Statistical Computing as an Introduction to Data Science

Colin Rundel

JSM 2016 - Chicago

Duke University
Department of Statistical Science

Sta 323 - Statistical Computing

Course details:

- Foundational computing course
- 2nd/3rd year elective for BSS
- Approximately 40 Students divided into teams of 4
- Biweekly team programming assignments
- Individual take-home midterms, team final project

Learning Objectives

1. R programming and ecosystem

   (R + Tidyverse)

2. Reproducible Research

   (rmarkdown + knitr + *make*)

3. Software Engineering / Collaboration

   (shell + git + github)

Infrastructure

Dedicated departmental server

- RStudio Server Pro
- Individual departmental accounts
- System wide install of default packages

Github Organization

- 1 Organization / class
- 1 private repo / team / assignment
- Shared public repos (e.g. examples)
- CI / Testing via Wercker

Why github?

All assignment (and project) related work is maintained on github

- Forces students to use version control (git)
- Simplifies course administration
  - Code / documentation / scaffolding in one place
  - Easy to grab files (pull)
  - Easy to distribute files (push)
  - Built-in team permissions
- Searchability
- Accountability
- Continuous integration tools

Course Sketch

HW1 - FizzBuzz (Workflow basics)

HW2 - Graph Data Structures (Base R, testing)

HW3 - La Quinta is Spanish for next to Denny's (Web APIs, scraping, make)

HW4 - Karl Broman's Socks (Shiny, profiling, parallelization)

HW5 - Parking Wars: Manhattan (Data munging, prediction)

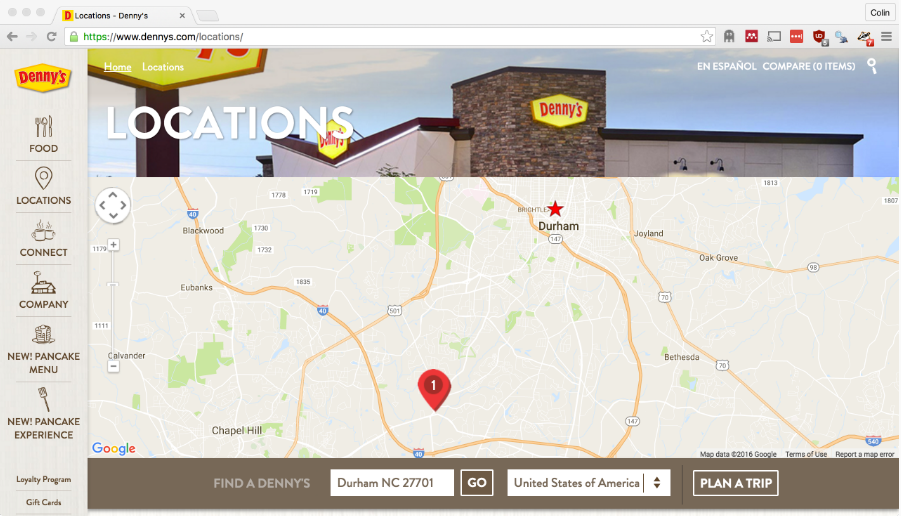
HW6 - How big is your data? (Hadoop, Spark)

HW3 - La Quinta is Spanish for next to Denny's

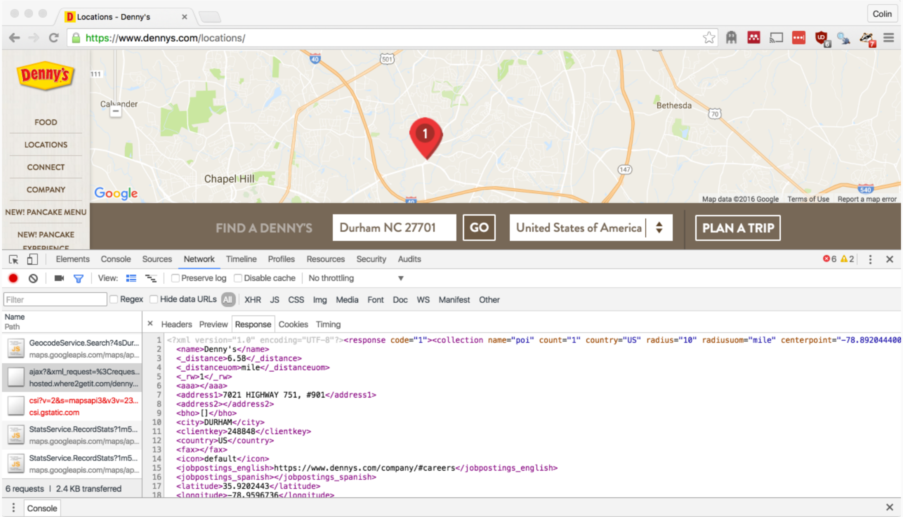
Assignment is based on a post by John Reiser on his new jersey geographer blog.

