National Parks Data Storytelling

Springboard Capstone 1 Project

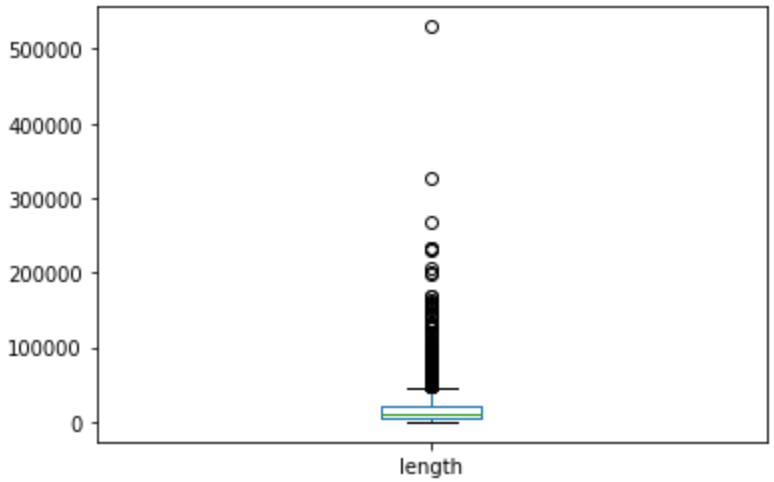
Evan Hintz

This dataset includes all national park trails of the United States. My end goal is to discover what factors have the biggest effect on trail popularity. After some cleaning and reorganizing of the data to make it easier to analyze I moved on to exploratory data analysis and data storytelling as follows (some cleaning was done during this section as well as more inconsistencies and inaccuracies were found).

First, I took a look at how the trails are distributed throughout the states. Location and accessibility will be very important in trail popularity and how often they're used. California has a huge amount of trails, about 20% of the total. This makes sense as a large portion of the state (which is quite large) is national park land. California has a great location which also gives it generally nice weather and is already a highly-populated area. We'll look at it's trail use and popularity later.

I was surprised to see some states with as few as 1 or 2 total trails. This was where Georgia was found to have some bad data, an "international park" that should not be present which was quickly removed from the set.

Another huge factor will be trail length, people have varying levels of skills and physical conditioning when it comes to hiking. Some people go out to push their limits for weeks on end, while others are simply out for a casual sight-seeing stroll.



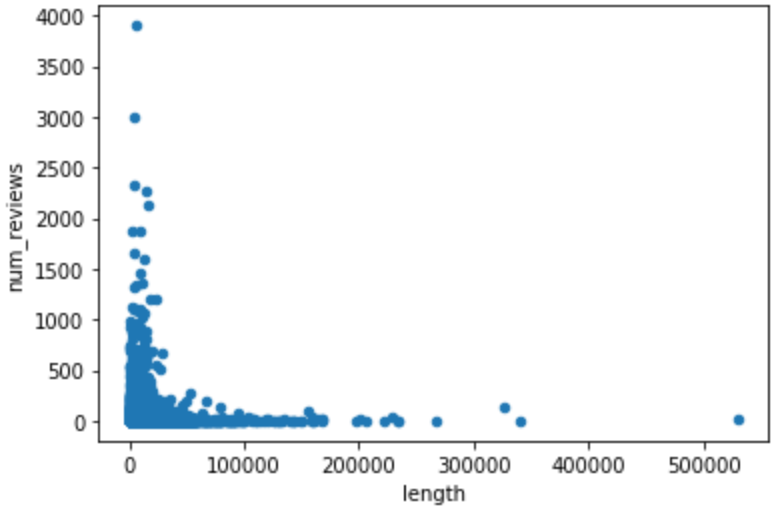
The trails have quite a large range of length, from hundreds of miles to one that is so short it is listed at a length of 0. We'll first look at that outlier that is over 500000 (approx yards). It turns out this 'trail' is actually a driving route that is about 300 miles

long.This further shows how the trails are used and the difference in the types of consumers. We'll have to be careful of some of these outliers when calculating other statistics. On the other end, the Newspaper Rock Trail has 0 length and seems to be a large rock that doesn't require any actual hiking but contains hundreds of Indian Petroglyphs (carvings). This is a fantastic trail option because it is accessible to almost everyone and has a unique feature that could attract people of all ages and interests in outdoor activities. Having such a particular and uncommon feature this also needs to be kept in mind when comparing trail popularity as it is not something that is included directly in the data set.

Another interesting variable to explore is the elevation gain for each trail. It has a large factor in trail difficulty and in turn can affect consumer's trail choice. The box plot is almost identical to that of length. A few very very large outliers and a high density through the lower half of the range (2000 to 5000 ft). This could be interpreted as the demand of trail type, more people want a 'do-able' trail that doesn't require a lot of exertion but it is also restricted by the topography within the parks.

A box plot for the number of reviews produces again, a very similar distribution. It is likely that the more popular trails receive a higher number of reviews, and as they obtain more reviews even more people are interested in exploring the trail.

The final plot we'll review here is the relationship between the number of reviews and the trail length.



It is clear the shorter trails are used more often and as such receive more reviews. It's important to remember that the number of reviews does not include whether it was positive or negative - some trails may have received a large number of reviews but has in fact been used less because of them.

This dataset invites a multitude of different analyses all with many insights and possible interpretations. I would like to continue exploring these by looking at the effects of the variables on the number of reviews, trail rating, trail usage, and maybe most importantly the popularity. In the end I would like to have completed a regressional analysis of the previously discussed variables (length, elevation gain) as well as all the features and activities included (these were not talked about at all here but include things like if the trail has waterfalls, wild-flowers or if biking, camping, etc are allowed).