Title: Analysis on adverse events related to foods, dietary supplements, and cosmetics from year 2004 – 2016 based on data provided by FDA

Summary

The adverse events related to foods, dietary supplements, and cosmetics are gaining more and more attention. This analysis is conducted based on the data extracted from the CFSAN Adverse Event Reporting System from year 2004 to 2016.

Upon both exploratory analysis and application of machine learning methods, it can be found that for people of all age groups (excluding people less than 1 year old), unconventional food and cosmetics are common top sources of adverse events for both female and male. It is also surprising that for male children and adults, cosmetics is the fourth top source of adverse events.

Furthermore, over the past 12 years, frequency of adverse events related to unconventional food and cosmetics on female increased rapidly. Even for male, there is a slight increasing trend of cosmetics related events.

As to the extreme outcomes of adverse events, analysis was done to inspect the occurrence of death or life threatening with gender, age and industrial categories. It can be concluded that industrial category is a not a significant factor causing death or life threatening. However, when zooming in the age group of 20-40 or 40-60, industrial category is a significant factor resulting in an extreme outcome.

The structure of the draft is as follows:

part 1: Introduction

part2: Data processing;

part 3: Preliminary analysis and Data visualization;

part 4: Modeling and Analysis (including a brief discussion about encoding method)

part 5: Future work;

part 6: Appendix

Part 1. Introduction

Nowadays, food safety problem is drawing more and more attention. It is worthwhile to conduct a deep analysis on the adverse events related to foods, dietary supplements and cosmetics. By conducting such analysis, one can find out which categories of products causing most adverse events, and which people are more vulnerable to a specific kind of food or products. Accordingly, regulations on certain industries can be properly developed and more ad-hoc advice can be given to customers.

The data extracted from the CFSAN Adverse Event Reporting System is a reliable source of data to conduct a research on adverse events related to foods, dietary supplements and cosmetics. The data is provided by FDA and can be downloaded from <https://www.fda.gov/Food/ComplianceEnforcement/ucm494015.htm>.

Part 2. Data pre-processing

A careful data preprocessing was conducted to the dataset downloaded from the website, in Python. The variable names and how each of them was cleaned and/or normalized are shown below:

1. RA\_Report #: converted from numeric to string

2. RA\_CAERS Created Date: defined as date

3. AEC\_Event Start Date: defined as date

4. PRI\_Product Role: renamed as to “Role”

5 PRI\_Reported Brand/Product Name: renamed as “Brand/Product”

6. PRI\_FDA Industry Code: converted from numeric to string

7. PRI\_FDA Industry Name: renamed as “Industry”; only kept the word before first “/” as industry name. For example: Soft Drink/Water converted as soft drink. For Vit/Min/Prot/Unconv Diet(Human/Animal), renamed as “Unconventional Food”.

8. CI\_Age Unit: renamed as “Age Unit”; trailing “(s)” deleted at cells applied; normalized together with “Age” (shown below).

*9. CI\_Age at Adverse Event*: renamed as “Age”. This variable was *normalized* based on “Age\_Unit”: if Age\_unit is “Decade”, Age is ten times of current value and Age Unit changed to “year”; if Age\_unit is “day”, then any age larger than 31 is divided by 30 to be converted to month, and Age\_unit is changed to “month” accordingly; if Age\_unit is “week”, ages greater than or equal to 4 is converted to “month” otherwise converted to days, so is “Age Unit” accordingly. If if Age\_unit is “month”, ages greater than 12 converted to “year”. If Age\_unit is “year”, any value greater than or equal to 100 deleted as outliers.

After normalizing, values in Age\_unit and their counts are: Year: 47309; Month: 998; Day:116. The three values stand for 3 groups of people that are approximately mutually exclusive (minor overlapping problem of the three sets): people with Age\_unit = year defined as “children and adults” (age greater than or equal to 1 year); people with Age\_unit = month defined as “Infants” (age less than 12 months but greater than 1 month); people with Age\_unit = day defined as “babies” (age less than 1 month).

