

LAB 4: VISUALISING DISTRIBUTION



Table of Contents

| | |
|---|----|
| Learning Outcome..... | 2 |
| Data Preparation for Task 1 & 2 | 2 |
| Task 1: Using a Summary Card..... | 3 |
| Task 2: Binning Measures | 7 |
| Data Preparation for Task 3 | 11 |
| Task 3: Build a Box Plot | 12 |
| Data Preparation for Task 4..... | 19 |
| Task 4: Creating a Population Pyramid | 19 |

Learning Outcome

At the end of this session, learners will be able to:

- Use the summary card in Tableau
- Apply binning measures
- Build a Box Plot
- Create a Population Pyramid

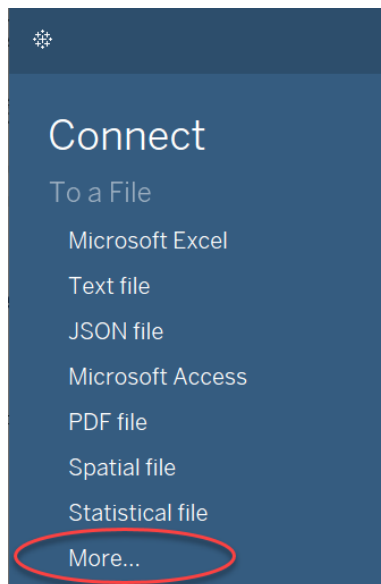
Data Preparation for Task 1 & 2

Create a Tableau workbook that connect to the **Sample - Superstore Subset (Excel) Tableau data source**.

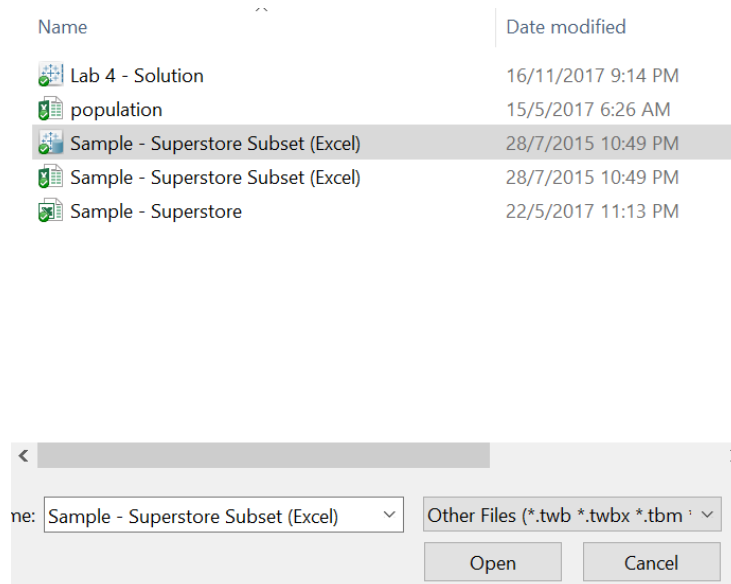
Step 1. Launch Tableau. Under Connect To a File, select **More...**

Step 2. From the file open window, select the **Sample - Superstore Subset (Excel) Tableau Datasource** file.

Step 1



Step 2



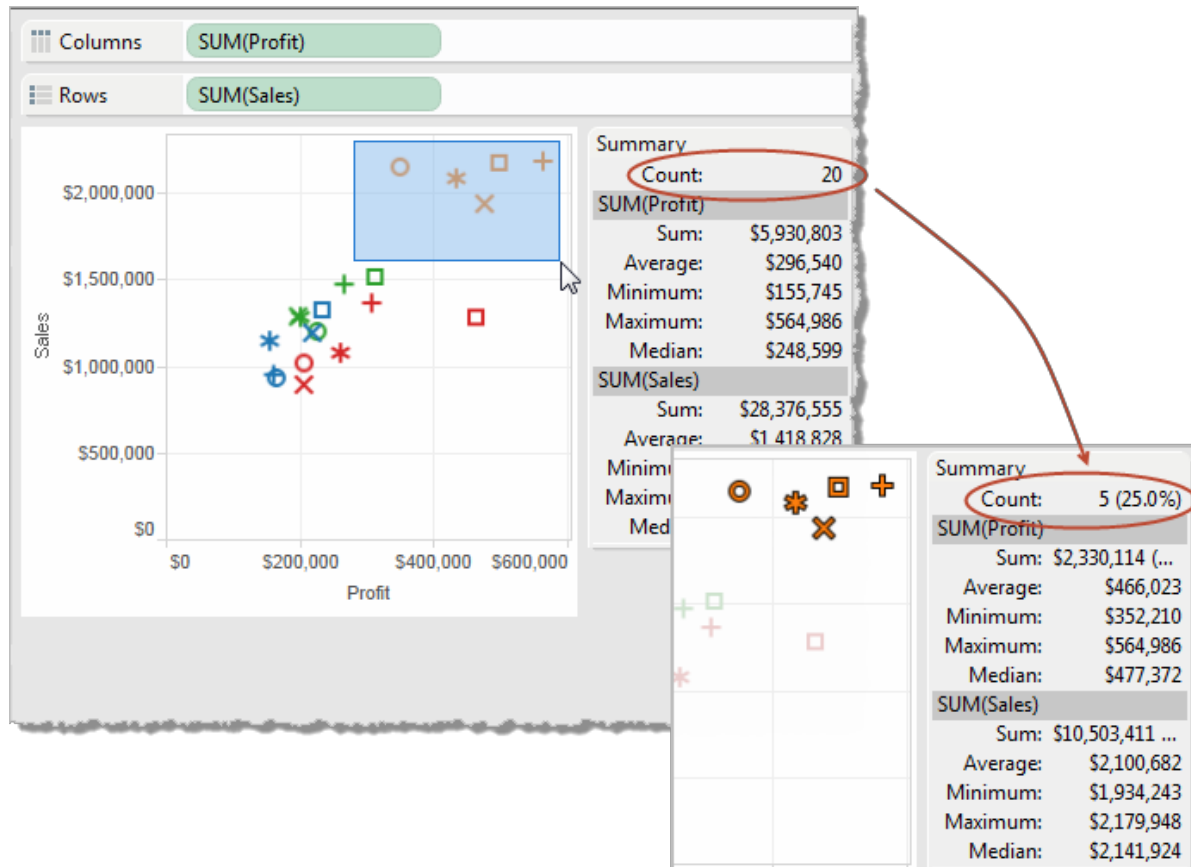
Task 1: Using a Summary Card

The summary card is a really quick way to view information about a selection or the entire data source. You can hide or show the Summary Card by selecting it on the View Cards toolbar menu



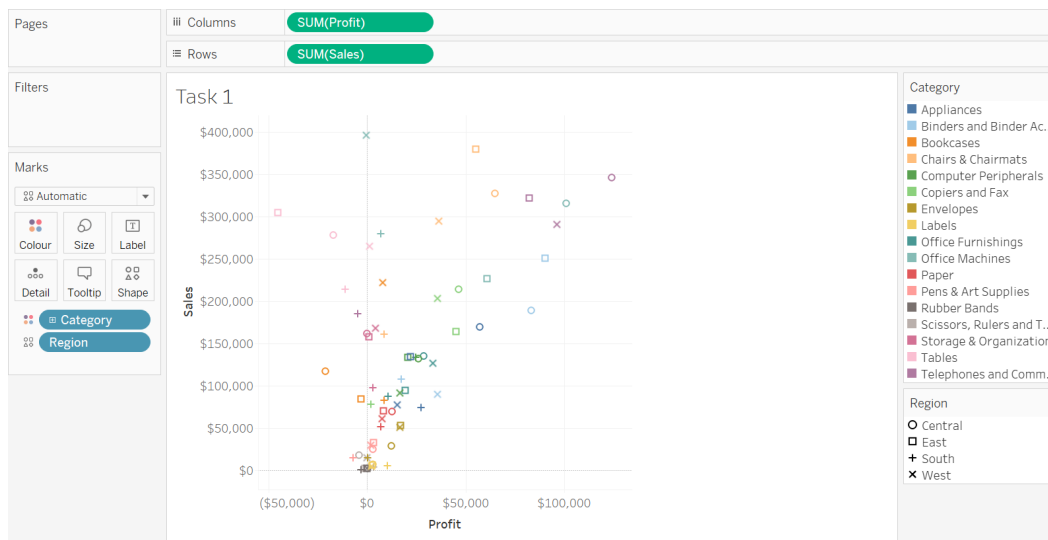
. You can also select **Worksheet** → **Show Summary**.

