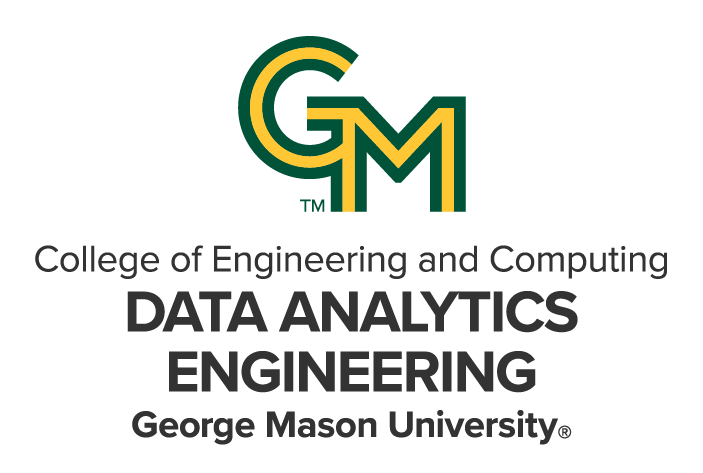
A black rectangle with a black background

Description automatically generated with low confidence

DAEN 690

Project Report

Gauri Tiwari

Puneeth Velivela

Rahul Guttapally

Sandeep Lingampelly

Shreya Rajani Shankar

Vivek Patil Paidigumal

Spring 2025

Extracting Relevant Records from Two Large Colonial Waterbird Datasets

**About the Cover**

This semester, the DAEN program is proud to spotlight one of our esteemed capstone partners, Daniel Erasmus—a visionary whose groundbreaking work influences leaders worldwide. As the founder and CEO of Erasmus.AI, Daniel is a renowned futurist and a pioneer in scenario planning, artificial intelligence, and strategic foresight. His innovative approach to blending AI with human-centric decision-making has profoundly shaped global conversations on technology, sustainability, and future-readiness. Through his thought leadership, Daniel continues to inspire organizations across the globe to embrace change and build resilient futures.

At Erasmus.AI, Daniel conceived and led the development of ClimateGPT—the world’s first foundational AI model family focused on climate change. Built on over a decade of collecting and processing planetary-scale datasets, this groundbreaking innovation leverages AI to uncover hidden connections in global news, from Human-Centered Extreme Weather Dashboards to maps of global innovations, risks, and breakthroughs. The Erasmus.AI platform exemplifies his commitment to using technology to inform and address some of the world’s most pressing challenges.

As co-founder of The Digital Thinking Network (DTN), Daniel has spent over 25 years leading large-scale scenario planning and transformation processes. His work has driven notable actions, such as initiating a response to food security challenges during COVID-19 that delivered 1 million meals within three months and has since provided over 60 million meals in Sub-Saharan Africa. His scenario processes have also anticipated major global events, including the Global Financial Crisis in 2006 and the Oil Price Collapse in 2012—each resulting in multi-billion-dollar benefits for his clients. In the public sector, DTN's transformative initiatives include the Rotterdam Advisory Board, which spearheaded the Rotterdam Climate Initiative in 2005 with the ambitious goal of halving CO2 emissions by 2025, and the creation of the 30-year global future scenarios Ci’Num.

An accomplished author, Daniel has written three books on innovation and the networked society, as well as numerous columns, including the Information Society column for the Financial Times Review. He has also held various prominent board positions and fellowships, including serving on the University of Stellenbosch’s Faculty of Science Advisory Board, Cambridge-based Titan Advanced Energy Solutions, and the supervisory board of the Quad9 Foundation. Through his visionary leadership, Daniel continues to shape the future across disciplines and industries.

Contents

Table of Contents

[Section 1: Introduction 3](#_Toc197284198)

[1.1 Report Purpose 3](#_Toc197284199)

[1.2 Report Readership 3](#_Toc197284200)

[1.3 Report Structure 3](#_Toc197284201)

[Section 2: Problem Definition 4](#_Toc197284202)

[2.1 Problem Space 4](#_Toc197284203)

[2.2 Research 5](#_Toc197284204)

[2.2.1 Research on Population Trends & Habitat Changes 5](#_Toc197284205)

[2.2.2 Impact of Climate Change & Human Activity 6](#_Toc197284206)

[2.2.3 Understanding the Avian Knowledge Network (AKN) & Data Validation 6](#_Toc197284207)

[2.2.4 Collaboration & Next Steps 6](#_Toc197284208)

[2.3 Solution Space 6](#_Toc197284209)

[2.4 Project Objectives 7](#_Toc197284210)

[2.5 Primary User Stories 8](#_Toc197284211)

[2.6 Product Vision 9](#_Toc197284212)

[2.6.1 Scenario #1- Enhancing Conservation Decision-Making 9](#_Toc197284213)

[2.6.2 Scenario #2 - Standardizing Ecological Research Data 9](#_Toc197284214)

[Section 3: Datasets 9](#_Toc197284215)

[3.1 Overview 9](#_Toc197284216)

[3.2 Field Descriptions 10](#_Toc197284217)

[3.3 Data Context 13](#_Toc197284218)

[3.4 Data Conditioning 14](#_Toc197284219)

[3.5 Data Quality Assessment 15](#_Toc197284220)

[3.6 Other Data Sources 17](#_Toc197284221)

[3.6.1 Ornithology Journal (IBP-AOS-LIST24.csv) – AOU Code Standardization 17](#_Toc197284222)

[3.7 Storage Medium 18](#_Toc197284223)

[3.8 Storage Security 18](#_Toc197284224)

[3.9 Storage Costs 19](#_Toc197284225)

[Section 4: Data Standardization 19](#_Toc197284226)

[4.1 Data Cleaning and Standardization 19](#_Toc197284227)

[4.1.1 Filling Missing State Values 19](#_Toc197284228)

[4.1.2 Filtering Records for the Atlantic Flyway  20](#_Toc197284229)

[4.1.3 Filtering of Non-Colonial WaterBirds Species  20](#_Toc197284230)

[4.1.4 Populating Missing Population Estimates  20](#_Toc197284231)

[4.1.5 Standardization of Categorical Variables  20](#_Toc197284232)

[4.1.6 Unification of Species Identification  21](#_Toc197284233)

[4.1.7 Correction of Misclassified Values  21](#_Toc197284234)

[4.1.8 Elimination of Redundant Records  21](#_Toc197284235)

[4.1.9 Consolidation of Survey Method Descriptions  22](#_Toc197284236)

[4.2 General Quality Enhancements 22](#_Toc197284237)

[4.3 Flagging of Records for Expert Review 22](#_Toc197284238)

[4.3.1 Flagging Categories 22](#_Toc197284239)

[4.3.2 Updates Tracking Fields 23](#_Toc197284240)

[Section 5: Data Visualization and Dashboard Development 23](#_Toc197284241)

[5.1 Dashboard 1: Atlantic Flyway Avian Population Overview 23](#_Toc197284242)

[5.2 Dashboard 2: Expanded View of Colony Distribution and Species Presence 28](#_Toc197284243)

[5.3 Dashboard 3: Improved Species-Level Colony Visualization and Exploration 31](#_Toc197284244)

[5.4 Summary of Visualization Outcomes 34](#_Toc197284245)

[Section 6: Findings 35](#_Toc197284246)

[Section 7: Summary 35](#_Toc197284247)

[Section 8: Future Work 35](#_Toc197284248)

[Appendix A: Domain Background 39](#_Toc197284249)

[Appendix B: Glossary 43](#_Toc197284250)

[Appendix C: GitHub Repository 44](#_Toc197284251)

[Overview 44](#_Toc197284252)

[GitHub Repository Link 44](#_Toc197284253)

[GitHub Repository Contents 44](#_Toc197284254)

[Appendix D: Risks 45](#_Toc197284255)

[Sprint 1 Risks 45](#_Toc197284256)

[Sprint 2 Risks 46](#_Toc197284257)

[Sprint 3 Risks 47](#_Toc197284258)

[Sprint 4 Risks 49](#_Toc197284259)

[Sprint 5 Risks 50](#_Toc197284260)

[Appendix E: Agile Development 52](#_Toc197284261)

[Scrum Framework Team Approach 52](#_Toc197284262)

[Sprint 1 Lessons Learned 52](#_Toc197284263)

[Sprint 2 Lessons Learned 53](#_Toc197284264)

[Sprint 3 Lessons Learned 53](#_Toc197284265)

[Sprint 4 Lessons Learned 54](#_Toc197284266)

[Sprint 5 Lessons Learned 54](#_Toc197284267)

[Works Cited 56](#_Toc197284268)

Table of Figures

[Figure 1: Flyway routes of USA 4](#_Toc197283936)

[Figure 2: Interactive Tableau Dashboard 1 for Atlantic Flyway Avian Population Overview 24](#_Toc197283937)

[Figure 3: State/Province Selector Map 24](#_Toc197283938)

[Figure 4: Filter by Year to View Historical Survey Trends for Selected State 25](#_Toc197283939)

[Figure 5: Year-Based Species Filtering Table 25](#_Toc197283940)

[Figure 6: Species Population Trend Across Years 26](#_Toc197283941)

[Figure 7: Annual Survey Effort Visualization 26](#_Toc197283942)

[Figure 8: Animated Colony Distribution Over Time 27](#_Toc197283943)

[Figure 9: Monthly Observations and Survey Methods for Laughing Gull (2018) 27](#_Toc197283944)

[Figure 10: Survey Methods for Laughing Gull in 2018 28](#_Toc197283945)

[Figure 11: Interactive Tableau Dashboard 2 for Colony Distribution 29](#_Toc197283946)

[Figure 12: Map Showing Colony Distribution of Selected Species in that State and year, with Tooltip Displaying Location, Species, Colony ID, and Population Estimate 30](#_Toc197283947)

[Figure 13: Interactive Tableau Dashboard 3 for Colony Locations, Species Composition, and Population Insights 31](#_Toc197283948)

[Figure 14: Colony Distribution Map with Interactive Species and Year Filters 32](#_Toc197283949)

[Figure 15: Bar Chart Displaying Species Composition and Population at a Selected Colony Site 33](#_Toc197283950)

[Figure 16: Bubble Chart Displaying Top 10 Colonies by Population for Selected Species in 2007 33](#_Toc197283951)

[Figure 17: Sprint project dates. 52](#_Toc197283952)

Table of Tables

[Table 1: Glossary Table 43](#_Toc197283918)

[Table 2: Sprint 1 Risks 45](#_Toc197283919)

[Table 3: Sprint 2 Risks 46](#_Toc197283920)

[Table 4: Sprint 3 Risks 48](#_Toc197283921)

[Table 5: Sprint 4 Risks 50](#_Toc197283922)

[Table 6: Sprint 5 Risks 51](#_Toc197283923)

**This page intentionally left blank**

Abstract

Colonial waterbird populations serve as global indicators of wetland health, biodiversity, and climate change impacts, making their long-term monitoring critical to ecosystem assessment. Their breeding and migration patterns offer key insights into ecosystem conditions. North America’s Atlantic Flyway, a major migratory corridor, supports many species through coordinated efforts across Atlantic Flyway States and Canadian provinces. This project, in partnership with the Atlantic Flyway Council’s Nongame Migratory Bird Technical Section (NMBTS), uses long-term data to inform regional conservation strategies. However, the historical dataset provided exhibited substantial challenges, including inconsistent data formats, outdated American Ornithological Society (AOS) species codes, inclusion of non-Atlantic Flyway States, redundant fields, misclassified observer data, and incomplete metadata. These limitations significantly reduced the dataset’s utility for trend analysis and limited its value for conservation decision-making. To address this, a comprehnsive data cleaning and transformation effort was undertaken using Python Libraries, which included validating geographic entries via geolocation, filtering records specific to the Atlantic Flyway, standardizing categorical values, correcting misclassifications, and updating species identifiers using current ornithological taxonomy. Records missing population estimates, unit types, or recorded outside breeding months were flagged and summarized by species and location for partner review. Following the cleaning phase, an interactive Tableau Public dashboard was developed to visualize population trends, colony distributions, and survey methodologies. Partner-driven enhancements enabled additional species-specific visualizations to support biological review. Initial testing on Virginia's Laughing Gull population suggested a potential decline over time. To verify this trend, breeding periods, colony distribution, and survey effort were examined. The consistency across these indicators confirmed the decline as valid. These findings underscore the urgent need for standardized survey methodologies across states to ensure reliable long-term monitoring, guide conservation priorities, and support sustainable species recovery throughout the Atlantic Flyway.

**This page intentionally left blank**

Report

# Introduction

## Report Purpose

This report explains the progress made in the project, including the research done, identification of critical data sources, and steps taken to prepare for working with the **Atlantic Flyway Council’s Nongame Migratory Bird Technical Section (NMBTS)**. The goal is to ensure that reliable and relevant **Colonial WaterBird (CWB)** data from historical datasets can be cleaned, standardized, and integrated into the **Avian Knowledge Network (AKN)** for conservation and research.

## Report Readership

This report is intended for **the project team, faculty advisors, and partners at the Nongame Migratory Bird Technical Section (NMBTS)**. It provides a structured overview of our project’s objectives, methodologies, and findings to ensure alignment among all stakeholders. The **project team** will use this report to document progress, challenges, and solutions, guiding future development phases. **Faculty advisors** will review the report to assess research depth, technical execution, and adherence to best practices.

