Stat. 651 Homework 1

Lydia Gibson

October 19, 2022

```
library(pacman)
p_load(tidyverse, macleish, nasaweather, palmerpenguins, mdsr)
```

Problem 4 (Medium):

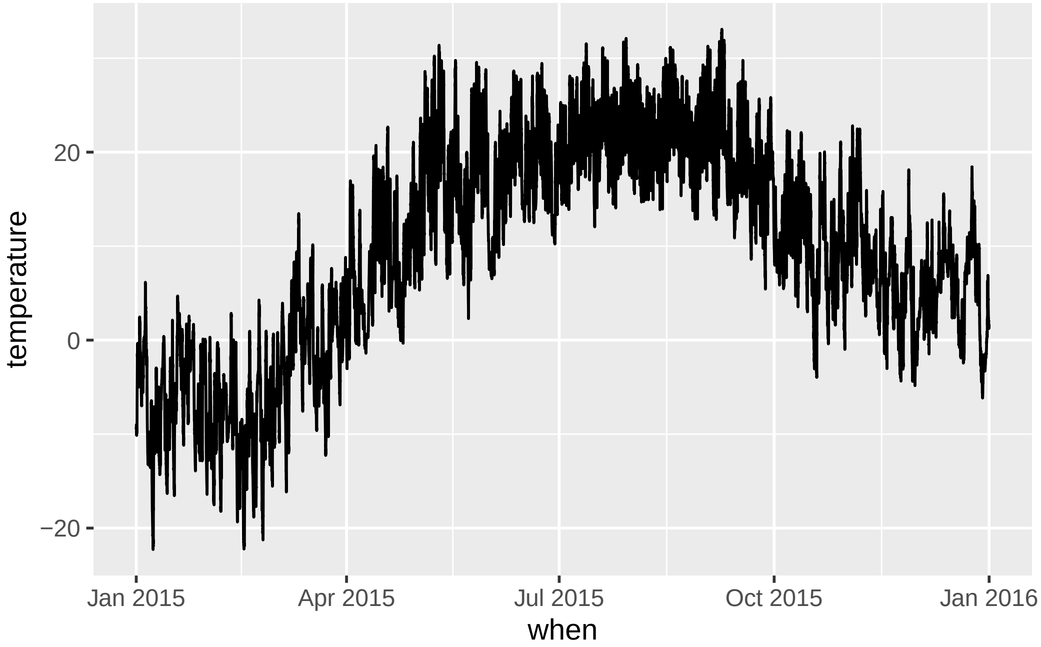
The `macleish` package contains weather data collected every 10 minutes in 2015 from two weather stations in Whately, MA.

```
head(whately_2015)
```

```
# A tibble: 6 x 8
  when                temperat~1 wind_~2 wind_~3 rel_h~4 press~5 solar~6 rainf~7
  <dttm>                   <dbl>   <dbl>   <dbl>   <dbl>   <int>   <dbl>   <dbl>
1 2015-01-01 00:00:00      -9.32    1.40    225.    54.6     985       0       0
2 2015-01-01 00:10:00      -9.46    1.51    248.    55.4     985       0       0
3 2015-01-01 00:20:00      -9.44    1.62    258.    56.2     985       0       0
4 2015-01-01 00:30:00      -9.3     1.14    244.    56.4     985       0       0
5 2015-01-01 00:40:00      -9.32    1.22    238.    56.9     984       0       0
6 2015-01-01 00:50:00      -9.34    1.09    242.    57.2     984       0       0
# ... with abbreviated variable names 1: temperature, 2: wind_speed,
#   3: wind_dir, 4: rel_humidity, 5: pressure, 6: solar_radiation, 7: rainfall
```

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

```
ggplot(data = whately_2015, mapping = aes(x = when, y = temperature)) +
  geom_line()
```



Problem 8 (Medium):

Using data from the nasaweather package, use the geom_path function to plot the path of each tropical storm in the storms data table. Use color to distinguish the storms from one another, and use faceting to plot each year in its own panel.

```
head(storms)
```

