

# R Data Visualization

Wednesday, 9 April 2025  
Presented by Satoshi Koiso

# Table of contents

---

## [Overview]

- R vs Excel vs STATA
- Figure Types and use cases

## [R]

- Basic operation
- 1. Histogram
- 2. Bar plot
  - 2.1. Stacked bar plot
- 3. Line graph
- 4. Scatter plot
- 5. Tornado plot
- 6. Heatmap

# Table of contents

---

## [Overview]

- **R vs Excel vs STATA**
- Figure Types and use cases

## [R]

- Basic operation
- 1. Histogram
- 2. Bar plot
  - 2.1. Stacked bar plot
- 3. Line graph
- 4. Scatter plot
- 5. Tornado plot
- 6. Heatmap

# R vs. Excel vs. STATA

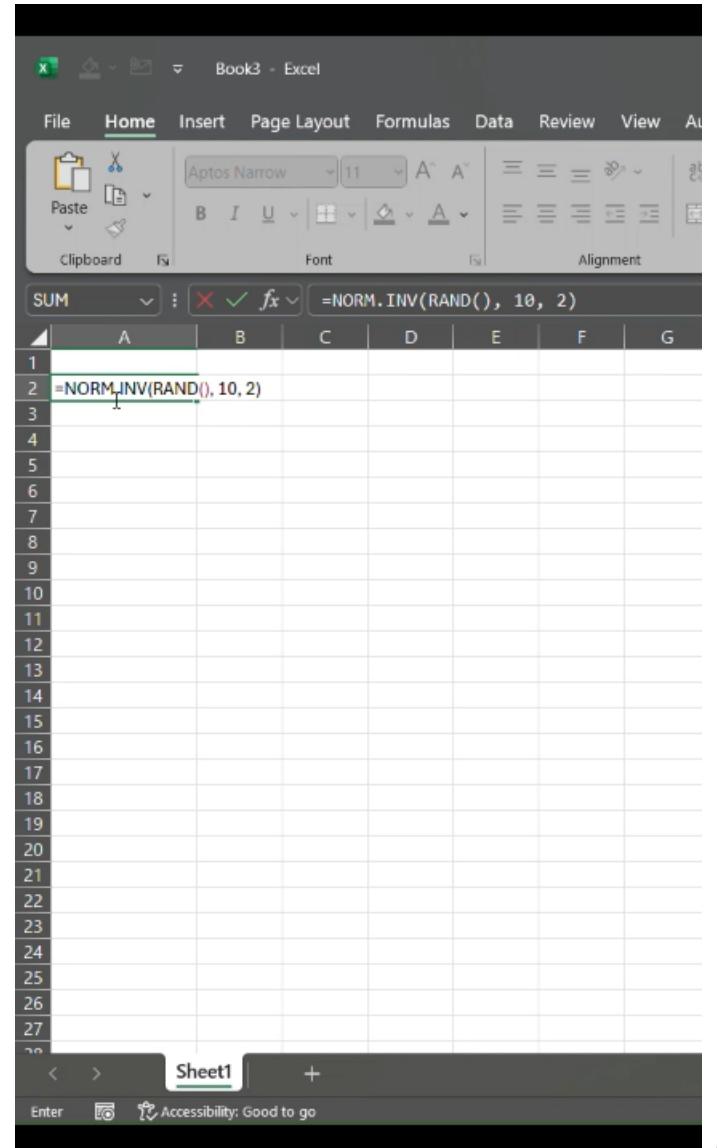
Software	Price	Pros	Cons
R	Free	<ul style="list-style-type: none"><li>• Versatile</li><li>• Better graphical tools</li><li>• Comprehensive documentation</li><li>• Strong community of users</li><li>• Large selection of libraries</li></ul>	<ul style="list-style-type: none"><li>• Steeper learning curve</li><li>• Longer scripts compared to STATA</li><li>• Most work will require installation of libraries</li><li>• Updates to R and libraries often affect running older scripts</li></ul>
Excel	Free for online or \$99.99/year	<ul style="list-style-type: none"><li>• Easiest to use</li></ul>	<ul style="list-style-type: none"><li>• Cannot handle large data files</li><li>• Prone to human error</li><li>• Hard to conduct <b>probabilistic sensitivity analysis</b></li></ul>
Google spreadsheet	Free		
STATA	\$840 per year (for a single-user)	<ul style="list-style-type: none"><li>• Easier to use than R (regression)</li><li>• More out-of-the-box statistical tools</li></ul>	<ul style="list-style-type: none"><li>• Limited data visualization capabilities compared to R</li><li>• Can only import one data set at a time</li></ul>

# Probabilistic Sensitivity Analysis

- In PSA, we sample numbers from a distribution to account for uncertainty.
- Example: Sampling 100 numbers from a normal distribution with a mean of 10 and SD of 2.
- R
  - Can reproduce the same result by setting a seed.
  - Just three lines of code.
- Excel/Google spreadsheet
  - Results are unstable.
  - Much more steps.

```
1 set.seed(123) # Optional: for reproducibility
2 sample_values <- rnorm(100, mean = 10, sd = 2)
3
4 # View the first few values
5 sample_values
6
7
```

```
7:1 [Top Level] > set.seed(123) # Optional: for reproducibility
> sample_values <- rnorm(100, mean = 10, sd = 2)
> # View the first few values
> sample_values
[1]  8.879049  9.539645 13.117417 10.141017 10.258575 13.430130
[7] 10.921832  7.469878  8.626294  9.108676 12.448164 10.719628
[13] 10.801543 10.221365  8.888318 13.573826 10.995701  6.066766
[19] 11.402712  9.054417  7.864353  9.564050  7.947991  8.542218
[25]  8.749921  6.626613 11.675574 10.306746  7.723726 12.507630
[31] 10.852928  9.409857 11.790251 11.756267 11.643162 11.377281
[37] 11.107835  9.876177  9.388075  9.239058  8.610586  9.584165
[43]  7.469207 14.337912 12.415924  7.753783  9.194230  9.066689
[49] 11.559930  9.833262 10.506637  9.942906  9.914259 12.737205
[55]  9.548458 13.032941  6.902494 11.169227 10.247708 10.431883
[61] 10.759279  8.995353  9.333585  7.962849  7.856418 10.607057
[67] 10.896420 10.106008 11.844535 14.100169  9.017938  5.381662
[73] 12.011477  8.581598  8.623983 12.051143  9.430454  7.558565
[79] 10.362607  9.722217 10.011528 10.770561  9.258680 11.288753
[85]  9.559027 10.663564 12.193678 10.870363  9.348137 12.297615
[91] 11.987008 11.096794 10.477463  8.744188 12.721305  8.799481
[97] 14.374666 13.065221  9.528599  7.947158
```



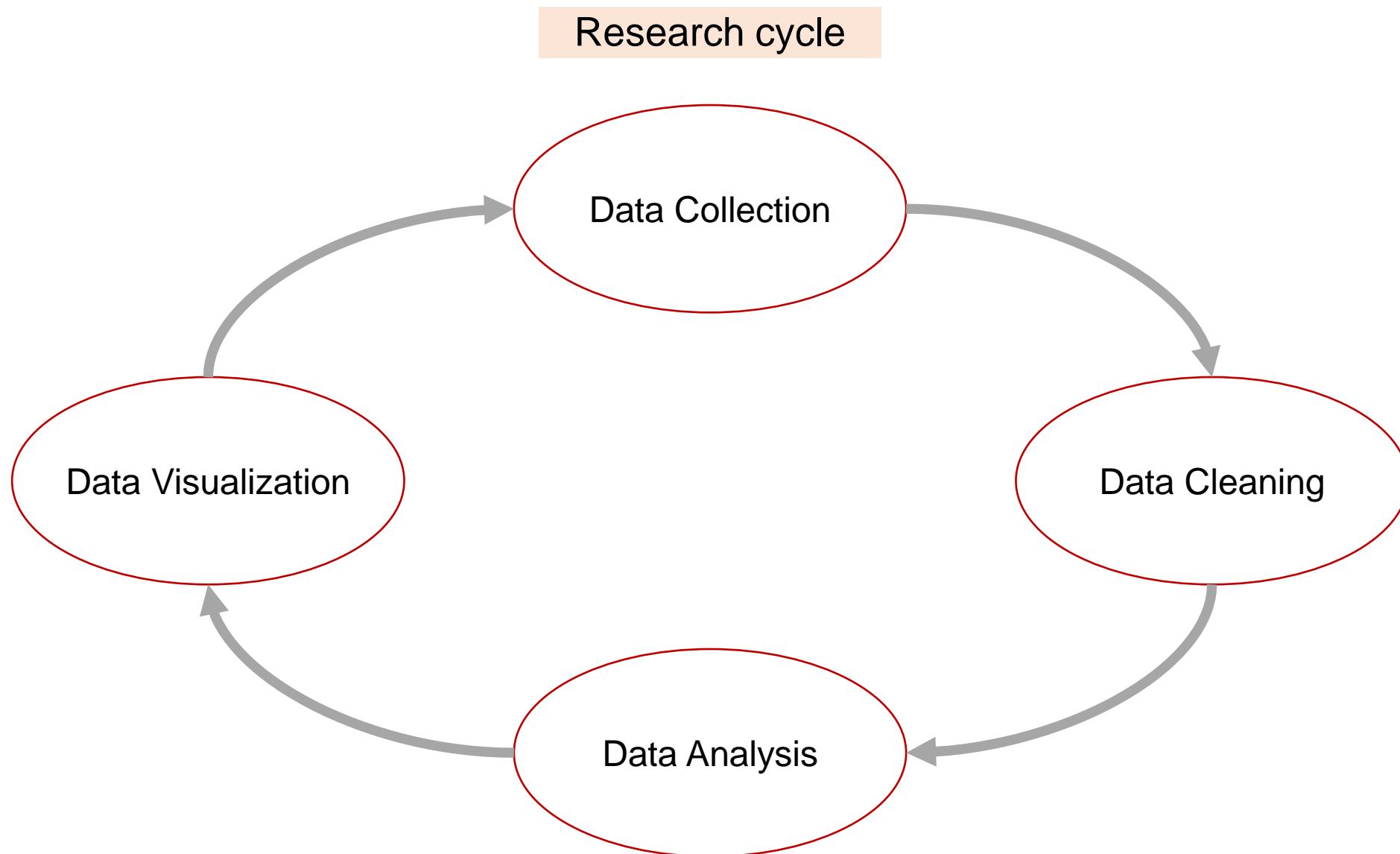
# What should R be used for?

---

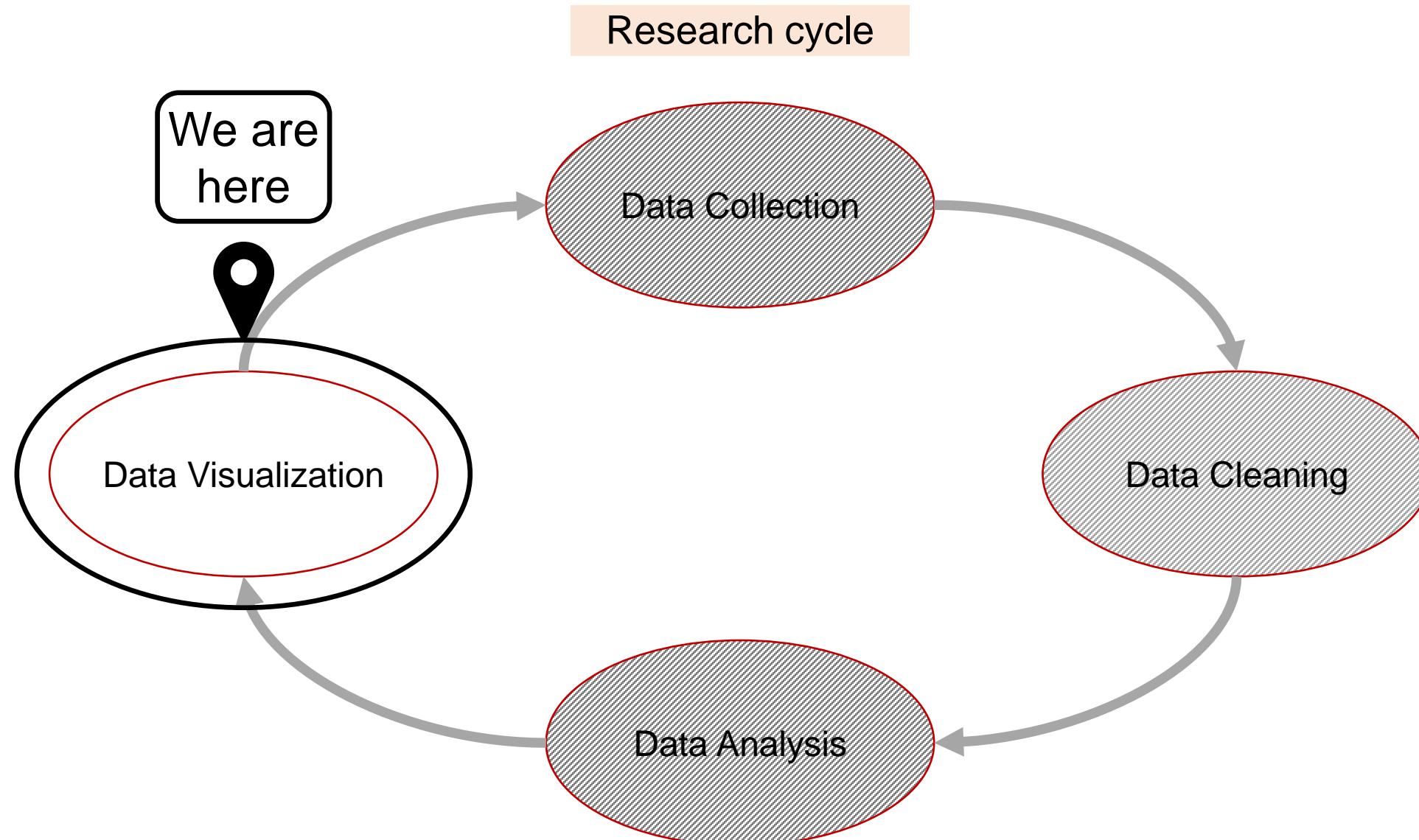
- Data analysis and figure creation for one specific project
- Sensitivity analyses involving many runs
- Integrating data with online
- Geospatial analyses

# Scope of this session

---



# Scope of this session



# Table of contents

---

## [Overview]

- R vs Excel vs STATA
- **Figure Types and use cases**

## [R]

- Basic operation
- 1. Histogram
- 2. Bar plot
  - 2.1. Stacked bar plot
- 3. Line graph
- 4. Scatter plot
- 5. Tornado plot
- 6. Heatmap

# Histogram

Use case:

- Show distributions of data.
- The height of a bar represents frequency of occurrence of values within each bin.

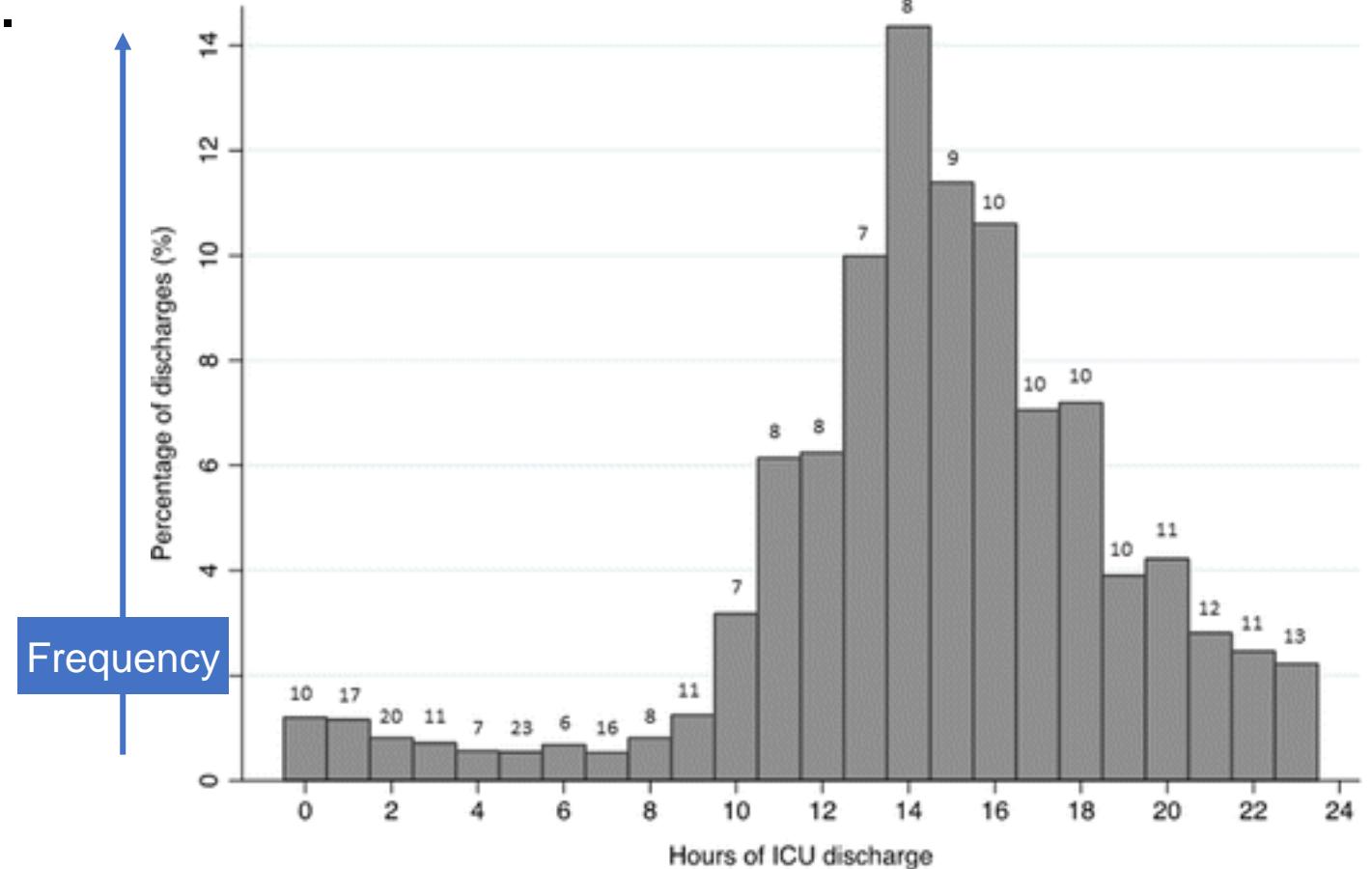


Figure1. Histogram showing the distribution of hours of ICU discharge for the study population as well as the unadjusted mortality rates distributed according to hour of discharge  
From Azevedo, L.C., de Souza, I.A., Zygund, D.A. et al. Association Between Nighttime Discharge from the Intensive Care Unit and Hospital Mortality: A Multi-Center Retrospective Cohort Study. BMC Health Serv Res 15, 378 (2015). <https://doi.org/10.1186/s12913-015-1044-4>

# Bar graph

Use case:

- Compare data across different groups.

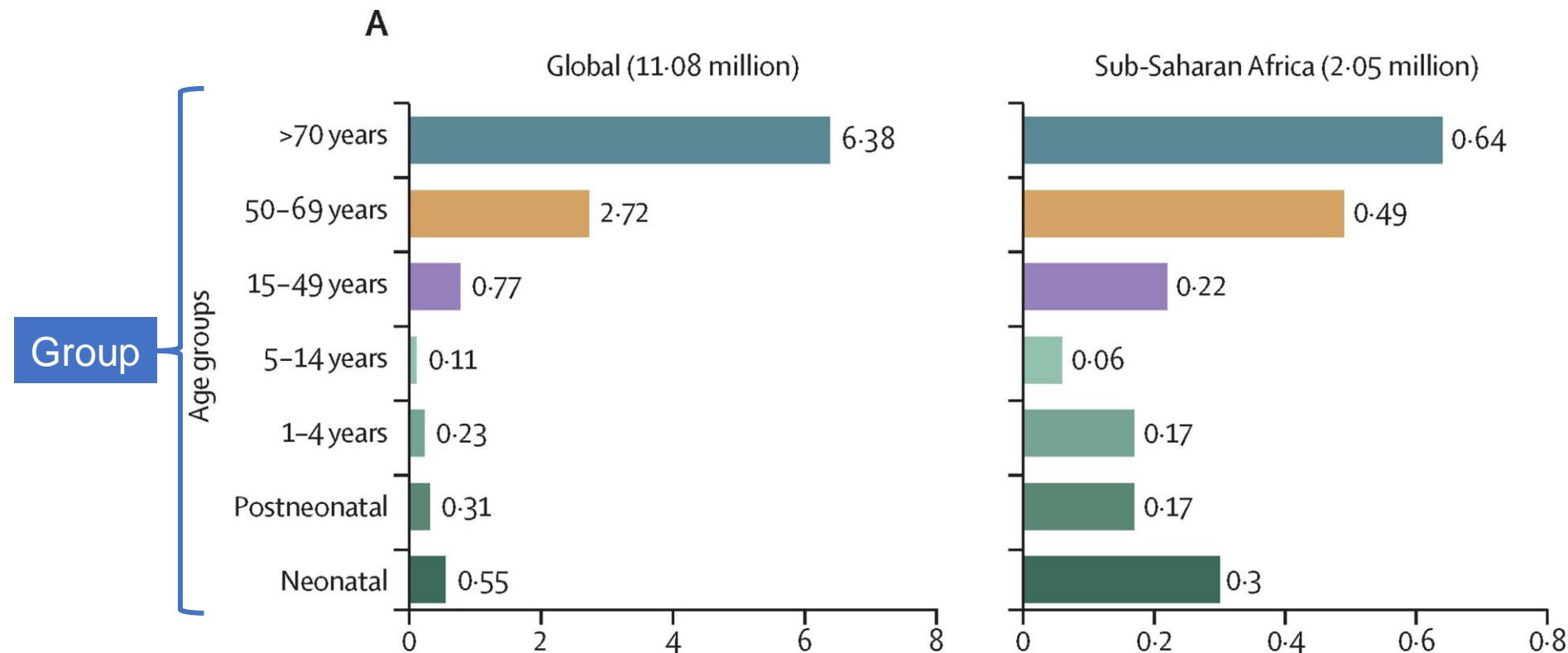


Figure 8. Cumulative deaths averted (in millions) in two different scenarios, by age, global versus sub-Saharan Africa, 2025–2050 (A) Gram-negative drug scenario.

From GBD 2021 Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance 1990-2021: a systematic analysis with forecasts to 2050. Lancet. 2024;404(10459):1199-1226. doi:10.1016/S0140-6736(24)01867-1

# Stacked bar graph

## Use case:

- Compare data across subgroups in different groups.
- e.g. Across a group of strategies, showing which subgroup of costs occupies how much of the total costs.

