Luke Weber
MAT 421- 17427
Professor H. Wang
Apr 16th 2023

Bees to Honey

Introduction

I found a data set on Kaggle[1] covering several metrics related to honey bee production and its value in the United States from the year 1995 up until 2021. Across several US states for each year the data set contains relevant information on the amount of bee colonies, the honey yield per colony, and the total amount of honey produced. As well as the average price per pound of honey and the total value of the honey production. I am interested in seeing the relationship between colony size and the amount of honey produced as well as colony size and the total value of the honey. I am also interested in the trend per year and seeing if honey production, colony size, or value is rising each year.

I will be using RStudio to analyze the data set and create the linear regression model for each relationship with the entire US and individual states. The GitHub and R file are shared down in the references[2]. The R file details each step of the process and we will go through the file in the linear regression model section. RStuido is a great program that will be able to handle the regression model and provide a better visual representation of the data and the regression line than jupyter. It will also be able to do the exact same things as python or jupyter but I'll be able to keep the flow and code more organized in its presentation.

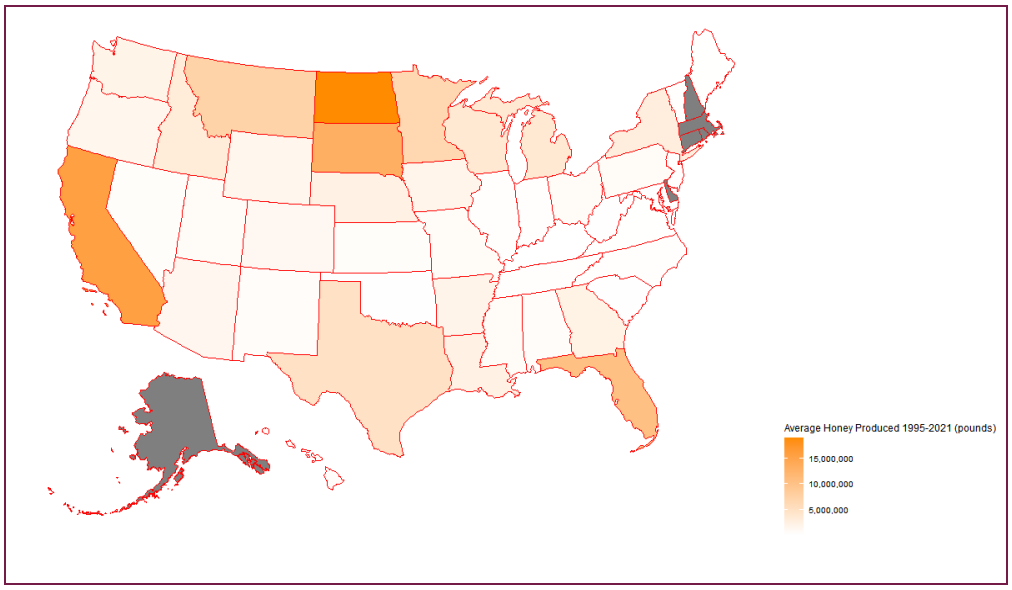
The Data

The first thing after opening the data set in R that is catching is that the data set only includes forty four of the fifty united states. Alaska, Delaware, Connecticut, Massachusetts, New Hamshire, and Rhode Island are absent from the data set. It is not completely damaging we can continue on with the model, just note that those states are not represented in any of the models. The second and most glaring problem is that the data set is not necessarily independent. Take for example the two data points Alabama 1995 and Arizona 1995 these may be independent but we will have Alabama again for the year 1996. Alabama 1996 is not independat from Alabama 1995. The same state is not independent in sequential years. This violates one of the assumptions of a the linear regression model. We can create independence by taking the averages of each state for every year. So instead of having Alabama 1995, Alabama 1996, … Alabama 2021 we just have Alabama. The code below shows how we will create the new data set with forty four entries containing independent means for each state. We will also run another block of code following the same logic but concerning each year to create another independent data set that contains the means for each year. We will start with the state mean models, because we can create a geographic heat map of the means for the fourty four represented states.

