Madison Smith January 26, 2018

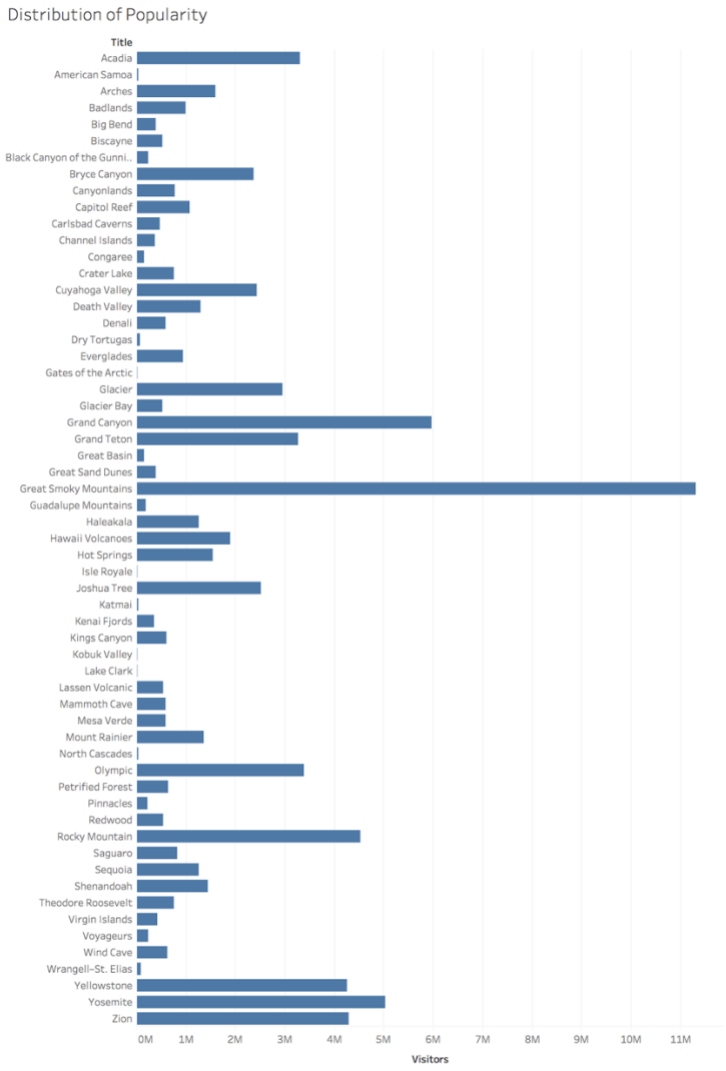
Homework 2: Exploratory Data Analysis

I found this data on data.world. Kevin Nayar compiled the data from the US National Park website. (link to data) I was lucky and the data set was already in a usable format with column titles and a good number of entries, so I did not have to do anything to the data set.

The data set has a row dedicated to each national park. The columns in the data include size, coordinates, and whether or not the national park is a world heritage park. The data set includes all 59 United States national parks.

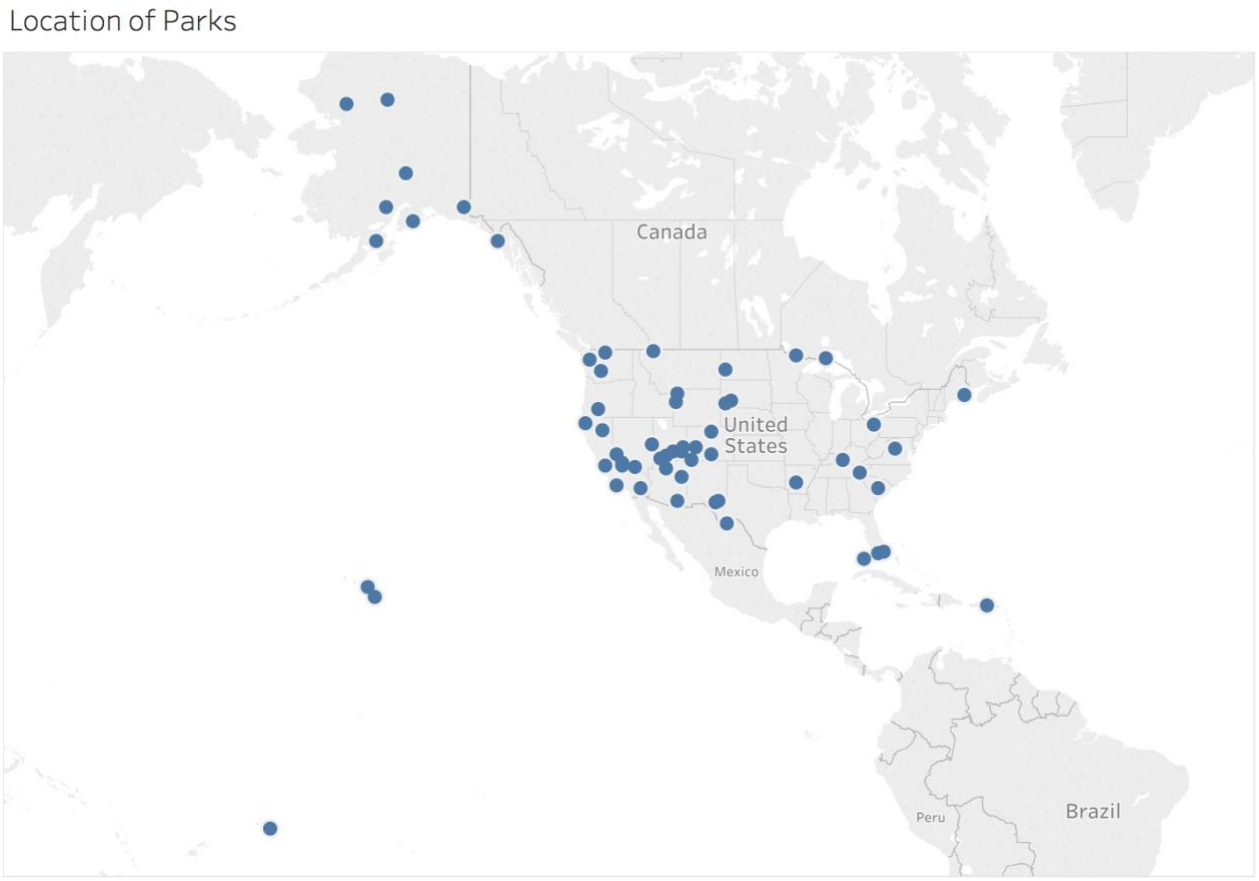
I started the exploratory analysis by creating visualizations that showed the distribution of the data. I made sure the data seemed plausible and that there were not obvious gaps, mistakes, or duplicates.

