R Essential Training

By: Ammar Jabakji

What is R?

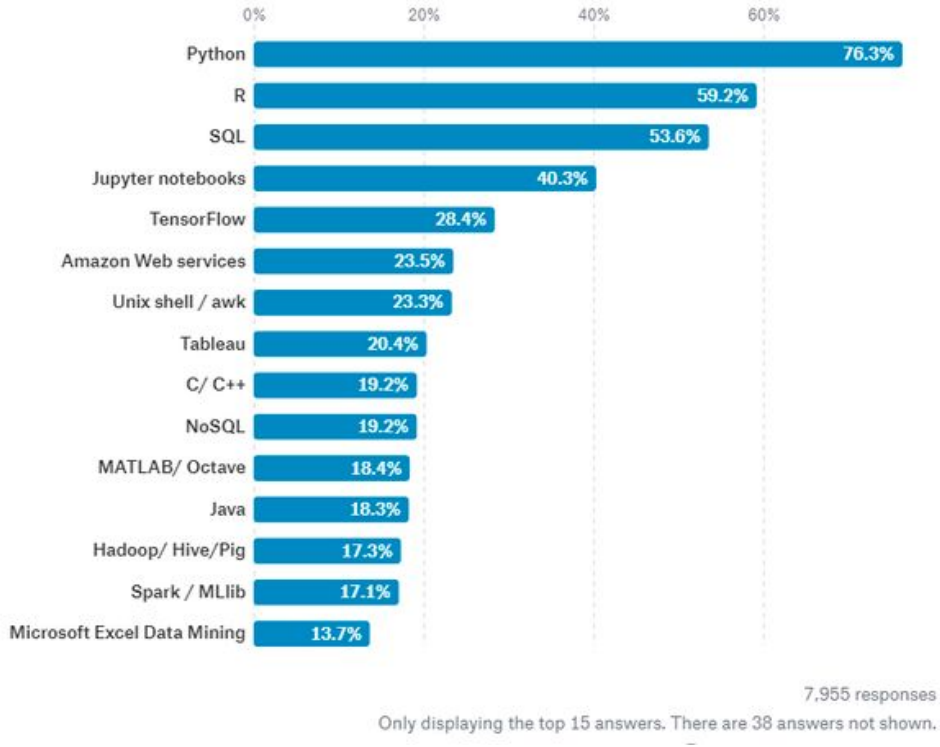
- R was created by Ross Ihaka and Robert Gentleman
- Encouraged to make R free software
- R is not Statistical languages but a programming language that works well with statistics. Not a Software.
- R is extensible; can be expanded by installing “packages”
- R is command-line driven.
- Primary using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

Agenda

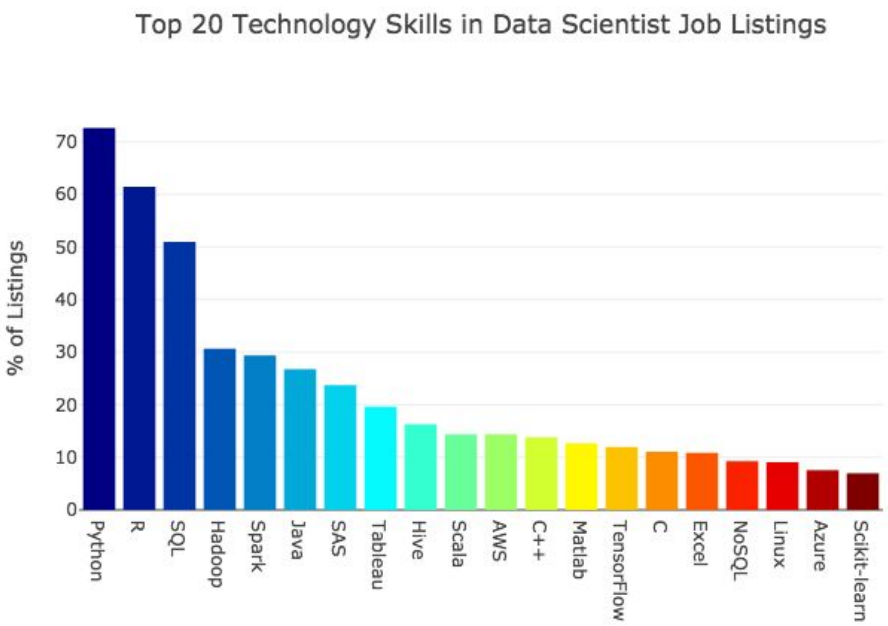
- Day:1 Introduction to R and Rstudio
- Day:2 Manipulation data using R for Data Science
- Day:3 Data virtualization
- Day:4 Data wrangling in R
- Day:5 Exploratory data

