Case study

Hadley Wickham

Assistant Professor / Dobelman Family Junior Chair

Department of Statistics / Rice University

November 2010

Getting started

```
options(stringsAsFactors = FALSE)
library(plyr)
library(ggplot2)

both <- read.csv("baby-both.csv")
```

Questions

Can we separate dual-sex names from errors?

How has usage changed for dual-sex names?

1. Focus on smaller subset
2. Develop summary statistic
3. Classify names

First task

Too many names (~7000): need to identify smaller subset (~50) likely to be interesting.

Prototype classification on these names, and then extend to full dataset.

First task

Too many names (~7000): need to identify smaller subset (~50) likely to be interesting.

Prototype classification on these names, and then extend to full dataset.

For this task, what attributes of a name are likely to be useful?

Your turn

Summarise each name with the number of years its made the list for both boys and girls, the average proportion of babies given that name.

Which names would you include for further investigation?

```
both_sum <- ddply(both, "name", summarise,
  years = length(name),
  avg_usage = mean(boy + girl) / 2
)

# No point at looking at names that only appear once
both_sum <- subset(both_sum, years > 1)

qplot(years, avg_usage, data = both_sum)
```

```
# Now save our selections

selected_names <- subset(both_sum,
  years > 50 & avg_usage > 0.0005)$name

selected <- subset(both, name %in% selected_names)

nrow(selected) / nrow(both)
```

Your turn

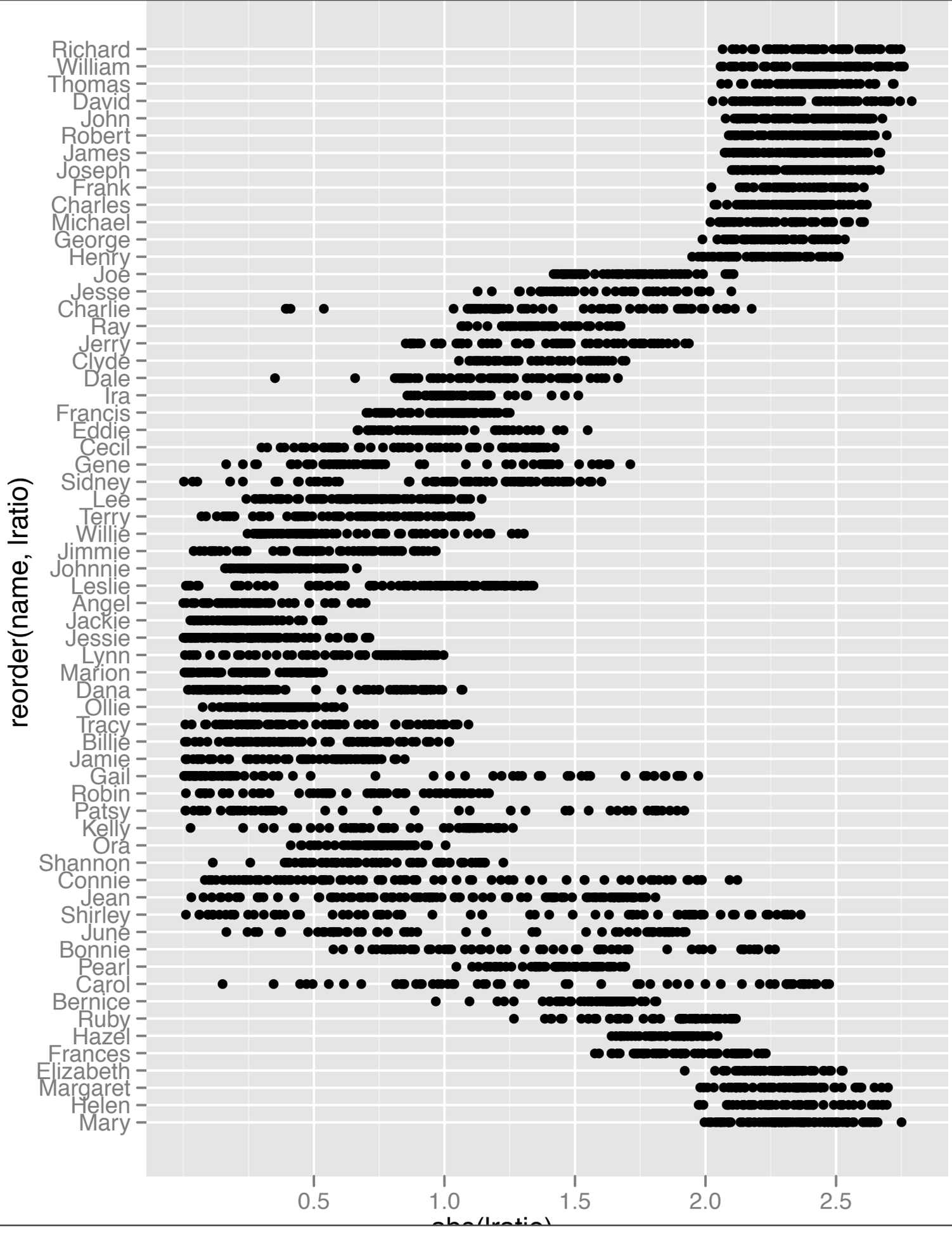
Explore how the gender assignment of these names has changed over time.

