Practice ggplot2

Math 241, Week 1

```
# it's good practice to check that all the packages required are loaded and installed
libs <- c('tidyverse','dplyr','ggplot2','knitr','viridis','mdsr', 'macleish','babynames')
for(l in libs){
  if(!require(l,character.only = TRUE, quietly = TRUE)){
    message( sprintf('Did not have the required package << %s >> installed. Downloading now ... ',l))
    install.packages(l)
  }
  library(l, character.only = TRUE, quietly = TRUE)
}
```

Goals of this in-class activity:

- Practice creating and refining graphs with `ggplot2`.
- Consider the strengths and weaknesses of various `geom`s and `aes`thetics for telling a data story.

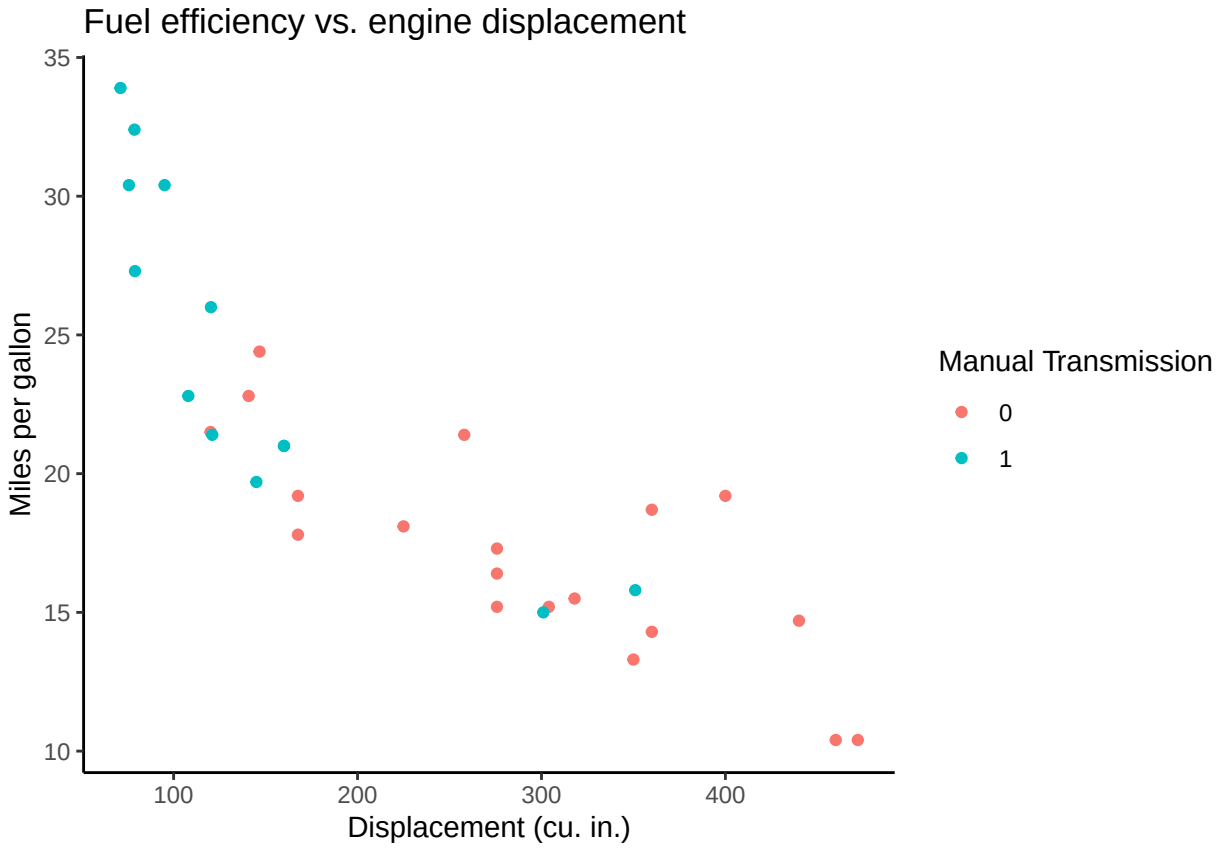
Notes:

- When creating your graphs, consider context (i.e. axis labels, title, . . .)!
- If I provide partially completed code, I will put `eval = FALSE` in the chunk. Make sure to change that to `eval = TRUE` once you have completed the code in the chunk.
- Be prepared to ask for help from me, Tory, and your classmates! We scratch the surface of `ggplot2` in class. But I encourage you to really dig in and make your graphs your own (i.e. don't rely on defaults).

