STA2453 Exploratory Data Analysis for Detecting Stellar Flares

Isaac Baguisa

2025-02-12

Introduction and Data Overview

This exploratory data analysis (EDA) aims to uncover the underlying patterns and characteristics of brightness measurements from three different stars (TIC 0131799991, TIC 031381302, and TIC 129646813), as part of a larger project to detect stellar flares using data from the Transiting Exoplanet Survey Satellite (TESS). Each step in this analysis is to address specific questions that collectively aim to improve our understanding of stellar flares.

The dataset contains time series measurements of stellar brightness and associated errors from three stars. The key variables include: time, the observation timestamp in days (Barycentric Julian Date). PDCSAP_FLUX, the photometric flux after pre-search data conditioning, indicating the star's observed brightness (electrons per second). PDCSAP_FLUX_ERR, the error associated with each flux measurement (electrons per second).

Methods and Data Analysis

Comparing Imputed Values with Non-missing Data

The data contains gaps due to the satellite turning off, therefore imputing missing values and comparing them with actual observations allows us to validate the imputation method and understand its impact on potential flare detection. Effective imputation that closely follows the true data patterns is important for analysis to ensure that potential flares are not missed during gaps in data collection. Prior to imputation, the number of missing values for each star are in table 1

Table 1: Number of Missing Values in PDCSAP_FLUX by Star

Dataset	Available	Missing
TIC 031381302	17033	686
TIC 129646813	18188	91
TIC 0131799991	13034	338

The imputation method used is the Kalman filtering algorithm. The Kalman filter works well becauses it uses an adaptive process that estimates the state of a linear dynamic system from a series of noisy measurements. It can effectively track the sudden increases in brightness and subsequent decays, even when there are gaps in the data. This makes it robust missing not at random (MNAR) data, as it does not require the missing data to be unrelated to the unobserved data values. Figure 1 shows the imputed values closely follows the original flux, therefore this imputation method will used throughout the analysis.

Identifying Patterns or Cycles in Brightness

Visualizing flux over time allows us to detect patterns or irregularities such as sudden spikes that could denote flares. Sudden peaks in these plots may indicate flares, while consistent patterns could suggest underlying stellar processes or cycles. In figure 1, for TIC 031381302, the flux shows consistent variability with occasional sharp spikes, suggesting frequent, moderate fluctuations in brightness which might be indicative of stellar flares. TIC 129646813 exhibits a relatively stable flux level with very distinct, sharp peaks likely pointing to