Versions of R

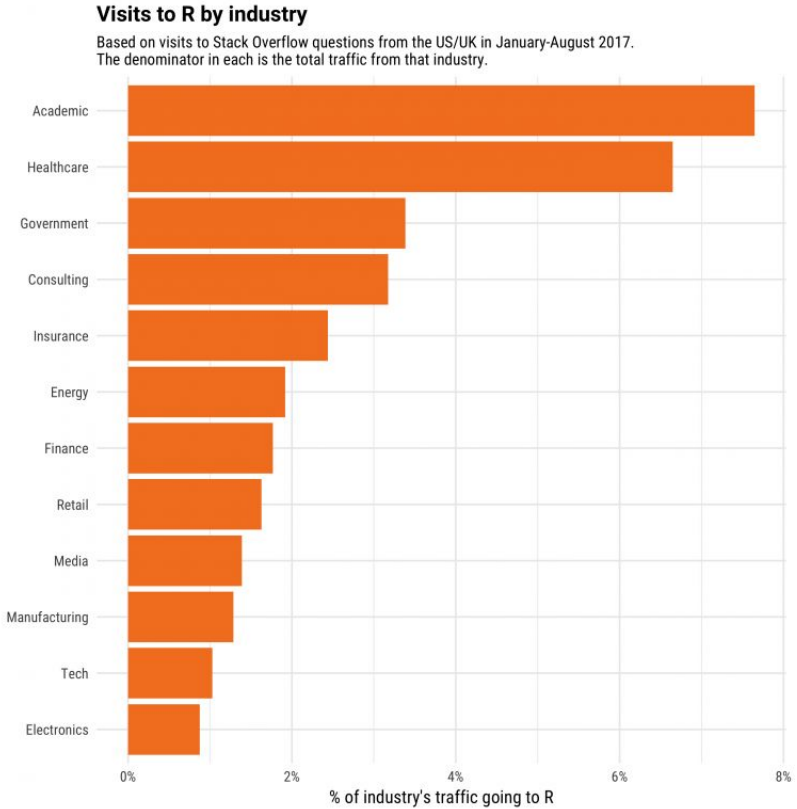
R has seen quick adoption in the past 10-20 years.



Programming Languages Most Used and Recommended by Data Scientists



Data Scientist Job Listings



Who uses R most?

