서 울 대 학 교
SEOUL NATIONAL UNIVERSITY

서울대학교
데이터사이언스대학원
Graduate School of Data Science
Seoul National University

M3239.003100: Data Analysis and Visualization
Lecture 1

Introduction

Hyunwoo Park
Graduate School of Data Science
Seoul National University

Agenda

- Course Logistics
  - Instructor
  - Grading Scheme
  - Course Policy
- Course Overview
  - Class Schedule
  - Course History and Credit
  - Stata
  - Python
    - Data Collection and Cleaning
    - Static Visualization with Matplotlib
    - Dashboard Design
  - Web Programming
    - Interactive Visual Analysis System with d3.js
  - Why Visualization?
- Class Survey (Due 9/13)
  - Submitted via Google Forms
  - Link to be sent out by this weekend
- Things To Do
  - Install Stata
  - Install the Anaconda distribution of Python 3
  - Launch Jupyter Notebook
  - Get to know the terminal interface
  - Pick a text/code editor

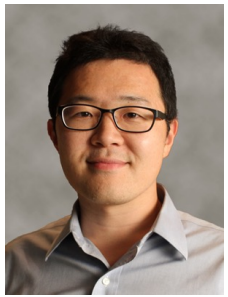
Course Logistics

Before we begin

- Course syllabus is posted on eTL.
- All course announcements will be posted on eTL.
- Teaching philosophy
  - I will try to make this course relevant to you.
  - I should teach what I believe is actually helpful for you moving forward.
  - I believe in learning by doing, especially for creating something with programming.
  - You should learn from my perspective and organization of thoughts, not just crystalized knowledge.
- Expectation
  - The course is not math heavy but programming heavy.
  - Workload will be very demanding, but I believe you will take away a lot of hands-on skills on data visualization when you complete the course.
  - You are expected to be a self-starter.

Instructor

- About me
  - Associate Professor
    Graduate School of Data Science
    Seoul National University
  - Email: hyunwoopark@snu.ac.kr
    Office: 942-417
    Office Hours: By appointment
  - For more information about me, visit http://hyunwoopark.com.



- Before SNU
  - Assistant Professor
    Dept. of Operations and Business Analytics
    Fisher College of Business
    The Ohio State University
    Affiliated with Translational Data Analytics Institute
  - Ph.D. in Industrial Engineering
    from Georgia Tech
  - Master in Information Mgt. and Systems
    from UC Berkeley
  - B.S. in Electrical Engineering from SNU

Grading scheme

- Midterm (40%)
- Participation (20%)
  - eTL Activities & Class Survey
- Group Project (40%)
  - Project Proposal (10%)
  - Final Presentation (15%)
  - Output Evaluation (15%)
  - Individual scores may be adjusted based on peer evaluation.

Homework (30%; 6% each)

- No late submission is accepted unless otherwise notified.
- All submissions must be individual work.
- Direct sharing of the code for homework is expressly and strictly prohibited and constitutes an academic misconduct.
- A high-level discussion is permitted and encouraged.

Group project (30%)

- A team of 5-7 people will work together to build an interactive visual analysis system using d3.js.
- I will form groups based on Class Survey AND your requests.
- Please send your group formation request separately. Your request for group formation may not be fully accommodated.
- Your final output may be showcased online.
- Free-riders will be punished by peer evaluation. A student with an extremely and consistently bad peer evaluation may receive a 0 for Group Project, even if the group score is high.
- Deliverables
  - Project proposal (a pdf document)
  - Working interactive visual analysis system
  - System demo video
  - 4+ screenshots highlighting the features of your visual analysis system
  - Peer evaluation

Participation > Class Survey (5%; Due 9/13)

- You will answer survey questions about yourself and your interest.
- It will be conducted using Google Forms.
- Survey answers will be compiled into a dataset to be used for this class.
- Survey link will be sent out by this week.
- Completing this survey properly is 5% of the course grade.
  (Some deductions may be applied if instructions were not followed.)

Participation > Online eTL Activities (15%)

- Content-related questions should be first posted on the eTL [Q&A] or [Sharing] board, so that other students can answer for you.
- You earn 1% for each question or answer you post.
- Only questions and answers posted with real name will be counted towards Online Q&A Activities grade.
- You earn credit for maximum one question or answer per day.
- Spammers and trollers may receive a 0, regardless of their past activities.

Grade appeal policy

- All grade appeals should be made within two weeks after the grade for the item is released.
- Final course grade (i.e., letter grade) is not negotiable for any reason.
  - Including, but not limited to, financial burden, job offer deadline, graduation plan, etc.

Office hours policy

- Office hours are offered by appointment only.
- I am unable to offer office hours to give one-on-one handholding sessions.
- You must email me describing briefly why you need a one-on-one conversation with me before scheduling an office hours session.
- Each office hours session is limited by 30 minutes. This is also to avoid a small number of students monopolizing my time on this course.

Language policy

- This course is taught in Korean. All materials will be in English and some Korean.

TA

- Kyu Tae Shim (심규태)
- ktshim@snu.ac.kr

Classroom

- Lectures will be in 942-302 in the following days.
  - 9/1 (Thu)
  - 9/6 (Tue)
  - 9/8 (Thu)
- After that, I will send an announcement on when the class will be in the classroom or just over zoom.
- Even if the class will be in the classroom, it will be broadcasted over zoom.

