Capstone Project: Battle of Neighborhoods

Segmentation and clustering of National Parks in United States

By,

Manasi Khare

Battle of National Park neighborhoods in the United States

Table of Contents

1. Introduction

1.1 Background

The national parks in the United States have been called “the best idea” by writer and historian Wallace Stegner. Beginning in the 1800s, the scenic natural wonders of the West, majestic mountains and trees in Yosemite, and arid ruins in Casa Grande, have inspired individual Americans to call for their preservation, and asking their government to create something called “National Parks”.

In 1872, Yellowstone became the world’s first national park, and today there are almost 400 national parks across the nation.

With so much wilderness available to explore, along with innumerable opportunities to care for the environment, preserve history and revitalize communities, National Parks in the United states attract tens of thousands of visitors every year.

1.2 Problem

Planning visit to a national park involves exploring venues in and around the park area and choosing the best options for individual(s). This can be a tedious work as it can involve looking for several websites, as well as social networking sites to get recommendations, ratings etc.

1.3 Stakeholders

This project is aimed at helping those who would like to gather venue resource information and equip the visitors with more in-depth analysis of venues around the National Parks in United States.

2. Data acquisition and cleaning

2.1 Data acquisition

A comprehensive dataset for US National Parks is available for public domain on public.opendatasoft.com website here. I downloaded the complete Excel dataset and loaded it into the Pandas dataframe. The dataset contains complete list of National parks, national monuments, national recreation areas, historic site etc. with their geographical location, ID, and other metadata. This is ~ 67MB of data.

2.2 Data cleaning and wrangling

First, I checked for any duplicate entries so that our results are not biased. In this dataset, there were no duplicate entries; therefore, no additional work was required to remove duplicates.

Second, the geographical coordinates of each of the national parks were in the (latitude, longitude) format under a column names 'Geo Point'. I split this column into two columns, namely Latitude and Longitude. This will be helpful further in data analysis.

The raw dataset also consisted of lot of additional information about each of the parks, such as shape of the park, ID, Code, Created by field etc. Since this additional data was not necessary for data analysis, I deleted those columns from our dataframe.

The dataset also consisted about 31 missing values. Since I wanted to analyze only factual data, missing values could not be predicted or assigned. Therefore, I simply ignored entire rows that had missing value(s).

Next, I checked if the data types in the pandas dataframe are consistent and see if we need to correct any datatypes. It appeared that the Latitude and Longitude values in the pandas dataframe were of 'Object' type. I converted them to 'Float', so that they can be used while mapping the data and exploring for venues.

2.3 Feature selection

I have chosen to limit the scope of this project to national parks only. Therefore, any other features, such as national monuments, Historic sites, seashores etc. had to be filtered out from the dataframe.

After data cleaning and feature selection, I had data for 119 national parks that I could use for analysis.

See below for a screenshot of what the dataframe looks like:

	index	State	Region	Type	Parkname	Latitude	Longitude
0	5	VA	NE	National Historical Park	Appomattox Court House	37.380221	-78.798570
1	6	NY	NE	National Historical Park	Women's Rights	42.908170	-76.816558
2	8	CA	PW	National Park	Pinnacles	36.490396	-121.181168
3	9	DE	NE	National Historical Park	First State	39.830758	-75.563539
4	16	OH	MW	National Park	Cuyahoga Valley	41.259102	-81.570988
...	...	...	...	...	...	...	...
114	410	PA	NE	National Military Park	Gettysburg	39.815352	-77.232477
115	412	WV	NC	National Historical Park	Harpers Ferry	39.318705	-77.741902
116	414	OR	PW	National Park	Crater Lake	42.941064	-122.132749
117	416	TX	IM	National Park	Big Bend	29.297380	-103.229435
118	419	AK	AK	National Historical Park	Klondike Gold Rush	59.610202	-135.261768

