R: Innovating at the Bureau of Labor Statistics

Arcenis Rojas

Economist

Division of Consumer Expenditure Surveys

Federal Committee on Statistical Methodology
March 2018



Overview

- IPP: Division of International Prices
- **PPI**: Division of Industrial Prices and Price Indexes
- **CE**: Division of Consumer Expenditure Survey
- OCWC: Office of Compensation and Working Conditions
- OSMR: Office of Survey Methods and Research



Overview

- Automation (IPP)
- Quality control (PPI)
- Real-time response rates (OCWC)
- Data visualization (CE)
- Other R Shiny applications
- R packages



R Shiny Applications

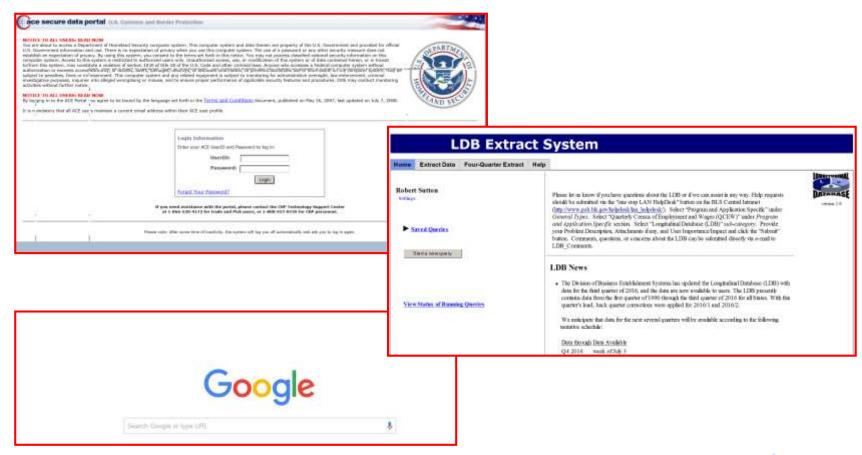


Sample Refinement Automation

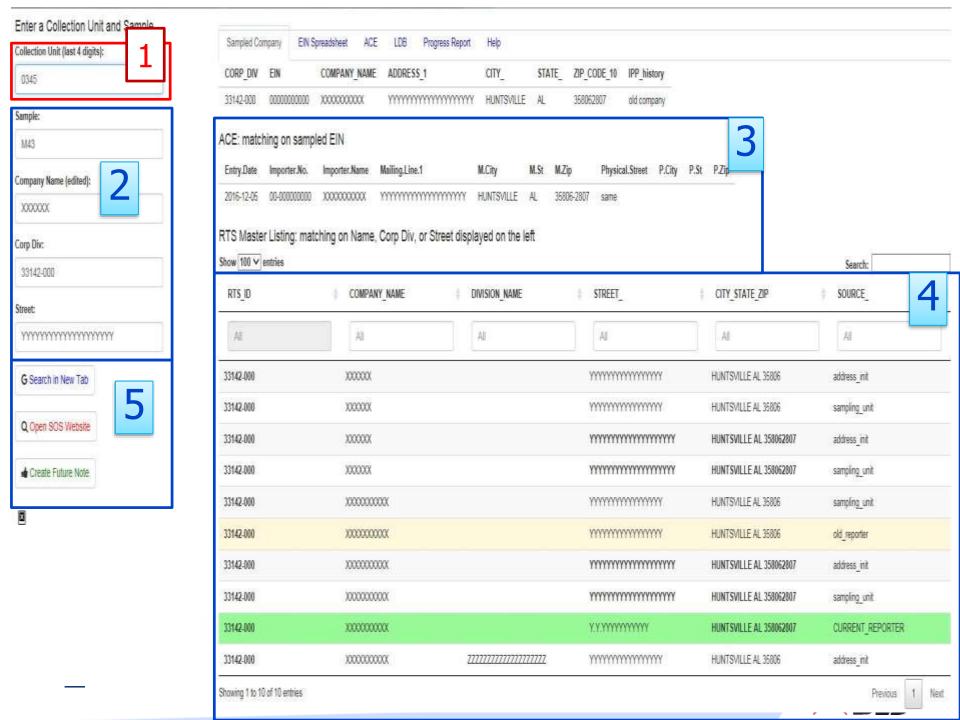
- International Prices Program
 - Receive data from Census and Customs
 - ► Must verify Establishment ID Number (EIN), name, and address to provide to field economists
 - ▶ 1700 export collections units per sample
 - ▶ 2400 import collection units per sample
 - ▶ 6 IPP sample team members
 - ▶ 16 copies, 20 pastes, and 46 clicks per unit