10. CI\_Gender: renamed as “Gender”

11. AEC\_One Row Outcomes: no change

12. SYM\_One Row Coded Symptoms: no change

Also, for the purpose of data visualization and analysis, *a group of derived variables* were created based on original variables:

1. Report\_Year: year part of RA\_CAERS Created Date.

2. Age\_interval: created based on Age and Age Unit. For children and adults, it has categories of '0-20', '21-40', '41-60','61-80','81-100'; For infants, it has values of '0-3', '3-6', '6-9','9-12'; For babies, it has values of '0-5', '5-10', '10-15','15-20', '20-25', '25-30'.

3. life\_threatening: an indicator variable created based on “AEC\_One Row Outcomes”. If an outcome contains “DEATH” or “LIFE THREATENING”, then it is assigned value 1, otherwise 0.

**Sample data**

Part 3: Preliminary analysis and Data visualization

(Note: Because we are only allowed to upload 2 plots, I choose encoding as many variables as possible in a single plot, although I made over 30 simple plots in total. In a real project report, I tend not to include many variables in a single plot like the plots I submit this time)

The top 10 industrial categories causing adverse events are shown in Table 1.

Table 1: Top 10 industrial categories causing adverse events

(by gender and age group)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Babies | Infants | Children and adults |
| Female | Baby Food Prod,  Unconventional\_food,  Cosmetics,  Dietary Conv Food,  Soft Drink,  Vegetables,  Egg | Baby Food Prod,  Unconventional\_food,  Cosmetics,  Dietary Conv Food,  Soft Drink,  Bakery Prod,  Milk,  Fruit,  Nuts,  Coffee | Unconventional\_food,  Cosmetics,  Nuts,  Dietary Conv Food,  Bakery Prod,  Vegetables,  Fishery,  Soft Drink,  Milk,  Fruit |
| Male | Baby Food Prod,  Unconventional\_food,  Cosmetics,  Soft Drink,  Dietary Conv Food,  Ice Cream Prod | Baby Food Prod,  Unconventional\_food,  Dietary Conv Food,  Cosmetics,  Milk,  Soft Drink,  Fruit,  Nuts,  Bakery Prod,  Vegetables | Unconventional\_food,  Fishery,  Nuts,  Cosmetics,  Bakery Prod,  Soft Drink,  Fruit,  Vegetables,  Dietary Conv Food,  Milk |
| Overall | Baby Food Prod,  Unconventional\_food,  Cosmetics,  Soft Drink,  Dietary Conv Food,  Vegetables,  Ice Cream Prod,  Egg | Baby Food Prod, Unconventional\_food,  Dietary Conv Food,  Cosmetics,  Soft Drink,  Milk,  Bakery Prod,  Fruit,  Vegetables,  Nuts | Unconventional\_food,  Cosmetics,  Nuts,  Fishery,  Bakery Prod,  Dietary Conv Food,  Soft Drink,  Fruit,  Vegetables,  Milk |

It can be seen from Table 1 that in terms of industrial categories causing adverse events, there is no obvious difference between male and female. Furthermore, Unconventional food and Cosmetics are common top sources of adverse events for both female and male. It is also surprising that for male children and adults, cosmetics is the fourth top source of adverse events.

To further explore the frequency of adverse events on people of different genders and ages, a stacked bar chart showing frequency of adverse events by gender and industrial categories on children and adults group was generated (Figure 1).

