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 you own program to crawl
  - From API
  - Extract from HTML or JSON
  - Python: Scrapy, Beautiful Soup
  - 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

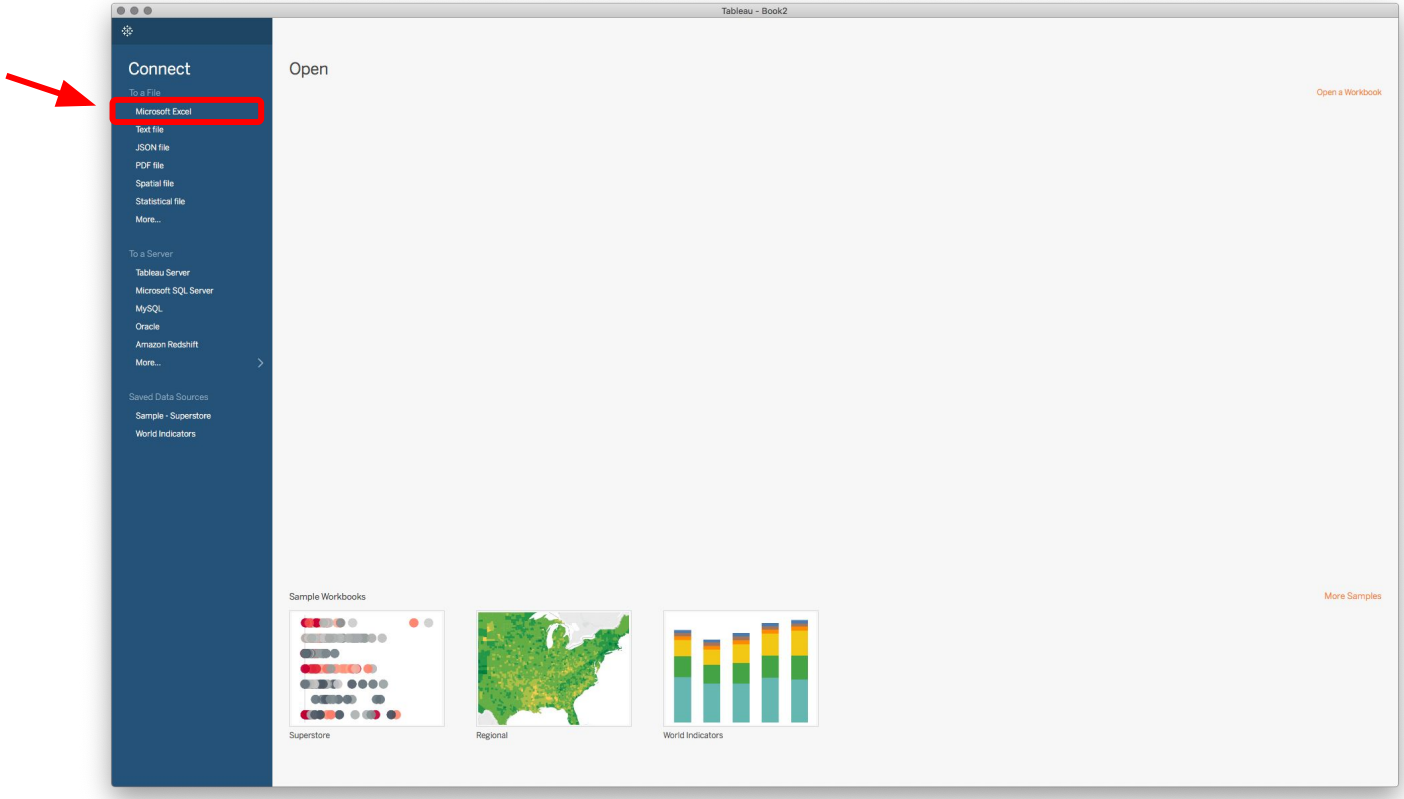


Tableau - Book2
Connect
To a File
Microsoft Excel
Text file
JSON file
PDF file
Spatial file
Statistical file
More...
To a Server
Tableau Server
Microsoft SQL Server
MySQL
Oracle
Amazon Redshift
More...
Saved Data Sources
Sample - Superstore
World Indicators
Open
Open a Workbook
Sample Workbooks
More Samples
Superstore
Regional
World Indicators

