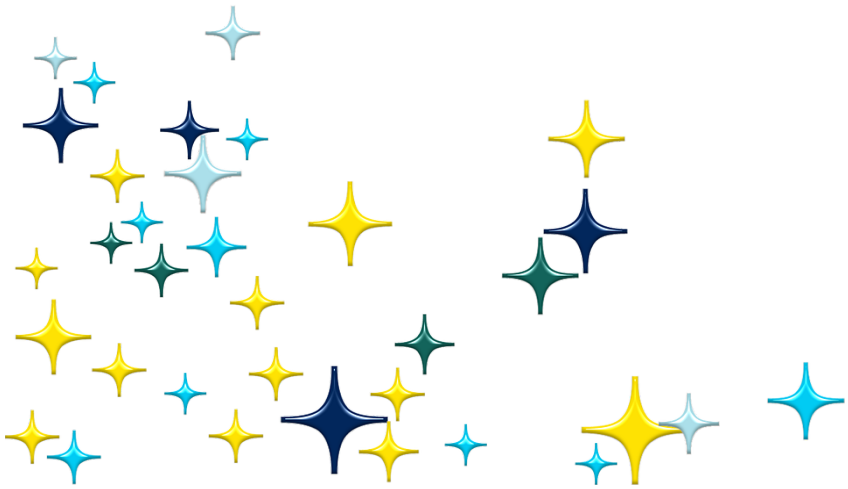
Process Book

The Most Stellar 480x Project

By the out of this world team:

Alexis Caira, Ian Coolidge,

Mary Barsoum, & Sophia Strano



Overview and Motivation:

Our project is a fun and playful representation of UFO abduction reports worldwide. When brainstorming and looking at data sets, we wanted something that was recently updated, came cleaned already (for sake of time and sanity,) and something that people would have fun with. Within working with A4's features of maps, we realized we could try and make some fun visualizations of the city data given in our data set and utilize some of the map features given. Our specific learning objectives for the project was the following:

I. Gain exposure and experience with working with large data sets
II. Practice making different visuals within D3
III. Find, import, and work with JSON mappings
IV. Gain familiarity with Tableau and their public hosting of data sets and embedded visualizations for website

Related Work:

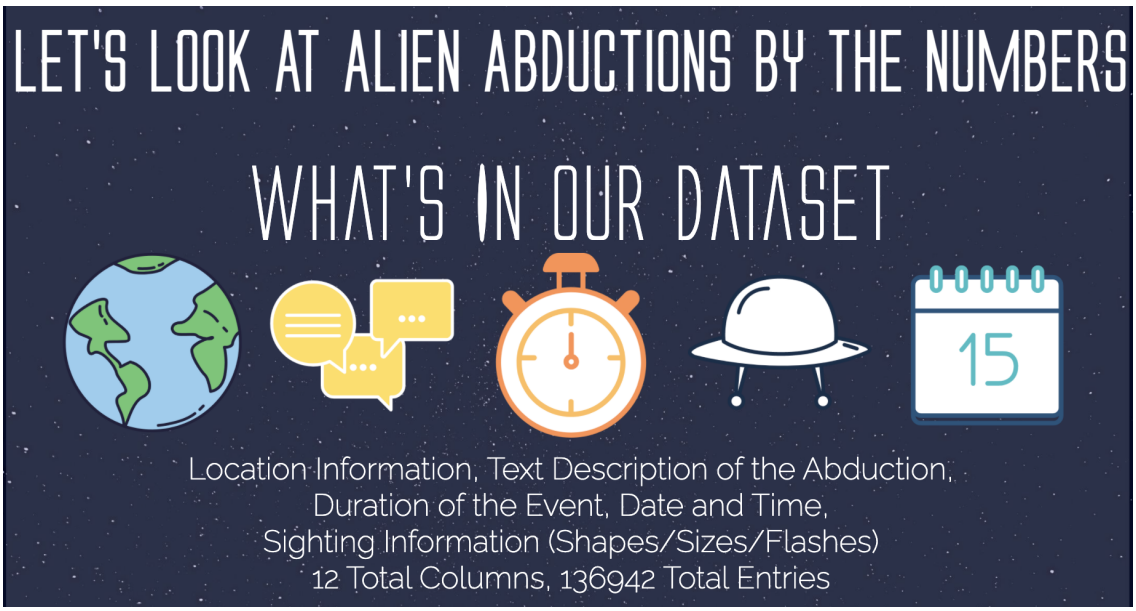
- Anything that inspired you, such as a paper, a website, visualizations we discussed in class
    - We were inspired by the data we explored on https://nuforc.org/ , a website dedicated to the exploration and collection of UFO abduction Data.
    - We largely drew on work related to Assignment 2/A4 in order to create our interactive visualizations

