Exploratory Data Analysis Report

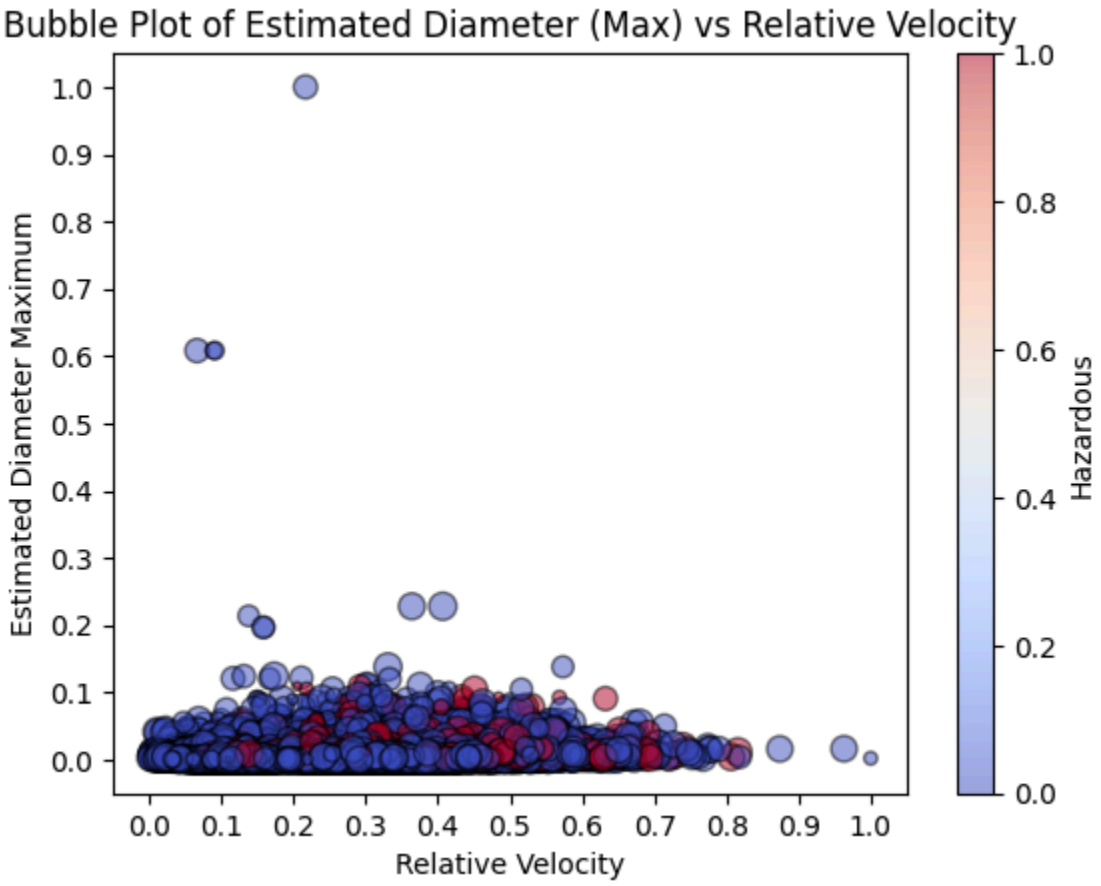
Introduction:

With the use of visualizations and analytical insights, the "Exploratory Data Analysis Report" delivers a thorough analysis of a dataset on asteroids. It identifies major trends and characteristics in the data. Beginning with a scatter plot that illustrates the connections between relative velocity, estimated maximum diameter, and them being hazardous, it demonstrates the predominance of low relative velocities and smaller asteroid sizes. The following sections of the paper highlight the top 25 maximum diameter asteroids and the top 25 fastest-moving asteroids, respectively, in order to provide readers a better grasp of the largest and fastest celestial bodies. With right-skewed patterns in estimated diameter and relative velocity, histograms show the distribution of continuous variables in the sample, highlighting the predominance of smaller asteroids and lower velocities.

