Adverse Food Events Analytics

By:
Connor Beauchamp,
Jorge Castano,
Lewis Furlan-Lowry,
Maxwell Tyson

Table Of Contents

Title Page **1**

Table Of Contents **2**

Abstract **3**

Introduction **4**

The Data Set 4

Method **5**

Data Preparation 5

Analytics Tools 5

Orange 5

Weka 5

Rapid Miner 5

Watson 6

Visualization Models 6

Associations 6

Clustering 6

Predictive Algorithms 6

Results **6**

Industries 6

Outcomes 6

Demographics 7

Products 7

Timing 7

Conclusion **7**

Surprising Discoveries 7

What else? 7

References **8**

Abstract

Foodborne illnesses are something that has affected most people at one point in their lives. Being personally affected by adverse food events, our team decided to utilize big data and industry leading analytics tools, to perform an analysis on over 90,000 adverse food events in the United States from the years 2004 to 2017. Through this analysis, we hoped to find associations between age groups, genders, product names, industries, and several more variables contributing to a variety of symptoms and outcomes. With the utilization of a comprehensive data set, and the processing power of a supercomputer running a sophisticated AI developed by IBM called 'Watson', we made several interesting discoveries which can potentially be used to avoid adverse food events in the future. This paper provides a walk through of our entire process; from identifying and obtaining the data set, through to the final analysis of our results. We will outline the several different data processing and analytic tools utilized throughout our process, and why each of these was either successful or unsuccessful in providing us with the services we required. Finally, using the different visualization methods and analysis techniques taught in "Database II: Advanced Database", we will discuss the various results we obtained, why these results are interesting, and how we can possibly utilize this information to make better decisions about the types of products we choose to consume.

Introduction

In the United States, adverse food events and product complaint reports are submitted to the FDA and are stored in the CFSAN Adverse Event Reporting System database. With food being a necessity, we felt it would be interesting to analyze and identify problem products and other factors regarding individuals and the food industry. Through the use of Watson, we were able to identify problematic products and gain insight on past events, such as salmonella outbreaks in peanut butter. This paper will walks through our teams entire process; from how we went about preparing the data, to the results we collected, with explained visualizations.

The Data Set

The Adverse Food Events dataset consists of approximately 214,610 product-related user-reported adverse medical events through 2004 to 2017. This dataset is from the CFSAN Adverse Event Reporting System (CAERS) database which contains information from product complaint reports submitted to the FDA. Below is a table regarding the 12 columns found in the dataset.

Columns and corresponding description found in database	
Column Name	**Description**
RA_REPORT #	Unique report ID (Numeric)
RA_CAERS CREATED DATA	Date the report was created (DateTime)
AEC_Event Start Date	Data the event started (DateTime)
PRI_Product Role	Role the product played in the event (String)

PRI_Reported Brand	Full product name (string)
PRI_FDA Industry Code	Industry product code (numeric)
PRI_FDA Industry Name	Industry Name (string)
CI_Age at Adverse Event	The victim's age at the time of incident (numeric)
CI_Age Unit	The unit for the age value, such as "years" (string)
CI_Gender	Victims gender
AEC_One Row Outcome	Outcomes from the event, Hand-coded (string)
SYM_One Row Coded	Outcomes from the event, system coded (string)

The columns which took precedence while handling and analyzing the data included the product role, reported brand, industry name, adverse effect, age unit, gender, aec_one row outcome. The event start date column was used only in a couple of instances, and the unique report ID was not used in the analysis explicitly. These rows were chosen due to the significance of the potential in the correlations. Initially this was not evident, but as the analysis progressed, the outcome and symptoms columns became widely used while mostly being supported by age and industry name.

Upon further analysis of the data, a large amount of duplicate rows were found. This took the data set from 214,610 rows to 90,000. Once the duplicates were removed there were 33,545 reported combination of symptoms, 45,651 unique products, 283 unique outcomes, and over 41 different industries. In addition to the duplicate rows, the column names needed to be changed since they got to be a bit confusing while handling the data.

