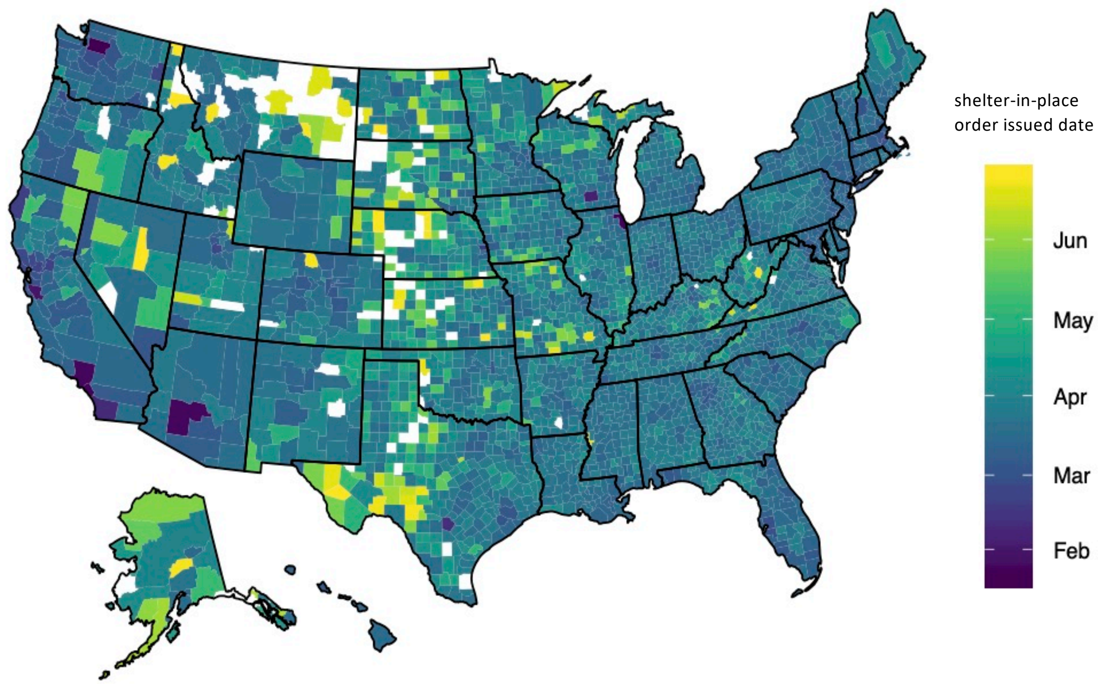
S&DS 177
YData: COVID-19 Behavioral Impacts



Youpei Yan
youpei.yan@yale.edu

Lecture 4 (Feb 25, Thursday 9:25 – 11:15)

Overview:

- Numpy vs. Pandas
- Panel Data
- More on plotting under dataframe
- Correlation vs. Causality

Lecture 4 (Feb 25, Thursday 9:25 – 11:15) 1. Review of Numpy

Let's visit a new dataset: number of visits in a county (**poi_vpc.csv**)

NAICS code of the industry

$$\text{dvc\#\#\#\#} = \frac{\text{total visits to an industry \#\#\#\#}}{\text{device count}}$$

Date & County ID

date	geoid	dvc7225	dvc7139	dvc4461	dvc4531	dvc4511	dvc4471	dvc7121	dvc4539	dvc4533	dvc4522	dvc4481
30jan2020	1001	0.306592	0.0731662	0.0649954	0.00649954	0.0222841	0.0766945	0.0252553	0.0222841	0.0250696	0.00278552	0.00464253
27jan2020	1001	0.237883	0.0792943	0.0759517	0.00761374	0.018013	0.0726091	0.0245125	0.0185701	0.0224698	0.00111421	0.00334262
31jan2020	1001	0.425255	0.0768802	0.0713092	0.00891365	0.0311978	0.102136	0.0207985	0.0293408	0.03974	0.00464253	0.00668524
02feb2020	1001	0.215227	0.0245125	0.0263695	0.00259981	0.0185701	0.0653668	0.0102136	0.0146704	0.0237697	0.00297122	0.00297122
28jan2020	1001	0.251253	0.0835655	0.0696379	0.00668524	0.0181987	0.0694522	0.0224698	0.0206128	0.0189415	0.00185701	0.00631383
01feb2020	1001	0.37883	0.105664	0.0488394	0.00408542	0.0451253	0.0854225	0.00464253	0.0356546	0.0529248	0.00427112	0.00742804
29jan2020	1001	0.261096	0.0662953	0.0705664	0.00501393	0.0193129	0.0670381	0.0157846	0.01987	0.0213556	0.00222841	0.00557103
27jan2020	1003	0.340055	0.129606	0.0494959	0.00902841	0.0603575	0.124381	0.0767644	0.0261687	0.0245188	0.0208066	0.034418
31jan2020	1003	0.535793	0.129239	0.0487168	0.0112282	0.0803391	0.154216	0.0703483	0.0385885	0.0293767	0.036297	0.0499083
29jan2020	1003	0.342301	0.105133	0.0405591	0.00692026	0.051879	0.103346	0.0692026	0.0261228	0.0218148	0.0181943	0.0291476

