Honey Production US

Soumya Paul

29/07/2021

Context

In 2006, global concern was raised over the rapid decline in the honeybee population, an integral component to American honey agriculture. Large numbers of hives were lost to Colony Collapse Disorder, a phenomenon of disappearing worker bees causing the remaining hive colony to collapse. Speculation to the cause of this disorder points to hive diseases and pesticides harming the pollinators, though no overall consensus has been reached. Twelve years later, some industries are observing recovery but the American honey industry is still largely struggling. The U.S. used to locally produce over half the honey it consumes per year. Now, honey mostly comes from overseas, with 350 of the 400 million pounds of honey consumed every year originating from imports. This dataset provides insight into honey production supply and demand in America by state from 1998 to 2012.

Content

The National Agricultural Statistics Service (NASS) is the primary data reporting body for the US Department of Agriculture (USDA). NASS's mission is to "provide timely, accurate, and useful statistics in service to U.S. agriculture". From datasets to census surveys, their data covers virtually all aspects of U.S. agriculture. Honey production is one of the datasets offered.

- *numcol*: Number of honey producing colonies. Honey producing colonies are the maximum number of colonies from which honey was taken during the year. It is possible to take honey from colonies which did not survive the entire year
- *yieldpercol*: Honey yield per colony. Unit is pounds
- *totalprod*: Total production (numcol × yieldpercol). Unit is pounds
- *stocks*: Refers to stocks held by producers. Unit is pounds
- *priceperlb*: Refers to average price per pound based on expanded sales. Unit is dollars.
- *prodvalue*: Value of production (totalprod × priceperlb). Unit is dollars.

Other useful information: Certain states are excluded every year (ex. CT) to avoid disclosing data for individual operations. Due to rounding, total colonies multiplied by total yield may not equal production. Also, summation of states will not equal U.S. level value of production.

Inspiration

- How has honey production yield changed from 1998 to 2012?
- Over time, which states produce the most honey? Which produce the least? Which have experienced the most change in honey yield?
- Does the data show any trends in terms of the number of honey producing colonies and yield per colony before 2006, which was when concern over Colony Collapse Disorder spread nationwide?

- Are there any patterns that can be observed between total honey production and value of production every year? How has value of production, which in some sense could be tied to demand, changed every year?

```
#loading the data set
library(readxl)
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
honeyproduction <- read_excel("honeyproduction.xlsx")
head(honeyproduction)
```

```
## # A tibble: 6 x 8
##   state numcol yieldpercol totalprod   stocks priceperlb prodvalue  year
##   <chr>  <dbl>       <dbl>     <dbl>    <dbl>      <dbl>     <dbl> <dbl>
## 1 AL     16000          71   1136000   159000       0.72    818000  1998
## 2 AZ     55000          60   3300000  1485000       0.64   2112000  1998
## 3 AR     53000          65   3445000  1688000       0.59   2033000  1998
## 4 CA    450000          83  37350000 12326000       0.62  23157000  1998
## 5 CO     27000          72   1944000  1594000       0.7    1361000  1998
## 6 FL    230000          98  22540000  4508000       0.64  14426000  1998
```

```
#dimension of the data
dim(honeyproduction)
```

