OPEN DATA TOOLKIT

Lauren Barrett
Nicole Hoesterey
Mariana Quevedo Vallejo
Junhe Yang

School of International and Public Affairs | Columbia University

Purpose

This Open Data Toolkit is intended to provide the public with information to make better use of the open data that is available. It covers the process of working with data from data sourcing to presenting datasets to different audiences including decision makers.

Overview

A team of Columbia University graduate students created this Open Data Toolkit for Open Development Initiative (ODI) as part of a consultancy on behalf of the School of International and Public Affairs (SIPA) and the East-West Management Institute in Spring 2015. The team traveled to Phnom Penh and Hanoi to assist in the creation of Donor and Development Assistance sections for the Open Development Cambodia (ODC), Open Development Mekong (ODM), and Open Development Vietnam (ODV) websites. In field, the team led stakeholder interviews, conducted workshops on data sourcing and visualization, and initiated the launch of this Open Data Toolkit to empower researchers and the general public to make better use of data.

This toolkit gives a comprehensive introduction and how-to of each stage of the creative use of open data - from describing what it is, to finding it and extracting it, all the way through to writing about it and visualizing it.

In the appendices, contents of two workshop trainings (using RStudio and Tableau) are included.

Table of Contents

CHAPTER I

Introduction to Open Data

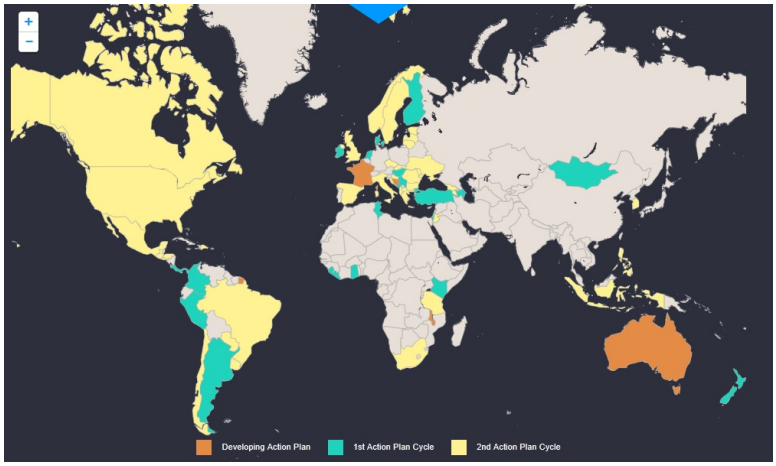
What is Open Data?

The world creates and stores more data every two days than it did since the beginning of time up until 2003.[1] Being able to access all this data freely has grown into a salient policy issue for governments especially and developing countries are facing unique challenges of all this new data with regards to its citizens - how much data should they share with their citizens and how? Who has access to it? Non-profits such as the Open Development Initiative (ODI) are seeking to leverage this new wealth of information to build up an open data and information network that the public can freely access.

Open data is data that is free to use, re-use and distribute. It has grown in popularity along with open-source software in which programmers can collaborate, share, design and use code already available to the general public. The support and popularity of open data rests on the assumptions that open data makes people and organizations more knowledgeable, that people have the freedom to access this information, and that it is free to access.

Many *governments* are signing on to Open Government Partnership as a way of disclosing their data to their citizens free of charge or licensing obligations. Below is an action plan cycle map of governments around the world.[2] Open data promotes transparency and encourages citizens to hold their governments accountable.

Businesses and *nonprofits* are using open data to disclose information about their organizations, increase transparency, and collaborate with other organizations through sharing their data.



Research centers and *academic institutions* publish their data so that the general public can access this open data in a way that strives to be accurate, timely, and accessible.

[1] According to Google's CEO Eric Schmidt
[2] Open Government Partnership, http://www.opengovpartnership.org/

CHAPTER II

Sourcing Data

The Research Basics of Sourcing Data

1. Determine the question you are trying to answer.
2. Seek credible sources.
3. Know how to interview.
4. Understand the format of the data files.

1. Determine the question you are trying to answer.

Who is your audience?
Why is the audience interested in this data?
What data is available?

For example, university students, government officials, or journalists might want to know about dams in Brazil. A first question they may want to know is "How many dams are there?" Put this into context with the rest of the world - is it a relatively large or small number compared with other countries? How many dams are completed and how many are in progress? What are the effects of the dams? On local communities? On energy needs/expenditures? More will be discussed in latter sections on narrating your data, but you need to have a question in mind that you want to try to answer.

2. Seek credible sources.

Primary v. Secondary - Many times an article you read online will have been adapted from another article, or is a summary of data recently published online. Read through the article carefully and note any references to other articles, websites, or specific individuals that the article was getting its information from. Seek out the original source of information.

Accurate - This means no partial or biased data. Who is your source here? Sometimes organizations publish data that push its own agenda, or leave out information that may be potentially harmful to its reputation or the message it supports. Critiquing the data you find through academic sources is helpful because the author(s) often addresses all sides of biases. Be wary of data that looks too good to be true or that is published by institutions that have an explicit agenda.

Triangulate - This is standard in reporting. Triangulating your information means finding three credible sources that can independently attest to the information or data's validity.

3. Know how to interview.

This is especially important if the data you are seeking to source is not already available. There are several types of in-person interviews:

Structured - Questions are recited exactly as they are prepared, and cues such as body language, awareness of non-prompting, and other verbal and behavioral cues are noted.

Semi-Structured - Most popular. Interviewers follow question guidelines, however not word by word. They may paraphrase or improvise based on the direction the meeting/interview is taking. General goals, questions, and desired topics are understood before the interview.

Unstructured - Most informal. These interviews are typically resourced as conversations, in which the interviewer may ask questions with no specific guideline to follow. Questions are generally open-ended and the interview is flexible as it can take any direction the two parties choose to follow. Sometimes, this is the best way to discover information that the interviewer may not have known to even ask.

Focus Groups - These are sessions of 2 or more persons in which the interviewer would like to hear responses from community members, organization, or specific groups as a whole. Special attention should be paid to those who claim to speak on behalf of the group, talk over others, or those that are especially silent. It is the responsibility of the moderator to diplomatically include everyone in the discussion, as well as take notes on what is said, not said, and other behavioral cues.

4. Understand the format of the data files.

Data is stored and presented in a variety of file formats. Some formats are easy for humans to read, but difficult for computers, while others are easy for computers to read, but may initially look incomprehensible to a human. Understanding the file format of your data will help you know what you can do with the data and if you need to take steps to save the data in a different format before using it. Some data sources are more "open" than others, allowing more people to access the data and use it for their own analysis. The following are common formats found while conducting research:

PDF (.pdf) - Portable document format files contain a fixed-layout flat document that includes the text, images, and graphics contained in the original. Saving as a .pdf allows the file to be opened on any computer, regardless of the application, software, or operating system used to create the file. PDFs are valued for their readability and aesthetic display. Data contained within PDF files is difficult for computers to read, even though it is easy for humans to see. A program to extract data from PDFs (see *Scraping Data* section) will be needed to convert tables within a PDF to a more manipulable format. It is not recommended to save open data, such as data tables, in PDF format, although it is a nice format for publishing reports or articles.

Excel (.xls or .xlsx) - Microsoft Excel is a spreadsheet program that allows for recording data in rows, columns, and cells. The program can be used for calculations, formulas, basic graphs and visualization, pivot tables, and macro programming. Several spreadsheets can be saved within a single Excel workbook. This format is often used for presenting open data since it is widely understood by people and allows for basic analysis within the program, and can be easily imported into databases and other data visualization tools. However, it is important to note that Excel is not an open-source software.

Comma-separated values (.csv) - This file type is a text file that contains information in a table structured format where the data is separated by commas. It is often confused with an Excel file since .csv files often open automatically in Excel on most computers. CSV files are useful when importing/exporting large tables between format types or databases (for example, when exporting a table from a database to a spreadsheet program that cannot directly read data from the database the data can be exported as .csv by the database and imported as a .csv by the spreadsheet program). When opening a .csv file in Excel, be aware that formulas and formatting that you add to the data will not be saved unless the spreadsheet is saved in an .xls format. Only a single sheet in a workbook application such as Excel can be saved as a .csv at a time.

JPEG (.jpg)- This is the most common file format for digital images. Saving a photo as a JPEG allows the user to choose a level of compression, trading off image size for image quality. Images can be compressed to 1/10 the size of the original data without perceptible loss in quality, making it an ideal format for web or email which require smaller file sizes. Most file types are "lossless", meaning that when the file is reopened all the original data will remain the same. However, JPEG is a "lossy" format, meaning that when an image is compressed, the original data is modified and will lose some of its detail.

Text (.txt) - .txt files are considered a universal format since they can be opened and read by any text editing or word processing software. These files contain text without formatting (such as bold or italics) and generally match that of a system terminal or simple text editor. Text is generally stored in ASCII format (American Standard Code for

Information Interchange) which is based on the English alphabet, although there are modern character encoding schemes used to support other languages that based on ASCII.

XML (.xml) - XML (eXtensible Markup Language) is a language for describing data and documents in a format that both humans and machines can read. XML is often used for the interchange of data over the internet; XML consists of a series of tags (e.g. <country>), with attributes and text associated with each tag. HTML, which is used to describe the contents of a webpage, uses a pre-defined set of XML tags, but there are no restrictions on the tag names in XML files used to describe datasets. An XML reader is required to analyze XML-formatted information and transfer it to an application for display and reading.

JSON (.json) - JSON stands for JavaScript Object Notation. It is used to transmit data from a server to an application, and can be used as an alternative to XML (some find it easier to understand than XML). It is considered a 'lightweight' data interchange format that is human-readable and text based. There are some websites, such as json-csv.com, that can convert JSON files to CSV files for easier viewing (See *Scraping Data* section).

Example of JSON versus XML data interchange format. JSON on the left, XML on the right[3]

http://localhost:8080/Json/SyncReply/Contacts

```
{
  - Contacts: [
      - {
            FirstName: "Demis",
            LastName: "Bellot",
            Email: "demis.bellot@gmail.com"
        },
      - {
            FirstName: "Steve",
            LastName: "Jobs",
            Email: "steve@apple.com"
        },
      - {
            FirstName: "Steve",
            LastName: "Ballmer",
            Email: "steve@microsoft.com"
        },
      - {
            FirstName: "Eric",
            LastName: "Schmidt",
            Email: "eric@google.com"
        },
      - {
            FirstName: "Larry",
            LastName: "Ellison",
            Email: "larry@oracle.com"
        }
    ]
}
```

http://localhost:8080/Xml/SyncReply/Contacts

```
<ContactsResponse xmlns:i="http://www.w3.org/20
  <Contacts>
    <Contact>
      <Email>demis.bellot@gmail.com</Email>
      <FirstName>Demis</FirstName>
      <LastName>Bellot</LastName>
    </Contact>
    <Contact>
      <Email>steve@apple.com</Email>
      <FirstName>Steve</FirstName>
      <LastName>Jobs</LastName>
    </Contact>
    <Contact>
      <Email>steve@microsoft.com</Email>
      <FirstName>Steve</FirstName>
      <LastName>Ballmer</LastName>
    </Contact>
    <Contact>
      <Email>eric@google.com</Email>
      <FirstName>Eric</FirstName>
      <LastName>Schmidt</LastName>
    </Contact>
    <Contact>
      <Email>larry@oracle.com</Email>
      <FirstName>Larry</FirstName>
      <LastName>Ellison</LastName>
    </Contact>
  </Contacts>
</ContactsResponse>
```

[3] Shandra Locken, "The What, Why and How of JSON for EDI Integration Specialists," Aurora EDI Alliance, 2013.

CHAPTER III

Scraping Data

"Website scraping is a dark art."[4]

What is data scraping?

