



HKUST
VISLAB

COMP 4462

Data Visualization Tutorial

Leo Yu Ho, Lo
Ming Yao

Tuesday 19 February, 2019
<https://bit.ly/vis-t02>

Visualization process

Prepare data

- Get data
 - Download, crawl, collect
- Load data
 - Load data into visualization software
- Transform, join and aggregate
 - Make data to the form that ready to be drawn
- Filter
 - Clean up data and remove irrelevant information

Draw data (visualize)

- Visual encoding design
 - It's what you learn from the lectures
 - Marks and channels
 - Position, color, size, shape, etc.
- Interactions
 - Pan and zoom, select and filter
 - Click, drag and drop, scroll, and keyboard input, etc.

Get data

- Download data prepared by the others
 - Kaggle Dataset
 - World Bank
 - And many more
- Crawl from the web
 - Write your own program to crawl
 - From API
 - Extract from HTML or JSON
 - Python: Scrapy, BeautifulSoup
 - Nodejs: Cheerio
- Collect from users
 - Most costly
 - Takes time and efforts
 - Best in quality

Load data

- Most common
 - csv: comma separated value
 - tsv: tab separated value
 - xlsx: Excel
- Databases
 - SQL: Oracle, MySQL, PostgreSQL, MS SQL, etc.
 - Structured, normalized
 - NoSQL: MongoDB
 - Document based
- PDF
 - As far as I know, only Tableau supports import from PDF
- Other source
 - [Google Cloud Public Datasets](#)
 - Only available through Google Cloud
 - Too big to be downloaded

Clean data

- Data is always dirty
 - Missing values
 - Typo
 - Overloaded fields (mixing continuous numbers with text)
 - Mismatch primary keys / external keys
 - Duplicate entries
 - Missing data for several days
 - Equipment failure / bugs in crawler programs / website is down
 - Non-sense error in data, e.g. integer overflow, or just not making any sense
 - Emoji / language / accent decoration
 - Identical typeface but different in unicode
- Depends on severity, it can be very nasty to deal with
- Data normalization
 - [Google Text Normalization Challenge](#)

Transform, join and aggregate

- Manipulate data to the form for visualization
 - Wide form
 - Long form
 - Derive attributes: percentage changes, year-to-year changes
- Join
 - Linking up multiple table or data sources
 - Inner join, left join, right join, outer join
 - Commonly join on ID
 - Sometimes on date
 - Sometimes on multiple attributes
- Aggregate
 - Statistical: counting, sum, average, median, etc.
 - Grouping: binning, frequency, time slicing
 - Moving average, running sum

Ranking	2018	2017	2016	2015
CS	14	19	14	8
CHEM	23	27	28	25

Subject	Ranking	Year
CS	14	2018
CS	19	2017
CS	14	2016
CS	8	2015
CHEM	23	2018
CHEM	27	2017
CHEM	28	2016
CHEM	25	2015

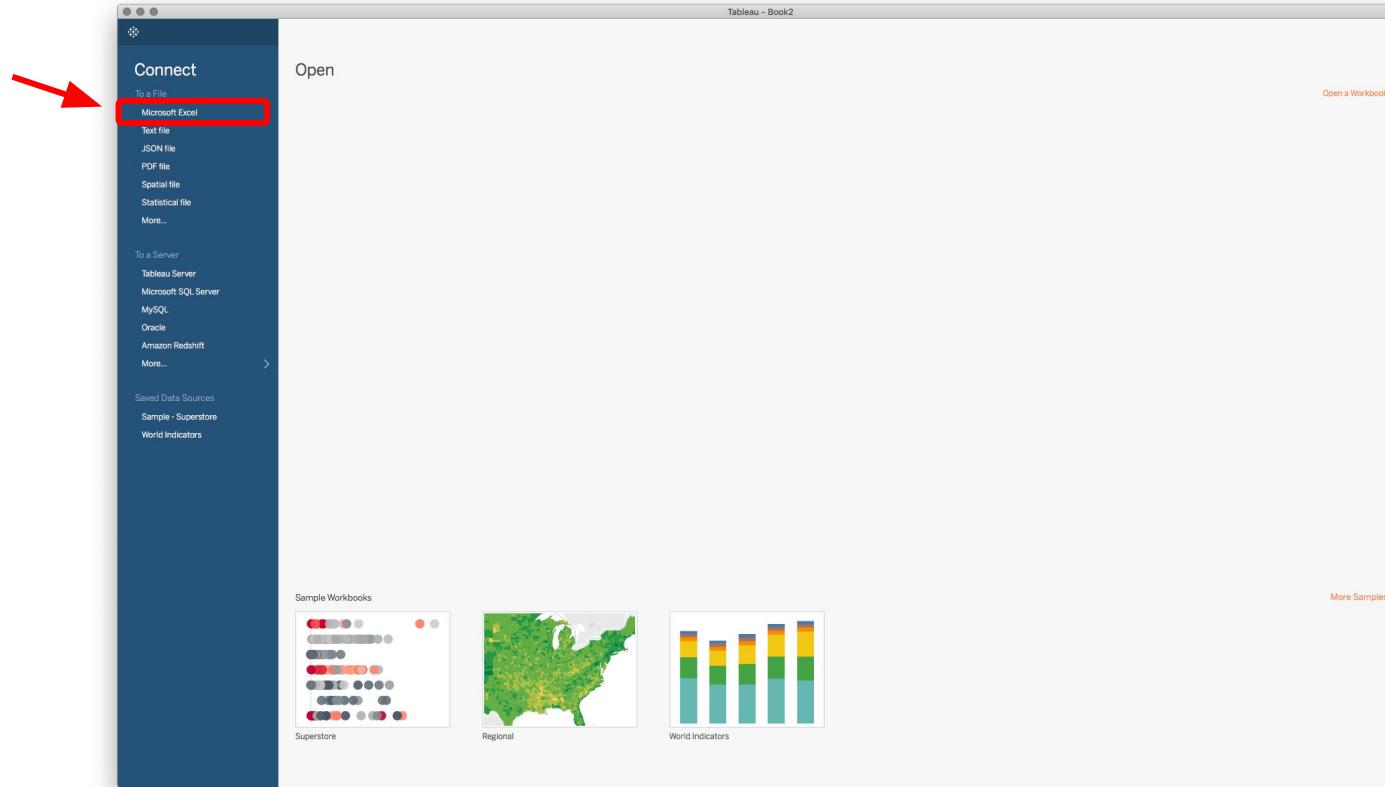
Filter

- Reduce the number of items to show
- Focus only on relevant data, clean up irrelevant data
 - Base on user interest
 - Or users' level of authority
 - Not everyone can access all the data
 - Time relevancy
 - Outdated data are no longer relevant to real-time analysis
 - Geographic relevance
 - You don't care about restaurants outside Hong Kong (unless you're going to travel)
- Hard to show all with a limited screen size
 - Especially on mobile device
 - Reduce cluttering, more “clickable” on screen to show item details
- Zoom in to a specific small subset of data
 - Then you can show more detail of each item
 - Google Maps, zoom in to show more detailed terrain

Tableau

- Tableau Public
 - Free
 - All saved works are public
 - Publicly viewable, downloadable
 - Must connect to the internet in order to save
 - Less data connectors
- Tableau Desktop
 - Free for students, need verification
 - Can save locally, use without connecting to the internet
 - More data connectors
- Tableau Server
 - Standalone, dedicated server
 - Enterprise level, expensive

Load Data



Load Data

Tableau - Book2

Connections

global_superstore_2016 Microsoft Excel

Sheets

Orders

People

Returns

New Union

Use Data Interpreter

Data Interpreter might be able to clean your Microsoft Excel workbook.

