Decline of the Honey Bee Population

Luke Andrade
Submitted to: Amelie Marian, PhD
Department of Computer Science, Rutgers University
December 12, 2022

Introduction

A large proportion of the world is aware that the honey bee population is in sweep decline. However, not everyone understands the impact that honey bees have on our daily lives or how serious the bee decline is. According to the USDA, a healthy pollinator population is vital to producing marketable commodities, they support healthy ecosystems needed for clean air, stable soils, and a diverse wildlife, and their pollination efforts increases crop production and quality for a wide variety of foods, including fruits, nuts, vegetables, legumes, oilseeds, and forage crops. The decline in honey bee population, honey production, and crop production will be investigated using data from the USDA.

USDA Information Source: https://www.usda.gov/media/blog/2020/06/24/pollinators-crossroads

Data: https://quickstats.nass.usda.gov/

Tools Used

All work will be done in R with the family of packages Tidyverse for data manipulation and visualization

```
library(tidyverse)
```

The Data

This first data set was downloaded from USDA quick stats and will be mutated to contain the variables year, US state, honey production in lbs, honey production in lbs per colony, and the amount of colonies.

```
df1 <- read_csv('USDAdata.csv') %>%
  filter(Domain == 'TOTAL',
         Period == 'MARKETING YEAR',
         `Data Item` %in% c('HONEY - PRODUCTION, MEASURED IN LB', 'HONEY - PRODUCTION, MEASURED IN LB /
  mutate(Value = str_replace_all(Value, ',', '') %>%
           as.numeric(),
         State = str_to_title(State)) %>%
  drop_na(Value) %>%
  select(c(2,6,17,20)) %>%
  arrange(Year, State) %>%
  pivot_wider(names_from = `Data Item`, values_from = Value)
names(df1) <- c('Year', 'State', 'Honey(lbs)', 'Honey(lbs/colony)', 'Colonies')
head(df1)
```

```
## # A tibble: 6 x 5
##    Year State       `Honey(lbs)` `Honey(lbs/colony)` Colonies
##   <dbl> <chr>              <dbl>               <dbl>    <dbl>
## 1  1987 Alabama          1610000                  35    46000
## 2  1987 Arizona          3760000                  47    80000
## 3  1987 Arkansas         2001000                  69    29000
## 4  1987 California      17820000                  33   540000
## 5  1987 Colorado         3212000                  73    44000
## 6  1987 Connecticut        68000                  34     2000
```

The second data set was downloaded from USDA and will be mutated to contain the variables year, US state, and pounds of apples produced.

```
df2 <- read_csv('USDAdata2.csv') %>%
  filter(`Data Item` == 'APPLES - PRODUCTION, MEASURED IN LB') %>%
  select(c(2,6,20)) %>%
  mutate(State = str_to_title(State),
         Value = str_replace_all(Value, ',', '') %>%
           as.numeric()) %>%
  filter(Value > 0) %>%
  drop_na(Value)
names(df2) <- c('Year', 'State', 'Apple(lbs)')
head(df2)
```

```
## # A tibble: 6 x 3
##    Year State        `Apple(lbs)`
##   <dbl> <chr>               <dbl>
## 1  2022 California      240000000
## 2  2022 Michigan       1100000000
## 3  2022 New York       1450000000
## 4  2022 Oregon          175000000
## 5  2022 Pennsylvania    460000000
## 6  2022 Virginia        185000000
```

The next step is to join the first and third data sets into one using an inner join.

```
df3 <- inner_join(df1, df2, by = c('Year', 'State'))
head(df3)
```

```
## # A tibble: 6 x 6
##    Year State          `Honey(lbs)` `Honey(lbs/colony)` Colonies `Apple(lbs)`
##   <dbl> <chr>                 <dbl>                <dbl>    <dbl>        <dbl>
## 1  1987 Kansas              2346000                   51    46000     12000000
## 2  1987 New Jersey           850000                   34    25000     80000000
## 3  1987 New Mexico           950000                   50    19000     12600000
## 4  1987 Pennsylvania        1872000                   39    48000    500000000
## 5  1987 South Carolina       510000                   34    15000     45000000
## 6  1987 Virginia            1200000                   48    25000    455000000
```

Finally, we will merge the original 2 data sets but using a left join.