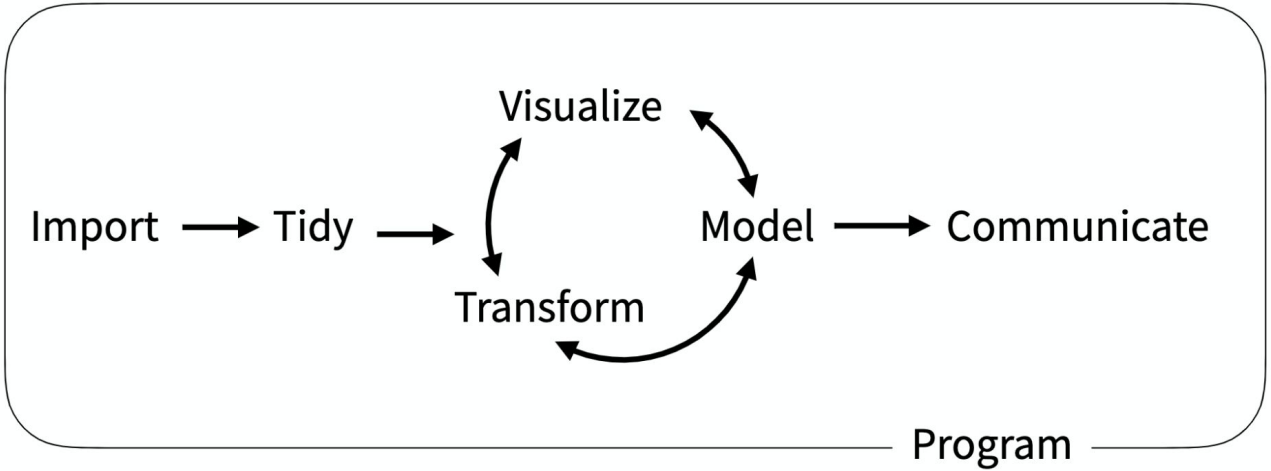
Companies that use R for Analytics

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

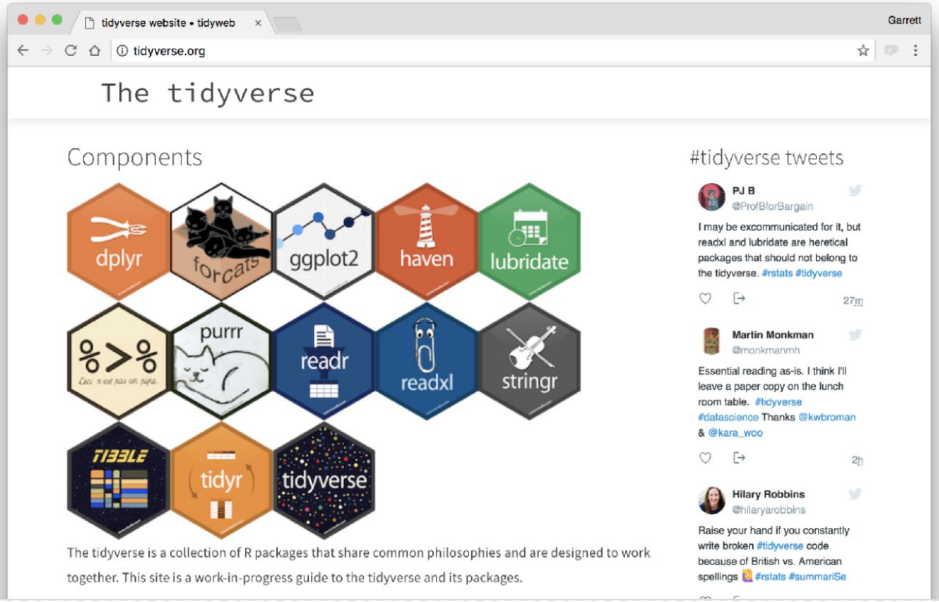
tidyverse

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



- **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

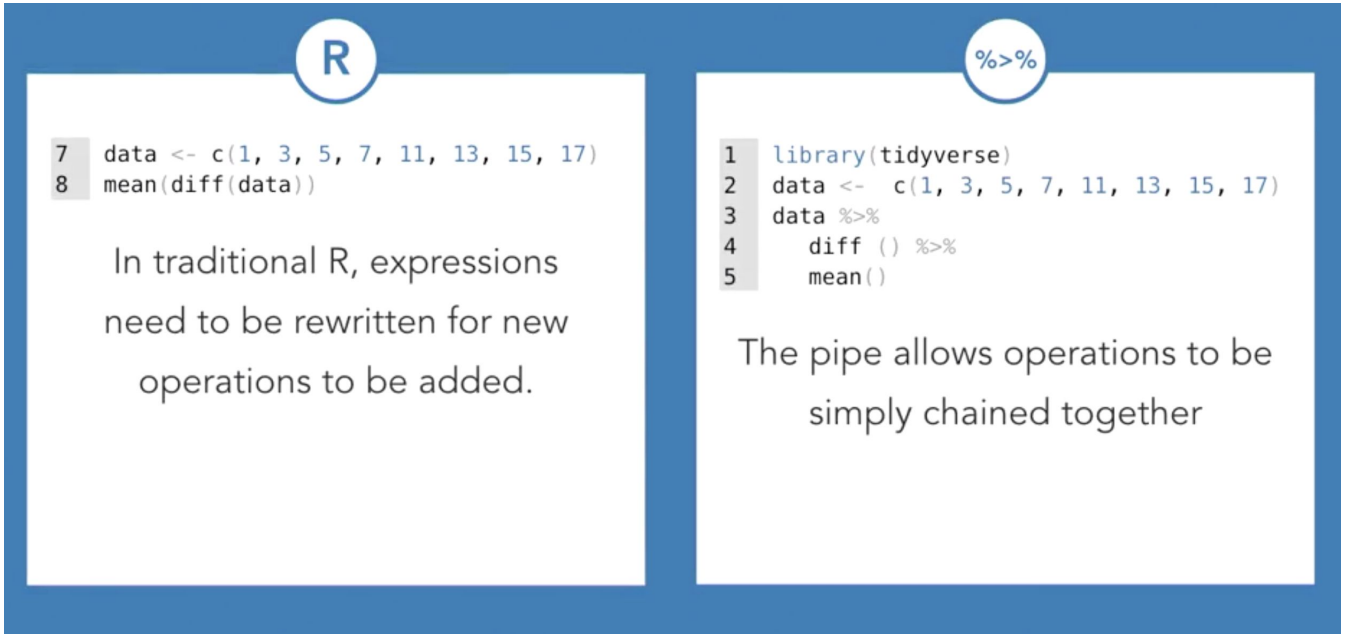
tidyverse: Data Visualization

- 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"