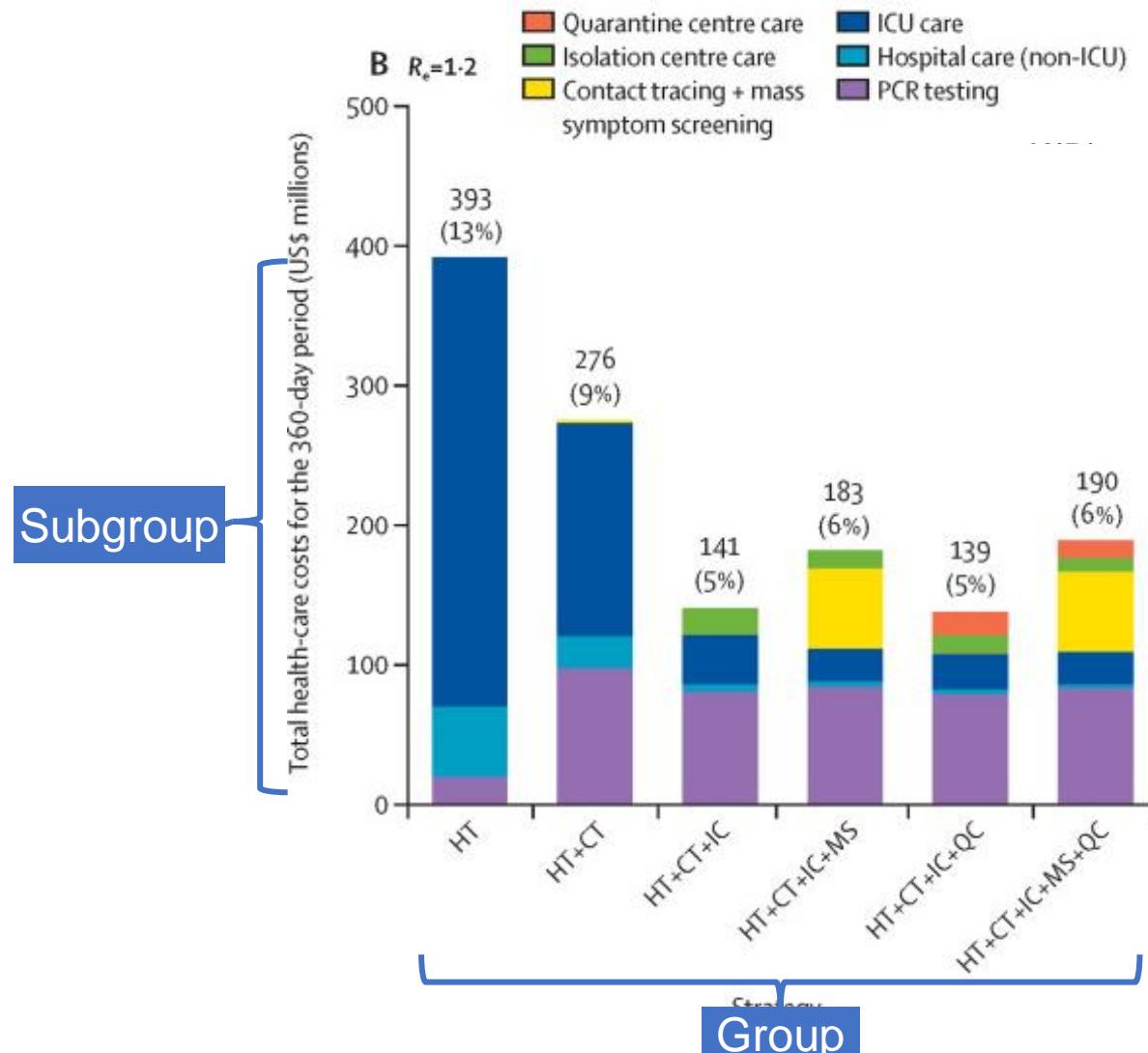


Figure 2. Budget impact analysis of contributors to health-care costs of COVID-19 intervention strategies applied to the population of KwaZulu-Natal province, South Africa (11 million people)  
From Reddy KP, Shebl FM, Foote JHA, et al. Cost-effectiveness of public health strategies for COVID-19 epidemic control in South Africa: a microsimulation modelling study. Lancet Glob Health. 2021;9(2):e120-e129. doi:10.1016/S2214-109X(20)30452-6

# Scatter plot

Use case:

- Explore a relationship between two variables.
- Visualize trends in data and shape of data.

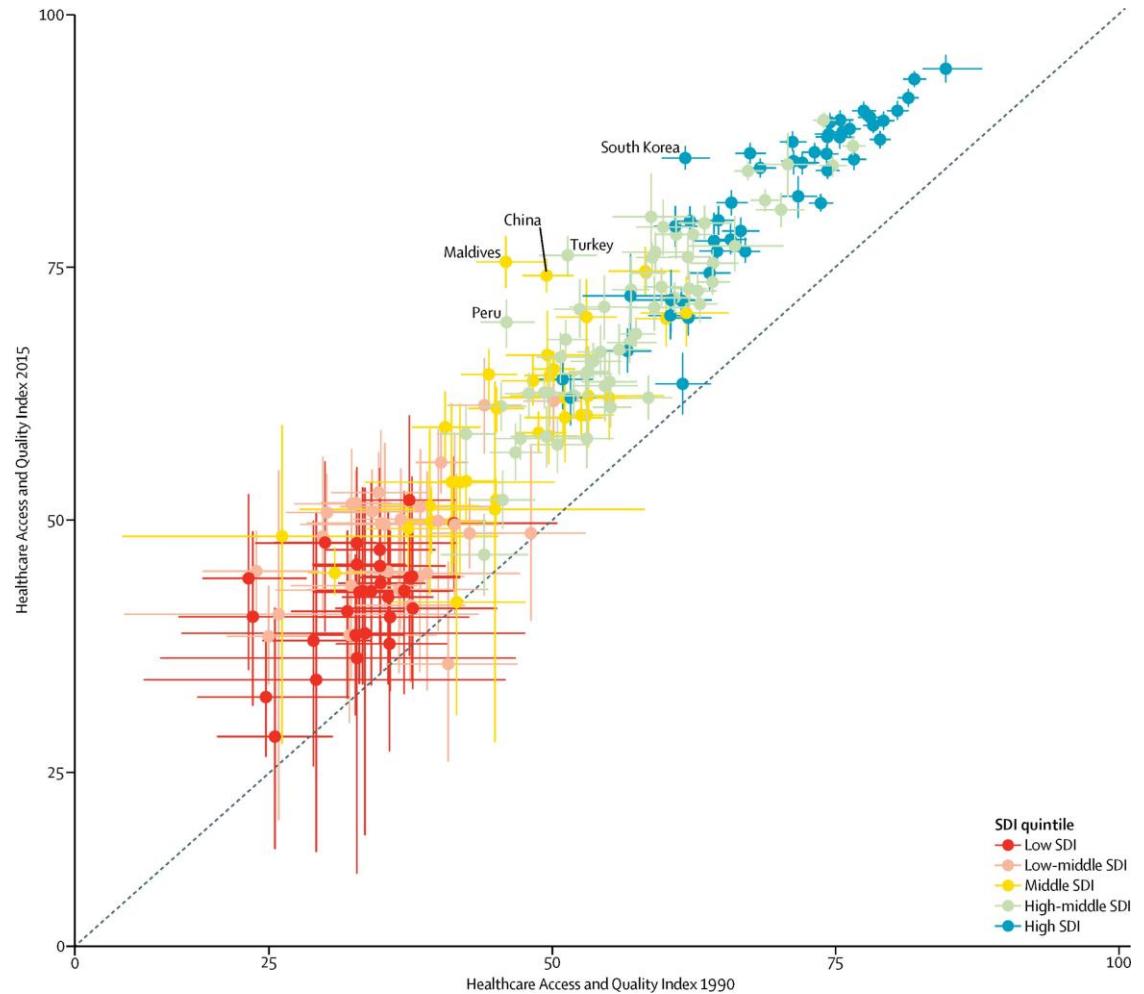


Figure 3 Comparison of 1990 and 2015 HAQ Index estimates, with uncertainty, by country or territory

From Healthcare Access and Quality Index based on mortality from causes amenable to personal health care in 195 countries and territories, 1990–2015: a novel analysis from the Global Burden of Disease Study 2015 Barber, Ryan M et al. The Lancet, Volume 390, Issue 10091, 231 - 266

# Cost-effectiveness scatter plot

Use case:

- Show how parameter uncertainty impacts cost-effectiveness.
- Each dot shows the result from each recalculation in a probabilistic sensitivity analysis.

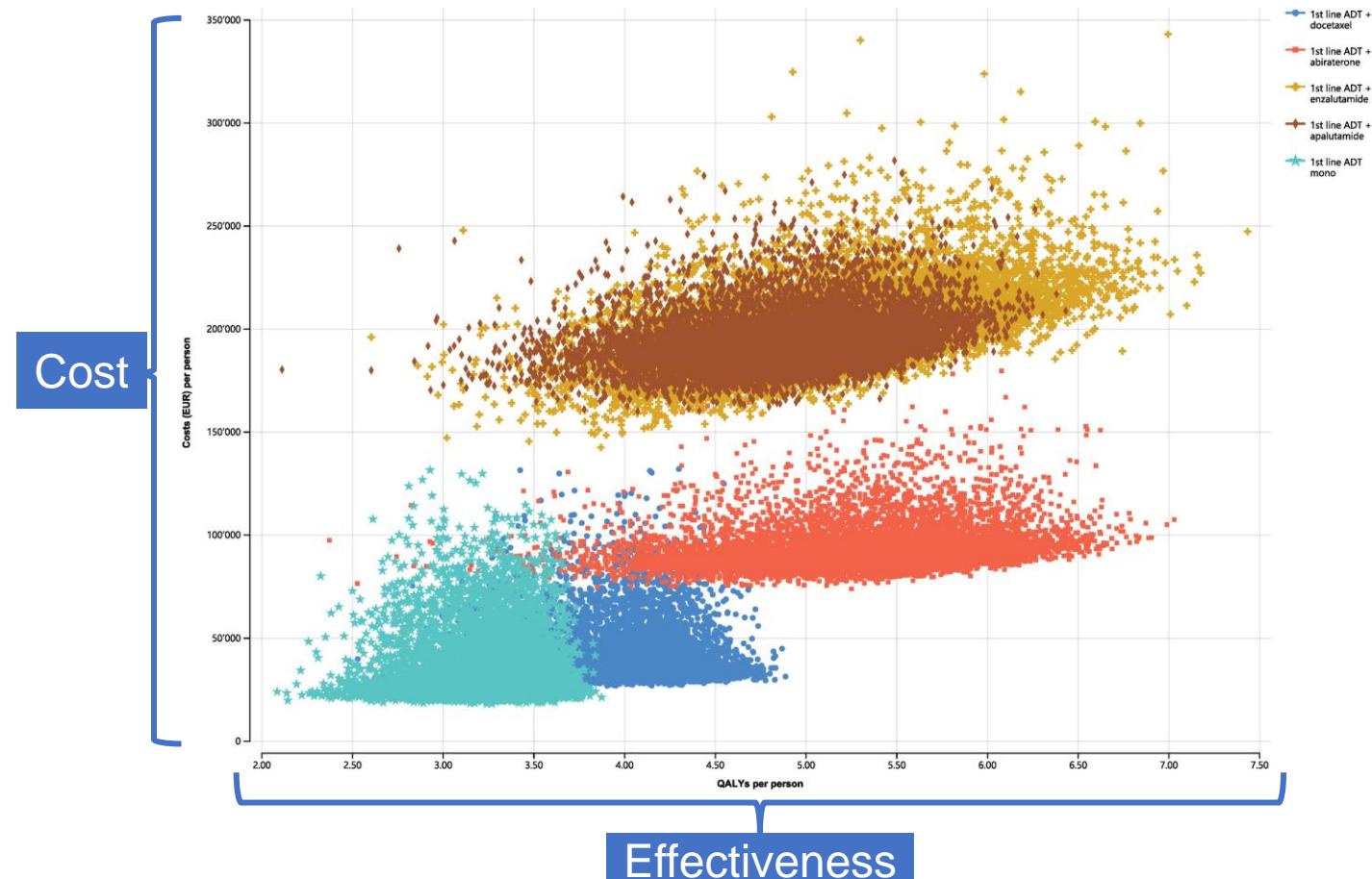


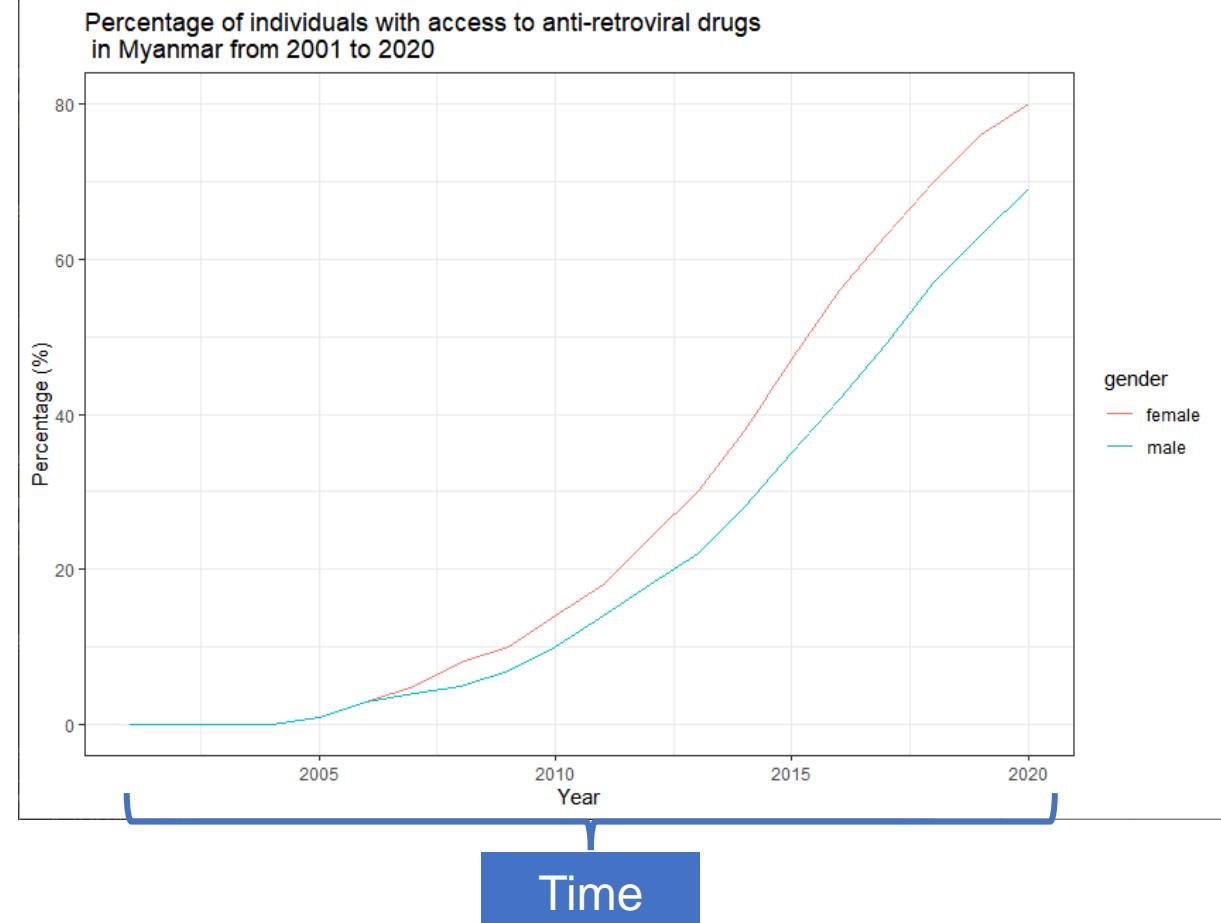
Fig 3. Cost-effectiveness scatter plot based on probabilistic sensitivity analysis.

From Barbier MC, Tomonaga Y, Menges D, Yebo HG, Haile SR, Puhan MA, et al. (2022) Survival modelling and cost-effectiveness analysis of treatments for newly diagnosed metastatic hormone-sensitive prostate cancer. PLoS ONE 17(11): e0277282. <https://doi.org/10.1371/journal.pone.0277282>

# Line graph

Use case:

- Visualize values of variables with a gradual change of a continuous variable (often time).



# Survival curve

Use case:

- Plot model output of proportion alive over time.
- Compare several intervention arms.

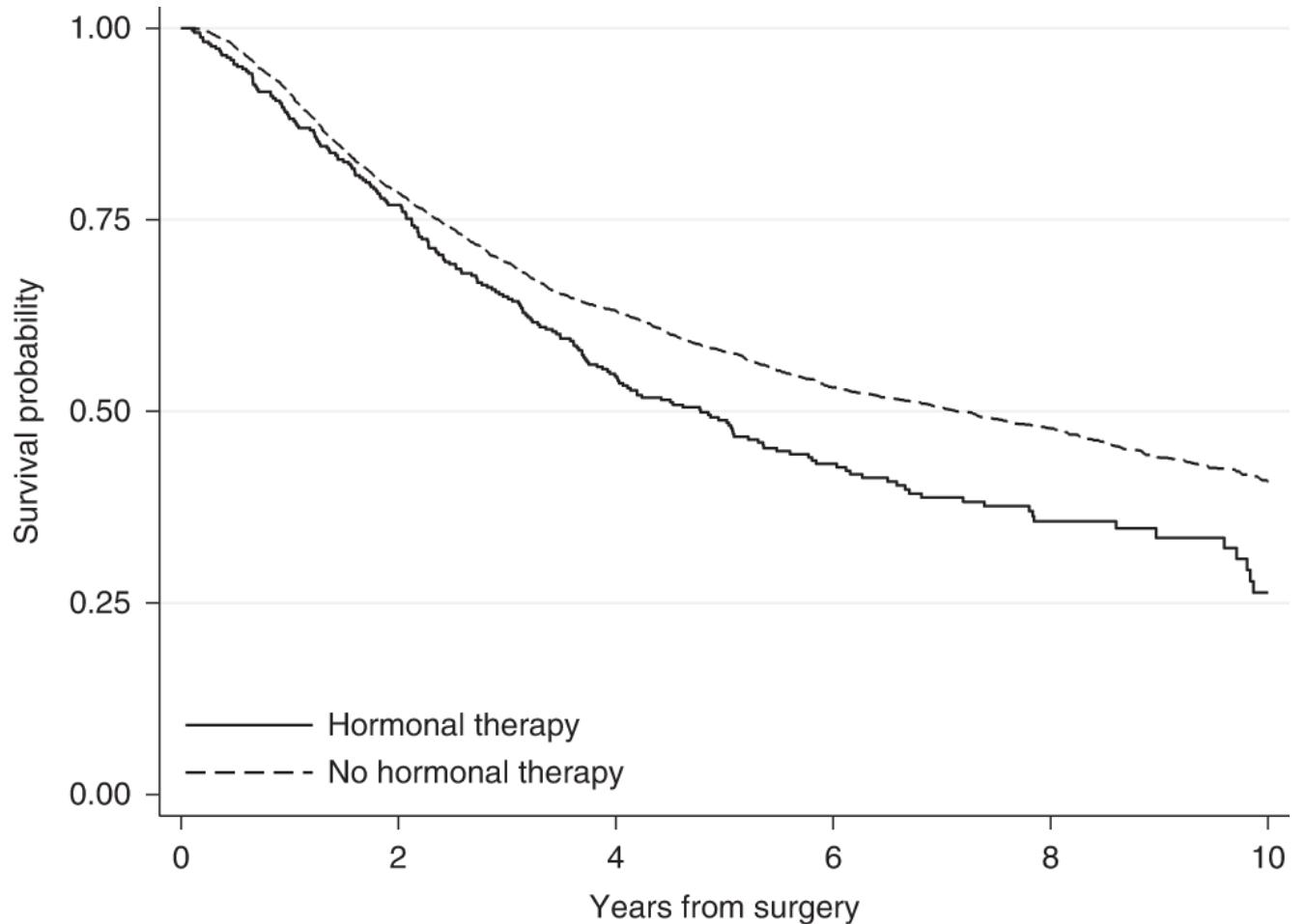


Fig. 1: Kaplan–Meier survival curves.

From Syriopoulou, E., Wästerlid, T., Lambert, P.C. et al. Standardised survival probabilities: a useful and informative tool for reporting regression models for survival data. Br J Cancer 127, 1808–1815 (2022). <https://doi.org/10.1038/s41416-022-01949-6>

# Cost-Effectiveness frontier

Use case:

- Compare competing strategies' cost-effectiveness.
- Line connects successive points representing cost-effective strategies at different values of the cost-effectiveness threshold.
- Points not on the frontier are not considered cost-effective at any threshold.

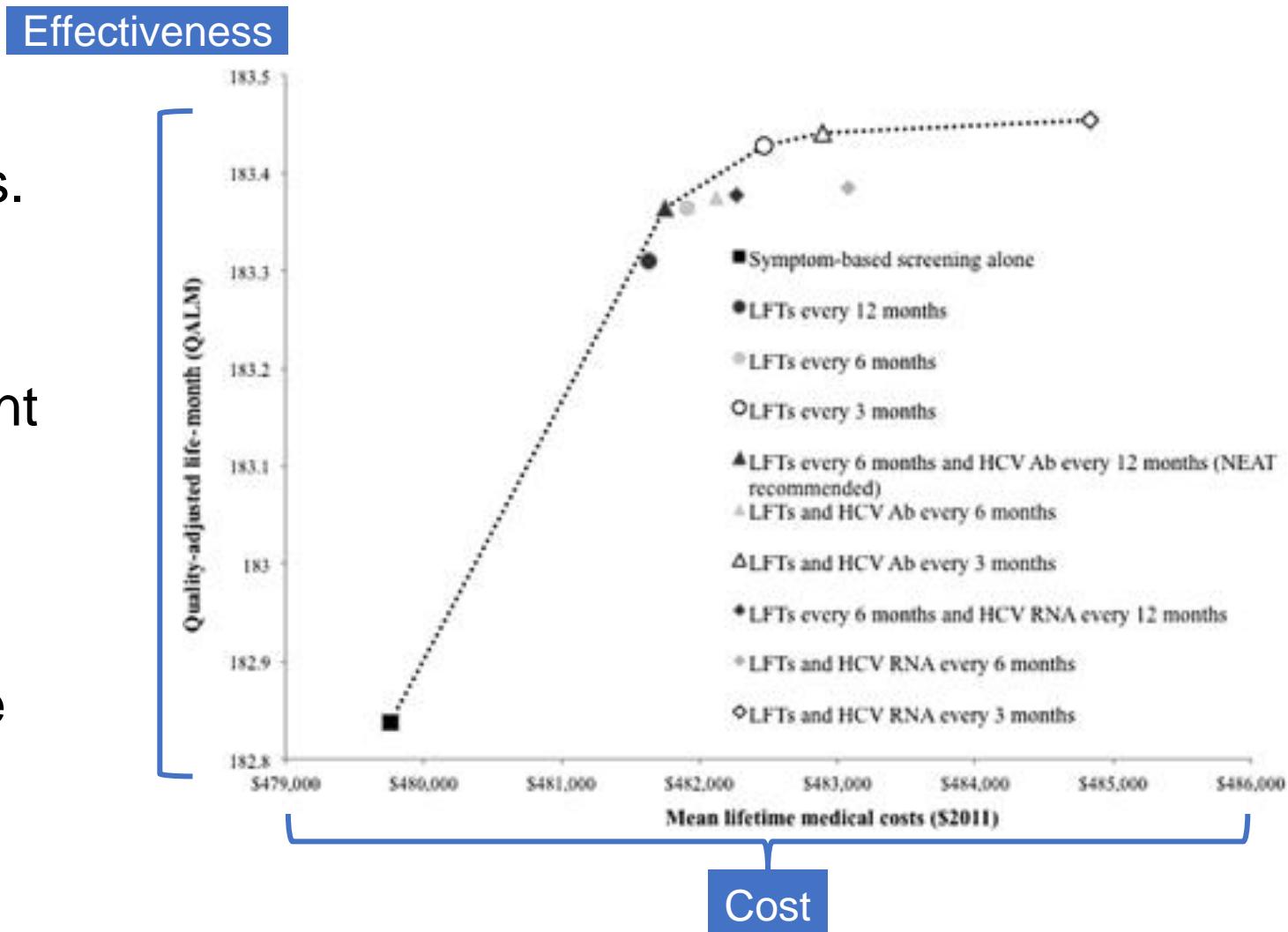


Figure 1. Efficiency frontier of a cost-effectiveness analysis of screening for acute hepatitis C virus (HCV) infection in human immunodeficiency virus (HIV)-infected men who have sex with men. From Benjamin P. Linas, Angela Y. Wong, Bruce R. Schackman, Arthur Y. Kim, Kenneth A. Freedberg, Cost-effective Screening for Acute Hepatitis C Virus Infection in HIV-Infected Men Who Have Sex With Men, Clinical Infectious Diseases, Volume 55, Issue 2, 15 July 2012, Pages 279–290, <https://doi.org/10.1093/cid/cis382>

# How to interpret Cost-Effectiveness frontier

Use case:

- The slope between two points represents the incremental cost-effectiveness ratio (ICER).
- In the right graph,
  - A: \$43,700/QALY
  - B: \$129,700/QALY
- Depending on the decision-maker's preference, the optimal strategies would be different.