Problem 1 (Easy):

Consider the following data graphic.

```
ggplot(mtcars, aes(x = disp, y = mpg, color = factor(am))) +
geom_point() +
  labs(x = "Displacement (cu. in.)",
       y = "Miles per gallon",
       color = "Manual Transmission",
       title = "Fuel efficiency vs. engine displacement") +
  theme_classic()
```



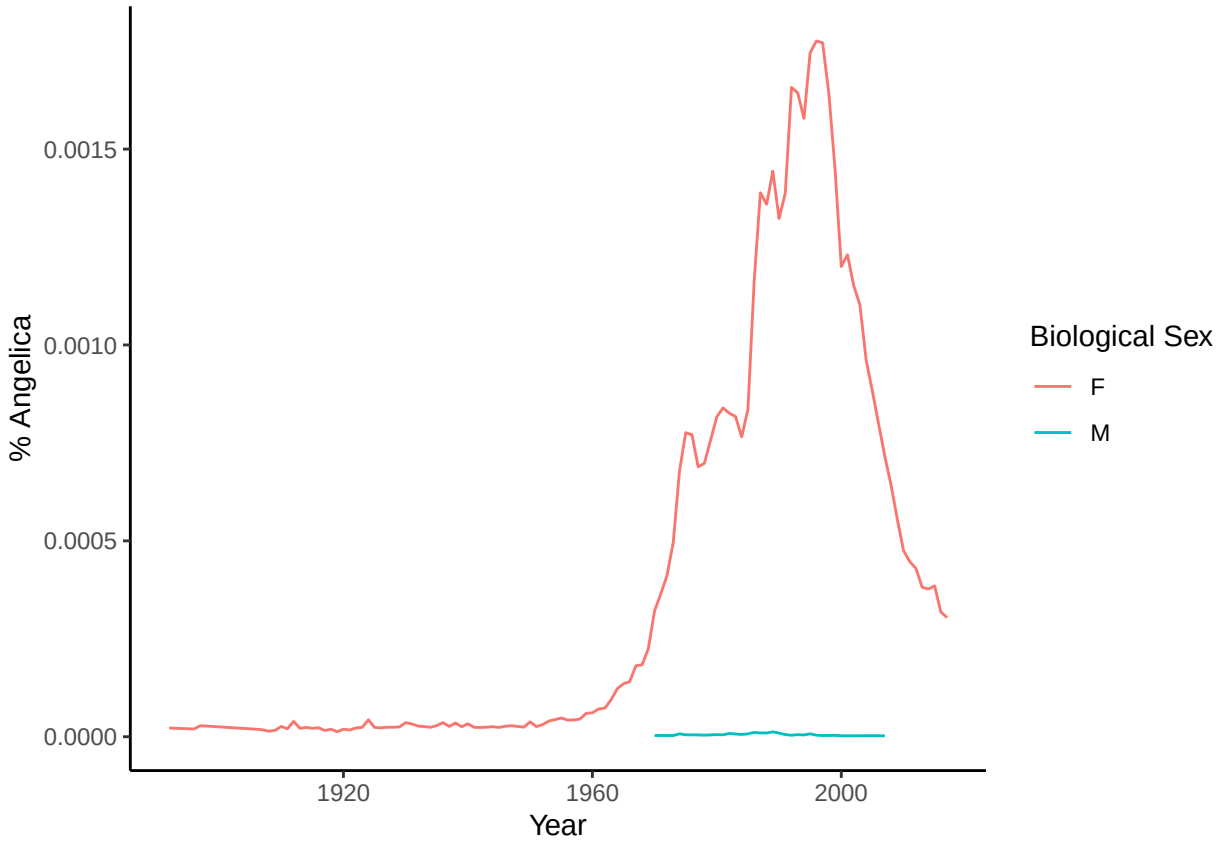
The am variable takes the value 0 if the car has automatic transmission and 1 if the car has manual transmission. How could you differentiate the cars in the graphic based on their transmission type?

