Autocoding open-ended text into item codes in the Consumer Expenditure Diary Survey

Federal Committee on Statistical Methodology 2026 Annual Conference
March 12, 2026

Consumer Expenditure (CE) Surveys Quick Facts

The surveys are the only federal government data collection effort to obtain information on the complete range of consumers' expenditures, income, and demographic characteristics, directly from consumers.

2
Surveys

The **Interview** survey collects detailed data on major and/or recurring expenditures for periods of 3 months or longer; the **Diary** survey collects records for smaller, more frequently purchased items.

Data are collected by Census for BLS: **Interview** via personal interview, **Diary** via a respondent self-administered diary.

CE data are used by the **Consumer Price Index** to weight its price indexes, and by researchers, other governmental agency statistics and private sector organizations.

The Diary Survey

ILLUSTRATIVE EXAMPLE

Clothing, Shoes, Jewelry, and Accessories

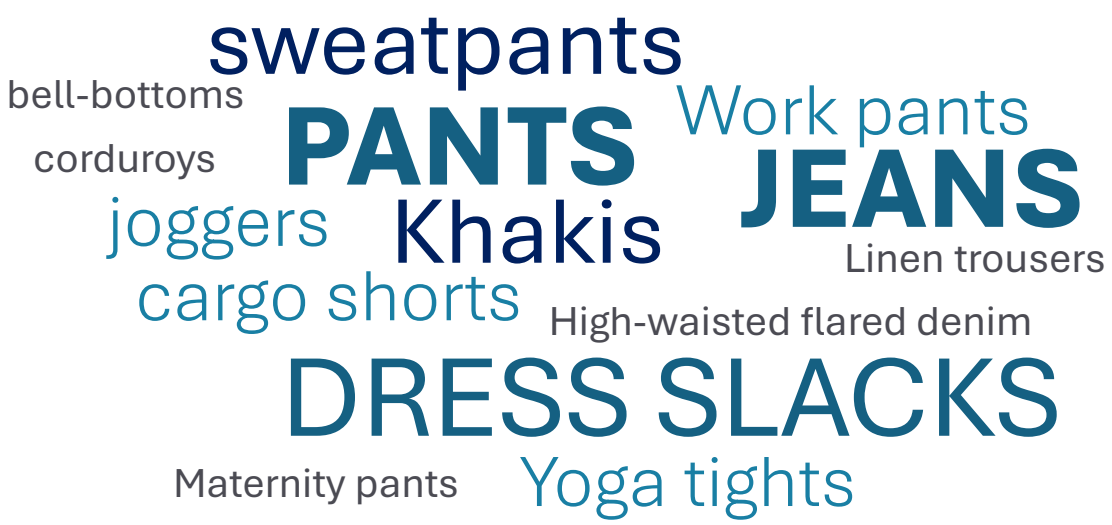
What did you buy or pay for?	Cost without tax	Was the item for: Child Under 2	Boy 2-15	Girl 2-15	Man 16 & over	Woman 16 & over	Name of Store or Website where purchased
dress shirts	*75 00*	1	2	3	4	5 X	*Dillards.com*
running shoes	*69 00*	1	2	3	4	5 X	↓
wallet	*29 00*	1	2	3	4 X	5	↓
baseball cap	*14 99*	1	2 X	3	4	5	*Target*
bib	*3 50*	1 X	2	3	4	5	*Sweet Dreams boutique*
necklace	*250 00*	1	2	3	4	5 X	*Olde Towne jewelry*
non-prescription sunglasses	*59 00*	1	2	3	4	5 X	*Walmart.com*
~~*child's costume (returned for refund)*~~	~~*15 00*~~	1 X	2	3	4	5	~~*Partysupply.com*~~

Daily expenses are **recorded directly by the respondent** over two consecutive one-week periods

Four expenditure sections:

- Clothing, shoes, jewelry, and accessories
- Food and drinks for home consumption
- Meals outside the home
- All other expenditures

Item Coding Example



All are grouped together in

PANTS AND SHORTS

Item code **410240**

CE Diary Autocoder Overview

← MEALS OUTSIDE THE HOME → ← ALL OTHER EXPENDITURE SECTIONS →

{Used for additional validation against ML predictions}

RULES-BASED

Applies **if-then logic** to link the Diary entry to a corresponding item code

DIRECT MATCH

Looks for **identical text string matches** between the Diary entry and item code descriptions

MACHINE LEARNING

Predicts an item code using a **classification algorithm** learned from labeled historical data

SIMILARITY RATIO

Looks for **very similar text strings** between the Diary entry and item code descriptions

COMBO CODES

Looks for best possible matches for **multiple expenses within one Diary entry** *(e.g., "shirts and pants")*

*An estimated 10-20% of this subset will have low probability of a correct prediction and be **flagged for human review***

Method Details and Results

DIRECT MATCH

- Multi-step processing of the incoming Diary entry
- Compare against a maintained robust dictionary
- ~50% of records can be direct matched

MACHINE LEARNING

- Random Forest model built for each of the 4 record types using 2 years of training data
- Accuracy, Precision, Recall, and F1 used to evaluate model performance
- Low confidence predictions are flagged for human review

SIMILARITY RATIO

- For all records with an ML prediction, similarity ratio is calculated for the processed Diary entry and the predicted item code description
- *Formula:*

$$\frac{2 * \text{number of matching characters}}{\text{Total number of characters}}$$

Outcomes

- **Reduce processing time:** Sped up item code assignment by more than 70%
- **Uphold data integrity:** No increase in volume of item code misclassifications
- **Cost savings:** Reduced the cost of Diary digitization by 13%

Practical Considerations

- **Human intervention** is an important component of the model building and output review processes
- **Model size matters** given the infrastructure
- **Deployment planning** must be thoughtful, involve all stakeholders, and offer generous time frames
- **Troubleshooting in production:** the ability to react in a timely manner
- **Planning for the future:** server needs and staff skillsets to support Autocoder

Thank you!

pollock.melissa@bls.gov