Final Project Paper

DSC550-302T

Katie Briggs

Recalls on consumer products happen when a large quantity of items is no longer safe for use. Items that are frequently recalled include vehicles, medications, foods, and products (i.e. small choking hazards for infants). On average, 400 products per year are recalled by the Consumer Product Safety Commission (CPSC). In 2016, clothing products were associated with 427,500 injuries that resulted in an Emergency Room visit. Currently, there are 641 people with Cyclospora infections who have reported consuming a recalled product of Fresh Express Salad mix. Cyclosporiasis is a parasite infection. Out of those 641, 34 people have been hospitalized. The Centers for Disease Control and Prevention’s mission statement is CDC 24/7: Saving Lives, Protecting People. While most recalls are voluntary from the manufacturer, CPCS works to get recall awareness out to people. However, there are critics that state that there are shortfalls in the overall communication process of making the public aware of recalls. Most recalls are posted on government websites or on the CDC website. Making the consumer responsible for policing the consumer goods that they have purchased.

I chose recalls as my proposed final project topic due to the thought of, why is there not a better process for identifying recalls quicker and a more efficient way of notifying the person that bought the recalled item. While preparing for my project management professional career I learned that it is a basic human need to feel safe and secure. According to Maslow’s Hierarchy of Needs Theory, from his 1943 paper “*A Theory of Human Motivation*” [4], psychological needs and safety needs are one of the more basic needs that humans look for. Without fulfilling those needs, they cannot progress to the top of pyramid where they attain self-actualization and achieving one’s full potential. With the way that they country is today, we are on high alert with our own safety requirements, mainly due to Covid-19 pandemic.

My questions that I wanted to answer during this project were:

1. What are the most frequent items that get recalled?
2. Most specifically what food items, medication items and other products that are of interest to the public.

These questions are important because, if these frequent item sets are known, data science could provide a way to mitigate the risk of the recalled items staying on the shelves of the markets. Having those products removed as soon as possible or flagging recalled items at the cash register to prevent from being purchased. In my research, I found that there are inventory systems in the meat industry that utilize RDIF packaging. This is used to track the package across the country. With knowing frequent items that are recalled, such as salad or spinach mixes, this inventory system could expedite removal of recalled items off the shelves. It looks as though I am not the only person thinking about a solution to this problem. SICK, an Application and Technology group, believe that there are several different ways RFID can be used to track packages and minimize the scale and financial damage of recalls for companies. Their focus is on the supply chain process and software development for a company, reducing financial damage and the company’s confidence level among consumers. My focus and approach are based on providing safety and where it can be placed for the consumers.

My approach was to find data over the past two years and find the frequency of items recalled. I used open data sets from <https://www.fda.gov/about-fda/open-government-fda-data-sets/recalls-data-sets>. These both were in an Extensible Markup Language (XML) format. I saved Recalls2020 and Recalls2019 into different XML files and uploaded them to Jupyter Notebook. I went back and forth with whether an XML would be a good format to work with, as my experience is limited on scraping data. So, I looked for CSV files with data and did not find any that were in a real time format as the open data sets on the Food and Drug Administrations website. I decided to learn about BeautifulSoup and work with the initial data sets that I chose in the beginning.

The libraries that I used for the coding of this project were:

|  |
| --- |
| **Libraries Imported** |
| bs4 BeautifulSoup |
| xml.etree.ElementTree as et |
| pandas as pd |
| matplotlib.pyplot as plt |
| mlxtrend.frequent\_patterns  apriori  association\_rules |

I read in Recalls2020.XML to the notebook using the format, recall\_data = open('recalls2020.xml', encoding="utf8"). Then passed the stored data inside the BeautifulSoup parser, storing the returned object. After storing the data, I found all tags associated with </recall>, as this was how each recalled item was separated within the XML file. The tags were broken down even more, with .find\_all function to locate the product and its description. Seeing that when broken down it was it was still in a tree format; I used the element tree parser to make the data easier to read. Column were created for a future dataframe, as well as rows. The columns defined to use for the data analysis were Brand, Company, ProductDescription and Reason. Pandas was used to place the parsed XML into a dataframe named df\_xml and an index was set as seen in Figure 1.

