# Lesson 5
# Tools and Techniques

## Overview

This lesson can be overwhelming to beginners, so we will divide this into four parts: data hypothesis; data analysis; data processing; and data cleaning.

Duration: 4 hours (2 weeks)

## Introduction

For DEVC127, our approach in producing data-driven stories is the "data hypothesis approach" — a method of constructing a hypothesis and answering questions that can prove the veracity of the hypothesis by looking at available data, analyzing data, cleaning data, and presenting data (Internews). As a beginner, having a problem-first approach in writing your data-driven stories will keep you from being overwhelmed with a sea of data.

Like in a scientific method, you begin with a question or a problem that you wish to answer. As a development communicator, you will think of questions with a social lens. For example, how can we measure gender inequality with data? How can we come up with healthcare solutions with data? How can we find systemic biases in the education system with data? How can we prove global inequity with data? How can we protect our youth from vulnerability with data?

In this module, you will learn an approach to building data-driven stories, pick up the tools you need, and start practicing some techniques for beginners.

## Objectives

At the end of this lesson, you should be able to:

1. State the four questions when doing a data hypothesis approach
2. Define terms in basic analysis;
3. Describe how softwares and applications are used for processing data;
4. Clean data and define data aggregation

## Lesson proper

Discussion flow:

1. Data hypothesis approach and terminologies
2. Analyzing data with softwares and applications: spreadsheets, visualization softwares, and other tools for mining and running statistics
3. Sorting, selecting, grouping, and calculating
4. Tips on cleaning raw data and the importance of data aggregation in the development process

## DATA HYPOTHESIS

London, 1854.

A cholera outbreak hit a district in London. More than 100 people died in just three days from August 31 to September 2. By September 10, 500 had died. John Snow, a doctor in 19th century England, collected data to prove his theory that the dreaded infectious disease that caused severe diarrhea and deaths spread through germ cells when people consumed infected water or food. Back then, the "germ theory" ran opposed to "miasma theory" which posited the cholera spread through particles in the air (Winston, 2021).

Snow looked at the death rates and found that most of the deaths were in areas around the Broad Street water pump in the Soho district in London. Using a microscope, he saw certain particles when he examined some water from the said pump and was convinced that the pump was contaminated. Later, it was indeed discovered that pipes connected to toilets had leaked into, and therefore contaminated, the well on Broad Street.

Some "anomalies" or deviations from Snow's observations occurred. They did not fit Snow's Broad Street pump theory, but the doctor explained most of these, according to reports. One anomaly was the absence of death in a place called "The Lion Brewery" near the Broad Street pump. If Snow's germ theory were true, someone would have died there, but it was said that none of the 70 brewery staff died. It turned out, however, that the employees preferred beer, and the brewery had its own water pump, so their survival rate from the epidemic was 100%.

Snow is now considered by many to be the founder of modern epidemiology, the branch of medicine and public health that attempts to determine the causes and risks related to diseases.

Data journalism can be done like this. That is, starting with a hypothesis, then looking at data to verify.

Data hypothesis approach in journalism hypothesis is an investigative method based on Mark Lee Hunter's "story based inquiry" approach, which structures the investigation around a hypothesis. Like in our John Snow anecdote, public health or healthcare is one sector that can be of interest to a data journalist when launching a data hypothesis approach. Other sectors can be agriculture, economy, education, environment, gender, governance, and human rights. After contemplating the sector of choice, a journalist may then start thinking of an issue under the sector, hypothesize based on questions, look for evidence in (or lack of) data, and communicate the findings in a story.

This is what separates data journalism from data science. Whereas, "data science" is the study of how data can be turned into a resource (McGrath, 2018), data journalism is a step further, turning data into a story, which can raise awareness, even drive change. In the story-based inquiry manual, creating hypotheses is one of the initial steps in weaving a narrative and telling a journalistic story:

- We discover a subject.
- **We create a hypothesis to verify.**
- We seek open source data to verify the hypothesis.
- We seek human sources.
- As we collect the data, we organise it – so that it is easier to examine, compose into a story, and check.
- We put the data in a narrative order and compose the story.
- We do quality control to make sure the story is right.
- We publish the story, promote and defend it.

## Four questions

Asking questions may come handy when crafting a hypothesis. But, what type of questions? Internews proposes four categories:

1. **Problem** questions: how big is the problem? How expensive is it or how much does it cost a community? Is the problem getting better or worse?
2. **Impact** questions: Who is affected by the problem? How? Are some groups more affected than others?
3. **Cause** questions: What are the causes of the problem? Whose fault is it? What factors have made it worse?
4. **Solution** questions: What is the solution to the problem? How can we measure effectiveness?

**Problem questions**. If you go outside and look around you, you may instantly spot a social problem that affects you or members of your family. If you are in a rural town, a visible problem may be lack of access roads that connect farms to markets. If you are in a highly urbanized city, an obvious problem may be heavy traffic. A potential hypothesis from this issue can be: Heavy traffic in EDSA has not improved despite the Build Build Build project. After recognizing a problem in your community that you think needs urgent action, you may proceed to listing specific *problem questions* such as the following:

- How serious is the traffic problem in Metro Manila?
- How much does daily congestion in EDSA cost the Philippine economy?
- During the pandemic/after Build Build Build, is the traffic problem getting better or worse?

