QSS20: Modern Statistical Computing

Unit 01: Intro and Setup

Goal for today's session

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

Goal for today's session

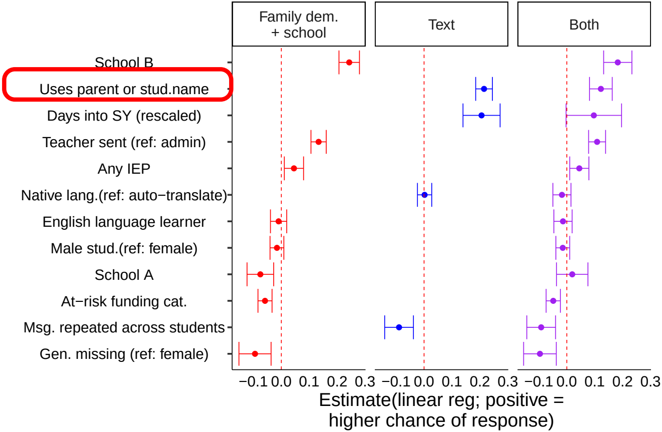
- **Course goals**
- Intros
- Break
- Nuts and bolts
- Residual tech setup

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 use of command line
  2. **Programming in messy contexts:** applied tasks in Python (data wrangling; basic text analysis); 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:

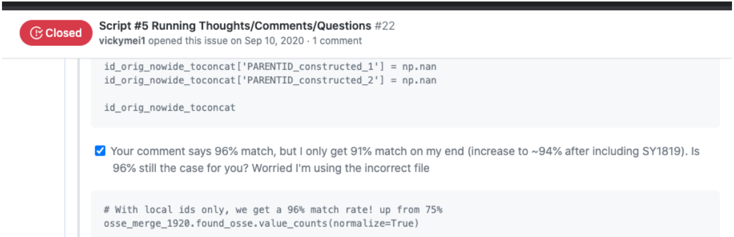


Beyond the statistics, series of workflow and programming tasks before running regression

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

3. **Reconcile different decisions in data cleaning:**



Script #5 Running Thoughts/Comments/Questions #22

vickymei1 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
```

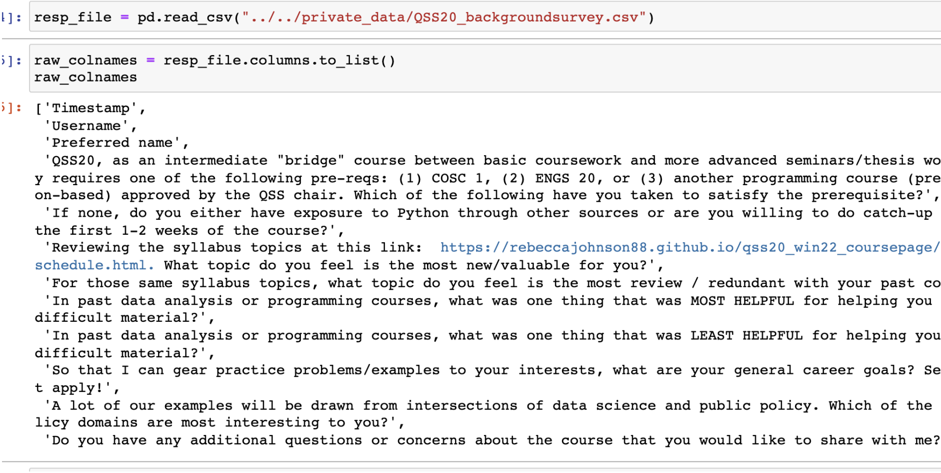
☑ Your comment says 96% match, but I only get 91% match on my end (increase to ~94% after including SY1819). Is 96% still the case for you? Worried I'm using the incorrect file

```
# 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...



```
4]: resp_file = pd.read_csv("../../private_data/QSS20_backgroundsurvey.csv")
```

