Honey Production in the US

Module PSY6422

18 July 2024

Data Origins

The raw data for this project was downloaded from Kaggle and was originally created by Mohit Poudel. The units of measure for honey production are in pounds (lbs). More information about each of the other variables can be found in the codebook_honey_prod.docx file.

Research Question

Global demand for honey is ever increasing, as the population grows and more people strive to eat healthier and more organically (Garcia et al., 2018). Bee populations on the other hand, are decreasing due to several factors including loss of habitat and destruction of wildlife. Even colonies of managed bees, i.e. those used commercially for honey production, are declining due to damaging management practices used in beekeeping (Panziera et al., 2022).

My plots aim to visualise how honey production has changed over a few years using one of the largest population of consumers and producers, the US. Based upon this knowledge, it is expected that there will be a decrease in production as years progress.

Data Preparation

The code below shows my steps for prepping my data and data wrangling. This includes the setup for loading the required libraries and a sample of the raw data. During the process I created a new column for the dataset of state name abbreviations, so that my final plots would be easier to read. In doing this I discovered some states were not consistently present through the dataset and some were not present at all. The missing states are shown below and were removed for consistency. The raw data was processed so that only the relevant columns needed for my plots were extracted, this left Productions, Year and State abbreviations.

```
#libraries required for running the data
library(tidyverse)
library(readr)
library(here)
library(tinytex)
library(dataMaid)
library(magrittr)
library(dplyr)
library(ggstar)
library(extrafont)
```

```r
#setting relative file paths
here::i_am("Honey Production markdown.Rmd")

data_dir <- here("data")
figs_dir <- here("figs")
raw_dir <- here("raw")
```

```r
honey_prod <- read_csv("raw/US_honey_production_dataset.csv", show_col_types = FALSE)
head(honey_prod)
```

```r
#creating state abbreviations for the state names, to make it look neater when they are written on the
all_states <- c( "Alabama", "Alaska", "Arizona", "Arkansas", "California",
                 "Colorado", "Connecticut", "Delaware", "Florida", "Georgia",
                 "Hawaii", "Idaho", "Illinois", "Indiana", "Iowa", "Kansas",
                 "Kentucky", "Louisiana", "Maine", "Maryland",
                 "Massachusetts", "Michigan", "Minnesota", "Mississippi",
                 "Missouri", "Montana", "Nebraska", "Nevada",
                 "NewHampshire", "NewJersey", "NewMexico", "NewYork",
                 "NorthCarolina", "NorthDakota", "Ohio", "Oklahoma",
                 "Oregon", "Pennsylvania", "RhodeIsland", "SouthCarolina",
                 "SouthDakota", "Tennessee", "Texas", "Utah",
                 "Vermont", "Virginia", "Washington", "WestVirginia",
                 "Wisconsin", "Wyoming")

state_abbr <- c(
  "Alabama" = "AL", "Arizona" = "AZ", "Arkansas" = "AR",
  "California" = "CA", "Colorado" = "CO", "Florida" = "FL", "Georgia" = "GA", "Hawaii" = "HI", "Idaho" =
  "Illinois" = "IL", "Indiana" = "IN", "Iowa" = "IA", "Kansas" = "KS",
  "Kentucky" = "KY", "Louisiana" = "LA", "Maine" = "ME", "Michigan" = "MI", "Minnesota" = "MN",
  "Mississippi" = "MS",
  "Missouri" = "MO", "Montana" = "MT", "Nebraska" = "NE", "NewJersey" = "NJ", "NewYork" = "NY",
  "NorthCarolina" = "NC", "NorthDakota" = "ND", "Ohio" = "OH", "Oregon" = "OR", "Pennsylvania" = "PA",
  "Vermont" = "VT", "Virginia" = "VA", "Washington" = "WA", "WestVirginia" = "WV",
  "Wisconsin" = "WI", "Wyoming" = "WY")

#checking which states are missing across the data and filtering the dataset to remove any state that i
honey_cleaned <- honey_prod %>%
  group_by(state) %>%
  filter(n() >= 12)
missing_states <- setdiff(all_states, honey_cleaned$state)
missing_states
```

```
##  [1] "Alaska"        "Connecticut"   "Delaware"      "Maryland"      "Massachusetts" "Nevada"
##  [7] "NewHampshire"  "NewMexico"     "Oklahoma"      "RhodeIsland"   "SouthCarolina"
```

```r
#a tidier data set with a new column for state abbreviations
honey_cleaned <- honey_cleaned %>%
  mutate(state_abbr = state_abbr[state])

#View the result
head(honey_cleaned)
```

```
#export cleaned data
write.csv(honey_cleaned, file= here("data","cleaned_honey_data.csv"))
```

```
# change the data from being grouped by state to being grouped by new column, state_abbr
honey_cleaned %>%
  ungroup() %>%
  group_by(state_abbr) %>%

#select only the "state_abbr","productions" and "year" columns and show data only from years 2019-2021
  select(state_abbr, productions, year) %>%
  filter(year > 2018)

#define the data frame to the above criteria
df <- honey_cleaned %>%
  ungroup() %>%
  group_by(state_abbr) %>%
  select(state_abbr, productions, year) %>%
  filter(year > 2018)

#order the column "state_abbr" alphabetically
df[order(df$state_abbr),]

#move column "productions" to last position
df %>%
  select(-productions, productions)
```

```
#finding the min and max values of productions
summary(df$productions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   44000  110000  261000  879752  844000 8668000
```

```
#view the data frame that will be used for the plots
head(df)
```

