Assignment 6

Lauren Lui  
Level 6000

# Abstract and Introduction (2%, +½ page)

Avian Influenza, more commonly known as Bird Flu, spread rapidly through commercial chicken farms in the United States in 2022. This outbreak lowered egg inventory, subsequently raising egg and chicken prices across the country. Because of this spike in egg prices, my groupmates in my X-Informatics class and I became curious about what causes commercial bird flu. We decided to create an information system for avian flu and some variables that affect the spread of the disease. Through an article on the 2014-2015 avian influenza outbreak in the United States poultry sector [1], we figured out that wild bird flu cases and migratory patterns are a major predictors of commercial bird flu cases. Once the team explored our literature reviews, we decided that the scope of our research would be avian flu cases, and any other related features for the continental United States of America, and for the year 2022. We decided to focus on this time period and location because of the spike in commercial cases in the US during the year.

In order to satiate my own curiosity, and to help provide insight into how avian flu spreads, I chose to focus this project on the relationship between commercial bird flu cases and wild bird flu cases. Because of historical research on the spread of avian flu between different species of migratory birds [1] , my hypothesis for this analysis was that there is a noticeable correlation between wild and commercial bird flu cases. I also believed that the time of year would also play a significant role in predicting cases. While I wasn’t sure how accurate a model of commercial cases could be with the limited features in this report, I attempted to create an accurate model that could predict the occurrence of a commercial bird flu outbreak in a given area from historical or current wild bird flu cases. In addition to a prediction model, I also performed clustering analysis to see if there are any distinguishable types of cases that could be subjects of future research.

# Data Description and Exploratory Data Analysis (3%, +1 page)

Once I decided on the goal for the research and the scope of the data, I returned to my literature reviews to find the most reliable and accessible source of bird flu data. Although there were many news articles about bird flu, very few of them had actual sources for their data. The few articles that did cite their sources or provide links for more information all pointed to the USDA for commercial cases, as it affects both chicken and egg stock nationwide. Once at the USDA website [2], I quickly found the commercial bird flu dataset for 2022 as both a Comma Separated Value (CSV) file and visualized in a static map of the United States at the state level. The CSV file contained the Outbreak Date, State, County, Facility Type, and Number of birds affected (Figure 2.1). Although there was no metadata provided about how the data was collected, or how the dates were formatted, the USDA website did provide a definition for some phrases used in the visualization.



Figure . USDA Commercial Bird Flu 2022 Data

On the USDA commercial bird flu website, there is a link to a similar site for wild bird flu cases [3]. This site contains an interactive table visualization of the wild avian flu cases and a link to download the dataset. It is interesting to note that although both sites are affiliated with the same organization, their interfaces are vastly different, and this caused some uncertainty for a user navigating between the two. The wild avian flu data was also stored in a CSV file, and contained the Date Detected, State, County, HPAI Strain, Bird Species, Sampling method, and submitting agency (Figure 2.2­­­­­). Unlike the commercial dataset, this one contains more well-documented metadata on the site, including types of bird flu strains and their abbreviations, as well as the shorthand notation for various reporting agencies.