When you select data in the view, the Summary Card updates to show you information only for the data within the selection:

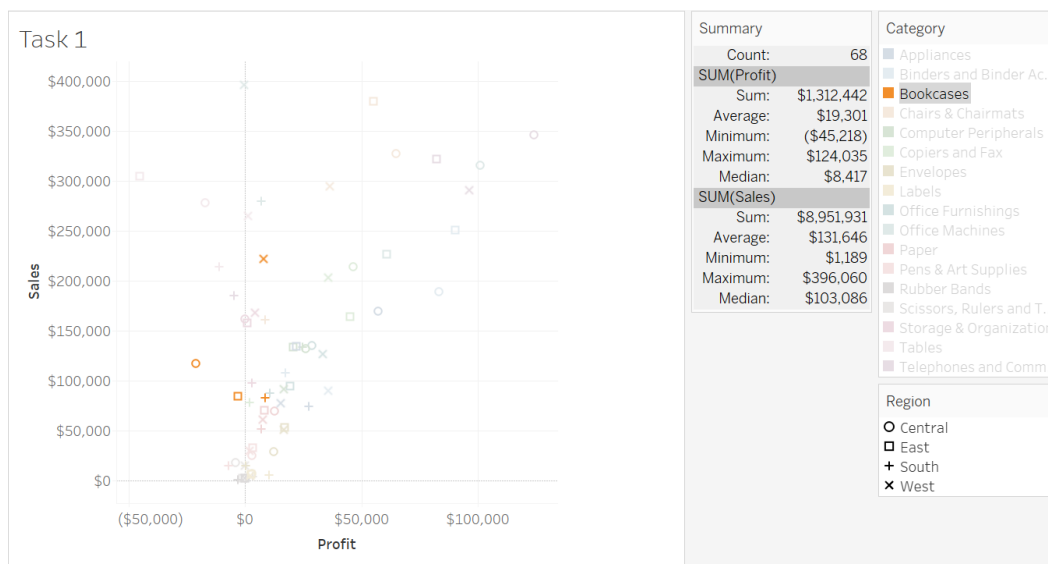


1. Open a new sheet and rename it as **Task 1**.
2. Make sure **Sample - Superstore Subset (Excel)** data source is selected.
3. Drag **Profit** to the **Columns** shelf.
4. Drag **Sales** to the **Rows** shelf.
5. Drag **Category** to the **Marks** card and drop it on the **Colour** shelf.

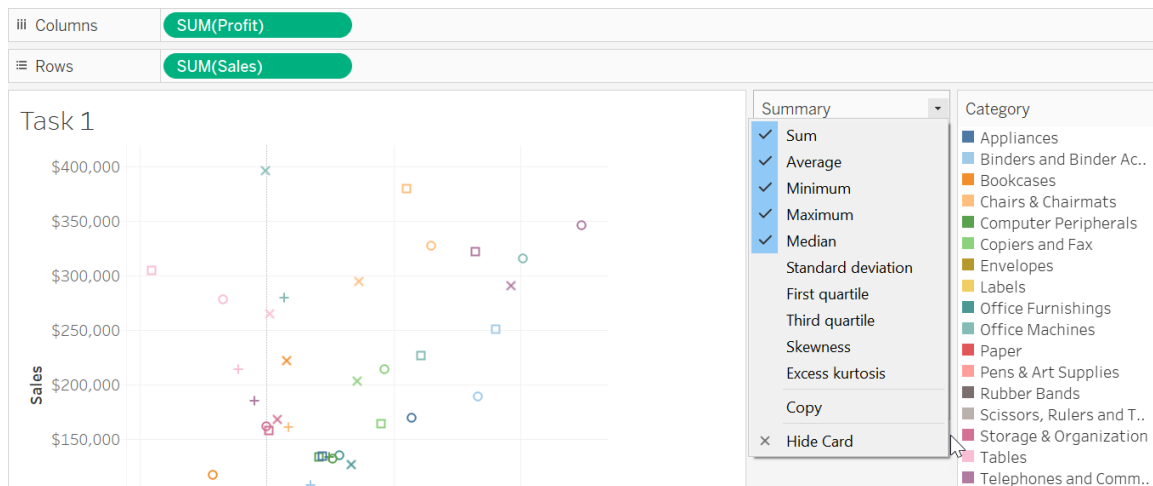
6. Drag **Region** to the **Marks** card, and drop it on the **Shape** shelf.



7. Select **Worksheet** menu → **Show Summary**. Drag and place the Summary Card on the left of the Category legend.
8. Select a Category and Summary Card will be updated to show the relevant information.



9. Click on the down arrow key on the Summary Card and more options are available for further analysis.



By default, the Summary Card shows Sum, Average, Minimum, Maximum, and Median values for the data. (Average is computed by summing all relevant values and then dividing by the total number of values. Median is computed by sorting values from lowest to highest and then selecting the middle value.)

You can use the drop-down menu for the Summary Card to show additional statistics:

- **Standard Deviation**

A measure of data spread around its average, measured in the same units as the data itself. The sample standard deviation is an unbiased estimate of the population standard deviation given a slight correction. This standard deviation includes the correction.

- **First Quartile**

A measure of location which is commonly used with other quartiles to provide a robust measure of spread. Robust in this case means not as sensitive to outliers as the standard deviation. The first quartile is the 25th percentile, typically the lower line in a boxplot

- **Third Quartile**

A measure of location which is commonly used with other quartiles to provide a robust measure of spread. Robust in this case means not as sensitive to outliers as the standard deviation. The third quartile is the 75th percentile, typically the upper line in a boxplot.

- **Skewness**

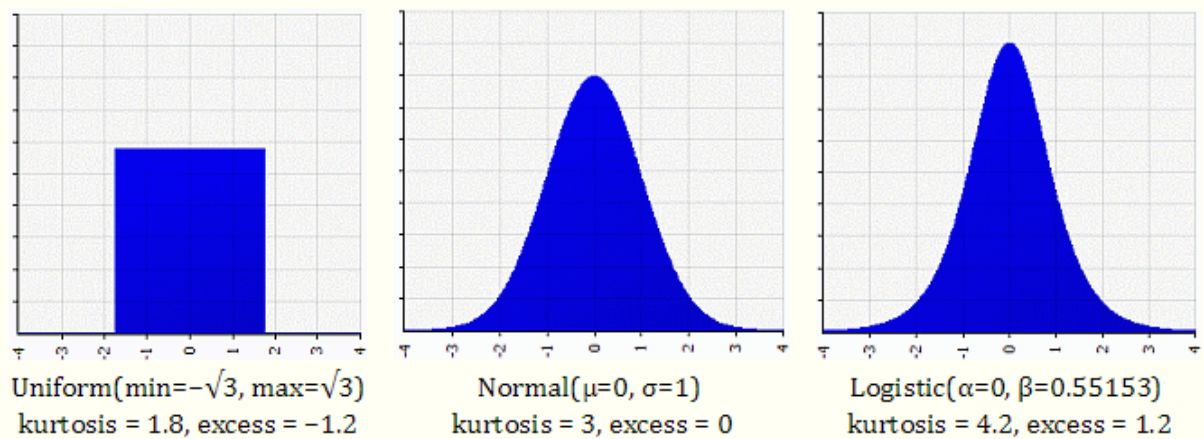
A measure of the tendency of your data to have extreme values to one side. Positive skewness means the extreme values are to the right, while negative skewness means the extreme values are to the left.

- **Excess Kurtosis**

A measure of the tendency of your data to have more extreme or outlying values than a normal distribution. A normal distribution has a kurtosis of 3 so this value is kurtosis minus three.

