A close-up of a logo

AI-generated content may be incorrect.

**🌍Project Overview**

This project examines the impact of climate change on migratory bird behavior across decades of global data. Using large-scale datasets from eBird (bird observations), NOAA (climate data), and MODIS (land-cover imagery), we analyzed how environmental variables such as temperature, wind speed, and precipitation affect migration timing, spatial routes, and habitat shifts.

Our goal was to build an end-to-end data analytics pipeline—from data collection to visualization and predictive modeling—demonstrating how data science can inform ecological conservation and climate policy.

**Website:**

<https://sites.google.com/view/ecodataanalytics/home>

1. **Executive summary:**

This capstone project explores the influence of climate change on bird migration patterns using decades of real-world data. As migratory birds depend on environmental cues such as temperature and habitat availability, shifts in climate can significantly disrupt their migration behaviors, impacting biodiversity and conservation efforts. To address this issue, we analyzed large-scale datasets from eBird (bird observations), NOAA (climate data), and MODIS (land cover data) to study changes in bird migration timing, spatial routes, and correlations with environmental variables.

The project was executed in three milestones. Milestone 1 involved data collection, cleaning, and integration. In Milestone 2, we conducted temporal, spatial, and correlation analyses using statistical methods like linear regression and Pearson correlation. In Milestone 3, we applied machine learning models (Random Forest and Gradient Boosting) to predict future migration timing based on climate variables such as temperature, precipitation, wind speed, and humidity.

Key findings include species-specific migration shifts (e.g., Green Kingfisher arriving earlier, Black-and-white Warbler delaying), weak but notable climate correlations (especially with wind speed), and identification of regional migration hotspots with variable habitat loss. While the Random Forest model outperformed Gradient Boosting in prediction accuracy, its low R² indicated room for improvement through enhanced datasets and feature engineering.

Our results contribute to ecological research and offer insights that can aid policy makers and conservationists in protecting migratory routes. The project highlights the importance of combining environmental data analytics with machine learning to address climate-related biodiversity challenges.

**Tables of contents**

1. Executive Summary.....................................................................2
2. Background   
     2.1 Business and Project Background ................................4  
     2.2 Project Goal and Scope ................................................4  
     2.3 Technical Background...................................................4
3. Project Outcomes and Achievements   
     3.1 Overview of Outcomes..................................................5  
     3.2 Milestone 1: Data Collection and Integration................5  
     3.3 Milestone 2: Analytical Modeling.................................6  
     3.4 Milestone 3: Predictive Modeling and Visualization.....7

3.5 Sample Screenshots/Explanations: ...............................8

1. Project Planning and Management Summary   
     4.1 Overview......................................................................14  
     4.2 Project Process by Milestones.....................................14
2. Team Reflection on Project Experience   
     5.1 Success Factors............................................................17  
     5.2 Collaboration and Communication..............................17  
     5.3 Challenges and Areas to Improve................................19
3. Appendix   
     6.1 Project Website URL...................................................19  
     6.2 Project Files List.........................................................19  
     6.3 Progress and Report Files List....................................19
4. **Background:**

**2.1) Business and Project Background:**  
Climate change is altering ecosystems worldwide. Migratory birds, which rely on environmental cues, are especially vulnerable. Mismatches in migration timing due to temperature or habitat changes can affect survival and biodiversity. Our project addresses these concerns by studying long-term bird migration data and correlating them with climate trends to derive actionable insights.

**2.2) Project Goal and Scope:**  
The goal was to investigate how climate variables influence migration patterns using long term data and to develop tools for predicting future trends. The scope included:

* Analyzing temporal migration changes
* Mapping spatial migration hotspots
* Assessing climate correlations
* Modeling predictions using machine learning
  1. **Technical Background:**

We used real-world datasets from eBird, NOAA, and MODIS. Technologies and tools included:

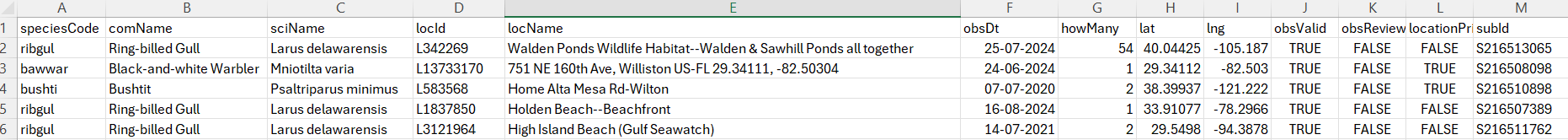
* Python (Pandas, Scikit-learn, Folium, Plotly)
* Google Colab
* DBSCAN clustering
* Random Forest and Gradient Boosting models
* Linear regression and Pearson correlation for statistical analysis