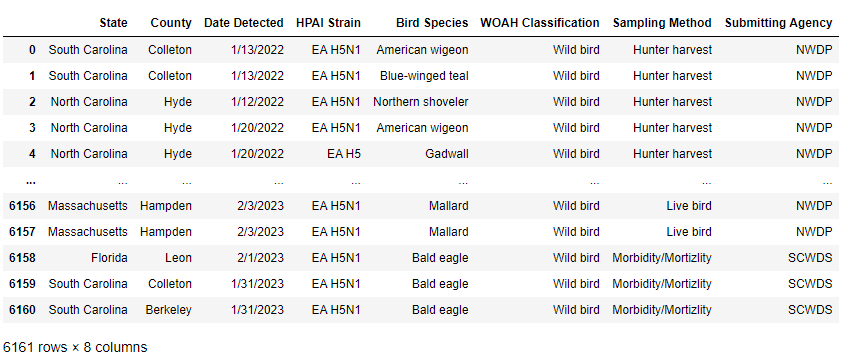


Figure . USDA Wild Bird Flu 2022 Dataset

After my group found the USDA sources, we also wanted see what other potentia data sources were available to us. One other source we found was the World Organization for Animal Health (WOAH) [4], which had avian flu cases from 2005 to current time at the country level. This source had an impressive date range, but was not granular enough or our interests. We also explored the World Animal Health Information System [5], and while the data was well documented and robust, the data was at the state level, and not as granular as we wanted it to be.

# Analysis (5%, +2 pages)

Once I decided on the data sources for this project, the first step was to perform exploratory data analysis on the two data sets. In this initial analysis, I had two goals: to recognize and perform any necessary cleanup and formatting of the data and to understand the distribution of each feature for better model selection. Luckily, the data provided didn’t need any data cleanup. There were no values, and for both categorical values and numerical values there were no invalid values. As for data formatting, I noticed that the dates for each dataset were formatted differently, and needed to reformat them into a unified format for data merging. In addition to this, I knew that I wanted to analyze data at the day and month level, so I chose to format the dates as a Month and Day column. After checking for data cleanup and formatting, I decided to visualize the categorical and numerical features for each dataset separately.

For the commercial dataset, the categorical values are State, County, Month, Day, and Flock Type. The only numerical value is Flock Size. For State, County, Month, and Flock Type, I visualized the distributions in bar charts displaying the number of outbreaks per category, and for Flock Size, I visualized it in a histogram. I chose not to visualize county or day distributions because there are too many categories, and both features are subcategories of other features. The State feature (Figure 3.1) shows a very uneven distribution of case numbers, with a large majority of cases in Minnesota and South Dakota. The Month feature (Figure 3.2) had a bimodal distribution with the two centers in April and October. The Flock Type bar chart (Figure 3.3) shows that most cases are the World Organization of Animal Health (WOAH) Poultry, WOAH Non-Poultry, and Commercial Turkey cases. Flock size (Figure 3.4) is extremely skewed to the right, with the majority of flock sizes under 400, but with outliers as large as 5.3 million. This is explainable by the different types of commercial locations, ranging from local farms to large corporate factories.

