Technical Appendix

For <working citation>Table of Contents

[Data Description 1](#_Toc458347030)

[Data Available According to Year 1](#_Toc458347031)

[Data Definitions 1](#_Toc458347032)

[Data Gathering 2](#_Toc458347033)

[Step 1: Download Data 3](#_Toc458347034)

[Step 2: Split Data into Common Groups 3](#_Toc458347035)

[Step 3: Formatting the Groups: 4](#_Toc458347036)

[Step 4: Splitting the Fields 5](#_Toc458347037)

[Step 5: Merge Data 5](#_Toc458347038)

[Step 6: Clean Up 5](#_Toc458347039)

[Step 7: Analysis 6](#_Toc458347040)

[Step 8: Merge Summary Tables 6](#_Toc458347041)

[Step 9: Clean Up 6](#_Toc458347042)

[Step 10: Analysis 6](#_Toc458347043)

[Data Preparation 7](#_Toc458347044)

[Data Cataloguing 7](#_Toc458347045)

[Species 8](#_Toc458347046)

[Reason 8](#_Toc458347047)

# Data Description

The data used in this analysis was obtained through a collection of records provided by the United States Department of Agriculture. The USDA provides public access to food safety recalls by providing access to an archive of press releases. The dates for the archives range from 1994 through 2016 and are categorized by year. Shortcomings observed in the summary records provided for each year include a varying list of categories to describe the reasons for recalls, and missing amounts of the pounds recovered in the years 2000 through half of 2005. Additionally, the USDA provides a summary statement to accompany each year, which include a breakdown of the aforementioned metrics:

* Year of Recall
* Total number of recalls and the total pounds recalled, listed by
  + Class (I, II, III)
  + Reason (STEC, Listeria, Undeclared Allergen, etc.)
  + Species (Beef, Mixed, Pork, Poultry)

## Data Available According to Year

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Detail | 2006 - 2015 | 2005 | 2004 - 2000 | 1994 – 1999 |
| Recall Number | Y | Y | Y | Y |
| Recall Date (opened) | Y | Y | Y | Y |
| Recall Date (closed) | N | N | N | Y |
| Recalled (lbs.) | Y | Y | Y | Y |
| Recovered (lbs.) | Y | P | N | Y |
| Recall Class | Y | Y | Y | Y |
| Product | Y | Y | Y | Y |
| Product Description | Y | Y | Y | Y |
| Problem Type | Y | Y | Y | Y |
| Annual Summary Table | Y | Y | N | N |
| Discovery Method | Y | Y | Y | Y |
| Distribution | N | Y | Y | Y |
| Recall Notification Level | N | Y | Y | Y |

Y = Yes N = No P = Partial

## Data Definitions

**Recall Number:** A uniquely assigned identifier for each recall issued

**Recall Date (opened):** Date marking the beginning of the recall and press release issue

**Recall Date (closed):** Date when the recall was closed.

**Recalled (lbs.):** The total weight of the recall measured by pounds

**Recovered (lbs.):** The total weight of the amount recovered measured by pounds

**Recall Class:** The severity of the recall categorized by Class I, II, and III

**Product:** A short description of the item being recalled, e.g. “Ground Beef” or “Pates & Mousses”.

**Product Description:** A longer description about the product, including any identification codes.

**Problem Type:** A description containing the reason for the recall, such as the product being misbranded or containing allergens.

**Annual Summary Table:** A summary table was created for some years, containing number breakdowns for various categorizations assigned by a USDA analysist. The groups identified as the recall reasons change each year.

**Discovery Method:** A description about how the problem was originally discovered.

**Distribution:** A list of the states where the products were distributed.

**Recall Notification Level:** Which consumers or agencies were notified and how were they notified.

# Data Gathering

The food recall data was published by the USDA onto their website and requires browsing by the specific year for the recall. However, the data is not easily downloaded into a self-contained spreadsheet or database file and instead presents some challenges: the data is formatted differently depending on the year, and there are discrepancies between the records listed on the main page for each year and records listed on an available summary page.

With food recall records spread out between 24 different years and structured in various formats, a script was written in the Python programming language to download each record, parse, and place the results into a database. Any footnotes added to the page by USDA officials was placed into a separate text file for referencing.

