Springboard

Final Springboard Capstone project report

How to promote the honey production?

Yuanchun Wang

06/28/2020

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?

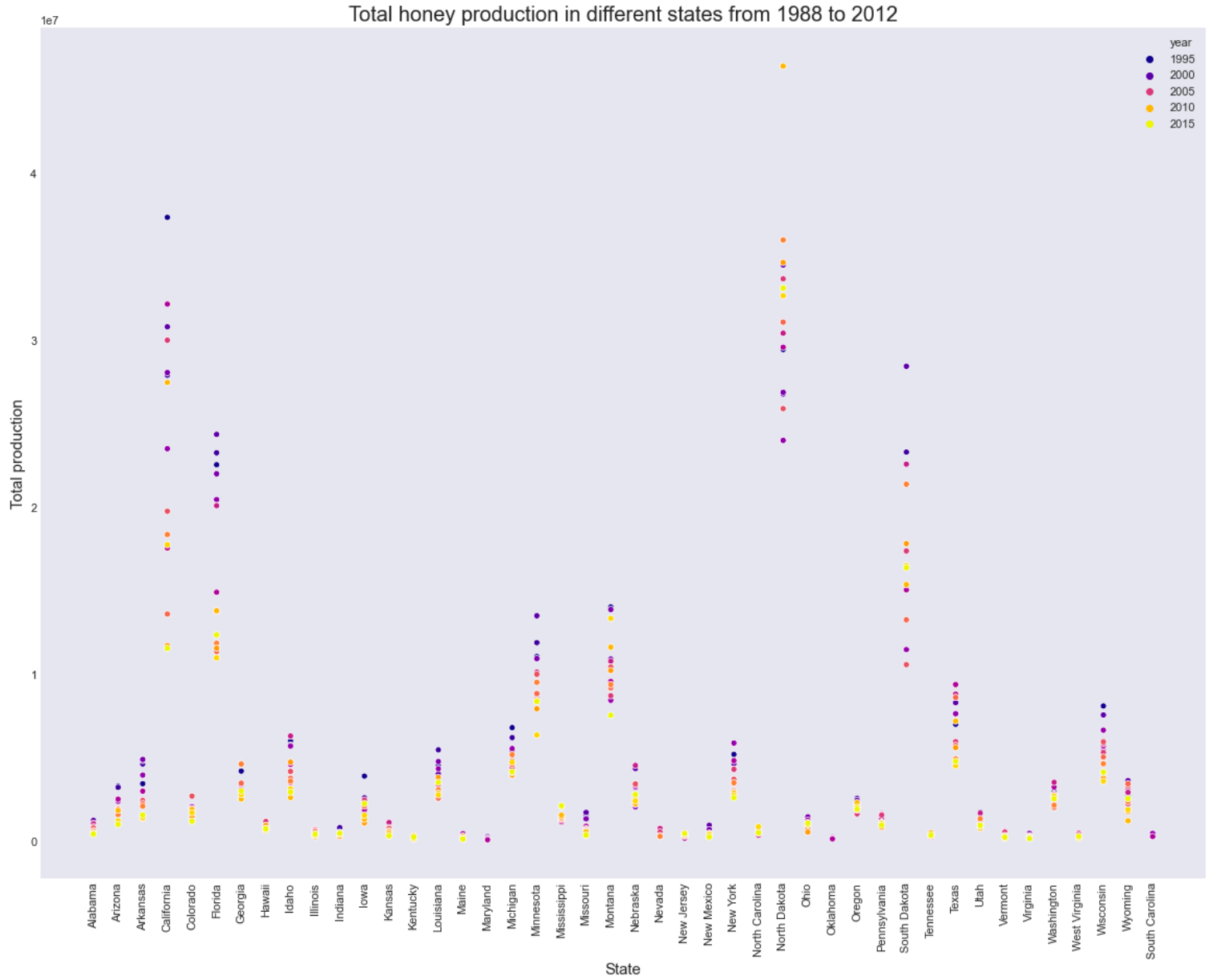
- The beekeepers: give the suggestions on how to increase the honey production and how to apply neotics to promote honeybee colony number increasing
- the customers who consume the honey: provide historical data analysis and production prediction to let everyone know that how the honey production will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

Data Wrangling

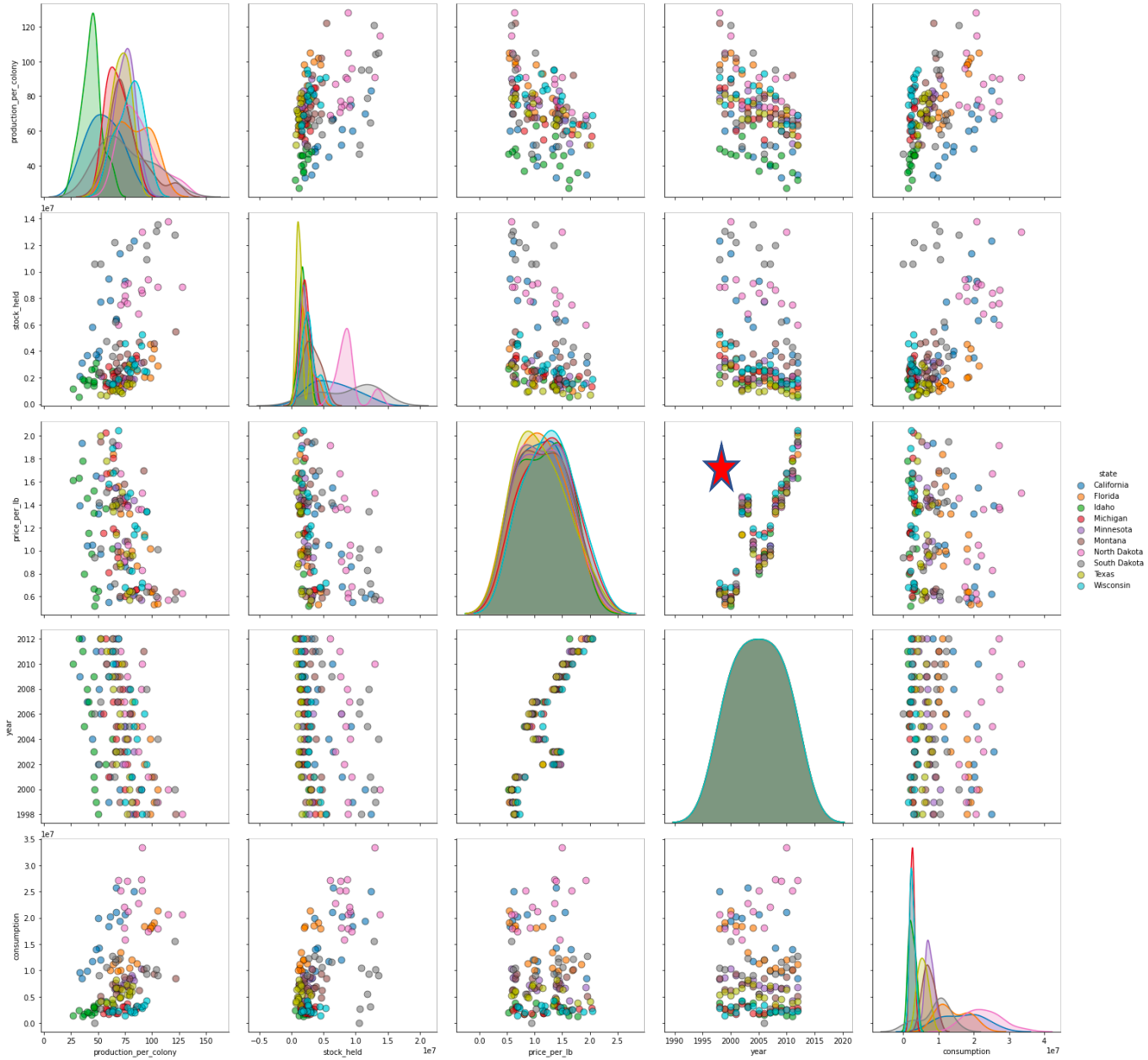
1. Rename the columns to make the data easy to read.
2. Drop useless column ‘FLPS’.
3. Fill the empty space with 0.
4. Replace ‘state_code’ with full state name.
5. Add ‘consumption’ column.
6. After applying the above steps, there are 626 rows left; for neonics data, after wrangling, there are 1132 rows left.

Data Visualization

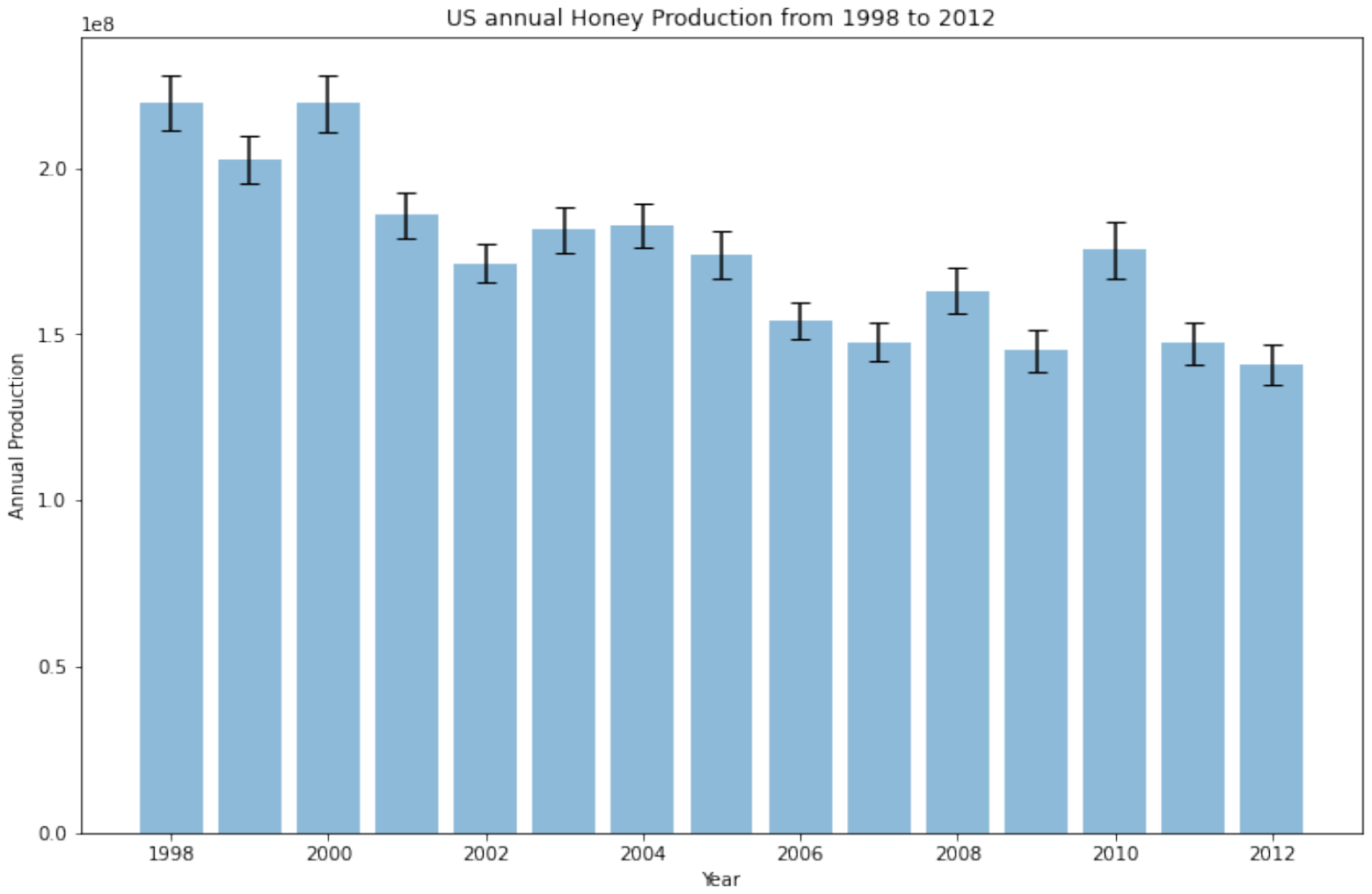
The honey production decrease in most of the states.



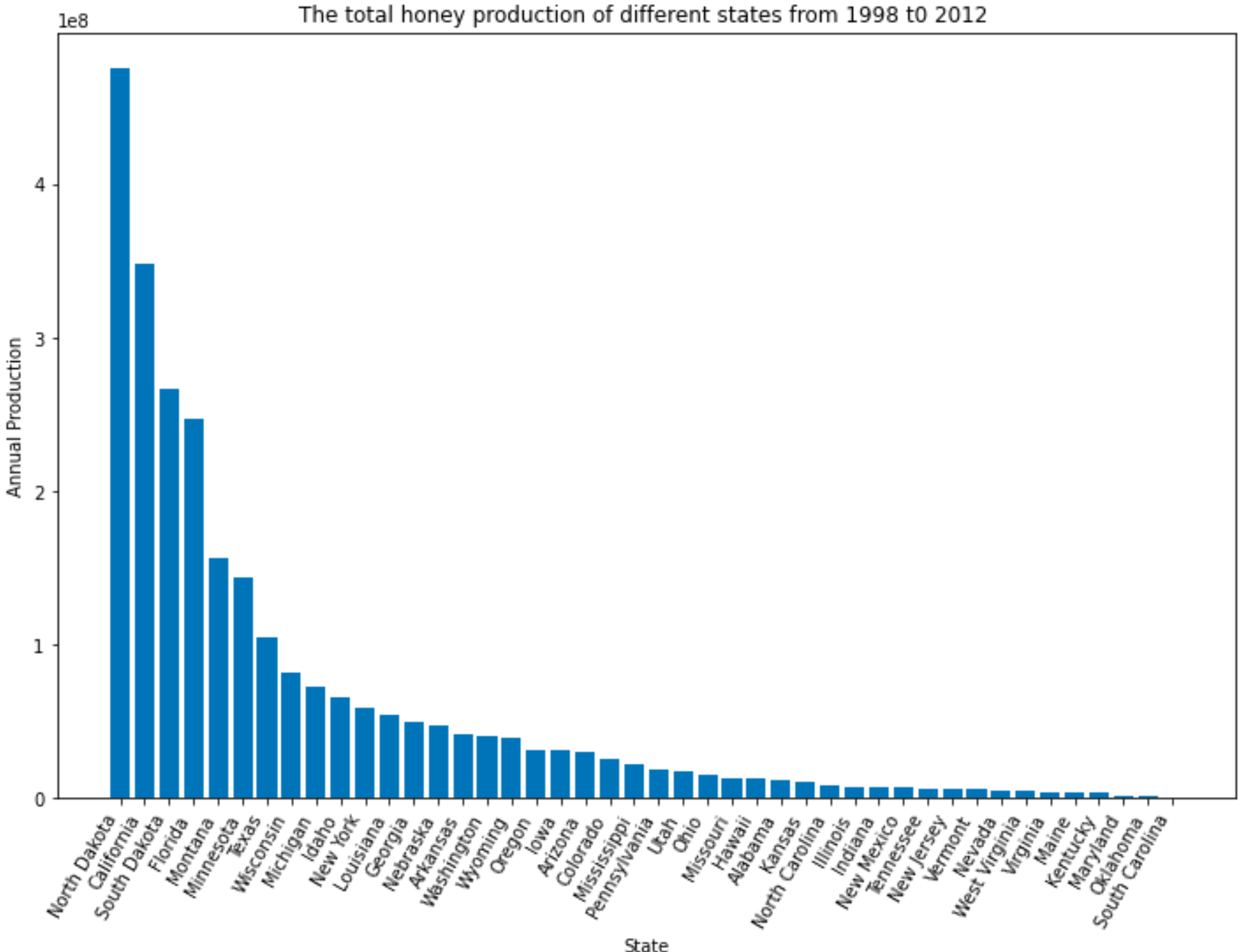
The Pairwise plot between the main features.



