

# Exploratory Analysis on Honey Production in USA

Data Set Link: <https://www.kaggle.com/jessicali9530/honey-production>

## Team Members

Name: Ganesha K S

SRN: PES1201701731

Email ID: [toganesha13@gmail.com](mailto:toganesha13@gmail.com)

Contact No: 8277637836

Name: Arpan Ghoshal

SRN: PES1201701700240

Email ID: [arpanghoshal77@gmail.com](mailto:arpanghoshal77@gmail.com)

## Abstract

- This assignment is about an analysis on the honey production in the states of USA.
- The data set contains the information on the production values of Honey on different colonies and the year in which that number of colonies existed
- The analytic code contains the simple box plots which compares the production of honey in various states and we can thus see the variations in the production from 1998 to 2012

## Data Set

- numcol: Number of honey producing colonies. Honey producing colonies are the maximum number of colonies from which honey was taken during the year. It is possible to take honey from colonies which did not survive the entire year
- yieldpercol: Honey yield per colony. Unit is pounds
- totalprod: Total production (numcol x yieldpercol). Unit is pounds
- stocks: Refers to stocks held by producers. Unit is pounds
- priceperlb: Refers to average price per pound based on expanded sales. Unit is dollars.
- prodvalue: Value of production (totalprod x priceperlb). Unit is dollars.
- Other useful information: Certain states are excluded every year (ex. CT) to avoid disclosing data for individual operations. Due to rounding, total colonies multiplied by total yield may not equal production. Also, summation of states will not equal U.S. level value of production

## Introduction

- In 2006, global concern was raised over the rapid decline in the honeybee population, an integral component to American honey agriculture.
- Large numbers of hives were lost to Colony Collapse Disorder, a phenomenon of disappearing worker bees causing the remaining hive colony to collapse.
- Speculation to the cause of this disorder points to hive diseases and pesticides harming the pollinators, though no overall consensus has been reached.
- Twelve years later, some industries are observing recovery but the American honey industry is still largely struggling.
- The U.S. used to locally produce over half the honey it consumes per year.
- Now, honey mostly comes from overseas, with 350 of the 400 million pounds of honey consumed every year originating from imports.
- This dataset provides insight into honey production supply and demand in America by state from 1998 to 2012.

## Questions that this analysis helps to answer:

- How has honey production yield changed from 1998 to 2012?
- Over time, which states produce the most honey? Which produce the least? Which have experienced the most change in honey yield?
- Does the data show any trends in terms of the number of honey producing colonies and yield per colony before 2006, which was when concern over Colony Collapse Disorder spread nationwide?
- Are there any patterns that can be observed between total honey production and value of production every year? How has value of production, which in some sense could be tied to demand, changed every year?

## Exploratory Analysis OR Insights Drawn

1. The three boxplots that are plotted in the kernel tells us about the production values, price per pound, yield per colony.
2. Another boxplot plotted gives us an idea about the yield per colony of the honey from different states throughout the year. This data set also includes the outliers to give an idea about the practicalities in the spread of the data.

3. The values of the production of honey is appended to a list and the name of the states to another list which is then used to plot a bar graph which helps us to identify which state in the USA produces honey which has the highest production value.
4. The values of the price of honey per pound and the yield per colony are used to plot a scatter plot which identifies the amount of honey produced which has varying values of pounds (which tells us the which honey is worth buying).
5. The scatter plot also contains the joint plot along with it to help us understand the spread of the individual data in the x axis and the y axis which was achieved using a seaborn package.
6. We also specify the use of a Histogram in the dataset by plotting the histogram for the same column i.e., yield per colony.
7. Another relationship can be drawn between stocks held by the producers of honey and the total production (yield per colony X number of colonies) by drawing a joint plot which turns out to be an approximately linear graph from which we can conclude that the more the honey was produced the more financially successful the producers were and hence they contain more stocks.
8. The xticks function was used in the bar graph to correctly specify the states against the production values of honey in the graph which normally wouldn't be possible

## Conclusion:

- Honey is becoming rarer and more expensive
- The number of colonies is decreasing in most of the states
- Top 5 states with positive correlation with respect to year: North Dakota, Hawaii, Montana, Oregon, South Dakota
- Top 5 states with negative correlation with respect to year: Missouri, South Carolina, Maryland, Arkansas, Alabama
- The production per colony is decreasing in most of the states
- Top 5 states with positive correlation with respect to year: Mississippi, South Carolina, Maine, Oklahoma, Virginia
- Top 5 states with negative correlation with respect to year: Florida, Oregon, Minnesota, Indiana, Michigan
- The total production is decreasing in most of the states
- Top 5 states with positive correlation with respect to year: Mississippi, North Dakota, North Carolina, Kentucky, New Jersey
- Top 5 states with negative correlation with respect to year: Maryland, Alabama, Missouri, Arizona, Florida