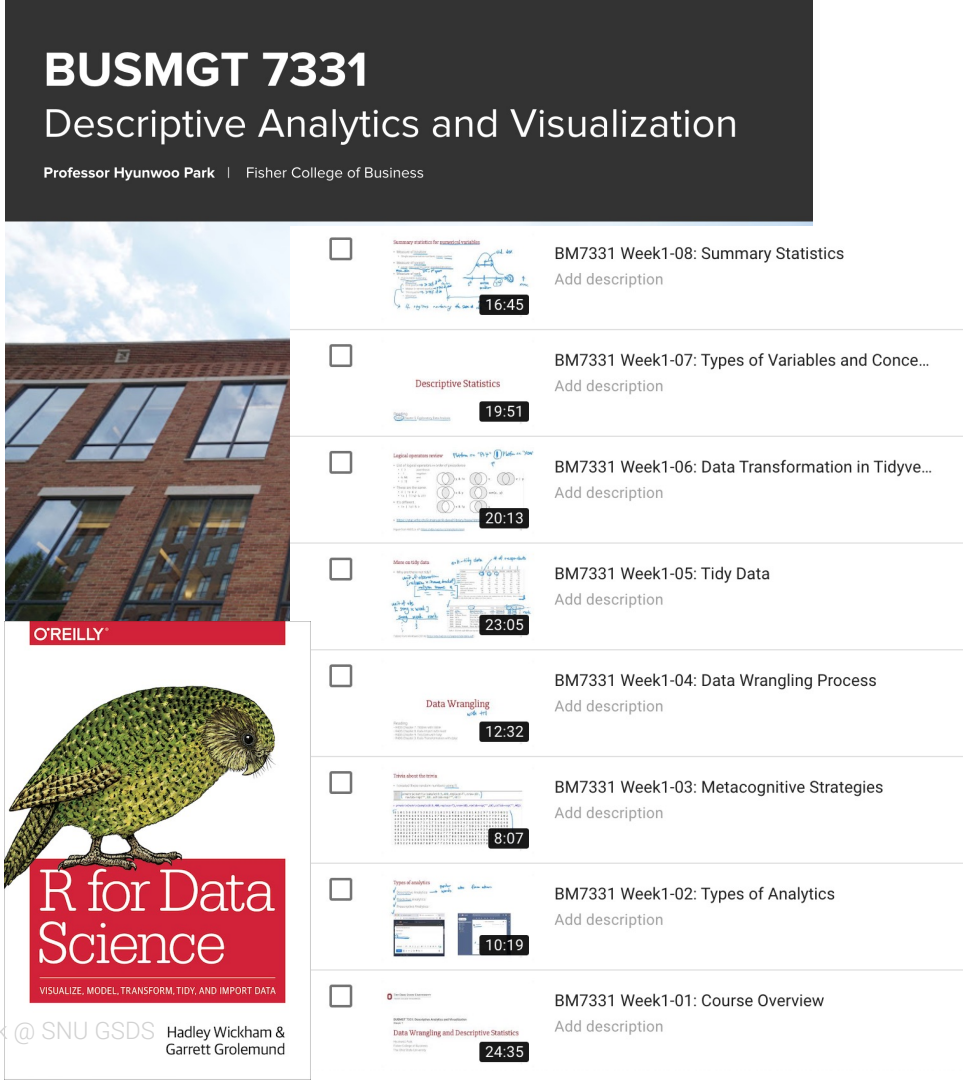
After all

- I am a new faculty at SNU Graduate School of Data Science who started last year.
- It means this course is still in the making.
- The downside is that there will be inevitably some rough corners and uncertainties here and there.
- The upside is that you will become part of making this course and laying the foundation. I welcome your feedback on how I should design and run this course even during the semester.
- Please be respectful to each other (including me) and I hope all of us enjoy this course this semester.

Course Overview

Brief history of the course

- I developed “Descriptive Analytics and Visualization” course for the Fisher College of Business at The Ohio State University.
- I taught the course for three times, and it became a very popular course there.
- The problem is that the course was developed to teach visualization in R. (By the way, R is also a great choice for data visualization for common charts and building a dashboard. Try it out if you are interested.)



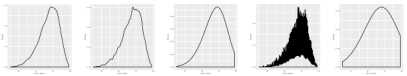
What students learned in that course

- Topic 1: Data Wrangling and Descriptive Statistics
- Topic 2: More Data Wrangling, Covariation, and Customizing Visualization
- Topic 3: Describing Textual Data
- Topic 4: Geospatial Visualization
- Topic 5: Network Visualization
- Topic 6: Interactivity and Dashboard Design
- Topic 7: Visualizing Models, Dimensionality Reduction, and Clustering

Density plot

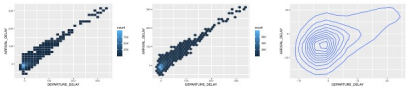
- Density plot smooths out and normalizes a histogram with moving window of width called "bandwidth".

```
ggplot(df) + geom_density(aes(x=Critic_Score))
ggplot(df) + geom_density(aes(x=Critic_Score), bw=1)
ggplot(df) + geom_density(aes(x=Critic_Score), bw=10)
ggplot(df) + geom_density(aes(x=Critic_Score), adjust=.1)
ggplot(df) + geom_density(aes(x=Critic_Score), adjust=10)
```



[3] Counting after binning: bin2d, hex, density_2d

```
ggplot(more_filtered_flights) +
  geom_bin2d(aes(x=DEPARTURE_DELAY, y=ARRIVAL_DELAY))
ggplot(more_filtered_flights) +
  geom_hex(aes(x=DEPARTURE_DELAY, y=ARRIVAL_DELAY))
ggplot(more_filtered_flights) +
  geom_density_2d(aes(x=DEPARTURE_DELAY, y=ARRIVAL_DELAY))
```

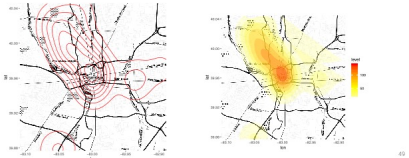


Word cloud

```
library(wordcloud)
wc <- tidy_amz %>% count(word, sort=T)
wordcloud(wc$word, wc$n, min.freq=50, colors=brewer.pal(5, "Dark2"))
?wordcloud
```

