Honey production in the USA

Ujan Dasgupta
Roll No.: MDS202249
Email: ujan@cmi.ac.in

December 9, 2022

Abstract

In this project we have taken a dataset containing data on the Honey Production in the USA from 1998 to 2012.We use visualization techniques in R to evaluate the change in the production of honey over the years in the US.

Introduction

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.

Definitions of Key Variables

- *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

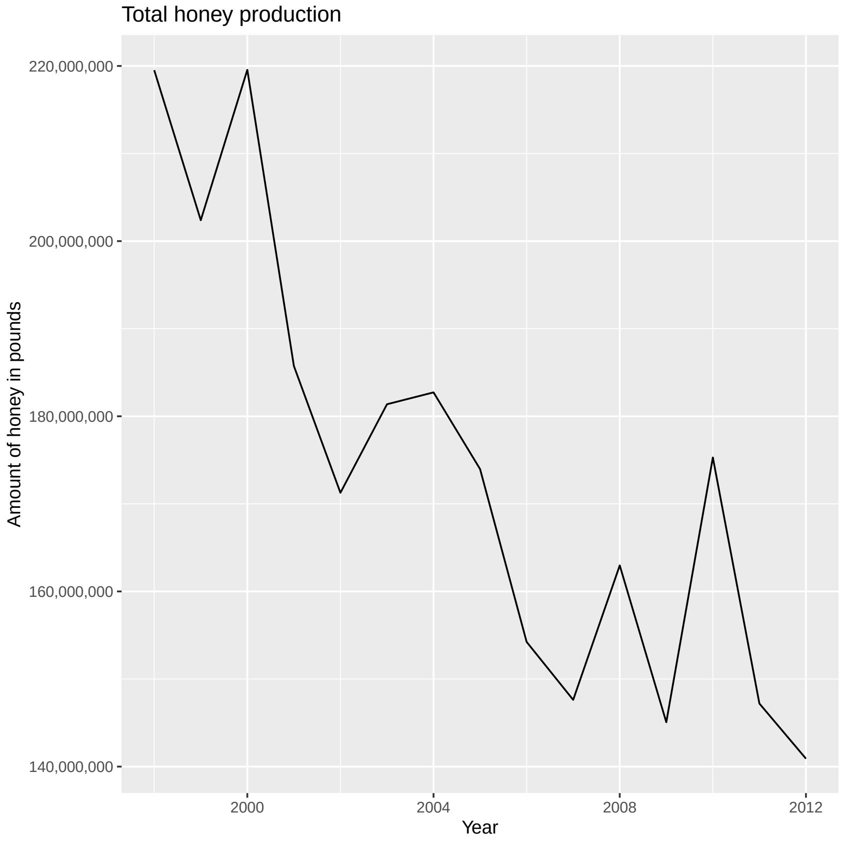
- *totalprod*: Total production (numcol x yieldpercol). Unit is pounds
- *stocks*: Refers to stocks held by producers. Unit is pounds
- *priceperlb*: Refers to average price per pound based on expanded sales. Unit is dollars.
- *prodvalue*: Value of production (totalprod x priceperlb). Unit is dollars.

```
library(tidyverse)

## - Attaching packages -------------------- tidyverse 1.3.2 -
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## - Conflicts --------------------- tidyverse_conflicts() -
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(tidytext)
library(scales)

##
## Attaching package:  ’scales’
##
## The following object is masked from ’package:purrr’:
##
##     discard
##
## The following object is masked from ’package:readr’:
##
##     col_factor

data <- read.csv("E:/visualisation-project/data/honeyproduction.csv")
```

