Strawberries

Ajay Krishnakumar

2023-10-23

Introduction: The data, the motivations

When was the last time you looked at your strawberries - not just a cursory examination before you ate them - but when you really looked? At the labels, at the warnings and cautions and USDA certifications and the chemicals they spray on them? I had certainly never considered any of those factors until this assignment, until we were told that perhaps this was something to look into, until Haviland Wright painted for us the extent to which strawberries are not to be trusted.

- What sort of chemicals are being used on strawberries?
- How hazardous are these chemicals?
- Where are these chemicals used most?

Considering the consistent placing of Strawberries on the 'Dirty Dozen'list, (https://www.ewg.org/foodnews/dirt dozen.php) these some questions we should be thinking about. As consumers we should be concerned about the extensive use of pesticide in daily foods. As human beings with an instinct for self-preservation, we should be concerned about the chemicals we are putting in our bodies. So what can the data tell us? How alarmed should we be? I'll try to answer that first question. The second is left as an exercise for the reader.

Before getting into the weeds with the pesticides, we shall also look at data on market information. Is there anything we can say about prices of strawberries, their end uses? It would certainly be a useful lens with which to color our later analysis.

The data itself is from the USDA NASS database, cleaned through much effort on Haviland's part(though perhaps not one hundredth of the effort we should put into cleaning our strawberries). It is this data - collected by survey instead of census and focusing on non organic strawberries - that serves as the jumping off point for what follows.

Data Cleaning and Organization

The first order of business is to split the data that remains from the earlier cleaning into market data and chemical data. This is done thus:

```r
#Loading the libraries we shall need
library(knitr)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(stringr)
library(ggplot2)
library("Hmisc")
```

```

Attaching package: 'Hmisc'

The following objects are masked from 'package:dplyr':

    src, summarize

The following objects are masked from 'package:base':

    format.pval, units
```

```r
strwb_survey<- read.csv("strwb_survey.csv", header = TRUE)

#Extracting the data relevant to chemicals using the Pesticide or Market Column
```

```
strwb_survey_c<- strwb_survey |> filter(Pesticide_or_Market ==
                                          " BEARING - APPLICATIONS")
strwb_mkt<- strwb_survey |> filter(Pesticide_or_Market !=
                                     " BEARING - APPLICATIONS")
```

The data being read in here is the data that Haviland cleaned, with a couple columns named by me and written into a csv file.

Cleaning Market Data

Let us consider first, the market data. A view statement shows us what it looks like and what needs to be done to clean it. I won't print that View statement here but suffice it to say there are changes that can be made here to clean the data. First, regarding the column I have named Price or Weight: It transpires that this is hardly an appropriate name.

```
strwb_mkt |> distinct(Price_or_Weight.)
```

```
  Price_or_Weight.
1   PRICE RECEIVED
2             <NA>
3       PRODUCTION
```

```
strwb_mkt <- strwb_mkt |> rename(PriceProduction = Price_or_Weight.)
```

We can see that this column takes three distinct values: "PRICE RECEIVED", NA and "PRODUCTION". I have renamed this column more sensibly above. While I have called it PriceProduction, it is important to note that the NA values aren't just dead space in every case. In the case of the chemical subset of the data, this should rightly be NA. But this dataset doesn't have the chemicals data. So what about the NAs here?

A quick examination tells us that Price or Production can be interpolated from one of the next two columns.

What about the name of the Pesticide or Market column? This seemed fitting when I used the column to subset chemical and market data. We can check distinct Domain Categories to confirm that this is a valid way to subset our data. But in the market subset, what information does this column now convey?

But now we see the Purpose column doesn't always have purpose. Sometimes it contains what is very clearly the Measurement Unit, while the Measurement Unit itself is NA. Let's fix this too.

```r
strwb_mkt$Measurement_Unit<- ifelse(is.na(strwb_mkt$Measurement_Unit),
                                strwb_mkt$Purpose,strwb_mkt$Measurement_Unit)
strwb_mkt$Purpose<- str_replace(strwb_mkt$Purpose,"- PRICE RECEIVED","")
strwb_mkt$Purpose<- str_replace(strwb_mkt$Purpose,"- PRODUCTION","")

strwb_mkt$Purpose<- ifelse(str_detect(strwb_mkt$Purpose,"MEASURED")==TRUE,
                           "AGGREGATE",
                           strwb_mkt$Purpose)

strwb_mkt$Aggregate_type<- ifelse(strwb_mkt$Purpose == "Aggregate",
                                  strwb_mkt$Measurement_Unit,
                                  strwb_mkt$Aggregate_type)
```

