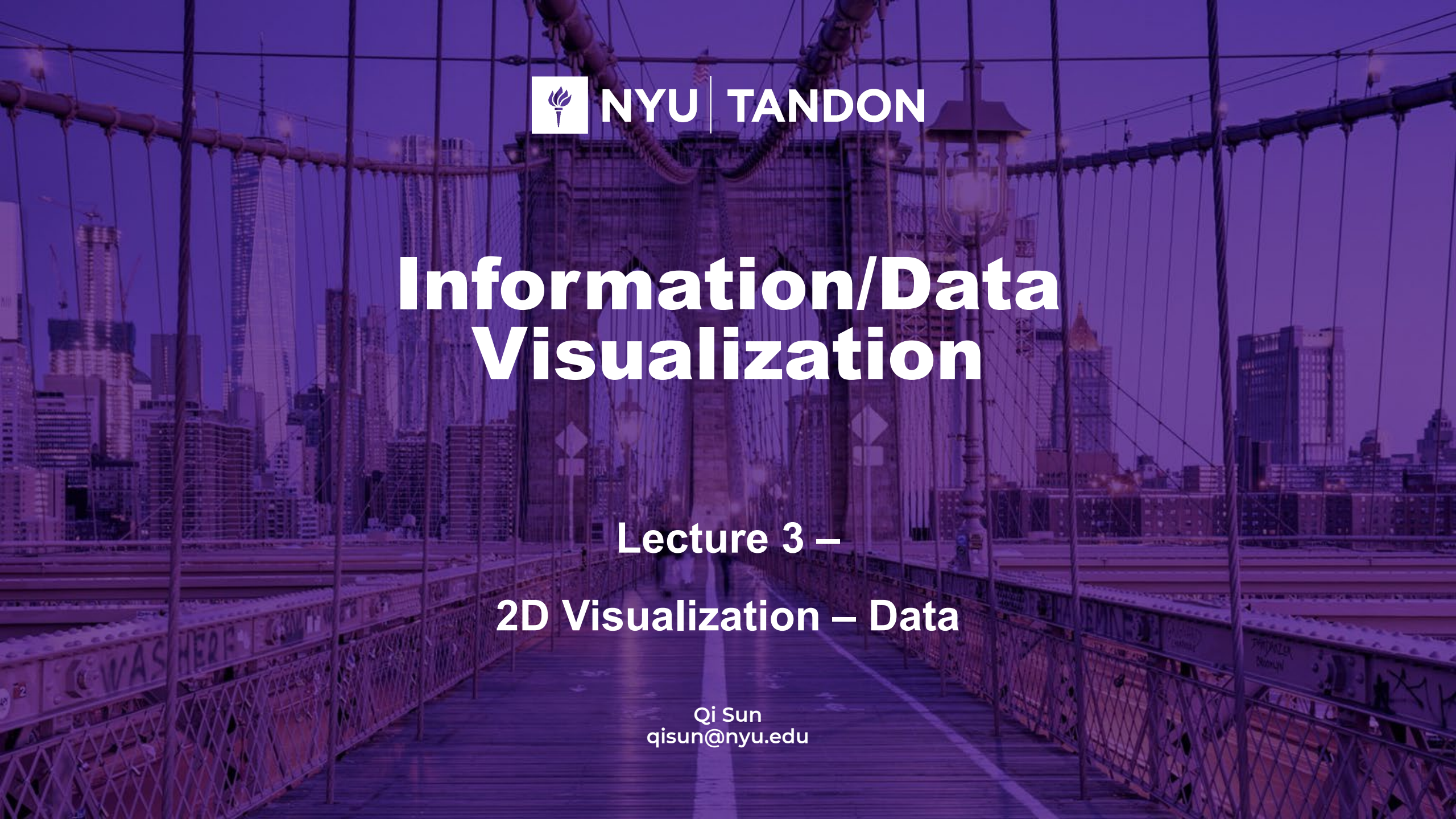


NYU TANDON

Information/Data Visualization

Lecture 3 –

2D Visualization – Data

Qi Sun
qisun@nyu.edu

Logistics

- Assignment 1
  - Do NOT use matplotlib pre-defined colormap

Where We Are

- Basic mathematical tools

- Color: definition, space, map, and design
- 2D visualization: data -> spatial, temporal, network

