Exploring Potentially Hazardous Near-Earth Objects

Data Science Project

Kiara A. Richardson

kar15@rice.edu

Department of Computational Applied Mathematics and Operations Research

William Marsh Rice University

RCEL 506 – Applied Statistics and Data Science for Engineering Leaders

Professor Edgar Avalos Gauna

Spring 2023

Table of Contents

Project Introduction

Project Background

Agencies like NASA track asteroids, not only because they might pose a threat to humanity by colliding with Earth, but because they can provide us with information about the history of our solar system, and even be useful for mining raw materials in space (Asteroid Watch, 2023). A near-Earth object (NEO) is any small Solar System body whose orbit brings it into proximity with Earth's orbital plane. By convention, a Solar System body is a NEO if its closest approach to the Sun is less than 1.3 astronomical units (AU). If a NEO's orbit crosses the Earth's orbit, and the object is larger than 140 meters (460 ft) across, it is considered a potentially hazardous object (PHO) (Asteroid Watch, 2023). Most known PHOs and NEOs are asteroids, but a small fraction are comets. There are more asteroid PHOs and NEOs because Earth is close to the asteroid belt, where the majority of the asteroids in our solar system preside. Comets are different and act like miniature planets with their own orbits around the Sun. Having them cross Earth's orbital plane is scarce because their elliptical paths are so large, they can rarely be seen near the inner planets.

When studying the properties of asteroids and comets, it's imperative to understand the Tisserand Parameter, also called the Jupiter Tisserand Invariant. It is a measure of the orbital motion of a comet or asteroid with respect to a large perturbing body, usually Jupiter when talking about our solar system. For a small body with semi-major axis **a**, orbital eccentricity **e**, and orbital inclination **i**, relative to the orbit of a perturbing larger body with semimajor axis $\boldsymbol{a_J}$, the parameter is defined as follows:

$$T_J = \frac{a_J}{a} + 2\cos(i)\sqrt{\frac{a}{a_J}(1-e^2)}$$

The Tisserand parameter considers the semimajor axis, eccentricity, and inclination of the small body's orbit, and remains broadly constant during the small body's lifetime. It is a form of the restricted three-body problem and is useful in identifying small bodies observed before and after encounters with planets, as its numerical value remains largely unchanged by the encounter (Tisserand Criterion, 2023). The Tisserand parameter can also be used to classify planet-crossing bodies. Jupiter-family comets have Tisserand parameters between 2.0 and 3.0, and Halley-type and long-period comets have values less than 2.0. Asteroids generally have values greater than 3.0. However, there are both some periodic comets whose orbits have evolved to values greater than 3 and some asteroids with values less than 3 (Tisserand Criterion, 2023). Many of the latter have been shown to be likely extinct or inactive comet nuclei. The parameter is named after the French astronomer François Félix Tisserand (1845–1896) (François Felix Tisserand, 2023).

To summarize,

$$T_J < 3 \rightarrow comets$$

$$T_J > 3 \rightarrow asteroids$$

$$T_J < 2 \rightarrow Halley\ comets$$

$$2 < T_J < 3 \rightarrow Jupiter\ comets$$

For this project, only general comets and asteroids will be studied.

Project Topic

This project will study:

1. What Near-Earth Objects (NEOs) are asteroids, and which are comets
2. What comet and asteroid Near-Earth Objects (NEOs) are Potentially Hazardous Objects (PHOs)
3. What Potentially Hazardous Objects (PHOs) look like graphically
4. What seasons does Earth experience the most Potentially Hazardous Objects (PHO) interactions

Project Hypothesis

This project will hypothesize:

1. There are more asteroid Near-Earth Objects (NEOs)
2. There are more asteroid Near-Earth Objects (NEO) that are also Potentially Hazardous Objects (PHOs)
3. The Semi-Major axis of a Potentially Hazardous Object (PHO) should decrease as it approaches the Sun, and its Orbital Speed (or Mean Motion) should increase as it approaches the Sun
4. Earth will experience the most Potentially Hazardous Object (PHO) interactions in the Fall months

Literature Review

For the Literature Review, three relevant sources were found to base this project off. Each source is comprised of a piece of this project. There are not many resources available that have the same results as this project, but there are relative sources that encompass parts of this project. These are:

1. Origin and Evolution of Near Earth Asteroids by A. Morbidelli - Cambridge University
    a. https://www.cambridge.org/core/journals/international-astronomical-union-colloquium/article/origin-and-evolution-of-near-earth-asteroids/92D813DF88D343249AB4B7B2ED4CBC35
2. Comets in the near-Earth object population by Francesca DeMeo and Richard P. Binzel - Massachusetts Institute of Technology
    a. https://www.sciencedirect.com/science/article/pii/S0019103507005258?casa_token=FLxJBL2mTwAAAAAA:GRiTLAYj2H67GUUZJbk0A60H6-sWyVXoboqT8_kVByVgmmFZi1azOJSAZaRVIslWgen4jcxG4A

3. Potential impact detection for Near-Earth asteroids: the case of 99942 Apophis (2004 MN4) - Cambridge University
    a. https://www.cambridge.org/core/journals/proceedings-of-the-international-astronomical-union/article/potential-impact-detection-for-nearearth-asteroids-the-case-of-99942-apophis-2004-mn4/C22E1705065ED5E02D8CE6F89D68D3B1

Statistical Analysis

Data Importing

The secondary dataset was found on the Kaggle website:

- https://www.kaggle.com/datasets/shrutimehta/nasa-asteroids-classification?resource=download&select=nasa.csv

After more research, the primary dataset was found to be located at NASA's Jet Propulsion Laboratory (JPL) Near-Earth Object (NEO) library:

- http://neo.jpl.nasa.gov/

Using the library is very complicated and there is little instruction on how to navigate through it. However, there is a plethora of primary datasets available there (including this one). The data collected there is much more detailed and expands over a longer amount of time.

The secondary dataset used for this project covers the years 1995-2016. The dataframe size is 4687 rows × 40 columns and the data types for the columns were bool(1), float64(30), int64(5), object(4). The memory usage is 1.4+ Megabytes (MB). After importing the necessary modules into Python, the data is imported into the Jupyter Notebook and is seen below:

	Neo Reference ID	Name	Absolute Magnitude	Est Dia in KM(min)	Est Dia in KM(max)	Est Dia in M(min)	Est Dia in M(max)	Est Dia in Miles(min)	Est Dia in Miles(max)	Est Dia in Feet(min)	Est Dia in Feet(max)	Close Approach Date	Epoch Date Close Approach
0	3703080	3703080	21.600	0.127220	0.284472	127.219879	284.472297	0.079051	0.176763	417.388066	933.308089	1995-01-01	788947200000
1	3723955	3723955	21.300	0.146068	0.326618	146.067964	326.617897	0.090762	0.202951	479.225620	1071.581063	1995-01-01	788947200000
2	2446862	2446862	20.300	0.231502	0.517654	231.502122	517.654482	0.143849	0.321655	759.521423	1698.341531	1995-01-08	789552000000
3	3092506	3092506	27.400	0.008801	0.019681	8.801465	19.680675	0.005469	0.012229	28.876199	64.569144	1995-01-15	790156800000
4	3514799	3514799	21.600	0.127220	0.284472	127.219879	284.472297	0.079051	0.176763	417.388066	933.308089	1995-01-15	790156800000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4682	3759007	3759007	23.900	0.044112	0.098637	44.111820	98.637028	0.027410	0.061290	144.723824	323.612307	2016-09-08	1473318000000
4683	3759295	3759295	28.200	0.006089	0.013616	6.089126	13.615700	0.003784	0.008460	19.977449	44.670934	2016-09-08	1473318000000
4684	3759714	3759714	22.700	0.076658	0.171412	76.657557	171.411509	0.047633	0.106510	251.501180	562.373736	2016-09-08	1473318000000
4685	3759720	3759720	21.800	0.116026	0.259442	116.025908	259.441818	0.072095	0.161210	380.662441	851.187094	2016-09-08	1473318000000
4686	3772978	3772978	19.109	0.400641	0.895860	400.640618	895.859655	0.248946	0.556661	1314.437764	2939.172192	2016-09-08	1473318000000

After review, a large amount of columns was deemed unnecessary, so the data needed to be cleaned.

