Visualizing, Exploring, and Forecasting National Park Visitation Traffic

Anh Tran Bryan Truong Chase Diaz Olivia Hadlaw Priyal Patel

Summary & Motivations

Despite the growing popularity of U.S. National parks, there currently lacks a single source that provides a comprehensive visualization of each U.S. national park and their visitation. Our project provides a solution that allows users to easily view and interpret historical visitation numbers for multiple parks at once, as well as forecast future visitation numbers.

The National Park Service Dataset

The NPS shares park visitation data publicly on their website here. The dataset contains over 28,000 rows, with each row detailing the name of the national park, date (month and year), ranging from 1979-2020. Parks that were missing monthly data were removed. To visualize the data, we also added columns for the latitude and longitude of each row's national park.

Data Exploration & Trends

We discovered several trends that impacted our modeling and data processing decisions. When looking at average visitation per park based on month, we found that summer months tend to have the highest visitation, with July and August having the highest visitation (Figure 1). Since seasonality is clearly a factor that impacts visitation, we opted to train both a Holt-Winters and Seasonal ARIMA (SARIMA) model. When considering the date ranges for training and testing, we had to consider the impact of the COVID-19 pandemic on visitation (Figure 2). In 2020, a sharp drop in the rolling mean and standard deviation are observed, though the trend returns to its pre-pandemic pattern towards the end of 2021. We opted to omit 2020-2021 data from the training data, as the pandemic visitation numbers are not indicative of 2022 (and beyond) post-pandemic visitation habits.

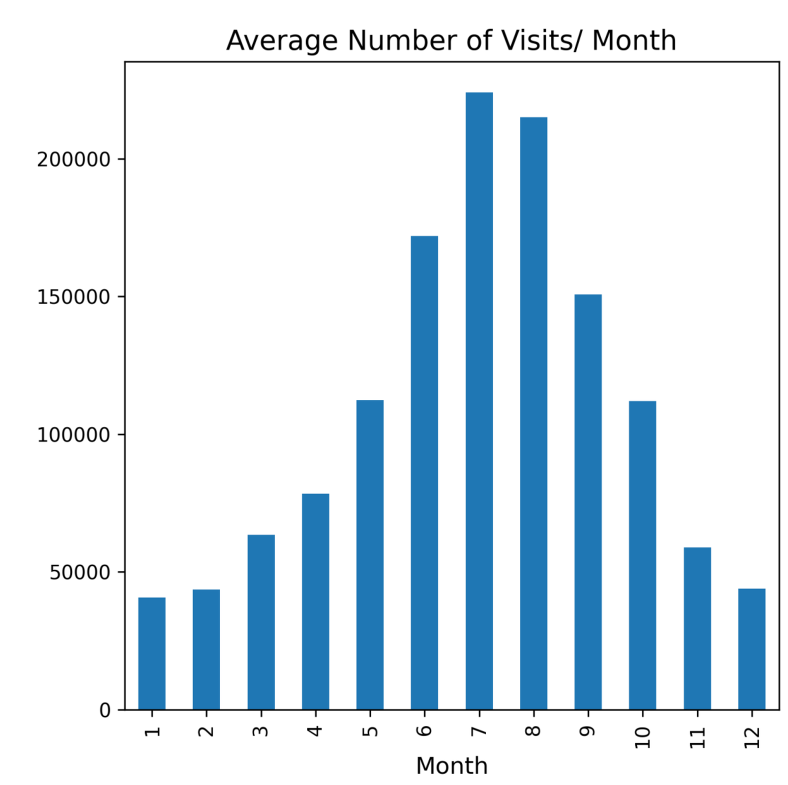

Figure 1.

