# NASA Exoplanet Data Analysis in R

Trang Ly

2023-10-04

## Introduction

### NASA Planetary Systems Dataset

Acknowledgement: "This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program."

This archive is designed to provide researchers, astronomers, and enthusiasts with easy access to a comprehensive catalog of data related to exoplanets and their host stars.

Two data sets pertaining to the Planetary Systems are available:

- Planetary Systems, and
- Planetary Systems Composite Data

For the analyses in this project, I will be using the Planetary Systems data. ver. Wed Oct 4 12:36:20 2023

## Analyses

There are over 200 variables and 30,000 observations. For my analyses, I will focus on the following variables, their observations, and relationships:

- discoverymethod: Discovery Method
- disc_year: Discovery Year
- pl_radj: Planet Radius [Jupiter Radius]
- pl_massj: Planet Mass [Jupiter Mass]
- st_rad: Stellar Radius [Solar Radius]
- st_mass: Stellar Mass [Solar mass]

#### Load Packages

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(ExoR)
```

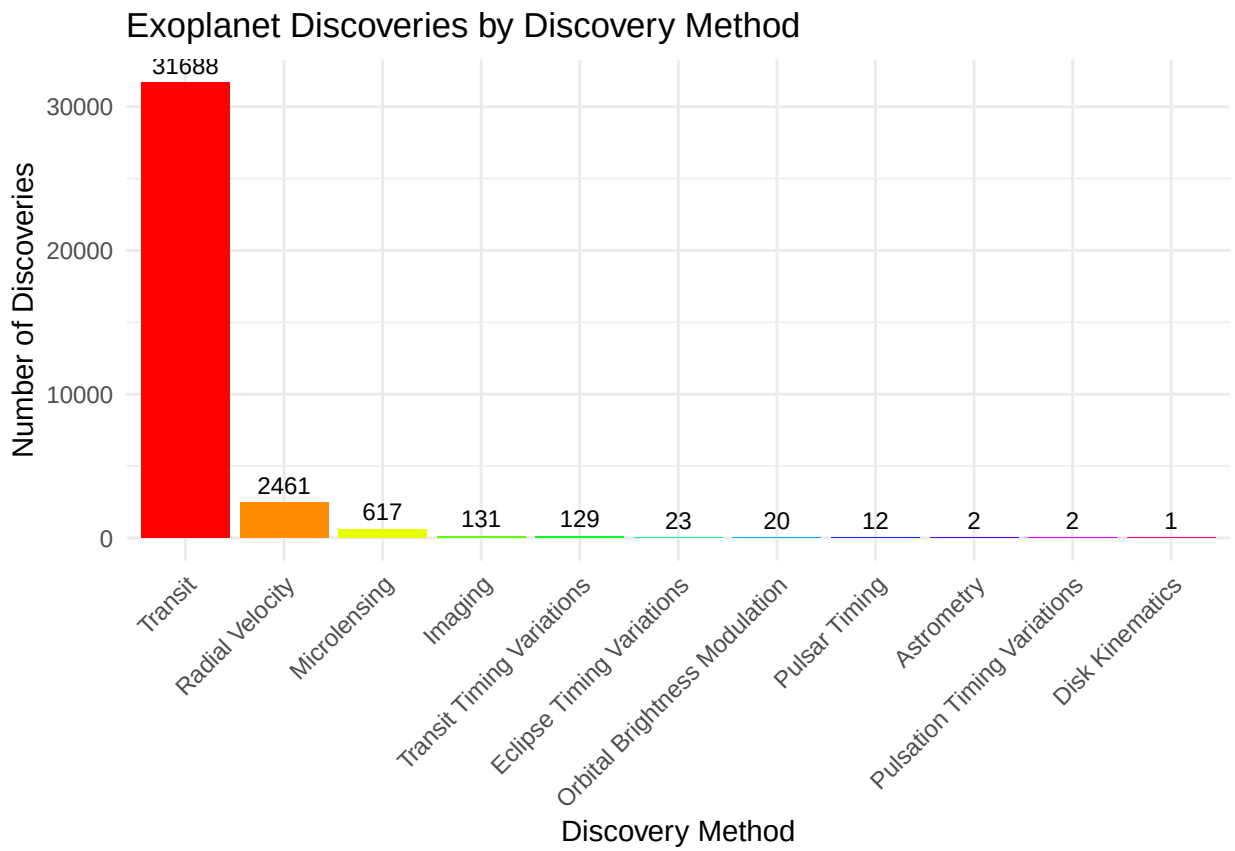
### Load Data

