Data manipulation with dplyr

Maruf Ahmed Bhuiyan

8/5/2020

```r
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
```

```r
counties = readRDS("counties.rds")

str(counties)
```

```
## tibble [3,138 x 40] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ census_id         : chr [1:3138] "1001" "1003" "1005" "1007" ...
##  $ state             : chr [1:3138] "Alabama" "Alabama" "Alabama" "Alabama" ...
##  $ county            : chr [1:3138] "Autauga" "Baldwin" "Barbour" "Bibb" ...
##  $ region            : chr [1:3138] "South" "South" "South" "South" ...
##  $ metro             : chr [1:3138] "Metro" "Metro" "Nonmetro" "Metro" ...
##  $ population        : num [1:3138] 55221 195121 26932 22604 57710 ...
##  $ men               : num [1:3138] 26745 95314 14497 12073 28512 ...
##  $ women             : num [1:3138] 28476 99807 12435 10531 29198 ...
##  $ hispanic          : num [1:3138] 2.6 4.5 4.6 2.2 8.6 4.4 1.2 3.5 0.4 1.5 ...
##  $ white             : num [1:3138] 75.8 83.1 46.2 74.5 87.9 22.2 53.3 73 57.3 91.7 ...
##  $ black             : num [1:3138] 18.5 9.5 46.7 21.4 1.5 70.7 43.8 20.3 40.3 4.8 ...
##  $ native            : num [1:3138] 0.4 0.6 0.2 0.4 0.3 1.2 0.1 0.2 0.2 0.6 ...
##  $ asian             : num [1:3138] 1 0.7 0.4 0.1 0.1 0.2 0.4 0.9 0.8 0.3 ...
##  $ pacific           : num [1:3138] 0 0 0 0 0 0 0 0 0 0 ...
##  $ citizens          : num [1:3138] 40725 147695 20714 17495 42345 ...
##  $ income            : num [1:3138] 51281 50254 32964 38678 45813 ...
##  $ income_err        : num [1:3138] 2391 1263 2973 3995 3141 ...
##  $ income_per_cap    : num [1:3138] 24974 27317 16824 18431 20532 ...
##  $ income_per_cap_err: num [1:3138] 1080 711 798 1618 708 ...
##  $ poverty           : num [1:3138] 12.9 13.4 26.7 16.8 16.7 24.6 25.4 20.5 21.6 19.2 ...
##  $ child_poverty     : num [1:3138] 18.6 19.2 45.3 27.9 27.2 38.4 39.2 31.6 37.2 30.1 ...
##  $ professional      : num [1:3138] 33.2 33.1 26.8 21.5 28.5 18.8 27.5 27.3 23.3 29.3 ...
```

```
##  $ service            : num [1:3138] 17 17.7 16.1 17.9 14.1 15 16.6 17.7 14.5 16 ...
##  $ office             : num [1:3138] 24.2 27.1 23.1 17.8 23.9 19.7 21.9 24.2 26.3 19.5 ...
##  $ construction       : num [1:3138] 8.6 10.8 10.8 19 13.5 20.1 10.3 10.5 11.5 13.7 ...
##  $ production         : num [1:3138] 17.1 11.2 23.1 23.7 19.9 26.4 23.7 20.4 24.4 21.5 ...
##  $ drive              : num [1:3138] 87.5 84.7 83.8 83.2 84.9 74.9 84.5 85.3 85.1 83.9 ...
##  $ carpool            : num [1:3138] 8.8 8.8 10.9 13.5 11.2 14.9 12.4 9.4 11.9 12.1 ...
##  $ transit            : num [1:3138] 0.1 0.1 0.4 0.5 0.4 0.7 0 0.2 0.2 0.2 ...
##  $ walk               : num [1:3138] 0.5 1 1.8 0.6 0.9 5 0.8 1.2 0.3 0.6 ...
##  $ other_transp       : num [1:3138] 1.3 1.4 1.5 1.5 0.4 1.7 0.6 1.2 0.4 0.7 ...
##  $ work_at_home       : num [1:3138] 1.8 3.9 1.6 0.7 2.3 2.8 1.7 2.7 2.1 2.5 ...
##  $ mean_commute       : num [1:3138] 26.5 26.4 24.1 28.8 34.9 27.5 24.6 24.1 25.1 27.4 ...
##  $ employed           : num [1:3138] 23986 85953 8597 8294 22189 ...
##  $ private_work       : num [1:3138] 73.6 81.5 71.8 76.8 82 79.5 77.4 74.1 85.1 73.1 ...
##  $ public_work        : num [1:3138] 20.9 12.3 20.8 16.1 13.5 15.1 16.2 20.8 12.1 18.5 ...
##  $ self_employed      : num [1:3138] 5.5 5.8 7.3 6.7 4.2 5.4 6.2 5 2.8 7.9 ...
##  $ family_work        : num [1:3138] 0 0.4 0.1 0.4 0.4 0 0.2 0.1 0 0.5 ...
##  $ unemployment       : num [1:3138] 7.6 7.5 17.6 8.3 7.7 18 10.9 12.3 8.9 7.9 ...
##  $ land_area          : num [1:3138] 594 1590 885 623 645 ...
```

```r
counties %>% select(state, county, population, unemployment)
```

```
## # A tibble: 3,138 x 4
##    state   county   population unemployment
##    <chr>   <chr>         <dbl>        <dbl>
##  1 Alabama Autauga       55221          7.6
##  2 Alabama Baldwin      195121          7.5
##  3 Alabama Barbour       26932         17.6
##  4 Alabama Bibb          22604          8.3
##  5 Alabama Blount        57710          7.7
##  6 Alabama Bullock       10678         18
##  7 Alabama Butler        20354         10.9
##  8 Alabama Calhoun      116648         12.3
##  9 Alabama Chambers      34079          8.9
## 10 Alabama Cherokee      26008          7.9
## # ... with 3,128 more rows
```

