PROJECT 1 – National Park Species Data Analysis

GROUP 4

Introduction

The dataset provides information on species observed in 15 most visited parks in the USA, the information includes park name, species name, numbers observed, conservation status, etc. This is a portion of the data extracted from a centralized online platform developed by the **U.S. National Park Service (NPS)**. Its main purpose is to provide access to a wide range of natural and cultural resource data and information related to national parks across the United States.

With information on species in the park, surveys can be conducted to see how rich the flora and fauna are, and the number of individuals observed can be used to verify food chain theories or investigate abnormalities in the ecosystem. The conservation status of species can also be investigated to find ways to conserve and restore the ecosystem.

One problem with the dataset is that there is no time series for the recorded data, so it is difficult to detect trends in changes in fauna and flora or to combine the dataset with time-varying data such as CO2 levels. The dataset also does not have the location of the park, so if you want to survey vegetation based on geographic location, you will need to consult another data source.

Question 1: Does the species data from 15 most visited parks reflect the pyramid of number/biomass (Food chain pyramid)?

1. Introduction

Based on studies and observations, food chains indicate that the higher an animal is in the chain, the fewer individuals there are - and vice versa. Our group found that it is possible to use observational data from the dataset to support this theory.

We extracted key information relevant to the question, including the national park name (*ParkName*), species category name (*CategoryName*), common name of the species (*CommonNames*), and the number of observed individuals (*References*). With this data, we were able to create graphs illustrating the population of each species and evaluate whether the results align with the concept of the pyramid of numbers.

2. Approach

To make the chart interactive and visually engaging, we created multiple charts using Dash (Python) and Power BI.

- Scatter Plot: A scatter plot was created to display species versus observed counts, with color mapping by category. This plot allows users to explore abundance trends across categories and detect food chain patterns.

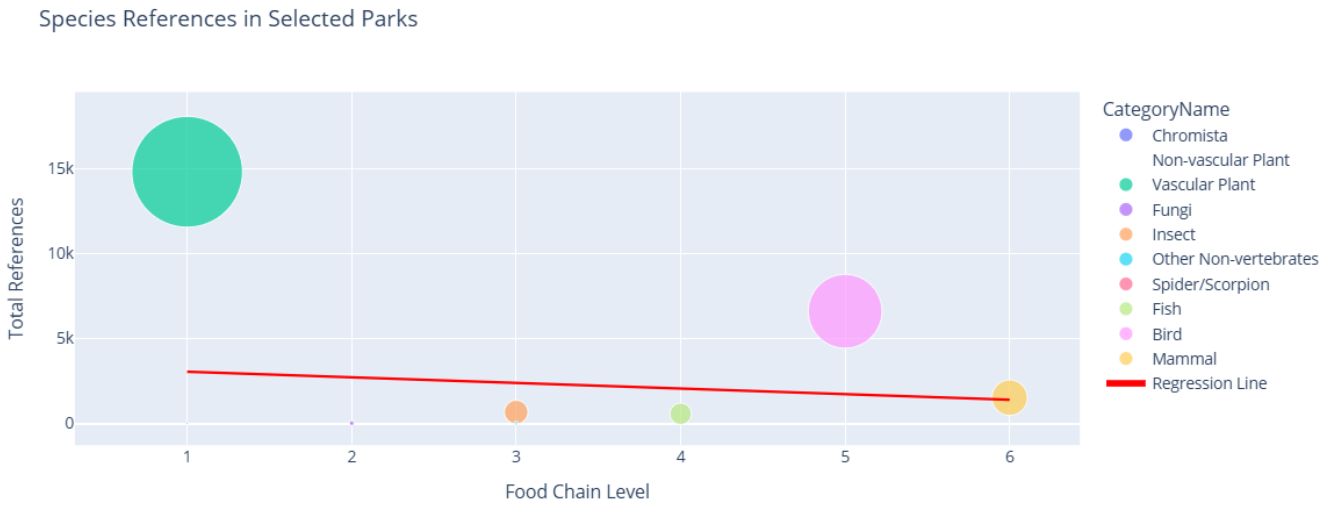
- Bar Chart: We used a bar chart to show the top 10 most observed species, filtered interactively from the scatter plot. This feature enhances interactivity and allows readers to explore each species in more detail and understand its position within the food chain.