- Human perception and Geometries of 3D
- 3D visualization

- Application case studies

2D Data Visualization - Typology

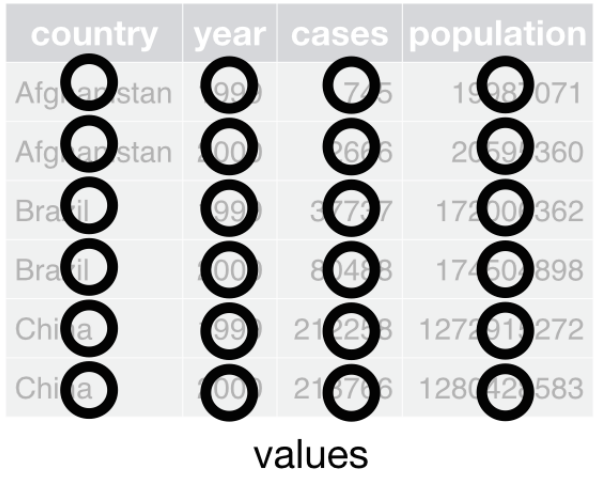
Insight needs	Data scales	Analyses	Visualizations	Graphic symbols	Graphic variables	Interactions
Categorize/cluster	Nominal	Statistical	Table	Geometric symbols	Spatial	Zoom
Order, rank, sort	Ordinal	Temporal	Chart	Point	Position	Search and locate
Distributions (also outliers)	Interval	Geospatial	Graph	Line	Retinal	Filter
Comparisons	Ratio	Topical	Map	Area	Form	Details on demand
Trends (process and time)		Relational	Tree	Surface	Color	History
Geospatial			Network	Volume	Optics	Extract
Compositions (also of text)				Linguistic symbols	Motion	Link and brush
Correlations/relationships				Text		Projection
				Numerals		Distortion
				Punctuation marks		
				Pictorial symbols		
				Images		
				Icons		
				Statistical glyphs		

2D Data Representation

- Each **variable** must have its own **column**.
- Each **observation** must have its own **row**.
- Each **value** must have its own **cell**.
- Each **type** of observational unit form a **table**.

Rule of thumb:

- Variables often map to aesthetics.
- Observations can be compared.



2D Data Representation

- **Variables**: measure some underlining attribute.
- **Observations**: collection of all measurements for variables related to a single unit.
- **Values**: represent the measurement of a variable in one observation.

2D Data – Representation and Processing

Country	Year	Population	Case
A	2020	10	1
A	2021	11	2
B	2020	50	3
B	2021	55	5

2D Data – Representation and Processing

Country	Year	Type	Value
A	2020	Population	10
A	2021	Population	11
A	2020	Case	1
A	2021	Case	2
B	2020	Population	50
…	…	…	…

