Kerry Oostdyk

Revano Harahap

Shahla Shahnawaz

William Berry

Project 1 for Group 11: U.S. National Parks

Introduction:

The first national park in the United States of America was Yellowstone National Park. It was established by the United States Congress on March 1, 1872. The act was signed into law by President Ulysses S. Grant. This started a movement to preserve public spaces for the pleasure of the public. This movement spread throughout the world to over 100 nations and helped create over 1,200 national parks and preserves.

Originally, Yellowstone National park was under the control of the Department of the Interior. As the years progressed, more national parks and monuments were authorized. Some were under the control of the Department of the Interior. Other parks and monuments were under the control of the War Department (later the Department of Defense) and the Forest Service of the Department of Agriculture.

In 1916, President Woodrow Wilson signed into law the act that created the National Park Service. This new federal bureau was under the Department of the Interior and tasked with protecting the national parks and monuments they managed at the time. An Executive Order in 1933 transferred numerous national monuments and military sites from the Forest Service and the War Department to the National Park Service.

"The National Park System of the United States now comprises more than 400 areas covering more than 84 million acres in 50 states, the District of Columbia, American Samoa,

Guam, Puerto Rico, Saipan, and the Virgin Islands. These areas are of such national significance as to justify special recognition and protection in accordance with various acts of Congress."

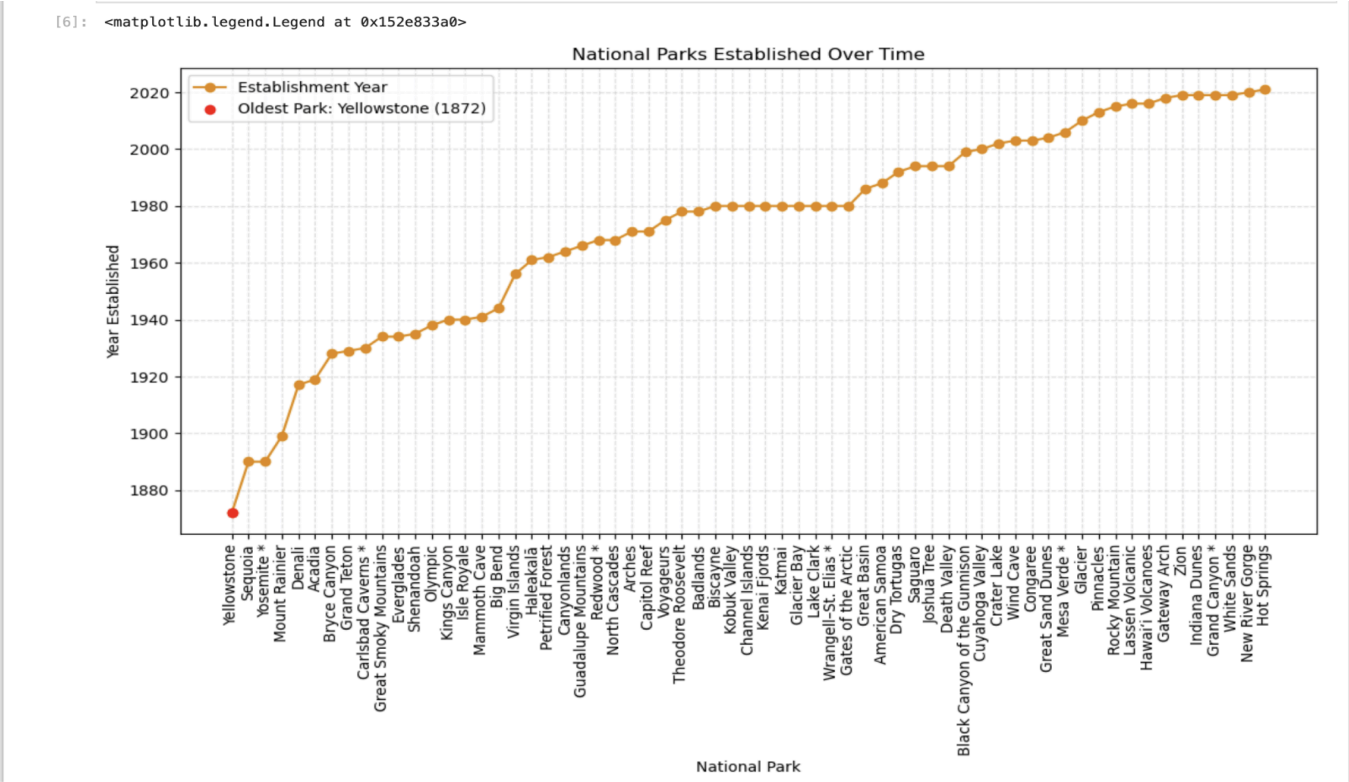
Data Cleaning:

In this project, we performed data cleaning on a dataset of national parks to make it more structured and ready for analysis. The process began by importing the dataset into Jupyter Notebook using pandas "read_csv()" function. Initially, we attempted to merge this dataset with another that contained latitude and longitude values. However, the merge failed due to mismatched column names, so we adjusted our approach. We then focused on cleaning the existing dataset by parsing the "Date established as park[7][12]" column, which was originally an integer, into a proper "datetime" format using pandas "pd.to_datetime()" function with the format "%m/%d/%Y". We also renamed several columns for clarity and consistency, such as changing "Date established as park[7][12]" to "date_established", "Recreation visitors (2021)[11]" to "visitors", and "Area (2021)[13]" to "acres". These changes helped standardize the column names and made the dataset more user-friendly. Additionally, we ensured that column names were in lowercase and used underscores instead of spaces to maintain consistency. These steps collectively improved the structure of the dataset, enabling more effective analysis and visualization in future steps.

Driving Question 1: What are the oldest parks in the country?

This is the graph we came up with, showcasing the establishment dates of national parks across the United States. From the data, we found that the oldest national park is Yellowstone, which was established in March of 1872. This makes Yellowstone not only the first national park in the U.S., but also a significant landmark in the preservation of natural landscapes.

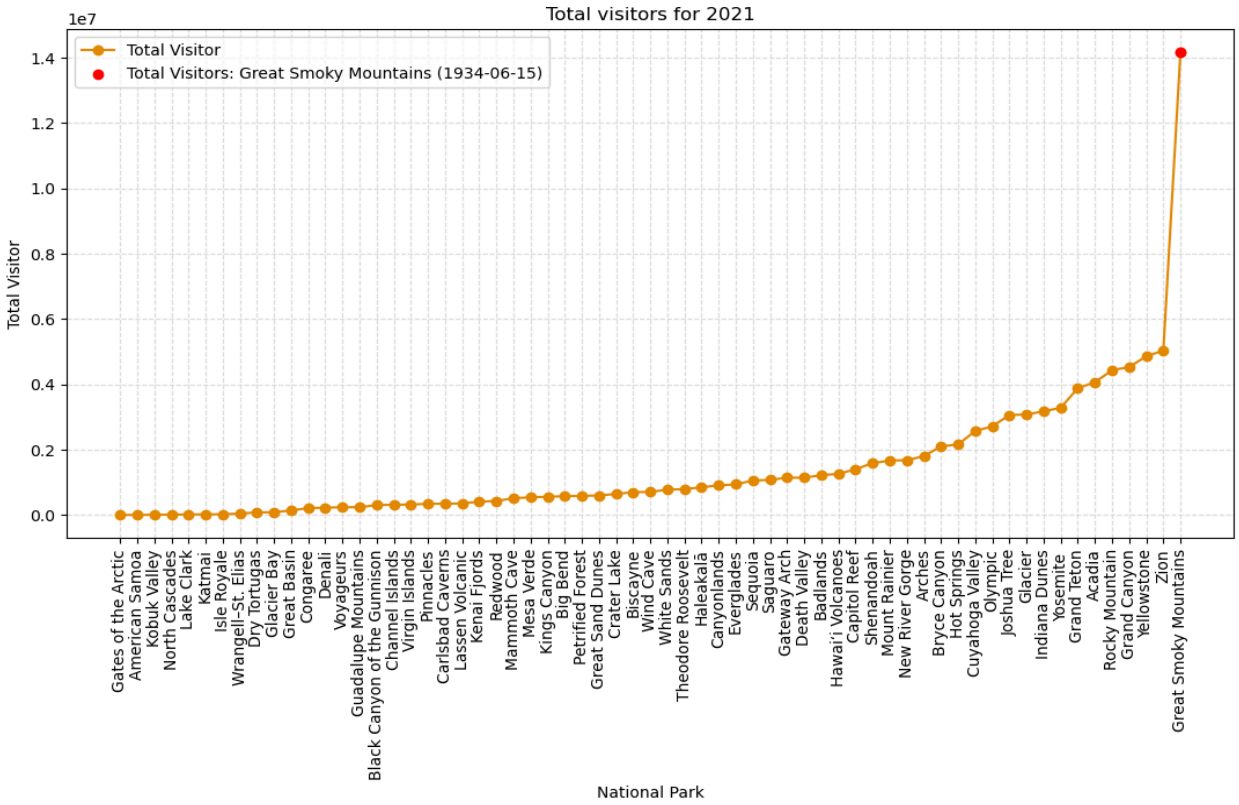
Figure 1



Follow-up Question 1: If the establishment year was related to visitor count?

This is our line chart showing the number of visitors to national parks. The most Visited park is the Great Smoky Mountains National Park, which was established in 1934. Despite having the highest visitor count, it is neither the oldest nor the newest park in the dataset, highlighting that visitor numbers can be influenced by factors other than the park's age.