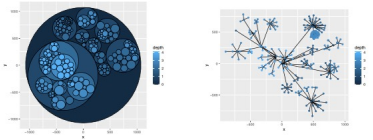
Visualizing density with contour and heatmap

```
basemap + geom_density2d(data=sbohll2, aes(lon, lat), color="red")
basemap +
 stat_density2d(data=sbohll2, aes(fill=..level..), geom="polygon", alpha=.3)+
 scale_fill_gradient2(low="white", mid="yellow", high="red", midpoint=50)
```

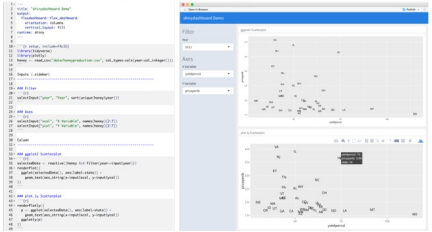


Circlepack layout

```
ggraph(grf, "circlepack", weight="size") + geom_node_circle(aes(fill =
depth), size = 0.25, n = 50) + coord_fixed()
ggraph(grf, "circlepack", weight="size") + geom_edge_link() +
geom_node_point(aes(color=depth)) + coord_fixed()
```

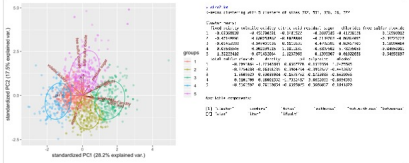


Interactive data exploration with shinydashboard



k-means clustering in R

```
wine2.km <- kmeans(scale(wine2 %>% select(-highq,-quality)), 5)
wine2.km
ggbiplot(wine2.pca, alpha=.3, groups=as.factor(wine2.km$cluster), ellipse=T)
```



Converting the content for Python and adding d3.js

- Python is a great choice for data visualization if you want more fine-grained customization in your visualization.
- Python is also great for streamlining your data visualization with other computer science or software engineering code.
- Python can also build a web server for an interactive visual analysis system using web programming and d3.js.
- Credits and special thanks go to Georgia Tech Vis Group friends.


Rahul Basole


John Stasko


Alex Endert


Polo Chau


Chad Stolper

Two-track system (back in 2021)

- Python
  - Data collection and cleaning
  - Matplotlib
  - Various high-quality static visualization with customizations
  - Dashboard

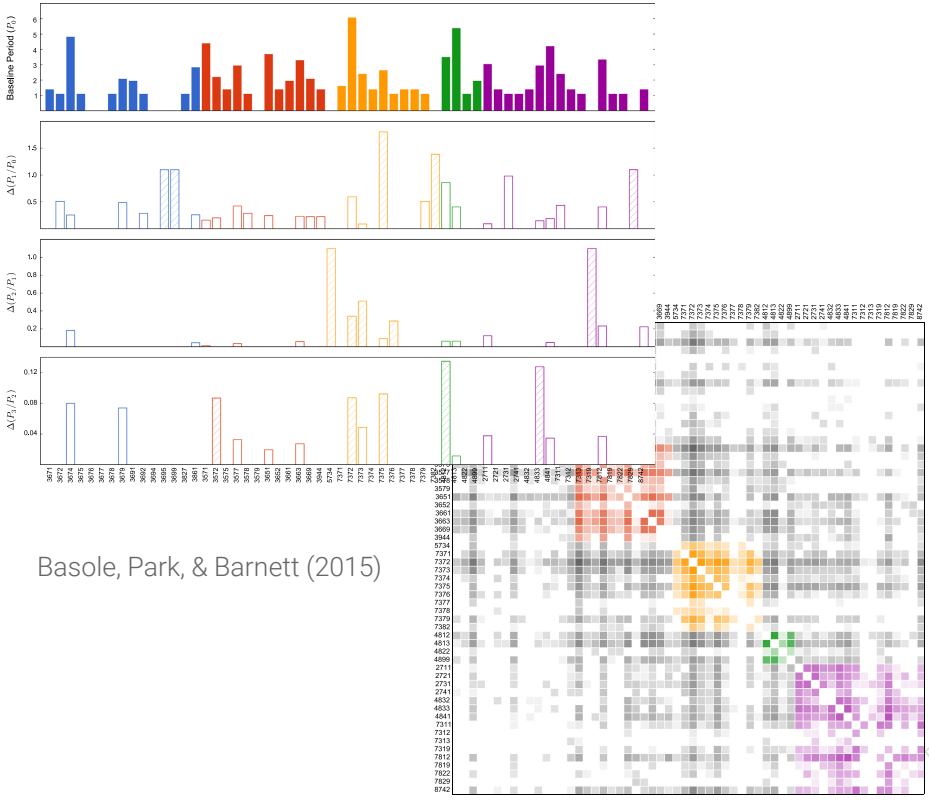
  - Evaluation: Homework + Midterm

- d3.js
  - Web programming basics
  - Back-end server with Python
  - Front-end HTML, CSS, JavaScript
  - Front-end framework: Bootstrap
  - d3.js as a JavaScript library

