

Class 3: Grammar of data analysis - Part I

API-201

Quantitative Analysis and Empirical Methods I

Profs. Goel and Taylor

Harvard Kennedy School

Agenda

1. Introduce the grammar of data analysis [`tidyverse`]
2. Apply this language to study discrimination in police stops

Data frames

It's often convenient to organize data into **data frames**, which are tables of information. We typically call the rows of a data frame *observations*, and call the columns of a data frame *variables*.

Data frames

To construct a data frame, we use the **tibble** function.

[Can use the **data.frame** function instead.]

```
pets <- tibble(  
  name = c('wolfie', 'bmo', 'blue', 'elmo'),  
  is_dog = c(TRUE, TRUE, FALSE, FALSE),  
  age = c(2, 3, 1.5, 4)  
)  
pets
```

```
A tibble: 4 x 3  
  name   is_dog    age  
  <chr>  <lgl> <dbl>  
1 wolfie TRUE     2.0  
2 bmo    TRUE     3.0  
3 blue   FALSE    1.5  
4 elmo   FALSE    4.0
```

Five verbs of data manipulation

- **filter** returns a subset of rows [observations] in a data frame based on some condition.
- **select** returns a subset of columns [variables] of a data frame.
- **mutate** adds or modifies columns [variables] of a data frame.
- **summarize** collapses multiple values into a single value.
- **arrange** orders the rows [observations] in a data frame based on some criterion.

Data manipulation

Consider the data frame **w eo** with country-level information.

```
head(w eo)
```

A tibble: 6 × 4			
country	continent	pop_2020	rgdp_2020
<chr>	<chr>	<dbl>	<dbl>
Afghanistan	Asia	37.000	69202.210
Albania	Europe	2.865	36628.799
Algeria	Africa	43.863	621442.475

Angola Africa 30.793 187947.539
Antigua and Barbuda North America 0.094 2376.677
Argentina South America 45.561 909620.490

- Starting with the **w eo** data frame, which verb would we use to return a data frame of countries in Asia?

filter

select

mutate

summarize

arrange



- Suppose we'd like a data frame of Asian countries, with an extra column showing GDP per capita. Which two verbs would we use?

filter and select

select and mutate

filter and mutate

mutate and summarize



Start the presentation to see live content. For screen share software, share the entire screen. Get help at pollev.com/app

The group_by operation

The five verbs of data manipulation are particularly powerful when we combine them with the `group_by` operation.

Under the hood, R **splits** the data into groups, **applies** one or more of the five verbs to each group, and then **combines** the results for each group into a new data frame.

The group_by operation

Police stops

	Date	Precinct	Race	Weapon
Stop 1	2012-07-15	3	White	TRUE
Stop 2	2012-07-15	3	Black	TRUE
Stop 3	2012-07-15	1	White	FALSE
Stop 4	2012-07-15	2	Black	TRUE
Stop 5	2012-07-15	2	Black	FALSE
Stop 6	2012-07-15	3	Black	FALSE

The group_by operation

Police stops

Hit rate **by** race

Hit rate **by** precinct

Hit rate **by** race and precinct

	Date	Precinct	Race	Weapon
Stop 1	2012-07-15	3	White	TRUE
Stop 2	2012-07-15	3	Black	TRUE
Stop 3	2012-07-15	1	White	FALSE
Stop 4	2012-07-15	2	Black	FALSE
Stop 5	2012-07-15	2	Black	FALSE
Stop 6	2012-07-15	3	Black	FALSE



Precinct	Race	Hit rate
1	White	0
2	Black	0
3	Black	0.5
3	White	1

- To compute the average GDP across countries in each continent, which verb would we combine with `group_by`?

filter

select

mutate

summarize

arrange



Key ideas

The five verbs of data manipulation – plus the group_by operation – allow us to carry out sophisticated descriptive statistical analyses

Putting it all together

Now we'll use what we learned to investigate racial discrimination in police stops.

<https://bit.ly/API201-dplyr>