Table of Contents

CS167 Final Project - Bird Analysis

Group 13

Task Distribution:

- Task 1: Yishao Wang
- Task 2: Daniel Vonrhein
- Task 3: Scott Vo

Introduction

The bird analysis project utilized a crowd-sourced bird observation dataset provided by Cornell Lab of Ornithology. Utilizing this dataset we were able to practice various data-cleaning techniques, different forms of data manipulation, and apply visual representations to the data we created. The data framework we settled on using is BEAST, as it provides a different combination of tools used for data-cleaning (SQL), and dealing with geo-spatial data. SQL is also useful here to select specific data for spatial and temporal visualization.

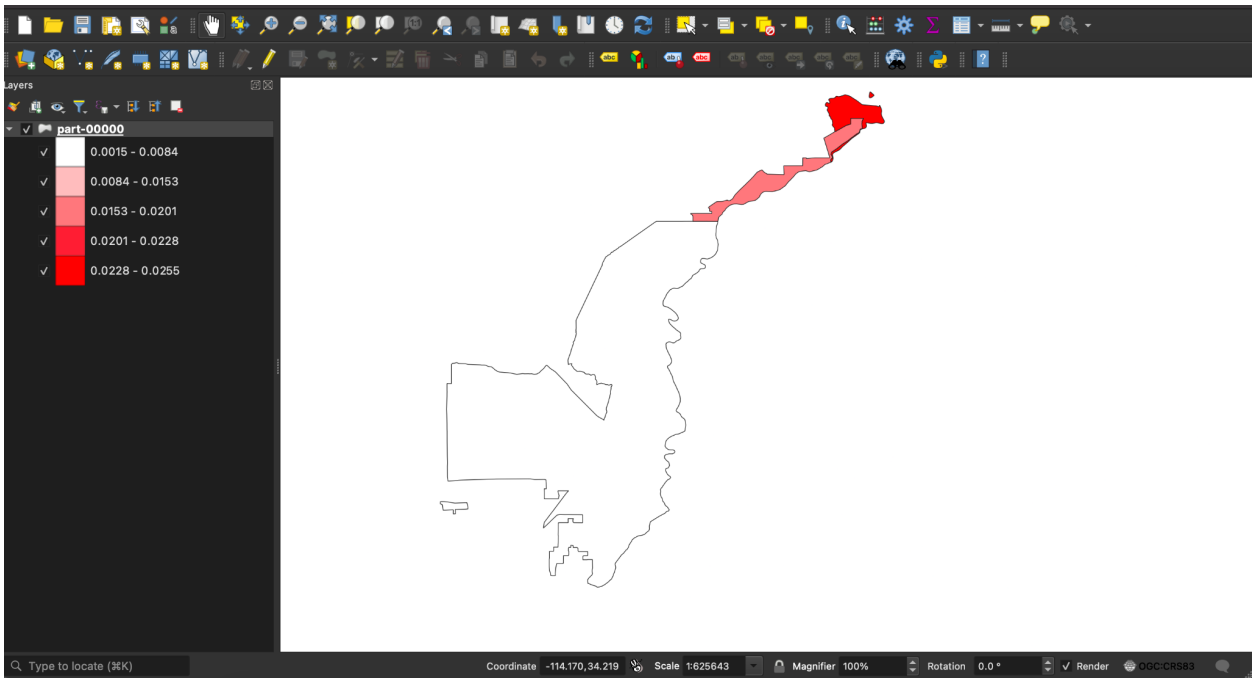
Task 1:

The parquet data format is very useful for this project as the dataset have numerous columns with repeating data value (i.e. Zip Codes, Species, Names, etc.), and can be made smaller utilizing the compression the parquet data format provides, as you can see in the table below the level of compression that can be achieved is significant. In actual big data, this level of compression is most likely needed to be efficient.

