Econ 106

Lecture 4
Fall 2024

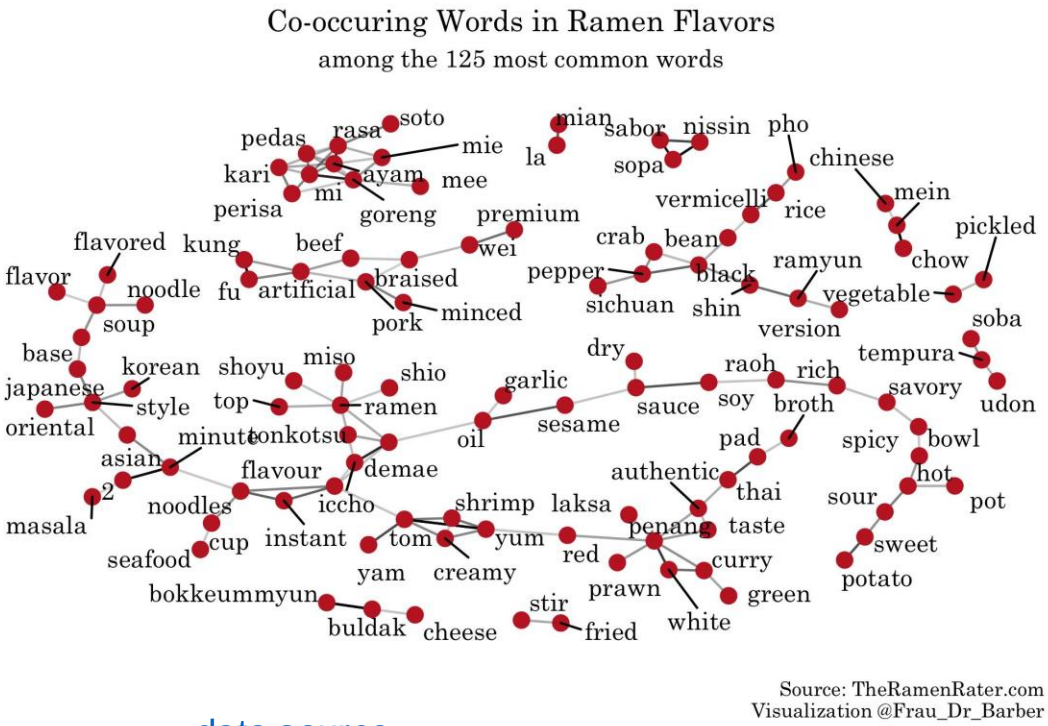
Large part of these slides are adapted from Nick Hagerty at Montana State University and
Introduction to Data Science by Rafael A. Irizarry, used under CC BY-NC-SA 4.0, and "Data Science for Economists" by Grant R. McDermott, used under the MIT License.

Reminders

- Lab 1 will be posted Sunday, due one week later
  - one-page writeup with tables
  - R script

https://pollev.com/vsovero

#tidytuesday



[data source]

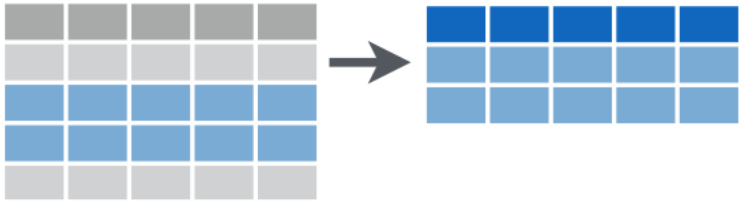
[code to generate graph]

