Flight of the Bumblebee

Ciara Maher and Nooha Mohammed

Summary of Questions and Results

1. Looking at multiple diseases and threats, in which state(s) has the honey bee population been hit hardest since 2015?
   Using geopandas we will create a map visualization showing the decline of colonies for each state, with color showing the range of colony loss
2. How did the greatest threats to honey bee populations change over 2015-2021?
   Looking at the specific honey bee stressors and threats to population, we'll be able to see the biggest threat to honey bees in each year available. This will show how the threats have morphed over the years and what is currently, or most recently, the biggest threat.
3. Looking to the future, how can we predict the honey bee populations to change in 5 years?
   Using an ML model trained on data on honey bee populations across the different states and years, we will be able to predict future population sizes. This will be achieved through the scikit-learn library.
4. How has the price of honey fluctuated and what stressors may have the biggest impact on honey prices?
   Comparing the prices of honey and high stressors over time will allow us to see if threats to honey bees have affected the pricing for honey, and which ones are the most dangerous for the market.

Motivation

Honey bee populations have been greatly threatened over the past decade, and as a vital pollinator this change threatens crop growth and ecosystems. Looking at the states that have had the most threats to their honey bee populations, we will be able to see which states need to begin addressing the issues that may cause further decline. We can also see the trends of which threats have been the most significant over time and which will be most significant to address in the contemporary moment. Being able to predict how populations will decline in the future will also allow us to see which areas will need the most help and how dangerous this issue can become in the future.

Dataset

The dataset for this project was created in Kaggle, and originally sourced from the USDA NASS program's annual data on honey bees. This is a collection of three csv files focusing on colony

amounts, specific threats or stressors to colony population, and honey production in the US. It has columns specifying states, years (between 2015 and 2021), and the seasons indicated by data collection “quarters". While the data has been pre-processed into Kaggle, there is a small portion of null values in the dataset that will have to be filtered through.
Link to this data set can be found here:
NASS Honey Bee 2015-2021 | Kaggle

Challenge Goals

The first challenge goal we plan to meet is working with **multiple datasets**. Our data includes three different csv files that will all be valuable to answering our research questions. These files contain data on bee colonies, honey production, and bee stressors. We believe we will be using all three of these files to answer questions about how stressors affect bee colonies and honey production, in what areas particular stressors (like environmental changes) are most common, and overall measuring of bee health.

The second challenge goal we plan to meet is **learning a new library**. We hope to create unique and insightful visualizations that go beyond the visualization techniques we have learned so far. We hope to use this knowledge of a new library to tell a story with what the data means and help inform others about our research topic. This will include interactive visualizations that go along with the theme of our project, which will likely be done by using altair or plotly.

A third challenge goal we would like to meet is working with a **machine learning** model. We want our research project to help predict the health of honeybees in the future. Therefore, our goal is to create a machine learning model that can predict declines or increases in honeybee populations and honey production, based on past trends and stressors in the environment. We will work with different models and hyperparameters to answer our third research question about whether or not we can predict future honeybee population trends.

Method

Research Question 1:

To answer the question: *Looking at multiple diseases and threats, in which state(s) has the honey bee population been hit hardest since 2015?*

1. Clean each dataset to remove any unnecessary rows and replace missing values with Nan
2. Join bee_stressors and bee_colony files together by state
3. Group the data by state and sum up the percentage lost.
4. Report max percentage lost and which state.
5. Group data by state and sum up stressor percentages.
6. Report row with the index of the state that had the highest percentage of colonies lost with summed up stressors.
7. Report largest stressor.

Research Question 2:

To answer the question: *How did the greatest threats to honey bee populations change over 2015-2021?*

1. Filter stressor data to remove unnecessary rows and replace missing values with NaN.
2. Group by year and sum up percentages for each stressor.
3. Report highest percentage stressor for each year.
4. Create a line chart plotting a line for each stressor and its total percentage for each year.

Research Question 3:

To answer the question: *Looking to the future, how can we predict the honey bee populations to change in 5 years?*

1. Create a Regression Classifier model with percent_loss as the label, and State, each stressor column, and year as the features.
2. Split data into 10% testing, 10% validation, and 80% training.
3. At each max_depth, train the model, create predictions, and test its error for both the validation and test data.
4. Compare errors and choose max depth to minimizes error.

Research Question 4:

To answer the question: *How has the price of honey fluctuated and what stressors may have the biggest impact on honey prices?*

1. Group by year and find the average price of honey for that year.
2. Group by year and find the max percentage of colonies lost from each stressor.
3. Plot average price of honey along with the highest percentage for each stressor on a plot.

Work Plan

We will go about this project as follows:

1. Set up: This task involves reading in all of our datasets as CSVs into our development environment. We will also import all of the libraries we will be using for this project.
    a. Estimated Time: 2 hours
2. Follow steps for Q1 and create plots to visualize findings
    a. Estimated Time: 4 hours
3. Follow steps for Q2 and Q4.
    a. Estimated Time: 6 hours
4. Follow steps for Q3
    a. Estimated time: 8 hours
5. Write report and create poster
    a. Estimated Time: 6 hours