

STAT 209

Data Computing and Visualization

May 25, 2021

Colin Reimer Dawson

Outline

“Data Science”

Intros

Some Terminology

Course Outline

Outline

"Data Science"

Intros

Some Terminology

Course Outline

Data is the new black

Some Cool Things you can do with data

Recommendation Systems

Frequently Bought Together



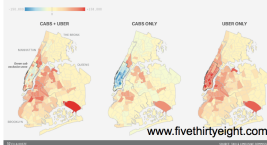
Customers Who Bought This Item Also Bought



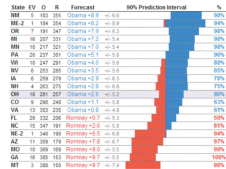
Data-Driven Journalism

Are Users Supplementing Or Replacing Cabs?

Change in number of Uber and taxi pickups by taxi zone, April-June 2014 versus April-June 2015

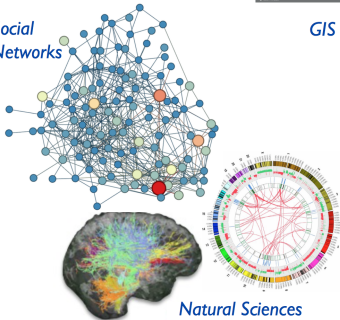


Competitive State Summary



Political Science

Social Networks

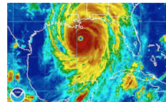


Natural Sciences

GIS / Development / Public Policy



Meteorology / Climate Science

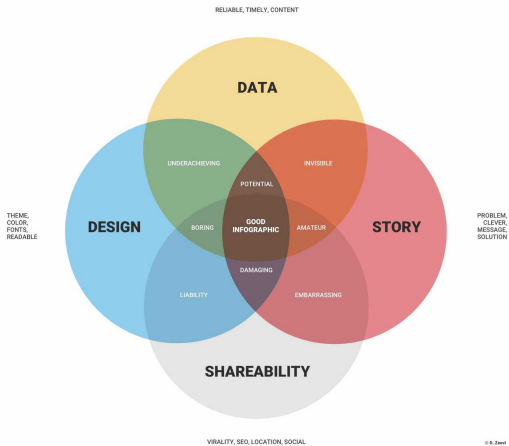


Sports

Finance



DATA VISUALIZATION



Outline

"Data Science"

Intros

Some Terminology

Course Outline

Brainstorm

What is the difference between "data" and "information"?

Outline

"Data Science"

Intros

Some Terminology

Course Outline

Cases

Cases When we collect data, we write down some measurements or characteristics of our **cases** — the individual “entities” that make up our dataset.

- The people in a survey or research study
- Plots of land in an agricultural experiment
- Days, in a weather dataset

Categorical vs. Quantitative Variables

For each case we record one or more **variables**. One of the most basic distinctions is between **categorical** (or “qualitative”) and **quantitative** data.

Categorical: “Qualitative” variable that divides cases into groups

Quantitative: Measures something on a scale; arithmetic makes sense

Data Frames

A standard form for a dataset is a grid, called a **data frame**, where each row is a *case*, and each column is a *variable*.

ID	Major	Height (in.)
1	Neuroscience	67
2	CS	71
	...	
21	Economics	64

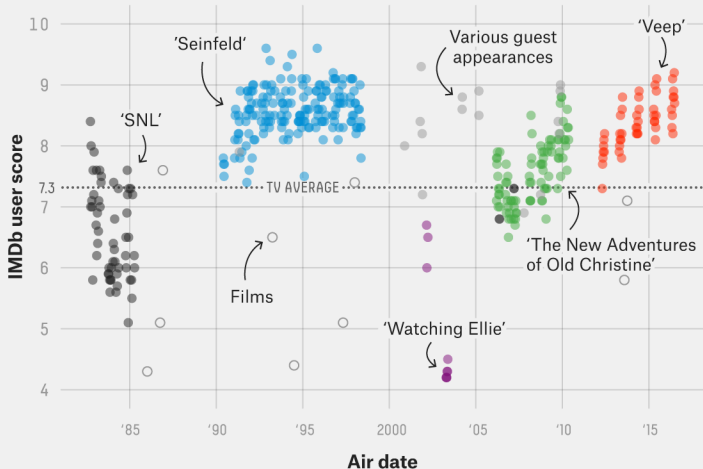
Deconstructing Visualizations

For each of the following visualizations:

1. What are the cases (think "rows" of a dataset)?
2. What variables are depicted (think "columns" of a dataset)?
3. What graphical element (position, color, etc.) is used to encode each variable?
4. What insights can you get from the visualization that would be tough to get just by looking at the raw data?