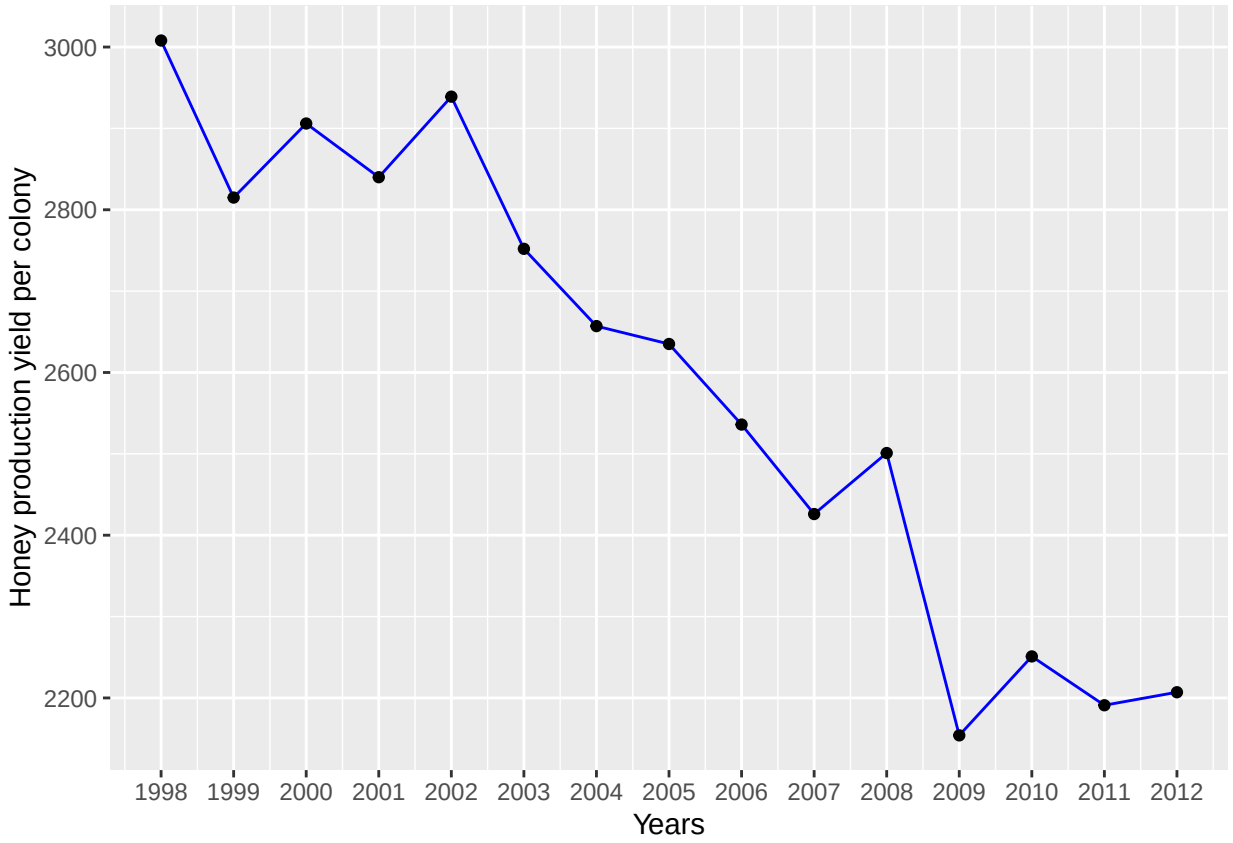
```
## [1] 626   8
```

```
#States in US having honey producing colonies
unique(honeyproduction$state)
```

```
##  [1] "AL" "AZ" "AR" "CA" "CO" "FL" "GA" "HI" "ID" "IL" "IN" "IA" "KS" "KY" "LA"
## [16] "ME" "MD" "MI" "MN" "MS" "MO" "MT" "NE" "NV" "NJ" "NM" "NY" "NC" "ND" "OH"
## [31] "OK" "OR" "PA" "SD" "TN" "TX" "UT" "VT" "VA" "WA" "WV" "WI" "WY" "SC"
```

```
#Production yield from year 1998-2012
p=c()
for (i in 1:15) {
 p[i]=sum(filter(honeyproduction,year==1998+(i-1))$yieldpercol)
}
y=seq(1998,2012,1)
library(ggplot2)
library(scales)
ggplot(NULL,aes(x=y,y=p))+
  geom_line(color="blue")+
  geom_point()+
  xlab("Years")+
  ylab("Honey production yield per colony")+
```

```
  scale_x_continuous(breaks = y)
```