3. Analysis

3.1. Scatter plot for food chain pattern

```
# Scatter plot
@app.callback(
    Output('scatter-plot', 'figure'),
    Input('park-dropdown', 'value')
)
def update_scatter(selected_parks):
    data = filter_data(selected_parks)
    data = data.sort_values("FoodChainLevel")

    fig = px.scatter(
        data,
        x='FoodChainLevel',
        y='References',
        title=f"Species References in Selected Parks",
        labels={'FoodChainLevel': 'Food Chain Level', 'References': 'Total References'},
        size='References',
        size_max=60,  # Increase this value to make dots larger
        color='CategoryName',  # Add color
        hover_data=['CategoryName']
    )

    # Compute regression line
    x = data['FoodChainLevel']
    y = data['References']
    if len(x) > 1:  # Ensure enough data points for regression
        slope, intercept = np.polyfit(x, y, 1)
        regression_y = slope * x + intercept

        fig.add_trace(
            go.Scatter(
                x=x, y=regression_y,
                mode='lines',
                name='Regression Line',
                line=dict(color='red'))
        )
    return fig
```

Result:

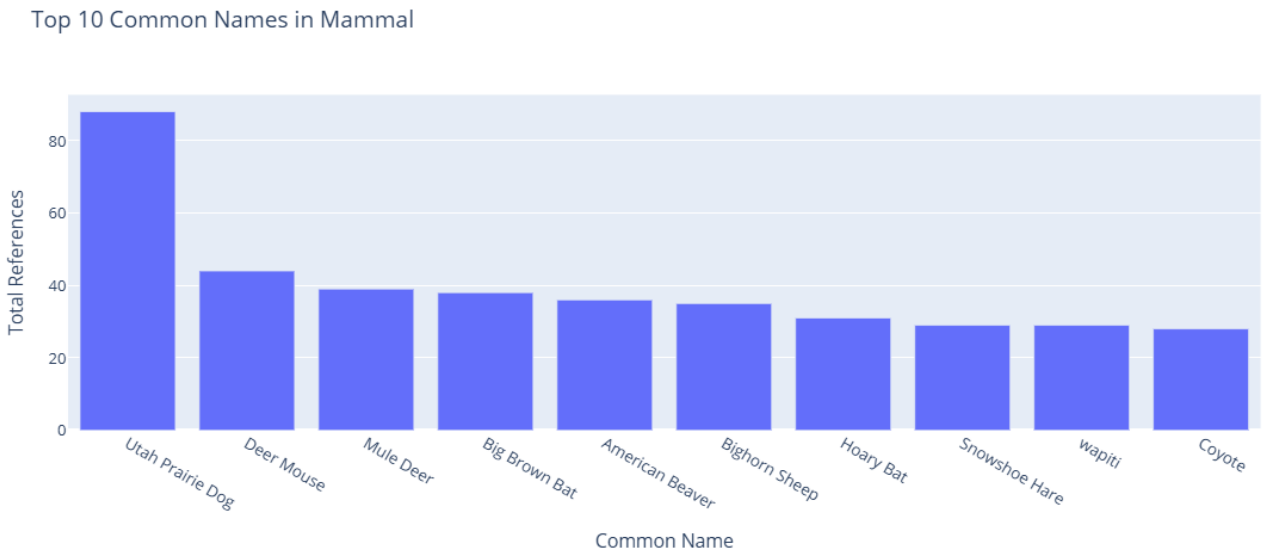


This interactive scatter plot displays each species along the x-axis and their observed counts on the y-axis. Species are color-coded by their biological category to reflect different food chain levels. A regression line overlays the scatter to highlight trends, supporting the hypothesis that species at higher trophic levels tend to appear in fewer numbers. This format allows for easy exploration of the pyramid of numbers theory in national parks.

