Bye Bye Birdie

Project Report

Aastha Jha
(301555582)

Owais Hetavkar
(301558712)

Roodra Kanwar
(301477001)

INTRODUCTION

The fluctuating weather patterns along with the rise in temperature has brought about adverse changes in the natural terrains of birds present around the globe. This shortfall has particularly affected British Columbia bird sightings as well. It is reported that BC has lost around 20 percent of its habitat over the past three generations which is quite steep. In this project, we mainly aim at finding how weather has affected these sightings and which species have become endangered across British Columbia, Canada.

METHODOLOGY

1.) Data Collection

The data for the project is collected from two sources. These are:
1) Global Historical Climatology Network (GHCN) Dataset.
2) Cornell's Ebird API.

Dataset Description

The Ebird dataset was constructed by making calls to Cornell's Ebird API in the form of JSON format which consists of Species code, Common Name, Scientific name, Latitude, Longitude, Observation date, Count being the important keys.
The GHCN dataset has two files - stations and weather observations. The stations.txt consists of Station ID, Latitude & Longitude of station and Station name. Weather is a zipped file consisting of CSVs with the Station ID, Date, Observation, Value being the important columns for our project.

2.) Tech Stack & Cluster Preparation

Two technologies from Google Cloud Platform have been used extensively for this project:

- Cloud Storage: A cloud storage bucket is used to hold all datasets as well as intermediate files created while preprocessing them. We are able to use python scripts to facilitate direct file transfer between the SFU cluster and Cloud Storage Bucket to move large datasets.
- DataProc Compute Cluster: A private Hadoop cluster on GCP with the following specifications:
  All machines running - Debian 10, Hadoop 2.10, Spark 2.4
  Standard (1 master, 4 workers) Config
  Master: 16vCPUs (Intel 5th Gen), 128GB RAM, 100GB Storage
  Workers x 4: 2vCPUs (AMD EPYC), 16GB RAM, 60GB Storage
  Anaconda & Jupyter Notebook Support.

3.) Data Preparation and Preprocessing

a.) Ebird

- Generating the complete date wise Ebird data is possible only by making separate calls for each date in that period. For us to have a complete dataset from 1959-2021, we made nearly 23,000 requests to the API. To do this efficiently we made use of python's multiprocessing package to parallelize the work. This led to the creation of separate json for each date which could then be combined to get year wise jsons for bird records.

b.) GHCN

Stations - Reading the stations text file and formatting it with a regex for the rows value. Filtering the entire dataset for only British Columbia weather stations giving a dataset of all weather stations in BC and their coordinates.

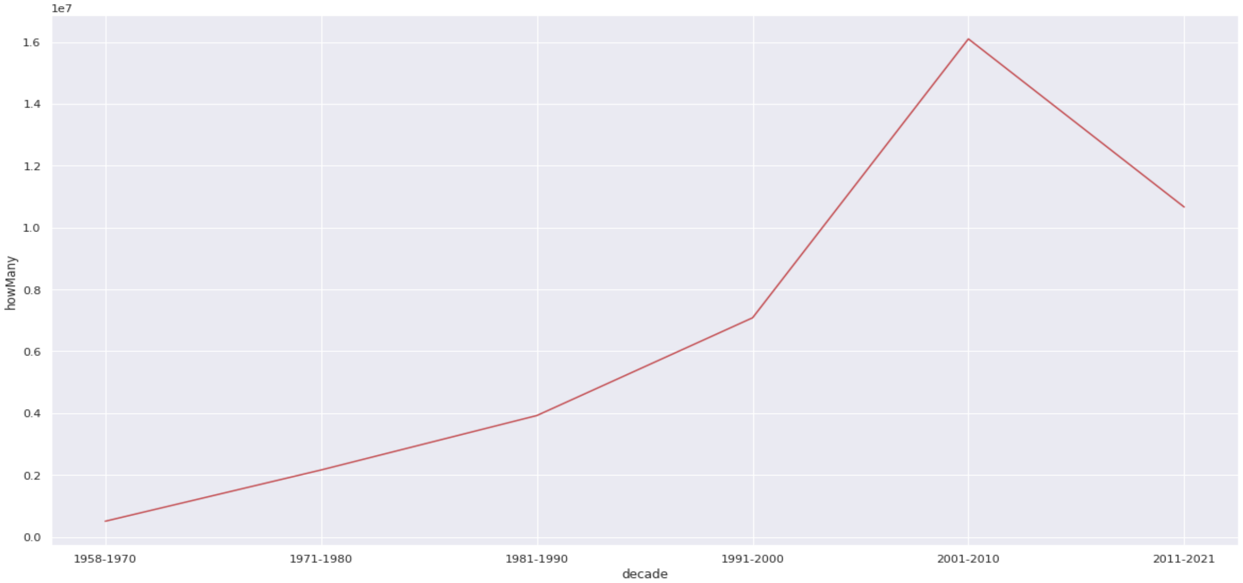
Weather - Reading the file by defined schema and filtering for null values, British Columbia weather stations, picking year count from 1959 onwards. Pivoting the weather data so that for each date and station, we have one row corresponding to minimum temperature,maximum temperature, snowfall and precipitation (as columns).