Questions:

The goal of our project was to create an appealing visualization of UFO and alien reports that came from across the United States. We wanted to look for trends like popular states and cities, particularly to note if any one region reported a higher number of abductions, and why that may be. In creating our representation, we also wanted to note the number of reports per capita in a given state, as this can give additional insight as to whether or not a certain state seems to have more citizens reporting alien abductions, rather than the pure number of abductions per state. While analyzing our data, we started to consider how we could expand on our report as it was US-centric, however since this is a niche topic and it’s debatable as to what constitutes as an alien sighting it’s difficult to acquire what could be considered reputable information. Due to inevitable missing data as the abductions are self reported, as well as potential misinformation and population variance from state to state, we mainly focus on the question: How could we most accurately source and represent this data, and is it even possible to verify?

Data:

Using our alien theme, we found a data set on data.world that shared the description of an alien sighting, the location or the sighting, and the shape of the sighting among other things. This was found on the website [data.world,](data.world) and was created by Tim Renner. It was originally scraped from The National UFO Research Center (NUFORC), then cleaned by Renner. We used the cleaned CSV file and used it as is. We narrowed down what data we would display to be only in the U.S. sightings and information rather than the whole world, since the cities listed were mixed and states of the US were often switched to counties, which was tricky to sort though. The data also featured longitude and latitude coordinates, which allowed us to import our data into Tableau and give the city location and the reported sighting. Originally, the team looked at using the duration data given in the set, but this proved unsuccessful since it was not fully formatted and instead roughly pulled directly from the text. This meant that any words people typed in like “approx” or any ranges had that value, which gave almost every entry a unique value. Within our website, we included a high level overview of the dataset before presenting the visualization, so our viewers understood the basics of where the data was coming from, shown in the image below.



Exploratory Data Analysis:

We decided to use bar charts to look at our data along with a map of all the alien sightings. After looking at our data further, we decided that it would be nice to have multiple different bar charts so the user could look at the data set in different ways. When looking through the data, a major drawback to any text analysis we could have done was the formatting of the text. This was harder to format because if we were to do a world cloud or another form of displaying top words, we would need to

1. Run some form of machine learning and data sorting API to get the information and
2. Filter out buffer words such as “the, like, as, etc” that do not provide actual inside information to what people are reporting as common occurrences in an abduction, since it’s just filler words.

Furthermore, we found this to be outside the scope of the class, so we moved on from a text display and pursued more numerical data values. Most people do not like word clouds anyways, so it was not a huge loss.

The original platform data.world includes a chart building tool that imports the original data set and allows you to choose variables as x/y tools, similar to Tableau but more primitive. This took less effort, which made it a good option for preliminary testing of what each visual option would look like (lines, graphs, bar charts, box plots, etc.) This allowed us to look at what the visual could look like if we were to implement it in D3 or another more robust tool without the extra time, which helped us practice the concept of trying a basic visual before putting extra time and effort into actually creating it. We were able to more effectively preview various chart types for our data before taking the time to create them in D3.

When we started putting our data into D3, we noted that there were long wait times for simple computations because of the size of the file being fetched from S3. To help resolve this, we moved the shape data into its own file and took the count of records to make a smaller file to load.

Design Evolution:

We initially considered creating an entirely different set of Visualizations adjacent to club penguin, but decided the UFO data was more interesting to a wide audience and would more easily lend itself to a variety of maps and graphs.