Data Sources







Enter a Collection Unit and Sample

Collection Unit (last 4 digits):
Sample:
X43
Company Name (edited):
Corp Div:
character(0)
Street:
logical(0)
Search in New Tab
Open Google Maps
Open SOS Website
Create Future Note

Left Side



Right Side



ACE: matching on sampled EIN

Entry.Date	Importer.No.	Importer.Name	Mailing.Line.1	M.City	M.St	M.Zip	Physical.Street	P.City	P.St	P.Zip

RTS Master Listing: matching on Name, Corp Div, or Street displayed on the left



Search Results

Enter a Collection Unit and Sample	Company Name Changes								
Collection Unit (last 4 digits):	No company name changes in the reporter or old_co_name_fan tables for Corp Dir 33142-000								
0345j ×	Google Search: company name, city, and state								
Sample:	Search Images those these those those those those								
M43	XXXXXX YYYYYYYYYYYYYYYY HUNTSVILLE AL 35806								
Company Name (edited):									
XXXXXX	All Images Videos News Shopping Mape Books								
Coep Div:									
33142-000	fi mendis								
Street:	Any time Carrabba's Italian Grill in Huntsville, AL Paet hour https://www.carrabbas.com/locations/al/huntsville •								
***********	Past 24 hours Looking for a great Italian restaurant? Bring your family and friends to the Carrabba's Italian Past week Grill location in Huntsville today and enjoy classic Italian dishes!								
G Search in New Tab	Past month Past year Reliphie May:: Stand Up Live Huntsville standuplivehuntsville laughstub com/event ctm?id=476375 •								
Q Open SOS Website	All results Raiphie May - Stand Up Live Huntsville. Stand Up Live Huntsville, AL Two lem minimum. to show time. Please call the Box Office at xxx xxx xxxx.								
▲ Create Future Note	Josh Blue : Stand Up Live Huntsville standuplive huntsville laughstub com/event.cfm?id=476363 + Joeh Blue : Stand Up Live Huntsville . Stand Up Live Huntsville, AL Two item								
8	Kountry Wayne :: Stand Up Live Huntsville standuplivehuntsville iaughstub com/event cfm?id=480893 + Kountry Wayne :: Stand Up Live Huntsville :. April 05, 2017 6:30 PM. Stand Up Live Huntsville :. April 05, 2017 6:30 PM. Stand Up Live Huntsville :. Huntsville :. April 05, 2017 6:30 PM. Stand Up Live Huntsville :. April 05, 2017 6:30 PM. Stand Up Live Huntsville :. April 05, 2017 6:30 PM. Stand Up Live Huntsville :. April 05, 2017 6:30 PM. Stand Up Live Huntsville :. April 05, 2017 6:30 PM. Stand Up Live								
	Chingo Bling: Stand Up Live Huntsville standuplivehuntsville taughstub.com/event.cfm?id=481551 = Chingo Bling: Stand Up Live Huntsville. Stand Up Live Huntsville, Huntsville, AL Two item minimum to show time. Please call the Box Office at xxxx xxxx xxxx								
	Searches related to XXXXXX YYYYYYYYYYYYYYYYYYY HUNTSVILLE AL 35806								
	carrabba's funtavile al coupons stand up live funtavile menu								
	carrabba's italian grill huntsville al 35801 italian restaurants huntsville al								
	carrabba's parkway place mall outback huntoville at								
	bonefish griff huntsville alabama stand up live huntsville al menu								
	Advanced search Search Help Send feedback								
	Google Home Advertising Programs Business Solutions Privacy Terms								



Export Addresses at a Glance



Samp	led Company	EIN Spreadsheet	ACE LDB	Progress Report Help										
IN S	oreadsheet.	matching on san	npled EIN											
pct	export_id	eistrata	name	address	city	st	zip	eistrecs	addrecs	split	flag1	flag2	problem	Mike.s,Commer
33%								2398	800		1	1		
21%								2398	510					
11%								2398	260					
7%								2398	177					
5%								2398	127					
4%								2398	93					
3%								2398	82					



