Data Intensive Computing – Project Phase 1

REPORT

Problem Statement:

Title: Asteroid trajectories analysis and impact to earth predictions using NASA JPL Asteroid Dataset.

Abstract:

The space race has changed our understanding of the universe at an unimaginable pace. Even today, scientists spend their time and effort uncovering the secrets of space to gain more insights and to protect Earth from any potential threats. In support of this effort, our objective in this project is to gain more understanding about asteroids, their trajectories and whether they pose a potential threat to Earth.

We will be addressing the following key questions:

1. Identify any patterns or correlations between asteroid's properties and their orbital parameters.
2. Identify whether the asteroids are potentially hazardous or not.
3. Predict future asteroid encounters.
4. Predict the position of asteroids where they are nearest and the farthest away from the sun
5. Identify the classification of asteroids and their proximity to the different planets in the solar system.
6. Predict the minimal orbital intersection distance to assess the potential risks of collisions of asteroids with other astronomical objects.

Background:

On our quest to find a problem statement for this project, we came across various datasets on natural disasters, oceans, land, and atmosphere from sources such as *Earthdata*, *Berkeley* earth etc. The datasets were out of the scope of our project. On further exploration we came across the domain of astronomy, which made us understand the significance of celestial bodies on these calamities and on our planet's atmosphere. This newfound knowledge made us stick to the domain of astronomy and led us to the dataset about asteroids, maintained by NASA JPL. On further understanding of this dataset, we came to know that an asteroid impact on Earth can be as catastrophic as destroying an entire city. The presence of an asteroid near the Earth can also cause natural calamities such as tsunamis because of the gravitational pull. This makes our objective to classify asteroids are potentially hazardous, highly significant.

Domain Significance:

An asteroid passed by Earth on February 22nd, 2024. It is said to be the size of a bus. The astronomers from Arizona were the first to see the asteroid. NASA has now started to track this asteroid to predict its return in the future. Likewise, NASA tracks many asteroids they have noticed because asteroids pose a potential threat to our planet, making it crucial to study their characteristics and trajectories to develop strategies to minimize the impact of a crash. In our attempt to build models and algorithms to identify hazardous asteroids, we hope to contribute to their efforts into planetary defense and space exploration.

Source of the Dataset:

The NASA JPL Asteroid dataset is from the NASA JPL Small-Body Database. The dataset contains the list of the asteroids seen by NASA. It has features such as eccentricity, inclination, minimum orbit intersection distance (MOID), absolute magnitude, semi-major axis of the orbit etc. The site provides the names of each feature, after which additional investigation explained its meaning and relevance. In the process, we learnt about Keplar's laws, Newton's gravitational laws and basic calculations used by NASA scientists to classify the asteroid as dangerous.

Data Cleaning Steps:

1. **Drop Duplicates:** We used 'drop_duplicates' from pandas library to identify the duplicate rows in the datasets and delete them. As there are no duplicate rows in the datasets, there are no rows removed.
2. **Rename Columns:** The process involves updating the column names to provide more descriptive and meaningful labels, enhancing clarity and comprehension of the dataset. The definition of the columns is provided in the JPL dataset source.
3. **Removing Features:** In this step we eliminate columns from the dataset that are deemed redundant or irrelevant to the problem statement, thereby enhancing the dataset's relevance and focusing on pertinent information for analysis or modeling purposes.
4. **Removing Rows:** The process involves systematically removing rows from the dataset that contain null values specifically in the columns corresponding to "neo_flag" or "pha_flag." These flags serve as crucial indicators distinguishing hazardous asteroids, thereby ensuring a dataset focused solely on pertinent information related to asteroid classification and threat assessment.
5. **Missing Value Imputation:** In this procedure, absent or null values within the dataset are addressed by filling them using either the mean value of the respective column or by employing interpolation techniques. This approach ensures a comprehensive dataset by substituting missing data points with estimated values derived from the available information, thereby maintaining the integrity and completeness of the dataset for subsequent analysis or modeling tasks.
6. **Features Data Type Formatting:** This process involves converting specific data types within the dataset to strings. By doing so, it enables the execution of various data cleaning actions in subsequent steps. This proactive approach ensures compatibility and facilitates seamless data manipulation, allowing for effective cleansing and preprocessing of the dataset to enhance its quality and usability for further analysis or modeling endeavors.
7. **Correcting Feature Values:** Removing primary designation values prepended to the full name of asteroids.
8. **Handling Outliers:**
    a. Outlier Identification: Outliers within certain features were detected using the interquartile range (IQR) method, a robust statistical technique for outlier detection.
    b. Outlier Removal: Subsequently, these outliers were systematically eliminated from the dataset to mitigate any potential distortions they may cause in subsequent analyses or models.
    c. Verification via Box Plots: To validate the effectiveness of the outlier removal process, box plots were utilized to visually inspect the distribution of the features. This step ensured that the outliers were successfully addressed, thereby enhancing the integrity and reliability of the dataset for further analysis or modeling purposes.

    Note: We are not performing outlier for eccentricity and inclination as they are Keplerian elements and described an asteroid state at a specific epoch.

9. **Encoding Categorical Data:** The process involved converting string-type categorical data from columns such as "pha_flag", "neo_flag", and "classification" into numerical values. This transformation was accomplished through a combination of replacement and factorization techniques. By replacing categorical strings with corresponding numerical representations and factorizing them into unique codes, the dataset's categorical features were effectively encoded into a format suitable for analysis and modeling tasks.
10. **Added New Columns with Standardized Data:** New columns were introduced to the dataset:
    a. "per_seconds": This column represents the conversion of "per_days" into seconds, thereby standardizing the measurement of perihelion distances.
    b. "perihelion_point" and "aphelion_point": The dataset now includes calculated points for both perihelion and aphelion distances, ensuring comprehensive coverage of orbital parameters for further analysis and modeling.

11. **Interchanging Column Positions:**
    a. "period_seconds" column is shifted such that it is right next to "period_days" and "period_years" for consistency.
    b. Similarly, "neo_flag" and "pha_flag" are shifted such that they are right next to their numerical encoding. They are also shifted to the end of the dataset as they are classifications.

