# Springboard- Data Science Course

# Capstone Project 2

## Analyzing Camping Sites for Bookings Availability

**Priyanka Panhalkar**

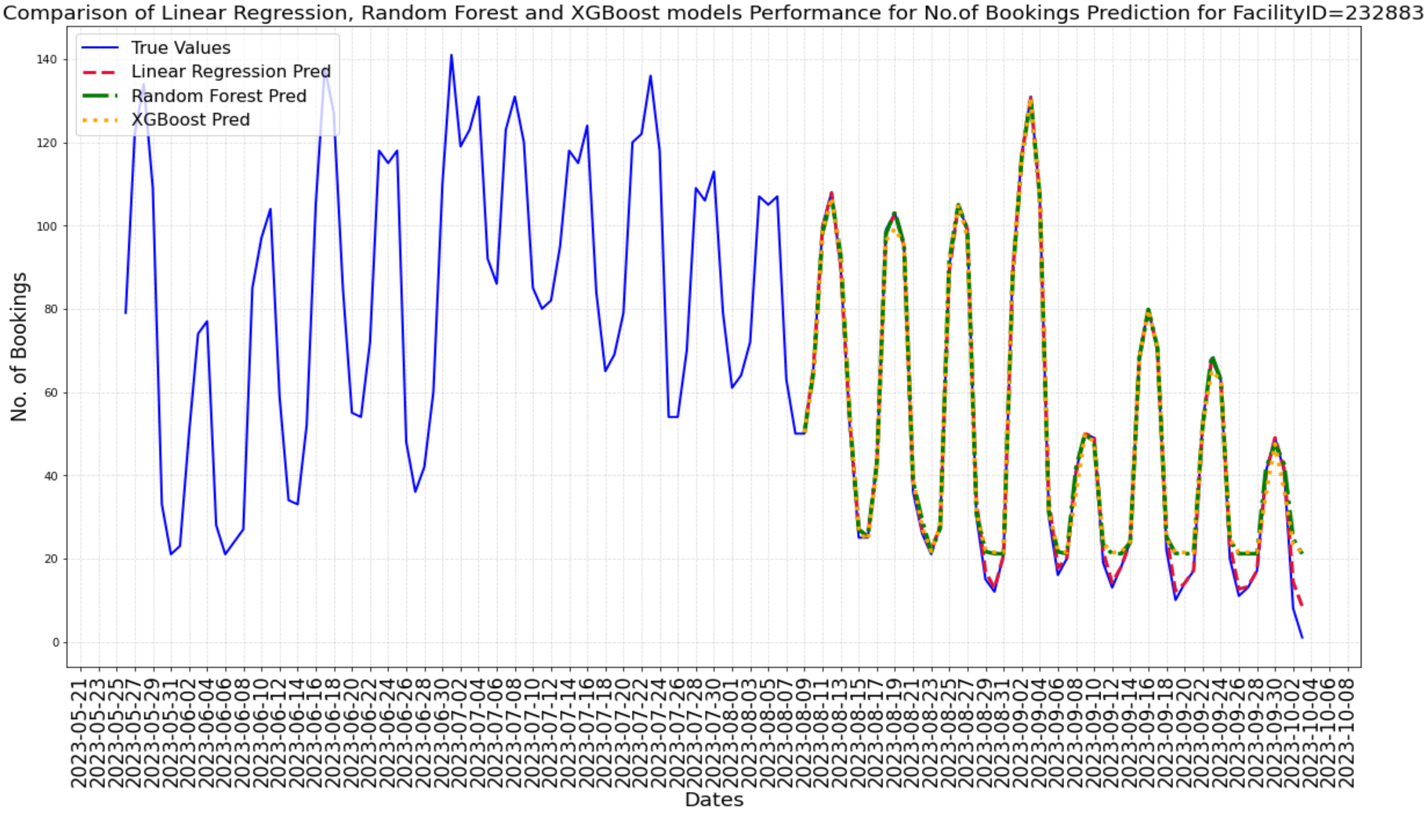
**February, 2025**

# Introduction

**Business Problem**: Booking a desired campsite can be challenging in peak season. People who want the camping sites often don't get it because of overcrowding at certain sites and some campsites are always less crowded. To predict which campsite is available to book on which time of the year depending on a given FacilityId.