Data scraping is a technique that allows you to get data that is usable and user-friendly. By scraping, you take the data from a human-readable source (i.e a .pdf document or webpage) and make it readable to a machine (i.e a .csv file that can be read by Tableau or R).

Cardinal Rules of Data-Scraping

1. Do not scrape if you do not have to: always look for a copy of the dataset in some format that you can easily convert to a .csv file using Excel or other conversion tool (import.io). Chances are that the data is out there in some form that will make your life easier.

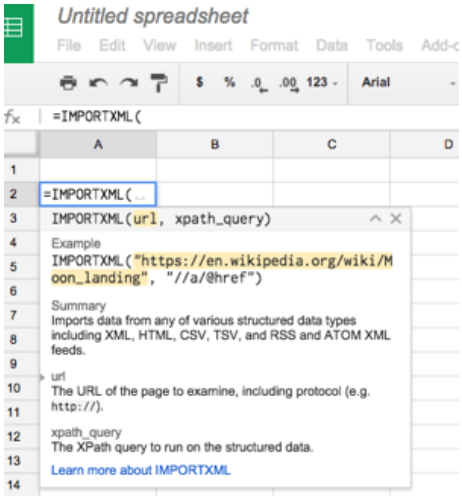
2. Look for an option to download a .csv file directly from the source. Most of the sites that carry information about development have ways to guide readers to build their own report and download the data that they need.

Design the scraper

What type of file do you have?

<u>*A .csv file: you are done*</u> (go to the data format section).

<u>*An .xml file*</u>: use a converter tool. You can also use google spreadsheets command =importxml

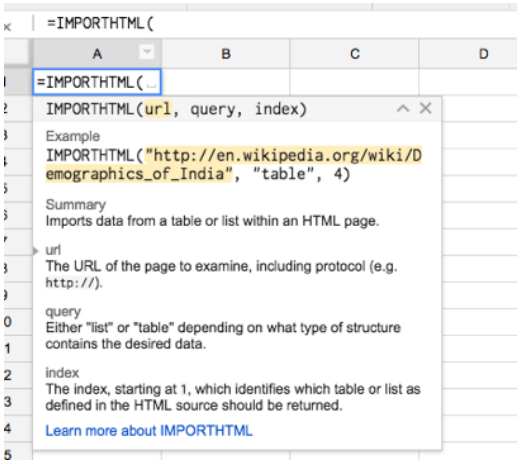


To use it you will need to be familiar with XPATHS[5] (they are used to navigate through the elements of an .xml document). XML documents are trees of nodes. This can be thought of as similar to a family tree with various connections and relationships. For example:

[4] Sara-Jayne Terp
[5] XPATHS Resource: http://www.w3schools.com/xpath/xpath_syntax.asp

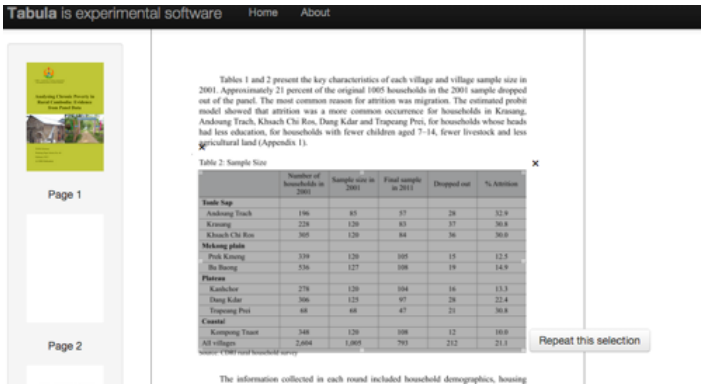
// development: means get all the nodes that are called "development"
@lang=eng: means get all the nodes that have the attribute of the English language.

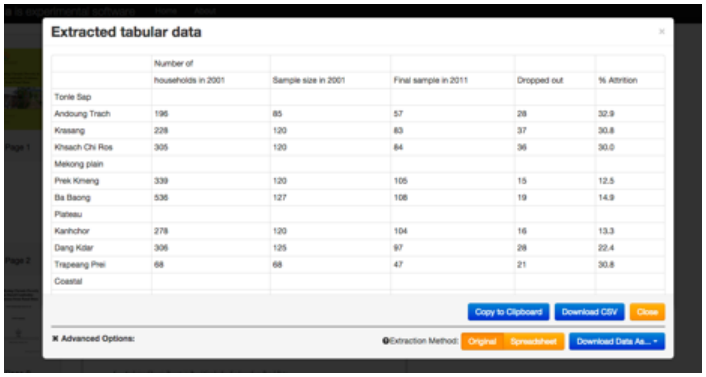
A .json file: use a converter tool such as http://www.convertcsv.com/json-to-csv.htm (there are many others out there that can be found through a Google search). You have to be careful because .json files do not always fit into the .csv table format.



A html table: use the =importhtml command in google. You can export a table or a list imbedded in a html document by writing the url, the type of element you want (table or list) and the number of the table or list.

A table in a .pdf file: use Tabula. You can download Tabula at tabula.technology and follow the download instructions. Once you have installed Tabula, just open it, upload your .pdf document and manually select the table you need to export.





Once you have selected your data, wait for Tabula to extract it and download as .csv:

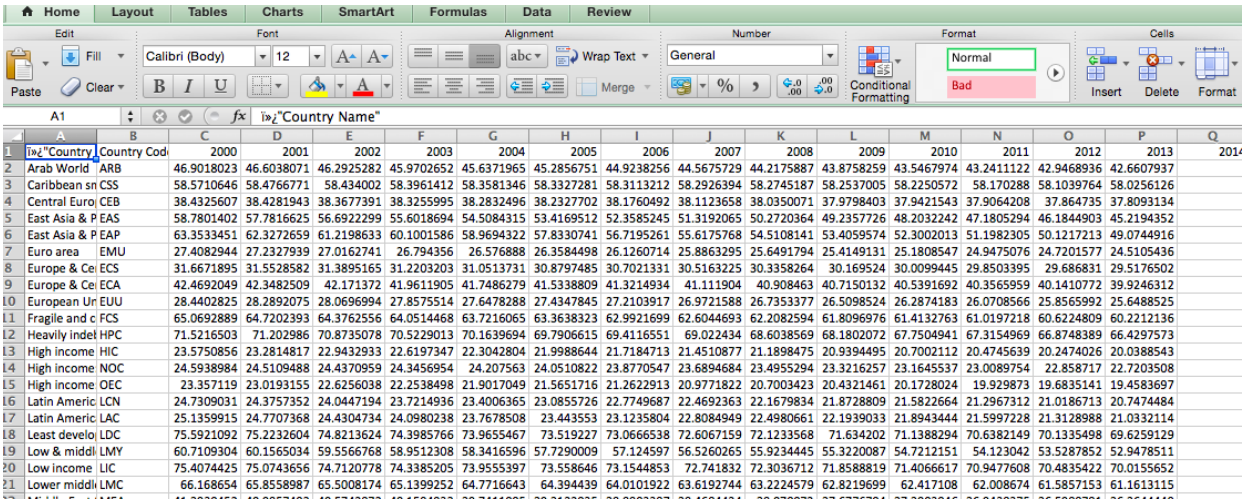
APIs

What's an API?

An Application Programming Interface (API) is a software intermediary that makes it possible for application programs to interact with each other and share data.[6] Each API has its own unique instructions to access data and will return a unique set of data structures when data is requested. An API can be used by applications to access data sets in machine-readable formats like csv, xml, and json. Some examples are:

World Bank's data API: http://data.worldbank.org/node/9
Twitter's search API: https://dev.twitter.com/overview/api
API included in most CKAN datastores: http://docs.ckan.org/en/latest/api/index.html

Example: go to http://api.worldbank.org/countries/all/indicators/SP.RUR.TOTL.ZS?date=2000:2015&format=csv. It will immediately download the csv file for the rural population in all countries from 2000-2015.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	"Country	Country Cod	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	201
2	Arab World	ARB	46.9018023	46.6038071	46.2925282	45.9702652	45.6371965	45.2856751	44.9238256	44.5675729	44.2175887	43.8758259	43.5467974	43.2411122	42.9468936	42.6607937	
3	Caribbean sn	CSS	58.5710646	58.4766771	58.434002	58.3961412	58.3581346	58.3327281	58.3113212	58.2926394	58.2745187	58.2537005	58.2250572	58.170288	58.1039764	58.0256126	
4	Central Euro	CEB	38.4325607	38.4281943	38.3677391	38.3255995	38.2832496	38.2327702	38.1760492	38.1123658	38.0350071	37.9798403	37.9421543	37.9064208	37.864735	37.8093134	
5	East Asia & P	EAS	58.7801402	57.7816625	56.6922299	55.6018694	54.5084315	53.4169512	52.3585245	51.3192065	50.2720364	49.2357726	48.2032242	47.1805294	46.1844903	45.2194352	
6	East Asia & P	EAP	63.3533451	62.3272659	61.2198633	60.1001586	58.9694322	57.8330741	56.7195261	55.6175768	54.5108141	53.4059574	52.3002013	51.1982305	50.1217213	49.0744916	
7	Euro area	EMU	27.4082944	27.2327939	27.0162741	26.794356	26.576888	26.3584498	26.1260714	25.8863295	25.6491794	25.4149131	25.1808547	24.9475076	24.7201577	24.5105436	
8	Europe & Ce	ECS	31.6671895	31.5528582	31.3895165	31.2203203	31.0513731	30.8797485	30.7021331	30.5163225	30.3358264	30.169524	30.0099445	29.8503395	29.686831	29.5176502	
9	Europe & Ce	ECA	42.4692049	42.3482509	42.171372	41.9611905	41.7486279	41.5338809	41.3214934	41.111904	40.908463	40.7150132	40.5391692	40.3565959	40.1410772	39.9246312	
10	European Ur	EUU	28.4402825	28.2892075	28.0696994	27.8575514	27.6478288	27.4347845	27.2103917	26.9721588	26.7353377	26.5098524	26.2874183	26.0708566	25.8565992	25.6488525	
11	Fragile and c	FCS	65.0692889	64.7202393	64.3762556	64.0514468	63.7216065	63.3638323	62.9921699	62.6044693	62.2082594	61.8096976	61.4132763	61.0197218	60.6224809	60.2212136	
12	Heavily inde	HPC	71.5216503	71.202986	70.8735078	70.5229013	70.1639694	69.7906615	69.4116551	69.022434	68.6038569	68.1802072	67.7504941	67.3154969	66.8748389	66.4297573	
13	High income	HIC	23.5750856	23.2814817	22.9432933	22.6197347	22.3042804	21.9988644	21.7184713	21.4510877	21.1898475	20.9394495	20.7002112	20.4745639	20.2474026	20.0388543	
14	High income	NOC	24.5938984	24.5109488	24.4370959	24.3456954	24.207563	24.0510822	23.8770547	23.6894684	23.4955294	23.3216257	23.1645537	23.0089754	22.858717	22.7203508	
15	High income	OEC	23.357119	23.0193155	22.6256038	22.2538498	21.9017049	21.5651716	21.2622913	20.9771822	20.7003423	20.4321461	20.1728024	19.929873	19.6835141	19.4583697	
16	Latin Americ	LCN	24.7309031	24.3757352	24.0447194	23.7214936	23.4006365	23.0855726	22.7749687	22.4692363	22.1679834	21.8728809	21.5822664	21.2967312	21.0186713	20.7474484	
17	Latin Americ	LAC	25.1359915	24.7707368	24.4304734	24.0980238	23.7678508	23.443553	23.1235804	22.8084949	22.4980661	22.1939033	21.8943444	21.5997228	21.3128988	21.0332114	
18	Least develo	LDC	75.5921092	75.2232604	74.8213624	74.3985766	73.9655467	73.519227	73.0666538	72.6067159	72.1233568	71.634202	71.1388294	70.6382149	70.1335498	69.6259129	
19	Low & middl	LMY	60.7109304	60.1565034	59.5566768	58.9512308	58.3416596	57.7290009	57.124597	56.5260265	55.9234445	55.3220087	54.7212151	54.123042	53.5287852	52.9478511	
20	Low income	LIC	75.4074425	75.0743656	74.7120778	74.3385205	73.9555397	73.558646	73.1544853	72.741832	72.3036712	71.8588819	71.4066617	70.9477608	70.4835422	70.0155652	
21	Lower middl	LMC	66.168654	65.8558987	65.5008174	65.1399252	64.7716643	64.394439	64.0101922	63.6192744	63.2224579	62.8219699	62.417108	62.008674	61.5857153	61.1613115	

[6] Open API definition

CHAPTER IV

Exploring and Managing Data

Introduction to Data Exploration

What is data exploration?[7]

Data exploration is usually the first step after you have sourced data with the right format. Users conduct data exploration to understand the information they gathered and to begin analyzing the data. Data in real life often comes in a non-rigid manner that is not user-friendly. For analysis, you will need to narrow down the large bulks of raw data that you collected and prepare the data for moving forward to data analysis, narration, and visualization.

