An Introduction to Data Frames

Understanding the organization power of tabular data for data analysis

---

name	height	mass	hair_color	skin_color	eye_color	birth_year	gender	homeworld	species
Luke Skywalker	172	77	blond	fair	blue	19	male	Tatooine	Human
C-3PO	167	75	NA	gold	yellow	112	NA	Tatooine	Droid
R2-D2	96	32	NA	white, blue	red	33	NA	Naboo	Droid
Darth Vader	202	136	none	white	yellow	41.9	male	Tatooine	Human
Leia Organa	150	49	brown	light	brown	19	female	Alderaan	Human
Owen Lars	178	120	brown, grey	light	blue	52	male	Tatooine	Human
Beru Whitesun lars	165	75	brown	light	blue	47	female	Tatooine	Human
R5-D4	97	32	NA	white, red	red	NA	NA	Tatooine	Droid
Biggs Darklighter	183	84	black	light	brown	24	male	Tatooine	Human

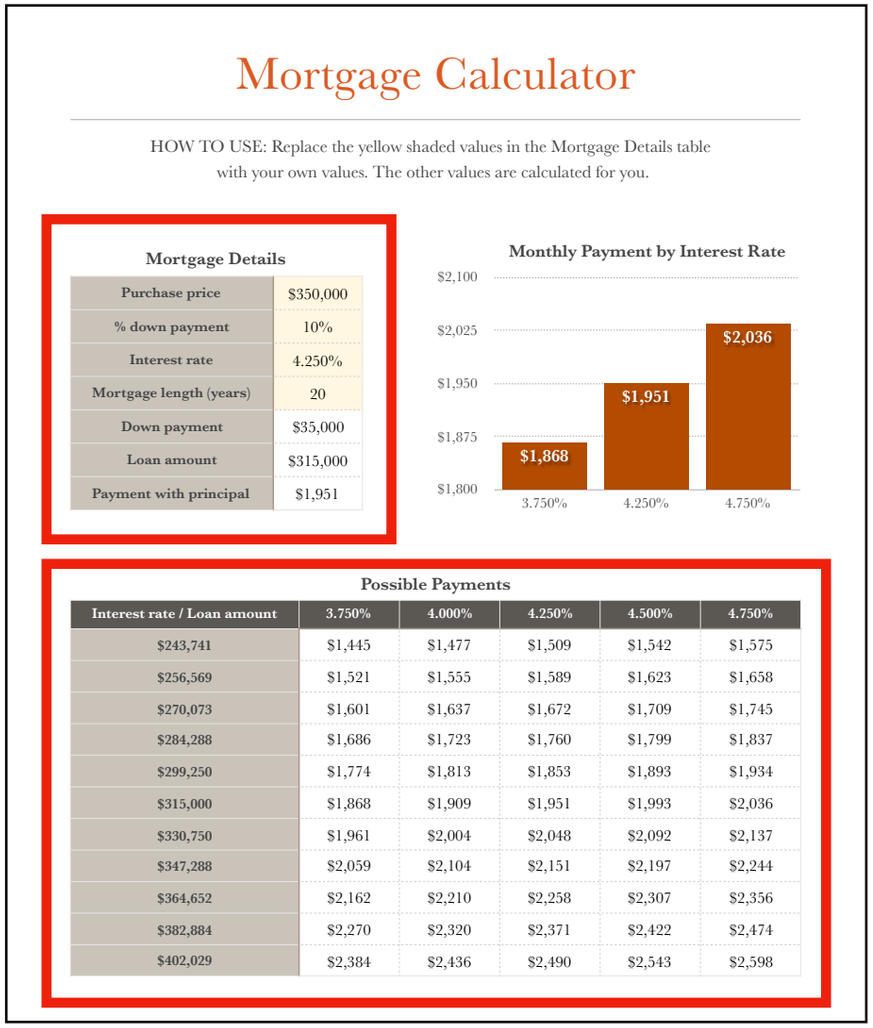
Ryan Benz

Data Tables are Everywhere!

in formatted text files (e.g. csv files)

```
"Sepal.Length","Sepal.Width","Petal.Length","Petal.Width","Species"
5.1,3.5,1.4,0.2,"setosa"
4.9,3,1.4,0.2,"setosa"
4.7,3.2,1.3,0.2,"setosa"
4.6,3.1,1.5,0.2,"setosa"
5,3.6,1.4,0.2,"setosa"
5.4,3.9,1.7,0.4,"setosa"
4.6,3.4,1.4,0.3,"setosa"
5,3.4,1.5,0.2,"setosa"
```

inside of spreadsheets



Mortgage Calculator

HOW TO USE: Replace the yellow shaded values in the Mortgage Details table with your own values. The other values are calculated for you.