The data was gathered using four different scripts written in Python to gather the data, according to similarities in the data structures:

* 1994 to 1999
* 2000 to 2004
* 2005
* 2006 to 2015

## Step 1: Download Data

**Place Each table into a separate Excel sheet: one sheet per year**

**Is Summary Available?**

**No**

**Browse Every Year of Data**

**Begin Data Download Process**

**Save Tables to Excel Document**

**End Data Download Process**

**Yes**

**Copy Table from Web Page**

**Download Summary Table**

## Step 2: Split Data into Common Groups

**Copy and Paste Tables with similar layouts into new Excel Document**

**Browse Every Excel Table**

**Begin Splitting into Common Groups**

**End Splitting into Common Groups**

**Save each Excel document by group**

## Step 3: Formatting the Groups:

**Save Excel document as CSV file**

**Open in Excel**

**Begin Formatting 1994 to 1999**

**Save CSV file**

**Apply Regular Expressions**

**Open CSV in Plain Text Editor**

**Save CSV file**

**Clean Data using OpenRefine**

**End Formatting for 1994 to 1999**

**Import CSV into OpenRefine**

**Selenium: Identify appropriate values using XPath and RegEx Queries**

**Python: Begin loop through each year in group**

**Python: Start iNotebook**

**Begin Formatting Groups**

**Open in Excel**

Begin loop through each group

**Java: Start Selenium Web Driver**

**Connect Python to Excel using XLWings Python Library**

**Python: Copy record values into Excel sheet**

**Selenium: send command to web browser to load web page with table data for year**

**Save as CSV file**

**Python: End year loop**

**End category loop**

**End Formatting Groups**

## Step 4: Splitting the Fields

**Load each CSV into OpenRefine**

Begin loop through each group

**Begin Splitting Fields in Groups**

**End category loop**

**End Splitting Fields in Groups**

**Save as CSV file**

**Create new columns and match values with RegEx queries**

## Step 5: Merge Data

**Load each CSV into SQLite DB**

Begin loop through each group

**Begin Merging Data Process**

**End Merging Data Process**

**Export Single Table into CSV**

**Join All Tables into Single Table**

**End category loop**

## Step 6: Clean Up

**Add “Reason” Column and Cluster**

**Export Data into CSV file**

**Add “Type” Column and add Types**

**Convert Dates to Date Field, Clean Up Misformatting**

**Convert Recalled Amount to Numbers, Drop Non-Numbers**

**Correct Spellings for Recall Class**

**Load CSV into OpenRefine**

**Convert Recovered Amount to Numbers**

**Begin Clean Up Process**

**End Clean Up Process**

## Step 7: Analysis

**Jupyter: Create Charts using Bokeh**

**Jupyter: Load Data into Pandas**

**Python: Load CSV into Jupyter Notebook**

**Begin Analysis Process**

**Decide to Gather Additional Records**

**Realize Incomplete Records**

**Inspect and Review Data**

**End Analysis Process**

## Step 8: Merge Summary Tables

**Repeat Step 4**

**Repeat Step 5**

**Export Data as CSV for Step 9**

**End Merging Summary Tables**

**Begin Merging Summary Tables**

**Match Existing Data with Summary Data. Create New Records where Match Not Found**

**Repeat Step 3 with Python code aimed at summary tables**

## Step 9: Clean Up

**Begin Clean Up Process**

**Load CSV into OpenRefine**

**Correct Spellings for Recall Class**

**Convert Recalled Amount to Numbers, Drop Non-Numbers**

**Convert Recovered Amount to Numbers**

**Convert Dates to Date Field, Clean Up Misformatting**

**Add “Type” Column and add Types**

**Add “Reason” Column and Cluster**

**End Clean Up Process**

**Export Data into CSV file**

## Step 10: Analysis

**Copy tables from 1994 to 1998 into CSV**

**Realize that 1994 to 1998 is missing**

**Review Records**

**Begin Analysis Process**

**Attach data to end of existing data from Step 9**

**Repeat Step 4**

**Repeat Step 5**

**Create Summary Stats**

**End Analysis Process**

# Data Preparation

Each category for yearly data was placed into a separate database table so that efforts to prepare each dataset for analysis could be compartmentalized. Provenance for any changes made to the data was tracked through a series of entries made to a log file. A file and folder hierarchy was created on the computer for exporting data before any significant changes were made. The system allows for each record to be traced throughout every state of change.

We used a software tool named OpenRefine to assist with making bulk changes to the record. For example, when converting the descriptive amounts of food recalls from text (4.2 million lbs) to numerical (4200000), OpenRefine allowed us to identify records by creating filters, tagging matching records, and then applying changes.