Often, you will find data from different sources in different formats. By exploring the data, you will form a general idea about the basic attributes of the dataset that include: what the variables are (how many observations there are), if the dataset contains missing values, if the data make sense in real life, etc.

What questions do we have in mind when exploring data?

The key to exploring data is to prepare the data for further analysis. You will first need to look at the license and owner of data and make sure that you will be able to use it and share your findings. Second, you should look at the format of data and see if you can work with the specific data format. When proceeding to the third step, you can delve deeper into the content of data and see if the dataset includes the information you want, and if it is relevant to the question you have in mind. A more detailed process of data exploration is included below.

[7] Explanation from Techopedia, http://www.techopedia.com/definition/28789/data-exploration

Exploring and Managing Data in Excel

Excel is a spreadsheet program that is used for storing, organizing, and manipulating data. Many open data sources provide data in Excel format, which is easily downloadable and allows for others to analyze the data without having to convert the data to a different format. Excel can be used to sort, explore, manage, and analyze data. Again, keep in mind that Excel is not an open-source software.

How is Excel organized?

Each spreadsheet is made up of a grid of cells formed by columns (identified by letters) and rows (identified by numbers). Each cell has a unique reference made using the column letter and row number, such as B6 or AA10. Cells can contain data as text, numbers, or formulas. Formulas make calculations based on information contained in other cells, or aggregate data across a range of cells to calculate statistics such as the average, median, or sum. An Excel workbook can be comprised of multiple spreadsheets.

Exploring data in Excel

When first opening data in Excel, take a moment to become familiar with the contents:

- *What variables are presented?* Are these the variables you expected or need?
- *Are there missing values?*
- *Is the data formatted in a crosstab?* A crosstab will have one variable in the columns and another in the rows, such as countries listed in the rows, years listed in the columns, and each cell containing the value of official development assistance (ODA) received by that country in that year (see example of *Incorrect format* below).
- *Do variables have consistent coding?* For example, the variable *respondents sex* may have text values such as female or male in some cells, while binary variables (0 or 1) in others. Another example would be multiple spellings of a country name: Democratic Republic of the Congo, DRC, Dem. Rep. Congo.
- *Are there multiple spreadsheets in the workbook?* Check along the bottom of the screen to see if there are multiple tabs (each representing a different spreadsheet). If so, click through to see what data is presented in each spreadsheet.

Once familiar with the data presented in the workbook, it is time to decide if the data needs to be formatted or cleaned in any way. The standard way to format data for analysis is to list the variables across the top of the spreadsheet in Row 1. Each observation is then listed in each row. This means that data presented in a crosstab format will need to be reformatted for analysis.

Incorrect format: Country name in row, year in column

	A	B	C	D	E	F
1	Country Name	1990	1991	1992	1993	1994
2	Cambodia	41310000	89710000	200520000	300510000	313350000
3	Vietnam	180550000	227300000	565220000	251610000	903300000
4	Lao PDR	149070000	139750000	162640000	200030000	214200000
5	Myanmar	160770000	177010000	113980000	100150000	167070000
6	Thailand	795580000	713560000	737820000	580690000	576600000

Correct format: Variable names in Row 1

	A	B	C
1	Year	Country	ODA
2	1990	Cambodia	41310000
3	1990	Vietnam	180550000
4	1990	Laos	149070000
5	1990	Myanmar	160770000
6	1990	Thailand	795580000
7	1990	Total	1327280000
8	1991	Cambodia	89710000
9	1991	Vietnam	227300000
10	1991	Laos	139750000
11	1991	Myanmar	177010000
12	1991	Thailand	713560000
13	1991	Total	1347330000
14	1992	Cambodia	200520000
15	1992	Vietnam	565220000
16	1992	Laos	162640000

Managing Data in Excel

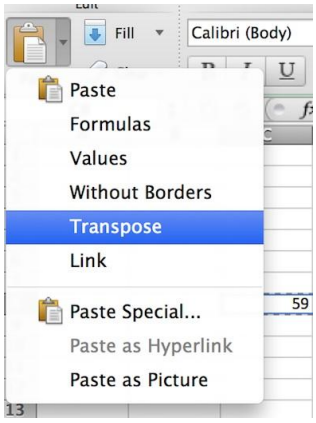
Import data

To import data that is not currently saved in Excel format, open up Excel and navigate to "File" and choose "Import". From there, select the file format of the data source you would like to open in Excel. If it is a text file and each data value is separated by a common character (this is called a delimiter: it could be a comma, colon, period, etc.), you can define the delimiter when importing the data and Excel will separate the data values into individual cells.

Similarly, if you open data in Excel and there are multiple data values within a single cell that are separated by a delimiter, go to the "Data" tab and choose "Text to columns". This command can distribute the contents of a single cell across multiple cells.

Format data

If the data is not formatted correctly (if it does not have the variable names in Row 1 and only one observation per row), it is necessary to reformat the data before proceeding with analysis. Helpful tips for formatting data:



- Start a new sheet with variable names in Row 1. Data will be copy and pasted to this new sheet.
- To repeat information in multiple cells (such as a country name), type the information in a cell. Hover over the cell until the small black square appears in the bottom right corner. Drag the square down to cover the cells where the information is to be repeated.
- If reformatting data in a crosstab format, the paste command "transpose" will transfer the data from horizontal to vertical, or vice versa.

Putting data into tables

You will have some additional controls over data if you put it into a table (such as sorting and filtering as described below). To put data into a table, select the area of the data and click on the "Table" tab and select "New". You will see that your variable names are now treated as table headers.

Sort data

Select a column of alphanumeric data in a range of cells, or make sure that the active cell is in a table column containing alphanumeric data. On the Data tab, in the Sort & Filter group, do one of the following:

1) To sort in ascending alphanumeric order, click Sort A to Z.
2) To sort in descending alphanumeric order, click Sort Z to A.

Creating filters

When you put your data in a table, filtering controls are added to the table headers automatically. For quick filtering:

1)Click the arrow in the table header of the column you want to filter.
2) In the list of text or numbers, uncheck the **(Select All)** box at the top of the list, and then check the boxes of the items you want to show in your table.
3) Click **OK**.

Pivot tables in Excel

Creating pivot tables in Excel is a great way to analyze and summarize large data sets. To create a pivot table, highlight the data to be included in the pivot table and go to the "Insert" tab. Select "Pivot table". This will bring up a dialogue box - click OK. The new pivot table will appear on a new worksheet. On the right hand side of the worksheet will be a window to control what variables appear in the pivot table. When desired variables are checked, they will appear in the pivot table and Excel will automatically sum the totals in each category. In order to change from sum to some over measure (such as average), right click on the column header "Sum of [variable name]" and select "Field Settings." Options will be given to display the sum, count, average, standard deviation, product, and variance.[8]

Formulas in Excel

A formula performs calculations or other actions on the data in your worksheet. A formula always starts with an equal sign (=), which can be followed by numbers, math operators (such as a plus or minus sign), and functions, which can really expand the power of a formula.

[8] More resources on creating pivot tables can be found at Office Support

Exploring and Managing Data with R

What is R and RStudio?
R is a language and environment for statistical computing and graphics. R can be considered a different implementation of S, a statistical programming language for organizing, visualizing, and analyzing data.

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian/Ubuntu, RedHat/CentOS, and SUSE Linux).

When do we use R?
R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and more) and graphical techniques, and is highly extensible (users and developers can expand R and add to its capabilities). The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

For detailed instructions on how to explore data with R, please refer to Appendix A.

CHAPTER V

Data Storage

"Metadata is key to ensuring that resources will survive and continue to be accessible into the future."[9]

Metadata

What is Metadata?

Often referred to as "information about information," metadata is a description of data. Generally it includes answers to the most basic questions about data: how was this data collected (the means of collection)? When was this data collected? What was the purpose of collection? Who was the creator of this data? What can be said about the quality of the data? Are there any peculiar characteristics to this data?

Types of Metadata

Structural Metadata: describes how the objects that compose data are put together. For example, how pages are ordered to form chapters in a book.

Descriptive Metadata: "describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords."[10]

Administrative Metadata: it refers to the aspects of data that concern data management, such as when was it created, the types of copyrights and licensing and how to archive it and process it.

[9] Niso. Understanding Metadata. http://www.niso.org/publications/press/UnderstandingMetadata.pdf
[10] Niso. Understanding Metadata. http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Data Warehousing

What is Data Warehousing?
It is a system used for storing data for reporting and analysis. They are compilations of one or more data sources in which both historical and current data is stored. It provides an opportunity to aggregate data from different sources, improving data quality. Data warehousing originated from various types of enterprises compiling data from their different divisions, but is now used by any organization, company, or individual who has large amounts of data to organize and store. Data warehousing was invented by William H. Inmon, "who first described a data warehouse as being a subject-oriented, integrated, time-variant and nonvolatile collection of data that supports management's decision-making process."[11]

CKAN

What is CKAN?
CKAN is a data management tool developed and managed by the Open Knowledge Foundation to make data accessible. It provides a platform for sharing, publishing, finding and using data. CKAN offers a possibility to warehouse data and make it open to the public according to the standards of open data.

Who uses CKAN?
CKAN is open sourced, which means it is used by both governments and non-profit organizations who would like to make their data open, public, accessible and shareable. CKAN makes it easier for data communities to share and collaborate with each other.

Datahub
Datahub is a data management platform based on CKAN that provides "free access to many of CKAN's core features, letting you search for data, register published datasets, create and manage groups of datasets, and get updates from datasets and groups you're interested in."[12]

[11] Data warehouse definition. http://searchsqlserver.techtarget.com/definition/data-warehouse
[12] About datahub. http://datahub.io/about

CHAPTER VI

Narrating Data

Data narration is more than simply describing data. It is user-centric through its focus on addressing the main questions of the audience and presenting information in a way that is easy to comprehend. Beyond that, it leaves an impression on readers by presenting information that shows them a pattern they may not have known or noticed on their own.

There are a few key questions to ask yourself as you start to narrate data:

1. Who is your audience?
2. What type of information are they seeking?
3. What are the main questions they are seeking to answer?
4. What narration formats (reports, tables, charts, infographics, numbers, statistics, verbal explanations) are they familiar with?

In order to illustrate an example of narrating data, this section will go through the process of writing the Donor and Development Assistance (DDA) landing page for the Open Development Mekong (ODM) website. Keep in mind that each narration you are designing is unique - this is only one example of approaching data narration.