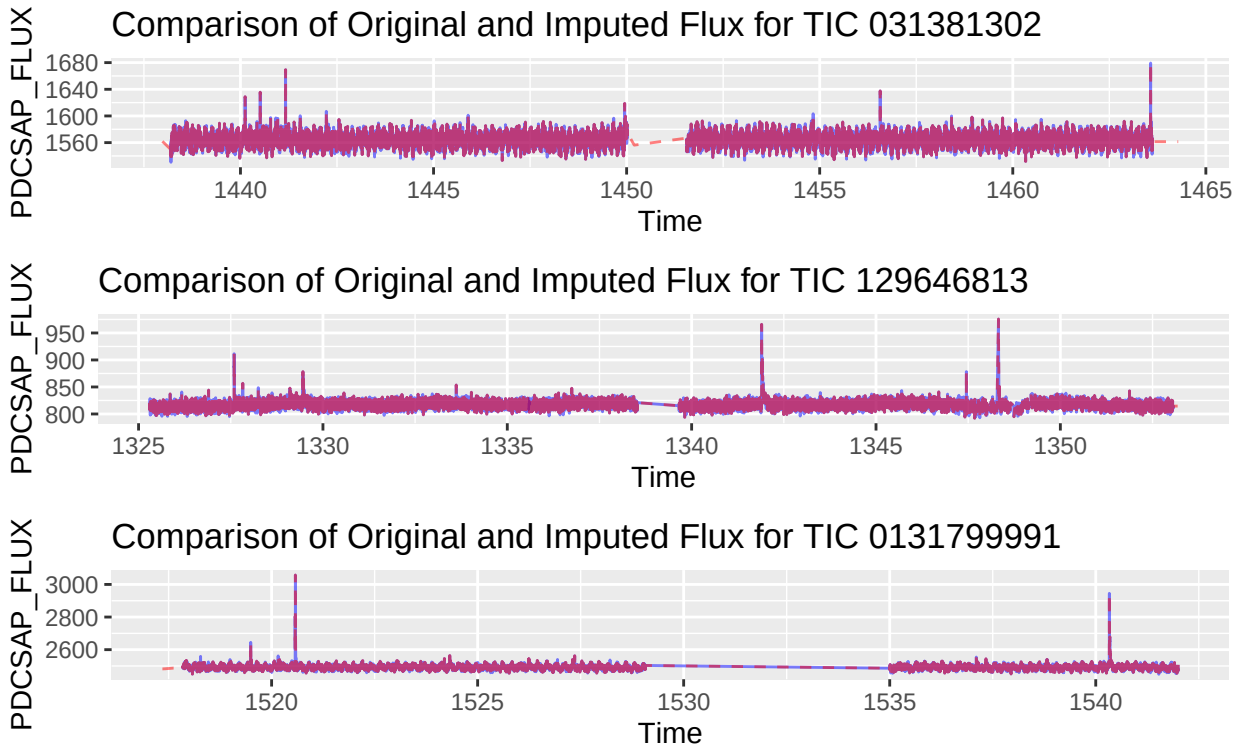

Figure 1: Comparison of Original and Imputed PDCSAP_FLUX

infrequent but significant stellar flare. Lastly, TIC 0131799991 maintains a very stable flux with only a few noticeable spikes, indicating rare events of heightened stellar flares.

General Distribution of PDCSAP Flux

The histogram plots provide a better understanding of the flux distribution, highlighting typical brightness levels. By examining the distribution's skewness and kurtosis, we can infer the frequency and intensity of anomalous brightness events.

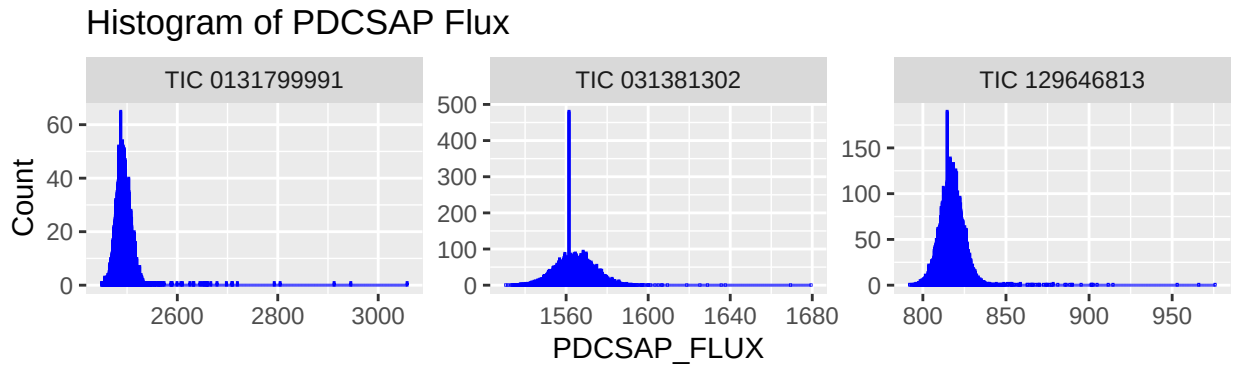

Figure 2: Distribution of PDCSAP Flux by Star

The long tail towards the higher flux values suggests the presence of occasional but significant spikes in brightness, which are characteristics of stellar flares. The skewness could imply that while flares are relatively rare compared to normal stellar activity, they are significant when they occur, as shown by the tail extending to higher values. All three histograms in figure 2 suggest relatively stable stellar brightness with a consistent flux level dominating the observations. The lack of wide variance across all three histograms indicates stable

conditions with occasional bursts of higher energy. Additionally, TIC 0131799991 and TIC 129646813 have extreme peaks and shows minimal spread, indicating that most flux measurements are very close to a typical value with very few variations. TIC 031381302 also has an extreme peak, but with a larger variation than the other two stars.

