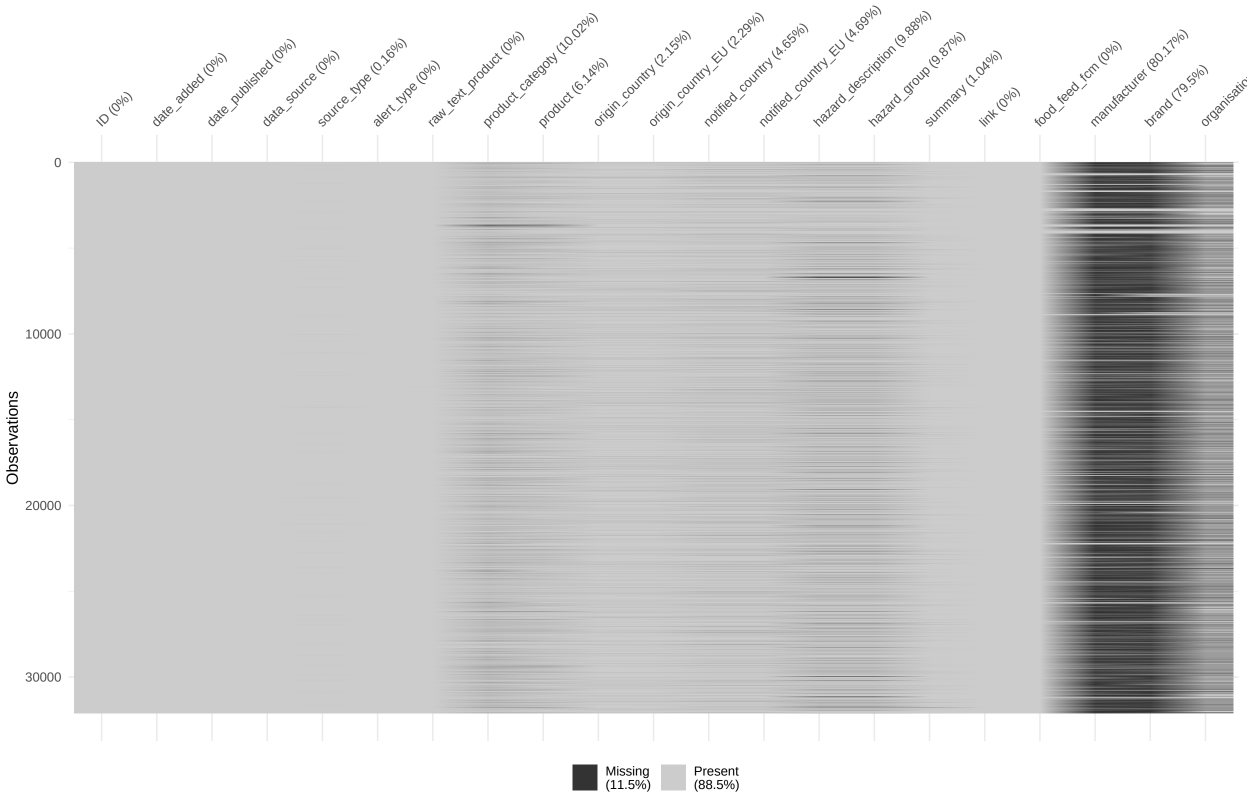
Food hazard data EDA

Marina Vabistsevits

20/07/2020

```
raw.dta<-read_csv("../data/FSA_data_competition_2020.csv", na = c("NA", ":", "unclassified", "unknown")
         rename("ID" = "ID",
                "date_added" = "Date Added",
                "date_published" = "Date of Publishing",
                "data_source" = "Data Source",
                "source_type" = "Source Type",
                "alert_type" = "Alert Type",
                "raw_text_product" = "Raw Product Phrase",
                "product_categoty" = "Product Category",
                "product" = "Commodity / Product",
                "origin_country" = "Country of Origin",
                "origin_country_EU" = "Eu/non-EU Country of Origin",
                "notified_country" = "Notified by",
                "notified_country_EU" = "EU/non-EU Notifying Country",
                "incident_title" = "Incident Title",
                "hazard_description" = "Hazard Description", # can extract about ecoli fro here
                "hazard_group" = "Hazard Group",
                "summary" = "Summary",
                "link" = "Link",
                "food_feed_fcm" = "Food; Feed or FCM",
                "manufacturer" = "Manufacturer",
                "brand" = "Brand",
                "organisation" = "Organisations",
                "food_or_not" = "Is A Food Article" )

# basic tidy

data <- raw.dta %>%
        select(-food_or_not, -incident_title) %>%
        mutate(food_feed_fcm = ifelse(food_feed_fcm == 'FCM', 'fcm',
                               ifelse(food_feed_fcm =='Food', 'food', food_feed_fcm))) %>%
        filter(food_feed_fcm != "fcm")
```

```
data %>% arrange(date_published) %>% vis_miss()
```



```
# new cols
data <- data %>%
        # tidy up dates using lubridate
        mutate(date_added = dmy(date_added),
                date_published = dmy(date_published)) %>%
        mutate(date_added_year = year(date_added),
               date_published_year = year(date_published)) %>%
        mutate(date_published_month = ifelse(nchar(month(date_published)) == 2, month(date_published), p
               # create year_month
               date_published_year_month = paste0(date_published_year, "-", date_published_month),
               #create year_quarter
               date_published_quarter = ifelse(date_published_month %in% c("01", "02", "03"), "Q1",
                                        ifelse(date_published_month %in% c("04", "05", "06"), "Q2",
                                        ifelse(date_published_month %in% c("07", "08", "09"), "Q3",
                                        ifelse(date_published_month %in% c("10", "11", "12"), "Q4", NA)))
               date_published_year_quarter = paste0(date_published_year, "-", date_published_quarter) )
        # create incident ID
        separate(ID, into= c("ID", "ID_incident"), sep= "-", remove=F) %>%
        mutate(ID_incident = ifelse(is.na(ID_incident), ID, ID_incident))

data %>% count(date_added_year)
```

```
## # A tibble: 2 x 2
##   date_added_year     n
##             <dbl> <int>
## 1            2019 21876
## 2            2020 10244
```

```
data %>% count(date_published_year)
```

```
## # A tibble: 6 x 2
##   date_published_year     n
##                 <dbl> <int>
## 1                2016  4306
## 2                2017  6497
## 3                2018  8228
## 4                2019 10149
## 5                2020  2937
## 6                  NA     3
```

```
data %>% count(product) %>% arrange(-n)
```

```
## # A tibble: 3,490 x 2
##    product              n
##    <chr>            <int>
##  1 <NA>              1973
##  2 chicken           1951
##  3 bakery product    1935
##  4 beef              1681
##  5 meat product       995
##  6 pepper             816
##  7 food supplement    720
##  8 cheese             660
##  9 pork               644
## 10 sesame             540
## # ... with 3,480 more rows
```

```
data %>% count(alert_type) %>% arrange(-n)
```

```
## # A tibble: 10 x 2
##    alert_type                     n
##    <chr>                      <int>
##  1 recall                     14064
##  2 border rejection            5504
##  3 alert                       3649
##  4 information for attention   2692
##  5 update                      2082
##  6 information for follow-up   1941
##  7 outbreak                    1863
##  8 warning                      182
##  9 information                  109
## 10 lookout                       34
```

```
data %>% count(data_source) %>% arrange(-n)
```

```
## # A tibble: 43 x 2
##    data_source                                     n
##    <chr>                                       <int>
##  1 RASFF Portal                                11906
##  2 CDPH Recalls (Canada)                        5075
##  3 Food Poisoning Bulletin (US)                 2133
##  4 Ministry of Health - Border Rejections (Japan)  1850
##  5 FoodSafetyNews.com                           1495
##  6 FSA Alerts & Recalls (UK)                    1427
##  7 MAPAQ (Canada)                                 1230
##  8 FDA Recalls (USA)                               904
##  9 Product Recalls Website: Oulah (France)         896
## 10 AFSCA Recalls (Belgium)                         614
## # ... with 33 more rows
```

