Descriptive Analytics

Exploring and Visualizing Data

Veterans Analytics Course

September 16-17, 2020

Provided by: CANA Advisors

Case Study
National Parks

- National park visitors by year
- Data fields:
  - Year
  - National park name
  - Region
  - State
  - Visitors
  - Visit Type

Data: structured vs. unstructured

- **Structured**
  - Lists, data frames, spreadsheets, databases, ‘big’ data
  - May contain -
    - Numeric values
    - Logicals (True/False)
    - Factors
    - Strings with set format
- **Unstructured**
  - Multiple formats (no rigid structure)
  - May contain -
    - Images
    - Free Text
    - Speech
    - Others

Data Attributes

- **Accuracy** – How correct is the data?
- **Confidence** – what are the ‘error bars’ around the data provided?
- **Authority** – how authoritative is the source of the data?

Question - Is it better to have Accurate or Authoritative data?

Tidy Data

Optimally organized data

Tidy data

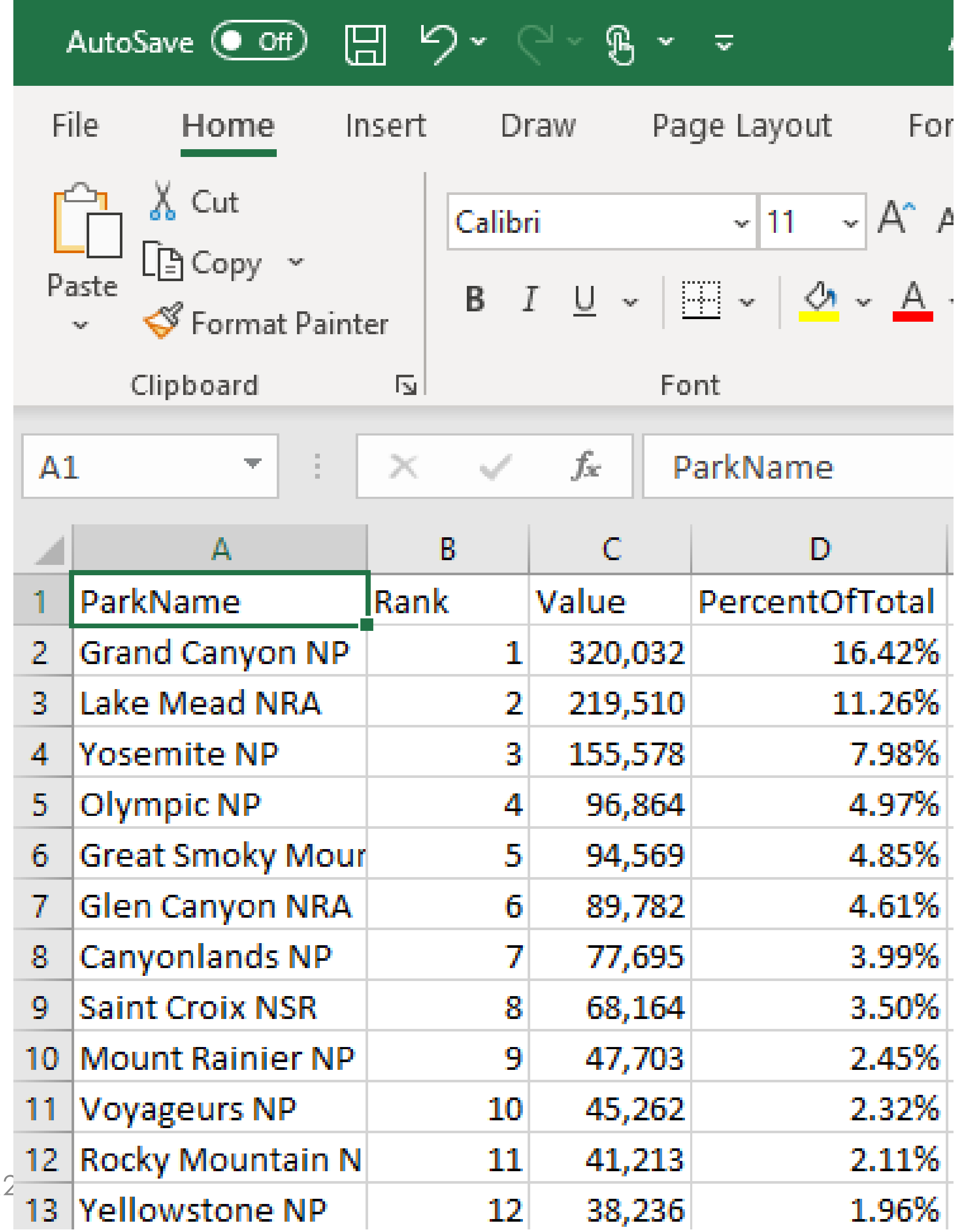
- Each variable must have its own column
- Each observation must have its own row
- Each value must have its own cell

For Excel, this means the first row is the header / list of field names and each row underneath is a record / observation

- No blank rows
- No merged cells

This data is tidy

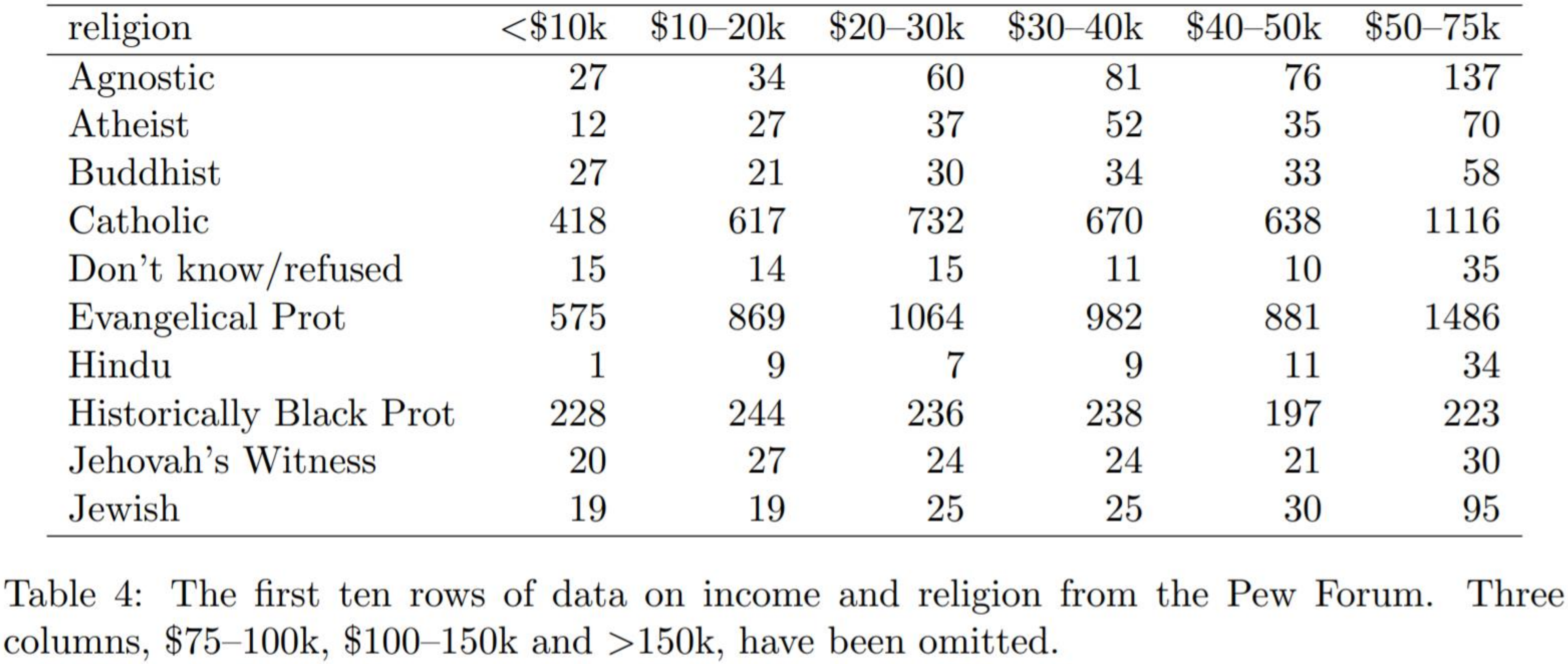
- Each variable has its own column
- Each observation has its own row
- Each value has its own cell