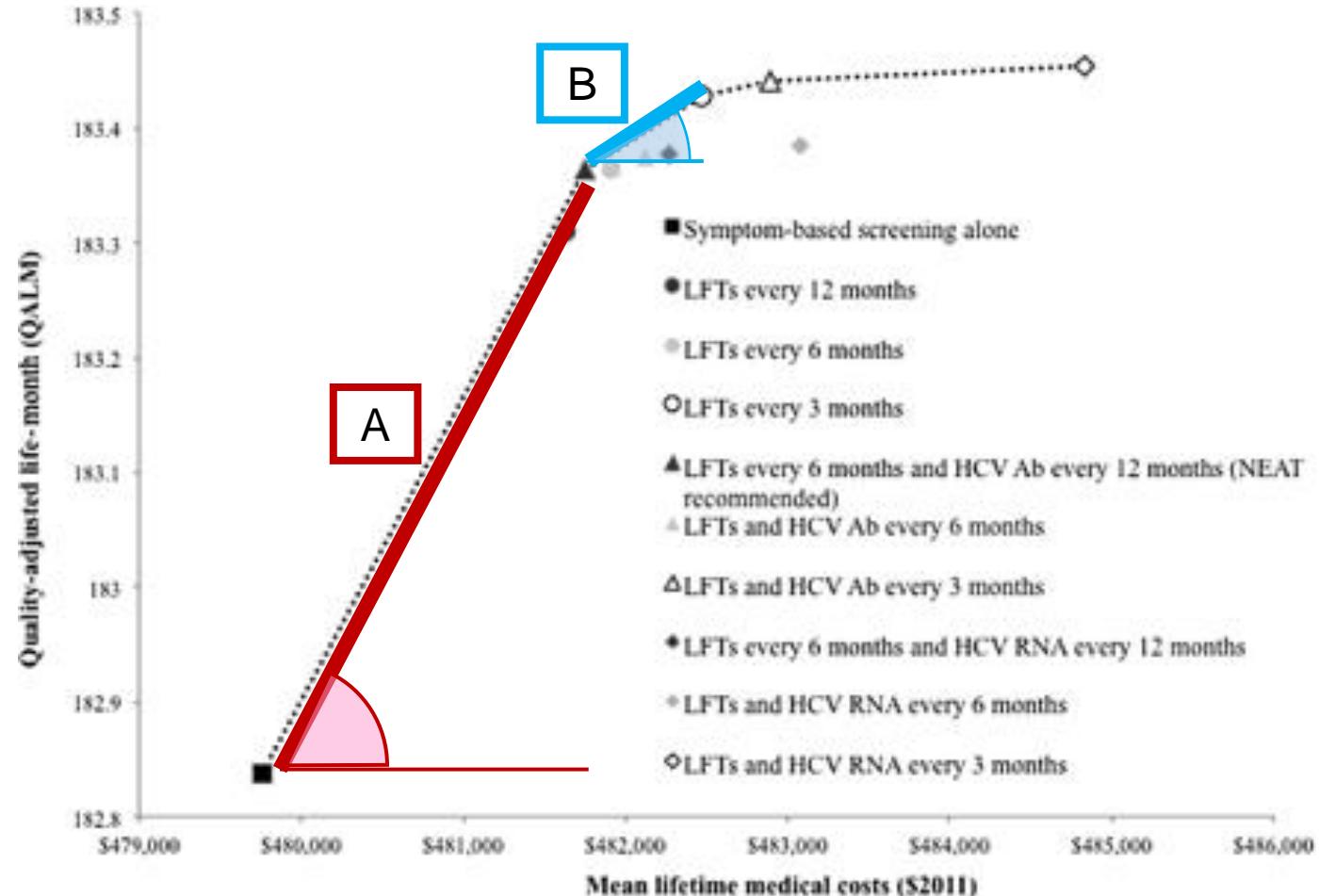


Figure 1. Efficiency frontier of a cost-effectiveness analysis of screening for acute hepatitis C virus (HCV) infection in human immunodeficiency virus (HIV)-infected men who have sex with men. From Benjamin P. Linas, Angela Y. Wong, Bruce R. Schackman, Arthur Y. Kim, Kenneth A. Freedberg, Cost-effective Screening for Acute Hepatitis C Virus Infection in HIV-Infected Men Who Have Sex With Men, Clinical Infectious Diseases, Volume 55, Issue 2, 15 July 2012, Pages 279–290, <https://doi.org/10.1093/cid/cis382>

# Cost-effectiveness acceptability curve

Use case:

- Shows the percentage of calculations that favor each strategy over a Willingness-to-pay range.
- As WTP is higher, more effective strategies are more likely to be chosen.

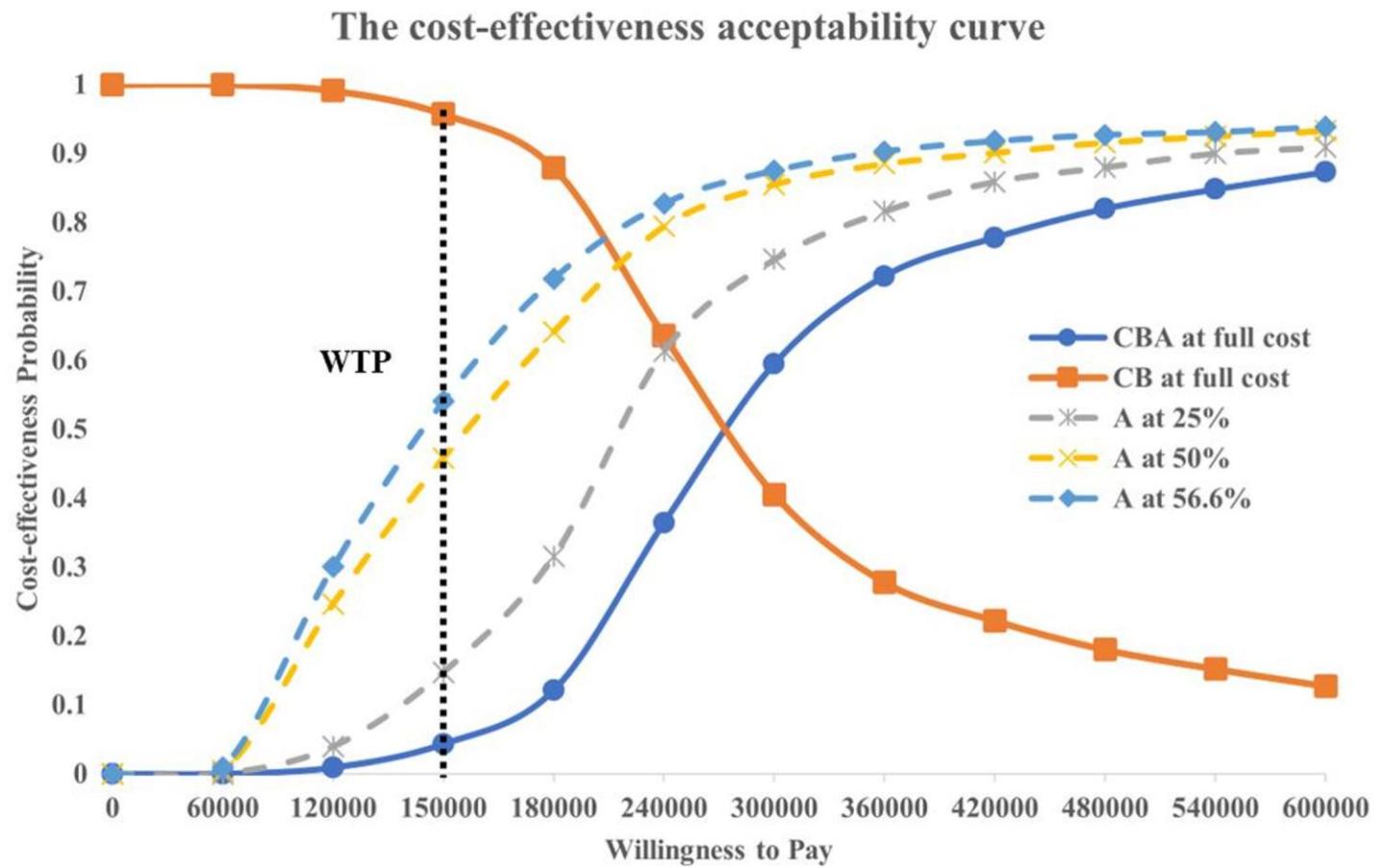


Figure 3. The cost-effectiveness acceptability curve of different atezolizumab cost.

From Lei, Jianying, MSc, et al. "First-Line Treatment With Atezolizumab Plus Bevacizumab and Chemotherapy for US Patients With Metastatic, Persistent, or Recurrent Cervical Cancer: A Cost-Effectiveness Analysis." *Value in Health*, vol. 27, no. 11, 2024, pp. 1528–34, <https://doi.org/10.1016/j.jval.2024.07.013>.

# What is (deterministic) sensitivity analysis?

- SA evaluates the effects of input uncertainty on the conclusion.

Input parameters	Values
HIV Testing cost	\$5
HIV Prevalence	1%
HIV Testing sensitivity	80%
PreP cost	\$20
ART cost	\$50

Conclusion:

Testing every three years is cost-effective.

If the input values are not certain, how would this conclusion be different?

## One-way SA:

- Changing a single parameter value one by one.
- E.g.
  - 1) Testing cost: \$1
  - 2) Testing cost: \$10
  - 3) Prevalence: 0.1%
  - 4) Prevalence: 10%

...

## Two-way SA:

- Changing two parameter values together.
- E.g.
  - 1) Testing cost: \$1 & Prevalence: 0.1%
  - 2) Testing cost: \$1 & Prevalence: 10%
  - 3) Testing cost: \$10 & Prevalence: 0.1%
  - 4) Testing cost: \$10 & Prevalence: 10%

....

# Tornado graph (used in one-way sensitivity analyses)

Use case:

- Show the effect on a specific model output of changing inputs to the maximum or minimum value in a range.

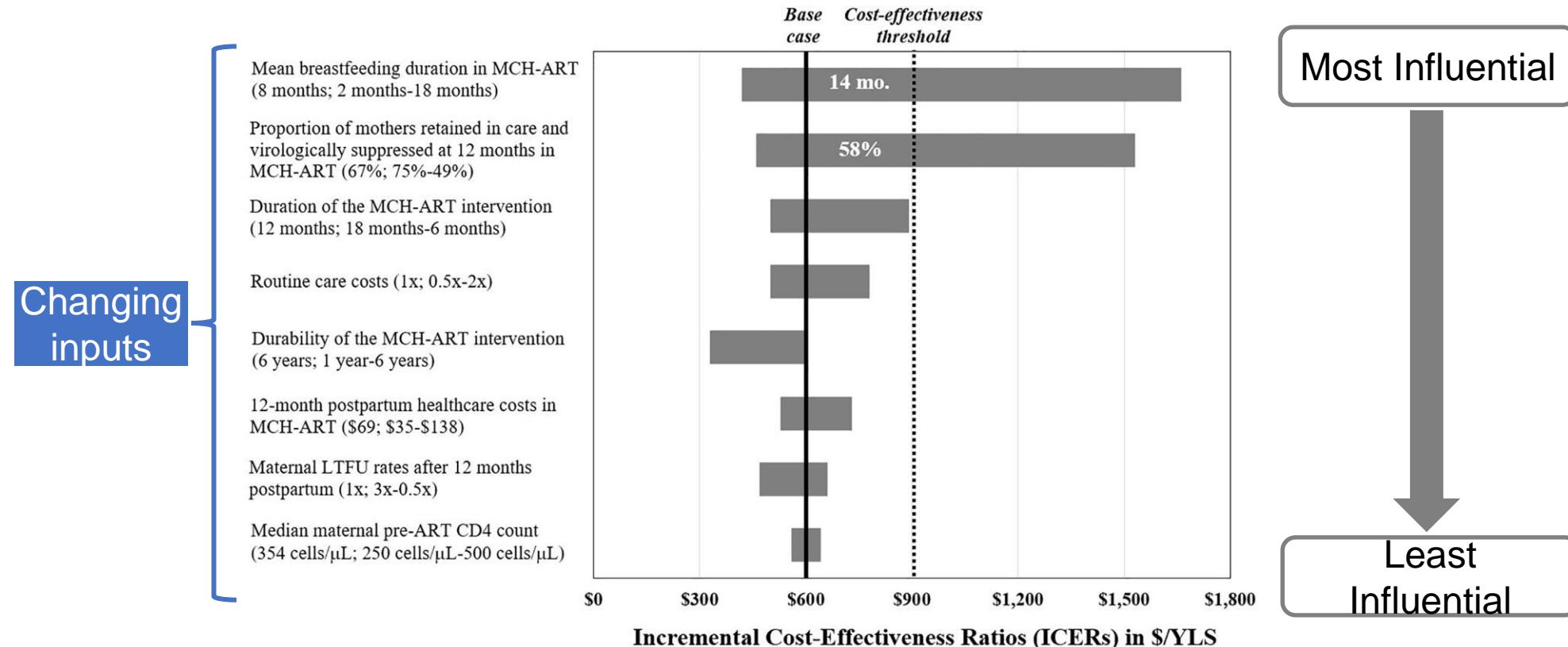


Fig 2. Tornado diagram of one-way sensitivity analyses.

From Dugdale CM, Phillips TK, Myer L, Hyle EP, Brittain K, Freedberg KA, et al. (2019) Cost-effectiveness of integrating postpartum antiretroviral therapy and infant care into maternal & child health services in South Africa. PLoS ONE 14(11): e0225104. <https://doi.org/10.1371/journal.pone.0225104>

# Heatmap (used in two-way sensitivity analyses)

Use case:

- Show which decision to choose as a function of two input parameters in two-way sensitivity analysis.
- The color of the cell shows the optimal decision conditioning on the selected values of the two input parameters.
- Typically compare the efficiency and cost of different strategies.

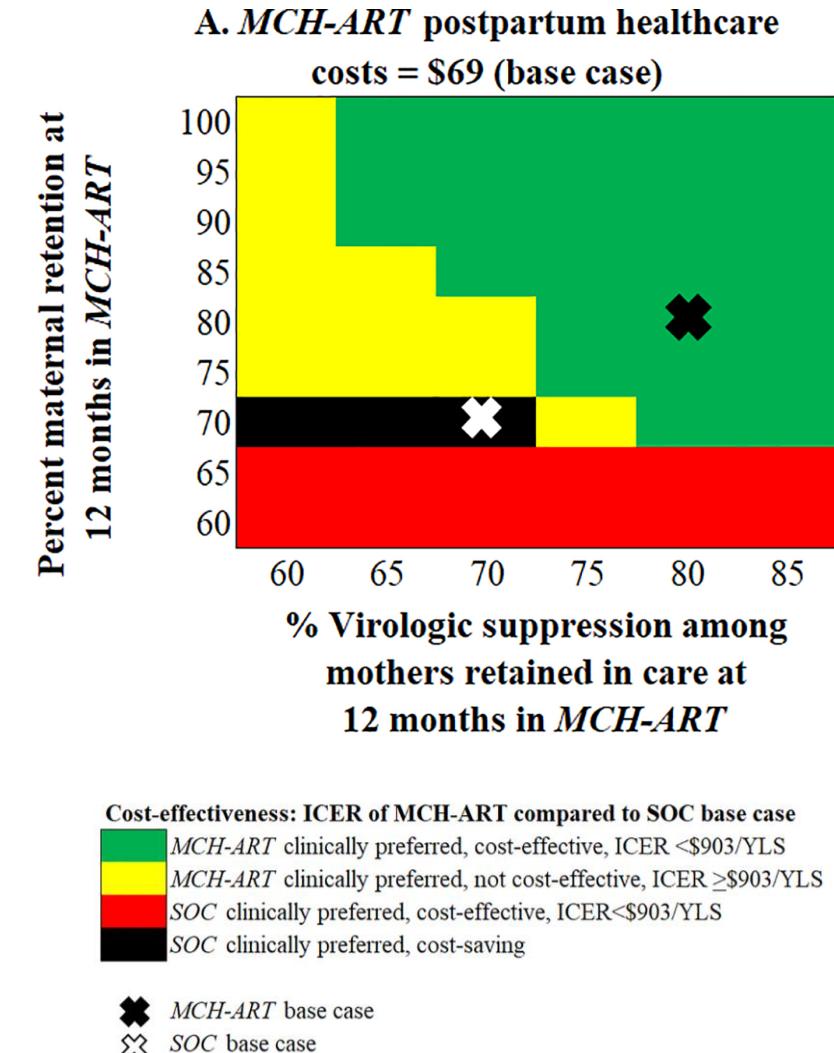


Fig 3. Multi-way sensitivity analyses.

From Dugdale CM, Phillips TK, Myer L, Hyle EP, Brittain K, Freedberg KA, et al. (2019) Cost-effectiveness of integrating postpartum antiretroviral therapy and infant care into maternal & child health services in South Africa. PLoS ONE 14(11): e0225104. <https://doi.org/10.1371/journal.pone.0225104>

# Forest plot

## Use case:

- Compare results in meta-analyses and systematic reviews.
- Display results (often odds/hazard ratios and relative risk) with the lower and upper bounds of each study.

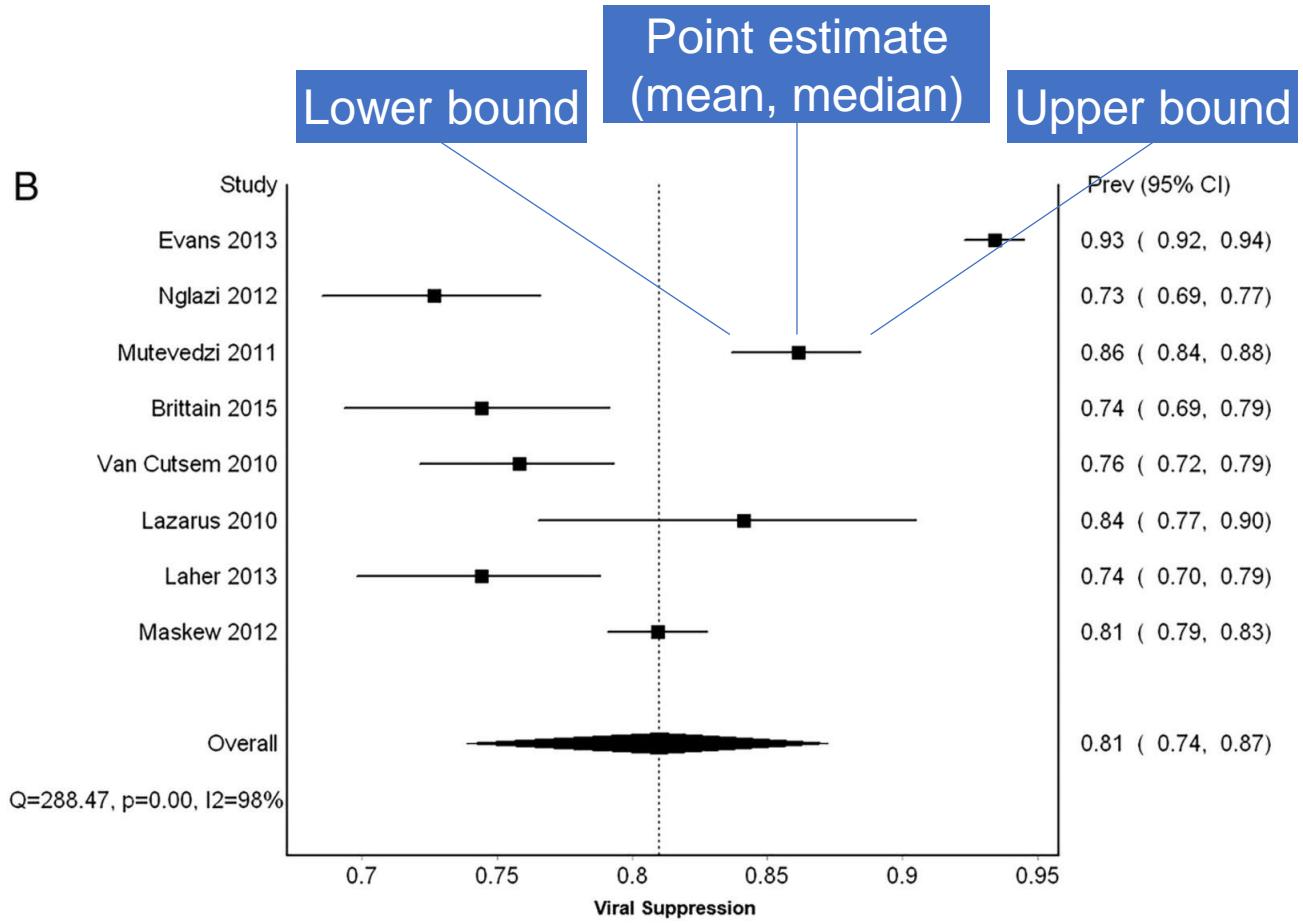


Figure 3 (B) Meta-analysis viral suppression: Forest plot of the proportion of virally suppressed HIV-infected adolescents and young adults in South Africa.

From Brian C Zanoni, Mohenndran Archary, Sarah Buchan, Ingrid T Katz, Jessica E Haberer - Systematic review and meta-analysis of the adolescent HIV continuum of care in South Africa: the Cresting Wave: BMJ Global Health 2016;1:e000004.