I graphed the number of annual visitors to each of the national parks and sorted the names of the parks alphabetically to make it easy to find the data given a name of a park.



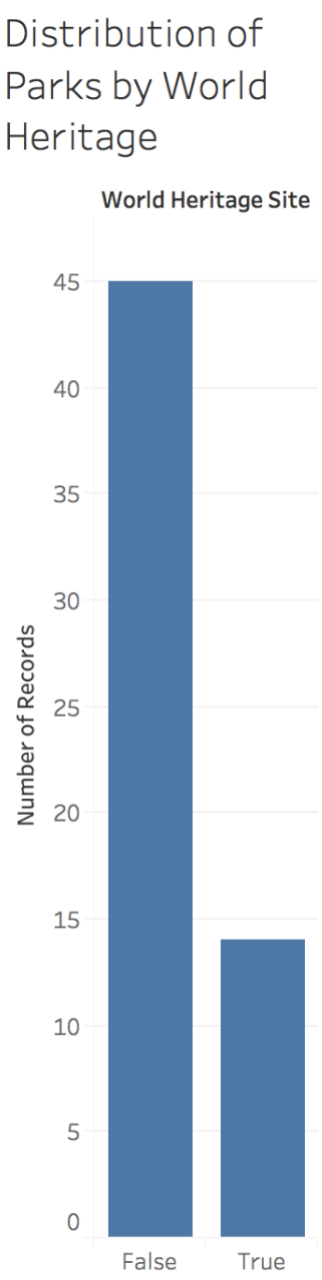
I noticed that the parks that are well known, like the Grand Canyon, Yellowstone, and Zion, rank high in the data, which makes the data set seem plausible. I did not realize that the Great Smoky Mountains National Park was so popular, but a Google search tells me that it is a popular park.

I graphed the longitude and latitude coordinates of each park on the map to see if there was obvious missing data.



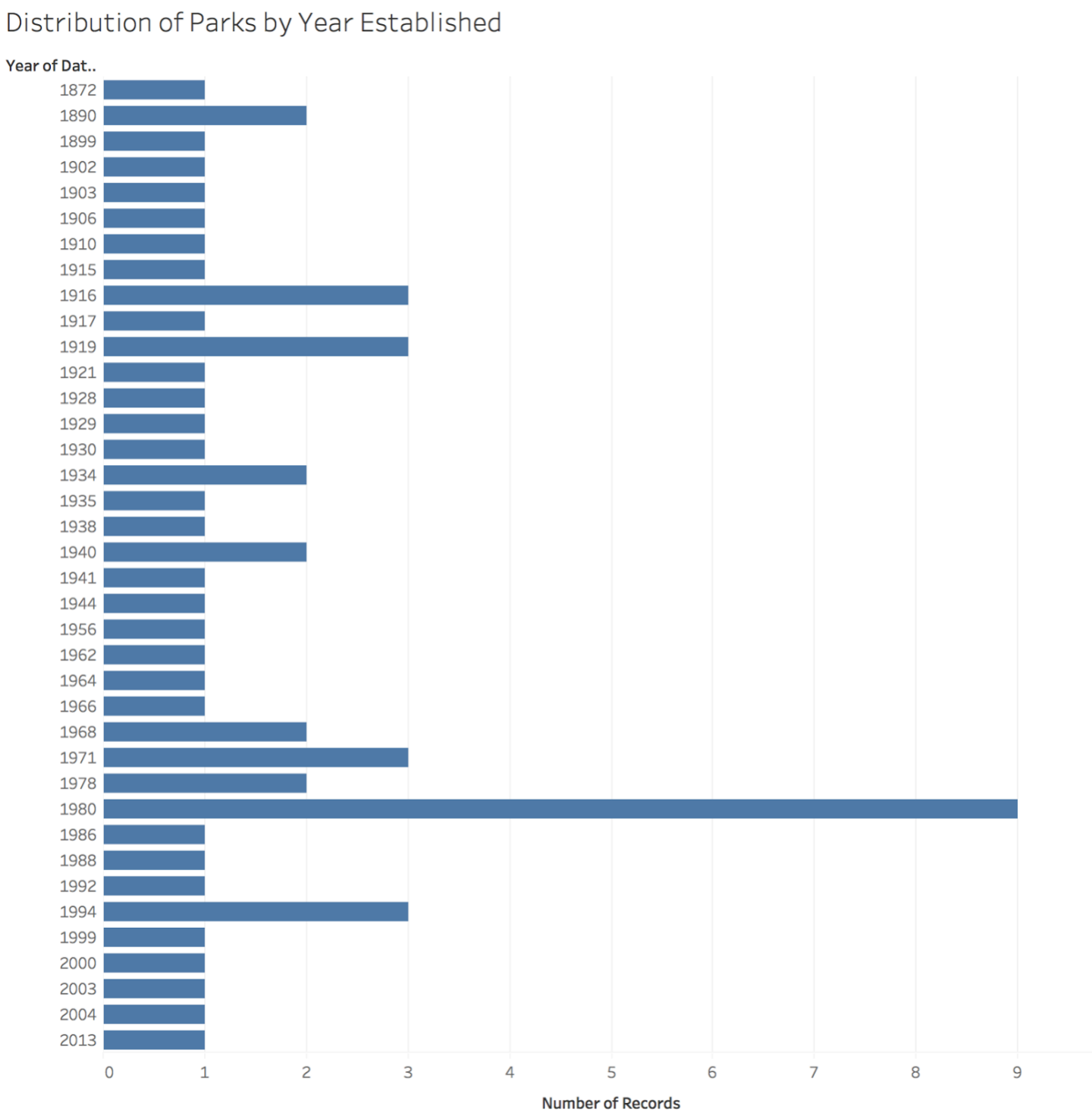

The locations of the national parks seem pretty accurate. There are national parks scattered across the country. The national parks I am familiar with, like Mt. Rainier National Park and Acadia National Park, are all represented. However, I noticed that the data also includes national parks in US territories, which I thought was important to note.