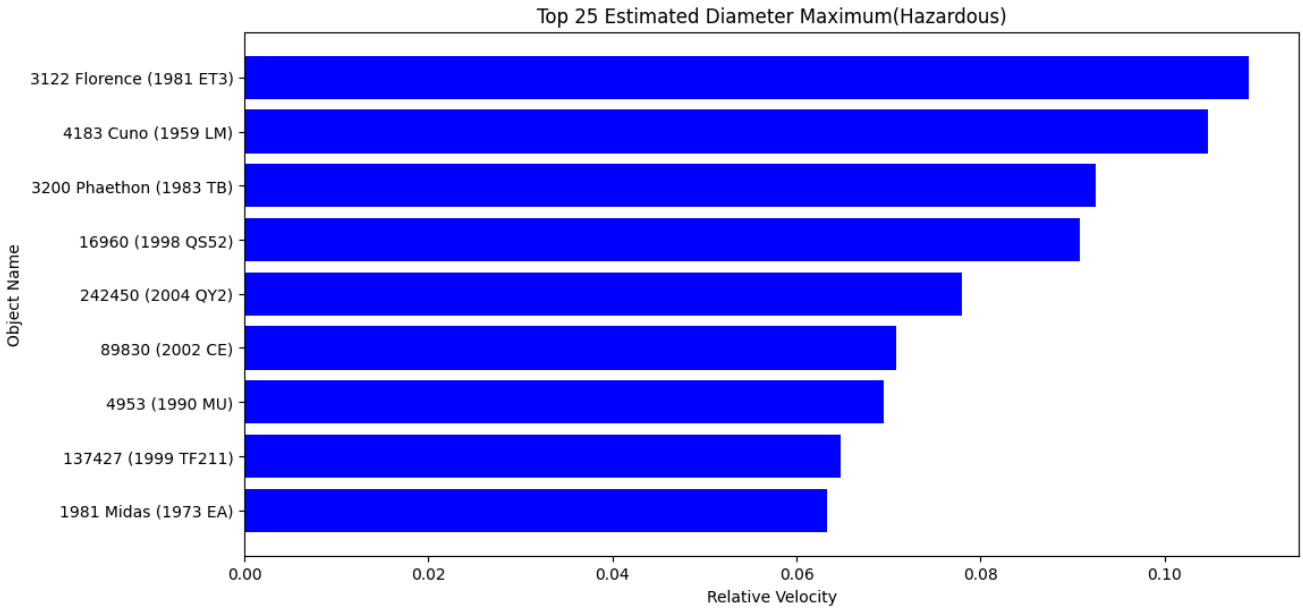
Outliers in the data are clearly highlighted, especially in the estimated diameter and relative velocity variables. A scatter matrix offers insights into the interaction of numerous parameters. The dataset is divided into two pie charts that show the relative velocities and missed distances of the hazardous and non-hazardous asteroids, respectively. This clearly shows that all hazardous asteroids have high values for both of these qualities. This study acts as a starting point for additional investigation into asteroid tracking and planetary defense.

The "Exploratory Data Analysis Report" explores the asteroid dataset and highlights significant themes and linkages. A scatter plot showing the predominance of low relative velocities and tiny asteroid sizes is presented first, then information on the largest and fastest moving asteroids is then provided. The presence of outliers is discussed while the histograms illustrate the distributions of continuous variables. Using a scatter matrix, attribute correlations are investigated with an emphasis on asteroids' potential for harm. The dataset is clearly separated into hazardous and non-hazardous asteroids using two pie charts, illustrating the characteristics of dangerous objects. The foundation for future research in asteroid tracking and planetary defense is laid forth in this report.