Load Data

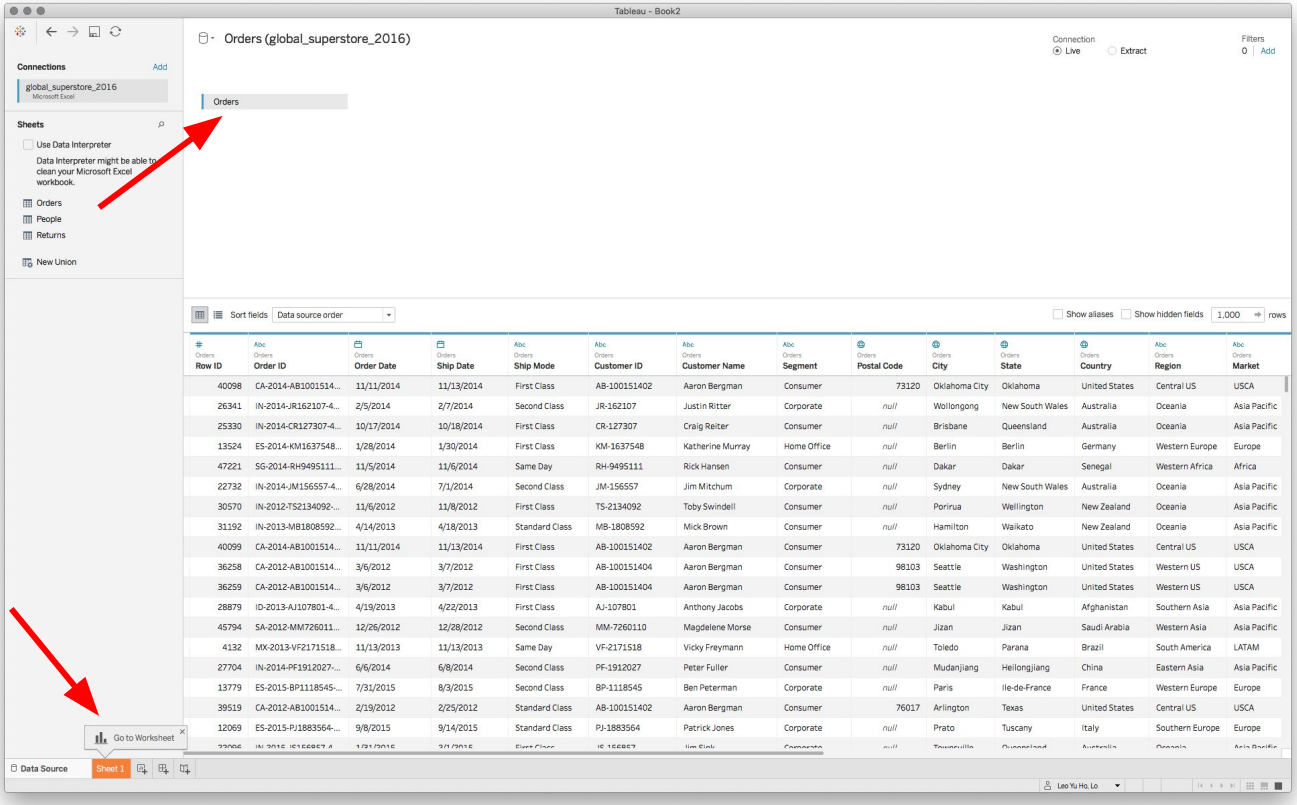
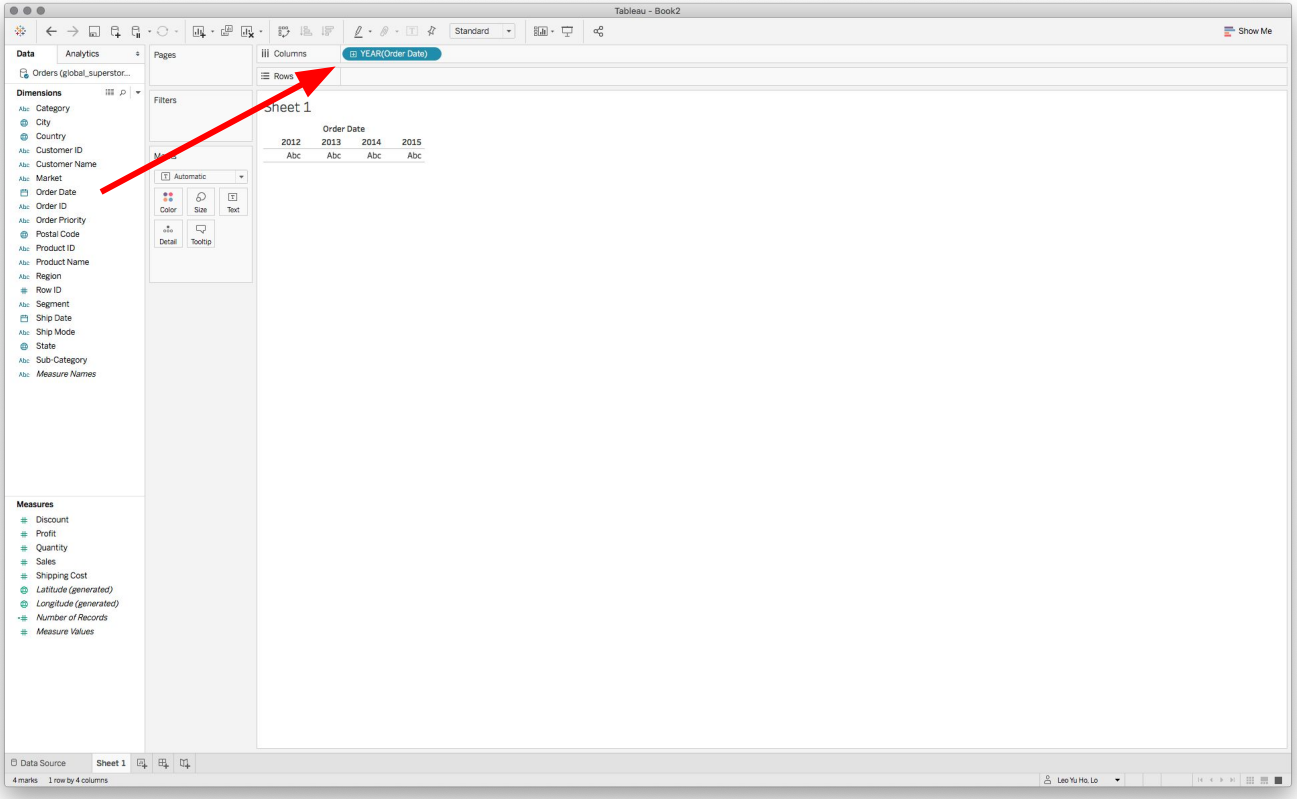
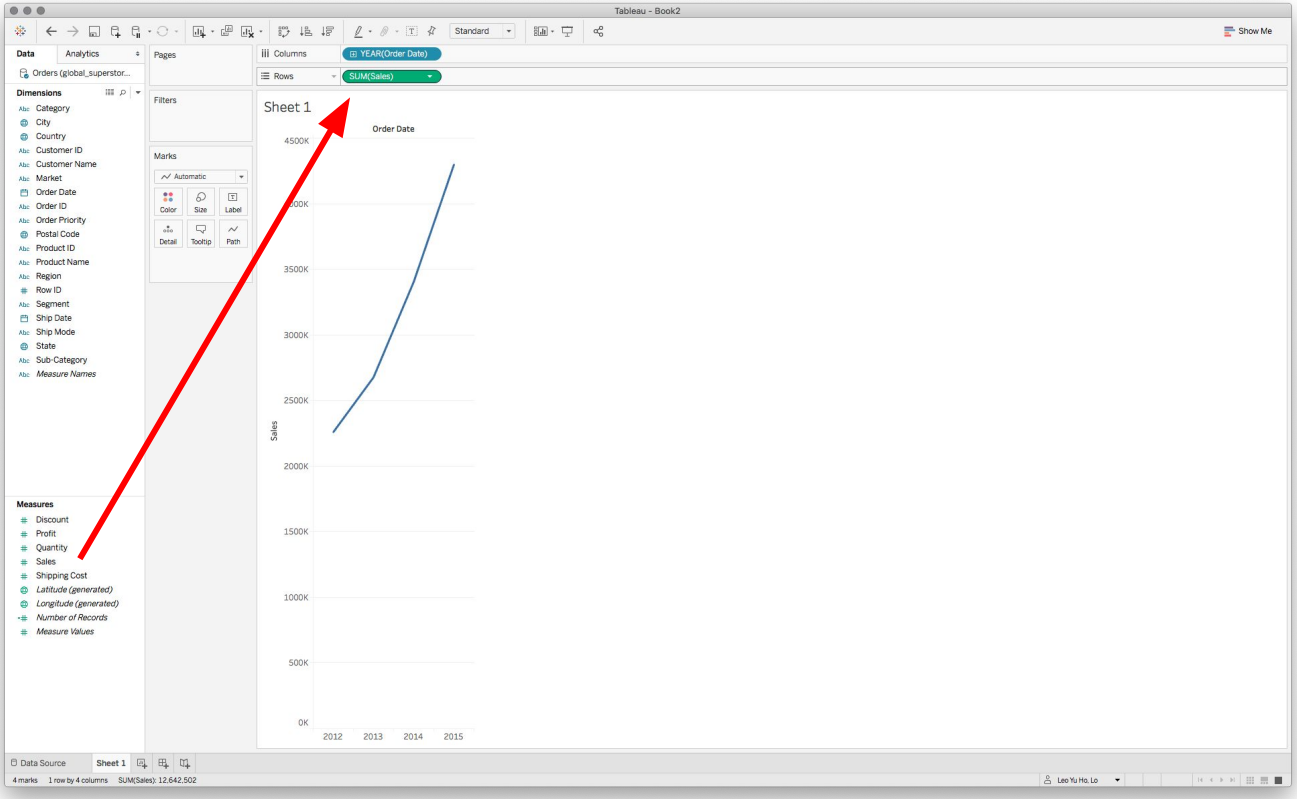


Tableau - Book2
Orders (global_superstore_2016)
Connection
Live
Extract
Filters
0
Add
Connections
Add
global_superstore_2016
Microsoft Excel
Sheets
Use Data Interpreter
Data Interpreter might be able to clean your Microsoft Excel workbook.
Orders
People
Returns
New Union
Orders
Sort fields
Data source order
Show aliases
Show hidden fields
1,000
rows
Row ID
Order ID
Order Date
Ship Date
Ship Mode
Customer ID
Customer Name
Segment
Postal Code
City
State
Country
Region
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-JR162107-4... 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-CR127307-4... 10/17/2014 10/18/2014 First Class CR-127307 Craig Reiter Consumer null Brisbane Queensland Australia Oceania Asia Pacific
13524 ES-2014-KM1637548... 1/28/2014 1/30/2014 First Class KM-1637548 Katherine Murray Home Office null Berlin Berlin Germany Western Europe Europe
47221 SG-2014-RH9495111... 11/5/2014 11/6/2014 Same Day RH-9495111 Rick Hansen Consumer null Dakar Dakar Senegal Western Africa Africa
22732 IN-2014-JM156557-4... 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-TS2134092-... 11/6/2012 11/8/2012 First Class TS-2134092 Toby Swindell Consumer null Porirua Wellington New Zealand Oceania Asia Pacific
31192 IN-2013-MB1808592... 4/14/2013 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-AJ107801-4... 4/19/2013 4/22/2013 First Class AJ-107801 Anthony Jacobs Corporate null Kabul Kabul Afghanistan Southern Asia Asia Pacific
45794 SA-2012-MM726011... 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-VF2171518... 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 Asia Asia Pacific
13779 ES-2015-BP1118545-... 7/31/2015 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-PJ1883564-... 9/8/2015 9/14/2015 Standard Class PJ-1883564 Patrick Jones Corporate null Prato Tuscany Italy Southern Europe Europe
Go to Worksheet
Data Source
Sheet 1
Leo Yu Ho, Lo