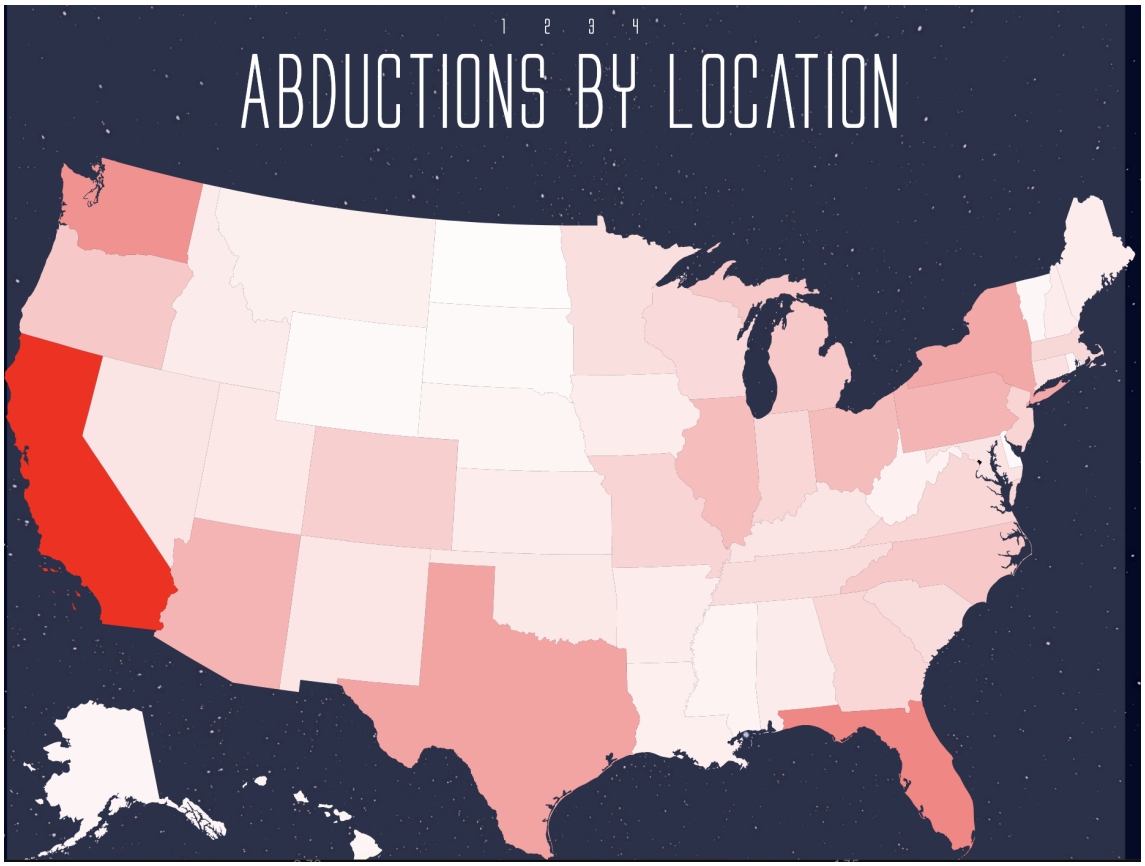
In creating our visualizations, we decided to include a bar chart based on the heavy emphasis placed on the findings of the Cleveland-Mcgill paper throughout the course, which supports the hypothesis that bar charts are exceedingly more likely to be accurately perceived.

When we created the first mapping, initially we looked at using a world map to present data from around the world, but we found that it would not work with the data and would cause more headaches than it was worth. We changed this to a per capita probability map per state, which allowed us to play with the map features and JSON readings.