  - Evaluation: Group Project + Midterm

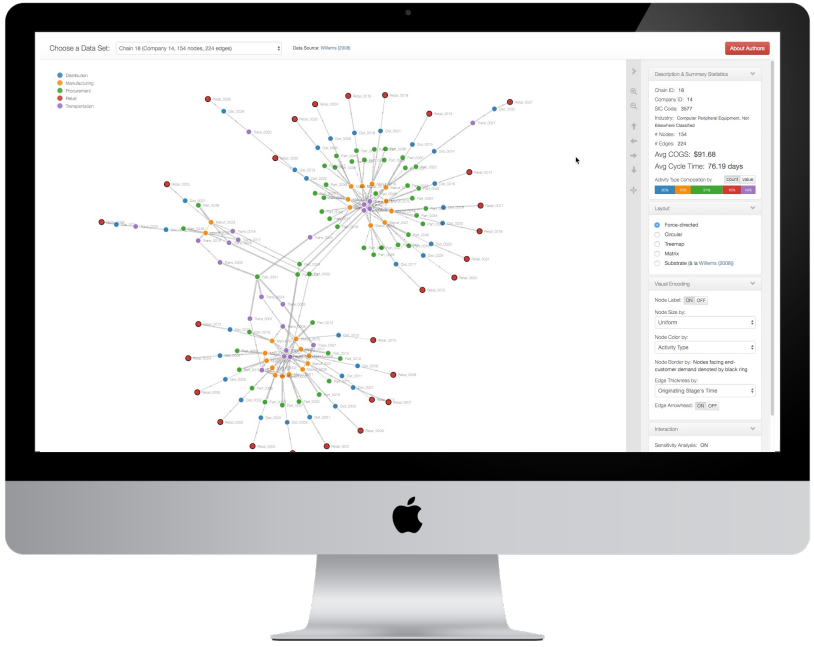
My aspiration

- Python



Basole, Park, & Barnett (2015)

- d3.js

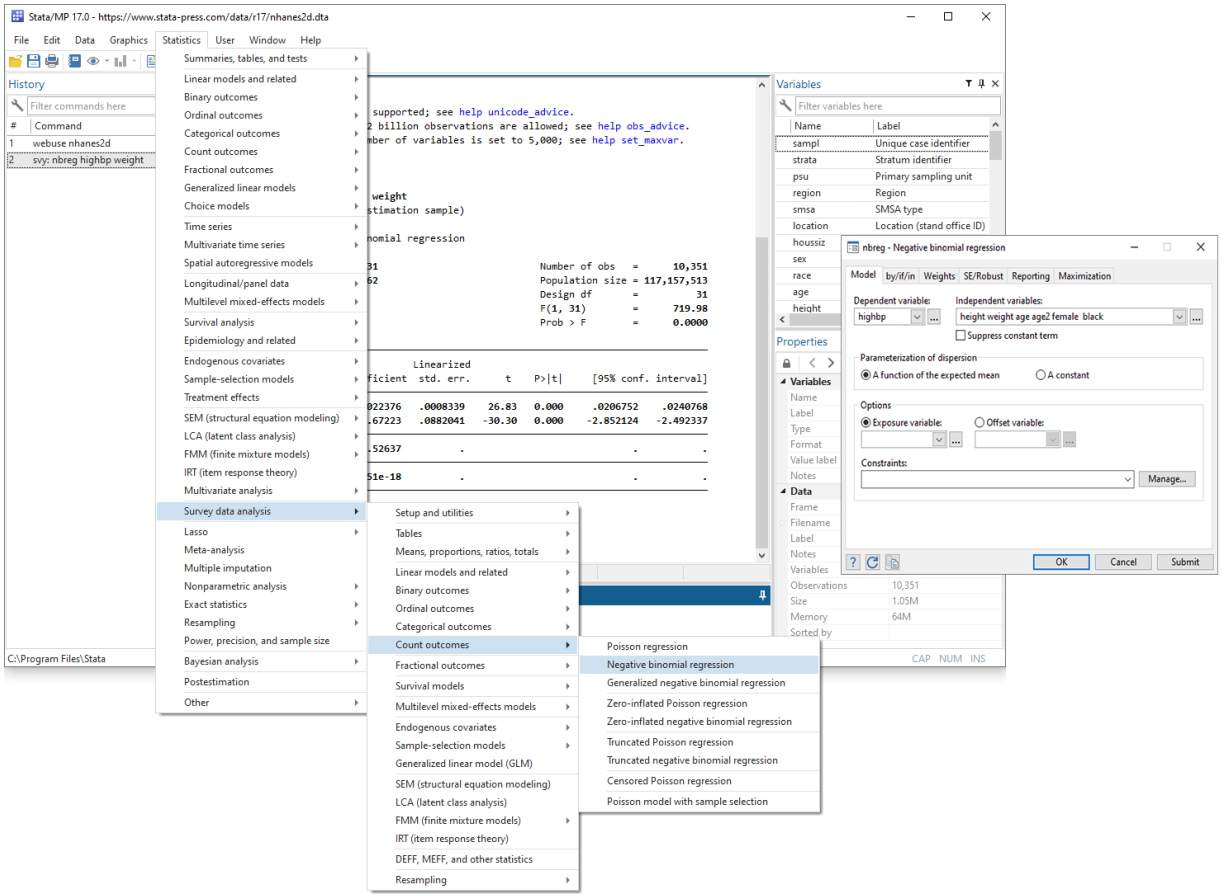


https://visualscm.herokuapp.com/

Park, Bellamy, & Basole (2016)

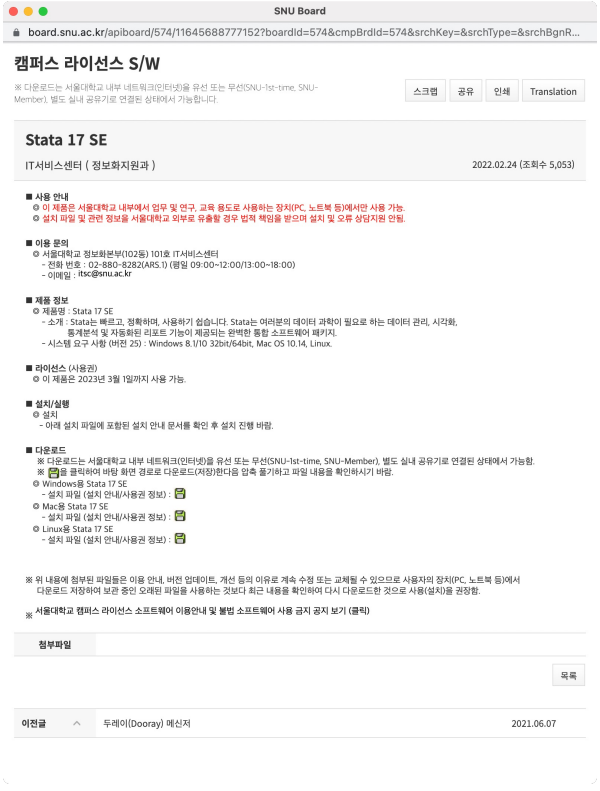
Three-track system (NEW in 2022)