Additionally, **NMBTS partners** will leverage this report to understand how our work contributes to conservation efforts, particularly in enhancing data quality for long-term monitoring of colonial waterbirds. By maintaining clarity and accuracy in our documentation, we aim to support **collaborative decision-making and data-driven conservation strategies**. This report serves as a key reference for tracking the project's development and ensuring that all involved parties remain informed of its direction and impact.

## Report Structure

This report is structured to provide a clear understanding of the project’s objectives, challenges, and progress. It begins with the **Problem Definition**, outlining the issues with historical colonial waterbird data, including inconsistencies and missing metadata. The **Methodology** section details the datasets used, data validation steps, and integration into the **Avian Knowledge Network (AKN)**. The report also covers the **Development of a Validation Tool** to improve future data collection. Finally, it concludes with **Project Progress and Next Steps**, summarizing key achievements and future work to ensure data accuracy for conservation planning.

# Problem Definition

## Problem Space

Effective conservation planning for colonial waterbirds in the Atlantic Flyway relies on accurate, well-organized, and standardized datasets. However, existing records suffer from inconsistencies, incomplete metadata, outdated species codes, and the inclusion of irrelevant data from non-Atlantic Flyway states. These issues reduce the dataset's reliability and hinder their integration into larger conservation networks like the Avian Knowledge Network (AKN).

A map of different colors of the north america

Description automatically generated

Figure 1: Flyway routes of USA

This above map shows North America's four major migratory bird flyways: Pacific (red), Central (orange), Mississippi (green), and Atlantic (blue). These flyways represent the primary routes birds take during seasonal migration between breeding and wintering grounds.

To address these challenges, this project focuses on **cleaning, standardizing, and validating colonial waterbird datasets**. Key tasks include **removing redundant or irrelevant records, ensuring proper classification of species using correct AOU codes, correcting misclassified observer data, and improving metadata completeness**. Additionally, outdated species codes will be updated, and the dataset will be restructured to enhance its accessibility and interoperability.

By refining these datasets, we aim to **improve data accuracy, consistency, and usability**, ensuring they serve as a **valuable resource for conservation research, policymaking, and habitat management** within the Atlantic Flyway.

A major concern within this problem space is the **impact of climate change and human expansion** on colonial waterbirds. **Sea-level rise** is a significant threat, especially to species that nest in low-lying coastal areas. Many breeding colonies are now at risk of flooding, forcing birds to either relocate or abandon traditional nesting sites. **Urbanization and industrial expansion** further exacerbate the issue by encroaching on natural habitats, increasing disturbances, and reducing available food sources. Without a structured and **reliable data framework**, it becomes nearly impossible to monitor these changes effectively and adapt conservation strategies accordingly.

Given the **magnitude and complexity** of the problem, solving it entirely is beyond the scope of this project. However, our work aims to address a **specific and critical subset**—the **extraction, validation, and integration of historical waterbird records** to enhance data quality and usability. By developing **decision rules to filter and clean unreliable records**, we can help remove duplicate entries and incorrect observations while preserving valuable historical insights. Additionally, we will create a **data validation tool** to standardize future data collection, ensuring that **new records meet quality standards before being integrated into AKN**.

This project focuses on **cleaning and validating historical datasets**, but our work also provides a foundation for **future conservation efforts**. Once we improve the quality of available data, additional research can focus on:

* **Predictive modeling** to assess how colonial waterbirds will respond to future climate conditions.
* **Automated validation algorithms** to enhance the efficiency of data integration.
* **Citizen science contributions** to support long-term data collection and public engagement in conservation.

By narrowing our focus to a **manageable part of the broader issue**, we ensure a **realistic and impactful approach** while creating a **scalable framework** that conservationists can build upon. Addressing **data quality and reliability** is a fundamental step toward **ensuring the accuracy of long-term conservation assessments and strategies**.

## Research

To address the challenges of **data reliability and conservation planning** for colonial waterbirds in the Atlantic Flyway, our team conducted extensive research across multiple areas. Our research efforts were organized into three key themes: **species population trends, environmental impact factors, and data management in conservation databases**. This research provided us with a clear understanding of **historical data inconsistencies, the impact of climate change, and the role of data validation in improving conservation efforts**.

### Research on Population Trends & Habitat Changes

Our first area of research focused on **colonial waterbird population trends and their connection to habitat changes**. By reviewing studies on **species distribution, nesting behaviors, and long-term population declines**, we identified key factors contributing to the **reduction of colonial waterbird numbers**. Specifically, we examined:

* **Declining species populations** due to habitat fragmentation and human disturbances.
* The **loss of critical nesting sites** in coastal regions, primarily caused by wetland degradation.
* **Variability in breeding success** due to environmental changes and increased predation risks.

Additionally, we explored **seasonal migration routes and site fidelity**, assessing how environmental stressors influence colonial waterbird **nesting preferences, movement patterns, and survival rates**. Understanding these trends is crucial for **integrating reliable historical data** into conservation models.

### Impact of Climate Change & Human Activity

Our second research focus involved analyzing the **effects of climate change and human intervention on colonial waterbirds**. Rising **sea levels, extreme weather events, and habitat destruction** are increasingly **forcing waterbirds to adapt or relocate**, which disrupts **long-established migratory and breeding behaviors**. Key topics we investigated include:

* **How climate change alters breeding success** and site selection.
* The **effects of urbanization, tourism, and offshore wind development** on nesting colonies.
* **Increased predation risks** due to habitat loss, forcing species into areas with less protection.

Through this research, we identified **patterns of environmental stress** that impact both **short-term behaviors** and **long-term survival**, reinforcing the importance of **accurate conservation data**.

### Understanding the Avian Knowledge Network (AKN) & Data Validation

To ensure **effective data management**, we studied the **Avian Knowledge Network (AKN) database** to understand how historical and contemporary datasets are structured, stored, and validated. Our findings included:

* **Common inconsistencies in historical datasets**, such as missing metadata and duplicate records.
* The **importance of standardized data formats** for integration into modern conservation systems.
* How **decision rules and quality assurance methods** can improve **data accuracy and reliability**.

Additionally, we examined **data validation techniques** used in conservation research, focusing on **automated filtering methods, data cleaning strategies, and standardization procedures**. This research will guide our development of a **data validation framework**, ensuring that only **high-quality, verified records** are used for future conservation planning.

### Collaboration & Next Steps

Throughout our research, we engaged with **scientific literature, conservation reports, and online databases** to establish a **comprehensive knowledge base**. As we proceed with **dataset analysis and validation**, our findings will shape the **decision rules and quality control measures** we implement. Moving forward, we will continue refining our **data validation methods**, ensuring **seamless integration** of reliable historical records into the AKN.

This research lays the foundation for our project, **bridging the gap between historical datasets and modern conservation needs**. By ensuring data integrity, we aim to support **long-term monitoring and conservation efforts** for colonial waterbirds in the Atlantic Flyway.

## Solution Space

Our solution aims to **enhance the reliability and accessibility of historical colonial waterbird (CWB) data** by implementing a **data validation, cleaning, and integration framework**. This approach ensures that **decision-makers, conservationists, and researchers** can rely on accurate, high-quality datasets to guide long-term conservation planning and environmental policy.

By standardizing and integrating historical datasets into **the Avian Knowledge Network (AKN)**, our system delivers value in the following ways:

* **Improved Conservation Insights** – Clean and validated data will enable **more accurate population trend analysis**, helping conservationists track species distribution changes over time.
* **Data-Driven Decision-Making** – Reliable datasets will support **habitat management, climate change adaptation, and offshore wind development** while minimizing ecological risks.
* **Enhanced Data Integrity** – Standardized validation rules will **eliminate duplicate records, correct inconsistencies, and ensure metadata completeness**, making historical records more useful.
* **Automated Data Validation** – Our **validation tool** will streamline the **data cleaning and verification process**, ensuring **long-term accuracy in monitoring efforts** and reducing manual errors.

By creating a **structured data pipeline**, our system ensures that **historical waterbird data is no longer fragmented or unreliable** but instead serves as a **foundation for informed conservation strategies and ecological forecasting**. Over time, this project will help **stakeholders make data-backed decisions that contribute to the preservation of colonial waterbird populations and their habitats**.

## Project Objectives

This project aims to **improve the quality, accessibility, and reliability of historical colonial waterbird (CWB) datasets** by developing a structured **data validation and integration framework**. By cleaning and standardizing historical records, the project will ensure that conservationists and policymakers have **accurate, high-quality data** for long-term monitoring and decision-making.

1. **Expected Learnings from the Project**

Through this project, the team expects to gain **expertise in data validation techniques, decision rule implementation, and dataset integration** for ecological research. Additionally, the team will develop a better understanding of **challenges related to managing historical conservation data**, including **inconsistencies, missing metadata, and duplicate records**.

1. **Anticipated Solution Outcomes**

The primary outcome of this project is a **validated and standardized historical dataset**, which will be **integrated into the Avian Knowledge Network (AKN)** for broader accessibility. This will ensure that **population trend analyses, conservation policies, and habitat management strategies** are based on **accurate and consistent historical records**. Additionally, a **data validation tool (Shiny App)** will be developed to **automate quality control** for future dataset submissions, ensuring that **newly collected data meets standardized quality criteria before integration**.

1. **Enhanced Understanding of the Problem Space**

The team expects to develop **a deeper understanding of data inconsistencies in historical ecological datasets** and their impact on **conservation planning**. Working closely with the **North American Migratory Bird Technical Section (NMBTS)**, the team will refine **data filtering and validation criteria** to ensure **historical records remain useful for contemporary research and conservation efforts**.

1. **Value Delivered to Partners**

This project will provide **validated historical datasets** that improve **conservation monitoring, ecological forecasting, and habitat protection strategies**. By ensuring **data reliability and accessibility**, the project will support **conservationists, policymakers, and researchers** in making **informed decisions** regarding **habitat preservation, climate adaptation, and species protection initiatives**.

1. **Project Objectives Summary**

* **Objective 1:** Establish **decision rules** to filter, clean, and validate historical CWB datasets.
* **Objective 2:** Collaborate with **NMBTS stakeholders** to refine data selection criteria.
* **Objective 3:** Integrate **high-quality, validated historical records** into the **AKN database**.
* **Objective 4:** Develop a **data validation tool (Shiny App)** to maintain data integrity in future submissions.
* **Objective 5:** Implement a **Quality Assurance/Quality Control (QAQC) framework** to ensure **long-term dataset reliability**.

By accomplishing these objectives, the project will **significantly enhance the accuracy of waterbird population monitoring**, support **scientific research**, and contribute to **data-driven conservation strategies** aimed at **protecting colonial waterbird populations**.

## Primary User Stories

Based on the **user context and value proposition**, the team has developed the following **primary user stories** to guide our project. These stories reflect the needs of **conservationists, researchers, and data managers** who rely on accurate historical data to make informed decisions about colonial waterbird conservation.

**User Story 1: Ensuring Access to Reliable Conservation Data**

*"As a conservation scientist, I want access to a validated and standardized historical dataset so I can analyze colonial waterbird population trends, habitat shifts, and species distribution with confidence. This will allow me to make accurate conservation decisions, policy recommendations, and habitat protection plans based on high-quality data, free from inconsistencies and errors."*

**User Story 2: Integrating Historical Data for Long-Term Research**

*"As a researcher studying climate change's impacts on waterbirds, I want to ensure that historical datasets are cleaned and integrated into a unified, standardized format so I can compare past and present trends. This will enable me to conduct long-term ecological studies, examining how factors such as rising sea levels, human activity, and environmental changes influence migration patterns, breeding success, and species survival rates over time."*

**User Story 3: Consolidating and Validating Data for Database Management**

*"As a data manager working with the Avian Knowledge Network (AKN), I want a system that consolidates fragmented, inconsistent records into a unified dataset using decision rules so that I can ensure data integrity. This will help maintain a clean, reliable, and standardized dataset, enabling conservationists and researchers to confidently use the data for monitoring, decision-making, and policy development."*

**User Story 4: Automating Quality Control for Future Data Submissions**

*"As a field biologist collecting new survey data, I want a data validation tool that automatically checks for errors, missing values, and inconsistencies before submission. This will ensure that newly collected data meets standardized quality criteria, preventing inaccurate records from entering the AKN database and improving long-term data accuracy."*

These user stories define how different stakeholders interact with the system and illustrate the value of data validation and integration for conservation efforts. By ensuring data reliability, this project will support accurate ecological research, habitat protection, and informed decision-making.

## Product Vision

Effective conservation of colonial waterbirds relies on accurate, high-quality data to track population trends, habitat changes, and the impact of climate change. However, historical datasets often contain inconsistencies, missing metadata, and duplicate records, making them unreliable for research and decision-making. A solution that validates, cleans, and integrates historical data ensures it meets modern quality standards for long-term ecological monitoring. By providing a structured data validation framework, data reliability can be enhanced, supporting conservation planning, and enabling more accurate scientific research.

### Scenario #1- Enhancing Conservation Decision-Making

For conservation scientists and environmental policymakers who rely on accurate historical data to analyze colonial waterbird population trends and habitat changes, the Avian Knowledge Network (AKN) serves as a centralized ecological database that stores and manages both historical and contemporary waterbird records. A solution that provides validated, standardized datasets ensures clean, high-quality data for species distribution analysis, habitat protection planning, and long-term environmental impact studies.

