So you want to be a researcher?
Principles and Practical Data Tools to help you Fly Transparently

UCLA CCPR Statistics Core Workshop

Michael Tzen
michael@ccpr.ucla.edu

Fall 2017

CCPR Stats

http://www.ccpr.ucla.edu/CCPRWebsite/services/statistics-and-methods

Stats Consulting

Data Q&A, Stats Advice, Funded Projects

11:00 - 12:00 PM (T)

2:00 - 3:00 PM (R)

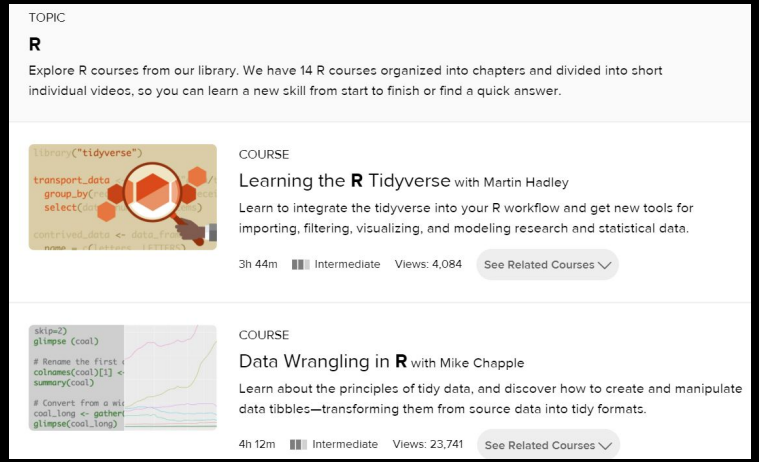
4284 Public Affairs

IDRE

https://idre.ucla.edu/calendar

Lynda.com

https://oit.ucla.edu/lynda-com



Thank You (Before I Forget)

Helpful References (Free)

