Applied Statistics for Data Scientists with R

About Me

Md. Ahsanul Islam

Analysis Executive at Kantar Market Research

M.Sc. & B.Sc. in Statistics at University of Chittagong

Ex-executive (Analysis) at Luminaries Research Ltd.

LinkedIn GitHub

The capacity to learn is a gift
The ability to learn is a skill
The willingness to learn is a choice.

- Brian Patrick Herbert

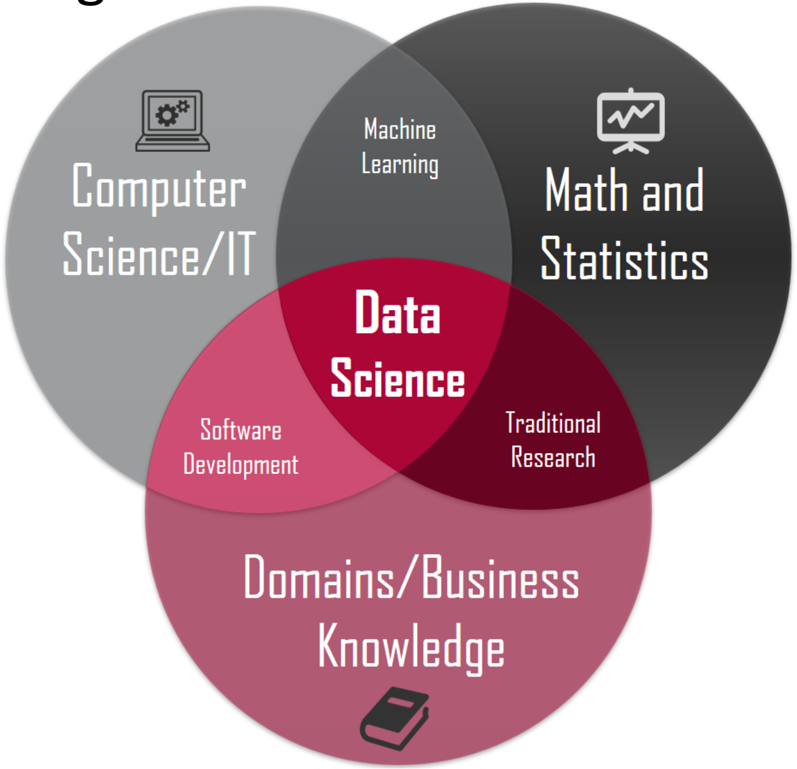
Course Prerequisite

- Common sense
- Basic computer skills
- Programming skill is not required
- Ability to use search engines
- Willingness to learn

Role of Statistics in Data Science

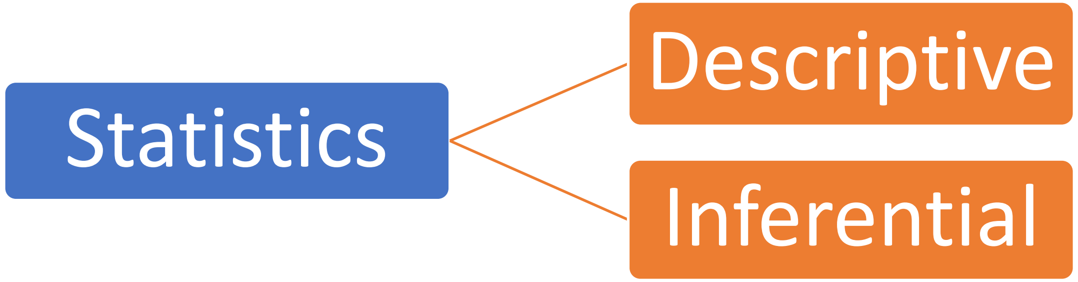
“Statistics is the foundation of predictive modeling, machine learning and AI.”

- Exploratory data analysis (EDA) and data preprocessing.
- Statistical Estimation and Optimization Techniques.
- Sampling Techniques.
- Experimental Design.
- Hypothesis Testing.
- Statistical Modeling.



What is Statistics?

- Statistics is a branch of mathematics that deals with the collection, organization, analysis, interpretation, and presentation of data.
- The focus is in quantitative data to extract meaningful information.



Note: “Statistics” and “Statistic” are not the same thing.