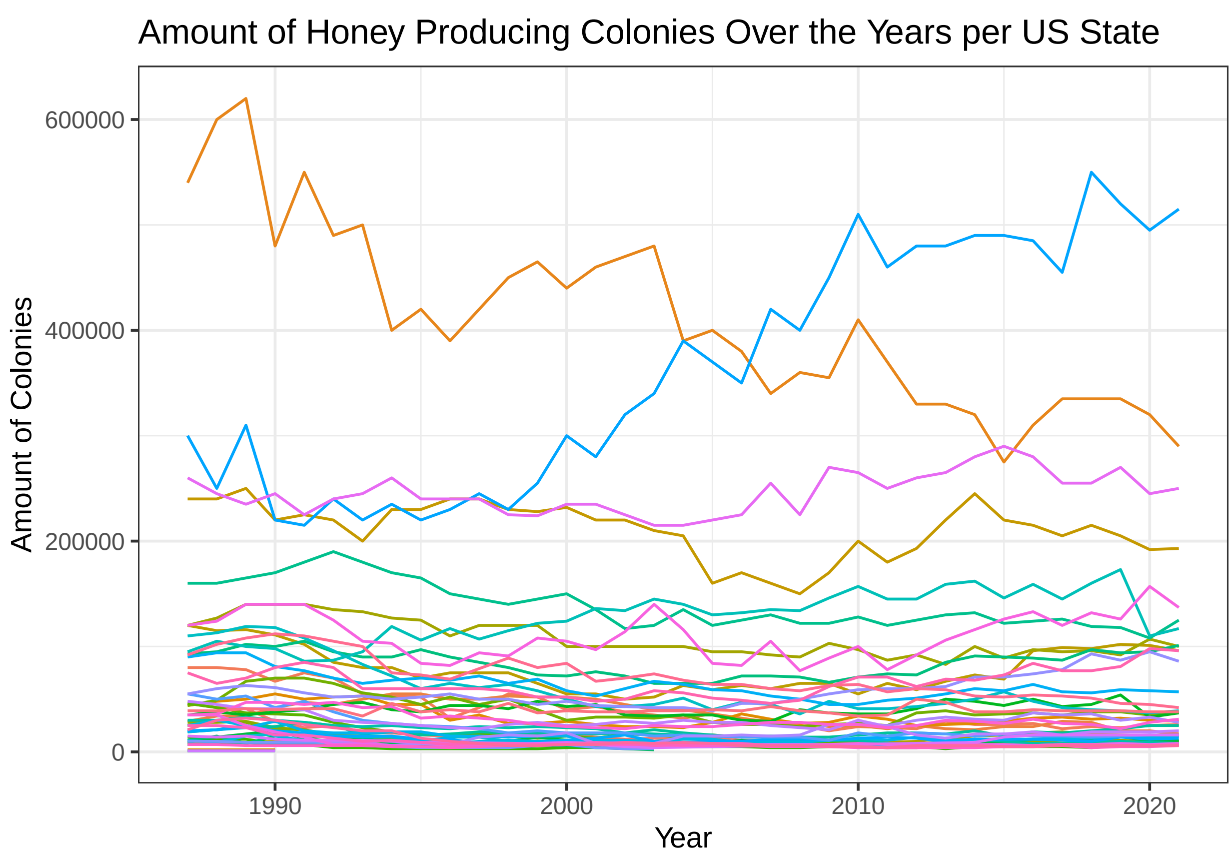
```
df4 <- df1 %>%
  left_join(df2, by = c('Year','State'))
head(df4)
```

```
## # A tibble: 6 x 6
##    Year State       `Honey(lbs)` `Honey(lbs/colony)` Colonies `Apple(lbs)`
##   <dbl> <chr>              <dbl>                <dbl>    <dbl>        <dbl>
## 1  1987 Alabama          1610000                   35    46000           NA
## 2  1987 Arizona          3760000                   47    80000           NA
## 3  1987 Arkansas         2001000                   69    29000           NA
## 4  1987 California      17820000                   33   540000           NA
## 5  1987 Colorado         3212000                   73    44000           NA
## 6  1987 Connecticut        68000                   34     2000           NA
```