"The Elements of Data Analytic Style"
- Jeff Leek

https://leanpub.com/datastyle

"Tidy Data" - Hadley Wickham

http://www.jstatsoft.org/v59/i10

Open Science Framework

https://osf.io/

Feedback Survey

https://goo.gl/forms/qURJcYpE7Pf8fEdn1

Outline: Transparent Data Workflow

Data - Big/Small x Designed/Undesigned

1. Obtain
2. Input
3. Wrangle
4. Analyze

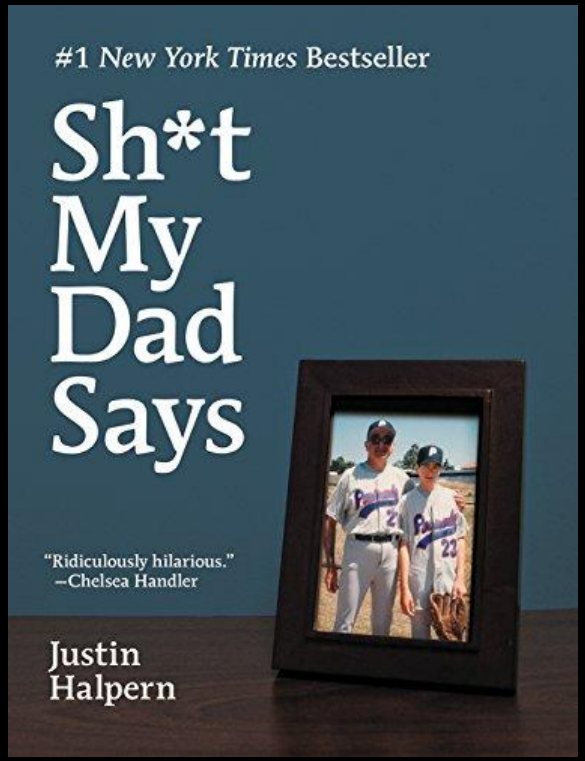
Outline: Transparent Data Workflow

Data - Big/Small x Designed/Undesigned

1. Obtain
2. Input
3. Wrangle
4. Analyze

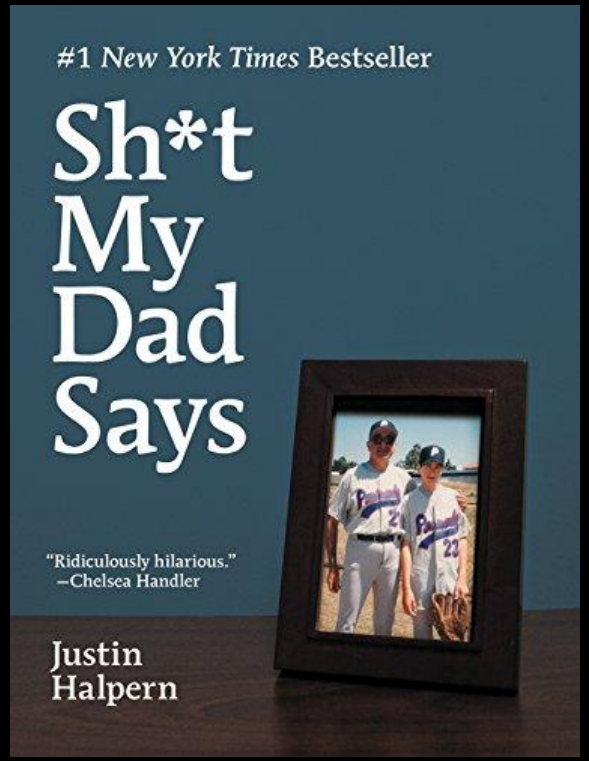
0. Do Each Transparently

Why Transparency?



- 6th Grade Science Project
- Dad is a scientist at UCSD
- 6th Grader fabricates data about family dog recognizing shapes
- Scientist dad finds out...

Why Transparency?



"You have shamed the entire scientific community. Fucking Einstein, everybody..."

"...I'm sorry I had to be so hard on you, but I don't want people thinking you're a lying sack of shit. You ain't. You're a quality human being. Now go to your room, you're grounded."

Human Being: Transparency for All

- For Yourself
  - Remind yourself what you did
  - This is what I did 3 months ago

- For Collaborators
  - Remind your collaborators what you did
  - Colleagues do not know the intimate details you do

- For Public Dissemination
  - “Show your Work” for Journal / Media Outlets
  - Public knows less than your colleagues about intimate details

New Low Burden Tools

User → Jupyter / Rstudio → [R]

A. Transparency Engine

- Jupyter Notebook (universal)
- Rstudio Knitr Notebook (R)
- dyndoc (Stata 15 $$$)

B. Data Analysis Software

- [R]
- Python
- Stata
- other

0a. Jupyter for Literate Code Snippets

jupyter Jupyter_R_Stata Last Checkpoint: a day ago (autosaved)

File Edit View Insert Cell Kernel Help

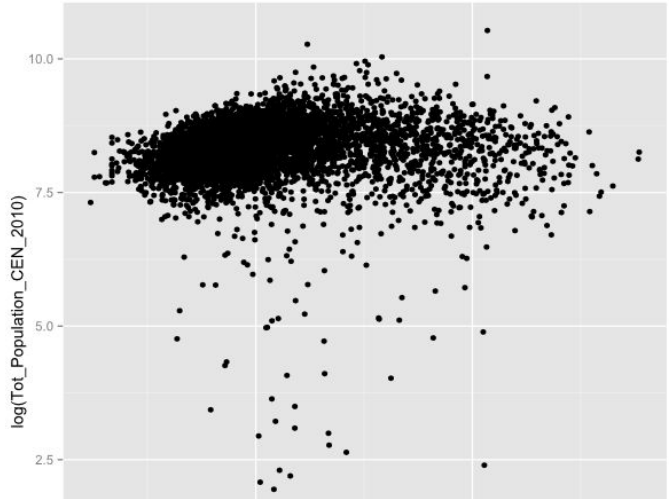
Code Cell Toolbar: None

On Your Own: Request Tract Size Variable, Make scatterplot of Population by Tract Size

http://api.census.gov/data/2015/pdb/tract/variables/LAND_AREA.json

In [21]:
```
ggplot(data_pdb_df_pk, aes(x=log(LAND_AREA),y=log(Tot_Population_CEN_2010))) +
geom_point()

ggplot(data_pdb_df_pk, aes(x=(LAND_AREA),y=(Tot_Population_CEN_2010))) +
geom_point()
```