3.2. Dynamic bar chart for species detail

```
# Bar chart
@app.callback(
    Output('bar-chart', 'figure'),
    Input('scatter-plot', 'clickData')
)
def update_bar_chart(clickData):
    if clickData is None:
        return px.bar(title="Select a Category to see Top 10 Common Names")

    # Access click data
    category_name = clickData['points'][0]['customdata'][0]
    print("selected_category: ", category_name)

    filtered_df = df[df['CategoryName'] == category_name].copy()

    # Check if CommonNames exist
    if 'CommonNames' in filtered_df.columns:
        # Extract first common name from comma-separated values
        filtered_df['PrimaryCommonName'] = filtered_df['CommonNames'].str.split(',').str[0]
    else:
        # If there is no CommonNames
        filtered_df['PrimaryCommonName'] = filtered_df['ScientificName']

    # Get top 10 Common names
    top_common_names = (filtered_df.groupby('PrimaryCommonName')['References'].sum().reset_index().nlargest(10, 'References'))

    fig = px.bar(
        top_common_names,
        x='PrimaryCommonName',
        y='References',
        title=f"Top 10 Common Names in {category_name}",
        labels={'PrimaryCommonName': 'Common Name', 'References': 'Total References'}
    )
    return fig
```

Result:



This bar chart dynamically updates to show the top 10 most observed species in a specific category selected in the scatter plot. It allows users to drill down into species trends and verify if dominant species align with expected food chain patterns. This enhances interactivity and supports understanding of how species abundance varies within categories.

4. Discussion

Some conclusions that can be drawn after observing and analyzing the graph are as follows:

The number of species in the scatter plot reflects the nature of the food chain pyramid: the higher a species is in the food chain, the fewer individuals there are. The regression line drawn from the data also supports this trend. However, there are outliers, such as Yellowstone National Park, which show a species distribution that deviates from the typical food chain pattern. This suggests either a bias in data recording for this park or a unique characteristic in the distribution of flora and fauna there.

When analyzing the most common species in the bar chart, some species are not placed correctly within the food chain. For example, mule deer and deer mice are listed as mammals and, due to a coarse classification system, are ranked at the top of the chain—even though they are herbivores and not typically top-level predators. This highlights the need for more detailed species classification to create a food chain hierarchy that more accurately reflects natural ecosystems.

Question 2: What proportion of species are endangered, threatened, or of special concern?

1. Introduction

Biodiversity is a critical mission of the National Park Service. Understanding which species are at risk and how they are distributed across different parks is essential for effective conservation management. This analysis explores the proportion of species classified as endangered, threatened, or of special concern within the most visited national parks in the United States using the NPS Species dataset. The

dataset contains comprehensive information on 61,119 species documented across 15 major national parks, including their taxonomic classification and conservation status. This question identifying conservation patterns across different species categories and park locations can highlight where conservation efforts are most urgently needed and help evaluate the effectiveness of existing protection measures.

2. Approach

To analyze the proportion of species with vulnerable conservation status, we create three complementary visualizations:

First, we create a stacked bar chart showing the proportion of species in each conservation status category (endangered, threatened, special concern, and other) across different biological taxonomic groups. This type of visualization is ideal for comparing relative proportions across multiple categories simultaneously while also showing the absolute distribution. The stacked nature of the bars will allow us to see both the total proportion of at-risk species and the breakdown by severity level. We use color mapping to distinguish between different conservation statuses, with more intense colors representing more severe conservation concerns.