**Impact questions**. Suppose you are keen to probe another problem, say, mental health problems. *Impact questions* focus on who/what group is affected in a population. During the pandemic, you may have noticed that the retail and service sectors have been terribly hit. Meaning, sales ladies and masseuses have gone out of work when lockdowns were imposed, possibly leading to distress and mental health problems. You may hypothesize then that the mental health of women have been affected more than that of men during the pandemic. If you wish to focus on the youth, your impact questions may be:

- Who among students (college? High school? elementary?) in Region 4A are mentally/emotionally affected by the COVID-19 lockdowns?
- How are their mental health affected?
- Are students more affected than their working parents?

**Cause questions**. Just like John Snow, a data journalist can also investigate the cause of a problem, with help from data. Suppose you got curious about the rise of teenage pregnancies and wanted to know the root of this population trend. One hypothesis could be: Lack of reproduction health programs in barangays could lead to higher rates of teenage pregnancies in the area. In coming up with a sound hypothesis, cause questions that can be asked are:

- What is causing the boom in pregnancy rates among teenagers?
- Who is at fault?
- What made the problem worse during the pandemic?

**Solutions questions**. Looking for solutions, especially to complex issues like inequality, may be a challenge. So, asking concrete *solutions questions* may work better in a data story. For example: What is the solution to increase intergenerational mobility in Laguna, or how can we make it easier for lower-income families in Laguna to be part of "higher opportunity" neighborhoods? So, a data story defined by solutions questions may ask *How can we measure upward mobility in Laguna*? One possible hypothesis is that good education can drive upward social mobility in a community. The journalist can then start looking for data and any correlation, for instance, between the number of highly accredited schools and the median income in a community. Other solutions questions may be:

- What is the solution to gender inequality in rural areas? How can women have a higher social and financial capital?
- What is the solution to youths' growing mental health problems? How can we measure the effectiveness of a mental health awareness drive?

## Basic terminologies in producing a data-driven story

Once you are ready with your chosen sector, story topic, questions, and hypothesis, it is high time to start looking for data that can be analyzed, processed, and cleaned. But, what does it mean for data to undergo these steps? Here are some terminologies.

**Data analysis** -- also known as "data analytics" -- is "the practice of converting collected data into information that is useful for decision-making," however, the raw data typically

undergo data processing and data cleaning before it can be explored for insights (McGrath, 2018).

**Video analytics in shopping stores: A sample data analysis**

Videos are data, too. And some apps and software specialize in video data. In the US, there is a software and application called "Percolata" which uses video analytics in the retail industry to track which areas in a store receives the most attention ("foot traffic"). Using video footages as data, retailers can flag individuals using face recognition, track activities in a store, and identify repeat visits.

Uniqlo and 7-Eleven are two companies that use Percolata to assess the productivity of their staff and predict retail sales by floor or area in their store, based on hundreds of hours of videos. They use the software in their decision-making processes (marketing, staffing, etc).

How do you think is it different or similar to police surveillance in some cities around the world?

*Review of statistics concepts*

| Technique | Definition |
|---|---|
| Data | Numerical and non-numerical observations |
| Population | The group containing all possible entities of concern |
| Sample | A part of the population examined or studied |
| Observation | Each separate collection of one bit of data |
| Study | Collecting data and using statistics to make an inference about it |
| inference | An educated, statistically supported "guess" about a group of data |
| Descriptive Statistics | Values that describe (e.g. center, spread, shape) data sets |
| Inferential Statistics | Making educated guesses, testing theories, modeling observations' relationships, and predicting outcomes with data analysis |