Orders

Orders (global_superstore_2016)

Connection

Live Extract

Filters

0 | Add

Sort fields Data source order

Show aliases Show hidden fields 1,000 rows

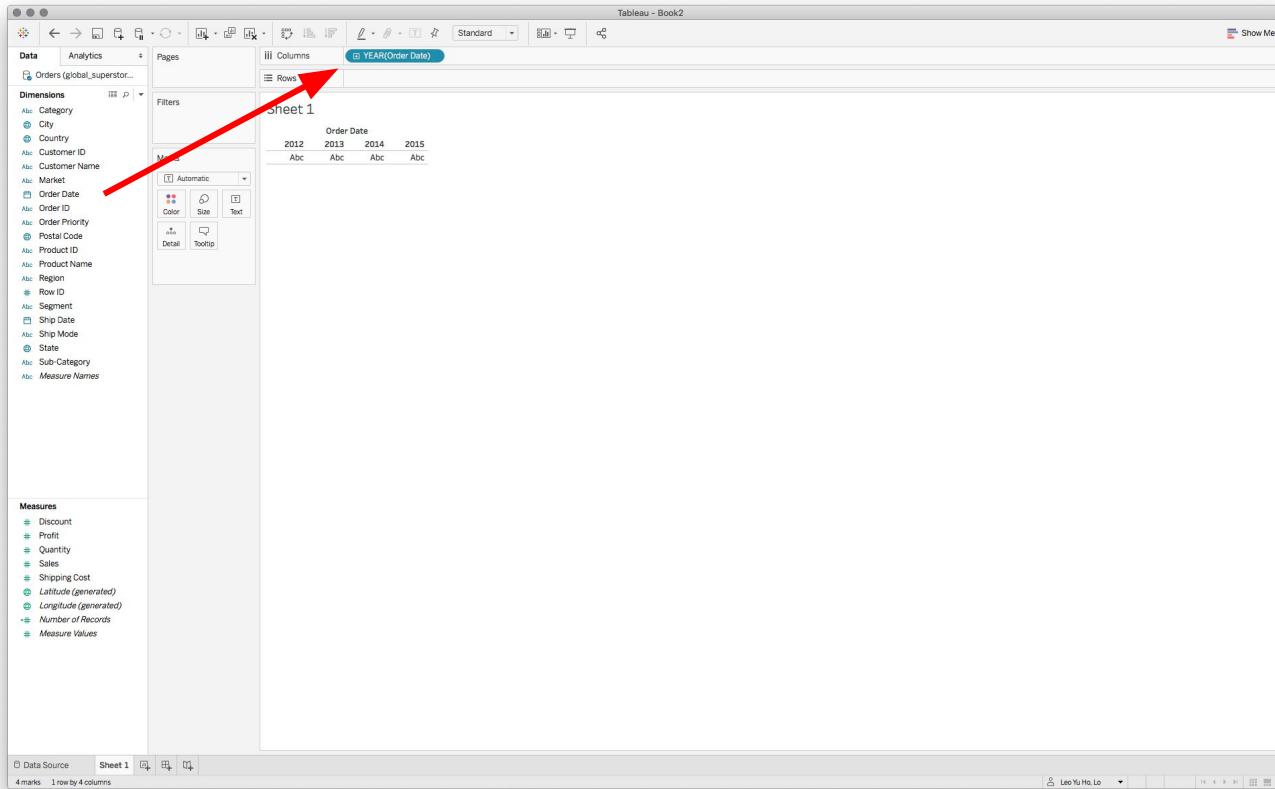
#	Orders Row ID	Abc Orders Order ID	Orders Order Date	Orders Ship Date	Abc Orders Ship Mode	Abc Orders Customer ID	Abc Orders Customer Name	Abc Orders Segment	Orders Postal Code	Orders City	Orders State	Orders Country	Abc Orders Region	Abc Orders Market
40098	CA-2014-AB1001514...	11/11/2014	11/13/2014	First Class	AB-100151402	Aaron Bergman	Consumer	73120	Oklahoma City	Oklahoma	United States	Central US	USCA	
26341	IN-2014-JR162107...	2/5/2014	2/7/2014	Second Class	JR-162107	Justin Ritter	Corporate	null	Wollongong	New South Wales	Australia	Oceania	Asia Pacific	
25330	IN-2014-CR127307...	10/17/2014	10/18/2014	First Class	CR-127307	Craig Reiter	Consumer	null	Brisbane	Queensland	Australia	Oceania	Asia Pacific	
13524	ES-2014-KM163754...	1/28/2014	1/30/2014	First Class	KM-1637548	Katherine Murray	Home Office	Berlin	Berlin	Germany	Western Europe	Europe		
47221	SG-2014-RH949511...	11/5/2014	11/6/2014	Same Day	RH-9495111	Rick Hansen	Consumer	null	Dakar	Dakar	Senegal	Western Africa	Africa	
22732	IN-2014-JM156557...	6/28/2014	7/1/2014	Second Class	JM-156557	Jim Mitchum	Corporate	null	Sydney	New South Wales	Australia	Oceania	Asia Pacific	
30570	IN-2012-TS213409...	11/6/2012	11/8/2012	First Class	TS-2134092	Toby Swindell	Consumer	null	Porirua	Wellington	New Zealand	Oceania	Asia Pacific	
31192	IN-2013-MB180859...	4/10/2013	4/18/2013	Standard Class	MB-1808592	Mick Brown	Consumer	null	Hamilton	Waikato	New Zealand	Oceania	Asia Pacific	
40099	CA-2014-AB1001514...	11/11/2014	11/13/2014	First Class	AB-100151402	Aaron Bergman	Consumer	73120	Oklahoma City	Oklahoma	United States	Central US	USCA	
36258	CA-2012-AB1001514...	3/6/2012	3/7/2012	First Class	AB-100151404	Aaron Bergman	Consumer	98103	Seattle	Washington	United States	Western US	USCA	
36259	CA-2012-AB1001514...	3/6/2012	3/7/2012	First Class	AB-100151404	Aaron Bergman	Consumer	98103	Seattle	Washington	United States	Western US	USCA	
28879	ID-2013-AU107801...	4/19/2013	4/22/2013	First Class	AU-107801	Anthony Jacobs	Corporate	null	Kabul	Kabul	Afghanistan	Southern Asia	Asia Pacific	
45794	SA-2012-MM72601...	12/26/2012	12/28/2012	Second Class	MM-7260110	Magdelene Morse	Consumer	null	Jizan	Jizan	Saudi Arabia	Western Asia	Asia Pacific	
4132	MX-2013-V2171518...	11/13/2013	11/13/2013	Same Day	VF-2171518	Vicky Freymann	Home Office	null	Toledo	Parana	Brazil	South America	LATAM	
27704	IN-2014-PF1912027...	6/6/2014	6/8/2014	Second Class	PF-1912027	Peter Fuller	Consumer	null	Mudanjiang	Heilongjiang	China	Eastern US	Asia Pacific	
13779	ES-2015-BP1118545...	7/3/2015	8/3/2015	Second Class	BP-1118545	Ben Peterman	Corporate	null	Paris	Ile-de-France	France	Western Europe	Europe	
39519	CA-2012-AB1001514...	2/19/2012	2/25/2012	Standard Class	AB-100151402	Aaron Bergman	Consumer	76017	Arlington	Texas	United States	Central US	USCA	
12069	ES-2015-PJ1888356...	9/8/2015	9/14/2015	Standard Class	PJ-18883564	Patrick Jones	Corporate	null	Prato	Tuscany	Italy	Southern Europe	Europe	
93046	IN-2012-IE16687...	9/1/2012	9/1/2012	First Class	IE-16687	Ham Clark	Consumer	null	Tumut	Queensland	Australia	Oceania	Asia Pacific	