A scenario where statistical knowledge helps

- Suppose you are a senior data analyst at a manufacturing company.
- Your boss notices that Machine A's products have fewer defects than Machine B's, but it's not clear if this difference is real or just random variation.
- Your job is to find out if Machine B truly has a problem.
- You may require to use an appropriate sampling technique and statistical test to test the hypothesis.

Another scenario where statistical knowledge helps

- You are a data analyst for an e-commerce company.
- Your boss wants to predict how much revenue the company will generate during the upcoming Ramadan and what factors might influence the trend.
- This will help in planning marketing campaigns, and inventory management.
- You may require to use an appropriate explainable predictive model.

Statistics vs Applied Statistics

- **Statistics** is about **creating** the "tools,"
  whereas **Applied Statistics** is about **using** those "tools" to address real-world challenges.

Major Tools Used in Applied Statistics

Why Use R instead of Other Programming Languages?

- Specifically designed for statistics by statisticians.
- Thousands of packages for statistical analysis.
- Excellent for data visualization.
- Easy to learn.
- Mostly academic and research focused.

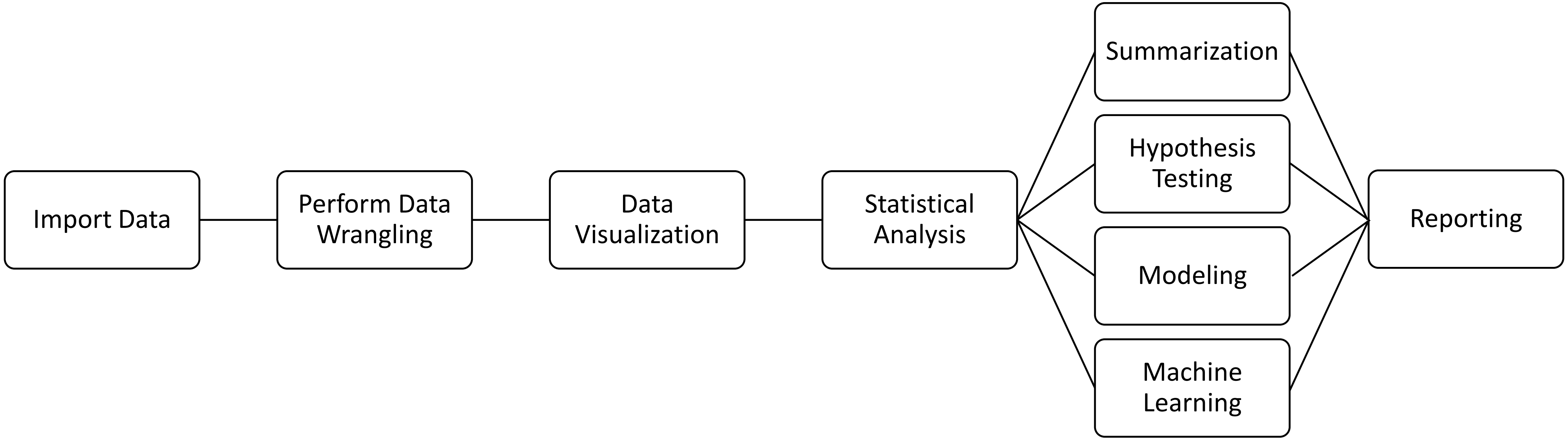
What Will You Learn?

- R programming fundamentals
- Data manipulation and visualization
- Applied statistical techniques
- Building predictive models
- Developing Shiny apps

- Extra: SPSS basics for analysis.

Refer to course outline for details.

