معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

A Gentle Introduction to R

QCRI P2P Training Series

Francisco (Paco) Guzmán
Scientist
Arabic Language Technologies

Available at: https://github.com/guzmanhe/rtutorial2016

Before we start

- Do you have R installed?
- Do you have R studio installed?
- Have you downloaded the course material?

60% of the content is not on these slides

https://github.com/guzmanhe/rtutorial2016

https://github.com/guzmanhe/rtutorial2016/archive/master.zip

Data Science

Statistics
Machine Learning
Data Cleaning
Data Analysis
Data Mining
Pattern Recognition
Visualization

What is **R** ?

R is a language and environment for statistical computing and graphics.

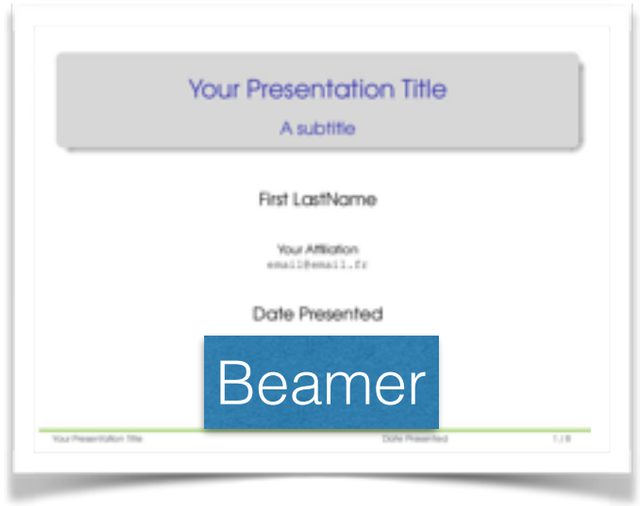
R is the preferred tool of many data scientists

An analogy

Word processing

LaTeX

Presentations



Data Analysis

What can you do with **R** ?