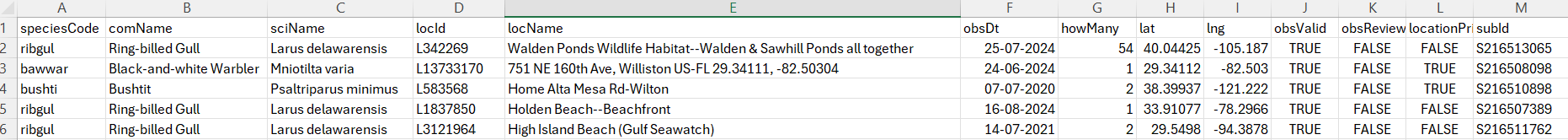
1. **Project outcomes and achievements**
   1. **Overview:**

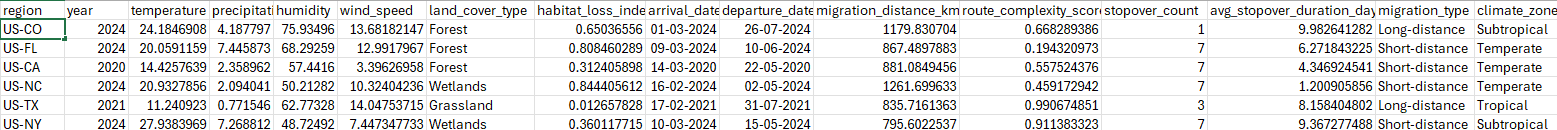
Our team successfully developed a data-driven framework to evaluate how climate factors impact bird migration. The outcomes and deliverables of the project span across three key milestones, encompassing data engineering, advanced statistical analysis, spatial visualization, and predictive modeling. All tasks were aligned with conservation-focused research and ecological awareness.

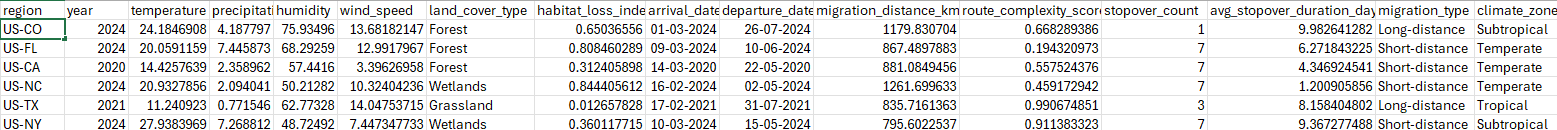
* Cleaned and merged bird, climate, and land cover data.
* Conducted statistical analyses (temporal, spatial, correlation).
* Built predictive models to forecast migration timing.
* Developed interactive maps and dashboards for visualization.
  1. **Milestone 1: Data Collection and Integration**
* **Data Sources:** Acquired datasets from eBird (bird sightings), NOAA GHCN (temperature, precipitation), and MODIS (land cover classification).
* **Data Cleaning:** Standardized date and numerical formats, removed duplicates, filled missing values, and normalized data across all sources.
* **Data Merging:** Joined datasets on date and geographic coordinates to enable spatiotemporal alignment of bird sightings with climate data.
* **Challenges:**
  + Handling large CSV files with thousands of records.
  + Inconsistent formats in bird observations.
  + Temporal misalignment across sources.
* **Deliverables:**
  + Cleaned datasets and master dataset for integrated analysis.
  + Preprocessing code in milestone1\_&2.ipynb.
  + Summary slides outlining data pipeline.
  1. **Milestone 2: Statistical and Spatial Analyses**
* **Temporal Analysis:**
  + Performed linear regression on arrival day-of-year (DOY) to detect shifts over time.
  + Found species-specific trends—e.g., Green Kingfisher showed significantly earlier arrivals.
  + Visuals: Line plots, box plots, violin plots, and histograms.
* **Correlation Analysis:**
  + Applied Pearson correlation between migration DOY and climate variables (Temp, Precip, Wind Speed).
  + Wind speed showed a moderate inverse correlation with some species (e.g., -0.21 for Green Kingfisher).
  + Generated heatmaps and scatter plots to visualize correlation strength and direction.
* **Spatial Analysis:**
  + Used DBSCAN clustering to identify migration hotspots.
  + Plotted bird density and habitat loss using Folium and Plotly Mapbox.
  + Mapped species occurrence across regions with habitat degradation color-coded.
* **Deliverables:**
  + Graphs and statistical outputs.
  + climate\_correlations.csv summarizing species-variable correlations.
  + Interactive HTML maps showing hotspots.
  1. **Milestone 3: Predictive Modeling and Visualization**
* Machine Learning Models:
  + Trained Random Forest and Gradient Boosting models to predict migration DOY.
  + Input features: TMAX, TMIN, PRCP, wind speed, humidity, year.
  + Performance:
    - Random Forest: MSE = 189.29, R² ≈ 0.005, MAE = 7.88 days
    - Gradient Boosting: MSE = 195.04, MAE ≈ 8.2 days
  + Random Forest had slightly better performance and interpretability.
* Feature Importance:
  + Temperature (TMAX and TMIN) were the most influential features.
  + Wind speed and year also had measurable, but lower impact.
* Visualization and Dashboard:
  + Developed visual tools such as scatter plots (Actual vs Predicted DOY), bar charts (feature importance), and temporal overlays.
  + Created an interactive migration dashboard to present arrival estimates and trends geographically.
  + Visuals implemented using Plotly, Matplotlib, and Folium.
* Deliverables:
  + Final notebook: milestone3.ipynb
  + final\_bird\_migration\_dataset.csv
  + Output visuals and model metrics
  + Presentation summarizing modeling process and results
  1. **Sample Screenshots/Explanations:**

