Adverse Food Events - Exploratory Data Analysis

Tsun Hei Tai

24/10/2021

Exploratory data analysis conducted using data extracted from from the US Food and Drug Administration (FDA) Center for Food Safety and Applied Nutrition (CFSAN) Adverse Event Reporting System (CAERS).The data extracted contains approximately 90000 reactions recorded from 2004 to mid 2017.

The adverse event reports about a product and the total number of adverse event reports for that product in CAERS only reflect information AS REPORTED and do not represent any conclusion by FDA about whether the product actually caused the adverse events. For any given report, there is no certainty that a suspected product caused a reaction.

Acknowledgements - This dataset was gathered by the US Food and Drug Administration and downloaded from https://www.kaggle.com/fda/adverse-food-events

Key findings

- Product that led to most deaths were raw oysters
- Supplements were the most common products that led to adverse reactions
- More females reported adverse reactions compared to males
- Diarrhoea , Vomiting , Nausea , Abdominal Pain were the most common symptoms

Setup working environment

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggplot2)
```

Import data

```
food <- read.csv("CAERS_ASCII_2004_2017Q2.csv")
```

Data manipulation

```
food$RA_CAERS.Created.Date <- as.Date(food$RA_CAERS.Created.Date, "%m/%d/%y")
food$AEC_Event.Start.Date <- as.Date(food$AEC_Event.Start.Date, "%m/%d/%y")
colnames(food)
head(food)
str(food)
```

Column names are not user friendly, change them to names that are easier to use.

```
names(food) <- c('report_no','created_date','start_date','product_role','product_name','industry_code',
                 'outcomes','symptoms')
names(food)
```

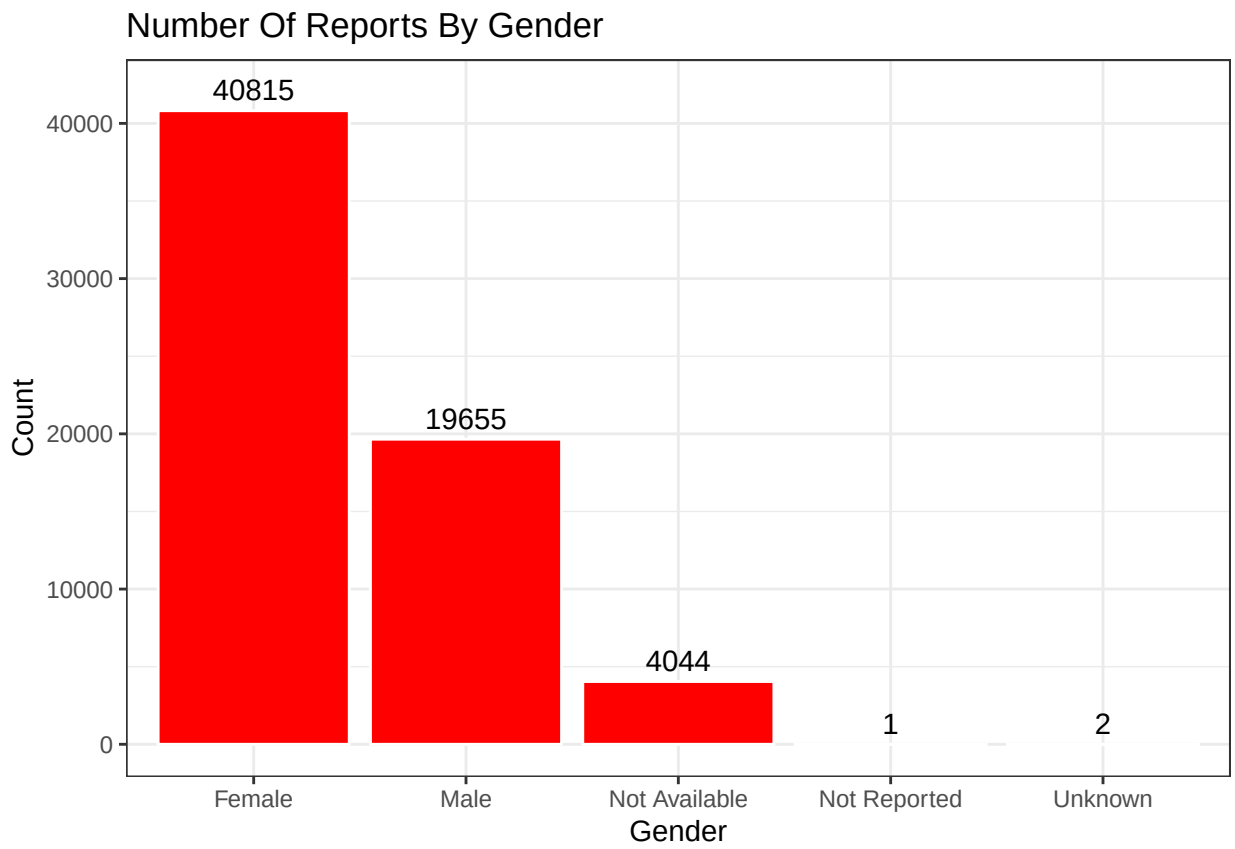
```
##  [1] "report_no"            "created_date"         "start_date"
##  [4] "product_role"         "product_name"         "industry_code"
##  [7] "industry_name"        "age_at_adverse_event" "age_unit"
## [10] "gender"               "outcomes"             "symptoms"
```

We know from the README file with this dataset that there are duplicates as reports for the same issue might have been logged multiple times. Remove them from the data set as we know that each report number is unique.

```
food_v1 <- food %>%
  distinct(report_no, .keep_all = TRUE)
```

Number of reports by gender

```
  ggplot(food_v1, aes(x = gender)) +
  geom_bar(colour="white", fill = 'red') +
  stat_count(geom = "text", aes(label = ..count..), vjust = -0.5) +
  coord_cartesian(ylim = c(0, 42000)) +
  labs(x = 'Gender', y = 'Count',
       title = 'Number Of Reports By Gender') +
    theme_bw()
```

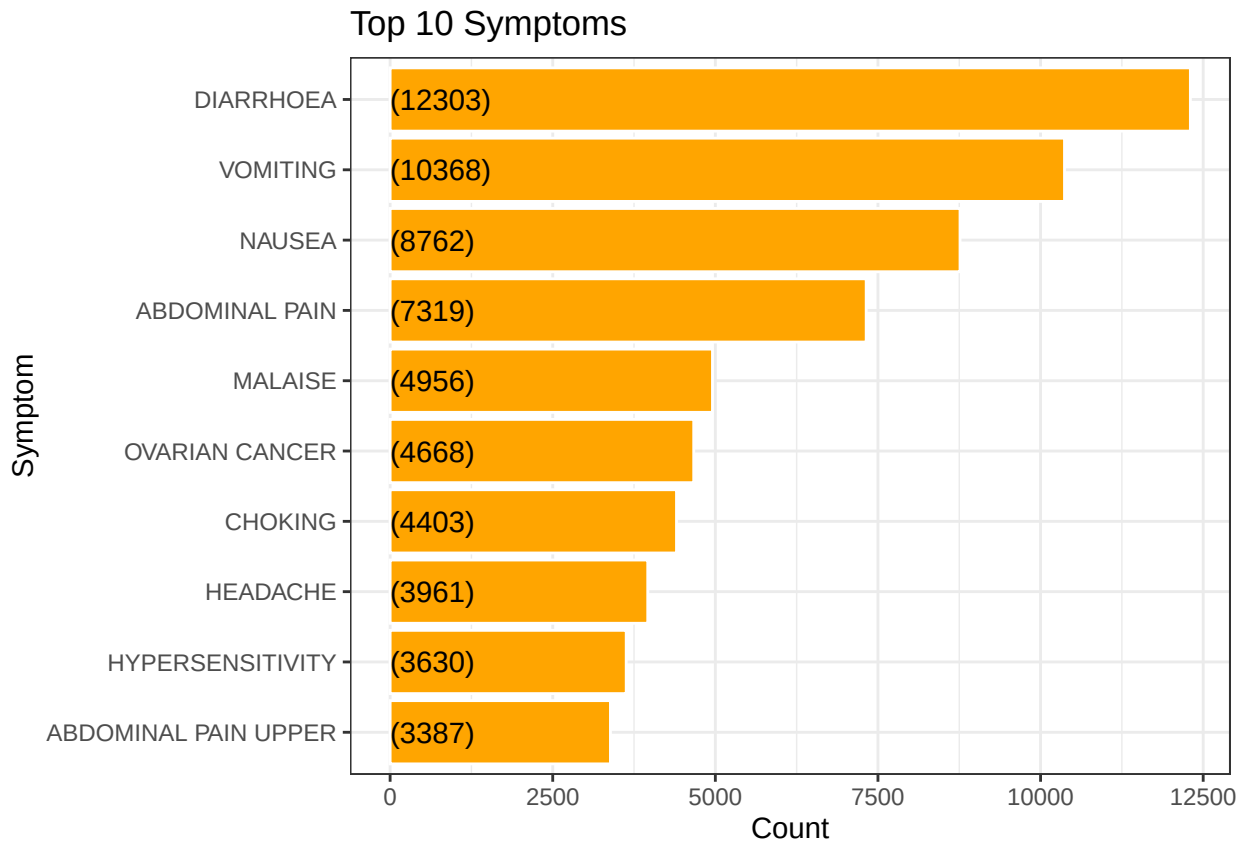


Interesting that most of the reports of adverse events were made by females.

Ten most common symptoms

