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. Initially conceived in the mid 1990's there was much skepticism that there were exoplanets and that they could be identified as such using the transiting method.

Exoplanet Count

Kepler:

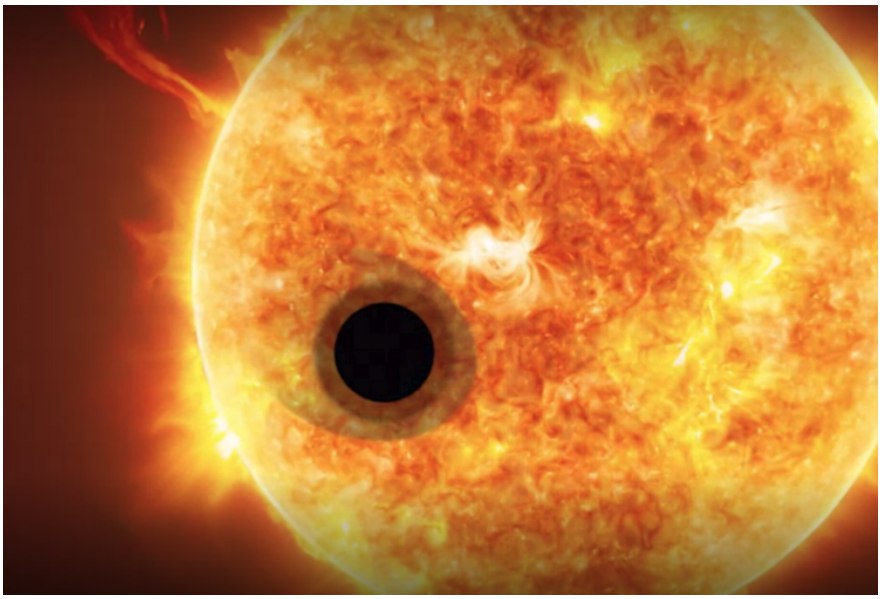
Candidates: 2,244
Confirmed: 2,327
Small Habitable Zone Confirmed: 30

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. The aim is find earth sized planets around suns that have an orbit period that puts them in the 'goldilocks zone', where they can have both an atmosphere and water.



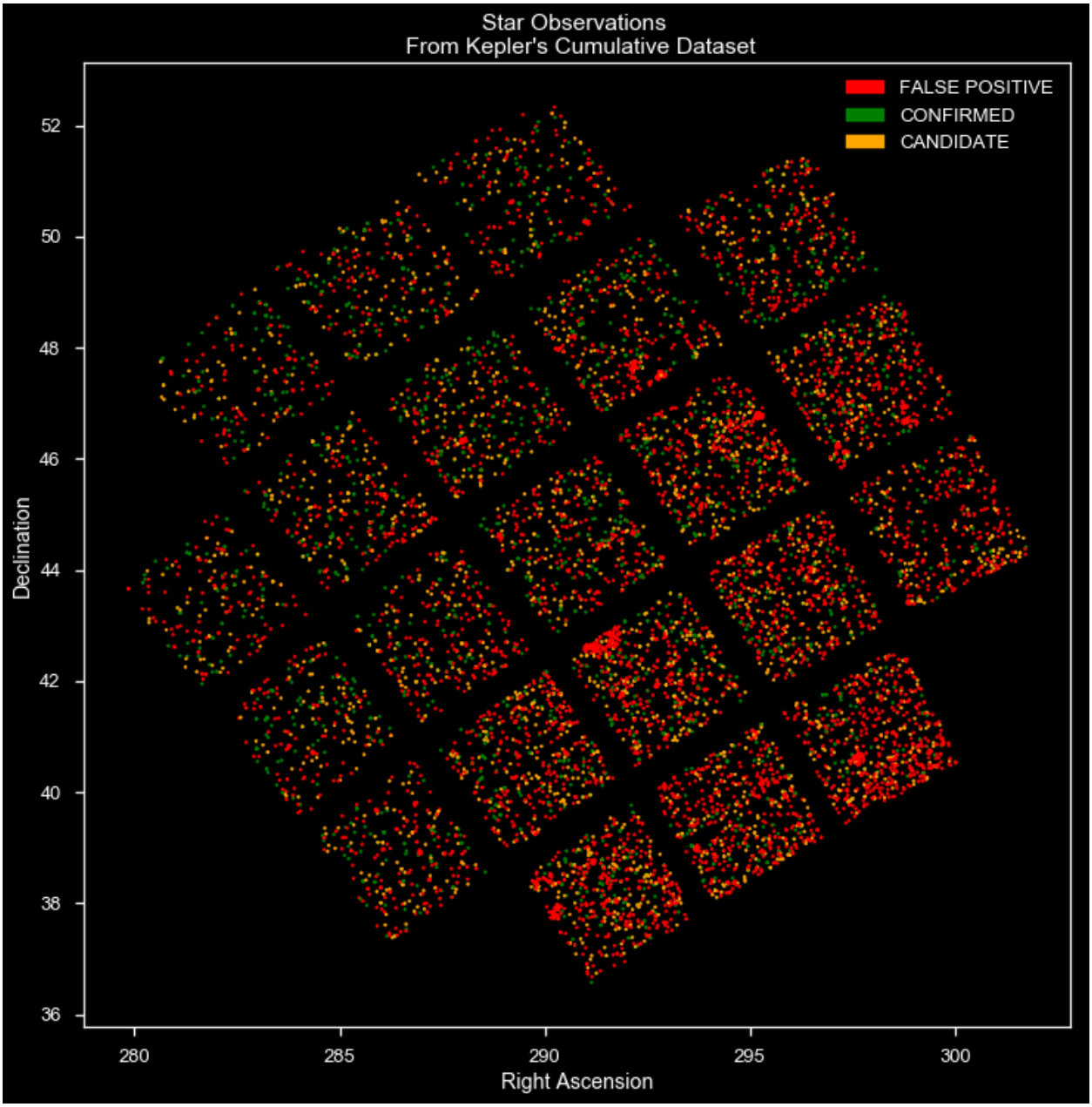
Data from Kepler Mission 2 'Hunt for Exoplanet' Kaggle competition