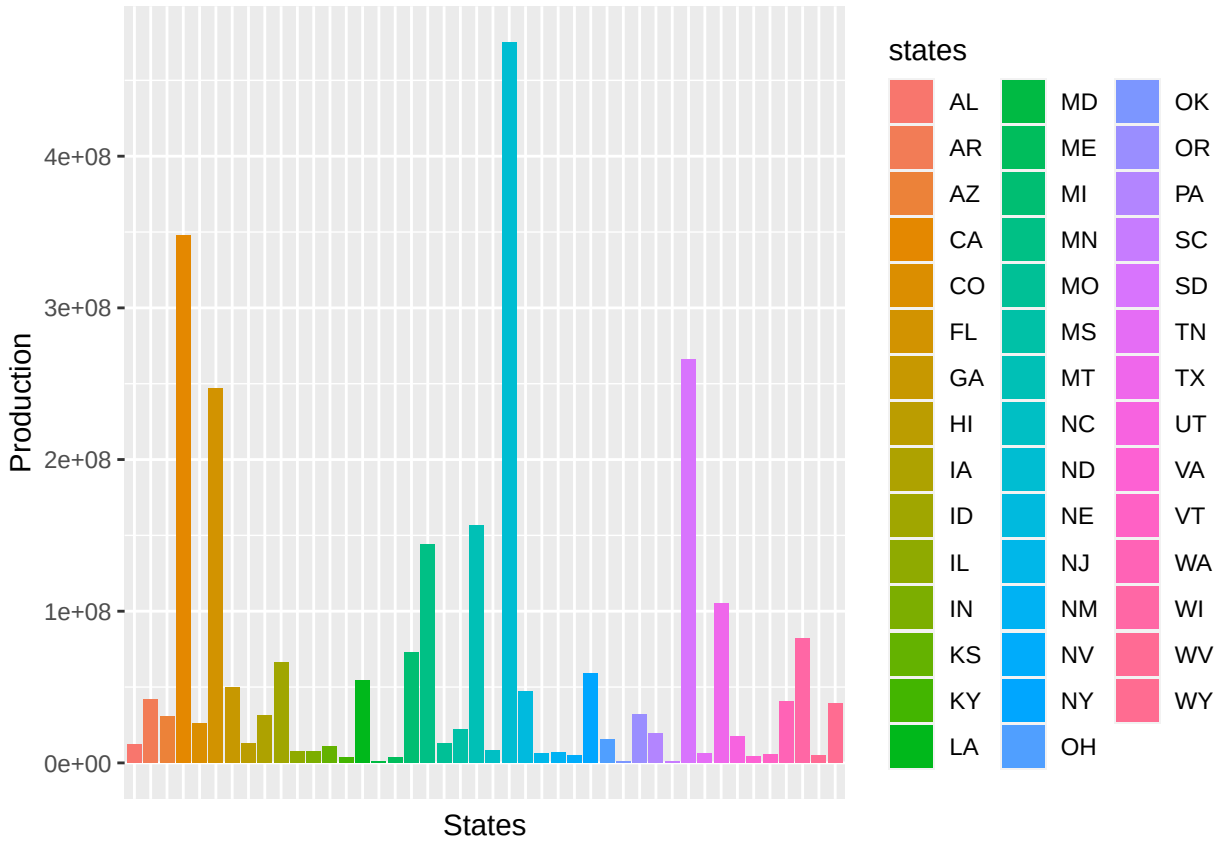

```
#Production on different states over years
states=unique(honeyproduction$state)
p1=c()
for (i in 1:length(states)) {
  p1[i]=sum(filter(honeyproduction,state==states[i])$totalprod)
}
#State having the maximum production
states[which.max(p1)]
```

```
## [1] "ND"
```

```
#State having the minimum production
states[which.min(p1)]
```

```
## [1] "SC"
```

```
#Visualization of the data
library(ggplot2)
ggplot(NULL,aes(x=states,y=p1,fill=states))+
  geom_bar(stat="identity")+
  ylab("Production")+
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())+
  xlab("States")
```



