

The count verb

DATA MANIPULATION WITH DPLYR



Chris Cardillo
Data Scientist

Count

```
counties %>%  
  count()
```

```
# A tibble: 1 x 1  
      n  
  <int>  
1  3138
```

Count variable

```
counties %>%  
  count(state)
```

```
# A tibble: 50 x 2  
  state      n  
  <chr>    <int>  
1 Alabama    67  
2 Alaska    28  
3 Arizona    15  
4 Arkansas   75  
5 California  58  
6 Colorado   64  
7 Connecticut  8  
8 Delaware    3  
9 Florida    67  
10 Georgia   159  
# ... with 40 more rows
```

Count and sort

```
counties %>%  
  count(state, sort = TRUE)
```

```
# A tibble: 50 x 2  
  state      n  
  <chr>    <int>  
1 Texas    253  
2 Georgia  159  
3 Virginia 133  
4 Kentucky 120  
5 Missouri 115  
6 Kansas   105  
7 Illinois 102  
8 North Carolina 100  
9 Iowa      99  
10 Tennessee 95  
# ... with 40 more rows
```

Count population

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

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

Add weight

```
counties %>%  
  count(state, wt = population, sort = TRUE)
```

```
# A tibble: 50 x 2  
  state          n  
  <chr>        <dbl>  
1 California 38421464  
2 Texas      26538497  
3 New York   19673174  
4 Florida    19645772  
5 Illinois   12873761  
6 Pennsylvania 12779559  
7 Ohio       11575977  
8 Georgia    10006693  
9 Michigan    9900571  
10 North Carolina 9845333  
# ... with 40 more rows
```

Let's practice!

DATA MANIPULATION WITH DPLYR

Group by and summarize

DATA MANIPULATION WITH DPLYR



Chris Cardillo
Data Scientist

Summarize

```
counties %>%  
  summarize(total_population = sum(population))
```

```
# A tibble: 1 x 1  
  total_population  
      <dbl>  
1      315845353
```

Aggregate and summarize

```
counties %>%  
  summarize(total_population = sum(population),  
            average_unemployment = mean(unemployment))
```

```
# A tibble: 1 x 2  
  total_population average_unemployment  
      <dbl>          <dbl>  
1    315845353          7.80
```

Summary functions

- `sum()`
- `mean()`
- `median()`
- `min()`
- `max()`
- `n()`

Aggregate within groups

```
counties %>%  
  group_by(state) %>%  
  summarize(total_pop = sum(population),  
            average_unemployment = sum(unemployment))
```

```
# A tibble: 50 x 3  
  state      total_pop average_unemployment  
  <chr>      <dbl>          <dbl>  
1 Alabama    4830620         758.  
2 Alaska     725461         257.  
3 Arizona    6641928         180.  
4 Arkansas   2958208         674.  
5 California 38421464         626.  
6 Colorado   5278906         477.  
7 Connecticut 3593222          65.3  
8 Delaware   926454          23.8  
9 Florida    19645772         696.  
10 Georgia   10006693        1586.  
# ... with 40 more rows
```

Arrange

```
counties %>%
  group_by(state) %>%
  summarize(total_pop = sum(population),
            average_unemployment = mean(unemployment)) %>%
  arrange(desc(average_unemployment))
```

```
# A tibble: 50 x 3
  state      total_pop average_unemployment
  <chr>      <dbl>         <dbl>
1 Mississippi 2988081         12.0
2 Arizona     6641928         12.0
3 South Carolina 4777576         11.3
4 Alabama     4830620         11.3
5 California  38421464         10.8
6 Nevada      2798636         10.5
7 North Carolina 9845333         10.5
8 Florida     19645772         10.4
9 Georgia     10006693          9.97
10 Michigan   9900571          9.96
# ... with 40 more rows
```

Metro column

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

```
# A tibble: 3,138 x 4  
  state  metro  county  population  
  <chr>  <chr>  <chr>      <dbl>  
1 Alabama Metro   Autauga    55221  
2 Alabama Metro   Baldwin  195121  
3 Alabama Nonmetro Barbour   26932  
4 Alabama Metro   Bibb     22604  
5 Alabama Metro   Blount   57710  
6 Alabama Nonmetro Bullock   10678  
7 Alabama Nonmetro Butler    20354  
8 Alabama Metro   Calhoun  116648  
9 Alabama Nonmetro Chambers   34079  
10 Alabama Nonmetro Cherokee   26008  
# ... with 3,128 more rows
```