- Stata (New in 2022)
- Python
- D3.js



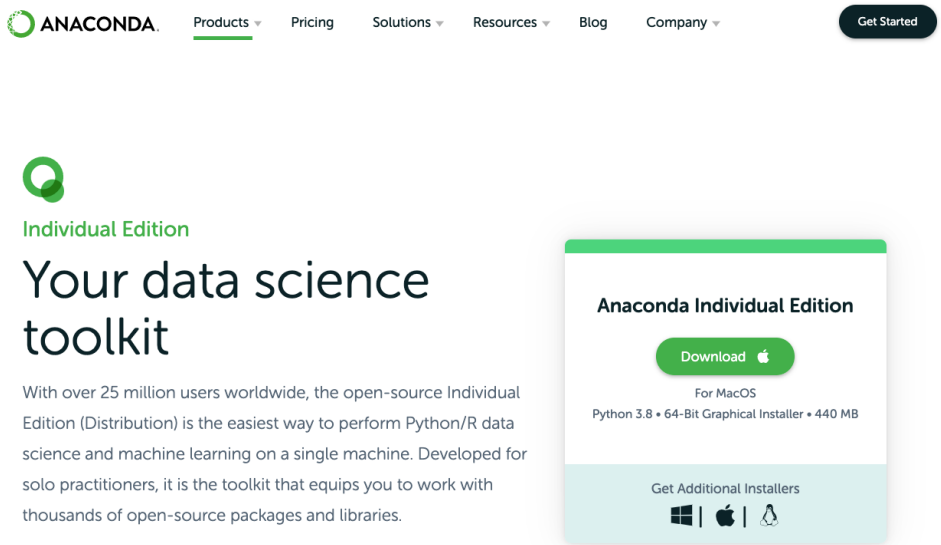
Things To Do

Install Stata

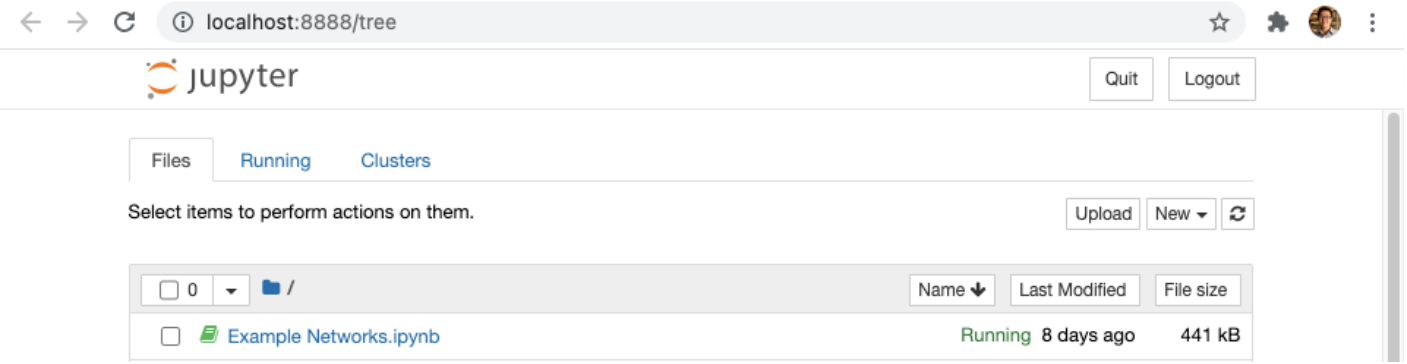


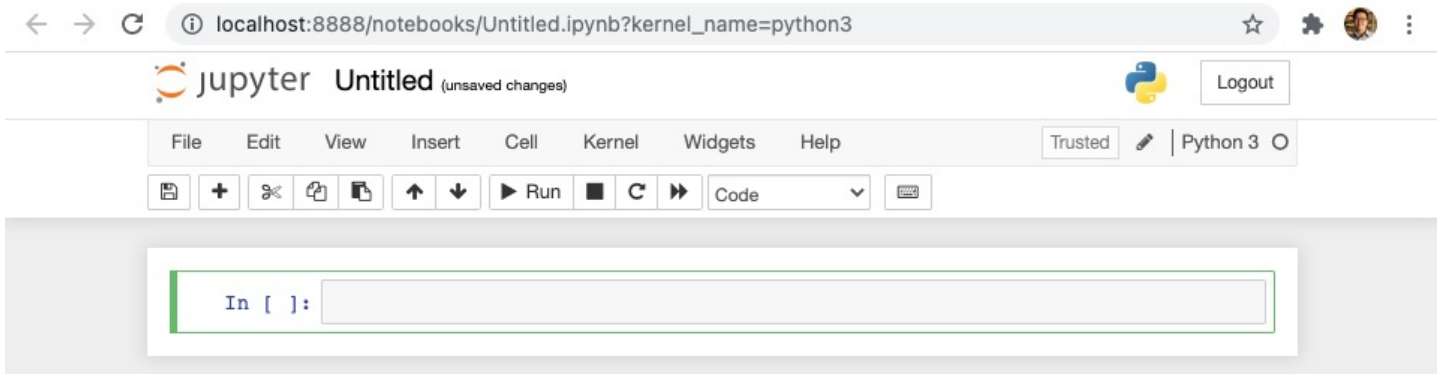
Install Python 3 Anaconda distribution

- https://www.anaconda.com/products/individual
- Python 3.8 (Version 2021.05)
- Included packages:
  - https://docs.anaconda.com/anaconda/packages/py3.8_osx-64/



