615_Assignment3_Honey

Dave Anderson, Sky Liu, Tingrui Huang, Xiang Xu

October 3, 2018

Introduction

The struggle of the honeybee population in our country is a well-known idea. The public understands the importance of the presence of honeybees for our own existence, and that the population of bees has seen a dramatic decrease. The objectives of our exploration of honey data were to understand where honey is commercially produced in the United States and what patterns, if any, we could find in the production over time.

Data

Our data source for this exploration was the National Agricultural Statistics Service site using their 'quick stats' tool[1]. We chose to look at the survey data on honey, which included more years than the census. We also chose the variables like 'Bee Colonies', 'Loss', 'Deadout', 'Production in lbs.', 'Production per Colony', and 'Price'.

```
#Loading data files
Honey <- read_csv("Honey.csv", col_types = cols(
  Year = col_integer(),`State ANSI` = col_integer(),
  watershed_code = col_integer(),Value = col_number()))
Deadout <- read_csv("Deadout.csv",col_types = cols(
  Year = col_integer(),`State ANSI` = col_integer(),
  watershed_code = col_integer(),Value = col_number()))
Price_per_lb <- read_csv("Price per lb.csv",col_types = cols(
  Year = col_integer(),`State ANSI` = col_integer(),
  watershed_code = col_integer(),Value = col_double()))
Production_per_Colony <- read_csv("Production per Colony.csv",col_types = cols(
  Year = col_integer(),`State ANSI` = col_integer(),
  watershed_code = col_integer(),Value = col_double()))
Honey_value_annual <- read_csv('Honey_value.csv', col_types = cols(
  Year = col_integer(),Value = col_double()))
CPI <- read_csv( '1987_2017CPI.csv', col_types = cols(
  Year = col_integer(),CPI = col_double()))
honey_loss_dt <- read_csv('Honey_Loss_6_States.csv', col_types = cols(
  Year = col_integer(),Value = col_number()))
```

Data Cleaning

In cleaning the dataset, we often excluded the 'US TOTAL' because we were interested in analyzing different states. There were some missing values for various variables, and some states were only included in a few years of survey data, but because our focus would be drawn to the six top-producing states, our conclusions were not affected by these inconsistencies. We combined the sets of data into a usable data frame and grouped

the data in various ways to form our visualizations. We also added three new variables; number of colonies per state and colonies lost per total colonies, and adjusted price value.

```
#sort each data set to variables we want.
Honey <- dplyr::select(Honey,Year,State,Value)
Deadout <- dplyr::select(Deadout,Year,Period,State,Value)
Price_per_lb <- dplyr::select(Price_per_lb,Year,State,Value)
Production_per_Colony <- dplyr::select(Production_per_Colony,Year,State,Value)
Honey_value_annual <- dplyr::select(Honey_value_annual,Year,Value)

#Filter out totals, group each variable by state, average values from each year
Production <- filter(Honey, State != "US TOTAL") %>%
  group_by(State) %>% summarise(Average_production = mean(Value)/2000)
Loss <- filter(Deadout, State != "US TOTAL") %>%
  group_by(State) %>% summarise(Average_loss = mean(Value))
Price <- filter(Price_per_lb, State != "US TOTAL") %>%
  group_by(State) %>% summarise(Average_price = mean(Value))
Colony_production <- filter(Production_per_Colony, State != "US TOTAL") %>%
  group_by(State) %>% summarise(Average_per_colony = mean(Value))

#Combine into one set. Add new variables to show number of colonies and loss/colony
Honey_by_State <- full_join(Production, Loss, by = "State")
Honey_by_State <- full_join(Honey_by_State, Price, by = "State")
Honey_by_State <- full_join(Honey_by_State, Colony_production, by = "State") %>%
  mutate(Colonies = Average_production*2000/Average_per_colony) %>%
  mutate(Loss_per_colony = Average_loss/Colonies)

# pick 6 top states with highest production and complete data
Honey_State <- Honey %>%  group_by(State)
unique(Honey$State)
Honey_sixstate <- Honey_State %>%
  filter(State %in%
  c("CALIFORNIA", "FLORIDA" ,"SOUTH DAKOTA", "NORTH DAKOTA","MONTANA", "MINNESOTA")) %>%
  arrange(State, Year)

# pick the 6 top states mentioned above to analyze the price trend
Price_state <- Price_per_lb %>%  group_by(State)
unique(Price_state$State)
Price_sixstate <- Price_state %>%
  filter(State %in%
  c("CALIFORNIA", "FLORIDA" ,"SOUTH DAKOTA", "NORTH DAKOTA","MONTANA", "MINNESOTA")) %>%
  arrange(State, Year)

#Honey lost in 6 states
#Sum by year
#(Since we only have the data in 1st & 2nd quarter 2018, we will exclude 2018 data)
honey_loss_dt$Value <- as.numeric(gsub(",","",honey_loss_dt$Value))
honey_2017 <- honey_loss_dt %>% select(Year,State,Value) %>% filter(Year==2017) %>%
  group_by(Year,State) %>% summarise(total=sum(Value))
honey_2016 <- honey_loss_dt %>% select(Year,State,Value) %>% filter(Year==2016) %>%
  group_by(Year,State) %>% summarise(total=sum(Value))
honey_2015 <- honey_loss_dt %>% select(Year,State,Value) %>% filter(Year==2015) %>%
  group_by(Year,State) %>% summarise(total=sum(Value))
```