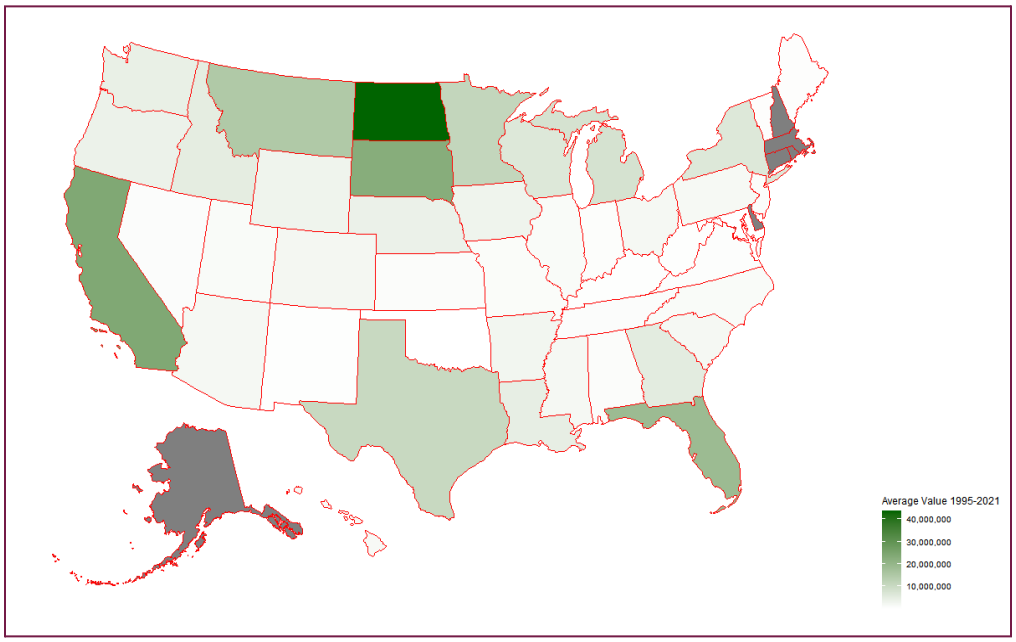
```
StateProd = HoneyProd %>%
  group_by(state) %>%
    summarise(
      mean_prod=mean(production),
      mean_value=mean(value_of_production),
      mean_size=mean(colonies_number)
    )
View(StateProd)
```

```
YearProd = HoneyProd %>%
  group_by(year) %>%
    summarise(
      mean_prod=mean(production),
      mean_value=mean(value_of_production),
      mean_size=mean(colonies_number)
  )
View(YearProd)
```

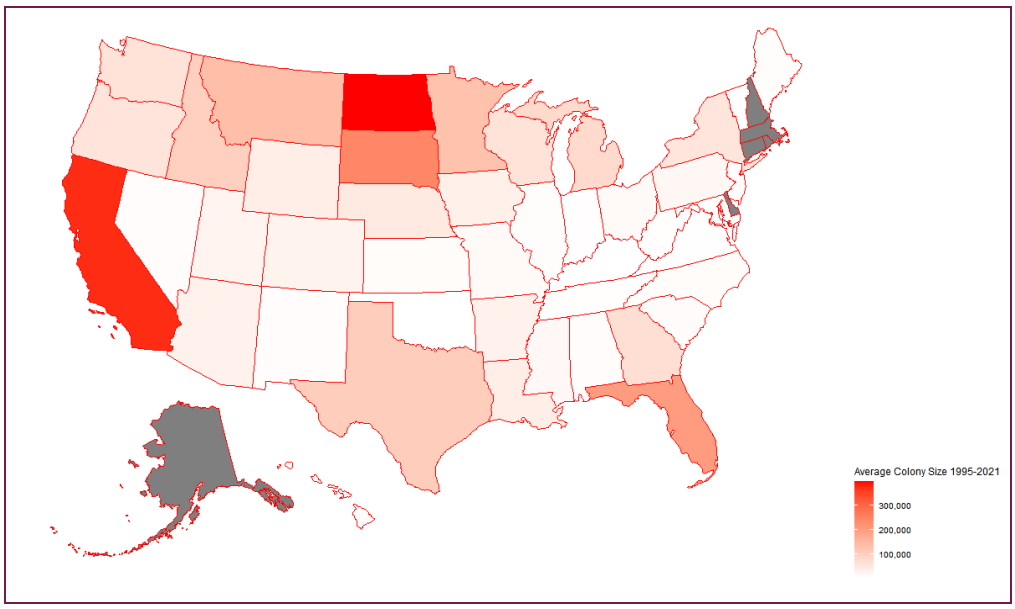
There are three heat maps, one for the average amount of honey produced (pounds), another for the average value of the honey (dollars), and lastly for the average number of honey producing bee colonies. The heat maps provide a way to visualize all of the data without an axis or plot points. It is also really simple to make connections to the data set.



