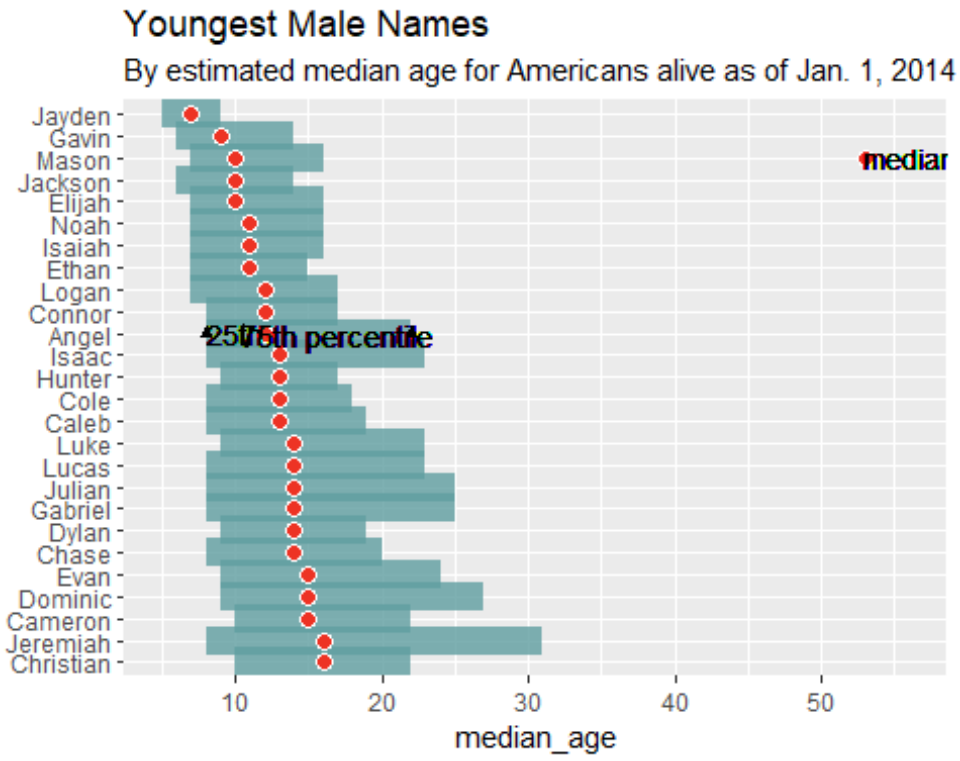
Data Wrangling and ggplot

Yogindra Raghav

October 11, 2018

1. Re-create the data graphic, "Youngest Male Names" (again, your result will be different from the chart below). You can recycle some of the codes above. In particular, the youngest men names are given by the ascending order or median_age. Your chart should be restricted to birth names given to at least 100,000 male Americans since 1900. Use filter() to filter names with at least 100,000 est_num_alive. Can you make the color of the bars Carolina blue?

```
library(Hmisc)
library(mdsr)
library(babynames)

BabynamesDist = make_babynames_dist()

## Warning: package 'bindrcpp' was built under R version 3.4.4

male_1900 = BabynamesDist %>% filter(sex=="M")%>%group_by(name)%>%
mutate(N=n())%>%summarise(est_num_alive=sum(est_alive_today),
q1_age = wtd.quantile(age_today, est_alive_today, probs = 0.25),
median_age =wtd.quantile(age_today,est_alive_today, probs = 0.5),
q3_age = wtd.quantile(age_today, est_alive_today, probs = 0.75) ) %>%
filter(est_num_alive>99999.) %>%
arrange(median_age) %>% head(26)

plot = ggplot(data = male_1900, aes(x = reorder(name, -median_age), y =median
_age))+xlab(NULL)+ggtitle("Youngest Male Names")+labs(subtitle = "By estimate
d median age for Americans alive as of Jan. 1, 2014")

plot + geom_linerange(aes(ymin = q1_age, ymax = q3_age), color = "cadetblue",
size = 5, alpha = 0.8) + geom_point(fill = "#ed3324", colour = "white", size
= 3, shape = 21) +geom_point(aes(y =53, x= 24), fill = "#ed3324", colour = "w
hite", size = 3, shape =21) + geom_text(aes(y = 56, x = 24), label = "median"
, cex = 3.8)+ geom_text(aes(y = 10, x = 16), label = "25th", cex = 3.8) + geo
m_text(aes(y = 17, x = 16), shape = 17, label = "75th percentile")+ geom_poin
t(aes(y = 8, x =16), shape = 17)+ geom_point(y =22, x = 16, shape =17)+ coord
_flip()

## Warning: Ignoring unknown parameters: shape
```