Basics

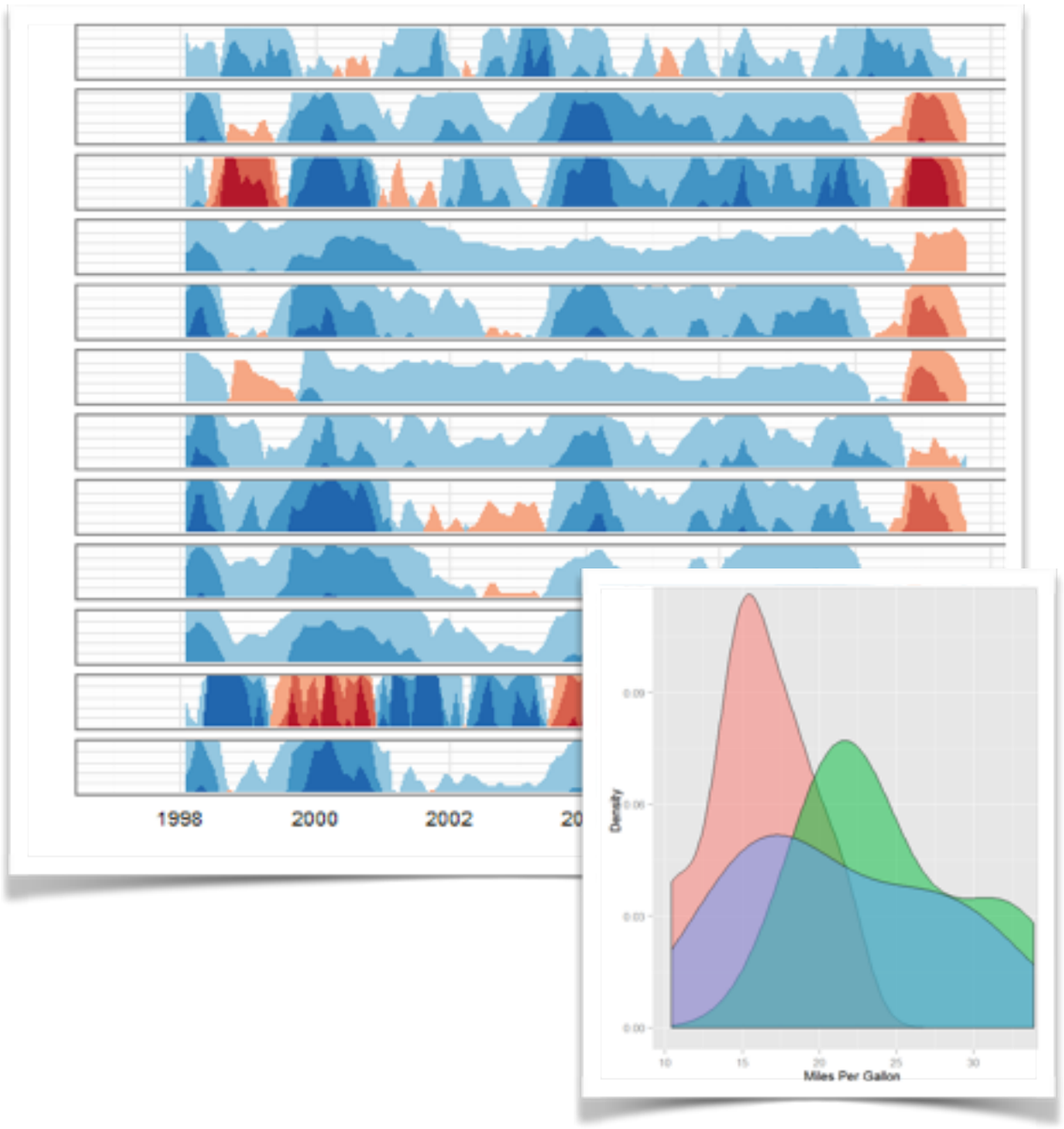
- Data cleaning
- Data pre-processing
- Data summarization

Advanced

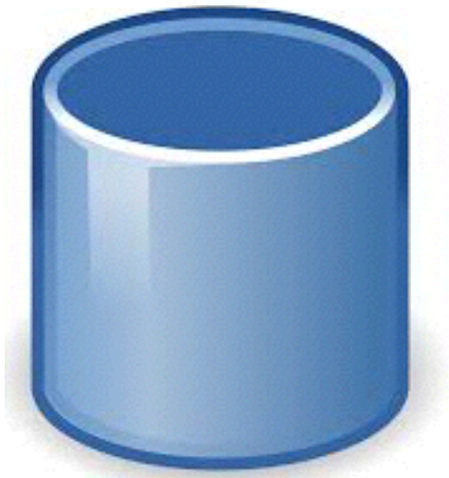
- Statistical analysis
- Visualization
- Machine learning

Advantages of R

- Works anywhere (portable)
- Can be extended easily
- Has large community
- Looks great
- Enhances reproducibility
- Is free!



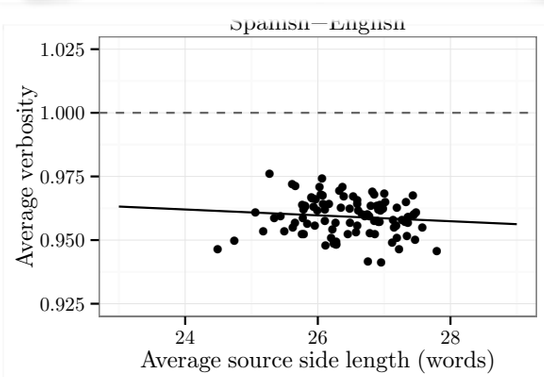
This tutorial



data

paper

tuning	Arabic-English (multi-r short	mid	long
MERT			
short	**47.26***	50.71	50.8
mid	46.53	**51.11***	51.3
long	46.23	50.84	**51.7**
max-min	1.04	0.40	0.9
loss if using *closest*	0.00	0.00	0.0
PRO-fix			
short	**46.74**	50.57	50.9



average verbosity (y axis) for 100 random samples, (Left: Arabic-English, multi- and single-reference) rence) data.

