1)

Predicting National Park Service Trail Popularity With Physical, Behavioral, and Reported Measures

Report by Sammy Rose Dobrozsi

October 7th, 2022

With over 21,000 miles of recreational trails managed by the United States National Parks Service, it is critical to understand how many trail users are going to different trails and for what reasons in order to allocate resources and direct future development. The present data was scraped from the website AllTrails roughly three years ago and includes behavioral measures of trail users (e.g. trail usage levels), reported measures by trail users (e.g. trail user reviews), and physical attributes of trails (length, elevation). AllTrails is a trail database, hosting GPS files of trails entered by users supported by metadata, focused on a phone app that can assist with trail selection and navigation. The different types of data are advantageous for gaining a more holistic view of what makes trails popular (and presumably more desirable to users).

2)

This data was retrieved from Kaggle: https://www.kaggle.com/datasets/planejane/national-park-trails

The original data consists of 3,313 trail listings from AllTrails. Certain non-usable variables were removed from the .csv in Excel prior to importing into R, such as Name, Country, State, Lat and Long coordinates, Features, Activities, and Units. Trail ID was converted to string data, and Visitor Usage missing values were filled using rounded mean imputation.

The study at hand seeks to predict trail popularity (y), measured by a compound index, using seven different physical, behavioral, or social measures of trail use. As popularity is no longer visible on AllTrails since this data was scraped, it is not possible to know how it is computed, since there may have been other trail attributes added/removed since then that contributed to its calculation. The seven measures of trails and trail use are:

Physical Measures

Trail Length measured in meters,

Trail Elevation Gain measured in meters,

Trail Difficulty Rating in a scale of odd numbers from 1 to 7, set manually by AllTrails "based on trail condition, steepness of grades, gain and loss of elevation, and the amount and kinds of natural barriers that must be traversed,"

Route Type which describes whether trails are a loop back to the start, and out-and-back that returns along the same path to the start, or a point-to-point that does not loop or return to the start,

Behavioral Measures

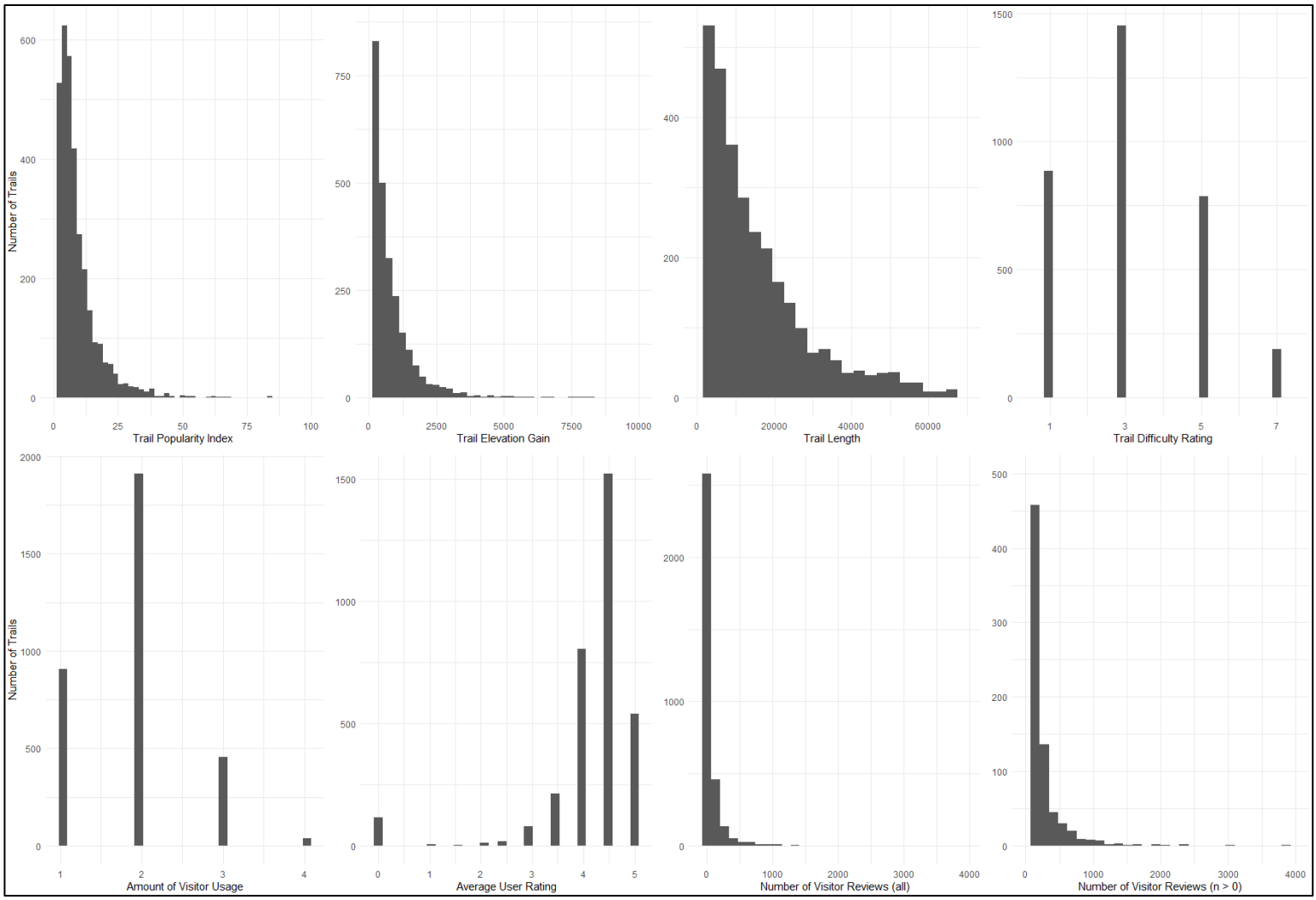
Visitor Usage which rates the number of trail check-in's on a scale from 1 to 4,

