Web Scraping in Support of the Census Bureau's Public Sector Programs

Federal Committee on Statistical Methodology Research and Policy Conference
Washington, DC
October 27, 2022

Hector Ferronato[1,2] and Brian Dumbacher[1]
[1]U.S. Census Bureau
[2]Reveal Global Consulting

Outline

- Introduction
- Scraping Assisted by Learning (SABLE)
- Quarterly Summary of State and Local Tax Revenue (QTAX)
- State Government Finances (STATEFIN)
- Conclusions and Future Work

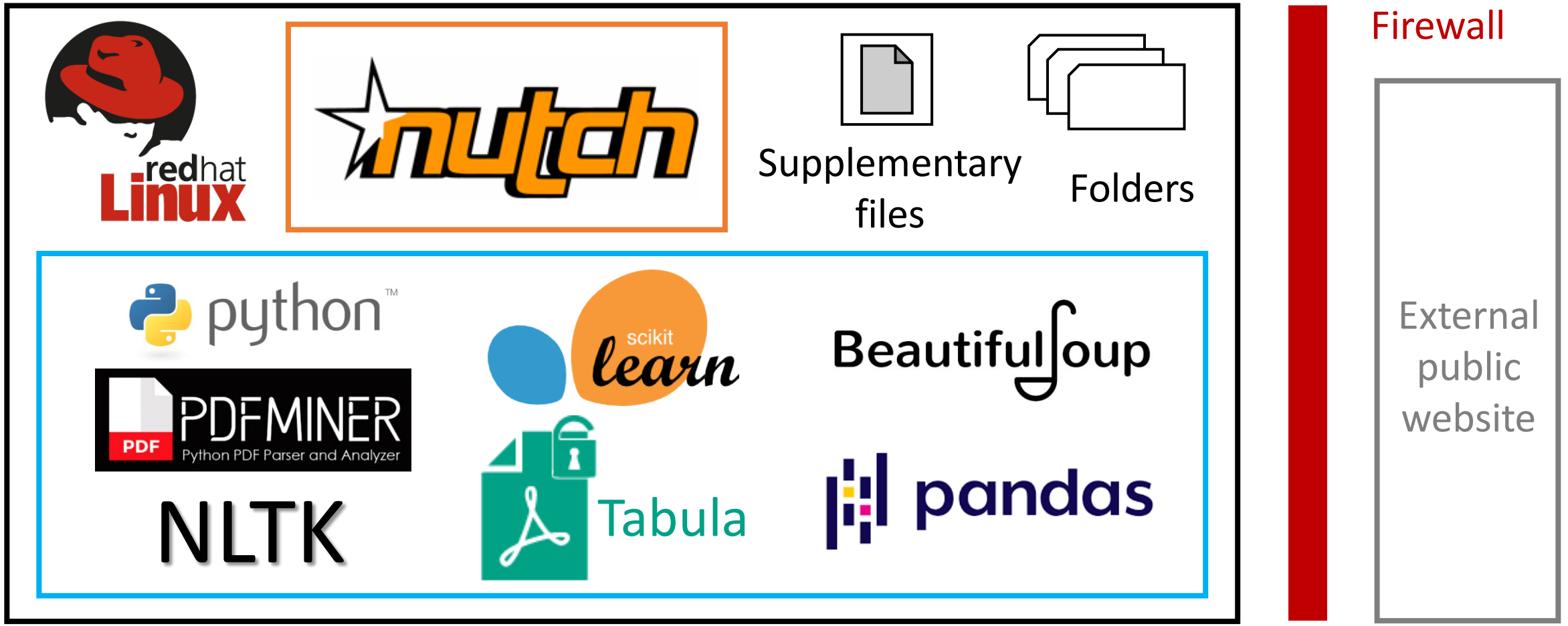
United States® Census Bureau

Introduction

- **The Economic Directorate of the U.S. Census Bureau conducts various surveys of state and local governments**
- **Respondent data can often be found on government websites** in the form of PDF and EXCEL reports.
- **Automated data collection from these websites can reduce burden** for both respondents and Census Bureau analysts.