- Storing the dataframes as CSV in GCS Bucket.
- Joining Stations and Weather (using Stations as Broadcast) based on the Stations ID to get comprehensive details about the latitude, longitude, Station ID, Observation and Value.

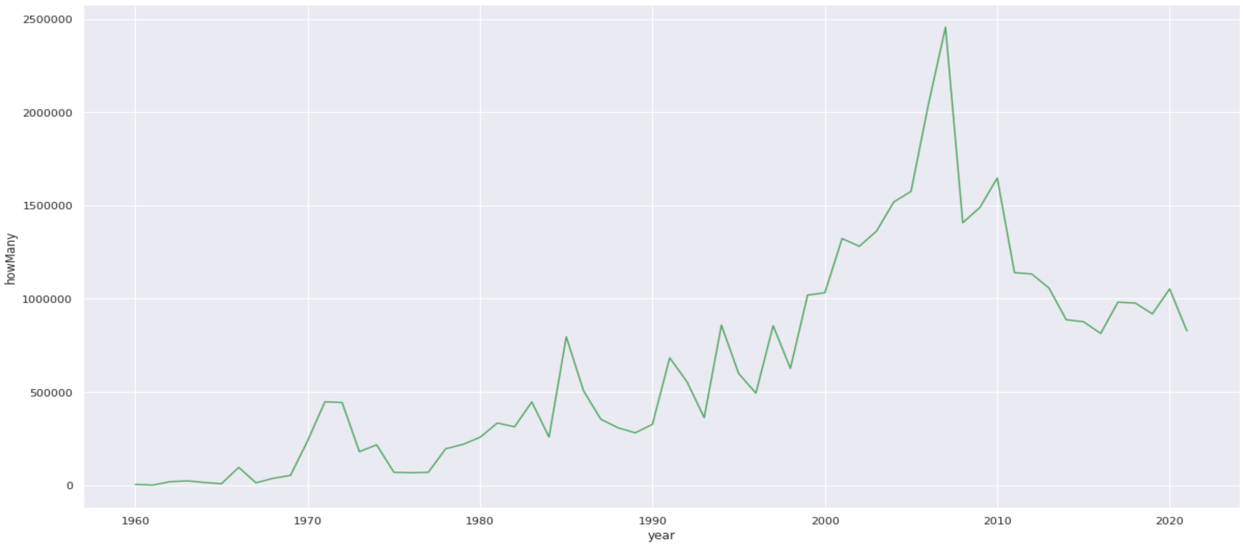
c.) GHCN & Ebird Join

- To join both of these we iterate on the year using pySpark through both the data sets.
- We cross join both the datasets while making sure the date is equal.
- We calculate the haversine distance between the weather station and the ebird data point.
- We aggregate it on this distance to find the station closest to the ebird data point.
- This allows us to retain only the entry with the closest weather data and get rid of the extra entries created due to cross join.