Benefits of Automation

- 80-100 hours per sample of time savings
 - Much less clicking
 - Better and more thorough sample review
 - More time to review more problematic collection units

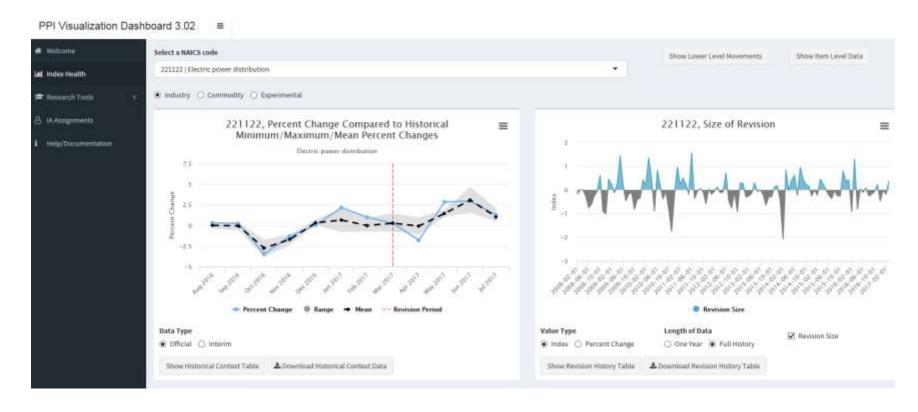


Sample Refinement Automation

- Ara Khatchadourian: khatchadourian.ara@bls.gov
- Rob Sutton: <u>sutton.robert@bls.gov</u>