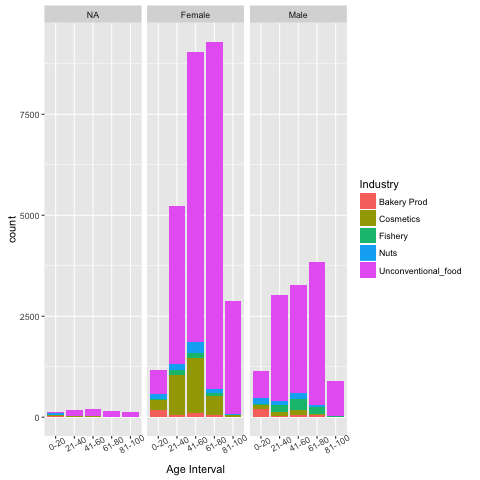


Figure 1: Frequency of adverse events on children and adults group by industry

As shown in Figure 1, among the top 5 industries causing adverse events, unconventional food is a curse to both male and female. In particular, female suffers more frequently from unconventional food than male, especially female of age 60 to 80 and 40-60. In addition, cosmetics is a second source of adverse events for female of different age groups. For male, apart from unconventional food, they are mainly bothered by fishery, and the frequency is higher than that to female. The frequency of adverse events caused by nuts is almost same for both female and male.

A line graph showing the trend of frequency of adverse events of the same top 5 industries is shown as Figure 2.

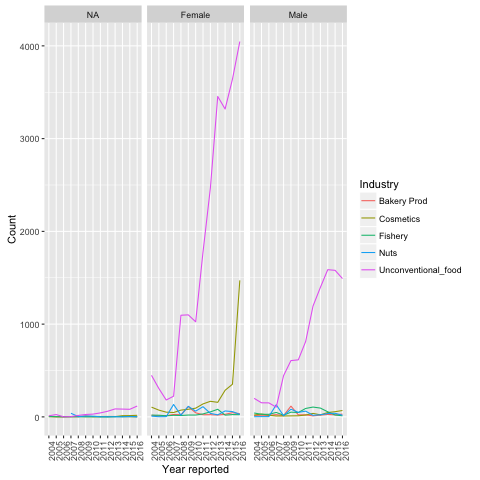
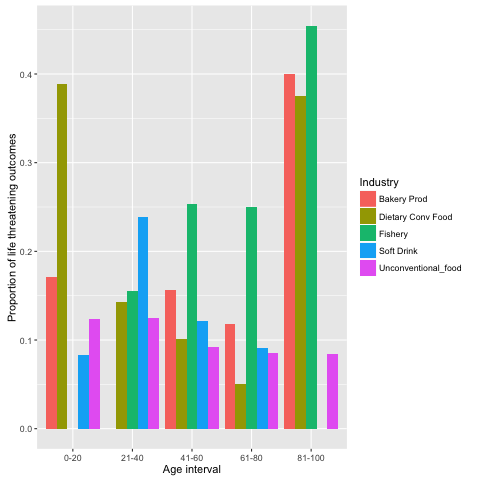


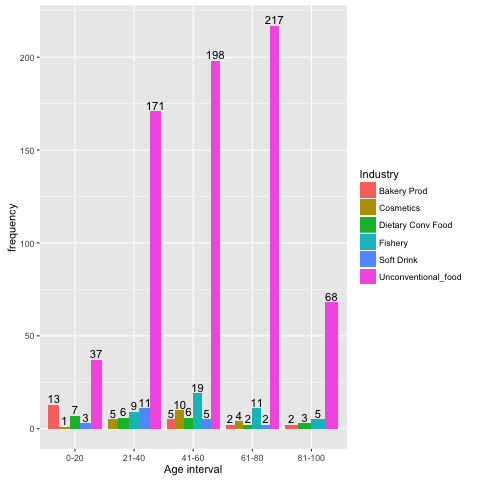
Figure 2: Trend of frequency of adverse events

on children and adults group by industry

It is surprising that frequency of adverse events related to unconventional food and cosmetics on female increased rapidly from year 2004 to 2016. Even for male, there is a slight increasing trend of cosmetics related events. The frequency of rest of industries fluctuated in the same time period, but no transparent patterns found.

**Life-threatening**

****



Maniwordle

Part 4: Modeling and Analysis

The analysis focuses on the outcome of adverse events. Particularly, if the outcome is death or life threatening, one would care about which factors take most part.

