

# Statistical techniques in Tableau



# Data Preparation

# Data preparation

ANALYZING DATA IN TABLEAU



<https://www.kaggle.com/yingwurenjian/chicago-divvy-bicycle-sharing-data>

# Data preparation

Ask yourself...

- Do any fields need to be refined?
- Are there calculated fields we can create to more effectively tell our data story?
- Does the data contain fields that will allow for summaries or grouping at a higher level?
- Are there sufficient categorical fields to **slice and dice** your data?



# Divvy dataset: stations table



- `id` : ID attached to each station
- `name` : station name
- `latitude` : station latitude
- `longitude` : station longitude

# Divvy dataset: trips table

- *Trips taken between Jan - June, 2019*
- `trip id` : ID attached to each trip
- `bikeid` : ID attached to each bike
- `tripduration` : time of trip in seconds
- `starttime` : day and time trip started (CST)
- `endtime` : day and time trip ended (CST)
- `from station id` : station ID of trip start
- `from_station_name` : station name of start
- `to station id` : station ID of trip end



- `to station name` : station name of end
- `usertype` : *customer* or *subscriber*
- `birthyear` : birth year of rider
- `gender` : gender of rider



# Dimension and measure recap

## Dimensions:

- Categorical or qualitative data

## Measures:

- Numerical data that can be aggregated

**We want to move fields strategically between these two types:**

- Move numeric fields that shouldn't be aggregated to the Dimensions section



The screenshot displays the Tableau Public interface with the 'Prep & Create Data' workspace active. The left sidebar shows a list of fields organized into two tables: 'Stations' and 'Trips'. The 'Stations' table includes fields like 'Id', 'Name', 'Docks', 'Latitude', and 'Longitude'. The 'Trips' table includes fields like 'End Time', 'From Station Id', 'From Station Name', 'Gender', 'Start Time', 'To Station Id', 'To Station Name', 'Trip Id', 'Usertype', 'Bike Id', 'Birthyear', 'Tripduration', and 'Trips (Count)'. The main workspace is titled 'Prep & Create Data' and contains two large empty areas with the text 'Drop field here'. The top toolbar includes various icons for navigation, data manipulation, and visualization. The bottom status bar shows 'Data Source' and 'Prep & Create Data' tabs, along with a 'Show Me' button and a 'Go to Settings to activate Windows' message.

**Data** | Analytics | Pages | Columns | Rows

Search | Filters | Marks

**Tables**

- Stations**
  - # Id
  - Abc Name
  - # Docks
  - Latitude
  - Longitude
- Trips**
  - End Time
  - # From Station Id
  - Abc From Station Name
  - Abc Gender
  - Start Time
  - # To Station Id
  - Abc To Station Name
  - # Trip Id
  - Abc Usertype
  - # Bike Id
  - # Birthyear
  - # Tripduration
  - # Trips (Count)

Drop field here

Drop field here

Activate Windows  
Go to Settings to activate Windows.

# Table of Contents



1. Univariate exploratory data analysis:  
3
2. Measures of spread and confidence intervals
3. Bivariate exploratory data analysis
4. Forecasting and clustering

# exploratory data analysis



# Exploratory Data Analysis (EDA)

- Main characteristics of your data
- Spot extreme values
- Suggest hypotheses
- Assess assumptions

**General goal: get an idea of the overall structure of your data**

## Univariate EDA

- Summary table
- Bar plot
- Histogram
- Box plot

# Tables & bar plots

Visualize the distribution of a single, categorical variable

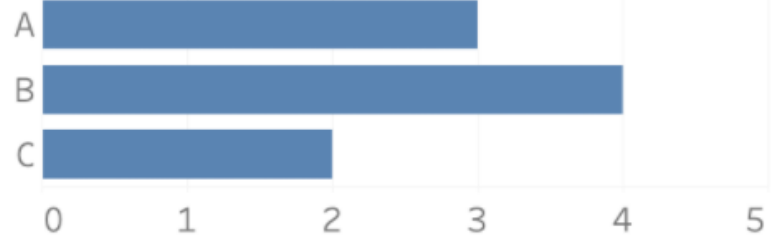
Category

A  
B  
A  
B  
B  
C  
A  
B  
C

Category

A	3
B	4
C	2

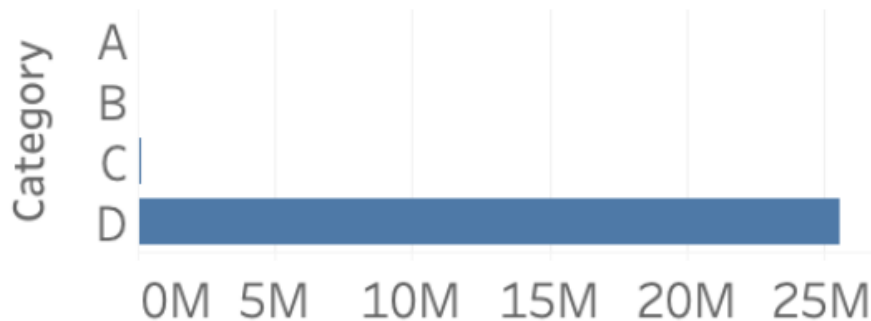
Category



# When to use a table vs. a plot

- Focus is on individual values (snapshot) and not on trends
- Dataset contains few values
- Small difference between values is crucial
- Data is presented in a non-interactive way

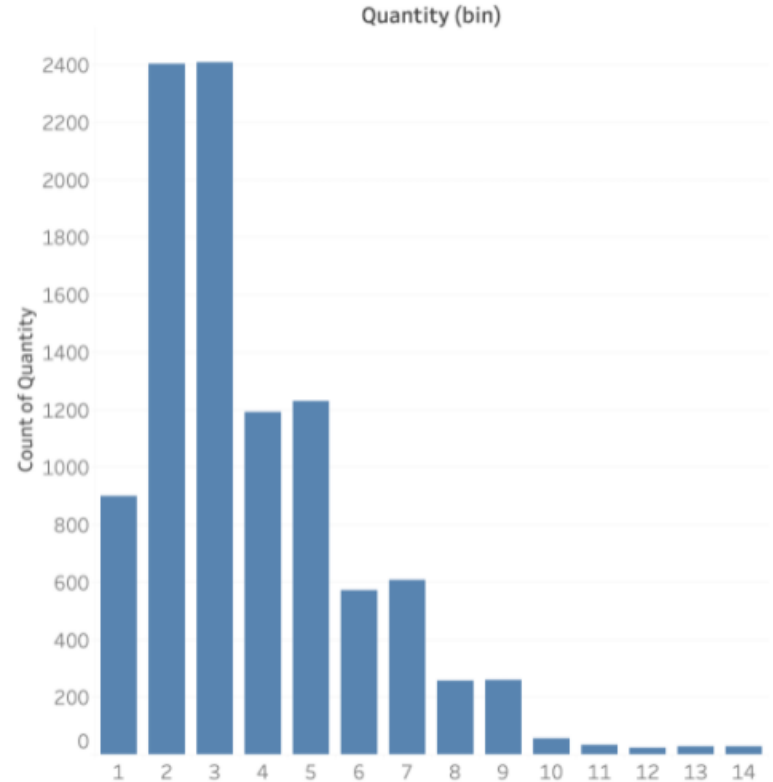
Category	
A	20
B	400
C	160.000
D	25.600.000



# Histograms

Visualize the distribution of a single, continuous variable

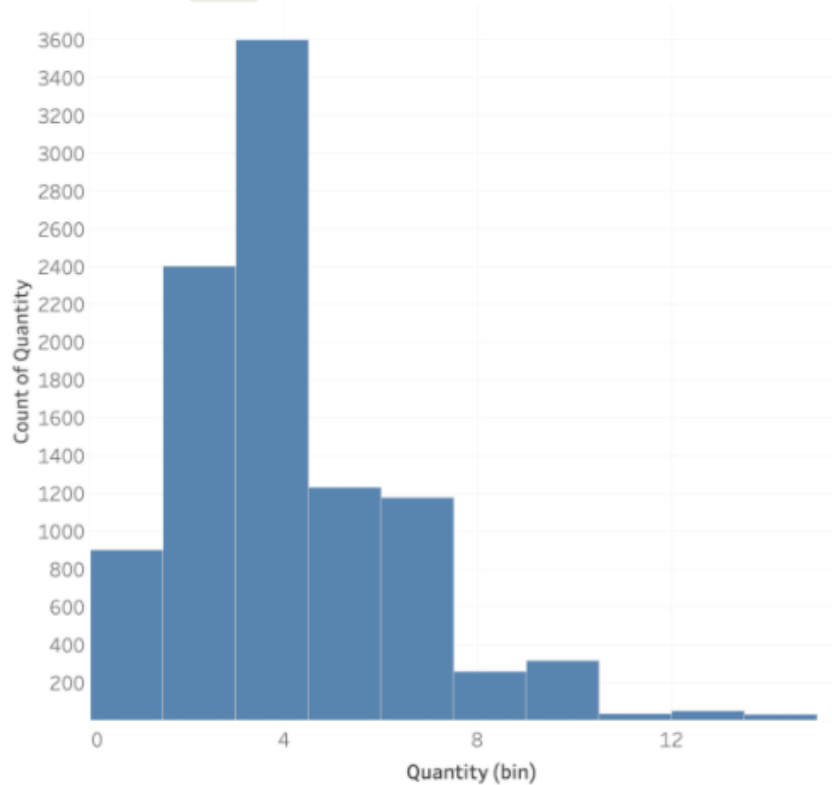
- Lowest/highest value
- Most common value(s)
- Splitting variable in bins