A1 ParkName

	A	B	C	D
1	ParkName	Rank	Value	PercentOfTotal
2	Grand Canyon NP	1	320,032	16.42%
3	Lake Mead NRA	2	219,510	11.26%
4	Yosemite NP	3	155,578	7.98%
5	Olympic NP	4	96,864	4.97%
6	Great Smoky Mour	5	94,569	4.85%
7	Glen Canyon NRA	6	89,782	4.61%
8	Canyonlands NP	7	77,695	3.99%
9	Saint Croix NSR	8	68,164	3.50%
10	Mount Rainier NP	9	47,703	2.45%
11	Voyageurs NP	10	45,262	2.32%
12	Rocky Mountain N	11	41,213	2.11%
13	Yellowstone NP	12	38,236	1.96%

Tidy or untidy?



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

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, $75–100k, $100–150k and >150k, have been omitted.

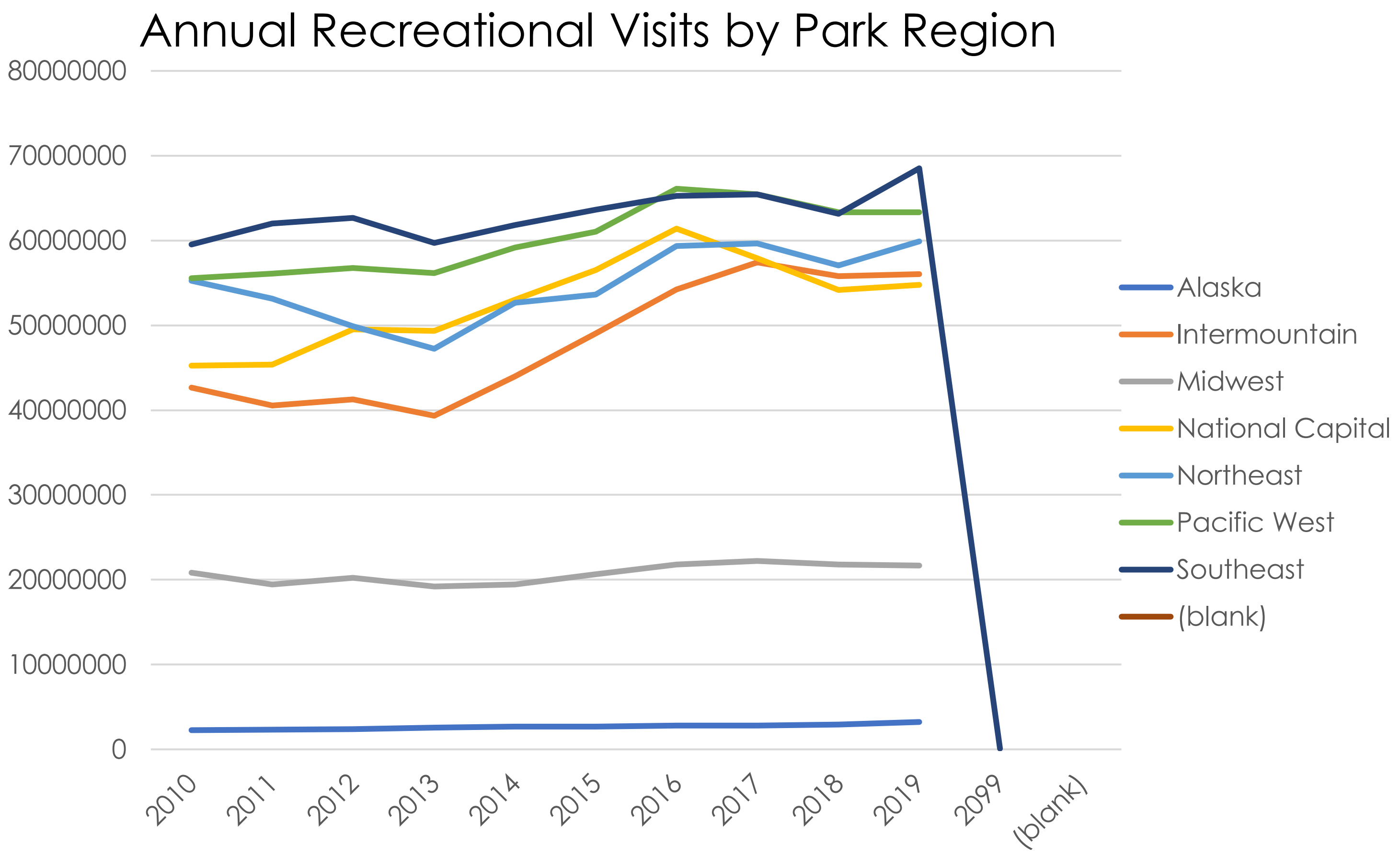
Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1 - 23. doi:http://dx.doi.org/10.18637/jss.v059.i10

Common data issues

- Improperly stored data
- Duplicate records
- Outliers
- Missing data
- Invalid entries

Garbage data, garbage visual

- What underlying issues impact this graph?



