Jay Frothingham
AST 200
Project 3

Are brightness and orbits linked? An exploration of comets

Background

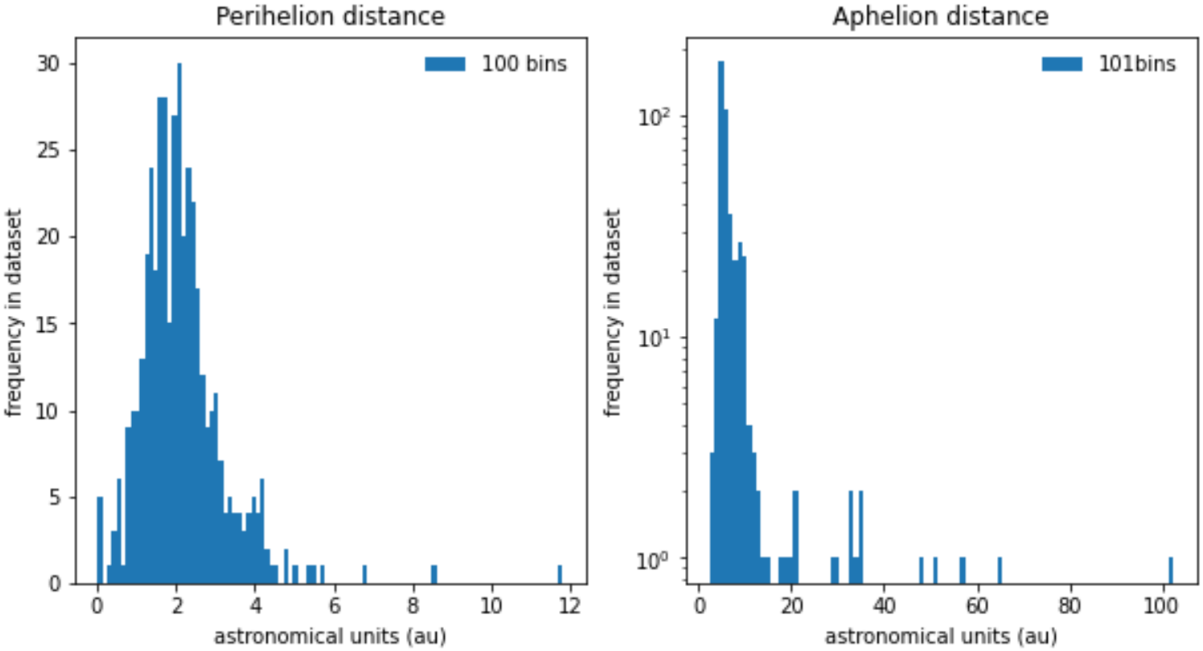
The data are drawn from the NASA Jet Propulsion Laboratory (JPL) small-body database. They are data describing comets rather than asteroids or other small bodies. Comets are small bodies generally made mostly of ice, dust, and carbon. As they approach the sun, they warm up and release gas, which forms a coma and/or a tail around the main core of the comet, its nucleus. Comets orbit the sun in a range of elliptical orbits,

The JPL small-body database utilizes NASA JPL's Horizons system to compute astronomical orbits for comets and calculate some physical parameters based on astrometrical and radar measurements. The data are collected from multiple instruments and professional as well as amateur observations.

The dataset used for this investigation contains information on 431 comets. It is limited to numbered comets, and excludes comet fragments. This is a volume-limited sample. According to the JPL small-body database website, "Numbers are only assigned to secure short-period (<200 year) comets." This means that the comets included in the dataset are more likely to have multiple observations and measurements used to determine their properties. While longer-period comets are valuable and interesting to consider, their longer periods mean that they may only have been observed once or twice.

This investigation will focus on the perihelion, aphelion, orbital period, orbital class, total magnitude, and nuclear magnitude. Perihelion describes the distance to the Sun when a comet is at the point in its orbit nearest to the Sun. The comets in this dataset have perihelion distances ranging from 0.039 astronomical units (au) to 11.784 au. Aphelion describes the distance to the Sun when a comet is at the point in its orbit farthest from the Sun. The comets in this dataset have aphelion distances ranging from 2.44 au to 101.73 au. Perihelion and aphelion can be calculated or modeled from other, observed orbital parameters.

Figure 1. Distributions of perihelion and aphelion distance within the dataset.



The perihelion distance is close to a Gaussian distribution with a mean of 2 au and a standard deviation of 1 au. The aphelion distance has more extreme outliers, but also approximates a Gaussian distribution, this time with both mean and standard deviation of about 7-8 au. Bins for each histogram were chosen to cover all integer values between the minimum and maximum distances for each variable.

Orbital period is the amount of time it takes a comet to complete a single orbit. The comets in this dataset have orbital periods ranging from 3.24 years to 365 years. The orbit classes of comets are defined by their orbital periods, and denoted with three-letter acronyms. The capitalization of the letters in the acronym is significant in differentiating between similarly-named classes--for instance, the two so-called Jupiter-family comet classifications JFc and JFC are two different classifications with slightly different definitions.