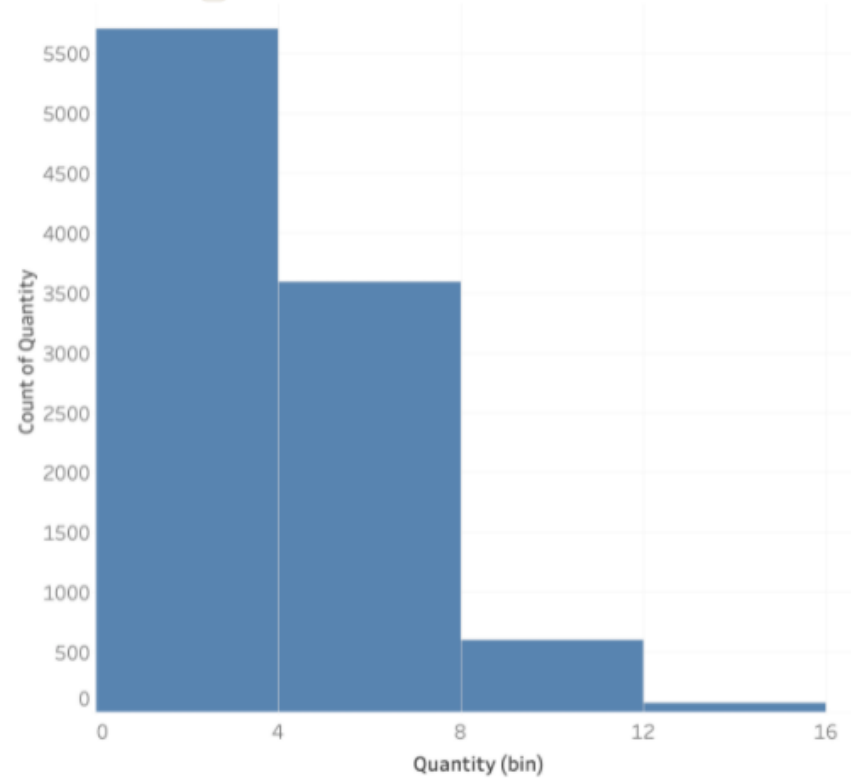


# Size of bins

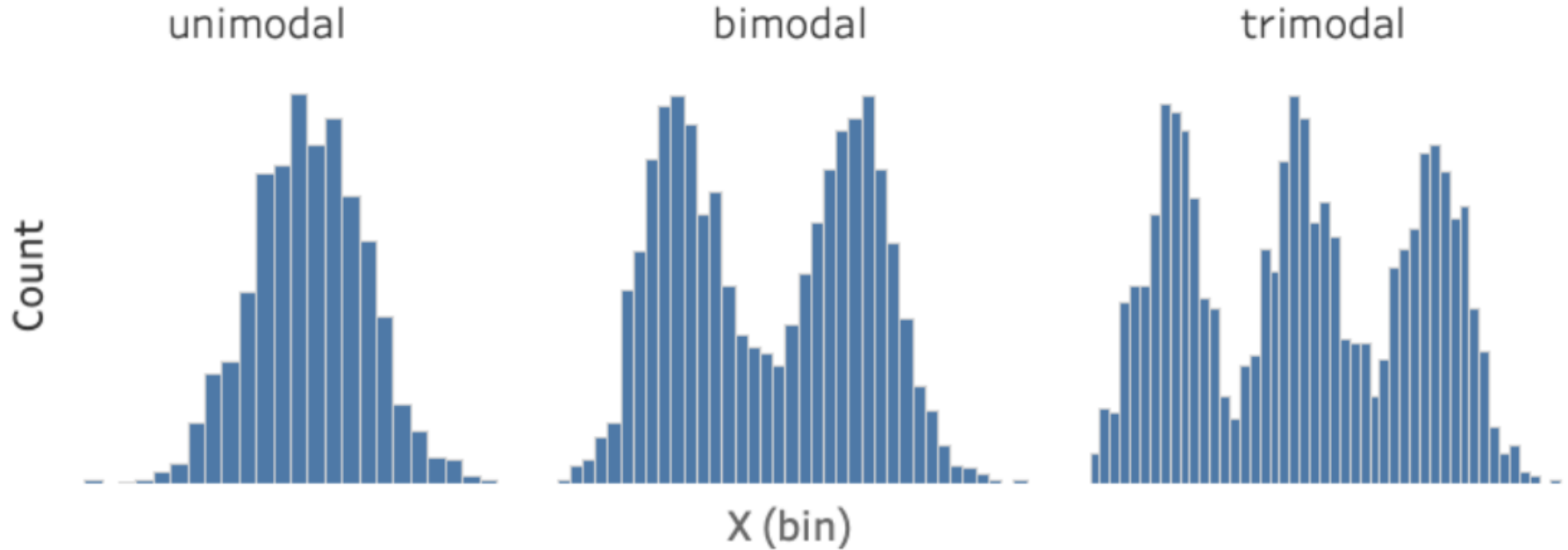
Binwidth = 1.5

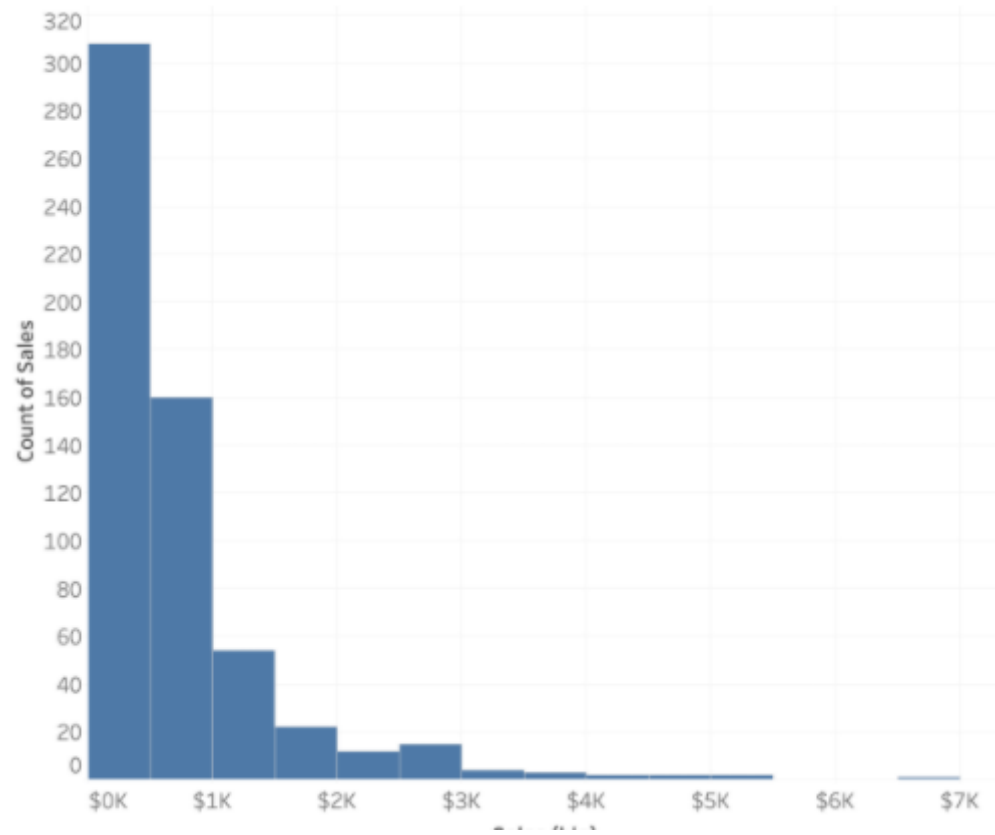


Binwidth = 4



# Modality





Worksheet: Sample - Superstore

Standard

Show Me

Data Analytics

Sample - Superstore

Search

Tables

Orders

Customer Name

Location

Order Date

Order ID

Product

Profit (bin)

Segment

Ship Date

Ship Mode

Top Customers by P...

Discount

Profit

Quantity

Sales

Orders (Count)

People

Person

People (Count)

Returns

Returned

Returns (Count)

Measure Names

Profit Ratio

Parameters

Profit Bin Size

Top Customers

Columns

Rows

Filters

Marks

Automatic

Color

Size

Text

Detail

Tooltip

Sheet 1

View Data: Sample - Superstore

9,994 rows

Show aliases

Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Postal Code	Product Name
Furniture	Henderson	United States	Claire Gite	Bush	11/8/2019	CA-2019-152156	42420	Bush Son...
Furniture	Henderson	United States	Claire Gite	Hon	11/8/2019	CA-2019-152156	42420	Hon Delu...
Office Supplies	Los Angeles	United States	Darrin Van Huff	Universal	6/12/2019	CA-2019-138688	90036	Self-Adhe...
Furniture	Fort Lauderdale	United States	Sean O'Donnell	Bretford	10/11/2018	US-2018-108966	33311	Bretford i...
Office Supplies	Fort Lauderdale	United States	Sean O'Donnell	Eldon	10/11/2018	US-2018-108966	33311	Eldon Fol...
Furniture	Los Angeles	United States	Brosina Hoffman	Eldon	6/9/2017	CA-2017-115812	90032	Eldon Exp...
Office Supplies	Los Angeles	United States	Brosina Hoffman	Newell	6/9/2017	CA-2017-115812	90032	Newell 32
Technology	Los Angeles	United States	Brosina Hoffman	Mitel	6/9/2017	CA-2017-115812	90032	Mitel 532i
Office Supplies	Los Angeles	United States	Brosina Hoffman	DXL	6/9/2017	CA-2017-115812	90032	DXL Angle
Office Supplies	Los Angeles	United States	Brosina Hoffman	Belkin	6/9/2017	CA-2017-115812	90032	Belkin FSX
Furniture	Los Angeles	United States	Brosina Hoffman	Chromcraft	6/9/2017	CA-2017-115812	90032	Chromcra
Technology	Los Angeles	United States	Brosina Hoffman	Other	6/9/2017	CA-2017-115812	90032	Konftel 2i
Office Supplies	Concord	United States	Andrew Allen	Xerox	4/15/2020	CA-2020-114412	28027	Xerox 19i
Office Supplies	Seattle	United States	Irene Maddox	Fellowes	12/5/2019	CA-2019-161389	98103	Fellowes i
Office Supplies	Fort Worth	United States	Harold Pawlan	Holmes	11/22/2018	US-2018-118983	76106	Holmes R...
Office Supplies	Fort Worth	United States	Harold Pawlan	Storex	11/22/2018	US-2018-118983	76106	Storex D...
Office Supplies	Madison	United States	Pete Kriz	Other	11/11/2017	CA-2017-105893	53711	Star-D-St
Office Supplies	West Jordan	United States	Alejandro Grove	Fellowes	5/13/2017	CA-2017-167164	84084	Fellowes :
Office Supplies	San Francisco	United States	Zuschuss Donatelli	Newell	8/27/2017	CA-2017-143336	94109	Newell 34
Technology	San Francisco	United States	Zuschuss Donatelli	Cisco	8/27/2017	CA-2017-143336	94109	Cisco SPA
Office Supplies	San Francisco	United States	Zuschuss Donatelli	Wilson Jones	8/27/2017	CA-2017-143336	94109	Wilson Jo

Orders People Returns

9,994 rows

Activate Windows  
Go to Settings to activate Windows.



Dashboard Story Analysis Map Format Window Help

Standard

Show Me

Data Analytics

Sample - Superstore

Search

Tables

- Order ID
- Product
- Profit (bin)
- Segment
- Ship Date
- Ship Mode
- Top Customers by P...
- Discount
- Profit
- Quantity
- Sales
- Orders (Count)

People

- Person
- People (Count)

Returns

- Returned
- Returns (Count)

Measure Names

- Profit Ratio
- Latitude (generated)
- Longitude (generated)
- Measure Values

Parameters

- Profit Bin Size
- Top Customers

Columns: Measure Names

Rows: Ship Mode

Filters: Measure Names: Co...

