Final Capstone 2 project report

Part I. Project proposal

Honeybees and Neonic Pesticides Data

Background:
Honey is an important food source. The consumption of honey and bee larvae likely provided significant amounts of energy, supplementing meat and plant food. In 2006, beekeepers globally were struck by honey bee colony collapse disorder (CCD). The best way to to kill CCD is the use of a family of pesticides called neonics; however, the excess neonics may kill bees over extened periods. Thus, predicting the honey production and track the correlational evidence between the usage of neonics and honeybee colonies are very useful.

1. **What is the problem you want to solve?**

   Two problems: one, predict the honey production based on colony numbers. Two, find out the correlational evidence between the usage of neonics and the numbers of honeybee colonies and figure out how to maximize the colony number by controlling neonic usage.

2. **Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that** they wouldn't have done otherwise?

   The beekeepers and the customers who consume the honey would be my clinet. My production prediction would give suggestions to beekeepers to obtain the highest value of sales; and the prediction on how to use neonics to obtain the most honey colonies would benefit the beekeeps to keep the bee healthy and maximize the honey production; finally, All these factors will promise the enough honey providing in the market for consumers.

3. **What data are you using? How will you acquire the data?**

   I am using the honey bees colonies and neonicotinoids data, which comes from Kaggle website: https://www.kaggle.com/kevinzmith/honey-with-neonic-pesticide#vHoneyNeonic_v03.csv

4. **Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.**

   My approach outline is data importing, data wrangling, data visualization and model predicting; The detail are as follows:

a. For the honey production prediction: Based on the production data and the usage of neonics data during 1998-2016, three different inference tools, frequentist interference, bootstrap interference and Bayesian interference would be employed and evaluated to find the best model to predict the production.

b. For the correlational evidence between the usage of neonics and the numbers of honeybee colonies: the kendall correlation method would be used to predict the correlation between five neonics and the number of honeybee colonies.

5. **What are your deliverables? Typically, this includes code, a paper, or a slide deck.**

   There are mainly two deliverables for this project:

   a. Jupyter notebook that includes all my raw code and reasoning for the decisions I made.
   b. PowerPoint presentation that summarizes the key results from the project and future directions that would be interesting to pursue.

Part II. Data wrangling

After loading the honey production data, we conduct the following steps:

1. Rename the columns to make the data easy to understand.
2. Add column ‘state’ to the data.
3. Replace ‘state_code’ with full state name.
4. Add ‘consumption’ column by using the value of ‘total production’ minus the value of ‘stock_held’.

After applying the above steps, there are 626 rows left.

After loading the neonic usage data, we conduct the following steps:

1. Rename the columns to make it easy to understand.
2. Drop several rows with empty cells and column ‘FLPS’.
3. Change ‘object’ datatype to ‘category’ datatype.
4. Fill the empty cells with 0.

for neonics data, after wrangling, there are 1132 rows left.

Part III. Initial findings (EDA part)

During exploratory data analysis, we ask the following questions to understand more about our data:

1. How does honey production change?
2. What’s the honeybee colony number changing trend?
3. After applying neonics to honeybee, which combination of neonics promote the colony numbers most?
4. How to obtain maximum production value?

Initial findings are including:

1. We analyze the total honey production in different states from 1998 to 2012 and find out ‘North Dekota’ has the highest honey production in 2010 and visualize the results in figure 1. In most states, the honey production decrease during this period.