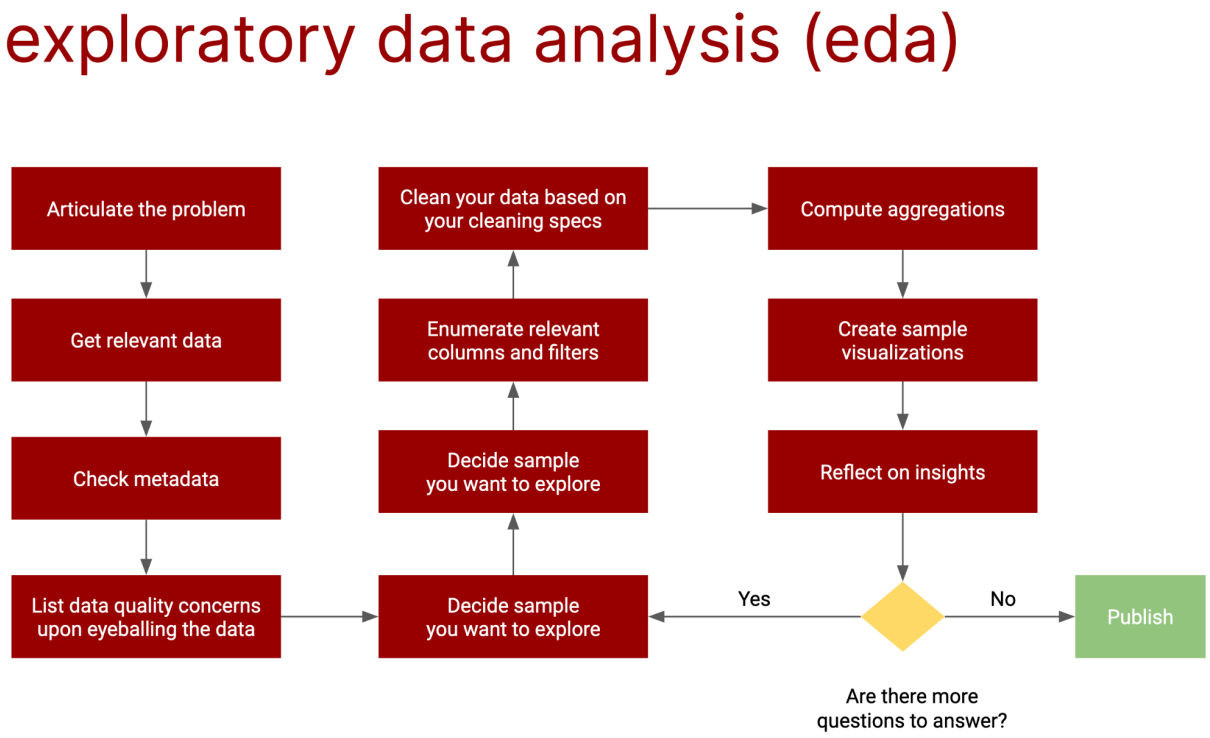
| | |
|---|---|
| Descriptive observations | Data that describes qualities rather than amount |
| Random variables | Numerical or descriptive observations that happen by chance |
| Data set | A group of collected observed |
| Quantitative data | Numerical data |
| Qualitative data | Non-numerical data e.g. text |

**Data processing** means organizing raw data into a structured format (e.g., arranging data into rows and columns in a table).

**Data cleaning** means stripping the organized data of incomplete, duplicated, and erroneous characters or items.

If you remember, it was mentioned in the last module that data is never clean. 80% of the job is data janitorials–cool analysis stuff is only about 20%. Data cleaning, processing and analysis are iterative processes—the first pass will always lead to more questions. Keep in mind that to know how a dataset can be cleaned, you should also understand how it is collected, processed, and stored into the format you have now.

After processing and cleaning, the data journalist can then explore the data to observe and discover its characteristics. For some stories, further data or refining may be needed. Journalists might also need to calculate descriptive statistics (e.g., mean) to bring out a story in the data.




*Exploratory data analysis process flow*

==========

## Learning activity

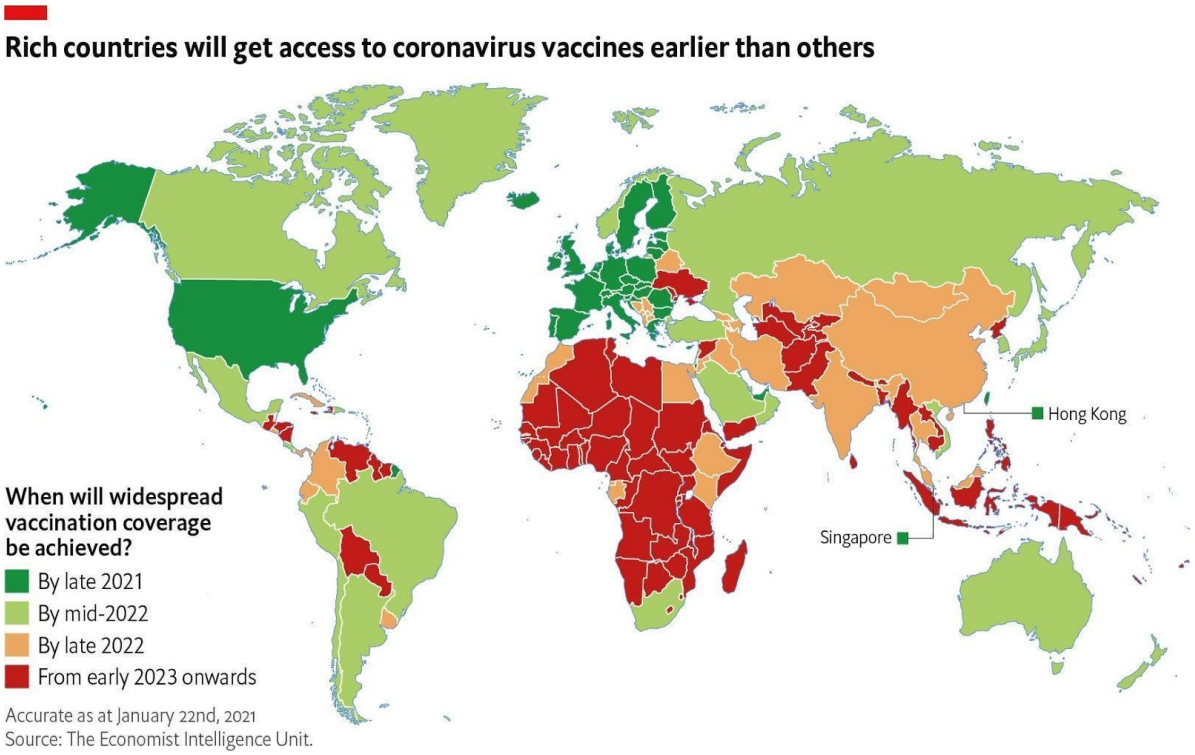
**Read** Chapter 2 of *Story-based inquiry: a manual for investigative journalists*: "Using hypotheses: the core of investigative method." Look at the "graphic way" of looking at the process of setting out the hypothesis, understand the advantages of hypothesis driven investigations and how hypotheses work, and check out the case study on Baby Doe. In what ways can a data journalist strengthen a hypothesis?