Unlike existing datasets that are fragmented, inconsistent, and contain duplicate or missing records, a validation and integration system clean, organizes, and standardizes historical data to improve reliability. This ensures conservationists can confidently use the data for monitoring and decision-making without the risk of inaccurate conclusions. However, historical data gaps may still exist, requiring conservationists to supplement validated datasets with additional field studies and research insights for comprehensive analysis.

### Scenario #2 - Standardizing Ecological Research Data

For researchers studying the impact of climate change on colonial waterbirds, who require both historical and contemporary datasets to analyze long-term environmental trends, a data validation framework serves as a system that ensures datasets are clean, accurate, and standardized. A structured approach to validation eliminates inconsistencies, corrects errors, and ensures metadata completeness, enabling researchers to conduct reliable comparative studies on migration patterns, breeding success, and the effects of climate change on species distribution.

Unlike manually processed datasets that are prone to human error, missing values, and unreliable records, an automated validation process ensures higher data integrity while making ecological research more efficient and data driven. However, while the system provides validated historical records, researchers still need to interpret and model trends using specialized conservation methodologies to generate actionable insights.

# Datasets

## Overview

The dataset used in this project is the **Atlantic Flyway Colonial Waterbird Survey Dataset (AFCWB Dataset)**.It supports ecological study and conservation planning by providing thorough records of colonial waterbird populations in different Atlantic Flyway regions. This dataset offers information on population trends, species distribution, and environmental factors that have an impact on waterbirds over a long period of time.

With 50 columns and 67,049 rows, the dataset has a total file size of 8,120 Kilobytes. The data encompasses several places throughout the Atlantic Flyway States and spans the years 1850 to 2019. The dataset is crucial for incorporating historical data into contemporary conservation databases and assessing population trends over time.

The dataset is divided into six main components to guarantee efficient analysis:

* **Survey Information:** Gathers the primary survey information, such as VisitNum, Observer, SurveyDate, and ColonyName. This section aids in determining the date, time, and location of each survey's administration.
* **Location & Geographic Data:** Enables accurate colony location mapping by providing spatial data like Latitude, Longitude, County, State, and UTMCoordinates.
* **Species & Population Data:** Keeps track of bird species observations, including Species, CountType, UnitCounted, and PopulationEstimate fields to monitor species diversity and population levels.
* **Data Collection Methodology:** Using elements like SurveyMethod, DataQuality, and ReliabilityCode to guarantee consistency and dependability, the data collection methodology describes how the data was gathered and evaluated.
* **Metadata & Record Tracking:** Provides insight regarding dataset maintenance and provenance using fields like whoCreated, dateCreated, and source\_note.
* **Administrative & Contact Information:** Offers contact information so users can get in touch with data producers for more details or explanations.

Early findings point to a few issues that must be resolved during the data cleansing procedure. These include duplicate entries for the same species at the same places, format inconsistencies (e.g., different date formats and species name discrepancies), and missing metadata (e.g., incomplete ColonyCodes and missing timestamps). By resolving these problems, the dataset will be dependable, clean, and prepared for inclusion in conservation databases.

All things considered, the Atlantic Flyway Colonial Waterbird Survey Dataset is a comprehensive and useful tool for learning about waterbird populations, assisting with conservation choices, and promoting the long-term viability of avian species throughout the Atlantic Flyway.

## Field Descriptions

The **Field Descriptions** provides a detailed breakdown of the attributes within the dataset, explaining their purpose, format, and relevance to the project. Understanding these fields is crucial for ensuring **data consistency, accuracy, and proper integration** during the cleaning and validation process. This section outlines key dataset variables, highlights **important metadata considerations**, and identifies potential **ambiguities or missing values** that require further clarification. By documenting field descriptions comprehensively, the team ensures a structured approach to **data standardization and quality control**.

Below are the fields included in the dataset:

* **ColonyName (Type: string)** – The name of the bird colony or survey location. This field identifies the site where observations were made. Example: *"Hardwood Island"*.
* **ColonyCode (Type: string)** – A unique code assigned to each colony. This code is used to differentiate between colonies that may have similar names. Example: *"59-236"*.
* **SurveyDate (Type: integer)** – The year on which the survey was conducted. It is a redundant field which contains only data related to survey year. This field may be blank if only the year is provided. Example: *(2018)*.
* **Year (Type: integer)** – The year the survey was conducted. This field is critical for temporal analyses and trend assessments. Example: *1978*.
* **Month (Type: integer)** – The month during which the survey was conducted. May be blank if the detailed date is not provided. Example: *(6)*.
* **Day (Type: integer)** – The day of the month on which the survey was conducted. This field is optional if only the year (or partial date) is available. Example: *(31)*.
* **StartTime (Type: integer)** – The starting time of the survey event. If provided, it should be in 24-hour integer format. Example: *(1530)*.
* **EndTime (Type: string or datetime)** – The ending time of the survey event. Similar to StartTime, it should follow a consistent 24-hour time format. Example: *(1620)*.
* **Observer (Type: string)** – The name or identifier of the person who conducted the survey. This field helps in tracking data collection responsibility. If there is no observer data, the field can be “Unknown” . Example: (*"Vosmik"*).
* **VisitNum (Type: integer)** – The visit or survey iteration number. This helps to differentiate multiple surveys at the same location over time. Example: *(48)*.
* **Latitude (Type: float)** – The geographic latitude of the survey location, expressed in decimal degrees. Example: (*44.306)*.
* **Longitude (Type: float)** – The geographic longitude of the survey location, expressed in decimal degrees. Example: (*-68.4517)*.
* **TownCity (Type: string)** – The town or city where the survey was conducted. This helps provide additional location context. Example: (*"Tremont")*.
* **County (Type: string)** – The county in which the survey site is located. Example: (*"Hancock")*.
* **State (Type: string)** – The state where the survey was performed, using standard abbreviations. Example: *"ME"*.
* **TopoQuad (Type: string)** – The topographic quadrangle name corresponding to the survey site. This field may be empty if not applicable. Example: *(Achilles)*.
* **Species (Type: string)** – The abbreviated code or Common name for the species observed. Example: *"GBHE"* (representing a specific bird species).
* **CountType (Type: string)** – Describes the nature of the count; for example, whether the count is an estimate or a direct count. Example: *"Estimate"*.
* **UnitCounted (Type: string)** – The unit of measurement for the count (e.g., Individuals, Nests). Example: *"Nests"*.
* **CorrectedCount (Type: integer)** – The count value after any corrections have been applied. This field represents the final, validated count. Example: *43*.
* **UncorrectedCount (Type: integer)** – The original count value before any corrections. This field may be blank if no uncorrected value is provided. Example: *(54)*.
* **PopulationEstimate (Type: integer)** – An overall estimate of the population based on the survey. This field is used when direct counts are supplemented by estimation techniques. Example: *(605)*.
* **SurveyMethod (Type: string)** – The method or technique used to conduct the survey (e.g., Ground, Aerial). This field provides insight into the data collection methodology. Example: (*"Ground"*)
* **DataQuality (Type: string)** – A qualitative assessment of the data quality. This helps gauge the reliability of the survey data. Example: (*"Good")*.
* **ReliabilityCode (Type: integer)** – A numeric code that indicates the reliability or confidence in the survey data. Example: (*9)*.
* **Comments (Type: string)** – Additional notes or comments regarding the survey event, including any observations on the data or field conditions. Example: (*"Source: Eliot Paine; 67 live young, 7 dead young. Location comm: Blue Hill Bay - Hardwood Island")*.
* **CorrectionDetails (Type: string)** – Information on any corrections made by the observer or data manager after the survey. Example: (*"No correction by Observer or Data Manager.")*.
* **UTMCoordinates (Type: string)** – The Universal Transverse Mercator (UTM) coordinates for the survey location as a combined field, if available. Example: *(12N 317003 3642964)*.
* **UTMX (Type: integer)** – The UTM easting value for the survey location. Example: *(317003)*.
* **UTMY (Type: integer)** – The UTM northing value for the survey location. Example: *(3642964)*.
* **UTMZone (Type: integer or string)** – The UTM zone identifier for the location. Example: *(12)*.
* **QuadName (Type: string)** – The name of the quadrangle associated with the survey location. Example: (*"Bartlett Island"* or *"North Haven West")*.
* **QuadNum (Type: string)** – A code or number representing the quadrangle. Example: (*"44068-C4"* or *"44068-B8")*.
* **BCR (Type: integer)** – The Bird Conservation Region number indicating the ecological region. Example: (*14)*.
* **whoCreated (Type: string)** – Identifier for the individual or system that created the record. Example: (*"srb")*.
* **dateCreated (Type: datetime)** – The date the record was created, typically in a standard format (MM/DD/YYYY). Example: (*"5/14/2015")*.
* **pubCode (Type: integer or string)** – A code representing the publication or source of the record. Example: (*225)*.
* **visit\_nr (Type: integer)** – A secondary visit number field that may provide additional ordering or identification for visits. Example: *(242632)*.
* **site\_nr (Type: integer)** – A numeric code to identify the survey site within a larger study or region. Example: *(903335)*.
* **series\_nr (Type: integer)** – A series number that groups related survey records together. Example: *(71089)*.
* **par\_code (Type: string)** – A parameter code used for grouping or linking related records. Example: *(pop03c)*.
* **AOU\_code (Type: string)** – The American Ornithologists’ Union (AOU) four-letter species code representing the bird species observed.
* **ColonyCode2 (Type: string)** – An alternative or secondary colony code that may be used for cross-referencing colonies. Example: *(*297 069.0*)*.
* **source\_note (Type: string)** – Additional notes from the data source that provide context or clarification about the record. Example: (“*Grand Manan Whale and Seabird Research Station”)*.
* **eWKT\_4326 (Type: string)** – A field containing geometry or location data in Well-Known Text (WKT) format using the EPSG:4326 spatial reference system.
* **recID (Type: integer)** – A unique record identifier for the dataset. Example: (“*2480*”).
* **Method\_to\_determine\_sp (Type: string)** – Describes the method used to determine the species, such as visual confirmation or other identification techniques.
* **Contact (Type: string)** – Contact information for follow-up or additional inquiries about the survey record.
* **included\_in (Type: string)** – An indicator or tag showing which dataset or subset this record is included in. Example: (*"unique\_to\_this\_file")*.

## Data Context

The dataset under analysis comprises **67,048 unique records** distributed across **50 columns**, capturing extensive information related to **species**, **colonies**, **surveys**, **locations**, and various **survey methods**. Compiled by **Dr. Zachary Loman** as part of a project funded by the **U.S.G.S Science Support Program** (2018–2020), the dataset aims to provide insights into **colonial waterbird populations** within the **Atlantic Flyway**. The data integrates contributions from multiple sources, including historical datasets and field surveys.

1. **Data Structure and Content**

* **Data Types:** The dataset includes a mix of **integers**, **floats**, **strings**, and **date fields**. While comprehensive, there is variability in the completeness and accuracy of some metadata fields.
* **Key Variables:**
* **Species:** Identifies observed bird species, with both common and scientific names (e.g., Pelagic Cormorant - Phalacrocorax pelagicus), along with **AOU codes**.
* **Location Data:** Includes **Latitude**, **Longitude**, **UTM coordinates**, and **Bird Conservation Region (BCR)** values to capture precise geospatial information.
* **Survey Information:** Covers **survey dates**, **survey methods** (e.g., Aerial, Ground, Boat, Photo), and **observer data**.
* **Count Data:** Records both **CorrectedCount** and **UncorrectedCount** to reflect raw and adjusted bird counts based on survey conditions.
* **Data Quality and Reliability:** Fields like **Data Quality** (ranging from Precise to Unreliable) and **Count Reliability Codes** (1–9) assess data accuracy and methodological integrity.

1. **Metadata Overview**

While the dataset includes metadata, it exhibits certain limitations:

* **Inconsistent naming conventions** in fields such as **Species** and **State**.
* **Missing or incomplete entries** in fields like **Observer Names**, **Detection Cue**, and **eWKT\_4326**.
* **Limited documentation** for some variables, complicating the interpretation of specific data points.

The **PubCode** field and associated tabs provide details on data sources, but certain records, particularly from historical datasets (e.g., **Historic\_CWB\_extracted**), may lack full metadata coverage.

1. **Contextual Relevance**

The dataset focuses on the **Atlantic Flyway** and the monitoring of **colonial waterbird populations**. This context ensures that analyses align with relevant ecological and geographic scopes, providing insights into species distribution, population trends, and habitat usage.

1. **Data Usage Considerations**

* **Scope and Limitations:**
* Excludes data from the **SE Seabird Atlas** and **Gulf of Maine** post-2013, unless specified.
* Some records are unique to this dataset, while others overlap with historical extractions (**221 records** shared with **Historic\_CWB\_extracted**).
* The dataset was **last updated on January 19, 2022** and may not reflect the most recent survey data.
* **Analytical Applications:**
* Suitable for studies on species distribution, population dynamics, and conservation planning within the **Atlantic Flyway**.
* Despite metadata gaps, the dataset provides a strong foundation for ecological and geographical analyses, especially when combined with other data sources for completeness.

## Data Conditioning

The 67,048 records in the collection, which have 50 columns, provide details regarding species, colonies, surveys, locations, count types, survey techniques, and data quality. In addition to the various data types—integers, floats, texts, and dates—many columns contain missing values.

