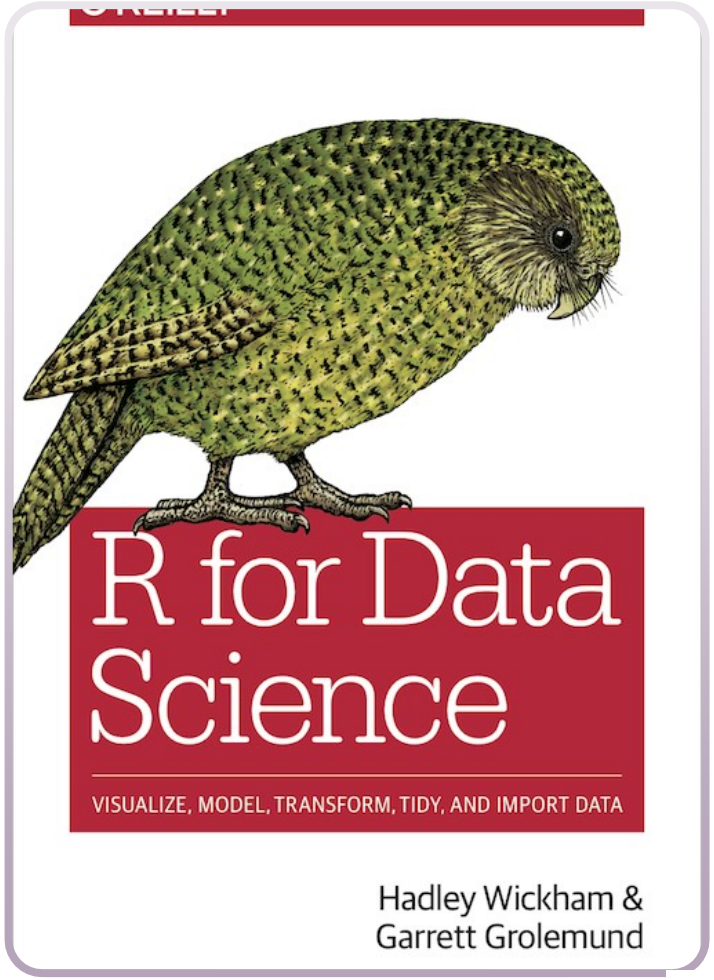
Data Transformation and Introduction to Tidyverse

Divya Seernani, R-Ladies Freiburg

Freiburg

Resources

- Book – Wickham and Grolemund
- 3 part complete ggplot tutorial by r-statistics.co
- Various material put up by r-ladies on github
- Dpylr and ggplot- https://www.onceupondata.com/2019/01/04/datafest-tbilisi/