Data Cleaning

The data cleaning process was relatively short. NASA does a very good job of excluding NaN values in their datasets. Therefore, the only columns needed to be removed instead of rows. The columns removed were unnecessary for analysis. This process left the project with a new dataframe with 4687 rows × 18 columns. The data types for the columns were: bool(1), float64(16), object(1). The memory usage is 627.2+ KB. The new dataframe is as seen below:

	Absolute Magnitude	Est Dia in KM(max)	Close Approach Date	Relative Velocity km per hr	Miss Dist. (Astronomical)	Jupiter Tisserand Invariant	Epoch Osculation	Eccentricity	Semi Major Axis	Inclination	Asc Node Longitude	Orbital Period
0	21.600	0.284472	1995-01-01	22017.003799	0.419483	4.634	2458000.5	0.425549	1.407011	6.025981	314.373913	609.599786
1	21.300	0.326618	1995-01-01	65210.346095	0.383014	5.457	2458000.5	0.351674	1.107776	28.412996	136.717242	425.869294
2	20.300	0.517654	1995-01-08	27326.560182	0.050956	4.557	2458000.5	0.348248	1.458824	4.237961	259.475979	643.580228
3	27.400	0.019681	1995-01-15	40225.948191	0.285322	5.093	2458000.5	0.216578	1.255903	7.905894	57.173266	514.082140
4	21.600	0.284472	1995-01-15	35426.991794	0.407832	5.154	2458000.5	0.210448	1.225615	16.793382	84.629307	495.597821
...	...	...	...	...	...	...	...	...	...	...	...	...
4682	23.900	0.098637	2016-09-08	79755.354273	0.041361	5.156	2457637.5	0.361512	1.161429	39.880491	164.183305	457.179984
4683	28.200	0.013616	2016-09-08	11610.539577	0.006469	5.742	2458000.5	0.073200	1.075134	5.360249	345.225230	407.185767
4684	22.700	0.171412	2016-09-08	25889.910626	0.061009	4.410	2458000.5	0.368055	1.528234	4.405467	37.026468	690.054279
4685	21.800	0.259442	2016-09-08	40867.522309	0.260760	4.477	2458000.5	0.202565	1.486600	21.080244	163.802910	662.048343
4686	19.109	0.895860	2016-09-08	129408.666253	0.462372	4.108	2458000.5	0.405642	1.474045	53.574923	187.642183	653.679098

Data Description

Using Python to take descriptions of the data was the next step. This ensures that the "data about the data" is known before results are collected. Describe, median, and variance were used to understand this.

Using the calling df.describe(), Python outputs:

	Absolute Magnitude	Est Dia in KM(max)	Relative Velocity km per hr	Miss Dist. (Astronomical)	Jupiter Tisserand Invariant	Epoch Osculation	Eccentricity	Semi Major Axis	Inclination	Asc Node Longitude	Orbital Period
count	4687.000000	4687.000000	4687.000000	4687.000000	4687.000000	4.687000e+03	4687.000000	4687.000000	4687.000000	4687.000000	4687.000000
mean	22.267865	0.457509	50294.919829	0.256778	5.056111	2.457724e+06	0.382569	1.400264	13.373844	172.157275	635.582076
std	2.890972	0.826391	26255.601377	0.145798	1.237818	9.202975e+02	0.180444	0.524154	10.936227	103.276777	370.954727
min	11.160000	0.002260	1207.814804	0.000178	2.196000	2.450164e+06	0.007522	0.615920	0.014513	0.001941	176.557161
25%	20.100000	0.074824	30358.313370	0.133420	4.049500	2.458000e+06	0.240858	1.000635	4.962341	83.081208	365.605031
50%	21.900000	0.247765	46504.401181	0.265029	5.071000	2.458000e+06	0.372450	1.240981	10.311836	172.625393	504.947292
75%	24.500000	0.567597	65079.535405	0.384154	6.019000	2.458000e+06	0.512411	1.678364	19.511681	255.026909	794.195972
max	32.100000	34.836938	160681.487851	0.499884	9.025000	2.458020e+06	0.960261	5.072008	75.406667	359.905890	4172.231343

Using the calling df.median() and df.var(), Python outputs:

```
The Median of the cleaned dataset is:
 Absolute Magnitude            2.190000e+01
Est Dia in KM(max)             2.477650e-01
Relative Velocity km per hr    4.650440e+04
Miss Dist.(Astronomical)       2.650286e-01
Jupiter Tisserand Invariant    5.071000e+00
Epoch Osculation               2.458000e+06
Eccentricity                   3.724502e-01
Semi Major Axis                1.240981e+00
Inclination                    1.031184e+01
Asc Node Longitude             1.726254e+02
Orbital Period                 5.049473e+02
Perihelion Distance            8.331526e-01
Perihelion Arg                 1.897616e+02
Aphelion Dist                  1.618195e+00
Mean Anomaly                   1.857189e+02
Mean Motion                    7.129457e-01
Hazardous                      0.000000e+00
dtype: float64
```

```
The Variance of the cleaned dataset is:
 Absolute Magnitude            8.357719e+00
Est Dia in KM(max)             6.829225e-01
Relative Velocity km per hr    6.893566e+08
Miss Dist.(Astronomical)       2.125711e-02
Jupiter Tisserand Invariant    1.532194e+00
Epoch Osculation               8.469474e+05
Eccentricity                   3.255996e-02
Semi Major Axis                2.747373e-01
Inclination                    1.196011e+02
Asc Node Longitude             1.066609e+04
Orbital Period                 1.376074e+05
Perihelion Distance            5.859259e-02
Perihelion Arg                 1.071495e+04
Aphelion Dist                  9.053893e-01
Mean Anomaly                   1.155660e+04
Mean Motion                    1.173933e-01
Hazardous                      1.351647e-01
dtype: float64
```

