DATA ANALYTICS

Unit 1:The R Programming Environment

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 1:The R Programming Environment

Mamatha H R

Department of Computer Science and Engineering

R PROGRAMMING LANGUAGE

- R is a free, open-source programming language and software environment for **statistical computing, bioinformatics, visualization and general computing.**

  R is an **interpreted language.**

  Ross Ihaka & Robert Gentleman in 1993

  Extensive catalog of statistical and graphical methods

  Environment for statistical analysis, graphics representation and reporting

DATA ANALYTICS

What is R?

- **Most widely used data analysis software**

  Used by 2M+ data scientists, statisticians and analysts

- **Most powerful statistical programming language**

  Flexible, extensible and comprehensive for productivity

- **Create beautiful and unique data visualizations**

  As seen in New York Times, Twitter and Flowing Data

- **Thriving open-source community**

  Leading edge of analytics research

What is R used for?

Statistical Computing

Data analysis

Machine learning algorithm

Visualization

Scientific research

Why use R for data analytics, statistical computing and graphics?

- R is open source and free!
- R is popular – and increasing in popularity
- R runs on all platforms
- Thousands of R packages.
- R is being used by the biggest tech giants
- R is interactive language and expressive syntax
- R has easy-to-use interface.

Applications of R Programming in Real World

R – allows to collect data in real-time, perform statistical and predictive analysis, create visualizations and communicate actionable results to stakeholders

More than 9100 packages of statistical function.

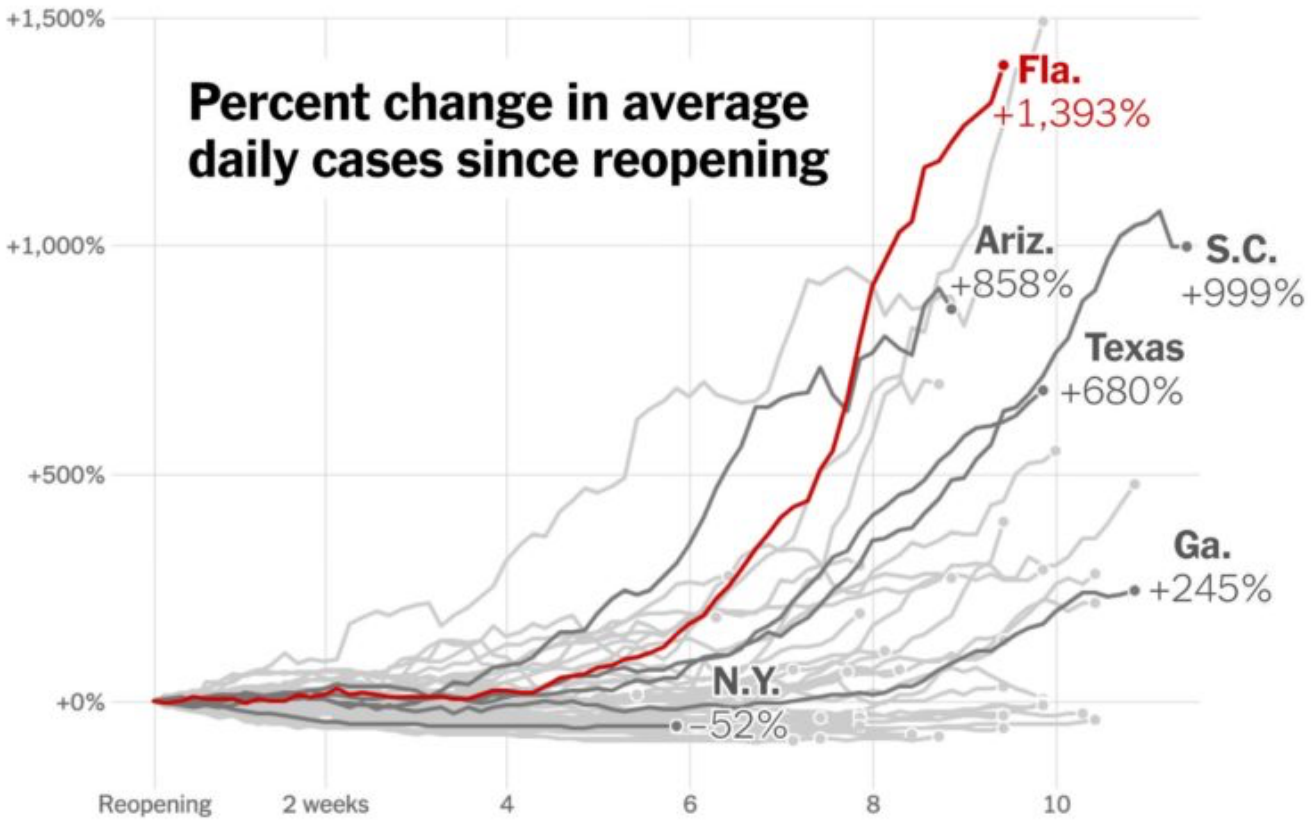
R's expressive syntax allows to quickly import, clean and analyze data from various data sources.