United States® Census Bureau

SABLE



United States® Census Bureau

QTAX

- QTAX collects **quarterly tax revenue data from state and local** governments.
- Much of this **tax revenue data is publicly available on government websites** on monthly basis.
- **Lack of standardization** of URLs and report formats
- Using **SABLE to automate the data collection**

United States® Census Bureau

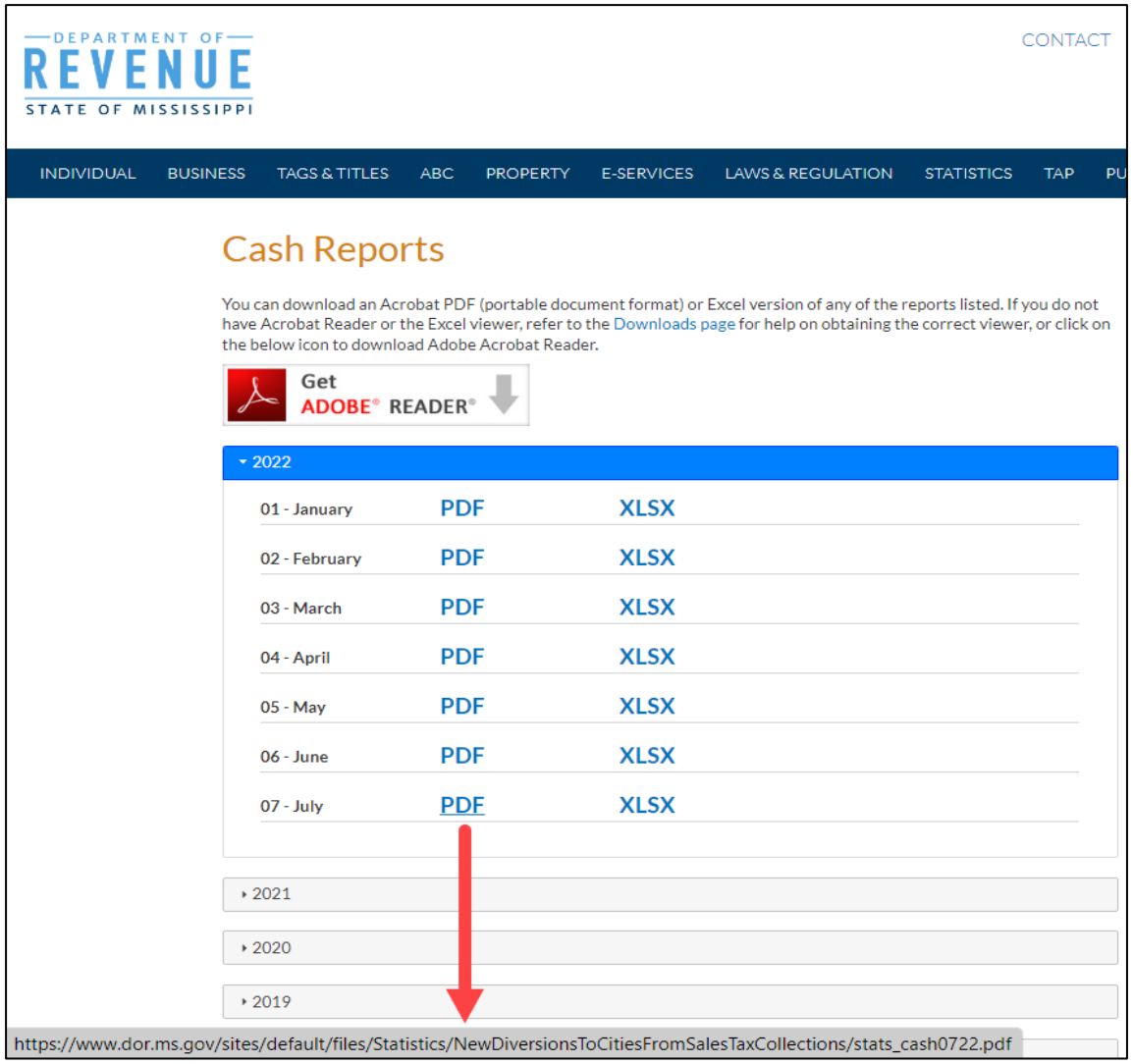
QTAX Downloading Option 1: URL Patterns

- The analysts determine what PDFs to download for each state, and the researchers then propose candidate URLs that are functions of the user-specified year and month. These functions are referred to as URL patterns.
  - https://aoa.vermont.gov/sites/aoa/files/revenue-economy/RevenueRpts/DECEMBER%20-%20FY20%20Revenue%20Press%20Release%201-24-20.pdf

United States® Census Bureau

QTAX Downloading Option 2: Archive Page

- Identify the main archive page that the state uses to publish the tax reports. In this case the Python program visits the archive and searches for the report URL.
- Download with WGET.



United States® Census Bureau

QTAX Scraping

- Analysts tell the researchers what taxes to scrape for each state. The researchers then develop a template that takes the TXT version of the PDF as input and scrapes the following data:
  - Tax names
  - Tax values
  - Time period (e.g., monthly, or fiscal year to date)
  - Units (e.g., dollars or thousands of dollars)

STATE OF MAINE
Undedicated Revenues - General Fund
For the Fourth Month Ended October 31, 2019
For the Fiscal Year Ending June 30, 2020
Comparison to Budget

	Actual	Budget
Sales and Use Tax	$ 149,179,583	$ 144,864,644
