Springboard

Springboard Capstone 2 project report 2

How to promote the honey production?

Yuanchun Wang

06/25/2020

Honey and Honeybee

1. Honey is an important food source and the production is decreasing.
2. Honeybee colony collapse disorder is getting worse.
3. How to use neonics pesticide to save the honeybee and

Increse the honey production

Outline

- Data acquiring
- Data wrangling
- Data visualization
- Machine learning analysis
- Summary
- What’s next

Data acquiring

- Honey production data is from Kaggle website:

https://www.kaggle.com/jessicali9530/honey-production

- Honeybee neonics data is from Kaggle website:

https://www.kaggle.com/kevinzmith/honey-with-neonic-pesticide#vHoneyNeonic_v03.csv

- Related references are downloaded from the google scholar.

Who is this project for?

- The beekeepers: give the suggestions on how to increase the honey production and how to apply neotics to promote honeybee colony number increasing
- the customers who consume the honey: provide historical data analysis and production prediction to let everyone know that how the honey production will develop in future.
- Finally, this project will give suggestions on how to promise the enough honey providing in the market for consumers.

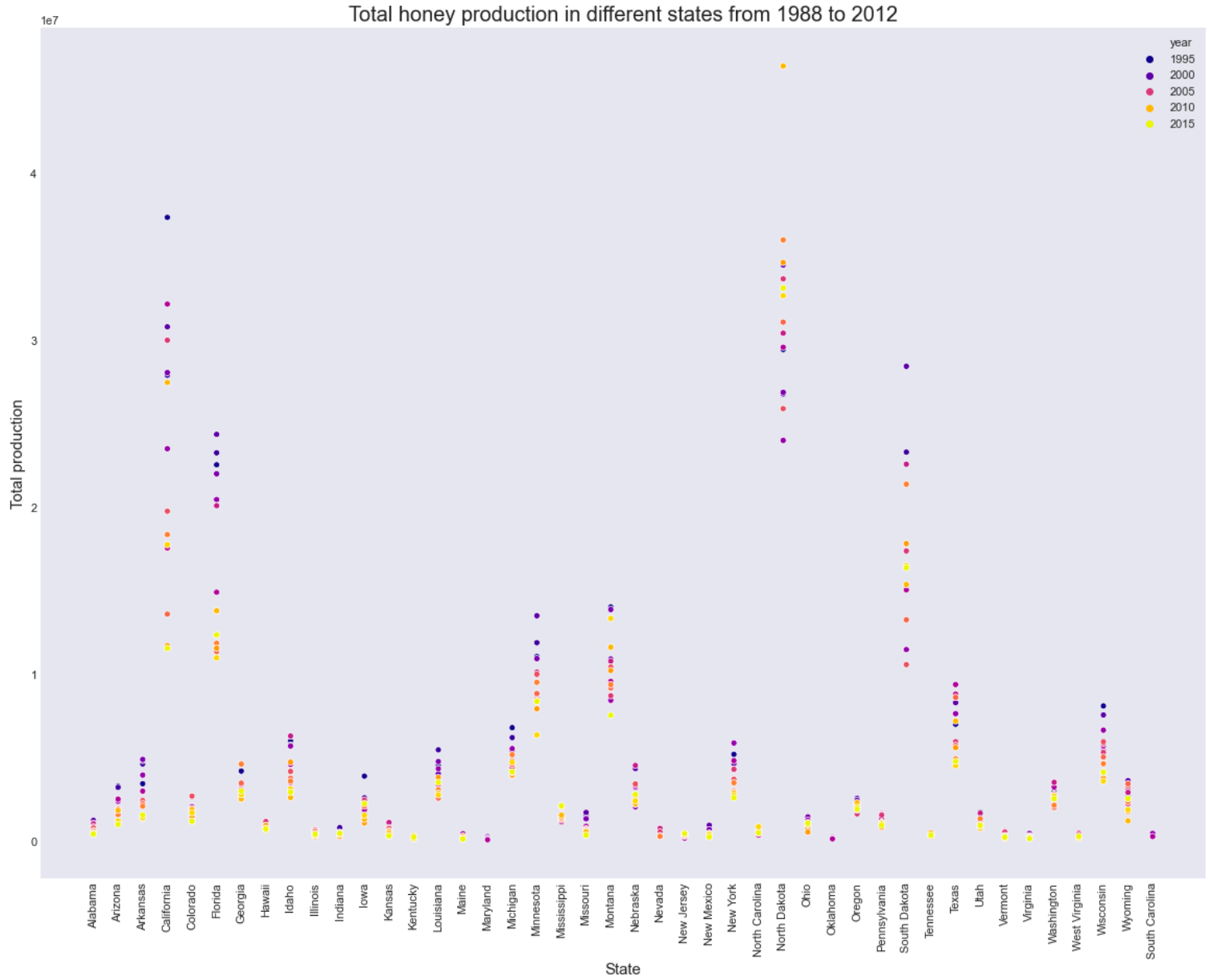
Data Wrangling

1. Rename the columns to make the data easy to read.
2. Drop useless column ‘FLPS’.
3. Fill the empty space with 0.
4. Replace ‘state_code’ with full state name.

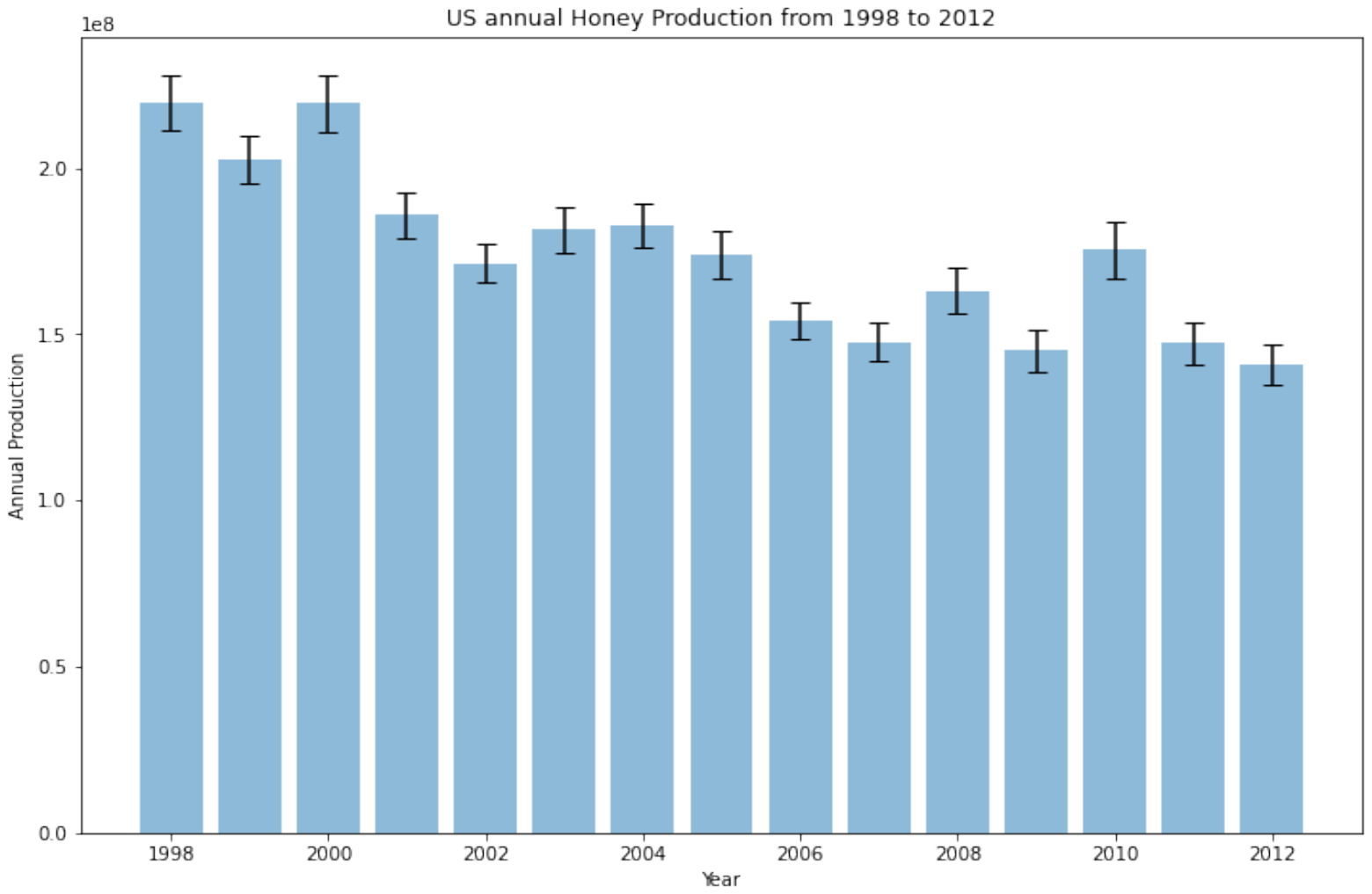
After applying the above steps, there are 626 rows left; for neonics data, after wrangling, there are 1132 rows left.