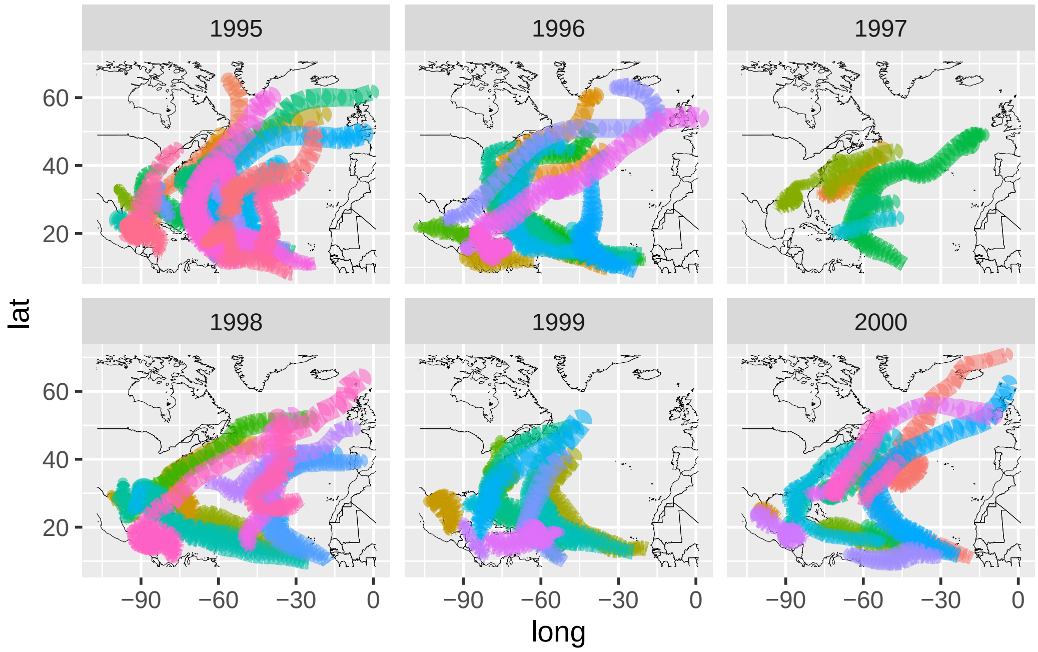
```
# A tibble: 6 x 11
  name     year month   day  hour   lat  long pressure  wind type          seasday
  <chr>   <int> <int> <int> <int> <dbl> <dbl>    <int> <int> <chr>           <int>
1 Allison  1995     6     3     0  17.4 -84.3     1005    30 Tropical D~         3
2 Allison  1995     6     3     6  18.3 -84.9     1004    30 Tropical D~         3
3 Allison  1995     6     3    12  19.3 -85.7     1003    35 Tropical S~         3
```

```
4 Allison  1995     6     3    18  20.6 -85.8     1001    40 Tropical S~        3
5 Allison  1995     6     4     0  22   -86        997    50 Tropical S~        4
6 Allison  1995     6     4     6  23.3 -86.3      995    60 Tropical S~        4
```

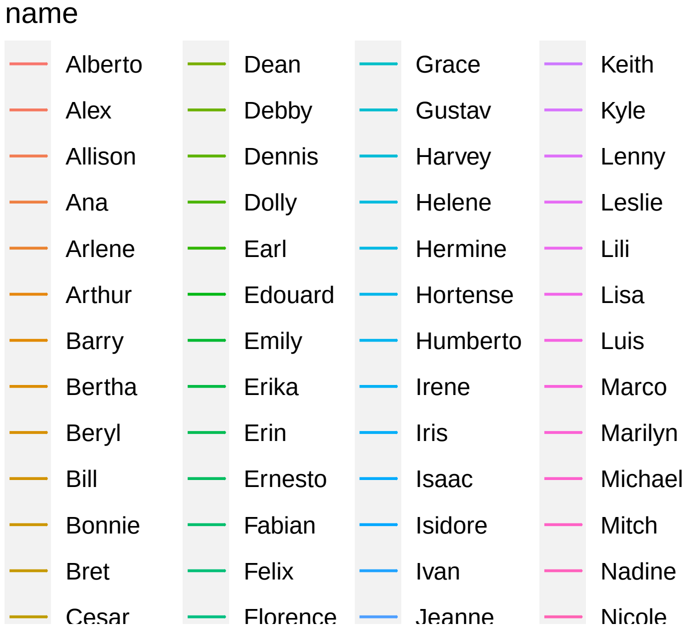
```
bbox <- storms %>%
  select(lat, long) %>%
  map_df(range)          # using the purrr R package

bbox
```

```
# A tibble: 2 x 2
    lat  long
  <dbl> <dbl>
1   8.3 -107.
2  70.7    1
```

```
base_map <- map_data("world") %>% ggplot( aes(x = long, y = lat)) +
  geom_path(aes(group = group), color = "black", size = 0.1) +
  lims(x = bbox$long, y = bbox$lat)

storms <- storms %>%
  unite("the_date", c(year, month, day), sep="-", remove="FALSE") %>%
  mutate(the_date = lubridate::ymd(the_date))

base_map <- base_map +
  geom_path(data = storms,
            aes(color = name, alpha = 0.01, size = wind, show.legend = FALSE),
            arrow = arrow(length = unit(0.005, "inches"))) +
  facet_wrap(~year)

base_map + theme(legend.position = "none")
```



```
legend<-cowplot::get_legend(base_map)
cowplot::plot_grid(legend)
```