R also has charting capabilities, which means you can plot your data and create interesting visualizations from any dataset.

DATA ANALYTICS

As seen in New York Times, Twitter and Flowing Data

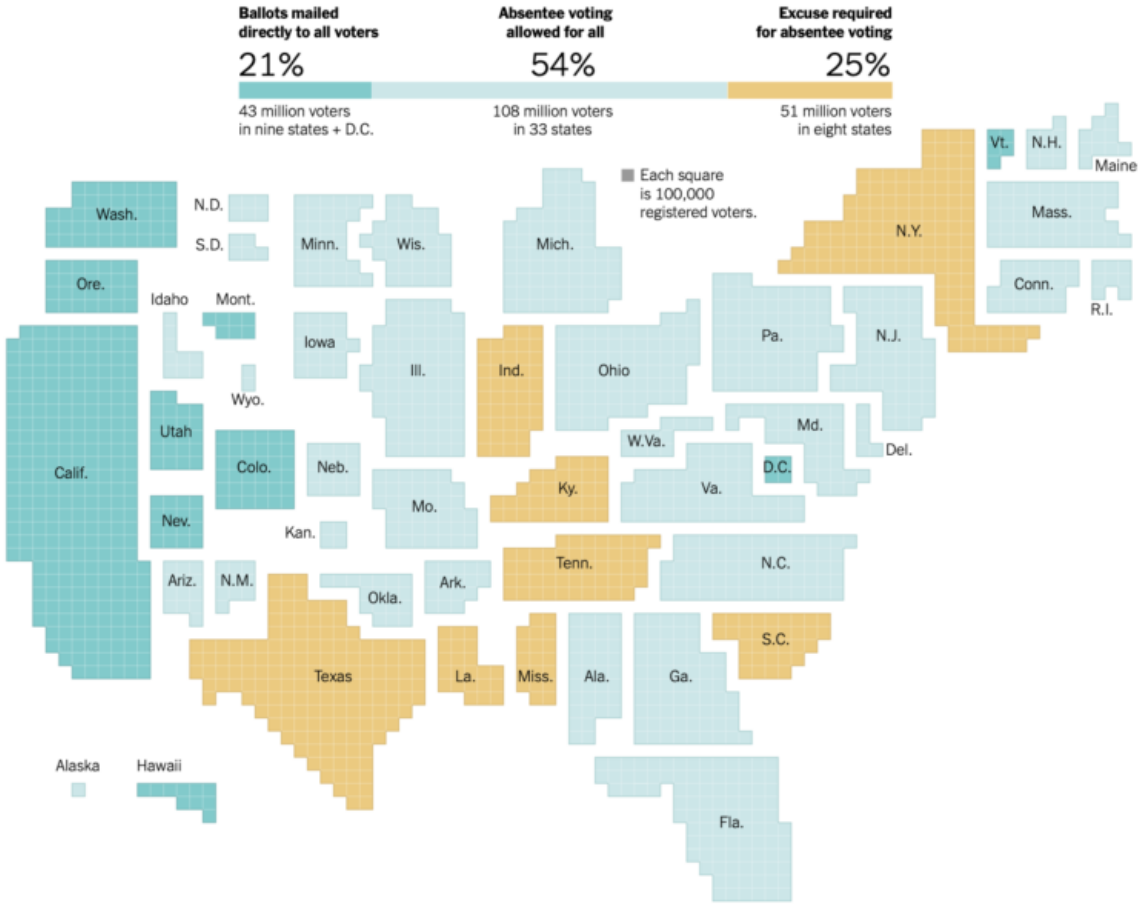
Increase in cases since states reopened



Source: https://flowingdata.com/2020/07/10/increase-in-cases-since-states-reopened/

As seen in New York Times, Twitter and Flowing Data

Who can vote by mail ?

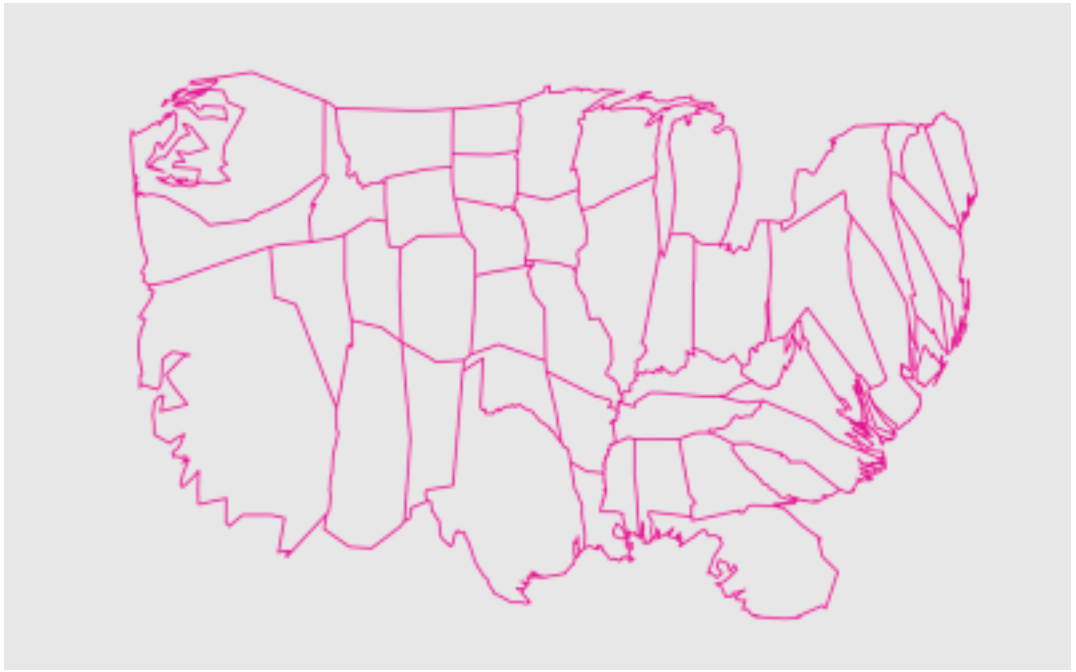


There's going to be a lot more voting by mail this year. The New York Times shows what each state is doing. It's a cartogram. So it must be election season.

Source: https://flowingdata.com/2020/08/13/who-can-vote-by-mail/

Cartogram

A **cartogram** is a map in which the geometry of regions is distorted in order to convey the information of an alternate variable



Source: https://flowingdata.com/charttype/cartogram/

Applications of R Programming in Real World

R - used in predictive analytics and machine learning.

Packages for ML tasks like linear and non-linear regression, decision trees, linear and non-linear classification and many more.