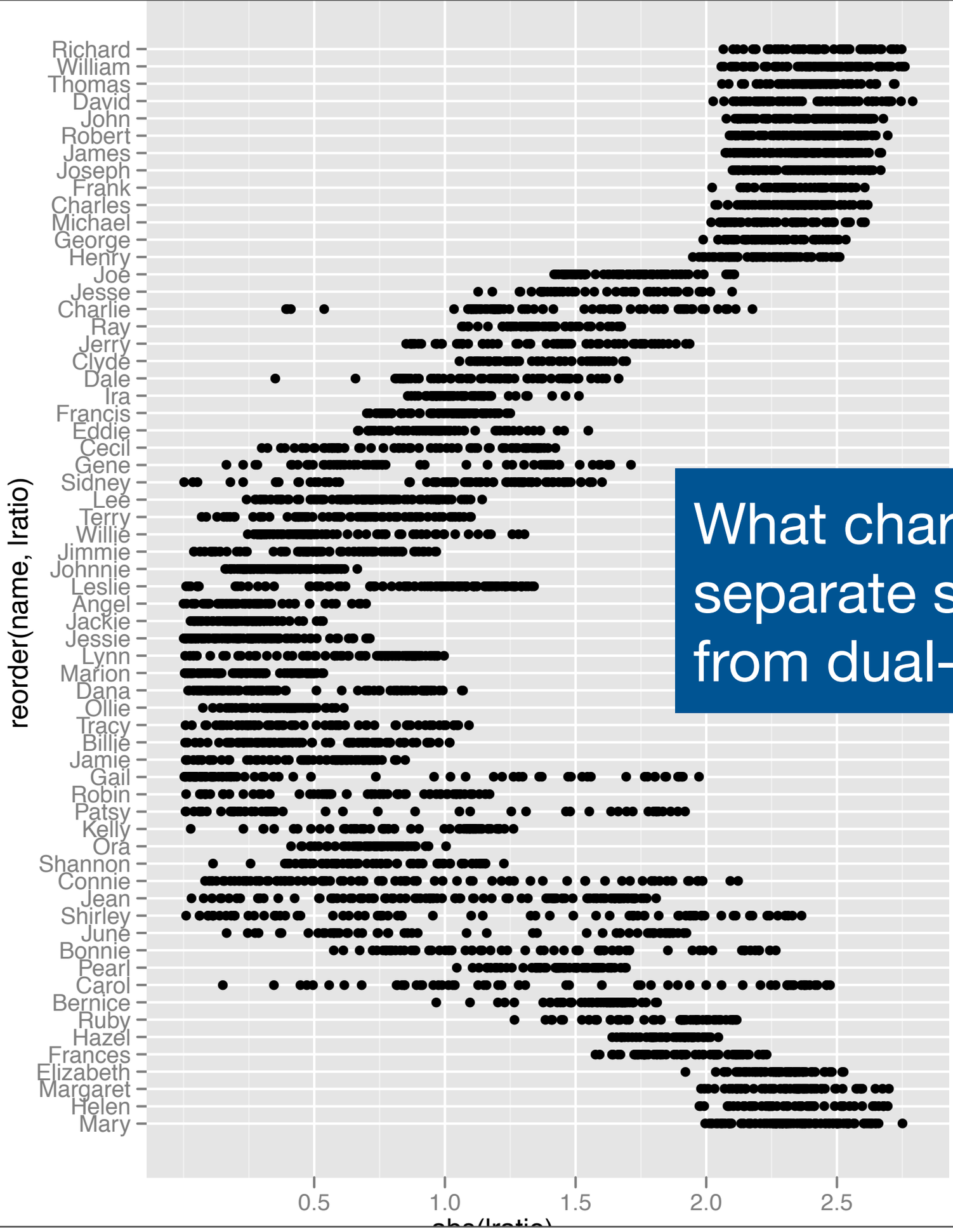
What is a good statistic to use to compare boy popularity to girl popularity?

```
qplot(year, boy - girl, data = selected,
  geom = "line", group = name)
qplot(year, abs(boy - girl), data = selected,
  geom = "line", group = name,
  colour = sign(boy - girl))

qplot(year, boy / girl, data = selected,
  geom = "line", group = name)
qplot(year, log10(boy / girl), data = selected,
  geom = "line", group = name)

selected$lratio <- with(selected, log10(boy / girl))
qplot(lratio, name, data = selected)
qplot(lratio, reorder(name, lratio), data = selected)
qplot(abs(lratio), reorder(name, lratio),
  data = selected)
```


