Wrangling Lab

Zachary Dougherty

October 30, 2018

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(ggplot2)
#install.packages("babynames")
library(babynames)
babynames_df <- babynames
#install.packages("ggplot2movies")
library(ggplot2movies)
movies_df <- movies
```

Baby Names

Question 1

```
count(babynames_df, year)
```

```
## # A tibble: 136 x 2
##     year    nn
##    <dbl> <int>
##  1  1880  2000
##  2  1881  1935
##  3  1882  2127
##  4  1883  2084
##  5  1884  2297
##  6  1885  2294
##  7  1886  2392
##  8  1887  2373
##  9  1888  2651
## 10  1889  2590
## # ... with 126 more rows
```

```
summarise(group_by(babynames_df, year), total_births = sum(n))
```

```
## # A tibble: 136 x 2
##     year total_births
##    <dbl>        <int>
##  1  1880       201482
##  2  1881       192696
##  3  1882       221534
##  4  1883       216945
##  5  1884       243463
##  6  1885       240854
##  7  1886       255319
##  8  1887       247396
##  9  1888       299474
## 10  1889       288948
## # ... with 126 more rows
```

Question 2

```
summarise(group_by(babynames_df,year), unique_name = length(unique(name)))
```

```
## # A tibble: 136 x 2
##     year unique_name
##    <dbl>       <int>
##  1  1880        1889
##  2  1881        1830
##  3  1882        2012
##  4  1883        1962
##  5  1884        2158
##  6  1885        2139
##  7  1886        2225
##  8  1887        2215
##  9  1888        2454
## 10  1889        2390
## # ... with 126 more rows
```

Question 3

```
sex_year_pair <- unite(babynames_df, col = year_sex, year, sex, sep = ", ")
count_sex_per_year <- count(sex_year_pair, year_sex)
colnames(count_sex_per_year) <- c("year_sex","births")
```

Question 4

```
separate(sex_year_pair, year_sex, into = c("year","sex"), sep = "\\, ")
```

```
## # A tibble: 1,858,689 x 5
##    year  sex   name          n   prop
##    <chr> <chr> <chr>     <int>  <dbl>
##  1 1880  F     Mary       7065 0.0724
##  2 1880  F     Anna       2604 0.0267
##  3 1880  F     Emma       2003 0.0205
##  4 1880  F     Elizabeth  1939 0.0199
##  5 1880  F     Minnie     1746 0.0179
##  6 1880  F     Margaret   1578 0.0162
##  7 1880  F     Ida        1472 0.0151
##  8 1880  F     Alice      1414 0.0145
```

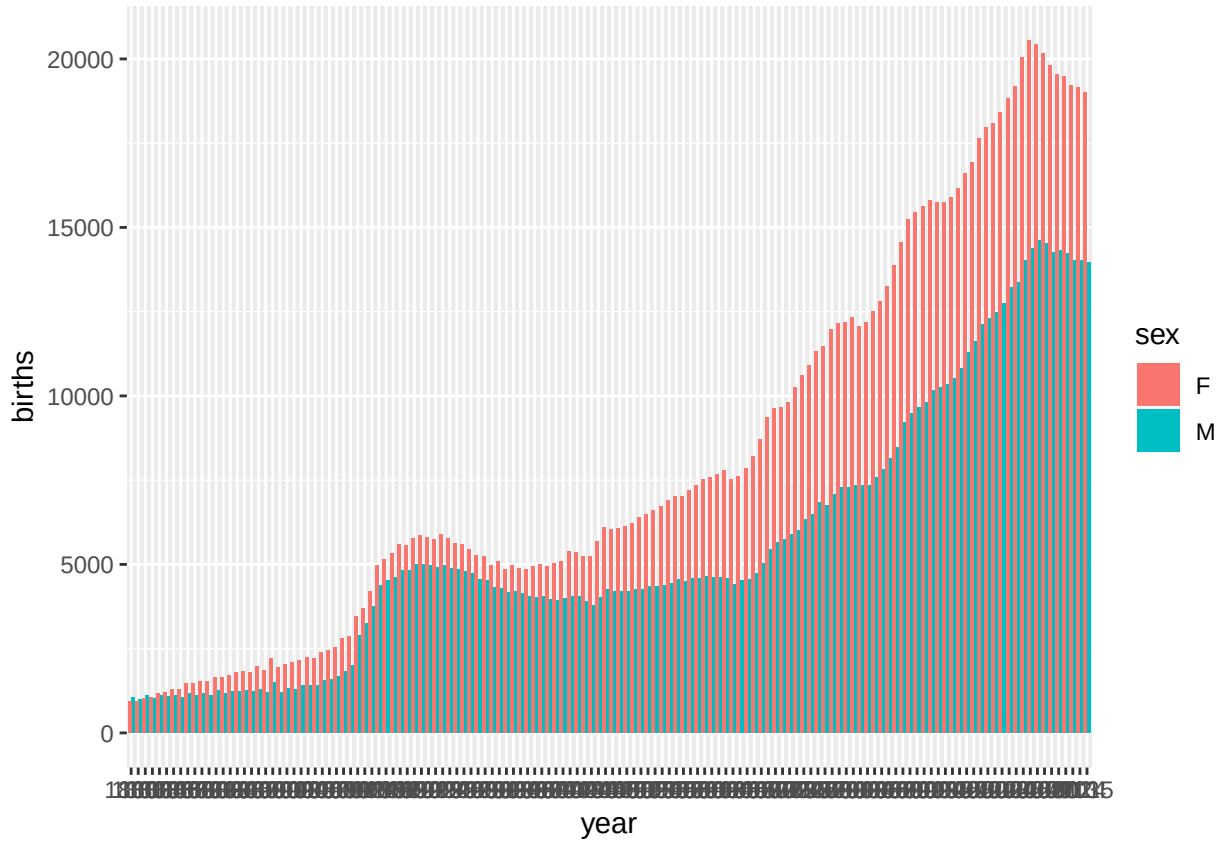
```
##  9 1880  F     Bertha      1320 0.0135
## 10 1880  F     Sarah       1288 0.0132
## # ... with 1,858,679 more rows
```

Question 5

