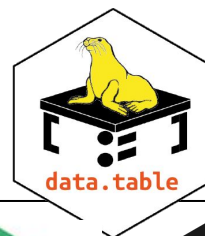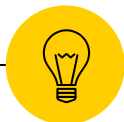# R Packages for Data Cleaning

R - Ladies Gaborone

25/01/2022

Overview

Who I am

Messy to tidy data

R Packages for Data Cleaning

    Janitor

    Naniar

    Amelia

    Data wizard

    data.table

Demo in RStudio

Simisani Ndaba

Member of R-Ladies Gaborone,

Teaching Assistant since 2016, Department of Computer Science, University of Botswana www.ub.bw
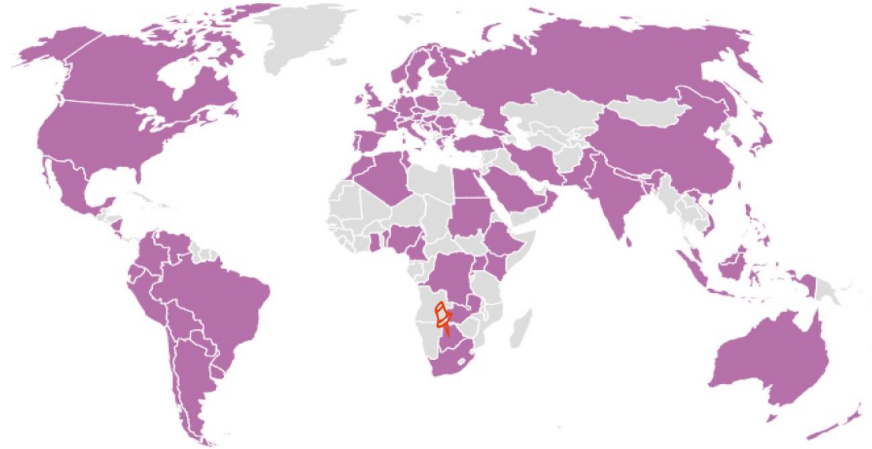
Msc in Computer Information Systems

@RLadiesGaborone/@simisani10

simisani.ndaba013@gmail.com

3

Where I'm from

DNA evidence points to this region as the origin of all modern humans, study says

From messy to tidy data

Data preparation is not objective

Depends on the problem you want to solve:
      Data organisation for facilitation
      Data analysis tools

Tasks for tidy data analysis
      Manipulation, visualisation and modelling

Data cleaning involves
      Exploration, visualisation and fixing errors (structural,outliers,NA,Validation
      rules)

TidyTuesday

A weekly data project in R from the R4DS online learning community

variables   observations   values

https://codata.org/initiatives/data-skills/codata-connect/webinar-series-research-skills-enhancement/webinar-4-importance-of-data-cleaning/

R Packages for Data Cleaning

janitor

naniar

amelia

datawizard

data.table

📌 Janitor_package

exploring and cleaning data

Functions: tabulations,duplications,

column names and formatting

The adorn_ functions dress up the results of these tabulations calls fast,

basic reporting

https://cran.r-project.org/web/packages/janitor/vignettes/janitor.html

Naniar_package

Naniar = NA in r

Developed by Nick Terney

A suite of tools for generating visualisation of missing values,

   imputations and summarising missing data

Data structures used to track missing data

https://naniar.njtierney.com/articles/getting-started-w-naniar.html

Multiple Imputation using amelia_package

imputation=replacement of missing values

Method:

- Joint Model=Gaussian/Normal Distribution
- Conditional Model=linear regression
- PCA(Principal Component Analysis)

Amelia_package

Joint Modeling

amelia() uses the

- bootstrap, resampling algorithm and
- EM(Expectation Maximum) algorithm, maximum likelihood based model

Data.table_package

Alternative version to data.frame
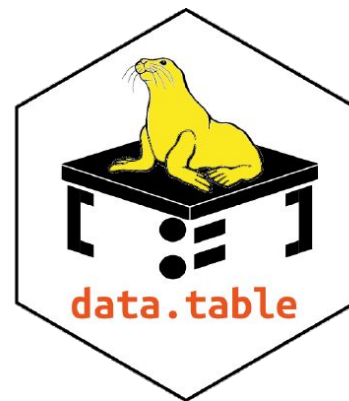
Used on tabular data

incredibly fast

DT syntax

$$DT[i , j, by]$$

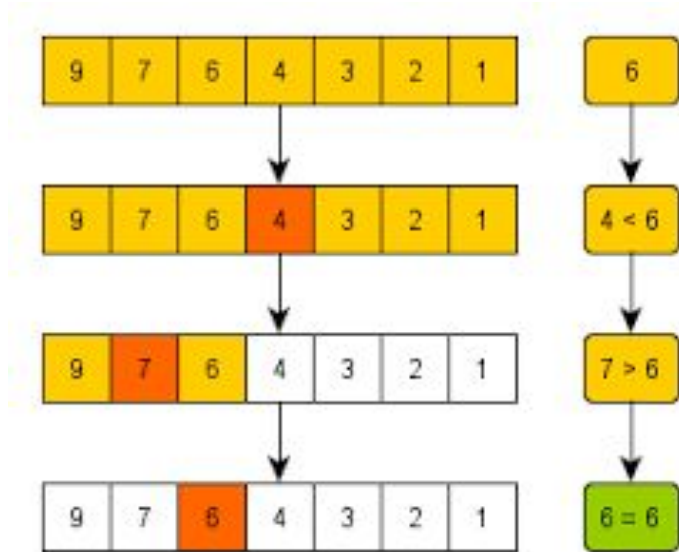   row, column, group by   /SQL

Group using by

Binary search using Keys(reference)

https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html

Binary Search....Data.table package cont..

Datawizard_package

A lightweight package to easily manipulate, clean, transform and prepare data for analysis

Data wrangling, transformation, visualisation, text formatting, data properties

https://easystats.github.io/datawizard/

📌 Demo in RStudio

```
install.packages(janitor)
install.packages(naniar)
install.packages(datawizard)
install.packages(data.table)

#extra functionality
install.packages(amelia)
install.packages(dplyr)
install.packages(readr)
```

Thanks!

https://www.meetup.com/rladies-gaborone/

YouTube

https://github.com/rladies/meetup_presentations_gaborone