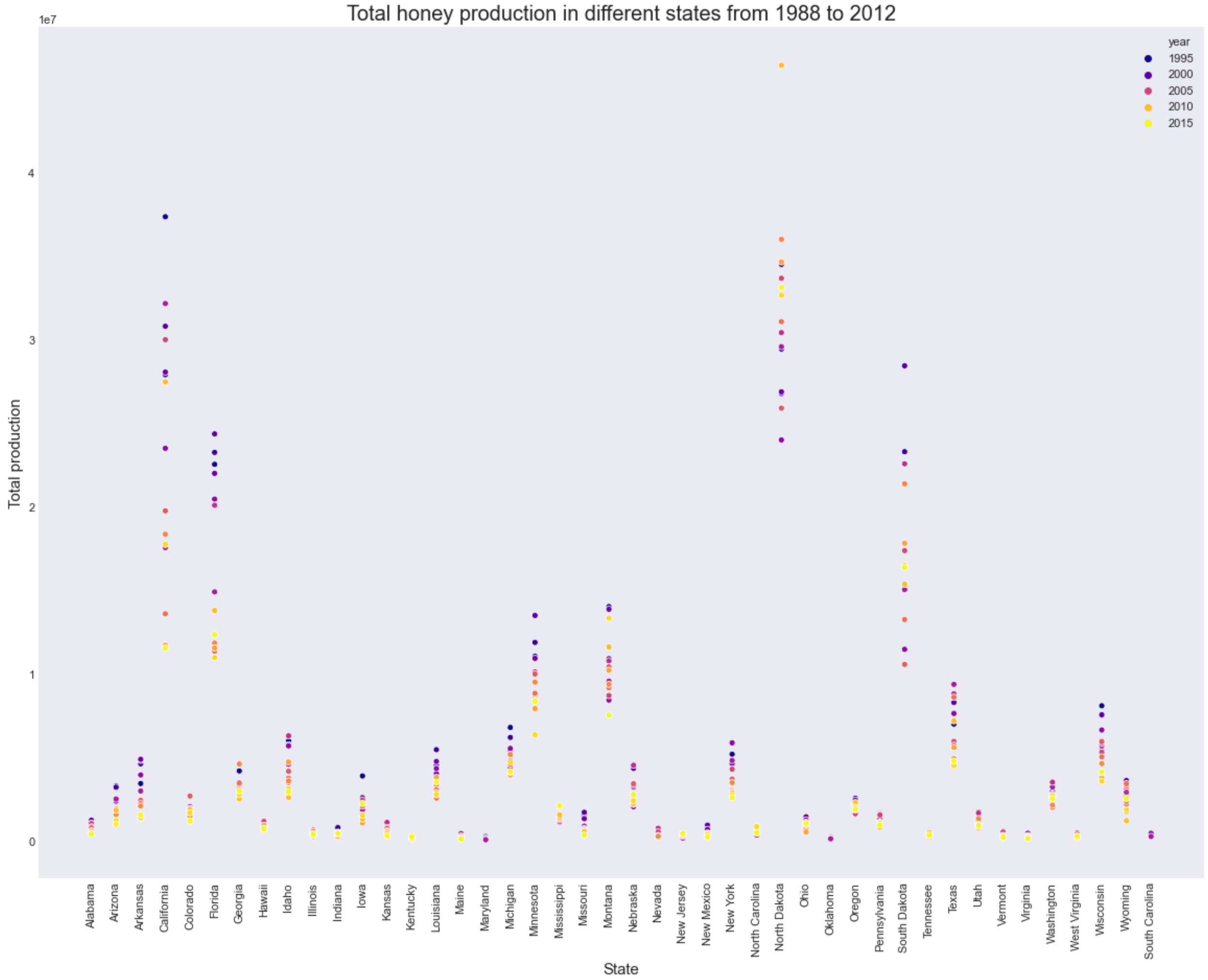


Figure 1. The total honey production in different states of the USA from 1998 to 2012.

2. Then we checked the total honey production in the whole USA to see how the honey production changing. We find that during 1998 and 2012, the USA produces the maximum honey in 2000 (Figure 2). After 2000, the production exhibits the decreasing trend very clearly.
3. We analyze the honey stock, total production value and consumption and summary the results in Figure 3. The stock and consumption exhibit the plain decreasing; however, the total production value shows the zigzag increasing trend. Among all the states, ‘North Dakota’ has the most production during this period and ‘South Carolina’ has the minimum one. It is interesting to find the reason.

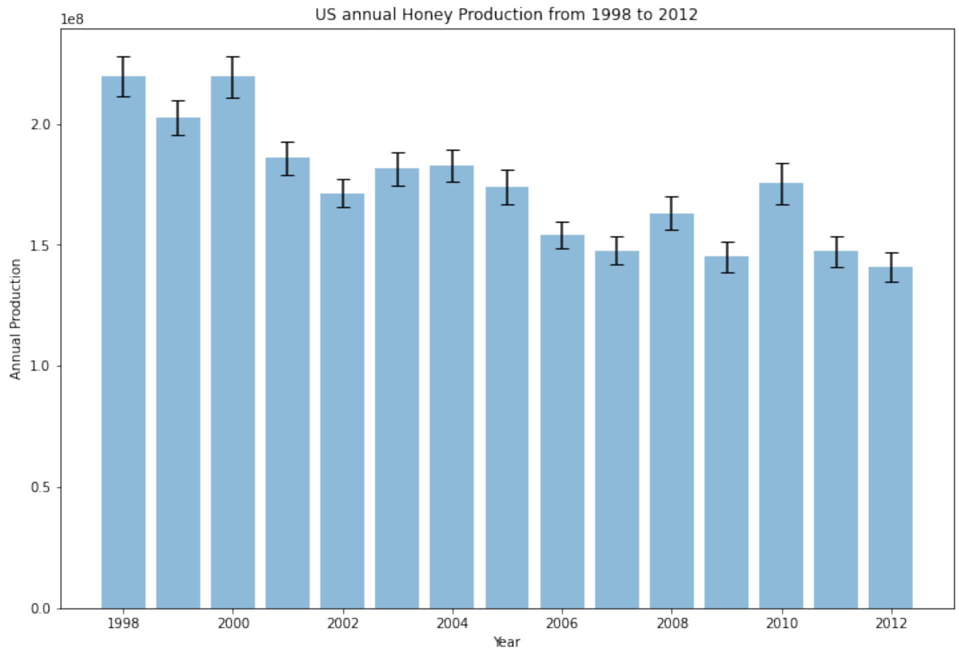


Figure 2. The annual honey production during 1998 and 2012 in USA.

