Module 1 :Data, "Big Data", and Data Analytics

<https://youtu.be/qgyPf4Ln9wM>

<https://r4ds.had.co.nz/>

[Smith, Jerry A (2013). Six Lists of Lists for Data Scientists. December 28, 2015.](https://datascientistinsights.com/2015/12/28/six-lists-of-lists-for-data-scientists/)

[Links to an external site.](https://datascientistinsights.com/2015/12/28/six-lists-of-lists-for-data-scientists/)

<https://youtu.be/-NVMoOqDkLU>

# Lesson 3: Data Analysis Pipeline and CRISP-DM

<https://www.youtube.com/watch?v=nNc_q08yWxw&t=3s>

# Lesson 4: Exploring Data with R

<https://youtu.be/5Y7tfFCgZ6I>

* (90 min) Alternatively, you can work along with the video lecture and the two-part tutorial by Roger Peng that covers the same topics and is based on our text book

Peng, R. (Oct 14, 2013). Plotting with ggplot2: Part 1 [Video]. YouTube. <https://youtu.be/HeqHMM4ziXA>

[Links to an external site.](https://youtu.be/HeqHMM4ziXA)

Peng, R. (Oct 14, 2013). Plotting with ggplot2: Part 2 [Video]. YouTube. <https://youtu.be/n8kYa9vu1l8>

[Links to an external site.](https://youtu.be/n8kYa9vu1l8)

* Post questions, comments, work-around, discoveries for this week
* <http://tryr.codeschool.com/>
* [Links to an external site.](http://tryr.codeschool.com/)
* [R Essentials](https://d1b10bmlvqabco.cloudfront.net/attach/ivo6jwdccw617v/ijayba1a9vm6hc/ixuszg2t5ij4/REssentials.html)
* [Links to an external site.](https://d1b10bmlvqabco.cloudfront.net/attach/ivo6jwdccw617v/ijayba1a9vm6hc/ixuszg2t5ij4/REssentials.html)
* [Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics*,](http://vita.had.co.nz/papers/layered-grammar.pdf) *[19](http://vita.had.co.nz/papers/layered-grammar.pdf)*
* [Links to an external site.](http://vita.had.co.nz/papers/layered-grammar.pdf)
* [(1), 3-28](http://vita.had.co.nz/papers/layered-grammar.pdf)
* [Links to an external site.](http://vita.had.co.nz/papers/layered-grammar.pdf)

We will learn a lot of R in this course and you will have numerous opportunities to hone your data processing and analysis skills in R. However, some of you might want additional resources, particularly if you haven't programmed much before. Fortunately, Northeastern University subscribes to Lynda's online training courses. If you need additional background in R programming, you are urged to take the course "[Up and Running with R](http://www.lynda.com/R-tutorials/Up-Running-R/120612-2.html?srchtrk=index:1%0Alinktypeid:2%0Aq:R%0Apage:1%0As:relevance%0Asa:true%0Aproducttypeid:2)

[Links to an external site.](http://www.lynda.com/R-tutorials/Up-Running-R/120612-2.html?srchtrk=index:1%0Alinktypeid:2%0Aq:R%0Apage:1%0As:relevance%0Asa:true%0Aproducttypeid:2)

" by Barton Poulson on Lynda. Sign in to Lynda Online Training by going to [lynda.northeastern.edu](http://lynda.northeastern.edu/)

[Links to an external site.](http://lynda.northeastern.edu/)

or from Blackboard. Log on with your myNEU username and password. Note that the Google Chrome may not be fully compatible with Lynda, so you may need to switch to an alternate browser (Safari, Internet Explorer, or Firefox). You can also install the Lynda mobile app from their website or from the appropriate app store for your iOS or Android device. You will need to [log in on the apps with your myNEU credentials](https://kb.northeastern.edu:8443/portal/guestviewsolution.jsp?solutionid=041326314391488&isguest=true&gcc=neu)

[Links to an external site.](https://kb.northeastern.edu:8443/portal/guestviewsolution.jsp?solutionid=041326314391488&isguest=true&gcc=neu)

.

