This project looked at how different methods of discovering exoplanets—like Transit and Non-Transit techniques—relate to the properties of the planets and their stars. We used a real dataset from NASA's Exoplanet Archive and focused on a few key variables, such as orbital characteristics, planet size and temperature, and the properties of the host star. We applied a dimensionality reduction technique to group similar variables together and then compared how these grouped traits differ between the two discovery methods. The clearest difference appeared in orbital features, which varied more depending on how the planet was found, while other features like stellar temperature or planet size didn't show much variation.

To dig deeper, we created a simulation to mimic this kind of data. We generated fake exoplanet-like observations with three traits, plus an outcome influenced by some of them. We also randomly assigned each simulated planet to either the Transit or Non-Transit category. This allowed us to test how different prediction models perform under controlled conditions. The simulation was helpful because it gave us a way to know exactly which traits mattered, so we could see if the models were picking up on the right patterns.

We tested three models: regular linear regression, ridge regression, and lasso regression. These models were run repeatedly on both Transit and Non-Transit groups, and we recorded how well they predicted the outcome each time. In general, ridge and lasso models gave better results because they're designed to handle data where some traits might be irrelevant or correlated. The project showed that combining real-world analysis with simulations can help us better understand both the data and the strengths and weaknesses of different modeling techniques.