Big Data Analysis of NASA's 5 Millenium Solar Eclipse Database

Soham Phanse
Indian Institute of Technology Bombay, 19D170030@iitb.ac.in

Abstract **- Solar eclipses are a topic of interest among astronomers, astrologers, and the general public. There were and will be about 11898 eclipses in the 5 millennia from 2000 BC to 3000 AD. Data visualization and regression techniques offer a deep insight into how various parameters of a solar eclipse are related to each other. Physical models can be verified and can be updated based on the insights gained from the analysis. The study covers the major aspects of data analysis, including data cleaning, pre-processing, EDA, distribution fitting, regression, and machine learning-based data analytics.**

Index Terms - Big Data Analysis, Space Visualization, Data Analytics

Introduction

I. What is a Solar Eclipse?

A solar eclipse occurs when a portion of the Earth is engulfed in a shadow cast by the Moon, which fully or partially blocks sunlight. This occurs when the Sun, Moon, and Earth are aligned in a straight line, commonly called a syzygy by astronomers. Such alignment coincides with a new moon indicating the Moon is closest to the ecliptic plane - which is the plane of the Earth's elliptical orbit around the Sun. Eclipses are broadly classified into - Total, Annular, Partial and Hybrid eclipses. These further are subdivided into classes that satisfy some parameters. Eclipses are also classified as Central and Non-Central based on if the central line of the umbra touches the Earth's surface. In a total eclipse, the disk of the Sun is fully obscured by the Moon. In partial and annular eclipses, only part of the Sun is obscured.

II. NASA's 5 Millenium Solar Eclipse Database

NASA's 5 Millenium Solar Eclipse Database (Espenak and Meeus) is a catalog of solar eclipses over 5 millennia, i.e., from 2000 BC to 3000 AD, that summarizes the principal characteristics of each solar eclipse over the time interval. According to the catalog, Earth will experience 11,898 eclipses of the Sun during the 5000-year period from –1999 to +3000 (2000 BCE to 3000 CE). The coordinates of the Sun used in these predictions are based on the VSOP87 theory [Bretagnon and Francou, 1988]. The Moon's coordinates are based on the ELP-2000/82 theory [Chapront-Touze and Chapront, 1983]. Here is an excerpt from the dataset.

	1	-1999	6	12	0	11691	46438	-49456	5	T	-0.2701	1.0733	6	-33.3	74	344	247	397
1	1	-1999	6	12	0	11691	46438	-49456	5	T	-0.2701	1.0733	6	-33.3	74	344	247	397
2	1	-1999	12	5	176	85523	46426	-49450	10	A	-0.2317	0.9382	-32.9	10.8	76	21	236	404
3	1	-1998	6	1	187	65356	46415	-49444	15	T	0.4994	1.0284	46.2	83.4	60	151	111	135
4	1	-1998	11	25	177	21423	46403	-49438	20	A	-0.9045	0.9806	-67.8	-143.8	25	74	162	74
5	1	-1997	4	22	217	47996	46393	-49433	-13	P	-1.467	0.1611	-60.6	-106.4		281		

FIGURE I
FIRST 6 ROWS OF NASA'S 5 MILLENIUM SOLAR ECLIPSE DATABASE

We briefly describe each of the variables in the dataset. (NASA GSFC).

- **Catalog Number:** Sequential number of the eclipse in the catalog links to the map published in the Five Millennium Canon of Solar Eclipses: -1999 (2000 BC) to +3000 AD
- **Eclipse Date:**
  - Calendar Date at the instant of Greatest Eclipse.
  - Gregorian Calendar is used for dates after 1582 Oct 15. - Julian Calendar is used for dates before 1582 Oct 04.
- **TD of Greatest Eclipse:** Dynamical Time (TD) of the Greatest Eclipse, the instant when the axis of the Moon's shadow cone passes the closest Eclipse to Earth's center.
- **ΔT(s) or DT(s)**
  - Delta T (ΔT) is the arithmetic difference between Dynamical Time and Universal Time. It measures the accumulated clock error due to the variable rotation period of Earth.
  - The orbital positions of the Sun and Moon required by eclipse predictions are

