Lecture 1: Introduction to Data Analytics for Policy Analysis

Dr. Rachel Blum

October 24, 2019

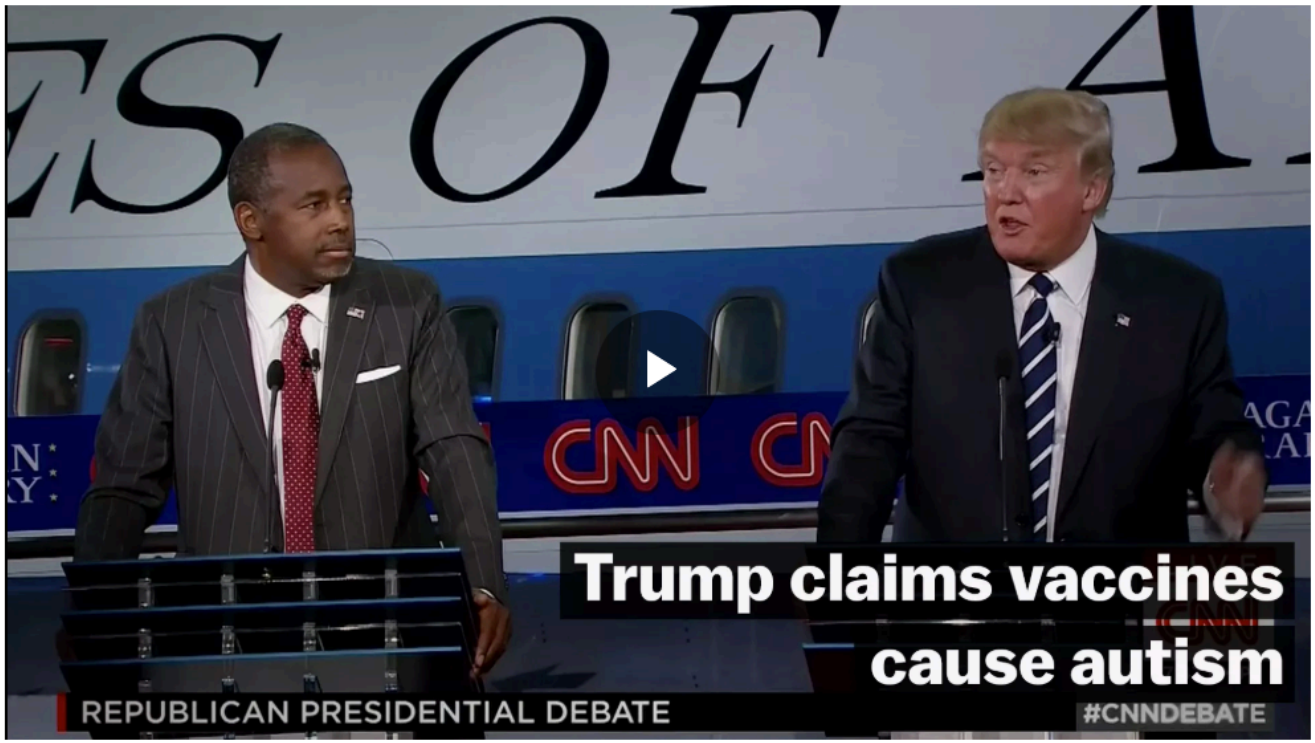
Table of Contents

Goals in this course

1. To understand challenges to cause and effect
2. To assess empirical claims
3. To conduct original analyses in R
4. To persuasively present results

