NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

PREDICT 420: Database Systems and Data Preparation Summer 2017

Instructor Name

Atef Bader,PhD

a-bader@northwestern.edu

Course Description

Behind every analytics project is an analytical data source. In this course, students explore the fundamentals of data management and data preparation. Students acquire hands-on experience with various data file formats, working with quantitative data and text, relational (SQL) database systems, and NoSQL database systems. They access, organize, clean, prepare, transform, and explore data, using database shells, query and scripting languages, and analytical software. This is a case-study- and project-based course with a strong programming component.

Prerequisites

PREDICT 400: Math for Modelers for students entering the MSPA program after fall 2014.

Learning Goals

- Articulate analytics as a core strategy using examples of successful predictive modeling/data mining applications in various industries.
- Formulate and manage plans to address business issues with analytics.
- Define key terms, concepts and issues in data management and database management systems with respect to predictive modeling
- Evaluate the constraints, limitations and structure of data through data cleansing, preparation and exploratory analysis to create an analytical database.
- Use object-oriented scripting software for data preparation.
- Transform data into actionable insights through data exploration.

Textbooks

Lubanovic, B. (2015). *Introducing Python: Modern Computing in Simple Packages.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-35936-2]

McKinney, W. (2013) *Python for Data Analysis: Agile Tools for Real-World Data*. Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-31979-3]

Osborne, J. W. (2013). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data.* Thousand Oaks, Calif.: Sage.
[ISBN-13: 978-1-4129-8801-8]

Reference Books (Recommended)

Barrett, D. J. (2012) *Linux Pocket Guide* (2nd ed.). Sebastopol, Calif.: O'Reilly.
[ISBN-13: 978-1-449-31669-3]

Gennick, J. (2011). *SQL Pocket Guide* (3rd ed.). Sebastopol, Calif.: O'Reilly. .
[ISBN-13: 978-1-449-39409-7]
Lutz, M. (2014). Python Pocket Reference (5th ed.). Sebastopol, Calif.: O'Reilly.
[ISBN-13: 978-1-449-35701-6]

Software and Systems (Python)

No software purchases are needed for this course. The course utilizes software that is freely available for PC/Windows, Mac/OS X, and Linux systems.

The course uses the Canopy integrated development environment for Python, which runs on Windows, Mac, and Linux systems. A free academic version of Canopy is available from Enthought, Inc. Follow these steps to obtain the Full Installer for Canopy.

(1) Go to https://www.enthought.com/products/canopy/academic and click on "Request your Academic License"

(2) Follow the instructions, starting by signing up for an Enthought account if you have not already done so. Be sure to use your Northwestern University e-mail address for your account, so that you will be automatically granted a free Academic License. You will need to be logged in with this license to download the Full Academic Installer.

If you have a problem obtaining an account, send a message to

epd.academic@enthought.com

(3) Download the full free academic version of Canopy (do NOT install the Express free version):

https://www.enthought.com/downloads/

The Full Installer includes over 150 Python modules in addition to a Python-aware editor and the IPython shell.

(4) For installation and startup instructions, refer to the Canopy documentation at

http://docs.enthought.com/canopy/quick-start.html

When starting Canopy the first time, if you are not already using another Python on your computer, then you should select for Canopy to be the default Python on your computer, so that you can invoke Python from the terminal/command line as well as through the Canopy integrated development environment. If you are already using a different Python on your computer, see this article to help you decide whether to select Canopy as your default Python:

https://support.enthought.com/entries/23646538-Make-Canopy-User-Python-be-your-default-Python

(5) After you start Canopy for the first time, login with your Enthought credentials so that you get appropriate access down the road (support, package updates...).

(6) If you have questions or problems, your first line of support is the Enthought knowledge base:

https://support.enthought.com/home

There are many sources for Python training. In addition to Python-focused textbooks, we use online training and tutorials in scientific Python from Enthought, Inc., available through the Canopy platform. Here are four very good books for learning more about Python:

Beazley, D. (2009). *Python Essential Reference* (4th ed.). Upper Saddle River, N.J.: Pearson/Addison-Wesley.