Go to Worksheet

Data Source Sheet 1

Leo Yu Ho Lo

Basic Plotting: Select Row and Column

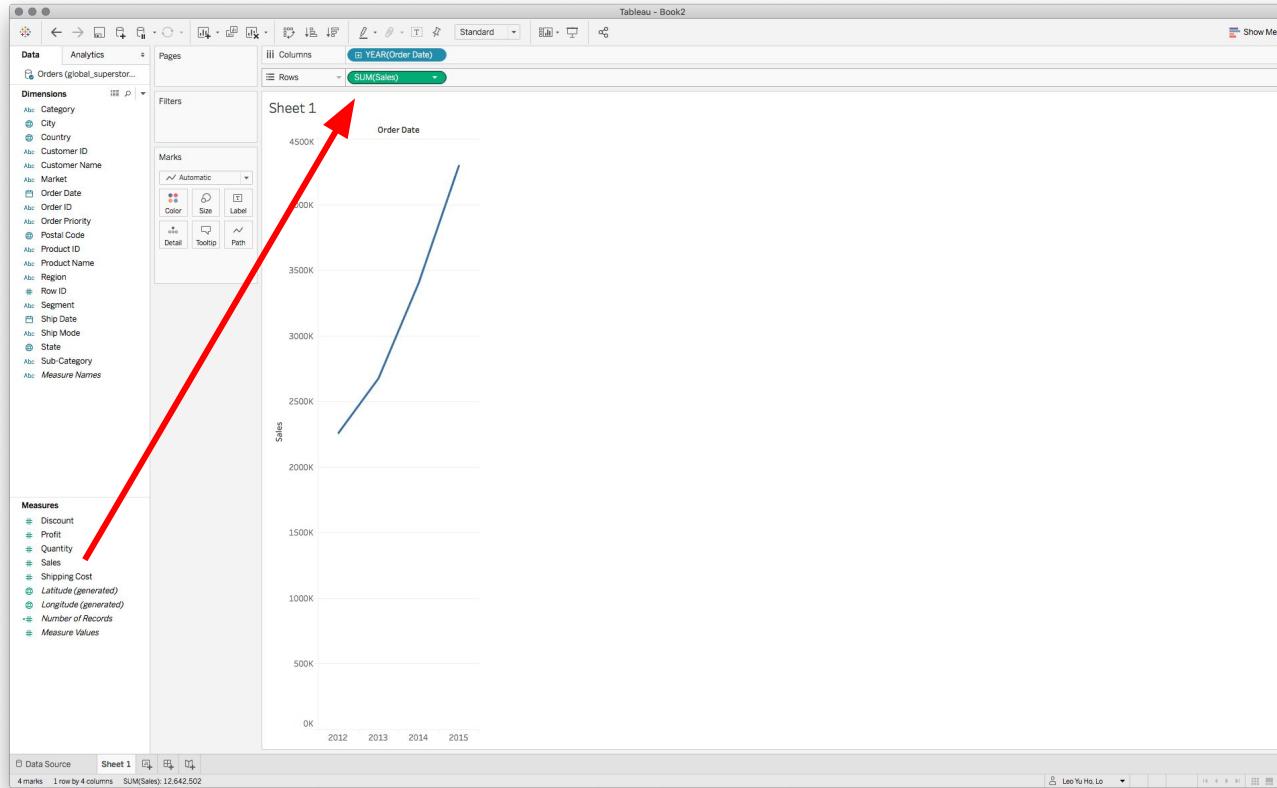


The screenshot shows the Tableau desktop interface with the following details:

- Top Bar:** Includes standard window controls (Minimize, Maximize, Close), a back/forward navigation bar, and a toolbar with various icons for data preparation and visualization.
- Header:** Displays "Tableau - Book2" and a "Show Me" button.
- Left Shelf:** Contains two sections:
 - Dimensions:** Lists "Category", "City", "Country", "Customer ID", "Customer Name", "Market", "Order Date", "Order ID", "Order Priority", "Postal Code", "Product ID", "Product Name", "Region", "Row ID", "Segment", "Ship Date", "Ship Mode", "State", "Sub-Category", and "Measure Names".
 - Measures:** Lists "Discount", "Profit", "Quantity", "Sales", "Shipping Cost", "Latitude (generated)", "Longitude (generated)", "Number of Records", and "Measure Values".
- Center Area:** Shows a data grid titled "Sheet 1" with the following data:

	Order Date	2012	2013	2014	2015
	Abc	Abc	Abc	Abc	
- Right Shelf:** Shows the "Rows" shelf with the "YEAR(Order Date)" dimension selected. A red arrow points to this shelf.
- Bottom:** Includes a status bar with "4 marks 1 row by 4 columns", a copyright notice "© 2012 Tableau Software, Inc.", and a navigation bar with icons for zoom and orientation.