Goals in this course

To understand challenges to cause and effect



Goals in this course

To understand and conduct analyses like this

Table 3
ANES full models

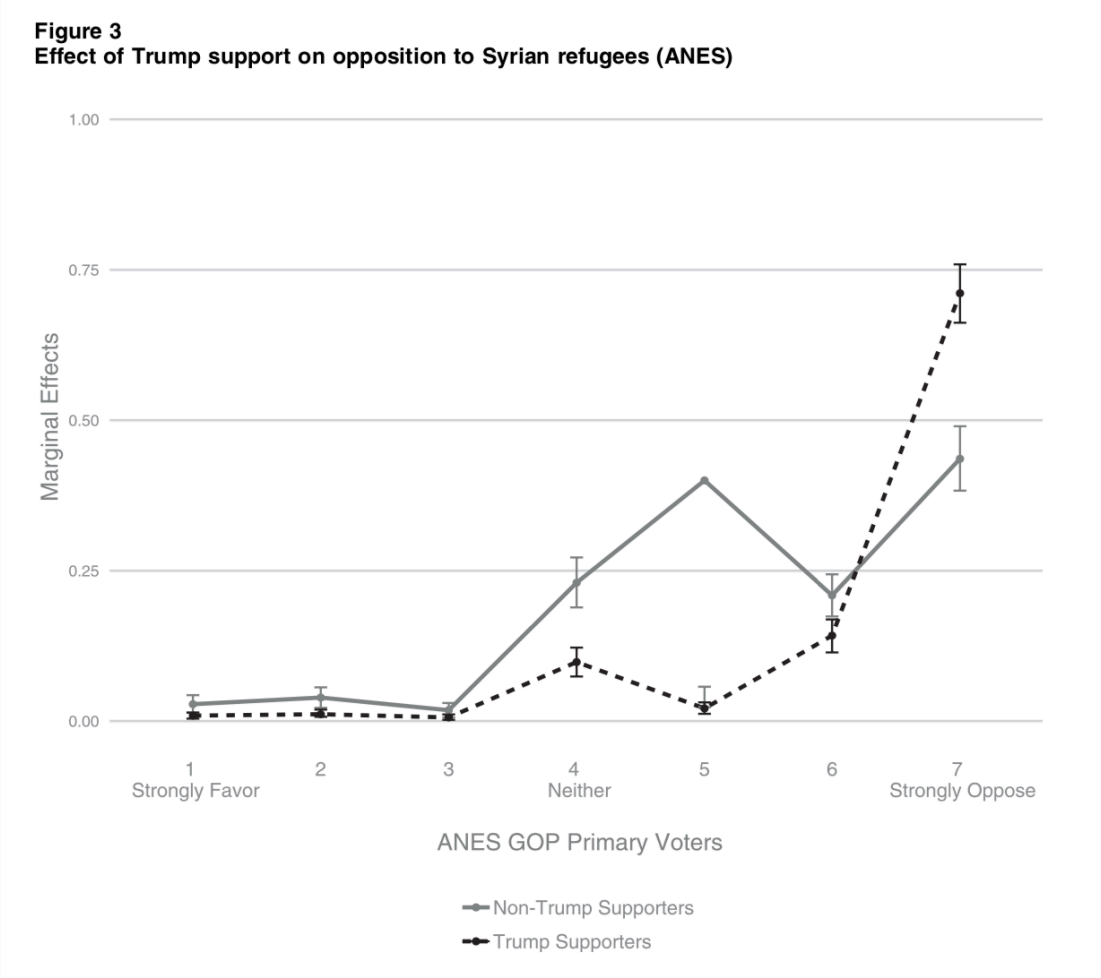
	Support for isolationism *Logit*	Opposition to Syria refugees *Ordered Logit*	Opposition to free trade agreements *Ordered Logit*
Trump Primary Choice	0.771*** (0.195)	1.158*** (0.161)	0.555*** (0.153)
Ideology	-0.085 (0.101)	0.513*** (0.084)	0.09 (0.08)
Education	-0.095 (0.1)	-0.105 (0.082)	-0.206** (0.078)
Income	-0.024 (0.013)	-0.008 (0.011)	-0.018 (0.011)
Female	-0.028*** (0.006)	0.01* (0.005)	-0.012 (0.005)
Age	-0.316 (0.191)	-0.031 (0.156)	-0.029 (0.149)
Not White	-0.107 (0.33)	-0.046 (0.274)	0.02 (0.272)
Log Likelihood	-351.84	-812.65	-991.6
N	684	688	594

***p $<$ 0.05, ** p $<$ 0.01, *** p $<$ 0.001

Standard errors in parentheses

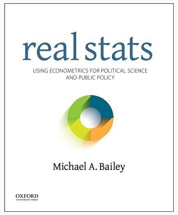
Goals in this course

To produce graphics like this

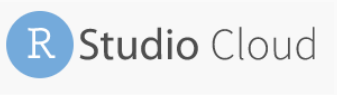


Course components

- **Text:** *Real Stats* by Michael Bailey.



- **Software:** RStudio (cloud).



- **Canvas:** Includes syllabus and all relevant info.
- **Class:** Lecture, then lab.
- **Assignments:**
  - Weekly Problem Sets
  - Hackathon
  - Quantitative Report
  - Final Exam

Table of Contents

Overview of Research Design

- We have a question about how the world works.
- We develop a theory about the world.
- That theory has observable implications.
- We need measures to fit those observable implications.
- We may want to **TEST** those implications to evaluate the theory.
- We make claims about the world based on our analyses.