Basic Plotting: Select Row and Column



Basic Plotting: Select Row and Column



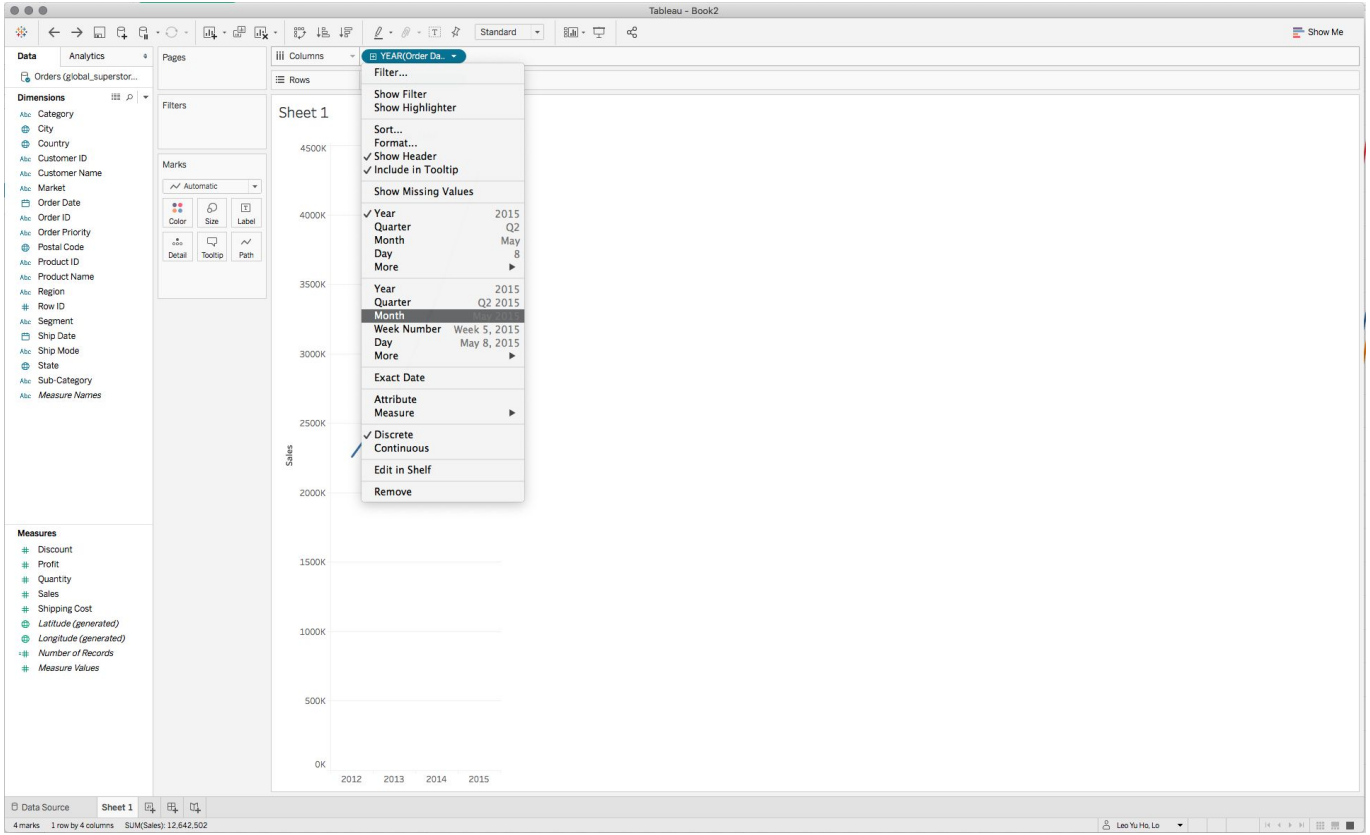
Basic Plotting: Adjust Date

"Dimensions" are **discrete** variables

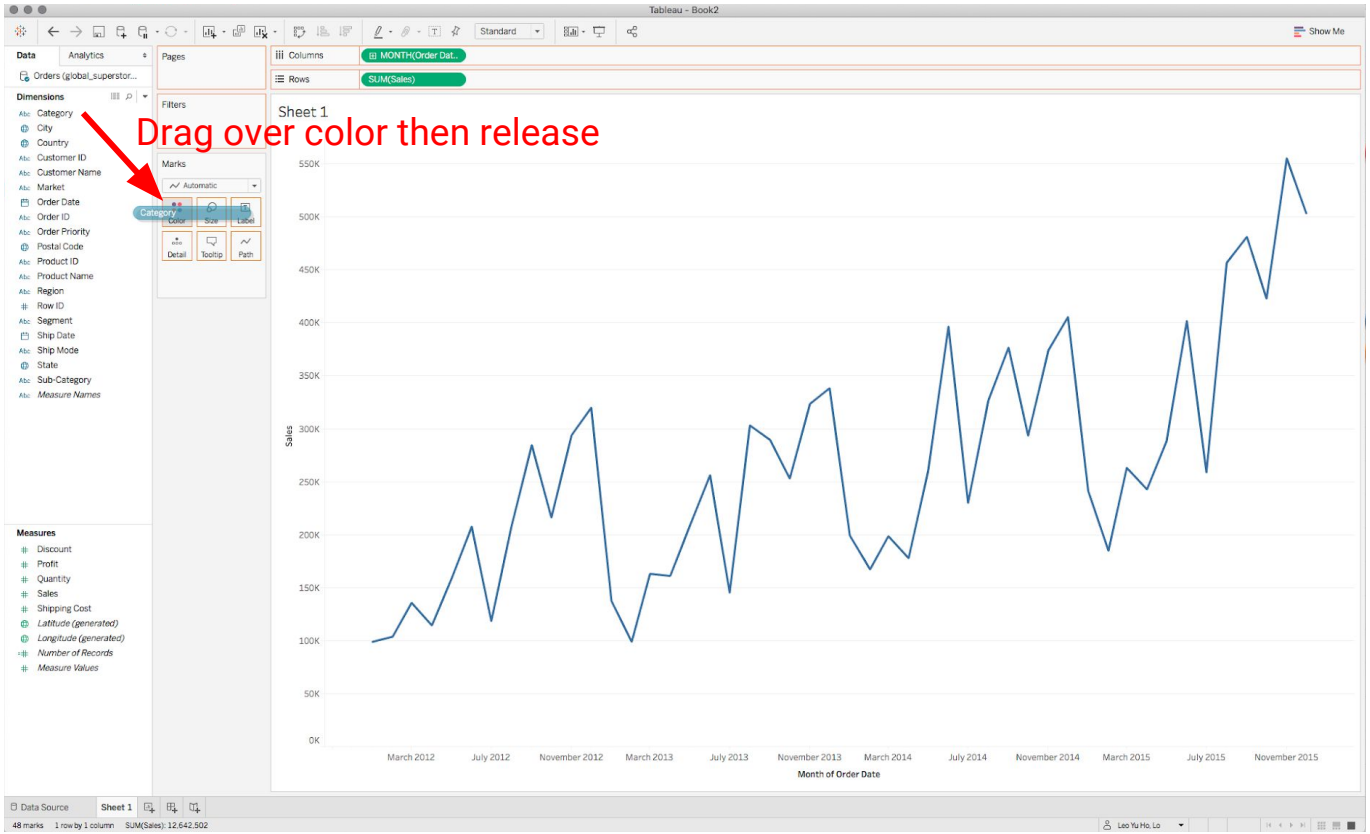
"Measures" are **continuous** variables

Date can be either **discrete** or **continuous**

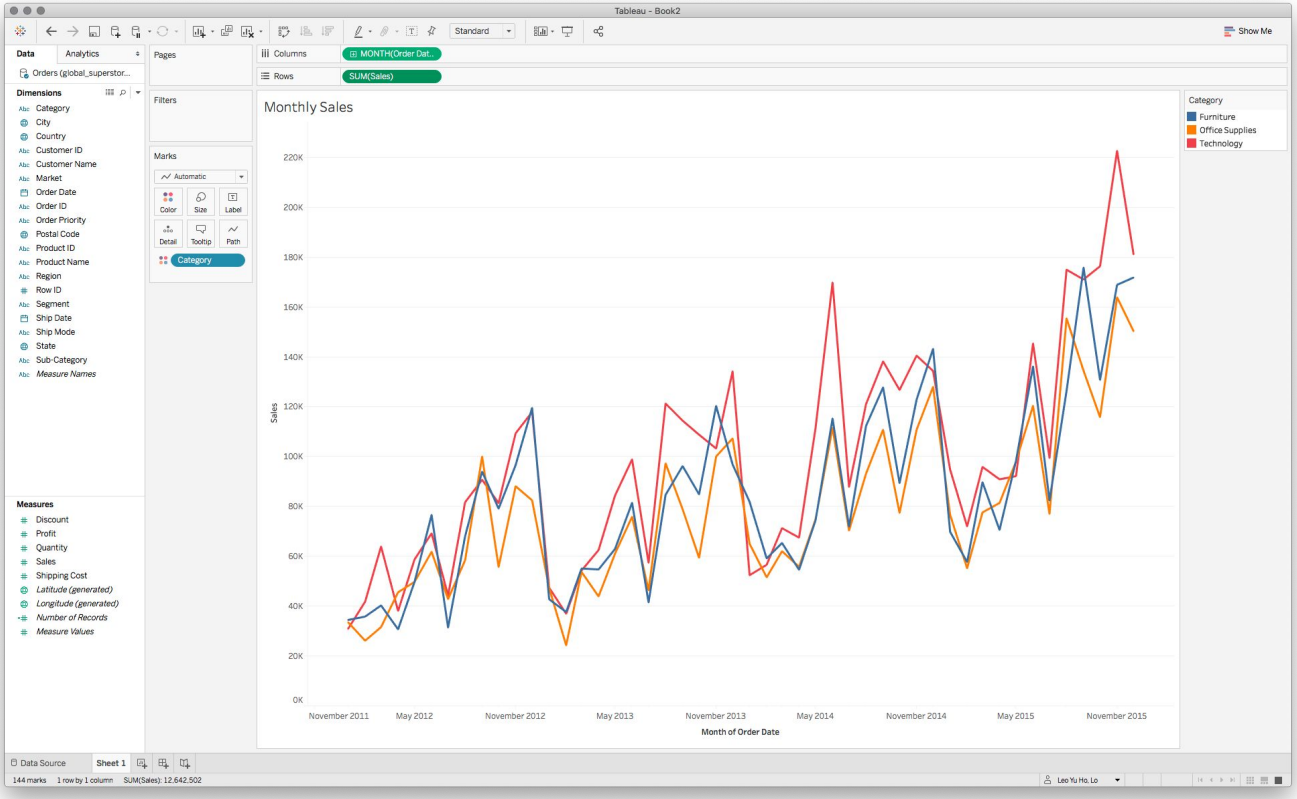
We need it to be **continuous** right now, you can try **discrete** and see what happens