4 Experiments and Evaluation

We experimented with single-reference and multi-reference tuning and testing datasets for two language pairs: Spanish-English and Arabic-

This tutorial

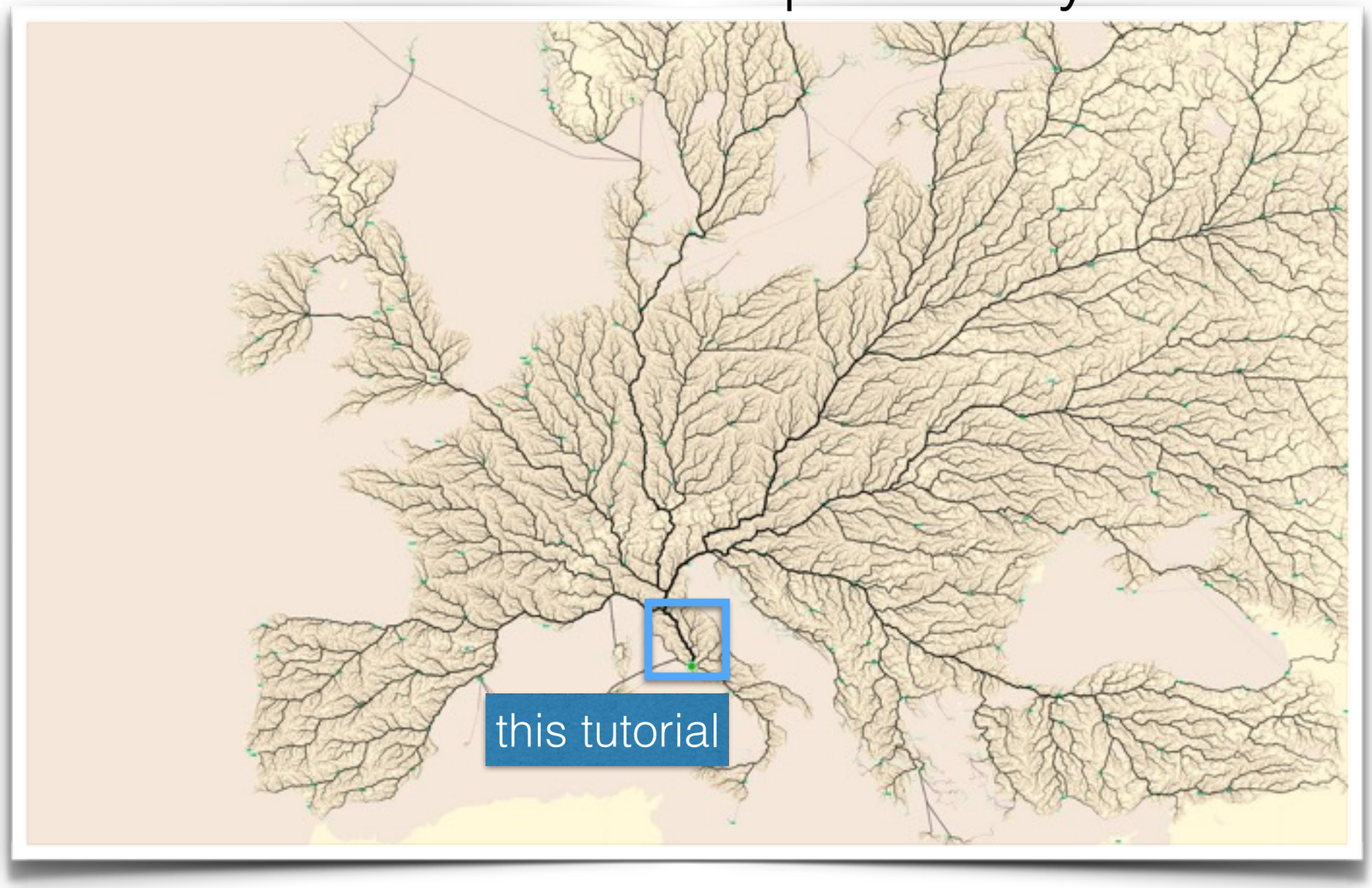
- A brief introduction to R
- Four hands-on recipes to learn to:
  - load data
  - summarize data
  - plot data
  - export graphs, tables

