Honey Bees Colonies in the US

dweorh

2023-02-23

Business requirement

A software company is interested in writing a software for beekeepers to help them manage colonies and apiaries.

The company needs to know how big and stable the market is.

Data source

For the analysis survey data collected by United States Department of Agriculture between 2012 and 2022 will be used.

Project setup

data file: USDA_NASS_honey_bees_2012_2022_survey.csv

```
# install.packages('tidyverse')
library(tidyverse)
# install.packages("dplyr")
library("dplyr")
# install.packages("janitor")
library("janitor")
# install.packages("scales")
library("scales")
# install.packages("RColorBrewer")
library("RColorBrewer")

# color pallete
qual_col_pals = brewer.pal.info[brewer.pal.info$category == 'qual',]
col_vector = unlist(mapply(brewer.pal, qual_col_pals$maxcolors, rownames(qual_col_pals)))

expected_items <- c('ADDED & REPLACED',
                    'INVENTORY, MAX',
                    'LOSS, COLONY COLLAPSE DISORDER',
                    'LOSS, DEADOUT')

expected_periods <- c(1:4)
```

Data cleanup

```
survey <- read_csv('USDA_NASS_honey_bees_2012_2022_survey.csv')

# select only needed columns from the set
# normalize names
# remove redundant data
# convert text values to less bug prone and saving resources integer values
survey_reduced <- survey %>%
  clean_names() %>%
  select(year, period, state, data_item, value) %>%
  mutate(data_item = str_replace(data_item, 'HONEY, BEE COLONIES - ', '')) %>%
  mutate(data_item = str_replace(data_item, ', MEASURED IN COLONIES', '')) %>%
  mutate(period = str_replace(period, 'JAN THRU MAR', '1')) %>%
  mutate(period = str_replace(period, 'APR THRU JUN', '2')) %>%
  mutate(period = str_replace(period, 'JUL THRU SEP', '3')) %>%
  mutate(period = str_replace(period, 'OCT THRU DEC', '4')) %>%
  mutate(period = as.numeric(period))

# for further use save cleanup data
write_csv(survey_reduced, 'USDA_NASS_honey_bees_2012_2022_survey_reduced.csv')

# remove unused variables and release memory
rm(survey)
gc()
```

```
##           used (Mb) gc trigger (Mb) max used  (Mb)
## Ncells 1217789 65.1    2320864  124  2320864 124.0
## Vcells 2095971 16.0    8388608   64  3423300  26.2
```

```
# check if there is correct number of periods (4) for every year and every state
survery_missing_periods <- survey_reduced %>%
  select(year,state,period) %>%
  group_by(year,state,period) %>%
  distinct() %>%
  group_by(year,state) %>%
  summarize(periods=n()) %>%
  filter(periods < length(expected_periods))

# if there are any missing periods for any year,
# we will exclude such a year from the further calculations
survey_missing_periods_years <- as.vector(t(survery_missing_periods %>% select(year) %>% distinct()))
```

Check if we have equal number of rows for each 'Data Item' type.

```
survey_reduced %>% group_by(data_item) %>% summarize(total_count=n())
```

```
## # A tibble: 4 x 2
##   data_item                      total_count
##   <chr>                                <int>
## 1 ADDED & REPLACED                      1334
```

```
## 2 INVENTORY, MAX                          1334
## 3 LOSS, COLONY COLLAPSE DISORDER           588
## 4 LOSS, DEADOUT                           1334
```

Many 'LOSS, COLONY COLLAPSE DISORDER' data is missing.

Chart with totals, should show us if that's an issue.