Above is the heat map for the average amount of honey produced (pounds) during 1995-2021 for each state. The grey states are the six states not represented in the data.



Above is the heat map for the average value of the production (dollars) during 1995-2021 for each state. It is very similar to the first heat map, about the honey produced.



Above is the heat map for the average number of producing bee colonies during 1995-2021 for each state. It is very similar to the previous two heat maps suggesting that there is some relationship between the number of bee colony, the amount of honey they produce, and their value. Which makes sense.

Linear Regression Per State

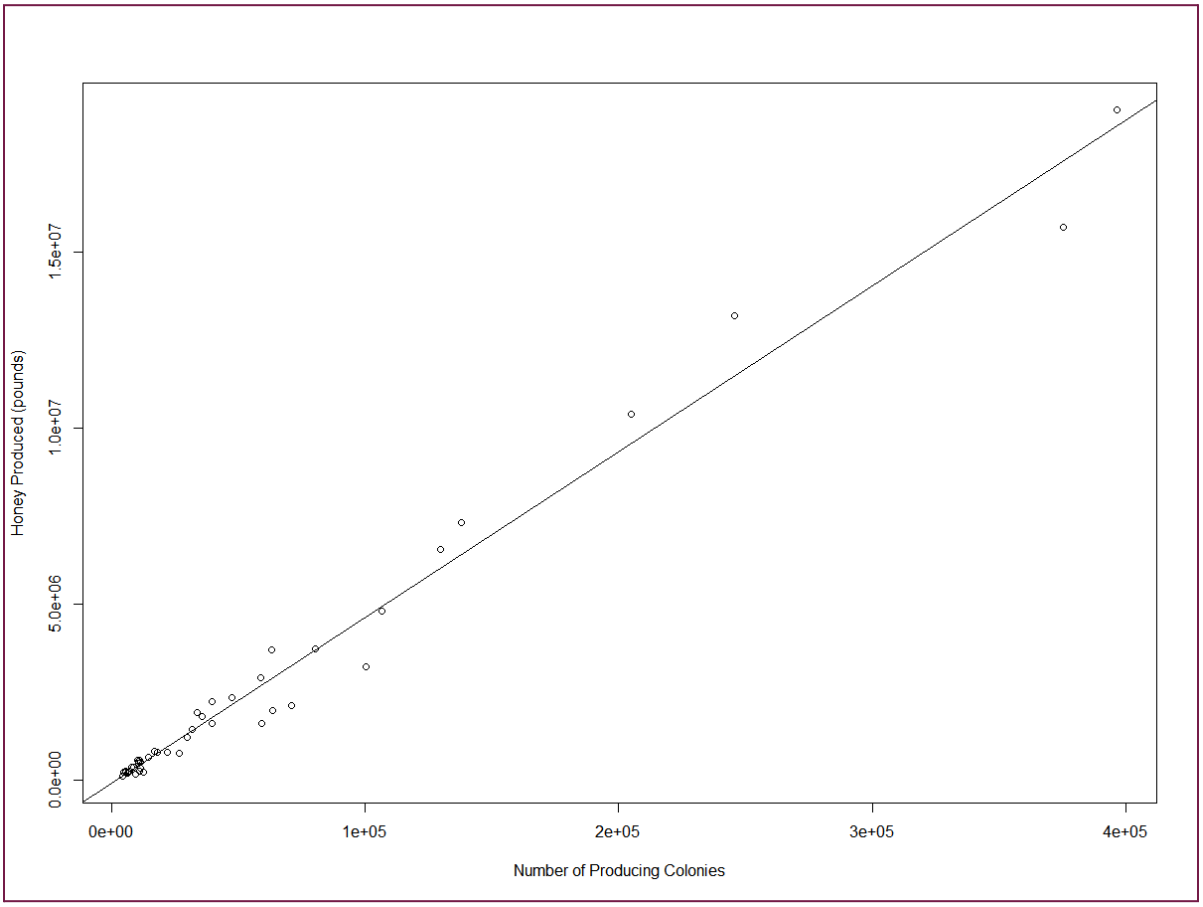
With a pretty good idea that there is some relationship between the three variables we want to confirm that their relationship is linear in order to do a linear regression. If not this paper is going to end abruptly. We want the absolute value of our correlation $|r|$ to be close to 1, the closer the better. If r is positive then there is a positive relationship, our line will go up; if r is negative then there is a negative relationship, our line will go down. Running the correlation between the average number of colonies and average honey produced gives, $r = 0.99$, which is incredible. This indicates that there is a very strong