1. **Using Business Rules to Filter**

To satisfy the needs of the client, we will use the following filters:

* Eliminate states not included in the Atlantic Flyway:
* We will check the dataset's State column against a predetermined list of states that are recognized by the Atlantic Flyway and eliminate any records that don't match.
* Eliminate species that are not colonial waterbirds or that are not a part of the Atlantic Flyway:
* A validated list of species that are part of both the Atlantic Flyway and colonial waterbird groups will be used to cross-check the Species column which has been provided by the client.
* Any species that is not on the list will be excluded from the dataset.

1. **Handling Missing Data**

* Columns with excessive missing values (e.g., eWKT\_4326, Method\_to\_determine\_sp, StartTime, EndTime) may be dropped if they are irrelevant to the analysis.
* Impute or remove key missing values in Latitude, Longitude, SurveyDate and Species to maintain data integrity.

1. **Creating a Common Data Format**

* For consistency, convert the year, month, and day into the appropriate date format.
* In the categorical areas (State, Species, SurveyMethod, and DataQuality), make sure that the values are consistent and that the spelling variations are correct.
* Transform the numerical fields (CorrectedCount, UncorrectedCount) into the corresponding float or integer types.

1. **Outlier and Anomaly Detection**

* Check for invalid values:
* Ensure Month values are between 1 and 12.
* Validate Day values (1–31).
* Verify Latitude and Longitude fall within realistic geographic boundaries.
* Detect outliers in CorrectedCount and UncorrectedCount using statistical methods to ensure no extreme or erroneous values skew analysis.

1. **Data Cleaning in Geospatial**

* Check to make sure latitude and longitude accurately correspond to the anticipated Atlantic Flyway region.
* Ensure spatial consistency by standardizing and validating UTMCoordinates, UTMX, UTMY, and UTMZone inputs.
* Verify that the BCR (Bird Conservation Region) values correspond to the appropriate geographical locations.

## Data Quality Assessment

The dataset's quality reflects its complexity, drawing from multiple sources, timeframes, and survey methodologies. Data quality is assessed based on **accuracy**, **completeness**, **consistency**, and **reliability**.

1. **Completeness**

* The dataset contains **66,827** unique records, meaning most records are not duplicated.
* 221 records also appear in another dataset called **Historic\_CWB\_extracted.**
* Some data is missing, particularly for New Hampshire MANEM surveys and the Gulf of Maine data after 2013. This means the dataset may not fully represent all regions and years.

1. **Accuracy and Reliability**

* The data was collected and compiled by Dr. Zachary Loman, Shannon Beliew, and Mark Wimer of USGS, ensuring credibility.
* Some records have been adjusted using correction methods, while others remain uncorrected. These are labelled in the dataset so users can differentiate between them.

1. **Validation and Quality Checks**

* A Reliability Code is assigned to each record to indicate how trustworthy the information is.
* The Data Quality column highlights potential issues with certain records.
* The Correction Details field explains when a data value has been modified, helping users understand why changes were made.

1. **Location Data Accuracy**

* The dataset includes latitude, longitude, and UTM coordinates, which help pinpoint the exact location of each recorded observation.
* The Visit Number field tracks multiple visits to the same site, helping reduce errors from duplicated entries.

1. **Potential Issues and Data Gaps**

The dataset consists of several key fields, each of which has been analyzed for inconsistencies, missing values, and potential data quality issues. Below is a summary of the observations for each column:

* **ColonyName:** Some records contain numerical values in the **ColonyName** field instead of proper Colony names. This inconsistency affects the identification and classification of colony names.
* **ColonyCode:** There are **37,353 missing values** in the **ColonyCode** column. Nearly half of the dataset lacks colony codes, which limits the ability to accurately associate records with specific colonies.
* **State:** A few records in the **State** column contain missing values, which affects the completeness of geographic data.
* **Species:** The **Species** column has missing values and contains non-waterfowl bird species that are not required for the dataset. Additionally, this field includes a mix of both **species common names** and **AOU codes** in the same column, leading to inconsistencies in data representation.
* **CountType, UnitCounted, DataQuality:** These three columns are categorical fields, but their values are not standardized. There are multiple variations of terms representing the same meaning, leading to redundancy and inconsistency in data classification.
* **PopulationEstimate:** The values in the **PopulationEstimate** column are highly similar to those in the **UncorrectedCount** column, indicating potential duplication or redundancy in the dataset.
* **Observer:** The **Observer** column is intended to contain names of individuals who conducted the surveys. However, it also includes **landmass names** and **waterbody names**, which are outliers and do not belong in this field.
* **QuadName, QuadNum:** The values in the **QuadName** and **QuadNum** columns appear to be interchanged, leading to incorrect assignments of quadrants.
* **AOU\_code:** The **AOU\_code** column should contain four-letter alphanumeric codes. However, it currently includes numerical codes, which do not align with the expected format.
* **Month:** The **Month** field contains data that does not represent valid months, leading to errors in temporal data analysis.

These observations highlight the inconsistencies and missing values in key fields of the dataset, requiring further review and standardization.

## Other Data Sources

The primary dataset integrates information from multiple sources that complement and expand upon the data being analyzed. These additional datasets provide historical context, broader geographic coverage, and validation references, ensuring a more comprehensive understanding of colonial waterbird trends.

* **Historic\_CWB\_extracted**
* A historical extract from the **Colonial Waterbird Database** (covering 1895–2004) compiled by **Shannon Beliew** and **Mark Wimer**.
* **221 records** overlap between this dataset and the **Historic\_CWB\_extracted** file.
* This source includes older data do not present in the current dataset, making it valuable for long-term trend analyses.
* **Gulf of Maine Seabird Working Group (GOMSWG) Database**
* Contains comprehensive data for the **Gulf of Maine** region, particularly post-2013, which is not fully covered in the current dataset.
* The GOMSWG database should be consulted for complete data through **2019** to ensure accurate trend analyses in the Gulf of Maine.
* **SE Seabird Atlas**
* This dataset, while not included in the primary dataset, contains relevant seabird survey data for the **Southeast United States**.
* It offers additional geographic coverage that complements the Atlantic Flyway data, though integration would require careful alignment of data formats and variables.
* **MANEM Surveys**
* Data from **MANEM** surveys up to **2013** is included in the current dataset, except for **New Hampshire**.
* For complete regional analysis, supplementary data for **NH** would be needed.
* **PubCode Field References**
* The **PubCode** field in the dataset links records to specific published or unpublished sources, which offer further context and validation for individual data points.
* Consulting these sources can clarify data provenance and support more in-depth analyses.

### Ornithology Journal (IBP-AOS-LIST24.csv) – AOU Code Standardization

To ensure accurate species identification and maintain data consistency, the **Ornithology Journal’s IBP-AOS-LIST24.csv dataset** is being used as a reference for validating and updating species AOU codes. This dataset provides **standardized 4-letter and 6-letter AOU codes**, allowing for precise classification of species within the Atlantic Flyway States. With **2,341 records across 10 columns**, it serves as a comprehensive source for species identification and taxonomy alignment.

The dataset enhances **accuracy and reliability** by offering unique species identifiers, including **SPEC (species codes), COMMONNAME (common species names), and SCINAME (scientific names)**. Since it is sourced from the **2024 Ornithology Journal**, it reflects the latest classification standards and covers the majority of species present in the study region. **Validation efforts will focus primarily on 4-letter AOU codes**, ensuring that species names are standardized across datasets.

One potential challenge is the **annual update of this dataset**, which may result in missing species that were previously recorded in older versions. In such cases, prior editions of the Ornithology Journal will be referenced to fill gaps. Additionally, some species codes may require manual verification due to **inconsistent formats** between historical and current datasets. Addressing these discrepancies will be crucial for maintaining **data integrity and long-term usability**.

By incorporating this dataset into the project workflow, we aim to **improve data standardization, facilitate accurate species classification, and ensure consistency across historical and contemporary records**.

## Storage Medium

The dataset for this project, the **Atlantic Flyway Colonial Waterbird Dataset**, was stored and managed using **Microsoft Office 365 tools**, including **OneDrive** and **Microsoft Teams**. OneDrive was used to store datasets, reports, and documentation, enabling seamless collaboration among team members while maintaining institutional security and access controls. Microsoft Teams served as the primary platform for **communication, file sharing, and version tracking** throughout the project.

Throughout the **data cleaning and standardization process**, the dataset was handled **entirely in Excel format**. Excel was chosen because it allowed the team to **easily review, manipulate, and validate large datasets using filters, formulas, conditional formatting, and pivot tables**. The familiar interface and flexibility of Excel supported both manual checks and structured transformations, helping ensure that inconsistencies were identified and resolved efficiently.

In addition to cloud storage, **local copies of Excel files** were temporarily used during active cleaning and processing tasks. Once updates were completed, cleaned files were synced back to OneDrive to maintain a centralized, version-controlled repository. This workflow ensured **data accessibility, security, collaborative editing, and compliance** with university data management requirements.

## Storage Security

Data security is a critical component of this project, ensuring that datasets remain **protected and accessible only to authorized members**. The team follows standard **data security protocols**, including:

* **Access Control:** OneDrive and Teams have restricted access settings, ensuring that only approved users can modify or retrieve files.
* **Version Control:** Using **Microsoft Teams and OneDrive’s built-in version history** to prevent accidental data loss and track changes.
* **Data Encryption:** Microsoft Office 365 provides end-to-end encryption for stored and shared files, securing data from unauthorized access.
* **Backup Strategy:** Automatic **OneDrive cloud backups** protect against data loss due to system failures or accidental deletions. These security measures ensure **data integrity, regulatory compliance, and uninterrupted workflow** throughout the project.

## Storage Costs

The **storage costs for this project are minimal**, as the team is utilizing **Microsoft OneDrive and Teams**, both of which are provided by the university. However, potential cost considerations include:

* **OneDrive Storage Limitations:** Each student has **1TB of free storage** under the university’s Office 365 plan, which is sufficient for project needs.
* **Teams File Storage Restrictions:** While Teams allows seamless collaboration, large datasets may require additional space allocation.
* **Local Storage Considerations:** Temporary data processing may require **local storage on personal devices**, though this cost is negligible. By leveraging **Microsoft Office 365 tools**, the team ensures **cost-effective, secure, and university-compliant data storage** while facilitating smooth collaboration. Future storage needs will be reassessed if the dataset size increases beyond current capacity.

# Data Standardization

This section highlights the structured approach taken to clean, standardize, and transform the dataset to support effective analysis and visualization. The team addressed key data issues such as missing values, inconsistent formats, and redundant records by applying targeted cleaning techniques, integrating external references, and leveraging tools like geolocation APIs. Following data refinement, an interactive dashboard was developed using Tableau to visualize trends, survey efforts, and species distribution across the Atlantic Flyway. This combined effort ensured the dataset was accurate, consistent, and ready for meaningful insights.

## Data Cleaning and Standardization

A reliable dataset is critical for producing accurate insights, especially when used to support conservation planning and species monitoring. To ensure the dataset met the standards required for meaningful analysis, a structured data cleaning process was implemented. This process involved identifying and resolving inconsistencies, handling missing or misclassified values, and eliminating redundancy. The team followed a sequential approach guided by conservation goals and feedback from project partners, ultimately enhancing the dataset’s overall quality, consistency, and readiness for visualization.

### Filling Missing State Values

**Problem Identified:** Some rows in the dataset had missing values in the State column, but these same rows contained valid Latitude and Longitude values. This inconsistency meant that state information could still be inferred using geolocation data, preventing potential gaps in analysis.

**Rationale for Action:** Missing state values could lead to incomplete or inaccurate regional insights. Since latitude and longitude data were available, leveraging a geolocation service allowed the team to fill in these gaps without relying on manual intervention, ensuring a more comprehensive and reliable dataset.

**Resolution Summary:** The team used the Geopy package with the Nominatim geolocation service in Python to determine the missing state values based on coordinates. For each missing entry, latitude and longitude were sent to the API, which returned the correct state name. This information was then updated in the dataset. To prevent exceeding API rate limits, a delay was introduced between requests. After processing, a new dataset was generated with complete state values, ensuring accuracy and consistency across all records.

### Filtering Records for the Atlantic Flyway

**Problem Identified:** The dataset contained records from multiple states and provinces, but only those within the Atlantic Flyway were relevant for analysis. Including data from outside this region could introduce noise and impact the accuracy of conservation insights.

**Rationale for Action:** To ensure the dataset aligned with conservation efforts focused on the Atlantic Flyway, the project team needed to retain only records from officially recognized states and provinces. Removing unrelated data helped streamline analysis and maintain data integrity.

**Resolution Summary:** The team loaded the provided list of valid Atlantic Flyway state abbreviations, cleaned it to correct inconsistencies (e.g., converting "PEI" to "PE"), and compared each record’s state against this list. Records that did not belong to the Atlantic Flyway were removed. To maintain transparency, two datasets were saved: one containing only valid Atlantic Flyway records and another with excluded records for reference. This refinement ensured that the dataset remained relevant and focused on the intended conservation work.

### Filtering of Non-Colonial WaterBirds Species

**Problem Identified:** The dataset included entries for bird species that were not part of the targeted colonial waterbird group defined by the Atlantic Flyway Council. These out-of-scope species could skew insights when evaluating species-specific trends or regional conservation priorities.

**Rationale for Action:** Filtering to include only the relevant species ensures that analytical outputs match the conservation scope of the project. Retaining irrelevant species would dilute the dataset and potentially obscure trends for the focal species. Entries were matched against a curated species list to confirm inclusion.