Visualisation

For my visualisation, I created two graphs, one which shows each year of production individually, and the other which combines them.

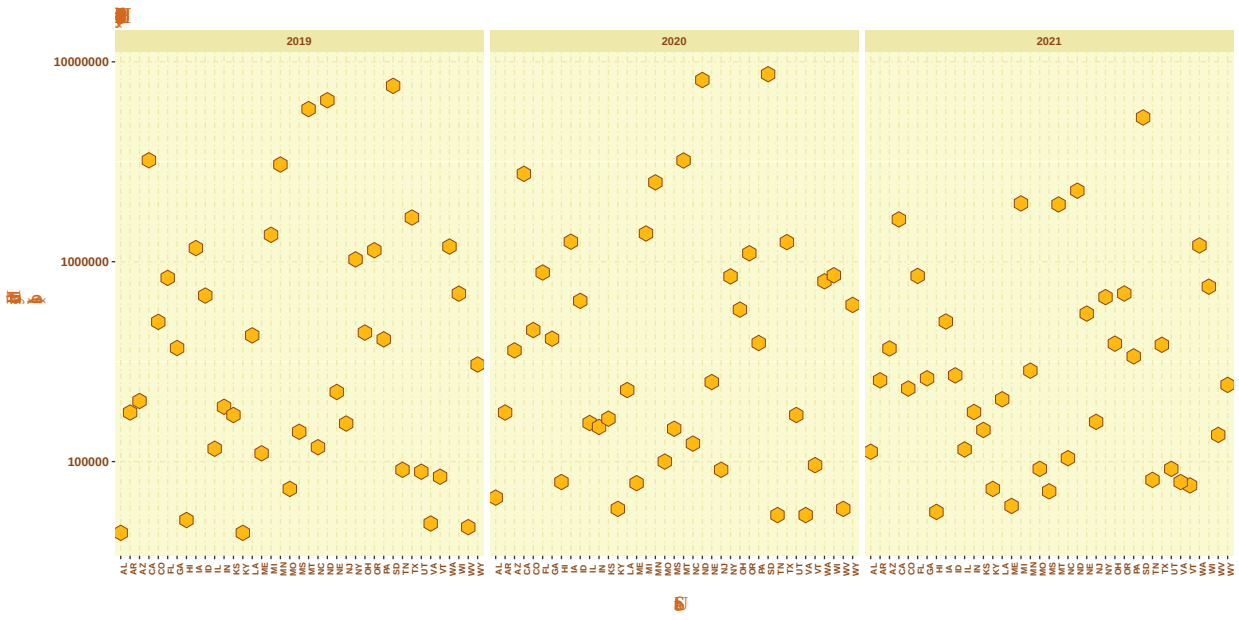
I felt that the first graph accurately portrayed a general pattern in production through the years across the whole country. However, it did not display inter/intra state differences well. Thus i created the second graph, which I believe presents these differences more clearly.