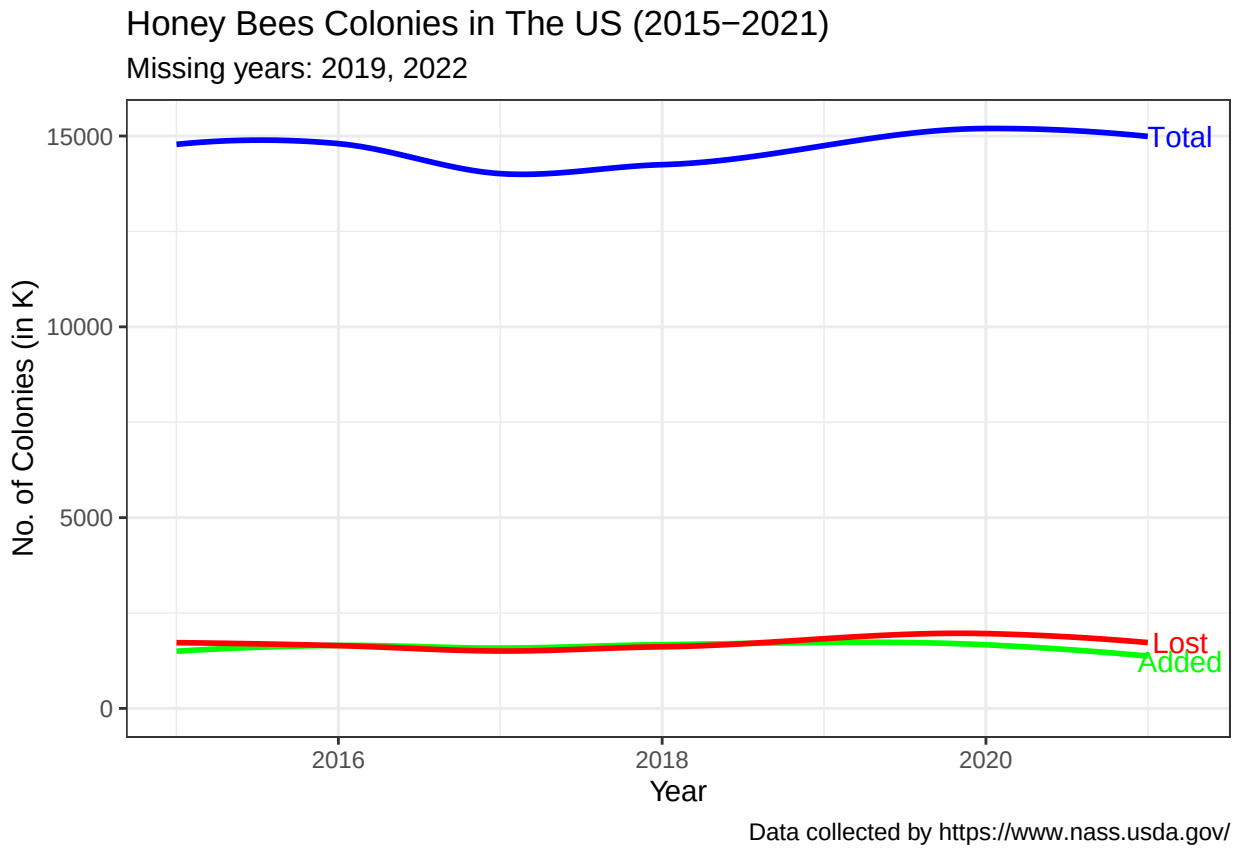
As this is a survey, in some states in some years 'losses' might be merged into one category.

```
survey_totals_per_state <- survey_reduced %>%
  filter(!year %in% survey_missing_periods_years) %>%
  filter(data_item == 'INVENTORY, MAX') %>%
  group_by(year,state) %>%
  summarize(total=sum(value))

survey_totals <- survey_reduced %>% group_by(year,data_item) %>% summarize(total = sum(value))

# a function to calculate totals of certain type
add_loss_total <- function(type = 'add') {
  data <- pick(data_item, total)
  # data <- cur_data()
  if (type == 'add') {
    res <- sum(data[data$data_item == 'ADDED & REPLACED',]$total)
  } else if (type == 'loss') {
    sub_data <- data[data$data_item %in% c('LOSS, DEADOUT', 'LOSS, COLONY COLLAPSE DISORDER'),]
    res <- sum(sub_data$total)
  } else if (type == 'inventory') {
    res <- sum(data[data$data_item == 'INVENTORY, MAX',]$total)
  }
  return(res)
}

survey_totals_add_loss <- survey_totals %>%
  filter(!year %in% survey_missing_periods_years) %>%
  group_by(year) %>%
  summarize(
    loss = add_loss_total('loss') / 1000,
    add = add_loss_total('add') / 1000,
    inventory = add_loss_total('inventory') / 1000
  )

totals_pivot <- pivot_longer(survey_totals_add_loss, cols = c(loss,add,inventory))

survey_last_valid_year = max(survey_totals_add_loss$year)

last_year_data <- totals_pivot %>% filter(year == survey_last_valid_year)
last_year_data_states <- survey_totals_per_state %>% filter(year == survey_last_valid_year)
```

