

QSS20: Modern Statistical Computing

Session 0101: Intro and Setup

Goal for today's session

- ▶ Course goals
- ▶ Intros
- ▶ Break
- ▶ Nuts and bolts
- ▶ Residual tech setup

Goal for today's session

- ▶ **Course goals**

- ▶ Intros

- ▶ Break

- ▶ Nuts and bolts

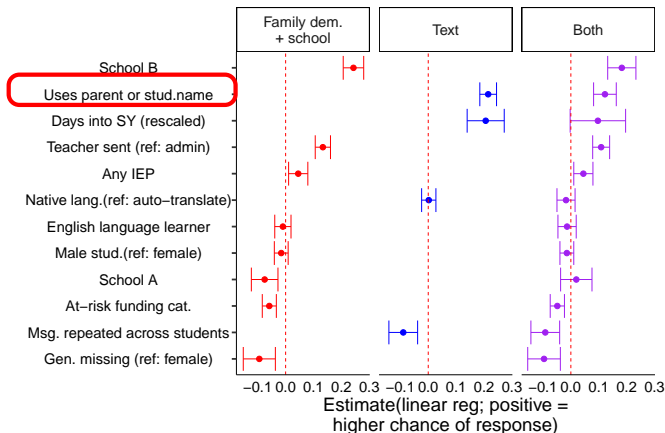
- ▶ Residual tech setup

Broad goals

- ▶ Build upon your introductory programming course and to equip you with the computing literacy to conduct social science research in the age of “big data.”
- ▶ Two components
 1. **Workflow tools:** Git/GitHub; LaTeX; basic bash scripting
 2. **Programming in messy contexts:** applied tasks in Python (data wrangling; basic text analysis; basic geospatial); some SQL

An example

Graph from a recent talk; **box in red** shows that parents are more likely to respond to text messages from teachers when the teacher uses the parent or their child's name:



Beyond the statistics, series of workflow and programming tasks before running regression

1. Acquire the data:

- ▶ **Ideal:** csv or database
- ▶ **Real:** excel file w/ variable number of tabs and spaces in column names; pdfs containing text; website

2. Clean the data:

p_name	s_name	msg_content
Rebecca Johnson	Jennifer John-son	Hi Ms. Johnson! Jenny did great on her math test.
Rebecca Johnson	Jennifer John-son	Hello Rebecca- I'm concerned about Jennifer's grades.

3. Reconcile different decisions in data cleaning:



Script #5 Running Thoughts/Comments/Questions #22

vickyme1 opened this issue on Sep 10, 2020 · 1 comment

```
id_orig_nowide_toconcat['PARENTID_constructed_1'] = np.nan
id_orig_nowide_toconcat['PARENTID_constructed_2'] = np.nan

id_orig_nowide_toconcat
```

- ☒ Your comment says 96% match, but I only get 91% match on my end (increase to ~94% after including SY1819). Is 96% still the case for you? Worried I'm using the incorrect file

```
# With local ids only, we get a 96% match rate! up from 75%
osse_merge_1920.found_osse.value_counts(normalize=True)
```

How does QSS20 fit with other courses you might have taken/will take?

- ▶ **Data wrangling and visualization but focus on R:**
 - ▶ QSS17 (Data visualization): tidyverse; ggplot
 - ▶ QBS181 (Data wrangling): R and SQL
- ▶ **Deeper dives into the statistics/analysis side:** stats prereqs; some courses in COSC more focused on machine learning (may have one session on binary classification if time)
- ▶ **Throughout:** focus on real-world policy applications and ethics of and policy context behind the data
- ▶ And to summarize your feedback on redundancies/newer topics...

