Analyzing Trend between Nebula's Proper Motions and its Distance from the Solar System's Barycenter

Introduction:

This research project examines how the absolute distance between a nebula and its solar system's barycenter affects the absolute average of its proper motions. This project will involve data analysis, data generation, preprocessing my data, fitting data with errors, and providing explanations for the model I end up using to represent my data.

Chosen Phenomenon and Data Source:

The phenomenon I chose to study is the distribution of nebulae in our solar system. I plan to pull data from SIMBAD Astronomical Database - CDS (Strasbourg) which contains data on each nebula's Object Type Identifier, Typed Identifier, Identifier, ICRS Coordinates, FK4 Coordinates, Gal Coordinates, Proper Motions, Radial Velocity, Redshift, CZ, and Notes. This dataset can be accessed at [SIMBAD Astronomical Database - CDS (Strasbourg)] (http://simbad.cds.unistra.fr/simbad/).

Equations to fit the data:

Exponential:

$\mu(x) = Ae^{\lambda x}$

$\mu(x)$: Nebula's Predicted Proper Motions (mas/yr), A: Absolute Average of the Proper Motions of the Nebula closest to the Solar System's Barycenter, e: Base of exponential, λ: Growth Constant, x:Parameter of Interest

Natural Log:

$\mu(x) = ln(x + \mathrm{h}) + \mathrm{k}$

$\mu(x)$: Nebula's Predicted Proper Motions (mas/yr), x: Parameter of Interest, h: Absolute Distance of Nebula Closest to the Solar System's Barycenter, k: Absolute Average of the Proper Motions of the Nebula closest to the Solar System's Barycenter

Linear Function:

$\mu(x) = B(x - h) + k$

B: Rate of Change, x: Parameter of Interest, h: Absolute Distance of Nebula Closest to the Solar System's Barycenter, k: Absolute Average of the Proper Motions of the Nebula closest to the Solar System's Barycenter

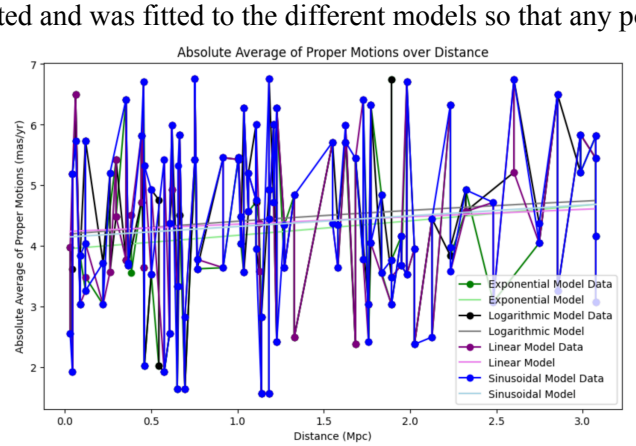
Sinusoidal Function:

$\mu(x) = -asin(bx + c) + d$

a: Rough Amplitude of Data, b: Rate of Change, x: Parameter of Interest, c: Phase Shift, d: Vertical Shift

Data Generation:

Random data was generated and was fitted to the different models so that any potential trends in the data could be visualized:



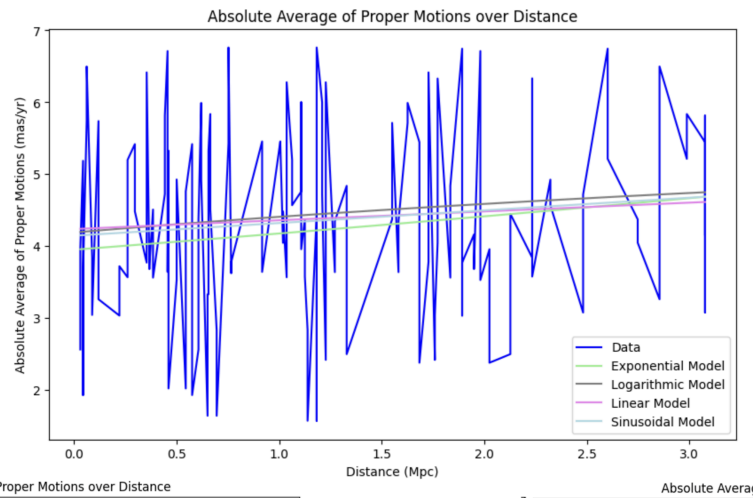
Data Filtering:

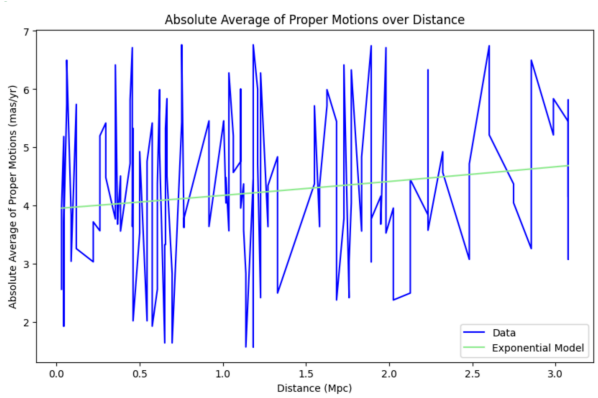
To ensure that I used quality data from the database, I applied various quality filters to filter out any outliers or missing data

