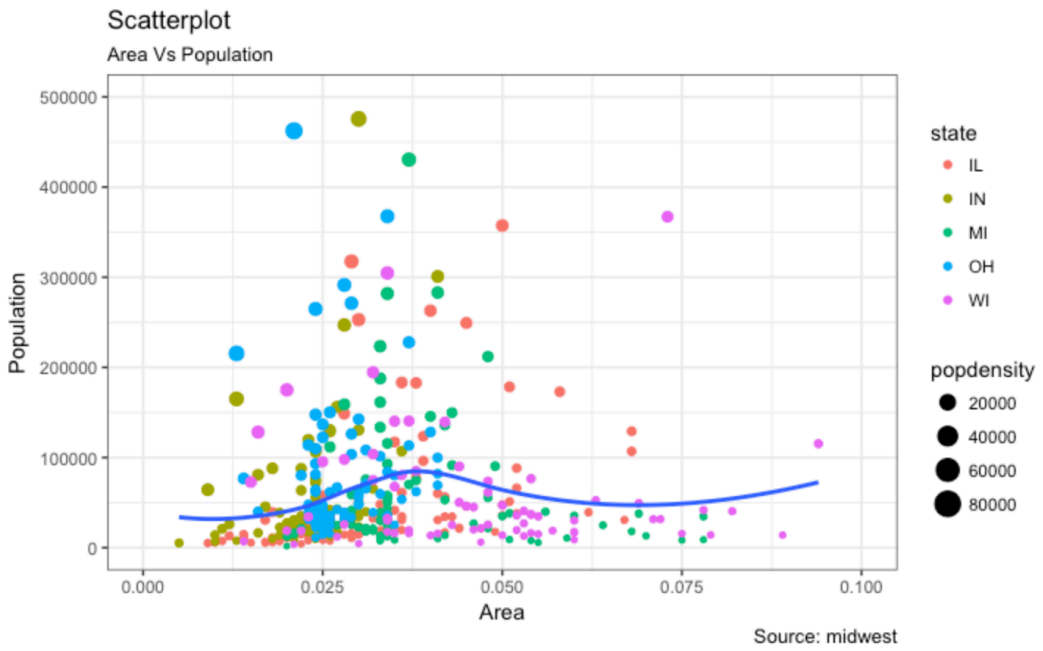
UNIVERSITY OF VIRGINIA SPRING SEMESTER 2019

Analysis of Political Data Using R

PLAD 4500

Tuesday/Thursday 4:00-5:15pm, Nau Hall 142



What is R and What Can it Do?

R, along with Python, is one of two primary coding languages used in the growing field of data science. It's used for statistical research in many fields such as political science.

Data analysis and statistics have been primary tools for research in political science for more than 60 years. In the early days, researchers had to program their analyses on stacks of punch cards and take them to the single and enormous computer on campus, feeding the cards one at a time through a slot and praying that none of the cards get jammed.

Starting about 20 years ago, researchers started to use software that can be operated digitally on a local machine. Most researchers used software named SPSS, SAS, or Stata.

Instructor:
Jonathan Kropko

Office
Gibson Hall 383

Office hours
Fridays, 9am - 11am, and by appointment

Email
jkropko@virginia.edu

Office phone
(434) 924-4660

Please download the following free software as soon as possible, or update them to the latest versions:

1

R

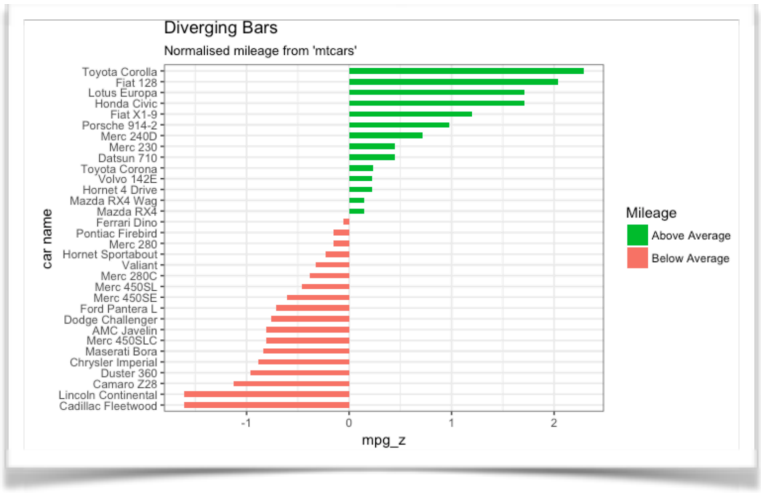
Available from the Comprehensive R Archive Network: https://cran.r-project.org/

2

R STUDIO

An excellent user interface for R available at: https://www.rstudio.com/

The problem with these software packages is that they are proprietary. That means two things. First, it means that these software packages are slow to update and incorporate new statistical methods. Second, it means that the software is *expensive*.