Data Visualization

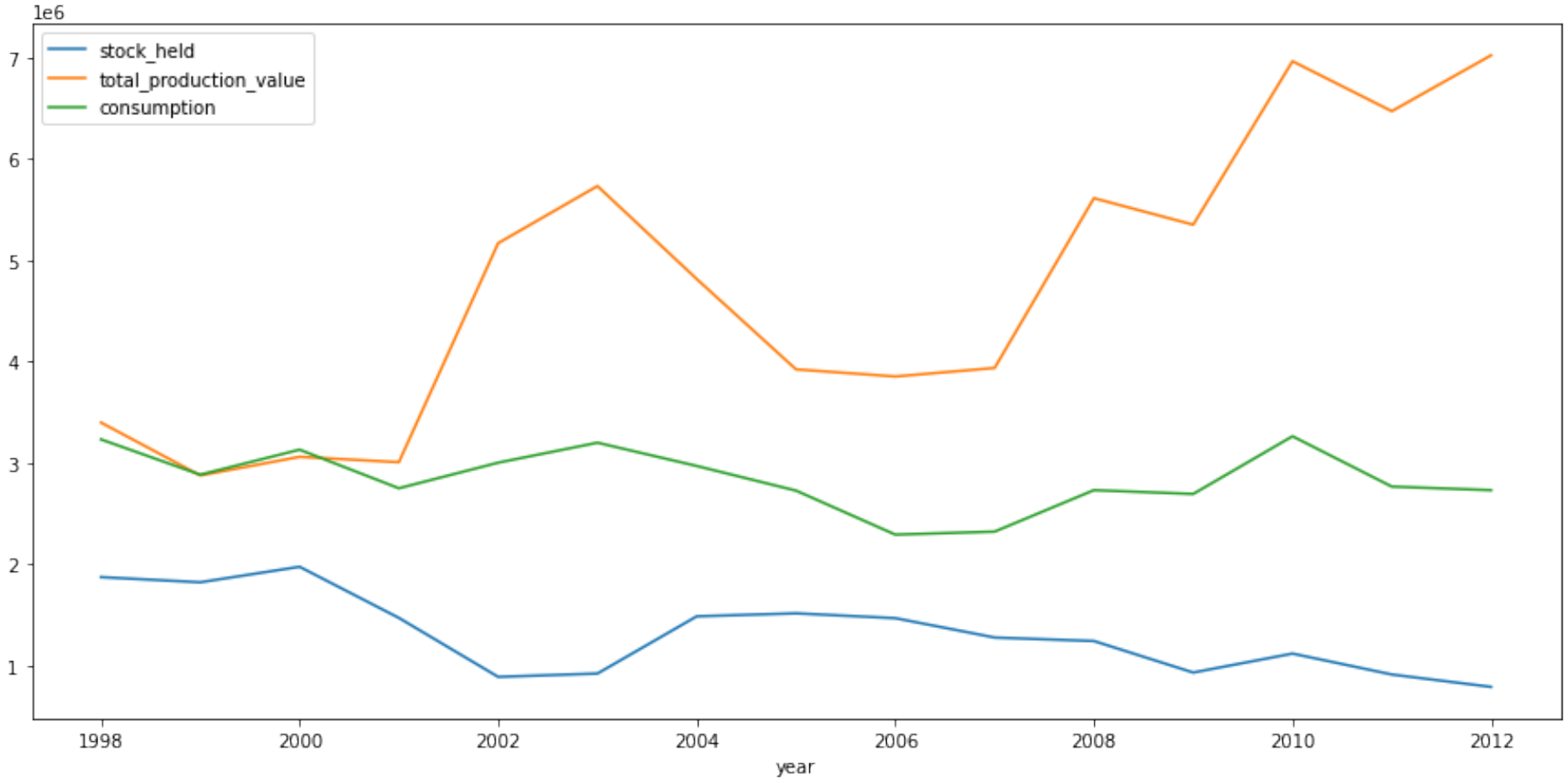
The honey production decrease in most of the states.



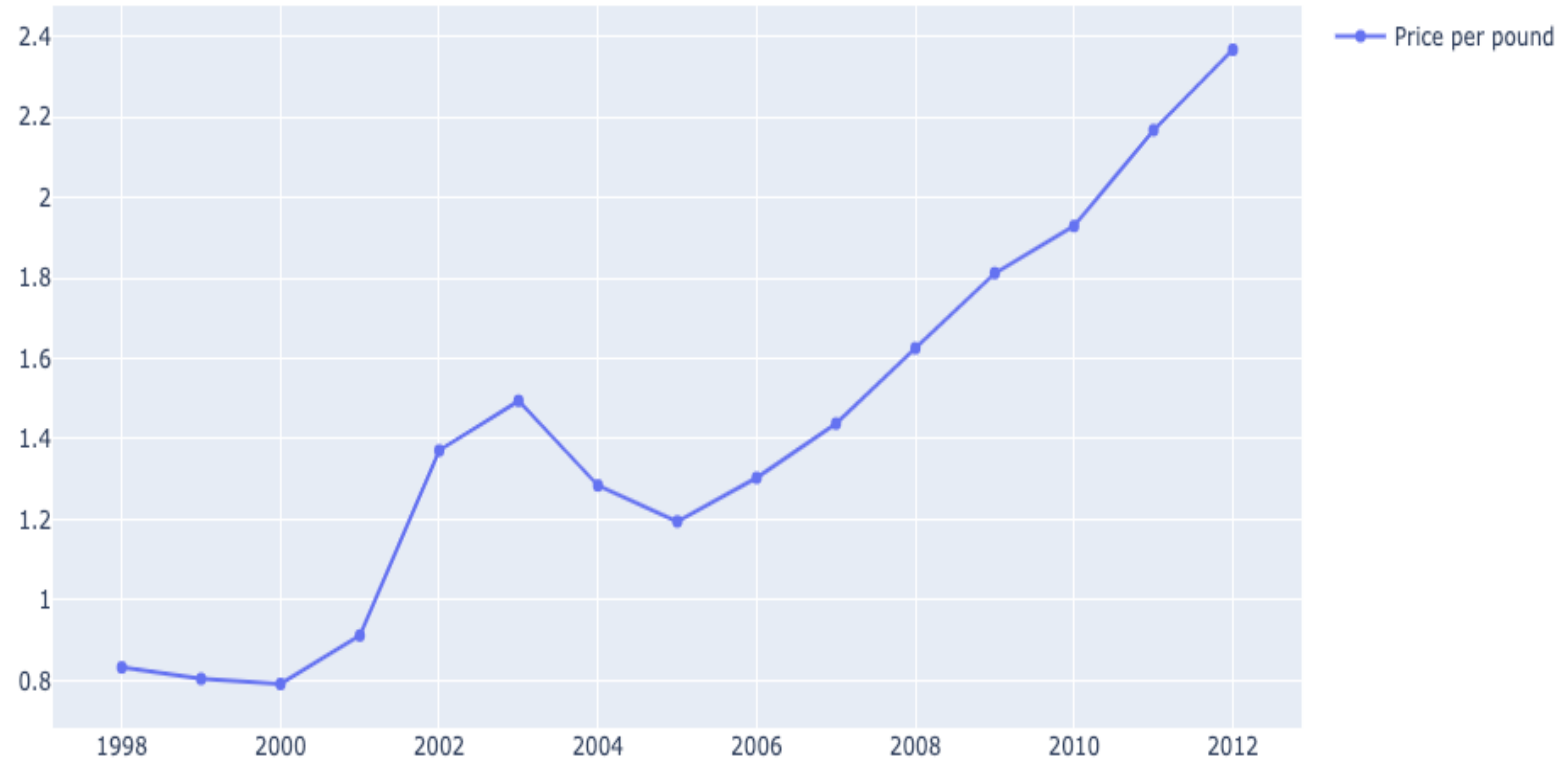
The USA produces the maximum honey in 2000



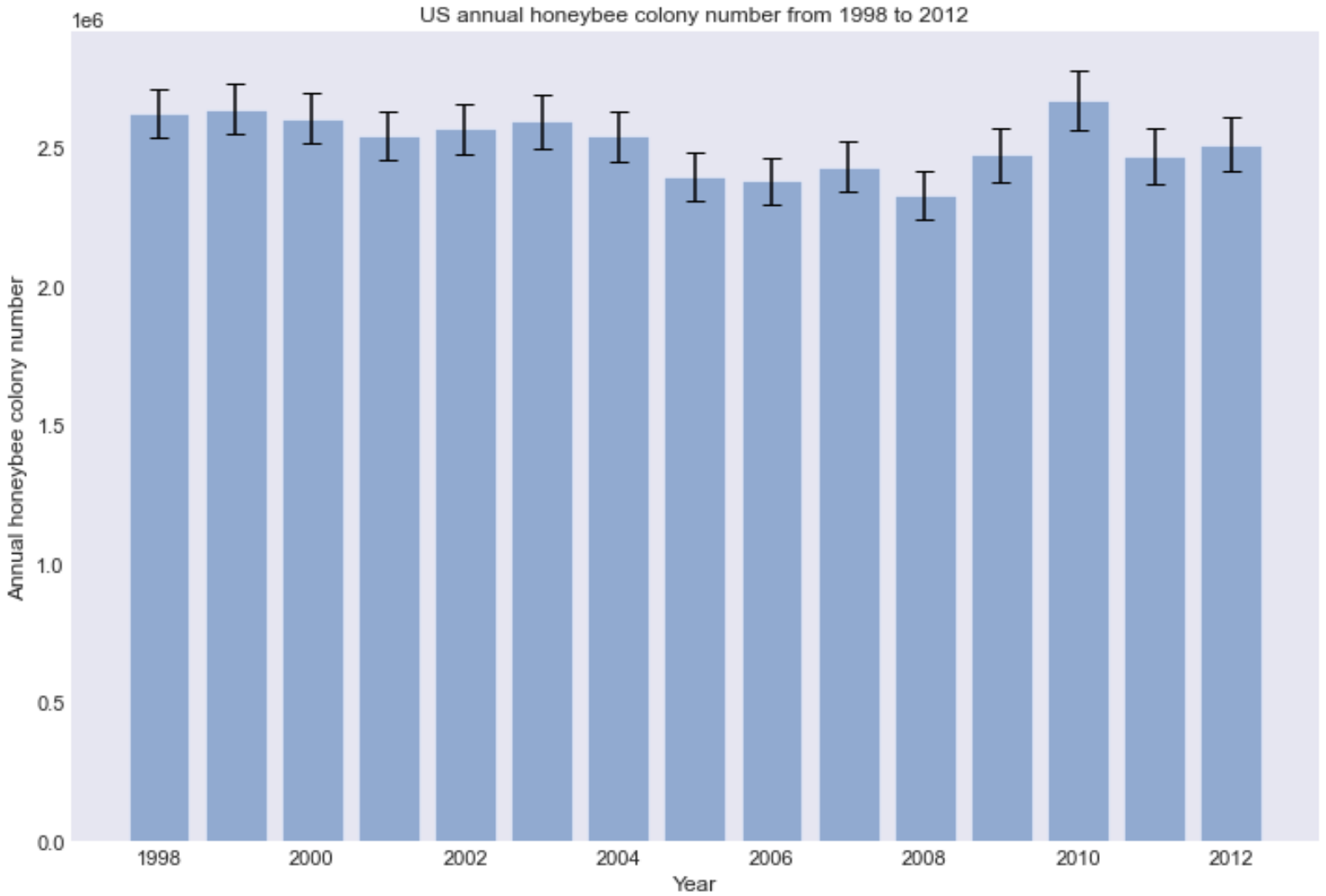
The honey stock, total production value and consumption changing trend