reorder(name, lratio)
Richard
William
Thomas
David
John
Robert
James
Joseph
Frank
Charles
Michael
George
Henry
Joe
Jesse
Charlie
Ray
Jerry
Clyde
Dale
Ira
Francis
Eddie
Cecil
Gene
Sidney
Lee
Terry
Willie
Jimmie
Johnnie
Leslie
Angel
Jackie
Jessie
Lynn
Marion
Dana
Ollie
Tracy
Billie
Jamie
Gail
Robin
Patsy
Kelly
Ora
Shannon
Connie
Jean
Shirley
June
Bonnie
Pearl
Carol
Bernice
Ruby
Hazel
Frances
Elizabeth
Margaret
Helen
Mary
0.5
1.0
1.5
2.0
2.5



What characteristics separate sex-errors from dual-sex names?

Your turn

Compute the mean and range of lratio for each name.

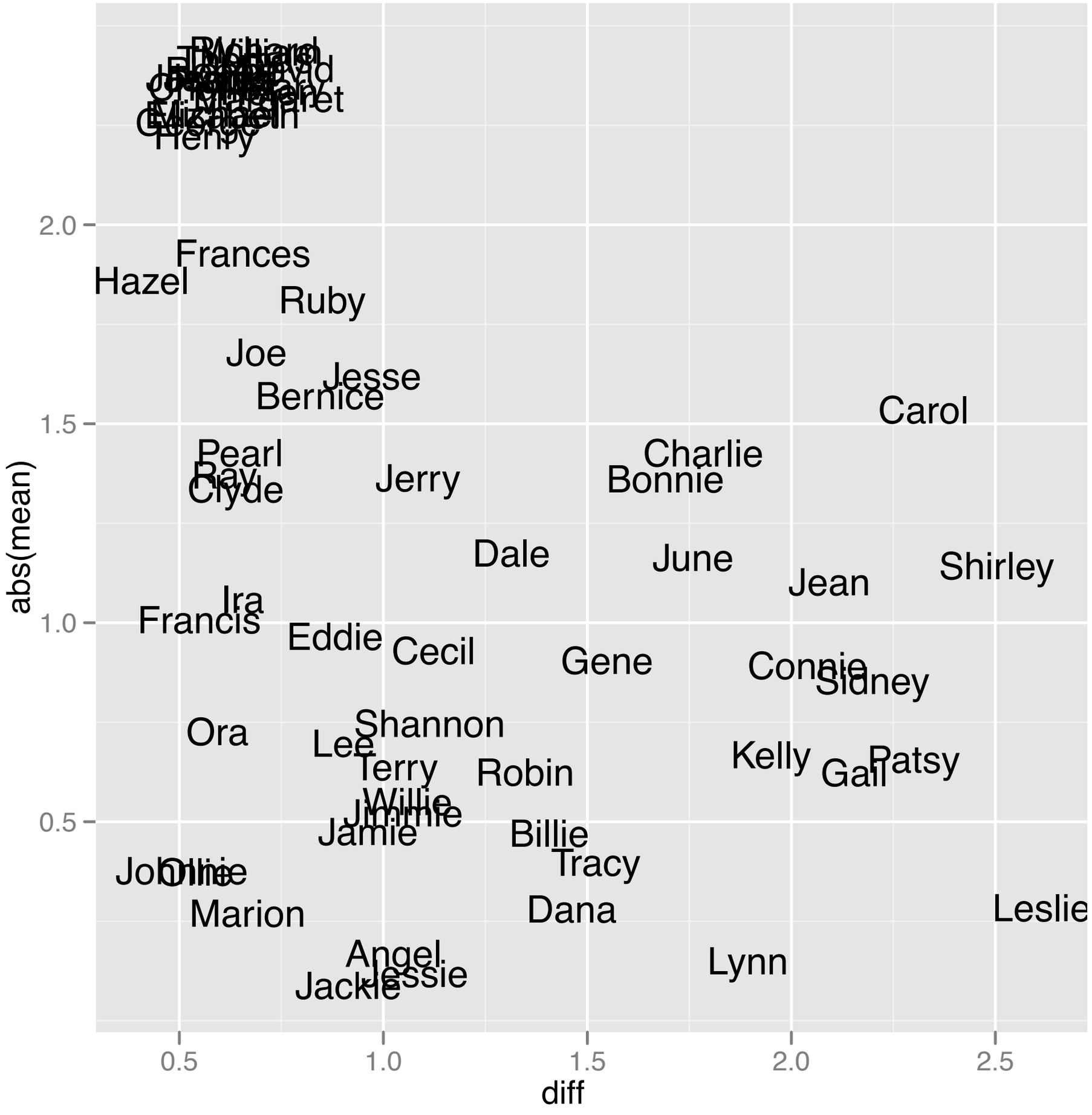
Plot and come up with cutoffs that you think separate the two groups.

