PROJECT PAPER
CAPP 30122 – Computer Science and Applications II

Team name

Project Parks Pulse (PPP)

Members (name+CNetIDs)

Minh Nghiem - mnghiem@uchicago.edu
Seongyeon Yang - seongyeon@uchicago.edu
Yi-Huai Chang - yhchang@uchicago.edu
Diyanet Nijiati - diyanet@uchicago.edu

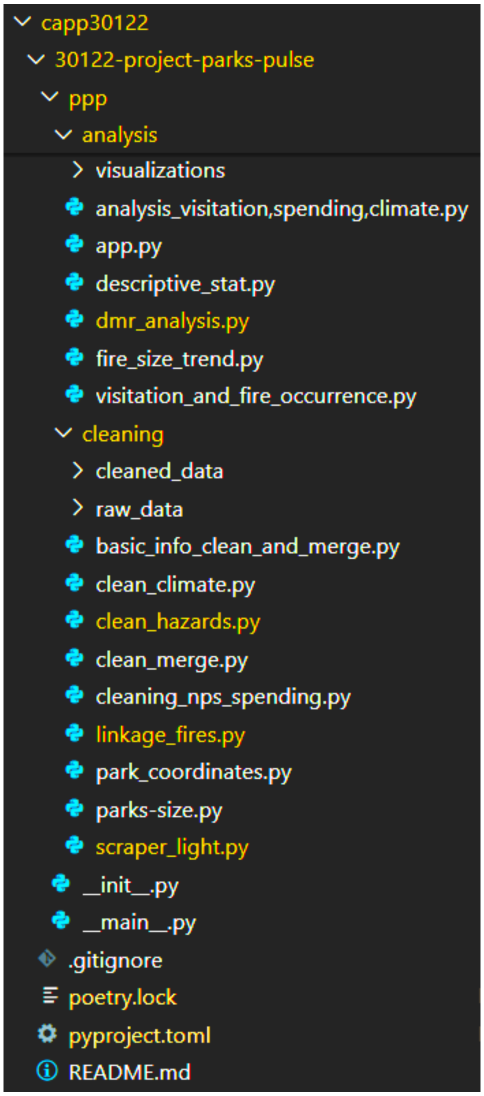
Project Abstract

One of the nation's most awe-inspiring assets, also known widely as America's best idea, national parks in the United States sees rapidly rising interest over the years, reflected in robust recreational activities, scientific pursuit, and conservation efforts. Besides preserving and stabilizing the Earth's ecosystems, national parks contribute economically, creating jobs and playing a pivotal role in a variety of sectors. While they are not always at odds, the duality of parks' role introduces a number of problems, and with the reality of climate change, tends to put parks themselves at risk.

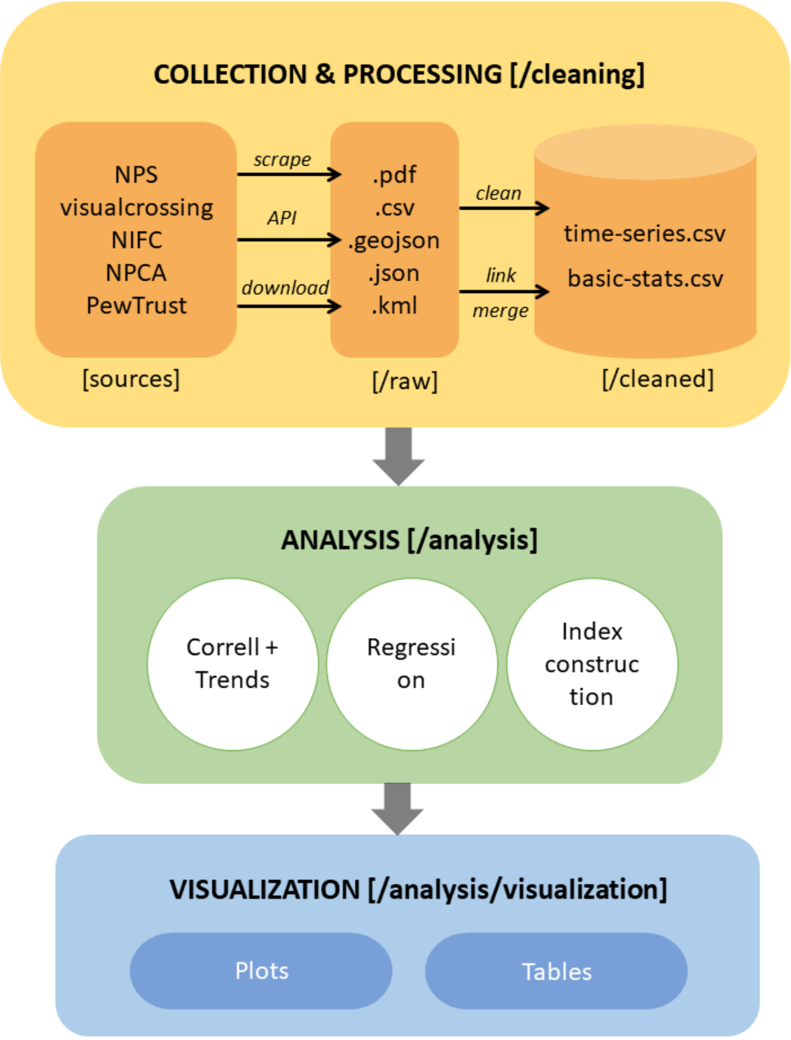
Recognizing how health is a broad term and is the product of the multifaceted problems that national parks face, the project aims to analyze the U.S. National Parks System (NPS) and other open-sourced datasets to construct a composite measure of individual parks' health status, called the Park Health Index (PHI). To compute PHI, we will examine and assign appropriate weight to different aspects of a park's current state, including park usage, climate data, park management, and hazards. From individual time series data by parks unit, we are also interested in exploring trends in visitation and impact of climate change among parks over 10 years (2011-2022). The goal of this project is threefold: (1) spreading awareness of national parks' health (snapshot and overtime), (2) uncovering patterns that might reveal useful information for assessment and evaluation of parks management, and (3) revealing prevalent problems to better inform policy-makers on areas on which to focus for better parks conservation and resources allocation.

