles Johanna
## 4     Ince Yannik  2000  94   Sommerfeld Felix  2000  95       Heisel Tim
## 5 Dicker Caroline  2000  24    Kathreiner Jule  1999  26     Bieler Geena
## 6   Heide Fabiola  1998 128 Hartmann Elisabeth  1997  21 Schiebisch Carla
##   year4 age1 age2 age3 age4
## 1  2000   15   15   14   14
## 2  2000   13   14   13   14
## 3  1999   14   15   14   15
## 4  1999   15   14   14   15
## 5  1997   16   14   15   17
## 6  2000   14   16   17   14
```

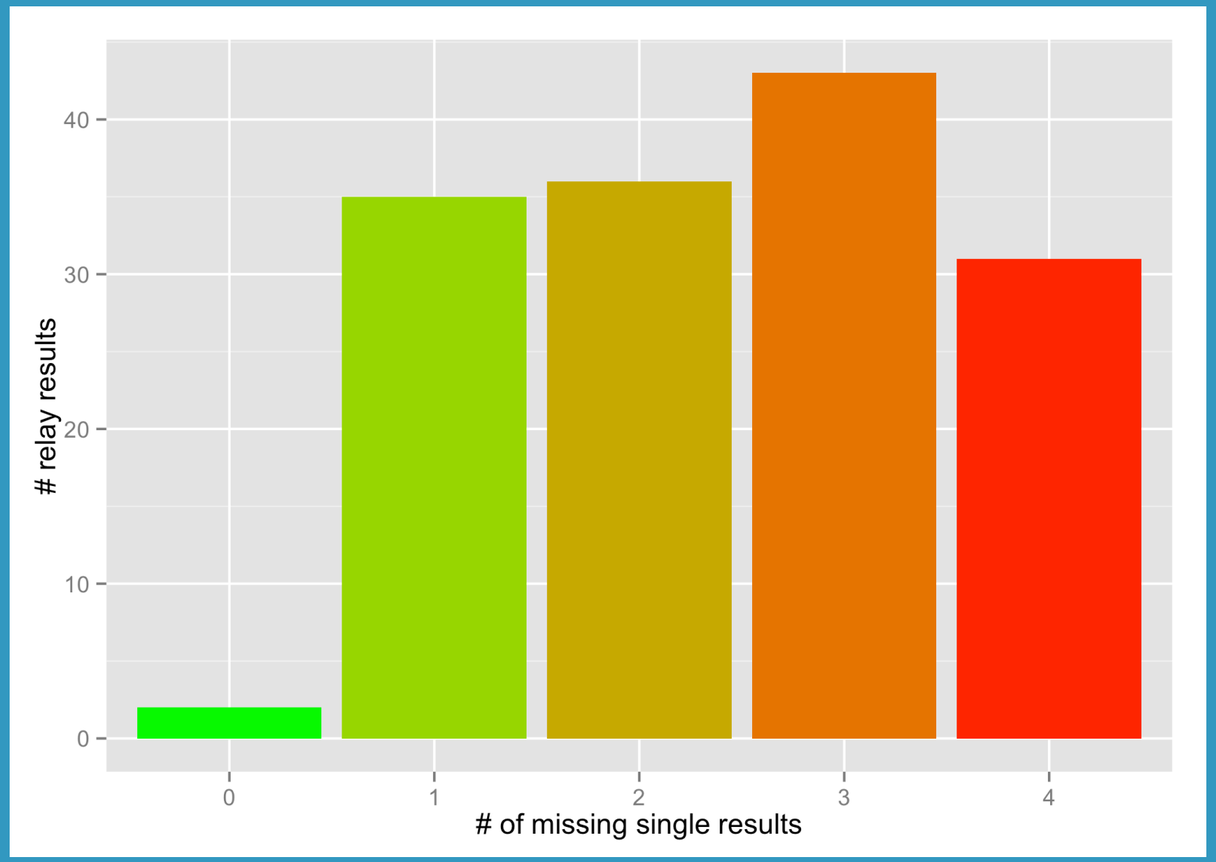

683
SCHAPER JAHNSTADION RHEINE SCHAPER
PATALLA.net