A chart to present a general country level overview of 'add', 'loss', and 'inventory'.

```
labels <- list(add = "Added", loss = "Lost", inventory = "Total")
colors <- list(add = "green", loss = "red", inventory = "blue")
```

```
chart_title <- paste0("Honey Bees Colonies in The US (",min(survey_reduced['year']),"-",survey_last_vali
chart_subtitle <- paste0("Missing years: ", paste(survey_missing_periods_years, collapse=", "))

ggplot(data=totals_pivot) +
  geom_smooth(mapping=aes(x=year, y=value, color=name)) +
  labs(title=chart_title,
       subtitle = chart_subtitle,
    caption="Data collected by https://www.nass.usda.gov/",
     x='Year', y='No. of Colonies (in K)') +
  expand_limits(y=0) +
  geom_text(
    data = last_year_data,
    aes(label = labels[name], x = year + 0.2, y = ifelse(last_year_data$name=='add', value*0.9, value),
    parse = TRUE
  ) +
  scale_color_manual(name="", labels = labels, values = colors) +
  theme_bw() +
  theme(legend.position = "none")
```



On the country level the total inventory stays more or less stable in 10y period.

The same with adds and losses.

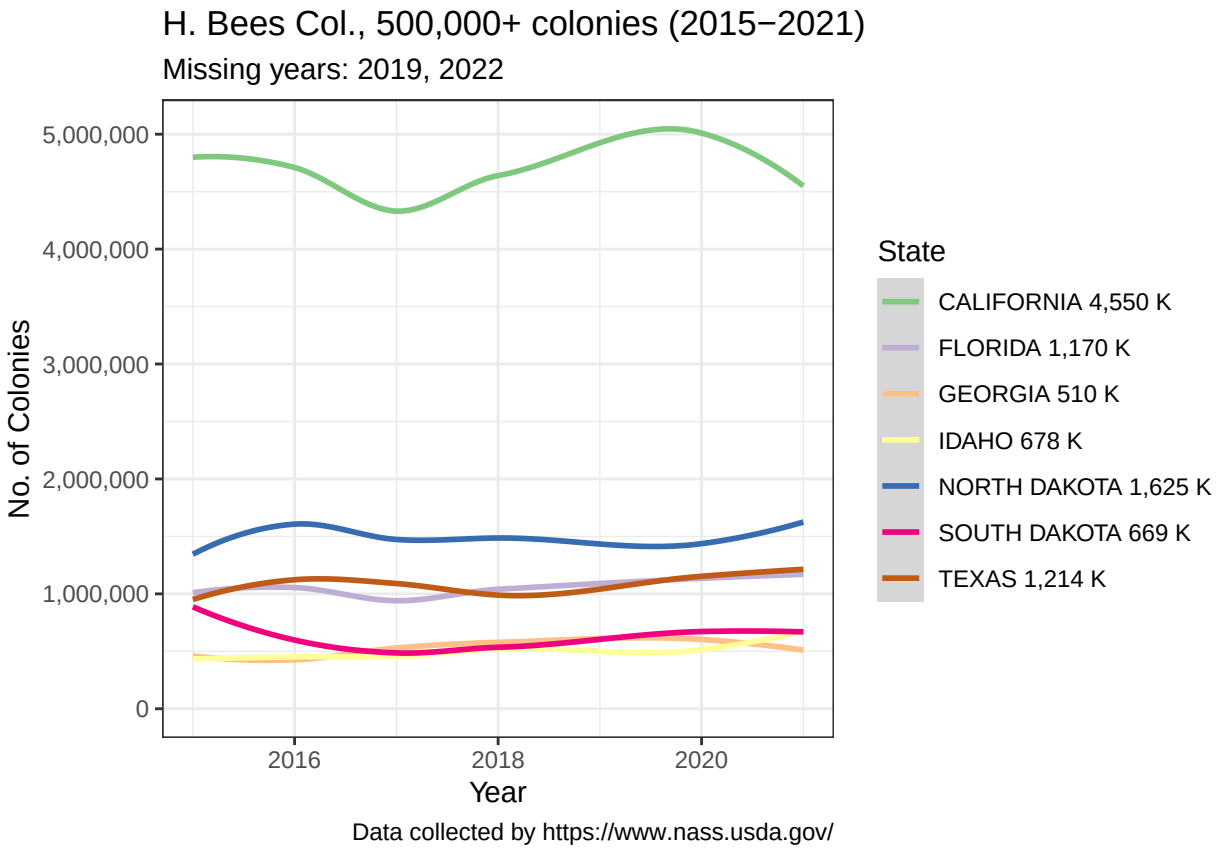
Detailed charts

To make charts more readable, states were divided, arbitrarily, into 4 categories:

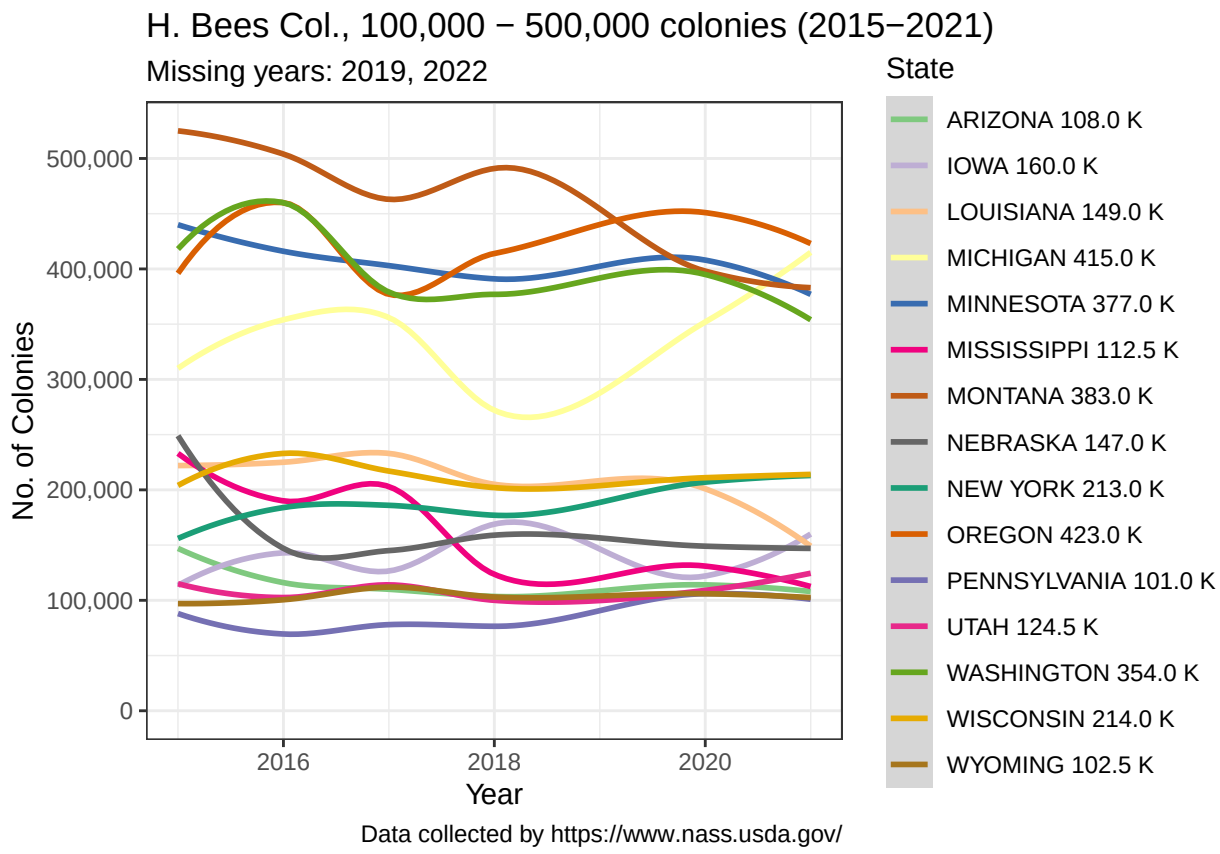
- up to 50,000 colonies
- 50,000 - 100,000 colonies
- 100,000 - 500,000 colonies
- 500,000 colonies and more

```
v_small_states = as.vector(
  t(
    survey_totals_per_state %>%
      filter(year == 2021 & total < 50000) %>%
        ungroup() %>% select(state)
    )
  )
small_states = as.vector(
  t(
    survey_totals_per_state %>%
      filter(year == 2021 & total >= 50000 & total < 100000) %>%
      ungroup() %>% select(state)
    )
  )
mid_states = as.vector(
  t(
    survey_totals_per_state %>%
      filter(year == 2021 & total >= 100000 & total < 500000) %>%
      ungroup() %>% select(state)
    )
  )
big_states = as.vector(
  t(
    survey_totals_per_state %>%
      filter(year == 2021 & total >= 500000) %>%
      ungroup() %>% select(state)
    )
  )
```

