Baby Names

Giuseppe A. Paleologo

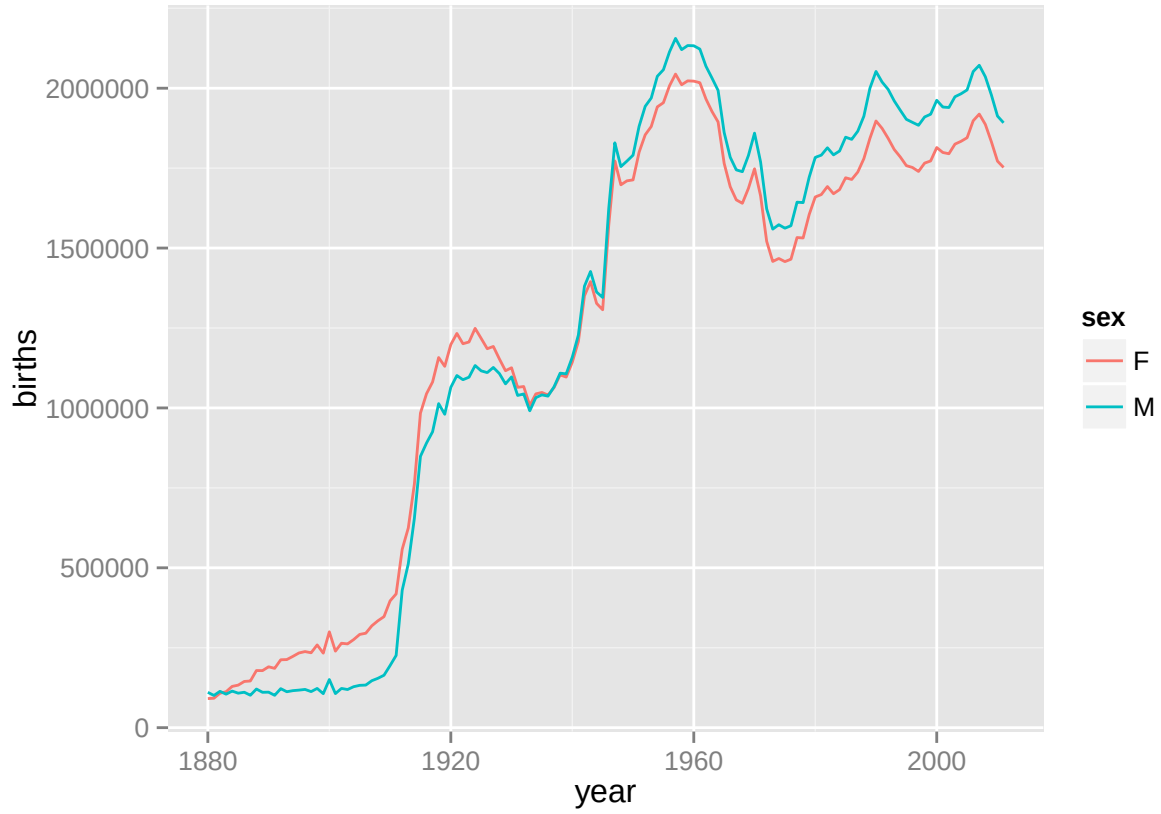
Tuesday, March 17, 2015

Read Files

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(data.table)
library(stringr)
library(magrittr)
setwd('C:/Users/Giuseppe/Dropbox (Personal)/babynames')

years <- 1880:2011
path <- sprintf('names/yob%d.txt', years)
columns <- c('name', 'sex', 'births')

reader <- function(yr){
  x<- fread(sprintf('names/yob%d.txt', yr), data.table=FALSE) %>% set_names(columns)
  x$year <- yr
  x
}

babynames <- lapply(years, reader) %>%  bind_rows
```

Plots births by sex and year

```
library(magrittr)
babynames %>%
  group_by(year, sex) %>%
  summarize(births=sum(births)) %>%
  ggplot(aes(x=year, y=births, color=sex)) + geom_line()
```



Fraction of total names, within sex and year

