

R Essential Training

By: Ammar Jabakji

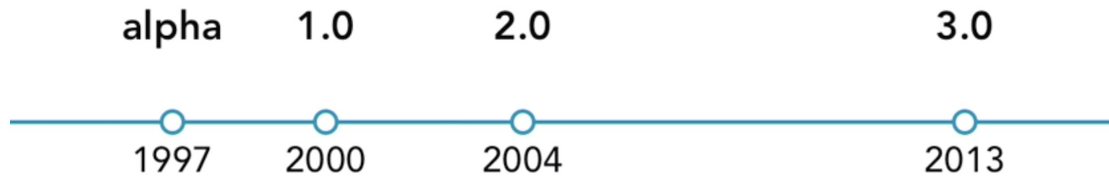
What is R?

- R was created by Ross Ihaka and Robert Gentleman
- Encouraged to make R free software
- R is not Statistical languages but a programming language that works well with statistics. Not a Software.
- R is extensible; can be expanded by installing “packages”
- R is command-line driven.
- Primary using for data analysis, data modeling , and visualization.

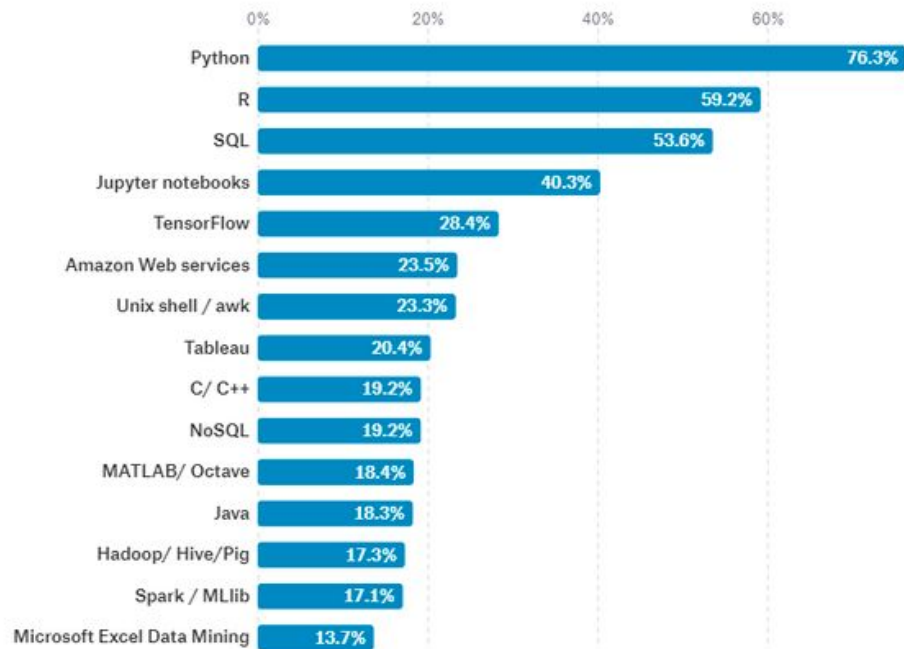
Agenda

- Day:1 Introduction to R and Rstudio
- Day:2 Manipulation data using R for Data Science
- Day:3 Data virtualization
- Day:4 Data wrangling in R
- Day:5 Exploratory data

Versions of R



R has seen quick adoption in the past 10-20 years.



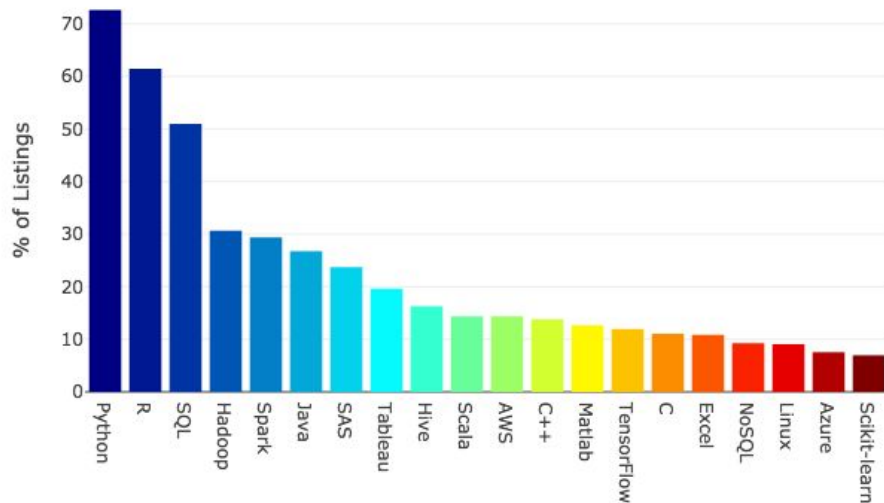
7,955 responses

Only displaying the top 15 answers. There are 38 answers not shown.