positive linear relationship between those two variables. We will run the rest of the code below to create the linear regression model, plot the data, and plot the regression line.

```
cor(StateProd$mean_size, StateProd$mean_prod)
USreg1=lm(StateProd$mean_prod~StateProd$mean_size)
plot(StateProd$mean_size, StateProd$mean_prod, xlab="Number of Producing Colonies", ylab="Honey Produced (pounds)")
abline(USreg1)
summary(USreg1)
```

This gives the graph below with the regression line. It has an adjusted $r^2 = 0.97$ and a p-value (2.2×10^{-16}). The regression line follows the equation,

$$\widehat{(Average\ Honey\ Produced)} = 47(Average\ Number\ of\ Colonies) - 85,780.$$



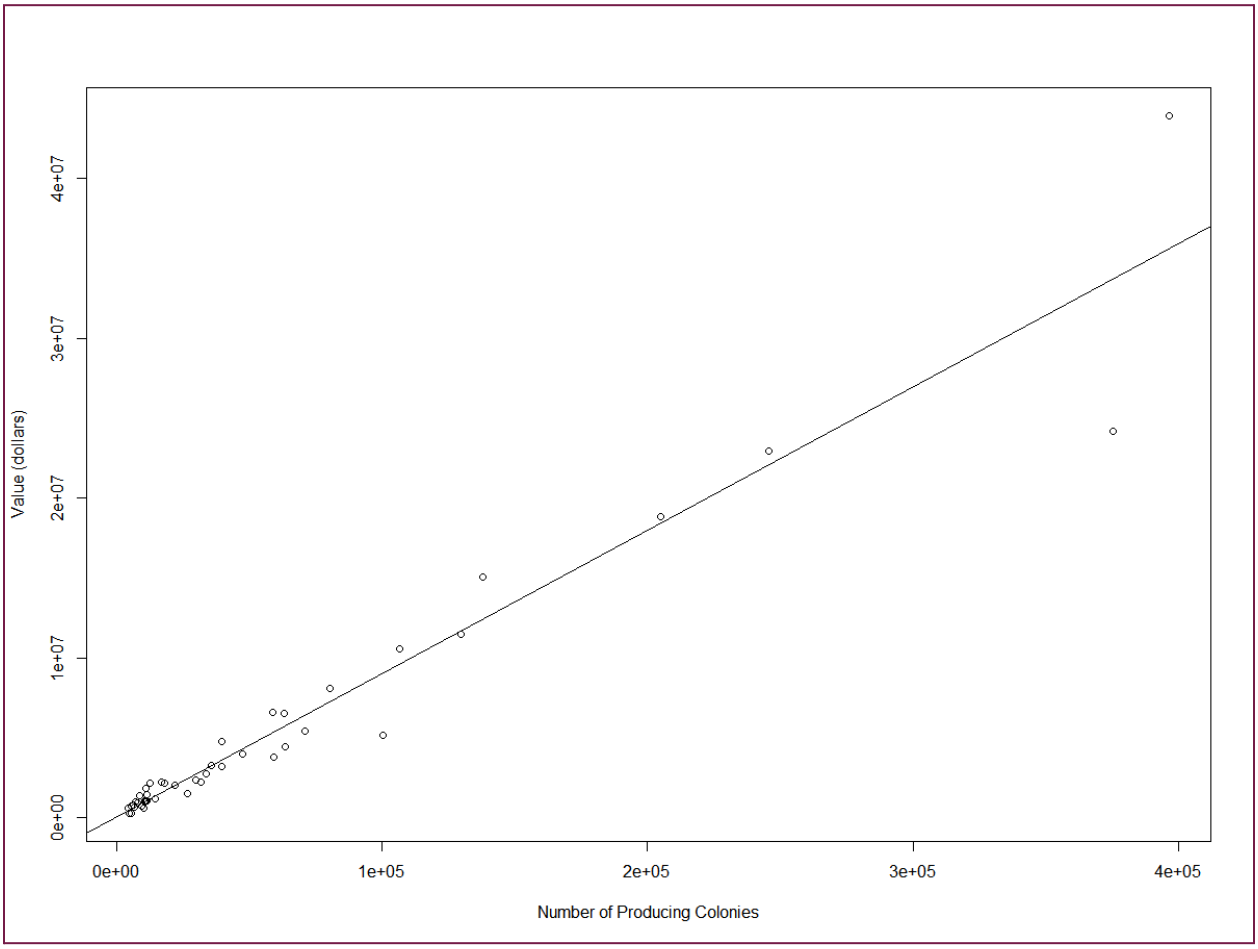
Running the correlation between the average number of colonies and average value of production gives, $r = 0.96$, which is also incredible. This indicates that there is a very strong positive linear relationship between those two variables as well. We will run the rest