Measurement Unit doesn't seem to always represent measurement unit and in fact probably contains aggregate type instead - for example where it says "Production Utilization". Let's fix that by switching the data in those two columns in such situations

```r
a<- strwb_mkt$Measurement_Unit[which(str_detect(strwb_mkt$Measurement_Unit,
                                                "MEASURED")==FALSE)]
b<- strwb_mkt$Aggregate_type[which(str_detect(strwb_mkt$Measurement_Unit,
                                              "MEASURED")==FALSE)]

strwb_mkt$Aggregate_type[which(str_detect(strwb_mkt$Measurement_Unit,
                                          "MEASURED")==FALSE)]<- a
strwb_mkt$Measurement_Unit[which(str_detect(strwb_mkt$Measurement_Unit,
                                            "MEASURED")==FALSE)]<- b

#Casting all non numeric values as NA
strwb_mkt$Value<- suppressWarnings(as.numeric(strwb_mkt$Value))
```

That's the market data cleaned.

Cleaning Chemical Data

We start by extracting all the chemical information we can from the data we have. This is stored in the Domain and Domain Category columns.

```r
#Getting the Chemical Type and putting it into its own column
strwb_survey_chem1 <- strwb_survey_c |> mutate(`Chemical Type` =
                        str_sub(Domain,str_locate(Domain,"CHEMICAL,")[,2]+1,),
                                                 .after= Domain.Category)
```

```
#Extracting Chemical Name into a column
strwb_survey_chem1<- strwb_survey_chem1 |> mutate(`Chemical Name`=
                                            str_sub(Domain.Category,
                                          str_locate(Domain.Category,":")[,2]+3,
                                        str_locate(Domain.Category,"=")[,1]-2),
                                                  .after = `Chemical Type`)
#Extracting Chemical ID into a column
strwb_survey_chem1<- strwb_survey_chem1 |> mutate(`Chemical ID`=
                          str_sub(Domain.Category,-7), .after= `Chemical Name`)

strwb_survey_chem1$`Chemical ID`<-
  str_sub(strwb_survey_chem1$`Chemical ID`,1,-2)

strwb_survey_chem1$`Chemical ID`<- ifelse(strwb_survey_chem1$Domain=="TOTAL",
                                          NA,strwb_survey_chem1$`Chemical ID`)

strwb_survey_chem1$`Chemical ID`<-
  ifelse(str_detect(strwb_survey_chem1$Domain.Category,"TOTAL")==TRUE,NA,
                                          strwb_survey_chem1$`Chemical ID`)

strwb_survey_chem1$`Chemical ID`<-
  str_replace(strwb_survey_chem1$`Chemical ID`,"=","")

strwb_survey_chem1$`Chemical ID`<-
  str_trim(strwb_survey_chem1$`Chemical ID`, side = 'both')

strwb_survey_chem<- strwb_survey_chem1
```

Now we will use the PC codes we've extracted and placed in Chemical ID and use the EPA Pesticide Chemical Search (https://ordspub.epa.gov/ords/pesticides/f?p=chemicalsearch:1) to find CAS codes where we can.

Now, these CAS codes are cross-referenced with the WHO's Classification of Pesticides by Hazard and Guidelines to Classification(https://iris.who.int/bitstream/handle/10665/332193/9789240005662-eng.pdf?sequence=1), pages 72 onwards. This gives us the following:

```
#Reading in UN information on hazards by CAS number
library(readxl)
hazards <- read_xlsx("WHO_codes.xlsx")
hazards$CAS_no <- str_trim(str_sub(hazards$CAS,1,
                                   str_locate(hazards$CAS," ")[,2]),"both")
```

```
codes<- hazards$CAS_no

hazards$Toxicity<- str_sub(hazards$CAS, str_locate(hazards$CAS,
                    fixed(hazards$CAS_no))[,2],)

hazards$Toxicity <- str_match(hazards$Toxicity, " \\s*(.*?)\\s* ")[,2]

hazards<- hazards |> select(-CAS)

hazards$Hazard <- ifelse(hazards$Toxicity=="Ia","Extremely hazardous",
                  ifelse(hazards$Toxicity=="Ib","Highly hazardous",
                  ifelse(hazards$Toxicity=="II","Moderately hazardous",
                  ifelse(hazards$Toxicity=="III", "Slightly hazardous",
                  ifelse(hazards$Toxicity =="U","Not acutely hazardous",
                  ifelse(hazards$Toxicity=="FM","Fumigant","Obsolete"))))))

hazard_info<- merge(x =chemical_list,y=hazards,by="CAS_no",
                    x.chemical_list=TRUE)

#adding in a few chemicals whose codes were missing by hand

hazard_info<-rbind(hazard_info,c("20543-04-8","COPPER OCTANOATE","
23306","U","Not acutely hazardous"))

hazard_info<-rbind(hazard_info,c("2180409-60-3","
CYFLUFENAMID","555550","U","Not acutely hazardous"))

hazard_info<-rbind(hazard_info,c("121552-61-2","CYPRODINIL","
288202","U","Not acutely hazardous"))

hazard_info<-rbind(hazard_info,c("70630-17-0","
MEFENOXAM","113502","III","Slightly hazardous"))

hazard_info<-rbind(hazard_info,c("","PAECILOMYCES FUMOSOR","115002","U",
                                 "Not acutely hazardous"))


strwb_chem_info<- merge(x=strwb_survey_chem,y=hazard_info, by= "Chemical Name",
                        all.x=TRUE)

strwb_chem_info$Value <- suppressWarnings(as.numeric(strwb_chem_info$Value))
```

```r
strwb_chem_info$Value <- ifelse(strwb_chem_info$Value == "
(NA)", NA,strwb_chem_info$Value)
```

