**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 in Nepal with the goal of providing insights into honey production, demand, and supply across the United States of America.

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 from a peak of 60 million in 1947 (Sanchez-Bayo and Wyckbuys, 2018). This project will study interactions between different factors measuring success of honey production with the aim of understanding which areas may be contributing to the decay of honey production.

**Methods:**

The dataset after initial exploration was “manipulated” to investigate trends within the data.

The necessary modules were imported into the jupyter notebook, followed by importing the dataset itself from the file. The dataset was then explored to identify the factors and how they related to each other, 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 amount for comparison of all datapoints on the same linear scale. This was done by grouping, transforming and sorting the dataset.

These ranks were then used in a function to compare their production success levels relative to other states by creating a multiplot of line graphs individually highlighting a single state of interest and their change in 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 to quantify their positional change over time. A ‘for’ loop was combined with if/else statements to highlight a single specific state per subplot. To generate the multiplot of all states, a ‘for’ loop was created for the steps, adding a new panel for each iteration step.

The average change in honey production ranking was quantified by grouping the dataset by state and aggregate. This produced the 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 first generally by a scatterplot before looking for geographical influence by sorting the states into more general regions of the U.S. A dictionary was created for sorting the states into their respective regions, which was then inverted for mapping onto the dataframe. The regional relationship to honey production was then 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 number of colonies was studied, indexing and aggregating the data to create a subset with the average number of colonies within each region 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:**

* Although California holds the title for greatest production of honey in one year overall, the current (2021) largest producer is South Dakota.
* There was an unexplained substantial drop in honey production across all of America in 2010 which would be very interesting to explore.
* There is a lot of variation in state ranking over time with four states even stopping honey production at various points prior to 2021 (fig.1).
* There are two patterns of relationships between production and value of production, both showing a positive correlation (one steeper and one more gentle)
* 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 but with most regions showing a general slight increase between 1995 and 2021 implying that the aforementioned drop in honey production is not associated with a drop in colony or bee number but that the colonies are producing less honey per bee.
* The average yield per colony is deteriorating over time (R2 = -0.35, p<0.05) supporting this theory.

**Conclusions:**

The overall production of honey in the US is drastically decreasing, however this is a separate problem to the current concerns regarding increased insect death. On top of the bees dying, individual bees/colonies are becoming less efficient, with reduced successful honey yield. Thus, it can be concluded that there is some detrimental impact on the health of honeybees in America which is making them less likely to produce honey. More information needs to be collected 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.

GitHub URL: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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.