Just an intro

aka. appetizer

Our goal

You shall be able to explore on your own

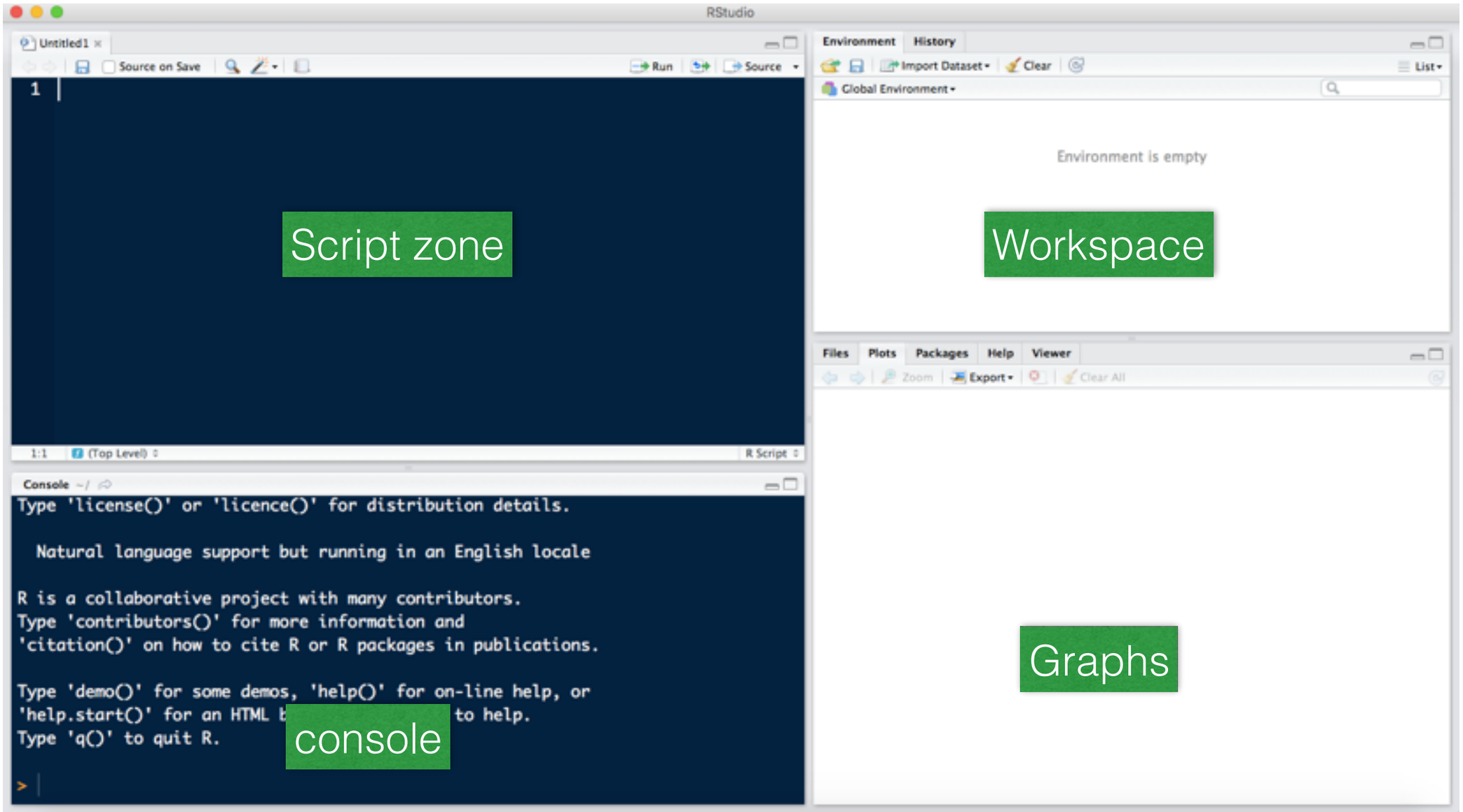


“Roads to Rome”, Benedikt Groß and Philipp Schmitt. Moovel Lab

1 2 3

R Basics

R Studio


RStudio
Untitled1
Source on Save
Run
Source
1
Script zone
1:1
(Top Level)
R Script
Console ~/
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 b
to help.
Type 'q()' to quit R.
>
console
Environment
History
Import Dataset
Clear
List
Global Environment
Environment is empty
Workspace
Files
Plots
Packages
Help
Viewer
Zoom
Export
Clear All
Graphs

R Basics

- Basic operations
  - Assignment
  - Arithmetics
- Data types
- Simple graphs

Basic ops

- Assignment
- Arithmetics

```
#numeric
my_num1<- 15
```

```
#character
my_string<- "Hello world"
```

```
#logical