Understanding your data Take a look at the counties dataset using the glimpse() function. What is the first value in the income variable?

```r
glimpse(counties)
```

```
## Rows: 3,138
## Columns: 40
## $ census_id          <chr> "1001", "1003", "1005", "1007", "1009", "1011", ...
## $ state              <chr> "Alabama", "Alabama", "Alabama", "Alabama", "Ala...
## $ county             <chr> "Autauga", "Baldwin", "Barbour", "Bibb", "Blount...
## $ region             <chr> "South", "South", "South", "South", "South", "So...
## $ metro              <chr> "Metro", "Metro", "Nonmetro", "Metro", "Metro", ...
## $ population         <dbl> 55221, 195121, 26932, 22604, 57710, 10678, 20354...
## $ men                <dbl> 26745, 95314, 14497, 12073, 28512, 5660, 9502, 5...
## $ women              <dbl> 28476, 99807, 12435, 10531, 29198, 5018, 10852, ...
## $ hispanic           <dbl> 2.6, 4.5, 4.6, 2.2, 8.6, 4.4, 1.2, 3.5, 0.4, 1.5...
```

```
## $ white              <dbl> 75.8, 83.1, 46.2, 74.5, 87.9, 22.2, 53.3, 73.0, ...
## $ black              <dbl> 18.5, 9.5, 46.7, 21.4, 1.5, 70.7, 43.8, 20.3, 40...
## $ native             <dbl> 0.4, 0.6, 0.2, 0.4, 0.3, 1.2, 0.1, 0.2, 0.2, 0.6...
## $ asian              <dbl> 1.0, 0.7, 0.4, 0.1, 0.1, 0.2, 0.4, 0.9, 0.8, 0.3...
## $ pacific            <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0...
## $ citizens           <dbl> 40725, 147695, 20714, 17495, 42345, 8057, 15581,...
## $ income             <dbl> 51281, 50254, 32964, 38678, 45813, 31938, 32229,...
## $ income_err         <dbl> 2391, 1263, 2973, 3995, 3141, 5884, 1793, 925, 2...
## $ income_per_cap     <dbl> 24974, 27317, 16824, 18431, 20532, 17580, 18390,...
## $ income_per_cap_err <dbl> 1080, 711, 798, 1618, 708, 2055, 714, 489, 1366,...
## $ poverty            <dbl> 12.9, 13.4, 26.7, 16.8, 16.7, 24.6, 25.4, 20.5, ...
## $ child_poverty      <dbl> 18.6, 19.2, 45.3, 27.9, 27.2, 38.4, 39.2, 31.6, ...
## $ professional       <dbl> 33.2, 33.1, 26.8, 21.5, 28.5, 18.8, 27.5, 27.3, ...
## $ service            <dbl> 17.0, 17.7, 16.1, 17.9, 14.1, 15.0, 16.6, 17.7, ...
## $ office             <dbl> 24.2, 27.1, 23.1, 17.8, 23.9, 19.7, 21.9, 24.2, ...
## $ construction       <dbl> 8.6, 10.8, 10.8, 19.0, 13.5, 20.1, 10.3, 10.5, 1...
## $ production         <dbl> 17.1, 11.2, 23.1, 23.7, 19.9, 26.4, 23.7, 20.4, ...
## $ drive              <dbl> 87.5, 84.7, 83.8, 83.2, 84.9, 74.9, 84.5, 85.3, ...
## $ carpool            <dbl> 8.8, 8.8, 10.9, 13.5, 11.2, 14.9, 12.4, 9.4, 11....
## $ transit            <dbl> 0.1, 0.1, 0.4, 0.5, 0.4, 0.7, 0.0, 0.2, 0.2, 0.2...
## $ walk               <dbl> 0.5, 1.0, 1.8, 0.6, 0.9, 5.0, 0.8, 1.2, 0.3, 0.6...
## $ other_transp       <dbl> 1.3, 1.4, 1.5, 1.5, 0.4, 1.7, 0.6, 1.2, 0.4, 0.7...
## $ work_at_home       <dbl> 1.8, 3.9, 1.6, 0.7, 2.3, 2.8, 1.7, 2.7, 2.1, 2.5...
## $ mean_commute       <dbl> 26.5, 26.4, 24.1, 28.8, 34.9, 27.5, 24.6, 24.1, ...
## $ employed           <dbl> 23986, 85953, 8597, 8294, 22189, 3865, 7813, 474...
## $ private_work       <dbl> 73.6, 81.5, 71.8, 76.8, 82.0, 79.5, 77.4, 74.1, ...
## $ public_work        <dbl> 20.9, 12.3, 20.8, 16.1, 13.5, 15.1, 16.2, 20.8, ...
## $ self_employed      <dbl> 5.5, 5.8, 7.3, 6.7, 4.2, 5.4, 6.2, 5.0, 2.8, 7.9...
## $ family_work        <dbl> 0.0, 0.4, 0.1, 0.4, 0.4, 0.0, 0.2, 0.1, 0.0, 0.5...
## $ unemployment       <dbl> 7.6, 7.5, 17.6, 8.3, 7.7, 18.0, 10.9, 12.3, 8.9,...
## $ land_area          <dbl> 594.44, 1589.78, 884.88, 622.58, 644.78, 622.81,...
```

```
# Answer: 51281
```