2D Data – Representation and Processing

Country	Year	Rate
A	2020	1 / 10
A	2021	2 / 11
B	2020	3 / 50
B	2021	5 / 55

Data Representation

Table 1: Population

Country	Year	Population
A	2020	10
A	2021	11
B	2020	50
B	2021	55

Table 2: Cases

Country	Year	Case
A	2020	1
A	2021	2
B	2020	3
B	2021	5

Data Representation

Table 1: Population

Country	2020	2021
A	10	11
B	50	55

Table 2: Cases

Country	2020	2021
A	1	2
B	3	5

Data Representation

Student	Test 1	Test 2
John	80	95
Jane	90	92
Mary	75	90

Student	Test	Score
John	1	80
John	2	95
Jane	1	90
Jane	2	92
Mary	1	75
Mary	2	90

Data Representation

Country	Year	Population	Case
A	2020	10	1
A	2021	11	2
B	2020	50	3
B	2021	55	5

Country	Year	Rate
A	2020	1 / 10
A	2021	2 / 11
B	2020	3 / 50
B	2021	5 / 55

Data Representation

religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Data Representation

year	artist	track	time	date.entered	wk1	wk2	wk3
2000	2 Pac	Baby Don't Cry	4:22	2000-02-26	87	82	72
2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	91	87	92
2000	3 Doors Down	Kryptonite	3:53	2000-04-08	81	70	68
2000	98^0	Give Me Just One Nig...	3:24	2000-08-19	51	39	34
2000	A*Teens	Dancing Queen	3:44	2000-07-08	97	97	96
2000	Aaliyah	I Don't Wanna	4:15	2000-01-29	84	62	51
2000	Aaliyah	Try Again	4:03	2000-03-18	59	53	38
2000	Adams, Yolanda	Open My Heart	5:30	2000-08-26	76	76	74

year	artist	time	track	date	week	rank
2000	2 Pac	4:22	Baby Don't Cry	2000-02-26	1	87
2000	2 Pac	4:22	Baby Don't Cry	2000-03-04	2	82
2000	2 Pac	4:22	Baby Don't Cry	2000-03-11	3	72
2000	2 Pac	4:22	Baby Don't Cry	2000-03-18	4	77
2000	2 Pac	4:22	Baby Don't Cry	2000-03-25	5	87
2000	2 Pac	4:22	Baby Don't Cry	2000-04-01	6	94
2000	2 Pac	4:22	Baby Don't Cry	2000-04-08	7	99
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-02	1	91

Basic Data Operations

- Pivoting
- Separating & uniting
- Selecting & filtering
- Mutating
- Binding & joining
- Missing data

Pivoting

Wide form

Country	2020	2021	2022
A	10	11	12
B	50	55	60

Pivot to longer

Pivot to wider

Long form

Country	Year	Population
A	2020	10
A	2021	11
A	2022	12
B	2020	50
B	2021	55
B	2022	60

Pivoting

Country	2020	2021	2022
A	10	11	12
B	50	55	60

Pivot to longer

Country	Year	Population
A	2020	10
A	2021	11
A	2022	12
B	2020	50
B	2021	55
B	2022	60

Pivot to longer input:

- Columns involved (e.g. 2020, 2021, 2022)
- New column representing names. (e.g. “Year”)
- New column representing values. (e.g. “Population”)

Pivoting

Country	2020	2021	2022
A	10	11	12
B	50	55	60

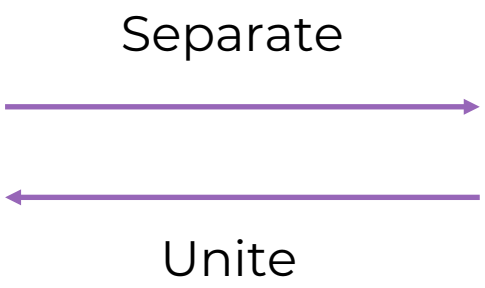
Country	Year	Population
A	2020	10
A	2021	11
A	2022	12
B	2020	50
B	2021	55
B	2022	60

Pivot to wider input:

- Column containing names. (e.g. “Year”)
- Column containing values. (e.g. “Population”)

Separating & Uniting

Country	Year	Rate
A	2020	1 / 10
A	2021	2 / 11
B	2020	3 / 50
B	2021	5 / 55



Country	Year	Population	Case
A	2020	10	1
A	2021	11	2
B	2020	50	3
B	2021	55	5

Separating & Uniting

Separate input:

- Column to be separated. (e.g. “Rate”)
- Names of new columns. (e.g. “Population”, “Case”)
- Separator. (e.g. “/” or 2)