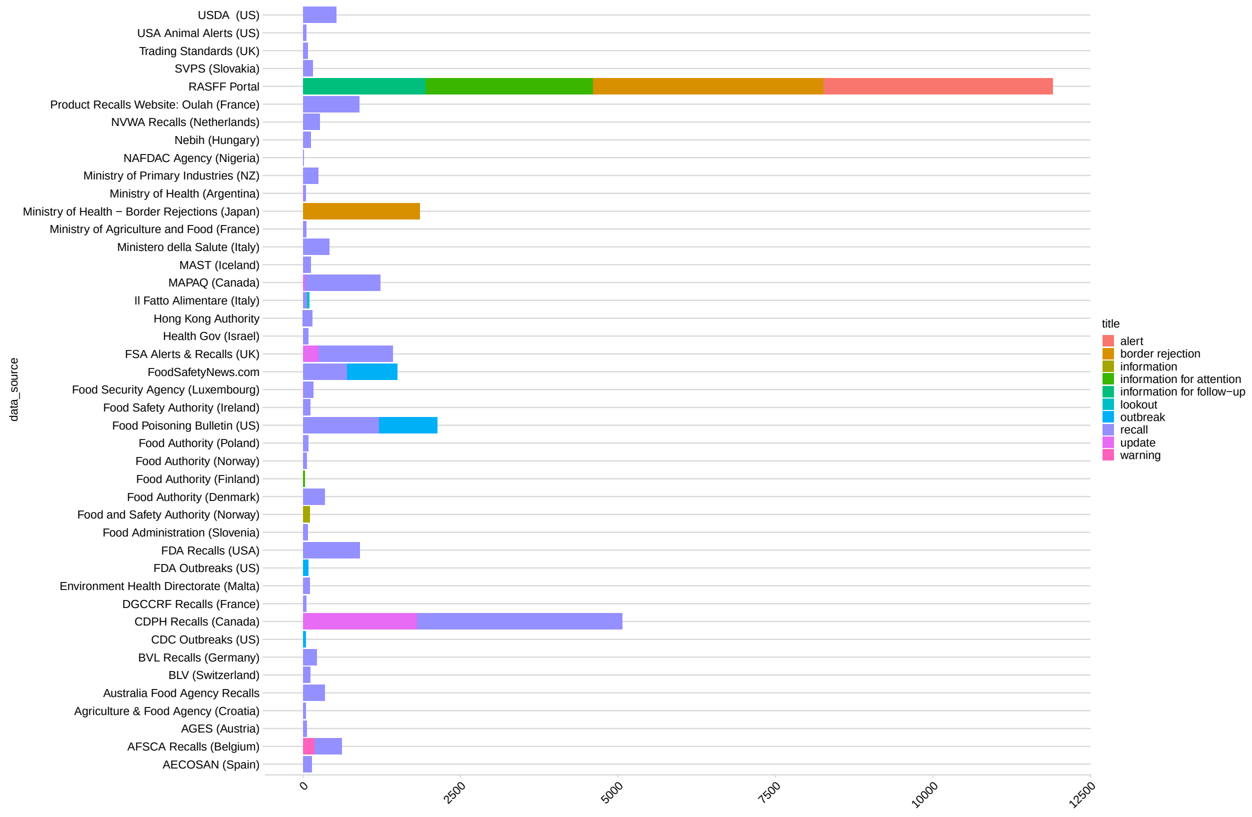
```
data %>% count(hazard_group) %>% arrange(-n)
```

```
## # A tibble: 35 x 2
##    hazard_group                       n
##    <chr>                          <int>
##  1 microbial contaminants (other)  5831
##  2 pathogenic micro-organisms      5588
##  3 allergens                       5221
##  4 <NA>                            3171
##  5 foreign bodies                  1896
##  6 composition                     1536
##  7 poor or insufficient controls   1192
##  8 pesticide residues              1149
##  9 heavy metals                     869
## 10 Fraud                            783
## # ... with 25 more rows
```

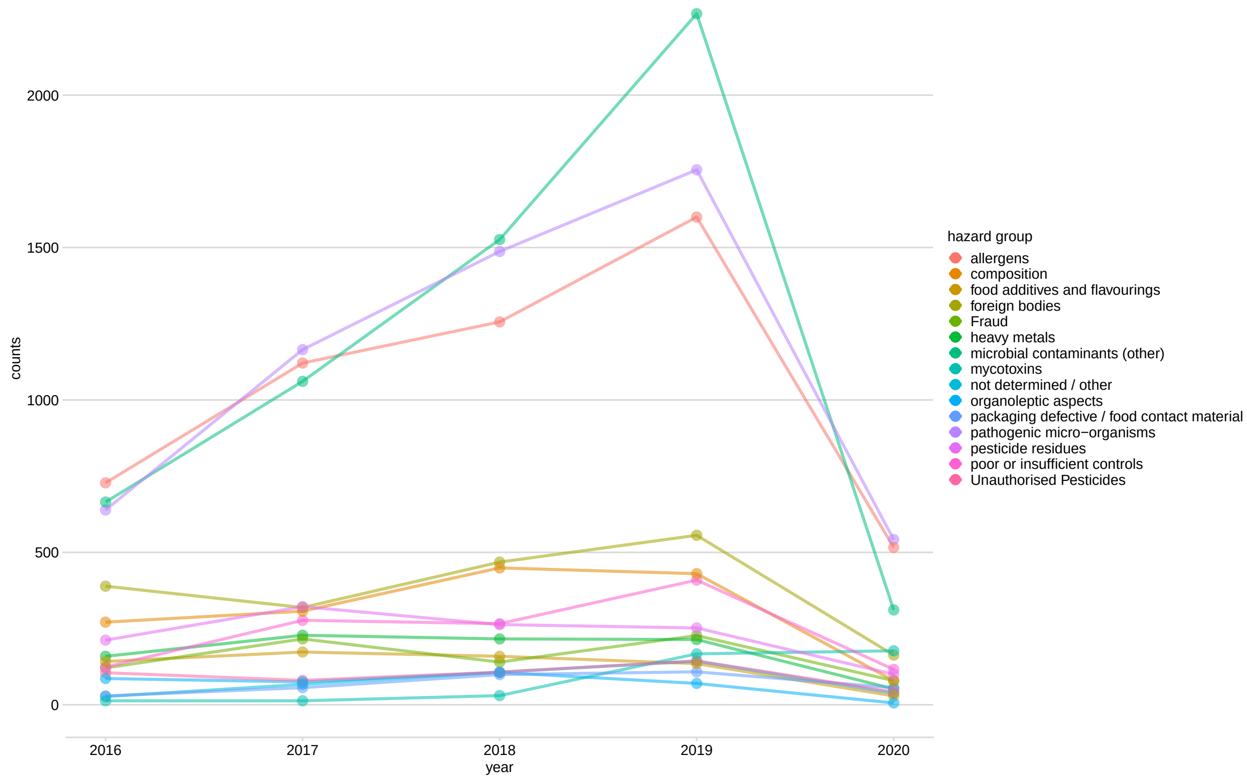
```
data %>% count(origin_country) %>% arrange(-n)
```

```
## # A tibble: 151 x 2
##    origin_country     n
##    <chr>          <int>
##  1 Canada          6508
##  2 USA             4763
##  3 United Kingdom  1966
##  4 France          1823
##  5 Italy           1130
##  6 China           1040
##  7 Belgium         1018
##  8 Poland           885
##  9 Netherlands      840
## 10 Turkey           826
## # ... with 141 more rows
```

