

# **Obtaining, Wrangling and Analysing Messy Data**

## **Unit 3**

### **Nepal Data Literacy Program, 2019**

Organized by



Supported by



# **Course Contents**

Module 1: Scraping Data from the Web

Module 2: Scraping Data from PDFs and Images

Module 3: Cleaning Data

Module 4: Strengthening your Data Reliability

Module 5: Finding Stories in the Facts by visualizing data

Module 6: Finding insights in data

# Module 1

# **Scraping Data from the Web**

# What is Scraping?

- Many times data is not easily accessible – although it does exist and is available on the web, but most times in formats that are not easily usable for analysis
- **Scraping** describes the method to extract data hidden in documents – such as Web Pages and PDFs and make it useable for further processing
- It is among the most useful skills if you set out to investigate data – and most of the time it's not especially challenging. For the most simple ways of scraping you don't even need to know how to write code

# What we will cover

## Scraping data from the Web

- Scraping using Scraper extension on Chrome
- Scraping using Table2Clipboard2 on Firefox
- Two lab exercises

# Getting Started

A screenshot of a web browser window. The address bar shows the URL [data.un.org/en/iso/np.html](http://data.un.org/en/iso/np.html) with a 'Not secure' warning. Below the address bar is a navigation bar with icons for back, forward, refresh, and home. To the right of the navigation bar is the word 'Nepal'. Underneath the navigation bar is the flag of Nepal. Below the flag is a list of four categories, each preceded by an orange triangle icon.

- ▶ General Information
- ▶ Economic indicators
- ▶ Social indicators
- ▶ Environment and infrastructure indicators

Source: <http://data.un.org/en/iso/np.html>

# Download Scraper on



1. Open Chrome browser
2. To download **Scraper Extension**, go to Chrome webstore
3. Search for 'Scraper'
4. In the 'Scraper' window that opens, click the blue + ADD TO CHROME button

The screenshot shows the 'Scraper' extension page on the Chrome Web Store. At the top, there's a large 'Scraping' icon. Below it, the title 'Scraping' is displayed, followed by 'Offered by: dvhvn'. A rating of 339 stars is shown. Below the title, there are tabs for 'Overview', 'Reviews', and 'Related'. The main content area shows a table comparing various JavaScript frameworks. The table includes columns for Framework, Version compared, Size, License, and Source language. The 'Amplo SDK' is highlighted in the table.

Framework	Version compared	Size	License	Source language
Amplo SDK	0.9.3 1 Jul 2013	Variables, Core size: 40 KB (minified & gzipped)	MIT & GPL	?
AngularJS	1.2.14 1 Mar 2014	96 KB (minified & gzipped)	Apache 2.0	JavaScript
CupQ (abandoned)	0.2 June 2012	26 KB (minified & gzipped)	Scrape similar...	?
DHTMLX	4.0 4 Jun 2014	Variables	Look Up in Dictionary	?
Dojo	1.10.3 08 Dec 2014	41 KB (minified & gzipped), 150 KB (minified, 598 KB (uncompressed))	BSD & AFL	JavaScript + HTML
Echo3	3.0.61 24 Mar 2011	?	MPL, LGPL or GPL	JavaScript and/or Java

[https://chrome.google.com/webstore/detail/scraping/mbigbapnjcgaffohm\\_bkdkleccacepngid?hl=en-GB&utm\\_source=chrome-ntp-launcher](https://chrome.google.com/webstore/detail/scraping/mbigbapnjcgaffohm_bkdkleccacepngid?hl=en-GB&utm_source=chrome-ntp-launcher)

# Using Chrome Extension:

1. Open a new Chrome window and go to:

<http://data.un.org/en/iso/np.html>

2. Open table with the heading "Economic Indicators"

3. Highlight the rows of the "Economic Indicators" table, then right-click and select Scrape similar.

4. In the Scraper window that opens, click the **Export to Google Docs**

Exchange rate (per US\$)			
Nepal			
	2005	2010	2018
GDP: Gross domestic product (million current US\$)	8.259	16.281	20.914 <sup>e</sup>
GDP growth rate (annual %, const. 2010 prices)	3.1	4.8	0.4 <sup>e</sup>
GDP per capita (current US\$)	322	602	722 <sup>e</sup>
Economy: Agriculture (% of Gross Value Added)	35.2	35.4	31.6 <sup>e</sup>
Economy: Industry (% of Gross Value Added)	17.1	15.1	14.2 <sup>e</sup>
Economy: Services and other activity (% of GVA)	47.7	49.5	54.2 <sup>e</sup>
Employment in agriculture <sup>f</sup> (% of employed)	76	74.8	71.3
Employment in industry <sup>f</sup> (% of employed)	4.7	7.5	8.2
Employment in services <sup>f</sup> (% of employed)	1	Look Up "GDP: Gross domestic product (million current US\$)..."	20.5
Unemployment rate <sup>f</sup> (% of labour force)	79.9 / 81.0	Copy	2.7
Labour force participation rate <sup>f</sup> (female/male pop. %)	79.9 / 81.0	Search Google for "GDP: Gross domestic product (million current US\$)..."	82.7 / 85.8
CPI: Consumer Price Index (2010=100)	2.2	Print...	117.8 <sup>d</sup>
Agricultural production index (2004-2006=100)	1.4	Scrape similar...	140 <sup>e</sup>
International trade: exports (million current US\$)	81.4	Inspect	741 <sup>d</sup>
International trade: imports (million current US\$)	2.2	Speech	10.036 <sup>d</sup>
International trade: balance (million current US\$)	-1.4	Services	-9.297 <sup>d</sup>
Balance of payments, current account (million US\$)	1		.815 <sup>d</sup>

Social indicators			
Nepal			
	2005	2010	2018
Population growth rate <sup>f</sup> (average annual %)	1.5	1.1	1.2 <sup>b</sup>
Urban population (% of total population)	15.1	16.8	19.7
Urban population growth rate <sup>f</sup> (average annual %)	4	3.1	3.2 <sup>b</sup>
Fertility rate, total <sup>f</sup> (live births per woman)	3.6	3	2.3 <sup>b</sup>
Life expectancy at birth <sup>f</sup> (females/males, years)	65.2 / 62.9	68.1 / 65.5	70.4 / 67.4 <sup>b</sup>
Population age distribution (0-14/60+ years old, %)	39.7 / 6.7	37.0 / 7.4	30.2 / 8.8 <sup>a</sup>
International migrant stock <sup>f</sup> (000s/ of total pop.)	679.5 / 2.6	578.7 / 2.1	502.7 / 1.7 <sup>d</sup>
Refugees and others of concern to UNHCR 000	538.8 <sup>k</sup>	891.3 <sup>k</sup>	24.4 <sup>d</sup>

# Download Table2clipboard



1. Open Firefox browser
2. Download an add-on called [Table2Clipboard2](#)
3. On this webpage, click Continue to Download
4. On the next page that opens, click + Add to Firefox

The screenshot shows the Firefox Add-ons page with the search bar set to "Table2Clipboard2". The results show the "Table2Clipboard2" extension by "Firefox user 14263523". The extension icon is a clipboard with a document. Below the title, it says "Allow to copy to clipboard an HTML table rows/columns selection correctly formatted. This is forked from <https://github.com/dafizilla/firefox-table2clipboard>". A blue "Add to Firefox" button is visible. Below the main card, there are two sections: "Rate your experience" and "About this extension". The "Rate your experience" section has a "Log in to rate this extension" button. The "About this extension" section describes the extension's purpose and its GitHub fork. At the bottom, there are links for "Report this add-on for abuse", "Read all 9 reviews", and a "Wallabagger" sidebar.

Firefox Add-ons

Explore Extensions Themes More... ▾

Table2Clipboard2 by Firefox user 14263523

Allow to copy to clipboard an HTML table rows/columns selection correctly formatted. This is forked from <https://github.com/dafizilla/firefox-table2clipboard>

+ Add to Firefox

Rate your experience

How are you enjoying Table2Clipboard2?

Log in to rate this extension

Report this add-on for abuse

Read all 9 reviews

About this extension

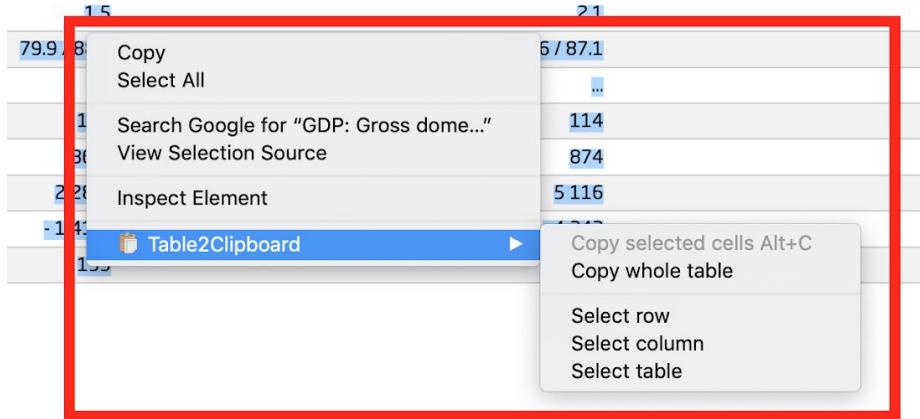
Allow to copy to clipboard an HTML table rows/columns select. This is forked from <https://github.com/dafizilla/firefox-table2clipboard>. See original author's website for details. <http://dafizilla.sourceforge.net/table2clip/>

Other users with this extension also installed

Wallabagger v2.6.1 3,641 users

# Using Firefox Extension

1. Open a new Firefox window and go to:  
<http://data.un.org/en/iso/np.html>
2. Open table with the heading "Economic Indicators"
3. Highlight table, then right-click and select Table2Clipboard > Copy whole table
4. Open a new spreadsheet. In the cell, paste the copied table (Ctrl+V)



# Lab 1: Scrape Data Using Browser Extension



Article Talk Read Edit View history

## List of Nepal government organizations

From Wikipedia, the free encyclopedia

The following is a list of Government Organizations of Nepal:<sup>[1]</sup>

Government Office	Website URL
Government of Nepal	<a href="http://www.nepal.gov.np">www.nepal.gov.np</a>
Office of the President & V. President	<a href="http://www.presidentofnepal.gov.np">www.presidentofnepal.gov.np</a>
Commission for the Investigation of Abuse of Authority (CIAA)	<a href="http://www.ciaa.gov.np">www.ciaa.gov.np</a>
Ministry of Federal Affairs and General Administration (MoFAGA)	<a href="http://www.mofaga.gov.np">www.mofaga.gov.np</a>
Election Commission	<a href="http://www.election.gov.np">www.election.gov.np</a>
Office of the Auditor General	<a href="http://www.oagnep.gov.np">www.oagnep.gov.np</a>
Public Service Commission	<a href="http://www.psc.gov.np">www.psc.gov.np</a>
Supreme Court	<a href="http://www.supremecourt.gov.np">www.supremecourt.gov.np</a>
National Human Rights Commission	<a href="http://www.nhrcnepal.org">www.nhrcnepal.org</a>
Office of the Prime Minister and Council of Ministers	<a href="http://www.opmcn.gov.np">www.opmcn.gov.np</a>
Ministry of Finance	<a href="http://www.mof.gov.np">www.mof.gov.np</a>
Financial Comptroller General Office	<a href="http://www.fcgo.gov.np">www.fcgo.gov.np</a>
Department of Custom	<a href="http://www.[http://customs.gov.np customs.gov.np]customs.gov.np">www.[http://customs.gov.np customs.gov.np]customs.gov.np</a>
Inland Revenue Department	<a href="http://www.ird.gov.np">www.ird.gov.np</a>
Department of Revenue Investigation	<a href="http://www.dri.gov.np">www.dri.gov.np</a>
Department of Revenue	<a href="http://www.ratc.gov.np">www.ratc.gov.np</a>
Ministry of Industry	<a href="http://www.moi.gov.np">www.moi.gov.np</a>
Department of Industry	<a href="http://www.doind.gov.np">www.doind.gov.np</a>
Department of Mines and Geology	<a href="http://www.dmanenai.gov.np">www.dmanenai.gov.np</a>

Data source: [https://en.wikipedia.org/wiki/List\\_of\\_Nepal\\_government\\_organizations](https://en.wikipedia.org/wiki/List_of_Nepal_government_organizations)

# Lab 2 :Scraping Data Using Google Spreadsheets

The screenshot shows a Google Spreadsheet interface. At the top, there is a table titled "Environment and infrastructure indicators" with data for the years 2005, 2010, and 2018. Below this, the main spreadsheet area has a title bar "Untitled spreadsheet" and a menu bar with File, Edit, View, Insert, Format, Data, Tools, Add-ons, Help, and a status bar indicating "All changes saved in Drive". The formula bar at the bottom contains the formula =IMPORTHTML("http://data.un.org/en/iso/np.html", "table", 5). The spreadsheet grid shows row 1 containing the formula and rows 2 through 4 empty.

	2005	2010	2018
Individuals using the Internet (per 100 inhabitants)	0.8	7.9 <sup>q</sup>	19.7 <sup>e</sup>
Research & Development expenditure (% of GDP)	...	0.3 <sup>r,s</sup>	...
Threatened species (number)	77 <sup>m</sup>	93	104 <sup>d</sup>
Forested area (% of land area)	25.4	25.4	25.4 <sup>b</sup>
CO <sub>2</sub> emission estimates (million tons/tons per capita)	3.1 / 0.1	5.1 / 0.2	8.0 / 0.3 <sup>n</sup>
Energy production, primary (Petajoules)	349	384	430 <sup>b</sup>
Energy supply per capita (Gigajoules)	14	17	18 <sup>b</sup>
Tourist/visitor arrivals at national borders <sup>t</sup> 000	375	603	753 <sup>e</sup>
Important sites for terrestrial biodiversity protected (%)	50.3	54.6	54.6
Pop. using improved drinking water (urban/rural, %)	93.1 / 80.2	92.0 / 86.0	90.9 / 91.8 <sup>b</sup>
Pop. using improved sanitation facilities (urban/rural, %)	47.8 / 26.7	51.9 / 35.1	56.0 / 43.5 <sup>b</sup>

Untitled spreadsheet star folder

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

fx =IMPORTHTML("http://data.un.org/en/iso/np.html", "table", 5)

	A	B	C	D	E	F	G
1	=IMPORTHTML("http://data.un.org/en/iso/np.html", "table", 5)						
2							
3							
4							

Data Source: <http://data.un.org/en/iso/np.html>

# Module 2

# **Scraping Data from PDFs and Images**

# What we will cover

**Extracting** data hidden in PDF documents to make it useable for further processing

- **Scraping data from PDF using online tools**
  - Sodapdf - Splitting PDFs
  - Zamzar
  - Cometdocs
- **Scraping data from PDf using offline tools**
  - Using Tabula
- **Scraping data from images**
  - Using Google doc OCR

# Getting Started

Before we move forward, here is a reminder about data formats:

- **Machine readable, structured:** generated by a computer, and are organized in rows and columns. e.g - CSV (comma-separated values), TSV (tab-separated values), Excel (.xls)
- **Unstructured:** These are sometimes generated by a computer, but are not organized as data table by computer. For example - PDF, Word, and bitmap images (GIF, JPEG, PNG, BMP)
- **Portable Document Format (PDF):** These files may include charts that contain data, but the data is saved in an unified document with text.
- **Excel file (XLS):** files save data as tables readable by Microsoft Excel
- **Comma separated values (CSV):** Plain text files with each data point separated by a comma

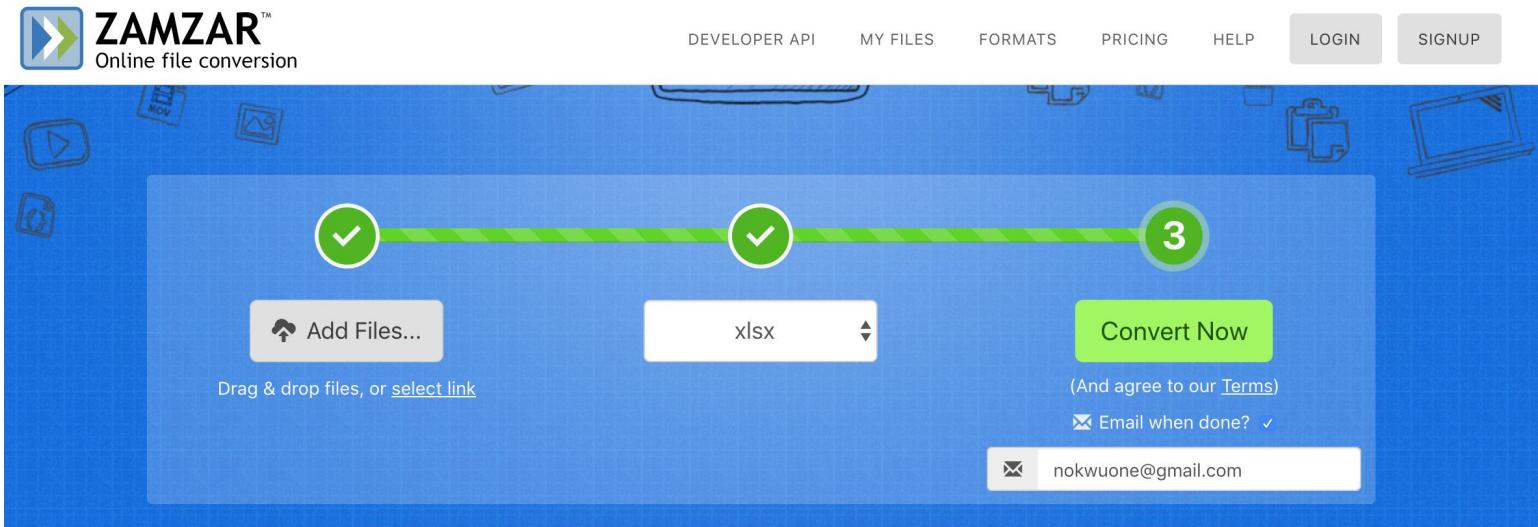
# Extracting the Required Page in PDFs

1. Go to  
<https://www.sodapdf.com/split-pdf/>
2. Upload the  
[Nepal-Labour-Force-Survey-2017\\_18-Report.pdf](#)
3. Select page 29, click Split! "Then view and download"
4. Save file with title Table 2.1

The screenshot shows the Soda PDF website's interface for splitting PDFs. At the top, there are three ways to upload files: dragging them onto a 'DROP FILES HERE' area, clicking a 'CHOOSE FILE' button, or selecting from Google Drive or Dropbox. Below this, a large central box contains two numbered steps: Step 1 shows a downward arrow pointing to 'VIEW & DOWNLOAD IN BROWSER', and Step 2 shows a box for 'Send file by email' with fields for 'Email address' and a checkbox for agreeing to receive communications. A note states that information will be handled in accordance with their Privacy Policy. A green 'SEND TO EMAIL' button is at the bottom. To the right, a sidebar titled 'Do More with Soda' lists other features: PDF Editor, Compress PDF, Protect PDF, and PDF Converter, each with its own icon.

<https://www.sodapdf.com/split-pdf/>

# Convert PDF to CSV online Using Zamzar



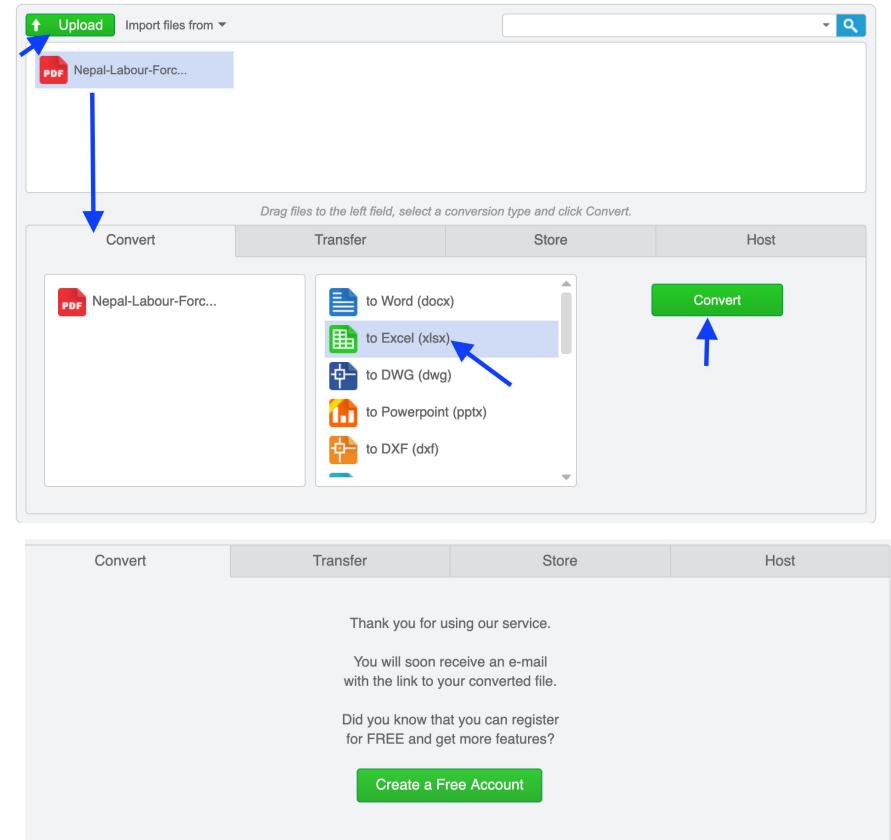
## Files to Convert

FILE NAME	FILE SIZE	PROGRESS

<http://www.zamzar.com/>

# Converting PDF to Excel Online with Cometdocs:

1. Open [www.cometdocs.com](http://www.cometdocs.com) Click the Go to the Web App button.
2. **Sign in** and Upload the [Nepal-Labour-Force-Survey-2017\\_18-Report.pdf](#) report.
3. Click convert, drag-and-drop the file to the empty box.
4. Click conversion type Word/Excel and click convert
5. File downloads into your computer. Open file with google sheets to view.



# Offline data Scraping using Tabula:

1. Install [Java](#)
2. Open [Tabula website](#) and install for your operating system.
3. Extract the downloaded zip file
4. Go into the “tabula” folder. Run the tabula.exe
5. A web browser will open – this is Tabula or go to:  
<http://127.0.0.1:8080/>

## Tabula



Tabula is a tool for liberating data tables locked inside PDF files.

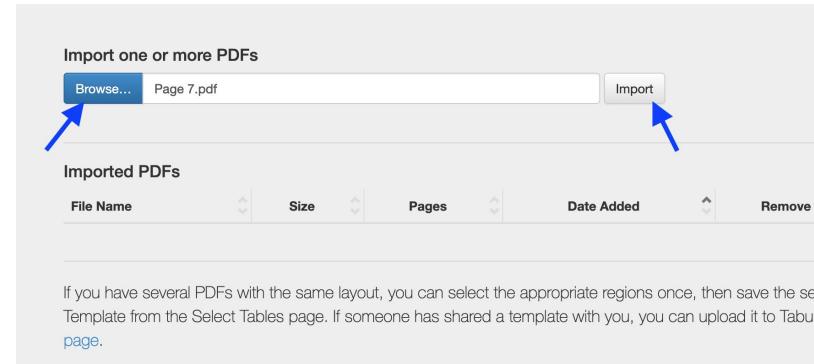
[View the Project on GitHub](#)  
tabulapdf/tabula



<https://tabula.technology/>

# Using Tabula:

1. Import the file you split earlier called **page 29.pdf**
2. Select the data table
3. Click preview and export
4. A new window opens
5. On the window select lattice and click export format in CSV



Clear All Selections Autodetect Tables Preview & Export Extracted Data

(urban/rural). A profile of household heads also forms part of this chapter.

