

# 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 Notebooks.

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.

Proposed 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 8-10)
  - 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 11-13, 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 (links will be added before the beginning of the semester)

Date	Topic	Lecture Notes	Data Source	Homework	Key Words
Feb 4	Data Introduction	Lec 1 Lab 0	<a href="#">NYTimes</a>	Project 1	table.function() string histogram line/scatter plot
Feb 11	Data Types and Tables	Lec 2 Lab 1	county/state background		Panel Data pandas groupby.()
Feb 18	Visualization in Panel Data	Lec 3 Lab 2	dwel-time at home; weather data	Project 1 Due	data.plot() df.boxplot() correlation vs. causality
Feb 25	More data visualization and Map Plot	HW1 answer Lec 4 Lab 3	Visitation of public places	Project 2	causality Plotly
Mar 4	Data Merge/Append	Lec 5 Lab 4	Policy chart And State characteristics		table.pivot() table.join() np.append()
Mar 11	Distribution and Hypotheses	Lec 6 Lab 5		Project 2 Due	p-value
Mar 18	Midterm				
Mar 25	Causality and Omitted Variable Bias, Selection Problem in Econometrics	HW2/ Midterm answer Lec 7		Project 3	selection bias experimental data randomization treatment effects OVb
Apr 1	Confidence Intervals	Lec 8 Lab 6			bootstrap
Apr 15	Prediction and Regression	Lec 9 Lab 7		Project 3 Due *	slope/intercept
Apr 22	Linear Regression	HW3 answer Lec 10 Lab 8		Project 4	R square residual fitted value OLS
Apr 29	Significance	Lec 11 Lab 9			true slope hypothesis testing t-value
May 6	Examples and Review	Lec 12 Lab 10		Project 4 Due	
	Final				

\* Apr 8<sup>th</sup> is Yale's 4<sup>th</sup> 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](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>