Number of Reviews which is the number of reviews left by trail users on a given trail's AllTrails listing,

Reported Measures

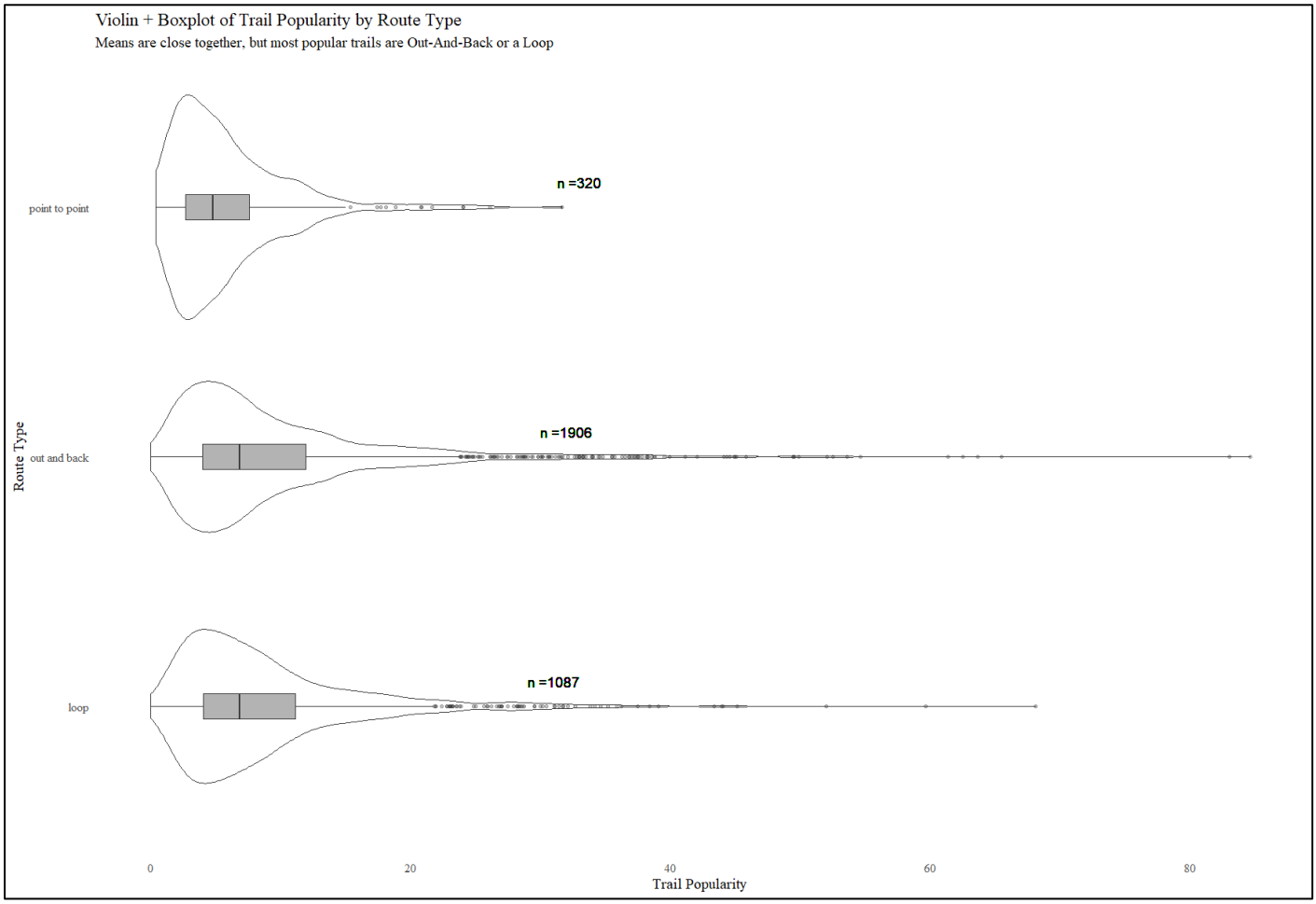
Average Rating reported by trail users on a scale from 1 (worst) to 5 (best) by 0.5-point steps.