Beazley, D. & Jones, B. K. (2013). *Python Cookbook* (3rd ed.). Sebastopol, Calif.: O'Reilly.

Chun, W. J. (2007). *Core Python Programming* (2nd ed.). Upper Saddle River, N.J.: Prentice Hall.

Hellmann, D. (2011). *The Python Standard Library by Example.* Upper Saddle River, N.J.: Pearson/Addison-Wesley. [ISBN-13: 978-0-321-76734-9]

A useful overview of the world of Python, *The Hitchhiker's Guide to Python* by Kenneth Reitz, is available online at <http://docs.python-guide.org/en/latest/>.

Useful Python documenta

tion, coding examples and advice sites include <https://www.python.org/>, <http://stackexchange.com/>, and <http://nullege.com/.>.

Software and Systems (Linux)

This course utilizes Linux-based servers for analytics and databases. The database servers may be accessed directly from your personal computer or from the Social Sciences Computing Cluster (SSCC) analytics servers.

The SSCC is a number of Linux computers in Evanston, Illinois with a wide variety of statistical software. It serves as a research facility for graduate students in the social sciences and business. The School of Professional Studies (SPS) has joined in the support of this facility so that its graduate students (especially those in Predictive Analytics) have access to the extensive software that it provides. General information about the SSCC is available at http://sscc.northwestern.edu.

As a student in PREDICT 420, you will be given an account on the SSCC. You will receive an e-mail from the SSCC system manager early in the term. This e-mail will provide a link for setting up your account. Going thru that link makes database entries that lead to the account creation. The process also puts the applicant thru the usage agreement process. Students should identify themselves as graduate students in SPS.

SSCC accounts are not associated with individual courses or instructors. They are for your use only and remain available as long as you maintain a valid Northwestern NetID. Do not share your SSCC account by giving your NetID and password to others. Your SSCC account is tied to your network identity. Communication between the student and SSCC management is through Northwestern University e-mail.

SSCC account holders have the ability to store files on their user accounts. SPS Predictive Analytics has its own log-in host: dornick, so use dornick for your work, rather than seldon or hardin. Files can be uploaded and downloaded using software tools employing secure file transfer protocol (SFTP). Filezilla, available for PC/Windows and Mac/OSX computers, is a free public-domain tool for file transfer. Additional information about file transfers is available at

http://www.it.northwestern.edu/research/sscc/filetransfer.html.

Using the SSCC means using Linux because the statistical software programs reside on the SSCC, not on the students' personal computers. Essential Linux commands include those for directory and file management, such as ls, mkdir, cd, and cp. Remember that Linux is case-sensitive. Students have direct access to the Social Sciences Computing Cluster (SSCC) for analytics and to PostgreSQL database servers at Northwestern University. Database servers utilize Red Had Linux (RHEL 6) and the most recent production versions of PostgreSQL supported on that Linux distribution.

In addition to the *Linux Pocket Guide* used as a reference book, there are many sources for learning more about Linux. Here is a good introduction to the Linux operating system:

Ward, D. (2015). *How Linux Works: What Every Superuser Should Know* (2nd ed.). San Francisco: No Starch Press. [ISBN-13: 978-1-59327-567-6]

Evaluation

The student's final grade will be determined as follows:

- Week 1:
    - Python Exercise #1 (30 points)
- Week 2:
    - Python Exercise #2 (30 points)
- Week 3:
    - Python Exercise #3 (30 points)
    - Individual Assignment 1 (50 points)
- Week 4:
    - Python Exercise #4 (30 points)
    - Individual Assignment 2 (50 points)
- Week 5:
    - Python Exercise #5 (30 points)
    - Individual Assignment 3 (100 points)
- Week 6:
    - Python Exercise #6 (30 points)
    - Individual Assignment 4 (50 points)
- Week 7:
    - Python Exercise #7 (30 points)
    - Individual Assignment 5 (50 points)
- Week 8:
    - Python Exercise #8 (30 points)
    - Individual Assignment 6 (50 points)
- Week 9:
    - Python Exercise #9 (30 points)
- Week 10:
    - Final Project Report (180 points)
- Week 1–10 Discussion Participation (200 points)

Total: 1000 pts.