Narrative with essential Markdown formatting

Markdown rendering is Immediate

WYSWYG

Software Code

Results of Software Code

0b. Rstudio for Reproducible Development

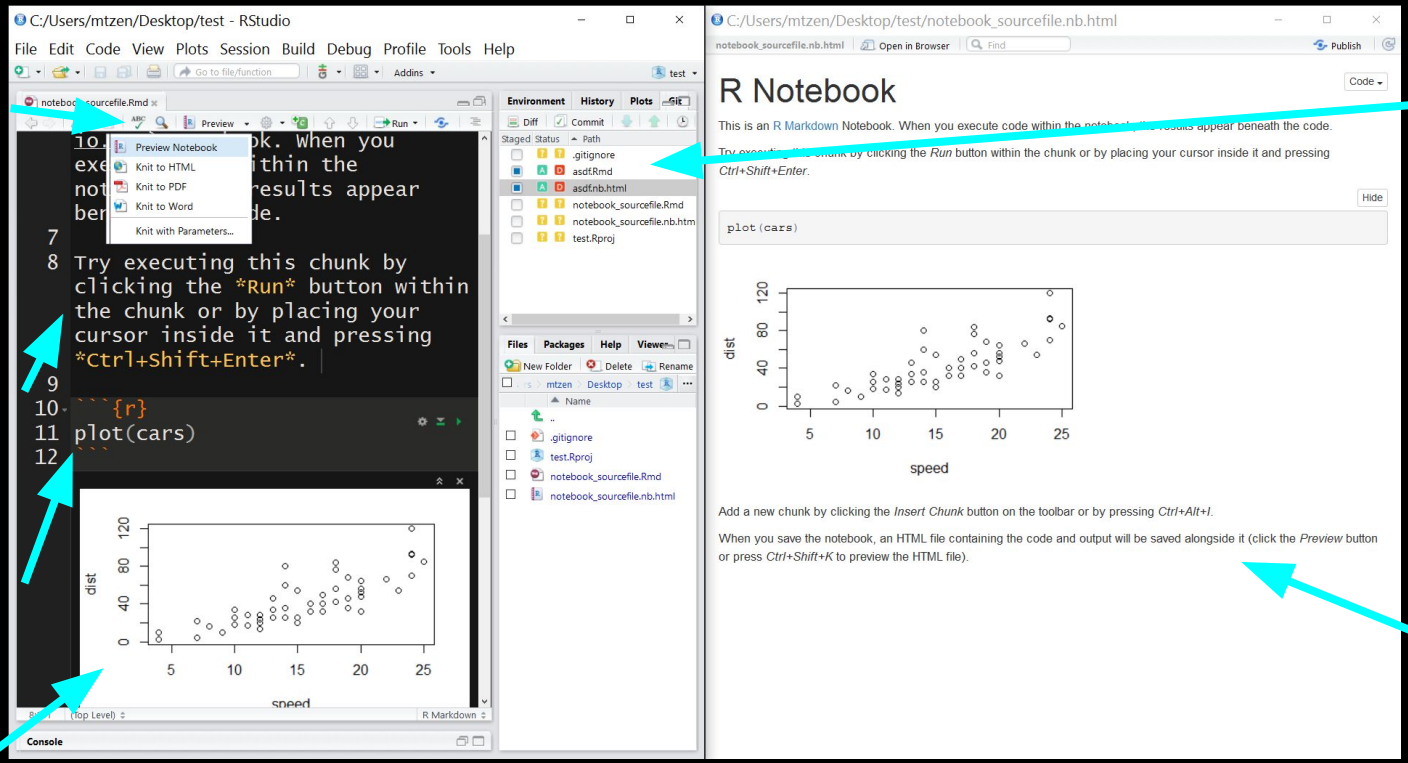
Markdown Compilation Step

NO WYSWYG

Markdown Syntax in .Rmd file

Code in .Rmd file

Preview Code Results



Git + Github

Code Results + Markdown Rendered Results

0c. How about Stata? 2016 Last Year

User → ??? → Stata

'Markdoc' for Stata - specialized package to write markdown syntax as comments in a stata script

Verdict: Premature, 'transparency' (in general) for stata is lagging behind.