3)

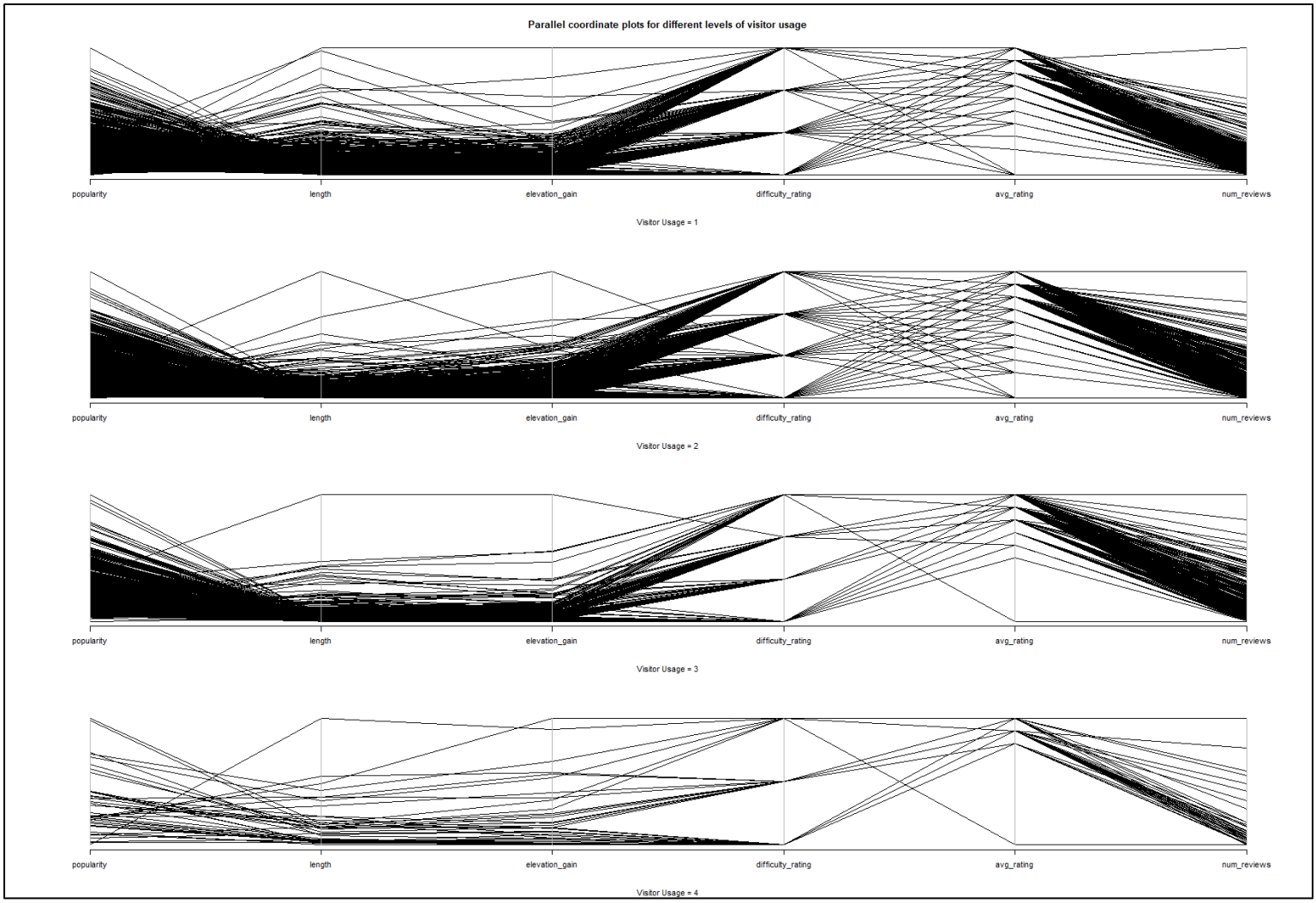


Most numeric variables are skewed right, signifying that most trails are low in popularity, have little uphill, are short, and have few relatively few reviews (the vast majority have zero; a zoomed-in histogram on the bottom right of number of reviews is included to give more insight into non-zero relative frequency). Average User Rating is strongly skewed left. Almost every trail on AllTrails *has* a rating, there are very few in the 0 column. When users do rate a trail, they appear to mostly report only their positive experiences. Trail Difficulty Rating and Amount of Visitor Usage are both