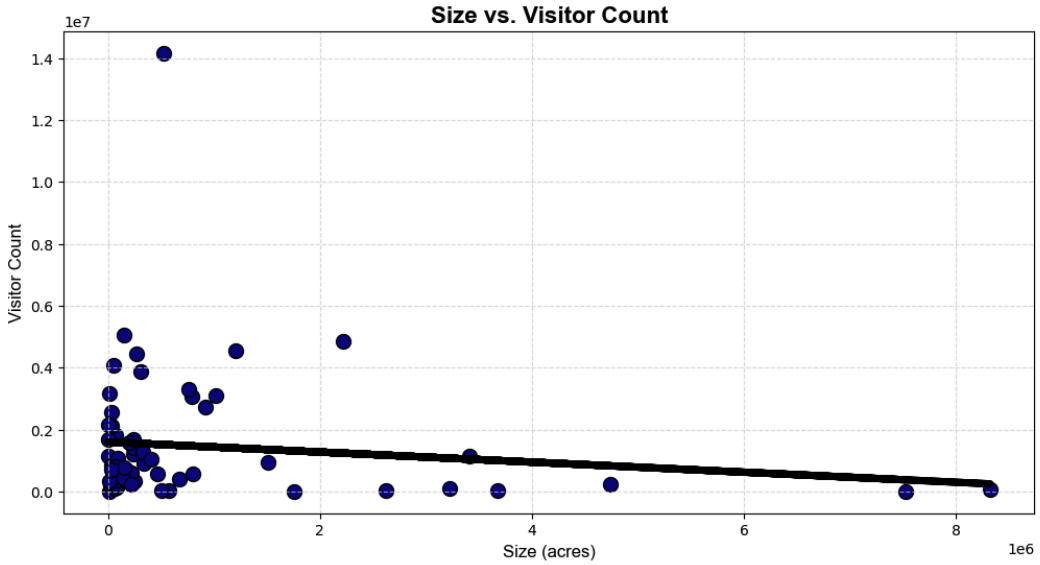
Figure 2



Driving Question 2: How does the size of the park affect the number of visitors?

One of our first thoughts in regards to our project on the national parks is whether or not the size or area of the park is a factor in terms of visitation. We ran a regression on the visitor count and park age. As our r-value of -0.12511101661783258 shows, there is no correlation between visitor count and size. If you did consider this a very small correlation, it would be a negative one. There is a slight chance that the larger the park, the lower the visitor count. Figure 3 is shown below.

Figure 3



We also had several outliers. The Great Smoky Mountain National Park had more than 14 million visitors, far more than any other park in the nation. The top three parks in size all happened to be in Alaska. Our next investigation led us to eliminating these 4 outliers from our dataset and checking the correlation after.

Using the locate function, I found the 3 parks in Alaska and found the Great Smoky Mountain park. You can see this in the code below, Figure 4.

Figure 4

```
In [18]: df.loc[df.visitors>12000000]
```

Out[18]:

	national_park	date_established	visitors	description	state	latitude	longitude	acres	year_established	u
27	Great Smoky Mountains	1934-06-15	14161548	The Great Smoky Mountains, part of the Appalac...	North Carolina, Tennessee	35.683333	-83.533333	522426.88	1934	

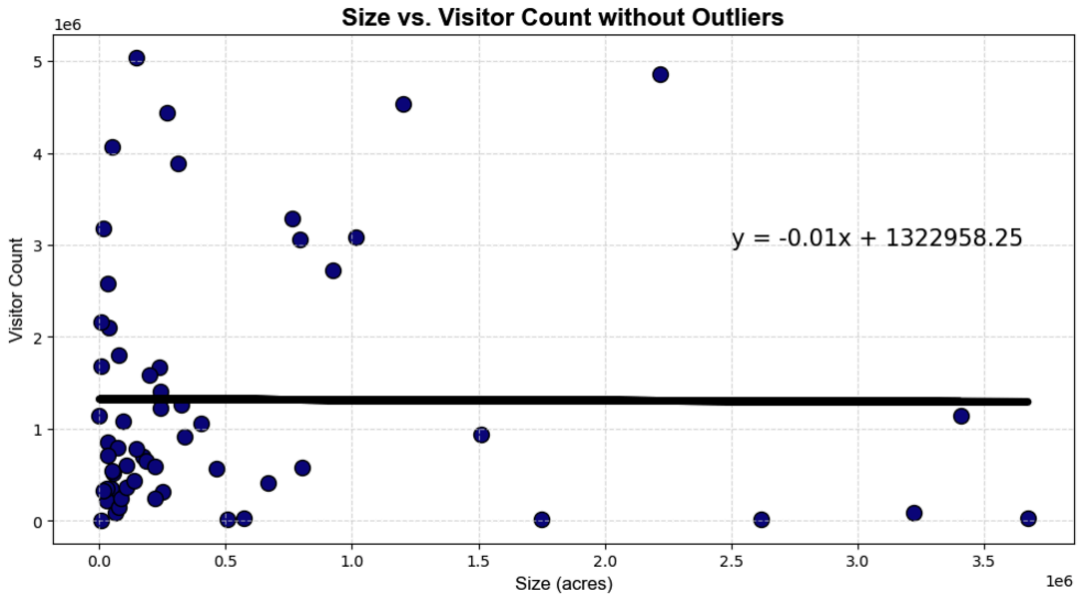
```
In [19]: df.loc[df.acres>4000000]
```

