

A hiker with a backpack is walking away on a dirt trail. The trail is surrounded by green bushes and a large field of purple wildflowers. In the background, there is a dense forest of evergreen trees and distant mountains under a blue sky with wispy clouds.

# Turransky Regression Project

An assessment on public National Park trail data



# Introduction

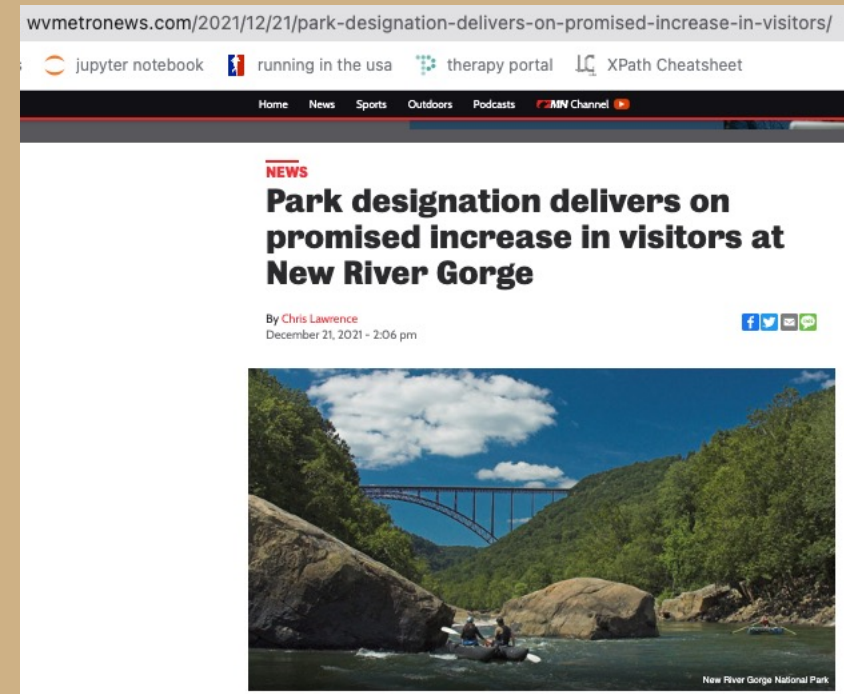


- There are 63 sites declared by Congress as a US National Parks
- National Park designation helps provide adequate protection of the area's resources.
- As new land areas are designated as National Parks, they see a surge in attendance.
- More hikers means more maintenance, security etc. are needed.
- Knowing what draws hikers to a trail will help new areas plan accordingly

# Hiker attendance is a big issue for old and new national parks!



Arches, Zion and other national parks saw such an influx in visitors during the pandemic they had to implement restrictive measures like timed entry tickets or a lottery to do certain hikes.



*From the above website:..."Officials promised there would be at least a 20 percent increase in traffic and visitation in the year the park designation is added. ...visitor numbers have jumped well over 20 percent for 2021."*

# Methodology

- Used Selenium to dynamically navigate HikingProject.com to get individual national park websites and websites for each trail at each national park
- Used beautiful soup to parse HTML code for desired data.

The screenshot displays the HikingProject.com website interface. At the top, the browser address bar shows the URL `hikingproject.com/search?q=national%20park`. Below the address bar, there are navigation links for 'metis', 'jupyter notebook', 'running in the usa', 'therapy portal', and 'XPath Cheatsheet'. The main header features the 'HIKING PROJECT' logo, navigation links for 'Trail Guide', 'Best Photos', and 'Top Hikes', and a search bar with the text 'Find trails, cities, etc' and a 'Sign In' button. The search results section shows a list of national parks with their respective trail counts. On the left side of the search results, there is a sidebar with filters for 'All (19,599)', 'Hikes, Trails and Gems (6,434)', 'Areas (425)', 'Users (36)', 'Photos (11,378)', 'More (1,314)', 'Clubs (9)', and 'Cities (3)'. The main content area displays the search results for 'national park'.