calculated using Terrestrial Dynamical Time (TD) because it is a uniform time scale.- However, world time zones and daily life are based on Universal Time (UT).
    - The difference between these two-time scales must be known to convert eclipse predictions from TD to UT.
    - The parameter delta-T (ΔT) is the arithmetic difference, in seconds, between the two as ΔT=TD−UT.
- **Lunation Number:** Lunation Number is the number of synodic months since the New Moon of 2000 Jan 06. The Brown Lunation Number can be determined by adding 953.
- **Saros Num:** Saros series number of eclipses. (Each eclipse in a Saros is separated by an interval of 18 years 11.3 days.)
- **Eclipse Type: First Character**
    - P = Partial Eclipse.
    - A = Annular Eclipse.
    - T = Total Eclipse.
    - H = Hybrid or Annular/Total Eclipse.
- **Eclipse Type: Second Character**
    - m = Middle eclipse of Saros series.
    - n = Central eclipse with no northern limit.
    - s = Central eclipse with no southern limit.
    - + = Non-central eclipse with no northern limit.
    - - = Non-central eclipse with no southern limit.
    - 2 = Hybrid path begins total and ends annular.
    - 3 = Hybrid path begins annular and ends total.
    - b = Saros series begins (first eclipse in series).
    - e = Saros series ends (last eclipse in series).
- **QLE**: Quincena Lunar Eclipse parameter identifies the type of lunar eclipse that precedes and/or succeeds a solar eclipse where:
    - n = penumbral lunar eclipse (Moon passes partly or completely within Earth's penumbral shadow)
    - p = partial lunar eclipse (Moon passes partly within Earth's umbral shadow)
    - t = total lunar eclipse (Moon passes completely within Earth's umbral shadow)
- **Gamma:** Distance of the shadow cone axis from the center of Earth (units of equatorial radii) at the instant of greatest eclipse.
- **Eclipse Magnitude:** Eclipse magnitude is the fraction of the Sun's diameter obscured by the Moon. For annular, total, and hybrid eclipses, this value is actually the diameter ratio of the Moon/Sun.
- **Latitude:**
    - The latitude where the greatest eclipse is seen.
    - Only eclipses with non-zero central duration are considered in the following analysis.
    - The geographic latitude and longitude corresponding to the position of the greatest eclipse.
    - Negative values correspond to the Southern Hemisphere and Positive Values to the Northern Hemisphere.
    - '0' corresponds to the Equator.
- Longitude (in degrees): Longitude where the greatest eclipse is seen.
- Sun Altitude (in degrees): Sun's altitude at greatest eclipse
- Sun Azimuth (in degrees): Sun's azimuth at the greatest eclipse.
- Path width (in km): Width of the path of totality or annularity Width at greatest eclipse (kilometers)
- Total Central Duration (in seconds)
    - Central Line Duration of total or annular phase at greatest eclipse.
    - For central eclipses (total, annular, or hybrid), the central line duration of the total or annular phase (in minutes and seconds) is given at the geographic position intersected by the axis of the lunar shadow cone at the instant of the greatest eclipse.
    - In the case of a total or hybrid eclipse, this duration is nearly the maximum duration of the total phase along the umbral path.
    - For an annular eclipse, the duration at the greatest eclipse may be near either the minimum or maximum duration of the annular phase along the path.

Exploratory Data Analysis

I. Data Visualization

We adopt the standard data analysis procedure and focus on exploratory data analysis to gain more insights and intuitively make sense of the data. Histograms, bar plots, line plots, scatter plots, and 2D histograms have been generously used. Then we fit the data with statistical distributions to understand the underlying data statistics and help predict parameters when unknown. We also obtain the probabilistic distribution of the data to find the probability of that parameter is equal to a certain value or, better off, lying in some range of values.

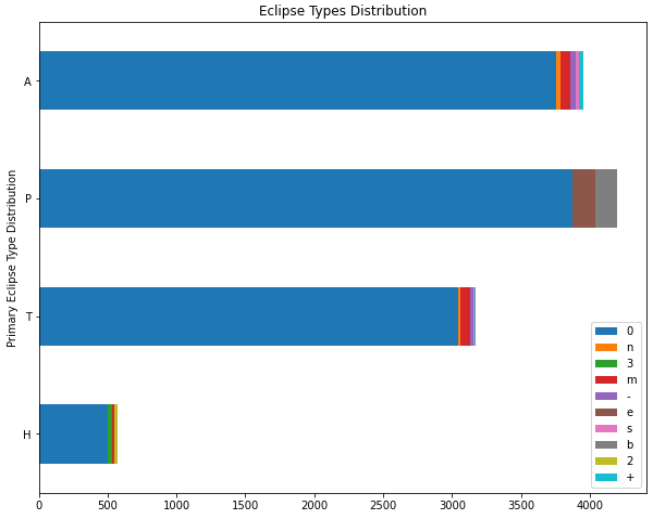

FIGURE II
DISTRIBUTION OF PRIMARY AND SECONDARY ECLIPSE TYPES

The plot describes the primary eclipse types and their further subdivisions into classes with a bar plot. It is evident that partial eclipses are the most common as compared to others.

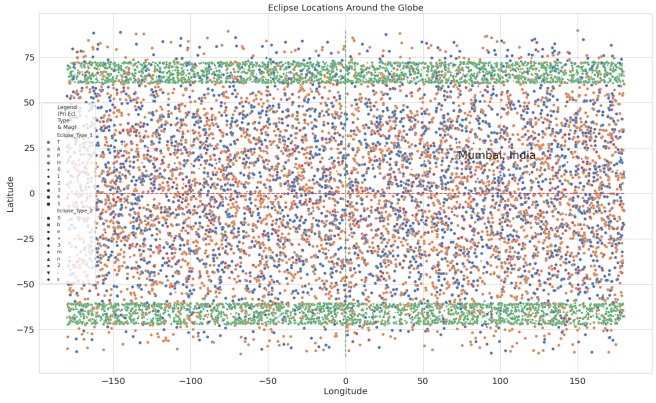

FIGURE III
DISTRIBUTION OF ECLIPSE LOCATIONS AROUND THE GLOBE

The locations across the globe that witness the greatest eclipse, i.e., locations where the axis of the moon's umbra intersects with the earth, are marked. The Plot is augmented by selecting different sizes and shapes for marking the points specifying eclipse magnitude and eclipse secondary type. A peculiar observation is the high concentration of partial eclipses near the poles. Mumbai's latitude and longitude location is mapped for better intuitive understanding.

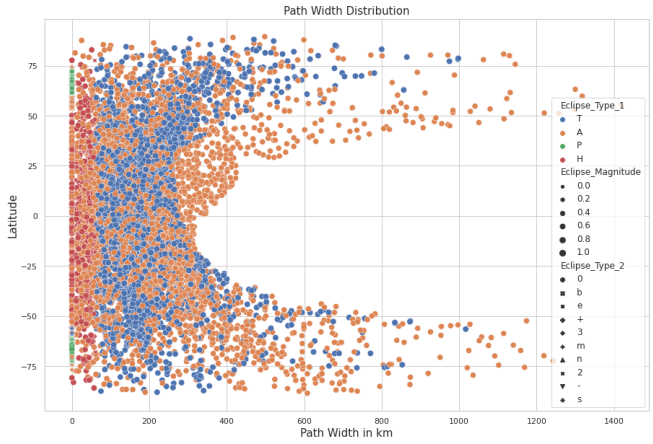

FIGURE IV
PATH WIDTH DISTRIBUTION WITH LATITUDE

The path width i.e. the distance swept on the earth's surface by the point of intersection of the axis of the moon's umbra with the earth, is found to be symmetrical in the northern and southern hemispheres. Partial eclipses are on the far left with almost zero path width, hinting at the small amount of path swept by the shadow axis near the poles. Hybrid, Total and Annular Eclipses have increasing widths in order, respectively.