Problem 2 (Easy):

Angelica Schuyler Church (1756–1814) was the daughter of New York Governer Philip Schuyler and sister of Elizabeth Schuyler Hamilton. Angelica, New York was named after her.

Using the `babynames` package generate a plot of the reported proportion of babies born with the name Angelica over time and interpret the figure.

```
data(babynames) # this will explicitly ask R to load the babynames dataset to your environment
angelica <- filter(babynames, name == "Angelica")
ggplot(data = angelica, aes(x = year, y = prop, color = sex)) +
geom_line() +
  labs(x = 'Year', y = '% Angelica', color = 'Biological Sex') +
  theme_classic()
```



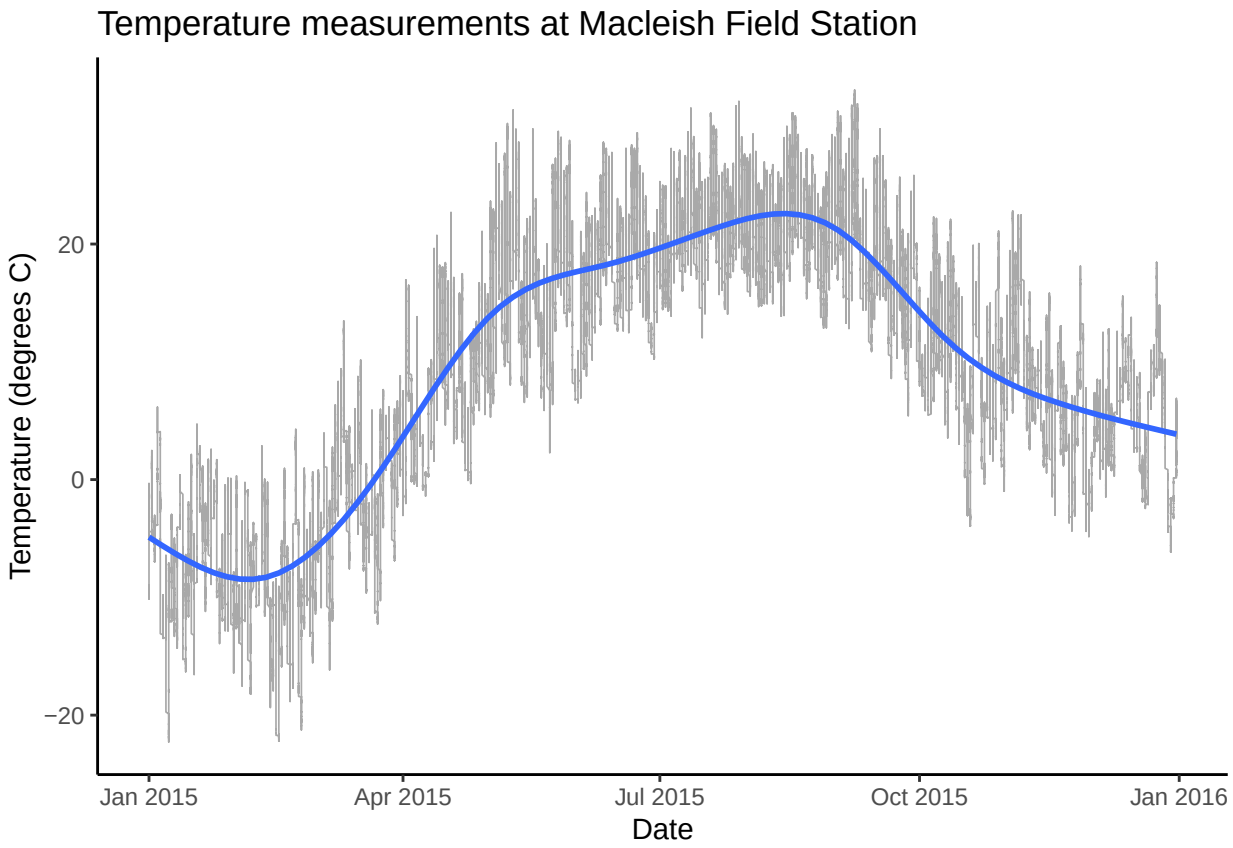
We see a huge increase in the proportion of girls named Angelica.