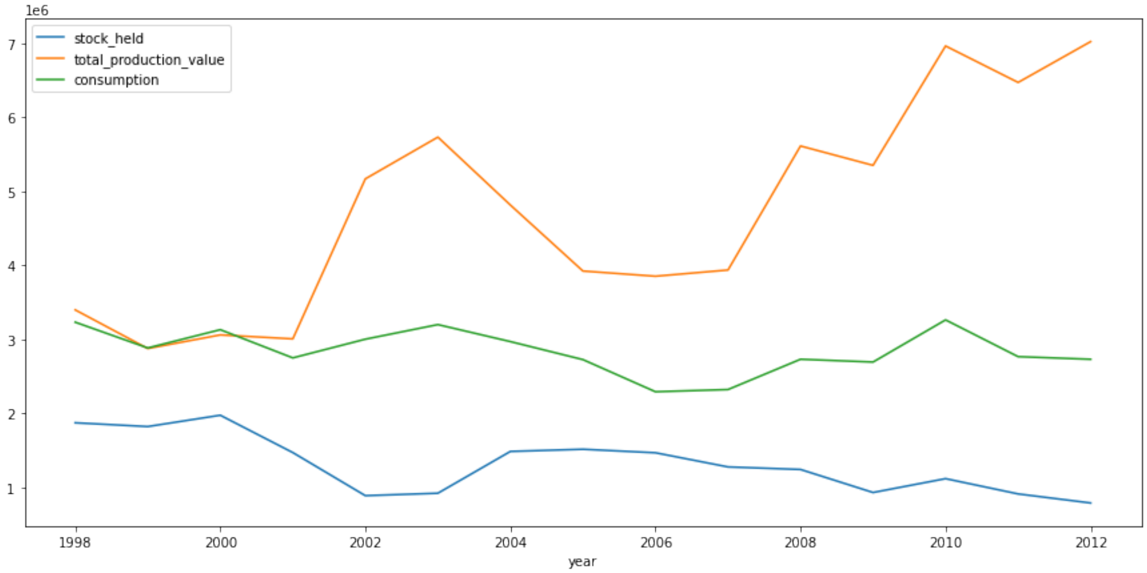


Figure 3. The stock held, total production value and consumption changing trend in the USA during 1998 to 2012.

4. Why does the production decrease? Let's watch the honey price per lb. It keeps increasing since 2004; during 1998 and 2004, it has the peak at 2003 by '$1.4973' (Figure 4). Because the price goes up, it promotes the zigzag increasing of total production value, even though the annual production decreases.

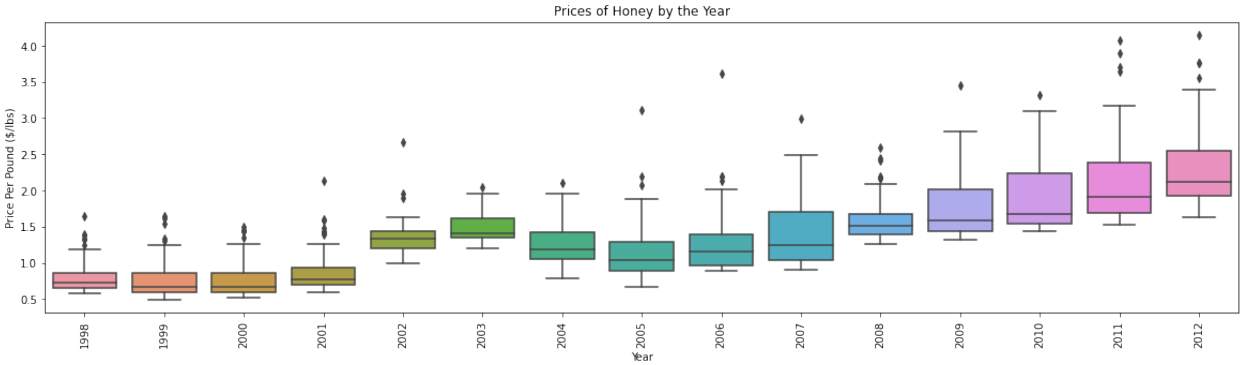
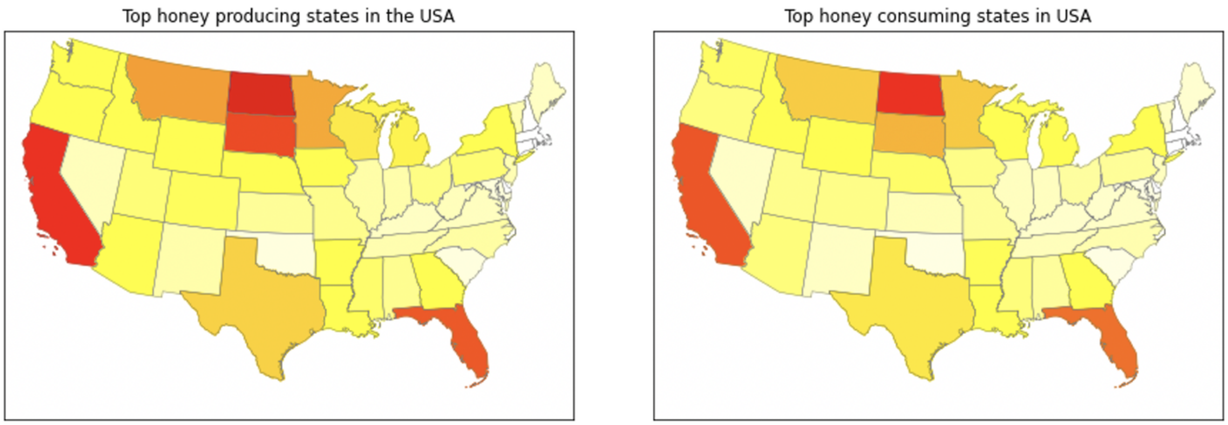


Figure 4. The evolution of the price of honey during 1998 and 2012.

