An Introduction to Data Analysis with R

Eralp DOGU, Post-Doctoral Research Associate
Healthcare Systems Engineering Institute (HSyE)
Northeastern University
Email: e.dogu@neu.edu
www.HSyE.org
Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

Drowning In Big Data - Finding Insight In A Digital Sea Of Information

For roughly a decade, businesses and governments have been collecting digital information about the world’s citizens; not just names and numbers, but loads of data. Big Data. The IDC (International Data Corporation) recently predicted that the Big Data industry will experience a 26.4% compound annual growth rate, reaching $41.5 billion by 2018. What this means is that more and more companies and organizations are paying top dollar to media agencies to be deluged in never-ending monsoons of big data and lots of money is being spent figuring out how to harness its power. Chris Mattmann, from the big-data initiative at the NASA/Propulsion Lab (JPL) in Pasadena, CA says, “NASA in total is probably managing several hundred petabytes, and somehow, an operation of all that data is in getting out the information that people need.”

Work out and flex your data muscles!

3 Career Tips for Becoming a Data Scientist

Seemingly overnight, data scientists have been pushed into the online spotlight. In the October 2012 edition of the Harvard Business Review, the position of “data scientist” was labeled as the sexiest job of the 21st Century. Job posting search engines like Indeed.com confirm the rise in popularity of this position, showing 15,000 percent growth over the last few years.

But what is a data scientist, and what does it take to become one? The role originated in companies such as LinkedIn, Google and Facebook, where analysts were tasked with managing databases with terabytes, petabytes, exabytes and sometimes zettabytes of data. These individuals are more than just data wizards; however, they’re what scientists BruceLee calls, “business analysts plus.” Business skills are almost as important as programming and statistical-analysis skills for data scientists, as they’re often the bridge between executives and the world of Data.

If you’re an aspiring data scientist, here are three secrets to career success.
R Development Core Team

http://www.r-project.org/
http://en.wikipedia.org/wiki/R_%28programming_language%29
R is a tool for...

- Data manipulation
  - Clean/manipulate data
  - Slice/subset data
- Analysis and modelling
  - Statistical modelling
  - Numerical simulation
- Data visualization
  - Summary graphics
  - Visualization fit of models

R basic interface is simple

- [http://cran.r-project.org/](http://cran.r-project.org/)
R Studio is an environment

- [http://www.rstudio.com/](http://www.rstudio.com/)

Hello, world!
Take a tour...

```r
Take a tour...

Console

1948 0.4948495
1955 1.1810416
1956 -9.4633838
1957 -9.0050505
1958 -20.0053053
1959 -17.0053053
1960 NA

$figure


$type

[1] "additive"

attr("class")

[1] "decomposed.ts" "plot(decompose(Apts))" "library(MASS)" "data(faithful)" "head(faithful)"

eruptions waiting

1 3.600 79
2 1.800 64
3 3.333 74
4 2.283 62
5 4.122 85
6 2.883 55

> dim(faithful)

[1] 272 2

> cor(faithful)

eruptions waiting

eruptions 1.0000000 0.9043112
waiting 0.9043112 1.0000000

> hist(faithful)

Error in hist.default(faithful) : 'x' must be numeric

> hist(faithful$eruptions)

> plot(faithful)
```

Take a tour...
Take a tour...

![Histogram of faithful eruptions](image1)

Take a tour...

![Scatter plot of faithful data](image2)
Using R as a calculator

• Simple math

• Storing results in variables

• Vectorized math

Using R as a calculator

• Basic statistics

• Advanced statistics
Using R as a simulator

- Use built-in functions for probability distributions
- Assume length of stay for patients at a certain institution follows an exponential distribution with mean=3 days

```r
> LOS <- rexp(1000, 1/3)
> hist(LOS)
```

Functions for Probability Distributions

- **d**dist: density function (pdf)
- **p**dist: cumulative density function (cdf)
- **q**dist: quantile function
- **r**dist: random number generator
Importing-exporting data

• Excel files
  ```r
  > read.csv("hsye.csv", header=TRUE)
  > write.csv(hsye, file="hsye.csv")
  ```

• Web
  ```r
  > url('http://coe.neu.edu/healthcare/test.txt')
  ```

• R objects
  ```r
  > load('hsye.RData')
  > data(hsye)
  > save(hsye, file="hsye.RData")
  ```