User → Jupyter → [R] → Stata

'Rstata' package to boomerang Stata output to/from R

Verdict: Premature, Erroneously stumbles out of the starting gate

https://github.com/lbraglia/RStata/issues/2

0c. How about Stata? 2017 Now

User → $$$ → Stata 15

https://www.stata.com/new-in-stata/markdown/

Upgrade purchases

Educational single-user from Stata/SE 14 (or earlier)

You must be affiliated with a degree-granting institution to receive educational pricing.

ALL PRICES IN USD

Stata/SE	Stata/MP 2-core	Stata/MP 4-core	Stata/MP 6-core	Stata/MP + cores
For large datasets.	Fast & for the largest datasets.	Faster.	Even faster.	Even faster.
$425/perpetual Buy	$725/perpetual Buy	$1,025/perpetual Buy	$1,225/perpetual Buy	Select cores

Product features	Stata/SE	Stata/MP
Maximum number of variables	32,767	120,000

Announcing Stata release 15

Learn more »

Hi Mike,

I hope you are well! Do we happen to have access through CCPR to STATA SE 14 or STATA SE 15? I have a large-ish data file that was created in newer STATA, so I need to save it as old in order to open it up on my computer which only has STATA SE 13.

Thank you,

[redacted]

Software Exercise:

Quick setup of Transparency Engine

Mental Exercise:

Reflect on how the Jupyter or Rstudio workflow can help transparency

1. Obtain → 2. Input → 3. Wrangle → 4. Analyze

0b. Organize

1. Obtain → 2. Input → 3. Wrangle → 4. Analyze

0b. Organize

Agree on sharing formats with team

A possible file structure

```
\docs
    \project
        \data
            \data_raw
            \data_proc
        \code_software
        \results
        \writeup
```

Save any output as both:

1. universal format
   - .csv
   - .txt
   - .md
2. software specific format
   - .dta
   - .sas7bdat
   - .Rdata
   - .docx

1. Obtain → 2. Input → 3. Wrangle → 4. Analyze

0b. Organize

Agree on sharing formats with team

A possible file structure

```
\docs
    \project
        \data
            \data_raw
            \data_proc
        \code_software
        \results
        \writeup
```

Save any output as both:

1. universal format
   - .csv
   - .txt
   - .md
2. software specific format
   - .dta
   - .sas7bdat
   - .Rdata
   - .docx

1. Obtain → 2. Input → 3. Wrangle → 4. Analyze

0b. Organize

Name your Babies, Name your Files

prefix + noun + suffix

Challenge yourself to choose 3 words

"If I were to only use 1 word to explain the concept to someone new, which is the most important word?"

choose that as your noun

everything else as suffix for context

prefix for 'order of operations'

a_slides_workshop_transparent.pdf

b_exercises_workshop_transparent.r

http://kbroman.org/steps2rr/pages/organize.html

steps toward reproducible research

Organize your data and code

Perhaps the most important step to take towards ease of reproducibility is to be *organized*. Ideally, the names of files and subdirectories are self-explanatory, so that one can tell at a glance what data files contain, what scripts do, and what came from what.