See Taking a Chance in the Classroom (Chance Vol. 29, Iss. 2, 2016) for a more detailed write up the statistical analysis aspect of this assignment.

Finding Denny's Locations



Finding the API



```
GeocodeService.Search?4sDur...
maps.googleapis.com/maps/ap...
ajax?&xml_request=%3Creques...
hosted.where2getit.com/denny...
csi?v=2&s=mapsapi3&v3v=23...
csi.gstatic.com
StatsService.RecordStats?1m5...
maps.googleapis.com/maps/ap...
StatsService.RecordStats?1m5...
maps.googleapis.com/maps/ap...
6 requests | 2.4 KB transferred
```

```xml
<?xml version="1.0" encoding="UTF-8"?><response code="1"><collection name="poi" count="1" country="US" radius="10" radiusuom="mile" centerpoint="-78.892044400
  <name>Denny's</name>
  <_distance>6.58</_distance>
  <_distanceuom>mile</_distanceuom>
  <_rw>1</_rw>
  <aaa></aaa>
  <address1>7021 HIGHWAY 751, #901</address1>
  <address2></address2>
  <bho>[]</bho>
  <city>DURHAM</city>
  <clientkey>248848</clientkey>
  <country>US</country>
  <fax></fax>
  <icon>default</icon>
  <jobpostings_english>https://www.dennys.com/company/#careers</jobpostings_english>
  <jobpostings_spanish></jobpostings_spanish>
  <latitude>35.9202443</latitude>
  <longitude>-78.9596736</longitude>
```

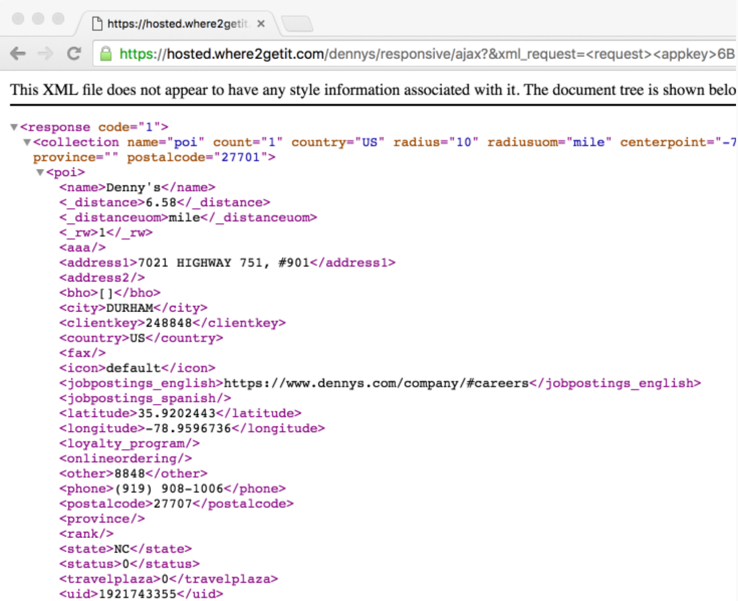
API Request

```
https://hosted.where2getit.com/dennys/responsive/ajax?&xml_request=
<request>
<appkey>6B962D40-03BA-11E5-BC31-9A51842CA48B</appkey>
    <formdata id="locatorsearch">
        <dataview>store_default</dataview>
        <limit>16</limit>
        <order>rank,_distance</order>
        <geolocs>
            <geoloc>
                <addressline>Durham NC 27701</addressline>
                <longitude>-78.89204440000003</longitude>
                <latitude>35.9981205</latitude>
                <country>US</country>
            </geoloc>
        </geolocs>
        <stateonly>1</stateonly>
        <searchradius>10|25|50|100</searchradius>
    </formdata>
</request>
```

