DATA COLLECTION AND DATA PROCESSING

ADVANCED DATA SCIENCE TRAINING I

"People resist a census, but give them a profile page and they'll spend all day telling you who they are."

Max Berry, Lexicon

OUTLINE

1. What Data To Collect: Sampling Theory and Study Design
2. Modern Data Collection: APIs and Web Scraping
3. Working with your Data: Data Wrangling
4. Getting Ready for Analysis: Data Cleaning
5. Making Your Data (More) Manageable: Data Transformation
6. Ensuring Good Data: Data Quality and Data Validation

REFERENCES

DATA COLLECTION AND DATA PROCESSING

REFERENCES

Chapman, A. [2005], *Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data*, Report for the Global Biodiversity Information Facility, Copenhagen.

van Buuren, S. [2012], *Flexible Imputation of Missing Data*, CRC Press, Boca Raton.

Orchard, T. and Woodbury, M. [1972], *A Missing Information Principle: Theory and Applications,* Proc. Sixth Berkeley Symp. on Math. Statist. and Prob., Berkeley.

Hagiwara, S. [2012], *Nonresponse Error in Survey Sampling – Comparison of Different Imputation Methods*, Honours Thesis, Carleton University, Ottawa.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J. and Solenberger, P. [2001], *A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models*, Survey Methodology, v.27, n.1, pp.85-95, Statistics Canada, Catalogue no. 12-001.

Survey Methods and Practices, Statistics Canada, Catalogue no.12-587-X.

REFERENCES

Rubin, D.B. [1987], *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Kutner, M., Nachtsheim, C., Neter, J. and Li, W. [2004], *Applied Linear Statistical Models*, 5th ed., McGraw-Hill/Irwin, New York.

Green, S. and Salkind, N. [2011], *Using SPSS for Windows and Macintosh – Analyzing and Understanding Data*, 6th ed., Prentice Hall, Upper Saddle River.

Wikipedia entry for Data Cleansing

Wikipedia entry for Imputation

Wikipedia entry for Outliers

Torgo, L. [2017], *Data Mining with R* (2nd edition), CRC Press.

McCallum, Q.E. [2013], *Bad Data Handbook*, O'Reilly.

REFERENCES

Kazil, J., Jarmul, K. [2016], *Data Wrangling with Python*, O'Reilly

de Jonge, E., van der Loo, M. [2013], *An Introduction to Data Cleaning with R*, Statistics Netherlands.

Pyle, D. [1999], *Data Preparation for Data Mining*, Morgan Kaufmann Publishers.

Weiss, S.M., Indurkhya, I. [1999], *Predictive Data Mining: A Practical Guide*, Morgan Kaufmann Publishers.

Buttrey, S.E. [2017], *A Data Scientist's Guide to Acquiring, Cleaning, and Managing Data in R*, Wiley.

Aggarwal, C.C. [2013], *Outlier Analysis*, Springer.

Chandola, V., Banerjee, A., Kumar, V. [2007], *Outlier detection: a survey*, Technical Report TR 07-017, Department of Computer Science and Engineering, University of Minnesota.

Hodge, V., Austin, J. [2004], A survey of outlier detection methodologies, *Artif.Intell.Rev.,* 22(2):85–126.

REFERENCES

Feng, L., Nowak, G., Welsh, A.H., O'Neill, T. [2014], *imputeR: a general imputation framework in R*.

Steiger, J.H. , Transformations to Linearity, lecture notes.

Wood, F., Remedial Measures Wrap-Up and Transformations, lecture notes.

Dougherty, J., Kohavi, R., Sahami, M. [1995], Supervised and unsupervised discretization of continuous features, in *Machine Learning: Proceedings of the Twelfth International Conference*, Prieditis, A., Russell, S. (eds), Morgan Kaufmann Publishers.

Orchard, T., Woodbury, M. [1972], A Missing Information Principle: Theory and Applications, Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.