Problem 9 (Medium):

Using the penguins data set from the palmerpenguins package:

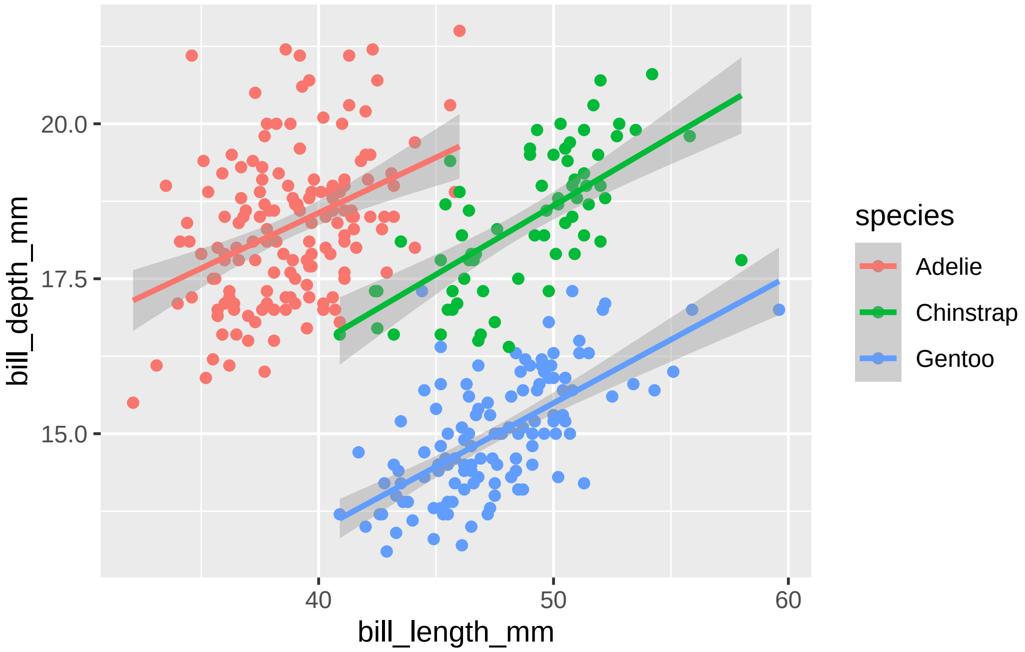
(a) Create a scatterplot of bill_length_mm against bill_depth_mm where individual species are colored and a regression line is added to each species. Add regression lines to all of your facets. What do you observe about the association of bill depth and bill length?

```
head(penguins)
```

```
# A tibble: 6 x 8
  species island    bill_length_mm bill_depth_mm flipper_l~1 body_~2 sex    year
  <fct>   <fct>              <dbl>         <dbl>       <int>   <int> <fct> <int>
1 Adelie  Torgersen           39.1          18.7         181    3750 male   2007
2 Adelie  Torgersen           39.5          17.4         186    3800 fema~  2007
3 Adelie  Torgersen           40.3          18           195    3250 fema~  2007
4 Adelie  Torgersen           NA            NA            NA      NA <NA>   2007
5 Adelie  Torgersen           36.7          19.3         193    3450 fema~  2007
6 Adelie  Torgersen           39.3          20.6         190    3650 male   2007
# ... with abbreviated variable names 1: flipper_length_mm, 2: body_mass_g
```

```
p1 <- penguins %>%
  ggplot(aes(x = bill_length_mm, # set aesthetics for x
             y = bill_depth_mm, # set aesthetics for y
             color = species)) + # color by species
  geom_point() + # create scatter plot
  geom_smooth(method = 'lm') # add regression line for each species


p1
```