Navigating data

• Data frames as matrices
  ```r
  > data(faithful)
  Error in `[.data.frame`(faithful, 1:10) : undefined columns selected
  > faithful[1:10,]
  eruptions waiting
  1 3.600 79
  2 1.800 54
  3 2.833 74
  4 2.283 62
  5 4.333 85
  6 2.863 55
  7 4.700 88
  8 3.600 85
  9 3.950 51
  10 4.350 85
  > faithful[faithful$waiting>90,]
  eruptions waiting
  100 3.783 90
  101 4.716 90
  102 4.400 92
  103 4.333 90
  104 4.650 90
  105 5.100 96
  106 4.083 93
  107 4.627 93
  108 4.133 91
  109 4.800 94
  110 4.450 90
  111 4.417 90
  ```
Creating subsets

• Subsetting data

• Transforming data

Using R as a Statistical Modeler

• R has a powerful modelling syntax and model alternatives

• Models are basically specified with a formulae like
  \[ y \sim x \]
  \[ \text{LOS} \sim \text{age} + \text{surgery\_type} \]

• Models are also guide the visualization of relationships in a graphical form
Using R as a Statistical Modeler

• R has a powerful modelling syntax and model alternatives

• Models are basically specified with a formulae like

  \[ y \sim x \]
  \[ \text{LOS} \sim \text{age} + \text{surgery\_type} \]

• Models are also guide the visualization of relationships in a graphical form
Using R as a Statistical Modeler

- Logistic model

```r
> survival.lm <- glm(Health_outcome~Age+Surgery_type+Hypertension, 
data=hsye, family=binomial("logit"), na.omit )
```

I need to click!
Rcmdr()
Data Visualization

- [http://shinyapps.stat.ubc.ca/r-graph-catalog/](http://shinyapps.stat.ubc.ca/r-graph-catalog/)
- [http://rgraphgallery.blogspot.com/](http://rgraphgallery.blogspot.com/)

```r
require(latticeExtra)
require(lattice)
data(SeatacWeather)
temperatures <- xyplot(min.temp + max.temp ~ day | month, 
data = SeatacWeather, type = "l", layout = c(3, 1))
rainfall <- xyplot(precip ~ day | month, data = SeatacWeather, type = "h", lwd = 4)
doubleYScale(temperatures, rainfall, style1 = 0, style2 = 3, add.ylab2 = TRUE, 
text = c("min. T", "max. T", "rain"), columns = 3)
```
Data Visualization

ggplot2()
Data Visualization

**ggplot2()**

![Graphs and plots demonstrating data visualization using ggplot2](image-url)

---

**Code Example**

```r
library(ggplot2)
library(grid)

# Use an example of the Gapminder data
# Prepare and format the data
gapminder <- read.csv(file = gap_url)

# Add countries to the data frame
my_countries <- c("Canada", "Rwanda", "Oman")

# Create a direct label data frame
labeled_data <- data.frame(year = 2000, 
                           lifeexp = gapminder$lifeexp[year == 2000], 
                           country = my_countries)

# Use ggplot to create the graph
ggplot() + geom_point(aes(x = year, y = lifeexp, group = country, 
                          shape = country), 
                     data = labeled_data) + 
  geom_text(aes(x = year, y = lifeexp, 
                label = country), 
            data = labeled_data) + 
  facet_wrap(~ year) + 
  coord_cartesian(ylim = c(0, 80))
```

---

**Footnote:**

Go to CSV file to download figure and code.
Going Interactive

shiny()

• [http://shiny.rstudio.com/gallery/google-charts.html](http://shiny.rstudio.com/gallery/google-charts.html)

Going Interactive

shiny()

• [http://shiny.rstudio.com/gallery/word-cloud.html](http://shiny.rstudio.com/gallery/word-cloud.html)
Statistical Reporting

Sweave()

> \text{lm} = \text{lm}(y \sim x)
The plot of $x$ versus $y$, together with the regression line is generated by

> plot(x, y)
> abline(lm)

and is now shown in Figure 1.

![Graph of Regression](image)

Figure 1: Regression of $y$ on $x$

The complete code (B/I/X and R) is

\begin{verbatim}
\begin{figure}[ht]
\centering
\setkeys{Gin}{width=0.5\textwidth}
<<echo=FALSE, fig=TRUE, width=4, height=4>>=
plot(x, y)
abline(lm)
\end{verbatim}
\end{figure}

Getting help

• Help function: \text{help(func)} or \text{?func}

• \url{http://cran.r-project.org/}
R Sources

• Web...
• [http://www.r-project.org/doc/bib/R-books.html](http://www.r-project.org/doc/bib/R-books.html)