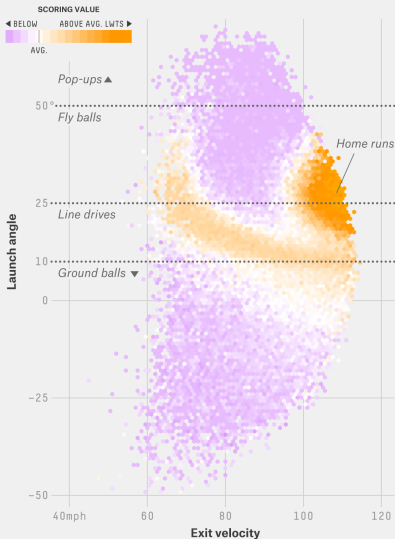
Julia Louis-Dreyfus is good at almost everything

IMDb ratings for appearances by Louis-Dreyfus



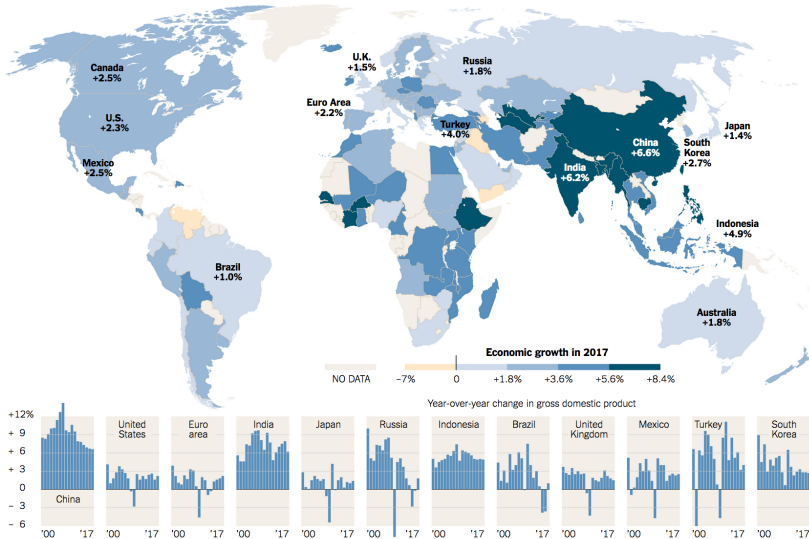
The sweet spot

Scoring value (LWTS) of batted balls based on launch angle and speed off the bat, 2015 MLB



Growth Across the Globe

For the first time since the financial crisis a decade ago, all of the world's major economies are growing.



Some figures are estimates

Source: The Conference Board; Bureau of Labor Statistics | By Karl Russell

Outline

"Data Science"

Intros

Some Terminology

Course Outline

Course Outline

- Part 0: Getting up and Running with R/RStudio (1 week)
- Part I: Basic Visualization (about 4 weeks)
- Part II: Data "Wrangling" (about 4 weeks)
- Part III: Large datasets and datasets with special structure (about 4 weeks)

On the Web

- Course Website: <http://colindawson.net/stat209>

On the Web

- Course Website: <http://colindawson.net/stat209>
 - Syllabus, schedule, homework, slides, code, etc., there

On the Web

- Course Website: <http://colindawson.net/stat209>
 - Syllabus, schedule, homework, slides, code, etc., there
- Slack: stat209s2021.slack.com. Recommended:
Download the desktop and/or mobile app

On the Web

- Course Website: <http://colindawson.net/stat209>
 - Syllabus, schedule, homework, slides, code, etc., there
- Slack: stat209s2021.slack.com. Recommended:
Download the desktop and/or mobile app
 - Convenient one-stop place for all course-related electronic communication

On the Web

- Course Website: <http://colindawson.net/stat209>
 - Syllabus, schedule, homework, slides, code, etc., there
- Slack: stat209s2021.slack.com. Recommended:
Download the desktop and/or mobile app
 - Convenient one-stop place for all course-related electronic communication
- RStudio: The main software interface we will use.

On the Web

- Course Website: <http://colindawson.net/stat209>
 - Syllabus, schedule, homework, slides, code, etc., there
- Slack: stat209s2021.slack.com. Recommended:
Download the desktop and/or mobile app
 - Convenient one-stop place for all course-related electronic communication
- RStudio: The main software interface we will use.
 - You will soon get an account on rstudio.oberlin.edu; you can also install R and RStudio on your own computer if you want.

