Intro to R for Data Science

Beginner's workshop

AbdulMajedRaja RS

About Me

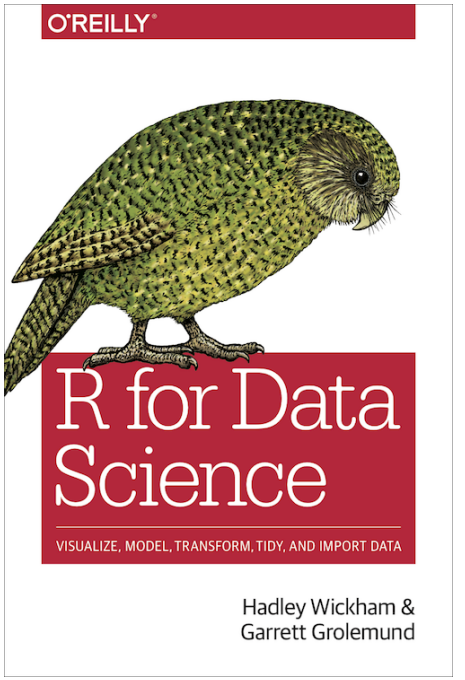
- Studied at
- Bengaluru R user group
- R Packages Developer (,)

Disclaimer:

- This workshop is going to make you a Data Scientist .
- The objective is to help you get a flavor of R and how it is used in Data Science
- Thus, get you ready to embark on your own journey to become a Data Scientist who uses R

Content:

This presentation's content is heavily borrowed from the book
by and



About R

- R is a language and environment for statistical computing and graphics. (Ref:)
- R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand
- R is Free
- R can be extended (easily) via .
- R is an interpreted language

R Interpreter / Console / GUI

Demo

About RStudio

- RStudio is a for R, released by the company

- RStudio and its team regularly contribute to R community by releasing new packages, such as:
  -
  -
  -

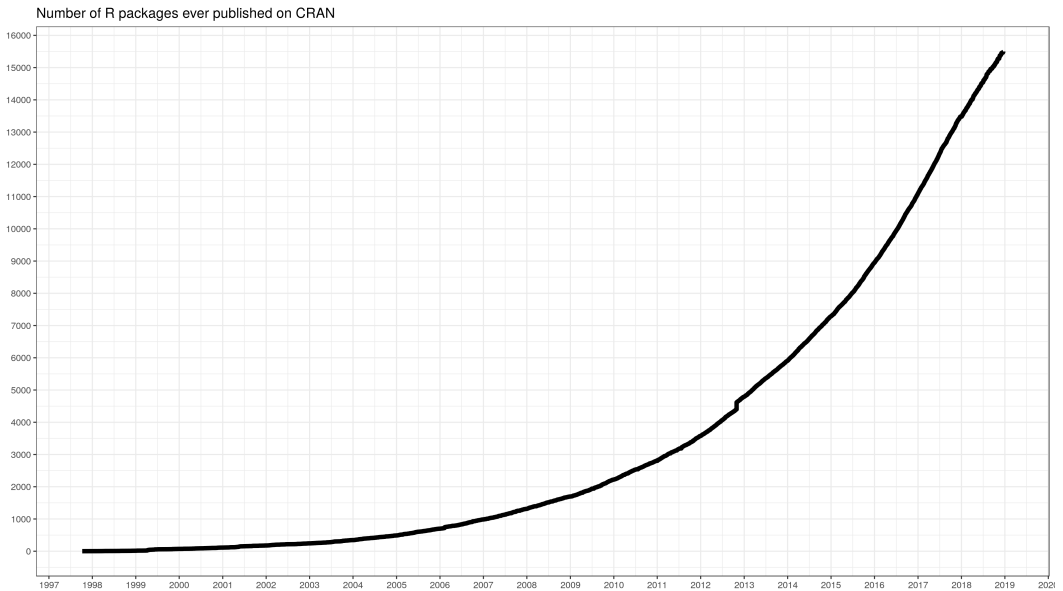
RStudio

Demo

R Ecosystem

Like , 's strength lies in its Ecosystem. - R Packages

Growth



Source: @daroczig

Basics of R Programming

Hello, World!

The traditional first step - :

Hello, World!

The traditional first step - :

That's one small step for a man, one giant leap for mankind

Neil Armstrong

Arithmetic Operations

Assignment Operators

Arrow (Less-than < and Minus -)

(Equal Sign)

Objects

- The entities R operates on are technically known as .