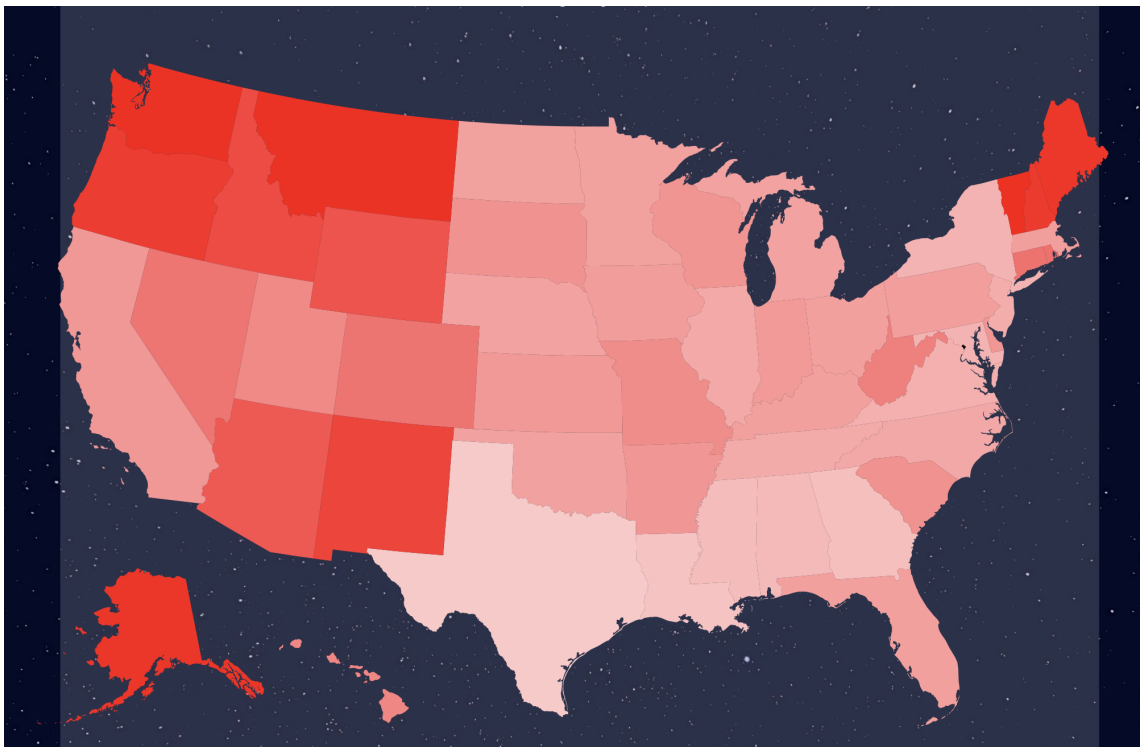
Another design choice that changed during the implementation was the use of a drop down feature to allow users to select graph views. This was changed due to the fact that we wanted more visual information at once. When testing the drop down, it was confusing to have the change on screen and the graphic change on render would be overwhelming, and only could be avoided if we removed them and made each a static choice, or visually have a smoothing tool to make the axis and point change better.

Implementation:

For our visualizations, we choose several types of interactive visuals from two different sources, like we did in A2. Our data hosting was done through Amazon S3 since our data was over 150 MB. Hosting on S3 allowed us to explore implementing CORS requests though the custom Express server we created for the site, enabling us to access outside data from the repository to use. This was useful to practice and understand, since larger data sets outside of the classroom are often too large to be hosted on Github and need an outside storage system. For our actual visualizations, we choose to use D3/JS and Tableau. This allowed us to gain experience with Tableau, which is popular in industry so we thought it important to explore. We also included two D3 charts, to demonstrate basic skills from the class. For our charts in D3 we used a heatmap of the US, imported from a JSON hosted on S3.