The reference standard is a normal distribution, which has a kurtosis of 3. In token of this, often the **excess kurtosis** is presented: excess kurtosis is simply **kurtosis-3**.

- A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≈ 3 (excess ≈ 0) is called **mesokurtic**.
- A distribution with kurtosis < 3 (excess kurtosis < 0) is called **platykurtic**. Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader.
- A distribution with kurtosis > 3 (excess kurtosis > 0) is called **leptokurtic**. Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper.



Adapted from http://onlinehelp.tableau.com/current/pro/desktop/en-us/inspectdata_summary.html

© 2003-2013 Tableau Software. All Rights Reserved

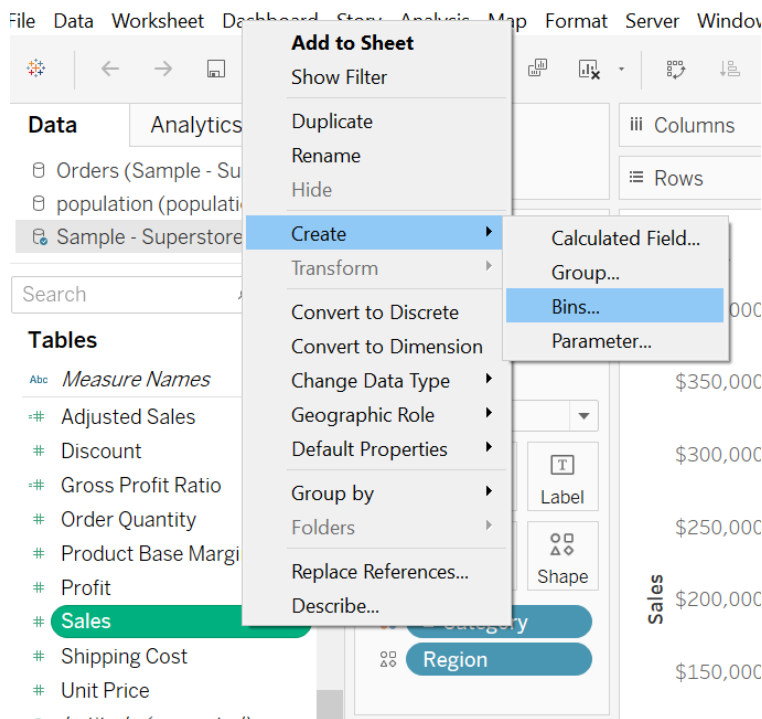
Task 2: Binning Measures

Measures are fields that typically contain numeric information, such as sales or budget figures. When you place a measure on a shelf in Tableau Desktop, it creates an axis.

However, sometimes you might want to organize the values of a measure into discrete groups. For example, suppose you have a measure that holds the ages of customers ranging from 18 to 90. Rather than break a view down by every age, you may want to analyze based on age groups (18 to 25, 26 to 33, and so on.). You can create these ranges by binning the data.

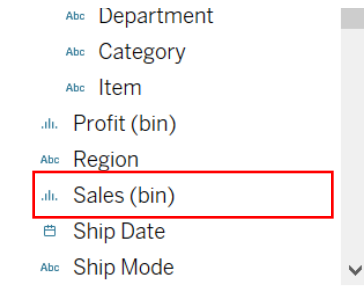
Say you are analyzing the sales performance for a retail store. One way to look at sales is in the form of a histogram, so you can see sales less than \$500, between \$500 and \$1000, and so on. To build a histogram, first you would bin the sales total values into categories.

1. Open a new sheet and rename it as **Task 2**.
2. Make sure **Sample - Superstore Subset (Excel)** data source is selected.
3. Right click **Sales** and select **Create Bins**.



- In the Create Bins dialog box, specify the size of the bins. In this example, type **500** in the **Size of bins** text box.

- Click **OK**. Tableau creates a new dimension called **Sales (bin)**.



- Drag **Sales (bin)** to the **Rows** shelf.

Notice that sales values are broken into \$500 bins. Each bin label designates the lower limit (inclusive) of the bin's range. For example, the bin labeled \$1,000 contains numbers equal to or greater than \$1,000, but less than \$1,500.

| | |
|---------|-------------|
| Columns | |
| Rows | Sales (bin) |

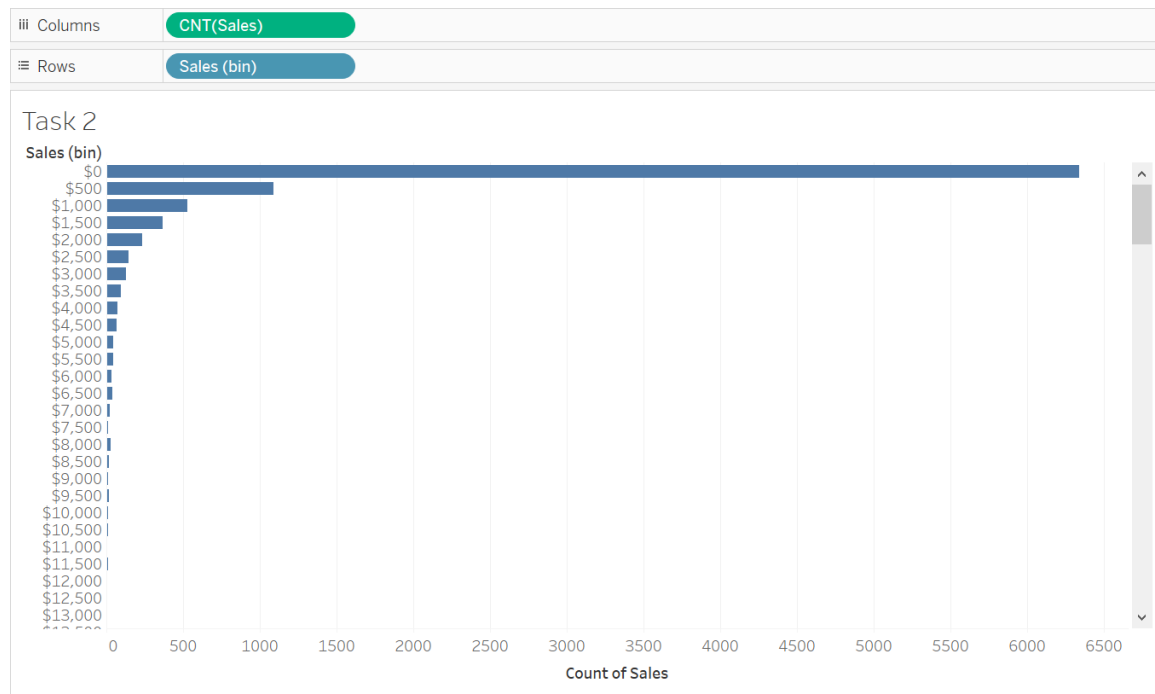
Task 2

| Sales (bin) | |
|-------------|-----|
| \$0 | Abc |
| \$500 | Abc |
| \$1,000 | Abc |
| \$1,500 | Abc |
| \$2,000 | Abc |
| \$2,500 | Abc |
| \$3,000 | Abc |
| \$3,500 | Abc |
| \$4,000 | Abc |
| \$4,500 | Abc |
| \$5,000 | Abc |
| \$5,500 | Abc |
| \$6,000 | Abc |
| \$6,500 | Abc |
| \$7,000 | Abc |
| \$7,500 | Abc |
| \$8,000 | Abc |

7. Drag **Sales** to the **Columns** shelf.

Right click **Sales** on the **Columns** shelf and select **Measure (Sum) → Count**.

In the final view, each bar represents the number of transactions with sales amounts within each bin. You can now see that most sales at this superstore are for less than \$500.

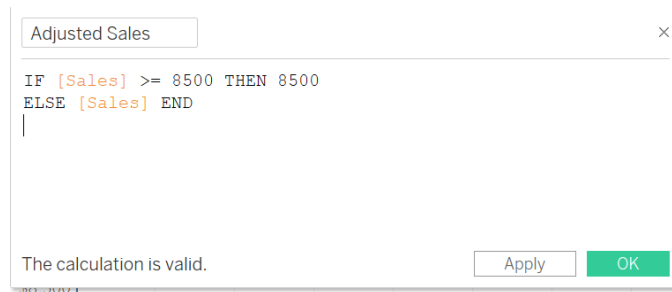


Create a calculated field to further customize bins

After you examine the result of the sales bin exercise in the previous section, you might determine that the values above \$8,500 are outliers and should be grouped together. To group them, you can create a calculated field, and then create a bin from the calculation.