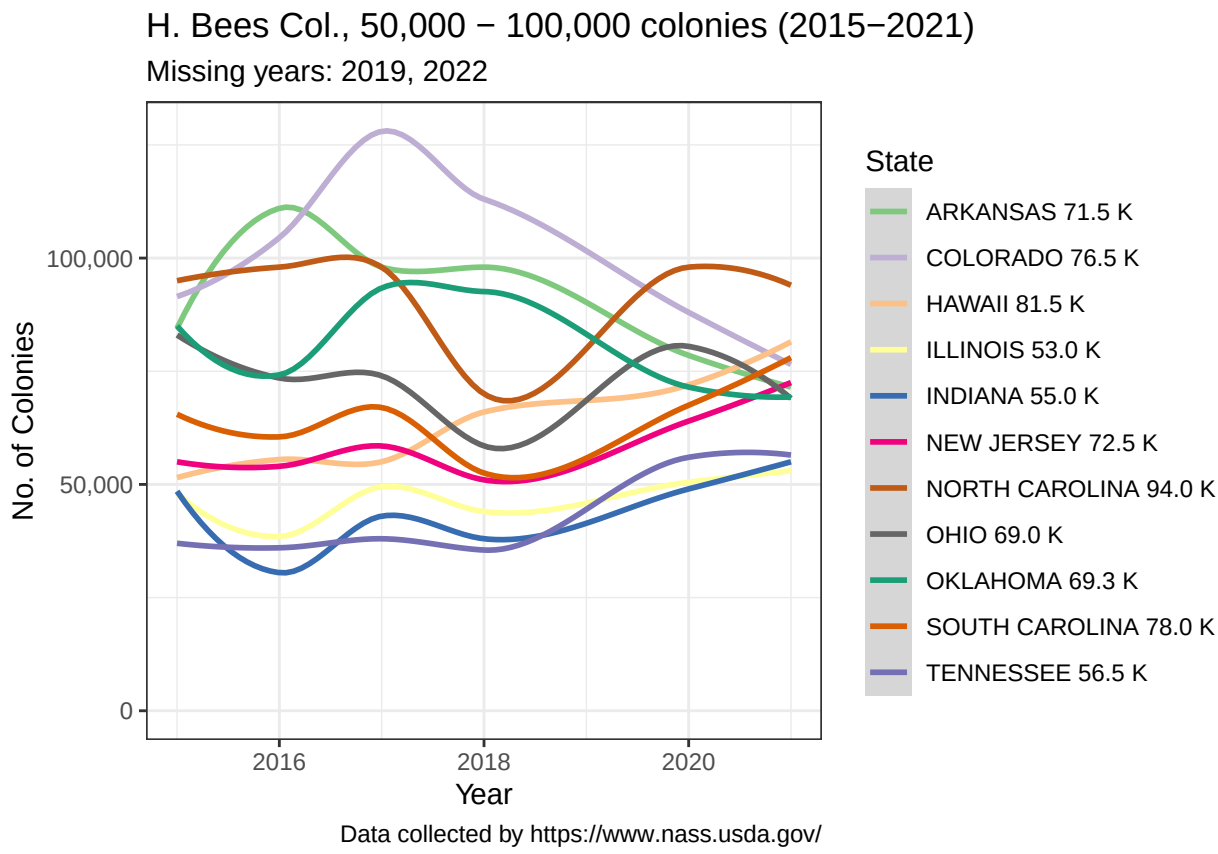
```
chart_big_title <- paste0("H. Bees Col., 500,000+ colonies (",min(survey_reduced['year']),"-",survey_las

plot_data <- survey_totals_per_state %>% filter(state %in% big_states)
ggplot(data=plot_data) +
  geom_smooth(mapping=aes(x=year, y=total, color=state)) +
  labs(title=chart_big_title,
       subtitle = chart_subtitle,
       caption="Data collected by https://www.nass.usda.gov/",
       x='Year', y='No. of Colonies') +
  expand_limits(y=0) +
  theme_bw() +
  scale_color_manual(name="State",
                     values = col_vector,
                     labels = paste(plot_data$state,
                                    scales::comma(plot_data[plot_data$year == survey_last_valid_year,]$
                                    'K')
                     ) +
  scale_y_continuous(labels = scales::comma)
```