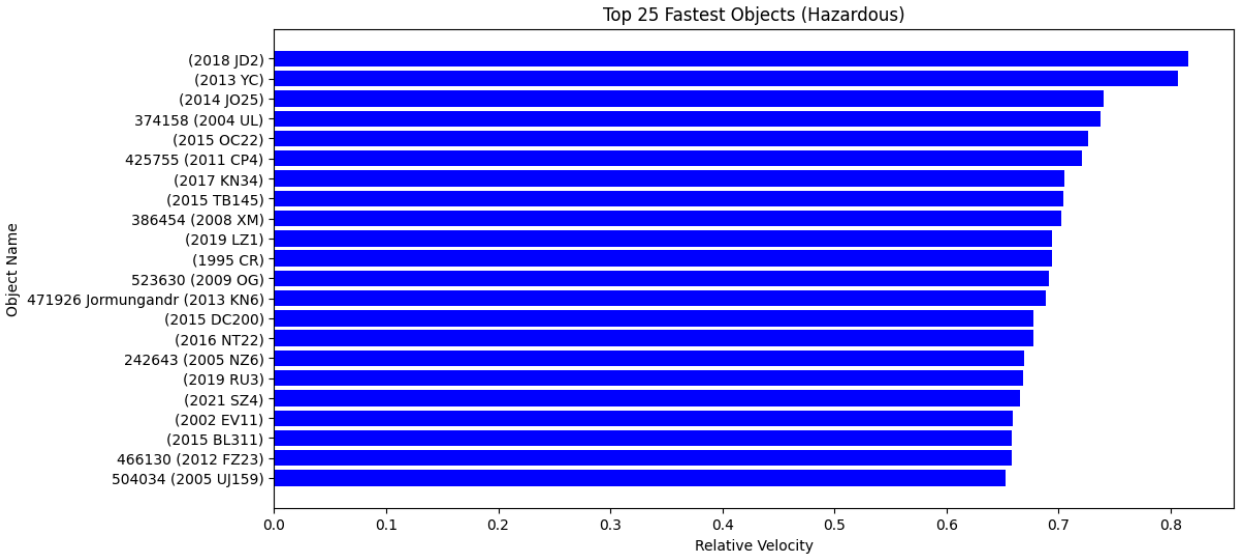
Analysis:



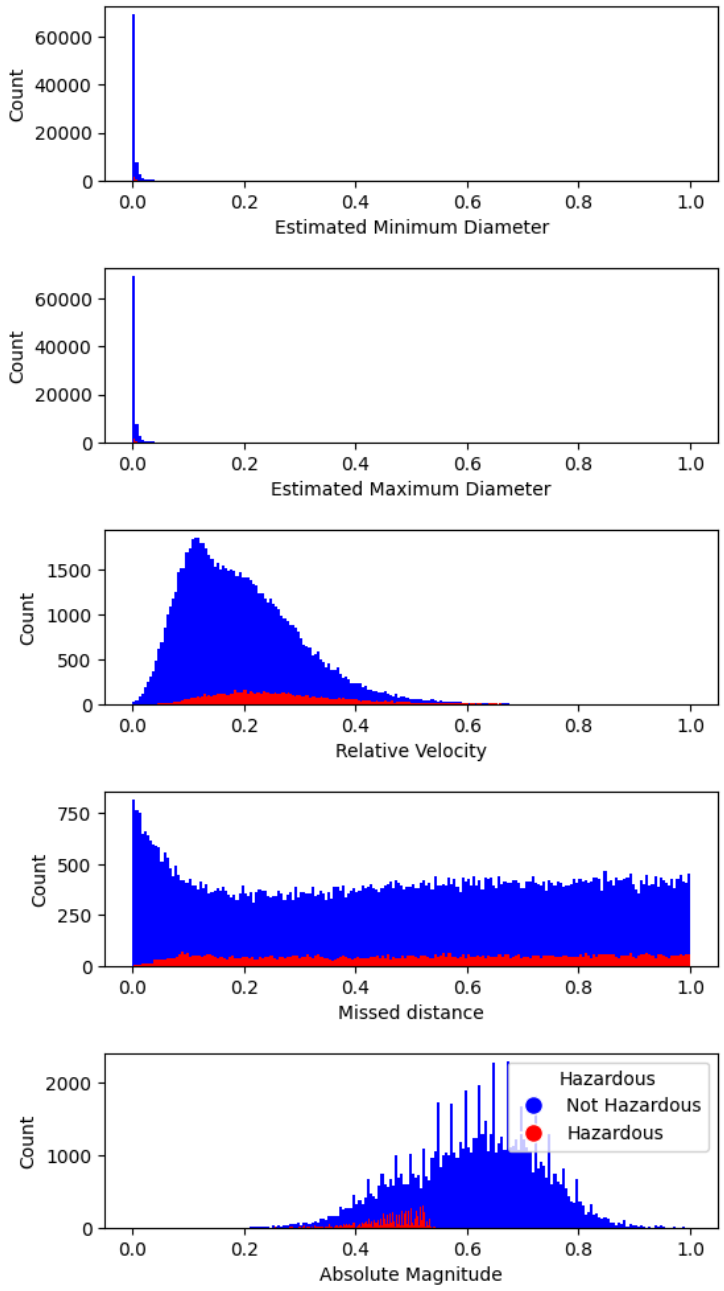
The above scatter plot tells us that the relative velocity of the majority of asteroids in the dataset ranges from (0.0,0.8). Most of the asteroids have a low estimated maximum diameter. The color of the bubbles show whether they are hazardous or not with the size of the bubble showing their missed distance from earth.