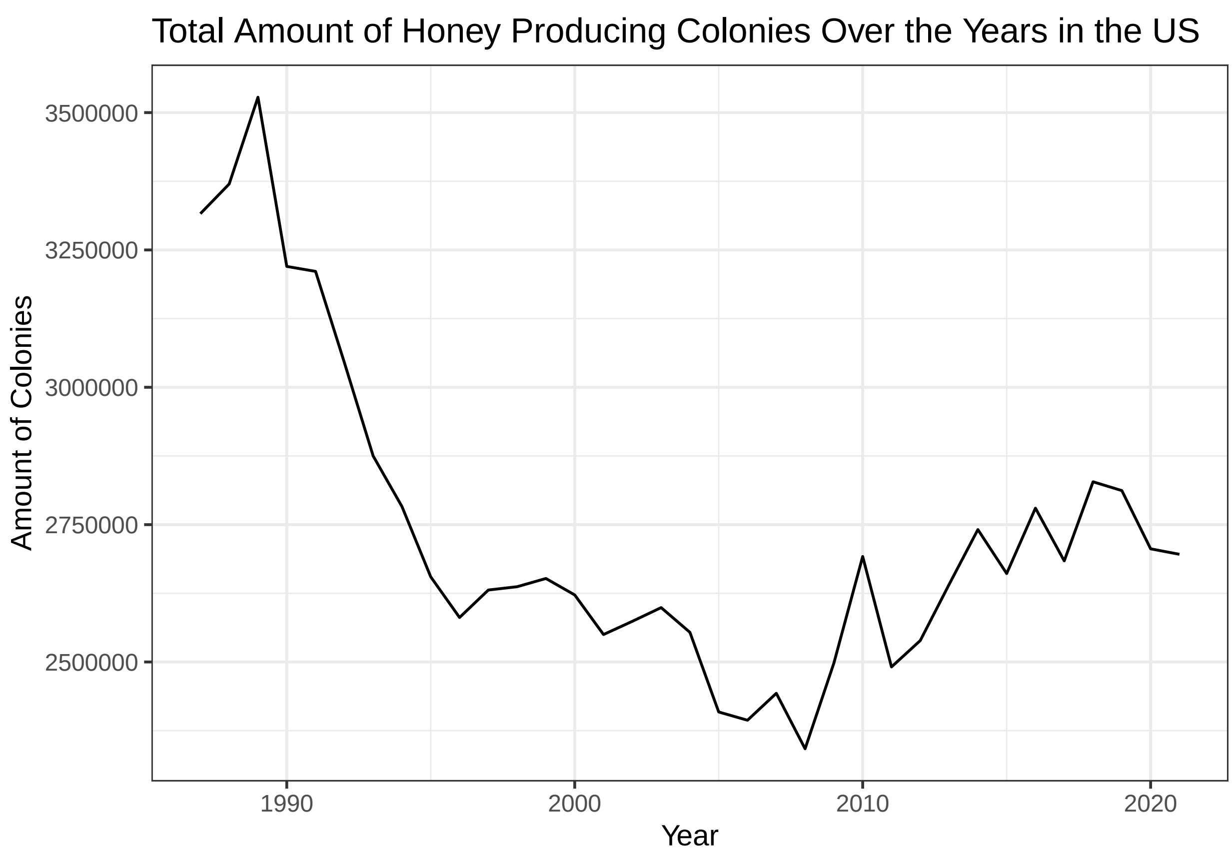
Visualizing the Data

First, lets see how the amount of colonies has changed over the years.

```
df4 %>%
  filter(State != 'Us Total') %>%
  ggplot(aes(x = Year, y = Colonies, color = State)) +
  geom_line() +
  theme_bw() +
  theme(legend.position = 'none') +
  ggtitle('Amount of Honey Producing Colonies Over the Years per US State') +
  xlab('Year') +
  ylab('Amount of Colonies')
```