Service Provider Tax	4,472,832	5,108,226
Individual Income Tax	145,540,731	146,057,541
Corporate Income Tax	9,790,420	10,000,000
Cigarette and Tobacco Tax	14,689,073	13,373,708
Insurance Companies Tax	8,959,727	8,800,287
Estate Tax	3,112,229	208,414

United States® Census Bureau

QTAX Scraping: Regex Examples

```
m_col = re.search(r"(actual|budget)\s+(budget|actual)", line)
if m_col:
    if m_col.group(1) == "budget" and m_col.group(2) == "actual":
        col = 2
    elif m_col.group(1) == "actual" and m_col.group(2) == "budget":
        col = 1
```

```
m = re.search(r"estate\s+tax\s+\$?\s*([\d,.()]+)\s+\$?\s*([\d,.()]+)",
line)
```

United States® Census Bureau

QTAX Results

Month	Downloading			Scraping		
	Success	Caution	Problem	Success	Caution	Problem
Jan 2022	23	2	1	22	1	0
Feb 2022	23	2	1	23	0	0
Mar 2022	23	2	1	23	0	0
Apr 2022	23	2	1	23	0	0
May 2022	24	2	0	24	0	0
Jun 2022	24	2	0	24	0	0

Total	140	12	4	139	1	0
Efficacy %	**90%**	8%	3%	**89%**	1%	0%
	139 scraped out of 140 downloaded = **99.3% scraping efficacy**					

Classification of results from 26 states in terms of success in downloading and success in scraping

United States® Census Bureau

STATEFIN

- STATEFIN **collects a wealth of data on state finances**:
  - **revenue** by source, **expenditures** by function, **debt** by term, and **assets** by purpose
- Some states provide this data in PDFs published on their websites on a yearly basis.
- The Census Bureau is prototyping **using SABLE for web scraping two California state finance reports**:
  - Legislative, Judicial, and Executive section of the Governor's Budget (report 0010)
  - Business, Consumer Services, and Housing Agency (report 1000)