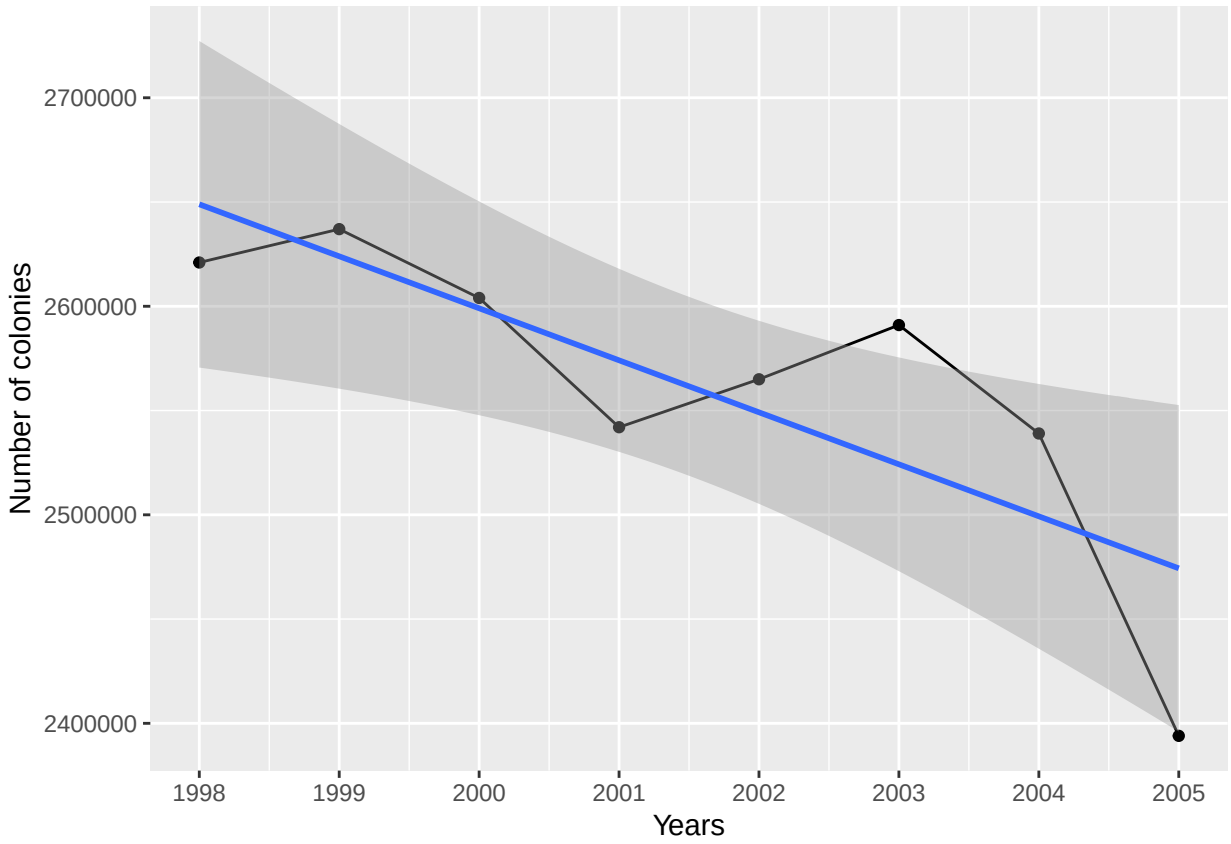

```
p2=c()
for (i in 1:length(states)) {
  p2[i]=var(filter(honeyproduction,state==states[i])$totalprod)
}
#The most change in honey yield happened in
states[which.max(p2)]
```

```
## [1] "CA"
```

Over time, "ND" produces the maximum honey, "SC" produces the least and "CA" experienced the most change in honey yield.