Basic Plotting: Select Row and Column



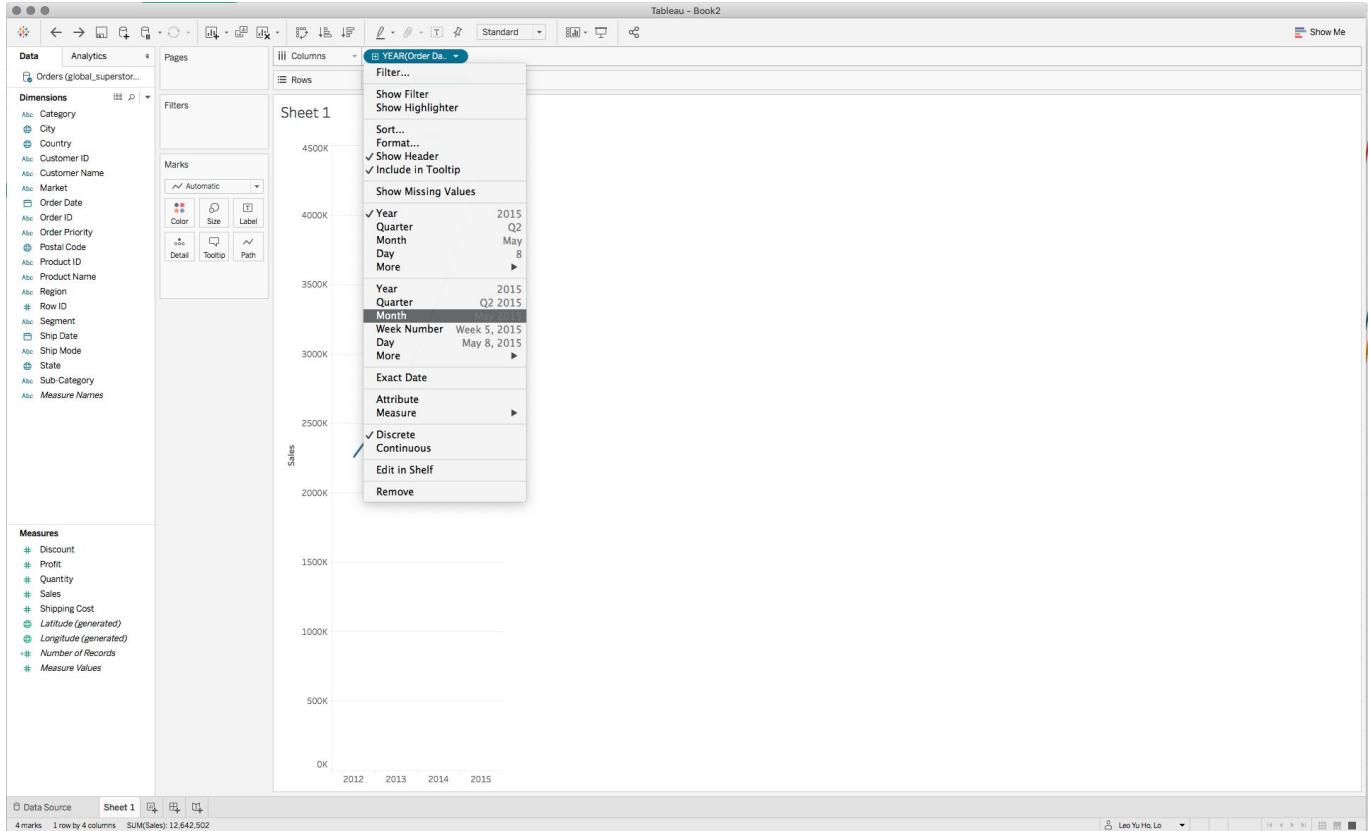
Basic Plotting: Adjust Date

“Dimensions” are discrete variables

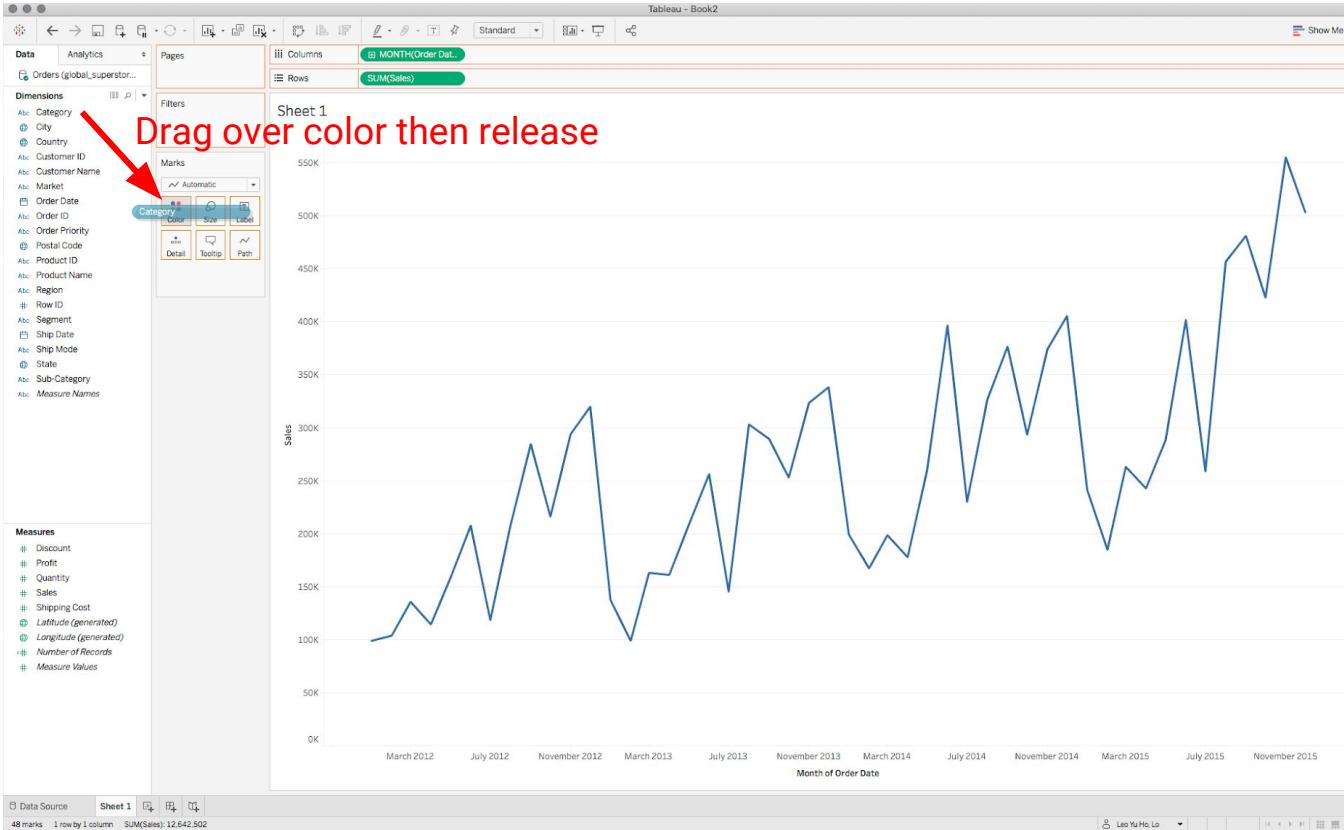
“Measures” are continuous variables

Date can be either discrete or continuous

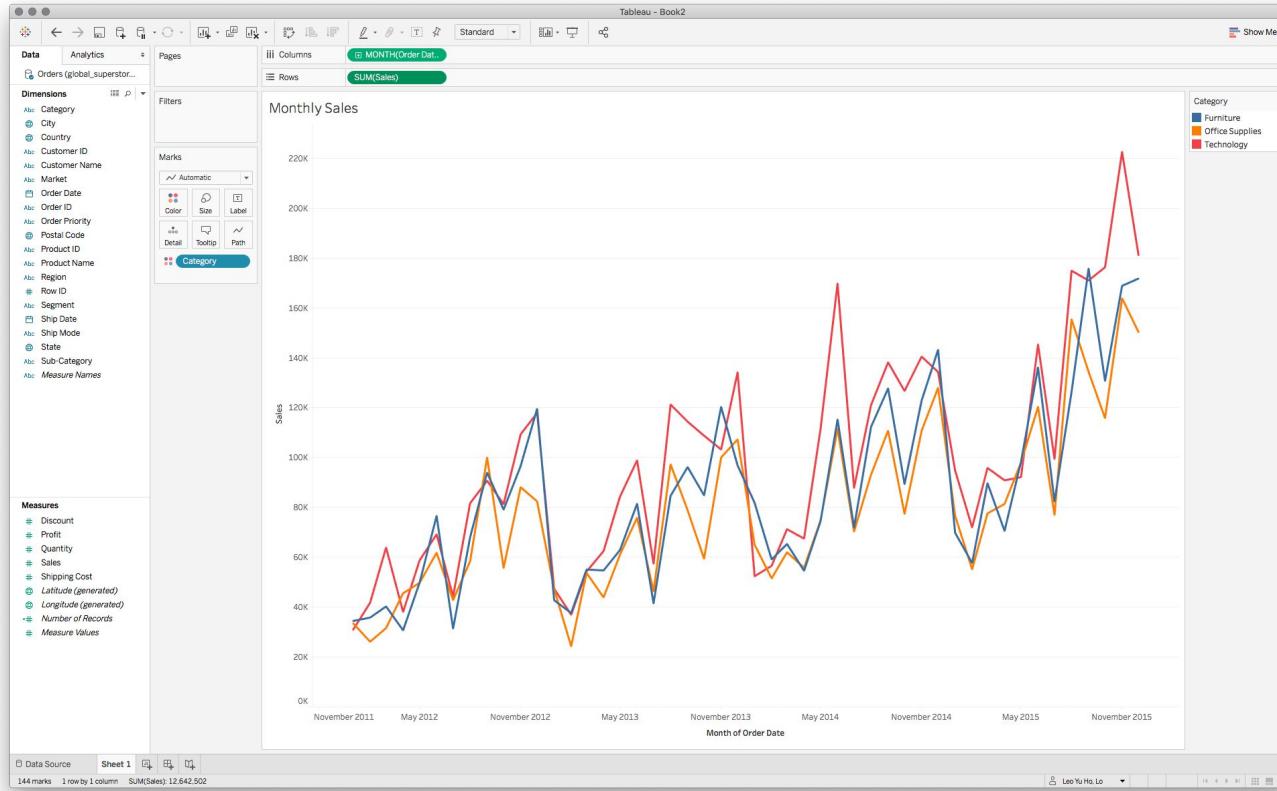
We need it to be continuous right now, you can try discrete and see what happens