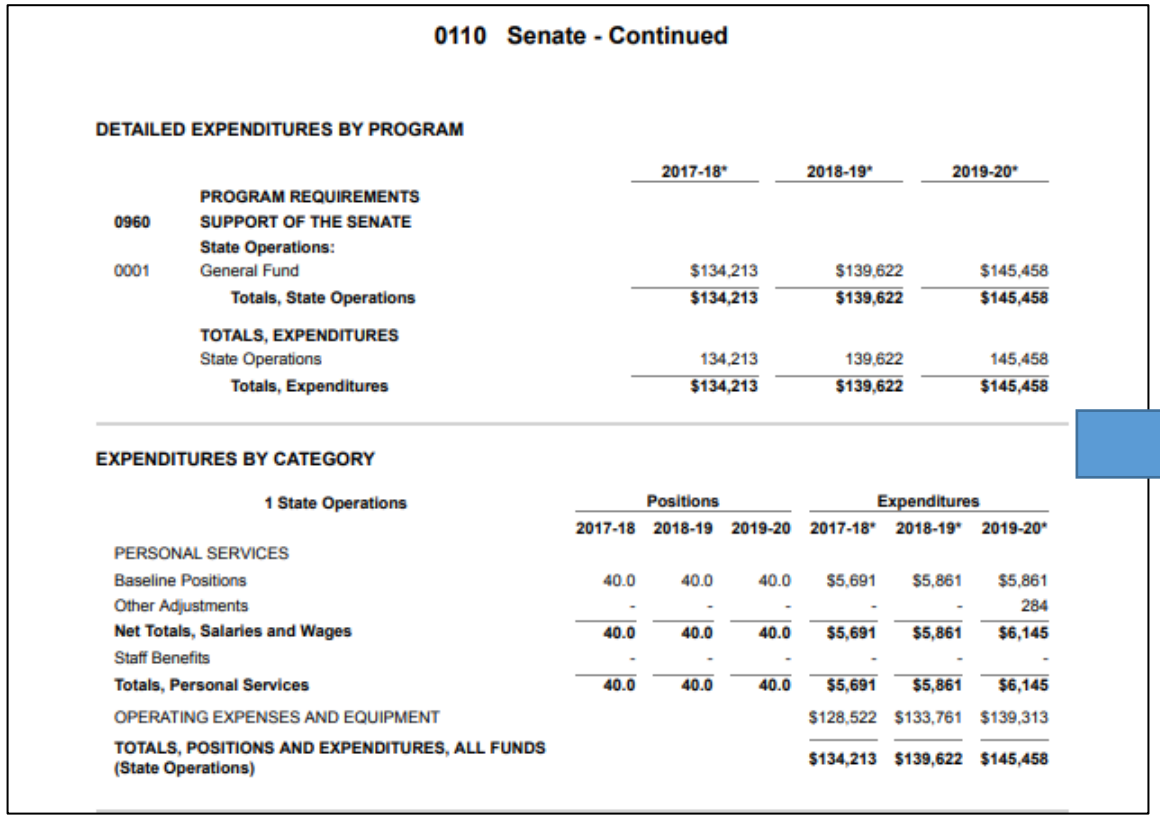
United States® Census Bureau

STATEFIN Downloading

- Each of the State Agencies in the Budget Detail section publishes a PDF report of the entire agency's budget.
- Currently, analysts manually download these reports.

United States® Census Bureau

STATEFIN PDF-to-EXCEL Conversion



0110 Senate - Continued

DETAILED EXPENDITURES BY PROGRAM

		2017-18*	2018-19*	2019-20*
	PROGRAM REQUIREMENTS			
0960	**SUPPORT OF THE SENATE**			
	State Operations:			
0001	General Fund	$134,213	$139,622	$145,458
	Totals, State Operations	**$134,213**	**$139,622**	**$145,458**
	TOTALS, EXPENDITURES			
	State Operations	134,213	139,622	145,458
	Totals, Expenditures	**$134,213**	**$139,622**	**$145,458**