Method

Data Preparation

Having had a significantly difficult experience with the first dataset chosen, we decided to choose a cleaner dataset to work with which we still found interesting. The adverse food dataset was downloaded directly from Kaggel and was relatively clean, but had column names which were deemed potentially confusing, so we decided to rename them directly in excel. The values in each column also contained little errors, with only a small amount of blank or null values. However, there were a very large amount of duplicate rows, which upon clean-up, reduced the dataset by more than half its original size. We initially thought that there may have been duplicate events reported by different people, which meant that deleting these rows would be skewing the results, but after careful analysis of the data, we concluded these anomalies to be just duplicates. During the clean-up process we noticed that many of the products in the dataset were labeled as ‘redacted’, and numerous reports were found to be not available, not reported, or had an unknown entry for gender. We decided to include these values due to the attributes held in the other columns, and also thought it would be interesting to see correlations between events with redacted attributes versus explicitly stated values.

Analytics Tools

Orange

Orange is a machine learning and data mining software used for data analysis through Python scripting and visual programming. This software contains a large toolbox of data mining

components such as: data management and preprocessing, classification, regression, association, clustering, and many more. Orange can make data mining simple with its GUI based tools that aid beginners who want to visualize patterns, and who don't have a technical coding background. At first, Orange seemed promising to us as it was simple to grasp, and results were outputted with ease. The visualization of data, however, was not as appealing as some of the other tools, and more intensive procedures would take multiple hours or crash the program. (Janez, 2013)

RapidMiner

RapidMiner, Inc develops an open source data science platform called RapidMiner Studio. RapidMiner Studio offers a visual programming environment for predictive analytics workflows that allows its users to utilize its graphical environment for big data analytics. RapidMiner Studio also provides extensions, such as text mining, web crawling, or integration with other analytical tools such as Weka ("Company Overview of RapidMiner, Inc.", n.d.). Being similar to Orange, RapidMiner Studio was simple to use, but outputted unattractive visualizations. RapidMiner also suffered from critical performance issues. Even with an overclocked i7 processor, and 16gb of ram, RapidMiner's extreme processing requirements would max out the system resources and crash the program with more complex visualizations. We were, however, able to utilize some of the basic models offered, but with the limited amount of processing power available to us locally, we were left with no choice but to find other alternatives for generating these models.

Watson

Developed by IBM, Watson Analytics is a smart data analysis and visualization service that allows users to quickly discover patterns in any dataset ("IBM Watson Analytics", n.d.).

Watson's AI is implemented on a cloud-based supercomputer, which meant we would not longer had to deal with the issue of limited local computing power. With features such as guided data discovery, automated predictive analytics, and cognitive capabilities, we felt that Watson was one of the most promising analytics software for us to use. Our first experience with Watson went poorly, as there was vague errors given, preventing any data set from being uploaded; even Watson's included data sets would fail to upload. After spending some time on IBM Watson forums, we found out that it was an issue on their end, and would be resolved very soon. The next morning our data set uploaded into Watson with 0 errors or warnings. We were quick to see Watson's easy-to-use nature, and its clean and meaningful data visualizations. Watson had a user interface not like any other of the tools we attempted to use, and this made the analysis relatively easy to conduct. We found some of the associations that Watson would generate on its own (based on what it thought our data set was trying to evaluate), where actually quite helpful, even if they required a bit of tweaking. The versatility of Watson allowed us to explore our data set on a level we were not expecting. With the ability to analyze the same variables through different visualization models, our perception of this type of analysis changed dramatically. Our struggles with the other tools we tried using for this analysis only accentuated the benefits of technologies like Watson. With its sophistication and intuitive interface, we could focus less on how to use the tool, and were able to focus more on exploring the the vast amount of information within the data-set. The Results section of this paper will dive deep into our findings, and discuss the meaningful information we were able to collect using Watson.

Visualization Models

Associations

Associations were a terrific starting point, for diving down into our data set. They gave insight into the correlations data which allowed us to have a clear starting point once our own insights ran short. The variety of symptoms and outcomes, however, correlated with the product_role, made it hard to visualize the data in a meaningful manner. Therefore, further filtering was required.

Clustering

We chose to use clustering to show the different relations amongst the data points which may help us to understand their correlation visually. We had good results while using network maps, which allowed us to breakdown columns where on the other hand, the domain would be difficult to classify.

