

Selecting

DATA MANIPULATION WITH DPLYR



Chris Cardillo

Data Scientist

Select

```
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
```

Select a range

```
counties %>%  
  select(state, county, drive:work_at_home)
```

```
# A tibble: 3,138 x 8  
  state    county    drive carpool transit  walk other_transp work_at_home  
  <chr>   <chr>   <dbl>  <dbl>  <dbl> <dbl>         <dbl>         <dbl>  
1 Alabama Autauga    87.5    8.8    0.1  0.5          1.3          1.8  
2 Alabama Baldwin   84.7    8.8    0.1  1           1.4          3.9  
3 Alabama Barbour   83.8   10.9    0.4  1.8          1.5          1.6  
4 Alabama Bibb      83.2   13.5    0.5  0.6          1.5          0.7  
5 Alabama Blount    84.9   11.2    0.4  0.9          0.4          2.3  
6 Alabama Bullock   74.9   14.9    0.7  5            1.7          2.8  
7 Alabama Butler    84.5   12.4    0    0.8          0.6          1.7  
8 Alabama Calhoun   85.3    9.4    0.2  1.2          1.2          2.7  
9 Alabama Chambers  85.1   11.9    0.2  0.3          0.4          2.1  
10 Alabama Cherokee 83.9   12.1    0.2  0.6          0.7          2.5  
# ... with 3,128 more rows
```

Select and arrange

```
counties %>%  
  select(state, county, drive:work_at_home) %>%  
  arrange(drive)
```

```
# A tibble: 3,138 x 8  
  state      county      drive carpool transit  walk other_transp work_at_home  
  <chr>    <chr>      <dbl>  <dbl>  <dbl> <dbl>      <dbl>      <dbl>  
1 New York New York      6.1    1.9   59.2  20.7      5.4      6.8  
2 Alaska  Northwest Arctic Borough 16.5   10.4    0.4  46.9     21.2     4.6  
3 Alaska  Aleutians East Borough  18.4    4.9    0.5  71.2     2.2     2.8  
4 New York Kings      18.6    4.4   61.7   8.8     2.5     3.9  
5 Alaska  North Slope Borough    20.1   17     2.8  37.9     7.9    14.3  
6 Alaska  Lake and Peninsula Borough 21.2    6.8    1.1  36.2    32.4     2.4  
7 New York Bronx      22.5    4.7   59.7   8     1.8     3.3  
8 Alaska  Nome Census Area      25.8   10     0.3  36.9    22.7     4.3  
9 Alaska  Bethel Census Area     26.5  12.7    0.5  33     22.6     4.8  
10 Alaska Yukon-Koyukuk Census Area 28.7    8.1    0.2  38.1    20.1     4.9  
# ... with 3,128 more rows
```

Contains

```
counties %>%  
  select(state, county, contains("work"))
```

```
# A tibble: 3,138 x 6  
  state    county work_at_home private_work public_work family_work  
  <chr>   <chr>         <dbl>         <dbl>         <dbl>         <dbl>  
1 Alabama Autauga         1.8          73.6          20.9           0  
2 Alabama Baldwin         3.9          81.5          12.3           0.4  
3 Alabama Barbour         1.6          71.8          20.8           0.1  
4 Alabama Bibb             0.7          76.8          16.1           0.4  
5 Alabama Blount          2.3           82           13.5           0.4  
6 Alabama Bullock         2.8          79.5          15.1           0  
7 Alabama Butler          1.7          77.4          16.2           0.2  
8 Alabama Calhoun         2.7          74.1          20.8           0.1  
9 Alabama Chambers        2.1          85.1          12.1           0  
10 Alabama Cherokee        2.5          73.1          18.5           0.5  
# ... with 3,128 more rows
```

Starts with

```
counties %>%  
  select(state, county, starts_with("income"))
```

```
# A tibble: 3,138 x 6  
  state    county  income income_err income_per_cap income_per_cap_err  
  <chr>   <chr>   <dbl>    <dbl>         <dbl>         <dbl>  
1 Alabama Autauga    51281     2391      24974         1080  
2 Alabama Baldwin   50254     1263      27317          711  
3 Alabama Barbour   32964     2973      16824          798  
4 Alabama Bibb      38678     3995      18431         1618  
5 Alabama Blount    45813     3141      20532          708  
6 Alabama Bullock   31938     5884      17580         2055  
7 Alabama Butler    32229     1793      18390          714  
8 Alabama Calhoun   41703      925      21374          489  
9 Alabama Chambers  34177     2949      21071         1366  
10 Alabama Cherokee  36296     1710      21811         1556  
# ... with 3,128 more rows
```