Next, I compared the number of national parks that are world heritage sites to the number of national parks that are not world heritage sites.



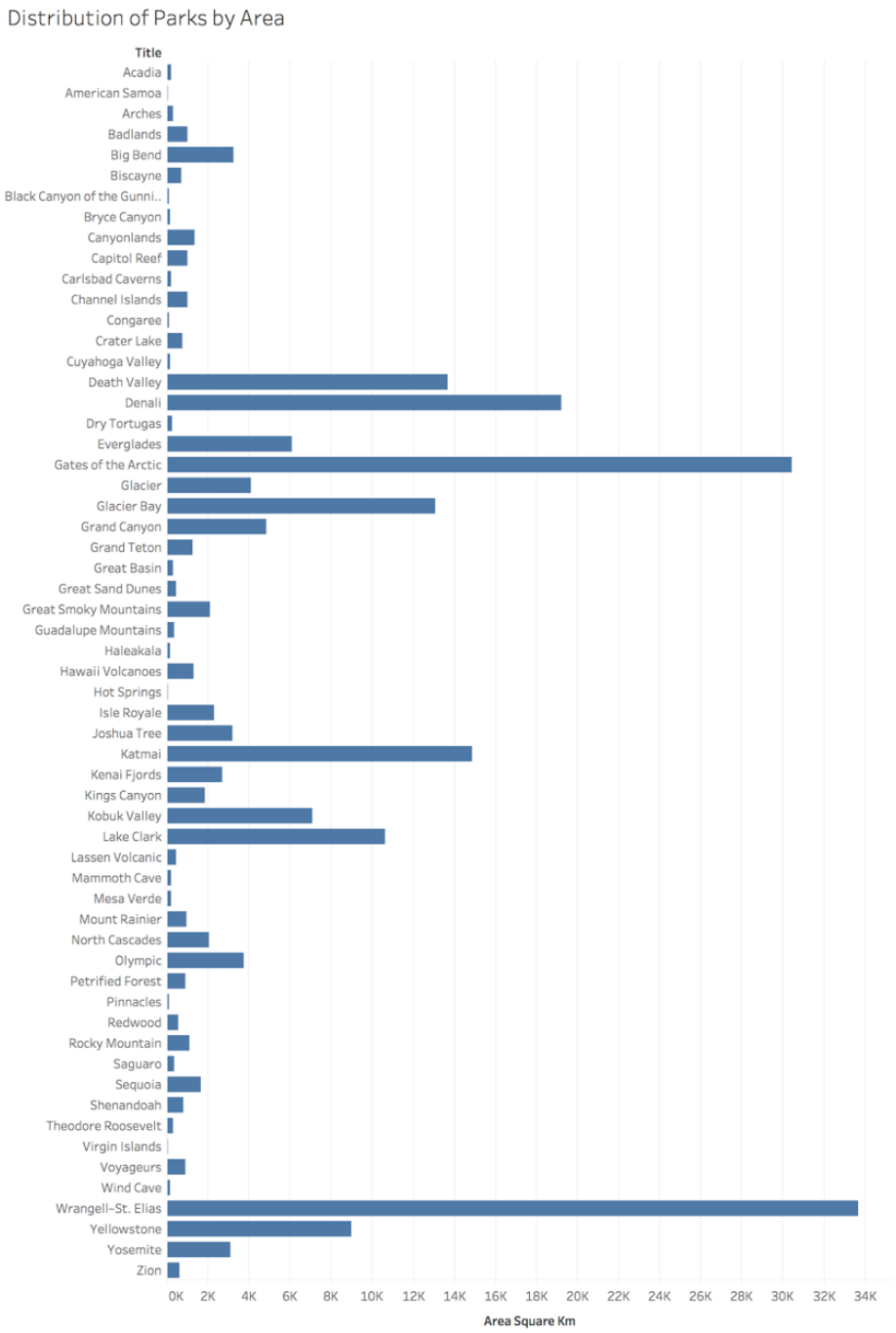

I did not realize that any national parks were world heritage sites, so I was surprised that nearly 15 national parks were designated heritage sites. But a Google search tells me that this data is plausible. I also thought it was important to note that the number of world heritage national parks were significantly outnumbered by the number of national parks that are not world heritage sites.

Then I graphed the number of parks that were established each year.




It is interesting that 1980 was a very popular year to establish national parks. But this data seems plausible- one or two parks every couple years seems accurate.

Then I graphed the area (sq. km) of each park and sorted the names of the parks alphabetically to make it easy to find the data given a name of a park.




I googled the size of these parks and the numbers seemed to match pretty well, so I do not think there is any reason to believe this data to be inaccurate. However, I thought it was interesting that the biggest parks were all parks I had never heard of.

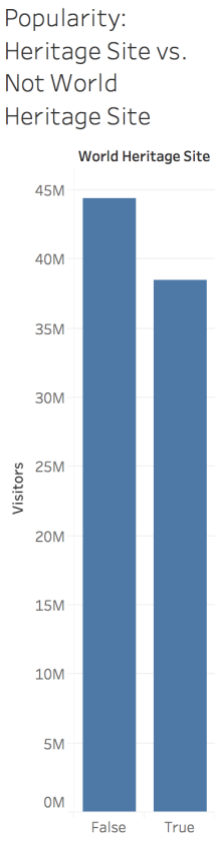
After I finished going through the distributions of the national park data, I decided that there was nothing that stuck out as missing, inaccurate, or duplicated. I decided I could trust this data set and started asking questions of it.