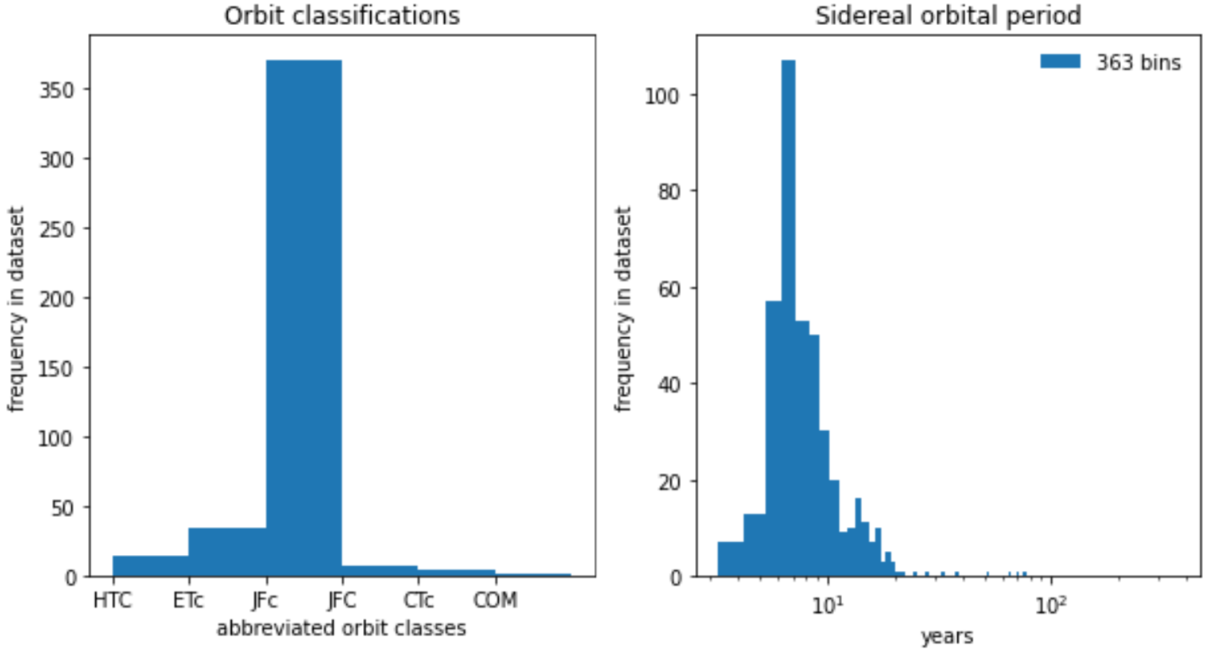
Figure 2: Table of orbital classifications present in the dataset.

Class	Name	Orbital period	Aphelion distance	# in dataset	Other notes
COM	n/a	n/a	n/a	1	Comets whose orbits don't fit into any existing classes
CTc	Chiron-type	$T > 3T_{Jupiter}$	$> ad_{Jupiter}$	4	
ETc	Encke-type	$T > 3T_{Jupiter}$	$< ad_{Jupiter}$	35	
HTC	Halley-type	20yr < T < 200 yr	n/a	14	

JFc	Jupiter-family	$2T_{Jupiter} < T < 3T_{Jupiter}$	n/a	370	As defined by Levison and Duncan
JFC	Jupiter-family	T < 20 years	n/a	20	As defined classically

*Orbital period of Jupiter $T_{Jupiter}$ = 11.87 years aphelion distance $ad_{Jupiter}$ = 5.46 au

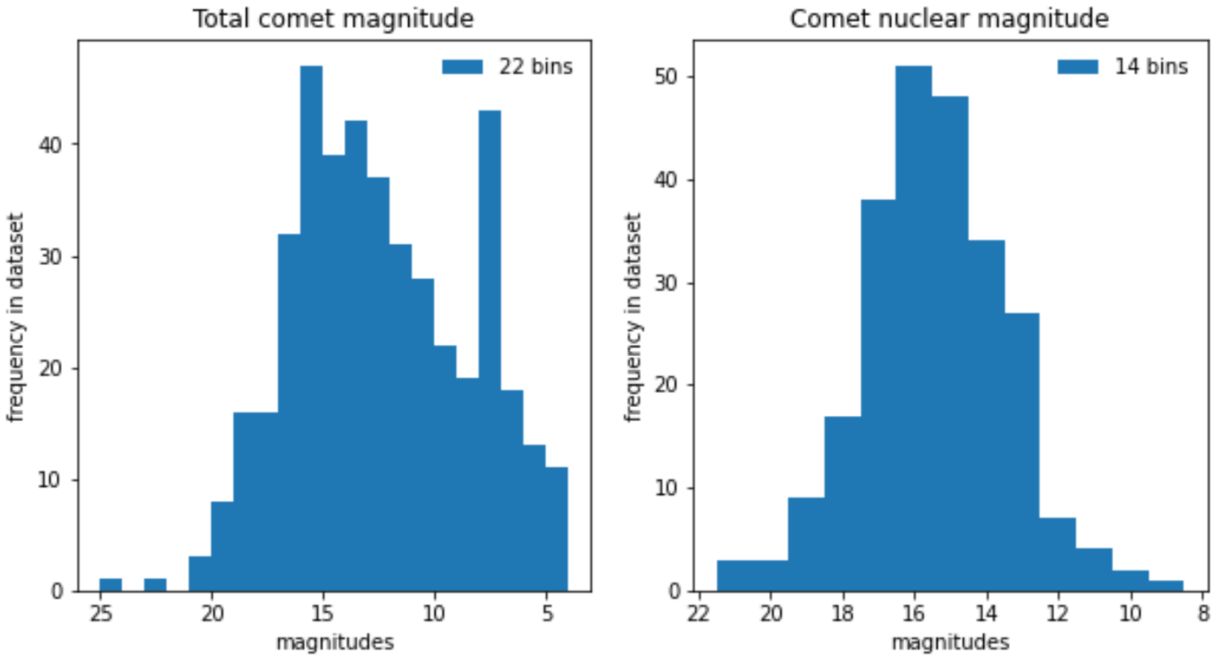
Figure 3. Distributions of orbital classification and sidereal orbital period within the dataset.



