Analysis of National Park Trails (as listed on AllTrails)

By Alyssa Diaz

September 2022

GoogleColab Github

Introduction

In recent years, hiking has exploded in popularity: on the widely used hiking website/app AllTrails "the number of hikes logged in 2020 was up 171.36% compared to 2019."[1] With increased trail usage come the need for accurate information as provided by various hiking apps, with AllTrails being the most-used worldwide with over 40 million users.[2] I chose a dataset compiling trails in National Parks as taken from AllTrails: I personally have experience exploring and using both and have found the accuracy of certain metrics (especially difficulty rating) does tend to vary.

In this analysis, I will examine what influences a hike's difficulty rating, and recommend how to make the difficulty ratings more transparent and understandable. By having a more finely honed rating system, everyone involved will benefit: if hikers have a better understanding of what they will face difficulty-wise on a hike, enjoyment and safety will increase; if AllTrails has a reliable and objective rating system in place, app user satisfaction will increase; for National Parks (or any parks and hiking areas) hikers will be safer if they have complete information at their disposal in deciding what hike to embark on.

Data and Methodology

The dataset "*National Park Trails:* Every trail in the National Parks Service gathered from alltrails.com"[3] was compiled by user "planejane" by scraping from AllTrails' web page by state.[4] Each hike listed contains trail statistics such as length, elevation gain, and difficulty rating, as well as user ratings and number of reviews posted on the AllTrails website. Another analysis of this dataset (posted on Kaggle) provided a wide overview of distribution of various metrics, with more of a focus on distribution of hikes in each state and park.[5]

Many of the variable names needed clarification; I was able to match most of the variables in the dataset to the corresponding metrics as listed on the AllTrails website. I also determined the units of length and elevation by comparing to the trail lengths for the same trail listing on AllTrails, and set up new columns for length and elevation in miles and feet (standard units for hiking in the US). Finally, I added a column for 'Steepness (% Grade)' in order to explore a possible correlation between hike steepness and difficulty rating.

After fixing the obvious errors and omissions in the data (such as incorrect state/country names and a number of null values), I checked the hikes to ensure that they are actually hikes in National Parks. By focusing on outliers, I identified and removed the erroneously included hikes that are 'scenic drives' and 'overlooks,' and those that are in areas outside of National Parks.

After cleaning, the dataset contains 3,271 trails in 55 National Parks and 30 states, with a total of 34,904 miles of trails and 6,850,686 feet (1,297 miles) of elevation gain.