Launch Jupyter Notebook



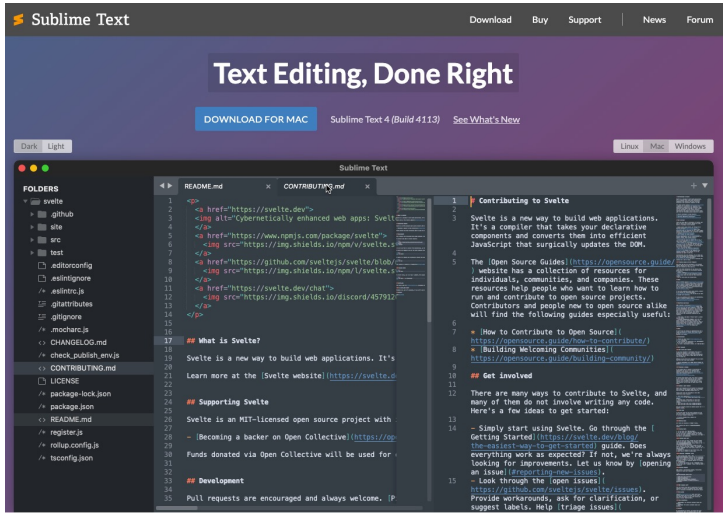


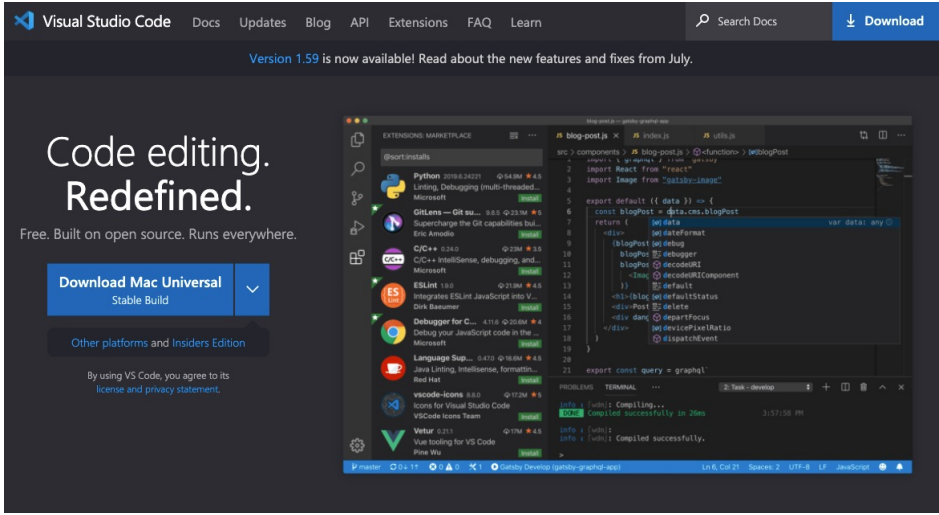
Get to know the terminal interface

```
(base) MacBook-Pro:python_code oksure$ which python
/Users/oksure/opt/anaconda3/bin/python
(base) MacBook-Pro:python_code oksure$ which jupyter
/Users/oksure/opt/anaconda3/bin/jupyter
(base) MacBook-Pro:python_code oksure$ jupyter console
Jupyter console 6.4.0

Python 3.8.8 (default, Apr 13 2021, 12:59:45)
Type 'copyright', 'credits' or 'license' for more information
IPython 7.22.0 -- An enhanced Interactive Python. Type '?' for help.

In [1]:
Do you really want to exit ([y]/n)?
Shutting down kernel
(base) MacBook-Pro:python_code oksure$ [IPKernelApp] WARNING | Parent appears to have exited, shutting down.

(base) MacBook-Pro:python_code oksure$ jupyter notebook
[I 2021-09-02 12:11:37.438 LabApp] JupyterLab extension loaded from /Users/oksure/opt/anaconda3/lib/python3.8/site-packages/jupyterlab
[I 2021-09-02 12:11:37.438 LabApp] JupyterLab application directory is /Users/oksure/opt/anaconda3/share/jupyter/lab
[I 12:11:37.442 NotebookApp] Serving notebooks from local directory: /Users/oksure/Dropbox/Research_SupplyNetwork/python_code
[I 12:11:37.442 NotebookApp] Jupyter Notebook 6.3.0 is running at:
[I 12:11:37.442 NotebookApp] http://localhost:8888/?token=88cbcfda9fdb3afc5d675d6f61b1a7c551e5493321c13a2d
[I 12:11:37.442 NotebookApp]  or http://127.0.0.1:8888/?token=88cbcfda9fdb3afc5d675d6f61b1a7c551e5493321c13a2d
[I 12:11:37.442 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 12:11:37.461 NotebookApp]

    To access the notebook, open this file in a browser:
        file:///Users/oksure/Library/Jupyter/runtime/nbserver-61601-open.html
    Or copy and paste one of these URLs:
        http://localhost:8888/?token=88cbcfda9fdb3afc5d675d6f61b1a7c551e5493321c13a2d
     or http://127.0.0.1:8888/?token=88cbcfda9fdb3afc5d675d6f61b1a7c551e5493321c13a2d
[0] 1:bash  2:bash- 3:bash  4:python3.8*                                    "MacBook-Pro.local" 12:11 02-Sep-21
```

Pick your code/text editor

- Sublime Text
  - https://www.sublimetext.com/
- VS Code
  - https://code.visualstudio.com/





Why Visualization?

Why visualization?

- How many 7's are there?