Example: Vector of numeric

Vectors

- Atomic Vectors - Homogeneous Data Type
    - logical
    - integer
    - double
    - character
    -
    -
- Lists - (Recursive Vectors) Heterogeneous Data Type
- is used to represent absence of a vector

- Factors are built on top of integer vectors.
- Dates and date-times are built on top of numeric vectors.
- Data frames and tibbles are built on top of lists.

Numeric Vector

Character Vector

Logical Vector

Coersion

Typecasting - Explicit

Typecasting - Implicit

Vector - Accessing

Vector Manipulation

Appending

Vector - Arithmetic

Factors

- In R, factors are used to work with categorical variables, variables that have a fixed and known set of possible values.
- Useful with Characters where non-Alphabetical Ordering is required

List

Lists are a step up in complexity from atomic vectors: each element can be any type, not just vectors.

List Accessing

Matrix

Dataframe

Tabular Structure

- dimension
- row.names
- col.names

Dataframe Manipulation

Loops & Iterators

For Loop

As you move forward, Check the family of functions - ,
, , .

For advanced functional programming, refer package

Logical Operations

%in% operator

Logical Operators

Conditions

Functions

Types

- Base-R functions (, ,)
- Package functions (,)
- User-defined functions

Packages

Package Installation & Loading

From CRAN (usually Stable Version)

Loading

Help

using

```

```

using ?

```

```

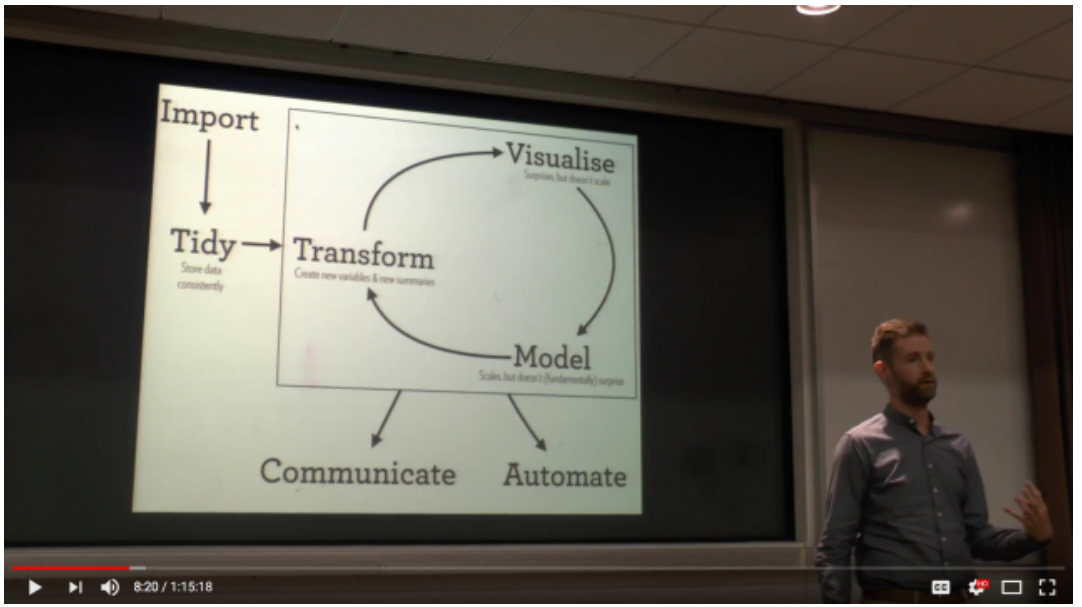
Help - Example

Packages Vignette

Data wrangling and Visualization using Tidyverse

Data Science Framework

There are now like, you know, a billion venn diagrams showing you what data science is. But to me I think the definition is pretty simple. Whenever you're struggling with data, trying to understand what's going on with data, whenever you're trying to turn that
. I think that's " - Hadley Wickham



Source: Hadley Wickham

Tidyverse

- An opinionated collection of R packages designed for data science.
- All packages share an underlying design

.

tidyverse packages

Loading the Library

Input Data

Reading the dataset

Basic Stats

Dimension (Rows Column)

Dataset Overview

Demo on RStudio

Data Questions (Business Problem)

- What's the percentage of Male and Female respondents?
- What are the top 5 countries?

dyplr verbs

- - adds new variables that are functions of existing variables
- - picks variables based on their names.
- - picks cases based on their values.
- - reduces multiple values down to a single summary.
- - changes the ordering of the rows.