Exploratory Data Analysis (EDA):

1. **Statistical Analysis:**

	spkid	absolute_magnitude	epoch	epoch_mjd	epoch_cal	eccentricity	semi_major_axis	perihelion_distance
count	8.884620e+05	888462.000000	8.884620e+05	888462.000000	8.884620e+05	888462.000000	888462.000000	888462.000000
mean	3.795082e+06	16.819612	2.458893e+06	58892.370693	2.019757e+07	0.150460	2.699311	2.296475
std	6.891651e+06	1.207740	6.445488e+02	644.548801	1.772542e+04	0.079076	0.385752	0.417994
min	2.000719e+06	13.468000	2.428098e+06	28097.000000	1.935102e+07	0.000060	1.888571	0.081820
25%	2.237041e+06	16.000000	2.459000e+06	59000.000000	2.020053e+07	0.092043	2.402974	1.991277
50%	2.462084e+06	16.900000	2.459000e+06	59000.000000	2.020053e+07	0.143897	2.654156	2.239008
75%	3.727859e+06	17.629000	2.459000e+06	59000.000000	2.020053e+07	0.197604	3.002403	2.579431
max	5.401720e+07	20.249000	2.459000e+06	59000.000000	2.020053e+07	0.968396	5.518980	5.301929

8 rows × 28 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 886661 entries, 0 to 886660
Data columns (total 32 columns):
 #   Column                     Non-Null Count   Dtype
---  ------                     --------------   -----
 0   id                         886661 non-null  object
 1   spkid                      886661 non-null  int64
 2   full_name                  886661 non-null  string
 3   primary_designation        886661 non-null  string
 4   neo_flag                   886661 non-null  object
 5   pha_flag                   886661 non-null  object
 6   absolute_magnitude         886661 non-null  float64
 7   orbit_id                   886661 non-null  object
 8   epoch                      886661 non-null  float64
 9   epoch_mjd                  886661 non-null  int64
 10  epoch_cal                  886661 non-null  float64
 11  equinox                    886661 non-null  object
 12  eccentricity               886661 non-null  float64
 13  semi_major_axis            886661 non-null  float64
 14  perihelion_distance        886661 non-null  float64
 15  inclination                886661 non-null  float64
 16  mean_motion                886661 non-null  float64
 17  time_of_perihelion_passage 886661 non-null  float64
 18  tp_cal                     886661 non-null  float64
 19  period_days                886661 non-null  float64
 20  period_years               886661 non-null  float64
 21  moid                       886661 non-null  float64
 22  moid_ld                    886661 non-null  float64
 23  sigma_e                    886661 non-null  float64
 24  sigma_a                    886661 non-null  float64
 25  sigma_q                    886661 non-null  float64
 26  sigma_i                    886661 non-null  float64