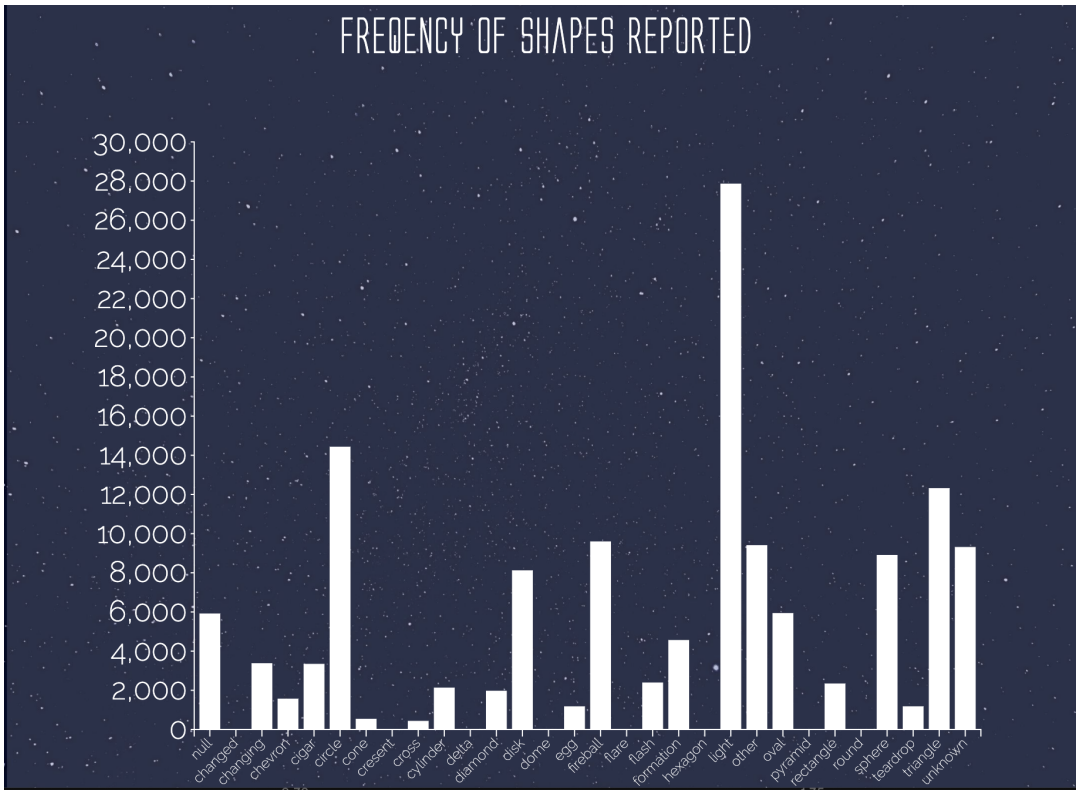
Above is the first map made in D3, which shows the amount of abductions per state via heat map, with the darker the shade of red, the more frequent the reports. We can see here that California and Florida had the most amount and the midwest was the least, but this does not show that there are more people on the west coast and therefore can make more reports, which leads us to the second map.



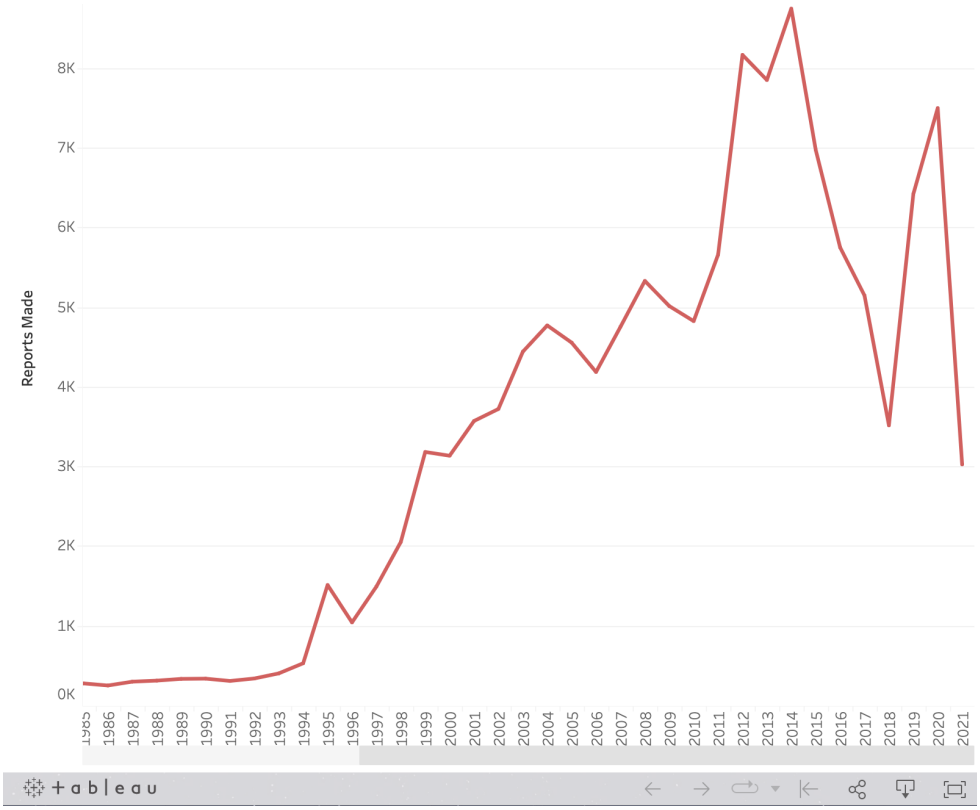
In each map, we also included dynamic tooltips which allow the user to see the number of sightings and the number of sightings per 10,000 people respectively. We thought this was important to allow the user another way to view the data rather than just through the map. In addition, to compute the color of each state based on its sighting count, we used a custom interpolation function that took two colors and a fraction between 0 and 1 as inputs and computed a gradient between the two colors based on this fraction. This way, instead of having to develop a complex color scheme, we simply could plug fractions of the maximum into this interpolation function.