API Request

```
https://hosted.where2getit.com/dennys/responsive/ajax?&xml_request=
<request>
<appkey>6B962D40-03BA-11E5-BC31-9A51842CA48B</appkey>
    <formdata id="locatorsearch">
        <dataview>store_default</dataview>
        <limit>16</limit>
        <order>rank,_distance</order>
        <geolocs>
            <geoloc>
                <addressline>Durham NC 27701</addressline>
                <longitude>-78.89204440000003</longitude
                <latitude>35.9981205</latitude>
                <country>US</country>
            </geoloc>
        </geolocs>
        <stateonly>1</stateonly>
        <searchradius>10|25|50|100</searchradius>
    </formdata>
</request>
```

API Results

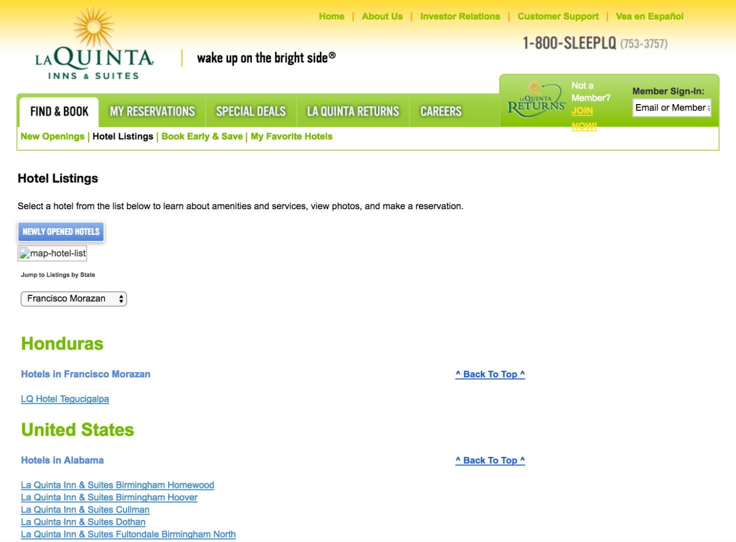


https://hosted.where2getit.com/dennys/responsive/ajax?&xml_request=<request><appkey>6B

This XML file does not appear to have any style information associated with it. The document tree is shown belo

```
▼<response code="1">
  ▼<collection name="poi" count="1" country="US" radius="10" radiusuom="mile" centerpoint="-7
    province="" postalcode="27701">
    ▼<poi>
        <name>Denny's</name>
        <_distance>6.58</_distance>
        <_distanceuom>mile</_distanceuom>
        <_rw>1</_rw>
        <aaa/>
        <address1>7021 HIGHWAY 751, #901</address1>
        <address2/>
        <bho>[]</bho>
        <city>DURHAM</city>
        <clientkey>248848</clientkey>
        <country>US</country>
        <fax/>
        <icon>default</icon>
        <jobpostings_english>https://www.dennys.com/company/#careers</jobpostings_english>
        <jobpostings_spanish/>
        <latitude>35.9202443</latitude>
        <longitude>-78.9596736</longitude>
        <loyalty_program/>
        <onlineordering/>
        <other>8848</other>
        <phone>(919) 908-1006</phone>
        <postalcode>27707</postalcode>
        <province/>
        <rank/>
        <state>NC</state>
        <status>0</status>
        <travelplaza>0</travelplaza>
        <uid>1921743355</uid>
```

Finding La Quinta Locations



Hotel Listings

Select a hotel from the list below to learn about amenities and services, view photos, and make a reservation.