Outline

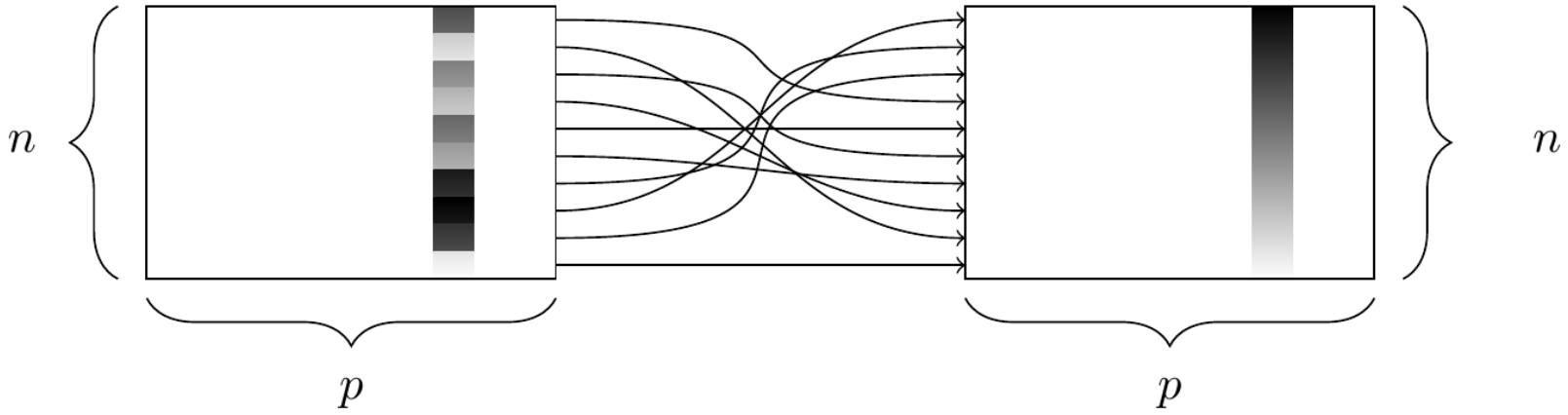
- summary tables
- frequency tables

filter()

Subset Observations (Rows)



arrange()



select()

Subset Variables (Columns)

mutate()

Make New Variables

Recap: key dplyr verbs

There are five key dplyr verbs that you need to learn.

1. filter: Filter (i.e. subset) rows based on their values.
2. arrange: Arrange (i.e. reorder) rows based on their values.
3. select: Select (i.e. subset) columns by their names:
4. mutate: Create new columns.
5. summarize: Collapse multiple rows into a single summary value

https://pollev.com/vsovero

Summarizing Variables

- How can we summarize/describe variables?
  - quantitative data
  - categorical data

Useful functions for quantitative variables

- Center: mean(), median()
- Spread: sd(), var()
- Range: min(), max(),
- Position: first(), last(),
- Total: sum()

Creating a summary table

Creating a summary table

- **summarize**()
- **Arguments:** a specific summary statistic
  - e.g.: **mean**(), **min**(), **max**(), **median**(), **sum**()
- **Output**: a table with the calculated summary statistic

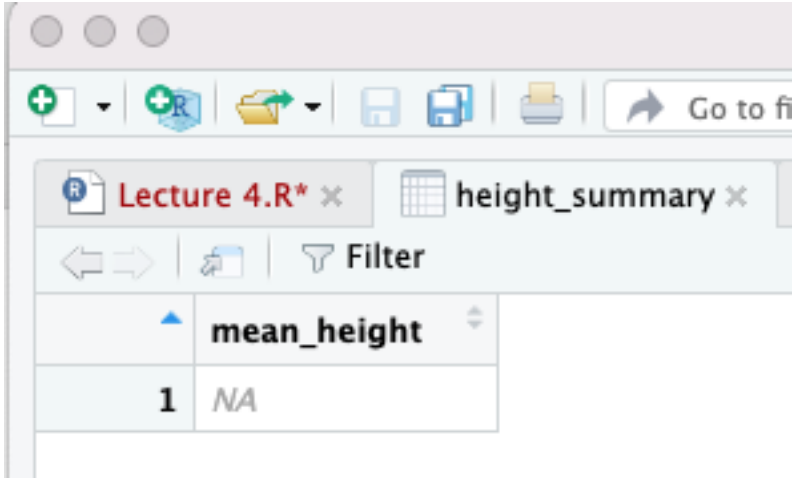
Creating a summary table

- the summarize function is similar to the mutate function
- you have to provide a name (mean_height) for where the new information will be stored

```
height_summary <- starwars %>%
                  summarize(mean_height = mean(height))
```

What happened?

```
height_summary <- starwars %>%
                summarize(mean_height = mean(height))
```



Missing Data is What Happened

- R can’t compute the mean of a numeric vector if there are any missing values (NA)
- We will have to tell it to ignore the rows that have missing values