Predictive Algorithms

We tried to use the predictive algorithm approach to get some insightful data which may give us some insight into the future. Although the many different mining tools offered different ways to approach this, we were not able to extract significant results from the data set using this approach. The results that were produced were either too broad or did not finish computing successfully, yielding an error as a result.

Results

Concomitant vs Suspect

A main variable that we needed to consider in each of our results was whether or not the adverse food event was caused by a ‘concomitant’ or a ‘suspect’ product. If the event was caused by a concomitant, this means that the adverse food event is naturally occurring. The graph below, We found that most of our data set fell on the side of the suspect, with just over 90% of entries being the suspect of the adverse food event.

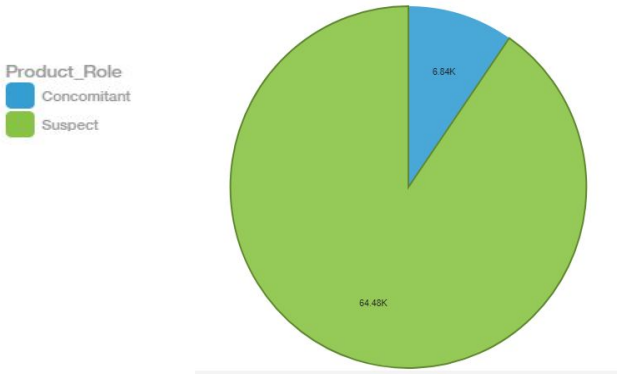

Figure 1: Product role

Industries

We began our analysis by exploring all of the different associations pertaining to the specific industries causing adverse food events. We felt this information was important to analyze because if a specific industry is a leading cause of adverse food events, we can begin to make more informed decisions about the types of products we choose to consume from that industry.

In each type of visualization we used, Vitamin/Mineral/Protein/Unconventional-diets would completely dominate all of the other industries for causes of adverse food events. Because

of this outlier, we decided to exclude this industry from our results so we could obtain a better view of the other industries causing these events. The word map below visualizes the industries causing the most adverse food events. This format makes this information extremely easy to understand and interpret, which prompted us with many questions as to why these industries were causing these events. What was most surprising to us, was that cosmetics was the number 2 leading industry causing of adverse food events.

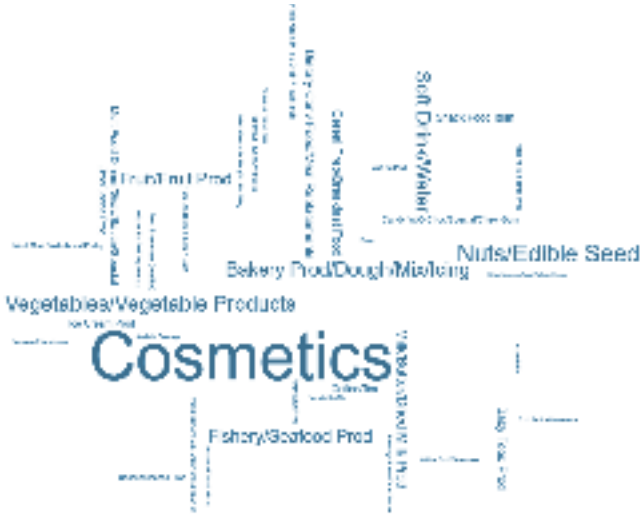

Figure 2: Industry occurrences represented in a word cloud (does not include Vitamin/Mineral/Protein/Unconventional-diets)

Another aspect pertaining to the industries included in this data started to raise our interests, and that was age. we wanted to correlate the average age affected by adverse food events as a result of products from specific industries. The following graph displays the average age by number of entries that affected by each industry. What was most interesting about this graph was that individuals over the age of 20 were more likely to be affected by each industry and not seniors or children.