**Question: Are there any common characteristics of popular national parks? If so, what are they?**

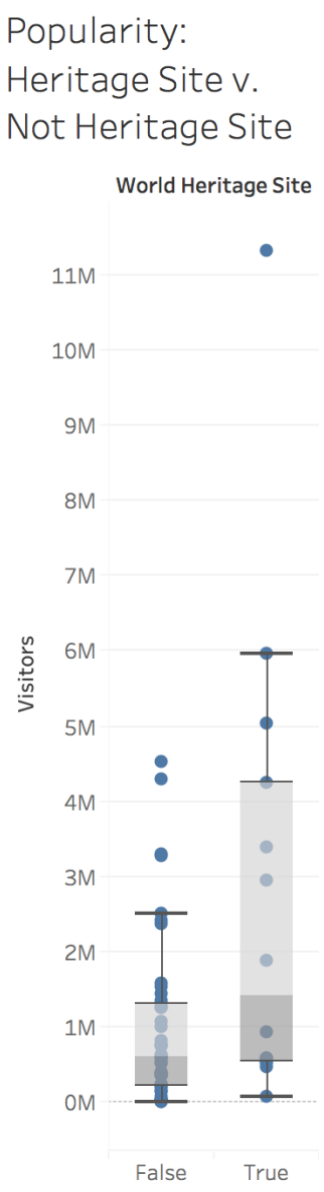
NOTE: For the purpose of this project, I am defining popularity as the number of annual visitors.

To answer this question, I tried to chart various characteristics of national parks against the number of visitors and tried to find patterns. The first characteristic I looked at was whether or not the national park was a world heritage site.

**Question: Is there any correlation between whether or not a national park is a world heritage site and the park's popularity?**



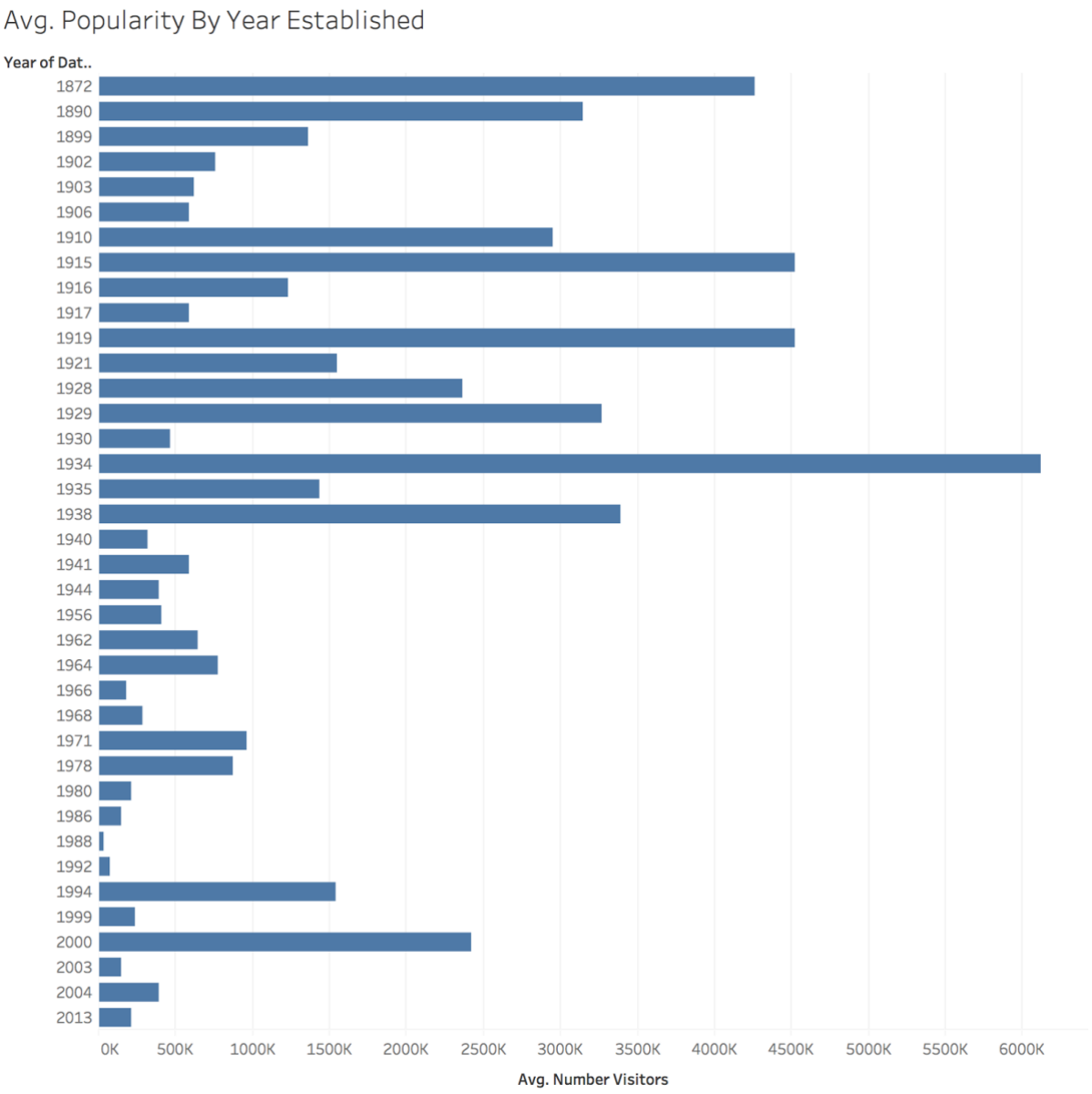

I created this visualization, but then I realized that this visualization showed the sum of the visitors to each type of park and I knew from a visualization above, that there were significantly more national parks that are not world heritage sites than are world heritage sites. So, it makes sense that national parks that are not world heritage sites have more visitors because there are more parks. So, I redid the visualization and disaggregated the number of visitors by park.




This visualization makes a lot more sense than the one above. While national parks that are not world heritage sites have more total visitors, national parks that are world heritage sites have more visitors on average. The median and upper interquartile range for the data of world heritage parks are significantly higher than the parks that are not world heritage sites. So, being a world heritage site is one characteristic of many popular national parks.