These tutorials by the kind folks at Datacamp are also wonderful and might help some of you:

* For basics in R: <https://www.datacamp.com/tracks/r-programming>
* [Links to an external site.](https://www.datacamp.com/tracks/r-programming)
* For visualization in R using ggplot2: <https://www.datacamp.com/courses/data-visualization-with-ggplot2-1>
* [Links to an external site.](https://www.datacamp.com/courses/data-visualization-with-ggplot2-1)

Module 2

# The CRISP-DM Model

<https://youtu.be/dzbajM-nKPw>

* (45 min) Read the following [article by Colin Shearer](https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf)
* [Links to an external site.](https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf)
* to get an overview of the *[CRISP-DM](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining)*
* *[Links to an external site.](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining)*
* model for data mining; it outlines the key steps a data scientist generally takes when mining data and building a predictive model.

### Additional Resources

* Article: [What is the CRISP-DM Methodology by SV Europe](http://www.sv-europe.com/crisp-dm-methodology/)
* [Links to an external site.](http://www.sv-europe.com/crisp-dm-methodology/)
* Video Tutorial: [Data Mining Process and CRISP-DM](https://youtu.be/nNc_q08yWxw)
* [Links to an external site.](https://youtu.be/nNc_q08yWxw)
* (7:45m)
* Video Tutorial: [Basics of Data Mining](https://youtu.be/u2oSiVOQRmg)
* [Links to an external site.](https://youtu.be/u2oSiVOQRmg)
* (9:49m)

<https://www.youtube.com/watch?v=8SGif63VW6E>

<https://www.youtube.com/watch?v=aywFompr1F4>

* Watch the [lecture by Hadley Wickham](https://youtu.be/8SGif63VW6E)
* [Links to an external site.](https://youtu.be/8SGif63VW6E)
* on dplyr, a common R package for easier data manipulation during the data preparation phase of a project.

This second lecture "[Introduction to the dplyr R package](https://youtu.be/aywFompr1F4)

[Links to an external site.](https://youtu.be/aywFompr1F4)

" by Roger Peng is a shorter (condensed) version for more experienced R programmers or those in a hurry.

* Read [Chapter 5 (Data Transformation)](http://r4ds.had.co.nz/transform.html)
* [Links to an external site.](http://r4ds.had.co.nz/transform.html)
* in *[Grolemund, G., Wickham, H. R for Data Science](http://r4ds.had.co.nz/)*
* *[Links to an external site.](http://r4ds.had.co.nz/)*
* and work along in R.

### Additional Resources

* ​[Cleaning and Wrangling Data with TidyR and dplyR](https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf)
* [Links to an external site.](https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf)
* Organising the data
* Read [Chapter 18 (Pipes)](http://r4ds.had.co.nz/pipes.html)
* [Links to an external site.](http://r4ds.had.co.nz/pipes.html)
* in *[Grolemund, G., Wickham, H. R for Data Science](http://r4ds.had.co.nz/)*
* *[Links to an external site.](http://r4ds.had.co.nz/)*
* and work along in R.
* (5 min) Watch the tutorial on [How to use pipes in dplyr - R](https://youtu.be/qxF_W0R4tfs)
* [Links to an external site.](https://youtu.be/qxF_W0R4tfs)

Watch the tutorial video on manipulating data frames and performing simple data exploration and preparation. Work along in R and reproduce the commands.

<https://youtu.be/VTOE6NLlsV8>

For those who want to learn more about plain R, should take a look at the following chapters from [Manas A. Pathak Beginning Data Science with R](http://link.springer.com/book/10.1007/978-3-319-12066-9)

[Links to an external site.](http://link.springer.com/book/10.1007/978-3-319-12066-9)

. Note that you don't have to read them deeply but rather follow alongside R and explore the different techniques in exploring a data set and becoming familiar with it. This is essential to determine what kind of data mining or machine learning techniques should be (and can be) applied.

Module 3 : Data exploration

Watch the tutorials and alongside read [Chapter 7 (Exploratory Data Analysis)](http://r4ds.had.co.nz/exploratory-data-analysis.html)

[Links to an external site.](http://r4ds.had.co.nz/exploratory-data-analysis.html)

in *[Grolemund, G., Wickham, H. R for Data Science](http://r4ds.had.co.nz/)*

*[Links to an external site.](http://r4ds.had.co.nz/)*

*.* Work along in R.