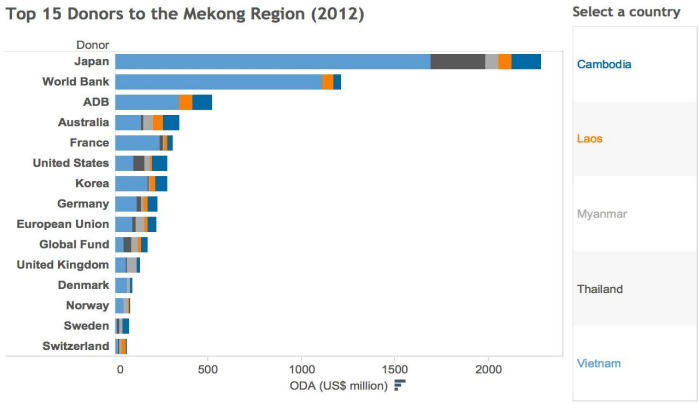
Steps to creating the DDA landing page:

1. *Define the audience*. Initially, the landing page will provide basic information about major donors and trends in development assistance. The audience will likely be development professionals who are interested in learning about donor trends over time.

2. *Determine what information the audience is seeking*. Think about why your target audience has navigated to your page. What information are they seeking? Defining the key questions will guarantee that only the most important and relevant information is presented and all major points are covered in the overview. Answer these questions in the overview in order of importance to the audience. We defined the key questions for the DDA landing page as:
    a. *What projects are currently receiving the most funding? Why?*
    b. *Who are the biggest donors to the region?*
    c. *How important is donor assistance to countries' economies?*
    d. *How is ODA changing over time?*

3. *Research the answers to the questions.* Where you research will depend on your research question and the questions of your audience. Look for reliable sources: avoid advocacy groups that may provide biased commentary and find sources that provide up-to-date data and data that is updated regularly. (See *Research 101* section for research tips).

4. *Present the answers with a narrative format familiar to the reader.* Think of how each question is best answered. Should it simply be a text explanation, or should visuals be used? What can be used to give numbers perspective and help the audience grasp the magnitude? When the answer requires context and explanation, a text answer may be best. For example:
    - "*What projects are currently receiving the most funding? Why?"* is answered through text. It requires some explanation of regional context.
    - "*Who are the biggest donors"* is answered both with bullet points *and* a Tableau visualization.[13] The bullet points present the answer at a glance for readers in a hurry who are quickly scanning the page, while the Tableau visualization allows for more interaction. Users can make comparisons between donors and countries.

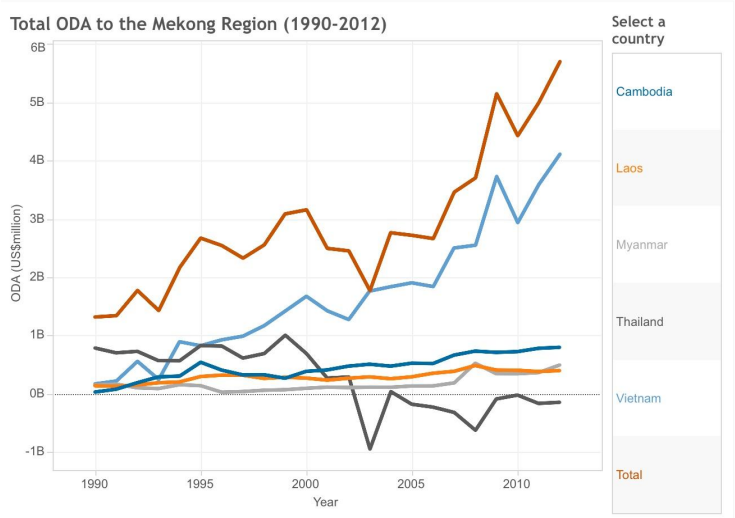


    - "*How important is donor assistance to countries' economies?"* is answered with numbers as the percentage of a country's GNI that comes from official development assistance. This puts assistance in terms of the broader economy. The ratio is given for each of the countries in the Mekong region and a comparison to the average ratio of ODA to GNI in Africa was given as well.

> **ODA flows contribute to the GNI of the nations in the Mekong region. In 2012, ODA comprised 6.16% of Cambodia's GNI, which classified it as having "medium aid dependency" on foreign aid.[4] Laos received 4.87% of its GNI from ODA, while Vietnam received 2.98%. Thailand is now repaying more of its ODA than it is receiving.[5] For comparison, the average ratio of ODA to GNI in Africa was 2.4% in 2010.[6]**

[13] View the interactive donor comparison visualization here

- "*How is ODA changing over time*" is answered with a time series graph.[14] This visualization is well suited for clearly illustrating trends over time and showing when countries start receiving more aid, or see a reduction in aid. Listing the years, amounts of aid, and countries would be cumbersome, but the visualization is clear. Explanations for certain trends are provided through text for added analysis.



In total, the overview page was only 500-600 words. Users can go deeper into the section to find more information, but the landing page provides only the most important information. The audiences' most pressing questions are answered with bullet points, numbered lists, and visualizations to draw the eye.

[14] View the interactive time series visualization here

Donor Profiles

The Donor and Development Assistance section will include profiles of the largest donors. Donor profiles provide an overview of the major multilateral and bilateral donor agencies working in the Mekong region. The profiles provide key information about the largest donors to the region: a brief background, focus sectors, major projects, regional and country-level strategy, and common critiques of the donor. These profiles can be longer than the landing page and go into more detail (1000-1500 words in length).

I. General overview
   A. Multilateral or bilateral donor? If multilateral, who are the members?
   B. When was the agency founded?
   C. What is their mission statement?
   D. Who does the agency give to?
   E. What forms of aid does the agency provide?
   F. What is the total funding provided by the agency?

II. External evaluations of the agency: what are others saying about the agency? What are their strengths and weaknesses? Be objective.

III. Mekong Region Overview: does the agency have a regional strategy?

IV. Country strategies: for each country, describe what the major sectors are that are supported by the donor, key projects, total funding, and projections for funding over time. Highlight any projects done by the agency that have received media attention, for either positive or negative reasons.

In addition to text, be sure to include visualizations and/or tables to help summarize the data. A table with 3-4 key parameters is a good way to compare donor activity across countries. Visualizations can help present numerical data. Infographics can help illustrate key processes of donor activity (such as the process of project design and implementation). Be creative with visualizations!

The SIPA Team completed four donor profiles. Profiles should be written for the ten largest donors:

✓ Japan (JICA)
✓World Bank
✓Asian Development Bank
✓Australia
❑ China
❑ France
❑ United States (USAID)
❑ Korea
❑ Germany
❑ European Union

Country	ODA from WB 2014	Sectoral Focus	WB Income Classification	Context
Cambodia	$67 million	Health Education Trade Development Water and Sanitation Teacher Quality Improvement Social Protection	Low	The World Bank froze lending to Cambodia in mid-2011, 2012 and 2013 following the refusal of the Cambodian government to distribute land titles after forced eviction for Boeung Kak lake residents surrounding a WB funded development project.
Laos	$110 million	Finance, Energy Natural Resource Management Poverty Reduction	Lower Middle	Laos is considered the battery of the Mekong. Natural resources and the energy derived from it accounts for over half of the country's wealth.
Myanmar	$282 million	Health Education Energy, Agriculture Water Resource Management Social Protection Telecommunications	Low	Myanmar is in a "triple transition" - political, economic, and social. It is transitioning from an authoritarian military regime to a democracy, from a centralized to a market-based economy, and from 60 years of conflict to peace.[4]
Thailand	$28 million	Energy Rural Development	Upper Middle	Thailand "has recently changed from borrower-lender to knowledge partners." It is a success story although it is the only country not to have been colonized, and has slower growth than others.
Vietnam	$1,551 million	Environmental Sustainability, Social Equity Macroeconomic Stability Anti-corruption	Lower Middle	Vietnam could be the next Lower Mekong country to pull itself into the Upper Middle income category, and join the ranks of the "developed world."

CHAPTER VII

Visualizing Data

Now that we have learned how to source, scrap, explore, manage, store and narrate, we can use our data to create visualizations.

What is data visualization?

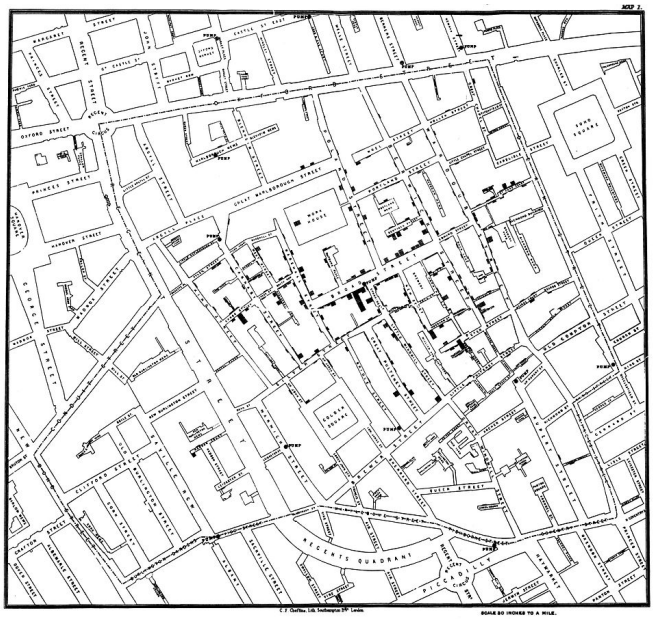
Data visualization is a way of communicating information visually that shows patterns, trends, or correlations. It displays highly complex data and information in a way that is quick and easy to understand. It also communicates information that may have been overlooked or under-lighted in text. Often, it is generated to be able to be understood by the broader public. It is broadly used in digital and data journalism.

Interactive data visualizations, such as Tableau and CartoDB softwares, give the user the capability to filter data in a way that is of most interest to them. This is extremely useful because the audience has more freedom to include or exclude data from very large datasets made available to the general public. As seen in the following famous John Snow visualization, excluding less relevant information can declutter the drawing and make the patterns more clearly visible.

Origins of Data Visualization

John Snow is credited with creating the first data visualization.

In 1854, John Snow created a visualization using a map designed by cartographer Charles Cheffins. Following a cholera outbreak in 1854, Snow and others talked to building residents and Snow mapped deaths in the Soho neighborhood of London. He drew bars inside on top of each building's location; each bar represented one death. From this map and the information gathered to create this map, he showed Broad Street pump to be the culprit of the outbreak in a way that was easy for people to understand.



Right: 1854 Broad Street pump, London

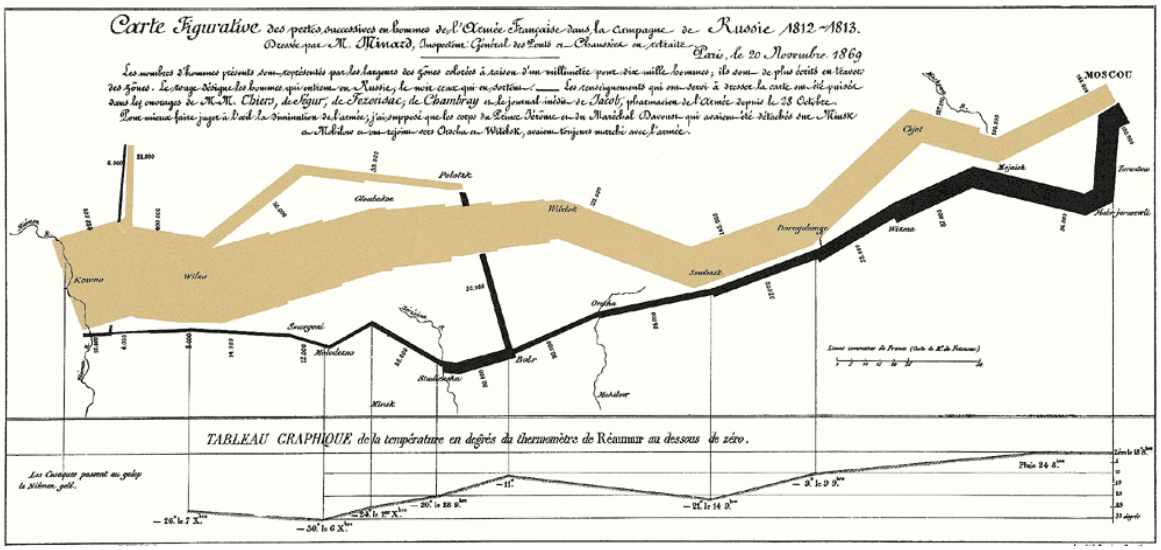
Visualizations can be created two ways: 1) drawings that seek to show an answer to a question with the answer already clearly defined, and 2) drawings that seek to draw associations and test theories and assumptions with no answer yet clearly defined. In this visualization, Snow was using the former method by trying to prove the "answer" to the origins of the deaths that cholera was water-borne.

