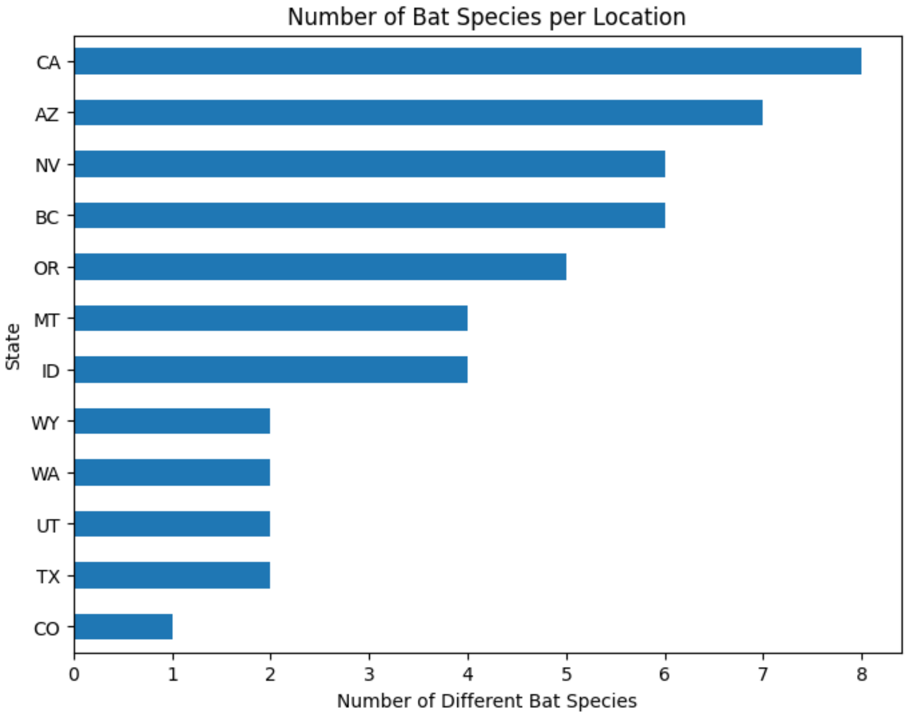
Monthly Progress Summary *(September)*

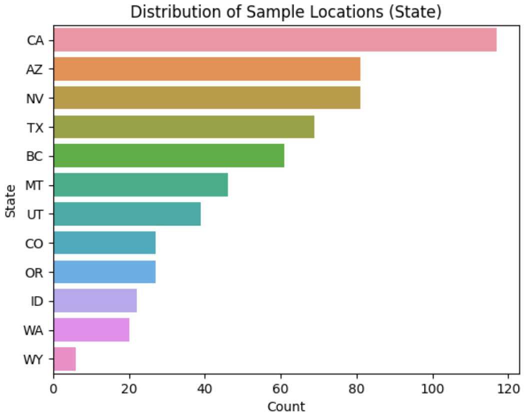
I. Data Understanding Summary

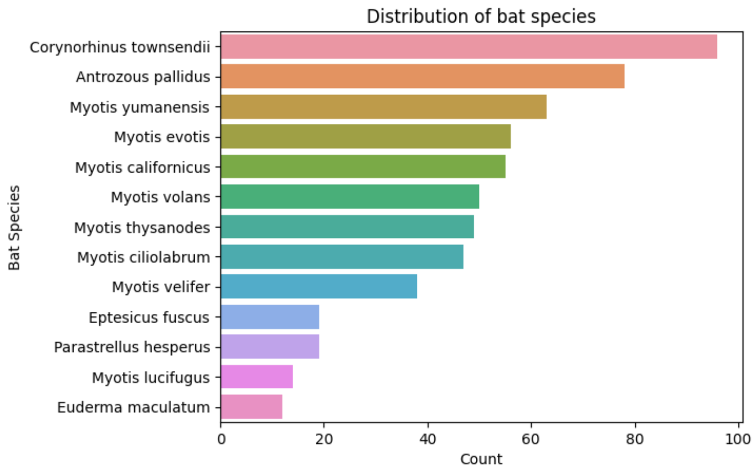
Question:
Please provide a summary of your team's "Data Understanding" accomplishments during the month of September. Remember to include any relevant links to your work (e.g., a Python notebook showcasing your team's Exploratory Data Analysis work – e.g., statistical analysis and visualizations related to key variables, patterns, relationships, data quality issues).