Selecting columns Select the following four columns from the counties variable: [x] state [x] county [x] population [x] poverty

You don't need to save the result to a variable. Select the columns listed from the counties variable.

```
counties %>% select(state, county, population, poverty)
```

```
## # A tibble: 3,138 x 4
##    state   county   population poverty
##    <chr>   <chr>         <dbl>   <dbl>
##  1 Alabama Autauga       55221    12.9
##  2 Alabama Baldwin      195121    13.4
##  3 Alabama Barbour       26932    26.7
##  4 Alabama Bibb          22604    16.8
##  5 Alabama Blount        57710    16.7
##  6 Alabama Bullock       10678    24.6
##  7 Alabama Butler        20354    25.4
##  8 Alabama Calhoun      116648    20.5
##  9 Alabama Chambers      34079    21.6
```

```
## 10 Alabama Cherokee      26008    19.2
## # ... with 3,128 more rows
```

```r
counties_selected <-
      counties %>% select(state, county, population, unemployment)

counties_selected
```

```
## # A tibble: 3,138 x 4
##    state   county   population unemployment
##    <chr>   <chr>         <dbl>        <dbl>
##  1 Alabama Autauga       55221          7.6
##  2 Alabama Baldwin      195121          7.5
##  3 Alabama Barbour       26932         17.6
##  4 Alabama Bibb          22604          8.3
##  5 Alabama Blount        57710          7.7
##  6 Alabama Bullock       10678         18
##  7 Alabama Butler        20354         10.9
##  8 Alabama Calhoun      116648         12.3
##  9 Alabama Chambers      34079          8.9
## 10 Alabama Cherokee      26008          7.9
## # ... with 3,128 more rows
```

```r
counties_selected %>% arrange(population)
```

```
## # A tibble: 3,138 x 4
##    state      county    population unemployment
##    <chr>      <chr>          <dbl>        <dbl>
##  1 Hawaii     Kalawao           85          0
##  2 Texas      King             267          5.1
##  3 Nebraska   McPherson        433          0.9
##  4 Montana    Petroleum        443          6.6
##  5 Nebraska   Arthur           448          4
##  6 Nebraska   Loup             548          0.7
##  7 Nebraska   Blaine           551          0.7
##  8 New Mexico Harding          565          6
##  9 Texas      Kenedy           565          0
## 10 Colorado   San Juan         606         13.8
## # ... with 3,128 more rows
```

```r
counties_selected %>% arrange(-population)
```

```
## # A tibble: 3,138 x 4
##    state      county      population unemployment
##    <chr>      <chr>            <dbl>        <dbl>
##  1 California Los Angeles   10038388         10
##  2 Illinois   Cook           5236393         10.7
##  3 Texas      Harris         4356362          7.5
##  4 Arizona    Maricopa       4018143          7.7
##  5 California San Diego      3223096          8.7
##  6 California Orange         3116069          7.6
##  7 Florida    Miami-Dade     2639042         10
```

```
##  8 New York   Kings          2595259         10
##  9 Texas      Dallas         2485003          7.6
## 10 New York   Queens         2301139          8.6
## # ... with 3,128 more rows
```

```
counties_selected %>% arrange(desc(population))
```

```
## # A tibble: 3,138 x 4
##    state      county      population unemployment
##    <chr>      <chr>            <dbl>        <dbl>
##  1 California Los Angeles   10038388         10
##  2 Illinois   Cook           5236393         10.7
##  3 Texas      Harris         4356362          7.5
##  4 Arizona    Maricopa       4018143          7.7
##  5 California San Diego      3223096          8.7
##  6 California Orange         3116069          7.6
##  7 Florida    Miami-Dade     2639042         10
##  8 New York   Kings          2595259         10
##  9 Texas      Dallas         2485003          7.6
## 10 New York   Queens         2301139          8.6
## # ... with 3,128 more rows
```

```
counties_selected %>%
  arrange(desc(population)) %>%
    filter(state == "New York")
```

```
## # A tibble: 62 x 4
##    state    county      population unemployment
##    <chr>    <chr>            <dbl>        <dbl>
##  1 New York Kings          2595259         10
##  2 New York Queens         2301139          8.6
##  3 New York New York       1629507          7.5
##  4 New York Suffolk        1501373          6.4
##  5 New York Bronx          1428357         14
##  6 New York Nassau         1354612          6.4
##  7 New York Westchester     967315          7.6
##  8 New York Erie            921584          7
##  9 New York Monroe          749356          7.7
## 10 New York Richmond        472481          6.9
## # ... with 52 more rows
```

```
counties_selected %>%
  arrange(desc(population)) %>%
    filter(state == "New York") %>%
      filter(unemployment < 6)
```

```
## # A tibble: 5 x 4
##   state    county     population unemployment
##   <chr>    <chr>           <dbl>        <dbl>
## 1 New York Tompkins       103855          5.9
## 2 New York Chemung         88267          5.4
## 3 New York Madison         72427          5.1
```

```
## 4 New York Livingston      64801          5.4
## 5 New York Seneca          35144          5.5
```

```
counties_selected %>%
  arrange(desc(population)) %>%
    filter(state == "New York", unemployment < 6)
```

```
## # A tibble: 5 x 4
##   state    county     population unemployment
##   <chr>    <chr>           <dbl>        <dbl>
## 1 New York Tompkins       103855          5.9
## 2 New York Chemung         88267          5.4
## 3 New York Madison         72427          5.1
## 4 New York Livingston      64801          5.4
## 5 New York Seneca          35144          5.5
```

Arranging observations Here you see the counties_selected dataset with a few interesting variables selected. These variables: private_work, public_work, self_employed describe whether people work for the government, for private companies, or for themselves. In these exercises, you'll sort these observations to find the most interesting cases.