skewed right; most trails receive low to moderate usage (there are over 21,000 miles of them, and not all of them are easily accessible to most people), and are low to moderate difficulty. The pattern of difficulty reflects

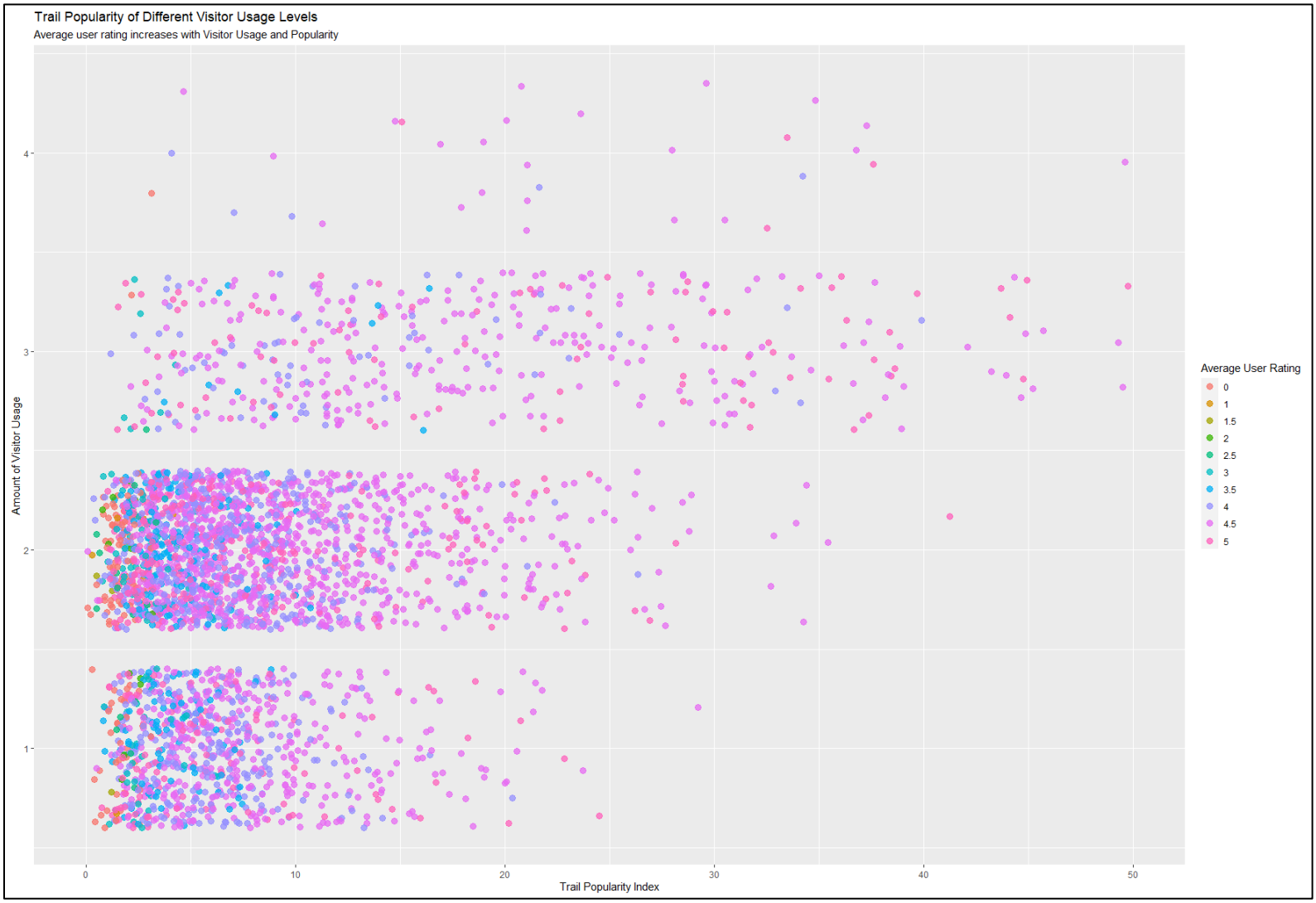
that of many outdoor recreation activities – there are typically fewer users of recreation resources that require more skill or specialized equipment to participate in. In the context of trails, these trails may be difficult or expensive to maintain, require expensive backpacking or climbing equipment, or could take place in higher risk settings that many users do not seek out.