Marks: Automatic

Color Size Text

Detail Tooltip

Measure Values

CNT(Orders)

Sheet 1

Count of Orders

Sorted descending by count of Orders within Ship Mode

Ship Mode	Count of Orders
Standard Class	5,968
Second Class	1,945
First Class	1,538
Same Day	543

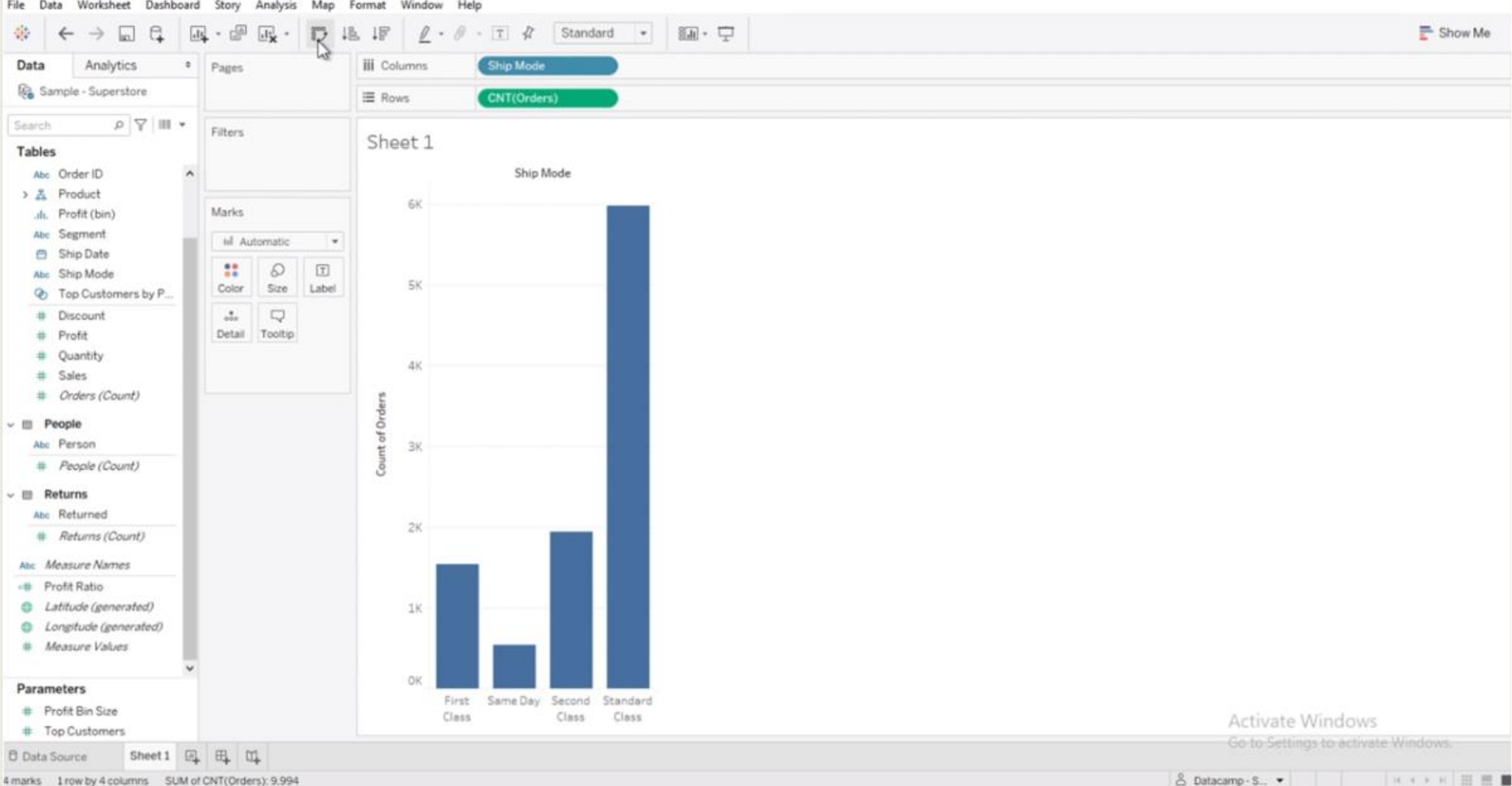
Activate Windows

Go to Settings to activate Windows.

Data Source Sheet 1

4 marks 4 rows by 1 column SUM of Measure Values: 9,994

Datcamp - S...







"Complete  
Exercise"



File Edit View Window Help

Standard

Show Me

Data Analytics

Sample - Superstore

Search

Tables

- Orders
  - Customer Name
  - Location
  - Order Date
  - Order ID
  - Product
  - Profit (bin)
  - Segment
  - Ship Date
  - Ship Mode
  - Top Customers by P...
  - Discount
  - Profit
  - Quantity
  - Sales
  - Orders (Count)
- People
  - Person
  - People (Count)
- Returns
  - Returned
  - Returns (Count)
- Measure Names
- Profit Ratio

Parameters

- Profit Bin Size
- Top Customers

Pages

Columns

Rows

Filters

Marks

Automatic

Color Size Text

Detail Tooltip

Sheet 1

Drop field here

Drop field here

Drop field here

field here

OK Cancel

Activate Windows

Go to Settings to activate Windows.

Standard

Data Analytics

Sample - Superstore

Search

Tables

- Orders
  - Customer Name
  - Location
  - Order Date
  - Order ID
  - Product
  - Profit (bin)
  - Quantity (bin)**
  - Segment
  - Ship Date
  - Ship Mode
  - Top Customers by P...
- People
  - Person
  - People (Count)
- Returns
  - Returned
  - Returns (Count)