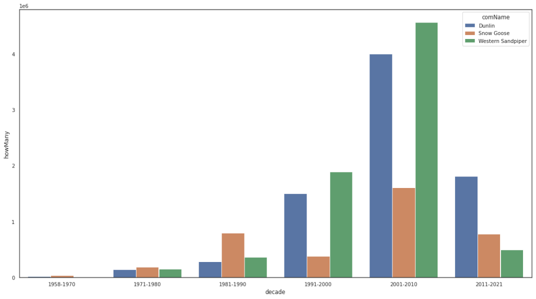
4.) Data Visualization & Analysis



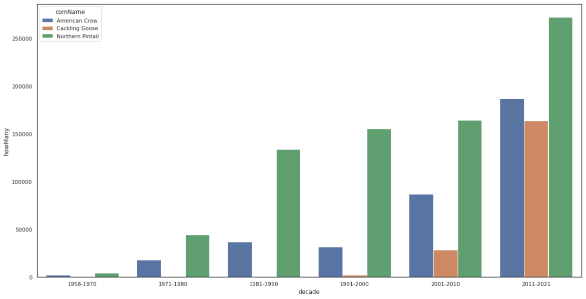
Decade Wise Count of Birds



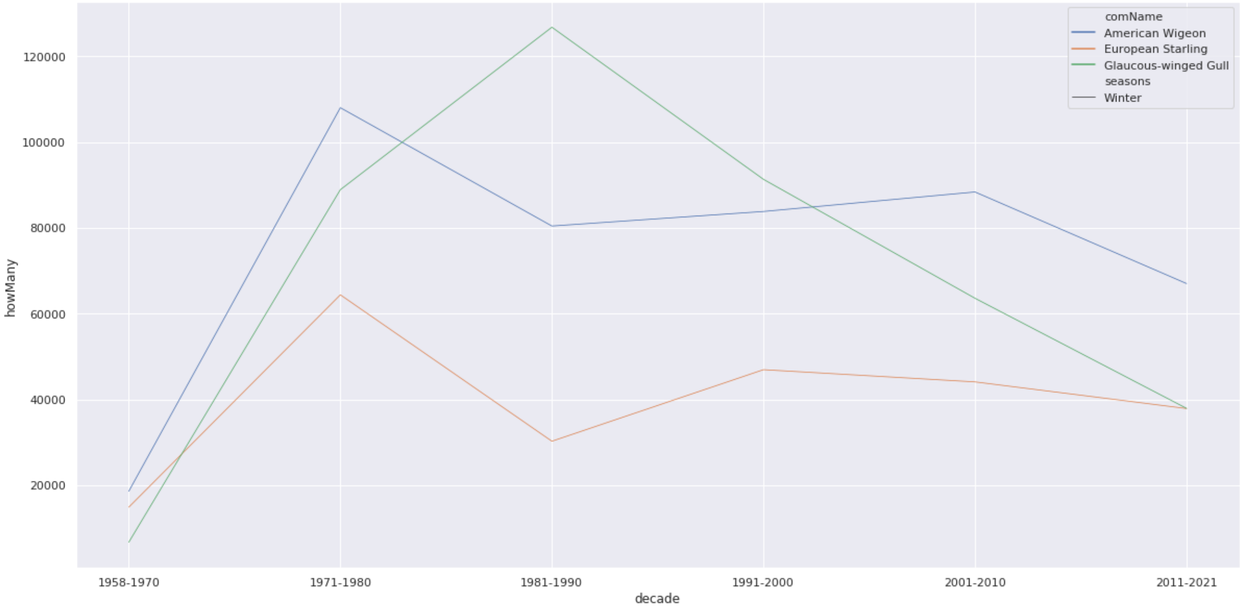
Year Wise Count of birds