**Sample Dataset:**

****





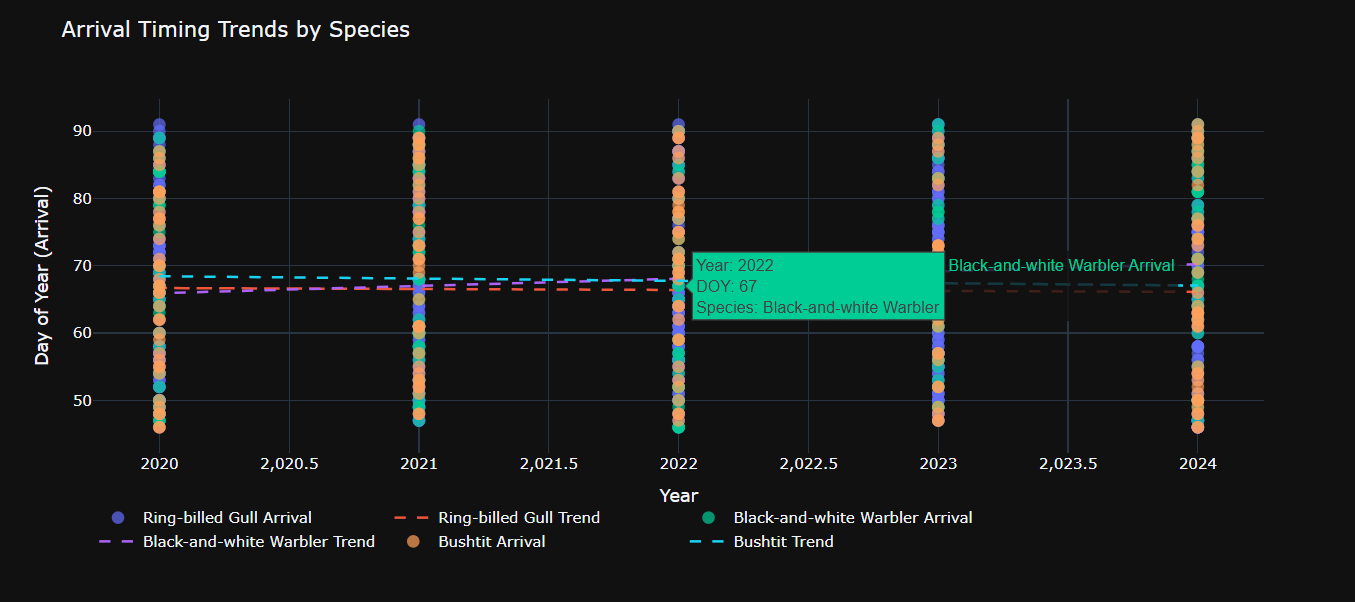


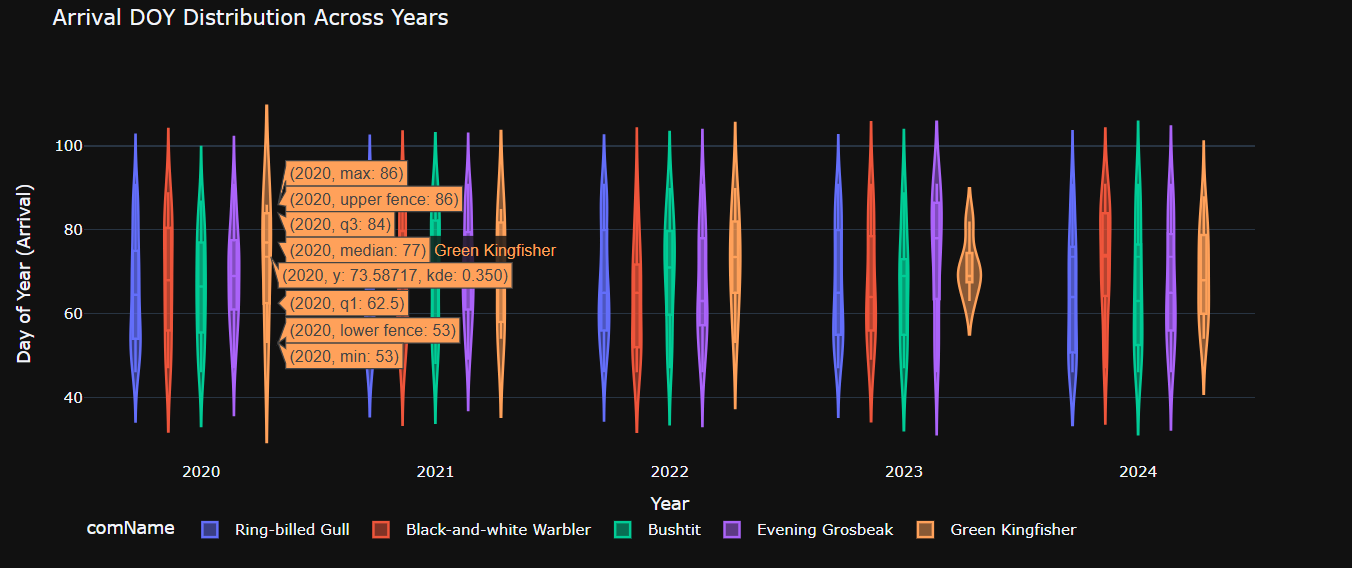
**Dataset Summary:**

The project utilized integrated datasets from eBird (bird sightings), NOAA (climate), and MODIS (land cover). The bird observation data included species names, count, date, and location coordinates. Climate data consisted of temperature, precipitation, humidity, wind speed, land cover type, and