**Recalled Items 2020**

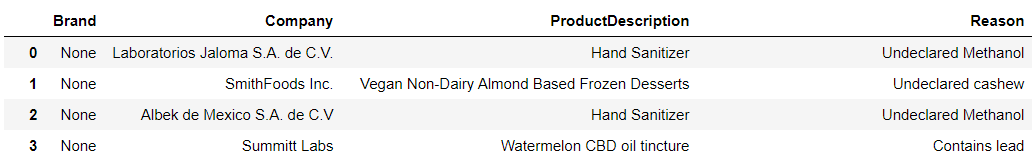


Figure 1: Parsed XML Dataframe Sample

I spent most of my time gathering, shaping, and preparing the recall data for each set. There was a lot of trail and error on parsing but found that element tree was the most helpful for my intent. I conducted value\_counts on the ProductDescription column to find the most frequent products currently recalled this year. The top ten items are seen below in Figure 2:

**Recalled Items 2020**

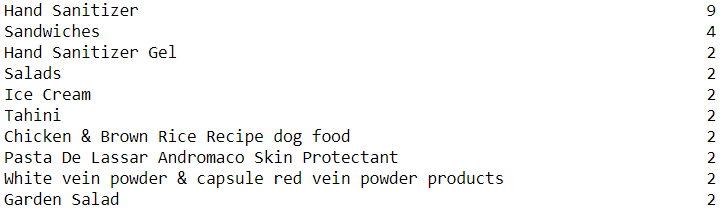


Figure 2: Frequent Counts from Recall 2020 Dataframe Sample

As we can see, safety with social distancing and frequent hand washing is a concern with the active coronavirus outbreak and more liquor companies race to make hand sanitizer [6] to compensate for its once shortage, it is not a surprise that it is the most recalled item to date. But I was surprised and had no idea that there were recalls on such a product. I thought Romaine Lettuce would be number one. From the list we can see that Hand Sanitizer Gel is coming in at number three on the list and Garden Salad falling to number ten. After seeing what frequent items were recalled, I wanted to know the Company’s that are having the most recalls affect them. I conducted a value\_count on the Company column, in Figure 3 it shows that Whole Foods Market is the company with the most recalls this year with eleven items being recalled. Fresh Express that has the current recall on the salad mix is at number three. Does this change my confidence level in shopping at Whole Foods Market? Yes, it does and then I think “Would this company benefit in partnering with an RFID package service and application to provide a more efficient recall process for its consumers?”

**Recalled Items 2020**

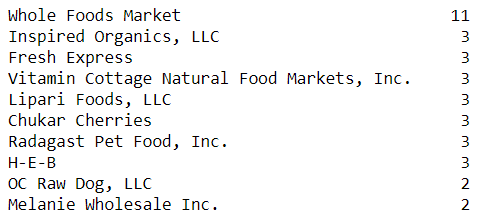


Figure 3: Company’s with the most Recalls Sample

I chose to do an Apriori Algorithm on my dataset. In the beginning of this semester I thought using a CHARM algorithm, but after data preparation and noticing that my data was all categorical and no numeric values to conduct the frequent itemsets on, I leaned toward the Apriori Algorithm. In this algorithm an item column is needed and an ID. I created an ID column by using df.index + 1. Baskets / Bins were created with the ProductionDescription. Hand Sanitizer was one attribute that I used to group by. The results indicated that if Hand Sanitizer was recalled, it was due to the Reason of Undeclared Methanol, contamination and sub potency ethyl alcohol. I then encoded the units to return either 0 or 1 to take place of the categorical data. These were set to be my basket sets. My frequent\_set variable was set with a min support of seven percentage. The rule variable had a lift and min threshold of 1.

Support was indicated to be 0.11, Confidence 0.68, Lift 4.5. The rule variable had a variable that equaled six and an Eighty percent confidence level.

Support shows how popular this item is in the dataset. It is measured by the proportion of how many transactions in which the item appears. In this case our item is Product. Measure of Confidence tells me how likely the product will be recalled and what it is recalled for (‘Reason’). The measure of lift controls how popular a product is.

For the dataset recalls2019.XML I imported the same libraries as I did for recalls2020.XML. Data gathering and analysis on this dataset went faster since I worked out the kinks with recalls2020. I conducted the same value\_count on the ProductDescription and Company columns. In Figure 4 are the top ten items that were recalled in the year 2019.