[x] Add a verb to sort the observations of the public_work variable in descending order.

```
counties_selected <- counties %>%
  select(state, county, population, private_work, public_work, self_employed)

# Add a verb to sort in descending order of public_work
counties_selected %>% arrange(desc(public_work))
```

```
## # A tibble: 3,138 x 6
##    state     county             population private_work public_work self_employed
##    <chr>     <chr>                   <dbl>        <dbl>       <dbl>         <dbl>
##  1 Hawaii    Kalawao                    85         25          64.1          10.9
##  2 Alaska    Yukon-Koyukuk Ce~        5644         33.3        61.7           5.1
##  3 Wisconsin Menominee                4451         36.8        59.1           3.7
##  4 North Da~ Sioux                    4380         32.9        56.8          10.2
##  5 South Da~ Todd                     9942         34.4        55             9.8
##  6 Alaska    Lake and Peninsu~        1474         42.2        51.6           6.1
##  7 Californ~ Lassen                  32645         42.6        50.5           6.8
##  8 South Da~ Buffalo                  2038         48.4        49.5           1.8
##  9 South Da~ Dewey                    5579         34.9        49.2          14.7
## 10 Texas     Kenedy                    565         51.9        48.1           0
## # ... with 3,128 more rows
```

Filtering for conditions You use the filter() verb to get only observations that match a particular condition, or match multiple conditions. [x] Find only the counties that have a population above one million (1000000). [x] Find only the counties in the state of California that also have a population above one million (1000000).

```
counties_selected <- counties %>%
  select(state, county, population)
```

```r
# Filter for counties with a population above 1000000
counties_selected %>% filter(population > 1000000)
```

```
## # A tibble: 41 x 3
##    state      county         population
##    <chr>      <chr>               <dbl>
##  1 Arizona    Maricopa          4018143
##  2 California Alameda           1584983
##  3 California Contra Costa      1096068
##  4 California Los Angeles      10038388
##  5 California Orange            3116069
##  6 California Riverside         2298032
##  7 California Sacramento        1465832
##  8 California San Bernardino    2094769
##  9 California San Diego         3223096
## 10 California Santa Clara       1868149
## # ... with 31 more rows
```

```r
# Filter for counties in the state of California that have a population above 1000000
counties_selected %>% filter(state == "California",
                             population > 1000000)
```

```
## # A tibble: 9 x 3
##   state      county         population
##   <chr>      <chr>               <dbl>
## 1 California Alameda           1584983
## 2 California Contra Costa      1096068
## 3 California Los Angeles      10038388
## 4 California Orange            3116069
## 5 California Riverside         2298032
## 6 California Sacramento        1465832
## 7 California San Bernardino    2094769
## 8 California San Diego         3223096
## 9 California Santa Clara       1868149
```

Filtering and arranging We're often interested in both filtering and sorting a dataset, to focus on observations of particular interest to you. Here, you'll find counties that are extreme examples of what fraction of the population works in the private sector.