 27  sigma_n                    886661 non-null  float64
 28  sigma_tp                   886661 non-null  float64
 29  sigma_per                  886661 non-null  float64
 30  classification             886661 non-null  int64
 31  pha_flag_numerical         886661 non-null  int64
dtypes: float64(21), int64(4), object(5), string(2)
memory usage: 216.5+ MB
```

The 'describe()' method provides a summary of statistics for key features including the mean, standard deviation, mode, median, maximum, and minimum values.

Furthermore, when utilizing the 'info()' function to examine the data frame, it reveals the data type associated with each column and serves as validation that there are no null values present after the completion of the data cleaning procedure.

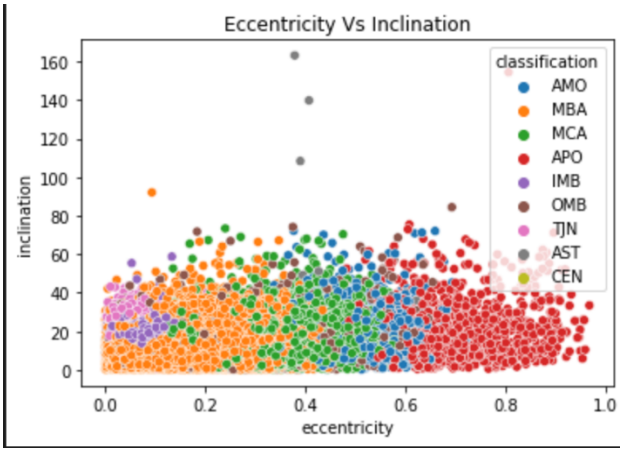
2. **Cross Tabulation:**

neo_flag	N	Y	All
pha_flag			
N	885609	2400	888009
Y	0	453	453
All	885609	2853	888462

The contingency table analysis provides valuable insights into the distribution and relationships between *Potentially Hazardous Asteroid (PHA)* and *Near-Earth Object (NEO)* flags within the asteroid dataset. By examining the frequency counts and patterns in the cross-tabulation, researchers gain a better understanding of the characteristics and classifications of asteroids, particularly those potentially posing hazards to Earth.
From this table, we can say that if an asteroid is not a Near-Earth object, then it cannot be a Potentially Hazardous Asteroid.

3. **Scatter Plots:**

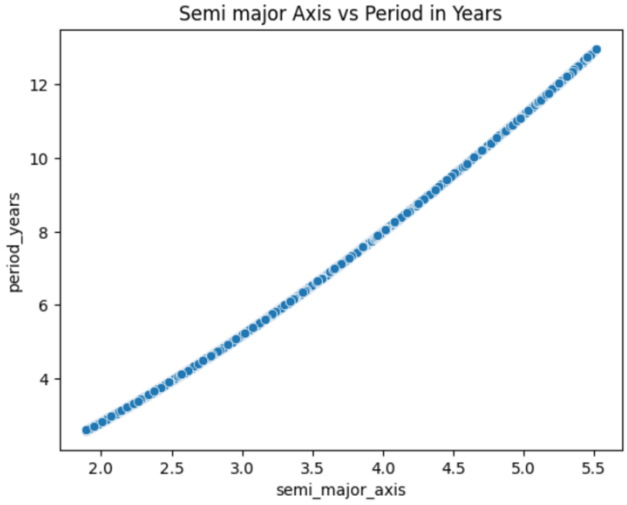
a. Eccentricity vs Inclination



Scatter plots between *eccentricity* and *inclination* show the relationship between the elliptical orbit and the tilt.

The relationship between *eccentricity* and *inclination* can provide clues about the asteroid's dynamical history, such as its interactions with other bodies or processes that influenced its orbit.

b. Semi Major Axis vs Period in Years



The relationship between the *semi-major axis* and the *period_time* in years is given by *Kepler's third law of planetary motion* which states that $T^2 = \frac{4\pi^2}{G(M1+M2)} a^3$

where,

T = Orbital period in seconds
G = the gravitational constant = 6.674 x 10^{-30}
M1 = the mass of the asteroid
M2 = the mass of the sun
a = the semi-major axis (unit: au)

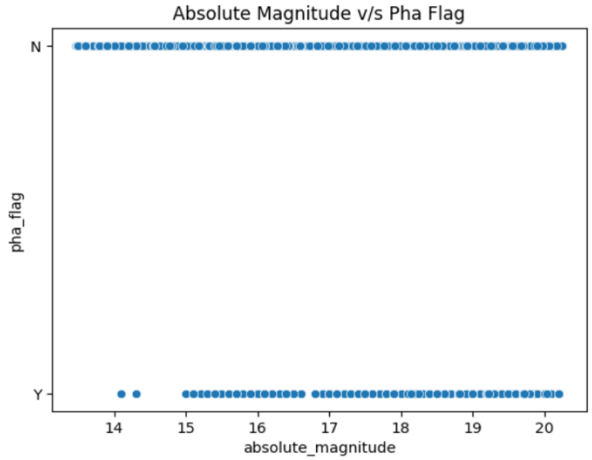
We know that M2 << M1 and M1 and G are constants, so we can say that $T^2 = a^3$

This is supported by the above scatter plot, which is like

$$y^2 = x^3$$

where, x = semi-major axis and y = period_years (T).

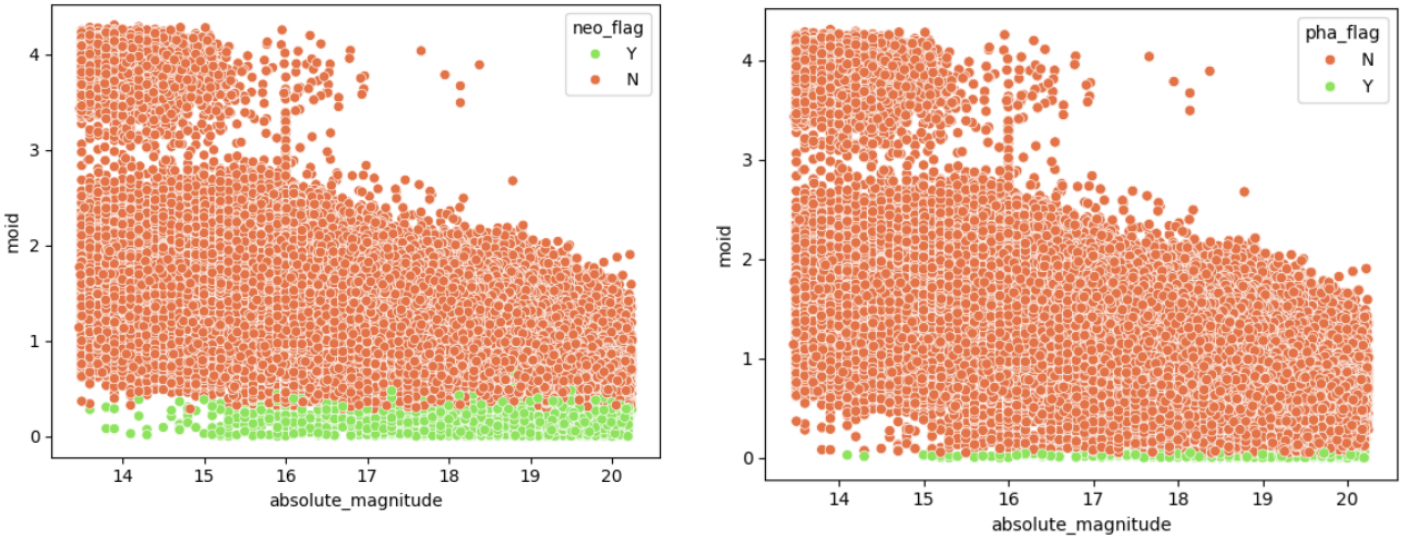
c. PHA Flag vs Absolute Magnitude