```
relays4x100m = relays4x100m %>%
    left_join(
      sprints100m[c("url", "nr", "performance")] %>%
        rename(nr1=nr) %>% rename(performance1=performance)
      , by=c("nr1", "url")
      ) %>%
    left_join(
      sprints100m[c("url", "nr", "performance")] %>%
        rename(nr2=nr) %>% rename(performance2=performance)
      , by=c("nr2", "url")
      ) %>%
    left_join(
      sprints100m[c("url", "nr", "performance")] %>%
        rename(nr3=nr) %>% rename(performance3=performance)
      , by=c("nr3", "url")
      ) %>%
    left_join(
      sprints100m[c("url", "nr", "performance")] %>%
        rename(nr4=nr) %>% rename(performance4=performance)
      , by=c("nr4", "url")
      )
```

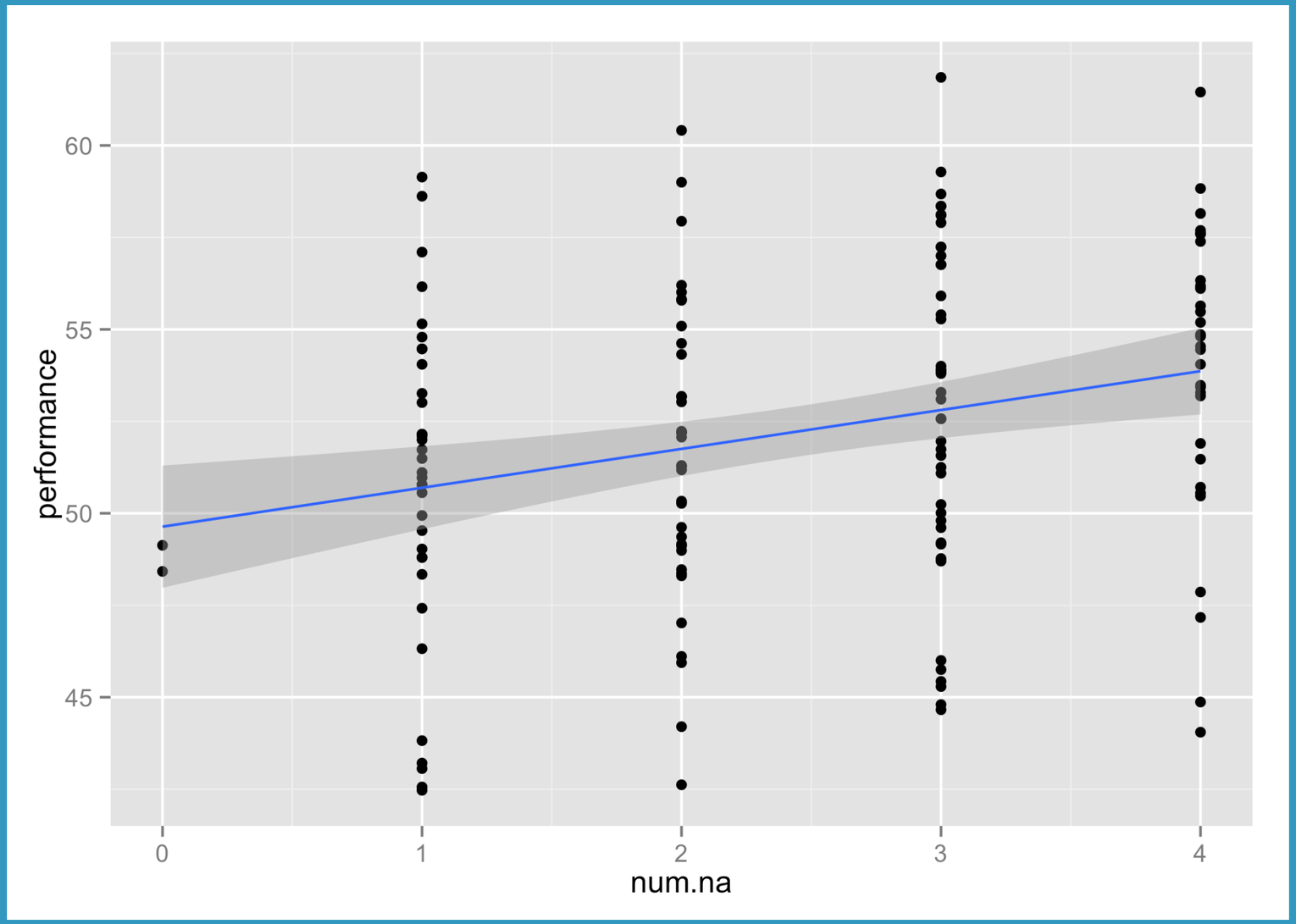
```
relays4x100m[1:20,c('performance', 'performance1', 'performance2', 'performance3', 'performance4')]
```

