**Trends in the production of honey across the United States of America between 1995 and 2021**

**Introduction**

A dataset on honey production using data sourced from the National Agricultural Statistics Service, USDA. This dataset can be found at <https://www.kaggle.com/datasets/mohitpoudel/us-honey-production-19952021> and was created by Mohit Poudel of the Agriculture and Forestry University. This dataset was chosen due to its biological relevance to the current concerns regarding the international decline of insects with >40% of species being threatened with extinction (Sanchez-Bayo and Wyckhuys, 2018). Over the last 60 years the number of honeybee colonies in the USA has shrunk by 3.5 million (Sanchez-Bayo and Wyckhuys, 2018). This project will study interactions between factors measuring successful honey production with the aim of understanding which areas may be contributing to the downfall of honeybees.

**Methods:**

The necessary modules were imported into the jupyter notebook, followed by the data from the zip folder. The dataset was explored to identify the factors and their relation to one another, before the dataset was cleaned and checked for null values. The overall and current states that produced the greatest amount of honey was determined with indexing and filtering. Trends in production were visualised by grouping the data by state and plotting a line graph. The dataset was then ranked by production to compare datapoints on a linear scale. This was done by grouping, transforming and sorting the data.

These ranks were used in a function to compare production success levels relative to other states by creating a multiplot of line graphs individually highlighting a single state of interest and their ranking over time (fig.1). A separate function “add\_label” was written to label the first (1995) and most recent (2021) rank for each state quantifying their positional change. A ‘for’ loop was combined with if/else statements to highlight a single state per subplot. To generate the multiplot, a ‘for’ loop was created for the steps, adding a new panel for each iteration step. Changes in honey production ranking were quantified by grouping the dataset by state and aggregating to get mean, min and max ranking values over time. The distribution of average ranking was then visualised using ‘.displot()’.

The relationship between production value and amount was then investigated generally by a scatterplot before looking for geographical influence by sorting the states into general US regions. A dictionary was created to sort states into four regions, which was then inverted to map onto the dataframe. The regional relationship to honey production was displayed in a scatterplot. This relationship was separated into the regionality of production amount and average price to understand the demand-supply relationship. Finally, the effect of the colony number was studied, indexing and aggregating the data to create a subset with the average regional colony number over time. The relationship between colony number and price of honey was then visualised using a scatterplot and quantified with a spearman’s rank test.

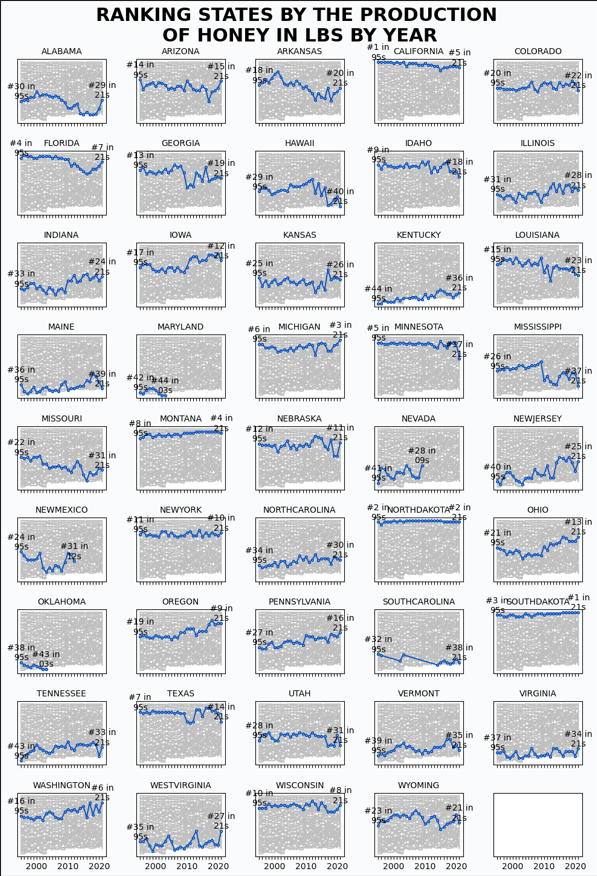
**Results:**

Indexing the raw data frame showed that although California holds produced the greatest amount of honey in one year overall, the current (2021) largest producer is South Dakota. Regardless of ranking all states experienced a substantial unexplained drop in production in 2010, which may be of interest when determining factors underlying the decline in colony success. There is variation in state ranking over time, with four states even stopping honey production prior to 2021 (fig.1). Production is positively correlated with production value but in two distinct patterns, which were thought to potentially be associated with distinct regions or differences in average price. Nearly opposite trends in average production and average price of honey were seen across the four regions. States in Northeast and Southern America price honey much higher despite producing significantly fewer lbs of honey annually. This aligns with the assumptions of regional supply and demand.

The number of colonies fluctuates over time, yet most regions show a general increase over the total period, implying the aforementioned drop in honey production may not be linked to a drop in colony or bee number and instead less honey is produced per bee. The average yield per colony deteriorating over time (R2 =-0.35, p<0.05) supports this theory.

**Figure 1: Multiplot comparing American states and their ranking in honey production.**

States in the honey dataset were ranked by the number of lbs of honey produced annually. This data is displayed across all states (pale grey lines) across time (1995-2021). Each subplot highlights the changes in the position each state holds in the honey production ranking. The original and most recent rankings are labelled in each subplot . Some plots stop earlier than 2021 which is made clear by separating the data in this manner.



**Problems of dataset:**

Some inaccuracies were encountered in the analysis of this dataset regarding the average price, coupled with several factors being dependent on each other meant it did not lend itself to comprehensive and sturdy model building however did provide some useful insights.

**Conclusions:**

The overall production of honey in the US is drastically decreasing, however this analysis shows that this is a separate problem to the current concerns regarding increased death. Individual bees/colonies are becoming less efficient, with reduced yield. Thus, something may be detrimentally impacting the health of honeybees in America thereby reducing their honey production. More information is needed on factors effecting honeybee health and honey production efficiency before a comprehensive conclusion can be reached, for example information on environment (temperature, humidity, rainfall), predation, breeding practices, population density (city vs rural) and pesticide use would all contribute towards building a more complete picture of what is happening to the bees in America. Establishing the underlying problem will give us something to target and hopefully begin to stop and potentially reverse the harm to the honeybee ecosystem and may even give us insight into the larger issue of insect death and extinction. Such a model could even potentially predict events such as the 2010 dip in production.

GitHub URL: <https://github.com/mos5939/Assessment-Honey_production>

**References:** Sánchez-Bayo. F. & Wyckhuys. K.A.G. (2019) ‘Worldwide decline of the entomofauna: A review of its drivers’, *Biological Conservation,* 232, pp. 8-27.