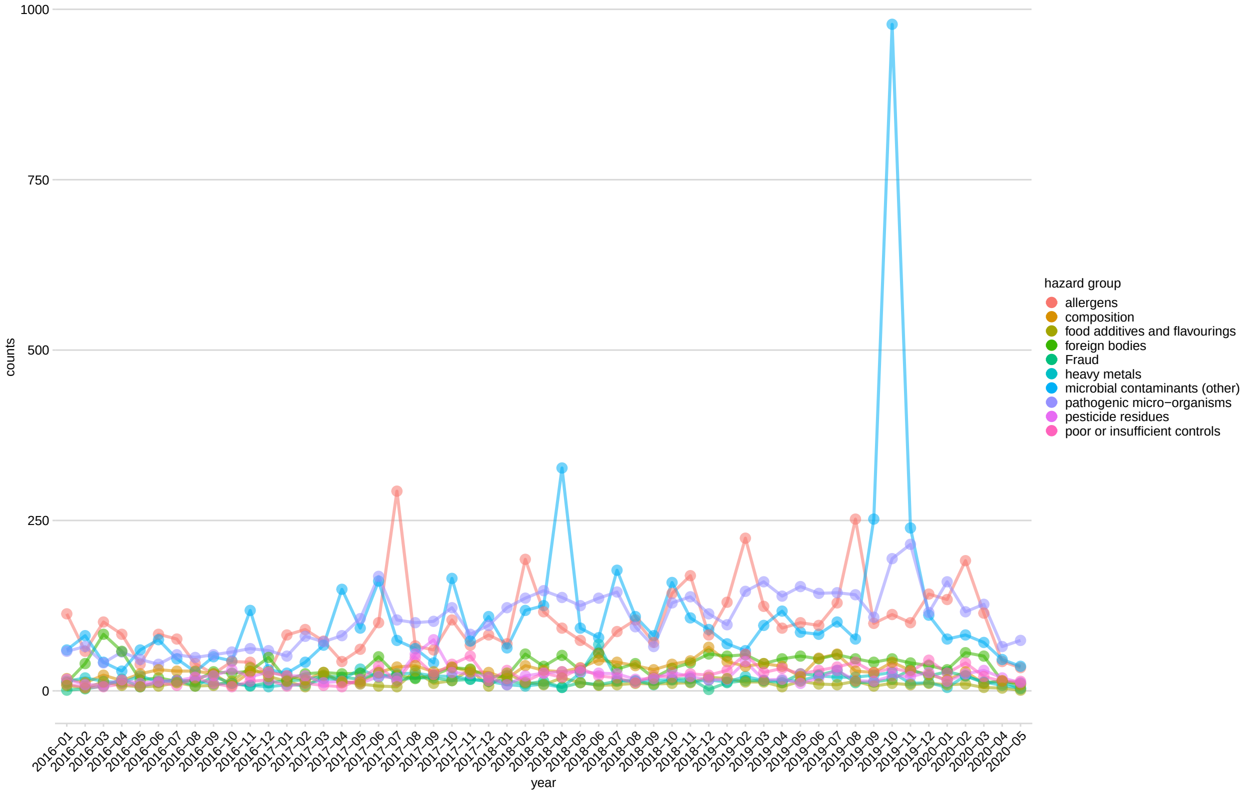
```
ggplot(data, aes(x = data_source, fill = alert_type)) +
  geom_bar()+
  theme_minimal_hgrid(10, rel_small = 1)+
  #facet_grid(~alert_type)+
  #scale_fill_manual(values=pal)+
  coord_flip()+
  labs(fill = "title", y="") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



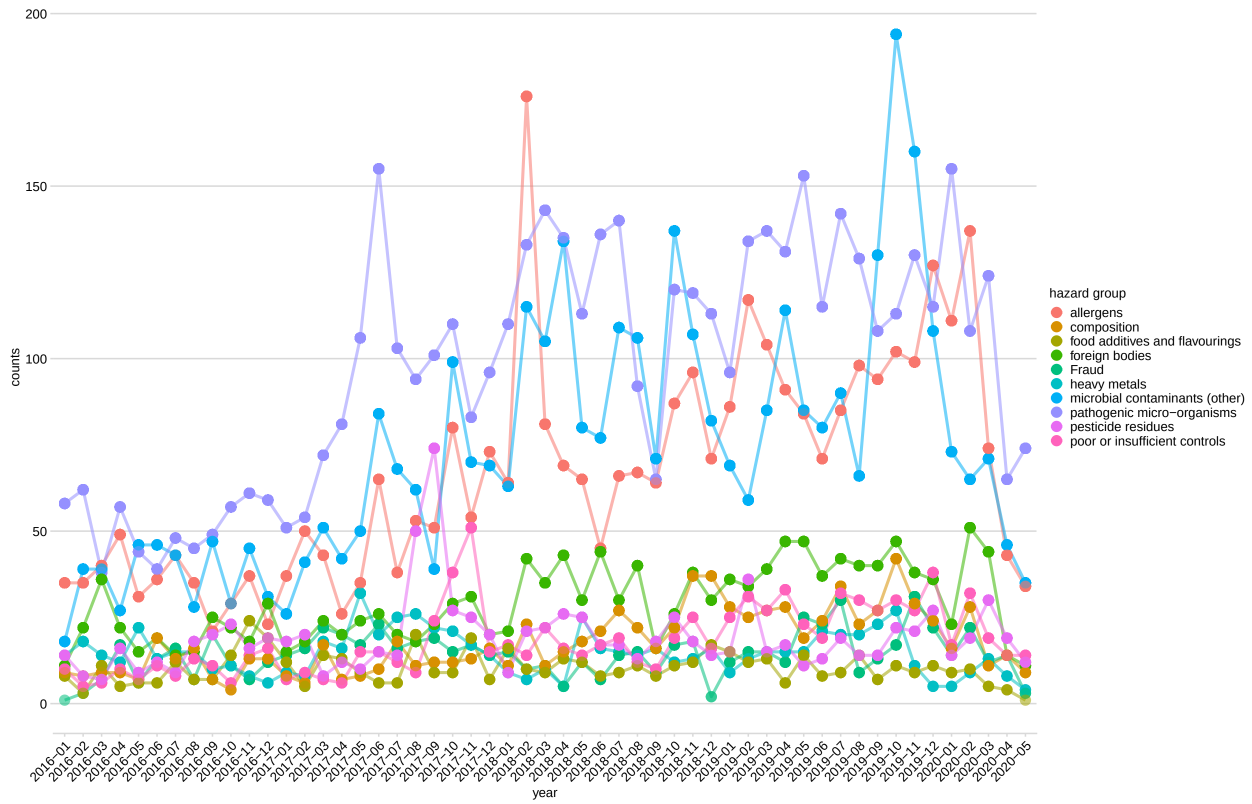
```
# Hazard categoty overview by Year
dat1<-data %>%
  filter(!is.na(hazard_group),
         date_published_year_month != "NA-NA") %>%
  group_by(hazard_group) %>%
  filter(n()>300) %>%
  ungroup %>%
  select(hazard_group, date_published_year )  %>%
  group_by(hazard_group, date_published_year) %>%
  mutate(count = n())%>%
  distinct()

pc <-ggplot(data = dat1, aes(x = date_published_year  , y = count, group = hazard_group)) +
  geom_line(aes(color = hazard_group, alpha = 1), size = 1) +
  geom_point(aes(color = hazard_group, alpha = 1), size = 3) +
  #scale_x_continuous(breaks = sort(unique(dat$date_published_year))[c(TRUE, FALSE)]  )+
  theme(legend.position = "right", ncol=1) +
  #scale_colour_manual(values=c(pal))+
  theme_minimal_hgrid(10, rel_small = 1) +
  labs(x = "year",  colour="hazard group",
       y = "counts",
       title = "")+
    guides(alpha = FALSE)
pc
```