Second, we create a horizontal bar chart showing the correlation between species categories and conservation status. This approach is inspired by the correlation analysis showing interesting relationships between different taxonomic groups and total biodiversity. The chart show how different taxonomic groups correlate with conservation status concerns, which might reveal unexpected patterns. For example, birds have a negative correlation (-0.1613) with total biodiversity, while insects and other invertebrates have very strong positive correlations (>0.99). This visualization will help identify which taxonomic groups might have disproportionate conservation concerns relative to their abundance.

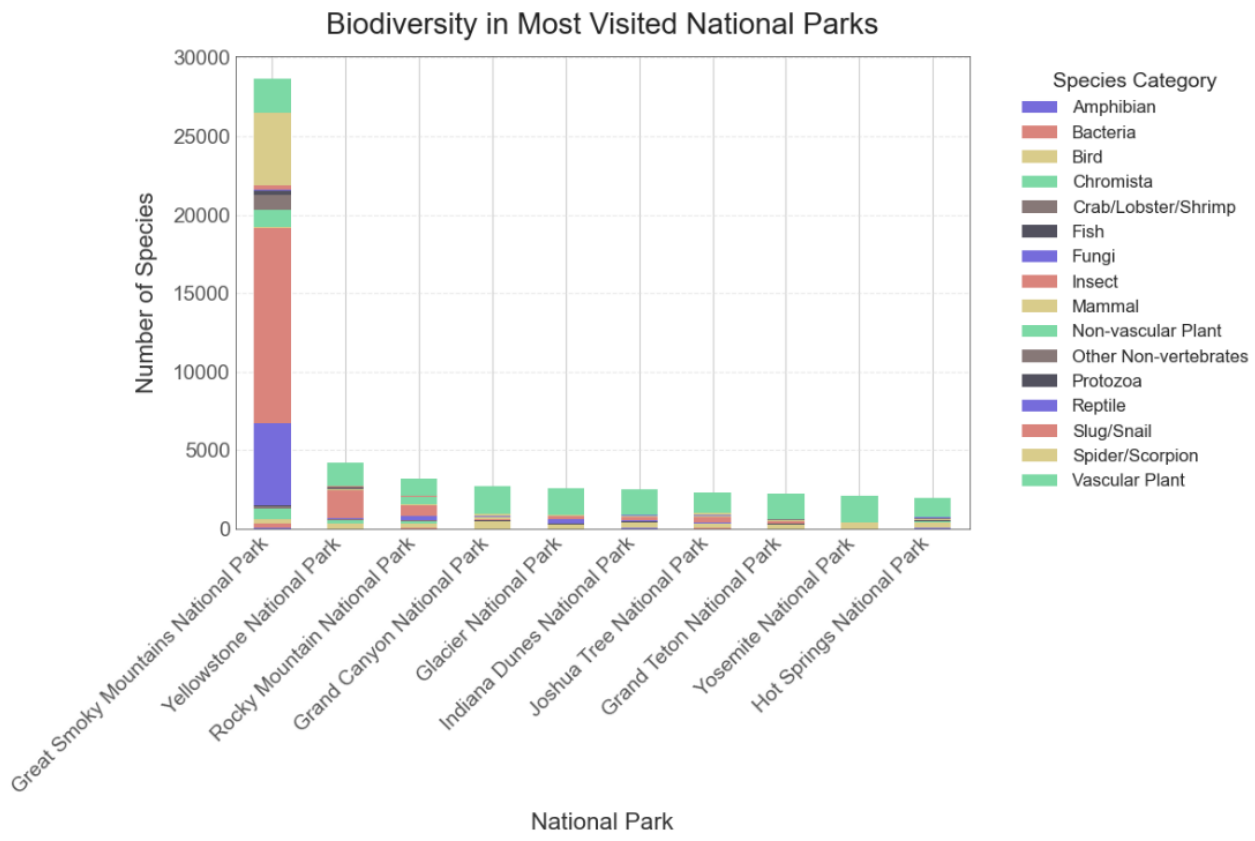
Third, we create a bubble chart analyzing conservation status by region and ecosystem type. This visualization is inspired by the species richness bubble chart that shows dramatic differences in biodiversity across regions and ecosystems. For example, the Appalachian region (home to Great Smoky Mountains) has exceptionally high species richness in temperate forests (28,670 species), while the Rocky Mountains show notable biodiversity in Alpine/Subalpine ecosystems (12,138 species across 4 parks). By examining how conservation statuses vary across these same ecosystem types, I can identify whether certain ecosystems have disproportionately high percentages of threatened species despite differences in overall species counts. This regional and ecosystem perspective will complement the taxonomic analysis and provide insight into whether conservation concerns are concentrated in particular ecological contexts.