* Read [Chapter 8 (Workflow)](http://r4ds.had.co.nz/workflow-projects.html)
* [Links to an external site.](http://r4ds.had.co.nz/workflow-projects.html)
* in *[Grolemund, G., Wickham, H. R for Data Science](http://r4ds.had.co.nz/)*
* *[Links to an external site.](http://r4ds.had.co.nz/)*
* *. W*ork along in R.

[Summary Statistics in R](https://www.r-bloggers.com/r-tutorial-series-summary-and-descriptive-statistics/)

* [Links to an external site.](https://www.r-bloggers.com/r-tutorial-series-summary-and-descriptive-statistics/)
* ​
* [Black, Kelly. R Tutorial: Statistics and Descriptive Analytics](http://www.cyclismo.org/tutorial/R/index.html)
* [Links to an external site.](http://www.cyclismo.org/tutorial/R/index.html)
* ​
* [Wickham, H. gplot2: Elegant Graphics for Data Analysis. Springer](https://link-springer-com.ezproxy.neu.edu/book/10.1007%2F978-3-319-24277-4)
* [Links to an external site.](https://link-springer-com.ezproxy.neu.edu/book/10.1007%2F978-3-319-24277-4)
* (downloadable PDF via Northeastern Common Login)

Tidying data with tibbles

* Read [Chapter 9 (Introduction to Data Wrangling)](http://r4ds.had.co.nz/wrangle-intro.html)
* [Links to an external site.](http://r4ds.had.co.nz/wrangle-intro.html)
* and [Chapter 10 (Tibbles)](http://r4ds.had.co.nz/tibbles.html)
* [Links to an external site.](http://r4ds.had.co.nz/tibbles.html)
* in *[Grolemund, G., Wickham, H. R for Data Science](http://r4ds.had.co.nz/)*
* *[Links to an external site.](http://r4ds.had.co.nz/)*
* *.* Work alongside in R.
* *​(45 min) Watch* the video tutorial on [Tibbles](https://youtu.be/0bcA3-6fKDk)
* [Links to an external site.](https://youtu.be/0bcA3-6fKDk)
* by Peter Hurford. Work alongside in R.

​[Data wrangling - Wikipedia](https://en.wikipedia.org/wiki/Data_wrangling)

[Links to an external site.](https://en.wikipedia.org/wiki/Data_wrangling)

* Importing data from file
* Read [Chapter 11 (Data Import)](http://r4ds.had.co.nz/data-import.html)
* [Links to an external site.](http://r4ds.had.co.nz/data-import.html)
* in *[Grolemund, G., Wickham, H. R for Data Science](http://r4ds.had.co.nz/)*
* *[Links to an external site.](http://r4ds.had.co.nz/)*
* *.* Work alongside in R. Note that *read.csv()* in base R is not the same as *read\_csv()* from the **tidyverse** package.
* [How to Open Big CSV Files](https://www.csvexplorer.com/blog/open-big-csv/)
* [Links to an external site.](https://www.csvexplorer.com/blog/open-big-csv/)

The work of data scientist

View the interviews to get a sense of what it's like to work as a data scientist. It's not always about math, and programming, and algorithms. Notice how much time they say is spent on shaping data and determining what the actual problem is that needs to be solved.

* IBM Data and AI (Aug 3, 2016). [Data science expert interview: Jennifer Shin](https://youtu.be/BRQhZFYDCiw)
* [Links to an external site.](https://youtu.be/BRQhZFYDCiw)
* . [Video]. YouTube. <https://youtu.be/BRQhZFYDCiw>
* [Links to an external site.](https://youtu.be/BRQhZFYDCiw)
* Intermountain Healthcare (Mar 1, 2017). [A Day in the Life of a Data Analyst](https://youtu.be/xbTqJiZ8nhA)
* [Links to an external site.](https://youtu.be/xbTqJiZ8nhA)
* . [Video]. YouTube. <https://youtu.be/xbTqJiZ8nhA>
* [Links to an external site.](https://youtu.be/xbTqJiZ8nhA)
* RCR Wireless News (Feb 17, 2015). [Gigs: A day in the life of a data scientist](https://youtu.be/EaptTxhh6sM)
* [Links to an external site.](https://youtu.be/EaptTxhh6sM)
* . [Video]. YouTube. <https://youtu.be/EaptTxhh6sM>
* [Links to an external site.](https://youtu.be/EaptTxhh6sM)
* MortarData2013 (Mar 6, 2013). [Dirty Secrets of Data Science by Hilary Mason](https://youtu.be/fZuDwiM1XBQ)
* [Links to an external site.](https://youtu.be/fZuDwiM1XBQ)
* . [Video]. YouTube. <https://youtu.be/fZuDwiM1XBQ>
* [Links to an external site.](https://youtu.be/fZuDwiM1XBQ)