5. How about the honey consumption by state? We visualize the state consumption and producing maps (Figure 5). Based on honey production-consuming analysis, 'North Dekota' and 'Californian' rank as the top 2 in both honey-producing and honey-consuming states. 'Florida' ranks in third in consuming and in fourth in producing honey; on the contrary, 'South Dekota' ranks in third in producing and in fourth in consuming honey.


Top honey producing states in the USA

Top honey consuming states in USA

Figure 5. Top honey producing and consuming state map in USA.

6. Let’s find out the correlation between the production and the number of honeybee colony with kendall method (Figure 6). We get the following conclusions:

- 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.

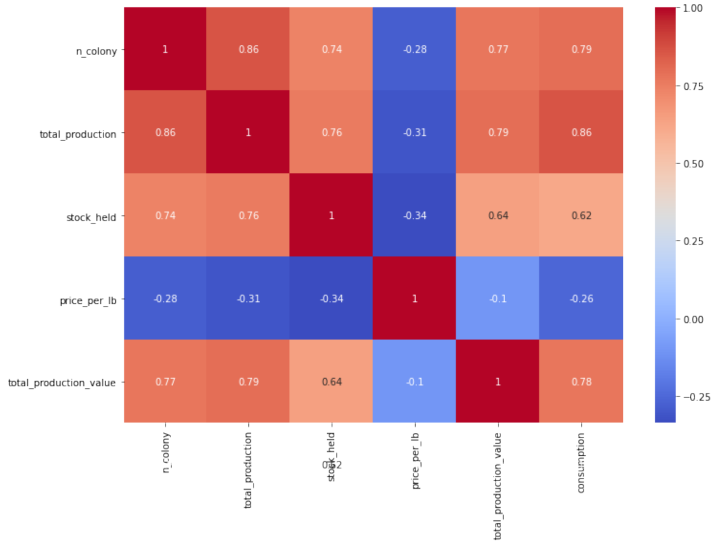


Figure 6. The heatmap of the correlation between honey colony numbers, total production, stock held, price and total production value.

7. The honey production has the strong correlation with honey colony number; how to increase colony number? Since we all know, the neonics applied in USA increase greatly since 2003 to control the colony collapse disorder (CDD). Then we analyze the correlation between five kinds of neonics and the colony number.
8. Firstly, we outline the neonic usage data(Figure 7) and find that the year ‘2003’ could be the turning point for using the neonic pesticides; since 2003, the curve for neonic usages starts increasing sharply. However, the colony numbers begin to decrease for several years and until 2006, colony numbers reach the bottom and start to increase to the peak (2010). In the following several years, it keeps small increasing trends. What’s astonished, the allneonic reaches peak at 2.87 million kg at 2014. Even though, it decreases after 2014, which still creates inestimable harm to the honeybee colonies in the long run. We will keep watching that.

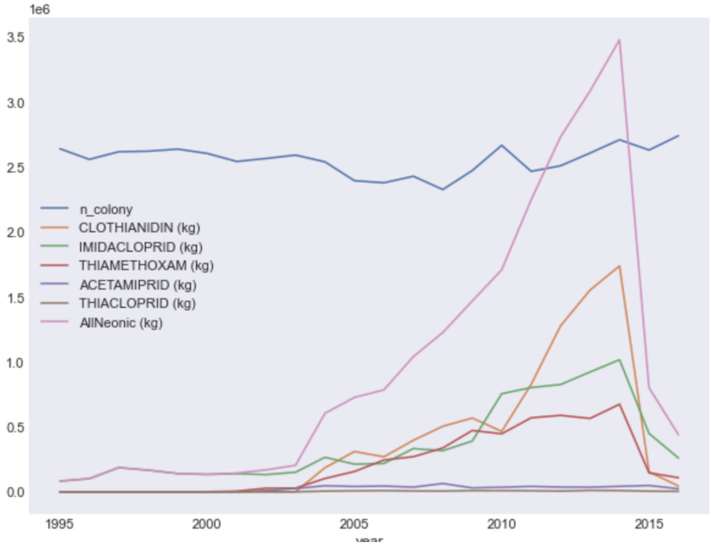


Figure 7. The changing trend of honeybee colony numbers and neonic usage.

9. Secondly, let' s check the neonic usage changing trend in different states. More than half of the states keep low neonic usage. In 1997, 'California' neonic usage already reaches 76719 kg, which ranks the top 1. Even though it decreases for several years, it still causes irreversible harm to the colonies. It double confirms that '2003' is the hinge for neonic usage in USA. Next, we focus on the neonic usages before 2003 and after 2003.