Programming Languages Most Used and Recommended by Data Scientists

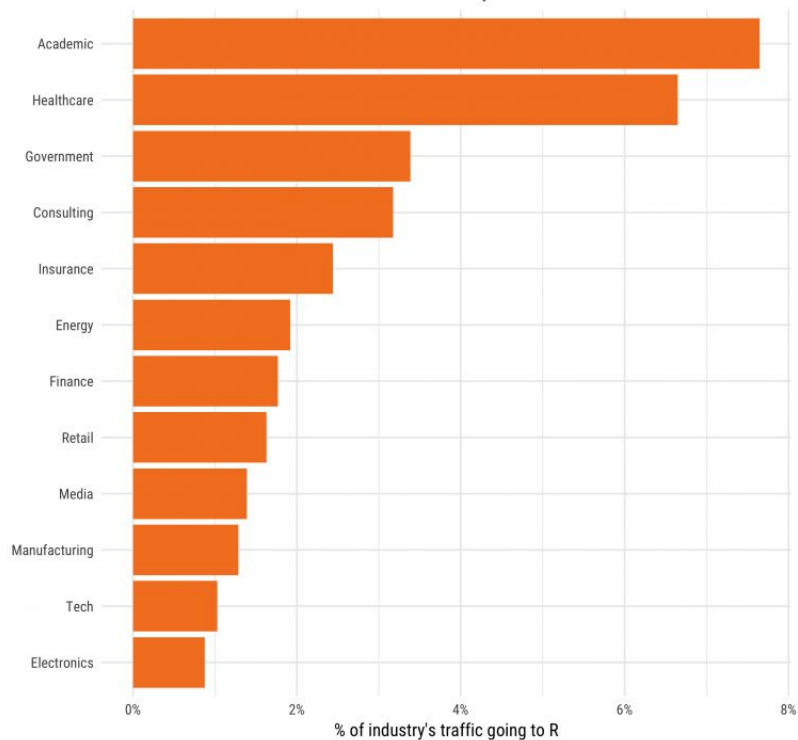
Data Scientist Job Listings

Top 20 Technology Skills in Data Scientist Job Listings



Visits to R by industry

Based on visits to Stack Overflow questions from the US/UK in January-August 2017.
The denominator in each is the total traffic from that industry.



Who uses R most?



Companies that use R for Analytics

NOVARTIS

Google

accenture
High performance. Delivered.

Thomas Cook

MERCK

TECH TechCrunch

ORBITZ

facebook

genpact

Bing



wipro

ANZ

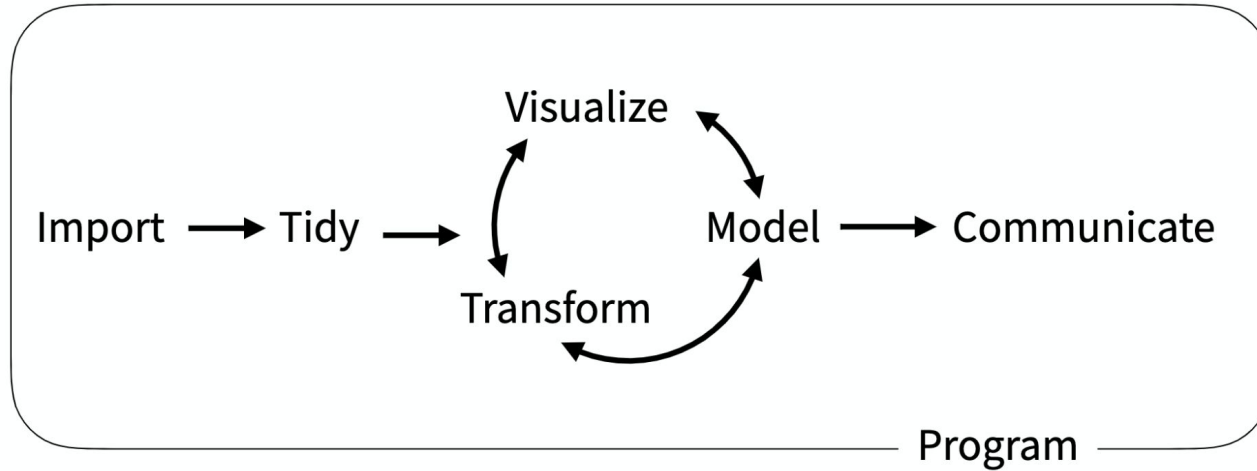
The
New York
Times

Who uses R most?