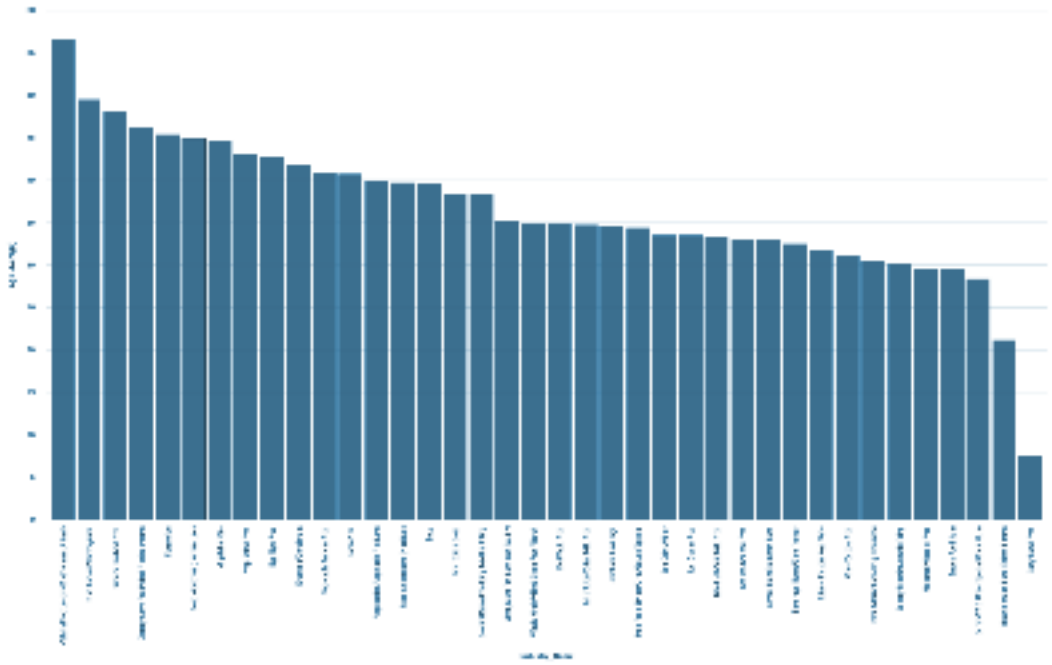

Figure 3: Age and Product correlation

Outcomes

The outcomes column in the dataset represents the final result of the adverse food event. This includes death, a visit to the ER, non serious illness etc. Our initial interest in this column was simply to see what the most common outcomes were. The graph below captures the top 7 most common outcomes resulting from adverse food events. Interestingly, and also fortunately, death did not appear in this result set. The most common outcome was non serious injuries, which could most likely be related to indigestion following food consumption.

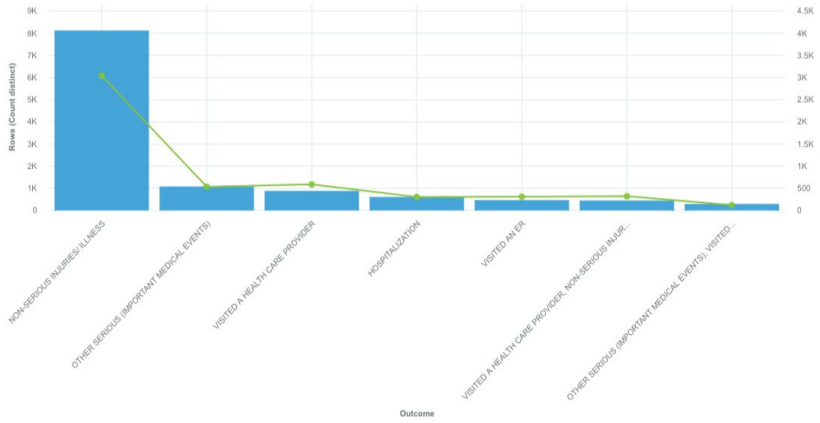

Figure 4: Top 7 most common outcomes resulting from adverse food events

We started getting curious about other outcomes within our data set, and what types of outcomes each industry was causing. With all the different types of outcomes available, we wanted to focus on some of the more extreme or rare ones resulting from these adverse food events by industry. The bubble chart shown on the following page displays the industries with the highest number of outcomes that included 'death'. As you can see, Cosmetics was one of the top 2 contributors. Prompting the question of why cosmetics were considered a food related product, and the reason why they would be one of the top contributors. This will be answered later.