Out[19]:

	national_park	date_established	visitors	description	state	latitude	longitude	acres	year_established	us_loc

We created a new data frame where we dropped these outliers. After dropping them, we ran the regression over again getting a new r-value of -0.005031628703678563. This is a very low correlation, even less than before the outliers were removed. In **Figure 5** below, there is a nearly flat regression line indicating no correlation.

Figure 5



Driving Questions 3.1: How are the national parks distributed across the nation?

Our next driving question led us to map the locations of our National Parks across the United States and analyze the visitor counts based on their location. Today, there are 63 National Parks in the United States of America and its territories. Another interesting fact is the six states with the most parks are in the western half of the United States.

See Figure 6 and 7.

Figure 6

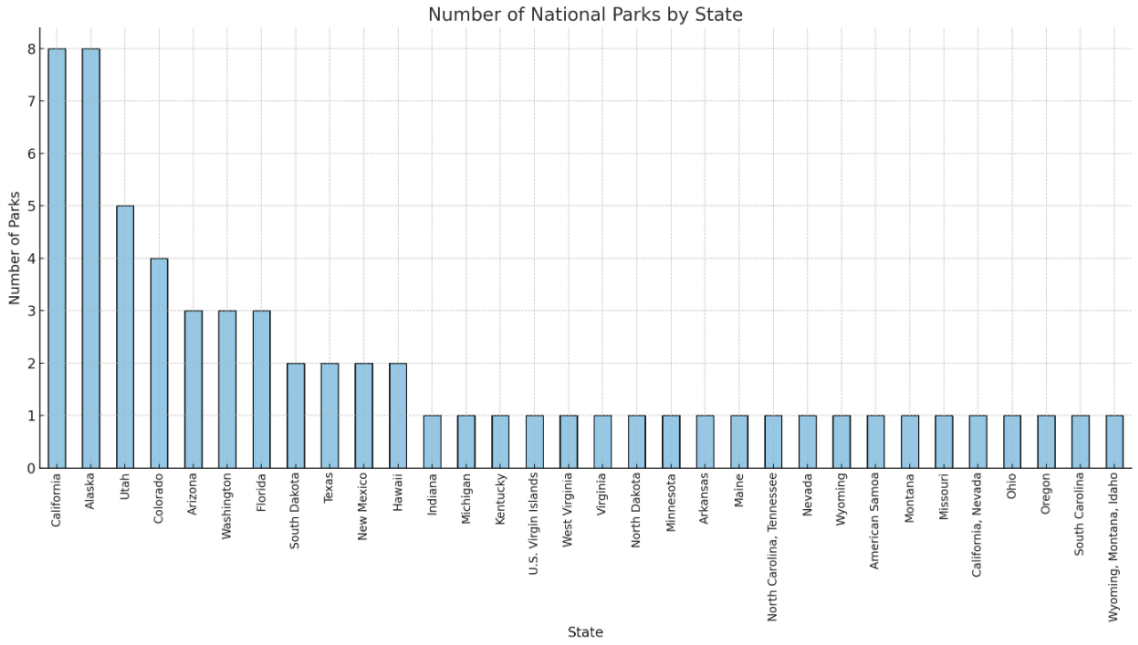
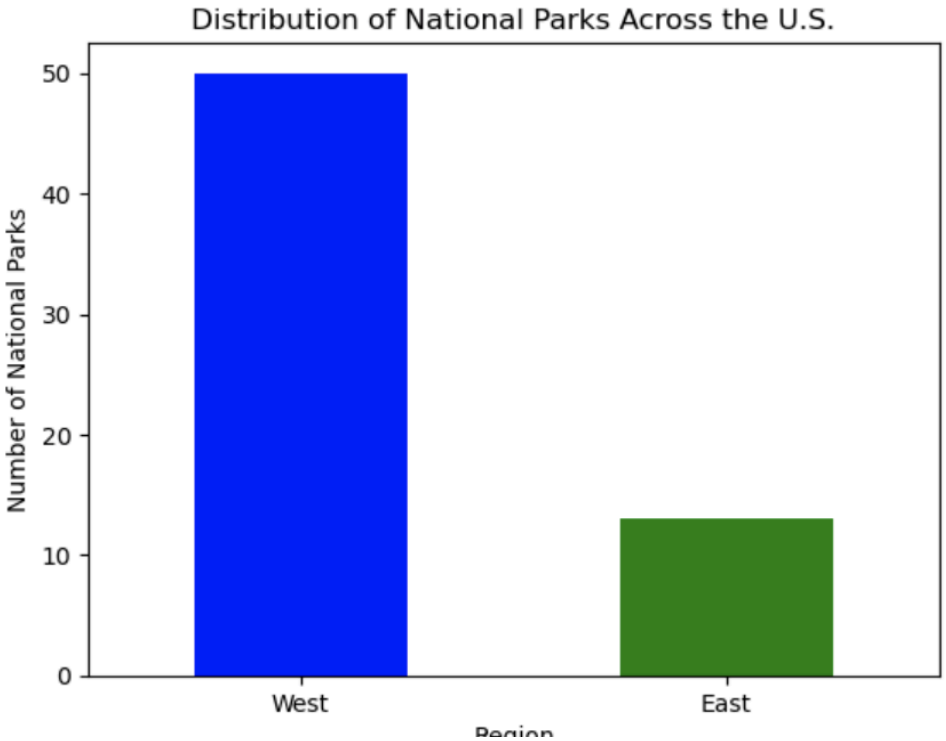