[1] Hiking in the US has Never Been More Popular[Study] Paul Ronto, 2021

[2] AllTrails.com

[3] https://www.kaggle.com/datasets/planejane/national-park-trails

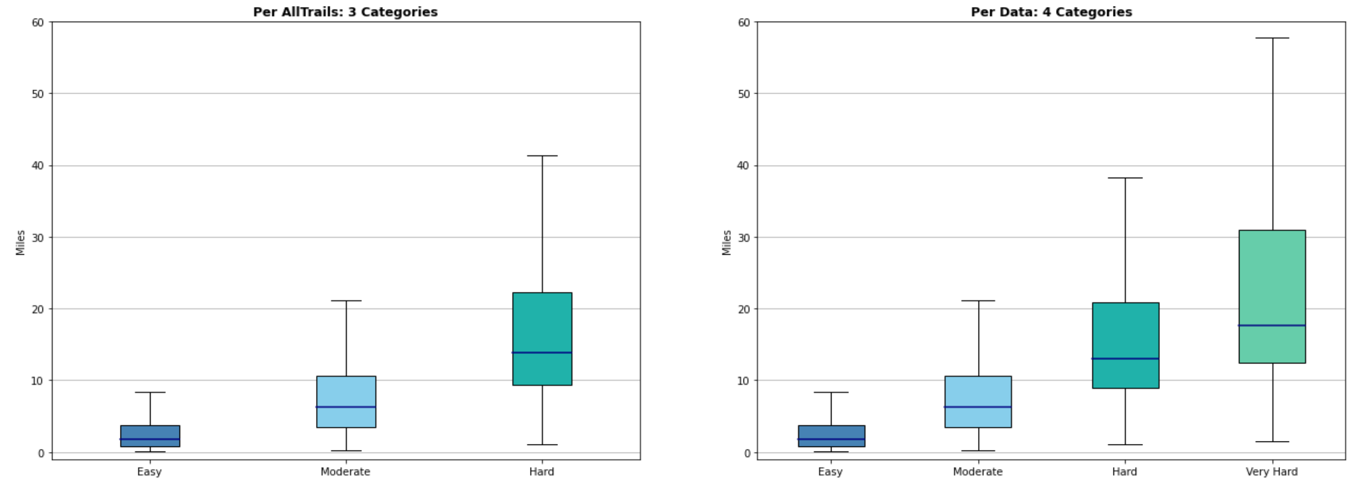
[4] https://github.com/j-ane/trail-data

[5] AllTrails in NP Analysis Jessica Qi Jiang

Analysis

While defining the difficulty rating system I found that the dataset contains 4 difficulty levels (1, 3, 5, and 7), while the AllTrails app utilizes only 3 difficulty ratings (Easy, Moderate and Hard). Research revealed that the numerical ratings in the dataset correspond to the AllTrails ratings where Easy=1, Moderate=3, and Hard=5&7. For the purposes of this analysis I defined a new rating of Very Hard=7 to explore if any significant differences exist between ratings 5&7 in the dataset.

Lengths of Hikes in Each Difficulty Category

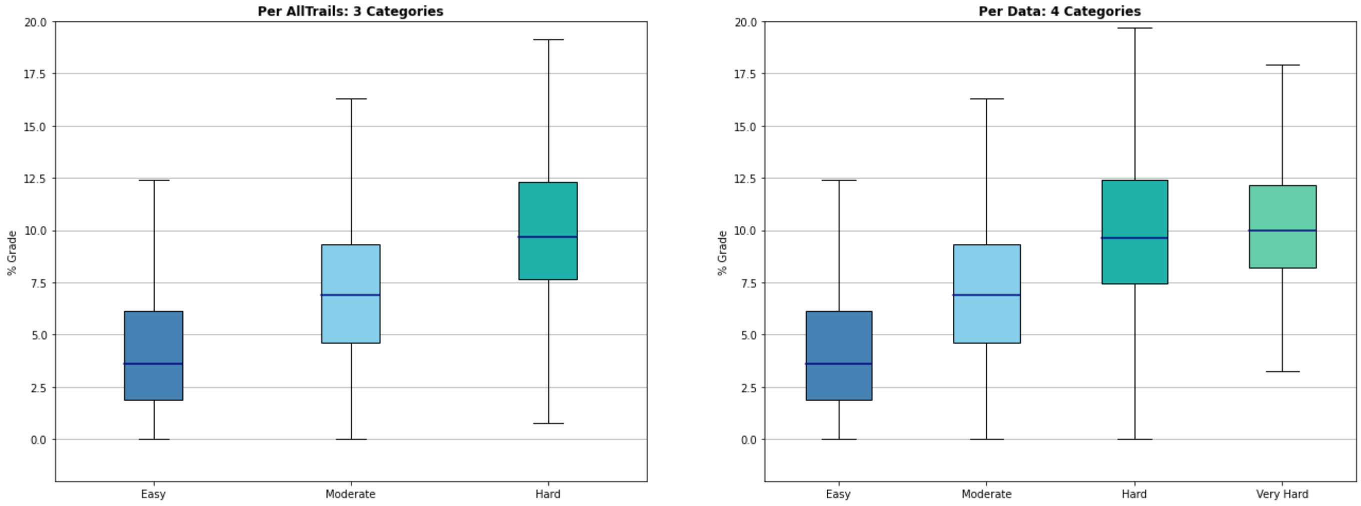


Per AllTrails: 3 Categories

Per Data: 4 Categories

For lengths of hikes categorized as Moderate and Hard by AllTrails, the distributions and means jump significantly between the two, while the distributions and means of the full 4 categories are more evenly spaced. The distribution for elevation gain between difficulty categories (not shown) was similar.