```
# Hazard categoty overview by Month (all records)

dat2<-data %>%
  filter(!is.na(hazard_group),
         date_published_year_month != "NA-NA") %>%
  group_by(hazard_group) %>%
  filter(n()>500) %>%
  ungroup %>%
  select(hazard_group, date_published_year_month )  %>%
  group_by(hazard_group, date_published_year_month) %>%
  mutate(count = n())%>%
  distinct()

pc <-ggplot(data = dat2, aes(x = date_published_year_month  , y = count, group = hazard_group)) +
  geom_line(aes(color = hazard_group, alpha = 1), size = 1) +
  geom_point(aes(color = hazard_group, alpha = 1), size = 3) +
  #scale_x_continuous(breaks = sort(unique(dat$date_published_year))[c(TRUE, FALSE)]  )+
  theme(legend.position = "right", ncol=1) +
  #scale_colour_manual(values=c(pal))+
  theme_minimal_hgrid(9, rel_small = 1) +
  labs(x = "year",  colour="hazard group",
       y = "counts",
       title = "")+
    guides(alpha = FALSE) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
pc
```

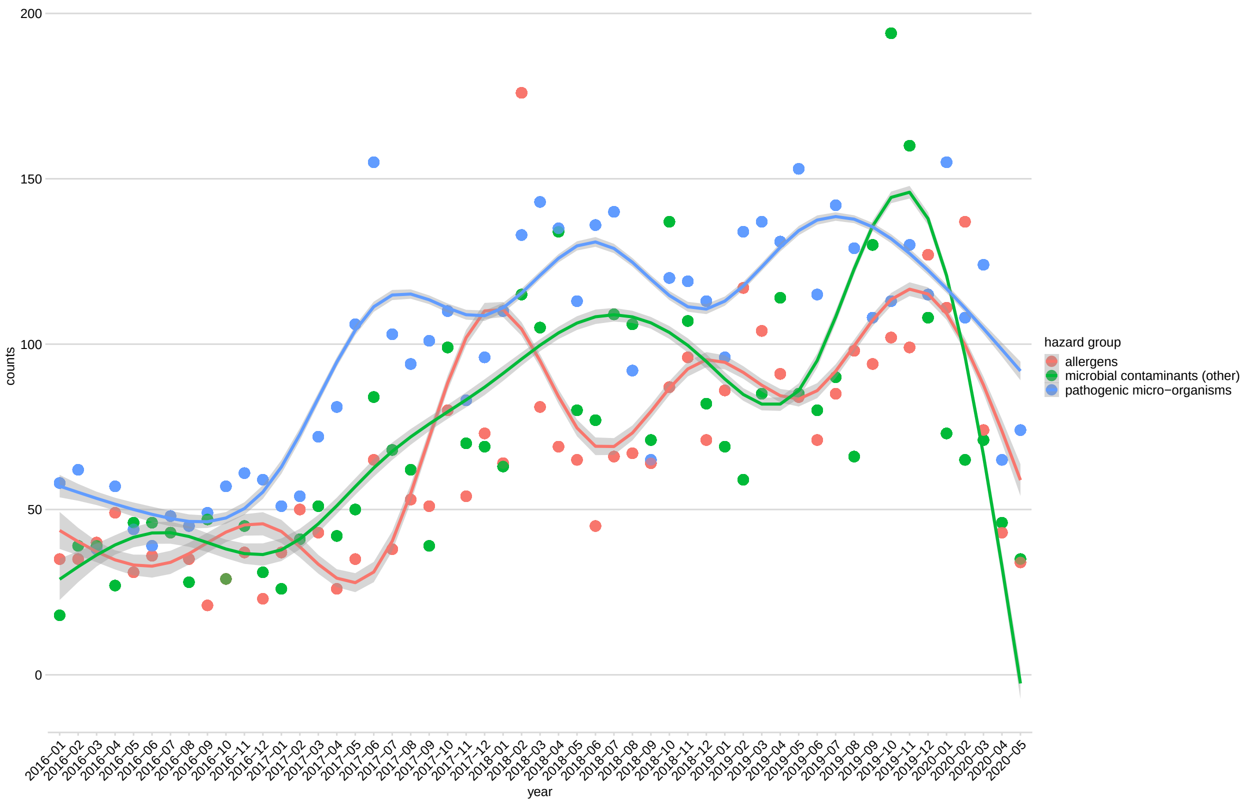


```
# Hazard categoty overview by Month (report multiple issue related to one event as one event)

dat3<-data %>%
  filter(!is.na(hazard_group),
         date_published_year_month != "NA-NA") %>%
  group_by(hazard_group) %>%
  filter(n()>500) %>%
  ungroup %>%
  select(hazard_group, date_published_year_month, ID_incident)  %>% distinct() %>%
  group_by(hazard_group, date_published_year_month) %>%
  mutate(count = n())%>%
  distinct()

pc3 <-ggplot(data = dat3, aes(x = date_published_year_month  , y = count, group = hazard_group)) +
  geom_line(aes(color = hazard_group, alpha = 1), size = 1) +
  geom_point(aes(color = hazard_group, alpha = 1), size = 3) +
  #scale_x_continuous(breaks = sort(unique(dat$date_published_year))[c(TRUE, FALSE)]  )+
  theme(legend.position = "right", ncol=1) +
  #scale_colour_manual(values=c(pal))+
  theme_minimal_hgrid(9, rel_small = 1) +
  labs(x = "year",  colour="hazard group",
       y = "counts",
       title = "")+
    guides(alpha = FALSE) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