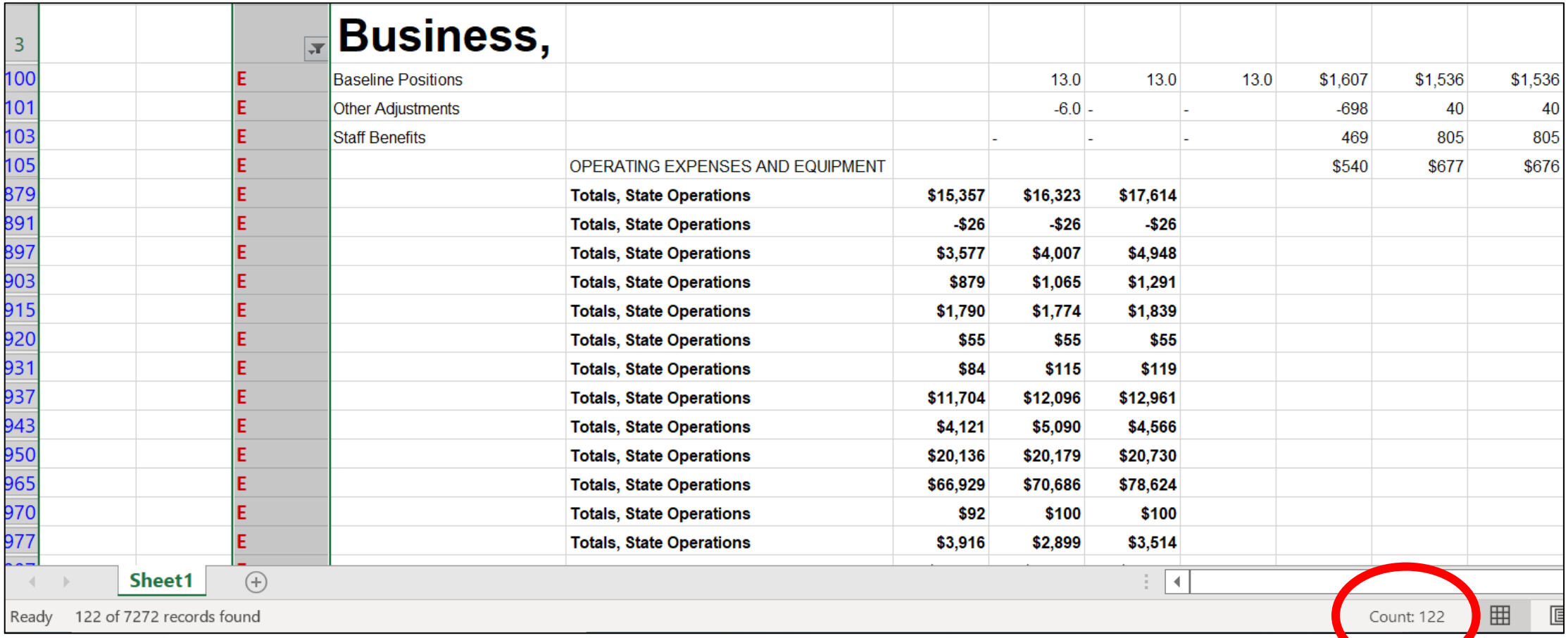
EXPENDITURES BY CATEGORY

1 State Operations	**Positions**			**Expenditures**		
	2017-18	**2018-19**	**2019-20**	**2017-18***	**2018-19***	**2019-20***
PERSONAL SERVICES						
Baseline Positions	40.0	40.0	40.0	$5,691	$5,861	$5,861
Other Adjustments	-	-	-	-	-	284
Net Totals, Salaries and Wages	**40.0**	**40.0**	**40.0**	**$5,691**	**$5,861**	**$6,145**
Staff Benefits	-	-	-	-	-	-
Totals, Personal Services	**40.0**	**40.0**	**40.0**	**$5,691**	**$5,861**	**$6,145**
OPERATING EXPENSES AND EQUIPMENT				$128,522	$133,761	$139,313
TOTALS, POSITIONS AND EXPENDITURES, ALL FUNDS (State Operations)				**$134,213**	**$139,622**	**$145,458**

