STAT 155 Final Presentation

Fahham Kurji

Introduction

- Question: Does using a certain discovery method correlate with different exoplanet parameters being measured?
- Context: There are multiple discovery methods, i.e. Transit and Radial velocity that heavily vary in how they discover planets. The question asks if these differences in discovery methods might correlate with the parameters of the planets they discover, i.e. mass, radius etc.

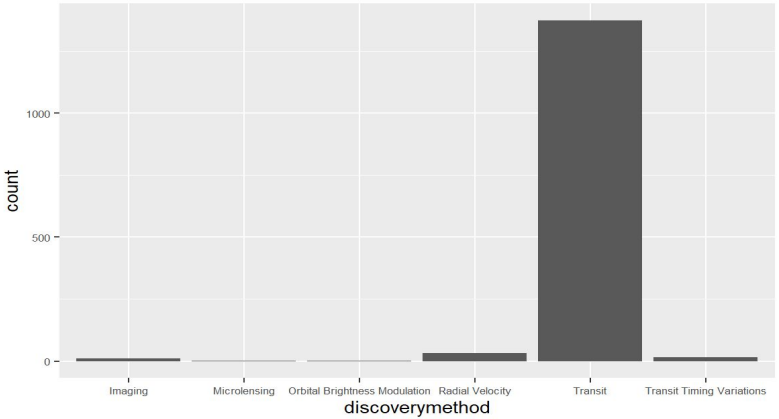
Data Wrangling

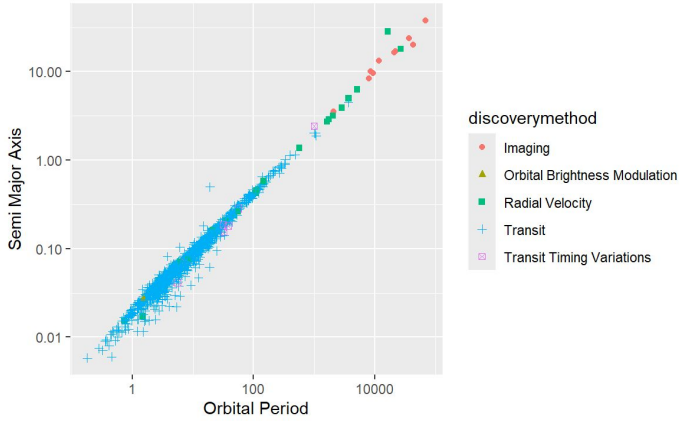
- Initial data source was NASA's exoplanet archive. Currently has 67 variables with 5912 data points.
- Most of the variables were filtered out leaving only relevant ones. Any NA values were also omitted decreasing the size of the dataset to 1435 variables from 5912.

Variable	Description
discoverymethod	Method used to discover exoplanet
pl_orbper	Orbital Period(Days)
pl_orbeccen	Eccentricity
pl_orbsmax	Semi-Major Axis(AU)
pl_massj	Planet Mass(Jupiter Masses)
pl_rade	Planet Radius(Earth Radii)
pl_eqt	Equilibrium Temperature(Kelvin)
st_rad	Stellar Radius(Solar Radii)
st_teff	Stellar Effective Temperature(Kelvin)
st_mass	Stellar Mass(Solar Masses)

Exploratory Data Analysis

- Univariate Analysis: Compared how much each discovery method was used.
- Bivariate Analysis: Plotted orbital period vs. semi major axis to understand general trends to look for.



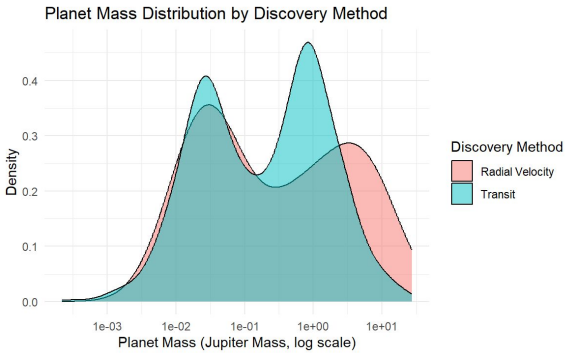


- Transit is by far the most used discovery method.