pc3
```

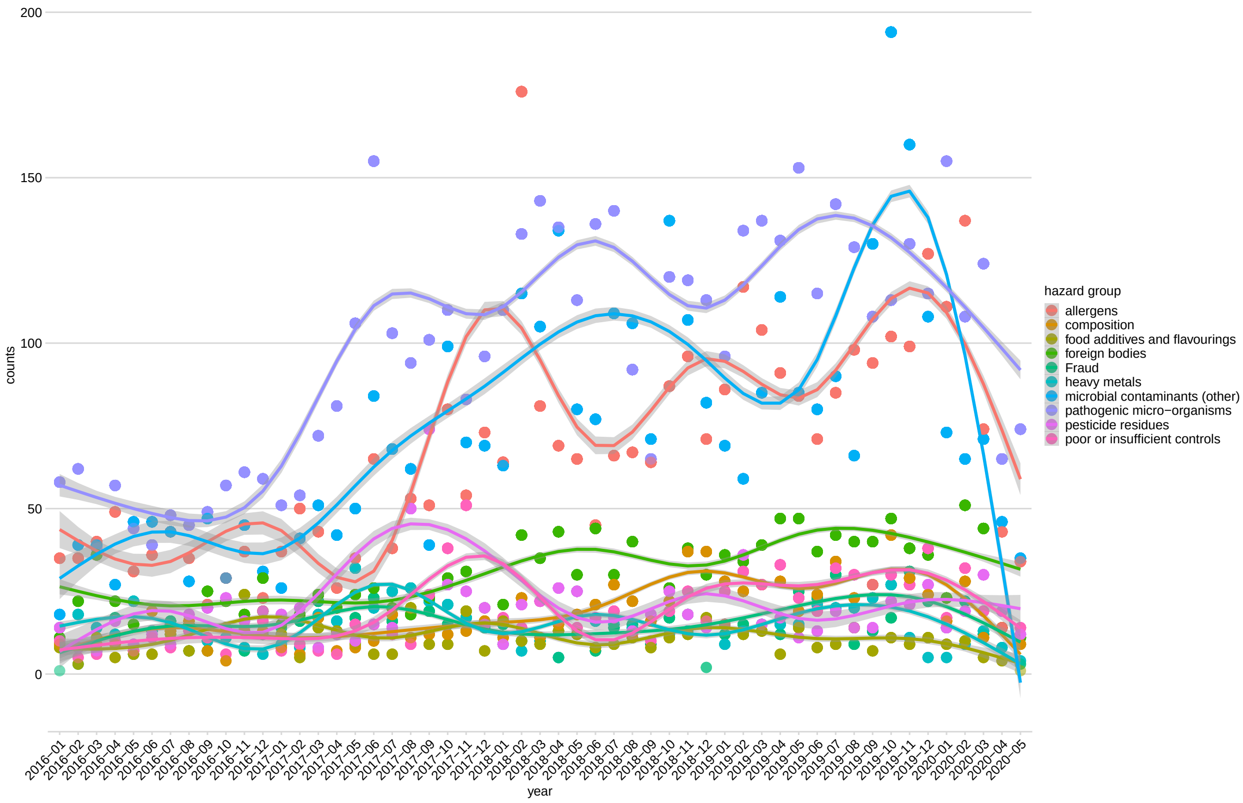


```
# Main 3 categories, with smooth lines

dat4<-data %>%
  filter(!is.na(hazard_group),
         date_published_year_month != "NA-NA") %>%
  group_by(hazard_group) %>%
  #filter(n()>500) %>%
  ungroup %>%
  select(hazard_group, date_published_year_month, ID_incident)  %>% distinct() %>%
  group_by(hazard_group, date_published_year_month) %>%
  filter(hazard_group %in% c("microbial contaminants (other)", "pathogenic micro-organisms", "allergens
  mutate(count = n())%>%
  distinct()

pc4 <-ggplot(data = dat4, aes(x = date_published_year_month  , y = count, group = hazard_group)) +
  #geom_line(aes(color = hazard_group, alpha = 1), size = 1) +
  geom_point(aes(color = hazard_group, alpha = 1), size = 3) +
  geom_smooth(aes(x = date_published_year_month  , y = count, color = hazard_group))+
  #scale_x_continuous(breaks = sort(unique(dat$date_published_year))[c(TRUE, FALSE)]  )+
  theme(legend.position = "right", ncol=1) +
  #scale_colour_manual(values=c(pal))+
  theme_minimal_hgrid(9, rel_small = 1) +
  labs(x = "year",  colour="hazard group",
       y = "counts",
       title = "")+
    guides(alpha = FALSE) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
pc4
```



```
# + other categories smooth line

dat5<-data %>%
  filter(!is.na(hazard_group),
         date_published_year_month != "NA-NA") %>%
  group_by(hazard_group) %>%
  filter(n()>500) %>%
  ungroup %>%
  select(hazard_group, date_published_year_month, ID_incident)  %>% distinct() %>%
  group_by(hazard_group, date_published_year_month) %>%
  #filter(hazard_group %in% c("microbial contaminants (other)", "pathogenic micro-organisms", "allergen
  mutate(count = n())%>%
  distinct()

pc5 <-ggplot(data = dat5, aes(x = date_published_year_month  , y = count, group = hazard_group)) +
  #geom_line(aes(color = hazard_group, alpha = 1), size = 1) +
  geom_point(aes(color = hazard_group, alpha = 1), size = 3) +
  geom_smooth(aes(x = date_published_year_month  , y = count, color = hazard_group))+
  #scale_x_continuous(breaks = sort(unique(dat$date_published_year))[c(TRUE, FALSE)]  )+
  theme(legend.position = "right", ncol=1) +
  #scale_colour_manual(values=c(pal))+
  theme_minimal_hgrid(9, rel_small = 1) +
  labs(x = "year",  colour="hazard group",
       y = "counts",
       title = "")+
    guides(alpha = FALSE) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
pc5
```

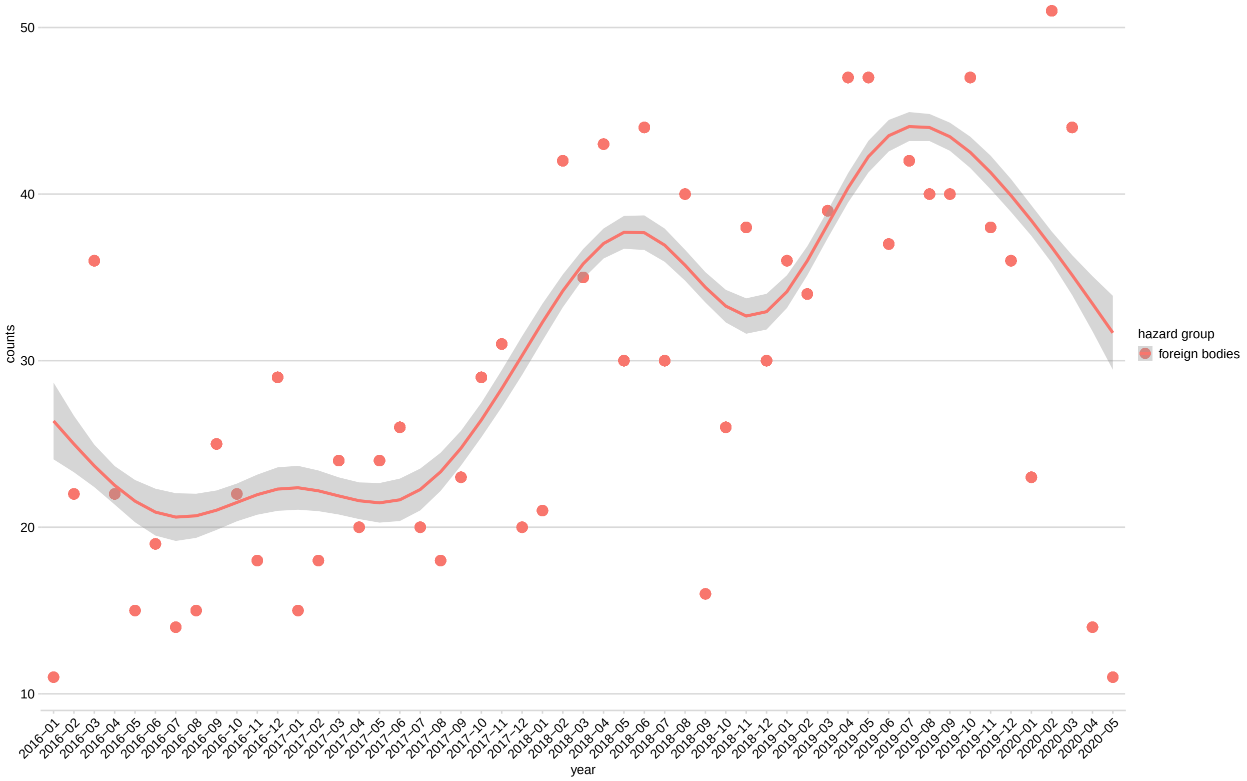


Foreign bodies - e.g. exploring plastic pollution

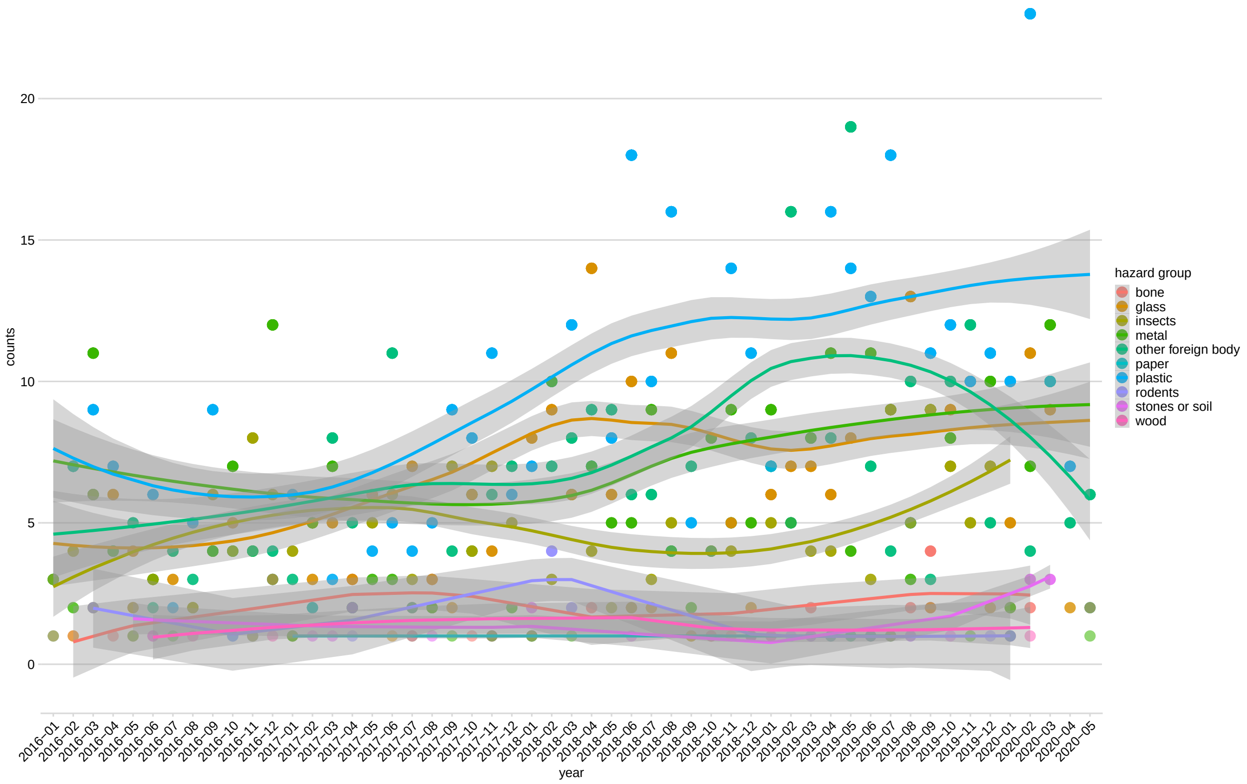
```
foreign <- data %>%
  filter(hazard_group == "foreign bodies") %>%
  mutate(contaminant = ifelse(str_detect(hazard_description, "(?i)plastic"), "plastic",
                       ifelse(str_detect( raw_text_product, "(?i)plastic"), "plastic",
                       ifelse(str_detect(hazard_description, "(?i)polystyrene"), "plastic",
                       ifelse(str_detect(hazard_description, "(?i)film"), "plastic",
                       ifelse(str_detect(hazard_description, "(?i)nylon"), "plastic",
                       ifelse(str_detect(raw_text_product, "(?i)rubber"), "plastic",
                       ifelse(str_detect(raw_text_product, "(?i)conveyor belt"), "plastic",
                       ifelse(str_detect(raw_text_product, "(?i)blue particles"), "plastic",
                       ifelse(str_detect(raw_text_product, "(?i)white and blue"), "plastic",
                       ifelse(str_detect(raw_text_product, "(?i)packaging tape"), "plastic",