```
# Load the NASA exoplanet data
exoplanets <- read.csv("PS_2023.10.04_12.36.20.csv", skip = 290)
# head(exoplanets)  # preview data
```

## Exoplanet Discovery

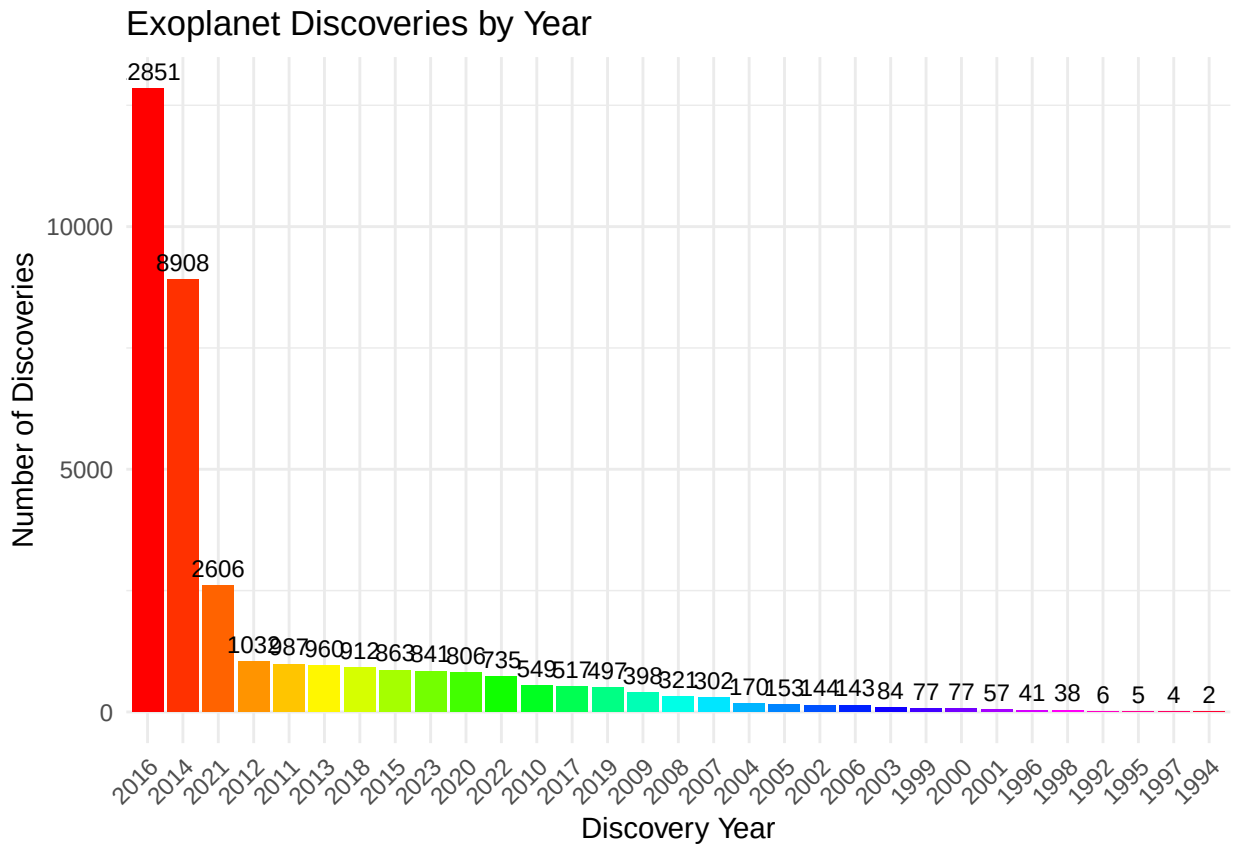
### Distribution of exoplanet discovery methods