The dataset I used is from the 2017 Kaggle completion. 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 has been pre-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.

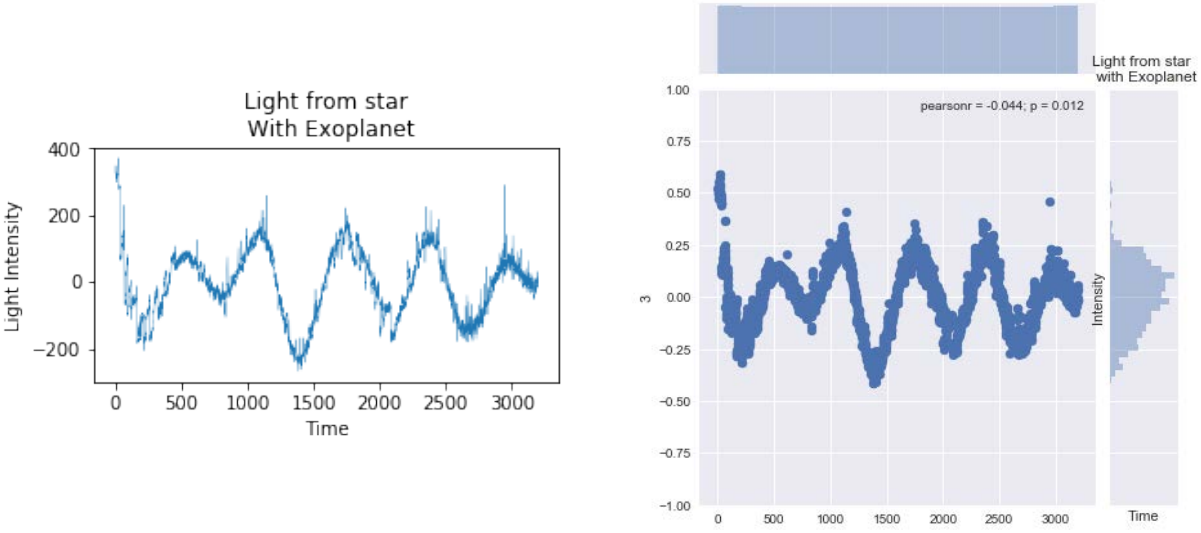
An exoplanet(s) transiting the star from our point of view will cause the light to dim. Depending on the size of the object, the angle, and duration of the transit the light will dim by a certain amount for a certain period. The spaceship is effectively “detecting a flea crossing the beam of a headlight 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