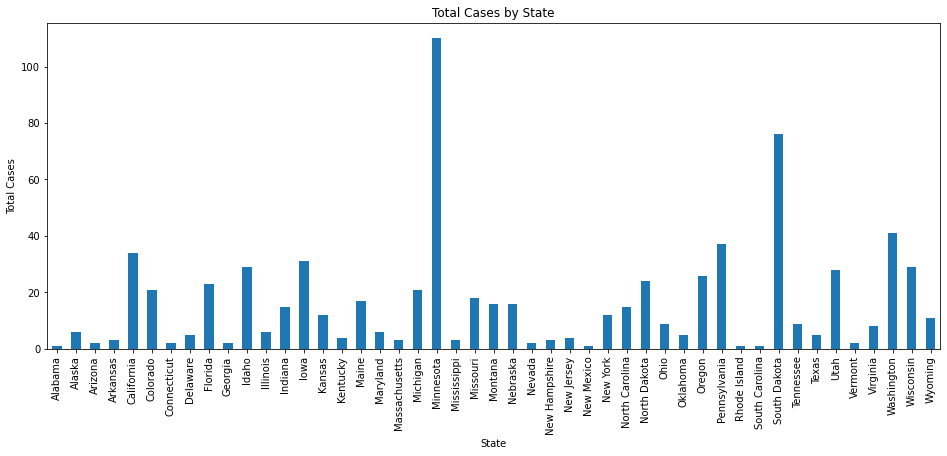


Figure . Commercial Cases by State

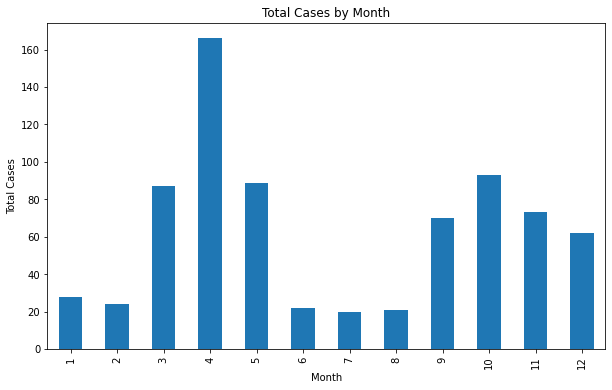


Figure . Commercial Cases by Month

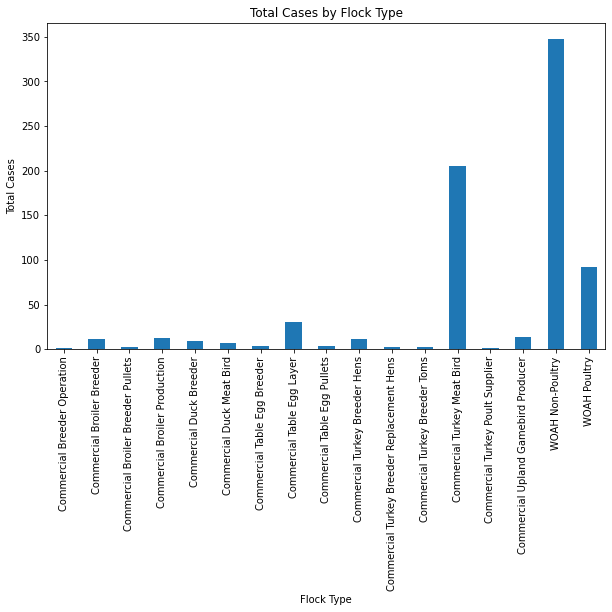


Figure . Commercial Cases by Flock Type

The wild bird flu dataset had an array of only categorical data, including the State, County, Month, Day, HPAI Strain, Bird Species, WOAH Classification, Sampling Method, and Submitting Agency. Similarly to the commercial dataset, I refrained from visualizing the County and Day features. Additionally, I didn’t include Bird Species and Submitting Agency due to the huge number of various classifications in each feature. For the remaining features, I created bar charts to visualize the distribution of cases per feature. The State distribution for wild bird flu (Figure 3.5) had a slightly more even distribution than the commercial dataset, but both showed Minnesota as the state with the most cases by far. The wild case month distribution (Figure 3.6) was also similar to the commercial dataset, with a bimodal distribution for the winter and spring months.

