### **Group #1 - Predictive Analysis of Near-Earth Comets using NASA JPL’s Small-Body Database**

Troy Allen, Roy Gabriel, Nattakorn Kittisut, Corey Smith & **Moe Kyaw Thu (Team Lead)**

**Project Proposal**:

The National Aeronautics and Space Administration (NASA) tracks numerous different astronomical phenomena and celestial bodies to gain insights into the workings of our universe and to predict potentially catastrophic events. In particular, NASA’s Jet Propulsion Laboratory (JPL) focuses substantial effort on tracking, categorizing, and documenting near-Earth asteroids and comets to perform various science analyses and predict possible Earth impact. This research will attempt to identify algorithmic correlations between various scientific parameters related to near-Earth comets (NECs) based on regression techniques learned in this course.

**Research Questions**

1. Which orbital elements have the most significant impact on predicting the perihelion distance (q) of near-Earth comets?
2. How accurately can a multiple linear regression model predict the semi-major axis (a) based on other orbital elements?
3. What is the relationship between the eccentricity (e) and the mean anomaly (M) of near-Earth comets, and can this relationship be accurately modeled using regression techniques?

**Data Description and Methodologies**

In light of the research questions, this study uses the dataset titled “Near-Earth Comets - Orbital Elements”. This dataset provides J2000 heliocentric ecliptic orbital elements for NECs. In total, it has 17 attributes (columns) along with 160 observations. The attributes provided in the dataset include essential orbital elements such as the semi-major axis, eccentricity, inclination, the longitude of the ascending node, the argument of perihelion, mean anomaly, and so on. Collectively, they are all crucial for understanding the trajectories and behavior of these comets as they pass close to Earth.

The following table describes the codebook for the dataset:

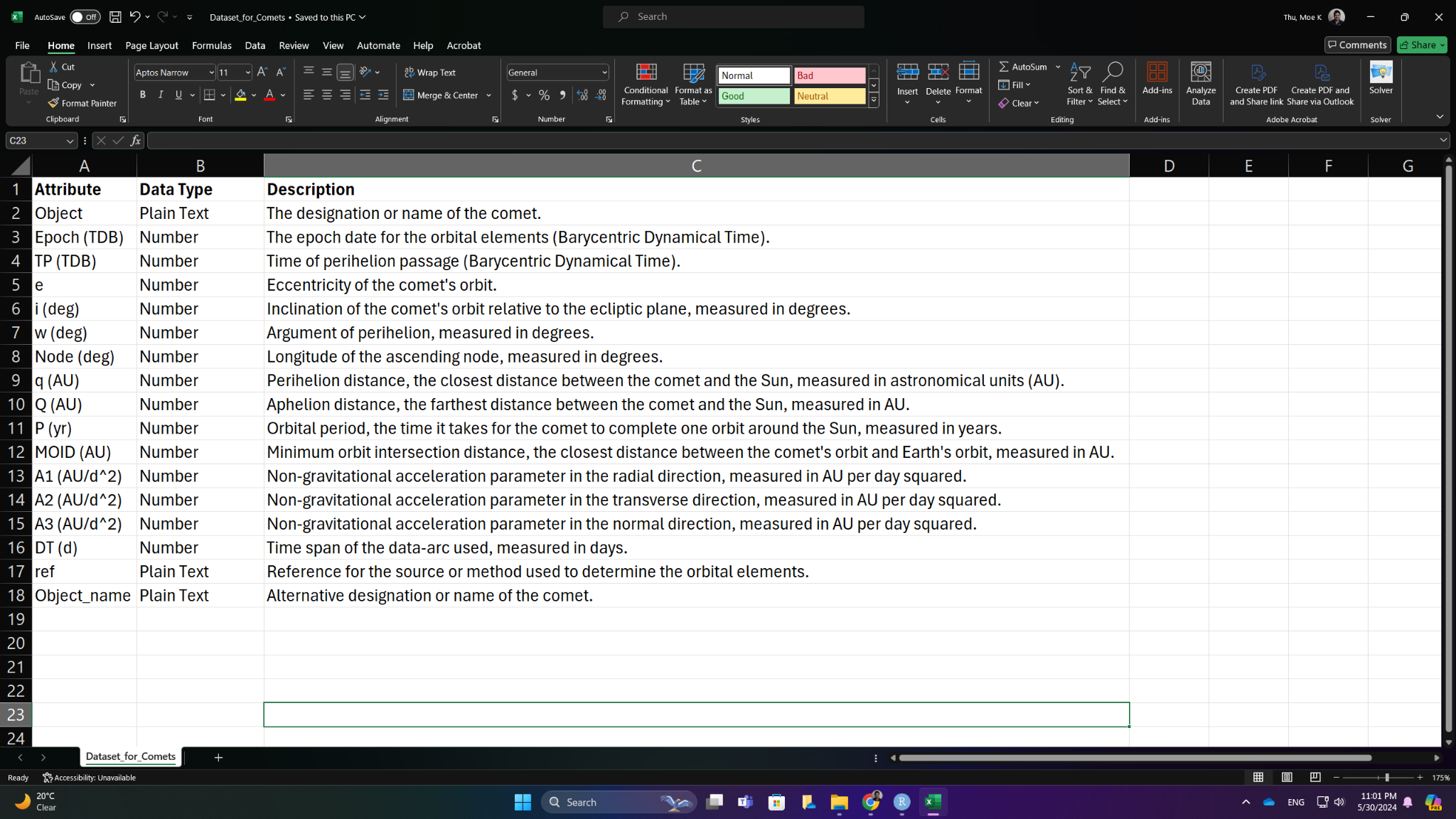


Fig. Codebook for Dataset

Based on this dataset, the methodology in this study involves several key steps. First and foremost, data cleaning and preprocessing, including the use of parsing, are performed to ensure the quality and integrity of the dataset. This includes handling missing values, detecting outliers, standardizing data format, and visualizing.

In the second stage, following data preprocessing, linear regression models are developed to understand the relationship between the response and the predictor variables. Model diagnostics, particularly detecting leverage points, normality, and non-constant variance testing for residuals are conducted to evaluate the performance and accuracy of the regression models. Additionally, various visualization techniques, particularly the use of Q-Q plots, are utilized to assess the normality of residuals and the overall goodness-of-fit of the models.

The final stage will involve applying inferential analysis techniques to the dataset to understand the relationship between the response and the predictor variables. In particular, this approach will focus on the use of stepwise regression and regularization methods, such as LASSO and Ridge regressions. These inferential techniques will help to uncover the patterns and trends within the dataset as well as feature importance and handling overfitting.

**Appendix**

This is to answer the questions as in the proposal requirement.

1. Are the data points time-dependent, i.e., were they collected at different time points (year, month, day, etc.)?

Yes, the data points are time-dependent. The orbital elements provided in the dataset are based on the J2000 epoch, and the time of perihelion (Tp) is given in Julian Date.

1. What kinds of data will be required to represent the independent variables, i.e., which variables are quantitative and which are qualitative? What are the possible categories of the qualitative variables? The number of variables as well as their types (qualitative/quantitative) should be listed. Explicitly define the variables.

There are two types of variables as below:

* Quantitative Variables:
  1. Semi-major axis (a)
  2. Eccentricity (e)
  3. Inclination (i)
  4. Longitude of the ascending node (Node)
  5. Argument of perihelion (Peri)
  6. Mean anomaly (M)
  7. Perihelion distance (q)
  8. Aphelion distance (Q)
  9. Orbital period (P)
  10. Mean daily motion (n)
  11. Time of perihelion (Tp)
  12. Epoch of the osculating elements (Epoch)
  13. Number of observations (NumObs)
  14. Number of oppositions (NumOpp)
* Qualitative Variables:
  1. Object Name
  2. Reference (Ref)
  3. Source

1. How will you collect the data (i.e., source of data)? Will you physically collect the data, or will you get the data from a source, e.g., the internet, a publication, etc? Since it may be difficult to collect physically, you can search the literature/internet, and you can use some published data by giving a direct reference (Web page, URL address, publication, etc.).

The data is collected from NASA’s Jet Propulsion Laboratory Small-Body Database. The dataset can be accessed directly via the provided link: NASA JPL Near-Earth Comets - Orbital Elements.

1. What is the sample size (at least approximately)? The appropriate sample size and the number of variables may change from one problem to another. Usually, you need at least 200 data points (roughly) to perform several tasks. You may even consider having around 10 variables (again, roughly).

The dataset comprises 160 observations with 17 attributes (variables) each. This sample size is sufficient for performing multiple regression tasks and ensuring robust analysis. If additional observations are required to improve the accuracy of the regressors, then the Small-Body Database Query can be used to get these extra points.