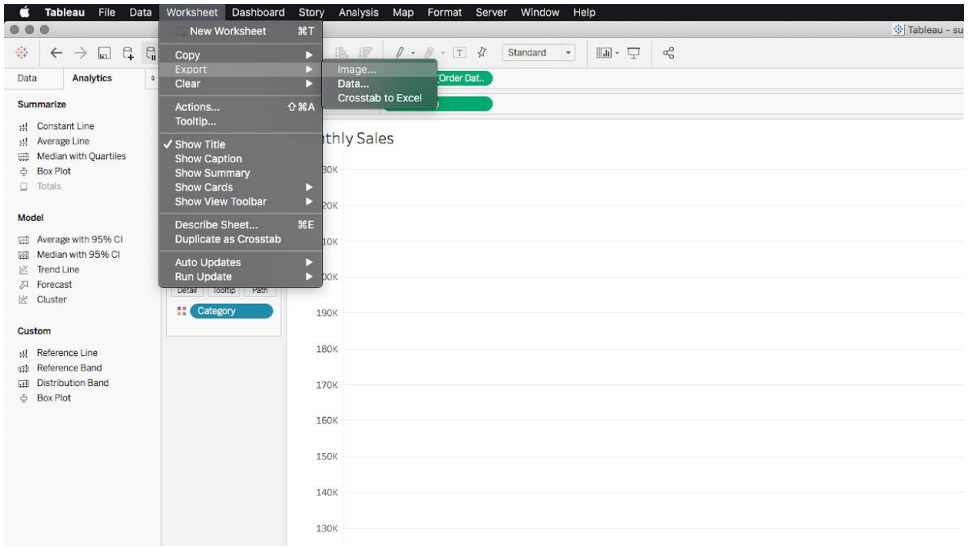
Basic Plotting: Marks with Color



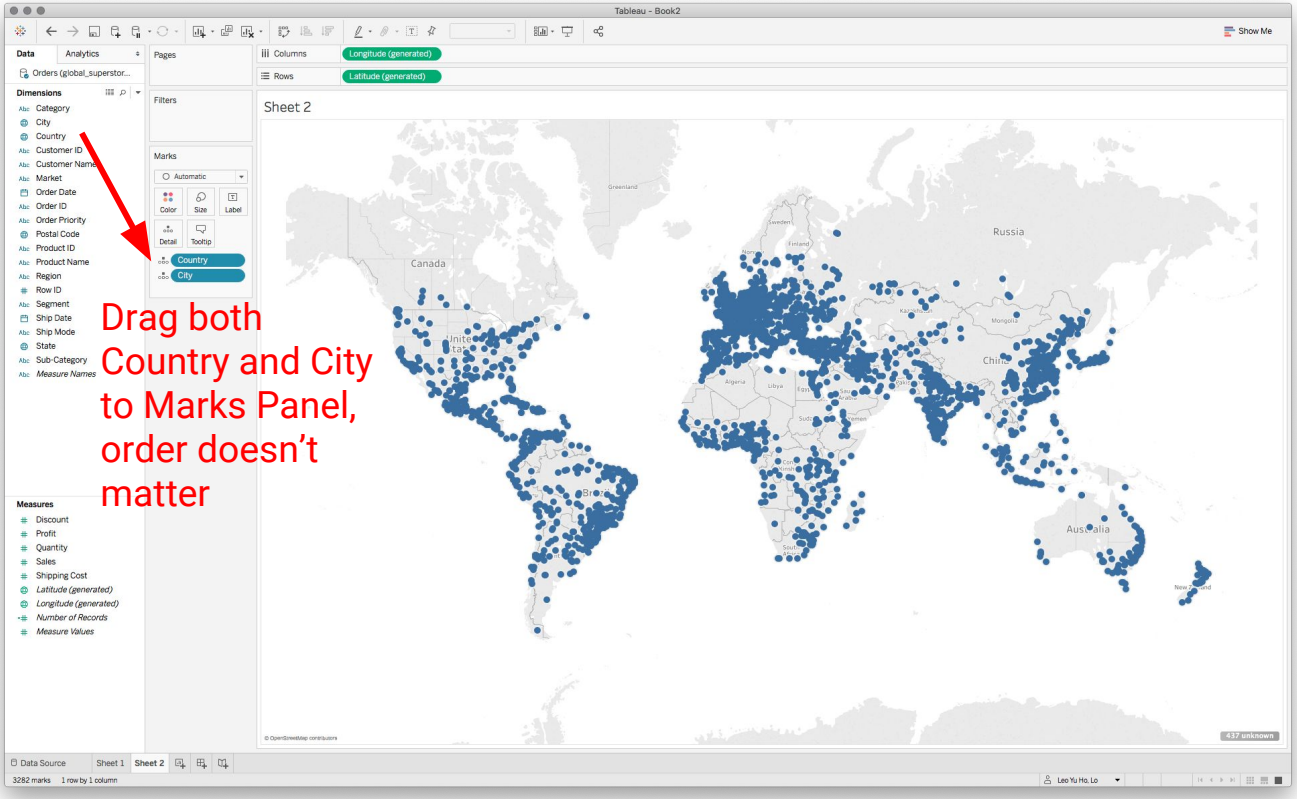
Basic Plotting: Ta-Da



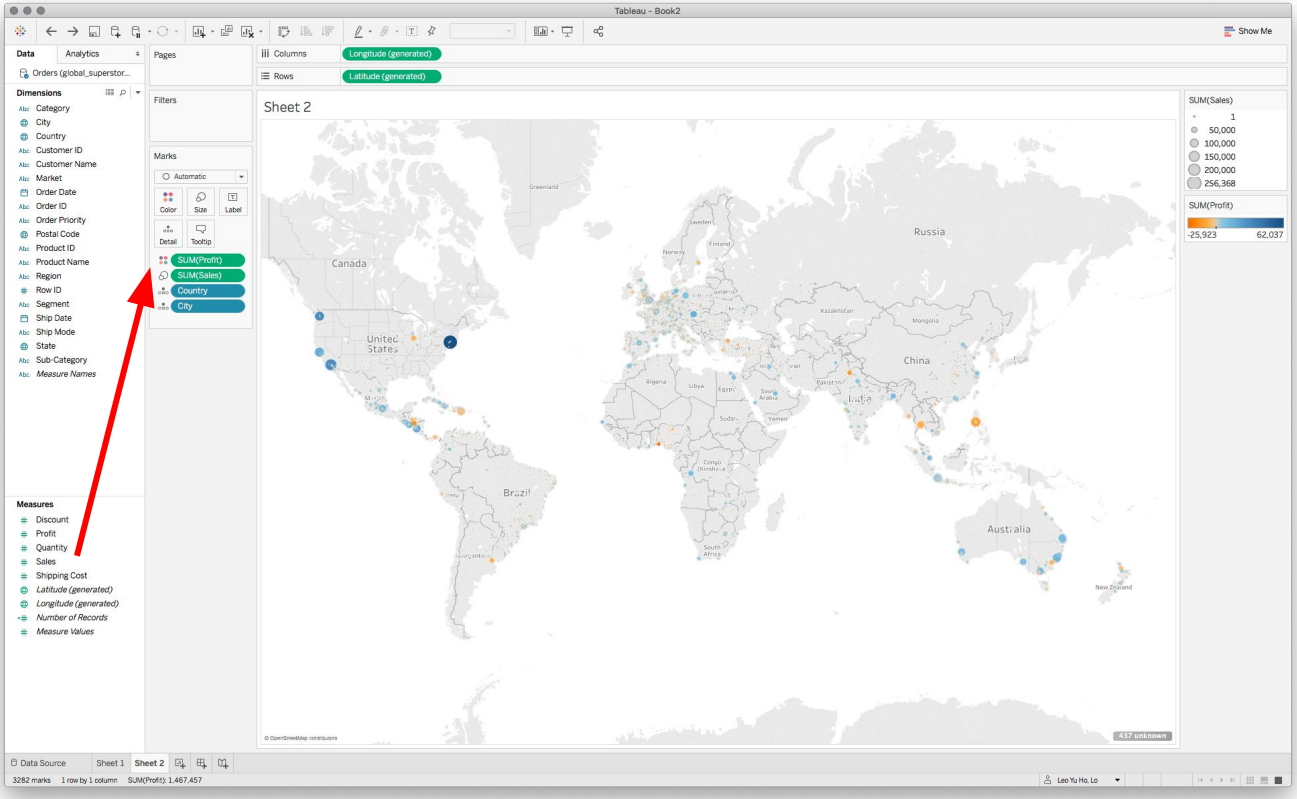