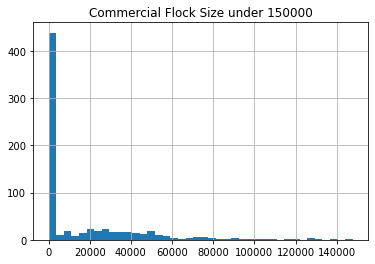


Figure . Commercial Cases by Flock Size

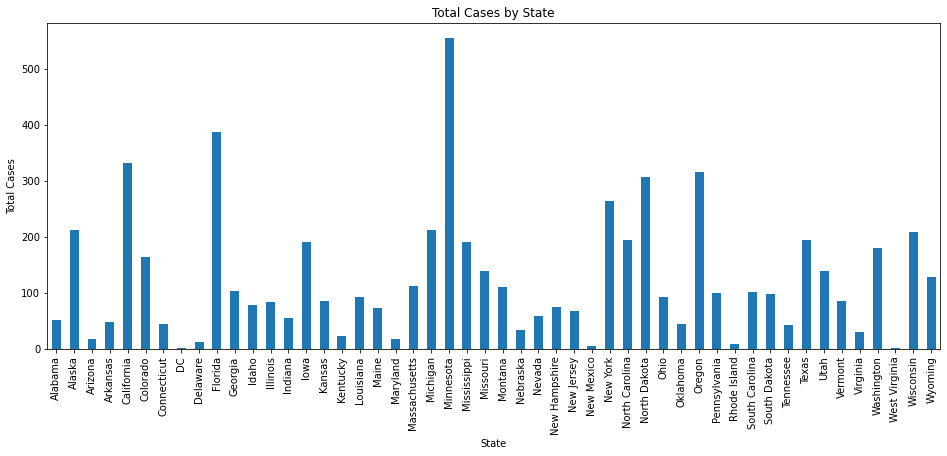


Figure . Wild Cases by State

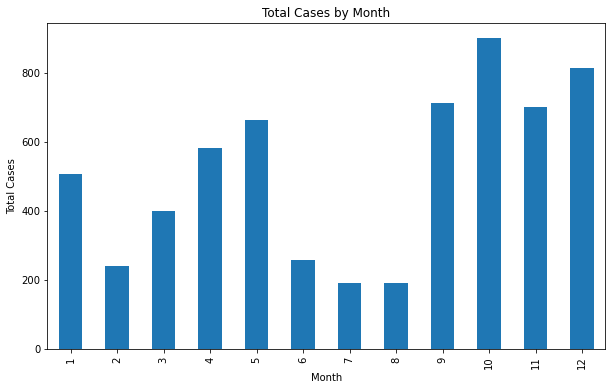


Figure . Wild Cases by Month

Once I completed the EDA on both commercial and wild datasets, I decided to create two merged datasets on the location and time. The first merged dataset would be merged at the day level, counting the total number of wild and commercial outbreaks in a given state and county in a single day. The next dataset was merged on the month level, with the same aggregated features selected. The reason for these two different levels of merging lies in the exploratory data analysis done earlier, which showed that there were not too many counties that had both commercial and wild cases recorded on the same date. This could be in part because of the methods of sampling, the lack of commercial poultry farms in every county, or a variety of other reasons. Each merged dataset contained the following features: State, County, Month, and possibly day. These are the features that each separate dataset would be merged on. The merged dataset would also contain the following aggregated values: Commercial Outbreak Count, Commercial Flock Size Mean, Commercial Flock Size Total, Is Commercial Outbreak, Wild Outbreak Count, and Is Wild Outbreak. Each of these features is numerical and would be used to build any models in the future.

# Model Development and Application of Models (12%, 6+ pages)

My goal when modeling the data was to utilize two different classification models and two clustering models. The classification models would serve as a measure of how well we can predict the existence of a commercial outbreak in given a state, county, month, and the existence of wild cases in that time. I this level of data merging because when I joined the data on the date level, there were only 19 days with both wild and commercial outbreaks. The two classification models used were KNN Classification and Random Forest Classification. I used KNN as a base level classifier to understand how well a simple algorithm could model the data. Random forest classification was chosen for two reasons: it is a more widely accurate model and it provides feature importance that could be used in the future for dimensionality reduction. As for the clustering models, my goal was to gain insight to the types of cases. I wanted to see if there were clear clusters that could be the subject of future research, or if there weren’t with the features I provided. The two models that I chose for this were K-means clustering and birch clustering. Similarly to the classification models, I wanted a baseline with K-means clustering and a spatial clustering algorithm, and was curious as to what the differences would be and if any insight could be gained from this.

To prepare the data for the classification models, I chose to limit the features to only the important ones and represent the categorical values as dummy variables via pandas’ get\_dummies function. In this way, I could represent each feature numerically for the sake of the algorithm. In addition to this, I chose to train my models on a standard 80-20 split, with 80% of the data used to train the model and the remaining 20% used to test the accuracy of the model. As I wasn’t sure which features were the most important, I chose to run each model twice at two different levels: One with states, month, and the existence of a wild outbreak, and one without the states. The models without the states ran significantly faster, and I was curious if the faster runtime would be worth the loss in accuracy.