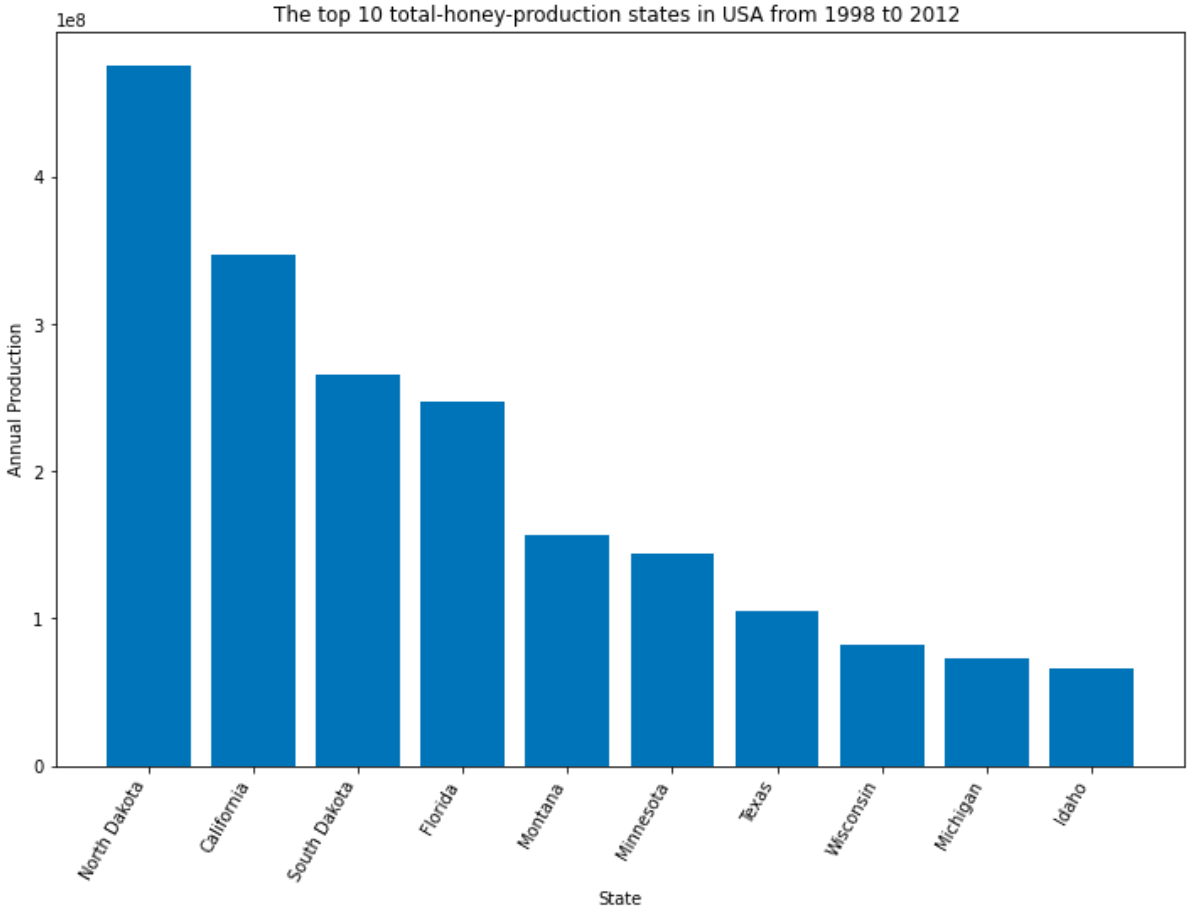
The USA produces the maximum honey at Year 2000

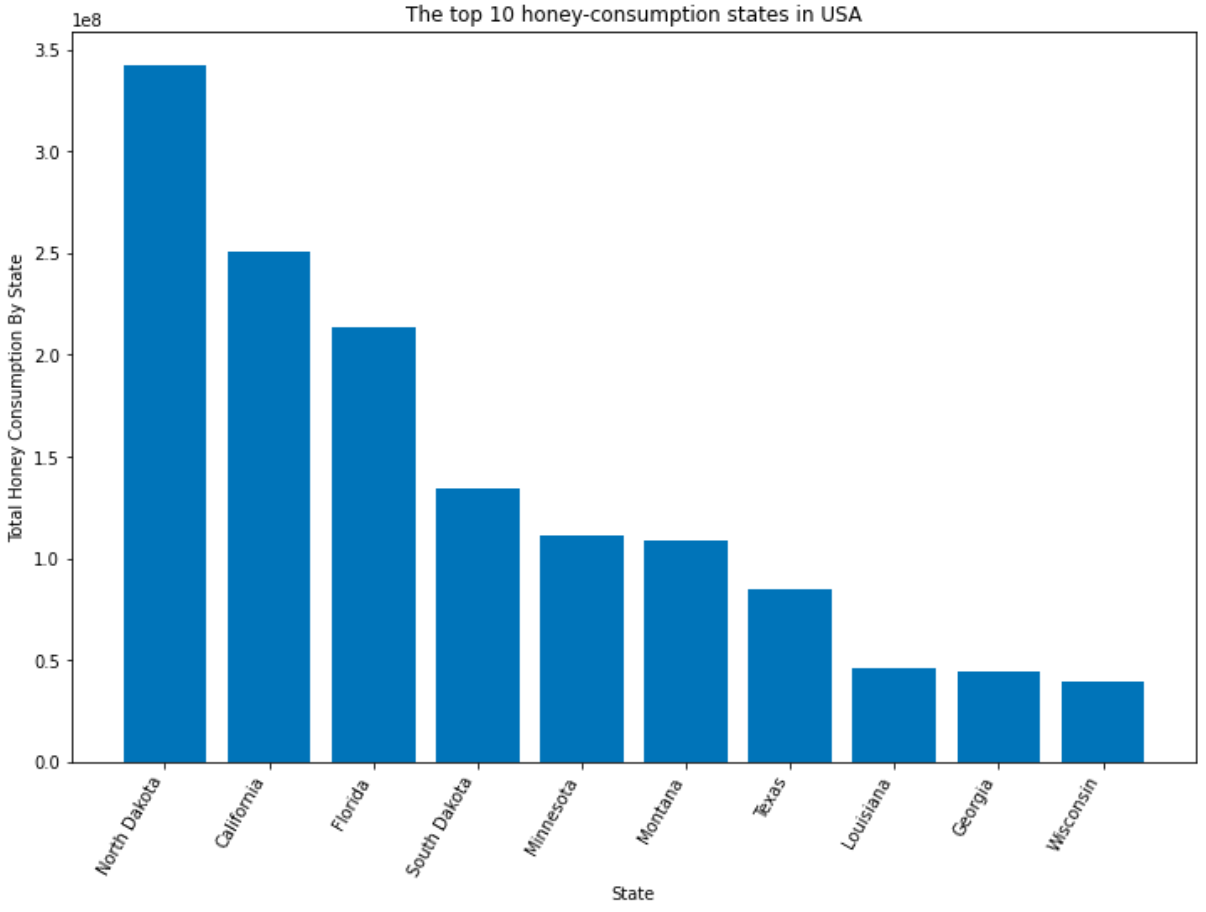


North Dekota has the topest level of honey production and South Carolina has the lowest level of honey production.



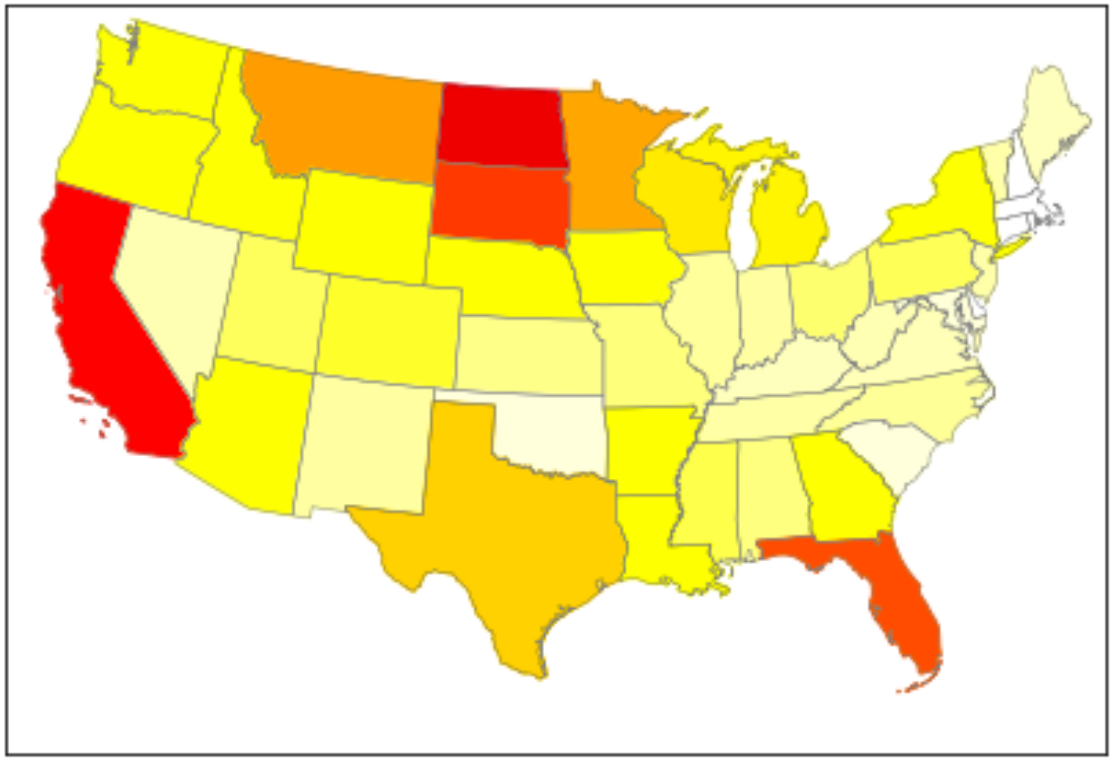
Top 10 states with honey production/ honey consumption