Table 2.1: Distribution of population by age, sex and locality

Age group (Years)	Nepal			Urban			Rural			thousands
	Male	Female	Total	Male	Female	Total	Male	Female	Total	
Total	13509	15513	29022	8578	9708	18285	4932	5805	10737	
00-04	1526	1399	2925	907	821	1728	618	578	1197	
05-09	1280	1119	2400	782	673	1455	498	446	944	
10-14	1493	1458	2953	938	871	1809	558	586	1144	
15-19	1418	1518	2936	915	965	1880	503	553	1056	
20-24	1116	1602	2718	786	1053	1838	331	549	880	
25-29	955	1345	2300	631	911	1542	324	434	758	
30-34	759	1130	1889	524	756	1280	236	374	610	
35-39	779	1130	1909	509	730	1239	270	400	670	
40-44	722	932	1654	470	589	1058	252	344	595	
45-49	717	858	1575	466	549	1016	251	308	559	
50-54	658	676	1333	420	424	844	237	252	489	
55-59	562	635	1197	346	365	711	216	270	486	
60-64	502	601	1103	285	355	640	217	246	463	
65+	1021	1110	2131	599	646	1246	422	463	885	percent
Total	100	100	100	100	100	100	100	100	100	
00-04	11.3	9.0	10.1	10.6	8.5	9.5	12.5	10.0	11.1	
05-09	9.5	7.2	8.3	9.1	6.9	8.0	10.1	7.7	8.8	
10-14	11.1	9.4	10.2	10.9	9.0	9.9	11.3	10.1	10.7	
15-19	10.5	9.8	10.1	10.7	9.9	10.3	10.2	9.5	9.8	
20-24	8.3	10.3	9.4	9.2	10.8	10.1	6.7	9.5	8.2	
25-29	7.1	8.7	7.9	7.4	9.4	8.4	6.6	7.5	7.1	
30-34	5.6	7.3	6.5	6.1	7.8	7.0	4.8	6.4	5.7	
35-39	5.8	7.3	6.6	5.9	7.5	6.8	5.5	6.9	6.2	
40-44	5.3	6.0	5.7	5.5	6.1	5.8	5.1	5.9	5.5	
45-49	5.3	5.5	5.4	5.4	5.7	5.6	5.1	5.3	5.2	
50-54	4.9	4.4	4.6	4.9	4.4	4.6	4.8	4.3	4.6	
55-59	4.2	4.1	4.1	4.0	3.8	3.9	4.4	4.7	4.5	
60-64	3.7	3.9	3.8	3.3	3.7	3.5	4.4	4.2	4.3	
65+	7.6	7.2	7.3	7.0	6.7	6.8	8.5	8.0	8.2	

# Using Tabula:

1. Your table to be exported looks like this.
2. File downloads to your computer as tabula-page 7.csv.
3. Open google sheet and import file and keep the default settings
4. Now you have successfully scraped this data.
5. Save your google sheet as the title of table 2.1 - **Distribution of population by age, sex and locality.** for future reference in our exercises

Is the extracted data incorrect?

You can revise your selected cells or try an alternate extraction method.

Revise Selected Cells

Data has been extracted from the cells you selected in the previous step. You can revise your selection(s) to add or remove cells.

◀ Revise selection(s)

Choose Alternate Extraction Method

The current preview uses the Stream extraction method. If the data is not mapped to the correct cells, try the Lattice method instead.

Stream    Lattice

Stream looks for whitespace between columns, while Lattice looks for boundary lines between columns.

Still look wrong?

Contact the developers and tell us what you tried to do that didn't work.

The screenshot shows the Tabula software interface. On the left, there's a preview of a table with data about age groups and gender. On the right, a file download dialog box is open, asking what to do with the CSV file 'tabula-Page 7.csv'. The 'Save File' option is selected. The file is described as a CSV Document (1.5 KB) from the URL http://127.0.0.1:8080.

Age group (Years)	Male	Total	Male	Female	Total
Total	13509	8285	4932	5805	10737
00-04	1526	728	618	578	1197
05-09	1280	455	498	446	944
10-14	1496	809	558	586	1144
15-19	1418	880	503	553	1056
20-24	1116	838	331	549	880
25-29	955	1 345	2300	631	911
30-34	759	1130	1889	524	756
35-39	779	1130	1909	509	730
40-44	722	932	1 654	470	589
45-49	717	858	1 575	466	549
50-54	658	676	1 333	420	424
55-59	562	635	1 197	346	365
60-64	502	601	1 103	285	355

<https://tabula.technology/>

# Extracting Data from Images Using



1. Open link of this [blog](#)
2. Right-click on Table and “**Save image as**”
3. Open Google Drive, click new from the drop-down menu, select **File upload**
4. Browse and select **saved file**
5. Search for the uploaded file and right-click on **the file** icon, select **Open with > Google Docs**
6. Google Doc with the table opens in new window. You will see that the data is distorted.



HOME BLOGS ABOUT US DATA EVENTS RESOURCES GET INVOLVED INI

## Education status in Dailekh district from data perspective

Posted on: September 23, 2014

How should we go about finding the status of education? I don't know but thought that going through the flash reports from Ministry of Education and getting the data from those reports, might give some insights. It was no easy activity. The flash report is more than 100 pages with all the data in tabular format in pdf, compressing data as much as possible.

A screenshot of a right-click context menu on a table cell. The menu items are: Open Image in New Tab, Save Image As..., Copy Image, Copy Image Address, Search Google for Image, Get Similar (Data Miner), Scrape similar..., Inspect, and Speech. A blue arrow points from the top right towards the 'Save Image As...' option.

Dist. Code	Dist. Name	Grade 6			Grade 7			Grade 8			Grade 9			Grade 10		
		Girls	Boys	Total	Girls	Boys	Total	Girls	Boys	Total	Girls	Boys	Total	Girls	Boys	Total
51.1	Mid western dev. reg.	34794	39449	74153	1779	3379	5158	160	247	407	553	948	1501	742	1377	2119
52	Dolpa	144	374	518	180	433	613	164	374	538	144	374	518	180	433	613
53	Jumla	160	247	407	160	247	407	160	247	407	160	247	407	160	247	407
54	Kakol	160	247	407	160	247	407	160	247	407	160	247	407	160	247	407
55	Mugu	160	247	407	160	247	407	160	247	407	160	247	407	160	247	407
56	Humla	160	247	407	160	247	407	160	247	407	160	247	407	160	247	407
56.1	Mid west. hill	16547	19001	35548	1779	3379	5158	160	247	407	160	247	407	160	247	407
57	Pyuthan	160	247	407	160	247	407	160	247	407	160	247	407	160	247	407
58	Rolpa	160	247	407	160	247	407	160	247	407	160	247	407	160	247	407
59	Rukum	160	247	407	160	247	407	160	247	407	160	247	407	160	247	407
60	Salyan	160	247	407	160	247	407	160	247	407	160	247	407	160	247	407
61	Surkhet	160	247	407	160	247	407	160	247	407	160	247	407	160	247	407
62	Dailekh	160	247	407	160	247	407	160	247	407	160	247	407	160	247	407
63	Jajarkot	160	247	407	160	247	407	160	247	407	160	247	407	160	247	407

It gives comprehensive data on the districts and on various indicators. Personally i don't see the utility of those data in the report. We ventured out to see the process of getting data for one district and chose Dailekh and see if anything useful what could be produced. For ease, we just collected enrolment data for all grades, segregated by gender and ethnic group (Dalit, Janajati and others).

We extracted data from annual flash reports from 2064 to 2069 and used Microsoft Excel for the analysis and

# **Exercise**

1. Go to this [report - Innovative Strategies in Higher Education for Accelerated Human Resource Development in South Asia: Nepal](#)
2. Extract Table 9 into google sheets
  1. Find the page where this table is located
  2. Split the PDF to extract the page
  3. Upload into Zamzar or Tabula and extract

Module 3

# **Cleaning Data**

# **Steps to take when cleaning data**

1. Understand the data in front of you and take notes of what you noticed and would like to fix.
2. Standardize the data formats this includes looking for errors like inconsistent age, dates, units formats etc.
3. Fix structural and formatting errors.
4. Remove irrelevant columns or rows and rearrange columns where necessary.
5. Pay attention to missing values by removing the variables that are missing or impute the missing values from based on the other variables
6. Check maths and fix any errors

# Cleaning Your Scrapped Data

## Table 2.1

1. Open the Distribution of Population by Age, sex and Locality
2. First make a copy of your data by right-click **Sheet 1** tab, and select **duplicate**
3. Right-click the sheet tab to **Rename**. Rename sheet 1 to **original** and the new tab to **working copy** rename sheets to avoid confusion.

The screenshot shows a Google Sheets document titled "Distribution of population by age, sex and locality". The sheet contains data for Nepal, categorized by age group (00-04 to 65+) and sex (Male, Female, Total). The columns are labeled A through E. A context menu is open over the first row, with "Duplicate" highlighted by a blue arrow. Another blue arrow points to the new tab bar at the bottom, which shows "page 29" and "Copy of page 29".

A	B	C	D	E
Age group		Nepal		Urban
(Years)	Male	Female	Total	Male
Total	13509	15513	29022	8578
00-04	1526	1399	2925	907
05-09	1280	1119	2400	782
10-14	1496	1458	2953	938
15-19		1518	2936	915
20-24		1602	2718	786 1 053
25-29			2300	631
30-34		1130	1889	524
35-39		1130	1909	509
40-44		932 1 654		470
45-49		858	1575	466
50-54		676	1333	420
55-59		635	1197	346
60-64		601	1103	285
65+		0	2131	599
Total		100	100	100
00-04	0	9	10.1	10.6
05-09	7.2	8.3	9.1	
10-14	9.4	10.2	10.9	
15-19	0.8	10.1	10.7	

# Understand the data in front of you.

Does the distribution of population data have information that may be useful for analysis to you or does it have any interesting observation?

1. Are there columns or values you don't understand?
2. Any missing or obviously wrong values?
3. Do we need to clean it?

Make note of your findings before you clean.

	A	B	C	D	E	F	G	H	I	J
1	Age group		Nepal			Urban			Rural	
2	(Years)	Male	Female	Total	Male	Female	Total	Male	Female	Total
3										thousands
4	Total	13509	15513	29022	8578	9708	18285	4932	5805	10737
5	00-04	1526	1399	2925	907	821	1728	618	578	1197
6	05-09	1280	1119	2400	782	673	1455	498	446	944
7	10-14	1496	1458	2953	938	871	1809	558	586	1144
8	15-19	1418	1518	2936	915	965	1880	503	553	1056
9	20-24	1116	1602	2718	786	1053	1838	331	549	880
10	25-29	955	1345	2300	631	911	1542	324	434	758
11	30-34	759	1130	1889	524	756	1280	236	374	610
12	35-39	779	1130	1909	509	730	1239	270	400	670
13	40-44	722	932	1 654	470	589	1058	252	344	595
14	45-49	717	858	1575	466	549	1016	251	308	559
15	50-54	658	676	1333	420	424	844	237	252	489
16	55-59	562	635	1197	346	365	711	216	270	486
17	60-64	502	601	1103	285	355	640	217	246	463
18	65+	1021	1 110	2131	599	646	1246	422	463	885
19										percent
20	Total	100	100	100	100	100	100	100	100	100
21	00-04	11.3	9	10.1	10.6	8.5	9.5	12.5	10	11.1
22	05-09	9.5	7.2	8.3	9.1	6.9	8	10.1	7.7	8.8
23	10-14	11.1	9.4	10.2	10.9	9	9.9	11.3	10.1	10.7
24	15-19	10.5	9.8	10.1	10.7	9.9	10.3	10.2	9.5	9.8
25	20-24	8.3	10.3	9.4	9.2	10.8	10.1	6.7	9.5	8.2
26	25-29	7.1	8.7	7.9	7.4	9.4	8.4	6.6	7.5	7.1

Data source: [Nepal Central Bureau of Statistics](#)

# Standardize the data formats

1. **Let's standardize the percentage values to have its own header and copy into a new page.**
2. In cell B20, input the formula =B2&"%" then enter and drag the result up to J20
3. Other things to note include: Representing your dates, age, unit, addresses in the right uniform format.

00-04	50+	00-1	1100	200	300	040	211	240	400
55+	1021 110	Male% x	2131	599	646	1246	422	463	885
Total	=B2&"%"	Female%	Total%	Male%	Female%	Total%	Male%	Female%	Total%
00-04		9	10.1	10.6	8.5	9.5	12.5	10	11.1
05-09	9.5	7.2	8.3	9.1	6.9	8	10.1	7.7	8.8
10-14	11.1	9.4	10.2	10.9	9	9.9	11.3	10.1	10.7

12/06/2019  
2019/06/12  
19/12/06  
12th June, 2019

Not Standardized

12/06/2019  
12/06/2019  
12/06/2019  
12/06/2019

Standardized

# Fix structural and formatting errors

1. Remove **merged cells**
2. Look for improper labelling and combine labels where necessary
3. **Combine row 1-2 to get uniformed column header. - Pg 36**
4. Now Let's fix the row labels.
5. **Remove duplicates** and rename where necessary.
6. **Find and Replace.**

Distribution of population by age, sex and locality

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive Share

Find and replace

Find: total

Replace with: sum

Search: Specific range 'Working copy!'

Match case

Match entire cell contents

Search using regular expressions Help

Also search within formulas

Replaced 7 instances of total with sum

Find Replace Replace all Done

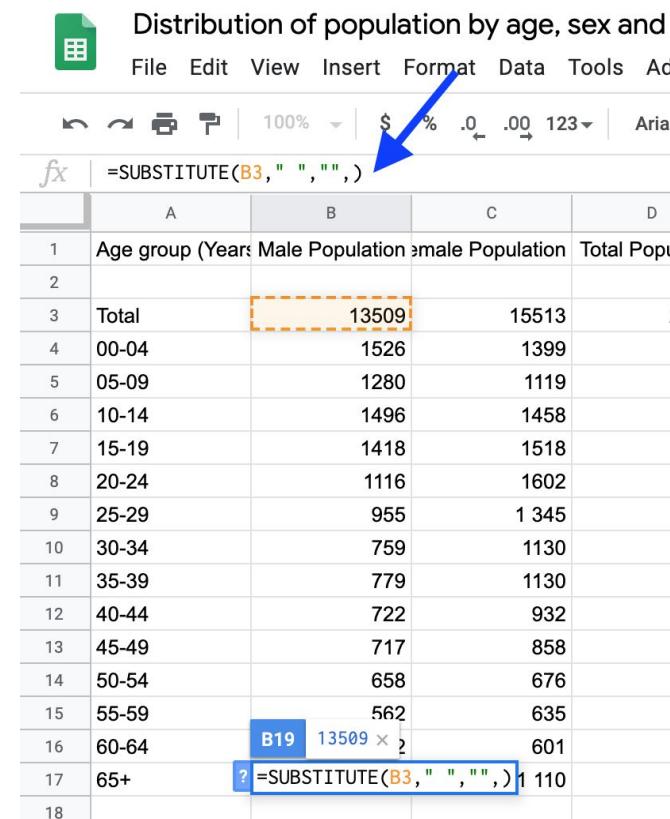
	A	B	C	D	E	F	G	H
1	Age group (Years)	Male Popu						Urban
2	sum							Rural Male
3	00-04							18285
4	05-09							4932
5	10-14							1728
6	15-19							618
7	20-24							1455
8	25-29							498
9	30-34							1809
10	35-39							558
11	40-44							1880
12	45-49							503
13	50-54							1838
14	55-59							331
15	60-64							1542
16	65+							324
17	sum	Male Popu						1280
18	00-04							236
19	05-09							1239
20	10-14							270
21	15-19	10.5	9.8	10.1	10.7	9.9	10.3	10.1
22	20-24	8.3	10.3	9.4	9.2	10.8	10.1	6.1

# Fix structural and formatting errors Cont'd

1. **Check for spelling errors** and inconsistent capitalization
2. Remove white spaces and empty cells using **substitute function**  
**=SUBSTITUTE(B3," ", "")**

## Others

3. Trim function: Trim function lets you remove trailing spaces  
**=TRIM()**
4. Split function lets you separate values by an identifier =**SPLIT()**
5. Concatenate function lets you join values =**CONCAT()**

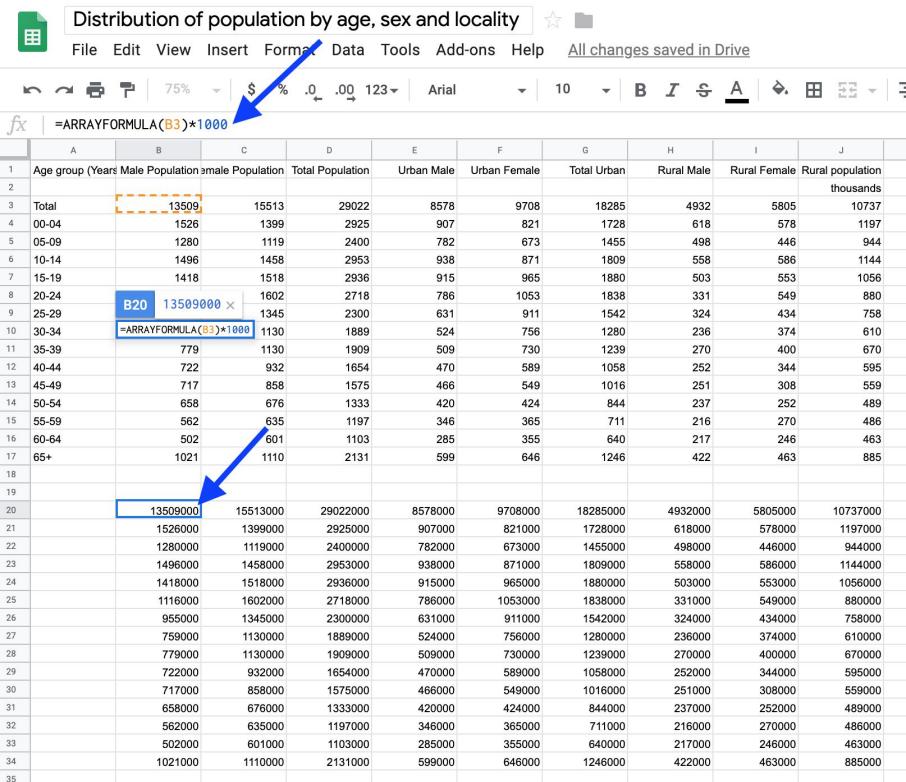


A screenshot of a Google Sheets spreadsheet titled "Distribution of population by age, sex and". The formula bar shows the formula `=SUBSTITUTE(B3, " ", "")`. A blue arrow points from the formula bar to the value 13509 in cell B3, which is highlighted with a dashed orange border. The spreadsheet contains data for age groups and their male and female populations.

	A	B	C	D
1	Age group (Years)	Male Population	Female Population	Total Popu
2				
3	Total	13509	15513	
4	00-04	1526	1399	
5	05-09	1280	1119	
6	10-14	1496	1458	
7	15-19	1418	1518	
8	20-24	1116	1602	
9	25-29	955	1 345	
10	30-34	759	1130	
11	35-39	779	1130	
12	40-44	722	932	
13	45-49	717	858	
14	50-54	658	676	
15	55-59	562	635	
16	60-64	13509	601	
17	65+	? =SUBSTITUTE(B3, " ", "")	1110	
18				

# Remove irrelevant columns/rows and rearrange columns where necessary

1. Here, you can move the percentage computed data to the next page
2. Cell J2 is showing that the values are to their thousands. Apply this to the whole data
3. In B18 input the arrayformula  
**=ARRAYFORMULA(B3)\*1000**  
This means to multiply the cell by 1000. Drag through the range to apply to the other cells.
4. **Paste special** to clear formula and delete the thousand row
5. You can further arrange the data by moving the **total row** beneath



A	B	C	D	E	F	G	H	I	J
Age group (Years)	Male Population	Female Population	Total Population	Urban Male	Urban Female	Total Urban	Rural Male	Rural Female	Rural population thousands
Total	13509	15513	29022	8578	9708	18285	4932	5805	10737
00-04	1526	1399	2925	907	821	1728	618	578	1197
05-09	1280	1119	2400	782	673	1455	498	446	944
10-14	1496	1458	2953	938	871	1809	558	586	1144
15-19	1418	1518	2936	915	965	1880	503	553	1056
20-24		1602	2718	786	1053	1838	331	549	880
25-29		1345	2300	631	911	1542	324	434	758
30-34		1130	1889	524	756	1280	236	374	610
35-39	779	1130	1909	509	730	1239	270	400	670
40-44	722	932	1654	470	589	1058	252	344	595
45-49	717	858	1575	466	549	1016	251	308	559
50-54	658	676	1333	420	424	844	237	252	489
55-59	562	635	1197	346	365	711	216	270	486
60-64	502	601	1103	285	355	640	217	246	463
65+	1021	1110	2131	599	646	1246	422	463	885
	13509000	15513000	29022000	8578000	9708000	18285000	4932000	5805000	10737000
	1526000	1399000	2925000	907000	821000	1728000	618000	578000	1197000
	1280000	1119000	2400000	782000	673000	1455000	498000	446000	944000
	1496000	1458000	2953000	938000	871000	1809000	558000	586000	1144000
	1418000	1518000	2936000	915000	965000	1880000	503000	533000	1056000
	1116000	1602000	2718000	786000	1053000	1838000	331000	549000	880000
	955000	1345000	2300000	631000	911000	1542000	324000	434000	758000
	759000	1130000	1889000	524000	756000	1280000	236000	374000	610000
	779000	1130000	1909000	509000	730000	1239000	270000	400000	670000
	722000	932000	1654000	470000	589000	1058000	252000	344000	595000
	717000	858000	1575000	466000	549000	1016000	251000	308000	559000
	658000	676000	1333000	420000	424000	844000	237000	252000	489000
	562000	635000	1197000	346000	365000	711000	216000	270000	486000
	502000	601000	1103000	285000	355000	640000	217000	246000	463000
	1021000	1110000	2131000	599000	646000	1246000	422000	463000	885000

# Check Maths

1. Insert a new row just below row 16.
2. Compute the sum for column B =**sum(B1:B15)**.
3. **You will see the disparity between the values.**
4. Go ahead and do this for the other totals by dragging the tip of the cell to the left to apply the formula in other cells.
5. This goes to show the importance of cleaning your datasets and verifying your Maths

Distribution of population by age, sex and locality

	A	B	C	D	E	F	G	H	I	J	K
1	Age group (Years)	Male Population	Female Population	Total Population	Urban Male	Urban Female	Total Urban	Rural Male	Rural Female	Rural population	
2	00-04	1526000	1399000	2925000	907000	821000	1728000	618000	578000	1197000	
3	05-09	1280000	1190000	2400000	782000	673000	1455000	498000	446000	944000	
4	10-14	1496000	1458000	2953000	938000	871000	1809000	558000	586000	1144000	
5	15-19	1418000	1518000	2936000	915000	965000	1880000	503000	553000	1056000	
6	20-24	1116000	1602000	2718000	786000	1053000	1838000	331000	549000	880000	
7	25-29	955000	1345000	2300000	631000	911000	1542000	324000	434000	758000	
8	30-34	759000	1130000	1889000	524000	756000	1280000	236000	374000	610000	
9	35-39	779000	1130000	1909000	509000	730000	1239000	270000	400000	670000	
10	40-44	722000	932000	1654000	470000	589000	1058000	252000	344000	595000	
11	45-49	717000	858000	1575000	466000	549000	1016000	251000	308000	559000	
12	50-54	658000	676000	1333000	420000	424000	844000	237000	252000	489000	
13	55-59	562000	635000	1197000	346000	365000	711000	216000	270000	486000	
14	60-64	502000	601000	1103000	285000	355000	640000	217000	246000	463000	
15	65+	1021000	1110000	2131000	599000	646000	1246000	422000	463000	885000	
16	Total	13509000	15513000	29022000	8578000	9708000	18285000	4932000	5805000	10737000	
17		13511000	15513000	29023000	8578000	9708000	18286000	4933000	5803000	10736000	
18											
19											
20											
21											
22											
23											

Data source: Nepal Central Bureau of Statistics

# Exercise

Take a look at the data on this link [here](#) and answer the following question.