While these software packages are still widely used, things are changing quickly. More and more researchers, both within academia and outside of it, are using open-source platforms for statistical computing. R has become the primary choice for statistics, and python has become the primary choice for other applications in data science.

There are several reasons why R is now our best option. First, R is and will continue to be absolutely 100% free. Second, R produces beautiful, logical, interactive graphics. Third, R is the most widely used software platform by serious statisticians, and when researchers develop a new statistical method, they usually will write a new R library to implement the method and release it to the public. So R's functionality increases at about the same rate as new research in statistics. Fourth, R very recently developed powerful and user-friendly tools for data management and for easily weaving code and results into a text document. Finally, knowledge of R is an important skill that will get the attention of employers.

R PACKAGES

R comes with a large set of commonly used functions for statistics and data management. But for more specialized topics we will need to download a package of additional functions from the CRAN repository. For example, one additional package we will use is *ggplot2*, which is a powerful engine for producing visualizations and graphics. To download this package open RStudio and follow these steps:

1. RStudio's interface includes several different windows. Find the window that says Console in the upper-left corner and click inside this window. You should see a cursor that will allow you to type.

2. Type install.packages("ggplot2") and push enter. The *ggplot2* package will now be downloaded. (You only have to do this one time after updating R)

3. To use the functions inside the *ggplot2* package, type library(ggplot2) in the console. Note the lack of quotation marks. Now you can use the functions inside *ggplot2*. (You will have to do this each time you start an R session to use *ggplot2*)

Course Organization

We are going to build up to cool things, but we have to take our time and start from the basics. That might get frustrating. Early on the problem sets may not be very interesting. Bear with me. It's better than accelerating the course and leaving people behind. This course is divided into five parts:

1. **Introduction to R and RStudio**: we'll get used to the user interface, learn about object oriented programming, and scripts. We'll talk about how to best use the help resources and post on webpages like Stack Overflow. And we'll learn about *knitr,* an R package for creating documents that weave text, code, and the results of the code all in one HTML or PDF document.

2. **Advanced data cleaning techniques**: most data are messy. And no further analysis is possible until the data have been cleaned. We'll work to reformat data to a tidy format by managing the observations and variables, reshaping, and merging datasets together. Data scientists think of this step as 90% of the work.

3. **Descriptive statistics and graphics**: we'll cover the functions to calculate means, medians, variances, quantiles, and correlations, and we will also learn to use *ggplot* and *plotly*, two packages to create beautiful and interactive data visualizations.

4. **Linear and logistic regression**: regressions are tools for assessing whether one variable has an effect on another. We'll learn how to run, interpret, and graph the results of these regressions. These are topics that we devote entire semesters to in graduate school. We'll focus on the basics, but we will also introduce some advanced ideas.

5. **Analyzing text as data**: all of the work we will do will lead up to an introduction to using text as data. In text analysis, we take a document, such as the U.S. Constitution, the presidential state of the union addresses, or speeches by delegates at the UN, and use R to extract information to use as data. This is a cutting edge topic in political science and a core topic in data science.

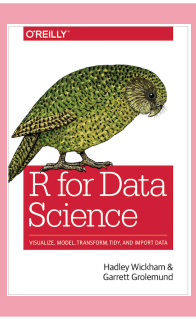
R: THE WILD WEST OF STATS COMPUTING

No one person or company creates R. It's a collective effort of thousands of researchers in many fields around the world. When a researcher develops a new technique to conduct statistics or work with data, he or she writes R code to perform this task and distributes it through CRAN or another repository. That's one of the best things about R.

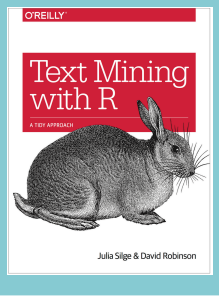
But the drawback is that there are often many, many ways to do the same thing in R. The approaches we will discuss are not the only way to perform the tasks we need to accomplish. Whenever possible, I will try to present the approaches in class that are easiest to teach and to understand, that use fewer lines of code, and run more quickly. You might find another way to do these things, and that's great. But if you stick to the ways we talk about in class, I can more easily follow your work and give you full credit.

Readings

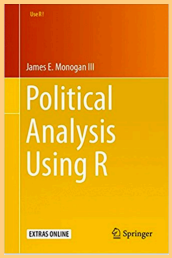
There are three textbooks for this course. Please complete the readings prior to class.



R for Data Science
by Hadley Wickham and Garrett Grolemund
A free, online copy is available here:
http://r4ds.had.co.nz/



Text Mining with R
by Julia Slige and David Robinson,
A free, online copy is available here:
https://www.tidytextmining.com/



Political Analysis Using R
by James E. Monogan
Available through Amazon: https://www.amazon.com/Political-Analysis-Using-James-Monogan/dp/3319234455/ref=sr_1_1?