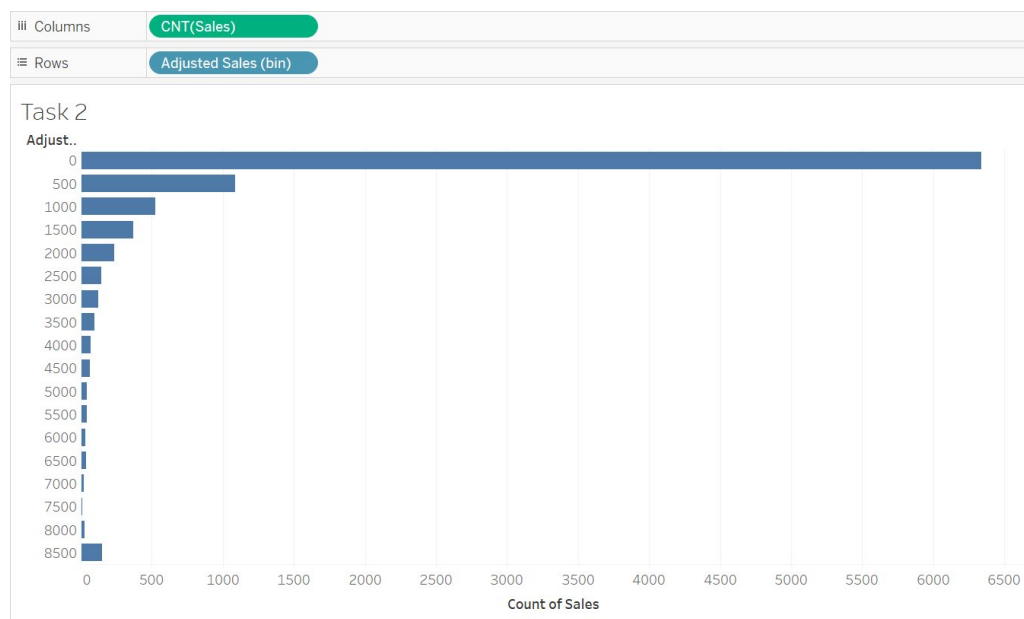
1. Right click the measure you are interested in (in this case, **Sales**) and select **Create Calculated Field**.
2. In the Calculated Field dialog box, complete the following steps.
 - a. Specify the name. This example uses **Adjusted Sales**.
 - b. In the **Formula** text box, build a formula to round the outliers to the value you want to use. This example uses the following formula:

```
IF [Sales] >= 8500 THEN 8500
ELSE [Sales] END
```



- c. Confirm that the status message indicates that the formula is valid, and then click **OK**.
3. Right click **Adjusted Sales** and select **Create Bins**.
4. In the **Create Bins** dialog box, for **Size of bins**, type **500**.
5. Drag **Adjusted Sales (bin)** on top of **Sales (bin)** on the **Rows** shelf.

The diagram shows all sales that are over \$8,500 at the \$8,500 level.




Note: Aggregated formulas are not supported in bins.

Adapted from <http://kb.tableausoftware.com/articles/knowledgebase/binning-measures-data-analysis>
© 2003-2013 Tableau Software. All Rights Reserved

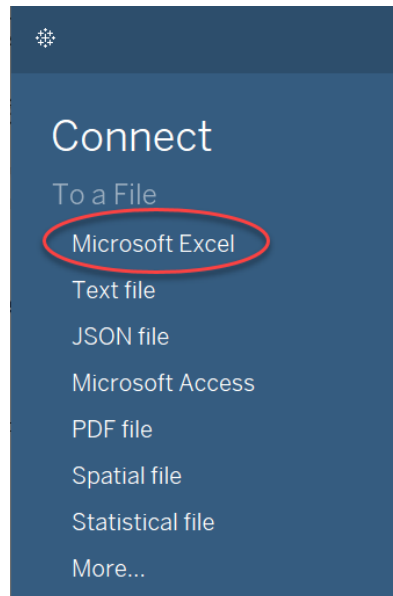
Data Preperation for Task 3

Create a Tableau workbook that connect to the **Sample - Superstore** data source.

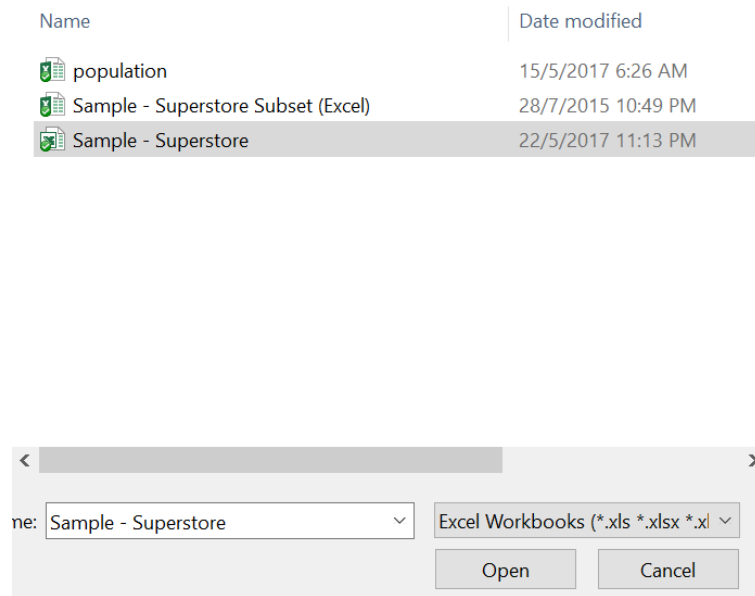
Step 1. Click on  icon to go to **Data Source** page. Under Connect To a File, select **Microsoft Excel**.

Step 2. From the file open window, select the **Sample - Superstore** excel file.

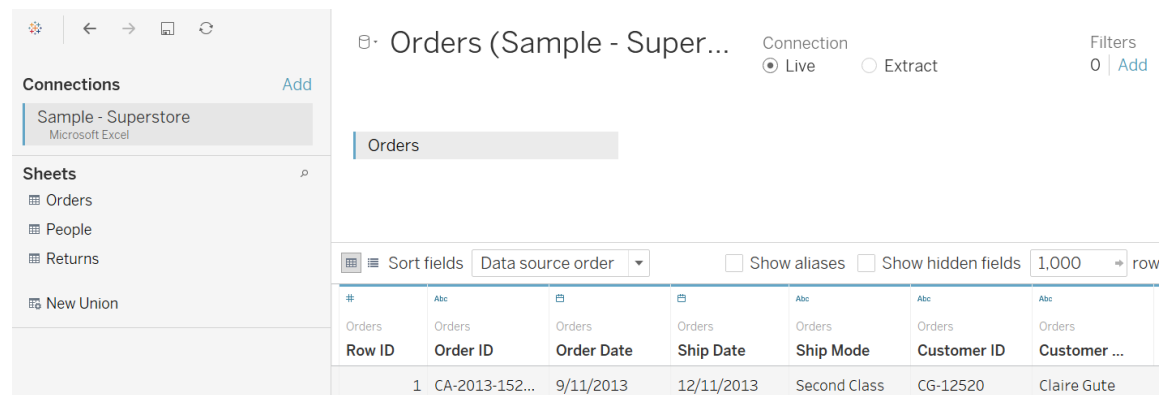
Step 1



Step 2



Step 3: Drag **Orders** to the *canvas*.