**Intended Stakeholders**: Recreation.gov is looking to improve their revenue by implementing tools for users. They can add visibility to the users to deter them from overcrowding famous sites. Also recommend facilities similar in nature which are expected to be less crowded for a given date.

**Data Science Results**: We generated a time series graph for a given **Facility ID** to forecast the expected number of bookings on a specific date. This allows users to identify which facilities are likely to be crowded on a given day. To enhance prediction accuracy, we compared three models—**Linear Regression, Random Forest, and XGBoost.** Among these, **Linear Regression** demonstrated the best performance based on **Mean Absolute Percentage Error (MAPE)**. An example graph is shown below.



Implementation details can be found in this Github repository. Link

# Approach

## Data Acquisition and Wrangling:

Dataset is available from Link. Recreation gov has a dataset available for the last 18 years. It is huge so we decided to sample down post pandemic data as after that people were more aware about outdoor recreational activities. After data acquisition, we wrangled the data and found the results below.

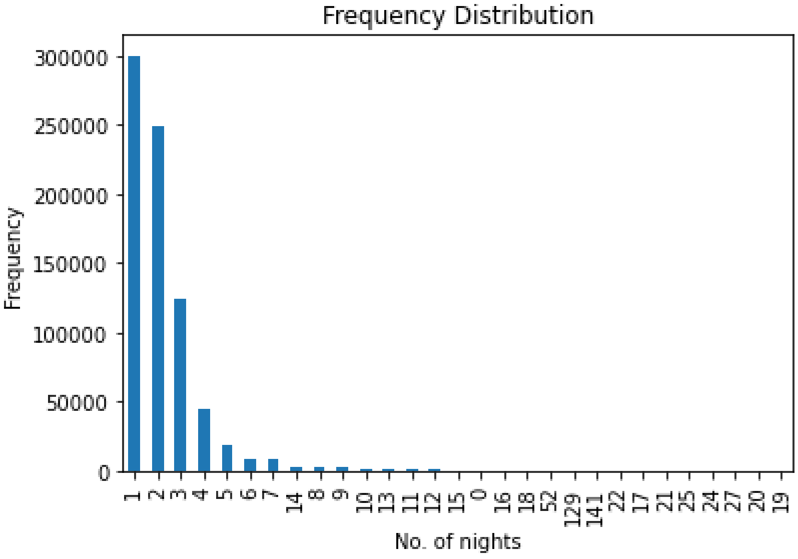
- There were about 35 columns and 8984096 rows for the year 2023 of the United States.
- We found the data types and null percentage in each column.
- The correlation of each column to another was found and a map was drawn for the parent child relationships between various columns. Link
- For our goal of time series forecasting we need to understand how the data is related with the time. eg frequency of data, weekly, monthly etc. We have order date, start date and end date for bookings in our dataset.
- The output of wrangling was a robust understanding of the meaning of each column, relation to other columns and our target (date) columns for building time series forecasting.

## Storytelling and Inferential Statistics:

Understanding booking trends at a camping site is crucial for optimizing resource allocation, predicting demand, and improving customer experience. This analysis explores historical booking data, applies inferential statistics, and visualizes trends to uncover insights.

