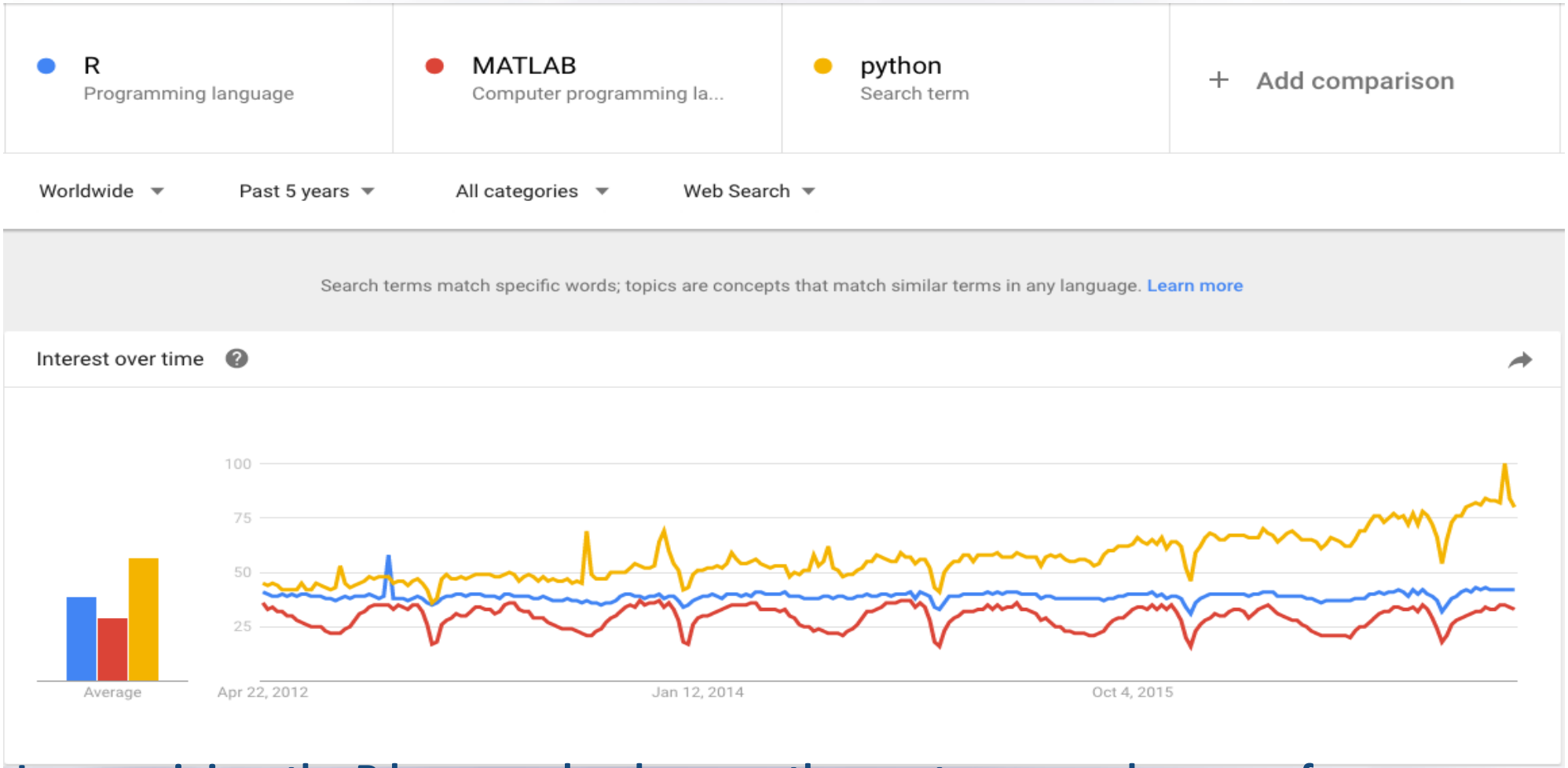
Data Analytic Methods Using R



In my opinion, the R language has become the most common language for communication in the fields of Statistics and and Data Analysis.

Data Analytic Methods Using R

Use basic analytics methods such as distributions, statistical tests and summary operations to investigate a data set.

Use R as a tool to perform basic data analytics, reporting and basic data visualization.

Data Analytic Methods Using R

- Putting the Data Analytics Lifecycle into Practice
- Use a strategy to approach any data analytics problem:
  - Discovery and Data Preparation
  - Model Building and Evaluation
- To begin to analyze the data you need:
  - 1. A tool that allows you to look at the data – that is “R”.
  - 2. Skill in basic statistics.

Data Analytic Methods Using R

Introduction to R

Using the R Graphical User Interface

Getting Data into (and out of) R

Data Types Used in R

Basic R Operations

Basic Statistics

Introduction to R

- open-source: R is a free software environment for statistical computing and graphics.
- oers tools to manage and analyze data.
- standard and many more statistical methods are implemented.
- possibility to write personalized code and to contribute new packages.