Problem 3 (Medium):

The `macleish` package contains weather data collected every 10 minutes in 2015 from two weather stations in Whately, MA.

```
whately_2015 <- whately_2015 %>% mutate(Date = as.Date(when))

ggplot(data = whately_2015, aes(x = Date, y = temperature)) +
  geom_line(size = 0.3, color = "darkgray") +
  labs(y = "Temperature (degrees C)", title = "Temperature measurements at Macleish Field Station") +
  geom_smooth() +
  scale_x_date() +
  theme_classic()
```

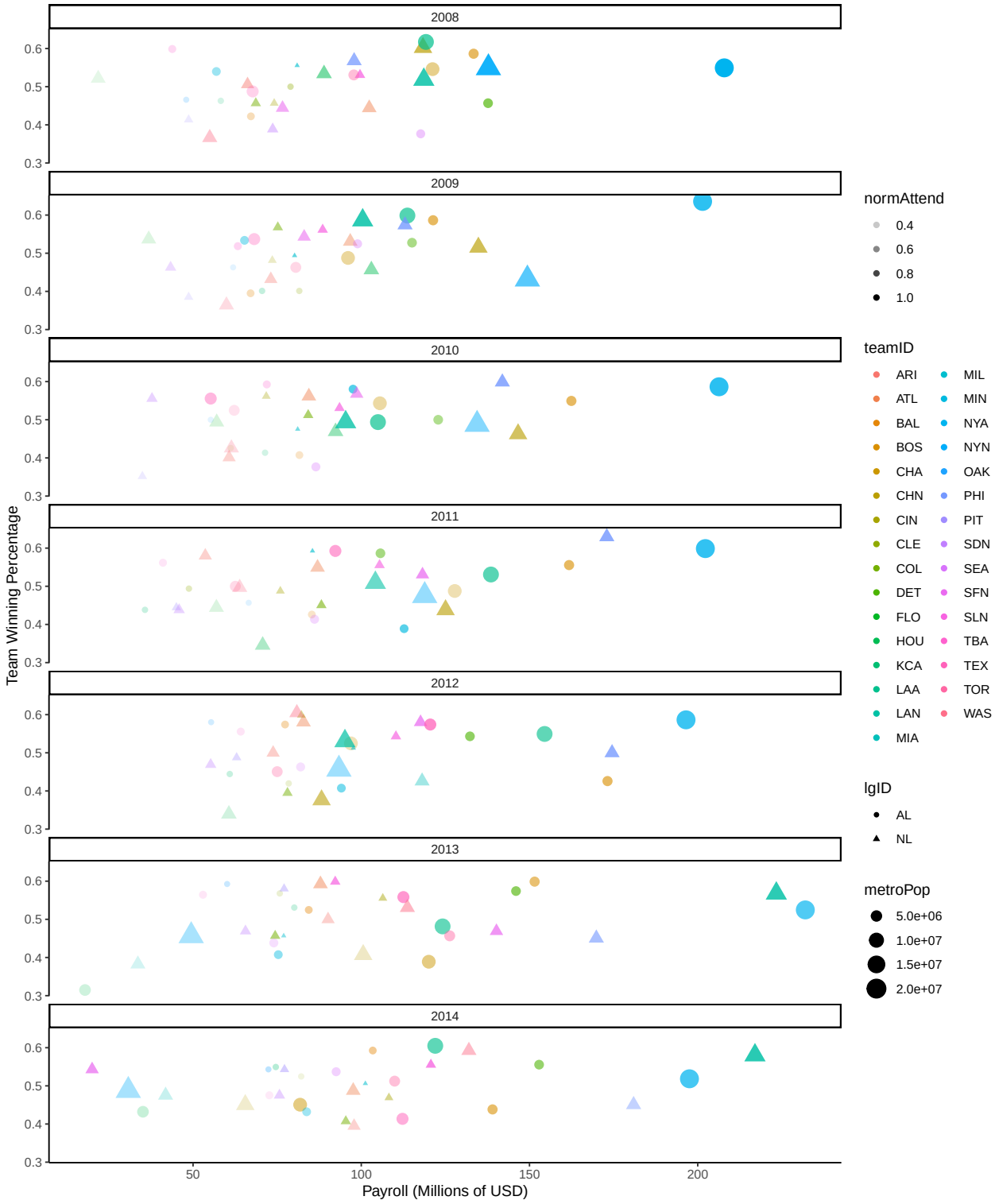


Using ggplot2, create a data graphic that displays the average temperature over each 10-minute interval (temperature) as a function of time (when).

Problem 4 (Medium):