# Table of contents

---

## [Overview]

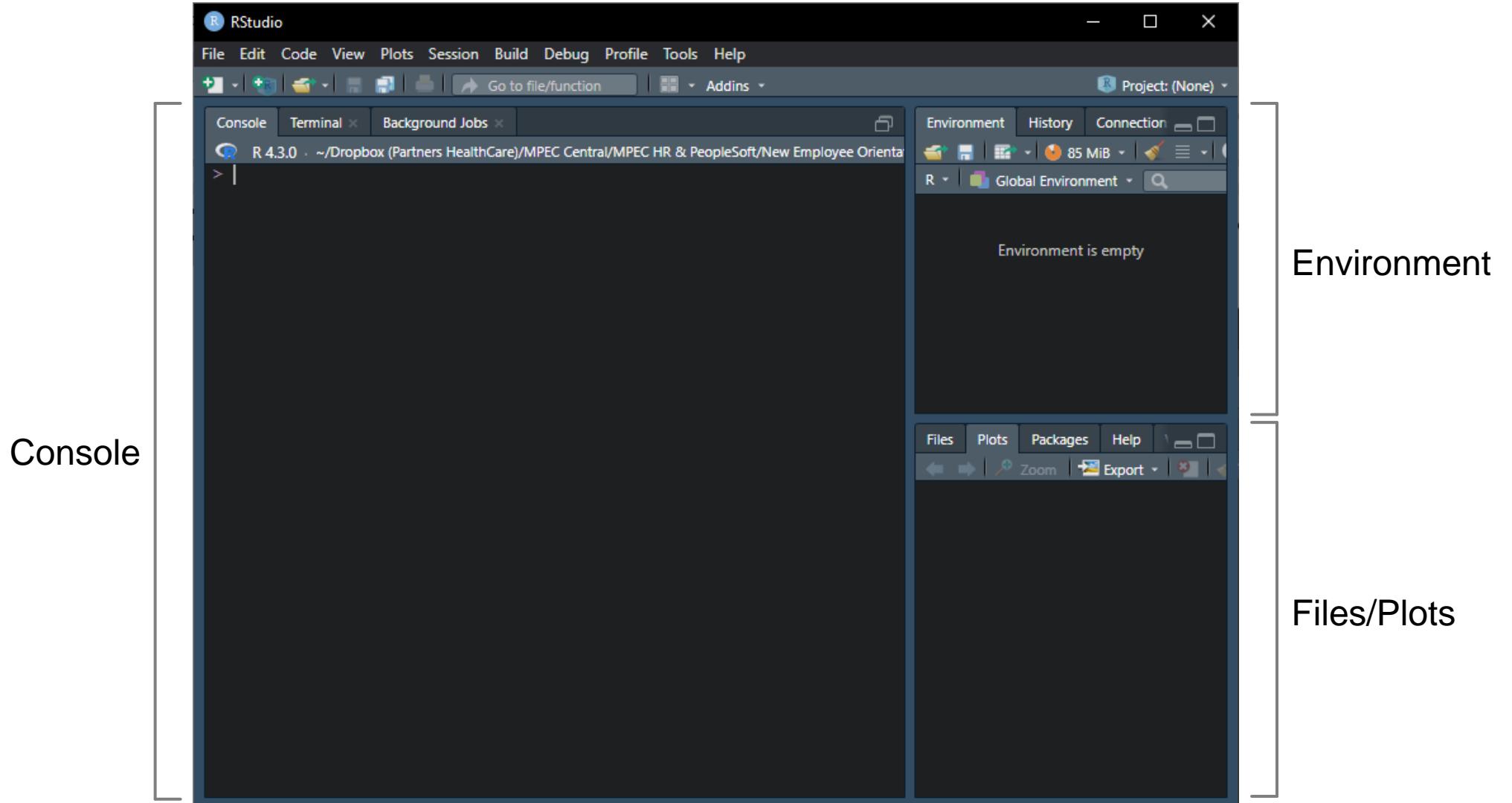
- R vs Excel vs STATA
- Figure Types and use cases

## [R]

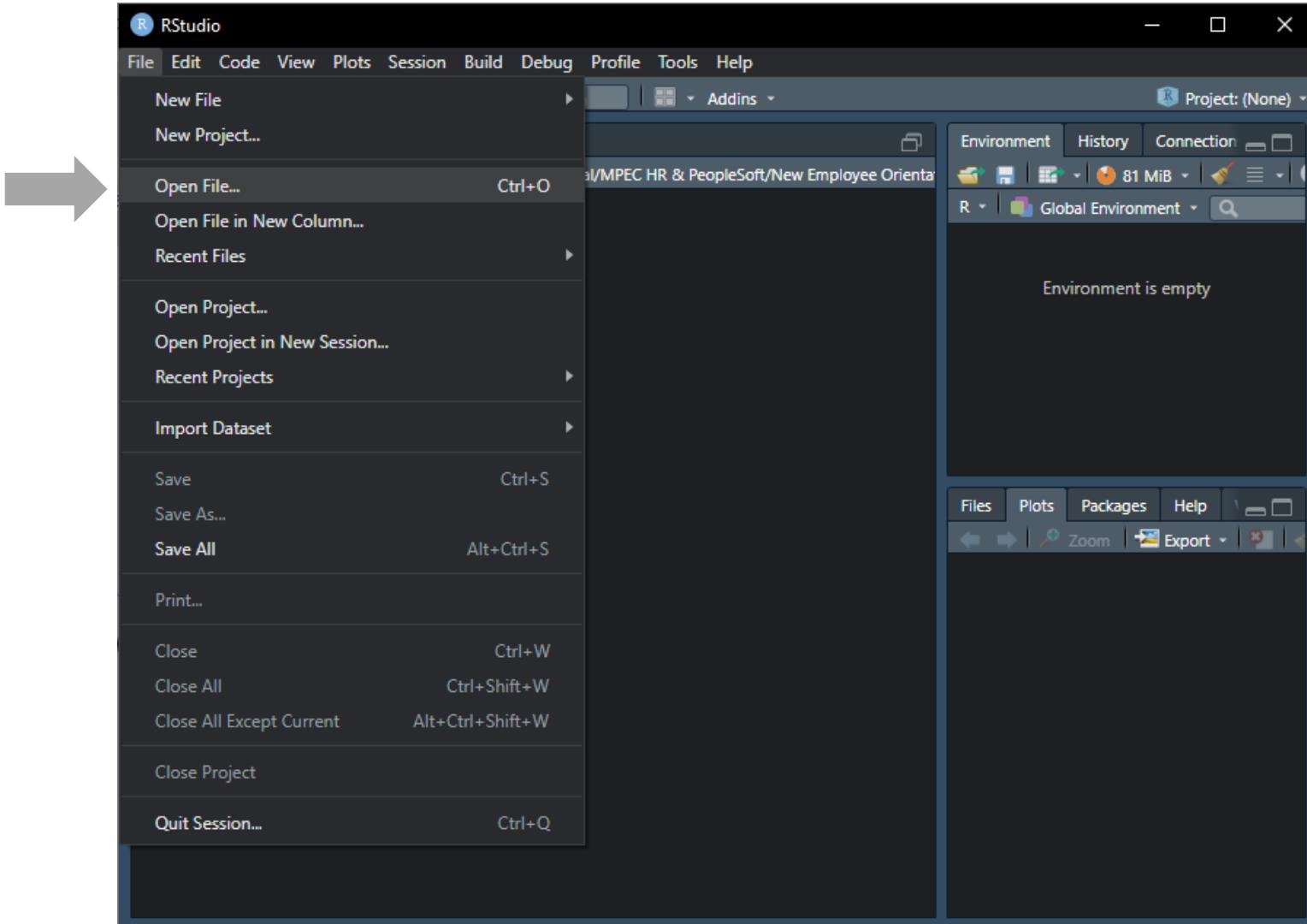
### • **Basic operation**

1. Histogram
2. Bar plot
  - 2.1. Stacked bar plot
3. Line graph
4. Scatter plot
5. Tornado plot
6. Heatmap

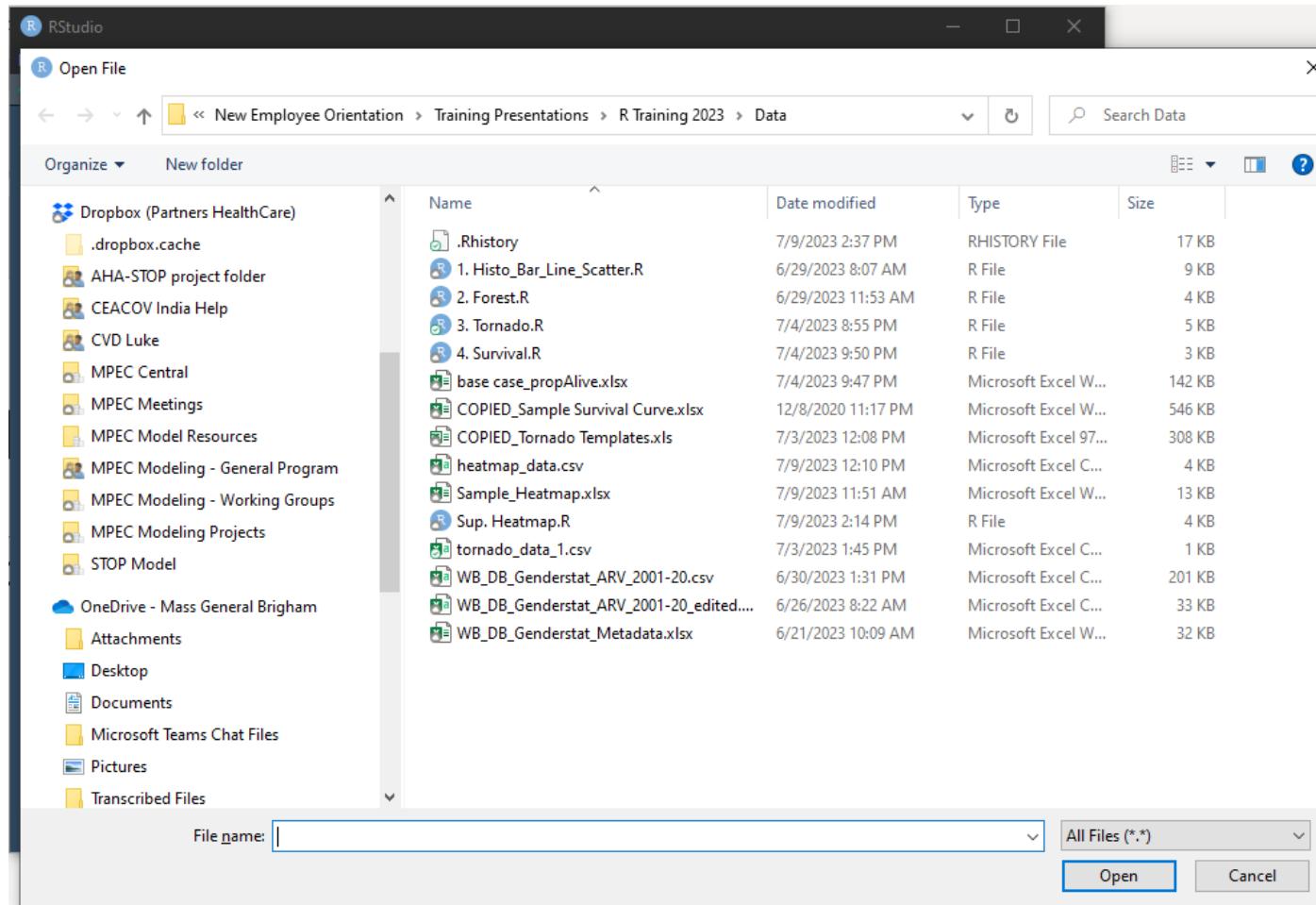
# RStudio Integrated Development Environment



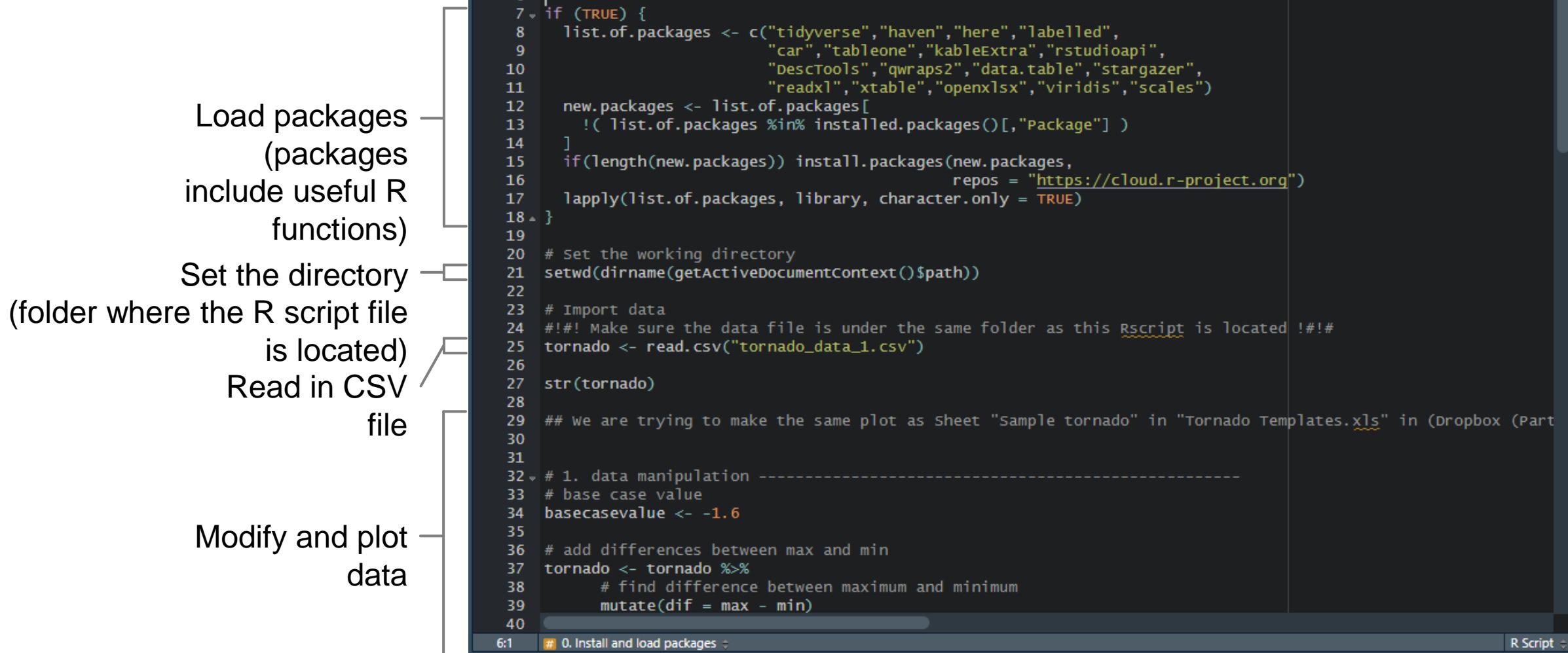
# Opening Script Step 1



# Opening Script Step 2



# Script components



# Running script

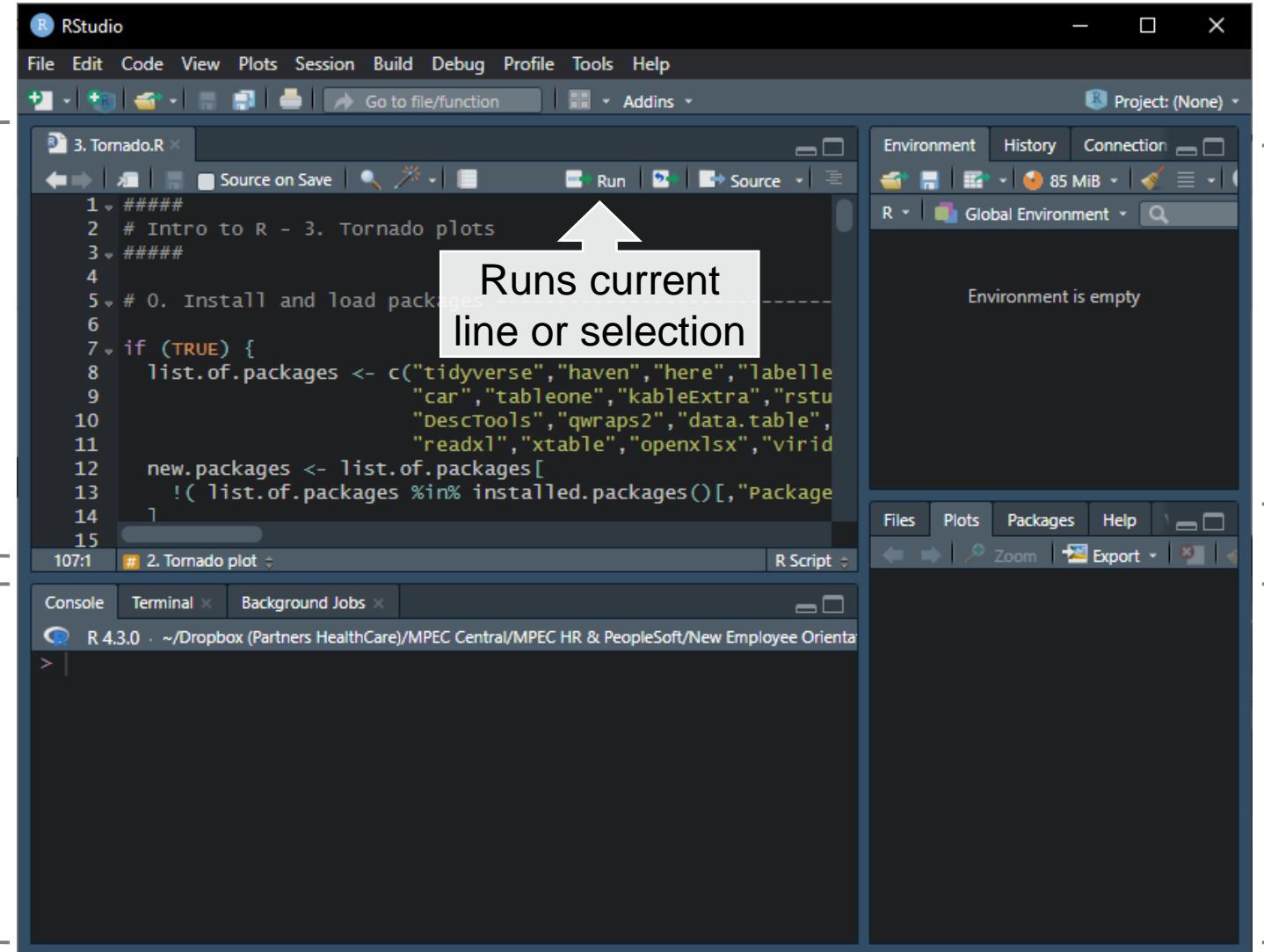
Source Editor

Runs current  
line or selection

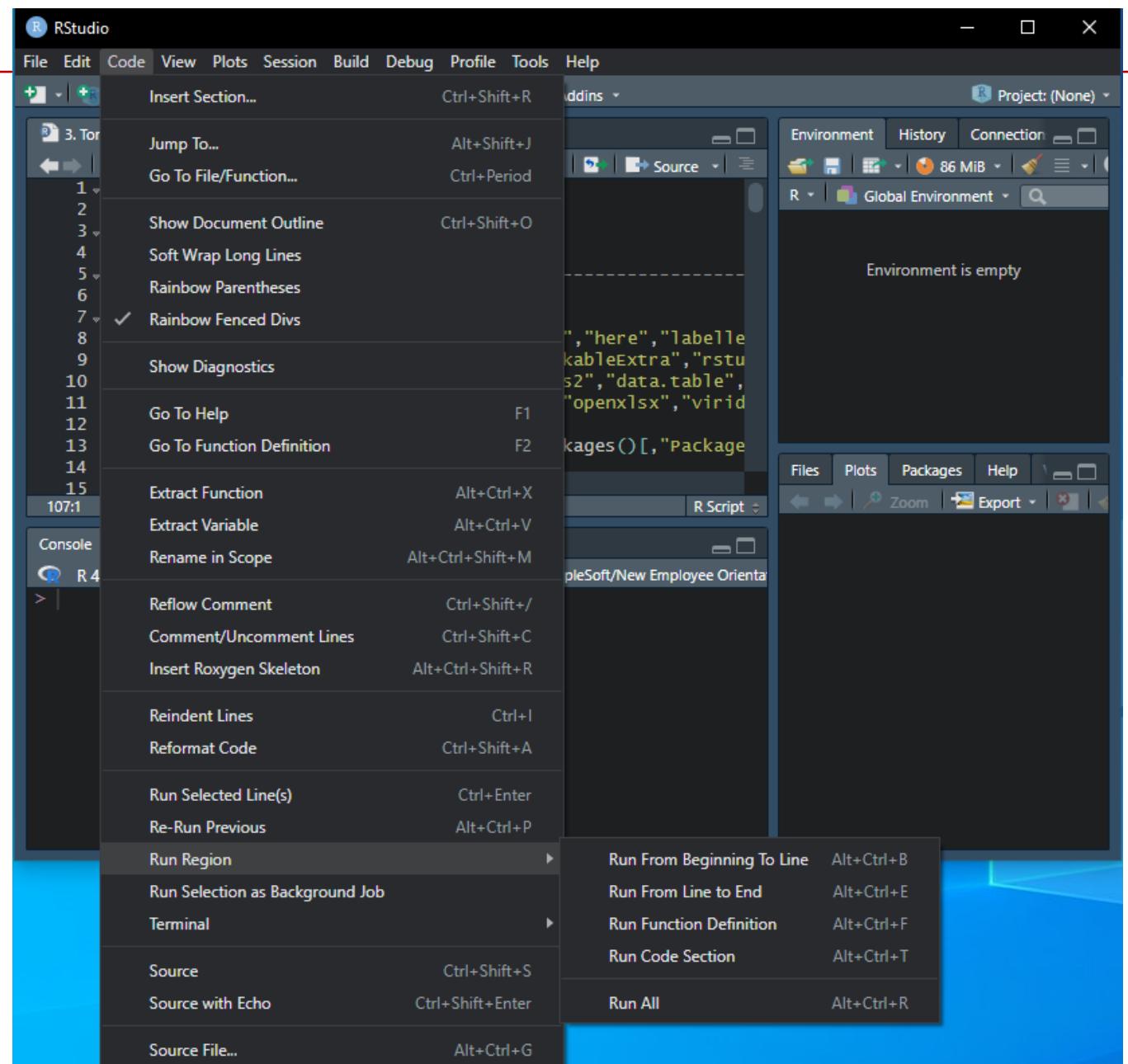
Console

Environment

Files/Plots



# Run options



# Output from Run All

Console

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

3.Tornado.R

```
repos = "https://cloud.r-project.org"
lapply(list.of.packages, library, character.only = TRUE)
# set the working directory
setwd(dirname(getActiveDocumentContext()$path))

# Import data
## Make sure the data file is under the same folder as this Rscript is located !##
tornado <- read.csv("tornado_data_1.csv")
str(tornado)

## We are trying to make the same plot as sheet "sample tornado" in "Tornado Templates.xls" in Dropbox (Part
30
31
32 # 1. data manipulation -----
33 # base case value
basecasevalue <- -1.6
35
36 # add differences between max and min
tornado <- tornado %>%
  # find difference between maximum and minimum
  mutate(dif = max - min)
40
41 # get the order of the parameters depending on the absolute values
42
```

0. Install and load packages

```
> # Set the working directory
> setwd(dirname(getActiveDocumentContext()$path))

> # Import data
> ## Make sure the data file is under the same folder as this Rscript is located !##
> tornado <- read.csv("tornado_data_1.csv")

> str(tornado)
'data.frame': 11 obs. of 3 variables:
 $ parameter: chr "Active TB Prevalence (0-80%" "chronic AIDS Mortality Misattribution (0-75%" "TB Treatment Success (0-100%" "Receiving TB Prophylaxis in SoC (100-0%" ...
 $ max      : num -11.91 -9.4 -1.98 -1.77 -2.33 ...
 $ min      : num -0.814 -1.6255 -0.0671 0.0352 -0.5761 ...

## We are trying to make the same plot as sheet "sample tornado" in "Tornado Templates.xls" in (Dropbox (Partners
Healthcare)\MPEC Central\Research .... [TRUNCATED]

> # add differences between max and min
> tornado <- tornado %>%
+   # find difference between maximum and minimum
+   mutate(dif = max - min)

> # get the order of the parameters depending on the absolute values
> param_order <- tornado %>%
+   # set the order of the parameters depending on .... [TRUNCATED]
```

Environment

History Connections Tutorial

Data

- tornado 11 obs. of 4 variables
- tornado\_plot 22 obs. of 8 variables

Values

- basecasevalue -1.6
- list.of.packages chr [1:17] "tidyverse" "haven" "here" "labelled" "car" "tableone" "kableExtra" "r\_
character (empty)
- new.packages character (empty)
- param\_order chr [1:11] "Initial CD4 (10-50)" "Latent TB Prevalence (0-60%" "Active TB Incide\_
width 0.95

Files Plots Packages Help Viewer Presentation

Effects of select parameters on absolute mortality decrease due to Empiric TB in the "Medium" Undiagnosed Active TB Prevalence Scenario (10.6%)

Base Case = -1.6

Absolute mortality decrease (%) due to Empiric TB

Environment

Files/Plots

# Environment pane

The screenshot shows the RStudio interface with the 'Environment' tab selected. The Global Environment pane displays various objects:

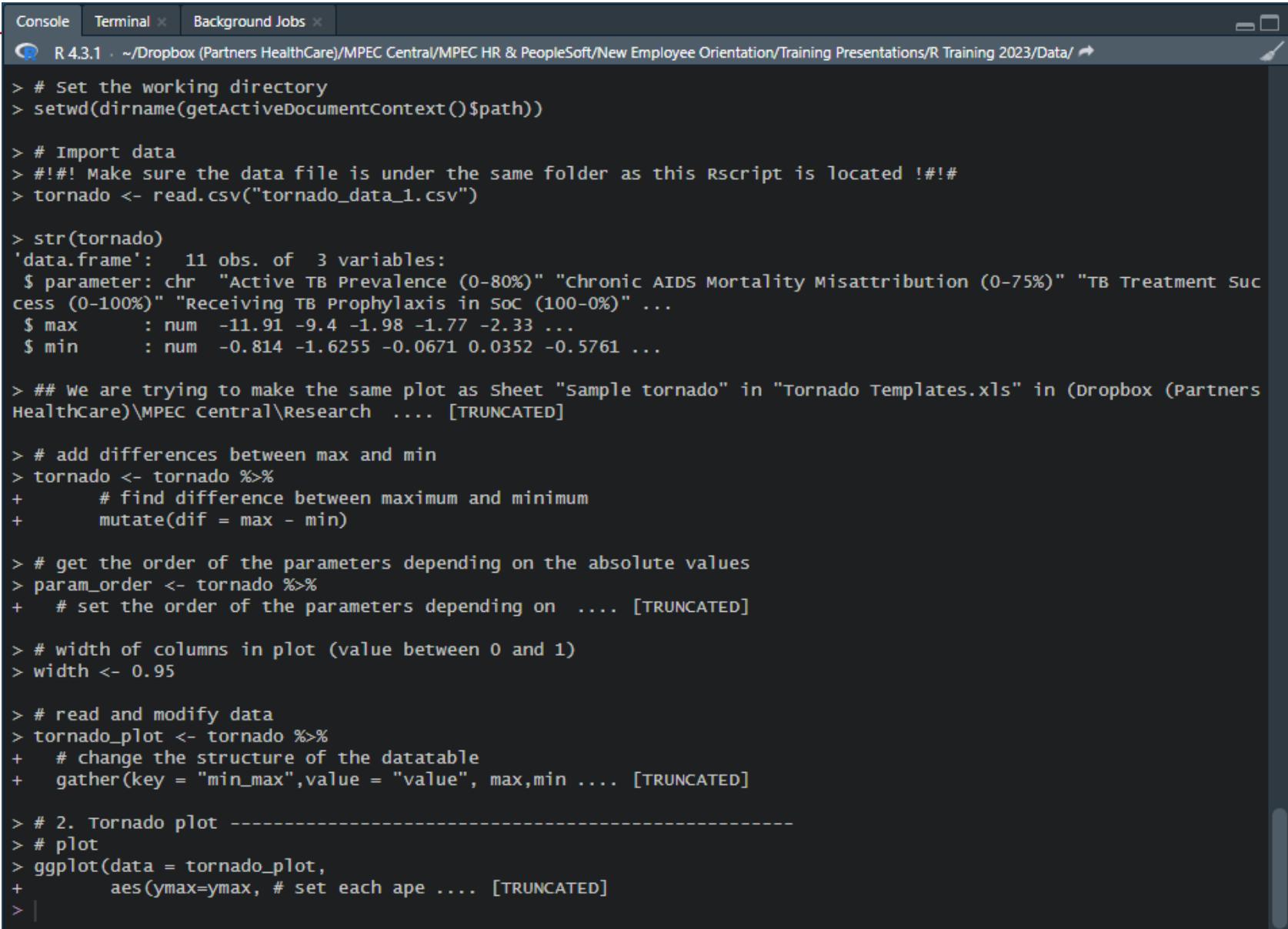
- Data**:
  - tornado: 11 obs. of 4 variables
  - tornado\_plot: 22 obs. of 8 variables
- values**:
  - basecasevalue: -1.6
  - list.of.packages: chr [1:17] "tidyverse" "haven" "here" "labelled" "car" "tableone" "kableExtra" "r...
  - new.packages: character (empty)
  - param\_order: chr [1:11] "Initial CD4 (10-50)" "Latent TB Prevalence (0-60%)" "Active TB Incide...
  - width: 0.95

Annotations on the left side group the objects:

- Tables of data** is associated with the 'tornado' and 'tornado\_plot' entries.
- Variables, vectors, lists, functions, etc.** is associated with the 'basecasevalue', 'list.of.packages', 'new.packages', 'param\_order', and 'width' entries.

