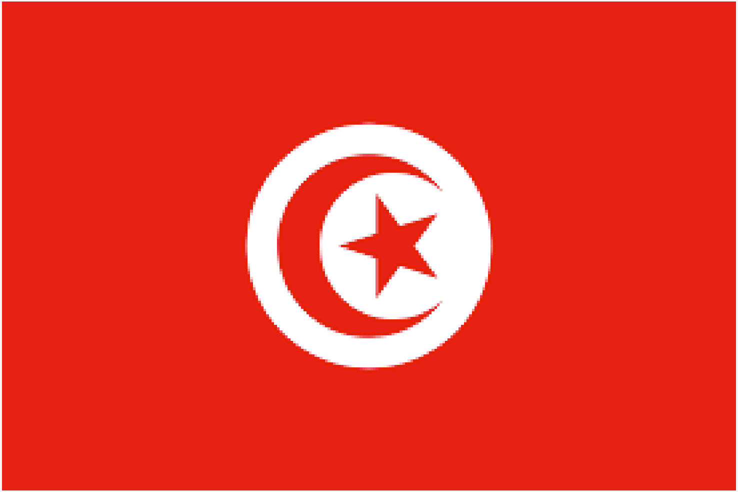
Data Science for Sports Studies

Dhafer Malouche

Yale University, USA
University of Carthage, Tunisia

Tsukuba University, Japan
March 2019

Tunisia?

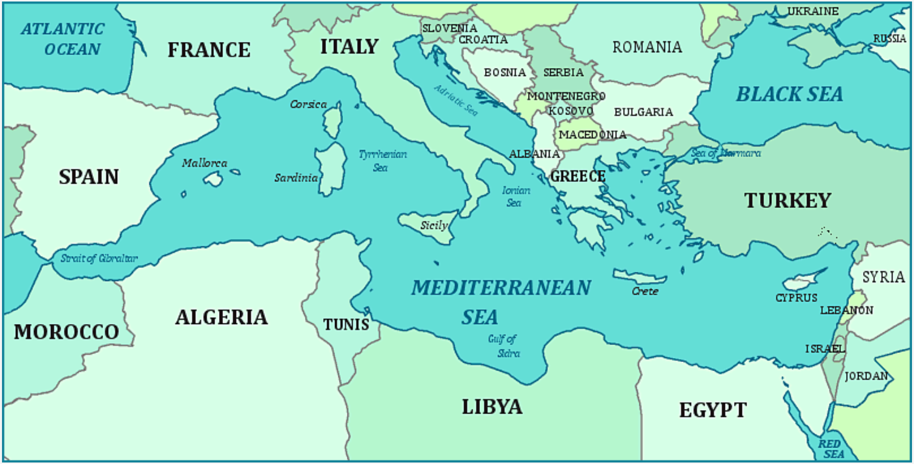


Tunisia?

Tunisia-Japan

Capitaland largest city	Tunis	Tokyo
Official languages	Arabic	
Spoken languages	Tunisian Arabic, Berber, French	Japanese
Ethnic groups	Arab-Berber 98%	98.5% Japanese
	Area	
Total	163,610 km2 (63,170 sq mi) (91st)	377,97 km2 (145,936 sq mi)(61st)
	Population	
2017 estimate	11,434,994 (79th)	126,440,000 (10th)
Density	63/km2 (163.2/sq mi) (133rd)	334/km2 (865.1/sq mi) (41st)
	GDP (PPP), 2018	
Total	$144.222 billion	$5.632 trillion (4th)
Per capita	$12,369	$44,550 (31st)
	GDP (nominal), 2018	
Total	$41.662 billion	$5.071 trillion[13] (3rd)
Per capita	$3,573	$40,106 (26th)
Gini (2017)	35.8 medium	37.9 medium 76th
HDI (2017)	0.735, high 95th	0.909, very high 19th

History


ATLANTIC OCEAN
FRANCE
ITALY
SLOVENIA
CROATIA
UKRAINE
RUSSIA
ROMANIA
BOSNIA
SERBIA
MONTENEGRO
KOSOVO
BULGARIA
MACEDONIA
BLACK SEA
Adriatic Sea
Corsica
SPAIN
Mallorca
Sardinia
Tyrrhenian Sea
ALBANIA
GREECE
Sea of Marmara
TURKEY
Ionian Sea
Sicily
Strait of Gibraltar
MEDITERRANEAN SEA
Crete
SYRIA
CYPRUS
LEBANON
MOROCCO
ALGERIA
TUNIS
Gulf of Sidra
ISRAEL
JORDAN
LIBYA
EGYPT
RED SEA

History

History

History

History

Who I am?

- Data Scientist and Statistician
- Teaching Statistics and several Data Science topics from more than 15 years
- PhD of Applied Mathematics and Statistics from Toulouse University, France

https://dhafermalouche.net

My Academic Position: Professor of Statistics

جامعة قرطاج
Université de Carthage

ESSAI

- Bayesian Statistics
- Time Series
- Big Data
- Advanced R/Python
- ...

Data Science

Lecturer at Yale University, July 2019

Lecturer at Yale University, July 2019

- Political and Social Scientists
- Climate Change
- Data Science, Survey Methodology . . .

What’s Data Science?

- It’s a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms.

- It unifies statistics, data analysis, machine learning and their related methods in order to understand and analyze actual phenomena with data.