habitat loss index. Additional derived fields such as migration distance, stopover count, and route complexity helped in behavior analysis. These datasets were cleaned, standardized, and merged by spatial and temporal keys to form a comprehensive base for our correlation analysis, spatial clustering, and predictive modeling.

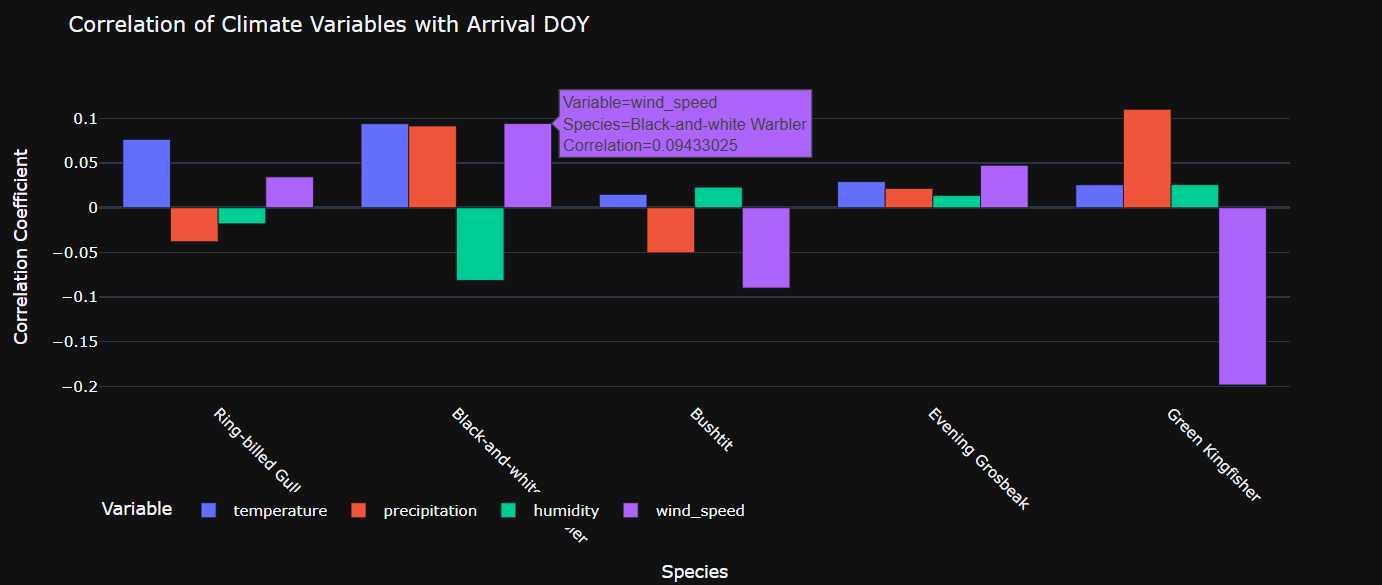
**Temporal Analysis – Migration Trends Visualization**