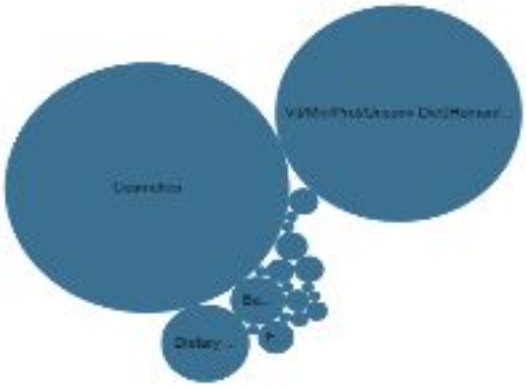

Figure 5: industries with the highest number of outcomes which included 'death'

Demographics

Gender vs RA_Report

The first demographic factor we analyzed was the gender of the affected individuals. We decided a pie chart, which would reflect this information best, and counted each gender from every row. We found that females report adverse food affects more than males.

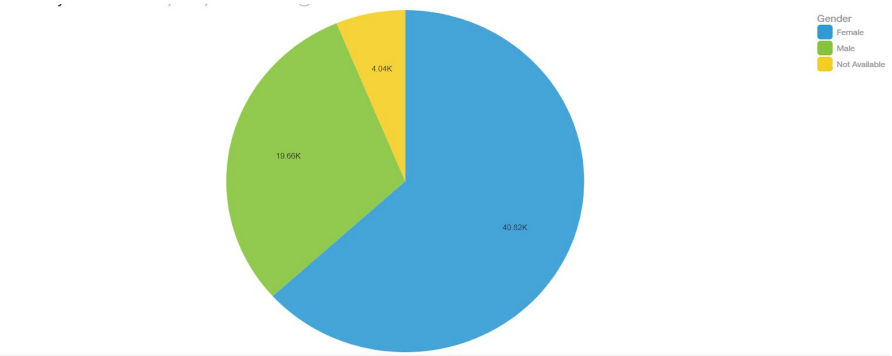

Figure 6: Gender vs Number of Entries

We found that as individuals aged, they were more susceptible to adverse food events. This is most likely a result of the digestive health disorders, medication side effects, or other age related issues that individuals experience as they age. Individuals of the age 72 reported adverse food effects the most. There was a sharp decrease of reports after age 72, as seen below, which is we thought could be due to individuals passing away from old age or other illnesses.

Figure 7: Average Age vs Number of Entries

Products

Products vs Number of Symptoms

There were many symptoms coinciding with the adverse effects of products. We decided to look into which products lead to the highest number of symptoms. Some of the symptoms found in our dataset include: rash, chest pain, choking, diarrhea, myalgia and nausea. We found that raw oysters produced the highest number of symptoms at 14 (which is a concomitant product, known for containing a norovirus), followed by Axona (6), and Centrum Silver Women (5). Below is a histogram that displays the top 11 products that cause the highest number of symptoms.

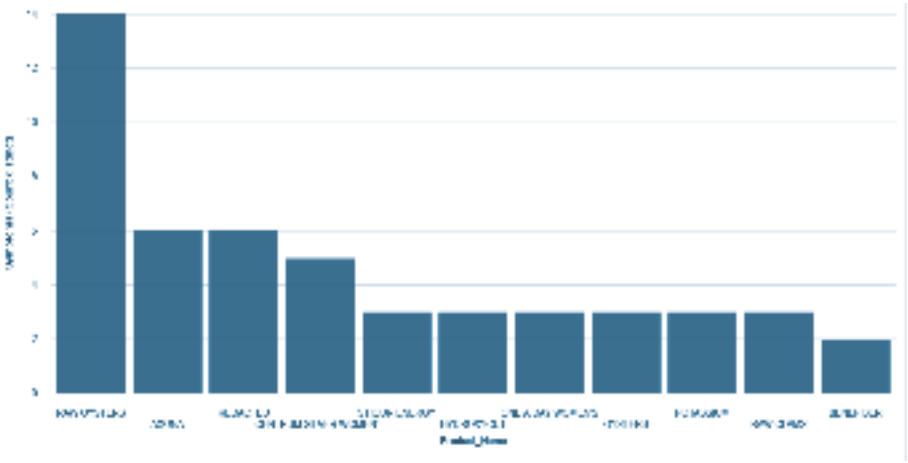

Figure 8: Products that caused the highest number of deaths

To determine the products that caused the highest number of deaths, we had to count each row that had the string “death” under the outcomes column. While there were not many deaths in the dataset, Diet Coke, Austin Peanut Butter Crackers, and even turmeric, were the products that caused the highest number of deaths. After some investigating, many of these deaths were a

result from salmonella outbreaks. Below is a heat map showing the products related to the highest number of deaths.