3. Analysis

3.1. Bar chart for biodiversity in most visited national parks:

```
# Create a stacked bar chart for biodiversity
fig, ax = plt.subplots(figsize=(12, 8))
ax.set_axisbelow(True)

biodiversity_data.plot(
    kind='bar',
    stacked=True,
    color=colors[:len(biodiversity_data.columns)],
    ax=ax
)

ax.set_title('Biodiversity in Most Visited National Parks', fontsize=20, pad=15, color='#222222')
ax.set_xlabel('National Park', fontsize=16, labelpad=10, color='#333333')
ax.set_ylabel('Number of Species', fontsize=16, labelpad=10, color='#333333')
plt.xticks(rotation=45, ha='right', fontsize=14, color='#555555')
plt.yticks(fontsize=14, color='#555555')
ax.legend(title='Species Category', fontsize=12, title_fontsize=14, bbox_to_anchor=(1.05, 1), loc='upper left')
ax.grid(axis='y', linestyle='--', alpha=0.7, color='#E0E0E0')

for spine in ax.spines.values():
    spine.set_color('#333333')
    spine.set_linewidth(0.5)

plt.tight_layout()
plt.savefig('biodiversity_in_most_visited_parks.png', dpi=300, bbox_inches='tight')
plt.show()
```

Result:

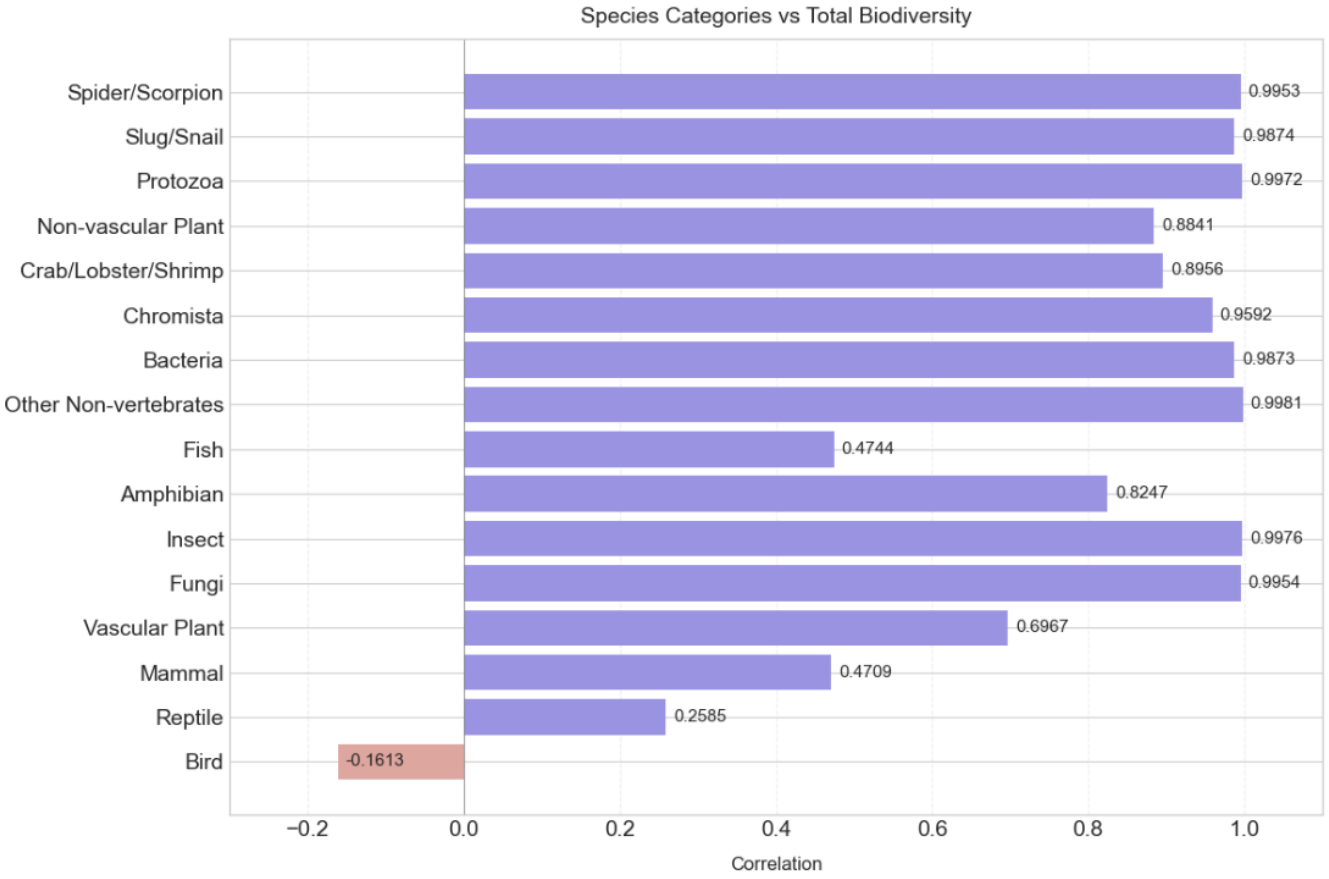


This stacked bar chart shows the proportion of species in each conservation status category (endangered, threatened, special concern, and other) across different biological taxonomic groups.

3.2. Horizontal bar chart for correlation between species categories and conservation status

```
# Count species by park and category
species_by_park_category = df.groupby(['ParkName', 'CategoryName']).size().unstack(fill_value=0)

# Calculate total species richness for each park
species_by_park_category['Total'] = species_by_park_category.sum(axis=1)

# Calculate correlations
correlations = {}
for category in species_by_park_category.columns:
    if category != 'Total':
        corr = np.corrcoef(species_by_park_category[category], species_by_park_category['Total'])[0, 1]
        correlations[category] = corr

# Convert to DataFrame
corr_df = pd.DataFrame(list(correlations.items()), columns=['Category', 'Correlation'])

# Define exact order from original image
category_order = [
    'Bird', 'Reptile', 'Mammal', 'Vascular Plant', 'Fungi', 'Insect', 'Amphibian', 'Fish', 'Algae', 'Other Non-vertebrates'
]

# Sort DataFrame by defined order
corr_df['Order'] = corr_df['Category'].map({cat: i for i, cat in enumerate(category_order)})
corr_df = corr_df.sort_values('Order').drop('Order', axis=1)

# Create chart with new format
plt.figure(figsize=(12, 8))
plt.rcParams.update({'font.size': 12, 'font.family': 'sans-serif'})

# Color array: Bird is #DDA69F, others are #9A93E2
colors = ['#DDA69F' if category == 'Bird' else '#9A93E2' for category in corr_df['Category']]

# Draw horizontal bar chart
bars = plt.barh(corr_df['Category'], corr_df['Correlation'], color=colors)

# Add correlation values next to each bar with 4 decimal places
for i, bar in enumerate(bars):
    plt.text(bar.get_width() + 0.01, bar.get_y() + bar.get_height()/2,
             f"{corr_df['Correlation'].iloc[i]:.4f}",
             va='center', fontsize=11)
```