Number of National Parks by State
Number of Parks
0
1
2
3
4
5
6
7
8
California
Alaska
Utah
Colorado
Arizona
Washington
Florida
South Dakota
Texas
New Mexico
Hawaii
Indiana
Michigan
Kentucky
U.S. Virgin Islands
West Virginia
Virginia
North Dakota
Minnesota
Arkansas
Maine
North Carolina, Tennessee
Nevada
Wyoming
American Samoa
Montana
Missouri
California, Nevada
Ohio
Oregon
South Carolina
Wyoming, Montana, Idaho
State

Figure 7


Distribution of National Parks Across the U.S.
Number of National Parks
0
10
20
30
40
50
West
East
Region

The cluster of visitors in the east is far closer together than the western half of the nation. We discovered this when we ran several maps using the code below in Figure 8. The resulting maps turned out to be a great visual. See Figure 9 and 10.

Figure 8

```
]:
# Configure the map plot_4
import hvplot.pandas
map = df.hvplot.points(
    "longitude",
    "latitude",
    geo = True,
    tiles = "CartoLight",
    frame_width = 1000,
    frame_height = 700,
    size = "visitors",
    scale = 0.01,
    color = "national_park"
)

# Display the map plot
map
```

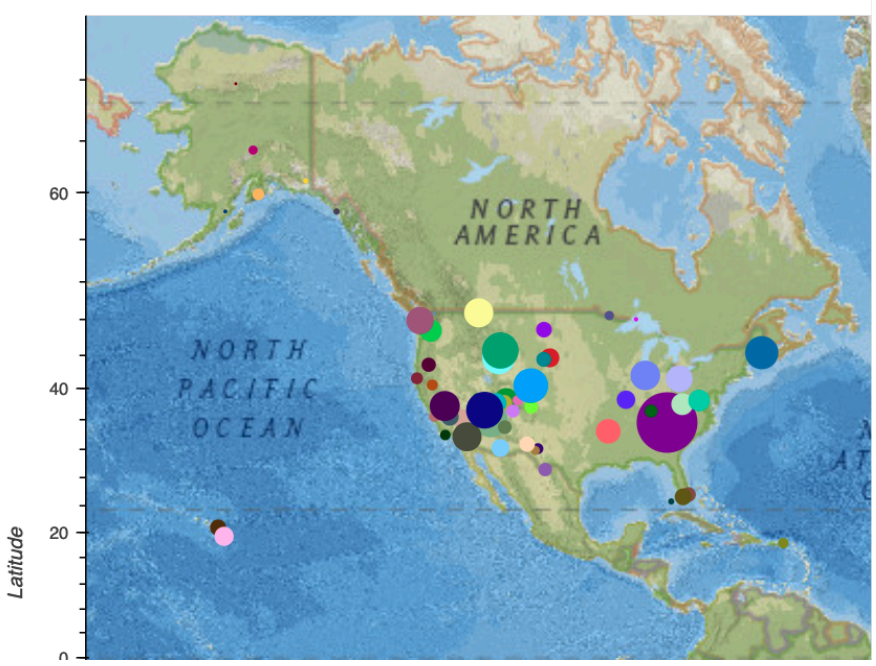
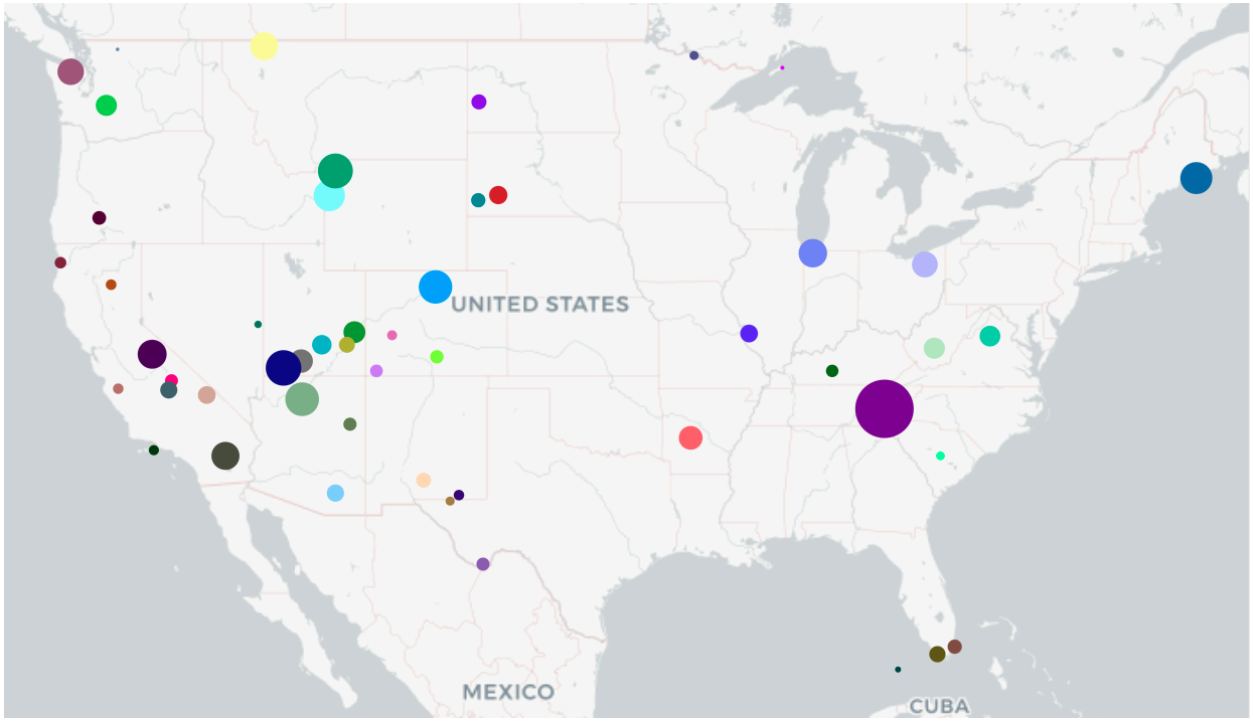


Figure 9

Figure 10



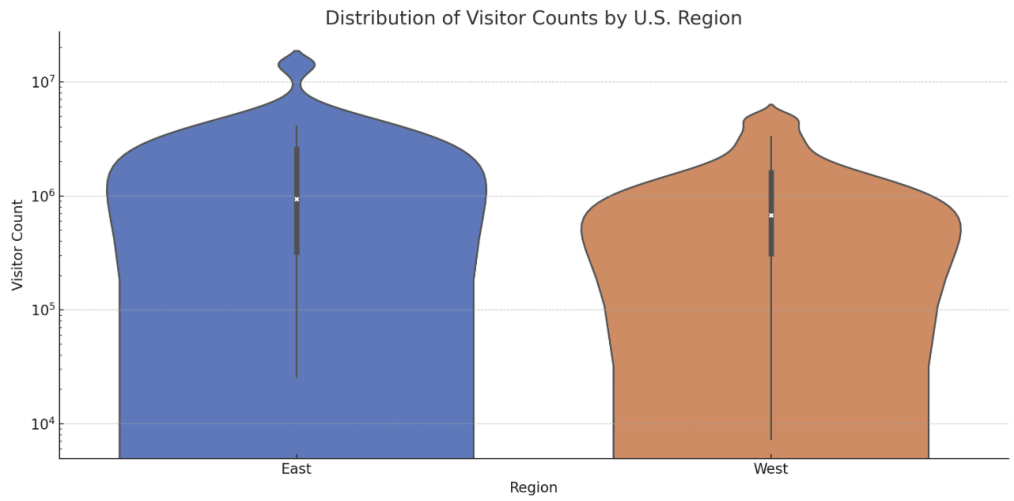
We also discovered that parks in the eastern half of the United States had more visitors compared to parks in the western half of the United States. We also noticed that the Western United States visitors had a more normal distribution than the East side.

See Figure 11.

Driving Question 3.2: Which states have the highest density of national parks?

While the violin chart doesn't directly address density, it can reveal how visitor counts vary by region. Parks in higher-density states (e.g, California, Utah) show diverse visitor counts, with some very popular parks contributing significantly to the overall spread seen in the violin plot.