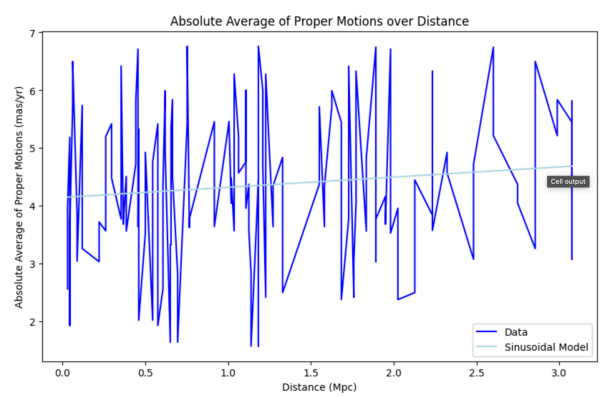
1. Filtered out data that did not have Proper Motions data and Redshift data
2. Filter data to those labeled with its Identifier being 'PN'
3. Used box plot to identify and remove outliers that would skew the data and don't represent the general trends of the population
4. Removed any data that had notes in case it would skew my data

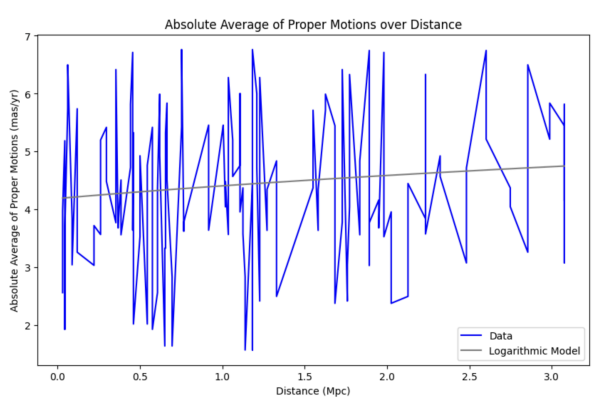
Data Fitting with Error:

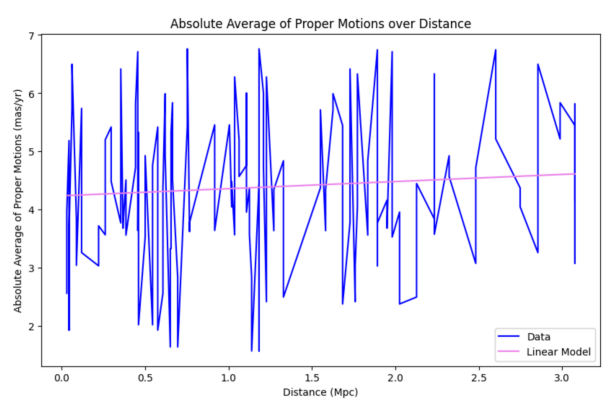
To analyze the correlation between a nebula's average distance from the solar system's barycenter and the absolute average of its proper motions, I fit different splits of the synthetic and existing data to the various mathematical models described above and calculated their respective Reduced Chi-Squared values to determine which model fits the data best.











Out of all of the models, the Reduced Chi-Squared value of the fitted Logarithmic Model most closely resembles the data as it is the value closest to 1. Therefore, the model that should be used to represent the data is the Logarithmic Model.

Explanation of Model Fit:

After fitting my models to their respective data, these were the resulting equations.

Exponential: $\mu(x) = 4.03380491(1.45012233)^{0.16184563x}$

4.03380491 corresponded to the Absolute Average of the Proper Motions for the nebula closest to the Solar System's Barycenter. 1.45012233 corresponds to the base of the exponential and since it is greater than 1 that means that there is an exponential growth in Proper Motions as distance from the Solar System's Barycenter increases. 0.16184563 corresponds to the growth constant of the exponential and tells how quickly the PM grows as distance increases.

Natural Log: $\mu(x) = ln(x + 23.97881601) + 1.11887011$

23.97881601 corresponds to a horizontal shift in the graph meaning that the data does not follow the curve of a logarithmic graph. 1.11887011 corresponds to the Absolute Average of the Proper Motions for the nebula closest to the Solar System's Barycenter. As there are no growth constants in this function, it assumes that the PMs' of the nebulas follow a logarithmic trend as the distance from the Solar Systems' Barycenter increases.

Linear Function: $\mu(x) = 0.26763252(x - (-2.21734118)) + 3.40020863$

0.26763252 corresponds to the rate of change of a nebula's PM as its distance from the Solar System's Barycenter increases and because it is greater than 0 that indicates that there is a positive association because the two variables. -2.21734118 corresponds to the horizontal shift of the function. 3.40020863 corresponds to the Absolute Average of the Proper Motions for the nebula closest to the Solar System's Barycenter.

Sinusoidal Function: $\mu(x) = -0.35515987sin(1.00694828x + 1.91250506) + 4.31822858$

0.35515987 corresponds to the rough Amplitude of the data or how much the PM varies. 1.00694828 corresponds to the period of the sinusoid or the distance it takes for the data to complete oscillate. 1.91250506 corresponds to the phase shift meaning the data does not start at PM = 0. 4.31822858 corresponds to the Absolute Average of the Proper Motions for the nebula closest to the Solar System's Barycenter.

Conclusion:

Because the data graph did not seem to have a consistent pattern, none of the models did a very good job of representing the data. While the Logarithmic Model was the closest to representing the data, its Reduced Chi-Squared value was still very far from the value of 1. Despite this, I did notice a consistent positive trend in the Proper Motions as the Distance increased. Throughout this project, there were many ups and downs. It was especially difficult getting started with the project as cleaning the data proved to be a challenge because of the odd symbols used to fill in missing data as well as because all of the values were of the Object type so I had to manipulate their types to use them in functions and visualizations. However, after I finished preparing my data, the rest was smooth sailing for the most part. In terms of future improvements, it would be nice to have more existing data to fit the models too. In addition, more research should be done on different potential models for representing the data as this project did not include many and none of the materials matched the data very well