- What errors or inconsistencies can you spot?
- What will you do to correct it?

Synthetic data for cleaning exercise

	A	B	C	D	E	F	G
1		From	To	Customer Id	Customer Age	Flight Date	
2	22323	Kathmandu	Biratnagar	75043	28	6/26/2019	
3	22324	Chitwan	Nepalgunj	51560	1990	6/26/2019	
4	<b>22325</b>	<b>Kathmandu</b>	<b>Chitwan</b>	<b>39940</b>	<b>43</b>	<b>6/26/2019</b>	
5	<b>22325</b>	<b>Kathmandu</b>	<b>Chitwan</b>	<b>39940</b>	<b>43</b>	<b>6/26/2019</b>	
6	<b>22325</b>	<b>Kathmandu</b>	<b>Chitwan</b>	<b>39940</b>	<b>43</b>	<b>6/26/2019</b>	
7	22326	Kathmandu	Chitwan	29680	41	6/26/2019	
8	22327	Chitwan	Kathmandu	2366	35	6/26/2019	
9	22328	Pokhara	Chitwan	24218	18	6/26/2019	
10	22329	Biratnagar	Kathmandu	41728	19	<b>26th July 19</b>	
11	22330	<b>Nepa Ig unj</b>	Chitwan	49519	44	6/26/2019	
12	22331	Kathmandu	Chitwan	75555	1984	6/26/2019	
13	<b>22332;</b>	Chitwan	Biratnagar	48880	40	<b>july 26th 2019</b>	
14	22333	<b>Pok</b>	Nepalgunj	22866	24	6/26/2019	
15	22334	Biratnagar	Pokhara	65166	31	6/26/2019	
16	<b>2233%</b>	Nepalgunj	Biratnagar	24917	28		
17	22336	Kathmandu	Nepalgunj	9472	37	6/26/2019	
18	22337	<b>Chitwan Kathmandu</b>		9891	22	6/26/2019	
19							
20							

# Cleaning Data with Workbench

The image shows the homepage of Workbench, a data analysis platform. The top navigation bar includes links for 'Blog', 'Sign in' (which is highlighted with a red box), and 'Sign up'. Below the header, a large central message reads 'Scrape, clean, analyze and visualize data without code.' with a 'Get started' button. On the left, there's a preview of a project titled 'Affordable housing crisis in San Francisco' by Pierre Conti, updated 2m ago - private. The project summary indicates it loads real estate data from SF open data, updates every day, and includes steps 1, 2, and 3. Step 1: 'Add from URL', Step 2: 'Convert to numbers', and Step 3: 'Group projects per neighborhood and calculate the amount of housing units in each of them.' To the right of the project summary is a bar chart titled 'San Francisco Real Estate Development' comparing 'Market rate units' (teal bars) and 'Affordable units' (pink bars). The chart shows data for three neighborhoods: Tenderloin, South of Market, and Financial District.

Neighborhood	Market rate units	Affordable units
Tenderloin	~5,700	~1,000
South of Market	~5,300	~1,000
Financial District	~4,000	~1,000

<https://workbenchdata.com/>

Module 4

# **Strengthening your Data Reliability**

## Scenario

Here, you will use google sheet to investigate the prevalence and treatment of diarrhea among children in Nepal. You will Identify:

- Q1: Which age group are most affected by diarrhea in Nepal?
- Q2: Which Province in Nepal has the highest number of children living with diarrhea?
- Q3: What is the relationship between the level of Education of the mothers and the no of affect Children with this disease?
- Q4: What gender is most affected by this disease?

Nkechicoker@gmail.com

# Task 1: Get Data

1. Open Prevalence and Treatment of Diarrhea in Nepal
2. Create a Duplicate of your original data.
3. Name your sheets.

The screenshot shows a Google Sheets spreadsheet with the title "Background characteristic". The first row contains headers: "Background", "Percentage w/ Number of ch", "Percentage fr Number of children with diarrhea". The second row contains "Age in months" and "children under age 5 with c". Subsequent rows show data for different age groups and their corresponding percentages. A context menu is open at cell B13 ("Source o"). The menu options include: Delete, Duplicate (highlighted with a blue arrow), Copy to, Rename, Change colour, Protect sheet..., Hide sheet, View comments, Move right, and Move left.

Background characteristic					
Background	Percentage w/ Number of ch	Percentage fr Number of children with diarrhea			
Age in months	children under age 5 with c				
<6	6	445	-67.6	27	
6-11	15.2	499	52	76	
12-23	9.9	1,034	77.2	102	
24-35	6.5	919	81.8	60	
36-47	6.2	968	48.9	60	
48-59	4.5	1,021	-52.2	46	
Sex					
Male	7.7	2,563	71.9	197	
Female	7.5	2,324	56.1	175	
Source o					
Improved		4,648	64.2	354	
Not impro		239		17	
Toilet fac					
Improved		2,810	64.5	182	
Unimprov		2,077	64.4	189	
Shared fa		923	73.6	74	
Unimprov		81		8	
Open defu		1,072	62.5	107	
Residenc					
Urban		2,649	59.8	207	
Rural		2,238	70.2	165	
Ecologic					
Mountain		342		18	

Data source: [National Demographic and Health Survey Data - 2016](#)

## Task 2: Clean data Using the Six Steps:

### Hint

1. Copy and duplicate your sheet
2. Understand your data and make a note of the rows and columns that may be most helpful and are inline with the lines of enquiries
3. Standardize the data formats
4. Unmerge cells where necessary
5. Fix your header and rows, rename ensure that your naming reflects the true content of the data
6. Rearrange and remove redundant columns and rows
7. Check maths

# Task 3: Analyse data to answer the following questions

- **Q1: What age group is most affected in Nepal?**

**Hint:** To answer the first question, you need to filter the data to see the age group of children most affected by diarrhea. To do this: **Filter** out the data of the age group and sort the values from A-Z

- **Q2: Which Province in Nepal has the highest number of children living with diarrhea?**

**Hint:** The procedure to answer the second question is similar to how you answered the first question. To do this: You need to filter the data to see the **province** and not **Age**. This shows you Province that the highest No of children with the disease at 112 Children.

# Task 3: Analyse data to answer the following questions

## **Q4: What gender is most affected by this disease?**

**Hint:** The procedure is similar to the last question – just that in this case you will filter the data and transpose it.

Sometimes it is helpful to transpose your dataset for better insight.

1. Filter out the data by gender.
2. Copy the selection of the filtered data,
3. Go to a new section of your sheet or open a new sheet, right click and paste special. on the drop down menu, click transpose
4. This will transpose the data and now you can sort from A-z.

# Lab: Analyse data

- Open [\*\*Migration Data- Nepal Labour Force Survey 2017/2018\*\*](#)
- Open the sheets
  - Distribution of migrants (all ages) by sex and province in which they currently reside
  - Age and sex distribution of migrants by current location

Now use the skills learnt so far, to clean and answer the following question:

- Which of the Provinces have the highest migration and which gender migrates has the higher migration?
- What is the average number of migrants by sex?
- Which age group has the highest migration of each gender?
- Analyse the migration distribution of people from 60+ age group



*Thank You*

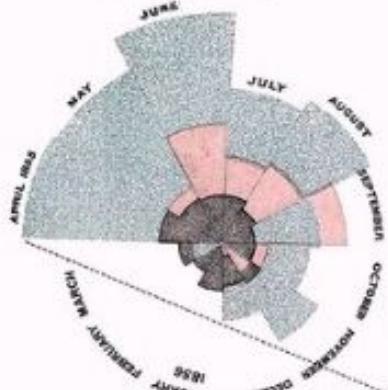
Module 5

# **Finding Stories in Facts by Visualizing Data**

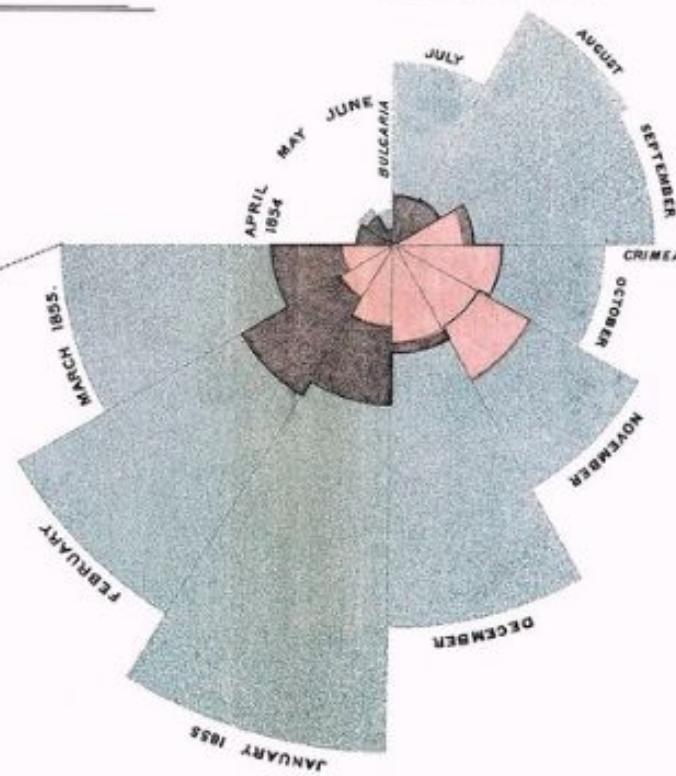
# **Part 1: Introduction to Exploratory Data Analysis**

2.  
APRIL 1855 to MARCH 1856.

DIAGRAM OF THE CAUSES OF MORTALITY  
IN THE ARMY IN THE EAST.



1.  
APRIL 1854 to MARCH 1855.



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

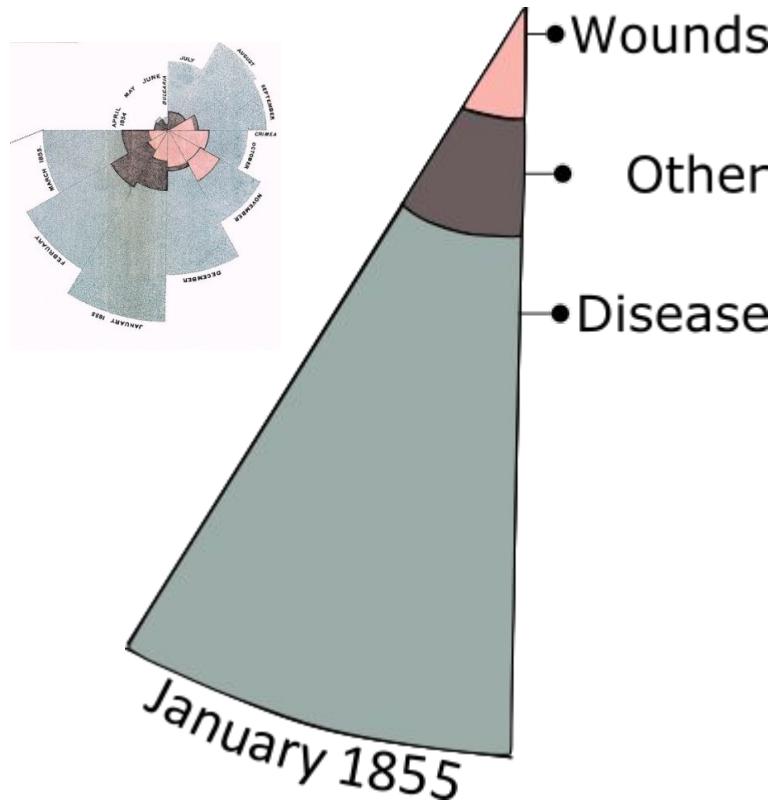
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventable or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes. The black line across the red triangle in Nov<sup>r</sup> 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1855, the blue coincides with the black.

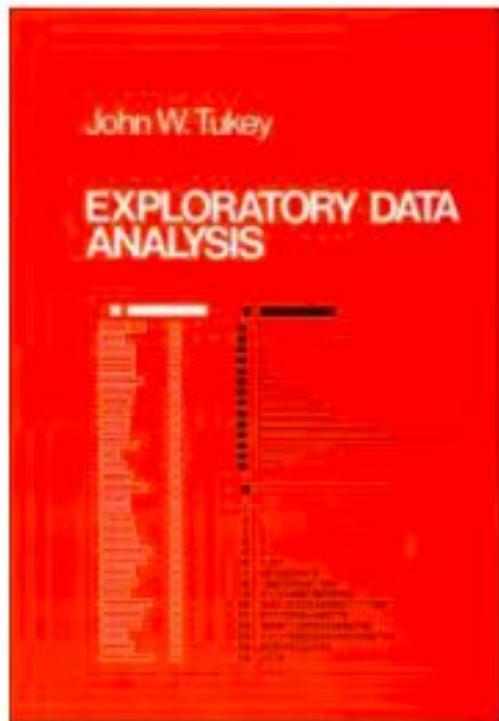
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

# What is EDA ?

- Disease, colored blue, killed far more soldiers than either “wounds” (red) or “other” (black)
- Reduced from 60% to 42% by February 1855
- From ~160 years ago ... shows the sheer impact of data, without all the technology behind it



# What is EDA ?



“**THE GREATEST VALUE  
OF A PICTURE IS WHEN  
IT FORCES US TO  
NOTICE WHAT WE NEVER  
EXPECTED TO SEE**”

“If data analysis is to be done well  
much of it must be a matter of  
judgement... theory will have to be  
GUIDE not COMMAND,”

# What is EDA ?

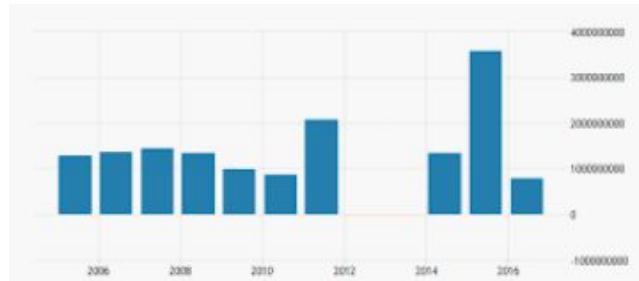
- An approach to analysing datasets to summarize their main characteristics
- A first look at data ...



# What is EDA ?

- A first look at data:

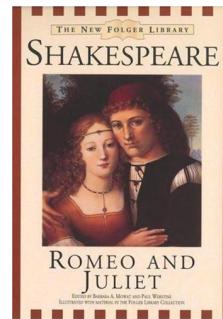
- Omissions
- Context
- Patterns
- Anomalies



# What is EDA ?

- A first look at data:

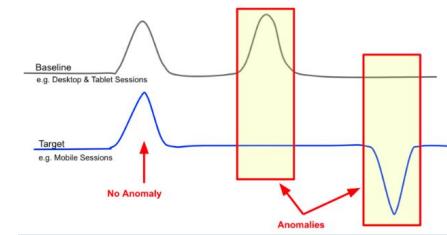
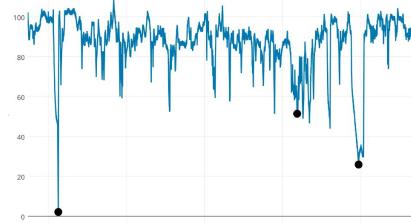
- Omissions
- Context
- Patterns
- Anomalies



# What is EDA ?

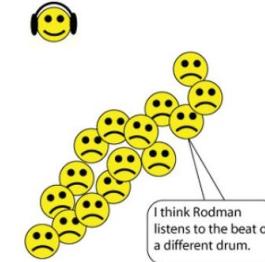
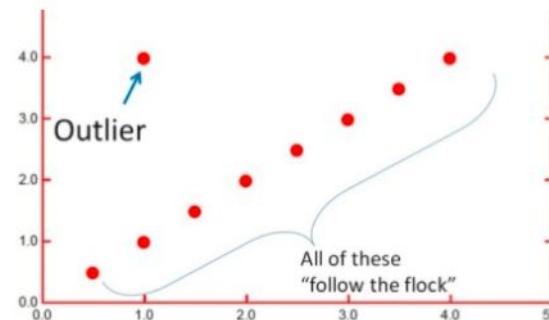
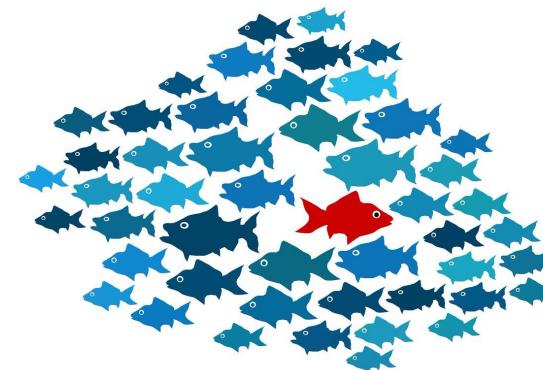
- A first look at data:

- Omissions
- Context
- Patterns
- Anomalies

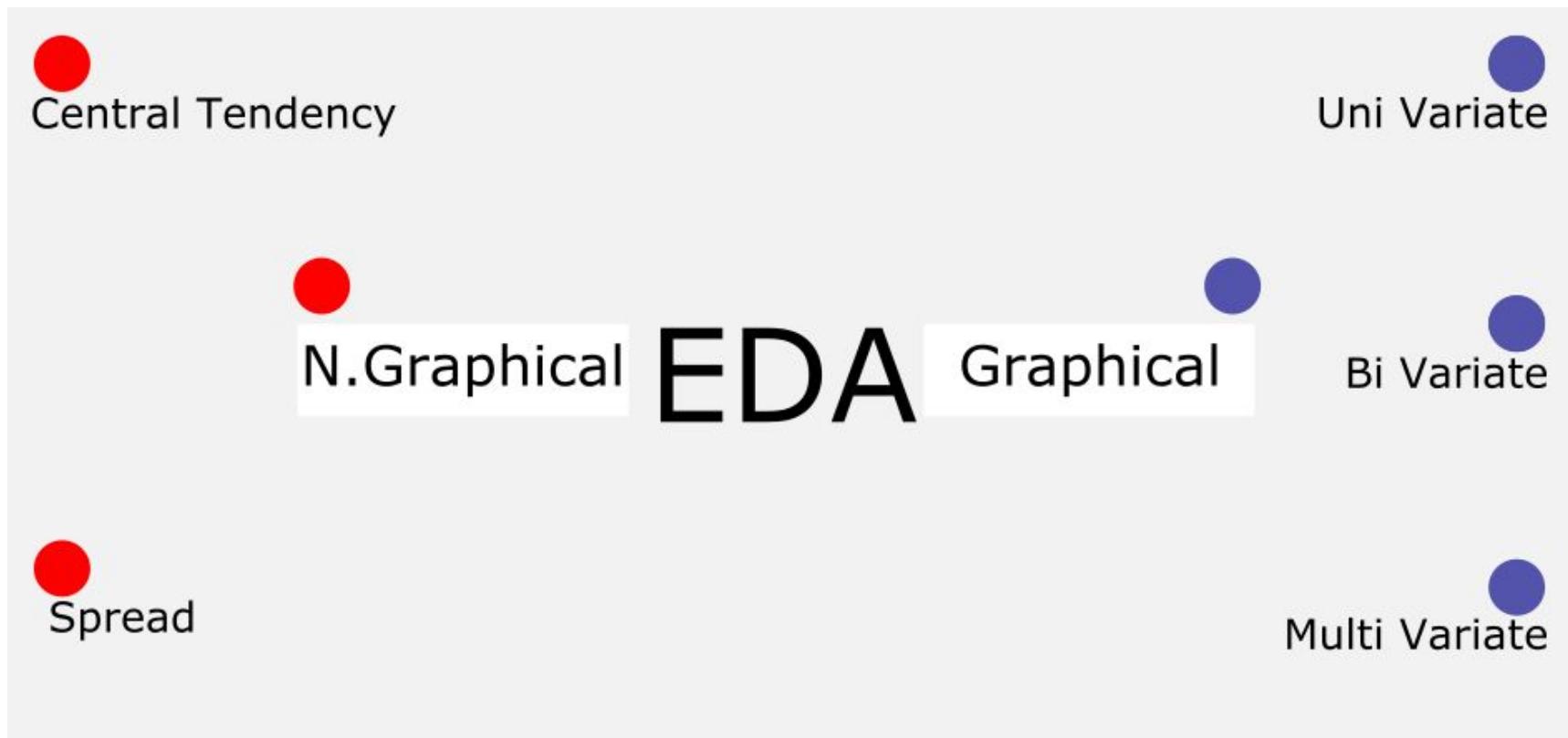


# Anomalies / Outliers

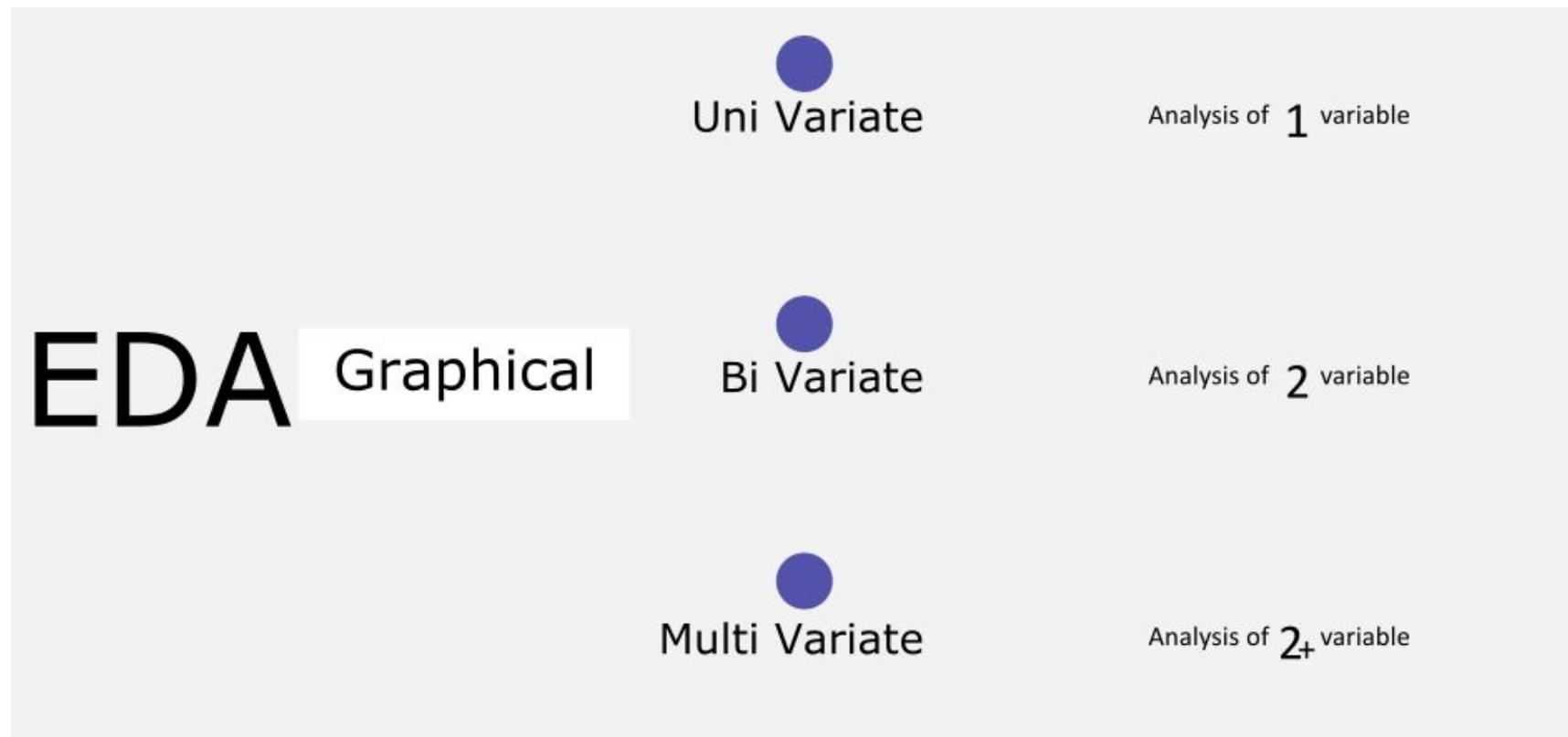
- An observation that diverges or differs significantly from an overall pattern in the data set



# Understanding EDA



# Understanding EDA



# Understanding EDA

QUALITATIVE



are

COUNTED \*



QUANTITATIVE



are

MEASURED \*



\* MOSTLY

# Understanding EDA

Determine Quantitative or Qualitative ?