From: www.r-project.org

JUPYTER

R vs. SAS vs. Julia vs. Python

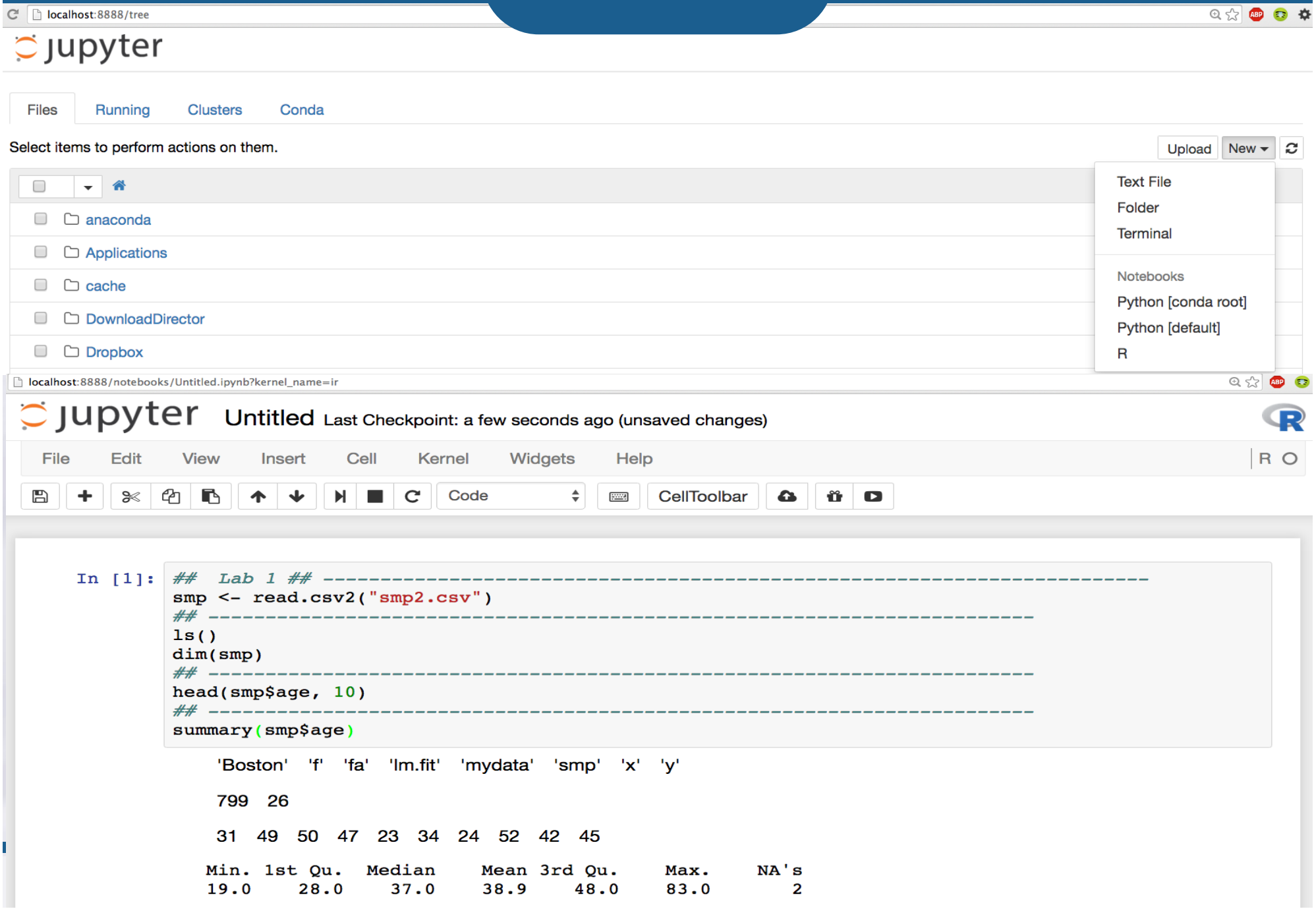
R is open source, SAS is a commercial product, Julia a very new dynamic programming language, ...

- **R** is free and available to everyone
- **R** code is open source and can be modified by everyone
- **R** is a complete and enclosed programming language
- **R** has a **big and active community**