```
##    performance performance1 performance2 performance3 performance4
## 1        55.28           NA           NA           NA        13.55
## 2        58.12        14.01           NA           NA           NA
## 3        57.94        14.54           NA           NA        14.93
## 4        52.20        12.89        13.08           NA           NA
## 5        52.15        14.25        13.53           NA        13.06
## 6        58.35        15.09           NA           NA           NA
## 7        50.78           NA        12.67         12.6        13.26
## 8        56.76           NA           NA           NA        15.24
## 9        52.57           NA        12.89           NA           NA
## 10       57.24           NA        14.61           NA           NA
## 11       54.47        14.59        13.52           NA        13.36
## 12       57.90           NA           NA         13.4           NA
## 13       52.20           NA        12.70           NA        12.60
## 14       56.20        13.90           NA         14.2           NA
## 15       59.00           NA        14.40           NA        14.60
## 16       57.10           NA        14.20         14.8        14.90
## 17       59.14        14.54        15.17           NA        14.19
## 18       51.74           NA           NA         13.1           NA
## 19       55.48           NA           NA           NA           NA
## 20       52.07           NA        13.16           NA        13.65
```

```
num.na = function(...){sum(is.na(...))}
relays4x100m$num.na = relays4x100m[c('performance1', 'performance2', 'performance3', 'performance4')] %>%
  apply(., 1, num.na)
ggplot(relays4x100m, aes(x=factor(num.na))) + geom_histogram(aes(fill=..x..)) +
  scale_fill_gradient("Count", low = "green", high = "red") + guides(fill=FALSE) +
  labs(y="# relay results", x="# of missing single results")
```



```
ggplot(relays4x100m, aes(x=num.na, y=performance)) + geom_point() + geom_smooth(method='lm')
```



```
relays4x100m$age.median = relays4x100m %>% select(age1:age4) %>% apply(1, median)
summary(lm(performance ~ gender + age.median + num.na, relays4x100m))
```

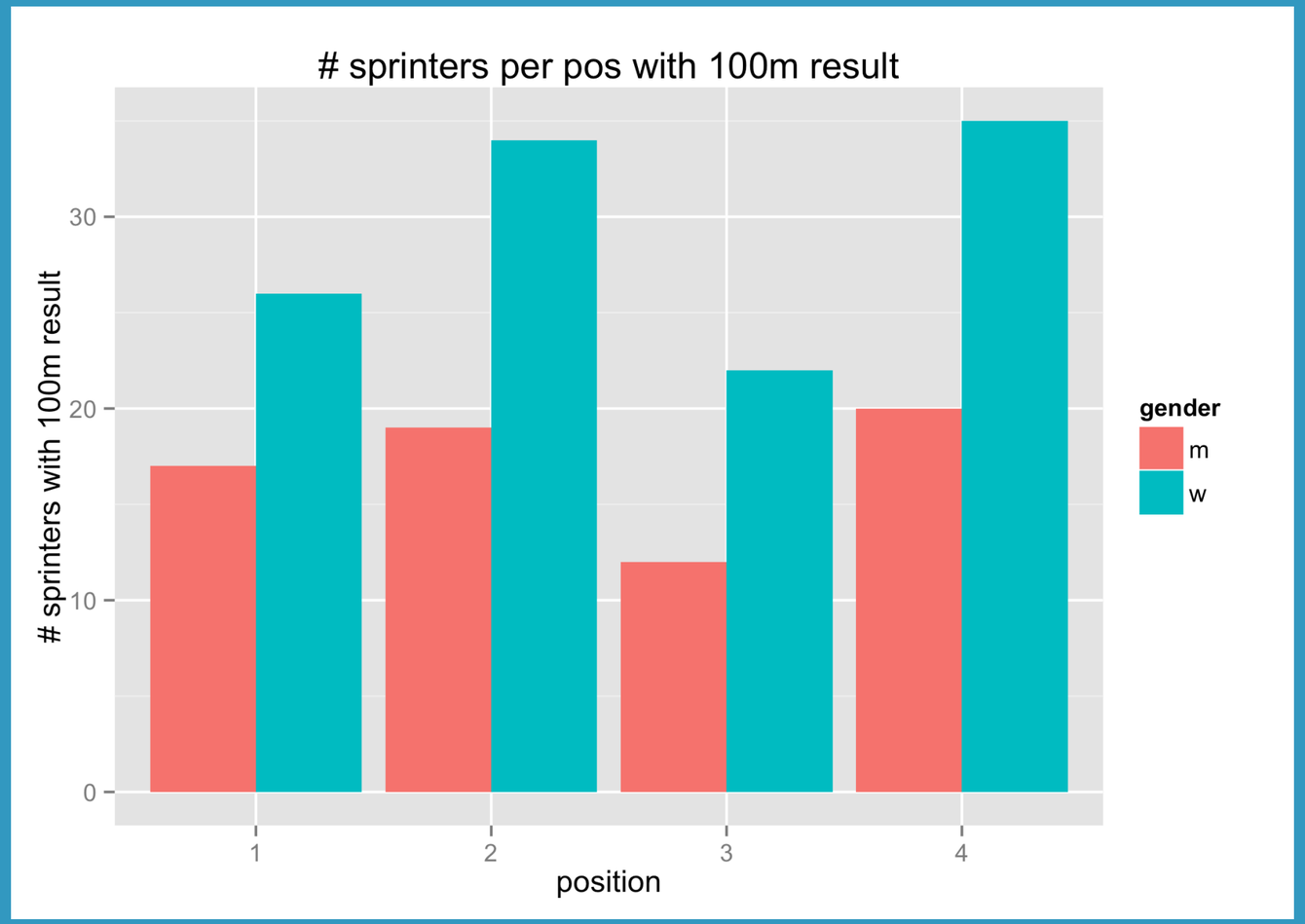
```
##
## Call:
## lm(formula = performance ~ gender + age.median + num.na, data = relays4x100m)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.121 -1.483 -0.293  1.408  6.740
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  67.9231     2.5397  26.744  < 2e-16 ***
## genderw       4.7310     0.4609  10.264  < 2e-16 ***
## age.median   -1.3572     0.1543  -8.794 4.17e-15 ***
## num.na        0.8262     0.1993   4.145 5.80e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.655 on 143 degrees of freedom
## Multiple R-squared:  0.6345, Adjusted R-squared:  0.6268
## F-statistic: 82.74 on 3 and 143 DF,  p-value: < 2.2e-16
```