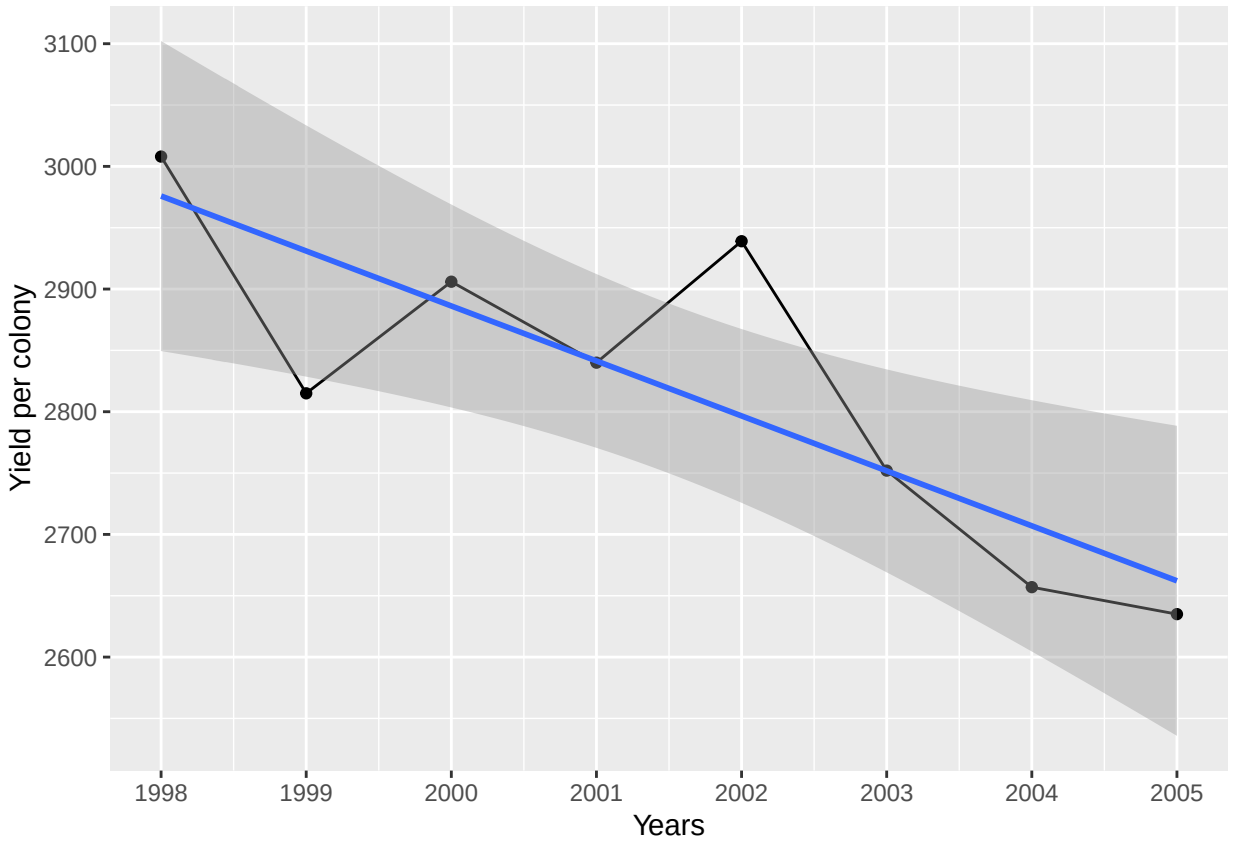
```
#Number of colonies for each year from 1998-2005
m1=c()
for (i in 1:8) {
  m1[i]=sum(filter(honeyproduction,year==1998+(i-1))$numcol)
}
Years=seq(1998,2005,1)