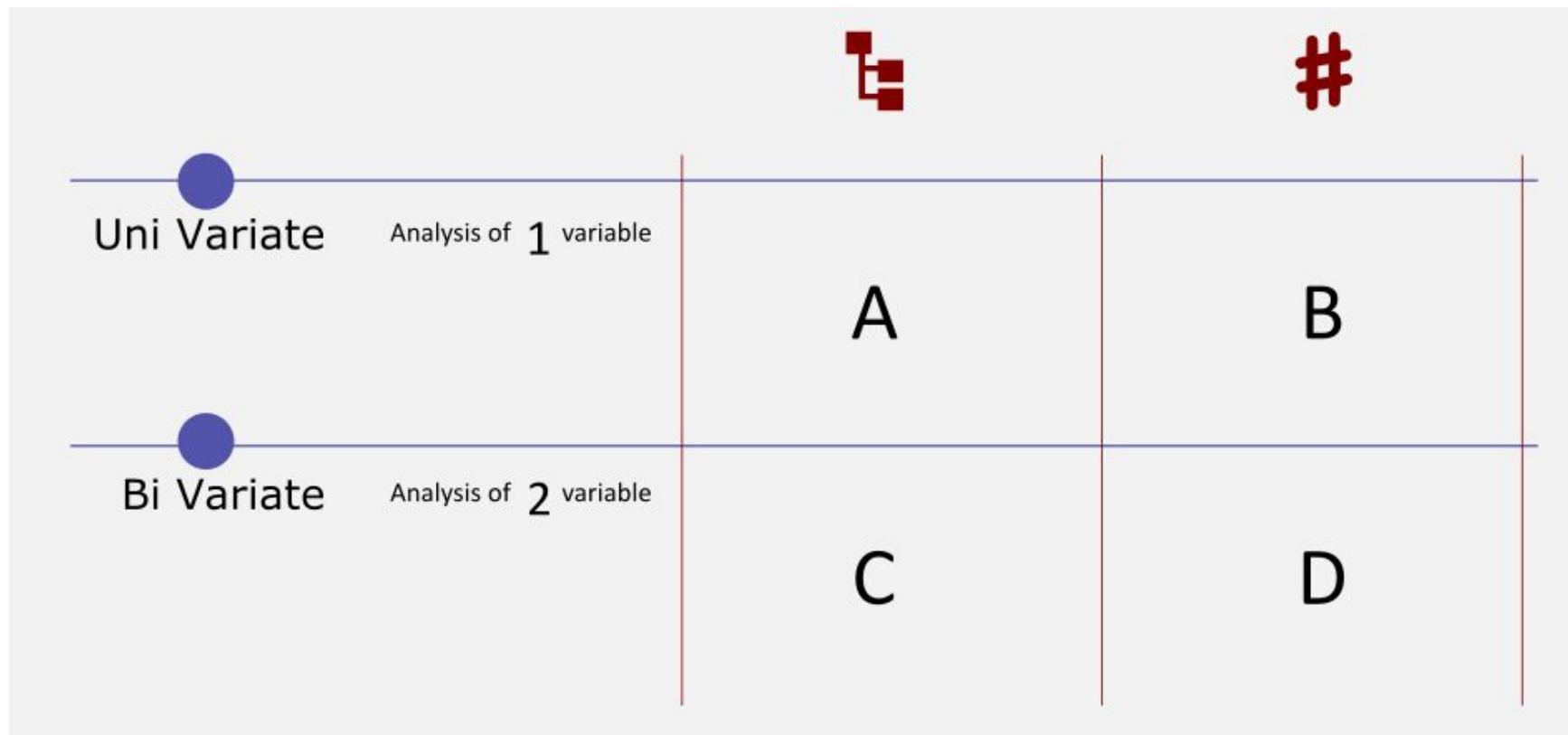
- Heights of Players in Nepal Football Team
- Hours of sleep on the previous weekend for WB Team
- Ravi's favorite ice cream
- Number of apps used by students of Kathmandu University
- Football Clubs represented by Lionel Messi

# Understanding EDA

Determine Quantitative or Qualitative ?

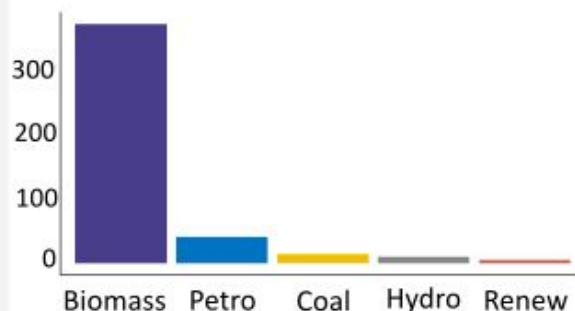
- Heights of Players in Nepal Football Team #
- Hours of sleep on the previous weekend for WB Team #
- Ravi's favorite ice cream
- Number of apps used by students of Kathmandu University #
- Football Clubs represented by Lionel Messi

# Understanding EDA

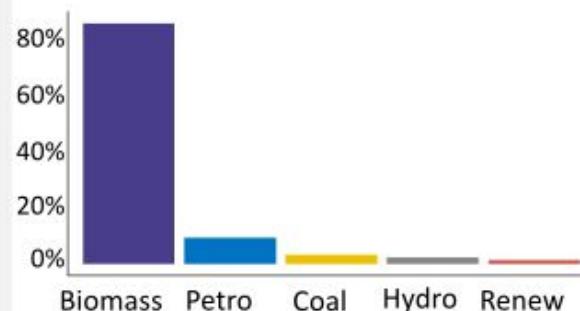


# Understanding EDA

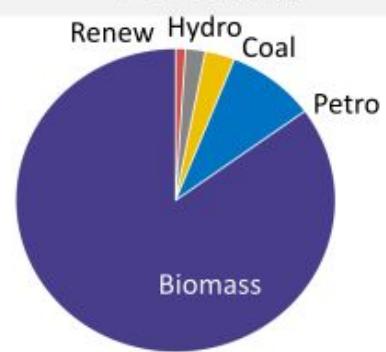
Bar Plot



Bar Plot



Pie Chart

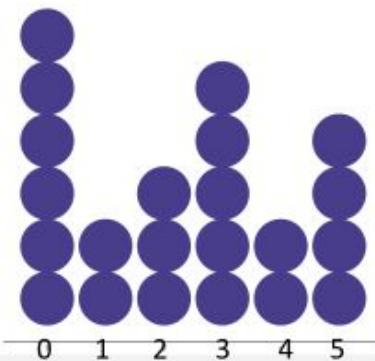


# Uni Var

Qualitative

# Understanding EDA

Dot PLOT

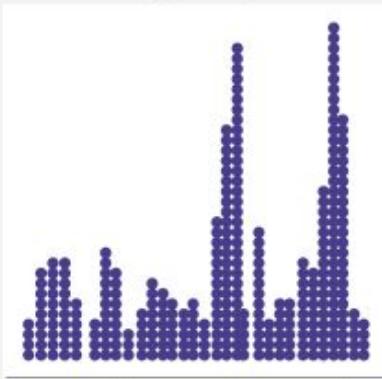


Minutes	0	1	2	3	4	5
People	6	2	3	5	2	4

Uni Var Quantitative

# Understanding EDA

Dot PLOT



Histogram



Histogram



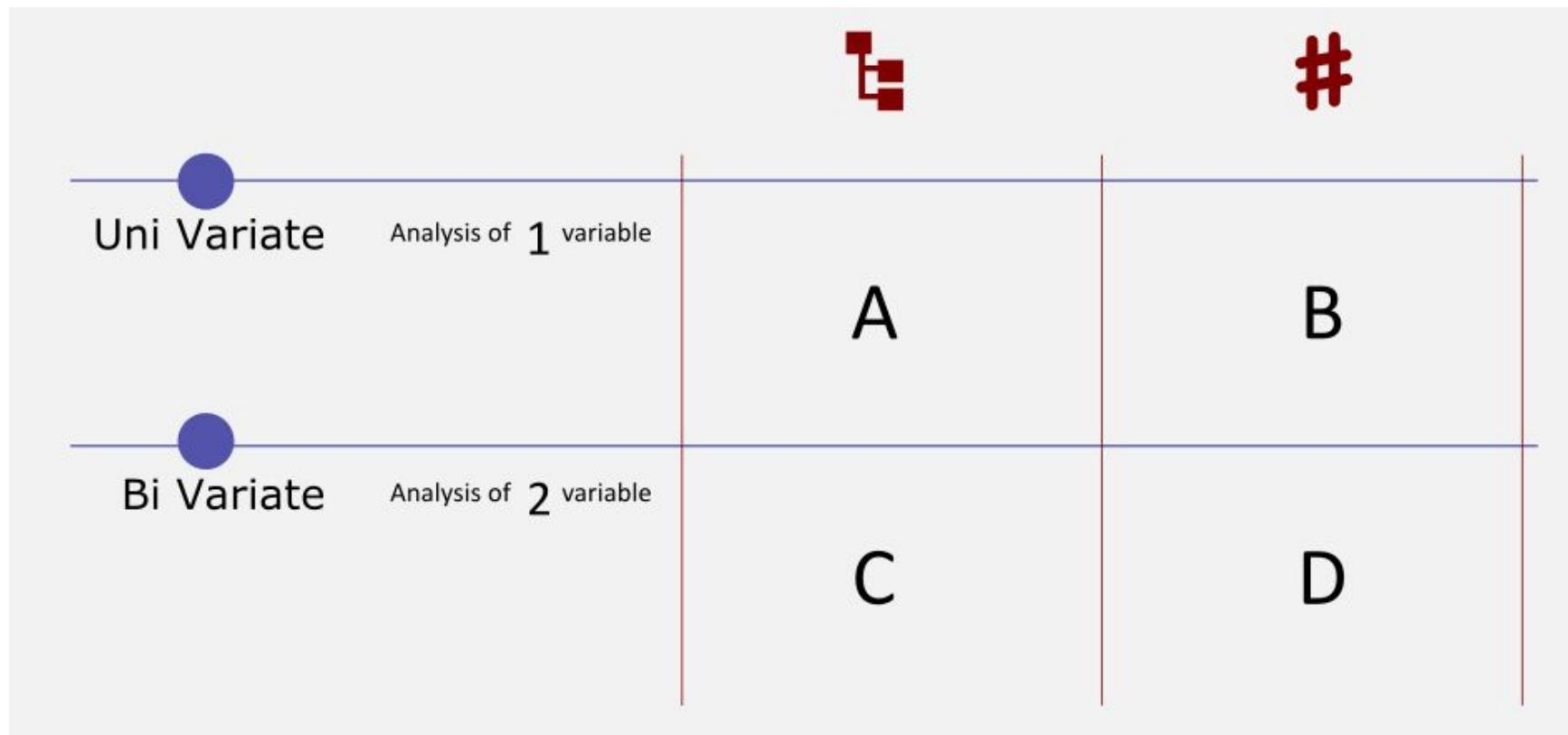
Histogram



Uni Var Quantitative

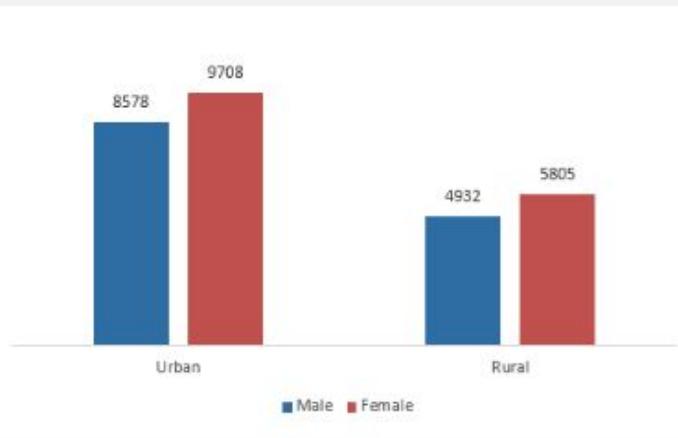
Age group (Years)	Nepal	
	Male	Female
Nepal	4703	1900
10-14	3	1
15-19	50	50
20-24	178	188
25-29	347	265
30-34	432	314
35-39	574	277
40-44	579	209
45-49	588	146
50-54	512	113
55-59	448	100
60-64	370	84
65+	622	153