Strings and regular expression

* Read [Chapter 14 (Strings)](http://r4ds.had.co.nz/strings.html)
* [Links to an external site.](http://r4ds.had.co.nz/strings.html)
* in *[Grolemund, G., Wickham, H. R for Data Science](http://r4ds.had.co.nz/)*
* *[Links to an external site.](http://r4ds.had.co.nz/)*
* *.* Work alongside in R. We will see this lesson again next week.
* (45 min) Optionally, watch the tutorial on [Regular Expressions in R](https://youtu.be/q8SzNKib5-4)
* [Links to an external site.](https://youtu.be/q8SzNKib5-4)
* by Roger Peng. This is a very "computer science" topic but helps tremendously in building text parsing functions. Again, we will see more of that next week. Follow along in R.

[Keenan, Tyler (2017). An Introduction to Predictive Analytics. B2C. February 27, 2017](http://www.business2community.com/brandviews/upwork/intro-predictive-analytics-01789589#B4FDZUQrVAVI2wGM.97)

Module 4 : Strings, Text Processing, and Regular Expressions

* Watch the tutorial video and read [Chapter 14 (Strings)](http://r4ds.had.co.nz/strings.html)
* [Links to an external site.](http://r4ds.had.co.nz/strings.html)
* in *[Grolemund, G., Wickham, H. R for Data Science](http://r4ds.had.co.nz/)*
* *[Links to an external site.](http://r4ds.had.co.nz/)*
* *.* Work alongside in R. Note that **stringr** is built on top of **stringi**; it is helpful to eventually learn about **stringi** if you are doing complex string processing.

<https://youtu.be/WwLBC_giFhU>

<https://youtu.be/igmZg0w2T9A>

Watch the tutorial on [Regular Expressions in R](https://youtu.be/q8SzNKib5-4)

[Links to an external site.](https://youtu.be/q8SzNKib5-4)

by Roger Peng. This is a very "computer science" topic but helps tremendously in building text parsing functions. Follow along in R.

* [Dealing with Regular Expressions](http://uc-r.github.io/regex#regex_syntax)
* [Links to an external site.](http://uc-r.github.io/regex#regex_syntax)
* - University of Cincinnati
* [Gaston Sanhez Handling Strings in R](http://www.gastonsanchez.com/r4strings/)
* [Links to an external site.](http://www.gastonsanchez.com/r4strings/)
* Chapters 7 - 9
* [Simple tutorial for Regular Expressions](https://regexone.com/lesson/introduction_abcs)
* [Links to an external site.](https://regexone.com/lesson/introduction_abcs)
* [Regex101.com](https://regex101.com/)
* [Links to an external site.](https://regex101.com/)
* - online regular expression tester.
* [Regexpal.com](http://www.regexpal.com/)
* [Links to an external site.](http://www.regexpal.com/)
* - another tester.
* [Regular expressions tutorial](https://regexone.com/lesson/introduction_abcs)
* [Links to an external site.](https://regexone.com/lesson/introduction_abcs)

# Lesson 2: Data Types, Objects, Mode, Class, and Factors

Watch the tutorial video and then read [Chapter 15 (Factors)](http://r4ds.had.co.nz/factors.html)

[Links to an external site.](http://r4ds.had.co.nz/factors.html)

in *[Grolemund, G., Wickham, H. R for Data Science](http://r4ds.had.co.nz/factors.html)*

*[Links to an external site.](http://r4ds.had.co.nz/factors.html)*

*.* Work alongside in R and be sure to recreate all plots and code examples.

​​​[R Bloggers: Data Types, Part 3: Factors](https://www.r-bloggers.com/data-types-part-3-factors/)

Date and time processing

<https://youtu.be/DVJ9KmZHZ9M>

Organizing code with functions

<https://youtu.be/ucl4jkZsFIo>

[Quick-R: User Written Functions](https://www.statmethods.net/management/userfunctions.html)

### Optional Work

View the tutorials to start learning about XML and how to read XML into R and process it. XML is a common data interchange format and is more common when data has non-tabular structure.