Summarizing Data

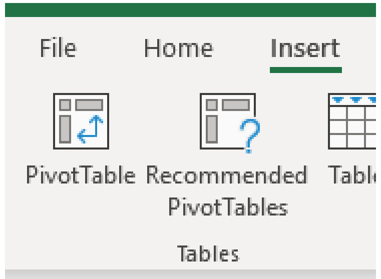
Slice and dice data with Pivot Tables

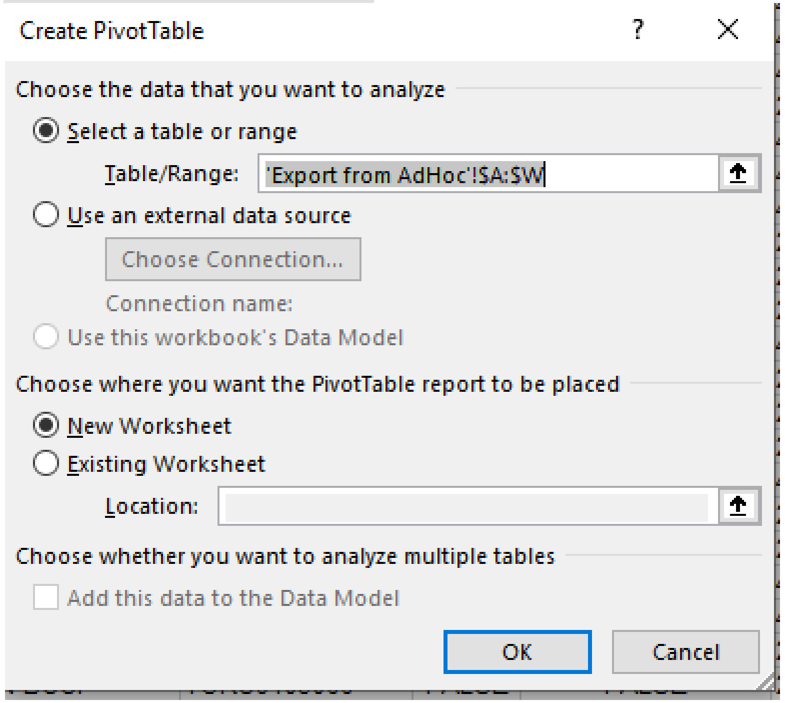
Functions compute summary statistics

- SUM()
- MIN()
- MAX()
- MEDIAN()
- AVERAGE()
- COUNTA()

PivotTables

- Allow you to quickly summarize data by groups
- Select Insert → PivotTable
- You typically want the default options:
  - Entire sheet as your range
  - PivotTable in a new window




Create PivotTable ? ×

Choose the data that you want to analyze

● Select a table or range

Table/Range: 'Export from AdHoc'!$A:$W

○ Use an external data source

Choose Connection...

Connection name:

○ Use this workbook's Data Model

Choose where you want the PivotTable report to be placed

● New Worksheet

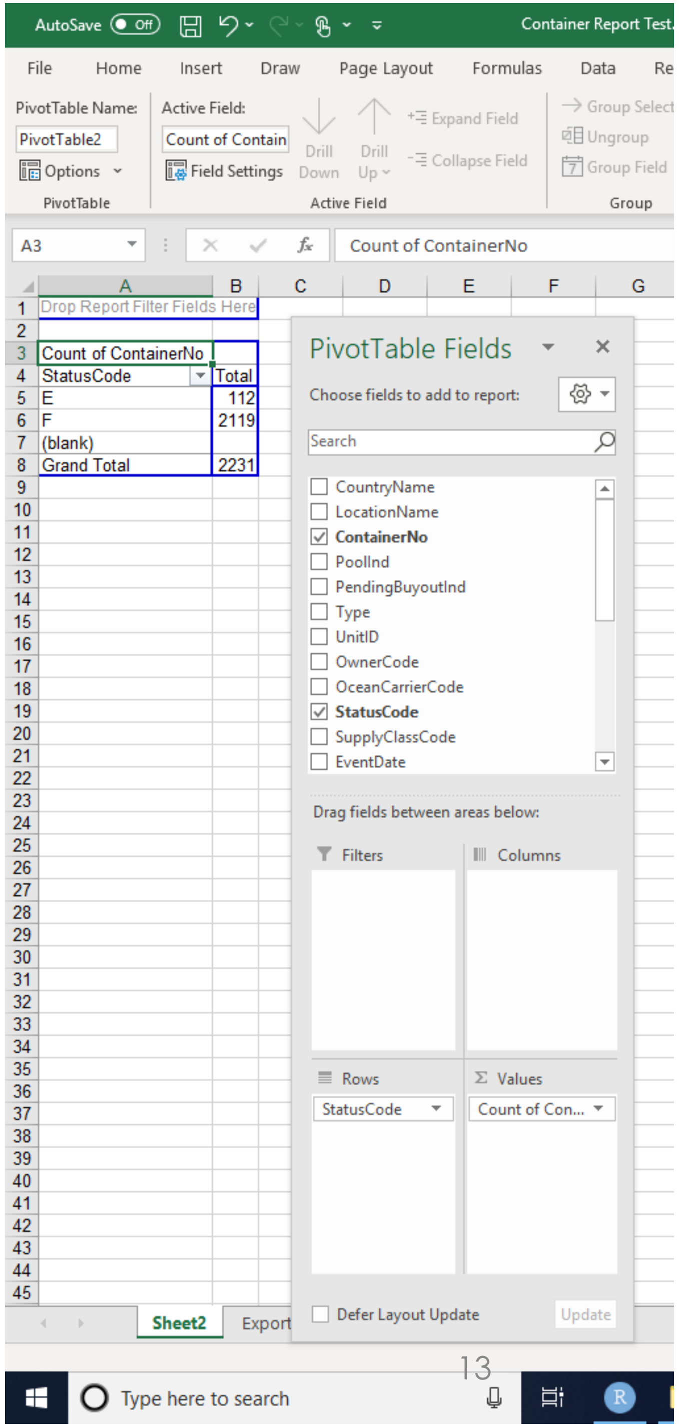
○ Existing Worksheet

Location:

Choose whether you want to analyze multiple tables

□ Add this data to the Data Model

OK Cancel



A3 fx Count of ContainerNo

	A	B
1	Drop Report Filter Fields Here	
2		
3	Count of ContainerNo	
4	StatusCode	Total
5	E	112
6	F	2119
7	(blank)	
8	Grand Total	2231

PivotTable Fields

Choose fields to add to report:

Search

□ CountryName
□ LocationName
☑ ContainerNo
□ PoolInd
□ PendingBuyoutInd
□ Type
□ UnitID
□ OwnerCode
□ OceanCarrierCode
☑ StatusCode
□ SupplyClassCode
□ EventDate

Drag fields between areas below:

Filters | Columns

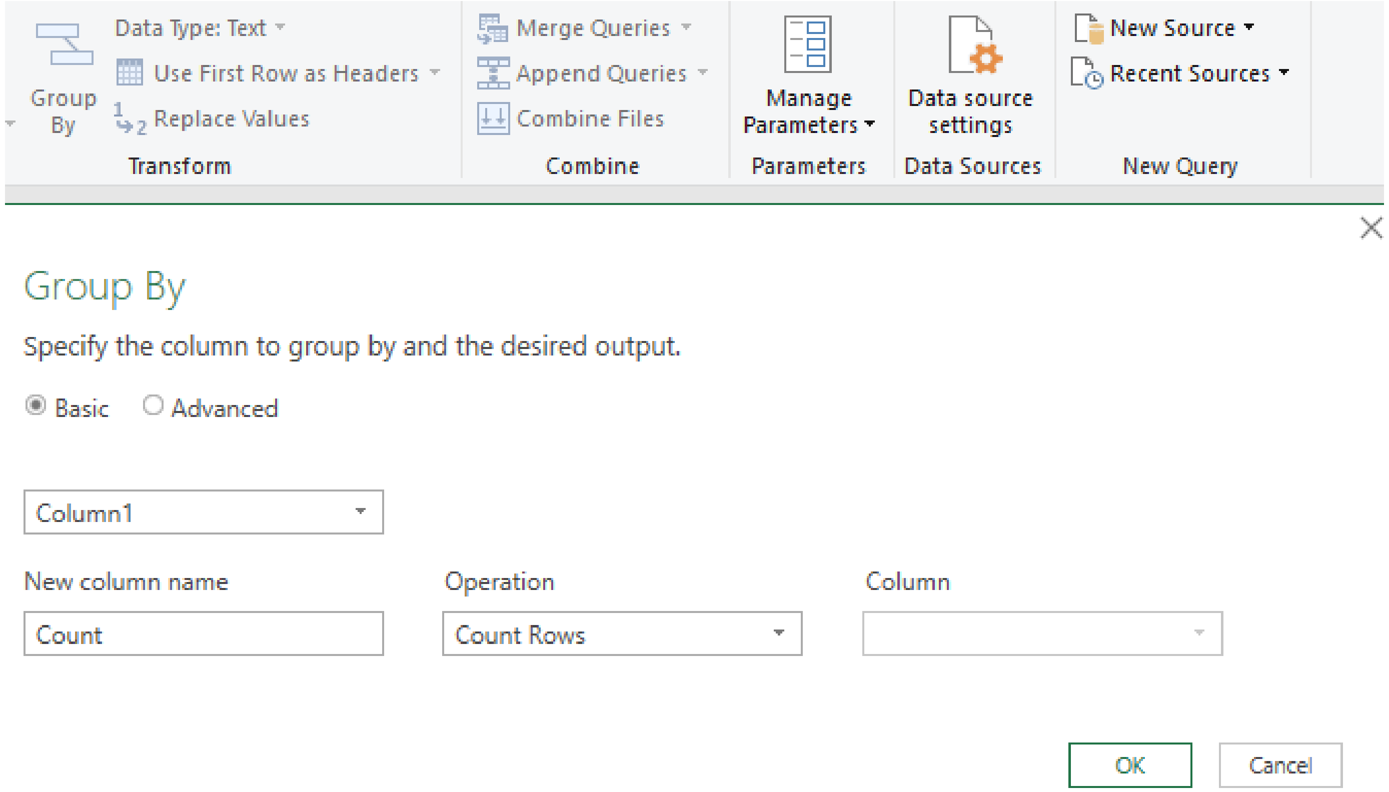
Rows: StatusCode

Values: Count of Con...

□ Defer Layout Update Update

Power Query + PivotTables

- Power Query offers an intuitive way to construct PivotTables



Group By

Specify the column to group by and the desired output.

● Basic ○ Advanced

Column1

New column name: Count

Operation: Count Rows

Column:

Practical Examples

Let’s see it in practice!

Visualizing Data

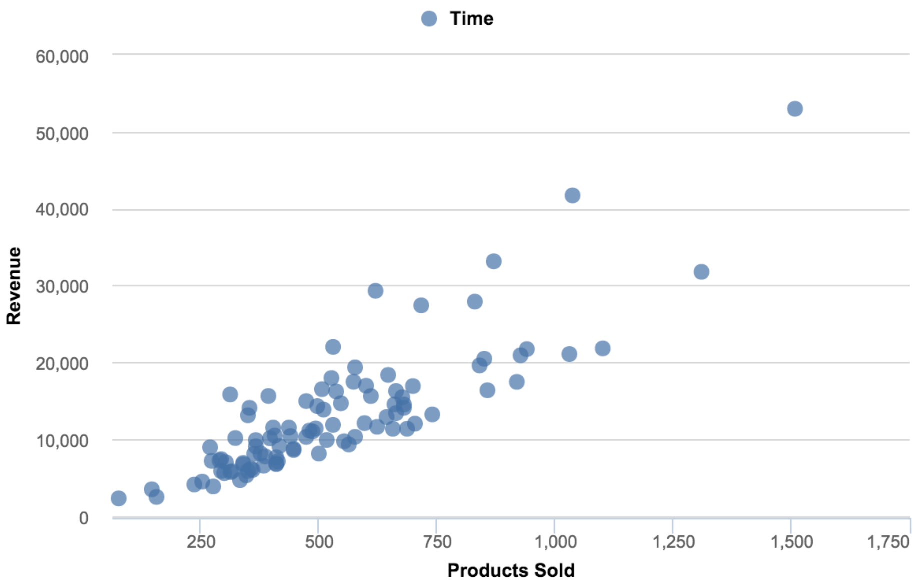
Uncovering relationships in complex data