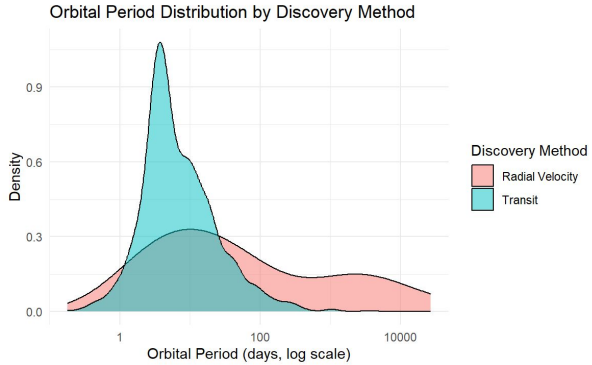
- The graph above shows a trend towards planets discovered with transit having generally lower orbital periods and semi major axes relative to other methods. Both the axes are put on a log scale.

Exploratory Data Analysis

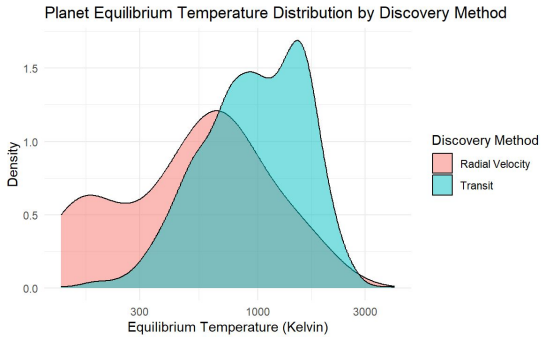
- Multivariate Analysis: Using transit and radial velocity, the two most common discovery methods, density plots were made with different variables to look for any relevant trends.



- Both transit and radial velocity follow similar trends to the mass of the planets they find, with transit have a higher density. This is likely due to the higher number of transit samples.



- Transit is heavily skewed towards a low orbital period while radial velocity covers a wider range of measurements.



- Radial velocity follows the trend of detecting lower to medium temperature planets while transit detects medium to high temperature planets.

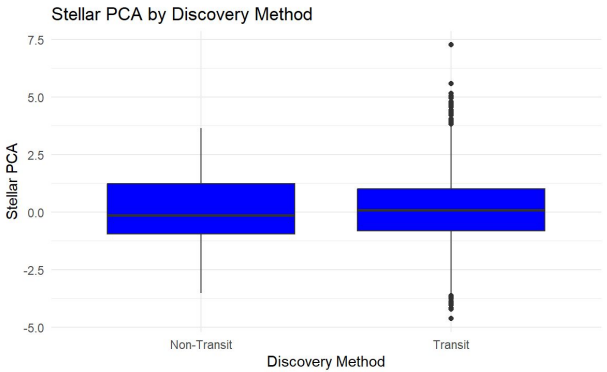
Modeling

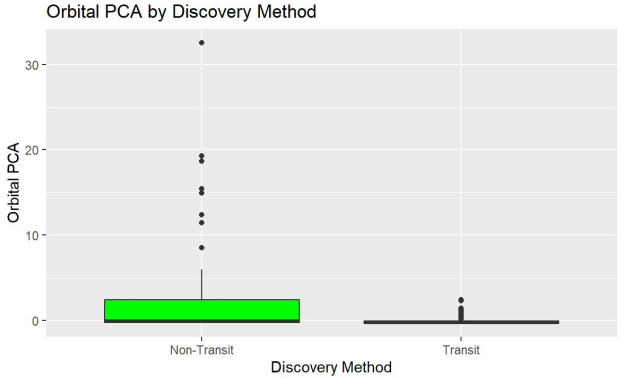
- PCA is the machine learning method used. Due to the large amount of transit samples, any other method was changed to the variable “Non-Transit” in order to create a statistically significant amount of data points.
- Prior to running PCA, orbital, planetary, and stellar parameters were manually grouped together in order to summarize each one’s correlation to a discovery method. PCA is ran three times with the first component of each capturing the maximum variance in the data.
- PCA also works in this dataset as it reduces the number of features present.