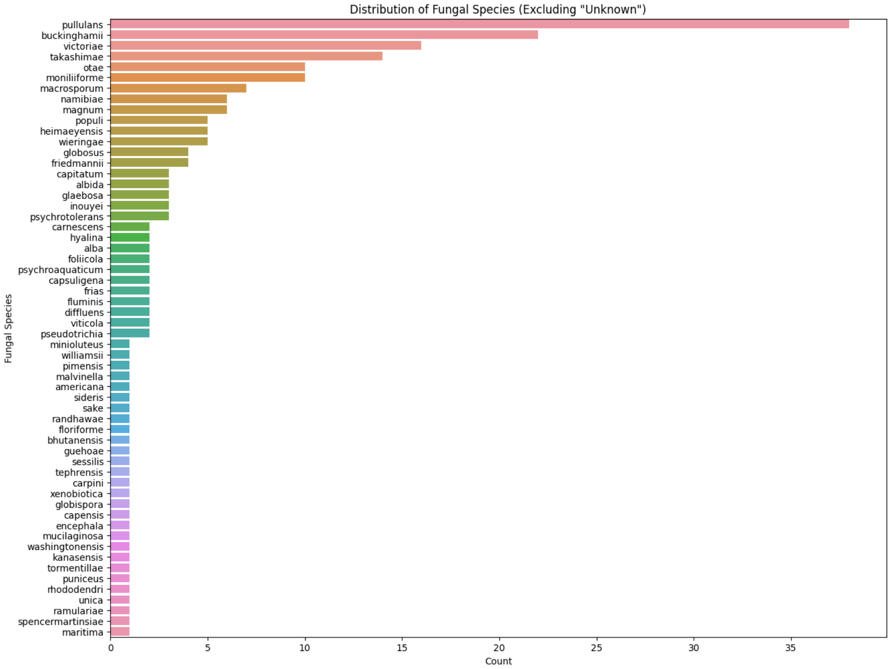
Most of the work this month was dedicated to locating usable data sources. A lot of our time was spent evaluating datasets for their usefulness in machine learning tasks. We learned a lot about how to properly pick a dataset for a particular task, such as the features and labeled data that must be present in a dataset, and how to identify common elements that can be used to concatenate datasets. A challenge was finding high-quality publicly available audio data, and filter that from the audio data results. We also encountered many setbacks in attempting to find publically available datasets.

During the data analysis phase, we looked at different fungal species present on the skins of bats. This preprocessing work can be found here. With this dataset, we looked at the distribution of different species of fungi across the bats, the distribution of the variations in bat species within the dataset, and the variation among locations. This revealed that California has the largest presence in this particular dataset. We were also able to determine the most prevalent fungal species, and the most prevalent bat species in the data.


Number of Bat Species per Location
State
CA
AZ
NV
BC
OR
MT
ID
WY
WA
UT
TX
CO
0
1
2
3
4
5
6
7
8
Number of Different Bat Species

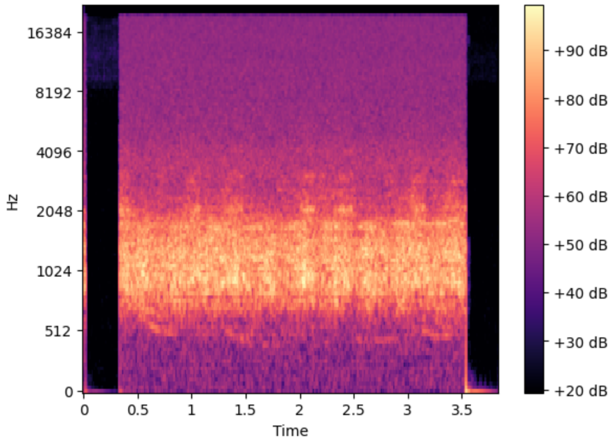

Distribution of Sample Locations (State)
State
CA
AZ
NV
TX
BC
MT
UT
CO
OR
ID
WA
WY
0
20
40
60
80
100
120
Count