The scatter plot between *Potentially Hazardous Asteroid (PHA)* flag and *absolute magnitude* show that the field 'absolute magnitude' alone is not sufficient to determine PHA flag

d. Absolute Magnitude and MOID with PHA / NEO Flag as Hue

An asteroid's *absolute magnitude* is the visual magnitude an observer would record if the asteroid were placed 1 astronomical unit (au) away, and 1 au from the Sun and at a zero-phase angle.
The *Minimum Orbit Intersection Distance (MOID)* is the closest distance between the orbit of an asteroid and the orbit of Earth. It is a measure of how closely the paths of the Earth and the asteroid approach each other.
Both *absolute magnitude* and *MOID* are important parameters in assessing the potential risk of impact of *near-Earth asteroids*. While they are independent parameters, scientists study them together along with other characteristics such as size, composition, and orbit to evaluate the potential threat posed by an asteroid.



Based on the scatter graphs on the top-left, it can be deduced that within a specific range of absolute magnitude, an asteroid may qualify as a *Near-Earth Objec*t if its *Minimum Orbit Intersection Distance (MOID)* falls below a certain decimal threshold.
The graph on the top-right illustrates that there are few *Potentially Hazardous Asteroids*, particularly when their *MOID* values are closer to zero. We can also say that *Potentially Hazardous Asteroids* are a subset of *Near-Earth Objects*, but not vice-versa.

e. Exploration of Asteroids in Solar System

One of the features observed in the dataset concerning asteroids is an orbit classification, which is particularly intriguing. This classification is represented by a three-character code, initially perplexing to decipher. However, the JPL Small-Body Database Search Engine offers a tool to translate this code into a descriptive name for the orbit class, facilitating subsequent research in Wikipedia.

The table below presents a summary of the orbit classifications, arranged by their respective distances from the sun.

The following graph below illustrates the perihelion and aphelion values of asteroids and planets, revealing a clustering pattern where asteroids of the same classification are grouped together.

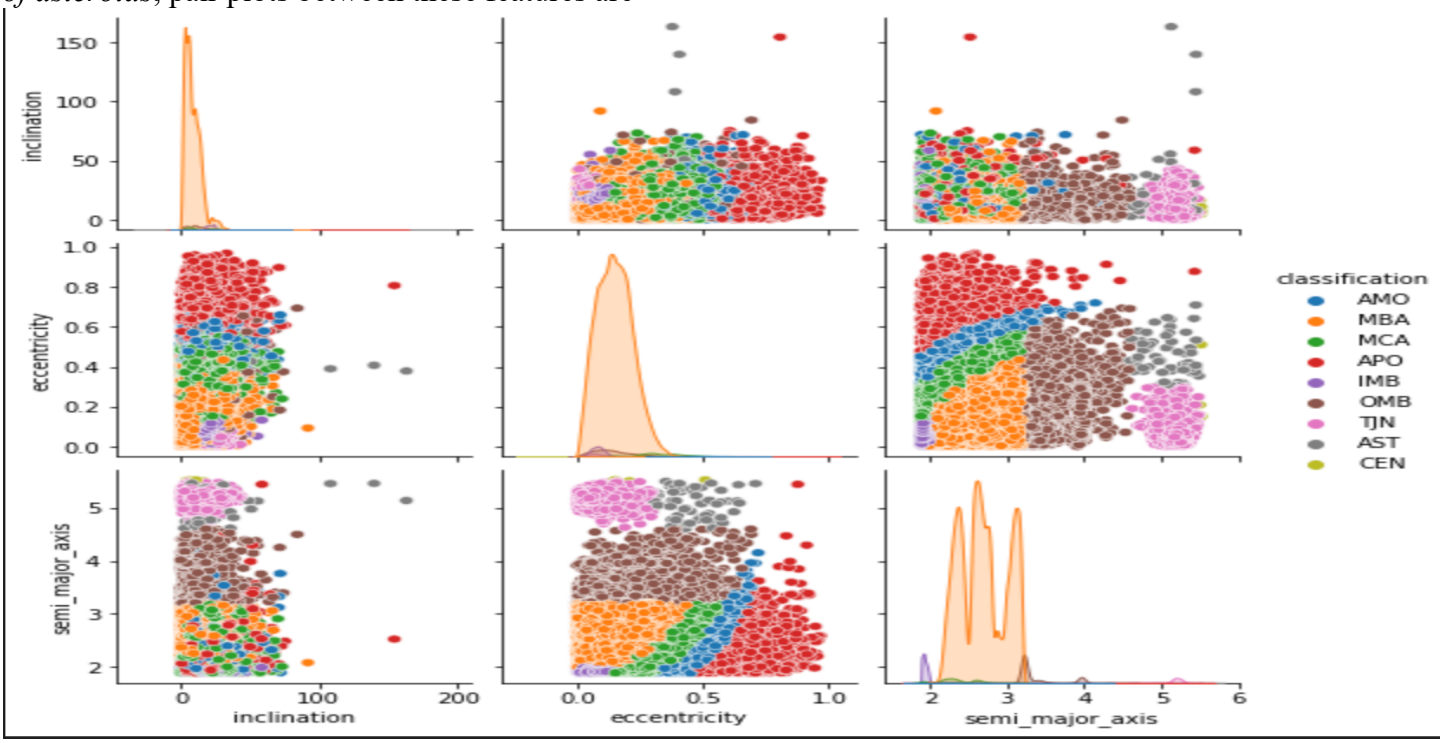
The *perihelion* and *aphelion point* of an asteroid is given by the relation:

*Perihelion = semi major axis * (1 – eccentricity)*

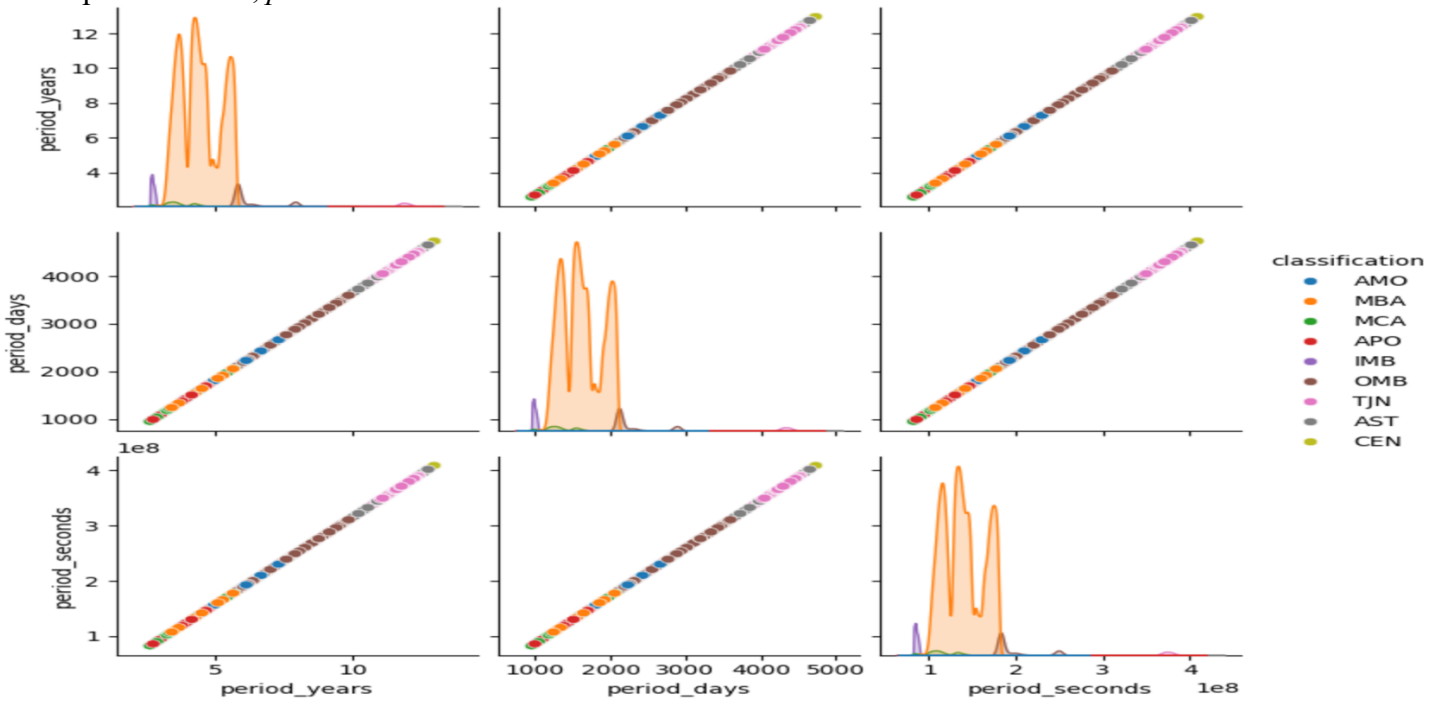
*Aphelion = semi major axis * (1 + eccentricity)*

Class	Name	Description
IEO	Arita	Resides entirely within the Earth's orbit
ATE	Aten	Earth-crossing asteroids, mostly within Earth's orbit