```
bpsymptoms <- function(food_v1)
{
symptoms <- str_split(food_v1$symptoms,',')

all_symptoms <- data.frame(matrix(unlist(symptoms),byrow=T),stringsAsFactors=FALSE)

colnames(all_symptoms) <- c("symptom_name")

#trimws removes leading and/or trailing whitespaces from a character string.
all_symptoms$symptom_name = trimws(all_symptoms$symptom_name)

all_symptoms %>%
  group_by(symptom_name) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count)) %>%
  ungroup() %>%
  mutate(symptom_name = reorder(symptom_name,Count)) %>%
  head(10) %>%

  ggplot(aes(x = symptom_name,y = Count)) +
  geom_bar(stat='identity',colour="white", fill = 'orange') +
  geom_text(aes(x = symptom_name, y = 1, label = paste0("(",Count,")",sep="")),
```

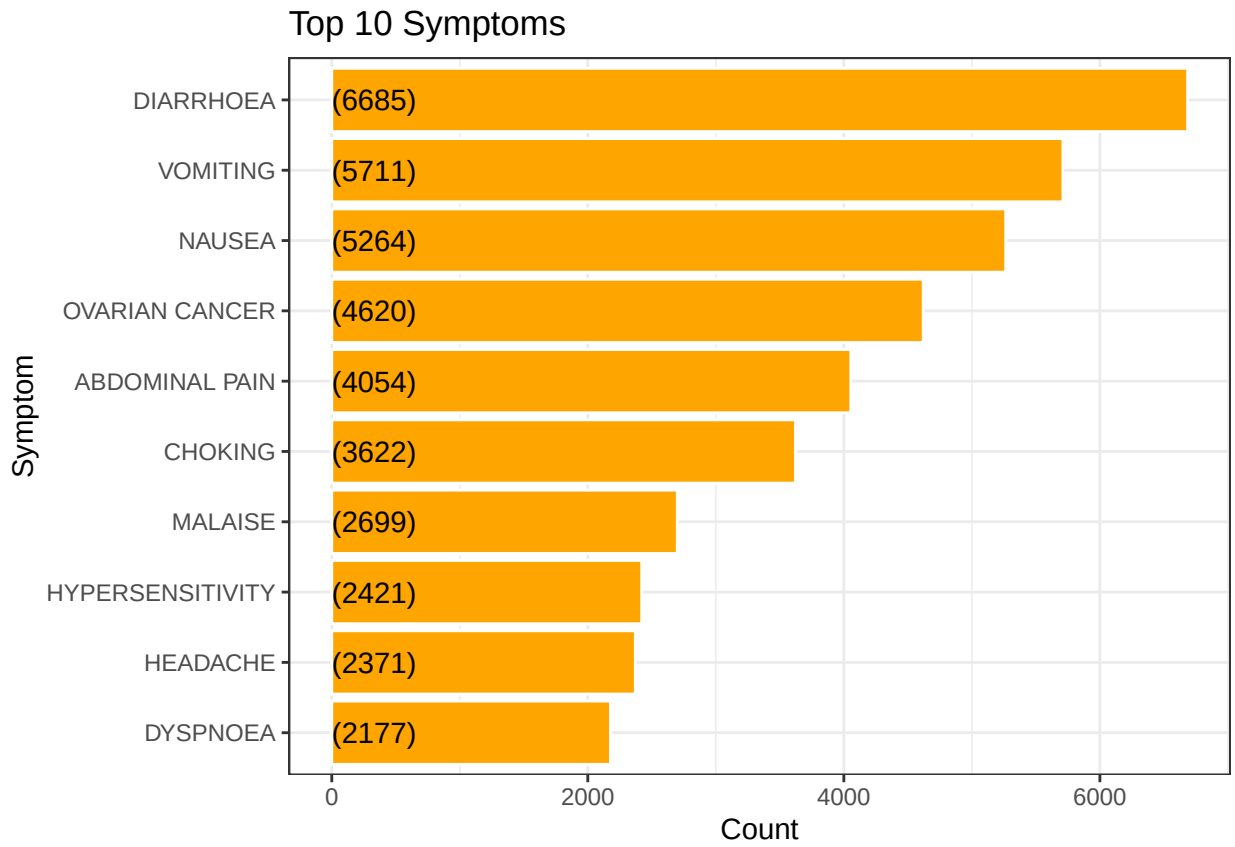
```
            hjust=0, vjust=.5, size = 4, colour = 'black') +
  labs(x = 'Symptom',
       y = 'Count',
       title = 'Top 10 Symptoms') +
  coord_flip() +
  theme_bw()
}

bpsymptoms(food_v1)
```



The most common symptoms people experienced were diarrhoea, vomiting, nausea and abdominal pain. These are quite common symptoms people experience when they have an adverse reaction to food.

Ten most common symptoms for females

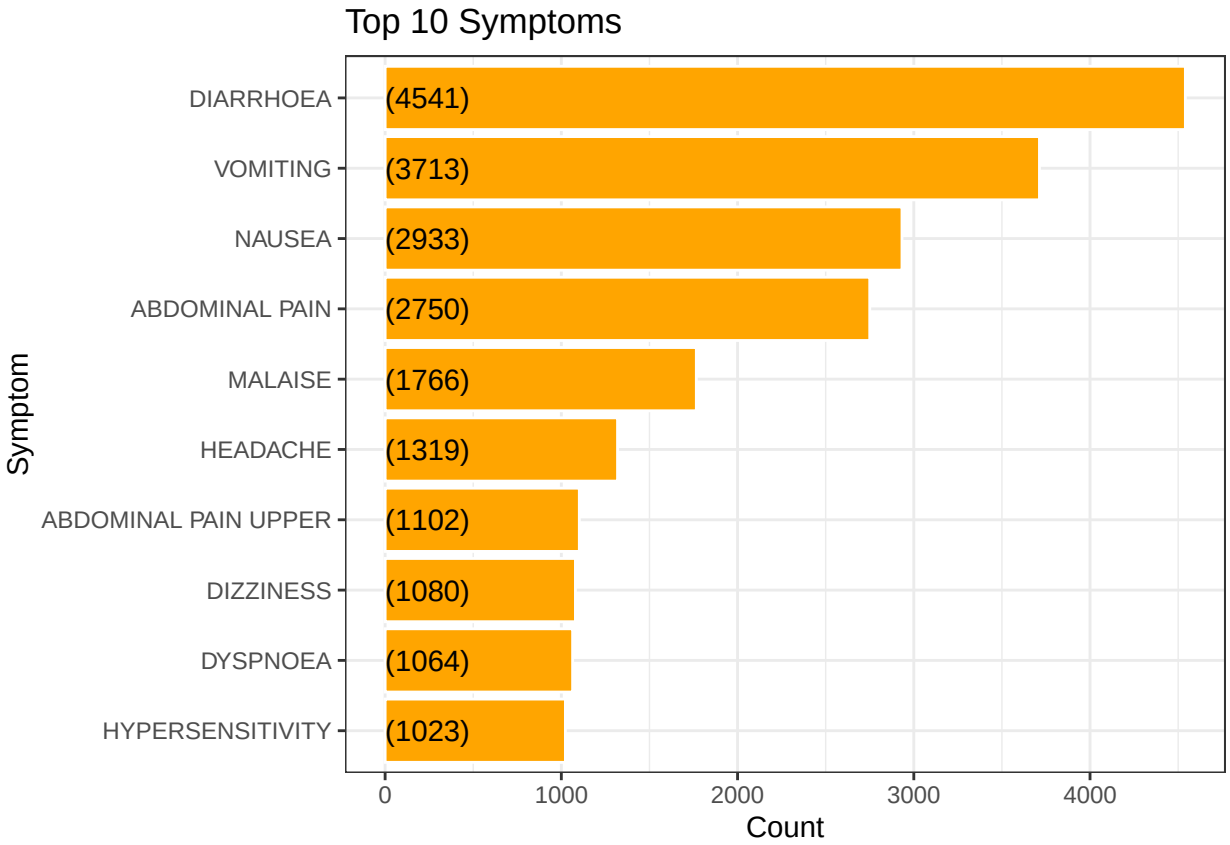
```
food_v1 %>%
  filter(gender == 'Female') %>%
  bpsymptoms()
```



Again diarrhoea, vomiting and nausea appear as the top symptoms for females. The interesting finding here is that ovarian cancer is fourth on this list.

Ten most common symptoms for males