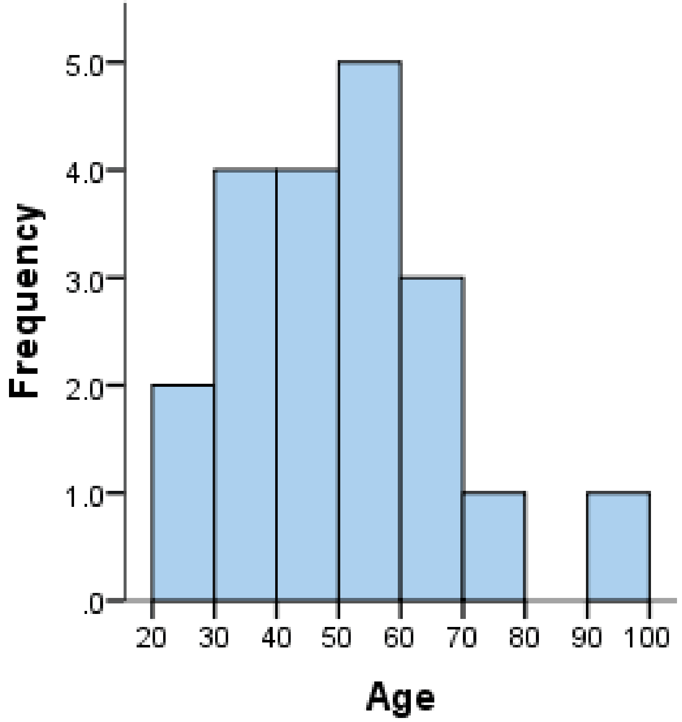
Discovering relationships in data

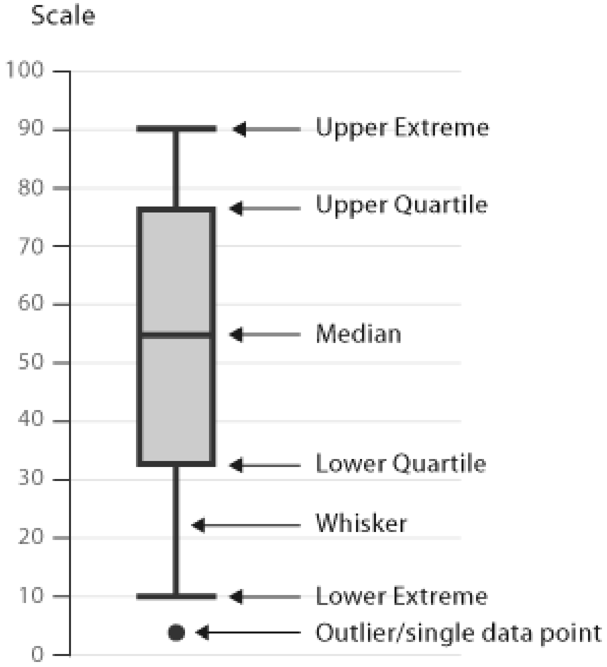
- **Collecting and summarizing data**
  - Basic statistics are helpful in summarizing data
  - Box plots, scatter plots, box and whisker plots provide compact representations of how data is distributed
  - Useful for exploring data but may not always be the best choice for communicating the data to your audience.

Discovering relationships in data

- **PLOT** the data in a meaningful way







Discovering relationships in data

A note on Machine Learning

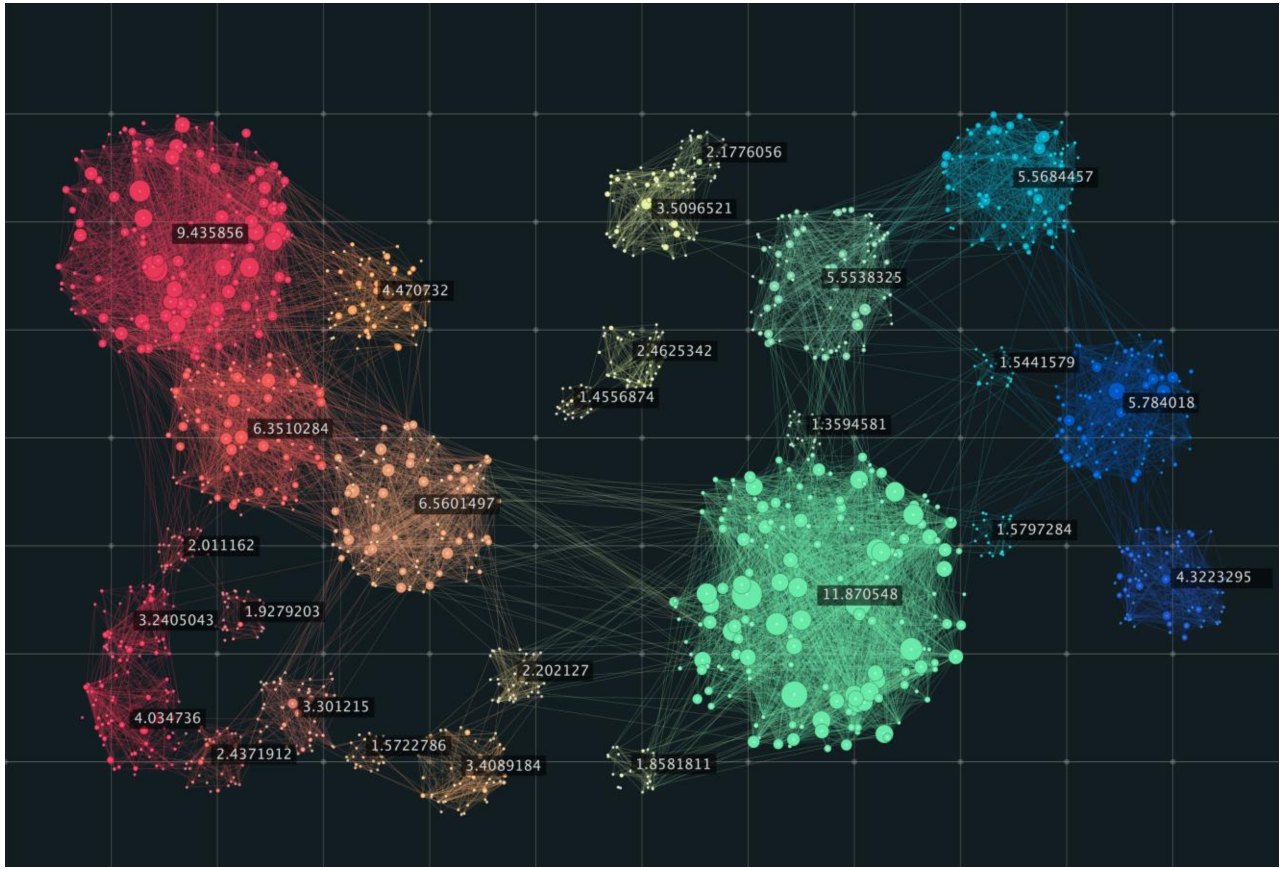
Supervised Learning

- Learn a function that approximates the relationship between input and output observable in the data
- **Labeled data**
- Examples:
  - Classification
  - Regression