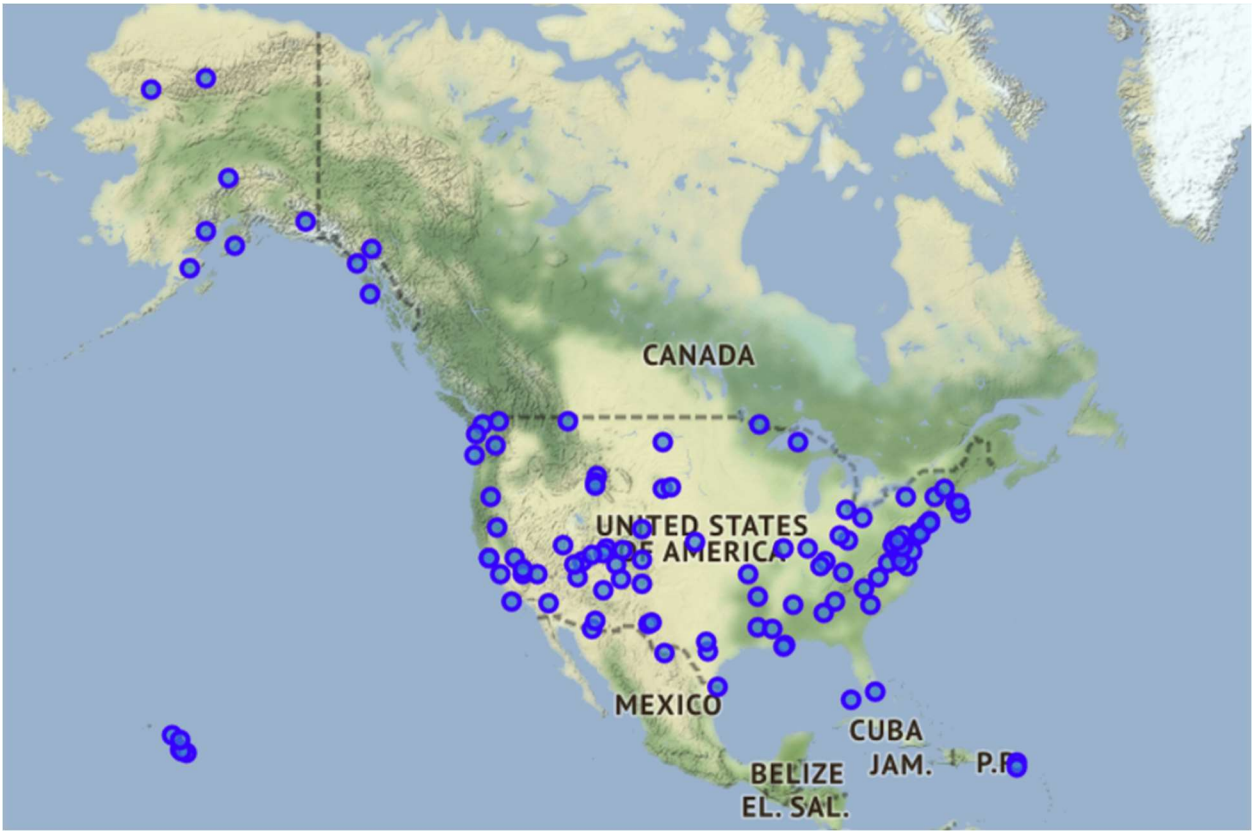
119 rows × 7 columns

3. Exploratory data analysis

3.1 Folium map

I begin the analysis by first mapping the park locations on the United States map using Folium. I used Stamen Terrain tileset to display national parks on the map.

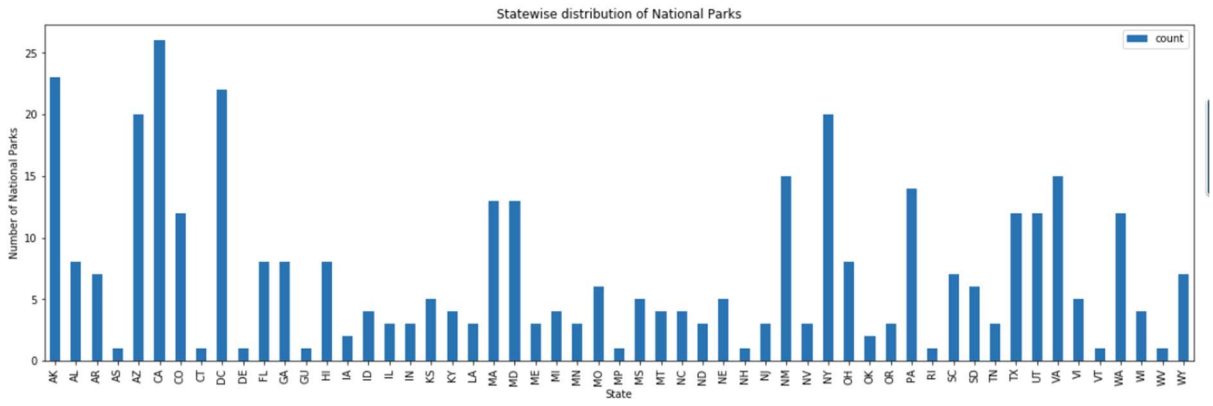
The map looks as follows:



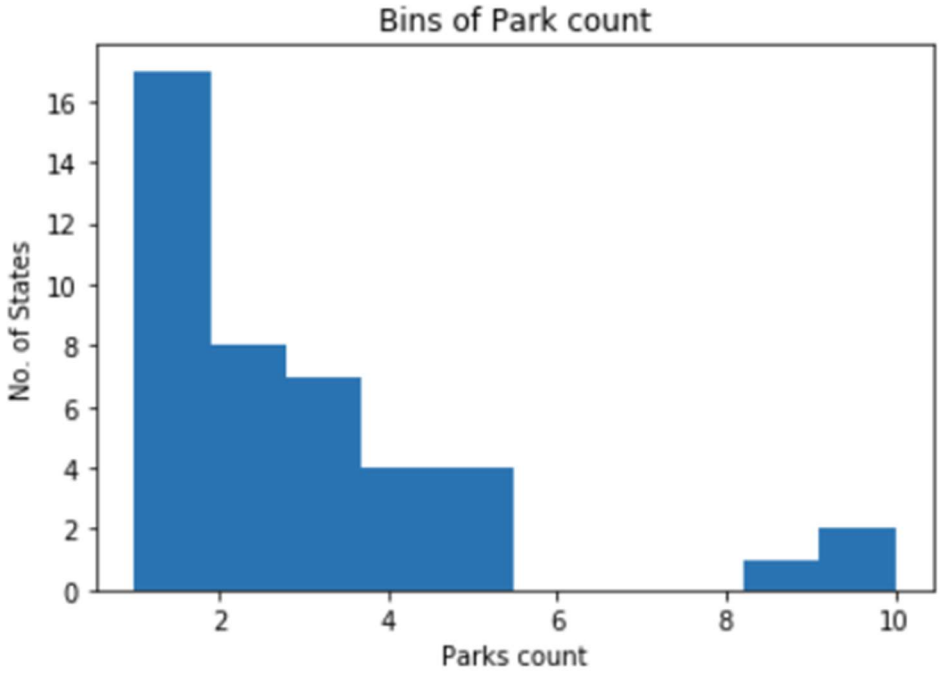
Visualizing national parks locations on the map this way clearly helps to see how some of the parks can be clustered together and may offer similar venues.

3.2 Data pre-processing

Next, I group the parks according to 'states' and identify distribution of number of parks in each state. Bar graph helps to show which states have more parks that the others.



Next, I sort them in the descending order of number of parks, and used binning to visualize.



From the bin histogram, it shows that there are many states having <4 parks, about 8 states with 4/5 parks and 3 states having more than 8 national parks.

4. Foursqaure API

4.1 Setting up URL for GET request

I employ Foursqaure APIs to explore the parks and its surroundings. First I set my Foursqaure credentials. Then I setup a URL to explore in the 500 meters radius of each of the park area using its latitude and longitude data and select first 100 results. The URL I use is as follows:

```
# create the API request URL
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    lat,
    lng,
    radius,
    LIMIT)
```

Where the parameters are -
CLIENT_ID = My Foursquare Client ID,
CLIENT_SECRET = My Foursquare Client Secret,
lat = the latitude of each park,
lng = the longitude of each park,
radius = 500,
LIMIT = 100.

After running this URL for all the parks, I could gather ~3500 venue results, and ~330 unique venue categories.

4.2 Exploring venues for each park

The next task is to analyze each park. For this I use one-hot encoding to encode categorical data using *get_dummies* function.

Then I grouped the results by park by taking the mean of the frequency of occurrence of each category to find the most common venue categories for each park.

Once the venues categories were grouped, then I print each park along with the top 5 most common venues and put them into *pandas* dataframe. The dataframe now looked like the following:

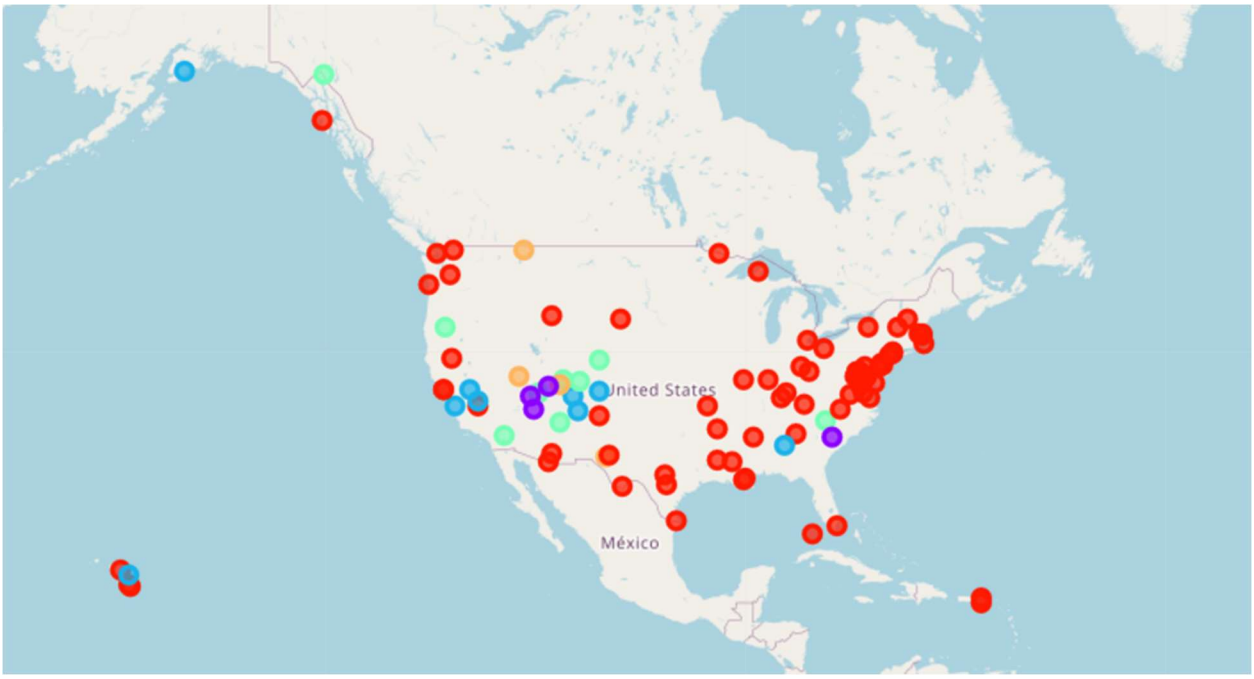
	Parkname	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abraham Lincoln Birthplace	Farm	Toy / Game Store	Construction & Landscaping	Historic Site	Food	Farmers Market	Fast Food Restaurant	Filipino Restaurant	Fish Market	Flea Market
1	Adams	American Restaurant	Chinese Restaurant	Seafood Restaurant	Pizza Place	Breakfast Spot	Coffee Shop	Italian Restaurant	Grocery Store	Cosmetics Shop	Ice Cream Shop
2	American Memorial Park	Hotel	Japanese Restaurant	Beach	Seafood Restaurant	Fast Food Restaurant	Bar	Pizza Place	Coffee Shop	Tea Room	Noodle House
3	Appomattox Court House	Fast Food Restaurant	Discount Store	History Museum	Park	American Restaurant	Italian Restaurant	Mexican Restaurant	Grocery Store	Coffee Shop	Big Box Store
4	Arches	Scenic Lookout	Trail	Park	National Park	Zoo Exhibit	Flower Shop	Farmers Market	Fast Food Restaurant	Filipino Restaurant	Fish Market

5. Clustering

5.1 K-means clustering

K-means clustering is useful for segmentation purposes. In this project, I chose K-means clustering since it can help me quickly discover insights from unlabeled data.

I use K-means algorithm from sci-kit learn package (sklearn.cluster.KMeans) to cluster these parks into 5 clusters and drop any missing values.



Virtually all the national parks appear to fall in the same cluster, with exception of the few. We can examine the clusters more closely to find out what distinguishes them.

5.2 Examining clusters

Once the clusters are created and mapped, I can examine each cluster and determine the discriminating venue categories that distinguish each cluster. The clusters look as follows:

Cluster number	No. of parks	Unique feature(s) of the cluster
1	76	Parks in this cluster are either of historical significance or military related
2	4	These parks offer wide range of trails and outdoor adventures
3	9	These parks mostly offer park adventures along with zoo exhibits and restaurants
4	9	Parks in this cluster offer scenic lookouts and trails
5	4	These parks have mountainous regions along with farms and restaurants

6. Recommendation Engine

Clustering gives a good overview of which parks offer which type of venues and which parks are similar. To further enhance the user experience, and to offer the user recommendations for the park, I have built a recommendation engine using content based filtering algorithm.

The recommendation algorithm works in 3 stages:

6.1 Finding unique venues

First, I find unique venues from the list of 1st to 5th common venues from section 4. These become the most common available venues for the recommendation engine.

6.2 Getting user preferences

Next, I get the preference data from the user in the order from 1st to 5th.

6.3 Dot product

Then I take the dot product of the user's preferences and available venues for a complete list of recommended park locations.

6.4 Exploring the recommended park

Once the list of recommendations is received, we can search for specific venue using Foursquare API. I can get the tips and ratings for the venue and map it using Folium.

7. Discussion

With this project, I attempt to analyze National Parks in the United States with the help of machine learning algorithms. I hope this work will find its uses among potential park visitors.

8. Conclusion

This project only explores national parks in the United States. The study can be further extended to all the parks in the United States and other countries alike.

9. References

- National Park Services: https://www.nps.gov/index.htm
- Opendata soft: https://public.opendatasoft.com/explore/?sort=modified
- Pandas user guide: https://pandas.pydata.org/docs/user_guide/index.html
- Matplotlib user guide: https://matplotlib.org/3.2.1/users/index.html
- Scikit-learn: https://scikit-learn.org/stable/