```
food_v1 %>%
  filter(gender == 'Male') %>%
  bpsymptoms()
```

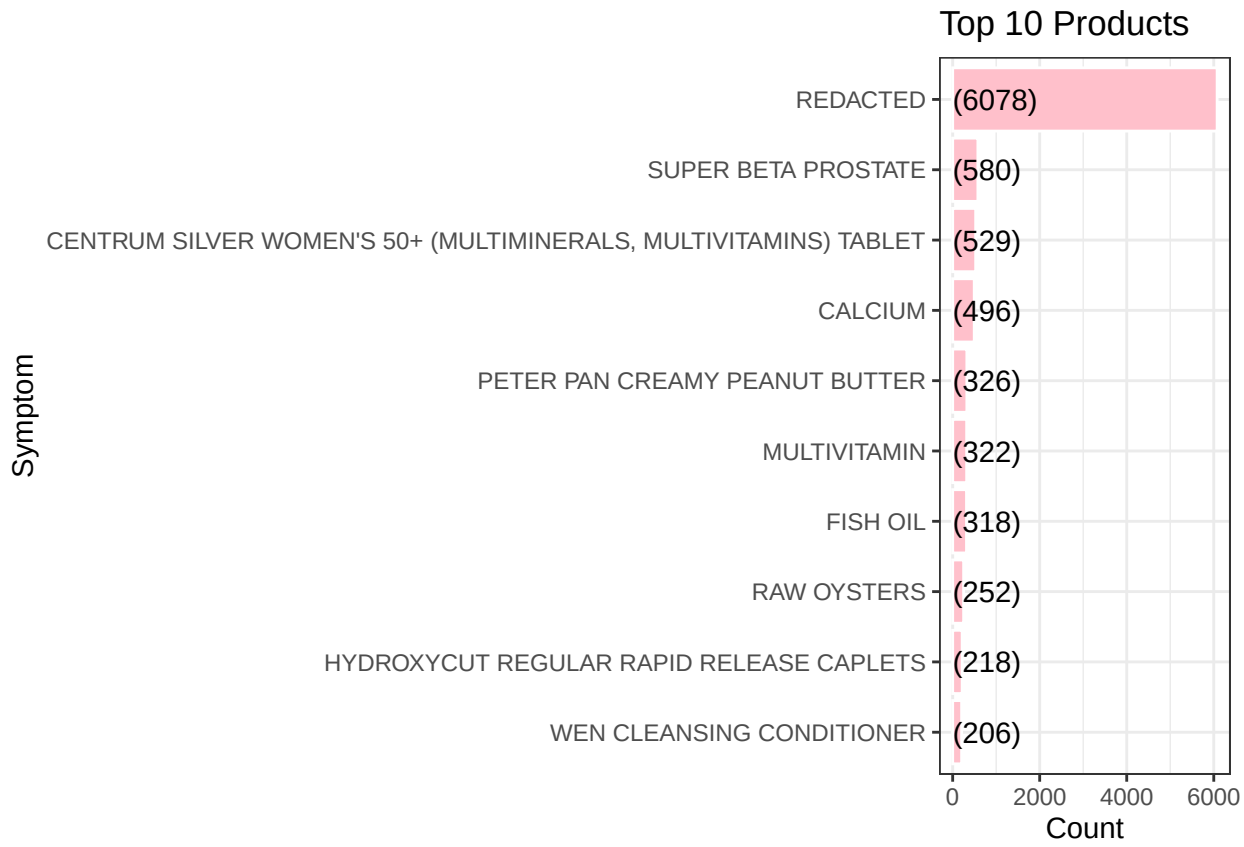


The top ten symptoms for males is quite similar to the top ten symptoms for the entire dataset, except for ovarian cancer.

Ten most common foods

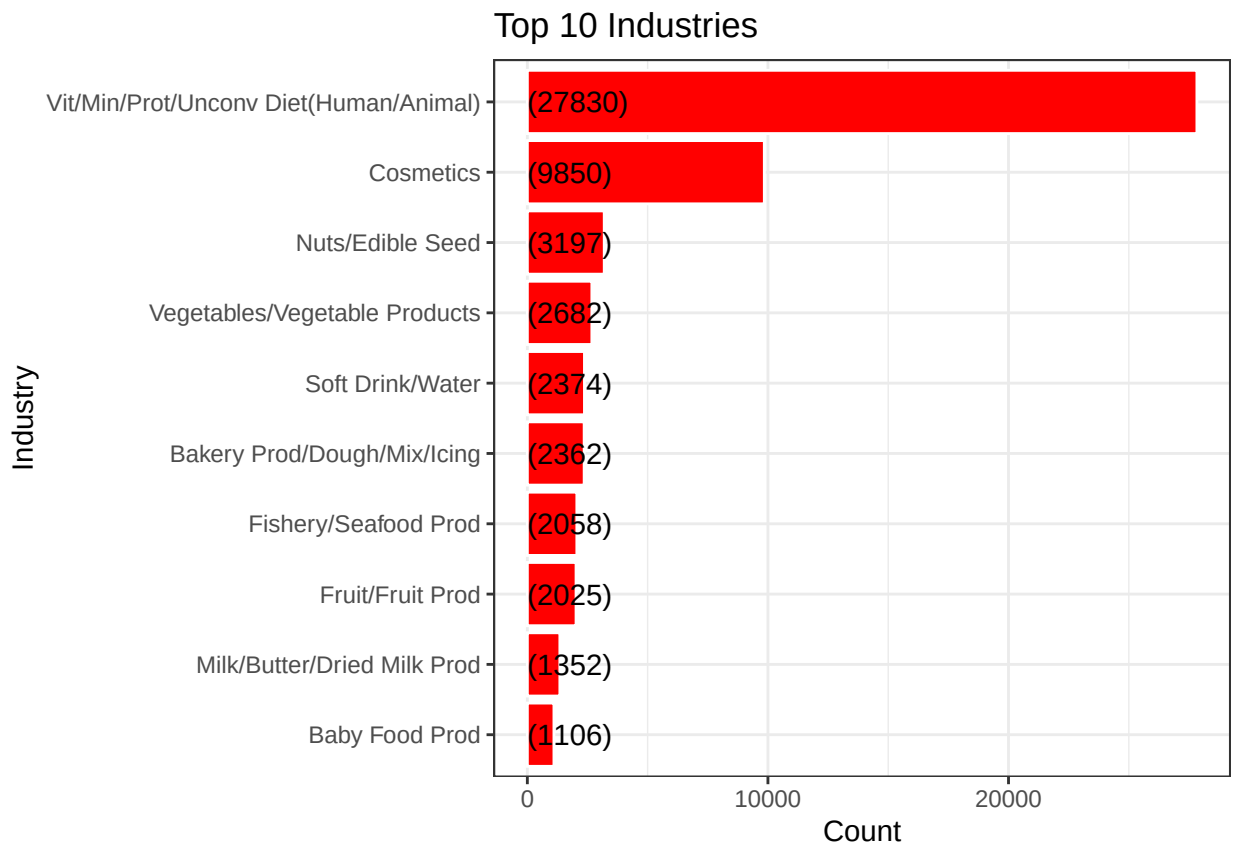
```
bpproducts <- function(food_v1)
{
food_v1 %>%
  group_by(product_name) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count)) %>%
  ungroup() %>%
  mutate(product_name = reorder(product_name,Count)) %>%
  head(10) %>%

  ggplot(aes(x = product_name,y = Count)) +
  geom_bar(stat='identity',colour="white", fill = 'pink') +
  geom_text(aes(x = product_name, y = 1, label = paste0("(",Count,")",sep="")),
            hjust=0, vjust=.5, size = 4, colour = 'black') +
  labs(x = 'Symptom',
       y = 'Count',
       title = 'Top 10 Products') +
  coord_flip() +
  theme_bw()
}
```