For both of these graphs I chose to use a Log scale for the Y axis, due to the considerable difference between the min and max values of honey production, this in itself due to some states having much fewer bee colonies than others.

Plot 1

```
#making a scattergraph
graph <- ggplot(data=df, aes(x = state_abbr, y = productions)) +
```

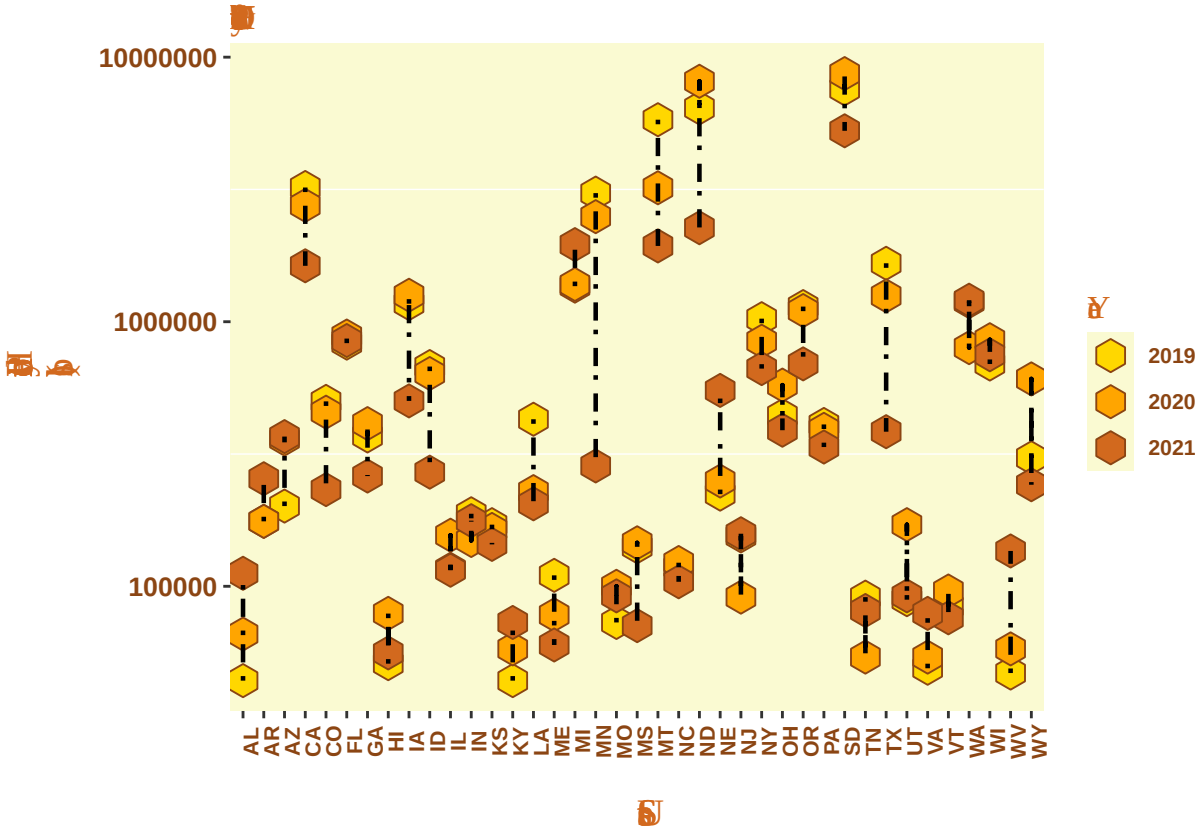
```
#changing the size, shape and colour of the plot points
  geom_star(starshape = 6, size = 4.5, fill= "darkgoldenrod1", colour = "chocolate4") +
  labs(x = "US States", y = "Honey Production\n(lbs)", title = "Honey Production in the US (2019-2021)"

#using facet wrap to compare all three years
  facet_wrap(~factor(year))+

#changing the appearance of the axes and title text
  theme(axis.text.x = element_text (size = 7, face = "bold", angle = 90, colour = "chocolate4"),
        axis.text.y = element_text(size= 10, face= "bold", colour = "chocolate4"),
        axis.title.y = element_text(margin = margin(0, 10, 0, 0)),
        axis.title.x = element_text(margin = margin(10, 0, 0, 0)),
        axis.title = element_text(family= "Times New Roman", face = "plain", size = 16, colour = "choco
        plot.title = element_text(size = 18, family= "Times New Roman", colour = "chocolate"),

#changing the appearance of the graph background
        panel.background = element_rect(fill = "lightgoldenrodyellow"),
        panel.grid.major = element_line(linewidth = 0.5, linetype = '1342', colour = 'palegoldenrod'),

        strip.background = element_rect(fill = "palegoldenrod"),
        strip.text = element_text(face = "bold", colour = "chocolate4"))

#making a logorithmic y axis scale
loggraph <- graph + scale_y_log10()
print(loggraph)
```




```
#saving the graph as a png
ggsave(filename = here(figs_dir, "initial_vis_210195998.png"))
```