ggplot(NULL,aes(x=Years,y=m1))+
  geom_line()+geom_point()+ylab("Number of colonies")+
  scale_x_continuous(breaks = Years)+geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



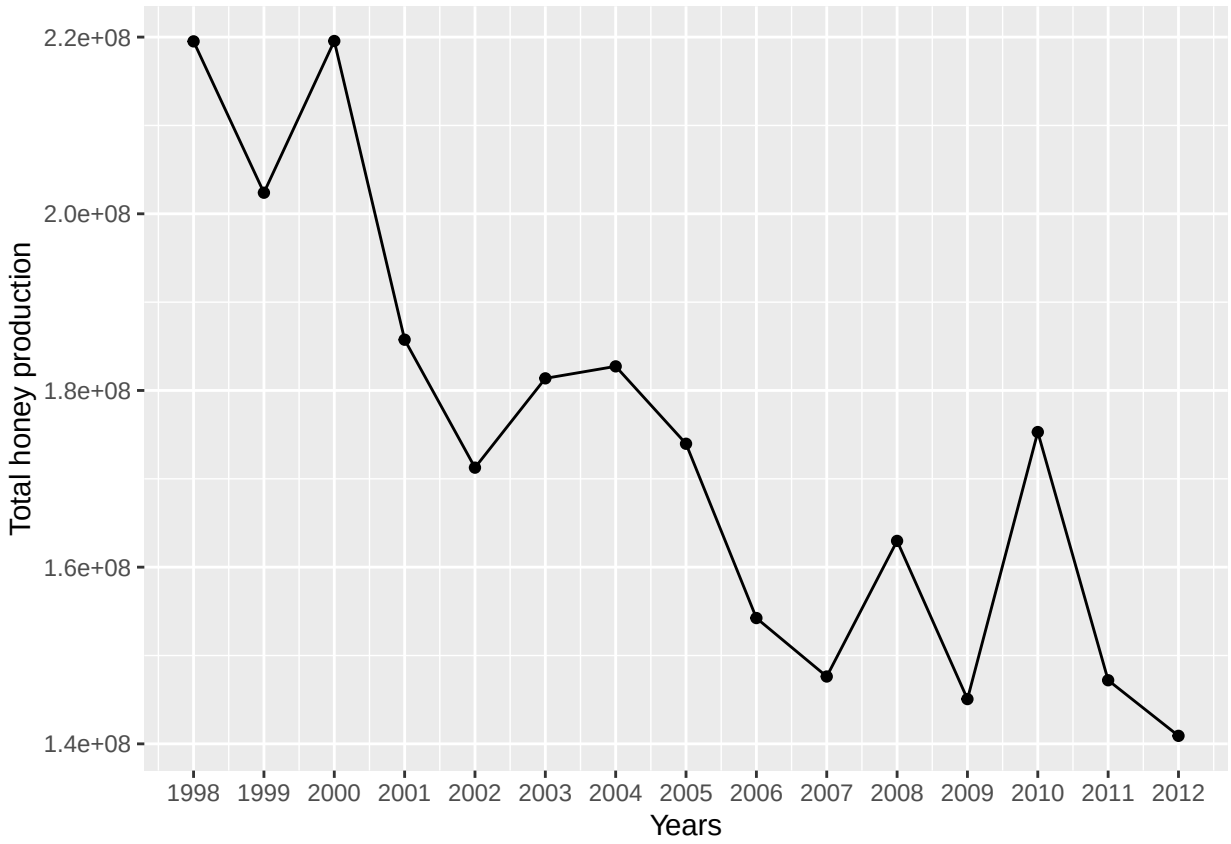