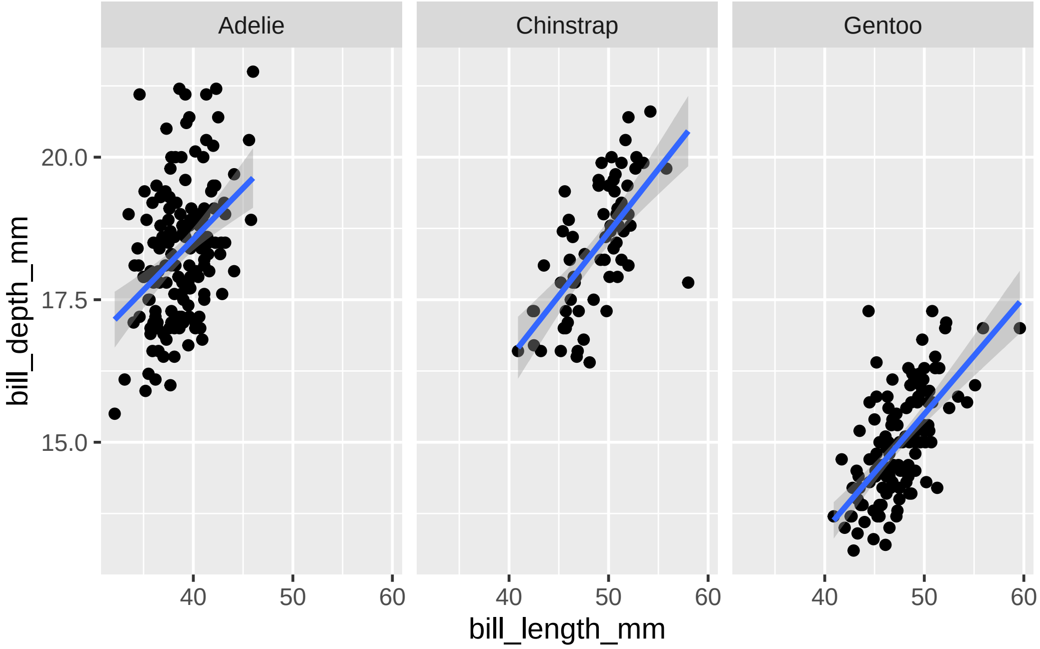
```
`geom_smooth()` using formula = 'y ~ x'
```



(b) Repeat the same scatterplot but now separate your plot into facets by `species`. How would you summarize the association between bill depth and bill length.