```
babynames %<>%
  group_by(year, sex) %>%
  mutate(prop=births/sum(births))
```

Top 1000 names by year, sex

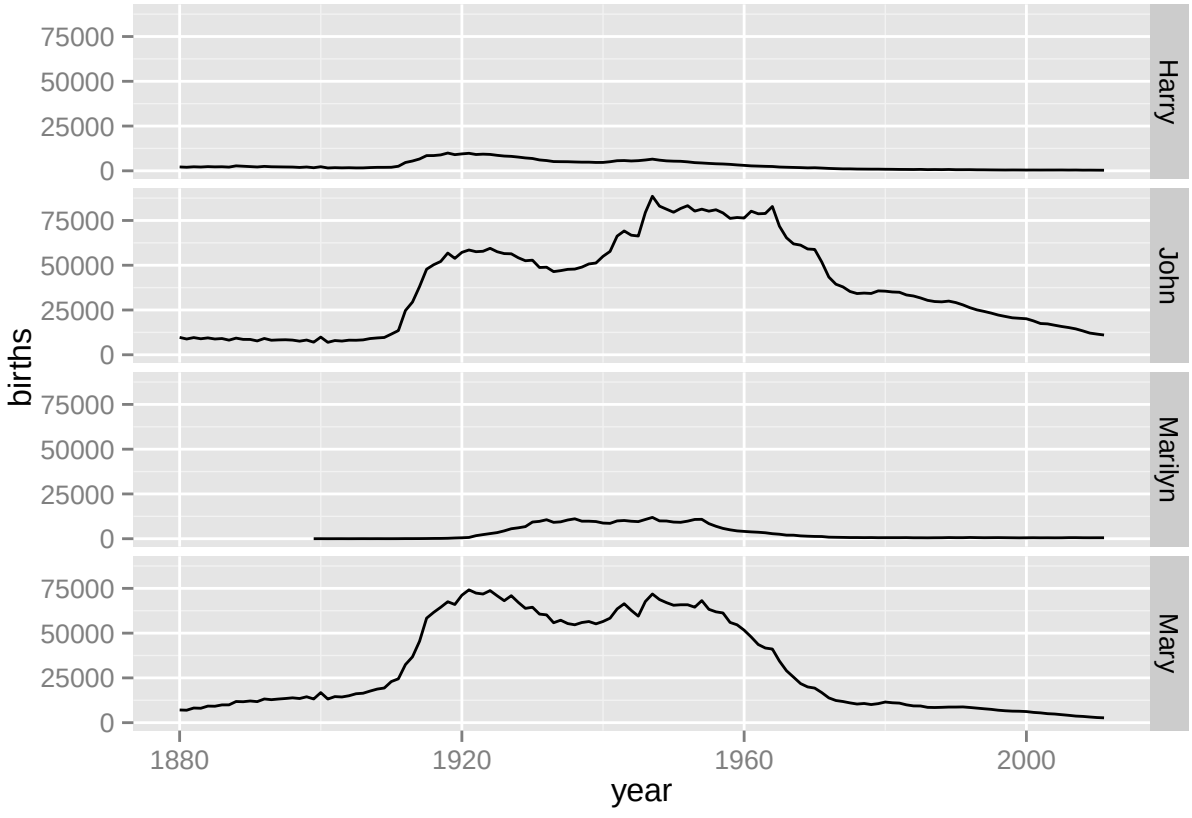
```
top1000 <- babynames %>%
  group_by(year, sex) %>%
  arrange(desc(births)) %>%
  filter(min_rank(desc(births)) <= 1000)
top1000
```

```
## Source: local data frame [267,352 x 5]
## Groups: year, sex
##
##         name sex births year       prop
## 1       Mary   F   7065 1880 0.07764334
## 2       Anna   F   2604 1880 0.02861759
## 3       Emma   F   2003 1880 0.02201268
```

