Exoplanet Search

The search for habitable planets outside our solar system has fascinated us for some time. Once theorized and now confirmed that there are many exoplanets (3,700 and counting), the search for life outside our solar system is now in full swing. Over 2,300 of them from Kepler space satellite.

Exoplanet Count

Kepler:

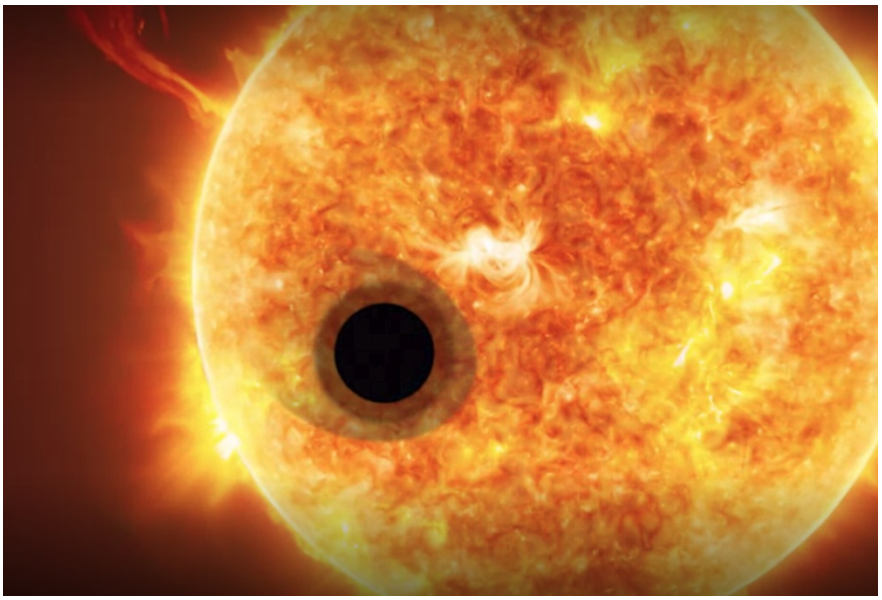
Candidates: 2,244
Confirmed: 2,327
Small Habitable Zone Confirmed: 30

K2:

Candidates: 480
Confirmed: 292

Exoplanets can be identified as Points of Interest for the Scientific Research Community

The Transit method identifies possible planets crossing the path of stars from the Kepler camera's point of view. This object of interest can then be directed to other telescopes for further analysis to verify one or more exoplanets. Other details such as chemical composition of the star and by inference the chemical composition and possible atmosphere of the exoplanet can also be determined.