Parameters

- Profit Bin Size
- Top Customers

Pages

Columns

Rows

Filters

Sheet 1

Marks

Automatic

Color Size Text

Detail Tooltip

Drop field here

Drop field here

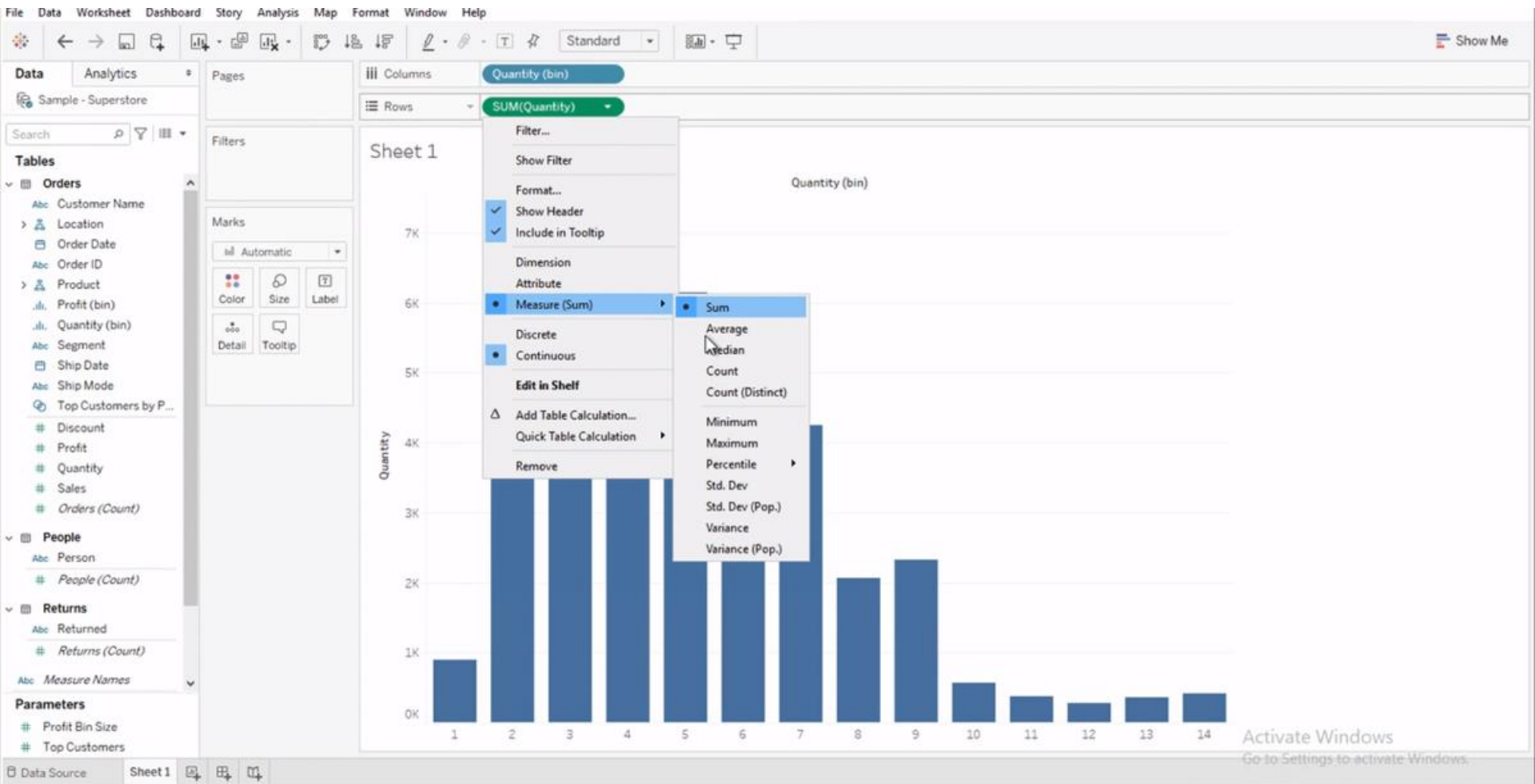
Drop field here

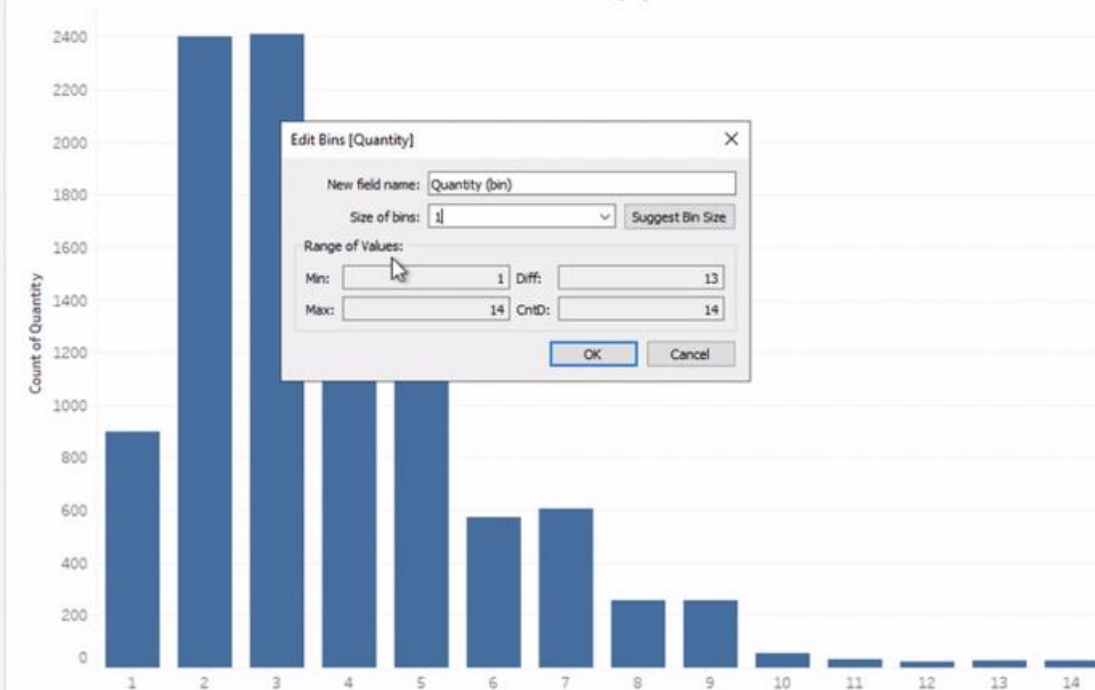
Activate Windows

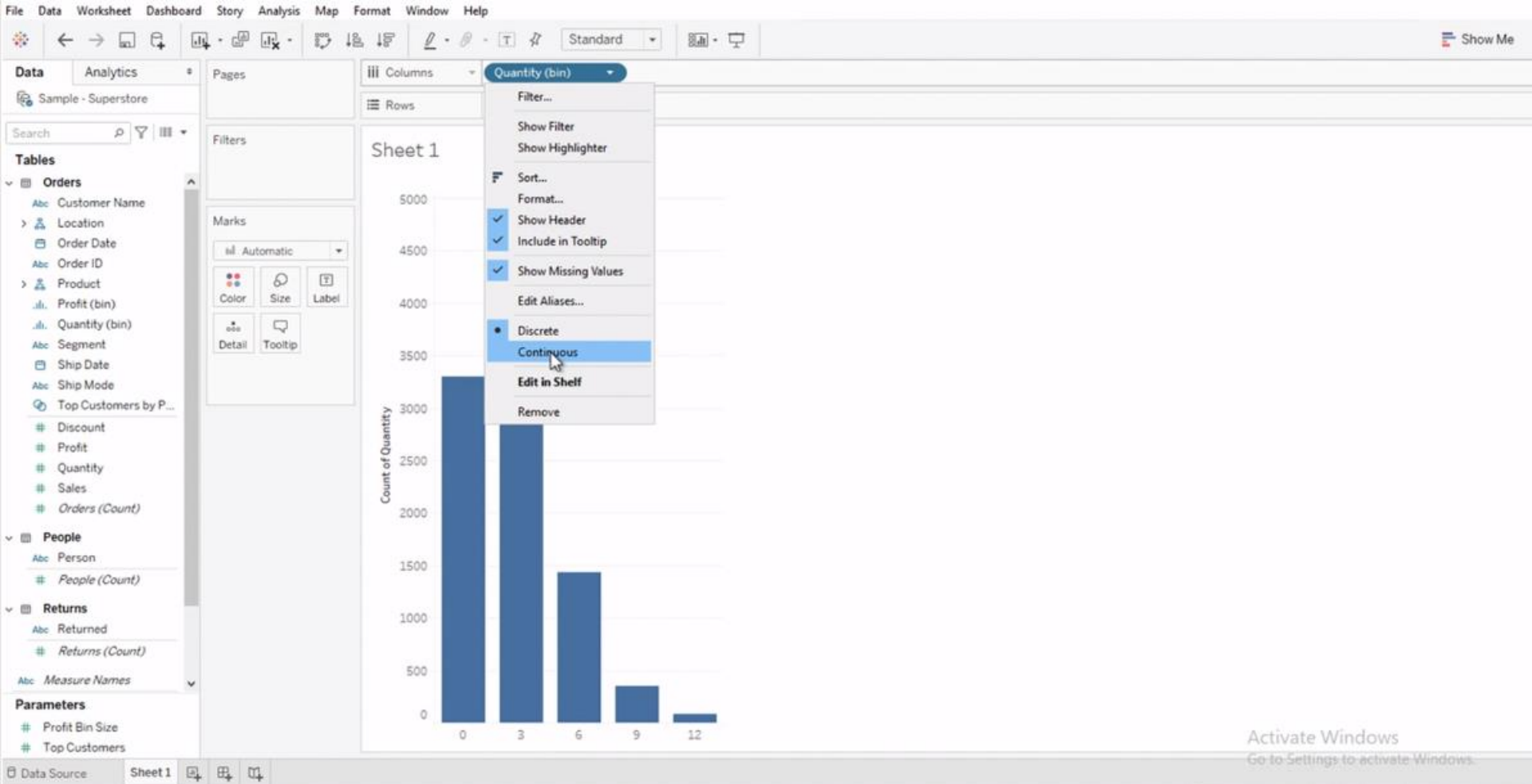
Go to Settings to activate Windows.

Data Source Sheet 1

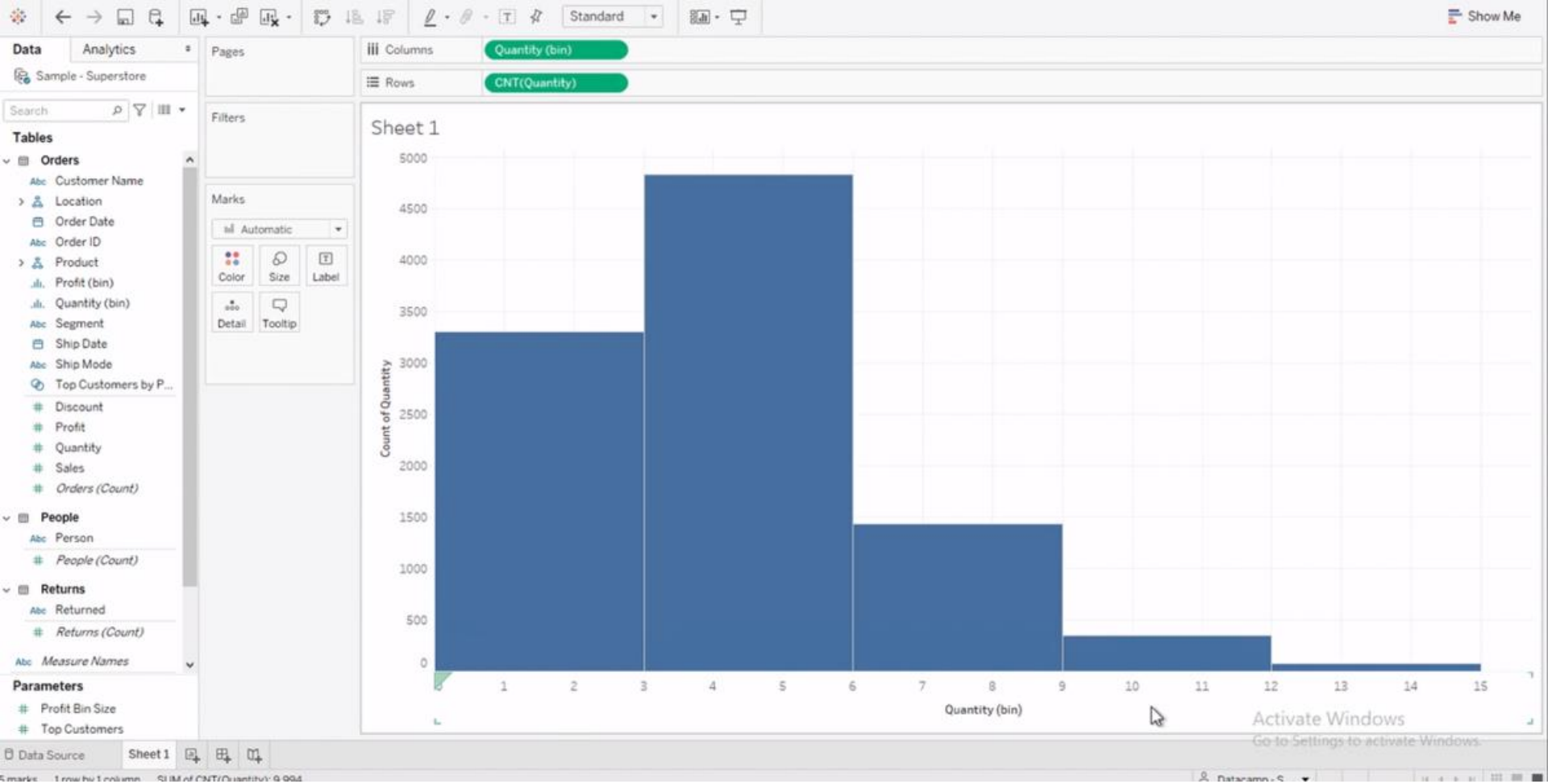
Datacamp - S...

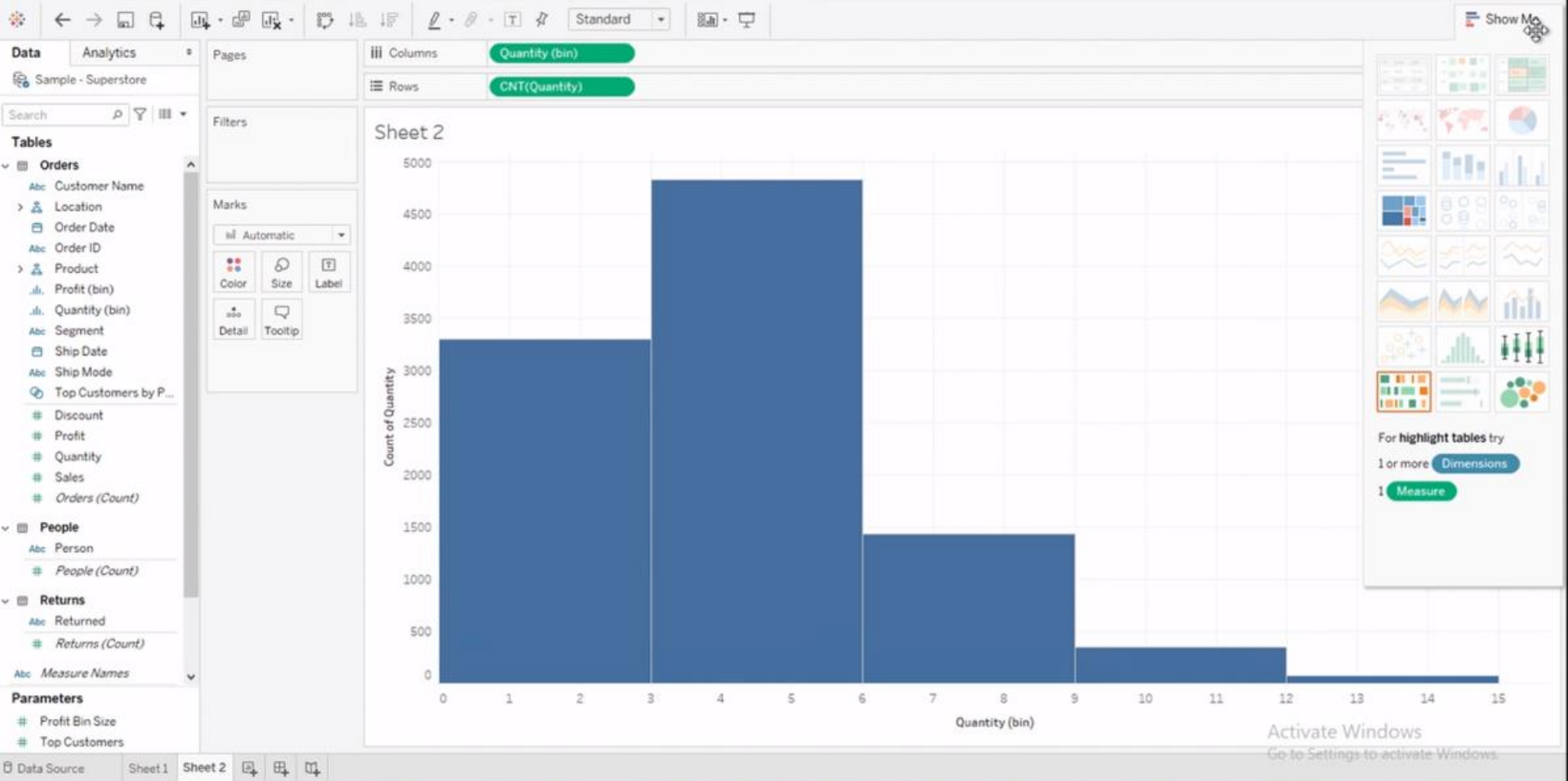














File Data Worksheet Dashboard Story Analysis Map Format Window Help Show Me

Data Analytics Pages

Columns Quantity (bin)

Rows CNT(Quantity)

Search

Tables

- Orders
  - Customer Name
  - Location
  - Order Date
  - Order ID
  - Product
  - Profit (bin)
  - Quantity (bin)
  - Segment
  - Ship Date
  - Ship Mode
  - Top Customers by P...
  - Discount
  - Profit
  - Quantity
  - Sales
  - Orders (Count)
- People
  - Person
  - People (Count)
- Returns
  - Returned
  - Returns (Count)
- Measure Names

Filters

Marks

Automatic

Color Size Label

Detail Tooltip

Sheet 2

Create Parameter

Name: Quantity (bin) Parameter Comment >>

Properties

Data type: Float

Current value: 2.95

Value when workbook opens: Current value

Display format: Automatic

Allowable values: ☐ All ☐ List ☒ Range

Range of values

☒ Minimum: 1

☒ Maximum: 5

☒ Step size: 1

☒ Fixed

Set values from

☐ When workbook opens

None

OK Cancel

Count of Quantity

Quantity (bin)

Activate Windows  
Go to Settings to activate Windows.