* DevNami (Jun 23, 2016). [Read XML Data in R](https://youtu.be/1cM_ZNZ9hhE)
* [Links to an external site.](https://youtu.be/1cM_ZNZ9hhE)
* . [Video]. YouTube. <https://youtu.be/1cM_ZNZ9hhE>
* [Links to an external site.](https://youtu.be/1cM_ZNZ9hhE)
* DevNami (Feb 1, 2017). [R Programming Read XML Node Details](https://youtu.be/VnWtvi4ZsA0)
* [Links to an external site.](https://youtu.be/VnWtvi4ZsA0)
* . [Video]. YouTube. <https://youtu.be/VnWtvi4ZsA0>
* [Links to an external site.](https://youtu.be/VnWtvi4ZsA0)
* DevNami (Jan 16, 2017). [R Programming Convert XML to Data Frame](https://youtu.be/im8MCBZdqFE)
* [Links to an external site.](https://youtu.be/im8MCBZdqFE)
* . [Video]. YouTube. <https://youtu.be/im8MCBZdqFE>
* [Links to an external site.](https://youtu.be/im8MCBZdqFE)

Module 5 : organizing data : relational and Hierarchical

Data organization approaches <https://youtu.be/3-nOlpzl8bM>

Tabular tidy data Watch the (short) video on [Principles of tidy data](https://youtu.be/oQuupzfX9OQ)

[Links to an external site.](https://youtu.be/oQuupzfX9OQ)

by Nick Carchedi and then read [Chapter 12 (Tidy Data)](http://r4ds.had.co.nz/tidy-data.html)

[Links to an external site.](http://r4ds.had.co.nz/tidy-data.html)

in *[Grolemund, G., Wickham, H. R for Data Science](http://r4ds.had.co.nz/)*

*[Links to an external site.](http://r4ds.had.co.nz/)*

*.* Work alongside in R. If you have seen relational databases before, then these ideas will seem familiar - it is essentially *normalization*

* *[Data Carpentry. Manipulating, analyzing and exporting data with tidyverse.](http://www.datacarpentry.org/R-ecology-lesson/03-dplyr.html)*
* *[Links to an external site.](http://www.datacarpentry.org/R-ecology-lesson/03-dplyr.html)*

Relational data <https://youtu.be/_J35v5HOkSE>

* [William Surles, (2017). Joining Data in R with dplyr. RPubs.](http://rpubs.com/williamsurles/293454)
* [Links to an external site.](http://rpubs.com/williamsurles/293454)

# Hierarchical Data and XML

<https://youtu.be/-3FDcUEFc84>

<https://youtu.be/Sy2dxrx6BC0>

* [W3Schools XPath Tutorial](https://www.w3schools.com/xml/xml_xpath.asp)
* [Links to an external site.](https://www.w3schools.com/xml/xml_xpath.asp)
* [Bosede, Tobi. Working with XML Data in R. June 13, 2014](http://www.informit.com/articles/article.aspx?p=2215520)
* [Links to an external site.](http://www.informit.com/articles/article.aspx?p=2215520)
* [R and the Web, Part II: XML in R. June 22, 2012](http://giventhedata.blogspot.com/2012/06/r-and-web-for-beginners-part-ii-xml-in.html)
* [Links to an external site.](http://giventhedata.blogspot.com/2012/06/r-and-web-for-beginners-part-ii-xml-in.html)
* ​
* [Lang, Duncan. An Introduction to the XML Package for R](https://cran.r-project.org/web/packages/XML/XML.pdf)
* [Links to an external site.](https://cran.r-project.org/web/packages/XML/XML.pdf)

### Slide Deck & Data Sets

* [Slide Deck](https://1drv.ms/p/s!AvFTFQKva2aZmyTkx1Nas6x2yA-3)
* [Links to an external site.](https://1drv.ms/p/s!AvFTFQKva2aZmyTkx1Nas6x2yA-3)
* [PubMed Data Set](https://da5020.weebly.com/uploads/8/6/5/9/8659576/pubmed_result_1_.xml)

Processing XML

Watch and work alongside these three tutorials to get another look at processing XML and more practice.