Grading Scale

A = 93%–100%
A- = 90%–92%
B+ = 87%–89
B = 83%–86%
B- = 80%–82%
C+ = 77%–79%
C = 73%–76%
C- = 70%–72%
F = 0%–69%

Discussions Etiquette

The purpose of the discussions is to allow students to freely exchange ideas. It is imperative to remain respectful of all viewpoints and positions and, when necessary, agree to respectfully disagree. While active and frequent participation is encouraged, cluttering discussions with inappropriate, irrelevant, or insignificant material will not earn additional points and may result in receiving less than full credit. Frequency is not unimportant, but the content of the message is paramount. All posted work should be original. Post Web links to the work of others; do not post images or files obtained from the others. Please remember to cite all sources. Discussion activity is graded week-by-week, with the end of each week being Sunday at 11:55 p.m. Central Time.

Discussion grading is associated with participation and student-to-student interaction in forums associated with week-by-week course content.

Attendance

This course will not meet at a particular time each week. All course goals, weekly learning objectives, and assessments are supported through classroom elements that can be accessed at any time. To measure class participation (or attendance), your participation in discussion is required, graded, and paramount to your success in this class. Please note that any scheduled synchronous or “live” meetings are considered supplemental and optional. While your attendance is highly encouraged, it is not required and you will not be graded on your attendance or participation.

Late Work

All assignments must be submitted in Canvas before the assigned due date. If a student turns in an assignment less than or equal to one week late, 50% credit will be deducted from the total score. Assignments turned in more than one week late will not receive credit. In the case of unexpected events, students must contact the instructor three days before the assignment due date in order to receive a one-week grace period. Students can only receive up to two grace periods in the course.

Learning Work Groups

Learning work groups are utilized in this course. Additional information about work groups will be provided by the instructor on the Canvas course site.

Academic Integrity at Northwestern

Students are required to comply with University regulations regarding academic integrity. If you are in doubt about what constitutes academic dishonesty, speak with your instructor or graduate coordinator before the assignment is due and/or examine the University Web site. Academic dishonesty includes, but is not limited to, cheating on an exam, obtaining an unfair advantage, and plagiarism (e.g., using material from readings without citing or copying another student's paper). Failure to maintain academic integrity will result in a grade sanction, possibly as severe as failing and being required to retake the course, and could lead to a suspension or expulsion from the program. Further penalties may apply. For more information, visit

www.scs.northwestern.edu/student/issues/academic_integrity.cfm

Plagiarism is one form of academic dishonesty. Students can familiarize themselves with the definition and examples of plagiarism, by visiting <www.northwestern.edu/uacc/plagiar.html>. A myriad of other sources can be found online. Some assignments in this course may be required to be submitted through SafeAssign, a plagiarism detection and education tool. You can find an explanation of the tool at <http://wiki.safeassign.com/display/SAFE/How+Does+SafeAssign+Work>. In brief, SafeAssign compares the submitted assignment to millions of documents in large databases. It then generates a report showing the extent to which text within a paper is similar to pre-existing sources. The user can see how or whether the flagged text is appropriately cited. SafeAssign also returns a percentage score, indicating the percentage of the submitted paper that is similar or identical to pre-existing sources. High scores are not necessarily bad, nor do they necessarily indicate plagiarism, since the score does not take into account how or whether material is cited. If a paper consisted of one long quote that was cited appropriately, it would score 100%. This would not be plagiarism, due to the appropriate citation. However, submitting one long quote would probably be a poor paper. Low scores are not necessarily good, nor do they necessarily indicate a lack of plagiarism. If a 50-page paper contained all original material, except for one short quote that was not cited, it might score around 1%. But, not citing a quotation is still plagiarism.