```
bpproducts(food_v1)
```



Ten most common industries

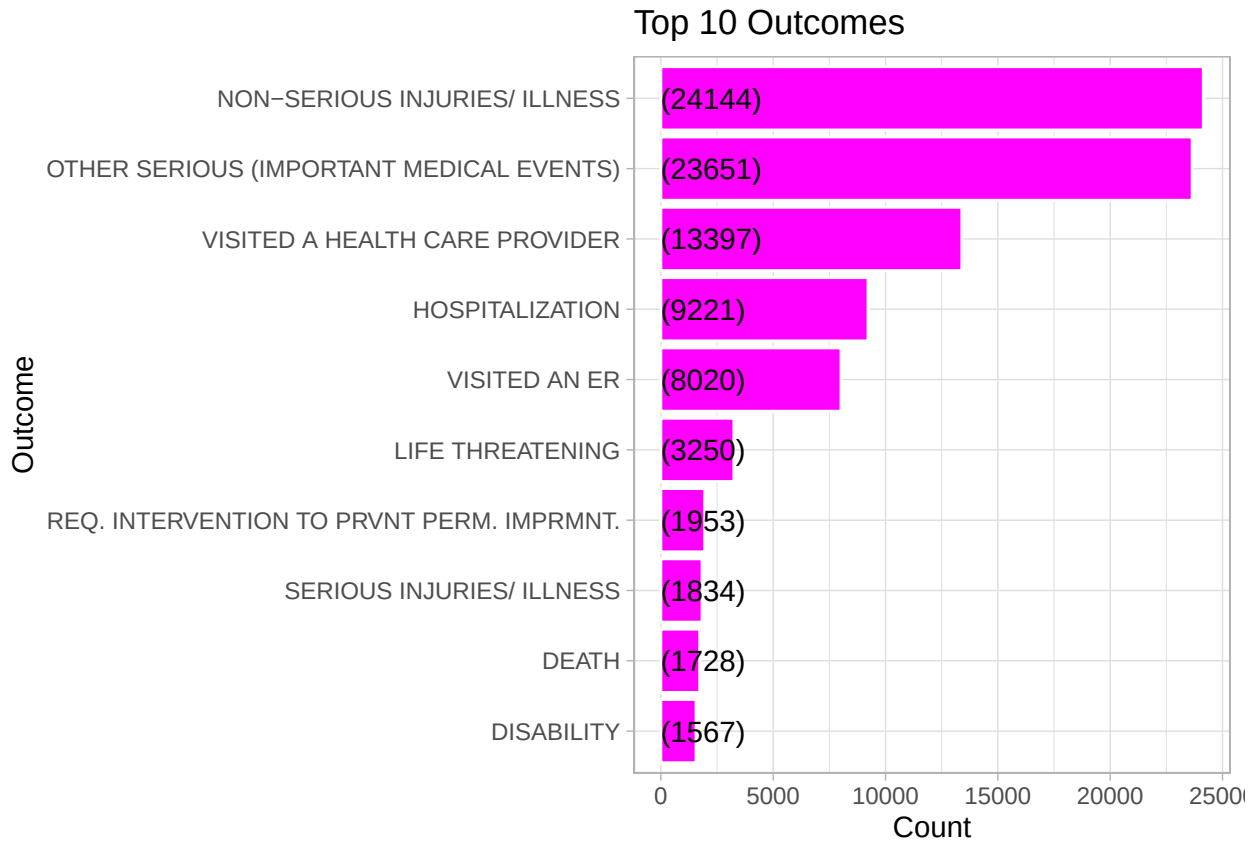
```
food_v1 %>%
  group_by(industry_name) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count)) %>%
  ungroup() %>%
  mutate(industry_name = reorder(industry_name,Count)) %>%
  head(10) %>%

  ggplot(aes(x = industry_name,y = Count)) +
  geom_bar(stat='identity',colour="white", fill = 'red') +
  geom_text(aes(x = industry_name, y = 3, label = paste0("(",Count,")",sep="")),
            hjust=0, vjust=.5, size = 4, colour = 'black') +
  labs(x = 'Industry',
       y = 'Count',
       title = 'Top 10 Industries') +
  coord_flip() +
  theme_bw()
```



Ten most common outcomes

```
outcomes = str_split(food_v1$outcomes,',')

all_outcomes <- data.frame(matrix(unlist(outcomes),byrow=T),stringsAsFactors = FALSE)

colnames(all_outcomes) = c("outcome_name")

all_outcomes$outcome_name = trimws(all_outcomes$outcome_name)

all_outcomes %>%
  group_by(outcome_name) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count)) %>%
  ungroup() %>%
  mutate(outcome_name = reorder(outcome_name,Count)) %>%
  head(10) %>%