To visualize how different variables in the dataset are related, we use a property of correlation - a statistical figure of merit taking values between 0 and 1, 1 being high correlation and 0 no correlation. It helps understand the data itself and its relations with other variables.

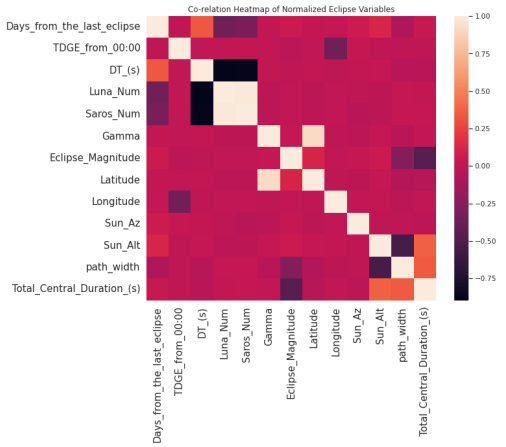

FIGURE V
CORRELATION HEATMAP OF DATA

II. Distribution Fitting

We fit various distributions to the data with the help of Maximum Likelihood Estimators to estimate how a variable is distributed along the permissible range and the probability of it lying in some interval. Histograms with density enabled have been used to demonstrate this feature.

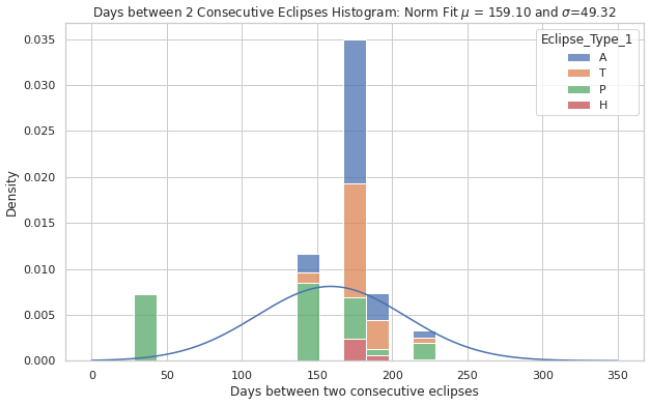

FIGURE VI
DISTRIBUTION OF NUMBER OF DAYS BETWEEN 2 CONSECUTIVE ECLIPSES

The number of days between two consecutive eclipses has been fitted with a normal distribution with mean μ = 159.10 days and standard deviation σ = 49.32 days. We can estimate that there are about 2.3 eclipses per year based on the average values. Considering 5000 years, the number is close to the number of entries in the catalog, validating our estimation.

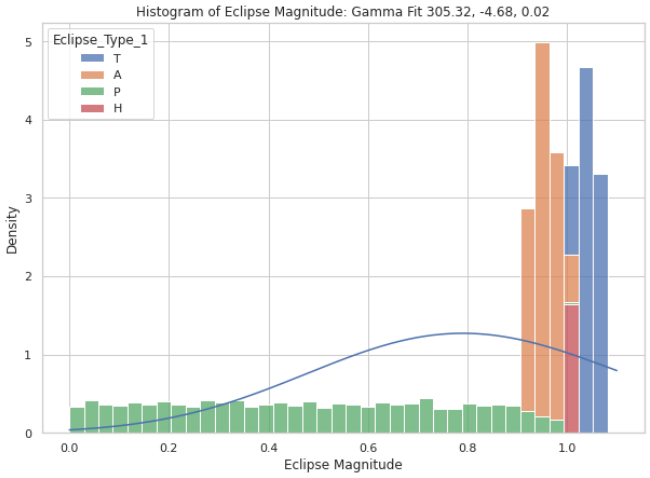

FIGURE VII
DISTRIBUTION OF ECLIPSE MAGNITUDE

The eclipse magnitude is a normalized measure of the amount of obscurity of the Sun. Unity and higher than that corresponding to total eclipses. This is found to be Gamma distributed with parameters (a = 305.32, b = -4.68, c = 0.02).

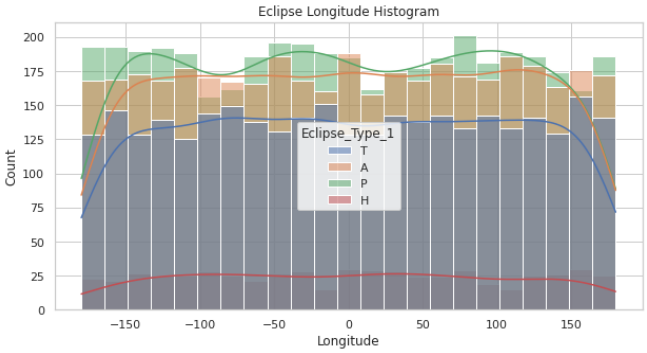

FIGURE VIII
DISTRIBUTION OF ECLIPSE POSITIONS (LONGITUDE)

The eclipse longitude, which locates the longitude of the location of the point of intersection of the moon's umbral shadow axis and the earth's surface, is almost uniformly distributed. More insight can be gained from Fig.II where the points are seen to be uniformly scattered.