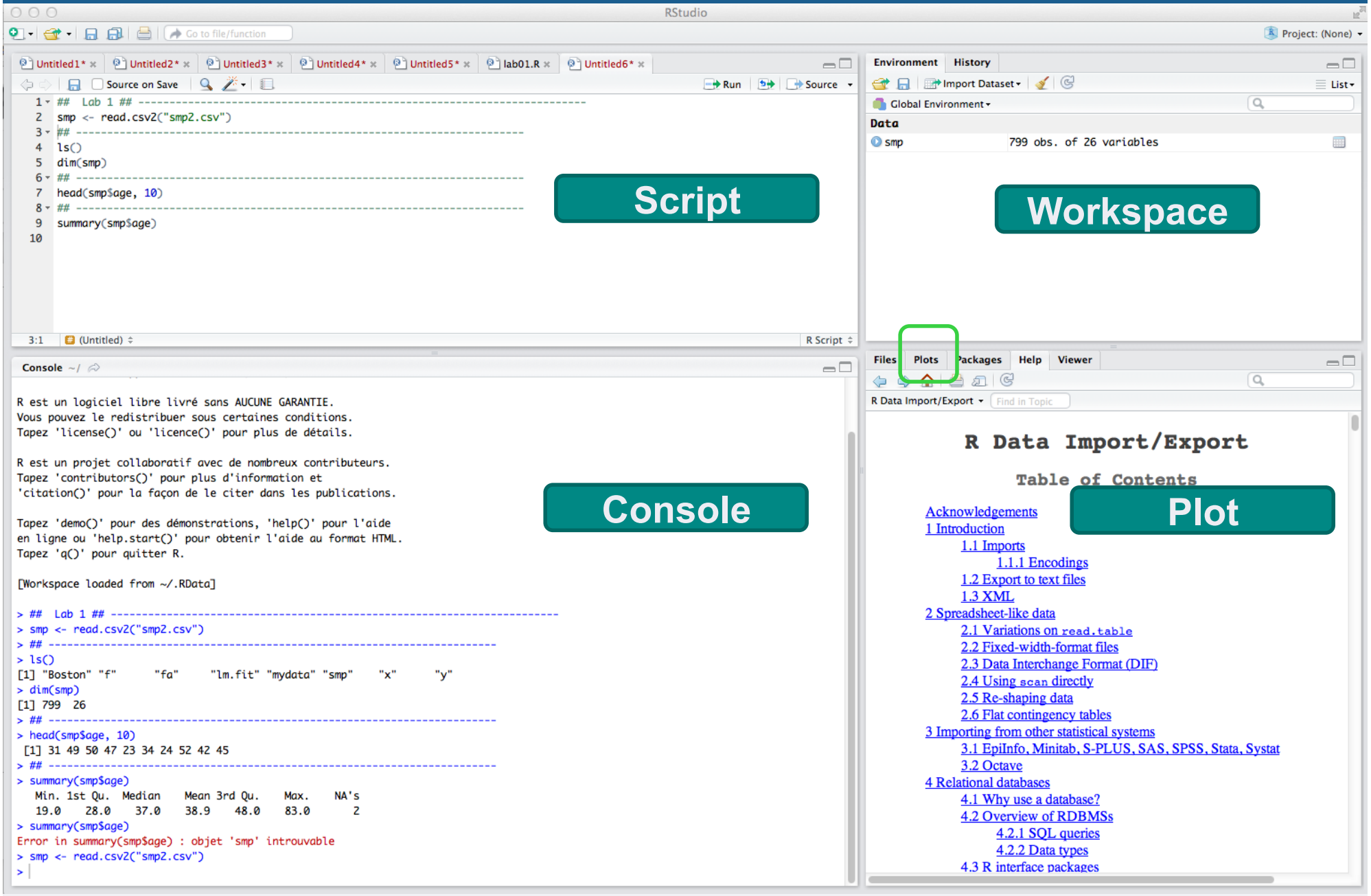
⇔

⇔

JUPYTER


localhost:8888/tree
jupyter
Files
Running
Clusters
Conda
Select items to perform actions on them.
Upload
New
Text File
Folder
Terminal
Notebooks
Python [conda root]
Python [default]
R
anaconda
Applications
cache
DownloadDirector
Dropbox
localhost:8888/notebooks/Untitled.ipynb?kernel_name=ir
jupyter
Untitled
Last Checkpoint: a few seconds ago (unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Code
CellToolbar
In [1]:
Lab 1 ## ----------
smp <- read.csv2("smp2.csv")
----
ls()
dim(smp)
----
head(smp$age, 10)
----
summary(smp$age)
'Boston' 'f' 'fa' 'lm.fit' 'mydata' 'smp' 'x' 'y'
799 26
31 49 50 47 23 34 24 52 42 45
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
19.0 28.0 37.0 38.9 48.0 83.0 2

Using the RStudio Graphical User Interface



Overview: Getting Data Into (and Out of) R

Getting Data Into R

- Type it in (if it's small)!
- Read from a data file
- Read from a database

Getting Data Out of R

- Save in a workspace
- Write a text file
- Save an object to the file system
- You can save plots as well!

Getting Data Into R: External Sources

Getting and Setting the Working Directory

```
# Get and print current working directory.
print(getwd())

# Set current working directory.
setwd("/desktop/stats")

# Get and print current working directory.
print(getwd())
```

Getting Data Into R: External Sources

```
# Create a data file: input.csv
data <- read.csv("input.csv")
retval <- subset( data, dept == "IT")
```

R supports multiple file formats

```
id,name,salary,start_date,dept
1,Rick,623.3,2012-01-01,IT
2,Dan,515.2,2013-09-23,Operations
3,Michelle,611,2014-11-15,IT
4,Ryan,729,2014-05-11,HR
5,Gary,843.25,2015-03-27,Finance
6,Nina,578,2013-05-21,IT
7,Simon,632.8,2013-07-30,Operations
8,Guru,722.5,2014-06-17,Finance
```

```
# Write filtered data into a new file.
write.csv(retval,"output.csv")
```

Getting Data Into R: External Sources

R supports multiple file formats

File name can be a URL

```
file <- read.table("http://www.data.org/file.csv", sep=",")
```

is the same as `read.csv(…)`

.TXT and .XLSX

```
car <- read.csv("http://win-vector.com/dfiles/car.data.csv")
```

Can read directly from a databas via ODBC interface

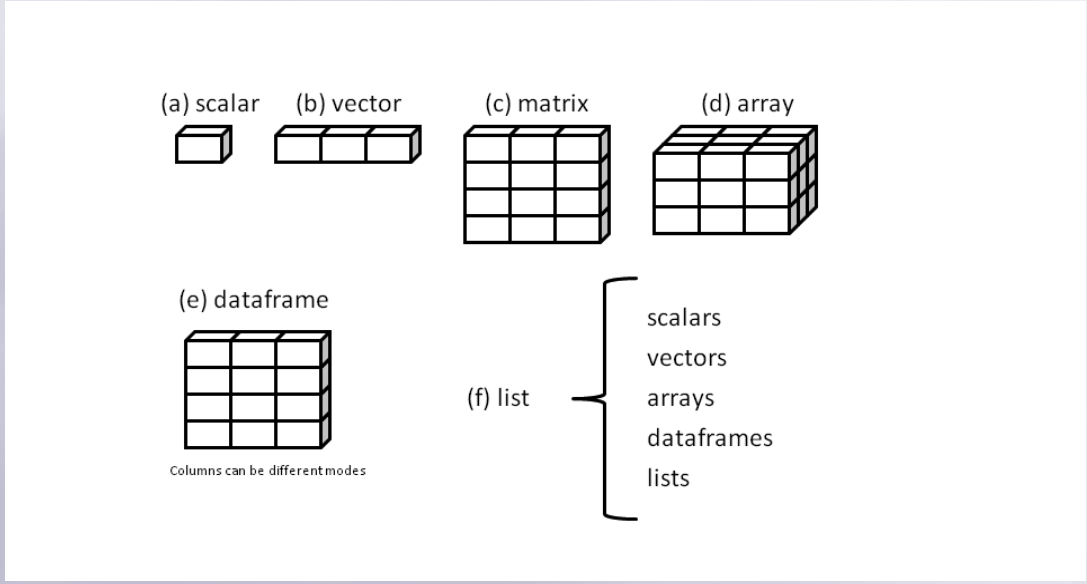
- `library(RODBC)`
- `Db <- odbcConnect("world", uid="root", pwd="")`

```
data<-read.csv(file.choose(),header = T)
# file.choose() used to point that file
```

Data Types Used in R

in R, the variables are not declared as some data type.
The variables are assigned with R-Objects.
The data type of the R-object becomes the data type of the variable.
There are many types of R-objects. The frequently used ones are:

- ✓ Scalar
- ✓ Vector
- ✓ Matrice
- ✓ Array
- ✓ Data Frame
- ✓ List



Data Types Used in R

```
# Create data type.
v <- 23.5
print(class(v))
```

"numeric"

```
v <- TRUE
print(class(v))
```

"logical"

```
rep (1,10)            # repeats the number 1 : 10 times
seq(0,20, by=4)    # sequence of integers between 0 and 20 by step =4

# Create a vector.
apple <- c('red','green',"yellow")
print(apple)
```

"red" "green" "yellow"

```
# Get the class of the vector.
print(class(apple))
```

"character"

```
# Create a list.
list1 <- list(c(2,5,3),21.3,sin)
print(list1)
print(class(list1))
```

" list "

Data Types Used in R

```
# Create an array.
a <- array(c('green','yellow'),dim = c(3,3))
print(a)
print(class(a))
```

” array "

```
# Create a matrix.
M = matrix( c(1,2,80,85,67,56), nrow = 2, ncol = 3)
print(M)
print(class(M))
```

"matrix"

```
# Create a vector.
apple_colors <- c('green','green','yellow','red','red','red','green’)
# Create a factor object.
factor_apple <- factor(apple_colors)
# Print the factor.
print(factor_apple)
print(nlevels(factor_apple))
```