He included deaths and water pump locations, and excluded other unnecessary information such as bakeries, hospitals, parks or numbers of residents, and horses, etc.

Snow was not the first one to create this map, but he is credited as being the first one to create this map that was *used to tell a story*.

Charles Minard created one of the most famous visualizations.

Below is a design of **Napoléon Bonaparte's famous march** to Moscow in 1812-1813 designed by Charles Minard in 1869. He started with 400,000 troops and returned with 10,000.[15]

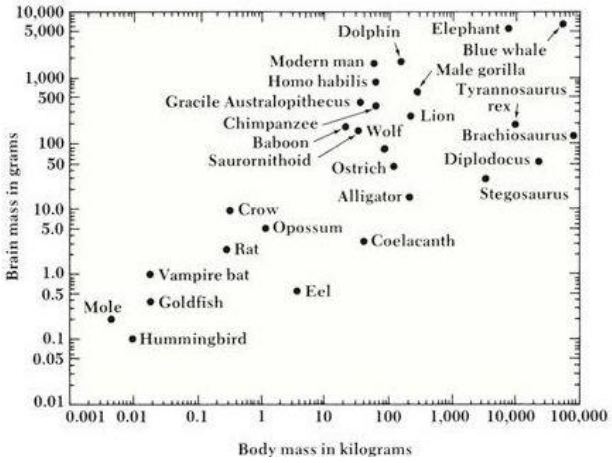


[15] This visualization, "Napoleon's March," was taken from Edward Tufte's website: http://www.edwardtufte.com/tufte/minard

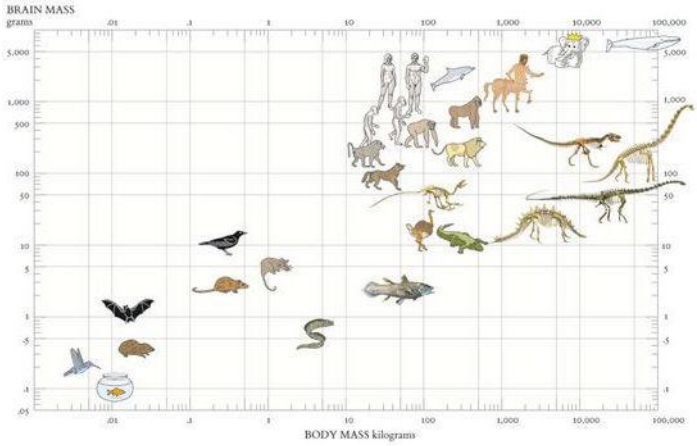
Edward Tufte is known as a pioneer in data visualization.

Edward Tufte is an American statistician born 1942. He is Professor Emeritus in Political Science, Statistics, and Computer Science at Yale University. He is known for his contributions to information and visual literacy. He first published *The Visual Display of Quantitative Information* in 1983, and has published four in a series of five books, the latest is *Beautiful Evidence*. He has been called the "da Vinci of Data" and the "Galileo of Graphics."[16] Below is a small *before* and *after* example of what smart visualization can do, the *before* is of Carl Sagan's original diagram, and the *after* is the one designed by Tufte.[17]

Before



After

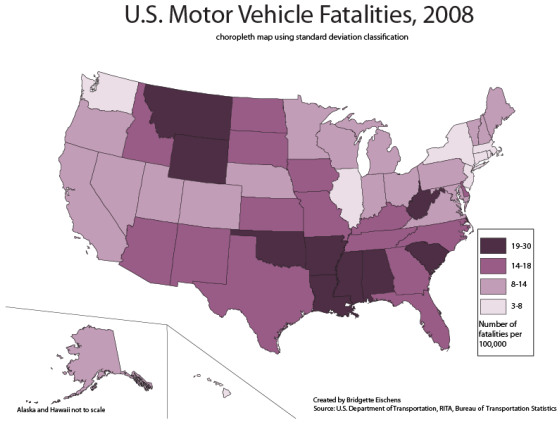


[16] "The da Vinci of Data," New York Times (March 30, 1998). http://www.nytimes.com/1998/03/30/business/the-da-vinci-of-data.html and Adam Aston's "Tufte's Invisible Yet Ubiquitous Influence," Bloomberg BusinessWeek (June 10, 2009). http://www.bloomberg.com/bw/stories/2009-06-10/tuftes-invisible-yet-ubiquitous-influencebusinessweek-business-news-stock-market-and-financial-advice

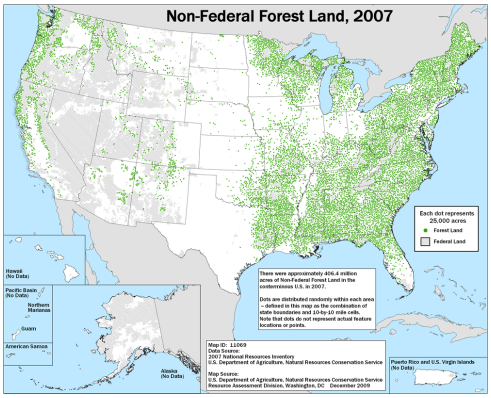
[17] Bloomberg, "The Vision of Edward Tufte." http://www.bloomberg.com/ss/09/06/0608_tufte/3.htm .

Types of Data Visualizations

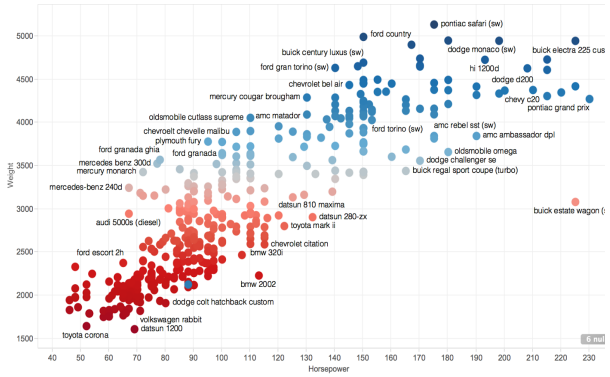
Choropleth



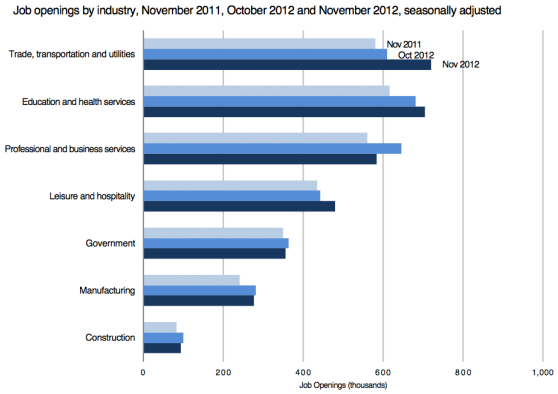
Dot distribution



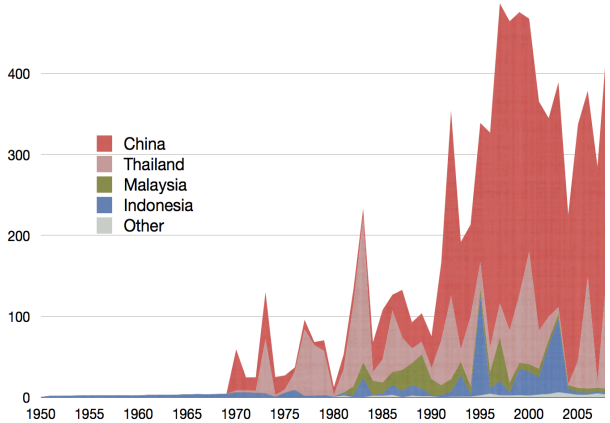
Scatterplot



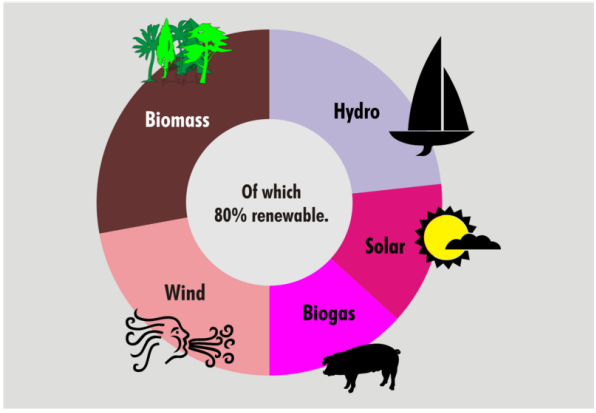
Bar graph



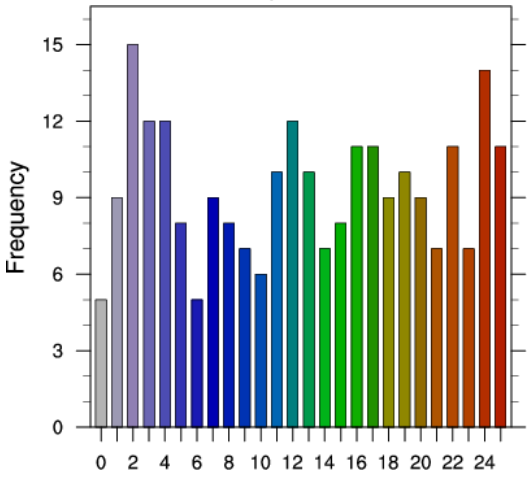
Time series



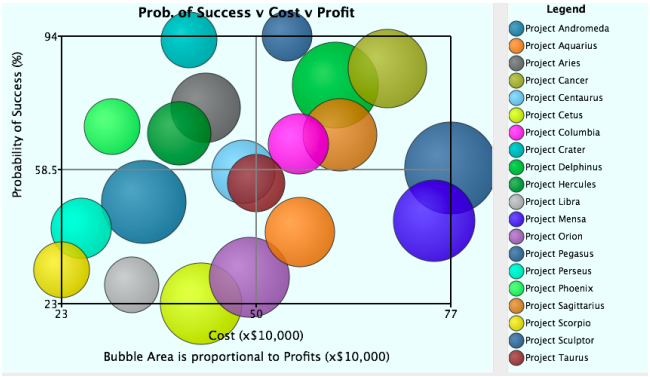
Pie Chart



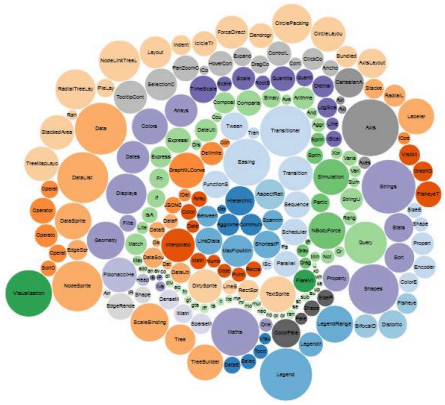
Histogram



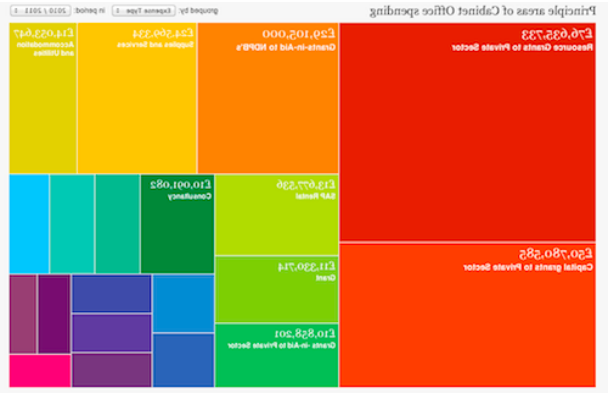
Bubble chart



Bubble cloud



Tree map[18]



The following six pages are designed to help you explore data visualization techniques and create visualizations of your own by using available free or open-source software.