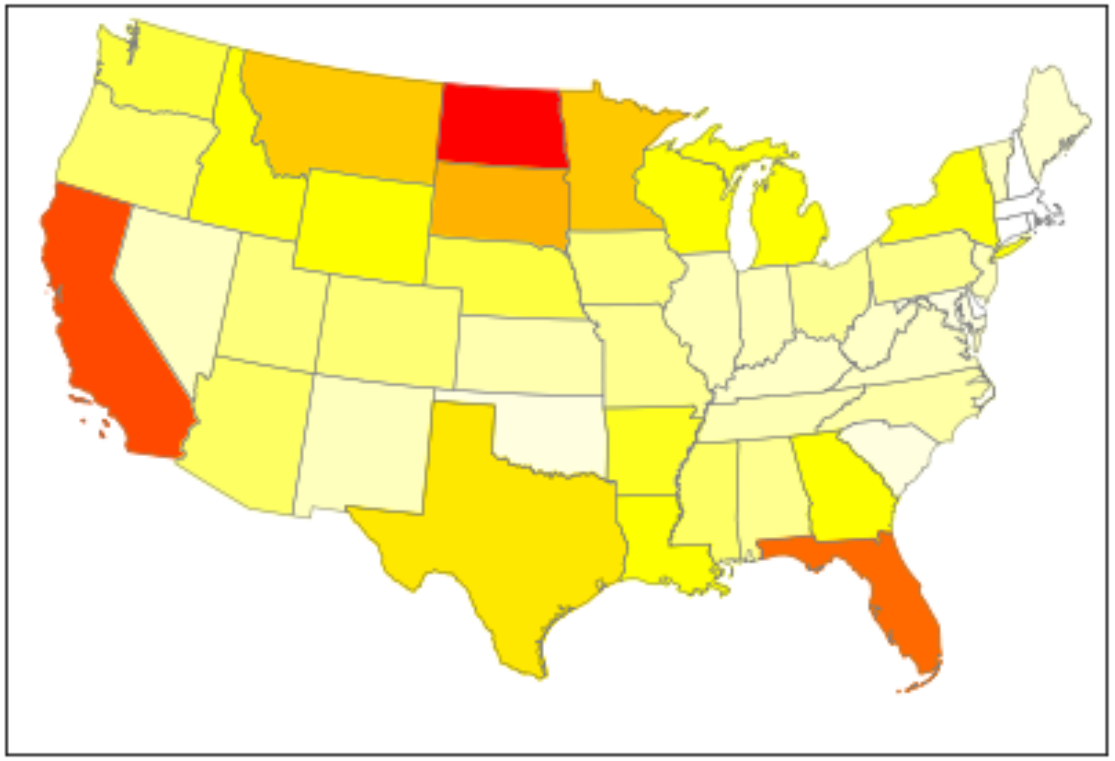


Top honey producing and consuming state map in USA

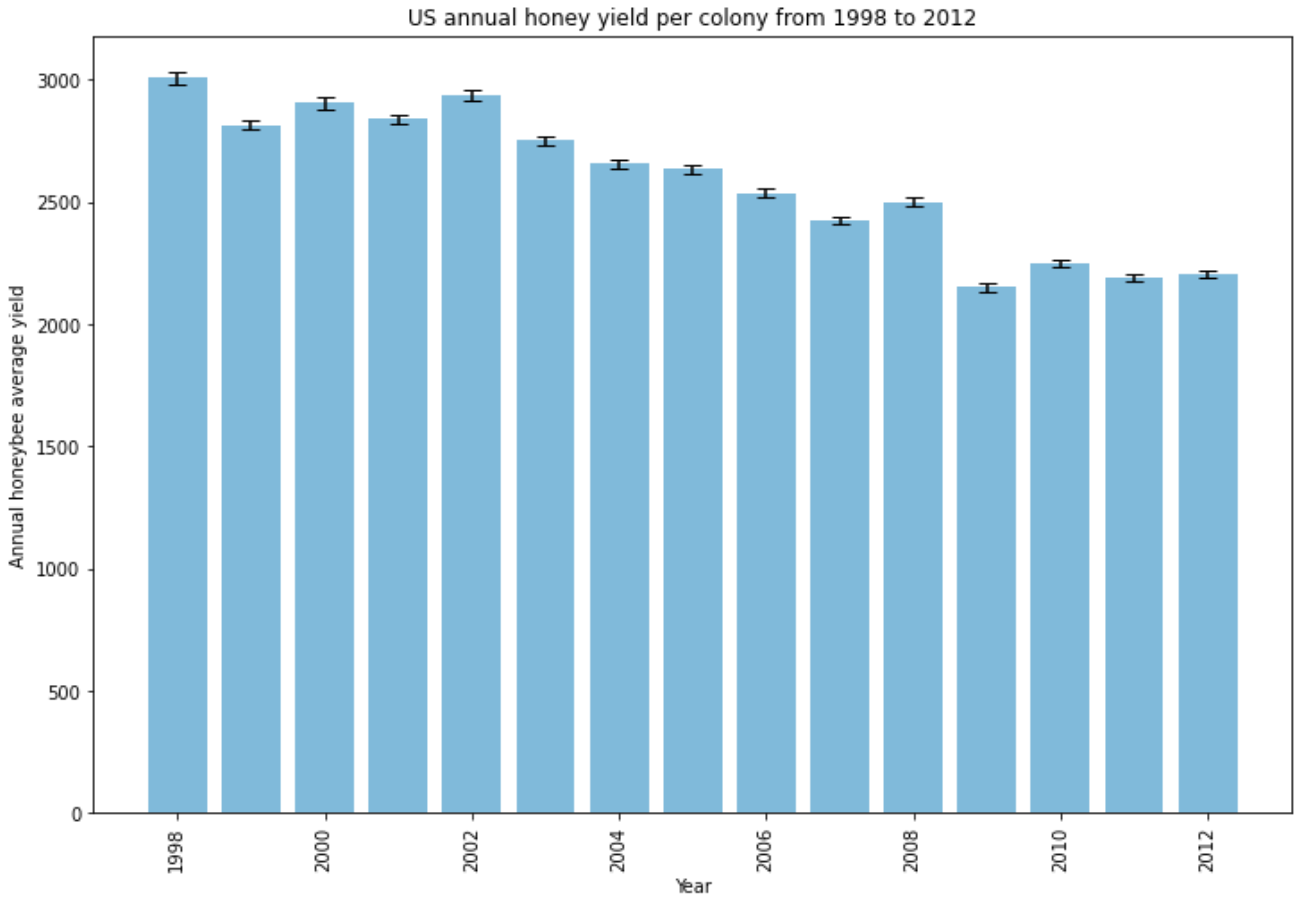
Top honey producing states in the USA



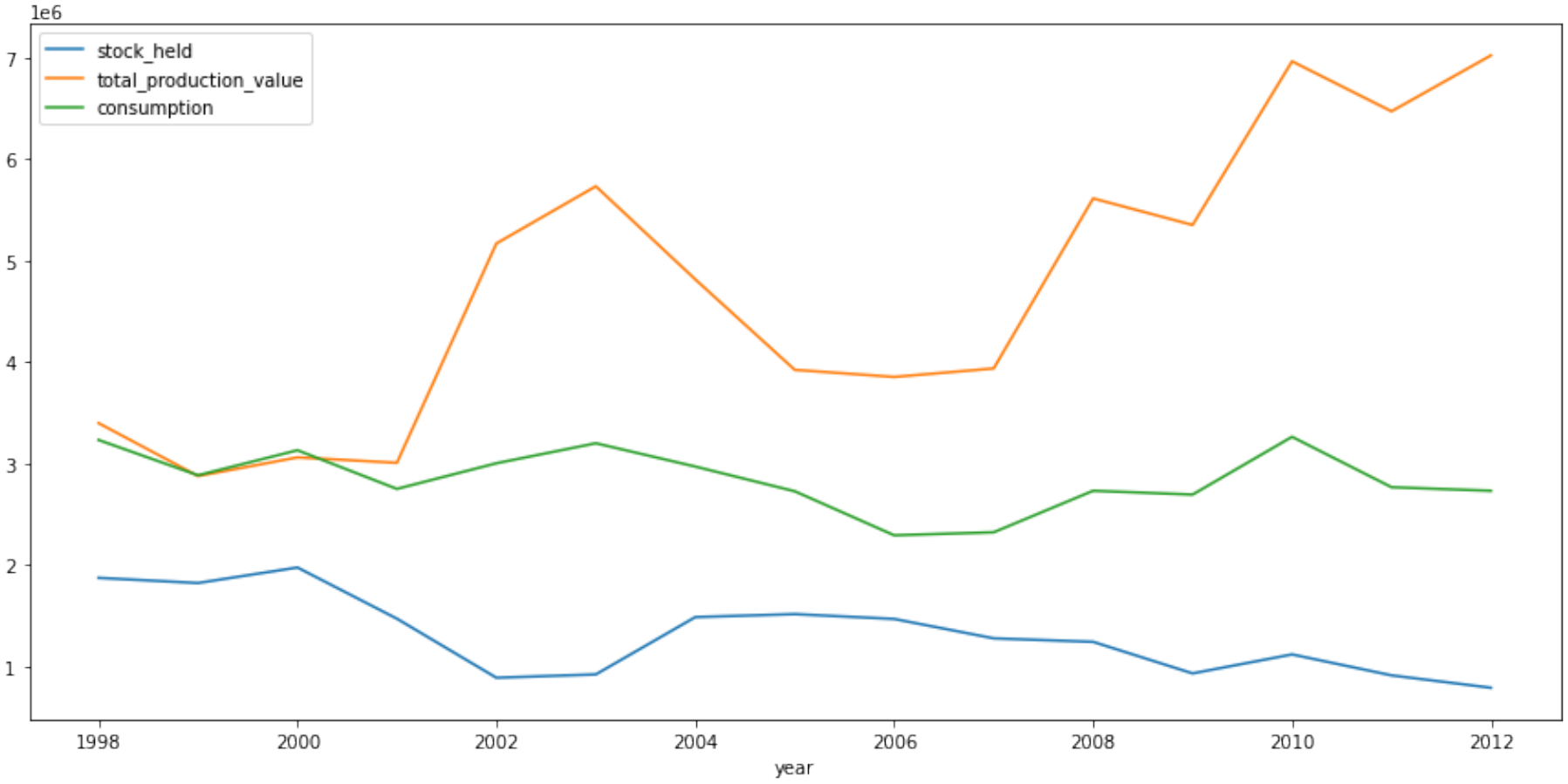
Top honey consuming states in USA



The annual honey yield per colony exhibits decreasing trend.

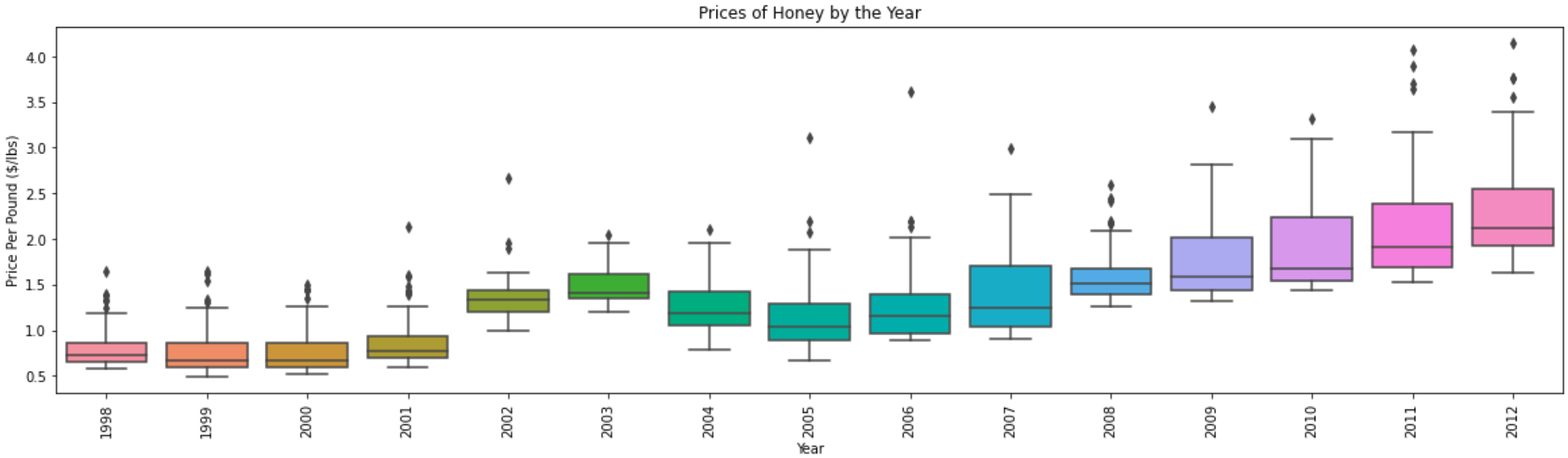


The honey stock, total production value and consumption changing trend

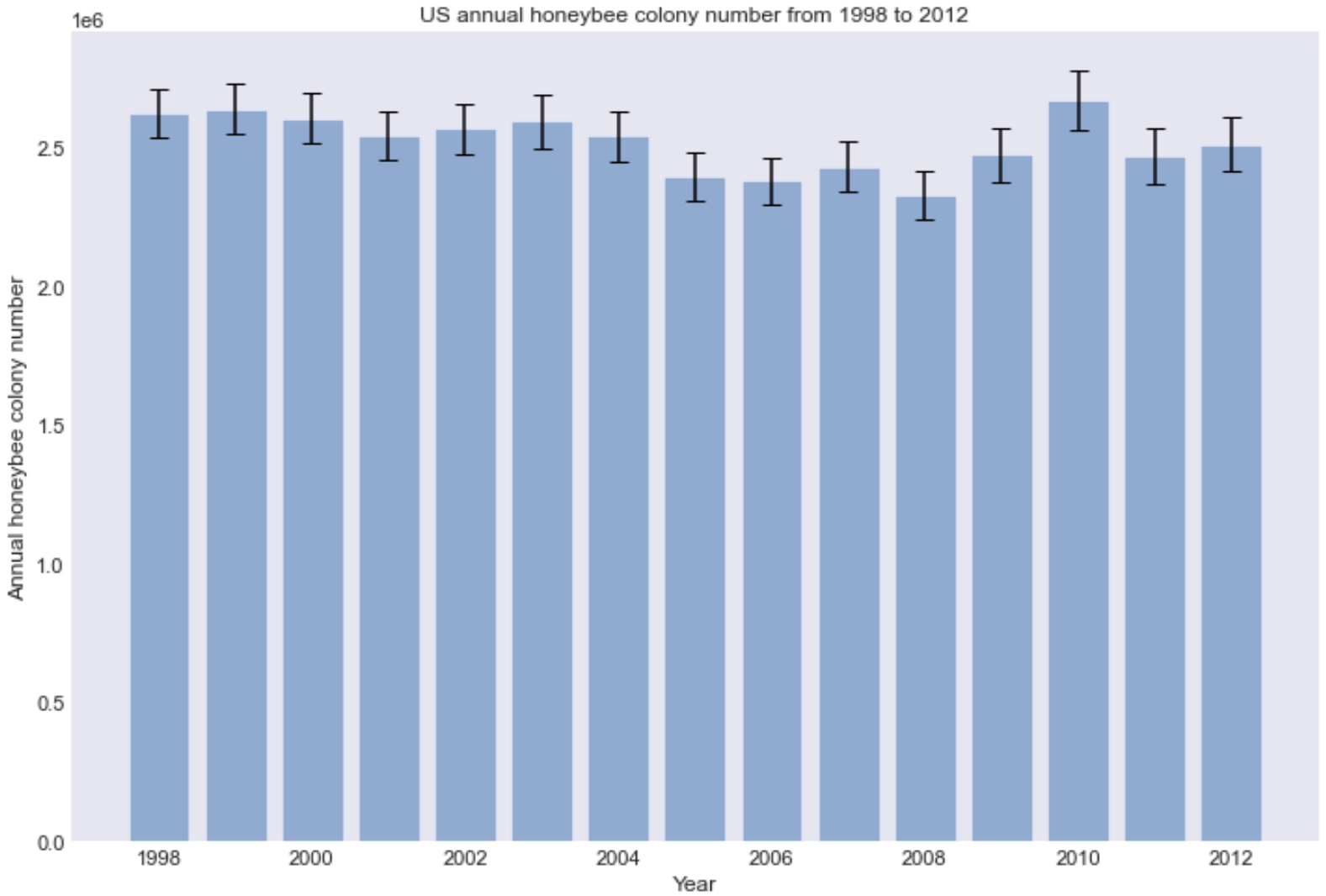


Why total production values goes up?

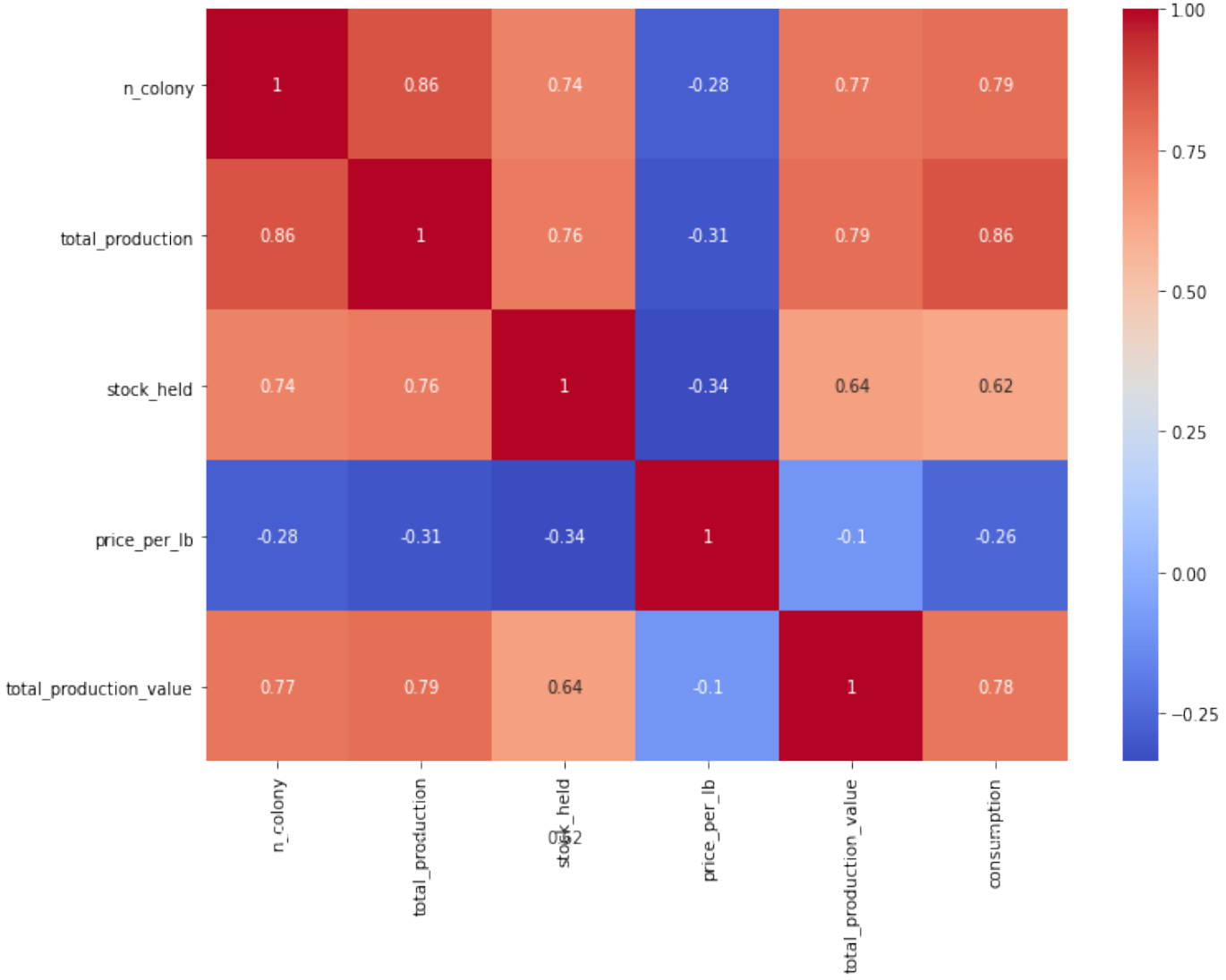
The honey price exhibits the almost increasing trend



US annual honeybee colony number are decreasing during 1998 and 2012.