* DevNami (Jun 23, 2016). [Read XML Data in R](https://youtu.be/1cM_ZNZ9hhE)
* [Links to an external site.](https://youtu.be/1cM_ZNZ9hhE)
* . [Video]. YouTube. <https://youtu.be/1cM_ZNZ9hhE>
* [Links to an external site.](https://youtu.be/1cM_ZNZ9hhE)
* DevNami (Feb 1, 2017). [R Programming Read XML Node Details](https://youtu.be/VnWtvi4ZsA0)
* [Links to an external site.](https://youtu.be/VnWtvi4ZsA0)
* . [Video]. YouTube. <https://youtu.be/VnWtvi4ZsA0>
* [Links to an external site.](https://youtu.be/VnWtvi4ZsA0)
* DevNami (Jan 16, 2017). [R Programming Convert XML to Data Frame](https://youtu.be/im8MCBZdqFE)
* [Links to an external site.](https://youtu.be/im8MCBZdqFE)
* . [Video]. YouTube. <https://youtu.be/im8MCBZdqFE>

Module 6 : retrieving data and relational databases

Ket concepts in relational database

<https://youtu.be/BfIIpEtIOok>

<https://www.youtube.com/watch?v=FCtGbrou5DM&t=1s>

[Links to an external site.](https://people.csail.mit.edu/rivest/HyafilRivest-ConstructingOptimalBinaryDecisionTreesIsNPComplete.pdf)

​[Tidy and 3rd Normal Form](https://priceonomics.com/hadley-wickham-the-man-who-revolutionized-r/)

# Lesson 2: SQLite, RSQLite, and Basic SQL Queries

<https://youtu.be/0SoPZSfvLcY>

<https://youtu.be/l8odlDm540o>

<https://www.youtube.com/watch?v=OlT3FispsMU>

* [​](https://sqlite.org/download.html)
* [Links to an external site.](https://sqlite.org/download.html)
* ​[SQLite 3 Database and Tools Download Site](https://sqlite.org/download.html)
* [Links to an external site.](https://sqlite.org/download.html)
* [SQLite 3 Commands](http://zetcode.com/db/sqlite/tool/)
* [Links to an external site.](http://zetcode.com/db/sqlite/tool/)
* [Wickham, H. RSQLite Quick Tutorial](https://cran.r-project.org/web/packages/RSQLite/vignettes/RSQLite.html)
* [Links to an external site.](https://cran.r-project.org/web/packages/RSQLite/vignettes/RSQLite.html)

SELECT list of colums

FROM list of tables

JOIN table name ON condition

JOIN other table ON other condition

WHERE condition for keeping each row

GROUP BY list of columns

HAVING condition for keeping each group

ORDER BY list of columns

# SQL Count

What is the difference between:

SELECT Count(\*) FROM User;

SELECT Count(UserID) FROM User;

SELECT Count(UserPoints) FROM User;

Count will count a row if every column in the parentheses is NULL. If any of the columns in the parentheses are not NULL, then the row is counted.

Count(\*) will look at every column for a non-null value. Since a row that is entirely nulls isn't much of a row, this means counting every row.

Count(UserID) will count every row where UserID is non-null. UserID is a primary key, so it must be not null (and unique, but that's not relevant here). Since it cannot be null, this will count every row.

Count(UserPoints) will count only the rows where UserPoints is not null, which means that some rows might not get counted (the ones that have a null UserPoints).

Be careful about columns that don't have null in your dataset versus columns that are constrained to be not null. What you think is true about the data might change on you, and the R that you wrote a year ago might start doing weird things when you feed it this month's data.

# Constraints vs. Expectations

I will start with a joke, one that is actually relevant to this topic.

At the end of a long week grading papers, two philosophy professors are sitting on a hill, watching sheep walk by. One comments:

"That sheep is gray."

The other responds, "Actually, we can only say that the sheep is gray on one side."

\The first smiles. "Only partly true. We can only say that the sheep is gray on one side on Fridays."

[pause for laughter] One of the reasons that SQL is used for transactional data is that it supports constraints. If you try to insert or update data to violate a constraint, an error is immediately thrown. Primary key constraints make sure that every primary key is unique and not null. Foreign key constraints make sure that every value in that foreign key column references an existing value. Unique constraints make sure you cannot insert or update a value that is not unique in that column. Check constraints make sure that every value in that column obey the constraint as defined.

