HW1

배현주

9/23/2019

1. Age distribution of American boys named Joseph

Load libraries

```
library(babynames)
library(mdsr)
library(dplyr)
library(Hmisc)
library(scales)
library(ggplot2)
```

Load dataset

```
baby_data <- make_babynames_dist()
str(baby_data)
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':   1639722 obs. of  9 variables:
 $ year           : num  1900 1900 1900 1900 1900 1900 1900 1900 1900 1900 ...
 $ sex            : chr  "F" "F" "F" "F" ...
 $ name           : chr  "Mary" "Helen" "Anna" "Margaret" ...
 $ n              : int  16706 6343 6114 5304 4765 4096 3920 3896 3856 3414 ...
 $ prop           : num  0.0526 0.02 0.0192 0.0167 0.015 ...
 $ alive_prob     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ count_thousands: num  16.71 6.34 6.11 5.3 4.76 ...
 $ age_today      : num  114 114 114 114 114 114 114 114 114 114 ...
 $ est_alive_today: num  0 0 0 0 0 0 0 0 0 0 ...
```

Plot

1. Filter the data and find the median year and number of Joseph alive.

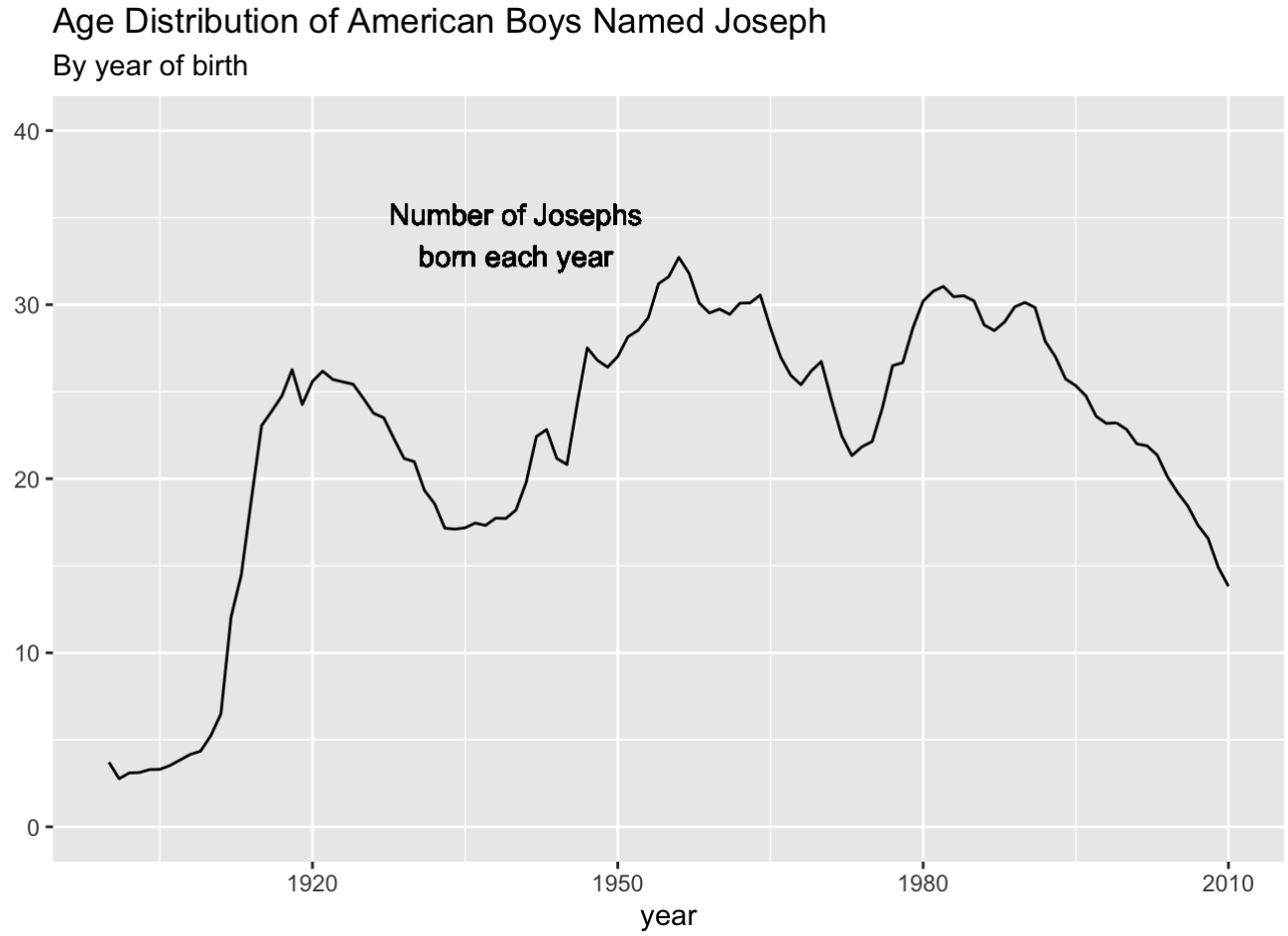
```
joseph <- filter(baby_data, name == "Joseph", sex == "M")
alive_n.median <- with(joseph, wtd.quantile(year, est_alive_today, prob=0.5))
alive_n.median
```

```
 50%
1975
```

> The median number of living still Joseph is 20847.19 and the birth of year is 1975.