Plots

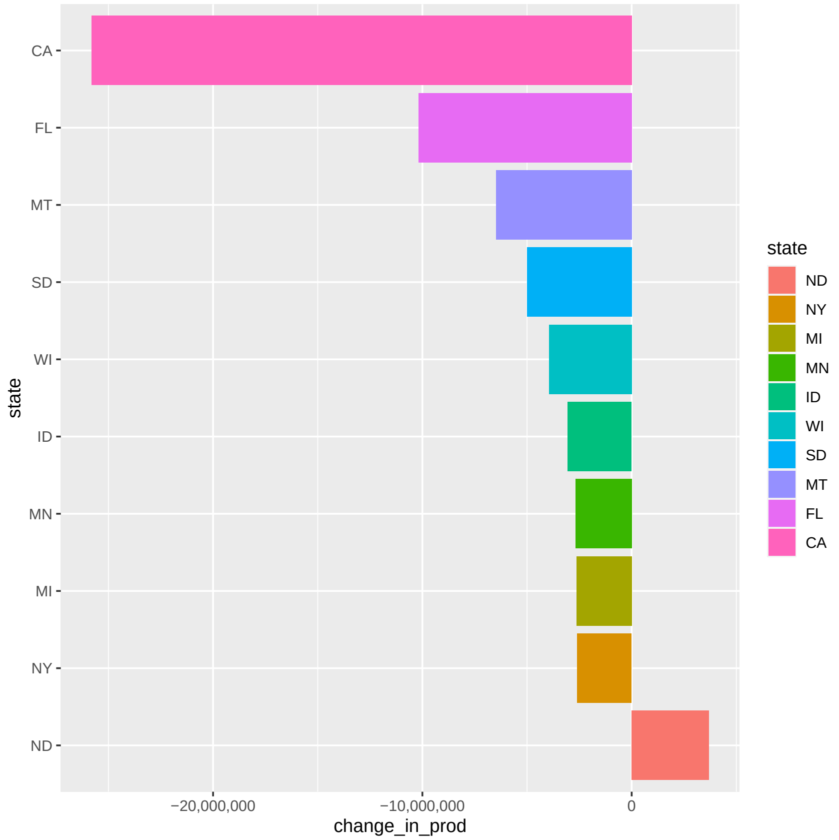
1. First we plot the Total Honey Produced in the USA by the listed stated over the years 1998-2012. We plot the data in a line graph.

```
data %>%
        group_by(year) %>%
        summarise(total_honey_prod = sum(totalprod)) %>%
        ggplot(aes(year, total_honey_prod)) +
                geom_line() +
                scale_y_continuous(labels = comma) +
                labs(title = "Total honey production",
                x = "Year",
                y = "Amount of honey in pounds")
```



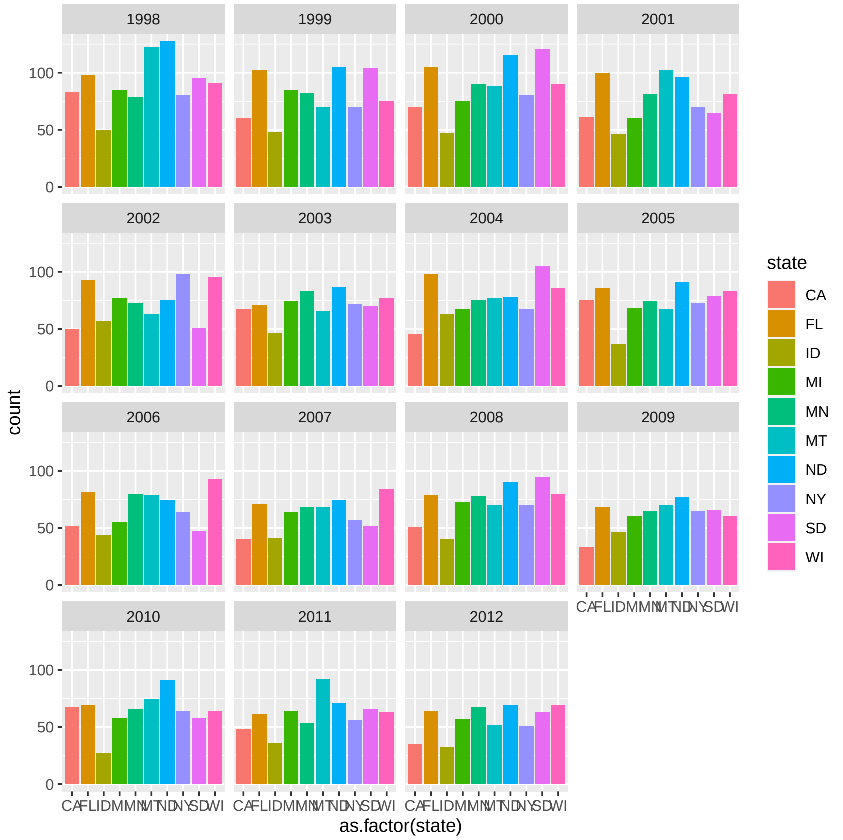
2. From the above graph it is visible that the overall production of Honey has decreased over time in the USA. In this plot, we filter out ten states with the maximum absolute change in the total production of honey fro 1998 to 2012 and we plot the change in a horizontal bar graph for these ten states

```
data %>%
        filter(year %in% c(max(year), min(year))) %>%
        select(state, totalprod, year) %>%
        pivot_wider(names_from = year, values_from = totalprod, names_prefix = "year_")
        group_by(state) %>%
        summarise(change_in_prod = year_2012 - year_1998) %>%
        filter(!is.na(change_in_prod)) %>%
        mutate(state = fct_lump(state, 10 , w = abs(change_in_prod))) %>%
        mutate(state = fct_reorder(state, -change_in_prod)) %>%
        filter(state!= "Other") %>%
        ggplot(aes(change_in_prod, state, fill=state)) +
                geom_col() +
                scale_x_continuous(labels = comma)
```