The histogram on the left shows that there are significantly more JFc comets in the dataset than other comet classifications. Bins were assigned such that there is one bin for each unique classification. It is not surprising that there are few CTc comets, as these have a longer orbital period and may not be included in this dataset (which is limited to numbered comets). To put the classification distribution into context, the histogram on the left shows the distribution of orbital periods. As expected, there are very few long-period comets and significantly more low-period comets present in the dataset. Bins were assigned to cover all integer values between the minimum and maximum period.

The total magnitude of a comet is the combined brightness of its nucleus and coma. It can be determined through photometry of the comet. The comets in this dataset have total magnitudes ranging from 4.0 to 24.3. In the magnitude system, smaller magnitudes denote brighter objects than larger magnitudes. The nuclear magnitude of a comet is the brightness of only its nucleus. The comets in this dataset have nuclear magnitudes ranging from 8.5 to 21.5. Nuclear magnitude is not determined directly, but it can be estimated from theoretical models of how much the nucleus contributes to the total magnitude.

Figure 4. Distributions of total and nuclear magnitude within the dataset.



The histogram on the left, showing the distribution of total magnitudes, very loosely follows a Gaussian distribution. The slope of the distribution for brighter magnitudes (<12) is more gradual, while the slope for dimmer magnitudes (>12) is more abrupt. There are two spikes around 8 magnitudes and 16 magnitudes, though the mean of the distribution is around 12 magnitudes. The histogram on the right, showing the distribution of nuclear magnitudes, follows a Gaussian distribution with a mean of 15 magnitudes and standard deviation of 2 magnitudes. Bins for both histograms were chosen to cover all integer values between the minimum and maximum magnitudes for each variable.

The dataset may be biased towards brighter comets with short orbital periods. Brighter comets are easier to detect by photometry. Shorter-period comets are more likely to have a higher number of confirmed observations, so there will be more of them present in the dataset, and the error associated with their data may be smaller.

To come up with my investigation, I chose a few variables that interested me from the dataset, printed statistical information about each, and plotted their distributions. I found that some of the variables really only had a few data points, and so would not be suitable for an investigation. For instance, many of the comets in the dataset did not have data for the rotational period, or the amount of time it takes for the comet to rotate about its axis.

I made initial scatter plots of aphelion distance and perihelion distance plotted against total magnitude and nuclear magnitude, separated into populations by orbital class. From these plots, I decided to investigate whether comet brightness is related to either its orbital period or aphelion and perihelion distances, and whether there are different relationships for total comet magnitude versus nuclear magnitude.

Procedure

To complete this investigation, I plotted the perihelion distance against nuclear and total magnitude and calculated correlation coefficients for each potential relationship to check whether it would be viable to proceed with curve fitting.

I fitted a quadratic function to the relationship between nuclear magnitude and total magnitude and evaluated its goodness of fit based on its residuals. I fed the original total magnitude data into this function to calculate nuclear magnitudes predicted by this relation.

I fitted several different functions to the relationship between nuclear magnitude and perihelion distance and evaluated which fit functions best described the relationship based on their residuals. I chose a fit function and fed my calculated nuclear magnitudes into this function to calculate perihelion distances predicted by this and the previously described relations.

I plotted the calculated perihelion distances and the actual perihelion distances against the total magnitudes to compare whether the combined models for relationships between nuclear and total magnitude and between nuclear magnitude and perihelion distance were reflected in the data.

I also did a Monte Carlo simulation on the nuclear magnitude and perihelion distance relation by generating one thousand sample datasets drawn from normal, bivariate distributions of nuclear magnitudes and perihelion distances based on the actual dataset's ranges and covariance. I then fit a line to each sample dataset and made a histogram of the slopes. I compared the distribution of sample slopes to the linear slope fit to the original dataset.