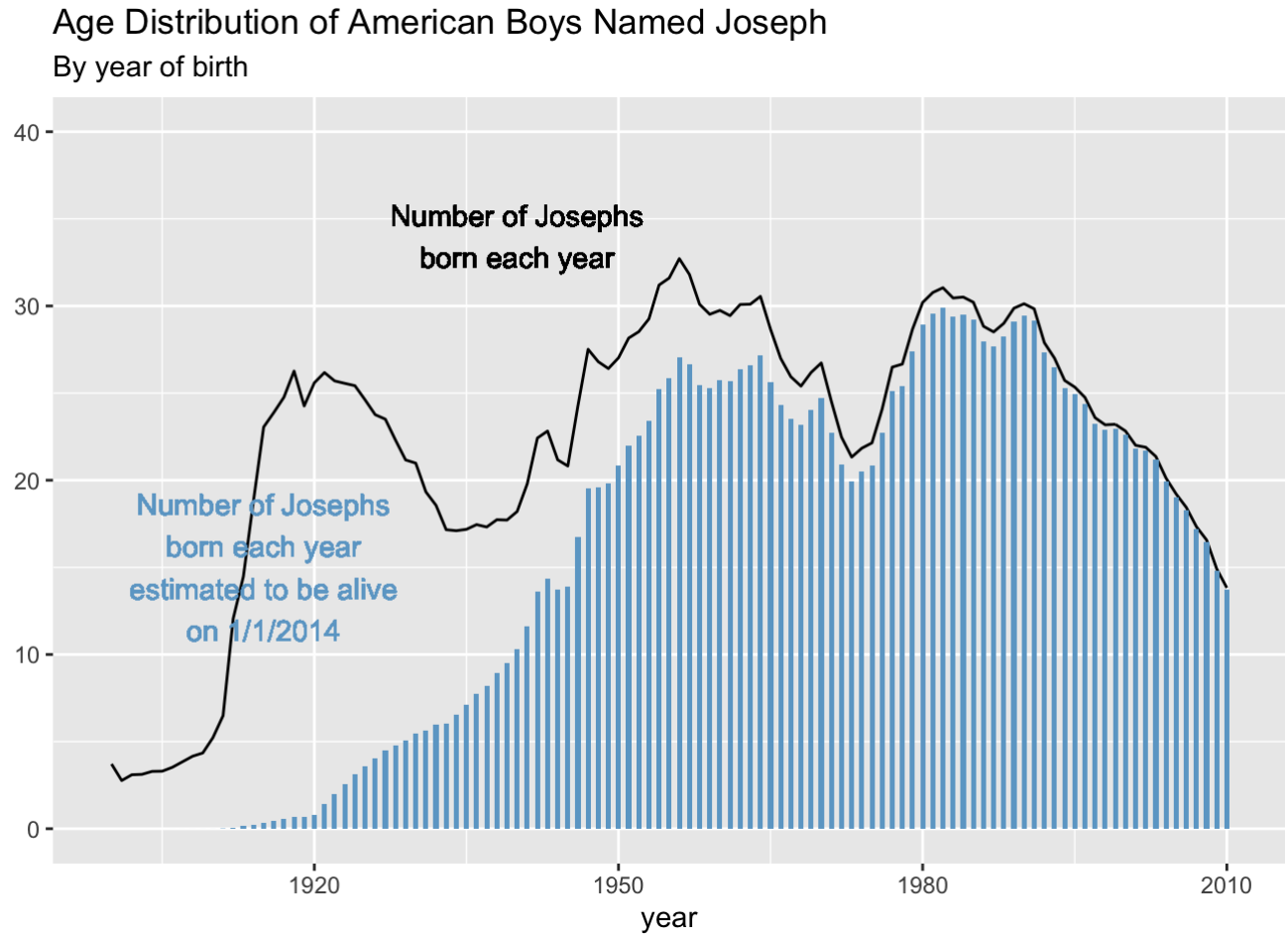
2. Plot the number of Josephs born each year

```
joseph_num <- ggplot(data=joseph) +
  geom_line(mapping = aes(x=year, y=n/1000)) +
  labs(title = "Age Distribution of American Boys Named Joseph",
       subtitle = "By year of birth") +
  ylab("") +
  coord_cartesian(ylim=c(0, 40), expand=TRUE) +
  geom_text(x = 1940, y = 34, label = "Number of Josephs\nborn each year")
joseph_num
```