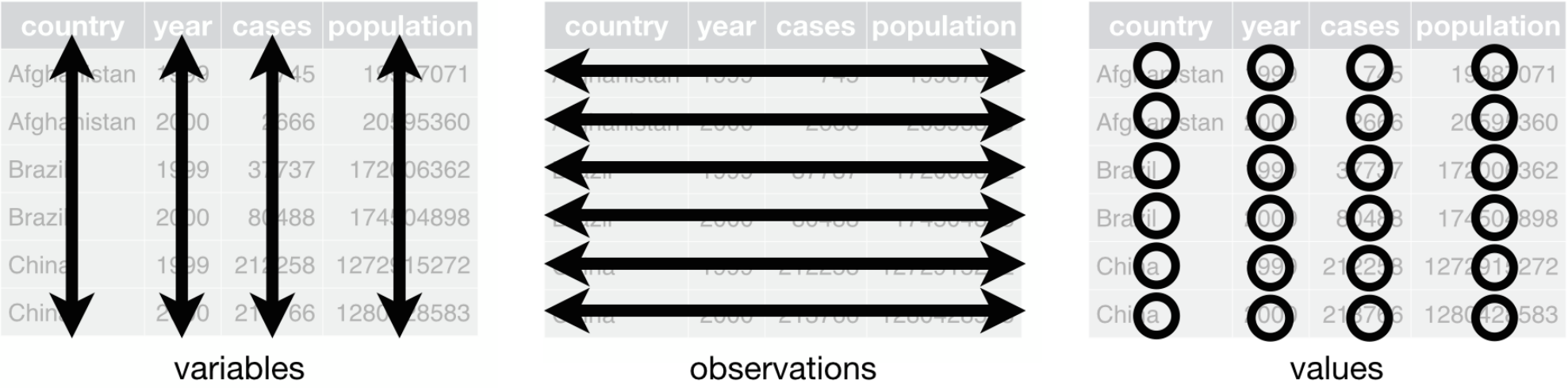
Freiburg

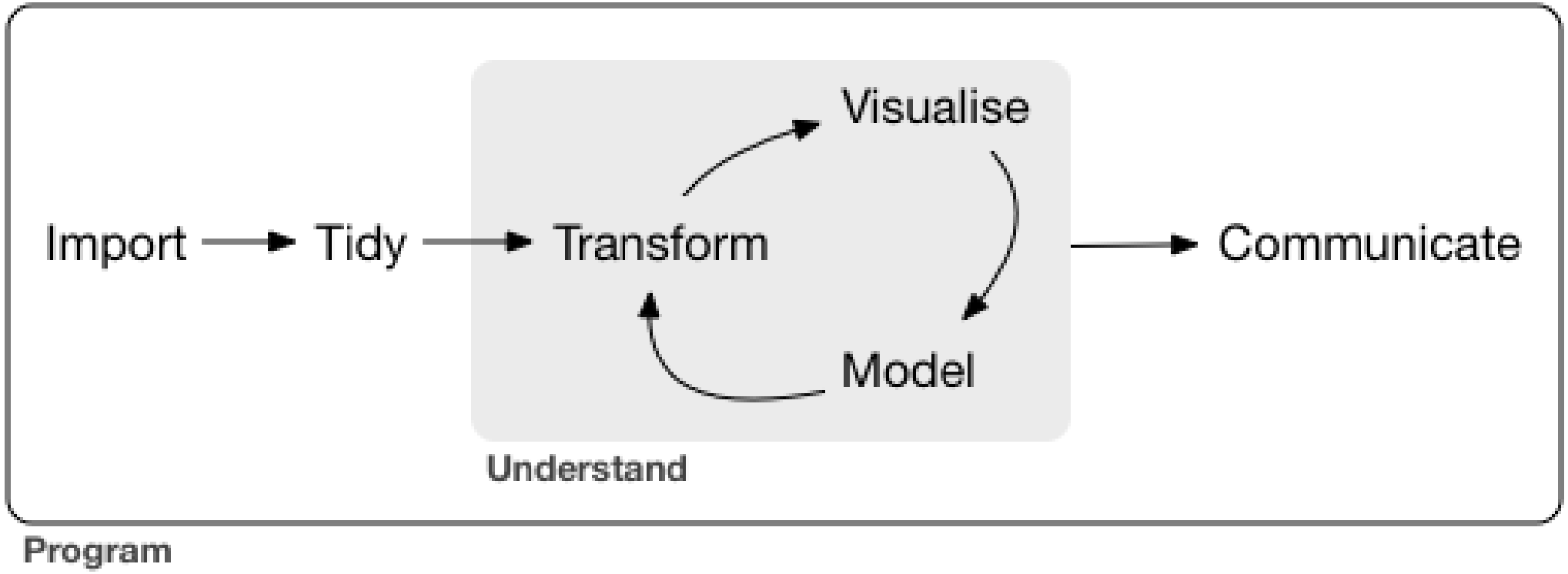
Tidyverse

- A number of packages that work well together
- Same underlying principles = same way of thinking about problems
- readr, dpylr, ggplot2, tibble, tidyr and purr packages

Freiburg

What is a Tidy Dataset?




Import
Tidy
Transform
Visualise
Model
Communicate
Understand
Program

Freiburg

Data Transformation

- The dplyr package
- Narrowing Observations
- Creating new variables
- Rename/Reorder observations
- Calculating summary statistics

Freiburg

The Indian Census, 2011

https://github.com/nishusharma1608/India-Census-2011-Analysis

Read into R-Studio

Look at the data – str()

Library(tidyverse)

Freiburg

The verbs of dplyr

- Pick Observations – filter()
- Reorder rows – arrange()
- Pick variables – select ()
- Create new variables – mutate()
- Collapse many values to a summary – summarise()
- Operate above functions on a group-by-group basis – group_by()

- Wickham and Grolemund

Freiburg

Make sure to use the right operators!

Freiburg

Filter – Western States of Gujarat and Maharshtra

- X<- filter (df, expressions)
- Careful – do you mean and?
- How are they written? – Verify spelling and case.

```
WestStates <- filter(india.districts.census.2011, State.name =='MAHARASHTRA'|
State.name =='GUJARAT')
```

Freiburg

Arrange districts based on agricultural workers

- X<- arrange (df, order_by column)
- Default is ascending – use desc() to arrange descending

```
AgriDist<- arrange(Maharashtra, Agricultural_Workers)
AgriDistDes<- arrange(Maharashtra, desc(Agricultural_Workers))
```

Freiburg

Mutate

- Absolute numbers don't tell us much.
- What percentage of households have computers or the internet?
- Use the entire dataset
- X<- mutate (df, variable=function of exsisting columns)

```
IndCen2011Calculations<- mutate(india.districts.census.2011,
    PercentInternet  =  Households_with_Internet / Households * 100,
    PercentComputer  =  Households_with_Computer / Households * 100)
```

Freiburg

Select

- We are only interested in data on households with latrines and bathing facilities
- X<- select (df, columns)

ModernHomes<-select(india.districts.census.2011, State.name, District.name, Households, Having_bathing_facility_Total_Households, Having_latrine_facility_within_the_premises_Total_Households)

Freiburg

Back to mutate

- So many numbers!!!! Mutate to standardize them to percentages

```
ModernHomes2<-mutate(ModernHomes, PercentToilet =
Having_latrine_facility_within_the_premises_Total_Households / Households * 100,
        PercentBath = Having_bathing_facility_Total_Households / Households * 100)
```

Freiburg

Let's visualize this!