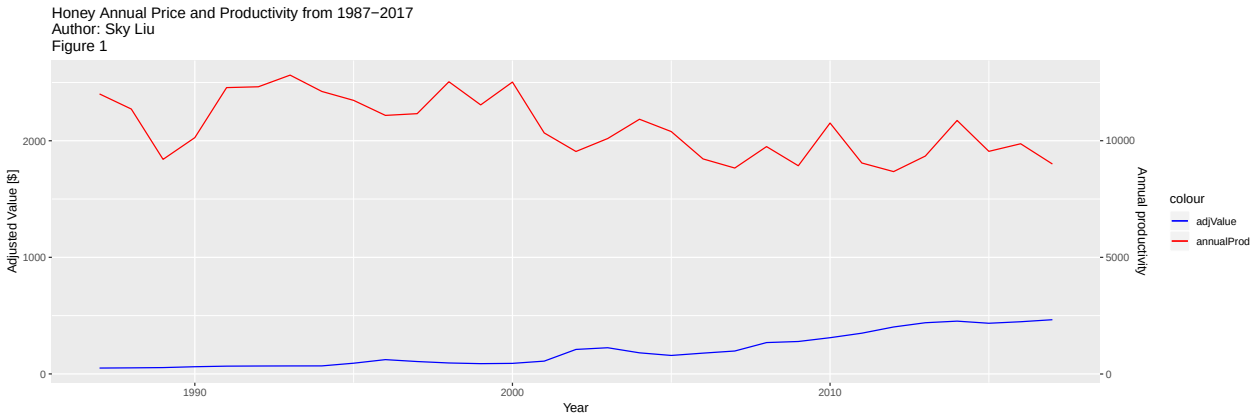
```r
# Total loss from 2015-2017
honey_total <- rbind(honey_2017,honey_2016,honey_2015)

#Adjust the annual honey value (price received) by 1987 inflation rate.
baseCPI <- rep(113.6, 21)
adjusted_Price <- as.data.frame(Honey_value_annual$Value * (CPI$CPI / baseCPI))
Honey_value_annual <- cbind(Honey_value_annual,adjusted_Price)
names(Honey_value_annual) <- c('Year','Value','adjValue')
#Add annual productivity
Annual_production <- filter(Honey, State != "US TOTAL") %>% group_by(Year)  %>%
  summarise(Average_production = mean(Value)/2000)
Annual_production <- arrange(Annual_production, desc(Year))
Honey_value_annual <- cbind(Honey_value_annual,Annual_production$Average_production)
names(Honey_value_annual) <- c('Year','Value','adjValue','annualProd')
```

Plots and Findings

Our results were both surprising and consistent with previous assumptions.

