**Why Tableau?**

“In 2014, Tableau (aka DATA) is the most talked about visual analytics company in the world. Last week their annual conference was held in Seattle with a record-breaking, impressive crowd of over 5500+ in attendance despite the addition of numerous regional conferences around the globe. A [litany of analytics ecosystem partners](http://tcc14.tableauconference.com/expo-hall) also participated including top big data players (Cloudera, Hortonworks, Amazon, IBM, MapR, Splunk, Google), a variety of ETL offerings (Alteryx, Informatica, Paxata, Trifacta, Clover ETL, Lavastorm), predictive solutions (Big ML, Rapid Insight), database vendors (HP Vertica, Teradata, Actian, Exasol), implementers, and other add-on offerings. The business intelligence and analytics industry analyst community was also there to cover what DATA had to say.

Without a doubt, this event has grown rapidly over the past few years much like the Tableau user base. Historically Tableau’s land-and-expand momentum has been exceptionally impressive with sales growth of over 80% to 100% each year – literally doubling sales year over year over year over year over year.” –Jen Underwood, Microsoft Sr Program Manager of Business Intelligence & Analytics. <http://www.jenunderwood.com/2014/09/15/tcc2014funbuthohum/>

**Introduction to Tableau**

This handout is an orientation to the Tableau software. Tableau is becoming the industry standard for data visualization. This is because:

* It can connect to a variety of data sources (local data files or data servers);
* It facilitates interactive data exploration and visualizations;
* It requires very little programming knowledge (none if you are using basic functionalities);
* Visualizations can be “published” online and shared via blogs or other websites.

There are actually several versions of Tableau, two of which we’ll be looking at in this course:

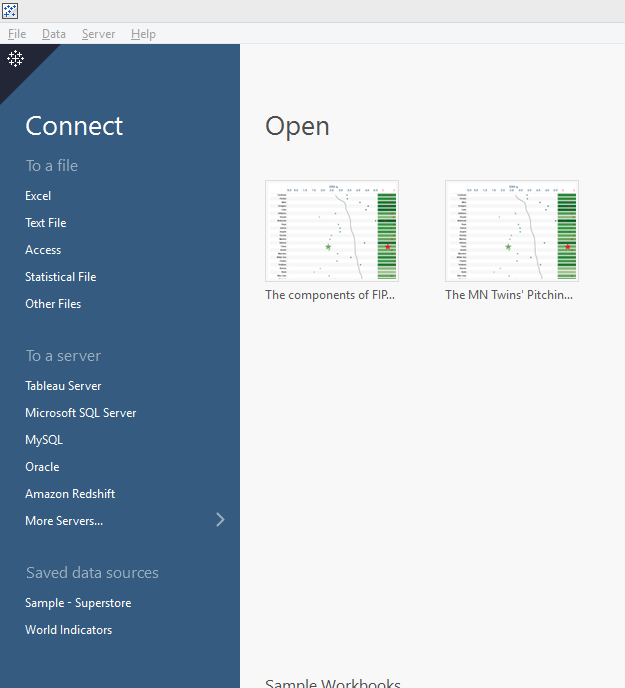
* Tableau Desktop
  + Tableau Desktop allows you to save your workbooks locally
  + Facilitates connection not only to locally-stored data but also data from a myriad of different databases such as Google Analytics or Amazon (if you have the credentials!).
  + It’s not usually free, but you get it because you’re a student!
* Tableau Public
  + Free to the general public.
  + You can only connect to data from a local Excel or text file
  + Workbooks are saved on Tableau Public’s server rather than locally
  + You can publish the visualizations you create with Tableau Desktop to Tableau Public, and share them.
  + Some functionality limited compared to Desktop (can’t connect to R in Tableau Public, for example)

Although Tableau has some data manipulation capabilities as we will see, **it is *not* a data management tool**: programs such as R, Python, even JMP are much better for those purposes. Thus, if you have data you want to visualize that needs cleaning first, it is best to do the cleaning outside of Tableau using a more suitable tool, then save the cleaned data as an Excel or .csv file and connect to it from Tableau.