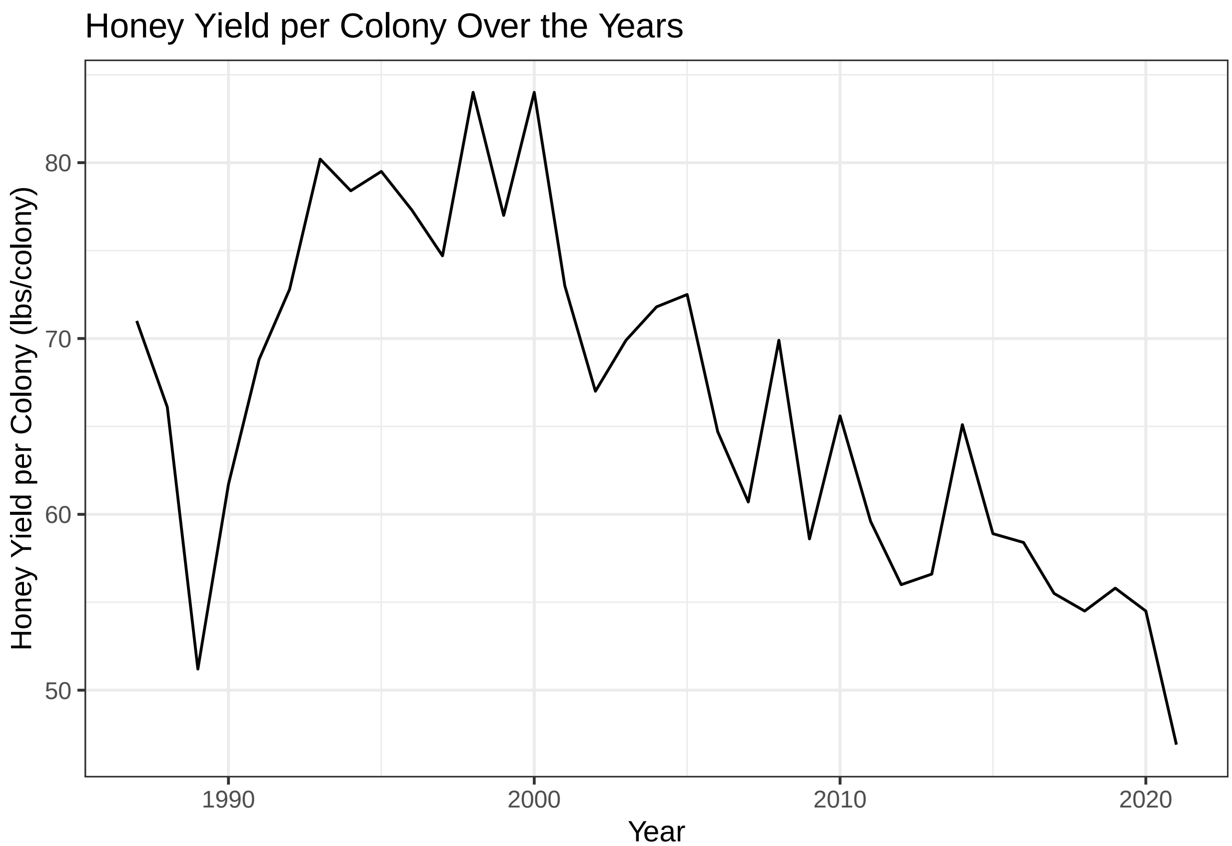


```r
df4 %>%
  filter(State == 'Us Total') %>%
  ggplot(aes(x = Year, y = Colonies)) +
  geom_line() +
  theme_bw() +
  ggtitle('Total Amount of Honey Producing Colonies Over the Years in the US') +
  xlab('Year') +
  ylab('Amount of Colonies')
```



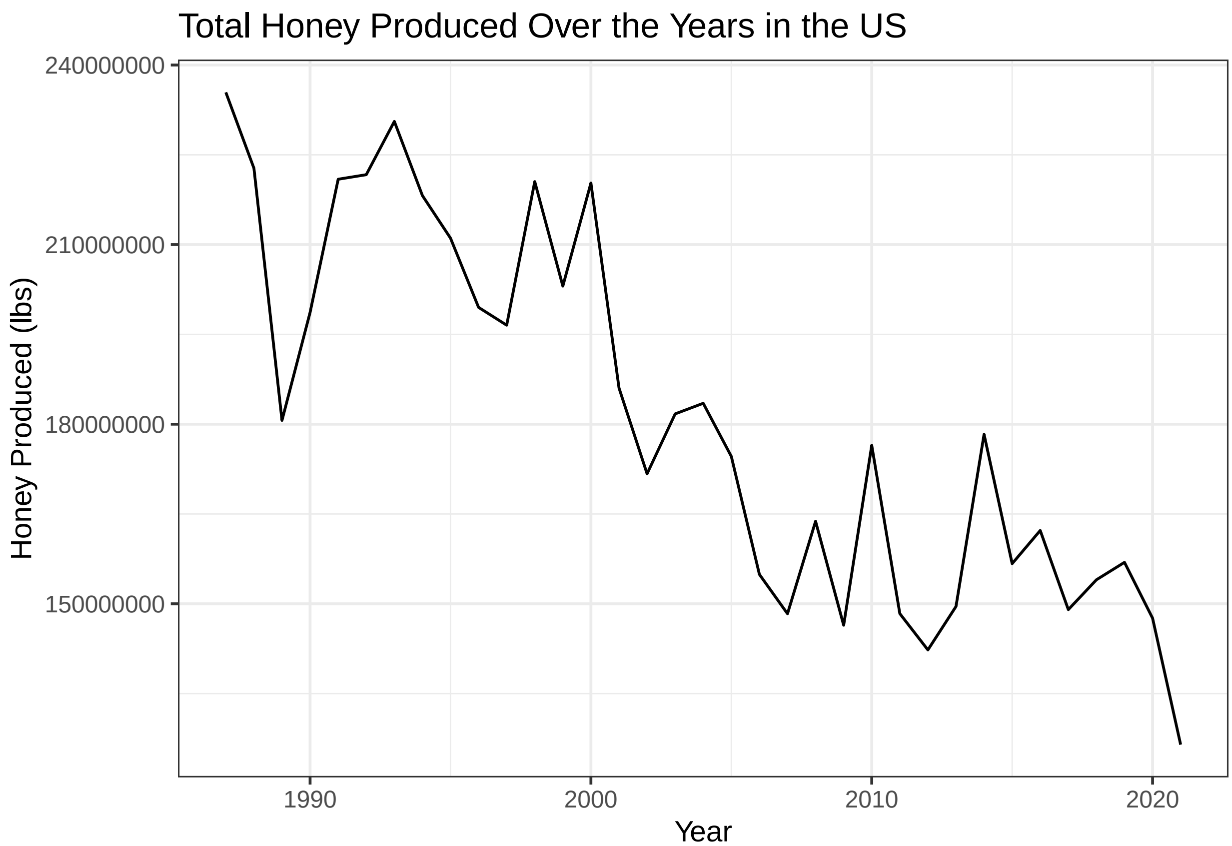
When looking at the data per state, it is difficult to come to one conclusion as each state has a different trend. However, the data for the US total shows that there is a clear decrease in total amount of colonies. However, this does not necessarily imply that there are less amount of bees but rather less amount of colonies. It could be possible that there are fewer colonies with more bees. Let's investigate the honey yield per colony.

```r
df4 %>%
  filter(State == 'Us Total') %>%
  ggplot(aes(x = Year, y = `Honey(lbs/colony)`)) +
  geom_line() +
  theme_bw() +
  ggtitle('Honey Yield per Colony Over the Years') +
  xlab('Year') +
  ylab('Honey Yield per Colony (lbs/colony)')
```