Correlation of Flux Errors

Understanding how measurement errors vary with flux values helps in assessing data reliability. A high correlation between flux and its error could indicate that high readings during a flare are less reliable. A positive correlation might suggest that error increases with flux intensity, which could complicate the detection of true flares from anomalies.

Table 2: PDCSAP Flux and Flux Error Correlation by Star

Star	Correlation
TIC 031381302	0.0831141
TIC 129646813	0.0020326
TIC 0131799991	0.1466368

The correlations between PDCSAP flux and the flux error in these stars are weak, suggesting that errors in the flux measurements are generally not strongly dependent on the flux values themselves. The low correlation suggests that the error associated with flux measurements does not vary significantly with the stellar brightness, indicating stable and consistent data quality across different flux levels.

Identifying Abrupt Changes (Potential Flares)

Even if the brightness of the stars are not normally distributed, the use of standard deviation as a measure of dispersion can still provide insight of points that deviate significantly. The thresholds set at 3 and 4 standard deviations above the mean flag the most extreme variations in brightness that are likely to be of interest - potential flares.

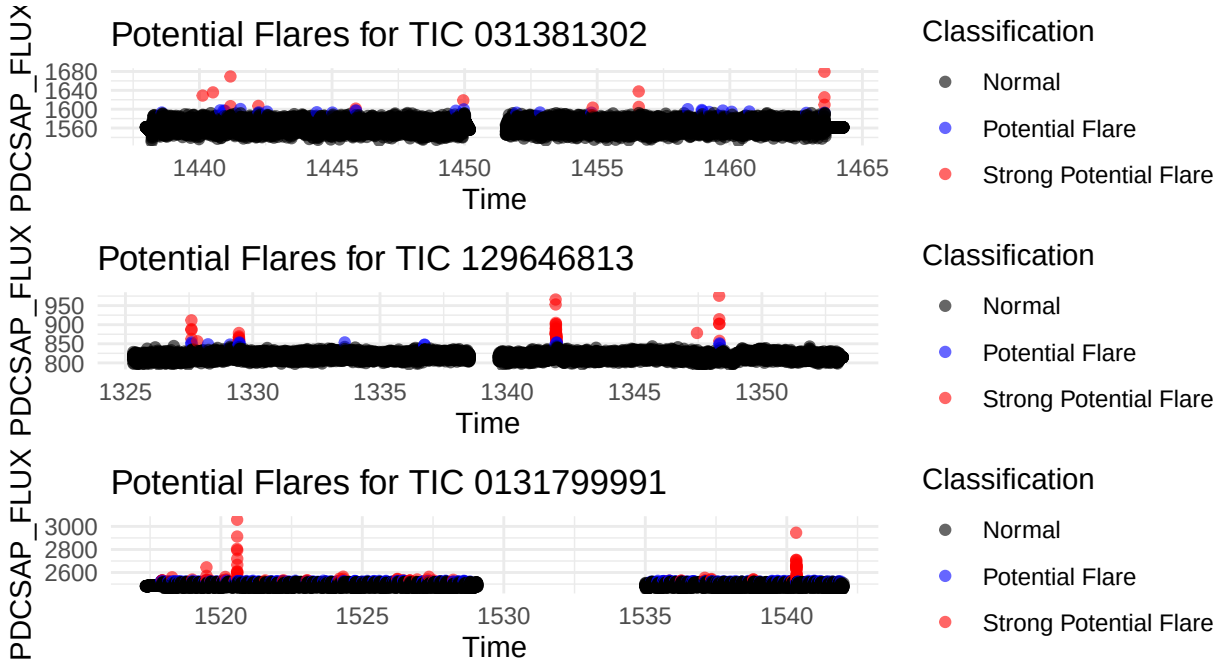

Figure 3: Potential Flares Using 3 and 4 SDs as Threshold

Clustering Brightness Events with DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) does not assume clusters to be of any specific shape. This makes it suitable for the stellar flares data where groups of data points might not be uniformly shaped. Stellar flares can occur in various patterns and intensities, and DBSCAN can identify these irregular patterns as clusters based on density. Identifying distinct clusters can help differentiate between normal stellar activity and anomalies that may signify flares.