**Watch** Fireside chat with data analysts on Data Cleaning 101. This is a synchronous session during lecture hours and recording will be released a day after.

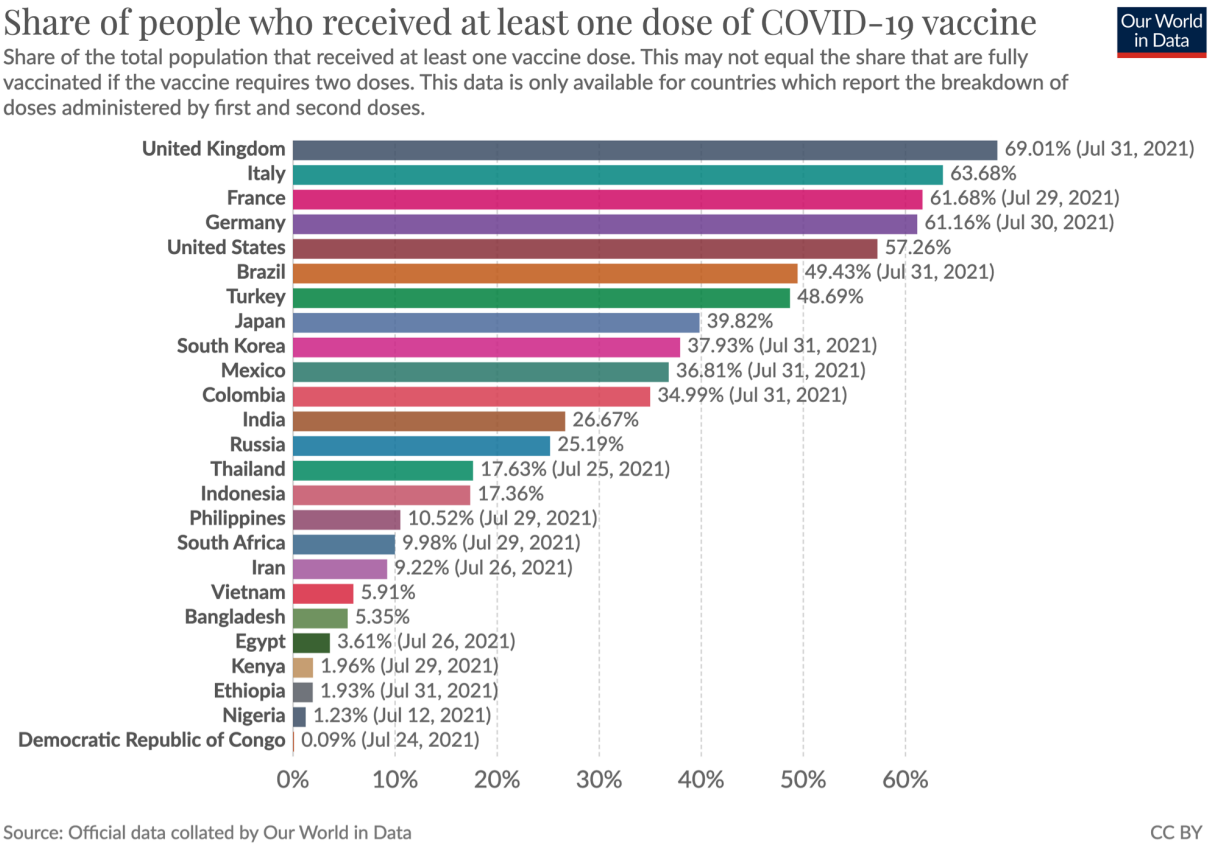
==========

## TOOLS

Early in 2021, just before massive vaccinations rolled out in several countries, *The Economist* released a data-driven scoop using only this map:




Rich countries around the world (United States, European countries, and Singapore) were expected to get widespread vaccination coverage earlier than other countries. Based on the data visualization, for example, the residents of the United Kingdom (green) will get coronavirus jabs by late 2021, while those in the Philippines (bloody red) will get widespread vaccination from early 2023 onwards.




Charts like these can be a powerful way of telling stories of inequalities and other social problems. What are the basic tools a data journalist will need to process data and create graphics such as these?

## Rundown of popular tools

**Excel**. This spreadsheet program is considered the most pervasive application when it comes to business analytics. Statisticians, financial reporters, data analysts, and almost anyone dealing with numbers rely on Microsoft Excel. Data journalists start the processing of data with Excel. Below are other tools that budding data journalists can start tinkering with.