my_bool<- TRUE
```

Data types

- Scalars
  - Numeric
  - Characters
  - Logical
- Other types
  - Vectors
  - Factors
  - Data Frames

Common functions

print(X**)**	prints in console the value of a variable
class(X)	gives you the data type of a variable
str(X)	tells you about the variable(s) class, size and values
summary(X)	gives you a summary of a variable. Specially useful for lists/vectors
c(x1,x2,…)	creates a vector by **combining** the given elements

Factors

- They represent categories
  - **level:** a numerical value
  - **label:** string representation

```
str(f_affiliations)
```

```
>>>  Factor w/ 2 levels "QF","QP": 1 2 2 1
```

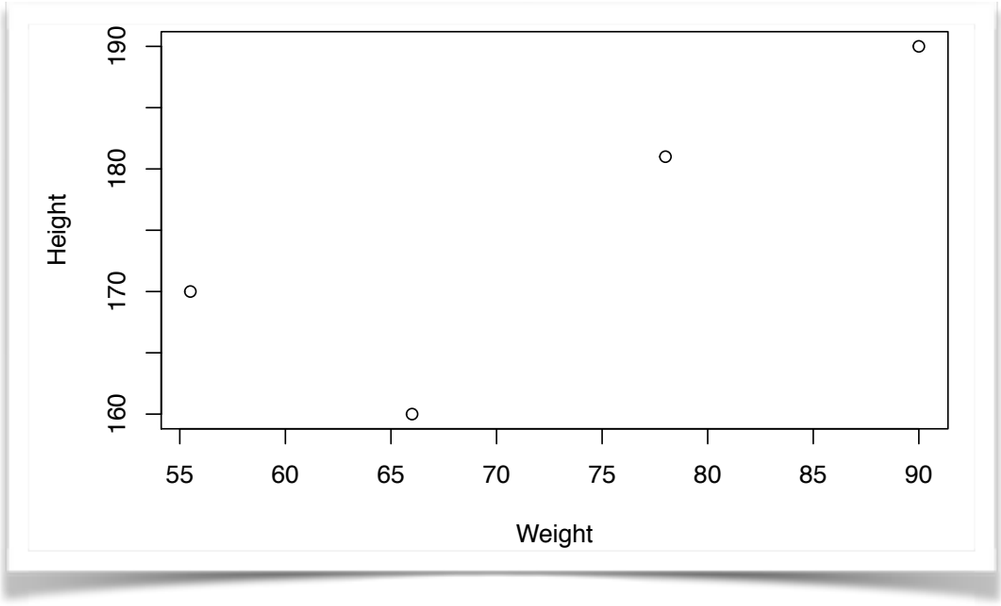
Data Frames

DFs are collections of variables (columns) with different measurements(rows)