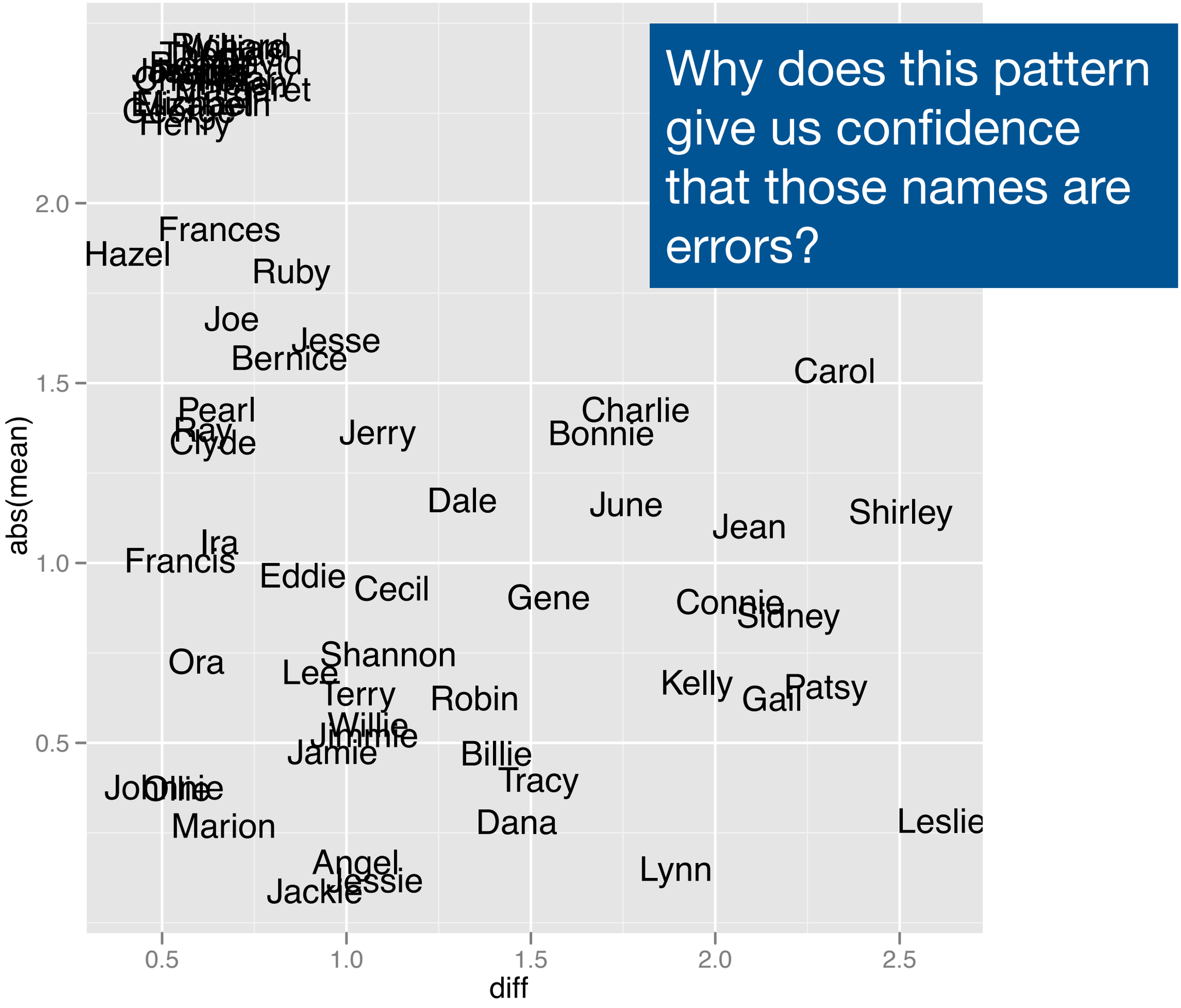
```
rng <- ddply(selected, "name", summarise,
  range = diff(range(lratio, na.rm = T)),
  mean = mean(lratio, na.rm = T)
)

qplot(range, abs(mean), data = rng)
qplot(range, abs(mean), data = rng, geom = "text",
  label = name)

rng$dual <- abs(rng$mean) < 2
arrange(rng, mean, dual)

selected <- join(selected, rng[c("name", "dual")])
```