Mortgage Details

Purchase price	$350,000
% down payment	10%
Interest rate	4.250%
Mortgage length (years)	20
Down payment	$35,000
Loan amount	$315,000
Payment with principal	$1,951

Possible Payments

Interest rate / Loan amount	3.750%	4.000%	4.250%	4.500%	4.750%
$243,741	$1,445	$1,477	$1,509	$1,542	$1,575
$256,569	$1,521	$1,555	$1,589	$1,623	$1,658
$270,073	$1,601	$1,637	$1,672	$1,709	$1,745
$284,288	$1,686	$1,723	$1,760	$1,799	$1,837
$299,250	$1,774	$1,813	$1,853	$1,893	$1,934
$315,000	$1,868	$1,909	$1,951	$1,993	$2,036
$330,750	$1,961	$2,004	$2,048	$2,092	$2,137
$347,288	$2,059	$2,104	$2,151	$2,197	$2,244
$364,652	$2,162	$2,210	$2,258	$2,307	$2,356
$382,884	$2,270	$2,320	$2,371	$2,422	$2,474
$402,029	$2,384	$2,436	$2,490	$2,543	$2,598

on television

and lot of other places too…

Tables Provide Important Organizational Structure

You can’t do much with a bunch of random data points

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa

Interpretable by both humans & computers

If we’re careful about how we store data in tables, they can make the *entire* process of working with data **MUCH** easier

The Data Frame: R's Most Important Data Structure

- R's main representation of **2D tables**
- **Organize the data** we work with
- Are a **fundamental part of R**
- **Influence how the language is used** and how it's developed over the last 20 years, part of why **R is great for data**
- **Fundamental to the `tidyverse`** collection of R packages and how they operate

The Data Frame:
R's Most Important Data Structure

Here's a data frame you may have seen before?

The infamous iris dataset

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
```

… (140 more rows)

The Data Frame:
R's Most Important Data Structure

Rows and columns (typically) have meaning

columns *measurements of the flowers*

rows → *observations of individual flowers*

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
```

The Data Frame:
R's Most Important Data Structure

Columns have names

columns → *"first" row gives the name of the columns*

called the ***header***

```
> iris
  Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1          5.1         3.5          1.4         0.2     setosa
2          4.9         3.0          1.4         0.2     setosa
3          4.7         3.2          1.3         0.2     setosa
4          4.6         3.1          1.5         0.2     setosa
5          5.0         3.6          1.4         0.2     setosa
6          5.4         3.9          1.7         0.4     setosa
7          4.6         3.4          1.4         0.3     setosa
8          5.0         3.4          1.5         0.2     setosa
9          4.4         2.9          1.4         0.2     setosa
10         4.9         3.1          1.5         0.1     setosa
```

The Data Frame:
R's Most Important Data Structure

Can contain different types of data

numbers

words

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
```

Quick Live Demo

Investigating the properties of a data frame in R

R Function	What it does
head, tail	show the first or last few rows of data
nrow, ncol, dim	Get the number of rows, columns, or both
names	gives the names of the columns
View	opens a nice table view of the data
$	get a single column from the data
math_function(*dataframe$column*)	compute a mathematical summary on a column in the data

Data Frames Become Really Powerful When They Are Tidy!

Tidy data: data that has been organized into a table in a consistent way that make working with data easier

Each row is an **observation**
an individual iris flower

Each column is a **variable**
lengths, widths and species

Every **value** lives in its own cell
the actual measurements

Each table has a specific **topic**
iris flower measurements

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
```

Real world data is often NOT tidy

An important part of working with data is learning how to make data tidy (i.e. ready for analysis)

Exercise 1

Reconstruct the Data Table

— **Header row**

} **Data rows**

Place these cells in their proper place in the table above

yellow	strawberry	FALSE
Need to Peel?	TRUE	Color
Fruit Name	red	banana

What is the topic of this table?