```
relays4x100m = relays4x100m %>% filter(num.na < 3)
```

We keep 73 relay results

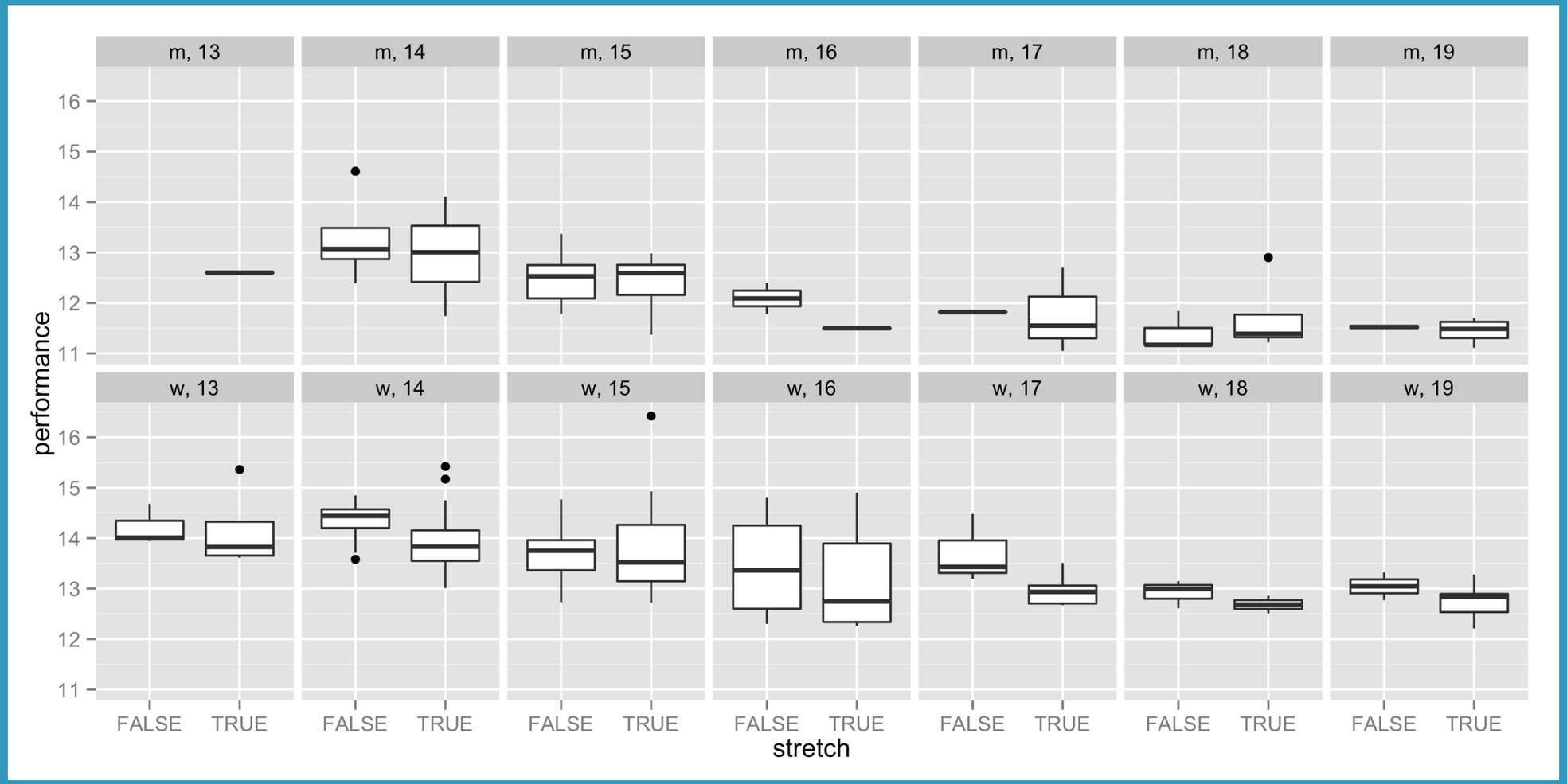
```
relays4x100m.long = relays4x100m[
  c('url', 'tournament', 'gender', 'performance',
    'age1', 'age2', 'age3', 'age4',
    'performance1', 'performance2', 'performance3', 'performance4')] %>%
  rename(age.1=age1) %>%
  rename(age.2=age2) %>%
  rename(age.3=age3) %>%
  rename(age.4=age4) %>%
  rename(performance.1=performance1) %>%
  rename(performance.2=performance2) %>%
  rename(performance.3=performance3) %>%
  rename(performance.4=performance4) %>%
  reshape(varying=5:12, direction='long', timevar='position')
```