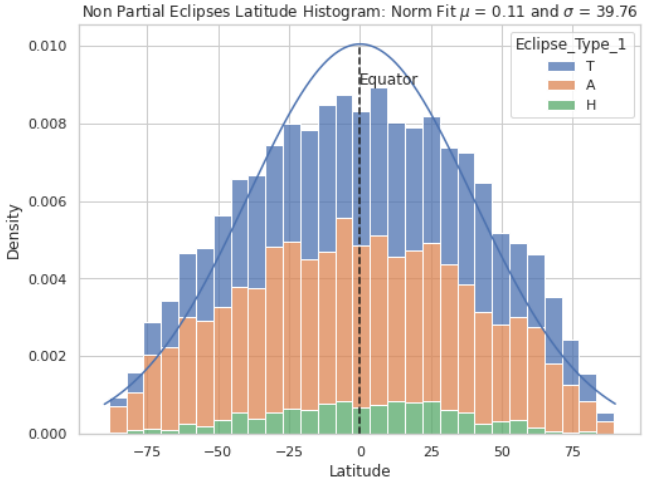

FIGURE IX
FITTING OF NON PARTIAL ECLIPSE POSITIONS (LATITUDE)

The latitude positions of Non-partial eclipses are nicely fitted with a Standard Normal Distribution with parameters μ = 0.11 N and standard deviation σ = 39.76. On the contrary, the latitude positions of the partial eclipses can be fitted with Double Weibull Distribution with parameters (a =18.82, b=0, c = 67.85) as shown below:

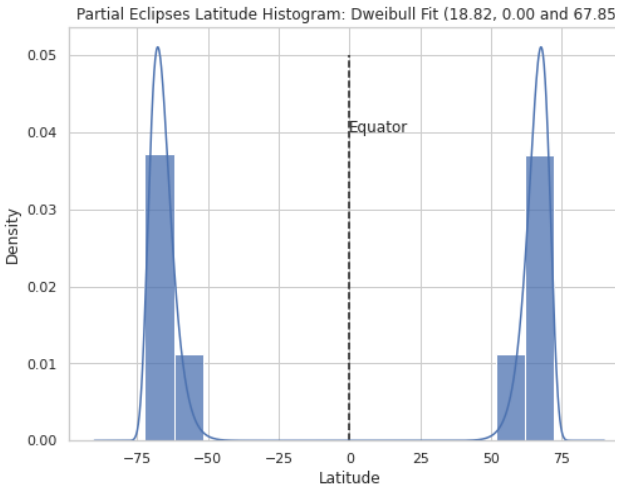

FIGURE X
DISTRIBUTION FITTING OF PARTIAL ECLIPSE POSITIONS (LATITUDE)

The Gamma parameter is the distance of the shadow cone axis from the center of Earth (units of equatorial radii) at the instant of the greatest eclipse. It is seen to be almost uniformly distributed.

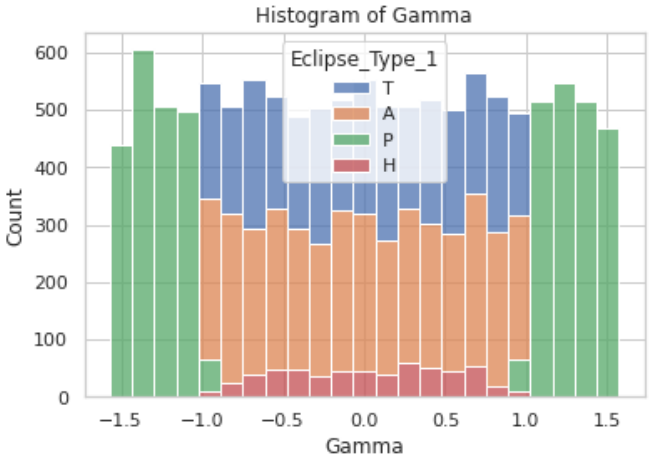

FIGURE XI
DISTRIBUTION OF ECLIPSE GAMMA

We can see beautiful patterns emerging from the scatter plot of Gamma and Latitude below. This hints at some physical governing laws relating these parameters to each other and not random bytes of data.

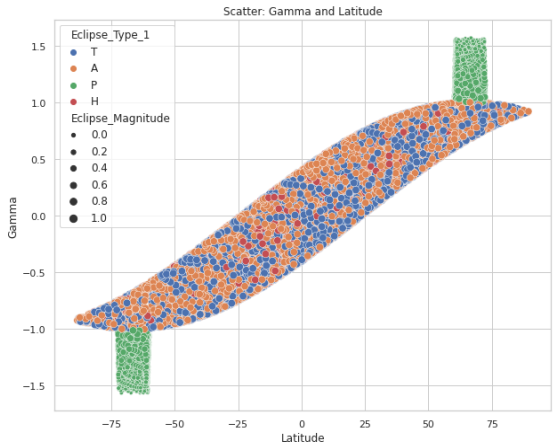

FIGURE XII
THE RELATION BETWEEN GAMMA AND ECLIPSE LATITUDE

In particular, the partial eclipses are concentrated near latitudes +90 and -90, which correspond to the North and South Poles, respectively.

Sun Azimuth, which is the Sun's azimuth at the greatest eclipse, is measured in degrees and can be seen to be uniformly distributed for partial eclipses.