Quick Facts on R

- R is the highest paid IT skill
- R most-used data science language after SQL
- R is used by 70% of data miners
- R is #15 of all programming languages
- R growing faster than any other data science language
- R is the #1 Google Search for Advanced Analytics software
- R has more than 2 million users worldwide

Alternatives to R programming

- Python – Popular general purpose language
- SAS (Statistical Analysis System)
- SPSS – Software package for statistical analysis

History of R

- The **S** language has been developed since the late 1970s by **John Chambers** and colleagues at Bell Labs as a **language for programming with data.**

  - S language combines ideas from a variety sources (awk, lisp, APL,) and provides an environment for quantitative computations and visualization.

  - Provides an explicit and consistent structure for manipulating, analyzing statistically, and visualizing data.

- **S-Plus** is a commercialization of the Bell Labs framework. It is “S" plus “graphics".

History of R

- R - Open source statistical environment/platform developed by **R**obert Gentleman and **R**oss Ihaka (U of Auckland, NZ) during the 1990s.
- R is currently maintained by the R core-development team, a hard-working, international team of volunteer developers.
- The primary R system is available from the Comprehensive R Archive Network, also known as CRAN.
- CRAN also hosts many add-on packages that can be used to extend the functionality of R. Over 6,789 packages are available on CRAN that have been developed by users and programmers around the world.

Finding out the latest version of R:

- To find out what is the latest version of R, you can look at the CRAN (Comprehensive R Network) website, http://cran.r-project.org/.

Installing R on Windows:

- To install R on your Windows computer, follow the below steps:
- Go to http://ftp.heanet.ie/mirrors/cran.r-project.org. or https://cran.rstudio.com/bin/windows/base/
- Under “Download and Install R”, click on the “Windows” link.
- Download the required the .exe file. You should see a link saying something like “Download R 3.4.0 for Windows” (or R X.X.X, where X.X.X gives the version of R, eg. R 3.4.0). Click on that link.
- After downloading double-click on the R-3.4.0-win.exe to run it.
- You will be asked what language to install it in – choose English.

Download and Install RStudio on Windows

https://www.rstudio.com/products/rstudio/download/#download/
and click on DOWNLOAD RSTUDIO DESKTOP.

List of IDEs

- RStudio
- StatET for R (eclipse based)
- R-Brain IDE (RIDE)
- IntelliJ IDEA
- R Tool for Visual Studio

Setting your Working Directory

R getwd() Function

- Working directory is the directory where R finds all R file for reading and writing.
- getwd() function returns an absolute filepath representing the current working directory of the R process.
  - **getwd()**
  - **"C:/Users/*****/Documents"**

Setting your Working Directory

R setwd() Function

- setwd(dir) is used to set the working directory to dir.

```
setwd("e:/folder/")
```

R dir() Function

- dir() function lists all the files in a directory.

```
dir()
```

R ls() Function

- ls() is a function in R that lists all the object in the working environment. **ls()**
- It can be used in scenario where you want to clean the environment before running code. Below command will remove all the object from R environment.

```
rm(list = ls())
```

GETTING HELP WITH R

- To get general help just type the below command

  help.start()

- To access documentation for the standard lm (linear model) function, for example, enter the command.

  help(lm) / help("lm") / ?lm /?"lm“

- To see the list of pre-loaded data (datasets), type the function,

  data()

INSTALLING R PACKAGES

- The primary location for obtaining R packages is CRAN.

- Information about the available packages on CRAN with the ***available.packages()*** function.

```
a <- available.packages()
```

- Packages can be installed with the ***install.packages()*** function in R.