What are the observations?

What are the variables?

Exercise 1

Reconstruct the Data Table

Fruit	Color	Need to Peel?
strawberry	red	FALSE
banana	yellow	TRUE

— Header row

} Data rows

Place these cells in their proper place in the table above

yellow	strawberry	FALSE
Need to Peel?	TRUE	Color
Fruit Name	red	banana

What is the topic of this table?

data about fruits

What are the observations?

individual fruits

What are the variables?

Fruit Name, Color, Need to Peel?

Data Manipulation is a Key Part of Data Analysis

When we do data analysis, we are essentially just performing lots of different manipulations on the data to further understand it

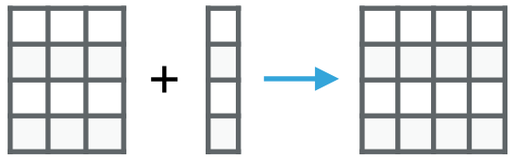
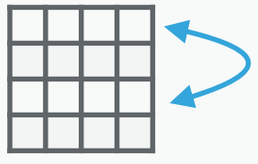
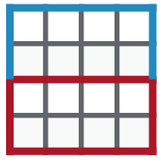
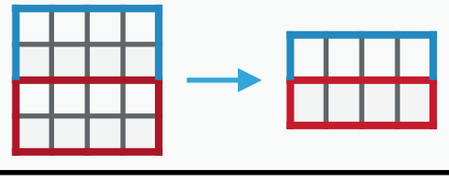
- Computing new values from the data
- Combining different data sets
- Investigating sub-groups in the data
- Computing data summarizations
- Building models
- Making plots and visualizations

The organizational structure of data frames facilitates these manipulations

Tidy data frames in particular help to make these manipulations easy to understand and perform

Data Manipulations with Tidy Data

A few fundamental data manipulations encompass most data analysis tasks

Manipulation		Why?
Add a column		compute new data from existing variables, join new data
Pick specific columns		focus in on specific variables
Subset to specific rows		focus on a specific sub-group in the data
Reorder/sort rows		understand the order of the data, find top/bottom observations
Group subsets		want to analyze sub-groups in your data
Summarize rows		compute summary values across multiple rows, useful with grouping

These manipulations typically operate on columns (i.e. the variables)

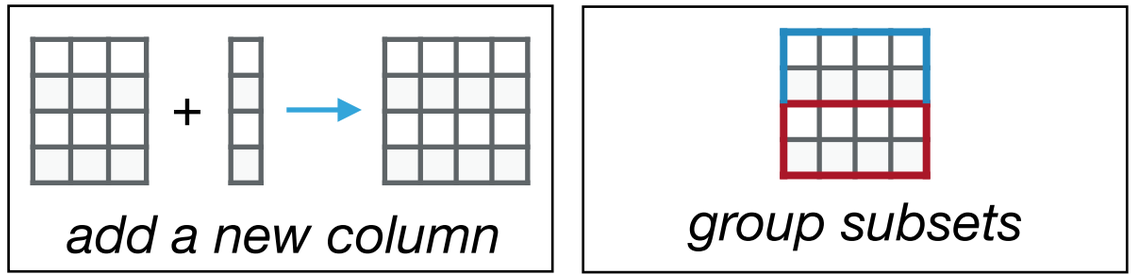
Exercise 2

Choose the Right Manipulation

Task

- You have too many variables and want to simplify the data
- You want to compute a new variable that is the sum of two existing variables
- You want to get rid of the observations with missing data
- You want to compute the mean test score for each student

Manipulation Choices



Exercise 2

Choose the Right Manipulation

Task

- You have too many variables and want to simplify the data

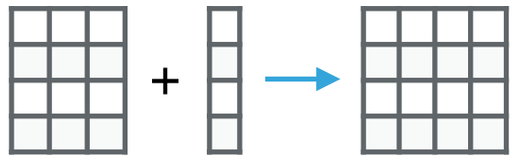
- You want to compute a new variable that is the sum of two existing variables