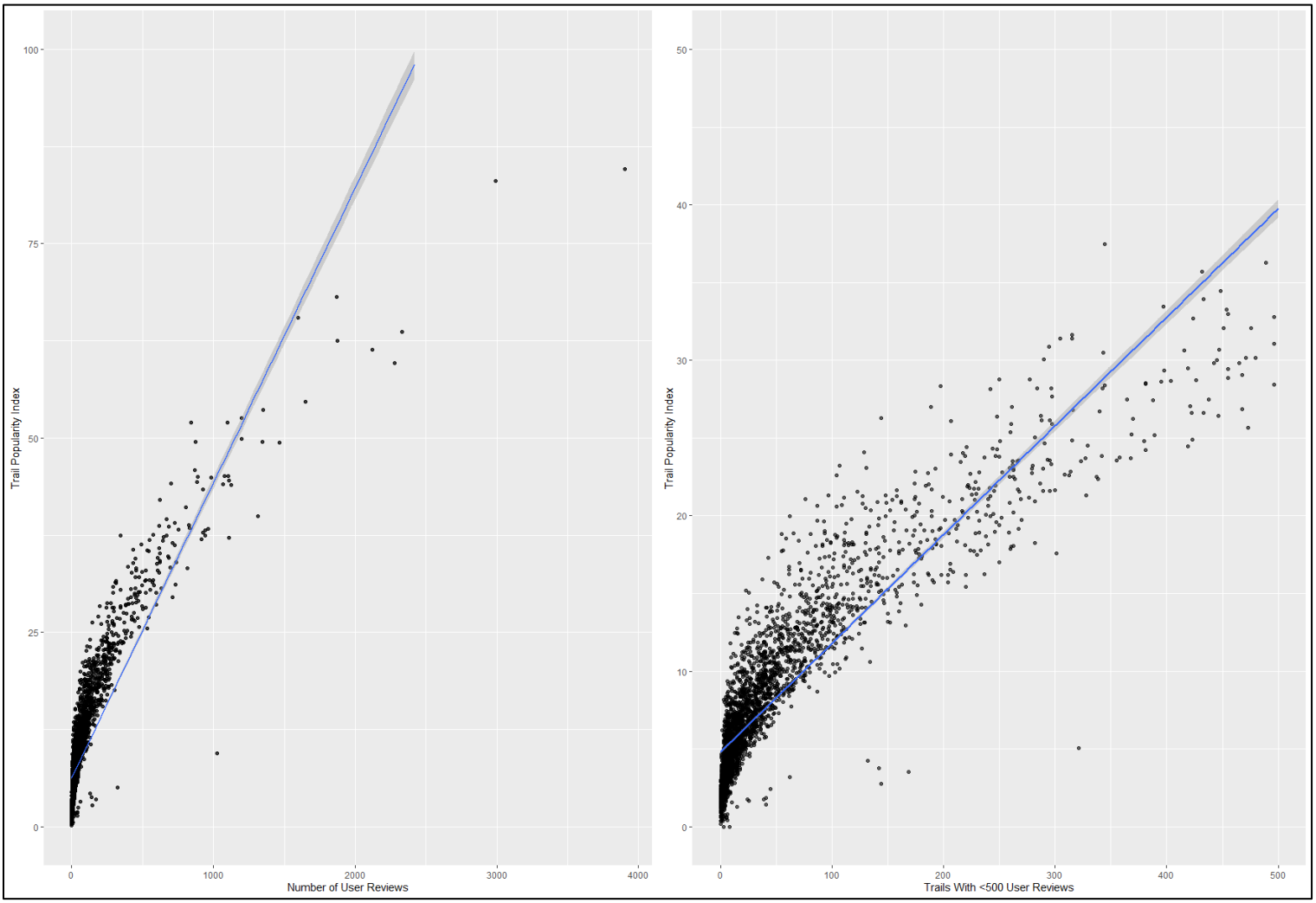
When comparing the distribution of the three different types of trail, their mean popularities are relatively low and close together. Point to point trails have a notable tendency to be less popular and do not have any high popularity representation in the data, whereas out and back contains the highest popularity trails. Many trails are built by design to give users access to a noteworthy landmark, such as a waterfall or scenic view. It makes sense that some of these are the highest popularity, since many people may only tend to report on their experiences on a trail given a peak experience like when seeing something especially beautiful.