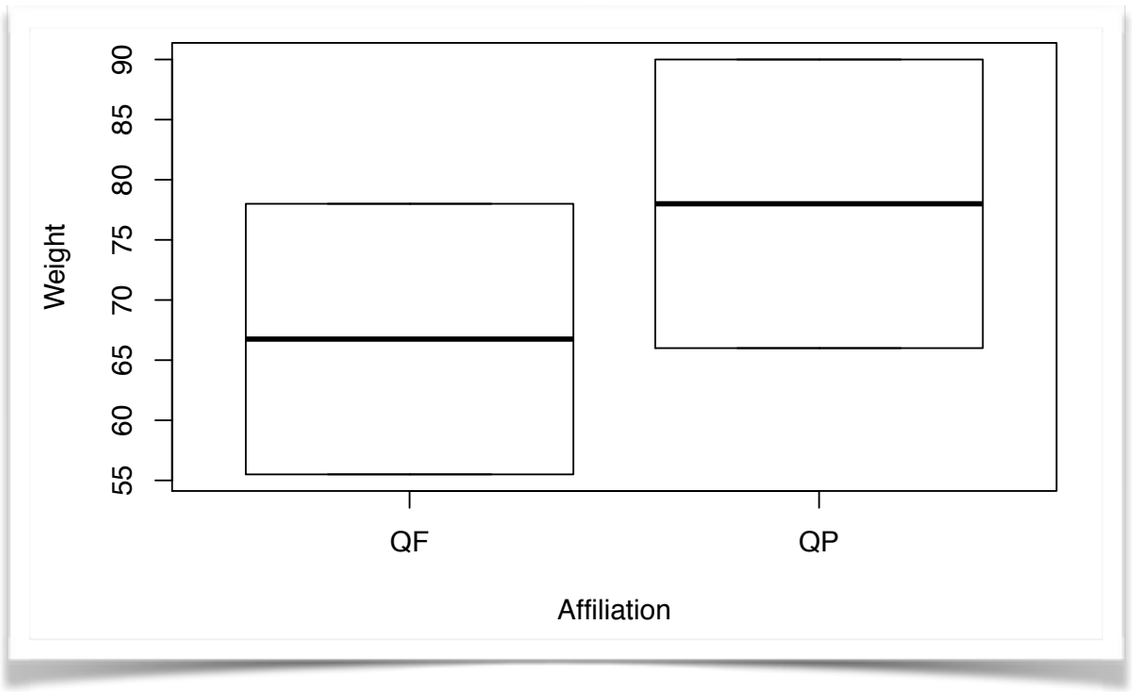
```
>>>   name affiliation weight height is_fit
>>> 1 John           QF   78.0    181   TRUE
>>> 2 Mary           QP   66.0    160  FALSE
>>> 3  Bob           QP   90.0    190   TRUE
>>> 4 Anna           QF   55.5    170  FALSE
```

Basic plots

- Num vs Num



- Num vs Factor



Exercise

- Add **Salma** and **Khaled** to your data frame
  - "Salma" (weight=61, height=166, is_fit=TRUE, affiliation="QSTP") and
  - "Khaled" (weight=75, height=175, is_fit=FALSE, affiliation="QSTP")
- Plot height vs. weight again


Speckmantel
Bacon, Reis,
Tomaten,
Pfeffer,
schneiden, etwas salzen,
wickeln. In der Pfanne gut
dann in den Backofen
20-30 Minuten.
in Scheiben schnei
cho
bunt, in Streife
ße Zwiebe