The data set `MLB_teams` in the `mdsr` package contains information about Major League Baseball teams from 2008–2014. There are several quantitative and a few categorical variables present. See how many variables you can illustrate on a single plot in R. The current record is 7. (Note: **This is not good graphical practice**—it is merely an exercise to help you understand how to use visual cues and aesthetics!)

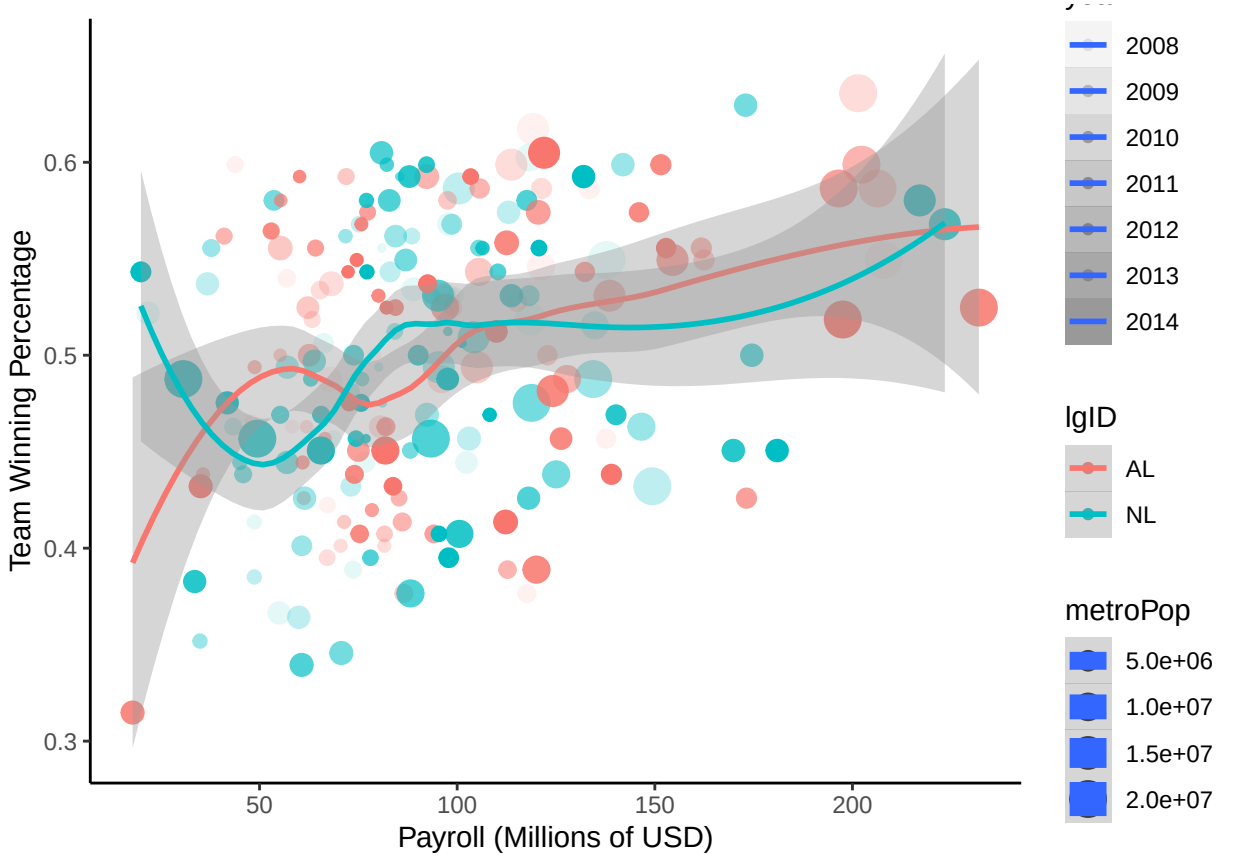
```
ggplot(MLB_teams) +
  geom_point(aes( x = payroll / 1000000, y = WPct,
                  color = teamID, size = metroPop, shape = lgID, alpha = normAttend)) +
  facet_wrap(~yearID, ncol = 1) +
  xlab("Payroll (Millions of USD)") +
  ylab("Team Winning Percentage") +
  theme_classic()
```