Data Science

By: Ammar Jabakji

(Applied) Data Science



1

Data import

2

Data
wrangling

3

Data
visualization

Tidyverse

By: Ammar Jabakji



The tidyverse is two things simultaneously:

- A collection of R packages
- An approach to how to do data science with the R language

tidyverse.org

The screenshot shows the tidyverse.org website. The browser tab is titled "tidyverse website • tidyweb" and the address bar shows "tidyverse.org". The page has a header "The tidyverse" and a section "Components" featuring a grid of 12 hexagonal icons for R packages: dplyr, forcats, ggplot2, haven, lubridate, %>% (with the text "Chain. It just makes code simpler."), purrr, readr, readxl, stringr, TIBBLE, tidyr, and tidyverse. Below the grid, a paragraph states: "The tidyverse is a collection of R packages that share common philosophies and are designed to work together. This site is a work-in-progress guide to the tidyverse and its packages." On the right, a section titled "#tidyverse tweets" displays three tweets from users PJ B, Martin Monkman, and Hilary Robbins, each with a retweet count.

tidyverse website • tidyweb x

tidyverse.org

The tidyverse

Components

The tidyverse is a collection of R packages that share common philosophies and are designed to work together. This site is a work-in-progress guide to the tidyverse and its packages.

#tidyverse tweets

PJ B @ProBorBargain
I may be excommunicated for it, but readxl and lubridate are heretical packages that should not belong to the tidyverse. #rstats #tidyverse
27m

Martin Monkman @monkmarmm
Essential reading as-is. I think I'll leave a paper copy on the lunch room table. #tidyverse #datascience Thanks @kwbroman & @kara_woo
2h

Hilary Robbins @hilaryarobbins
Raise your hand if you constantly write broken #tidyverse code because of British vs. American spellings 🇬🇧 #rstats #summerSe

- **tibble** replaces data frames with tibbles
- **readr** and **readxl** facilitate data import and export
- **dplyr** and **tidyr** perform data manipulation
- **stringr** manipulate text strings
- **ggplot2** enables data visualization

```
library("tidyverse")
```

does the equivalent of

```
library("ggplot2")  
library("dplyr")  
library("tidyr")  
library("readr")  
library("purrr")  
library("tibble")  
library("stringr")  
library("forcats")
```

What Makes tidyverse Different?

- The core of the tidyverse is developed by RStudio
- Long-term support is assured by RStudio's own dependence on the tools
- The tidyverse is developed openly on GitHub

1

Data import

2

Data
wrangling

3

Data
visualization

tidyverse: Data Import

- readr completely overhauls the process for importing rectangular data (.csv, .tsv, etc.)
1. Significantly faster
 2. More intelligent (imports dates as dates, and numbers as numbers)
 3. Never converts strings to factors

tidyverse: Data Wrangling

- The tidyverse utilizes pipes (`%>%`) to provide a logical framework for chaining together common data wrangling tasks
- The tidyr library provides tools for reshaping and transforming data ready to be manipulated in the tidyverse
- dplyr is the workhorse for subsetting, filtering, summarizing and generally wrangling your data

tidyverse: Data Visualization

- ggplot2 is a powerful, consistent grammar of graphics, allowing complex **static** charts to be built easily
- htmlwidget allows R developers to build rich, interactive charts for the web
- Shiny allows completely custom interactive web apps to be built using R



`%>%`

`%>%` is pronounced “pipe”

The pipe operator is “syntactic sugar,” and it’s the workhorse of the tidyverse

R

```
7 data <- c(1, 3, 5, 7, 11, 13, 15, 17)
8 mean(diff(data))
```

In traditional R, expressions
need to be rewritten for new
operations to be added.

%>%

```
1 library(tidyverse)
2 data <- c(1, 3, 5, 7, 11, 13, 15, 17)
3 data %>%
4   diff () %>%
5   mean()
```

The pipe allows operations to be
simply chained together

What Are data.frames and Tibbles?

- data.frame is base R's "standard rectangular data store"
- Tibble is the tidyverse's "standard rectangular data store"
- data.table is a highly optimized "rectangular data store"
- Matrices are different and not designed for "arbitrary rectangular data"

data.frame vs. Tibble

data.frame

- Created by base R functions
- Print output doesn't include column info
- Converts strings to factors
- `class(df)`
`[1] "data.frame"`

Tibble

- Created by tidy verse functions
- Print output is prettier and includes column info
- Strings remain as strings
- `class(tibble)`
`[1] "tbl_df" "tbl" "data.frame"`