First, we discovered that honey production has seen an overall decrease over the last 30 years. Based on the law of supply and demand, we expect to see that the price of honey to be increasing along with the decreasing in production. In order to factor out the influence of inflation, we adjusted the price by using the price 1987 as a baseline. The graph proved our assumption that the price of honey has increased over the same period. See Figure 1.

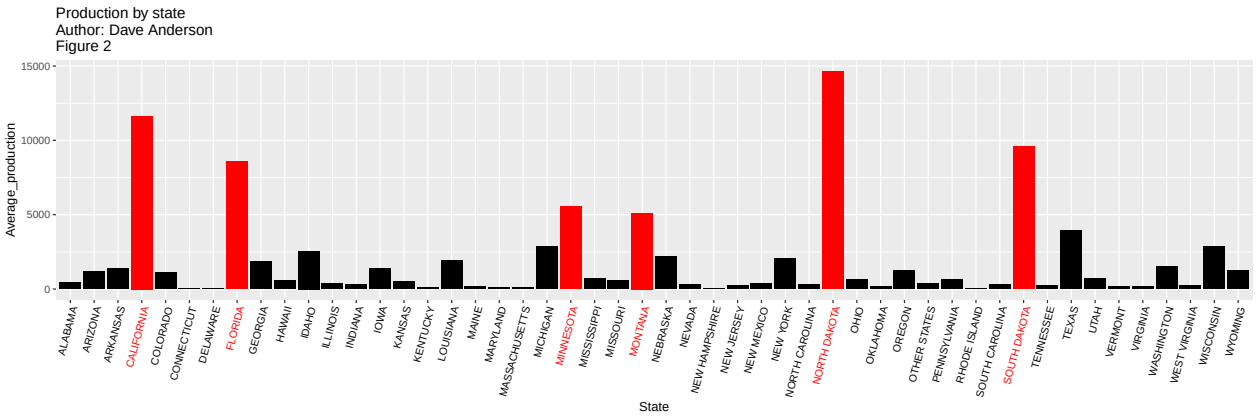
```r
#Honey annual price and productivity from 1987-2017
ggplot(data = Honey_value_annual, aes(x = Year)) +
  geom_line(aes(y = adjValue, colour = "adjValue")) +
  geom_line(aes(y = annualProd, colour = "annualProd")) +
  ggtitle(
    "Honey Annual Price and Productivity from 1987-2017 \nAuthor: Sky Liu \nFigure 1") +
  scale_y_continuous(sec.axis = sec_axis(~.*5, name = "Annual productivity")) +
  scale_colour_manual(values = c("blue", "red")) +
  labs(y = "Adjusted Value [$]",x = "Year")
```



Next, we chose to focus on two questions through our investigation of Honey data. Where is honey produced in the U.S., and how has production changed over time. We graphed the average annual production for each