To prepare for EDA let's do a few more things to the chemical data.

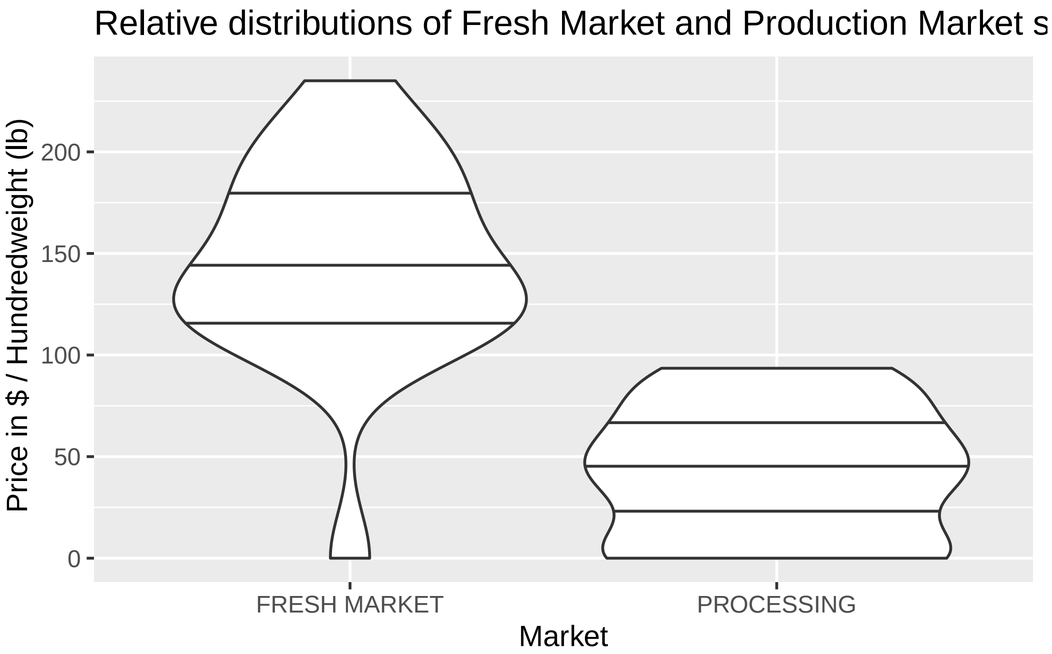
```r
#We are going to filter our data set so we look only at those
#entries for which we have hazard information

chem<- strwb_chem_info |> filter(is.na(Hazard)!=TRUE)

#Data with the aggregate type = AVG. What does AVG actually show us?

avg_chem <- chem |> filter(Aggregate_type==" AVG")
non_avg_chem <- chem |> filter(is.na(Aggregate_type)==TRUE)
```

A look at the two data frames created as above gives us interesting information. It would appear that the 'non_avg-chem' data shows the total weight in pounds of a particular fertilizer applied in that particular Year. The 'avg-chem' data looks at various 'average' data. I would contend that average is a little bit of a misnomer here and calling these 'per unit' data would be more informative. We have lb/acre/year, lb/acre/application and the 'number'