TIME

These 5 Facts Explain Trump's Very Un-Republican Foreign Policy

BY IAN BREMMER JULY 22, 2016

Donald Trump is far more interested in winning over crowds than adhering to any kind of Republican foreign policy orthodoxy. These five facts detail the divergence between a traditional Republican approach to the world and the views of a candidate selling a foreign policy as iconoclastic as the man himself.

Developing a research question

Research Question: Is there a divide among Republican voters on Trump's foreign policy?

Developing a research question

Research Question: Is there a divide among Republican voters on Trump's foreign policy?

What kind of question is this? Causal? Descriptive?

Hypotheses

Research Question: Is there a divide among Republican voters on Trump's foreign policy?

What are our expectations/hypotheses?

Hypotheses

Research Question: Is there a divide among Republican voters on Trump's foreign policy?

> **Hypothesis:** Voters who supported Trump will have different foreign policy preferences than other Republicans.
>
> **Null:** There will be no difference in the foreign policy preferences of Trump supporters and other Republicans.

Operationalizing the research question

Operationalization means transforming concepts into an quantitative measures.

Identify concepts

↓

Developing indicators

↓

Translate indicators to variables

Turning concepts into variables

Concepts: Republican voters, Trump supporters, foreign policy.

1. What is our unit of analysis?
2. What is our population of interest?
3. What data can we use?
4. What variables will we use?

Turning concepts into variables

Concepts: Republican voters, Trump supporters, foreign policy.

1. Unit of analysis = voters.
2. Population = U.S. voters, or Republican voters in 2016.
3. Data = 2016 American National Elections Study.
4. Variables = Republican primary vote choice and foreign policy questions on Syrian refugees, trade, and isolationism.

Unit of analysis defined

Unit of analysis = smallest component of your study.

Can you think of examples?

Unit of analysis examples

- Voters
- Counties
- States
- Countries
- Wars

Population

Population $\Rightarrow$ **Sampling Frame** $\Rightarrow$ **Sample**

- **Population** = what we would ideally like to examine if we could get the data (e.g., all states, all citizens).
- **Sampling frame** = the units that could possibly be sampled given our method.
- **Sample** = a subset of the population (random samples are preferable).

Data, variables, and attributes

- **Data** = individual items of information.
- **Variables** = measurable characteristics of our units of measurement.
- **Attributes** = values a variable can take on.

Variable Types

1. **Categorical** (aka *nominal*): can be put in categories. Order is arbitrary.
    - **Ordinal** (aka *ranked*): categorical, but with a clear order.
    - **Binary** (aka *dummy* or *indicator*): two categories, reference and comparison. Order is arbitrary.
2. **Continuous** (aka *interval*): potentially infinite number of values. Order matters.
    - **Ratio**: similar to continuous/interval variables, but with meaningful zero. Order matters.

ANES presidential primary variable

V161021a

Label: PRE: For which candidate did R vote in Presidential prim
Item name: PREVOTE_PRIMVWHO

Question: IF R VOTED IN A PRESIDENTIAL PRIMARY OR CAUCUS: Looking at page [PRELOAD: page] in the booklet. In the Presidential primary or caucus, who did you vote for? SHOW RESPONDENT BOOKLET For Web administration, reference to the respondent booklet was omitted.

Unweighted Frequencies

	FTF	Web	Total
1. Hillary Clinton	143	436	579
2. Bernie Sanders	98	294	392
3. Another Democrat	3	16	19
4. Donald Trump	113	333	446
5. Ted Cruz	48	114	162
6. John Kasich	29	85	114
7. Marco Rubio	17	68	85
8. Another Republican	16	36	52
9. Someone else who is not a Republican or Democrat	2	22	24
-1. Inapplicable	706	1682	2388
-8. Don't know	3	0	3
-9. Refused	2	4	6

ANES isolationism variable

V161153

Label: PRE: Country would be better off if we just stayed home
Item name: USWORLD_USISOL

Question: Do you agree or disagree with this statement: 'This country would be better off if we just stayed home and did not concern ourselves with problems in other parts of the world.'