Introducing %>% Pipe Operator

- The pipe, , comes from the magrittr package by Stefan Milton Bache
- is given as the

Example

```

```

Although doesn't make much sense to use in this context, Hope it explains the function.

Percentage of Male and Female

- Column name -

Pseudo-code

- the dataframe on column
- the values
- calculate value from the s

% of Male and Female - Group By & Count - Method 1

% of Male and Female - Group By & Count - Method 2

% of Male and Female - Group By & Count - Sorted

% of Male and Female - Percentage

% of Male and Female - Nice_Looking_Table

Female	4010	0.17
Male	19430	0.81
Prefer not to say	340	0.01
Prefer to self-describe	79	0.00

But, Wait!!!

Go Back and See

If you have only and ?

Time for some cleaning

In the form of ing

% of Male and Female - Filtered_Nice

Female	4010	0.17
Male	19430	0.83

An Awkward column name, isn't it??!

% of Male and Female - All_Nice_Table

Top 5 Countries

- Column name -

Pseudo-code

- number of respondents from each country
- countries in descending order based on their count value
- in the list is the output

Top 5 Countries - Code

United States of America	4716
India	4417
China	1644
Other	1036
Russia	879

Is a country name???

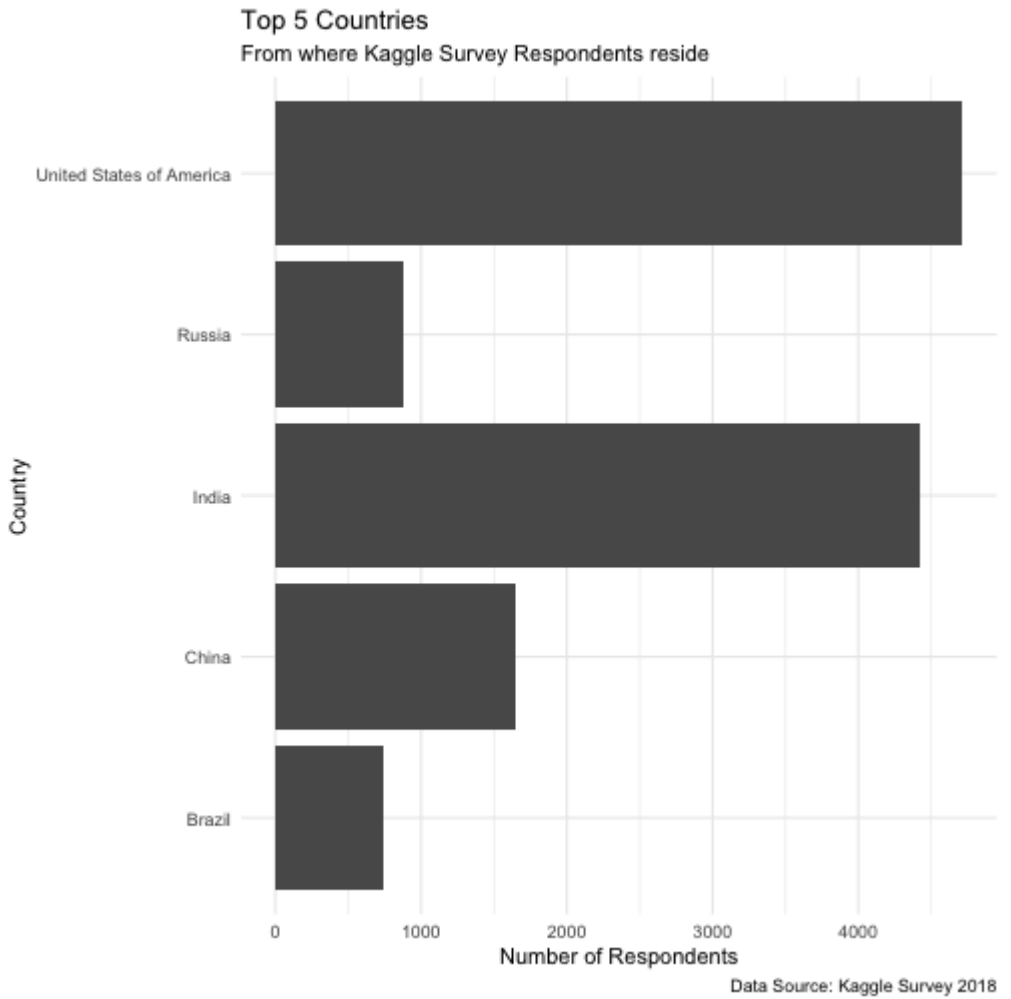
Top 5 Countries

United States of America	4716
India	4417
China	1644
Russia	879
Brazil	736

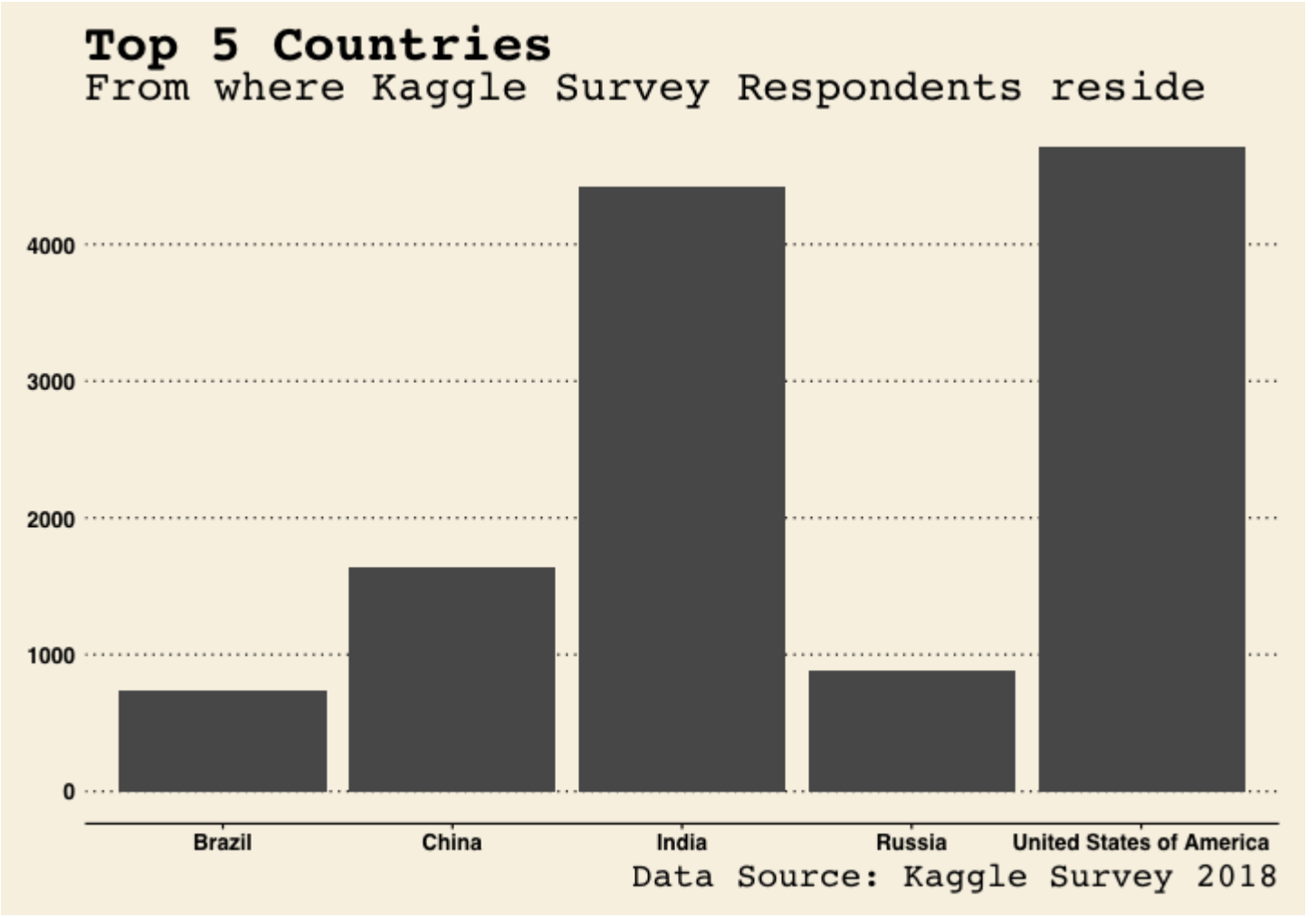
Table is nice, but a visually appealing plot is Nicer

Top 5 Countries - Plot #1

Top 5 Countries - Plot #2



Top 5 Countries - Plot #3 Themed



Documentation and Reporting using R Markdown

Demo

Project Demo

Object Detection in 3 Lines of R Code

using Tiny YOLO

-Project Demo-

References

- R for Data Science
- R-Bloggers

Thanks!

Slides created via the R package .

The chakra comes from remark.js, , and R Markdown.