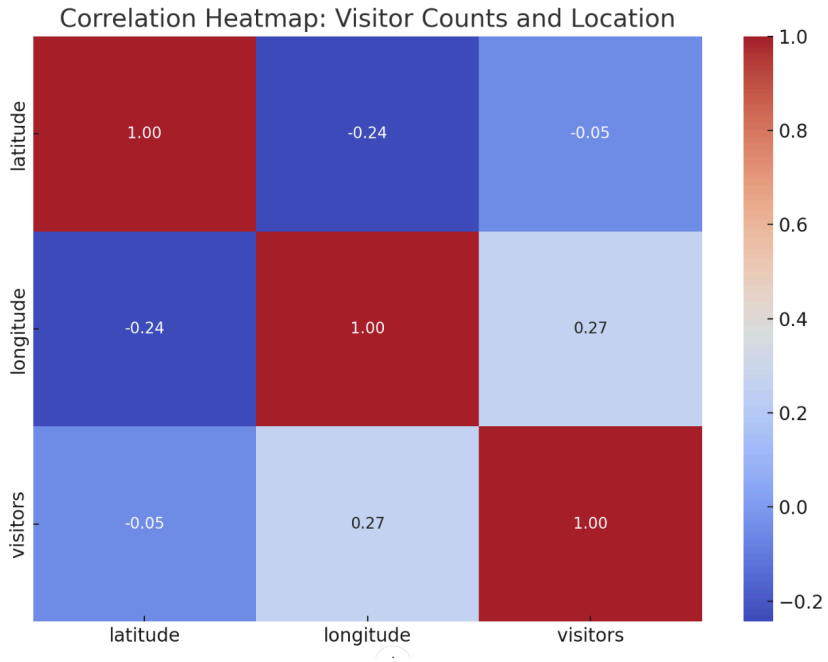
Figure 11



Driving Question: 3.3 Are visitor counts and park location related? (Heatmap)

The heatmap of latitude, longitude, and visitor counts also revealed that there is no strong correlation between park location and visitor numbers. Visitor counts are probably more influenced by factors like accessibility, popularity, and nearby population centers rather than geographic coordination.

