**Which Food Allergy has the highest Death Rate?**

Blog Post: <https://medium.com/devindatt/which-food-has-the-highest-death-rate-bd543b67729c>

Can you guess what it is? I will give you a little spoiler, its not what I thought it would be, which was the dreaded peanut!

If you are anything like my niece you have experienced the anxiousness going to a gathering and looking at some piece of food and wondering if there is something in here that she might have an adverse reaction to. You see my niece is got one of those ‘I am allergic to everything under the sun’ symptoms.

I had always wanted to do an analysis on food data, so when we were asked as part of the Udacity Data Science nanodegree program to do a analysis on a dataset of our choosing, I immediately looked on Kaggle for a dataset to use. I found one that was data extracted from the US Food and Drug Administration (FDA) Center for Food Safety and Applied Nutrition (CFSAN) Adverse Event Reporting System (CAERS). It consists of a little over 90,000 reports of allergies, their symptoms, and eventual outcomes.

To tackle this project I used the **CRISP-DM** process that is the [most popular and used](https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html) by Data Analysis worth their salt. If you are not familiar with it, acronym stands for ‘**CR**oss-**I**ndustry **P**rocess for **D**ata **M**ining’ and the process steps are laid out below:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Data Modeling
5. Evaluation
6. Deployment

**Business Understanding:**  This first step is the most important, which is to think of some pertinent questions important for decision-making. Looking at the features of the dataset (ie. columns) I thought the following questions I would like answered about food allergies if I was a business owner, a health advocate, or even a parent. A couple of questions I immediately thought of were the following:

**Question 1:**What was the most frequently reported food for allergies and what was the resulting outcome?

**Question 2:**What was the most reported food allergies symptom?

**Question 3:**What foods result in death for ‘elderly’ vs. the same reactions for ‘young people’?

**Data Understanding:** This next step is how to translate your questions into data analysis. There are two ways to go about this, look at your questions and then gather the data, or use the data you have. We followed the latter as the dataset had the columns we need.

**Preparing the Data:** This is the step that takes 80-90% of the process. These steps are outside the scope of this blog but can be seen in my GitHub repository link at the end of this blog if interested.

**Modeling the Data:** This step is to create a model of some sort (ie. regression) and use the data to make predictions. This step can be skipped if you can derive your analysis without modeling which is what I decided to do in this project.

**Evaluation:** This step is where I evaluate the results of my findings from the wrangled data and provide some insights. I will do this at the end for each of the questions asked above.

**Deployment:** This is where we move the approach into production or by using the results to persuade others within a company to act on the results. Here I used the approach of documenting these findings in this blog.

We will start with the first question and walk thru this process.

**Question 1:**What was the most frequently reported food for allergies and what was the resulting outcome?

For the first question we initially used the dataset columns to ‘understand the data’. We would use the ‘Outcome’ and ‘Brand-Product’ and ‘Industry Name’ for this question. Finding the most reported food might not be that difficult to find, as the entries are single in nature. However, this is not so for the 'symptoms, as they are one column that can hold multiple results. Linking the top result for food allergy to the top outcome will require some wrangling of the data and splitting out the Outcomes columns.

This is where we ‘prepare the data’. To get the most reported foods outcome, we will need to work backwards as the Outcomes needs to be separated. In order to understand what was the most frequently medical outcome from a reaction to a food allergy, we will first look at the "Outcomes" individually. Once we get the Outcomes split into either a list or their own individual columns, we can the just simply sort the new dataset but top food allergy and extract its associated medical outcome.

Just to get an idea of how congested the feature was before here is a sample of a portion of the Outcomes column that consists of nearly 300 uniquely entries shown below :

SERIOUS INJURIES/ ILLNESS, HOSPITALIZATION VISITED A HEALTH CARE PROVIDER, HOSPITALIZATION VISITED AN ER, VISITED A HEALTH CARE PROVIDER, OTHER SERIOUS (IMPORTANT MEDICAL EVENTS) NON-SERIOUS INJURIES/ ILLNESS, VISITED AN ER VISITED A HEALTH CARE PROVIDER, VISITED AN ER VISITED AN ER, VISITED A HEALTH CARE PROVIDER, HOSPITALIZATION SERIOUS INJURIES/ ILLNESS VISITED AN ER, OTHER SERIOUS (IMPORTANT MEDICAL EVENTS), HOSPITALIZATION