Task 3: Build a Box Plot

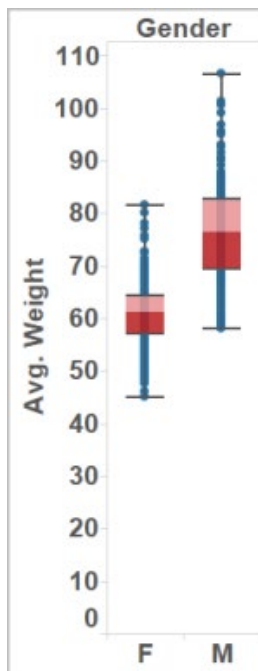
Here is a box-and-whisker visual:



The box represents the values between the first and third quartile (also known as the interquartile range= $Q_3 - Q_1$). The whiskers represent the distances between the lowest value to the first quartile and the third quartile to the highest value. Each quartile has a specific numeric value, determined from the data set. You start by determining the median of the data set. That is where the box turns from grey to light grey. Then, the upper and lower quartiles are determined.

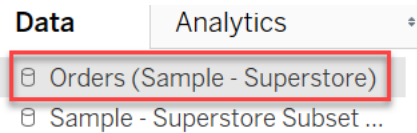
In Tableau, box plots also known as box-and-whisker plots, to show the distribution of values along an axis. Boxes indicate the middle 50 percent of the data (that is, the middle two quartiles of the data's distribution).

You can configure lines, called whiskers, to display all points within 1.5 times the interquartile range (in other words, all points within 1.5 times the width of the adjoining box), or all points at the maximum extent of the data, as shown in the following image:

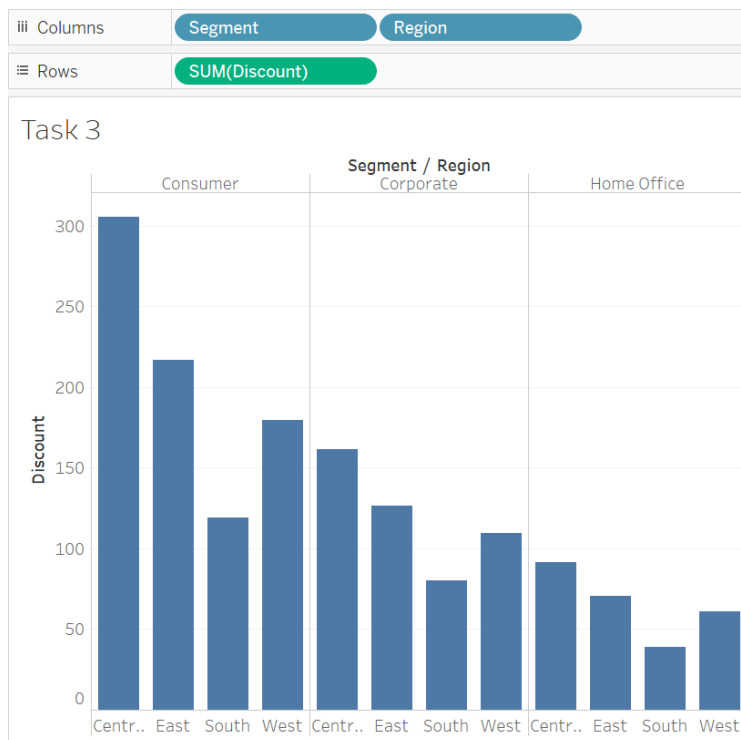


To create a box plot that shows discounts by region and segment, follow these steps:

1. Open a new sheet and rename it as **Task 3**.
2. Make sure **Orders (Sample – Superstore)** is selected.



3. Drag the **Segment** to the **Columns** shelf.
The measure is aggregated as a sum and row headers appear, identifying three segments.
4. Drag the **Discount** to the **Rows** shelf.
Tableau creates a vertical axis and displays a bar chart—the default chart type when there is a dimension on the **Columns** shelf and a measure on the **Rows** shelf.
5. Drag the **Region** to the **Columns** shelf and place it to the right of **Segment**.
Now you have a two-level hierarchy of dimensions from left to right in the view, with regions (listed along the bottom) nested within segments (listed across the top)



6. Click **Show Me** in the toolbar, then select the box-and-whisker plot chart type.

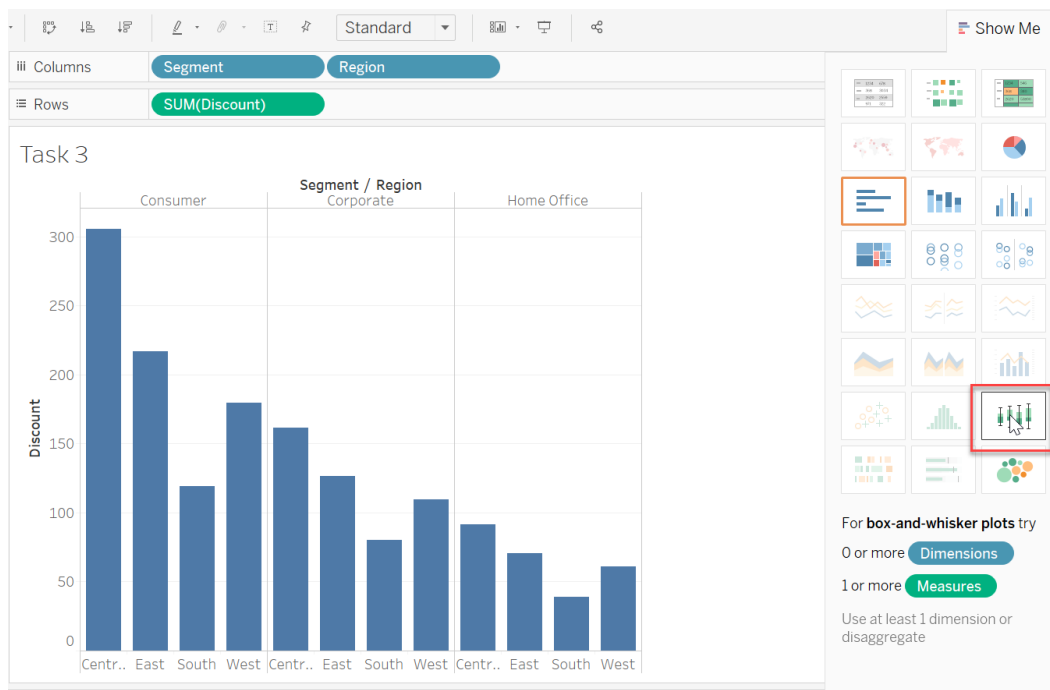
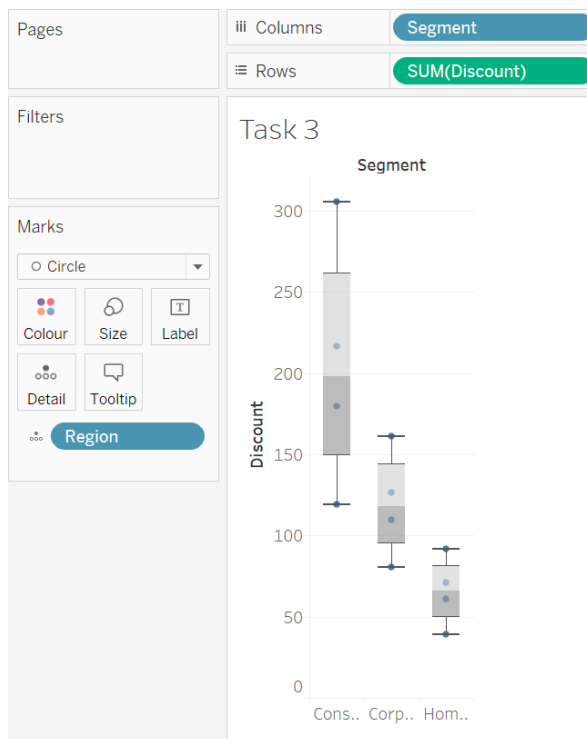


Tableau displays the following box plot:



Notice that there are only a few marks in each box plot. Also, Tableau reassigned **Region** from the **Columns** shelf to the **Marks** card. When you changed the chart type to a box plot, Tableau determined what the individual marks in the plot should represent. It determined that the marks should represent regions. We'll change that.

7. Drag **Region** from the **Marks** card back to **Columns**, to the right of **Segment**.



The horizontal lines are flattened box plots, which is what happens when box plots are based on a single mark.