Henry
Frances
Hazel
Ruby
Joe
Jesse
Bernice
Carol
Pearl
Ray
Clyde
Jerry
Charlie
Bonnie
Dale
June
Jean
Shirley
Ira
Francis
Eddie
Cecil
Gene
Connie
Sidney
Ora
Shannon
Lee
Terry
Robin
Kelly
Gail
Patsy
Willie
Jimmie
Jamie
Billie
Tracy
Marion
Dana
Leslie
Angel
Jessie
Jackie
Lynn
2.0
1.5
1.0
0.5
abs(mean)
0.5
1.0
1.5
2.0
2.5
diff


Why does this pattern give us confidence that those names are errors?
abs(mean)
diff
2.0
1.5
1.0
0.5
0.5
1.0
1.5
2.0
2.5
Frances
Hazel
Ruby
Joe
Jesse
Bernice
Carol
Pearl
Ray
Clyde
Jerry
Charlie
Bonnie
Dale
June
Jean
Shirley
Ira
Francis
Eddie
Cecil
Gene
Connie
Sidney
Ora
Shannon
Lee
Terry
Robin
Kelly
Gail
Patsy
Willie
Jimmie
Jamie
Billie
Tracy
Dana
Leslie
Marion
Angel
Jessie
Jackie
Lynn

```
qplot(year, lratio, data = selected, geom = "line",
  group = name) + facet_wrap(~ dual)

qplot(year, lratio, data = subset(selected, dual),
  geom = "line") + facet_wrap(~ name)

qplot(year, boy / (boy + girl),
  data = subset(selected, dual), geom = "line") +
  facet_wrap(~ name)
```

Your turn

Apply this threshold to all names, not just the few we focussed in on. Does it still seem like a good classification?

What can you say about trends in errors over time?

```
both$lratio <- with(both, log10(boy / girl))
rng <- ddply(both, "name", summarise,
  range = sd(lratio, na.rm = T),
  mean = median(lratio, na.rm = T)
)
rng$dual <- abs(rng$mean) < 1.75
arrange(rng, mean, dual)
both <- join(both, rng[c("name", "dual")])

qplot(year, lratio, data = subset(both, !dual))
qplot(year, abs(lratio), data = subset(both, !dual),
  colour = factor(boy > girl)) +
  geom_smooth(size = 3)
```

Tuesday, 23 November 2010