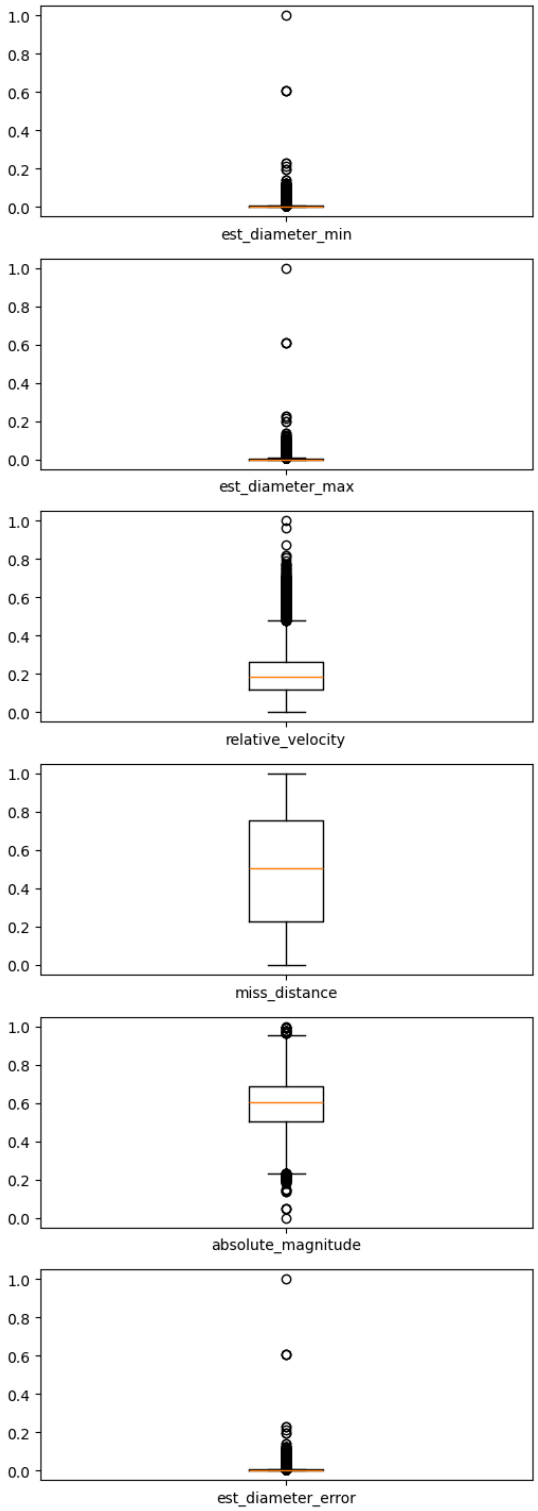
The above bar graph shows us the relative velocity of those hazardous asteroids which come under top 25 when sorted by estimated diameter maximum - having maximum or close to maximum “estimated diameter maximum” value