On the Web

- Course Website: <http://colindawson.net/stat209>
 - Syllabus, schedule, homework, slides, code, etc., there
- Slack: stat209s2021.slack.com. Recommended:
Download the desktop and/or mobile app
 - Convenient one-stop place for all course-related electronic communication
- RStudio: The main software interface we will use.
 - You will soon get an account on rstudio.oberlin.edu; you can also install R and RStudio on your own computer if you want.
- GitHub (later): Good way to track code changes/share code

On the Web

- Course Website: <http://colindawson.net/stat209>
 - Syllabus, schedule, homework, slides, code, etc., there
- Slack: stat209s2021.slack.com. Recommended:
Download the desktop and/or mobile app
 - Convenient one-stop place for all course-related electronic communication
- RStudio: The main software interface we will use.
 - You will soon get an account on rstudio.oberlin.edu; you can also install R and RStudio on your own computer if you want.
- GitHub (later): Good way to track code changes/share code
- Everything will be submitted and returned electronically, either via the RStudio server or (later) via GitHub

Assignments

- Frequent (~ 18) labs, mostly done in-class
- Occasional (~ 5) takehome quizzes on basic concepts
- Two group projects creating "data journalism" style "blog posts"
- Individual final project along similar lines

Grade Breakdown

Course grade based on:

- Demonstrated mastery of 20 discrete learning objectives (60%)
- Good faith, timely completion of assigned work (30%)
- Peer feedback on group projects (10%)

Structure of Class

- Largely, a “flipped class” structure

Structure of Class

- Largely, a “flipped class” structure
 - Read and watch posted mini-lecture videos before class

Structure of Class

- Largely, a “flipped class” structure
 - Read and watch posted mini-lecture videos before class
 - Start class with Q&A, then work on the lab for the day

Structure of Class

- Largely, a “flipped class” structure
 - Read and watch posted mini-lecture videos before class
 - Start class with Q&A, then work on the lab for the day
- Project milestones

Structure of Class

- Largely, a “flipped class” structure
 - Read and watch posted mini-lecture videos before class
 - Start class with Q&A, then work on the lab for the day
- Project milestones
 - In class “workshop day”, to work out the kinks together

Structure of Class

- Largely, a “flipped class” structure
 - Read and watch posted mini-lecture videos before class
 - Start class with Q&A, then work on the lab for the day
- Project milestones
 - In class “workshop day”, to work out the kinks together
 - In class “peer feedback” day, giving short informal presentations of draft visualizations “round robin” style

Structure of Class

- Largely, a “flipped class” structure
 - Read and watch posted mini-lecture videos before class
 - Start class with Q&A, then work on the lab for the day
- Project milestones
 - In class “workshop day”, to work out the kinks together
 - In class “peer feedback” day, giving short informal presentations of draft visualizations “round robin” style
 - Full draft due shortly after

Structure of Class

- Largely, a “flipped class” structure
 - Read and watch posted mini-lecture videos before class
 - Start class with Q&A, then work on the lab for the day
- Project milestones
 - In class “workshop day”, to work out the kinks together
 - In class “peer feedback” day, giving short informal presentations of draft visualizations “round robin” style
 - Full draft due shortly after
 - Written peer feedback soon after that

Structure of Class

- Largely, a “flipped class” structure
 - Read and watch posted mini-lecture videos before class
 - Start class with Q&A, then work on the lab for the day
- Project milestones
 - In class “workshop day”, to work out the kinks together
 - In class “peer feedback” day, giving short informal presentations of draft visualizations “round robin” style
 - Full draft due shortly after
 - Written peer feedback soon after that
 - Final draft due about a week after that

Structure of Class

- Largely, a “flipped class” structure
 - Read and watch posted mini-lecture videos before class
 - Start class with Q&A, then work on the lab for the day
- Project milestones
 - In class “workshop day”, to work out the kinks together
 - In class “peer feedback” day, giving short informal presentations of draft visualizations “round robin” style
 - Full draft due shortly after
 - Written peer feedback soon after that
 - Final draft due about a week after that
- Project 3: Some workshop time during the last week or so of classes, but not the full cycle of draft and revision.

To Do This Week

- Sign up for a 10 minute "meet and greet" during office hours (appointment link at the course webpage)

To Do This Week

- Sign up for a 10 minute "meet and greet" during office hours (appointment link at the course webpage)
- Fill out the background survey on the course website

To Do This Week

- Sign up for a 10 minute "meet and greet" during office hours (appointment link at the course webpage)
- Fill out the background survey on the course website
- Join the Slack group (stat209s2021.slack.com)

To Do This Week

- Sign up for a 10 minute “meet and greet” during office hours (appointment link at the course webpage)
- Fill out the background survey on the course website
- Join the Slack group (stat209s2021.slack.com)
- Lab 1 Thursday to get comfortable with R/RStudio (bring a laptop!)

To Do This Week

- Sign up for a 10 minute “meet and greet” during office hours (appointment link at the course webpage)
- Fill out the background survey on the course website
- Join the Slack group (stat209s2021.slack.com)
- Lab 1 Thursday to get comfortable with R/RStudio (bring a laptop!)
- Next week: Basic elements of data visualization