From the 90,000 entries in the Outcome row we found they are a mix of multiple smaller entries (ie. Visited an ER..etc.). When we wrangled this data we found these 298 outcomes was comprised of only 11 core outcome results. These are listed below:

['OTHER SERIOUS (IMPORTANT MEDICAL EVENTS)',

'VISITED AN ER',

'NON-SERIOUS INJURIES/ ILLNESS',

'HOSPITALIZATION',

'VISITED A HEALTH CARE PROVIDER',

'DISABILITY',

'SERIOUS INJURIES/ ILLNESS',

'DEATH',

'LIFE THREATENING',

'REQ. INTERVENTION TO PRVNT PERM. IMPRMNT.',

'CONGENITAL ANOMALY]

Now we will go back thru the original 'outcome column and count up all the times each of the 11 possible values are mentioned within each of the nearly 300 entries and put them in each of their own columns (ie. extending the column count by eleven columns).

This will turn out to be a very useful exercise to gather some insight on what is really happening here as it give a clearer picture as these numbers would be a representation of actual mentions of an outcome in any row instead of association one row with one outcome.

Once we had the dataset columns for report food and all unique medical outcomes in their own columns we can do a simple query to find top food reported and see the outcome it is linked to.

The top ten results show that

0: Vit/Min/Prot/Unconv Diet(Human/Animal) : 48501

1: Cosmetics : 11733

2: Nuts/Edible Seed : 3383

3: Vegetables/Vegetable Products : 3115

4: Soft Drink/Water : 2591

5: Bakery Prod/Dough/Mix/Icing : 2543

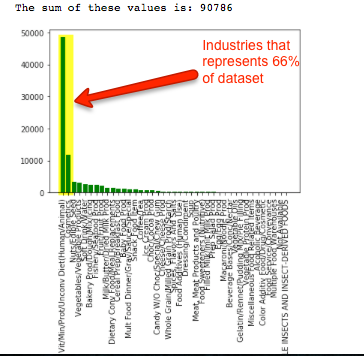
6: Fishery/Seafood Prod : 2382

7: Fruit/Fruit Prod : 2239

8: Milk/Butter/Dried Milk Prod : 1515

9: Dietary Conv Food/Meal Replacements : 1420

10: Cereal Prep/Breakfast Food : 1264



As we mentioned before and see from above, it is clear that the top 2 categories 'Vit/Min/Prot/Unconv Diet(Human/Animal)' and ‘Cosmetics’ categories by far the dominate reported industries (representing 66% of the whole dataset) so we should concentrate our analysis here even though there some interesting questions we could still ask about the other 40 categories as it pertains to the eleven outcomes but for the limit of this project (blog) we save that for another analysis.

We now have to somehow group all the reported outcomes for each of the Industry products that were reported with a food allergy. We are really only interested in the food that had the highest reports, which in this case is definitely surrounding the vitamins/minerals category (ie. 'Vit/Min/Prot/Unconv Diet(Human/Animal)'). We will now want to see how that was distributed over all the 11 outcomes and see what was the highest, which is shown below:

0: OTHER SERIOUS (IMPORTANT MEDICAL EVENTS) : 28350.0

1: HOSPITALIZATION : 13951.0

2: VISITED A HEALTH CARE PROVIDER : 10061.0

3: VISITED AN ER : 8218.0

4: NON-SERIOUS INJURIES/ ILLNESS : 3736.0

5: LIFE THREATENING : 3686.0

6: SERIOUS INJURIES/ ILLNESS : 3354.0

7: DISABILITY : 1814.0

8: REQ. INTERVENTION TO PRVNT PERM. IMPRMNT. : 1687.0

9: DEATH : 755.0

10: CONGENITAL ANOMALY : 53.0

0: OTHER SERIOUS (IMPORTANT MEDICAL EVENTS) : 6371.0

1: VISITED A HEALTH CARE PROVIDER : 5082.0

2: NON-SERIOUS INJURIES/ ILLNESS : 2923.0

3: DEATH : 969.0

4: VISITED AN ER : 577.0

5: HOSPITALIZATION : 564.0

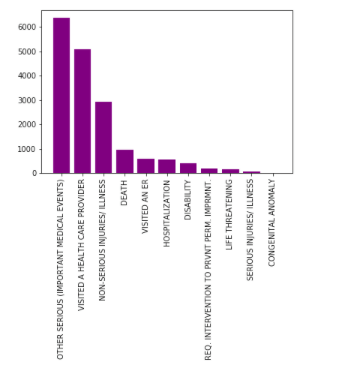
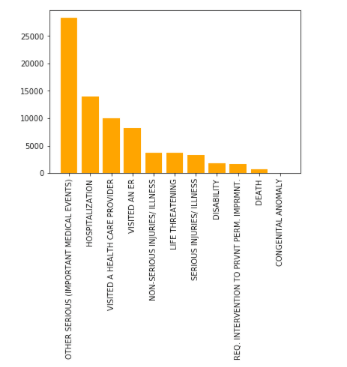
6: DISABILITY : 397.0

7: REQ. INTERVENTION TO PRVNT PERM. IMPRMNT. : 183.0

8: LIFE THREATENING : 158.0

9: SERIOUS INJURIES/ ILLNESS : 50.0

10: CONGENITAL ANOMALY : 3.0



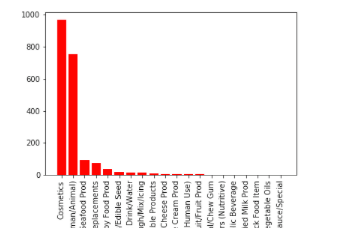
So what we see above is we have drilled down from looking at over 48,000 rows that pertain to the 'Vitamin, Minerals...etc. category and within those rows we have divided how that category reflects in our 11 reported outcomes. We see the highest reported Outcome by far is 'Other Serious (Important Medical Events)' with a whopping 28,350 reported mentions.

We also did the same for the 'Cosmetics' and noticed there were 6,371 times it was linked to the 'Other Serious (Important Medical Events)' outcome which is only a 22% of the mentions of the 'Vitamins/Supplement' category. Something interesting to note here, when analyzing 'vitamins' the outcome of 'death' was reported only ~1% of the time (and lands in 10th or 11th spot on the ranking above). However, when we switch to 'Cosmetics' we see that death rate rise to ~6% (and lands 4th of 11th ranking above). This is quite an interesting find as we normally hear so much on food allergies and vital or lethal affects but here was see vitamins (which can still be considered a 'food') and products that are not the normal items we would normally associate with 'food allergies.

And just for curiosity, I decided to see how these two categories would be reflected if we sorted the 66% data above by only outcomes that resulted in ‘death’. Below you can see the two categories combined account for 85% of all deaths reported as an outcomes:

0: Cosmetics : 969.0

1: Vit/Min/Prot/Unconv Diet(Human/Animal) : 755.0



Now I drilled down even further on the Industry to find the top products that resulted in death in both in the two Outcome categories (Other Serious & Cosmetics) and show those products below:

**OUTCOME: OTHER SERIOUS (IMPORTANT MEDICAL EVENTS)**

**Vitamins/Minerals:**

CENTRUM SILVER WOMEN'S 50+ 574.0

**Cosmetics:**

REDACTED 4214.0

WEN CLEANSING CONDITIONER 192.0

We can see **"Centrum Silver Women 50+"** was the top vitamin/supplement reported that resulted in 'other serious or important medical events' while the "**Benefiber with Wheat Powder**" was the top product in that same category that resulted/reported as an outcome as 'death'.

**OUTCOME: DEATH**

**Vitamins/Minerals:**

BENEFIBER POWDER (WHEAT DEXTRIN) 11.0

**Cosmetics:**

REDACTED 939.0

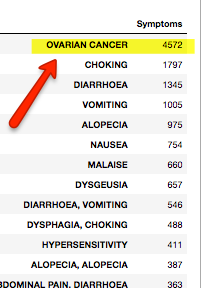
SHOWER TO SHOWER POWDER 6.0

Moving to the cosmetic category, we see "**Wen Cleansing Conditioner**" was the top cosmetic product reported that resulted in 'other serious or important medical events' while the "**Shower to Shower Powder**" was the top product in that same category that resulted/reported as an outcome as 'death'.