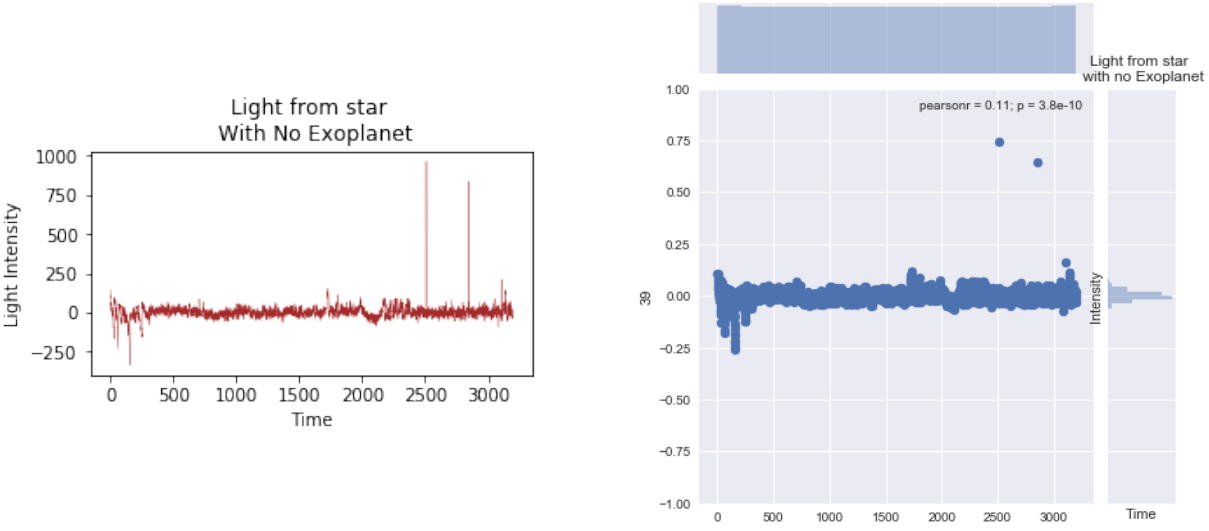
Star#	FLUX.1	FLUX.3	FLUX.3	FLUX.4	FLUX.5	FLUX.6	FLUX.7	FLUX.8	FLUX.9
37	-141.22	-81.79	-52.28	-32.45	-1.55	-35.61	-23.28	19.45	53.11
38	-35.62	-28.55	-27.29	-28.94	-15.13	-51.06	2.67	-5.21	9.67
39	142.40	137.03	93.65	105.64	98.22	99.06	86.40	60.78	45.18
40	-167.02	-137.65	-150.05	-136.85	-98.73	-103.14	-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.

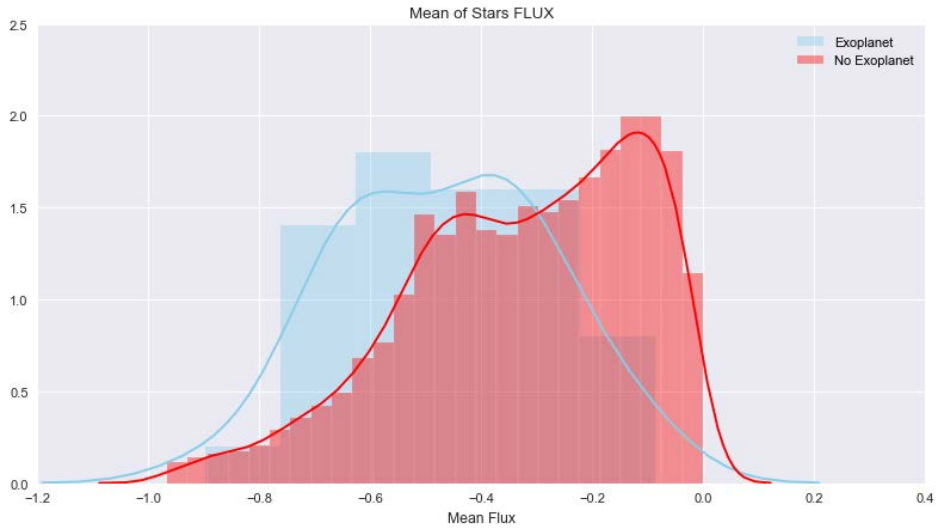
Time Series and Density Plot
Star with an Exoplanet