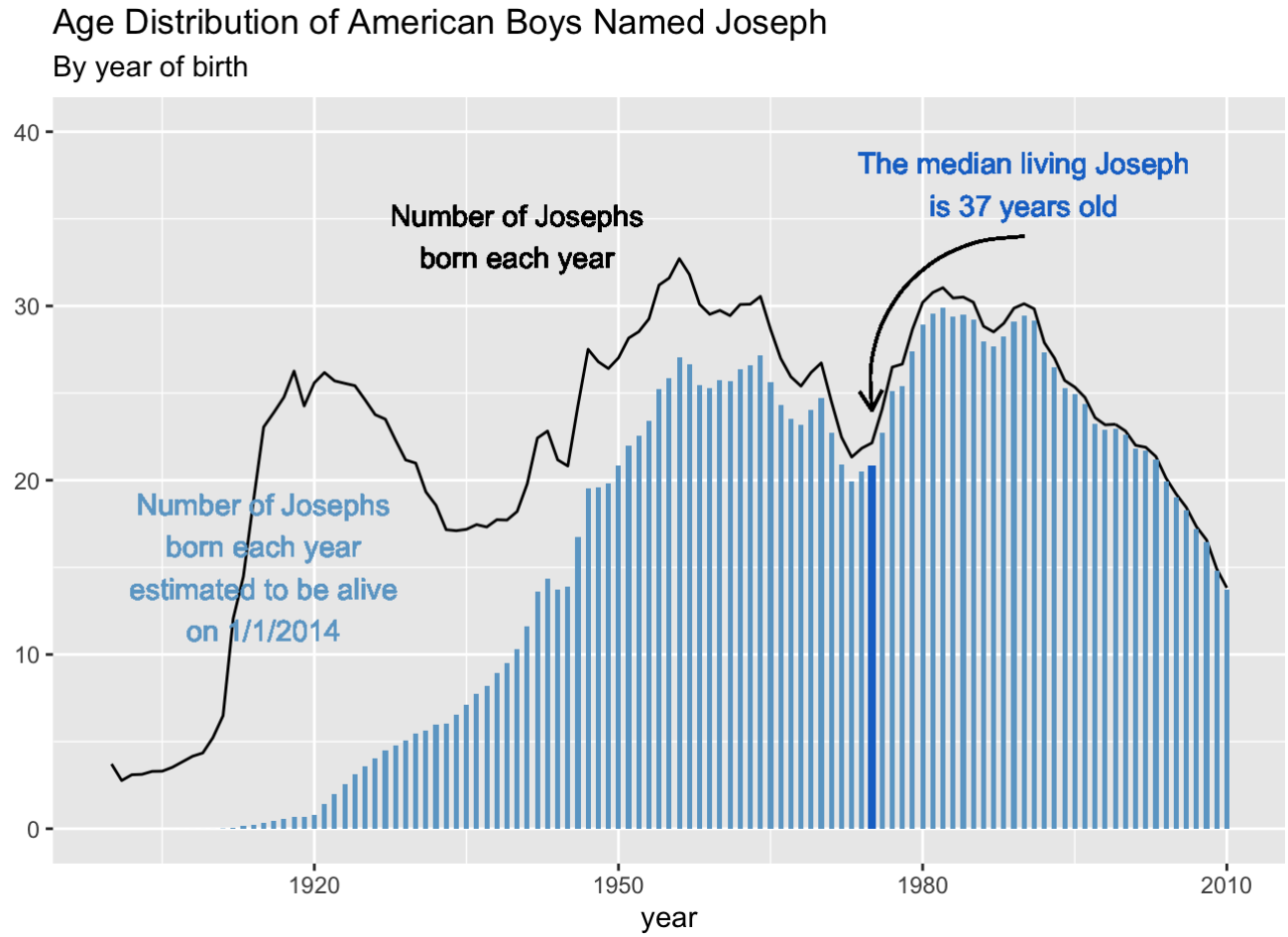
3. Plot the number of Joseph alive

```
joseph_num <- joseph_num +
  geom_bar(mapping = aes(x=year, y=est_alive_today/1000),
           stat="identity", fill="skyblue3", width=0.5) +
  geom_text(x = 1915, y = 15,
            label = "Number of Josephs\nborn each year\nestimated to be alive\non 1/
1/2014",
            colour = "skyblue3")
joseph_num
```