```
#Yield per colony for each year from 1998-2005
m2=c()
for (i in 1:8) {
  m2[i]=sum(filter(honeyproduction,year==1998+(i-1))$yieldpercol)
}
ggplot(NULL,aes(x=Years,y=m2))+
  geom_line()+geom_point()+ylab("Yield per colony")+
  scale_x_continuous(breaks = Years)+geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

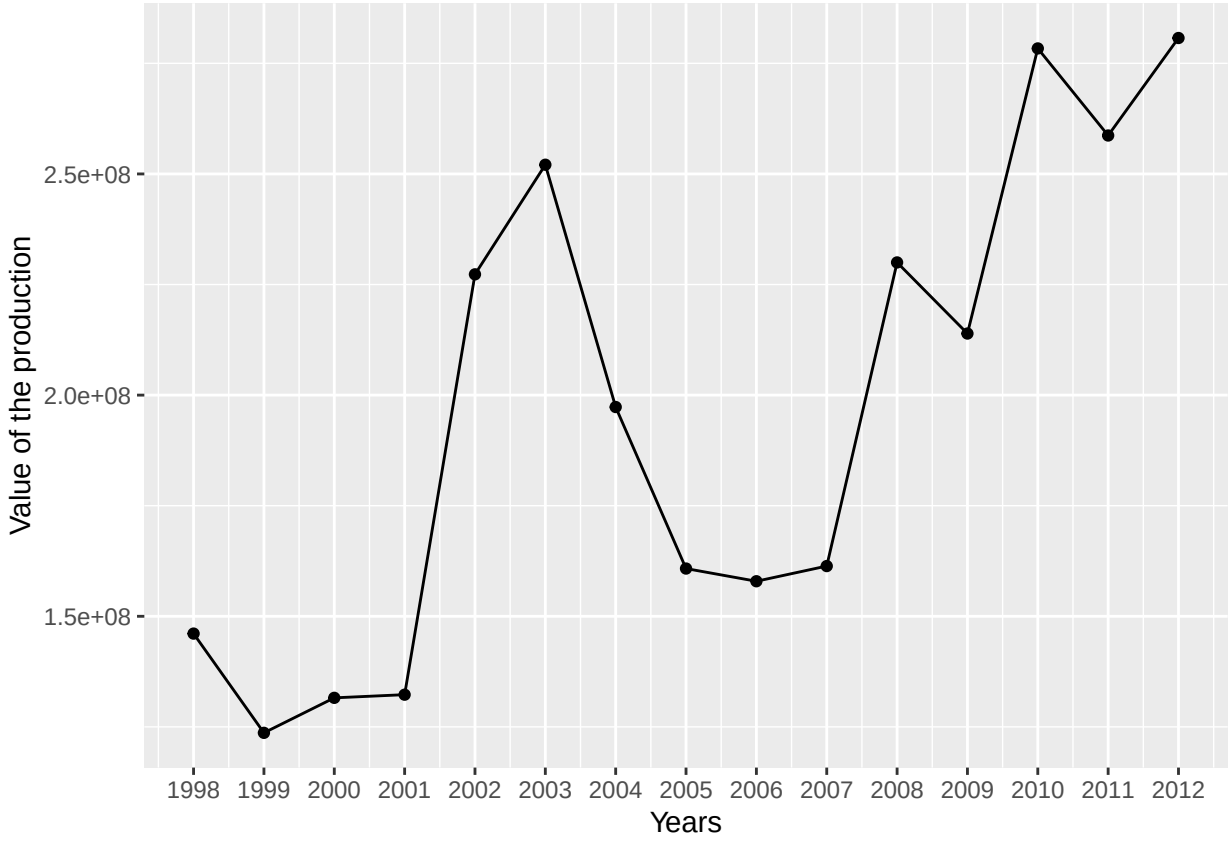


There is a decreasing trend in the number of honey producing colonies and yield per colony before 2006 which is a major concern for a state.

```
#Total honey production over different years
n1=c()
for (i in 1:15) {
  n1[i]=sum(filter(honeyproduction,year==1998+(i-1))$totalprod)
}
Years=seq(1998,2012,1)
ggplot(NULL,aes(x=Years,y=n1))+
  geom_line()+geom_point()+ylab("Total honey production")+
  scale_x_continuous(breaks = Years)
```