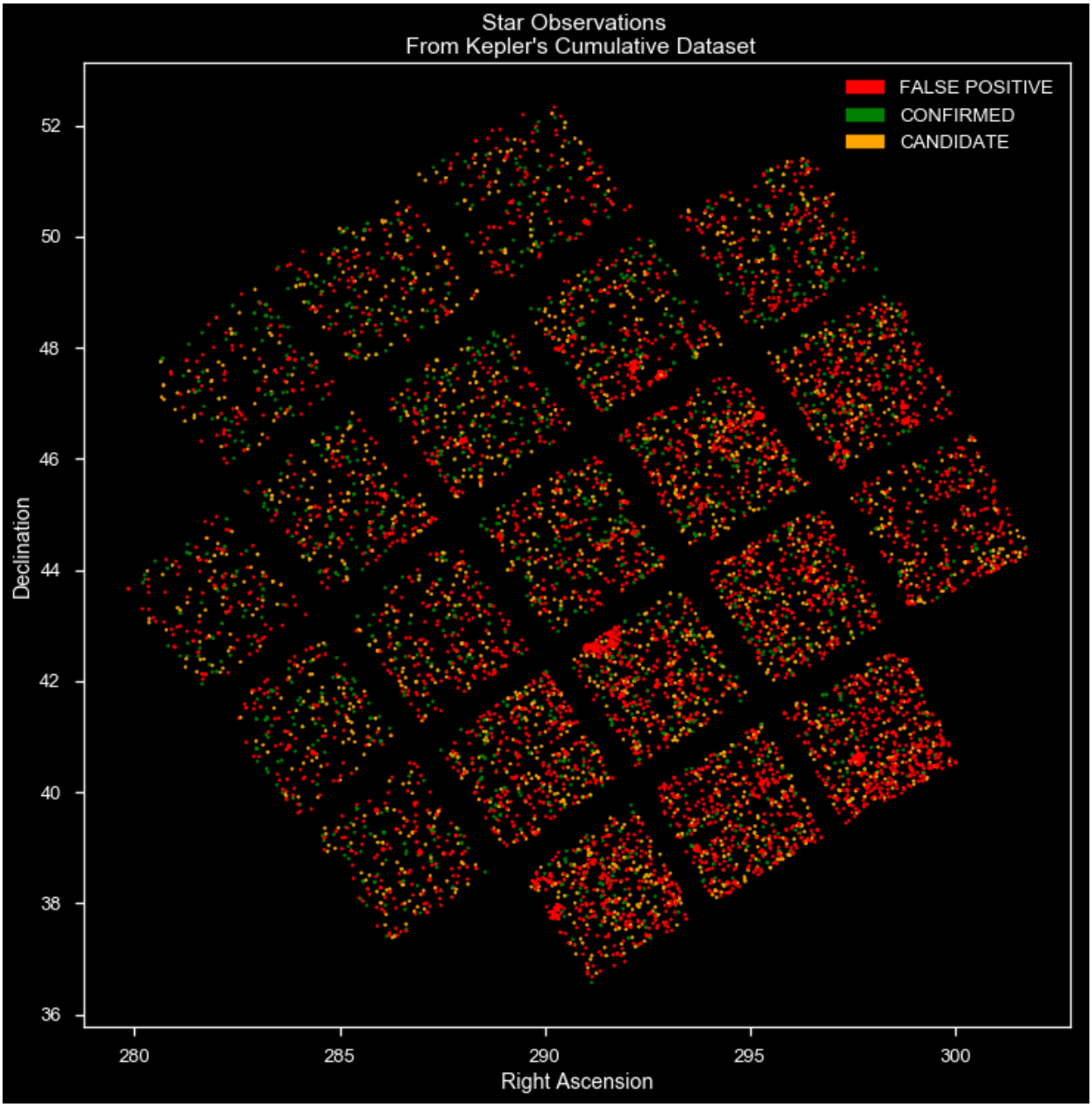
Data from Kepler Mission 2 'Hunt for Exoplanet' Kaggle competition

The data set contains over 5500 stars with 42 confirmed exoplanets. It is a time series with 3198 measurements of light intensity at 30 minute intervals (80 days). The data had been cleaned for the competition to remove known artifacts from the Kepler camera. For the competition it was split into a training and test set with confirmed exoplanets of 37 and 5 respectively.

Potential Dataset from Campaigns after Kepler Breakdown

Another dataset available is the cumulative data from subsequent Kepler missions called Kepler 2 campaigns. This data that was taken after the initial malfunction of the navigation gears of the ship. It was still able to take images of other parts of the sky using the pressure from the sun as a gear. This dataset includes the star names, planet names and more details of the stars including their position in the sky.

An interesting visual from this dataset are the sky coordinates of the observations:



Exploratory Data Analysis

Normalizing and Exploring the Data

With the large range of flux in light intensity I normalized the data to allow a better comparison.

A single dimming over the 80 day period may be a slower orbiting planet or other star activity. Two low intensity readings provide no additional information as they may not be related to each other, but three dimming equally spaced apart are a strong contender for an exoplanet.

Changes in intensity of the light from a star can be due to solar flares, sun spots or the rotation of the star. The light comes from stars at different distances, of different brightness and temperatures and of different sizes with different rotations.