... (382064 rows omitted)

Lecture 4 (Feb 25, Thursday 9:25 – 11:15) 1. Review of Numpy

1. Why we clean the data with this additional step?

- We collect the trip information with people's smartphone devices. Some counties have 1000+ devices, but some counties have only several.

- This makes the aggregate trips biased, as we expect to see more trips in counties with more devices participating in the survey.

- To "normalize" the data, we use the visits per device count for all the counties. And that's also why we see 0.3 or 0.07 trip (rather than integers) in a day.

$$\text{dvc\#\#\#\#} = \frac{\text{total visits to an industry \#\#\#\#}}{\text{device count}}$$

Lecture 4 (Feb 25, Thursday 9:25 – 11:15) 1. Review of Numpy

1. Why we clean the data with this additional step?

- We collect the trip information with people's smartphone devices. Some counties have 1000+ devices, but some counties have only several.

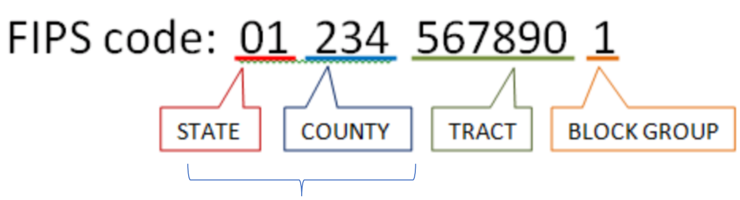
- This makes the aggregate trips biased, as we expect to see more trips in counties with more devices participating in the survey.

- To "normalize" the data, we use the visits per device count for all the counties. And that's also why we see 0.3 or 0.07 trip (rather than integers) in a day.

$$\text{dvc\#\#\#\#} = \frac{\text{total visits to an industry \#\#\#\#}}{\text{device count}}$$