```
## 4  Elizabeth   F   1939 1880 0.02130933
## 5     Minnie   F   1746 1880 0.01918829
## 6   Margaret   F   1578 1880 0.01734199
## 7        Ida   F   1472 1880 0.01617707
## 8      Alice   F   1414 1880 0.01553966
## 9     Bertha   F   1320 1880 0.01450661
## 10     Sarah   F   1288 1880 0.01415493
## ..       ... ...    ...  ...        ...
```

Plots the number of babies named John, Harry, Mary, Marilyn over time

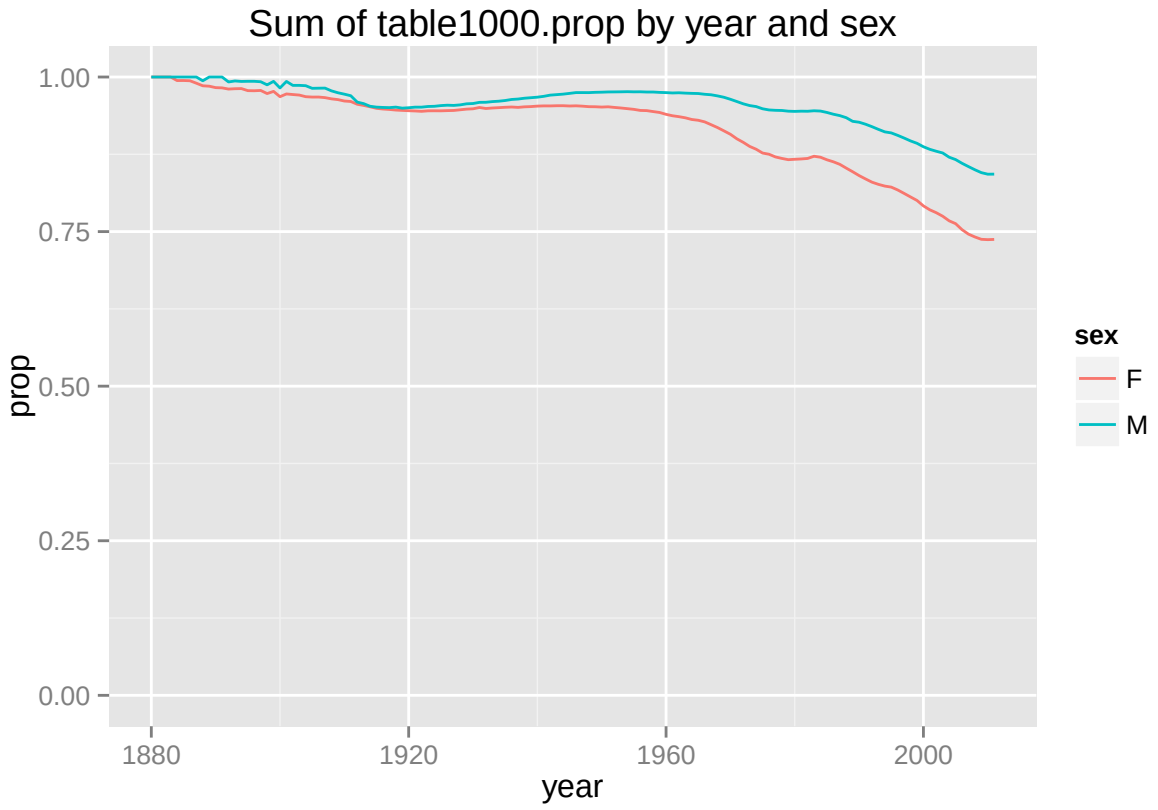
```
babynames %>%
  filter(name %in% c('John', 'Harry', 'Mary', 'Marilyn')) %>%
  group_by(year, name) %>%
  summarize(births=sum(births)) %>%
  ggplot(aes(x=year, y=births)) + geom_line() + facet_grid(name ~ .)
```



Plots the proportion of the top 1000 names as a percentage of total