Export Image (Not available in Tableau Public)



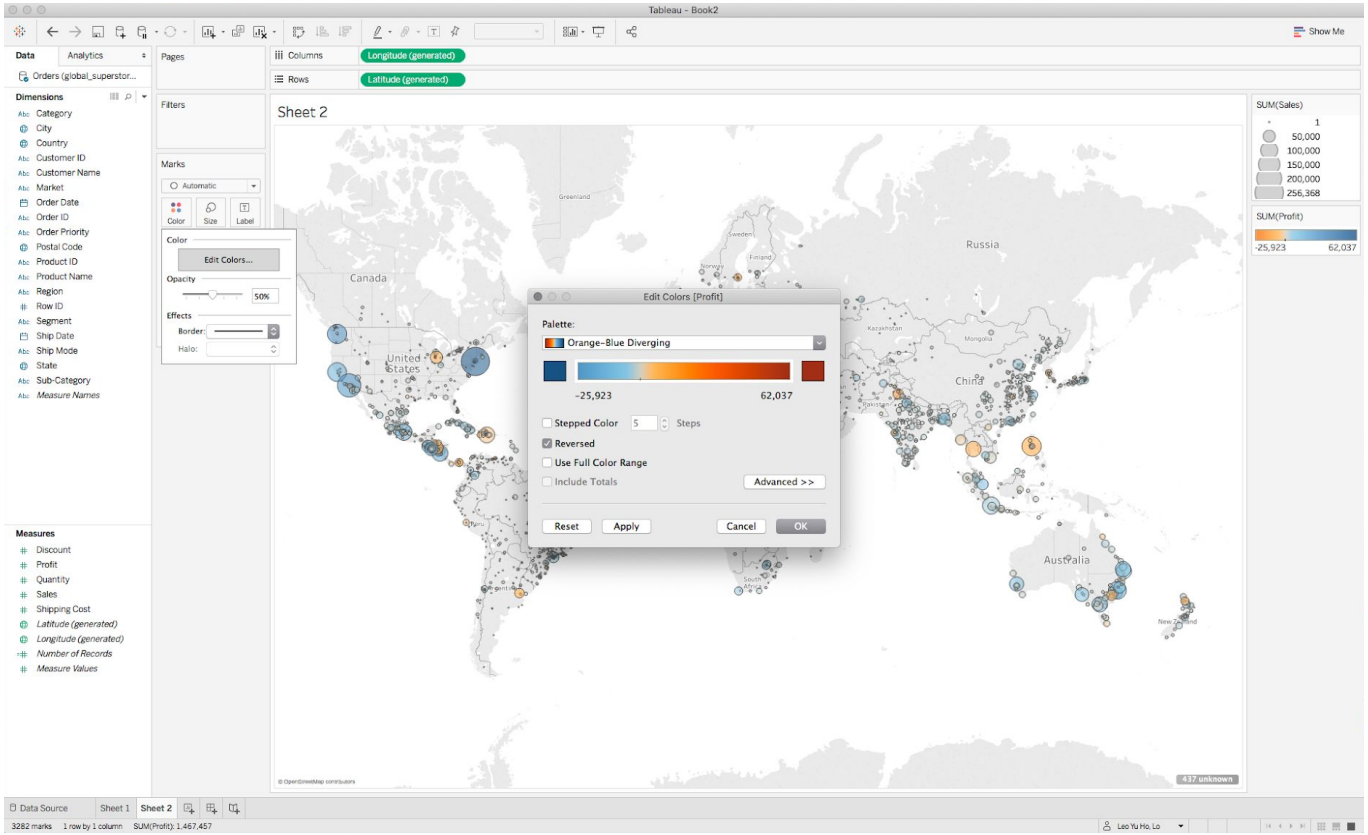
Plotting with Map



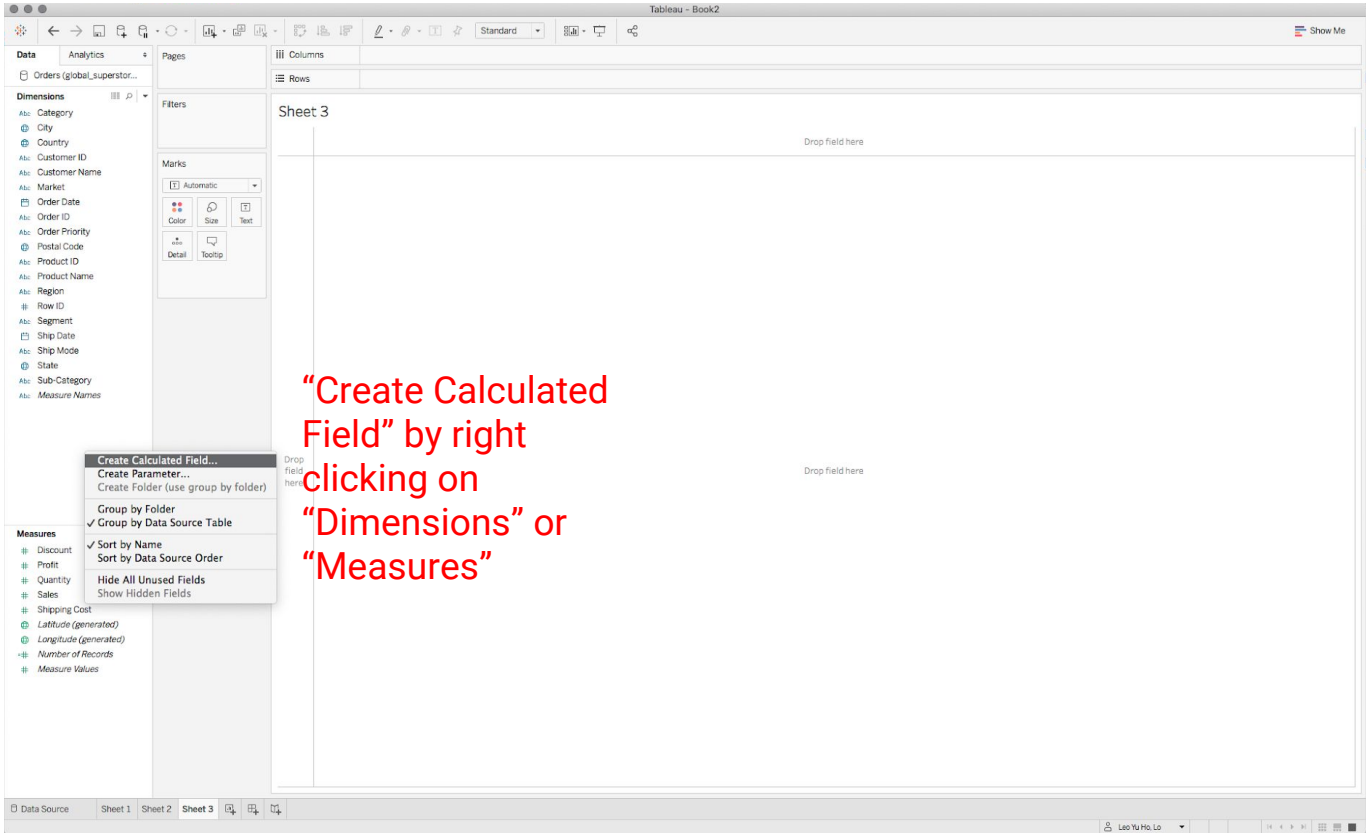
Plotting with Map: Encode with Size and Color



