Problem statement formation - Capstone 2

Hypothesis: In Wisconsin, Minnesota, Illinois, Michigan, Iowa, and Indiana, the largest cause (by percentage) of bee hive die-off is different from other states, requiring a focus on different disease prevention tactics than are standard in other states.

Context: Colony collapse disorder has been viewed as the largest threat to bee populations. However, a rise in bee hive die-off due to varroa mites may be present in the states of focus. If true, this would make it more economical in money and time to focus on the prevention of mite infestation over prevention of the causes of colony collapse disorder.

Criteria for success: Beekeepers in the northern Midwest states will know which disease or pest to focus their efforts and funds on and will know what increase in percentage of hive survival they can expect with effective mitigation efforts.

Scope of solution space: Data exists for 2016 through June 2021, broken into quarterly segments. Data is not typically collected from hobbyist beekeepers, who customarily have only one or two hives. Hobbyist beekeepers will have to extrapolate based on the data from larger beekeepers. Data is partitioned by state. Given the large variation in climate and conditions across large states, especially in California (large state, with a very large bee population due to its agricultural focus), having data broken up more would be more useful.

Constraints: A wide variety of stressors can cause the loss of a bee hive. For instance, varroa mites can weaken a colony, making it more vulnerable to pesticide use or severe weather. Additionally, sometimes an entire hive just leaves with no prior indication and no reason left behind as to cause. Therefore, the conclusions are reliable only with those caveats.

Stakeholders: Beekeepers, retailers that sell bee supplies, farmers near beekeeping operations that use pesticides on their fields

Deliverables:
Slide deck and report
Github repo with code

Initial thoughts on approaches and concerns:

- Data from sites with fewer than 5 colonies will have to be analyzed separately from sites with more than 5 colonies (see note on data cleaning below)
- Map percentages of reasons for hive loss over time
- Look for trends of reasons for hive loss: Seasonal trends? Trends across states? Trends between large operations and small operations? Trends from year to year?
- Are Wisconsin, Minnesota, Michigan, Indiana, Illinois, and Iowa similar enough to each other and different enough from other states that it makes sense to make recommendations for all six states as a unit? If not, what states could be grouped together with Wisconsin, if any?
- Use machine learning to predict forward in time. Does the trend of data from 2016 to 2021 support predicting forward in time?

Data sources:
USDA data as collected by Cornell University:

https://usda.library.cornell.edu/concern/publications/rn301137d?locale=en
https://www.kaggle.com/ellies15/bee-colony-data-cleaning-usda-data

Dependent variable:

- Percent of colonies lost in a 3-month period (e.g. Jan-Mar, Apr-Jun, etc)

Independent variables:

- State
- Quarter
- Year
- Percent lost (total of all lost)
- Percent of hives lost to varroa mites
- Percent of hives lost to other pests

- Percent of hives lost to pesticides
- Percent of hives lost to diseases
- Percent of hives lost to other known-cause failures
- Percent of hives lost to unknown failures
- Total number of hives (by state, maximum per quarter)
- Hives added during quarter
- Site has less than 5 colonies or more than 5 colonies

Data cleaning issues:

Sites with less than 5 colonies have data collected annually, while sites with more than 5 colonies have data collected quarterly