  ggplot(aes(x = outcome_name, y = Count)) +
  geom_bar(stat = 'identity',colour='white',fill='magenta') +
  geom_text(aes(x = outcome_name, y = 1, label = paste0("(",Count,")",sep="")),
            hjust=0,vjust=.5,size=4,colour = 'black') +
  labs(x = 'Outcome',
       y = 'Count',
       title = 'Top 10 Outcomes') +
```

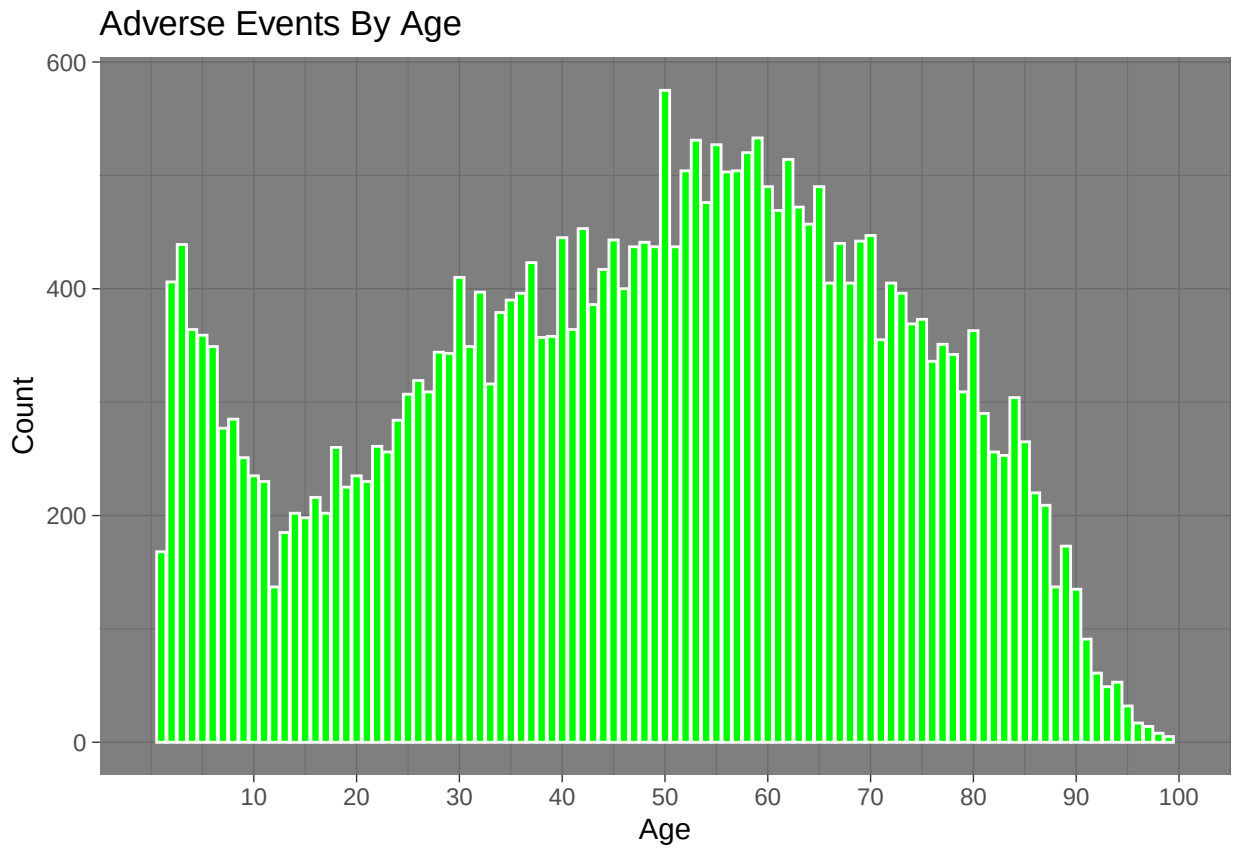
```
  coord_flip() +
  theme_light()
```



Top outcomes were non-serious injuries/illnesses, other serious important medical events or visited a health care provider. Also quite a high number of hospitalisations which includes visiting an ER.

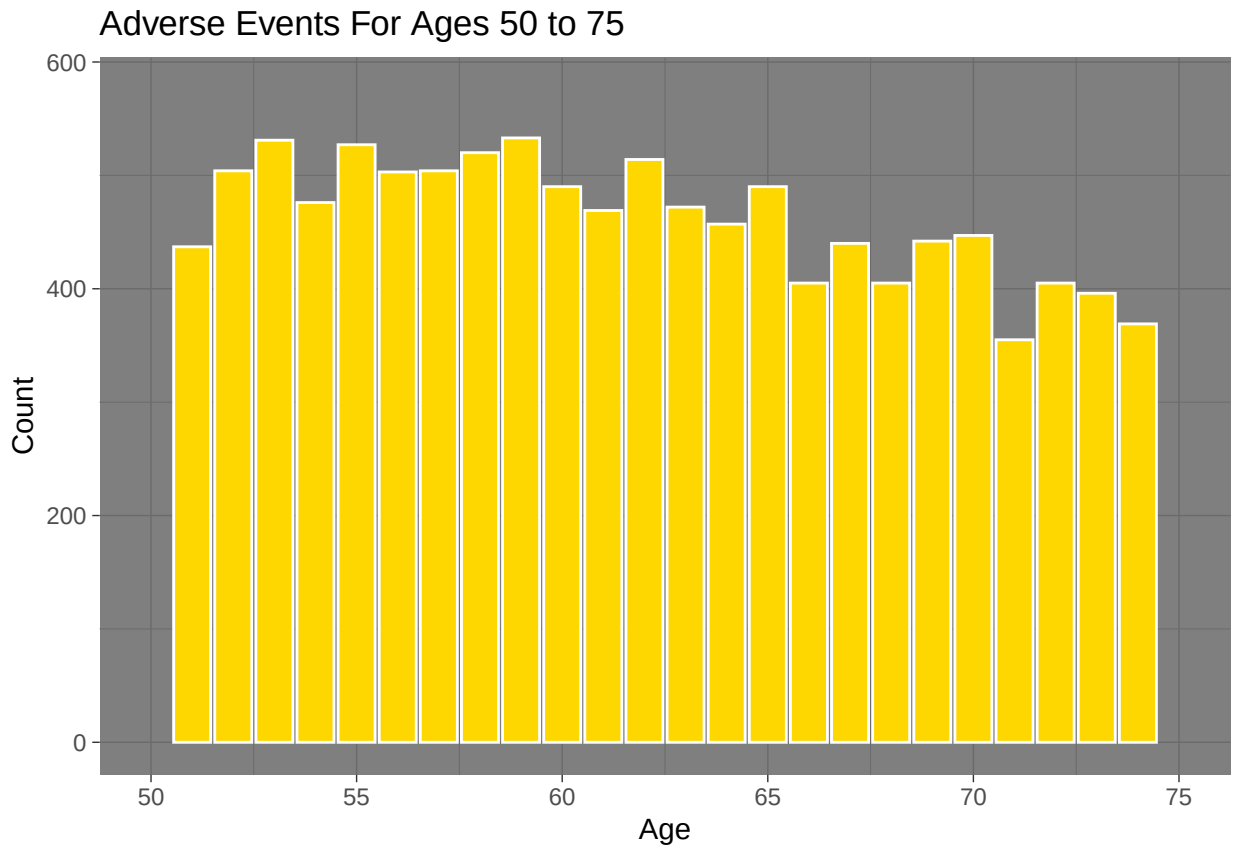
Reports by age

```
food_v1 %>%
  filter(!is.na(age_at_adverse_event)) %>%
  group_by(age_at_adverse_event) %>%
  summarise(Count = n()) %>%
  ungroup() %>%

  ggplot(aes(x = age_at_adverse_event, y = Count)) +
  geom_bar(stat = 'identity', colour = 'white', fill = 'green') +
  scale_x_continuous(limits = c(0,100), breaks = c(10,20,30,40,50,60,70,80,90,100)) +
  labs(x = 'Age', y = 'Count', title = 'Adverse Events By Age') +
  theme_dark()
```



By looking at the number of adverse event reports by age, this graph shows that quite a few of the reports are between the 50 - 70 age bracket.

```
food_v1 %>%
  filter(!is.na(age_at_adverse_event)) %>%
  group_by(age_at_adverse_event) %>%
  summarise(Count = n()) %>%
  ungroup() %>%