# Console pane

- Displays commands that have been run or are being run
- Can directly type and run commands in the console



The screenshot shows the RStudio interface with the 'Console' tab selected. The console window displays R code and its output. The code is used to set the working directory, import data from a CSV file, and perform various data manipulations like calculating differences between max and min values and creating a plot. The output includes the structure of the imported data frame and truncated sections of the code.

```
R 4.3.1 - ~/Dropbox (Partners HealthCare)/MPEC Central/MPEC HR & PeopleSoft/New Employee Orientation/Training Presentations/R Training 2023/Data/ ↗

> # Set the working directory
> setwd(dirname(getActiveDocumentContext()$path))

> # Import data
> #!!! Make sure the data file is under the same folder as this Rscript is located !#!#
> tornado <- read.csv("tornado_data_1.csv")

> str(tornado)
'data.frame': 11 obs. of 3 variables:
 $ parameter: chr "Active TB Prevalence (0-80)" "Chronic AIDS Mortality Misattribution (0-75)" "TB Treatment Success (0-100)" "Receiving TB Prophylaxis in soc (100-0)" ...
 $ max       : num -11.91 -9.4 -1.98 -1.77 -2.33 ...
 $ min       : num -0.814 -1.6255 -0.0671 0.0352 -0.5761 ...

> ## we are trying to make the same plot as sheet "sample tornado" in "Tornado Templates.xls" in (Dropbox (Partners HealthCare)\MPEC Central\Research .... [TRUNCATED]

> # add differences between max and min
> tornado <- tornado %>%
+   # find difference between maximum and minimum
+   mutate(dif = max - min)

> # get the order of the parameters depending on the absolute values
> param_order <- tornado %>%
+   # set the order of the parameters depending on .... [TRUNCATED]

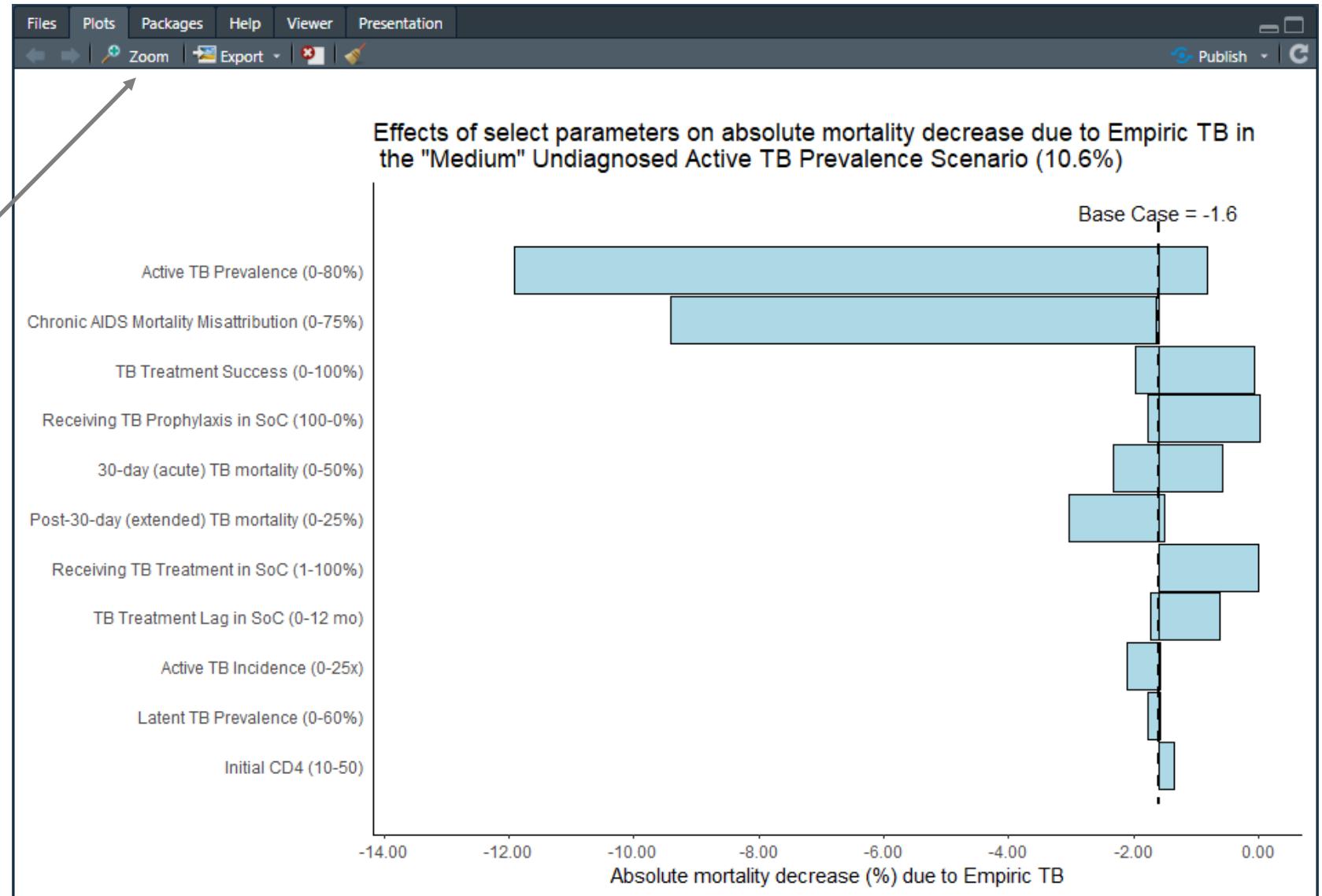
> # width of columns in plot (value between 0 and 1)
> width <- 0.95

> # read and modify data
> tornado_plot <- tornado %>%
+   # change the structure of the datatable
+   gather(key = "min_max", value = "value", max,min .... [TRUNCATED]

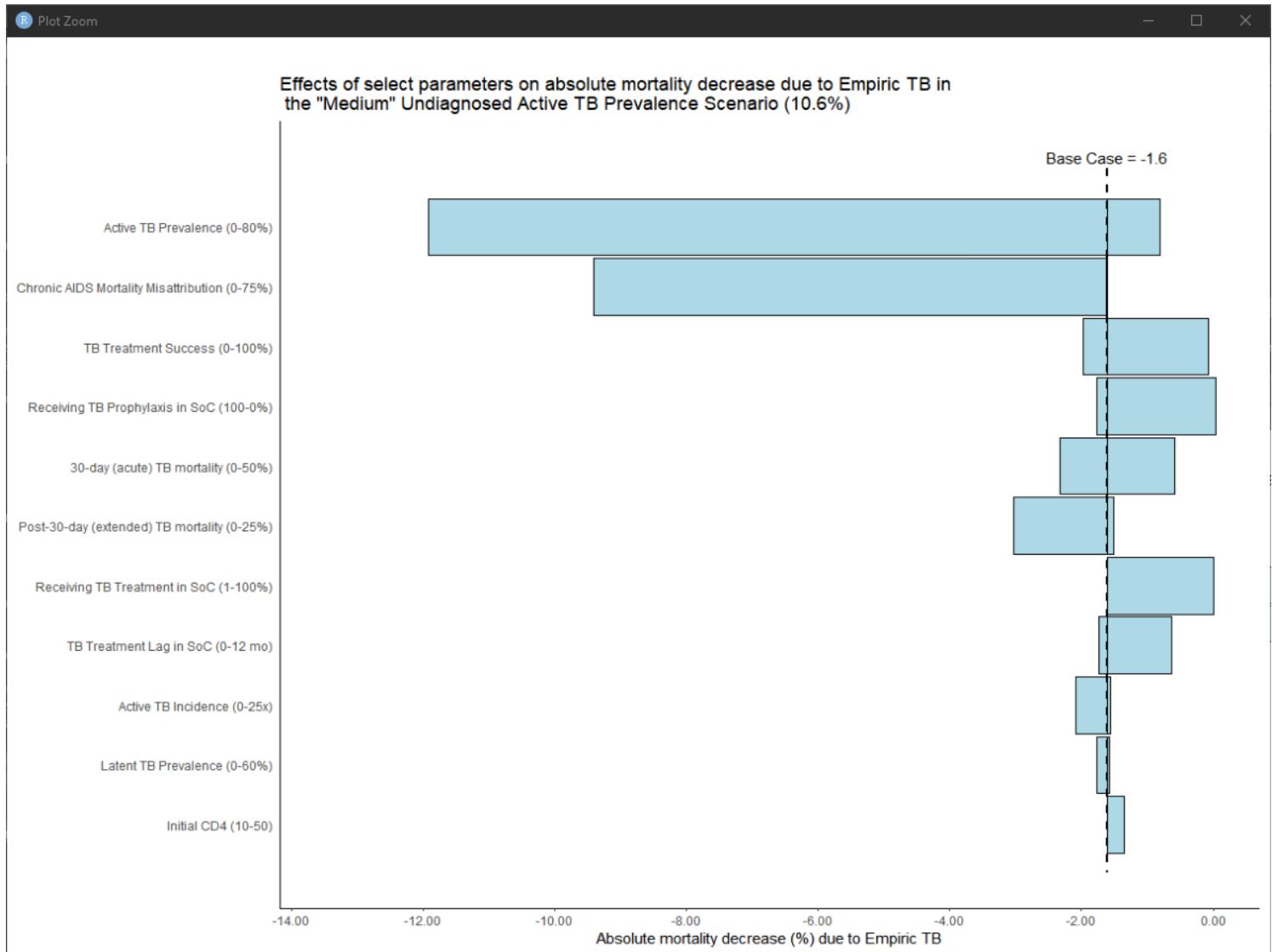
> # 2. Tornado plot -----
> # plot
> ggplot(data = tornado_plot,
+         aes(ymax=ymax, # set each ape .... [TRUNCATED]
> |
```

# Files/Plots pane

Opens figure in  
a separate  
window



# Plot Zoom



# Basic Operation in R

Command	Meaning	Example
Line starting with #	Comments that don't influence functions or operations.	Line 23 and 24 in "R_Data_Vis.R" <pre>23 # read data from a csv file 24 # Here the data set has the percent of the population with access to ARV by    gender of countries from 2001 to 2020 (World Bank) 25 arv &lt;- read.csv("Supp.WB_DB_Genderstat_ARV_2001-20.csv", header = T, encoding =    "UTF-8")</pre>
a <- b	Assign b to a. b can be the results of functions and calculations. e.g., a <- 1 + 1, then typing "a" will give 2.	Line 25 in "R_Data_Vis.R" <pre>23 # read data from a csv file 24 # Here the data set has the percent of the population with access to ARV by    gender of countries from 2001 to 2020 (World Bank) 25 arv &lt;- read.csv("Supp.WB_DB_Genderstat_ARV_2001-20.csv", header = T, encoding =    "UTF-8")</pre> 
function(arguments)	The <b>function</b> works by following specified <b>arguments</b>	Line 25 <b>read.csv</b> in "R_Data_Vis.R" <pre>23 # read data from a csv file 24 # Here the data set has the percent of the population with access to ARV by    gender of countries from 2001 to 2020 (World Bank) 25 arv &lt;- <b>read.csv("Supp.WB_DB_Genderstat_ARV_2001-20.csv", header = T, encoding =    "UTF-8")</b></pre>
%>% or  >	To link one function to another.	Line 221 in "R_Data_Vis.R" <pre>220 # change the shape of the data table to put the percentages of female and male in    different columns 221 arv_2020_spread &lt;- arv_2020 <b>%&gt;%</b> 222   <b>select(-c(SeriesCode,n)) %&gt;%</b> 223   spread(key = gender, value = value)</pre>

# Frequently used functions

Function	Purpose
<b>read.csv()</b>	To read data from csv. file
<b>read_excel()</b>	To read data from MS Excel file
<b>str()</b>	To check the data and type of data of each column
<b>view()</b>	To check the whole data in a new tab, which looks similar to an Excel format
<b>ggplot()</b>	To make plots. Only ggplot function doesn't work. You need to specify what type of graphs you want to make with additional functions with “+”. (Not “%>%”!)

**Google whenever you come across unknown functions!**

# Table of contents

---

## [Overview]

- R vs Excel vs STATA
- Figure Types and use cases

## [R]

- Basic operation

### **1. Histogram**

2. Bar plot
  - 2.1. Stacked bar plot
3. Line graph
4. Scatter plot
5. Tornado plot
6. Heatmap

# 0. Install and load packages (and data!)

- 0. Install and load packages (and data!)
  - Don't change anything until Line 19 "setwd(dirname(...))"
  - Import data from csv. data on percentage of people with access to ARV 2001-2020 (World Bank): read.csv()
  - Check the imported data
    - str()

```
'data.frame': 2440 obs. of 9 variables:  
 $ n      : int 1 2 9 10 11 12 25 26 29 30 ...  
 $ SeriesName : chr "Access to anti-retroviral drugs" "Access to anti-retroviral dr  
 anti-retroviral drugs" ...  
 $ gender   : chr "female" "male" "female" "male" ...  
 $ unit     : chr "%" "%" "%" "%" ...  
 $ SeriesCode: chr "SH.HIV.ARVC.FE.ZS" "SH.HIV.ARVC.MA.ZS" "SH.HIV.ARVC.FE.ZS" "SH  
 $ CountryName: chr "Afghanistan" "Afghanistan" "Argentina" "Argentina" ...  
 $ CountryCode: chr "AFG" "AFG" "ARG" "ARG" ...  
 $ year     : int 2001 2001 2001 2001 2001 2001 2001 2001 2001 2001 ...  
 $ value    : int 0 0 20 23 54 50 0 0 0 0 ...
```

A	B	C	D	E	F	G	H	I
1	SeriesName	gender	unit	SeriesCode	CountryName	CountryCode	year	value
2	1 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Afghanistan	AFG	2001	0
3	2 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Afghanistan	AFG	2001	0
4	9 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Argentina	ARG	2001	20
5	10 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Argentina	ARG	2001	23
6	11 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Australia	AUS	2001	54
7	12 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Australia	AUS	2001	50
8	25 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Benin	BEN	2001	0
9	26 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Benin	BEN	2001	0
10	29 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Bolivia	BOL	2001	0
11	30 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Bolivia	BOL	2001	0
12	31 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Botswana	BWA	2001	1
13	32 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Botswana	BWA	2001	1
14	35 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Burkina Faso	BFA	2001	0
15	36 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Burkina Faso	BFA	2001	0
16	39 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Cabo Verde	CPV	2001	0
17	40 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Cabo Verde	CPV	2001	0
18	49 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Chile	CHL	2001	12
19	50 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Chile	CHL	2001	12
20	51 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Colombia	COL	2001	0
21	52 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Colombia	COL	2001	0
22	61 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Cote d'Ivoire	CIV	2001	0
23	62 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Cote d'Ivoire	CIV	2001	0
24	65 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Cuba	CUB	2001	4
25	66 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Cuba	CUB	2001	3
26	69 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Denmark	DNK	2001	47
27	70 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Denmark	DNK	2001	46
28	85 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Denmark	DNK	2001	0

# 0. Install and load packages (and data!)

- view()

Screenshot of a Jupyter Notebook interface showing a data frame:

```

Untitled2* x travel_modeling.Rmd x 1. Histo_Bar_Line_Scatter.R* x arv x Untitled1* x
  ▲ n SeriesName gender unit SeriesCode CountryName CountryCode year value
  1 1 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Afghanistan AFG 2001 0
  2 2 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Afghanistan AFG 2001 0
  3 9 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Argentina ARG 2001 20
  4 10 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Argentina ARG 2001 23
  5 11 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Australia AUS 2001 54
  6 12 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Australia AUS 2001 50
  7 25 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Benin BEN 2001 0
  8 26 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Benin BEN 2001 0
  9 29 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Bolivia BOL 2001 0
  10 30 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Bolivia BOL 2001 0
  11 31 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Botswana BWA 2001 1
  12 32 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Botswana BWA 2001 1
  13 35 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Burkina Faso BFA 2001 0
  14 36 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Burkina Faso BFA 2001 0
  15 39 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Cabo Verde CPV 2001 0
  16 40 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Cabo Verde CPV 2001 0
  17 49 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Chile CHL 2001 12
  18 50 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Chile CHL 2001 12
  19 51 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Colombia COL 2001 0
  20 52 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Colombia COL 2001 0
  21 61 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Cote d'Ivoire CIV 2001 0
  22 62 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Cote d'Ivoire CIV 2001 0
  23 65 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Cuba CUB 2001 4
  24 66 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Cuba CUB 2001 3
  25 69 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Denmark DNK 2001 47
  26 70 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Denmark DNK 2001 46
  27
  
```

Screenshot of a Jupyter Notebook interface showing a data frame:

```

A   B   C   D   E   F   G   H   I
1  n SeriesName gender unit SeriesCode CountryName countryCode year value
  2  1 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Afghanistan AFG 2001 0
  3  2 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Afghanistan AFG 2001 0
  4  9 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Argentina ARG 2001 20
  5 10 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Argentina ARG 2001 23
  6 11 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Australia AUS 2001 54
  7 12 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Australia AUS 2001 50
  8 25 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Benin BEN 2001 0
  9 26 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Benin BEN 2001 0
  10 29 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Bolivia BOL 2001 0
  11 30 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Bolivia BOL 2001 0
  12 31 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Botswana BWA 2001 1
  13 32 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Botswana BWA 2001 1
  14 35 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Burkina Faso BFA 2001 0
  15 36 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Burkina Faso BFA 2001 0
  16 39 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Cabo Verde CPV 2001 0
  17 40 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Cabo Verde CPV 2001 0
  18 49 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Chile CHL 2001 12
  19 50 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Chile CHL 2001 12
  20 51 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Colombia COL 2001 0
  21 52 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Colombia COL 2001 0
  22 61 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Cote d'Ivoire CIV 2001 0
  23 62 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Cote d'Ivoire CIV 2001 0
  24 65 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Cuba CUB 2001 4
  25 66 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Cuba CUB 2001 3
  26 69 Access to anti-retroviral drugs male % SH.HIV.ARTC.MA.ZS Denmark DNK 2001 47
  27 70 Access to anti-retroviral drugs female % SH.HIV.ARTC.FE.ZS Denmark DNK 2001 46
  28
  
```

# 1. Histogram

- 1. Histogram
  - Aim: To visualize the distribution of countries by percentage of females who have access to ARV in the world in 2020
  - Subset data with “female” in “gender” column and “2020” in “year” column

```
38 # select data of female in 2020
39 arv_fe_2020 <- arv[arv$gender=="female" & arv$year == 2020, ]
```
  - (Check data)

Some ideas for playing around:

- How about “male” in “gender”?
- How about other years?

# 1. Histogram

- 1. Histogram

- Plot the data: here **ggplot()** comes!

- Specify the data ("arv\_fe\_2020") and which variable is used on the x-axis ("value").
    - Specify the type of graph: **geom\_histogram()**
    - Save the plot in a png. format

```
84 # save the figure in png  
85 ggsave("hist.png", width = 8, height = 6)
```

```
44 # plot  
45 ggplot(data=arv_fe_2020, # specify the data you want to use  
46         aes(x=value)) + # specify the column that you want to show in the x-axis  
47         # specify the type of your graph  
48         geom_histogram(  
49             fill = "grey", # which color is used to fill the bins  
50             color = "black", # which color is used to draw the borders of bins  
51             bins = 10,       # number of bins  
52         ) +  
53         # add title  
54         ggtitle("Distribution of the percentage of female who have access to anti-retr  
+ "#"\n" means change a line  
55         # add x-axis label  
56         xlab("Percentage (%))" ) +  
57         # add y-axis label  
58         ylab("Number of countries" ) +  
59         # customize the background theme  
60         theme_bw()
```

Some ideas for playing around:

- How about other colors inside bins?
- How about other colors for the borders of the bins?
- How about different numbers of bins?
- How about different titles and labels in the x/y-axes?