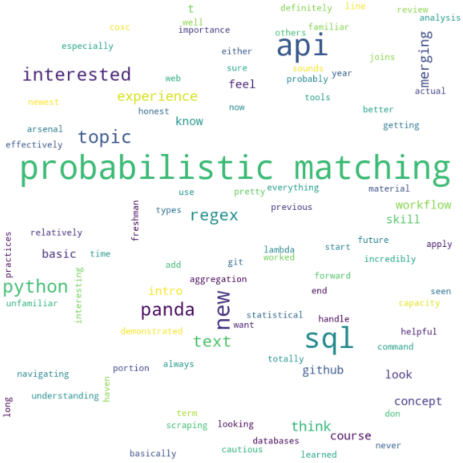
```
5]: raw_colnames = resp_file.columns.to_list()
    raw_colnames
```

```
5]: ['Timestamp',
     'Username',
     'Preferred name',
     'QSS20, as an intermediate "bridge" course between basic coursework and more advanced seminars/thesis wo
    y requires one of the following pre-reqs: (1) COSC 1, (2) ENGS 20, or (3) another programming course (pre
    on-based) approved by the QSS chair. Which of the following have you taken to satisfy the prerequisite?',
     'If none, do you either have exposure to Python through other sources or are you willing to do catch-up
    the first 1-2 weeks of the course?',
     'Reviewing the syllabus topics at this link:  https://rebeccajohnson88.github.io/qss20_win22_coursepage/
    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 co
     'In past data analysis or programming courses, what was one thing that was MOST HELPFUL for helping you
    difficult material?',
     'In past data analysis or programming courses, what was one thing that was LEAST HELPFUL for helping you
    difficult material?',
     'So that I can gear practice problems/examples to your interests, what are your general career goals? Se
    t apply!',
     'A lot of our examples will be drawn from intersections of data science and public policy. Which of the
    licy domains are most interesting to you?',
     'Do you have any additional questions or concerns about the course that you would like to share with me?
```

To reiterate the workflow before data are usable...

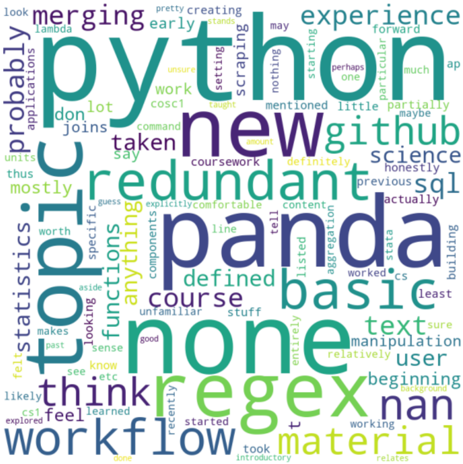
```python
def clean_onecol(one_col: str,
                 cutoff = 5):
    """
    Take in a messy column name and return a
    cleaned one

    @param one_col: Messy column name
    @param cutoff: number of tokens to cut the string at (default 5)

    @return:  clean column name
    """

    l = one_col.lower()
    l_nosp = re.sub(r"\s+|\,|\/|\(|\)|\?|\.", '_', l)

    ## tokenize
    l_nosp_token = l_nosp.split("_")

    ## if longer, add some remainder back in to
    ## differentiate similar q's
    if len(l_nosp_token) > cutoff+5:
        random.seed(2021)
        l_short = "_".join(l_nosp_token[:cutoff]) + \
                    "_".join(random.sample(l_nosp_token[cutoff:], 5))

    ## otherwise keep short
    else:
        l_short = "_".join(l_nosp_token[:cutoff])
    return(l_short)
```

Based on your feedback, topics that need MORE coverage



- Probabilistic / fuzzy matching
- SQL
- APIs, web scraping, and other data acquisition methods
- More on git, text as data, and regex

Based on your feedback, topics that need LESS coverage



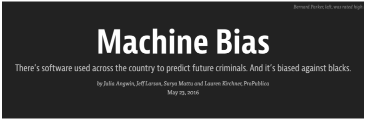
- ► Good handle already on python basics from COSC1 or other prereqs (eg user-defined functions)
- ► Course won't rehash statistics training from QSS15, ECON10, or other coursework