The honey price exhibits the almost increasing trend

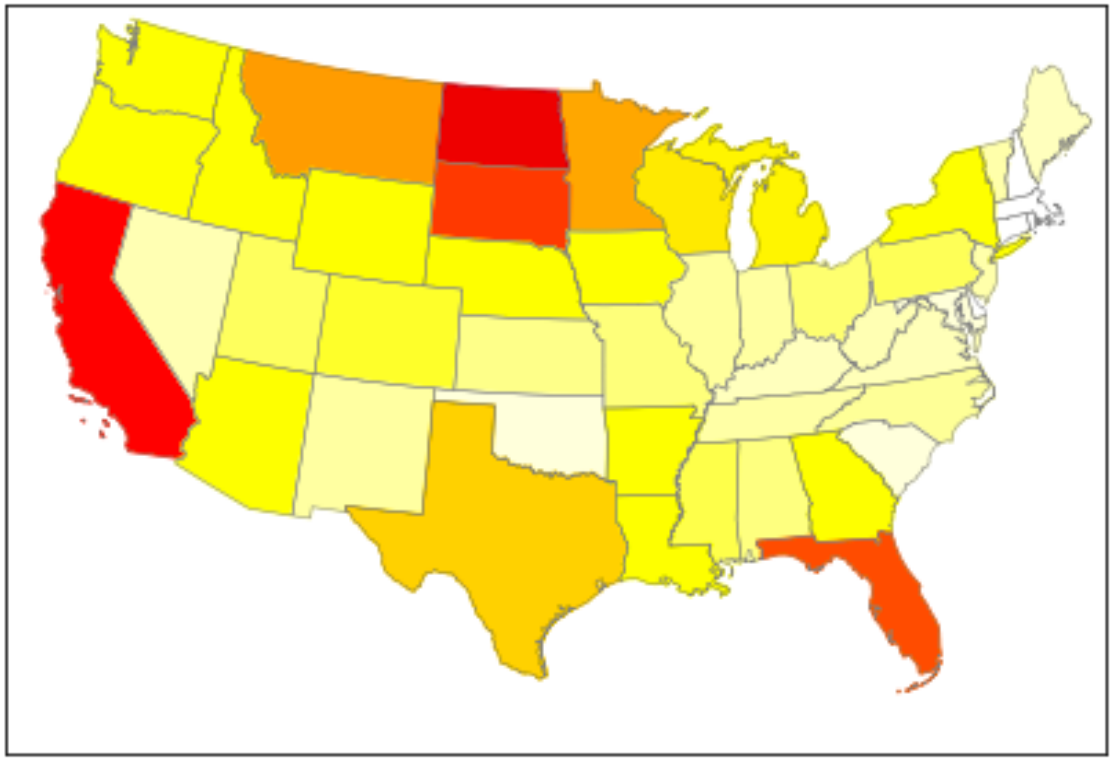


US annual honeybee colony number are relatively stable during 1998 and 2012.