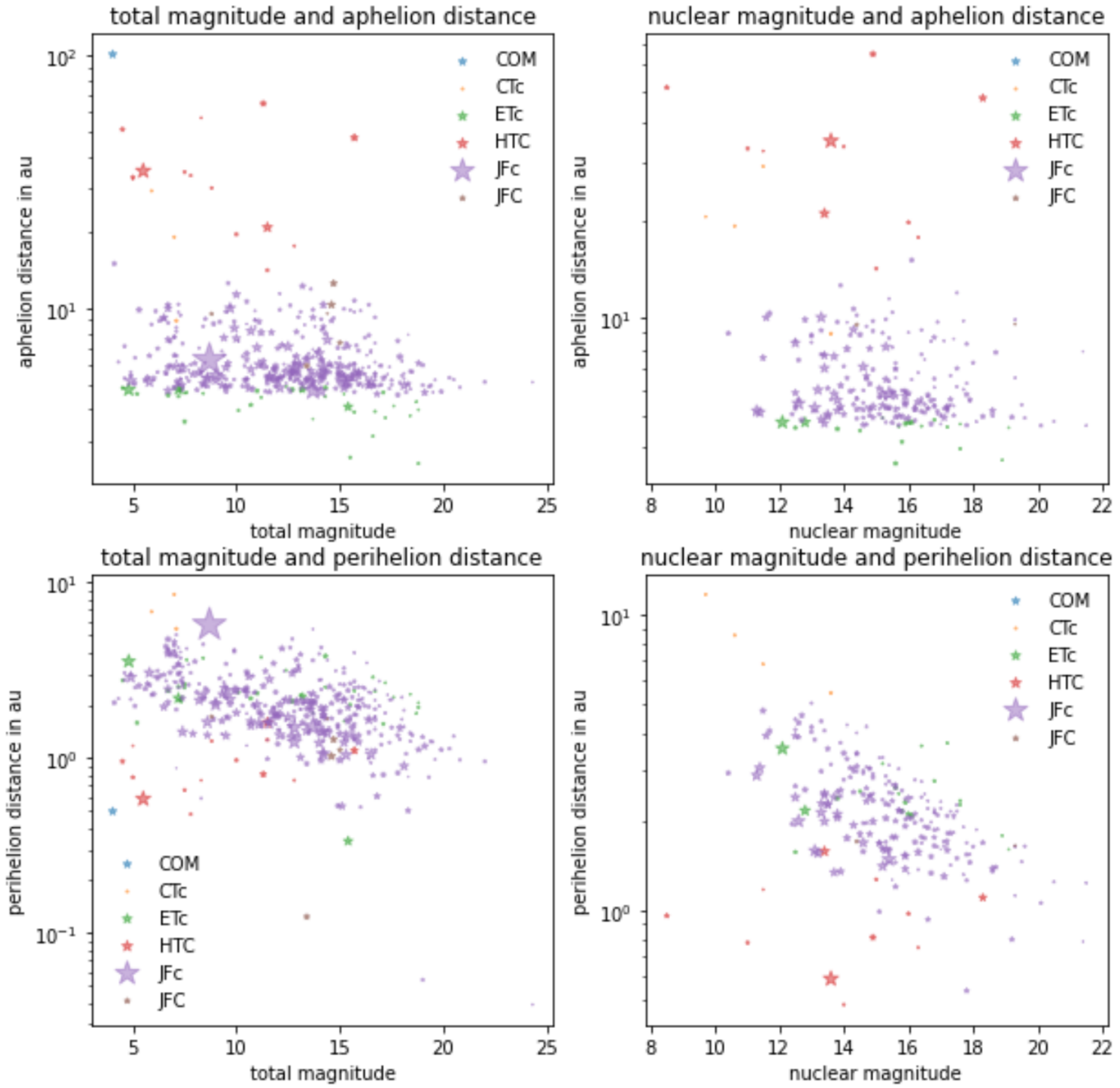
The code used to manipulate the data, calculate statistics, and generate plots was written in Python. It utilized the Python libraries `pandas` to load and manipulate the dataset, `numpy` and `scipy` for general numeric and statistical operations, and `matplotlib` for plotting, as well as a few custom-written functions to define linear and quadratic functions for curve fitting.

Another custom-written filtering function was used to separate the dataset into different populations by orbital class. This way, each population can be plotted separately in a different color to indicate whether any relations are also dependent on orbital period.

For curve fitting, a mask was applied to ignore NaN values present in the dataset. The scipy stats function `curve_fit` was used to optimize parameters for each of the curve fits.

Discussion and Analysis

Figure 5. Scatterplots comparing different combinations of magnitudes and distances.

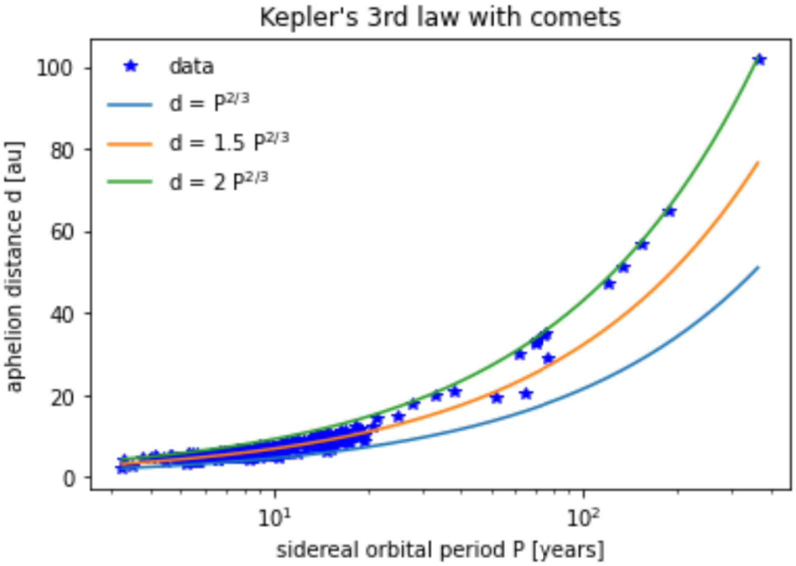


The size of each data point is proportional to the number of observations made of that particular comet. Well-observed comets are represented with larger data points than poorly-observed comets. Color is used to separate the data into different populations by orbital classification.