Map two features the likelihood of you getting abducted based on the US census data from 2019, which shows a better visual of per population density instead of raw report totals. For our other D3

visual, we created a bar chart that shows the amount of records and the reported shapes. This was made the same way we created the visuals from A2 in terms of importing data from a locally hosted git file (the smaller CSV) and creating x and y axis markers. This is seen in the photo pictured below.

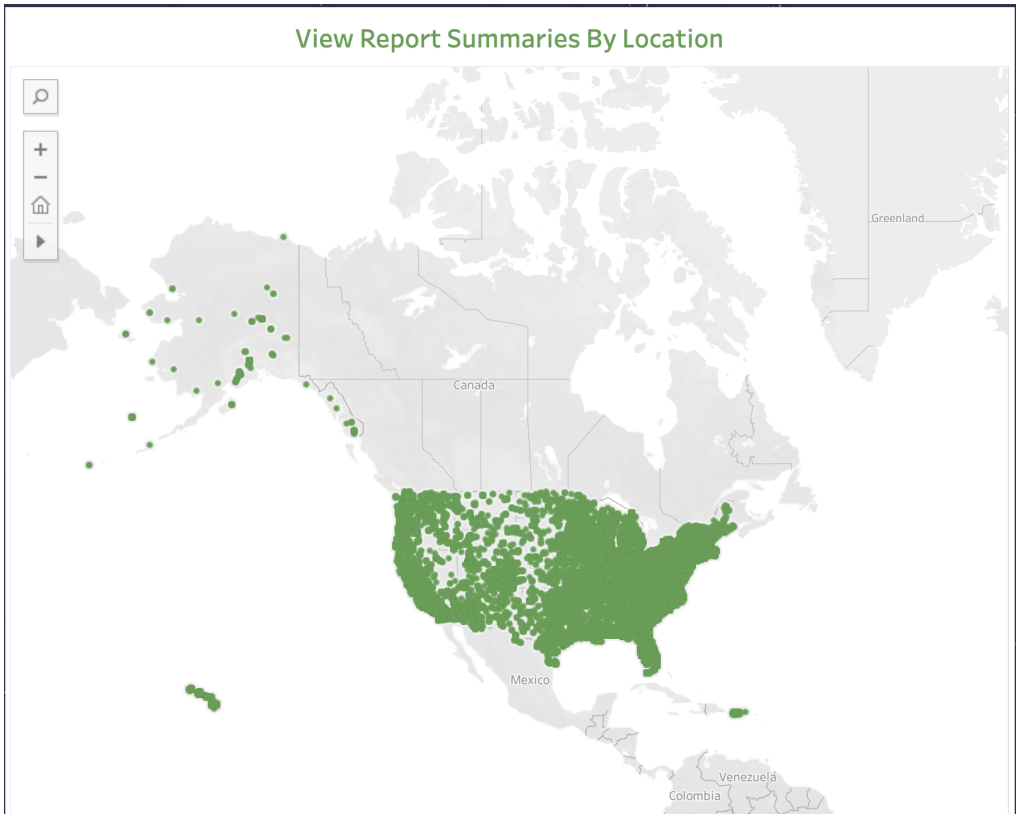


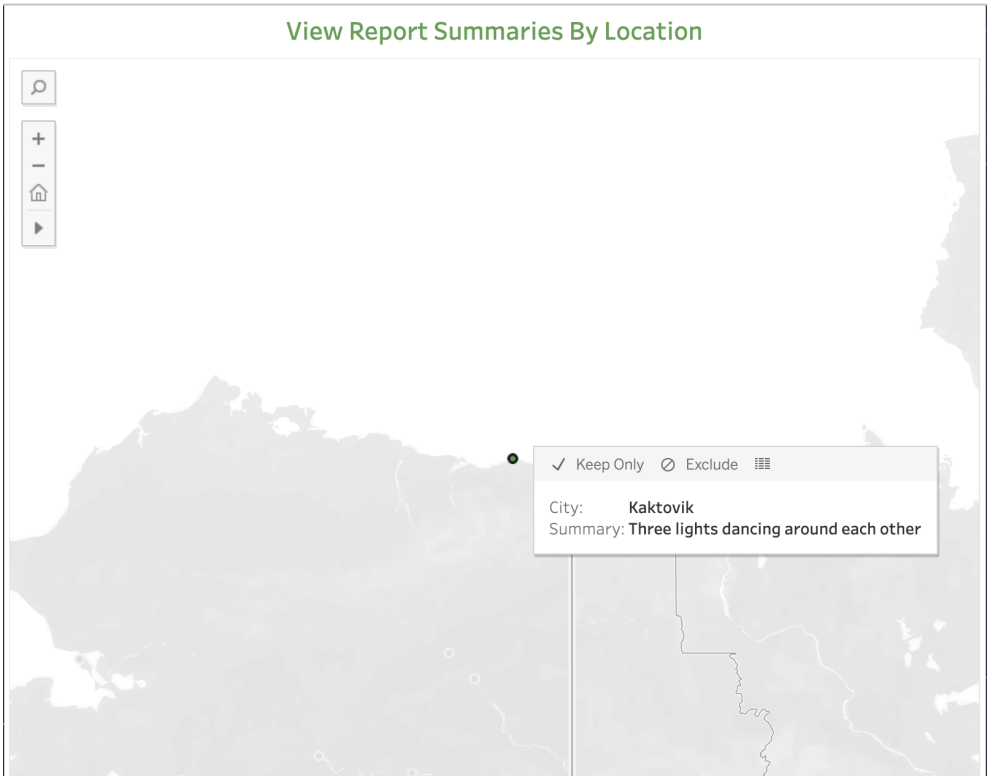
Finally, for our interactive Tableau visualizations, we included two graphs that are publicly hosted on their servers (anyone, with or without an account can view them). The first one is frequency of report over time (years) as seen below.



We decided it would also be interesting to see the evolution in report frequency over the years. It's good to note that the lower report frequency in the early 90's that makes steady gains into the mid 00's is likely largely influenced by the increased use of the internet nationwide. At this point, we heavily question why the number of UFO sightings suddenly dropped off in mid 2018, and if this has something to do with a lull in data connection from the NUFORC website. Additionally, there may be a correlation between the infamously planned "Area 51 Raid" of 2019, which led to much more alien-adjacent buzz online.

For the second, we created a map that allows users to zoom in and out on labels on the city and see the corresponding reports. This was a better use of Tableau, since there was a good amount of information on the tooltip to display and would be a longer processing time for D3, as seen with the amount of time it took to generate a map above. An example of what a user sees on click is pictured below.


View Report Summaries By Location
Greenland
Canada
Mexico
Venezuela
Colombia


View Report Summaries By Location
Keep Only
Exclude
City: Kaktovik
Summary: Three lights dancing around each other

Evaluation:

- Big data sets take a long time to load from S3, which we can see can get expensive to host and access for large companies over time when you are using a larger scale.
- Human data like text is hard, people report things relative to their own experiences so it is harder to catagorize their feels/what they see/subjective things and there is no way to cross check this information.
- Many people think they see UFOs on the west coast or Florida. Surprisingly low number for people in the midwest considering the amount of crop circles out there and wide open farms like classic alien movies love to use.
- Further ideas include
    - World wide map and creating new columns to expand and further sort data (time only category, continent, duration)
    - This would be fastest done with some python scripts or something to develop to find words and values faster.
    - Including photos of the reports included, if that data was to be added in the future. It would be a cool tooltip hover to see photos from the person reporting it so you could judge for yourself what it looks like.