SafeAssign includes an option in which the student can submit a paper and see the resultant report before submitting a final copy to the instructor. This ideally will help students better understand and avoid plagiarism.

Other Processes and Policies

Please refer to your SPS student handbook at <www.scs.northwestern.edu/grad/information/handbook.cfm> for additional course and program processes and policies.

Course Schedule

Important Note: Changes may occur to the syllabus at the instructor's discretion. When changes are made, students will be notified via an announcement in Canvas.

Week 1: Introducing Software and Systems

Learning Objectives

After this week the student will be able to:

- Identify properties of open-source scripting languages.
- Install an integrated development environment (IDE) for editing and executing programs/scripts.

Course Content

Textbook Readings

Lubanovic, B. (2015). *Introducing Python: Modern Computing in Simple Packages.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-35936-2]

Chapter 1: A Taste of Py (pages 1–14) and Chapter 2: Py Ingredients: Numbers, Strings, and Variables (pages 15–39)

McKinney, W. (2013) *Python for Data Analysis: Agile Tools for Real-World Data*. Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-31979-3]

Chapter 1: Preliminaries (pages 1–6) [Ignore the discussion about installation because we are using Enthought Canopy.] and Chapter 3: IPython: An Interactive Computing and Development Environment (pages 45–78)

Course Reserves

Connolly, T. M. and Begg, C. E. (2015). *Database Systems: A Practical Approach to Design, Implementation, and Management* (6th ed.). Upper Saddle River, N.J.: Pearson. [ISBN-13: 978-0-13-294326-0]

Chapter 1: Introduction to Databases (pages 3–33)

Ward, D. (2015). *How Linux Works: What Every Superuser Should Know* (2nd ed.). San Francisco: No Starch Press. [ISBN-13: 978-1-59327-567-6]

Chapter 1: The Big Picture (pages 1–10) and

Chapter 2: Basic Commands and Directory Hierarchy (pages 11–43)

Software Exercises

Download and install Enthought Canopy.
Python Practice: Getting Started

Discussions

Each week you are required to participate in discussion. Your participation in both posting comments and responding to other students' comments is graded. For this week’s discussion topic, visit the discussions in Canvas. Participation graded through the end of the week, Sunday at 11:55 p.m.

Week 2: Working with Files

Learning Objectives

After this week the student will be able to:

- Distinguish among file formats: plain text, comma-delimited text, JSON, XML.
- Access files within hierarchical structures on remote and local systems.
- Read and write files on remote and local systems.
- Transform from one file format to another on remote and local systems.

Course Content

Textbook Reading

Lubanovic, B. (2015). *Introducing Python: Modern Computing in Simple Packages.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-35936-2]

Chapter 3: Py Filling: Lists, Tuples, Dictionaries, and Sets (pages 41–67) and Chapter 8: Data Has to Go Somewhere (pages 173–193 up to section on relational databases)

McKinney, W. (2013) *Python for Data Analysis: Agile Tools for Real-World Data*. Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-31979-3]

Chapter 5: Getting Started with pandas (pages 111–154) and Chapter 6: Data Loading, Storage, and File Formats (pages 155–174 up to section on interacting with databases)

Course Reserves

Q. Ethan McCallan. (2012). *Bad Data Handbook: Mapping the World of Data Problems.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-32188-8]

Chapter 12 (pages 151–162) When Databases Attack: A Guide to When to Stick to Files.

Software Exercises

Python Practice: Reading and Writing Data Files, Beginning pandas

Discussions

Each week you are required to participate in discussion. Your participation in both posting comments and responding to other students' comments is graded. For this week's discussion topic, visit discussions in Canvas. Participation graded through the end of the week, Sunday at 11:55 p.m.

Week 3: Understanding Relational Databases

Learning Objectives

After this week the student will be able to:

- Define relational database terms: table, record, schema, index, key, view, normalize.
- Describe the functions that a database management system should provide.
- Explain why relational database systems are well suited for transaction processing.
- Access data from a relational database using an administrative shell.