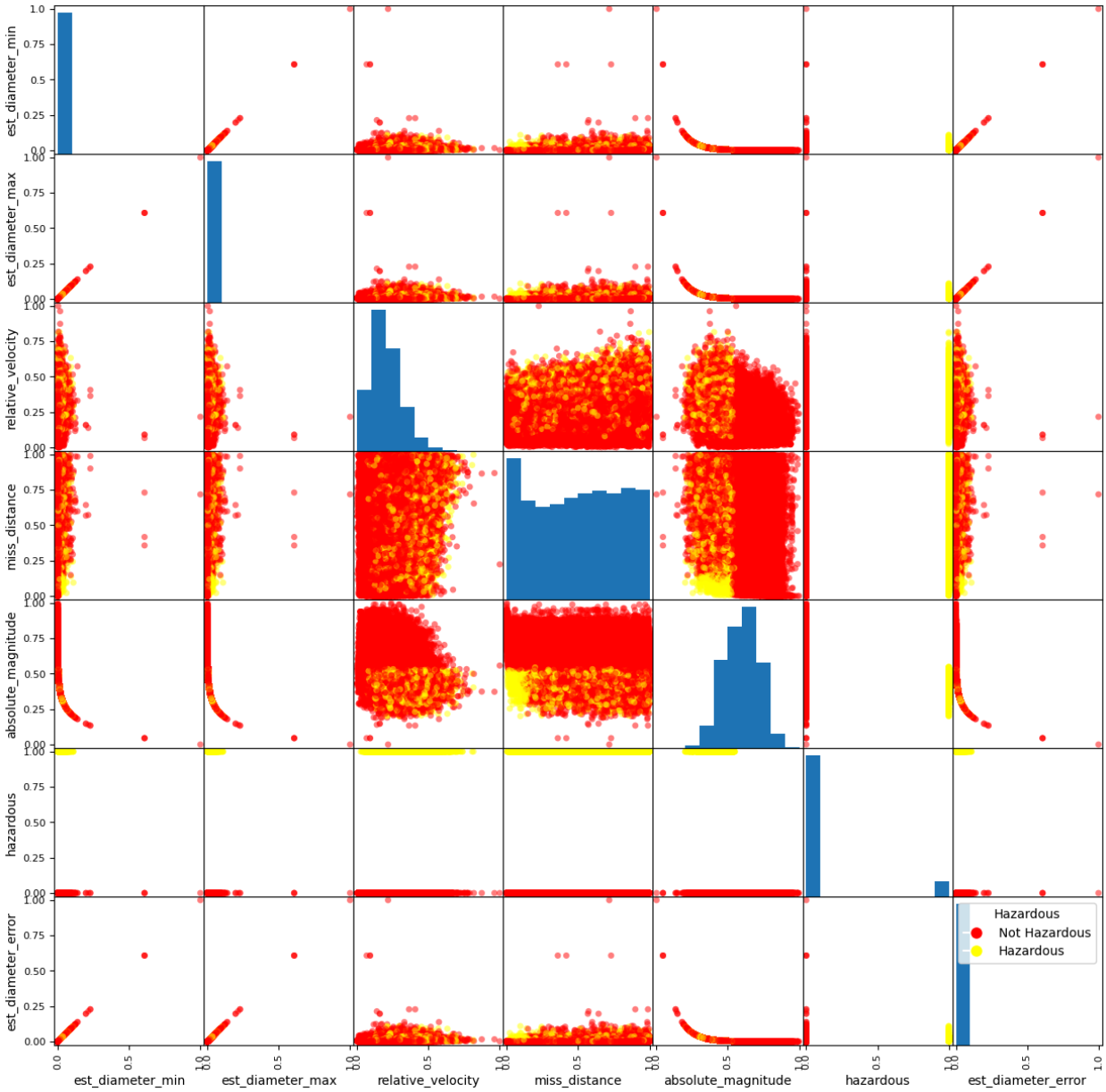
The above bar graph shows the names and relative velocity values of the top 25 fastest asteroids.