Unweighted Frequencies

	FTF	Web	Total
1. Agree	358	962	1320
2. Disagree	798	2112	2910
-8. Don't know	19	0	19
-9. Refused	5	16	21

ANES Syrian refugees variable

V161214x

Label: PRE: SUMMARY - Allow Syrian refugees
Item name: Not applicable; administrative or derived variable

Question: Not applicable

Unweighted Frequencies

	FTF	Web	Total
1. Favor a great deal	108	281	389
2. Favor a moderate amount	141	293	434
3. Favor a little	57	101	158
4. Neither favor nor oppose	254	943	1197
5. Oppose a little	56	103	159
6. Oppose a moderate amount	157	366	523
7. Oppose a great deal	386	988	1374
-8. Don't know	17	0	17
-9. Refused	4	15	19

ANES trade variable

V162176a

Label: POST: How strongly favor/oppose free trade agreements w/other countries
Item name: FREETRADE_AGRMTSTR

Question: IF R FAVORS U.S FREE TRADE AGREEMENTS WITH OTHER COUNTRIES/ IF R OPPOSES U.S. FREE TRADE AGREEMENTS WITH OTHER COUNTRIES: How strongly do you [favor/oppose] it?

Unweighted Frequencies

	FTF	Web	Total
1. A great deal	180	422	602
2. Moderately	337	824	1161
3. A little	131	252	383
-1. Inapplicable	409	1092	1501
-6. No post-election interview	122	414	536
-7. No post data, incomplete IW	0	86	86
-8. Don't know	1	0	1

What do we want to know?

Do voters who support Trump have different foreign policy preferences than other voters?

1. Probability of this relationship existing
2. Whether a relationship exists between two concepts (variables)
3. Whether a certain variable *causes* a certain outcome

Start here: description

Summary statistics show trends in our data.

Table 1
Summary statistics compared for Republican primary participants

	ANES	SCDS
Average Age	57	55
Average Education	Some college	Some post-graduate training
Percent Female	47.65	31.59
Average Ideology (1–7)	5.48	6.05
Average Income	$65,000	$74,000
Average Party ID (1–7)	5.97	5.38
Percent White	90	90.07
N	859	2,397

Cross-tabs

We can use cross-tabs to compare two variables.

Table B1: Position on whether the US should stay at home by primary candidate choice (ANES)

	Disagree (0)	Agree (1)
Cruz	85.7% (138)	14.3% (23)
Kasich	80.4% (90)	19.6% (22)
Rubio	77.4% (65)	22.6% (19)
Trump	68.2% (303)	31.8% (141)

Chi2: 22.39, Pr=0.000

Statistical inference

We want to make **statistical inferences**. As defined in *Real Stats* (pg. 120):

Statistical inference refers to the process of reaching conclusions based on the data.

Types of statistical inference

- **Hypothesis testing** (t-tests): are the observed data consistent with a certain claim (hypothesis)?
- **Point estimates** (regression): a best guess for a parameter.
    - *Parameters*: values that tell you about your entire population.
- **Confidence intervals**: best guess for an interval that traps the value of a parameter.
- **Bayesian inference**: probability of an outcome based on our prior beliefs and observed outcomes.

Hypothesis testing

We might want to test whether two groups differ in a statistically significant way.

- Do two groups differ from one another? $\rightarrow$ t-test
- Are two variables independent of one another? $\rightarrow \chi^2$
- Are two variables correlated, and if so, is the correlation positive or negative?

Tools: statistical models

Regression models allow us to quantify the relationship between variables while *controlling* for potentially confounding factors.

- Whether a relationship exists between two variables,
- how strong this relationship is, and
- what we expect Y (our outcome or dependent variable) will be for any given value of X (our treatment or independent variable).

Correlation or causation?

Making an inference about our data is not the same thing as making a *causal* inference.

- A causal inference is a conclusion about what made something happen.
- Correlation refers to a relationship between two factors. We might suspect that this relationship is causal, but we cannot establish this due to obstacles such as:
  - Randomness (stochasticity)
  - Endogeneity
  - Measurement error
  - Complexity of human behavior

Beating endogeneity: experiments

Randomized treatment lets us pinpoint source of exogeneous variation.

- Identify a relevant population that we randomly split into two groups.
- **Treatment group** receives the policy intervention.
- **Control group** does not.
- After treatment is assigned and administered, we compare the outcomes of both groups.
- If treatment group behaved differently than the control group, we believe the treatment had an effect. If not, then we believe it did not.

Experiments and validity

Randomized experiments rule out the possibility that the outcome is determined by factors other than the variable, or potential cause, of interest.