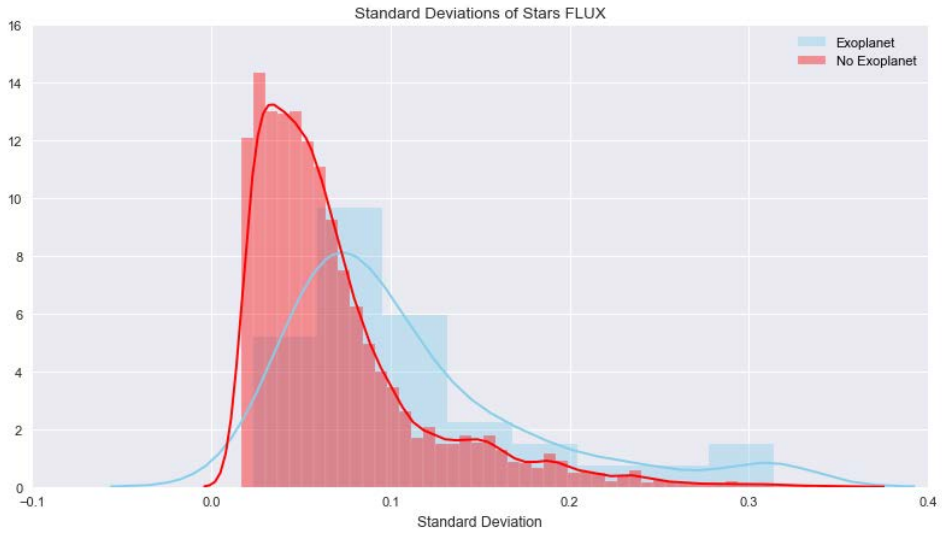


Time Series and Density Plot
Star with no Exoplanet

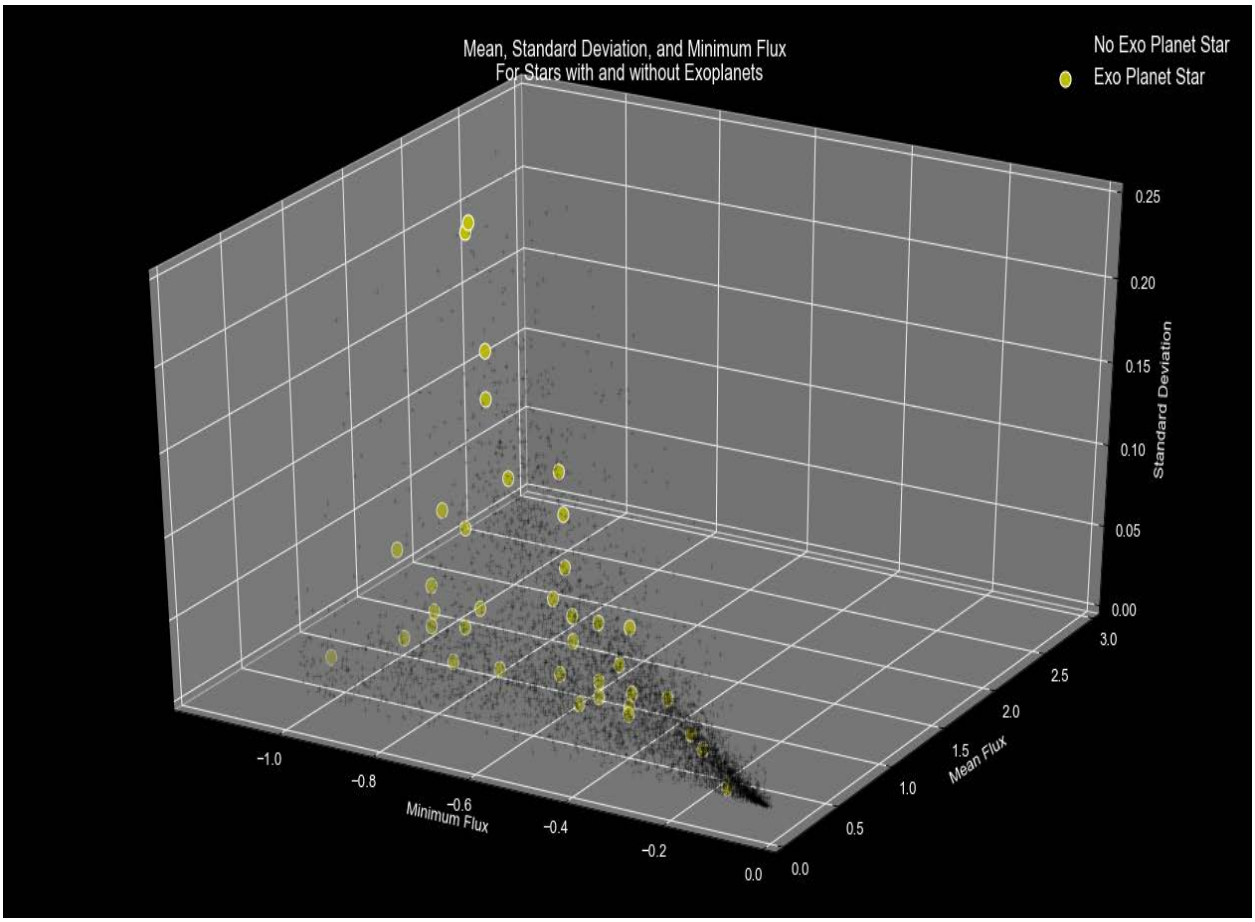


Plotting the means and standard deviations of the Flux of the stars:

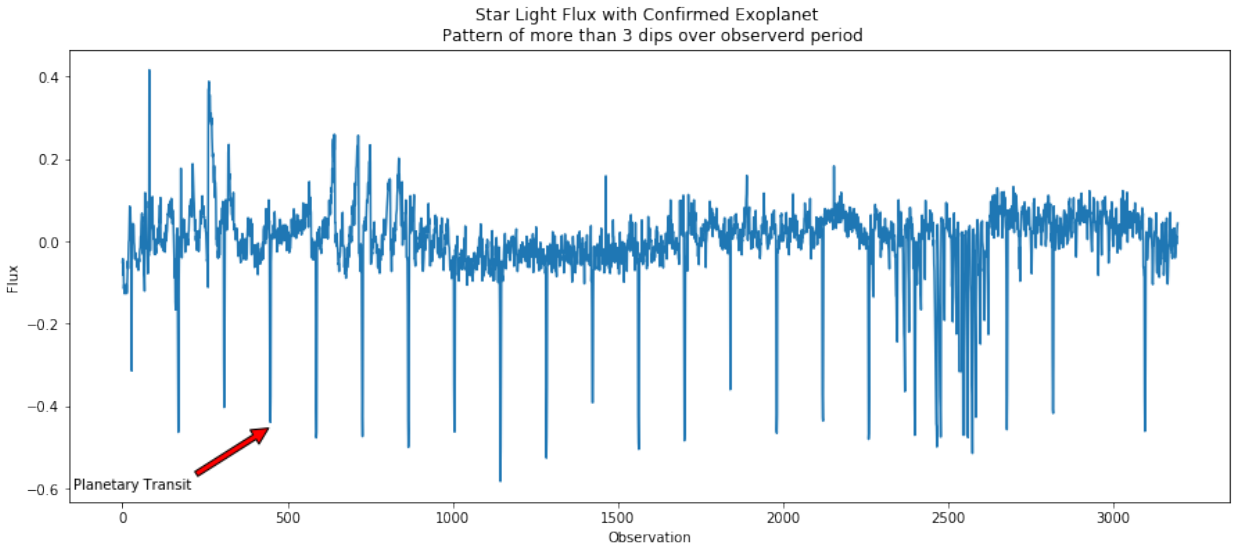




If the standard deviations, means and the minimums are plotted out in 3 dimensions, we can see that the stars with no exoplanets have a smaller standard deviation and their minimum value for flux is closer to the mean.

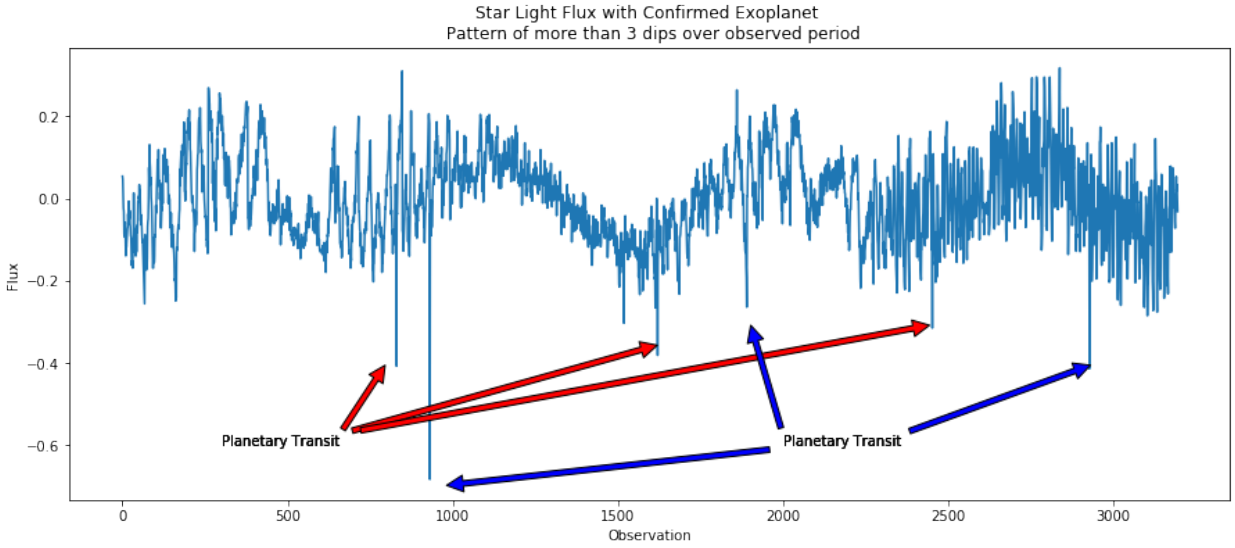


In these plots it is easier to see where the exoplanet indicators might be:



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.

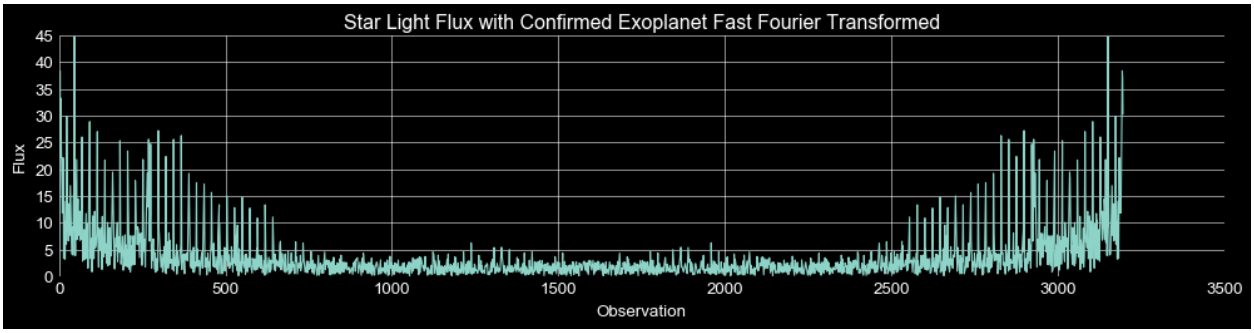
The next plot suggests the presence of more than one planet with 2 separate dips