APO	Apollo	Earth-crossing asteroids, mostly beyond Earth's orbit
AMO	Amor	Near-Earth asteroids predominantly outside Earth's orbit
MCA	Mars Crossing Asteroid	A class of asteroids whose orbits intersect with that of Mars, potentially posing a risk of collision with the planet.
IMB	Inner Main-belt Asteroid	located within the asteroid belt between the orbits of Mars and Jupiter
MBA	Main-belt Asteroid	Rocky objects located within the asteroid belt between Mars and Jupiter, providing insights into solar system formation and potential Earth impact threats.
OMB	Outer Main-belt Asteroid	Rocky bodies located in the outer region of the asteroid belt between the orbits of Mars and Jupiter
AST	Asteroid (other)	
TJN	Jupiter Trojan	Reside in Jupiter's orbit, positioned at the L4 or L5 Lagrange Points
CEN	Centaur	Small solar system bodies between Jupiter and Neptune
TNO	Trans Neptunian Object	Minor planets orbiting beyond Neptune at a greater average distance

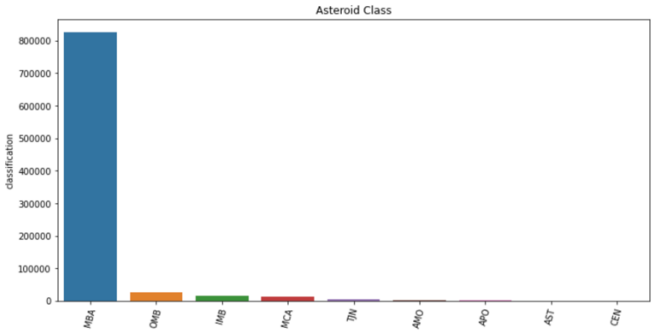
4. **Pair Plots:** Pair Plots visualizes given data to find the relationship between where the variables can be continuous or categorical.
As per the above-mentioned relationship between *eccentricity* and *inclination*, *semi major axis* and *orbital time of asteroids*, pair plots between these features are



The plot above suggests that classification can be effectively determined by assessing the values of *eccentricity* and *semi-major axis* rather than *inclination*. This is evident from the well-defined clusters of asteroids corresponding to their classes in the *eccentricity vs. semi-major axis graph*.
While the graph below shows that classification of clusters cannot take place based on how long the asteroid takes to complete its orbit, *period*.

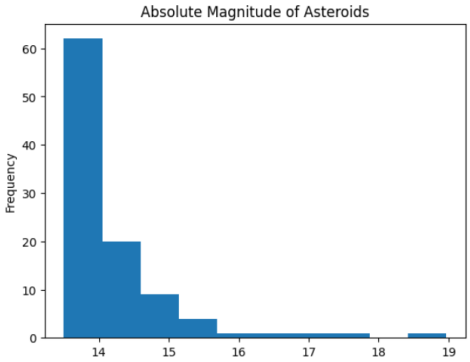


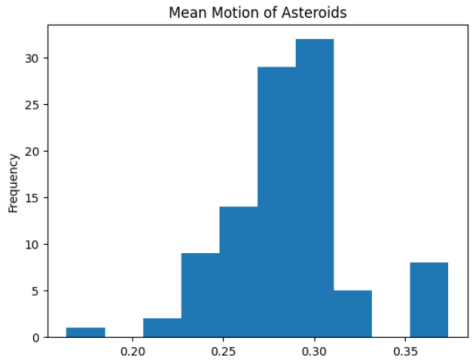
5. **Bar Graph and Histograms:**
The bar graphs are plotted with this dataset to show the frequency of the asteroids as per their orbital classification, which helps to classify the asteroids nearer to earth.



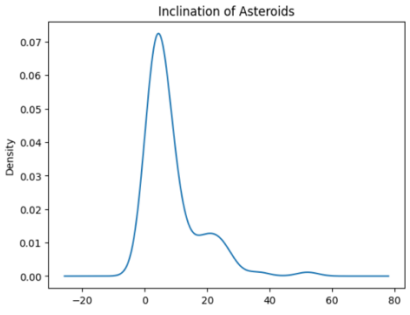
Based on the bar graph on the left, it is evident that most asteroids are classified as MBA (Main-belt Asteroids), followed by OMB (Outer Main-Belt Asteroids) in second place.

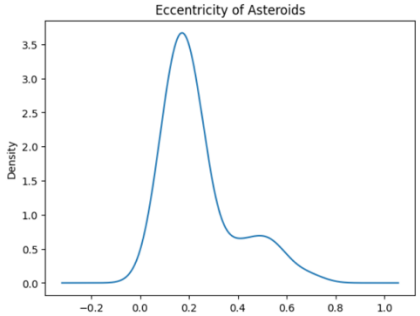
The below histograms for the first 100 asteroids, we can depict that most of the time the value of *Absolute Magnitude* is in range of 13 to 14 and the value of *Mean Motion* is in range of 0.27 to 0.3.

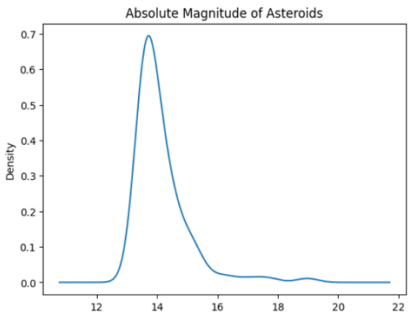


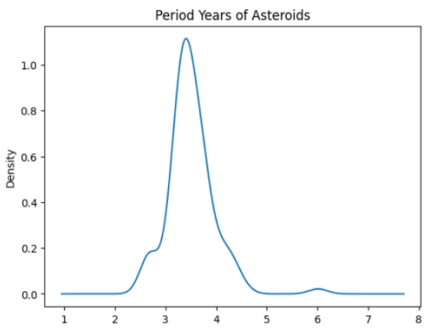


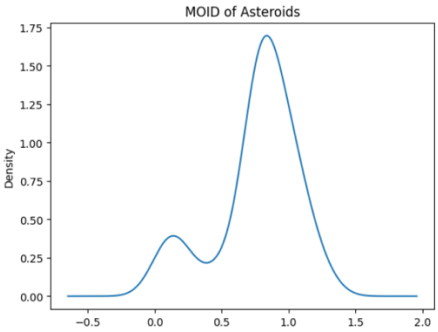