Data Correlation

Pandas Profiling was also considered when understanding all the variables. Pandas Profiling shows an overview of the dataset and gives information on the statistics and variable types. It was very helpful for the project. Pandas Profiling covers multiple pages, too many to show in this document, but the full report can be found in the Jupyter Notebook submitted with this.

Dataset statistics

Number of variables	18
Number of observations	4687
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	627.2 KiB
Average record size in memory	137.0 B

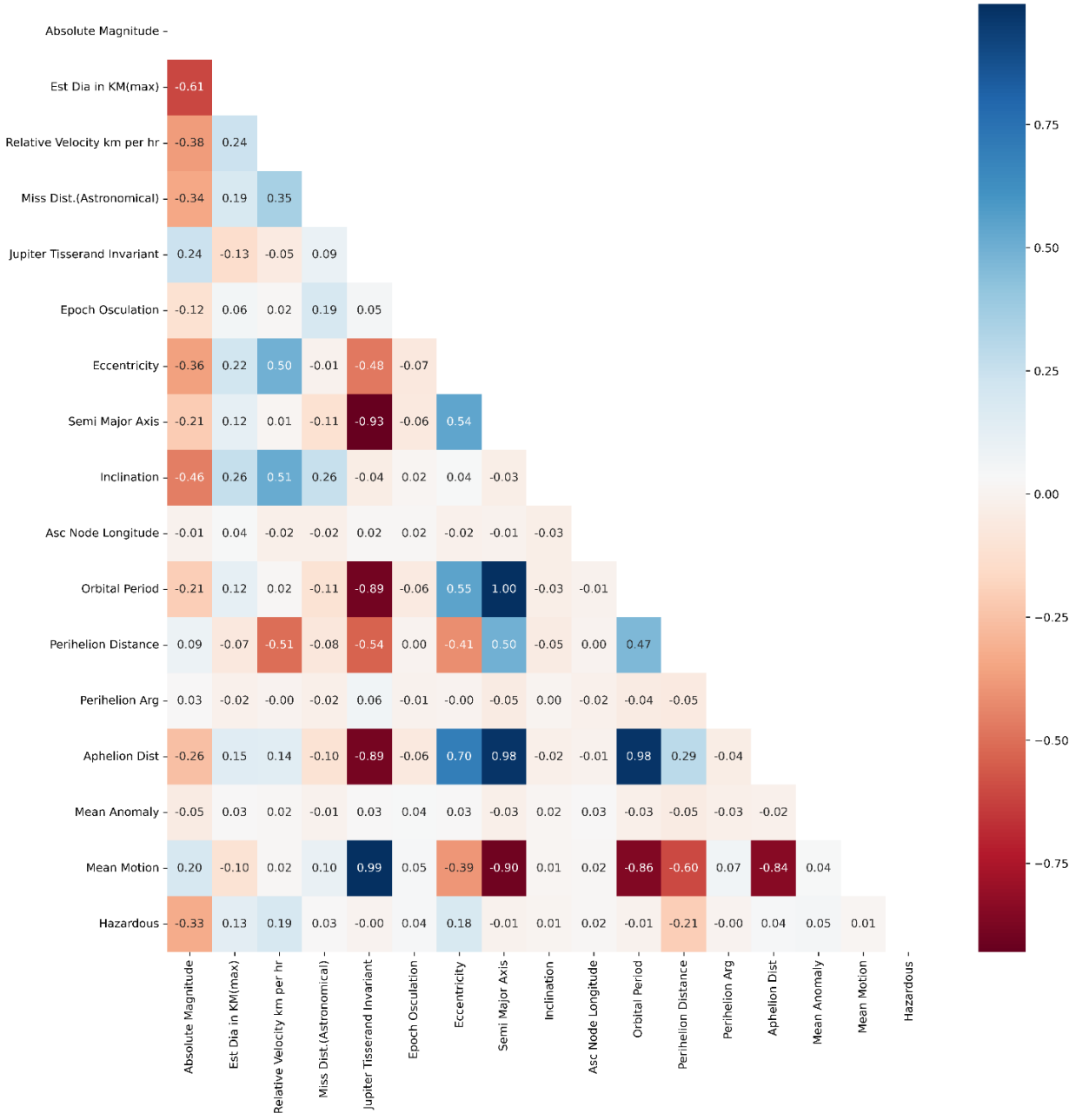
Variable types

Numeric	16
Categorical	1
Boolean	1

The correlation of the remaining variables was checked next. This was to see if any variables were highly correlated or colinear. The following about correlation was kept in mind:

- weak correlation - below ± 0.25
- moderate correlation - between ± 0.25 to ± 0.75
- strong correlation - between ± 0.75 to ± 1

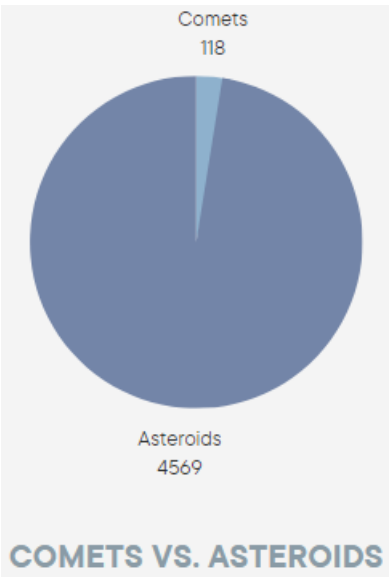
The correlation matrix is showed below:

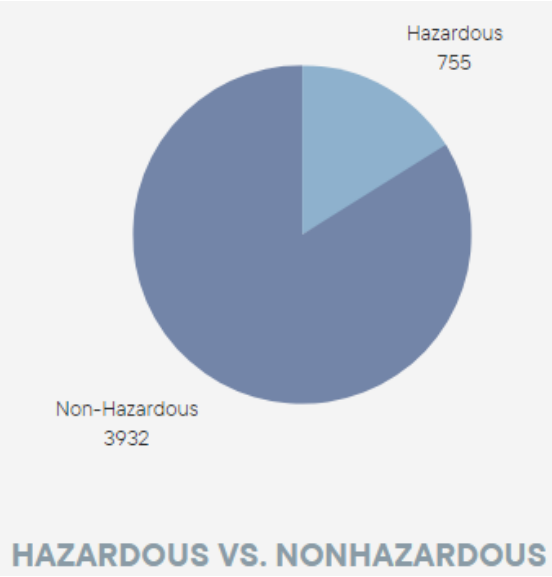


Results

Topic 1

Topic 1 questions which NEOs are asteroids, and which are comets. This is found by creating sub-dataframes from the cleaned version. The dataframe is split into comets vs. asteroids and hazardous objects vs. non-hazardous objects, leaving us with 4 sub-dataframes. Using the Tisserand column, the dataframe is split by categorizing values greater than 3 for asteroids and less than 3 for comets. Then, by categorizing the Hazardous column, the next dataframe is grouped by rows that say True for Hazardous and False for Non-Hazardous. These are the results:



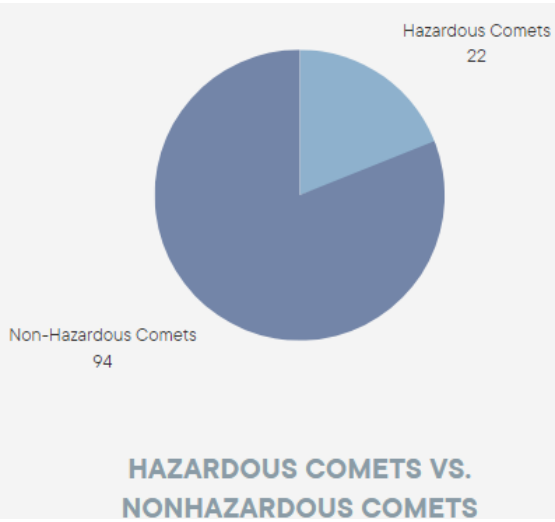


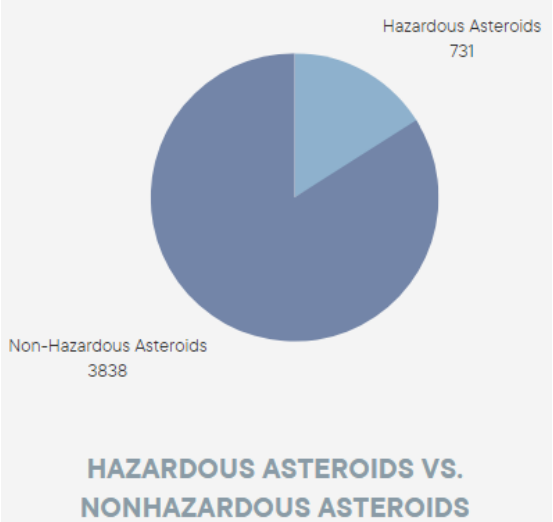
At this point, the first hypothesis was proven correctly. There were more asteroids in this dataset than comets.

Here, there is a large amount of asteroids in the dataset and there is a large amount of Non-Hazardous objects in the dataset. However, at this point, how many of those Hazardous and Non-Hazardous objects are either asteroids or comets? That is what Topic 2 discovers.

Topic 2

Topic 2 questions which comet and asteroid NEOs are also PHOs. Again, this is found by creating more sub-dataframes from the Hazardous vs. Non-Hazardous results. By looking at this dataframe only, again, group the objects into Hazardous and Non-Hazardous asteroids and comets using the Tisserand column. These are the results:





At this point, the second hypothesis was proven correctly. There were more Hazardous asteroids in this dataset.

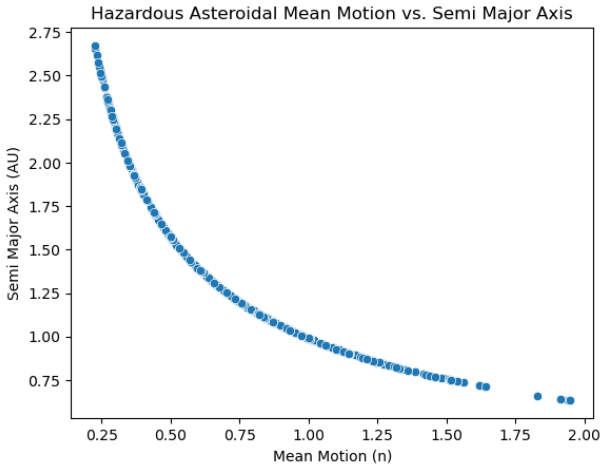
The asteroids in the dataset are much more in comparison to the comets, so by probability, it makes sense there are more Hazardous asteroids. From this point froward in the project, only Hazardous asteroids and Hazardous comets were being studied. Next was to see how these objects behaved graphically.

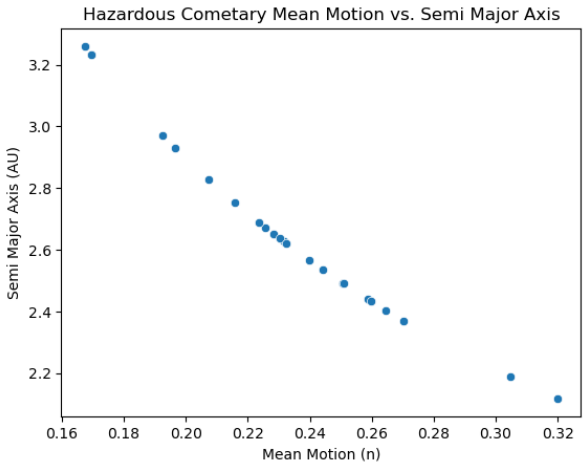
Topic 3

Topic 3 questions what PHOs look like graphically. There were 3 variables considered for this topic:

1. Mean Motion (units of **n**)
    a. Mean motion units are defined as $n = \frac{2\pi}{P}$. Here, **P** is the orbital period. Mean motion is the angular speed required for a body to complete one orbit, assuming constant speed in a circular orbit which completes at the same time as the variable speed, elliptical orbit of the actual body.
2. Semi-Major Axis (units of **AU**)
    a. For reference, 1 AU is 150 million kilometers. The Semi-Major Axis is half the longest diameter of an orbit.
3. Inclination (units of degrees °)
    a. Inclination, $i = \cos^{-1}(\frac{h_z}{|h|})$, is computed from the orbital momentum vector **h** and h_z is the z-component of **h**. Orbital inclination measures the tilt of an object's orbit around a celestial body. It is expressed as the angle between a reference plane and the orbital plane or axis of direction of the orbiting object.

