Analysis Plan for Kepler Data

1 Specify the Outcome and Set of Predictors of Interest

- **Outcome**: The outcome variable of interest is the **semimajoraxis**, which is continuous.
- **Predictors**: Predictors include parameters such as mass, radius, period, eccentricity, hoststar_mass, hoststar_radius, hoststar_metallicity, etc.
- **Missing data**: will be omitted from the analysis and a complete-case analysis will be coducted.

2 Specify the Functional Form of the Relationship between the Outcome and Predictors

- Iteration 1. Consider a Gaussian regression model with a simple linear predictor for each predictor variable initially.
- Iteration 2. Explore a Gaussian regression model with a log-transformed response and simple linear predictors for each predictor variable.
- Iteration 3. In a model without transformation of the response, include more complex functional forms through
  - Bayesian fractional polynomials
  - Bayesian symbolic regression
  - Bayesian generalized nonlinear models
- By default assume fixed effects only. However, assess the reasonability of adding random effects for planets within the same planetary system. Or those discovered in the same range of years.

3 Decide on the Prior Distributions for all Parameters in the Model

- Utilize Jeffreys' priors as the default option for the parameters
- Consider robust g-prioprs as the alternative in case this part of the plan is reiterated
- Consider adjusting prior inclusion probabilities in case of having knowledge, but otherwise use default values of probability of inclusion of 0.5 for every covariate.

4 Choose and Run Inference Algorithm

- If using the FBMS package `https://cran.r-project.org/web/packages/FBMS/index.html`, run the GMJMCMC or MJMCMC algorithm depending on whether you decide to work with the linear models or more flexible models like BGNLM or fractional polynomials.
- Choose tuning parameters carefully, particularly the number of populations and parallel threads, by default use 20 threads and 20 populations.

5 Check the Convergence Plots

- Use internal GMJMCMC diagnostics from the FBMS package to assess convergence via the `diagn_plot` function. If MJMCMC is used, consider standard Rhats for the inclusion pribabilities.
- **If** convergence is lacking (unstable summary accross last populations for the best marginal log posterior), **revisit step 4** and rerun the inference **with increased** numbers of threads, populations, and iterations per generation of GMJMCMC.

6 Predictive Performance Checking

- Use predictive checks to identify model shortcomings regarding variable selection.
- Evaluate predictive performance using plots, Root Mean Squared Error (RMSE), or R^2.
- When comparing models with transformed responses to those without transformations, transform the response back to the original scale before computing predictive scores for fair comparisons with models that were not transofmed.
- If there is enough data, use a hold-out test set for checking the predictive performance or another reasonable validation scheme.

7 Discussion

- Choose the final model using the median probability rule or the most probable model in terms of the posterior, motivate this choice.
- Interpret the resulting model and investigate how changes in independent variables influence the response.
- Compare the model results to the true underlying law (if feasible).

- **If** results are **unsatisfactory**, return to **step 2** and **reiterate**. The stopping criteria is getting explainable model with good posterior predictive performance.

Suggestions in Case Reiterations from Step 2 are not Helpful

- Incorporate informative priors based on domain knowledge or previous research findings to enhance the Bayesian analysis.
- Consider sensitivity analyses to assess the impact of prior choices on the results and conclusions. E.g. adjust reasonably prior inclusion probabilities.
- Explore alternative Bayesian model structures, such as hierarchical models or Bayesian nonparametric models, to capture potential complex relationships more effectively.
- Utilize Bayesian model averaging techniques to incorporate model uncertainty and provide more robust inference.
- Consider the use of Bayesian model comparison methods, such as Bayes factors or information criteria, to compare competing models and select the most appropriate one.