Recipes

Recipe 1

In this recipe, you'll learn:

- How to obtain data **summaries** (average, standard dev, median)
- How to create a basic **histogram**
- How to create a basic annotation **line**
- How to create a basic **scatter plot**

New commands

data(mtcars)	imports default dataset **mtcars** from R
help(command)	provides info about *command*
head(df)	shows the first lines in a *df*
mean(X)	calculates average
sd(X)	calculates standard deviation
summary(df)	gives a summary with max, min, median, quartiles for the *df*
hist(X, breaks)	plots an histogram for variable X breaks**:**number of breaks (bars) xlab: x-axis title main: graph title

New commands (2)

hist(X, breaks)	plots an histogram for variable X breaks:number of breaks (bars) xlab: x-axis title main: graph title
abline(v=X)	draws a simple vertical line at X (or horizontal line with h=Y) col=color of line (e.g. red, black) lty= **l**ine s**ty**le (e.g. 1=solid, 2=dash)
cor(X,Y)	calculates correlation between two vectors

Exercise

- Calculate summaries for HP and wt
- Plot histogram of HP
- Plot scatterplot of HP vs MPG and calculate correlation

Recipe 2

In this recipe, you'll learn:

- Install a new package
- Load a package
- Get average and summary statistics by group
- Dump a latex table

New commands

install.package(pkg)	downloads and installs package **pkg**
library(pkg)	loads contents of library
ddply(df,groups,type, results)	slices data by groups and applies summarizing functions to data frame groups= which factors to use for grouping type= we use **summarize** results= variables and summarizing functions
xtable(df,caption,label)	generates a latex table with the data in a data frame

Break 10 min

Recipe 3

In this recipe, you'll learn:

- How to install packages (reinforce)
- How to load data from a file
- How to plot nice scatter plots with ggplot
- How to save your graph as a jpg and pdf (command line)

New commands

read.table(file)	imports data from a file header=True/False. Tells if the file has a header col.names=vector with the name of the columns sep=column separator
pdf(file)	starts saving graphical output to pdf device width= control width aspect height= control height aspect
dev.off()	closes pdf file and goes back to old graphics device

New commands

ggplot(df)	Plots different types of graphs aes= sets the aesthetics of the graph e.g. x=variable to be used in x axis col=variable to be used as color
+ geom_point()	adds a layer of points to the plot
+ geom_smooth()	adds a trend line
+xlim(), +ylim()	adds the range for the x or y axes
+xlab(),+ylab()	adds the x and y axes titles
+ggtitle()	adds a main title to the graph
+scale_color_manual	allows to define the colors to be used, manually

Recipe 4

In this recipe, you'll learn:

- How to load data from a text file (reinforce)
- How to create summaries with ddply (reinforce)
- How to plot bar plots with ggplot
- How to plot different views using facets
- How to save your graph as a pdf (reinforce)

New commands

+geom_bar()	adds a bar plot layer to ggplot stat: “identity” to use variable values instead of counts position: “dodge” to un-stack different groups of bars and put them side by side
+facet_grid()	creates a graph with different views by factors
+scale_fill_manual	allows to define the fill colors to be used, manually

Wrapping up

What we've covered

- A brief introduction to R
- Four hands-on recipes to learn to:
  - load data
  - summarize data
  - plot data
  - export graphs, tables

Additional resources

- Quick-R reference guide:
  http://www.statmethods.net

- JHU Coursera Data Science specialization:
  https://www.coursera.org/specializations/jhu-data-science

- Data Camp R programming:
  https://www.datacamp.com/courses?learn=r_programming

- ggplot cheat sheet:
  https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

- Cross-validated:
  http://stats.stackexchange.com

- Stack overflow

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

END

Questions?: QCRI-Rtutorial-2016@googlegroups.com

Material available at: https://github.com/guzmanhe/rtutorial2016