# Understanding EDA

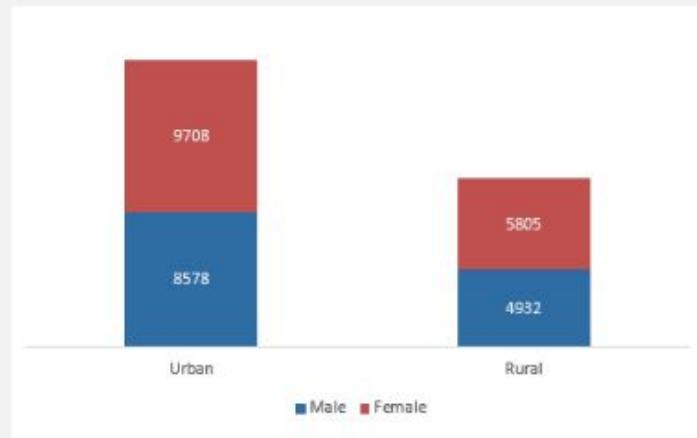


# Understanding EDA

Bar PLOT



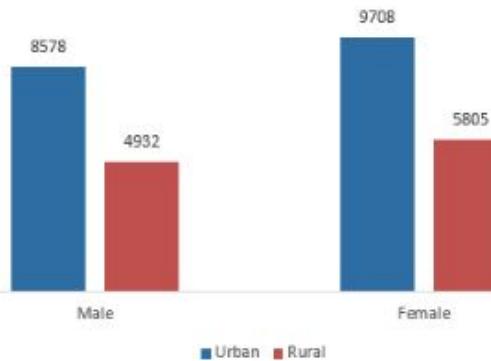
Bar PLOT



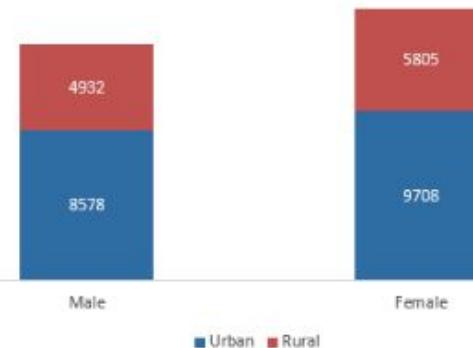
Bi Var Qualitative

# Understanding EDA

Bar PLOT

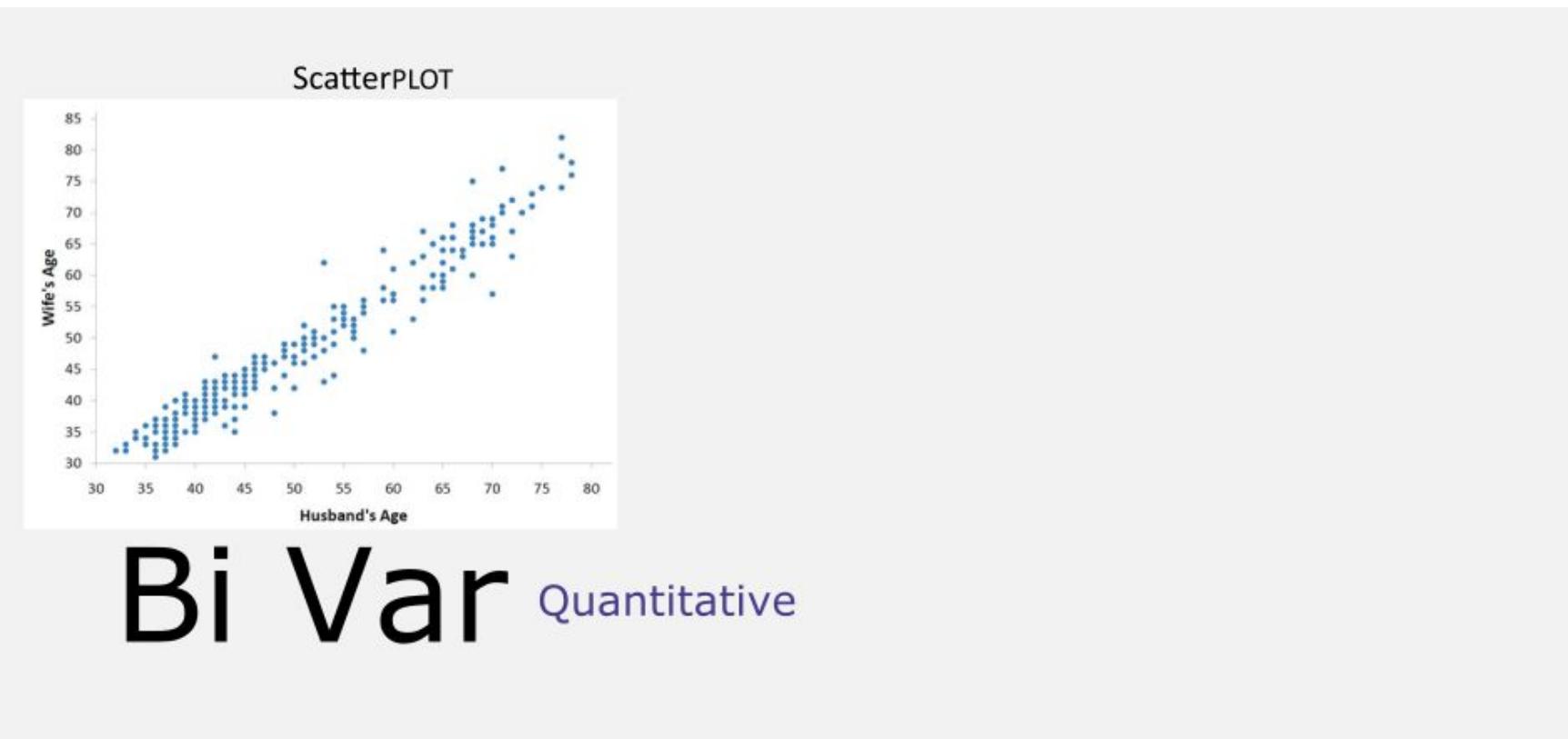


Bar PLOT

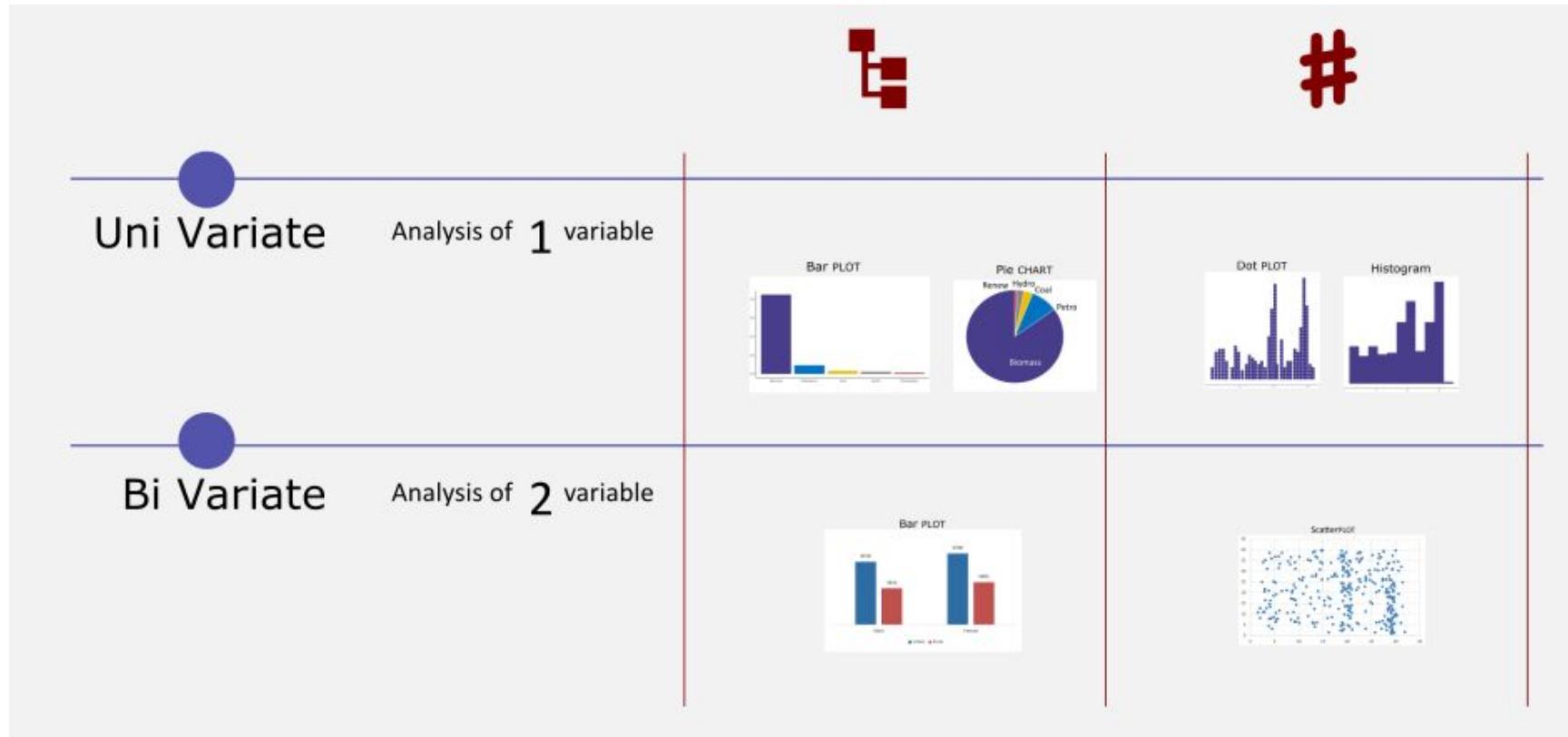


Bi Var Qualitative

# Understanding EDA



# Understanding EDA





*Thank You*

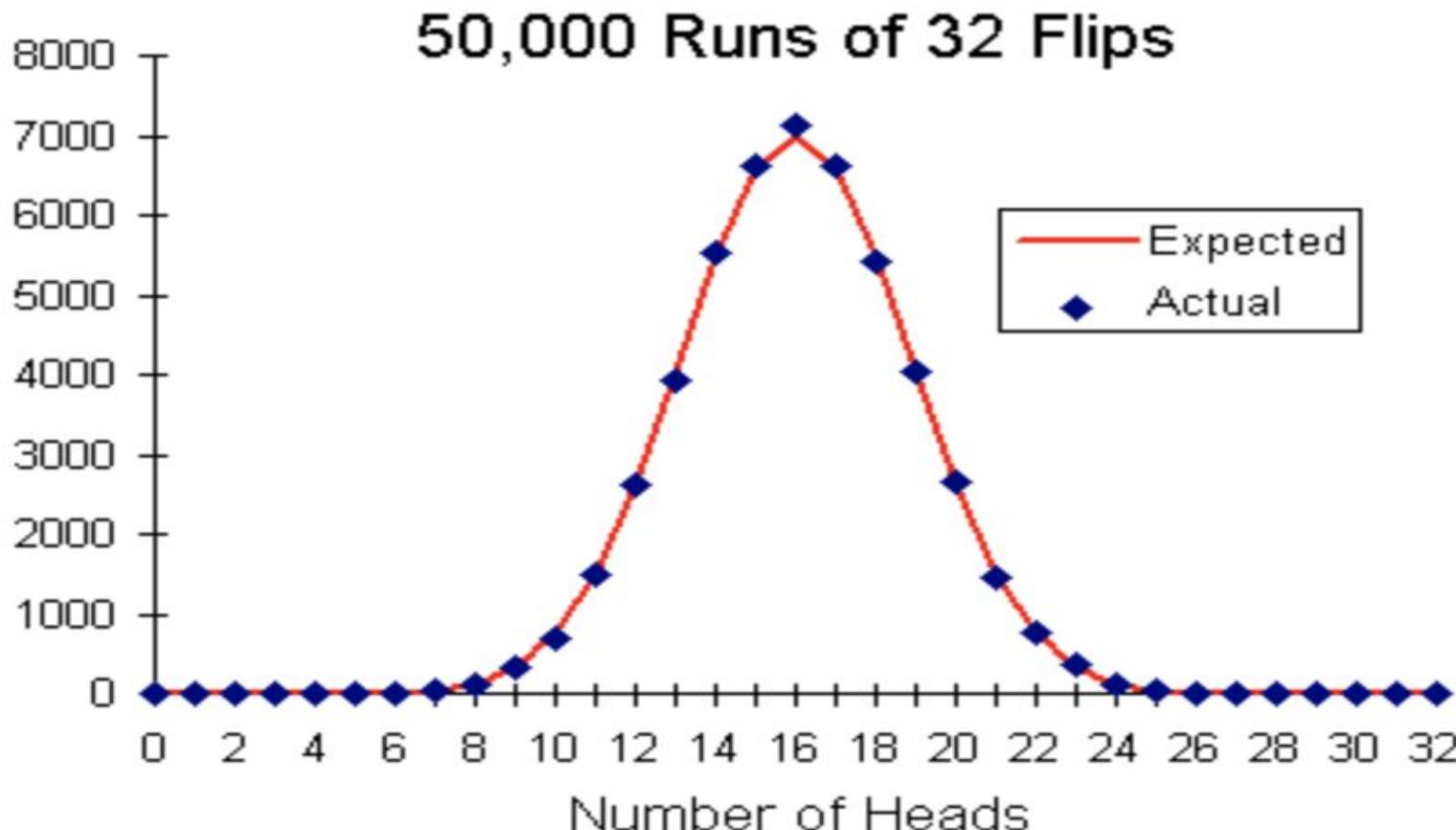
# Numerical Summaries of Data

- Central Tendency Measures
- Variability Measures

# Flipping a Coin Exercise



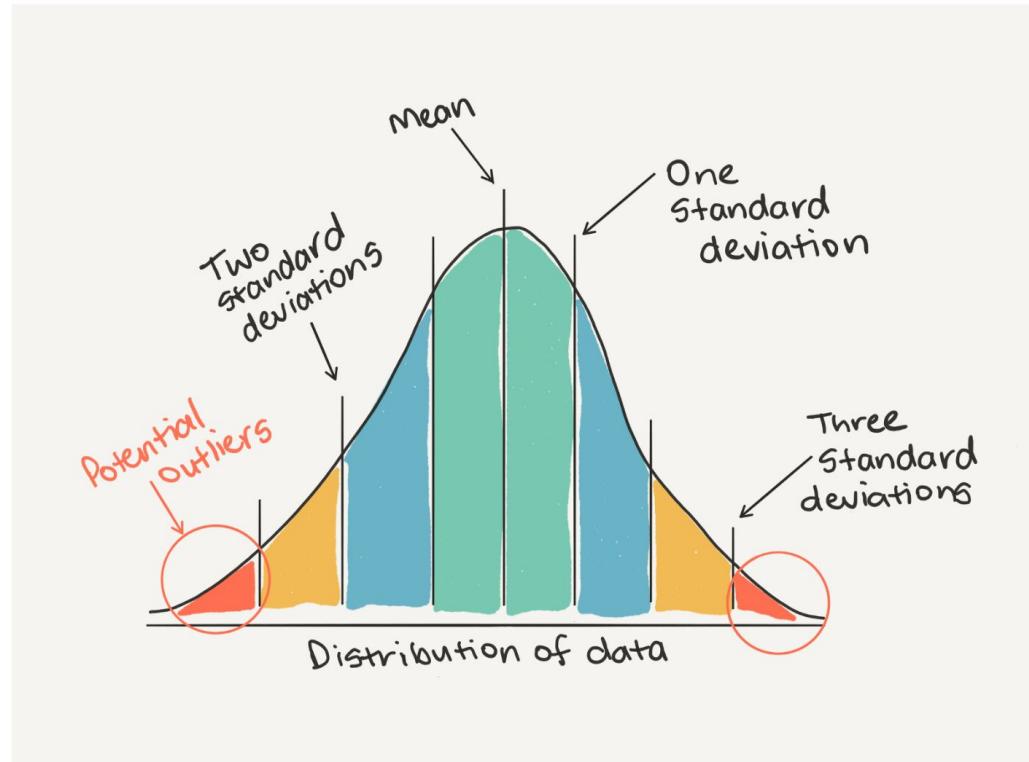
Let's flip 32 coins, count the number of heads. Repeat 50,000 times.



Source: <https://www.fourmilab.ch/rpkp/experiments/statistics.html>

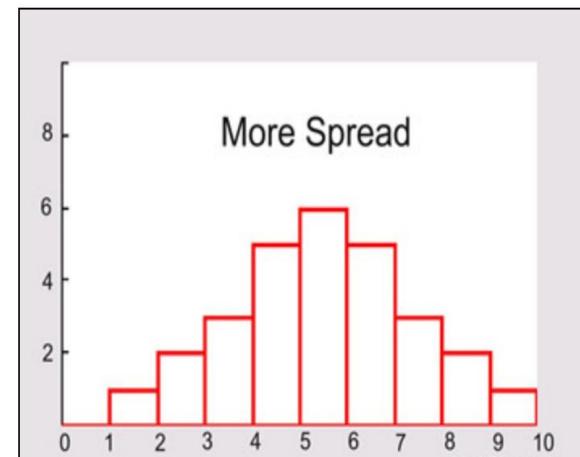
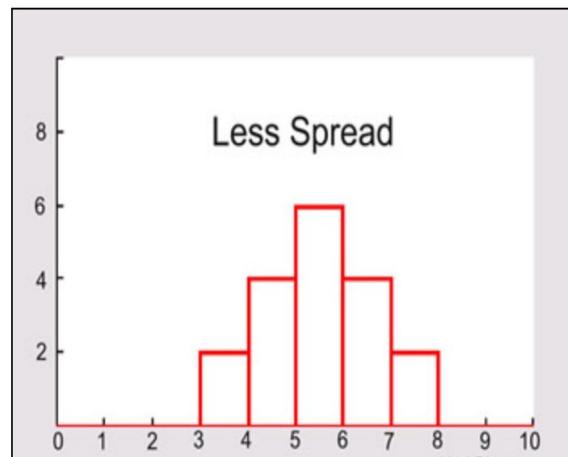
# Definition of terms: Central Tendency

- This is a value that describes where most of the set of a data falls or clusters. Mean, Median, Mode are used to measure this.



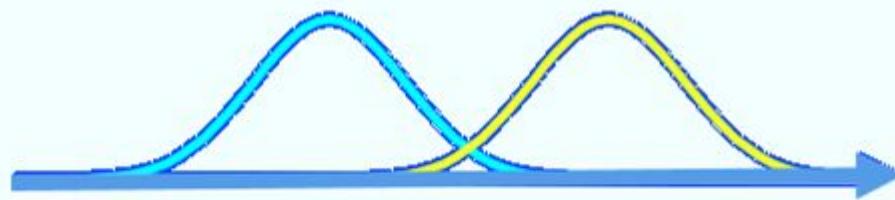
# Definition of terms: Spread

It is an indicator of how far away from the center/mean/median data values are and is measured by variance, standard deviation, and interquartile range



# Numerical Summaries of Data

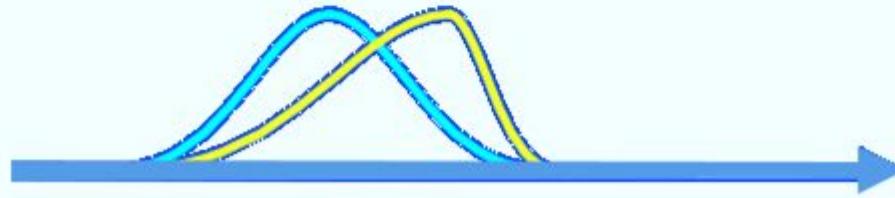
Central Tendency  
(LOCATION)



Variation  
(DISPERSION)



Shape



# Central Tendency Measures

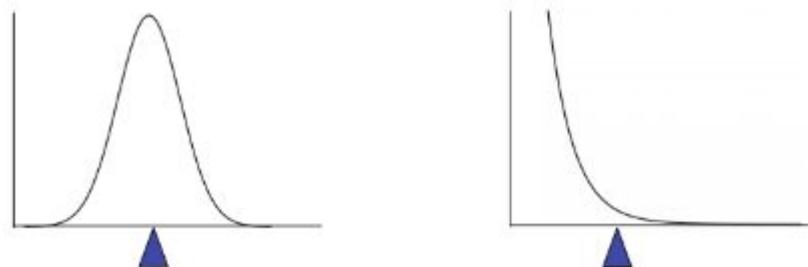
- Mean/Average
- Median
- Mode

# Mean/Average

- Sum of data points divided by Number of data points

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

- A balance point for the distribution



# Median

- Midpoint of the list if the measures are listed in ascending order
- The point on the measurement scale below which 50% of the score are located

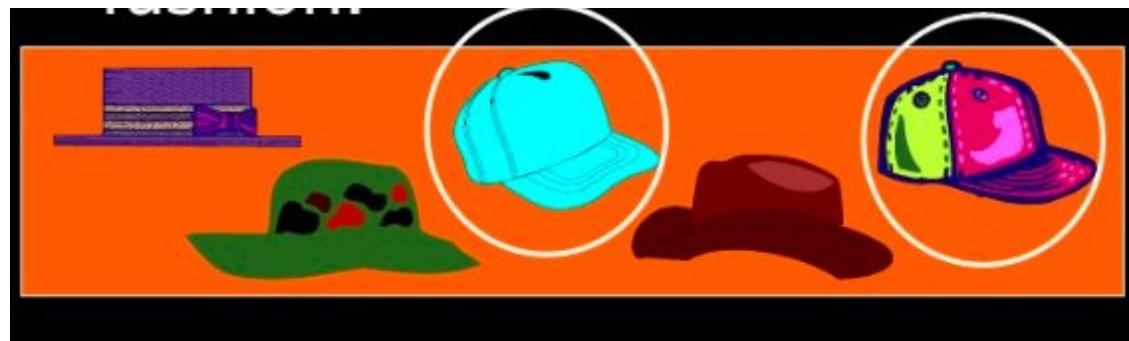
Age of participants: 17 19 21 22 23 23 23 38

$$\text{Median} = (22+23)/2 = 22.5$$

**NOTE: IF THE NUMBER OF DATA POINTS IS EVEN, THE MEDIAN IS THE MEAN OF THE MIDDLE TWO NUMBERS**

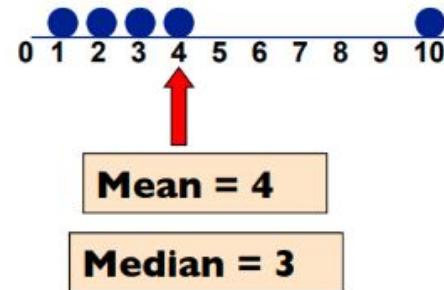
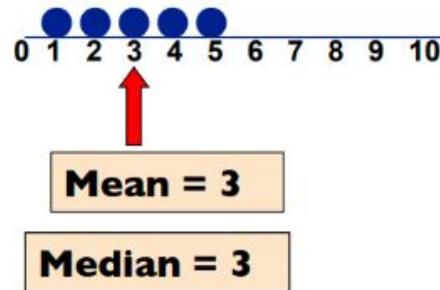
# Mode

- The value that has the greatest frequency
- It is not unique

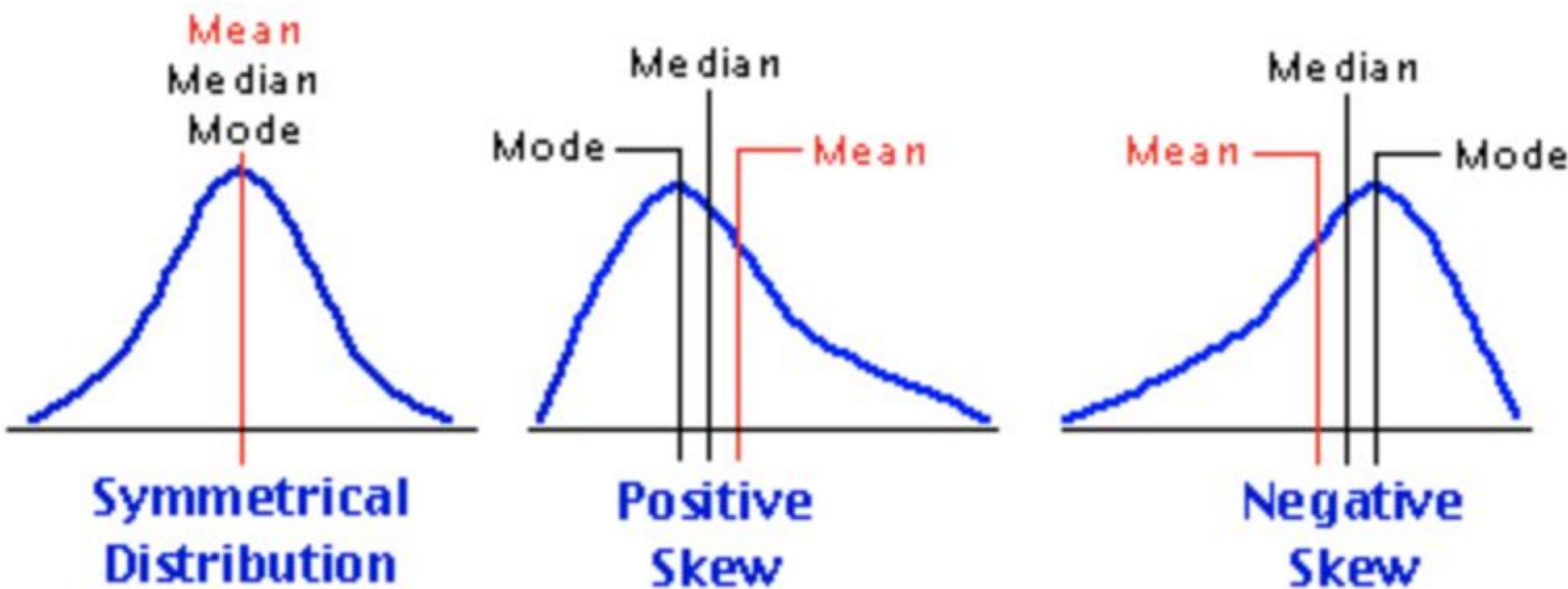


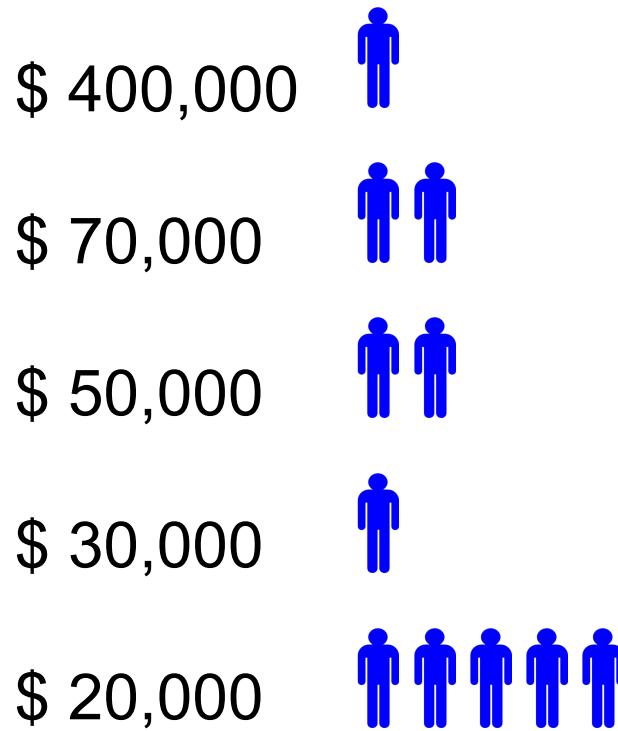
# Mean or Median ?

- MEAN is best for **symmetric distributions** without outliers
- MEDIAN is useful for **skewed distributions** or those with outliers



# Skewed Data





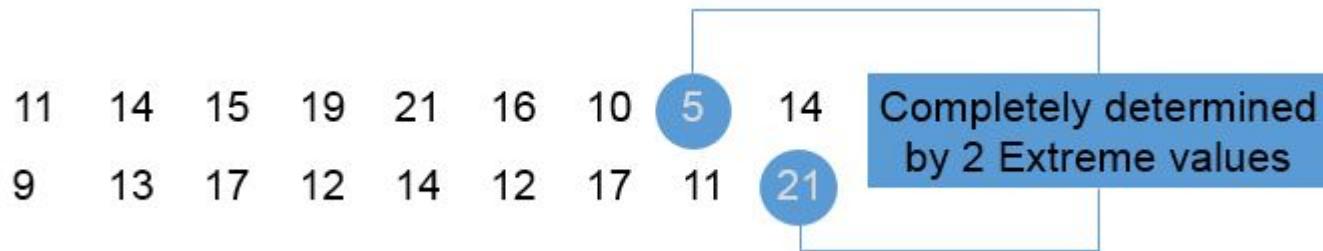
- Company board claims average pay is \$70,000
- Employees cite low pay - is it possible ?

# Variability Measures

- Range
- Percentiles
- Standard Deviation
- Variance

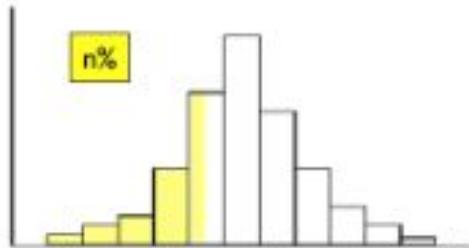
# Range

- Easy to compute and understand
- Non-Robust measure - influenced by extreme values
- Minimum value of the data set
- Maximum value of the data set



# Percentiles

- In general, the **n<sup>th</sup> percentile** is a value such that **n%** of the observations fall below it



$Q_1 = 25^{\text{th}}$  percentile

Median =  $50^{\text{th}}$  percentile

$Q_3 = 75^{\text{th}}$  percentile

$Q_1 := 25^{\text{th}}$  percentile

**25% of data fall below this value**

$Q_2 := 50^{\text{th}}$  percentile

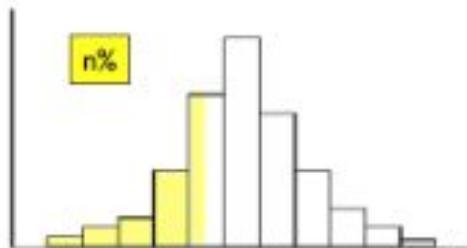
**50% of data fall below this value**

$Q_3 := 75^{\text{th}}$  percentile

**75% of data fall below this value**

# Percentiles

- In general, the **n<sup>th</sup> percentile** is a value such that **n%** of the observations fall below it



$Q_1 = 25^{\text{th}}$  percentile

Median =  $50^{\text{th}}$  percentile

$Q_2 = 75^{\text{th}}$  percentile

$Q_1 := 25^{\text{th}}$  percentile

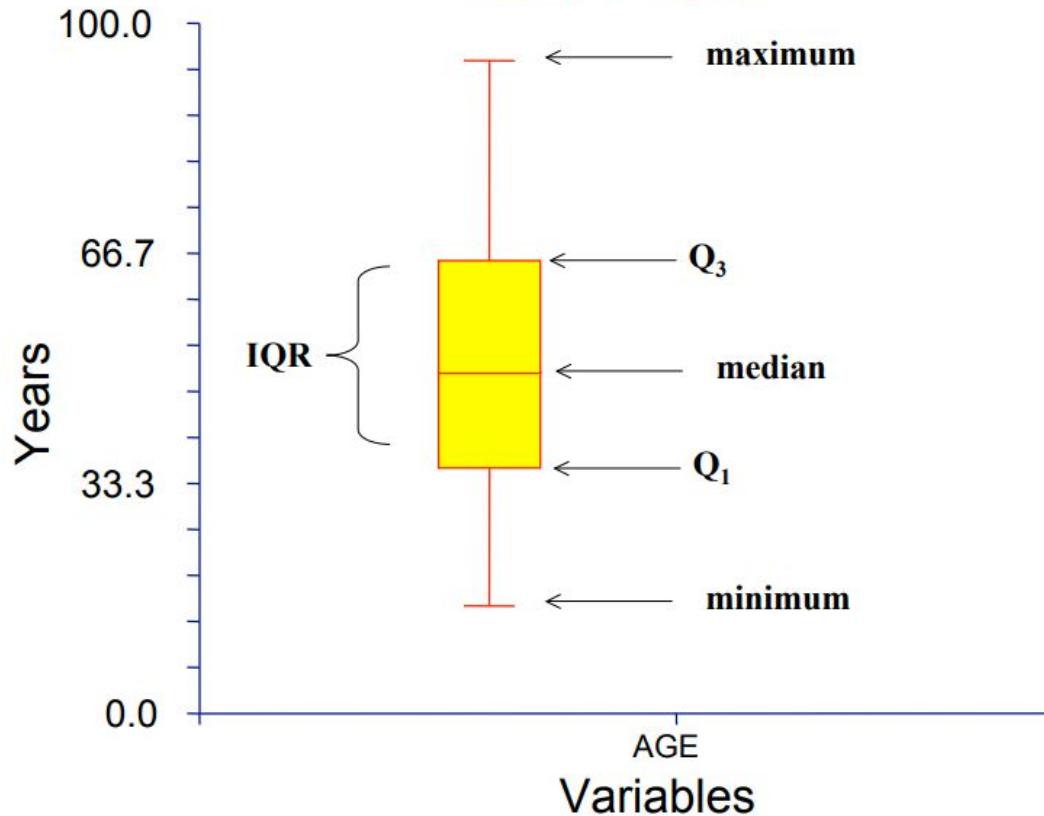
**Median to the left of the median**

$Q_2 := \text{MEDIAN}$

$Q_3 := 75^{\text{th}}$  percentile

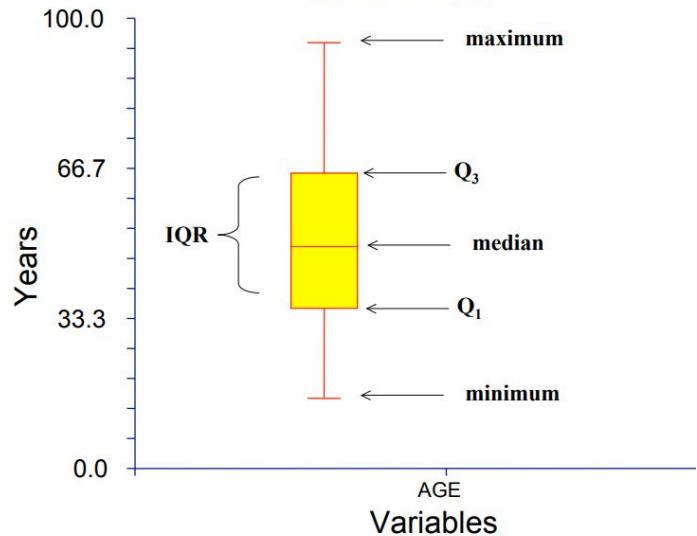
**Median to the right of the median**

# Box Plot

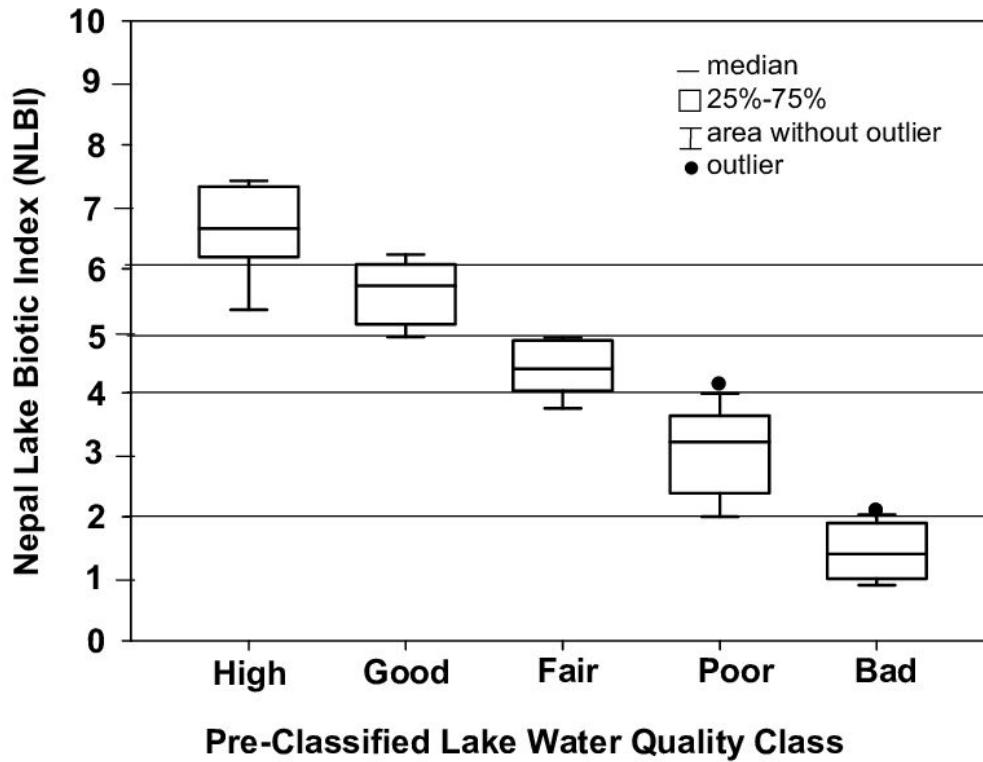


# Box Plot

1. Range
2. Median
3. Q1
4. Q3



# Box Plot



# Box Plot

- Make a box plot for the following data:
  - Minimum: 1
  - Q1: 7
  - Median: 10
  - Q3: 17
  - Maximum: 30
  - n: 100

