S&DS 177/577 Syllabus
YData: COVID-19 Behavioral Impacts

Instructor: Youpei Yan (youpei.yan@yale.edu)

- Introduction

Welcome to SDS177! This is a connector course to the main data science course SDS123. This series of YData courses aim to enhance students' knowledge and capabilities in the fundamental ideas and skills of data science. YData is an introduction to data science that emphasizes computational and programming skills along with inferential thinking. The course is designed to be accessible to students with little or no background in computing, programming, or statistics. At the same time, it is meant to be engaging for more technically oriented students, through the extensive use of examples and hands-on data analysis. The course is based on the Python programming language and a special-purpose cloud computing platform for students to edit and execute their code in Jupyter Notebook.

For our SDS177, we will focus on a special topic on COVID-19's Behavioral Impacts. People alter behavior in response to infectious disease risk, yet the 2020 COVID-19 epidemic is the first modern epidemic in the United States where most people were ordered by governments to reduce time in public. With modern data science, we are able to investigate how non-pharmaceutical interventions interact with voluntary or mandated behavioral shifts during epidemics. In this class, we use COVID-19 case reports and mobile device data during the pandemic to evaluate COVID-19 orders on people's representative behaviors.

- Data Sources

1. New York Times: COVID-19 case reports.
   Daily county's case and death reports.

2. List of COVID-19 policies.
   Based on the Raifman's database, which tracks the dates when each US state implemented new social safety net, economic, and physical distancing policies in response to the COVID-19 pandemic, combined with data on existing health and social policies and information on state characteristics.

3. Safegraph: Mobile Device Data.
   SafeGraph reports the median home dwell time and public visitation (for each industry) by census block group. Because of the data confidentiality, I will produce a device weighted average for each county or state (and maybe aggregate by week) for our course.

4. Weather Data.
   We construct a set of county-level weather control variables by aggregating 4km gridded estimates of maximum and minimum daily temperature, maximum and minimum relative humidity, precipitation amount, surface solar radiation (a measure of sunlight), and wind speed. Observations are by county day. The data are based on http://www.climatologylab.org/gridmet.html.

The datasets (given the sizes) can be accessed through the following Google Drive link:
https://drive.google.com/drive/folders/1SLPLMpgdzcn79iMBOYhh9CmOZaw7odpE
Please do not share the datasets with others unless you contact me and receive my permission.

- Course Structure

Weekly meeting online by Zoom.
Course time: **Th 9:25am-11:15pm**
Office hour: **Tu/Th 1-2pm** or by email (youpei.yan@yale.edu)

- Course Grade

We will have 4 projects (15 points each) + a midterm (20 points) + a final (20 points). All projects and both the exams will be completed online through the Jupyter Notebook. I highly encourage group discussion, but plagiarism is unacceptable.

You are able to use coding techniques we learn in class to answer the following questions with the given dataset. Note that each major question will be broken down to 10-15 small questions to practice your coding skills and critical thinking. The expected time to finish each homework assignments is 10 hours. Please do not save it till the due day!

Tentative topics of the 4 homework assignments:

1. COVID-19 Case/Death Reports vs Home Dwelling Time & Trips outside. (Week 1-3)

    a. Which states seem suffer the epidemic most/least? Do people go outside as before? Find the trips to places per device count with grouping and sorting techniques.
    b. Any possible causes based on the provided county or state-level characteristics? Use plots to show your findings.
    c. Simple raw comparison between a treatment group and a control group. What could bias the results?

2. Allocation of home dwelling time under the pandemic policies. (Week 4-6)

    a. Can we visualize a pattern of COVID19's state-level policies and the daily reported cases in those states? Try merging policy data with the case data.
    b. Merge the home-dwelling time data with the case reports and the policy lists. Plot the policy map and compare with the case reports.
    c. What factors would influence people's time allocation decisions? Use the correlation function to show your hypotheses with the given data.

Week 7 will have an extended office hour on Tuesday and the Midterm on Thursday.

3. Mask mandates and the fatigue from Stay-at-home/Shelter-in-place Policy (Week 8-11)

    a. What could be the sources of omitted variable biases if we plan to examine the effects of face mask mandates to people's allocation of time at home?
    b. Interpret the hypotheses and the *given* regression results and the confidence intervals.

4. Linear regression: effects of a selected COVID-19 policy by yourself. (Week 12-14, due at the end of the date for final)

    a. What other factors you should pay attention to, given the policy analysis you plan to conduct?
    b. Do some summary statistics to prove your thoughts.
    c. Run the regression and interpret the results. Does the policy have a significant impact to people's behavior (such as time to stay at home, or visitation to parks, restaurants, or warehouses)?

- Calendar Spring 2021

	Date	Topic	Lecture Notes	Homework	Key Words
1	4-Feb	Course Introduction & Data Import	Lec 1&2; Lab 0	Project 1	read_table(); read_url()
2	11-Feb	Data Loading; Select, Sort, Calculation	Lec 1&2; Lab 0		table.select(); table.sort(); column.mean()
3	18-Feb	Data manipulation; Grouping; Visualization	Lec 3; Lab 1	Project 1 Due	table.where(); table.group(); data.plot(); data.hist()
4	25-Feb	Panel Data; pandas	Lec 4; Lab 2	Project 2	pandas; df.groupby(); df.plot(); df.boxplot(); correlation vs. causality
5	4-Mar	More data visualization and Map Plot	Lec 5; Lab 3		df.loc[]; Plotly
6	11-Mar	Merge & Append Data; Review	Lec 6; Lab 4	Project 2 Due	pd.merge(); df.append()
7	**18-Mar**	**Midterm**			
8	25-Mar	Hypothesis Testing	Lec 7; Lab 5	Project 3	p-value; distribution; selection bias; experimental data; randomization; treatment effects; OVB
9	1-Apr	CI & Bootstrap	Lec 8; Lab 5		confidence intervals; bootstrap; vaccine
10	8-Apr	BREAKDAY			
11	15-Apr	Prediction & Regression	Lec 9; Lab 6	Project 3 Due	slope/intercept; significance; t-value
12	22-Apr	Linear Regression	Lec 10; Lab 6	Project 4	R square; residual; fitted value; OLS
13	29-Apr	Fixed Effects	Lec 11; Lab 6		statsmodels; model.fit()
14	6-May	Examples & Review		Project 4 Due	
	16-May	**Final at 2PM**			

* Apr 8th is Yale's 4th break day; we will not have class or homework due on that day.

- Lecture Notes

I will update my course notes, online practice (during the class), and our homework assignments on GitHub: https://github.com/youpeiyan/YData_SDS177

- Reference

Textbook: https://www.inferentialthinking.com/chapters/intro
YData123: http://ydata123.org/sp20/calendar.html

Case reports: https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv
Policy database: https://github.com/USCOVIDpolicy/COVID-19-US-State-Policy-Database
Safegraph data: https://www.safegraph.com/covid-19-data-consortium