Data Wrangling II

INFO 201

Today’s Objectives

Review DPLYR's ***grammar*** of data manipulation

Discuss proper ***data shape*** for analysis

Learn how to ***group observations*** and compute summary information

Understand how to ***join data frames*** together

Grammar of Data Manipulation

Why is it helpful to use ***a grammar*** of data manipulation?

Grammar of Data Manipulation

Select particular columns

Filter down to specific rows

Arrange (sort) your dataset by values

Mutate your dataframe to add a column

Summarise your dataframe (calculate summary info, mean)

Today's Dataset

	year	month	day	dep_time	dep_delay	arr_time	arr_delay	carrier	tailnum	flight	origin	dest	air_time	distance	hour	minute
1	2013	1	1	517	2	830	11	UA	N14228	1545	EWR	IAH	227	1400	5	17
2	2013	1	1	533	4	850	20	UA	N24211	1714	LGA	IAH	227	1416	5	33
3	2013	1	1	542	2	923	33	AA	N619AA	1141	JFK	MIA	160	1089	5	42
4	2013	1	1	544	-1	1004	-18	B6	N804JB	725	JFK	BQN	183	1576	5	44
5	2013	1	1	554	-6	812	-25	DL	N668DN	461	LGA	ATL	116	762	5	54
6	2013	1	1	554	-4	740	12	UA	N39463	1696	EWR	ORD	150	719	5	54
7	2013	1	1	555	-5	913	19	B6	N516JB	507	EWR	FLL	158	1065	5	55
8	2013	1	1	557	-3	709	-14	EV	N829AS	5708	LGA	IAD	53	229	5	57
9	2013	1	1	557	-3	838	-8	B6	N593JB	79	JFK	MCO	140	944	5	57
10	2013	1	1	558	-2	753	8	AA	N3ALAA	301	LGA	ORD	138	733	5	58
11	2013	1	1	558	-2	849	-2	B6	N793JB	49	JFK	PBI	149	1028	5	58
12	2013	1	1	558	-2	853	-3	B6	N657JB	71	JFK	TPA	158	1005	5	58
13	2013	1	1	558	-2	924	7	UA	N29129	194	JFK	LAX	345	2475	5	58
14	2013	1	1	558	-2	923	-14	UA	N53441	1124	EWR	SFO	361	2565	5	58
15	2013	1	1	559	-1	941	31	AA	N3DUAA	707	LGA	DFW	257	1389	5	59
16	2013	1	1	559	0	702	-4	B6	N708JB	1806	JFK	BOS	44	187	5	59
17	2013	1	1	559	-1	854	-8	UA	N76515	1187	EWR	LAS	337	2227	5	59
18	2013	1	1	600	0	851	-7	B6	N595JB	371	LGA	FLL	152	1076	6	0

Today's Dataset

module 9 exercise-4

Grouped Operations

What are reasons that you would want to calculate summary information for ***groups of observations***?

What are reasons that you would **not** want to calculate summary information for ***groups of observations***?

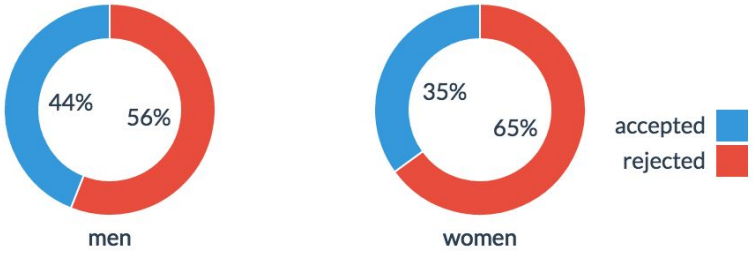
Simpson's Paradox

Tweet

Girls gone average.
Averages gone wild.

In 1973, the University of California-Berkeley was sued for sex discrimination. The numbers looked pretty incriminating: the graduate schools had just accepted 44% of male applicants but only 35% of female applicants. When researchers looked at the evidence, though, they uncovered something surprising:

> If the data are properly pooled...there is a small but statistically significant bias in favor of women.
>
> — (p. 403)