Exploratory Data Analysis - Market Data

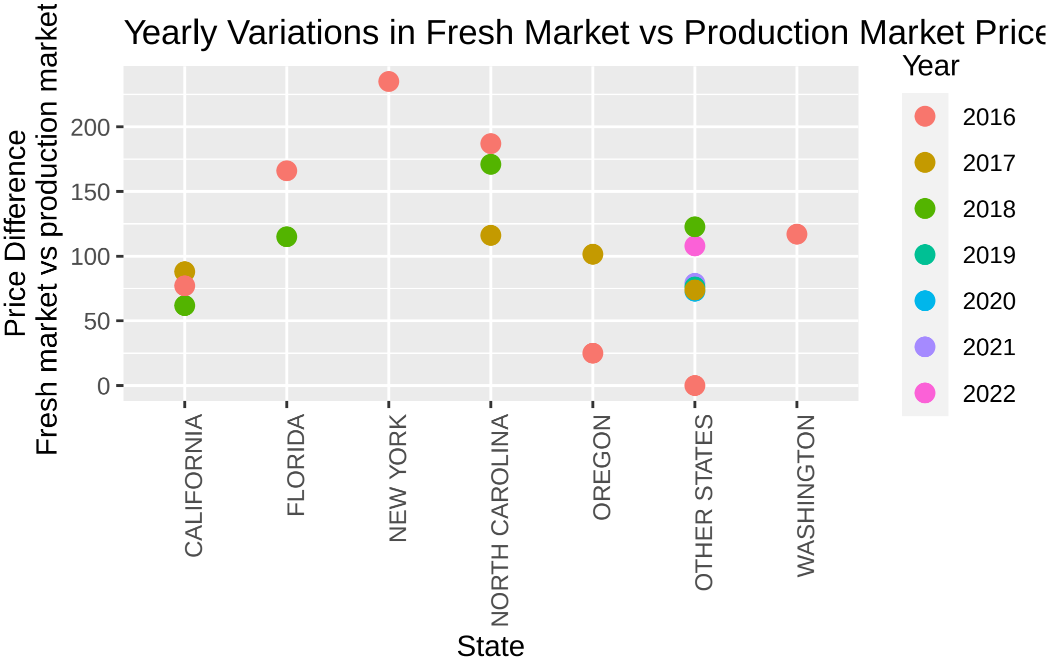
Preliminary Questions:

- What strawberries are most expensive? Intuitively you'd expect fresh market strawberries to have the highest markup. Is this true?
- How do prices vary across state?



We can tell from this graph that the overwhelming majority of fresh market strawberries are sold for significantly more than processing market strawberries. Just compare the median price per hundredweight(the middle lines in the violin plots). What's interesting is that the interquartile range for the two seems very similar at $40/cwt. The minimum and maximum values for fresh market strawberries are miles apart however. This might be due to transportation costs or the mark-up at places like farmers markets, say.

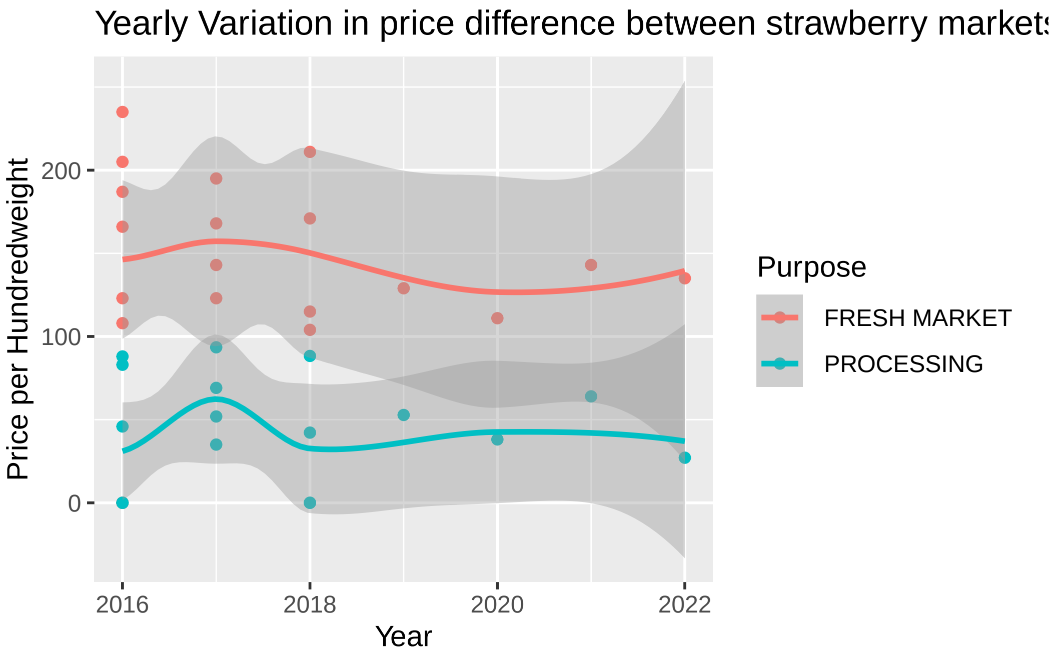
What can we tell about how the difference in fresh market and processing varies between states?