```
top1000 %>%
  group_by(year, sex) %>%
```

```
  summarize(prop=sum(prop)) %>%
  ggplot(aes(x=year, y=prop, color=sex)) +
    geom_line() +
    ggtitle('Sum of table1000.prop by year and sex') +
    scale_y_continuous(limits=c(0, 1))
```



How many boy names comprise 50% of the total in 2010?

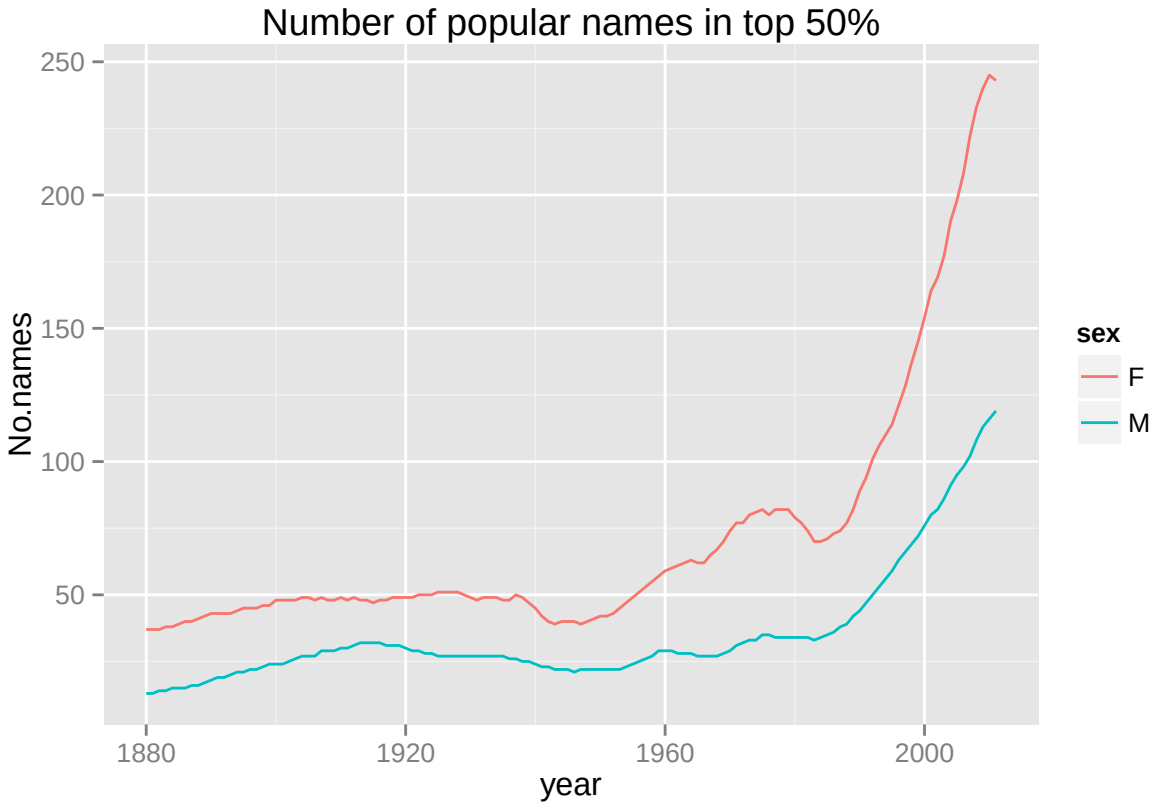
```
top1000 %>%
  filter(year == 2010, sex =='M') %>%
  arrange(desc(births)) %>%
  mutate(totprop = cumsum(prop)) %>%
  filter(totprop <= .50) %>%
  nrow
```