Course Website

All grades, labs, and readings will be posted on the UVa Collab site for the course, accessible at https://collab.itc.virginia.edu/portal. If you are officially enrolled in the course, the course website should already be accessible to you. If you are not officially enrolled, please speak to me so I can arrange for you to have access to the course material.

Equipment

If you own a laptop computer capable of running R, please bring it to class so that you can work on labs in class. If you do not have a computer capable of running R, let me know so we can make other arrangements.

Assessment

Your grade will be determined by three factors:

- Your performance on ten **problem sets** (60% of the grade)
- Your **final exam** (35% of the grade)
- Completing enough coursework on datacamp.com to earn 10,000 XP by April 30 at 4pm (5% of the grade)

There is no midterm exam in this course. The final grades will be determined from the final percents according to the table on the right. Percents lower than 60 will receive failing grades.

Percent range	The letter grade will be no lower than
> 92%	A
88% - 92%	A-
84% - 88%	B+
80% - 84%	B
76% - 80%	B-
72% - 76%	C+
68% - 72%	C
64% - 68%	C-
60% - 64%	D

Problem Sets

There will be 10 problem sets assigned throughout the semester. These assignments will help you practice the techniques in R we will discuss in class. You will work in **groups of 3** on these assignments. You and your group will write a lab report together using the R Markdown language and the *knitr* package in R, saved as an HTML file, to be submitted on Collab before class on the due date. I will assign groups, and I will rotate who gets grouped with whom throughout the semester. I only need one lab report per group. Labs will be graded out of 10 points each, based on the following criteria:

- Accuracy (4 points) — Are the techniques employed correctly?
- Communication and Completeness (4 points) — is every part of the problem set completed? Are the results communicated clearly and accurately? Does the lab report demonstrate that the authors understand the intuition of the techniques and code?
- Formatting (2 points) — is the problem set report formatted cleanly? How close are the tables and graphics to publishable quality?

Problem set	Assigned	Due date
1	Jan 29	Feb 5
2	Feb 5	Feb 12
3	Feb 12	Feb 19
4	Feb 19	Feb 26
5	Feb 26	March 5
6	March 7	March 21
7	March 28	April 4
8	April 4	April 11
9	April 16	April 23
10	April 23	April 30

Why So Much Group Work?

In industry, and in many academic settings, almost all data work is done collaboratively. People work as parts of teams, and have to divide the work efficiently and help each other to solve problems. Learning how to collaborate is just as important as learning the coding skills. It requires patience and an ability to communicate clearly. Please take this skill seriously, and avoid the

temptations of freeloading off of groupmates' work, or doing all the work by yourself to avoid the necessary challenge of helping your groupmates get up to speed. Although you are free to divide the work for each problem set between the three of you, a better strategy is to work on as much of the lab as you can together because most questions will be impossible to answer without correctly performing the tasks required by previous questions, and because you will need to practice all of the skills on each problem set.

Whether each group member does his or her fair share of the work and whether group members help each other to understand the work is not something I can enforce, or care to. I'm counting on you to help build a culture in this class where our focus is on learning the material as best we can, on helping each other to understand the material.

Final Exam

At the end of the semester you will complete a take-home and open-note final exam. Why take-home and open-note? Because in the real world, you will have access to textbooks, help pages, and the vast expertise of google to solve problems. That said, the exam will be very challenging, and will test you in every skill we cover over the course of the semester. I will give you multiple data sources, some of which may be quite messy. You will need to load, clean, and combine the data, perform an analysis and compile a clear report on the results of the analysis. This project will focus on the principles of the **reproducible research movement** to make data analysis as transparent as possible. I will provide more detail about the exam later in the semester.

Datacamp

You are also responsible for taking online courses on datacamp.com. Datacamp is a for profit company, and their courses are not in general free of charge, but we are using a program designed for college students in which all of you have free access to all of the datacamp courses for the next six months. In the future, you will need to be able to teach yourself the skills you need to use new methods, and websites like datacamp are wonderful resources.

You may take any courses you want, so choose courses that are of interest to you. Many of the courses focus on R, but you can also train yourself in python if you want to. When you complete a course, datacamp awards you "XP" points. You are responsible for obtaining at least 10,000 XP points before the last day of class on April 30. That equates to about two courses, which will take about 8 hours in total. The "Introduction to R" course, for example, takes 4 hours and is worth 6,200 XP. As the professor, I will be able to see how much XP each student has earned, so no additional documentation is needed. Students who meet this requirement will receive full credit for this part of the grade, and students with less than 10,000 XP will receive no credit. Please check as soon as possible to make sure you can log on without problems.