To reiterate the workflow before data are usable...

```

: resp_file = pd.read_csv("../public_data/QSS20_remotepreferences (Responses) - Form Responses 1.csv")

: raw_colnames = resp_file.columns.to_list()
raw_colnames

: ['Timestamp',
  'Email Address',
  'Score',
  'What is your preferred name?',
  'Will you be on campus or remote this term?',
  'Do you expect to have reliable internet access for accessing Slack and the Dartmouth VPN (if needed)?',
  'Do you expect to have reliable internet access to participate in the synchronous sessions (5:00-6:50 PM EST, Tuesdays and Thursdays)?',
  'What kind of device will you be able to use for programming assignments? Select all that apply',
  'Please tell me the time zone in which you will be living during the term. If different than Eastern time, how many hours behind or ahead are you? *',
  'Do you have any concerns about accessing resources, including basic needs (food, shelter, medical care), psychological care and counseling, or access to technology that you wish to share with me?',
  'Reviewing the syllabus topics at this link: https://rebeccajohnson88.github.io/qss20/docs/course\_schedule.html. What topic do you feel is the most new/valuable for you?',
  'For those same syllabus topics, what topic do you feel is the most review / redundant with your past coursework?',
  'In past data analysis or programming courses, what was one thing that was MOST HELPFUL for helping you get through difficult material?',
  'In past data analysis or programming courses, what was one thing that was LEAST HELPFUL for helping you get through difficult material?',
  'So that I can gear practice problems/examples to your interests, what are your general career goals? Select all that apply!',
  'A lot of our examples will be drawn from intersections of data science and public policy. Which of the following policy domains are most interesting to you?',
  'Do you have any additional questions or concerns about moving to online classes that you would like to share with me?']

```


To reiterate the workflow before data are usable...

```

1 def clean_onecol(one_col: str,
2                   cutoff = 5):
3     """
4     Take in a messy column name and return a
5     cleaned one
6
7     @param one_col: Messy column name
8     @param cutoff: number of tokens to cut the string at (default 5)
9
10    @return: clean column name
11    """
12
13    l = one_col.lower()
14    l_nosp = re.sub(r"\s+|\.|\\|\/|\(|\\)\|?|\.", '-', l)
15
16    ## tokenize
17    l_nosp_token = l_nosp.split("-")
18
19    ## if longer, add some remainder back in to
20    ## differentiate similar q's
21    if len(l_nosp_token) > cutoff+5:
22        random.seed(2021)
23        l_short = "-".join(l_nosp_token[:cutoff]) + \
24            "-".join(random.sample(l_nosp_token[cutoff:], 5))
25
26    ## otherwise keep short
27    else:
28        l_short = "-".join(l_nosp_token[:cutoff])
29    return(l_short)

```

Based on your feedback, topics that need MORE coverage



- ▶ Lots of interest in SQL
- ▶ APIs, web scraping, and other data acquisition methods
- ▶ More on git, command line, and other non-Anaconda ways of interacting with Python



- ▶ Good handle already on ggplot (we'll still have some viz stuff using the plotnine wrapper but can then move more quickly to interactive visualization)
- ▶ Python basics from COSC 1 (may condense first few weeks to get to newer content more quickly/focus on more complex data science applications)

Overarching goal: transparency and reproducibility



Why do those matter? Data science in high-stakes contexts

Misuses of data science:

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

AUTOMATING INEQUALITY

HOW HIGH-TECH TOOLS PROFILE,
POLICE, AND PUNISH THE POOR

Promoting responsible and equitable data science:



DATA FOR
BLACK LIVES



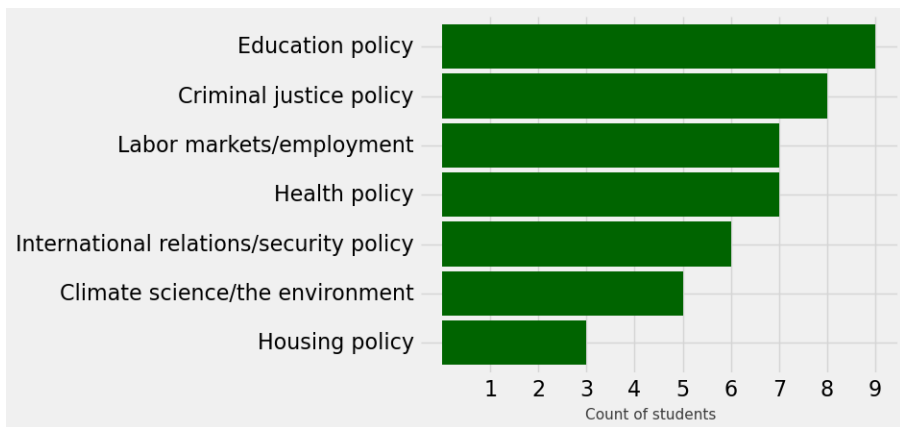
DATA SCIENCE
FOR
SOCIAL GOOD



Goal for today's session

- ▶ Course goals
- ▶ **Intros**
- ▶ Break
- ▶ Nuts and bolts
- ▶ Residual tech setup










General policy interests



Going around....