**Know your data tools. Use different tools for different tasks.**

**Connecting to data**

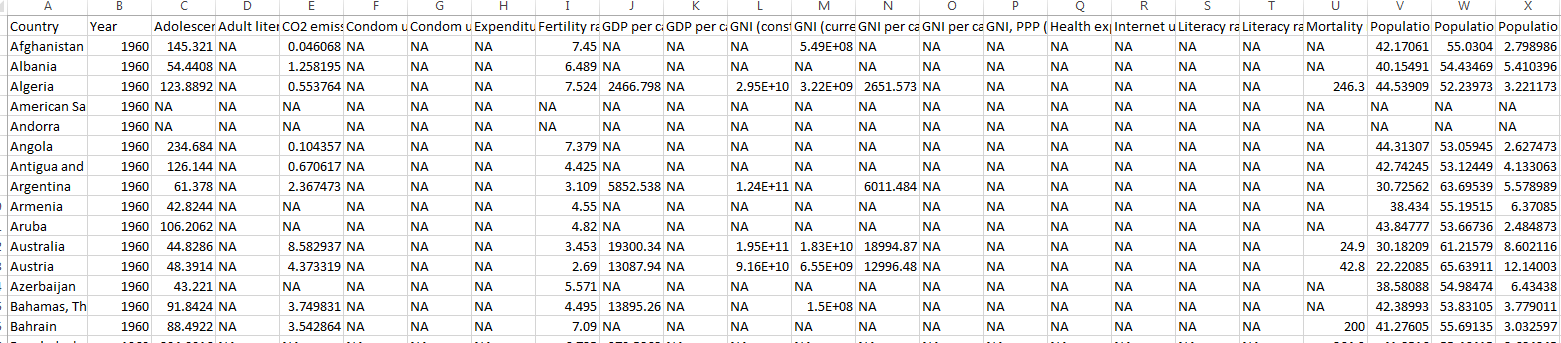
The first time you open Tableau Desktop you will see the “Home Page”:



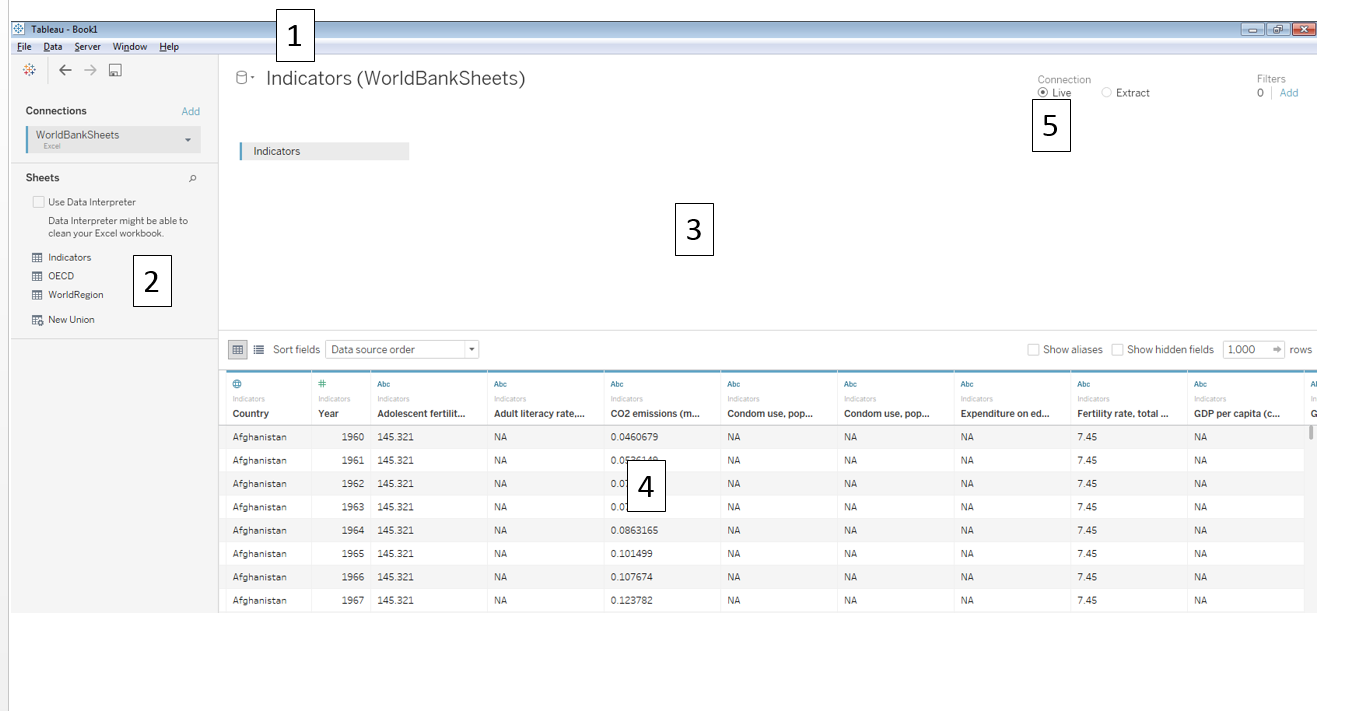
For this course, we will be connecting to data in local files. Tableau allows connection to:

* Excel files: these must have a .xls or .xlsx extension
* Text files: these include tab- (.txt or .tab) or comma-delimited files (.csv)
* Statistical files: SAS files (.sas7bdat); SPSS files (.sav); and R files (.Rdata). This is especially nice if you’ve done some data cleaning in one of these languages and want to create visualizations in Tableau using the cleaned data.

We’ll start by connecting to the World Bank data. This is a data set on economic development indices since 1960 for countries around the world. Here’s a snippet of the first sheet:

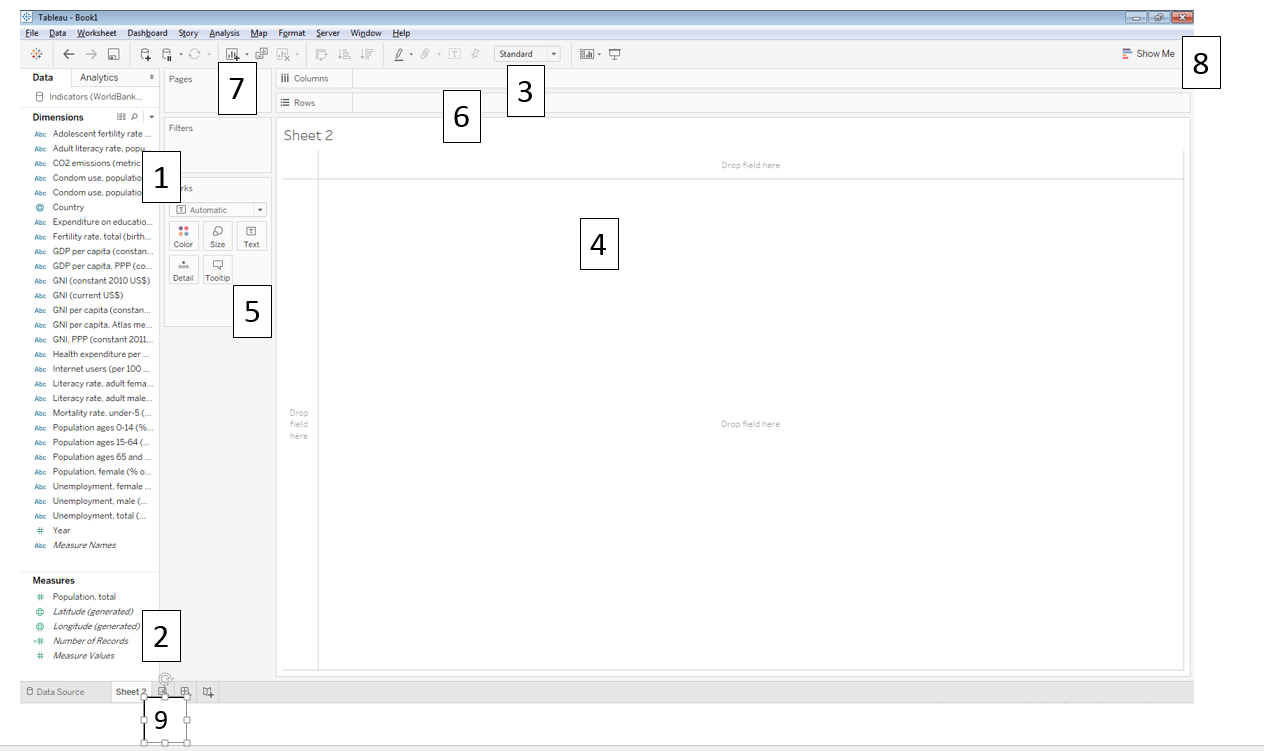


After saving this file to your local computer, click “Connect to Excel” and browse to it. Note you will see the available worksheets