4. Highlight the median number of Joseph alive

```
joseph_num <- joseph_num +
  geom_bar(mapping = aes(x=alive_n.median, y=ifelse(year == alive_n.median, est_alive
_today/1000, 0)),
           stat="identity", fill="dodgerblue3", width=0.7) +
  geom_text(x = 1990, y = 37,
            label = "The median living Joseph\nis 37 years old",
            colour = "dodgerblue3") +
  geom_curve(x = 1990, xend = 1975, y = 34, yend = 24,
             arrow = arrow(length = unit(0.3,"cm")), curvature = 0.5) + ylim(0, 42)
joseph_num
```



2. Median ages for males with 25 most common names

Load datasets

```
rm(list=ls())
baby_data <- make_babynames_dist()
colnames(baby_data)
```

```
[1] "year"            "sex"             "name"            "n"
[5] "prop"            "alive_prob"      "count_thousands" "age_today"
[9] "est_alive_today"
```

1. Filter the 25 most common names

```
male <- baby_data %>% filter(sex == "M")
top25_male <- male %>%
  group_by(name) %>%
  summarise(N=sum(n)) %>%
  arrange(desc(N)) %>%
  head(25)

male <- male %>%
  filter(name %in% top25_male$name)
```

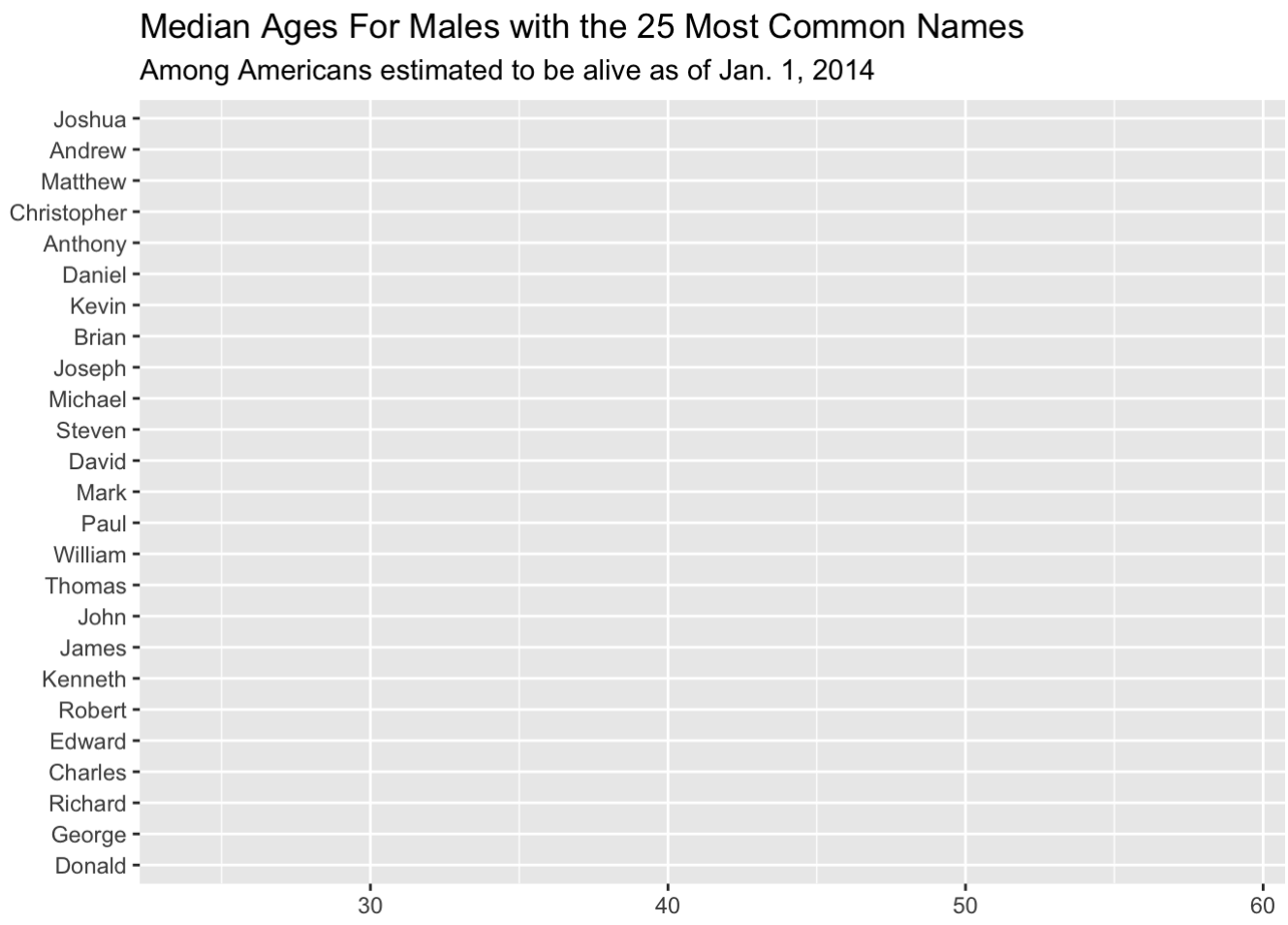
2. Make variables of Q1, median, Q3 age

```
male_quant <- male %>%
  group_by(name) %>%
  summarise(q1_age = wtd.quantile(age_today, est_alive_today, probs = 0.25),
            med_age = wtd.quantile(age_today, est_alive_today, probs = 0.5),
            q3_age = wtd.quantile(age_today, est_alive_today, probs = 0.75))
```