**Google Sheets**. This looks almost like Excel, only that it is free, online, and shareable. The last bit is the most interesting feature because multiple users can work on the same worksheets at the same time, which can come quite handy when collaborating with a group of fellow students or journalists.

**Programming languages**. R, D3 (JavaScript), and Python sound like Star Wars characters, but these are languages used in coding programs that can aid data journalists in processing and presenting data. These programming languages have a steep learning curve but allow for greater flexibility. Meaning, they are customizable; one can control specific elements of the graphs created.

Python is considered the more popular one among those whose work revolves around data. Python is a programming language that can scrape unstructured data such as a text document in paragraph format; such scraping requires a developer to "teach" the computer how to extract the text into a data format. Other examples of unstructured data are videos, audio, social media feeds, and email. Python, with its syntax close to human-readable language, offers readily available packages best used for data wrangling, mining, and visualization. R is a programming language and software used for data analysis and visualization. Often used with R is the RStudio Integrated Development Environment. RStudio has "a code editor, debugging features, and visualization tools that make R easier to use" (McGrath, 2018).

**Other tools.** For larger projects, data journalists may need the help of database management such as Microsoft Access or SQL (another programming language). For statistical analysis, tools such as **SPSS or STATA** can be used. When extracting or scraping data from PDFs, tools that can be used are **Tabula, CometDoc, PDFtoExcel, and Zamzar**. For giant image files (in GIF, JPEG, PNG, and BMP formats) that contain data, Optical Character Recognition Software may be used to recognize the text, or data, in the scanned bitmap images.

For beginners, all you need is a good old spreadsheet program. Spreadsheet skills are transferable to programming for analytics.

## Visualization softwares

Visualizing data in a graphical format enhances one's ability to see patterns and trends. Data visualization is closely related to information graphics ("infographics") as both are about communicating data visually. Never has there been so many visualization tools before; some of them require license, but many are free.

**Datawrapper** and **Flourish** are two of the most popular and easy-to-use data visualization applications among journalists today. With Datawrapper, you can build your own chart for free and no sign-up required. Tables are responsive and customizable. You can select from 20 chart types to show your data (e.g., bar, split, stacked, scatter plot, donut, pies). Flourish has templates for visualizations and stories. You can customize your design and download your output without the need for coding or installing software because it is web-based.

**Tableau** is another popular data viz software and it can be useful for exploratory analysis as it creates multiple views and eye catching graphs from data. It can be leveraged for the exploratory type of journalism via the Story Points feature. Some coding may be required, and an account comes for a fee (though you can try it for free, and there is a free 1-year license if you are a student). You also need to download the software to use their dashboard.

**Adobe Illustrator** by itself or together with graphs created in an app like Excel or via a programming language can also be used for easy manipulation of graph elements. License is required to open this graphic design software.

**Powerpoint** can be a data visualization tool, too. You may copy-paste your graph or chart from Excel and use Powerpoint's design applications to modify some elements of the visual such as titles or axis labels. To draw attention to a specific data point, you may do so through a textbox in Powerpoint. The animation feature of Powerpoint can also be useful in showing a progression in data.

## CLEANING RAW DATA

> "You've likely waited so long for the dataset that it's tempting to jump right in and get to work, but the first few hours of most data projects should involve cleaning up the data to make sure it's usable" (Constantaras).

Remember: Even experts say that data analysis can be extremely difficult to double check, so one must work slowly and take frequent breaks. Even in cleaning data, or putting data in a standard form, one must take time.

Cleaning is not just about removing empty rows and checking empty cells. Data standardization ensures that all the data input is in the same, consistent format. Dates, ages, addresses, and measurements are usually entered into a database and these should be standardized. Each type of data should have its own cell; meaning, merged cells should be avoided.

Microsoft published "Top ten ways to clean your data" to guide Excel users and here are the most useful ones:

1. Import the data.
2. Create a backup copy of the data in a separate workbook.
3. Ensure that all columns and rows are visible, with no blank rows.
4. Spell check and standardize the spelling and format if there are abbreviations (use Find and Replace if necessary).
5. Remove duplicate rows or duplicate values.
6. Change the case of the text if text comes in mixed cases (e.g., emails, product codes, proper case) by typing =LOWER or =UPPER or =PROPER in a cell next to the one to be corrected, followed by the cell location. For example: =PROPER(A2) [this changes JUAN CRUZ to Juan Cruz].
7. In fixing numbers and number signs, you may convert a number to text, convert a text string that represents a number to a number, round up/off a number to a specified number of decimals, or apply a currency symbol.
8. In fixing dates and times, you may convert between different time units or format a sequential serial number as a date.
9. Merge columns by selecting two or more adjacent cells you want to emerge, then on the Home tab, select Merge & Center (or Merge Cells). You can also select cells to split data. Concatenate is another function to join two or more text strings into one.
10. When manipulating a column, insert a new column next to the column that needs cleaning; fill down the formula in the new column if necessary, then remove the original column.

