Data Analysis Using R and RStudio

Ben Ward, Audit Principal of Data Analytics
California State Auditor's Office

NSAA IT Conference

Madison, Wisconsin| September 28, 2021

Agenda

Role of my office's Data Analytics Team

Using R for data analytics

- **R and RStudio**
- **Data Analytics Process**
- **Resources**

Data Analytics Team

- Consists of 11 team members
- Works in conjunction with audit teams to conduct audits requested and approved by the legislature or audits mandated by law
- Analyzes large sets of data to address audit objectives

R and RStudio

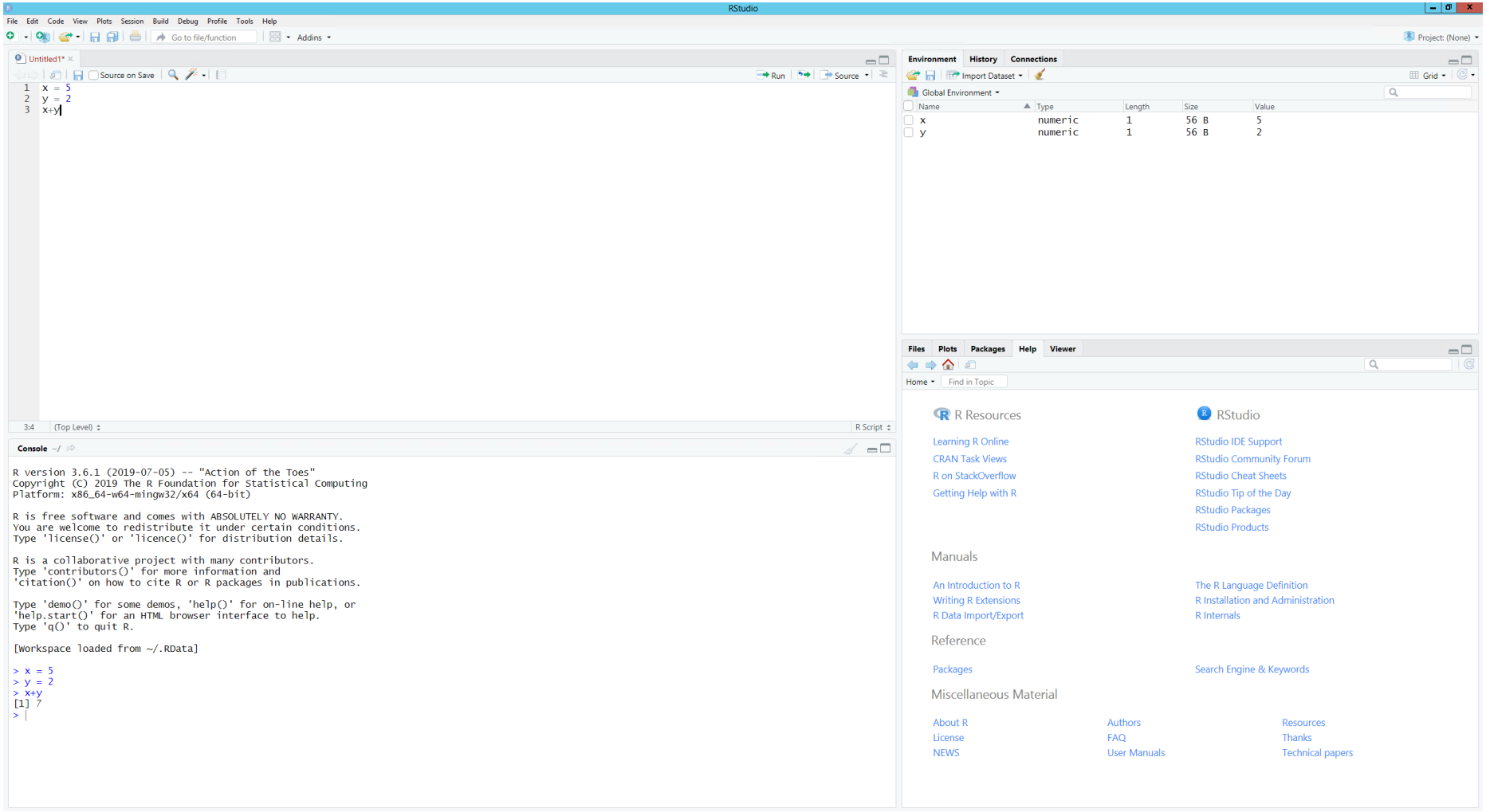
What's the difference?