Box plots are intended to show a distribution of data, and that can be difficult when data is aggregated, as in the current view.

8. To disaggregate data, select **Analysis** → **Aggregate Measures**.

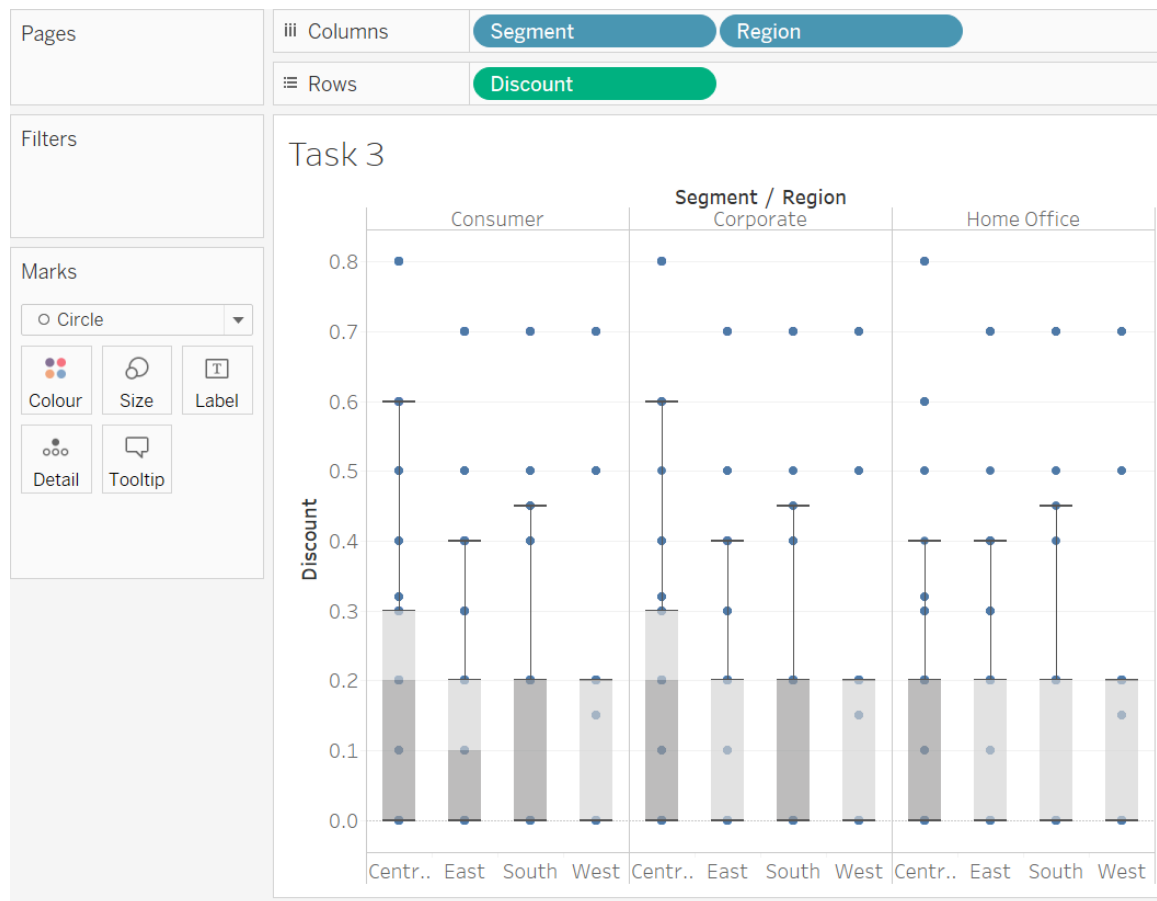
This command turns aggregation on or off, and because data is aggregated by default in Tableau, the first time you select this command, it disaggregates the data.

Disaggregating your data means that Tableau will display a separate mark for every row data value in your data source. This can be useful for **analyzing measures** that you may want to use both independently and dependently in the view.


For example, you may be analyzing the results from a product satisfaction survey with the Age of participants along one axis. You can aggregate the Age field to determine the

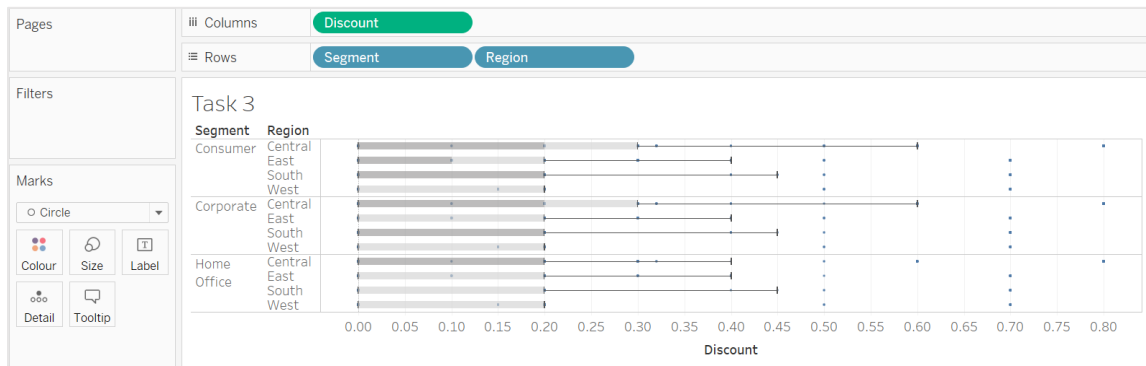
average age of participants or disaggregate the data to determine at what age participants were most satisfied with the product. Disaggregating data can be useful when you are viewing data as a scatter plot.

Now, instead of a single mark for each column in the view, you see a range of marks, one for each row in your data source.



- The view now shows the information we want to see. The remaining steps make the view more readable and appealing.

10. Click the **Swap**  button to swap the axes: The box plots now flow from left-to-right:



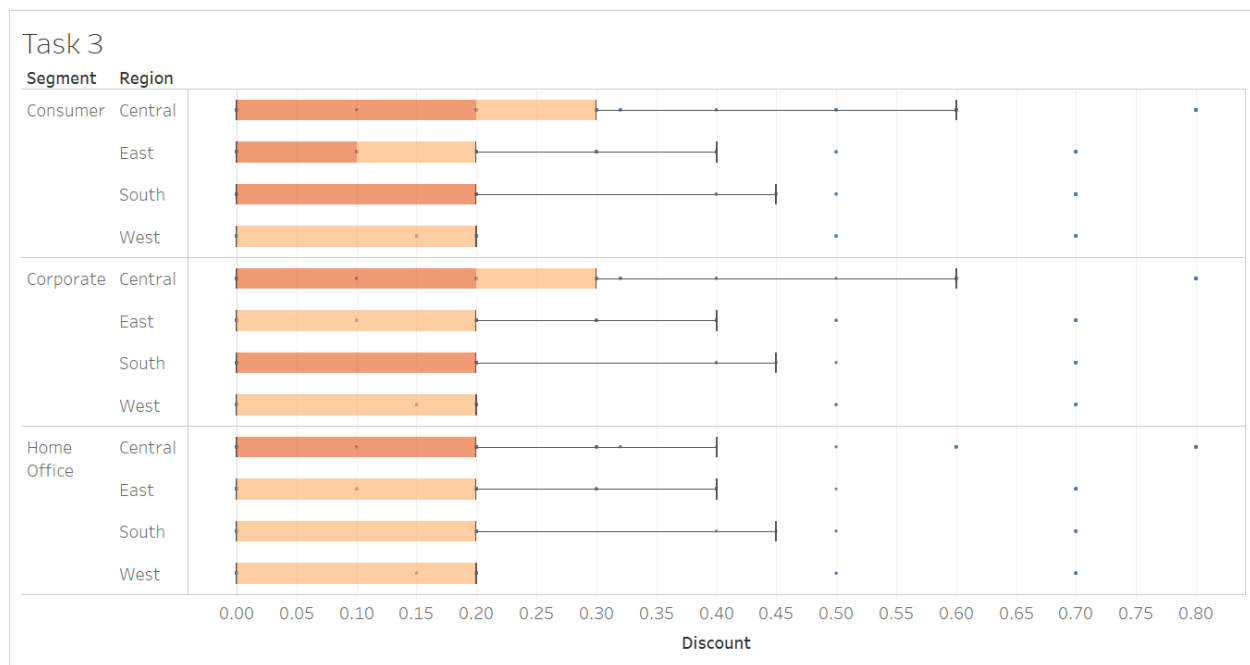
11. Right click the bottom axis and select **Edit Reference Line**.
12. In Edit Reference Line, Band, or Box dialog box, in the Fill drop-down list, select a colour scheme.