```
## [1] 116
```

Plots number of most popular names used by 50% of boys and girls over time

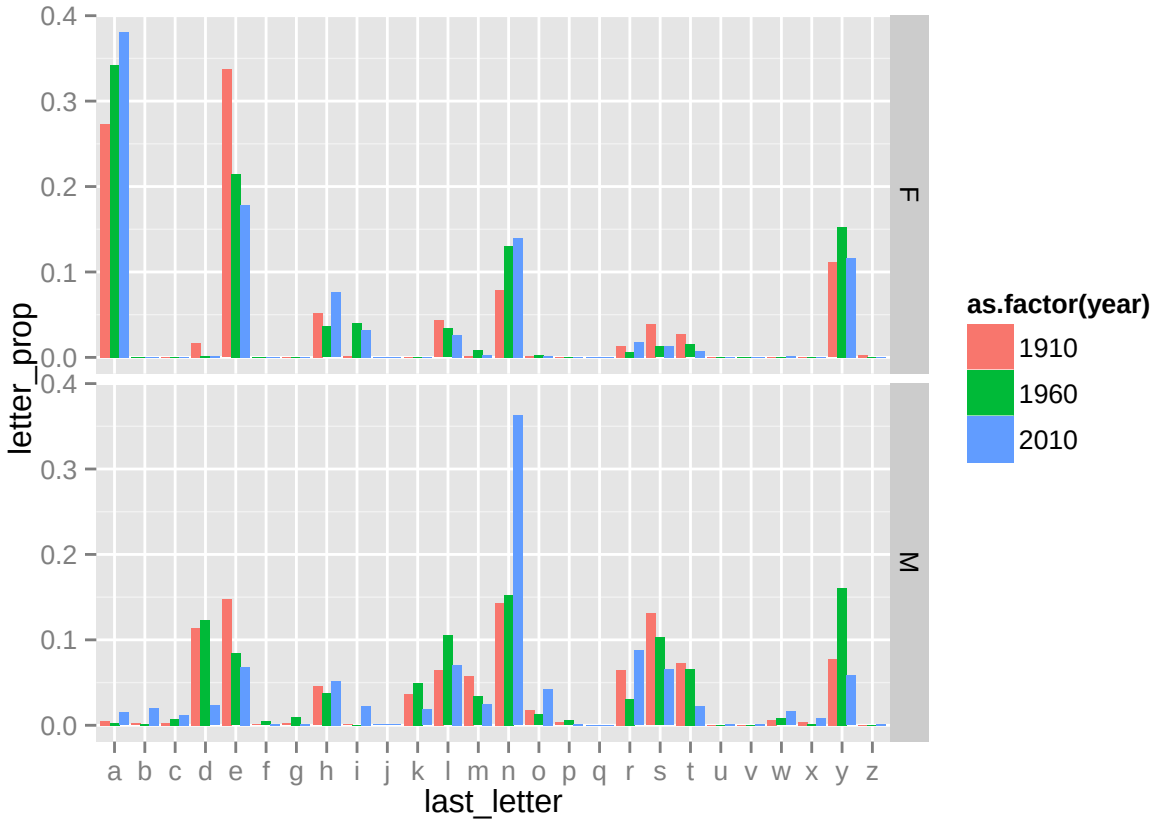
```
get_quantile_count <- function(x, qtle=0.5)
  x %>% sort(decreasing=TRUE) %>% cumsum %>% {.<= qtle} %>% sum
```