Honduras

Hotels in Francisco Morazan

^ Back To Top ^

LQ Hotel Tegucigalpa

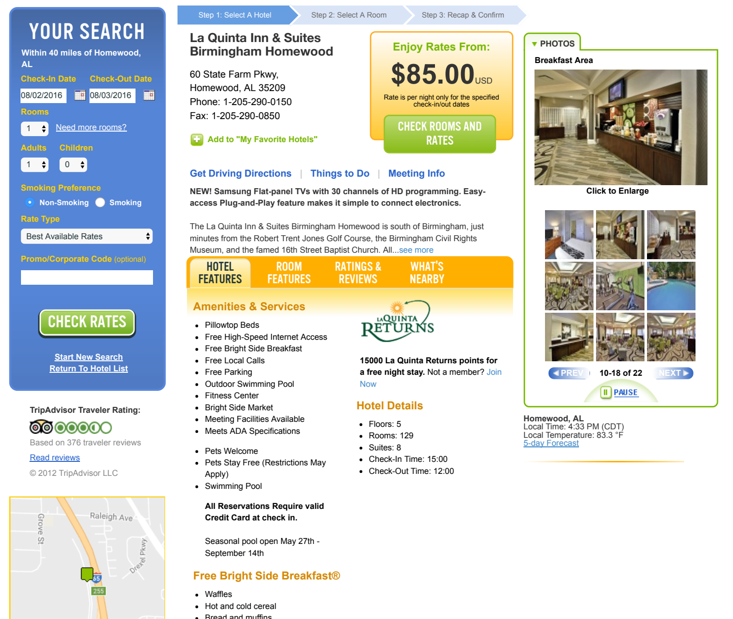
United States

Hotels in Alabama

^ Back To Top ^

La Quinta Inn & Suites Birmingham Homewood
La Quinta Inn & Suites Birmingham Hoover
La Quinta Inn & Suites Cullman
La Quinta Inn & Suites Dothan
La Quinta Inn & Suites Fultondale Birmingham North

Finding La Quinta Location Information



YOUR SEARCH

Within 40 miles of Homewood, AL

Check-In Date	Check-Out Date
08/02/2016	08/03/2016

Rooms: 1 — Need more rooms?

Adults	Children
1	0

Smoking Preference: ● Non-Smoking ○ Smoking

Rate Type: Best Available Rates

Promo/Corporate Code (optional)

CHECK RATES

Start New Search
Return To Hotel List

TripAdvisor Traveler Rating:

Based on 376 traveler reviews

Read reviews

© 2012 TripAdvisor LLC

Step 1: Select A Hotel — Step 2: Select A Room — Step 3: Recap & Confirm

La Quinta Inn & Suites Birmingham Homewood

60 State Farm Pkwy,
Homewood, AL 35209
Phone: 1-205-290-0150
Fax: 1-205-290-0850

Add to "My Favorite Hotels"

Enjoy Rates From:
$85.00 USD
Rate is per night only for the specified check-in/out dates

CHECK ROOMS AND RATES

Get Driving Directions | Things to Do | Meeting Info

NEW! Samsung Flat-panel TVs with 30 channels of HD programming. Easy-access Plug-and-Play feature makes it simple to connect electronics.

The La Quinta Inn & Suites Birmingham Homewood is south of Birmingham, just minutes from the Robert Trent Jones Golf Course, the Birmingham Civil Rights Museum, and the famed 16th Street Baptist Church. All...see more

HOTEL FEATURES | ROOM FEATURES | RATINGS & REVIEWS | WHAT'S NEARBY

Amenities & Services

- Pillowtop Beds
- Free High-Speed Internet Access
- Free Bright Side Breakfast
- Free Local Calls
- Free Parking
- Outdoor Swimming Pool
- Fitness Center
- Bright Side Market
- Meeting Facilities Available
- Meets ADA Specifications

- Pets Welcome
- Pets Stay Free (Restrictions May Apply)
- Swimming Pool

All Reservations Require valid Credit Card at check in.

Seasonal pool open May 27th - September 14th

Free Bright Side Breakfast®

- Waffles
- Hot and cold cereal
- Bread and muffins

LA QUINTA RETURNS

15000 La Quinta Returns points for a free night stay. Not a member? Join Now

Hotel Details

- Floors: 5
- Rooms: 129
- Suites: 8
- Check-In Time: 15:00
- Check-Out Time: 12:00

PHOTOS

Breakfast Area

Click to Enlarge

PREV 10-18 of 22 NEXT

PAUSE

Homewood, AL
Local Time: 4:33 PM (CDT)
Local Temperature: 83.3 °F
5-day Forecast

Reproducibility

Up to this point all reproducibility is based on individual Rmd documents (code and analysis in the same place)

Reproducibility

Up to this point all reproducibility is based on individual Rmd documents (code and analysis in the same place)

- Not feasible for this task - (irresponsible) grabbing of Denny's and La Quinta pages takes several minutes.