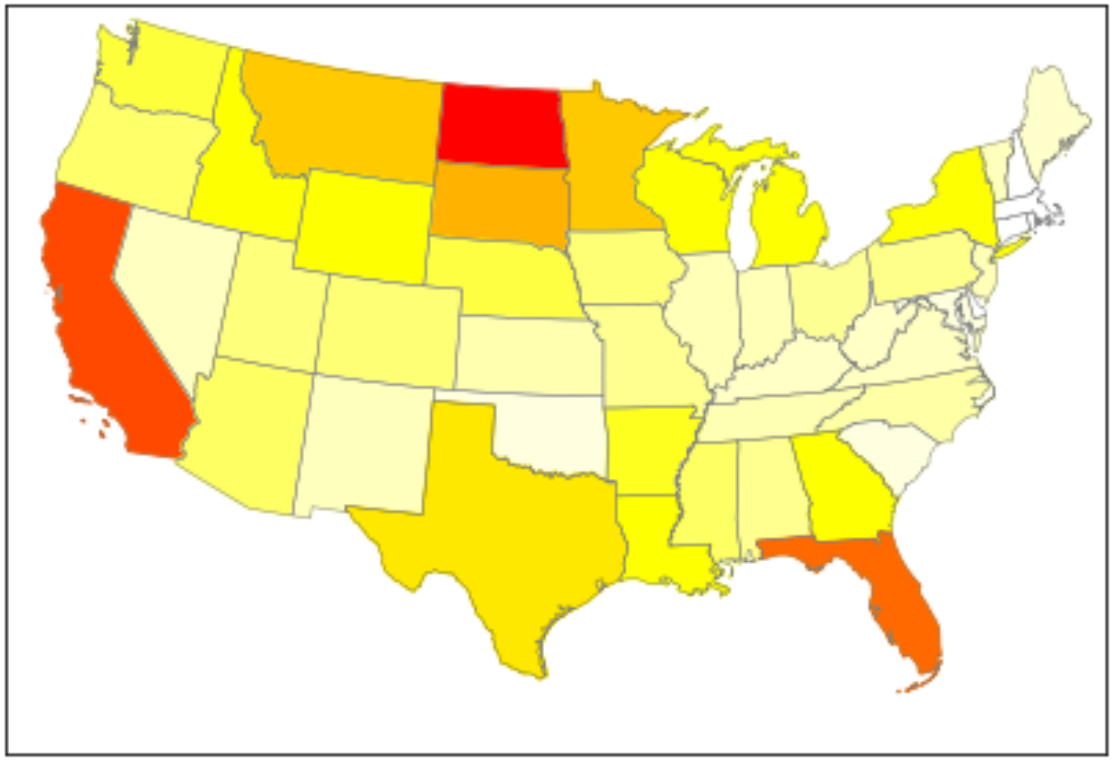


Top honey producing and consuming state map in USA

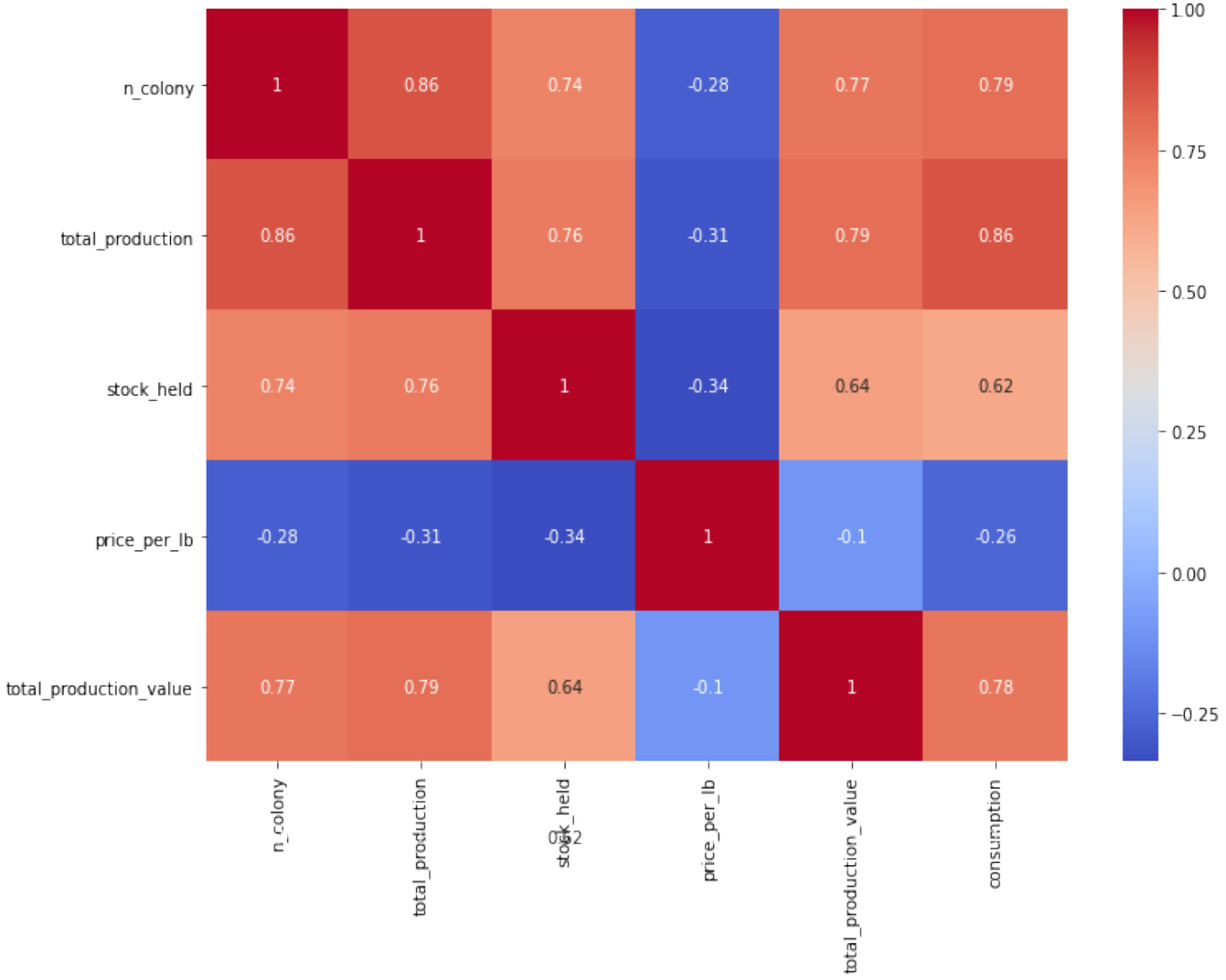
Top honey producing states in the USA



Top honey consuming states in USA



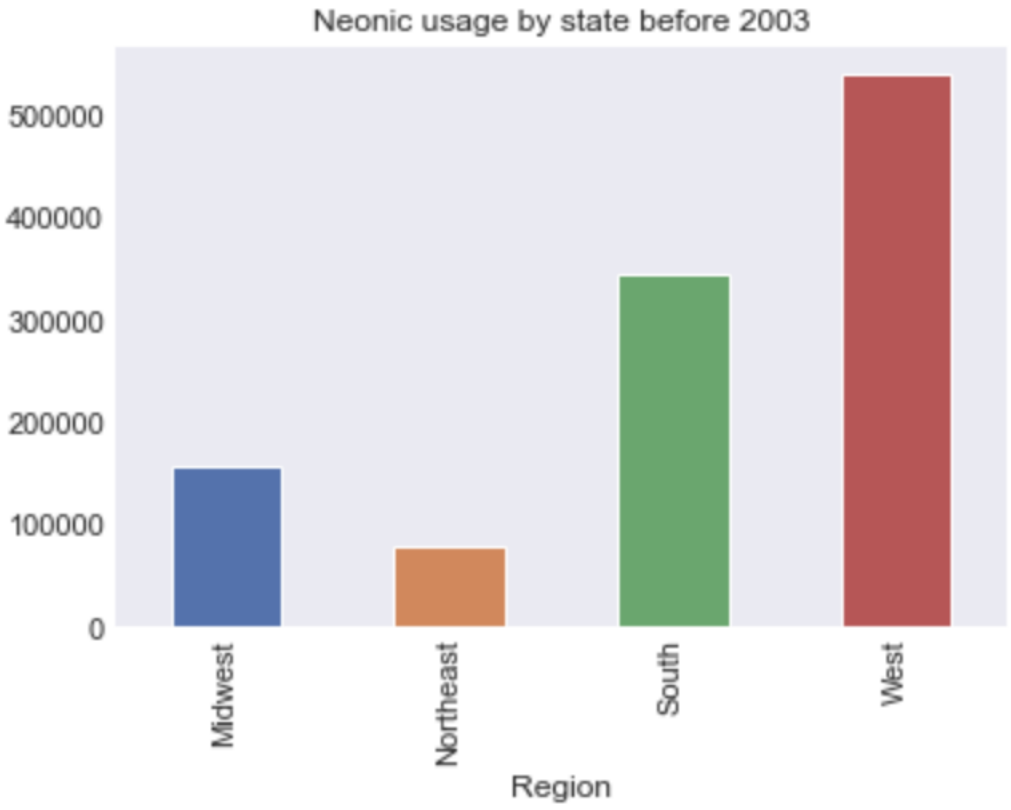
The kendall correlation between honey price and production

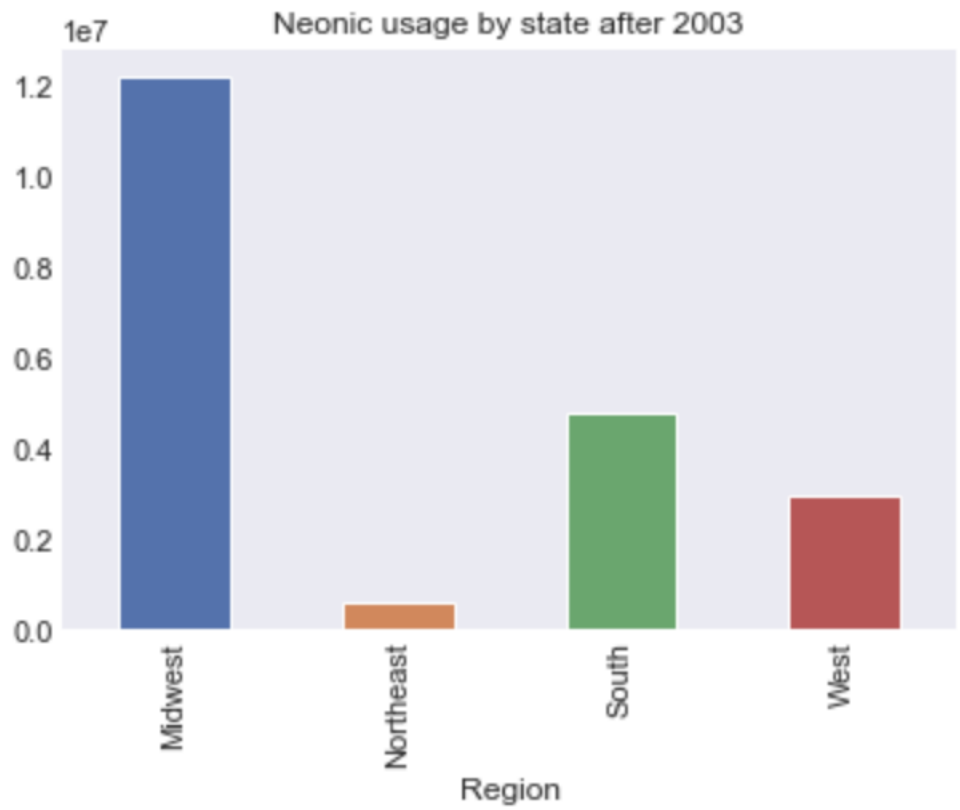


Summary I

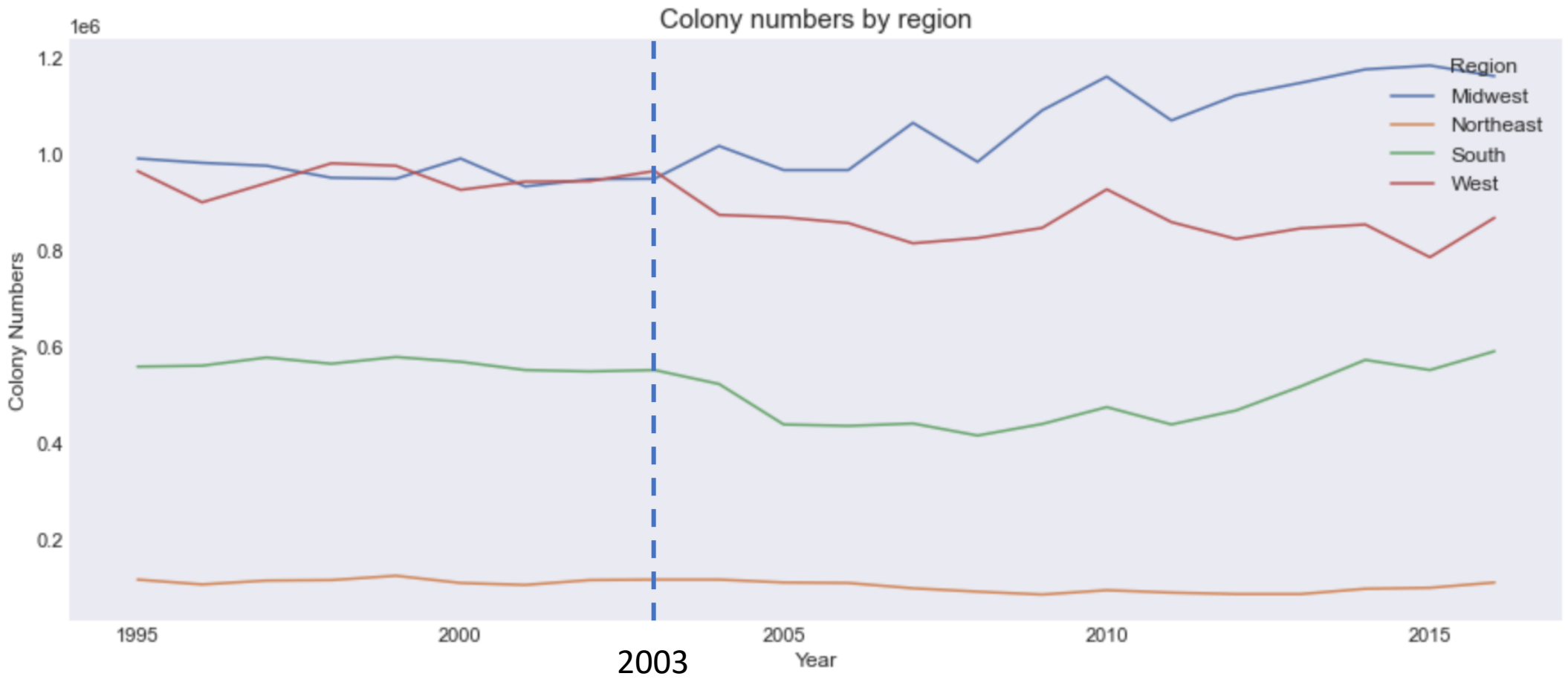
- Honey price per pound has negative correlation with 'number of colony', 'total production' and 'stocks' at the correction value of '-0.28', '-0.31', '-0.34', which indicates that when the honey colony become less or total production goes down or stocks decreases, the honey price per pound increases.
- Colony number has strong correlation with 'total production'(0.86), 'stock held'(0.74), 'total production value'(0.77) and 'consumption'(0.79); with the colony number increasing, the honey production, stock and total production value all goes up.
- Consumption has strong correlation with 'total production'(0.86), too; it indicates that the more total production, the more consumption; if we want to increase the consumption, we have to improve production.
- The colony number plays a key in role in influencing the production. The effect of neonics need to be understood more deeply.
- Let' s check how to use neonics properly to increase the colony number.