Data Source Sheet 1 Sheet 2

5 marks 1 row by 1 column SUM of CNT(Quantity): 9.994

1-33

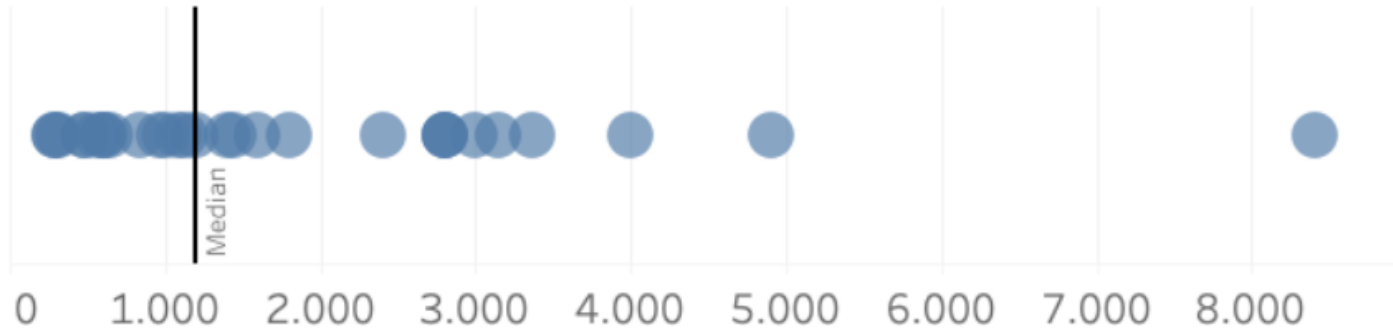




"Complete  
Exercise"

# Box plots

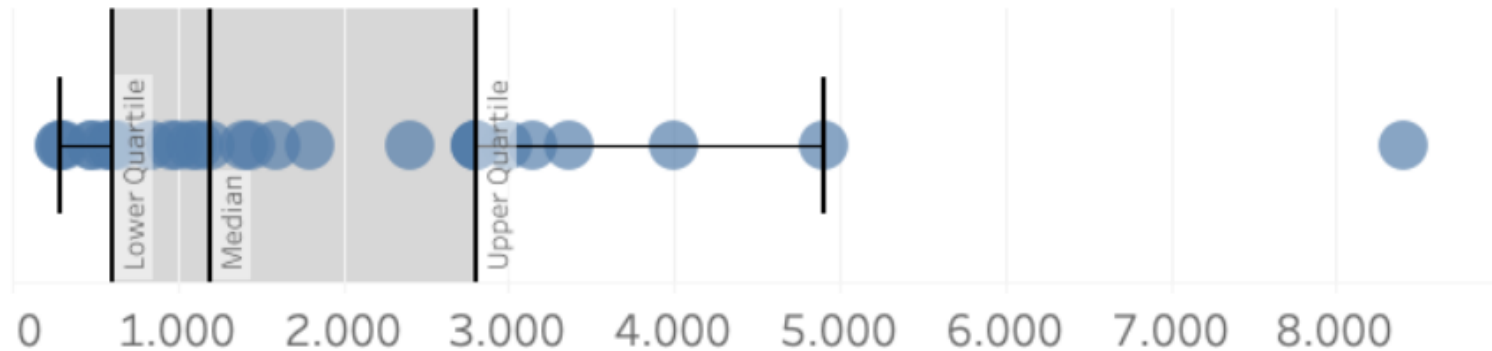
Visualize the distribution of a single, continuous variable





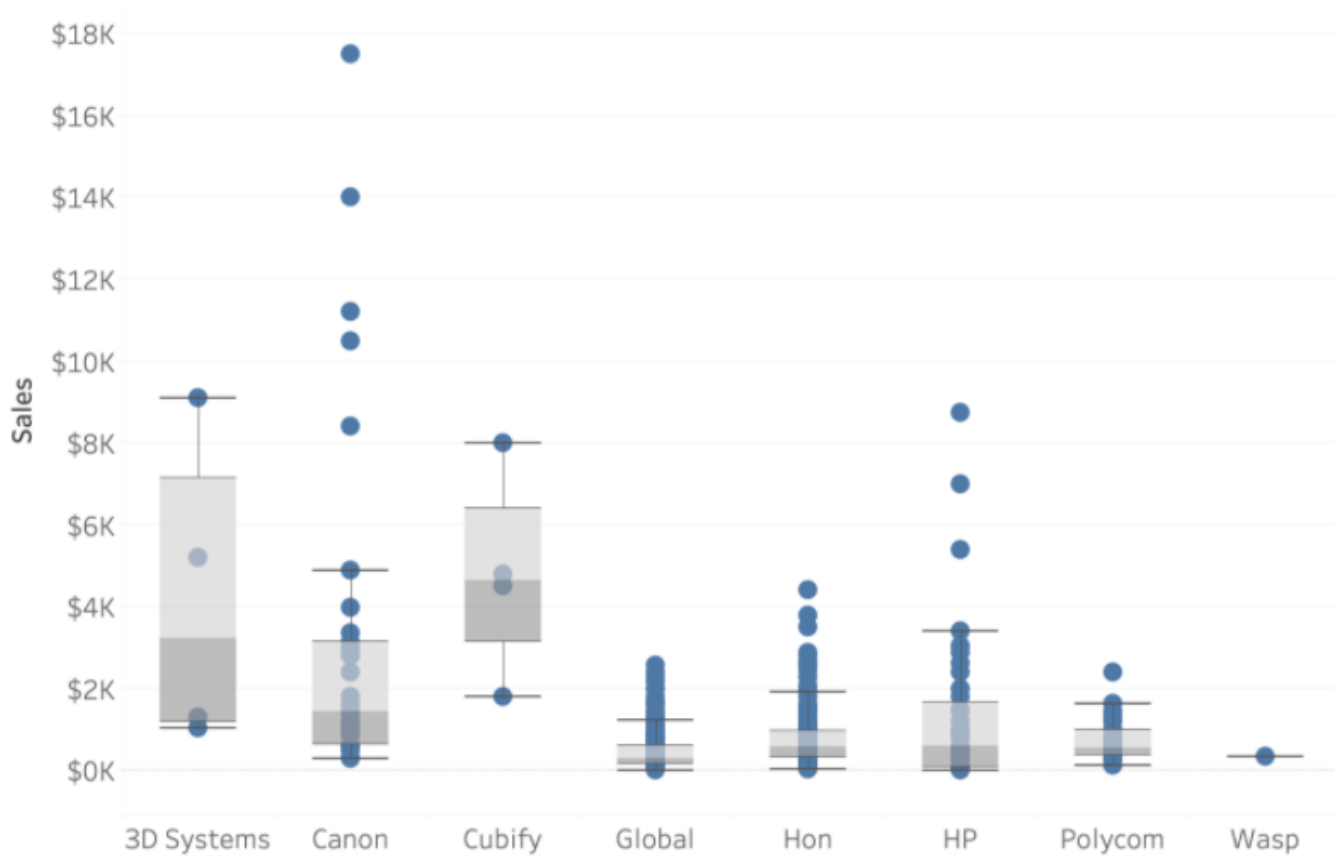
# Box plots

Visualize the distribution of a single, continuous variable



- Distance between lower quartile and upper quartile is the interquartile range (IQR)
- Whiskers: length of  $1.5 \times \text{IQR}$
- Outlier: extreme value outside whiskers