Using line plots, scatter plots, and violin plots, we visualized migration timing (DOY – Day of Year) across decades. For instance, the Green Kingfisher showed earlier arrival trends, while the Black-and-white Warbler exhibited delays. Linear regression lines provided quantifiable trends per species, supported by violin plots showing annual distribution spread.



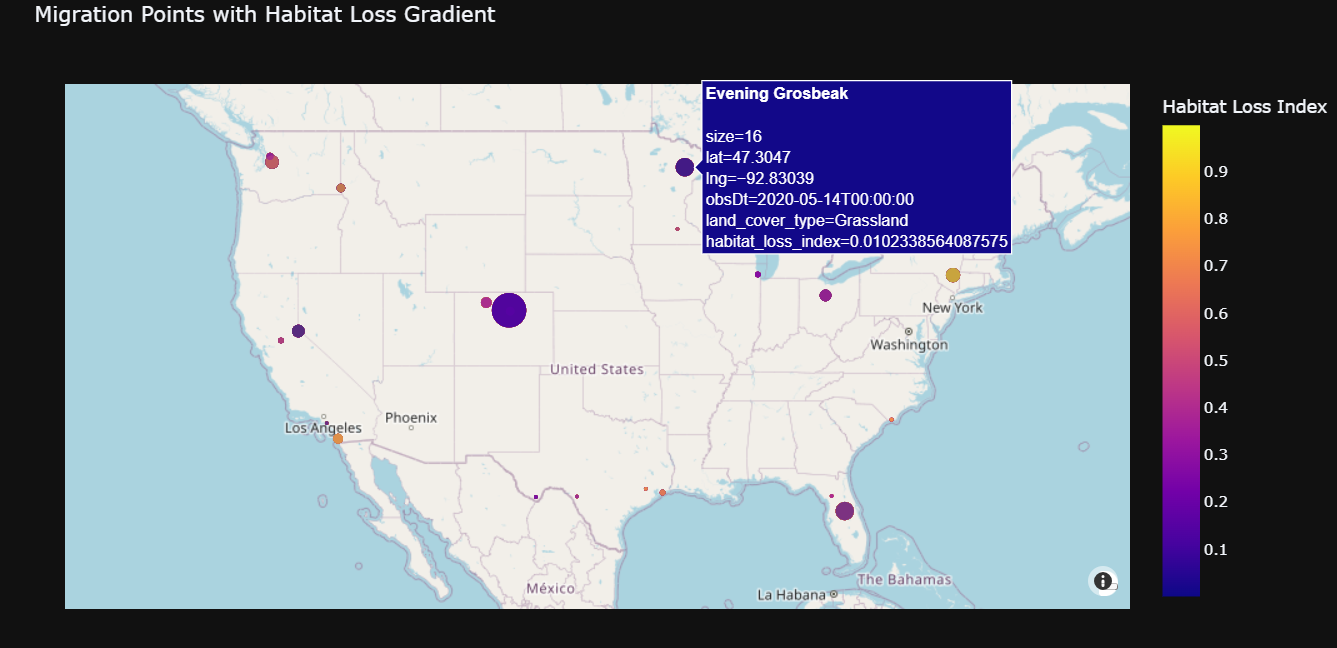


**Correlation Analysis – Climate vs Migration Timing**

****A set of bar charts and heatmaps demonstrated correlations between environmental factors (temperature, precipitation, wind speed) and bird arrival dates. Although most correlations were weak (|r| < 0.2), wind speed showed a modest negative relationship with some species. This visual evidence supported the analytical findings presented in the correlation CSV outputs.