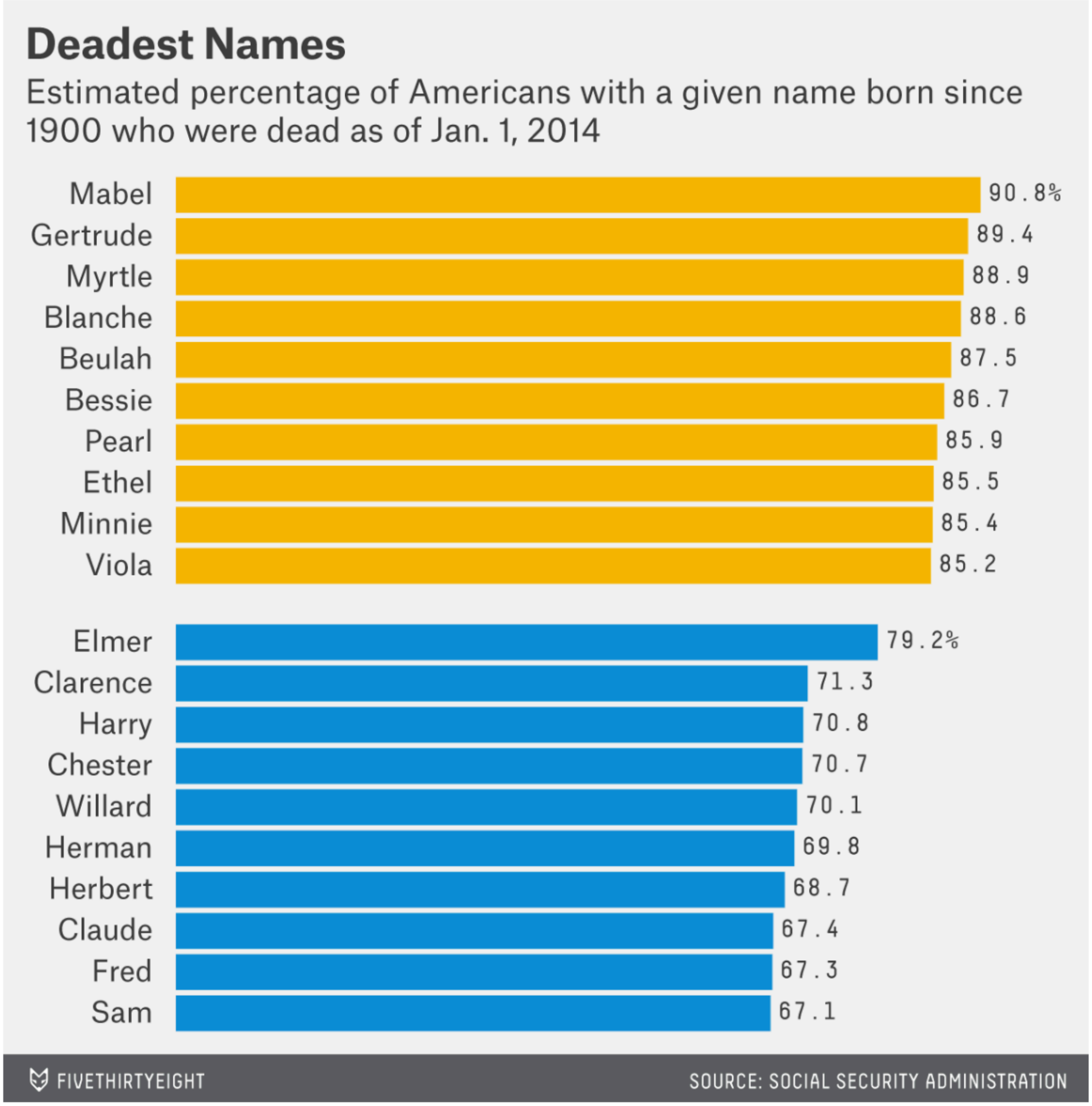
```
p2 <- penguins %>%
  ggplot(aes(x = bill_length_mm, y = bill_depth_mm)) + #set aesthetics
  geom_point() + #create scatterplot
  geom_smooth(method = 'lm') + #add regression line
  facet_wrap( ~ species) #facet by species

p2
```

```
`geom_smooth()` using formula = 'y ~ x'
```


Adelie
Chinstrap
Gentoo
bill_depth_mm
20.0
17.5
15.0
40
50
60
40
50
60
40
50
60
bill_length_mm

Problem 10 (Hard):

Use the `make_babynames_dist()` function in the `mdsr` package to recreate the "Deadest Names" graphic from FiveThirtyEight (https://fivethirtyeight.com/features/how-to-tell-someones-age-when-all-you-know-is-her-name).


Deadest Names

Estimated percentage of Americans with a given name born since 1900 who were dead as of Jan. 1, 2014

SOURCE: SOCIAL SECURITY ADMINISTRATION

```r
babynames_dist <- make_babynames_dist()
head(babynames_dist)
```

```
# A tibble: 6 x 9
   year sex   name          n   prop alive_prob count_thousands age_to~1 est_a~2
  <dbl> <chr> <chr>     <int>  <dbl>      <dbl>           <dbl>    <dbl>   <dbl>
1  1900 F     Mary      16706 0.0526          0           16.7       114       0
2  1900 F     Helen      6343 0.0200          0            6.34      114       0
3  1900 F     Anna       6114 0.0192          0            6.11      114       0
4  1900 F     Margaret   5304 0.0167          0            5.30      114       0
5  1900 F     Ruth       4765 0.0150          0            4.76      114       0
6  1900 F     Elizabeth  4096 0.0129          0            4.10      114       0
# ... with abbreviated variable names 1: age_today, 2: est_alive_today
```

```r
deadest <- babynames_dist %>%
  filter(year >= 1900) %>% #filter by years greater than or equal to 1900
  group_by(name, sex) %>% # group by name and sex
  summarise(N = n(), # count observations
            total_est_alive_today = sum(est_alive_today), #create column of total estimate
            total = sum(n)) %>%
  mutate(percent_dead = 1 - (total_est_alive_today / total)) %>% #create column of percent
  filter(total > 50000) %>% #filter out rows less than or equal to 50000
  arrange(desc(percent_dead)) %>% #arrange in descending order by percentage dead
  group_by(sex) %>% #group by sex
  top_n(10) #
```