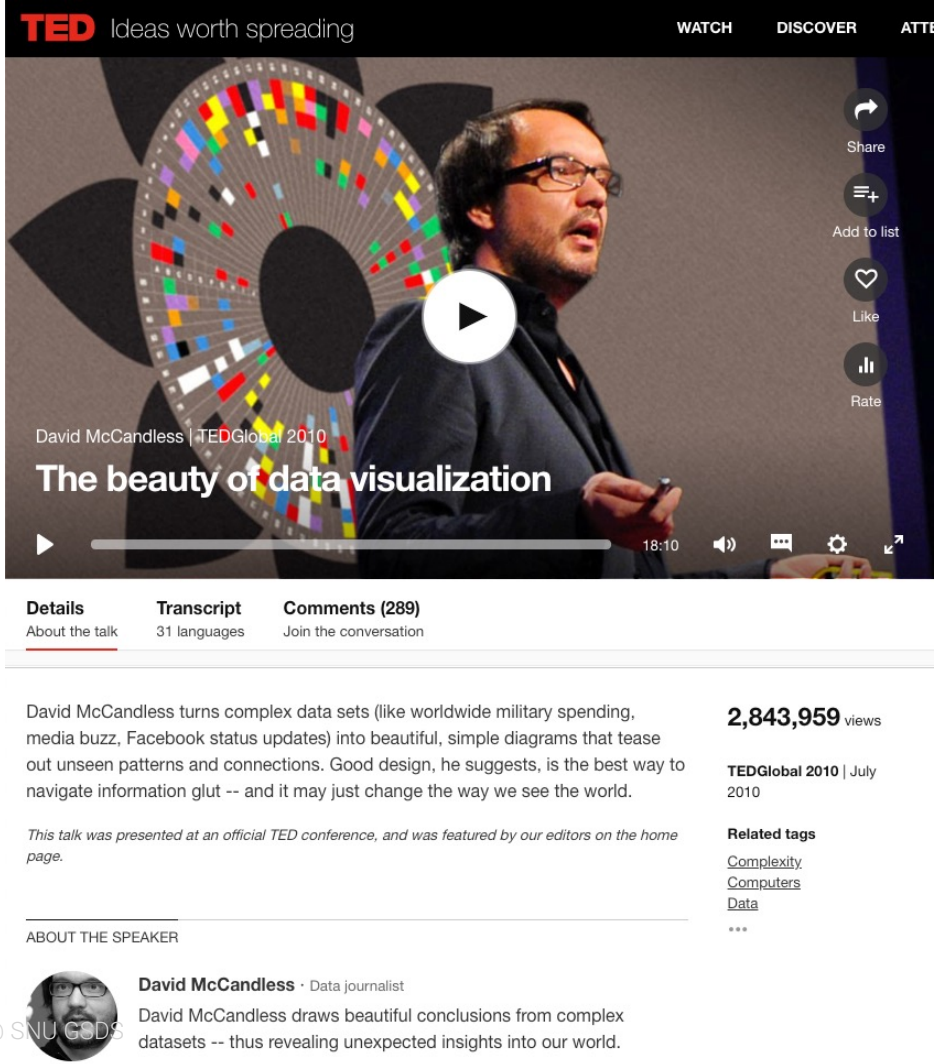
6 1 0 1 5 6 3 8 3 5 3 8 3 2 3 1 8 1 9 1 0 2 1 6 3 3 8 1 6 3 2 9 3 1 8 9 5 0 5 1
4 9 8 9 4 6 5 5 9 5 9 8 5 1 9 3 0 6 1 8 4 0 0 1 4 9 8 3 1 5 9 4 5 3 3 9 3 1 0 2
4 6 1 5 6 3 0 3 6 9 3 7 9 4 8 2 4 8 3 3 4 4 8 6 9 8 0 0 6 0 5 2 1 2 3 9 8 4 1 5
3 3 5 1 5 4 0 1 9 3 3 1 5 7 3 1 9 8 1 5 3 6 9 2 4 2 1 6 5 8 2 5 7 1 5 0 5 6 4 8
9 0 5 9 3 9 4 1 2 2 3 5 2 9 5 9 9 2 3 2 3 6 9 4 2 3 9 0 2 5 3 4 0 8 4 3 5 8 9 8
4 8 1 6 2 6 3 2 3 3 3 3 4 1 9 9 2 2 5 6 4 3 3 2 2 1 9 6 6 4 0 3 9 0 1 2 0 0 2 9
6 3 0 3 2 3 6 0 3 1 2 6 6 3 5 8 2 3 3 3 5 8 0 8 9 9 4 1 2 7 8 3 3 5 6 8 3 4 6 3
4 9 9 0 6 4 3 4 4 5 5 3 0 0 5 3 0 3 5 7 0 6 1 8 0 1 0 0 6 1 2 2 9 8 8 6 6 2 3 6
9 8 1 1 3 5 6 3 8 5 9 4 9 4 2 3 3 1 3 2 6 1 3 6 2 6 8 0 9 3 5 9 8 1 0 4 9 2 9 1
1 0 5 2 2 4 2 0 9 0 3 8 0 3 6 3 3 2 5 4 9 1 4 1 1 4 1 5 8 5 5 3 8 2 3 6 2 2 3 9

Power of visualization

- How about now?