- You want to get rid of the observations with missing data

- You want to compute the mean test score for each student

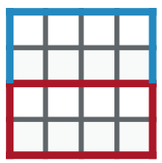
Manipulations

select specific cols



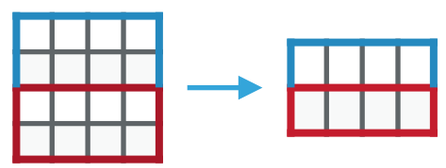
add a new column

filter to specific rows



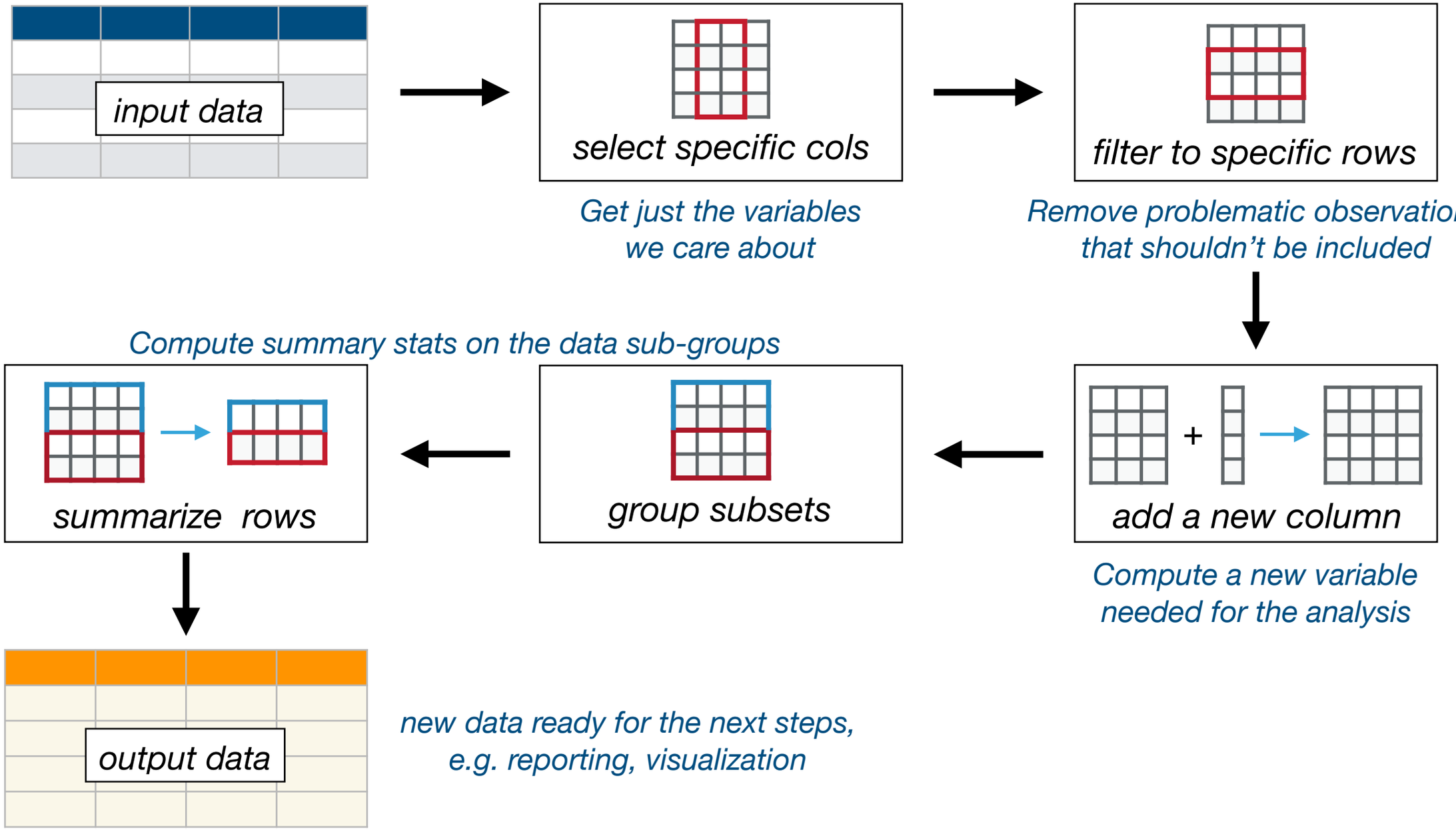
group subsets

+



summarize rows

Manipulations can be Combined to do More Complex Things



Wrapping It All Up

- Data tables are a key organizational structure that facilitate the data analysis process