Country	Year	Rate
A	2020	1 / 10
A	2021	2 / 11
B	2020	3 / 50
B	2021	5 / 55

Separate

Country	Year	Population	Case
A	2020	10	1
A	2021	11	2
B	2020	50	3
B	2021	55	5

Separating & Uniting

Unite input:

- New column name. (e.g. ”Rate”)
- Columns to be united. (e.g. “Population”, “Case”)
- Separator. (e.g. “/”)

Country	Year	Rate
A	2020	1 / 10
A	2021	2 / 11
B	2020	3 / 50
B	2021	5 / 55

Unite

Country	Year	Population	Case
A	2020	10	1
A	2021	11	2
B	2020	50	3
B	2021	55	5

Separating & Uniting

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

Selecting

Select input:

- A set of columns. (e.g. “Country”, “Year”, “Population”)

Country	Year	Population	Case
A	2020	10	1
A	2021	11	2
B	2020	50	3
B	2021	55	5

Select

Country	Year	Population
A	2020	10
A	2021	11
B	2020	50
B	2021	55

Filtering

Filer input:

- An expression that evaluates to TRUE or FALSE per row.
  E.g. year == 2020

Country	Year	Population	Case
A	2020	10	1
A	2021	11	2
B	2020	50	3
B	2021	55	5

Filter

Country	Year	Population	Case
A	2020	10	1
B	2020	50	3

Mutating

Mutate for adding new columns:

- E.g. century = floor(year / 100) + 1

Country	Year	Population
A	2020	10
A	2021	11
B	2020	50
B	2021	55

Mutate →

Country	Year	Century	Population
A	2020	21	10
A	2021	21	11
B	2020	21	50
B	2021	21	55

Mutating

Mutate for changing existing columns:

- E.g. population = population * 1,000,000

Country	Year	Population
A	2020	10
A	2021	11
B	2020	50
B	2021	55

Mutate →

Country	Year	Population
A	2020	10,000,000
A	2021	11,000,000
B	2020	50,000,000
B	2021	55,000,000

Binding

Country	Year	Population
A	2020	10
B	2020	50
B	2021	55

Bind rows

Country	Year	Population
A	2020	10
A	2021	11
B	2020	50
B	2021	55

Binding

Countr y	Year
A	2020
A	2021
B	2020
B	2021

Population
10
11
50
55

Bind columns →

Country	Year	Population
A	2020	10
A	2021	11
B	2020	50
B	2021	55

Joining

Joining two tables by variables, i.e. variables used to identify same rows in each table.