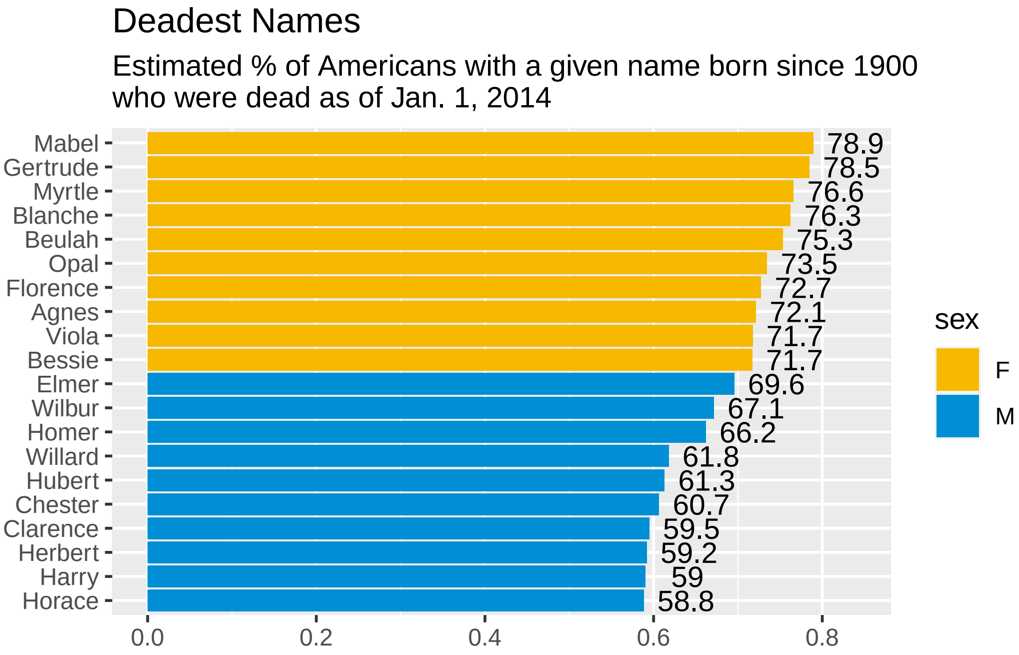
```
`summarise()` has grouped output by 'name'. You can override using the
`.groups` argument.
Selecting by percent_dead
```

```r
head(deadest)
```

```
# A tibble: 6 x 6
# Groups:   sex [1]
  name     sex       N total_est_alive_today  total percent_dead
  <chr>    <chr> <int>                 <dbl>  <int>        <dbl>
1 Mabel    F       111                20238.  96044        0.789
2 Gertrude F       111                31365. 145703        0.785
3 Myrtle   F        99                25492. 108943        0.766
```

```
4 Blanche  F        111                 16511.  69526        0.763
5 Beulah   F        111                 15647.  63367        0.753
6 Opal     F        111                 17471.  65823        0.735
```

```
ggplot(deadest, aes(reorder(name, percent_dead), percent_dead, fill = sex)) +
  geom_bar(stat = "identity") +
  geom_text(aes(y = percent_dead + 0.05), label = paste(round(deadest$percent_dead * 100,
  coord_flip() +
  ggtitle("Deadest Names", subtitle = "Estimated % of Americans with a given name born sin
scale_x_discrete(NULL) + scale_y_continuous(NULL) +
scale_fill_manual(values = c("#f6b900", "#008fd5"))
```


Deadest Names
Estimated % of Americans with a given name born since 1900
who were dead as of Jan. 1, 2014
Mabel
Gertrude
Myrtle
Blanche
Beulah
Opal
Florence
Agnes
Viola
Bessie
Elmer
Wilbur
Homer
Willard
Hubert
Chester
Clarence
Herbert
Harry
Horace
sex
F
M