  ggplot(aes(x = age_at_adverse_event, y = Count)) +
  geom_bar(stat = 'identity', colour = 'white', fill = 'gold') +
  scale_x_continuous(limits = c(50,75)) +
  labs(x = 'Age', y = 'Count', title = 'Adverse Events For Ages 50 to 75') +
  theme_dark()
```

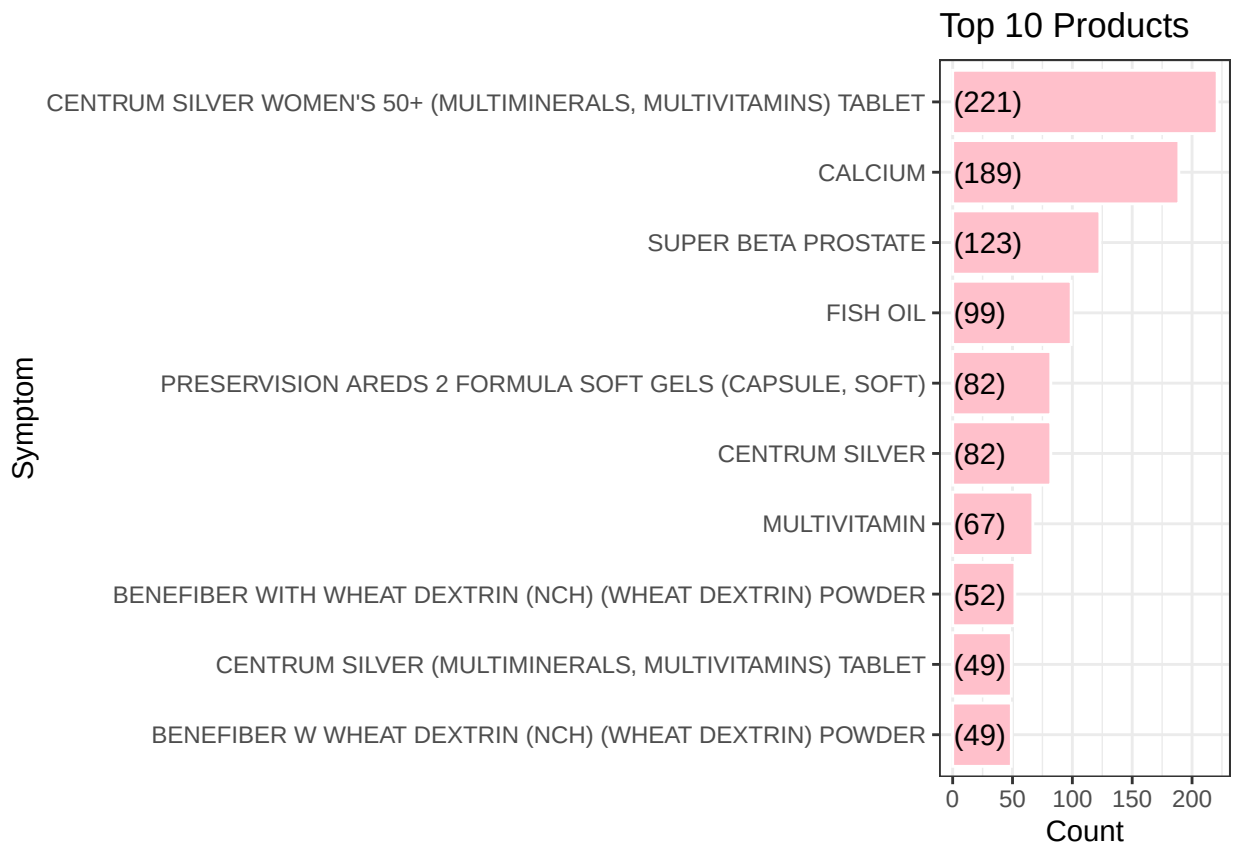


The 50 to 75 year old age bracket almost had 400 adverse events per age in this age bracket. From the graph, there are fewer in the 70 to 75 age bracket, so I will group other data analysis using ages 70 and above.

Products by age

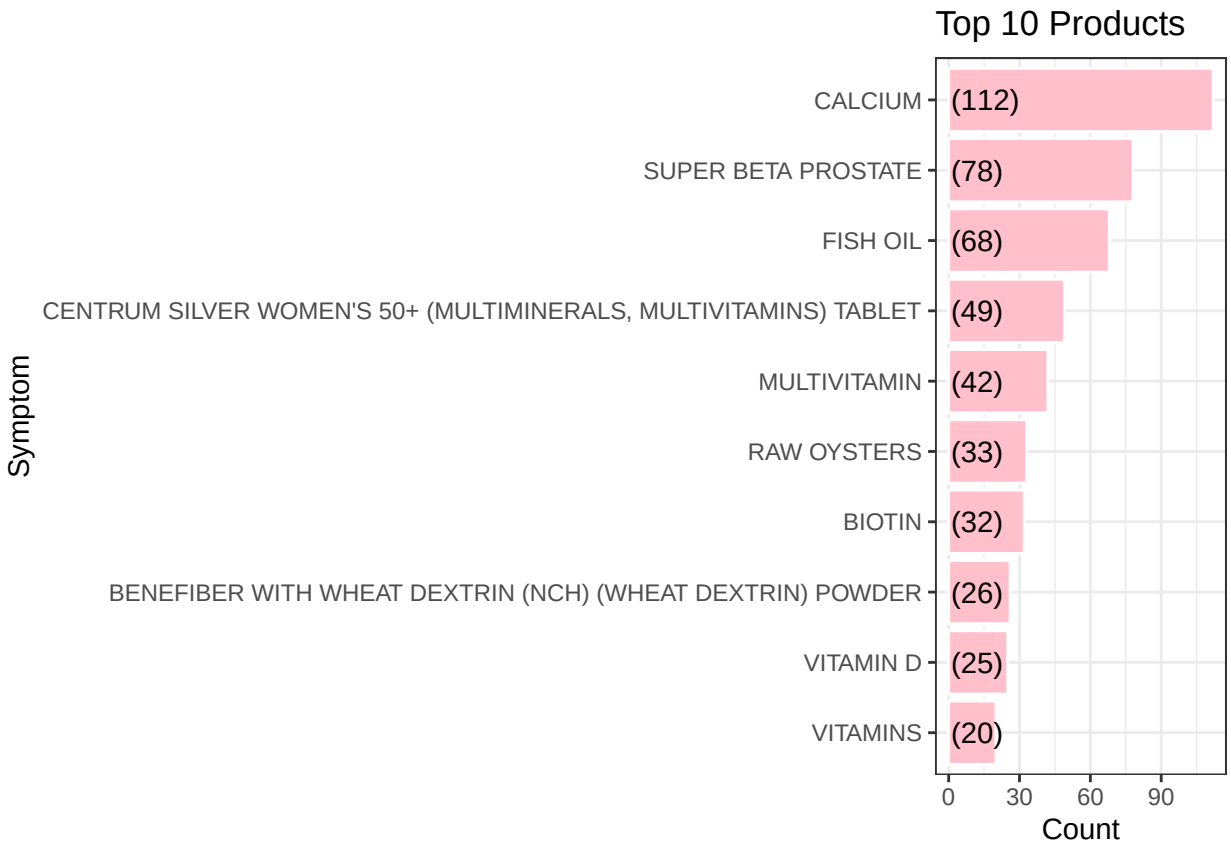
Age 70 and above adverse event products

```
food_v1 %>%
  mutate(age_at_adverse_event = as.numeric(age_at_adverse_event)) %>%
    filter(age_at_adverse_event >=70) %>%
  filter(product_name != 'REDACTED') %>%
  bpproducts()
```



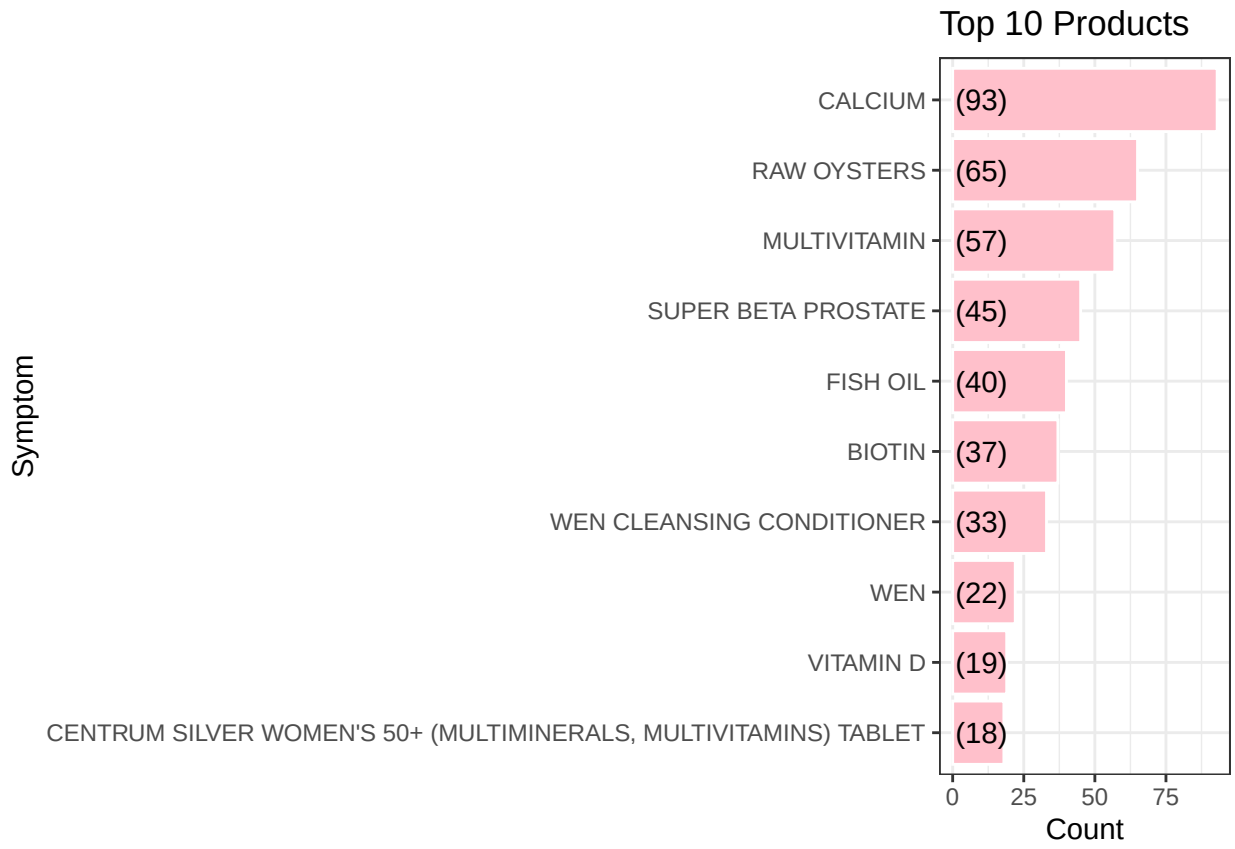
Age 60 to 69 adverse event products

