LING 580R, Section 3
Linguistic Data Analysis
Fall 2022
MWF 11:00–11:50am
B013 JFSB

Instructors:	Earl Kjar Brown, PhD	Joey Stanley, PhD
office:	4052 JFSB	4059 JFSB
email:	ekbrown@byu.edu	joey_stanley@byu.edu
phone:	801-422-3970	(801) 422-918
office hours:	Wed. & Thurs. 2-3 pm, or by appt.	Wed. and Thurs. 2:00–4:00
Zoom:	Dr. Brown turns on his Zoom mtg during his office hours: https://byu.zoom.us/j/8014223970	http://byu.zoom.us/my/joeystan

Course Purpose:
As a result of taking this course, students will be skilled at manipulating and visualizing data, and performing statistical analyses on linguistic data.

Course Description:
This course offers students an in-depth introduction to data analysis within linguistics. The students will become skilled practitioners of the R programming language, a language specifically designed for data analysis and statistics. The students will manipulate data in tabular datasets (like tables in spreadsheet software) by filtering and arranging rows, selecting columns and creating new ones, among other data manipulation techniques. Students will also create informative visualizations of data, and perform appropriate statistical analyses. Students do not need to have previous computer programming experience, but previous experience will be helpful.

Learning Outcomes:
       Students will:
1. Manipulate and process tabular datasets;
2. Create high-quality and informative visualizations;
3. Produce descriptive summary information about datasets;
4. Perform and interpret appropriate inferential statistical tests.

Materials:
1. The Learning Management System (LMS) Canvas
2. Online textbook R for Data Science
3. Online textbook Regression Modeling for Linguistic Data
4. Online textbooks Fundamentals of Data Visualization and Data Visualization

Readings:

In order to be prepared for class and to have a first approach to new topics, students complete assigned readings before class on the days indicated in the CMS.

Response Papers:
Students write response papers to readings and topics presented in class. The length of these response papers should be between 300 - 600 words; students should include the word count after finishing writing. The formatting of the paper should follow the [Generic Style Rules for Linguistics,](#) which uses the [Unified Style Sheet for Linguistics](#) for the references and in-text citations. Papers turned in late receive a one (1) point penalty for each 24-hour period late, until 7 days after the original due date, after which time the paper is no longer accepted. Exceptions to this late policy must be approved by the instructors **before** the due date. Unless otherwise indicated, response papers should have at least the following four (4) clearly labeled sections: Summary, Connections, Applications, Questions:

- Summary: One or two paragraphs of summary of the reading and/or topic;
- Connections: One or two paragraphs identifying connections to previous topics or to experience or knowledge from outside the course;
- Applications: One or two paragraphs of possible applications of the reading or topic to students' previous or current research projects;
- Questions: One or two questions.

Research Participant:

A small part of your overall grade comes from participating in language-related research in our department. These studies typically take around one hour of your time. You can choose any language-related study that is offering course credit for participation. We will inform you of participation opportunities as they become available. If you are unable or do not want to participate in a study, as an alternative you can attend a tour of the research labs in the Linguistics Department and write a one-page reflection of what you learn and how the research methods discussed could be applied to questions of interest to this course. Details on the scheduling of the tour will be provided later in the semester.

Coding Assignments:
Students complete coding assignments in order to solidify concepts and knowledge learned during the previous week or two. Instructions for the assignments are made available in the LMS. Most commonly, students will submit an R (.r) or R Markdown (.Rmd) file, and a CSV file.

Final Project:
Students complete a final project by analyzing meaningful data on a topic of their choice, perhaps after collecting the data. The final project should stretch the students' coding and data analysis abilities; obviously, the students will use what they learn throughout the semester, but this final project should **not** be simply a redo of previous coding assignments.

Final Exam:
Students take a comprehensive final exam in the CMS during this final exams period. See the university's Final Exam Policy [here.](#)

Late submissions policy:
Students are highly encouraged to start the assignments early in the week in order to have time to work through problems that inevitably occur. Weekly assignments that are turned in late receive a penalty of one point out of ten points per 24-hour period, up to 7 days, after which time the assignment is not accepted. Students are encouraged to start the weekly assignments early in the week in order to have time to work through problems that they encounter.

Grade Components and Final Grade Calculation:
Note: No rounding is applied when calculating the final letter grade, for example, an 89.99% is a B+ while a 90.00% is an A-.

Be a research participant = 2%
Response Papers = 13%
Coding Assignments = 40%
Final Project:
     Oral Presentation = 10%
     Written Paper = 15%
Final Exam = 20%
	93-100% = A
	90-92% = A-
	87-89% = B+
	83-86% = B
	80-82% = B-
	77-79% = C+
		73-76% = C
		70-72% = C-
		67-69% = D+
		63-66% = D
		60-62% = D-
		0-59% = E

Tentative Schedule of Topics:
Note: Check the LMS for homework assignments and their due dates.