KNN Classification provided decent initial results for datasets with and without states. The first run through, including states, had an accuracy score of 0.8417. This model was very good at correctly predicting if a month would have no commercial outbreaks (Figure 4.1), but was not as good at correctly predicting when a commercial outbreak would take place, only correctly guessing around 75% as opposed to the 90% accuracy of the true negative values. This is because, more often than not, there aren’t any commercial outbreaks, and the model is not sophisticated enough to deal with that. The KNN model without states had a similar performance, with an accuracy score of 0.7884. This model (Figure 4.2) is worse, however, at predicting both when an outbreak would happen and when it would not. In this case, the loss in about 8% accuracy is not worth using a slightly faster model, so it would be better to use the model with states.

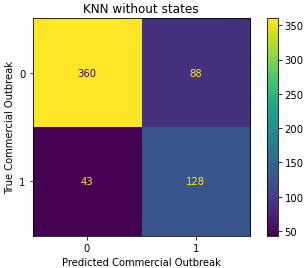


Figure . KNN Clustering Confusion Matrix without States

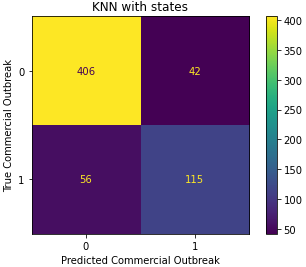


Figure . KNN Clustering Confusion Matrix with States

Random Forest classification did a much better job at modeling the data than KNN did. The first model (Figure 4.3), including states, had a significantly higher accuracy score than either KNN models, at 0.8918. It did, however, run into similar issues of over predicting no outbreaks as the other models. In addition to the accuracy score, the model provided feature importances, indicating how important each feature was in building the predictive model. Figure 4.4 shows the top four feature importances for this model, and we can see that it has the existence of a wild outbreak first, followed shortly by month and then some states are there as well. This is a good thing, and shows us that there is indeed some correlation between wild and commercial bird flu cases. In our second model, which did not include states, we decided to include dummy variables for each month as well, to see which months are the most important indicators of a commercial case. This model performed slightly worse than the random forest model with states, with an accuracy score of 0.8821. This score is still greater than that of the KNN models, but if we look at the confusion matrix (Figure 4.5), we can see that the model is worse than all models at predicting true commercial outbreaks. This is an indicator that the model should in fact use states in the future. The feature importance graph (Figure 4.6) shows us that the most important feature is yet again the wild outbreak, with each individual month being less important. This might be a sign in the future to avoid using dummy variables to represent the months.