[x] Filter for counties in the state of Texas that have more than ten thousand people (10000), and sort them in descending order of the percentage of people employed in private work.

```r
counties_selected <- counties %>%
  select(state, county, population, private_work, public_work, self_employed)

# Filter for Texas and more than 10000 people; sort in descending order of private_work
counties_selected %>% filter(state == "Texas",
                             population > 10000) %>%
                      arrange(-private_work)
```

```
## # A tibble: 169 x 6
##    state county  population private_work public_work self_employed
```

```
##    <chr> <chr>        <dbl>        <dbl>       <dbl>         <dbl>
##  1 Texas Gregg       123178         84.7         9.8           5.4
##  2 Texas Collin      862215         84.1        10             5.8
##  3 Texas Dallas     2485003         83.9         9.5           6.4
##  4 Texas Harris     4356362         83.4        10.1           6.3
##  5 Texas Andrews      16775         83.1         9.6           6.8
##  6 Texas Tarrant    1914526         83.1        11.4           5.4
##  7 Texas Titus        32553         82.5        10             7.4
##  8 Texas Denton      731851         82.2        11.9           5.7
##  9 Texas Ector       149557         82          11.2           6.7
## 10 Texas Moore        22281         82          11.7           5.9
## # ... with 159 more rows
```

```
counties_selected <- counties %>%
            select(state, county, population, unemployment)

counties_selected %>%
              mutate(unemployed_population = population * unemployment / 100)
```

```
## # A tibble: 3,138 x 5
##    state   county   population unemployment unemployed_population
##    <chr>   <chr>         <dbl>        <dbl>                 <dbl>
##  1 Alabama Autauga       55221          7.6                 4197.
##  2 Alabama Baldwin      195121          7.5                14634.
##  3 Alabama Barbour       26932         17.6                 4740.
##  4 Alabama Bibb          22604          8.3                 1876.
##  5 Alabama Blount        57710          7.7                 4444.
##  6 Alabama Bullock       10678         18                   1922.
##  7 Alabama Butler        20354         10.9                 2219.
##  8 Alabama Calhoun      116648         12.3                14348.
##  9 Alabama Chambers      34079          8.9                 3033.
## 10 Alabama Cherokee      26008          7.9                 2055.
## # ... with 3,128 more rows
```

```
counties_selected %>%
              mutate(unemployed_population = population * unemployment / 100) %>%
              arrange(desc(unemployed_population))
```

```
## # A tibble: 3,138 x 5
##    state      county         population unemployment unemployed_population
##    <chr>      <chr>               <dbl>        <dbl>                 <dbl>
##  1 California Los Angeles      10038388         10                1003839.
##  2 Illinois   Cook              5236393         10.7               560294.
##  3 Texas      Harris            4356362          7.5               326727.
##  4 Arizona    Maricopa          4018143          7.7               309397.
##  5 California Riverside         2298032         12.9               296446.
##  6 California San Diego         3223096          8.7               280409.
##  7 Michigan   Wayne             1778969         14.9               265066.
##  8 California San Bernardino    2094769         12.6               263941.
##  9 Florida    Miami-Dade        2639042         10                 263904.
## 10 New York   Kings             2595259         10                 259526.
## # ... with 3,128 more rows
```