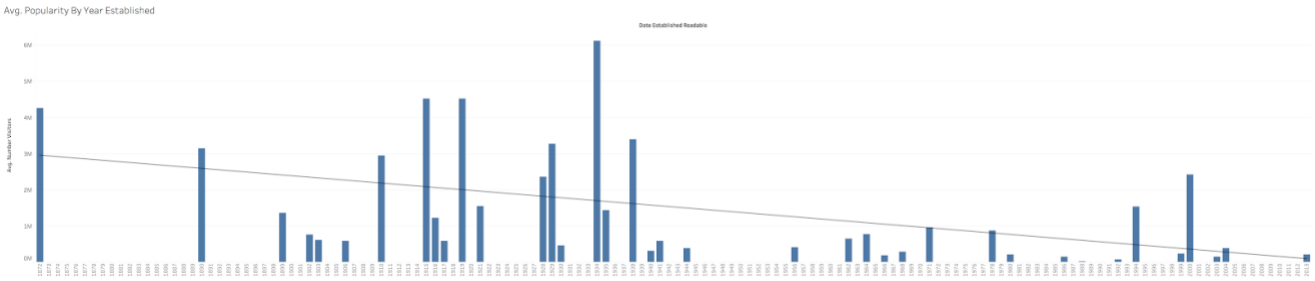
Next, I wanted to see if there was any correlation between the year the national park was established and the parks average popularity.

**Question: Is there any correlation between the year a national park was established and its current popularity?**




In order to create this visualization, I had to create a calculated field that calculates the average number of visitors, not the total number of visitors. I did this by using the AVG() function built into Tableau.

I made the above graph and thought I saw a slight downward trend, where parks established earlier are, on average, more popular today. So, I used the Tableau built-in trend line and made this new visualization. This visualization also adds in the years where no national park was established.

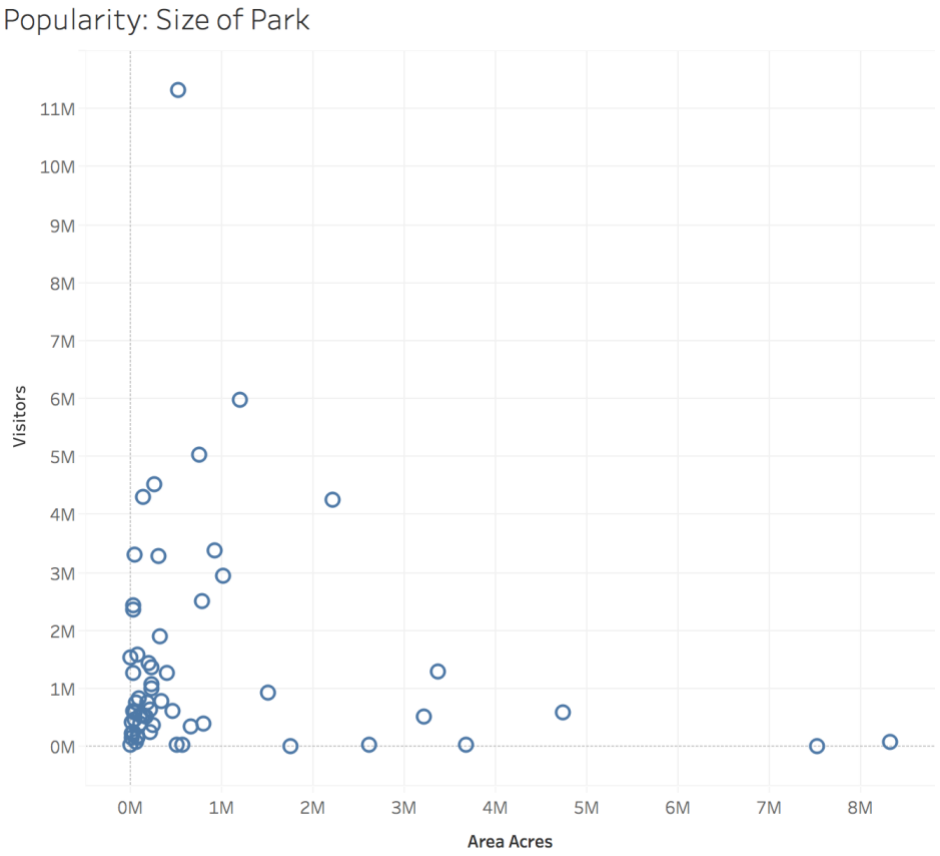



The first thing I noticed was that there is an obvious downward trend in popularity over time. I thought this trend was even more noticeable when the years where no national park was established were added to the graph. So, I learned another characteristic about popular national parks- that they tend to have been established earlier.

