

Show Me the Data

And the Source Code!

Kelly McConville



May 20, 2017

Data, data, data

- ▶ Saw lots of wonderful talks about exposing our students to adequately messy data.

Data, data, data

- ▶ Saw lots of wonderful talks about exposing our students to adequately messy data.
- ▶ I want to follow Chris Wild's advice and

Data, data, data

- ▶ Saw lots of wonderful talks about exposing our students to adequately messy data.
- ▶ I want to follow Chris Wild's advice and
 - ▶ *"Populate the imagination with possibilities."*

Data, data, data

- ▶ Saw lots of wonderful talks about exposing our students to adequately messy data.
- ▶ I want to follow Chris Wild's advice and
 - ▶ *"Populate the imagination with possibilities."*
- ▶ But, I also feel a bit over-whelmed.

Data, data, data

- ▶ Saw lots of wonderful talks about exposing our students to adequately messy data.
- ▶ I want to follow Chris Wild's advice and
 - ▶ *"Populate the imagination with possibilities."*
- ▶ But, I also feel a bit over-whelmed.
 - ▶ Especially, when Deb Nolan said it takes about ten articles to find one with data that fits the bill.

Data, data, data

- ▶ Saw lots of wonderful talks about exposing our students to adequately messy data.
- ▶ I want to follow Chris Wild's advice and
 - ▶ *"Populate the imagination with possibilities."*
- ▶ But, I also feel a bit over-whelmed.
 - ▶ Especially, when Deb Nolan said it takes about ten articles to find one with data that fits the bill.
- ▶ "Show me the data and the source code!"

Source code and the statistical community

- ▶ Why should we, the statistical community, value open source code?

Source code and the statistical community

- ▶ Why should we, the statistical community, value open source code?
 - ▶ Source code + Data = Reproducible workflow.

Source code and the statistical community

- ▶ Why should we, the statistical community, value open source code?
 - ▶ Source code + Data = Reproducible workflow.
 - ▶ *“Train new analysts whose only workflow is a reproducible one.”* –Mine Cetinkaya-Rundel

Source code and the statistical community

- ▶ Why should we, the statistical community, value open source code?
 - ▶ Source code + Data = Reproducible workflow.
 - ▶ *“Train new analysts whose only workflow is a reproducible one.”* –Mine Cetinkaya-Rundel
 - ▶ More people will use these great examples in their classes.

Source code and the statistical community

- ▶ Why should we, the statistical community, value open source code?
 - ▶ Source code + Data = Reproducible workflow.
 - ▶ *“Train new analysts whose only workflow is a reproducible one.”* –Mine Cetinkaya-Rundel
 - ▶ More people will use these great examples in their classes.
 - ▶ Equalizes access, even to people outside our classrooms.

Source code and the statistical community

- ▶ Why should we, the statistical community, value open source code?
 - ▶ Source code + Data = Reproducible workflow.
 - ▶ *“Train new analysts whose only workflow is a reproducible one.”* –Mine Cetinkaya-Rundel
 - ▶ More people will use these great examples in their classes.
 - ▶ Equalizes access, even to people outside our classrooms.
 - ▶ If you share your code, others will make it better!

Source code and the statistical community

- ▶ But, sharing code takes time and effort that often doesn't seem to help one's career.

Source code and the statistical community

- ▶ But, sharing code takes time and effort that often doesn't seem to help one's career.
- ▶ How do we show that we value this work?

Source code and the statistical community

- ▶ But, sharing code takes time and effort that often doesn't seem to help one's career.
- ▶ How do we show that we value this work?
 - ▶ Share code and improve shared code.

Source code and the statistical community

- ▶ But, sharing code takes time and effort that often doesn't seem to help one's career.
- ▶ How do we show that we value this work?
 - ▶ Share code and improve shared code.
 - ▶ Educate our institutions about why it matters.

Source code and the statistical community

- ▶ But, sharing code takes time and effort that often doesn't seem to help one's career.
- ▶ How do we show that we value this work?
 - ▶ Share code and improve shared code.
 - ▶ Educate our institutions about why it matters.
 - ▶ Write about it in tenure, promotion, and award letters.

Source code and the statistical community

- ▶ But, sharing code takes time and effort that often doesn't seem to help one's career.
- ▶ How do we show that we value this work?
 - ▶ Share code and improve shared code.
 - ▶ Educate our institutions about why it matters.
 - ▶ Write about it in tenure, promotion, and award letters.
 - ▶ Make it part of the culture by teaching our students to share their code.

Source code and the statistical community

- ▶ But, sharing code takes time and effort that often doesn't seem to help one's career.
- ▶ How do we show that we value this work?
 - ▶ Share code and improve shared code.
 - ▶ Educate our institutions about why it matters.
 - ▶ Write about it in tenure, promotion, and award letters.
 - ▶ Make it part of the culture by teaching our students to share their code.
 - ▶ Give credit to those doing open-source work. To name a few...