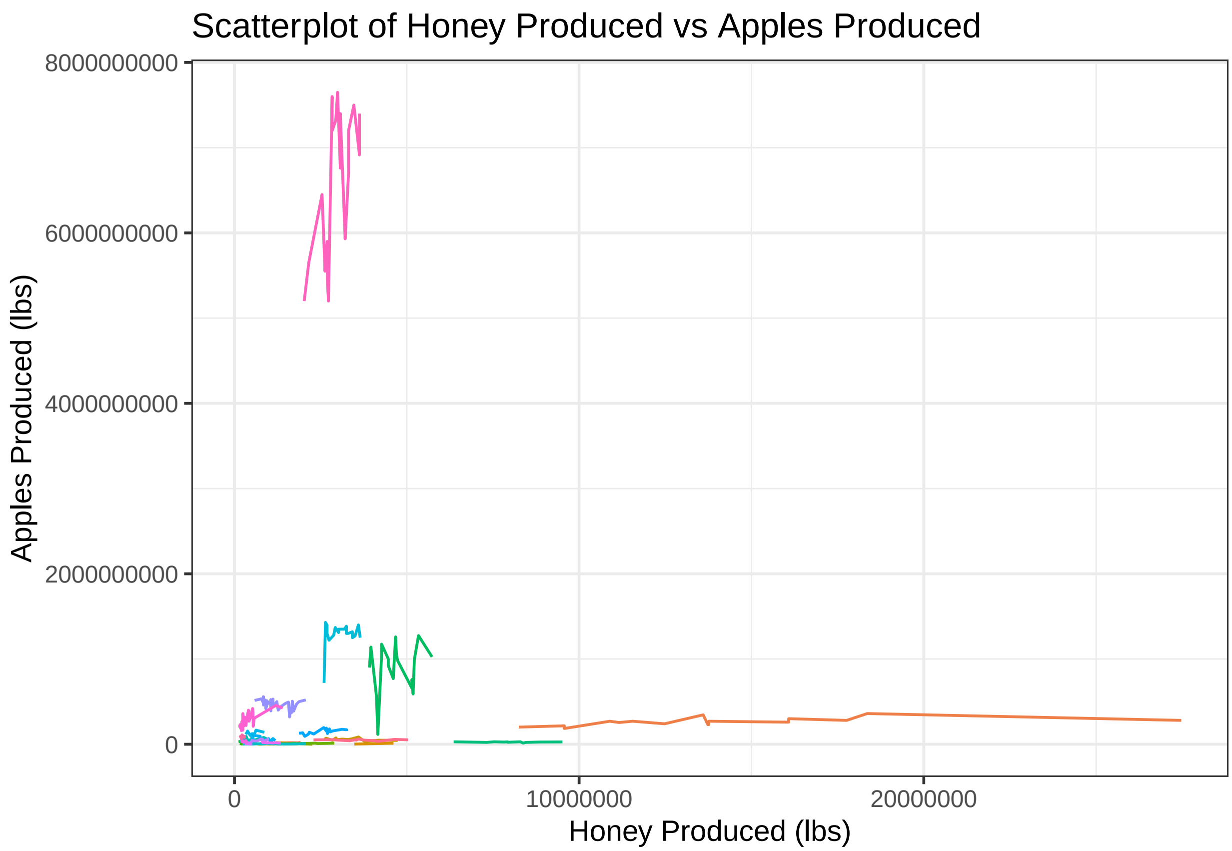
This plot is interesting because we see that the yield per colony is at its highest around the year 2000 but then falls off. This contradicts what we would expect since the amount of honey producing colonies had a major drop right before the year 2000. Could it be possible that when the amount of bees decreases they end up having to pollinate more to produce more honey to compensate for the lack of workers? I am curious as the what the total honey production per year is.

```
df4 %>%
  filter(State == 'Us Total') %>%
  ggplot(aes(x = Year, y = `Honey(lbs)`)) +
  geom_line() +
  theme_bw() +
  ggtitle('Total Honey Produced Over the Years in the US') +
  xlab('Year') +
  ylab('Honey Produced (lbs)')
```