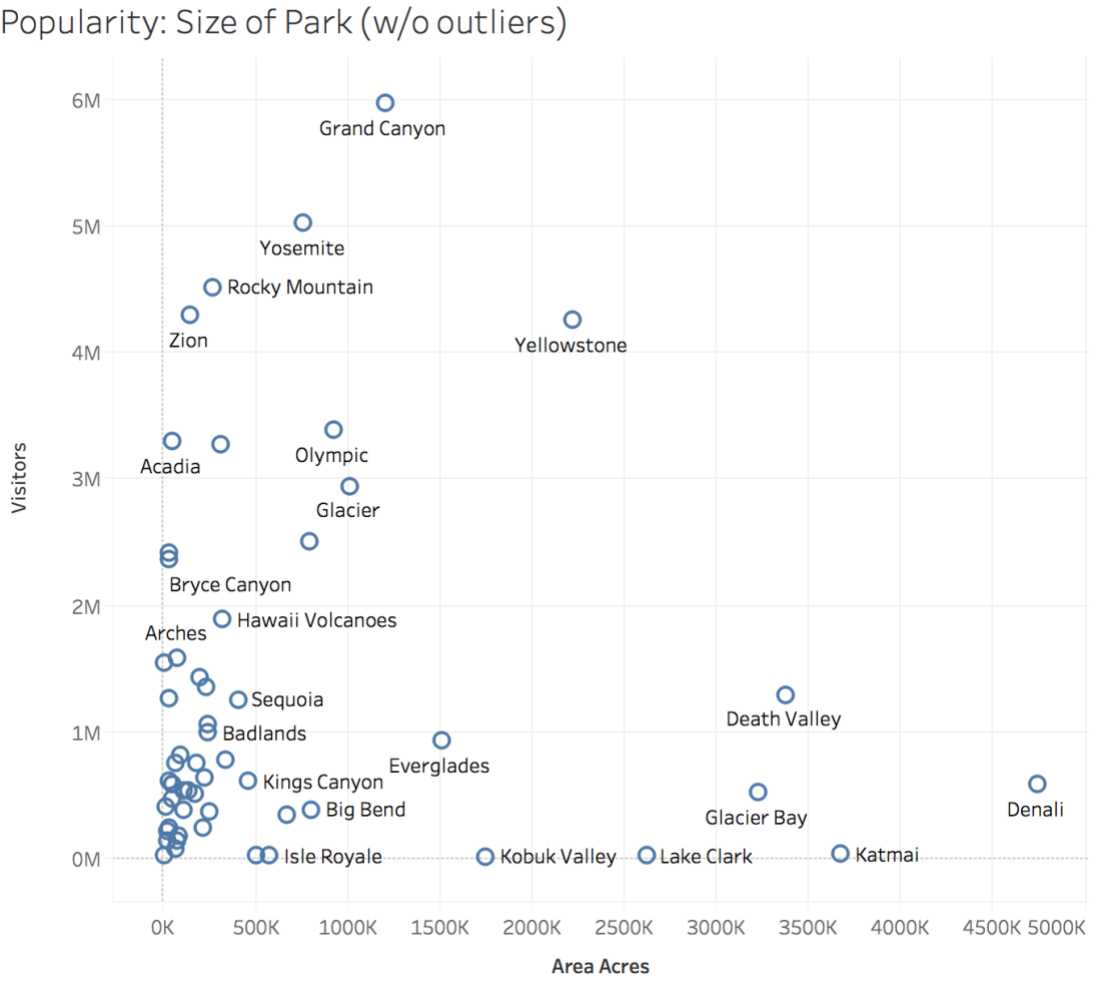
Next, I wanted to see if there was any correlation between the size of the park and the parks average popularity.

**Question: Is there any correlation between the size of a national park and its popularity?**

In order to answer this question, I looked at the area of the park (in acres) and graphed it against the number of annual visitors.



I noticed that there are a lot of outliers in this data, but I thought that there might be some sort of loose upward trend where larger parks have more visitors. In order to look into this further, I decided to filter out the more extreme outliers in order to get a closer look at the cluster of data points near the origin of the graph.




After looking at this graph, I decided that there might be some sort of very loose upward trend with a lot of outliers. I decided to use Tableau built in trend line function to see if it agreed with me.

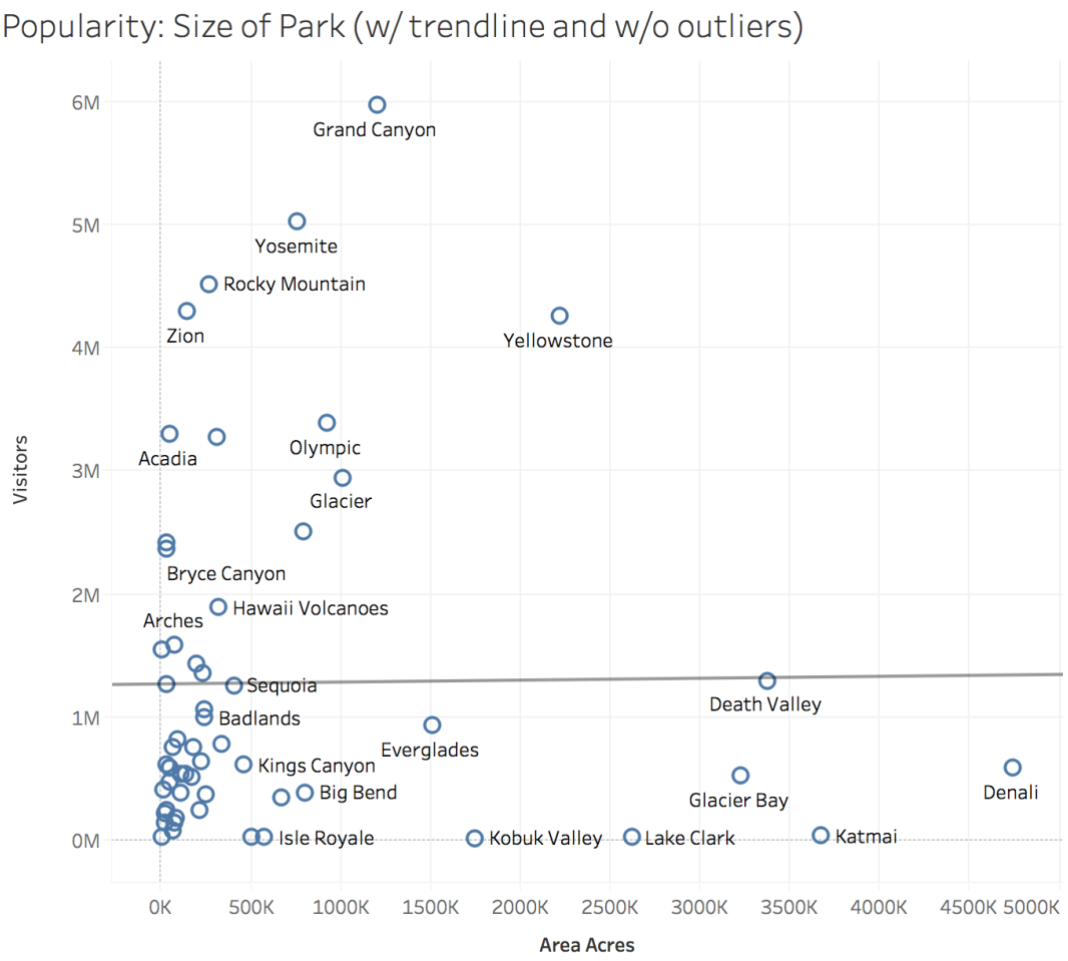
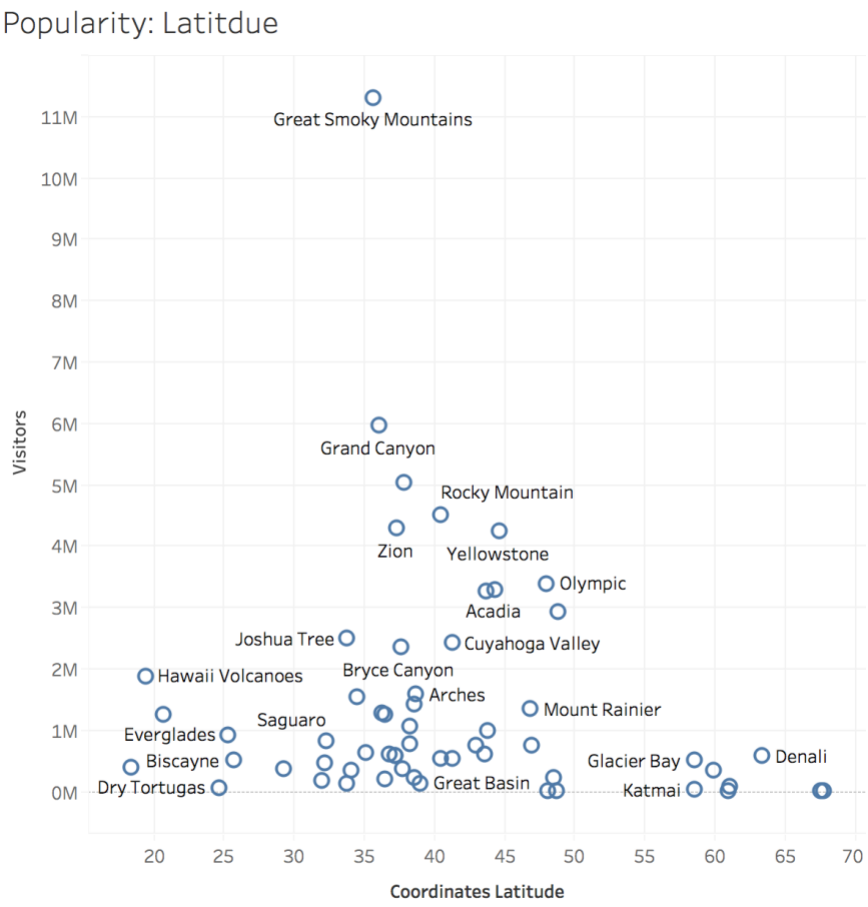