Usual Analysis Workflow in R



Import Data

Perform Data Wrangling

Data Visualization

Statistical Analysis

Summarization

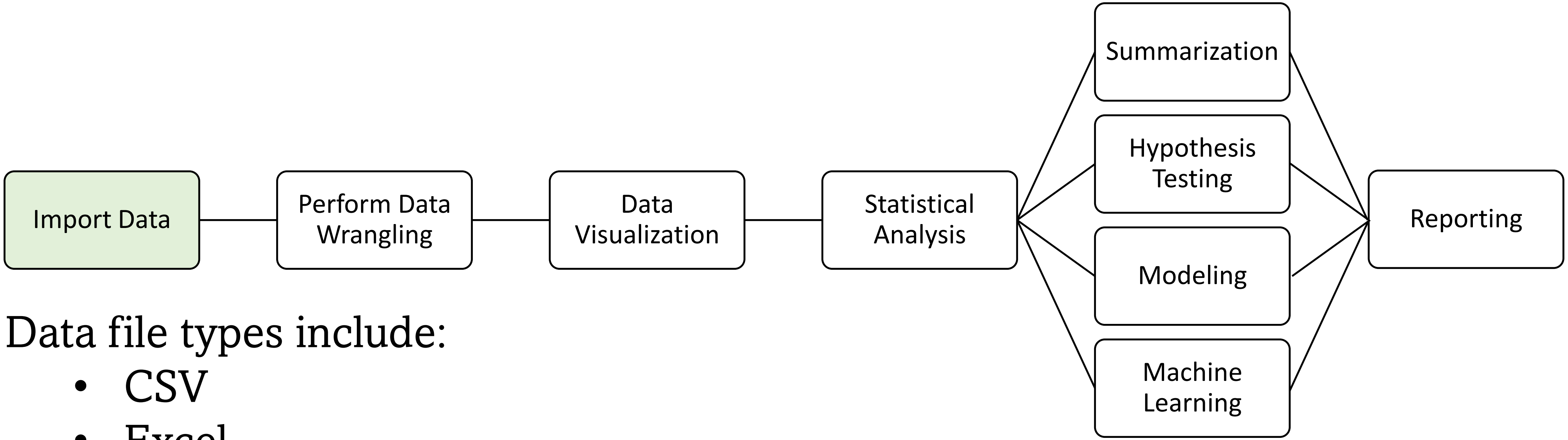
Hypothesis Testing

Modeling

Machine Learning

Reporting

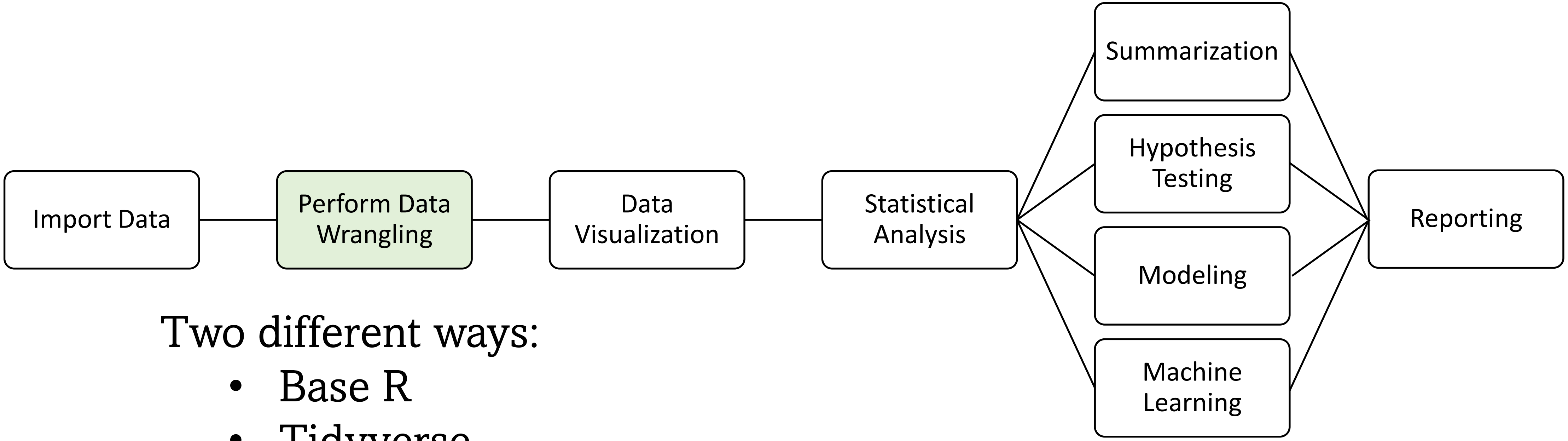
Usual Analysis Workflow in R



Data file types include:

- CSV
- Excel
- SPSS
- STATA
- JSON … …

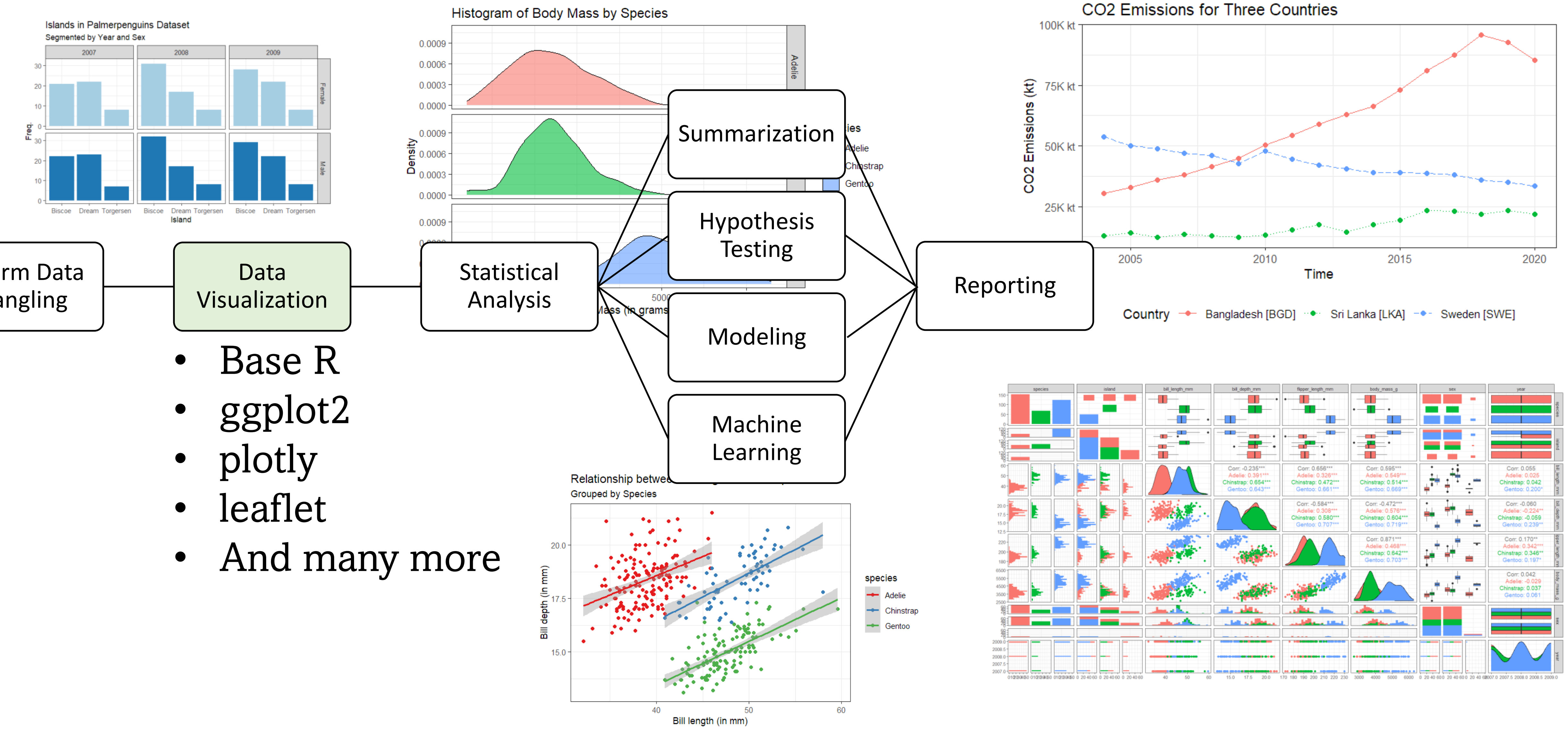
Usual Analysis Workflow in R



Two different ways:

- Base R
- Tidyverse

Usual Analysis Workflow in R



rm Data
angling

Data Visualization

Statistical Analysis

Summarization

Hypothesis Testing

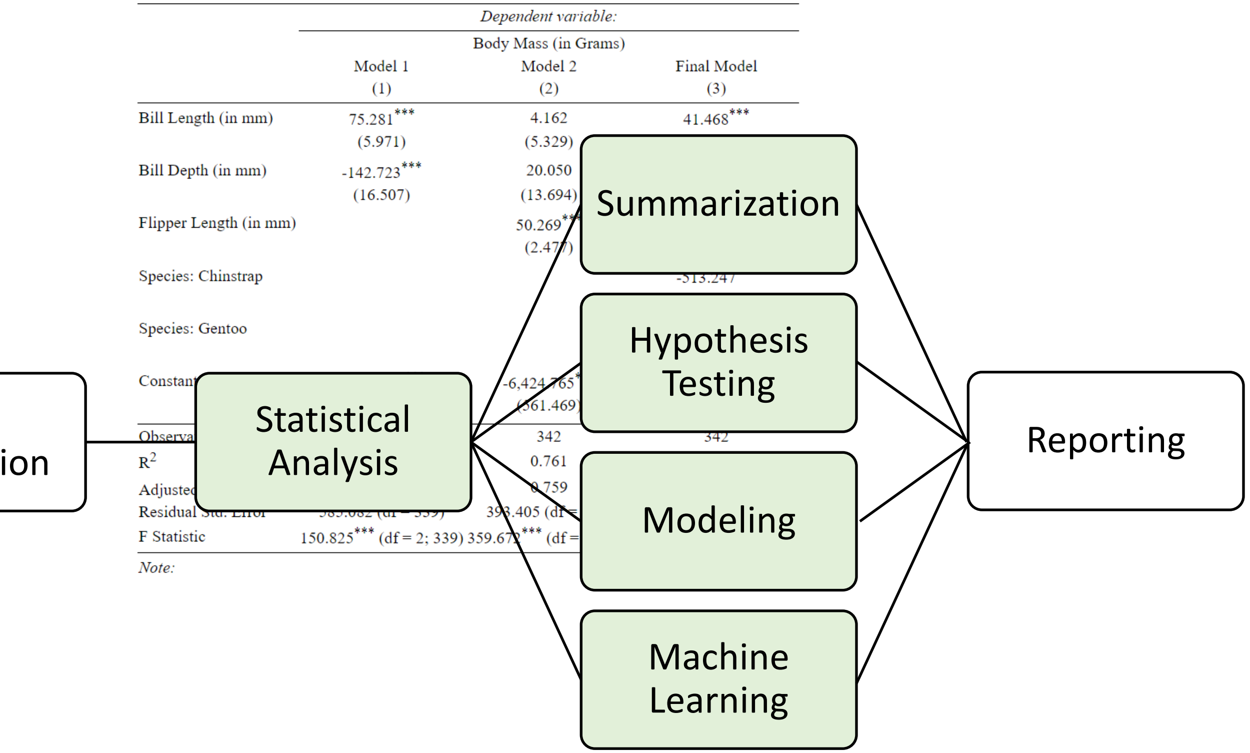
Modeling

Machine Learning

Reporting

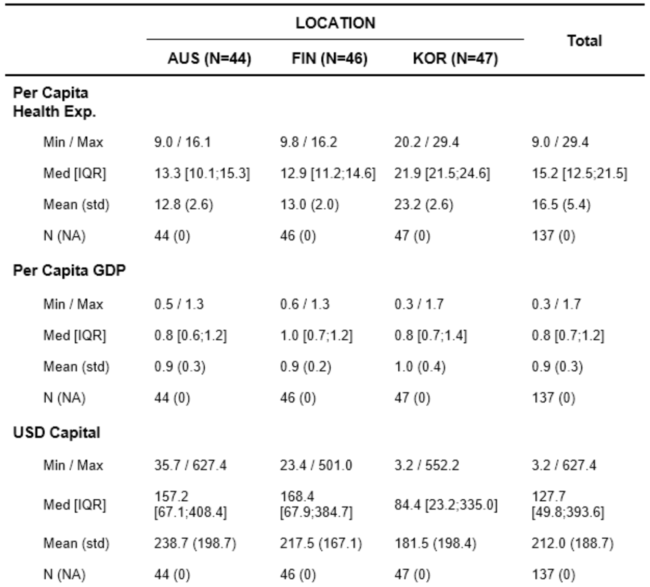
- Base R
- ggplot2
- plotly
- leaflet
- And many more