```
ggplot(na.omit(relays4x100m.long), aes(x=factor(position), fill=gender)) +
  geom_histogram( position="dodge") +
  scale_x_discrete(labels=c("1","2","3","4")) +
  labs(title = "# sprinters per pos with 100m result", y="# sprinters with 100m result", x="position")
```



Do coaches apply the wisdom?

```
relays4x100m.long$stretch = relays4x100m.long$position == 2 | relays4x100m.long$position == 4
ggplot(relays4x100m.long, aes(x=stretch, y=performance)) +
  geom_boxplot() + facet_wrap(~ gender + age, nrow=2)
```

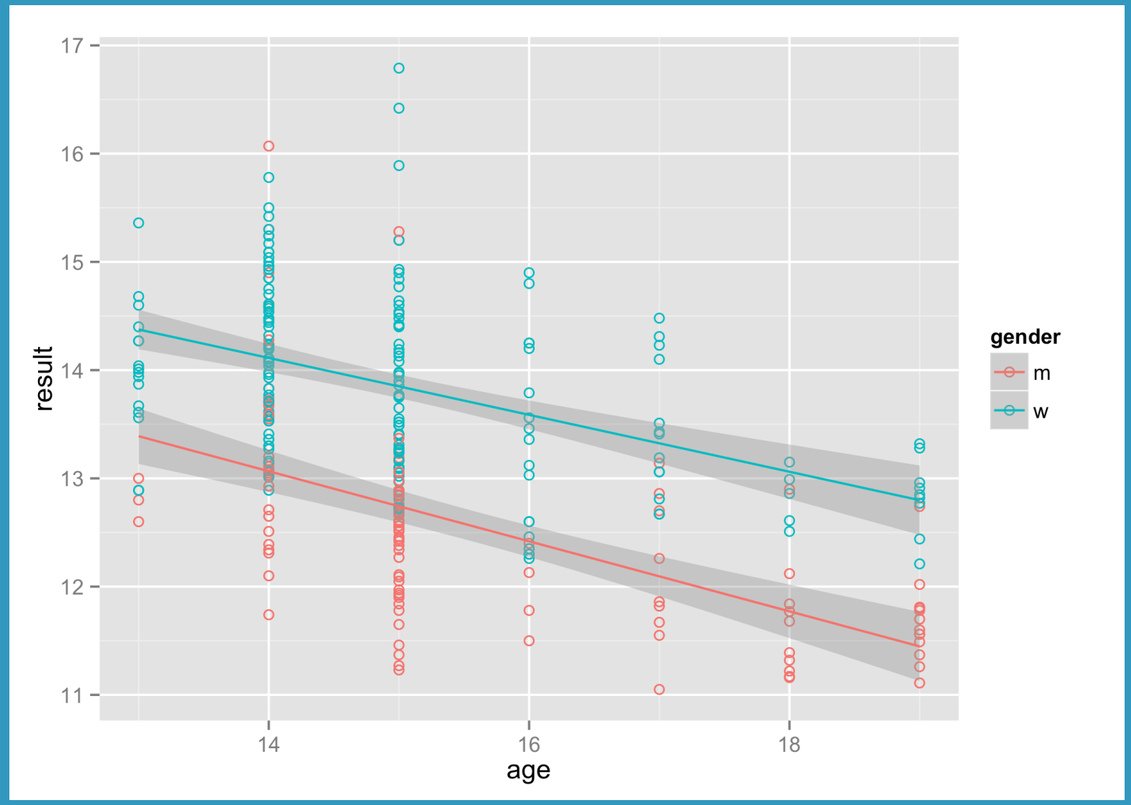