Adjust Color and Size

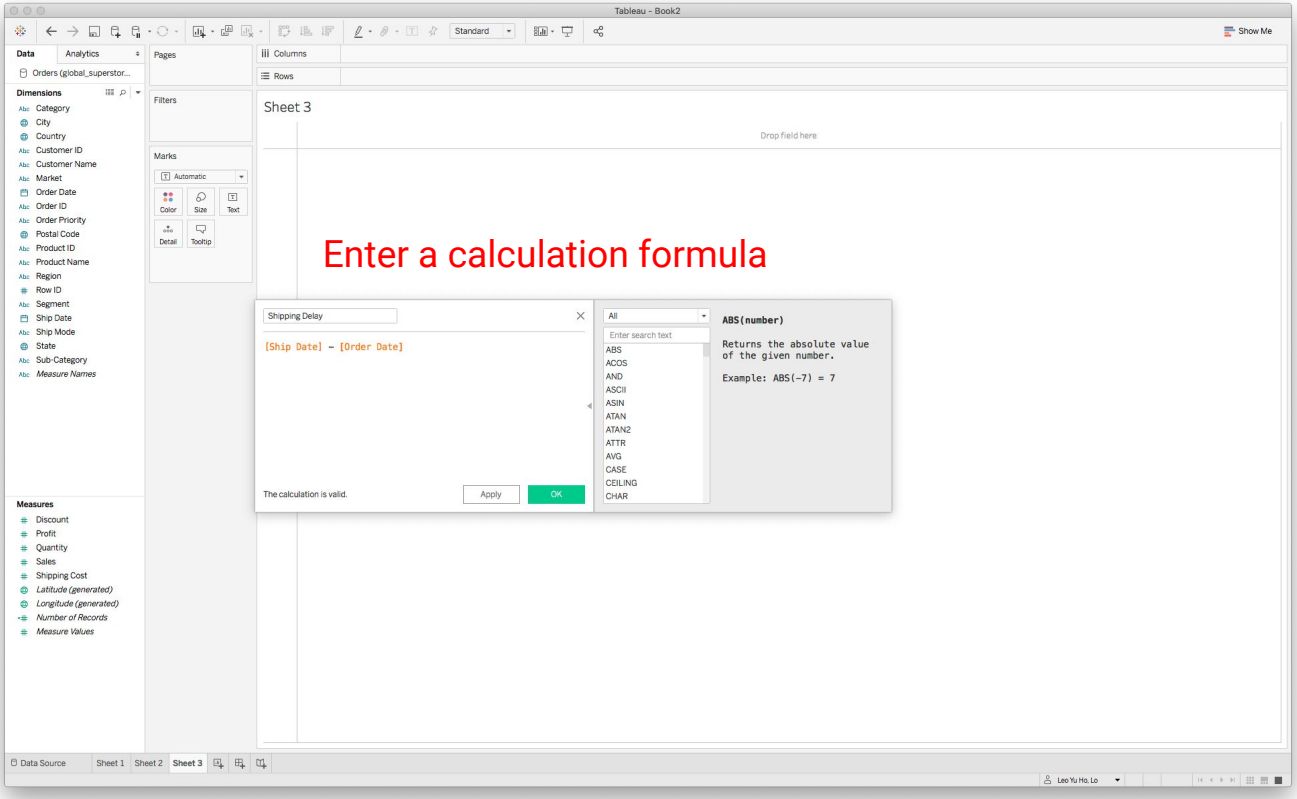


Calculated Field



"Create Calculated Field" by right clicking on "Dimensions" or "Measures"