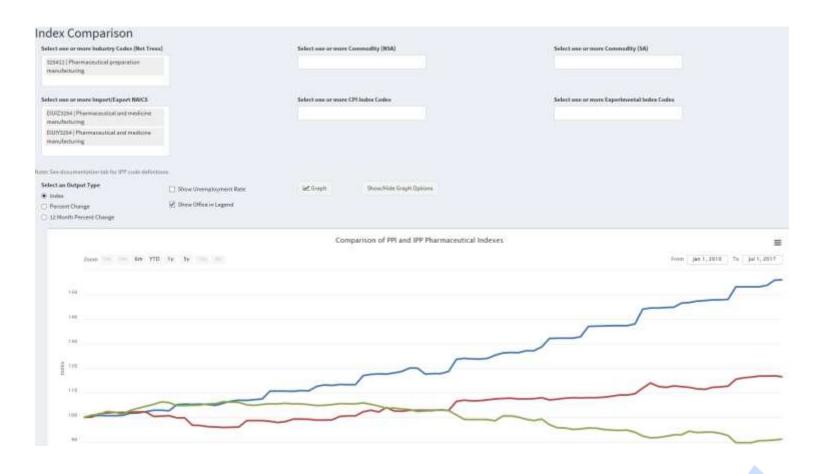


Industrial Prices Visualization Dashboard





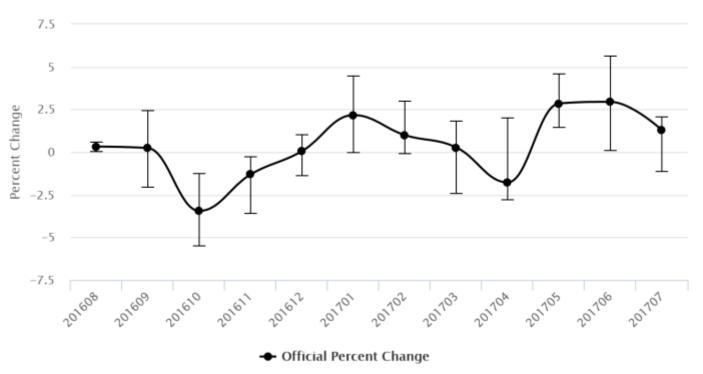
Index Comparisons





Index Review and Revision

95% Confidence Intervals of 221122 | Electric power distribution



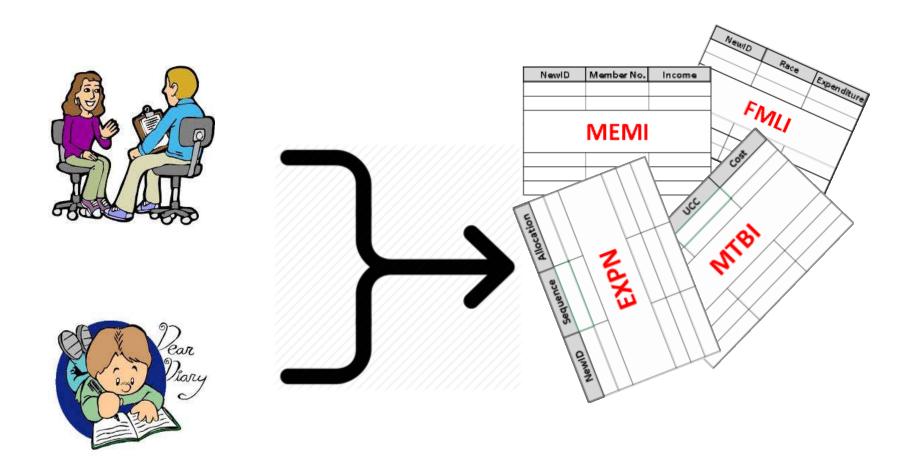
EMBARGOED DATA - NOT FOR PUBLIC RELEASE



Visualization Dashboard

- Neil Wagner: <u>wagner.neil@bls.gov</u>
- Steve York: york.stephen@bls.gov







CE Public-Use Microdata (PUMD)

- Public-Use Microdata
 - ► Family-level characteristics
 - Expenditures by Universal Classification Code (UCC)
 - Member-level characteristics
 - Expenditures and their characteristics by type of expenditure (EXPN... > 50 files each year!)
 - And more!



Files Required for Analysis

Family Characteristics File (34,177 Observations)



Expenditures File (1,720,755 Observations)

	newid ‡	ref_mo	исс ‡	cost ‡
1	02792065	01	210110	650.0000
2	02792075	01	210110	258.3333
3	02792115	01	210110	1370.0000
4	02792125	01	210110	426.0000
5	02792245	01	210110	3189.0000
6	02792265	01	210110	380.0000
7	02792295	01	210110	918.3333
8	02792385	01	210110	675.0000
9	02792455	01	210110	850.0000
10	02792585	01	210110	1500.0000
11	02792815	01	210110	880.0000
12	02793085	01	210110	1050.0000
13	02793345	01	210110	850.0000
14	02793355	01	210110	816.0000
15	02793455	01	210110	419.0000
16	02793525	01	210110	146.0000
17	02793645	01	210110	475.0000
18	02793775	01	210110	800.0000
19	02793805	01	210110	860.0000
20	02793815	01	210110	800.0000
21	02794175	01	210110	196.0000



Required Resources / Skills

```
136 #
137 #
                         Compute Annual Mean Estimates
138 #
141 # Merge Interview CU weights and expenditures
142 int df <- left join(
143
        fmli,
144
        expend %>% filter(ucc %in% getUCCs(expenditure, stub)) %>%
145
           group by(newid) %>% summarise(cost = sum(cost)),
146
        by = "newid"
147 ) %>%
148
        mutate_each_(
149
           funs(replace(., is.na(.), 0)),
           vars = c("cost", pasteO("wtrep", str pad(1:44, 2, "left", 0)))
151
152
153 # Compute an Interview annual mean estimate
154 int gm <- int df %>%
155
        mutate(wt cost = cost * fin1wt21) %>%
157
           grand mean = sum(wt cost, na.rm = TRUE) / sum(calwt)
158
      ) %>% unlist() %>% unname()
160 # Merge Diary CU weights and expenditures
161 dia df <- left join(
162
        fmld.
163
        expend %>% filter(ucc %in% getUCCs(expenditure, stub)) %>%
164
           group by(newid) %>% summarise(cost = sum(cost)),
       by = "newid"
165
166 ) %>%
167
        mutate each (
168
           funs(replace(., is.na(.), 0)),
169
           vars = c("cost", pasteO("wtrep", str_pad(1:44, 2, "left", 0)))
170
171
172 # Compute a Diary annual mean estimate
173 dia gm <- dia df %>%
174
        mutate(wt cost = cost * finlwt21) %>%
175
        summarise(grand mean = sum(wt cost, na.rm = TRUE) / sum(popwt)) %>%
176
        unlist() %>% unname()
```