As expected, there is a gradual but fluctuating decrease in total honey production in the US. Let's see if there is a trend between pounds of honey produced and pounds of apples produced.

```r
df3 %>%
  ggplot(aes(x = `Honey(lbs)`, y = `Apple(lbs)`, color = State)) +
  geom_line() +
  theme_bw() +
  theme(legend.position = 'none') +
  ggtitle('Scatterplot of Honey Produced vs Apples Produced') +
  xlab('Honey Produced (lbs)') +
  ylab('Apples Produced (lbs)')
```



It is incredibly difficult to see any trends in this data as the amount of data is extremely limited. Also, since this data did not come from experimentation, there is no control so many outside factors can be on influence.

Predicting the Future

Although predictions should only be made within the range of the data in regression, extrapolation makes sense for time series data for being able to make future predictions. I am interested in predicting the future honey production, amount of honey yielding colonies, and price per pound of honey from 2022 to 2030.

```
reg1 <- df4 %>%
  filter(State == 'Us Total')
Predictions <- as.data.frame(c(2022:2030))
names(Predictions) <- 'Year'
```

First let's see how much the amount of colonies and production of honey has decreased since 1987.

```
1 - (reg1$`Honey(lbs)`[35] / reg1$`Honey(lbs)`[1])
```

```
## [1] 0.4628434
```

```
1 - (reg1$Colonies[35] / reg1$Colonies[1])
```

```
## [1] 0.1869723
```

Since 1987, honey production has decreased 46% and the amount of colonies has decreased 19%.

```
model1 <- lm(`Honey(lbs)` ~ Year, data = reg1)
summary(model1)
```