Date	Topics
Week 1 Aug 29– Sep 2	- Intro to course and to R and RStudio (cheatsheet here), and R projects - Programming basics (in R): data types, `if else`, `for` loops, defining functions - R structures: data frame (and `tibble`), list, matrix, vector
Week 2 Sep 5–9	- File I/O, especially TXT (slurp and line-by-line), CSV, Excel, SPSS, SAS, Stata, Feather, JSON, XML (cheatsheet here) - Regexes, especially capture groups and lookaround (cheatsheet here) - String manipulation with `stringr` package (cheatsheet here)
Week 3 Sep 12–16	- From TXT files to tabular dataset - Frequency lists from files on hard drive ~~Webscrape frequencies from Davies corpora~~ - Webscrape frequency lists from Wiktionary
Week 4 Sep 19–23	- Sourcing Python modules in R with reticulate - Table joins (e.g., w/ `left_join()`) - Data manipulation with `dplyr` (cheatsheet here) and `tidyr` (cheatsheet here) - create new columns with `mutate` - keep only certain columns with `select`

	○ keep only certain rows with `filter` ○ sort rows with `arrange` (and `desc`) ○ create summary info with `summarise` (and `group_by`) ○ convert long-format datasets to wide with `pivot_wider` ○ convert wide-format datasets to long with `pivot_longer` ○ split a column with `separate` ○ combine several columns into one with `unite`
Week 5 Sep 26–30	● Data manipulation continued ● Fri. Sep. 30th: Guest lecture: Rich Ross (UVA)

Week 6
Oct 3–7
	● Data visualization with `ggplot2` (chapter here, cheatsheet here)
	   ○ barplot
	   ○ scatterplot
	   ○ boxplot
	   ○ faceting
	   ○ smoothing (loess) lines
	   ○ regression lines
	   ○ hierarchical cluster analysis (dendrograms)

Week 7 Oct 10–14	● Data visualization continued ● Combine R code and output into an HTML page with knitr

Week 8
Oct 17–21
	● Praat scripting
	   ○ Intro to the Software
	   ○ I/O, loops, TextGrids
	   ○ Automatic extraction of formants and other acoustic measurements

Week 9
Oct 24–28
	● Sociophonetic Data Processing
	   ○ Forced alignment with Montreal Forced Aligner
	   ○ Outlier detection
	   ○ Vowel normalization

Week 10
Oct 31–
Nov 4
	● Descriptive stats
	   ○ Central tendencies (mean, median, mode)
	   ○ Dispersion (range, standard deviation, variance)
	● Inferential stats
	   ○ t-tests, ANOVA, chi-squared, correlation, non-parametric tests

Week 11 Nov 7–11	● Inferential stats continued ○ Linear regression w/o and w/ random effects
Week 12 Nov 14–18	● Inferential stats continued ○ Logistic regression w/o and w/ random effects

	● Cluster Analysis (hierarchical agglomerative clustering, K-Means) ● Dimensionality Reduction ○ Principal Component Analysis
Week 13 Nov 21–22	● Generalized additive models (GAMs)
Week 14 Nov 28– Dec 2	Self-directed time to work on final project. Your instructors will be in the classroom during class time to help students, as needed.
Week 15 Dec 5–7	Final project oral presentations
Finals Dec 12–16	Final exam: At-home exam taken in the CMS anytime during the finals period (Dec. 12th, 12:00 am - Dec. 16th, 11:59 pm)

Academic Honesty:
The first injunction of the BYU Honor Code is the call to be honest. Students come to the university not only to improve their minds, gain knowledge, and develop skills that will assist them in their life's work, but also to build character. President David O. McKay taught that "character is the highest aim of education" (The Aims of a BYU Education, p. 6). It is the purpose of the BYU Academic Honesty Policy to assist in fulfilling that aim. BYU students should seek to be totally honest in their dealings with others. They should complete their own work and be evaluated based upon that work. They should avoid academic dishonesty and misconduct in all its forms, including but not limited to plagiarism, fabrication or falsification, cheating, and other academic misconduct.

Honor Code:
In keeping with the principles of the BYU Honor Code, students are expected to be honest in all of their academic work. Academic honesty means, most fundamentally, that any work you present as your own must in fact be your own work and not that of another. Violations of this principle may result in a failing grade in the course and additional disciplinary action by the university. Students are also expected to adhere to the Dress and Grooming Standards. Adherence demonstrates respect for yourself and others and ensures an effective learning and working environment. It is the university's expectation, and my own expectation in class, that each student will abide by all Honor Code standards. Please call the Honor Code Office at 422-2847 if you have questions about those standards.

Preventing Sexual Harassment:
Title IX of the Education Amendments of 1972 prohibits sex discrimination against any participant in an educational program or activity that receives federal funds. The act is intended to eliminate sex discrimination in education. Title IX covers discrimination in programs, admissions, activities, and student-to-student sexual harassment. BYU's policy against sexual

harassment extends not only to employees of the university, but to students as well. If you encounter unlawful sexual harassment or gender-based discrimination, please talk to your professor; contact the Equal Employment Office at 422-5895 or 367-5689 (24-hours); or contact the Honor Code Office at 422-2847.

Students with Disabilities:
Brigham Young University is committed to providing a working and learning atmosphere that reasonably accommodates qualified persons with disabilities. If you have any disability which may impair your ability to complete this course successfully, please contact the Services for Students with Disabilities Office (422-2767). Reasonable academic accommodations are reviewed for all students who have qualified, documented disabilities. Services are coordinated with the student and instructor by the SSD Office. If you need assistance or if you feel you have been unlawfully discriminated against on the basis of disability, you may seek resolution through established grievance policy and procedures by contacting the Equal Employment Office at 422-5895, D-285 ASB.