## SPREADSHEET TECHNIQUES

To see how processed data is used in a global industry, let's look at Netflix.

Netflix is one success story not just in the streaming and media industry, but also in the world of big data. Science writer Brian Clegg said that we, the viewers, see the outcome of Netflix's extensive data analysis in the recommendation system of the platform. Netflix, which was once a DVD rental service, has transformed itself as a result of the big data age. Its system attempts to predict the likes and dislikes of the viewers by analyzing its mass of customer data. Netflix knows who is watching what, when, and where. When viewers "like" *Game of Thrones* or *Stranger Things*, for instance, Netflix taps its data to recommend movies and series with the same director or the same actors or the same genre.

More than using data to predict preferences, Netflix also commissions movies or series based on success hits. An example is the *House of Cards*. Traditionally, a media company releases a pilot first, then tests it with different audiences to see if the series will push past one season. But, with analysis of data from the Netflix library, Netflix determined that the *House of Cards* will be a success; hence, instead of producing a pilot first, they went ahead and poured in $100 million up front for 26 episodes in the first two series.

Datasets, especially large ones in familiar formats, can be overwhelming. Usually though, these can be converted into user-friendly formats (e.g., comma-separated values transferred to neat rows and columns).

Once you have the data in spreadsheets, you may then proceed to sorting, selecting, grouping, and computing data.

### Sorting

We cannot work on raw, disorganized data. We need to convert plain text file data into a format that a computer software can understand so that it can organize data into rows, columns, and cells. One sheet in an Excel workbook contains 65,000 columns and 1,000,000 rows. So, we are dealing with tens of millions of cells if we fill up a sheet.

Oftentimes we need to organize these cells and sort them out alphabetically or numerically from largest to smallest or vice versa. Sometimes we may need to sort a dataset or rows by chronological order (date or time). In cases like this, the date/time format should be consistent (e.g., YYYY/MM/DD).

Suppose you gained an access to Netflix's database and wanted to put some kind of order, you can do so by sorting data by arranging the list of subscribers' names (A to Z or Z to A), or dates (oldest to newest or vice versa), or movies' number of views (smallest to largest, or largest to smallest). Sorting data helps a user understand and find data better.

Sorting can be done in several ways. One is by selecting a cell in the column to be sorted, and clicking Data tab > Sort & Filter group > Ascending/Descending (or Sort Smallest to Largest, Oldest to Newest, etc,).

## Selecting

Selecting cells in a spreadsheet means clicking on a cell or more. When selecting numerous cells, just click on a cell and drag the pointer until the necessary cells are highlighted. When dealing with large data sets, it's impractical to scroll through hundreds or thousands of cells to select certain values so here are easy tips:

- When selecting non-adjacent cells, you can hold Ctrl and select the cells or rows/columns of cells.
- Ctrl+A when selecting the entire worksheet.

You can visit Microsoft Excel's support website to see how to freeze panes to lock rows and columns, hide or show rows/columns, transpose or rotate from rows to columns and vice versa, and add or remove rows/columns.

## Grouping

Suppose you have organized your Netflix data by city (one sheet per city), but you wanted to perform tasks on multiple worksheets at the same time -- you can do so by grouping worksheets together. Grouping worksheets means that if you perform functions or do some changes in one sheet, these will be applied to the rest of the group. Grouping them together, to calculate the average of views across 100 cities for instance, saves time.

Grouping and doing changes across the group/s work best if the worksheets already have identical data structures; meaning, the rows and columns have the same locations. As well as grouping selected or all worksheets, ungrouping can also be done.

Grouping and ungrouping can be done by pressing and holding Ctrl, then clicking the worksheet tabs to be grouped or ungrouped. You can also right-click a worksheet tab, Select All Sheets or Ungroup Sheets.

## Calculating

Calculating data for data journalism usually employs the standard statistical calculations: finding the average, the middle value, the range, and minimums and maximums in a dataset. "Descriptive statistics," they call it.

Analysing data -- statistically speaking -- means looking at a typical value for the data (mean, median, mode), the spread of data, the nature of data, and the identification of unusual data points. A mathematician said "It is human nature to try to summarize data with a single number," and that the typical value for a data set is usually the mean or the median, depending on the skewness of data. We call an unusual data point an "outlier" which means any data point that is more than two standard deviations from the mean. Identifying why a data point is unusual or why outliers occurred in a dataset can often help us better understand a dataset.

Now to calculate data in Excel, use the "formulas" in Excel. An easy way to do this is this:

1. Select a cell where you want to place the output.
2. Type the equal sign (=).
3. Select the cell/s to be calculated. (Or, type the cell address. E.g., C3)
4. Enter an operator (+ or - or / or *)
5. Press Enter.

You can also enter a formula that has a built-in function. For example, =AVE for getting the average:

1. Select an empty cell.
2. Type an equal sign and then the function (AVE)

3. Type an open parenthesis.
4. Select the range of cells to be computed.
5. Type a closing parenthesis.
6. Press Enter.

The rest of useful formulas and functions can be found here.

These spreadsheet techniques are similar to how you implement it on Google Sheets. You can find the references on sorting, grouping, and calculating in the embedded links.