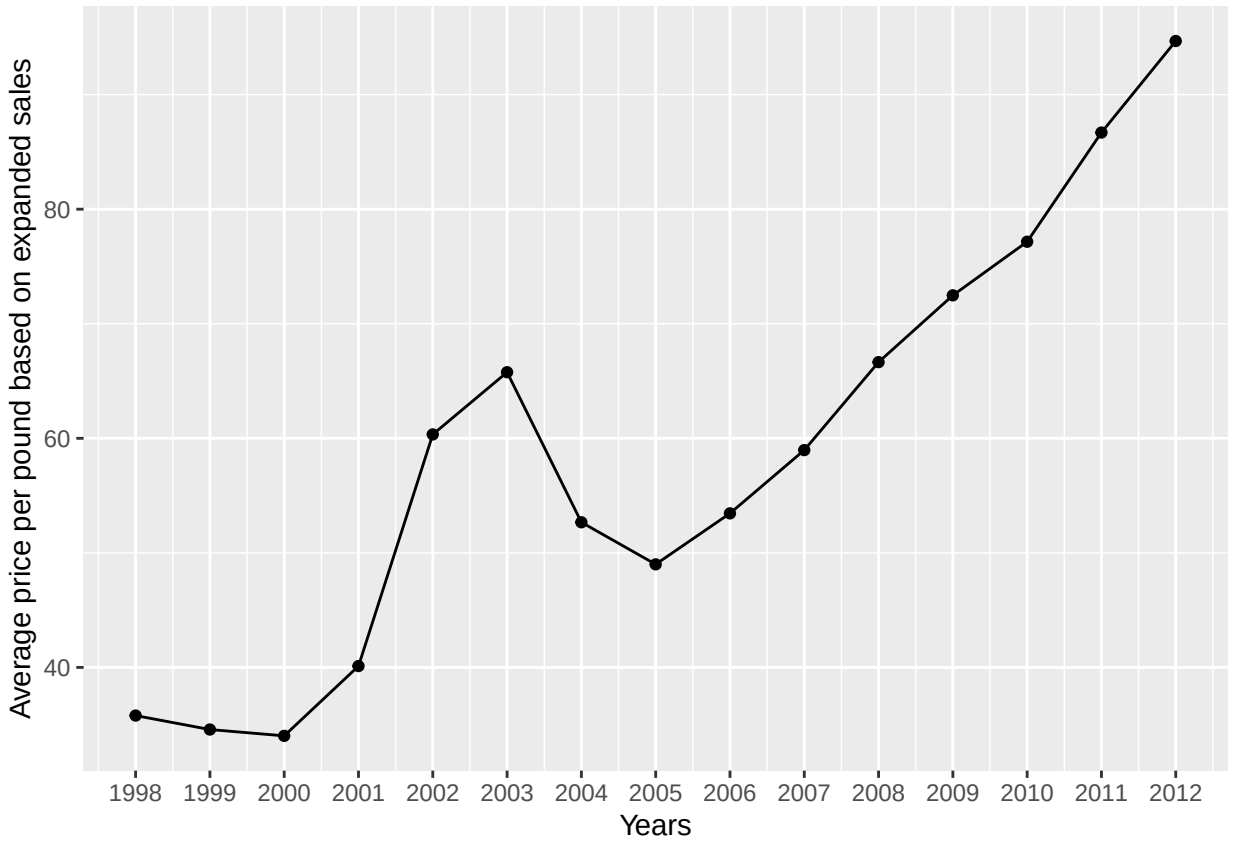



```
n2=c()
for (i in 1:15) {
  n2[i]=sum(filter(honeyproduction,year==1998+(i-1))$prodvalue)
}
ggplot(NULL,aes(x=Years,y=n2))+
  geom_line()+geom_point()+ylab("Value of the production")+
  scale_x_continuous(breaks = Years)
```



As we can see that the total honey production is decreasing whereas the value of the production which in some sense tied with the demand is increasing, we will definitely witness a price hike.

```
Years=seq(1998,2012,1)
n3=c()
for (i in 1:15) {
  n3[i]=sum(filter(honeyproduction,year==1998+(i-1))$priceperlb)
}
ggplot(NULL,aes(x=Years,y=n3))+
  geom_line()+geom_point()+ylab("Average price per pound based on expanded sales")+
  scale_x_continuous(breaks = Years)
```


Average price per pound based on expanded sales
80
60
40
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
Years