Plot

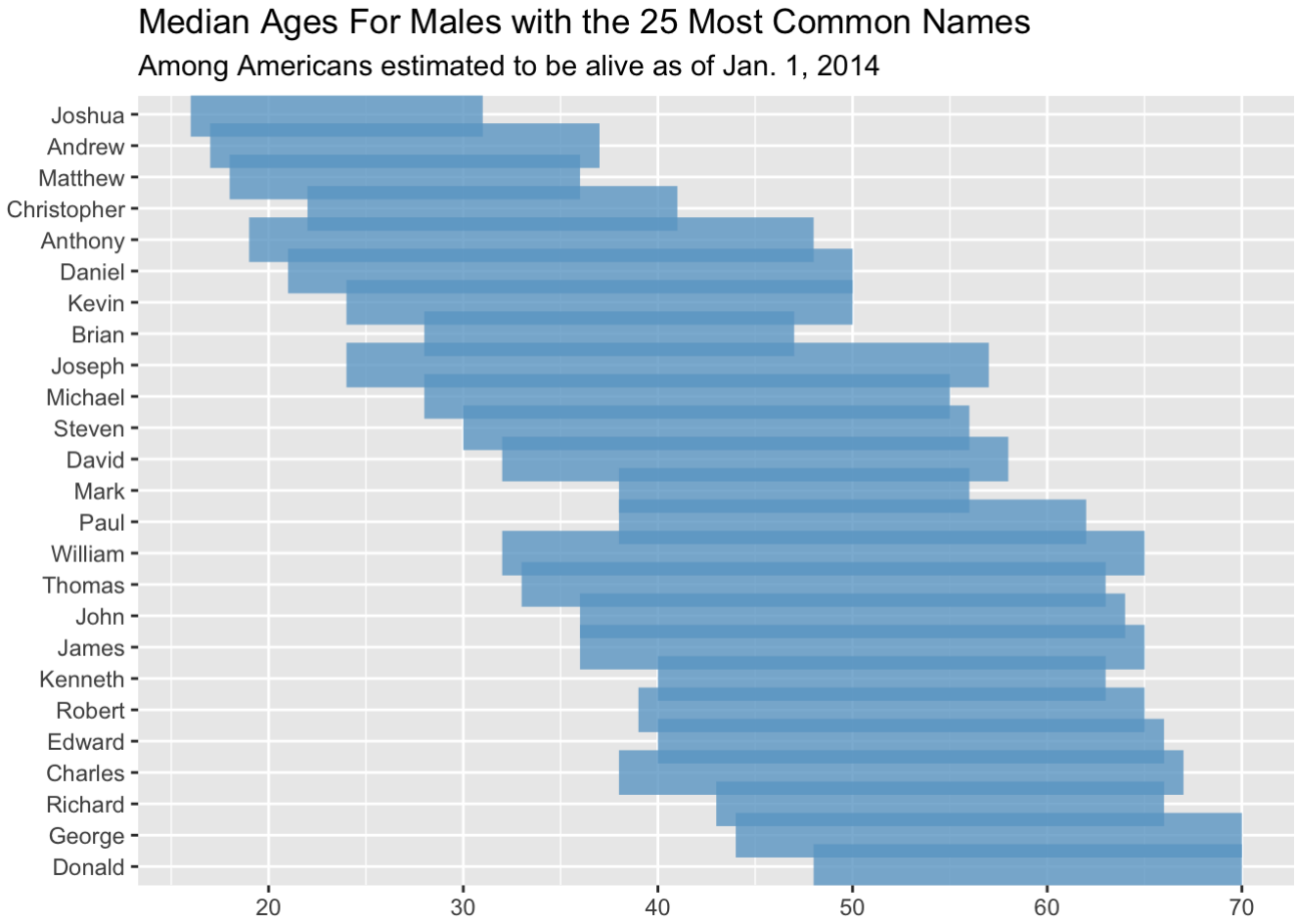
1. Make the background of the plot and flip the coordinate

```
top25_male_plot <- ggplot(male_quant, aes(x = reorder(name, - med_age), y = med_age))
+
  xlab(NULL) +
  ylab(NULL) +
  labs(title = "Median Ages For Males with the 25 Most Common Names",
          subtitle = "Among Americans estimated to be alive as of Jan. 1, 2014") +
    coord_flip()
top25_male_plot
```



