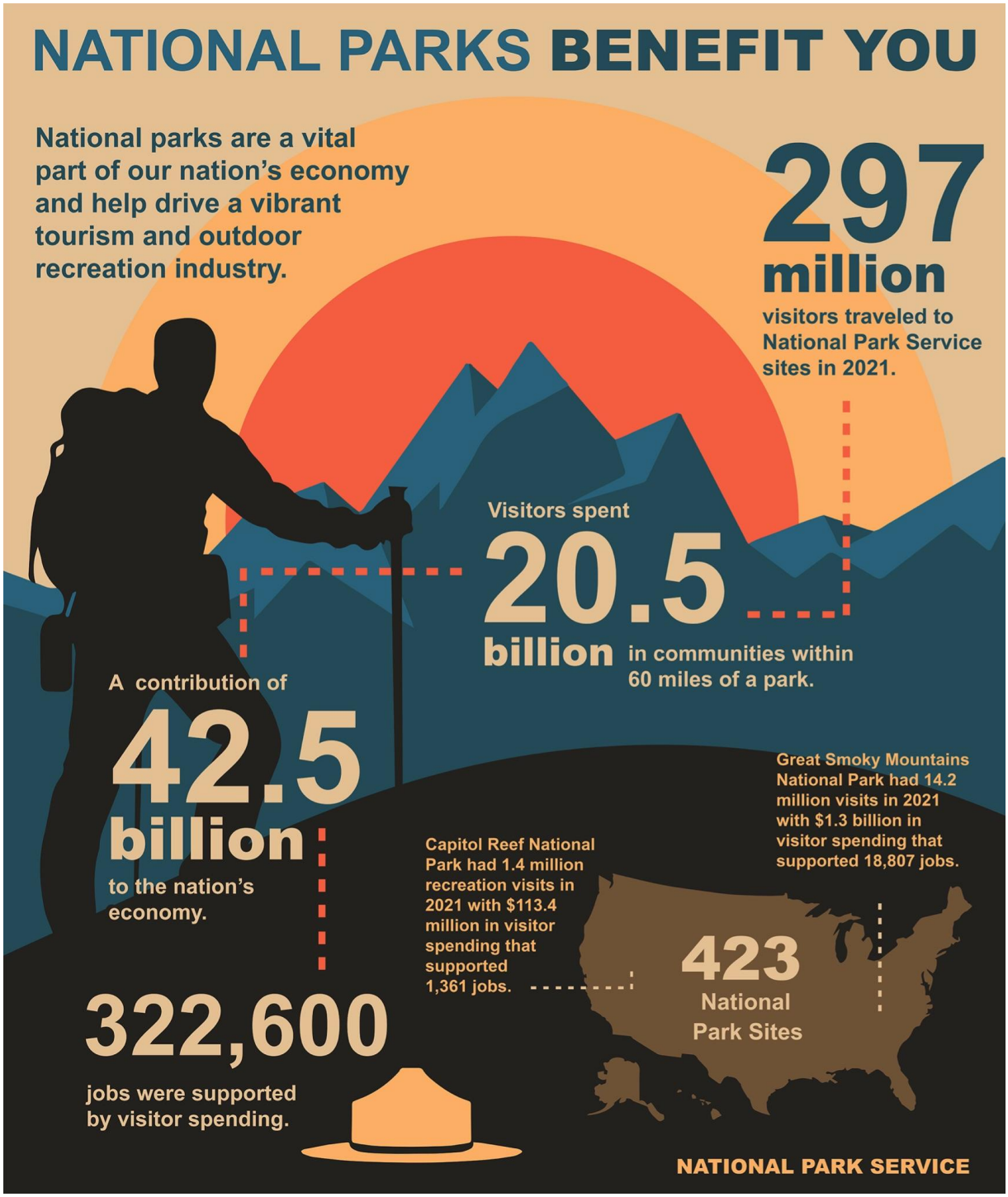
Exploratory Data Analysis: Team 7

United States National Parks

Hunter Dawley, Samantha Armijo, Ji Noh, Yahan Yang



Introduction

As data science students who live in a world where climate change is discussed more, we are naturally curious to see how the weather has changed over the years and how it has affected different aspects of our life and environment. We have all seen how climate change affects our world. To this day, we hear how typically it isn't this cold or humid around this time of year, but because of climate change, it is occurring. During COVID, when the world shut down people took advantage of their free time to visit national parks. Some parks "particularly those near urban areas, experienced increased numbers of visitors as people sought safe areas for recreation and exercise"[2].

With there being 423 national park sites in the United States[5], they "are a vital part of our nation's economy and help drive a vibrant tourism and outdoor recreation industry"[5]. In 2017 alone, 330,882,751 people visited United States National Parks[4]. That same year this industry contributed 18.2 billion to the nation's economy and supported 306,000 local jobs[4]. It is easy to see that these National Parks are a great source of revenue and jobs for the US government and local communities. Our team wanted to see the trends revolving around National Parks in the past years. We narrowed in on 2017 to see how the parks were operating before COVID-19 hit the world. Further, we wanted to see how weather patterns would affect National Parks' visitation rates over a 10-year span, 2012 to 2017. In Figure 1 we can see that different areas of the country had different visitation rates during the year 2017. By looking at these components, our team could see the peak season of parks in different parts of the country, California, Colorado, and Florida. We picked these 3 different states because each state experiences different weather patterns, visitation rates, and possibly biodiversity.

As climate change became more prominent it has impacted each of these states differently. Would weather patterns harm biodiversity rates? What kind of weather do visitors usually want to visit in? What part of the country (west, middle, or east) has the greatest variation in weather and biodiversity? These are all key questions our group strived to answer in this report with the data that we explored.

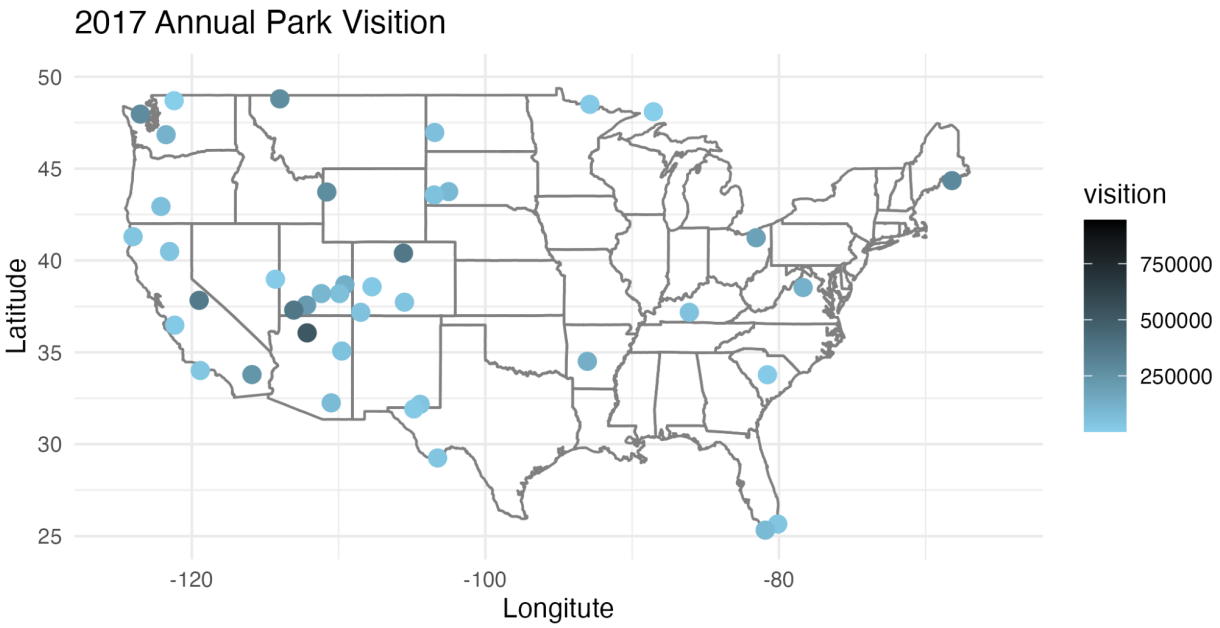


Figure 1: All state parks visitation rates in 2017 from raw dataset, National Parks Visitation Data

Data

Preprocessing Data

One of the datasets used to explore findings was the national park visitation dataset. This dataset came from the U.S National Park Service which publishes data about visitors to its parks and other historical sites. It contains annual monthly visitors per park since 1979. It includes different variables for each national park such as overnight stays by different types of lodging such as tents, RVs, etc to recreational and non recreational visit counts. This dataset included 35 variables/columns and 31,883 observations/rows. Each park has a unique unit code and year that was used to filter the desired parks and years. Using the naniar package in R it was concluded that there was no missing data in the raw national parks visitation data. After creating a data frame that contained the desired parks in 2017, we created different graphs that showed visitation rates for each state (see Figure 2), for each state's parks (see Figure 3), for all the parks together, and all the park visits by month in 2017 (see Figure 4). As we can see from doing preprocessing work on the visitation data, some parks in a certain state has more visitors, California has the most

visitors, and there are certain months of the year that people will visit National Parks. Overall, in 2017 Colorado had 1,461,270, California had 1,436,373, and Flordia had 506,599 average visits per park.