## Data Cataloguing

**Most of the records we captured were not tagged by the USDA into specific groupings for species and reason. This required manually reviewing and tagging each record using the information available.**

For tagging species, we looked for specific keywords (pork, veal, beef, etc.), but in the event a product was identified as “sausage” without any further information, then we tagged the species as “mixed”. Best attempts were made to determine the species of the product by reviewing all of the information available for that specific record. A review of the manually tagged records compared to the summary tables provided by the USDA showed a tendency to favor identifying species as “mixed”.

A similar approach was used for categorizing the reason for recall. Many records could be easily identified as belonging to a specific reason, such as containing a known allergen. Though there were many cases in which the reason for recall could span multiple categories; in such cases, the best matching category was chosen instead of placing the record into more than category and counting the records twice.

We created a record table for tracking the keywords used for identifying the species and reason for recall:

### Species

**Beef:**

beef, hamburger, steak, bison, cheeseburger, veal, buffalo, cattle, ribeye, lamb, ovine.

**Mixed:**

beef and pork, beef and poultry, beef and chicken, beef and ham, chicken and ham, turkey and ham, turkey and pork, pork and veal, chicken and pork, Vienna sausage, bacon and beef, bacon and cheeseburger, hot dog, sausage, meat, meatball chorizo, pate, salami, bologna, franks, bolgna, weisswurst, eggroll, head cheese, quiche, dog, mortadella, various, oascar mayer low fat lunchable cracker stacker, chili, carne, blintze, scrapple, burger, pot roast, enchilada, stew, potsticker, spaghettios, bratwurst, roast, salad, beans, souse, soup, Campbell’s, braunschwiger, pizza, noodle, frozen boneless rib-shaped patties with bbq sauce, smoked, basterma, ravioli, burrito, weiner, ready-to-eat, mortadella, jambalaya, baby food, products, liver, pastry, calzone, pasta kielbasa, links, lasagna, entrée, meal, hot-pot, pot pie, pot sticker, tamale, boudin

**Pork:**

pork, bacon, ham, prosciutto, pig, baby back rib

**Poultry:**

chicken, turkey, chicken, duck, poultry, egg

### Reason

**Extraneous Material:**

plastic, metal, foreign, extraneous, drug residue, ammonia, pesticide, chemical contamination, glass, residue, non-food grade lubricating oil

**Listeria**

**Listeria, Listeria monocytogenes, L. monocytogenes**

**Other:**

chlorfenvinphos, irradiation, import, unapproved, insanitary, ineligible, uninspected, excess nitrites, incorrectly declared as “sodium”, nitrite leels in excess, may contain specified risk materials, incorrectly labeled, species violation, botulinum, identity, spongiform, inspection, labeled as turkey breast but contained cured pork, adult, incorrectly lists beef as only meat ingredient, beef franks packages contain franks made with chicken and pork, recalls pork products due to misbranding, recalls chicken and pork products due to misbranding, recalls beef and pork products due to misbranding, ham rather than turkey, mistakenly labeled, mislabeled, abuse, prohibited, staph, unapproved substance, mistakenly placed in packages labeled as containing cheeseburger stuffed sandwiches, incorrect nutrition facts information panel, mislabeling of individual pouches that were packed with lasagna with meat sauce, spicy vegetable potstickers due to a mislabeling of individual pouches were packed with chicken potstickers, these products are considered misbranded because the USDA mark of inspection does not appear on the package labels …, mispackaged and labeled as Salisbury steak products, the product label does not include an accurate list of ingredients due to a printing error, the product was produced using the wrong inspection legend, product was mistakenly labeled as corned beef on the product label, the products contain sodium nitrite which is missing the word nitrite on the product label, … some packages may not bear the USDA mark of inspection, … products were slaughtered under religious exemption which was not declared on the product label, the product contains pork which is not declared on the product label

**Processing Defect:**

under, formulation error, recalls pork products due to possible processing deviation, process

**Salmonella:**

salmonella, salmonella & listeria

**STEC:**

e. coli

**Undeclared Allergen:**

allergen, soy, whey, not declaring eggs, gluten, contains soy and wheat, … chicken and turkey were wrapped in hog and sheep casings … hog and sheep casing may elicit allergic reactions …

**Undeclared Substance:**

may contain, msg, sodium bisulfite, undeclared ingredient, does not list sodium nitrite, nitrite was omitted from label, may contain edible and/or inedible beef that are not declared on the label, may contain pork that is not listed, contains chicken but is not declared, … mislabeling since monosodium glutamate (MSG) was not declared on label, … misbranded and may contain potassium sorbate …, … may contain … not declared on the label …, msg, company employee discovered that the ingredient monosodium glutamate (MSG) was used …