Source code and the statistical community

- ▶ But, sharing code takes time and effort that often doesn't seem to help one's career.
- ▶ How do we show that we value this work?
 - ▶ Share code and improve shared code.
 - ▶ Educate our institutions about why it matters.
 - ▶ Write about it in tenure, promotion, and award letters.
 - ▶ Make it part of the culture by teaching our students to share their code.
 - ▶ Give credit to those doing open-source work. To name a few...
 - ▶ Chester Ismay and Albert Y. Kim's "ModernDive"

Source code and the statistical community

- ▶ But, sharing code takes time and effort that often doesn't seem to help one's career.
- ▶ How do we show that we value this work?
 - ▶ Share code and improve shared code.
 - ▶ Educate our institutions about why it matters.
 - ▶ Write about it in tenure, promotion, and award letters.
 - ▶ Make it part of the culture by teaching our students to share their code.
 - ▶ Give credit to those doing open-source work. To name a few...
 - ▶ Chester Ismay and Albert Y. Kim's "ModernDive"
 - ▶ UC Berkeley's Data 8 Course Materials

Source code and the statistical community

- ▶ But, sharing code takes time and effort that often doesn't seem to help one's career.
- ▶ How do we show that we value this work?
 - ▶ Share code and improve shared code.
 - ▶ Educate our institutions about why it matters.
 - ▶ Write about it in tenure, promotion, and award letters.
 - ▶ Make it part of the culture by teaching our students to share their code.
 - ▶ Give credit to those doing open-source work. To name a few...
 - ▶ Chester Ismay and Albert Y. Kim's "ModernDive"
 - ▶ UC Berkeley's Data 8 Course Materials
 - ▶ David Diez, Christopher Barr, Mine Cetinkaya-Rundel's "OpenIntro"

Source code and the statistical community

- ▶ But, sharing code takes time and effort that often doesn't seem to help one's career.
- ▶ How do we show that we value this work?
 - ▶ Share code and improve shared code.
 - ▶ Educate our institutions about why it matters.
 - ▶ Write about it in tenure, promotion, and award letters.
 - ▶ Make it part of the culture by teaching our students to share their code.
 - ▶ Give credit to those doing open-source work. To name a few...
 - ▶ Chester Ismay and Albert Y. Kim's "ModernDive"
 - ▶ UC Berkeley's Data 8 Course Materials
 - ▶ David Diez, Christopher Barr, Mine Cetinkaya-Rundel's "OpenIntro"
 - ▶ (To come) Mine Cetinkaya-Rundel's "Data Science gateway course in a box"

Source code and the statistical community

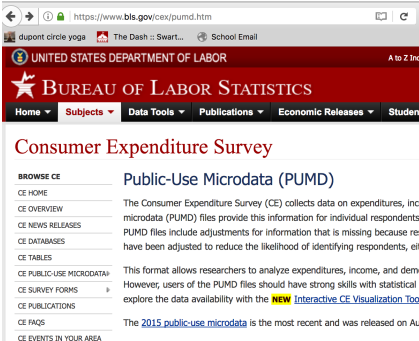
- ▶ But, sharing code takes time and effort that often doesn't seem to help one's career.
- ▶ How do we show that we value this work?
 - ▶ Share code and improve shared code.
 - ▶ Educate our institutions about why it matters.
 - ▶ Write about it in tenure, promotion, and award letters.
 - ▶ Make it part of the culture by teaching our students to share their code.
 - ▶ Give credit to those doing open-source work. To name a few...
 - ▶ Chester Ismay and Albert Y. Kim's "ModernDive"
 - ▶ UC Berkeley's Data 8 Course Materials
 - ▶ David Diez, Christopher Barr, Mine Cetinkaya-Rundel's "OpenIntro"
 - ▶ (To come) Mine Cetinkaya-Rundel's "Data Science gateway course in a box"
 - ▶ Jenny Bryan's "Happy Git and GitHub for the user"

Source code and the stat ed community

- ▶ But, sharing code takes time and effort that often doesn't seem to help one's career.
- ▶ How do we show that we value this work?
 - ▶ Share code and improve shared code.
 - ▶ Educate our institutions about why it matters.
 - ▶ Write about it in tenure, promotion, and award letters.
 - ▶ Make it part of the culture by teaching our students to share their code.
 - ▶ Give credit to those doing open-source work. To name a few...
 - ▶ Chester Ismay and Albert Y. Kim's "ModernDive"
 - ▶ UC Berkeley's Data 8 Course Materials
 - ▶ David Diez, Christopher Barr, Mine Cetinkaya-Rundel's "OpenIntro"
 - ▶ (To come) Mine Cetinkaya-Rundel's "Data Science gateway course in a box"
 - ▶ Jenny Bryan's "Happy Git and GitHub for the user"
 - ▶ Anyone who posts code to github or their website

Show me the data and the source code

- ▶ I call on all of us, myself included, to think about one set of materials (data + code) you could share on github or your website.