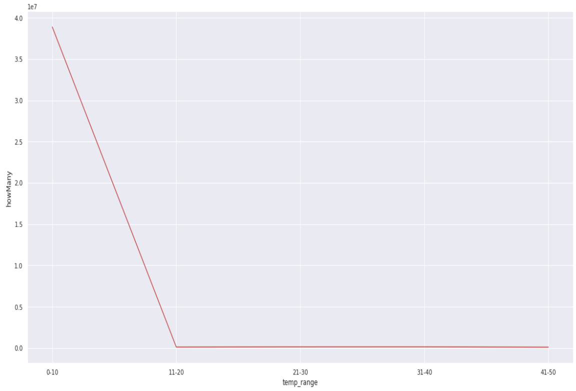
Endangered birds over the decades



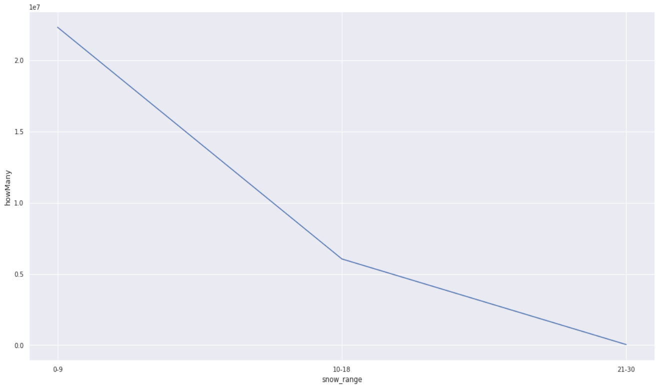
Increasing birds over the decades



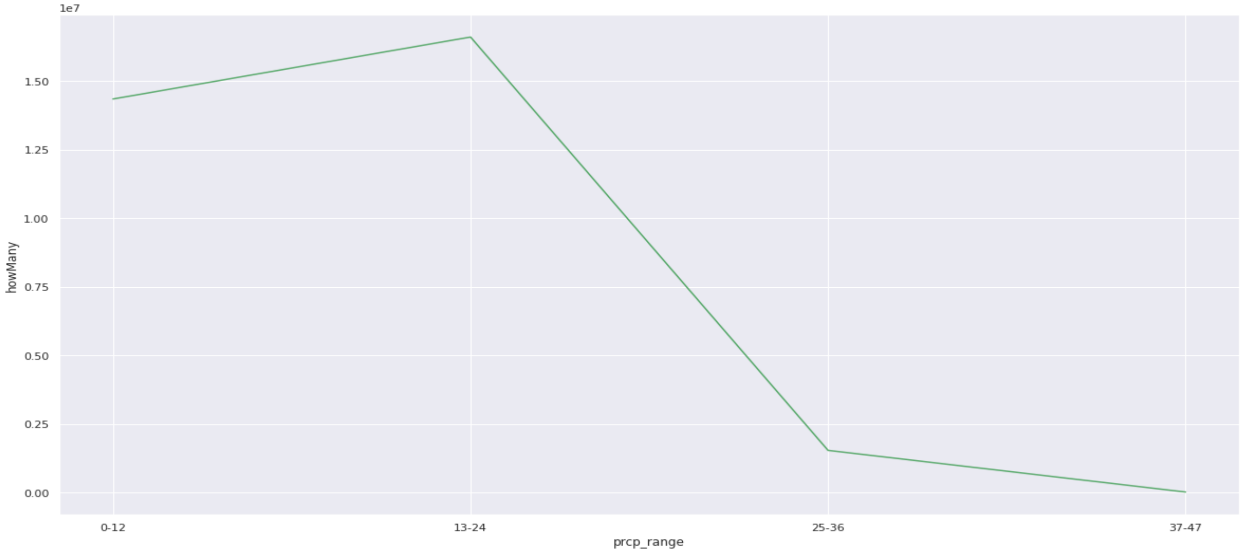
Winter Species decrease over the decades