[18]Choropleth. http://www.geobecks.net/myp-humanities/year-9/understanding-hazards/003---choropleth-mapping
Dot distribution. http://www.nrcs.usda.gov/wps/portal/nrcs/detail/sc/technical/dma/nri/?cid=stelprdb1083124
Scatterpolot. https://datagab.files.wordpress.com/2014/09/screen-shot-2014-09-12-at-4-19-42-pm.png
Bar Graph. http://thewhyaxis.info/content/42-defaults/bls-in2.gif
Time Series. http://commons.wikimedia.org/wiki/File:Jellyfish_production_time_series.png
Pie Chart. http://en.wikipedia.org/wiki/File:Example_of_a_doughnut_chart.png
Histogram. https://www.ncl.ucar.edu/Applications/histo.shtml
Bubble chart. http://www.bubblechartpro.com/wp-content/uploads/2012/04/Bubble_Chart_25_Example.png
Bubble Cloud. ttp://www.govhack.org/wp-content/uploads/How-to-participate-in-GovHack_html_m90d8020.jpg
Tree map. http://www.networkworld.com/article/2222214/opensource-subnet/miso-project-offers-open-source-tools-for-data-visualization.html

Excel

What is Excel[19]

Excel is a Microsoft spreadsheet application, one of the most powerful and widely-used data processing software. With Excel, you will be able to create tables, format tables, sort and filter table data, and use formulas with tables. This tool enables you to manipulate and visualize data in a simple, handy manner.

Who uses Excel

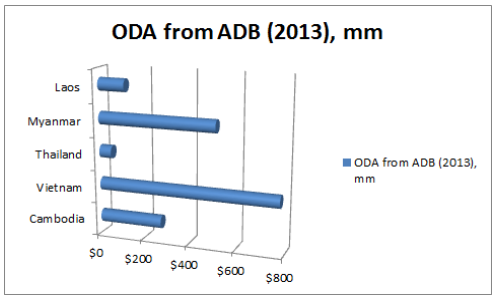
Excel is used by students, scholars, businesses, citizens etc. to track the data trends, perform calculations and build data models. It is often installed by default as part of the Microsoft Office Suite on computers. Excel workbooks contain individual worksheets, which you can use to create lists and spreadsheets. More advanced applications of Excel include creating pivot tables (see *Exploring and Managing Data with Excel* section) and generating basic visuals.

When to use Excel

Excel is widely used. People have used Excel for all sorts of purposes, from financial modeling in big companies to serious statistical analysis. In financial industry, Excel's ability to create new spreadsheets where users can define and apply formulas to extend the model and create forecasts makes it highly valuable. Excel is also frequently used for common information organization and tracking such as a list of staff names, project status reports, sales records, and invoicing. Additionally, Excel is a useful tool for scientific and statistical analysis for larger datasets. Excel's statistical formulas and graphing can help researchers perform variance analysis, correlation and regressions, chi-square testing, and more.[20]

Data visualization with Excel

Excel allows users to visualize the data currently stored in the worksheet into different types of graphics. By combining Excel formulas and visualization, users can even create interactive data visualizations with this tool. While making charts in Excel is fairly easy, most of the time we use Excel visualizations to explore and discover trends in the data. For commonly used, and more complex data visualization tools, please refer to the open-source software that are mentioned below.



Excel enables you to create different data visualization models including bar charts, pie charts, line charts, scatter plots etc, all in 2-D or 3-D format. Here we provide an example of how to create a bar chart. The steps are as follows: 1) Select the data that you want to plot in the bar chart. 2)On the **Insert** tab, in the **Charts** group, click bar type you would like to create. 3) Click the chart area of the chart. This displays the **Chart Tools**, adding the **Design**, **Layout**, and **Format** tabs. 4)Right-clicking the chart area, you will be able to change the chart type, re-select data and move the chart. Below is an example of bar-chart visualization using Excel.

[19]For additional resources, please visit Microsoft Office Support: https://support.office.com/
[20] "What is Excel Used For", http://www.opengatesw.net/ms-excel-tutorials/What-is-Excel-Used-For.htm

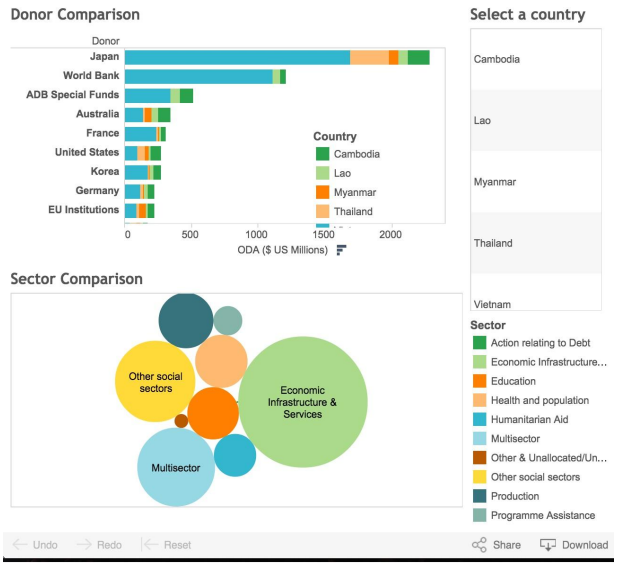
Tableau Public

What is Tableau Public?

Tableau is an interactive data visualization software. A free version of their program is available through Tableau Public, which is available for organizations to use as an introductory service.[21] With Tableau, you can connect to data, create visualizations such as bar charts, line charts, pie graphs, circle charts, histograms, scatterplots, maps, tree maps, and more, and then combine multiple visualizations into an interactive dashboard. Once your data is on the web through Tableau Public, anyone can download it and create their own visualizations from it.

Who uses Tableau Public?

Tableau is great because you do not need any programming knowledge to use it. It is used by writers, bloggers, students, professors, critics, citizens and more. Content created with Tableau Public can be embedded directly into websites, or shared via links.



When to use Tableau Public

Tableau is best used with quantitative data in order to visualize relationships in interesting ways. It is great for comparisons and allows users to filter results on certain characteristics. It is useful for when multiple variables can be compared simultaneously since you can combine two or more visualizations into a dashboard for greater interactivity. You can also create "Story Points" with Tableau to guide users through a narrative to help reveal a certain pattern or information you would like to highlight.

Getting started with Tableau Public

The very first step is to format your data correctly. Tableau needs raw data. This means the very first row of your Excel sheet has the column headers (these will be the variable names), each subsequent row is one observation, and the data begins in cell A1 (no titles, descriptions, etc). Also make sure to take out rows with totals – this is not a data point and Tableau can easily create totals, averages, and more once the data is uploaded.

Once the data is formatted, you can upload your data to Tableau. You will notice that Tableau automatically classifies variables as either *dimensions* or *measures*. A *dimension* is a qualitative, categorical, or independent variable. Examples of dimensions are country, province, year, or product title. A *measure* is a quantitative or dependent variable, such as official development assistance received, sales, latitude and longitude, number of projects. Drag and drop dimensions and measures into the workspace to begin creating visualizations. The *Show Me* tab will display options for visualizations that are possible using the selected variables.

[21] Tableau Public gives organizations the opportunity to explore its capabilities. If an organization wants to put data online for the public, please contact Tableau Public at info-public@tableausoftware.com to discuss a commercial relationship.

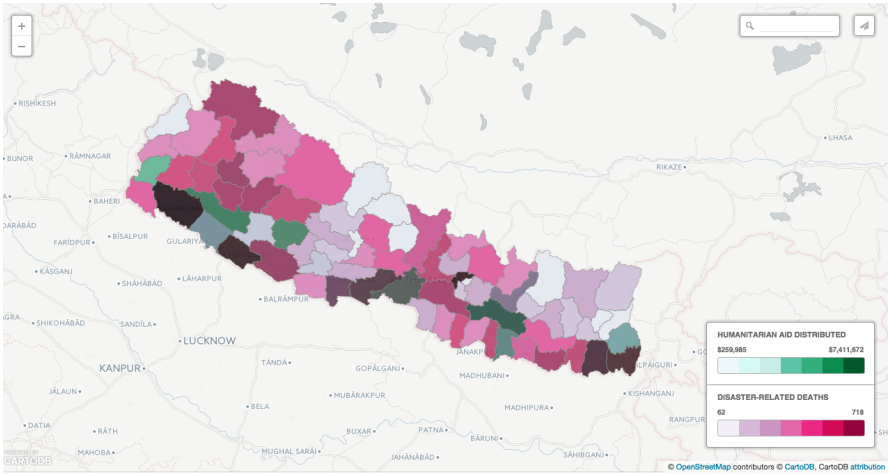
CartoDB

What is CartoDB?

CartoDB is interactive mapping software. *Carto* from "cartography" – the study and practice of map-making, and *DB* from "database." CartoDB allows users to upload data, and use others' datasets to map trends geographically. With CartoDB, you can connect to data, create visualizations and then combine multiple tables and maps into an interactive dashboard that lets users filter across variables. Once your data is shared with the web, anyone can interact with the data, download the dashboard, and create their own visualizations from it.

Who uses CartoDB?

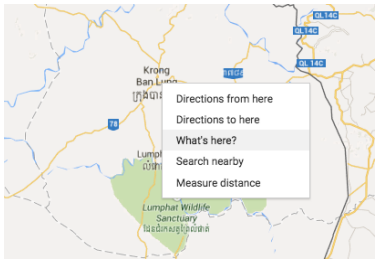
It is used by journalists, academics, government officials, NGOs, and more. Content created with CartoDB can be embedded directly into websites, or shared via links.

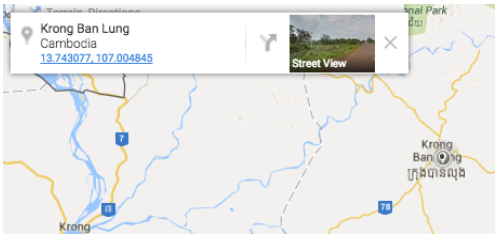


To the right is a map of Disaster-Related Deaths and Humanitarian Aid Received, mapped by district in Nepal in February 2015. Each color represents a variable and the opacity represents the severity.

How to use CartoDB

CartoDB uses Excel to upload its data. If you do not have your latitude and longitude points already, you can use CartoDB's data based on cities, towns, etc. or you can use GoogleMaps; the latter tends to give more precise locations. To do this, type in your point on GoogleMaps, and right click on the drop down menu to "What's Here?" This will drop a pin into your map. The latitude (first) and longitude (second) will appear in blue under your search. Use those to create columns for your geo points. You can upload data by clicking on "new table." After you upload your data, there are two views, the Table view and the Map view that allow users to alter their data in real time, and customize their own maps.