Levels: green red yellow

```
# Create the data frame.
BMI <-     data.frame(gender = c("Male", "Male","Female"),
 height = c(152, 171.5, 165),
 weight = c(81,93, 78),
Age = c(42,38,26))
```

```
print(BMI)
   gender  height weight  Age
1    Male   152.0     81    42
2    Male   171.5     93    38
3 Female   165.0     78    26
```

```
print(class(BMI))
```

"data.frame"

Analyzing and Exploring the Data

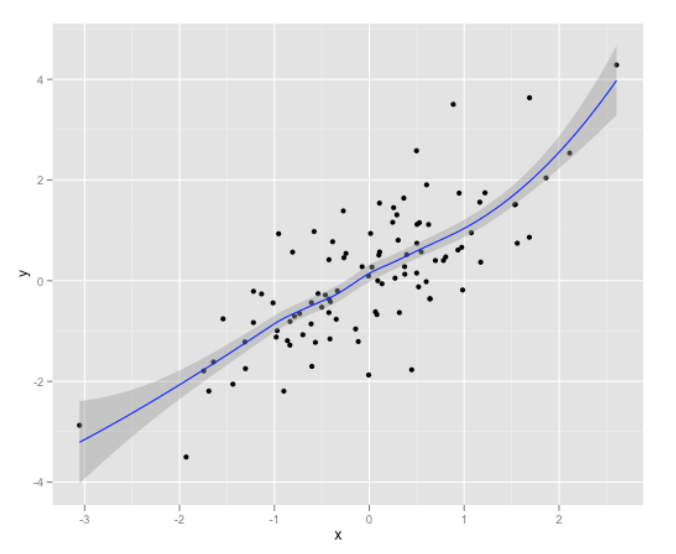
- Examining Analyzing a single or pairs of variables
- Examining the distribution of a single variable
- Analyzing the relationship between two variables
- Data exploration versus data presentation

Visualization

Summary statistics give us some sense of the data:

- Mean vs. Median.
- Standard deviation.
- Quartiles, Min/Max.
- Correlations between variables.

```
summary(data)
       x                    y
 Min.   :-3.05439   Min.   :-3.50179
 1st Qu.:-0.61055   1st Qu.:-0.75968
 Median : 0.04666   Median : 0.07340
 Mean   :-0.01105   Mean   : 0.09383
 3rd Qu.: 0.56067   3rd Qu.: 0.88114
 Max.   : 2.60614   Max.   : 4.28693
```



Visualization gives us
a more holistic sense

Visualization

ADD EXAMPLEs summary and plots

THEME 3: GRAPHICS page 347:The essential R reference

Chap2ggplot.pdf

Chapter 2 R ggplot2 Examples
Bret Larget

http://www.cookbook-r.com/Graphs/Plotting_distributions_(ggplot2)/

https://www.tutorialspoint.com/r/r_scatterplots.htm

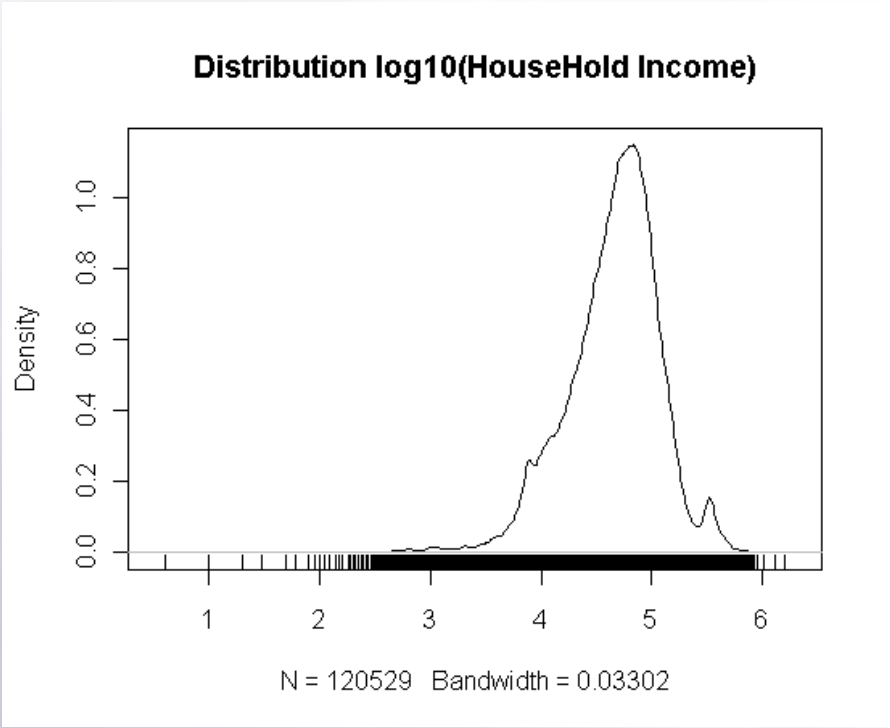
https://www.tutorialspoint.com/r/r_bar_charts.htm

Examining the Distribution of a Single Variable

Graphing a single variable

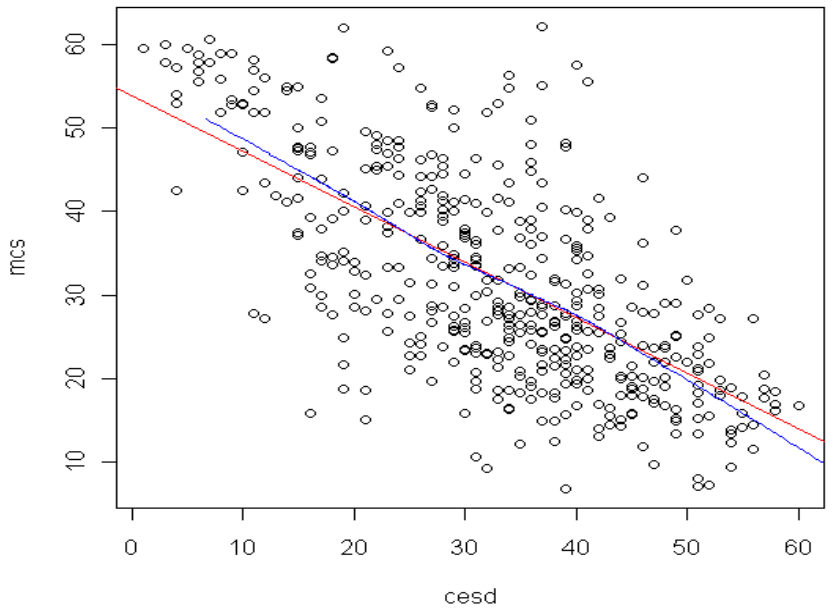
- plot(sort( )) – for low volume data
- Hist( ) – a histogram
- plot(density()) – densityplot
  - A "continuous histogram“