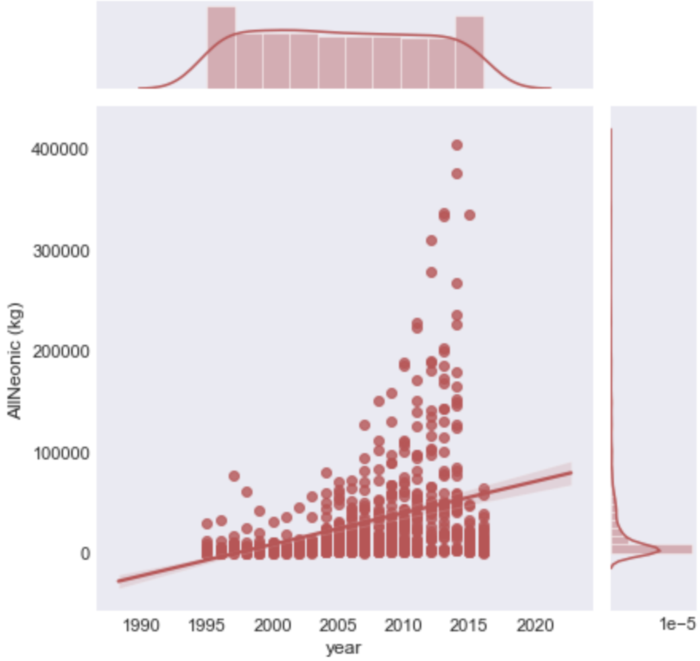


Figure 8. The neonic usage changing trend among all the states in the USA.

10. We divide the states into four regions: Midwest, Northeast, South and West, and visualize the data (Figure 9, 10). In the USA, more than 50% of the honeybee locate at MidWest and West area. Before 2003, the colonies of all these four regions keep small floatation. After 2003, the neonic usage produces different effect to these regions.
    - The Northeast region uses the least neonics and keeps stable colony numbers.
    - The South region reduces the neonic usage after 2003 and its colony numbers decreases since 2003 and increases after 2011.
    - The Midwest region benefits from the neonics the most and the colony numbers increase after 2003 since the neonics usage has a 100-fold increasing.
    - The West region suffers the neonics usage at a large amount before 2003; it increase usage by 6-fold, which doesn't reverse the decreasing colony numbers trend.

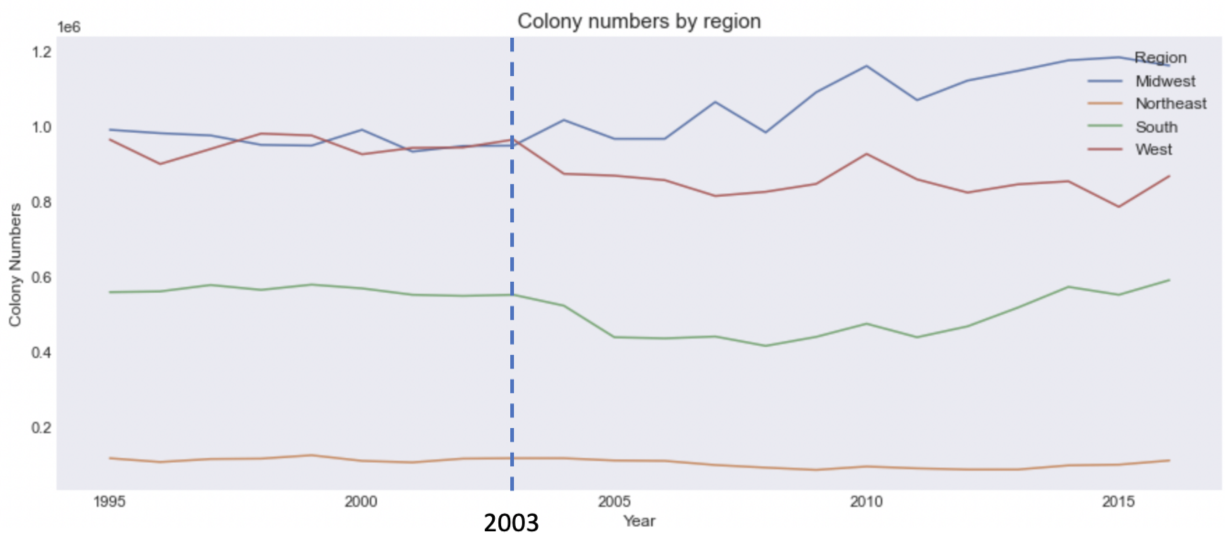


Figure 9. The colony numbers changing trend by region.