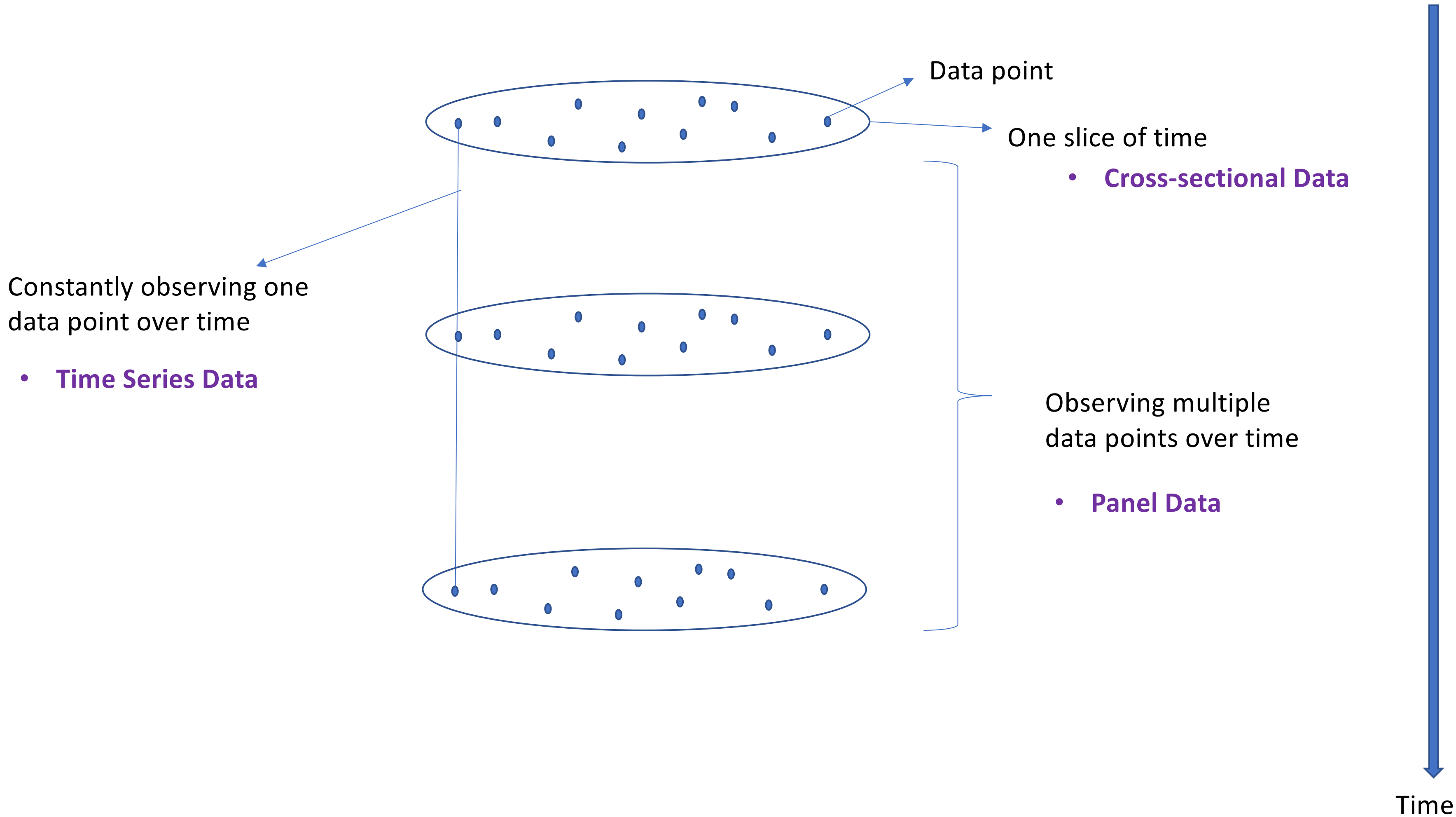
2. What are the geoid? How to tell the county and the state information based on it?

State ID = the integer part of (geoid/1000)
Or we use np.floor() function



Our geoid is the first 5 digits of a long code.
It's unique for each county

Lecture 4 (Feb 25, Thursday 9:25 – 11:15) 2. Panel Data & Pandas



Lecture 4 (Feb 25, Thursday 9:25 – 11:15) 2. Panel Data & Pandas

Function comparison between dataframe (Pandas) and arrays (Numpy)

- They have many similarities to deal with datasets, but they have their own advances:

Comparison	Pandas	NumPy
Works with	tabular data	numerical data
Tools	Series, DataFrame, etc.	Arrays
Performance	500K+ rows	50K rows or less
Memory	consume a lot	consume less
Objects	2d table object called DataFrame	amulti-dimensional array

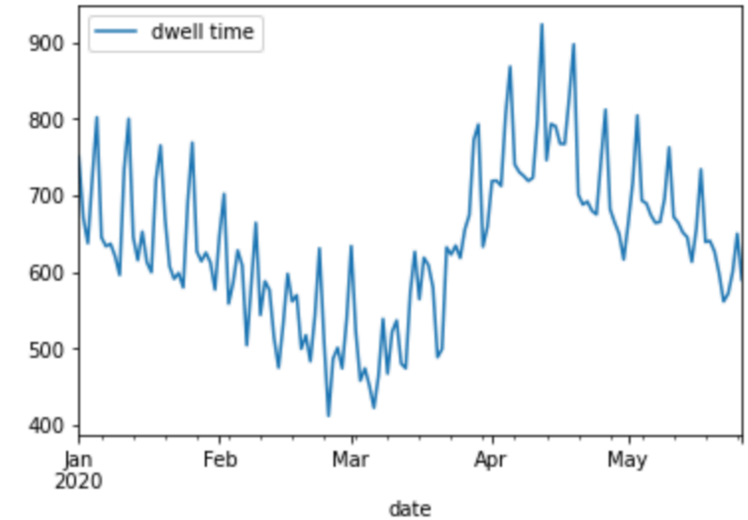
- There are several differences when we call a function for the same purpose of data management.
- We'll expand the following table gradually when we learn new techniques.