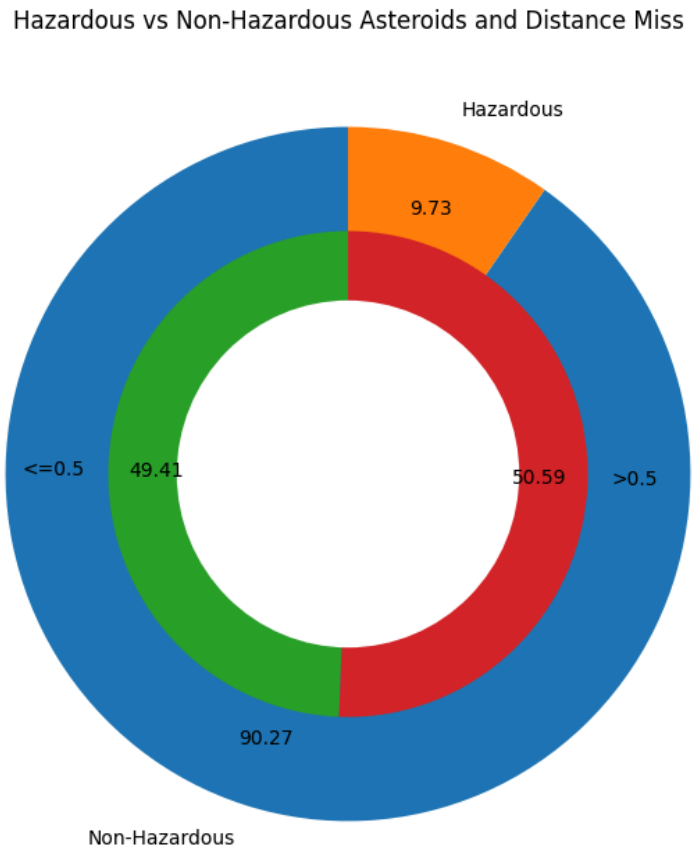
Above are the histograms of all the continuous variables of the data set. We can see that the estimated diameter (max and min) is highly right skewed meaning that most of the near earth objects are small in size. Relative velocity is also right skewed. Most of the hazardous asteroids have low relative velocity. Missed distance for both hazardous and non hazardous are evenly distributed. Absolute magnitude (luminosity) is slightly left skewed. This means that most of the asteroids have relatively high luminosity.



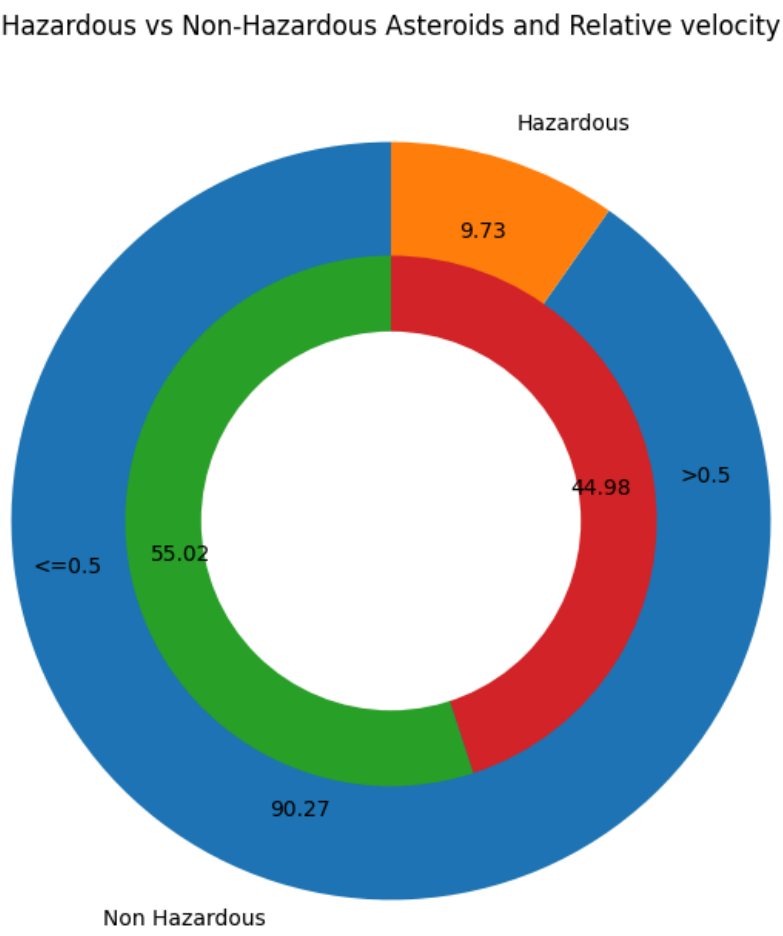
We can see that the estimated diameter (max, min and error) is highly right skewed. For all three most of the outliers lie within the range (0.0, 0.2) and very few outliers are greater than 0.5. Relative velocities are also right skewed with mean at approximately 0.2. Most of the outliers lie in [0.5,0.8). Miss distance has normal distribution with no outliers. Absolute magnitude (luminosity) the data slightly left skewed with outliers within (0.95,1.0] and (0.0,0.25]