```
food_v1 %>%
  mutate(age_at_adverse_event = as.numeric(age_at_adverse_event)) %>%
    filter(age_at_adverse_event <70) %>%
  filter(age_at_adverse_event >=60) %>%
  filter(product_name != 'REDACTED') %>%
  bpproducts()
```



Age 50 to 59 adverse event products

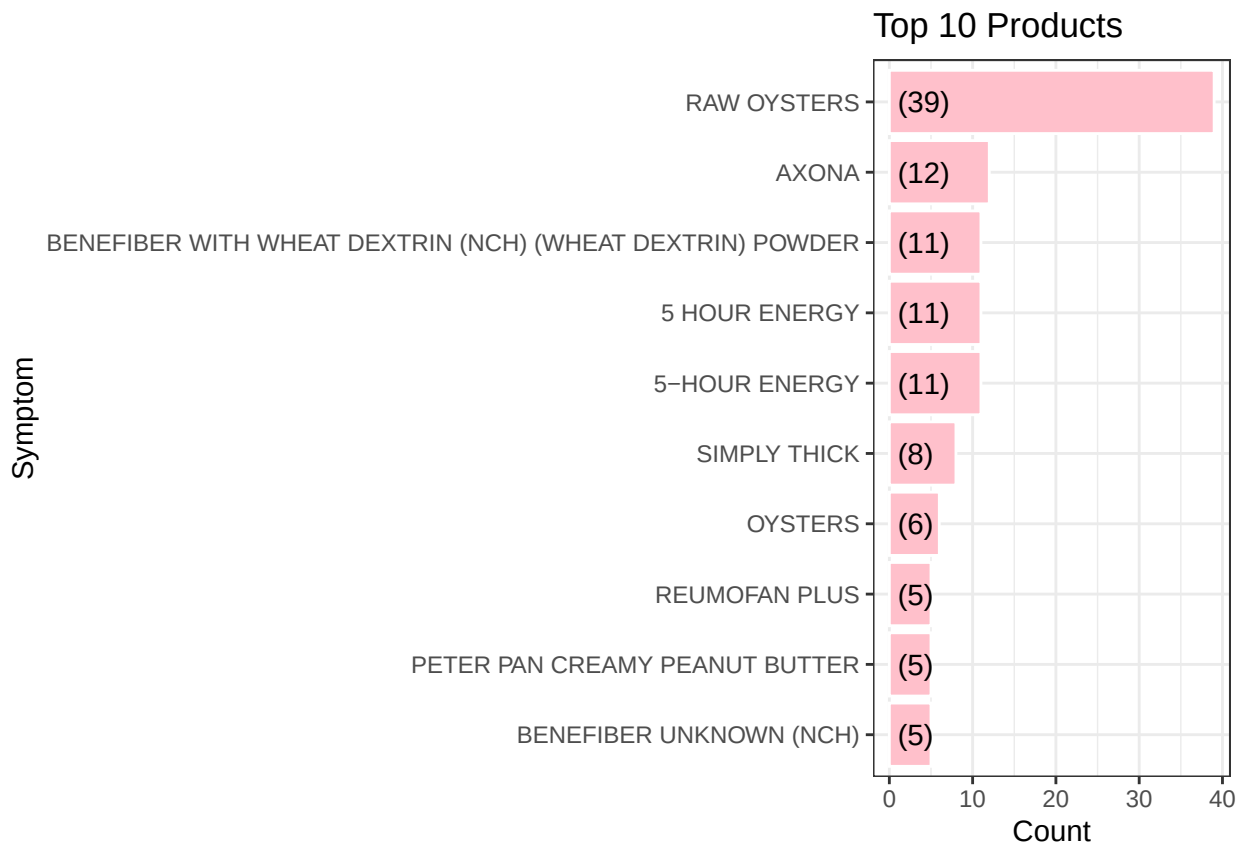
```
food_v1 %>%
  mutate(age_at_adverse_event = as.numeric(age_at_adverse_event)) %>%
    filter(age_at_adverse_event <60) %>%
  filter(age_at_adverse_event >=50) %>%
  filter(product_name != 'REDACTED') %>%
  bpproducts()
```



Products that led to serious adverse reactions

Products which had death as the adverse reaction

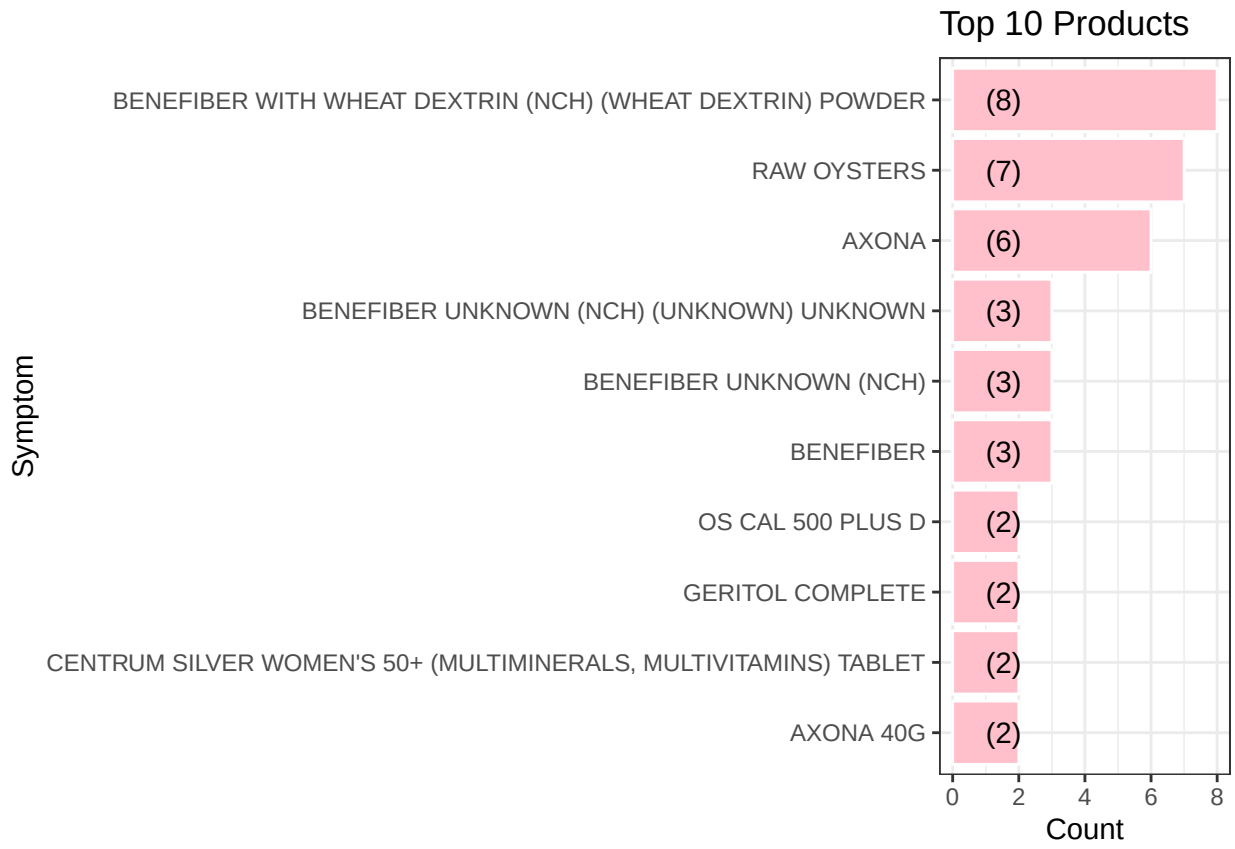
```
targetoutcome = 'DEATH'

death_food  = food_v1 %>%
  filter(str_detect(outcomes,targetoutcome))