```
lm(performance ~ age + gender + stretch, relays4x100m.long) %>% summary
```

```
##
## Call:
## lm(formula = performance ~ age + gender + stretch, data = relays4x100m.long)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23459 -0.43110 -0.03625  0.34444  2.73939
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.77353    0.46736  35.890   <2e-16 ***
## age         -0.28049    0.02933  -9.563   <2e-16 ***
## genderw      1.24882    0.09641  12.953   <2e-16 ***
## stretchTRUE -0.13446    0.09374  -1.434    0.153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6284 on 181 degrees of freedom
##   (107 observations deleted due to missingness)
## Multiple R-squared:  0.6169, Adjusted R-squared:  0.6106
## F-statistic: 97.16 on 3 and 181 DF,  p-value: < 2.2e-16
```

IMPUTATION

```
ggplot(sprints100m, aes(x=age, y=performance)) +
  geom_point(aes(color=gender), shape=1) +
  geom_smooth(aes(color=gender), method=lm) +
  xlab('age') + ylab('result')
```



```
reg = lm(performance ~ age + gender, sprints100m)
relays4x100m$performance1[is.na(relays4x100m$performance1)] =
  reg %>% predict(relays4x100m[is.na(relays4x100m$performance1),] %>%
  rename(age=age1) %>% rename(performance=performance1))
relays4x100m$performance2[is.na(relays4x100m$performance2)] =
  reg %>% predict(relays4x100m[is.na(relays4x100m$performance2),] %>%
  rename(age=age2) %>% rename(performance=performance2))
relays4x100m$performance3[is.na(relays4x100m$performance3)] =
  reg %>% predict(relays4x100m[is.na(relays4x100m$performance3),] %>%
  rename(age=age3) %>% rename(performance=performance3))
relays4x100m$performance4[is.na(relays4x100m$performance4)] =
  reg %>% predict(relays4x100m[is.na(relays4x100m$performance4),] %>%
  rename(age=age4) %>% rename(performance=performance4))
```

Should coaches apply the wisdom?

```
reg2 = lm(performance ~ performance1+performance2+performance3+performance4 + gender, relays4x100m)
summary(reg2)
```

```
##
## Call:
## lm(formula = performance ~ performance1 + performance2 + performance3 +
##     performance4 + gender, data = relays4x100m)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4544 -0.6929  0.0561  0.9767  3.4621
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -20.1768     4.1124  -4.906 6.23e-06 ***
## performance1   1.5773     0.3886   4.059 0.000131 ***
## performance2   1.7372     0.3287   5.284 1.48e-06 ***
## performance3   0.7398     0.3658   2.022 0.047127 *
## performance4   1.4305     0.3144   4.550 2.32e-05 ***
## genderw       -1.9981     0.5461  -3.659 0.000500 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.411 on 67 degrees of freedom
## Multiple R-squared:  0.8919, Adjusted R-squared:  0.8839
## F-statistic: 110.6 on 5 and 67 DF,  p-value: < 2.2e-16
```

Thank you! Questions?