- **Encapsulate everything within one directory**. Have a single directory for a project, containing all of the data, code, and results for that project. This makes it easier to find things, or to zip it all up and hand it off to someone else.
- **Separate raw data from derived data** and other data summaries. I prefer to have a subdirectory `RawData/` and then another subdirectory `Data/`, or perhaps two other subdirectories `DerivedData/` (containing reformatted, reorganized, or cleaned data files) and `DataSummaries/` (containing summary information, like lists of subjects or genetic markers, or summary statistics extracted from the primary data in order to make a particular graph). This makes it easier to tell the nature of the data in a file, by its location within the project directory.
- **Separate the data from the code**. I prefer to put code and data in separate subdirectories. I'll have an `R/` subdirectory and perhaps also `Python/` and `Ruby/` subdirectories.
- **Use relative paths** (never absolute paths). If you encapsulate all data and code within a single project directory, then you can refer to data files with relative paths (e.g., `../RawData/some_file.csv`). If you were to use an *absolute* path (like `~/Projects/SomeProject/RawData/some_file.csv` or `C:\Users\SomeOne\Projects\SomeProject\RawData\some_file.csv`) then anyone who wanted to reproduce your results but had the project placed in some other location would have to go in and edit all of those directory/file names.
- **Choose file names carefully**. I try not to change the names of raw data files that I get from a collaborator (though I'm often tempted to replace spaces with underscores). But scripts need names, and files with derived or cleaned data need names. Be as clear and explicit as possible. The same holds for the variables and functions within your scripts.

1. Obtain → 2. Input → 3. Wrangle → 4. Analyze

0b. Organize

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, Eivind Hovig

Rule 1: For Every Result, Keep Track of How It Was Produced

Rule 2: Avoid Manual Data Manipulation Steps

Rule 3: Archive the Exact Versions of All External Programs Used

Rule 4: Version Control All Custom Scripts

Rule 5: Record All

Rule 3: Archive the Exact Versions of All External Programs Used

In order to exactly reproduce a given result, it may be necessary to use programs in the exact versions used originally. Also, as both input and output formats may change between versions, a newer version of a program may not even run without modifying its inputs. Even having noted which version was used of a given program, it is not always trivial to get hold of a program in anything but the current version. Archiving the exact versions of programs actually used may thus save a lot of hassle at later stages. In some cases, all that is needed is to store a single executable or source code file. In other cases, a given program may again have specific requirements to other installed programs/packages, or dependencies to specific operating system components. To ensure future availability, the only viable solution may then be to store a full virtual machine image of the operating system and program. As a minimum, you should note the exact names and versions of the main programs you use.

http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285#s4

1. Obtain → 2. Input → 3. Wrangle → 4. Analyze

0b. Organize

Software Exercise:

Organize before starting analysis

Mental Exercise:

Reflect on how organization helps transparency

1. Obtain - Data Stream

For Immediate Release May 09, 2013

Executive Order -- Making Open and Machine Readable the New Default for Government Information

EXECUTIVE ORDER

- Computer File
  - .csv (universal)
  - .dta (stata)
- Web API
  - Census
  - Twitter

1. Obtain → 2. Input → 3. Wrangle → 4. Analyze

https://api.census.gov/data.html

Jupyter_R_Stata | UCLA Projects - Google Dr... | CCPR_Stat_Wkshp_FALL20... | http://api.cen...0d74861754e934 | Census Data API: /data/2013/p... | Census Data API: /data

api.census.gov/data.html

ccpremail CCPR box uclagoogle TRS IDRE Classes Rstudio evernote GitHub Feedly GelmanBlog RB FD census staff Air CSSIP Videos AYSO StackEdit JSS