13. In the Fit box, select **Entire View**.

You may have noticed that Tableau plots several data values outside the whiskers though the whiskers are supposed to represent the minimum and the maximum of the values. By default, the whiskers extend to 1.5 times the interquartile range (IQR) from the edge of the box. $IQR = Q3 - Q1$ so the range between the upper quartile and the lower quartile. So the upper whisker is by default at the $Q3 + (Q3 - Q1) * 1.5$ value while the lower whisker is automatically at the $Q1 - (Q3 - Q1) * 1.5$ data value. This can be simply changed to extend to the actual minimum / maximum by editing the reference lines along the axis.

For more on these options, see [Adding Box Plots](#).

Now your view is complete:




You can see that the discount was the same for all segments in the West. You can also see that the interquartile range (from the 25th percentile to the 75th percentile) for discount was greatest in the Central region for the consumer and corporate segments.

Adapted from http://onlinehelp.tableau.com/current/pro/desktop/en-us/buildexamples_boxplot.html

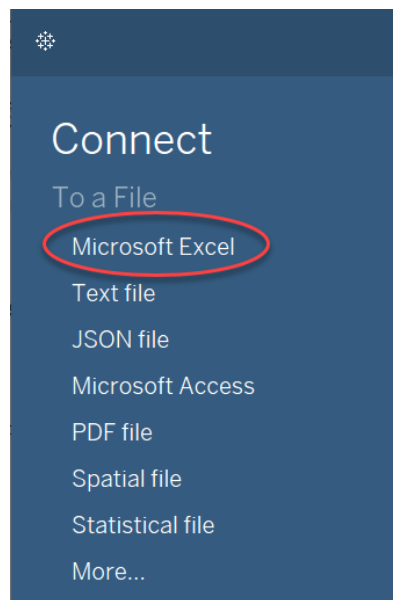
© 2003-2013 Tableau Software. All Rights Reserved

Data Preperation for Task 4

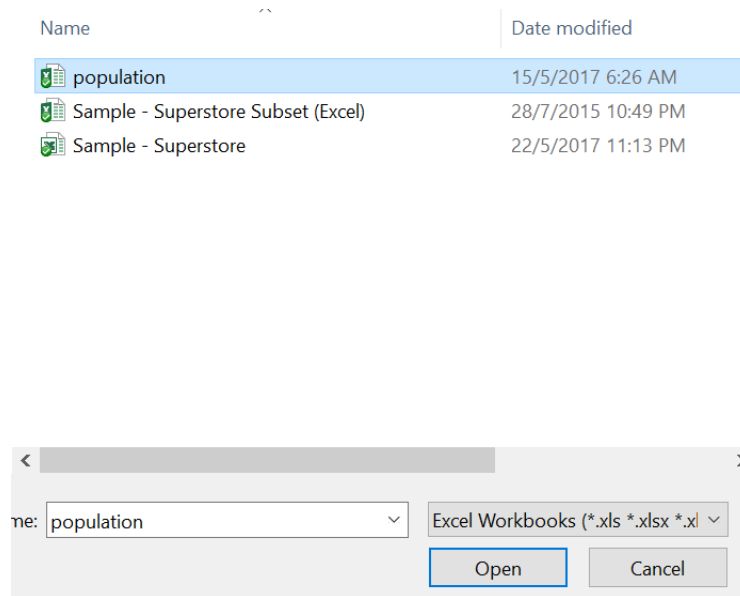
Step 1. Click on  icon to go to **Data Source** page. Under Connect To a File, select **Microsoft Excel**.

Step 2. From the file open window, select the **population** excel file.

Step 1



Step 2

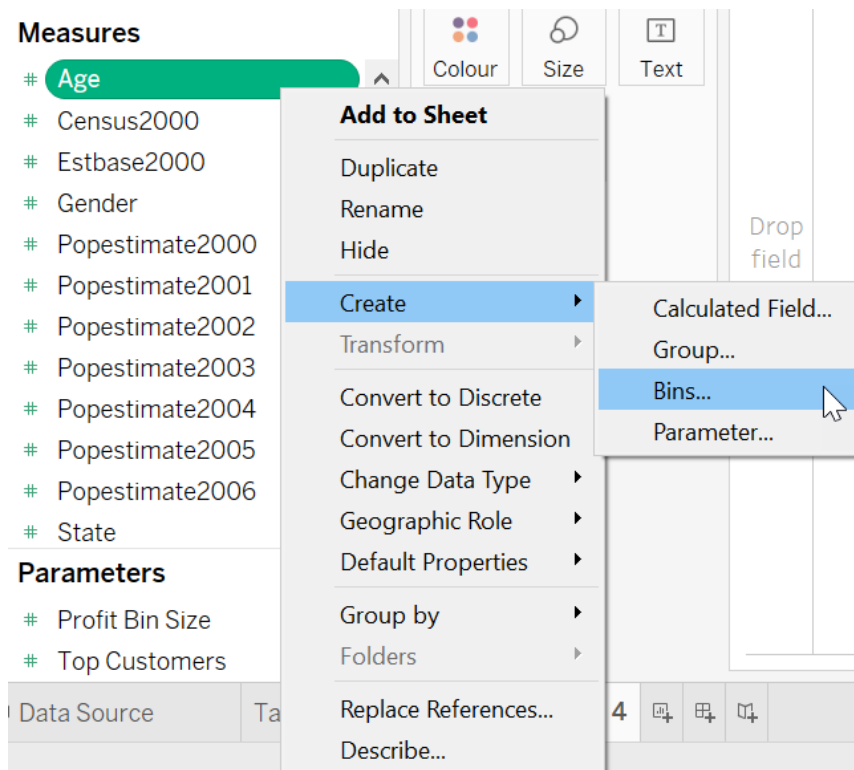


Task 4: Creating a Population Pyramid

A population pyramid, also known as an age structure diagram, is a graphical illustration that shows the distribution of various age groups in a population. One common population group that is effectively visualized through this type of diagram is male and female populations by age. To create a population pyramid, you must separate the population into males and females and then create “bins” or age groups.

1. Open a new sheet and rename it as **Task 4**.
2. Make sure the **population** data source is selected.

- Right click the **Age** field and then select **Create Bins**.



- In the Create Bins dialog box, type a bin size based on the age groups that you are interested in.

The screenshot shows a dialog box titled 'Edit Bins [Age]'. It contains the following fields and controls:

- New field name:** A text box containing 'Age (bin)'.
- Size of bins:** A dropdown menu showing '10' and a 'Suggest Bin Size' button.
- Range of Values:** A section containing:
 - Min:** A text box with '0'.
 - Diff:** A text box with '86'.
 - Max:** A text box with '86'.
 - CntD:** A text box with '86'.
- Buttons:** 'OK' and 'Cancel' buttons at the bottom.

5. Drag the **Age (bin)** field to the **Rows** shelf.

Task 4

| Age (bin) | |
|-----------|-----|
| 0 | Abc |
| 10 | Abc |
| 20 | Abc |
| 30 | Abc |
| 40 | Abc |
| 50 | Abc |
| 60 | Abc |
| 70 | Abc |
| 80 | Abc |

6. Select **Analysis** → **Create Calculated Field**.
7. In the Calculated Field dialog box, make the following selections to create a formula for the male population:
 - a. In the **Name** textbox, type **Male Population**.
 - b. In the **formula** textbox type the following formula and click **OK**.

If [Gender] = 1 then [ESTBASE2000] End

In this case, the census data has defined the Gender value for male as "1." The field "ESTBASE2000" contains estimated population values.