```
new_sex_per_year <- separate(count_sex_per_year, year_sex,
                             into = c("year","sex"),
                             sep = "\\, ")
new_sex_per_year
```

```
## # A tibble: 272 x 3
##    year  sex   births
##    <chr> <chr>  <int>
##  1 1880  F        942
##  2 1880  M       1058
##  3 1881  F        938
##  4 1881  M        997
##  5 1882  F       1028
##  6 1882  M       1099
##  7 1883  F       1054
##  8 1883  M       1030
##  9 1884  F       1172
## 10 1884  M       1125
## # ... with 262 more rows
```

```
ggplot(new_sex_per_year, aes(x = year, y = births, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge")
```



Movies

Question Six

```
str(movies_df)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    58788 obs. of  24 variables:
##  $ title      : chr  "$" "$1000 a Touchdown" "$21 a Day Once a Month" "$40,000" ...
##  $ year       : int  1971 1939 1941 1996 1975 2000 2002 2002 1987 1917 ...
##  $ length     : int  121 71 7 70 71 91 93 25 97 61 ...
##  $ budget     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ rating     : num  6.4 6 8.2 8.2 3.4 4.3 5.3 6.7 6.6 6 ...
##  $ votes      : int  348 20 5 6 17 45 200 24 18 51 ...
##  $ r1         : num  4.5 0 0 14.5 24.5 4.5 4.5 4.5 4.5 4.5 ...
##  $ r2         : num  4.5 14.5 0 0 4.5 4.5 0 4.5 4.5 0 ...
##  $ r3         : num  4.5 4.5 0 0 0 4.5 4.5 4.5 4.5 4.5 ...
##  $ r4         : num  4.5 24.5 0 0 14.5 14.5 4.5 4.5 0 4.5 ...
##  $ r5         : num  14.5 14.5 0 0 14.5 14.5 24.5 4.5 0 4.5 ...
##  $ r6         : num  24.5 14.5 24.5 0 4.5 14.5 24.5 14.5 0 44.5 ...
##  $ r7         : num  24.5 14.5 0 0 0 4.5 14.5 14.5 34.5 14.5 ...
##  $ r8         : num  14.5 4.5 44.5 0 0 4.5 4.5 14.5 14.5 4.5 ...
##  $ r9         : num  4.5 4.5 24.5 34.5 0 14.5 4.5 4.5 4.5 4.5 ...
##  $ r10        : num  4.5 14.5 24.5 45.5 24.5 14.5 14.5 14.5 24.5 4.5 ...
##  $ mpaa       : chr  "" "" "" "" ...
##  $ Action     : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Animation  : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ Comedy     : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ Drama      : int  1 0 0 0 0 1 1 0 1 0 ...
##  $ Documentary: int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Romance    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Short      : int  0 0 1 0 0 0 0 1 0 0 ...
```

Question 7