*Examples of aggregation techniques in Spreadsheets*

| Technique | Definition | Best practice | How to do it |
|---|---|---|---|
| Average/Mean | The sum of all values divided by the total number of values | Most commonly used measure of central tendency | `=AVERAGE(range_start: range_end)` |
| Median | The middle number in an ordered dataset | Best used when few outliers are expected and missing values are present | `=MEDIAN(range_start: range_end)` |
| Mode | The most frequent value | Best used when dealing with categorical data | `=MODE(range_start: range_end)` |
| Range | Simple measure of spread | Best to illustrate the difference between the minimum vs. maximum values in the population | `=MAX(range_start: range_end) - MIN(range_start: range_end)` |
| Quartiles | Division of ordered set into 4 groups<br>• Lower quartile - 25th percentile<br>• Second quartile - median<br>• Upper quartile - 75th percentile<br>• Interquartile range - upper Q3 and lower quartiles (approach that eliminates outlier effect) | Best used when outliers are expected within the population | `=QUARTILE(range_start: range_end, quartile_number)` |

### Bend your tool

Author and data analyst Cole Knaflic, who used to work at Google, had some encouraging words about selecting the tools for processing data:

> Try not to let your tools be a limiting factor when it comes to communicating effectively with data. Pick one and get to know it as best you can… Any frustration you encounter will be worth it when you can bend your tool to your will!

It does take time to look at data and determine the appropriate ways to show it. And, it takes more time to put them together into a cohesive and captivating narrative. What's important is that you, the storyteller, have access to the basic tools. Again, the tools are just one part of the process and must not hinder you if they look intimidating. After all, as Knaflic said, "Our tools do not know the story we aim to tell."

==========

### Learning activity

**Read** Excel dynamic array functions: what data journalists need to know (https://datajournalism.com/read/longreads/excel-dynamic-array-functions-what-data-journalists-need-to-know)

==========

## Data disaggregation

These Excel tricks are useful in data disaggregation. But what is it and why is it important? Aggregated data means combining data and lumping them together to provide a "big picture" of socioeconomic conditions, concealing the real state of social inequities. Breaking down the data into certain categories can show the disparities that exist between population groups (ADB, 2021).

Data disaggregation is defined as:

> the breakdown of observations, usually within a common branch of a hierarchy, to a more detailed level to that at which detailed observations are taken. With standard hierarchical classifications, statistics for related categories can be split

(disaggregated) when finer details are required and made possible by the codes given to the primary observations (UN Glossary of Classification Terms)

When properly disaggregated, data can identify the vulnerable, the disadvantaged, the marginalized. Gender-disaggregated data, for instance, enables policymakers and decision makers to design effective intervention programs for women, as these data help in understanding why certain segments of the population are left behind in the development process.

==========

## Learning activities

**Read** how data disaggregation can help in global efforts to "Leave No One Behind" using this Inclusive Data Charter flyer:
https://www.data4sdgs.org/sites/default/files/2018-08/IDC_onepager_Final.pdf

**Watch** Live tutorial on data processing, cleaning, and analysis using Google Sheets. This is a synchronous session during lecture hours and recording will be released a day after. The live tutorial is intended to entertain questions real-time while doing the data cleaning activity.

==========

## SUMMARY OF LESSON 5

When doing data journalism, one can use the **data hypothesis approach** and ask the four questions:

- **Problem** questions: how big is the problem? How expensive is it? Is the problem getting better or worse?
- **Impact** questions: Who is affected by the problem? How? Are some groups more affected than others? How old are they? Where do they live?
- **Cause** questions: What are the causes of the problem? Whose fault is it? What factors have made it worse?
- **Solution** questions: What is the solution to the problem? How can we measure effectiveness?

Sectors that data journalists can focus on are agriculture, economy, education, environment, gender, governance, health, and human rights.

When importing data sets from external sources, these collections of tabular data are often stored in comma-separated values (CSV) file and must be converted into a form that can be sorted, selected, grouped, and calculated. While the **mean** is used for normal distribution, the **median** is generally used for skewed distributions.

**Data analysis** is the practice of converting collected data into information that is useful for decision-making. **Data processing** means organizing raw data into a structured format, usually in a tabular form. **Data cleaning** means fixing incomplete, duplicated, and erroneous characters or items in worksheets.

Softwares and applications such as Python and R are used for processing data. These programming languages provide built-in utility functions that allow for extraction of statistics from a range of values.

**Data disaggregation** is breaking down data by sub-categories to reveal inequalities that may not be fully shown in an aggregated data.

## REFERENCES


Asian Development Bank. (May 2021). **Practical guidebook on data disaggegation for the sustainable development goals**. Retrieved August 2021 from https://www.adb.org/sites/default/files/publication/698116/guidebook-data-disaggregation-sdgs.pdf

Clegg, Brian. (2017). **Big Data: How the information revolution is transforming our lives**. London: Hot Science.

Constantaras, Eva. (2020). **Internews Jordan Data Journalism Training Manual.** Retrieved August 2021 from https://arij.net/wp-content/uploads/2020/07/Data-Journalism-Training-Manual.pdf

Knaflic, Cole Nussbaumer. (2015). **Storytelling with data: a data visualization guide for business professionals**. Canada: Wiley.