**Resolution Summary:** Entries with missing or unmatchable species identifiers were flagged for review. Where ambiguity remained, records were cautiously retained but labeled for future clarification.

### Populating Missing Population Estimates

**Problem Identified:** Some records had missing values in the PopulationEstimate column, which is essential for analyzing bird populations and trends. Without these values, certain observations would lack context, reducing the accuracy of population assessments.

**Rationale for Action:** Since population estimates are crucial for conservation analysis, filling in missing values using a structured approach ensured data completeness while maintaining consistency. Using CorrectedCount or UncorrectedCount as fallback values, based on predefined business rules, helped preserve the integrity of population estimates without introducing arbitrary assumptions.

**Resolution Summary:** Following the business rules provided by the partner, the team filled missing PopulationEstimate values using CorrectedCount if available; otherwise, UncorrectedCount was used. If both were missing, the value remained blank. This logic was applied row by row using df.apply() in pandas. After processing, a new dataset with updated population estimates was generated, ensuring a more comprehensive and reliable dataset for analysis.

### Standardization of Categorical Variables

**Problem Identified:** The dataset featured inconsistencies in how categorical information was recorded, particularly in the fields DataQuality, CountType, and UnitCounted. For example, "Good Estimate" and "good" were both used to describe data quality. Some terms appeared in lowercase, uppercase, or with minor spelling variations. These inconsistencies compromise the ability to group and analyze data accurately.

**Rationale for Action:** Standardizing these fields helps ensure consistent groupings when analyzing trends or calculating statistics. For example, if "Actual Count" and "actual" were treated as separate categories, any summary or statistical insight would be misleading. By consolidating categories into a unified terminology, clarity is enhanced, and misclassification is reduced.

**Resolution Summary:** In addition to the three primary fields, similar reviews were conducted for SurveyMethod and Observer, ensuring whitespace, spelling, and capitalization were consistent across all textual fields.

### Unification of Species Identification

**Problem Identified:** Species data in the dataset was recorded using inconsistent formats. While some entries used full common names (e.g., "Great Egret"), others used 4-letter AOU codes (e.g., "GREG"). In many cases, these codes were placed in incorrect columns or inconsistently used across rows. This lack of uniformity made it difficult to identify species-specific trends over time.

**Rationale for Action:** Accurate species identification is essential for ecological analysis. Using a trusted reference list, species names were validated, and their corresponding AOU codes were either added or corrected. Furthermore, rows that contained outdated or unrecognized species identifiers were flagged for review or filtered out. Reconciliation of these inconsistencies enables species-level comparisons without ambiguity.

**Resolution Summary:** Where AOU codes were embedded in the Species column, these values were extracted and relocated into a new column. Inversely, where codes were missing, they were supplemented based on validated common names using external references.

### Correction of Misclassified Values

**Problem Identified:** Some entries in the dataset contained values that were clearly misclassified. For instance, geographic landmass labels were mistakenly included in the Observer field, likely due to data entry errors. Similarly, legacy AOU codes appeared in the Species field instead of a dedicated column.

**Rationale for Action:** Columns should be internally consistent in the type of data they contain. Moving landmass data into a new Landmass column not only corrects the structure but also preserves this potentially useful geographic information. Removing non-observer entries from the Observer field improves its reliability and analytical relevance.

**Resolution Summary:** Additional efforts were made to remove non-standard symbols and formatting artifacts from free-text fields. Examples include strange encoding characters (like ­ or —) which were introduced during export or manual entry. These were systematically removed to ensure consistency across platforms.

### Elimination of Redundant Records

**Problem Identified:** In some cases, the dataset appeared to contain duplicated observations for the same event. A common pattern emerged in which counts of "Individuals" and "Nests" were recorded together with a predictable ratio—usually, two individuals per nest. This redundancy can inflate population estimates if not addressed.

**Rationale for Action:** Keeping both entries in such cases would result in double-counting. For example, a location might report 100 individuals and 50 nests, where the latter simply reflects a derived count from the former. By identifying these paired observations and removing the redundant one (typically the "Nests" entry), population totals are preserved without inflation.

**Resolution Summary:** To maintain consistency, remaining "Nests" records were scaled up by a factor of two where appropriate, ensuring population estimates aligned with true observational intent.

### Consolidation of Survey Method Descriptions

**Problem Identified:** Survey methods were entered using diverse, inconsistent phrasing and included several encoding errors. Common issues included stray characters, inconsistent terminology, and overlapping or unclear method descriptions.

**Rationale for Action:** To analyze trends in data collection techniques or assess biases associated with particular methods, survey types must be categorized clearly. All unique method entries were cleaned and then grouped into a consistent set of categories (e.g., "Aerial", "Boat", "Ground", "Drone", etc.). This categorization enables meaningful comparison across surveys.

**Resolution Summary:** Some custom values that began with "Other -" were preserved as it is to retain context-specific detail provided by observers, while standardizing the remainder for analytical consistency.

## General Quality Enhancements

**Problem Identified:** The dataset contained miscellaneous formatting, naming, and missing data issues that could collectively interfere with analysis.

**Rationale for Action:** Ensuring data hygiene across all fields contributes to the integrity and reliability of the entire dataset. Even small inconsistencies can result in parsing errors or misinterpretations.

**Resolution Summary:** Observer names were reviewed for consistency. Trailing spaces, redundant punctuation, and embedded formatting characters were removed. Placeholder values such as "Unknown" were assigned where essential fields lacked information. Erroneous entries were omitted from the final cleaned version.

Through the targeted cleaning and transformation procedures described above, the dataset has been substantially improved in terms of accuracy, consistency, and completeness. The process enabled the resolution of critical issues such as ambiguous species identification, misaligned records, and data entry errors. With standardized formats and clearly defined structures, the dataset is now optimized for effective analysis, visualization, and decision-making within the context of bird conservation in the Atlantic Flyway. This structured approach lays a solid foundation for the project's subsequent phases involving advanced analysis and interactive dashboards.

## Flagging of Records for Expert Review

As part of the Atlantic Flyway dataset preparation process, an extensive cleaning procedure was performed to correct and standardize entries where possible. However, several fields required expert interpretation or additional context that could not be resolved solely through automated cleaning. These areas often involved incomplete, ambiguous, or inconsistent information that could best be reviewed and addressed by the Atlantic Flyway Council members responsible for data collection in their respective states.

To facilitate this collaborative review process, a flagging criteria was implemented. This system systematically identified and marked records with potential data quality concerns. The goal of flagging was not to discard records, but to highlight them for further attention and possible revision by state-level domain experts.

### Flagging Categories

* Unknown Units Counted - Records with `UnitCounted = Unknown` were flagged due to ambiguity in the measurement unit, which can affect population interpretation.
* Blank Units Counted - Records lacking any value in the `UnitCounted` field were separately flagged, indicating missing information critical for contextualizing the population data.
* Blank Population Estimates - Records with no value in the `PopulationEstimate` field were flagged, as these represent gaps in the core quantitative data used for trend and abundance analysis.
* Adults and Individuals Counted - Records that specified `UnitCounted = Adults` or `UnitCounted = Individuals` were flagged for clarity, as these can vary in interpretation across species and survey protocols. Flagging ensures that these records can be verified for accuracy and relevance.
* Non-breeding Season Records - Records collected during the non-breeding season (October through March) were flagged. These surveys may reflect different ecological behaviors or observer intents and therefore warrant expert review before integration into breeding season analyses.
* Unknown Species Codes - Records using unknown or generic AOU codes such as `UNCO`, `UNGU`, or `UNHE` (e.g., 'Unknown Cormorants', 'Unknown Gulls') were flagged for identification confirmation or refinement.

### Updates Tracking Fields

To support a transparent and auditable review process, several reviewer-assistive fields were incorporated into the dataset:

* ReviewStatus - Indicates whether a flagged record has been reviewed.
* Decision - A dropdown menu allowing reviewers to select an outcome (`Include`, `Edit`, or `Discard`).
* FlagReason - Automatically populated to describe why a record was flagged.
* OriginalValue and UpdatedValue - Capture data before and after reviewer edits, enabling version tracking.
* ReviewDate and Reviewer - Allow documentation of the review timeline and the responsible personnel.

# Data Visualization and Dashboard Development

To translate complex ecological data into accessible insights, a series of interactive dashboards were developed using Tableau Public. These dashboards serve as dynamic visual tools to explore cleaned survey data of colonial waterbird populations across the Atlantic Flyway. The visualization component of this project was integral in helping biologists to understand species trends, distribution patterns, and survey methodologies across regions and years. Designed based on the feedback received from partners and challenges identified during exploratory data analysis, each dashboard was developed to offer unique analytical perspectives and intuitive navigation. The visualizations complement the cleaned dataset by offering clear summaries of trends and making the information more actionable for conservation planning.

## Dashboard 1: Atlantic Flyway Avian Population Overview

**Purpose and Overview**

This dashboard is designed to analyse and explore the avian population trends across the Atlantic Flyway region, which includes multiple U.S. states and Canadian provinces. It brings together various interactive components that allow users to investigate bird species population, colony distribution, survey trends, and survey methods over time.

[Tableau Public Link Dashboard1](https://public.tableau.com/shared/MMJ7QFD5Q?:display_count=n&:origin=viz_share_link)

A screenshot of a dashboard

AI-generated content may be incorrect.

Figure : Interactive Tableau Dashboard 1 for Atlantic Flyway Avian Population Overview

**Key Features and Interactions:**

* **State/Province Map (Top Left):**
  + Acts as the primary filter for the entire dashboard.
  + Users start by selecting a specific state or province to focus their analysis.

A map of canada with different colored states

AI-generated content may be incorrect.

Figure : State/Province Selector Map

* **Year Filter:**
  + Located beside the map.
  + Displays available years based on the selected region.
  + Filters all visualizations to show data for the chosen year.

A screenshot of a map

Description automatically generated

Figure : Filter by Year to View Historical Survey Trends for Selected State

* **Avian Population Table:**
  + Lists all bird species observed in the selected state and year.
  + Displays estimated population for each species in descending order.
  + Each row acts as a filter—clicking a species updates related charts.

A screenshot of a computer

AI-generated content may be incorrect.

Figure : Year-Based Species Filtering Table

* **Species Population Trend Line (Top Right):**
  + Shows how the population of the selected species has changed over time in the chosen state.
  + The x-axis (year) is clickable; selecting a data point filters the visuals below based on that year.

A graph showing the growth of the year

AI-generated content may be incorrect.

Figure : Species Population Trend Across Years

* **Survey Sites Graph (Bottom Centre):**
  + Displays the number of colonies survey was conducted over the years in the selected state for selected species.
  + Helps correlate survey frequency with population estimates.

A graph with numbers and a bar

AI-generated content may be incorrect.

Figure : Annual Survey Effort Visualization

* **Colony Distribution Map (Bottom Left):**
  + Visualizes how colony locations of the selected species have shifted geographically over time.
  + Includes a year slider and play controls to animate changes forward or backward.

A map of land with many points

AI-generated content may be incorrect.

Figure : Animated Colony Distribution Over Time

* **Observed Months Tree map (Bottom Centre):**
  + Highlights the months during which surveys for the selected species were conducted in the selected year.
  + Useful for identifying seasonal patterns in data collection.

A close-up of an orange sign

AI-generated content may be incorrect.

Figure : Monthly Observations and Survey Methods for Laughing Gull (2018)

* **Survey Methods Treemap (Bottom Right):**
  + Shows the type of survey method used (e.g., Aerial, Ground) for the selected species and year.
  + Helps explain potential variation in population counts based on methodology.

A yellow rectangular object with blue text

AI-generated content may be incorrect.

Figure : Survey Methods for Laughing Gull in 2018

**Dashboard Logic and Filter Flow**

* User selects a state from the Atlantic Flyway map.
* Year filter updates dynamically based on available data for that state.
* Species table displays population estimates for that specific year and location.
* Clicking a species row filters the trend and detail views for that species.
* Clicking a year on the species trend line updates both the Survey Method and Observed Months visualizations below.
* Colony map animation helps visualize spatial movement of species over time.

***Note:*** If a new filter is not selected, some charts (like Survey Method or Months) may briefly retain the previously selected species or year’s data.

**Conclusion**

Dashboard 1 provides a comprehensive and interactive visual exploration of avian population data across the Atlantic Flyway. From overall colony trends to individual species breakdowns, and methodological impacts. The tool serves both as an analytical aid and a visual storytelling platform for ecological researchers and public health or conservation agencies.

## Dashboard 2: Expanded View of Colony Distribution and Species Presence

**Motivation and Purpose**

In response to partner feedback requesting a more expansive and detailed view of colony distributions, Dashboard 2 was developed to focus on both **spatial location** and **species composition** at the colony level. The goal was to allow biologists and researchers to explore colony size and inter-species presence more clearly within a selected year and location. By presenting a larger geographic canvas, this dashboard supports a deeper ecological understanding of regional nesting sites and species diversity.