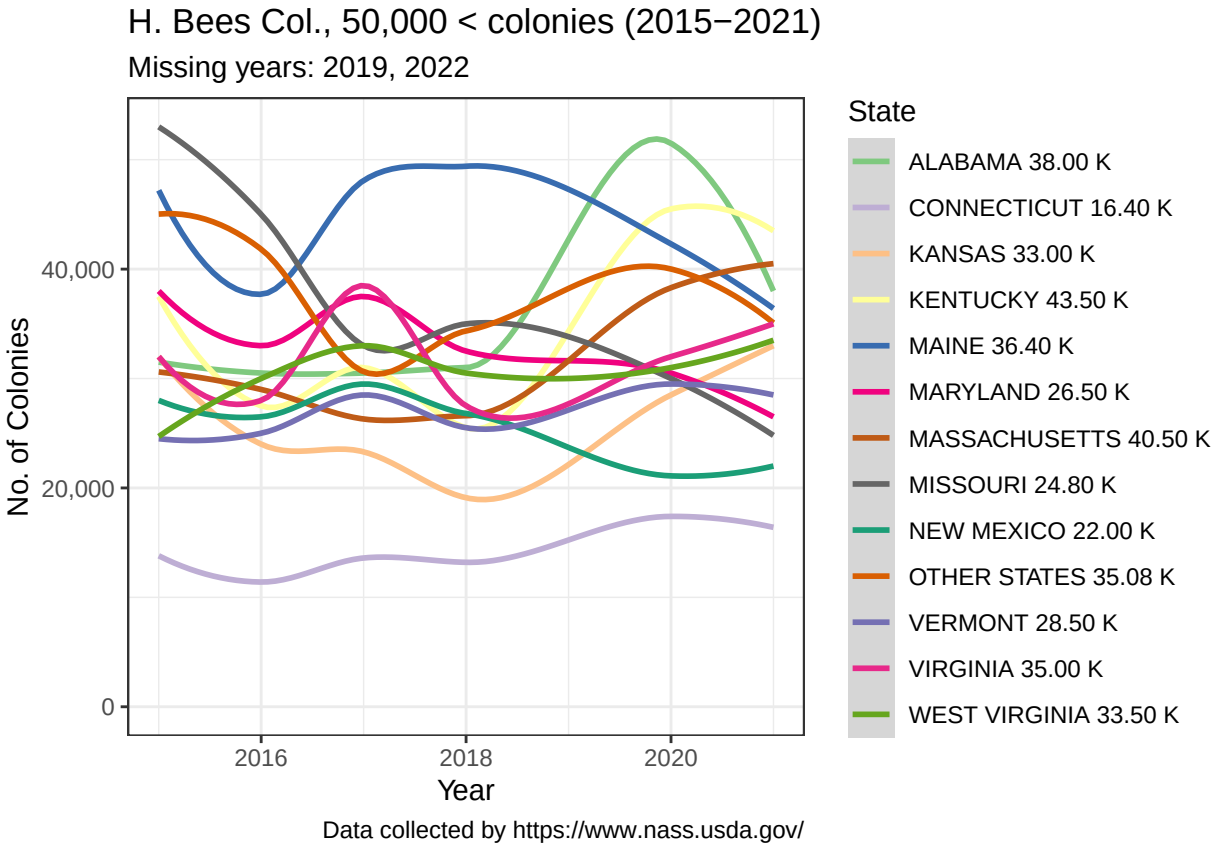

```
chart_mid_title <- paste0("H. Bees Col., 100,000 - 500,000 colonies (",min(survey_reduced['year']),"-",s

plot_data <- survey_totals_per_state %>% filter(state %in% mid_states)
ggplot(data=plot_data) +
  geom_smooth(mapping=aes(x=year, y=total, color=state)) +
  labs(title=chart_mid_title,
       subtitle = chart_subtitle,
       caption="Data collected by https://www.nass.usda.gov/",
       x='Year', y='No. of Colonies') +
  expand_limits(y=0) +
  theme_bw() +
  scale_color_manual(name="State",
                     values = col_vector,
                     labels = paste(plot_data$state,
                                    scales::comma(plot_data[plot_data$year == survey_last_valid_year,]$
                                    'K')
  ) +
  scale_y_continuous(labels = scales::comma)
```




```
chart_small_title <- paste0("H. Bees Col., 50,000 - 100,000 colonies (",min(survey_reduced['year']),"-"

plot_data <- survey_totals_per_state %>% filter(state %in% small_states)
ggplot(data=plot_data) +
  geom_smooth(mapping=aes(x=year, y=total, color=state)) +
  labs(title=chart_small_title,
       subtitle = chart_subtitle,
       caption="Data collected by https://www.nass.usda.gov/",
       x='Year', y='No. of Colonies') +
  expand_limits(y=0) +
  theme_bw() +
  scale_color_manual(name="State",
                     values = col_vector,