THEME

COORDINATES

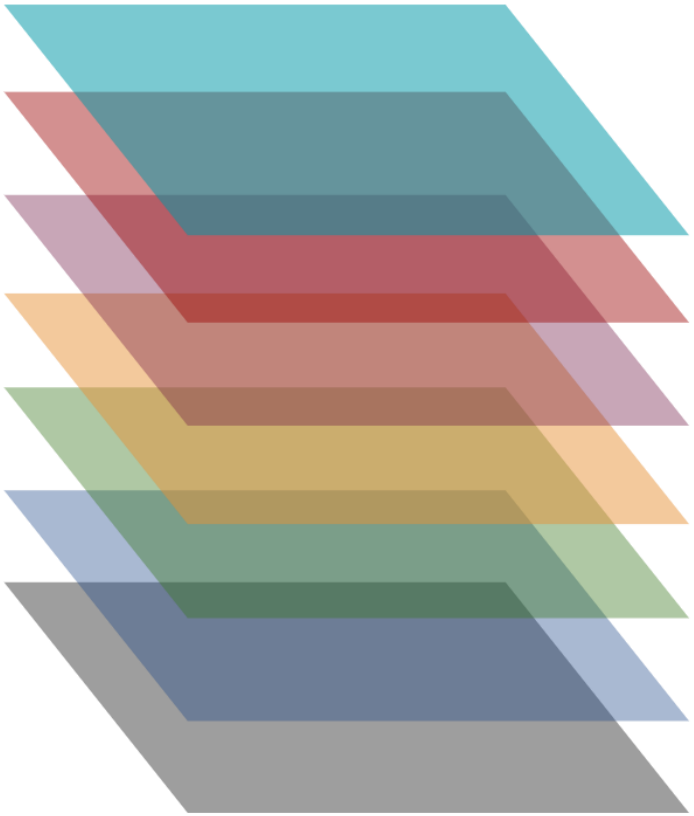
STATISTICS

FACETS

GEOMETRIES

AESTHETICS

DATA



https://github.com/OmaymaS/datafest_tidyverse_workshop/blob/master/slides/data_visualization_with_ggplot2.pdf

Freiburg

```
library(ggplot2)

ModernHomesPlot <- ggplot(ModernHomes2, aes(x=PercentToilet, y=PercentBath)) +
geom_point(aes(col=State.name, size=Households)) +
geom_smooth(method="glm", se=T) +
labs(subtitle="Toilets and Baths",
     y="Percentage with Bath",
     x="Percentage with Toilet",
     title="Scatterplot",
     caption = "Indian Census 2011")

plot(ModernHomesPlot)
```

http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

Freiburg

Summarize

- Mean literacy of India
- Mean literacy by state – group_by

```
summarize(india.districts.census.2011, PercentLiterate = mean(Literate/Population*100))

Literacy<-group_by(india.districts.census.2011, State.name)

LiteracyByState<-summarize(Literacy, PercentLiterate = mean(Literate/Population*100))
```

Freiburg

Exercice : Literacy Hygiene Relationship

```
LiteracyHygiene <- select(india.districts.census.2011, State.name, District.name, Literate, Households, Having_latrine_facility_within_the_premises_Total_Households)

LiteracyHygiene_Calculated<-mutate(LiteracyHygiene, PercentToilet = Having_latrine_facility_within_the_premises_Total_Households / Households * 100, AverageLiterate = Literate/Households)


LiteracyHygienePlot <- ggplot(LiteracyHygiene_Calculated, aes(x=PercentToilet, y=AverageLiterate)) +
  geom_point(aes(col=State.name, size=Households)) +
  geom_smooth(method="glm", se=T) +
  labs(subtitle="Toilets and Baths",
       y="Average Literate People per Household",
       x="Percentage with Toilet",
       title="Scatterplot",
       caption = "Indian Census 2011")


plot(LiteracyHygienePlot)
```

Freiburg

Exercice : Literacy Hygiene Relationship

```
library(plotly)
ggplotly(LiteracyHygienePlot)
```

Freiburg

Happy Working!

All slides and code available on R-Ladies Freiburg Github account

Twitter @RLadiesFreiburg

Email – freiburg@rladies.org

Next Meetup – 3rd July – Even more tidyverse!

Freiburg