- ▶ Name
- ▶ Favorite class at Dartmouth thus far and why?
- ▶ If you could have any data source at your disposal, what would it be and what's a question you would ask?

A bit about me

Where	What	Languages
	Psychology; economics; MA in ethics/philosophy; internships in consulting	
	Research fellow at NIH dept of bioethics	None
 PRINCETON UNIVERSITY	PhD in sociology, demography, and social policy	
	Data scientist	 python™
 THE LAB @ DC	 Office of Evaluation Sciences	

Goal for today's session

- ▶ Course goals
- ▶ Intros
- ▶ Break
- ▶ **Nuts and bolts**
- ▶ Residual tech setup

Course website: most authoritative guide

Please make sure to read the following pages most closely (can click on links in posted slides):

1. **Evaluation and grading:** https://rebeccajohnson88.github.io/qss20/docs/eval_grades_py.html- covers four late days and exact grade breakdown
2. **Software setup:** https://rebeccajohnson88.github.io/qss20/docs/software_setup.html
3. **Social impact practicum context:** https://rebeccajohnson88.github.io/qss20/docs/sip_finalproject.html
4. **Course schedule (more subject to change):**
https://rebeccajohnson88.github.io/qss20/docs/course_schedule.html

Enrollment and waitlist

- ▶ **If enrolled:** due to the intensive nature of the course, please try to decide as soon as possible if you're going to stay enrolled so that we can use all spots
- ▶ **If waitlisted:** will update you as slots change; next offered Winter 2022

Course components

1. **Most important - synchronous sessions:** lab-based rather than lecture-based; hands-on practice with more advanced applications / work on problem sets
2. **Slack**
3. **Office hours**
4. DataCamp for review/basic syntax
5. Two problem sets
6. Social Impact Practicum/final project

Synchronous sessions structure of typical in-class session

Time window	What
5:00-5:30	Slides or review of tutorial on course website; DataCamp questions
5:30-5:45	Break and get into breakout rooms
5:45-6:30	Work with assigned group on in-class tutorial or problem set in breakout rooms; I'll circulate around
6:30-6:50	Reconvene as a group and go over questions; outline any prep for next class

Might deviate as we have visitors (currently, only visitors related to SIP practicum; might have guest speakers working in data science and public policy if there's interest and if we have spare course time)

Slack: course communication

- ▶ #general_qss20 for announcements
- ▶ Join by clicking link on Canvas
- ▶ Please add an image and preferred pronouns to your profile
- ▶ **Expectations:**
 - ▶ If in doubt, always default to a public channel so that others can benefit from your question
 - ▶ **DMs to me:** only for family emergencies and other personal issues
 - ▶ I will respond within 24 hours on weekdays; by Monday AM on weekends; before a problem set is due, will respond to all questions posted before **3 pm** on due date but not questions between 3 pm and midnight when due
 - ▶ Means that I've seen your message and am thinking:



Office hours

► My office hours

- When2meet for people's availability (will then set stable time starting this week): <https://www.when2meet.com/?11426443-RkPV5>
- Sign up via Calendly link I'll post/pin on Slack. Two formats:
 1. **1 hour of virtual office hours on zoom**; most likely 4 pm EST Fridays
 2. **1 hour of in-person office hours**: if interest; Blunt Hall room 205; general campus quarantine rules apply and need to either arrive at beginning of hour or middle of hour
- Jianjun Hua office hours: Mondays 2-4 pm EST and available on email

Course components

1. Most important - synchronous sessions: lab-based rather than lecture-based; hands-on practice with more advanced applications / work on problem sets
2. Slack
3. Office hours
4. **DataCamp for review/basic syntax**
5. **Two problem sets**
6. Social Impact Practicum/final project

DataCamp: make sure to join via our specific course pages so assignments show up

Post on [#datacamp_questions](#) if you need your name-based Dartmouth email added rather than your id-based email