                       ifelse(str_detect(hazard_description, "(?i)glass"), "glass",
                       ifelse(str_detect( raw_text_product, "(?i)glass"), "glass",

                       ifelse(str_detect(hazard_description, "(?i)metal"), "metal",
                       ifelse(str_detect( raw_text_product, "(?i)metal"), "metal",
                       ifelse(str_detect( raw_text_product, "(?i)blade"), "metal",
```

```r
                        ifelse(str_detect( raw_text_product, "(?i)aluminum"), "metal",
                        ifelse(str_detect( raw_text_product, "(?i)iron"), "metal",
                        ifelse(str_detect( raw_text_product, "(?i)sharp"), "metal",
                        ifelse(str_detect( raw_text_product, "(?i)needle"), "metal",
                        ifelse(str_detect(hazard_description, "(?i)wires"), "metal",
                        ifelse(str_detect(hazard_description, "(?i)nails"), "metal",
                        ifelse(str_detect(summary, "(?i)foil"), "metal",

                        ifelse(str_detect(hazard_description, "(?i)insect"), "insects",
                        ifelse(str_detect(raw_text_product, "(?i)insect"), "insects",

                        ifelse(str_detect(hazard_description, "(?i)bone"), "bone",
                        ifelse(str_detect( raw_text_product, "(?i)bone"), "bone",

                        ifelse(str_detect(hazard_description, "(?i)wood"), "wood",
                        ifelse(str_detect(raw_text_product, "(?i)wood"), "wood",

                        ifelse(str_detect(raw_text_product, "(?i)paper"), "paper",
                        ifelse(str_detect(raw_text_product, "(?i)carton"), "paper",
                        ifelse(str_detect(raw_text_product, "(?i)cardboard"), "paper",

                        ifelse(str_detect(raw_text_product, "(?i)soil"), "stones or soil",
                        ifelse(str_detect(raw_text_product, "(?i)sand "), "stones or soil",
                        ifelse(str_detect(raw_text_product, "(?i)pebble"), "stones or soil",
                        ifelse(str_detect(raw_text_product, "(?i)stone"), "stones or soil",
                        ifelse(str_detect(raw_text_product, "(?i)gravel"), "stones or soil",
                        ifelse(str_detect(raw_text_product, "(?i)rock"), "stones or soil",

                        ifelse(str_detect(raw_text_product, "(?i)mice"), "rodents",
                        ifelse(str_detect(raw_text_product, "(?i)mouse"), "rodents",
                        ifelse(str_detect(raw_text_product, "(?i)rodent"), "rodents",

                        "other foreign body")))))))))))))))))))))))))))))))))))))))))
```

```r
# all foreign bodies

dat7<-data %>%
  filter(!is.na(hazard_group),
         date_published_year_month != "NA-NA") %>%
  group_by(hazard_group) %>%
  filter(n()>500) %>%
  ungroup %>%
  select(hazard_group, date_published_year_month, ID_incident) %>% distinct() %>%
  group_by(hazard_group, date_published_year_month) %>%
  filter(hazard_group == "foreign bodies") %>%
  mutate(count = n())%>%
  distinct()

pc7 <-ggplot(data = dat7, aes(x = date_published_year_month  , y = count, group = hazard_group)) +
  #geom_line(aes(color = hazard_group, alpha = 1), size = 1) +
  geom_point(aes(color = hazard_group, alpha = 1), size = 3) +
  geom_smooth(aes(x = date_published_year_month  , y = count, color = hazard_group))+
  #scale_x_continuous(breaks = sort(unique(dat$date_published_year))[c(TRUE, FALSE)]  )+
  theme(legend.position = "right", ncol=1) +
  #scale_colour_manual(values=c(pal))+
  theme_minimal_hgrid(9, rel_small = 1) +
  labs(x = "year",  colour="hazard group",
       y = "counts",
       title = "")+
    guides(alpha = FALSE) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