Purpose	Dataframe	Arrays
Import csv file	pd.read_csv('path')	Table.read_table('path')
Select rows based on values	Tablename.loc[option]	Tablename.where('var', option)
Sort	Tablename.sort_values('var')	Tablename.sort('var')
Find group mean	Tablename.groupby(['var']).mean()	Tablename.group('var', np.mean)
Add columns	Tablename['new var'] = …	Tablename.with_columns('var', val)

Lecture 4 (Feb 25, Thursday 9:25 – 11:15) 2. Panel Data & Pandas

Now let's practice!

Try to find the highest average home-dwelling time by date over the whole country, during the study time.



To do so, we need to tell the computer what to do step by step. Let's break the question into small pieces:

1. We need to group the data by date over the whole country. – which function?

2. We want to find the group mean. – which function?

3. We should sort the mean of home-dwelling time in a descending way. Or more, we can plot it. – which function?

Lecture 4 (Feb 25, Thursday 9:25 – 11:15) 3. More on data plotting

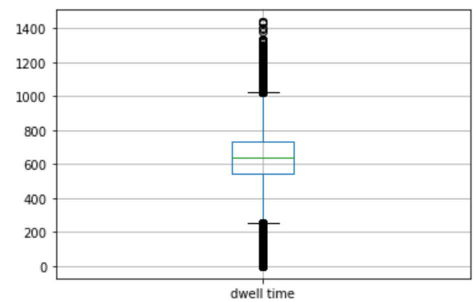
We’ve said it many times, but visualizing data is always the best way to provide straightforward information.

You may have seen a table of summary statistics, which usually looks like this:

Variable Name	# Observation	Mean	Standard Dev.	Min	Max
cases	456,288	124.5581	2053.61	0	204111
deaths	456,288	6.941337	186.4858	0	20795
dwell_time	456,288	637.243	159.5047	0	1438

Compare the 3rd row (dwell_time) in above table with the box plot here:

```
nytpd.boxplot(column=["dwell time"])
<matplotlib.axes._subplots.AxesSubplot at 0x7ffaa13629a0>
```



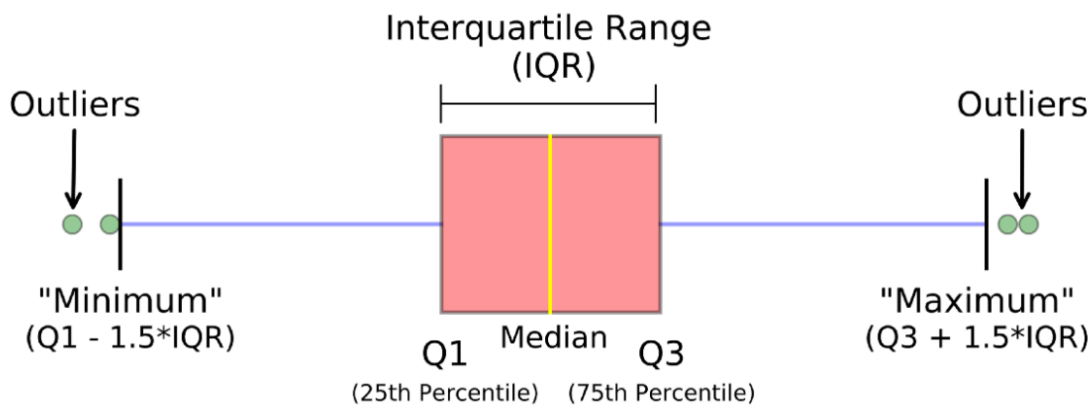
How to read boxplots?

What are Q1, Q3, and the inter-quantile range (IQR)?

What are ethe outliers?