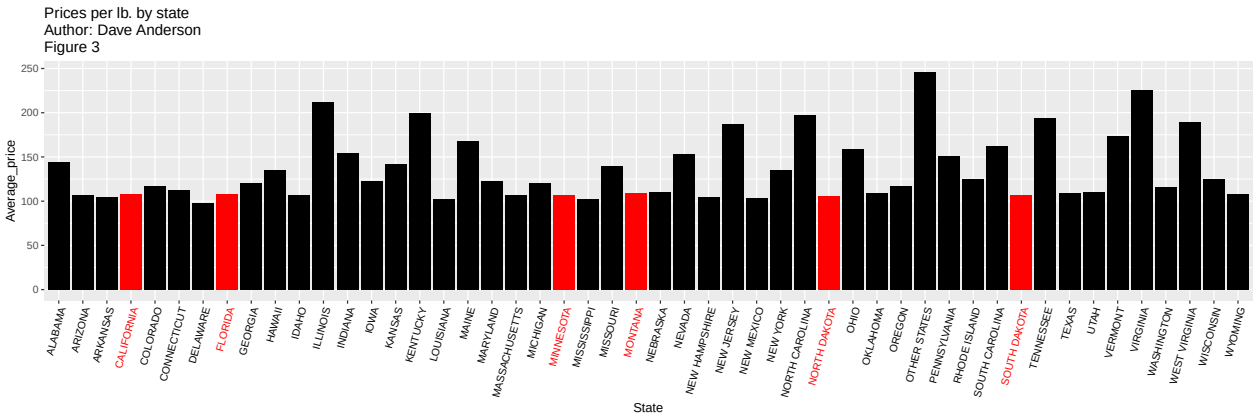
state and noticed that very few states drive the honey market including, in order, North Dakota, California, South Dakota, Florida, Minnesota, and Montana. The bar graph above (see figure 2) was the first simple step to answer the first question, showing average yearly production in honey in tons per state. Clearly, a few states drive the total U.S. market. We chose to analyze the top six states, highlighted in the graph, further.

```
#Total Production by state, largest six states indicated in red.
ggplot(data = Honey_by_State, mapping = aes(State,
  Average_production, fill=ifelse(Average_production > 5000,"A", "B")))+
  geom_col() + scale_fill_manual(
  guide=FALSE, values=c("red", "black")) +
  theme(axis.text.x = element_text(
    color = ifelse(Honey_by_State$Average_production > 5000, "red", "black"),
    angle = 75, hjust = 1)) +
  ggtitle("Production by state \nAuthor: Dave Anderson \nFigure 2")
```



The second bar chart of the same theme shows the average price of honey for each state. Again, our top six producing states are highlighted. We expected these states to have lower prices, an indicator of a healthy supply vs. demand market.

```
#Display of prices per lb. by state.
#Top six producing states still in red to show their low prices.
ggplot(data = Honey_by_State, mapping = aes(
  State ,Average_price, fill = ifelse(Average_production > 5000, "A", "B"))) +
  geom_col() +
  scale_fill_manual(guide=FALSE, values=c("red", "black")) +
  theme(axis.text.x = element_text(
    color = ifelse(Honey_by_State$Average_production > 5000, "red", "black"),
    angle = 75, hjust = 1)) +
  ggtitle("Prices per lb. by state \nAuthor: Dave Anderson \nFigure 3")
```

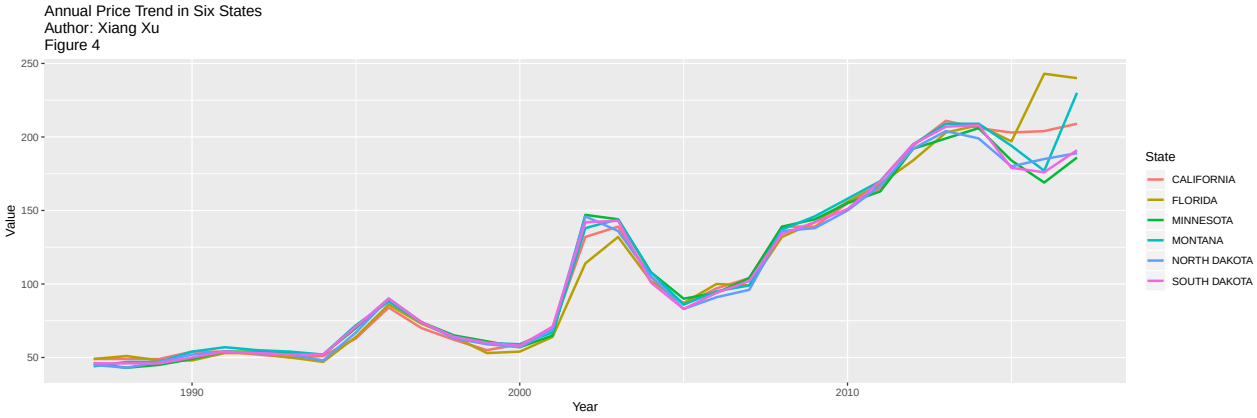


As expected, the top-producing states have some of the lowest prices, which indicates a healthy supply and demand market. (See figure 3)

Next, we wanted to take a further look into the production of the six top-producing states. With national production decreasing over time, we expected to see a similar trend with each of our six states.