My Assignments

TITLE ▾	TYPE ▴	ASSIGNEES ▴	ASSIGNER ▴	ASSIGNED ▴	DUE BY ▴	STATUS ▴	
Python Lists	Chapter	Organization	Rebecca Johnson	Mar 19, 2021	Apr 6, 17:00 EDT	IN PROGRESS	>
Python Basics	Chapter	Organization	Rebecca Johnson	Mar 19, 2021	Apr 6, 17:00 EDT	IN PROGRESS	>
Loops	Chapter	Organization	Rebecca Johnson	Mar 19, 2021	Apr 6, 18:00 EDT	IN PROGRESS	>
Data Manipulation with pandas	Course	Organization	Rebecca Johnson	Mar 19, 2021	Apr 6, 17:00 EDT	IN PROGRESS	>

Meant as auxiliary tool/playing a minor role so that you're prepared for in-class activities and so we don't need to review basic syntax. So graded on completion-only basis and only 5% of grade, but if you'd prefer to skip, can reapportion the 5% to pset 2

Two problem sets

- ▶ Will release 1 week before pset 1; at least 1.5 weeks before pset 2
- ▶ **Problem set one: due Thursday 04.15 at midnight:** Group submits a single submission for the group problems; you submit your own for the individual problem(s). Will have anonymous feedback survey on group participation that I weight into group portion of grade.
- ▶ **Problem set two: due Thursday 05.13 at midnight**
- ▶ **For each:**
 - ▶ Start well in advance and space out the parts (Pset 1 should be largely review from COSC 1 and the initial DataCamp modules)
 - ▶ May devote some class time pre deadline to work on the pset/answering questions
 - ▶ May provide intermediate/cleaned data so that getting stuck on early parts doesn't impede later parts

Course components

1. Most important - synchronous sessions: lab-based rather than lecture-based; hands-on practice with more advanced applications / work on problem sets
2. Slack
3. Office hours
4. DataCamp for review/basic syntax
5. Two problem sets
6. **Social Impact Practicum/final project**

What's a SIP?

- ▶ Sponsored by Dartmouth Center for Social Impact
- ▶ From them:
A Social Impact Practicum (SIP) is a project-based experiential learning opportunity connecting undergraduate courses at Dartmouth with community needs identified by nonprofit organizations throughout the Upper Valley.
In other words, a SIP is a real-world project with real-world impact.
- ▶ Can find a database of other SIPs here: <https://students.dartmouth.edu/social-impact/programs-initiatives/students/social-impact-practicums-sips/social-impact-practicum-sip-course>
- ▶ Ashley Doolittle (SIP director) happy to meet with anyone interested

Partner organization: Texas RioGrande Legal Aid

<https://www.trla.org/who-we-are>



[Legal Services](#) [About](#) [News](#) [Careers](#) [Pro Bono](#)



Who We Are

Social impact practicum: protecting the legal rights of temporary guestworkers

- ▶ Will learn much more from our visitors over the next couple weeks!
- ▶ **Broadly:**
 - ▶ US Department of Labor (DOL) gives what are called H-2 visas to authorized employers
 - ▶ Employers then hire foreign workers for these temporary visas; provide housing/transportation but the worker's visa is tied to that employer
 - ▶ If not well monitored, potential for violations of employees' rights (e.g., not paying them wages their due; overwork; unsafe conditions), since employees can face retaliation/deportation if they report issues

What TRLA data scientists have done thus far

<https://trla.shinyapps.io/H2Data/>

Scraper to pull daily job postings data on places employing H-2 guestworkers

H-2A and H-2B Job Posting Data

Connected to su-vpn.stanford.edu.