# **Part 2: Visualizing Data**

# Data Visualization

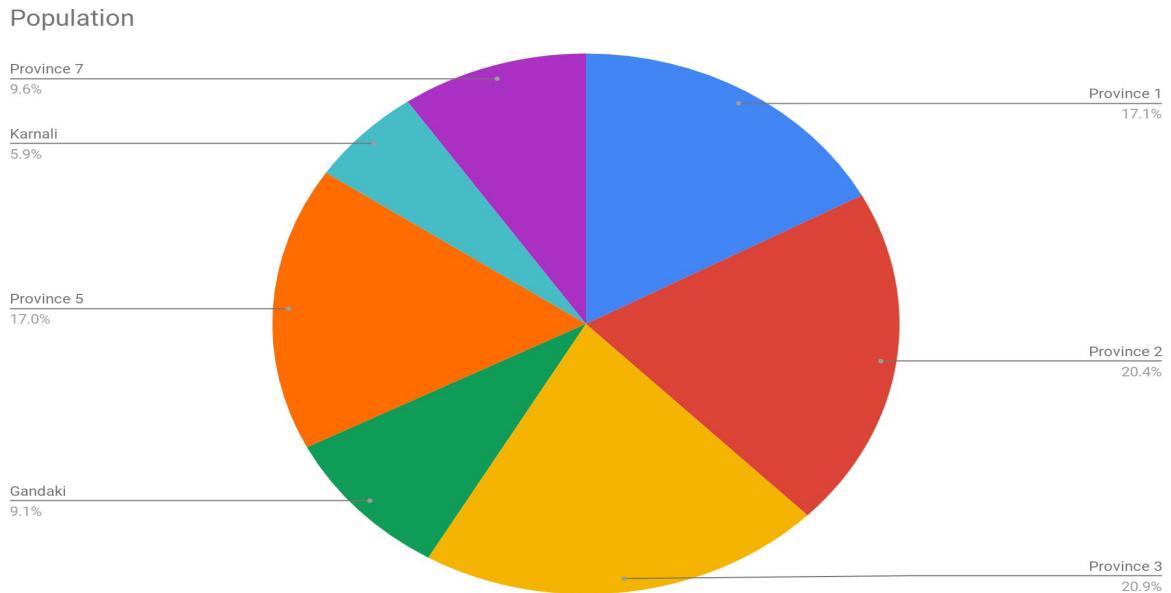
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. here are many kinds of chart and plots that helps us understand data better. Here are a few examples:

- Line charts
- Scattered plots
- Bar charts and Stacked bar charts
- Pie charts
- Histograms
- Infographics and Maps

**Pie Chart:** A chart that uses pie slices to show relative sizes of data, percentage or proportion. They are used for categorical data

SN	Province	Population
1	Province 1	4,534,943
2	Province 2	5,404,145
3	Province 3	5,529,452
4	Gandaki	2,403,757
5	Province 5	4,499,272
6	Karnali	1,570,418
7	Province 7	2,552,517
	Nepal	26,494,504

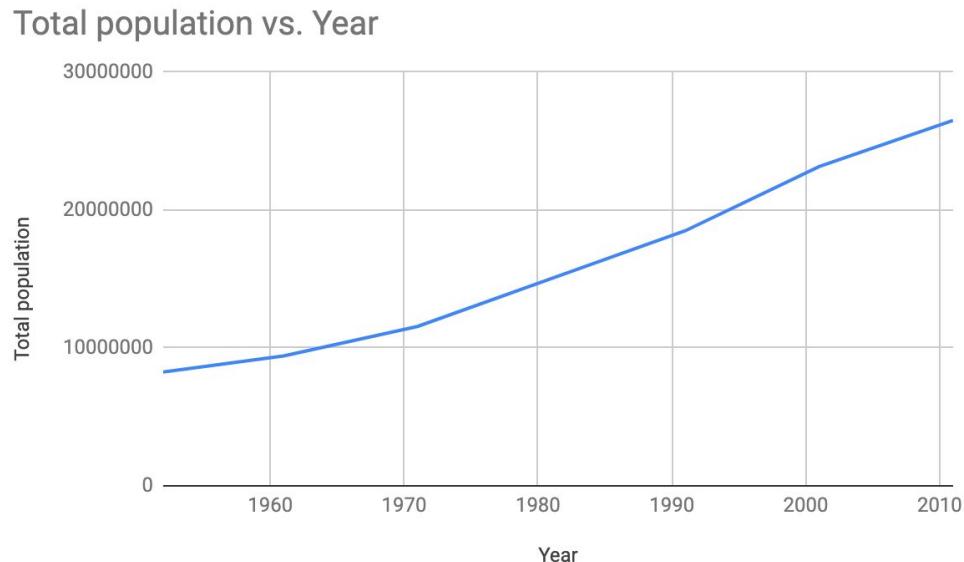
Source: CBS 2011



**Line Chart:** It is used to visualize changes and trends in data over time. It is mostly used for Numerical data

<b>Year</b>	<b>Total population</b>
1952	8256625
1961	9412962
1971	11555983
1981	15022839
1991	18491097
2001	23151423
2011	26494504

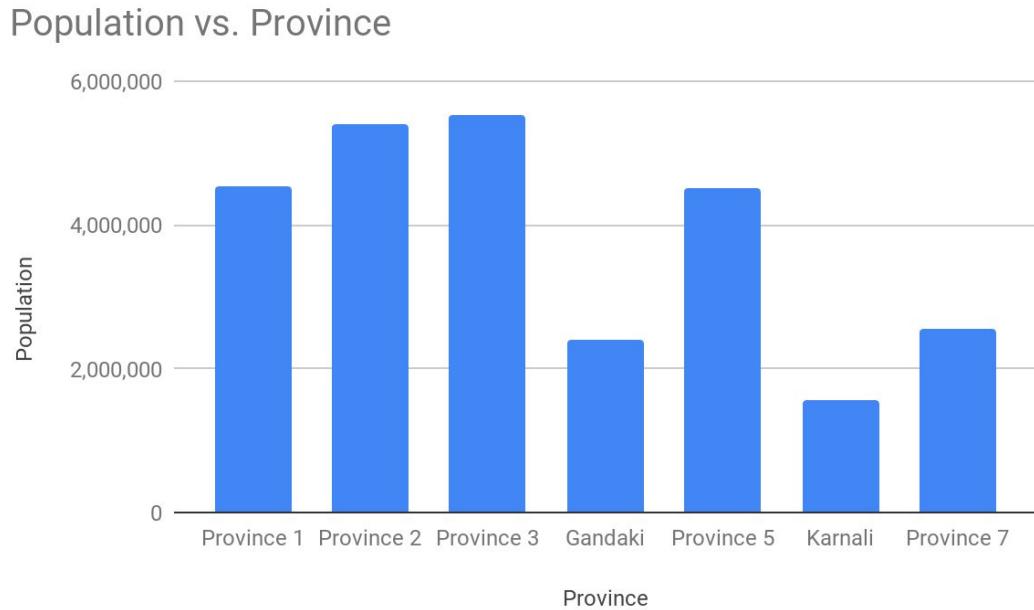
Source: Nepal in Data



# Bar Chart: A graphical display of data using bars of different heights used to show comparison in categorical data

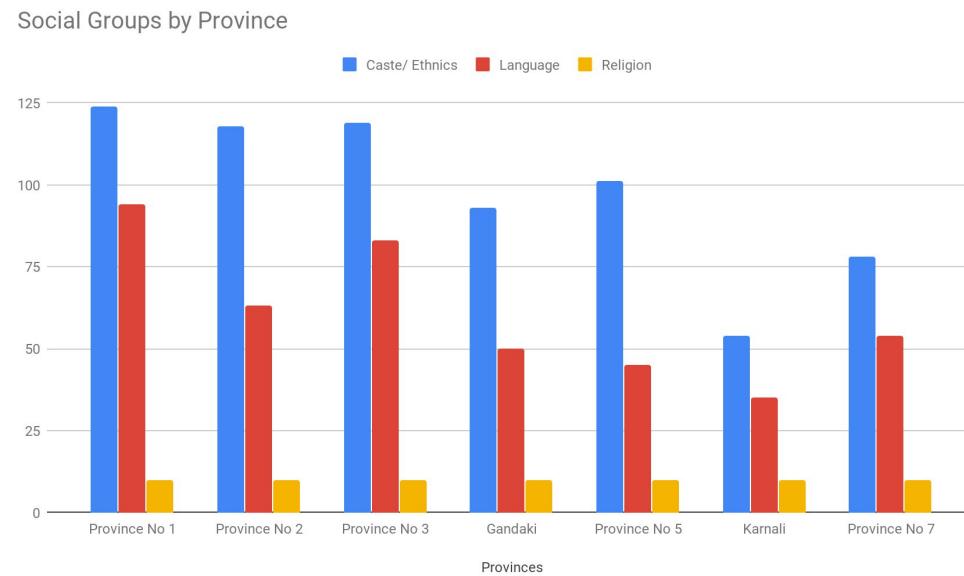
SN	Province	Population
1	Province 1	4,534,943
2	Province 2	5,404,145
3	Province 3	5,529,452
4	Gandaki	2,403,757
5	Province 5	4,499,272
6	Karnali	1,570,418
7	Province 7	2,552,517
	Nepal	26,494,504

Source: CBS 2011



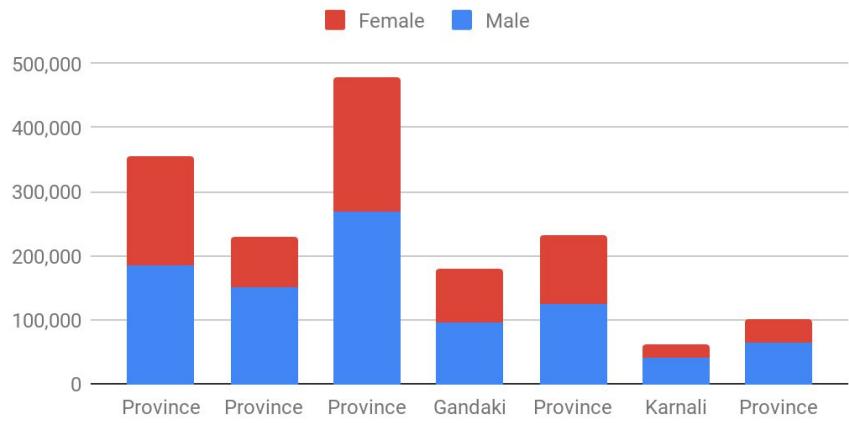
# Grouped Bar Chart: A graphical display of data using bars of different heights

Description	Nepal	Province No 1	Province No 2	Province No 3	Gandaki	Province No 5	Karnali	Province No 7
Caste/ Ethnics	125	124	118	119	93	101	54	78
Language	123	94	63	83	50	45	35	54
Religion	10	10	10	10	10	10	10	10
Source: CBS, 2011								



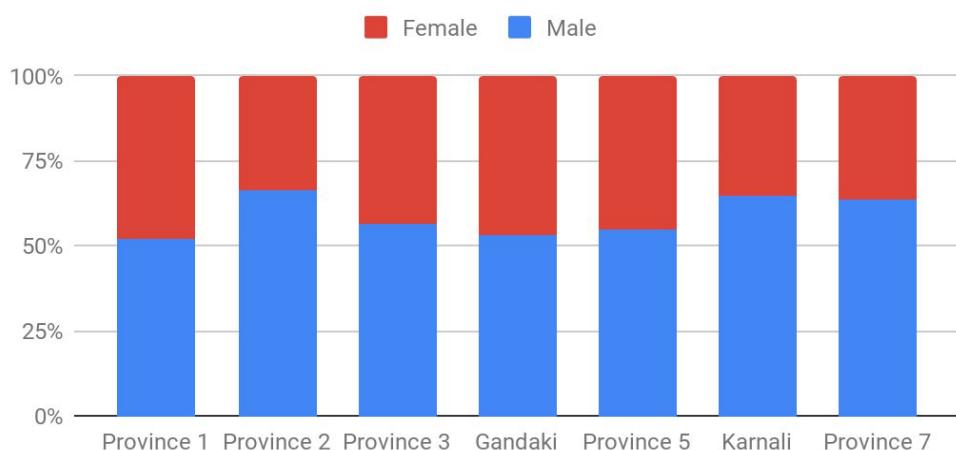
# Stacked Bar Chart: A bar graph with segments in each bar, and each segment represents different categories

Total population that have completed SLC or equiv.



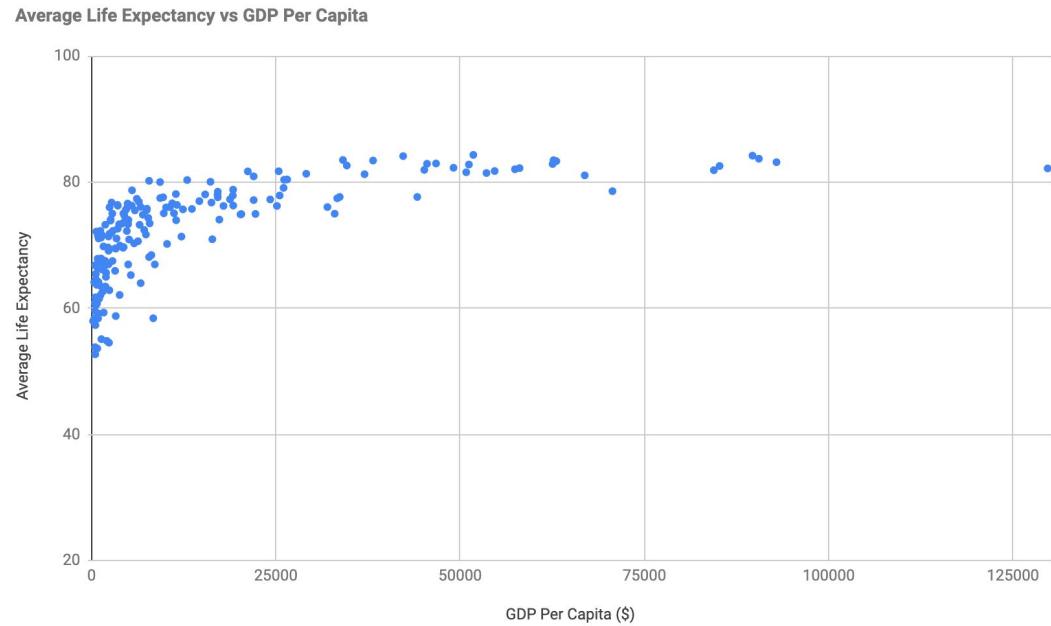
Source: [Education in figures 2017](#)

Percentage of population that have completed SLC or equiv.



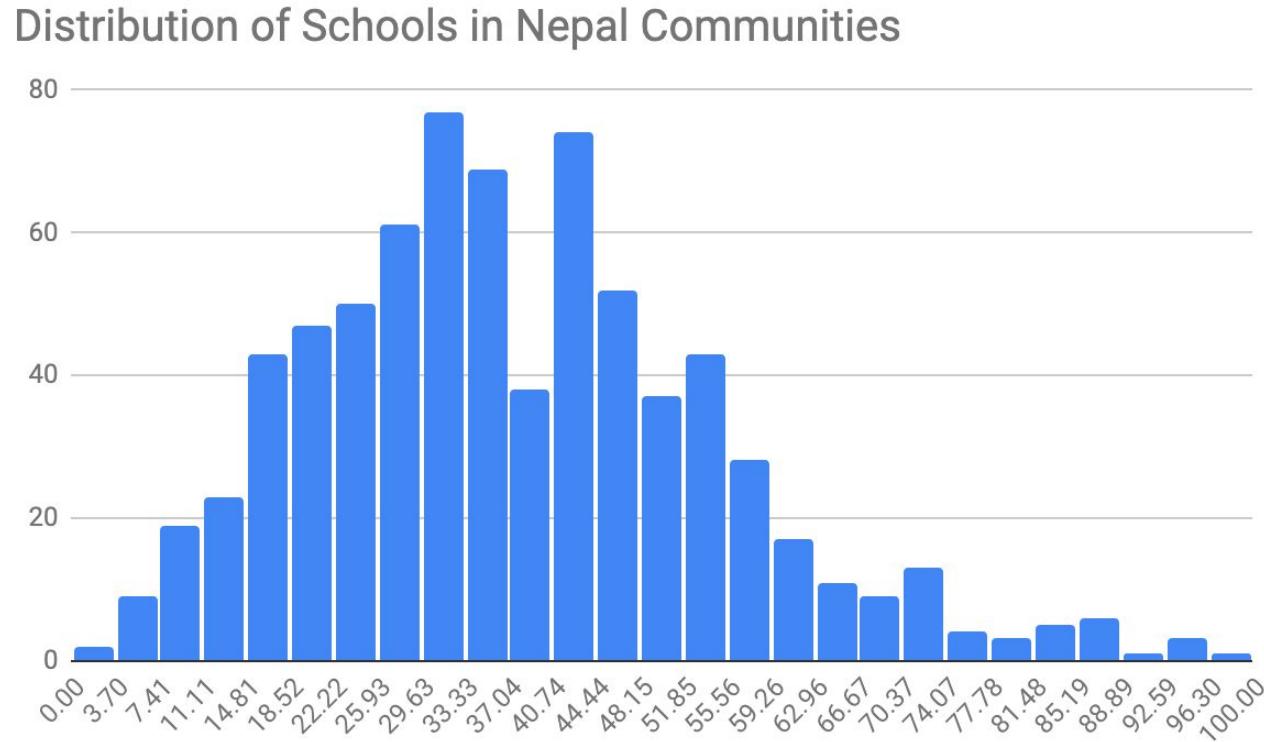
**Scatter Plot:** It is a two-dimensional data visualization that is used to find the relationship between two variables, often quantities.

Country	GDP Per Capita	Average Life Expectancy
United States	65058.27	79.772
China	10945.80	76.64
Japan	42270.55	84.118
Germany	53576.82	81.436
India	2305.21	69.118
France	46732.39	82.946
United Kingdom	45140.75	81.932
Italy	38189.68	83.416
Brazil	10625.84	76.018
Canada	51194.73	82.782



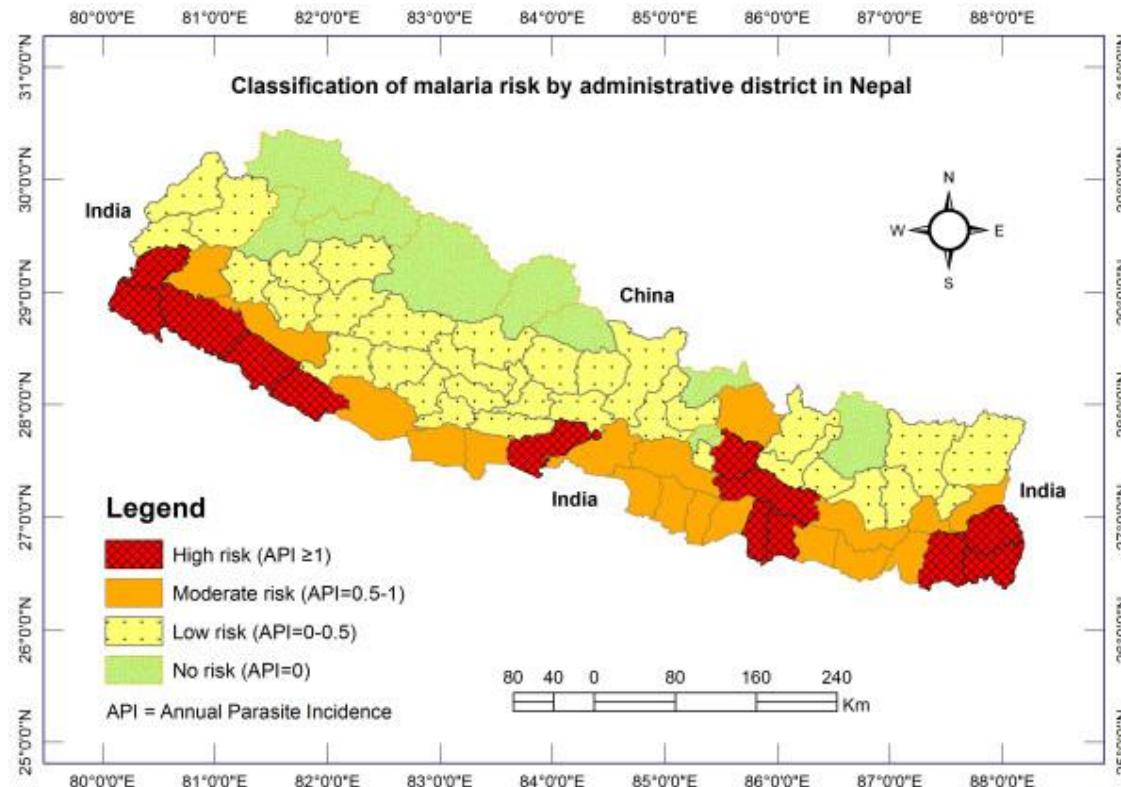
Data Source:  
[worldpopulationreview.com](http://worldpopulationreview.com)

Histogram: Used for numeric data in which each bar represents the frequency (count) allowing the inspection of the data for its underlying distribution



[Data Source: Education in figures 2017](#)

Map: This uses spatial data and geographic data to visualize data across several locations at a glance



# Lab 6: Create a Visualization in Google Sheets

1. Open your cleaned dataset - [\*\*Prevalence and Treatment of Diarrhea in Nepal\*\*](#)
2. Look through the sheets and **create charts showing**
  - a. **Q1** Which age group are most affected by diarrhea in Nepal?
  - b. Q2: Which Province in Nepal has the highest number of children living with diarrhea?
  - c. Q3: What is the relationship between the level of Education of the mothers and the no of affect Children with this disease?
  - d. Q4: What gender is most affected by this disease?