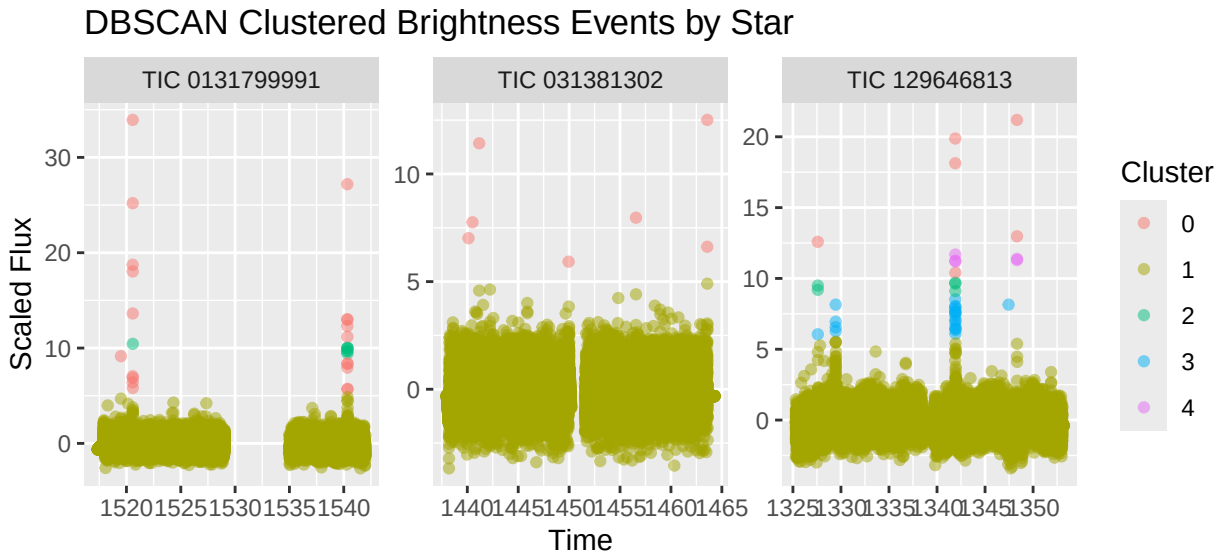

Figure 4: DBSCAN Clusters by Star

Table 3: Number of Points in Cluster by Star

Star_ID	Cluster	Count
TIC 0131799991	0	20
TIC 0131799991	1	13345
TIC 0131799991	2	7
TIC 031381302	0	7
TIC 031381302	1	17712
TIC 129646813	0	6
TIC 129646813	1	18242
TIC 129646813	2	5
TIC 129646813	3	21
TIC 129646813	4	5

ARMA Model to Analyze Residuals

Given the randomness of stellar flares, residuals are analyzed using a auto regressive moving average model (ARMA). Applying an ARMA model to the time series data helps in smoothing out short-term fluctuations and highlighting longer-term trends or cycles. This can effectively model the random shocks that represent deviations from a typical pattern, such as ones caused by stellar flares. This blend allows it to capture both the direct influence of the previous value (AR component) and the effects of previous errors (MA component) in the time series. In this residual analysis I will focus on TIC 129646813 specifically, as other stars may require some parameter tuning.

