Matthew Younce

Deliverable #1: Milestone 3

2 April 2021

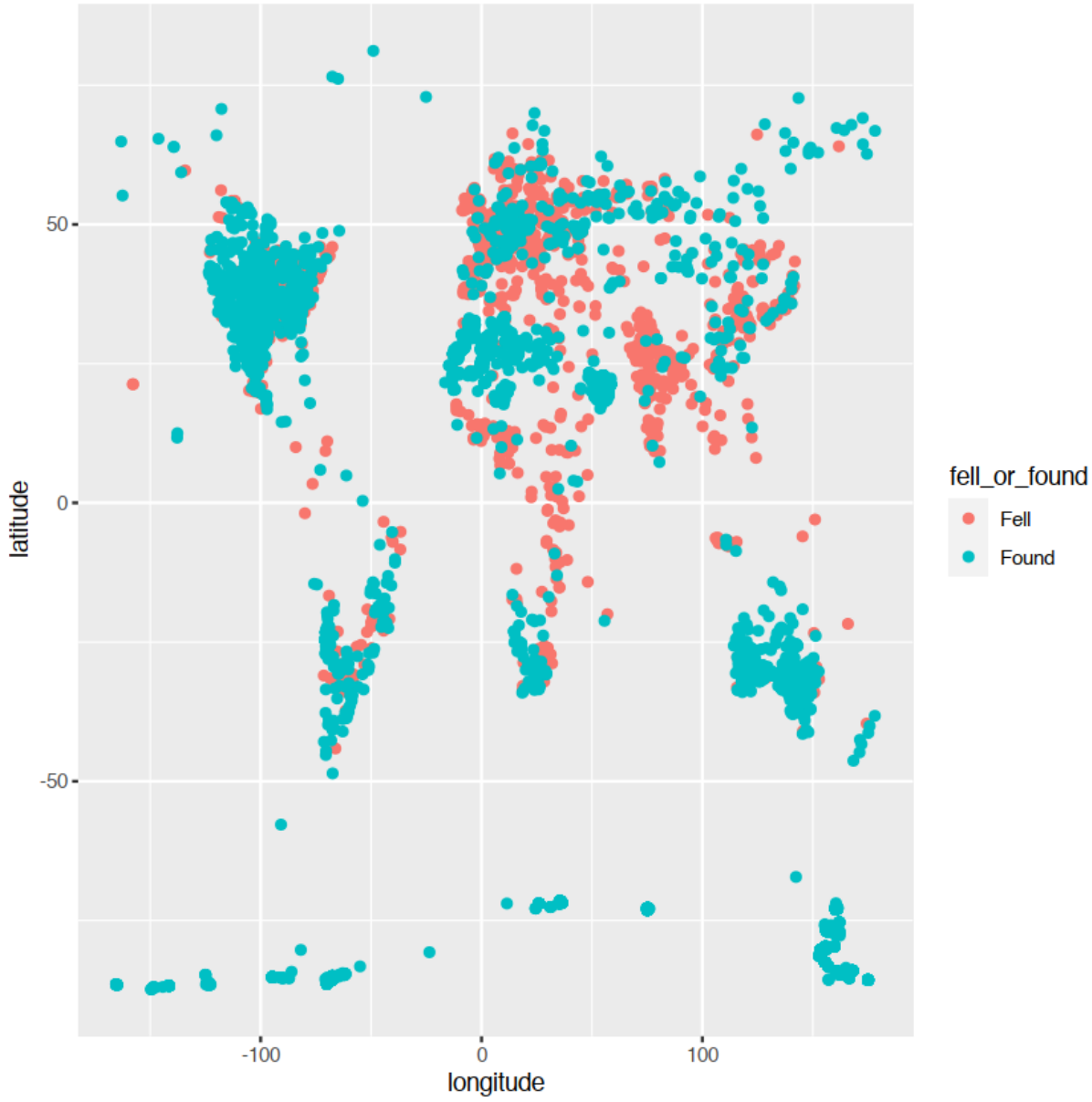
This study aims to investigate a dataset of meteorites that have been located and identified throughout human history. The initial data set included 45,716 individual meteorite-like objects. For the initial exploratory analysis, please view my Jupyter Notebook at:

https://github.com/mwy912/capstone/blob/main/Data_analysis_and_cleanup_R.ipynb

I imported the database and began working to clean the data. My first step was to convert the year column into an actual year, as it was recorded as a timestamp at 12 am on January 1 of the year that it was collected. Then I renamed a few columns to be more informative to what they contained. One attribute of interest was nametype, which could either be “valid” or “relict” — the latter indicating that the object was later discovered not to be a meteorite. These 75 relicts the first items I removed from the dataset.