death_food %>%
  filter(product_name != 'REDACTED') %>%
  bpproducts()
```



What a pity, I do like a good half dozen of raw oysters every now and again!

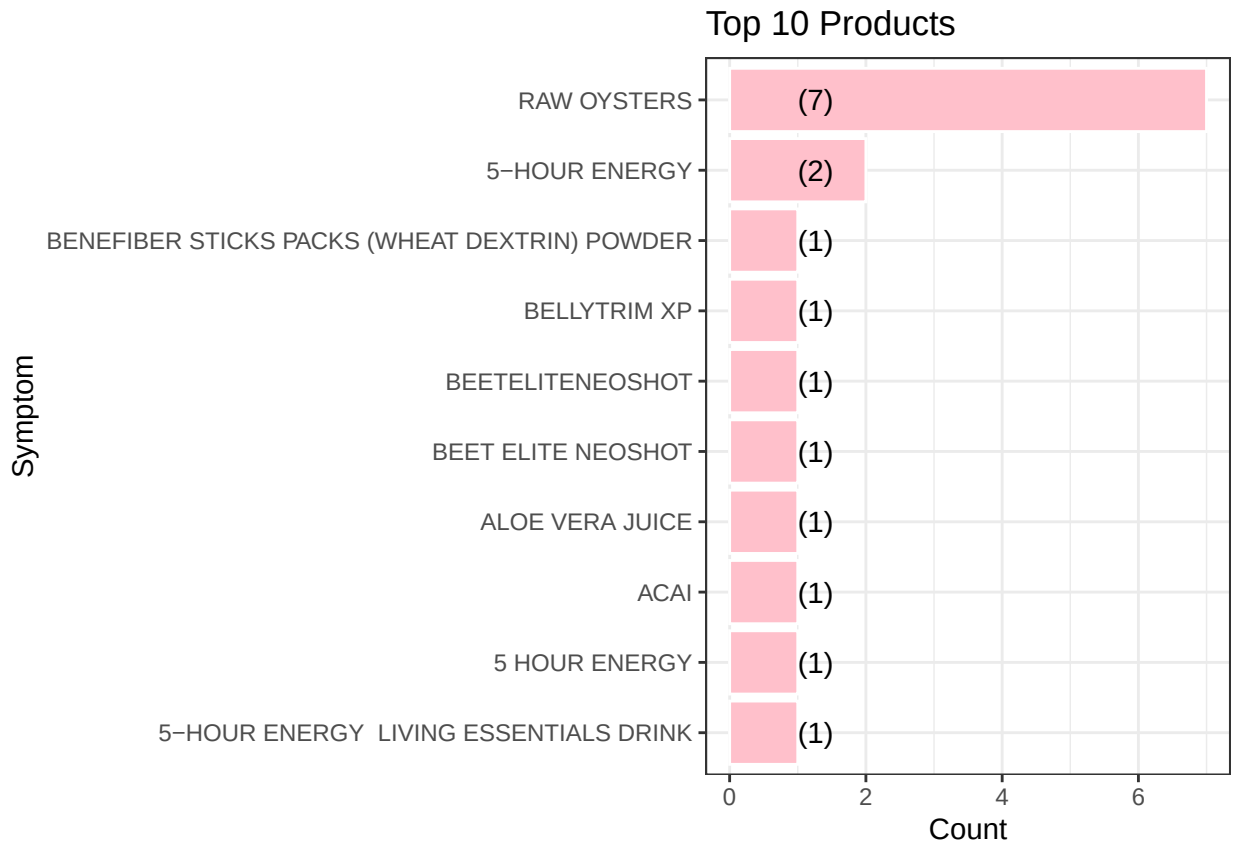
Products which had death as the adverse reaction for those that were 70 or older

```
targetoutcome = 'DEATH'

death_food  = food_v1 %>%
  filter(str_detect(outcomes,targetoutcome))

death_food %>%
  mutate(age_at_adverse_event = as.numeric(age_at_adverse_event)) %>%
    filter(age_at_adverse_event >=70) %>%
  filter(product_name != 'REDACTED') %>%
  bpproducts()
```



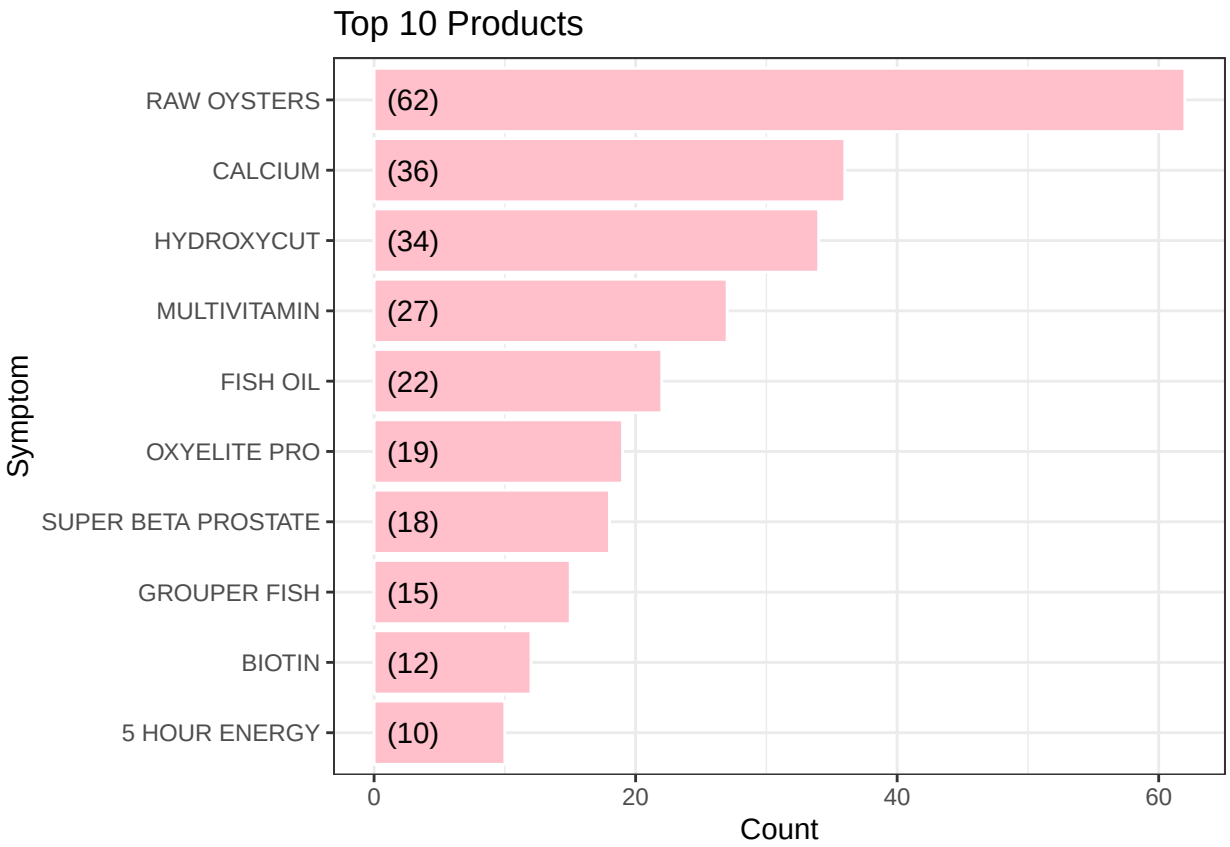
Products which had death as the adverse reaction for those that were 60 to 69

```
targetoutcome = 'DEATH'

death_food  = food_v1 %>%
  filter(str_detect(outcomes,targetoutcome))

death_food %>%
  mutate(age_at_adverse_event = as.numeric(age_at_adverse_event)) %>%
    filter(age_at_adverse_event <70) %>%
  filter(age_at_adverse_event >=60) %>%
  filter(product_name != 'REDACTED') %>%
  bpproducts()
```



Products that led to a life threatening reaction

```
targetoutcome = 'LIFE THREATENING'

death_food  = food_v1 %>%
  filter(str_detect(outcomes,targetoutcome))

death_food %>%
  filter(product_name != 'REDACTED') %>%
  bpproducts()
```



Again, raw oysters are the top product that led to a life threatening or death results.