- R is a statistical programming language
- RStudio is a user-friendly interface for R
- Both are open source (they are free!)
- Both have lots of online resources

R Packages

- Packages are similar to apps on a phone
- R users build packages to extend R's functionality
- RStudio develops and maintains a number of widely used packages (tidyverse packages)

RStudio Interface

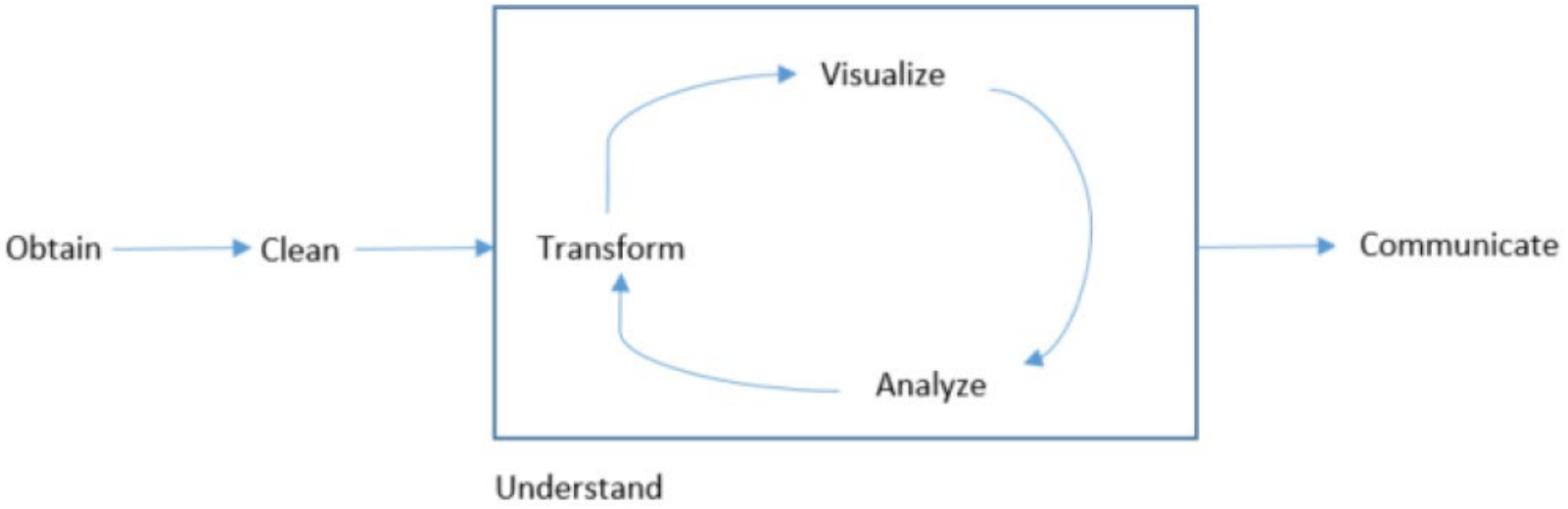


Data Analytics Process

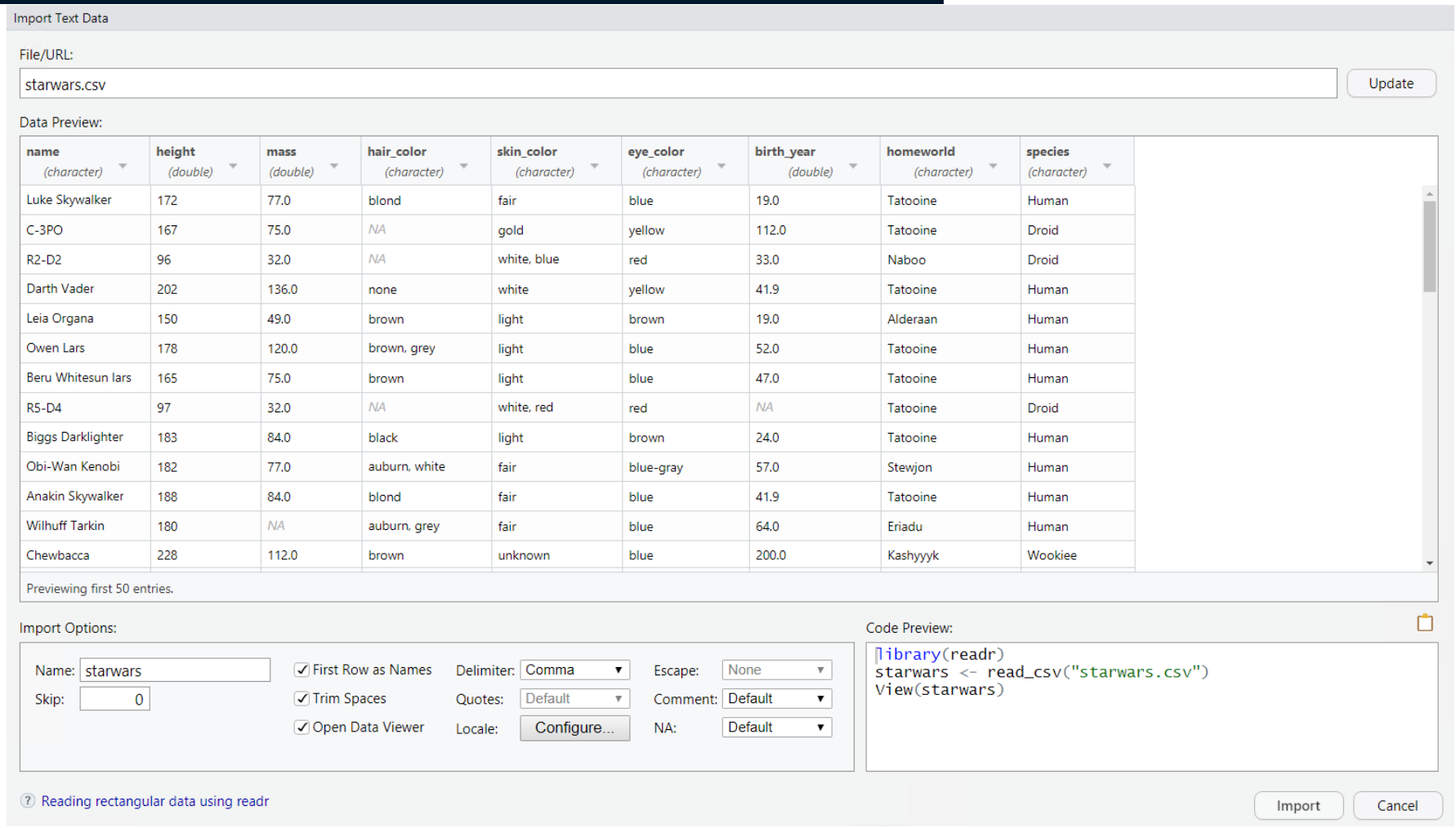
"Data analytics is the tradecraft of distilling hidden insights and enabling people to rapidly act on those insights."