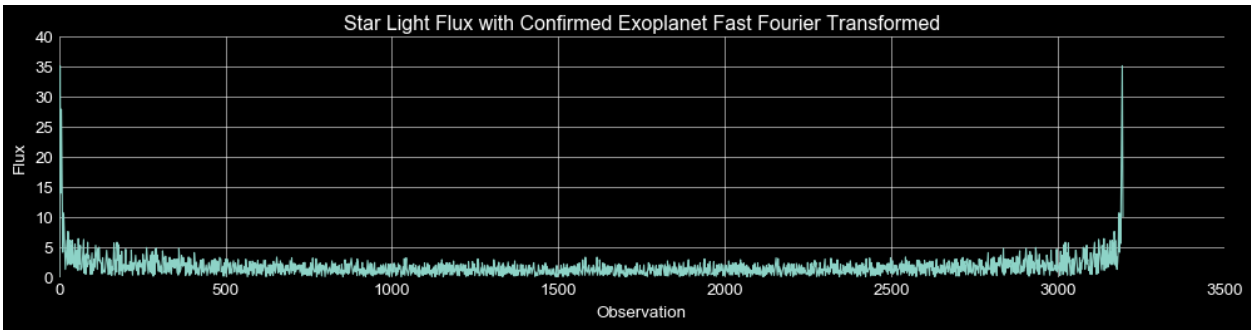


Applying the Fast Fourier Transformation allows the signal allows it to be broken down to its visible frequencies. Plotting both these allows us to see the patterns present in the flux measurements (see below).

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.

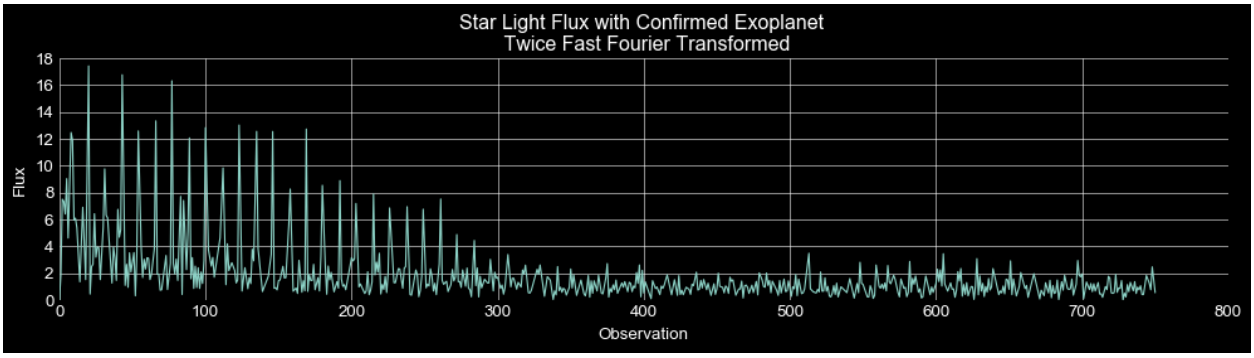
For comparison here are two wave forms that illustrating the pattern of harmonics when an exoplanet is and is not present after the Fast Fourier Transformation:

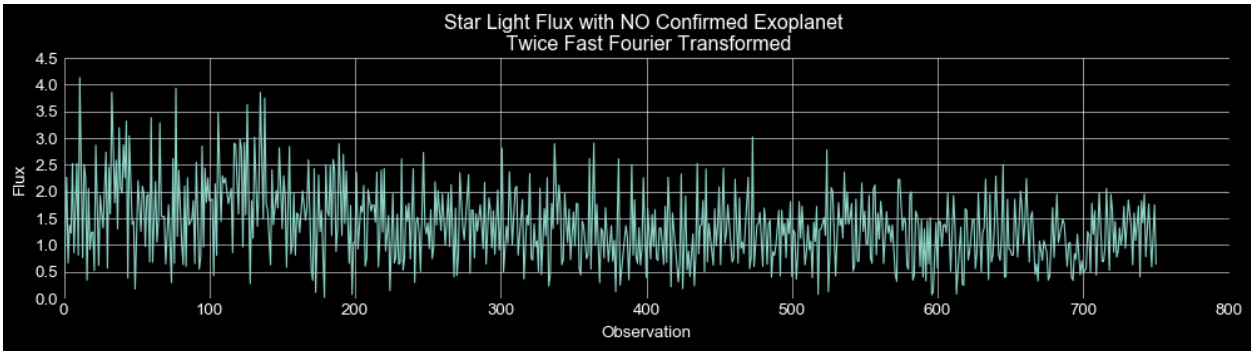




The wave created is also clearly symmetric. To reduce the number of features we can remove half of them leaving ~1600 features. In addition to see if more information could highlighted I re-applied the Fast Fourier Transformation and then culled half the features again. Since the first readings are the 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

A suite six models was instantiated and a range hyperparameters applied. These ran through the Gridsearch with my focus highest mean Recall (True Positive). This is done to avoid 'accuracy paradox' where there is a high accuracy score that is weighted towards non-exoplanets.