- The general term for this is **internal validity**.

However, a randomized experiment addresses effect of a given treatment at a given time and place. This effect may not continue in different times and places.

- The general term for this is **external validity**.

Table of Contents

What is “data analytics?”

Roughly, the science of collecting, organizing, and examining data in order to draw conclusions from it. Often pertains to:

- Raw or unstructured data,
- big data in the form of large-N datasets or computationally intensive techniques,
- machine learning, and
- innovative data visualization techniques.

Data analytics and politics

- Corpora of congressional communication: 203 op-eds, 49,668 press releases, and 19,352 e-newsletters for House members in the 113th Congress.
- Extensive data collection, processing, and organization required prior to using automated content analysis to break each corpora into its main topics.

Data formats

Broadly, data can be structured or unstructured. But even unstructured data has to be organized in some sort of file system.

- Easier across platforms: `.csv` (to some extent `.xslx`), `.txt`, `JSON` (language-independent format).
- Specific to certain programs: `.dta` (Stata), `.sas` (Sas), `.sav` (SPSS), `RData` (R).
- Web-based formats: `JSON`
- Some programs, such as `Python` and `R` can read any or all of these languages.

Data and your computer

Mac:

`/users/blumrm/dropbox/Miami/POL604_FA2019/datasets/examples_ps_1.csv`

Windows:

`C:/users/blumrm/dropbox/Miami/POL604_FA2019/datasets/examples_ps_1.csv`

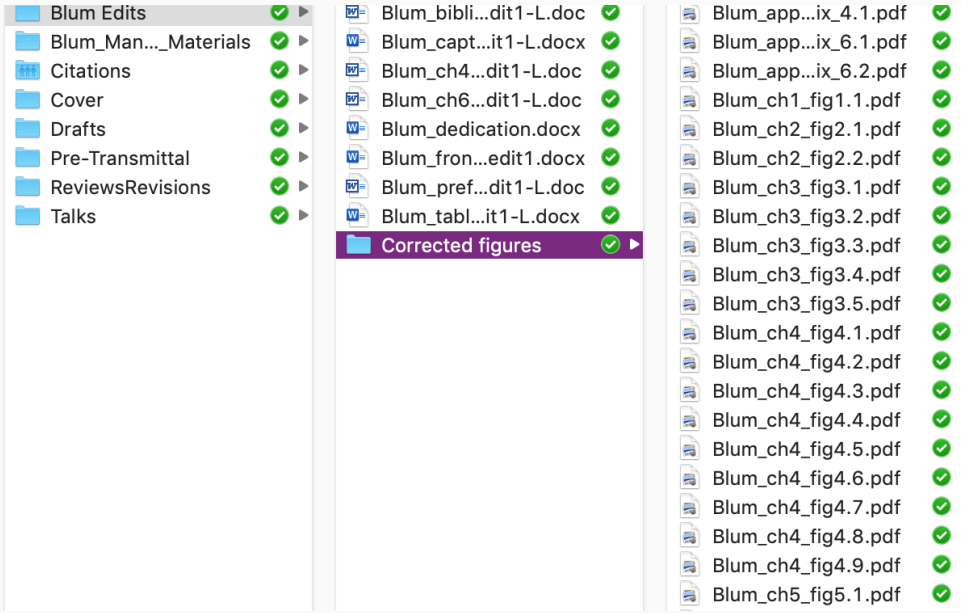
Data and your computer

Right click on any file to find the following:

- File path: `/users/blumrm/dropbox/Miami/POL604_FA2019/datasets/`
- File name: `examples_ps_1`
- File type: `.csv`

