STAT 184: Introduction to R

Instructor: Soumya Mukherjee
Content credit: Dr. Matthew Beckman and Olivia Beck

July 3, 2023

Objectives

- Quick introduction
- Motivation of the course
- Course Materials
- Syllabus fly-over
- Get R and RStudio Installed

Examples of data

Iris data

- This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

- This is a dataset that is directly available in R.

- Randomly selected rows from the data:

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 142          6.9         3.1          5.1         2.3  virginica
## 68           5.8         2.7          4.1         1.0 versicolor
## 129          6.4         2.8          5.6         2.1  virginica
## 43           4.4         3.2          1.3         0.2     setosa
## 14           4.3         3.0          1.1         0.1     setosa
## 51           7.0         3.2          4.7         1.4 versicolor
## 85           5.4         3.0          4.5         1.5 versicolor
## 21           5.4         3.4          1.7         0.2     setosa
## 106          7.6         3.0          6.6         2.1  virginica
## 74           6.1         2.8          4.7         1.2 versicolor
```

Gapminder

- Income & Life Expectancy:
  - *https://www.gapminder.org/tools/#$chart-type=bubbles&url=v1*
  - *How many variables are being displayed here?*
  - *What are they?*
  - *What did you learn by watching the animation?*
- Dollar Street
  - *https://www.gapminder.org/dollar-street/matrix*
  - *Are we even looking at "data" here?*

Medicare Spending

- Newspaper article here
- Data available here.

The logic of this course: Less volume, more creativity.

- A minimal set of simple tools can be combined in powerful ways
- Individually these tools are introduced as simple "data moves"
- The complexity of data use and presentation comes from combining these simple tools in order to achieve our specific purposes.

Individual lego bricks are simple.[1]

A complex model made of lego br

The logic of this course: Less volume, more creativity.

We will start with some infrastructure for these techniques:

- basics of "tidy" data
- commanding the computer to handle data and present the story we want to tell
  - *focus on "exploratory data analysis"*
  - *we will combine graphs, numerical summaries, and narrative*
  - *we defer to future STAT courses for formal "modeling"*

In coming weeks, we will study

- relationships between data and graphical presentations
- proper form of data to make a graphic (or a model)
- how to "wrangle" data we have into the form needed for a graphic (or a model)
- use of layering to develop rich graphics

Examples from many fields

Some data sets we will access for examples.

```
BabyNames                Names of children as recorded by the US Social
                            Security Administration.
CountryCentroids         Geographic location of countries
CountryData              Many variables on countries from the 2014 CIA factbook.
CountryGroups            Membership in Country Groups
DirectRecoveryGroups
MedicareCharges
MedicareProviders
MigrationFlows           Human Migration between Countries
Minneapolis2013          Ballots in the 2013 Mayoral election in Minneapolis
NCI60                    Gene expression in cancer.
NCI60cells               Cell Line descriptions in the NCI-60 dataset
WorldCities              Cities and their populations
ZipDemography            Demographic information for most US ZIP Codes (Postal Codes)
ZipGeography             Geographic information by US Zip Codes (Postal Codes)
registeredVoters         A sample of the voter registration list for Wake County,
                            North Carolina in Fall 2010.
```

Exploratory data analysis (EDA)

Due to the nature of the course and our goals, we will sometimes characterize our work from the paradigm of *Exploratory data analysis (EDA)*. Some goals for high-quality EDA include:

- Become acquainted with the data
- Explore intuition for research question
- Discover features in the data that impact modeling decisions

Orientation to Class Resources

- Lots of essential stuff launches from Canvas directly
  - *Textbook (Data Computing eBook)*
  - *Canvas Discussions*
  - *RStudio Cloud*
- R and RStudio
  - ***R** does the computations*
  - ***RStudio** is an interface to R that makes it easier to document your work, access nice features of R, and more*
  - ***Rstudio Cloud** is a cloud based platform that allows you to run RStudio on your browser*
- git and GitHub
  - ***git** is version control software used to facilitate collaboration and project evolution*
  - ***GitHub** is a remote git repository hosting service*
- Access to R & RStudio (It is important that you do these ASAP)
  - *Installing on your own computer. You need both of these:*
    - R software
    - RStudio desktop software.
    - Packages: Paste the following R commands into your RStudio Console and run it (press Enter):
  - *RStudio Cloud: https://posit.cloud/*
    - Choose **Sign up**
    - Choose the Free plan
    - Sign up using your psu email ID, set a password and enter your first and last name
    - Pro: Access all your R and RStudio stuff from any computer anywhere
    - Con: Takes some practice to get used to uploading files to & exporting files from the cloud interface. Also, memory allocations are limited and difficult to work with large data sets here.
- R Packages / Libraries: these are software modules that extend the functionality of R.
  - *We need to **install** the package just once to download it on your computer.*
  - *We need to load the package/library to use it each time we start a new R session*
- All of my course notes/slides are posted on Canvas (under modules) and GitHub: <>

Assignments and Questions

- Assignments, grades, etc will be handled with Canvas
- Use Canvas Discussions for most questions about course content, e.g. “How do I rename a variable?”; see syllabus for detail
  - *I’ll redirect you to post things on Canvas Discussions when other students would benefit from the question and the answer*
  - *Email is generally for private issues that aren’t relevant to other students*

1. Source :“Lego Color Bricks” by Alan Chia - Lego Color Bricks. Licensed under CC BY-SA 2.0 via Wikimedia Commons↩

2. Source: *Trafalgar Legoland 2003* by Kaihsu Tai - Kaihsu Tai. Licensed under CC BY-SA 3.0 via Wikimedia Commons↩