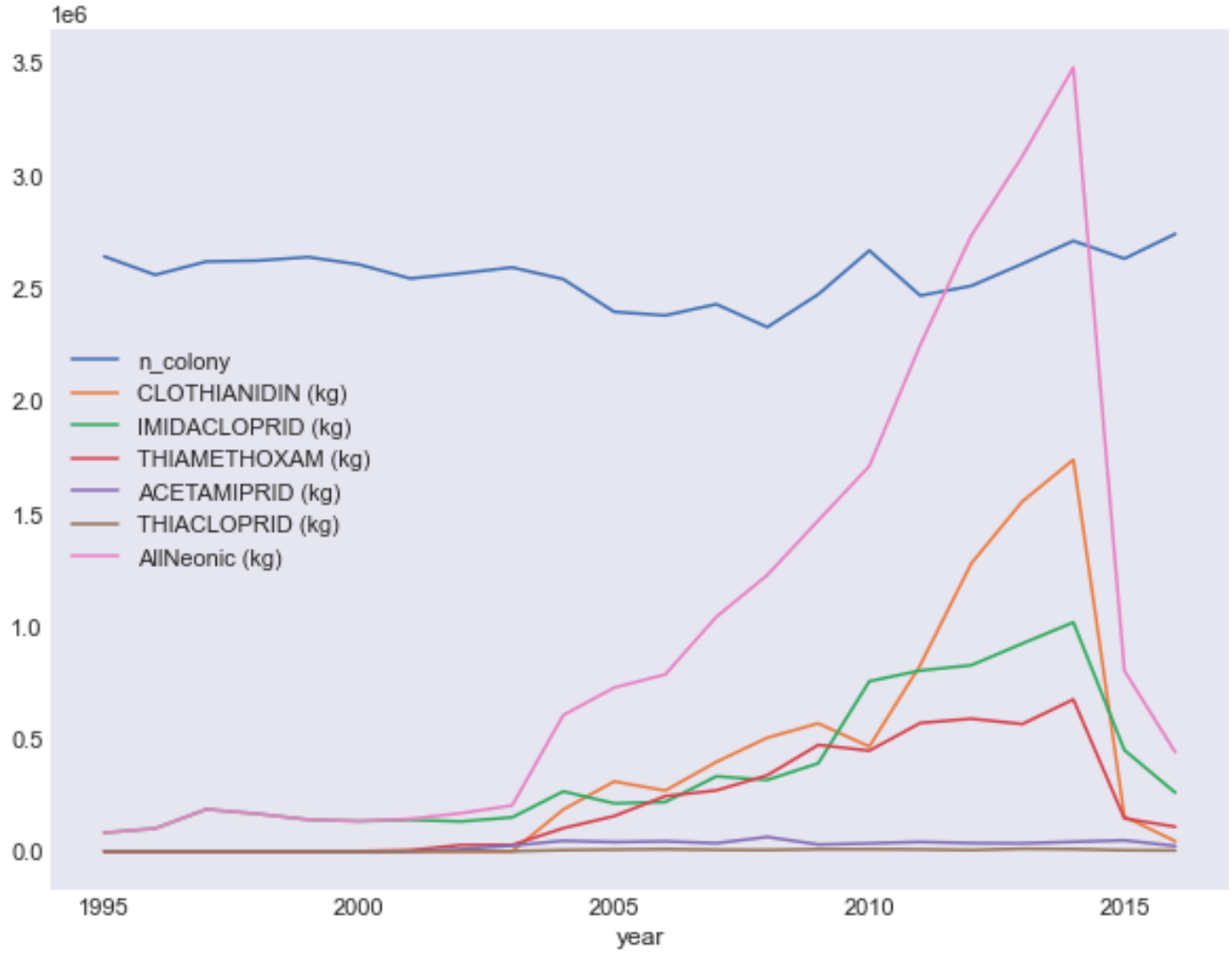
The kendall correlation between honey price and production



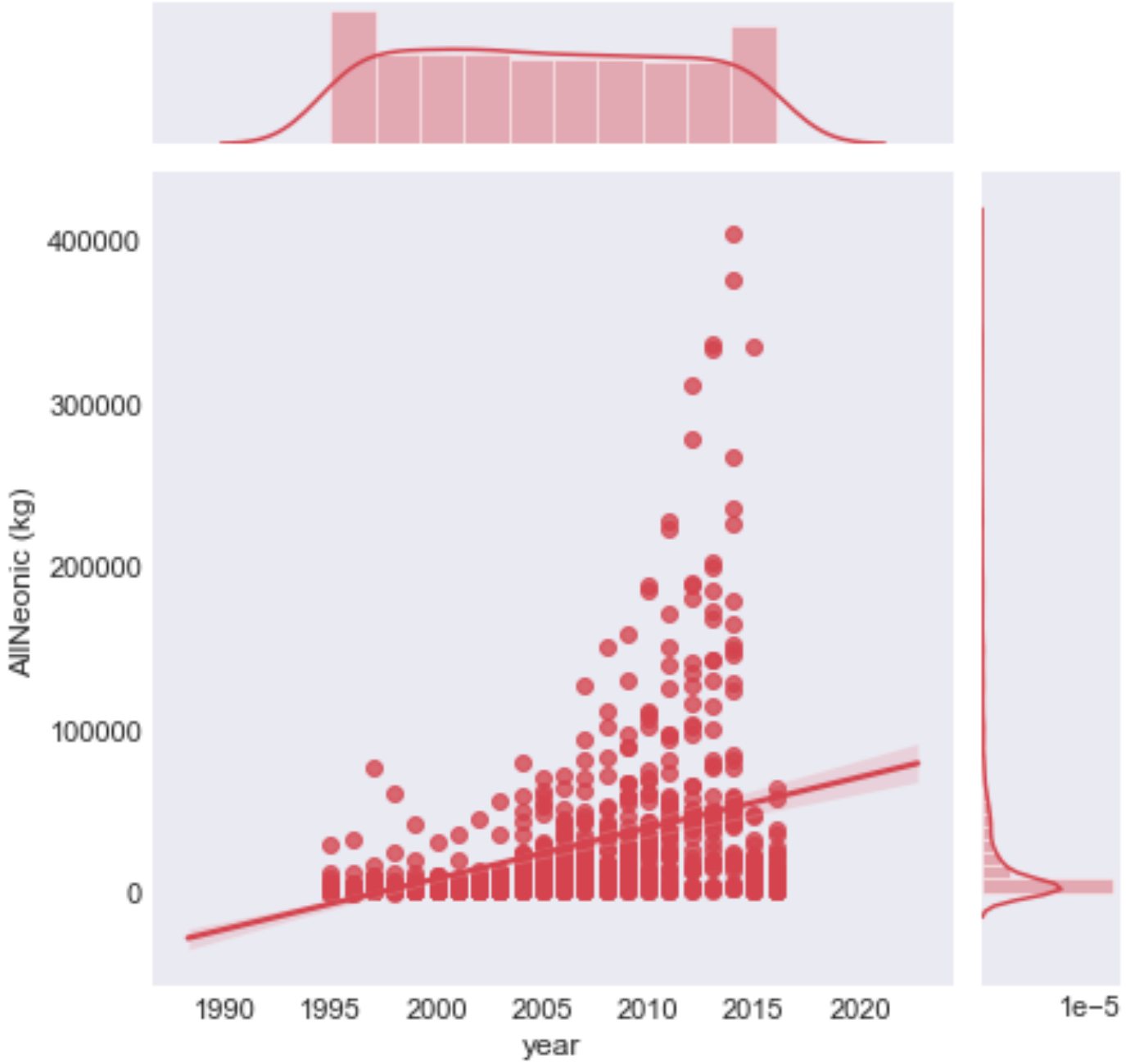
Summary I

- 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 numbers.

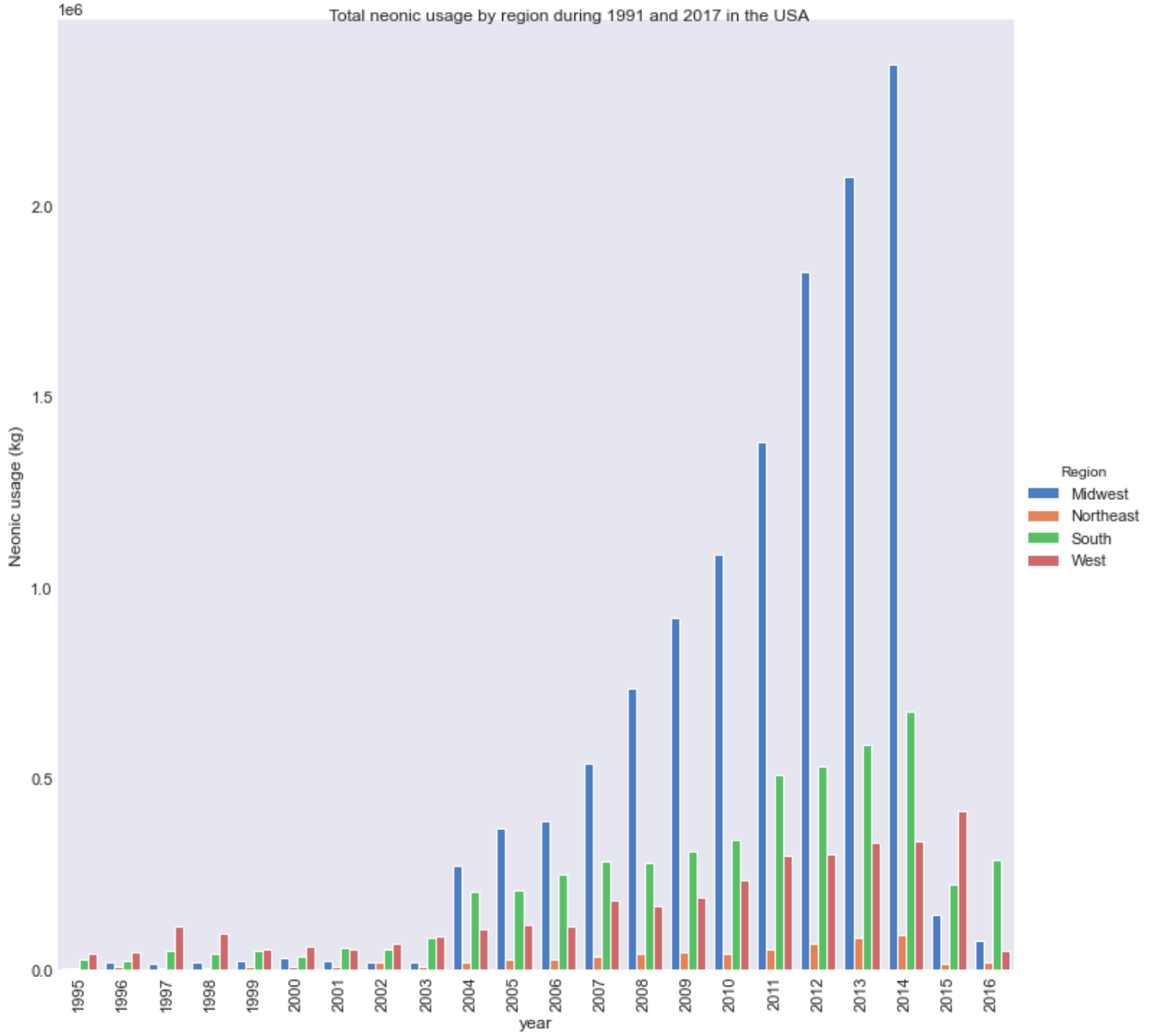
The five kinds of neonics usage is increasing sharply.