of the code below to create the linear regression model, plot the data, and plot the regression line.

```
cor(StateProd$mean_size, StateProd$mean_value)
USreg2=lm(StateProd$mean_value~StateProd$mean_size)
plot(StateProd$mean_size, StateProd$mean_value, xlab="Number of Producing Colonies", ylab="Value (dollars)")
abline(USreg2)
summary(USreg2)
```

This gives the graph below with the regression line. It has an adjusted $r^2 = 0.93$ and a p-value (2.2×10^{-16}). The regression line follows the equation,

$$\widehat{(Average\ Value\ of\ Production)} = 89.6(Average\ Number\ of\ Colonies) + 74,738.$$



We are going to skip doing the correlation regression model between average honey produced and average value of production because both of those variables have a linear relationship with the average number of colonies. So they must have a linear relationship

with each other. The average amount of honey you have multiplied by the average price would be the average value. Trivially.

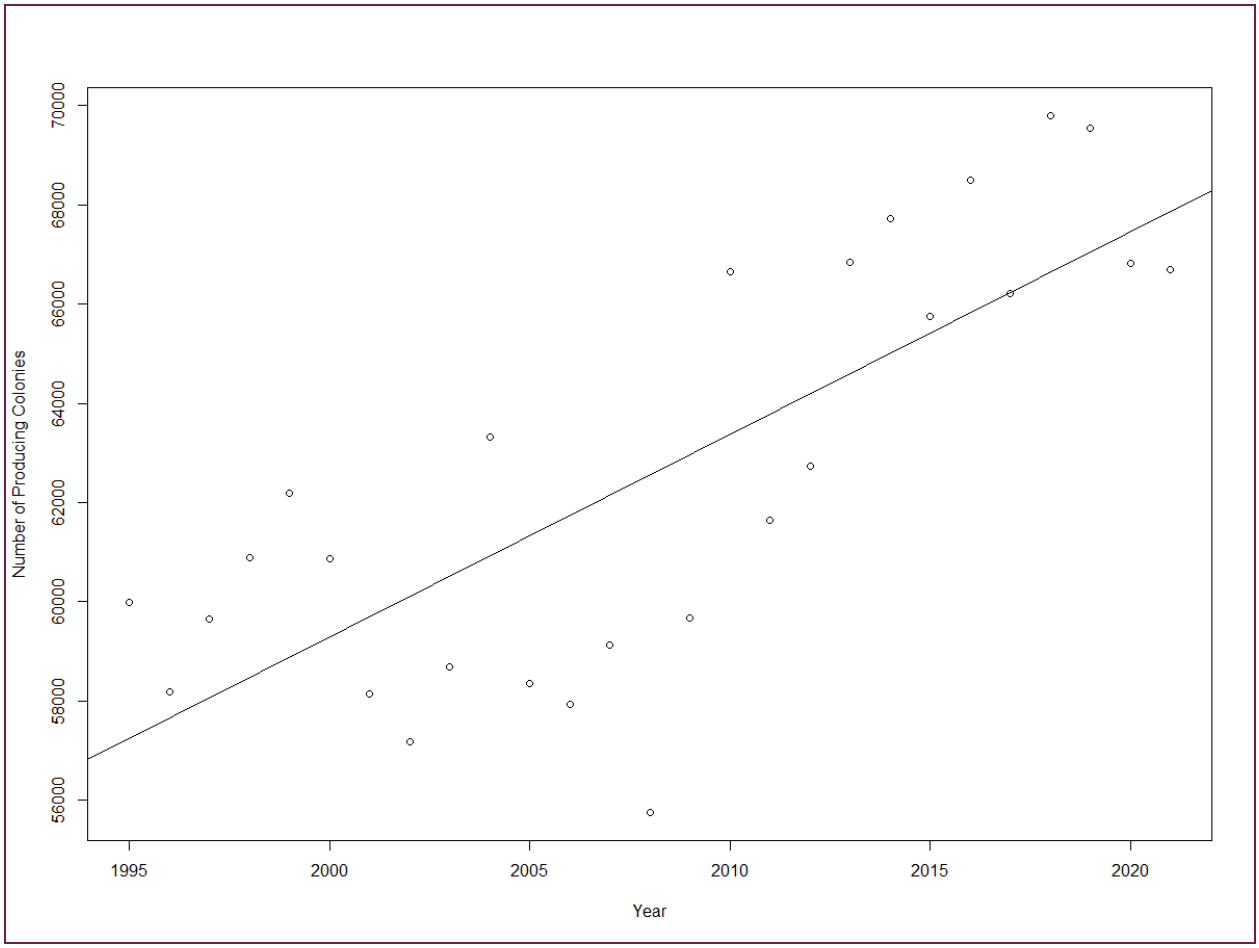
Linear Regression Per Year

There is no need to create a visual of average per year data set since it is one point per year. We could only make a bar graph or plot, and when we look at the linear regression we will make a plot for it. So visualizing the data before doing that is redundant. Additionally the data is already in order by the year. First we will look at the correlation between the year and the average number of colonies. It gives $r = 0.76$, which is good. This indicates a strong positive relationship between the two variables. Running the rest of the code below, gives us the regression model, plots the data, and plots the regression line.

```
cor(YearProd$year, YearProd$mean_size)
Yreg1=lm(YearProd$mean_size~YearProd$year)
plot(YearProd$year, YearProd$mean_size, xlab="Year", ylab="Number of Producing Colonies")
abline(Yreg1)
summary(Yreg1)
```

This gives the graph below with the regression line. It has an adjusted $r^2 = 0.56$, and a p-value (3.4×10^{-6}). The regression line follows the equation,