Male Population population (population)

If [Gender] = 1 then [ESTBASE2000] End

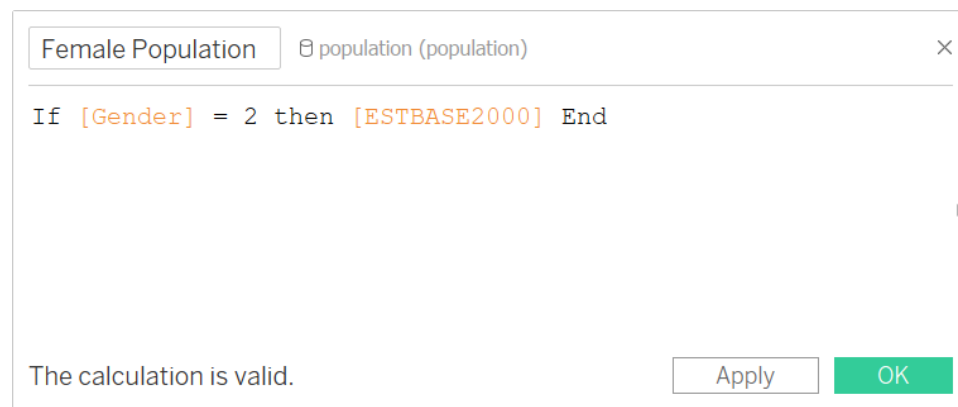
The calculation is valid.

Apply OK

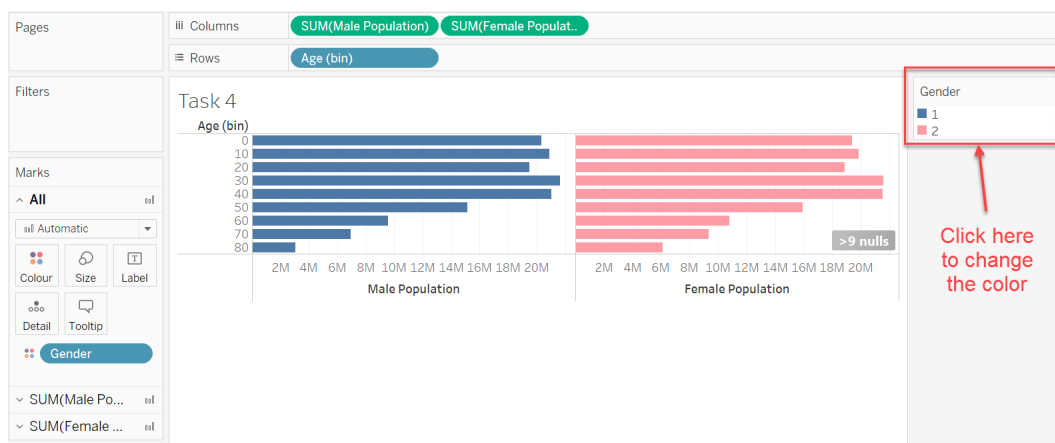
8. Select **Analysis** → **Create Calculated Field**.
9. In the Calculated Field dialog box, make the following selections to create a formula for the female population:
 - a. In the **Name** text box, type **Female Population**.
 - b. In the **formula** text box type the following formula and click **OK**.

If [Gender] = 2 then [ESTBASE2000] End

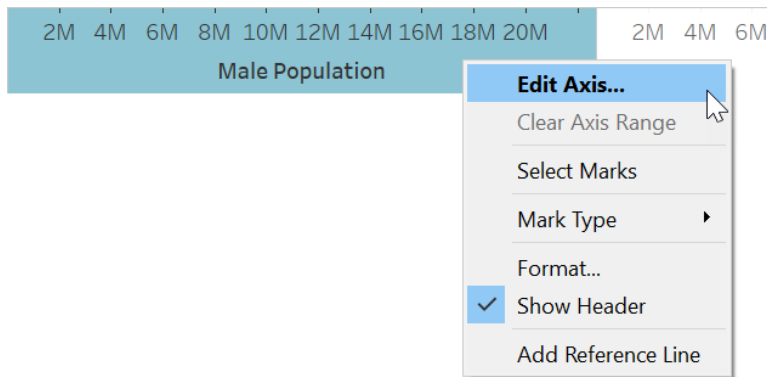
In this case, the census data has defined the Gender value for female as "2". The field "ESTBASE2000" contains estimated population values.



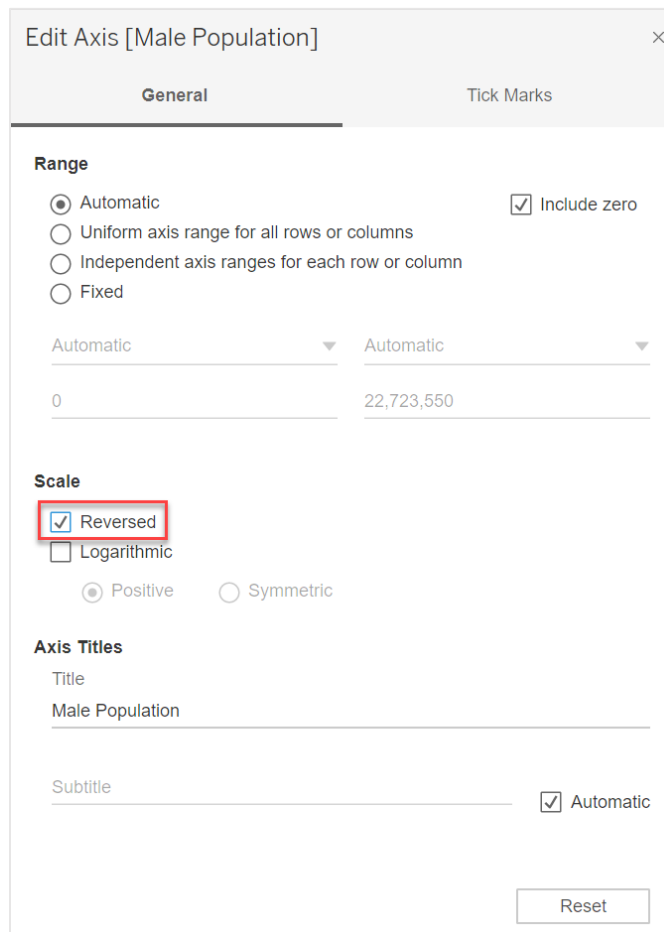
10. Drag the **Male Population** and **Female Population** fields to the **Columns** shelf.
11. Right click **Gender** to convert it to **Dimension**.
12. Drag **Gender** and place it in the **Colour Marks** card.



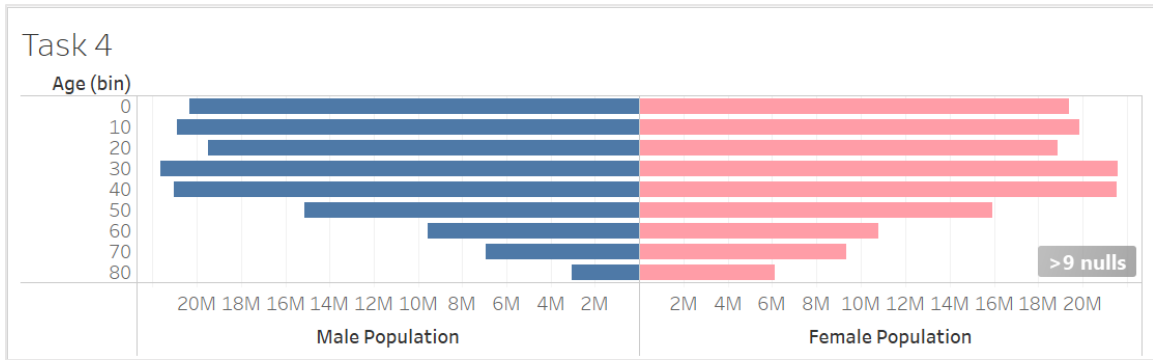
13. Right click the axis for **Male Population** and select **Edit Axis**.




14. In the Edit Axis dialog box, select the check box next to **Reversed** to reverse the order of values on the axis.



15. Click **OK**.



16. Click  to sort the population in ascending order.