- Example
  - Frequency table of household income



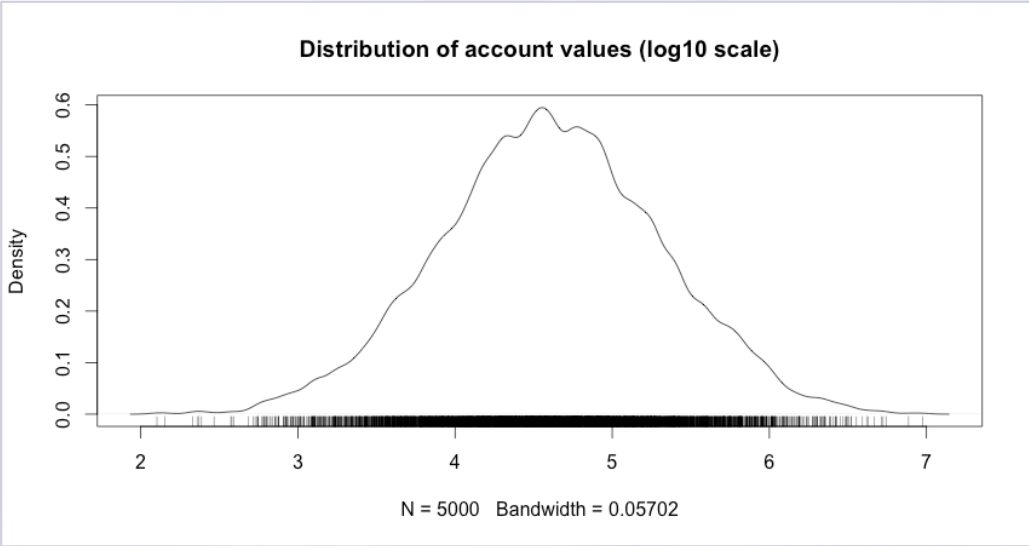
Two Variables: What are we looking for?

- Is there a relationship between the two variables?
  - Linear? Quadratic?
  - Exponential?
    - Try semi-log or log-log plots
  - Is it a cloud?
    - Round? Concentrated? Multiple Clusters?
- How?
  - Scatterplots
- Example
  - Red line: linear fit
  - Blue line: LOESS
  - Fairly linear relationship,
  - but with wide variance



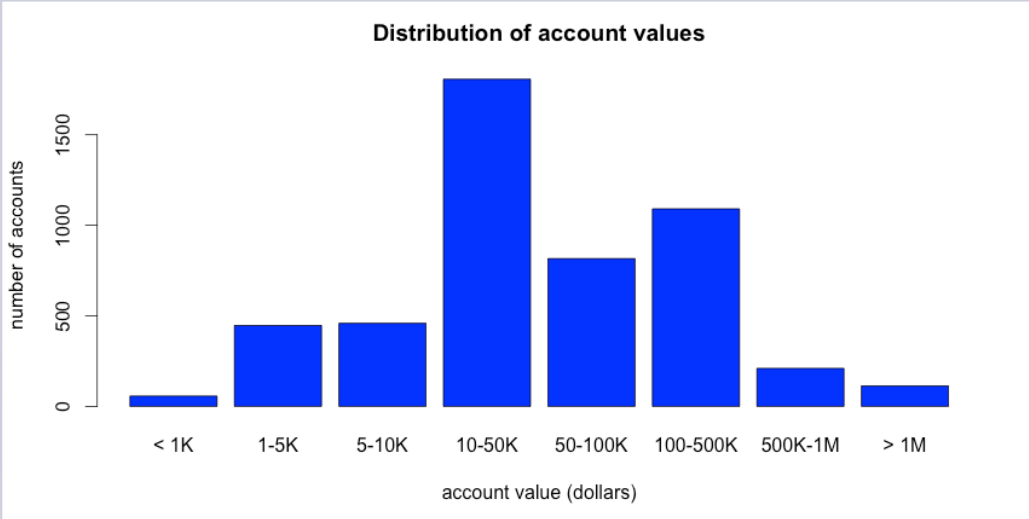

```
{
    plot(mcs ~ cesd)
  abline(lm(mcs ~ cesd), lcol="red")
  lines(lowess(mcs ~ cesd), lcol="blue")
}
```

Data Exploration vs. Presentation



Data Exploration:

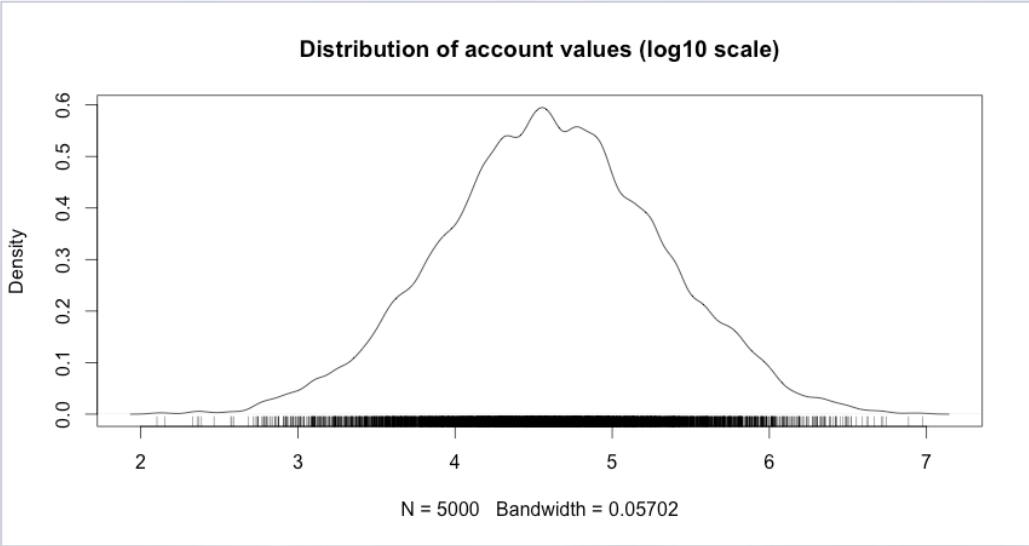
This tells you what you need to know.