Reproducibility

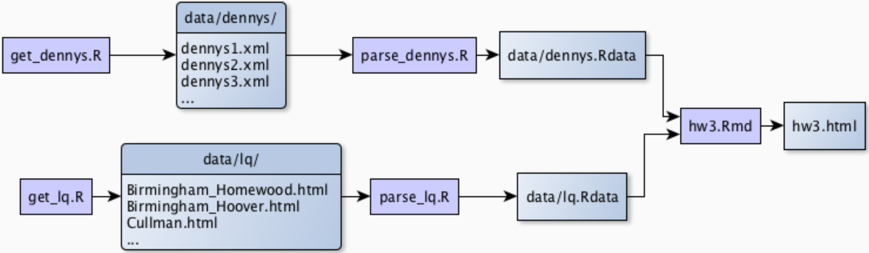
Up to this point all reproducibility is based on individual Rmd documents (code and analysis in the same place)

- Not feasible for this task - (irresponsible) grabbing of Denny’s and La Quinta pages takes several minutes.
- This is a “solved” problem for software development - build tools (e.g. make)

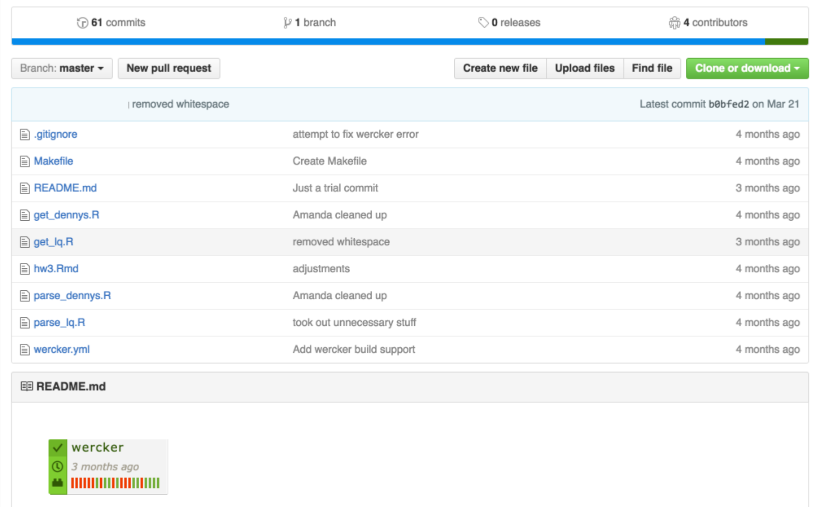
Reproducibility

Up to this point all reproducibility is based on individual Rmd documents (code and analysis in the same place)

- Not feasible for this task - (irresponsible) grabbing of Denny's and La Quinta pages takes several minutes.
- This is a "solved" problem for software development - build tools (e.g. make)



Github Repo

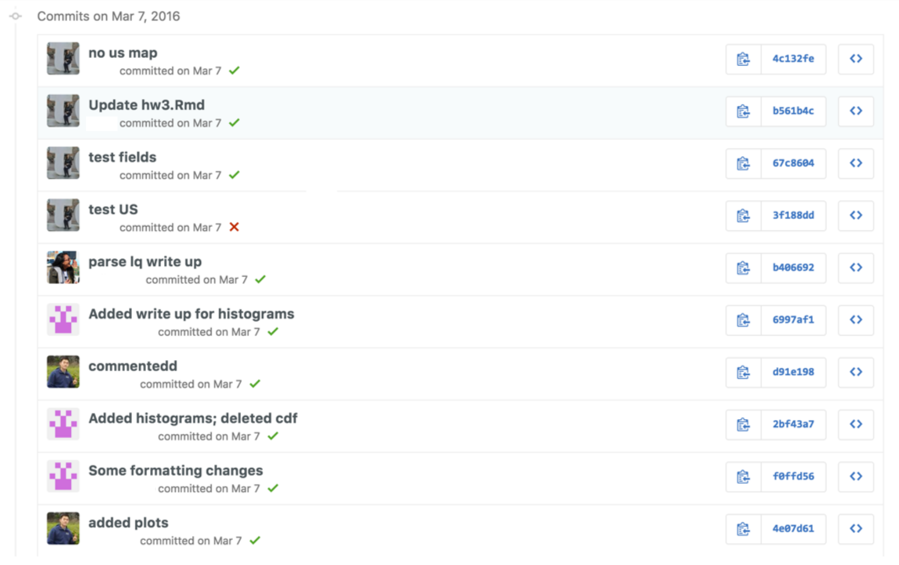


removed whitespace — Latest commit b0bfed2 on Mar 21

.gitignore	attempt to fix wercker error	4 months ago
Makefile	Create Makefile	4 months ago
README.md	Just a trial commit	3 months ago
get_dennys.R	Amanda cleaned up	4 months ago
get_lq.R	removed whitespace	3 months ago
hw3.Rmd	adjustments	4 months ago
parse_dennys.R	Amanda cleaned up	4 months ago
parse_lq.R	took out unnecessary stuff	4 months ago
wercker.yml	Add wercker build support	4 months ago