[Tableau Public Link Dashboard 2](https://public.tableau.com/app/profile/shreya.rajani.shankar/viz/ColonyDistributionasperpopulationsizeperyear/Dashboard1)

A screenshot of a computer screen

Description automatically generated

Figure : Interactive Tableau Dashboard 2 for Colony Distribution

**Key Features and Interactions:**

* **Colony Distribution Map (Left Side):**
* Displays all colony locations within the selected state for the chosen species and year.
* Each circle is sized proportionally based on the population estimate, enabling a quick visual assessment of colony size and density.
* Supports interactive filtering through three dropdown controls:
  + **State** – Filters colonies based on the selected U.S. state or Canadian province.
  + **Year** – Narrows data to a specific survey year.
  + **Species** – Filters observations by bird species for focused analysis.
* Aids in identifying key nesting sites, spatial population trends, and regional species concentrations.

A screenshot of a computer screen

Description automatically generated

Figure : Map Showing Colony Distribution of Selected Species in that State and year, with Tooltip Displaying Location, Species, Colony ID, and Population Estimate

**Dashboard Logic and Filter Flow**

* **Users start by selecting a state, year, and species** from the dropdown filters on the right side of the dashboard. These filters help narrow the data to focus on a specific location and time.
* **The colony map updates based on the selected filters**, showing colony locations for the chosen species in that state and year. Each circle on the map represents a colony, with its size based on the population estimate.
* **Hovering over a colony reveals more details**, such as the colony name, coordinates, year, and population count. This helps users better understand the distribution and significance of each site.

This simple flow allows users to interact with the data easily and supports clear insights for ecological research and conservation efforts.

**Limitations and Design Motivation for Enhancement**

While Dashboard 2 effectively visualized colony distribution and species-level insights, it presented a key limitation in high-density areas. Colonies located close to one another appeared overlapped on the map, as population size was directly represented by circle size. This often led to **visual clutter** and made it difficult for users to distinguish between neighboring colonies or interpret their true population values accurately.

To address this issue and enhance clarity, **Dashboard 3** was designed—introducing refined spatial representation and interaction-based filtering to improve readability in densely populated regions.

## Dashboard 3: Improved Species-Level Colony Visualization and Exploration

**Motivation and Purpose**

**Dashboard 3** was developed as an improved version of **Dashboard 2** to address challenges in viewing **colony locations** in areas where many colonies are **closely spaced**. In Dashboard 2, **colony size** was shown using different-sized circles based on **population**, which caused the circles to **overlap** on the map. This made it hard for users to clearly see and understand the data in **densely populated zones**. To solve this issue, Dashboard 3 uses **fixed-size points** for colony locations, making the map **cleaner** and **easier to read**. It also **separates the display of population size** from the map, which helps reduce confusion. This dashboard allows users to **explore where colonies are located** for different species, see **which species are found together** in a colony, and view the **top 10 colonies** with the highest populations for a **selected species** in a specific year. These updates make it easier for **biologists and researchers** to explore the data, understand **colony patterns**, and make **informed decisions** for **conservation planning** across the **Atlantic Flyway**.

[Tableau Public Link Dashboard 3](https://public.tableau.com/app/profile/vivek.patil.paidigumal/viz/shared/6Q5TYBS3R)

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure : Interactive Tableau Dashboard 3 for Colony Locations, Species Composition, and Population Insights

**Key Features and Interactions:**

* **Colony Locations Map (Left Side):**
* Features a **species dropdown filter** populated with all bird species observed within the Atlantic Flyway.
* Users begin by selecting a **specific species** and **year** using the filter controls.
* The map then displays all **colony sites** where the selected species was observed during that year.
* Each **colony marker** on the map acts as an interactive selector. When a colony is clicked, it dynamically updates the bar chart to show the **list of other species** coexisting at that site alongside the selected species.
* This map enables users to explore **species-specific spatial distribution** and understand **multi-species presence** within individual colonies.

A screenshot of a computer

Description automatically generated

Figure : Colony Distribution Map with Interactive Species and Year Filters

* **Species in Selected Colony – Bar Chart (Top Right):**
* Displays all bird species found in the selected colony for the chosen year.
* Each bar represents a species, with the height showing how many birds were recorded.
* Clicking on a bar for any species will update the dashboard to show the **Top 10 colonies** where that species had the highest population in the same year.

A graph with numbers and a bar

AI-generated content may be incorrect.

Figure : Bar Chart Displaying Species Composition and Population at a Selected Colony Site

* **Top 10 Colonies for Selected Species – Bubble Chart (Bottom Right):**
* Highlights the **top 10 colonies** with the highest population estimates for the selected species in the chosen year.
* Bubbles are sized based on population and labelled with colony name, year, state, and estimates.
* Helps identify key habitats and concentration zones for that species.

A chart of the top 10 colonies for laughing gull in 2007

AI-generated content may be incorrect.

Figure : Bubble Chart Displaying Top 10 Colonies by Population for Selected Species in 2007

**Dashboard Logic and Filter Flow**

* **The user starts** by selecting a species and year using the dropdown and slider on the map.
* **Colonies are plotted** on the map showing where that species was observed that year.
* **Clicking a colony** on the map updates the bar chart to show all species found at that site.
* **Clicking a species bar** in the chart filters the bubble chart to show the top 10 colonies with the highest population counts for that species.

**Conclusion**

Dashboard 3 offers a clearer and more effective way for biologists to explore colony-level bird population data. By separating the location map from population size visuals, it avoids clutter and improves readability—especially in areas with many nearby colonies. The interactive filters and drill-down features allow biologists to view where species are found, what other species share those sites, and which colonies have the largest populations. This simplified, focused design supports better understanding of species distribution and helps guide more accurate conservation decisions across the Atlantic Flyway.

**Why Tableau Desktop Was Chosen:**

* Tableau Desktop allows multi-sheet interactivity, which was essential for building a responsive and flexible user experience.
* It supports dual-axis charts, dynamic filters, and animated maps, which were key to visualizing both spatial and temporal data.
* Tableau provides a no-code interface, allowing complex calculated fields, filter actions, and visuals to be created efficiently.
* The final dashboard was published to Tableau Public, making it easily accessible for partners and stakeholders without requiring paid licenses.

## Summary of Visualization Outcomes

The interactive dashboards developed in this project provide a valuable way to explore and understand colonial waterbird population data. **Dashboard 1** offers a broad overview of species trends, survey activity, and population estimates across the Atlantic Flyway. **Dashboard 2** focuses on colony-level distribution by species and population size, allowing for clearer interpretation across states and years. **Dashboard 3** enables detailed analysis of species diversity at specific colonies and highlights the top population sites for each species.

Together, these dashboards help **Biologists** view trends, assess survey efforts, and identify key habitats. Features like filtering, mapping, and species-specific views make it easier to compare regions, detect changes, and support ongoing data review. These visual tools enhance collaboration and provide a strong foundation for long-term ecological monitoring and conservation planning across the Atlantic Flyway.

# Findings

The analysis of the Atlantic Flyway Colonial Waterbird (CWB) dataset uncovered several critical data quality issues that significantly reduced the dataset’s usability for long-term conservation and ecological trend analysis. Key issues included inconsistent data formats, outdated species codes, non-Atlantic flyway region entries and missing metadata in fields such as population estimates and units counted. The team addressed these challenges through a structured data cleaning and standardization, automated geolocation and filling in missing data based on logical rules.

As a result, the cleaned dataset now makes it easier to explore the data reliably of colony distribution and species trends. Notably, our dashboard testing using Virginia’s Laughing Gull data suggested a genuine decline in the population across multiple survey years. This was validated using evidence such as breeding season alignment, spatial colony consistency, and normalized survey effort.

These findings underscore the value of clean historical records in detecting long-term ecological patterns. Furthermore, our Tableau dashboards offer partners an intuitive interface to explore colony-specific metrics, seasonal trends, and survey metadata, improving accessibility and decision-making for biologists and conservation planners​.

# Summary

The project successfully delivered a cleaned, validated, and standardized dataset of colonial waterbird surveys spanning over a century across the Atlantic Flyway. Our efforts resolved missing or inconsistent data entries, updated old species codes using modern ornithological taxonomy and flagged ambiguous records for expert review. In parallel, we developed two interactive Tableau dashboards to visualize key metrics—population trends, colony distribution, survey effort and species-level insights.

Our work demonstrated that historical survey records, once cleaned and standardized, provide substantial value for understanding population trends and guiding conservation strategies. For example, declines in certain species like the Laughing Gull could now be substantiated with data-driven evidence. The outputs of this project serve both immediate analytic needs and broader conservation objectives by supporting informed management decisions, ecological forecasting, and planning for adaptive survey methods in the future​.

# Future Work

While significant steps were made in cleaning and visualizing historical CWB data, several essential next steps are proposed to amplify the long-term impact of this project:

1. **Expert Collaboration**  
   The current flagged records require resolution by state subject-matter experts. These include records with missing population values, ambiguous species codes and non-breeding season entries. Working directly with these experts will help finalize the dataset for scientific use.
2. **Data Integration with 2023–2024 Survey Results**  
   Once cleaned, the historical dataset should be merged with results from the 2023–2024 Colonial Waterbird Survey. This integration will allow for robust longitudinal analyses spanning over 150 years and enable a comparative study of population dynamics at regional and Flyway scales.
3. **Advanced Trend Analysis and Predictive Modeling**  
   Future work should incorporate advanced statistical techniques and modeling to forecast colony shifts and species vulnerabilities under climate stressors. These efforts would directly support policy decisions on habitat conservation and adaptive management.
4. **Standardized Review Framework**  
   A formal protocol for tracking the review and resolution of flagged records should be implemented. This will ensure transparency, auditability, and accountability in expert-led reviews.
5. **Scalability and Public Engagement**  
   The current visualization tools should be expanded for broader public access and enhanced usability. Engaging citizen scientists in long-term monitoring and flagging could further increase the dataset's reliability and outreach potential.

These next steps ensure that the foundational work completed in this project can be scaled, extended, and used as a decision-support tool for biodiversity conservation along the Atlantic Flyway.​

**This page intentionally left blank**

Appendix

Appendix A: Domain Background

**Introduction**

The Atlantic Flyway is one of the most important migratory routes in the world, stretching from the Arctic tundra of northern Canada all the way down to the Caribbean and South America. Every year, millions of birds rely on this corridor to migrate between their breeding and wintering grounds. Among these travelers are **colonial waterbirds (CWBs)**—species such as herons, egrets, gulls, pelicans, and terns that breed in colonies, often in coastal wetlands, estuaries, and barrier islands. These birds are not only fascinating to observe but also play critical roles in maintaining healthy ecosystems.

Unfortunately, many colonial waterbird populations have been declining across the Atlantic Flyway in recent decades. Factors such as habitat destruction, climate change, human disturbance, and predation are threatening their survival. At the same time, effective conservation efforts require accurate and reliable data to track population trends, monitor habitat changes, and assess risks. Yet, historical datasets are often incomplete, inconsistent, or poorly documented, making it challenging to integrate past and present data for meaningful conservation decisions.

This background section provides an overview of the ecological importance of colonial waterbirds, the threats they face, challenges in data management, and ongoing efforts to improve conservation through better data and monitoring.

**The Role of Colonial Waterbirds in Ecosystems**

Colonial waterbirds contribute to ecosystem health in multiple ways. Many species, including the **Great Blue Heron (*Ardea herodias*)** and **Royal Tern (*Thalasseus maximus*)**, serve as predators that regulate populations of fish, crustaceans, and invertebrates. By keeping prey populations in check, they help maintain balance in aquatic food webs and prevent the collapse of species interactions within their habitats.

Beyond their role as predators, these birds act as **bioindicators**—their population health reflects the condition of the habitats they rely on. When waterbird populations decline, it often signals broader environmental problems such as pollution, wetland degradation, or declining fish stocks. Their presence or absence offers valuable clues about ecosystem quality and the cumulative effects of multiple environmental stressors.

Additionally, large nesting colonies create unique habitats that benefit other species. Nesting activity contributes to the formation of **microhabitats** for insects, plants, and other birds. Some nesting colonies become biodiversity hotspots, supporting a range of organisms beyond the waterbirds themselves. Their guano (droppings) can also enrich nutrient cycles within these ecosystems, providing fertilizer for coastal vegetation that stabilizes shorelines.

These ecological contributions make CWBs not just species of interest but **keystone components** of coastal ecosystems. Their loss could trigger cascading effects across the food web, leading to reduced ecosystem resilience and function.

**Population Declines and Threats**

Long-term surveys along the Atlantic Flyway have documented alarming declines in many colonial waterbird species. For example:

* **Cattle Egret** populations have dropped by over **98%** from 1,459 pairs in 1993 to just 25 pairs in 2023.
* **Glossy Ibis** have declined by **82%**, falling from 1,008 pairs to 182 pairs.
* **Laughing Gull** numbers have decreased by **72%** over the same period.
* **Herring Gull** populations have shrunk by **77%.**

While some species, like the **White Ibis** and **Brown Pelican**, initially showed population increases, more recent data suggest even these species are facing declines as environmental pressures mount.