Areas	Sort by: Default
<b>Yellowstone National Park</b> WY > Northwestern Wyoming	267 Trails
<b>Indiana Dunes National Park</b> IN > Northern Indiana	43 Trails
<b>Redwood National Park</b> CA > North Coast	105 Trails
<b>Sequoia National Park</b> CA > High Sierra	123 Trails



Sample trail  
page and some  
details scraped:

hikingproject.com/trail/7024343/rimrock-to-uplands-loop

metis jupyter notebook running in the usa therapy portal XPath Cheatsheet

HIKING PROJECT

Trail Guide Best Photos Top Hikes

Find trails, cities, etc

Sign In

Rimrock to Uplands Loop

RECOMMENDED ROUTE

INTERMEDIATE

★★★★★ 4.0 (12)

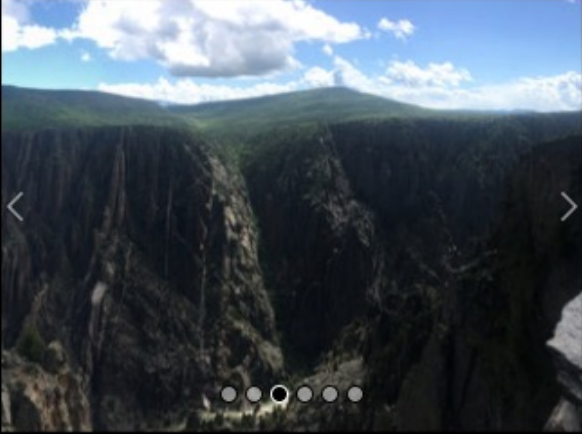
Areas

CO

Central Rockies

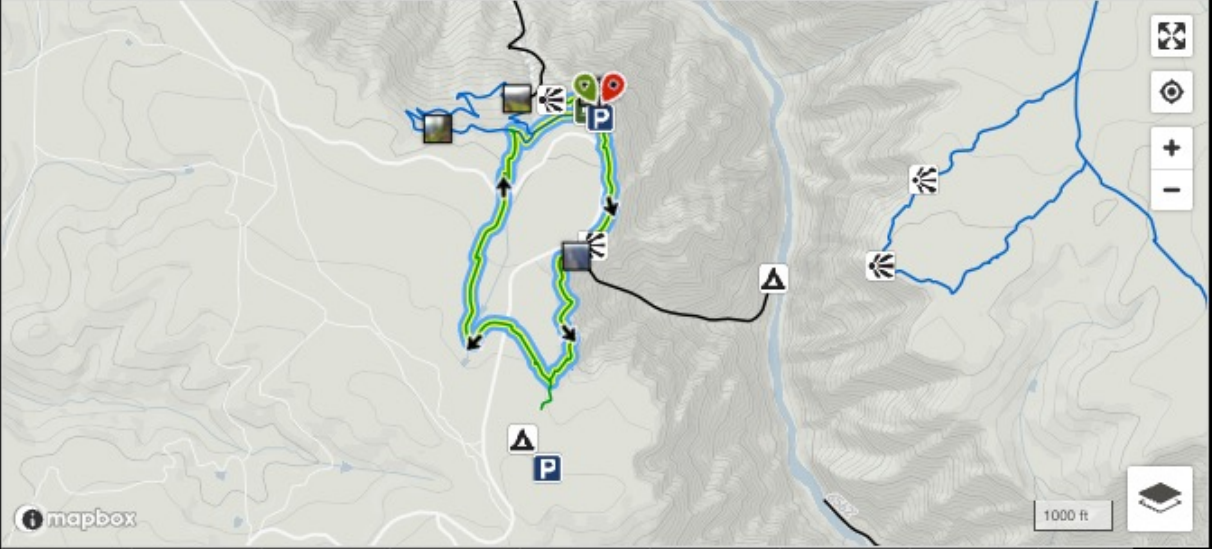
Black Canyon of the Gunnison National Park

South Rim



"An easy and accessible route leading to some of the most scenic portions of the park."

