Capstone 2 project report 1

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.
   Two, find out the correlational evidence between the usage of neonics and the numbers of honeybee colonies.

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 data, we conduct the following steps to clean up data.

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.

Part III. Initial findings

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

1. How to get the most honey production?
2. What's the honeybee colony number changing trend?
3. After applying neonics to honeybee, which combination of neonics promote the number 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.

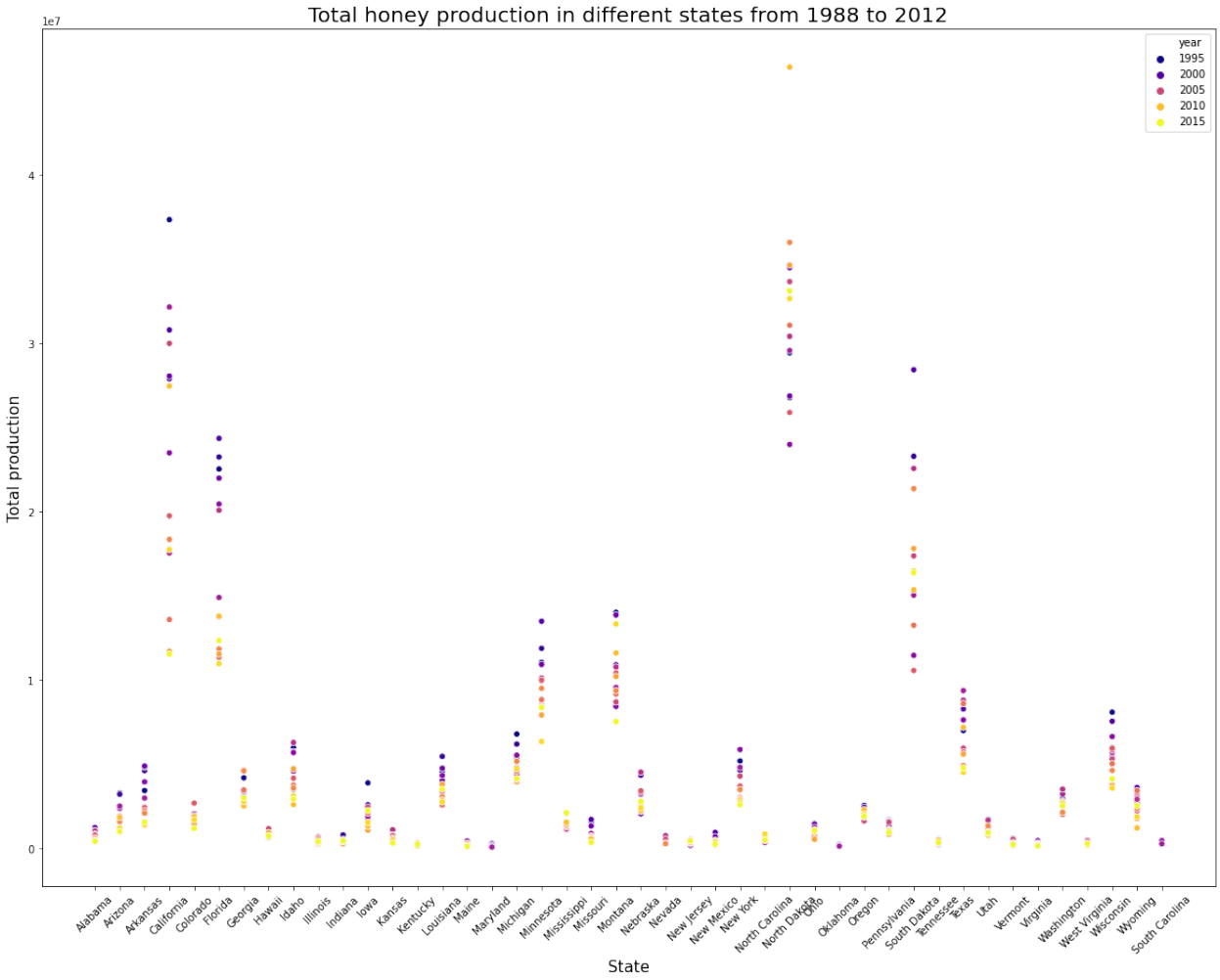


Figure 1. The total honey production in different states of the USA from 1998 to 2012.

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.