Introduction

CE Visualization

Methods

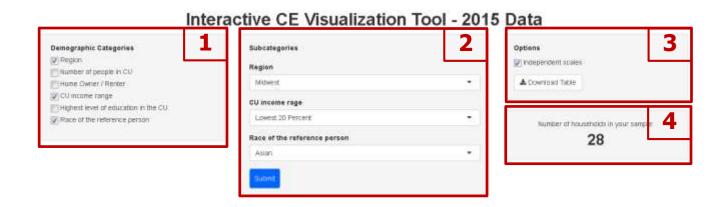
Introduction: Comparisons of reported expenditures

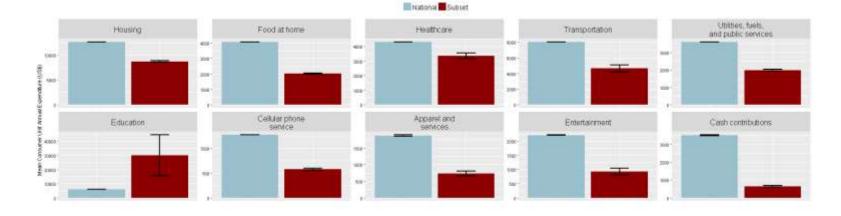
The **Consumer Expenditure Survey (CE)** program consists of two surveys, the Interview and D expenditures, income, and consumer unit (families and single consumers) characteristics. The sur U.S. consumer units (CUs), which we also refer to as households or families.

This application is intended to provide the user an introduction to CE data through an interactive is accompanying table showing comparisons of expenditures between a selected subsample of the (

Click on the CE Visualization tab above to use the application







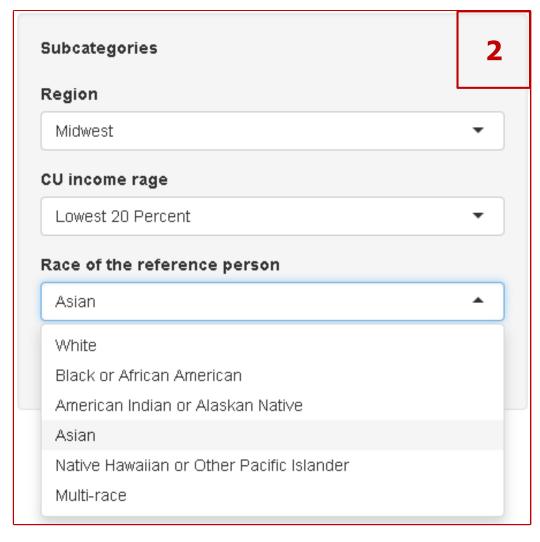


Demographic Categories

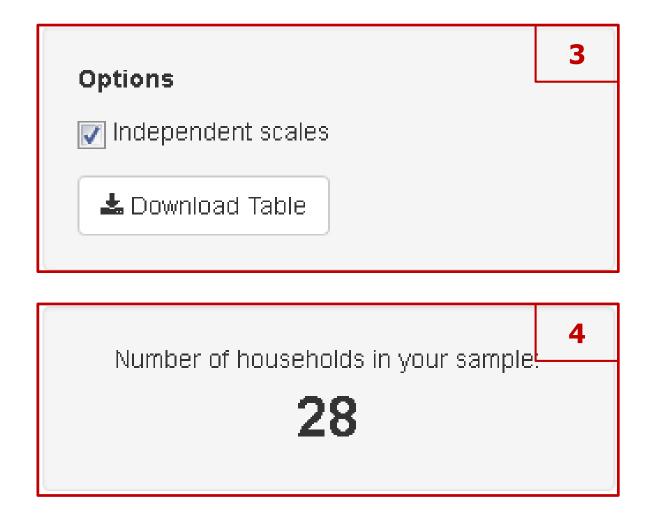
1

- 📝 Region
- Number of people in CU
- Mome Owner / Renter
- 📝 CU income range
- Highest level of education in the CU
- Race of the reference person.