Several interrelated threats are driving these trends:

* **Habitat Loss:** Coastal development, sea-level rise, and erosion have reduced nesting and foraging habitats. Marshes, islands, and wetlands are disappearing, leaving fewer safe breeding sites. Restoration efforts in some regions, such as creating artificial nesting platforms, have provided partial solutions, but these interventions are not always enough to offset habitat loss at scale.
* **Extreme Weather:** Hurricanes and storms frequently destroy nests and breeding colonies. Rising temperatures and altered weather patterns also affect the timing of migration and nesting, creating mismatches between breeding cycles and food availability. The increasing frequency of intense storms due to climate change adds further risk to already vulnerable colonies.
* **Human Disturbance:** Recreational activities, urban expansion, and tourism near nesting areas disrupt breeding colonies. Birds may abandon nests due to noise, lights, or human presence, leading to lower reproductive success. For instance, colonies located near popular beaches or boating areas often experience higher nest failure rates due to accidental or intentional disturbance.
* **Predation:** Expanding human settlements increase access for predators like raccoons, foxes, and feral cats. These predators prey on eggs and chicks, reducing the number of young birds that survive to adulthood. In some colonies, predator exclusion fences and targeted trapping have been used to mitigate predation pressure.
* **Emerging Threats:** Offshore wind development poses additional risks, including **collision mortality** (birds flying into turbines), **habitat displacement**, and disruption of migration routes. While renewable energy offers environmental benefits, its placement near critical migratory corridors raises concerns about unintended impacts on bird populations.

Each of these threats compounds the others, making conservation planning for CWBs a highly complex challenge.

**Data Challenges in Conservation**

To protect colonial waterbirds effectively, conservationists need accurate, standardized, and up-to-date data. Unfortunately, many existing datasets suffer from critical issues:

* **Inconsistent Data Collection:** Different states and organizations have used varying survey methods, counting protocols, and data formats over the years. These inconsistencies make it difficult to compare results across regions or decades.
* **Duplicate Records:** Overlapping surveys conducted by multiple agencies sometimes count the same colonies or nesting sites more than once, leading to inflated population estimates.
* **Missing Metadata:** Many historical records lack important details such as GPS coordinates, observation conditions, or observer identity. Without this metadata, validating and contextualizing the data becomes challenging.
* **Disparate Data Systems:** Data are stored across multiple platforms and formats, with little standardization or integration. This fragmentation limits the ability to conduct broad-scale analyses.

These data challenges hinder efforts to track population trends, assess habitat use, and model conservation scenarios. For example, inconsistent species codes across states make it difficult to merge datasets, and missing location data reduce the precision of habitat modeling efforts.

**Efforts Toward Data Standardization and Integration**

Recognizing these challenges, conservationists and data scientists are working to improve the quality and accessibility of colonial waterbird data:

* **Data Cleaning and Validation:** Teams are reviewing historical records to remove duplicates, correct errors, and standardize species names and codes.
* **Metadata Enhancement:** Missing metadata fields are being supplemented where possible using external sources or expert review.
* **Integration into Central Databases:** Cleaned and validated data are being incorporated into shared platforms such as the **Avian Knowledge Network (AKN)** to facilitate broader analyses and collaboration.
* **Use of Technology:** Tools like GIS mapping, satellite imagery, and AI-based data validation are being explored to enhance monitoring and data quality. For example, machine learning algorithms can help detect anomalies in large datasets, reducing manual effort while improving consistency.

By improving data accuracy and accessibility, conservationists hope to better identify population trends, habitat needs, and emerging threats.

These efforts not only benefit research but also provide a stronger foundation for conservation policy and public engagement. The integration of standardized data enables coordinated action across states, agencies, and conservation organizations.

**Implications for Conservation and Policy**

Reliable data are critical for informing conservation policies and management actions. For colonial waterbirds, data-driven insights can guide decisions such as:

* Prioritizing restoration of key breeding sites
* Establishing protective buffer zones around nesting colonies
* Managing human activities near sensitive habitats
* Identifying safe zones for offshore wind development
* Allocating limited resources to the most vulnerable species or habitats

At the same time, public engagement and education play important roles. Citizen science initiatives, such as volunteer-based bird counts, have contributed valuable data and raised awareness about the plight of colonial waterbirds. These programs also foster community stewardship, encouraging local support for habitat protection measures.

Collaborative efforts between researchers, policymakers, land managers, and community stakeholders will be essential for implementing effective, long-term conservation strategies. By aligning scientific data with policy action and public participation, conservationists can work toward sustaining colonial waterbird populations and their critical habitats across the Atlantic Flyway.

**Conclusion**

Colonial waterbirds are vital to the health and biodiversity of coastal ecosystems along the Atlantic Flyway. Yet they face mounting pressures from habitat loss, climate change, human disturbance, and new industrial threats. Addressing these challenges depends on having accurate, standardized data to guide conservation action.

Ongoing efforts to clean and integrate historical datasets represent critical steps forward. By strengthening data foundations, conservationists can more effectively track population trends, monitor habitat changes, and advocate for policies that ensure the survival of these species for generations to come.

The protection of colonial waterbirds is not just an effort to preserve individual species, but an investment in the health of coastal ecosystems and the services they provide to both wildlife and human communities.

Appendix B: Glossary

|  |  |
| --- | --- |
| Term | Definition |
| Atlantic Flyway | A major migratory corridor used by millions of birds from Canada to South America. |
| Colonial Waterbirds (CWBs) | Birds that nest in dense colonies, including herons, egrets, gulls, pelicans, and terns. |
| Bioindicators | Species whose population trends reflect the health of their ecosystems. |
| Avian Knowledge Network (AKN) | A centralized database for integrating historical and current bird observation data. |
| Metadata | Supplemental information about data, including location, observer, date, and survey conditions. |
| Metadata Standardization | The process of ensuring metadata follow a consistent structure for comparability. |
| Data Cleaning | The process of correcting errors, removing duplicates, and standardizing datasets. |
| Data Validation | Verifying data accuracy, completeness, and consistency to improve reliability. |
| Duplicate Records | Data entries that count the same observation more than once. |
| Habitat Loss | The reduction or destruction of natural areas used by wildlife for breeding or foraging. |
| Offshore Wind Development | Construction of wind turbines in marine environments, potentially impacting migratory birds. |
| Predator Exclusion Fence | A barrier designed to prevent predators from accessing nesting colonies. |
| Buffer Zone | A protected area established around sensitive habitats to reduce human disturbance. |
| Extreme Weather Events | Severe climatic events like hurricanes or storms that disrupt ecosystems and nesting sites. |
| Citizen Science | Public participation in scientific research, such as volunteering for bird counts. |
| GIS Mapping | Using Geographic Information Systems to visualize and analyze spatial data. |
| Satellite Imagery | Images captured from satellites used for monitoring environmental changes and habitats. |
| AI-based Data Validation | The use of artificial intelligence tools to automatically detect errors and inconsistencies. |

Table : Glossary Table

Appendix C: GitHub Repository

Overview

This GitHub repository serves as the central hub for all project documentation, source code, data processing scripts, and dataset management related to the **"Extracting Relevant Records from Two Large Colonial Waterbird Datasets"** project. The repository ensures transparency, collaboration, and streamlined version control throughout the project lifecycle. It is structured to provide a well-documented, professional approach to managing and integrating historical colonial waterbird (CWB) data from the **Atlantic Flyway datasets** into conservation planning efforts.

GitHub Repository Link

🔗 [GitHub Repository: Blue Heron Project](https://github.com/dae690/blue_heron)

GitHub Repository Contents

The repository includes the following key components:

* **Project Overview**: A detailed README file that describes the project’s objectives, scope, and methodology.
* **Client Information**: Provides details about the project's partners, including the **Atlantic Flyway Council (AFC) and the North American Migratory Bird Technical Section (NMBTS)**.
* **Development Approach**: Documents the workflow for processing, validating, and integrating datasets using **Python, R, and SQL-based data pipelines**.
* **Data Processing Scripts**: Contains scripts for **data validation, cleaning, and transformation** before integration into the Avian Knowledge Network (AKN).
* **Datasets & Metadata**: Includes relevant datasets used for **historical waterbird data extraction and validation**.
* **Agile Documentation**: Sprint updates, user stories, and team contributions following **Agile methodologies** for project management.

This repository ensures that all project contributors have access to the latest updates, facilitating efficient collaboration and seamless integration of validated datasets into conservation efforts.

Appendix D: Risks

Sprint 1 Risks

In **Sprint 1**, our team focused on **understanding the project scope, aligning with client expectations, and identifying key dataset challenges**. Through discussions and analysis, we identified several risks related to **scope clarity, metadata gaps, and requirement changes**, which could impact project planning and execution. The risk table for Sprint 1 primarily addressed these challenges to establish a **structured foundation for data processing in Sprint 2**.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk Name | Description | Probability | Impact | Mitigation |
| Scope Misalignment | Unclear project objectives or scope confusion from initial planning. | High | High | Finalize scope with the client and document in project plan. |
| Requirement Changes | Client may alter data standards after planning is complete. | Medium | Medium | Maintain a change log and review scope adjustments regularly. |
| Knowledge Gaps | Limited understanding of project terminology and client requirement. | Medium | Medium | Review reference materials (BirdPOP.org, Journal of Ornithology). |

Table : Sprint 1 Risks

Lessons Learned

* **Clearly Defining the Project Scope Early:** Establishing well-documented objectives with client input helped prevent misalignment.
* **Identifying Critical Metadata Fields Upfront:** Early assessment of **missing or unspecified metadata** reduced uncertainty in planning.
* **Tracking Requirement Changes Systematically:** Maintaining a structured **change log** helped manage evolving client requirements effectively.
* **Bridging Knowledge Gaps Through Continuous Learning:** Regularly reviewing **reference materials** (e.g., BirdPOP.org, Journal of Ornithology) improved the team’s understanding of project terminology and expectations.

Areas for Improvement

* **More Structured Scope Planning:** Some minor inefficiencies could have been avoided with **better initial documentation and alignment on deliverables**.
* **Early Client Engagement for Metadata Clarification:** More **frequent communication with stakeholders** could have helped resolve metadata uncertainties earlier.
* **Proactive Risk Identification:** Addressing potential requirement changes earlier in the sprint would have improved adaptability to client modifications.

Sprint 2 Risks

In **Sprint 2**, our team transitioned from **project scope and planning (Sprint 1) to data exploration and cleaning**. This phase introduced new risks related to **data integrity, inconsistencies, and biases**. We identified key challenges in **data completeness, standardization, and validation** to ensure the dataset is reliable for further analysis. The risk table for Sprint 2 focuses on addressing these issues to maintain **data accuracy and consistency**.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk Name | Description | Probability | Impact | Mitigation |
| Delayed Data Corrections | Any required corrections to missing or incorrect data depend on state-level approvals, which may take an extended period. | High | High | Maintain a log of flagged data and track client responses to implement corrections efficiently when approved. |
| Data Entry Errors from States | Incorrect or duplicate values may be entered by state officials, affecting the quality of analysis. | High | High | Implement a verification step before finalizing data ingestion. |
| AOU Code Changes | The AOU Codes updates annually and may affect data quality. | High | High | Cross-verify with BirdPOP.org and update dataset documentation accordingly. |

Table : Sprint 2 Risks

Lessons Learned

* **Handling Missing Data Early:** Identifying missing fields early in the process allows for timely mitigation through imputation or human review.
* **Standardizing Data Collection Methods:** Establishing clear guidelines for survey methodologies reduces inconsistencies in species observations.
* **Ensuring Data Accuracy:** Implementing **automated validation checks** minimizes the risk of incorrect data entries before integration.
* **Managing Duplicate Records Effectively:** Cross-referencing records based on **location, date, and species** helps eliminate redundancy.

Areas for Improvement

* **Developing a More Structured Data Cleaning Workflow:** A clear step-by-step **data validation and transformation plan** could streamline the process.
* **Early Identification of Biases in Data Collection:** Proactively **assessing sampling biases** and implementing adjustments can improve data reliability.
* **Enhanced Collaboration with Stakeholders:** More **frequent communication** with clients for metadata clarification would have reduced ambiguity in dataset fields.

Sprint 3 Risks

In **Sprint 3**, our team transitioned from **data exploration and cleaning (Sprint 2) to data validation and visualization**. This phase introduced new risks related to **data accuracy, unit inconsistencies, and survey effort variations**, which could impact population estimates and trend analysis. Our focus was on **standardizing units, updating metadata, and validating survey methodologies** to ensure reliable data integration. Additionally, we worked on **mapping Flyway states, trend analysis, and species distribution visualization** to enhance data usability. The **risk table for Sprint 3** highlights key challenges and mitigation strategies aimed at improving **data consistency and analytical accuracy**.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk Name | Description | Probability | Impact | Mitigation |
| Inconsistent Unit Types | The population count in the "UnitCounted" column reported in different units (nests vs. individuals), making it difficult to aggregate and compare across states. | High | High | Standardize the units before aggregating the data. For instance, each nest is counted as two individual units. |
| Visualization Misinterpretation Due to Data Entry Limitations | The "Population Estimates" column shows 0 values for entries collected via Ariel survey methods, due to known data entry limitations—it's difficult to estimate population accurately from aerial views. | Medium | Low | Communicate data limitations transparently in dashboard documentation or presentation. |
| Inconsistent or Untracked Review of Flagged Records | Without a structured tracking system, flagged records sent to state contacts may be inconsistently reviewed, edited without justification, or left unresolved — leading to data integrity issues, untraceable changes, and loss of accountability. | Medium | Low | Implement a centralized review tracking system with fields for Flag Reason, Reviewer, Review Status, Decision (Include/Discard/Edit), Original & Updated Values, and Review Date, enabling transparent review, auditability, and communication between data owners and reviewers. |
| Duplicate Counts from Repeated Surveys | Multiple records for the same species in the same colony and month by the same observer may represent refined or repeated counts, leading to potential overestimation of population. | High | High | Group records by observer, species, colony, and month, and retain only the maximum estimate to avoid inflated totals. |

Table : Sprint 3 Risks

Lessons Learned

* **Understanding the Impact of Unit Standardization:** Converting all population counts to a consistent unit (e.g., individuals instead of nests) ensures accurate aggregation and cross-state comparisons.
* **Recognizing the Effect of Survey Effort Differences:** Identifying variations in survey effort helped the team understand how data collection inconsistencies can impact population estimates.
* **Establishing Clear Data Validation Rules:** Implementing predefined validation checks helps identify and correct inconsistencies in unit types before analysis.
* **Proactively Addressing Data Discrepancies:** Early detection of unit inconsistencies and survey effort differences enabled timely mitigation strategies.