Getting Started

You can sign up for a free CartoDB account here, which allows you to upload 75.0 megabytes of your own data. When you sign up for an account, CartoDB also allows you to use public data to explore and design your own maps using chloropleths, categories, and more. You can publish your designs to the web.

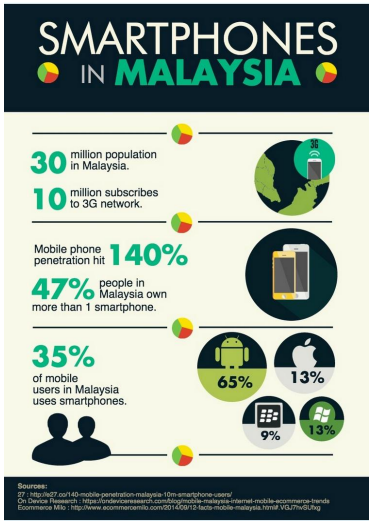
Piktochart

What is Piktochart?

Piktochart is an online tool for creating graphics. With the website, you can easily build infographics, reports, posters, and simple data visualizations that look stylish and modern. The user can easily combine text, graphics, pictures, and charts within a single visual to display information. Users can customize a template or theme, or they can create their own visual from scratch. The website has an easy-to-use drag and drop interface for editing. The visual can be embedded with HTML (allowing for clickable elements), or saved in PDF, PNG, or JPEG formats.

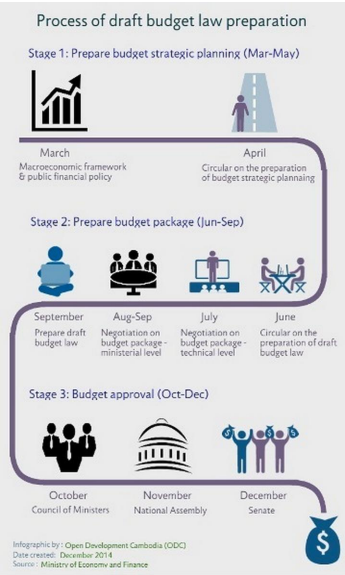
Who uses Piktochart?

Piktochart can be used by anyone with any level of graphic design experience. Users can easily produce professional-grade infographics with only a few clicks. Beautiful infographics can be produced without a graphic designer. Forbes Magazine named Piktochart the infographic tool of choice for the "graphically challenged" or for those who need to produce a graphic quickly.



When to use Piktochart

Since text, images, graphics, and charts can be combined within a single visual, Piktochart is ideal for producing infographics. Infographics display information, data, or knowledge quickly and clearly. Infographics are great when you have multiple key facts about a single topic; by combining those facts through charts and text in a single visual, the reader is given a complete picture of a complex topic.



Getting started with Piktochart

Users can sign up for a free account with Piktochart online at www.magic.piktochart.com[22] Once registered for an account, the easiest place to begin is to choose a pre-made infographic template. After opening the template, explore the features of customizing text, adding pictures, choosing illustrated graphics to incorporate and manipulating the size and color of elements. Once familiar with the interface, you can create beautiful infographics, reports, and presentations from scratch.

Piktochart can produce some simple data visualizations through bar, line, area, dot, pie, stacked venn, doughnut, progress bar, and bubble charts. Different types of charts can be used within a single visual to illustrate information in various ways. Try presenting data with at least one or two types of charts along with text to highlight key statistics or to briefly explain the information presented in the charts. Keep your visuals simple and clean - the goal is to present complex information in a simple way.

[22] Nonprofits and social enterprises can register for a PRO account for $39.99 USD per year (go to http://piktochart.com/pricing/nonprofit/ to request consideration for nonprofit pricing.)

HighCharts Cloud

What is HighCharts?
HighCharts cloud is an online visualization software that allows you to create data visualization using the templates available. The templates are not very flexible but they allow users to create, edit and share basic charts. High Software has a more complicated software that must be licensed.

Who uses HighCharts Cloud?
HighCharts Cloud is specially designed for "non-techies" which means that you do not have to know how to write code and is easily shareable on blogs, websites and social media. They are served in Amazon.

Getting Started with HighCharts Cloud:
Visit: https://cloud.highcharts.com/ and click on "Get Started Now". Then, drag your data in .csv or .xls format.

Click on "continue to template" to pick the appropriate chart for your data.

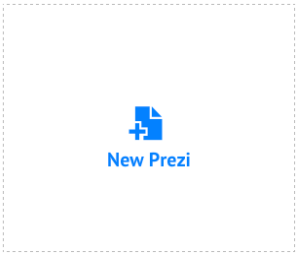
Then customize it by adding a title, axes titles, changing the appearance, editing the label, changing the tooltip, etc.:

Click on share to download your chart in the format you want or share it on social media, or to get the code to embed it on a website:

Prezi

Why and When to use Prezi?

Prezi is a cloud based presentation software, which means that you would be able to access your presentation from any computer connected to the internet. Prezi is very easy to use and helps you create engaging presentations.



Creating a Prezi

To start using Prezi, you need to create an account first. Then, click New Prezi. You can start from a blank presentation or use a one of the preloaded templates to create your presentation. Prezi works like a canvas, you can add content and navigate through it using the zoom buttons located in the right-hand side of your Prezi.

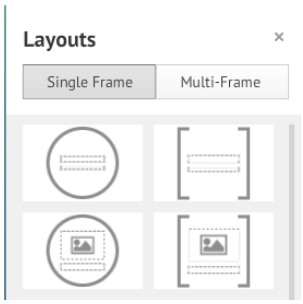
Text and images

To add text, you can click anywhere on your canvas and start typing, on top og your text you will see a formatting bar which will allow you to change the font, the size or the style of the text.

To add images, right click on the insert button in the top menu of prezi, and select what you would like to add to your presentation, it gives you the option of adding images, symbols and shapes, a youtube video or content uploaded from your own computer.

Layouts

To organize the design of the presentation, Prezi has the option of inserting a layout: click on the insert button in the top menu and pick from a single-frame or a multiframe and drag to the canvas.



Borrowing Content

Prezi allows you to take content from other presentations that are "reusable". To do this, just click "save a copy" to save it in your own library, from there you can use it to update the content.

Appendix A: Exploring Data with R[23]

Getting to know RStudio

- Console → The space where you type your command and run to see the results. Commands in the console cannot be directly saved.
- Script → A text file that enables you to save and run R commands.
- Environment → Environment stores the data frames and values you are currently working with on RStudio.
- Help, plot → Help of the commands, plots of the data, list of packages etc.

Working Directories

It is crucial to get to know where your working directory locates and set the correct working directory to the file that contains your datasets before importing data. If you are working with the wrong working directory, RStudio will not be able to locate your datasets.

Relevant commands:

- Getting your current working directory: getwd()
- Setting working directory: setwd(), alternatively you can go to session → set working directory and choose directory.

Always remember to hit the "run" button to see the results of your code!

Libraries

As a programming language, R allows us to write and share libraries (or more intuitively, packages) that manipulate data. For example, the "plyr" package provides some easy codes to aggregate data, and the "ggplot2" package has become the most commonly used one in visualizing data with R. To install a package, type install.packages ("name_of_the_package") to get RStudio download the package, and use library (name_of_the_package) to load the package every time before using it (You only have to install the package once though).

Getting help

RStudio is extremely powerful at providing help on how to use commands and what will the proper commands be when you look for them. When trying to get help on how to use a certain command, say, setwd(), type ?setwd(). If you do not know the name of the command and would like to look it up in RStudio, use two question marks followed by the content of your intended commands such as ??standarddeviation (to get the command of calculating standard deviation) and let RStudio help you.

[23] Structure of this part is taken from the course notes of *Data Visualization* taught by Dr. Eliot Cohen at Columbia University. Github page: https://github.com/Ecohen4/data-viz

Importing Data

- csv. → The most commonly used command is read.csv("name_of_the_file.csv"). Please note that csv. is always our preferred data format to work with. It is open source, and takes a much smaller space on your computer thus quicker and easier to proceed with.
- txt. → The most commonly used command is read.table ("name_of_the_file.txt", header=FALSE). More detailed instructions, see the help in RStudio. Different from read.csv, read.table expects no header of the data and blank space (instead of comma) to separate the characters.
- xlsx. → R does not have built-in commands to input an xlsx file. Instead, we install a package "xlsx" and use read.xlsx to read the file. Please note that you probably will have to update your java on your computer before successfully using this package. It is preferable to save the xlsx. into a csv. file and then import it.
- scan directly from a website → the scan() function allows us to directly read data from a website or text file. Try ?scan() for more information on how to use this command.

Data Structures

In below sections we will be using RStudio's built-in data frame, mtcars, to practice with the commands. Type mtcars in your script, hit "run", and we will be able to see the data appearing in the console. Before moving to data structures, we need to learn some basic concepts about R.

What is a vector: A vector is a sequence of data elements of the same basic type. A column vector appears as a column within a data frame.

What is a data frame: A data frame is a list of same-length column vectors. It looks like a matrix where each column represents a measurement, and each row represents an observation. A data frame can contain vectors of different data types.

Assigning values to an object. For example we have our mtcars data frame and we would like to create a new object "sample_data" that has exactly the same data as mtcars, we use the left arrow. sample_data <- mtcars. Assigning values to an object become especially useful when you have long names of your original data sets, and a good practice for you to observe your data frame step by step alongside with your commands.

How is the data structured? (structure)

To take a look at the structure of data, here are several commonly used commands:

sample_data <- mtcars
head (sample_data) shows us the first six rows of the data frame.
tail (sample_data) shows us the last six rows of the data frame.
str (sample_data) presents us with a more comprehensive view of the data frame that includes the number of rows and columns of the data, and the types of the vectors.

```
str(sample_data)
'data.frame': 32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3
24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62
95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76
3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
```

What type of data is it? (class)
The class () command helps us to look at the data type, or the class of the elements.

```
> class (50)
## [1] "numeric"
> class (FALSE)
## [1] "logical"
> class ("SIPA")
## [1] "character"
```

There are nine types of data that are commonly used in R:

- numeric
- integer
- complex (defined via the pure imaginary value *i*)
- logical

- list (generic vector containing other objects)
- raw (hold raw bytes)
- expression
- factor (Factors in R are stored as a vector of integer values with a corresponding set of character values to use when the factor is displayed. Both numeric and character variables can be made into factors, but a factor's levels will always be character values. You can see the possible levels for a factor through the levels command.)

How big is the data? (dimensions)
We use dim(), nrow() and ncol() to check the size of our data frame. In our sample data mtcars, we have 32 rows and 11 columns.

```
> dim(sample_data)
## [1] 32 11
> nrow(sample_data)
## [1] 32
> ncol(sample_data)
## [1] 11
```

What are the names of the column vectors?
We use the colnames function, or simply the names function.

```
> colnames(sample_data)
 ## [1] "mpg"  "cyl"  "disp" "hp"   "drat"
"wt"   "qsec" "vs"   "am"   "gear"
[11] "carb"
> names(sample_data)
 ## [1] "mpg"  "cyl"  "disp" "hp"   "drat"
"wt"   "qsec" "vs"   "am"   "gear"
[11] "carb"
```

How long is the period of record? (time series data)
For time series data, it is helpful to create a "Date" or POSIXct variable to check the range of the time frame. Date class helps us to look back and forth between Date formats, while POSIXct is used to track sub-daily time records. Try ?as.Date and ?POSIXct to see how to convert integers into a Date format. . You may refer to the range() function to check the range of the time series data.

Looking at the records (rows)
Let's create a small sample under our sample data.
small_sample <- sample_data[1:10,] (We take the first ten rows and all columns)

```
> small_sample<-sample_data [1:10,]
> small_sample[1,]
##           mpg cyl disp hp drat wt  qsec  vs am gear
carb
## Mazda RX4  21  6  160 110 3.9 2.62 16.46  0  1  4
    4
```

We looked at the first row of our small sample in this case.