2. Create a new variable with value $|0.5 - boys/total|$, using mutate(). List the top 10 years in which the name "Jackie" was given to M and F babies most equally. (Hint: arrange the variable created above.)

```
BabynamesDist %>% filter(name == "Jackie") %>% group_by(year) %>% mutate(boys
= sum(ifelse(sex == "M", n, 0)), total = sum(n)) %>% mutate(prop_by_sex =abs(
0.5 -boys/total)) %>% arrange(prop_by_sex) %>% summarise(prop_by_sex_summ = p
rop_by_sex[1]) %>% arrange(prop_by_sex_summ) %>% head(10)
```

```
## # A tibble: 10 x 2
##     year prop_by_sex_summ
##    <dbl>            <dbl>
##  1  2006          0.00211
##  2  1997          0.00258
##  3  1925          0.00829
##  4  1999          0.00943
##  5  1956          0.0169
##  6  1927          0.0192
##  7  1926          0.0202
##  8  2003          0.0213
##  9  2002          0.0256
## 10  1955          0.0266
```

3. Which year had the highest number of births?

```
highest_no_of_births = babynames %>% group_by(year) %>% summarise(N = sum(n))
%>% arrange(desc(N)) %>% head(1)

highest_no_of_births
```

```
## # A tibble: 1 x 2
##    year       N
##   <dbl>   <int>
## 1  1957 4200146
```

```
highest_no_of_births$year
```

```
## [1] 1957
```

4. In a single pipeline, compute the earliest and latest year that each name appears.

```
babynames %>% group_by(name) %>% summarise(earliest = min(year), latest = max
(year))
```

```
## # A tibble: 95,025 x 3
##    name      earliest latest
##    <chr>        <dbl>  <dbl>
##  1 Aaban         2007   2015
##  2 Aabha         2011   2015
##  3 Aabid         2003   2003
##  4 Aabriella     2008   2015
##  5 Aada          2015   2015
##  6 Aadam         1987   2015
##  7 Aadan         2003   2015
##  8 Aadarsh       2000   2015
##  9 Aaden         2001   2015
## 10 Aadesh        2005   2011
## # ... with 95,015 more rows
```

5. Among popular names (let's say at least 1% of the births in a given year), which name is the youngest - meaning that its first appearance as a popular name is the most recent?

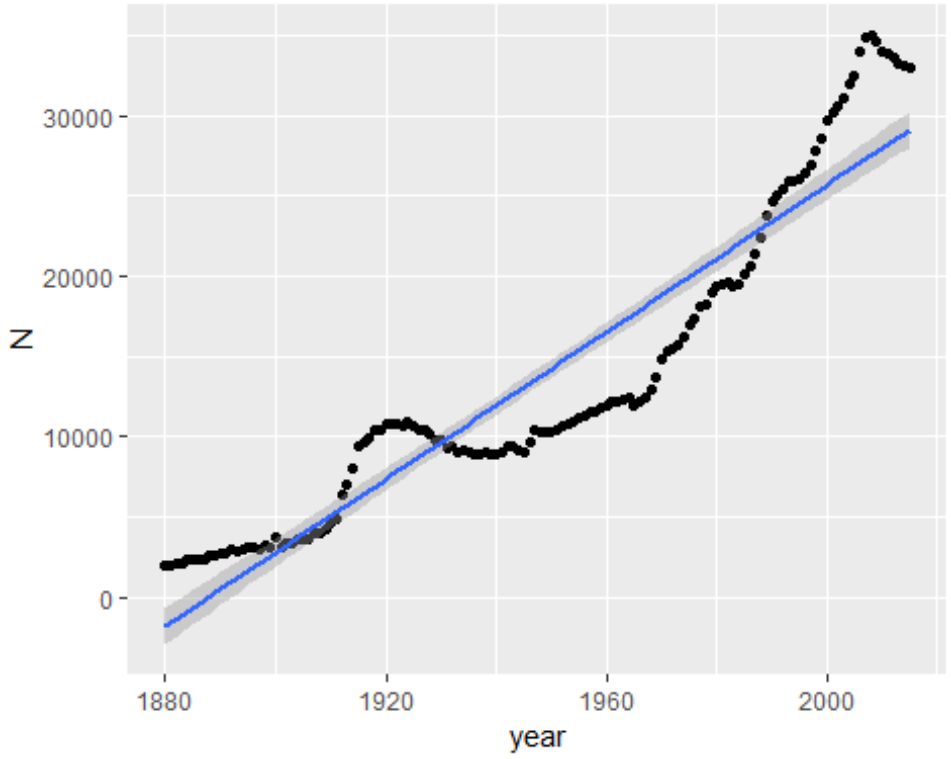
```
babynames %>% filter(prop>=0.01) %>% group_by(name) %>% summarise(minimum = m
in(year)) %>% arrange(desc(minimum)) %>% head(1)
```

```
## # A tibble: 1 x 2
##   name   minimum
##   <chr>    <dbl>
## 1 Olivia    2014
```

6. It seems like there is more diversity of names now than in the past. How have the number of names used changed over time? Has it been the same for boys and girls?