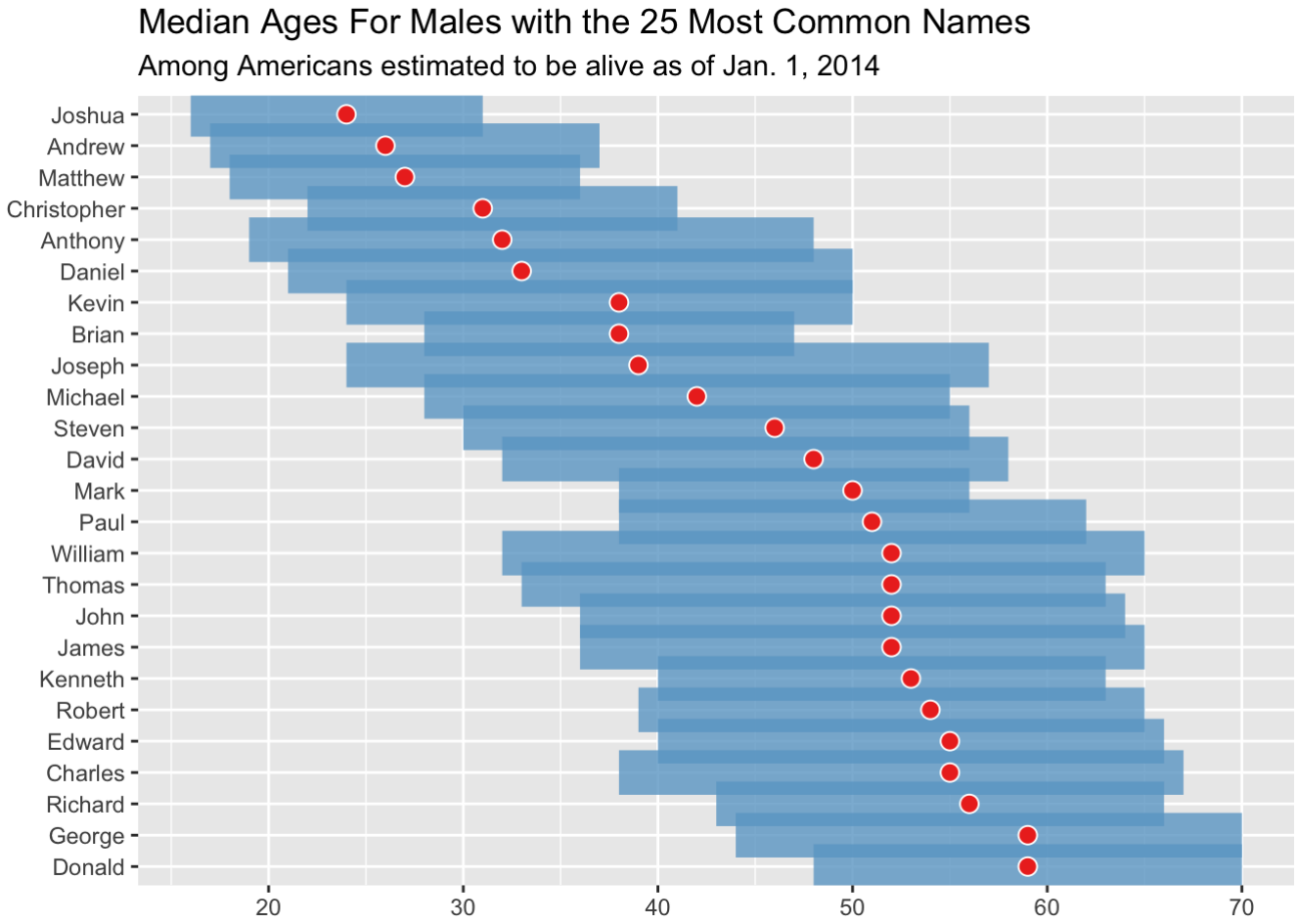
2. Add Q1 - Q3 boxplot to the plot

```
top25_male_plot <- top25_male_plot +
  geom_linerange(aes(ymin = q1_age, ymax = q3_age),
                 color = "skyblue3", size = 8, alpha = 0.8)
top25_male_plot
```