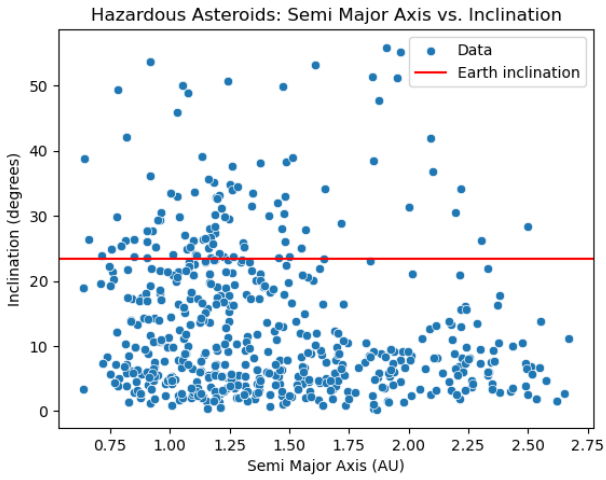
Using these 3 variables, graphs were made to study how they related to asteroids and comets. The first graphic results were 2D graphs that studied the Mean Motion vs. Semi-Major Axis for comets and asteroids. They are as seen below:

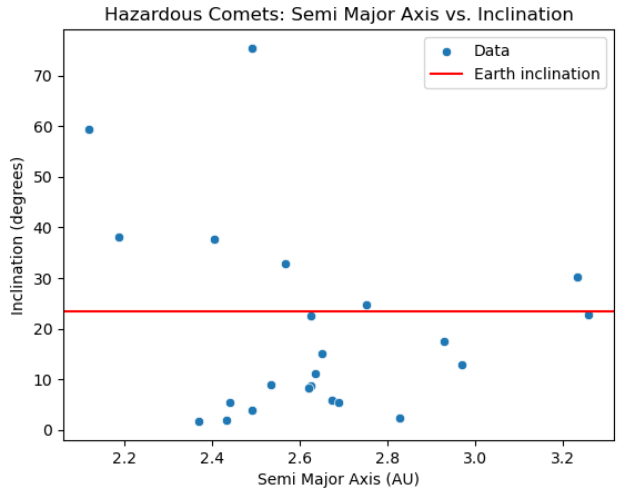




Here, it is shown that asteroids and comets both have an increase in their Mean Motion as their Semi-Major Axis decreases. This is because they are being pulled closer to the Sun and gaining speed before reaching their peak in their elliptical orbits.

Next, the Semi-Major Axis vs. Inclination 2D graphs were made to study the relationship between the small body orbital angle and orbital distance. The results are seen below:





The red line represents Earth's inclination which is 23.4°. For asteroids, there are multiple points where they cross directly through Earth's orbit, thus proving these objects are PHOs. As for the comet, there are only two that cross Earth's path. This is to be expected, as comets have very large orbits and long orbital periods. Extracting the tabular form of this data gives a more detailed view of the data points. The years are cataloged on the table as well.

The asteroid close encounter table below shows the first 10 entries, but there is a total of 36 instances of asteroids crossing Earth's path.

	Inclination	Semi Major Axis	Close Approach Date	Mean Motion
0	23.766741	1.225054	1995-03-22	0.726894
1	23.277205	1.182149	1995-04-22	0.766824
2	23.861945	0.714212	1996-01-22	1.632913
3	22.858798	1.145920	1996-11-08	0.803475
4	22.503754	1.103617	1997-03-22	0.850113
5	23.309957	1.190598	1997-05-22	0.758676
6	23.745884	0.850927	1998-11-22	1.255642
7	23.746184	0.850943	1998-11-22	1.255605
8	23.277205	1.182149	2000-07-08	0.766824
9	23.897140	1.110193	2001-09-15	0.842570
10	22.893293	1.297289	2002-09-08	0.667035

From the years 1995 to 2016, there are only 4 years with no asteroid passings. That is 1999, 2004, 2008, and 2016. Every other year there is at least 1 asteroid passing.

The comet close encounter table below shows the total number of instances of comets crossing Earth's path. This number is very minimal, as it is only two.

	Inclination	Semi Major Axis	Close Approach Date	Mean Motion
0	22.456259	2.626090	1999-04-15	0.231601
1	22.659883	3.259846	2010-02-08	0.167459

From the table, there are 2 instances of comets passing Earth's plane. One is in April 1999 and the other is in February 2010. For the year 1999, though there were no asteroid crossings, there was a comet crossing. Therefore, we can infer that from 1995 to 2016, Earth experienced no asteroid or comets crossing in the years 2004, 2008, and 2016.