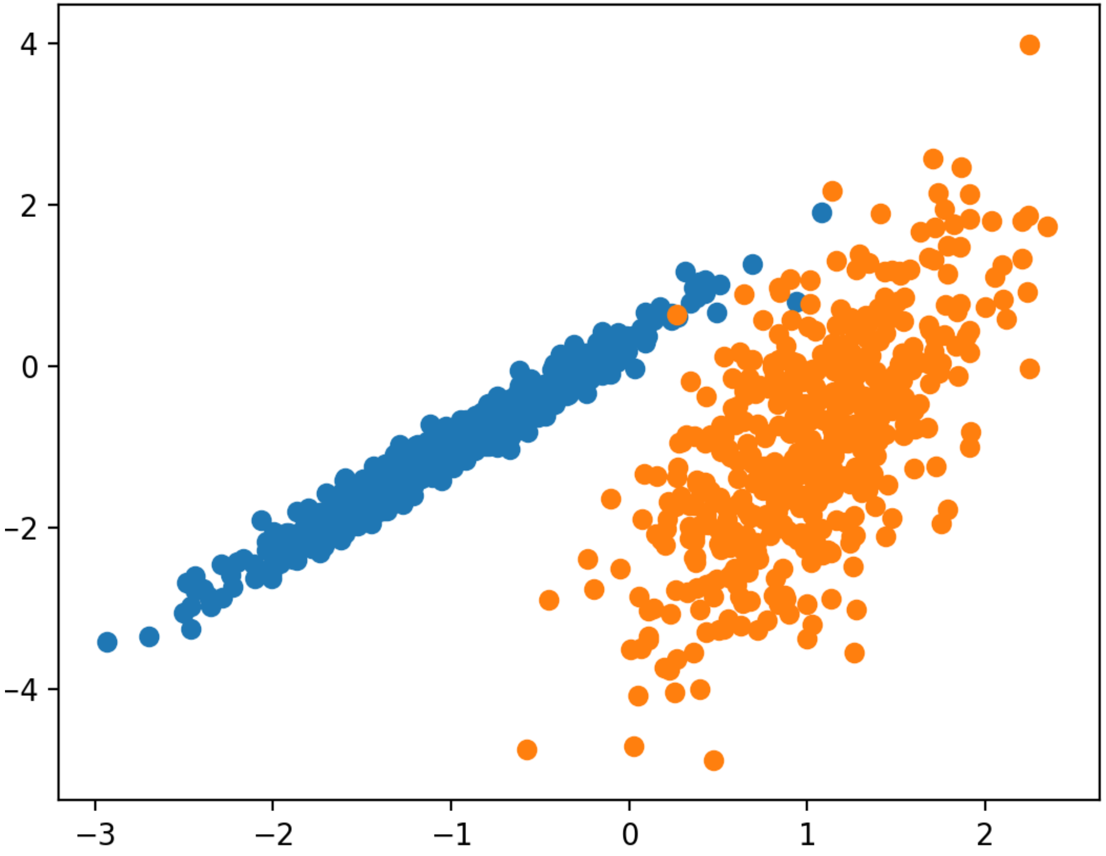
Unsupervised Learning

- Uncover the natural structure present within a set of data points
- **Unlabeled data**
- Examples:
  - Clustering

Clustering



https://www.mygreatlearning.com/blog/clustering-algorithms-in-machine-learning/



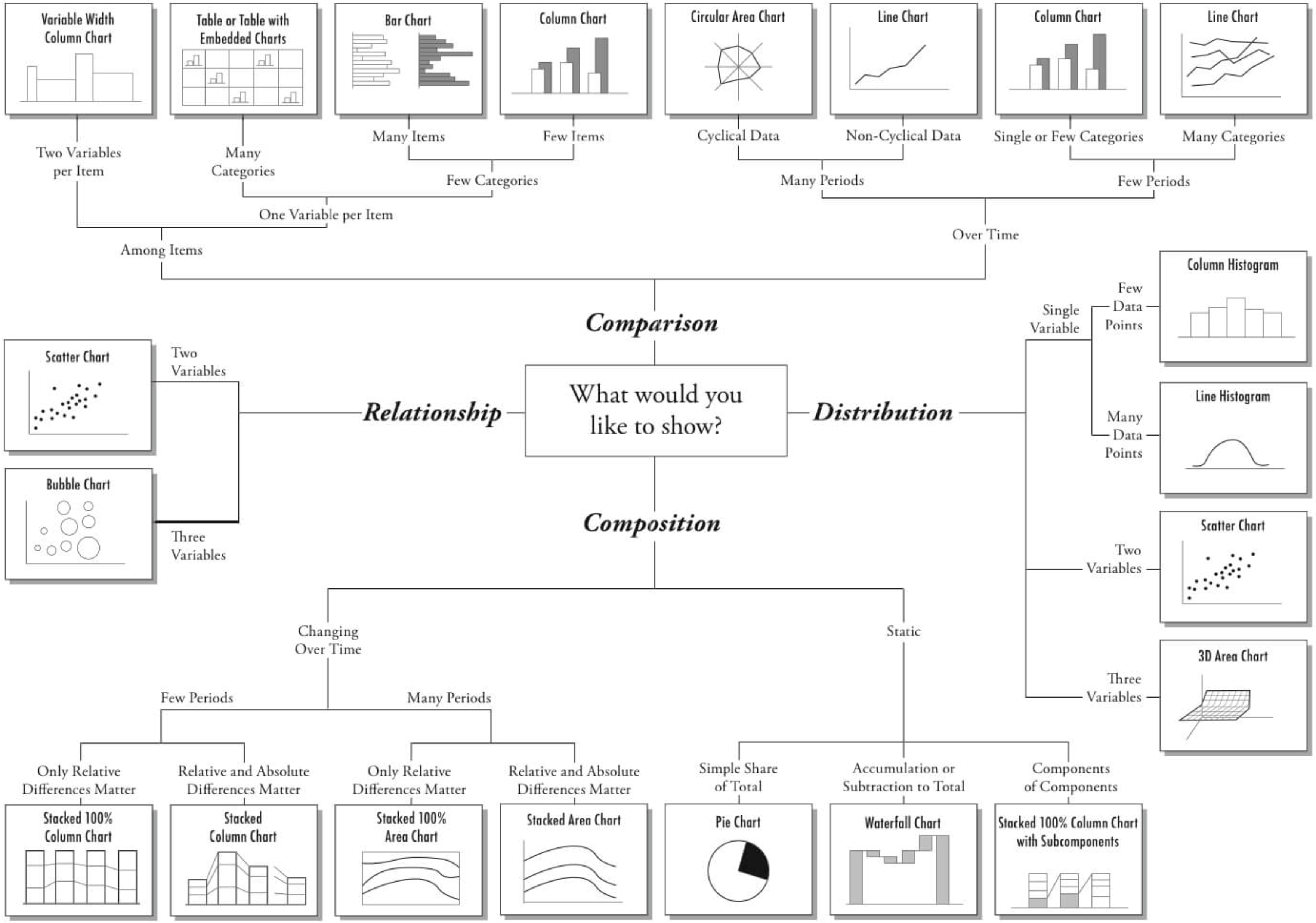
https://machinelearningmastery.com/clustering-algorithms-with-python/

Storytelling with Data

Communicating meaning from complex data

Chart Suggestions—A Thought-Starter

www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com



What would you like to show?