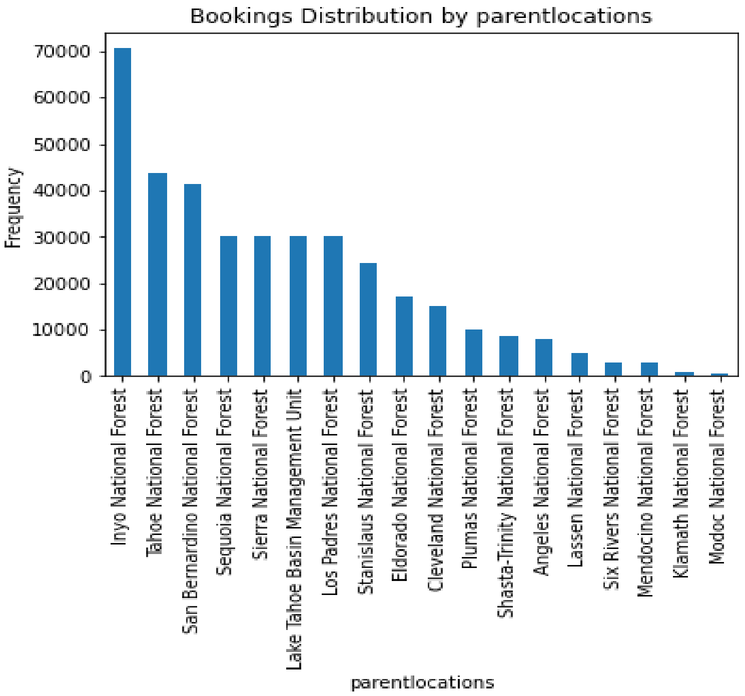
We decided to study a subset of data specific to California State in the US.

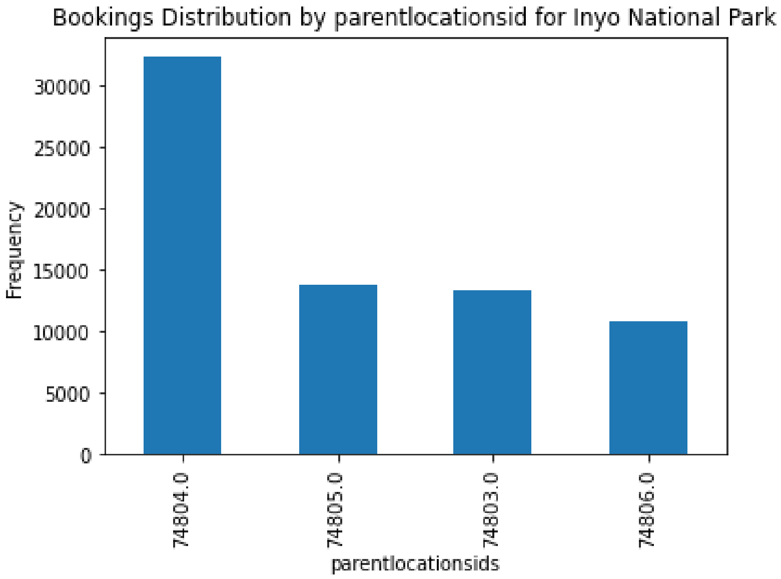
- Frequency distributions were analyzed for all columns relevant to our project goal. For example, the frequency distribution of **'Number of Nights for Camping'** (shown below) indicates that the majority of bookings were made for **1, 2, and 3 nights**.



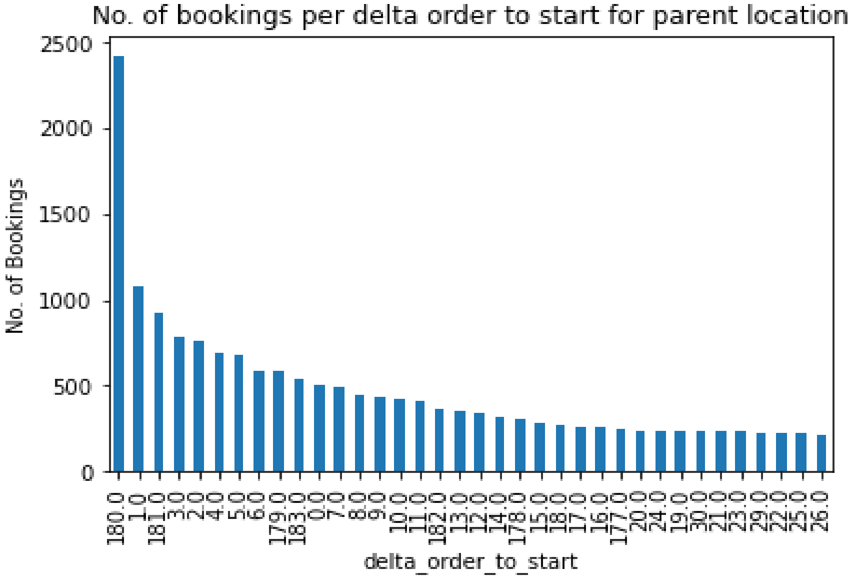
Similarly, frequency distributions were plotted for other columns, such as **Number of People, Booking Agency, and Equipment Used**, to gain further insights into the data.