- Taka Ariga, GAO's Chief Data Scientist

Data Analytics Process



Obtaining Data



Import Text Data

File/URL:

starwars.csv

Update

Data Preview:

name (character)	height (double)	mass (double)	hair_color (character)	skin_color (character)	eye_color (character)	birth_year (double)	homeworld (character)	species (character)
Luke Skywalker	172	77.0	blond	fair	blue	19.0	Tatooine	Human
C-3PO	167	75.0	NA	gold	yellow	112.0	Tatooine	Droid
R2-D2	96	32.0	NA	white, blue	red	33.0	Naboo	Droid
Darth Vader	202	136.0	none	white	yellow	41.9	Tatooine	Human
Leia Organa	150	49.0	brown	light	brown	19.0	Alderaan	Human
Owen Lars	178	120.0	brown, grey	light	blue	52.0	Tatooine	Human
Beru Whitesun lars	165	75.0	brown	light	blue	47.0	Tatooine	Human
R5-D4	97	32.0	NA	white, red	red	NA	Tatooine	Droid
Biggs Darklighter	183	84.0	black	light	brown	24.0	Tatooine	Human
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray	57.0	Stewjon	Human
Anakin Skywalker	188	84.0	blond	fair	blue	41.9	Tatooine	Human
Wilhuff Tarkin	180	NA	auburn, grey	fair	blue	64.0	Eriadu	Human
Chewbacca	228	112.0	brown	unknown	blue	200.0	Kashyyyk	Wookiee

Previewing first 50 entries.

Import Options:

Name: starwars

Skip: 0

First Row as Names

Trim Spaces

Open Data Viewer

Delimiter: Comma

Quotes: Default

Locale: Configure...

Escape: None

Comment: Default

NA: Default

Code Preview:

```
library(readr)
starwars <- read_csv("starwars.csv")
View(starwars)
```