This is a good way to make sure you don't end up with a purchase transcation for user 635 is there is no user 635.

When you're working in R, you might make certain assumptions about data, and you might rely upon constraints in the SQL database from which you grabbed your data, but these are not the same thing. For example, we know that UserID is unique if it is a primary key. We can trust that. We might assume that UserName is unique because who would put in two users with the same username? That's a bad assumption, and most of us would be aware of that. You might write a query to see if UserName is unique, but then we end up with a dangerous situation. We write R code (possibly with some SQL embedded) that depends on UserName to be unique for the code to work properly. Remember the sheep. UserName might be unique in this data, but if it does not have a unique constraint, then it might not be unique in tomorrow's data. Or next year's data, by which point you have long forgotten about the very subtle bug you introduced based on your uniqueness assumption.

Moral of the story: Your data might happen to follow certain patterns today, but don't write code that counts on that unless there is a SQL constraint enforcing it. The sheep is only gray on one side on Fridays.

Module 7 : collecting data from web scraping and API’s

Web scraping concepts

<https://youtu.be/YQC1FmIQO2g>

* [Introduction to HTML](https://www.w3schools.com/html/html_intro.asp)
* [Links to an external site.](https://www.w3schools.com/html/html_intro.asp)
* [Wikipedia on Web Scraping](http://en.wikipedia.org/wiki/Web_scraping)

Web scraping pages in R with rvest

<https://youtu.be/iXLTWWvz5eQ>

<https://www.youtube.com/watch?v=jyqnbUNEO00>

Read this [tutorial on Rvest](http://www.computerworld.com/article/2909560/business-intelligence/web-scraping-with-r-and-rvest-includes-video-code.html)

* [Tutorial by Prasad Raut: Scraping Data from Dell with Chrome and](https://prasad-raut.github.io/web-scraping-tutorial/) *[rvest](https://prasad-raut.github.io/web-scraping-tutorial/)*
* [Links to an external site.](https://prasad-raut.github.io/web-scraping-tutorial/)
* .
* ​[XPath Introduction](https://drive.google.com/open?id=1rwMHe49oiIofyGXuKFeDUjBHbTOWbjpWzkVLAMVBl6E)
* [Links to an external site.](https://drive.google.com/open?id=1rwMHe49oiIofyGXuKFeDUjBHbTOWbjpWzkVLAMVBl6E)
* [HTTP, HTML and rvest](http://rpubs.com/ktmud/rvest)
* [Links to an external site.](http://rpubs.com/ktmud/rvest)
* [Web Scraping with rves](http://rpubs.com/ktmud/rvest-example)

Getting data from web API’s

<https://youtu.be/OondjHAFDMQ>

# Case Study: Web Scraping in R for Text Analytics

<https://www.youtube.com/watch?v=4TPL76rF5g8&t=1s>

<https://www.youtube.com/watch?v=4vuw0AsHeGw&t=892s>

<https://www.youtube.com/watch?v=Od8gfNOOS9o>

Module 8 : Building the analytic store

Th role of analytics store

<https://youtu.be/KDuuKR4gL6c>

Read the [overview of NoSQL databases by Tim Perdue](https://www.lifewire.com/nosql-an-overview-of-nosql-databases-2495393)

Database genres

<https://www.youtube.com/watch?v=uD3p_rZPBUQ>

After viewing the lecture, read ​​[A Comparison of NoSQL Database Management Systems & Models](https://www.digitalocean.com/community/tutorials/a-comparison-of-nosql-database-management-systems-and-models)

[Links to an external site.](https://www.digitalocean.com/community/tutorials/a-comparison-of-nosql-database-management-systems-and-models)

and then read [Sadalage, P. (2014). NoSQL Databases: An Overview. ThoughtWord, Oct 3, 2014](http://www.thoughtworks.com/insights/blog/nosql-databases-overview)

* ​[Peng, Roger. Interacting with data using the filehash package for R](https://cran.r-project.org/web/packages/filehash/vignettes/filehash.pdf)
* [Links to an external site.](https://cran.r-project.org/web/packages/filehash/vignettes/filehash.pdf)

Key value database

<https://youtu.be/hz1DvAMgfms>

Columnar database : Columnar database are, as the name implies, column-oriented rather than the relational models row-orientation. It is relatively easy and computationally inexpensive to add new columns of data. Each record is identified through a key and each record contains zero or more columns (fields). The columns do not have to be of the same data type across rows and, in fact, may have different numbers of columns making sparse data efficient to store. Columnar databases are about half-way between relational and key-value databases. HBase, along with Cassandra and Hypertable, are among the most popular columnar database implementations.