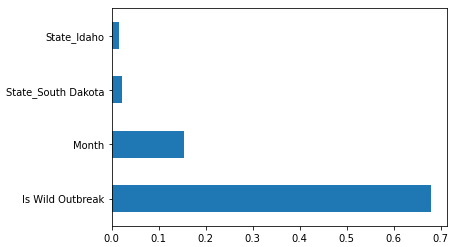


Figure . Random Forest Classifier Feature Importance with States

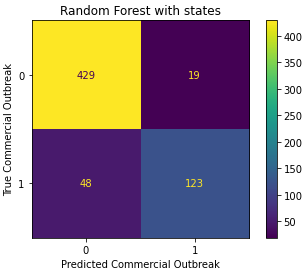


Figure . Random Forest Classification Confusion Matrix with States

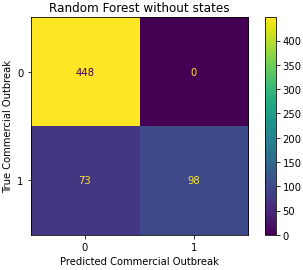


Figure . Random Forest Classification Confusion Matrix without States

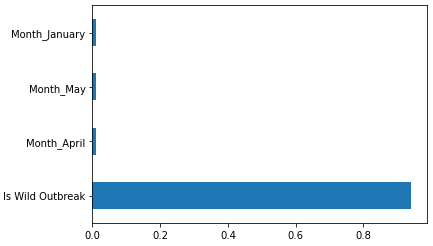


Figure . Random Forest Classification Feature Importance without States

For clustering, I decided to run the data at two levels: with just numeric data, and with categorical data as well. For the numeric data, the features used were: Commercial Outbreak Count, Total Commercial Cases, Wild Outbreak Count, and Month. For the categorical and numeric data, the numerical features were the same as the previous model, with the addition of both Month and State represented by dummy variables. For each model, I performed the elbow method to find the optimal number of clusters to use, and for each model on each dataset, the optimal number of clusters was three. Additionally, for each clustering model, I chose to visualize these clusters by finding the first two principle components from Principle Component Analysis (PCA).

Kmeans Clustering did not show any distinct features for either level of analysis. Both levels of analysis took between 8-10 seconds to run. For the Clustering with just numeric features (Figure 4.7), there are no clear clusters that are discernable from the naked eye. The principle components have some discernable relationship, but there’s not much to learn from the features. One thing to note is that there are some points in the database that seem to be outliers, and those points, in green, might be objects of interest for future study. Similarly to the numeric only clustering, the numeric and categorical principle components (Figure 4.8) don’t give very much insight to the data. It is interesting to note, however, that the three clusters are all same size between both models, so it might be of interest in the future to take note of the smaller clusters for each model and see what makes those clusters different from the large red cluster.

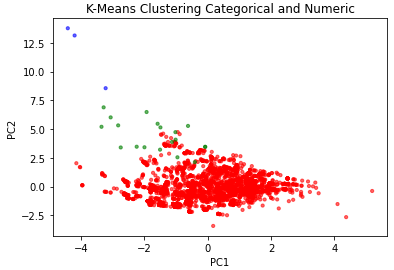


Figure . K-Means Clustering Categorical and Numeric Features

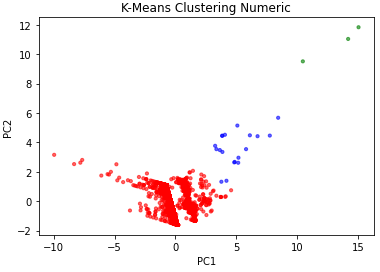


Figure . K-Means Clustering Numeric Features

Birch clustering, interestingly, provided results very similar to those of K-Means clustering. Unlike my hypothesis that one of the two models would fit the data better, both Birch and K-means provided almost identical results. The largest cluster for all four models included the same 3072 points, and the only difference between the birch and k-means algorithms was that the smaller two clusters had one cluster switch between them. This highlights the fact that there is a noticeable difference in these points, and that they are important to look at to gain further insight as to how this disease behaves in the United States.

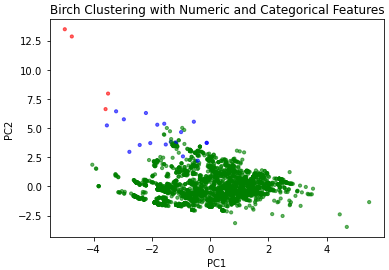


Figure . Birch Clustering with Numeric and Categorical Features

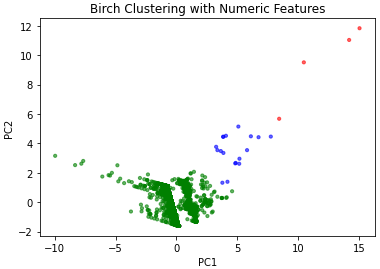


Figure . Birch Clustering with Numeric Features

# Conclusions and Discussion (3%, +1 page)

The analysis in this report aims to answer the following questions: How well can we use location, month, and concurrent wild bird flu cases to predict commercial bird flu cases? What kind of classification models best model this dataset? What amount of information is necessary to accurately predict commercial bird flu cases? What information can we gain from analyzing clustering models of the data? Is there a better type of clustering for this dataset?