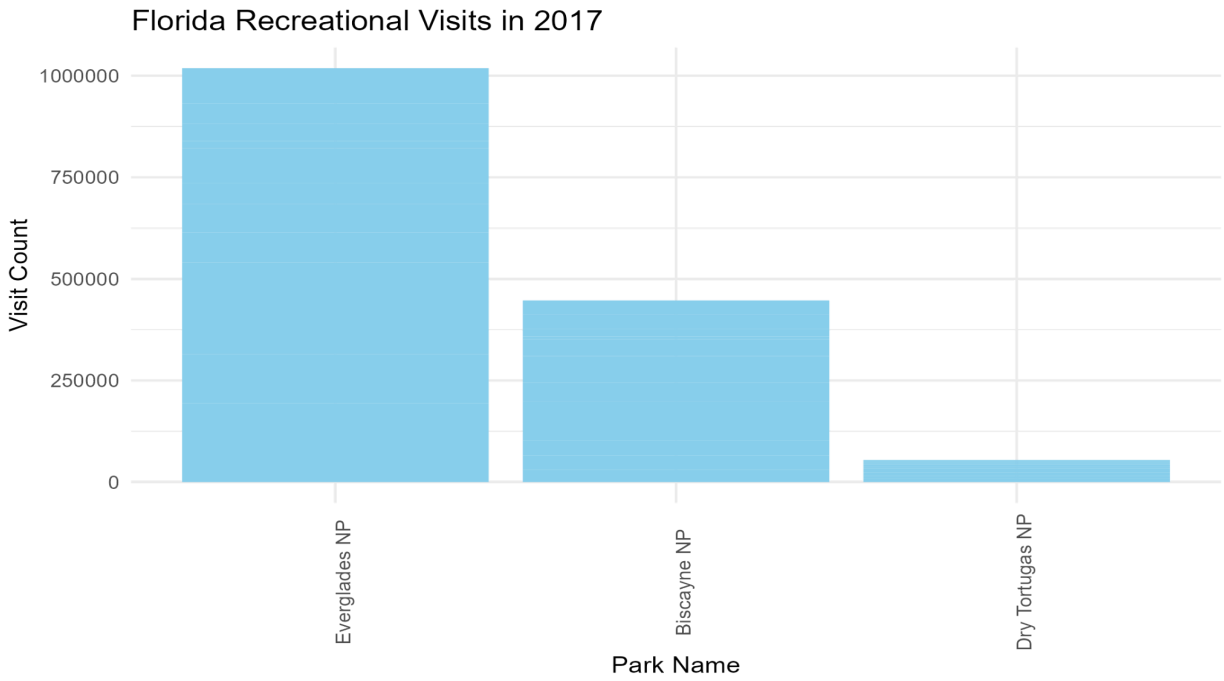


Figure 2: Florida's 2017 recreational visitation count for each park. This is an example of what was created for each state.

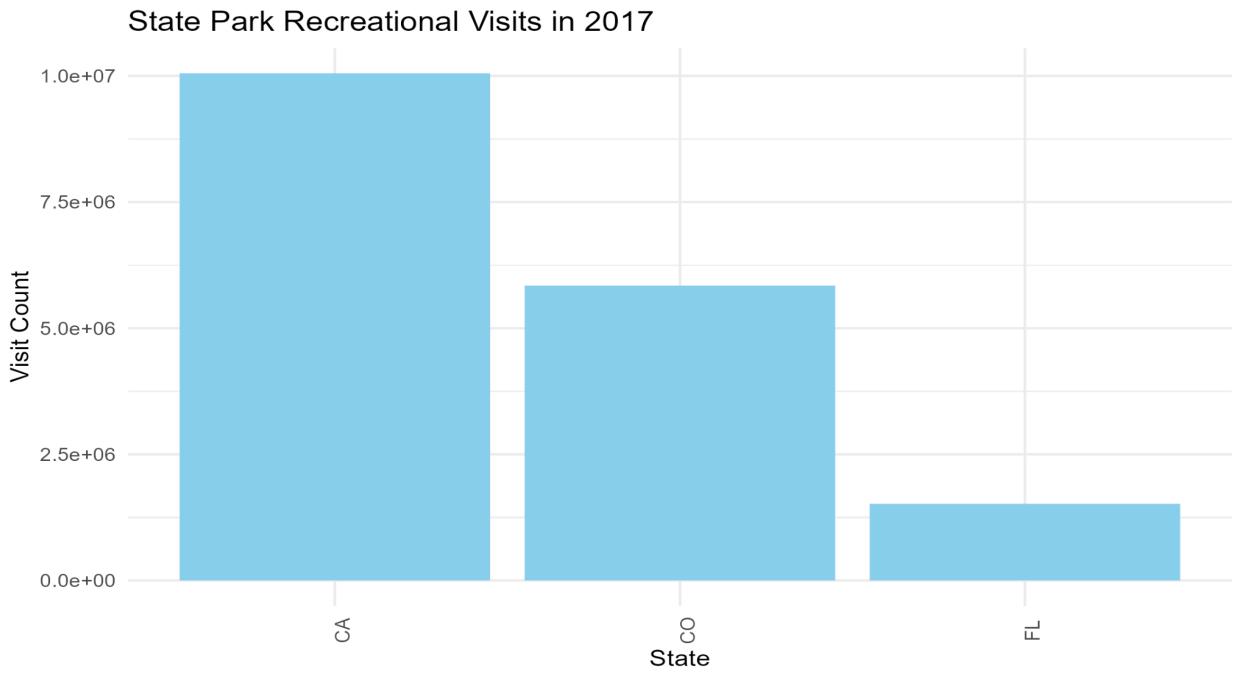


Figure 3: Sum of recreational visits by state.

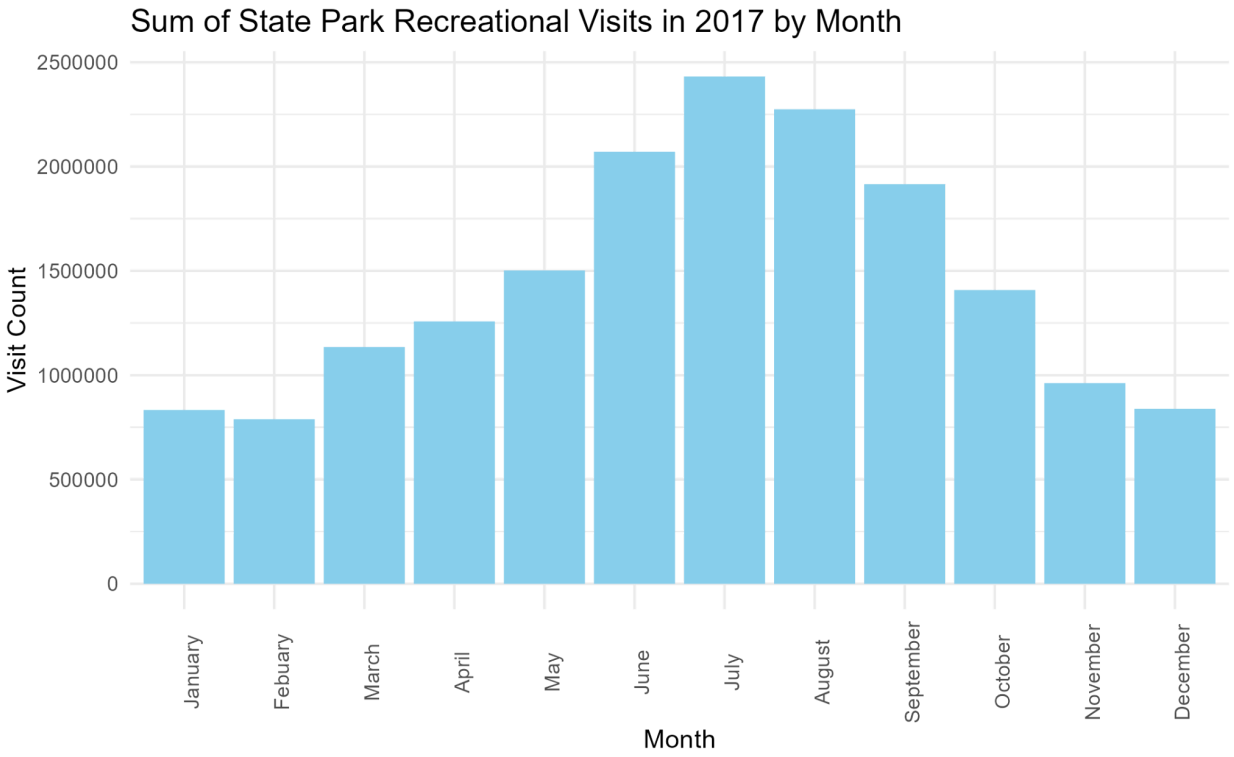


Figure 4: Recreational visit count for all parks by month in 2017.