Presentation:

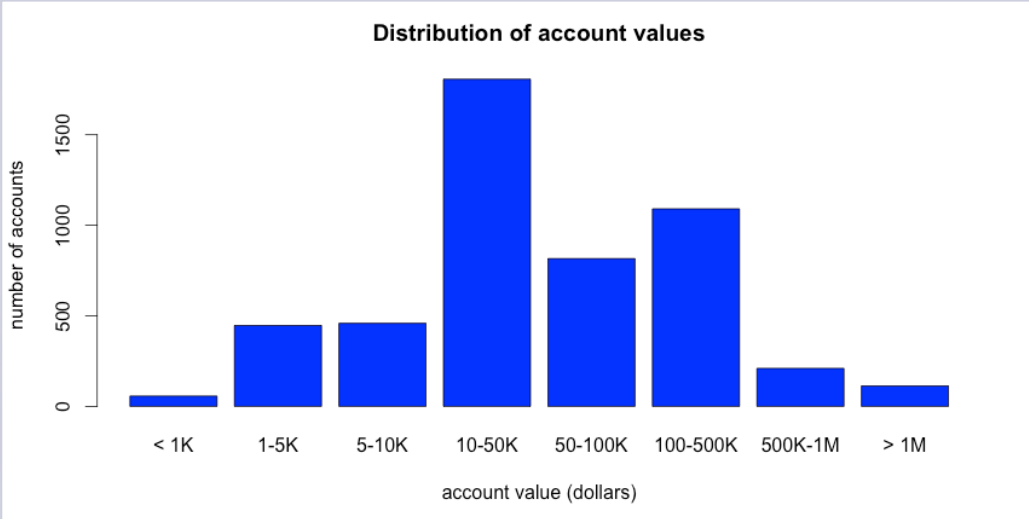
This tells the stakeholders what they need to know.

Data Exploration vs. Presentation



Data Exploration:

This tells you what you need to know.



Presentation:

This tells the stakeholders what they need to know.

Statistics for Model Building and Evaluation

A statistical analysis may be descriptive, simply reporting, visualizing and summarizing a data set, but usually it is also inferential.

- Statistics in the Analytic Lifecycle
- Descriptive Statistics
- Linear Regression
- Hypothesis Testing
- Anova
- Correlations

Statistics in the Analytic Lifecycle

- Understand Data Using Descriptive Statistics
- Model Building and Planning
  - Can I predict the outcome with the inputs that I have?
  - Which inputs?
- Model Evaluation
  - Is the model accurate?
  - Does it perform better than "the obvious guess"
  - Does it perform better than another candidate model?
- Model Deployment
  - Do my predictions make a difference?
    - Are we preventing customer churn?
    - Have we raised profits?

Descriptive Statistics

https://www.tutorialspoint.com/r/r_mean_median_mode.htm

Function	R Code
View the data	head(x); tail(x)
View a summary of the data	**summary(x)**
Compute basic statistics	sd(x); var(x); range(x)
Correlation	cor(x, y)

Using summary statistics to explore data

Data exploration uses a combination of summary statistics “means and medians, variances, and counts” and visualization , or graphs of the data.

In R, you’ll typically use the summary command to take your first look at the data.

EXAMPLE

Suppose your goal is to build a model to predict which of your customers don’t have health insurance;

1- You’ve collected a dataset of customers whose health insurance status you know.

2- identified some customer properties that you believe help predict the probability of insurance coverage: age, employment status, etc..

3- put all your data into a single data frame called custdata that you’ve input into R

Descriptive Statistics

```
In [4]: custdata <- read.delim("~/custdata.tsv")
        summary(custdata)
```

```
     custid          sex     is.employed         income
 Min.   :   2068   F:440   Mode :logical   Min.   : -8700
 1st Qu.: 345667   M:560   FALSE:73        1st Qu.: 14600
 Median : 693403           TRUE :599       Median : 35000
 Mean   : 698500           NA's :328       Mean   : 53505
 3rd Qu.:1044606                           3rd Qu.: 67000
 Max.   :1414286                           Max.   :615000

              marital.stat health.ins                                housing.type
 Divorced/Separated:155    Mode :logical   Homeowner free and clear    :157
 Married           :516    FALSE:159       Homeowner with mortgage/loan:412
 Never Married     :233    TRUE :841       Occupied with no rent       : 11
 Widowed           : 96    NA's :0         Rented                      :364
                                           NA's                        : 56

 recent.move      num.vehicles        age              state.of.res
 Mode :logical   Min.   :0.000   Min.   :  0.0   California   :100
 FALSE:820       1st Qu.:1.000   1st Qu.: 38.0   New York     : 71
 TRUE :124       Median :2.000   Median : 50.0   Pennsylvania: 70
 NA's :56        Mean   :1.916   Mean   : 51.7   Texas        : 56
                 3rd Qu.:2.000   3rd Qu.: 64.0   Michigan     : 52
                 Max.   :6.000   Max.   :146.7   Ohio         : 51
                 NA's   :56                      (Other)      :600
```

Descriptive Statistics

you're looking for several common issues: missing values, invalid values and outliers, and data ranges that are too wide or too narrow.