To find the relationship between worst outcome, i.e. death or life threatening, and industry, gender and age, a decision tree was built to explore the possible relationship.

A decision tree with derived variable “life-threatening” as label, and industry, gender and age as predictors gives the feature importance scores as follows:

Industry: 0.30880775, Age: 0.65859297, Gender: 0.03259929

Note that for predictor “industry”, the default ordinary encoding was applied although one-hot encoding is more proper for this variable. The model with “industry” one-hot encoded, has almost same prediction ability as the model with ordinary encoding. Yet ordinary encoding can give a more direct measurement of relative importance of predictors.

Based on the feature importance scores given above, one can see that Age is the top factor that results in outcome of death or life threatening. The result coincides with our common sense because the elders are more vulnerable to adverse events, thus have higher rate of an extreme outcome.

To alleviate the impact of age factor, two more decision trees were built fixing the age group at 20-40 and 40- 60 respectively. The relative importance of features are shown below:

Model 2 (age group 20-40):

Industry: 0.40440189, Age: 0.5121695 , Gender: 0.08342861

Model 3 (age group 40-60):

Industry: 0.47385953, Age: 0.48824346, Gender: 0.03789701

The result above indicates when fixing age group at a certain range, the impact of industry will become obvious. Especially, for people of age 40-60, the impact of industry and age is almost identical to yielding an extreme outcome.

Part 5: Future work

Future work will focus on following aspects:

First of all, industries that result in significant higher rate of death/ life threatening should to find out, if any (hopefully there is none). Since there are over 30 industries, filtering out certain industries based on data visualization could work well in analysis.

Secondly, more analysis should be done on baby and infants (defined in this report) although less data available so far. Especially, more attention should be paid if there is any increasing trend of adverse events.

Thirdly, for industries that have caused most adverse events such as unconventional food and cosmetics, exploration of brand/product is crucial. A text exploration by virtue of websites like ManiWordle should be a good starting point.

Part 6: Appendix

Feature importance of predictors in model with one-hot encoded “industry” (a total of 136 features)