We can utilize the table below (Table 5.1) to answer the first few questions on classification models. The table shows each model, dataset, and three measures of model performance for each model. Accuracy score is a common, general measure of model performance. Precision and Recall are more specific, and can give more insight to how many of your predicted commercial outbreaks are correct and how many of the true commercial outbreaks you predict respectively. We can look at the accuracy score to answer the first and second questions. The highest accuracy score is 0.8918, meaning that 89.18% of the time, we will correctly predict whether or not there will be a commercial outbreak in a county in a certain month. We can also see that overall, the Random Forest models had higher accuracy scores than the KNN models, showing that Random Forest fits this dataset better. We can use the accuracy score in conjunction with the precision and recall to show that for both KNN and Random forest, the models that included location performed significantly better than those without.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Dataset | Accuracy Score | Precision | Recall |
| KNN Classification | 80-20 Split Merged Data with States | 0.8417 | 0.7325 | 0.6725 |
| KNN Classification | 80-20 Split Merged Data without States | 0.7884 | 0.5926 | 0.7485 |
| Random Forest Classification | 80-20 Split Merged Data with States | 0.8918 | 0.8662 | 0.7193 |
| Random Forest Classification | 80-20 Split Merged Data without States | 0.8821 | 1 | 0.5731 |

Table 5.1 Classification Accuracy Table

For the clustering models, we did learn some important information on a handful of cases that we should focus on for future research. For any case that isn’t in the largest cluster, we can see what is different about these cases and potentially learn what makes these data points different. If these points represent a location with many cases, it might be interesting to look into why. The second clustering question can be answered by the similarity between the clusters found by all four models. Since there wasn’t any significant difference between the models, we can conclude that, at least between K-means and birch clustering, there is no significantly more insightful model.

As I went through the project, I felt like I gained more insight as to how the datasets were related, and learned more about what I would like to do in the future with the information I learned in this process. As for the classification models, I would be interested to see how much accuracy could be improved by including precipitation and climate data, as many literature reviews stated that these features could be important to the transmission of the disease. I wasn’t expecting the time, location, and wild bird flu case data to have such a high accuracy score, so I would be curious to see how granular we could get with our predictions with more information. For the clustering, I would like to do more research into what makes the cases in the smaller clusters special, and if we could learn anything from these cases to prevent the disease’s spread in the future. In conclusion, I’m happy with my results, but am interested to see what other types of analysis can be done with these datasets in the future.

# References

|  |  |
| --- | --- |
| [1] | M. M. a. A. M. Sean Ramos, "Impacts of the 2014-2015 Highly Pathogenic Avian Influenza Outbreak on the U.S. Poultry Sector," *Economic Research Service,* p. 20, 2017. |
| [2] | USDA, "2022-2023 Confirmations of Highly Pathogenic Avian Influenza in Commercial and Backyard Flocks," USDA, 18 January 2023. [Online]. Available: https://www.aphis.usda.gov/aphis/ourfocus/animalhealth/animal-disease-information/avian/avian-influenza/hpai-2022/2022-hpai-commercial-backyard-flocks. [Accessed 3 February 2023]. |
| [3] | USDA, "2022-2023 Detections of Highly Pathogenic Avian Influenza in Wild Birds," USDA, 12 April 2023. [Online]. Available: https://www.aphis.usda.gov/aphis/ourfocus/animalhealth/animal-disease-information/avian/avian-influenza/hpai-2022/2022-hpai-wild-birds. [Accessed 3 February 2023]. |
| [4] | WOAH, "Events management," 2023. [Online]. Available: https://wahis.woah.org/#/event-management?viewAll=true. [Accessed 10 February 2023]. |
| [5] | Food and Agricultural Organization of the United Nations, "Diseases," 2023. [Online]. Available: https://empres-i.apps.fao.org/diseases. [Accessed 8 February 2023]. |