README.md

Github Commits



Commits on Mar 7, 2016

no us map
committed on Mar 7 ✓ 4c132fe

Update hw3.Rmd
committed on Mar 7 ✓ b561b4c

test fields
committed on Mar 7 ✓ 67c8604

test US
committed on Mar 7 ✗ 3f188dd

parse lq write up
committed on Mar 7 ✓ b406692

Added write up for histograms
committed on Mar 7 ✓ 6997af1

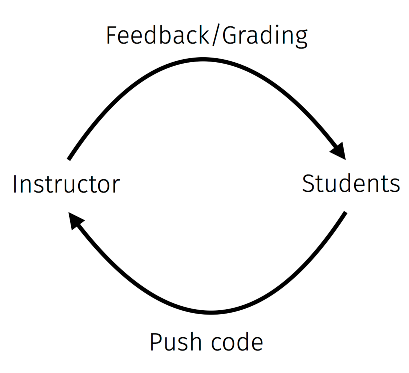
commentedd
committed on Mar 7 ✓ d91e198

Added histograms; deleted cdf
committed on Mar 7 ✓ 2bf43a7

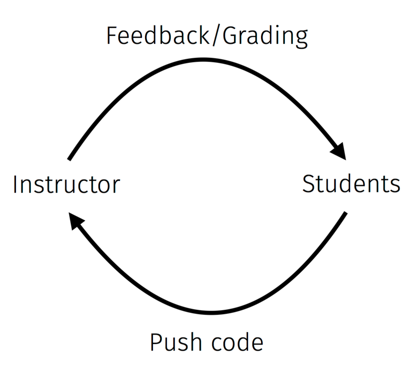
Some formatting changes
committed on Mar 7 ✓ f0ffd56

added plots
committed on Mar 7 ✓ 4e07d61

Feedback loop



Feedback loop



Github is great for feedback and accountability but doesn’t address scalability of the instructor and TAs (we are the rate limiting step).

A painfully common conversation

Student: We've submitted HW3!

+1 Day

Me: Your Rmd file doesn't knit, you used **`setwd`** *with an absolute path.*

+1 Day

Student: Ok we fixed that, does it work now?