Calculated Field



Aggregate

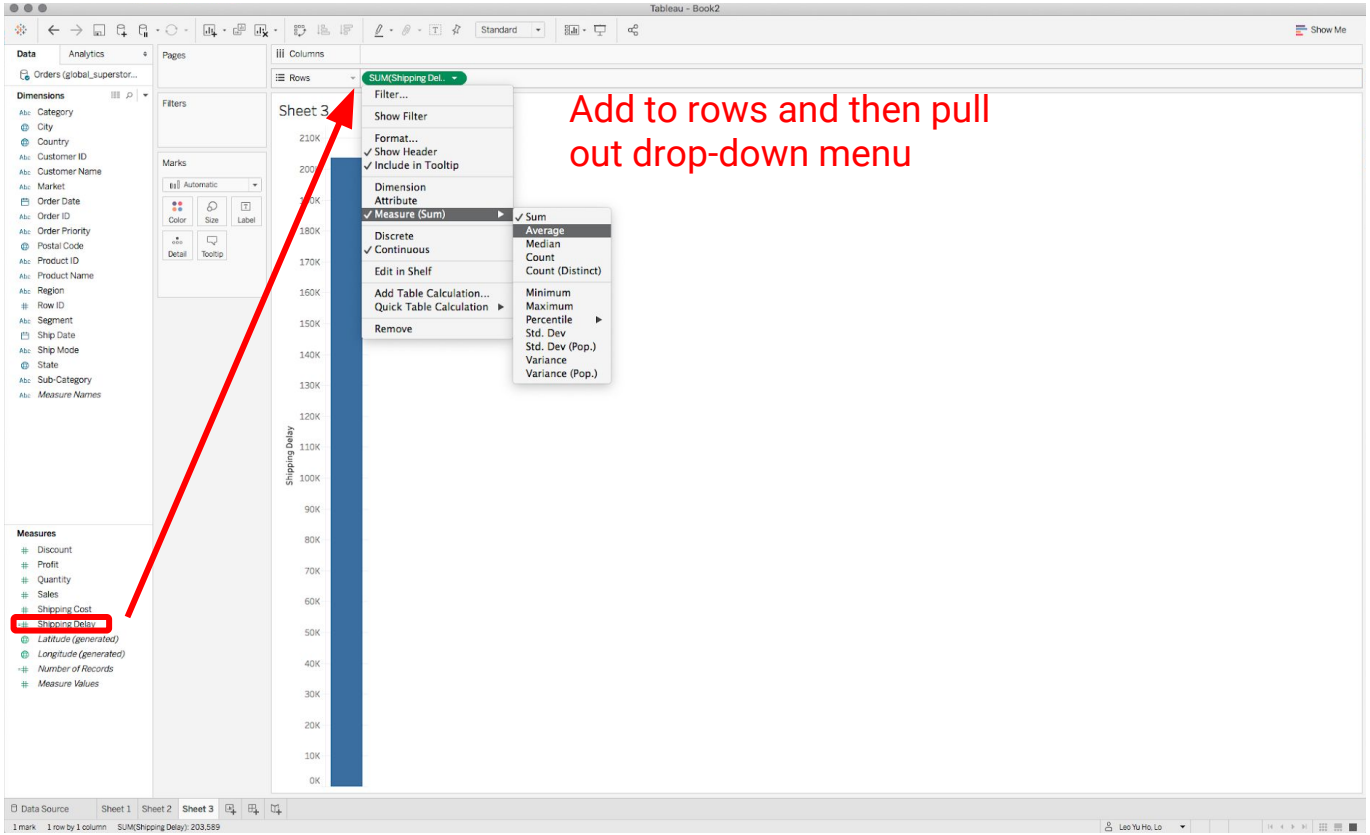
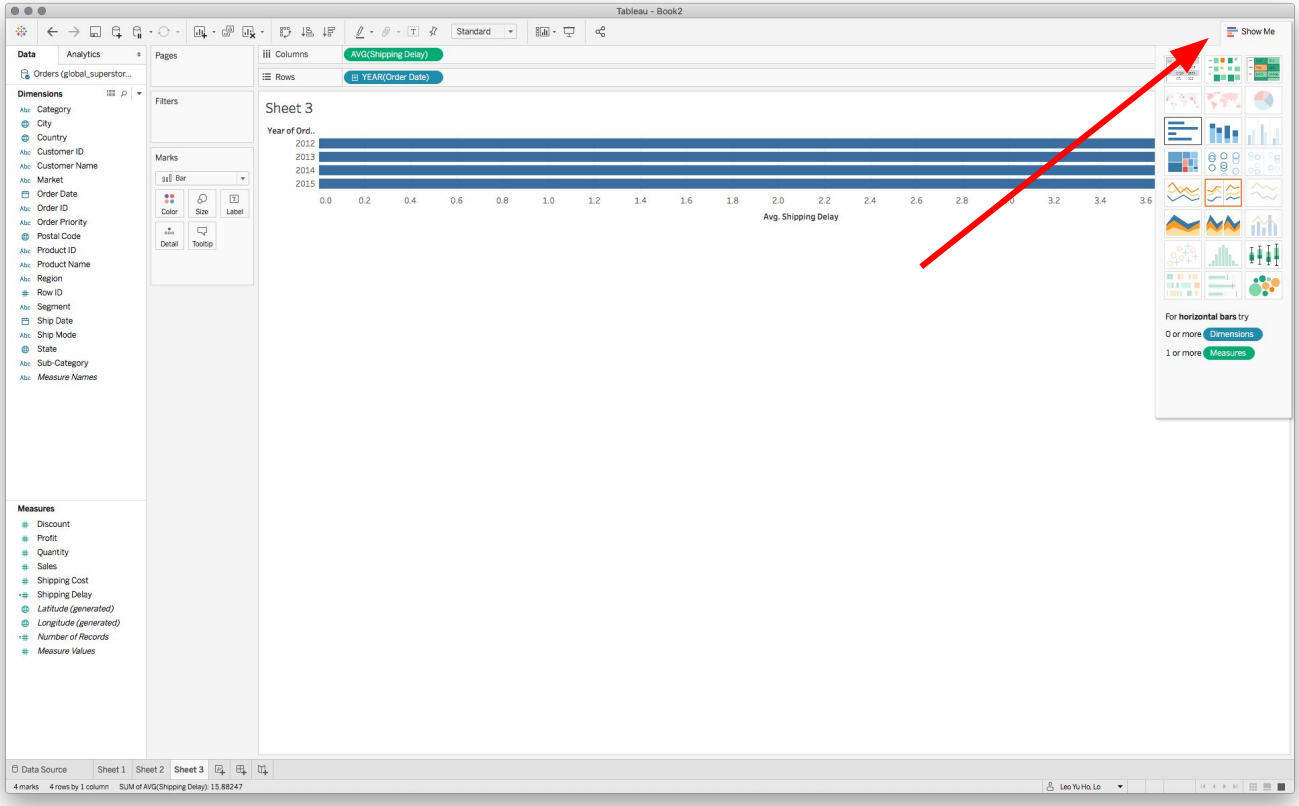


Tableau - Book2
Add to rows and then pull out drop-down menu
Sum
Average
Median
Count
Count (Distinct)
Minimum
Maximum
Percentile
Std. Dev
Std. Dev (Pop.)
Variance
Variance (Pop.)

Plotting is easy using “Show Me”



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

Try different combinations for different plots!

Filter

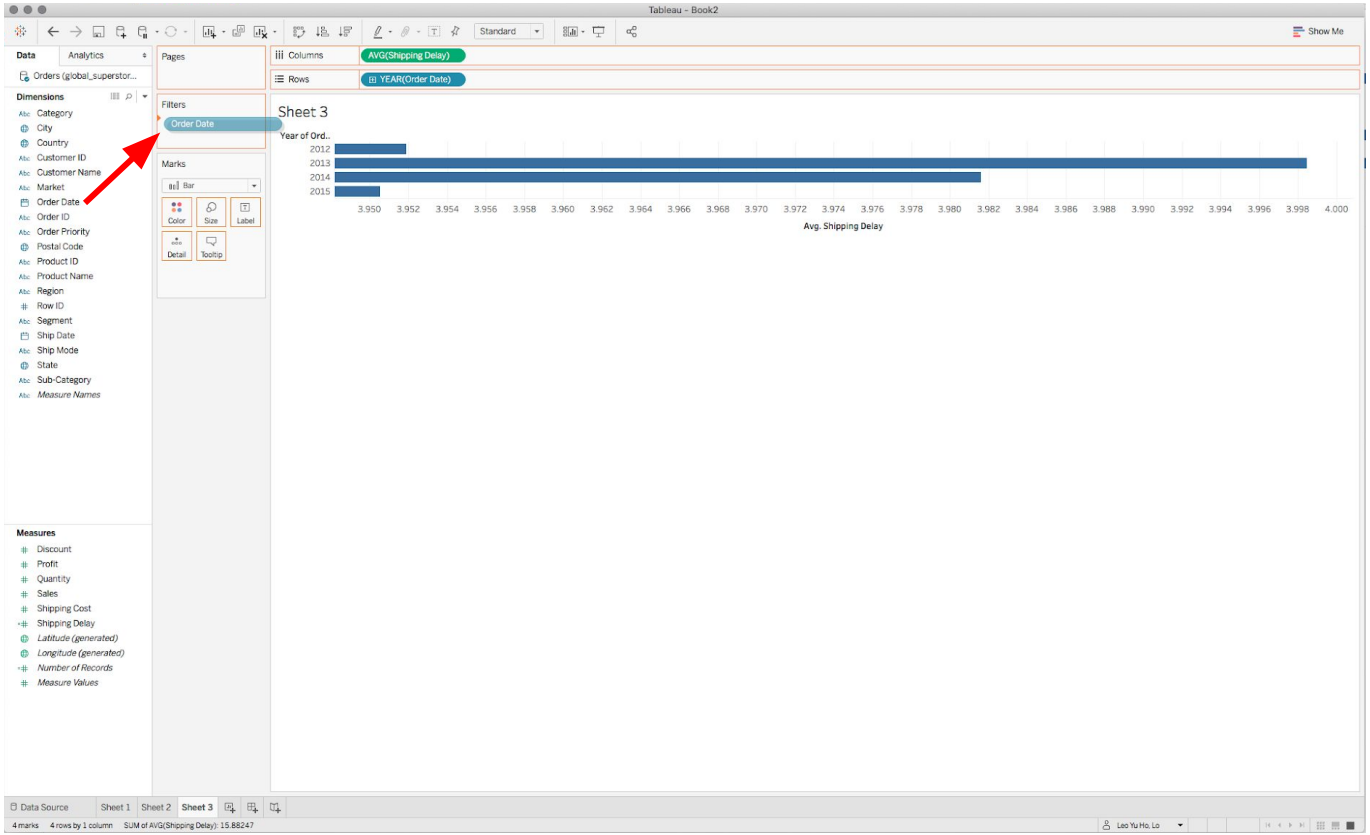


Tableau - Book2
Standard
Show Me
Data
Analytics
Orders (global_superstor...
Dimensions
Category
City
Country
Customer ID
Customer Name
Market
Order Date
Order ID
Order Priority
Postal Code
Product ID
Product Name
Region
Row ID
Segment
Ship Date
Ship Mode
State
Sub-Category
Measure Names
Measures
Discount
Profit
Quantity
Sales
Shipping Cost
Shipping Delay
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
Order Date
Marks
Bar
Color
Size
Label
Detail
Tooltip
Columns
AVG(Shipping Delay)
Rows
YEAR(Order Date)
Sheet 3
Year of Ord..
2012
2013
2014
2015
3.950
3.952
3.954
3.956
3.958
3.960
3.962
3.964
3.966
3.968
3.970
3.972
3.974
3.976
3.978
3.980
3.982
3.984
3.986
3.988
3.990
3.992
3.994
3.996
3.998
4.000
Avg. Shipping Delay
Data Source
Sheet 1
Sheet 2
Sheet 3
4 marks
4 rows by 1 column
SUM of AVG(Shipping Delay): 15.88247