Problem 5 (Medium):

Use the `MLB_teams` data in the `mdsr` package again to create an informative data graphic that illustrates the relationship between winning percentage and payroll in context.

```
ggplot(data = MLB_teams, aes(
  x = payroll / 1000000, y = WPct,
  alpha = yearID,
  color = lgID, size = metroPop)) +
  geom_point() +
  geom_smooth() +
  xlab("Payroll (Millions of USD)") +
  ylab("Team Winning Percentage") +
  theme_classic()
```



Problem 6 (Hard):

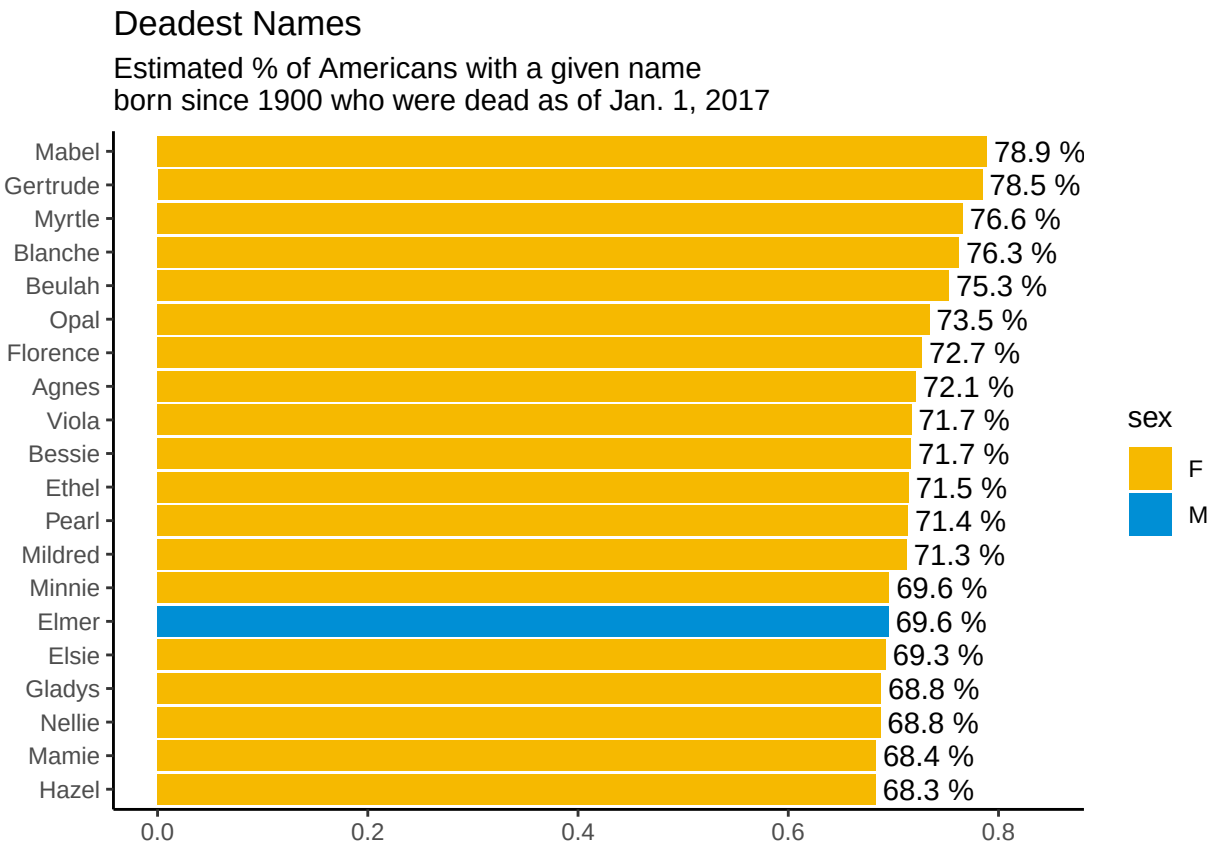
Use the function `make_babynames_dist()` in the `mdsr` package to recreate the "Deadest Names" graphic from FiveThirtyEight.

```
babynames_dist <- make_babynames_dist()
glimpse(babynames_dist)
```

```
## Rows: 1,639,722
## Columns: 9
## $ year            <dbl> 1900, 1900, 1900, 1900, 1900, 1900, 1900, 1900, 1900, ~
## $ sex             <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F",~
## $ name            <chr> "Mary", "Helen", "Anna", "Margaret", "Ruth", "Elizabet~
## $ n               <int> 16706, 6343, 6114, 5304, 4765, 4096, 3920, 3896, 3856,~
```

```
## $ prop            <dbl> 0.05257559, 0.01996211, 0.01924142, 0.01669226, 0.0149~
## $ alive_prob      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ count_thousands <dbl> 16.706, 6.343, 6.114, 5.304, 4.765, 4.096, 3.920, 3.89~
## $ age_today       <dbl> 114, 114, 114, 114, 114, 114, 114, 114, 114, 114, 114,~
## $ est_alive_today <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