Power of human vision

- Unparalleled information processing capacity compared to other senses
  - Sight: 1250MB/s
  - Touch: 125MB/s
  - Hearing / smelling: 12.5MB/s
- https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization or https://youtu.be/pLqjQ55tz-U
- An article about his talk: https://www.theguardian.com/science/punctuated-equilibrium/2010/dec/30/2



Trivia about the trivia

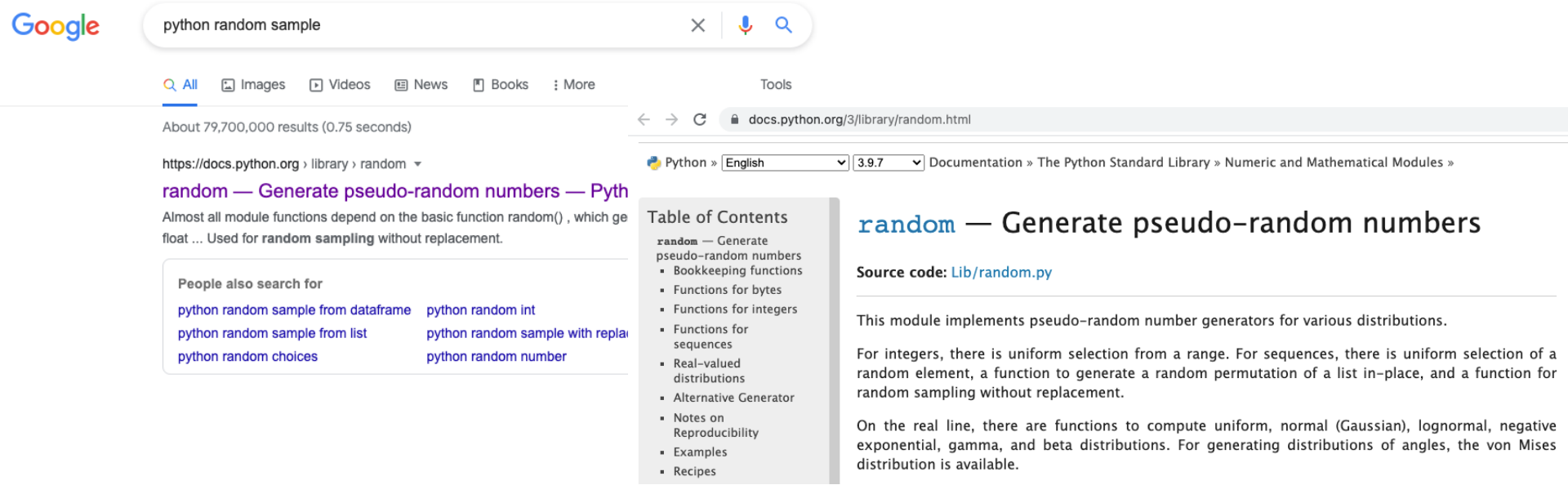
- I created these random numbers using Python.

```
1  import random
2  a = random.choices(range(10), k=400)
3  for i in range(10): print(' '.join(str(v) for v in a[i*40:(i+1)*40]))
```

```
In [6]: for i in range(10): print(' '.join(str(v) for v in a[i*40:(i+1)*40]))
5 6 1 0 9 9 2 2 8 4 6 0 1 2 6 7 8 3 8 4 3 2 6 4 9 7 9 7 5 3 4 8 9 0 0 9 4 2 6 0
6 8 5 1 4 6 6 8 9 5 7 2 0 5 6 5 3 0 7 5 1 4 7 3 8 6 7 5 9 3 7 9 5 9 4 6 6 4 8 4
0 0 2 6 5 7 7 1 8 4 9 4 7 3 9 7 9 4 4 1 8 0 9 0 1 8 3 6 0 3 0 8 6 7 1 3 3 7 4 6
6 2 7 4 3 1 9 5 2 9 8 8 6 1 4 2 7 4 6 9 6 8 0 8 8 2 2 3 2 8 1 9 1 9 7 3 9 8 9 4
7 9 1 8 7 5 8 8 6 1 1 4 3 8 5 4 6 3 1 4 5 9 3 4 4 7 8 1 5 2 0 8 1 2 1 2 2 5 9 5
5 8 7 6 0 0 7 2 9 3 2 8 5 6 6 7 2 0 1 9 0 2 0 7 8 5 0 7 1 4 9 5 5 6 3 5 8 4 9 7
7 9 2 1 3 2 0 7 2 5 0 2 5 3 0 6 9 0 9 8 8 3 7 7 2 3 7 6 0 4 4 8 4 8 2 7 4 4 5 8
9 2 4 6 9 9 4 4 7 3 3 1 9 4 0 7 0 3 4 5 8 5 8 2 9 0 4 8 7 4 1 4 3 4 0 7 7 6 6 5
1 8 2 8 8 2 6 4 5 8 2 7 7 7 6 3 2 9 7 3 5 9 7 4 6 1 9 6 3 8 2 4 4 0 4 4 7 4 4 3
6 0 0 9 6 0 0 9 1 6 4 9 8 8 4 0 8 1 3 1 0 5 1 0 3 3 2 5 7 7 9 5 8 9 6 9 2 7 5 4
```

How did I figure this out?

- I didn't solve it right away. So I Googled.
- https://docs.python.org/3/library/random.html



Not quite there yet… And finally!

- What I wanted is to randomly sample integers between 0 and 9 “with replacement.”
- random.sample is sampling without replacement.
- I kept search the documentation and found there’s another function called “choices.”

```
>>> sample([10, 20, 30, 40, 50], k=4)    # Four samples without replacement
[40, 10, 50, 30]
```

Simulations:

```
>>> # Six roulette wheel spins (weighted sampling with replacement)
>>> choices(['red', 'black', 'green'], [18, 18, 2], k=6)
['red', 'green', 'black', 'black', 'red', 'black']
```

Metacognition

- Meaning
  - Thinking about own thinking
  - Knowing about own knowing
  - Awareness of one's awareness
- Metacognitive knowledge
  - (1) Content knowledge (declarative knowledge): Knowing what you know and what you don't know
  - (2) Task knowledge (procedural knowledge): Knowing the procedure to find a solution for the problem
  - (3) Strategic knowledge (conditional knowledge): Knowing what to do to learn how to solve the problem
- Articulating what you do NOT know is paramount
  in accelerating self-learning in the era of Googling everything.

Key Takeaway

- Knowing what to search for is really really important.
- Google is your friend. Stack Overflow is usually a Google’s go-to friend.
- Python documentation is another Google’s friend.

Thank you!