Census API: Datasets in /data and its descendants

Title	Description	Vintage	Dataset Name	Geography List	Variable List	Tag List	Examples	Developer Documentation	API Base URL
1990 Decennial: Summary File 1	Data files available from Census 2000 and the 2010 Census. This file presents 100-percent population and housing figures for the total population, for 63 race categories, and for many other race and Hispanic or Latino categories. This includes age, sex, households, household relationship, housing units, occupancy status, and tenure (whether the residence is owned or rented). Also included are selected characteristics for a limited number of race and Hispanic or Latino categories. The data are available for the U.S., regions, divisions, states, counties, county subdivisions, places, census tracts, block groups, blocks, metropolitan areas (2000), core based statistical areas (2010), American Indian and Alaska Native areas, tribal subdivisions, Hawaiian home lands, congressional districts, and ZIP Code Tabulation Areas. Data are available down to the block level for many tabulations, but only to the census-tract level for others. The 2010 Census SF 1 has some tables on the population in group quarters that are available only to the county level. Available on DVD and American FactFinder. The Census 2000 Summary File 1 data were released in three stages. Individual state files and two national files were released. The state-level data were released first, followed by the Advance National File, which covered the same data subjects, but includes national level summary data for areas that cross state boundaries such as whole metropolitan areas, whole American Indian areas, etc. The Final National File contains the same data subjects and geographic areas as the Advance National File, but adds the first available urban/rural and urbanized area data. For the most current release dates for these files, see the "Census 2000 Release Schedule" link on the AFFAmerican FactFinder Main Page. The 2010 Census Summary File 1 data were released in stages. Individual state files and a national update were released in 2011. The SF 1 urban/rural update is planned for release in Fall 2012. This file contains the same data subjects as the previously released files, but for additional geography, including the urban and rural parts of the United States, regions, divisions, states, counties, and places; and urbanized areas and urban clusters. See the 2010 Census Data Products At A Glance	1990	sf1	geographies	variables	tags	N/A	documentation	http://api.census.gov/data/1990/sf1
1990 Decennial Census of Population and Housing - Summary File 3: Summary File 3	The census of population and housing, taken by the Census Bureau in years ending in 0 (zero). Article I of the Constitution requires that a census be taken every ten years for the purpose of reapportioning the U.S. House of Representatives. Title 13 of the U. S. Code provides the authorization for conducting the census in Puerto Rico and the Island Areas. After each decennial census, the results are released to the public in a variety of ways, including publishing multiple series of reports titled Census of Population and Housing. The abbreviation for these reports was CPH for some decades (including 1990 and 2010) and PHC for some decades (including 1970 and 2000).	1990	sf3	geographies	variables	tags	examples	documentation	http://api.census.gov/data/1990/sf3
2000 Decennial: Summary File 1	Data files available from Census 2000 and the 2010 Census. This file presents 100-percent population and housing figures for the total population, for 63 race categories, and for many other race and Hispanic or Latino categories. This includes age, sex, households, household relationship, housing units, occupancy status, and tenure (whether the residence is owned or rented). Also included are selected characteristics for a limited number of race and Hispanic or Latino categories. The data are available for the U.S., regions, divisions, states, counties, county subdivisions, places, census tracts, block groups, blocks, metropolitan areas (2000), core based statistical areas (2010), American Indian and Alaska Native areas, tribal subdivisions, Hawaiian home lands, congressional districts, and ZIP Code Tabulation Areas. Data are available down to the block level for many tabulations, but only to the census-tract level for others. The 2010 Census SF 1 has some tables on the population in group quarters that are available only to the county level. Available on DVD and American FactFinder. The Census 2000 Summary File 1 data were released in three stages. Individual state files and two national files were released. The state-level data were released first, followed by the Advance National File, which covered the same data subjects, but includes national level summary data for areas that cross state boundaries such as whole metropolitan areas, whole American Indian areas, etc. The Final National File contains the same data subjects and geographic areas as the Advance National File, but adds the first available urban/rural and urbanized area data. For the most current release dates for these files, see the "Census 2000 Release Schedule" link on the AFFAmerican FactFinder Main Page. The 2010 Census Summary File 1 data were released in stages. Individual state files and a national update were released in 2011. The SF 1 urban/rural update is planned for release in Fall 2012. This file contains the same data subjects as the previously released files, but for additional geography, including the urban and rural parts of the United States, regions, divisions, states, counties, and places; and urbanized areas and urban clusters. See the 2010 Census Data Products At A Glance	2000	sf1	geographies	variables	tags	N/A	documentation	http://api.census.gov/data/2000/sf1