Reading rectangular data using readr

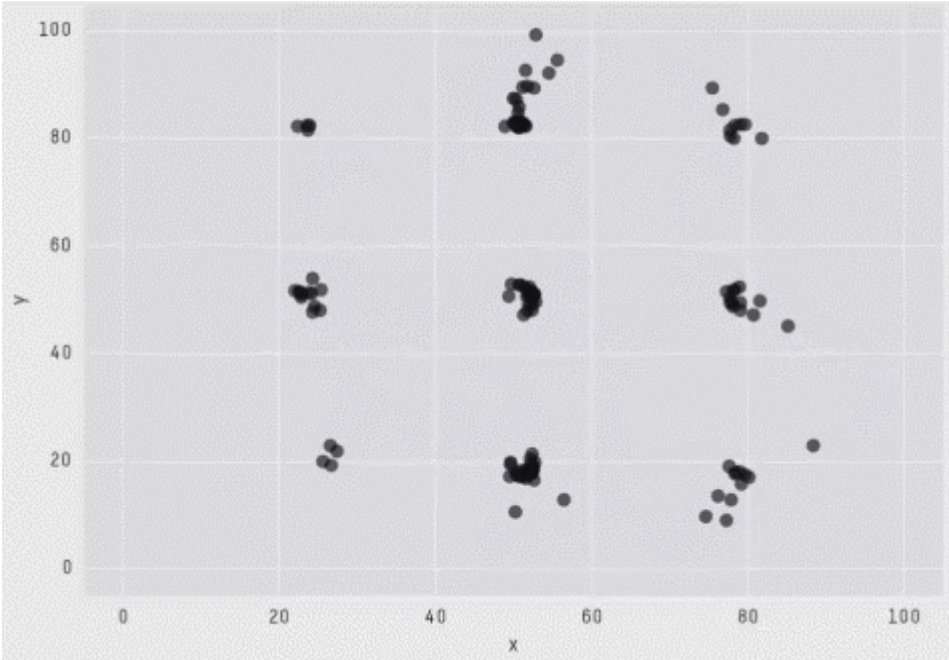
Import

Cancel

Understanding Data

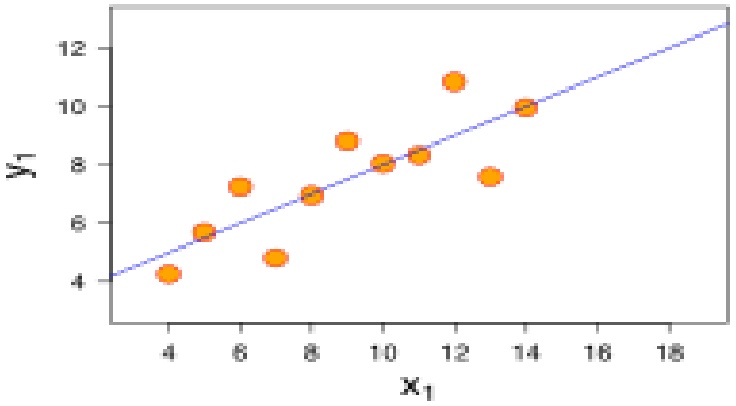
"The greatest value of a picture is when it forces us to notice what we never expected to see."

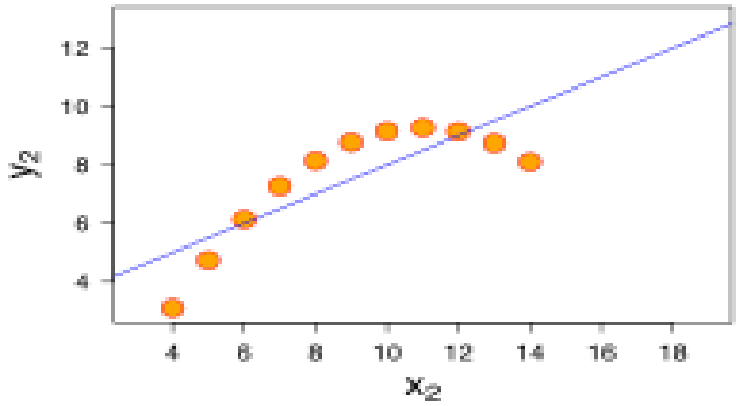
- John Tukey

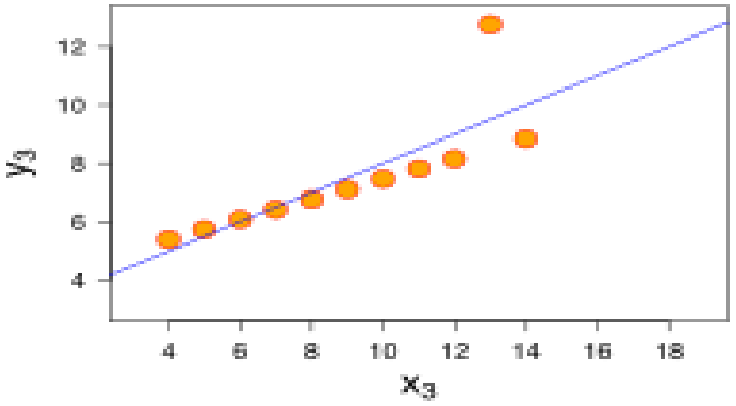


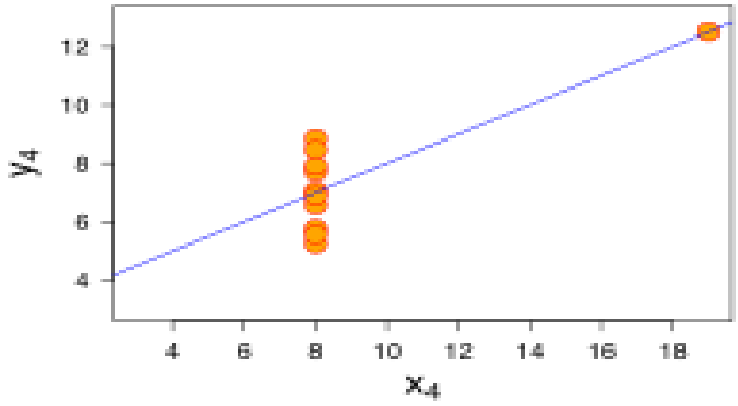
X Mean: 54.2665730
Y Mean: 47.8355311
X SD : 16.7693235
Y SD : 26.9303634
Corr. : -0.0610645