```
top1000 %>%
  group_by(year, sex) %>%
  summarize(No.names = get_quantile_count(prop)) %>%
  ggplot(aes(x=year, y=No.names, color=sex)) +
    geom_line() +
    ggtitle('Number of popular names in top 50%')
```



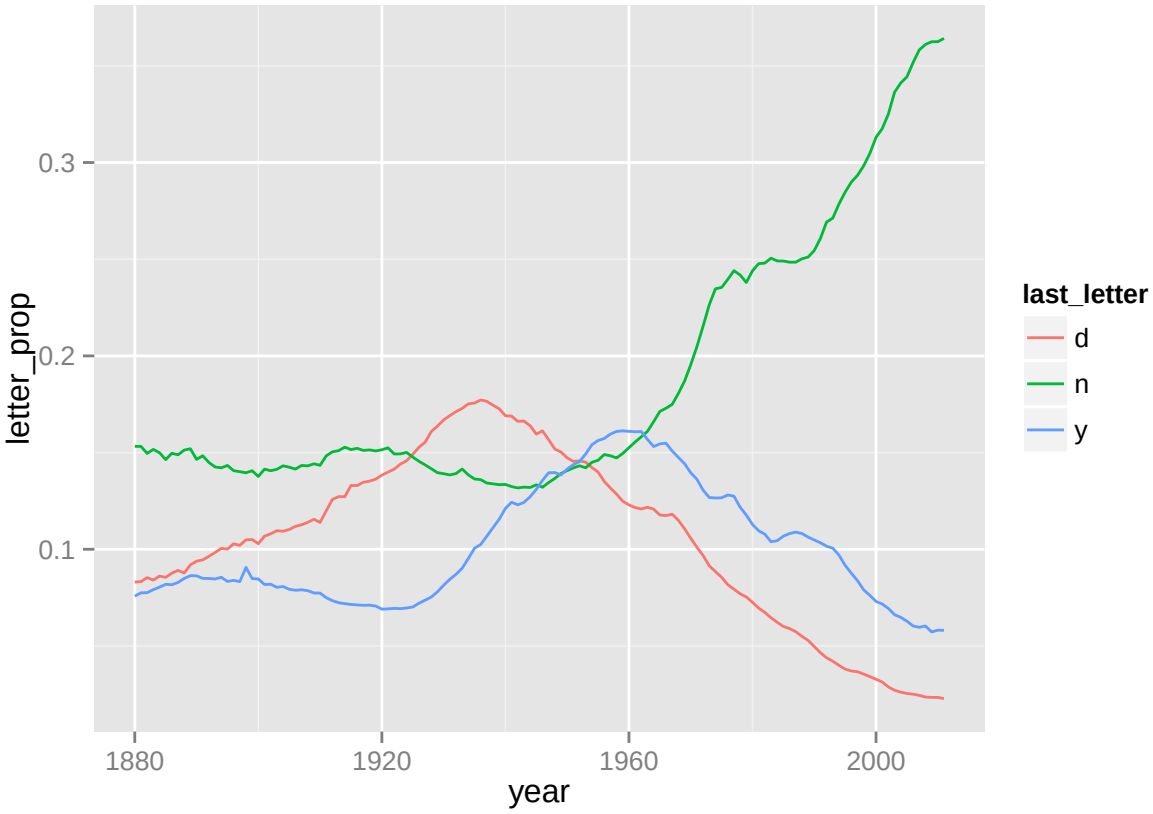
Plot the distribution of names by last letter for three time snapshots

```
letter_count <- babynames %>%
  mutate(last_letter = str_sub(name, start=-1L, end=-1L )) %>%
  group_by(year, sex, last_letter) %>%
  summarise(count=sum(births))

letter_prop <- letter_count %>%
  filter(year %in% c(1910, 1960, 2010)) %>%
  group_by(year, sex) %>%
  mutate(letter_prop=count/sum(count))

letter_prop %>%
  ggplot(aes(x=last_letter, y=letter_prop, fill=as.factor(year))) +
    geom_bar(stat='identity',position=position_dodge()) +
    facet_grid(sex ~ .)
```



Plots the proportion of boy names ending in 'd', 'n', and 'y' over time

```
letter_count %>%
  filter(sex=='M') %>%
  group_by(year) %>%
  mutate(letter_count=sum(count), letter_prop=count/sum(count)) %>%
  filter(last_letter %in% c('d', 'n', 'y')) %>%
  ggplot(aes(x=year, y=letter_prop, color=last_letter)) +
    geom_line()
```


0.3
0.2
0.1
letter_prop
1880
1920
1960
2000
year
last_letter
d
n
y