# When to use a box plot

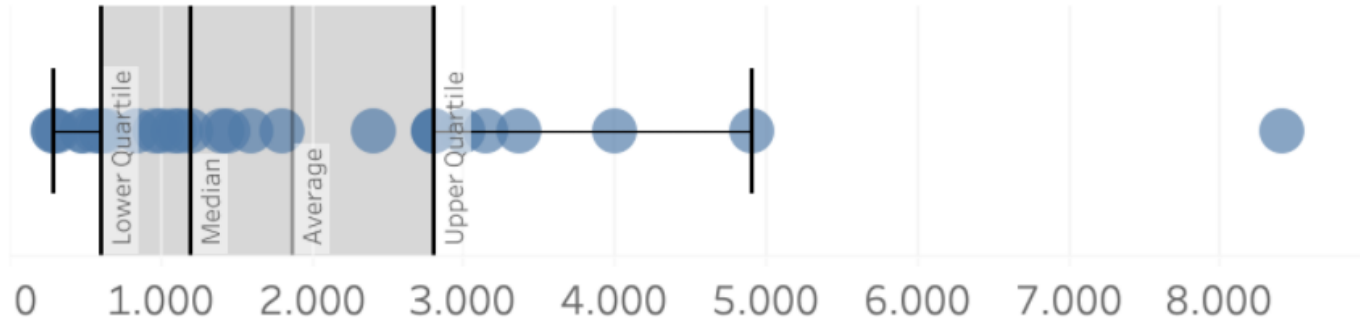


# When to use a box plot

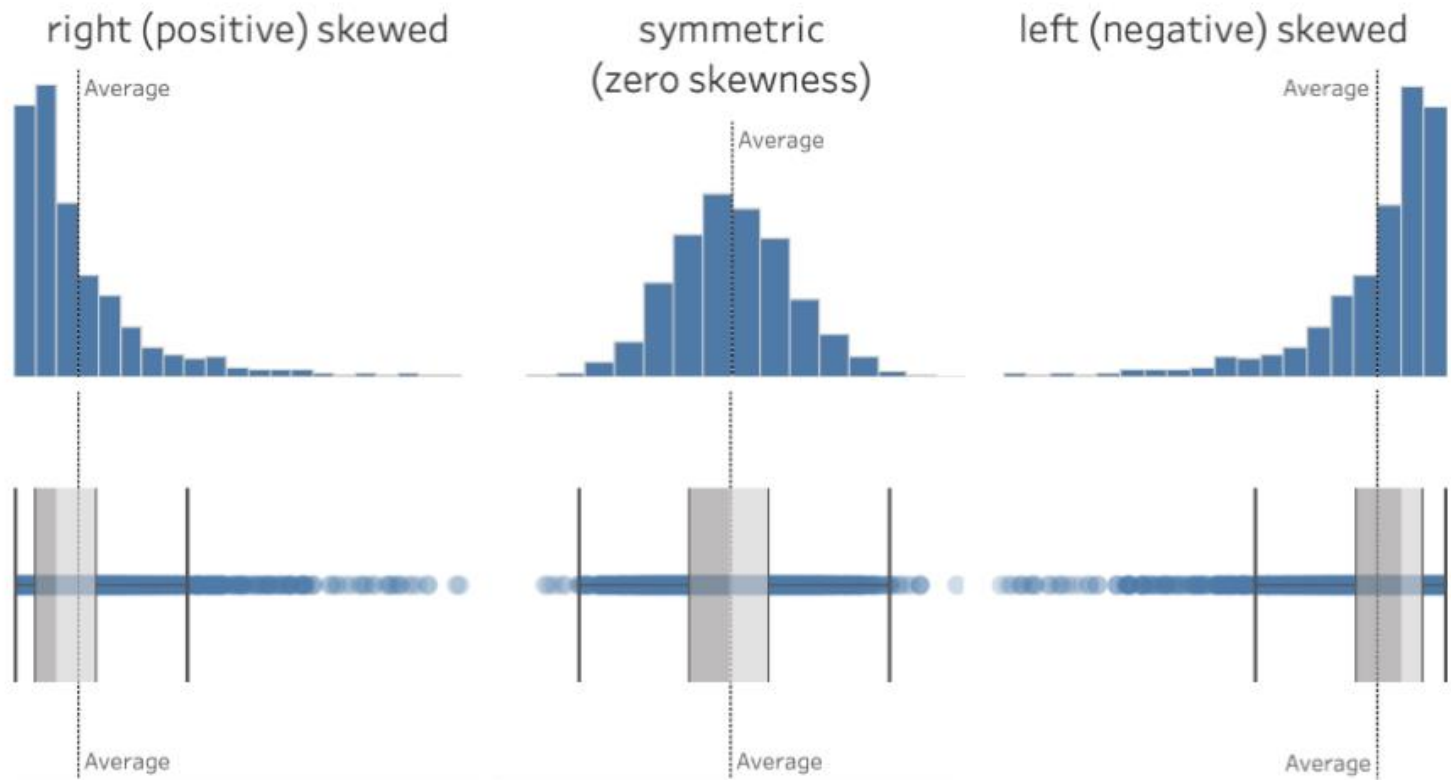
- Compare distributions among multiple categories
- Spot trends and differences between categories

# What about the mean?

- Average = arithmetic mean
- $\frac{a_1 + a_2 + a_3 + \dots + a_n}{n}$
- Average and mean are often used interchangeably

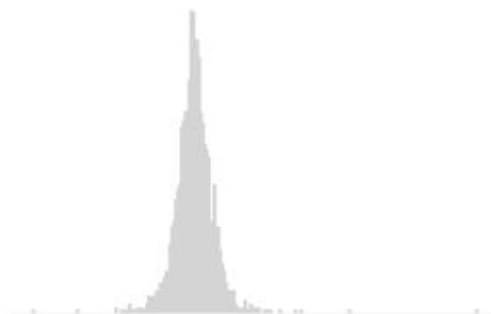


# Skewness

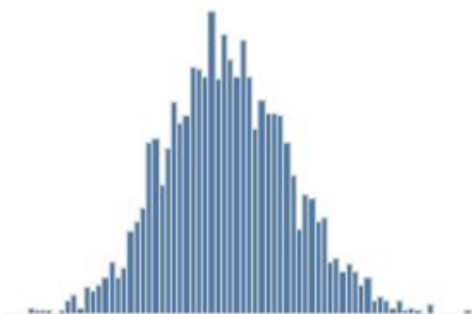


# Excess kurtosis

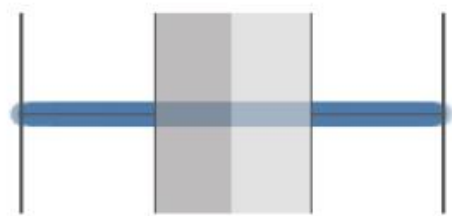
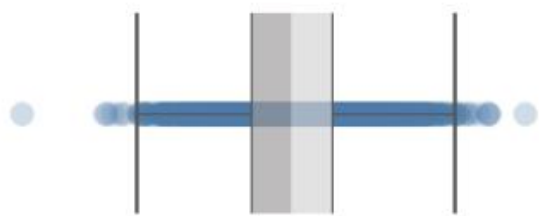
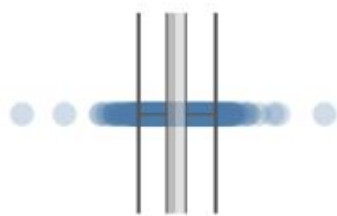
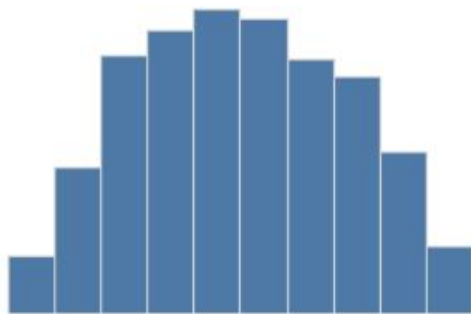
leptokurtic (positive)



mesokurtic



platykurtic (negative)