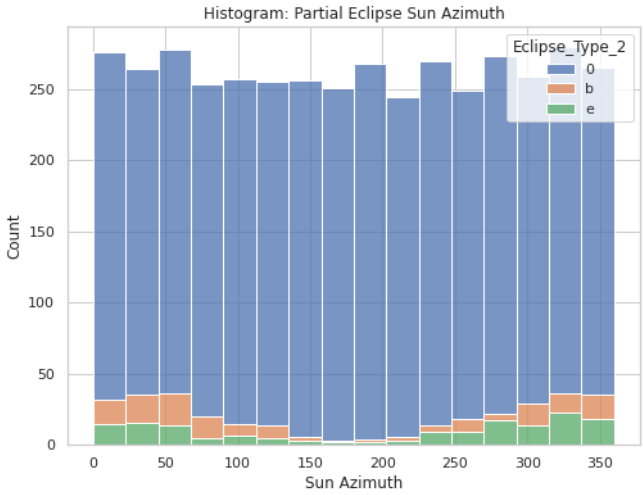

FIGURE XIII
DISTRIBUTION OF PARTIAL ECLIPSE SUN AZIMUTH

For Nonpartial eclipses, i.e the Total Hybrid and Annular eclipses the Sun azimuth is observed to be concentrated in regions around 0^0 and 180^0.

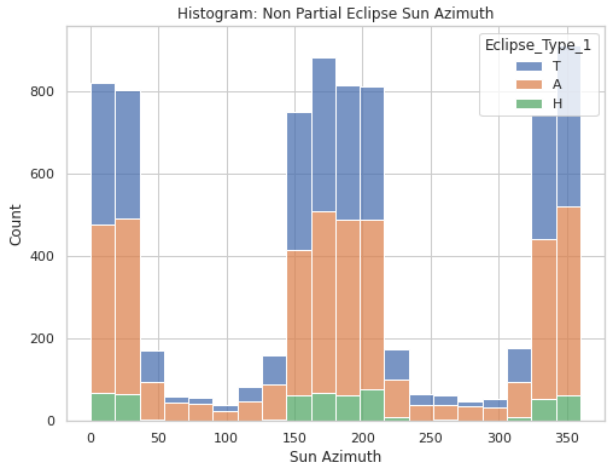

FIGURE XIV
DISTRIBUTION OF NON-PARTIAL ECLIPSE SUN AZIMUTH

The sun altitude, which is the sun's altitude at the time of the greatest eclipse (measured in degrees - as in the spherical coordinate system), is observed to be distributed as per the Gamma distribution with parameters (a = 1.97, b = 1.09, c = 108.81).

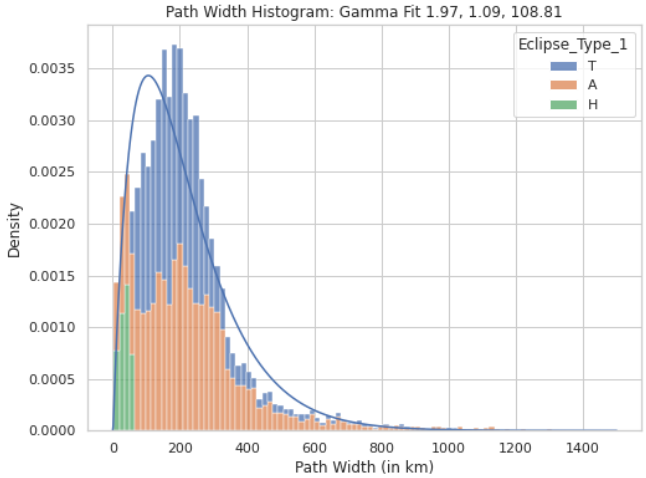

FIGURE XV
DISTRIBUTION OF NON-PARTIAL ECLIPSE PATH WIDTH

The central duration of eclipses (defined only for Hybrid, Total and Annular eclipses) is Gamma distributed with parameters (a = 8.59, -97.92, 45.30). Normal distribution was also fitted to the data with mean $\mu = 291.37$ and $\sigma = 132.20$.

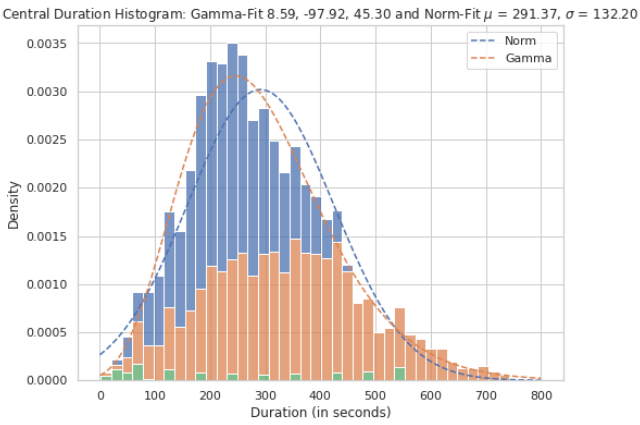

FIGURE XVI
DISTRIBUTION OF TOTAL CENTRAL DURATION (IN SECONDS)

REGRESSION AND CLASSIFICATION

I. Regression

We performed linear regression on the plots of parameters that seemed to have a linear relationship, Path Width v/s Total Central Duration, Eclipse Magnitude v/s Central Duration, and TDGE v/s Longitude. A thing to note about Figure XVII is that the data is split into two parts, one with Eclipse Magnitude > 1 and the other with Eclipse Magnitude < 1. Additionally, Central Duration > 5 seconds was used to fit the curves.