Overarching goal: transparency and reproducibility

DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
JORGE CHAM © 2014
WWW.PHDCOMICS.COM

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

Misuses of data science:



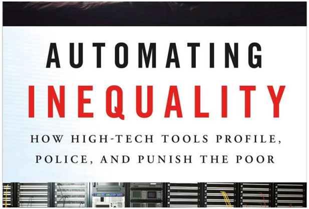
Bernard Parker, left, was rated high

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

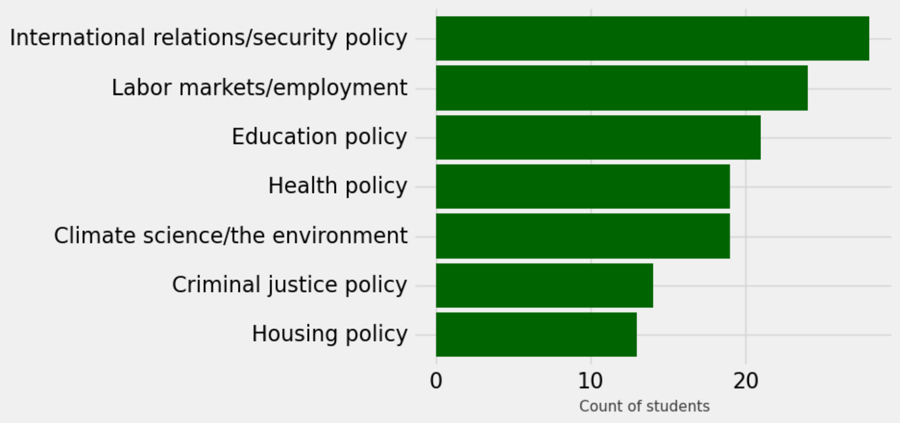
Promoting responsible and equitable data science:

DATA FOR BLACK LIVES

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	stata
BIOETHICS AT THE NIH	Research fellow at NIH dept of bioethics	None
PRINCETON UNIVERSITY	PhD in sociology, demography, and social policy	R
DATA SCIENCE FOR SOCIAL GOOD THE LAB @ DC OES Office of Evaluation Sciences	Data scientist	python™

Course TAs

- 3A: Jack Lauer: will introduce himself next week and is on slack!
- 6A: You-Chi (Eunice) Liu
- Each has taken a version of QSS20 (Eunice last spring; Jack via independent study)
- In addition to helping with problem sets/grading, resource via Slack and in class

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_win22_coursepage/docs/eval_grades_py.html- covers four late days and exact grade breakdown
2. **Software setup**: https://rebeccajohnson88.github.io/qss20_win22_coursepage/docs/software_setup.html– for now, focus on local Python and will cover JHUB next week
3. **Social impact practicum context**: https://rebeccajohnson88.github.io/qss20_win22_coursepage/docs/sip_finalproject.html
4. **Course schedule (more subject to change):** https://rebeccajohnson88.github.io/qss20_win22_coursepage/docs/course_schedule.html

Course components

1. **Most important - in-person class sessions:** lab-based rather than lecture-based; hands-on practice with more advanced applications / work on problem sets
2. **Slack-join via our Canvas page**
3. **Office hours**
4. DataCamp for review/basic syntax
5. Four problem sets
6. Social Impact Practicum/final project

Structure of typical in-class session

Time window	What
3:30-4:00 or 6:30-7:00	Slides or review of tutorial on course website; DataCamp questions
4:00-4:10 or 7:00-7:10	Break and get into small groups
4:10-5:00 or 7:10-8:00	Work with assigned group on in-class tutorial or problem set in breakout rooms; I'll circulate around
5:00-5:20 or 8:00-8:20	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 by next week's class
- **Expectations:**
  - If in doubt, always default to a public channel so that others can benefit from your question
  - **Order:** first tag your section's TA and then they can defer to me if they have problems answering
  - **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