Anscombe's Quartet









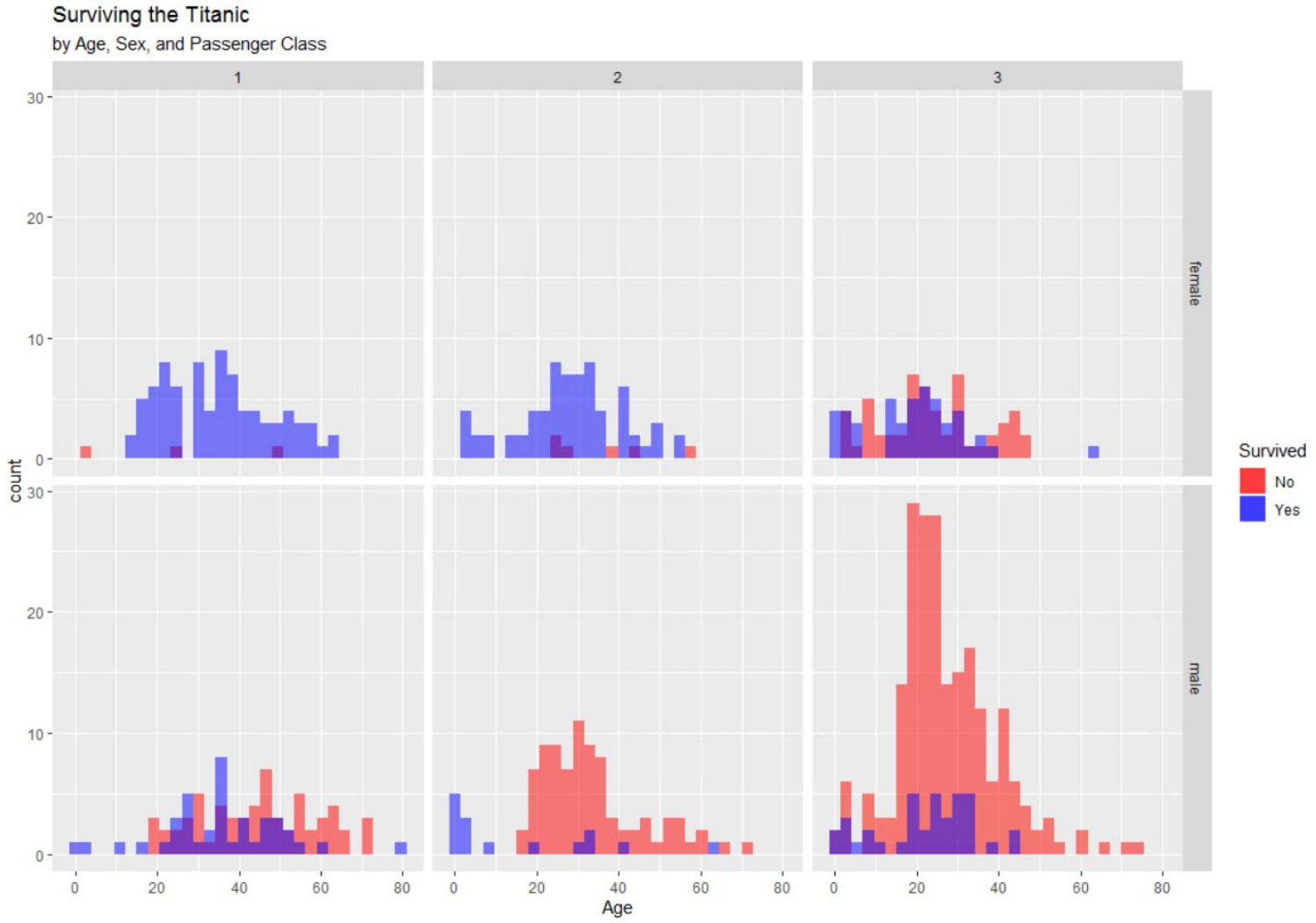
Graphing in R

- RStudio created the popular graphing package ggplot2

- Based on the Grammar of Graphics
  - Allows for layered approach to build graphics