We can see that California, relative to other states is relatively consistent in the difference in price of fresh market and production market strawberries. It is also on the lower side. North Carolina has the largest variation in the difference.

What's also interesting here is how in 2016, east coast states observed a higher difference in price relative to west coast states.

What about yearly variations in these prices?



I’ll begin the discussion of this graph with a caveat: There is a limited amount of information we can glean from this - this is reflected in the increasing area of the error ‘shadow’ around the line. In large part this is due to the dwindling number of points each consecutive year.

All the same, there is some information we can see. There is a decreasing trend in both prices but also a decrease in the distance between the smoothed regression lines.

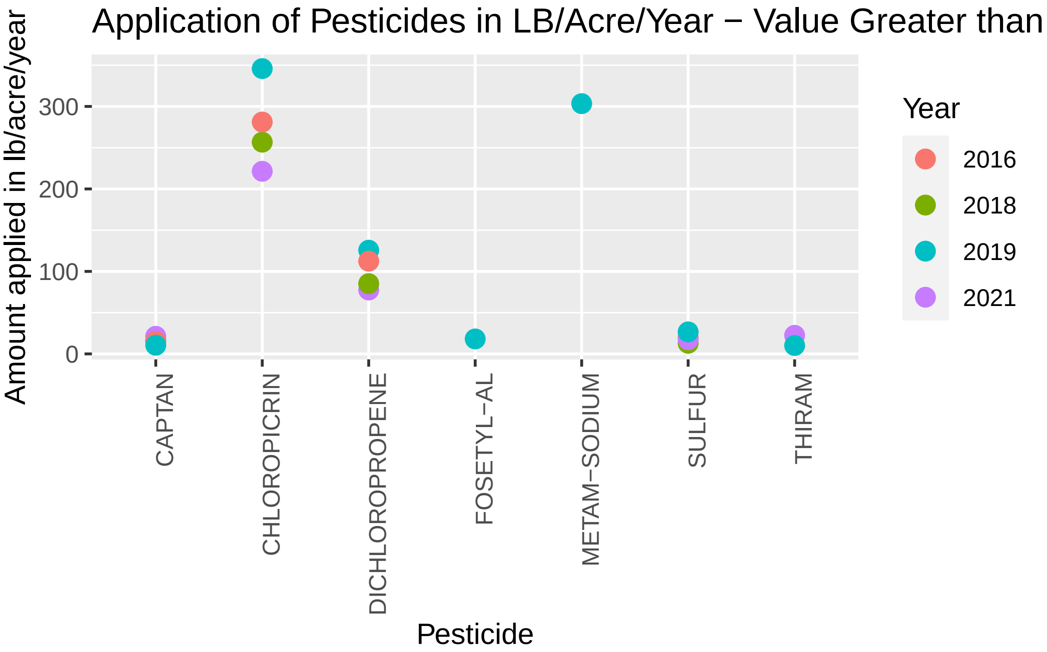
In none of the years do any of the fresh market prices go beneath those of even the highest production market. The overlap of the shadows and the increase in its area towards the end might indicate that perhaps we might find a point or two like that if we had more data for the more recent years and if the trend shown in the graph is backed up by that data. That is certainly interesting. Were that to be true, what would drive that reversal of the dynamic of the two prices?

Maybe there’s a world(not this world and not a world we can infer from the data above) where people buy fewer fresh strawberries. Why? Because who wants to eat chemically-tainted strawberries?