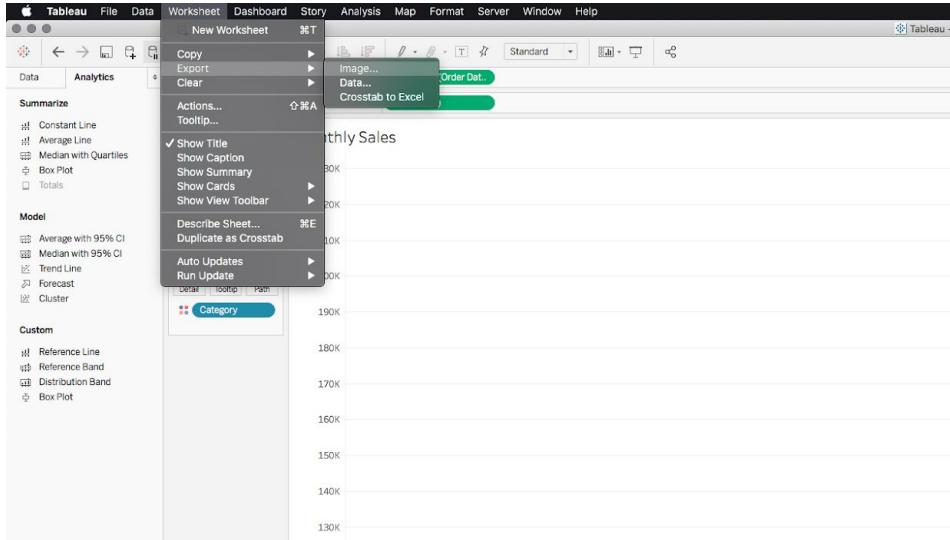
Basic Plotting: Marks with Color



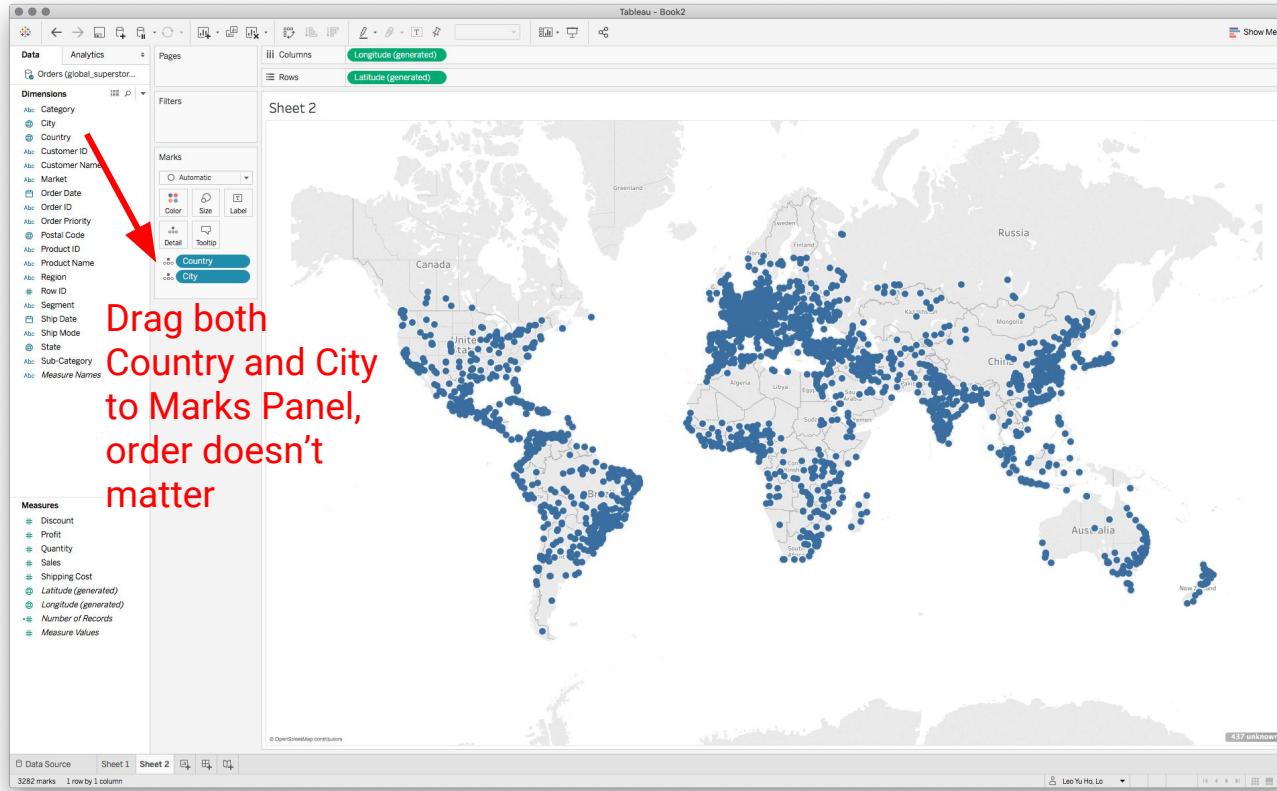
Basic Plotting: Ta-Da



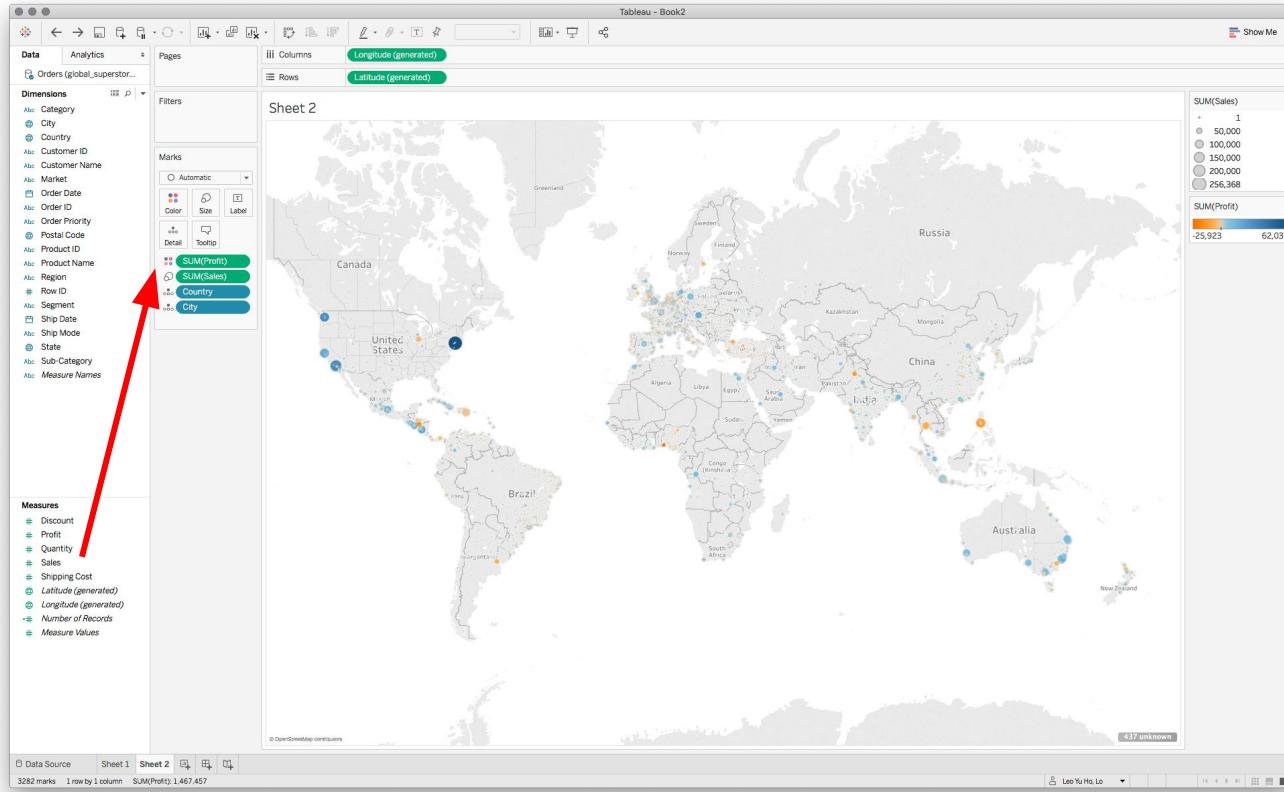
Export Image (Not available in Tableau Public)