In Figure 5, the top two plots display aphelion distance against total magnitude and nuclear magnitude. There seems to be no relationship between aphelion distance and either type of magnitude, but there is a clear distinction in aphelion distance between different comet classes. Most notably, ETc comets have low aphelion distances, JFc comets have moderate aphelion distances, and HTc comets have high aphelion distances. Comet classification can serve as a proxy for orbital period, as orbital period is how comets are classified.

Given the distinction between different comet classifications' aphelion distances, I considered the idea that aphelion distance and orbital period are directly related. For a comet, the aphelion distance can approximate the semi-major axis of its elliptical orbit. Kepler's Third Law, though usually applied to planetary orbits, states that the square of a body's orbital period is proportional to the cube of its semi-major axis.

Figure 6. Orbital period plotted against aphelion distance and compared to Kepler's 3rd Law.



The scatterplot shows orbital period and aphelion distance, with orbital period plotted on a log scale. Several potential relations described by Kepler's Third Law are also plotted. Since Kepler's Third Law is a proportional relationship, the exact constant of proportionality is unknown.

From visual inspection of the plotted fits, it appears that the comets in this dataset obey Kepler's Third Law with a proportionality constant between 1 and 2.

Returning to the scatterplots in Figure 5, the bottom two plots show a possible linear relationship between magnitude and perihelion distance. The slope of the relationship is different for each plotted type of magnitude, but both plots show the same qualitative trend of perihelion distance decreasing for increasing (i.e., dimmer) magnitudes.

I calculated the Pearson and Spearman correlation coefficients for both possible relations, and performed hypothesis tests. The Pearson correlation coefficient assumes both variables are normally distributed, while the Spearman correlation coefficient does not. For both correlation coefficients, a value of -1 indicates a perfect linear relationship with a negative slope between the two variables, while a value of 1 indicates a perfect linear relationship with a positive slope, and a value of 0 indicates no correlation. The hypothesis tests were performed with the null hypothesis of no correlation between variables.

For the relation between total magnitude and perihelion distance, the Pearson correlation coefficient is -0.502, with a p-value of 6.25e-17. The Spearman correlation coefficient is -0.583, with a p-value of 1.47e-23. In both cases, the correlation coefficient indicates a possible linear fit with a negative slope, but not a perfect fit. The p-values are low enough to indicate rejection of the null hypothesis, so there is some correlation between total magnitude and perihelion distance.

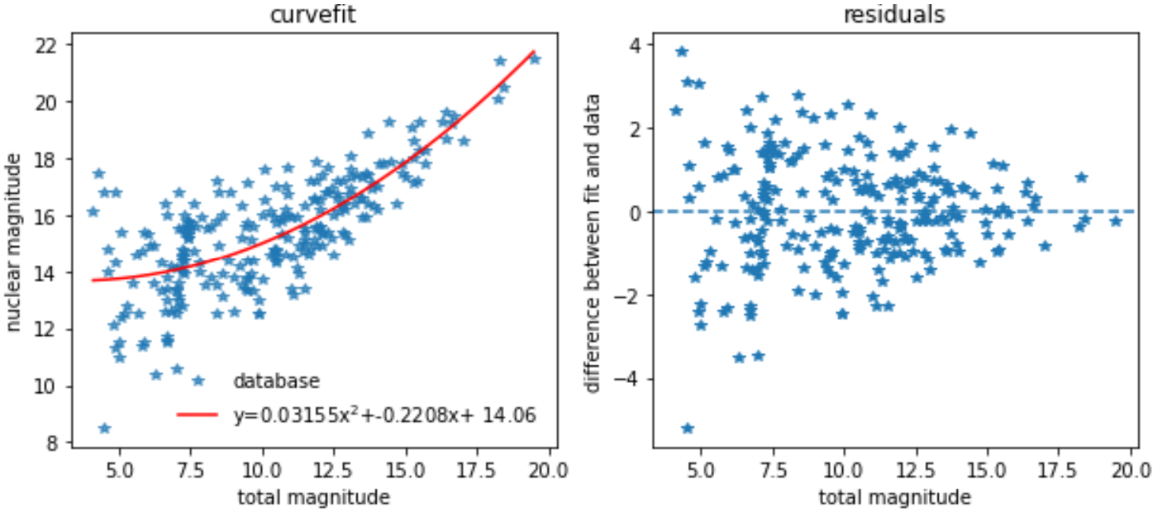
For the relation between nuclear magnitude and perihelion distance, the Pearson correlation coefficient is -0.511, with a p-value of 1.55e-17. The Spearman correlation coefficient is -0.505, with a p-value of 3.74e-17. In both cases, the correlation coefficient indicates a possible linear fit with a negative slope, but not a perfect fit. The p-values are low enough to indicate rejection of the null hypothesis, so there is some correlation between nuclear magnitude and perihelion distance.

The presence of a shared qualitative trend in the data for both nuclear and total magnitude suggests that there might be an additional relationship between nuclear magnitude and total magnitude that accounts for the quantitative differences in the trend.