# Table of contents

---

## [Overview]

- R vs Excel vs STATA
- Figure Types and use cases

## [R]

- Basic operation
- 1. Histogram
- 2. Bar plot**
  - 2.1. Stacked bar plot
- 3. Line graph
- 4. Scatter plot
- 5. Tornado plot
- 6. Heatmap

## 2. Bar chart

- 2. Bar chart
  - Aim: To visualize the percentage of females who have access to ARV in 2020 by country
  - Use the same data as the histogram (“arv\_fe\_2020”)
  - Plot the data: **ggplot()**
    - Specify the data (“arv\_fe\_2020”) and which variable is used on the x-axis (“CountryName”) and the y-axis (“value”).
    - Specify the type of graph: **geom\_bar()**

```
95 # plot
96 ggplot(data=arv_fe_2020, # specify data
97         aes(x=CountryName, # specify the column that you want to show in the x-axis
98             y=value)) +    # specify the column that you want to show in the y-axis
```

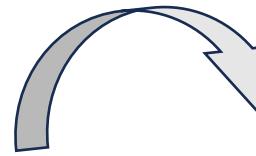
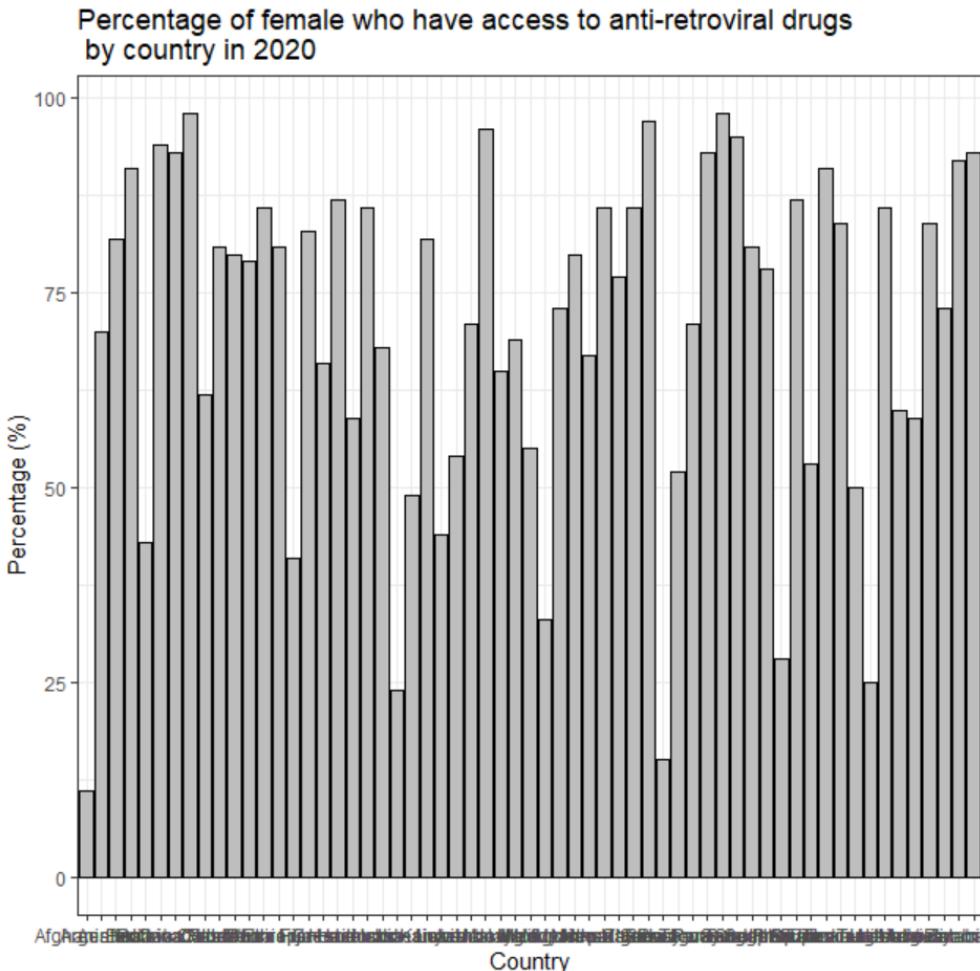
## 2. Bar chart

- 2. Bar chart
  - Specify the type of graph: **geom\_bar()**
    - **Don't forget “stat = “identity””!**
    - Without it, the plot will show the counts of the variable of the x-axis.

```
99  # specify the type of your graph
100 geom_bar(
101   fill = "grey", # which color is used to fill the bins
102   color = "black", # which color is used to draw the borders of bins
103   stat = "identity" # when you specify x and y in geom_bar, don't forget this argument
104 ) +
105 # add title
106 ggtile("Percentage of female who have access to anti-retroviral drugs \n by country in"
107 # add x-axis label
108 xlab("Country") +
109 # add y-axis label
110 ylab("Percentage (%)") +
111 # customize the background theme
112 theme_bw()
```

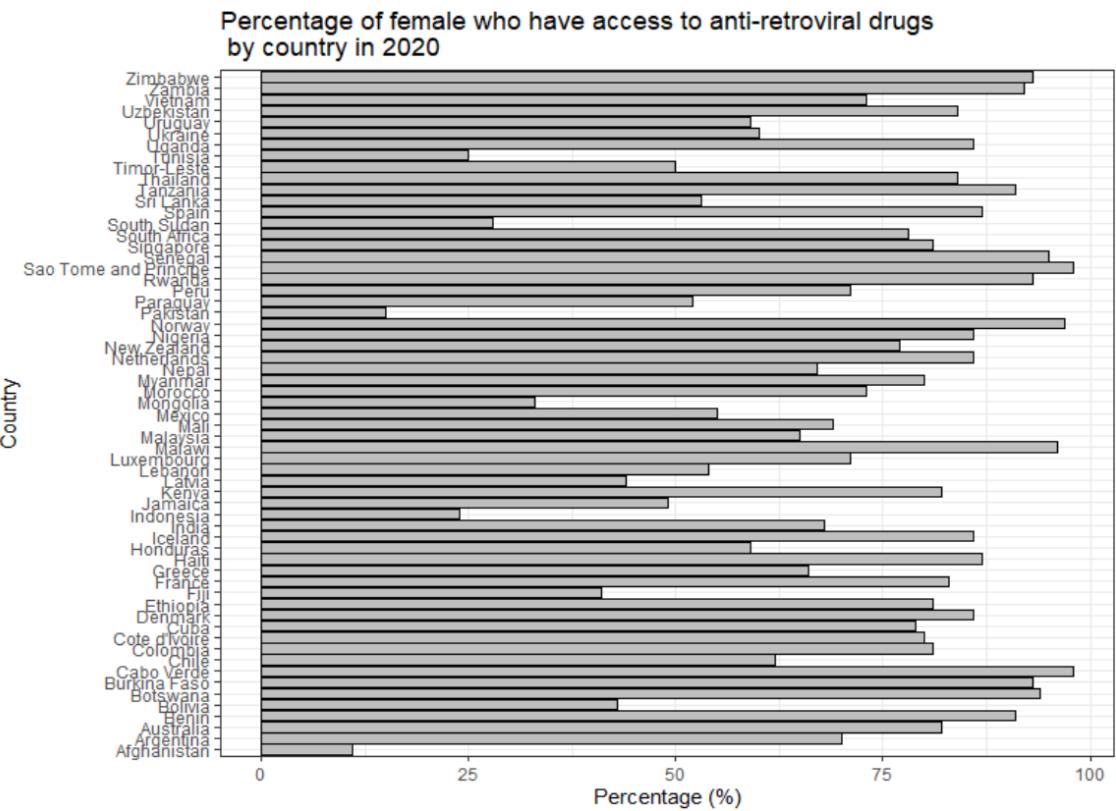
## 2. Bar chart

## • 2. Bar chart



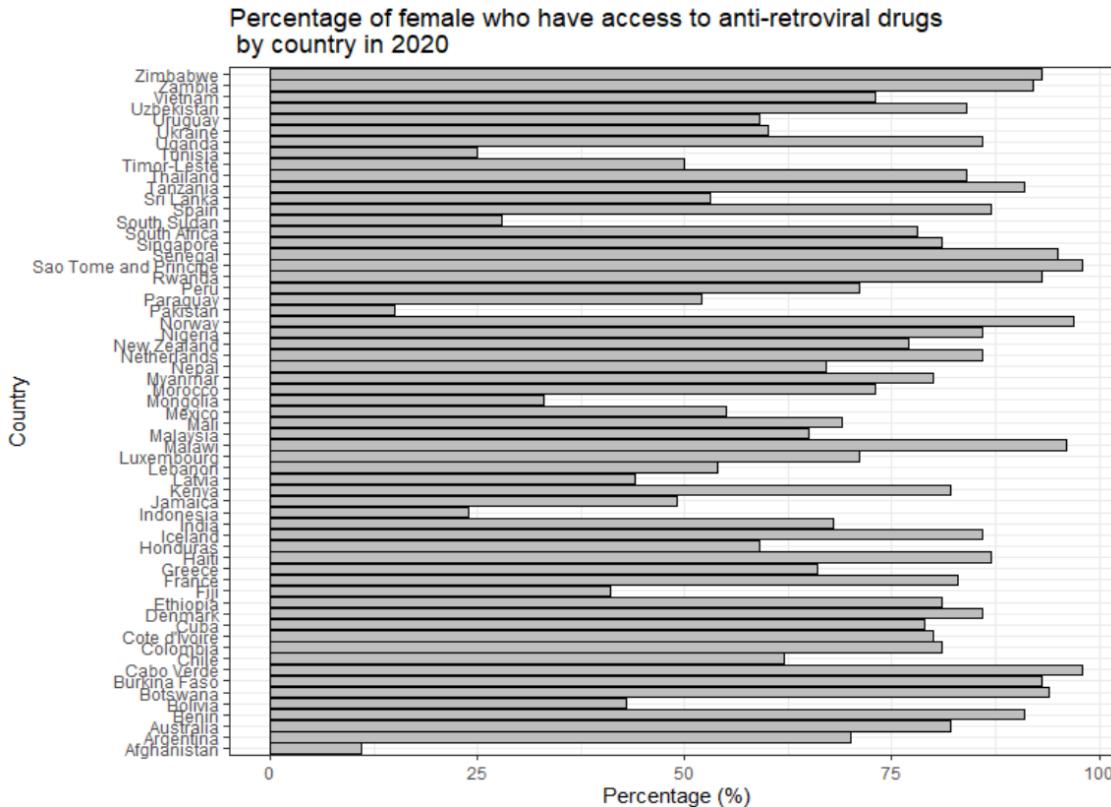
134

```
coord_flip() # to switch x and y
```

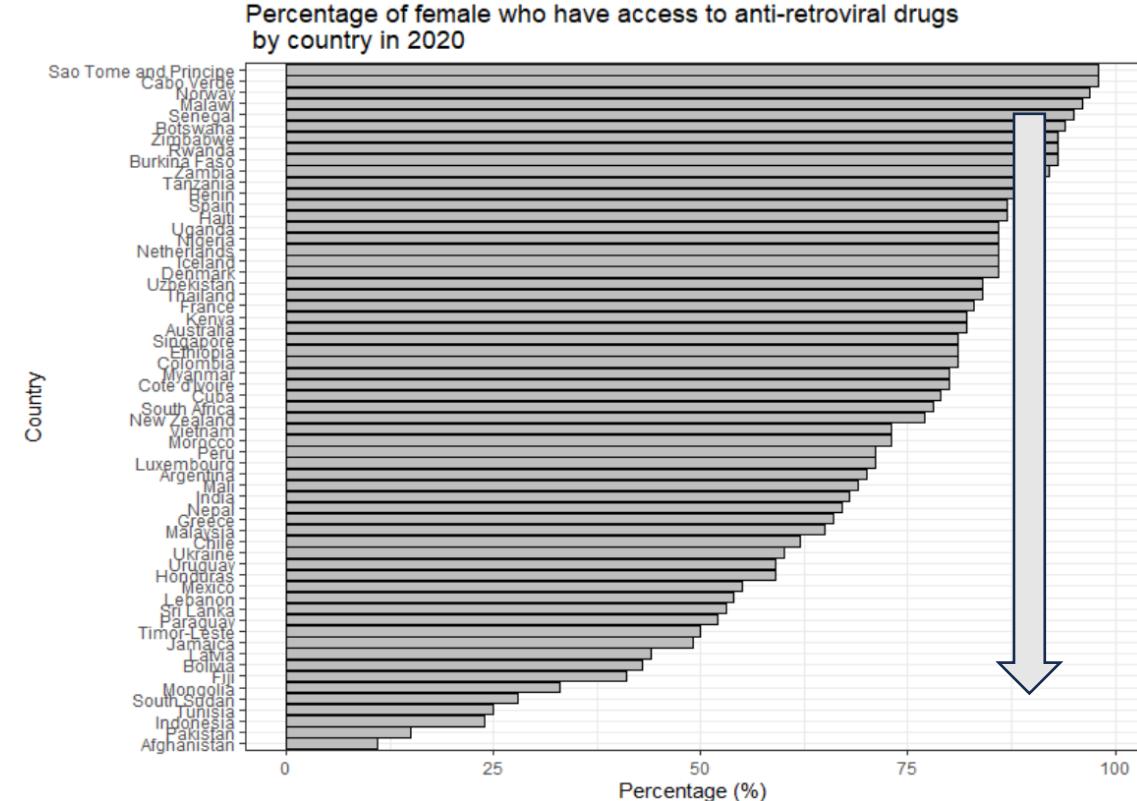


## 2. Bar chart

- 2. Bar chart



```
137 ggplot(data=arv_fe_2020, # specify data  
138   aes(x=reorder(CountryName, value), # x-axis, r  
139     y=value)) +    # specify the column that y
```



## 2.1. Stacked bar chart

- 2.1. Stacked bar chart
  - Aim: To visualize the percentage of females and males who have access to ARV in 2020 by country
  - Subset data with “2020” in “year” column

```
161 # select data of female & male in 2020
162 arv_2020 <- arv[arv$year == 2020, ]
```
  - (Check data)

Some ideas for playing around:

- How about other years?

# Table of contents

---

## [Overview]

- R vs Excel vs STATA
- Figure Types and use cases

## [R]

- Basic operation
- 1. Histogram
- 2. Bar plot
  - 2.1. Stacked bar plot**
- 3. Line graph
- 4. Scatter plot
- 5. Tornado plot
- 6. Heatmap

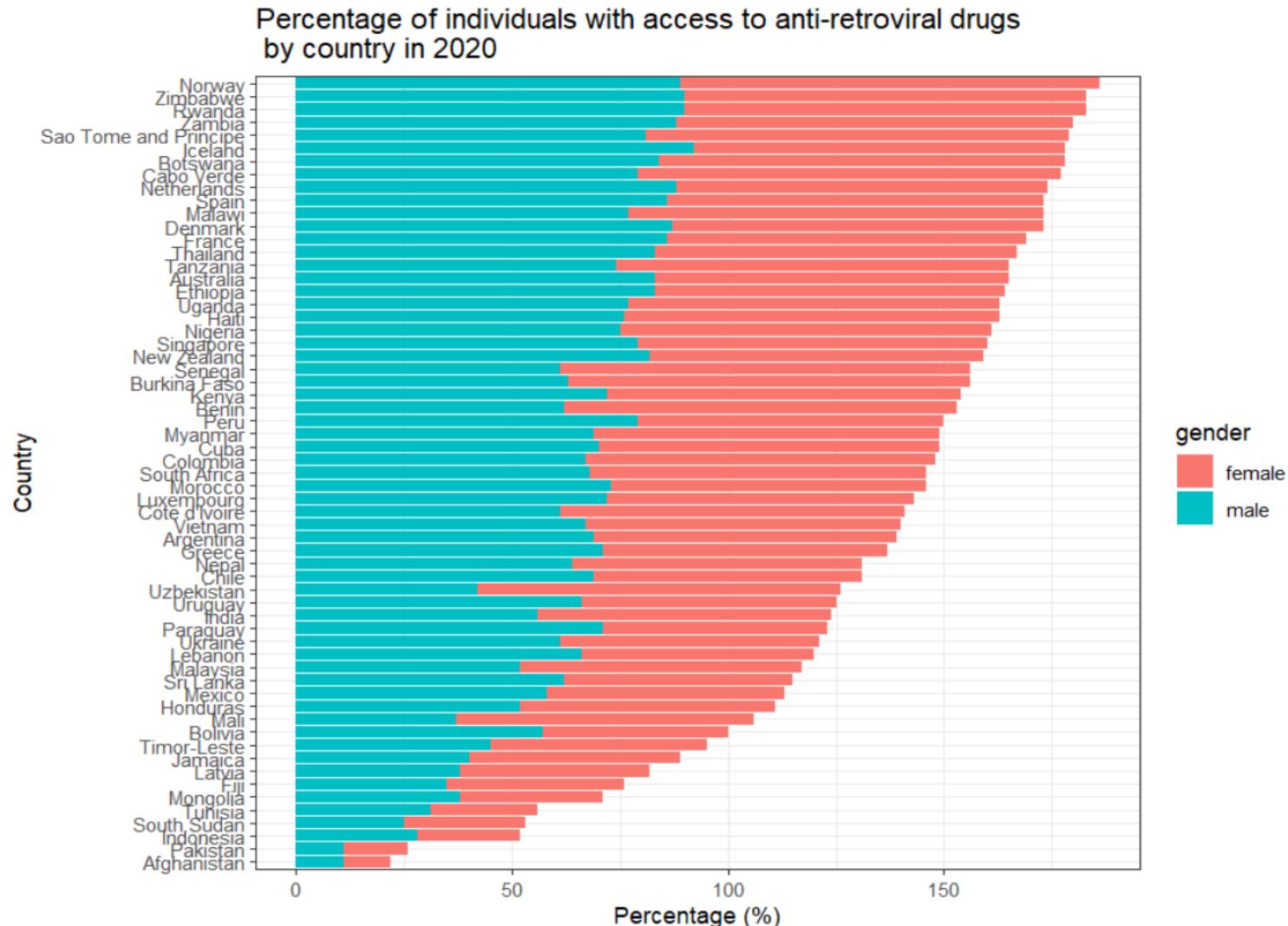
## 2.1. Stacked bar chart

- 2.1. Stacked bar chart
  - Plot the data: **ggplot()**
    - Specify the data (“arv\_fe\_2020”) and which variable is used on the x-axis (“CountryName” with `reorder()`), the y-axis (“value”), and the color to fill the bars (“gender”)
    - Specify the type of graph: **geom\_bar()**
      - Don’t forget “**position= “stack”**”!

```
167 # plot
168 ggplot(data=arv_2020, # specify the data
169   aes(x=reorder(CountryName, value), # x-axis,
170         y=value, # specify the column that you
171         fill = gender)) + # specify the subgroup
172 # specify the type of your graph
173 geom_bar(
174   stat = "identity", # when you specify x and y in
175   position = "stack" # when you make a stacked bar
176 ) +
177 # add title
178 ggtitle("Percentage of individuals with access to
179 # add x-axis label
180 xlab("Country") +
181 # add y-axis label
182 ylab("Percentage (%)") +
183 # customize the background theme
184 theme_bw() +
185 coord_flip() # to switch x and y
```

## 2.1. Stacked bar chart

- 2.1. Stacked bar chart



# Table of contents

---

## [Overview]

- R vs Excel vs STATA
- Figure Types and use cases

## [R]

- Basic operation
- 1. Histogram
- 2. Bar plot
  - 2.1. Stacked bar plot
- 3. Line graph**
- 4. Scatter plot
- 5. Tornado plot
- 6. Heatmap

### 3. Line chart

- 3. Line chart
  - Aim: To visualize the trend of the percentages of females and males who have access to ARV in Myanmar from 2001 to 2020
  - Subset data with “Myanmar” in “CountryName” column

```
192 # select data of Myanmar from 2001 to 2020  
193 arp_myanmar <- arp[arp$CountryName=="Myanmar", ]
```

- (Check data)

Some ideas for playing around:

- How about other countries?

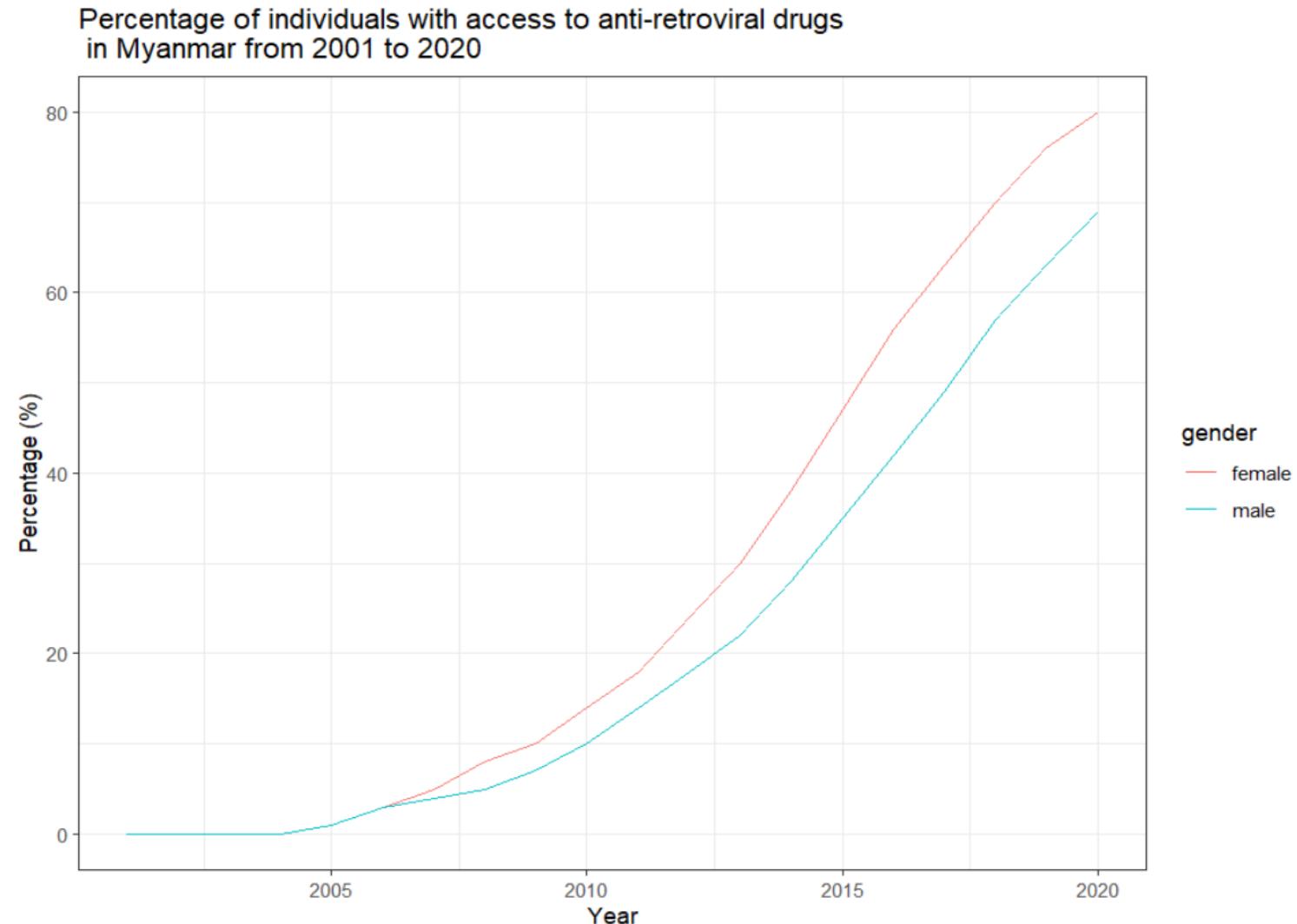
### 3. Line chart

- 3. Line chart
  - Plot the data: **ggplot()**!
    - Specify the data (“arv\_myanmar”) and which variable is used on the x-axis (“year”), the y-axis (“value”), how to group the lines (“gender”), and the color of the lines (“gender”)
    - Specify the type of graph: **geom\_line()**

```
198 # plot
199 ggplot(data=arv_myanmar, # specify data
200       aes(x=year, # specify the column that you
201             y=value, # # specify the column that
202             group = gender, # different lines b
203             color = gender)) + # different color
204   # specify the type of your graph
205   geom_line() +
206   # add title
207   ggtitle("Percentage of individuals with access
208   # add x-axis label
209   xlab("Year") +
210   # add y-axis label
211   ylab("Percentage (%)") +
212   # customize the background theme
213   theme_bw())
```

### 3. Line chart

- 3. Line chart



# Table of contents

---

## [Overview]

- R vs Excel vs STATA
- Figure Types and use cases

## [R]

- Basic operation
- 1. Histogram
- 2. Bar plot
  - 2.1. Stacked bar plot
- 3. Line graph
- 4. Scatter plot**
- 5. Tornado plot
- 6. Heatmap

# 4. Scatter plot

- 4. Scatter plot
  - Aim: To visualize the relationship between the percentages of females and males who have access to ARV in 2020
  - Change the data structure of “arv\_2020”

n	SeriesName	gender	unit	SeriesCode	CountryName	CountryCode	year	value
19	4827 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Afghanistan	AFG	2020	11
20	4828 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Afghanistan	AFG	2020	11
21	4835 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Argentina	ARG	2020	70
22	4836 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Argentina	ARG	2020	69
23	4837 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Australia	AUS	2020	83
24	4838 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Australia	AUS	2020	82
25	4851 Access to anti-retroviral drugs	female	%	SH.HIV.ARVC.FE.ZS	Benin	BEN	2020	91
26	4852 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Benin	BEN	2020	62
27	4855 Access to anti-retroviral drugs	male	%	SH.HIV.ARVC.MA.ZS	Bolivia	BOL	2020	57

## 4. Scatter plot

- 4. Scatter plot
  - Change the data structure of “arv\_2020” to “arv\_2020\_spread” with different columns with each gender

```
220 # change the shape of the data table to put the
221 arv_2020_spread <- arv_2020 %>%
222   select(-c(SeriesCode,n)) %>%
223   spread(key = gender, value = value)
```

The screenshot shows a data visualization interface with a scatter plot. The x-axis is labeled "female" and the y-axis is labeled "male". The plot area contains several data points, each representing a country. A legend indicates that blue dots represent females and orange dots represent males. The data points are roughly clustered into two groups: one group where both genders have high access (blue points near the top right) and another group where both genders have low access (orange points near the bottom left). A horizontal dashed line is drawn across the plot at approximately y=50.