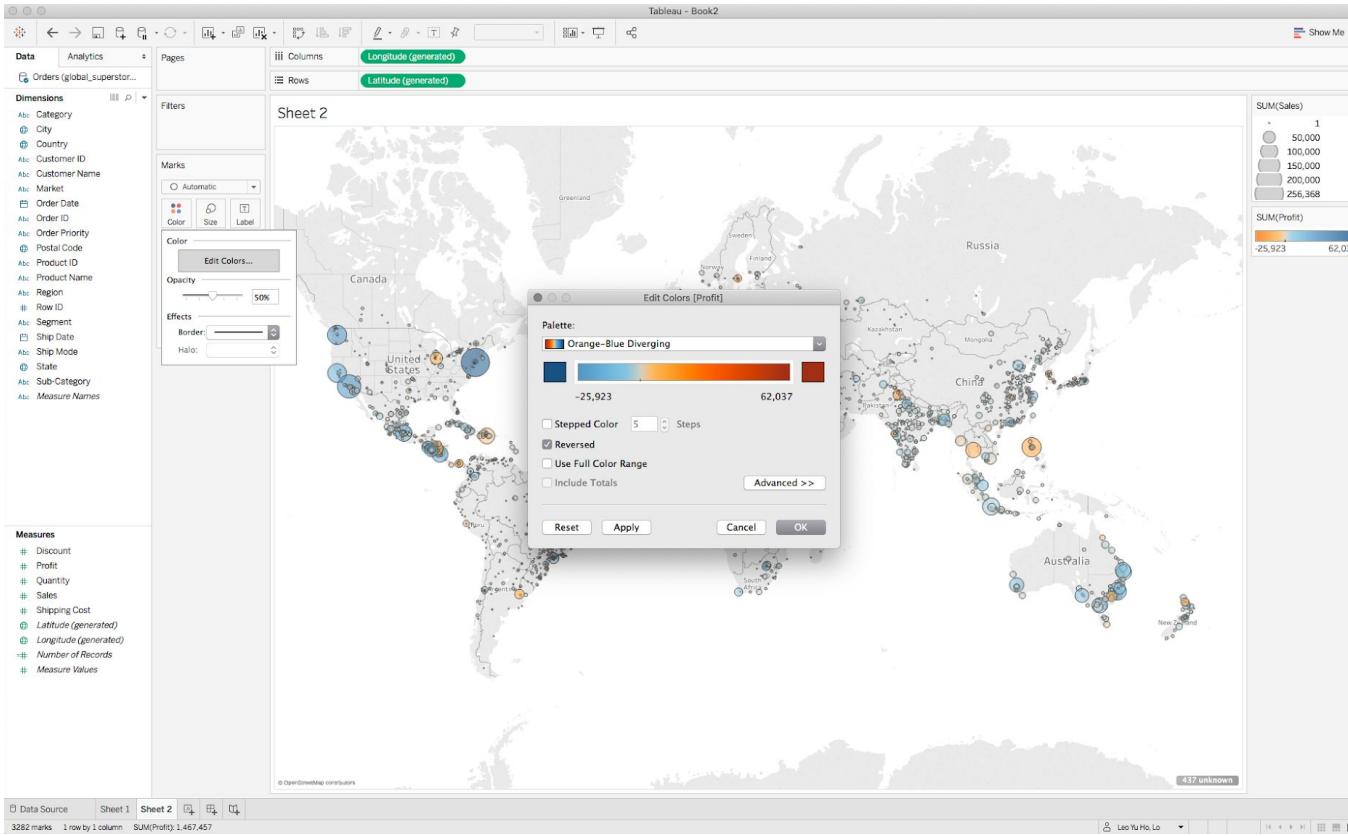
Plotting with Map



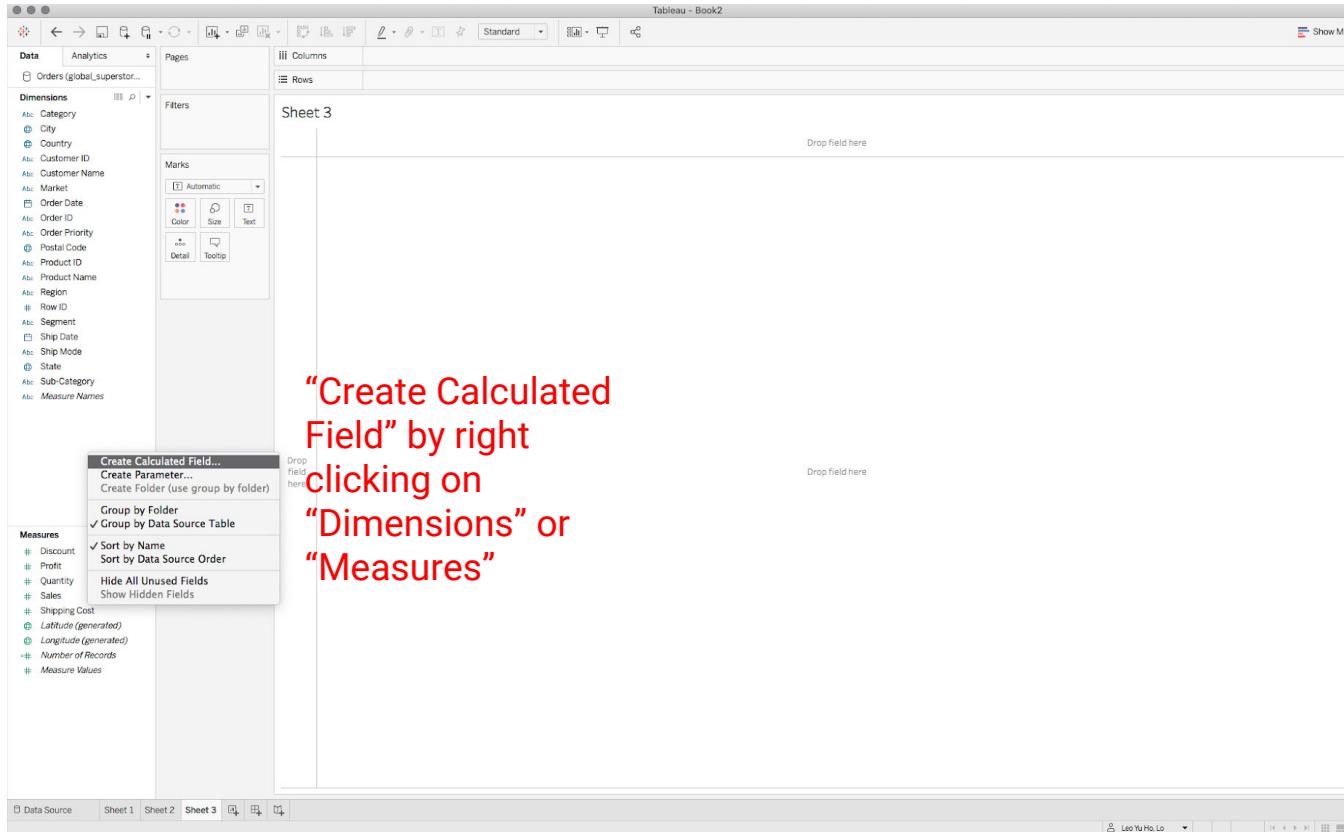
Plotting with Map: Encode with Size and Color



Adjust Color and Size



Calculated Field



The screenshot shows the Tableau interface with the following details:

- Top Bar:** Tableau - Book2, with standard menu and toolbar icons.
- Left Panel:** Data and Analytics dropdown, showing 'Orders (global_superstore)'. The Dimensions section is expanded, listing fields like Category, City, Country, Customer ID, etc. The Measures section is also expanded, listing fields like Discount, Profit, Quantity, Sales, etc.
- Middle Panel:** Shows 'Sheet 3' with a Marks card and a 'Drop field here' placeholder.
- Bottom Panel:** Shows tabs for Data Source, Sheet 1, Sheet 2, Sheet 3, and a search/filter area.
- Context Menu:** A right-click context menu is open over the 'Dimensions' and 'Measures' sections. The menu items include 'Create Calculated Field...', 'Create Parameter...', 'Create Folder (use group by folder)', 'Group by Folder', 'Group by Data Source Table', 'Sort by Name', 'Sort by Data Source Order', 'Hide All Unused Fields', and 'Show Hidden Fields'.

Text Overlay: A large red text box in the center of the interface contains the following instructions: "Create Calculated Field" by right clicking on "Dimensions" or "Measures"

Calculated Field

Tableau - Book2

Sheets: Sheet 3

Enter a calculation formula

Shipping Delay

[Ship Date] - [Order Date]

The calculation is valid.

OK

ABS(number)

Enter search text

Returns the absolute value of the given number.

Example: ABS(-7) = 7

ABS
ACOS
AND
ASCII
ATAN
ATAN2
ATTR
AVG
CASE
CEILING
CHAR

Standard

Drop Field here

Dimensions

- Category
- City
- Country
- Customer ID
- Customer Name
- Market
- Order Date
- Order ID
- Order Priority
- Postal Code
- Product ID
- Product Name
- Region
- Row ID
- Segment
- Ship Date
- Ship Mode
- State
- Sub-Category
- Measure Names

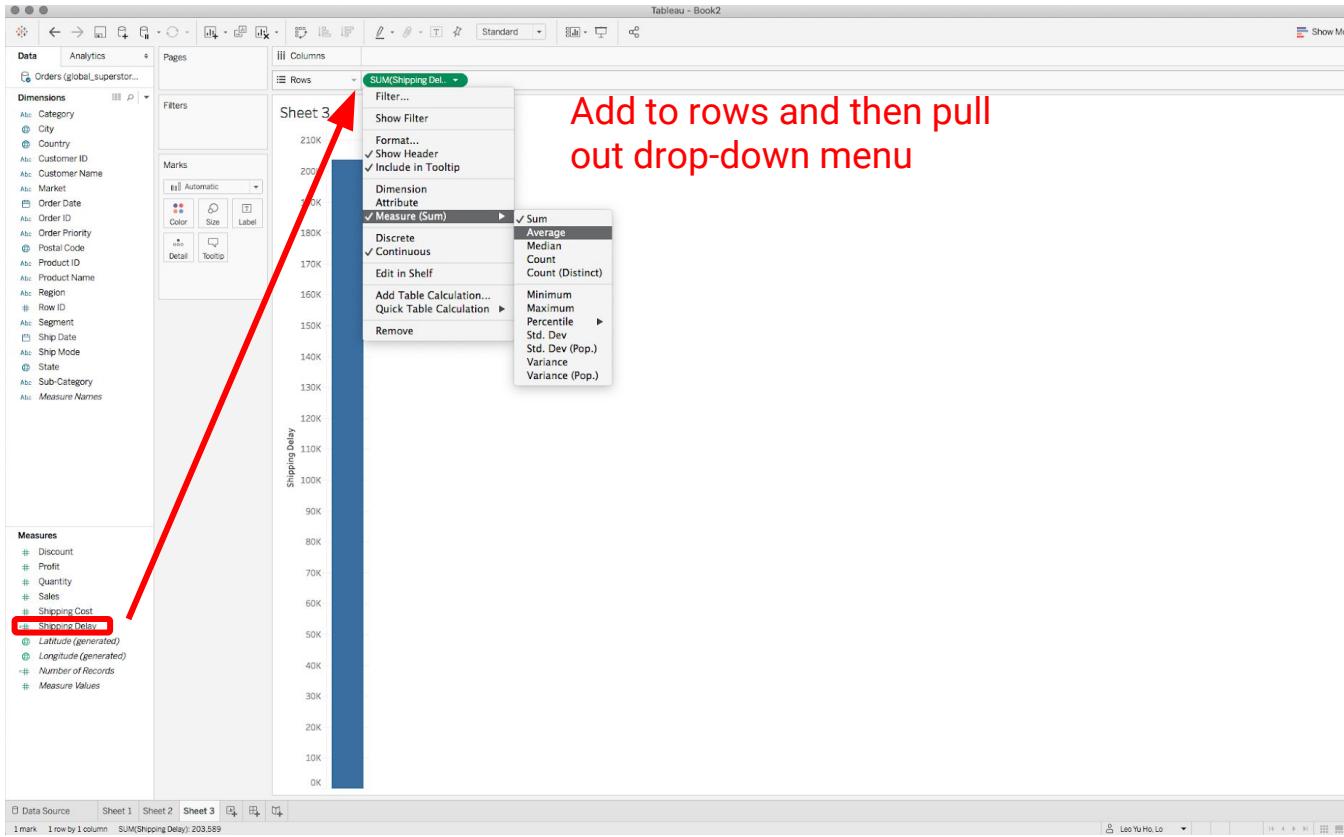
Measures

- Discount
- Profit
- Quantity
- Sales
- Shipping Cost
- Latitude (generated)
- Longitude (generated)
- Number of Records
- Measure Values