Our team also collected weather-related datasets, which contain temperatures, humidity, pressure, wind speed, wind direction, and weather description. The "Historical Hourly Weather Data 2012-2017" zip file from Kaggle contains seven different datasets; "city_attributes", "humidity", "pressure", "temperature", "weather_description", "wind_direction", and "wind_speed". Our group decided that among seven weather attributes, "temperature", "humidity", "weather_description", and "wind_speed" datasets would be useful for our research purpose. Each dataset in the "Historical Hourly Weather Data 2012-2017" zip file was measured on an hourly time frame, with 45253 observations. We used the data we had for the closest city to each National Park. From our dataset, we had Denver, Miami, Jacksonville, Los Angeles, San Francisco, and San Diego available to utilize for the states of Colorado, Florida, and California. Therefore, all of the Colorado national parks will be connected to Denver, and all of the Florida national parks will be with Miami. We had three cities for California, San Francisco, San Diego, and Los Angeles. We found that Yosemite, Pinnacles, Redwoods, and Lassen National Parks are closer to San

Francisco. Channel Islands, Joshua Tree, and Death Valley are closer to Los Angeles. Therefore, we filtered out San Francisco, Los Angeles, Denver, and Miami for the cities, as those are the cities that contain the national parks we are interested in. Afterward, we split the year, month, day, and time to separate variables for more precise analysis. To interpret the observation thoroughly, we transformed the temperature data from Kelvin (K) to Fahrenheit(°F). After all the filtering and splitting, we joined the "temperature", "wind_speed", "weather_description", and "humidity" datasets into one single data frame. "join_weather_12_17.csv" dataset contains 271519 rows and 9 columns, with columns being "year", "month", "day", "time", "City", "temperature", "weather_description", "wind_speed", and "humidity". Here, the dataset has information from October 1st of 2012 to November 30th of 2017.

With the "join_weather_12_17.csv" dataset, our team made box plots to compare the patterns or findings for each of the cities. Here, we plotted all the observations that are originally hourly measured in a box plot format. Therefore, extreme values are plotted as outliers, which are shown as a dot on the plot.

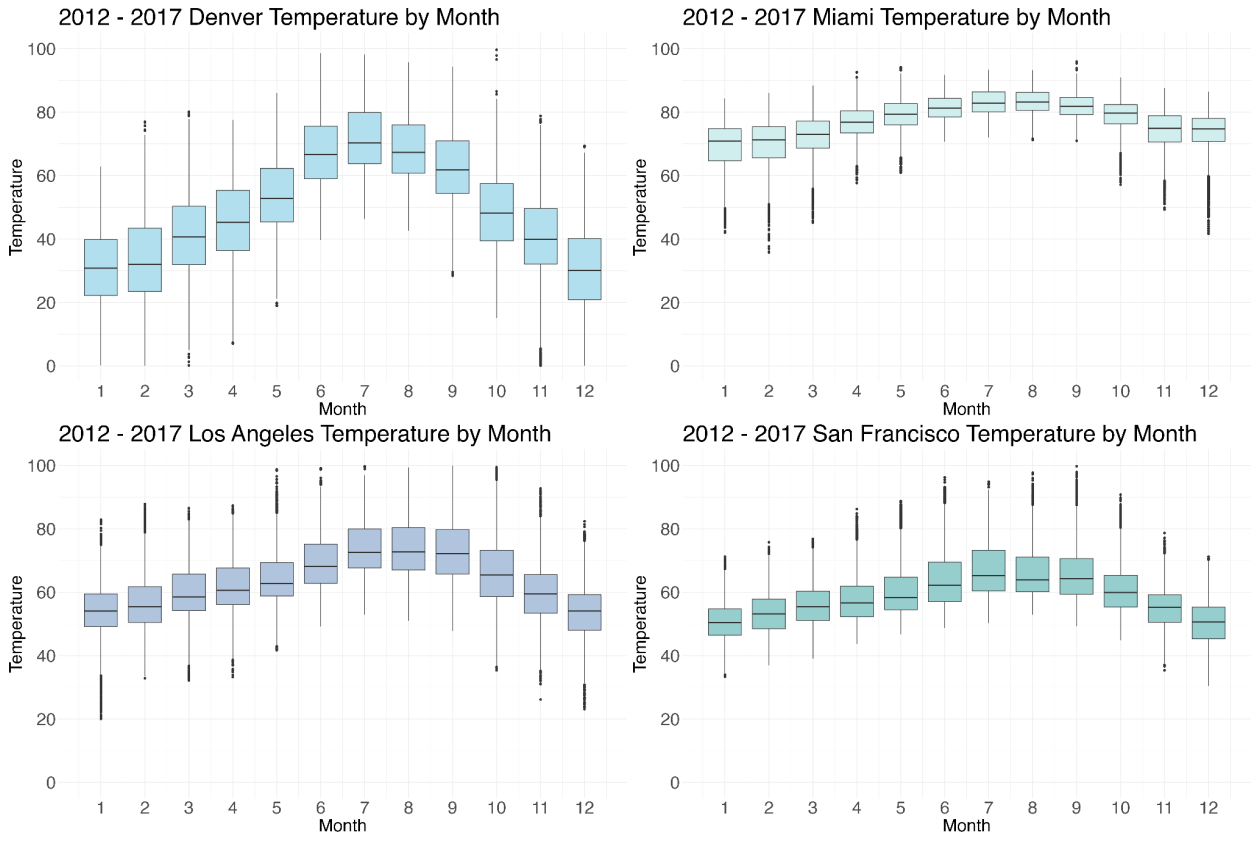


Figure 5: 2012-2017 Temperature by month for Denver, Miami, Los Angeles, and San Francisco

Looking at the visualizations of temperature, we identified that Denver, Colorado has the most fluctuation in terms of temperature among all four cities, Denver, Miami, Los Angeles, and San Francisco. Miami's temperature does not change drastically across the month, and it has the highest temperature, which ranges from 65°F to 80°F. The lowest temperature across the year for Miami, which is in January, still is the highest when compared with other cities' January temperatures. January temperature of Miami is even higher than every month's temperature in San Francisco (see Figure 5).

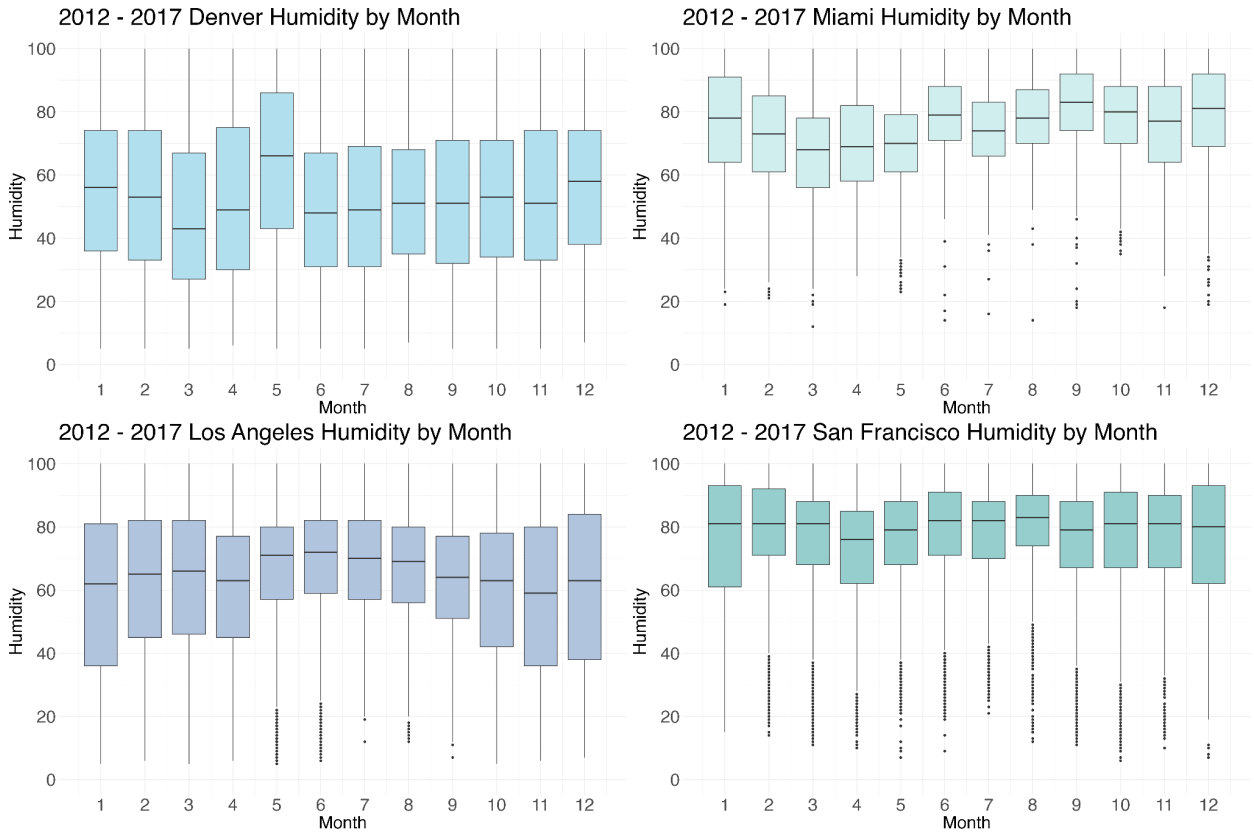


Figure 6: 2012-2017 Humidity by month for Denver, Miami, Los Angeles, and San Francisco

Humidity is mostly stable across year for all cities. Some interesting finding to note is that San Francisco, CA has a lot of outliers. Furthermore, San Francisco has the highest average humidity across the year, which range from 60% to 90% (see figure 6).