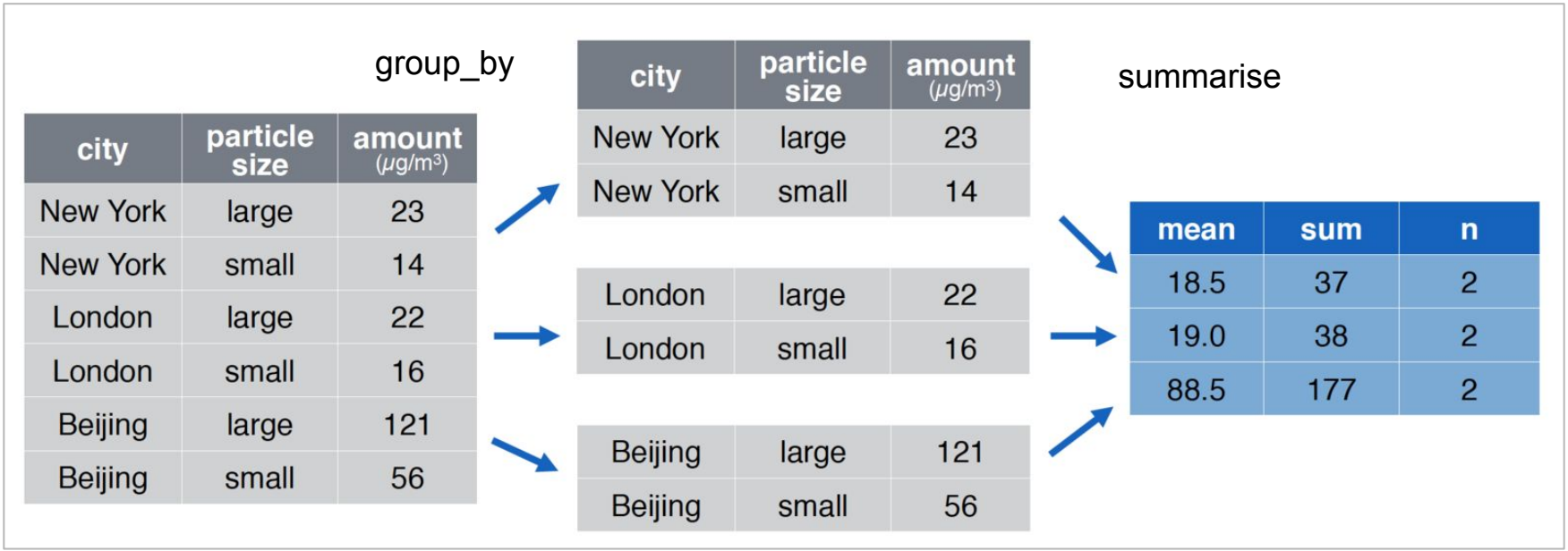


An interesting (but rare) aside

```r
# Make a data.frame
students <- data.frame(
  names=c('Mason', 'Tabi', 'Bryce'),
  math_exam1 = c(91, 82, 93),
  math_exam2 = c(88, 79, 77),
  spanish_exam1 = c(79, 88, 92),
  spanish_exam2 = c(99, 92, 92)
)

# Calculate summary stats
summarise(students,
          mean_math1 = mean(math_exam1),
          mean_math2 = mean(math_exam2),
          mean_math_scores=mean((math_exam1 + math_exam2) / 2)
)
```

Summarize: calculate one value from a set of values




```
# Group the pollution data.frame by city for comparison
pollution <- group_by(pollution, city) %>%
             summarise(mean = mean(amount), sum = sum(amount), n = n())
```

Group-by, then Summarize

module 9 exercise-5

Joins

	faa	name	lat	lon	alt	tz	dst
1	04G	Lansdowne Airport	41.130...	-80.61958	1044	-5	A
2	06A	Moton Field Municipal Airport	32.460...	-85.68003	264	-5	A
3	06C	Schaumburg Regional	41.989...	-88.10124	801	-6	A
4	06N	Randall Airport	41.431...	-74.39156	523	-5	A
5	09J	Jekyll Island Airport	31.074...	-81.42778	11	-4	A
6	0A9	Elizabethton Municipal Airport	36.371...	-82.17342	1593	-4	A
7	0G6	Williams County Airport	41.467...	-84.50678	730	-5	A
8	0G7	Finger Lakes Regional Airport	42.883...	-76.78123	492	-5	A
9	0P2	Shoestring Aviation Airfield	39.794...	-76.64719	1000	-5	U

Why is this airport information stored in a separate dataframe (`airports`)?

Joins

Allow you to combine **columns** from multiple data sources

Observations in each data source will have **identifying information** (1+ columns)

Foundation of working with **relational databases**

Multiple types of joins (link)

Left-Join Syntax

Join two data frames by shared identifier(s)

Return all rows in X, and all columns for x and y

```
# Join x and y by 'identifier'
joined <- left_join(x, y, by='identifier')
```

Documentation, cheatsheet

module 9 exercise-6

Upcoming...

By Tuesday: Be confident with **module 9**

Due Tuesday, 10/25 (***before class***): a4-data-wrangling