C	D	E	F	G	H
	Positions			Expenditures	
2020-21	2021-22	2022-23	2020-21*	2021-22*	2022-23*
1,873.6	2,541.3	2,600.2	$754,389	$692,123	$744,183
428.5	536.6	510.6	102,596	126,797	121,003
-	-	-	67,836	67,836	69,368
804.5	1,012.7	883.7	148,936	173,803	150,165
111.2	190.3	181.4	15,856	27,292	25,135
117.0	140.3	134.5	22,609	25,104	24,074
24,532.7	24,148.4	26,520.3	4,246,803	5,029,431	4,895,136
6,062.2	7,506.4	7,691.4	1,725,420	1,864,255	1,856,083
101.7	69.2	-	48,685	38,050	-
2,537.1	2,930.4	3,021.4	838,716	853,005	717,708
1,741.0	1,861.4	1,929.7	347,863	399,725	369,441
114.9	169.5	175.6	235,081	223,310	234,971

United States® Census Bureau

STATEFIN Labeling



3			Business,								
100		E	Baseline Positions			13.0	13.0	13.0	$1,607	$1,536	$1,536
101		E	Other Adjustments			-6.0	-	-	-698	40	40
103		E	Staff Benefits			-	-	-	469	805	805
105		E		OPERATING EXPENSES AND EQUIPMENT					$540	$677	$676
879		E		**Totals, State Operations**	**$15,357**	**$16,323**	**$17,614**				
891		E		**Totals, State Operations**	**-$26**	**-$26**	**-$26**				
897		E		**Totals, State Operations**	**$3,577**	**$4,007**	**$4,948**				
903		E		**Totals, State Operations**	**$879**	**$1,065**	**$1,291**				
915		E		**Totals, State Operations**	**$1,790**	**$1,774**	**$1,839**				
920		E		**Totals, State Operations**	**$55**	**$55**	**$55**				
931		E		**Totals, State Operations**	**$84**	**$115**	**$119**				
937		E		**Totals, State Operations**	**$11,704**	**$12,096**	**$12,961**				
943		E		**Totals, State Operations**	**$4,121**	**$5,090**	**$4,566**				
950		E		**Totals, State Operations**	**$20,136**	**$20,179**	**$20,730**				
965		E		**Totals, State Operations**	**$66,929**	**$70,686**	**$78,624**				
970		E		**Totals, State Operations**	**$92**	**$100**	**$100**				
977		E		**Totals, State Operations**	**$3,916**	**$2,899**	**$3,514**				

