Bird Analysis Big Data Report
CS167 - Group 25

Student Information:
Andrew Krikorian - akrik001 - 862373081
Nasser Ben Yedder - nben002 - 862378110
Zhenchao He - zhe086 - 862372355
Thomas Merritt - tmerr004- 862357156

Tasks:
1: Nasser Ben Yedder
2: Andrew Krikorian
3: Zhenchao He
4: Thomas Merritt

Introduction:

This project analyzes a dataset that represents bird observations all over the world. We chose the eBird_10k file as our final dataset submission. Our project consists of Spark and Beast integrations. The goal of the project is to make a parquet file in the first task and then pass this file onto tasks two, three, and four. In task two, we create a shapefile with all of the ratios of a specific bird (Mallard) joined with the zip code dataset. In task three, we selected a date range of 2/21/2015 to 5/22/2015 and plotted the data on a pie chart. In task four, we create a prediction model that can predict a bird category depending on a sighting. We used QGIS to visualize the shapefile produced in task two and google sheets to visualize the pie chart in task three.

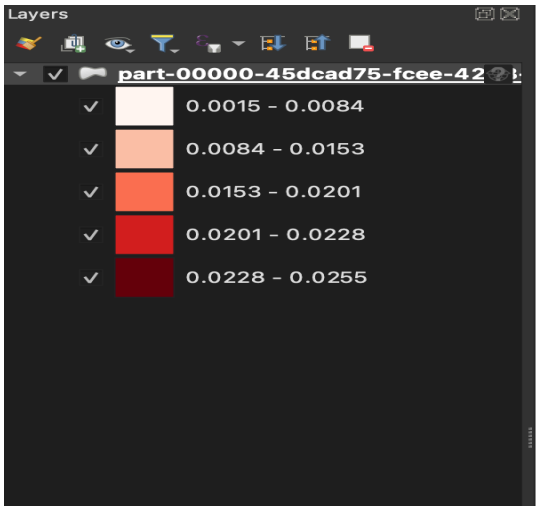
Task 1:

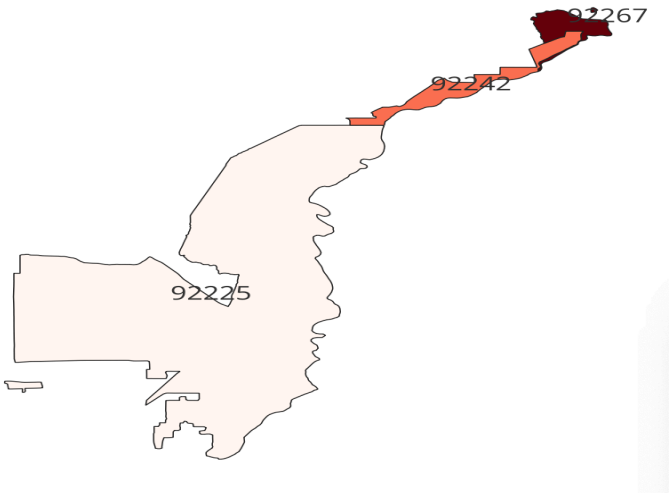
The Parquet format's efficient storage and processing abilities make it ideal for our project, enabling efficient storage and retrieval for large datasets like the bird observations we're handling. By converting our data to Parquet format, we can significantly reduce storage requirements while maintaining data integrity and allowing for faster processing. This is particularly advantageous for analytical tasks, such as those in our project, where we need to efficiently handle and analyze large volumes of data. As illustrated in the table below, Parquet files are substantially smaller compared to the original CSV files, both in zipped and uncompressed forms. This reduction in size translates to faster data access and lower storage costs, enhancing performance and allows for seamless integration with existing tools for visualization and analysis of bird observations for our project.

Dataset	CSV size (zipped)	CSV size (unzipped)	Parquet size
1k	42 KB	538 KB	22 KB
10k	258 KB	4 MB	144 KB
100k	2.5 MB	40.8 MB	1.1 MB

Task 2:

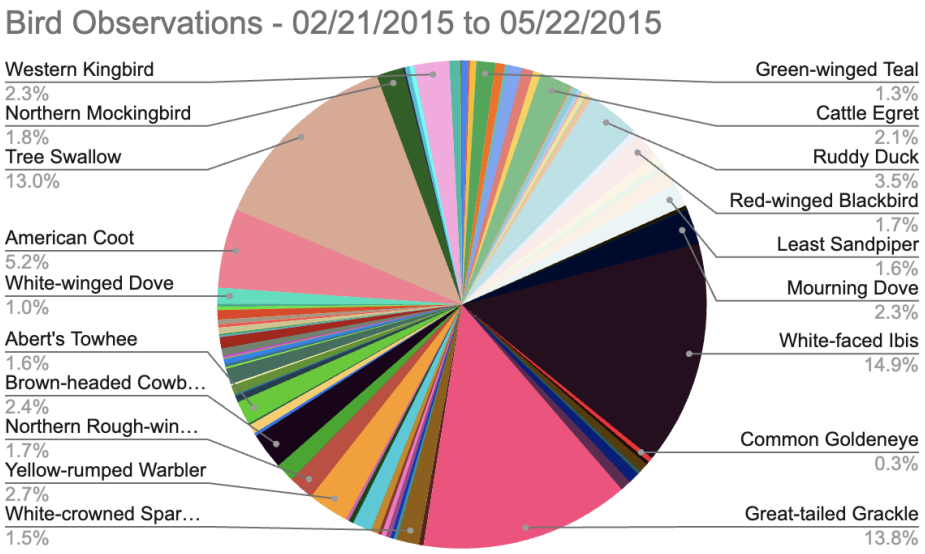
The purpose of task two was to get the ratio of a specified bird species in a specific zip code. I used Beast for the spatial data processing and QGIS to build the choropleth maps. I have attached three choropleth visualizations below pulled from the 10k file. The visualization below is of the Mallard species from the eBird_10k zip file.





Task 3:

Here are the results (pie chart) from the eBird 10K file for the date range of 2/21/2025 to 5/22/2025.



Task 4:

Here is the results (total time, accuracy, precision, and recall) for the 100,000 point dataset:

```
Total time: 4.436860125 seconds
Accuracy: 0.9710144927536232
Precision: 0.9826640635865301
Recall: 0.9710144927536232
```