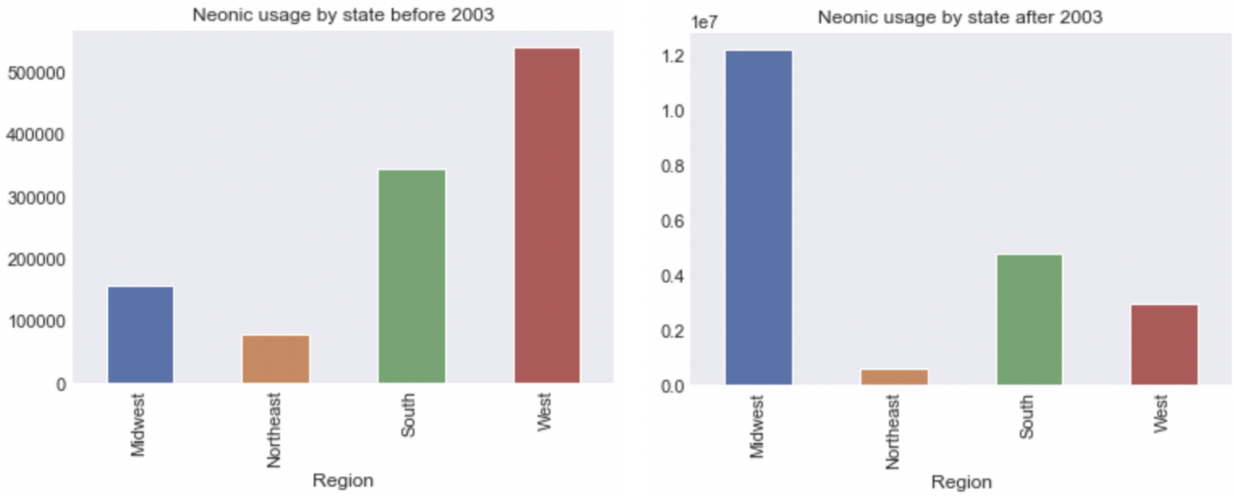


Figure 10. The neonics usage before and after 2003 by region.

11. All the five kinds of neonics exhibit different correlation trend with honeybee colony number; 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 and the heatmap is visualized in Figure 11;
a. 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.

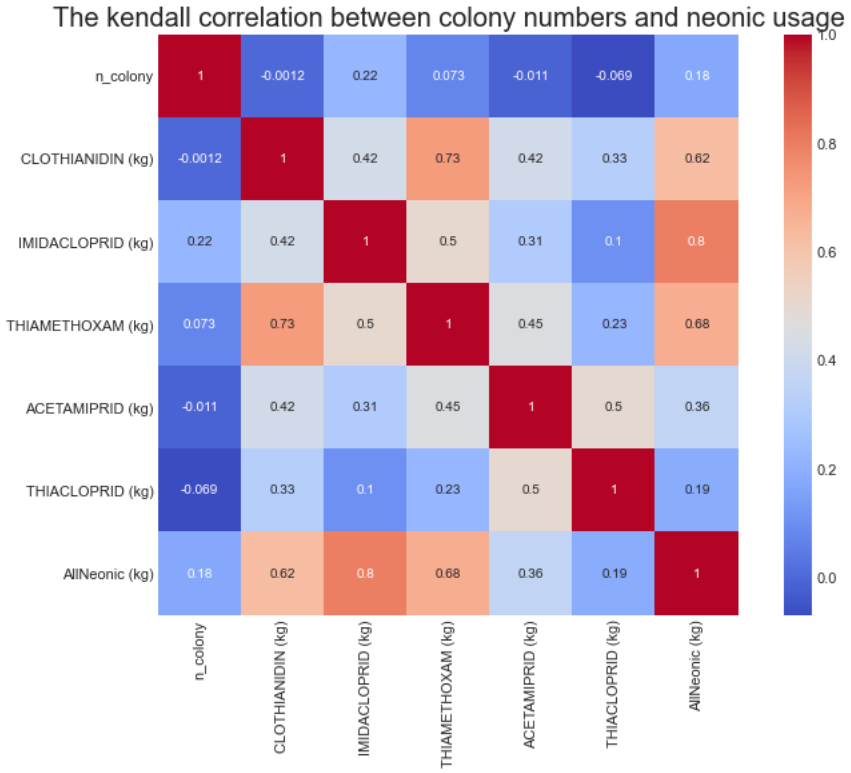


Figure 11. The correlation heatmap between colony numbers and the usage of five kinds of neonics.

Part IV. Machine learning analysis

Employ machine learning methods to predict the models can guide for making decision for honeybee keeper to avoid unnecessary losses. It includes two parts:

1. linear regression and logistic regression.
2. PCA analysis.

4.1 supervised learning

Firstly, we build the gradient boosting model on honeybee production data and visualize it in Figure 12. We identify that ‘total_prodcution’ is the most important variable. Consequently, we mainly focus on the correlation between the other features and ‘total_production’.

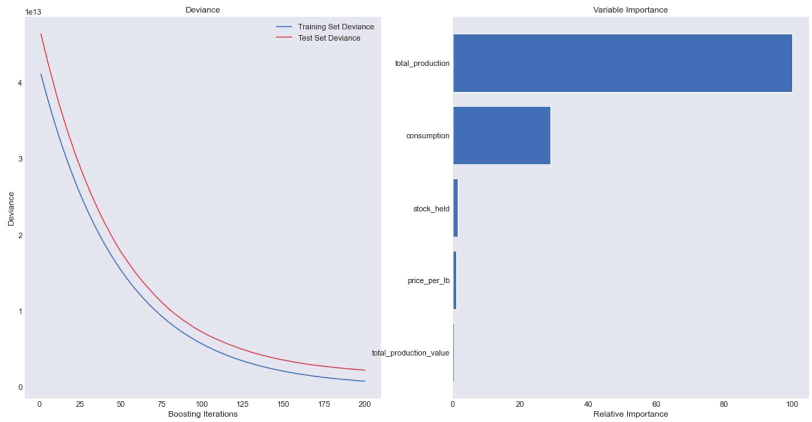


Figure 12. gradient boosting model plot.

The supervised learning analysis includes:

a. linear regression model;
b. decision tree model.
c. XGBoost model.

The linear regression model between the colony number and total production is shown in Figure 13. 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.