Usual Analysis Workflow in R

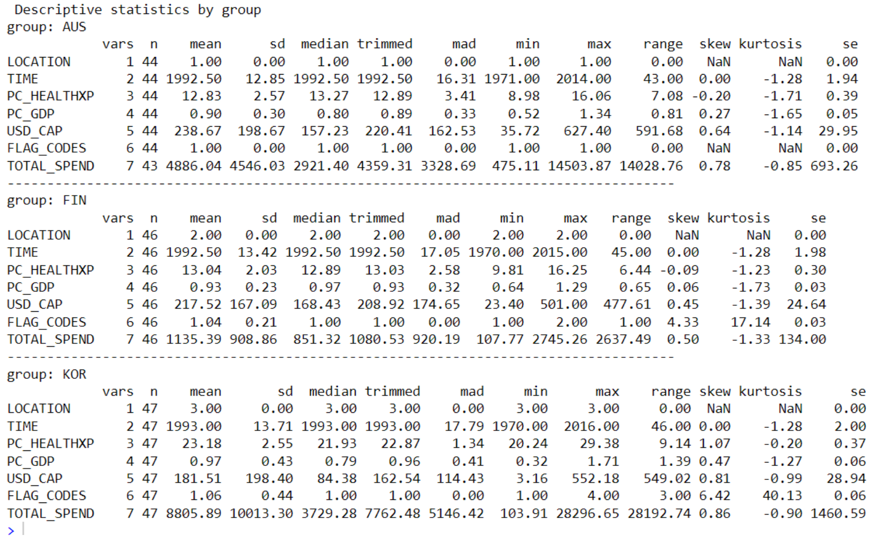


	Dependent variable:		
	Body Mass (in Grams)		
	Model 1	Model 2	Final Model
	(1)	(2)	(3)
Bill Length (in mm)	75.281***	4.162	41.468***
	(5.971)	(5.329)	
Bill Depth (in mm)	-142.723***	20.050	
	(16.507)	(13.694)	
Flipper Length (in mm)		50.269**	
		(2.477)	
Species: Chinstrap			-513.247
Species: Gentoo			
Constant		-6,424.765*	
		(561.469)	
Observa		342	342
R^2		0.761	
Adjusted		0.759	
Residual Std. Error	383.082 (df = 339)	393.405 (df =	
F Statistic	150.825*** (df = 2; 339)	359.672*** (df =	
Note:			

A couldn’t dare to list anything in this slide.

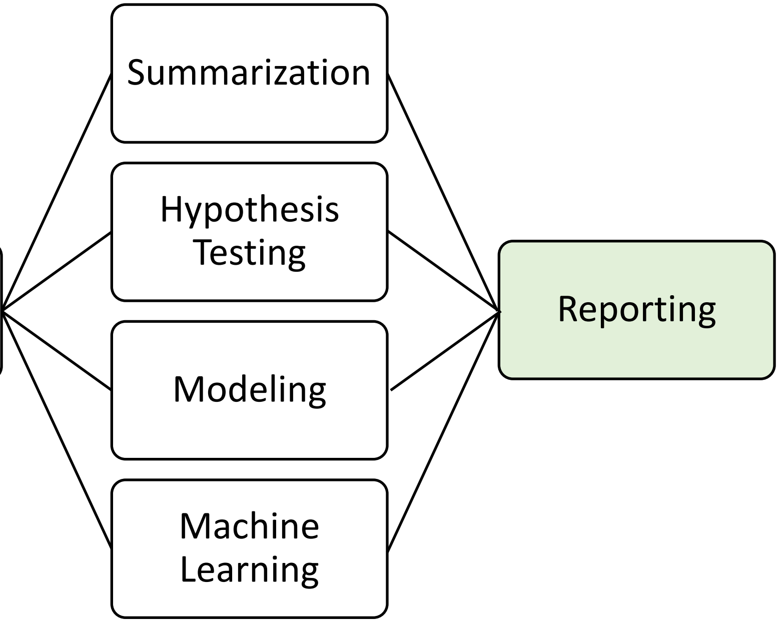


	LOCATION			Total
	AUS (N=44)	FIN (N=46)	KOR (N=47)	
Per Capita Health Exp.				
Min / Max	9.0 / 16.1	9.8 / 16.2	20.2 / 29.4	9.0 / 29.4
Med [IQR]	13.3 [10.1;15.3]	12.9 [11.2;14.6]	21.9 [21.5;24.6]	15.2 [12.5;21.5]
Mean (std)	12.8 (2.6)	13.0 (2.0)	23.2 (2.6)	16.5 (5.4)
N (NA)	44 (0)	46 (0)	47 (0)	137 (0)
Per Capita GDP				
Min / Max	0.5 / 1.3	0.6 / 1.3	0.3 / 1.7	0.3 / 1.7
Med [IQR]	0.8 [0.6;1.2]	1.0 [0.7;1.2]	0.8 [0.7;1.4]	0.8 [0.7;1.2]
Mean (std)	0.9 (0.3)	0.9 (0.2)	1.0 (0.4)	0.9 (0.3)
N (NA)	44 (0)	46 (0)	47 (0)	137 (0)
USD Capital				
Min / Max	35.7 / 627.4	23.4 / 501.0	3.2 / 552.2	3.2 / 627.4
Med [IQR]	157.2 [67.1;408.4]	168.4 [67.9;384.7]	84.4 [23.2;335.0]	127.7 [49.8;393.6]
Mean (std)	238.7 (198.7)	217.5 (167.1)	181.5 (198.4)	212.0 (188.7)
N (NA)	44 (0)	46 (0)	47 (0)	137 (0)