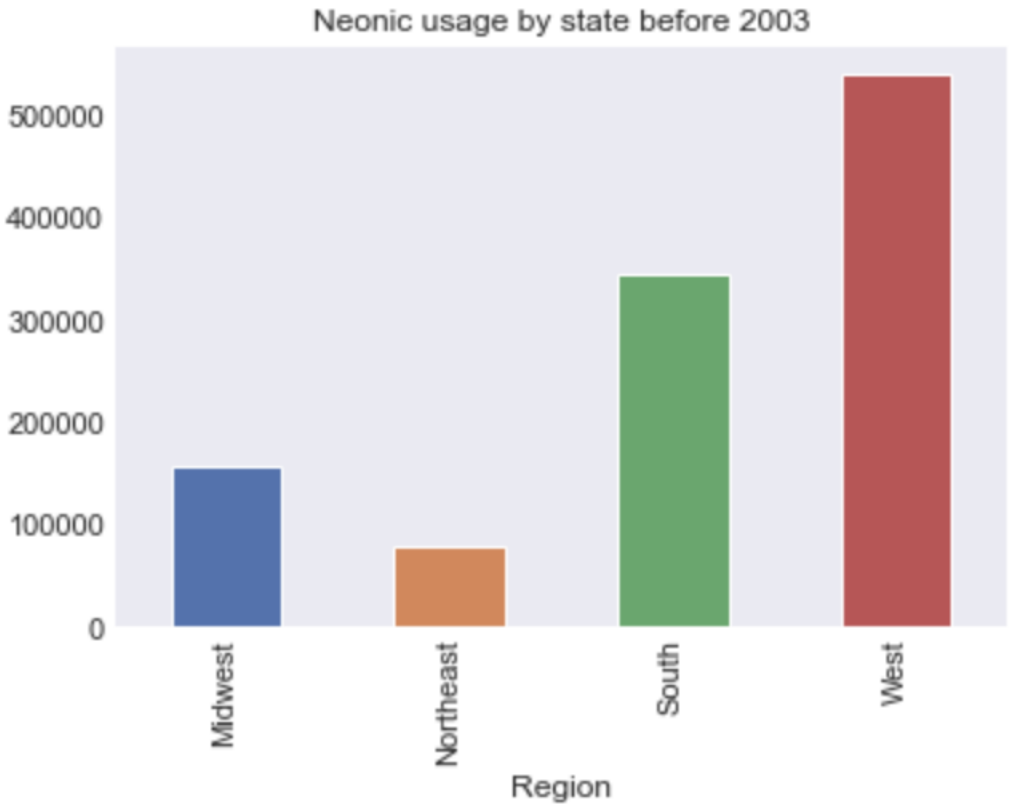
The annual neonics usage in all the states from the USA

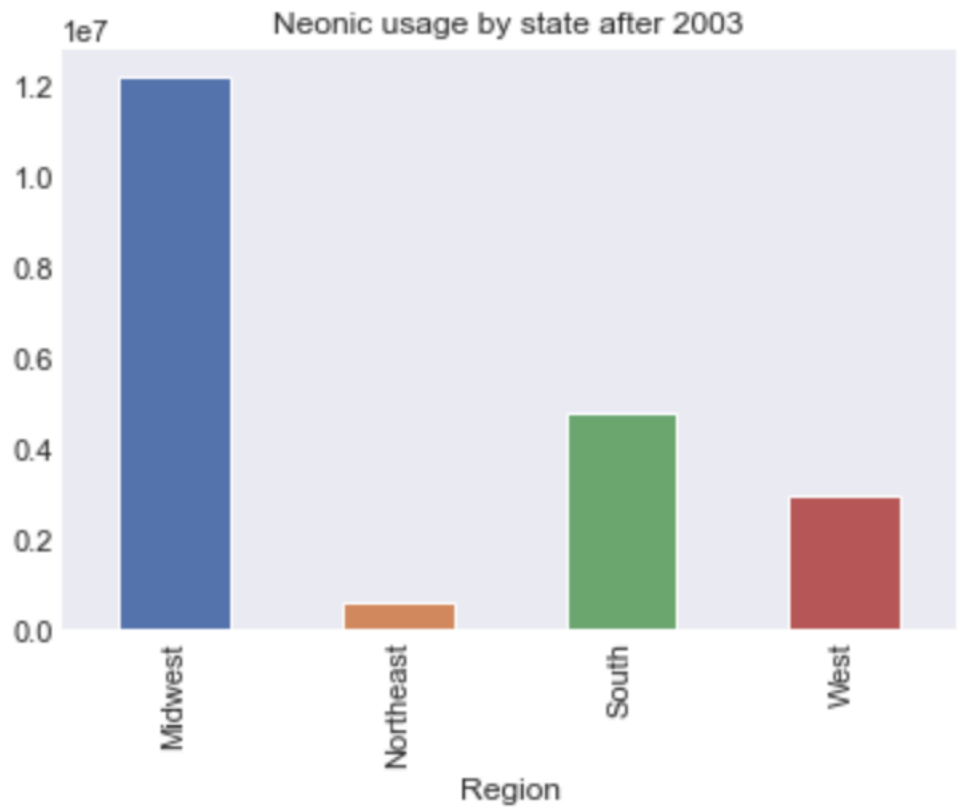


Total neonics usage by region during 1991 and 2017 in the USA

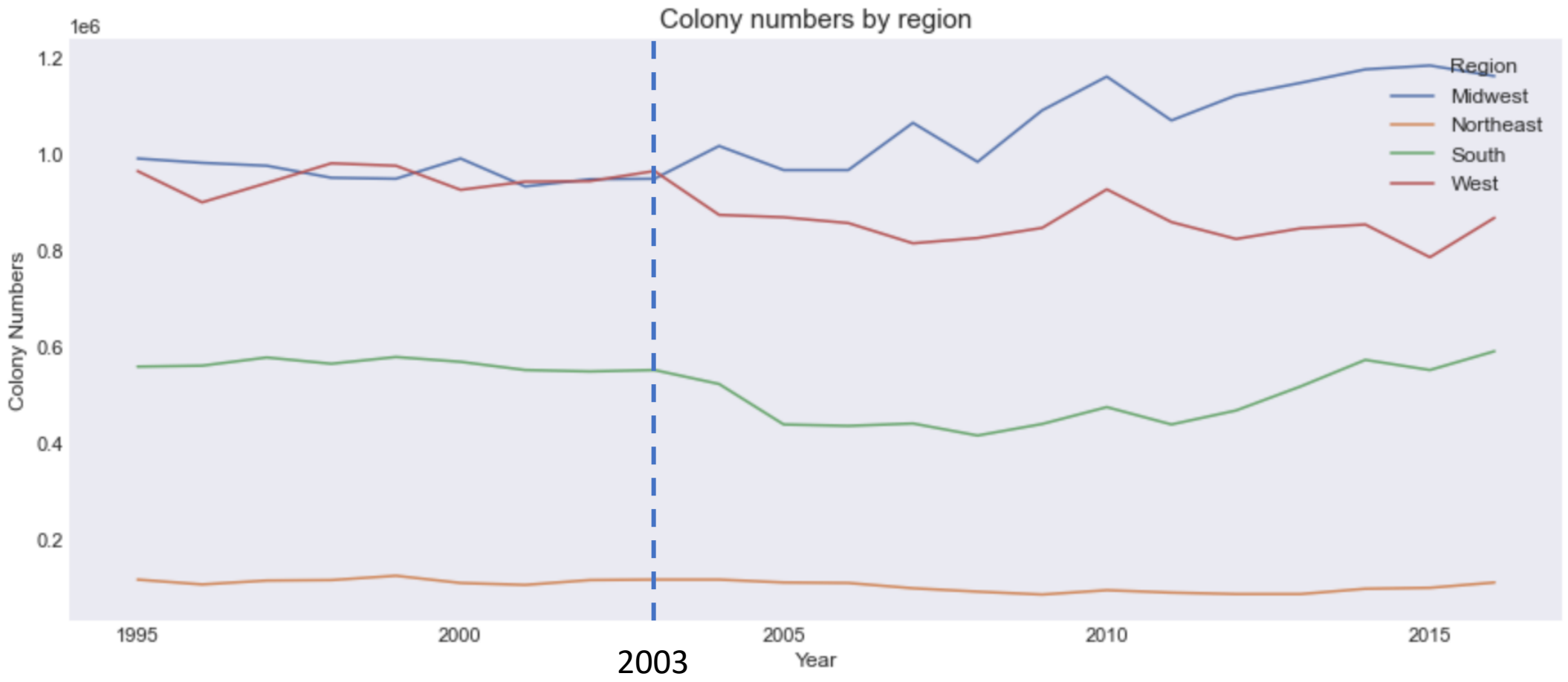


Neonic usage before 2003 and after 2003 by region





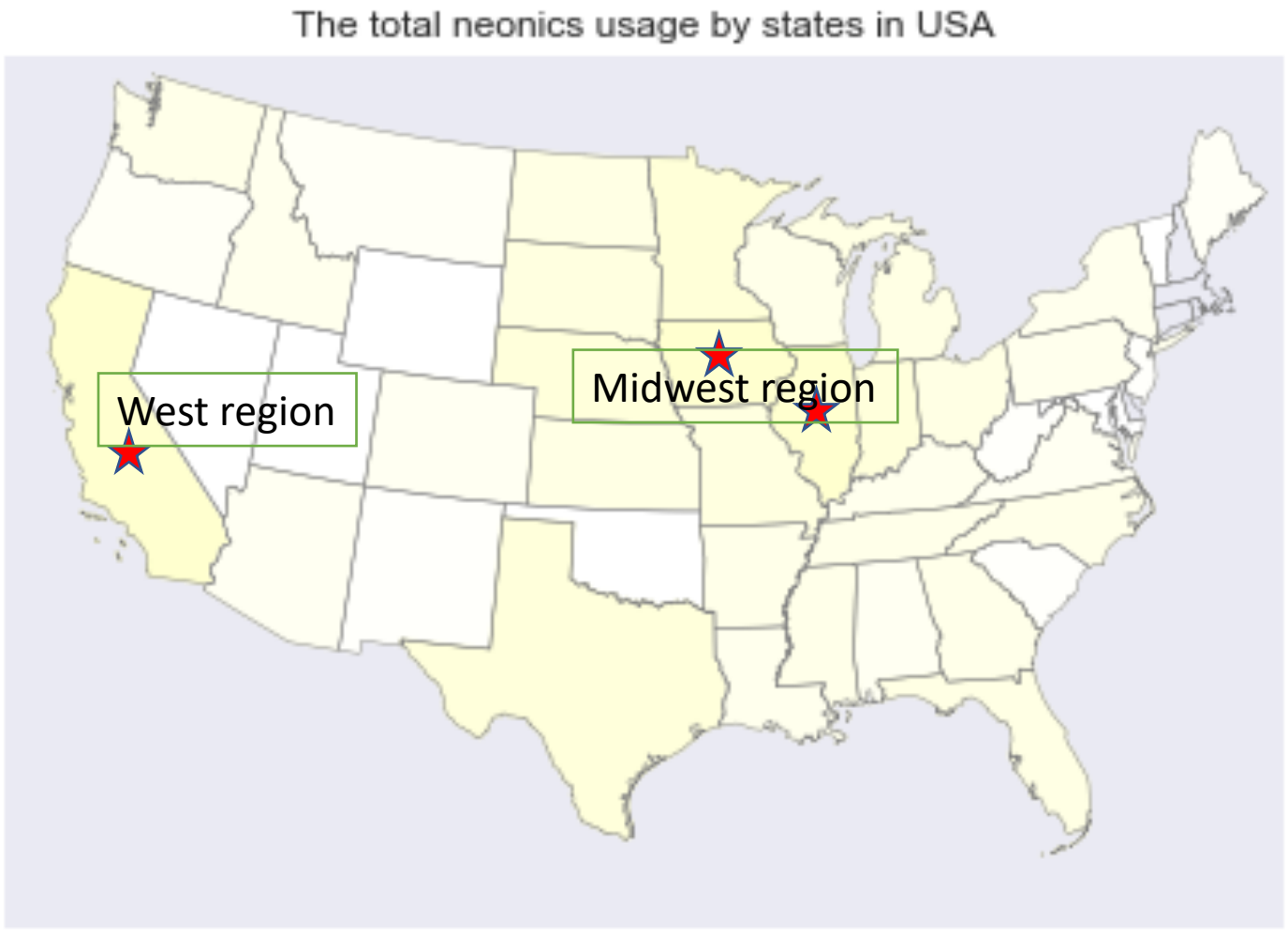
Total colony number changing before 2003 and after 2003 by region



The neonics usage map in the USA

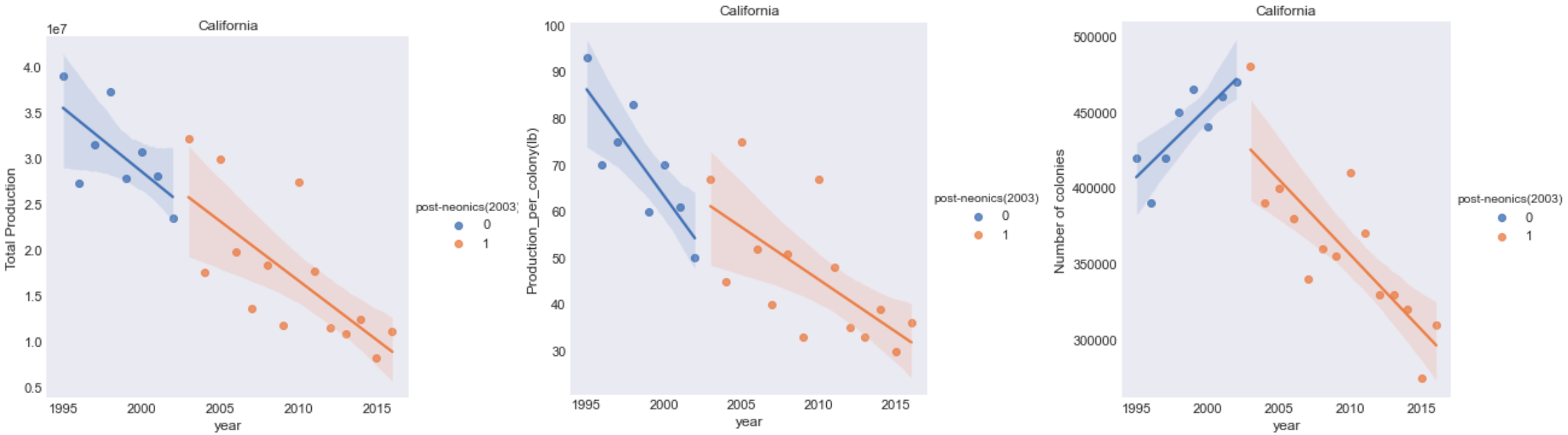
Top 3 neonic-usage states:

state	AllNeonic (kg)
California	1991179.1
Illinois	1978523.1
Iowa	1974038.9



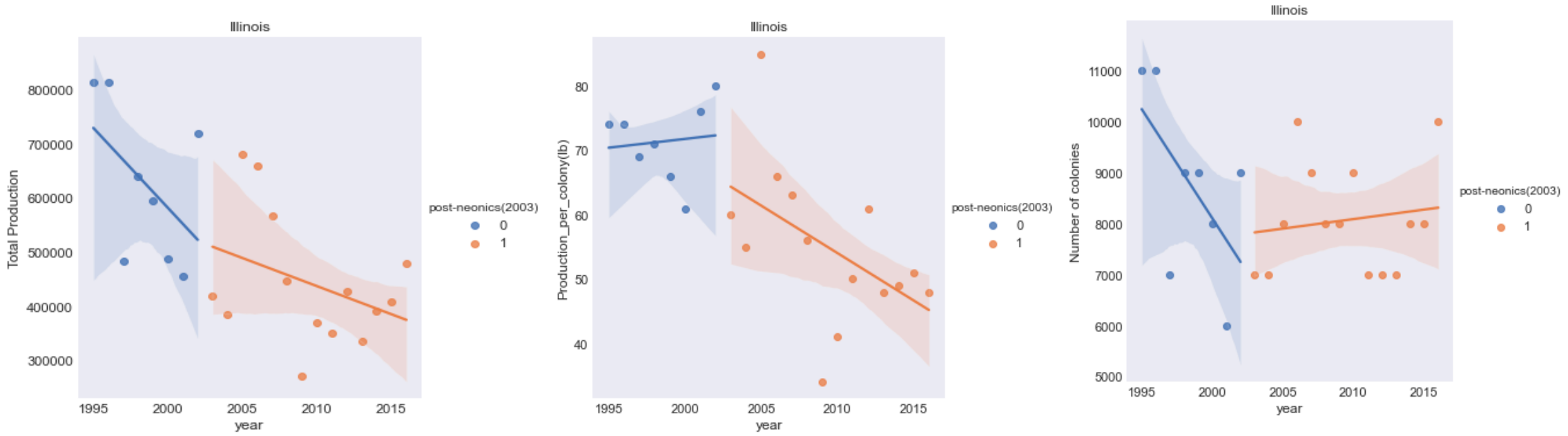
Note: The ‘star’ marks the top 3 states on neonics usage.