Tableau did not agree with me. I decided that there seems to be no correlation between the size of the park and the park's popularity.

**Question: Is there any correlation between a park's latitude coordinates (how north or south a park is) and the park's popularity?**

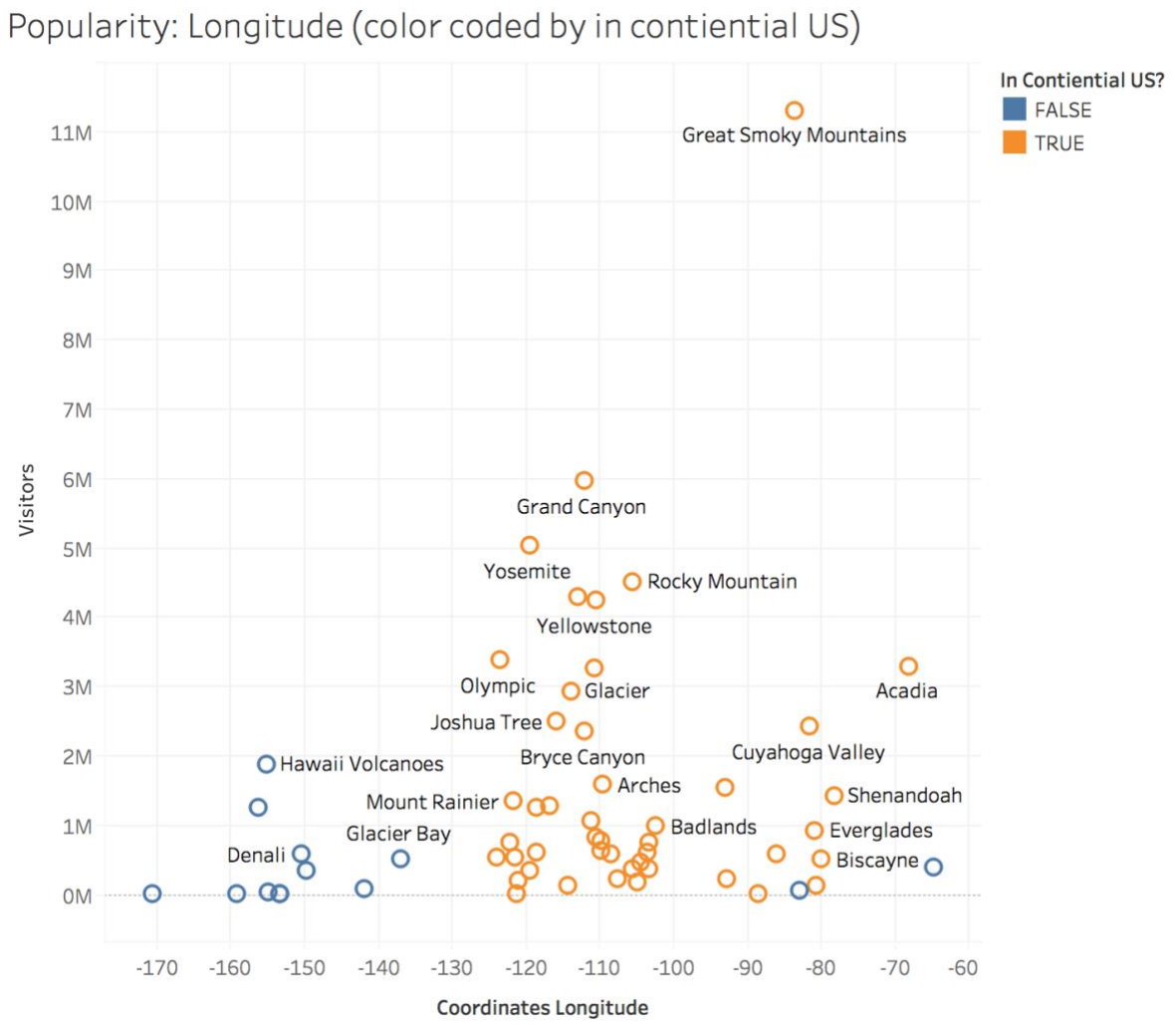
In order to answer this question, I decided to graph each parks latitude coordinates and the number of annual visitors of the park.




