# Project Scoping Document: Biodiversity Analysis of National Parks

## Project Title:

Biodiversity and Conservation Analysis in U.S. National Parks

## Project Summary:

Limited insights into species distribution hinder effective conservation efforts in national parks. This project aims to analyze data from the National Parks Service to better understand biodiversity trends and assess the conservation status of species across four major national parks. By combining data on species categories, conservation statuses, and observations, the project seeks to provide actionable insights into species protection and park-level biodiversity management. Over 200 species across these parks are potentially underrepresented in conservation strategies due to data limitations.

## Objectives:

1. **Analyze Conservation Status:** Investigate the distribution of conservation statuses among species and identify measurable trends in protection efforts.
2. **Endangered Species Trends:** Examine which categories of species (e.g., Mammals, Birds, Plants) are more likely to be endangered, aiming to improve identification accuracy of endangered species by 15%.
3. **Species Observations:** Explore observation data to identify dominant, rare, and endangered species within each park, and prioritize resources for species at the highest risk.
4. **Statistical Analysis:** Assess the significance of relationships between species categories and conservation statuses using robust statistical methods.
5. **Biodiversity Insights:** Compare species diversity across parks and evaluate factors that influence biodiversity.
6. **Observation Trends by Park:** Identify parks with the highest and lowest total species observations, exploring potential ecological challenges.
7. **Most Commonly Observed Species by Category:** Determine which types of species (e.g., Mammals, Birds) are most frequently observed in each park.
8. **Unique Species in Parks:** Calculate the number of unique species in each park to understand biodiversity richness.
9. **Endangered Species by Park:** Analyze which parks have the highest number of endangered species observations, identifying conservation hotspots.

10. **Correlation Between Observations and Conservation Status:** Investigate if there is a relationship between the number of observations and conservation statuses.
11. **Rare Species Identification:** Identify species with the lowest observation counts and highlight them as rare species requiring monitoring.
12. **Dominant Species by Park:** Determine the most dominant species in each park based on observation counts.

# Research Questions:

1. What is the distribution of species by conservation status?
2. Are certain types of species (e.g., Mammals, Birds) more likely to be endangered?
3. Which parks have the highest and lowest biodiversity based on observations and unique species counts?
4. Which species are most frequently or rarely observed in each park?
5. Is there a significant relationship between species categories and conservation statuses?
6. Which parks have the highest number of endangered species observations?
7. How does species diversity vary across parks?
8. Which species are dominant in each park based on observation counts?

# Data Sources:

1. **`species_info.csv`:**
    - Columns:
        - `category`: Species type (e.g., Mammal, Bird, Plant).
        - `scientific_name`: Scientific name of the species.
        - `common_names`: Common names for the species.
        - `conservation_status`: Status indicating whether the species is Endangered, Threatened, or requires no intervention.
2. **`observations.csv`:**
    - Columns:
        - `scientific_name`: Scientific name of the species (for merging with `species_info.csv`).
        - `park_name`: Name of the national park where the species was observed.
        - `observations`: Number of observations recorded for the species.
    - Granularity: Monthly data collection at the individual park level.

# Deliverables:

1. **Data Exploration:**
    - Initial exploration of datasets to understand their structure and key attributes.
2. **Visualizations:**
    - Charts showing distributions of conservation statuses, species observations, and biodiversity comparisons across parks.
3. **Statistical Analysis:**
    - Chi-Square tests to evaluate relationships between species categories and conservation statuses.
4. **Insights and Conclusions:**
    - Detailed answers to research questions with actionable insights for conservation planning.
5. **Recommendations:**
    - Suggestions for conservation priorities and biodiversity management.
6. **Concrete Interventions:**
    - Recommendations for prioritizing conservation resources based on findings.

# Methodology:

1. **Data Cleaning:**
    - Handle missing or inconsistent values in the datasets (e.g., replacing NaNs in `conservation_status` with "No Intervention").
2. **Data Integration:**
    - Merge the datasets on `scientific_name` to create a comprehensive view of observations and conservation statuses.
3. **Descriptive Statistics:**
    - Use counts, percentages, and aggregations to explore data trends.
4. **Visual Analysis:**
    - Create bar plots, heatmaps, and other visualizations to highlight key patterns.
5. **Statistical Testing:**
    - Apply Chi-Square tests to assess the significance of differences between species categories and conservation statuses.
6. **Ethical Considerations:**
    - Address risks such as revealing sensitive data about species locations that could lead to poaching.
    - Ensure equitable conservation efforts that do not inadvertently favor specific species.
7. **Stakeholder Engagement:**
    - Collaborate with park rangers, conservation biologists, and local communities to validate findings and inform actionable plans.

## Tools and Technologies:

- **Programming Language:** Python
- **Libraries:**
  - `pandas`: Data manipulation and analysis
  - `matplotlib` and `seaborn`: Data visualization
  - `numpy`: Numerical computations
  - `scipy.stats`: Statistical testing
- **Jupyter Notebook:** For organizing and presenting the analysis

## Timeline:

| Task | Estimated Time |
|---|---|
| Data Exploration | 2 days |
| Data Cleaning and Integration | 2 days |
| Descriptive Statistics | 1 day |
| Visualization and Analysis | 3 days |
| Statistical Testing | 2 days |
| Writing Conclusions | 2 days |
| Final Presentation Preparation | 2 days |

## Risks and Challenges:

1. **Data Quality Issues:** Missing or incomplete data might limit analysis.
2. **Generalization:** Observations may not represent actual species abundance due to uneven sampling efforts.
3. **Statistical Assumptions:** Chi-Square tests assume independence, which might not always hold in ecological datasets.

## Expected Outcomes:

- Detailed understanding of biodiversity patterns and conservation needs in the four national parks.
- Insights into which species and parks require focused conservation efforts.
- Visual and statistical evidence to support recommendations for biodiversity management.

---