**Recalled Items 2019**

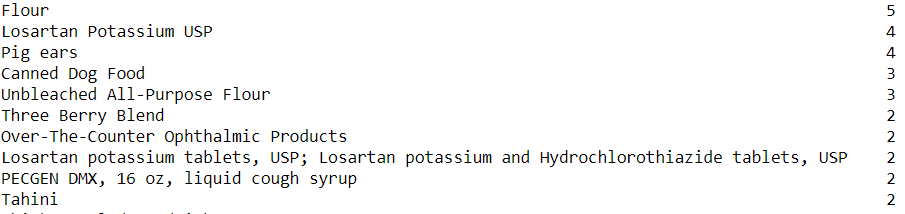


Figure 4: Frequent Counts from Recall 2019 Dataframe Sample

Flour and potassium tablets were among the most recalled products of the year.

**Recalled Items 2019**

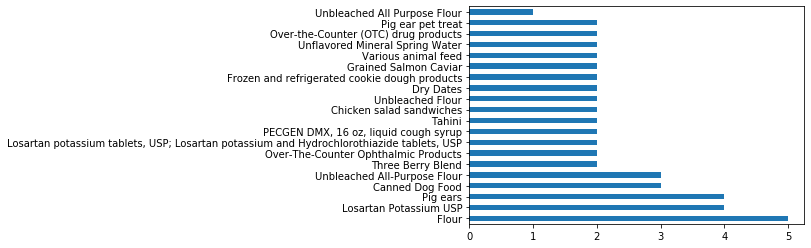


Figure 5: Bar Plot Top 20 Recalled Items 2019

Whole Foods Market was again the Company affected by recalled products. Figure 6 will show the top ten companies with recalls for the year 2019.

**Recalled Items 2019**

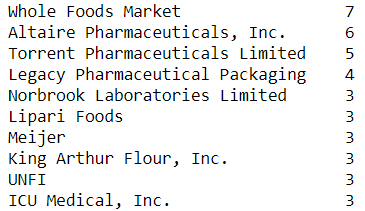


Figure 6: Company Recall 2019 Dataframe Sample

**Company Recalled Items 2019**

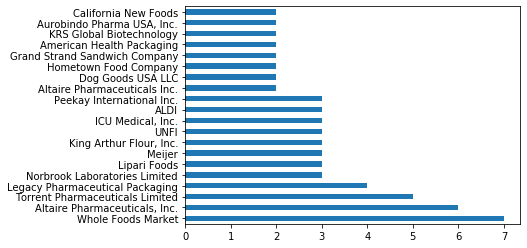


Figure 6: Company Recall 2019 Bar Plot

During the basket creation process in the Apriori Algorithm, the results showed that if Flour was recalled, it was due to E coli. There was one instance that it was salmonella. Again, I encoded the items to become integers of 0 or 1 for the basket sets. In one basket, the data indicated that the supermarket Aldi had two accounts of Flour being recalled for E Coli. The min support was set to seven percent on the frequent items. The association\_rules was applied lift metric with a min threshold of 1. Support was indicated to be 0.107, Confidence 0.72, Lift 4.8. The rule variable had a variable that equaled six and an Eighty percent confidence level.

I was not sure of what my results would be with both data sets. I assumed that spinach would be the most recalled item due to anecdotal evidence of salad bar hearsay. It was very astonishing to see that hand sanitizer was indeed the top item recalled for the year and that Whole Foods Market has been the Company for two years in a row with the most recalled items.

References:

[1] Duffin, R., (2020), Product Recalls in the United States – Statistics and Facts, Statista, Retrieved from: <https://www.statista.com/topics/3798/product-recalls-in-the-united-states/>

[2] Day, A., Schlesinger, J., (2017), How the CPSC keeps consumers safe from products that get recalled, Retrieved from: <https://www.cnbc.com/2017/09/09/how-the-cpsc-keeps-consumers-safe-from-products-that-get-recalled.html>

[3] CDC, (2020), Outbreak of Cyclospora Infections Linked to Bagged Salad Mix, Retrieved from: <https://www.cdc.gov/parasites/cyclosporiasis/outbreaks/2020/index.html>

[4] Wikipedia, (2020), Maslow’s Hierarchy of Needs, Retrieved from: <https://en.wikipedia.org/wiki/Maslow%27s_hierarchy_of_needs>

[5] SickUSAblog, (2016), Benefits of RFID for Track and Trace in Product Recall, Retrieved from: <https://sickusablog.com/benefits-rfid-track-trace-product-recall/>

[6] Wagner, M., (2020), New Habits, New Products: The Story Behind Liquor Companies Making Hand Sanitizer, Retrieved from: <https://www.worth.com/new-habits-new-products-the-story-behind-liquor-companies-making-hand-sanitizer/>

[7] Ng, A., (2016), Association Rules and the Apriori Algorithm: A Tutorial, Retrieved from: <https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html>