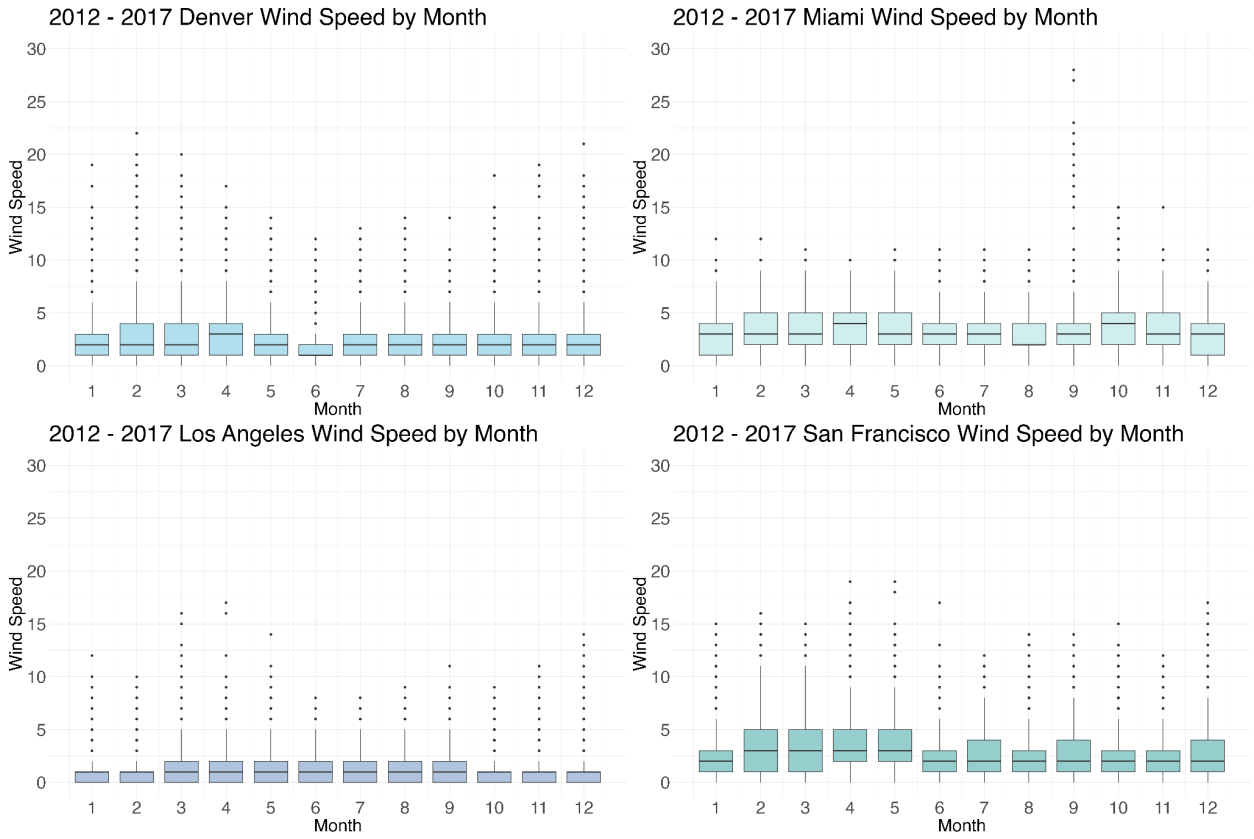


Figure 7: 2012-2017 Wind speed by month for Denver, Miami, Los Angeles, and San Francisco

Figure 7 shows the box plot for each city's wind speed across the year. Most of them have wind speeds between 0 miles per second to 5 miles per second. However, for wind speed, there are much more outliers for all cities compared to other weather attributes (see figure 7).

The last dataset we looked at was "Biodiversity Among the National Parks." This file was downloaded by Kaggle and included two datasets. One dataset had just the US parks and their respective latitude, longitude, park code, park name, state which it is in and the acreage. We cleaned this data to only include the seven national parks in California, four in Denver and three in Miami. We kept the latitude, longitude, state and park name variables while renaming the park names to match with the other datasets for an easier join. The other part of the biodiversity data is the species dataset. This dataset contained the park code and name. The species component contained the category, order and family of the species along with the common and scientific name for them. Lastly, the dataset recorded how many times the animal/plant was seen in 2017 (occurrence variable), along with the

record status and nativeness to the park. We were primarily interested in seeing how many species were in each park of our interest. We created a new data set that contained the park name and code along with the category of species. Since there were so many species and the distribution of nonvascular plants was much greater than animals, it was easier to narrow it down to a couple that we wanted to compare. Originally, the category variable contained observations for birds, mammals, nonvascular plants, reptiles, amphibians, fungi and insects. However, we decided to look at birds and mammals as we inferred they were the species most noticeable when people would visit national parks. Once we filtered for the 14 parks and just the mammals and birds in each park respectively, we found the counts and then averaged the birds and mammals per state. We named the new dataset “cleaned_parks” once we filtered for our state parks, their code, name and lat/long. The original parks dataset had 56 observations/rows with 6 variables/columns. The new “cleaned_parks” dataset had 14 observations (one for each park) and 6 variables. For the species dataset, the original one had 119,248 observations and 14 variables. Once we filtered, and added a new column the bird/mammal counts, our new dataset was called “species_final.” This dataset was narrowed down to 3 columns and 28 observations (one observation for each park and bird count and one observation for the park and mammal count). The three columns in this dataset were ‘park’, ‘category’ and ‘count.’ We finally joined the “cleaned_parks” to the “species_final” dataset through a full join so that we had the park code, name, latitude and longitude along with the biodiversity counts. This new dataset was called “join_species_park.”

To find the average number of birds and mammals in each national park of a state, we took “join_species_final” and created a new dataset of just the parks in each state. We then took the mean of the birds and mammals separately and put them in a new column. We did this for all of the three states. It is important to note that California’s total was divided by 7 whereas Colorado was only divided by 4 because of how many parks are in that state.

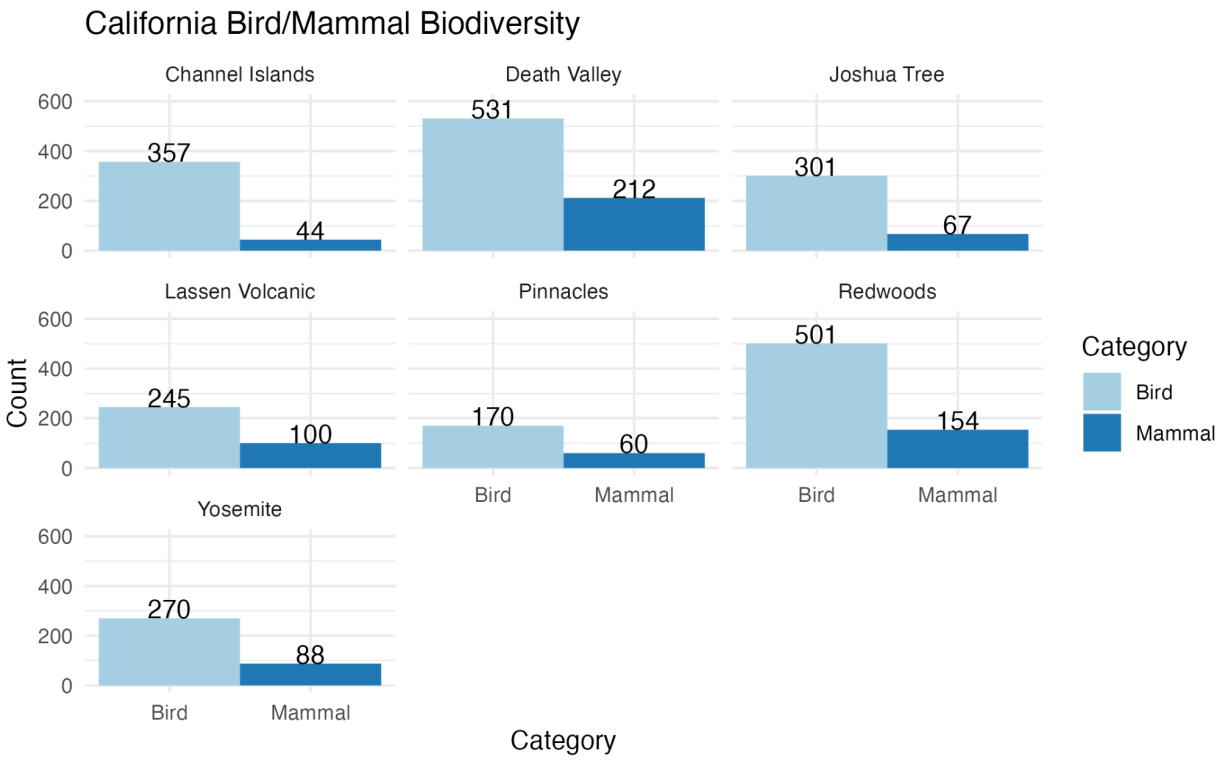


Figure 8: California Bird and Mammal Biodiversity

Among the seven California national parks that we looked at, all the parks had more birds than mammals present. Death Valley had the greatest biodiversity and the Pinnacles had the least. It was somewhat surprising that Death Valley had the highest count of birds and mammals as the National Park is in a very dry and rocky climate in California and Nevada. The average number of birds in California national parks was 339.29 and 103.57 mammals. California had the most biodiversity among all the states we looked at.