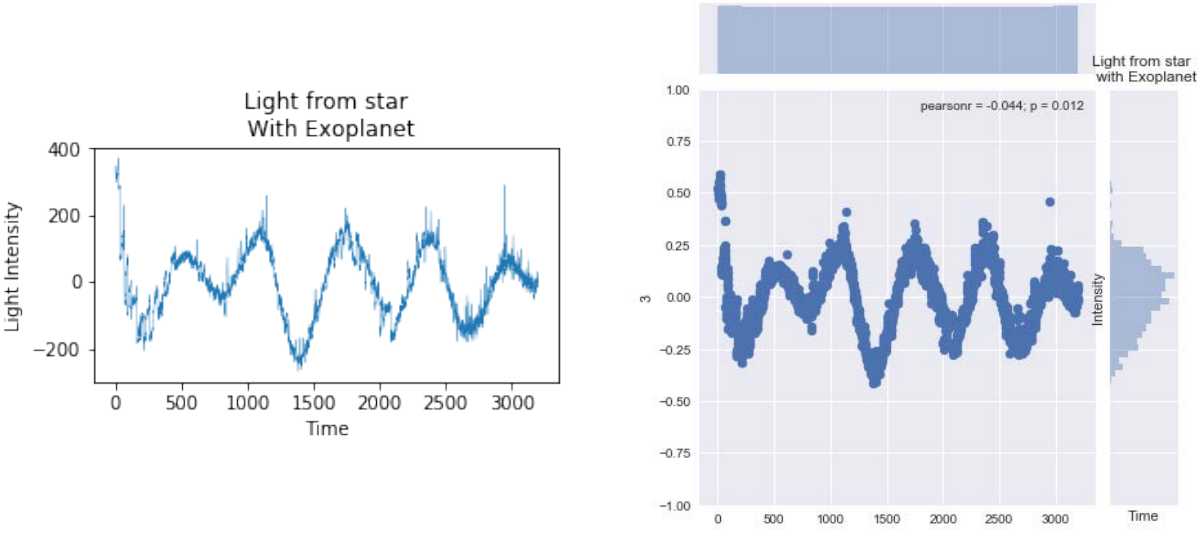
An exoplanet(s) transiting the star from our point of view will cause the light to dim. Depending on the size of the object, the angle, and duration of the transit the light will dim by a certain amount for a certain period. The spaceship is effectively "detecting a flea crossing the beam of a headlight several miles away". The light from the star also intensifies as the exoplanet reflects the light of the star as it is going around behind the star.

Light Intensity Variation by Star

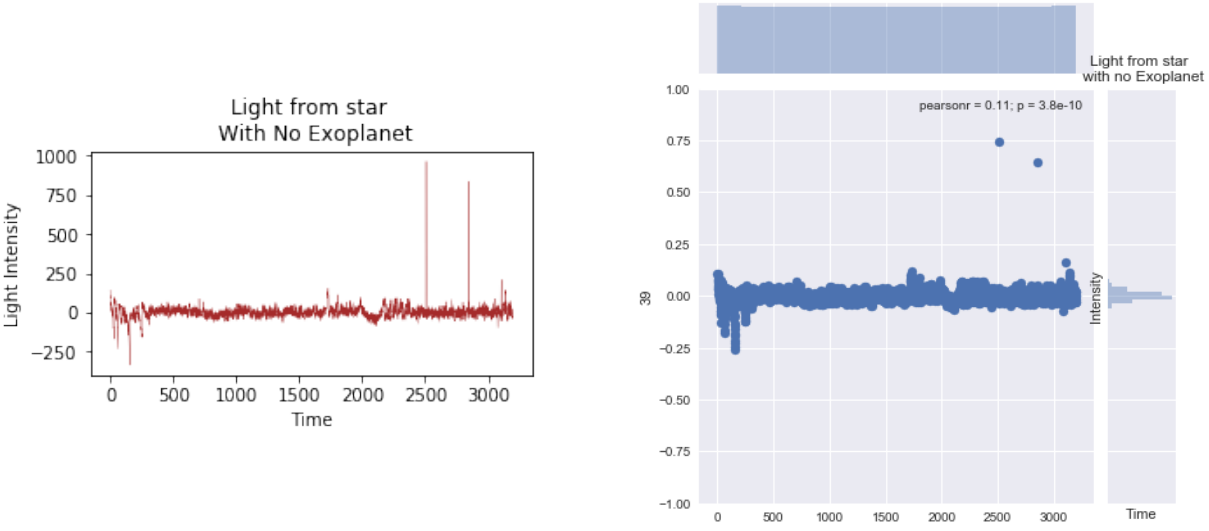
Star#	FLUX.1	FLUX.3	FLUX.3	FLUX.4	FLUX.5	FLUX.6	FLUX.7	FLUX.8	FLUX.9
37	-141.22	-81.79	-52.28	-32.45	-1.55	-35.61	-23.28	19.45	53.11
38	-35.62	-28.55	-27.29	-28.94	-15.13	-51.06	2.67	-5.21	9.67
39	142.40	137.03	93.65	105.64	98.22	99.06	86.40	60.78	45.18
40	-167.02	-137.65	-150.05	-136.85	-98.73	-103.14	-107.70	-123.19	-125.65
41	207.74	223.60	246.15	224.06	210.77	189.56	172.68	170.31	148.79

Looking at the time series plotted out, stars with exoplanets have a slightly different distribution (closer to two peaks) than those without, which have a clear single high peak. The peaks and troughs of the time series are more distinct.