I again calculated Pearson and Spearman correlation coefficients to evaluate whether there is correlation between total magnitude and nuclear magnitude, and hypothesis tests with a null hypothesis of no correlation between the variables. The Pearson correlation coefficient is 0.731, with a p-value of 7.605e-42. The Spearman correlation coefficient is 0.699, with a p-value of 5.371e-37. In both cases, the correlation coefficient indicates a linear fit with a negative slope, but not a perfect fit. The p-values are low enough to indicate rejection of the null hypothesis, so there is some correlation between nuclear magnitude and total magnitude.

From visual inspection of the plotted data, I decided to fit a quadratic function to the relationship between nuclear magnitude and total magnitude.

Figure 7. Total magnitude plotted against nuclear magnitude, with a quadratic curve fit.

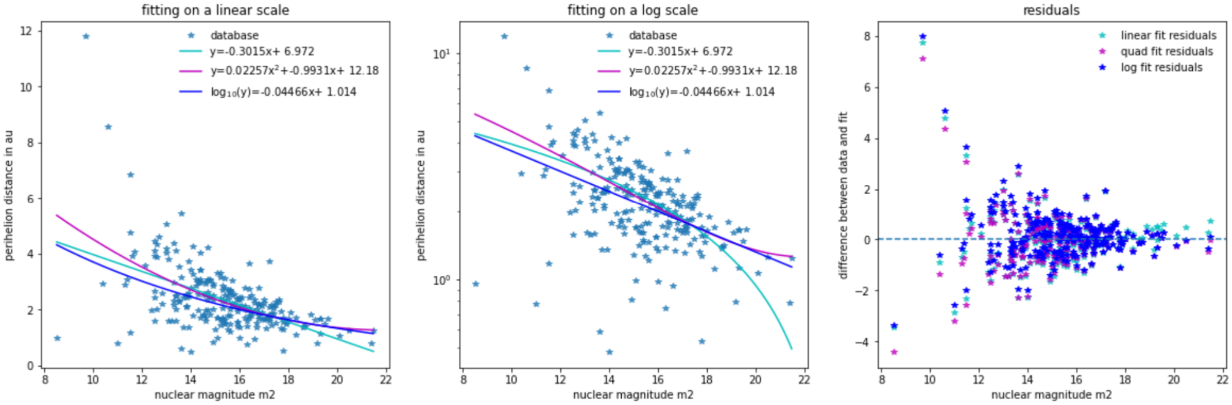


The figure shows the optimized quadratic fit function describing the relationship between nuclear magnitude and total magnitude, plotted over the data. Residuals indicate decreasing spread with increasing magnitude, but that the fit is centered within the data rather than skewing high or low.

There are fewer data points at higher magnitudes, as dimmer comets may be more difficult to detect. This means that the spread of data around the optimized curve fit decreases with increasing magnitudes, making it appear that the model is a better fit at higher magnitudes than at lower magnitudes. While this might be the case, it is also possible that more data collected at high magnitudes would display as much spread as data collected at low magnitudes. The residuals of the fit are evenly distributed above and below zero, indicating that the fit is not an overestimate or an underestimate. With this relationship, the approximate nuclear magnitude of a comet can be calculated from its total magnitude.

Several types of functions were plotted to try and describe the relationship between nuclear magnitude and perihelion distance. A simple linear fit between the two variables, a quadratic function, and a linear fit between the log (base 10) of the perihelion distance and the direct nuclear magnitude were plotted on both linear and log scales.

Figure 8. Nuclear magnitude plotted against perihelion distance, with several possible fits.



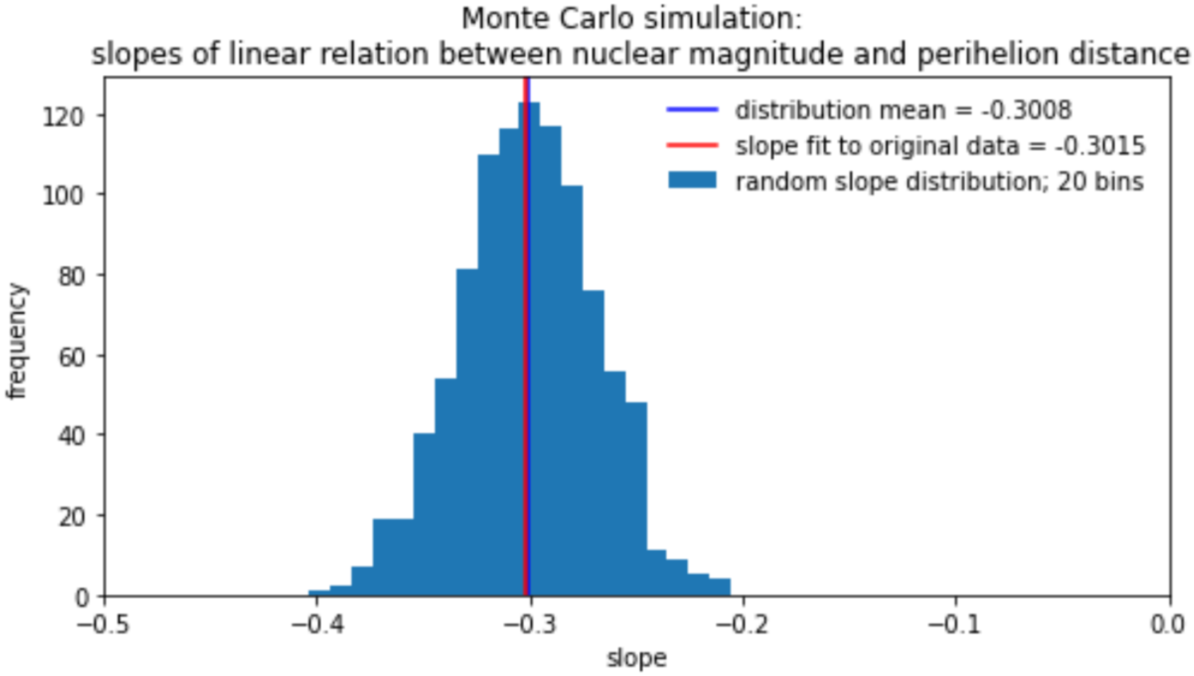
The figure shows several optimized possible fit functions describing the relationship between nuclear magnitude and total magnitude, plotted over the data. Residuals indicate decreasing spread with increasing magnitude, and that different fit functions skew higher or lower than others in different regions of the dataset.