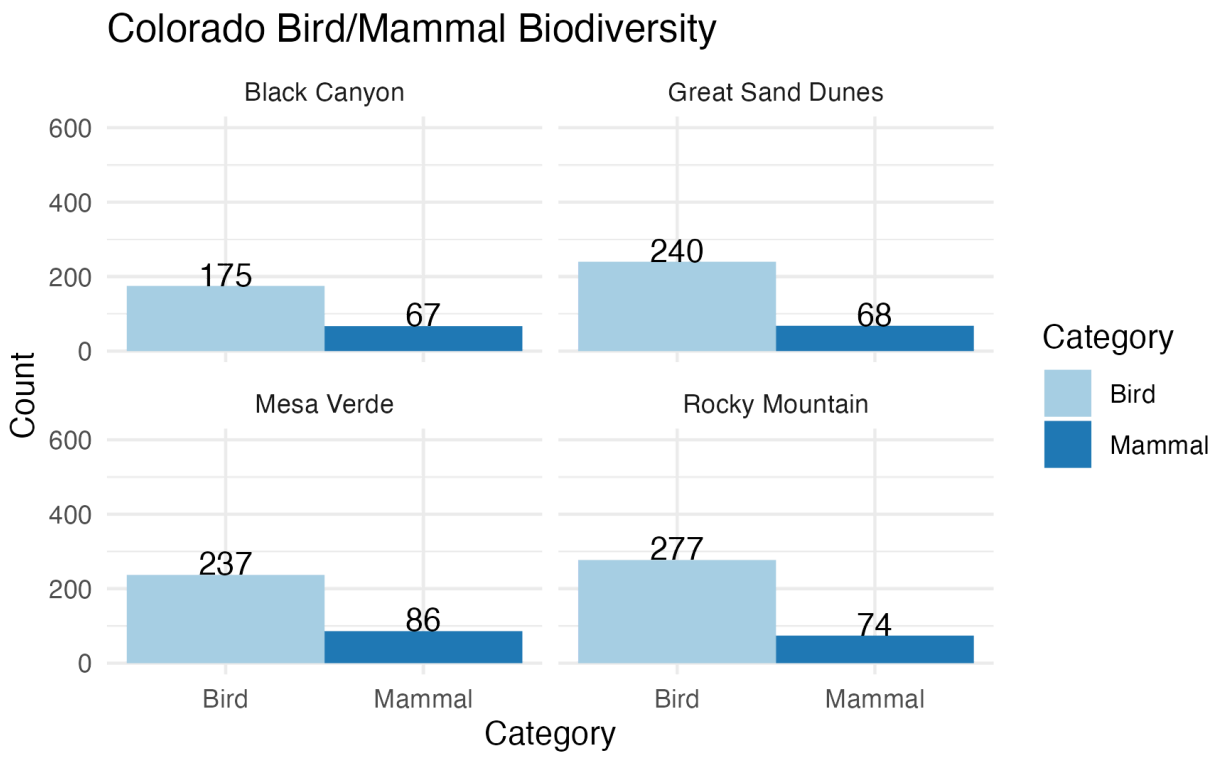


Figure 9: Colorado Bird and Mammal Biodiversity

The biodiversity among the Colorado National Parks was very similar. The amount of birds present in each park during 2017 was double the amount of mammals. The Rocky Mountains had the greatest biodiversity where the Black Canyon park had the least. That being said, Colorado had an average of 73.75 mammals and 232.25 birds per national park in 2017.

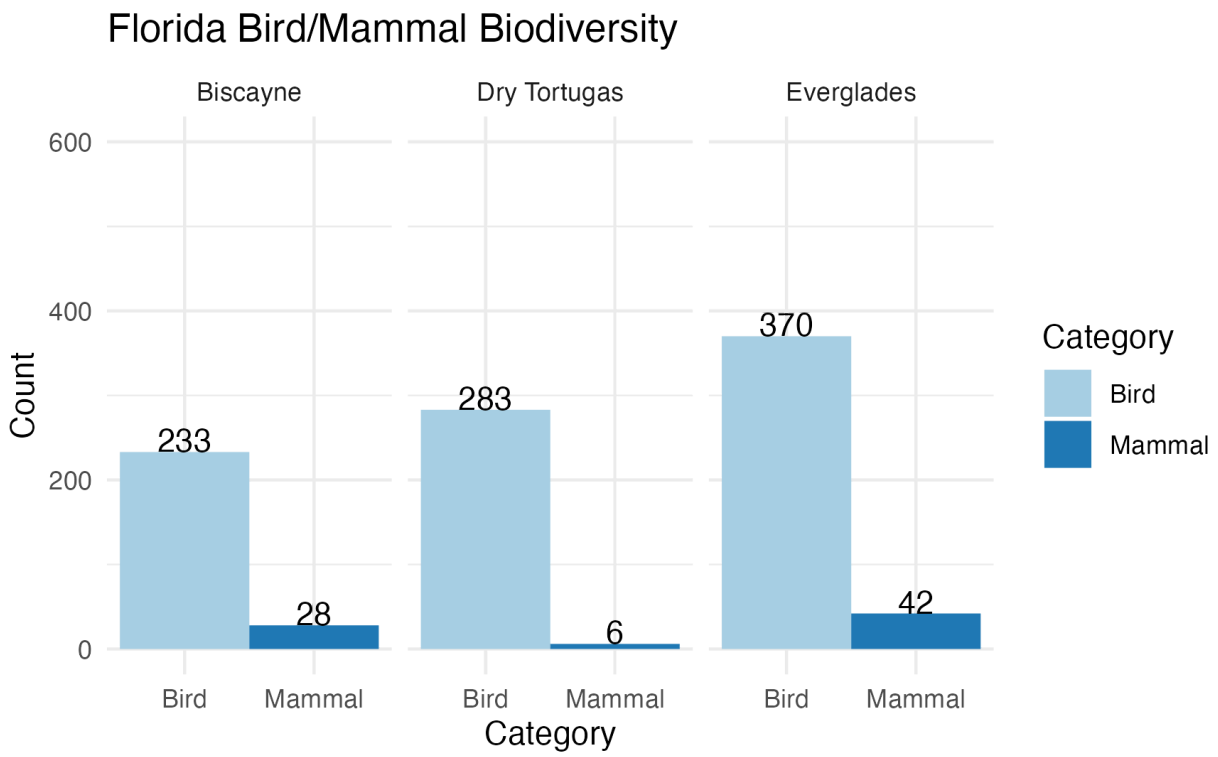


Figure 10: Florida Bird and Mammal Biodiversity

The Florida biodiversity among birds and mammals was the greatest. We can see that not only do Florida national parks have many more birds than mammals, but the Florida parks had more birds than the National Parks in both California and Colorado. On average, Florida parks only had 25.33 mammals per park but 295.33 birds. We can conclude that this has to do with the marsh/wetlands that the Florida national parks are in rather than a mountain climate.

Master Data

After preprocessing and cleaning all the data we attempted to create one master dataset but unfortunately our master exceeded rows in Excel so we created 2 final master datasets. One of the datasets contained 2017 weather and biodiversity. This was done by cleaning the species and park datasets and joining them by the Park Name. After joining those two we then added a city column because it was necessary for one of our analysis as we wanted to see the closest cities to the park. After that we joined biodiversity to 2017

weather by state. This created the final weather biodiversity csv file found in our clean data folder inside of data.

For our other master dataset we joined 2012-2017 weather to visitation. Prior to doing that we calculated the average temperature, windspeed, and humidity by month because the weather dataset variables include month, day, and year while visitation only has the month and year. After calculating the average per month we joined the weather dataset to visitation by month, year, and state.