6. **Bell Curves (Statistical Plot):** As per *John Tuckey*, Bell Curves or standard deviations plot are used to see if the standard deviation varies between diverse groups of data.
Bell curves plotted on this dataset show the mean distribution of the data for some noteworthy features such as eccentricity and inclination.



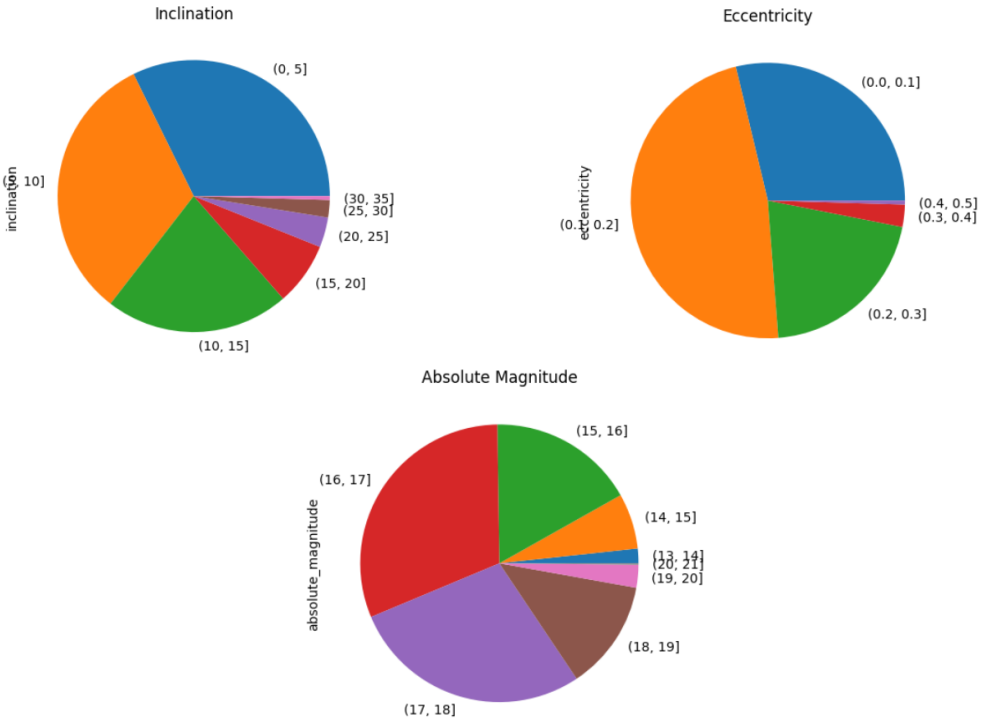








7. **Pie Charts:** Pie Charts are plotted on this dataset to visualize the distribution of asteroids as per the crucial features such as *eccentricity*, *inclination*, orbit class, *absolute magnitude*, and *minimum orbital distance*.

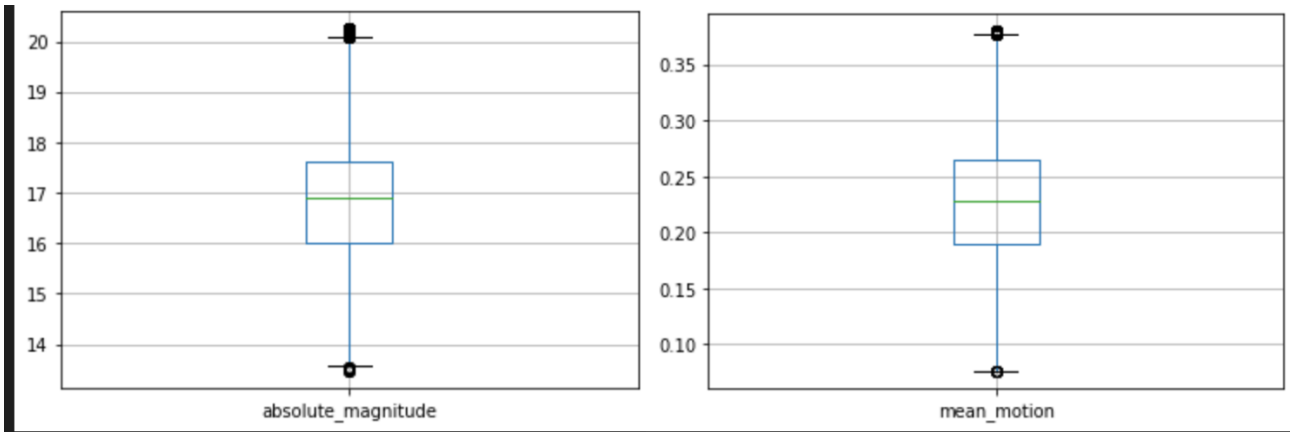


From the above pie charts we can say that,

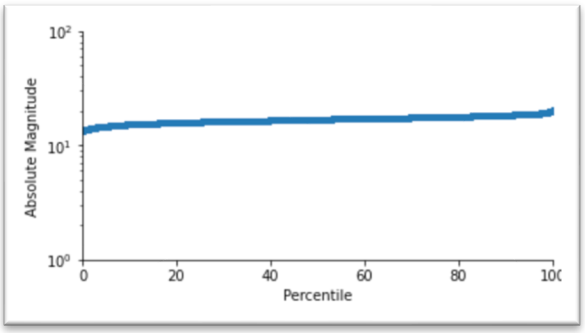
- More than 90% of the time, the value of *Eccentricity* lies from 0 to 0.3.
- More than 50% of the time, the value of *Absolute Magnitude* lies from 16 to 18.
- More than 50% of the time, the value of *Inclination* lies from 0 to 10.

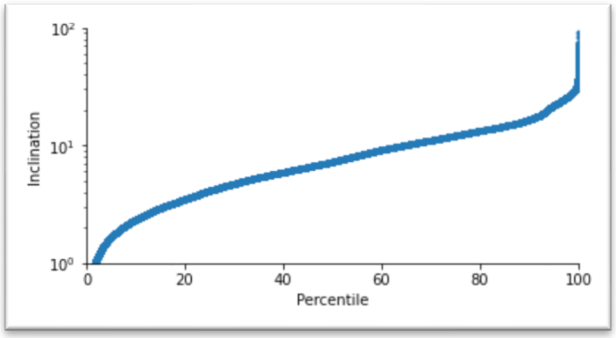
In summary, this pie chart provides a visual representation of the distribution of asteroid *inclinations* and *MOID* focusing on the top 10 most common inclinations and MOID values in the dataset. It is plotted to quickly grasp the relative proportions of different *inclination* values among the asteroids.

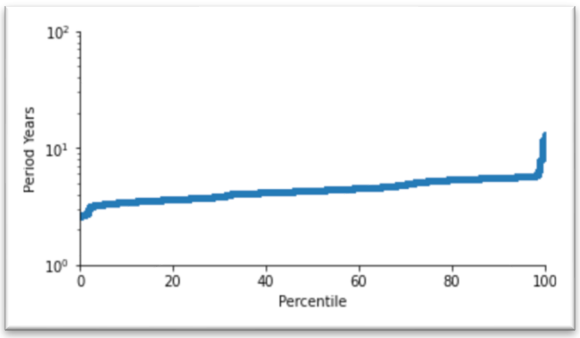
8. **Box Plots:** Box Plots are plotted on some noteworthy features such as *absolute magnitude* and *mean motion* to visualize the outliers and data with respect to the inter quartile range.

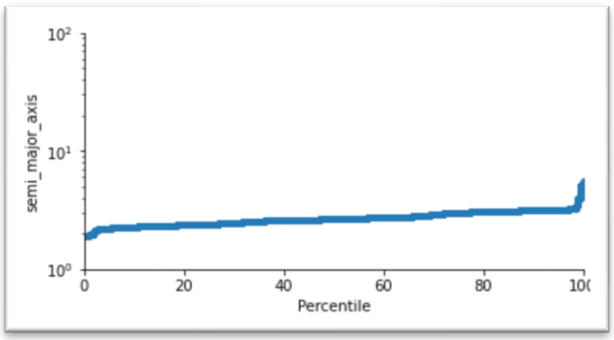


9. **Percentile Plot:** The percentile plots plotted on this dataset show the trend of the data as per the percentiles they are lying in. Percentile plots for some related features of this dataset such as *inclination*, *absolute magnitude*, *period years* and *semi major axis*.

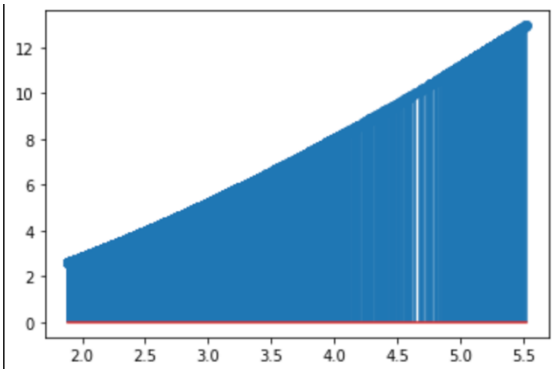








10. **Stem Plots:**



Since it is already known from *Kepler's third law of planetary motion* that *semi major axis* and *period time* in years are related by

$$T^2 = a^3.$$

where,
T = Period Time in Years
a = Semi major axis
The stem plot as x=*semi major axis* and y=*period years* is plotted to visualize the same polynomial relationship.

Step 11: Correlation of Columns: A correlation analysis was performed on the dataset, examining several key attributes related to asteroid characteristics. The attributes considered for correlation analysis include:

- SPKID (Small-body Planet Center unique identifier)
- Classification Numerical
- Eccentricity
- Semi-major Axis
- Perihelion Distance
- Inclination
- Mean Motion
- Time of Perihelion Passage
- Orbital Period (in days)