- An analysis was conducted to determine the number of unique values in various columns, such as **total unique parks, regions, site types, and parent locations**.
- Graphs were plotted to visualize the data distribution for each unique **Parent Location**, as shown in **Figure 1**. Additionally, similar distributions were analyzed to explore hierarchical relationships between other columns. For example, **Figure 2** illustrates the number of unique **Parent Location IDs** within a single parent location, such as **Inyo National Park**.





- Additional calculated columns were created to support our project goals. For example, the **difference between the order date and the start date of a booking** was computed (as shown below). This analysis revealed that the vast majority of bookings were made approximately **180 days in advance**.

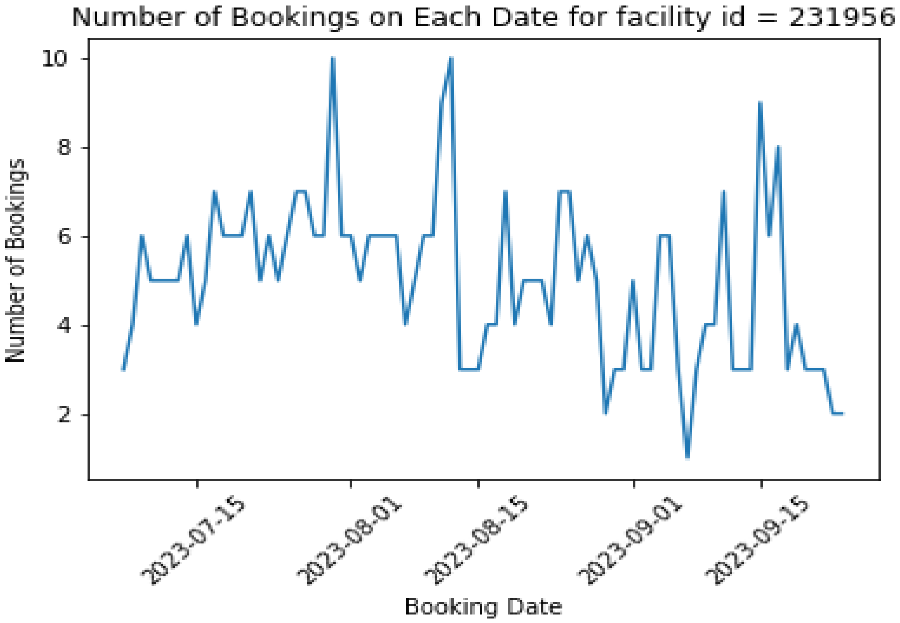


- Additional calculations were carried out to determine the number of sites booked on a given date. This analysis provides further insights into the occupancy levels and booking trends for each date.In this case since the target of our problem was not given explicitly, we generated the targets by performing below calculations.
- To better understand campsite booking trends and availability, we **transformed the original dataset** by expanding each booking's duration into **individual daily records**. This transformation enables a **time-series analysis** of how many bookings are active on any given date. Below is a detailed explanation of the process:

  The dataset contains information on bookings, including the **start date** (`startdate`) and the **number of days the booking lasts** (`delta_end_to_start`).The delta_end_to_start is a calculated column by taking delta between start and end dates. Instead of just knowing when a booking starts and how long it lasts, we aim to generate a dataset where each row represents **each individual day the booking is active**. Example below:
  If a booking starts on **February 1, 2024**, and lasts for **3 days**, the generated dates will be:
  2024-02-01
  2024-02-02
  2024-02-03
  2024-02-04