**Spatial Analysis – Mapping Migration Routes and Habitat Stress**

We used Folium and Plotly Mapbox to map species sightings, clustering them based on geographic coordinates and color-coding based on habitat degradation. Red markers indicated high habitat loss zones, while green represented stable areas. DBSCAN clustering highlighted dense migration hubs, particularly across western and eastern U.S. regions.

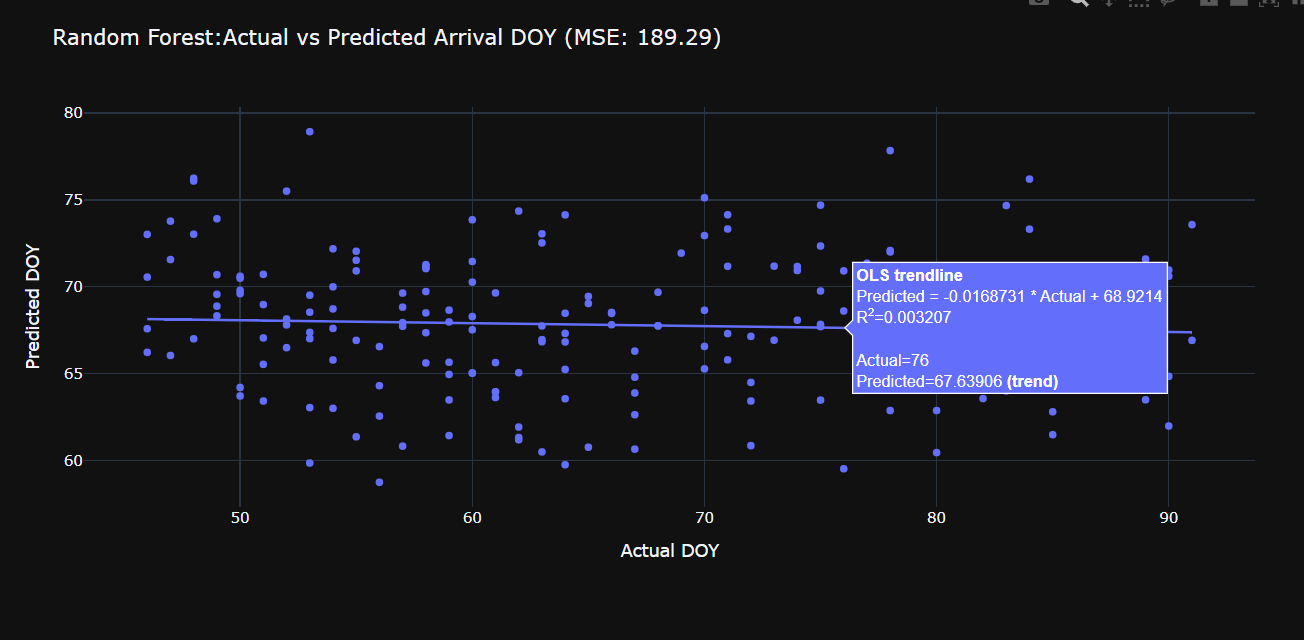


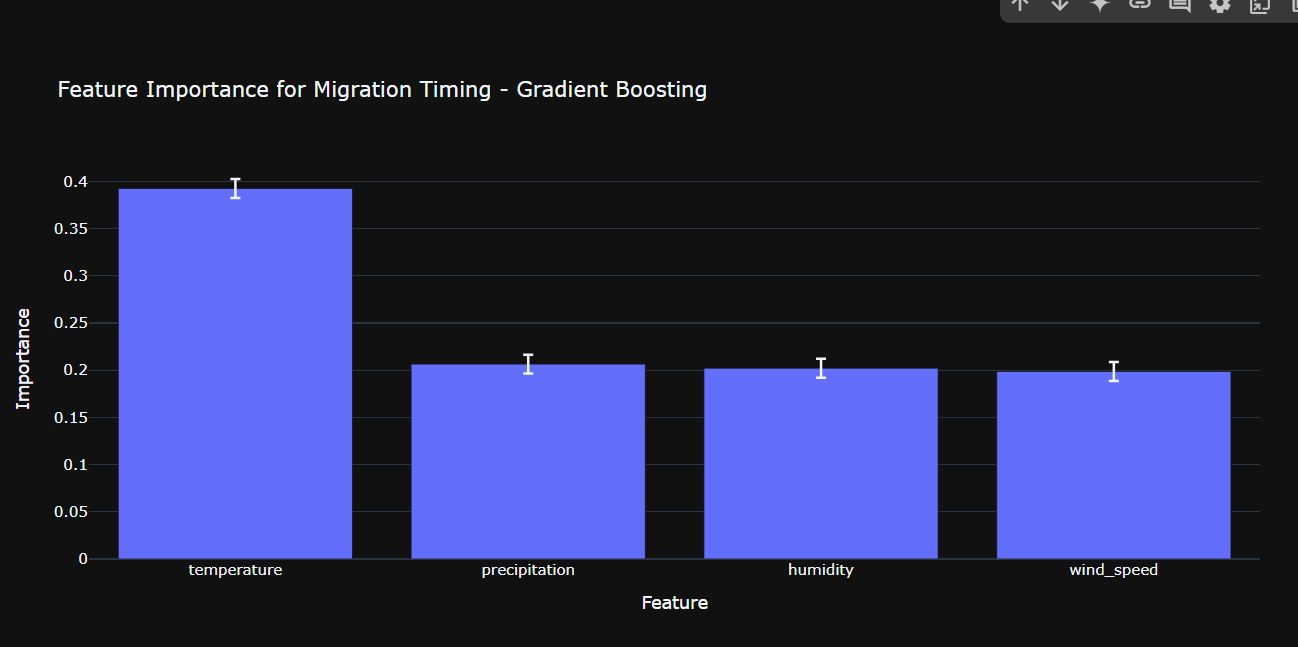
**Predictive Modeling – Machine Learning Outputs**

Random Forest and Gradient Boosting models were trained to predict arrival DOY based on climate features. Scatter plots compared actual vs. predicted values, revealing a better fit for Random Forest (lower MSE). Feature importance charts consistently ranked temperature as the most impactful predictor.

A screen shot of a graph

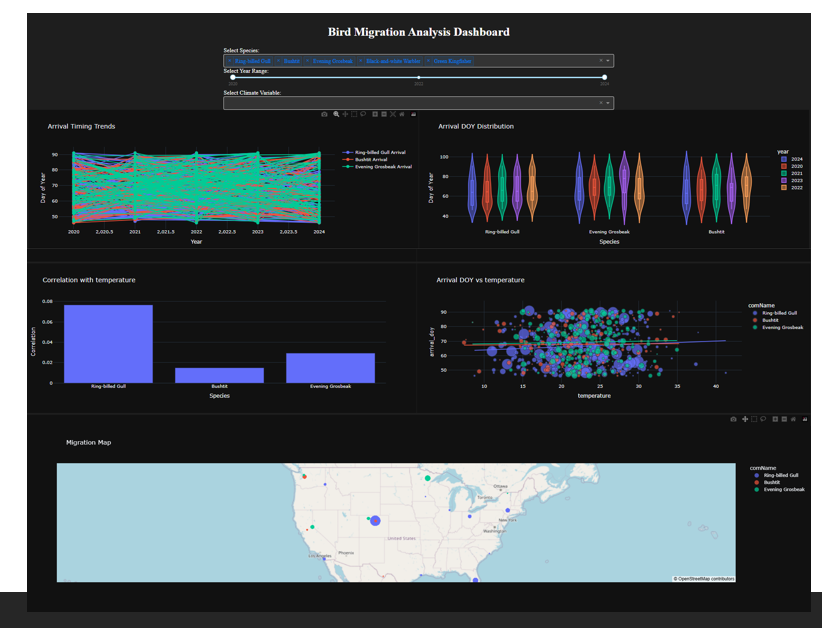
AI-generated content may be incorrect.**A screenshot of a computer screen

AI-generated content may be incorrect.**

**Final Dashboard**

We compiled visualizations into a lightweight interactive dashboard to present:

* Arrival trend lines over time
* Habitat loss maps
* Species-specific behavior summaries

****These visuals collectively made the insights accessible and actionable for conservation and research audiences.

**4) Project planning and management summary**

* 1. **Overview:**

The project was managed using structured planning, milestone-based tracking, and continuous communication. We used a Gantt chart to schedule all major deliverables and assign timelines, which was updated weekly based on progress. Task distribution was organized by individual strengths and experience levels—data preprocessing, modeling, analysis, and documentation were allocated accordingly. Weekly meetings via team and daily team updates via WhatsApp helped ensure accountability and alignment. Project documentation and file sharing were handled through Google Drive.