Neonic usage before 2003 and after 2003 by region





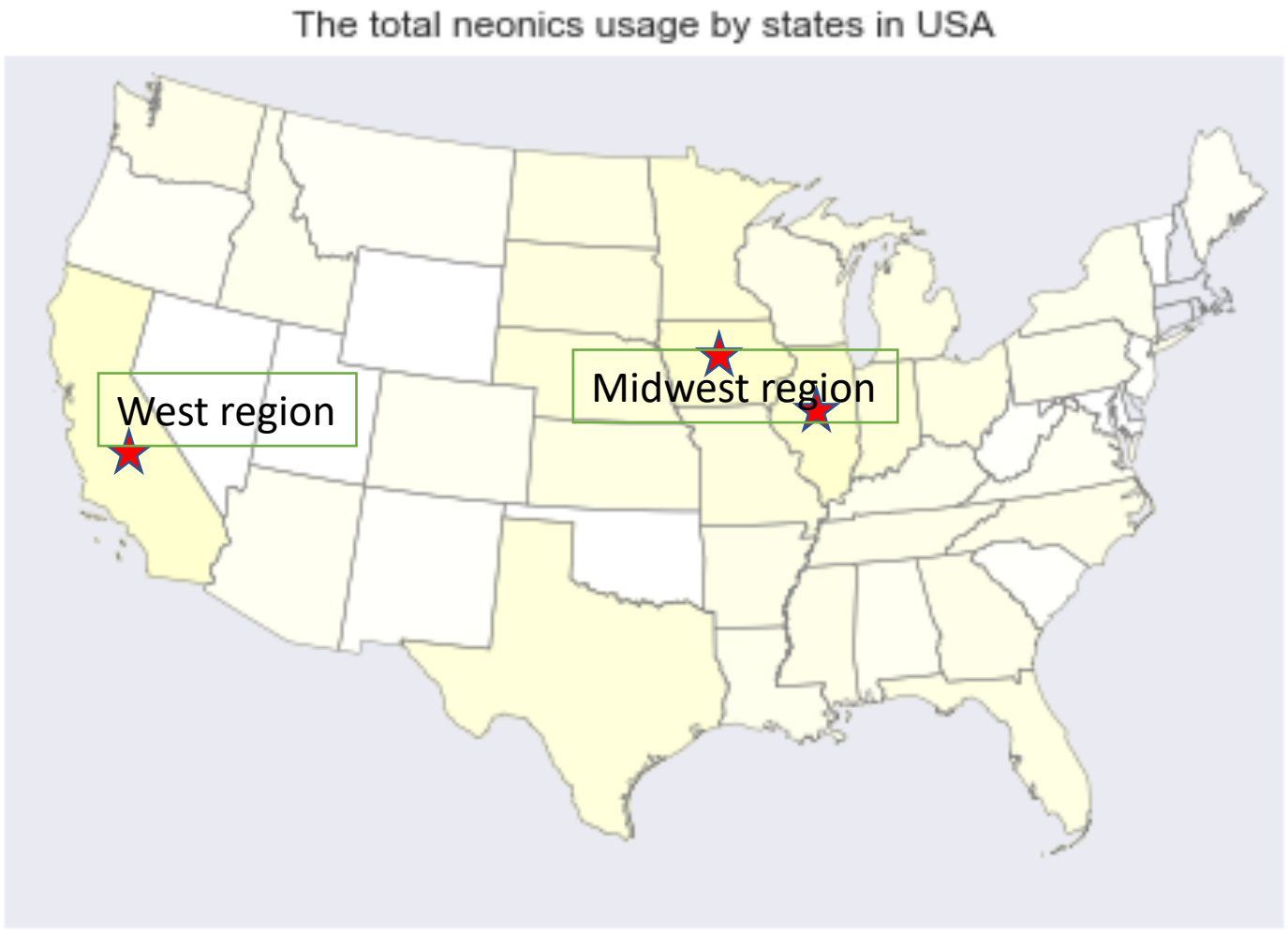
Total colony number changing before 2003 and after 2003 by region



The neonics usage map in the USA

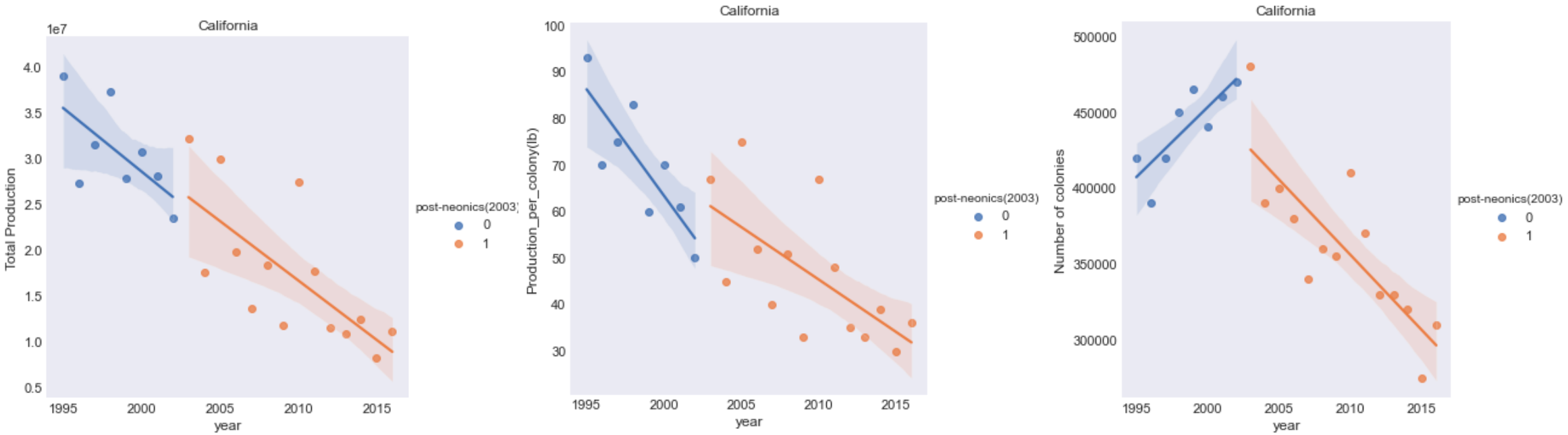
Top 3 neonic-usage states:

state	AllNeonic (kg)
California	1991179.1
Illinois	1978523.1
Iowa	1974038.9



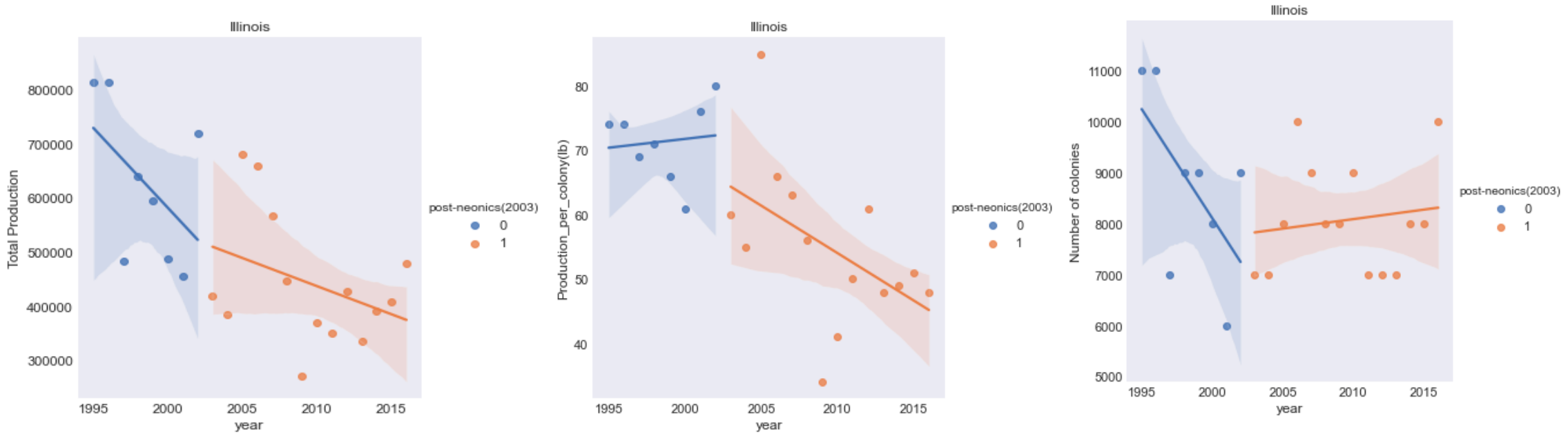
Note: The ‘star’ marks the top 3 states on neonics usage.

The neonics-usage effect in California



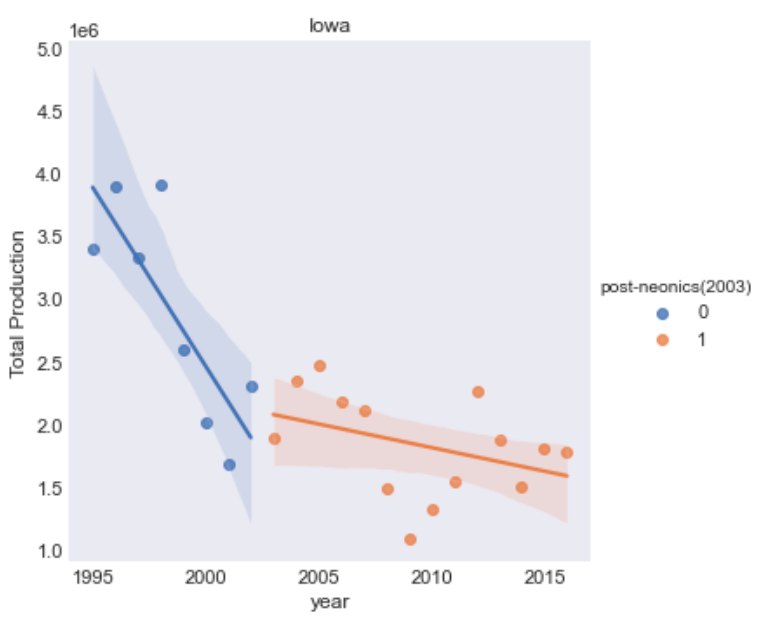
California's no. of colonies, yield per colony and total production have been decreasing consistently since their frequent heavy use of neonics in 1994.