The data is heavily imbalanced with 37 targets out of 5050 = 0.7%. To account for this the pipeline includes oversampling. Using SMOTE (Synthetic Minority Oversampling Technique) in the model will prevent overfitting and poor accuracy. I also ran the model without SMOTE to compare.

The main metric I chose was the *Recall score* to count the percentage of True Positives. The additional metric I also investigated was the *F1 score*, the weighted average of the Recall and *Precision scores* (Precision score being the amount of accurate positives predicted).

The six models were:

- **Decision Tree**
    - Selects features and branches out into trees towards a correct or incorrect classification. It can be prone to overfitting such that the model only works when subsequent data is similar to the original set.
- **Random Forest**
    - A collection of random decision trees that prevent the issue of overfitting described above by creating subsets of the features used. It can be slower computationally.
- **AdaBoost (Adaptive Boost)**
    - Boosting takes features that have weak learning (predictions that have slightly more than 50% being accurate) and combines them to create a stronger predictive model.
- **XGBoost (Extreme Gradient Boost)**
    - Gradient boosting sums an ensemble of classification and regression trees, Extreme Gradient Boosting uses a model with more regularization to control overfitting.
- **K Neighbors**
    - Classifies by finding the most similar data in a groups proximity. This would appear to work well with our dataset because of the time series nature of the data.
- **Bagging (Bootstrap Aggregating)**
    - An ensemble learning algorithm that uses sampling (bootstrapping) and aggregates weak models.

Top Five Highest Average Recall (True Positive)

Grid Search and Hyperparameter Scores

Model Name		Recall Score Summary			Hyperparameter Settings…				
Rank	Estimator	Min Recall score	Max Recall score	Mean Recall score	max depth	min samples leaf	min samples split	n estimators	n neighbors
1	KN	0.538	0.833	0.707	N/A	N/A	N/A	N/A	7
2	KN	0.461	0.833	0.681	N/A	N/A	N/A	N/A	5
3	KN	0.384	0.75	0.600	N/A	N/A	N/A	N/A	3
4	AdaBoost	0.153	0.667	0.384	N/A	N/A	N/A	18	N/A
5	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

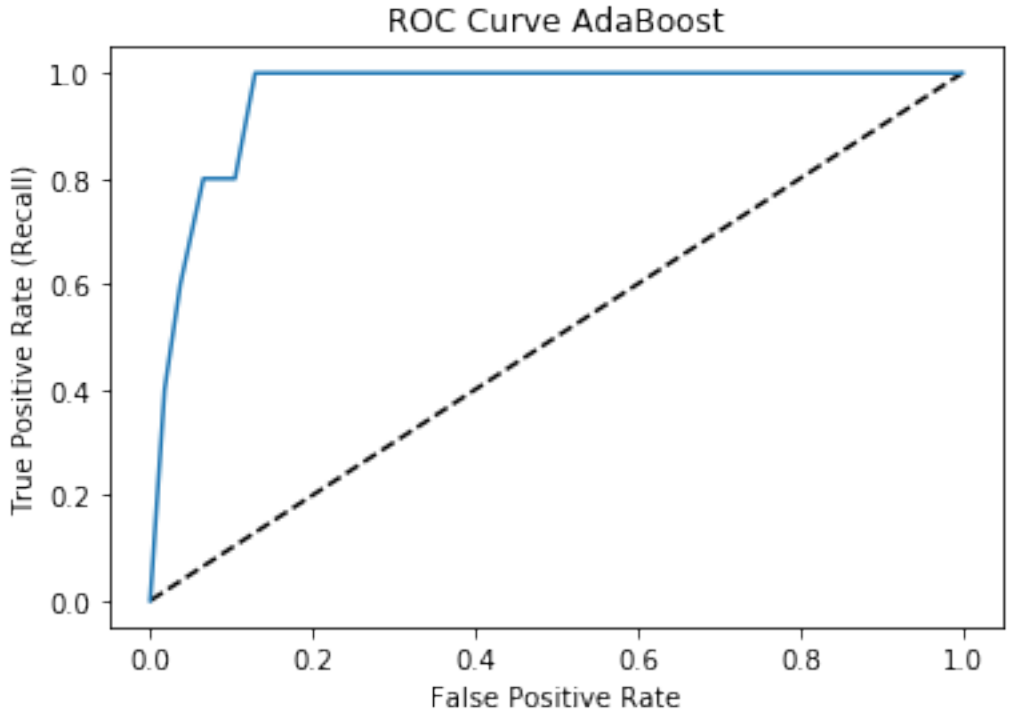
Running the model with the highest mean Recall score : Neighbors = 7

Recall score: 80.00%
F1 score: 14.55%

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

Classification Report: K Neighbors

	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



Here is a performance ranking of the 6 different classifiers, sorted by the highest Mean Recall Score.