## **Hint:**

1. Open dataset in google sheets and review the data.
2. Rearrange columns and rolls where necessary, highlight the **columns**.
3. Select **Insert**  **Charts** and select **2-D Column**  **Clustered Column** chart or pie chart A chart window displays on the spreadsheet.
4. Rename the chart where necessary

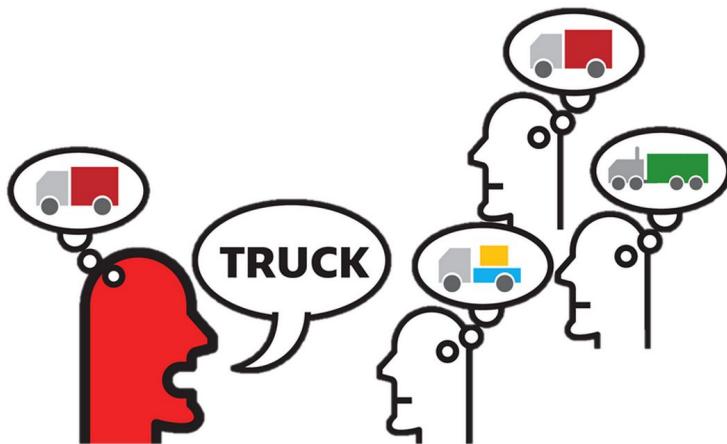
# Evaluation of Each Other's Visualizations

- Does the visualization tell a story or reveal an insight?
- Does the headline tell the reader what the story or trend is?
- Does the chart type match the data?
- Are there simple, clear fonts, colors and labels
- Did you cite the source of the data?

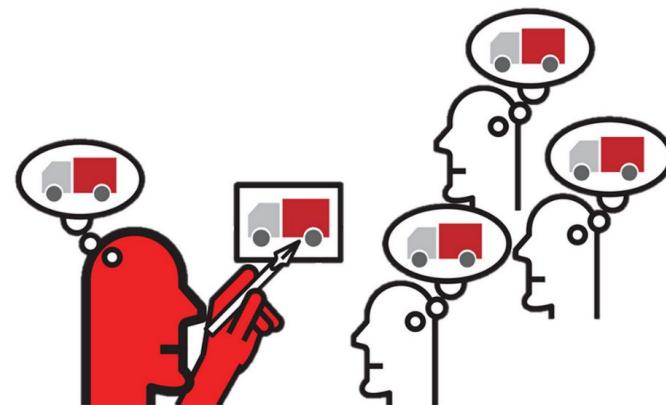
# **Principles and Ethics of Data Visualization**

# Styles of communication

## Verbal Communication



## Visual Communication



# Why do we use data visualization?

1. It gives a **clearer understanding** of your data or story.
2. It **amplifies** the message you want to get across.
3. It can aid **quick decision** analysis
4. You can **identify trends** quickly.
5. You can **share easily** with others.

# Staying Focused

- **What's the point?**

Make sure to include only the data you want the reader to remember

- **Simplify the numbers!**

When possible, reduce numbers to simplest form.

- **Make the angle clear**

Titles and labels should be specific, easy to understand and true to the data, including citing the source of the data

- **Set the scene!**

Styles and colors should aid understanding, not distract readers

# Basic Design Concepts

- **Simplicity:** Choose a maximum of three colors and fonts and stick with them consistently.
- **Brevity:** Keep text short and to the point.
- **Creativity:** Incorporate playful design that relates to the topic.
- **Two dimensions:** Avoid 3D graphics: they distort data and look.
- **Clarity:** Label clearly, specify units, use a legend when necessary.

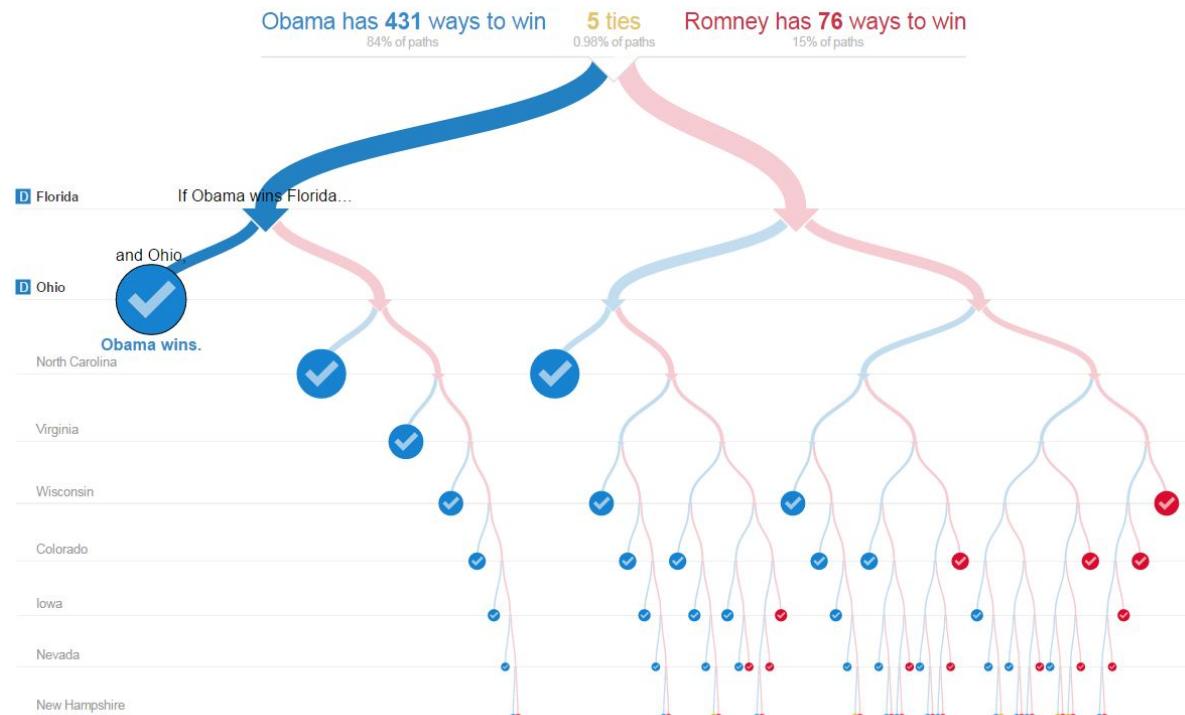
## Basic Design Concepts:

**Remove**  
to improve  
(the **data-ink** ratio)

# Look vs. Function:

## Questions

- What story does the data visualization tell?
- What details do you understand from the visualization?
- Is there text to help you understand?



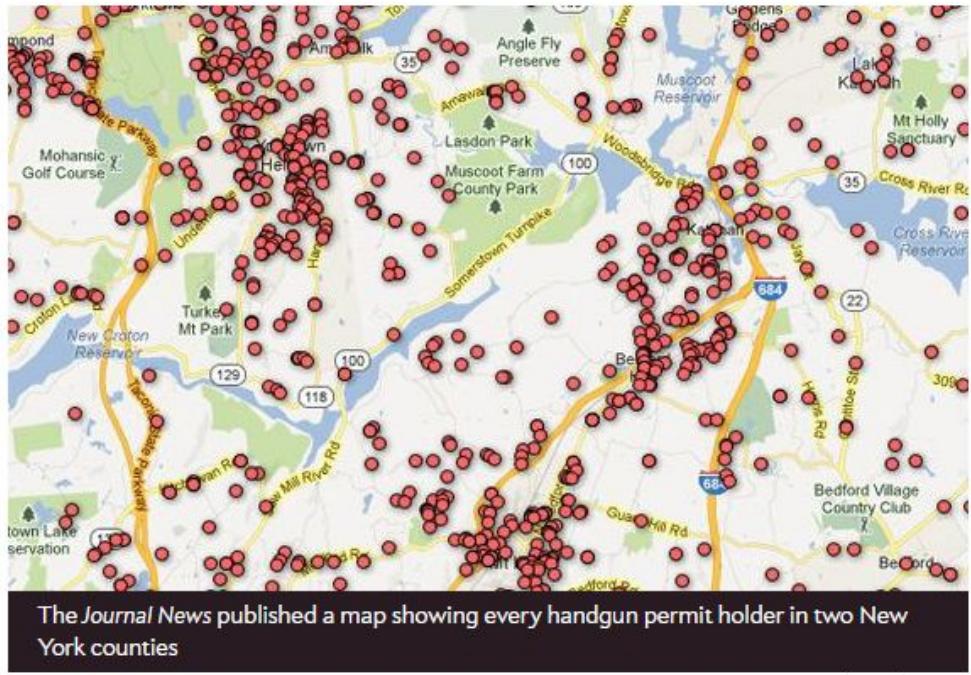
Source: [New York Times](#)

*"The greatest value of a picture is when it forces us to notice what we never expected to see."*

-John Tukey

# Ethics of Data Visualization: Privacy

- The Journal News published a map showing the name and address of people with gun license.
- This sparked a debate on privacy vs public safety in the media.
- What is the news value of publishing personal details here?

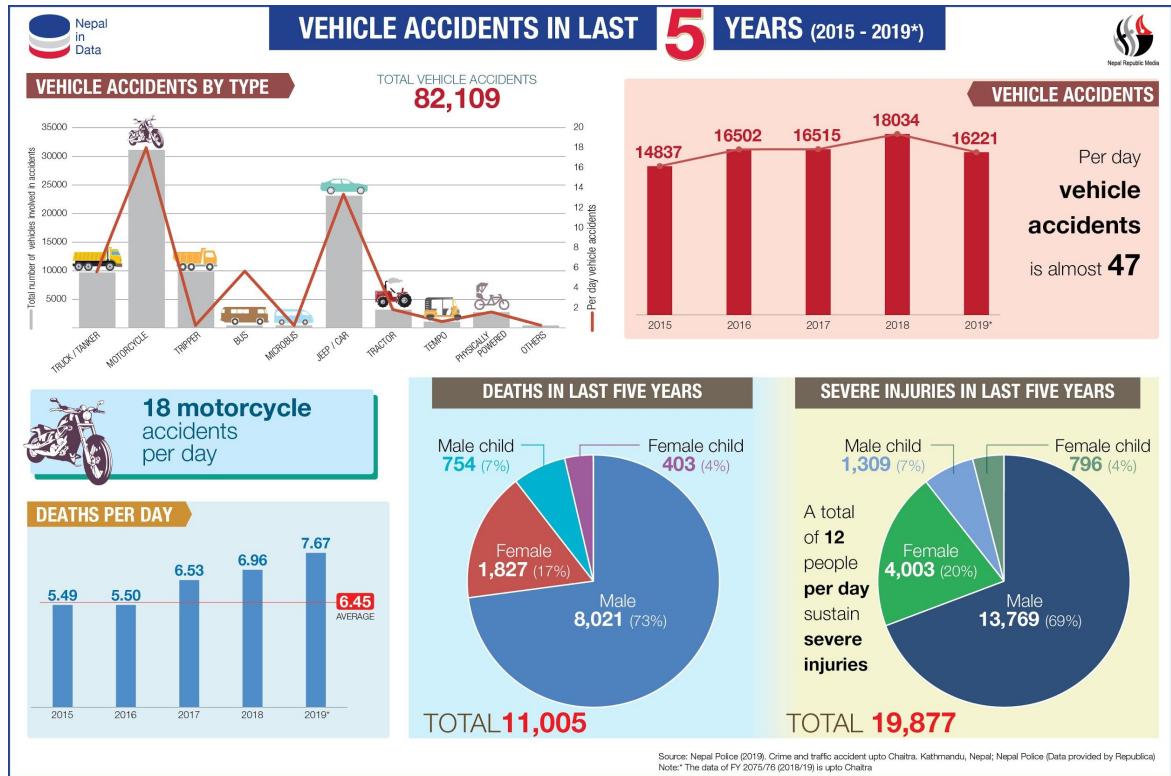


Courtesy of Google Maps.

# **Creating infographics using Infogram**

# Infographics = Information + Graphics

Graphical visual representations of information, data or knowledge intended to present information quickly and clearly.

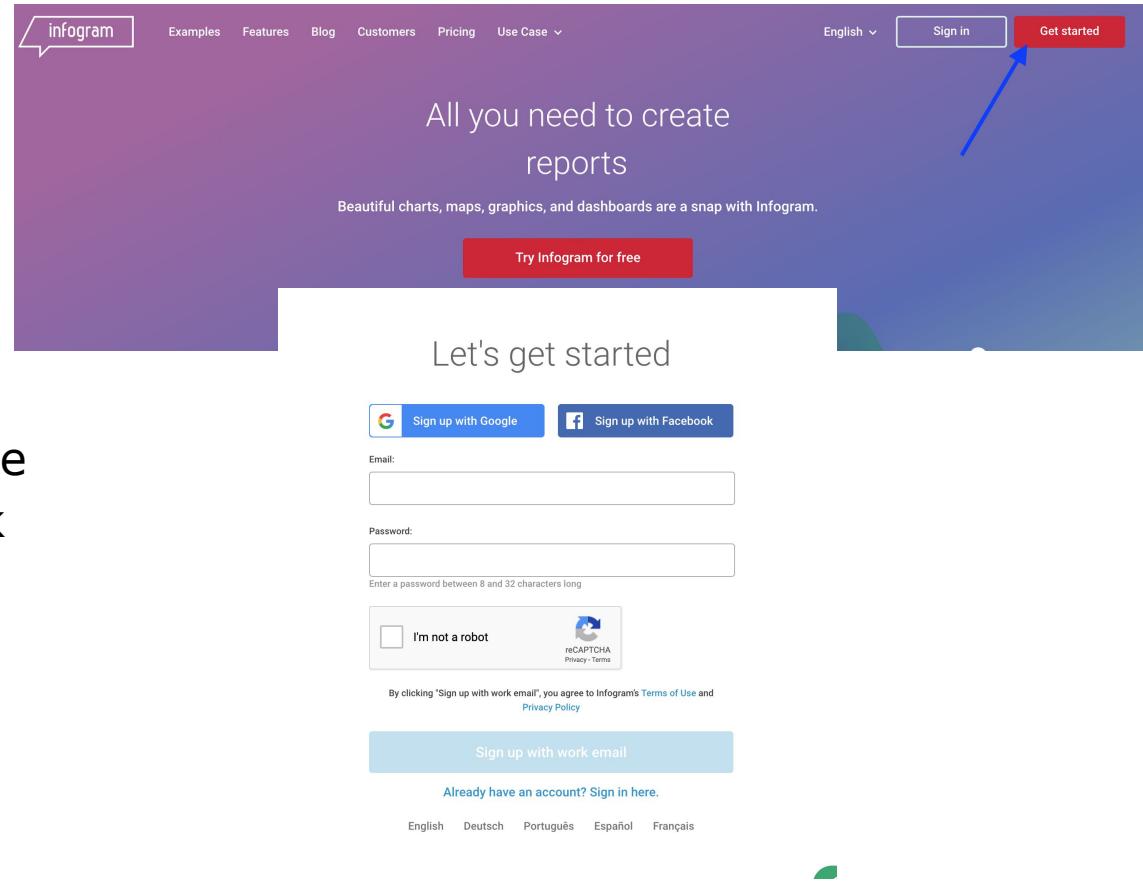


Source: [https://nepalindata.com/insight/vehicle-accidents-in-last-5-years-2015--2019\\*/](https://nepalindata.com/insight/vehicle-accidents-in-last-5-years-2015--2019*/)

# Getting Started with Infogram

1. Open the following website, and register (free) to login:  
<https://infogr.am/>

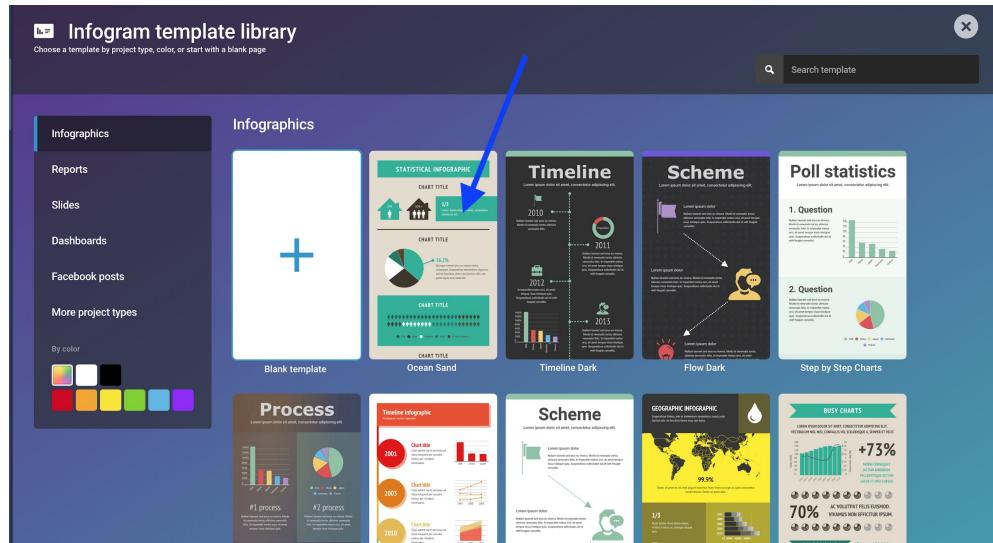
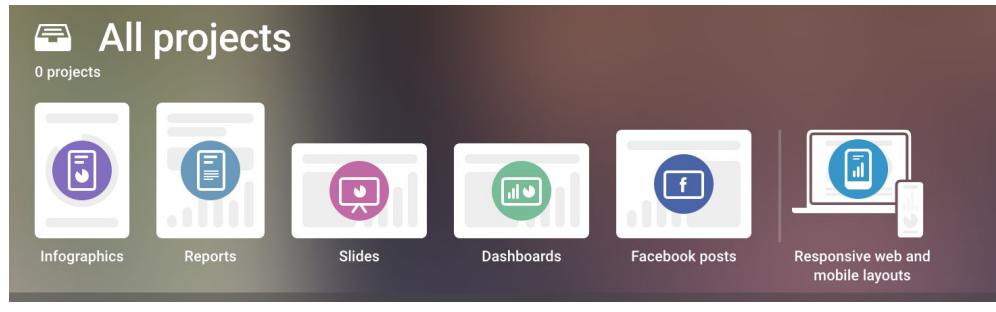
2. Sign up using your google account of your facebook account.
3. Set up your account by clicking next and sharing the kind of organization you work for



Source: [Infogram](#)

# Using Infogram

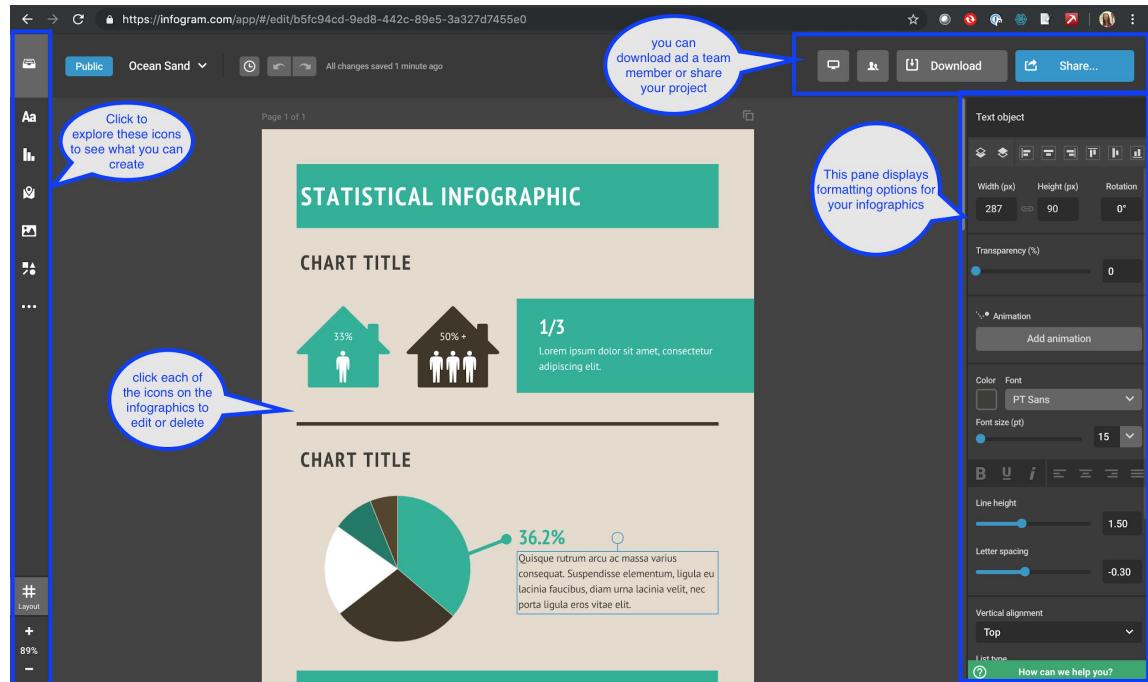
1. Once signed in, Select **either an infographics, report, or slides or social media post.**
2. Click infographics, let us choose a template 'statistical infographics' from the menu and click on **use.**
3. Create a title for the chart and identify whether the chart will be public (free) or private (which will require upgrade).



Source: [Infogram](#)

# Using Infogram

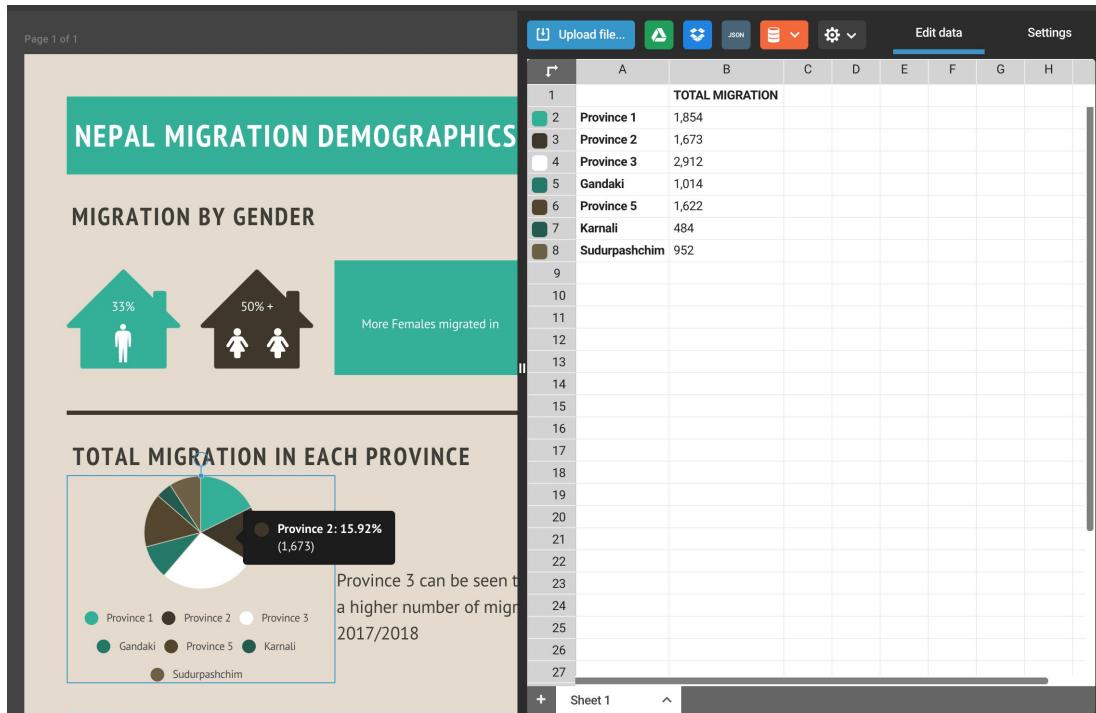
- NOTE: By default charts have some data, to change it, click on the chart, (it turn darker) and enable the option to edit.
- Change the chart titles by double clicking on the text box and naming it to reflect what infographics you want to show.