The next area I focused on was the latitude and longitude. I noticed there was a meteorite listed with a longitude of 354 degrees. This made little sense until I investigated and found this was a rock found on the surface of Mars, and was only there as a comparison to meteorites of Martian origin. I removed that one datapoint. I followed this up with the removal of 6,214 datapoints that were geolocated at (0,0) and would be of no use to trying to determine the location these objects impacted or were found. This left me with 31,892 datapoints.

Continuing to browse the data, I noticed that 4,761 data points all were located at (-71.5,35.67) which seems unlikely. After noting similar areas with likely bad geocoordinates, I decided to focus only on meteorites found between the Arctic Circle and Antarctic Circle. This removed an additional 22,111 meteorites, and left me with 9,781 meteorites to use in my analysis. Here is a basic graphic of the locations of those 9,781 meteorites:



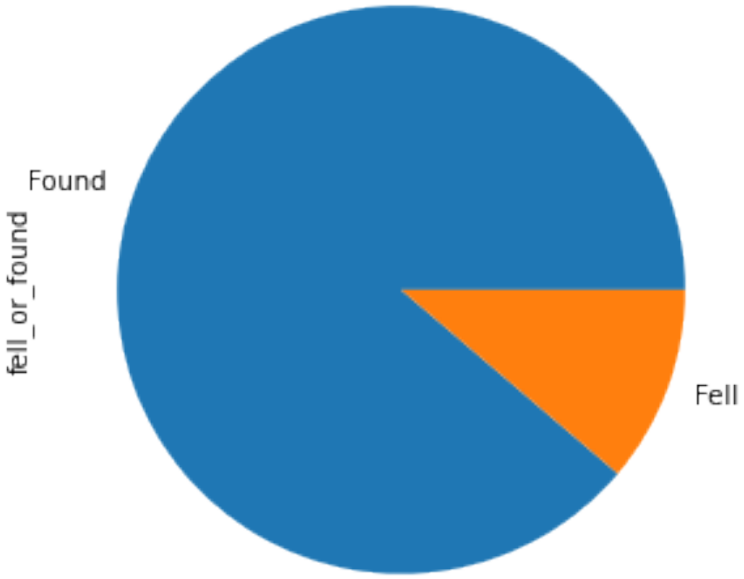
I next wanted to fill in any NAs. There were 67 meteorites without a mass given. I found the median (because of several large meteorite finds, the mean was very skewed) and filled in that value for the masses. For the 149 events without a year, I took the mean year of the dataset.

My next task was to begin a preliminary statistical analysis.

https://github.com/mwy912/capstone/blob/main/Prelim_statistics_py.ipynb

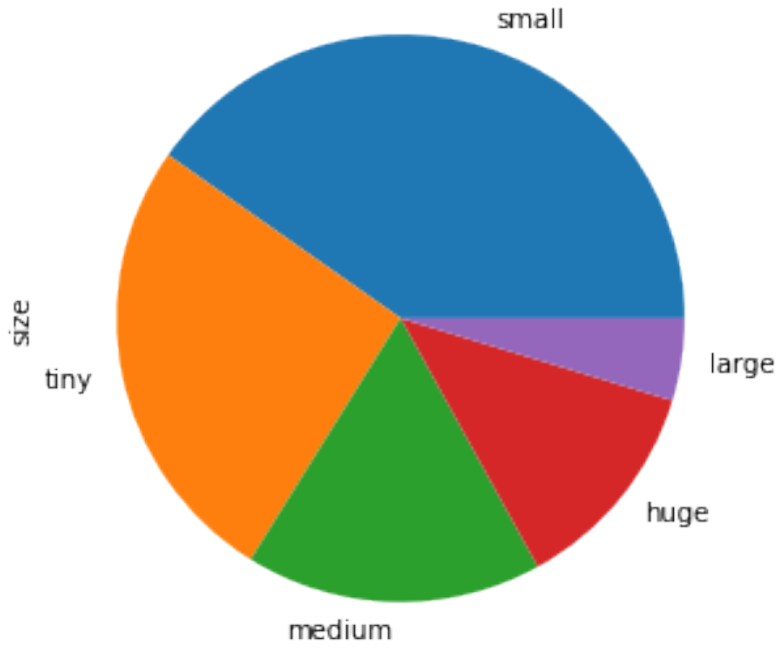
I looked at just getting counts of meteorites. First dividing up based on found or fell, I got:

8,686 found, 1095 fell

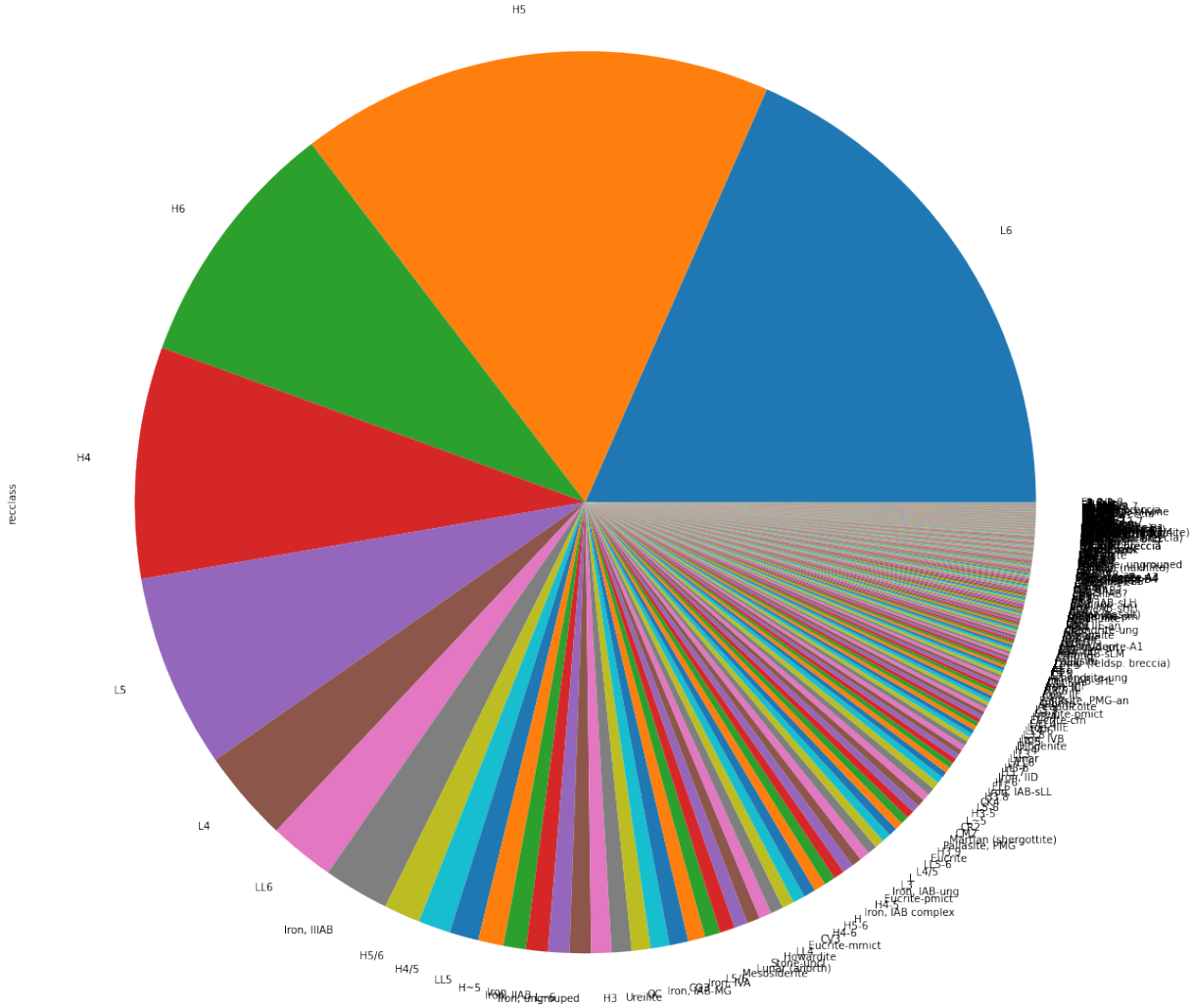


I next divided the meteorites into 5 categories based on size: Tiny (< 100g), small (100g - 1 kg), medium (1 kg - 5 kg), large (5 kg - 10 kg) and huge (greater that 10 kg). This was the breakdown:

small	3937
tiny	2533
medium	1651
huge	1202
large	458



Finally I looked at the descriptions in the “recclass” field. As you can see in the link, there are over 300 different recclasses, so I won’t post the entire count or table here. Here is the pie chart:



In the Jupyter Notebook at the link, you can see a horizontal bar chart as well as the full table of counts for this graph.

As an assessment of the suitability of machine learning and statistical modeling, I believe that, in order to perform this work, one area I need to tackle is this recclass category. These are far too many categories, and many of them are very similar. Once I can find a better classification system, I will be able to begin modeling and machine learning to begin predicting how often certain types of meteorites will impact and where they will fall.