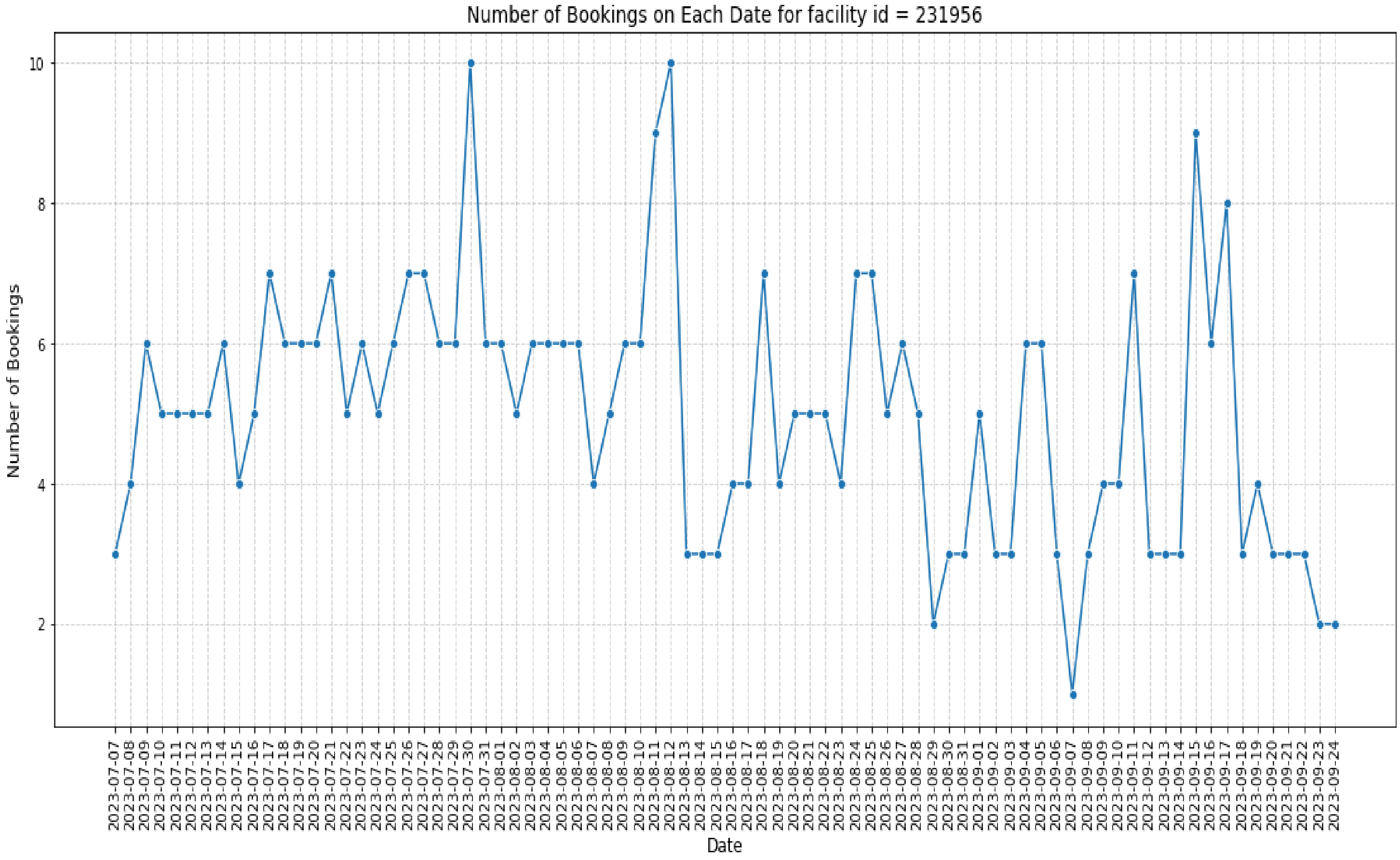
As per the above example the target (Number of Bookings) per facilityID was calculated for a given date.

- In the final stage of our **Exploratory Data Analysis (EDA)**, a **time series graph** was generated for a specific **Facility ID**, illustrating the number of bookings on a given date. This visualization helps in understanding booking trends and patterns over time. The output of EDA was a pickle file which we will use for the next step of Baseline Modeling.