Mikhaila Redovian



mapbox

1000 ft

Your Rating: ☆☆☆☆☆

Your Difficulty: ●●●●●

Your Favorites: Add to Favorites · Your List

2.2 Miles

8,335' High

404' Up

7% Avg Grade (4°)

Loop

7,975' Low

406' Down

53% Max Grade (28°)

# All Scrapped Details

- Features:

- |                                         |                                          |
|-----------------------------------------|------------------------------------------|
| 1. Park Name                            | 9. Distance hiker goes uphill            |
| 2. Trail Website                        | 10. Distance hiker goes downhill         |
| 3. Trail difficulty                     | 11. Trail type (loop, out and back etc.) |
| 4. Average Rating of the trail          | 12. Average grade (%)                    |
| 5. Number of people who rated the trail | 13. Max grade(%)                         |
| 6. Distance in miles                    | 14. Checkins                             |
| 7. High elevation                       | 15. State                                |
| 8. Low elevation                        | 16. Dog policy                           |

```
: np_trail_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 2938 entries, 0 to 2937
```

```
Data columns (total 18 columns):
```

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	2938 non-null	int64
1	index	2938 non-null	object
2	park_name	2938 non-null	object
3	website	2938 non-null	object
4	difficulty	2938 non-null	object
5	avg_rating	2938 non-null	float64
6	num_raters	2938 non-null	int64
7	distance_(miles)	2938 non-null	float64
8	high_(ft)	2938 non-null	int64
9	low_(ft)	2938 non-null	int64
10	up(ft)	2938 non-null	int64
11	down(ft)	2938 non-null	int64
12	trail_type	2938 non-null	object
13	average_grade(%)	2938 non-null	int64
14	max_grade(%)	2938 non-null	int64
15	checkins	2938 non-null	int64
16	State	2938 non-null	object
17	dog_policy	2881 non-null	object

# Regression Methodology Highlights

---

Performed Train/Validation/Test separation on model

---

Utilized dummy variables to assess categorical non-numeric values.

---

Grouped parks by regions

---

Measured improvements on R-Squared using stat model and Sklearn

---

Target was number of people who rated each trail. The more people who hike a trail, the more ratings a trail will have. Since the website did not have a 'completed' option this next best correlated well with our target.

# Improvements in R\_squared for Training and Validation data

## How it started

```
fit_train=train_model_t.fit()  
fit_train.summary()
```

3]:

OLS Regression Results

Dep. Variable:	num_raters	R-squared:	0.081
Model:	OLS	Adj. R-squared:	0.074
Method:	Least Squares	F-statistic:	11.55
Date:	Tue, 17 May 2022	Prob (F-statistic):	8.21e-16
Time:	15:41:31	Log-Likelihood:	-3950.5
No. Observations:	1060	AIC:	7919.
Df Residuals:	1051	BIC:	7964.
Df Model:	8		
Covariance Type:	nonrobust		

Initial Train R\_squared: 0.081

Initial Validate R\_squared: 0.070

## How it ended

```
2]: newX=X_train.drop(columns=['Trail_name', 'checkins', 'park_name', 'website', 'trail_type',  
test_model=sm_model_stats(newX,y_train)
```

OLS Regression Results

Dep. Variable:	num_raters	R-squared:	0.114
Model:	OLS	Adj. R-squared:	0.099
Method:	Least Squares	F-statistic:	7.458
Date:	Tue, 17 May 2022	Prob (F-statistic):	1.79e-18
Time:	15:42:05	Log-Likelihood:	-3930.9
No. Observations:	1060	AIC:	7900.
Df Residuals:	1041	BIC:	7994.
Df Model:	18		
Covariance Type:	nonrobust		

Final Train R\_squared: 0.114 (+0.033)

Initial Validate R\_squared: 0.097 (+0.027)



# Train + validate model vs test data results

- Train and Validation data combined resulted in a  $R^2$  score of 0.119.
- Running the above model on test data resulted in a  $R^2$  score of 0.124.
- Features that had the largest impact on number of raters:
  - Average rating
  - Distance
  - Location (Parks in Colorado Plateau region (including Zion, Grand Canyon, Arches etc.) did especially well



# Future Model Evaluation

- In December 2020, congress designated 72,808 acres of land in WV as New River Gorge National Park.
- Because of how recent this happened, Hikingproject.com did not have that park on the list of National Parks.
- This data could be scraped at a later time and compared against the trail model for further testing/validation of the model.