The screenshot shows a web browser window with the URL <https://www.bls.gov/cex/pumd.htm>. The page header includes the United States Department of Labor and Bureau of Labor Statistics logos, along with a navigation menu with items like Home, Subjects, Data Tools, Publications, Economic Releases, and Student. The main heading is "Consumer Expenditure Survey". Below this, there is a section for "Public-Use Microdata (PUMD)". A left-hand sidebar lists various links under the heading "BROWSE CE", including CE HOME, CE OVERVIEW, CE NEWS RELEASES, CE DATABASES, CE TABLES, CE PUBLIC-USE MICRODATA, CE SURVEY FORMS, CE PUBLICATIONS, CE FAQs, and CE EVENTS IN YOUR AREA. The main content area contains text explaining that the Consumer Expenditure Survey (CE) collects data on expenditures, income, and demographic information, and that PUMD files provide this information for individual respondents. It also mentions that PUMD files include adjustments for missing information and that users should have strong skills with statistical software. A link to the "NEW Interactive CE Visualization Tool" is highlighted in yellow. The text concludes by stating that the "2015 public-use microdata" is the most recent and was released on August 11, 2015.

BROWSE CE

- CE HOME
- CE OVERVIEW
- CE NEWS RELEASES
- CE DATABASES
- CE TABLES
- CE PUBLIC-USE MICRODATA
- CE SURVEY FORMS
- CE PUBLICATIONS
- CE FAQs
- CE EVENTS IN YOUR AREA

Public-Use Microdata (PUMD)

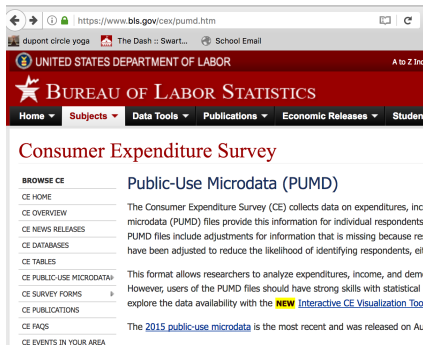
The Consumer Expenditure Survey (CE) collects data on expenditures, income, and demographic information for individual respondents. PUMD files provide this information for individual respondents, PUMD files include adjustments for information that is missing because respondents have been adjusted to reduce the likelihood of identifying respondents, either by name or by address.

This format allows researchers to analyze expenditures, income, and demographic information. However, users of the PUMD files should have strong skills with statistical software to explore the data availability with the **NEW** [Interactive CE Visualization Tool](#).

The [2015 public-use microdata](#) is the most recent and was released on August 11, 2015.

Show me the data and the source code

- ▶ I call on all of us, myself included, to think about one set of materials (data + code) you could share on github or your website.



The screenshot shows a web browser window with the URL <https://www.bls.gov/cex/pumd.htm>. The page header includes the United States Department of Labor and Bureau of Labor Statistics logos. A navigation menu contains links for Home, Subjects, Data Tools, Publications, Economic Releases, and Student. The main heading is "Consumer Expenditure Survey". Below this, there is a section for "Public-Use Microdata (PUMD)". A left-hand sidebar lists various options under "BROWSE CE", including CE HOME, CE OVERVIEW, CE NEWS RELEASES, CE DATABASES, CE TABLES, CE PUBLIC-USE MICRODATA, CE SURVEY FORMS, CE PUBLICATIONS, and CE FAQs. The main content area provides an overview of the PUMD files, stating that they collect data on expenditures and provide information for individual respondents. It also mentions that PUMD files include adjustments to reduce the likelihood of identifying respondents and that this format allows researchers to analyze expenditures, income, and demographic information. A link to the "NEW Interactive CE Visualization Tool" is provided. The page also notes that the 2015 public-use microdata is the most recent and was released on August 1, 2016.

- ▶ *“A new paradigm in textbooks? Versions, not editions?”* – Albert Kim/Chester Ismay

Where's that hockey puck going?

- ▶ What will be the theme of USCOTS 2019?

Where's that hockey puck going?

- ▶ What will be the theme of USCOTS 2019?
 - ▶ USCOTS 2019: Educating the next generation of statistical wind turbine technicians

Where's that hockey puck going?

- ▶ What will be the theme of USCOTS 2019?
 - ▶ USCOTS 2019: Educating the next generation of statistical wind turbine technicians
 - ▶ USCOTS 2019: Understanding variability with #NotBogStandard Deviations

Where's that hockey puck going?

- ▶ What will be the theme of USCOTS 2019?
 - ▶ USCOTS 2019: Educating the next generation of statistical wind turbine technicians
 - ▶ USCOTS 2019: Understanding variability with #NotBogStandard Deviations
- ▶ Where do YOU think the puck is heading?