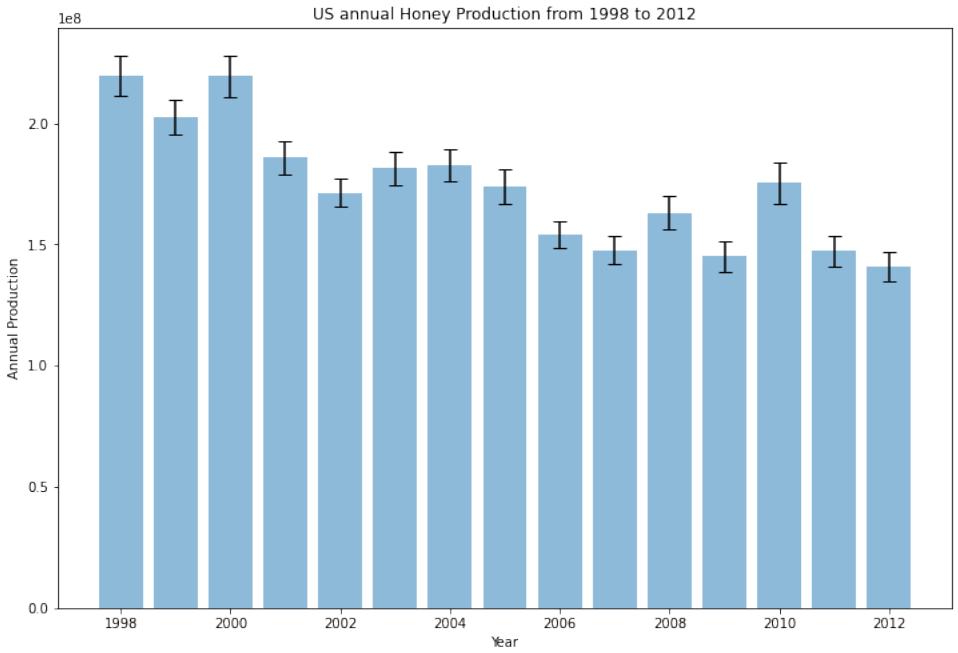


Figure 2. The US annual honey production during 1998 and 2012.

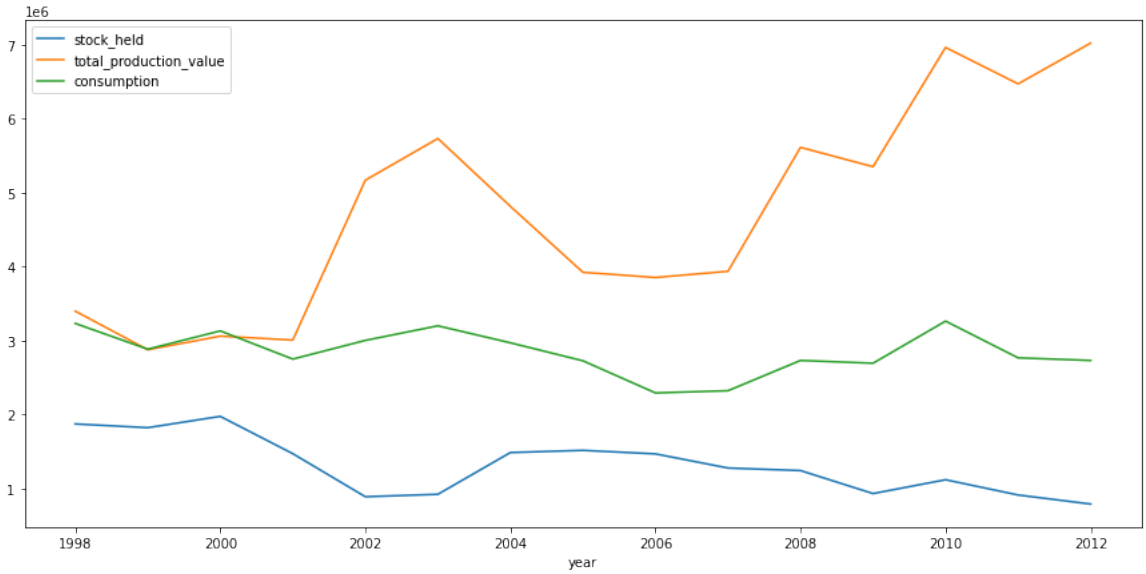


Figure 3. The stock held, total production value and consumption changing trend in the USA during 1998 to 2012.

4. Why does the production decrease? Let's watch the price per lb honey. 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.