Other helpers

- `contains()`
- `starts_with()`
- `ends_with()`
- `last_col()`

For more:

```
?select_helpers
```

Removing a variable

```
counties %>%  
  select(-census_id)
```

```
# A tibble: 3,138 x 39  
  state county region metro population  men women hispanic white black native asian pacific citizens income  
  <chr> <chr>  <chr> <chr>      <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>  
1 Alab... Autau... South Metro      55221 26745 28476      2.6  75.8  18.5    0.4    1      0    40725  51281  
2 Alab... Baldw... South Metro     195121 95314 99807      4.5  83.1   9.5    0.6    0.7    0   147695  50254  
3 Alab... Barbo... South Nonm...     26932 14497 12435      4.6  46.2  46.7    0.2    0.4    0    20714  32964  
4 Alab... Bibb    South Metro      22604 12073 10531      2.2  74.5  21.4    0.4    0.1    0    17495  38678  
5 Alab... Blount  South Metro      57710 28512 29198      8.6  87.9   1.5    0.3    0.1    0    42345  45813  
6 Alab... Bullo... South Nonm...     10678  5660  5018      4.4  22.2  70.7    1.2    0.2    0     8057  31938  
7 Alab... Butler  South Nonm...     20354  9502 10852      1.2  53.3  43.8    0.1    0.4    0    15581  32229  
8 Alab... Calho... South Metro     116648 56274 60374      3.5   73    20.3    0.2    0.9    0    88612  41703  
9 Alab... Chamb... South Nonm...      34079 16258 17821      0.4  57.3  40.3    0.2    0.8    0    26462  34177  
10 Alab... Chero... South Nonm...      26008 12975 13033      1.5  91.7   4.8    0.6    0.3    0    20600  36296  
# ... with 3,128 more rows, and 24 more variables: income_err <dbl>, income_per_cap <dbl>,  
# income_per_cap_err <dbl>, poverty <dbl>, child_poverty <dbl>, professional <dbl>, service <dbl>,  
# office <dbl>, construction <dbl>, production <dbl>, drive <dbl>, carpool <dbl>, transit <dbl>, walk <dbl>,  
# other_transp <dbl>, work_at_home <dbl>, mean_commute <dbl>, employed <dbl>, private_work <dbl>,  
# public_work <dbl>, self_employed <dbl>, family_work <dbl>, unemployment <dbl>, land_area <dbl>
```


Let's practice!

DATA MANIPULATION WITH DPLYR

The rename verb

DATA MANIPULATION WITH DPLYR



Chris Cardillo

Data Scientist

Select columns

```
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
```

Rename a column

```
counties_selected %>%  
  rename(unemployment_rate = unemployment)
```

```
# A tibble: 3,138 x 4  
  state    county    population unemployment_rate  
  <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
```

Combine verbs

```
counties_selected %>%  
  select(state, county, population, unemployment_rate = unemployment)
```

```
# A tibble: 3,138 x 4  
  state    county    population unemployment_rate  
  <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
```

Compare verbs

Select

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

Rename

```
counties %>%  
  select(state, county, population, unemployment) %>%  
  rename(unemployment_rate = unemployment)
```

Let's practice!

DATA MANIPULATION WITH DPLYR

The transmute verb

DATA MANIPULATION WITH DPLYR



Chris Cardillo

Data Scientist

Transmute

- Combination: select & mutate
- Returns a subset of columns that are transformed and changed

Select and calculate

```
counties %>%  
  transmute(state, county, fraction_men = men / population)
```

```
# A tibble: 3,138 x 3  
  state    county    fraction_men  
  <chr>   <chr>         <dbl>  
1 Alabama Autauga      0.484  
2 Alabama Baldwin     0.488  
3 Alabama Barbour     0.538  
4 Alabama Bibb        0.534  
5 Alabama Blount      0.494  
6 Alabama Bullock     0.530  
7 Alabama Butler      0.467  
8 Alabama Calhoun     0.482  
9 Alabama Chambers    0.477  
10 Alabama Cherokee   0.499  
# ... with 3,128 more rows
```

Select and calculate

```
counties %>%  
  transmute(state, county, population, unemployed_people = population * unemployment / 100)
```

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

Summary

	Keeps only specified variables	Keeps other variables
Can't change values	select	rename
Can change values	transmute	mutate

Summary

	Keeps only specified variables	Keeps other variables
Can't change values	<code>select</code>	<code>rename</code>
Can change values	<code>transmute</code>	<code>mutate</code>

Summary

	Keeps only specified variables	Keeps other variables
Can't change values	select	rename
Can change values	transmute	mutate

Summary

	Keeps only specified variables	Keeps other variables
Can't change values	<code>select</code>	<code>rename</code>
Can change values	<code>transmute</code>	<code>mutate</code>

Summary

	Keeps only specified variables	Keeps other variables
Can't change values	select	rename
Can change values	transmute	mutate

Summary

	Keeps only specified variables	Keeps other variables
Can't change values	select	rename
Can change values	transmute	mutate

Let's practice!

DATA MANIPULATION WITH DPLYR