Source: [Infogram](#)

# Using Infogram

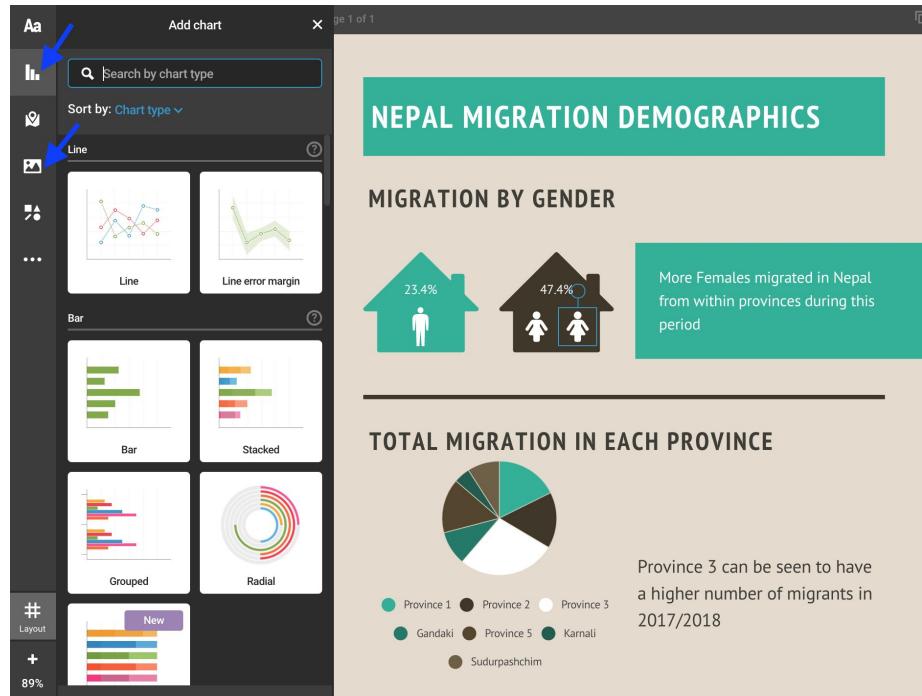
1. Update title by double clicking text box. **Name it Nepal Migration Demographics - We will be using data from your Exercise on Migration**
2. Update the data on the infographics with your own by clicking on the chart on the page, the **Edit data** window will open and you can now copy + paste from your spreadsheet.



Source: [Infogram](#)

# Using Infogram

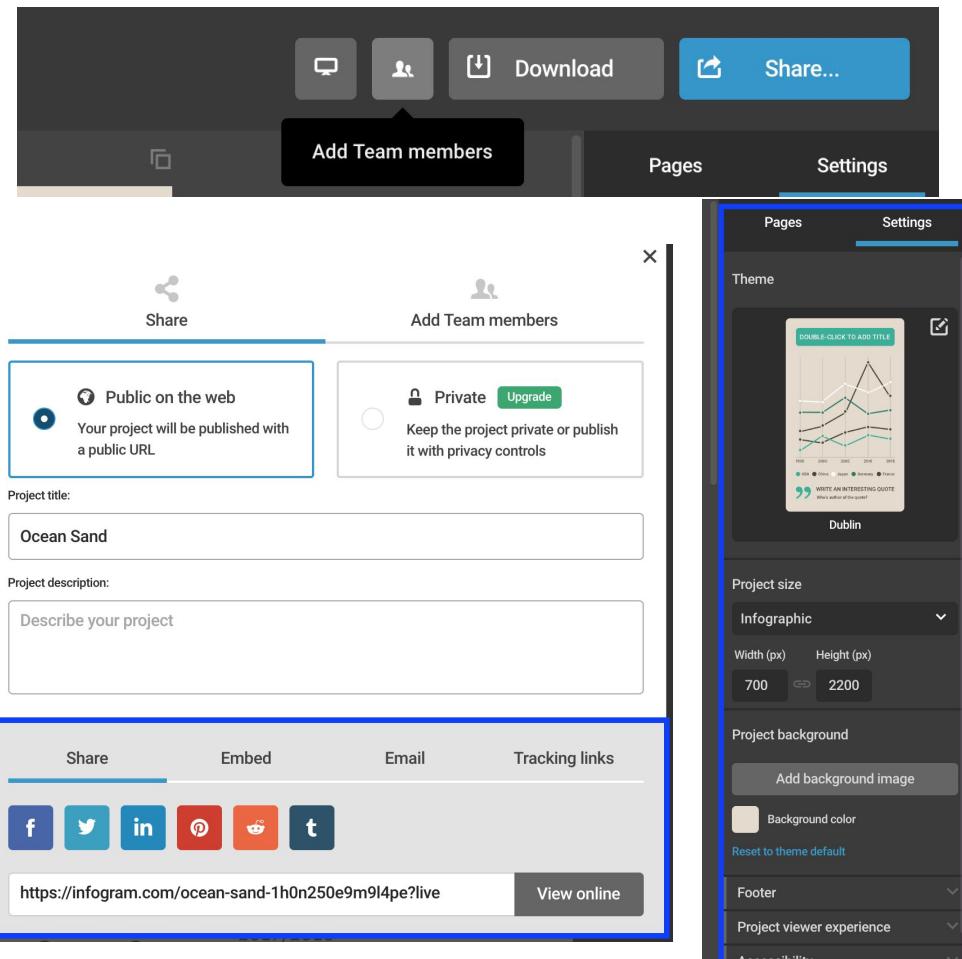
- Because Infogram works as storytelling application as well, it is possible to add charts, maps, text, videos, links, gifs etc.in the same visual package you can do this by navigating the left side menu of the page.
- Be sure to add source text under the chart. The source should include the title, data and link to the data source.



Source: [Infogram](#)

# Using Infogram

1. To update colors, font size, and formatting relating to update colors (theme), width, font, etc. Click on the object on the infographics to get settings related to that object or click the settings icon on the right hand side of the page
2. Click **Share** to publish to share on the web. On the **Share** window, any of the option to share, embed, or email that best suits your use.



Source: [Infogram](#)

# Lab : Creating Infographics

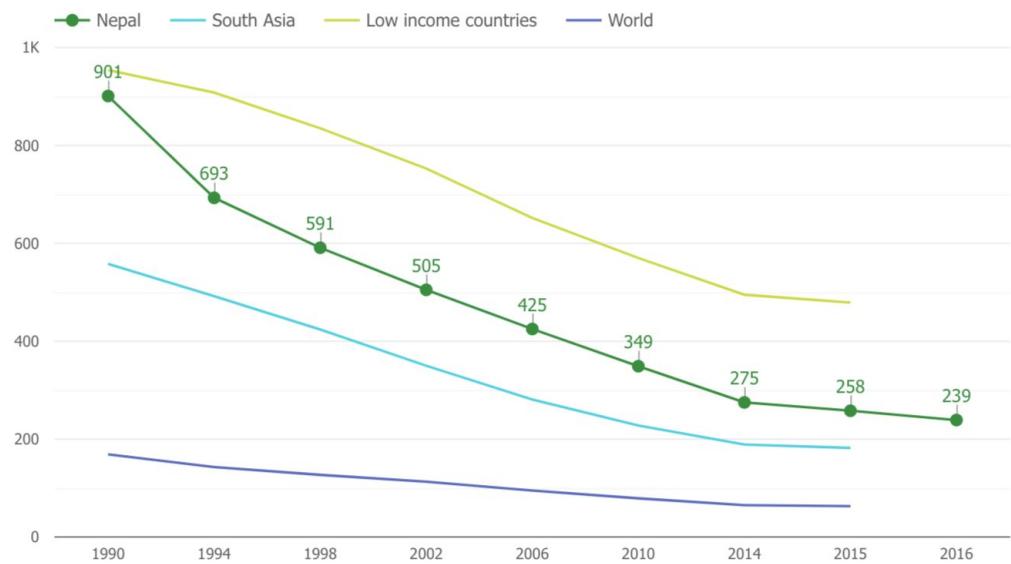
- Using your data on **Prevalence and Treatment of Diarrhea in Nepal**
- Using the Infogram, and the skills you have explored so far, create an infographics/story that raises awareness on findings.
- Feel free to research and add other data.
- You are allowed to be creative!

Module 6

# **Finding Insights in Data**

# SDGs Indicator 3.1.1: Maternal mortality ratio:

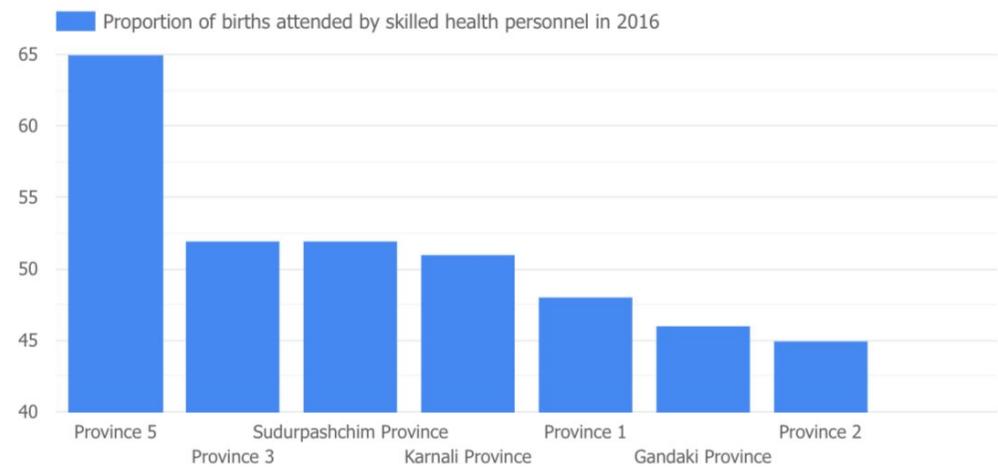
1. The ratio is the number of women who die from pregnancy-related causes while pregnant or within 42 days of pregnancy termination per 100,000 live births
2. The goal is to reduce maternal mortality to fewer than 70 per 100,000 live births by 2030
3. Summarize chart in one sentence



Source: [Visualizing Nepal's Health Progress](#)

# SDGs Indicator 3.1.2: Proportion of births attended by skilled health personnel

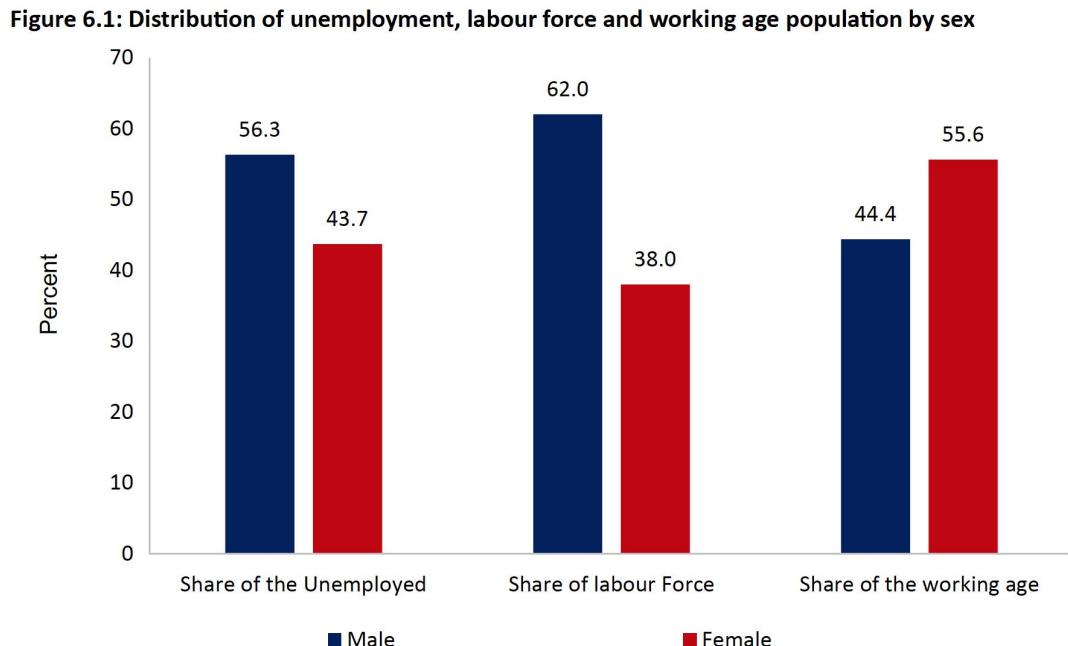
1. Births attended by skilled health staff are the percentage of deliveries attended by personnel trained to give the necessary supervision, care, and advice
2. There is no specific goal by the UN regarding this indicator
3. Give a one sentence analysis of the chart.



Source: [Visualizing Nepal's Health Progress](#)

# Profile of the unemployed people in Nepal by sex

Give a one sentence analysis of the chart.

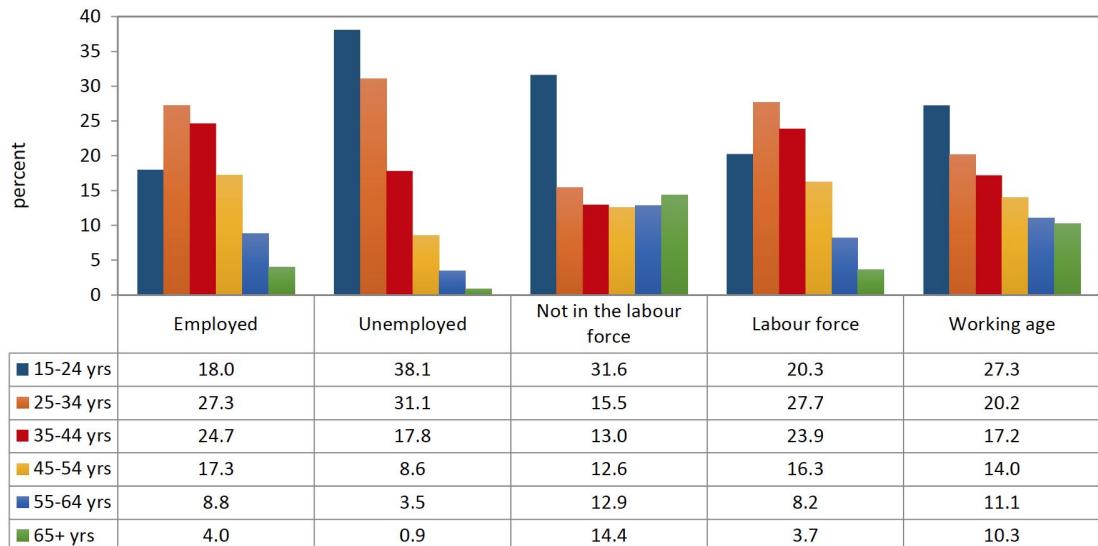


Data Source: [Report on the Nepal Labour Force Survey 2017/18](#)

# The age profile of persons in each component of Nepal's working-age population

Give a one sentence analysis of the chart.

Figure 3.2: The age profile of persons in each component of the working-age population

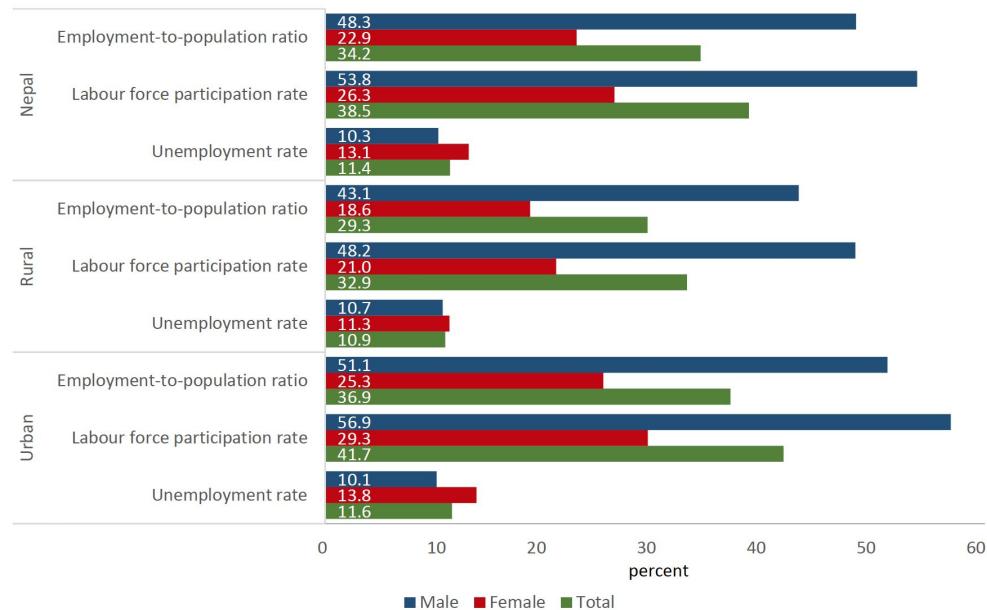


Data Source: Report on the Nepal Labour Force Survey 2017/18

# Key labour market indicators by sex and locality in Nepal

Give a one sentence analysis of the chart.

Figure 3.4: Key labour market indicators by sex and locality



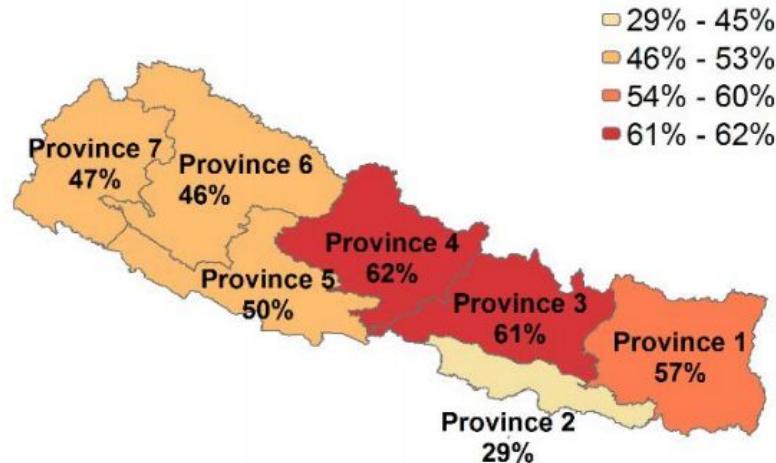
Source: Report on the Nepal Labour Force Survey  
2017/18

# Secondary Education in Nepal

**Figure 3.2 Secondary education by province**

*Percentage of women age 15-49 with secondary education complete or higher*

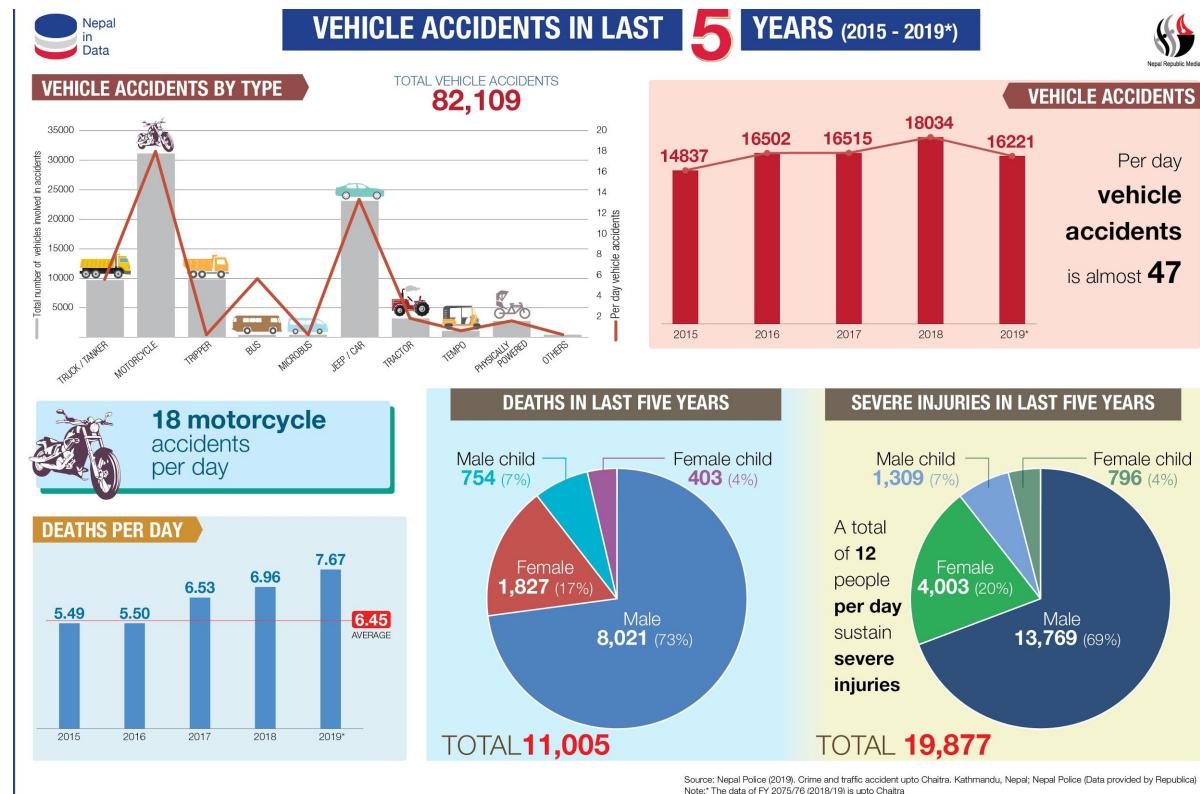
Give a one sentence analysis of the chart.



Source: National Demographic and Health Survey  
2016

# Vehicle Accidents in Nepal for 5 years

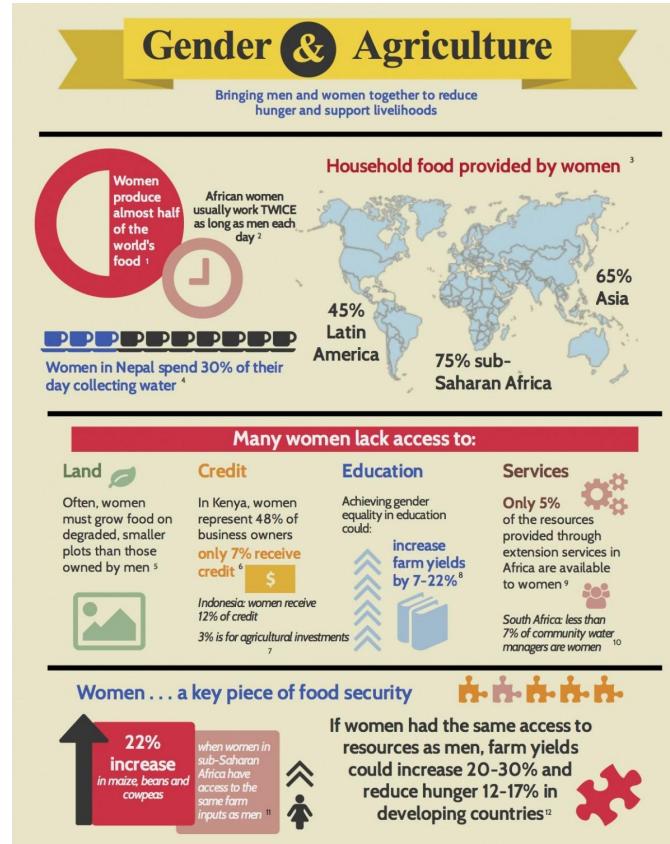
Give a one sentence analysis of the chart.



Source; [Nepal in Data](#)

# Lab 9:Summarizing Gender and Agriculture.

Give a fresh insight Headline to this infographic



RESEARCH  
PROGRAM ON  
Water, Land  
and  
Ecosystems



LED BY  
IWMI  
International Water Management Institute

More information available at <http://www.iwmi.org/world-womens-day/key-facts-on-gender/>

Data Source: Research Program on Water, land and Ecosystem

# **Questions**