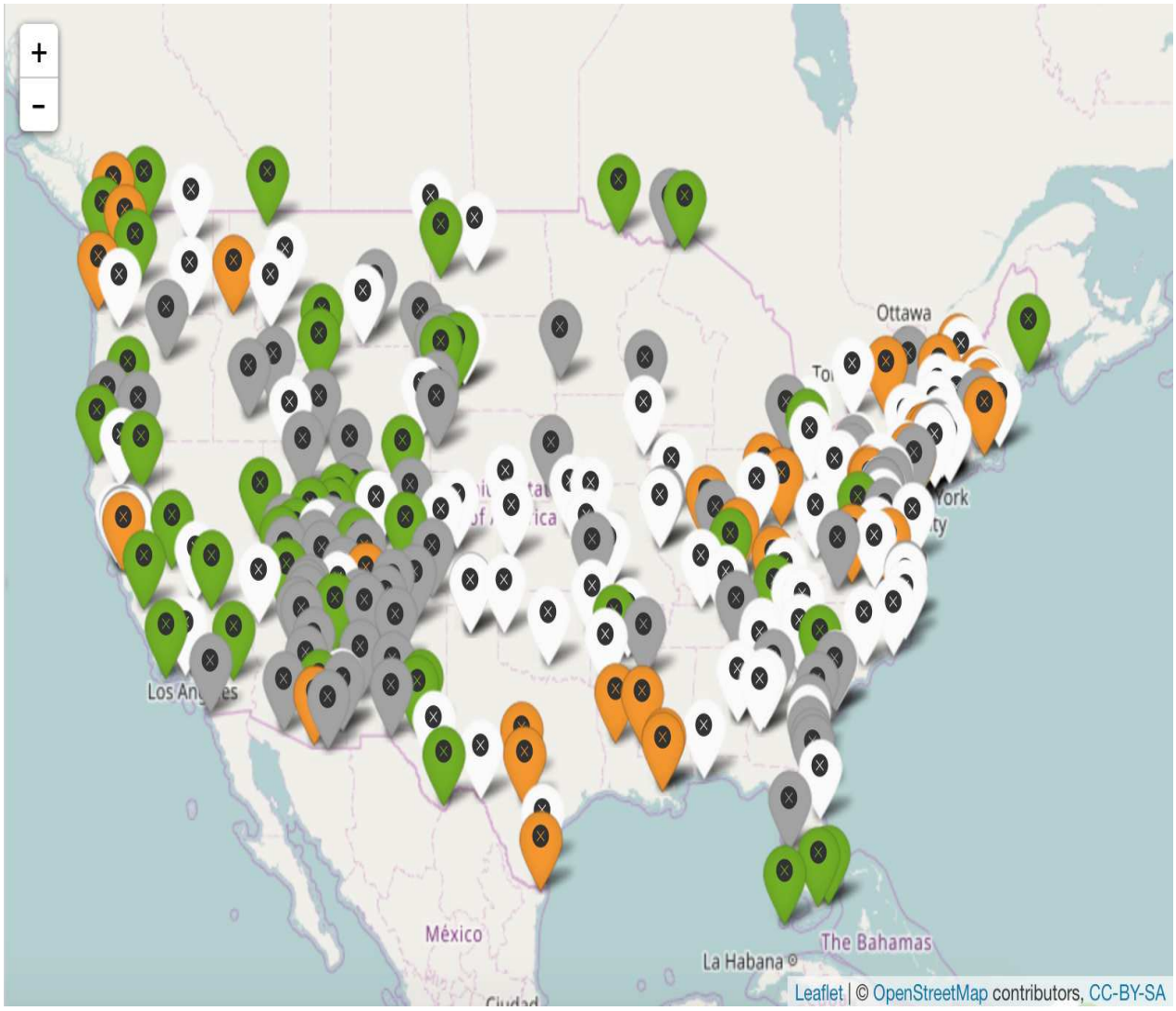
# FORECASTING NATIONAL PARKS VISITS

The U.S. National Parks System includes 417 areas including national parks, monuments, battlefields, military parks, historical parks, historical sites, lakeshores, seashores, recreation areas, scenic rivers and trails, and the White House (see map in Figure 1). Every year, hundreds of millions of recreational visitors come to the parks. What do we know about the parks that can affect the visitor counts? Can we forecast the monthly visits to a given park accurately? To derive insights and answer these questions, we take a look at the historical visits data and the parks information released by the National Parks Service (NPS).

**Figure 1:** A map of the U.S. National Parks System areas. Green: National Parks; Grey: National Memorial/National Monument; Orange: national Historical Park/Site; White: others. Made with *leaflet* package in R with NPS data.




For this problem, we obtained monthly visits data between 2010 and 2016 (source: https://irma.nps.gov/Stats/Reports/National). We also got park-specific data via the NPS API (https://www.nps.gov/subjects/developer/get-started.htm). The aggregated dataset park_visits.csv results in a total of 12 variables and 25587 observations. Each observation contains one record per park per month. Here's a detailed description of the variables:

- **ParkName**: The full name of the park.