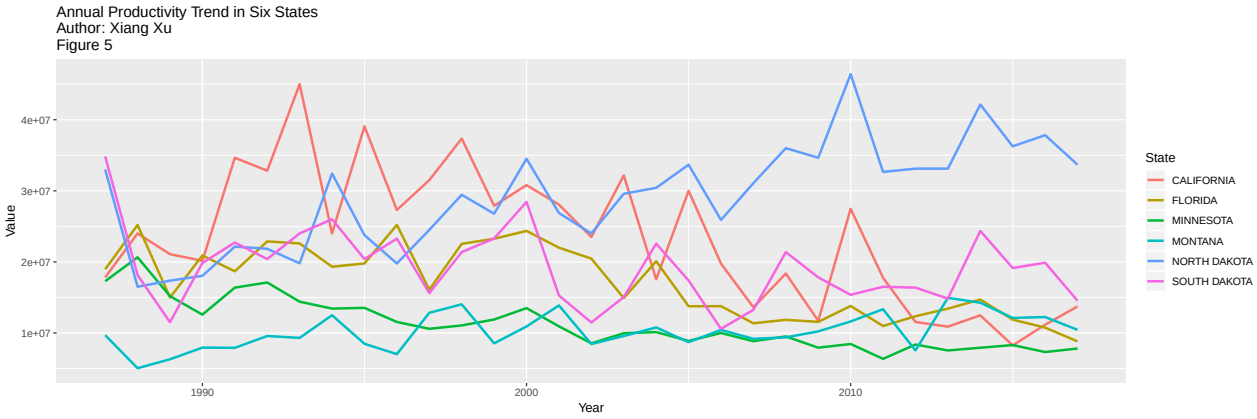
All six states have 31 years of data, which was good for our analysis. From figure 4 we could see that these six states have the same increasing trend in yearly average price, which is consistent with the price trend national wise.

```
#trend of annual price of 6 top productivity states
ggplot(Price_sixstate, aes(x=Year, y=Value ,color = State)) +
  geom_line(size = 1) +
  ggtitle("Annual Price Trend in Six States \nAuthor: Xiang Xu \nFigure 4")
```



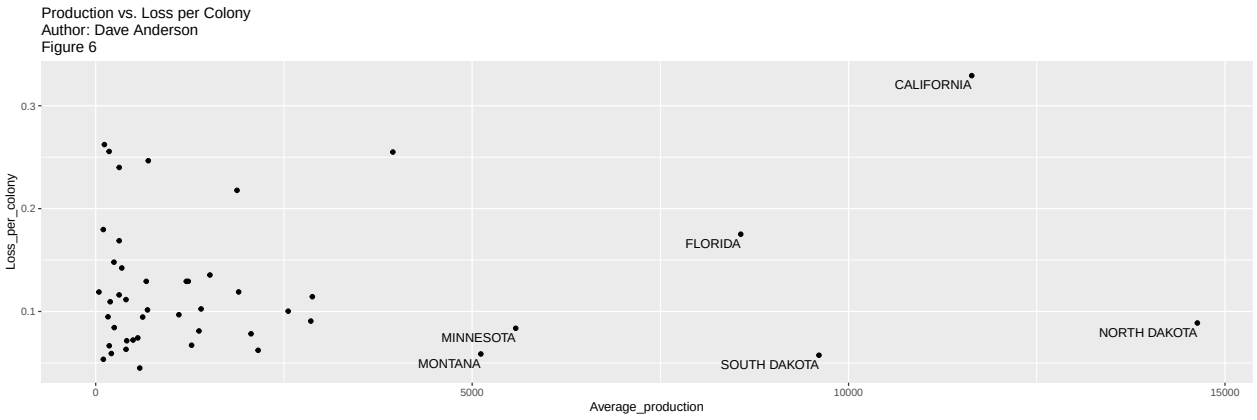
From figure 5, we can see that their yearly production fluctuated over years but in general all but North Dakotahave the downward trend in value.

```
#trend of annual productivity of 6 top productivity states
ggplot(Honey_sixstate, aes(x=Year, y=Value ,color = State)) +
  geom_line(size = 1) +
  ggtitle("Annual Productivity Trend in Six States \nAuthor: Xiang Xu \nFigure 5")
```