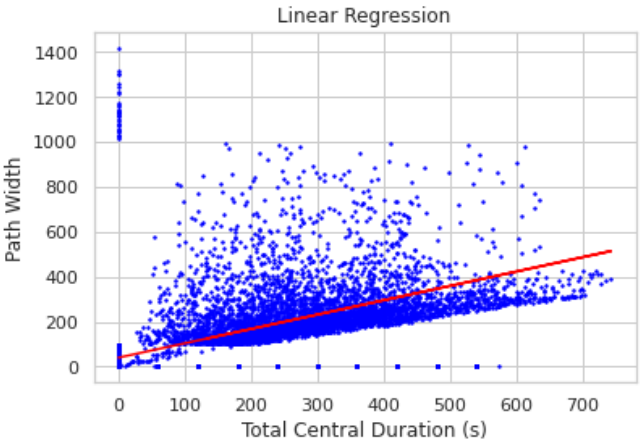

FIGURE XVII
LINEAR REGRESSION ON PATH WIDTH AND CENTRAL DURATION

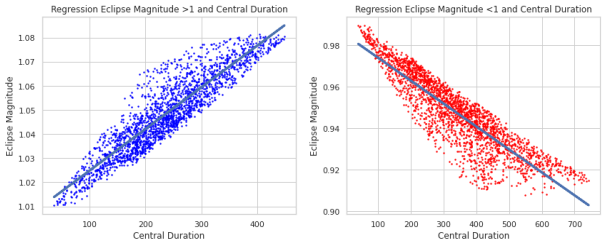

FIGURE XVIII
LINEAR REGRESSION ON ECLIPSE MAGNITUDE & CENTRAL DURATION

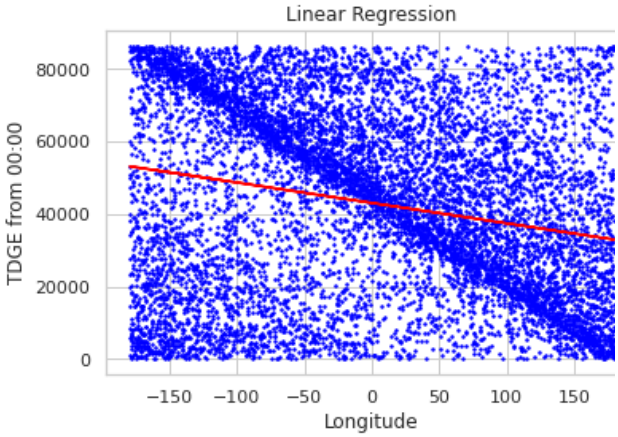

FIGURE XIX
LINEAR REGRESSION ON LONGITUDE AND TDGE_FROM_00:00

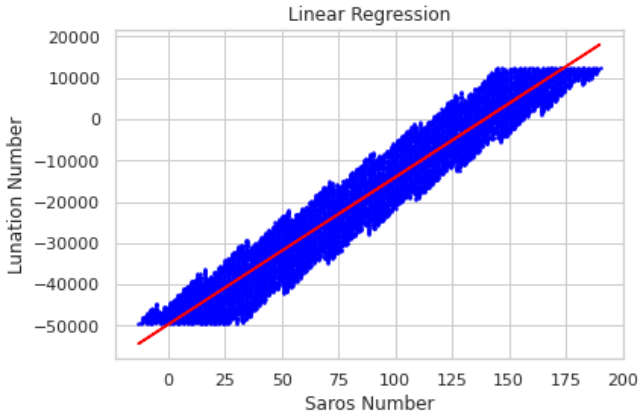

FIGURE XX
LINEAR REGRESSION ON LUNA NUMBER & SAROS SERIES NUMBER

II. Classification (Clustering)

We performed K-Means clustering for the given data to classify eclipses into 4 main categories, namely Partial eclipse, Annular eclipse, Total eclipse and Hybrid eclipse. After clustering the data based on the 4 categories, we have scatter plotted different parameters for each label and we can see clusters forming. The parameters were chosen such that their scatter plots were as spread out as possible, since such a pair of parameters will allow clustering to be visualized properly.

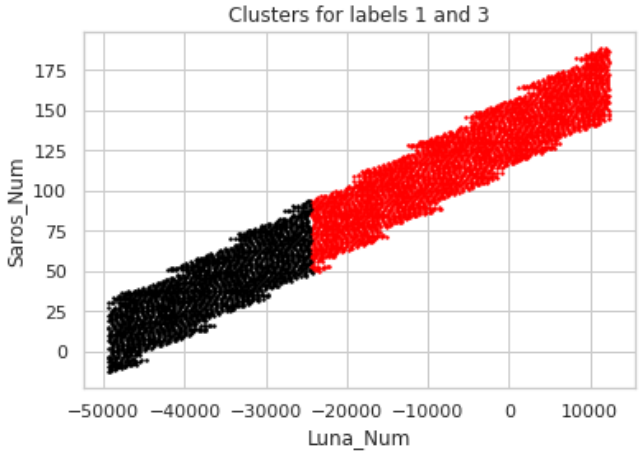

FIGURE XXI
CLUSTERING ON LUNA NUMBER & SAROS SERIES NUMBER

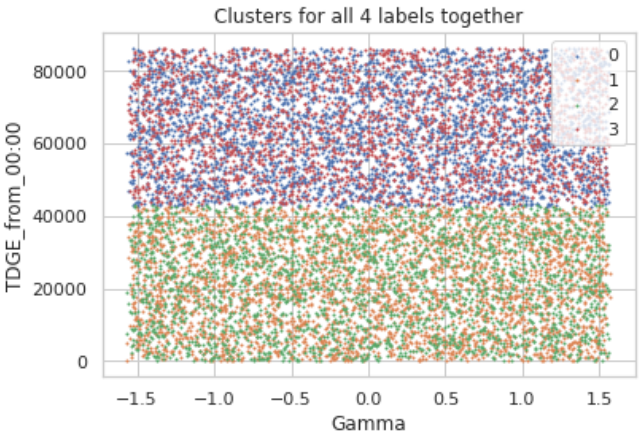

FIGURE XXII
CLUSTERING ON LUNA NUMBER & SAROS SERIES NUMBER

EMERGENCE OF PATTERNS

We plotted all the variables against each other and we got in total $^{13}C_2$ scatter plots. Some of them are mentioned here and we can clearly observe emergence of distinctively visible patterns in the data