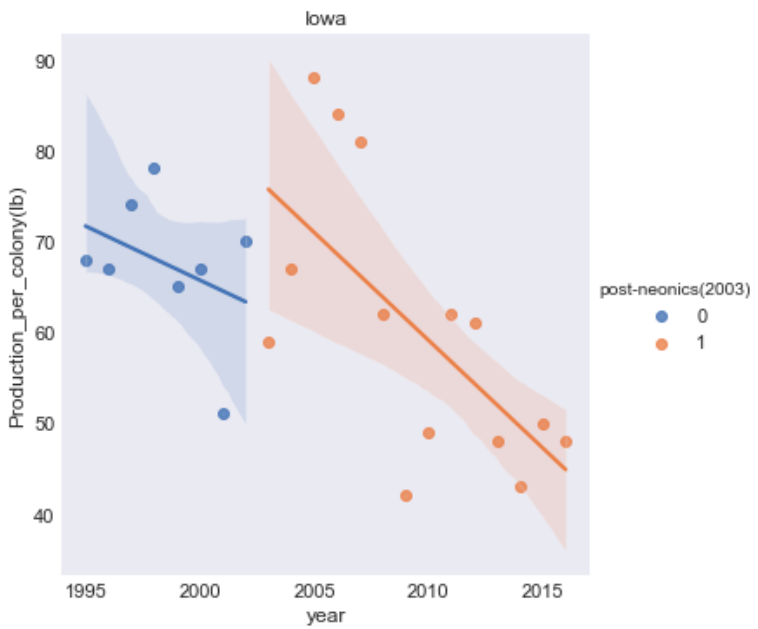
The neonics-usage effect in Illinois

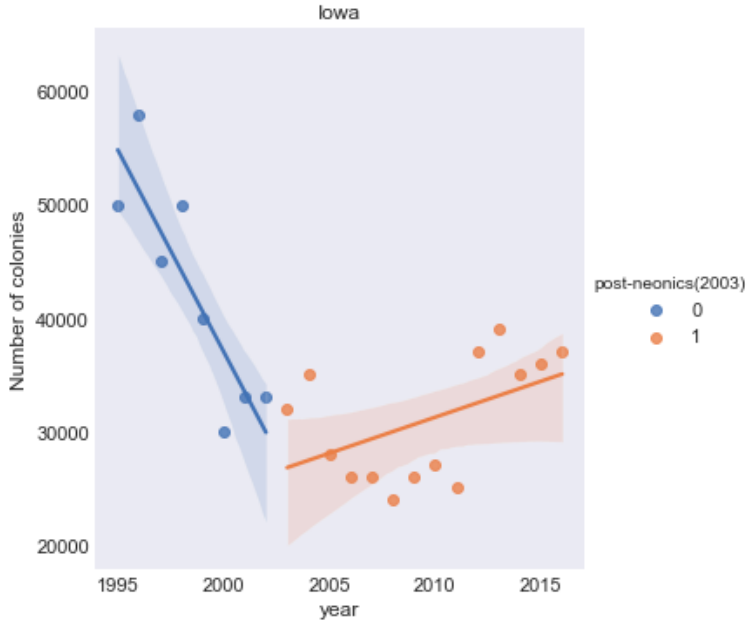


By applying neonics, Illinois's production per colony id sligtly decreasing after 2003 even though the colony number is increasing slowly.

The neonics-usage effect in Iowa

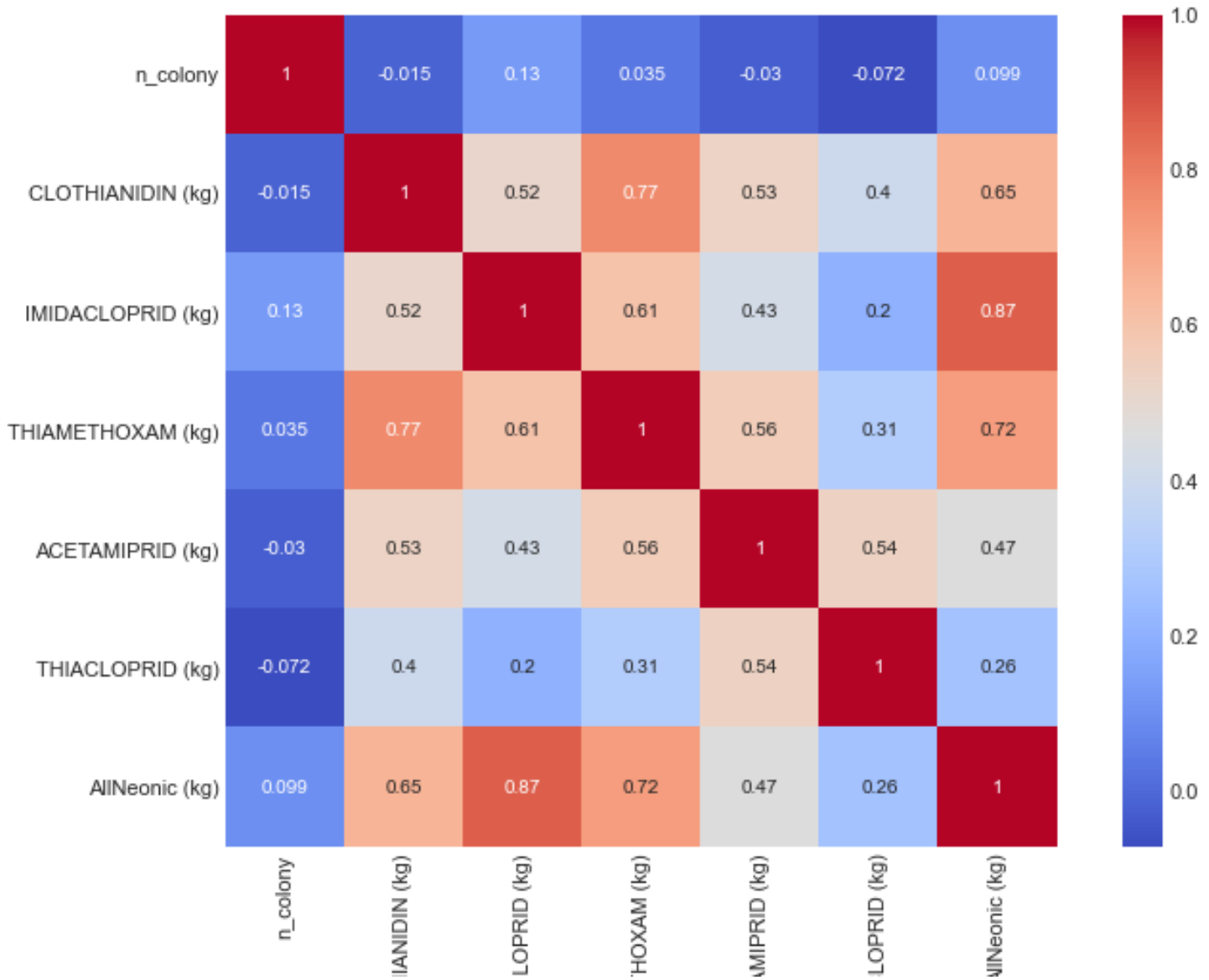






Similar to Illinois, Iowa's total production and number of colonies were decreasing before 2003., with yield per colony increasing.

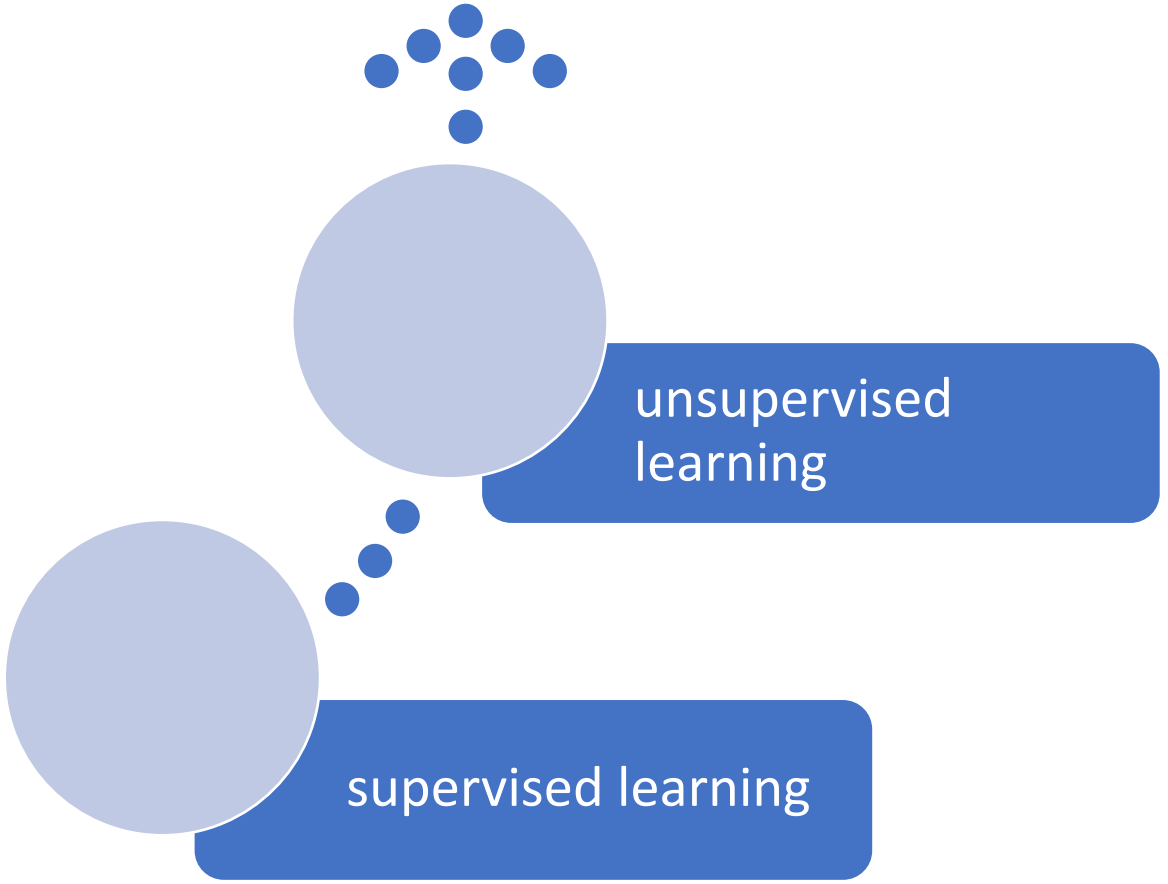
The kendall correlation between neonics application and honeybee colony numbers



Summary II

- The neonics are applied in USA since 2003 to control the colony collapse disorder (CDD). Then we analyze the correlation between five kinds of neonics and the colony number.
- All the five kinds of neonics exhibit different correlation trend with honeybee colony number; however, allneonic has positive correlation with the colony number at the value of 0.099, which indicates the application of neonic pesticide could promote the honeybee developing.
- Among the neonics, IMIDACLOPRID (corr=0.13) plays a key role in promoting honeybee developing; it also show strongest correlation with allNeonic at 0.87. Thus, Imidacloprid is the most import neonics in promoting honey propagation.
- The second important one is THIAMETHOXAM (corr=0.035). The rest of neonics all affect the honeybee colony negatively.