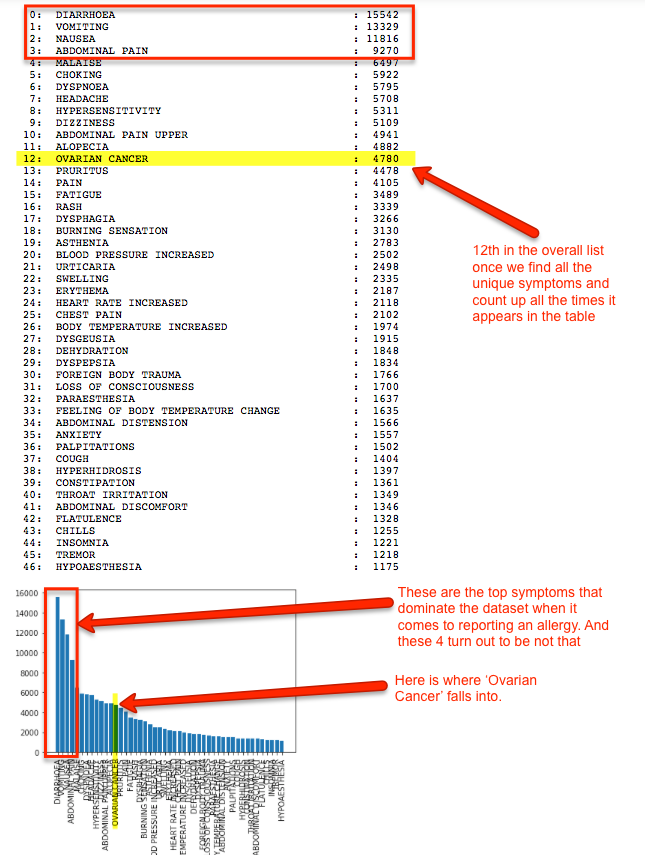
A note to make here about the products that topped the lists above. Since some of the top reported food seems it come up as 'REDACTED' I cannot find anything in the accompanying README documentation or company website or Internet in general to tell me if this is a product or is an indication of a designation to take literally... ie. as the product is too sensitive and has been 'redacted'. As as result, I will assume it means the latter, and proceed to look at the 2nd most reported food product. The product that is labelled 'REDACTED' was omitted from this analysis but if were to include those products they would far out weigh the ones mentioned above as we can see their number is nearly 22x larger then the 2nd place product.

**Question 2:**What was the most reported food allergies symptom?

To answer this question, I looked at the items in the Symptoms to see if we can do the same as we did we the Outcomes column items, but just filtering would result in the list below:



Looking just at the top symptom will illustrate how faulty this would be without pulling all the symptom entries apart and do the true count of each occurrence. However, after splitting out all the individual entries in that column we didn’t find 11 core entries like we did for Outcomes above we actually found 3857 individual symptom entries!! That was way too long of a list to display here so we only included the top 40. Doing this exercise shows that the once topped symptom of ‘Ovarian Cancer’ drops to 12th spot once properly separating out the reported entries.



What we found is that ‘Diarrhoea’ topped the list with 15542 mentions in the symptoms columns. In fact, the top four (4) symptoms mentioned are not considered life threating in the western world (ie. Diarrhoea, Vomiting, Nausea, and Abdominal Pain). An interesting stat I notice during this part of the analysis was that ‘Death’ was reported 458 times as a ‘symptom’ from a food allergy. This might suggest the reaction was violent and sudden and obviously this symptom contributes to the associated ‘death’ entries.

**Question 3:***What foods result in death for ‘elderly’ vs. the same reactions for ‘young' people?*

To answer this question I understood the data by looking at the 'Age Unit' and ‘Outcome’ fields to see what their unique values where they intersect. After reviewing and cleaning the data we found the top 10 products that resulted in ‘death’ for young individuals under 10 years old.