These possible fits are all very similar in how well they describe the data, as there is so much spread present. They all describe a trend of decreasing perihelion distance with increasing (dimming) nuclear magnitude. For high-magnitude (dimmer) comets above 18 magnitudes, the linear fit predicts a much lower perihelion distance than the other two models, which predict very similar perihelion distances. For mid-magnitude comets of about 14-18 magnitudes, all three fits predict similar perihelion distances, and residuals are about equally distributed above and below the fit. For low-magnitude (brighter) comets lower than about 14 magnitudes, the quadratic fit predicts the lowest perihelion distance of the three plotted models, while the log fit predicts the highest perihelion distance. The comets with higher perihelion distances are closer to the models.

Visually, the log fit appears to match the densest regions of data best in both the linear and log scale. It is less affected by the possible outlier population of comets with low (below 10 au) perihelion distance.

To evaluate whether the relationship between nuclear magnitude and perihelion is real, I performed a Monte Carlo simulation. For simplicity, I used the linear fit model and stored the slopes of one thousand randomly generated sample datasets based on the ranges and covariance of the original dataset.

Figure 9. Monte Carlo simulation of linear relationship between nuclear magnitude and perihelion

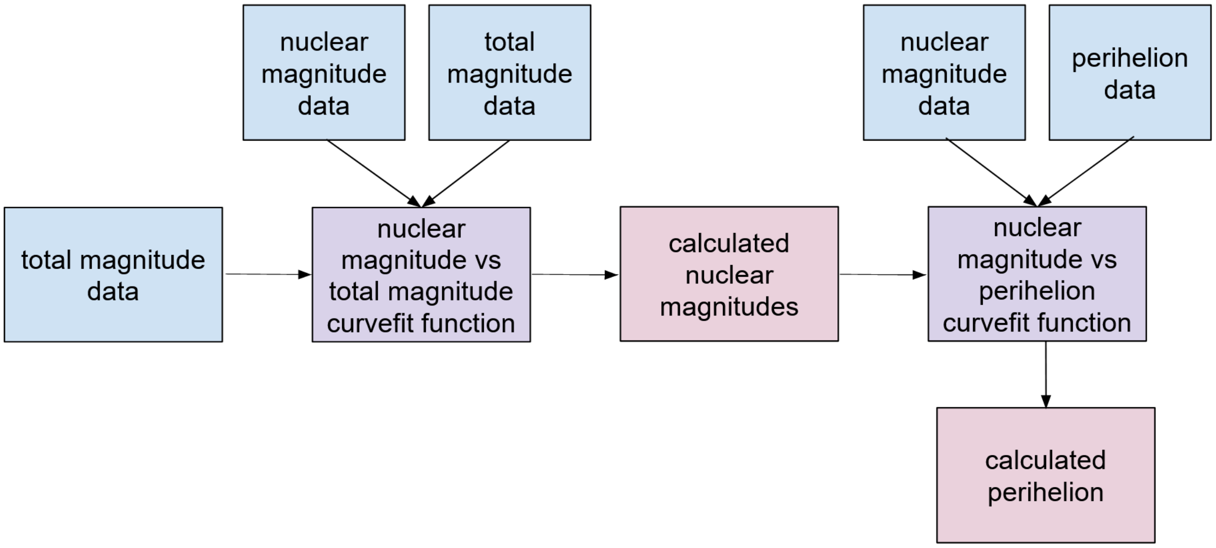


The histogram shows an approximately normal distribution of slopes, with a mean of about -0.3. The mean of the distribution is plotted in dark blue, and the slope of a line fit to the original dataset is plotted in red.

The slope of the line fit to the original dataset is so similar to the mean of the distribution of slopes produced in the Monte Carlo simulation that it is difficult to see both slopes indicated in FIgure 9. This indicates high agreement between the distribution and the originally optimized linear fit and suggests that the relationship between magnitude and perihelion distance is not due to random chance but is a real trend present in the dataset.

This trend is most distinct between nuclear magnitude and perihelion distance, but it is also present between total magnitude and perihelion distance. Rather than optimizing additional fit functions to describe the relationship, I used the data to evaluate the curve fits for previous relationships and attempt to describe the relationship between nuclear magnitude and perihelion distance in the process.