We also discovered how to process audio data in python, the work showcased here. The audio data is quite noisy, so significant noise reduction techniques had to be employed in order to be able to see the bat calls. A lot of time and effort was spent into researching the most current techniques and packages to achieve this. The audio data was then put into visual sonographs, to later be put into a CNN to identify the bat calls.

```
<matplotlib.colorbar.Colorbar at 0x7f1756714ac0>
```



II. Data Preparation Summary

Question:

Please provide a summary of your team's "Data Preparation" accomplishments during the month of September. Remember to include any relevant links to your work (e.g., a Python notebook showcasing your team's data preprocessing work – e.g., data cleaning, formatting, missing value imputation, outlier handling, feature engineering).

We have made a bit of progress in Data Preparation, but we are not done yet. We learned how to analyze audio data in Python (more on that in part I); however, we have encountered issues in acquiring acoustical data to use as our Challenge Advisor said; because of this, we are considering not using acoustical data to train on. Other than that, we have directly performed data cleaning tasks on one dataset such as filling in unknown values and found value counts for certain features to get a sense of the frequency and uniqueness of each feature. We have made plots showing the distribution of species, for example, and grouped certain features together by a category. See our work here; other datasets will be cleaned shortly before Saturday.

III. Lessons Learned and Challenges

Question:

Reflecting on the Data Understanding and Data Preparation phases, what were the key insights or challenges your team encountered? How did you address them? Share any important lessons learned that can help guide future steps in the project.

The key challenge our team encountered was locating the datasets. We were able to find a lot of data, but none of it was the audio data we were looking for. We also struggled with being able to find bat data within the east coast region. In general, most of the data we found was very specific to the research institution's research purpose so it was not as relevant to our project problem. Through data understanding, we were better able to assess whether the datasets were a good fit and through data preparation, we came across missing data that we might not have caught without that deep analysis on our dataset's metric. With these findings, we were able to communicate back and forth with our AI Studio TA and Challenge Advisor to refine our project scope according to available resources.

IV. Next Steps (Data Understanding and Prep)

Question:
Given your current progress, what additional tasks does your team need to complete in connection with the Data Understanding and Data Preparation phases of your project? What is your plan to complete these tasks?

Based on guidance received from our Challenge Advisor and AI Studio TA, we plan to conduct more in-depth analysis to more datasets that require further investigation or preprocessing steps. Since we are still waiting for possible acoustic datasets from the authority, we are making two preparations: 1) using several datasets we found as training data to predict factors that might cause WNS 2) learning how to process acoustic data if our Challenge Advisor acquires the acoustic datasets from the authority. We plan to continue performing EDA on several more datasets to see if they provide helpful insights to our problem, and combine them together for our model. We will also proactively ask for acoustic datasets so that we can settle down on the dataset choice by the end of the next meeting with our Challenge Advisor.

Additionally, we will finalize the feature selection process after the datasets are decided. This will ensure we retain the most relevant features for our predictive modeling. By completing these tasks, we will be able to enhance the quality and relevance of our data for subsequent stages of the project.

V. Request for AI Studio TA Support

Question(s):

What additional support do you need from your AI Studio TA? Please structure your response as specific questions, related to the Data Understanding and Data Preparation phases of your project. Consider areas where you may require specific guidance, clarifications, suggested approaches, or suggested resources. Your AI Studio TA will review these questions and work through them with you in an upcoming meeting or chat.

1. Given our failure to find usable acoustic data these past weeks, is it possible for us to proceed using non-acoustic data?
2. We have found acoustic data from other countries(Canada and England); is this usable or beneficial to our project to any extent?
3. Given where we are in our current progress(still without our final dataset), what's your outlook on our progress?
4. Is there anything we should keep in mind as we approach later stages of the project?
5. Once we obtain our final dataset, can you provide guidance on advanced statistical techniques or exploratory analysis methods that can help uncover more complex relationships/patterns in our dataset?

Underlying question: What features best predict if a colony will develop WNS?

###