Course Content

Textbook Reading

Lubanovic, B. (2015). *Introducing Python: Modern Computing in Simple Packages.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-35936-2]

Chapter 4: Py Crust: Code Structures (pages 69–107), Chapter 8: Data Has to Go Somewhere (pages 193–216), and Chapter 10: Systems (pages 241–259)

McKinney, W. (2013) *Python for Data Analysis: Agile Tools for Real-World Data*. Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-31979-3]

Chapter 6: Data Loading, Storage, and File Formats (pages 174–176)

Course Reserves

Connolly, T. M. and Begg, C. E. (2015). *Database Systems: A Practical Approach to Design, Implementation, and Management* (6th ed.). Upper Saddle River, N.J.: Pearson. [ISBN-13: 978-0-13-294326-0]

Chapter 4: The Relational Model (pages 101–118) and Chapter 22: Transaction Management (pages 619–623 defining transactions and their properties)

Course Reserves (Recommended)

Obe, R. and Hsu, L. (2012). *PostgreSQL Up and Running: A Practical Guide to the Advanced Open Source Database.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-32633-3]

Preface (pages ix–xii) and Chapter 1: The Basics (pages 1–8)

Worsley, J. C. and Drake, J. D. (2002). *Practical PostgreSQL.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-156592846-6]

Chapter 3: Understanding SQL (pages 33–89)

Software Exercises

Python Practice: SQLite, SQLAlchemy, Database Queries

Discussions

Each week you are required to participate in discussion. Your participation in both posting and responding to other students' comments is graded. For this week's discussion, visit discussions in Canvas. Participation graded through the end of the week, Sunday at 11:55 p.m.

Assignment

Individual Assignment 1 (50 points) due Sunday, at 11:55 p.m. Central Time

Week 4: Accessing and Manipulating Relational Data

Learning Objectives

After this week the student will be able to:

- Access data from a relational database using an object-oriented scripting language.
- Define structured query language (SQL) and explain its importance in working with databases.
- Select data from a relational database based on particular criteria.
- Join database tables to form views and tables.

Course Content

Textbook Readings

Lubanovic, B. (2015). *Introducing Python: Modern Computing in Simple Packages.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-35936-2]

Chapter 5: Py Boxes: Modules, Packages, and Programs (pages 109–122)

Course Reserves (Recommended)

Hellmann, D. (2011). *The Python Standard Library by Example.* Upper Saddle River, N.J.: Pearson/Addison-Wesley. [ISBN-13: 978-0-321-76734-9]

Chapter 7: Data Persistence and Exchange (pages 333–420)

Worsley, J. C. and Drake, J. D. (2002). *Practical PostgreSQL.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-156592846-6]

Chapter 4: Using SQL with PostgreSQL (pages 91–155)

Software Exercises

Python Practice: Merging Data Sources, Database Joins

Discussions

Each week you are required to participate in discussion. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic, visit discussions in Canvas. Participation graded through the end of the week, Sunday at 11:55 p.m.

Assignment

Individual Assignment 2 (50 points) due Sunday, at 11:55 p.m. Central Time

Week 5: Moving Beyond Relational Databases

Learning Objectives

After this week the student will be able to:

- List reasons for moving beyond relational database systems to NoSQL systems.
- Explain what is meant by a distributed file system and MapReduce.
- Explain the core information retrieval concepts used by search and analytics engines
- Access data from a NoSQL database using an administrative shell.
- Access data from a NoSQL database using an object-oriented scripting language.

Course Content

Textbook Reading

Lubanovic, B. (2015). *Introducing Python: Modern Computing in Simple Packages.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-35936-2]

Chapter 6: Oh, Oh: Objects and Classes (pages 123–171)

Course Reserves

Chodorow, K. (2013). *MongoDB: The Definitive Guide* (2nd ed). Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-34468-9]

Chapter 1: Introduction (pages 3–5),
Chapter 2: Getting Started (pages 7–28), and
Chapter 4: Querying (pages 53–77)

Franks, B. (2012). *Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics,* New York: Wiley. [ISBN-13: 978-1-118-20878-6]

Chapter 1: What Is Big Data and Why Does It Matter (pages 3–27) and
Chapter 4: The Evolution of Analytic Scalability (pages 87–119)

Gheorghe, R., Hinman, M.L., and Russo, R. (2016). Elasticsearch in Action. Shetler Island, N.Y.: Manning. [ISBN-13: 978-1617291623]

Chapter 1: Introducing Elasticsearch (pages 3–19)
Chapter 2: Diving into Functionality (pages 20–52)

Course Reserves (Recommended)

Dean, J. and Ghemaway, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM,* 51:1, 107–113.