- R uses Data Frames to hold (most) tabular data

- Tidy data frames are particularly important
  - Rows represent observations (the things we are measuring)
  - Columns represent variables (the specific attributes measured)
  - Each value is in its own cell (the actual measured values)

- A few fundamental data manipulations allow us to perform most data analysis tasks, powerful when combined together

Final Exercise

Here's a glimpse of the starwars data set from the dplyr package

name	height	mass	hair_color	skin_color	eye_color	birth_year	gender	homeworld	species
Luke Skywalker	172	77	blond	fair	blue	19	male	Tatooine	Human
C-3PO	167	75	NA	gold	yellow	112	NA	Tatooine	Droid
Darth Vader	202	136	none	white	yellow	41.9	male	Tatooine	Human
Chewbacca	228	112	brown	unknown	blue	200	male	Kashyyyk	Wookiee
Greedo	173	74	NA	green	black	44	male	Rodia	Rodian
Jabba Desilijic Tiure	175	1358	NA	green-tan, brown	orange	600	hermaphrodite	Nal Hutta	Hutt
Yoda	66	17	white	green	brown	896	male	NA	Yoda's species
Bossk	190	113	none	green	red	53	male	Trandosha	Trandoshan
Ackbar	180	83	none	brown mottle	orange	41	male	Mon Cala	Mon Calamari
Finis Valorum	170	NA	blond	fair	blue	91	male	Coruscant	Human

… many more rows

Final Exercise

Task: Compute the average *mass/height* for each *species*

name	height	mass	hair_color	skin_color	eye_color	birth_year	gender	homeworld	species
Luke Skywalker	172	77	blond	fair	blue	19	male	Tatooine	Human
C-3PO	167	75	NA	gold	yellow	112	NA	Tatooine	Droid
Darth Vader	202	136	none	white	yellow	41.9	male	Tatooine	Human
Chewbacca	228	112	brown	unknown	blue	200	male	Kashyyyk	Wookiee
Greedo	173	74	NA	green	black	44	male	Rodia	Rodian
Jabba Desilijic Tiure	175	1358	NA	green-tan, brown	orange	600	hermaphrodite	Nal Hutta	Hutt
Yoda	66	17	white	green	brown	896	male	NA	Yoda's species
Bossk	190	113	none	green	red	53	male	Trandosha	Trandoshan
Ackbar	180	83	none	brown mottle	orange	41	male	Mon Cala	Mon Calamari
Finis Valorum	170	NA	blond	fair	blue	91	male	Coruscant	Human

… many more rows

Final Exercise

Note: there are many species & missing values

name	height	mass	hair_color	skin_color	eye_color	birth_year	gender	homeworld	species
Luke Skywalker	172	77	blond	fair	blue	19	male	Tatooine	Human
C-3PO	167	75	NA	gold	yellow	112	NA	Tatooine	Droid
Darth Vader	202	136	none	white	yellow	41.9	male	Tatooine	Human
Chewbacca	228	112	brown	unknown	blue	200	male	Kashyyyk	Wookiee
Greedo	173	74	NA	green	black	44	male	Rodia	Rodian
Jabba Desilijic Tiure	175	1358	NA	green-tan, brown	orange	600	hermaphrodite	Nal Hutta	Hutt
Yoda	66	17	white	green	brown	896	male	NA	Yoda's species
Bossk	190	113	none	green	red	53	male	Trandosha	Trandoshan
Ackbar	180	83	none	brown mottle	orange	41	male	Mon Cala	Mon Calamari
Finis Valorum	170	NA	blond	fair	blue	91	male	Coruscant	Human