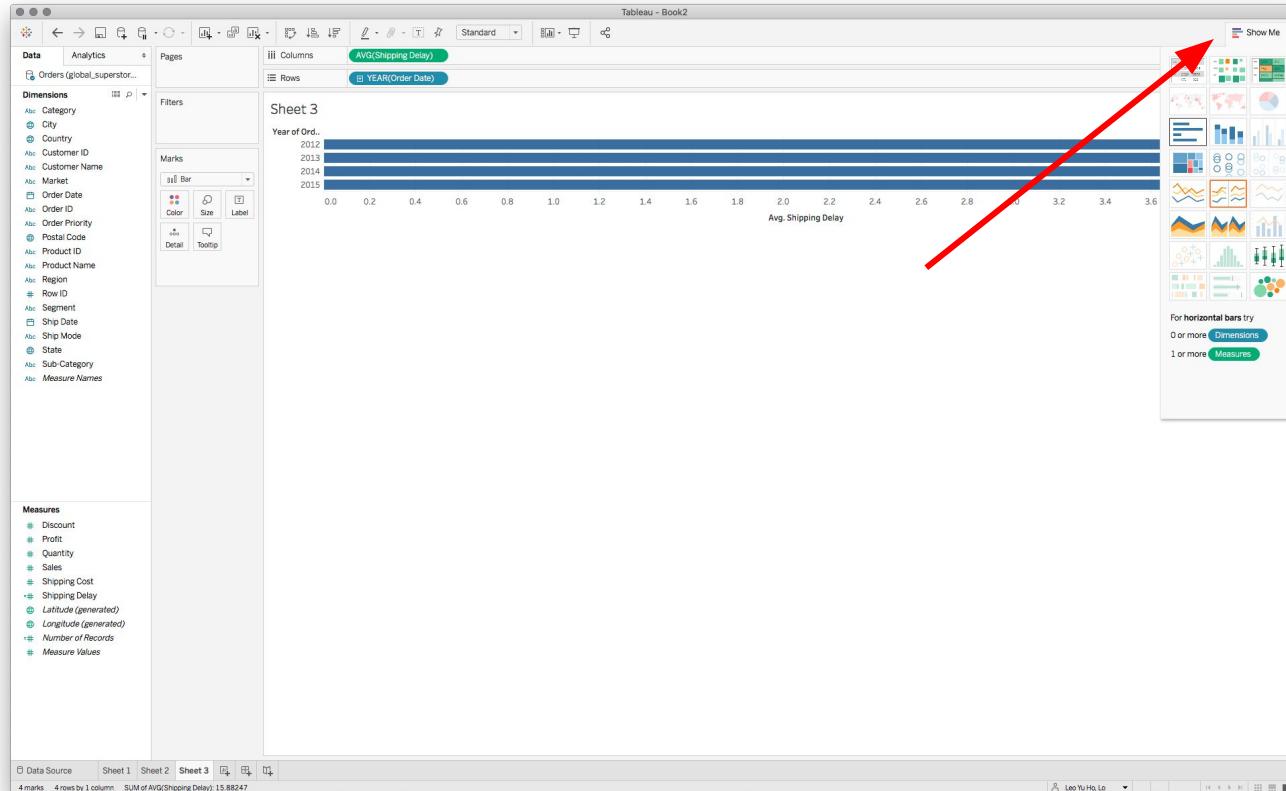
Data Analytics Pages Columns Rows

Leo Yu Ho Lo

Aggregate



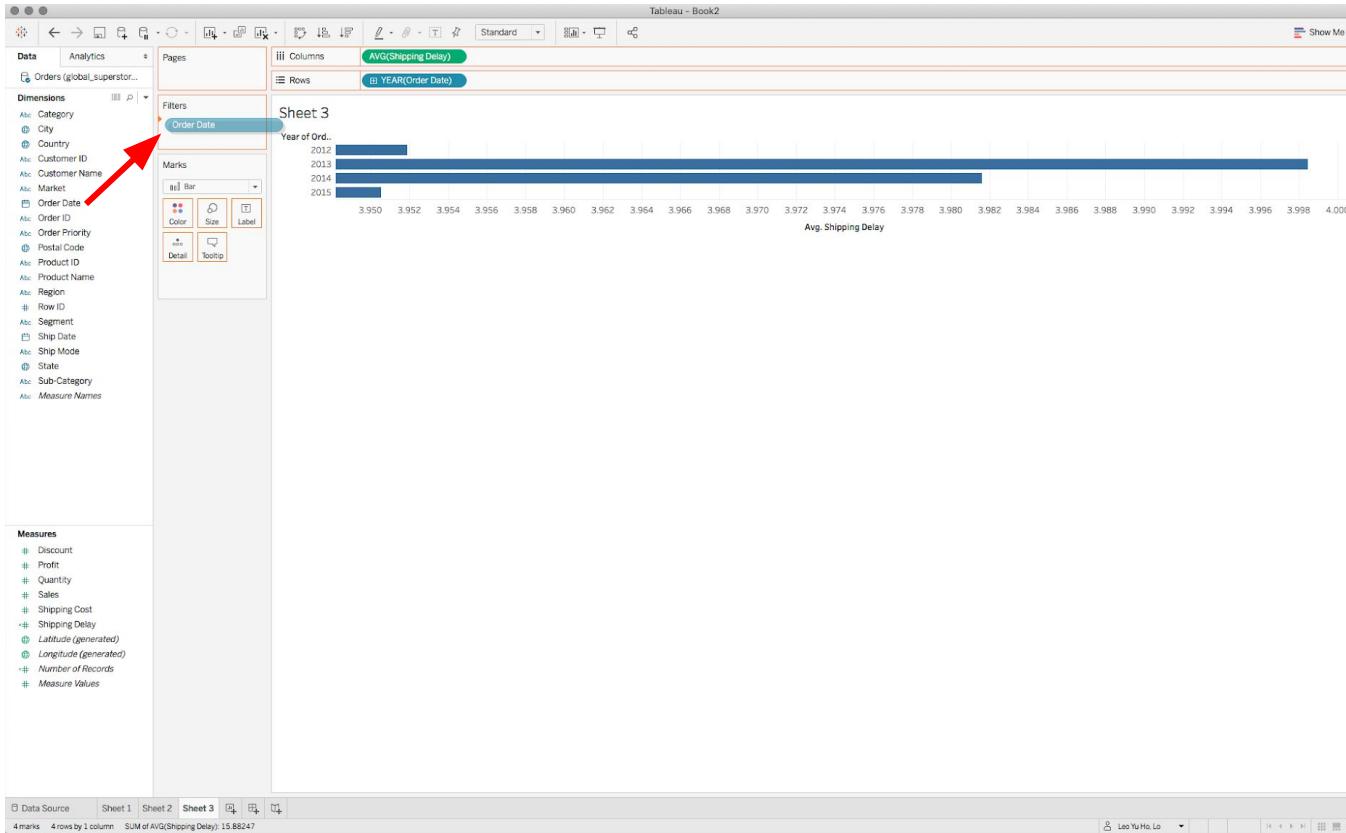
Plotting is easy using “Show Me”



Gray-out means not suitable for the current selected data types.

Try different combinations for different plots!

Filter



Filter

Tableau - Book2

Pages

Columns: AVG(Shipping Delay)

Rows: YEAR(Order Date)

Sheet 3

Year of Ord...