Height Percentile Calculator, by Age and Country, https://tall.life/height-percentile-calculator-age-country/

Dua, D., Karra Taniskidou, E. [2017], Liver Disorders dataset, UCI Machine Learning Repository.

REFERENCES

http://www.roymfrancis.com/scraping-instagram-choosing-hashtags/

Munzert, S., Rubba, C., Meissner, P., Nyhuis, D. [2015], *Automated Data Collection with R, A Practical Guide to Web Scraping and Text Mining*; Wiley

Mitchell, R. [2015], *Web Scraping with Python: Collecting Data From the Modern Web*, O'Reilly.

https://www.w3schools.com/xml/xpath_intro.asp

https://www.w3schools.com/

https://en.wikipedia.org/wiki/XHTML

https://medium.com/the-andela-way/introduction-to-web-scraping-using-selenium-7ec377a8cf72

https://pypi.python.org/pypi/selenium

REFERENCES

Guyon, I., Elisseeff, A., An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, 3(Mar):1157-1182, 2003.

Cawley, G.C., Talbot, N.L.C., Gene selection in cancer classification using sparse logistic regression with Bayesian regularization, *Bioinformatics*, (2006) 22 (19): 2348-2355.

Ambroise, C., McLachlan, G.J., Selection bias in gene extraction on the basis of microarray gene-expression data, *PNAS*, vol.99, no.10, pp.6562–6566, 2002.

Liu, H., Motoda, H. (eds), *Computational Methods of Feature Selection*, Chapman Hall/ CRC Press.

Kononenko, I., Kukar, M. [2007], *Machine Learning and Data Mining: Introduction to Principles and Algorithms*, ch.6, Horwood Publishing.

Lasso (statistics) on Wikipedia

Aggarwal, C.C. [2016], *Data Mining: the Textbook*, sec. 2.4.3, Springer.

REFERENCES

Robnik-Sikonja, M., Savicky, P., CORElearn package documentation, v1.51.2, CRAN.

Ng, A., Soo, K., Principal Component Analysis Tutorial, June 15, 2016.

Principal component analysis, on Wikipedia

Hastie, T., Tibshirani, R., Friedman, J. [2009], *The Elements of Statistical Learning (2nd ed.)*, ch.2, Springer.

Smith, L.I. [2002], A Tutorial on Principal Component Analysis

Shlens, J. [2014], A Tutorial on Principal Component Analysis, arXiv.org

Nonlinear dimensionality reduction, on Wikipedia

J. Leskovec, A. Rajaraman, J. Ullman [2015] *Mining of Massive Datasets*, Cambridge University Press.

REFERENCES

Skillicorn, D. [2007], *Understanding Complex Datasets: Data Mining with Matrix Decomposition*, Chapman and Hall/CRC Press.

CORElearn documentation

Feature selection, on Wikipedia

https://simplystatistics.org/2014/10/24/an-interactive-visualization-to-teach-about-the-curse-of-dimensionality/

Grolemund, G. [2015], *Data Wrangling with R: how to work with the structures of your data*, webinar, bit.ly/wrangling-webinar

https://www.rstudio.com/resources/cheatsheets/

Farrell, P., *STAT 4502 Survey Sampling Course Package*, Carleton University, Fall 2008

REFERENCES

Lessler, J. and Kalsbeek, W. [1992], *Nonsampling Errors in Surveys*, Wiley, New York

Oppenheim, N. [1992], *Questionnaire Design, Interviewing, and Attitude Measurement*, St. Martin's

Hidiroglou, M., Drew, J. and Gray, G. [1993], "A Framework for Measuring and Reducing non-response in Surveys," Survey Methodology, v.19, n.1, pp.81-94

Gower, A. [1994], "Questionnaire Design for Business Surveys," *Survey Methodology*, v.20, n.2

Survey Methods and Practices, Statistics Canada, Catalogue no.12-587-X

Boily, P., Schellinck, J., Hagiwara, S., *et al.* [in preparation], *Introduction to Quantitative Consulting.*