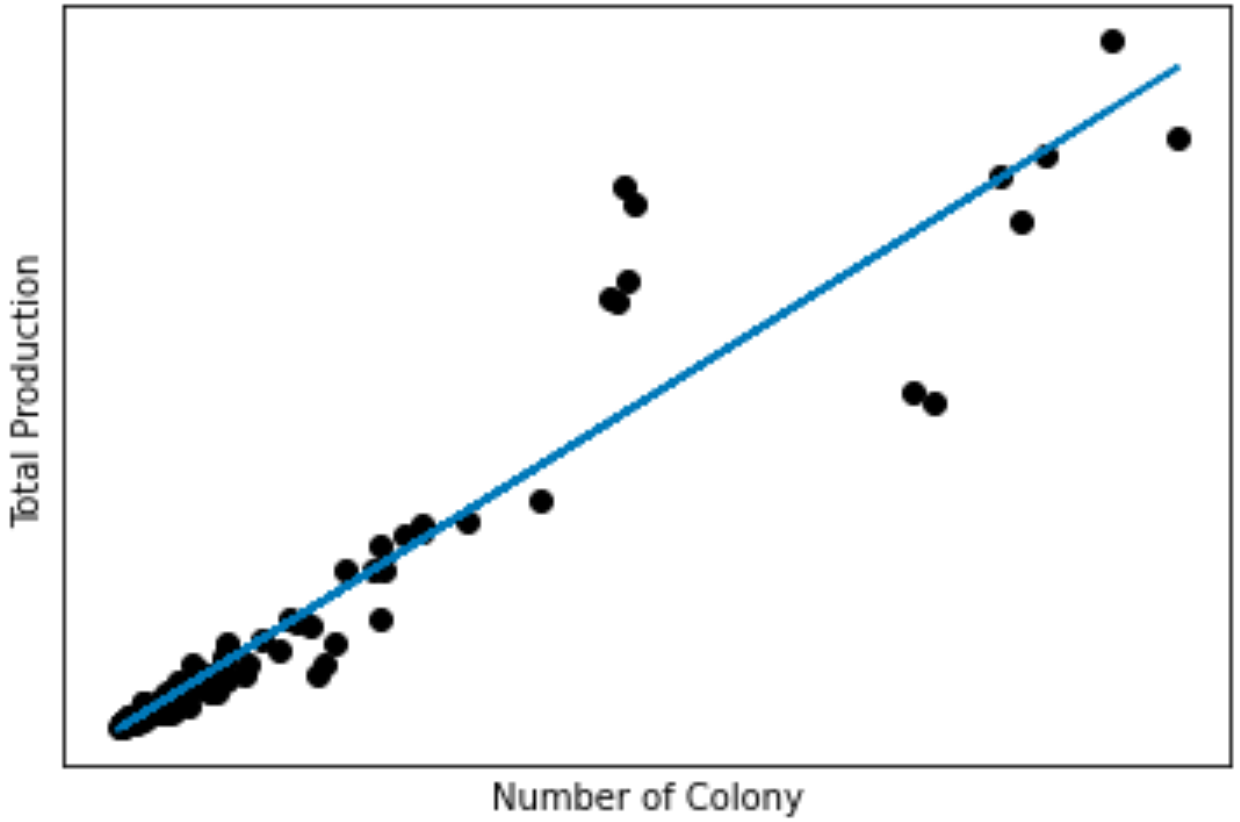
Machine learning analysis



Supervised learning

1. Linear Regression Model

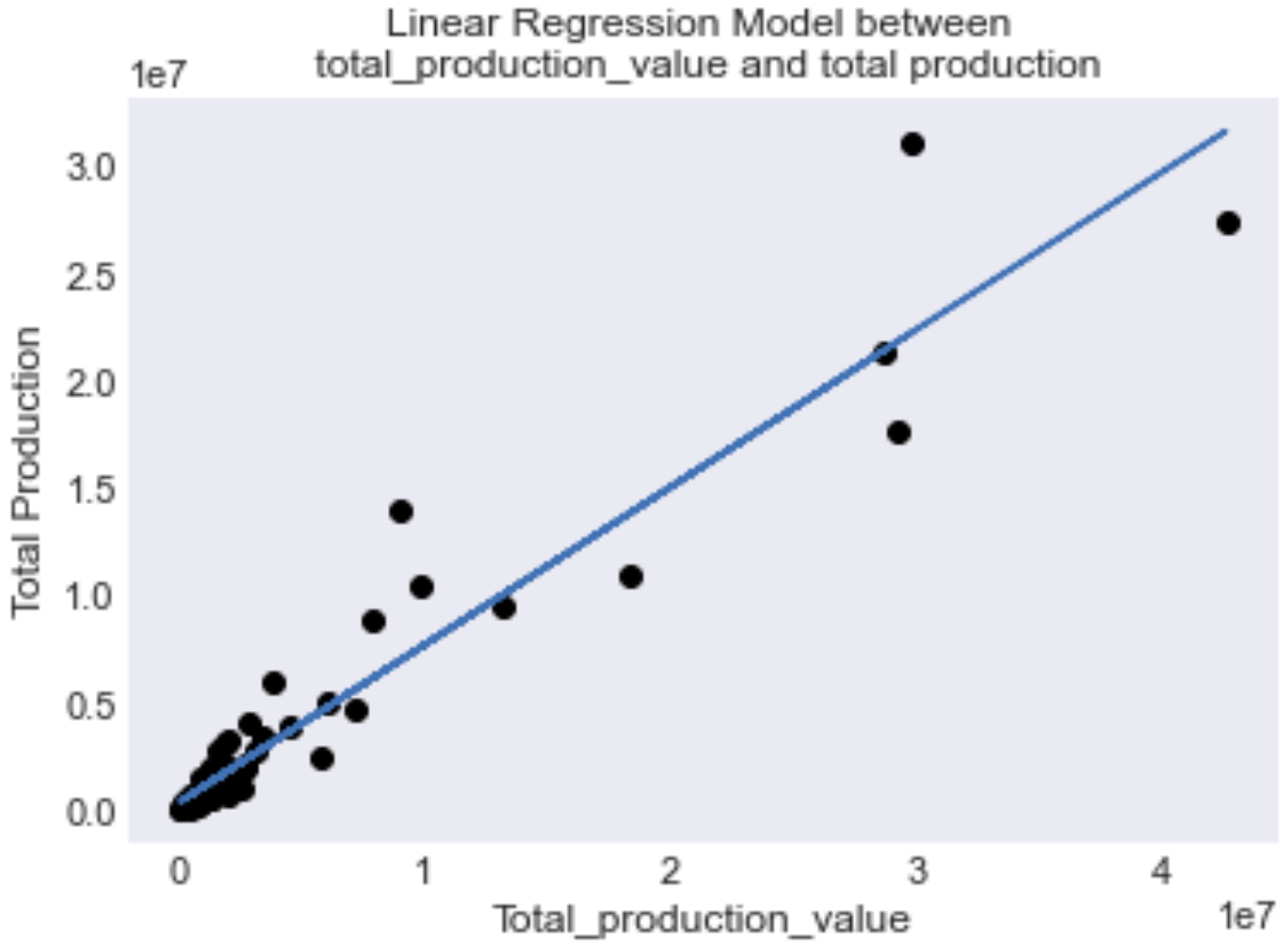
- R^2 score between colony number and total production is 0.9092 based on linear regression model.
- Linear regression model equation is Y=75.176269*x+(-217527.838246)



Supervised learning

1. Linear Regression Model

- R^2 score between production value and total production is 0.7764 based on linear regression model.
- Linear regression model equation is Y=0.735375*x+369372.252656



Supervised learning

- 2. **Decision Tree Model**

The R^2 score between y_test and y_predict is 0.9316 based on decision tree model.

Conclusion:

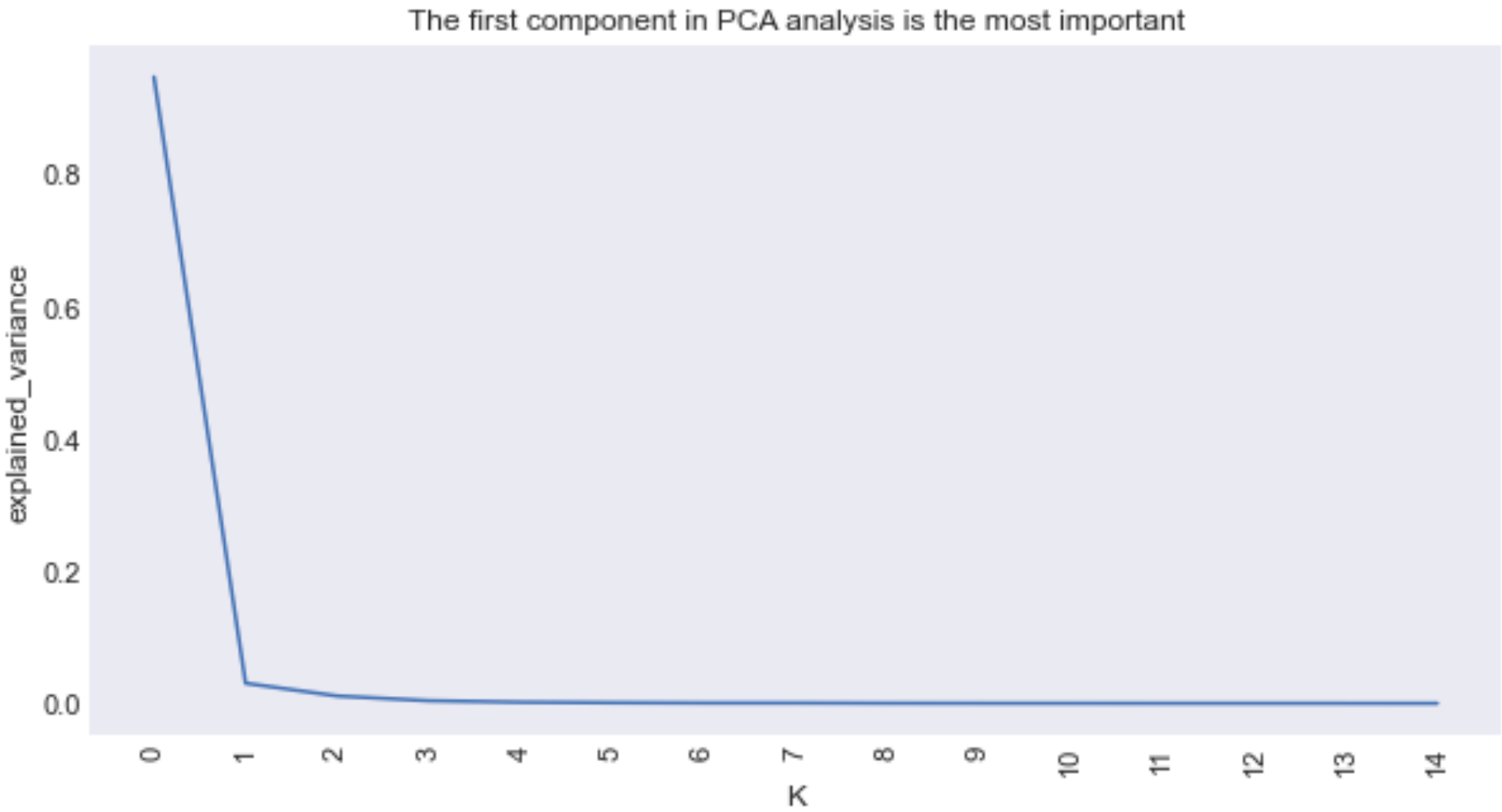
- The linear regression model (R^2=1) is more suitable for this honey production data than decision tree model (R^2=0.93).

Unsupervised learning

PCA analysis

- Principal Component Analysis (PCA) can help us reduce the dimensionality and features of our data.
- Here, PCA analysis is also employed to reduce a large set of variables into a much smaller one.

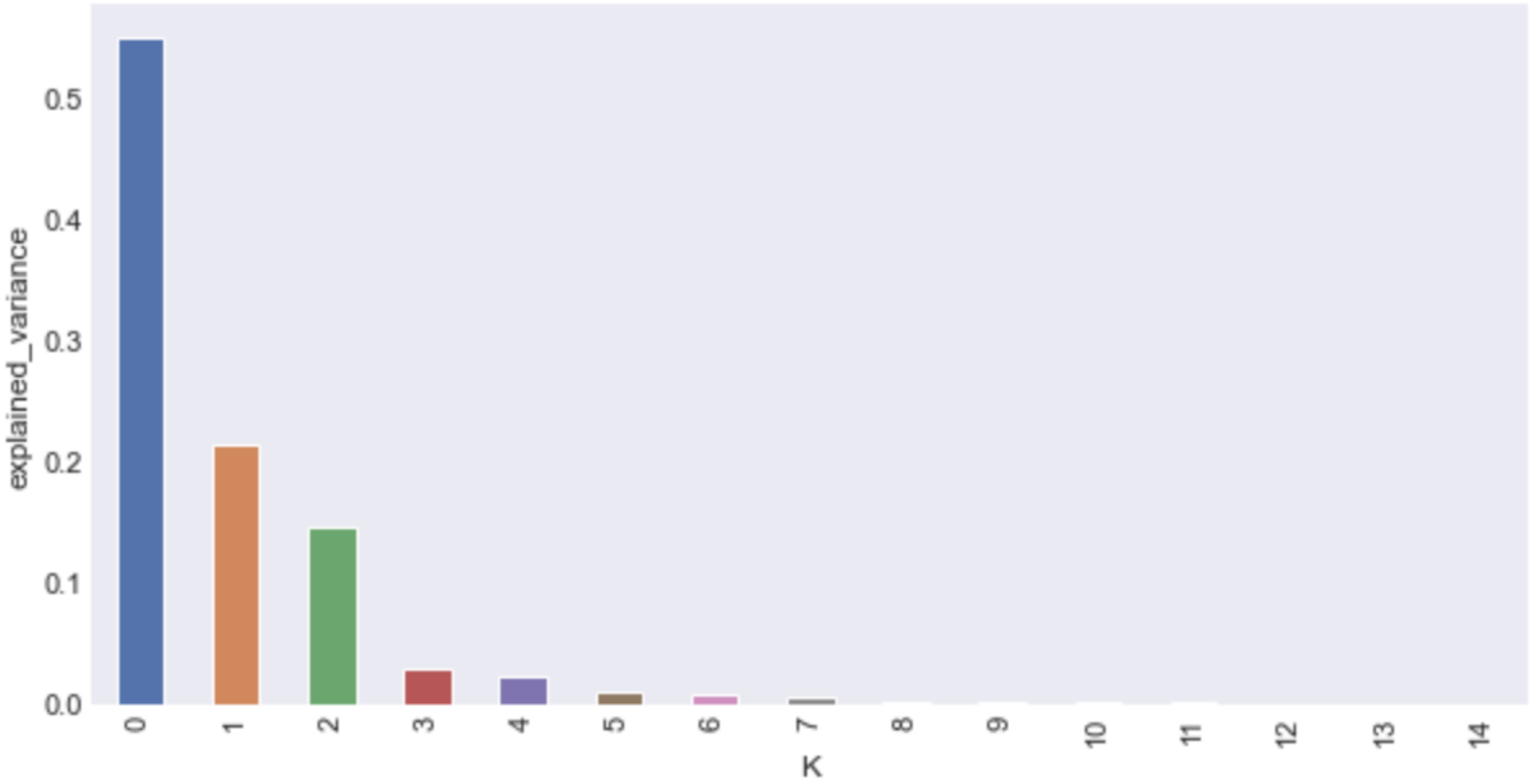
The First PCA analysis



we prepare a matrix including index ‘state’, column ‘year’ and ‘total_production’ as values.

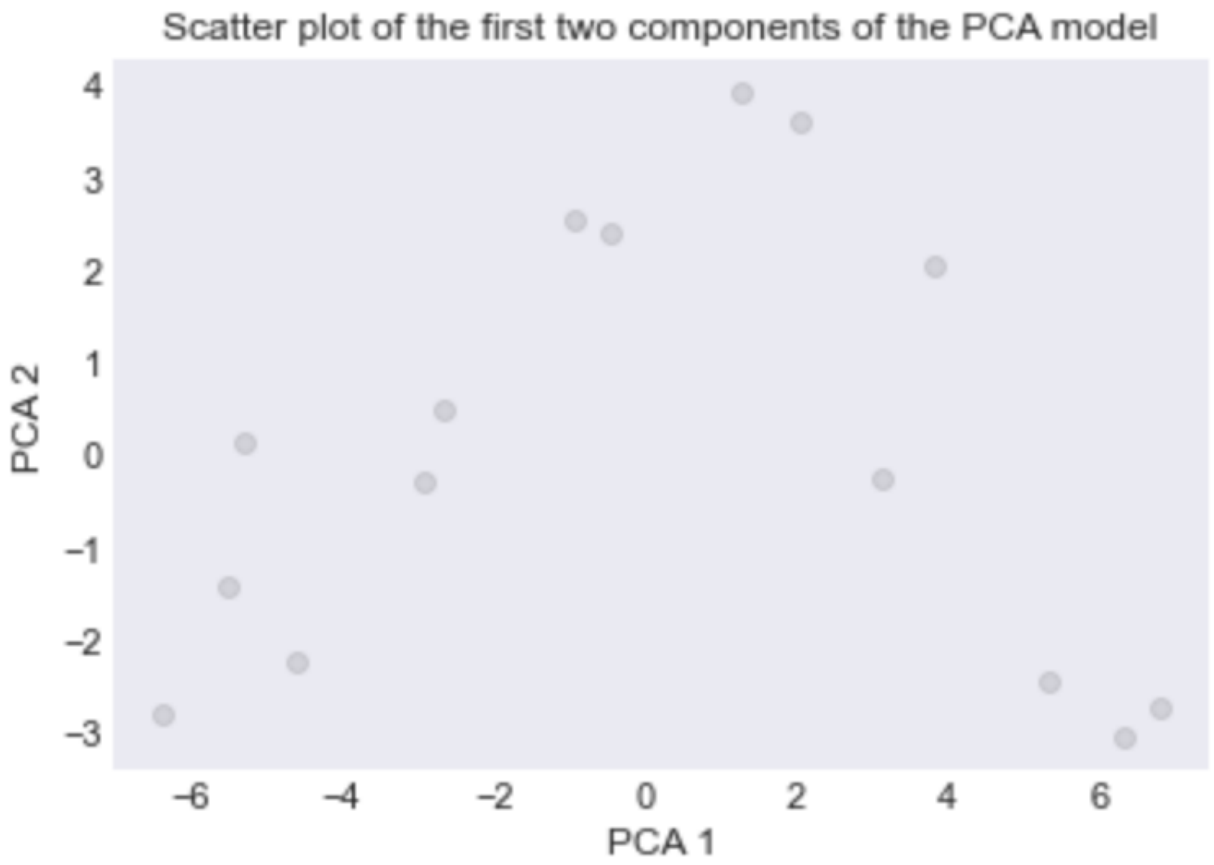
The Second PCA analysis

The first three components are very important among ‘state’ columns.



we prepare a matrix including index ‘year’, column ‘state’ and ‘total_production’ as values.

Between the first two ‘state’ components, there' s no clear clusters between them.



Summary

- Based on the honey production dataset analysis, the colony number has strong correlation with total production and total production values.
- Thus, figure out how to maximize the colony number will provide important suggestions to maximize the honey production, which helps guide honeybee management decisions in the United States.

What’s next

- We will focus on using supervised learning related method to train predictive models and employ cross validation to evaluate the model’ s metrics to find out the best model in neonic usage data.
- Summary the data analysis and present the final report.

Thanks for your attention!