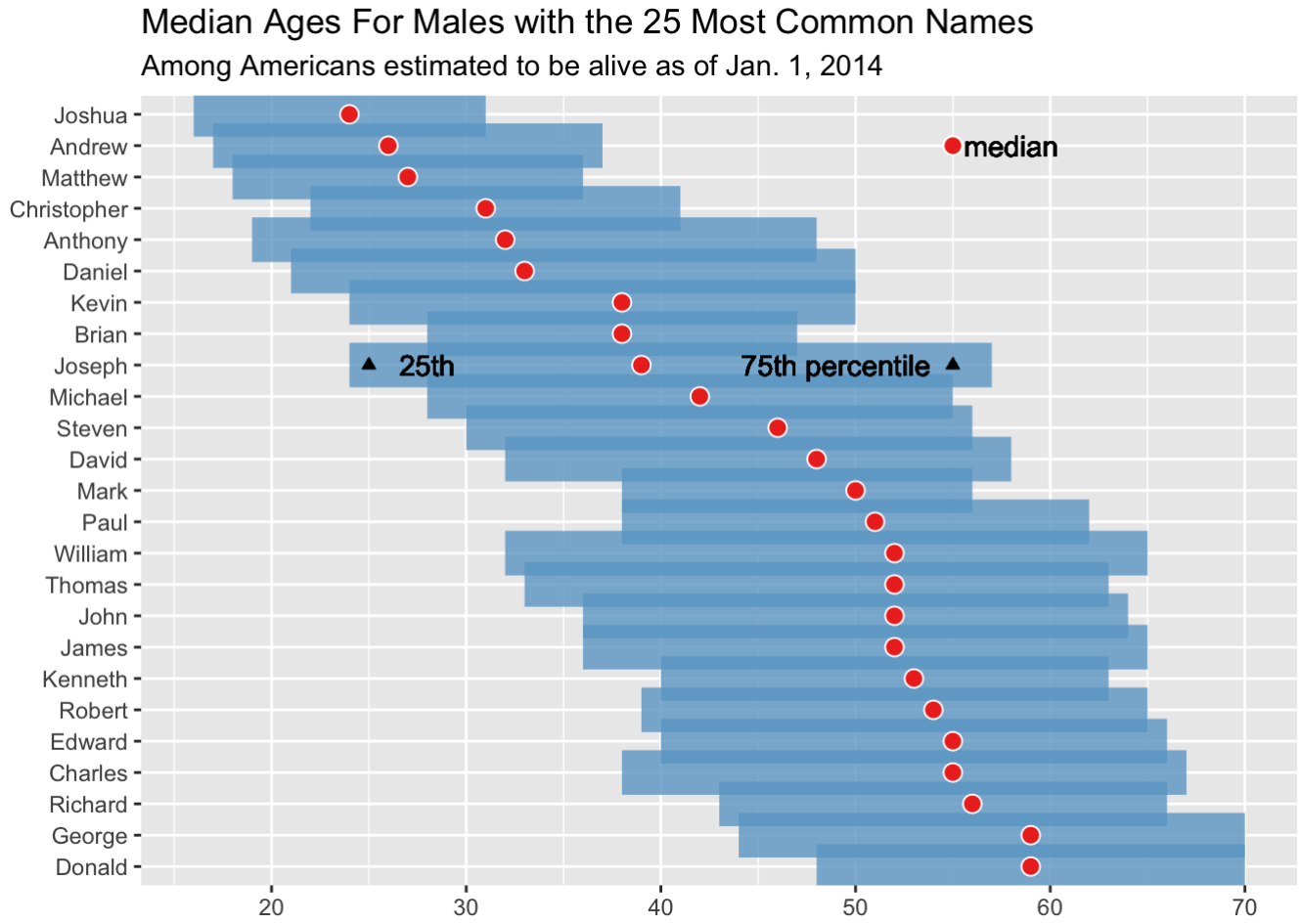
3. Add median point

```
top25_male_plot <- top25_male_plot +
  geom_point(fill = "#ed3324", color = "white", size = 3, shape = 21)
top25_male_plot
```