Areas for Improvement

* **Enhancing Documentation for Unit Conversions:** Keeping a structured log of how units are standardized will improve clarity and ensure consistency in future analyses.
* **Developing a Standard Survey Effort Adjustment Framework:** Implementing a predefined methodology for normalizing survey effort across datasets will enhance data reliability.
* **Strengthening Collaboration with Stakeholders:** More direct communication with data providers will help clarify unit definitions and survey methodologies, reducing inconsistencies.
* **Expanding Automated Quality Checks:** Adding more validation rules to flag unit mismatches and survey discrepancies will streamline data cleaning and integration.

Sprint 4 Risks

In Sprint 4, the team concentrated on **enhancing data visualization**, **refining data validation processes**, and **improving dashboard interactivity** in response to evolving client needs. As the project progressed to include more granular insights—such as **colony distribution by year** and **species-specific mapping**—challenges related to **visual clarity**, **data review accountability**, and maintaining **analytical simplicity** began to surface. Issues like **overlapping visual data**, increased **client requests**, and the absence of a system for **tracking data review outcomes** highlighted the need for careful planning. The team prioritized strategies to improve **visualization accuracy**, **usability**, and **data transparency** moving forward.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk Name | Description | Probability | Impact | Mitigation |
| Inconsistent or Untracked Review of Flagged Records | Without structured tracking criteria, flagged records sent to state contacts may be inconsistently reviewed, edited without justification, or left unresolved, leading to data integrity issues, untraceable changes, and loss of accountability. | Medium | Low | Implement centralized review tracking criteria with fields for Flag Reason, Reviewer, Decision (Include/Discard/Edit), Original & Updated Values, and Review Date, enabling transparent review, auditability, and communication between data owners and reviewers. |
| More visualizations to understand the colony distribution | Clients requested to understand the colony distribution map per year according to the size of PopulationEstimate and distribution across states for only species, which makes the existing dashboard more complex. | Medium | Medium | Creating visualizations in two individual sheets to represent the colony distribution map every year as the size of the PoulationEstimate and colony distribution across all the states only for the species. |
| Overlapping Data Points in Colony Distribution Map | While representing species distribution within each colony using circle markers sized by population estimate, multiple species at the same location led to overlapping data points. This caused visual clutter and made it difficult to accurately interpret population values, particularly in high-density areas where multiple species coexist. | Medium | Low | To improve clarity and focus, we limited the display to the top 5 colonies with the highest total population for the selected year. This allowed a cleaner visual presentation of species composition at key sites, while still preserving analytical value. Additional context can be accessed using filters or tooltips for detailed exploration. |

Table : Sprint 4 Risks

Lessons Learned

* **Balancing Detail and Simplicity in Visuals:** Separating complex visualizations into individual sheets helps manage dashboard clarity while still offering in-depth insights.
* **Visual Clarity Matters:** Overlapping visual elements can obscure key patterns, so prioritizing legibility (e.g., limiting visible data points) enhances user comprehension.
* **User Feedback Drives Meaningful Iteration:** Incorporating client feedback into dashboard features helps align the visual output with stakeholder needs.
* **Need for Structured Review Criteria:** Clearly defined review criteria are essential for traceability and consistency in collaborative data validation workflows.

Areas for Improvement

* **Enhance Tooltip Utility:** Improve tooltip content to provide deeper context for overlapping data without requiring separate visuals.
* **Document Visualization Decisions:** Maintain a log of visualization logic and filtering strategies to ensure transparency and reproducibility.
* **Develop Guidelines for Dashboard Scalability:** Create internal guidelines to handle future dashboard expansion while maintaining usability.
* **Standardize Review Workflow:** Finalize and formalize the review tracking process, integrating it with overall data governance procedures.

Sprint 5 Risks

In Sprint 5, our team concentrated on finalizing the written project report, polishing the presentation slides, and conducting rehearsals to prepare for the showcase. As we approached project closure, we encountered risks related to time management, communication clarity, slide-content alignment, and proofreading. The risk table below outlines these challenges and the mitigation strategies used to ensure a successful final delivery.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Time Management | Risk of exceeding allotted time during presentation delivery. | Low | Low | Conduct timed rehearsals, prioritize key points for conciseness. |
| Communication Clarity | Risk of unclear or mumbled speech affecting presentation. | Medium | Low | Practice speaking clearly, provide peer feedback, rehearse transitions. |
| Slide-Content Alignment | Risk of mismatch between slide visuals and spoken content. | Medium | Medium | Review slides with narration, adjust slides to match final script. |
| Report Proofreading | Risk of minor errors or inconsistencies in final written report. | Medium | Medium | Divide proofreading among team members and perform multiple checks. |

Table : Sprint 5 Risks

Lessons Learned

* **Timed Rehearsals Are Key**: Practicing under a timer helped ensure our presentation stayed within the allotted time.
* **Clarity Comes with Practice**: Multiple practice runs improved speaking clarity, pacing, and transitions.
* **Slide-Content Review Reduces Confusion**: Reviewing slides side-by-side with the script helped align visuals and narration.
* **Collaborative Proofreading Catches Errors**: Sharing the proofreading workload improved accuracy and formatting consistency.

Areas for Improvement

* **More Early Practice Sessions**: Starting rehearsals earlier would have reduced last-minute edits and nerves.
* **Incorporating Audience Feedback**: Gathering feedback from peers or mentors outside the team could further enhance clarity and engagement.
* **Dedicated Visual Design Review**: Allocating time to review slide design details (font size, readability) could improve visual polish.

Appendix E: Agile Development

Scrum Framework Team Approach

Our project follows an Agile framework using the Scrum methodology. Each sprint incorporates daily log meetings and continuous feedback from project partners. This approach helped us to adapt and progress in a relatively new project domain.

Scrum was chosen as our primary project management framework due to its ability to accommodate evolving requirements while maintaining structured workflows. While the transition to Agile was smooth in some areas, certain aspects required adaptation as the team adjusted to new collaboration dynamics.

Figure : Sprint project dates.

Sprint 1 Lessons Learned

Sprint 1 focused on **defining the project scope, aligning expectations, and identifying key risks. User Stories** were established through discussions, ensuring project clarity. **YouTrack** streamlined task management, and **daily stand-ups** improved collaboration.

What Worked Well:

* **Clear task tracking** through YouTrack.
* **Regular team meetings** ensured alignment.
* **Proactive risk identification** helped mitigate future issues.

Challenges Faced:

* **Scope refinement** required additional effort.
* **Frequent requirement changes** led to revisions.
* **Metadata definitions** needed further clarification.

Lessons Learned & Areas for Improvement:

* **Early scope finalization** prevents misalignment.
* **Improved documentation processes** enhance efficiency.
* **Proactive partners engagement** ensures clarity.

Sprint 2 Lessons Learned

Sprint 2 focused on **data exploration and initial cleaning**, refining metadata, and addressing inconsistencies. **User Stories** evolved based on client feedback, tackling **missing fields, duplicate records, and observer inconsistencies**.

What Worked Well:

* **Structured dataset review** enabled early issue detection.
* **Efficient task delegation** improved productivity.
* **Initial validation techniques** helped filter incorrect data.

Challenges Faced:

* **Survey inconsistencies** complicated standardization.
* **Metadata gaps required client clarification.**
* **Manual review processes slowed down initial cleaning.**

Lessons Learned & Areas for Improvement:

* **Structured metadata documentation** ensures consistency.
* **Defining a validation workflow** streamlines cleaning.
* **Regular client communication** helps resolve data uncertainties faster.

Sprint 3 Lessons Learned

Sprint 3 focused on data validation and visualization, addressing risks related to data accuracy, unit inconsistencies, and survey effort variations. Key efforts included standardizing units, updating metadata, validating survey methodologies, and enhancing species distribution and trend visualizations to improve data reliability.

What Worked Well:

* **Standardized unit conversions** improved data consistency.
* **Validation checks** helped detect inconsistencies early.
* **Addressing survey effort differences** refined population trends.

Challenges Faced:

* **Inconsistent unit reporting** complicated aggregation.
* **Survey effort variability** affected population estimates.
* **Duplicate survey records** risked inflating counts.

Lessons Learned & Areas for Improvement:

* **Clear unit standardization** ensures accurate comparisons.
* **Recognizing survey effort differences** improves trend analysis.
* **Better documentation of unit conversions** enhances transparency.
* **Developing a survey effort adjustment framework** strengthens reliability.
* **Improved stakeholder communication** reduces data inconsistencies.
* **Expanding automated checks** streamlines validation and cleaning.

Sprint 4 Lessons Learned

**Sprint 4** emphasized the completion of the record flagging process, refinement of cleaning procedures, and preparation of final deliverables for client review. A key focus during this sprint was the enhancement of **interactive dashboards** in Tableau to better communicate colony-level trends and support species-specific analysis. As user stories evolved, the team responded by prioritizing **clarity, transparency, and flexibility** particularly in how species distribution and population estimates were visualized. These efforts helped ensure that both the data and visual tools aligned with partners expectations and ecological insights.

What Worked Well:

* **Comprehensive flagging strategies** enabled consistent identification of data quality concerns.
* **Well-documented workflows** improved clarity and supported easier handoffs to project partners.
* **Enhanced visualizations in Tableau** effectively showcased colony distributions and population changes across states and years.

Challenges Faced:

* **Managing a high number of flagged entries** made documentation time-consuming.
* **Interpreting missing metadata** added complexity to the cleaning process.
* **Coordinating multiple script updates** in response to feedback led to occasional version control challenges.

Lessons Learned & Areas for Improvement:

* **Document processes** in parallel with cleaning activities to avoid last-minute rushes.
* **Build modular scripts** to allow faster and easier revisions when stakeholder requirements change.
* **Introduce early review checkpoints** for flagged records to catch issues sooner.
* **Validate dashboard outputs** with domain experts early on to align interpretations accurately.

Sprint 5 Lessons Learned

**Sprint 5** focused on finalizing the written project report, polishing the showcase presentation slides, and rehearsing for the final delivery. The team collaborated closely to **incorporate partner feedback**, ensure **slide-content alignment**, and maintain clarity under time constraints. **YouTrack** continued to support task management, while **team meetings** focused on refining deliverables for submission.

What Worked Well:

* **Multiple timed rehearsals** helped improve pacing and presentation flow.
* **Collaborative slide reviews** ensured visuals aligned with the spoken script.
* **Shared proofreading tasks** improved report quality and reduced errors.

Challenges Faced:

* **Tight deadlines** added pressure to balance report editing and presentation prep.
* **Aligning final slide edits** with last-minute script adjustments required extra coordination..
* **Maintaining speaking clarity** across rehearsals took repeated practice.

Lessons Learned & Areas for Improvement:

* **Starting rehearsals earlier** allows more time for refining content and delivery.
* **Dividing responsibilities** for report and slide reviews improves workflow efficiency.
* **Practicing with feedback** received from professor and other teams during rehearsals can further boost clarity and engagement.

References

Works Cited

[1] Atlantic Flyway NMBTS, Extracting Relevant Records from Two Large Colonial Waterbird Datasets (Revised), 2025.

[2] B. D. Watts, B. J. Paxton, R. Boettcher, and A. L. Wilke, Status and Distribution of Colonial Waterbirds in Coastal Virginia: 2023 Breeding Season, Center for Conservation Biology Technical Report Series, CCBTR-24-12, William & Mary, Williamsburg, VA, 2024.

[3] U.S. Fish & Wildlife Service, “Migratory Bird Program Administrative Flyways.” Available: <https://www.fws.gov/partner/migratory-bird-program-administrative-flyways>. Accessed: Feb. 2, 2025.

[4] U.S. Fish & Wildlife Service, “Atlantic Flyway Upcoming Meetings and Calendar of Events.” Available: <https://www.fws.gov/media/atlantic-flyway-upcoming-meetings-and-calendar-events>. Accessed: Feb. 2, 2025.

[5] U.S. Fish & Wildlife Service, “Migratory Bird Treaty Act of 1918.” Available: <https://www.fws.gov/law/migratory-bird-treaty-act-1918>. Accessed: Feb. 2, 2025.

[6] Avian Knowledge Network, “Home.” Available: <https://www.avianknowledge.net>. Accessed: Feb. 2, 2025.

[7] Cornell Lab of Ornithology, “All About Birds.” Available: <https://www.allaboutbirds.org/news/archives/>. Accessed: Feb. 2, 2025.

[8] Atlantic Flyway, CCBTR-24-12: Colonial Waterbirds in Coastal Virginia 2024, unpublished manuscript.

[9] The Institute for Bird Populations, "Bird Population Studies and Conservation," **BirdPop.org**, 2024. [Online]. Available: <https://www.birdpop.org/>. Accessed: Mar. 3, 2025.

**This page intentionally left blank**