Exploratory Analysis and Visualization

CS 5890

Jacob Butterfield, Nelson Miller, Dale Hulse

Dataset Overview

Our data set is the NASA collection of near-earth objects (NEOs). It contains about 300,000 objects; including their orbit, close approach date, velocity and magnitude. The dates in the data range from 1899 to 2201. Our goal is to analyze the data in order to find historical trends as well as future classifications regarding these objects. Finding optimal times to mine asteroids is also one application of our research.

Summary and Descriptive Statistics

The following tables describe the basic statistics of each class of asteroid. Included are the percentage of total asteroids of each type, as well as the minimum, maximum, average, and standard deviation of the entire dataset with respect to asteroid velocity, close approach distance and magnitude.

Class	Percentage of Total
AMO*	19.26%
APO*	53.31%
ATE*	26.70%
ETc*	0.01%
HTC*	0.00%
IEO*	0.53%
JFC*	0.01%
JFc*	0.18%
Grand Total	100.00%

Relative Velocity (KM/s)

Class	Min. V-Relative	Avg. V-Relative	Median V-Relative	Std. dev. of V-Relative	Max. V-Relative
AMO*	0.06	10.6274300418141	9.77	5.87718627083338	63.27
APO*	0.52	17.0380355340816	15.98	8.0224144954534	81.05
ATE*	0.14	15.6033625195389	14.85	6.54880331391627	49.77
ETc*	17.94	29.9946428571428	32.73	7.56975144273324	42.79
HTC*	18.44	48.6646666666667	49.94	18.6000694571156	78.5
IEO*	2.31	15.6608188585608	15.68	4.90305745318831	33.93
JFC*	29.95	40.9511428571429	38.48	8.40108493393722	57.94
JFc*	2.73	18.5623333333333	16.925	8.35735993280068	58.32
Grand Total	0.06	15.4211594618745	14.4	7.66071111722084	81.05

V-magnitude

Class	Min. H	Avg. H	Median H	Std. dev. of H	Max. H
AMO*	9.4	20.3218042694377	20.1	2.22554971724919	28.7
APO*	12.4	20.7212629583209	20.4	2.60641784190917	33.2
ATE*	14.5	21.8020161684244	21.4	2.74106440777886	32.1
ETc*					
HTC*					
IEO*	16.3	18.886786600496	19.8	1.39947861058908	25
JFC*					
JFc*					
Grand Total	9.4	20.9236542558087	20.5	2.63161846434044	33.2

Minimum Distance

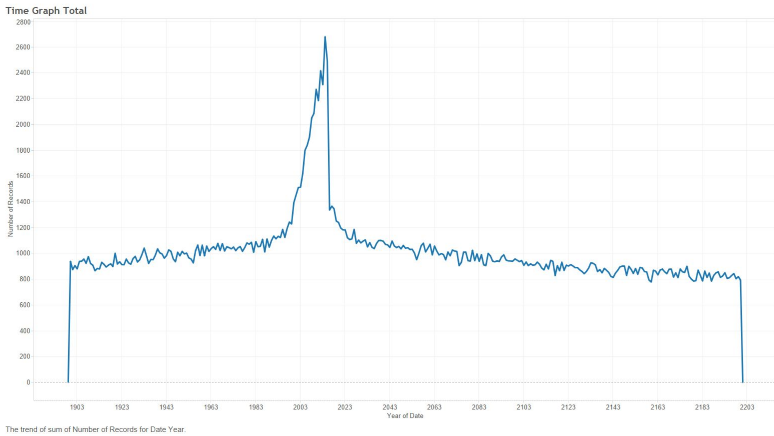
Class	Min. minimum-AU-distance	Avg. minimum-AU-distance	Median minimum-AU-distance	Std. dev. of minimum-AU-distance	Max. minimum-AU-distance
AMO*	4e-05	0.31283164395389	0.3244	0.120215432846011	0.4999
APO*	0	0.27801191864469	0.2896	0.136232250111328	0.5
ATE*	3e-05	0.280934576128366	0.2919	0.134586548246662	0.5
ETc*	0.1608	0.277092857142857	0.25755	0.0874758503333028	0.4778
HTC*	0.0916	0.31968	0.3772	0.134568767337957	0.4876
IEO*	0.0006	0.298618052109181	0.3203	0.131618054823354	0.4997
JFC*	0.0096	0.323437142857143	0.3566	0.128049310551729	0.4869
JFc*	0.0001	0.24573537037037	0.2402	0.134399133881786	0.4987
Grand Total	0	0.285557366195618	0.2983	0.133504153623325	0.5