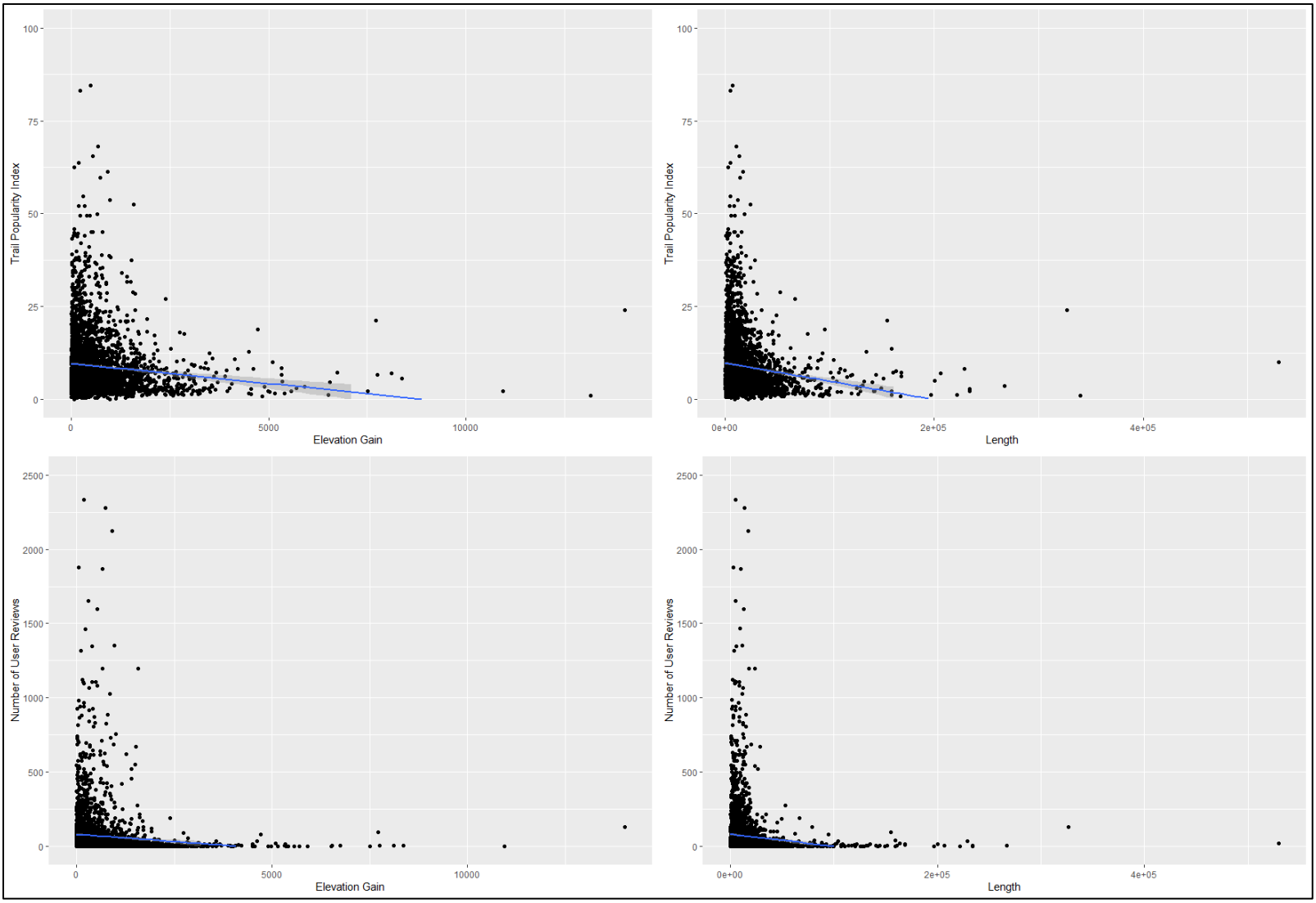
A parallel coordinate plot of the data broken down by the four visitor usage levels shows some important trends and aspects of the data. The highest usage trails have the highest average ratings and are skewed towards lower difficulty, reflecting the same trend with difficulty rating – most trail users are seeking out less difficult trails to recreate on. Conversely, lower usage trails tend to be the trails that are longer and higher difficulty. All visitor usage levels reflect similar distributions of popularity, though because there are fewer high usage level trails, the relative popularity is much higher. Higher visitor usage seems to correlate to fewer reviews, which is counterintuitive.