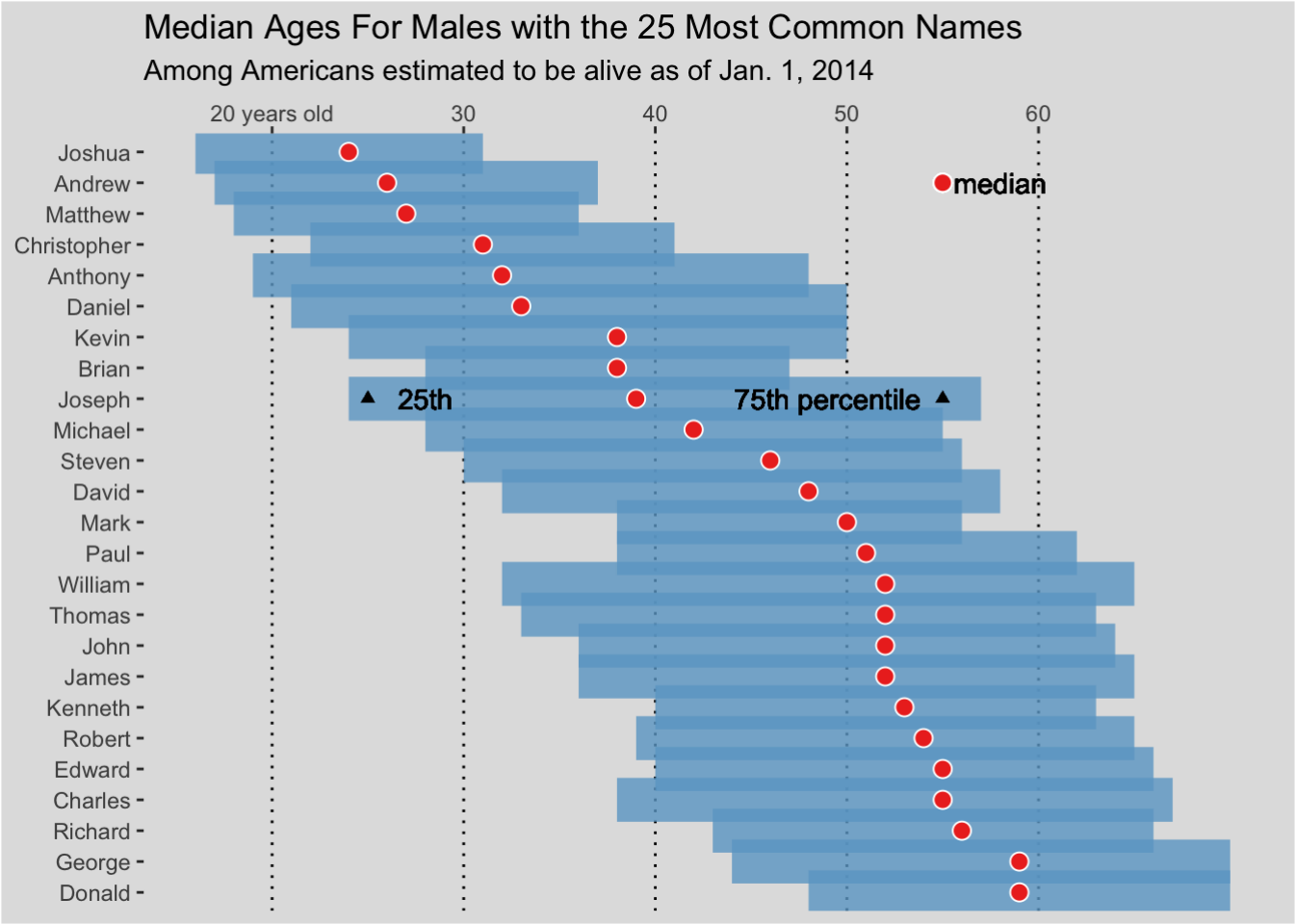
4. Add text

```
top25_male_plot <- top25_male_plot +
  geom_point(aes(y = 55, x = 24), fill = "#ed3324", colour = "white",
             size = 3, shape = 21) +
  geom_text(aes(y = 58, x = 24, label = "median")) +
  geom_text(aes(y = 28, x = 17, label = "25th")) +
  geom_text(aes(y = 49, x = 17, label = "75th percentile")) +
  geom_point(aes(y = 25, x = 17), shape = 17) +
  geom_point(aes(y = 55, x = 17), shape = 17)
top25_male_plot
```



5. Change the background and y lab

```
top25_male_plot <- top25_male_plot +
  theme(plot.background = element_rect(fill = "grey88"),
        panel.background = element_rect(fill = "grey88"),
        panel.grid.major.x = element_line(colour = "black", linetype = "dotted"),
        panel.grid.major.y = element_blank()) +
  scale_y_continuous(minor_breaks = seq(20, 60, 10),
                     breaks = seq(20, 60, 10),
                     labels = c('20 years old', '30', '40', '50', '60'),
                     position = 'right')
top25_male_plot
```


Median Ages For Males with the 25 Most Common Names
Among Americans estimated to be alive as of Jan. 1, 2014
20 years old
30
40
50
60
median
25th
75th percentile
Joshua
Andrew
Matthew
Christopher
Anthony
Daniel
Kevin
Brian
Joseph
Michael
Steven
David
Mark
Paul
William
Thomas
John
James
Kenneth
Robert
Edward
Charles
Richard
George
Donald