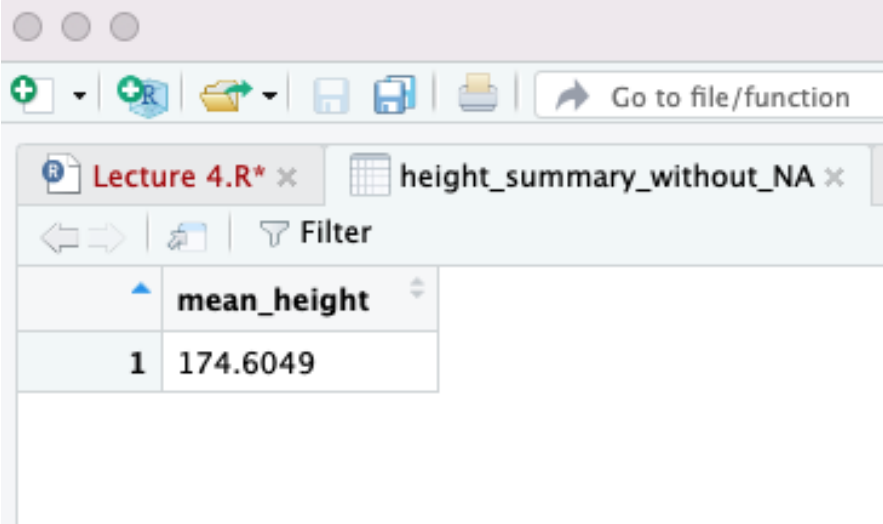
Lecture 4.R* × | height_summary × | starwars ×

Filter

	name	height	mass	hair_color
75	Wat Tambor	193	48.0	none
76	San Hill	191	NA	none
77	Shaak Ti	178	57.0	none
78	Grievous	216	159.0	none
79	Tarfful	234	136.0	brown
80	Raymus Antilles	188	79.0	brown
81	Sly Moore	178	48.0	none
82	Tion Medon	206	80.0	none
83	Finn	NA	NA	black
84	Rey	NA	NA	brown
85	Poe Dameron	NA	NA	brown
86	BB8	NA	NA	none
87	Captain Phasma	NA	NA	none

Fixed With an Additional Argument

```
height_summary_without_NA <- starwars %>%
        summarize(mean_height = mean(height, na.rm = TRUE ))
```



Base R vs. Dplyr

- We could have also used Base R for this calculation
- The code is actually a bit simpler
- However, it's going to store the information as a vector

```
height_summary_without_NA <- mean(starwars$height, na.rm = TRUE)
```

```
height_summary_without_NA <- starwars %>%
                summarize(mean_height =  mean(height, na.rm = TRUE ))
```

Creating a less silly summary table

- We can summarize more than one variable
- Add a comma, then write out the next statistic

```
ht_and_wt_summary <- starwars %>%
                    summarize(mean_height = mean(height , na.rm=TRUE),
                              mean_weight = mean(mass , na.rm=TRUE)
                              )
```

Exercise

- Calculate median and mean mass for characters in Naboo.
- What does this suggest about the distribution of mass?

https://pollev.com/vsovero

Summarizing Categorical Variables

- Econ 101 recap: we summarize/describe categorical variables with frequency tables

wage	Score	Rating	Gender	educ
11.42	3	average	male	16
3.91	3	average	female	12
8.76	3	average	male	16
7.69	4	above	male	16
5	3	average	female	16
3.89	3	average	female	12
3.45	5	above	female	12
4.03	4	above	male	16
5.14	2	below	male	17
3	3	average	male	16

Value	Count
below	1
average	6
high	3

Create a frequency table

- **count ()**
- **Arguments:** a categorical variable

- **Output**: a table with the frequency of each value of the categorical variable

Species Frequency Table

```
species_frequency<- starwars %>%
                    count(species)
```

	species	n
1	Aleena	1
2	Besalisk	1
3	Cerean	1
4	Chagrian	1
5	Clawdite	1
6	Droid	6
7	Dug	1
8	Ewok	1
9	Geonosian	1
10	Gungan	3
11	Human	35
12	Hutt	1

Species Frequency Table (ordered by frequency)

```
species_count_ordered<- starwars %>%
                        count(species) %>%
                        arrange(desc( n))
```

Lecture 4.R* × species_count_ordered ×

Filter

	species	n
1	Human	35
2	Droid	6
3	NA	4
4	Gungan	3
5	Kaminoan	2
6	Mirialan	2
7	Twi'lek	2
8	Wookiee	2
9	Zabrak	2
10	Aleena	1
11	Besalisk	1
12	Cerean	1
13	Chagrian	1