$(\widehat{Average\ Number\ of\ Colonies}) = 408(Year) - 758,201.$

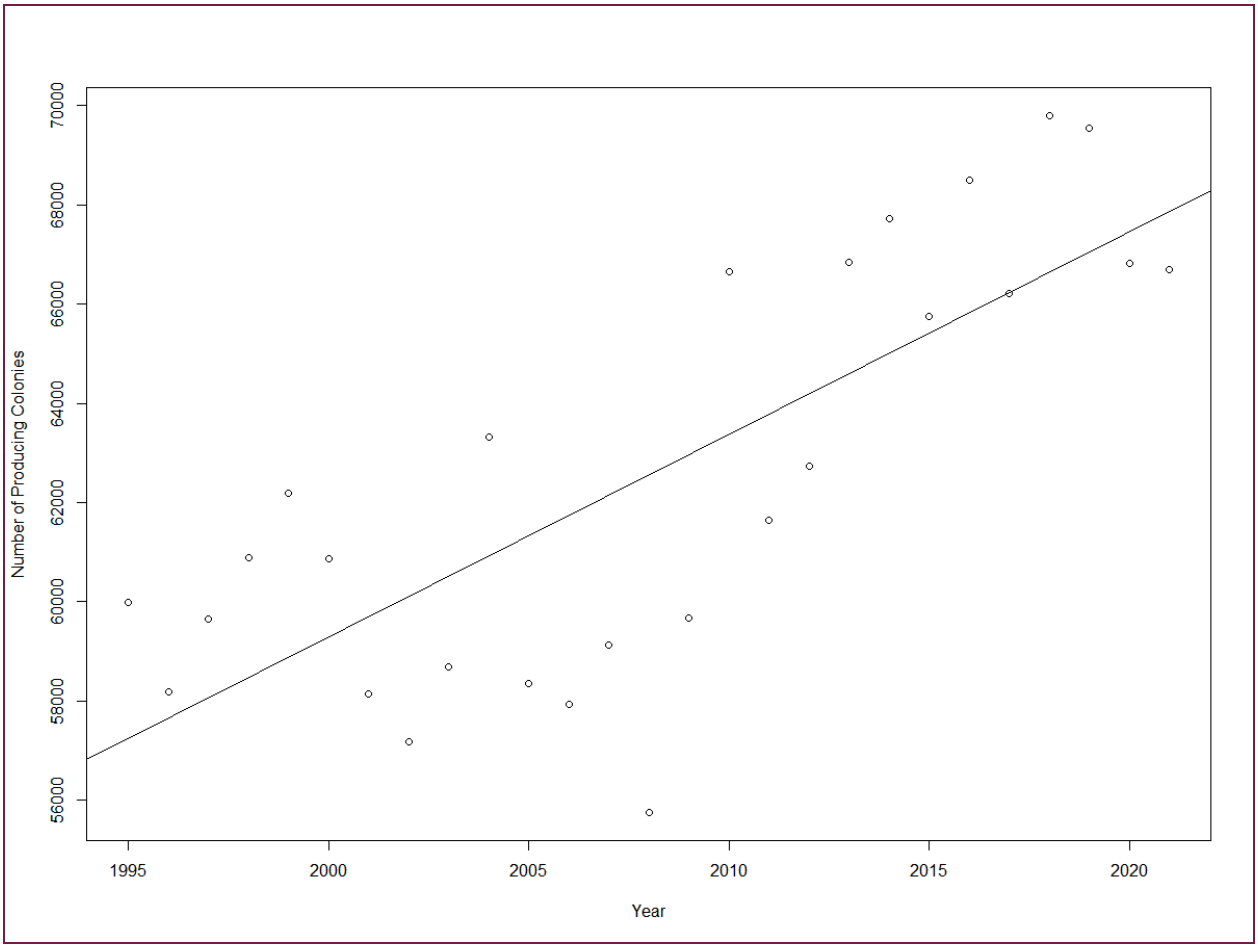


Next looking at the correlation between the year and the average value of production. It gives $r = 0.88$, which is good. This indicates a strong positive relationship between the two variables. Running the rest of the code below, gives us the regression model, plots the data, and plots the regression line.

```
cor(YearProd$year, YearProd$mean_value)
Yreg2=lm(YearProd$mean_value~YearProd$year)
plot(YearProd$year, YearProd$mean_value, xlab="Year", ylab="Value (dollars)")
abline(Yreg2)
summary(Yreg2)
```

This gives the graph below with the regression line. It has an adjusted $r^2 = 0.77$, and a p-value (8.3×10^{-10}). The regression line follows the equation,

