National Park Guide for Travelers

0. Team Members

Ziyuan Cai (cziyuan@upenn.edu)
Sisun Cheng (sisunc@upenn.edu)

1. Introduction

National parks are defined as a reserve of natural, semi-natural, or developed land that a sovereign state declares or owns, with the conservation of wild nature for posterity and as a symbol of national pride.

Since 1872 that the US government established the first "public park or pleasuring-ground for the benefit and enjoyment of the people", Yellowstone National Park, there has been 423 national park sites in the National Park System (NPS), while about 60 of those are what we often called national parks. On December 20, 2019 White Sands was re-designated White Sands National Monument as White Sands National Park, making it the 62nd designated national park in the National Park System.

For the sake of data accessibility, in this project, we only focus on the first 61 national parks. Each of these 61 national parks has its own characteristics, but they share the common that they attract a huge number of visitors each year. Despite the impact of COVID-19 on tourism, people are still fascinated by the beautiful natural landscapes in national parks. Thus, we aim to carry out a tourist's guide for national parks in the US for your reference.

2. Data

2.1 Data from Local Files

Table 1 Data from Local Files

Data File	Format	Data Source	Description
cb_2018_us_county_500k.shp	shapefile	https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html	US county boundaries from Census Bureau
cb_2018_us_state_500k.shp	shapefile	https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html	US state boundaries from Census Bureau
nps_boundary.shp	shapefile	https://irma.nps.gov/DataStore/Reference/Profile/2224545?lnv=True	National park boundary data
airport_code.csv	csv	http://www.airportcodes.org/	Airport code data
Crime_counties.csv	csv	https://jacobdkaplan.com/crime.html	Actual crime case numbers in each county where national parks locate

Normalized_10_year_visitors.csv	csv	https://public-nps.opendata.arcgis.com/	Visitor data of 2007-2016, manually normalized
southwest_price.csv	csv	https://www.southwest.com/	Air ticket price for Southwest Airlines, manually scraped
united_price.csv	csv	https://www.united.com/	Air ticket price for United Airlines, manually scraped
tickets.csv	csv	Manually searched and downloaded.	Whether or not a national park charges tickets
Top10_in_counties_final.csv	csv	Generated from 1_Ranking_National_Parks.ipynb and manually added the state, county, and codes for each park	Top 10 national parks with state, county, and FIPS codes

2.2 Data Directly from Codes

2.2.1 Twitter API

Twitter is a popular social media across the nation, and we can get use of the twitter contents to analyze people's feeling on national parks with the method of sentiment analysis. We got Twitter text with topics of each National Park's name from Twitter API. We used the names of every national park as the search word to get the twitter of maximum 1000.

2.2.2 Census data

To provide more information of the counties where the national parks locate, we use 'cenpy' package and census API to get median household income and population data with the spatial unit of census block groups.

2.2.3 OSM data

When visiting a new place, travelers often have problem finding parking spaces and restaurant. To get the local amenity data, we use 'osmnx' package to exploit OSM data. We select the amenities of "restaurant", "fast_food", "parking", and "parking_space" and divide them into groups of parkings and restaurants. By doing this, we hope to give the travelers a general sense of local amenities.

2.2.4 Weather data

For a better travel experience, it's important for travelers to take a quick look at the local weather data, especially the temperature and precipitation. To extract local weather data, we apply R scripts here. R package 'riem' allows us to get weather data from Automated Surface Observing System (ASOS) stations in the whole world thanks to the Iowa Environment Mesonet website. We prepare the weather data in R and import

it to jupyter notebook from .csv.

3. Methods

3.1 Sentiment Analysis with Twitter data

The goal of a sentiment analysis is to determine the attitude or emotional state of the person who sent a particular tweet. The outcome of polarity determines the attitude of the states and subjectivity determines the emotional part. In our work, we only need to focus on whether the national park is good or not through public, which is polarity, and we don't need the subjectivity part to see whether their remarks are subjective or objective. The ‘textblob’ package will find the words like "good" or "bad" to test the polarity of each tweet and will give out a number from -1 to 1 for quantification.