2000 Decennial: Summary File 3	This Census 2000 file presents data on the population and housing long form subjects such as income and education. It includes population totals for ancestry groups. It also includes selected characteristics for a limited number of race and Hispanic or Latino categories. The data are available for the U.S., regions, divisions, states, counties, county subdivisions, places, census tracts, block groups, metropolitan areas, American Indian and Alaska Native areas, tribal subdivisions, Hawaiian home lands, congressional districts, and Zip Code Tabulation Areas. Available on CD-ROM, DVD, and American FactFinder. After Census 2000, data on these subjects were produced from the American Community Survey or the Puerto Rico Community Survey.	2000	sf3	geographies	variables	tags	N/A	documentation	http://api.census.gov/data/2000/sf3
2002 Economic Census - All Sectors: Economy-Wide Key Statistics	The Economic Census is the U.S. Government's official five-year measure of American business and the economy. It is conducted by the U.S. Census Bureau, and response is required by law. In October through December 2012, forms were sent out to nearly 4 million businesses, including large, medium and small companies representing all U.S. locations and industries. Respondents were asked to provide a range of operational and performance data for their companies.	2002	ewks	geographies	variables	N/A	examples	documentation	http://api.census.gov/data/2002/ewks
2007 Economic Census - All Sectors: Economy-Wide Key Statistics	The Economic Census is the U.S. Government's official five-year measure of American business and the economy. It is conducted by the U.S. Census Bureau, and response is required by law. In October through December 2012, forms were sent out to nearly 4 million businesses, including large, medium and small companies representing all U.S. locations and industries. Respondents were asked to provide a range of operational and performance data for their companies.	2007	ewks	geographies	variables	N/A	examples	documentation	http://api.census.gov/data/2007/ewks
2008 County Business Patterns: Business Patterns	County Business Patterns (CBP) is an annual series that provides economic data by industry at the U.S., State, County and Metropolitan Area levels. This series includes the number of establishments, employment during the week of March 12, first quarter payroll, and annual payroll. CBP provides statistics for businesses with paid employees for the U.S., Puerto Rico, and the Island Areas.	2008	cbp	geographies	variables	N/A	examples	documentation	http://api.census.gov/data/2008/cbp
2008 Nonemployer Statistics: Non Employer Statistics	Nonemployer Statistics is an annual series that provides subnational economic data for businesses that have no paid employees and are subject to federal income tax. The data consist of the number of businesses and total receipts by industry. Most nonemployers are self-employed individuals operating unincorporated businesses (known as sole proprietorships), which may or may not be the owner's principal source of income. The majority of all business establishments in the United States are nonemployers, yet these firms average less than 4 percent of all sales and receipts nationally. Due to their small economic impact, these firms are excluded from most other Census Bureau business statistics (the primary exception being the Survey of Business Owners). The Nonemployers Statistics series is the primary resource available to study the scope and activities of nonemployers at a detailed geographic level. For complementary statistics on the firms that do have paid employees, refer to the County Business Patterns. Additional sources of data on small businesses include the Economic Census, and the Statistics of U.S. Businesses.	2008	nonemp	geographies	variables	N/A	examples	documentation	http://api.census.gov/data/2008/nonemp
2009 County Business Patterns: Business Patterns	County Business Patterns (CBP) is an annual series that provides economic data by industry at the U.S., State, County and Metropolitan Area levels. This series includes the number of establishments, employment during the week of March 12, first quarter payroll, and annual payroll. CBP provides statistics for businesses with paid employees for the U.S., Puerto Rico, and the Island Areas.	2009	cbp	geographies	variables	N/A	examples	documentation	http://api.census.gov/data/2009/cbp
2009 Nonemployer	Nonemployer Statistics is an annual series that provides subnational economic data for businesses that have no paid employees and are subject to federal income tax. The data consist of the number of businesses and total receipts by industry. Most nonemployers are self-employed individuals operating unincorporated businesses (known as sole proprietorships), which may or may not be the owner's principal source of income. The majority of all business establishments in the United States are nonemployers, yet these firms average less than 4 percent of all sales and receipts nationally. Due to their small economic impact, these firms are	2009	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	http://api.census.gov