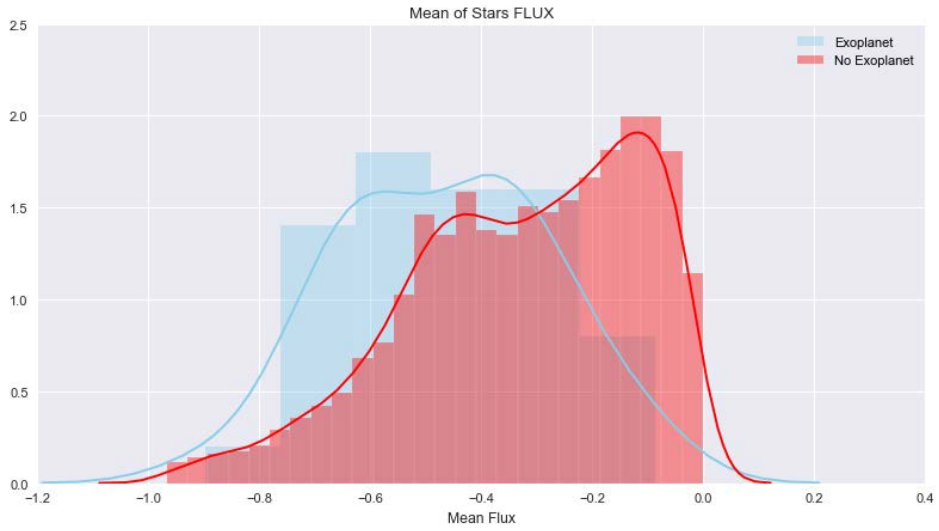
Time Series and Density Plot
Star with an Exoplanet