```
ratings <- select(movies_df, num_range("r",1:10))
```

Question 8

```
average_ratings_one <- summarise_all(ratings, funs(mean))
```

Question 9

```
average_ratings_two <- ratings %>% summarise_all(.,funs(mean))
```

Question 10

```
long_movies <- gather(ratings, key = "rater", value = "rating")
```

Question 11

```
longer_movies <- gather(movies_df, key = "genre", value = "encoding",
                        Action, Animation, Comedy, Drama, Documentary,
                        Romance, Short)
longer_movies
```

```
## # A tibble: 411,516 x 19
##    title  year length budget rating votes    r1    r2    r3    r4    r5
##    <chr> <int>  <int>  <int>  <dbl> <int> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 $      1971    121     NA    6.4   348   4.5   4.5   4.5   4.5  14.5
```

```
##  2 $100~  1939     71     NA    6      20   0    14.5   4.5  24.5  14.5
##  3 $21 ~  1941      7     NA    8.2     5   0     0     0     0     0
##  4 $40,~  1996     70     NA    8.2     6  14.5   0     0     0     0
##  5 $50,~  1975     71     NA    3.4    17  24.5   4.5   0    14.5  14.5
##  6 $pent  2000     91     NA    4.3    45   4.5   4.5   4.5  14.5  14.5
##  7 $win~  2002     93     NA    5.3   200   4.5   0     4.5   4.5  24.5
##  8 '15'   2002     25     NA    6.7    24   4.5   4.5   4.5   4.5   4.5
##  9 '38    1987     97     NA    6.6    18   4.5   4.5   4.5   0     0
## 10 '49-~  1917     61     NA    6      51   4.5   0     4.5   4.5   4.5
## # ... with 411,506 more rows, and 8 more variables: r6 <dbl>, r7 <dbl>,
## #   r8 <dbl>, r9 <dbl>, r10 <dbl>, mpaa <chr>, genre <chr>, encoding <int>
```

Question 12