1. The data file you are currently connected to. Click to rename it within Tableau.
2. Recently-used data sheets. This is also where names of multiple sheets will show up, if you connect to an Excel file with multiple tabs.
3. The data sheets you want to visualize: you can click-and-drag sheets from “2” into this space. This is where you would link multiple data sources; we’ll see this later.
4. A view of your data.
5. The connection type: live or extract. A Tableau Extract is a special type of data base that Tableau creates that is optimized for speed. You will want to create an extract (with a .tde extension) if you have a large data set; here we only have 30 rows so we’ll stick with Live.

Now, click on the “Sheet 1” tab down in the lower left corner, and you should see this view:



Before we investigate this view, a few important facts about Tableau vocabulary:

* “Fields” are what Tableau calls “Variables.” Country; Population; Year…these are all “fields” in Tableau-speak.

**Discrete vs Continuous**

* Discrete fields will lead to HEADERS, or DISCRETE BINS. Discrete fields, and their pills, are **BLUE.**
* Continuous fields will lead to CONTINUOUS AXES. Continuous fields, and their pills, are **GREEN**.

**Dimensions vs Measures**

* Dimensions are NOT AGGREGATED.
* Measures are fields that by default are AGGREGATED when visualized (summed, averaged). It makes sense to perform arithmetic operations on measures, but not on dimensions.
* **Do you think string fields should be represented as dimensions or as measures?**

Tableau automatically classifies all fields into either Dimensions or Measures. Any measure can be changed to by right-clicking the field (or fields), clicking the arrow that appears, and clicking “Convert to Dimension”; and vice versa. Changing from discrete to continuous can be done similarly.

Now on to the view:

1. All the dimensions in the data set
2. All the measures in the data set
3. Columns and Rows shelves: control grouping (for Dimensions) and axes (for Measures). Visualizations are created by dragging fields to these shelves.
4. The “canvas”, where visualizations appear. You can drag variables here instead of the Columns and Rows shelves.
5. The Marks card: this controls colors, shapes, sizes, labels, etc.
6. The Filters shelf: filter visualizations by Dimensions or Measures
7. The Pages shelf: filters by animating based on field
8. The “Show Me” card: this is kind of like a “shortcut” card, and allows you to view available visualizations for selected fields.
9. All Worksheets, Dashboards, and Stories currently in use; and buttons for new Worksheets/Dashboards/Stories.

Note there are a few fields that were not in the original data set:

* *Measure Names* (Dimension): Contains all the field names of the measures. You can imagine this as a vector of strings, containing all the names of the measures (quantitative variables). In this case, *Measure Names* contains the strings ‘Babip’, ‘BB’, ‘Home runs’, ‘ERA’, etc.
* *Number of Records* (Measure): Count of the entries in the data source. This is useful for showing how many of each dimension exist in the data set (e.g., how many rows there are for each league). Imagine this as a columns of 1’s
* *Measure Values* (Measure): This contains all the numeric values of all measures. Imagine it as a super long vector, with all numeric values in the data set. This is a useful field to use if you want to blend multiple fields in the same graph.

**There is one issue right off the bat:** because the original Excel sheet had “NAs” representing missing values, ***all of these fields are thought of as strings, and hence categorized as dimensions!***

The Worksheet is where the mechanics of graph creation take place. The *Dashboard* is where you can craft a message with your visualizations by:

* Combining worksheets
* Including text, images, or URLs with your visualization
* Interact with the data

Although you can perform some formatting in Dashboards, the bulk of the viz creation must be done on Worksheets.

The *Story* allows you to tell an even longer story, with multiple Worksheets or Dashboards. You cannot format vizzes in stories.

For example, examine the “[The Components of FIP](https://public.tableau.com/profile/silas.bergen#!/vizhome/ThecomponentsofFIP/FIPStory)” workbook from my Tableau Public site or the course webpage. Note how the graph and the table are each separate Worksheets, which together have been combined into an interactive Dashboard with a main title. This Dashboard has been compiled into a Story, with different default sorts and corresponding messages for each phase of the Story.