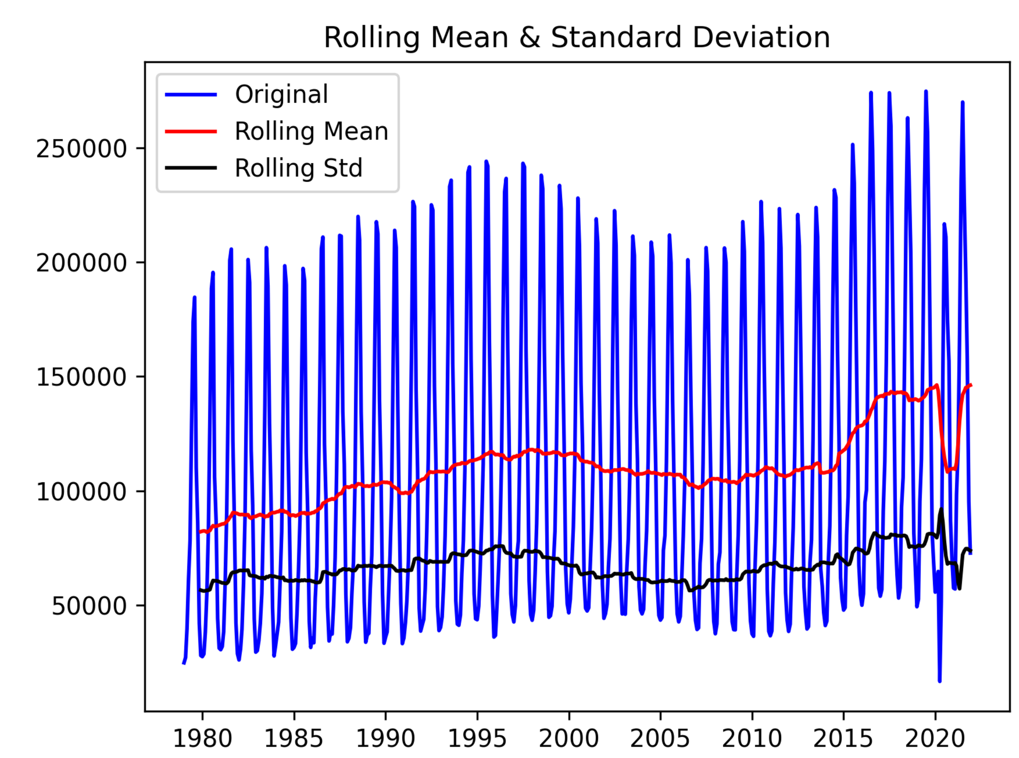

Figure 2.

Modeling Visitation

After cleaning and exploring the data, we then trained two different time-series models: a Holt-Winters and a SARIMA time-series model. Both were trained with data from 1979-2018 and tested against the 2019 data. In the SARIMA model, trend and seasonal elements were tuned as hyperparameters for each national park and then were used for forecasting. After evaluating the predictions from both models, we decided upon the SARIMA model, as it can capture the trends and seasonality and had a lower mean absolute percentage error than the Holt Winters model.

Creating a Data Visualization Tool

Having trained our model and tested it against 2019 data, we produced additional forecasts for each national park from 2020-2023. We then created a public Tableau dashboard to allow users to visualize and interact with the park data. The dashboard is hosted and accessible by anyone with internet access. The dashboard provides a variety of different views:

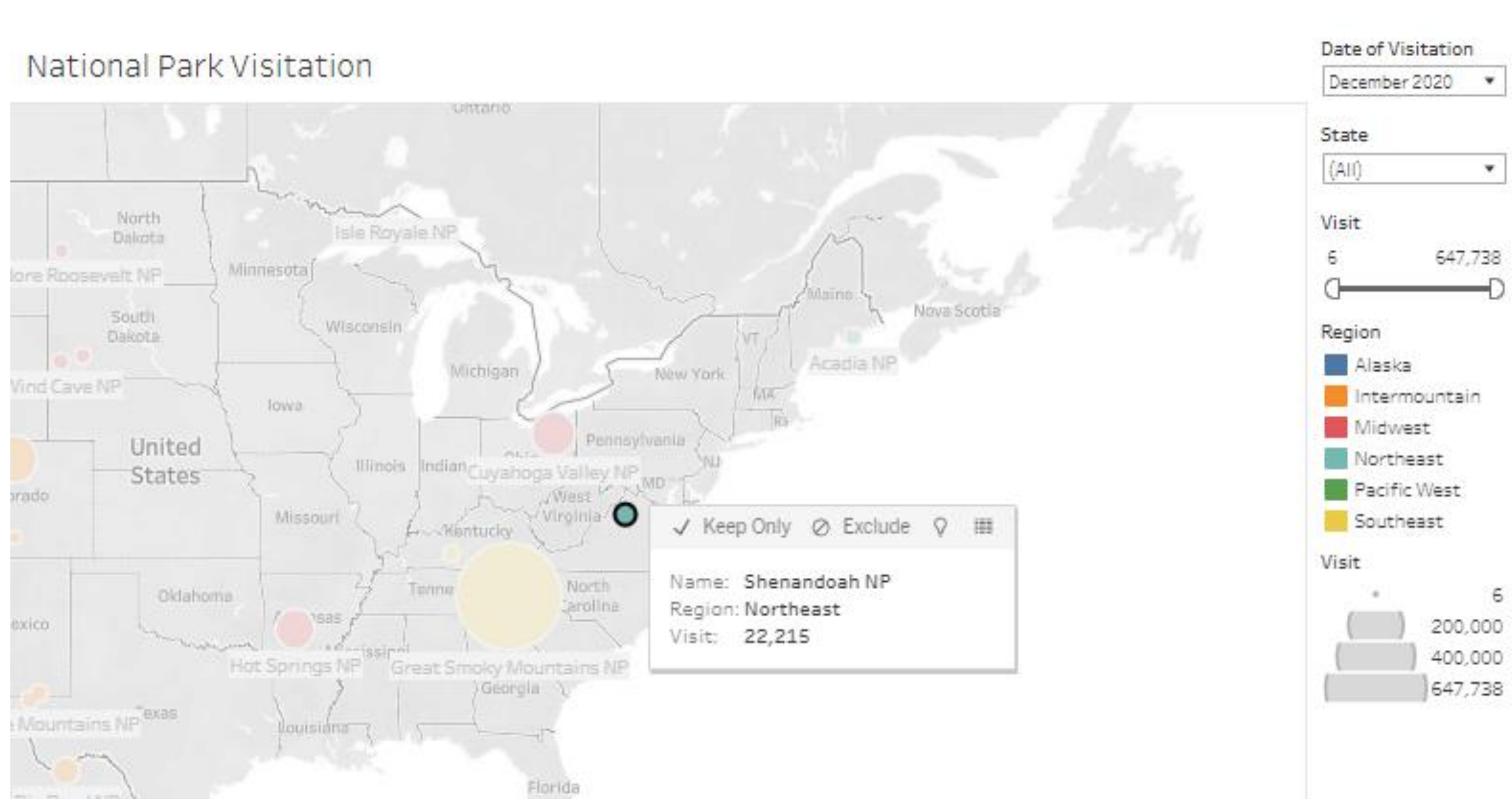

Figure 3: Tableau visualization. Visitation month/year is selected via drop-down menu. Colors indicate region, with the size of the circle indicating visitation numbers. Exact visitation/forecasted visitation is available on hover, with options to keep only/exclude on click.

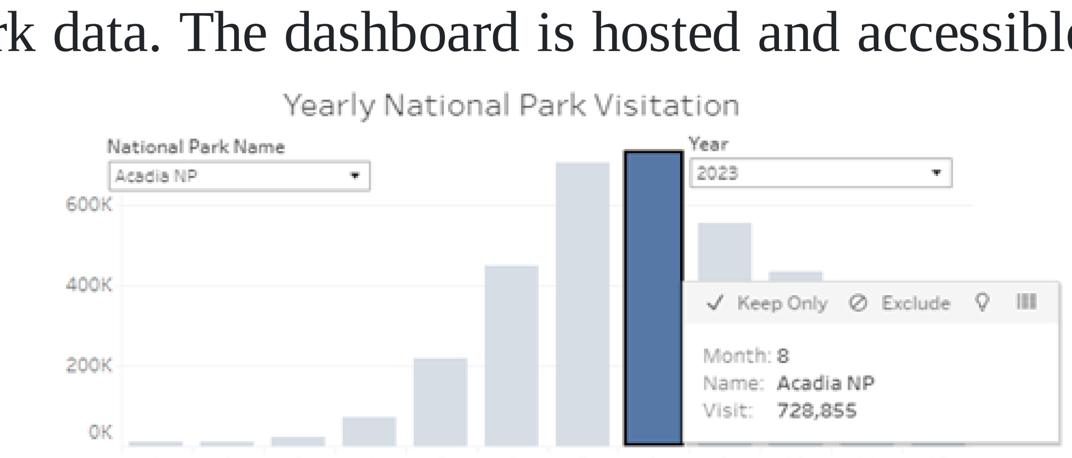

Figure 4: The Tableau visualization summarizes the yearly trend for each month based on the selected year and national park. The tooltip includes the month, national park name, and the exact visitation numbers/forecasted numbers.