Why has production of honey struggled over the years? It is difficult to draw conclusions to this question, with only data on lost colonies to work with. We created a scatterplot (figure 6) showing average production vs. average loss per colony to investigate further. There was no significant trend between the two variables, but it is interesting to note the difference between our big producers.

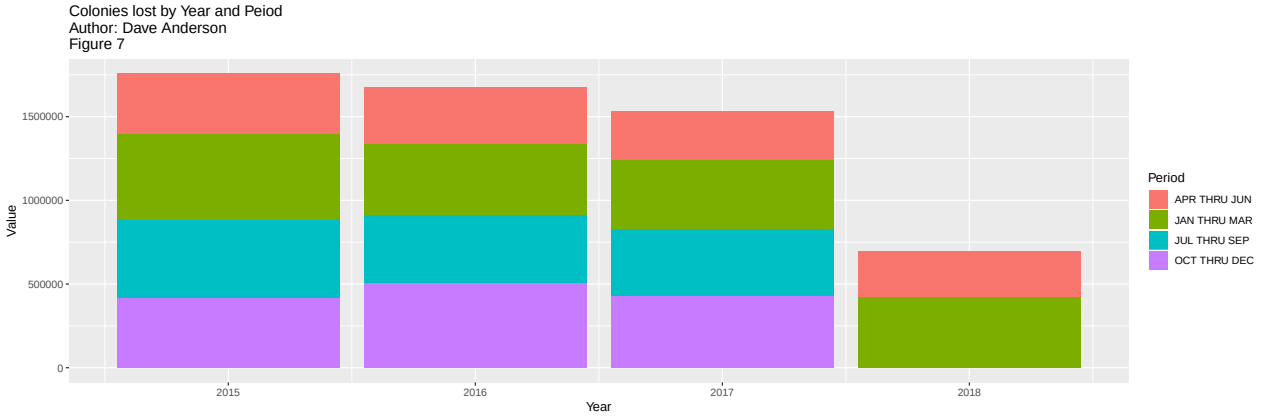
```
#Display of production vs. loss per colony with top states labeled.
ggplot(data = Honey_by_State, mapping =aes(
  Average_production,Loss_per_colony, label = State)) +
  geom_point() +
  geom_text(aes(label = ifelse(Average_production > 5000,as.character(State),'')),
  vjust = 1.5,hjust = 1) +
  ggtitle(
    "Production vs. Loss per Colony\nAuthor: Dave Anderson\nFigure 6")
```



California and Florida, our large, coastal states, have high rates of loss per colonies. California loses over 30% of their colonies. Meanwhile, the midwestern states that are large producers are all under the 10% loss benchmark. It may be reasonable to assume the climate of the midwest is actually more ideal for honeybees than the coastal states, but most states do not have enough production to help confirm this observation. Based on the data we do have, we can see that colonies lost has decreased, and that there are no periods of the year that stand out as dangerous times for our bee colonies.