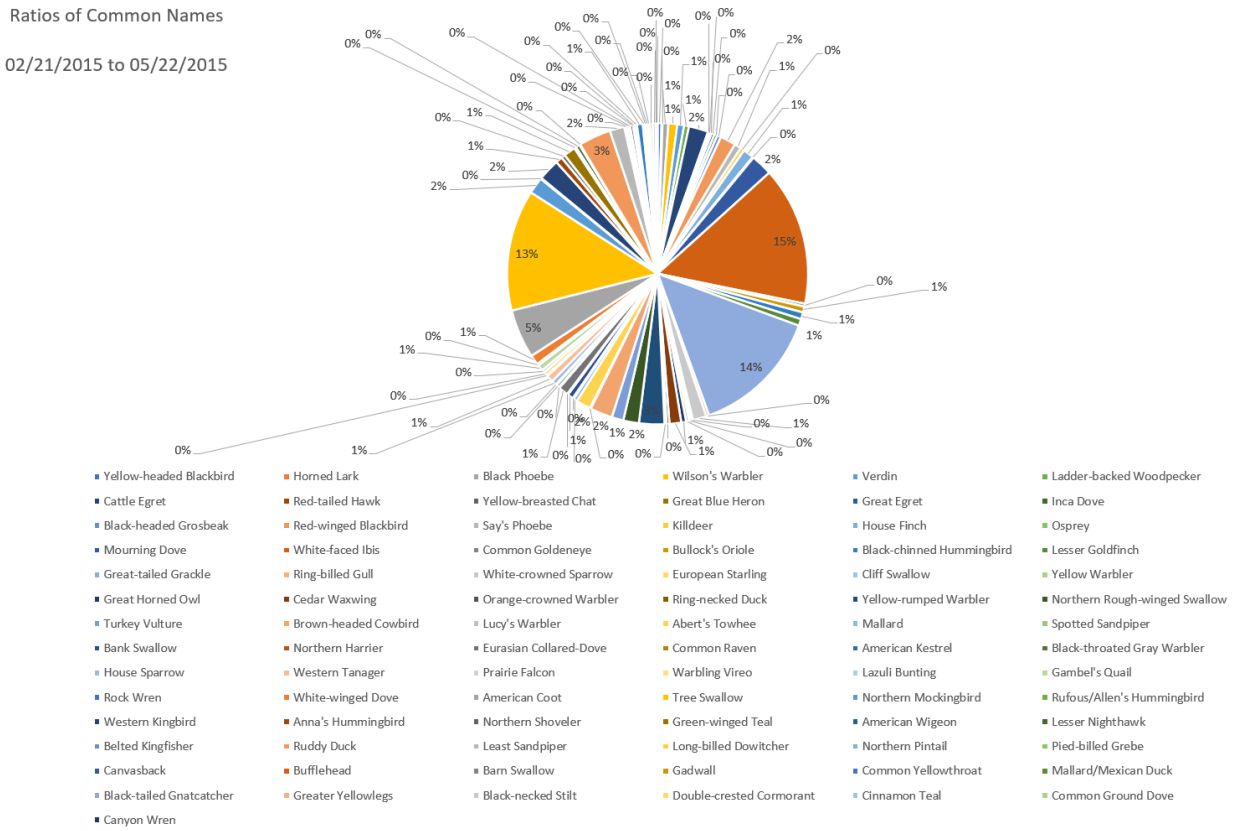
Dataset	CSV Size (KB)	Parquet Size (KB)
1K	526	33.3
10K	3,937	165
100K	39,853	1,475

Task 2:

Running Part2 on the 10k dataset parquet file, with the selected species as “Mallard” in the second argument, we get the following choropleth map for the ratio of Mallard observations over total observations for different ZipCodes.



Task 3:



From running the temporal visualization on the 10k dataset from Part 1, above is the output from 21 February 2015 to 22 May 2015. This was done with an SQL query to sum and group the observation counts by Common Name.