Data Cleaning and Preprocessing

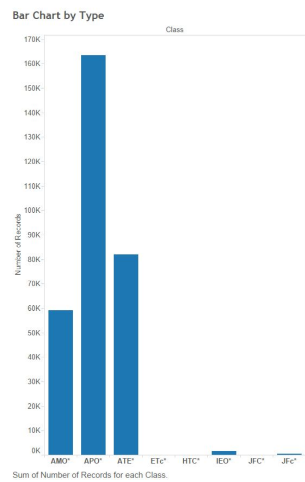
Data cleaning and preprocessing was a very difficult process. First of all, the data was only available via NASA’s NEO web page, and there were no options to download the dataset. Therefore, we had to download the webpage and read the data from the embedded table. Unfortunately, the webpage was far too large to load in a browser, therefore we had to download it using a “wget” command from the command line. The resulting download was about 300 MB in size. Once we had the webpage, we ran a Python script to parse out the styling and reduce the page size. We wrote a custom HTML parser in Java to convert the main table element to a csv file. Finally, we split up columns that had multiple unit types (some had distances in LD and AU), and removed unnecessary error margins in the dates using Pandas. We wrote the cleaned dataset to a final csv that is less than 30 MB in size.

Insights

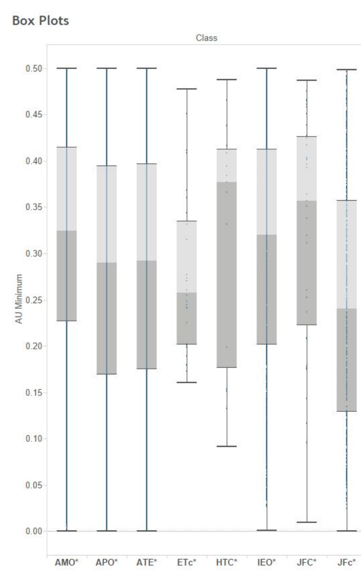
When performing our initial analysis, we discovered an interesting anomaly in the data. The chart below shows the total number of NEOs by year. A clear spike in total asteroids is seen between the year 2000 and 2015. We took this to mean that there are many more NEOs that are not currently known to or tracked by NASA. We predict that as each year passes, more and more of these objects will be detected and subsequently added to the dataset.



Another insight that we discovered through our analysis was how many objects there were of each type. The type of an NEO is determined by its orbit with respect to the earth; a description of these types can be found here: http://neo.jpl.nasa.gov/neo/groups.html. The bar chart indicates the total number of each type of near earth object. Three classes, Apollo, Aten and Amor, are the most common types of asteroids by far, with Apollo objects being over half of all objects in the dataset. All other classifications of asteroids are minimal.

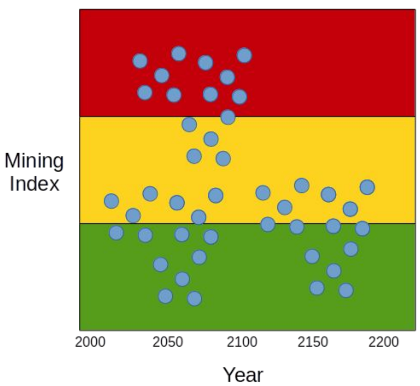


The last of our insights involves the distribution of the data. We created a series of boxplots comparing the distribution of close approach distance by asteroid classification. It can be seen that distribution is similar for most of the object types, especially with regards to minimum, maximum, and the positioning of the quartiles. The most common object classes, Apollo, Aten and Amor, are almost identical with their distributions. This means that the minimum distances of nearly

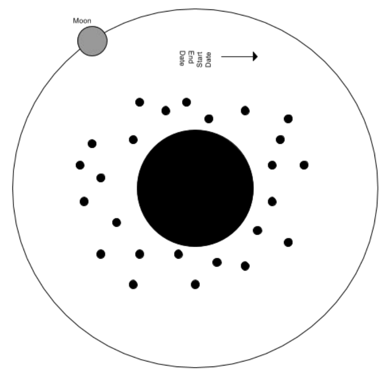


all asteroids are distributed similarly. Therefore, no single classification of asteroid is more likely to pass closer to earth.

Initial Screenshots and Sketch of Final Product



We are still generating our final product and do not have any screenshots currently. However, we do have a few sketches of what we will accomplish when it is complete. For example, the image to the left is a sketch of scatter plot of all of the asteroids over time compared to their "mining index." The mining index will be a calculated value that determines how accessible an asteroid is for resource extraction. Objects in the green region are the most accessible, while objects the yellow and red regions are not as suitable for asteroid mining.



A second chart that we are considering is a radial graph denoting the closest approaches as the "y" axis (distance from base line) against the dates they take place on as our "x" axis. Our "x" axis will start from the top and progress clockwise, as seen in the accompanied example.

Next Steps

This exploratory analysis has helped us to understand the data better and come up with some good sketches for what we want our final product too look like. For our next steps, we will need to prepare and format our data so it can work with D3.js. We plan on using D3.js to generate the final product of the sketches drawn above. We also plan on building a classification model to predict NEO types based off of distance, velocity, and size.