  - Users define:
    - What data to use
    - Graph aesthetics
    - Graphical primitives

Example Graphic



Example Audit

State Bar Audit

- The average number of days the State Bar took to complete its investigation phase increased by 56 percent from 2015 through 2020, reaching 190 days in 2020.
  - Could this be caused by outliers?
- By graphing the data distributions, we found the percentage of cases that were in the investigation phase for more than one year steadily increased from 1 percent in 2015 to 11 percent in 2020.

Documenting Your Work

- Rmarkdown allows you to explain your analysis and document the results

- Based on Literate Programming
  - Code chunks interspersed with natural language to explain the code and results

Rmarkdown Example

Ben Ward

9/28/2021

Step 1

We then import the data and look at the first few rows.

```
starwars_file <- read_csv("starwars.csv")
```

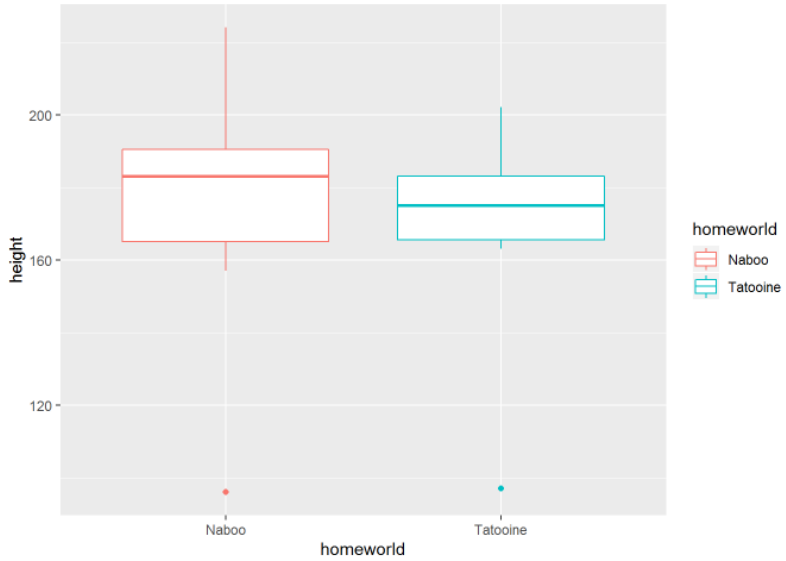
```
kable(head(starwars_file))
```

name	height	mass	hair_color	skin_color	eye_color	birth_year	homeworld	species
Luke Skywalker	172	77	blond	fair	blue	19.0	Tatooine	Human
C-3PO	167	75	NA	gold	yellow	112.0	Tatooine	Droid
R2-D2	96	32	NA	white, blue	red	33.0	Naboo	Droid
Darth Vader	202	136	none	white	yellow	41.9	Tatooine	Human
Leia Organa	150	49	brown	light	brown	19.0	Alderaan	Human
Owen Lars	178	120	brown, grey	light	blue	52.0	Tatooine	Human

Step 2

We then look at height by homeworld.

```
starwars_file %>%
  filter(homeworld %in% c("Naboo", "Tatooine")) %>%
  ggplot(aes(x=homeworld, y=height, color=homeworld)) +
  geom_boxplot()
```



Resources

Online Training

- Variety of free and inexpensive online courses, many taught by top universities
  - edX
  - Coursera
  - Data Camp

Cheatsheets

- RStudio has developed high quality cheatsheets for tidyverse packages
- Variety of topics, including:
  - Data import
  - Data transformation
  - Data visualization

https://www.rstudio.com/resources/cheatsheets/