- **ParkType**: The type of the park. For this study we restrict ourselves to the following more frequently visited types: National Battlefield, National Historic Site, National Historical Park, National Memorial, National Monument, National Park, National Recreation Area, and National Seashore.

- **Region**: The region of the park, including Alaska, Intermountain, Midwest, National Capital, Northeast, Pacific West, and Southeast.

- **State**: The abbreviation of the state where the park resides.

- **Year**, **Month**: the year and the month for the visits.

- **lat**, **long**: Latitude and longitude of the park.

- **Cost**: a simple extraction of the park's entrance fee. Some parks may have multiple levels of entrance fees (differ by transportation methods, age, military status, etc.); for this problem, we only extracted the first available cost information.

- **logVisits**: Natural logarithm of the recreational visits (with one added to the visits to avoid taking logs of zero) to the park in the given year and month.

- **laglogVisits**: the logVisits from last month.

-**laglogVisitsYear**: the logVisits from last year.

---

## Problem 1 - Number of National Parks in Jan 2016

2.0/2.0 points (graded)
Load park_visits.csv into a data frame called visits.

Let's first look at the visits in July 2016. Subset the observations to this year and month, name it visits2016jul. Work with this data subset for the next three problems.

Which park type has the most number of parks?

- (●) National Historic Site
- ( ) National Historical Park
- ( ) National Monument
- ( ) National Park

✔

Which specific park has the most number of visitors?

- ( ) Yellowstone NP
- ( ) Golden Gate NRA
- (●) Great Smoky Mountains NP
- ( ) Cape Cod NS

✔

Submit You have used 2 of 2 attempts

## Problem 2 - Relationship Between Region and Visits

0.0/3.0 points (graded)

Which region has the highest average log visits in July 2016?

- ◯ Intermountain
- ◉ National Capital
- ◯ Pacific West ✔
- ◯ Southeast

✖

What is the average log visits for the region in July 2016 with:

1. the highest average log visits?

14.1969 ✖ **Answer:** 10.767849

$14.1969$

2. the lowest average log visits?

10.0293 ✖ **Answer:** 9.374157

$10.0293$

**Explanation**
You can answer this question by using the tapply function on the visits by region using mean.

Submit You have used 3 of 3 attempts

ℹ Answers are displayed within the problem

## Problem 3 - Relationship Between Cost and Visits

2.0/2.0 points (graded)
What is the correlation between entrance fee (the variable cost) and the log visits in July 2016?

0.401061 ✔ **Answer:** 0.4010611

$0.401061$

Choose the most reasonable possible answer from the following statements:

- ○ Higher entrance fees are associated with lower log visits, likely because visitors are cost sensitive
- ◉ Higher entrance fees are associated with higher log visits, likely because more expensive parks are often more popular due to other features of the parks
- ○ There is no association between entrance fees and the log visits

✔

**Explanation**
Use the cor function to solve this question.

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

## Problem 4 - Time Series Plot of Visits

1.0/1.0 point (graded)
Let's now look at the time dimension of the data. Subset the original data (visits) to "Yellowstone NP" only and save as ys. Use the following code to plot the logVisits through the months between 2010 and 2016:

```
ys_ts=ts(ys$logVisits,start=c(2010,1),freq=12)
```

plot(ys_ts)

What observations do you make?

- [x] Between the years, the shapes are largely similar.
- [x] The log visits are highly cyclical, with the peaks in the summer time.
- [ ] There is a trend of substantial increase in log visits over recent years.

✔

Submit You have used 2 of 2 attempts

---

## Problem 5 - Missing Values

2.0/2.0 points (graded)

Note that there are some NA's in the data - you can run colSums(is.na(visits)) to see the summary.

Why do we have NA's in the laglogVisits and laglogVisitsYear? These variables were created by lagging the log visits by a month or by a year.

- ○ The dataset inevitably have missing data due to human entry negligence.
- ◉ These are lagged variables and the earlier data is not available for the first months.
- ○ The values were outliers and therefore removed.

✔

To deal with the missing values, we will simply remove the observations with the missing values first (there are more sophisticated ways to work with missing values, but for this purpose removing the observations is fine). Run the following:

visits = visits[rowSums(is.na(visits)) == 0, ]

How many observations are there in visits now?

21855 ✔

21855

Submit You have used 2 of 2 attempts

---

## Problem 6 - Predicting Visits

0.0/3.0 points (graded)

We are interested in predicting the log visits. Before doing the split, let's also make Month a factor variable by including the following:

visits$Month = as.factor(visits$Month)

Subset our dataset into a training and a testing set by splitting based on the year: training would contain 2010-2014 years of data, and testing would be 2015-2016 data.

Let's build now a simple linear regression model "mod" using the training set to predict the log visits. As a first step, we only use the laglogVisits variable (log visits from last month).

What's the coefficient of the laglogVisits variable?

✖ **Answer:** 0.927945

What's the out-of-sample R2 in the testing set for this simple model?

0.881 ✖ **Answer:** 0.8975923

0.881

**Explanation**
Run the linear regression with lm and look at the summary.
Then calculate the out-of-sample R2 using the test data.

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

## Verified Track Access

Graded assessments are available to Verified Track learners.