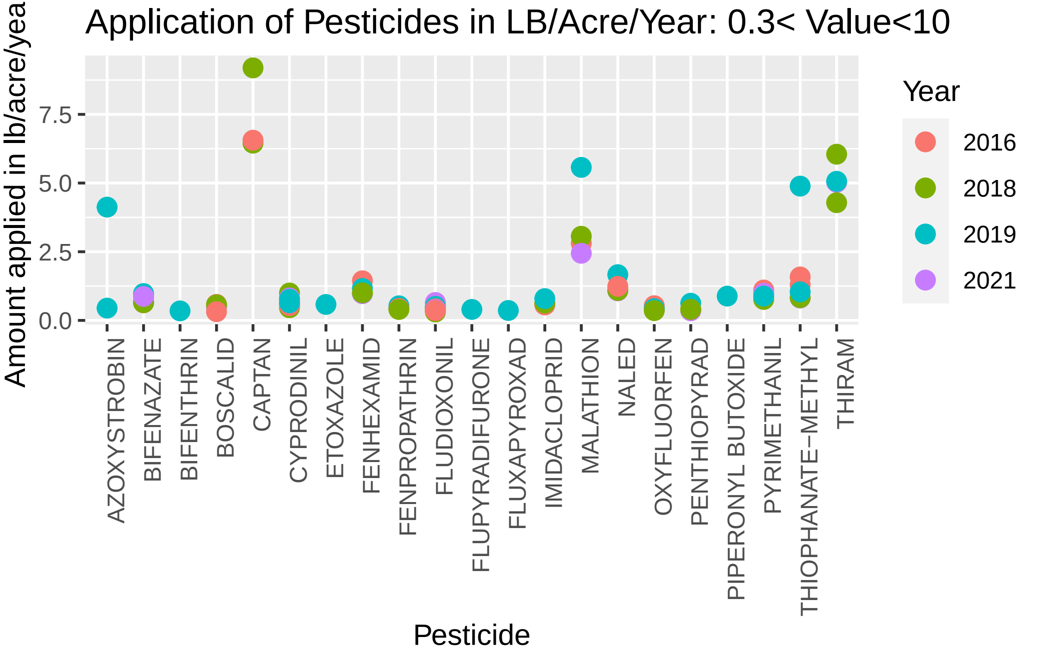
Exploratory Data Analysis - Pesticides on Strawberries

Let’s start by looking at the amount of each pesticide sprayed per year and per application. These tell us two different things. Per year tells us which pesticide is used the most. Per application gives some information about the relative efficacy of each pesticide.

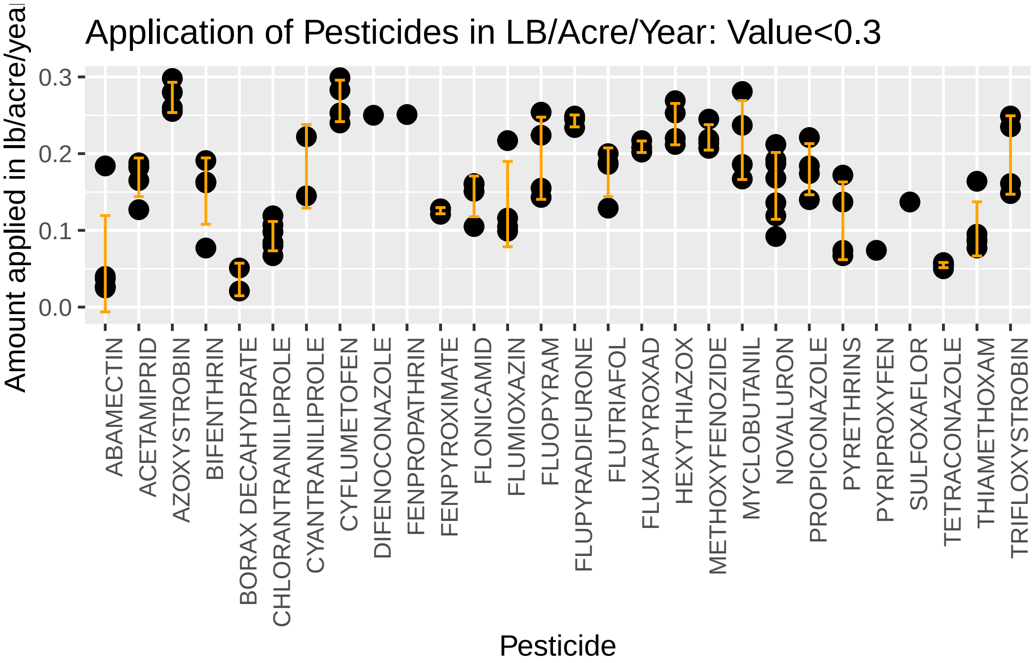
Starting with per year. If we try and plot all of our pesticides, it becomes impossible to glean anything clear because of the sheer range of the data. There are a few pesticides that are used in small quantities.I have I have a hunch that those will turn up in the per application graph so we'll graph their use per year separately.