… many more rows also many more NA’s in both height & mass

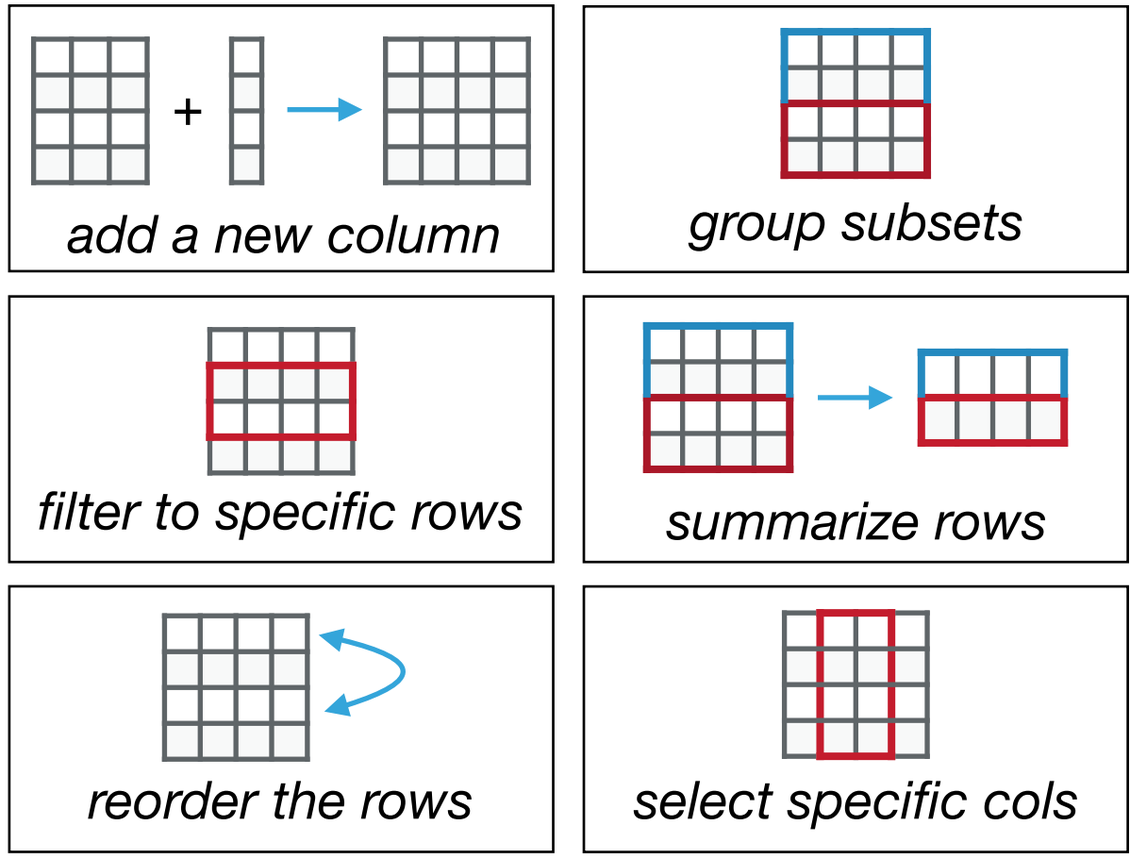
Final Exercise

Combine variable name / data manipulation pairs to create a pipeline to answer the question