Key Findings

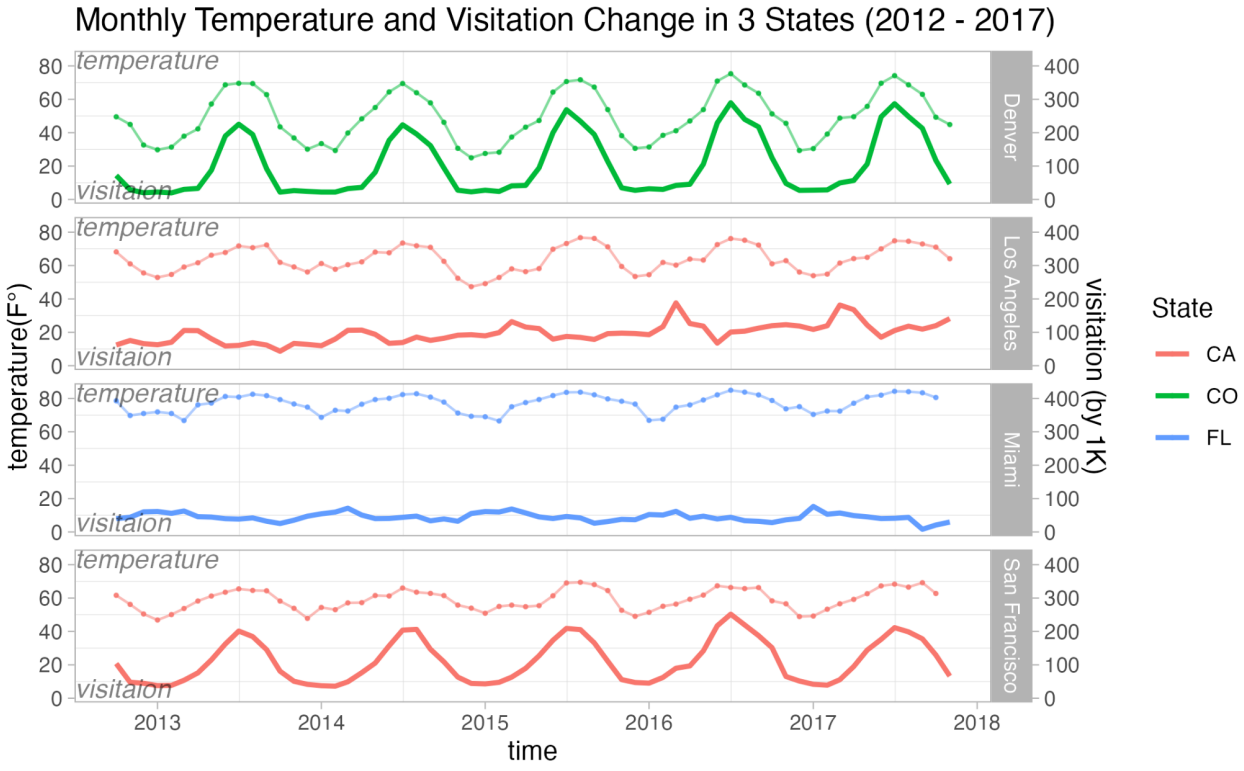


Figure 11: Time Series for Monthly Temperature and Visitation form 2012-2017

Here we created time series plots for weather and visitation to see if there is any pattern. Over time we can see that visitation rates and temperature follow each other. For Denver and San Francisco there is a clear pattern that as temperature increases so does visitation. This is in line with our intuition, a temperature around 65 to 77 degrees Fahrenheit is the optimal temperature for outdoor productivity. [cite7] Thus, the higher temperature the more visitation in those two cities. However, if we took a closer look at the

lines for Los Angeles and Miami, they show an opposite trend. The visitation rises slightly as temperature drops. This may be because the temperature in these areas is high all year round(over 60 degrees Fahrenheit), and people are more willing to go out when it is cold.

Apart from the temperature, humidity may also be one of the factors affecting visitation, because the temperature that the human body feels is different under different humidity. When humidity levels go up, people could sweat more based on life experience. Our bodies react as though the temperature has increased because excessive humidity makes the temperature feel hotter than it actually is. Even worse, high humidity prevents sweat from evaporating, leaving moisture on our skin that makes us feel even warmer. Now let's take a look if there is any pattern between humidity and park visitation.

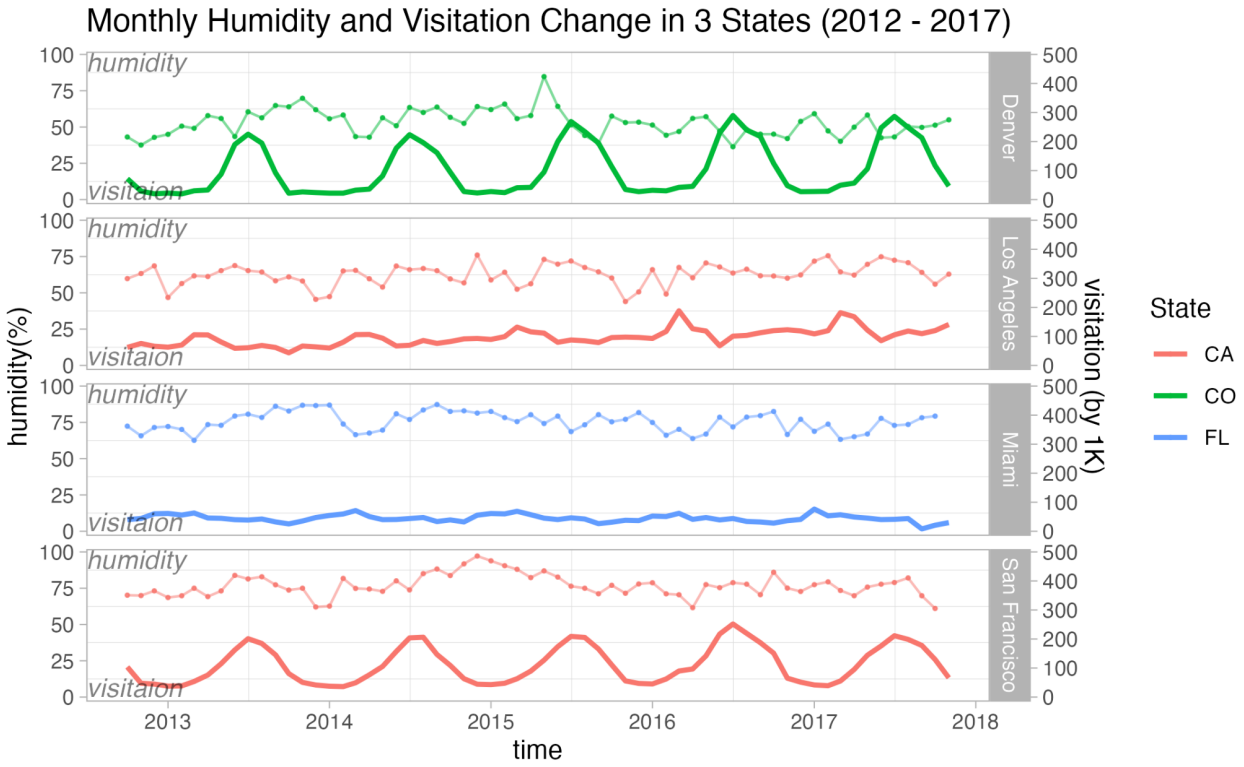


Figure 12: Time Series for Monthly Humidity Rate and Visitation form 2012-2017

There is a degree of correlation between humidity and visitation, as seen on the graph. The visitation falls as humidity increases in Denver and Los Angeles. Except for the fact that San Francisco experienced exceptionally high average humidity in 2015 as visitation dropped, there is no discernible pattern. In Miami, humidity fell in 2014 while visitation went up modestly, and from 2017 to 2018, visitation fell somewhat as humidity

continued to rise. Though the average humidity does not vary greatly from year to year, we may still draw the conclusion that there is some association between them.