Collaboration and Cheating

The following actions are examples of cheating:

- plagiarizing any part of any problem set or any part of the final exam, including prose, graphs, tables, and equations,

- directly copying answers from another student on the final exam or from another student in a different group on a problem set, or allowing answers to be copied,

- fabricating the results of statistical analyses.

If you are unsure about the difference between collaboration on problem sets and copying, please come speak to me before there are any potential problems.

Any student who is caught cheating on an assignment will be reported to the Honor Committee.

UVa Statlab and Meetup Groups

UVa offers a free consulting service for people doing data analysis in R (and other software packages). Please note, they will help you to understand R and accomplish tasks with code, but they **do not provide tutoring and they are not allowed to help you with the problem sets** specifically. To set up an appointment, email statlab@virginia.edu.

Also, please consider joining the UVa R Users Meetup group: https://www.meetup.com/UVa-R-Users-Group/; or the Charlottesville Data Science Meetup group: https://www.meetup.com/CharlottesvilleDataScience/. These groups host talks for individuals in the UVa and Charlottesville communities on innovative research using R and cutting edge data science methods. It's a wonderful community and you are very welcome to join it.

Schedule and Readings

R4DS = R for Data Science; PAUR = Political Analysis Using R; TMWR = Text Mining With R

Week	Date	Topic	Reading
1	Tuesday, January 15	Introduction: What we can do with R?	PAUR ch. 1, whenever you are able to get a copy
	Thursday, January 17	Data, R, R Studio, packages and CRAN	R4DS ch 1, 4
2	Tuesday, January 22	Objects and scripts	R4DS ch. 6
	Thursday, January 24	Getting help	https://stackoverflow.com/help/how-to-ask https://codeblog.jonskeet.uk/2018/03/17/stack-overflow-culture/ https://blog.codinghorror.com/what-does-stack-overflow-want-to-be-when-it-grows-up/
3	Tuesday, January 29	Using R markdown and *knitr* (PS 1 assigned)	R4DS ch. 27, 29, 30
	Thursday, January 31	Tidy data, opening and saving data files, managing rows and columns	R4DS section 5.1-5.5, 12.1, 12.2
4	Tuesday, February 5	Piping, collapsing, group calculations (PS 1 due, PS 2 assigned)	R4DS section 5.6-5.7, ch. 18
	Thursday, February 7	Managing categorical objects (factors)	R4DS ch. 15
5	Tuesday, February 12	Managing categorical objects (factors) (PS 2 due, PS 3 assigned)	
	Thursday, February 14	Strings, dates and times	R4DS ch. 14
6	Tuesday, February 19	Dates and times (PS 3 due, PS 4 assigned)	R4DS ch. 16
	Thursday, February 21	Reshaping	R4DS ch. 12
7	Tuesday, February 26	Merging (PS 4 due, PS 5 assigned)	R4DS ch. 13 (not 13.7)
	Thursday, February 28	Descriptive statistics	PAUR ch. 4, 5
8	Tuesday, March 5	*ggplot* (PS 5 due)	R4DS ch. 3, 28

Week	Date	Topic	Reading
8	Thursday, March 7	*ggplot*, continued, interactive graphics (PS 6 assigned)	https://plot.ly/r/#fundamentals https://cran.r-project.org/web/packages/googleVis/vignettes/googleVis_examples.html
9	Tuesday, March 12	SPRING BREAK	
	Thursday, March 14	SPRING BREAK	
10	Tuesday, March 19	Statistical models, bivariate linear regression	PAUR ch. 6
	Thursday, March 21	Interpreting coefficients, categorical X variables (PS 6 due)	
11	Tuesday, March 26	Inference and confidence	https://www.econometrics-with-r.org/ (Chapters 5, 7)
	Thursday, March 28	Multiple regression, control variables, model fit (PS 7 assigned)	https://www.econometrics-with-r.org/ (Chapter 6)
12	Tuesday, April 2	Logistic regression, probability plots	PAUR section 7.1
	Thursday, April 4	Logistic probability plots (PS 7 due, PS 8 assigned)	
13	Tuesday, April 9	Text as data	TMWR ch. 1, 3
	Thursday, April 11	Dealing with JSON and other crazy file types for text data (PS 8 due)	
14	Tuesday, April 16	Sentiment analysis (PS 9 assigned)	TMWR ch. 2
	Thursday, April 18	N-grams	TMWR ch. 4
15	Tuesday, April 23	Topic modeling (PS 9 due, PS 10 assigned)	TMWR ch. 6
	Thursday, April 25	Document-term matrices, interfacing with other text analysis packages	TMWR ch. 5
16	Tuesday, April 30	Wordfish and other amazing text methods (PS 10 due, Final exam assigned)	TBA