Steepness of Hikes in Each Difficulty Category

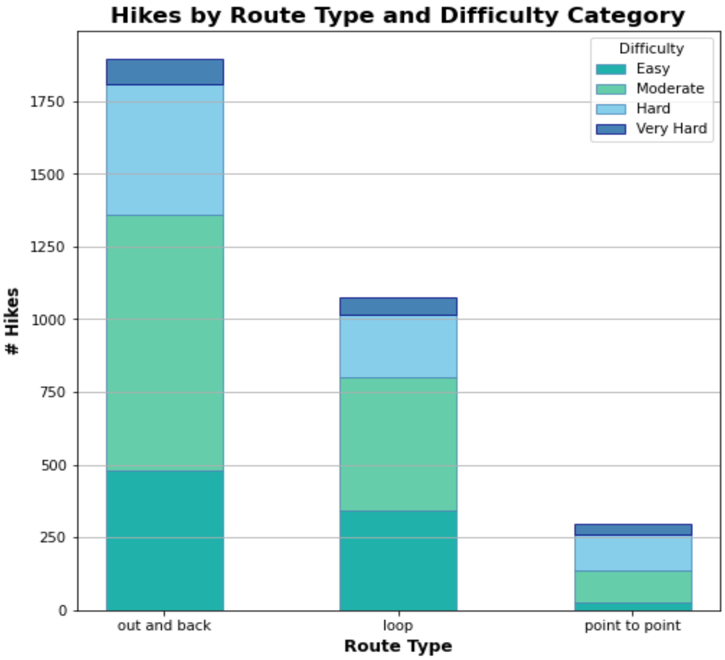


Per AllTrails: 3 Categories

Per Data: 4 Categories

In contrast, for steepness of hikes in each difficulty rating as categorized by AllTrails, the distributions and means are evenly spaced. When using the full 4 categories, the distribution of steepness in the Very Hard(7) category falls entirely within the range for Hard(5), and the means of Hard(5) and Very Hard(7) are very close to each other.

I also investigated possible correlations between difficulty rating and other metrics such as user ratings and hike locations (park and state). User ratings show no correlation to difficulty rating, and the proportion of hikes in each difficulty rating vary widely between parks. Most notably, the graph of the number of hikes in each route type shows that a much higher proportion of Hard and Very Hard hikes are point to point, with most Easy hikes being out and back or loop hikes.



Conclusion

Most significantly, difficulty rating as defined by AllTrails seems to be most closely tied to steepness: individually, length and elevation vary widely between the Hard(5) and Hard(7) categories, while the range of steepness in the two Hard categories is very similar. More analysis would be necessary to reveal the full extent of this correlation, and also to explore the correlation between ‘point to point’ hikes and the High difficulty rating.

Further analysis would also be needed to recommend a new rating system or ways to optimize the current system; a possible iteration could be to simply break the difficulty categories down into more than the 3 defined categories as is currently utilized by AllTrails. Alternatively, a secondary attribute could be added to each rating to indicate why the hike is rated at the determined level of difficulty (e.g. steepness or length) to better communicate what the user should expect on the trail and the level of exertion.

Clarified:

- National parks only or all park service areas?
  - Mostly national parks, includes a few non-park areas (errors), also possibly does not include all National Parks (currently 63 parks total, df contains only 60 including erroneous parks)
- Elevation, length in what units? (column 'units'=i for imperial, m for metric?)
  - Confirmed metric (meters) after comparing to listings on AllTrails
- Source and definition of visitor usage and popularity
  - Confirmed visitor usage: corresponds to "trail traffic" on AllTrails (Not listed in stats of app hike entry, but verbally in the hike description)
  - Contacted AllTrails and unclear how popularity calculated (won't reveal algorithm)

Cleaned:

- Fixed Hawaii/United States swapped as country/state
  - Fixed using .loc
- Changed Congaree Wilderness to Congaree National Park
  - Used to_replace
- Removed hikes in parks listed that are not actually National Parks
  - Manually checked against current National Park list
  - Used df=df['column'].str.contains('unwanted data'==False)
- Dropped columns:
  - units, city, country, trail_id (unnecessary)
  - popularity (unclear)
  - geoloc, activities, features (not using in this analysis)
- Converted and added as new columns: elevation gain to feet and length to miles
- Dealt w/ visitor usage NaN entries
  - Checked against app, no obvious pattern as to which hikes have NaN
  - Used df.fillna to replace NaN with 0
- Set up new 'Steepness (Slope) and 'Steepness (% Grade)' column, rounded to 2 dec places
  - Modified Steepness % Grade formula to show accurate results
  - Used np.arctan to make correct formula for angle of elevation
- Removed 13 scenic drives included erroneously as hikes
  - Looped thru df and checked against 'scenic drive' and 'drive', saved to 2 lists (drives, possible drives)
  - Manually checked list of possible drives and appended to drives list
  - Used df=df['column'].str.contains('|'.join(unwanted data list)'==False)
- Defined basic numerical difficulty ratings scale
  - Easy=1, Moderate=3, Hard=5 & 7
- Noted only 55 national parks listed: not missing data, but sign that some National Parks have no trails (eg most in Alaska are remote wilderness, Biscayne is marine park)

Explore possible correlations between:

- Steepness and difficulty rating
- Elevation and difficulty rating
- Length and difficulty rating

- User rating and difficulty rating