2012 3.952
2013 3.982
2014 3.998
2015 3.952

Filter [Year of Order Date]

General Condition Top

Select from list Custom value list Use all

Enter search text

2012
 2013
 2014
 2015

All None Exclude

Summary

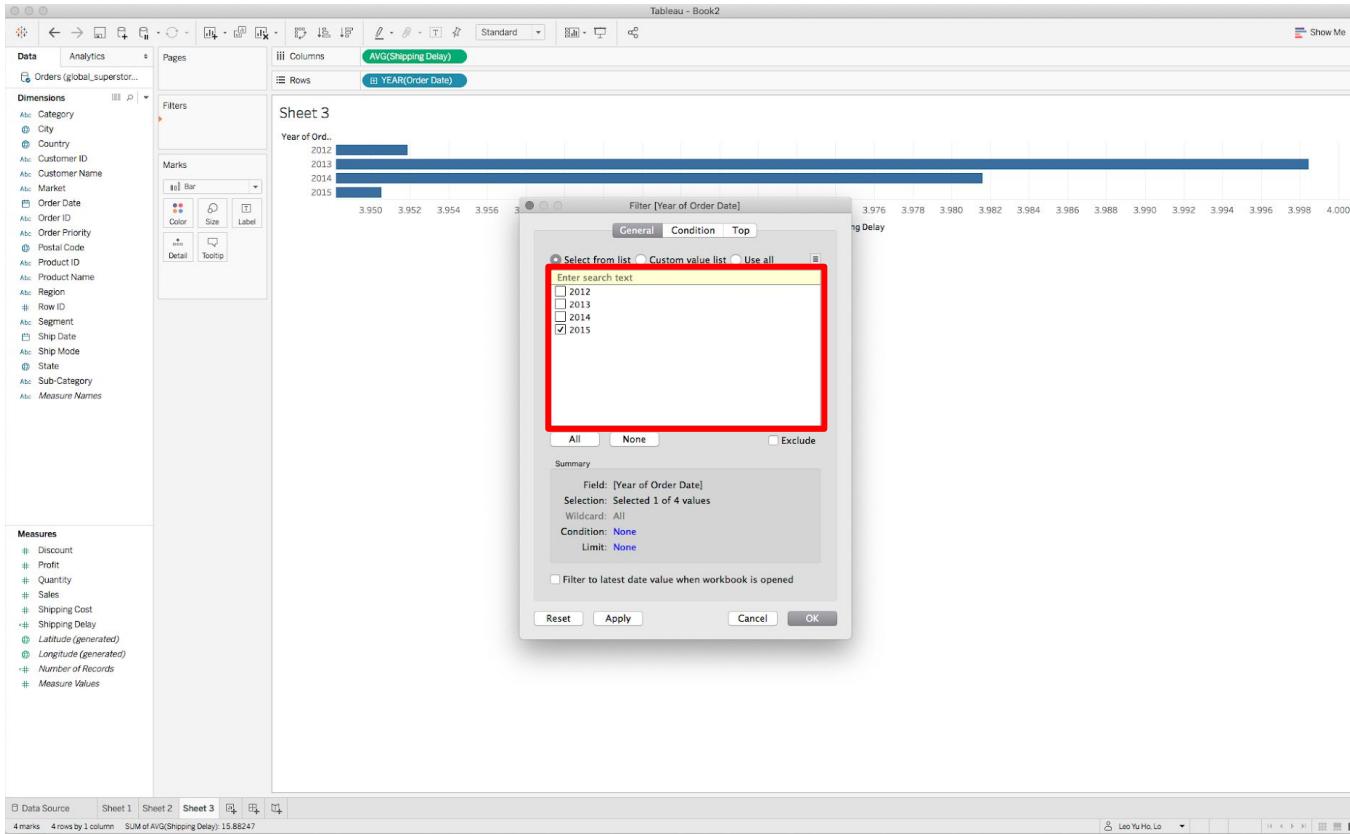
Field: [Year of Order Date]
Selection: Selected 1 of 4 values
Wildcard: All
Condition: None
Limit: None

Filter to latest date value when workbook is opened

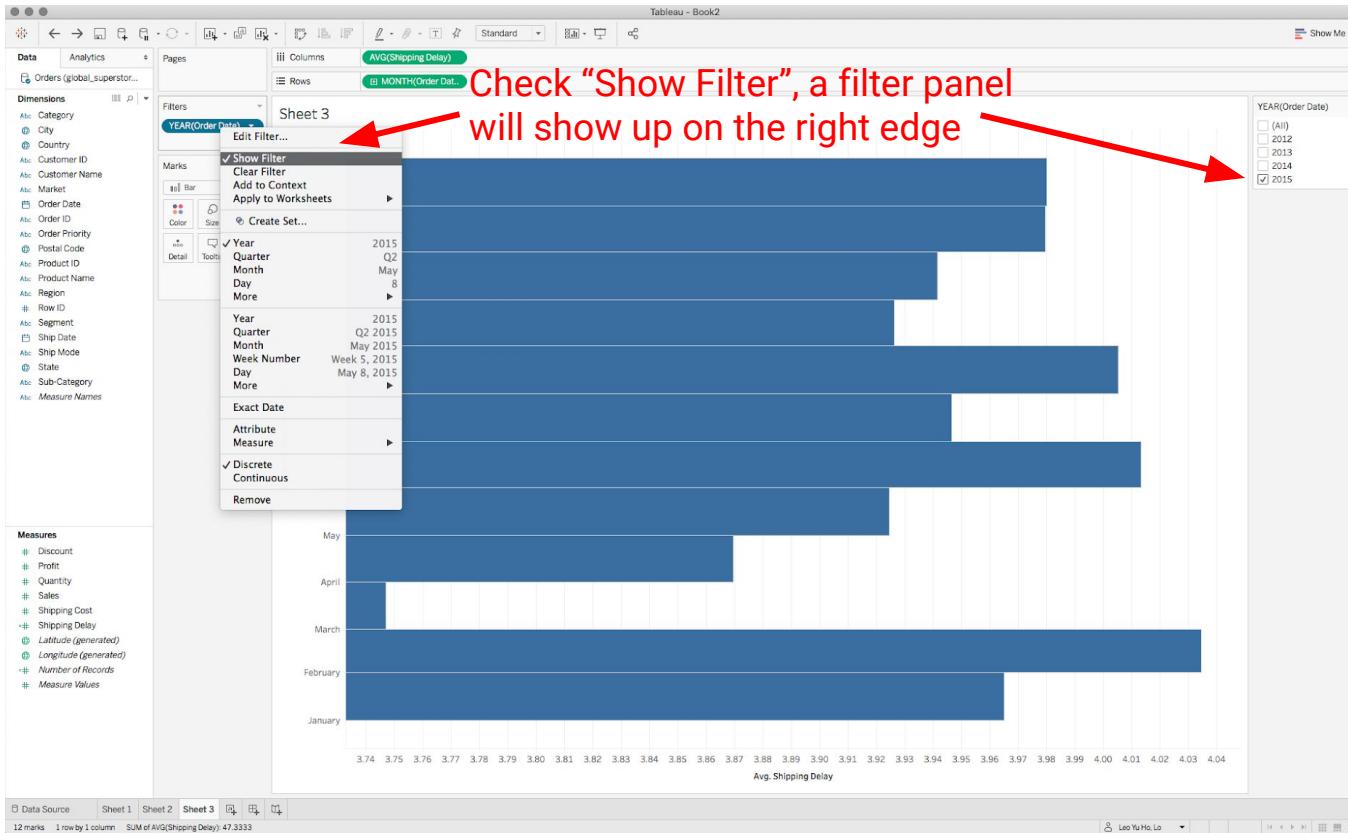
Reset Apply Cancel OK

Data Source Sheet 1 Sheet 2 Sheet 3 4 marks 4 rows by 1 column SUM of AVG(Shipping Delay): 15.88247

Leo Yu Ho Lo



Interactive filtering



Lab exercise

- Tasks
 - Download dataset from [GitHub](#)
 - Import data from Excel file
 - Create a line chart by drag and drop to columns, rows and color marks
 - Column: Month of Order Date
 - Row: Sum of Sales
 - Color: Category
 - Take a screenshot and upload to Canvas in .png format
 - Mac: cmd+shift+4
 - Windows: [Snipping Tool](#)
- Optional
 - Try using “Show me” to create different charts with different data
 - Plotting data on map, adjust color and size
 - Create “Calculated Field”
 - Add an interactive filter

More topics on Tableau

- Coursera course
 - <https://www.coursera.org/learn/analytics-tableau>
- Tableau training videos
 - <https://www.tableau.com/learn/training>
- Tableau Viz Gallery
 - <https://www.tableau.com/solutions/gallery>
- Other notable features of Tableau
 - Dashboard, Storyboard
 - Parameters
 - Grouping
 - Table join
 - Features in “Analytics” tab, e.g. Trend Line, Cluster
 - Quick table calculation (e.g. running sum)
 - Tableau Prep
 - Import data from pdf

Next Tutorial

Fantastic
Visualizations and
Datasets
Where to find them?

- Prepare for project and top-vis competition
- No lab exercise
- Help you to form groups