Topic 4

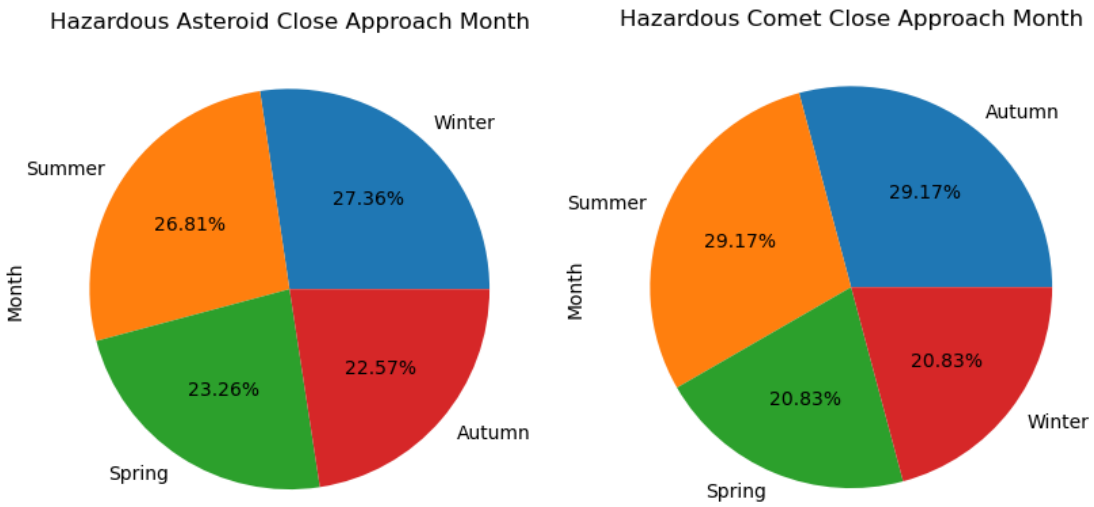
Topic 4 questions what seasons Earth experiences the most PHO interactions. For this round of results, only two datasets were used: hazardous_asteroids_df and hazardous_comets_df. Both dataframes were a result of Topic 2's results.

The seasons referenced in this project are in respect to North America. To get the seasons, the column labeled Close Approach Date was used. First, a function was defined named Seasons(x). The purpose of this function was to extract the months from the Close Approach Column and match them to a season in North America.

According to the meteorological definition, the seasons begin on the first day of the months that include the equinoxes and solstices. In the Seasons(x) definition, the months were labeled like so:

- Winter - months labeled [12,1,2]
- Spring - months labeled [3,4,5]
- Summer - months labeled [6,7,8]
- Fall - months labeled [9,10,11]

Python reads the Close Approach Date column as an object. Therefore, it was converted into a categorical object that can be identified in a pie chart. For this step, the Seasons(x) function was used in conjunction with a lambda function to convert the entire column. This resulted in the two pie charts below:



Out of 24 instances from 1995 to 2016, Earth experienced the most comet PHOs in Summer and Autumn. Out of 731 instances from 1995 to 2016, Earth experienced the most asteroid PHOs in Winter. The earlier stated hypothesis for Topic 4 was disproven.

Discussion and Conclusion

To summarize, the hypothesis for this project was proven to 75% percent accuracy.

1. Topic 1 - Proven
    a. There are more asteroid NEOs in this dataset
2. Topic 2 – Proven
    a. There are more asteroid NEOs that are also PHOs in this dataset
3. Topic 3 – Proven
    a. All PHOs should have a Semi-Major Axis within Earth's range
4. Topic 4 – Disproven

a. Earth does not experience the most PHO interactions in just the Fall. It is actually Summer, Fall, and Winter.

Therefore, it is safe to conclude that Earth is in danger of PHO occurrences nearly year-round.

Machine Learning Model

This project did not implement a Machine Learning Model. However, it did utilize unsupervised exploratory data analysis techniques to analyze the original dataset. It also used clustering methods when it came to grouping certain data variables together and extracting information from those results.

References

NASA. (n.d.). Asteroid Watch. NASA. Retrieved April 26, 2023, from https://www.jpl.nasa.gov/asteroid-watch

P., W. E. (n.d.). *François Felix Tisserand*. Nature News. Retrieved April 26, 2023, from https://www.nature.com/articles/054628a0

Tisserand Criterion. (n.d.). Retrieved April 26, 2023, from https://farside.ph.utexas.edu/teaching/celestial/Celestial/node82.html