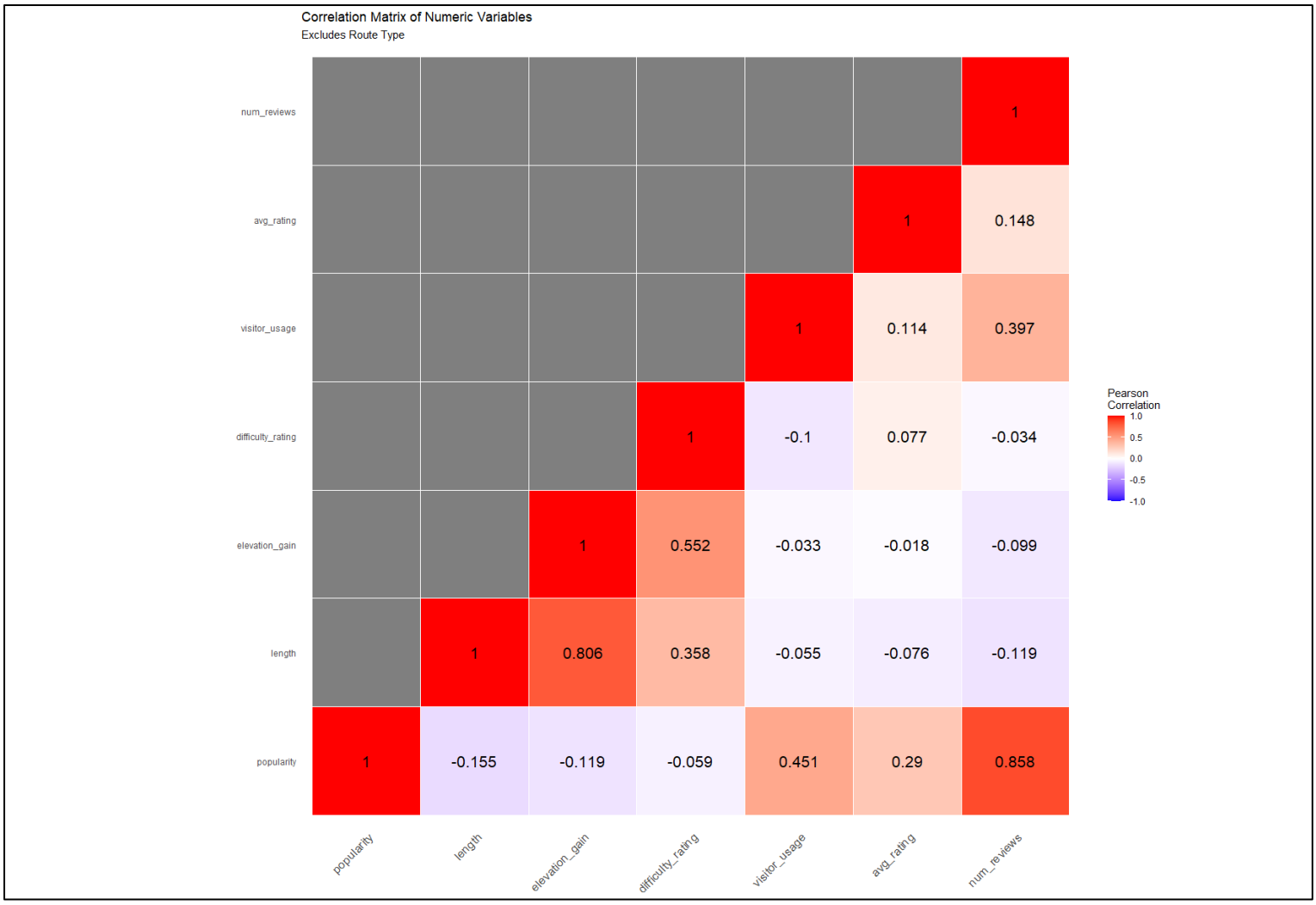
The highest value outliers are cut out to improve the interpretability of the majority of the data. When the popularity index is broken down by visitor usage, the relationship becomes more clear. Higher usage trails are fewer in number and contain the *highest popularity* trails. The sparsity of low popularity among higher usage trails contributes to a stronger positive correlation between popularity and visitor usage. Very few trails with a popularity index over 5 have an average rating below 3, which reveals that popularity takes into account perceived quality by users on top of engagement on the AllTrails platform. Low usage level trails cannot be popular even if they are rated highly.