pc7
```

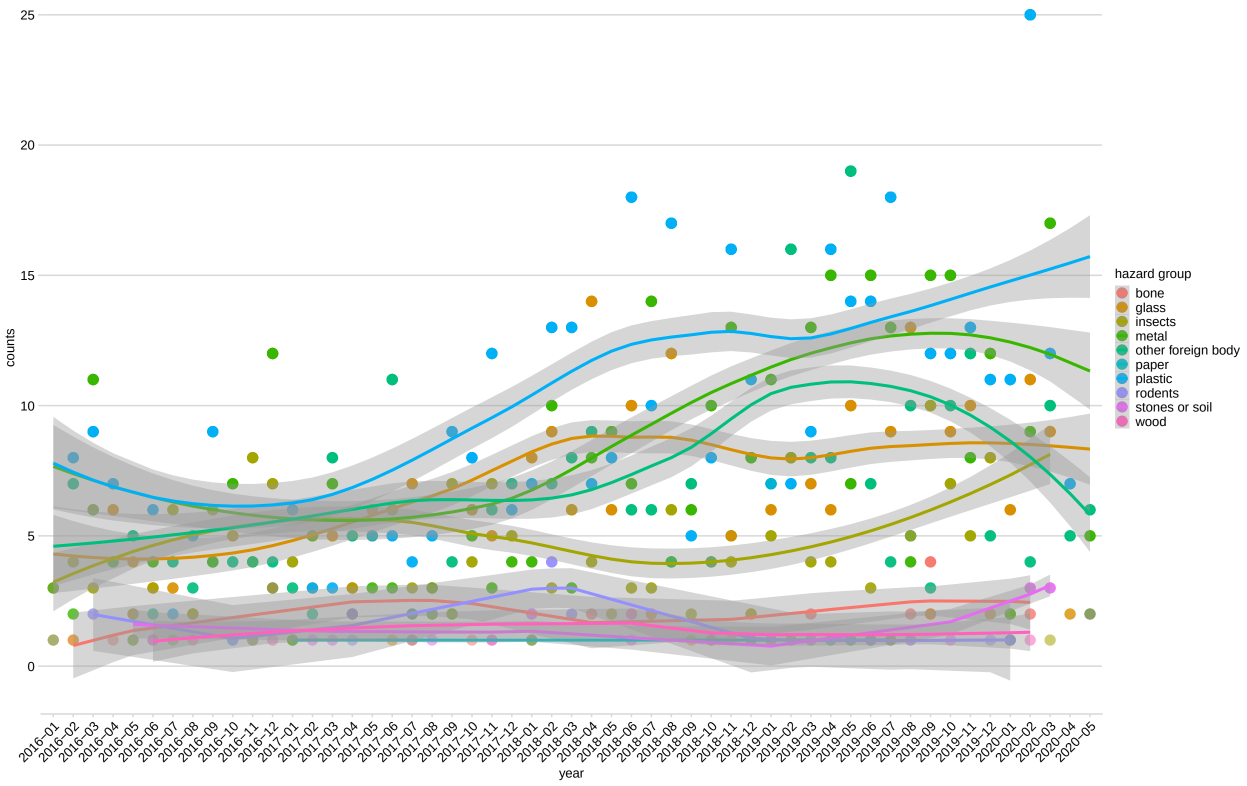


```
# by contaminant group

dat7<-foreign %>%
  filter(date_published_year_month != "NA-NA") %>%
  group_by(contaminant) %>%
  #filter(n()>500) %>%
  ungroup %>%
  select(contaminant, date_published_year_month, ID_incident) %>% distinct() %>%
  group_by(contaminant, date_published_year_month) %>%
  mutate(count = n())%>%
  distinct()

pc7 <-ggplot(data = dat7, aes(x = date_published_year_month  , y = count, group = contaminant)) +
  #geom_line(aes(color = contaminant, alpha = 1), size = 1) +
  geom_point(aes(color = contaminant, alpha = 1), size = 3) +
  geom_smooth(aes(x = date_published_year_month  , y = count, color = contaminant))+
  #scale_x_continuous(breaks = sort(unique(dat$date_published_year))[c(TRUE, FALSE)]  )+
  theme(legend.position = "right", ncol=1) +
  #scale_colour_manual(values=c(pal))+
  theme_minimal_hgrid(9, rel_small = 1) +
  labs(x = "year",  colour="hazard group",
       y = "counts",
       title = "")+
    guides(alpha = FALSE) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

pc7
```



```
## try to get contaminant from other groups

foreign_extra <- data %>%
  filter(hazard_group != "foreign bodies") %>%
  mutate(contaminant = ifelse(str_detect(hazard_description, "(?i)plastic"), "plastic",
                       ifelse(str_detect( raw_text_product, "(?i)plastic"), "plastic",
                       ifelse(str_detect(hazard_description, "(?i)polystyrene"), "plastic",
                       ifelse(str_detect(hazard_description, "(?i)film"), "plastic",
                       ifelse(str_detect(hazard_description, "(?i)nylon"), "plastic",
                       ifelse(str_detect(raw_text_product, "(?i)rubber"), "plastic",
                       ifelse(str_detect(raw_text_product, "(?i)conveyor belt"), "plastic",
                       ifelse(str_detect(raw_text_product, "(?i)blue particles"), "plastic",
                       ifelse(str_detect(raw_text_product, "(?i)white and blue"), "plastic",
                       ifelse(str_detect(raw_text_product, "(?i)packaging tape"), "plastic",

                       ifelse(str_detect(hazard_description, "(?i)metal"), "metal",
                       ifelse(str_detect( raw_text_product, "(?i)metal"), "metal",
                       ifelse(str_detect( raw_text_product, "(?i)blade"), "metal",
                       ifelse(str_detect( raw_text_product, "(?i)aluminum"), "metal",
                       ifelse(str_detect( raw_text_product, "(?i) iron"), "metal",
                       ifelse(str_detect( raw_text_product, "(?i)sharp"), "metal",
                       ifelse(str_detect( raw_text_product, "(?i)needle"), "metal",
                       ifelse(str_detect(hazard_description, "(?i)wires"), "metal",
                       ifelse(str_detect(hazard_description, "(?i)nails"), "metal",
                       ifelse(str_detect(summary, "(?i)foil"), "metal",

                       ifelse(str_detect(hazard_description, "(?i)glass"), "glass",
                       ifelse(str_detect( raw_text_product, "(?i)glass"), "glass",

                       ifelse(str_detect(hazard_description, "(?i)insect"), "insects",
                       ifelse(str_detect(raw_text_product, "(?i)insect"), "insects",


                       "other foreign body")))))))))))))))))))))))) %>%
  filter(contaminant != "other foreign body") %>%
  filter(!(contaminant == "metal" & grepl("heavy metal|metallic|Lead|alkaloid|metalaxyl", raw_text_produ
  filter(!(contaminant == "metal" & grepl("heavy metal|metallic|Lead|alkaloid", summary, ignore.case = T
  filter(!(contaminant == "metal" & hazard_description %in% c("undeclared peanut", "unauthorised substa
                                      "unauthorised substance iron glycinate chelate", "too high conten
                                      "salmonella outbreak" ,"salmonella spp sticks", "copper", "undesig
  filter(!(contaminant %in% c("plastic", "glass") & !grepl("piece|foreign|extraneous|find|bits|particle
```