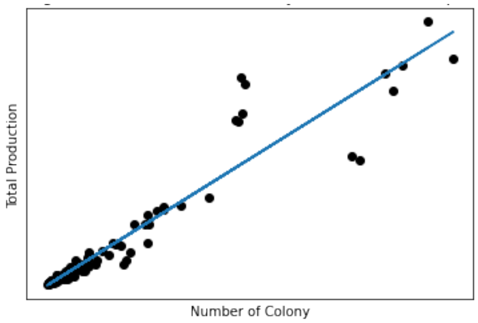


Figure 13. The linear regression model.

The R^2 of the linear regression model between production value and total production (Figure 14) is 0.7764. Linear regre**ssio**n model equation is Y=0.735375*x+369372.252656.

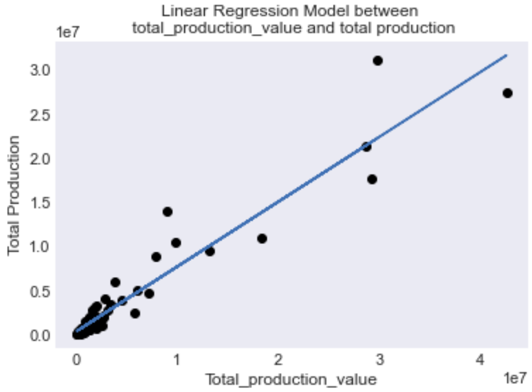


Figure 14. The linear regression model between total production value and total production.

R^2 score between price and total production is 0.09164 based on linear regression model (Figure 15). Linear regression model equation is Y=-

3864985.077840*x+(10028476.493532). Its slope is negative, which proved that the higher price, the lower total production.

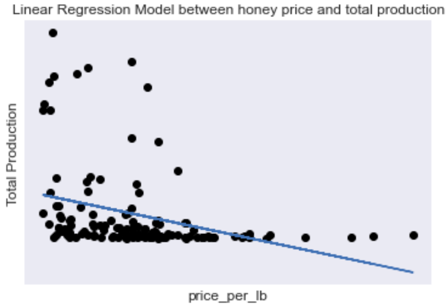


Figure 15. The linear regression model between honey price and total production.

R^2 score between stock held and total production is 0.8069 based on linear regression model (Figure 16). Linear regression model equation is Y=2.893702*x+(695483.505816)

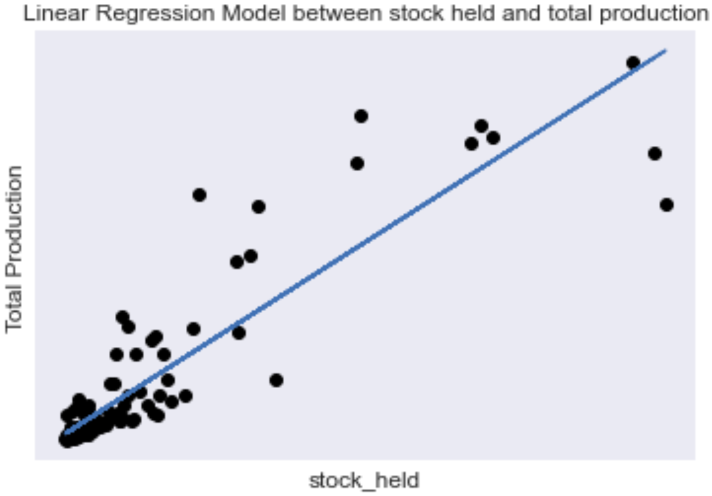


Figure 16. The linear regression model between stock held and total production.

R^2 score between stock held and total production is 0.9663 based on linear regression model (Figure 17). Linear regression model equation is Y=1.348415*x+(211645.595899)

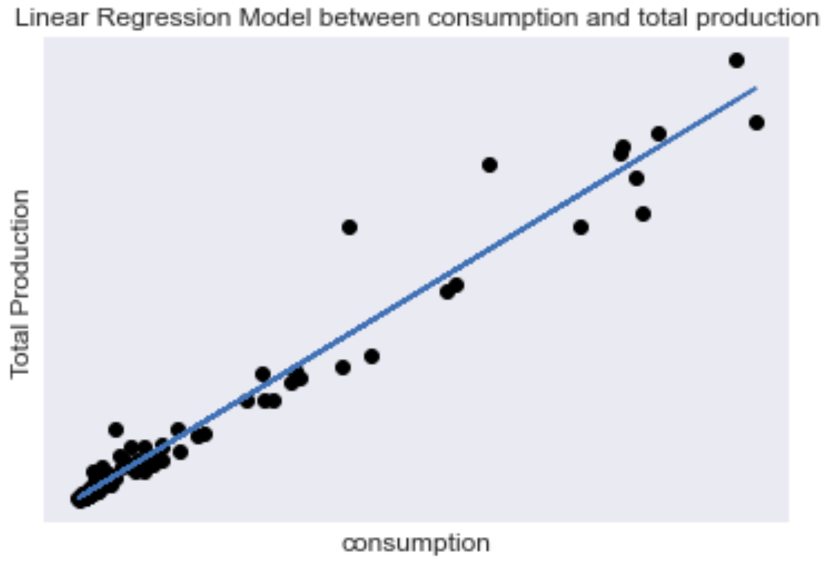


Figure 17. The linear regression model between consumption and total production.