Result:

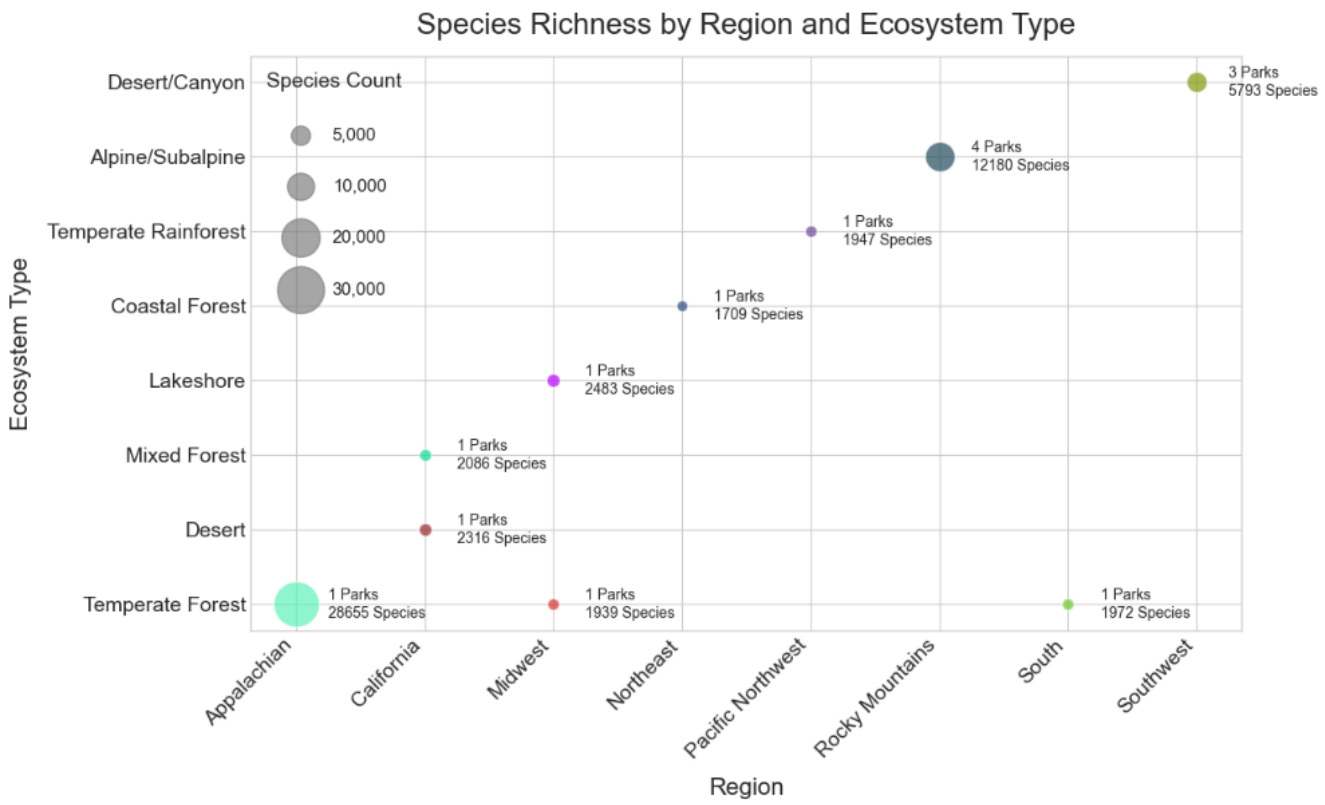


This horizontal bar chart shows the correlation between species categories and conservation status. This approach is inspired by correlation analysis showing interesting relationships between different taxonomic groups and total biodiversity