- Orbital Period (in years)
- Root Mean Square Residuals (RMS)

The correlation analysis was conducted after removing any rows with missing values (NaN) to ensure the accuracy and reliability of the correlation coefficients.
The correlation matrix generated from the analysis reveals the pairwise correlations between these attributes. Each cell in the correlation matrix represents the Pearson correlation coefficient, which ranges from -1 to 1.

- A correlation coefficient close to 1 indicates a strong positive linear relationship between the attributes.
- A coefficient close to -1 signifies a strong negative linear relationship.
- A coefficient around 0 implies little to no linear relationship between the attributes.

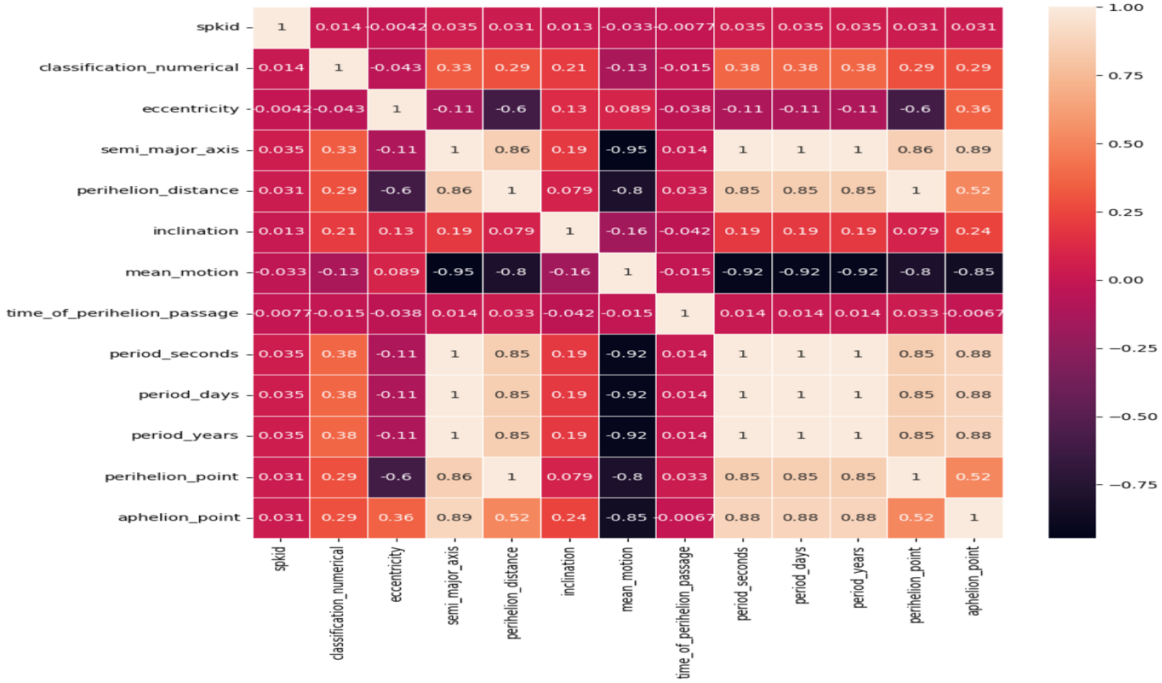
Interpreting these correlation coefficients provides insights into potential associations and dependencies among the asteroid characteristics under study.
The correlation analysis results, summarized in the 'astro_df_corr' data frame, serve as a valuable resource for further exploration and understanding of the underlying patterns and relationships within the asteroid dataset.

	spkid	classification_numerical	eccentricity	semi_major_axis	perihelion_distance	inclination	mean_motion	time_of_perihelion_passage	period_days	period_years	rms
spkid	1.000000	0.014026	-0.004227	0.035490	0.031135	0.013415	-0.032884	-0.007675	0.035428	0.035428	0.029076
sification_numerical	0.014026	1.000000	-0.043071	0.334034	0.289859	0.207656	-0.130932	-0.014702	0.375515	0.375515	-0.004001
eccentricity	-0.004227	-0.043071	1.000000	-0.108229	-0.601232	0.127470	0.088819	-0.038145	-0.109352	-0.109352	-0.136734
semi_major_axis	0.035490	0.334034	-0.108229	1.000000	0.855786	0.190853	-0.947330	0.013814	0.996227	0.996227	0.033271
perihelion_distance	0.031135	0.289859	-0.601232	0.855786	1.000000	0.079255	-0.799684	0.032952	0.854974	0.854974	0.096955
inclination	0.013415	0.207656	0.127470	0.190853	0.079255	1.000000	-0.158403	-0.041657	0.189952	0.189952	-0.060674
mean_motion	-0.032884	-0.130932	0.088819	-0.947330	-0.799684	-0.158403	1.000000	-0.014654	-0.916996	-0.916996	-0.036777
perihelion_passage	-0.007675	-0.014702	-0.038145	0.013814	0.032952	-0.041657	-0.014654	1.000000	0.014139	0.014139	0.175345
period_days	0.035428	0.375515	-0.109352	0.996227	0.854974	0.189952	-0.916996	0.014139	1.000000	1.000000	0.031671
period_years	0.035428	0.375515	-0.109352	0.996227	0.854974	0.189952	-0.916996	0.014139	1.000000	1.000000	0.031671
rms	0.029076	-0.004001	-0.136734	0.033271	0.096955	-0.060674	-0.036777	0.175345	0.031671	0.031671	1.000000

Step 12: Heatmap