```
babynames_dist %>%
  filter(year >= 1900) %>%
  group_by(name, sex) %>%
  summarize(
  N = n(),
  total_est_alive_today = sum(est_alive_today),
  total = sum(n)) %>%
mutate(pct_dead = 1 - (total_est_alive_today / total)) %>%
  filter(total > 50000) %>%
  arrange(desc(pct_dead)) %>%
  head(20) %>%
  ggplot(aes(x = reorder(name, pct_dead), y = pct_dead, fill = sex)) +
  geom_bar(stat = "identity") +
  geom_text(
  aes(
  y = pct_dead + 0.05,
  label = paste(round(pct_dead * 100, 1), "%")
  )
  ) +
  coord_flip() +
  ggtitle("Deadest Names",
  subtitle =
  "Estimated % of Americans with a given name\nborn since 1900 who were dead as of Jan. 1, 2017"
  ) +
  scale_x_discrete(NULL) +
  scale_y_continuous(NULL) +
  scale_fill_manual(values = c("#f6b900", "#008fd5")) +
  theme_classic()
```


Deadest Names
Estimated % of Americans with a given name
born since 1900 who were dead as of Jan. 1, 2017
Mabel
Gertrude
Myrtle
Blanche
Beulah
Opal
Florence
Agnes
Viola
Bessie
Ethel
Pearl
Mildred
Minnie
Elmer
Elsie
Gladys
Nellie
Mamie
Hazel
78.9 %
78.5 %
76.6 %
76.3 %
75.3 %
73.5 %
72.7 %
72.1 %
71.7 %
71.7 %
71.5 %
71.4 %
71.3 %
69.6 %
69.6 %
69.3 %
68.8 %
68.8 %
68.4 %
68.3 %
0.0
0.2
0.4
0.6
0.8
sex
F
M