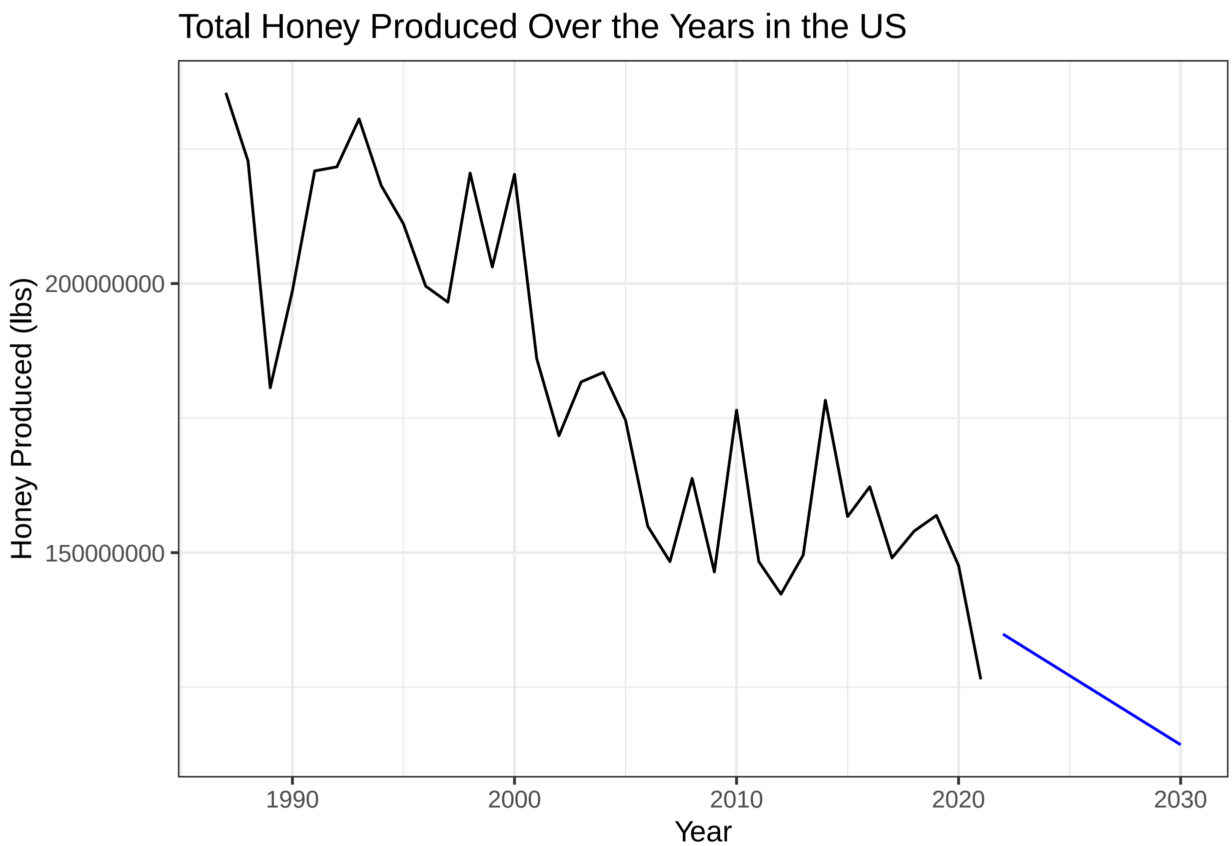
```
##
## Call:
## lm(formula = 'Honey(lbs)' ~ Year, data = reg1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -39011319  -9704990   1312010  10195187  28899810
##
## Coefficients:
##                Estimate Std. Error t value         Pr(>|t|)
## (Intercept) 5329409719  521216050   10.225 0.00000000000925 ***
## Year          -2569012     260085   -9.878 0.00000000002202 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15540000 on 33 degrees of freedom
## Multiple R-squared:  0.7473, Adjusted R-squared:  0.7396
## F-statistic: 97.57 on 1 and 33 DF,  p-value: 0.00000000002202
```

Our beta for Year has a p-val of approximately 0. This means we would reject the null hypothesis that beta = 0 and conclude that Year is significant in predicting pounds of honey produced. This model has an adjusted R-squared value of 0.7396 which indicates that this is a good model for predicting honey productions.

```
Predictions$`Honey(lbs)` <- predict(model1, Predictions)
```

Let's plot our predicted data with the actual data.

```
df4 %>%
  filter(State == 'Us Total') %>%
  ggplot(aes(x = Year, y = `Honey(lbs)`)) +
  geom_line() +
  theme_bw() +
  ggtitle('Total Honey Produced Over the Years in the US') +
  xlab('Year') +
  ylab('Honey Produced (lbs)') +
  geom_line(data = Predictions, aes(x = Year, y = `Honey(lbs)`), color = 'blue')
```




```
1 - (Predictions$`Honey(lbs)`[9] / reg1$`Honey(lbs)`[1])
```

```
## [1] 0.5144505
```

```
1 - (Predictions$`Honey(lbs)`[9] / reg1$`Honey(lbs)`[35])
```

```
## [1] 0.09607454
```

According to this our model, it is predicted that by 2030, total honey production will have decreased 51% since 1987 and 10% since last year.

Next we will predict the amount of honey producing bee colonies

```
model2 <- lm(Colonies ~ Year, data = reg1)
summary(model2)
```