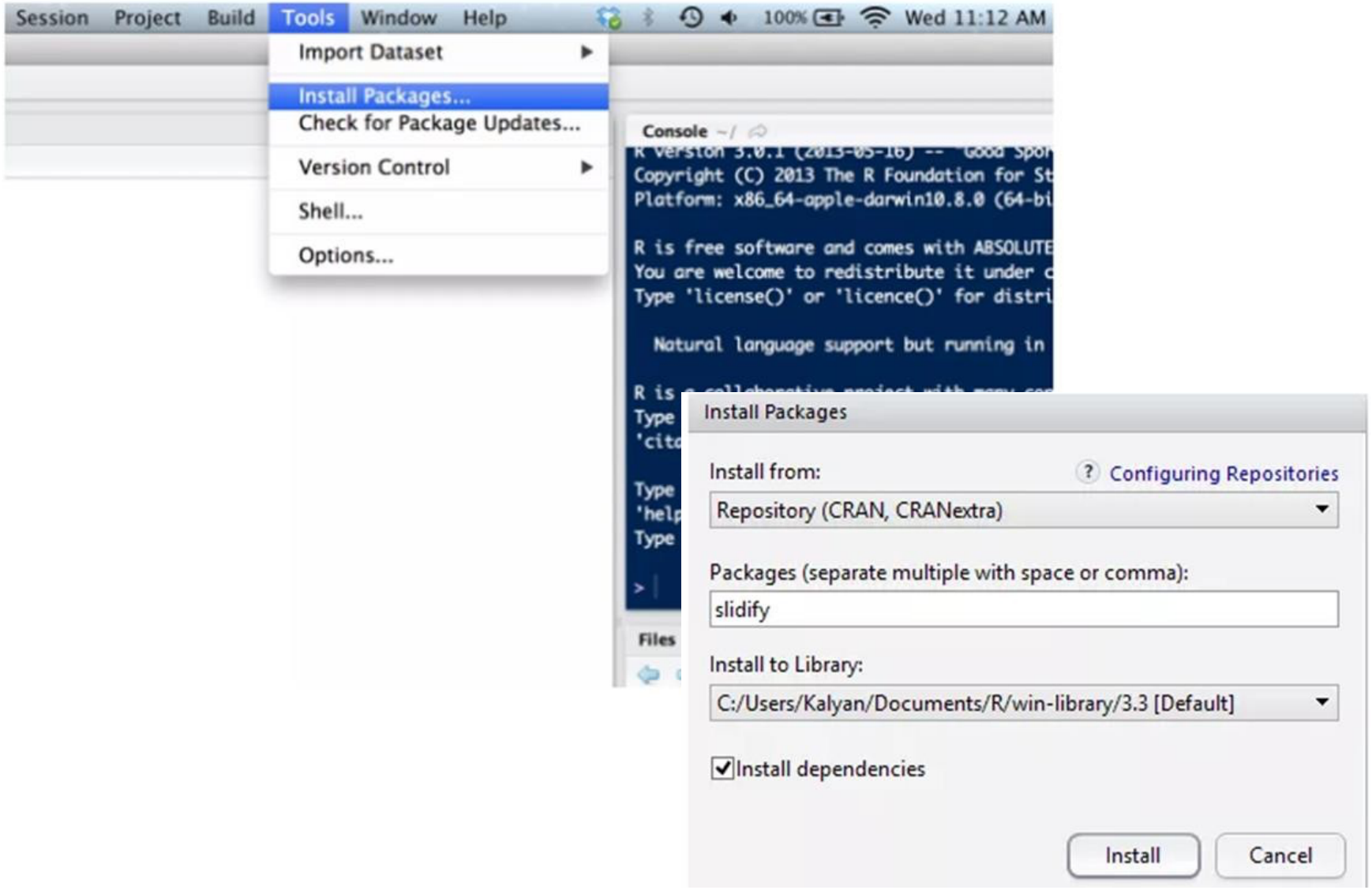
```
install.packages("ggplot2")
```

- Install multiple R packages at once with a single call to ***install.packages().*** Place the names of the R packages in a character vector.

```
install.packages(c("caret", "ggplot2", "dplyr"))
```

DATA ANALYTICS

Installing an R Package in RStudio



LOADING R PACKAGES

- Installing a package does not make it immediately available to you in R; you must load the package. The library() function is used to load packages into R.

library(ggplot2)

- After loading a package, the functions exported by that package will be attached to the top of the search() list (after the workspace).

library(ggplot2)
search()

- To save your workspace to a file, you may type

save.image() or use **Save Workspace…** in the **File** menu

- The default workspace file is called .RData

Exercise

Install R and R Studio

References

https://www.tutorialspoint.com/

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834