H2Data

Home

Basic Search ▾

Advanced Search (H-2A Only) ▾

← Change Data Selection

Download table as CSV



NUMBER OF RESULTS

10018



LAST UPDATED

03/27/2021

Show entries

Case
Number

Visa
type

Employer
Name

Job Title

Worksite
State

Date
Received
by DOL

Employment
Begin Date

Employment
End Date

Job Order Lin

Example post in NH for resort housekeeper

Job Details:

Occupational Code: **37201200 Maids and Housekeeping Cleaners**

Job Title: **Housekeeper**

Industry Code: **721110 - Hotels (except Casino Hotels) and Motels**

Number of Positions: **4**

Referrals: **9999**

Earliest Date to Display: **9/17/2020**

Last Date Job Order Will Display: **11/10/2020**

Type of Job: **Temporary**

Job Time Type: **Full Time (30 Hours or More)**

Duration: **Over 150 Days**

Special Job Category:

Job Duties and Skills:

Description:

Omni Mount Washington Resort, 310 Mount Washington Hotel Rd, Bretton Woods, NH 03575

4 Housekeepers needed for temporary, full-time employment from 12/01/20 to 3/31/21 in Bretton Woods, NH.

Job Duties: Maintain hotel/resort/villa in a clean and orderly manner. Clean guest rooms, in-room kitchens and living rooms, bathrooms, windows, conference facilities, halls, spa area and public spaces. Remove, sort, fold, carry and replace linens. Make beds, replenish supplies, set up guest room and meeting room furniture, pictures, and amenities according to resort standards. Mop, vacuum, extract/shampoo carpets, dust, clean/polish mirrors, dispose of refuse. Follow all sanitation policies and procedures.

No minimum education required.

1 month hotel/resort Housekeeper experience required.

Workers are subject to drug testing upon reasonable suspicion, paid by employer and applied equally to all workers, U.S. and foreign/H-2B.

Must be able to work a 5-day schedule, including weekends and holidays.

Applicants must complete an employment application.

But health and safety challenges

Bipartisan concern: <https://www.judiciary.senate.gov/press/dem/releases/senators-raise-concerns-over-h-2b-visa-abuses-that-enable-exploitation>

*Expert studies have demonstrated that the H-2B visa program, which provides no pathway to permanent legal status, leaves workers vulnerable to wage theft, abuse, and trafficking. H-2B workers are often at the mercy of their employers, which means **they may be afraid to speak out against poor working conditions**. Even when they do speak out, they **can struggle to access our justice system to protect themselves from retaliation by their employers**. This creates a perverse incentive for unscrupulous employers to hire H-2B workers instead of American workers.*

Broad project goals

- ▶ Past research has analyzed Department of Labor (DOL) enforcement data to investigate employers that have violated their workers' legal rights (e.g., withholding wages; unsafe conditions)
- ▶ Our project will link that data with the novel dataset collected by TRLA on job posting activity to investigate questions like: (1) what factors are correlated with getting caught violating workers' rights; (2) when an employer has faced legal sanctions, what does their posting activity look like afterwards; (3) are there ways to detect employers that try to evade program termination by changing their legal name?
- ▶ Ultimately guided by your interests and passions!

Goal for today's session

- ▶ Course goals
- ▶ Intros
- ▶ Break
- ▶ Nuts and bolts
- ▶ **Residual tech setup**

How to copy materials from our shared class folder to your personal jhub workspace

Will review more command line syntax on Thursday, but for now:

1. Navigate to <https://jhub.dartmouth.edu> and click on QSS20 option
2. Shared course materials (slides; in-class activities) are in the following read-only folder (shared/qss20/):

Select items to perform actions on them.



3. In the main directory (one level up from shared or the folder icon), create a folder to store editable files: `qss20_mywork`
4. Open terminal and run the following commands to copy the “testing.ipynb” file to that personal folder and check that it worked:
 - ▶ `cp shared/qss20/activities/testing.ipynb qss20_mywork/testing.ipynb`
 - ▶ `cd qss20_mywork`

Checklist (by Thursday classtime)

1. Are you on Slack and have you filled out your profile?
2. Are you set up on DataCamp?
3. Have you created a GitHub account?
 - 3.1 Fill out sheet here with your username: https://docs.google.com/spreadsheets/d/1ZyxiYudv0qWV4YYXwYD-oKLY0_ZdbZFVV1P77DfBC1A/edit#gid=399024829
4. Software setup: https://rebeccajohnson88.github.io/qss20/docs/software_setup.html
 - 4.1 Create an Overleaf account
 - 4.2 Local Python installation
 - 4.3 Terminal (already exists if on Mac! just make sure you can find it); terminal emulator