Country	Year	Population
A	2020	10
A	2021	11
B		
B		

Country	Year	Case
A	2020	1
A	2021	2

Join

Country	Year	Population	Case
A	2020	10	1
A	2021	11	2
B	2020	50	3
B	2021	55	5

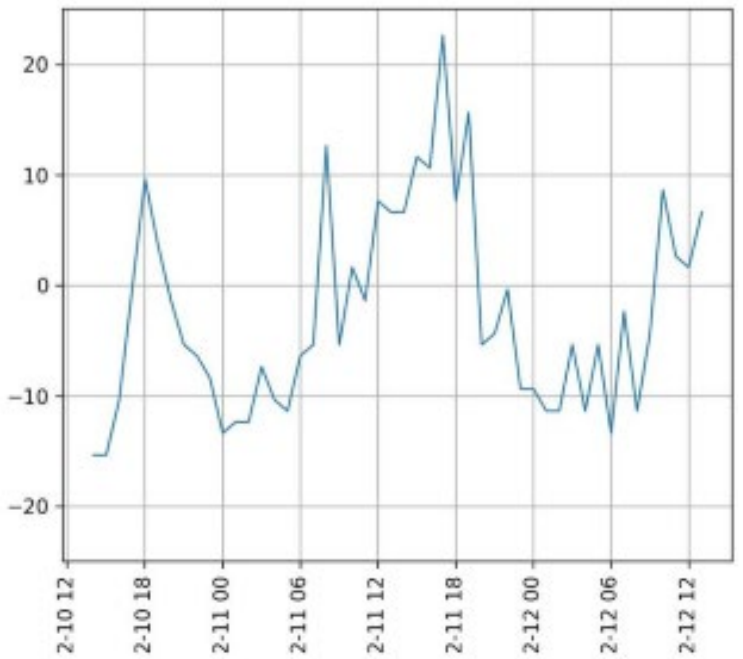
Missing Data

- Explicitly missing data: often marked with special value such as “N/A”.
- Implicitly missing data: empty field.
- Often need to convert implicitly missing data to explicitly missing data.

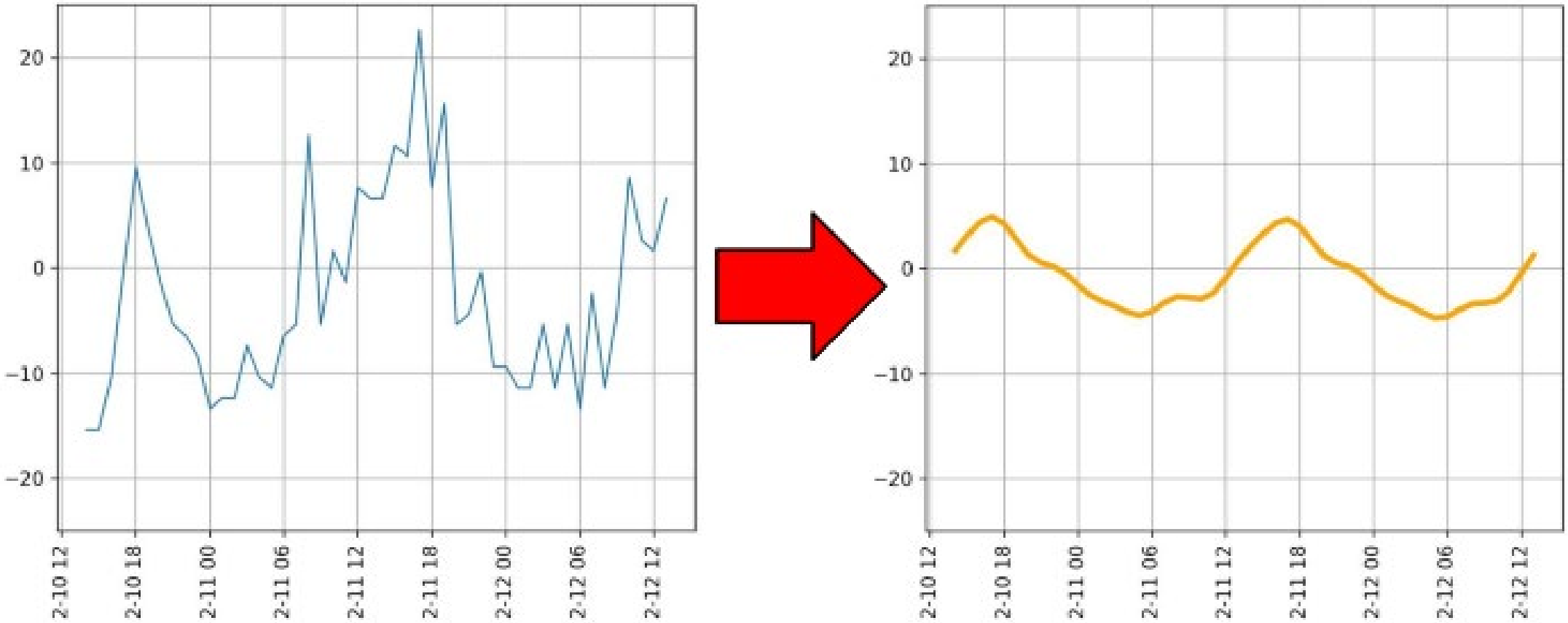
Country	Year	Population	Case
A	2020	10	1
A	2021	N/A	N/A
B	2020	50	3
B	2021	55	