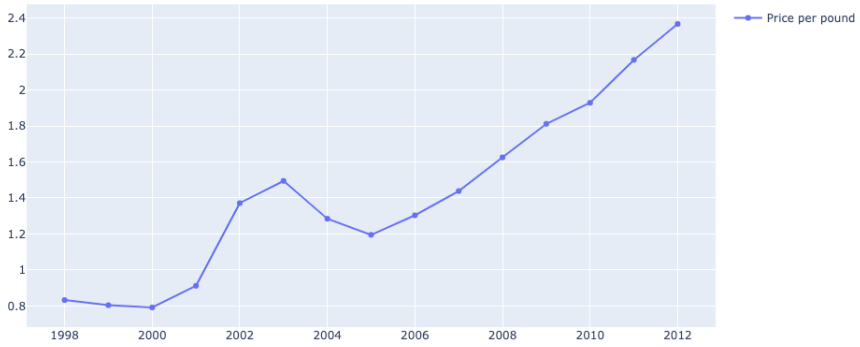


Figure 4. The evolution of the price of honey during 1998 and 2012.

5. Let's find out the correlation between the production and the number of honeybee colony with kendall method (Figure 5). 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. The effect of neonics need to be understood more deeply. Let' s check how to use neonics properly to increase the colony number.

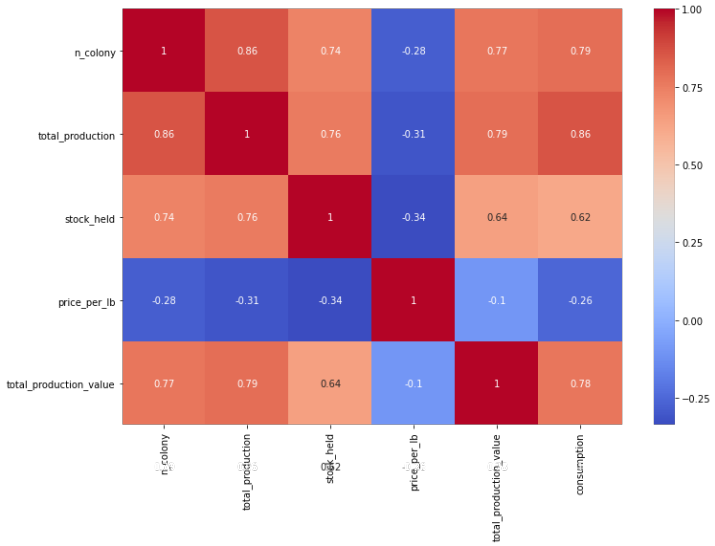
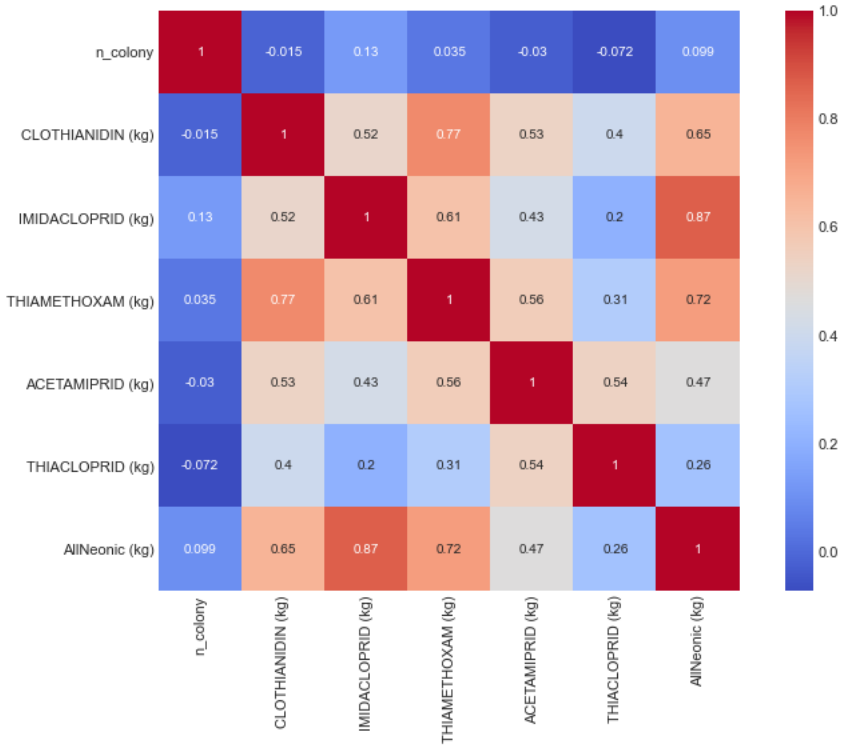


Figure 5. The heatmap of the correlation between honey colony numbers, total production, stock held, price and total production value.

6. The honey colony has the strong correlation with honey colony number; how to increase colony number? Since we all know, 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. The heatmap is visualized in Figure 6; All the five kinds of neotics exhibit different correlation trend with honeybee colony number; however, all the neonic 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.



Next Steps

1. 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.