Filter

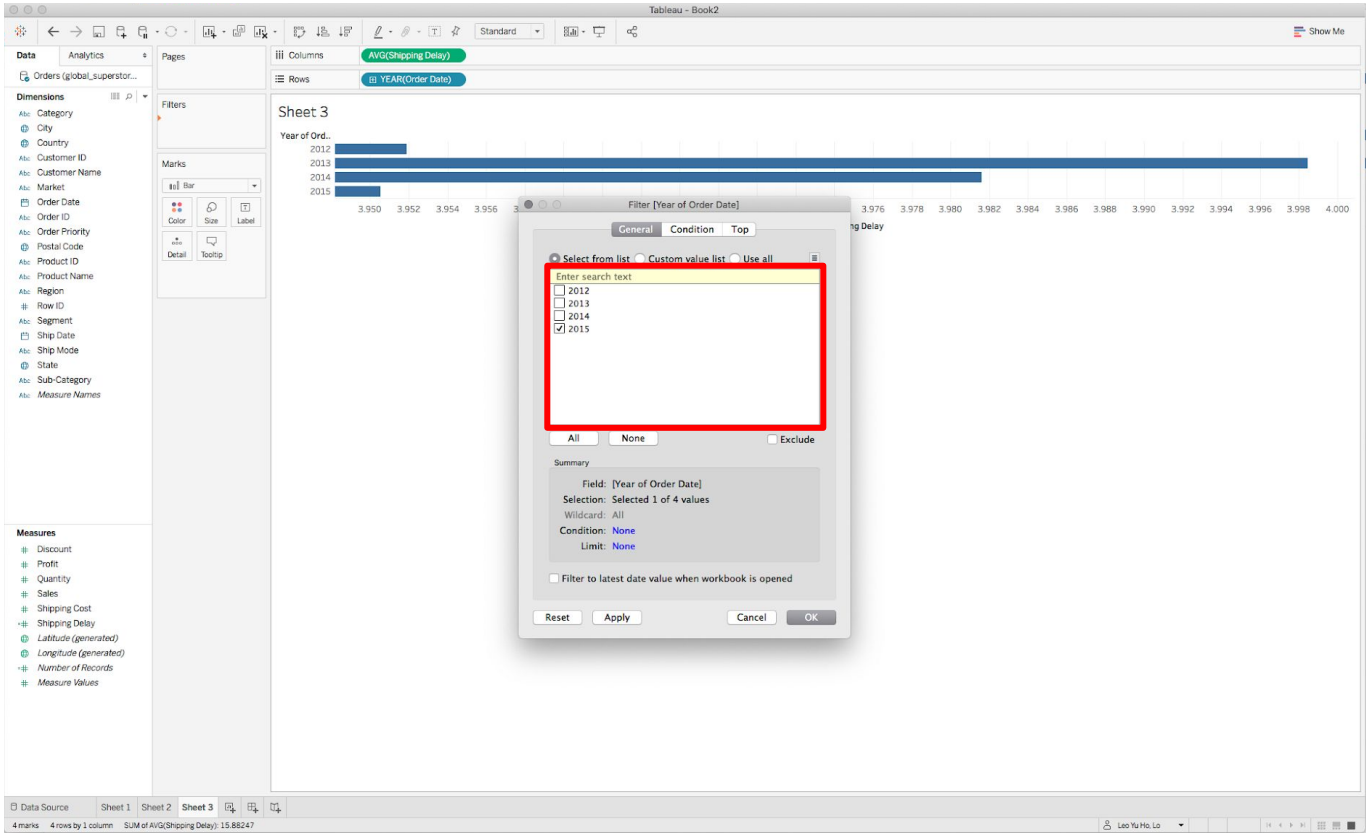
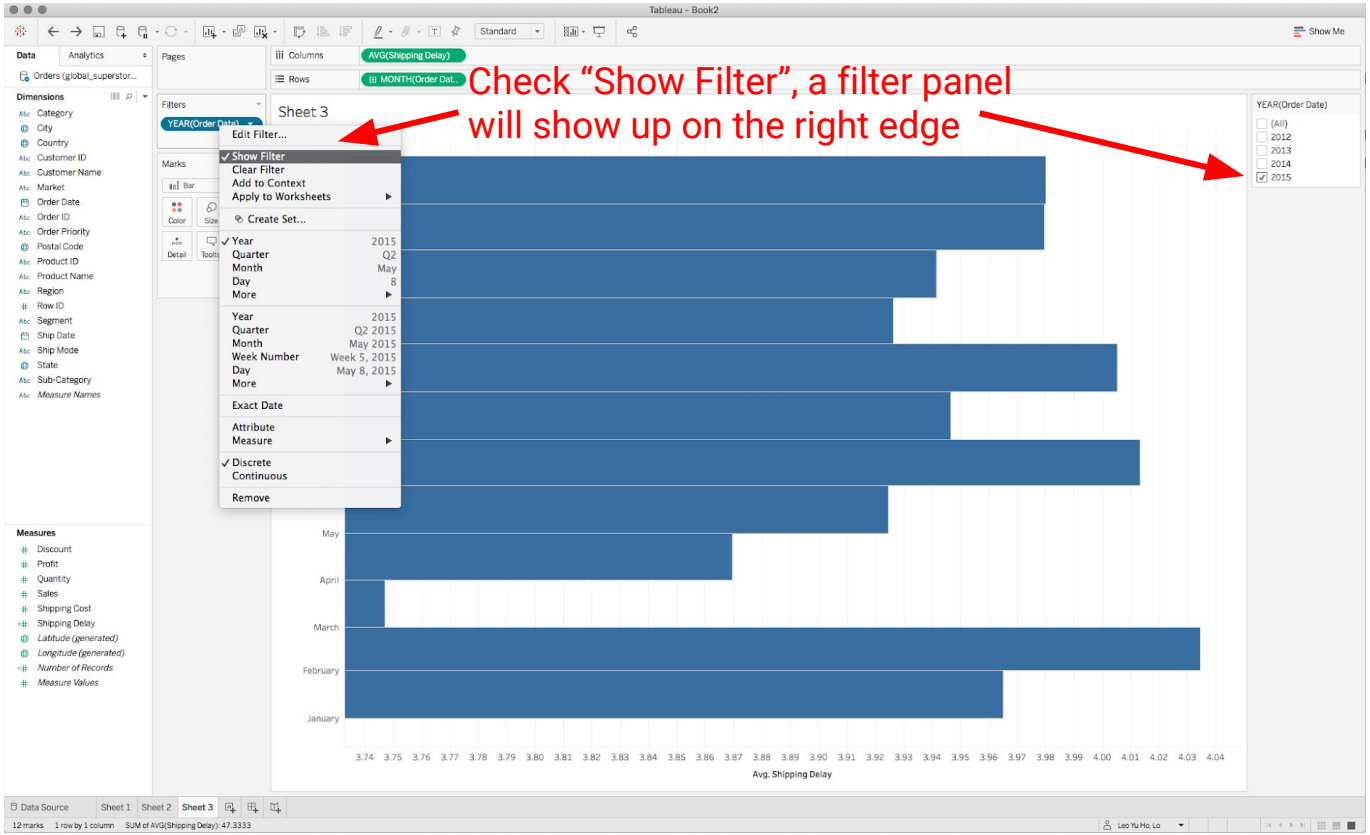


Tableau - Book2
Standard
Show Me
Data
Analytics
Orders (global_superstor...
Dimensions
Category
City
Country
Customer ID
Customer Name
Market
Order Date
Order ID
Order Priority
Postal Code
Product ID
Product Name
Region
Row ID
Segment
Ship Date
Ship Mode
State
Sub-Category
Measure Names
Measures
Discount
Profit
Quantity
Sales
Shipping Cost
Shipping Delay
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
Marks
Bar
Color
Size
Label
Detail
Tooltip
Columns
AVG(Shipping Delay)
Rows
YEAR(Order Date)
Sheet 3
Year of Ord..
2012
2013
2014
2015
3.950
3.952
3.954
3.956
3.976
3.978
3.980
3.982
3.984
3.986
3.988
3.990
3.992
3.994
3.996
3.998
4.000
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

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