Plot 2

```r
#defining custom colours
custom_colours <- c("2019" = "gold",
                    "2020" = "orange",
                    "2021" = "chocolate")

#making a graph with all three years combined
combinedgraph <- ggplot(data=df, aes(x = state_abbr,y = productions, fill =as.factor(year),)) +
  geom_star(starshape = 6, size = 4.5, colour = "chocolate4") +
  geom_line(aes(group = state_abbr), size= 0.8, colour = "black", linetype = '1342') +
  labs(x = "US States", y = "Honey Production\n(lbs)", fill = "Year", title = "The Difference in Honey
  scale_fill_manual(values = custom_colours) +

#changing the appearance of the axis and title text
  theme(axis.text.x = element_text (size = 8, face = "bold", angle = 90, colour = "chocolate4"),
        axis.text.y = element_text(size= 10, face= "bold", colour = "chocolate4"),
        plot.title = element_text(family= "Times New Roman", colour = "chocolate"),

#changing the appearance of the graph background
        panel.background = element_rect(fill = "lightgoldenrodyellow"),
        panel.grid.major = element_blank(),
#changing the appearance of the axes and legend
        axis.title = element_text(family= "Times New Roman", face = "plain", size = 14, colour = "chocol
        legend.title = element_text(family= "Times New Roman", size = 12, face= "plain", colour = "choc
        legend.text = element_text(size = 8, face = "bold", colour = "chocolate4") ,
        axis.title.y = element_text(margin = margin(0, 10, 0, 0)),
        axis.title.x = element_text(margin = margin(10, 0, 0, 0)))

#changing the y axis to a logarithmic scale
logcombinedgraph <- combinedgraph + scale_y_log10()
  print(logcombinedgraph)
```




```
ggsave(filename = here(figs_dir, "key_vis_210195998.png"))
```

Interpretation

The first of my two plots does show a trending pattern of decline over the three years. Possibly due to population decline of honey bees, this is backed up by literature suggesting that the US cannot keep up with the demand for honey with their own domestic supply, and that honey import in the US is constantly rising each year (Garcia et al., 2018). This in itself has led to other problems such as a phenomenon referred to as honey fraud, in which honey not meeting food regulation standard is being put onto the market.

Interestingly though, in my second graph it shows both increase and decrease in production. Generally, it seems as though the states producing the most amount of honey have decreased in production through the years whereas those who produced less to start with increased their production over time. Of course other factors may have contributed to the stats, both 2020 and 2021 were affected by the global pandemic which had an impact upon most consumption and production behaviours.

Summary

Overall I am pleased with my plots, I believe I presented them in a way which mirrored the topic of choice in a visually appealing manner. For example, my colour scheme chosen to mimic the colour of honey, the hexagonal shape of the data points to mimic honeycomb and even the grid lines on the first graph and the difference lines for the intra state changes in the second graph, which were chosen to emulate the appearance of bees flight paths.

If I carried out this project again I think I would aim to be more ambitious in the plots themselves, plotting data over all of the years for example, and perhaps creating interactive graphs. Originally I had wanted

to create a graph similar to my second graph but with an interactive tooltip feature that showed the state name, year and production amount in full when hovered over. Frustratingly, after trying for some time to do this I realised it would not work as plotly is not compatible with geom_star and I did not think the sacrifice of that was worth this element.I had also looked at mapping the data to a map of the US which I think would have been a good addition to my project.