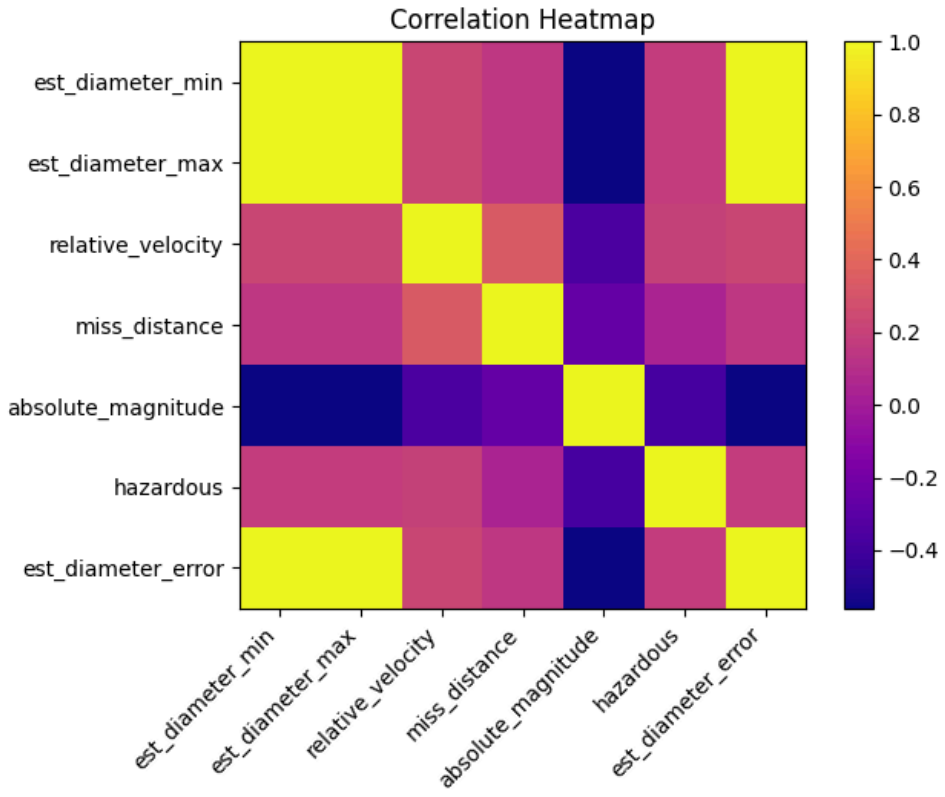
The above graph shows a scatter matrix, a grid of scatter plots, where each variable is plotted against every other variable. The points in each scatter plot are color-coded based on the ‘hazardous’ attribute as shown in the legend. The diagonal shows the histogram of each attribute.The scatter plot between relatove_velocity and miss_distance shows, the greater the relative velocity the more hazardous they are and vice versa.



The pie chart above shows the count of hazardous vs. non-hazardous asteroids in the dataset. For each outer slice (one: hazardous and second:non-hazardous), the inner circle shows how many of those asteroids have high value of missed distance and low value of missed distance.It can also be seen that all hazardous asteroids have high missed distances.



The pie chart above shows the count of hazardous vs. non-hazardous asteroids in the dataset. For each outer slice (one: hazardous and second:non-hazardous), the inner circle shows how many of those asteroids have high value and low values of relative velocity. It can also be seen that all hazardous asteroids have high relative velocities.



The matrix shows that there is very low correlation between absolute magnitude(luminosity) and diameter of the asteroid. There exists some small level of correlation between relative velocity and luminosity, relative velocity and hazardous, miss distance and relative velocity. The highest correlation is between relative velocity and miss distance. All of the relevant correlation values are not very high.

Conclusion:

In conclusion, the asteroid dataset is thoroughly reviewed in this exploratory data analysis report, which also identifies remarkable patterns and traits in the celestial objects it includes. We have learned important things about the distribution, outliers, and interactions of the variables in the dataset by using a range of analytical and visual methods. Particularly, the distinction between hazardous and non-hazardous asteroids has been a major subject.