Calculating the number of government employees In the video, you used the unemployment variable, which is a percentage, to calculate the number of unemployed people in each county. In this exercise, you'll do the same with another percentage variable: public_work. The code provided already selects the state, county, population, and public_work columns.

[x] Use mutate() to add a column called public_workers to the dataset, with the number of people employed in public (government) work. [x] Sort the new column in descending order.

```
counties_selected <- counties %>%
  select(state, county, population, public_work)

head(counties_selected)
```

```
## # A tibble: 6 x 4
##   state   county  population public_work
##   <chr>   <chr>        <dbl>       <dbl>
## 1 Alabama Autauga      55221        20.9
## 2 Alabama Baldwin     195121        12.3
## 3 Alabama Barbour      26932        20.8
## 4 Alabama Bibb         22604        16.1
## 5 Alabama Blount       57710        13.5
## 6 Alabama Bullock      10678        15.1
```

```
# Add a new column public_workers with the number of people employed in public work
counties_selected %>%
              mutate(public_workers = population * public_work / 100)
```

```
## # A tibble: 3,138 x 5
##    state   county   population public_work public_workers
##    <chr>   <chr>         <dbl>       <dbl>          <dbl>
##  1 Alabama Autauga       55221        20.9         11541.
##  2 Alabama Baldwin      195121        12.3         24000.
##  3 Alabama Barbour       26932        20.8          5602.
##  4 Alabama Bibb          22604        16.1          3639.
##  5 Alabama Blount        57710        13.5          7791.
##  6 Alabama Bullock       10678        15.1          1612.
##  7 Alabama Butler        20354        16.2          3297.
##  8 Alabama Calhoun      116648        20.8         24263.
##  9 Alabama Chambers      34079        12.1          4124.
## 10 Alabama Cherokee      26008        18.5          4811.
## # ... with 3,128 more rows
```

```
# Sort in descending order of the public_workers column
counties_selected %>%
              mutate(public_workers = population * public_work / 100) %>%
              arrange(-public_workers)
```

```
## # A tibble: 3,138 x 5
##    state      county         population public_work public_workers
##    <chr>      <chr>               <dbl>       <dbl>          <dbl>
##  1 California Los Angeles      10038388        11.5       1154415.
##  2 Illinois   Cook              5236393        11.5        602185.
##  3 California San Diego         3223096        14.8        477018.
```

```
##  4 Arizona    Maricopa          4018143        11.7         470123.
##  5 Texas      Harris            4356362        10.1         439993.
##  6 New York   Kings             2595259        14.4         373717.
##  7 California San Bernardino    2094769        16.7         349826.
##  8 California Riverside         2298032        14.9         342407.
##  9 California Sacramento        1465832        21.8         319551.
## 10 California Orange            3116069        10.2         317839.
## # ... with 3,128 more rows
```