Data workflow

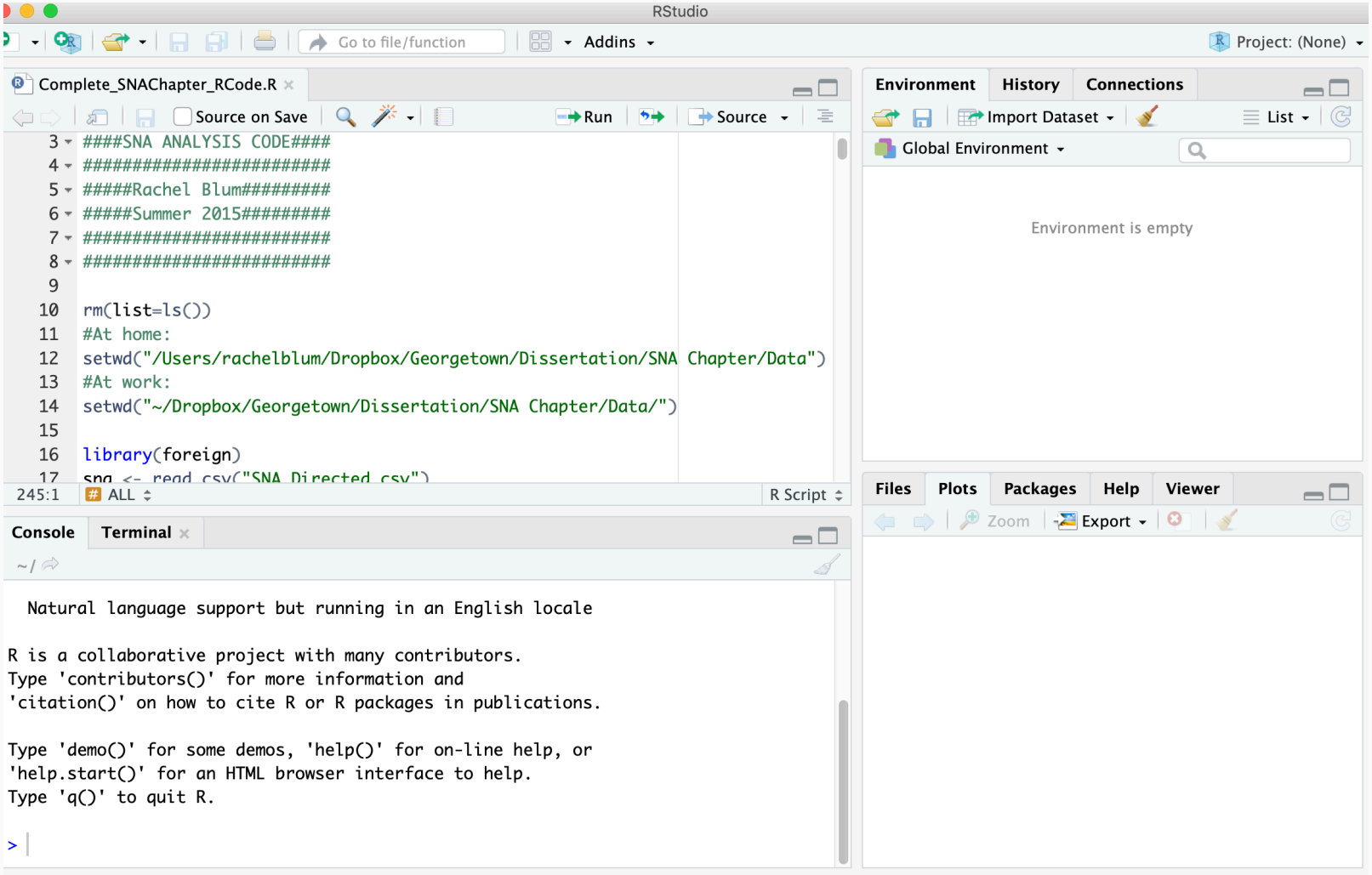
- Use a clear folder structure for your files.
- Group similar projects together.
- Name files clearly using similar conventions.
- Save files where they belong, not in *downloads* purgatory.



Using R for analysis

- R is a statistical language used for analysis.
- R is free.
- R makes it easy to create reproducible analyses.
- RStudio and RStudio cloud are useful interfaces that sit on top of R.

Let's open R...


RStudio
Go to file/function
Addins
Project: (None)
Complete_SNAChapter_RCode.R
Source on Save
Run
Source
###SNA ANALYSIS CODE####
#####Rachel Blum#########
#####Summer 2015#########
rm(list=ls())
#At home:
setwd("/Users/rachelblum/Dropbox/Georgetown/Dissertation/SNA Chapter/Data")
#At work:
setwd("~/Dropbox/Georgetown/Dissertation/SNA Chapter/Data/")
library(foreign)
245:1
ALL
R Script
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Zoom
Export
Console
Terminal
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

R scripts

- R scripts allow you to save and re-run everything you do.
- If no R script is open, go to files, new file, R script.
- Each R-script should be comprehensive (the code version of a project).
- Language in R scripts should be precise.
- Save your R-script where you save your data.

That's it!

By next class...

- Complete the R Studio tutorials on Canvas.
- Complete problem set 1.
- Read chapter 4 of *Real Stats*.