Comparison

Among Items

Two Variables per Item: Variable Width Column Chart

One Variable per Item

Many Categories: Table or Table with Embedded Charts

Few Categories

Many Items: Bar Chart

Few Items: Column Chart

Over Time

Many Periods

Cyclical Data: Circular Area Chart

Non-Cyclical Data: Line Chart

Few Periods

Single or Few Categories: Column Chart

Many Categories: Line Chart

Relationship

Two Variables: Scatter Chart

Three Variables: Bubble Chart

Distribution

Single Variable

Few Data Points: Column Histogram

Many Data Points: Line Histogram

Two Variables: Scatter Chart

Three Variables: 3D Area Chart

Composition

Changing Over Time

Few Periods

Only Relative Differences Matter: Stacked 100% Column Chart

Relative and Absolute Differences Matter: Stacked Column Chart

Many Periods

Only Relative Differences Matter: Stacked 100% Area Chart

Relative and Absolute Differences Matter: Stacked Area Chart

Static

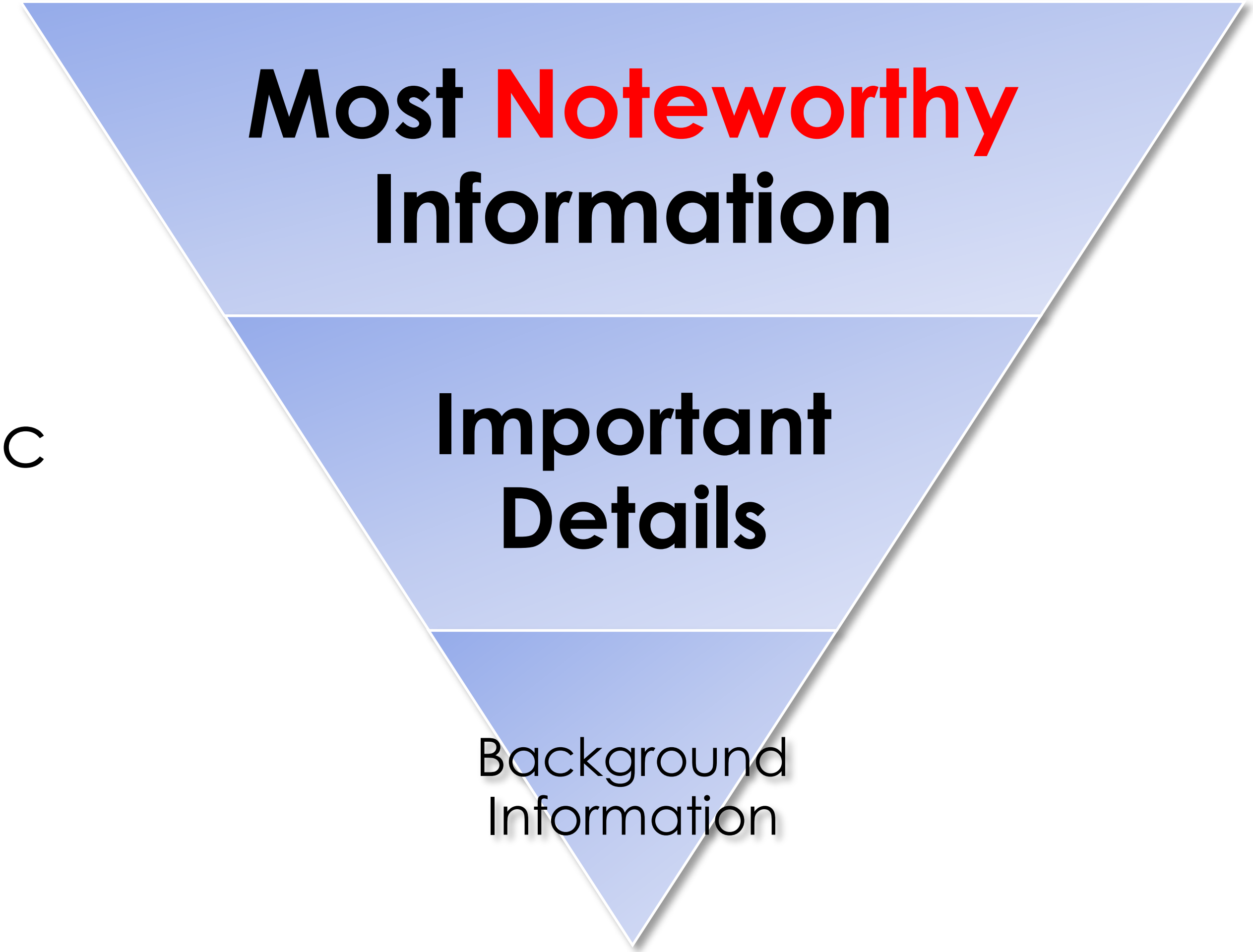
Simple Share of Total: Pie Chart

Accumulation or Subtraction to Total: Waterfall Chart

Components of Components: Stacked 100% Column Chart with Subcomponents

Dashboard Design

- Composed of **"elements"** (aka "cards")
- Often laid out on a grid.
- Each element has a specific purpose – tells its own story.
- Choose **size, color, and placement** of elements to draw attention to what is important.