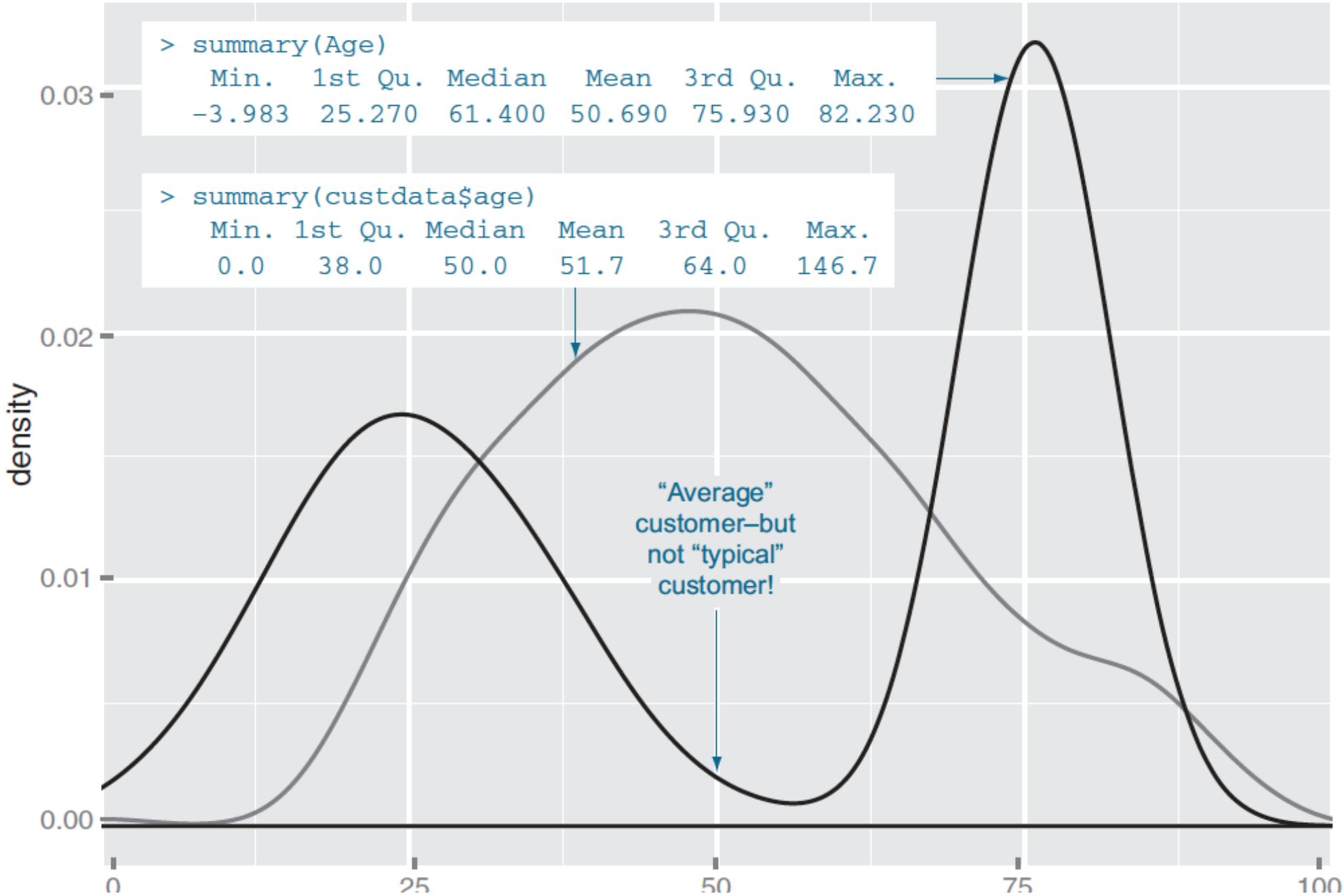
1: The variable is.employed is missing for about a third of the data. NA's :328
The variable income has negative values, which are potentially invalid. -8700

2: About 84% of the customers have health insurance. TRUE :841/1000

3: The variables housing.type, recent.move, and num.vehicles are each missing: 56 values.

4: The average value of the variable age seems plausible, but the minimum and maximum values seem unlikely. age -> min : 0.0

Spotting problems using graphics



R - Linear Regression

Regression analysis is a very widely used statistical tool to establish a relationship model between two variables.

$$y = ax + b$$

In R we create Relationship Model & get the Coefficients

```
x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)
y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)

# Apply the lm() function.
relation <- lm(y~x)
print(relation)

print(summary(relation))

 anova(relation)
```

R - Linear Regression

anova-lm.R

```
1  x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)
2  y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)
3
4  # Apply the lm() function.
5  relation <- lm(y~x)
6  print(summary(relation))
7  anova(relation)
8
```

5:20 (Top Level)

Console

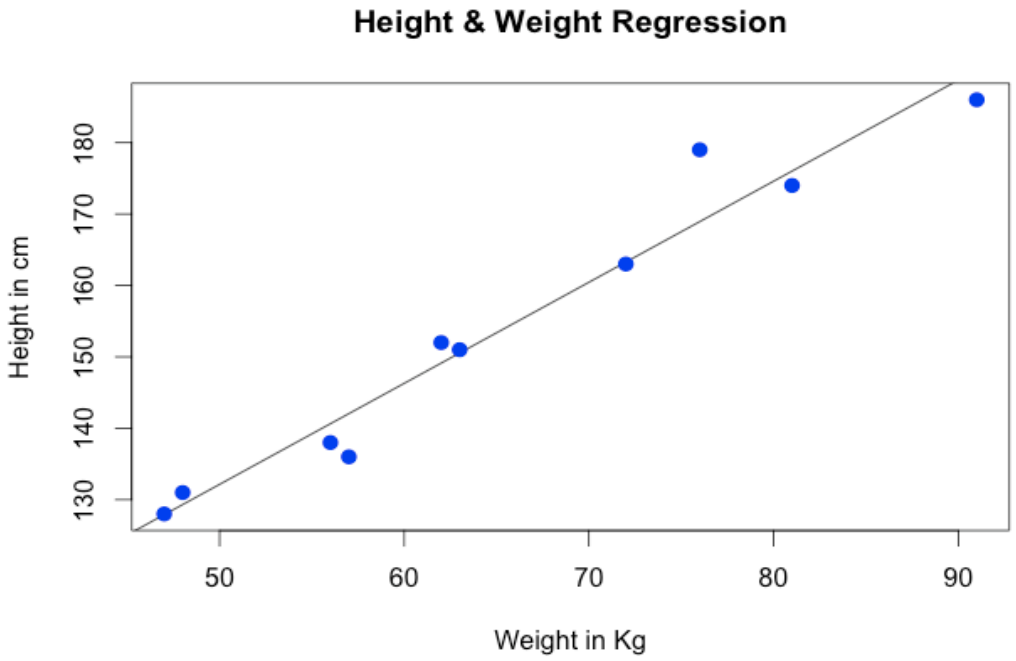
```
> relation <- lm(y~x)
> print(summary(relation))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3002 -1.6629  0.0412  1.8944  3.9775

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -38.45509    8.04901  -4.778  0.00139 **
x             0.67461    0.05191  12.997 1.16e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.253 on 8 degrees of freedom
Multiple R-squared: 0.9548, Adjusted R-squared: 0.9491
F-statistic: 168.9 on 1 and 8 DF,  p-value: 1.164e-06

> anova(relation)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 1787.45 1787.45  168.92 1.164e-06 ***
Residuals  8   84.65   10.58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