We also created a chart showing the number of colonies lost over the last three years for our top six states. All loss rates are fairly steady, which does not indicate disease or catastrophic events to increase loss over this time.(See figure 7)

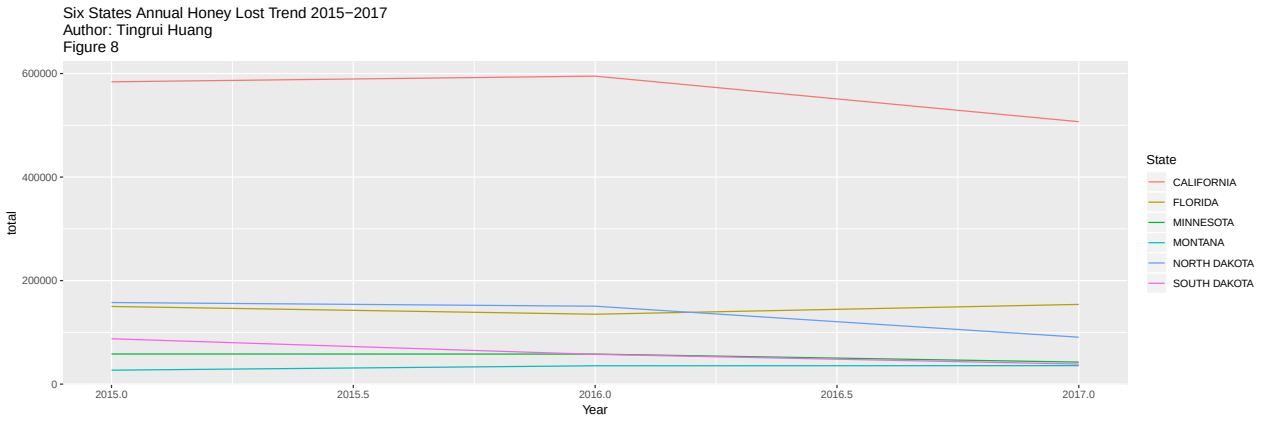
```r
#Display of colonies lost by year and peiod.
loss_by_year <- Deadout %>% filter(State == "US TOTAL") %>% group_by(Year)
ggplot(data = loss_by_year) +
  geom_col(mapping = aes(x = Year, y = Value, fill = Period)) +
  ggtitle("Colonies lost by Year and Peiod\nAuthor: Dave Anderson \nFigure 7")
```



To further understand the relations among honey production, loss and price, we take a look at the honey loss data for each year for each of the six states with highest productions. Our assumption is that for the states have higher net production (production - loss) would have a relatively lower price.

The data used for creating this graph is sourced from the 'Honey, Bee Colonies - Loss, Deadout' dataset from the USDA website. In this dataset, there are records for the honey loss in lbs from 2015 to the first half of 2018. We removed data in 2018 since it only contains the first six months data and the loss volume is obviously smaller than other years.

```r
options(scipen = 999)
# annual honey lost trend of 6 top productivity states from 2015-2017
ggplot(honey_total, aes(x=Year, y=total, color=State))+geom_line() +
  ggtitle("Six States Annual Honey Lost Trend 2015-2017\nAuthor: Tingrui Huang \nFigure 8")
```



As we can see in the graph, California has the highest loss all the time, but there was a decrease from 2016 to 2017. We assume the reason for the massive loss in California could be the result of extremely high temperature in Summer and the temperate marine climate. Florida and North Dakota have a very close loss in 2015 and the beginning of 2016, however, North Dakota has decreased its loss by 35% from 2016 to 2017 and it now has lower loss than Florida. We won't be surprised if there is a decrease in the price of honey and

an increase in the production of honey in North Dakota in the year of 2016-2017. The other three states have steady low loss compared with California, Florida and North. From the data and visualization, it???s hard to say if our assumption has met, since we only have two years of data. But we think we are on the right track, in the future, when we could get more data on loss, we could probably see the relations between net production (production - loss) and price.

Conclusion

Although we did not have enough statistical evidence to answer all of our honeybee questions, we learned a lot from our exploration. We confirmed previous knowledge that honeybee populations have struggled in recent years. We learned that six states are by far the top producers of honey in the United States. In analyzing these states, we learned that price of honey follows a typical pattern based on supply, and the number of colonies dying in each state varies quite a bit. We were not able to find much of a pattern to explain why colonies die, and further data collection would be necessary to further investigate this question.

Work Cited

[1] United States Department of Agriculture - National Agricultural Statistics Service https://www.nass.usda.gov