## **Baseline Modeling**:

We build a time series model with Number of bookings on each given date for a given Facility ID as shown below.



- Furthermore, we created **lagged features** using **minimum, maximum, mean, and standard deviation**, which were used as inputs for our models to enhance forecasting accuracy. We then compared the performance of three models—**Linear Regression, Random Forest, and XGBoost**—to identify the most effective approach.
- Functions were defined to perform **sequential train-test splitting** and **model evaluation**, enabling the generation of performance metrics for assessing model effectiveness.
- At the end of Baseline Modeling, a **time series plot** was generated to compare the **true number of bookings** with the predicted values from the three models—**Linear Regression, Random Forest, and XGBoost**. This visualization helps in understanding how each model performs against the actual data, highlighting trends, deviations, and overall model behavior. (as shown in the Data Science Results section above).
- All Models were used without hyperparameter tuning.

## **Extended Modeling**:

In this section, we **refactored** the previous code into **reusable functions** to efficiently generate **time series charts** displaying actual bookings alongside predictions from different models. Additionally, we performed a **comparison of algorithms** based on **Mean Absolute Percentage Error (MAPE)** to evaluate their performance.

At the end, the main objective was to **eliminate repetitive coding** by refactoring the logic into **reusable functions**. This approach allows us to dynamically generate **time series charts** and compute **model performance metrics** for each unique tuple value, ensuring efficiency and scalability in our analysis.

Below is a detailed set of functions created to eliminate repetitive coding by refactoring the logic into reusable functions which generated Time Series Charts. Link

1. Import the raw csv and filter the dataset by required variables.
2. Create Column transformations and add new calculated columns.
3. Create Lagged Features.
4. Perform sequential train and test split of the dataframe.
5. Function to see if time gaps exist and resolve it.
6. Define and Evaluate 3 Models- Linear Regression, Random Forest and XGBoost.
7. Finally, create a function by running a loop which has tuples as input for defining variables and generate a times series chart as output.

**Summary of Results**: The **Linear Regression Model** demonstrates superior **forecasting accuracy** compared to **Random Forest** and **XGBoost**, as reflected in its lower **Mean Absolute Percentage Error (MAPE)** and closer alignment with actual booking trends.

# Findings

Here is the Summary table comparing MAPE scores across five Facility IDs for the three models:

**Model Comparison for MAPE (Test Set) Score Across Facility IDs**

| Model | FacID 1 | FacID 2 | FacID 3 | FacID 4 | FacID 5 |
|---|---|---|---|---|---|
| Linear Regression | 0.41% | 0.97% | 2.61% | 0.88% | 1.8% |
| Random Forest | 13.35% | 17.26% | 6.68% | 2.16% | 2.5% |
| XGBoost | 13.42% | 17.89% | 7.36% | 2.71% | 3.49% |

| FacID | Dataset Size | Training Set Size | Test Set Size |
|---|---|---|---|
| 234329 | (21020) | (14713,5) | (6307,5) |
| 232858 | (23590) | (16513, 5) | (7077, 5) |
| 232883 | (9069) | (6348, 5) | (2721, 5) |
| 232781 | (13693) | (9585, 5) | (4108, 5) |
| 232782 | (12500) | (8750, 5) | (3750, 5) |

## Discussion of Results

### 1. Model Performance Comparison

The **comparison of MAPE scores** across different **Facility IDs** reveals key insights into the relative strengths and weaknesses of each model:

- **Linear Regression consistently delivers the best performance**, with MAPE scores below **3% across all Facility IDs**. This suggests that **booking demand follows a fairly linear pattern**, which the model captures effectively.

- **Random Forest and XGBoost exhibit significantly higher MAPE scores**, particularly for **FacID 1 and FacID 2**, where errors exceed **13-17%**. This indicates that these models might be **overfitting to noise** or struggling with the complexity of the data.
- **XGBoost slightly underperforms compared to Random Forest**, which could be due to **suboptimal hyperparameters** or **insufficient tree depth for generalization**.