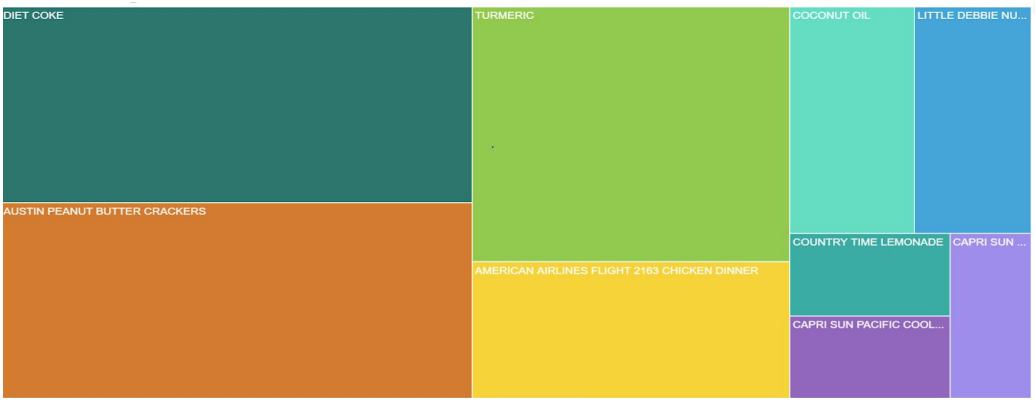

Figure 9: Products related to the highest number of deaths

Timing

We looked into the timing of the reports to determine if month had anything to do with the adverse food effects. Our findings show that the month (or even season) of the event occurrence, was not a large contributing factor, as shown in the bubble chart below.

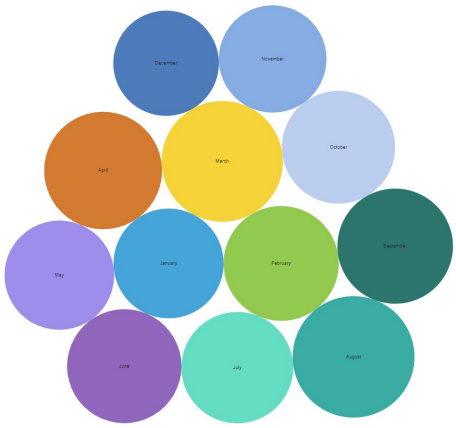

Figure 10: Average Age vs Year of Adverse Food Event

We found that the average age of individuals reporting adverse food effects were roughly between 40 and 50. The graph below visualizes the average age throughout the years. A

somewhat interesting finding, is there was a sharp decrease in the average age between 2007 and 2008, and then a sharp increase back to around the mean. Below is a graph that displays this information.