The neonics-usage effect in California



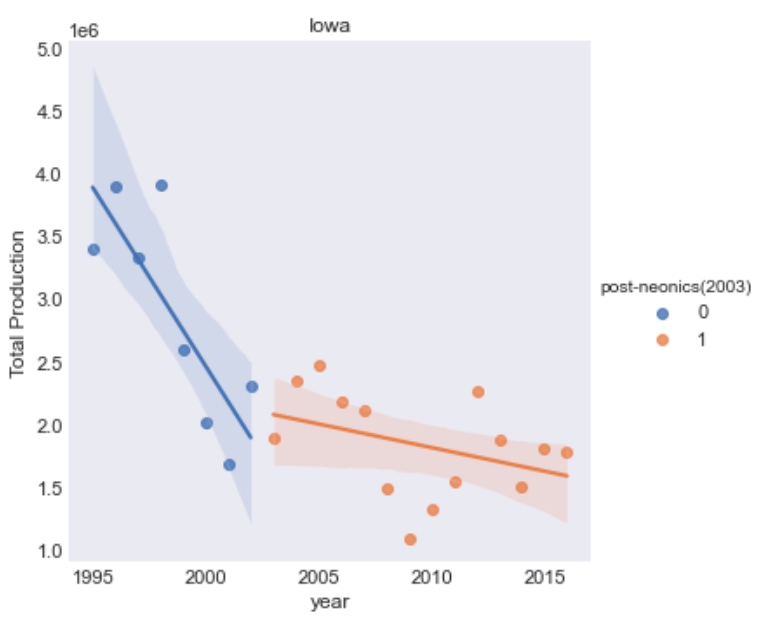
California's no. of colonies, yield per colony and total production have been decreasing consistently since their frequent heavy use of neonics in 1994.

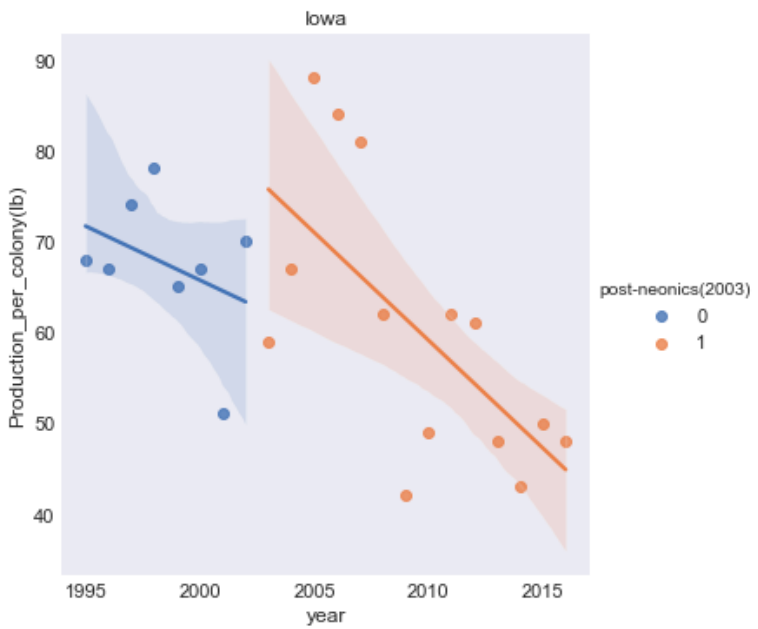
The neonics-usage effect in Illinois

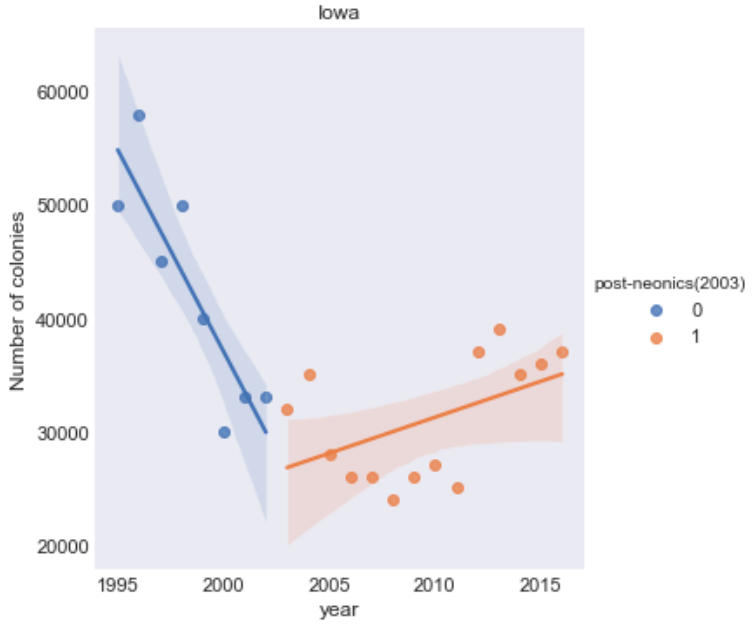


By applying neonics, Illinois's production per colony id sligtly decreasing after 2003 even though the colony number is increasing slowly.

The neonics-usage effect in Iowa





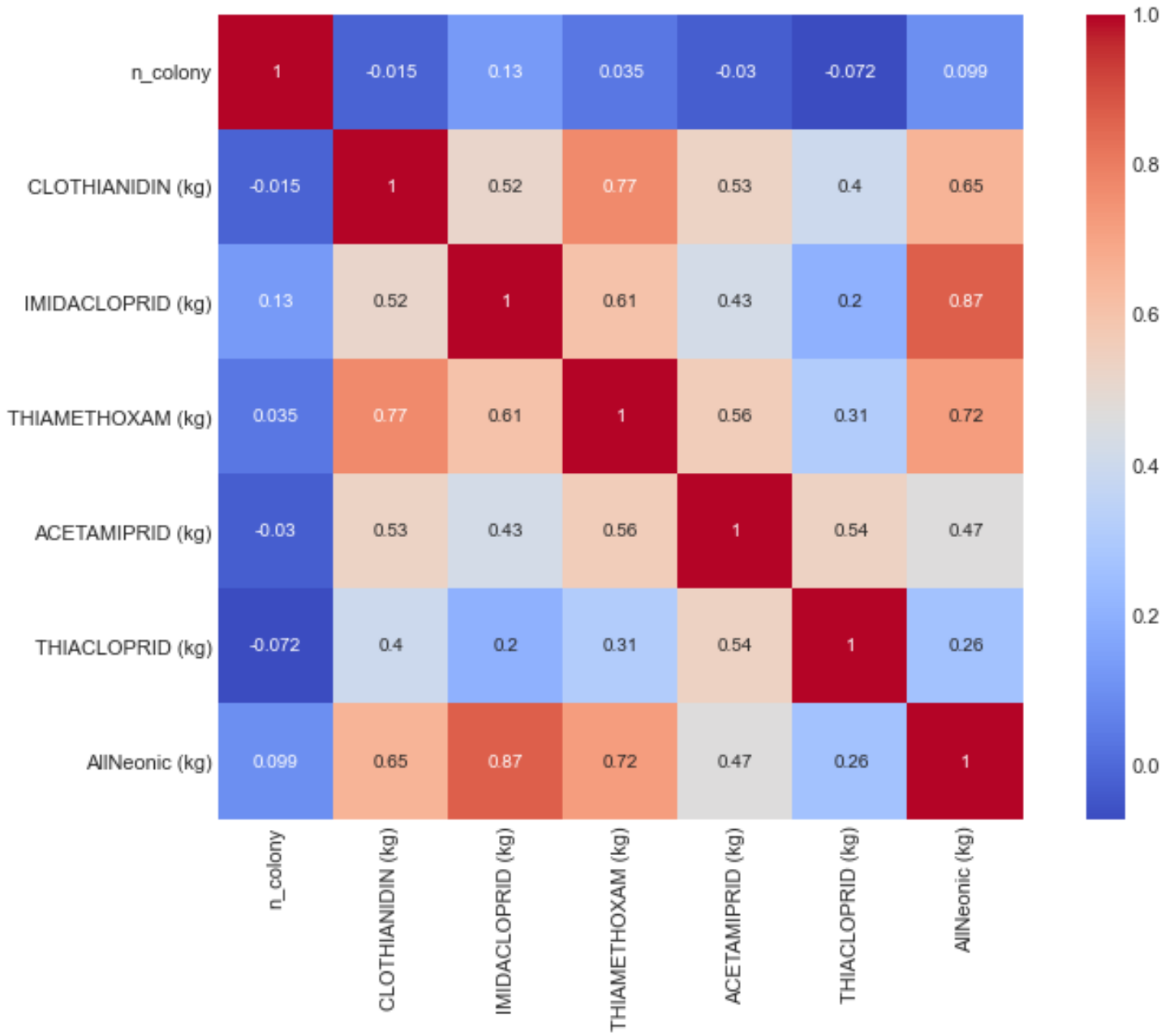


Similar to Illinois, Iowa's total production and number of colonies were decreasing before 2003., with yield per colony increasing.

Summary II

- 1. Only 'Midwest' area exhibits increasing trend of honeybee colony even though large amount of neonic usage after 2003.
- 2. Number of honeybee colony in 'Northeast' area is reiesatively stable and also this area only uses small amount of neonic pesticide.
- 3. 'South' region exhibits a slight decrease in number of honeybee colony after 2003 which coincides with a decrease in neonic usage, although colony numbers appear to be on the rise or maintaining a consistent number in the future.
- 4. 'West' region shows a decrease in the number of colonies and Californian should be responsible for the colony number decreasing due to its massive neonic usage before 2003.

The kendall correlation between neonics application and honeybee colony numbers

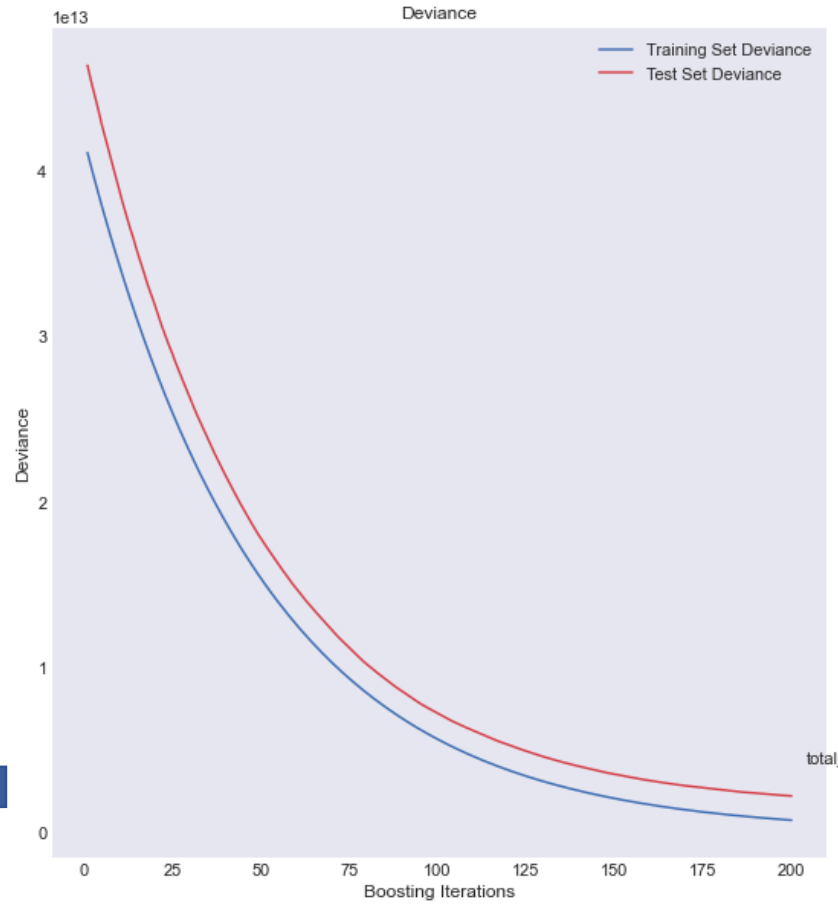


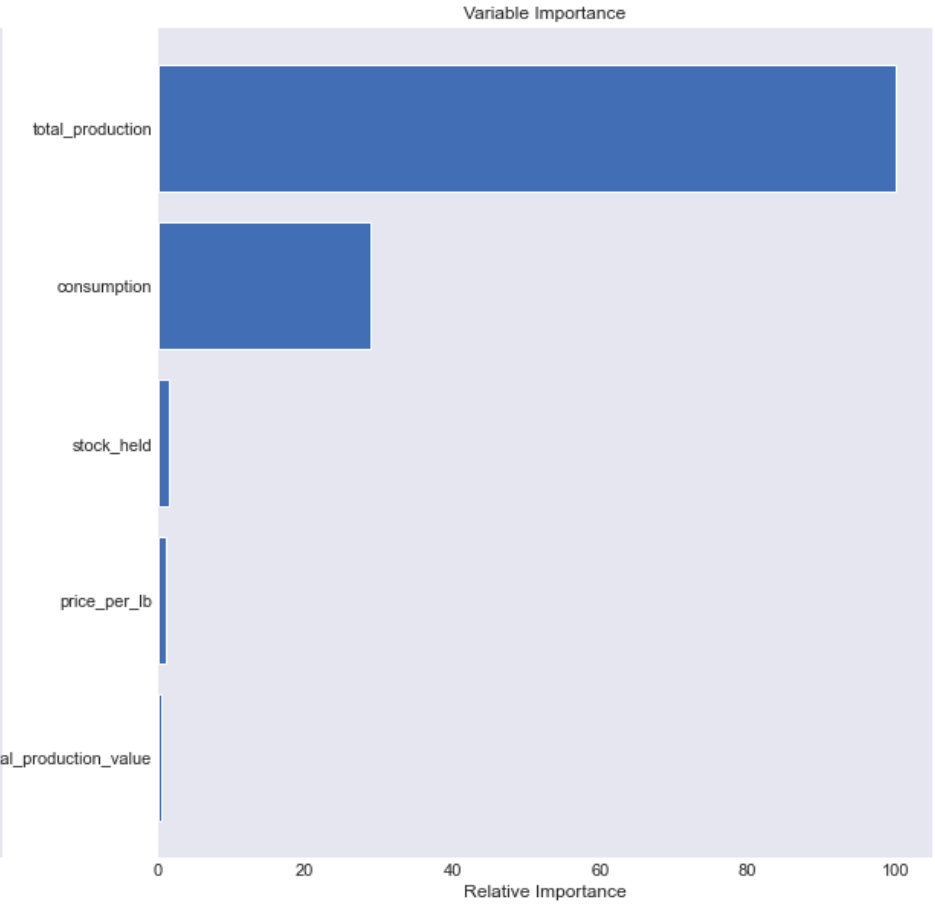
Summary III

- 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 0.18, which indicates the application of neonic pesticide could promote the honeybee developing.
- Among the neonics, IMIDACLOPRID (corr=0.22) plays a key role in promoting honeybee developing; it also show strongest correlation with allNeonic at 0.8. Thus, Imidacloprid is the most import neonics in promoting honey propagation.
- The second important one is THIAMETHOXAM (corr=0.073). The rest of neonics all affect the honeybee colony negatively.