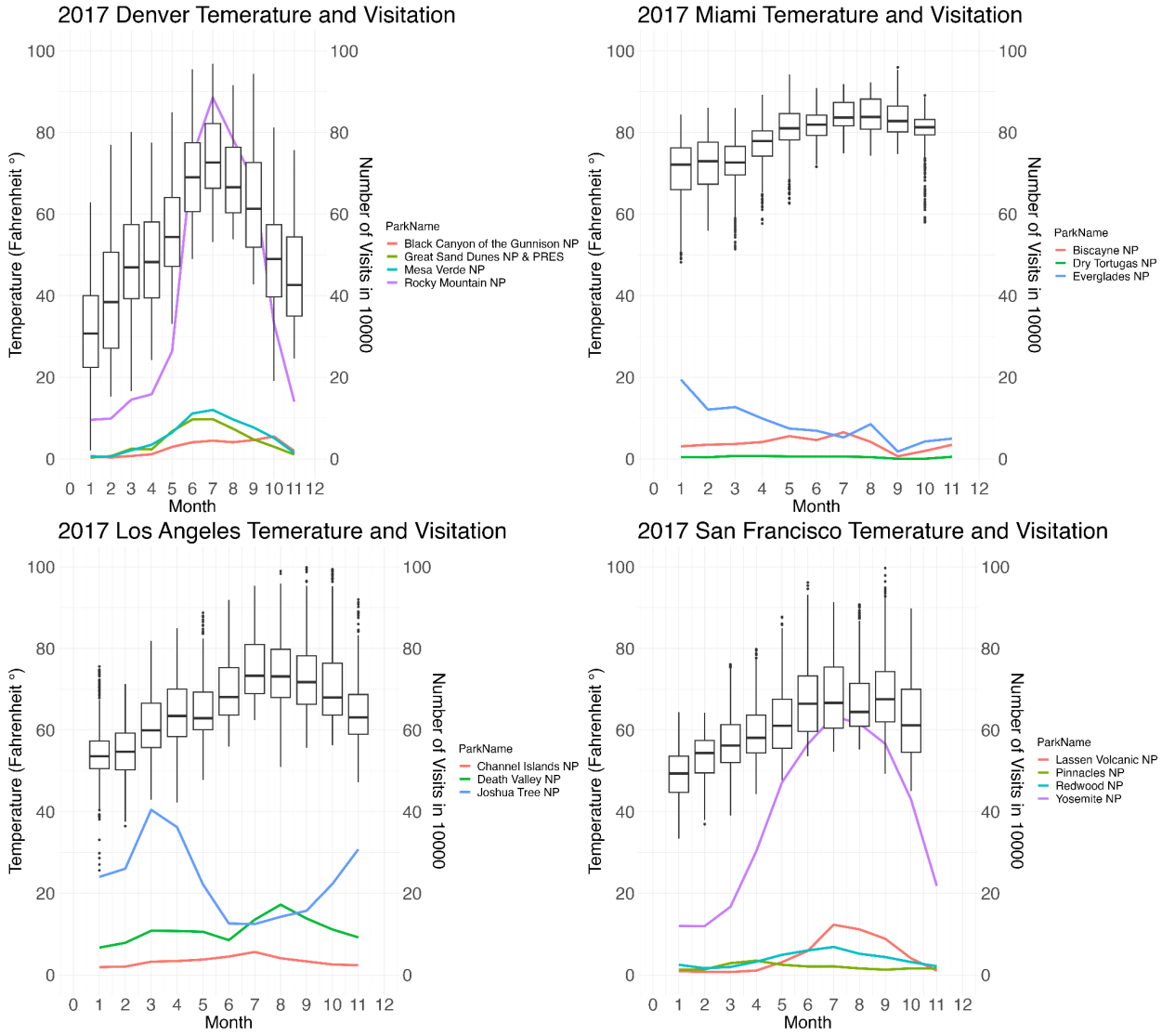


Figure 13: 2017 Visitation and Temperature Comparison for Each City

After looking at the time series analysis (see figure12), we concluded that among all the weather attributes, the temperature has the most relationship with the visitation to national parks. Therefore, we wanted to look closely at how the temperature pattern in 2017 is similar to the visitation pattern in 2017. Figure 13 shows the box plot for temperature that was measured in the hourly timeframe for each city, and on the same plot, the line shows the number of visits for each month. The combined plot for Denver, Colorado seems to have the strongest similar pattern between temperature and visitation. In 2017 Denver, when the temperature goes up, the visitation to every national park

increases. Similar observation continues with San Francisco, California, as the temperature pattern matches the visitation pattern well. In contrast, Miami, Florida plot shows a notable finding. In 2017 Miami, most of the visitations to Everglade National Park occurred during the winter season, January, February, and March. However, Biscayne National Park in Miami follows the Denver and San Francisco pattern; when temperature increases, visitation also increases. Los Angeles, California also shows unique patterns depending on the national park. For example, visitation to Joshua Tree National park is the highest in March, and lowest in July. In contrast, Channel Island and Death Valley national parks have the most visitation during summer time, July and August, and least during winter time, November, December, and January.

In conclusion, temperature pattern and visitation pattern seems to vary depending on the city. This finding urged us to look more precisely at the relationships between temperature and visitation.

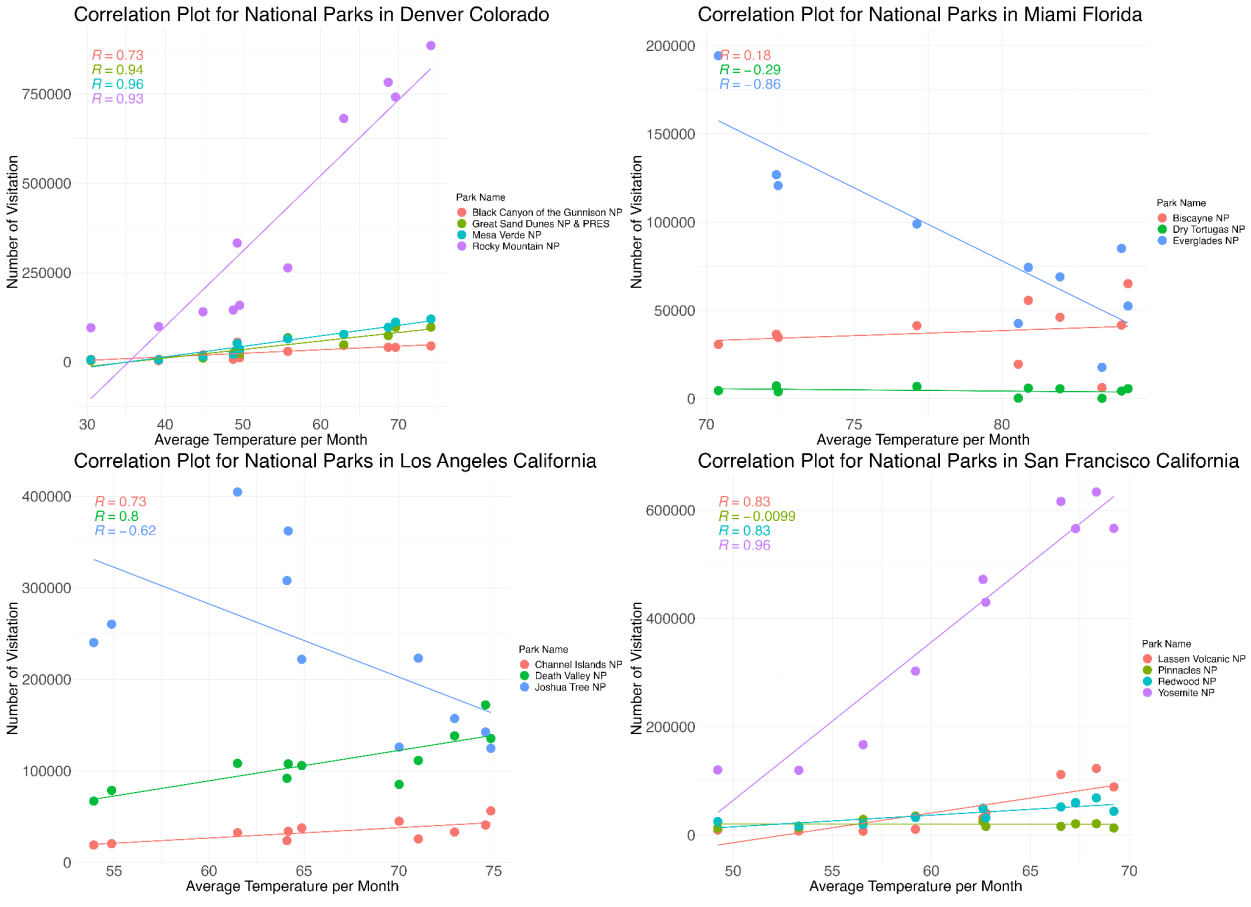


Figure 14: Correlation plot for 2017 Visitation and Temperature for each city

To elaborate on our temperature and visitation findings more precisely, figure 14 shows the correlation for each national park with the 2017 monthly average city temperature. Among all cities, Denver has the strongest positive correlation, with 0.73, 0.94, 0.96, and 0.93. Even though Black Canyon of the Gunnison National Park in Denver has the smallest correlation coefficient, 0.73, it is still very high. In the Miami correlation plot, only the Biscayne National Park has a positive correlation coefficient, with 0.18. Dry Tortugas and Everglades National Parks have a negative correlation coefficient being -0.29 and -0.86 respectively. In Miami, even the positive correlation coefficient value is very small, and Everglade National Park has a strong negative relationship with temperature and visitation. Los Angeles has both strong positive correlation coefficients and strong negative correlation coefficients. Channel Island National park has a 0.73 coefficient, Death Valley has a 0.8 correlation coefficient, and Joshua Tree has a -0.62 correlation coefficient. Although the type of relationships differs by national parks, all three parks have strong relationships with temperature and visitation. Lastly, most of the national parks in San Francisco have a positive relationship with temperature and visitation. Among four parks, three have a strong positive relationship; Lassen Volcanic National Park with R=0.83, Redwood National Park with R=0.83, and Yosemite National Park with R=0.96. One exception has been made with Pinnacles National Park in San Francisco. The correlation coefficient for Pinnacles is -0.00099, which is almost 0. For Pinnacles National Park, we can conclude that there is not much of a relationship between temperature and visitation.

In conclusion, the parks that had a negative correlation between visitation and temperature were the Everglades, Miami(-0.86), Dry Tortugas, Miami(-0.29), and Joshua Tree in Los Angeles (-0.62). Both Mesa Verde, Denver, and Yosemite, San Francisco, had the highest value of correlation (0.96). Since Denver has the most fluctuation in temperature, the parks in Denver have a higher positive correlation overall as people are less likely to visit when the temperature is below 50 degrees. This conclusion shows the reason why Miami Florida mostly has a negative correlation, as the coldest month in Miami is January with 70.41°F as an average temperature, which is the most desirable temperature for national park visits.

We draw maps to see the national park distribution in the United States and it is obvious that there are more parks in California, Colorado, and Florida. Also, the red color appears more in these states, which means they had a higher visitation in 2017.