The first thing that I noticed was that the data seemed to be somewhat bell shaped, where parks at the middle coordinates seemed to be more popular. I thought this was odd, but then I noticed that the parks at the upper end of the coordinates, like Denali and Glacier Bay, are all in Alaska and all the parks in the lower end of the coordinates value graph are all in Hawaii or the US territories in the Pacific Ocean. I thought that maybe being in the continental US might impact a parks popularity.

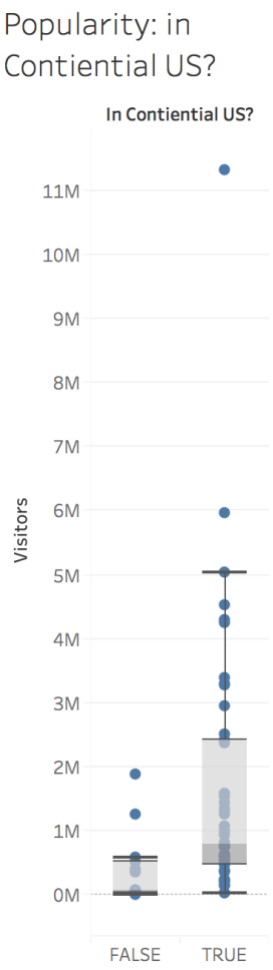
**Question: Is there a connection between being located in the continental United States and a parks popularity?**

I created a calculated field based on whether or not the park was in the continental United States. Then I used this field to color-code the parks.



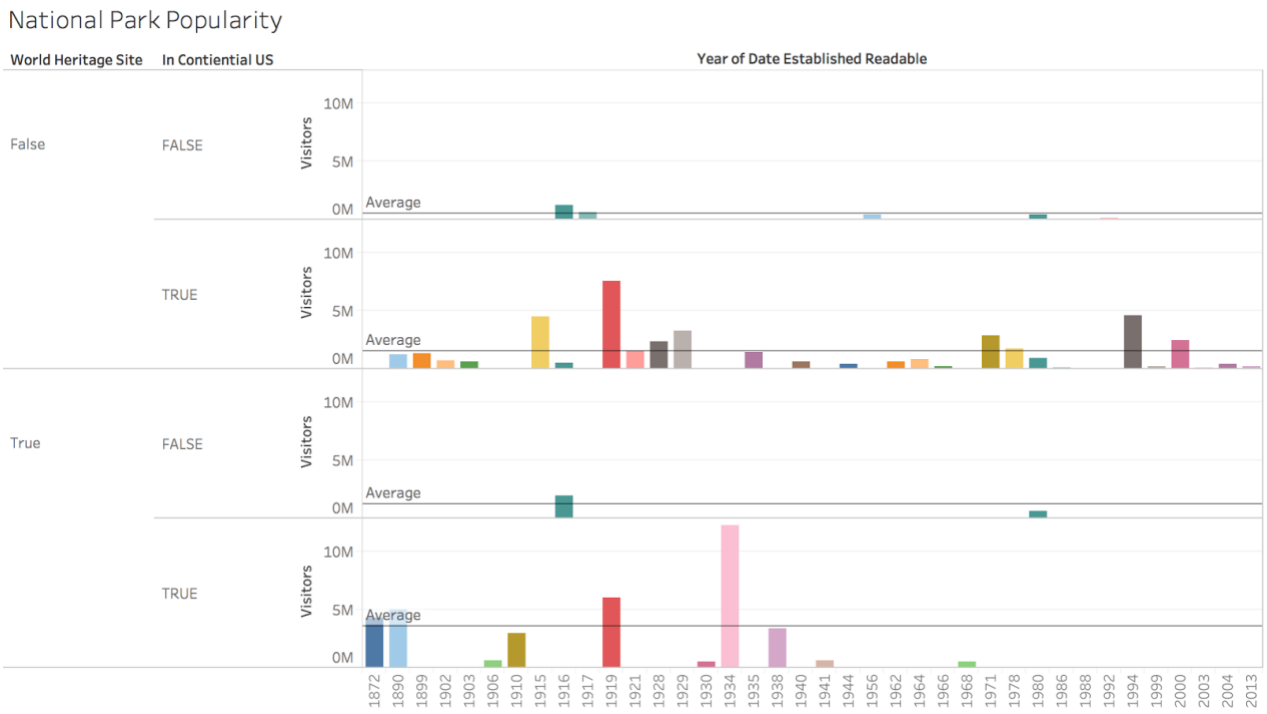

Once I color coded the different parks by whether or not the park is in the continental United States, it was easy to see that the most popular parks were ones that were in the continental United States.

Then, to make the difference in popularity between the parks in and not in the continental United States clearer, I made a box-and-whisker plot visualization, sorting the various parks.




It is clear that the parks in the continental United States are significantly more popular on average, than parks not in the continental United States. The median popularity value for the parks in the continental United States is greater than the most popular park not in the continental United States.

Circling back to my original question, the common characteristics of popular national parks include being located in the continental United States, established earlier in history (particularly in 1919), and being a world heritage site. I also learned that the size of the park does not seem to contribute to the popularity of the park.




Caption: This graph shows that national parks that are world heritage sites and in the continental United States are the most popular types of parks. The national parks that are not world heritage sites and not in the continental United States are the least popular types of parks.

I wanted to learn what characteristics popular national parks have in common, where popularity is defined as the number of annual visitors to the park. Through exploratory data analysis, it became apparent that being a world heritage site, being located in the continental United States, and being established early were all positively correlated with high popularity. This visualization illustrates these findings well because it shows the average popularity (number of visitors) for each of the nominal variables. The average trend line and identical scaling allow easy comparison between different factors and their associated popularity. It is difficult to show the gradual decline in popularity by the year the park was established, due to the lack of quantitative data and excess of nominal variables. But if looked at closely, it is apparent that the most popular parks were all established in the late 1800s and early 1900s while many of the least popular parks were established in the late 1900s and early 2000s.