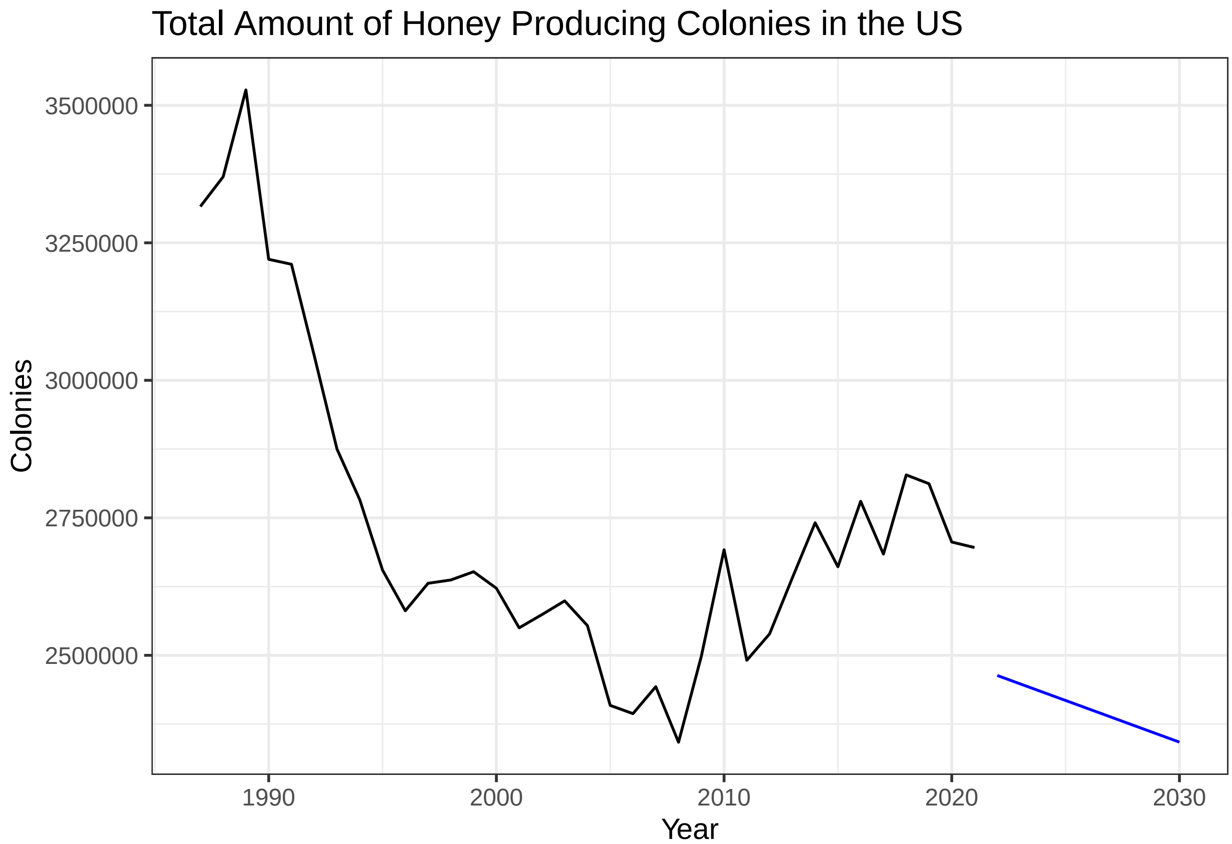
```
##
## Call:
## lm(formula = Colonies ~ Year, data = reg1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -333417 -191083  -75835  214903  564815
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33087899    8187086   4.041 0.000299 ***
## Year          -15146       4085  -3.707 0.000766 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 244100 on 33 degrees of freedom
## Multiple R-squared:  0.294,  Adjusted R-squared:  0.2726
## F-statistic: 13.74 on 1 and 33 DF,  p-value: 0.0007658
```

Just like the previous model, we reject the beta for Year and conclude that it is significant in predicting amount of colonies. However, this model is significantly weaker at an adjusted R-squared of 0.2726.

```
Predictions$`Colonies` <- predict(model2, Predictions)
```

Let's plot our predicted data with the actual data.

```
df4 %>%
  filter(State == 'Us Total') %>%
  ggplot(aes(x = Year, y = Colonies)) +
  geom_line() +
  theme_bw() +
  ggtitle('Total Amount of Honey Producing Colonies in the US') +
  xlab('Year') +
  ylab('Colonies') +
  geom_line(data = Predictions, aes(x = Year, y = Colonies), color = 'blue')
```



Even though we can see a quadratic relationship, I was hesitant to use polynomial regression since the model is very simple with only 1 predictor and polynomial regression can very easily lead to incorrect results when extrapolating.

```
1 - (Predictions$Colonies[9] / reg1$Colonies[1])
```

```
## [1] 0.2936632
```

```
1 - (Predictions$Colonies[9] / reg1$Colonies[35])
```

```
## [1] 0.1312267
```

Our model predicts that the amount of colonies in 2030 will have decreased by 29% since 1987 and 13% since last year.

```
Predictions
```

```
##   Year Honey(lbs) Colonies
## 1 2022  134867931  2463378
## 2 2023  132298919  2448232
## 3 2024  129729908  2433087
## 4 2025  127160896  2417941
## 5 2026  124591884  2402796
## 6 2027  122022872  2387650
## 7 2028  119453861  2372504
## 8 2029  116884849  2357359
## 9 2030  114315837  2342213
```

Results of the Analysis

The analysis shows that there was a huge drop in the total production of honey and amount of bee colonies. However, due to the lack of quantity and control in the data, it can't be determined if the amount of apples produced has been affected by the decrease in pollinators without bias. With supply decreasing and assuming demand stays the same, it is reasonable to expect an increase in honey prices independent from inflation.