Combined the polarity results and the normalized visitors from 2007-2016 in 50 and 50, we figured out the score of each national parks, which decides the ranking result.

3.2 K-means Clustering

Cluster analysis is a method that partitioning dataset into several groups of similarities when class label information stays unknown. In cluster analysis, association or similarity within the same class should be strong while it should be rather weak between groups.

K-means is one type of clustering algorithms. In K-means algorithm, the number of classes (k) is specified in advance, and the ‘sklearn’ package can help divide the data into k groups according to the input features. However, K-means has the disadvantage that it’s only suitable for numeric, continuous data in theory, although some people may apply it to binary and categorical data.

3.3 Hierarchical Clustering

Hierarchical clustering is a method by building a hierarchy of clusters. In hierarchical clustering, we do not need to specify the number of clusters before applying the algorithm, and the number of clusters can be determined by researchers as a result of hierarchical clustering. Hierarchical clustering is suitable for small data sets rather than large ones and can handle binary variables. Thus, in our final project, we apply hierarchical clustering in our clustering analysis Python also provides package ‘scipy.cluster.hierarchy’ to do hierarchical clustering..

4.Results

4.1 Sentiment Analysis and Ranking

Only a few of national parks don't contain enough tweets (1,000 tweets), which can be explained as they contain little discussion through public, and as a result, they are not popular at all. The top 10 national parks are listed in order as follows: Great Smoky Mountains National Park, Saguaro National Park, Petrified Forest National Park, Crater Lake National Park, Redwood National Park, Katmai National Park, Mesa Verde National Park, Grand Canyon National Park, Everglades National Park, Channel Islands National Park. Because twitter analysis is instant, so polarity analysis has certain volatility. In addition, we want to provide travel guidance during the Christmas period, so we chose to conduct twitter analysis in December. From the result, some really famous national parks are not in the list of top 10, maybe because they are not suitable to visit in winter, which can be indicated from the tweets post by the public.