```
new_longer_movies <- filter(longer_movies, encoding != 0, rating < 15)
```

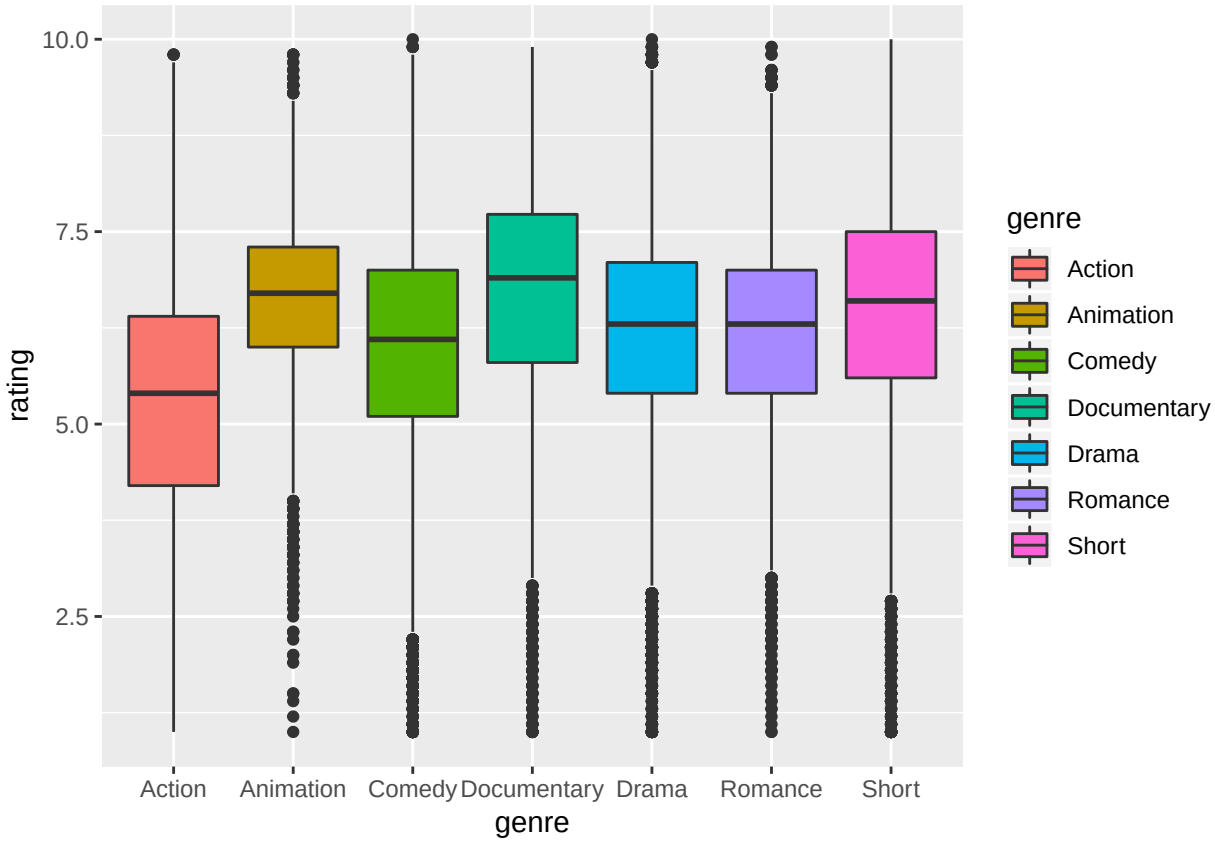
Question 13

```
longer_movies[longer_movies$encoding != 0 & longer_movies$rating < 15,]
```

```
## # A tibble: 65,134 x 19
##    title  year length budget rating votes    r1    r2    r3    r4    r5
##    <chr> <int>  <int>  <int>  <dbl> <int> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 $win~  2002     93     NA    5.3   200   4.5   0     4.5   4.5  24.5
##  2 'A' ~  1983    106     NA    7.1  1259   4.5   4.5   4.5   4.5   4.5
##  3 'A' ~  1987    101     NA    7.2   614   4.5   4.5   4.5   4.5   4.5
##  4 'Cro~  1988    110     NA    5    7252   4.5   4.5   4.5  14.5  24.5
##  5 'Gat~  1974     88     NA    3.5   100  14.5  14.5  24.5  14.5  14.5
##  6 'She~  1975     90     NA    5.5    91   4.5   4.5   4.5  14.5  14.5
##  7 ...A~  1981    113     NA    5.6   348   4.5   4.5   4.5   4.5  14.5
##  8 ...P~  1990     88     NA    4.7    11   4.5   0     4.5   4.5   4.5
##  9 ...t~  1970    100     NA    6     145   4.5   4.5   4.5  14.5  14.5
## 10 002 ~  1964     83     NA    3.6     6  34.5   0     0     0     0
## # ... with 65,124 more rows, and 8 more variables: r6 <dbl>, r7 <dbl>,
## #   r8 <dbl>, r9 <dbl>, r10 <dbl>, mpaa <chr>, genre <chr>, encoding <int>
```

Question 14

```
ggplot(new_longer_movies, aes(x = genre, y = rating, fill = genre)) +
  geom_boxplot()
```


10.0
7.5
5.0
2.5
rating
Action
Animation
Comedy
Documentary
Drama
Romance
Short
genre
genre
Action
Animation
Comedy
Documentary
Drama
Romance
Short