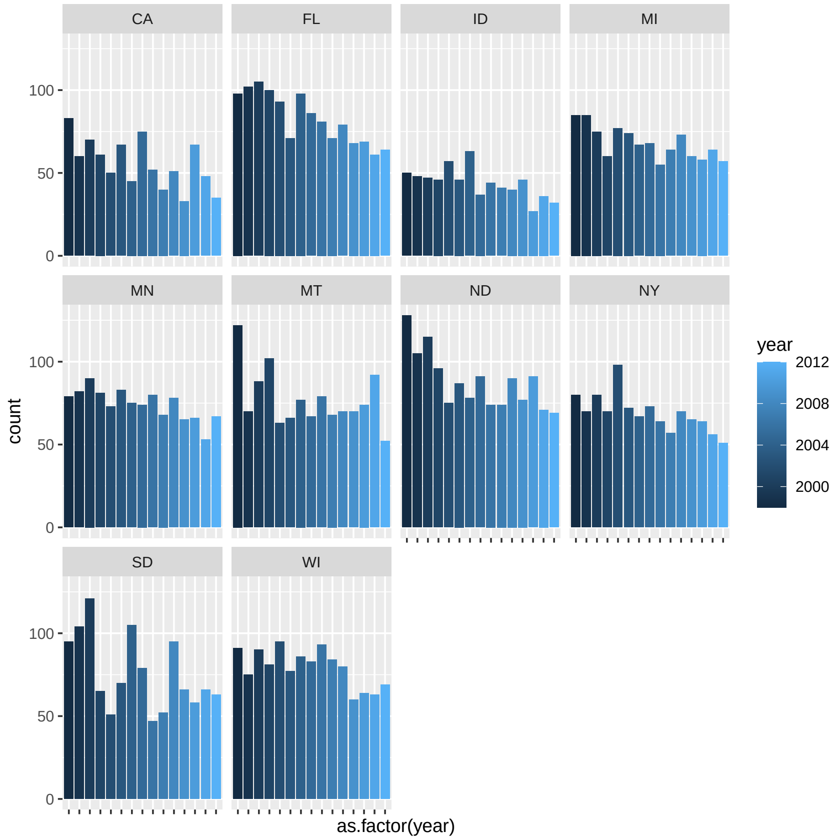
3. From the above graph we get the 10 states of the USA with the maximum absolute change in the production of honey. We now plot the vertical bar diagram for the yields per colony of these 10 states for each year over the period of 15 years (1998-2012).

```
imp_states <- data %>%
        filter(year %in% c(max(year), min(year))) %>%
        select(state, totalprod, year) %>%
        pivot_wider(names_from = year, values_from = totalprod, names_prefix = "year_")
        group_by(state) %>%
        summarise(change_in_prod = year_2012 - year_1998) %>%
        filter(!is.na(change_in_prod)) %>%
        mutate(state = fct_lump(state, 10 , w = abs(change_in_prod))) %>%
        mutate(state = fct_reorder(state, -change_in_prod)) %>%
        filter(state!= "Other") %>%
        distinct(state)

data %>%
        filter(state %in% as.character(as.array(imp_states$state))) %>%
        ggplot(aes(x = as.factor(state), weight = yieldpercol,fill = state))+
                geom_bar() +
                facet_wrap(~ year)
```



