Phoebe Ng, Kyle Roland, Gavin Sreesangkom, Michelle Lee (BE4)

Informatics 201 – Ott Toomet

Final Project Proposal

November 15th, 2018

Honeybees in Need

**Project Description –**

Data Sets:

For our final project we will be using several datasets. Our first dataset is called “Honeybees and Neonic Pesticides” and is taken from *Kaggle.com*. It includes information about the number of honey-producing colonies, the total production of honey, the amounts of specific types of neonic pesticides, the state, and the year. This dataset aggregates data collected by The National Agricultural Statistics Service (NASS), which is the primary data reporting group for the USDA. The NASS collects information on all aspects of U.S. agriculture including honey production, which is used in this Kaggle dataset. This dataset also takes information from the USGS through its pesticide national synthesis project which estimates the annual agricultural pesticide use. The estimates are calculated by using county crop acres collected from the previous Census of Agriculture. The USGS was created by an act of Congress in 1879 and it provides reliable scientific information to describe and understand the earth. It is the sole science agency for the Department of the Interior, which is the US federal executive department in charge of managing and conserving federal lands and natural resources.

Another dataset that we will use is the “Annual Bee Colony Lost Data” which is taken from *data.world* under the project “Honey Bees and Apiculture”. It includes information about the total annual loss of bees and the number of colonies specific to each state and year. The original data was retrieved from the United States Department of Agriculture.

The third dataset we will use is called “U.S. Pollution Data”, accessed from *Kaggle.com*. It includes information about the amounts of four major pollutants (nitrogen dioxide, sulphur dioxide, carbon monoxide and ozone) in different cities, states and years. The original data for this was taken from the U.S. Environmental Protection Agency (EPA). This will be later used in our project in comparison with previous dataset to understand a possible correlation with pollution levels and honeybee population.

Target Audience:

Our goal of using multiple datasets during this project is to understand factors about colony collapse disorder and raise awareness through meaningful visualizations. Our project will focus on targeting the U.S. Environmental Protection Agency. Although the EU has already banned neonic pesticides, the EPA has not taken any action toward mitigating the issue at this stage. Since we are taking datasets that include information about the neonic pesticides as well as pollution, we aim to identify patterns in the use of neonic pesticides and amounts of pollution that may be influencing the decline of bees to encourage a call to action from the EPA.

Questions for the Audience:

1. What states have the greatest average decline in bee population?
2. Is there a trend in which areas are declining the most in bee population, such as geographical location?
3. How do different pesticides correlate with bee populations?
4. Is there a relationship between areas with high pollution levels and bee population?

**Technical Description –**

Final Product Format:

We intend on using the Shiny app to make the visualization interactive and give the users the ability to explore through changing the parameters and variables that could potentially be influencing the decline of honeybees. For example, a user may be able to choose which pesticide to compare with honeybee population. We are also likely going to use html to display different 2d plots and maps of the datasets.

Reading Data:

We will be reading in all our data from static .csv files.

Data wrangling:

To answer the questions above and more, we will need to do a fair amount of data wrangling, which will include combining data sets and filtering to relevant data. During the filtering process, we will first need to identify the missing data and figure out whether we should drop the rows or replace those missing values with a mean, mode, or median value. In the data that has the date of the collection, we need to change the data type of the column to “date and time” so that we can correctly sort it by time and graph the change in the number of bee colonies over time to see the trend and the future of honey bees. Although we will need to filter the data quite a bit, they do not require any reshaping as they are in a well formatted in long-form and simple to work with.

Libraries:

Some of the major libraries that we are likely to use include shiny, ggplot2, and plotly.

Some other libraries that we think we may use include plot3D, leaflet, RGL (3D plot), and Esquisse.

Questions Answered with Statistical Analysis/Machine Learning:

We will be using the linear regression statistical analysis to see if there is a correlation between the pesticides and the decrease in bees colony. Also, we can compare the population of bee colonies with the different amounts and types of pollutants. To do this, finding the R-value is important for telling us how strong the correlations are.

For all the questions, as stated above in the data-wrangling section, we will use statistical analysis to decide which central tendency we should replace the cell with no data, or which row containing NA should be dropped. We will do this by looking at the missing cell and check to see if it is necessary to keep it or not.

Anticipated Challenges:

Major challenges would be to merge the datasets together in order to compare them. These data are from different organization and to compare them together we need to merge them. To merge the data correctly we need a shared common key which in this case will be the “date time” column. Some datasets have specific dates and some data only state the month. Thus, we need to make all the date and times uniform. We also anticipate facing a challenge when creating the interactive and the 3D visualization and how each will play into our entire project and what aspects will be best for interactions or 3D visualizations.

Data Sets:

Annual Colony Loss - <https://data.world/finley/honey-bees-and-apiculture>

Pollution - <https://www.kaggle.com/sogun3/uspollution>

Honey Bees and Neonic Pesticide - <https://www.kaggle.com/kevinzmith/honey-with-neonic-pesticide/home>

**Github Link:** <https://github.com/kcroland/INFO-201-Final-Project>