3.3. Bubbles chart for species richness by Region and Ecosystem type:

```
# Count species by park
species_by_park = df.groupby('ParkName')['SciName'].nunique().reset_index()
species_by_park.columns = ['ParkName', 'SpeciesCount']

# Add region and ecosystem information
species_by_park['Region'] = species_by_park['ParkName'].map(lambda x: park_info[x]['Region'])
species_by_park['Ecosystem'] = species_by_park['ParkName'].map(lambda x: park_info[x]['Ecosystem'])

# Count parks by region and ecosystem
parks_by_region_ecosystem = species_by_park.groupby(['Region', 'Ecosystem']).size().reset_index()
parks_by_region_ecosystem.columns = ['Region', 'Ecosystem', 'ParkCount']

# Sum species by region and ecosystem
species_by_region_ecosystem = species_by_park.groupby(['Region', 'Ecosystem'])['SpeciesCount'].sum().reset_index()

# Merge park count and species count
region_ecosystem_data = pd.merge(parks_by_region_ecosystem, species_by_region_ecosystem, on=['Region', 'Ecosystem'])

# Create figure
plt.figure(figsize=(12, 7))

# Get unique regions and ecosystems
regions = region_ecosystem_data['Region'].unique()
ecosystems = region_ecosystem_data['Ecosystem'].unique()

# Create positions
region_positions = {region: i for i, region in enumerate(regions)}
ecosystem_positions = {ecosystem: i for i, ecosystem in enumerate(ecosystems)}

# Generate random colors for each ecosystem
np.random.seed(42)  # For reproducibility
colors = [np.random.rand(3,) for _ in range(len(region_ecosystem_data))]

# Create the bubble chart
for i, row in region_ecosystem_data.iterrows():
    x = region_positions[row['Region']]
    y = ecosystem_positions[row['Ecosystem']]
    size = row['SpeciesCount'] / 30

    plt.scatter(x, y, s=size, color=colors[i], alpha=0.7, edgecolor='white')
    plt.annotate(f"{row['ParkCount']} Parks\n{int(row['SpeciesCount'])} Species",
                 (x + 0.25, y), ha='left', va='center', fontsize=10)
```

Result:



This bubble chart analyzes conservation status by region and ecosystem type. This visualization is inspired by the species richness bubble chart that shows dramatic differences in biodiversity across regions and ecosystems

4. Discussion

The analysis reveals birds have the highest proportion of conservation concerns despite their negative correlation with total biodiversity, suggesting they face disproportionate challenges relative to their contribution to overall species counts. While the Appalachian region's temperate forests (Great Smoky Mountains) have the highest species richness, Desert/Canyon ecosystems in the Southwest and Alpine/Subalpine ecosystems in the Rocky Mountains show higher percentages of vulnerable species. This highlights that biodiversity hotspots aren't necessarily conservation priority hotspots - specialized species in extreme environments often face greater threats due to narrow ecological niches. Overall, about 4-5% of species have conservation concerns, varying significantly across taxonomic groups, regions, and ecosystems, demonstrating the need for tailored conservation approaches rather than strategies based solely on species counts.

Conclusion

Our analysis of the National Park Service dataset confirmed the food chain pyramid theory while identifying that about 4-5% of species face conservation concerns, with birds being disproportionately vulnerable. We discovered that biodiversity hotspots don't necessarily align with conservation

priorities—specialized Desert/Canyon and Alpine/Subalpine ecosystems show higher percentages of threatened species than species-rich temperate forests.

Moving forward, we recommend: implementing more detailed species classification systems; developing targeted conservation strategies for birds; allocating additional resources to specialized ecosystems with high vulnerability rates; conducting temporal studies to track population changes; incorporating geographic data; and investigating outlier parks like Yellowstone for unique ecological patterns or data collection issues.

The dataset does not have information about the time of species recording, only static information, so it is difficult to combine with another dataset such as atmospheric CO2 over time to track the correlation between vegetation and CO2 because plants are very sensitive to this situation.