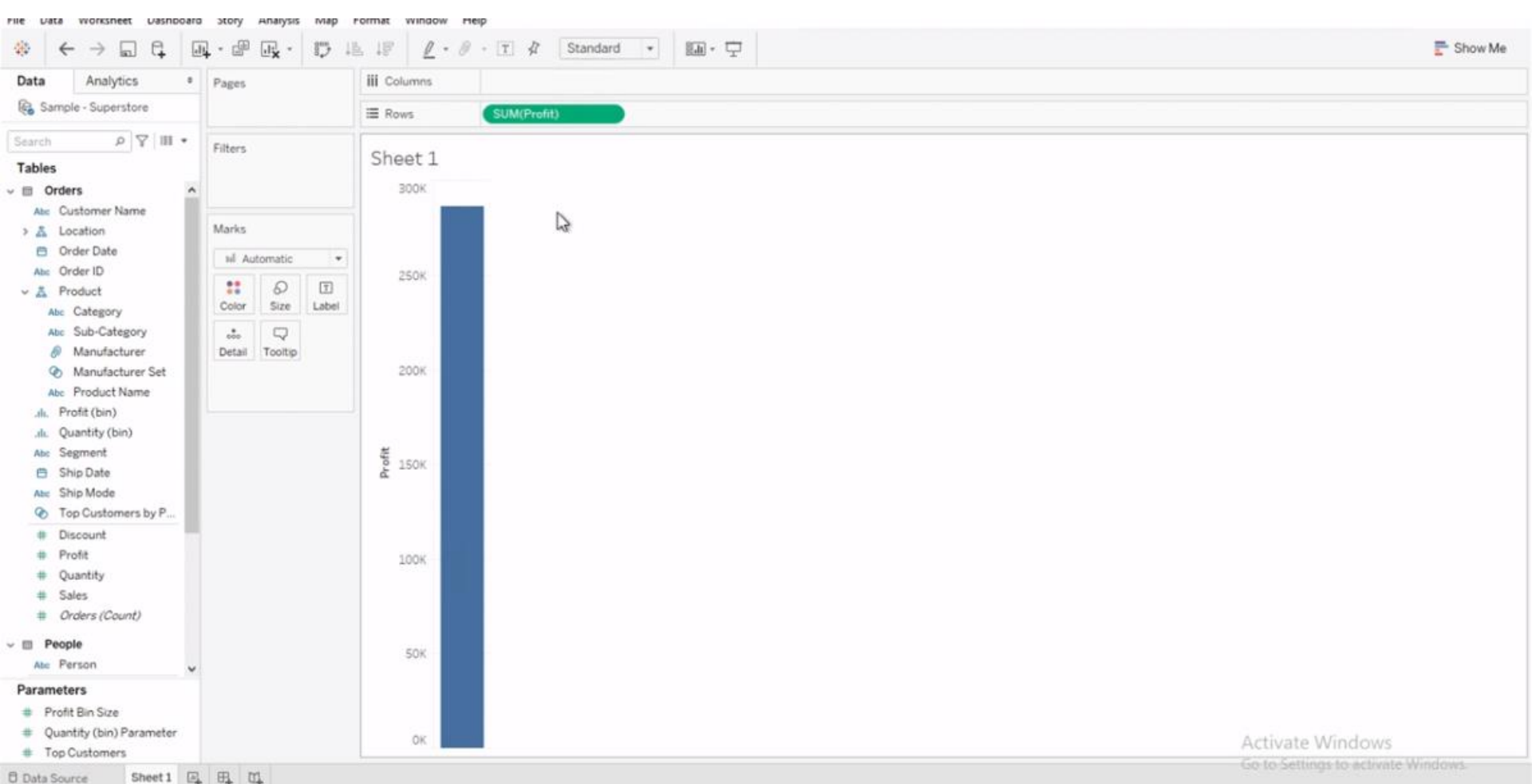


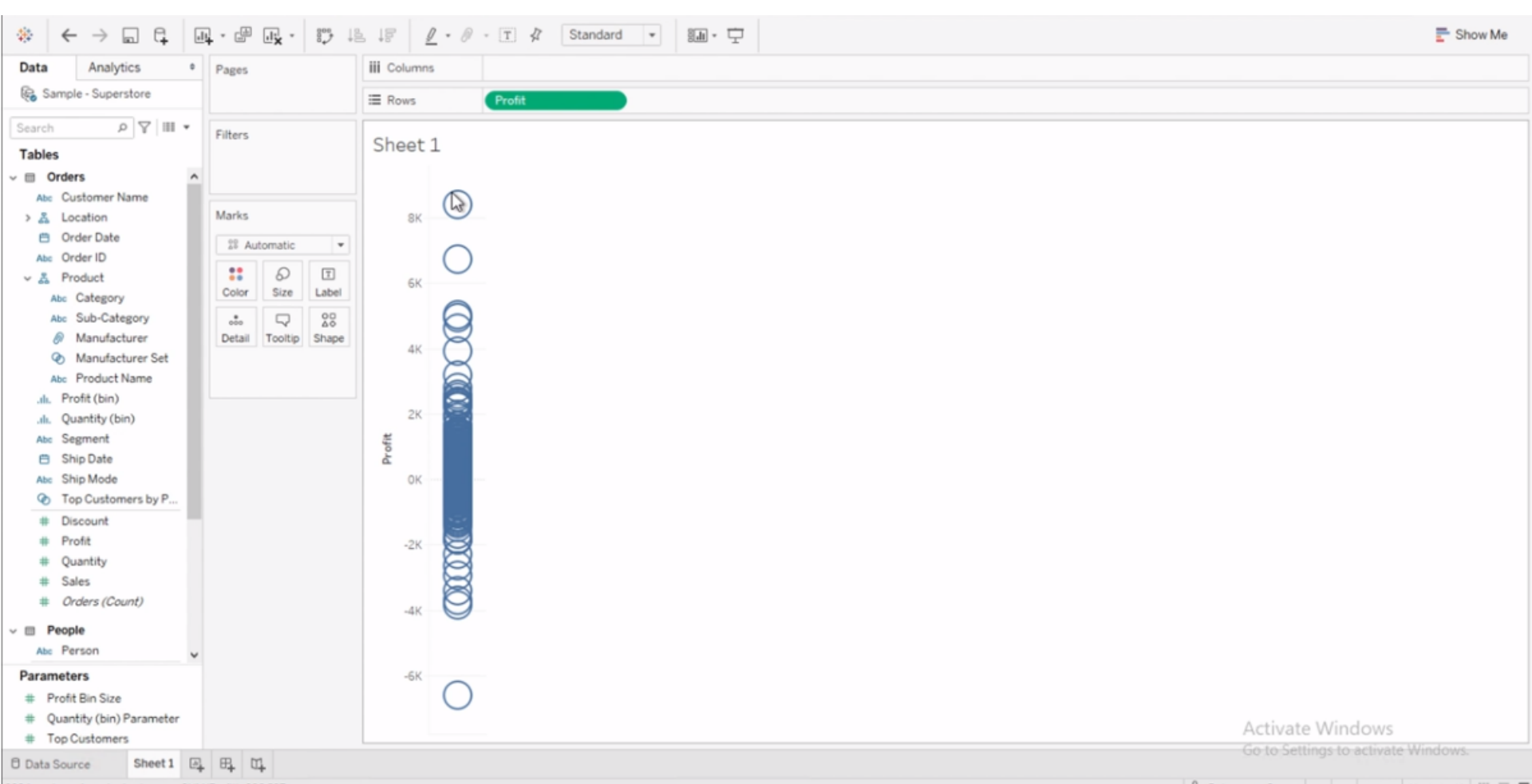


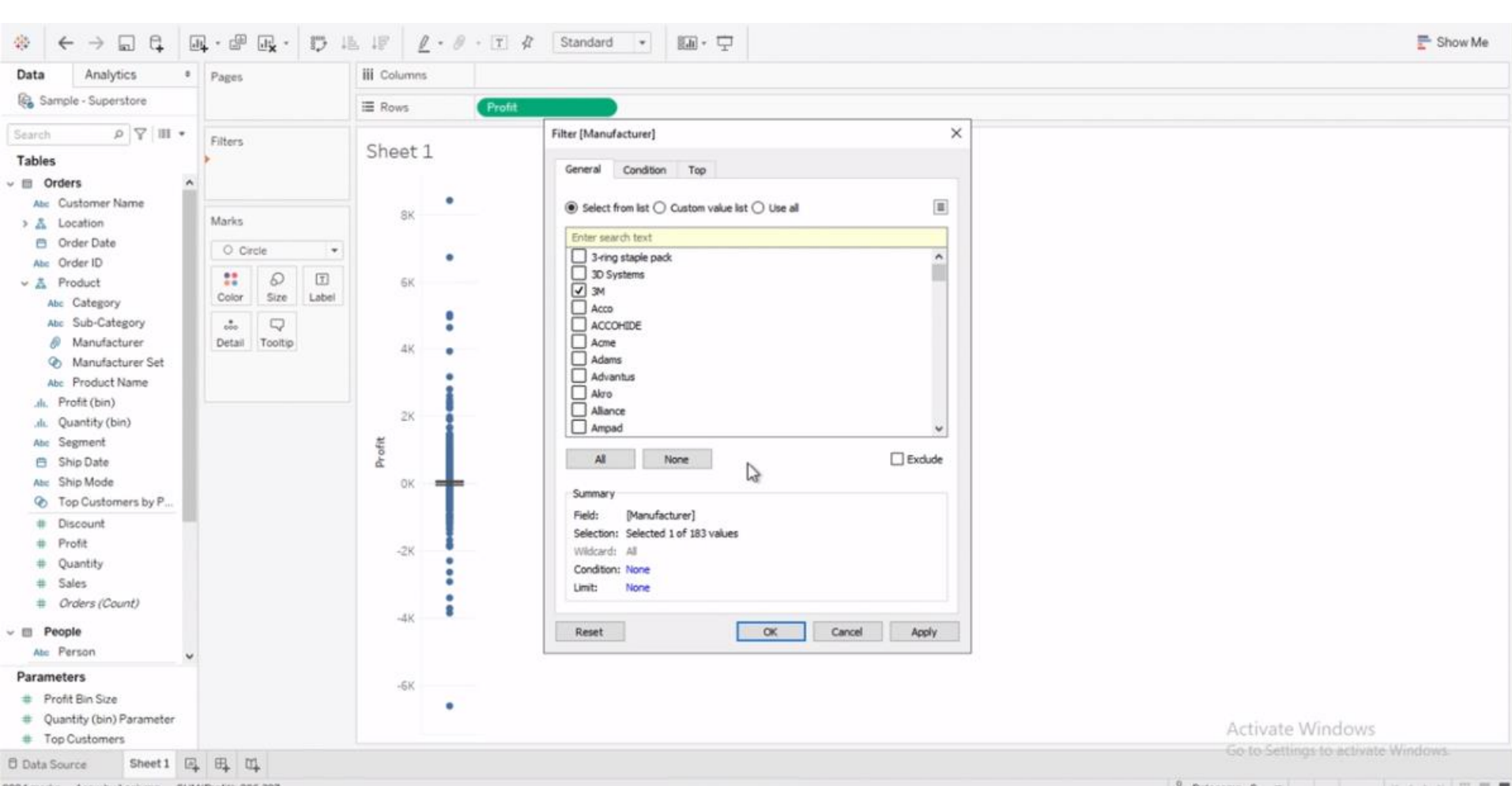
"Complete  
Exercise"

The screenshot shows the Tableau Desktop interface. On the left, the 'Data' pane displays the 'Sample - Superstore' data source. Under the 'Tables' section, the 'Orders' table is expanded, showing fields like Customer Name, Location, Order Date, Order ID, Product, Category, Sub-Category, Manufacturer, Manufacturer Set, Product Name, Profit (bin), Quantity (bin), Segment, Ship Date, Ship Mode, Top Customers by P..., Discount, Profit, Quantity, Sales, and Orders (Count). The 'Profit' field is currently selected and highlighted in green. Below the tables, the 'Parameters' section lists 'Profit Bin Size', 'Quantity (bin) Parameter', and 'Top Customers'. The main workspace, labeled 'Sheet 1', is empty and contains three 'Drop field here' prompts. The top toolbar includes various icons for navigation and formatting, and a 'Standard' dropdown menu. The bottom right corner features a watermark for 'Activate Windows' and a link to 'Go to Settings to activate Windows.'