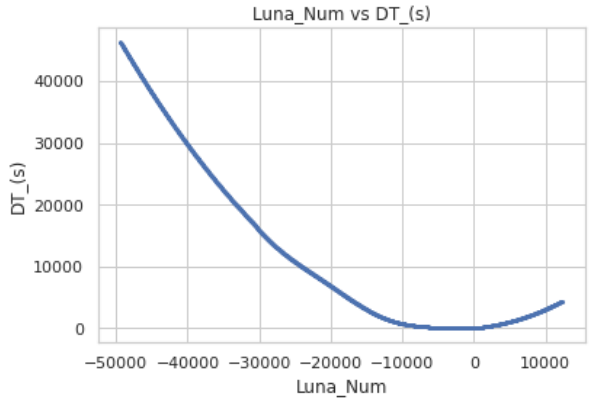

FIGURE XXIII
DISTRIBUTION OF LUNATION NUMBER AND DELTA T

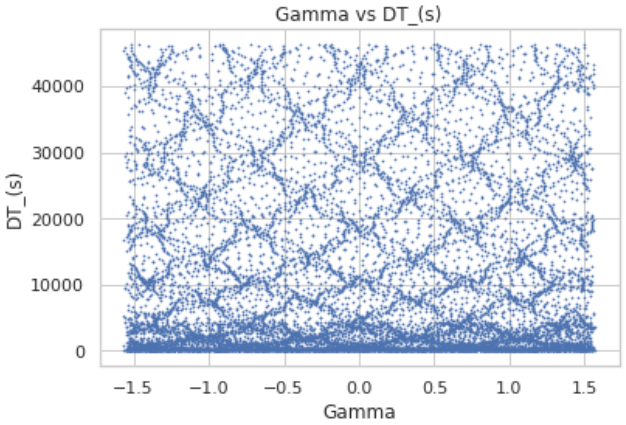

FIGURE XXIV
DISTRIBUTION OF GAMMA AND DELTA T

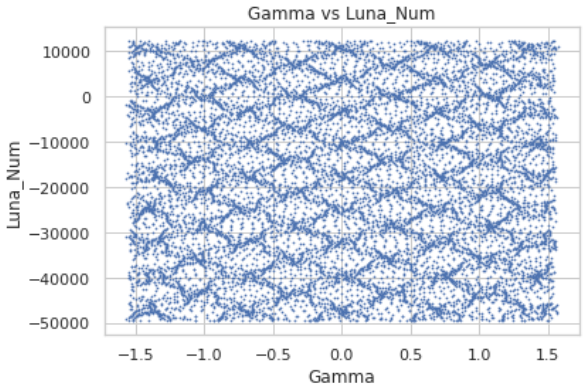

FIGURE XXV
DISTRIBUTION OF GAMMA AND LUNATION NUMBER

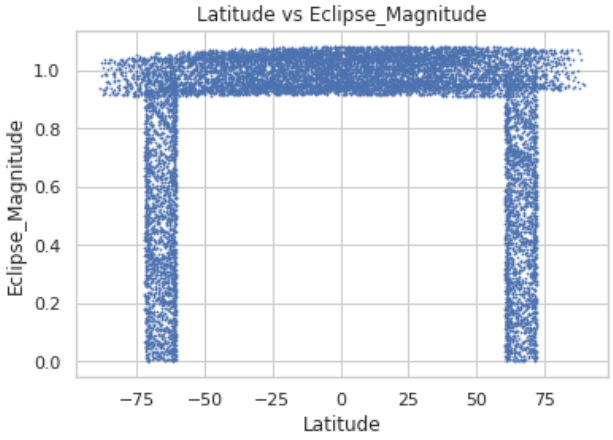

FIGURE XXVI
DISTRIBUTION OF LATITUDE AND ECLIPSE MAGNITUDE

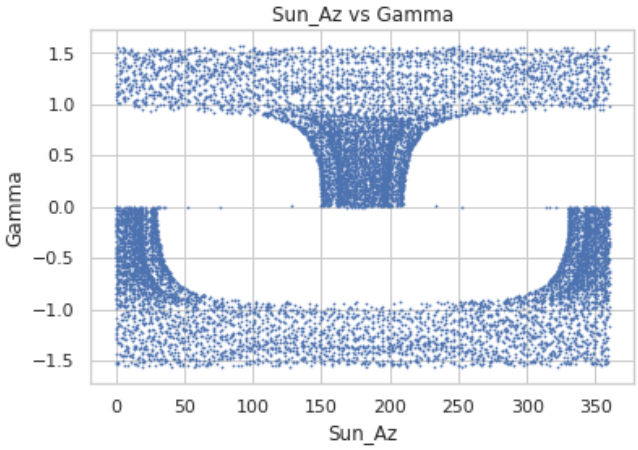

FIGURE XXVII
DISTRIBUTION OF GAMMA AND SUN AZIMUTH

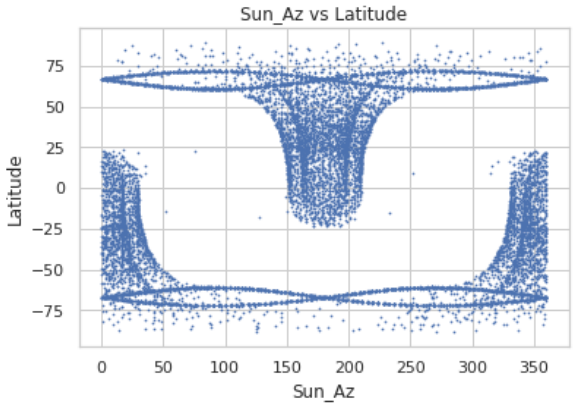

FIGURE XXVIII
DISTRIBUTION OF LATITUDE AND SUN AZIMUTH

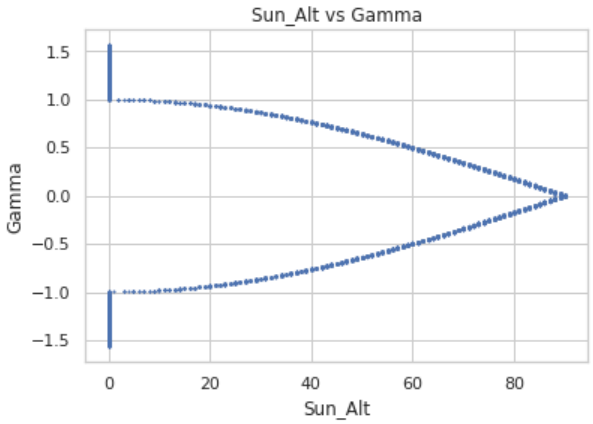

FIGURE XXIX
DISTRIBUTION OF GAMMA AND SUN ALTITUDE

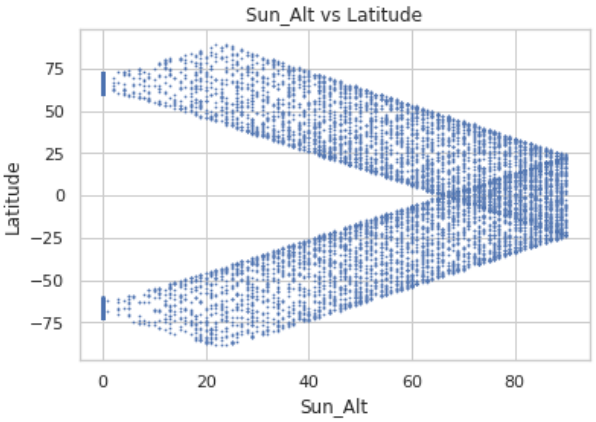

FIGURE XXX
DISTRIBUTION OF LATITUDE AND SUN ALTITUDE

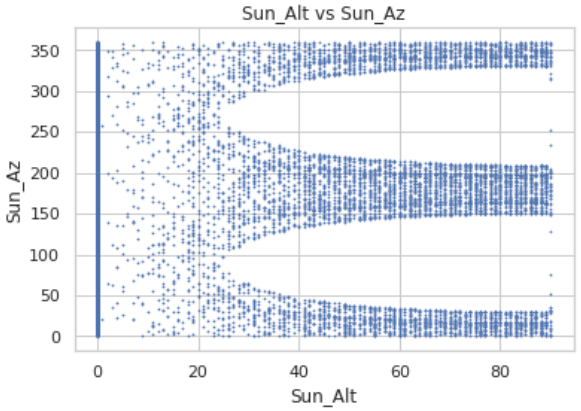

FIGURE XXXI
DISTRIBUTION OF SUN AZIMUTH AND SUN AZIMUTH

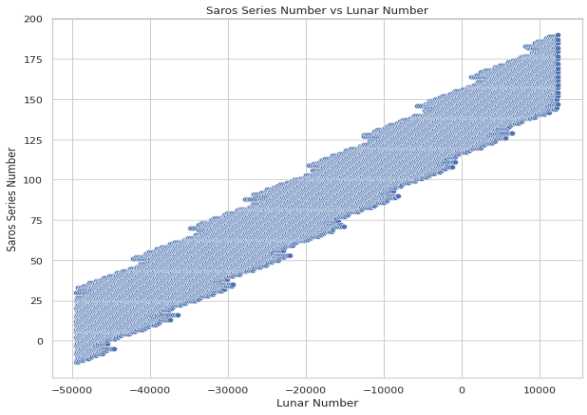

FIGURE XXXII
RELATION BETWEEN LUNAR NUMBER & SAROS SERIES NUMBER

REFERENCES

[1] "Eclipse Predictions by Fred Espenak and Jean Meeus (NASA's GSFC)", NASA Technical Publication TP-2006-21414, 2009, https://eclipse.gsfc.nasa.gov/SEcat5/catalog.html,Accessed November 4, 2021

[2] NASA GSFC. "NASA - Key to Catalog of Solar Eclipses." NASA - Key to Catalog of Solar Eclipses, 2007, https://eclipse.gsfc.nasa.gov/SEcat5/SEcatkey.html. Accessed 25 November 2021.

[3] "Ring of fire: Visualizing 5,000 years of solar eclipses." Ring of fire: Visualizing 5,000 years of solar eclipses, SAS Blogs, 19 July 2019, https://blogs.sas.com/content/sascom/2019/07/19/ring-of-fire-visualizing-5000-years-of-solar-eclipses/. Accessed 25 11 2021

[4] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2. (Publisher link).

[5] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), 261-272.

[6] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

[7] Michael L. Waskom (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, D., Brucher, M., Perrot, M., & Duchesnay, E.

(2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Contributors

1. Soham Phanse: Exploratory Data Analysis and Visualization, Distribution Fitting
2. Pranjal Gupta: Regression and Classification

Author Information

1. **Soham Phanse**, 19D170030, Junior Year Undergraduate, Department of Aerospace Engineering, Indian Institute of Technology Bombay.