Machine learning analysis

Build the gradient boosting regressor



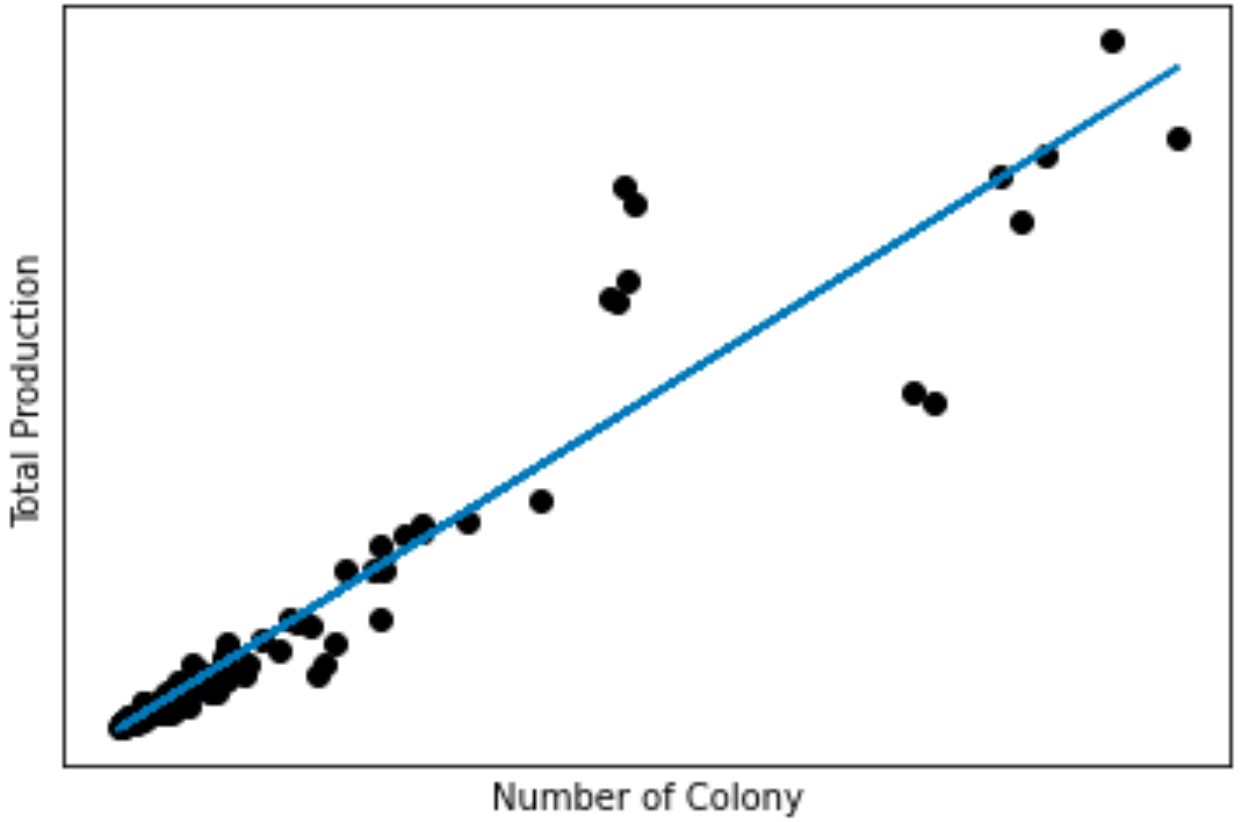


'Total production' is the most important variable.

Supervised learning

1. Linear Regression Model

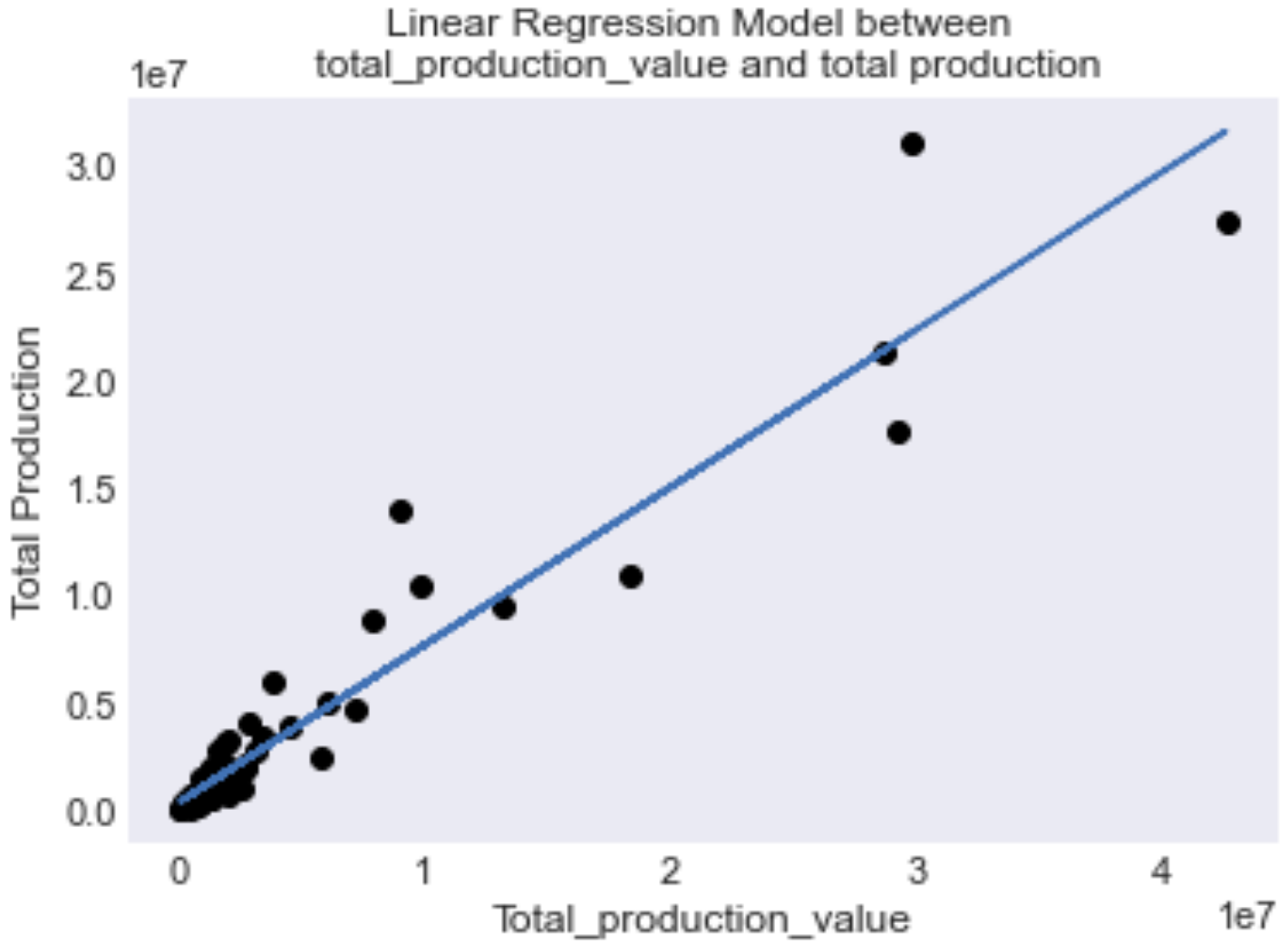
- R^2 score between colony number and total production is 0.9092 based on linear regression model.
- Linear regression model equation is Y=75.176269*x+(-217527.838246)



Supervised learning

1. Linear Regression Model

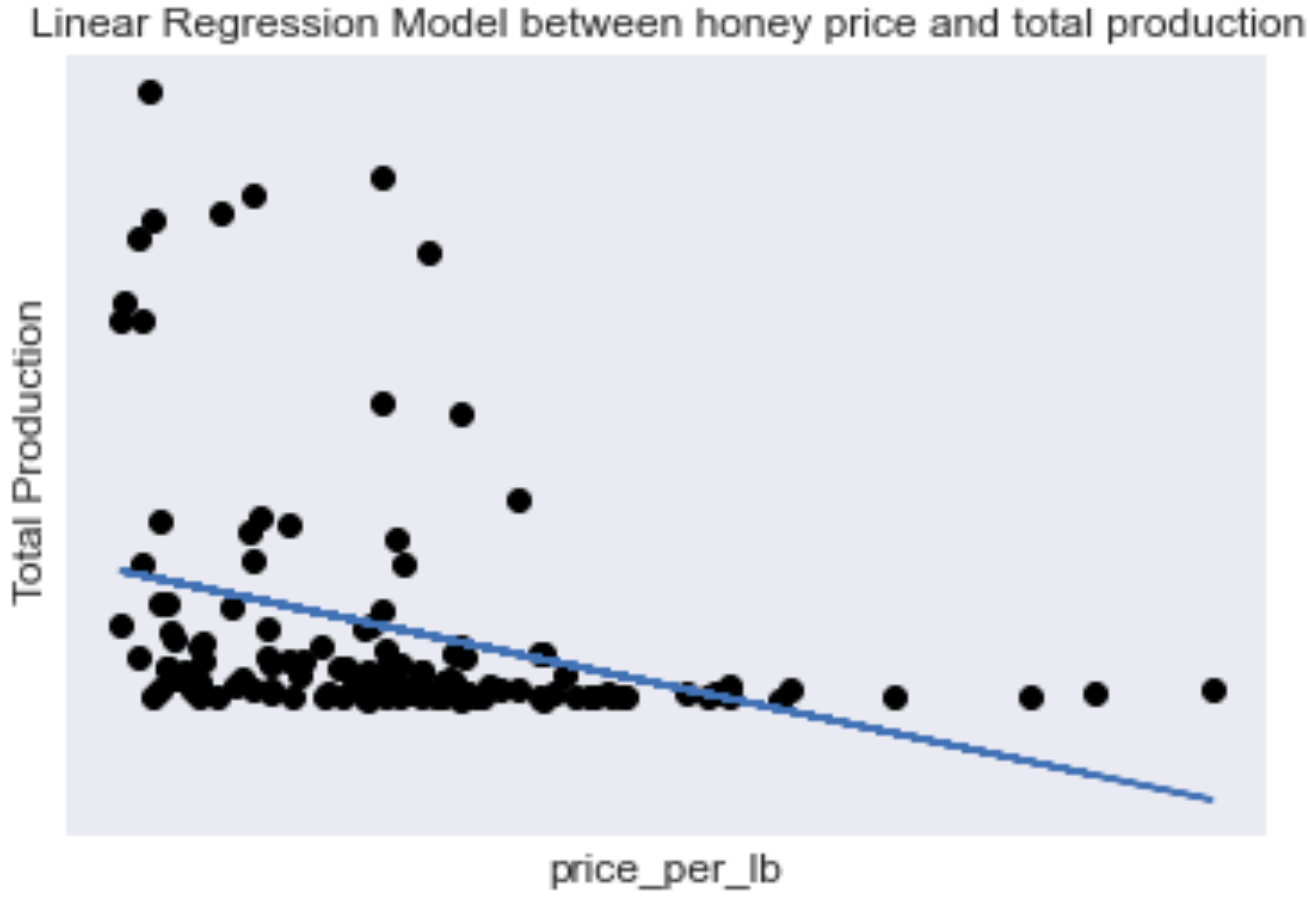
- R^2 score between production value and total production is 0.7764 based on linear regression model.
- Linear regression model equation is Y=0.735375*x+369372.252656



Supervised learning

1. Linear Regression Model

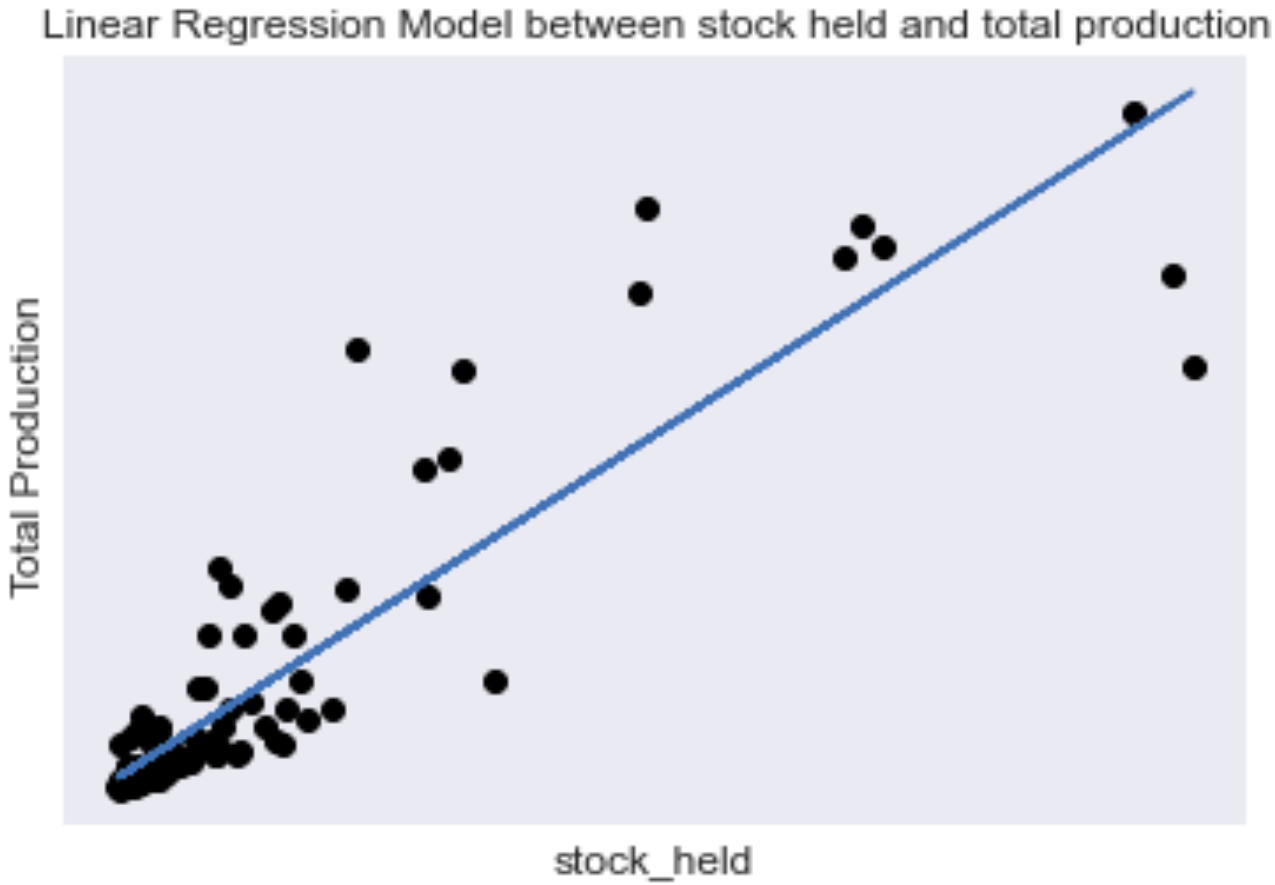
- R^2 score between price and total production is 0.09164 based on linear regression model.
- Linear regression model equation is Y=-3864985.077840*x+(10028476.493532)