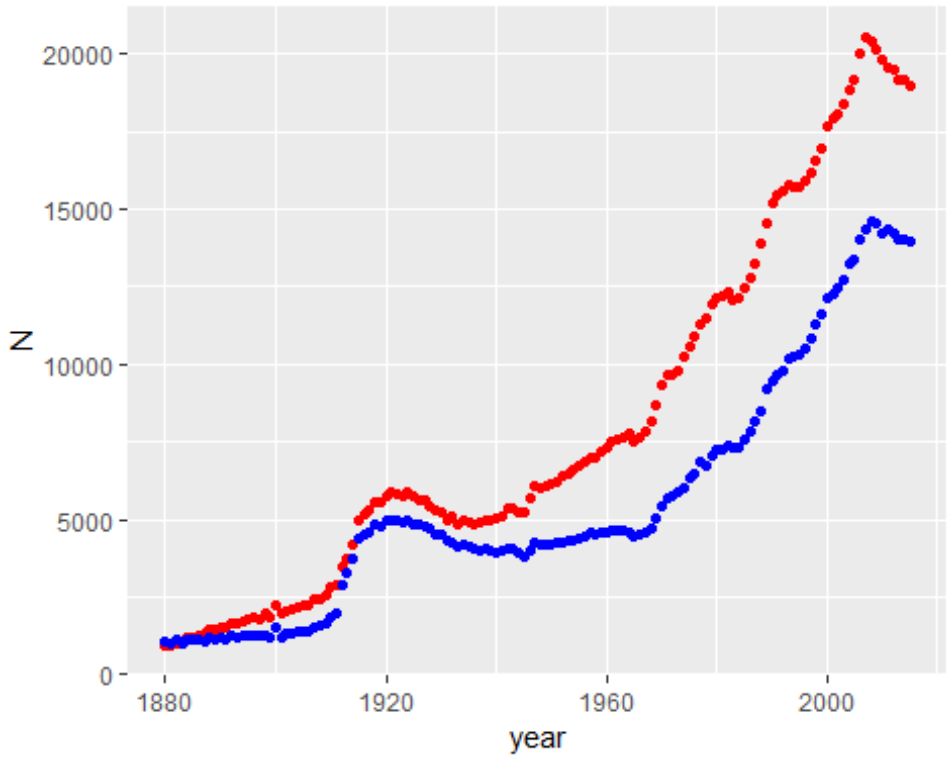
The number of names used over time has increased by quite a bit.

```
name_variation = babynames %>% group_by(year) %>% summarise(N=n())

ggplot(name_variation) + geom_point(mapping= aes(year, N))+ geom_smooth(mappi
ng = aes(year, N), method = "lm")
```



The variation of names used for girls is much larger than for boys especially in the later years (1990's onwards)

```
name_variation_females = babynames %>% filter(sex == "F") %>% group_by(year)
%>% summarise(N=n())
name_variation_males = babynames %>% filter(sex == "M") %>% group_by(year) %>
% summarise(N=n())
ggplot()+ geom_point(data = name_variation_females, mapping = aes(year, N), c
olor = "red")+ geom_point(data = name_variation_males, mapping = aes(year, N)
, color = "blue")
```



7. Find the most popular names of the 1990s.

```
the_90s = babynames %>% filter(year<2000&year>1989)
the_90s %>% group_by(name) %>% summarise(prop_in_1990s = sum(n)/sum(the_90s$n
)) %>% arrange(desc(prop_in_1990s))
```

```
## # A tibble: 45,921 x 2
##    name        prop_in_1990s
##    <chr>               <dbl>
##  1 Michael           0.0124
##  2 Christopher       0.00964
##  3 Matthew           0.00940
##  4 Joshua            0.00881
##  5 Jessica           0.00811
##  6 Ashley            0.00809
##  7 Jacob             0.00798
##  8 Nicholas          0.00736
##  9 Andrew            0.00730
## 10 Daniel            0.00729
## # ... with 45,911 more rows
```