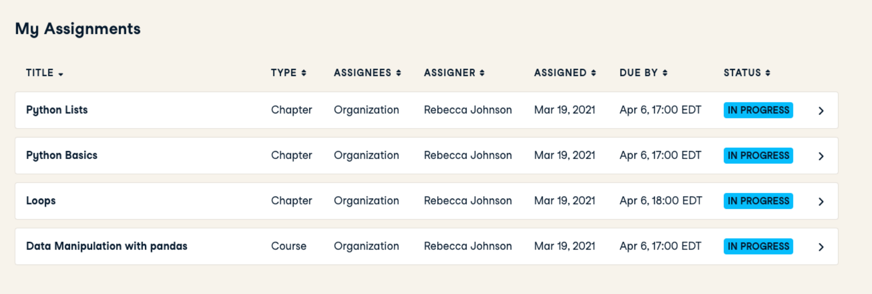
- **My office hours**
  - **Two slots (any section can go to either):** Wednesdays 12-1 pm; Thursdays 3-4 pm
  - Sign up in advance via this Calendly link I'll post/pin on Slack: `https://calendly.com/rebecca_a_johnson`. When signing up, indicate format:
    1. In person: my office (Blunt 301E)
    2. Zoom
- Will update on Slack with TA office hours

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. **Four 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 the first problem set

Four problem sets

- Pset one posted here (others may change) and will be on Canvas: https://github.com/rebeccajohnson88/qss20_slides_activities/blob/main/problemsets/01_pset1/pset1_blank.ipynb
- **Problem set one: due Friday 01-14 at 11:59 pm**
- **Others:** see schedule on course website
- **For each**:
  - Start well in advance (at least 3-4 days) 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. Four 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: UNH Center for Start Services

https://iod.unh.edu/projects/center-start-services

THE CENTER FOR
START SERVICES

Institute on Disability / UCED
University of New Hampshire

Overview

- Will learn much more from our visitors over the next few weeks!
- **Broadly:**
  - START has a database reflecting things like diagnoses, symptoms, clinical impressions, and interactions with emergency departments and law enforcement for individuals with IDD (intellectual and developmental disabilities)
  - We'll be merging data sources from different sources to explore patterns in mental health related outcomes for the patients during COVID-19

Example project questions

- COVID-19 and changes in suicidal ideation
- COVID-19 and racial disparities in interactions with law enforcement
- Natural language processing of clinical notes on patients
- Ultimately guided by your interests and passions!

Project examples from last spring

Students Inspired by QSS as a Tool for Social Change

Students are members of a team analyzing the treatment of guest workers in the U.S.

Project examples:

- Geo-visualization of locations of job sites relative to Census tract attributes (e.g., migration rates; unemployment): https://github.com/rebeccajohnson88/qss20_s21_proj/blob/main/memos/final_papers/dol_geocoding_writeup.pdf
- Causal analysis of relationship between inspection capacity and findings of legal issues: https://github.com/rebeccajohnson88/qss20_s21_proj/blob/main/memos/final_papers/dol_opmstaffing_writeup.pdf
- Natural language processing of job contracts: https://github.com/rebeccajohnson88/qss20_s21_proj/blob/main/memos/final_papers/dol_textasdata_writeup.pdf
- Supervised machine learning predicting investigations/violations: https://github.com/rebeccajohnson88/qss20_s21_proj/blob/main/memos/final_papers/dol_predictviol_writeup.pdf

Structure of project

- **Milestone 1:** memo or plan for what question you'll ask and analyses you'll run
- **Final outputs (see course website for more details):**
  - Final presentation (done in Beamer; LaTeX-based powerpoint software)
  - Short 10-page report (done in LaTeX)
  - Github repo and readme with all code to reproduce analyses

Goal for today's session

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

Next week

- Order of work: would recommend completing the DataCamp and then starting on problem set one unless you already feel comfortable with `pandas`
- Monday lecture/activity will focus on data wrangling relevant for psets one and two

Checklist (by Friday end of day)

1. Are you set up on DataCamp and working on the assignment due Monday noon?
   - Ping in #datacamp_questions with your email if need adding
2. Are you on Slack and have you filled out your profile (photo or avatar; pronouns)?
3. Have you created a GitHub account?
   - Will use in class on Wednesday 01-19 and will collect usernames beforehand
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