Overall structure of software

The diagram shows the workflow we undertook, following the data pipeline of the project. Within the square brackets are relative paths to relevant folders. Not shown on the diagram but constantly practiced throughout the process is documentation and storage, done both locally and remotely using git. The diagram can be cross-referenced with the program directory for clarity.



```
capp30122
  30122-project-parks-pulse
    ppp
      analysis
        visualizations
        analysis_visitation,spending,climate.py
        app.py
        descriptive_stat.py
        dmr_analysis.py
        fire_size_trend.py
        visitation_and_fire_occurrence.py
      cleaning
        cleaned_data
        raw_data
        basic_info_clean_and_merge.py
        clean_climate.py
        clean_hazards.py
        clean_merge.py
        cleaning_nps_spending.py
        linkage_fires.py
        park_coordinates.py
        parks-size.py
        scraper_light.py
      __init__.py
      __main__.py
    .gitignore
    poetry.lock
    pyproject.toml
    README.md
```



Member responsibilities

Member	Project phase	Tasks	Code files
Minh	Project scoping	Scoped and prepared project proposal	
	Data collection & processing	• Outlined relevant metrics and sources • Pulled and processed NP hazards data: light pollution, fires, deferred maintenance/repair	scraper_light.py linkage_fires.py clean_hazards.py
	Data analysis and visualization	• DMR regression on visit and spending data • Made partial regression plot • Sourced codes for factor analysis and outlined idea/data metrics for index construction • Reviewed and made suggestions for index creation process	dmr_analysis.py

	Finalization	• Put together analysis and outlined html report • Wrote project paper	analysis.html proj_paper.pdf
Seongyeon	Project scoping	Initialized git repo	
	Data collection & processing	• Collected and cleaned NPS spending data • Merged basic-stats .csv file	basic_info_clean_and_merge.py bleaning_nps_spending.py
	Data analysis and visualization	• Correlation and trend analysis on spending, climate, visitation data • Made correl heat map and other trend graphs • Statistical analysis: compared visitation, spending, and climate data before and during the COVID-19 period (2020-2022)	analysis_visitation,spending,climate.py
	Finalization	Wrote Readme.md	README.md
Yi-Huai	Data collection & processing	• Pulled and cleaned climate data: temp, precipitation, uvindex, visibility • Merged hazards data and climate data in time-series.csv file	Clean_climate,py, clean_merge.py
	Data analysis and visualization	• Descriptive statistics • Index construction • Made index table, heat map, cumulative distribution for composite index, scree plot, and regional comparison for factor analysis	descriptive_stats.py, factor_analysis.py
	Finalization	• Constructed the application • Designed output html.	app.py
Diyanet	Data collection & processing	• Collected and organized NPS basic information including location, size, and visitation data • Provided guidelines on using GitHub for version control	national_parks.py parks_size.py
	Data analysis and visualization	• Performed analysis on fire occurrence data and its correlation with park visitation • Visualized trend graphs	fire_size_trend.py visitation_and_fire_occurrence.py
	Finalization	• Finalized project directory structure • Ensured all codes are properly committed and pushed to GitHub	

Short guide on how to interact with the application and what it produces

The application produces a static html report of data visualization and short analyses of the Parks Health Index (PHI) and other findings that we uncovered from the data.

To run this application, you'll follow the steps below:

1. Clone the repository
2. Navigate to the repository in your local machine
3. Install the necessary dependencies/packages: run *poetry install* in the terminal
4. Activate the virtual environment: run *poetry shell* in the terminal
5. Launch the application to see the visualization and analyses: run *python -m ppp*

What the project tried to accomplish and what it actually accomplished.

The project set out to aggregate relevant data points that constitute a national park's health and construct a composite index that reflects the park's condition, relative to other national parks in the system. While we accomplished this, there are some flaws in the outcome that we did not hope for. First, the KMO value that indicates the usefulness of factor analysis returns a value slightly under 0.5 for 2019, indicating that this method of computing for the composite index might not be useful across all years. This is potentially due to the correlation among the factors that got picked (most likely climate data, as different aspects of the weather might be strongly correlated). Second, the number of data points we managed to pull in the end is not what we had hoped, leading to the fact that the index might not be as reflective as we had wanted. Across different categories, we had planned to collect a total of approx. 20 data points. However, due to time constraint and unavailability of datasets, we were only able to acquire 12 variables, of which 9 are usable to construct the index.

One of the outcomes of the project is to uncover some patterns that might be useful for park management. Apart from some interesting findings about trend over the 10-year period examined for individual parks and the whole national parks system, the project did reveal a potential pattern regarding deferred maintenance and visitation that might be helpful for park planning.

While our narrative is fairly well-constructed, we think visualization could have been improved with a more creative approach to make the report flows better and more engaging.