Rajaraman, A. and Ullman, J. D. (2012). *Mining of Massive Datasets.* Cambridge UK: Cambridge University Press. [ISBN-13: 978-1-107-01535-7]

Chapter 2: Large-Scale File Systems and Map-Reduce (pages 18–52)

Manning, C. D., Raghaven, P., and Schutze, H. (2008) Introduction to Information Retrieval. Cambridge, UK: Cambridge University Press. [ISBN-13: 978-0521865715]

Chapter 4: Index construction
Chapter 6: Scoring, term weighting & the vector space model
Chapter 8: Evaluation in information retrieval

Available online at http://nlp.stanford.edu/IR-book/information-retrieval-book.html

Software Exercises

Python Practice: JSON Data Format, Data Type Transformations

Discussions

Each week you are required to participate in discussion. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic, visit discussions in Canvas. Participation graded through the end of the week, Sunday at 11:55 p.m.

Assignment

Individual Assignment 3 (100 points) due Sunday, at 11:55 p.m. Central Time

Week 6: Accessing and Manipulating Text Data

Learning Objectives

After this week the student will be able to:

- Access and manipulate unstructured and semi-structured text data files.
- Access and manipulate text data within a relational database system.
- Access and manipulate text data within a NoSQL database system.
- Parse text data with regular expressions in an object-oriented scripting language.

Course Content

Textbook Reading

Lubanovic, B. (2015). *Introducing Python: Modern Computing in Simple Packages.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-35936-2]

Chapter 7: Mangle Data Like a Pro (pages 145–171)

Course Reserves

Levy, J. (2011). Bad Data Lurking in Plain Text. In Q. Ethan McCallan, ed., *Bad Data Handbook: Mapping the World of Data Problems.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-32188-8]

Chapter 4 (pages 53–68)

Course Reserves (Recommended)

Hellmann, D. (2011). *The Python Standard Library by Example.* Upper Saddle River, N.J.: Pearson/Addison-Wesley. [ISBN-13: 978-0-321-76734-9]

Chapter 1: Text (pages 3–68)

Software Exercises

Python Practice: Text Parsing, Regular Expressions

Discussions

Each week you are required to participate in discussion. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic, visit discussions in Canvas. Participation graded through the end of the week, Sunday at 11:55 p.m.

Assignment

Individual Assignment 4 (50 points) due Sunday, at 11:55 p.m. Central Time

Week 7: Selecting and Sampling Data

Learning Objectives

After this week the student will be able to:

- Define sampling terms: target population, sampling frame, sample, and representative sample.
- Distinguish among alternative forms of sampling, including random sampling, stratified sampling, and cluster sampling.
- Select data from databases following a sampling scheme.
- Address problems of under-coverage and sampling bias.

Course Content

Textbook Reading

Osborne, J. W. *(2013). Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data.* Thousand Oaks, Calif.: Sage. [ISBN-13: 978-1-4129-8801-8] Chapter 1: Why Data Cleaning Is Important: Debunking the Myth of Robustness (pages 1–16), Chapter 2: Power and Planning for Data Collection: Debunking the Myth of Adequate Power (pages 19–41), Chapter 3: Being True to the Target Population: Debunking the Myth of Representativeness (pages 43–69), and Chapter 4: Using Large Data Sets with Probability Sampling Frameworks: Debunking the Myth of Equality (pages 71–83)

Course Reserves (Recommended)

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology* (2nd ed). New York: Wiley. [ISBN-13: 978-0-470-46546-2]

Chapter 3: Target Populations, Sampling, Frames, and Coverage Error (pages 69–95),

Software Exercises

Python Practice: Indicator Variables, Categorical Data, Tables, Descriptive Statistics, Outliers

Discussions

Each week you are required to participate in discussion. Your participation in both posting and responding to other students' comments is graded. For this week’s discussion topic, visit discussions in Canvas. Participation graded through the end of the week, Sunday at 11:55 p.m.