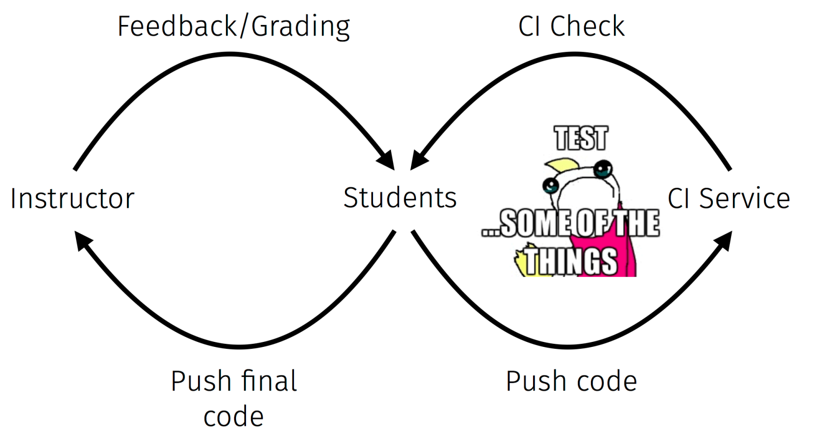
+1 Day

Me: Nope, you used `lme4` *without checking if it was installed.*

+1 Day

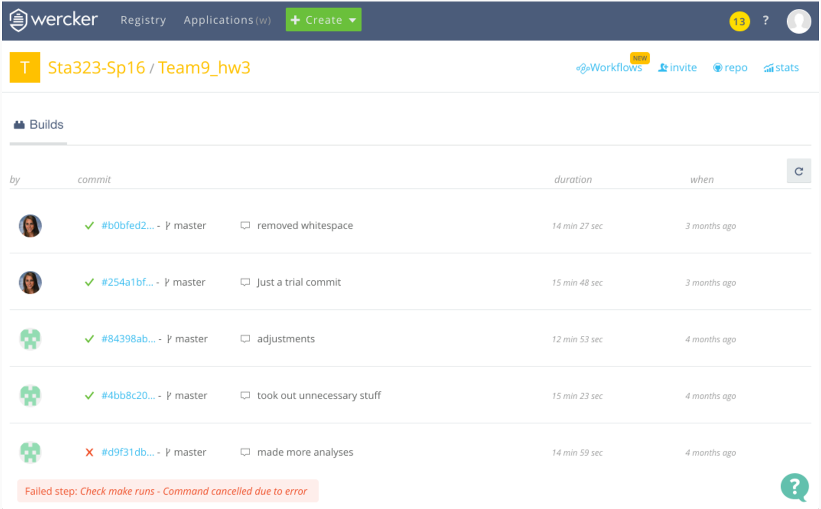
⋮

Course Process Cartoon - Improved



Goal is not to test for correctness - test for process / reproducibility.

Wercker

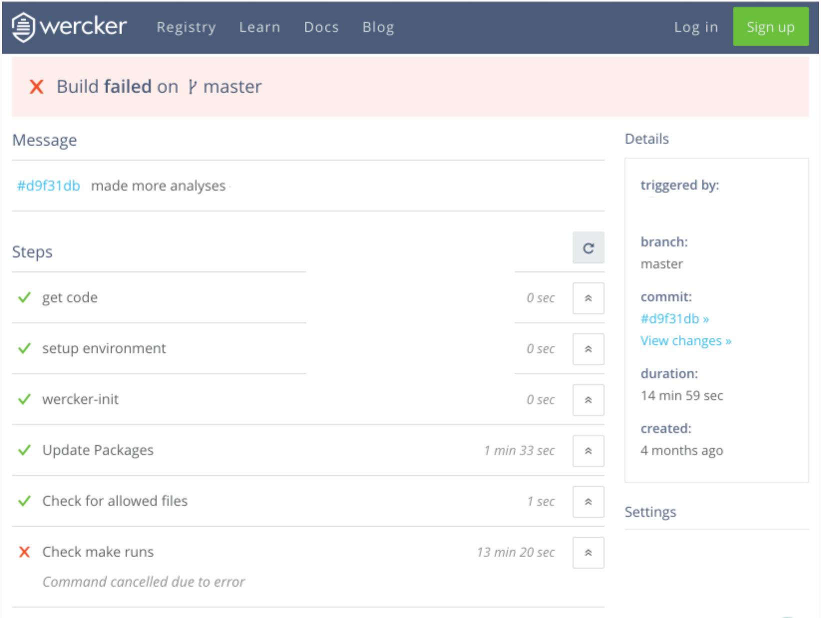


T Sta323-Sp16 / Team9_hw3

by		commit			duration	when
	✓	#b0bfed2... - master		removed whitespace	14 min 27 sec	3 months ago
	✓	#254a1bf... - master		Just a trial commit	15 min 48 sec	3 months ago
	✓	#84398ab... - master		adjustments	12 min 53 sec	4 months ago
	✓	#4bb8c20... - master		took out unnecessary stuff	15 min 23 sec	4 months ago
	✗	#d9f31db... - master		made more analyses	14 min 59 sec	4 months ago

Failed step: *Check make runs - Command cancelled due to error*

Wercker Steps



✗ Build **failed** on master

Message

#d9f31db made more analyses

Steps

✓	get code	0 sec
✓	setup environment	0 sec
✓	wercker-init	0 sec
✓	Update Packages	1 min 33 sec
✓	Check for allowed files	1 sec
✗	Check make runs *Command cancelled due to error*	13 min 20 sec