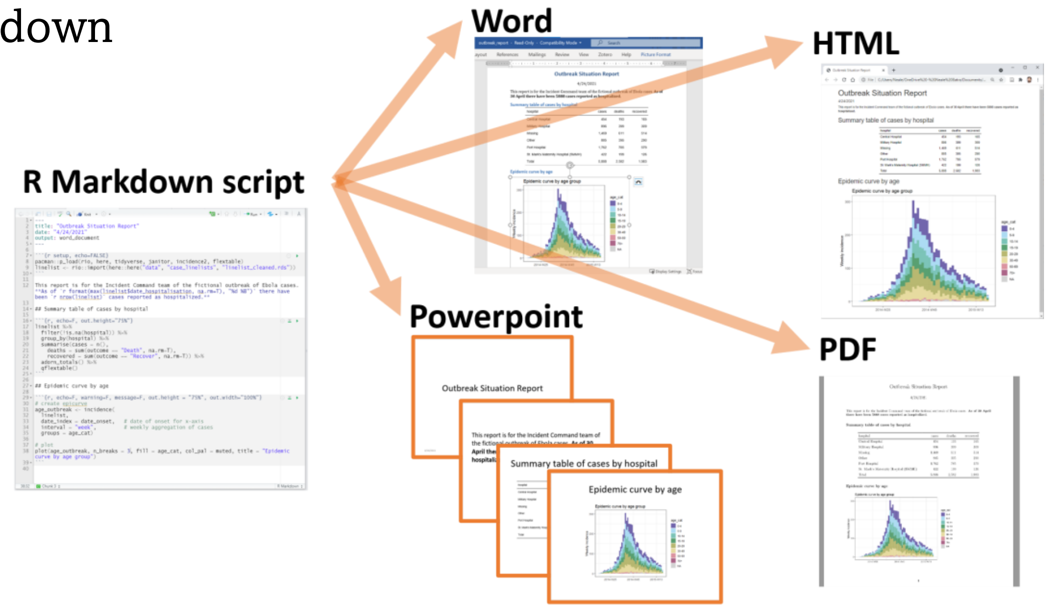


```
 Descriptive statistics by group 
group: AUS
            vars  n    mean      sd  median trimmed     mad     min      max    range  skew kurtosis     se
LOCATION       1 44    1.00    0.00    1.00    1.00    0.00    1.00     1.00     0.00   NaN      NaN   0.00
TIME           2 44 1992.50   12.85 1992.50 1992.50   16.31 1971.00  2014.00    43.00  0.00    -1.28   1.94
PC_HEALTHXP    3 44   12.83    2.57   13.27   12.89    3.41    8.98    16.06     7.08 -0.20    -1.71   0.39
PC_GDP         4 44    0.90    0.30    0.80    0.89    0.33    0.52     1.34     0.81  0.27    -1.65   0.05
USD_CAP        5 44  238.67  198.67  157.23  220.41  162.53   35.72   627.40   591.68  0.64    -1.14  29.95
FLAG_CODES     6 44    1.00    0.00    1.00    1.00    0.00    1.00     1.00     0.00   NaN      NaN   0.00
TOTAL_SPEND    7 43 4886.04 4546.03 2921.40 4359.31 3328.69  475.11 14503.87 14028.76  0.78    -0.85 693.26
------------------------------------------------------------------------------------------- 
group: FIN
            vars  n    mean     sd  median trimmed    mad     min     max  range  skew kurtosis     se
LOCATION       1 46    2.00   0.00    2.00    2.00   0.00    2.00    2.00   0.00   NaN      NaN   0.00
TIME           2 46 1992.50  13.42 1992.50 1992.50  17.05 1970.00 2015.00  45.00  0.00    -1.28   1.98
PC_HEALTHXP    3 46   13.04   2.03   12.89   13.03   2.58    9.81   16.25   6.44 -0.09    -1.23   0.30
PC_GDP         4 46    0.93   0.23    0.97    0.93   0.32    0.64    1.29   0.65  0.06    -1.73   0.03
USD_CAP        5 46  217.52 167.09  168.43  208.92 174.65   23.40  501.00 477.61  0.45    -1.39  24.64
FLAG_CODES     6 46    1.04   0.21    1.00    1.00   0.00    1.00    2.00   1.00  4.33    17.14   0.03
TOTAL_SPEND    7 46 1135.39 908.86  851.32 1080.53 920.19  107.77 2745.26 2637.49  0.50    -1.33 134.00
------------------------------------------------------------------------------------------- 
group: KOR
            vars  n    mean       sd  median trimmed     mad     min      max    range skew kurtosis      se
LOCATION       1 47    3.00     0.00    3.00    3.00    0.00    3.00     3.00     0.00  NaN      NaN    0.00
TIME           2 47 1993.00    13.71 1993.00 1993.00   17.79 1970.00  2016.00    46.00 0.00    -1.28    2.00
PC_HEALTHXP    3 47   23.18     2.55   21.93   22.87    1.34   20.24    29.38     9.14 1.07    -0.20    0.37
PC_GDP         4 47    0.97     0.43    0.79    0.96    0.41    0.32     1.71     1.39 0.47    -1.27    0.06
USD_CAP        5 47  181.51   198.40   84.38  162.54  114.43    3.16   552.18   549.02 0.81    -0.99   28.94
FLAG_CODES     6 47    1.06     0.44    1.00    1.00    0.00    1.00     4.00     3.00 6.42    40.13    0.06
TOTAL_SPEND    7 47 8805.89 10013.30 3729.28 7762.48 5146.42  103.91 28296.65 28192.74 0.86    -0.90 1460.59
> 
```