Handling Real-World Data

Missing values
In real world, we inevitably meet data with missing values. Some in R are shown as "NA", "-999", "10e-30" or blank " ". It is good practice to use **NA** to record missing values for the sake of consistency. NA can be any type of value. While NA stands for "Not Available", NaN means "Not a Number". A numeric expression that is correct in syntax but has not mathematical meaning (not a real number) will return NaN. We use is.na to detect whether there are missing values in our data frame. By using sum(is.na(data.frame)), we summarize the total number of NAs in the data frame.

```
> sum(is.na(sample_data))
## [1] 0
```

The above function shows that we currently do not have any missing values in our data. But we also need to look at the summary of the data frame to see if the data make sense to us. The most commonly used commands are summary () and range (). Say I would like to look at the summary statistics about the disp vector in our sampe_data data frame (here the $ sign indicates that disp is one of the vectors of the data frame sample_data).

```
> summary(sample_data$disp)
 ##  Min. 1st Qu.  Median   Mean  3rd Qu.   Max.
 ## 71.1   120.8   196.3   230.7   326.0   472.0
> range(sample_data$disp)
## [1]  71.1 472.0
```

After taking a look at the summary of disp, we found that the largest number 472.0 does not make sense in real life (I am making this up here but the key point is to be able to benchmark your data through looking at summary statistics). I would like to record elements in disp that equals to 472 a missing value.

```
> sample_data$disp[sample_data$disp==472.0] <- NA
> sum(is.na(sample_data$disp))
## [1] 1
```

Now we have manually introduced one missing value by taking out the value that does not make sense to us. The command na.omit will omit the entire row if there is a NA.

- Summary statistics
  Except for the summary function we've encountered above, table is also a command used for summarizing your data. summary is used for numerical or boolean variables while table is used for character or categorical variables.
- Mean, Standard Deviation
  We will use the mean and sd commands to get the mean and standard deviation of vectors.
- Subset, merge
  Subsetting and merging are common techniques that we use to manipulate with data. We use the subset and merge command, respectively.

```
> small_group<-subset (sample_data,disp == 71.1)
> small_group
                mpg cyl disp hp drat    wt qsec vs am gear
carb
Toyota Corolla 33.9   4 71.1 65 4.22 1.835 19.9  1  1    4
```

Here the subset function takes out a small fraction of our sample data frame where column disp has the value of 71.1.

The merge function can merge two data frames by common columns or row names, or do other versions of database join operations. Let's see a simple example here. Say we have two data frames that have one column same:

```
> group.A
   country value
1 Cambodia     A
2     Laos     B
3 Thailand     C
4  Myanmar     D
5  Vietnam     E
> group.B
   country value2
1 Cambodia      F
2     Laos      G
3 Thailand      H
4  Myanmar      I
5  Vietnam      J
```

Now we would like to merge these two data frames by the country vector that they have in common. Note that after merging the two data frames, there remains only one "country" column.

```
> group.C<-merge(group.A,group.B, by =
"country")
> group.C
   country value value2
1 Cambodia     A      F
2     Laos     B      G
3  Myanmar     D      I
4 Thailand     C      H
5  Vietnam     E      J
```

Observing data through simple visualization
Sometimes by graphing the data, researchers can easily get an idea of how the data spreads, if there exists a linear relationship, or spot outliers at a glance. Using hist function, you can draw a histogram of your data in r and discover potential outliers. Using the ggplot2 package, researchers can go further with visualizing the data by coming up with different forms of charts, comparing multiple charts etc.

Writing Data
Note that although we have manipulated with the data in RStudio, the original (imported) data files will not be changed. If you have imported an external file, for example, a csv., txt., or xlsx., you can write the current R object to an external file, which is not done automatically in R. Use the write.csv command to save your work to a separate csv. Meanwhile, you can also use the save command or file → save/save as to save your file into an R object and reopen your work in R next time.

- Resources on learning R
    - Coursera
    - Codecademy
    - R project for statistical computing
    - R Tutor
    - Code School
    - Cookbook for R

Appendix B: Tableau Workshop Instructions[24]

Retrieving the data

1) http://www.oecd.org/dac/stats/aid-at-a-glance.htm
    a) Go to Interactive summary charts by aid (ODA) recipients.
    b) Download the underlying data as a csv. file.
    c) From the data, we would like to find out 1) who were the top ten donors of each of the five countries in the Mekong region in 2013, 2) what sectors received the most funding for each of the five countries in the Mekong region in 2013.

Filtering the Data: Donors

2) Go to "Data" tab, and click on "Filter". On the header of each column, drop down menus will appear.
3) Click on the drop down menu of "RecipientNameE" and choose the five countries in the Mekong region: Cambodia, Laos, Thailand, Myanmar, Vietnam.
4) Click on the drop down menu of the "DonorName" column (F) and deselect the "(blanks)" option to filter out the missing values.
5) Now we are able to get a complete set of filtered data with non-blank donors that applies to the five countries in the Mekong region. By selecting all of the current data (ctrl+A), copying (ctrl+C) and pasting the data (ctrl+V) to a new sheet, we are creating a new sheet that is free of the filters with only the necessary information we need. We rename this new sheet as "Donors".

Connecting to Tableau: "Donors" worksheet

6) Open Tableau Public and click on the orange button "Open Data".
7) Under "In a file", choose "Text File" that allows us to connect to cvs. files. Choose the csv. workbook to connect.
8) Drag the "Donors" sheet to the orange "*Drag tables here*" space.
9) Click the orange button "Go to Worksheet".
10) Here we will be able to see the data we connected to (the "Donor" sheet), Dimensions and Measures of the data.
    a) Dimensions: By default, Tableau treats any field containing qualitative, categorical information as a dimension. A dimension is a field that can be considered an independent variable.
    b) Measures: By default, Tableau treats any field containing numeric (quantitative) information as a measure. A measure is a field that is a dependent variable; that is, its value is a function of one or more dimensions.

[24] This workshop was presented by the SIPA team to the staff of Open Development Cambodia team on Tuesday, January 18th in the ODC offices in Phnom Penh, Cambodia. It was also given to PanNature and Open Development Vietnam on Wednesday, March 18th in Hanoi, Vietnam.

11) To answer the question of "who were the top ten donors (of Cambodia, Vietnam etc.)?", we drag "Amount" and "Donor Name E" to "Columns" and "Rows" shelves, respectively. [25]
12) Right-click the "Donor Name E" on the Rows shelve and sort the donors by sum of amount in descending order.
13) Now we'd like to create an option that enables us to look at the top ten donors of each of the five countries. Here is when filters in tableau come in:
    a) Drag "Recipient Name E" into the "Filters" area. Select all of the five countries.
    b) Right-click on this filter, and click "Show Quick Filter". Now we will be able to use the filter we just created. Options include changing the filter into a single- or multi- selection one, and editing and changing the title of the filter.
    c) Now we also would like to include a filter that only shows the top ten donors, of for example, Vietnam. Drag "Amount" into the "Filters" area, click on "Sum" option, "next" and "ok" the filter.
    d) Right-click the Sum(Amount) filter and choose "Quick Table Calculation" option. Then choose "Rank" and select "1" and "10" as the range.
14) Rename the worksheet as "Donors".
15) Now we are able to create the first worksheet that enables people to look at the top ten donors of each of the five Mekong countries in 2013!

Filtering the Data: Sectors

16) Now we'd like to create a new sheet in our csv file that shows the sectoral amount and the shares of each sectors of the five countries.
17) Go to the master sheet that we worked on, and filter out the five countries in the Mekong region (see step 3)).
18) Click on the dropdown menu of "sectorname" and deselect the "(blanks)" option.
19) Copy and paste the filtered data to a new sheet renamed as "Sectors".
20) On the "Sectors" sheet, create a new column entitled "Total". To calculate the sum of each sectors in each of the five countries, use "sum" function on the amount. Sum the total amount for different sectors of the five countries. (You should be able to generate one total number for each of the five countries). To fill the remaining rows without messing up the total number data, lock the row number in our "sum" function by adding a $ sign. For example, if Cambodia has sectors from row 2 to row 11, with amounts in column D, we should modify the

[25] Users should try to explore different forms of visualization before deciding on which one to use. Under the "Marks" tab, users will be able to explore the possibilities of visualization, such as pie charts, line charts, horizontal/vertical bar charts etc. In this example, we think the best visualization form is a horizontal bar chart.

function into "sum(D$2:D$11)". Repeat the procedure with the five countries, and drag the function to the remaining blank rows.

21) Create a new column entitled "% of Total". In this column, we use functions to decide the share of the amount of each sector to the total. Use function "=amount/Total".
22) Now we will be able to have a csv worksheet that has our desired data.

Connecting to Tableau: "Sectors" worksheet.

23) On the same tableau file, start a new worksheet.
24) Click on "Connect to Data" option under the "Data" tab.
25) This time, choose the "Sectors" worksheet in our csv. file to connect to tableau. Go to the worksheet.
26) Drag the "% of Total" measure to "Columns" shelf.
27) Now we'd like to show different sectors by colors. To achieve this, drag the "Sectorname" dimension to "color" tab under the "Marks" area.
28) Then, to clearly label the different sectors with their shares, drag the "% of Total" measure to "Label" tab under the "Marks" area. Change the "format" of "SUM(% of Total) into "percentage".
29) Again, add the "Recipient Name E" dimension to the "Filters" area to create a country filter.
30) Rename the worksheet "sect
31) Now we have our second worksheet that visualizes the sectoral focus of the donations received by the five countries.

Presenting the tableau worksheets: Dashboard

32) Tableau presents (and publishes) the visualized worksheets in dashboards.
33) Click on the button to add a new Dashboard.
34) Drag and drop the "Donors" tableau worksheet to the top of the Dashboard.
35) Drag and drop the "Sectors" tableau worksheet to the bottom of the Dashboard.
36) Adjust the title of the Dashboard.

Creating one filter that controls two worksheets

37) Now we have a dashboard that has two worksheets, each controlled by a filter. We would like to create a filter that controls both of the sheets. One way to achieve this is to use Tableau functions and a parameter.
38) First, create a parameter entitled "Select Country:". To edit the parameter: select "String" in "Data Type" and then, click "Add from Field" and choose "Recipient Name E" from either the the "Donor" or "Sector" dataset. What parameter does is to create a list of characters that contains the five country names.

39) Then, in the "Donors" worksheet's Dimensions section, right-click to create a new calculated field. We name this new dimension as "country yes or no". In the "formula" part of the calculated field, write:
    a) if [Recipient Name E] = [Select Country:] then "yes"
    b) else "no"
    c) end
40) Save the "country yes or no" dimension. Use this dimension as the filter for worksheet "Donors" instead of "Recipient Name E". Choose only "yes" to create the filter.
41) Repeat the process and create a "country yes or no 2" for the "Sectors" worksheet. Use "country yes or no 2" to replace "Recipient Name E" as the sheet filter.
42) Now go back to the dashboard. Under the "Analysis" tab, click on the "Parameters" option to show the parameter "Select Country:".
43) Remove the two previous filters from the Dashboard.
44) Now, with the "Select Country:" parameter, we'll be able to control two worksheets.
45) Formatting the Dashboard: you can make the elements floating and fit them in vertical or horizontal boxes.
46) Add a title to the Dashboard by checking "Show Title" in the bottom left hand corner.

Publish the Dashboard through Tableau Public

47) Click "File" and "Save to Web". Log into your tableau public account and follow the instructions (If Tableau requires you to create a local extraction, just click "ok" to create a local tableau data file and then publish).
48) On the website that Tableau Publish generated, click on "Share" and get the weblink or html code of your online dashboard.
49) Share your work! An example of this visualization can be found here.