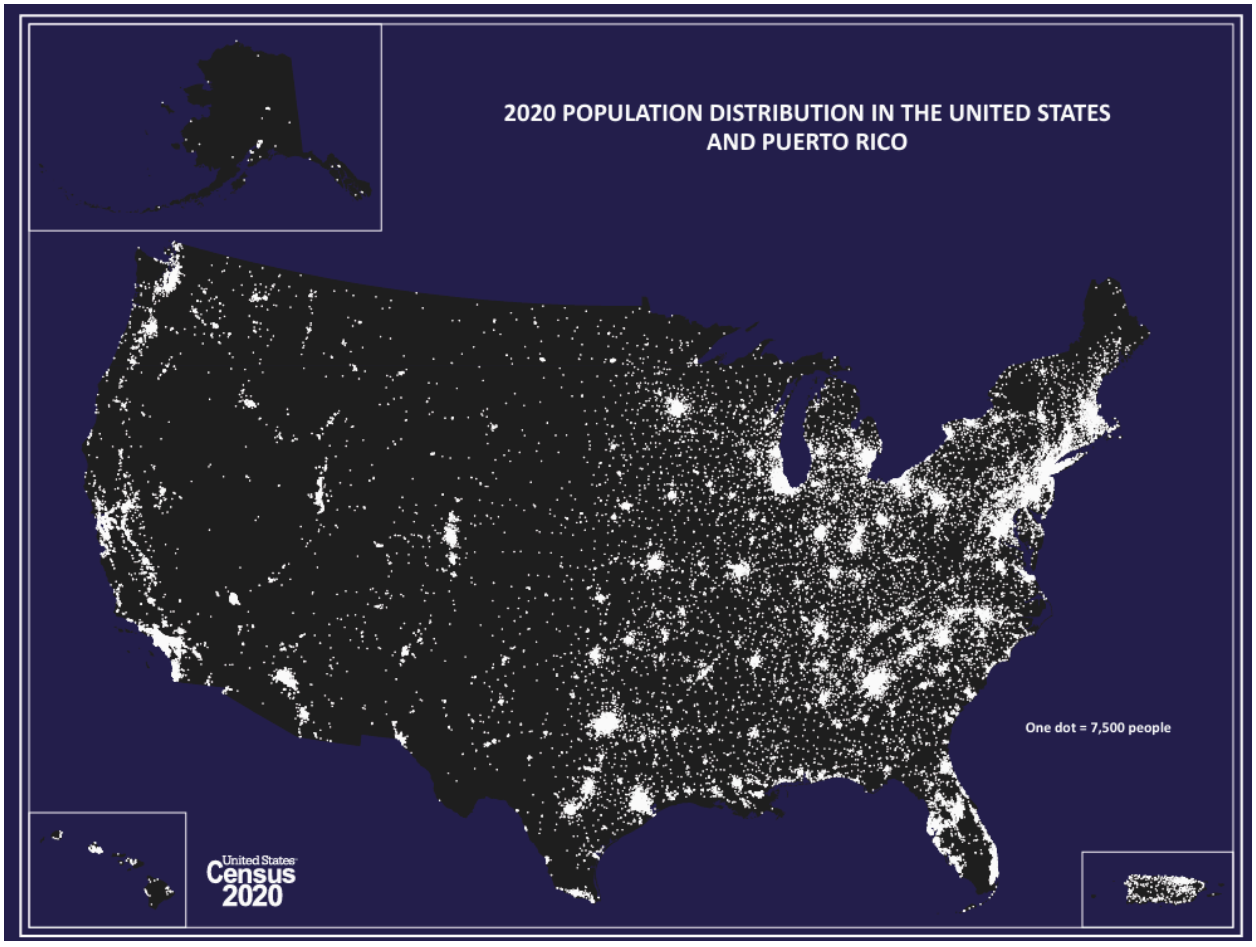
Figure 12



We can see from the U.S. Census Bureau Population Distribution that more people live in the eastern half of the United States compared to the western half. This may impact visitation of the eastern parks compared to the western parks. In terms of the Smoky Mountain National Park and its very large number of visitors is due to the park being close to two-thirds of the U.S. population. Numerous major cities are within a day's drive of the park (3)

See Figure 12 from the US Census Bureau website. (2)

Figure 12



T-Test

Our final analysis was calculating a T-test to compare the visitor mean between the east and western half of the nation (divided by the Mississippi river). We calculated the mean and ran the regression using Scipy statistics. The results showed a visitor mean of 2313120.23 for the eastern half of the nation and a visitor mean of 1243625.88 for the west. The results of the t-test are as follows, "T-testResult(statistic=1.0024858646380677, p value=0.3345930382425417, df=12.864694047021516)". Since our p value is at 0.33 we cannot reject the null hypothesis, meaning the visitor mean is not statistically different between the two sides of the country. This may be due to the large variance in the data.

Bias and Limitations

There were several biases and limitations to our data set. First there was a temporal bias because visitors counts are skewed by our data so close to covid (2021). We also had a timeframe bias being that our data is only from the year 2021. We don't know how much our visitor count is affected by Covid. One year of data is only a small snippet in time for our data conclusions leading to a timeframe bias. Finally, we may have had a visualization bias. Maps may distort perceptions of park proximity or size, especially in areas with dense clusters: East Coast vs. expansive areas in the West.

Conclusion and Future Work

In the future work, we could increase the scope by including data over more years. Since our data was from 2021, just after the Covid pandemic, we could try to see visitation level before and after the pandemic to see what impact it caused. Another aspect we could consider is economics.

Do times with higher inflation impact park visitation compared to periods of lower inflation?

In conclusion, we know the Great Smoky Mountain National Park was the most visited park in 2021. We know there are more parks in the west compared to the east. However, the parks in the east have higher visitation on average, but this is not a statistically significant data point. The oldest national park is Yellowstone. And, we know park age and size do not impact visitation.

Works Cited.

OpenAI. ChatGPT. OpenAI, www.openai.com. Accessed 25–26 Nov. 2024 and 2–3 Dec. 2024.

Slides Carnival. *Free PowerPoint Templates and Google Slides Themes*. Slides Carnival, www.slidescarnival.com. Accessed 25–26 Nov. 2024 and 2–3 Dec. 2024.

Kaggle. *The United States National Parks*. Kaggle, www.kaggle.com. Accessed 25–26 Nov. 2024 and 2–3 Dec. 2024.

United States Census Bureau. *2020 Population Distribution*. Census Bureau, 2020, www2.census.gov/geo/maps/DC2020/PopDist_Nighttime/2020_Pop_Distribution_Post_pg.pdf. Accessed 25–26 Nov. 2024 and 2–3 Dec. 2024.

National Park Service. "Quick NPS History." *National Park Service*, www.nps.gov/articles/quick-nps-history.htm. Accessed 25–26 Nov. 2024 and 2–3 Dec. 2024.

National Park Service. "Yellowstone Protection Act of 1872." *National Park Service*, www.nps.gov/yell/learn/management/yellowstoneprotectionact1872.htm. Accessed 25–26 Nov. 2024 and 2–3 Dec. 2024.

"Smoky Mountains: Most Visited National Park." *Smoky Mountain National Park*, Smoky Mountain National Park, smokymountainnationalpark.com/blog/smoky-mountains-most-visited-national-park/. Accessed 25–26 Nov. 2024 and 2–3 Dec. 2024.