Calculating the percentage of women in a county The dataset includes columns for the total number (not percentage) of men and women in each county. You could use this, along with the population variable, to compute the fraction of men (or women) within each county. In this exercise, you'll select the relevant columns yourself.

[x] Select the columns state, county, population, men, and women. [x] Add a new variable called proportion_women with the fraction of the county's population made up of women.

```
# Select the columns state, county, population, men, and women
counties_selected <- counties %>% select(state, county, population, men, women)
head(counties_selected)
```

```
## # A tibble: 6 x 5
##   state   county  population   men women
##   <chr>   <chr>        <dbl> <dbl> <dbl>
## 1 Alabama Autauga      55221 26745 28476
## 2 Alabama Baldwin     195121 95314 99807
## 3 Alabama Barbour      26932 14497 12435
## 4 Alabama Bibb         22604 12073 10531
## 5 Alabama Blount       57710 28512 29198
## 6 Alabama Bullock      10678  5660  5018
```

```
# Calculate proportion_women as the fraction of the population made up of women
counties_selected %>% mutate(proportion_women = women / population)
```

```
## # A tibble: 3,138 x 6
##    state   county   population   men women proportion_women
##    <chr>   <chr>         <dbl> <dbl> <dbl>            <dbl>
##  1 Alabama Autauga       55221 26745 28476            0.516
##  2 Alabama Baldwin      195121 95314 99807            0.512
##  3 Alabama Barbour       26932 14497 12435            0.462
##  4 Alabama Bibb          22604 12073 10531            0.466
##  5 Alabama Blount        57710 28512 29198            0.506
##  6 Alabama Bullock       10678  5660  5018            0.470
##  7 Alabama Butler        20354  9502 10852            0.533
##  8 Alabama Calhoun      116648 56274 60374            0.518
##  9 Alabama Chambers      34079 16258 17821            0.523
## 10 Alabama Cherokee      26008 12975 13033            0.501
## # ... with 3,128 more rows
```

Select, mutate, filter, and arrange In this exercise, you'll put together everything you've learned in this chapter (select(), mutate(), filter() and arrange()), to find the counties with the highest proportion of men.

[x] Select only the columns state, county, population, men, and women. [x] Add a variable proportion_men with the fraction of the county's population made up of men. [x] Filter for counties with a population of at least ten thousand (10000). [x] Arrange counties in descending order of their proportion of men.

```
counties %>%
  # Select the five columns
  select(state, county, population, men, women) %>%
  # Add the proportion_men variable
  mutate(proportion_men = men / population) %>%
  # Filter for population of at least 10,000
  filter(population >= 10000) %>%
  # Arrange proportion of men in descending order
  arrange(-proportion_men)
```

```
## # A tibble: 2,437 x 6
##    state      county          population   men women proportion_men
##    <chr>      <chr>                <dbl> <dbl> <dbl>          <dbl>
##  1 Virginia   Sussex               11864  8130  3734          0.685
##  2 California Lassen               32645 21818 10827          0.668
##  3 Georgia    Chattahoochee        11914  7940  3974          0.666
##  4 Louisiana  West Feliciana       15415 10228  5187          0.664
##  5 Florida    Union                15191  9830  5361          0.647
##  6 Texas      Jones                19978 12652  7326          0.633
##  7 Missouri   DeKalb               12782  8080  4702          0.632
##  8 Texas      Madison              13838  8648  5190          0.625
##  9 Virginia   Greensville          11760  7303  4457          0.621
## 10 Texas      Anderson             57915 35469 22446          0.612
## # ... with 2,427 more rows
```

```
# Sussex County in Virginia is more than two thirds male:
# this is because of two men's prisons in the county.
```