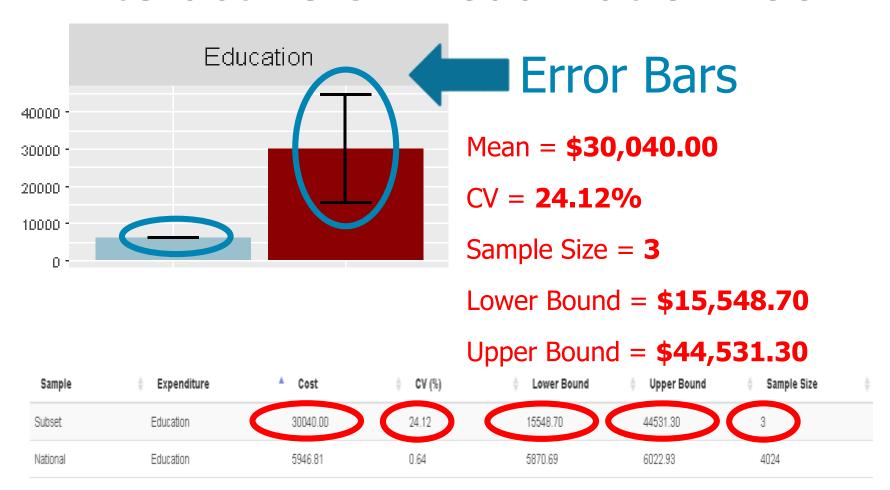














Benefits to the user

- Accessibility: The user can access the app for free as long as they have internet access on a device with a web browser
- Usability: The user operates only the clean, user-friendly UI to get data, results, and visualizations



Arcenis Rojas: rojas.arcenis@bls.gov

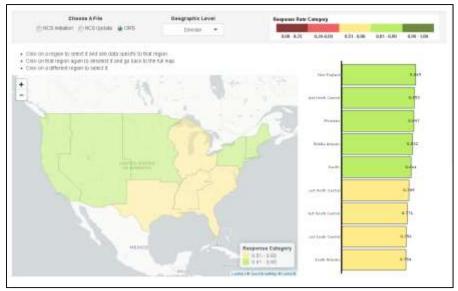


Real-time Response Rate Tool

- Office of Compensation and Working Conditions
- Provide real-time response rates to field offices
 - ► Focus on problem collection areas
 - Improved sample representativity

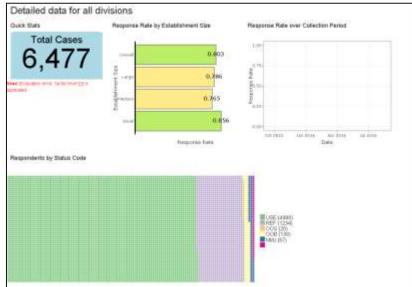


Real-time Response Rate Tool



Response rates by region and/or establishment size

Detailed summaries for each region





Real-time Response Rate Tool

- Brandon Kopp (OSMR):
 kopp.brandon@bls.gov
- Randall Powers (OSMR): powers.randall@bls.gov
- Arcenis Rojas (CE):
 rojas.arcenis@bls.gov



Other Shiny Applications

- Choropleth maps of unemployment data (OSMR)
- Energy Information Administration analyzer (PPI)
- Text analysis Shiny App (Survey Methods)



R Packages



R Packages

- rpms: Recursive Partitioning for Modeling Survey Data package (Survey Methods)
- growfunctions: Bayesian Non-Parametric
 Dependent Models for Time-Indexed
 Functional Data package (Survey Methods)



rpms

- Fits a linear model to survey data in each node obtained by recursively partitioning the data.
- Adjusts for complex sample design features used to obtain the data.
- Produces design-consistent coefficients to the least squares linear model between the dependent and independent variables.



rpms

- The main function returns the resulting binary tree with the linear model fit at every endnode.
- Daniell Toth (OSMR): <u>toth.daniell@bls.gov</u>



growfunctions

- Bayesian Non-Parametric Dependent Models for Time-Indexed Functional Data package (Survey Methods)
- Estimates a collection of time-indexed functions under either of Gaussian process (GP) or intrinsic Gaussian Markov random field (iGMRF) prior formulations



growfunctions

- Dirichlet process mixture allows sub-groupings of the functions to share the same covariance or precision parameters
- The GP and iGMRF formulations both support any number of additive covariance or precision terms, respectively, expressing either or both of multiple trend and seasonality.



growfunctions

■ Terrance Savitsky (OSMR):

savitsky.terrance@bls.gov



Challenges



Challenges

- Data confidentiality
- Need for an R server to make apps/programs public
- Can only put Shiny apps on a webpage via iFrames or setting up an account on a cloud server (i.e., Digital Ocean, R Studio)



Contact Information

Arcenis Rojas
Economist
Division of Consumer Expenditure Surveys
www.bls.gov/cex
202-691-6884
rojas.arcenis@bls.gov