Figure 10. Process used to calculate perihelion distances from total magnitude.

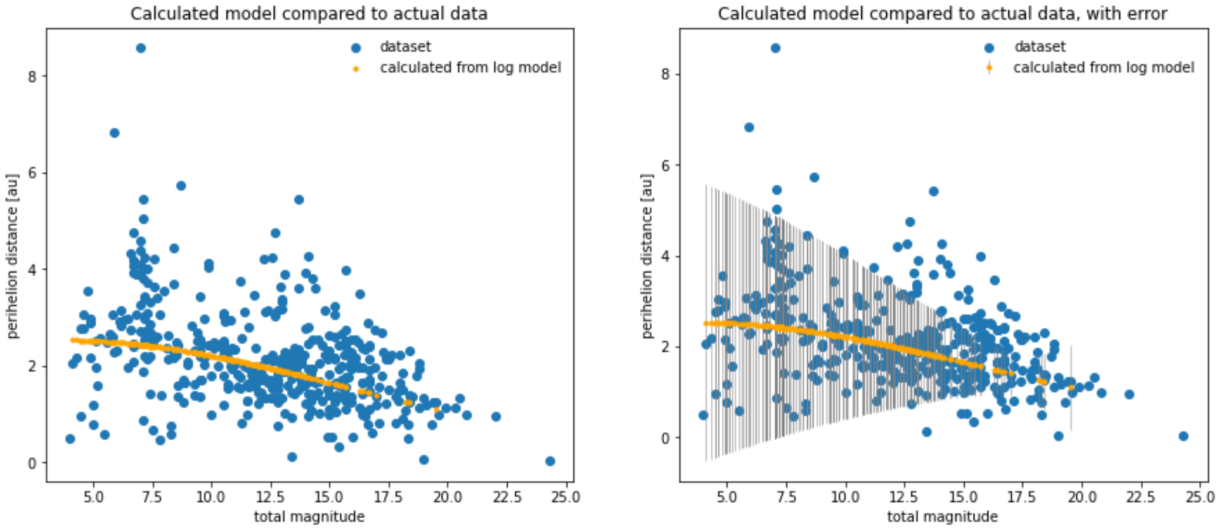

nuclear magnitude data
total magnitude data
nuclear magnitude data
perihelion data
total magnitude data
nuclear magnitude vs total magnitude curvefit function
calculated nuclear magnitudes
nuclear magnitude vs perihelion curvefit function
calculated perihelion

Blue indicates data drawn directly from the dataset. Purple indicates functions determined through optimized curve fitting. Red indicates calculated quantities meant to simulate data.

I fed total magnitude values from the original dataset into my optimized function describing the relationship between nuclear magnitude and total magnitude to calculate theoretical nuclear magnitude values. These are the nuclear magnitudes we would expect to see if the fit function accurately describes the relationship between nuclear and total magnitudes. I then fed those calculated nuclear magnitude values into the optimized log fit function describing the relationship between nuclear magnitude and perihelion distance to calculate theoretical perihelion distance values. These are the perihelion distances we would expect to see if 1) the first fit function accurately describes the relationship between nuclear and total magnitudes; 2) the second fit function accurately describes the relationship between nuclear magnitude and perihelion distance; and 3) there are no additional factors that account for the quantitative differences in the relationships between perihelion distance and each type of magnitude.

These calculated perihelion distances were then plotted against the original total magnitudes used at the start of the process, and compared to a scatter plot of the actual perihelion distances from the dataset plotted again against the total magnitudes from the dataset. Errors in the fit functions were estimated based on their residuals, and propagated in quadrature.

Figure 11. Total magnitude plotted against actual and calculated perihelion distances.



The figure shows perihelion distances from the dataset and calculated from several fit functions plotted against total magnitudes from the dataset. The plot on the right is shown with estimated uncertainties on the calculated data points.

At first glance, the simulated perihelion distances calculated through curve-fit models do not completely match the actual perihelion distances. The calculated data points also have large errors associated with them, making the shape of the modeled relation between total magnitude and perihelion distance less certain. There is spread present in the data, particularly for lower-magnitude (brighter) comets. However, the calculated data points visually appear centered and a good fit for the densest sections of the data, ignoring possible outliers with extremely high perihelion distances. This indicates that the models and assumptions used to calculate the perihelion distances may be accurate.

To conclude this investigation, the comets in this dataset display no relationship between total or nuclear magnitude and aphelion distance. They do display a relationship between the total and nuclear magnitudes, and between the perihelion distance and both types of magnitude. The relationship between nuclear and total magnitudes is likely a mixture of an actual astrophysical phenomenon and an artifact of any theoretical modeling and assumptions that went into determining the nuclear magnitudes included with the dataset.

Comets in this dataset do display a relationship between perihelion distance and magnitude, both nuclear and total. The brighter (lower-magnitude) comets in the dataset are those with larger perihelion distances, and the perihelion distances decrease with increasing (dimming) magnitudes. Physically, this may correspond to comets that pass closer to the sun losing material through their tails or other destructive processes. It may also be due to observational bias. Dimmer objects are more difficult to observe, and there very well may be high-magnitude comets with large perihelion distances whose orbits take them too far from the Earth to be easily observable. More work is needed to determine whether this is the case, perhaps by applying the principle that brightness decreases with distance squared.