For the decision tree model, the R^2 score between y_test and y_predict is 0.9316 based on decision tree model. The linear regression model (R^2=1) is more suitable for this honey production data than decision tree model(R^2=0.93) and XGBoost(R2=0.4897).

4.2 supervised learning: PCA analysis

As we all know, 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.

Firstly, we prepare the first matrix including index ‘state’, column ‘year’ and ‘total_production’ as values. After the PCA analysis, we obtain the figure 18 and find out the first component in year is the most important.

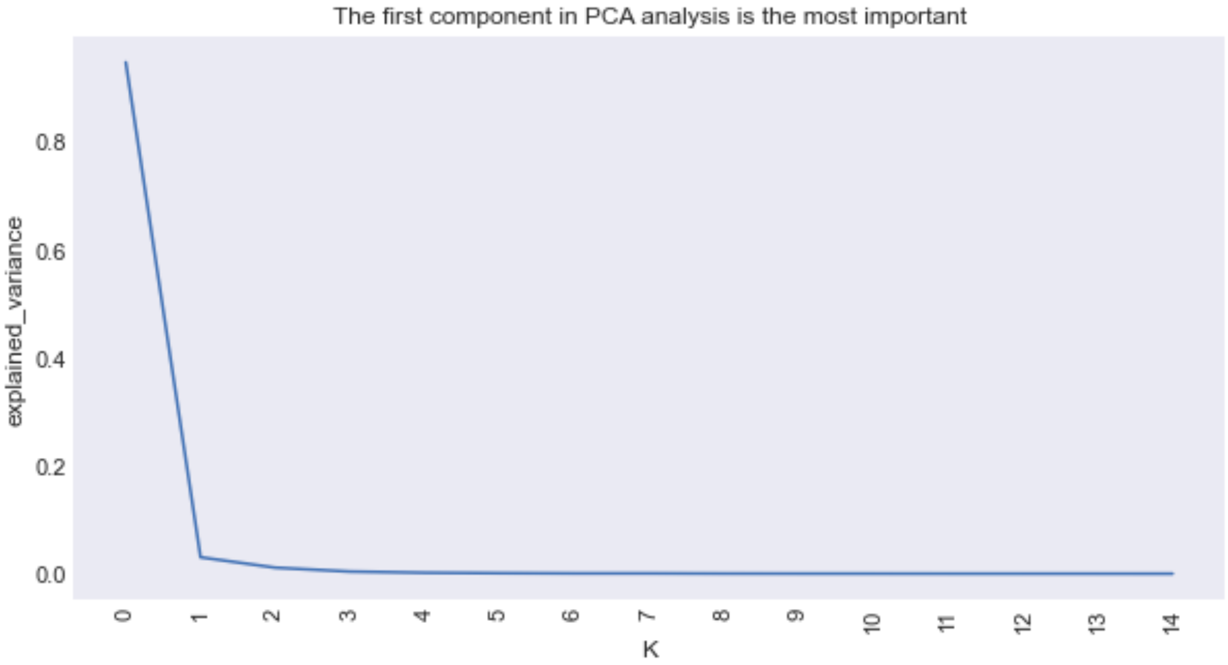


Figure 18. The PCA analysis of honey production by year.

Secondly, we prepare the second matrix including index ‘year’, column ‘state’ and ‘total_production’ as values and conduct the PCA analysis, which finds out the first three components are very import among different states (Figure 19).

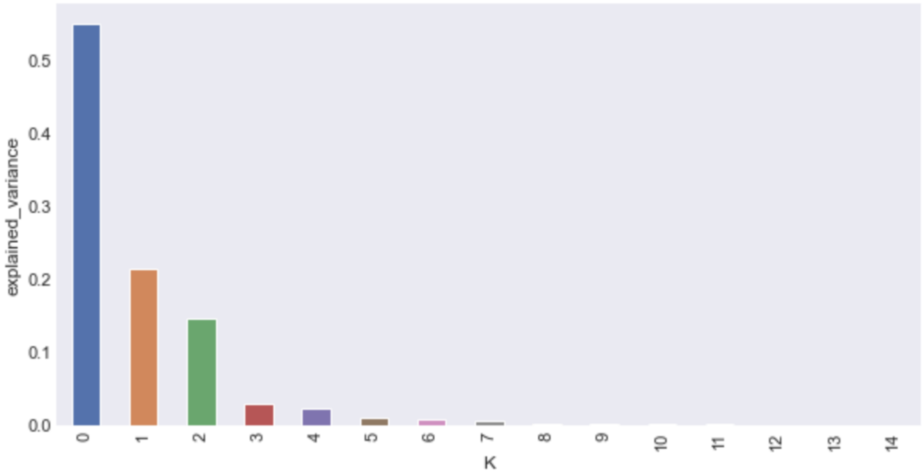


Figure 19. The PCA analysis of honey production by state.

Then, the cluster analysis of the first two components has been done and identified the first two components has no clear clusters between them.

4.3 supervised learning on neonic usage data

We build several models to fit the whole data and figure out:

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.
We analyze the modeling between single neonic feature and colony numbers and find:

1. R^2 score between Allneonic and colony number is 0.05542.
2. 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.
3. 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.
4. 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.
5. 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.

Summary

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

1. 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.

2. 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)

3. 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.

4. 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.