Details

triggered by:

branch:
master

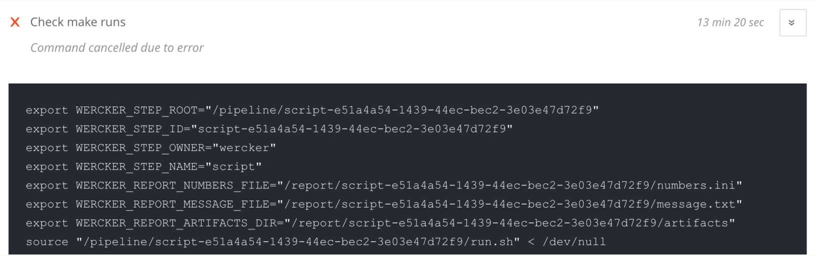
commit:
#d9f31db »
View changes »

duration:
14 min 59 sec

created:
4 months ago

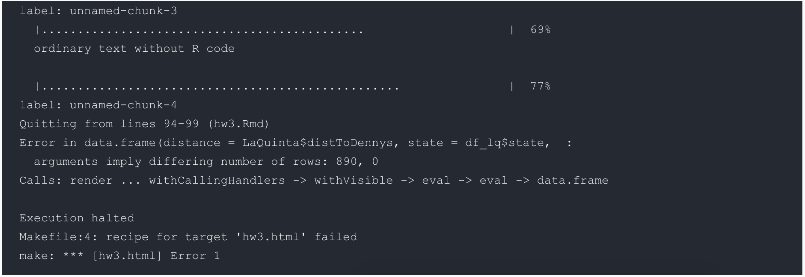
Settings

Wercker Error



```
export WERCKER_STEP_ROOT="/pipeline/script-e51a4a54-1439-44ec-bec2-3e03e47d72f9"
export WERCKER_STEP_ID="script-e51a4a54-1439-44ec-bec2-3e03e47d72f9"
export WERCKER_STEP_OWNER="wercker"
export WERCKER_STEP_NAME="script"
export WERCKER_REPORT_NUMBERS_FILE="/report/script-e51a4a54-1439-44ec-bec2-3e03e47d72f9/numbers.ini"
export WERCKER_REPORT_MESSAGE_FILE="/report/script-e51a4a54-1439-44ec-bec2-3e03e47d72f9/message.txt"
export WERCKER_REPORT_ARTIFACTS_DIR="/report/script-e51a4a54-1439-44ec-bec2-3e03e47d72f9/artifacts"
source "/pipeline/script-e51a4a54-1439-44ec-bec2-3e03e47d72f9/run.sh" < /dev/null
```

⋮

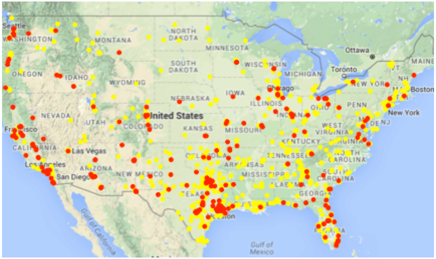


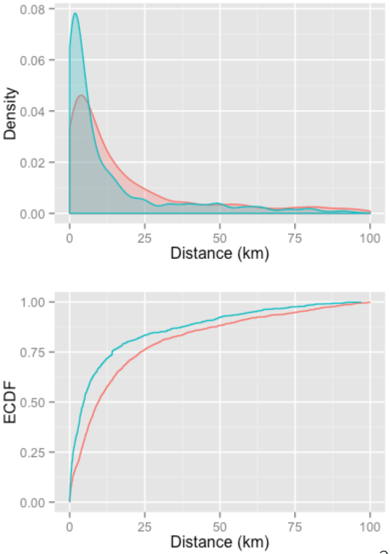
```
label: unnamed-chunk-3
  |.............................................                    |  69%
  ordinary text without R code

  |..................................................               |  77%
label: unnamed-chunk-4
Quitting from lines 94-99 (hw3.Rmd)
Error in data.frame(distance = LaQuinta$distToDennys, state = df_lq$state,   :
  arguments imply differing number of rows: 890, 0
Calls: render ... withCallingHandlers -> withVisible -> eval -> eval -> data.frame

Execution halted
Makefile:4: recipe for target 'hw3.html' failed
make: *** [hw3.html] Error 1
```

Final Analysis





Lessons Learned

- Use github* for everything
- Investments in automation pay off
- Don’t reinvent the wheel - borrow software engineering best practices
- Programming fundamentals are important - but tools and applications provide better motivation

Questions, Comments?

rundel@gmail.com

github.com/rundel/

github.com/rundel/Presentations/

bit.ly/Sta523_2014
bit.ly/Sta523_2015
bit.ly/Sta323_2016