SeriesName	unit	CountryName	CountryCode	year	female	male
Access to anti-retroviral drugs	%	Afghanistan	AFG	2020	11	11
Access to anti-retroviral drugs	%	Argentina	ARG	2020	70	69
Access to anti-retroviral drugs	%	Australia	AUS	2020	82	83
Access to anti-retroviral drugs	%	Benin	BEN	2020	91	62
Access to anti-retroviral drugs	%	Bolivia	BOL	2020	43	57
Access to anti-retroviral drugs	%	Botswana	BWA	2020	94	84
Access to anti-retroviral drugs	%	Burkina Faso	BFA	2020	93	63

# 4. Scatter plot

- 4. Scatter plot
  - Plot the data: **ggplot()**!
    - Specify the data (“arv\_2020\_spread”) and which variable is used on the x-axis (“female”), and the y-axis (“male”)
    - Specify the type of graph: **geom\_point()**

Some ideas for playing around:

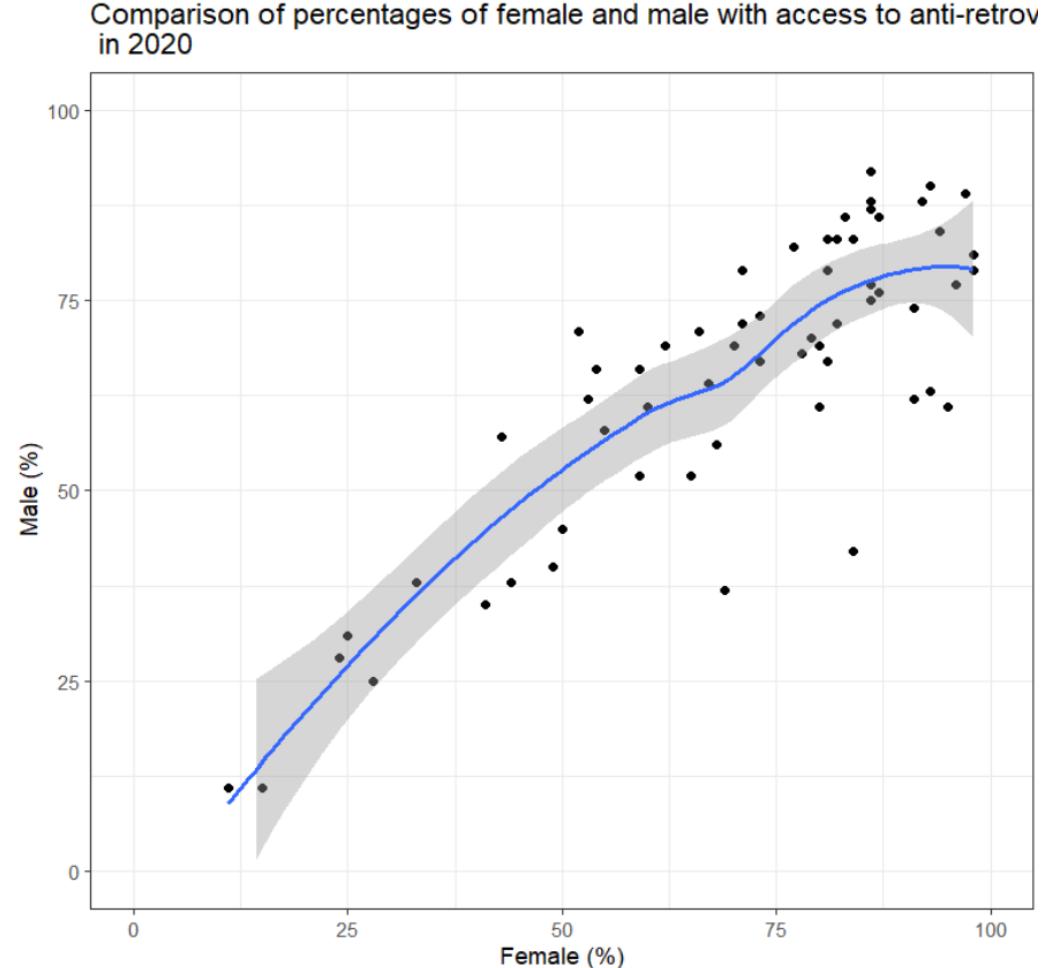
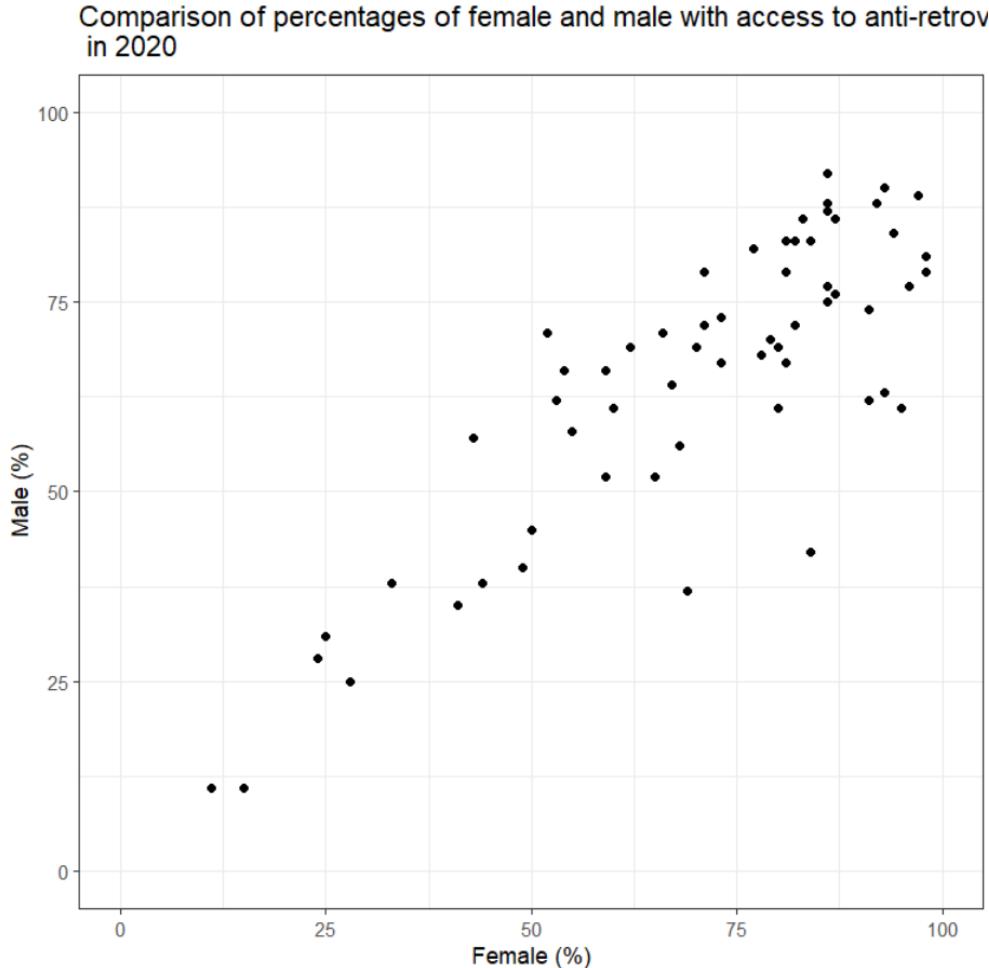
- How about flipping the x-axis and y-axis?

```
228 # plot
229 ggplot(data=arv_2020_spread, # specify data
230     aes(x=female, # specify the column
231           y=male)) + # specify the columns
232     # specify the type of your graph
233     geom_point() +
234     # make sure x-axis extends to 100%
235     xlim(0,100) +
236     # make sure y-axis extends to 100%
237     ylim(0,100) +
238     # add title
239     ggtitle("Comparison of percentages of")
240     # add x-axis label
241     xlab("Female (%)") +
242     # add y-axis label
243     ylab("Male (%)") +
244     # customize the background theme
245     theme_bw()
```

# 4. Scatter plot

- 4. Scatter plot

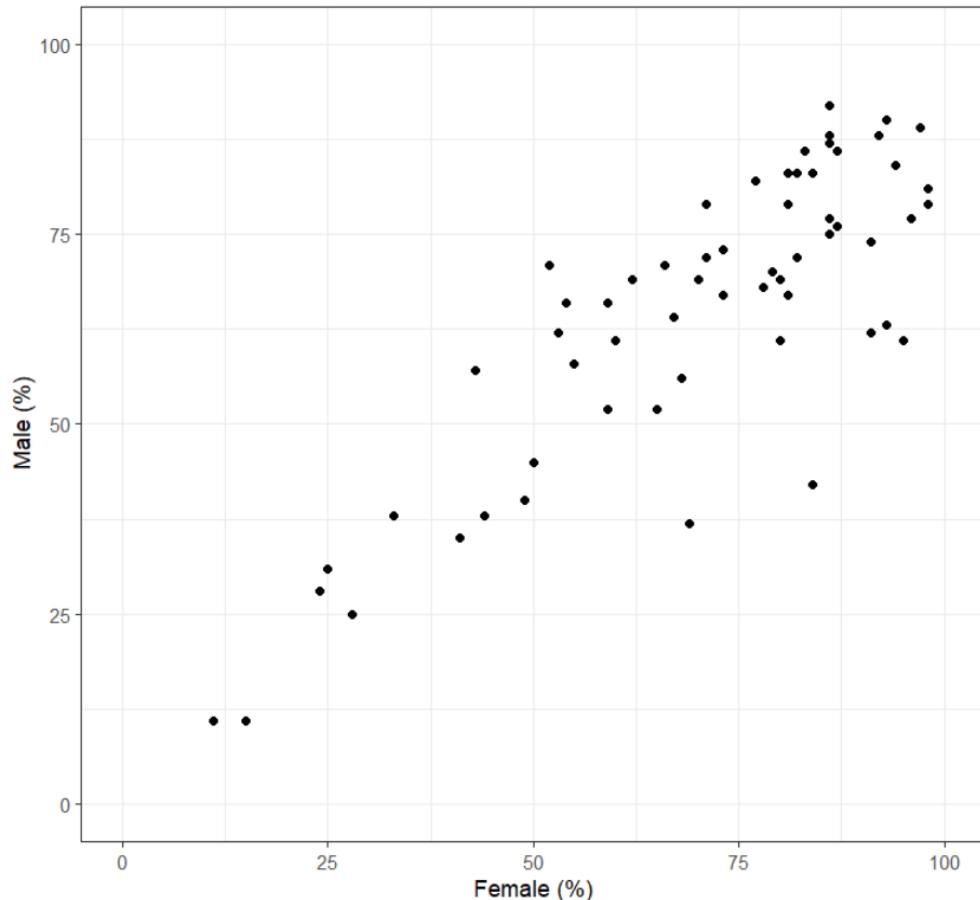
```
253 # add the linear trend and the confidence interval  
254 geom_smooth() +
```



# 4. Scatter plot

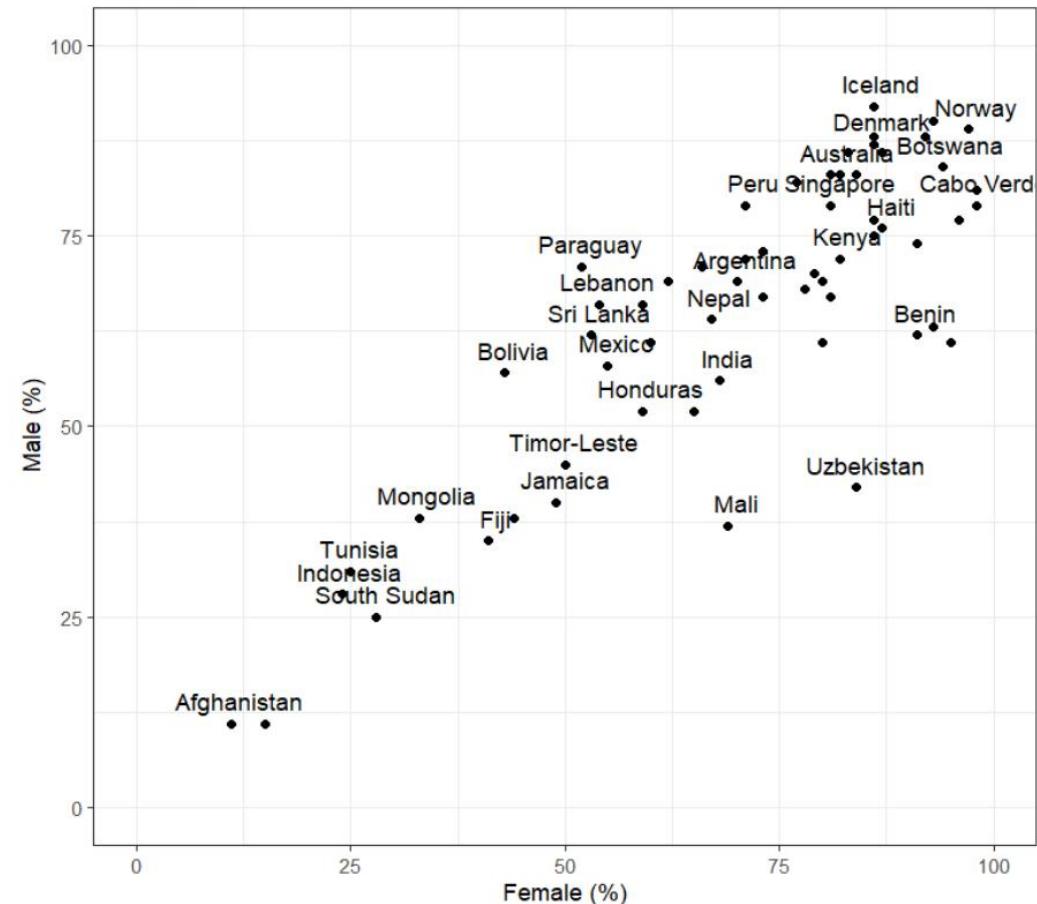
- 4. Scatter plot

Comparison of percentages of female and male with access to anti-retrov in 2020



```
274 # add country names
275 geom_text(
276   label = arv_2020_spread$CountryName, # specify texts
277   nudge_x = 1, nudge_y = 3,    # shift the texts along x and y axis
278   check_overlap = TRUE      # avoid overlap
279 ) +
```

Comparison of percentages of female and male with access to anti-retrov in 2020



# Table of contents

---

## [Overview]

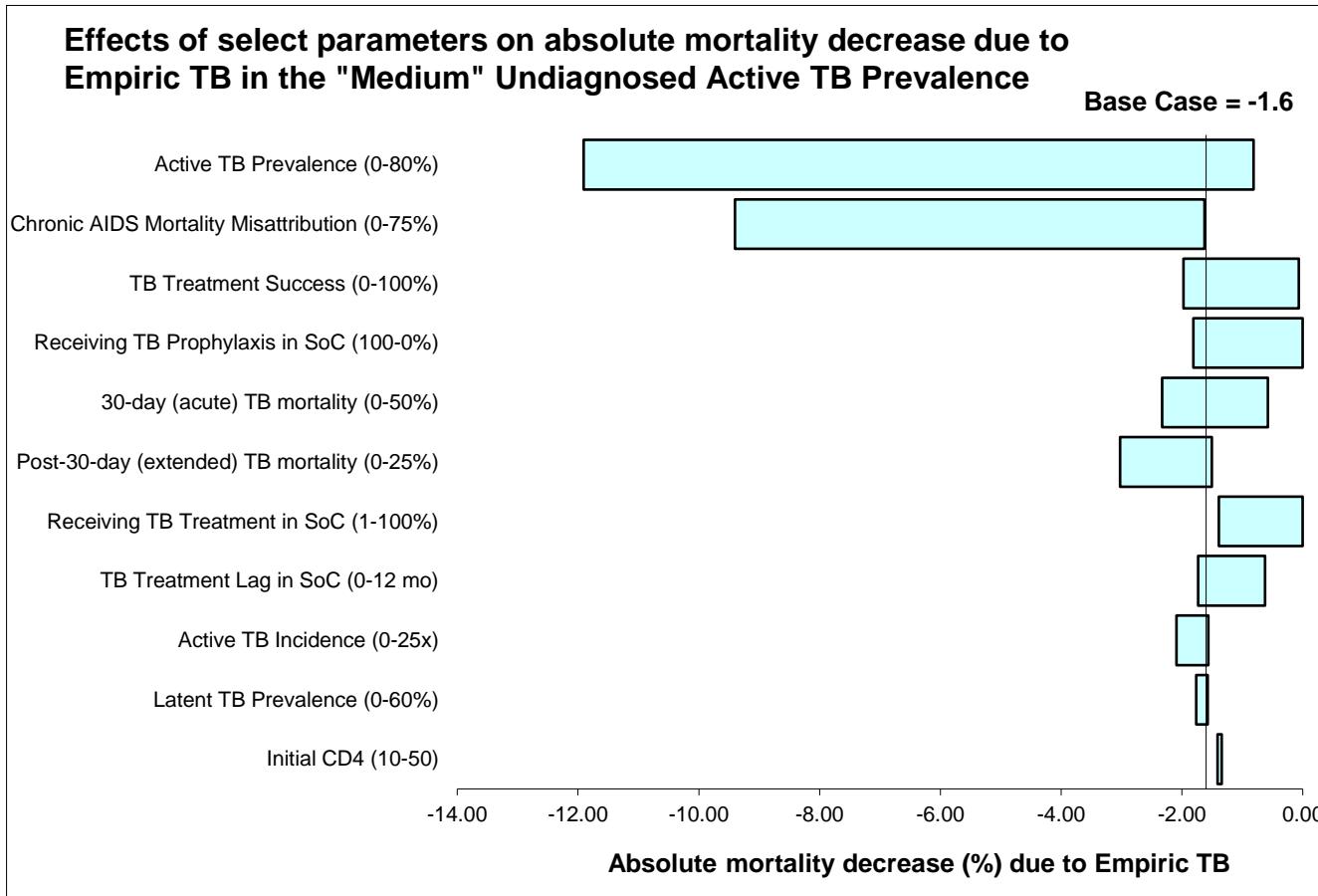
- R vs Excel vs STATA
- Figure Types and use cases

## [R]

- Basic operation
- 1. Histogram
- 2. Bar plot
  - 2.1. Stacked bar plot
- 3. Line graph
- 4. Scatter plot
- 5. Tornado plot**
- 6. Heatmap

## 5. Tornado plot

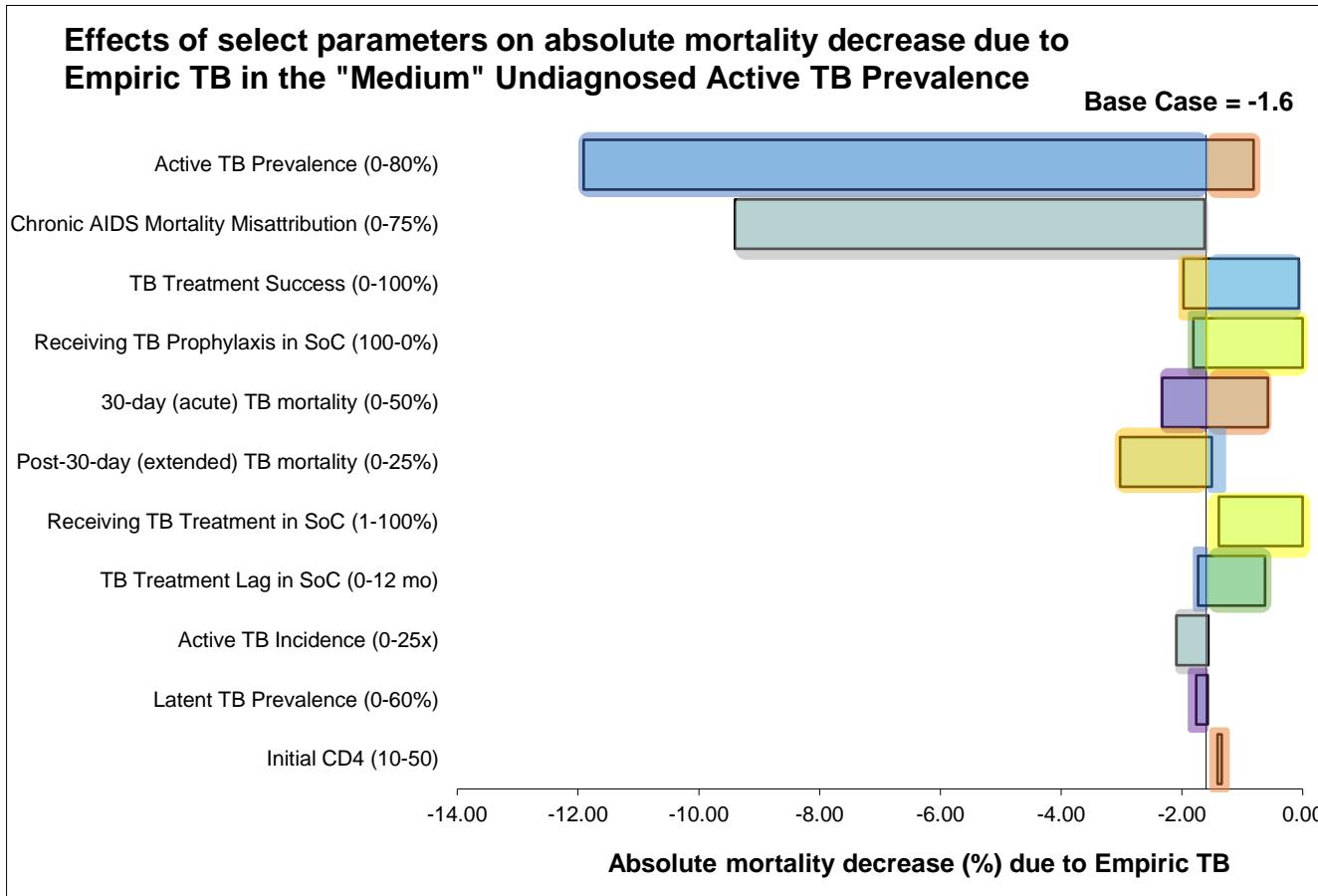
- Aim: To visualize the effect ranges on outputs when changing one input parameter



- Looks like a bar chart, but...