Assignment

Individual Assignment 5 (50 points) due Sunday, at 11:55 p.m. Central Time

Week 8: Cleaning Data

Learning Objectives

After this week the student will be able to:

- Identify bad data problems.
- Clean and update data items using an object-oriented scripting language.
- Screen data for potential problems, identifying outliers and miscoded data using an object-oriented scripting language.
- Address problems of missing data in surveys and databases.

Course Content

Textbook Reading

McKinney, W. (2013) *Python for Data Analysis: Agile Tools for Real-World Data*. Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-31979-3] Chapter 7: Data Wrangling: Clean, Transform, Merge, Reshape (pages 177–217)

Osborne, J. W. *(2013). Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data.* Thousand Oaks, Calif.: Sage. [ISBN-13: 978-1-4129-8801-8] Chapter 5: Screening Data for Potential Problems: Debunking the Myth of Perfect Data (pages 87–104), Chapter 6: Dealing with Missing or Incomplete Data: Debunking the Myth of Emptiness (pages 105–138), Chapter 7: Extreme and Influential Observations: Debunking the Myth of Equality (pages 139–168), Chapter 12: The Special Challenge of Cleaning Repeated Measures Data: Lots of Pits in Which to Fall (pages 253–259)

Course Reserves (Recommended)

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology* (2nd ed). New York: Wiley. [ISBN-13: 978-0-470-46546-2]

Chapter 10: Postcollection Processing of Survey Data (pages 329–367)

Software Exercises

Python Practice: pandas DataFrame Objects, Heterogeneous Data, Recoding/Mapping Data

Discussions

Each week you are required to participate in discussion. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic, visit discussions in Canvas. Participation graded through the end of the week, Sunday at 11:55 p.m.

Assignment

Individual Assignment 6 (50 points) due Sunday, at 11:55 p.m. Central Time

Week 9: Transforming and Organizing Data

Learning Objectives

After this week the student will be able to:

- Distinguish among cross-sectional, temporal/time series, spatial, panel, and spatio-temporal data.
- Perform data aggregation within an object-oriented scripting language.
- Recode and transform data fields using an object-oriented scripting language.
- Aggregate, group and reorganize data using an object-oriented scripting language.

Course Content

Textbook Readings

McKinney, W. (2013) *Python for Data Analysis: Agile Tools for Real-World Data*. Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-31979-3]

Chapter 9: Data Aggregation and Group Operations (pages 251–288) and
Chapter 10: Time Series (pages 289–328)

Osborne, J. W. *(2013). Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data.* Thousand Oaks, Calif.: Sage. [ISBN-13: 978-1-4129-8801-8]

Chapter 8: Improving the Normality of Variables through Box-Cox Transformations: Debunking the Myth of Distributional Irrelevance (pages 169–190) and
Chapter 11: Why Dichotomizing Continuous Variables is Rarely a Good Practice: Debunking the Myth of Categorization (pages 231–252)

Software Exercises

Python Practice: Data Aggregation, Grouping, Pivot Tables

Discussions

Each week you are required to participate in discussion. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic, visit discussions in Canvas. Participation graded through the end of the week, Sunday at 11:55 p.m.

Week 10: Review

Learning Objectives

No new learning objectives are introduced in this Week.

Course Content

None.

Discussions

Each week you are required to participate in discussion. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic, visit discussions in Canvas. Participation graded through the end of the week, Sunday at 11:55 p.m.

Assignment

Term Project Report (130 points) due Wednesday, at 11:55 p.m. Central Time