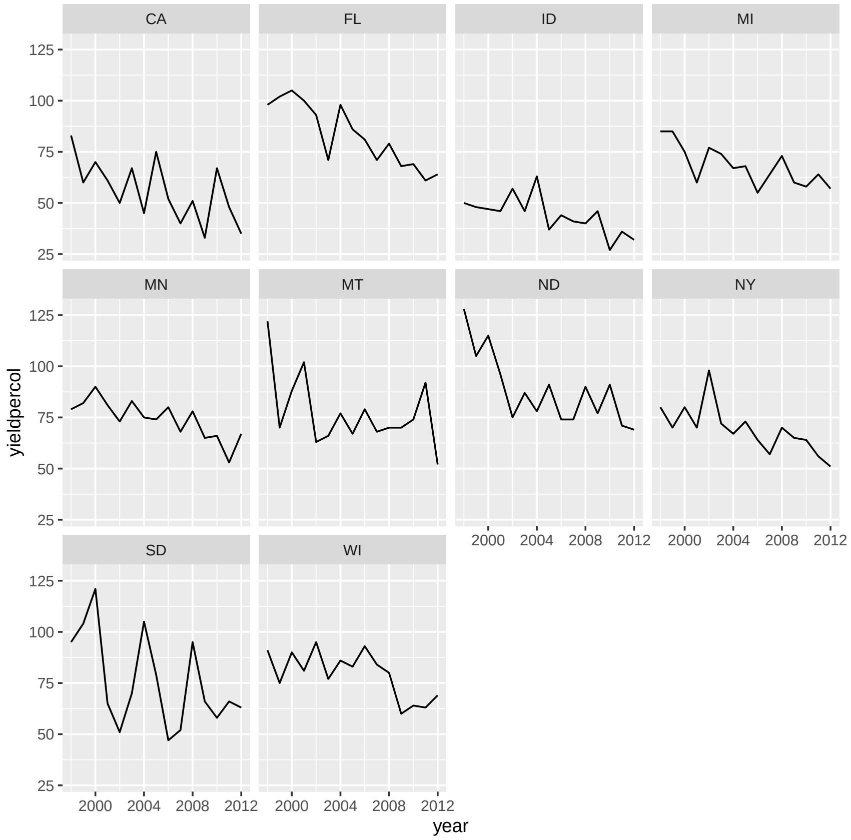
4. We now plot the vertical bar diagram for the yields per colony over the period of 15 years from 1998 to 2012 for each ofthe 10 states of the USA with the maximum absolute change in the production of honey.

```
data %>%
        filter(state %in% as.character(as.array(imp_states$state))) %>%
        ggplot(aes(x = as.factor(year), weight = yieldpercol,fill = year))+
                geom_bar() +
                facet_wrap(~ state)+
                theme(axis.text.x = element_blank())
```



5. We can see that the data here is based over time. So we do a time plot of the yields per colony over the period of 15 years from 1998 to 2012 for each ofthe 10 states of the USA with the maximum absolute change in the production of honey.

```
data %>%
        filter(state %in% as.character(as.array(imp_states$state))) %>%
        ggplot(aes(year, yieldpercol))+
                geom_line() +
                facet_wrap(~ state)
```



Links

1. Dataset
2. R Shiny Link
3. YouTube Link
4. Github Link