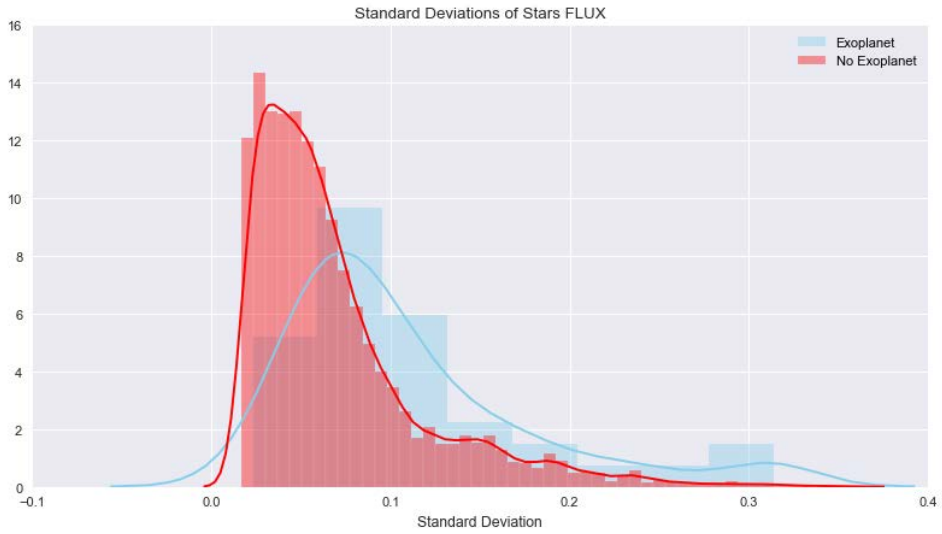


Time Series and Density Plot
Star with no Exoplanet

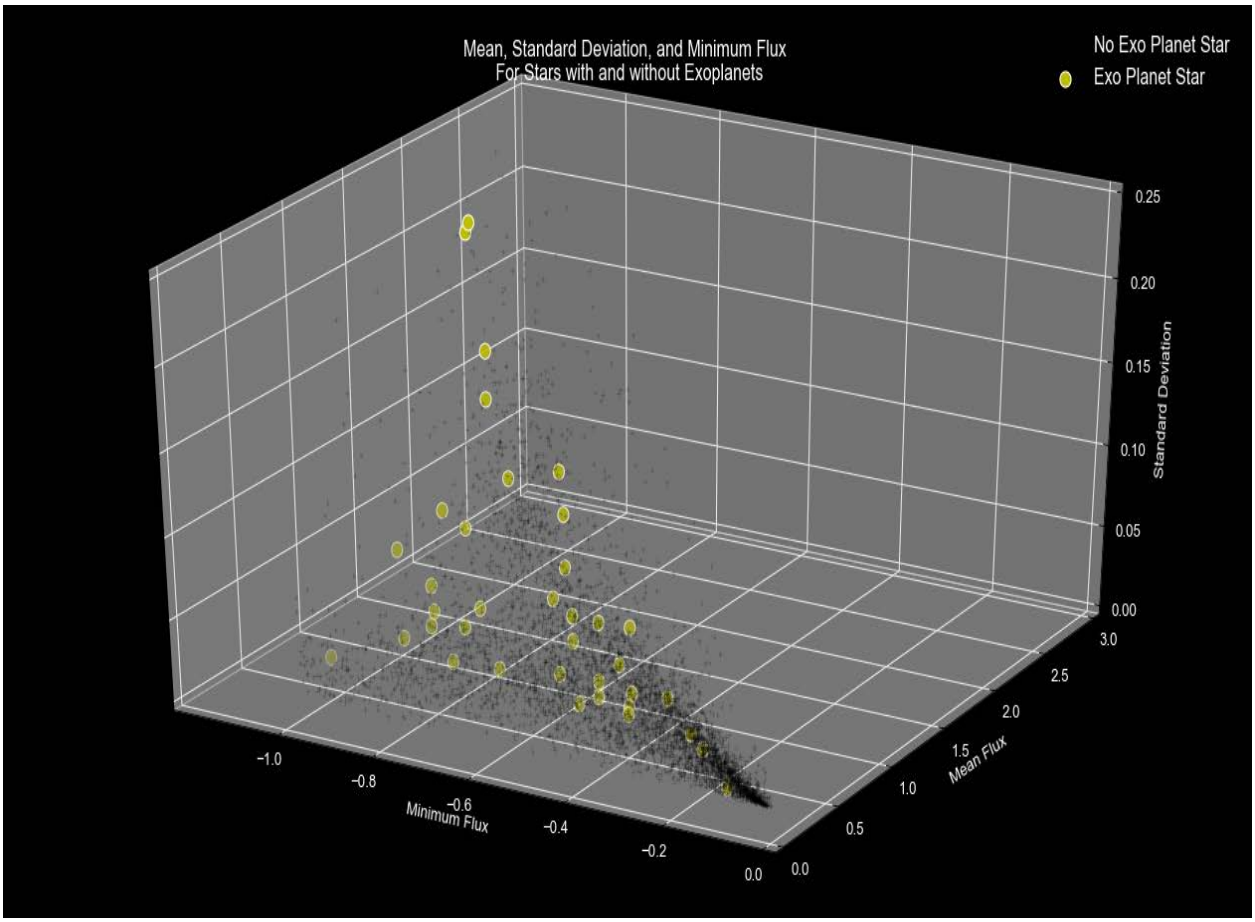


Plotting the means and standard deviations of the Flux of the stars:

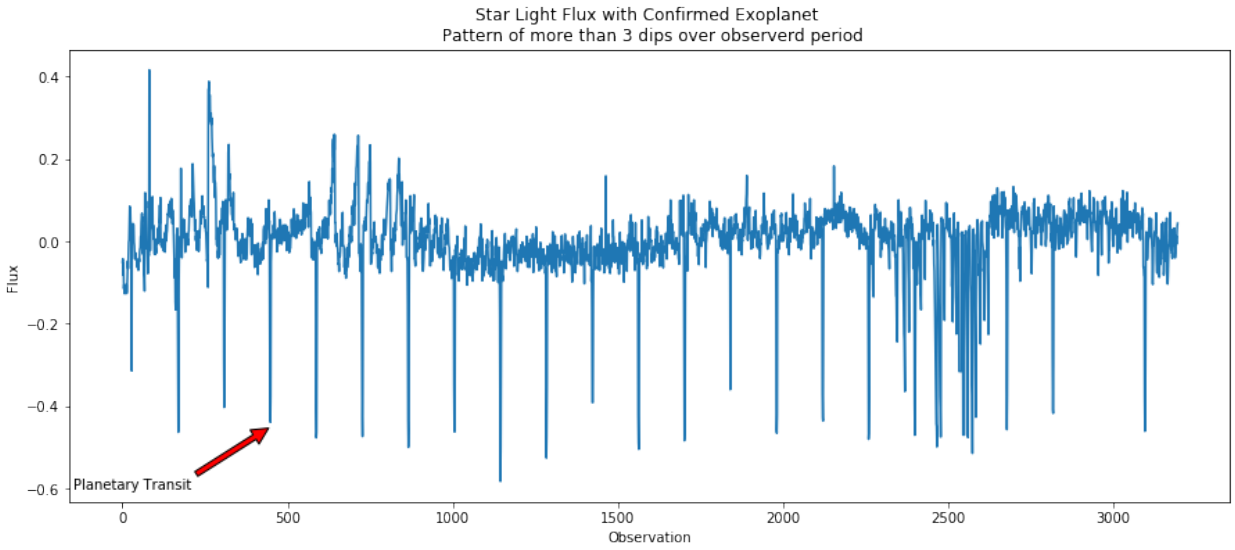




If the standard deviations, means and the minimums are plotted out in 3 dimensions, we can see that the stars with no exoplanets have a smaller standard deviation and their minimum value for flux is closer to the mean.

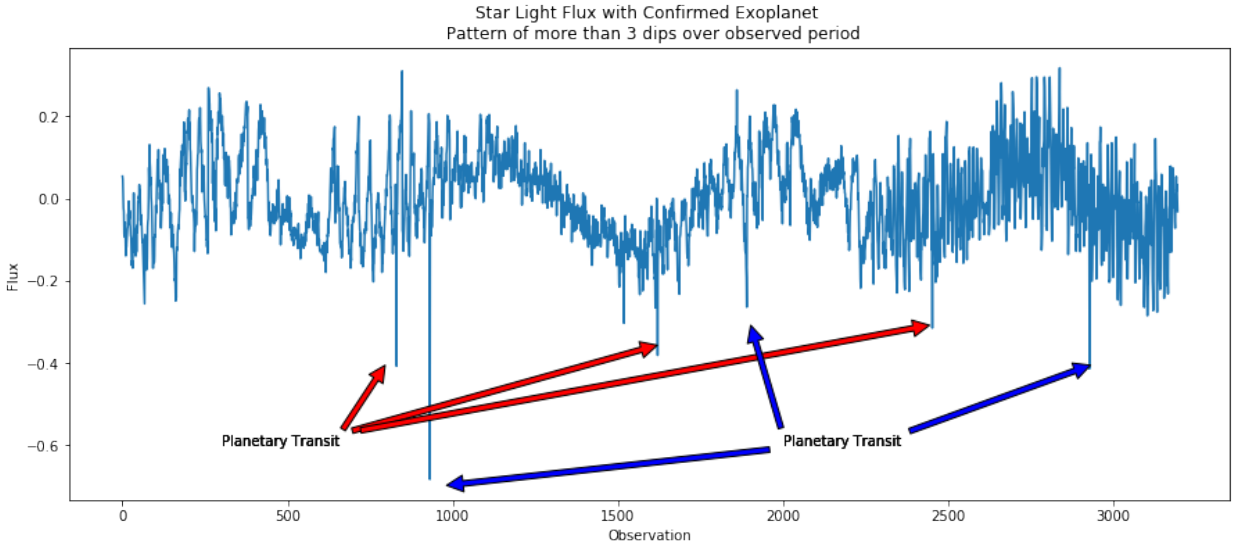


In these plots it is easier to see where the exoplanet indicators might be:



The above plot has 23 transits, which equates to an exoplanet with an orbital period of 3-4 days. The size of the dip also indicates a large very (and fast!) planet.