```
# Find weight of a person with height 170.
a <- data.frame(x = 170)
result <-  predict(relation,a)
print(result)
plot(y,x,col = "blue",main = "Height & Weight Regression",
abline(lm(x~y)),cex = 1.3,pch = 16,xlab = "Weight in Kg",ylab = "Height in cm")
```

```
> a
   x
1 170
> result
      1
76.22869
```

Files | Plots | Packages | Help

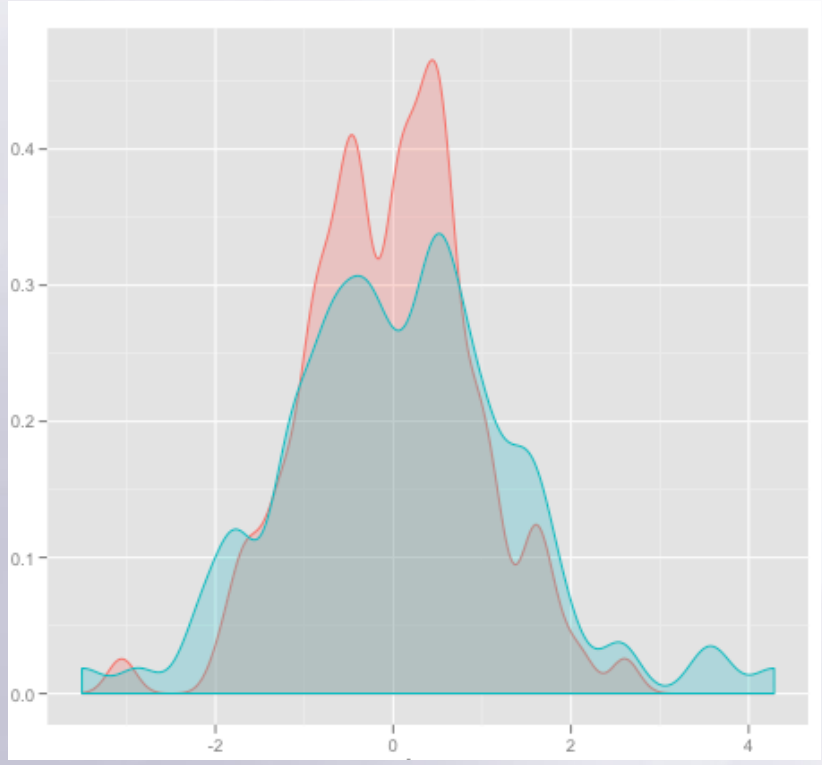
Zoom | Export | Clear All



Statistical Methods for Evaluation

Hypothesis Testing

- Fundamental question: "Is there a difference between the populations based on samples?“
  - Examples : Mean, Variance
- Null hypothesis : There is no difference
- Alternate hypothesis : There is a difference



One sample Student's t-test

Here is an example concerning daily energy intake in kJ for 11 women (Altman, 1991, p. 183). First, the values are placed in a data vector:

```
intake <- c(5260,5470,5640,6180,6390,6515,6805,7515,7515,8230,8770)
```

```
t.test(intake, mu=7725)
```

```
One Sample t-test
data: intake
t = -2.8208, df = 10, p-value = 0.01814
alternative hypothesis: true mean is not
equal to 7725
95 percent confidence interval:
5986.348 7520.925
sample estimates:
mean of x
6753.636
```

We can immediately see that $p < 0.05$ and thus that (using the customary 5% level of significance) data deviate significantly from the hypothesis that the mean is 7725.

One sample Student's t-test

Comparison of the sample mean with a known value, when the variance of the population is not known and n <30.

It was made an intelligence test in 10 subjects.
The average result of the population, is equal to 75.
The sample mean is significantly similar (in 95%) to the average population ?

the Student's t-test for a single sample have a pre-set function in R we can apply immediately.

It is the t.test (a, mu), we can see below applied.

One sample Student's t-test

It was made an intelligence test in 10 subjects.
The average result of the population, is equal to 75.
The sample mean is significantly similar (in 95%) to the average population ?

```
A  <  - c(65, 78, 88, 55, 48, 95, 66, 57, 79, 81)

t.test (a, mu=75)

        One Sample t-test

data:  A
t = -0.78303, df = 9, p-value = 0.4537
alternative hypothesis: true mean is not equal to 75
95 percent confidence interval:
 60.22187       82.17813
sample estimates:
mean of x
    71.2
```

R - Chi Square Test

Chi-Square test is a statistical method to determine if two categorical variables have a significant correlation between them.

```
# Chi-sq test
row1 = c(91,90,51)
row2 = c(150,200,155)
row3 = c(109,198,172)
data.table = rbind(row1,row2,row3)
chisq.test(data.table)
```

```
> data.table
     [,1] [,2] [,3]
row1   91   90   51
row2  150  200  155
row3  109  198  172
> chisq.test(data.table)

 Pearson's Chi-squared test

data:  data.table
X-squared = 25.086, df = 4, p-value = 4.835e-05
```

```
# Chi-sq test
setwd("/Users/ouzarf")
df <- read.csv("treatment.csv")
table(df$treatment, df$improvement)
chisq.test(df$treatment, df$improvement)
```

```
> df <- read.csv("treatment.csv")
> table(df$treatment, df$improvement)

              improved not-improved
  not-treated       26           29
  treated           35           15
> chisq.test(df$treatment, df$improvement)

 Pearson's Chi-squared test with Yates' continuity correction

data:  df$treatment and df$improvement
X-squared = 4.6626, df = 1, p-value = 0.03083
```

 # If we get a p-Value less than the significance level of 0.05,
 # we reject the null hypothesis and conclude that the two variables are in fact dependent.

 # which indicates a string correlation.

Summary

Key points covered in this part:

- How to use basic analytics methods such as distributions, statistical tests and summary operations to investigate a data set
- How to use R to apply visualization patterns to better understand the data, help develop a model and derive hypotheses, and determine if our actions had a practical affect.

THANK YOU

Hypothesis Testing

In this Hypothesis testing we use :

"t-test","z-test","Anova","F-test","chi-square”

to prove whether Ho is correct (or) H1 is correct.

5.1 One-sample t tes page 97

https://www.youtube.com/watch?v=RIhnNbPZC0A

https://www.r-bloggers.com/one-sample-students-t-test/

How PACKAGE in R

http://lumimath.univ-mrs.fr/~broglio/