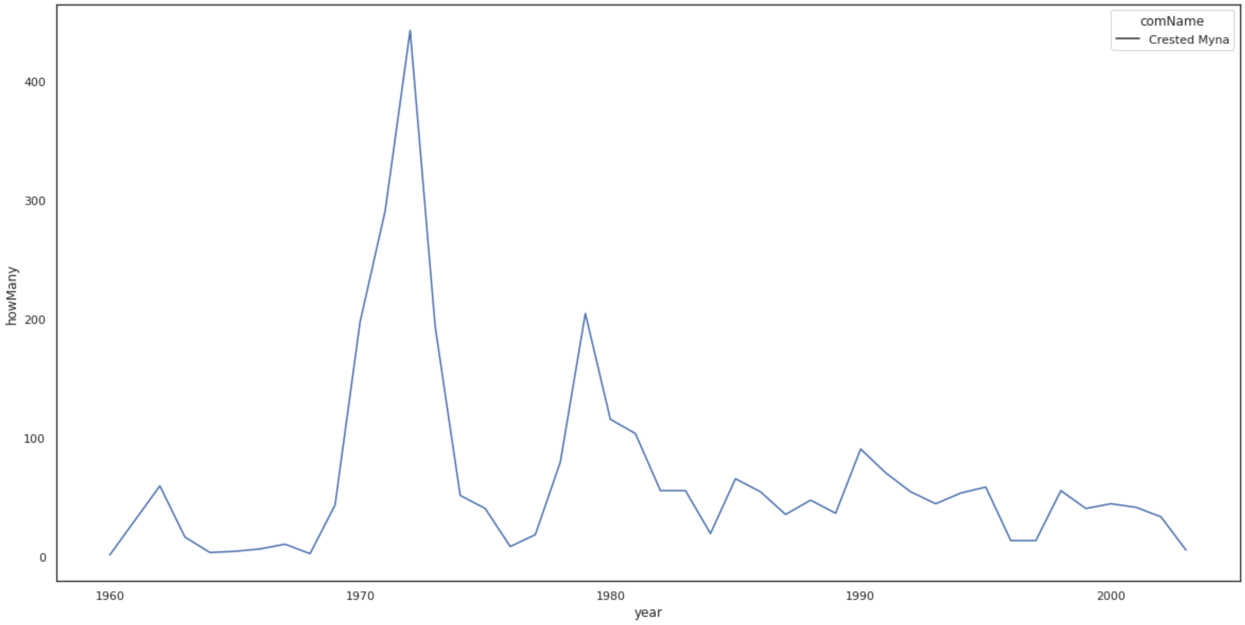
Bird Count over the Temperature range



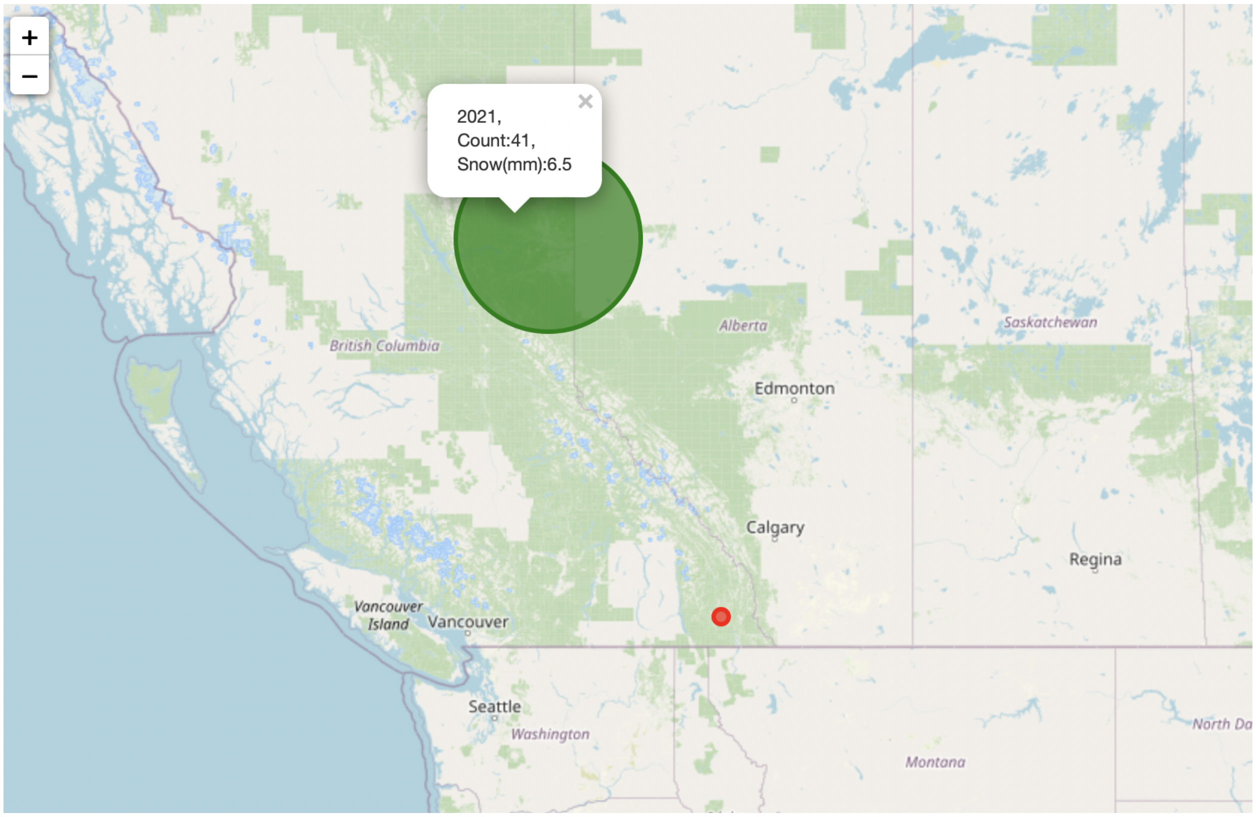
Bird Count over the Snow Range



Bird Count over the Rainfall range



Crested Myna was last spotted in 2003 which is of concern.

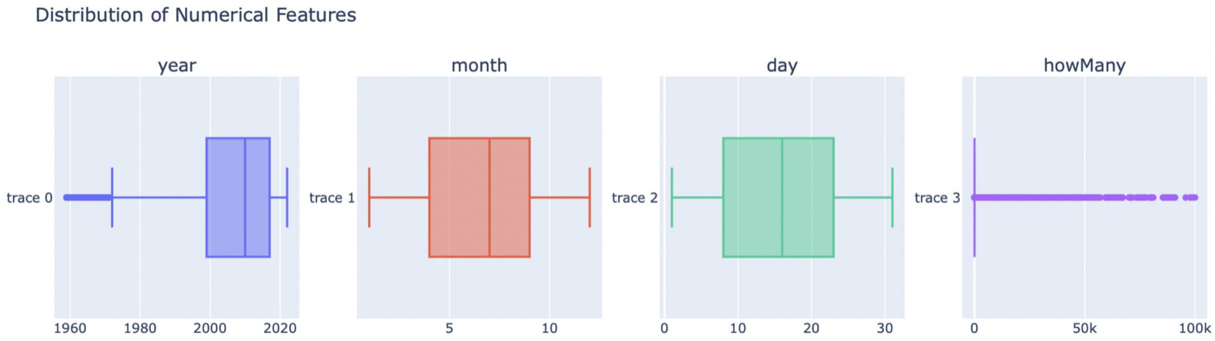


Movement of White-rumped Sandpiper across the year 2021 and 2012

The red circle in the above plot denotes the location of White-rumped Sandpiper in 2012 and the green circle denotes its location in 2021. The size of the circle depicts the total sightings. This bird was spotted around 63% of the time during summer months. Considering that in 2012 the snow(mm) was 11.8 and in 2021 it was 6.5 mm, it makes sense that the bird moved from a location with more snowfall to one with less snowfall.