Lee Hunter, Mark. (2011). **Story-based inquiry: a manual for investigative journalists**. UNESCO Publishing.

McGrath, Mike. (2018). **R for Data Analysi**s. UK: In Easy Steps.

Winston, Wayne. (2021). **Analytics stories: using data to make good things happen**. Canada: Wiley.

## Exercise 2: Cleaning, filtering, and analyzing data

**Overview**

Data plays a fundamental role in this brand of journalism—you must be able to extract interesting insights from raw data. The full pipeline of data analysis is a tedious one. Before you can arrive at an insight, it takes hours of data cleaning, processing, and analysis.

This exercise has the following major steps:

1. Data cleaning
2. Data processing
3. Exploratory data analysis

**Objectives**

At the end of the exercise, you should be able to:

1. Enumerate at least three questions you want to answer in exploring the data.
2. Explore the data step-by-step.
3. Analyze your data set based on the angle of the story you would want to work on.
4. Identify at least three initial insights from the data.

**Mechanics and Deadlines**

1. Review Excel tutorials on cleaning data on spreadsheets and apply the guidelines on your data set. Take note of the steps that your team has taken to scrub the sheets.
2. Have a look at basic statistics, filters, vlookups, if functions, and pivot tables. These are often the steps needed to manipulate data into different formats and find stories. As the key areas of data analysis, these five areas should be practiced over the next week.
3. Select the sections of your data that you will use for your story and come up with an angle/thesis statement based on this.

| | |
|---|---|
| Week 1: Data cleaning | Create a backup of your extracted data set. Check the following: spelling, duplicate values/rows/columns. Find other datasets that may complement your data and clean it, too. Remove unnecessary columns/rows/formulas/headings. In short: omit errors and standardize style. |
| Week 2: Data processing | Processing data is sorting and putting your data sets into a structured format. Merge rows or columns or even |

| | |
|---|---|
| | spreadsheets that need to go together, or split the units that need to be separated. Insert formulas when necessary for your visualization. Name rows and columns |
| Week 3: Exploratory data analysis | Make comparisons, do arithmetic, assess logic, observe patterns, summarize the content of data (e.g., mean, median, outliers).<br><br>For metadata analysis, answer:<br>1. Who/what is the source of your data?<br>2. How frequently is it published?<br>3. Where is it used?<br>4. How did the source organize their data?<br>5. How did you clean/filter/reorganize the data set?<br>6. Are there ethical concerns/issues that need to be raised?<br><br>Using any of these four categories of inquiry, write at least three questions that you seek to answer with your cleaned and processed data. You may mix categories:<br><br>- **Problem** questions: how big is the problem? How expensive is it? Is the problem getting better or worse?<br>- **Impact** questions: Who is affected by the problem? How? Are some groups more affected than others? How old are they? Where do they live?<br>- **Cause** questions: What are the causes of the problem? Whose fault is it? What factors have made it worse?<br>- **Solution** questions: What is the solution to the problem? How can we measure effectiveness? |

**Expected Outputs**

1. Cleaned spreadsheet
2. Filtered data
3. Initial data analysis

**Assessment**

| Criterion | Performance Level | | |
|---|---|---|---|
| | 8-10 points | 4-7 points | 1-3 points |

| | | | |
|---|---|---|---|
| Data cleaning (Multiplier: 4) | A backup is created. Spelling is checked and standardized. Values and components are complete. Duplicate rows and values removed. Cases and number formats are standardized. Columns/rows properly merged or split when necessary. Formula/column/row inserted when necessary. Bonus: Complementary datasets. | A backup is created. Spelling is checked and standardized, but not entirely. Some values and components are missing. One or more rows/values were unnecessarily duplicated. Cases are mixed (upper/lower) and some number formats are of different format. Some columns/rows are not properly merged or split. Some formulas/columns/rows need to be inserted when necessary. | No backup. Spelling is not checked and standardized. Values and components are missing. Duplicate rows and values are present. Cases and number formats are standardized. Columns/rows that should have been merged are split, and vice versa. Important formulas/columns/rows are missing. |
| Initial analysis of the data (Multiplier: 4) | A pattern, if not a story, is clearly identified. This can either be a newsworthy increase or decrease in values, correlation (which may be causation) of datasets/values/components, or an indication of a particular trend. Metadata analysis explains the source, manner/process, frequency, and purpose of dataset. | A pattern or a story is identified, but not newsworthy. Metadata analysis cites the source, manner/process, frequency, and purpose of dataset, but lacks some details. | No pattern or story is identified; not newsworthy. Metadata analysis is missing, i.e., no explanation who or what the source of data is, how the dataset was processed, how frequent the data comes out, and what the dataset is used for. |
| Human interest angle highlighted in the story pitch (Multiplier: 2) | Potential story in data shows a development-oriented problem or calls for urgent action or solution. | Potential story in data is newsworthy and has some development angle. | Potential story in data is not newsworthy and has no development angle. |
| ***Total*** | | | |

***Perfect score=100 points***