A rapid decline in the ACF values after lag 0 typically indicates that the ARMA(1,1) model is effectively capturing the autocorrelation structure of the data. From the ACF plot, we see that larger lags are contained

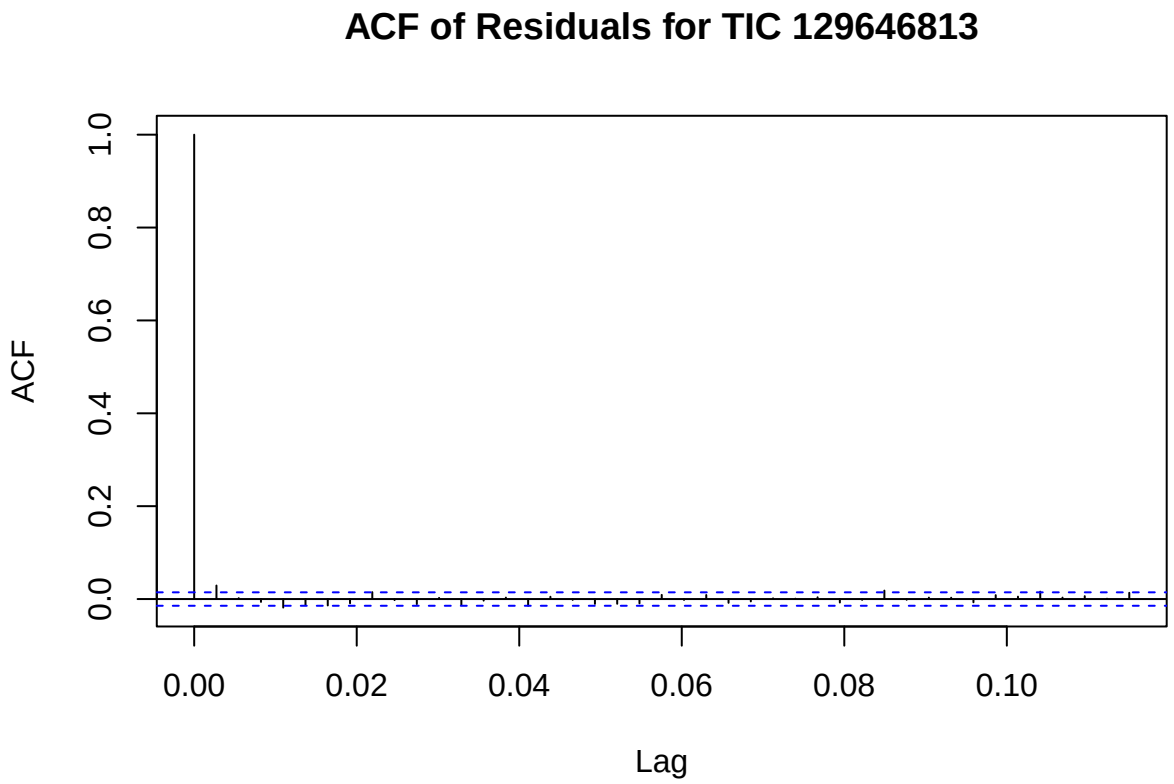

Figure 5: ARMA(1,1) Model Fitted to the Residuals of TIC 129646813

(or almost fully contained) in the blue boundary points of $\pm 2/\sqrt{(}n)$.

Conclusion

The exploratory data analysis provided critical insights into the flux behaviour of three stars observed by TESS. Each visualization was chosen to highlight different aspects of the data. Beginning with imputation of missing data, caused by satellite's turning off, we have applied a the Kalman filter for a continuous analysis. Additionally, visualization techniques has showed the general distribution of brightness values, patterns and potential cyclic behaviours across the stars. These visualizations have highlighted sudden peaks that potentially correspond to flares. The use of DBSCAN clustering was used in identifying irregular patterns of data points. This approach aligns with the sudden nature of flares, allowing us to cluster points between normal brightness variations and potential flares. Lastly, the implementation of an ARMA(1,1) model to analyze residuals, and this model helps understand the underlying processes behind stellar flares. The findings from this EDA will guide further statistical modeling for effective stellar flare detection.