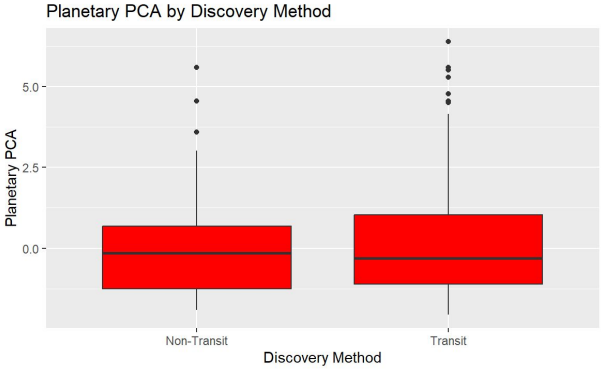
Variable	Description
discoverymethod	Method used to discover exoplanet
pl_orbper	Orbital Period(Days)
pl_orbeccen	Eccentricity
pl_orbsmax	Semi-Major Axis(AU)
pl_massj	Planet Mass(Jupiter Masses)
pl_rade	Planet Radius(Earth Radii)
pl_eqt	Equilibrium Temperature(Kelvin)
st_rad	Stellar Radius(Solar Radii)
st_teff	Stellar Effective Temperature(Kelvin)
st_mass	Stellar Mass(Solar Masses)

Modeling

- The results show the first principal component derived from each iteration of PCA done.







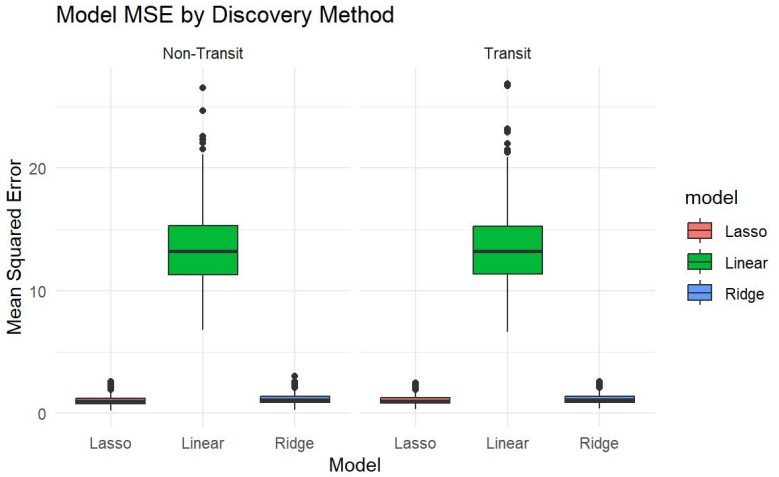
- The medians of both the Transit and Non-Transit groups are very similar, suggesting that the method of discovery does not strongly correlate with overall differences in stellar characteristics. However, the Transit group shows more outliers, which may reflect the larger sample size.

- The interquartile range for the Transit group is much smaller, suggesting that Transit methods tend to detect planets with similar orbital parameters. In contrast, the Non-Transit group has a wider interquartile range and more outliers, indicating greater variability in orbital parameters.

- The median and interquartile ranges of both discovery methods are similar suggesting that the discovery method does not correlate with differences in planetary parameters. Both methods also have some outliers.

Monte Carlo Simulation

- Purpose: To compare the predictive performance of linear, ridge, and lasso regression when it comes to predicting target variables.
- Design: Simulates three continuous variables that are the input, i.e. planet temperature with the output being a predicted target variable, i.e. planet radii. Performance of the regression models is also compared between discovery methods. The entire simulation is ran 1,000 times to get a better understanding of the distribution. How well a regression model did is determined using Mean Squared Error. In a real world situation, the model would be shown real data and could use that as a foundation to act as a predictive model.



- The table above suggests that linear regression is the worst of the three due to the large spread of its Mean Squared Error. Lasso and Ridge low interquartile range and median make them good candidates for being useful predictive models.

Summary and Reflection

- Question: Does using a certain discovery method correlate with different exoplanet parameters being measured?
- Not all exoplanet parameters are strongly correlated with the discovery method. Orbital parameters such as semi-major axis and orbital period show a clear correlation, with Non-Transit methods detecting a wider variety of orbits. In contrast, stellar and planetary parameters appear to have weaker or negligible correlations with the method of discovery, as indicated by the similar PCA distributions across Transit and Non-Transit.

Summary and Reflection

- Being my first time using GitHub, I learned how to navigate the UI as well as construct a repo that could be reproduced.
- I also learned the importance of trying new ideas in machine learning, i.e. my initial model used k-means clustering, however I got much better results with PCA.

Summary and Reflection

- There aren’t many ethical implications to this project.
- This project can help us understand biases present in the current exoplanet population, i.e. the majority of exoplanets having low orbital periods. Rather than the fact that these are the most common planet type, transit correlating to finding these types of planets would help explain why we’ve found a disproportionate number of them.
- The results of this project could be very different from today as the current exoplanet archive is being constantly updated. Around 250 new planets have been added since the beginning of this project in April.