To be a Data Scientist, you need

- Good mathematics, statistics and probability theory background
- Good knowledge (advanced) of Software like R and Python
- Big Data tools: Hadoop, Spark, H2o

Data Scientist work closely with other researchers

- Collect Data for his research:
  - Surveys or Experimental Design
  - Web scraping, web harvesting, and web data extraction
- Learning Patterns from Data: Machine Learning and Artificial Intelligence techniques and algorithms
- Build Software and Applications for Predictions, Visualizations or Learning patterns for future Data and Observations

Collecting Data: Surveys

- Face 2 Face
- Phone
- Internet (Survey Monkey, Lime Survey)

Collecting Data: Surveys, Sample Design

- Non-random and random sampling, quota sampling, simple random sampling

- Probability Proportional to Size (PPS) method

- Dealing with Sampling Problems: response rate, missing data, estimation of the sample-size

Collecting Data: Surveys, Questionnaire

- Use tablets, No more papers,
- Need to be online: Google forms (Small Questionnaires, KwikSurveys, LimeSurvey, Qualtrics...
- Work offline:
  - CSPro: https://www.census.gov/data/software/
  - SurveyToGo: https://www.dooblo.net/downloads/

Collecting Data: Surveys, Visualizing and Reports

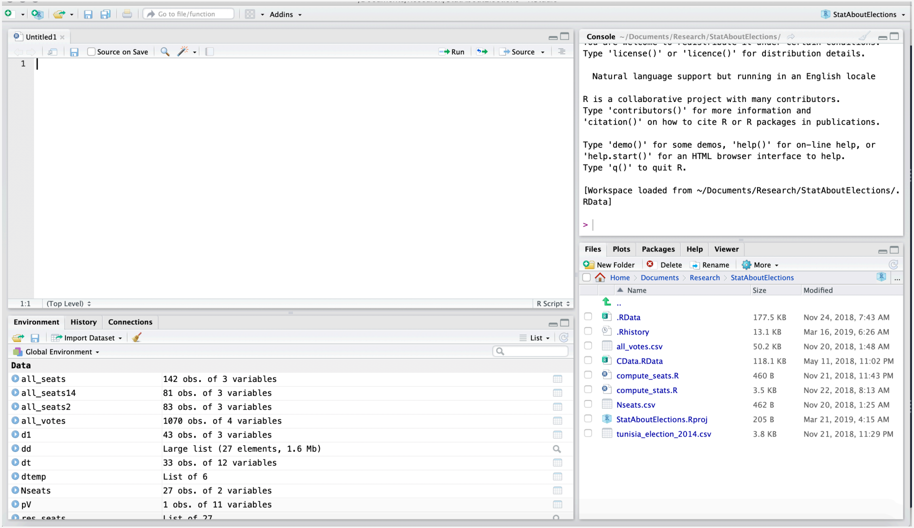
RStudio

https://www.rstudio.com

100 Free tutorials for learning R

https://www.listendata.com/p/r-programming-tutorials.html

Collecting Data: Surveys, Visualizing and Reports



```
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/Documents/Research/StatAboutElections/.
RData]

>
```

Data

all_seats	142 obs. of 3 variables
all_seats14	81 obs. of 3 variables
all_seats2	83 obs. of 3 variables
all_votes	1070 obs. of 4 variables
d1	43 obs. of 3 variables
dd	Large list (27 elements, 1.6 Mb)
dt	33 obs. of 12 variables
dtemp	List of 6
Nseats	27 obs. of 2 variables
pV	1 obs. of 11 variables
res_seats	List of 27

Name	Size	Modified
..		
.RData	177.5 KB	Nov 24, 2018, 7:43 AM
.Rhistory	13.1 KB	Mar 16, 2019, 6:26 AM
all_votes.csv	50.2 KB	Nov 20, 2018, 1:48 AM
CData.RData	118.1 KB	May 11, 2018, 11:02 PM
compute_seats.R	460 B	Nov 21, 2018, 11:43 PM
compute_stats.R	3.5 KB	Nov 22, 2018, 8:13 AM
Nseats.csv	462 B	Nov 20, 2018, 1:25 AM
StatAboutElections.Rproj	205 B	Mar 21, 2019, 4:15 AM
tunisia_election_2014.csv	3.8 KB	Nov 21, 2018, 11:29 PM

Collecting Data: Surveys, Visualizing and Reports

- `ggplot2` for data visualization

  https://malouche.github.io/slidesOftalks/index.htmldata-visualization

- `survey` Summary statistics, two-sample tests, rank tests, generalised linear models, cumula- tive link models, Cox models, loglinear models...

  http://r-survey.r-forge.r-project.org/survey/

- `sjPlot`

  http://www.strengejacke.de/sjPlot/

Collecting Data: Web Scraping

- Data from Wikipedia:
  - `WikidataR` This package serves as an API client for https://www.wikidata.org.
  - `WikipediaR`: Provides an interface to the Wikipedia web API.
- OpenData Website:
  - `knoema` This package works with datasets from knoema.com
  - `WDI` World Bank Development Indicators Data (800 Indicators from 1960)

Collecting Data: Web Scraping, `rvest`

- This package is useful in extracting the information you need from web pages.
- Some tutorials about `rvest`
  - https://www.analyticsvidhya.com/blog/2017/03/beginners-guide-on-web-scraping-in-r-using-rvest-with-hands-on-knowledge/
  - https://towardsdatascience.com/web-scraping-tutorial-in-r-5e71fd107f32

Collecting Data: Qualitative Data with R

- RQDA package to analyse interviews and for Qualitative Data Analysis

- Tutorials
  - http://rqda.r-forge.r-project.org
  - https://www.r-bloggers.com/qualitative-data-science-using-rqda-to-analyse-interviews/

Reporting with R

- Rmarkdown
- Shiny Interactive Data Visualization

  https://shiny.rstudio.com/gallery/

  https://github.com/rstudio/shiny-examples
- flexdashboard Dashboards with R

  https://rmarkdown.rstudio.com/flexdashboard/examples.html

Python

ANACONDA® CLOUD

https://dhafermalouche.net

Thank you!