Lecture 4 (Feb 25, Thursday 9:25 – 11:15) 3. More on data plotting

Boxplot:

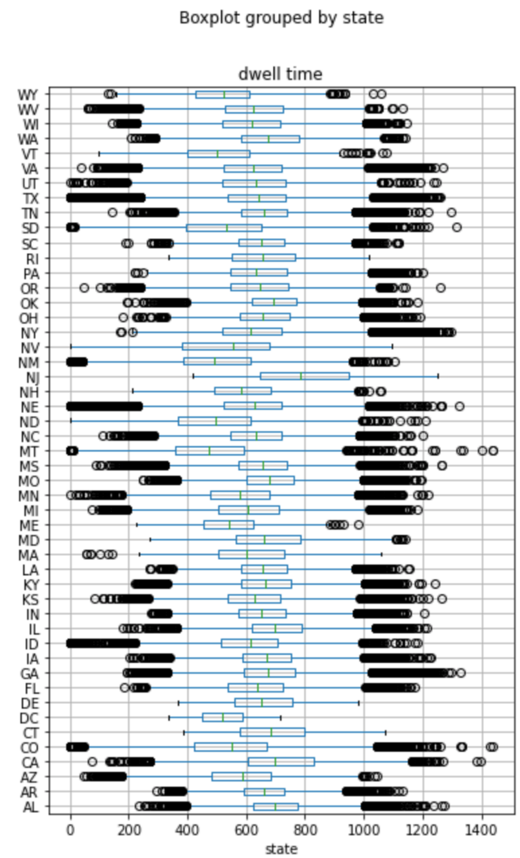


A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

Lecture 4 (Feb 25, Thursday 9:25 – 11:15) 3. More on data plotting

```
nytpd.boxplot(column=["dwell time"],by='state', vert=False, figsize=(6,10))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ffaa0cbc5e0>
```



We can always complicate the function to add more information in a figure, for instance, here we make the figure to show boxplot horizontally, specify the figure size, and group results by state.

- Which state has the highest median and IQR?
    - What could be the cause of it? (Recall what happened during the fist phase of COVID-19 for that state.)
- Which states have no outliers?
    - What could be the cause of it?

Lecture 4 (Feb 25, Thursday 9:25 – 11:15) 4. Correlation vs. Causality

- Sometimes, what we observe may not lead to a seemingly correct conclusion.
- Let’s go through some phenomenon and their conclusions first:
  - We observe a co-movement of cases and average dwelling home time.
    - (Y/N) More time to stay at home causes higher COVID-19 cases.
    - (Y/N) Higher cases outside made people choose to stay at home more voluntarily.
    - (Y/N) The policy maker observes increasing COVID-19 cases, and decide to have lockdown. People have no where to go but to stay at home.
    - (Y/N) Hypothetically, a natural disaster (say hurricane) hit the country when the cases started to increase, people hide in to avoid hurricane.

  - We observe the visits to grocery stores are increased when the COVID-19 cases increased.
    - (Y/N) More visits to grocery stores lead to a higher infection rate.
    - (Y/N) People are scared of the increasing case numbers and decided to store more food at home. So the visits to grocery stores would be reduced over the long run.
    - (Y/N) Other local stores are closed because of lockdown, and force people to visit certain grocery stores more.
    - (Y/N) People just simultaneously prepare for a hurricane.

Lecture 4 (Feb 25, Thursday 9:25 – 11:15) 4. Correlation vs. Causality

- As you can see, even if we find a strong correlation between two variables A and B, it is still hard for us to decipher the cause-and-effect.
    - It could be A directly causes B.
    - It could be B directly causes A.
    - It could be A directly causes C, and C indirectly influence B, or vice versa.
    - It could be A and D both influencing B simultaneously, and D's impact is larger comparing to A's impact.
    - It could be an event E influence A and B both, but we only observe A and B while ignoring D.
    - Etc.

```
nytpd_dategroup.corr()
```

	cases	deaths	dwell time
cases	1.000000	0.996325	0.320942
deaths	0.996325	1.000000	0.271765
dwell time	0.320942	0.271765	1.000000

- Now, let's observe the correlation between variables.
- Write down two seemingly correct conclusions based on the observed relationship here.
- Then tell a story that can flip the conclusion.

Lecture 4 (Feb 25, Thursday 9:25 – 11:15)

Index of **YData177_Lab2.ipynb**

1. Review with Visiting Data
    1.1 Average visiting counts to points of interest
    1.2 Daily mean of visit count
    1.3 FIPS code of county and state
2. Panel Data
    2.1 Import New York Times data using pandas
    2.2 Sorting & Grouping
    2.3 Data Plotting
        2.3.1 Line plot
        2.3.2 Box plot
3. Correlation & Causality

```
dataframe.rename(columns={"A":"B"})
dataframe.sort_values('varname')
dataframe.groupby(['varname'])

dataframe.plot(y=["var1","var2","var3"],use_index = True)
dataframe.boxplot(column=["var1","var2","var3"])

dataframe.corr()
```