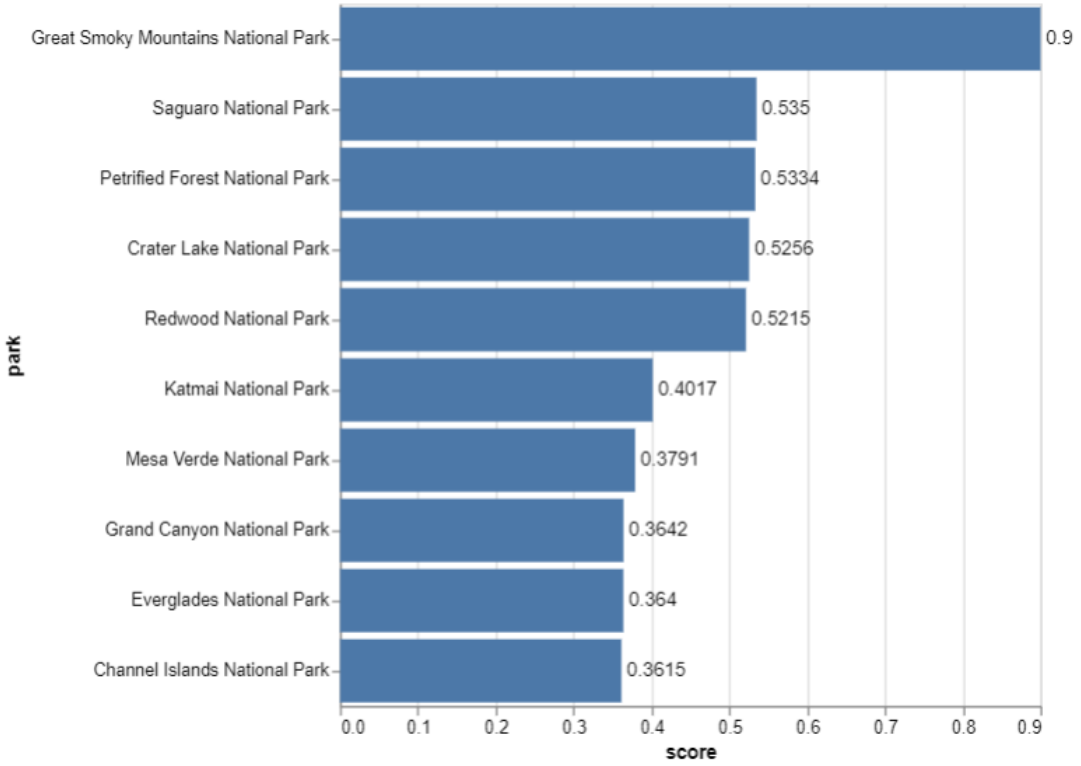

Figure 1 The ranking of the top 10 parks

4.2 Information and Surroundings of the Parks

There are national parks all over the United States, so their related attributes will be very different. Katmai National Park has the largest crime numbers nearby, Saguaro National Park has the greatest number of parkings around and Everglades National Park

has the most surrounding restaurants. There is no significant difference in median house income attributes, because the area closest to the national parks must contain a small population, and only the urban areas in the nearby county will have a relatively high income.

Table 2 Information and Surrounding Features of Each Park

Name	Median household income	Parking #	Restaurant #	Tickets	Max temperature	Min temperature	Precipitation	Wind speed	Crimes per 1000 population
Channel Islands National Park	87130.6	2468	938	No	67.40308	51.94462	0.454912	13.44615	3.951806
Crater Lake National Park	45011.75	238	55	Yes	47.46154	38.23846	1.02004	15.24615	13.71767
Everglades National Park	61926.72	6661	1841	Yes	79.58154	64.70769	0.82052	13.18462	21.05196
Grand Canyon National Park	51069.02	1104	341	Yes	67.8	45.99538	0.128252	11.53846	9.702908
Petrified Forest National Park	36470.41	458	125	No	67.8	45.99538	0.128252	11.53846	5.107604
Great Smoky Mountains National Park	47624.34	1035	337	No	53.00615	36.52	1.399892	13.84615	23.28171
Katmai National Park	66896.33	525	112	No	1.049123	-5.68947	0.006386	5.912281	29.83075
Mesa Verde National Park	49601.23	114	31	Yes	39.58154	17.74923	0.07404	12.73846	9.258192
Redwood National Park	47756.43	802	161	No	58.19231	47.09077	0.527146	14.69231	11.87141
Saguaro National Park	53585.05	13962	876	Yes	68.06769	40.52462	0.090218	14.41538	11.7102

4.3 Cluster Result

K-means and hierarchical analysis return the same clustering result and all 10 national parks are divided into 3 small groups. The groups are listed as follows,

- Group 0: Everglades National Park, Katmai National Park, Saguaro National Park
- Group 1: Crater Lake National Park, Grand Canyon National Park, Petrified Forest National Park, Great Smoky Mountains National Park, Mesa Verde National Park, Redwood National Park
- Group 2: Channel Islands National Park

Parks in group 0 is common in high crime rate, many local restaurants, and moderate median household income. Group 1 has the similarities of low median household

income, few local restaurants and other amenities, and moderate crime rate. Group 2 shows a high local income and low crime rate.

With the clustering result above, we hope to give our users a general sense of which national parks are similar to each other and the common characteristics shared in each group, so that we can help them decide their destination this Christmas holiday.

4.4 Analysis on Air Ticket Price

To help travelers better plan their trips to national parks during the Christmas holiday, we collect the air ticket price data from Southwest Airlines and United Airlines. We select PHL as the origin and the nearest commercial airport of each national park as the destination. The ticket price data was collected on Dec 6th, 2021 from their official website.