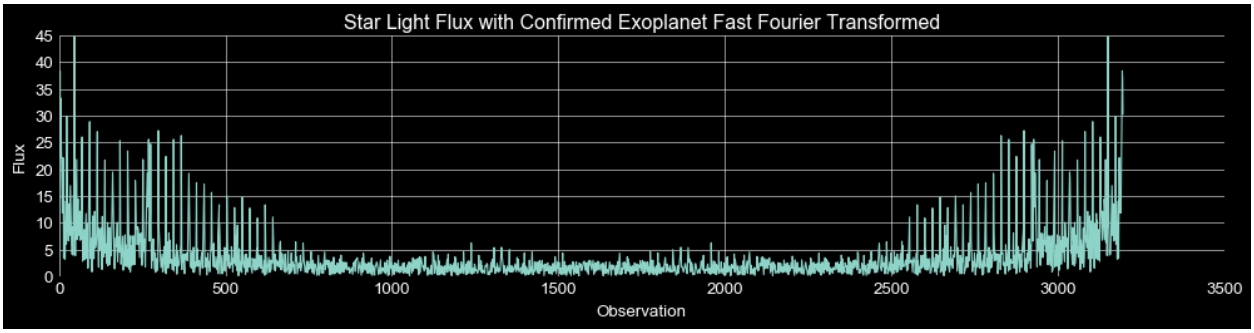
The next plot suggests the presence of more than one planet with 2 separate dips

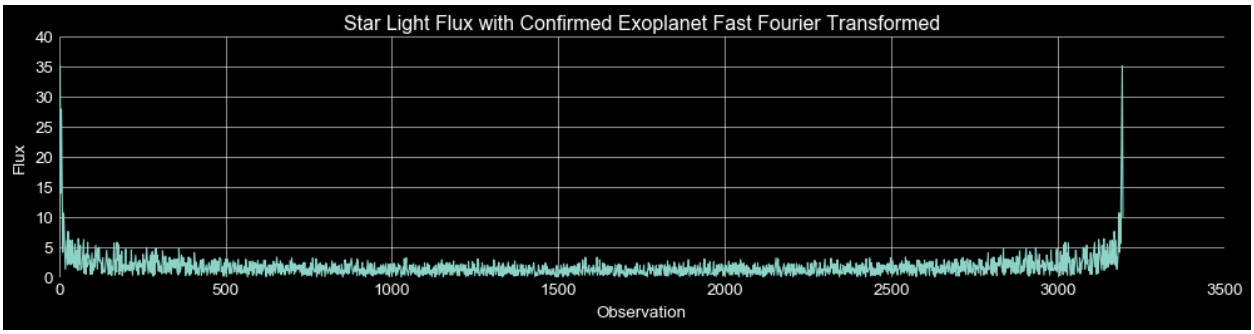


Applying the Fast Fourier Transformation allows the signal allows it to be broken down to its visible frequencies. Plotting both these allows us to see the patterns present in the flux measurements (see below).

With the Fourier Transformation the first peak is the overall indicator of the frequency of the series. Subsequent harmonics (peaks) indicate the presence of an exoplanet.

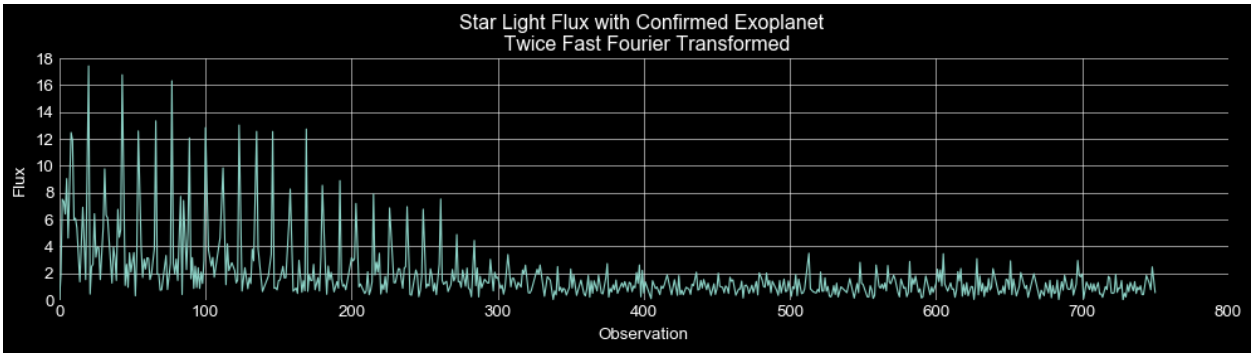
For comparison here are two wave forms that illustrating the pattern of harmonics when an exoplanet is and is not present after the Fast Fourier Transformation:

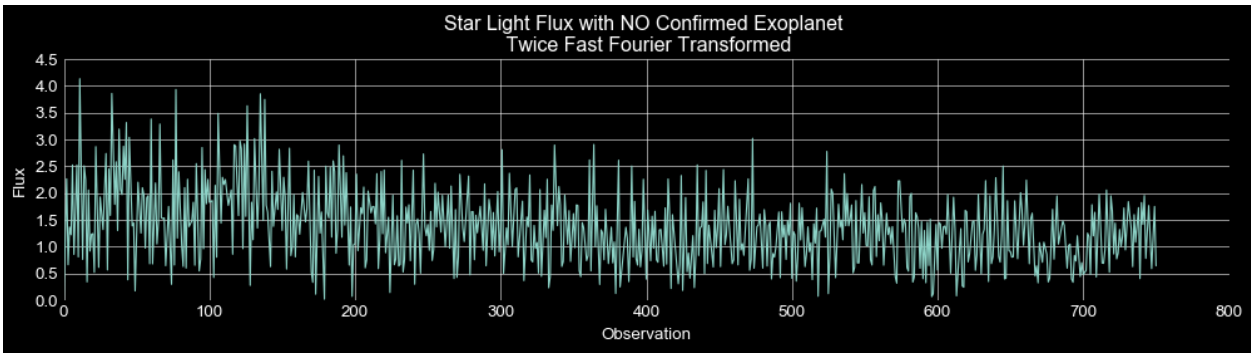




The wave created is also clearly symmetric. To reduce the number of features we can remove half of them leaving ~1600 features. In addition to see if more information could highlighted I re-applied the Fast Fourier Transformation and then culled half the features again. Since the first readings are the basic frequency of the star, the remaining harmonics have the information we are interested in for features about exoplanet – I dropped those 50 features. This leaves a dataset of only 750 features down from 3178.

Here are two examples of transformed data, one with and one without an exoplanet transiting.





Classifiers and Gridsearch

Defining models and their hyper-parameters range that will run through the Gridsearch focusing on string recall so as not to tune the model for non-exoplanets

- Random Forest
- AdaBoost
- XGBoost
- Decision Tree
- K Neighbors
- Bagging

	estimator	min score	max score	mean score	max depth	min samples leaf	min samples split	n estimators	n neighbors
69	KN	0.538	0.833	0.707	N/A	N/A	N/A	N/A	7
68	KN	0.461	0.833	0.681	N/A	N/A	N/A	N/A	5
67	KN	0.384	0.75	0.600	N/A	N/A	N/A	N/A	3
49	AdaBoost	0.153	0.667	0.384	N/A	N/A	N/A	18	N/A
11	Random Forest	0.153	0.667	0.384	3	10	15	30	N/A

The strongest mean sore was K Nearest - setting the parameters to n_neighbors = 7

Running the model with Neighbors = 7:

Recall score: 80.00%
F1 score: 14.55%

Classification Report: KNeighbors

	precision	Recall	f1-score	Support
0 No Exoplanet	1.00	0.92	0.96	565
1 Exoplanet	0.08	0.80	0.15	5
avg / total	0.99	0.92	0.95	570

With a recall of 80% 4 out of 5 stars with exoplanets planets were correctly identified.

Conclusions & Acknowledgements

The double Fast Fourier Transformation did not yield anywhere near as good results than those from the Kaggle competition. The shape of the distributions held some promise but did not deliver.

Further work could be done investigating the alternate dataset to see how well they can predict given less imbalance. As well as plotting out an AUROC curve for the different models to see how well they each do.

The Future, Summer 2018 and beyond:

The power will soon drain from Kepler and the campaigns come to a close. The TESS spaceship will go live later in the summer of 2018 and results will soon be pipelined to citizen scientists. More data will be coming along with more opportunities.

Acknowledgements:

Juan Felipe for the code used to create the grid search cross validation incorporating the different classifiers: https://www.kaggle.com/jfcgon