1.09510255e-02, 4.12025703e-03, 1.38646178e-02,

0.00000000e+00, 1.15656338e-02, 1.84637082e-02,

3.82686412e-04, 1.88321703e-03, 1.29617612e-02,

0.00000000e+00, 2.01263856e-02, 2.64286624e-02,

0.00000000e+00, 7.74484404e-03, 6.50566900e-03,

9.70761357e-03, 7.20051684e-03, 1.21671047e-02,

8.13208624e-03, 1.78023757e-02, 0.00000000e+00,

0.00000000e+00, 1.59736847e-02, 0.00000000e+00,

1.44828584e-02, 0.00000000e+00, 8.34129802e-03,

0.00000000e+00, 1.56614216e-02, 2.11755709e-02,

1.06456402e-03, 0.00000000e+00, 3.46366705e-02,

4.70888786e-03, 0.00000000e+00, 1.20438282e-02,

1.00754463e-02, 5.97097971e-03, 1.92913288e-02,

0.00000000e+00, 7.02617060e-03, 9.84744338e-03,

1.21575674e-02, 0.00000000e+00, 1.74930211e-02,

8.95480117e-03, 0.00000000e+00, 4.24861216e-05,

8.37092425e-05, 4.10725141e-03, 0.00000000e+00,

6.70055188e-03, 4.42057402e-05, 8.57446699e-05,

4.45295918e-05, 1.80360293e-02, 2.33317716e-02,

8.63962172e-03, 4.10658753e-03, 1.39463021e-04,

1.57513281e-03, 1.81513655e-04, 1.16143029e-02,

1.36321599e-03, 2.68388291e-03, 6.11545649e-03,

1.54785209e-02, 3.06998218e-03, 9.51159232e-03,

1.81471111e-02, 1.38560781e-02, 2.41195720e-03,

1.45890266e-02, 8.58514587e-03, 1.10642261e-02,

3.17885390e-04, 1.03448827e-02, 6.95238539e-03,

1.32014553e-02, 6.18658529e-03, 1.11758342e-03,

4.15254436e-03, 1.77216427e-02, 3.48260692e-03,

1.50242008e-03, 2.38029820e-03, 1.00138709e-02,

1.03623206e-02, 7.35987244e-03, 2.27566534e-04,

4.53051519e-03, 2.33009232e-03, 1.46595722e-02,

4.56062677e-03, 6.15684670e-03, 3.10166201e-03,

1.21974694e-02, 1.45336927e-02, 8.61120194e-04,

7.91422666e-04, 1.92775823e-02, 6.47480502e-03,

4.50127855e-03, 1.13589032e-02, 2.46347188e-02,

1.16115560e-02, 1.48301512e-02, 1.88172491e-03,

3.17483375e-03, 7.78291322e-03, 5.47678809e-03,

1.67213697e-02, 5.33012139e-03, 1.04412278e-02,

9.32712401e-07, 2.58277392e-03, 1.47522000e-05,

0.00000000e+00, 1.16519577e-02, 1.19982549e-02,

2.04183560e-03, 1.93192454e-03, 1.82760997e-03,

8.55219868e-03, 1.69271812e-03, 6.77052256e-03,

1.55587977e-03, 2.94209100e-03, 1.89367209e-03,

1.36874380e-04, 8.47177876e-05, 4.71967569e-03,

8.78552487e-05, 2.51782704e-02, 5.63357099e-03,

1.16341928e-02])

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Industry | Age\_interval | freq.x | freq.y | prop |
| 1 | Bakery Prod | 0-20 | 76 | 13 | 0.171052632 |
| 2 | Bakery Prod | 21-40 | 21 | NA | NA |
| 3 | Bakery Prod | 41-60 | 32 | 5 | 0.15625 |
| 4 | Bakery Prod | 61-80 | 17 | 2 | 0.117647059 |
| 5 | Bakery Prod | 81-100 | 5 | 2 | 0.4 |
| 6 | Dietary Conv Food | 0-20 | 18 | 7 | 0.388888889 |
| 7 | Dietary Conv Food | 21-40 | 42 | 6 | 0.142857143 |
| 8 | Dietary Conv Food | 41-60 | 59 | 6 | 0.101694915 |
| 9 | Dietary Conv Food | 61-80 | 40 | 2 | 0.05 |
| 10 | Dietary Conv Food | 81-100 | 8 | 3 | 0.375 |
| 11 | Fishery | 0-20 | 14 | NA | NA |
| 12 | Fishery | 21-40 | 58 | 9 | 0.155172414 |
| 13 | Fishery | 41-60 | 75 | 19 | 0.253333333 |
| 14 | Fishery | 61-80 | 44 | 11 | 0.25 |
| 15 | Fishery | 81-100 | 11 | 5 | 0.454545455 |
| 16 | Soft Drink | 0-20 | 36 | 3 | 0.083333333 |
| 17 | Soft Drink | 21-40 | 46 | 11 | 0.239130435 |
| 18 | Soft Drink | 41-60 | 41 | 5 | 0.12195122 |
| 19 | Soft Drink | 61-80 | 22 | 2 | 0.090909091 |
| 20 | Soft Drink | 81-100 | 2 | NA | NA |
| 21 | Unconventional\_food | 0-20 | 298 | 37 | 0.124161074 |
| 22 | Unconventional\_food | 21-40 | 1370 | 171 | 0.124817518 |
| 23 | Unconventional\_food | 41-60 | 2140 | 198 | 0.092523364 |
| 24 | Unconventional\_food | 61-80 | 2538 | 217 | 0.085500394 |
| 25 | Unconventional\_food | 81-100 | 809 | 68 | 0.084054388 |