Selecting

DATA MANIPULATION WITH DPLYR

James Chapman
Curriculum Manager, DataCamp

datacamp

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()`
- `matches()`

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>  <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

James Chapman
Curriculum Manager, DataCamp

datacamp

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 relocate verb

DATA MANIPULATION WITH DPLYR

James Chapman
Curriculum Manager, DataCamp

datacamp

Relocate

- Change column positions
- Quick and efficient syntax

county	census_id	region	state
Autauga	1001	South	Alabama
Baldwin	1003	South	Alabama

census_id	region	state	county
1001	South	Alabama	Autauga
1003	South	Alabama	Baldwin

```
counties
```

```
# A tibble: 3,138 × 40
   census_id state   county   region metro population   men women hispanic white
   <chr>     <chr>   <chr>    <chr>  <chr>       <dbl> <dbl> <dbl>    <dbl> <dbl>
 1 1001      Alabama Autauga  South  Metro       55221 26745 28476      2.6  75.8
 2 1003      Alabama Baldwin  South  Metro      195121 95314 99807      4.5  83.1
 3 1005      Alabama Barbour  South  Nonm…       26932 14497 12435      4.6  46.2
 4 1007      Alabama Bibb     South  Metro       22604 12073 10531      2.2  74.5
 5 1009      Alabama Blount   South  Metro       57710 28512 29198      8.6  87.9
 6 1011      Alabama Bullock  South  Nonm…       10678  5660  5018      4.4  22.2
 7 1013      Alabama Butler   South  Nonm…       20354  9502 10852      1.2  53.3
 8 1015      Alabama Calhoun  South  Metro      116648 56274 60374      3.5  73
 9 1017      Alabama Chambers South  Nonm…       34079 16258 17821      0.4  57.3
10 1019      Alabama Cherokee South  Nonm…       26008 12975 13033      1.5  91.7
# … with 3,128 more rows, and 30 more variables: black <dbl>, native <dbl>, …
```

Relocating to .before

```
counties %>%
  relocate(region, .before = state)
```

```
# A tibble: 3,138 × 40
   census_id region state   county   metro population   men women hispanic white
   <chr>     <chr>  <chr>   <chr>    <chr>      <dbl> <dbl> <dbl>    <dbl> <dbl>
 1 1001      South  Alabama Autauga  Metro      55221 26745 28476      2.6  75.8
 2 1003      South  Alabama Baldwin  Metro     195121 95314 99807      4.5  83.1
 3 1005      South  Alabama Barbour  Nonm…      26932 14497 12435      4.6  46.2
 4 1007      South  Alabama Bibb     Metro      22604 12073 10531      2.2  74.5
 5 1009      South  Alabama Blount   Metro      57710 28512 29198      8.6  87.9
 6 1011      South  Alabama Bullock  Nonm…      10678  5660  5018      4.4  22.2
 7 1013      South  Alabama Butler   Nonm…      20354  9502 10852      1.2  53.3
 8 1015      South  Alabama Calhoun  Metro     116648 56274 60374      3.5  73
 9 1017      South  Alabama Chambers Nonm…      34079 16258 17821      0.4  57.3
10 1019      South  Alabama Cherokee Nonm…      26008 12975 13033      1.5  91.7
# … with 3,128 more rows, and 30 more variables: black <dbl>, native <dbl>, …
```

Relocating to .after

```
counties %>%
  relocate(state, .after = region)
```

```
# A tibble: 3,138 × 40
   census_id county   region state   metro population   men women hispanic white
   <chr>     <chr>    <chr>  <chr>   <chr>      <dbl> <dbl> <dbl>    <dbl> <dbl>
 1 1001      Autauga  South  Alabama Metro      55221 26745 28476      2.6  75.8
 2 1003      Baldwin  South  Alabama Metro     195121 95314 99807      4.5  83.1
 3 1005      Barbour  South  Alabama Nonm…      26932 14497 12435      4.6  46.2
 4 1007      Bibb     South  Alabama Metro      22604 12073 10531      2.2  74.5
 5 1009      Blount   South  Alabama Metro      57710 28512 29198      8.6  87.9
 6 1011      Bullock  South  Alabama Nonm…      10678  5660  5018      4.4  22.2
 7 1013      Butler   South  Alabama Nonm…      20354  9502 10852      1.2  53.3
 8 1015      Calhoun  South  Alabama Metro     116648 56274 60374      3.5  73
 9 1017      Chambers South  Alabama Nonm…      34079 16258 17821      0.4  57.3
10 1019      Cherokee South  Alabama Nonm…      26008 12975 13033      1.5  91.7
# … with 3,128 more rows, and 30 more variables: black <dbl>, native <dbl>, …
```

relocate() + select helpers

```
counties %>%
  relocate(census_id, .after = last_col())
```

```
# A tibble: 3,138 × 40
   state  county region metro population   men women hispanic white black native
   <chr>  <chr>  <chr>  <chr>      <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl>  <dbl>
 1 Alaba… Autau… South  Metro      55221 26745 28476      2.6  75.8  18.5    0.4
 2 Alaba… Baldw… South  Metro     195121 95314 99807      4.5  83.1   9.5    0.6
 3 Alaba… Barbo… South  Nonm…      26932 14497 12435      4.6  46.2  46.7    0.2
 4 Alaba… Bibb   South  Metro      22604 12073 10531      2.2  74.5  21.4    0.4
 5 Alaba… Blount South  Metro      57710 28512 29198      8.6  87.9   1.5    0.3
 6 Alaba… Bullo… South  Nonm…      10678  5660  5018      4.4  22.2  70.7    1.2
 7 Alaba… Butler South  Nonm…      20354  9502 10852      1.2  53.3  43.8    0.1
 8 Alaba… Calho… South  Metro     116648 56274 60374      3.5  73    20.3    0.2
 9 Alaba… Chamb… South  Nonm…      34079 16258 17821      0.4  57.3  40.3    0.2
10 Alaba… Chero… South  Nonm…      26008 12975 13033      1.5  91.7   4.8    0.6
# … with 3,128 more rows, and 29 more variables: asian <dbl>, … census_id <chr>
```

select() vs. relocate()?

`select()`

- Keeps only the columns specified
- Good for subsetting and moving

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

`relocate()`

- Keeps all columns
- Best choice for moving-only

```
counties %>%
  relocate(region, .before = state)
```

Let's practice!

DATA MANIPULATION WITH DPLYR