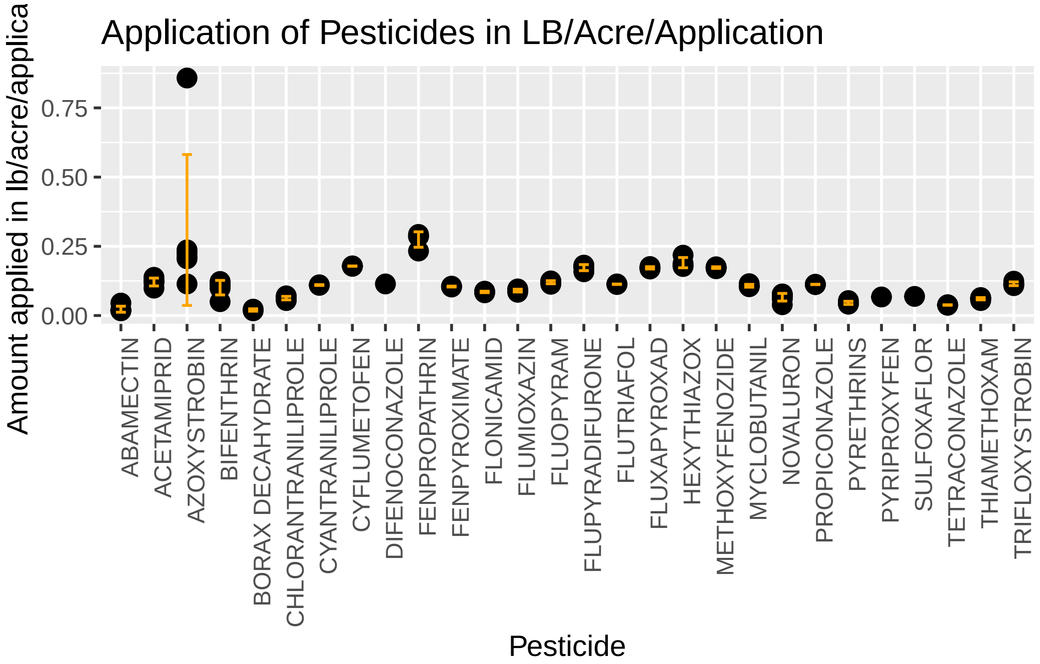
What about for the pesticides with smaller values?



Now for the most interesting of this series of plots, the chemicals(and there are a lot of them) which are applied in smaller quantities per acre per year. I've gotten rid of colors here so we can see the variability in points more easily without being distracted



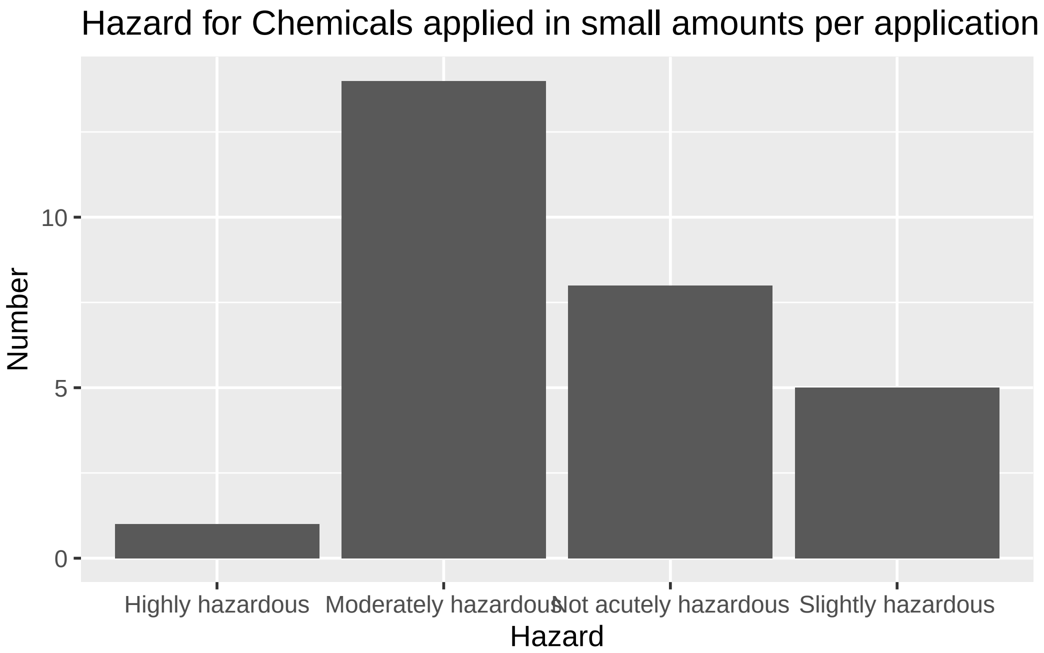
So let's use the same chemicals as in the graph above but let's look at their quantity per application.