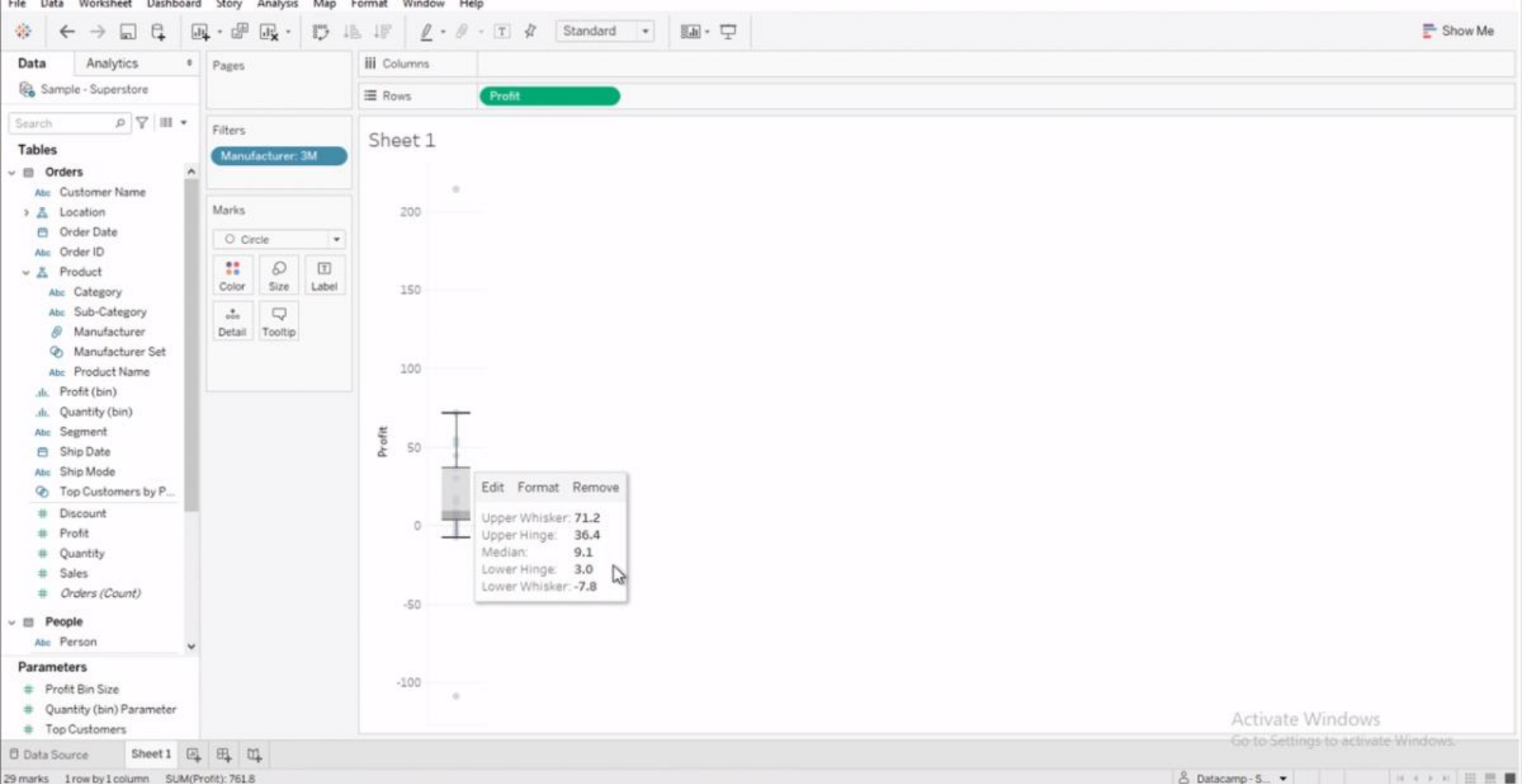


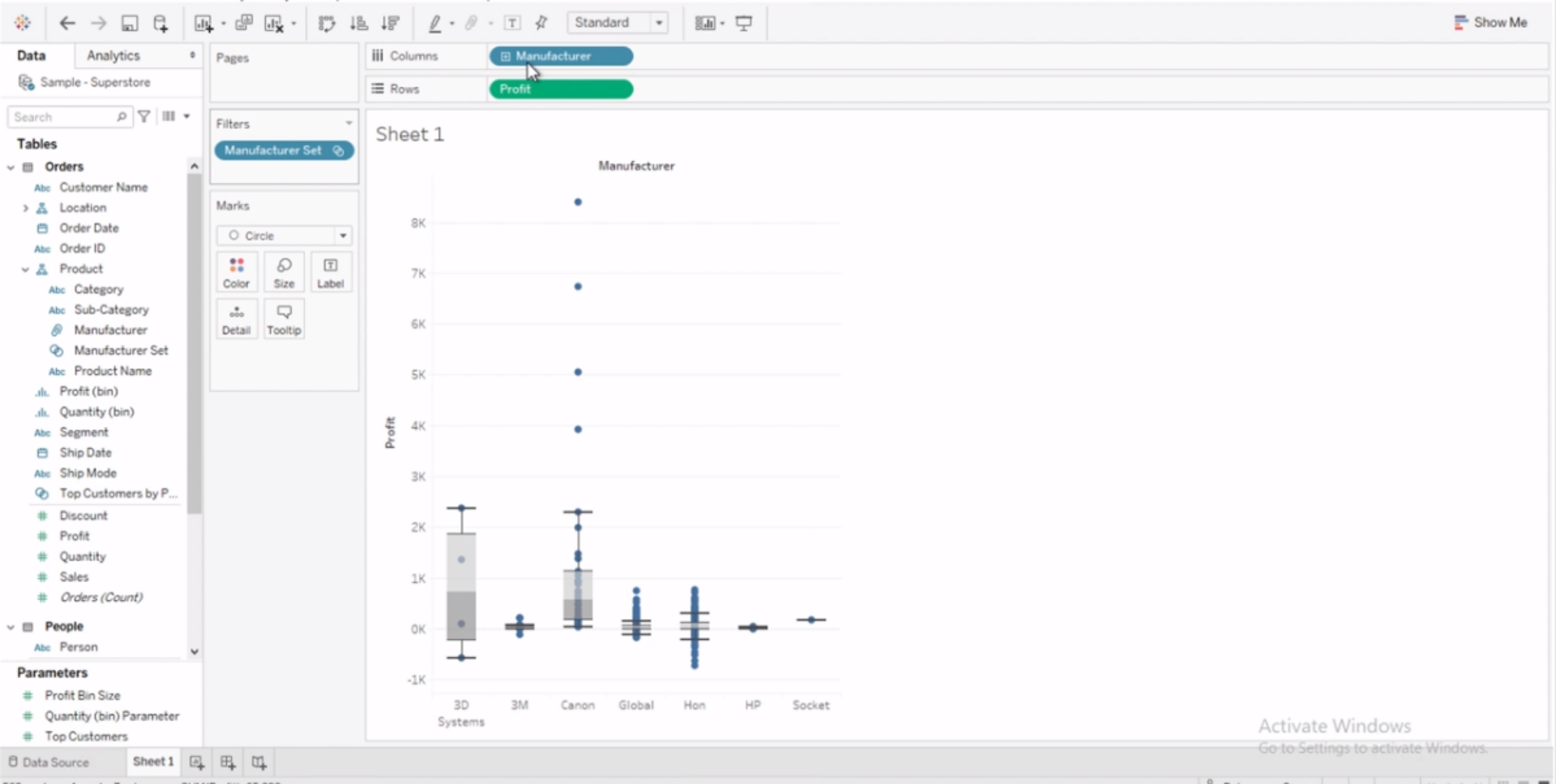


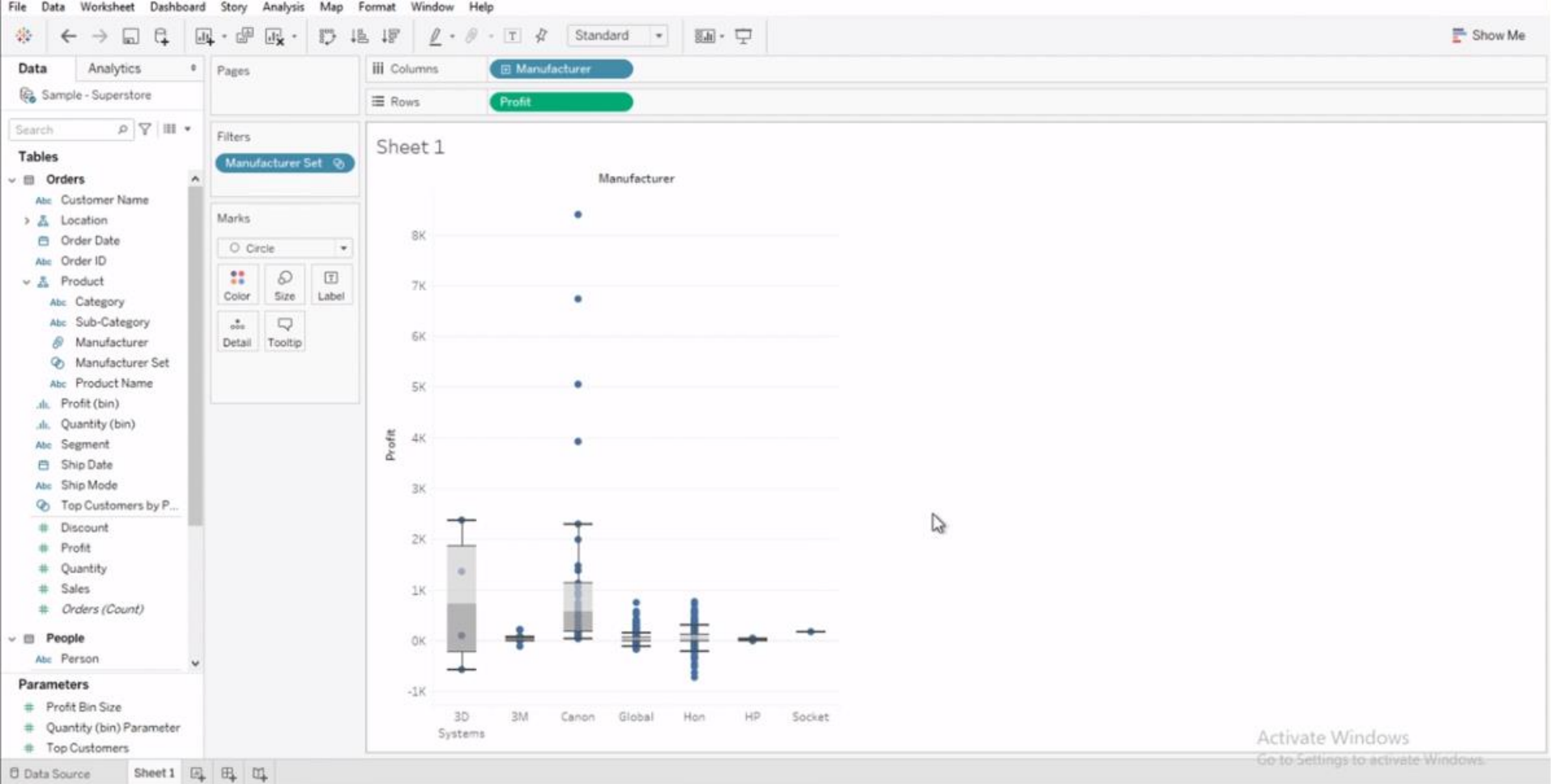


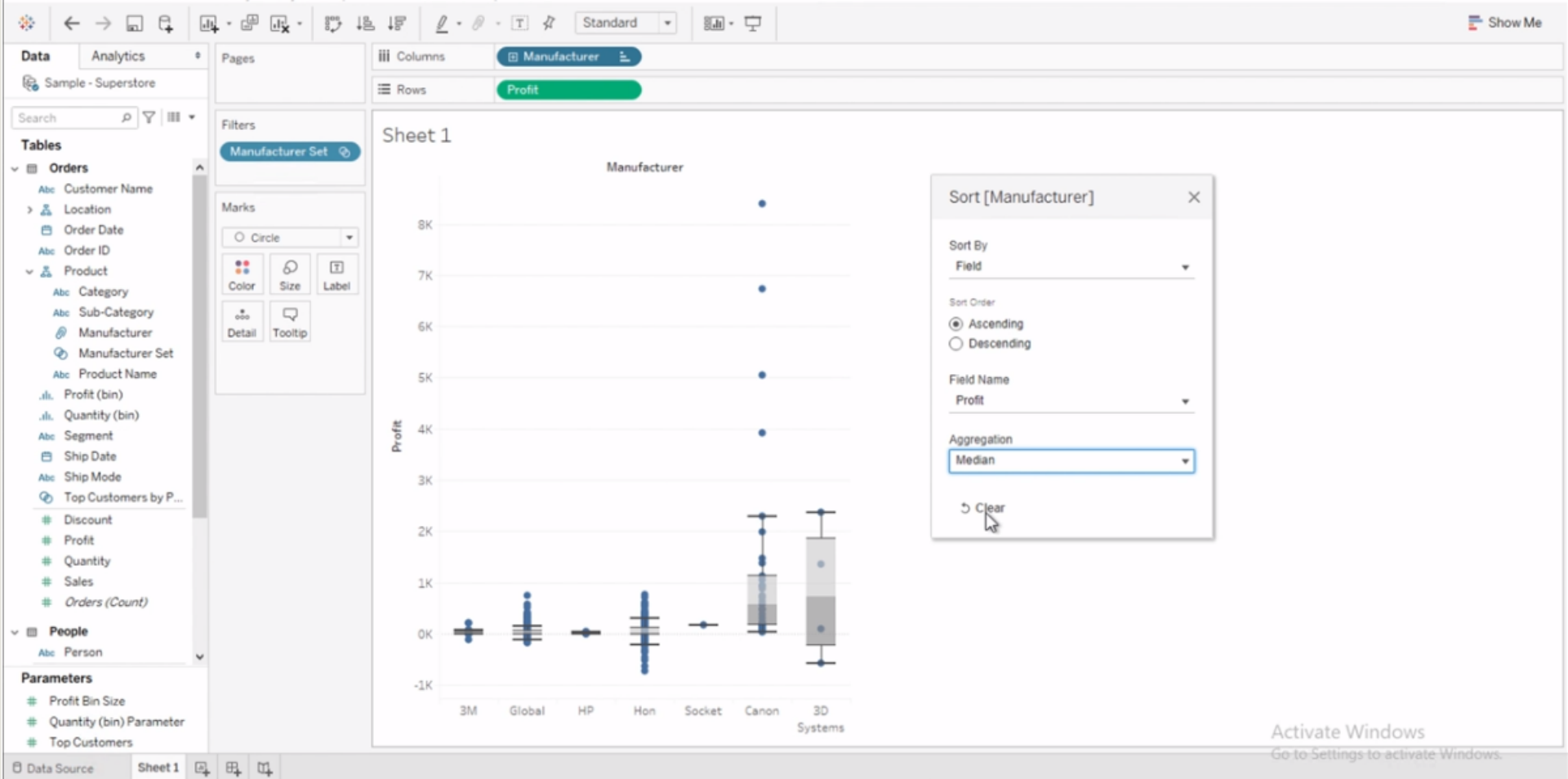


Activate Windows  
Go to Settings to activate Windows.











# Complete Lab"