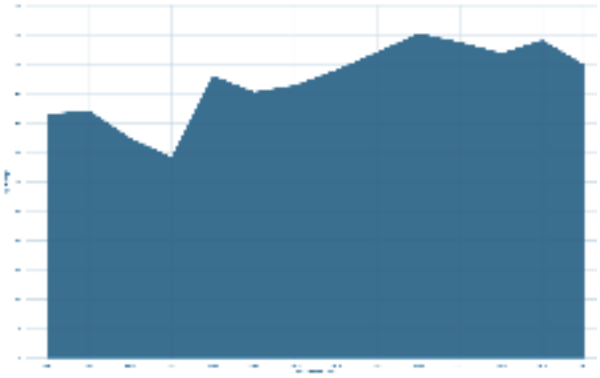

Figure 11: Average age throughout years where information was collected

Conclusion

Surprising Discoveries

The most interesting outcome from this specific analysis was that Cosmetics was the number 2 leading cause for adverse food events. We explored this further with research, and found out that it wasn't the consumption of these products, but simply adverse reactions to their ingredients. Adverse side effects or events relating to cosmetic ingredients are reported to the same CFSAN database as food related incidents says Alexandra Sifferlin from Time Health (2017). She interviews Dr. Steve Xu who uncovers the fact that cosmetics companies are able to bypass regulations implemented in other industries and are not required to disclose any adverse side effects from ingredients used in their products to the FDA. Dr. Steve Xu found through his research of a version of the CFSAN database, that the 3 main products causing adverse events in the cosmetic industry are hair care, skin care and tattoos (Sifferlin, 2017).

Another surprising discovery was the constant emergence of peanut butter in our results. More specifically, Peter Pan's and Great Value's peanut butter were reported hundreds of times in 2007, resulting in their products showing up in many of our visualizations. After some investigation, we found that this was the result of Salmonella Tennessee contamination. Both brands are manufactured by ConAgra Foods and this bacterium that causes foodborne illness was believed to be present in only jars with the product code 2111 ("Food Poisoning in the News: Peter Pan / Great Value Peanut Butter", nd).

Lastly, we wanted to figure out what types of bacteria where causing these events and to discovered that "91 percent of adverse food events are associated to the following: Norovirus (commonly found in oysters, fruits, and vegetables); Salmonella (commonly found in eggs, meat, and dairy products); Clostridium perfringens (found in meat and poultry); Campylobacter (found in undercooked meat and contaminated water); Staphylococcus (found in animal products such as cream, eggs, and milk)"(Krans, B. 2017). We also discovered that the people most at risk for these events are, infants and children, pregnant women, older adults, people with chronic conditions. (Krans, B. 2017) This is evident in our demographics analysis of our data set where we look at the ages at which these adverse food events occur the most. We are unable to speak to the statistics of Pregnant Women being more susceptible to food related illnesses, even if our results showed women are affected more than men, due to lack of information pertaining to current medical conditions.

What else?

When looking at what else could have been done with this dataset, the most limiting factor is the time span during which the data was collected. While 13 years is enough to perform

mining and generate meaningful results for the purposes of this project, it would be useful to track this data across a longer time span. We feel that it would be interesting to track symptoms across different generations to draw some conclusions about how modern food products are affecting different generations. Despite this constraint we were able to gain a some in this regard by tracking ages of those reporting these adverse events. Another compelling area for analysis would be to track and record follow up data with those reporting the incidents. It would be interesting to see if people were less likely to have recurring instances of illness after first making the report. This type of knowledge is critical for doctors as it allows them to create specialized treatment plans tailored to preventing illnesses that are likely to reoccur.

A way to complement the information from this dataset would be to access medical history records, lawsuits, or perhaps data from other countries depending on the scope desired. This dataset would also benefit from columns which perhaps pointed out specific SKU codes for the products helping to identify specific brands. Depending on the granularity batch trackers may help to identify specific dates in which a product has been found to be contaminated such as the peanut butter.

References

Company Overview of RapidMiner, Inc.. Bloomberg. Retrieved April, 2 2018, from https://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapId=266292394

Food Poisoning in the News: Peter Pan / Great Value Peanut Butter. Findlaw. Retrieved, 8, April, 2018 from http://injury.findlaw.com/product-liability/food-poisoning-in-the-news-peter-pan-great-value-peanut-butter.html

IBM Watson Analytics. *IBM*. Retrieved April, 2, 2018, from https://www.ibm.com/us-en/marketplace/watson-analytics

Janez, D. (2013, August). Orange: Data Mining Toolbox in Python. Retrieved April 9, 2018, from http://eprints.fri.uni-lj.si/2267/1/2013-Demsar-Orange-JMLR.pdf

Krans, B., Murrell, D., (2017). What to Eat After Food Poisoning. Retrieved March, 26, 2018, from:https://www.healthline.com/health/food-nutrition/What-to-eaT-after-food-poisoning

Sifferlin, A. (2017). The Hidden Dangers Of Makeup And Shampoo. Retrieved March, 29, 2018 from: http://time.com/4832688/makeup-shampoo-toxic/