Software Exercise:

Obtain 2015 Census Planning JSON data via API

Mental Exercise:

Reflect on how the software steps make data obtainment a transparent process

2. Input - Data into Software

Human Says Tomato

Computer Says Potato

Hurdle towards Finish Line

Foreground: Step 2

- **Short Term Goal - Tabular Format for Processing**
  - **Match human-computer expectations**
  - **Specify What is:**
    - **“Character” text**
    - **“Numeric” number**

Background: Step 4 in mind

- **Long Term Goal - Tabular Format for Analysis**
  - **Steer towards later analysis**
  - **Specify What is:**
    - **Entity-unit**
    - **Measure-variables**

Clear the Hurdle? Assertion Checks

The read-in data should agree with user expectation

+ Arnold handshakes Terminator
- Arnold arm wrestles Terminator

end-user domain knowledge with companion data guide

Basic Axioms

- Age in [0,200]
- Probability in [0,1]
- Sex in {Male,Female}
- ~ 70,000 Tracts
- ~ 3,000 Counties
- ~ 50 States
- Other Snapple Facts

Software Exercise:

With obtained API data, Input as tabular text and numeric data (check Snapple Facts too)

Mental Exercise:

Reflect on how the software steps make reading in data a transparent process

3. Wrangle - Analysis Tabular Format

Majority of analyses need the entities and measures to be extracted from indexable tabular formats

Steer the input data (2) towards analysis format (4) by wrangling into “Coordinate” or “Tidy” tabular formats

Arnold’s favorite pit-stop

1. Obtain → 2. Input → 3. Wrangle → 4. Analyze

The Coordinate, The Tidy, ...and The Ugly

Name	Year	Variable Name	Variable Value
Arnold Schwarzenegger	2030	Age	34
Arnold Schwarzenegger	2030	Sex	M
Arnold Schwarzenegger	2040	Age	44
Arnold Schwarzenegger	2040	Age	M
Sofia Vergara	2030	Age	30
Sofia Vergara	2030	Sex	F
Sofia Vergara	2040	Age	40
Sofia Vergara	2040	Sex	F

Format

- Each Row is a Coordinate Triplet (i, j, k)
  - i : entity
  - j : measure name
  - k : value of entity's measure

Positives

- Extremely Basic
- Quick Lookup of Specific Triplet

Negatives

- Hard to Glance for Group Patterns

Name	Year	Age	Sex
Arnold Schwarzenegger	2030	34	M
Arnold Schwarzenegger	2040	44	M
Sofia Vergara	2030	30	F
Sofia Vergara	2040	40	F

Format

- Each Row is a Unique Entity
- Measures as Columns

Positives:

- Ready for Analysis Routines
- Somewhat Glanceable

Negatives

- Very Minor drawback of storage

	Age-2030	Age-2040	Sex-2030	Sex-2040
Arnold Schwarzenegger	34	44	M	M
Sofia Vergara	30	40	F	F

Format

- No Consistent Format, Up to the End-User

Positives

- Flexibly Glanceable for Humans
  - "Consumer Report" (Belongs After Step 4)

Negatives

- Useless for Analysis Software
  - Want Low-Level Indexible Table (Before 4)

Software Exercise:

With properly inputted data, Wrangle to analysis ready tabular formats

Mental Exercise:

Reflect on how the wrangled data structures can help transparency

4. Analyze - Data with Software(s)

Armed with Tabular Entities and
Measures, ready for any analysis routine

Plot, Summary, Model

[R] Stata

Plot a Graphic

[R]

Summarize a Table

[R] and Stata

Fit a Regression Model

[R] and Stata

Software Exercise:

Analyze the data

Mental Exercise:

Reflect on how you got to this analysis step. Compare the amount of code in Step 4 with Steps 1-3

Thank You

Helpful References (Free)

"The Elements of Data Analytic Style" - Jeff Leek

https://leanpub.com/datastyle

"Tidy Data" - Hadley Wickham

http://www.jstatsoft.org/v59/i10

Open Science Framework

https://osf.io/

Feedback Survey

https://goo.gl/forms/qURJcYpE7Pf8fEdn1