Usual Analysis Workflow in R



- PDF from rmarkdown
- Word file from rmarkdown
- HTML file from rmarkdown
- Publish to rpubs.com
- R Shiny Webapp
- Flexdashboard
- Slide using xarigan



~~~
---
title: "Outbreak Situation Report"
date: "4/24/2021"
output: word_document
---

```{r setup, echo=FALSE}
pacman::p_load(rio, here, tidyverse, janitor, incidence2, flextable)
linelist <- rio::import(here::here("data", "case_linelists", "linelist_cleaned.rds"))
~~~

This report is for the Incident Command team of the fictional outbreak of Ebola cases.
**As of `r format(max(linelist$date_hospitalisation, na.rm=T), "%d %B")` there have
been `r nrow(linelist)` cases reported as hospitalized.**

Summary table of cases by hospital

```{r, echo=F, out.height="75%"}
linelist %>%
  filter(!is.na(hospital)) %>%
  group_by(hospital) %>%
  summarise(cases = n(),
    deaths = sum(outcome == "Death", na.rm=T),
    recovered = sum(outcome == "Recover", na.rm=T)) %>%
  adorn_totals() %>%
  qflextable()
```

Epidemic curve by age

```{r, echo=F, warning=F, message=F, out.height = "75%", out.width="100%"}
# create epicurve
age_outbreak <- incidence(
  linelist,
  date_index = date_onset,   # date of onset for x-axis
  interval = "week",         # weekly aggregation of cases
  groups = age_cat)

# plot
plot(age_outbreak, n_breaks = 3, fill = age_cat, col_pal = muted, title = "Epidemic
curve by age group")
```
```
```

Some Things to Remember During Live Classes

- Take short notes if required
- Stay muted and turn off camera if not asked to turn on
- Ask questions in the chat.
  Answers will be given at the end of each part of the lessons.
- Stay focused on the classes.
  DO NOT BROWSE FACEBOOK or do others tasks during class time.

To Learn Effectively

- Please attend the live classes regularly
- Reserve some time of your day to practice the codes
- Discussing problems with your peers (classmates) is encouraged
- Share what you have learned with others after each module
- Do projects, solve real problems in your field
- Take good care of your health

We learn more by looking for the answer to a question and not finding it than we do from learning the answer itself.

- Lloyd Alexander