Variables

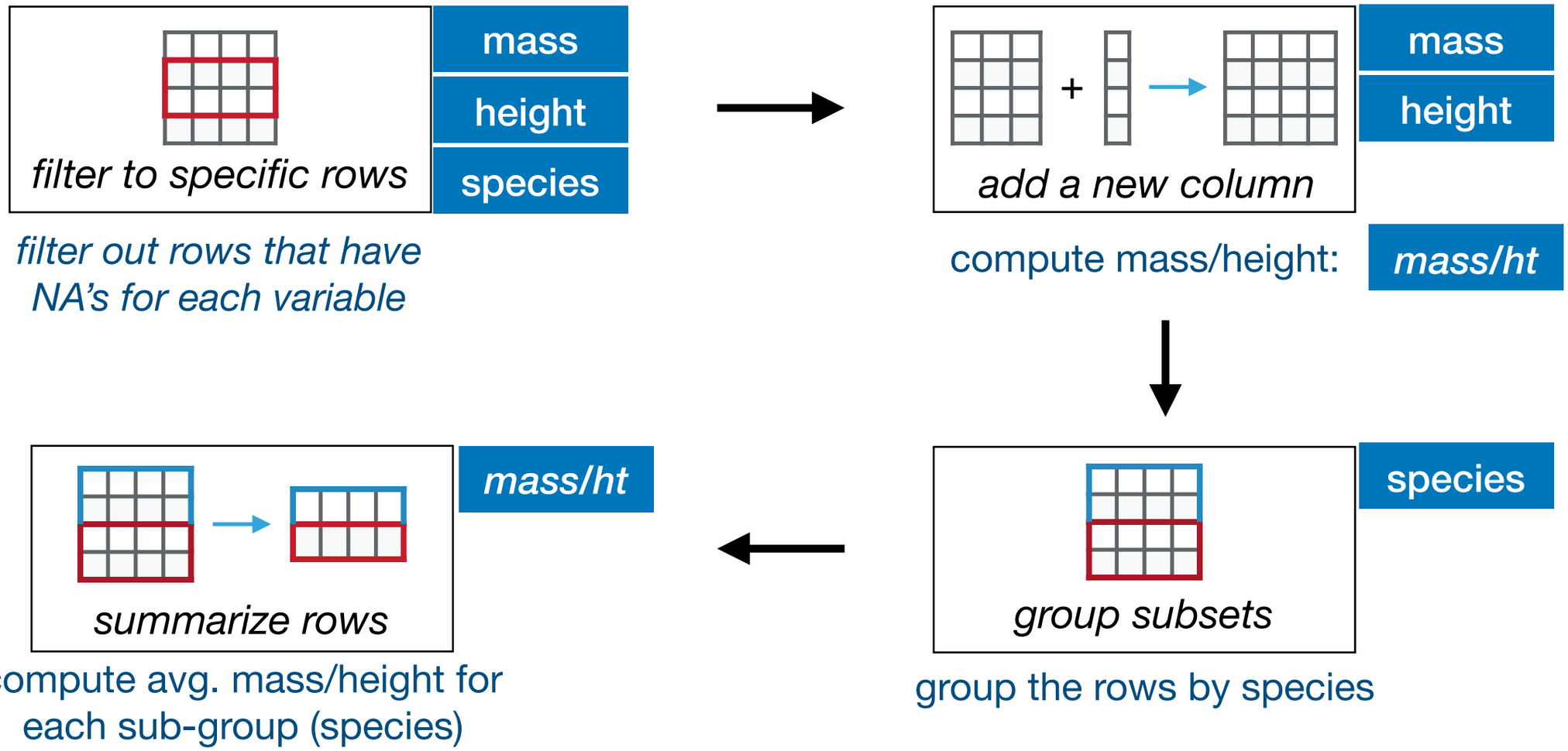
- **mass**
- **height**
- **species**
- ***mass/ht*** *this is a new derived column not in the original data*

Manipulations



Final Exercise

Task: Compute the average *mass/height* for each *species*



Other Exercises

Thinking About Variables

1. Choose a topic (any topic you'd like)

Examples: measuring student performance, traffic congestion, restaurant quality

2. Think of 5 variables you could measure to study the topic

Topic:

Variables

1.
2.
3.
4.
5.

Thinking About Variables

1. Choose a topic (any topic you'd like)

Examples: measuring student performance, traffic congestion, restaurant quality

2. Think of 5 variables you could measure to study the topic

Topic: measuring student performance

Variables

1. grade
2. age
3. class size
4. exam score(s)
5. school location

Thinking About Variables

1. Choose a topic (any topic you'd like)

Examples: measuring student performance, traffic congestion, restaurant quality

2. Think of 5 variables you could measure to study the topic

Topic: traffic congestion

Variables

1. location
2. time of day
3. date
4. number of accidents in the area
5. number of cars on the road

Thinking About Variables

1. Choose a topic (any topic you'd like)

Examples: measuring student performance, traffic congestion, restaurant quality

2. Think of 5 variables you could measure to study the topic

Topic: restaurant quality

Variables

1. location
2. type of food
3. how long in business
4. number of health violations
5. Yelp rating