We followed Agile-inspired practices by breaking the project into sprints (milestones), having review sessions after each milestone, and incorporating instructor feedback. GitHub was used for collaborative code tracking, ensuring all members could contribute to and review notebooks.

* 1. **Project Process / Milestones:**

The project followed three major stages:

* Milestone 1:
  + Objective: Gather and preprocess bird, climate, and land cover data.
  + Activities: API integration, dataset download, cleaning (handling missing values, standardizing formats).
  + Achievements: Created clean and merged datasets; shared dataset schema.
  + Challenges: API inconsistencies, data format mismatches, and large file handling.
* Milestone 2:
  + Objective: Conduct exploratory analysis—temporal, spatial, and correlation.
  + Activities: Built visualizations (scatter, line, violin, heatmap), performed clustering and correlation analysis.
  + Achievements: Uncovered species-specific migration shifts and weak-to-moderate climate correlations.
  + Challenges: Geospatial join issues, interpreting low R² in correlation results.
* Milestone 3:
  + Objective: Predict migration timing using machine learning and compile final deliverables.
  + Activities: Trained Random Forest and Gradient Boosting models; built dashboards; prepared presentations and final report.
  + Achievements: Generated useful model outputs and visual tools; completed final documentation.
  + Challenges: Limited R² in predictions; aligning insights across multiple datasets.

**Workload Summary:**

Based on the Gantt chart and task logs:

**Milestone 1:** ~40 total team hours (data collection, cleaning, merging).

**Milestone 2:** ~50 team hours (analysis, visualizations, correlation metrics)

**Milestone 3:** ~60 team hours (model development, results interpretation, reporting)

**Documentation & Presentation:** ~20 hours (writing, slides, report formatting)

The average workload per member was approximately 34–36 hours due to leading coordination, modeling, and reporting tasks.

**5) Team Reflection on Project Experience:**

* 1. **Project Success Factors:**  
     Our project’s success was driven by clear milestone planning, effective division of tasks, and consistent communication. Leveraging each member’s strengths—for example, coding, visualization, and documentation—ensured timely progress. Having a team lead helped streamline decisions and align tasks with deadlines. Instructor feedback after each milestone also guided us in refining our approach and improving the depth of our analysis.
  2. **Team Collaboration and Communication Experiences:**
* **General Collaboration Experiences:**  
  We collaborated using a hybrid approach—scheduled weekly Zoom calls combined with frequent updates via WhatsApp. Responsibilities were assigned based on expertise, which prevented overlap and ensured focused contributions. One of our key successes was early alignment on team roles and deliverables. However, during the first milestone, some delays occurred due to late clarification on data integration logic, which was later resolved through quick sync calls.
* **Meeting Arrangements and Experiences:**  
  Weekly team meetings (via Microsoft teams) were held every Sunday evening to track milestone completion, raise blockers, and review assigned tasks. These meetings helped maintain consistency and keep everyone accountable. In addition, ad-hoc 1:1 calls were arranged whenever a member needed support or technical clarification.
* **Collaboration System Use:**
  + **Google Drive** was used for centralized file storage (datasets, presentations, logs, reports). It ensured version control and real-time collaboration on documentation.
  + **GitHub** was used for code sharing and collaborative notebook development. This enabled each member to work on separate branches and avoid conflicts.
  + **WhatsApp** provided a fast and informal channel for daily coordination and reminders.
  + **Zoom** was essential for structured meetings, milestone presentations, and collaborative troubleshooting.

Among these, Google Drive and WhatsApp were the most useful in terms of accessibility and ease-of-use, while GitHub ensured our code was well-managed and secure.

* **Other Experiences:**  
  Peer learning played a major role in enhancing our output—members with stronger coding skills helped others debug notebooks, while those with design experience assisted with presentation formatting. One area that occasionally slowed us down was aligning code versions between Colab and GitHub, especially during merge conflicts. We adapted by assigning a “merge coordinator” role when needed.

**Challenges:**

**Technical:**

* Merging datasets with different formats and scales was complex.
* Handling large files in Colab caused performance issues.
* Predictive models had low accuracy due to limited features.
* Geospatial visualizations slowed down due to data density.

**Non-Technical:**

* Coordinating schedules was difficult at times.
* Document formatting lacked consistency early on.
* Managing time during academic workload peaks was challenging.

**5.3) Areas to Improve:**

* Begin documentation and version control earlier.
* Explore more advanced models for prediction.
* Improve storytelling in dashboards for non-technical users.
* Conduct peer reviews after each milestone for early feedback.
* Anticipate risks through early technical feasibility checks.