Trail popularity climbs very steeply with the number of user reviews. This effect may largely be in part due to the lack of trails having many negative reviews, as there are only a handful of trails with many reviews that have a low popularity index. The regression line closely follows the shape of the data in the zoomed-in graph, suggesting relatively low influence of the highest-reviewed outliers.



There are so few trails of especially high length or elevation gain in the data that they do not exert much influence on the popularity or number of reviews. Because trails with especially high elevation gain may require climbing equipment, or especially long trails may require backpacking equipment, they won't receive large amounts of engagement on the AllTrails platform and therefore have higher popularity. This is surprising when taking into account the importance of peak experiences in outdoor recreation; trails that involve overcoming some difficulty or that lead to especially high or far away places that can contain exceptional beauty may be attractive to dedicated hikers, but they are not popular based on how AllTrails measures it.



A correlation matrix of variables (excluding the non-numeric variable Route Type) reveals some moderately and some significantly correlated variables. Elevation Gain and Length are moderately correlated with difficulty rating, because those are both variables taken into account when calculating difficulty. Elevation Gain and Length are strongly correlated with each other, which indicates that the dataset has few long trails that are flat, and few high elevation trails that are short. Popularity is most strongly correlated with Number of Reviews, which suggests that the Popularity Index is driven in large part by engagement on AllTrails. The strongest negative correlations are between Length and Elevation Gain, and popularity, meaning that longer trails and trails with more uphill are less popular. Popularity is only barely negatively correlated with Difficulty Rating. While none of the relationships are strong enough to draw conclusive inference, it's possible that because Popularity is in part calculated through

engagement on the AllTrails platform, difficult but high engagement trails skew this relationship more neutral. The moderate positive correlation between Visitor Usage and Number of Reviews makes sense: visitors typically only review trails they physically go to, and trails with high usage will likely get higher numbers of reviews. In general, there is little autocorrelation among the predictor variables, but also only strong correlation among very few variables with the response variable. The predictive power of this dataset is undiscernible with basic visualization alone.

4)

In general, the conclusions that can be drawn from this visual analysis have high face validity. Trails that see higher usage, have more reviews, and are easier to traverse and more accessible to more people are more popular. Some more difficult trails are more popular and have high engagement on AllTrails, suggesting that the population of dedicated hikers *can* be high, but based on this dataset alone it is hard to distinguish any elements of those that could be extrapolated to management action.

Future analysis should include the geographic portion of the data, as location within the US may help focus further research on what aspects of certain regions make hiking more popular. More granularity in physical attributes would also increase the usefulness of the analysis; because out and back trails are the most popular, knowing what lies on the far end of them or along the way could improve the predictive power of a popularity model.

Given the relatively low frequency of low average rating reviews, different types of data may be required in order to resolve any perceived user issues on those trails, or prevent negative experiences on other trails. Qualitative interviewing would best reveal what attributes lead to negative experiences on trails.