A heatmap visualization was created to illustrate the correlation matrix derived from the asteroid dataset (astro_df_corr). The heatmap, plotted using the Seaborn library in Python, provides a clear and concise depiction of the pairwise correlations among various attributes of the asteroids under study.
From the below heatmap, we can say that

- *Period_days* is derived from Period_years as the correlation is equal to 1.
- *Perihelion_distance* is well associated with semi-major axis and Period days with the correlation value of 0.86 and 0.85 respectively.
- *Perihelion point* and *aphelion point* are well correlated with semi-major axis and period days.



Step 13: Group Asteroids by classes

	classification	count	class_name
0	MBA	825852	Main-belt Asteroid
1	OMB	26533	Outer Main-belt Asteroid
2	IMB	14505	Inner Main-belt Asteroid
3	MCA	13492	Mars Crossing Asteroid
4	TJN	5154	Jupiter Trojan
5	AMO	1657	Amor
6	APO	1196	Apollo
7	AST	69	Asteroid (other)
8	CEN	4	Centaur

Group asteroids by classes: MBA, OMB, IMB, MCA, TJN, AMO, APO, AST, CEN.
The left table gives the count of asteroids in each class.

classification	period_years
IMB	2.688386
MCA	3.644061
APO	3.657458
AMO	3.830969
MBA	4.412025
OMB	6.223524
AST	11.400317
TJN	11.843731
CEN	12.938571

classification	semi_major_axis
IMB	1.933299
MCA	2.360735
APO	2.361147
AMO	2.436203
MBA	2.681729
OMB	3.377694
AST	5.062244
TJN	5.195534
CEN	5.511272

These two tables show the mean value of period years and semi-major axis for all the grouped asteroids, respectively.

From the scatter plots we know that asteroids in APO and AMO class are the nearest to the earth.

```
Total counts of Apollos and Amor
APO    1196
AMO    1657
Name: classification, dtype: int64
```

```
Counts of Potentially Hazardous Asteroids
APO    406
AMO     47
Name: classification, dtype: int64
```

```
Counts of Near Earth Objects
AMO    1657
APO    1196
Name: classification, dtype: int64
```

The data above presents the total count of asteroids, along with the counts of Potentially Hazardous Asteroids and Near-Earth Objects, categorized under APO and AMO classifications.

Summary

Overall, the data cleaning and EDA processes provided valuable insights into asteroid characteristics and behaviors, contributing to our understanding of space phenomena, and enhancing our ability to mitigate potential risks from asteroid impacts.

References

1. https://www.earthdata.nasa.gov/
2. https://berkeleyearth.org/data/
3. https://nypost.com/2024/02/22/lifestyle/an-asteroid-the-size-of-a-bus-zoomed-past-earth-wednesday-how-close-it-get/
4. https://ssd.jpl.nasa.gov/sbdb_query.cgi
5. https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/
6. https://matplotlib.org/3.1.1/index.html