```
# add all contaminants

dat8<-bind_rows(foreign, foreign_extra) %>%
  filter(date_published_year_month != "NA-NA") %>%
  group_by(contaminant) %>%
  #filter(n()>500) %>%
  ungroup %>%
  select(contaminant, date_published_year_month, ID_incident) %>% distinct() %>%
  group_by(contaminant, date_published_year_month) %>%
  mutate(count = n())%>%
  distinct()

pc8 <-ggplot(data = dat8, aes(x = date_published_year_month  , y = count, group = contaminant)) +
  #geom_line(aes(color = contaminant, alpha = 1), size = 1) +
  geom_point(aes(color = contaminant, alpha = 1), size = 3) +
  geom_smooth(aes(x = date_published_year_month  , y = count, color = contaminant))+
  #scale_x_continuous(breaks = sort(unique(dat$date_published_year))[c(TRUE, FALSE)]  )+
  theme(legend.position = "right", ncol=1) +
  #scale_colour_manual(values=c(pal))+
  theme_minimal_hgrid(9, rel_small = 1) +
  labs(x = "year",  colour="hazard group",
       y = "counts",
       title = "")+
    guides(alpha = FALSE) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

pc8
```



```
# only 4 cataegories

dat9<-bind_rows(foreign, foreign_extra) %>%
  filter(date_published_year_month != "NA-NA") %>%
  filter(date_published_year_quarter   != "2020-Q2") %>%
  filter(contaminant %in% c("plastic", "metal", "glass", "insects")) %>%
  group_by(contaminant) %>%
  #filter(n()>500) %>%
  ungroup %>%
  select(contaminant, date_published_year_month, date_published_year, date_published_year_quarter, ID_i
  group_by(contaminant, date_published_year_month, date_published_year, date_published_year_quarter) %>
  mutate(count = n())%>%
  distinct()

pc9 <-ggplot(data = dat9, aes(x = date_published_year_month   , y = count, group = contaminant)) +
  #geom_line(aes(color = contaminant, alpha = 1), size = 1) +
  geom_point(aes(color = contaminant, alpha = 1), size = 2) +
  geom_smooth(aes(x = date_published_year_month   , y = count, color = contaminant))+
  #scale_x_continuous(breaks = sort(unique(dat$date_published_year))[c(TRUE, FALSE)]  )+
  theme(legend.position = "right", ncol=1) +
  #scale_colour_manual(values=c(pal))+
  theme_minimal_hgrid(9, rel_small = 1) +
  labs(x = "year",  colour="hazard group",
       y = "counts",
       title = "")+
    guides(alpha = FALSE) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
pc10 <-ggplot(data = dat9, aes(x = date_published_year_quarter  , y = count, group = contaminant)) +
  #geom_line(aes(color = contaminant, alpha = 1), size = 1) +
  geom_point(aes(color = contaminant, alpha = 1), size = 2) +
  geom_smooth(aes(x = date_published_year_quarter  , y = count, color = contaminant))+
  #scale_x_continuous(breaks = sort(unique(dat$date_published_year))[c(TRUE, FALSE)]  )+
  theme(legend.position = "right", ncol=1) +
  #scale_colour_manual(values=c(pal))+
  theme_minimal_hgrid(9, rel_small = 1) +
  labs(x = "year",  colour="hazard group",
       y = "counts",
       title = "")+
    guides(alpha = FALSE) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
foreign_all<-bind_rows(foreign, foreign_extra)
dim(foreign_all)
```

```
## [1] 2126   29
```

```r
foreign_all %>% count(origin_country, sort=T)
```

```
## # A tibble: 78 x 2
##    origin_country     n
##    <chr>          <int>
##  1 Canada           542
##  2 USA              240
##  3 United Kingdom   216
##  4 Germany          166
##  5 France           156
##  6 Belgium           94
##  7 Netherlands       87
##  8 Italy             76
##  9 Denmark           64
## 10 Austria           38
## # ... with 68 more rows
```

```r
foreign_all %>% count(notified_country, sort=T)
```

```
## # A tibble: 36 x 2
##    notified_country     n
##    <chr>            <int>
##  1 Canada             532
##  2 United Kingdom     241
##  3 Germany            216
##  4 USA                200
##  5 France             141
##  6 Denmark            119
##  7 <NA>                98
##  8 Netherlands         83
##  9 Belgium             82
## 10 Italy               75
## # ... with 26 more rows
```

Add coordinates data

```
add_coordinates <- function(input){

  ## get coordinates
  # get world map
  wmap <- getMap(resolution="high")
  # get centroids
  centroids <- gCentroid(wmap, byid=TRUE)
  # get a data.frame with centroids
  geo_df <- as.data.frame(centroids)
  colnames(geo_df) <- c("long", "lat")
  geo_df <- geo_df %>% tibble::rownames_to_column("country")


  # update names in data
  input<- input %>%
    mutate( origin_country = case_when(origin_country == "USA" ~ "United States of America",
                                            origin_country =="Hong Kong" ~ "Hong Kong S.A.R.",
                                            origin_country =="Serbia" ~ "Republic of Serbia",
                                            origin_country =="Bosnia Herzegovina" ~ "Bosnia and Herzego
                                            origin_country =="Tanzania" ~ "United Republic of Tanzania"
                                            TRUE ~ origin_country)) %>%
     mutate( notified_country = case_when(notified_country == "USA" ~ "United States of America",
                                            notified_country =="Hong Kong" ~ "Hong Kong S.A.R.",
                                            notified_country =="Serbia" ~ "Republic of Serbia",
                                            notified_country =="Bosnia Herzegovina" ~ "Bosnia and Herzeg
                                            notified_country =="Tanzania" ~ "United Republic of Tanzani
                                            TRUE ~ notified_country))  %>%
    filter(!origin_country %in% c("Palestinian Territories", "INFOSAN" , "Commission Services", NA) | !


  # join with coords data
  output <- input %>%
    left_join(., geo_df, by = c("origin_country" = "country")) %>%
    rename("lat.origin" = "lat",
           "long.origin" = "long") %>%
    left_join(., geo_df, by = c("notified_country" = "country")) %>%
    rename("lat.notified" = "lat",
           "long.notified" = "long") %>% drop_na()

  return(output)
}



test <- add_coordinates(data)
```

```
foreign_all %>% count(alert_type, sort=T)
```

```
## # A tibble: 10 x 2
##    alert_type                     n
##    <chr>                      <int>
##  1 recall                      1400
##  2 alert                      389
##  3 information for follow-up   141
##  4 border rejection            80
##  5 information for attention   79
##  6 update                      21
##  7 lookout                      6
##  8 outbreak                     6
##  9 information                  3
## 10 warning                      1
```

```r
foreign_all %>% count(product_categoty, sort=T)
```

```
## # A tibble: 31 x 2
##    product_categoty                                    n
##    <chr>                                           <int>
##  1 fruits and vegetables                             380
##  2 meat and meat products (other than poultry)       303
##  3 cereals and bakery products                       261
##  4 poultry meat and poultry meat products            176
##  5 <NA>                                              175
##  6 milk and milk products                            160
##  7 nuts; nut products and seeds                      119
##  8 prepared dishes and snacks                         91
##  9 cocoa and cocoa preparations; coffee and tea       79
## 10 soups; broths; sauces and condiments               71
## # ... with 21 more rows
```

```r
foreign_all %>% count(contaminant, sort=T)
```

```
## # A tibble: 10 x 2
##    contaminant             n
##    <chr>               <int>
##  1 metal                 622
##  2 plastic               502
##  3 other foreign body    351
##  4 glass                 307
##  5 insects               221
##  6 wood                   37
##  7 rodents                35
##  8 bone                   28
##  9 stones or soil         18
## 10 paper                   5
```

```r
foreign_all<-foreign_all %>%
  mutate(contaminant2 = ifelse(contaminant %in% c("wood", "rodents", "bone","stones or soil", "paper"),


foreign_all %>%
  select(-link, -brand, -manufacturer, -raw_text_product, -organisation, -date_added, -date_added_year)
  add_coordinates() %>%
  write_tsv("../data/food_hazards_foreign_bodies.csv")
```

save all with coords

```
data %>%
  select(-link, -brand, -manufacturer, -raw_text_product, -organisation, -date_added, -date_added_year)
  add_coordinates() %>%
  write_tsv("../data/food_hazards_data_all.csv")
```