Given these observations, **Linear Regression emerges as the best model for forecasting campsite bookings**, as it provides **the most reliable predictions** with **the best performance for the time series that were investigated** .

## 2. Potential Business Implications

The results have **direct applications for campsite management and planning**:

- **Improved Booking Availability Predictions**: By accurately forecasting demand, campsite administrators can **better allocate resources**, such as staff and facilities, during high-demand periods.
- **Optimized Pricing & Promotions:** Understanding peak booking patterns allows for **dynamic pricing strategies**, offering incentives for bookings during low-demand periods.
- **Capacity Planning & Expansion:** Insights from the forecasts can guide decisions on **adding more campsites, improving reservation policies, and adjusting booking windows** to enhance visitor experience.
- **Operational Efficiency:** With a **reliable demand prediction model,** campsite operators can **reduce last-minute cancellations and overbookings**, ensuring a smoother booking process for customers.

# Conclusions and Future Work

## Conclusions

This project aimed to forecast **campsite booking availability** to aid in **capacity planning, resource allocation, and customer experience optimization** for **National Forest campsites in California.**

For the time series we investigated, the **Linear Regression model** provides the **most accurate forecasts** for campsite booking availability. Its **low error rate** ensures **trustworthy predictions**, making it a potentially valuable tool for campsite operators to **enhance planning, optimize operations, and improve customer satisfaction**.

## Future Work

Given more time and resources, the following enhancements could improve the forecasting system:

**Feature Engineering Enhancements:**

- Incorporate **external factors** such as **weather data, holidays, and local events** to improve prediction accuracy.
- Introduce **more lagged variables** to capture deeper time-dependent trends.

**Geographical Expansion:**

- Extend the analysis to **other regions beyond California** to validate model performance across different National Forests.

**Hyperparameter Optimization:**

- Fine-tune **Random Forest** and **XGBoost** models using **Grid Search** or **Bayesian Optimization** to improve their performance.

**Exploration of Additional Models:**

- Test **Prophet (by Facebook)** for a more robust time series forecasting approach.

These improvements would enhance the predictive power of the model and further align it with real-world **operational and business needs.**

# Recommendations for the Clients

Based on the findings of our analysis, we propose the following **actionable recommendations** to help campsite operators **optimize bookings, improve planning, and enhance customer satisfaction:**

- **Implement Demand-Based Booking Strategies:**
    - Use the **Linear Regression model's predictions** to adjust reservation policies, such as **increasing campsite availability** during peak booking periods and offering discounts for low-demand periods.
    - Identify high-demand dates and encourage early bookings to **prevent overbooking and last-minute cancellations.**
- **Enhance Operational Planning & Resource Allocation:**
    - Utilize the forecasted booking trends to **optimize staffing levels, maintenance schedules, and facility availability** based on projected demand.
    - Allocate additional resources to **popular facilities** while redistributing excess capacity from **low-demand locations** to balance bookings across different sites.
- **Develop a Dynamic Pricing & Promotion Strategy:**
    - Introduce **dynamic pricing** based on predicted demand, offering discounts for less crowded dates to **improve occupancy rates.**
    - Use the insights from the model to create **targeted marketing campaigns,** promoting facilities with lower forecasted demand through special offers and bundles.

By implementing these strategies, campsite operators can **maximize revenue, improve the customer booking experience, and ensure efficient resource utilization.**

# Consulted Resources

1. For dataset- https://ridb.recreation.gov/download
2. For Learning
   https://www.youtube.com/@statquest
   https://www.analyticsvidhya.com/blog/2021/05/greykite-time-series-forecasting-in-python/
   https://robjhyndman.com/hyndsight/#category=forecasting
   https://www.youtube.com/@krishnaik06
   https://www.youtube.com/@codebasics
   https://www.youtube.com/@Stats4Everyone
   https://www.youtube.com/@machinelearningplus
3. Datacamp Resources
   https://app.datacamp.com/learn/courses/introduction-to-time-series-analysis-in-python
   https://app.datacamp.com/learn/courses/machine-learning-for-time-series-data-in-python