Group by

```
counties %>%  
  group_by(state, metro) %>%  
  summarize(total_pop = sum(population))
```

```
# A tibble: 97 x 3  
# Groups:   state [50]  
   state      metro total_pop  
   <chr>    <chr>    <dbl>  
1 Alabama Metro    3671377  
2 Alabama Nonmetro 1159243  
3 Alaska Metro    494990  
4 Alaska Nonmetro 230471  
5 Arizona Metro    6295145  
6 Arizona Nonmetro 346783  
7 Arkansas Metro    1806867  
8 Arkansas Nonmetro 1151341  
9 California Metro    37587429  
10 California Nonmetro 834035  
# ... with 87 more rows
```

Ungroup

```
counties %>%  
  group_by(state, metro) %>%  
  summarize(total_pop = sum(population)) %>%  
  ungroup()
```

```
# A tibble: 97 x 3  
  state      metro total_pop  
  <chr>    <chr>    <dbl>  
1 Alabama Metro      3671377  
2 Alabama Nonmetro  1159243  
3 Alaska  Metro      494990  
4 Alaska  Nonmetro   230471  
5 Arizona Metro     6295145  
6 Arizona Nonmetro   346783  
7 Arkansas Metro     1806867  
8 Arkansas Nonmetro  1151341  
9 California Metro    37587429  
10 California Nonmetro   834035  
# ... with 87 more rows
```


Let's practice!

DATA MANIPULATION WITH DPLYR

The top_n verb

DATA MANIPULATION WITH DPLYR



Chris Cardillo
Data Scientist

top_n

```
counties_selected <- counties %>%  
  select(state, county, population, unemployment, income)  
  
counties_selected %>%  
  group_by(state) %>%  
  top_n(1, population)
```

top_n

```
# A tibble: 50 x 5
# Groups:   state [50]
  state      county      population unemployment income
  <chr>      <chr>          <dbl>          <dbl>    <dbl>
1 Alabama   Jefferson        659026          9.1    45610
2 Alaska    Anchorage Municipality 299107          6.7    78326
3 Arizona   Maricopa         4018143         7.7    54229
4 Arkansas  Pulaski          390463          7.5    46140
5 California Los Angeles     10038388        10     56196
6 Colorado  El Paso          655024          8.4    58206
7 Connecticut Fairfield       939983          9     84233
8 Delaware  New Castle       549643          7.4    65476
9 Florida   Miami-Dade      2639042         10     43129
10 Georgia   Fulton          983903          9.9    57207
# ... with 40 more rows
```

Highest unemployment

```
counties_selected %>%  
  group_by(state) %>%  
  top_n(1, unemployment)
```

```
# A tibble: 51 x 5  
# Groups:   state [50]  
  state      county      population unemployment income  
  <chr>    <chr>          <dbl>         <dbl>   <dbl>  
1 Alabama Conecuh          12865          22.6   24900  
2 Alaska  Northwest Arctic Borough    7732          21.9   63648  
3 Arizona Navajo          107656          19.8   35921  
4 Arkansas Phillips         20391          18.1   26844  
5 California Imperial       178206          17.4   41079  
6 Colorado Crowley           5551           27    31151  
7 Connecticut New Haven      862224           9.5   61640  
8 Delaware Kent          169509           8.4   54976  
9 Florida Hamilton         14395          15.8   35048  
10 Georgia Taylor           8401          20.6   28143  
# ... with 41 more rows
```

Number of observations

```
counties_selected %>%  
  group_by(state) %>%  
  top_n(3, unemployment)
```

```
# A tibble: 153 x 5  
# Groups:   state [50]  
   state county      population unemployment income  
   <chr> <chr>          <dbl>         <dbl> <dbl>  
1 Alabama Conecuh      12865          22.6  24900  
2 Alabama Monroe      22217          20.7  27257  
3 Alabama Wilcox      11235          20.8  23750  
4 Alaska Bethel Census Area 17776          17.6  51012  
5 Alaska Northwest Arctic Borough 7732          21.9  63648  
6 Alaska Yukon-Koyukuk Census Area 5644          18.2  38491  
7 Arizona Apache      72124          18.2  31757  
8 Arizona Graham      37407          14.1  45964  
9 Arizona Navajo     107656          19.8  35921  
10 Arkansas Desha      12379          17.7  27197  
# ... with 143 more rows
```

Let's practice!

DATA MANIPULATION WITH DPLYR