Value Data Processing – Frequency

Country	Year	Population	Case
A	2020	10	1
A	2021	11	2
B	2020	50	3
B	2021	55	5



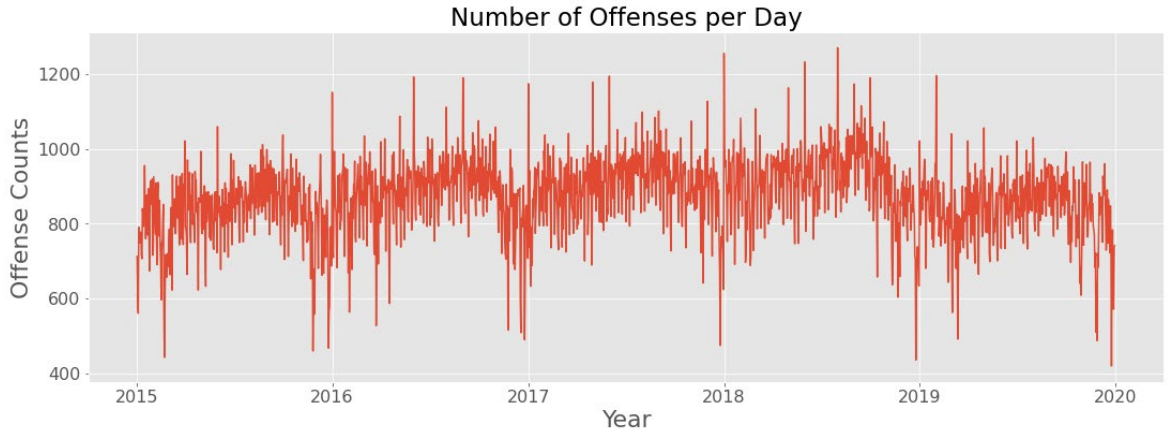
Discuss: Frequency-Based Data Filtering

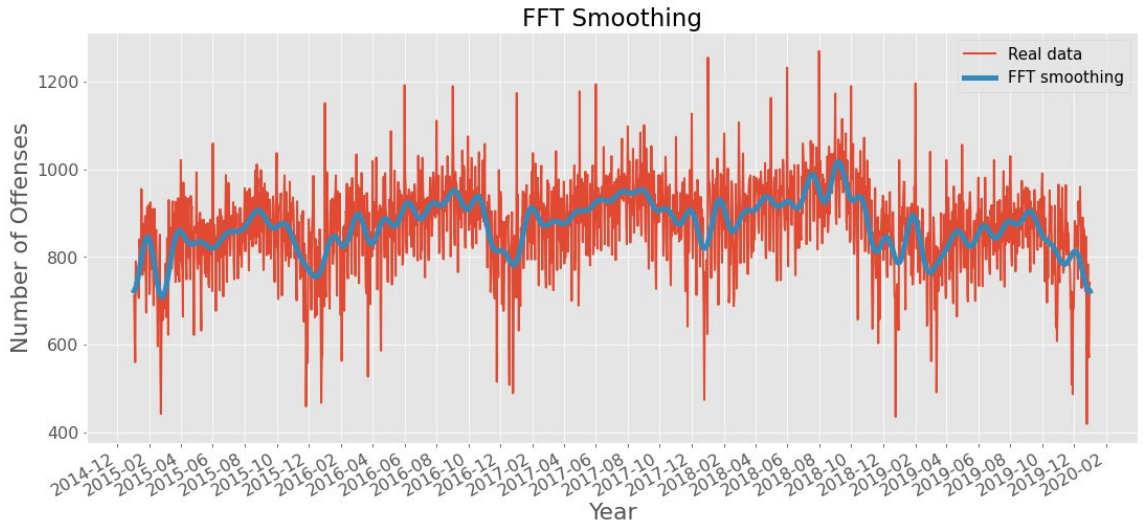


Discuss: Data Visualization Use Cases

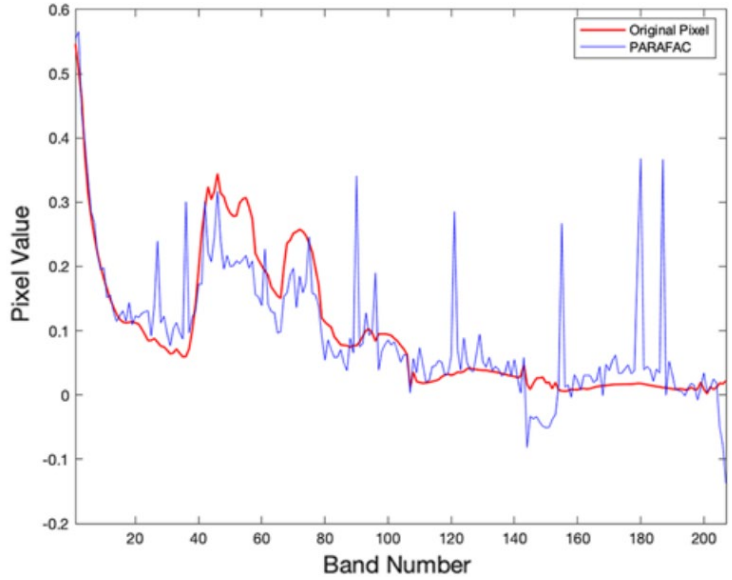
high frequency trading

crime analysis of noisy data





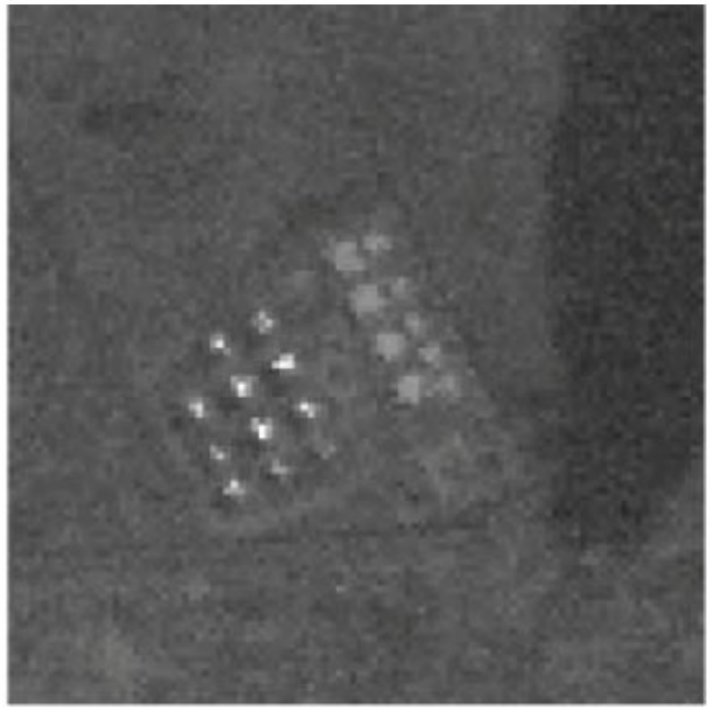
Urban Data Case Study



Ground Truth

Sensed

Processed



Python Implementation

```
>>> from scipy.fft import fft, ifft
>>> x = np.array([1.0, 2.0, 1.0, -1.0, 1.5])
>>> y = fft(x)
>>> y
array([ 4.5       +0.j        ,  2.08155948-1.65109876j,
       -1.83155948+1.60822041j, -1.83155948-1.60822041j,
        2.08155948+1.65109876j])
>>> yinv = ifft(y)
>>> yinv
array([ 1.0+0.j,  2.0+0.j,  1.0+0.j, -1.0+0.j,  1.5+0.j])
```