Supervised learning

1. Linear Regression Model

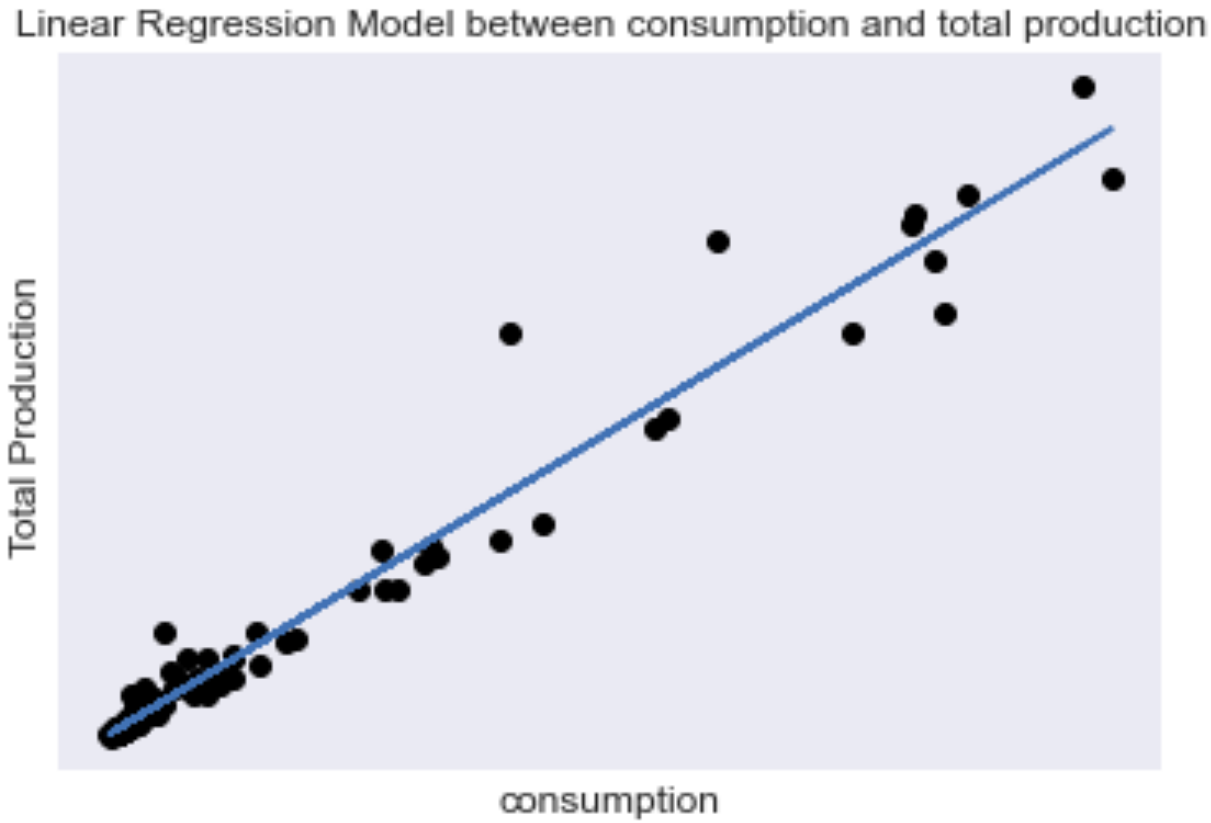
- R^2 score between stock held and total production is 0.8069 based on linear regression model.
- Linear regression model equation is Y=2.893702*x+(695483.505816)



Supervised learning

1. Linear Regression Model

- R^2 score between stock held and total production is 0.9663 based on linear regression model.
- Linear regression model equation is Y=1.348415*x+(211645.595899)



Supervised learning

- 2. **Decision Tree Model**

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

- **3. XGBoost Model**
- The R^2 score between y_test and y_predict is 0.4897 based on XGBoost 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).

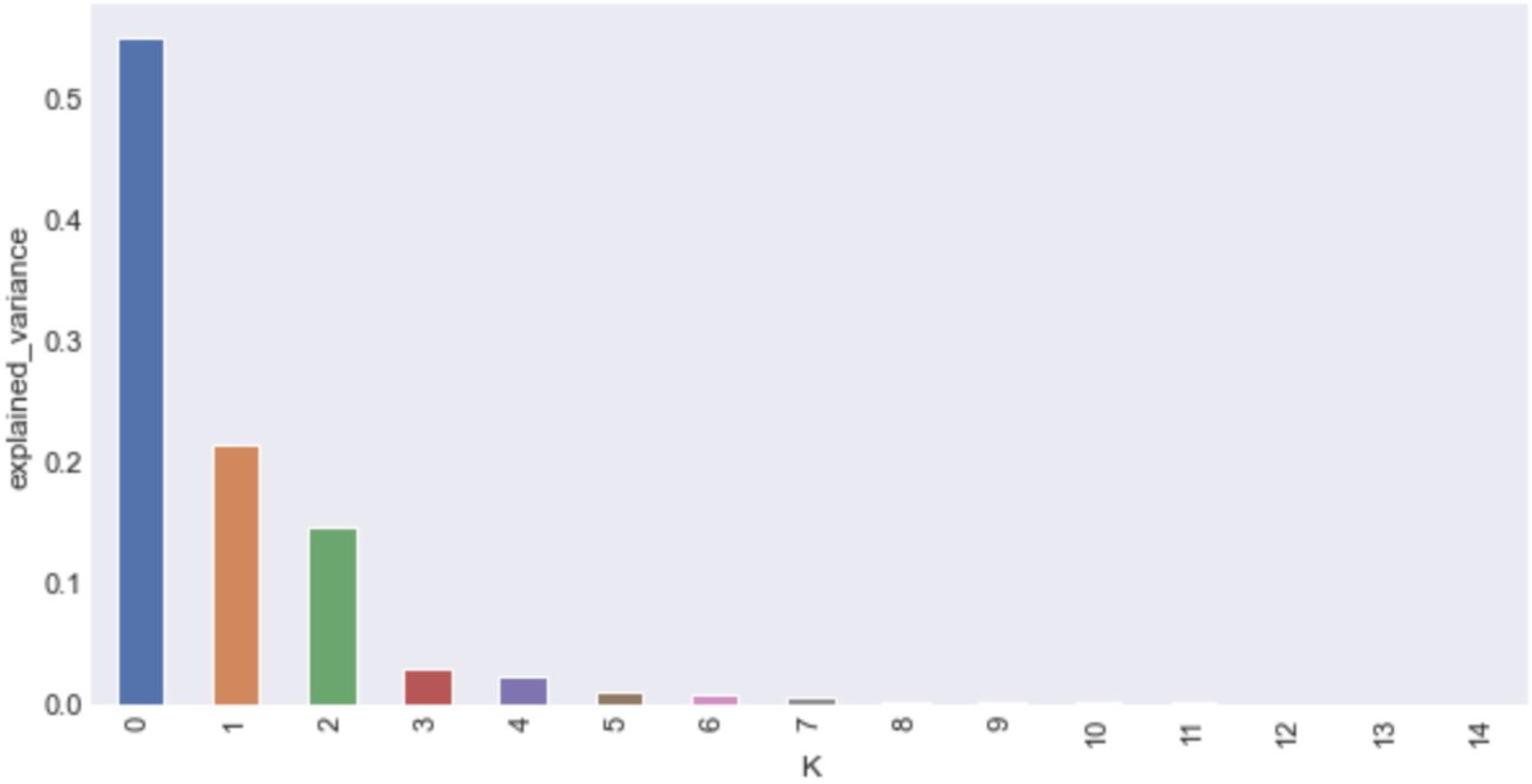
Unsupervised learning

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.

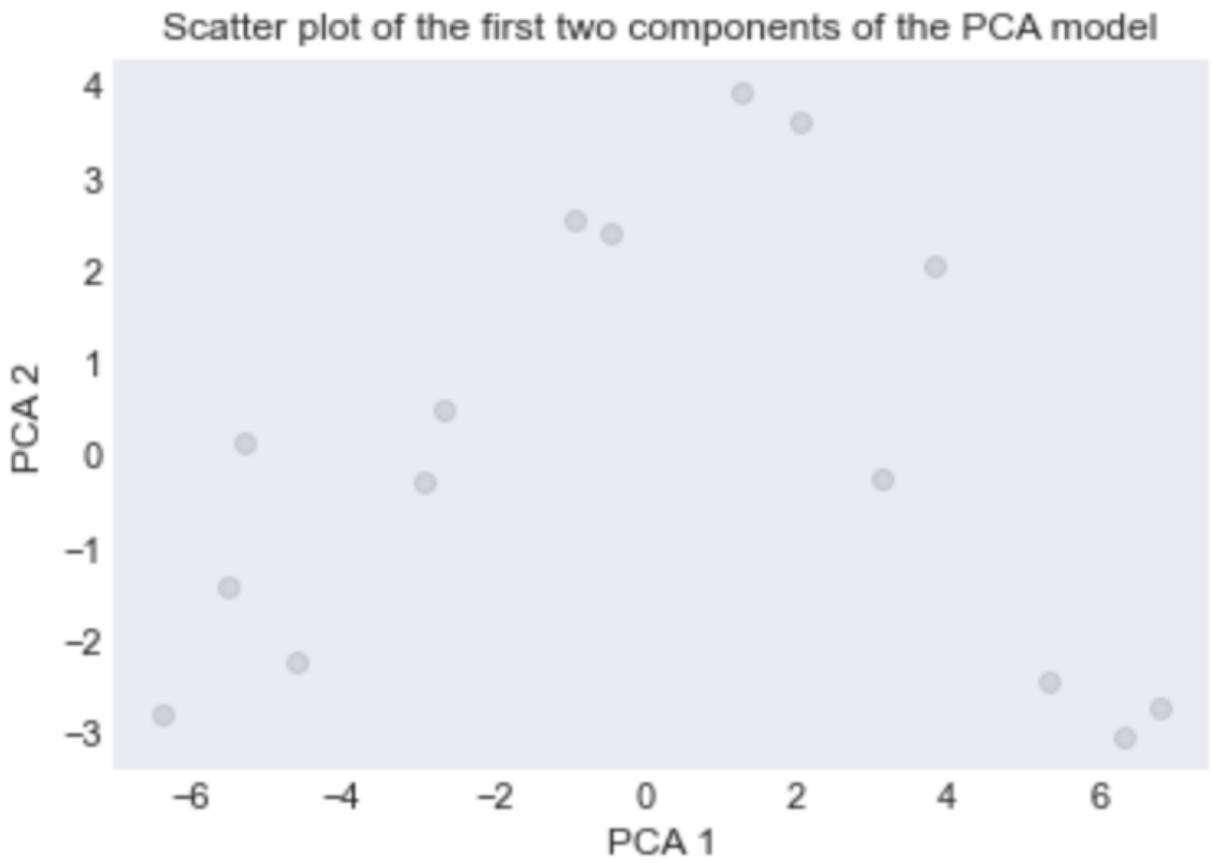
The PCA analysis

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 IV

- Based on the honey production dataset analysis, the colony number has strong correlation with total production and total production value.
- 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.

Supervised learning

- 1. R^2 score between y_test (number of colony) and y_predict is 0.3026 based on linear regression model.
- 2. The accuracy of logistic regression classifier on test set is 0.01.
- 3. The R^2 score between y_test and y_predict is -0.072 based on XGBoost model.
- 4. The R^2 score between y_test and y_predict is 0.4901 based on decision tree model.
- **The decision tree model is the best one for fitting the neonics usage data.**

More linear regression model analysis about honey production

- R^2 score between Allneonic and colony number is 0.05542.
- R^2 score between IMIDACLOPRID and colony number is 0.2605, which is the most beneficial neonics. It indicates in a defined range**(0.2~1*10^6 kg)**, the more IMIDACLOPRID usage, the more colony numbers.
- R^2 score between CLOTHIANIDIN and colony number is -0.004, which indicates there' s almost no correlation between them. Thus, CLOTHIANIDIN usage can be decreased during honeybee growing process.
- R^2 score between ACETAMIPRID and colony number is 0.0893; the kendall corr is -0.03; so we should decrease the usage amount of this neonic pesticide.
- R^2 score between THIACLOPRID and colony number is -0.005; the kendall correlation is -0.072; Thus, we should stop using this neonic to avoid the negative influences on colonies.

Final conclusion

- neonics usage has positive correlation with honeybee colony numbers (corr=0.18) and the colony number also has postive correlation with honey production (corr=0.86), total production value(corr=0.77), stock held(corr=0.74), consumption (corr=0.79). Thus, the decreasing colony-number trend will lead to all the features listed here decreasing. Properly neonics usage can promote healthy honeybee colony developing.
- Based on linear regression analysis, we obtained several equations for predicting the related features: a. Linear regression model equation between colony number and total production is Y=75.176269*x+(-217527.838246) b. Linear regression model equation between total production value and total production is Y=0.899496*x+321513.360978 c. Linear regression model equation between stock helad and total production is Y=2.893702*x+(695483.505816) d. Linear regression model equation between price and total production is Y=-3864985.077840*x+(10028476.493532) e. Linear regression model equation between consumption and total production is Y=1.348415*x+(211645.595899)

Final conclusion

- The honey price has negative correlation with total production and with less and less prodcution, the price can be predicted to keep going up in the near future.
- Among the four regions, only MidWest region exhibit a very clear zigzag increasing-colony-number trend and also the neonics usage of Midwest region after 2003 is the most. In the short period, the neonics help the colony growing. However, we can't make conclusions that the other region also needs to increase the neonics usage to increase the production. 'California' can be a good example. Before 2003, it used too much neonics and in the following 15 years, its colony number keeps decreasing.
- Among all the five neonics, IMIDACLOPRID should be the priority; ACETAMIPRID and THIACLOPRID usage should be less and less, finally, decrease to 0.

Thanks for your attention!