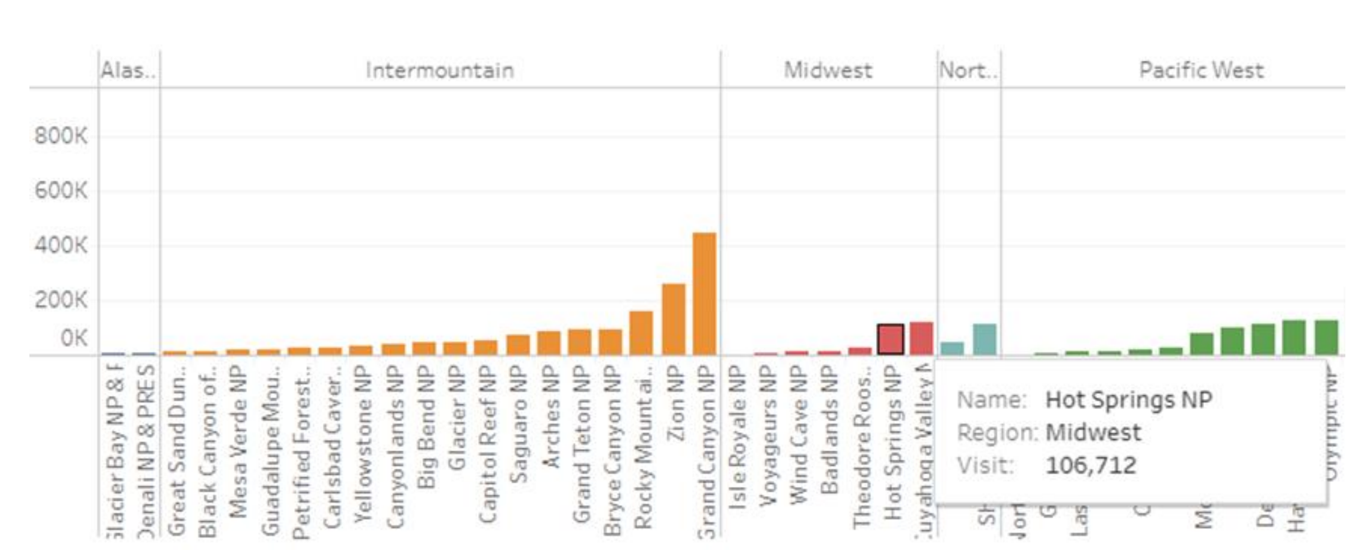

Figure 5: The Tableau visualization also summarizes the most popular national parks for the selected date of visitation by region. Hovering shows exact visitation numbers/forecasted numbers.

Experiments + Evaluation

To evaluate the performance of the model against the 2019 data (which was our test set), several accuracy metrics were used. Most notably, these included Mean Absolute Percentage Error (MAPE), Mean Absolute Deviation (MAD), Mean Forecast Error (MFE), and Mean Absolute Error (MAE). MAPE, which is scale-agnostic, is the primary metric that we used to evaluate the performance of our model. The overall mean MAPE of all national park forecasts in 2019 was 17%, which is below the 20% threshold that we used to define "good" or "acceptable" predictive accuracy. While we always would prefer a lower measure of error (we initially defined "success" as a MAPE below 10%), we are nevertheless pleased with our results, as we do not want to overfit our model to the 2019 data. The MAPE of 17% indicates that our model does indeed provide meaningful predictions. To evaluate the visualization and user interface, we asked individuals from our personal networks who are interested in traveling to national parks to participate in a survey after utilizing the visualization. 95.2% of participants said they would utilize the visualization to make travel decisions.