## 5. Tornado plot

- Aim: To visualize the effect ranges on outputs when changing one input parameter



- Looks like a bar chart, but...
- What we're doing is **to put rectangles!**
- Game plan:
  - Import result data of min and max
  - Find differences (will be side lengths)
  - Identify the positions of each apex of rectangles

## 5. Tornado plot

- Needed data:
  - Input parameters that varied in your one-way sensitivity analysis
  - Maximum output values by the input parameter
  - Minimum output values by the input parameter
  - (Base case value)

	A	B	C
1	parameter	max	min
2	Active TB Prevalence (0-80%)	-11.9092	-0.814
3	Chronic AIDS Mortality Misattribution (0-75%)	-9.4012	-1.6255
4	TB Treatment Success (0-100%)	-1.9762	-0.0671
5	Receiving TB Prophylaxis in SoC (100-0%)	-1.7739	0.0352
6	30-day (acute) TB mortality (0-50%)	-2.3272	-0.5761
7	Post-30-day (extended) TB mortality (0-25%)	-3.0225	-1.504
8	Receiving TB Treatment in SoC (1-100%)	-1.3884	0
9	TB Treatment Lag in SoC (0-12 mo)	-1.7323	-0.6258
10	Active TB Incidence (0-25x)	-2.092	-1.5625
11	Latent TB Prevalence (0-60%)	-1.7641	-1.5715
12	Initial CD4 (10-50)	-1.4125	-1.3429

## 5. Tornado plot

- Calculate the differences between the maximum and minimum values

```
303 # add differences between max and min  
304 tornado <- tornado %>%  
305     # find difference between maximum and minimum  
306     mutate(dif = max - min)
```

parameter	max	min	dif
1 Active TB Prevalence (0-80%)	-11.9092	-0.8140	-11.0952
2 Chronic AIDS Mortality Misattribution (0-75%)	-9.4012	-1.6255	-7.7757
3 TB Treatment Success (0-100%)	-1.9762	-0.0671	-1.9091
4 Receiving TB Prophylaxis in SoC (100-0%)	-1.7739	0.0352	-1.8091

- Identify the order of parameters depending on the absolute values of differences

```
308 # get the order of the parameters depending on the absolute values  
309 param_order <- tornado %>%  
310     # set the order of the parameters depending on the absolute values  
311     arrange(desc(dif)) %>%  
312     mutate(parameter = factor(x = parameter, levels = parameter)) %>%  
313     select(parameter) %>%  
314     # change the data type from list to vector  
315     unlist() %>%  
316     levels()  
317
```

Most Influential



Least Influential

# 5. Tornado plot

- Identify the positions of each apex of each rectangle

```
323 # read and modify data  
324 tornado_plot <- tornado %>%  
325   # change the structure of the datatable  
326   gather(key = "min_max", value = "value", max, min) %>%
```

Separate one parameter into two boxes

parameter	max	min
Active TB Prevalence (0-80%)	-11.9092	-0.8140
Chronic AIDS Mortality Misattribution (0-75%)	-9.4012	-1.6255
TB Treatment Success (0-100%)	-1.9762	-0.0671
Receiving TB Prophylaxis in SoC (100-0%)	-1.7739	0.0352
30-day (acute) TB mortality (0-50%)	-2.3272	-0.5761
Post-30-day (extended) TB mortality (0-25%)	-3.0225	-1.504
Receiving TB Treatment in SoC (1-100%)	-1.3884	0
TB Treatment Lag in SoC (0-12 mo)	-1.7323	-0.6258
Active TB Incidence (0-25x)	-2.092	-1.5625
Latent TB Prevalence (0-60%)	-1.7641	-1.5715
Initial CD4 (10-50)	-1.4125	-1.3429

parameter	dif	min_max	value	ymin	ymax	xmin	xmax
Active TB Prevalence (0-80%)	-11.0952	max	-11.9092	-11.9092	-1.6000	10.525	11.475
Chronic AIDS Mortality Misattribution (0-75%)	-7.7757	max	-9.4012	-9.4012	-1.6000	9.525	10.475
TB Treatment Success (0-100%)	-1.9091	max	-1.9762	-1.9762	-1.6000	8.525	9.475
Receiving TB Prophylaxis in SoC (100-0%)	-1.8091	max	-1.7739	-1.7739	-1.6000	7.525	8.475
30-day (acute) TB mortality (0-50%)	-1.7511	max	-2.3272	-2.3272	-1.6000	6.525	7.475
Post-30-day (extended) TB mortality (0-25%)	-1.5185	max	-3.0225	-3.0225	-1.6000	5.525	6.475
Receiving TB Treatment in SoC (1-100%)	-1.3884	max	-1.3884	-1.6000	-1.3884	4.525	5.475
TB Treatment Lag in SoC (0-12 mo)	-1.1065	max	-1.7323	-1.7323	-1.6000	3.525	4.475
Active TB Incidence (0-25x)	-0.5295	max	-2.0920	-2.0920	-1.6000	2.525	3.475
Latent TB Prevalence (0-60%)	-0.1926	max	-1.7641	-1.7641	-1.6000	1.525	2.475
Initial CD4 (10-50)	-0.0696	max	-1.4125	-1.6000	-1.4125	0.525	1.475
Active TB Prevalence (0-80%)	-11.0952	min	-0.8140	-1.6000	-0.8140	10.525	11.475
Chronic AIDS Mortality Misattribution (0-75%)	-7.7757	min	-1.6255	-1.6255	-1.6000	9.525	10.475

# 5. Tornado plot

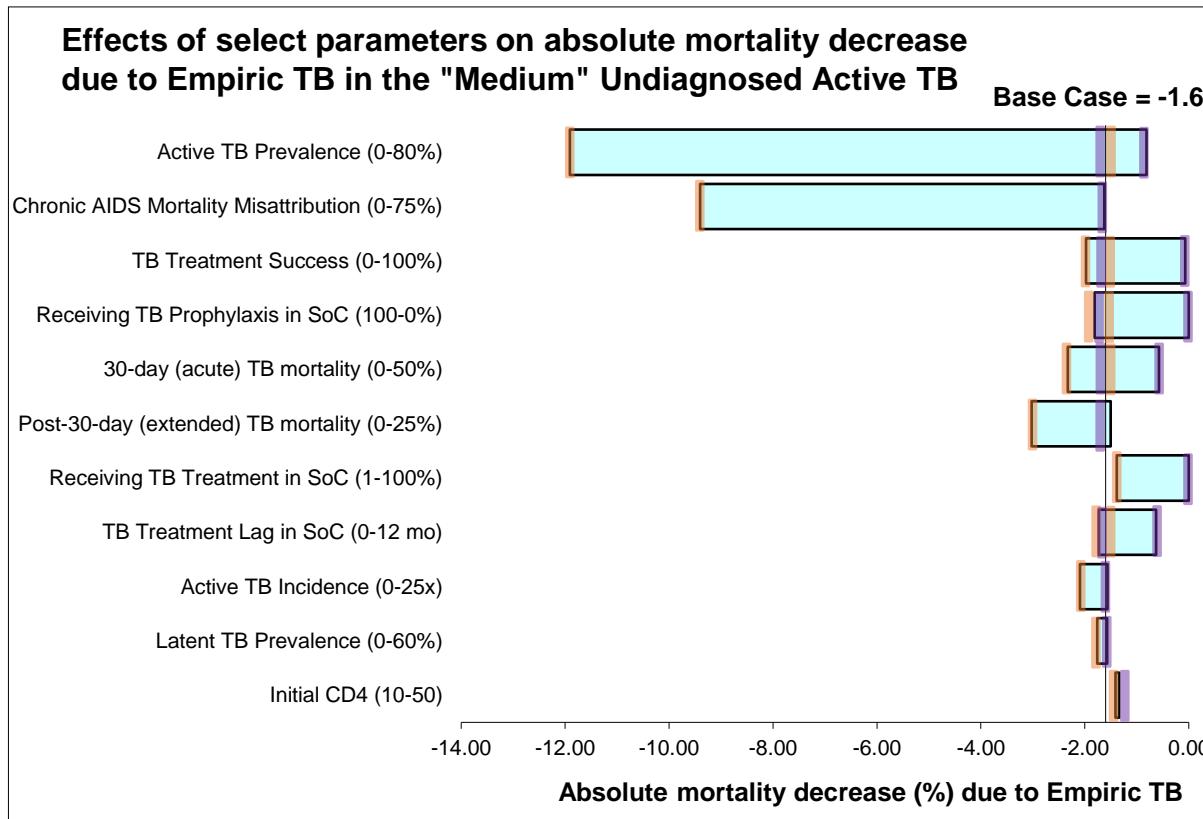
- Identify the positions of each apex of each rectangle

```
27 # calculate the positions of apexes of rectangles
28 mutate(parameter=factor(parameter,levels = param_order),
29   ymin=pmin(value, basecasevalue),
30   ymax=pmax(value, basecasevalue),
31   xmin=as.numeric(parameter)-width/2,
32   xmax=as.numeric(parameter)+width/2)
```

Keep parameter order

Get position of rectangle vertical sides  
depending on base case values.

Get position of rectangle horizontal sides.



parameter	dif	min_max	value	ymin	ymax	xmin	xmax
1 Active TB Prevalence (0-80%)	-11.0952	max	-11.9092	-11.9092	-1.6000	10.525	11.475
2 Chronic AIDS Mortality Misattribution (0-75%)	-7.7757	max	-9.4012	-9.4012	-1.6000	9.525	10.475
3 TB Treatment Success (0-100%)	-1.9091	max	-1.9762	-1.9762	-1.6000	8.525	9.475
4 Receiving TB Prophylaxis in SoC (100-0%)	-1.8091	max	-1.7739	-1.7739	-1.6000	7.525	8.475
5 30-day (acute) TB mortality (0-50%)	-1.7511	max	-2.3272	-2.3272	-1.6000	6.525	7.475
6 Post-30-day (extended) TB mortality (0-25%)	-1.5185	max	-3.0225	-3.0225	-1.6000	5.525	6.475
7 Receiving TB Treatment in SoC (1-100%)	-1.3884	max	-1.3884	-1.6000	-1.3884	4.525	5.475
8 TB Treatment Lag in SoC (0-12 mo)	-1.1065	max	-1.7323	-1.7323	-1.6000	3.525	4.475
9 Active TB Incidence (0-25x)	-0.5295	max	-2.0920	-2.0920	-1.6000	2.525	3.475
10 Latent TB Prevalence (0-60%)	-0.1926	max	-1.7641	-1.7641	-1.6000	1.525	2.475
11 Initial CD4 (10-50)	-0.0696	max	-1.4125	-1.6000	-1.4125	0.525	1.475
12 Active TB Prevalence (0-80%)	-11.0952	min	-0.8140	-1.6000	-0.8140	10.525	11.475
13 Chronic AIDS Mortality Misattribution (0-75%)	-7.7757	min	-1.6255	-1.6255	-1.6000	9.525	10.475

## 5. Tornado plot

- Plot the data: **ggplot()**!
  - Specify the data (“tornado\_plot”) and which variable is used for the maximum and minimum values on the x-axis and the y-axis (“xmax”, “xmin”, “ymax”, and “ymin”)
  - Specify the type of graph: **geom\_rect()**
  - Specify the color to fill the rectangles and the borders of the rectangles: **scale\_fill\_manual()** and **scale\_color\_manual()**

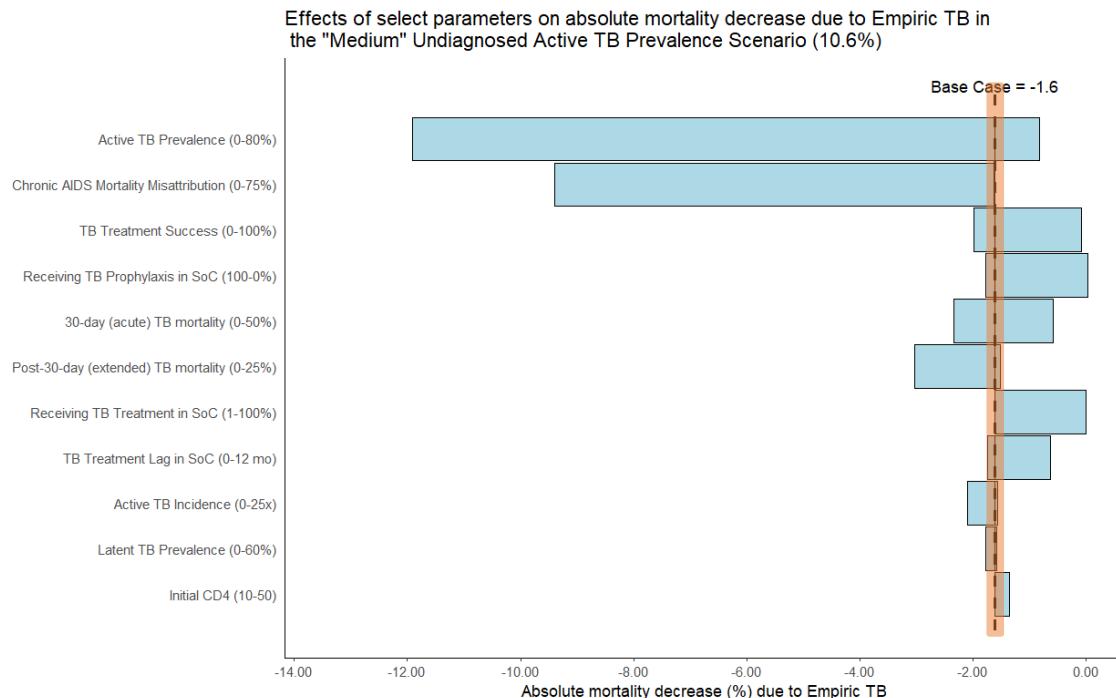
```
338 # plot
339 ggplot(data = tornado_plot,
340         aes(ymax=ymax, # set each apex
341             ymin=ymin,
342             xmax=xmax,
343             xmin=xmin)) +
344     # specify the type of your graph and fill and border color
345     geom_rect(aes(fill="", color="")) +
346     scale_fill_manual(values = "lightblue") + # color inside the rect
347     scale_color_manual(values = "black") + # color of the rect borders
```

# 5. Tornado plot

- Add a vertical line to show the base case result with an annotation:

**geom\_segment()**

```
367 # add vertical line to show the baseline result
368 geom_segment(aes(x=max(xmax) + 0.5,
369                 xend=0.2,
370                 y=basecasevalue,
371                 yend=basecasevalue),
372                 linetype = 2, # dashed line
373                 linewidth = 1) +
374 # add annotation of the vertical line
375 annotate("text", # set annotation type
376           x = max(tornado_plot$xmax) + 0.7, y = basecasevalue, # set
377           label = paste0("Base Case = ", basecasevalue)) +
```



# Table of contents

---

## [Overview]

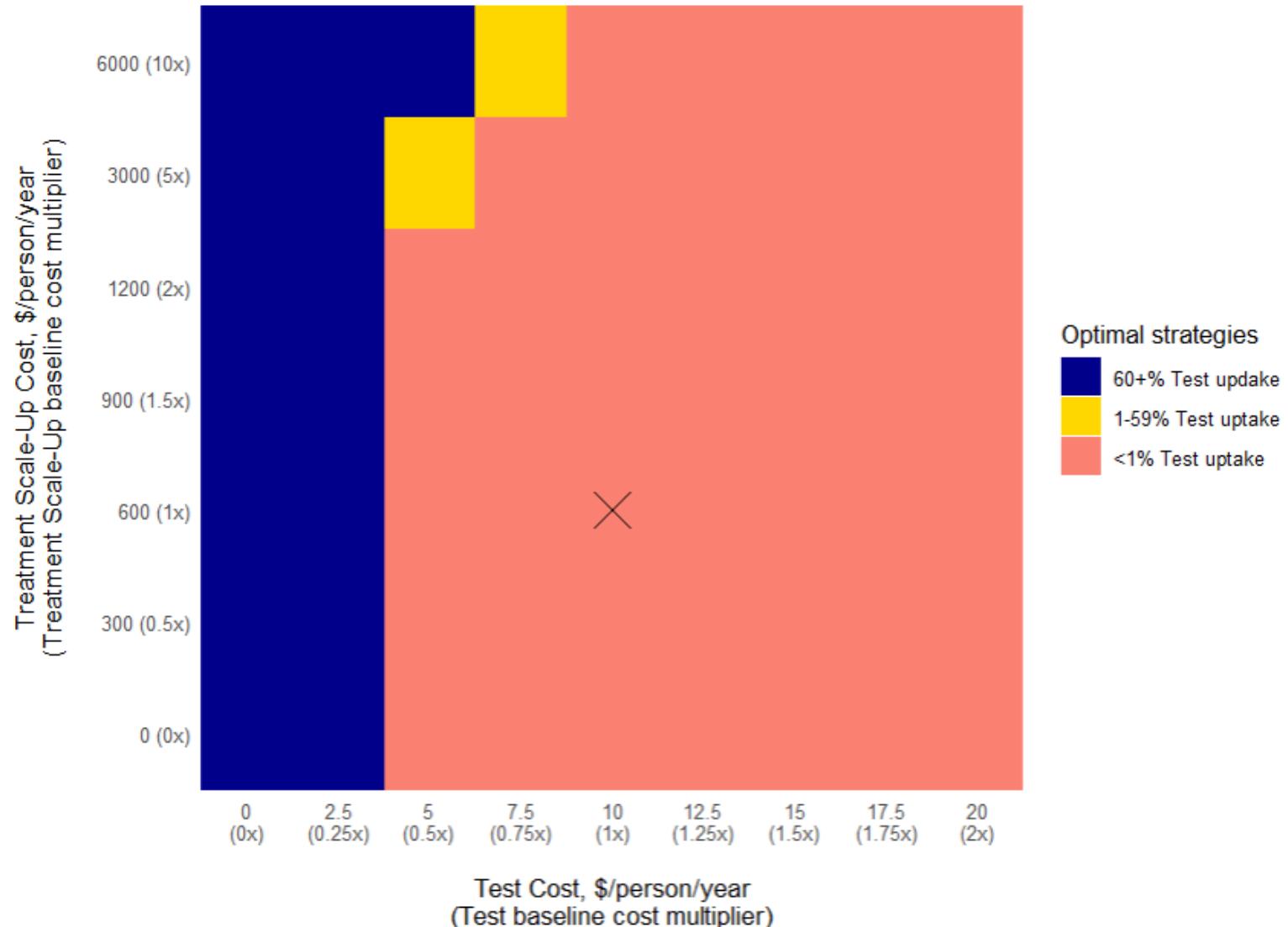
- R vs Excel vs STATA
- Figure Types and use cases

## [R]

- Basic operation
- 1. Histogram
- 2. Bar plot
  - 2.1. Stacked bar plot
- 3. Line graph
- 4. Scatter plot
- 5. Tornado plot
- 6. Heatmap**

# 6. Heatmap

- Aim: To visualize the most cost-effective strategies depending on variation of two conditions

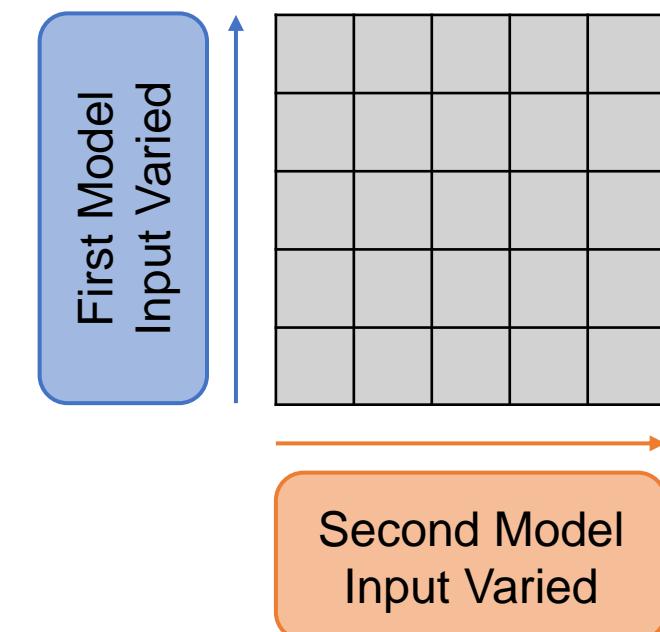


# 6. Heatmap

- Needed data:
  - Optimal strategies identified by each permutation of the variations of two conditions
  - (Base case conditions and result)

	Tx_cost_times	Test_cost_times	Optimal_strategy
1	0	0	60+% Test uptake
2	0	0.25	60+% Test uptake
3	0	0.5	<1% Test uptake
4	0	0.75	<1% Test uptake
5	0	1	<1% Test uptake
6	0	1.25	<1% Test uptake
7	0	1.5	<1% Test uptake
8	0	1.75	<1% Test uptake
9	0	2	<1% Test uptake
10	0.5	0	60+% Test uptake

```
27 heat <- read.csv("2.heatmap_data.csv")
```



## 6. Heatmap

- Modifying data for plotting (cont'):

- Use factor levels to plot Test and Treatment scale up cost in ascending order
- Create x and y axis labels

```
393 # create "levels" for the x- and y-axis to show the result in ascending order
394 heat <- heat %>%
395   mutate(Tx_cost_times = factor(Tx_cost_times, levels = unique(heat$Tx_cost_times)),
396         Test_cost_times = factor(Test_cost_times, levels = unique(heat$Test_cost_times)
397       )))
398 # create x axis labels
399 heat$x_labels = paste0(
400   as.numeric(as.character(heat$Test_cost_times))*Test_cost_base,
401   "\n(",heat$Test_cost_times, "x)")
402
403 # create y axis labels
404 heat$y_labels = paste0(
405   as.numeric(as.character(heat$Tx_cost_times))*Tx_cost_base,
406   " (",heat$Tx_cost_times, "x)")
```

# 6. Heatmap

	Tx_cost_times	Test_cost_times	Optimal_strategy	x_labels	y_labels
1	0	0	60+% Test update	0 (0x)	0 (0x)
2	0	0.25	60+% Test update	2.5 (0.25x)	0 (0x)
3	0	0.5	<1% Test uptake	5 (0.5x)	0 (0x)
4	0	0.75	<1% Test uptake	7.5 (0.75x)	0 (0x)
5	0	1	<1% Test uptake	10 (1x)	0 (0x)
6	0	1.25	<1% Test uptake	12.5 (1.25x)	0 (0x)
7	0	1.5	<1% Test uptake	15 (1.5x)	0 (0x)
8	0	1.75	<1% Test uptake	17.5 (1.75x)	0 (0x)
9	0	2	<1% Test uptake	20 (2x)	0 (0x)
10	0.5	0	60+% Test update	0 (0x)	300 (0.5x)
11	0.5	0.25	60+% Test update	2.5 (0.25x)	300 (0.5x)
12	0.5	0.5	<1% Test uptake	5 (0.5x)	300 (0.5x)
13	0.5	0.75	<1% Test uptake	7.5 (0.75x)	300 (0.5x)
14	0.5	1	<1% Test uptake	10 (1x)	300 (0.5x)
15	0.5	1.25	<1% Test uptake	12.5 (1.25x)	300 (0.5x)
16	0.5	1.5	<1% Test uptake	15 (1.5x)	300 (0.5x)

# 6. Heatmap

Specify the data (“heat”) and which variable is used on the x and y axis.

Create heatmap

Draw “X” as base case

Plot title and axes labels

Specify color for each optimal strategy

Format figure

```
413 # plot
414 ggplot(data = heat, # specify the data
415     aes(x=Test_cost_times, # specify the x-axis
416         y=Tx_cost_times, # specify the y-axis
417         fill = Optimal_strategy)) + # specify the color showing the optimal strategy
418     geom_tile() +
419     # add "X" at the base case
420     geom_point(
421         data = heat[(heat$Tx_cost_times == 1) & (heat$Test_cost_times == 1),],
422         aes(x = Test_cost_times, y = Tx_cost_times), fill = "NA",
423         color = "black", size = 7, shape = 4
424     ) +
425     theme_classic() +
426     # add plot title
427     ggtitle("Test cost and Treatment scale-up cost") +
428     # x axis label
429     xlab("Test Cost, $/person/year\n(Test baseline cost multiplier)") +
430     # y axis label
431     ylab("Treatment Scale-Up Cost, $/person/year\n(Treatment Scale-Up baseline cost
multiplier)") +
432     # set strategy colors, reverse legend order, name legend "Optimal strategies"
433     scale_fill_manual(values = c("salmon", "gold", "darkblue"),
434                         guide = guide_legend(reverse = TRUE),
435                         name = "Optimal strategies") +
436     # adjust distance between x axis labels and plot, add x axis tick labels
437     scale_x_discrete(expand = c(0,0),
438                         labels = unique(heat$x_labels)) +
439     # adjust distance between y axis labels and plot, add y axis tick labels
440     scale_y_discrete(expand = c(0,0),
441                         labels = unique(heat$y_labels)) +
442     # remove axis ticks and lines, add margin between axes and axes titles
443     theme(axis.ticks = element_blank(),
444           axis.line = element_blank(),
445           axis.title.y = element_text(margin = unit(c(0,5,0,0), "mm")),
446           axis.title.x = element_text(margin = unit(c(5,0,0,0), "mm")))
```

# Resources

---

- [Data Visualization Cheatsheet](#)
- [Cheatsheet Updates](#)
- [R Color Cheatsheet](#)
- [ggplot2 Documentation](#)
- [Coloring for Colorblindness \(davidmathlogic.com\)](#)