Classifier Ranking	Classifier	Min Recall Score	Max Recall Score	Mean Recall Score
1st	KNClassifier	0.538462	0.833333	**0.707265**
4th	AdaBoostClassifier	0.307692	0.75	0.463675
5th	RandomForestClassifier	0.0833333	0.666667	0.429487
8th	XGBClassifier	0.25	0.583333	0.405983
42nd	DecisionTree	0.25	0.384615	0.32265
50th	BaggingClassifer	0	0.333333	0.194444

K Neighbors and AdaBoost out performed the other classifiers when the data was passed through the Fourier Transformation twice.

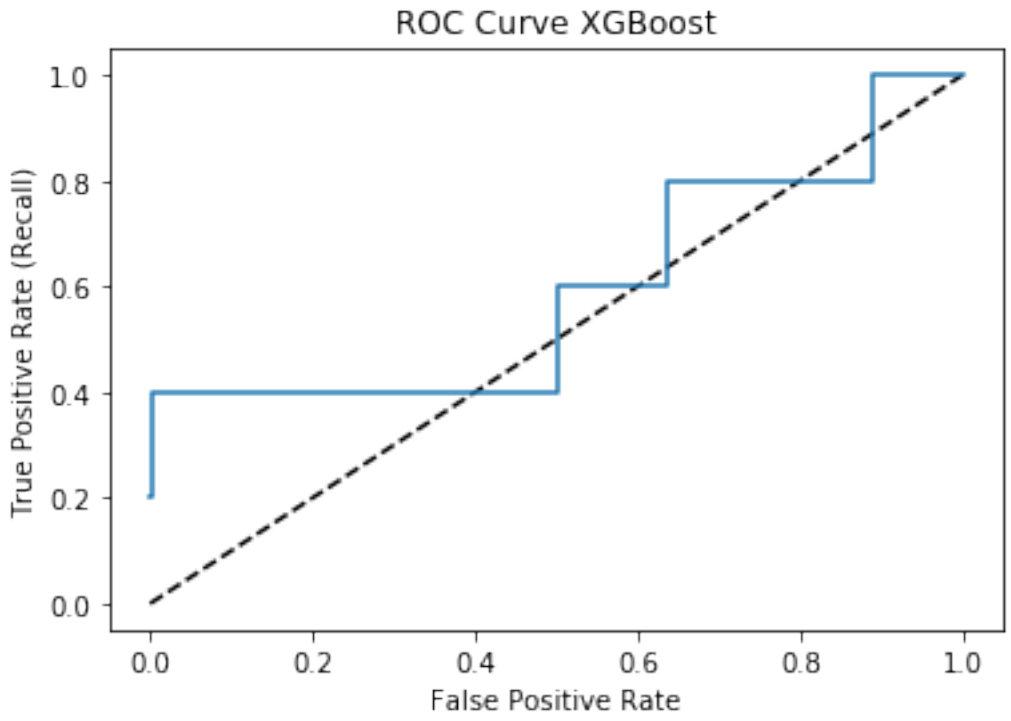
Running the same model to score against best F1 score, XGBoost was in the top five. But a high F1 score is not as useful as we are trying to maximize true positives.

Classifier	Min F1 Score	Max F1 Score	Mean F1 Score
XGBClassifier	0.352941	**0.636364**	0.481283
XGBClassifier	0.25	0.6	0.372222
XGBClassifier	0.285714	0.56	0.415238
XGBClassifier	0.363636	0.555556	0.431397
XGBClassifier	0.375	0.545455	0.480731

Running the XGBoost classifier with the values from the best F1 score provides a low recall score of: `40.00%`

Running the classifier grading model against the same dataset without the oversampling produces very poor average recall results.

Classifier	Min Recall	Max Recall	Mean Recall
XGBClassifier	0.153846	0.416667	0.301282
XGBClassifier	0.153846	0.416667	0.301282
XGBClassifier	0.153846	0.416667	0.301282
XGBClassifier	0.230769	0.333333	0.271368
AdaBoostClassifier	0.153846	0.333333	0.245726
AdaBoostClassifier	0.153846	0.333333	0.273504
AdaBoostClassifier	0.153846	0.333333	0.273504
XGBClassifier	0.0833333	0.333333	0.241453



Conclusions & Acknowledgements

The K neighbors algorithm with neighbors set to 7 achieved the highest recall score/true positives due to the nature of the transformed data with its cycles. For overall F1 accuracy the XGBoost method worked well for the same reason, due to the nature of the weak features. The decision tree algorithm might work better with untransformed data.

Compared to the Kaggle competition submissions, the double Fast Fourier Transformation did not yield anywhere near as good results than those from the Kaggle competition.

The shape of the distributions of light flux held some promise, so potentially if they are distinct enough from the non-exoplanet distributions that may yield some results, or analyzing the peaks and troughs patterns through more complex transformations.

Further work could also be done investigating the alternate dataset to see how well a prediction can be with a larger target set and less imbalance if the light flux data becomes available.

The Future : Summer 2018 and beyond:

The power will soon drain from Kepler's batteries and the campaigns will come to a close. The TESS (Transiting Exoplanet Survey Satellite) launched in April 2018 will go live later in the summer and results will soon be pipelined to citizen scientists. Much 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