5.) Data Modeling

Created a machine learning model in pySpark to predict the count of birds.



Box Plots of Numerical Columns

After doing a bit of exploratory data analysis, we noticed the following trends in the data and realized we need to clean the data a bit:

1. Data has 0 null values
2. Categorical columns present with at most 72517 categories(locId). - Encoded these using StringIndexer
3. "howMany" column has very high skewness (right skewed data) - Applied log transformation to it T
4. Scaling needs to be done as in the boxplot you can see that all columns have different scales. - Robust scaler might be better because of the presence of outliers. However, since we were using spark 2.4 on GCP, RobustScaler was not available in spark 2.4 and we had to use StandardScaler
5. "year" column has some outliers i.e. years from 1959 to 1971 (from boxplot). Moreover, from the exploratory data analysis, we can see that there are very few entries for this interval. Ignored these years as data recording in the beginning years wouldn't be that accurate. Also ignored 2022 as when we downloaded the dataset 2022 was still not over.

This would be a regression problem and hence used Linear Regression to model the data. The RMSE score on validation for the LinearRegression model is 623.261.

PROBLEMS & CHALLENGES FACED

- Downloading historical ebird data efficiently, as single threaded python scripting was taking nearly 10 minutes to download data for only 1 year.
- Facilitating quick transfer of files between SFU cluster and Google Cloud Storage
- Due to using GCP free trial, a very limited amount of compute resources were provided. To allocate these effectively to have a decent cluster required trial and error.
- Pivoting weather data to have all weather features in 1 row.
- Joining ebird data and weather data to make sure only the weather data from the closest weather station is taken and entries are not duplicated.
- Limited to libraries available in spark 2.4 (i.e. not the latest version of spark). Had to use spark 2.4 as it was compatible with jupyter notebooks in GCP.

CONCLUSION

The overall analysis shows that the bird count has started decreasing prior to 2009 and these could be because of many underlying factors. As ebird API is a crowdsourced platform, the sightings can never be accurate but even if we leave space for trial and error, the continual drop shows that subsequent species sightings have gone down especially for the winter birds. This could mean disturbance in migration patterns for birds habitual to colder climates. However, the species for remaining seasons such as spring, summer and fall show an increasing trend. Moreover, the temperature, snow and rainfall correlates that sightings have indeed gone down when the weather conditions have been extreme. This means that harsh weather conditions pose threats to the bird population sightings.

PROJECT SUMMARY

Category	Points
Getting the data : Extracted data from Global Historical Climatology Network (GHCN) and Cornell's Ebird API.	3
ETL: Pivoting the weather data. Observation date format for both weather and ebird had to be tweaked. Joined the weather and ebird data on location coordinates.	2
Problem: Problem statement was based on the hypothesis of declining winter birds in BC due to the effects of climate change for which we were able to provide sufficient proof.	1
Algorithmic Work: Computed the nearest weather station for each ebird row using Haversine distance with pySpark to iterate through years.	4
Bigness/Parallelization - 16GB Dataset, GCP Computer Cluster with 24 vCPUs. Data Scraping was done with Multithreading.	4
UI	0
Visualization: Cleaned the data and applied data transformations on the datasets, converting them to pandas dataframe to use analytical libraries - seaborn and matplotlib.	3
Technologies: New technologies such as Google Cloud Platform - Cloud Storage, DataProc Compute Cluster as well as Python Multiprocessing.	3