Species Frequency Table (remove NA)

```
species_count_remove_NA<- starwars %>%
                count(species) %>%
                 arrange(desc( n)) %>%
                 filter(!is.na (species))
```

	species	n
1	Human	35
2	Droid	6
3	Gungan	3
4	Kaminoan	2
5	Mirialan	2
6	Twi'lek	2
7	Wookiee	2
8	Zabrak	2
9	Aleena	1
10	Besalisk	1
11	Cerean	1
12	Chagrian	1
13	Clawdite	1
14	Dug	1
15	Ewok	1

Showing 1 to 15 of 37 entries, 2 total columns

Shortening our tables with **slice**()

- **slice_head**(n) select the first n rows
- **slice_tail**(n) select the last n rows.
- **slice_sample**(n) randomly select n rows.
- **slice_min**(x, n) select n rows with the smallest values of variable x
- **slice_max**(x, n) select n rows with the largest values of variable x

Top 10 List of Most Frequent Species

```
top_ten_species<- starwars %>%
                  filter(!is.na (species)) %>%
                 count(species) %>%
                  arrange(desc( n)) %>%
                 slice_head(n=10)
```

Exercise

- filter out missing values of eye color
- order the eye colors in a frequency table from highest to lowest frequency
- report the top ten eye colors by frequency

Frequency Table (two variables)

```
species_gender_frequency<- starwars %>%
                count(species, gender)
```

	species	gender	n
1	Aleena	masculine	1
2	Besalisk	masculine	1
3	Cerean	masculine	1
4	Chagrian	masculine	1
5	Clawdite	feminine	1
6	Droid	feminine	1
7	Droid	masculine	5
8	Dug	masculine	1
9	Ewok	masculine	1
10	Geonosian	masculine	1
11	Gungan	masculine	3

Recap so far: How to Summarize/Describe Variables

- Quantitative variables:
  - mean, median, etc.
  - use **summarize**()

- Categorical variables:
  - frequencies (use **count**())

group_by()

Grouped summary table

- **group_by**()
- **Arguments**: set of variables
- **Output**: a table with the calculated summary statistic for each combination of unique values in the variables inside **group_by**()

Grouped summary table

```
weight_by_sex <- starwars %>%
                 group_by( sex) %>%
                  summarize(mean_wt = mean(mass, na.rm = TRUE))
```

Output

	sex	mean_wt
1	female	54.68889
2	hermaphroditic	1358.00000
3	male	81.00455
4	none	69.75000
5	NA	48.00000

Grouped summary table (more than one group)

```
weight_by_gender_sex <- starwars %>%
                        group_by( gender, sex) %>%
                         summarize(mean_wt = mean(mass, na.rm = TRUE))
```

Output

Lecture 4.R* × weight_by_gender_sex ×

Filter

	gender	sex	mean_wt
1	feminine	female	54.68889
2	feminine	none	NaN
3	masculine	hermaphroditic	1358.00000
4	masculine	male	80.21905
5	masculine	none	69.75000
6	NA	NA	81.00000

what happened here?

Using group_by() to Create a Frequency Table

```
characters_by_homeworld<- starwars %>%
                          group_by(homeworld) %>%
                           summarize(count=n())
```

characters_by_homeworld

Filter

	homeworld	count
1	Alderaan	3
2	Aleen Minor	1
3	Bespin	1
4	Bestine IV	1
5	Cato Neimoidia	1
6	Cerea	1
7	Champala	1
8	Chandrila	1
9	Concord Dawn	1
10	Corellia	2
11	Coruscant	3
12	Dathomir	1

Exercise

- Let’s load some data from a tidytuesday challenge
- we will use the **read_csv**() function from the tidyverse package
- it can read in csv files from your computer or from a URL

tidytuesday data

Data Dictionary

jobs_gender.csv

Data Dictionary

variable	class	description
year	integer	Year
occupation	character	Specific job/career
major_category	character	Broad category of occupation
minor_category	character	Fine category of occupation
total_workers	double	Total estimated full-time workers > 16 years old
workers_male	double	Estimated MALE full-time workers > 16 years old
workers_female	double	Estimated FEMALE full-time workers > 16 years old
percent_female	double	The percent of females for specific occupation
total_earnings	double	Total estimated median earnings for full-time workers > 16 years old
total_earnings_male	double	Estimated MALE median earnings for full-time workers > 16 years old
total_earnings_female	double	Estimated FEMALE median earnings for full-time workers > 16 years old
wage_percent_of_male	double	Female wages as percent of male wages - NA for occupations with small sample size