                     labels = paste(plot_data$state,
                                    scales::comma(plot_data[plot_data$year == survey_last_valid_year,]$
                                    'K')
  ) +
  scale_y_continuous(labels = scales::comma)
```




```
chart_v_small_title <- paste0("H. Bees Col., 50,000 < colonies (",min(survey_reduced['year']),"-",survey

plot_data <- survey_totals_per_state %>% filter(state %in% v_small_states)
ggplot(data=plot_data) +
  geom_smooth(mapping=aes(x=year, y=total, color=state)) +
  labs(title=chart_v_small_title,
       subtitle = chart_subtitle,
       caption="Data collected by https://www.nass.usda.gov/",
       x='Year', y='No. of Colonies') +
  expand_limits(y=0) +
  theme_bw() +
  scale_color_manual(name="State",
                     values = col_vector,
                     labels = paste(plot_data$state,
                                    scales::comma(plot_data[plot_data$year == survey_last_valid_year,]$
                                    'K')
  ) +
  scale_y_continuous(labels = scales::comma)
```



Conclusions

1. Honey Bees market looks stable across 10y, according to surveys.
2. The biggest beekeeping market is in California with ca. 30% of total US H. Bees colonies.
3. That market neither grows nor shrinks, so competition can be hard.
4. Further research about already existing software for beekeepers is needed.