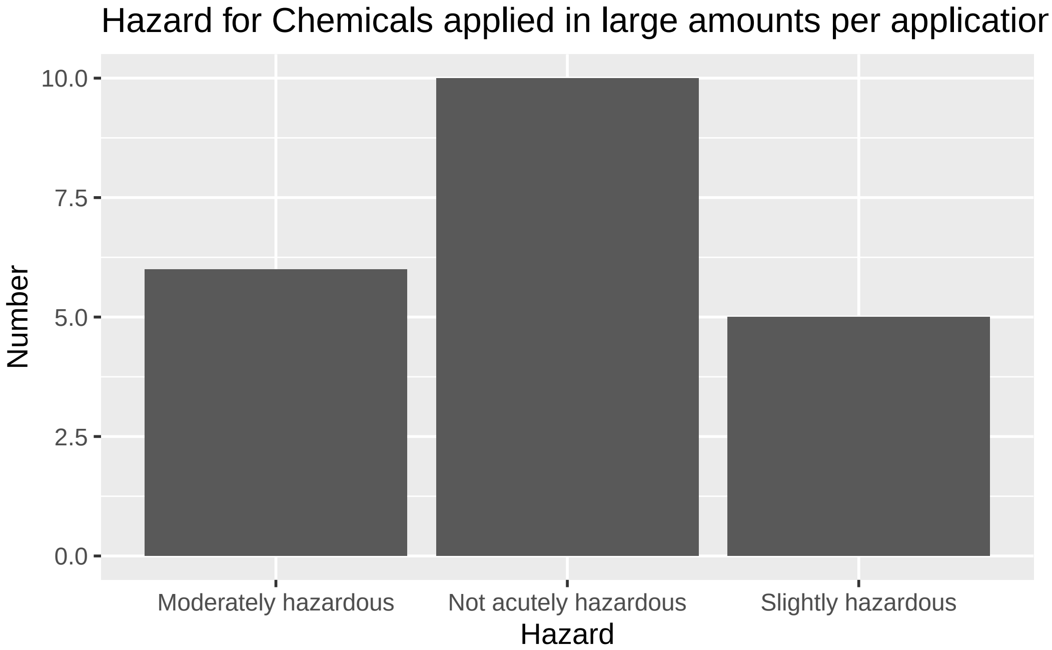
“‘

This data shows that all these compounds are used in incredibly small quantities per application. The relative amount used per application is extremely similar to the relative differences in amount used per year. What does this tell us? Are these pesticides very effective at doing their job? I would say that the data shows that, given the information we have(we could confirm this with information on yields of strawberry crops sprayed with different quantities of each pesticide but that’s beyond the scope of this report). So why are they effective? Could it be because they are toxic.

Before we look at that, its worth noting that the application of azoxystrobin has a large error bar because of one point where it was applied in a large quantity.



So a decent number of these chemicals are either highly hazardous or moderately hazardous. More importantly only eight of them are not acutely hazardous. But this does put some paid to the theory that these would all be highly hazardous chemicals. Why don't we look at chemicals that are applied in large amounts. What is their toxicity like?



This is interesting. It makes sense that a large number of these are not acutely hazardous.

What this suggests is that the small number of severely hazardous chemicals applied is good because perhaps the large number of them are banned. All the same the large number of moderately hazardous chemicals applied is concerning.

Another thing to think about that might bear future exploration: What kind of spectrum is moderately hazardous? We have moderately hazardous chemicals applied in large quantities and in small. Are some much more hazardous than others?

Conculding Remarks

Not good news I'm afraid. Regardless of the happy news that severely hazardous chemicals are not frequently applied and not in large quantities, very many other toxic pesticides do seem to be applied to strawberry crops and some of them in large quantitites. There is more delving that needs to be done:

- What constitutes moderately hazardous? How does the toxicity of these chemicals vary?
- How does the application amount of each chemical relate to the lethal dose(if it exists) for that chemical? How long before it is no longer fatal?
- What is the environmental impact of these chemicals. Those that aren't toxic to humans could well be toxic to other animals. In fact while looking at the toxicity, I found that several of these chemicals are severely, acutely toxic to aquatic life which is a big problem if the pesticides get swept into local water bodies.