Document and graph database : Document databases store their information in "document" structures. Documents are nested structured generally stored as JSON or XML objects as those structures provide inherent nesting capability. A document database imposes few formal structures as long as the information can be expressed as a "document". Most document database are queried using JavaScript rather than SQL. Many of them, such as *MongoDB*, support fast concurrent query execution using **MapReduce**. *CouchDBs* major strength is in its resilience to hardware and network failure making it an ideal choice for mobile environments.

Graph databases are less commonly used compared to key-value, columnar, and document databases. However, they are an attractive choice for storing "network" information such as that occurring in "social networks". Graph databases store nodes of information with relationships between the nodes. One of the most popular graph databases is Neo4J. Neo4J allows extremely fast traversal of networks to find relevant information.

Introduction to mongodb <https://youtu.be/Z1wwRrrhl_k>

[Example Code: Accessing MongoDB in R](https://da5020.weebly.com/uploads/8/6/5/9/8659576/examplemongo.r)

Module 9 : Forecasting future values

Intriduction to time series forecasting

<https://youtu.be/Pc1E301z1es>

<https://youtu.be/8GIavF4TbLY>

<https://youtu.be/dF3aYF5qI3w>

Basic forecasting : worked example

<https://youtu.be/a3qMCXW8lOI>

Forecasting with linear regression

<https://youtu.be/ZP0M5ESMrio>

<https://youtu.be/1L2CcAvJ85A>

Module evaluation and comparison

<https://youtu.be/_zMfRDy13TM>

[Woolridge, Jeffrey (2015). Introductory Econometrics. Chapters 10 - 12 on Time Series Regression](http://www.eco.uc3m.es/jgonzalo/teaching/TecnicasEconometricas/WooldridgeCh10-12.pdf)

[Links to an external site.](http://www.eco.uc3m.es/jgonzalo/teaching/TecnicasEconometricas/WooldridgeCh10-12.pdf)

Prediction interval and seasonality

<https://youtu.be/qgx7tdyJ3Jg>

<https://youtu.be/EbLJO1kM3jI>

Module 10 : casual predictive models

Linear regreesion and correlation

<https://youtu.be/KZM1rN4_PLU>

Linear regression

<https://youtu.be/2Xl1EbLnjQc>

<https://youtu.be/CvU-BlDTEqc>

<https://www.youtube.com/watch?v=RaBVUXvmUYI>

Additive cyclicality adjusted models <https://youtu.be/FHI1-zm34Io>

Feature selection <https://www.youtube.com/watch?v=HP3RhjLhRjY>

Dummy coding

Module 11: machine learning with knn

Understanding nearest neighbour classification

<https://youtu.be/zst_kuQjehA>

<https://youtu.be/D11wB7pGCwg>

Normalizing features and distance measure

<https://youtu.be/MebK48EvWYs>

<https://youtu.be/uis8J_2iXsw>

# Guest Lecture: k Nearest Neighbor Classification Algorithm <https://www.youtube.com/watch?v=4ObVzTuFivY>

# Lesson 4: Worked Example II: Using kNN from the caret Package

* Work through the example presented in this [tutorial on kNN Using caret R package](https://rpubs.com/njvijay/16444)
* [Links to an external site.](https://rpubs.com/njvijay/16444)
* using the [Wine dataset](https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data)
* [Links to an external site.](https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data)
* . Ensure that you install the [***caret*** package](http://topepo.github.io/caret/index.html)
* [Links to an external site.](http://topepo.github.io/caret/index.html)
* in your R environment before you work through the code. Here's some code you can use to download the data into R directly from the URL:

dataurl <- "https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data"  
download.file(url = dataurl, destfile = "wine.data")  
wine\_df <- read.csv("wine.data", header = FALSE)

<https://www.youtube.com/watch?v=7Jbb2ItbTC4>