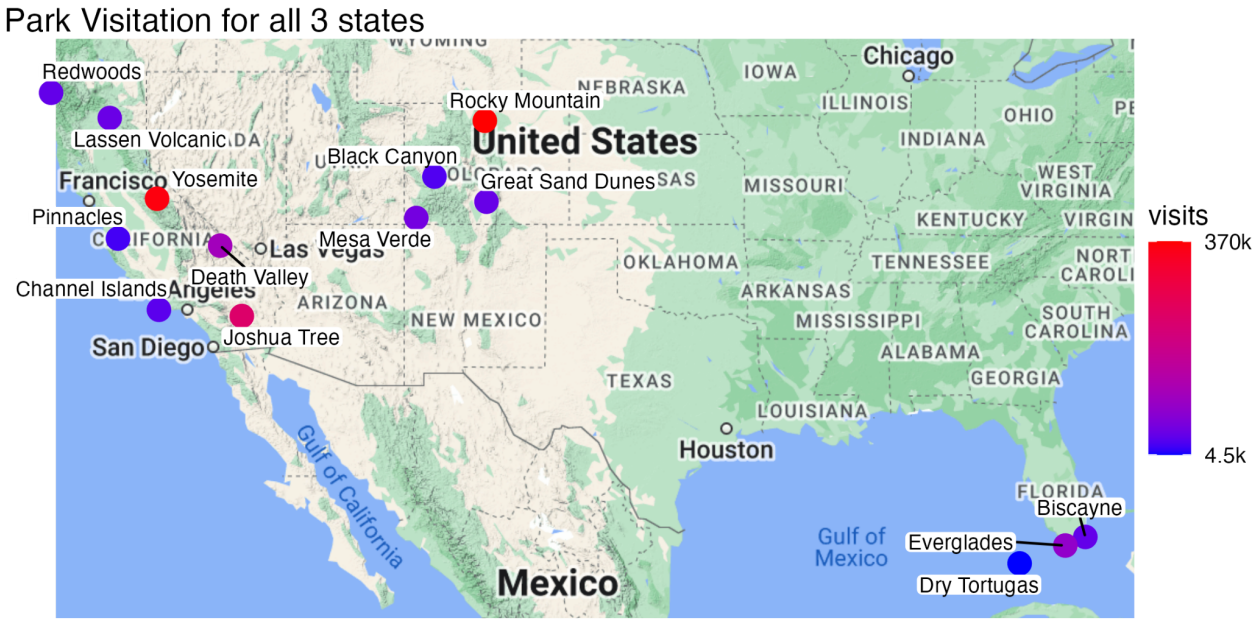


Figure 15: Visitation Rate Among All National Parks in 2017

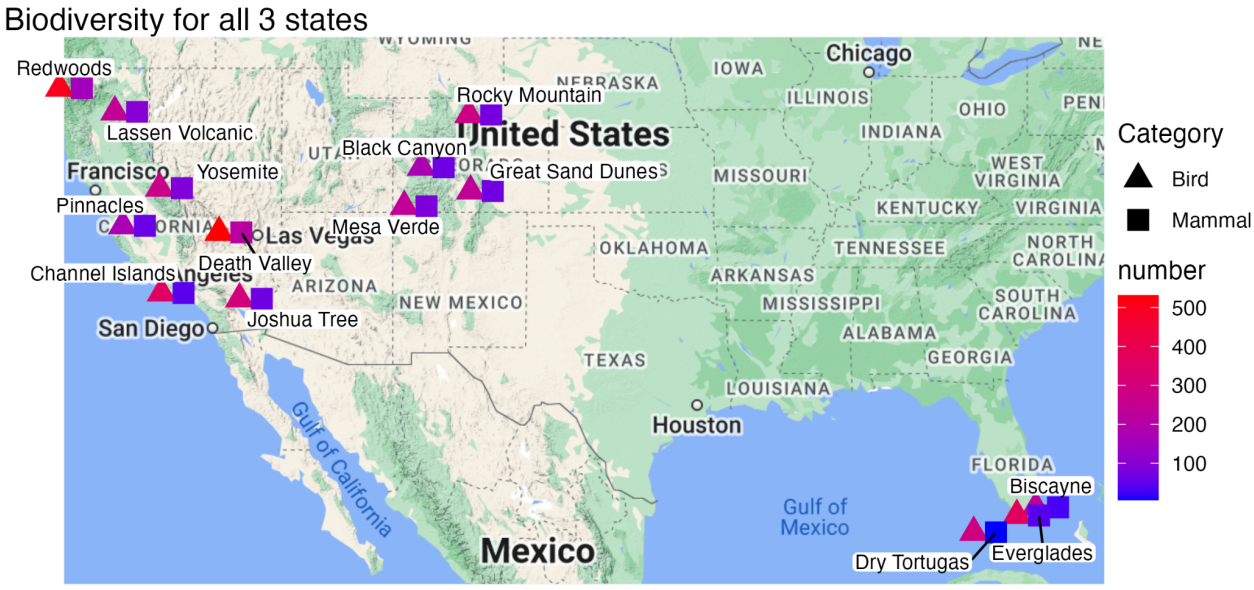


Figure 16: Bird and Mammal Counts in All National Parks in 2017

For mapping, we used Google Map API to visualize the United States map and show the distribution of parks on the map by points. The color scales from blue to red indicate

the number of visitors, birds, and mammals in each park of our interest. The red color refers to a higher amount and blue refers to a less amount. In figure 15, we can tell that the Rocky mountains, Yosemite and Joshua Tree have the highest visitation rate. And the parks in California have a higher visitation among all three states. It's obvious that in figure 16, the Redwoods, Death Valley and Everglades national parks have the largest variety of birds while the first two national parks also have the highest biodiversity of mammals. Birds do better in mild and warmer climates, while mammals do better in cooler climates like Colorado. By comparing these two figures, we can conclude that people tend to visit national parks with higher biodiversity in terms of birds and mammals in general.

Conclusion

It is not surprising to see that United States National Parks are visited a lot throughout the year. Our country is divided about how to protect the parks and how much money should be invested into them. Through looking at the various datasets, we found that National Parks serve as a valuable job and financial source for all the states that they are in but also our country as a whole. In order to generate jobs and revenue from the parks, it is important to know when they are busiest to determine how much staffing is needed and what drives people to the national parks and when. Weather and biodiversity are key components in visitation rates. We found that visitation rates were just slightly higher in Colorado than in California. The most visited national park in 2017 was the Rocky Mountains in Colorado but the next top three parks visited were all in California (Yosemite, Joshua Tree and Death Valley). The visitation rates were highest in all of the parks when the climate was mild and the temperature was between 50 and 75 degrees. Miami, FL had the most extreme weather out of the three four cities that we looked at. Miami had a bunch higher average temperature and Denver had much more fluctuation in the temperature-much colder months than Florida and California. Since Denver has the most fluctuation in temperature, the parks have a higher positive correlation overall because people are less likely to visit when the temperature is below 50 degrees. That being said, Florida parks experience the highest visitation during their coldest months of January and February because the temperature is in the 70s. When it comes to biodiversity, we see that it is greatest in mild climates with four seasons like California but birds are greatest in warmer climates like Florida rather than mammals that do well in colder climates like Colorado. We can conclude that visitation and biodiversity is lowest when there are extreme weather conditions. We also wanted to see how visitation changed over the five years

between 2012 and 2017. The visitation rates in Florida parks and the parks closest to Los Angeles stayed very consistent over the years but the Colorado parks and the ones nearest to San Francisco experience great fluctuation in visitation rates. The visitation would peak and fall in a recurring pattern. In general, we can see that the National Parks are not at risk of losing visitor interest and biodiversity. Further, more people are prone to visit national parks that have more biodiversity present like birds and mammals. If you are planning a trip to the US National Parks, expect for them to be busiest when the weather is mild!

Some next steps for this project could be to look at a broader scale and investigate more parks and analyze the data they have to further enhance our findings. Furthermore, since our analyses are based on 2012-2017, we can expand our timeframe to see if the patterns and findings remain the same. Especially in year 2019 to 2021, when Covid-19 occured, the number of visits and even the weather are more likely to change. Therefore, collecting data on that timeframe for a further analysis would be interesting. Another step our team could potentially take is to look at the weather description dataset. Due to the limited timeline, we had not covered the text analysis for the weather description. Therefore, our team would like to look at descriptive data to see if there are any trending weather description that could potentially affect the visitation and biodiversity of National Parks.

Resources

1. Beniaguev, David (2017).Historical Hourly Weather Data 2012-2017, Version 1. Retriever November 17, 2022 from https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data
2. Elhard, Jay. "Covid-19 Pandemic Causes Impacts and Opportunities for U.S. National Parks." *National Parks Service*, U.S. Department of the Interior, 12 Apr. 2021, https://www.nps.gov/acad/learn/news/pandemic-causes-impacts-and-opportunities-for-u-s-national-parks.htm#:~:text=Findings%20in%20the%20paper%20show,in%20others%20it%20remained%20low.
3. National Park Dataset (2017). Biodiversity in National Parks, Version 1. Retrieved November 17, 2022 from https://www.kaggle.com/datasets/nationalparkservice/park-biodiversity

4. “National park Visitor Spending Contributed $42.5 Billion to the U.S. Economy.” *National Parks Service*, Natural Resource Stewardship and Science Directorate, 23 June 2023, https://www.nps.gov/orgs/1778/vse2021.htm
5. “National Park Visitor Spending Contributed $28.6 Billion to U.S. Economy in 2020.” *National Parks Service*, Office of Communications, 10 April 2021, https://www.nps.gov/orgs/1207/vse2020.htm
6. National Park Visitation Dataset obtained from Module 5 Google Document.
7. https://www.outsideonline.com/health/wellness/whats-best-temperature-productivity/[cite for outdoor activities]