0: Baby Food Prod : 26.0

1: Dietary Conv Food/Meal Replacements : 14.0

2: Vit/Min/Prot/Unconv Diet(Human/Animal) : 12.0

3: Milk/Butter/Dried Milk Prod : 2.0

4: Nuts/Edible Seed : 2.0

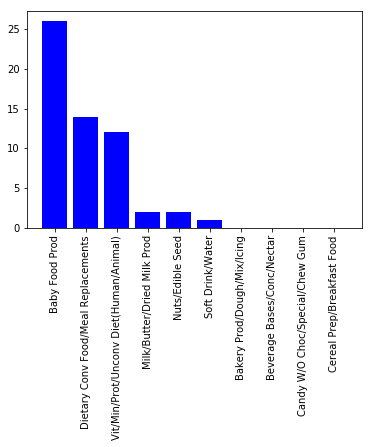
5: Soft Drink/Water : 1.0

6: Bakery Prod/Dough/Mix/Icing : 0.0

7: Beverage Bases/Conc/Nectar : 0.0

8: Candy W/O Choc/Special/Chew Gum : 0.0

9: Cereal Prep/Breakfast Food : 0.0



Then we looked at the same analysis filtering but for individuals over 50 years old.

0: Vit/Min/Prot/Unconv Diet(Human/Animal) : 329.0

1: Cosmetics : 198.0

2: Fishery/Seafood Prod : 66.0

3: Dietary Conv Food/Meal Replacements : 28.0

4: Bakery Prod/Dough/Mix/Icing : 9.0

5: Nuts/Edible Seed : 9.0

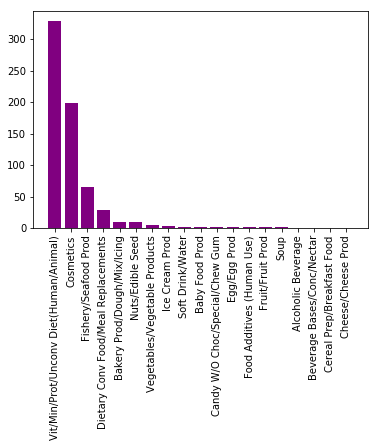
6: Vegetables/Vegetable Products : 5.0

7: Ice Cream Prod : 3.0

8: Soft Drink/Water : 2.0

9: Baby Food Prod : 1.0

10: Candy W/O Choc/Special/Chew Gum : 1.0



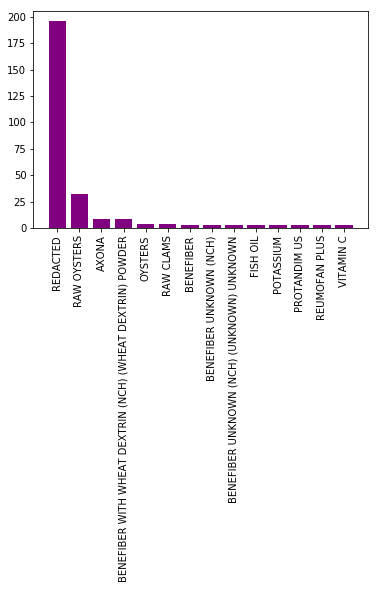
Looking at each age group I filtered by Brand Products that resulted in a symptom of ‘death’ and found the following top products:

Filter list on ages that would be over 50 years (which includes age units of years but excludes units of

# months, weeks or days and decades) for Brand Names that resulted in a 'Death' outcome.

0: REDACTED : 196.0

1: RAW OYSTERS : 32.0



We see the top industry and top brand product that resulted in deaths for youth under ten (10) years and younger was 'Baby Food Prod' and 'SIMPLY THICK' respectively. And we We see the top industry and top brand product that resulted in deaths for older people over fifty (50) years was 'Vit/Min/Prot/Unconv Diet(Human/Animal)' and 'RAW OYSTERS' respectively. Note, the product in ''old' group is 2nd most reactive product after 'REDACTED' which we skipped as that product doesn't add any useful context to the product.

Conclusion, we can see the industries and brand products that are related to deaths in young pertain to products that would be ones that would mostly likely associated with one that mothers would be giving to infants. And the contrast there is a totally separate set of products that relate to death in 'older' people as these products would rarely be used for the same age group.

Note, here is that the conclusion is slightly skewed in the numbers towards the elderly (ie. looks like more older people are more susceptible to food allergies. But this is not case and is more likely the result on been able to replace and clean the age units correctly. Having a more homogeneous age units in years would result in a more accurate analysis, but I feel this rough analysis was sufficient to provide a general insight to the products to avoid in both the young and the elderly.