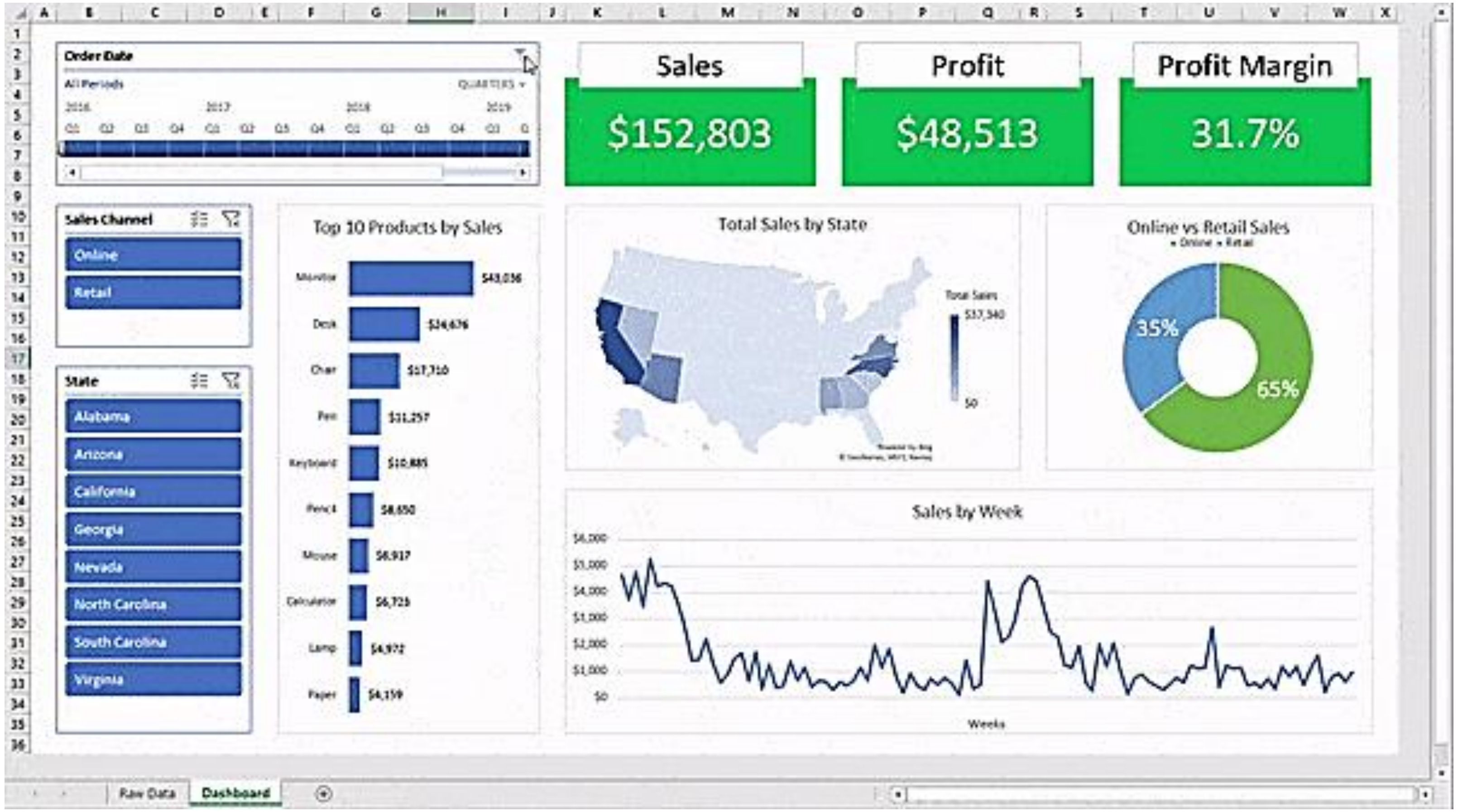
Most Noteworthy Information

Important Details

Background Information

Types of Elements

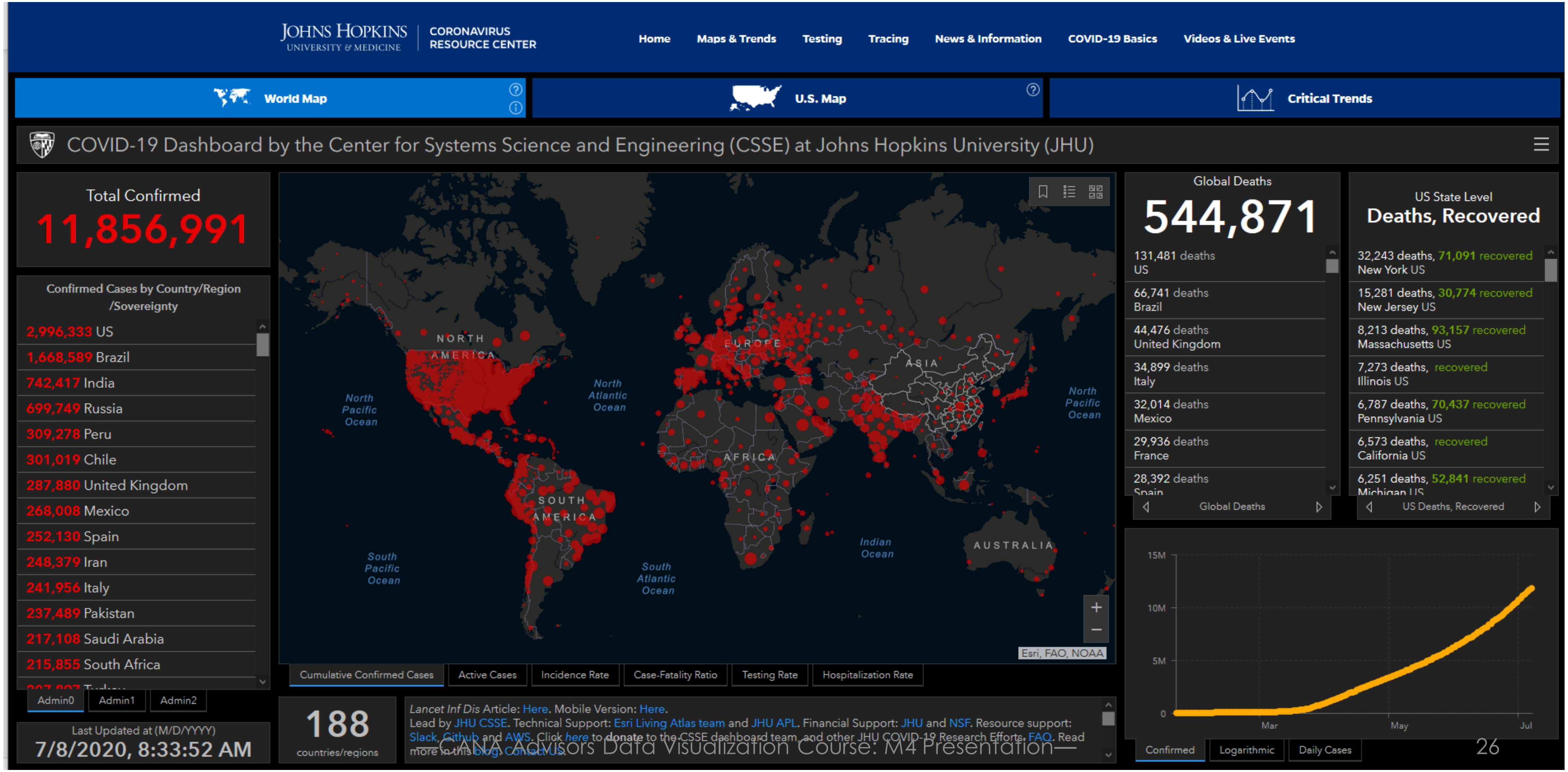
- Big Numbers
- Charts / Graphs
- Maps
- Heat map
- Slicers
- Tables
- Sparklines



Dashboard Best Practices

- Consider visual hierarchy
- Consider your goal
- Provide Context
- Keep it Simple
- Provide user controls
- Use the right visuals
- Select relevant KPIs
- Iterate & Improve

Dashboard Critique



COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)

Total Confirmed

11,856,991

Confirmed Cases by Country/Region /Sovereignty

2,996,333	US
1,668,589	Brazil
742,417	India
699,749	Russia
309,278	Peru
301,019	Chile
287,880	United Kingdom
268,008	Mexico
252,130	Spain
248,379	Iran
241,956	Italy
237,489	Pakistan
217,108	Saudi Arabia
215,855	South Africa
207,897	Turkey

Admin0 Admin1 Admin2

Last Updated at (M/D/YYYY)

7/8/2020, 8:33:52 AM

Cumulative Confirmed Cases | Active Cases | Incidence Rate | Case-Fatality Ratio | Testing Rate | Hospitalization Rate

188

countries/regions

Lancet Inf Dis Article: Here. Mobile Version: Here.
Lead by JHU CSSE. Technical Support: Esri Living Atlas team and JHU APL. Financial Support: JHU and NSF. Resource support: Slack, Github and AWS. Click *here* to **donate** to the CSSE dashboard team, and other JHU COVID-19 Research Efforts. FAQ. Read more in this blog. Contact US.

Global Deaths

544,871

131,481 deaths	US
66,741 deaths	Brazil
44,476 deaths	United Kingdom
34,899 deaths	Italy
32,014 deaths	Mexico
29,936 deaths	France
28,392 deaths	Spain

Global Deaths

US State Level

Deaths, Recovered

32,243 deaths, 71,091 recovered	New York US
15,281 deaths, 30,774 recovered	New Jersey US
8,213 deaths, 93,157 recovered	Massachusetts US
7,273 deaths, recovered	Illinois US
6,787 deaths, 70,437 recovered	Pennsylvania US
6,573 deaths, recovered	California US
6,251 deaths, 52,841 recovered	Michigan US

US Deaths, Recovered

Confirmed | Logarithmic | Daily Cases

Practical Examples

Let’s see it in practice!