Sheet1

Ready 122 of 7272 records found Count: 122

122 labeled objects of interest

United States® Census Bureau

STATEFIN Scraping

	A	B	C	D	E	F	G
106	E66	2120 Alcoholic Beverage Control Appe	1650 ADMINISTRATIVE REVIEW	Baseline Positions	$ 276.00	$ 381.00	$ 424.00
107	E66	2120 Alcoholic Beverage Control Appe	1650 ADMINISTRATIVE REVIEW	Other Adjustments	$ 284.00	$ 178.00	$ 238.00
108	E66	2120 Alcoholic Beverage Control Appe	1650 ADMINISTRATIVE REVIEW	Staff Benefits	$ 293.00	$ 298.00	$ 331.00
109	E66	2120 Alcoholic Beverage Control Appe	1650 ADMINISTRATIVE REVIEW	OPERATING EXPENSES AND EQUIPMENT	$ 244.00	$ 348.00	$ 348.00
110	B50	2240 Department of Housing and Com	1650 ADMINISTRATIVE REVIEW	Federal Trust Fund	$ 136,646.00	$ 3,400,660.00	$ 137,245.00
111	E66	2240 Department of Housing and Com	1660 CODES AND STANDARDS PROGRAM	Totals, State Operations	$ 35,570.00	$ 39,058.00	$ 40,455.00
112	M66	2240 Department of Housing and Com	1660 CODES AND STANDARDS PROGRAM	Totals, Local Assistance	$ -	$ 250.00	$ 250.00
113	E89	2240 Department of Housing and Com	1665 FINANCIAL ASSISTANCE PROGRAM	Totals, State Operations	$ 288,375.00	$ 202,348.00	$ 112,243.00
114	M89	2240 Department of Housing and Com	1665 FINANCIAL ASSISTANCE PROGRAM	Totals, Local Assistance	$ 2,625,537.00	$ 6,343,481.00	$ 6,296,511.00
115	E50	2240 Department of Housing and Com	1670 HOUSING POLICY DEVELOPMENT PR	Totals, State Operations	$ 10,266.00	$ 13,497.00	$ 18,559.00
116	M50	2240 Department of Housing and Com	1670 HOUSING POLICY DEVELOPMENT PR	Totals, Local Assistance	$ 25,424.00	$ 266,833.00	$ 613,750.00
117	E50	2240 Department of Housing and Com	1675 CALIFORNIA HOUSING FINANCE AGE	Totals, State Operations	$ 33,852.00	$ 36,149.00	$ 37,892.00
118	M50	2240 Department of Housing and Com	1680 LOAN REPAYMENTS PROGRAM	Totals, Local Assistance	$ (5,856.00)	$ (1,944.00)	$ (1,944.00)
119	E50	2240 Department of Housing and Com	1685 HPD DISTRIBUTED ADMINISTRATIOI	Totals, State Operations	$ (19.00)	$ (179.00)	$ (180.00)
120	E66	2320 Department of Real Estate	1700 DEPARTMENT OF REAL ESTATE	Baseline Positions	$ 22,551.00	$ 22,999.00	$ 22,999.00
121	E66	2320 Department of Real Estate	1700 DEPARTMENT OF REAL ESTATE	Other Adjustments	$ 2,152.00	$ (1,572.00)	$ 1,212.00
122	E66	2320 Department of Real Estate	1700 DEPARTMENT OF REAL ESTATE	Staff Benefits	$ 13,126.00	$ 12,419.00	$ 13,482.00
123	E66	2320 Department of Real Estate	1700 DEPARTMENT OF REAL ESTATE	OPERATING EXPENSES AND EQUIPMENT	$ 13,593.00	$ 18,906.00	$ 19,014.00
124							
125							
126							

1000_guide_with_values_and_trus

Ready

Count: 122

122 labeled objects scraped

United States® Census Bureau

STATEFIN Results

- **122 out of 122 (100%)** labeled objects were scraped from report 1000
- **193 out of 306 (63%)** labeled objects were scraped for report 0010
- These results indicate the scraping program needs to be more flexible in order to find the remaining objects and collect their values.
- Overall, these are positive results for automatic scraping the dollar values for each combination of department, program, and object.

United States® Census Bureau

Conclusions

- The **QTAX** project involves scraping a **relatively small number of tax revenue items from 26 different state** government websites and reports.
- The **STATEFIN** prototype involves **scraping many financial items from a limited number of documents** on the California state government website.
- Results so far for **both QTAX and STATEFIN are positive**. The downloading and scraping methodologies can handle inconsistencies in URLs and document layouts.
- **Refinement is needed**, but progress has been made towards automating a substantial portion of what used to be manual work.

United States® Census Bureau

Future Work

- Based on the encouraging results obtained so far for QTAX and STATEFIN, there is also potential in **applying similar methodology to other public sector surveys**. For example, other surveys of state and local governments collect **data on public retirement systems and employment and payroll**.
- Lastly, to be transparent with respondents and the greater public, there are plans to update the publicly available SABLE GitHub repository with the most recent Python code. This repository is located at https://github.com/uscensusbureau/SABLE and received its last major update after the initial PDF methods for QTAX were developed.

United States® Census Bureau

Contact Information

- Hector.R.Ferronato@census.gov
- Brian.Dumbacher@census.gov

United States® Census Bureau