$(\widehat{Average\ Value\ of\ Production}) = 236,043(Year) - 468,246,160.$

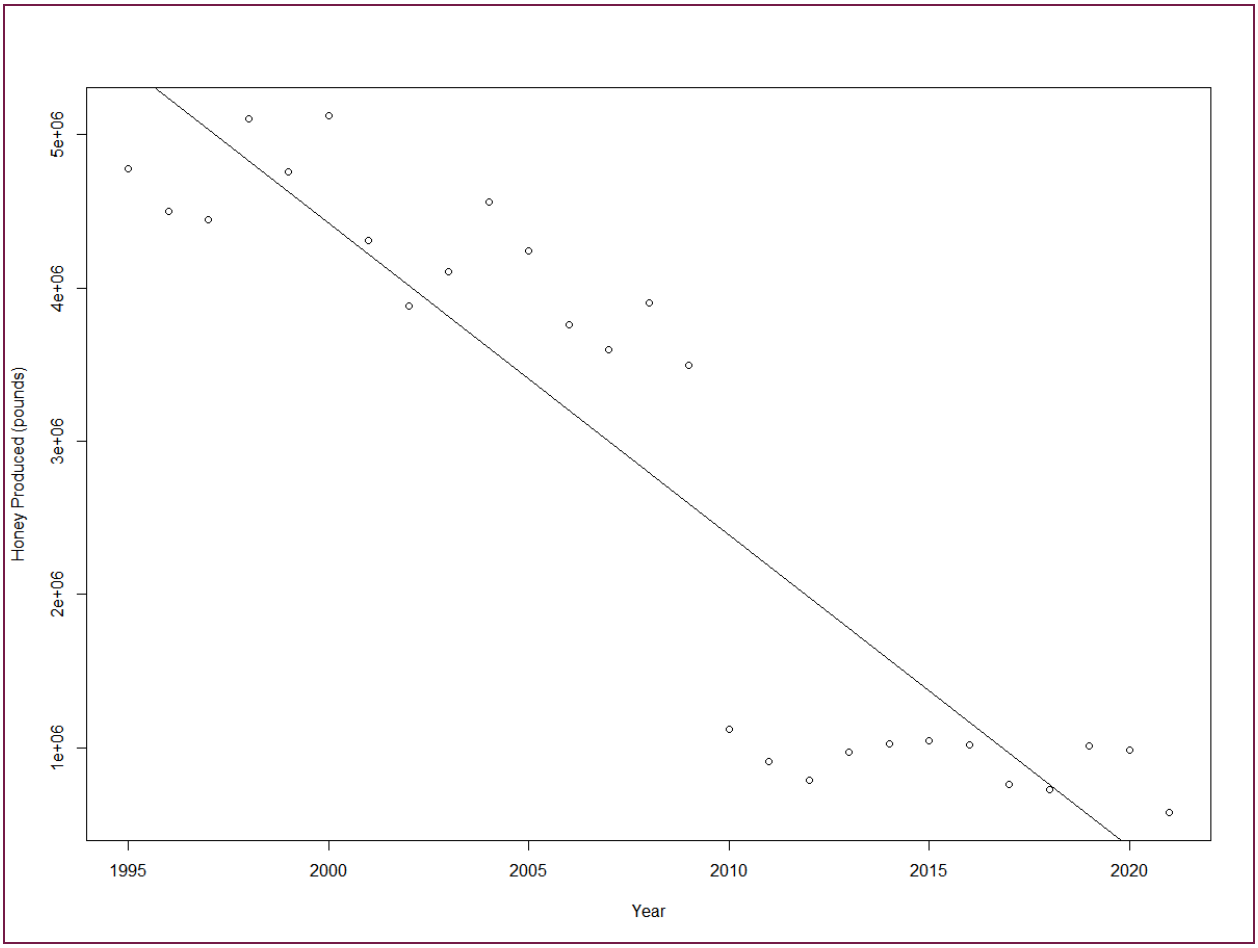


Finally taking a look at the correlation between the year and the average amount of honey produced. It gives $r = -0.91$, which is very good. This indicates a very strong negative relationship between the two variables. Running the rest of the code below, gives us the regression model, plots the data, and plots the regression line.

```
cor(YearProd$year, YearProd$mean_value)
Yreg2=lm(YearProd$mean_value~YearProd$year)
plot(YearProd$year, YearProd$mean_value, xlab="Year", ylab="Value (dollars)")
abline(Yreg2)
summary(Yreg2)
```

This gives the graph below with the regression line. It has an adjusted $r^2 = 0.83$, and a p-value (1.9×10^{-11}). The regression line follows the equation,

$(Average\ \widehat{Honey}\ Produced) = -203,402(Year) + 411,227,237.$



Conclusion

It is interesting to note that the average amount of colonies and the average value of production trend upwards each year, but the average amount of honey trends downward each year. This seems to contradict the state grouped linear models, but that is not necessarily the case. When averaging the data for each state and not accounting for the year, the data loses its year to year story. For example looking at North Dakota and

averaging the variables for each year, you end up with North Dakota having the most producing colonies and the most honey produced. It does not matter if one of those variables decreased year to year , on average the states with more colonies produce more honey. When looking at the the average data for every year you lose that state level focus in trade for a broad nation wide story. I feel like the year averaged models are the most interesting to look at and they tell the best story, but the state averaged models are the most useful. I think that the data set itself is still very interesting and has potential for more testing and modeling. The heat map is an extremely helpful visualization tool for this data as it allows a real life geographical representation. It is very easy to read and understand the story of the data. Overall the linear regression models for the averages are strong and allow good insight into the relationships between the number of colonies, honey produced, and production value. The linear regression models also provide a good look at how these variables have changed over time, and where they will be in the future.

References

[1]https://www.kaggle.com/datasets/mohitpoudel/us-honey-production-19952021

[2]https://github.com/Lukew69/Honey-Production