As is plotted below, there's a rising trend of the air ticket price as the time comes near Christmas. Prices on some routes reach a peak from the weekend of 12.18/19, while others keep rising until Christmas Day. Generally speaking, United Airlines has a cheaper ticket price than Southwest Airlines, and some routes and time of Southwest Airlines are missing because it does not perform flights. The changes in ticket prices for both airlines can be seen more visually from the interactive maps and plots in our webpage.

Figure 2 Air Ticket Price of Southwest Airlines

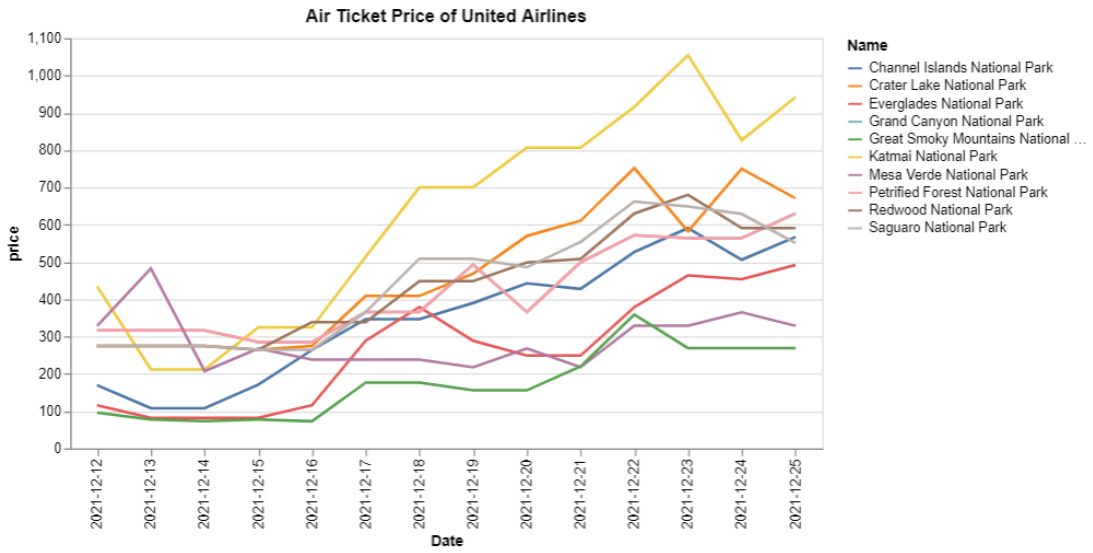

Figure 3 Air Ticket Price of United Airlines

5. Conclusions

Based on the analysis above, we can finally make reasonable recommendations to the residents of Philadelphia who plan for a national-park-trip during the Christmas holiday. If you have high requirements for the quality of travel, like those who must stay in a five-star hotel when traveling, Channel Islands National Park is your best choice. The reason why it is divided into a separate category is the median house income around this park is very high, and the crime rate is the lowest. Therefore, you can definitely meet your various needs locally and experience quite a good service.

On the contrary, if you have no requirements for quality, and the only thing you care about is your wallet and want an economical trip, then Great Smoky Mountains National Park and Redwood National Park will be the best choice. The surrounding median house income of these two parks is at a lowest level of 10 national parks, and neither of them require tickets, so you can definitely spend your Christmas trip with the least cost. However, given the distance from Philadelphia, Great Smoky Mountains National Park has the cheapest airfare, and this one would be the first choice. If you want to go to Redwood National Park, remember to buy a ticket in advance and try to avoid the peak of Christmas.

If you are a person who likes warm weather, then Everglades National Park is the place you need to visit during Christmas holidays. In contrast, if you like to challenge extreme weather, then you will definitely want to go to Katmai National Park. A large part of the reason of both of the two parks are divided into category 0 is that the climates of the two parks are very different from those of other parks.

Although some parks are not mentioned in our recommendation, these are also unique. Considered that they are all top 10 selected by the public, you can freely choose any of the park you want to go to for Christmas holiday relaxation.