```
# A bar plot of exoplanet discovery methods in descending order
create_barplot(exoplanets, discoverymethod, count, "Exoplanet Discoveries by Discovery Method", "Discov
```



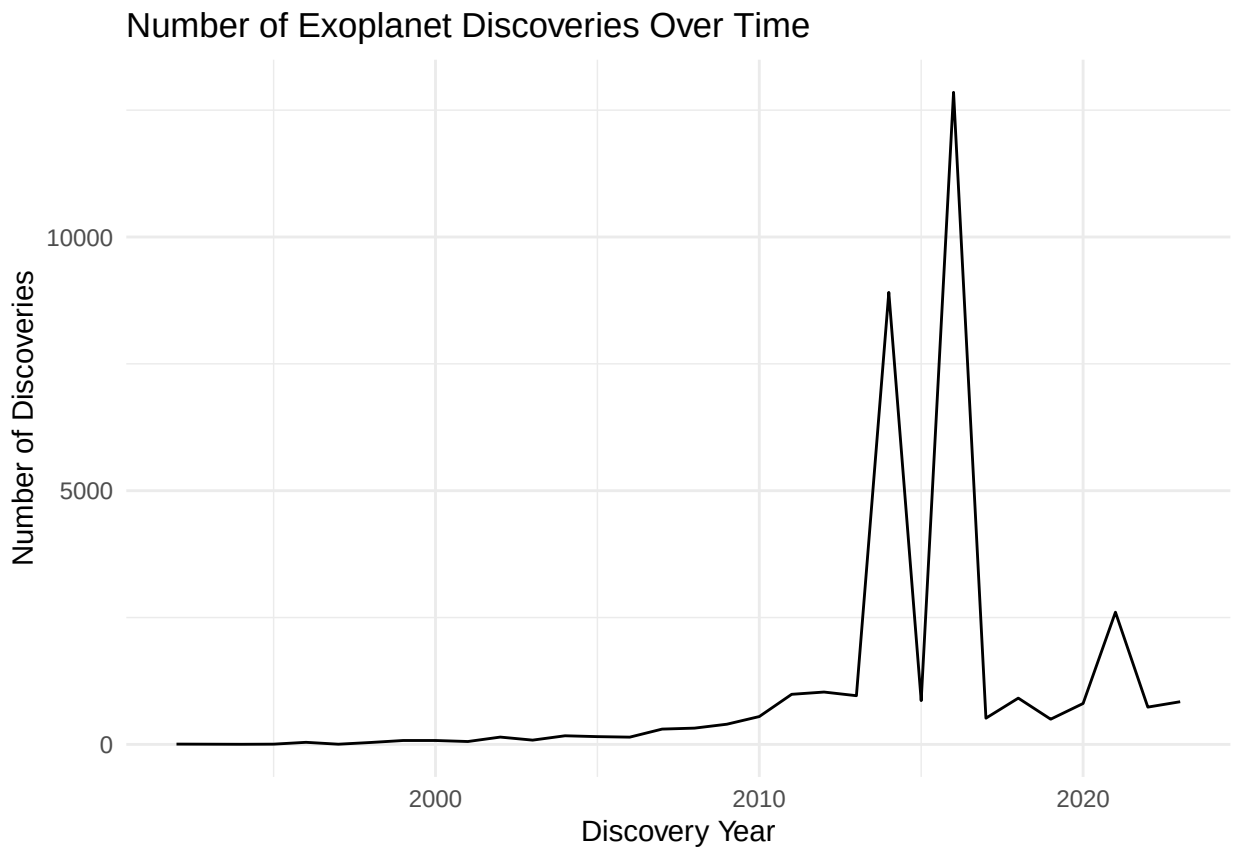
### Distribution of exoplanet discovery years

```
# A bar plot of exoplanet discoveries by year in descending order
create_barplot(exoplanets, disc_year, count, "Exoplanet Discoveries by Year", "Discovery Year", "Number
```




### Trends in exoplanet discovery

```
# A look at exoplanet discoveries throughout the years
# Line plot
create_line_plot(exoplanets, disc_year, count, "Number of Exoplanet Discoveries Over Time", "Discovery
```




**The most used method by year:**

```
most_used_methods <- exoplanets %>%
  group_by(disc_year, discoverymethod) %>%
  summarise(count = n()) %>%
  arrange(disc_year, desc(count)) %>%
  slice(1) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'disc_year'. You can override using the
## `.groups` argument.
```

```
# Print the results
print(most_used_methods)
```

```
## # A tibble: 31 x 3
##    disc_year discoverymethod count
##        <int> <chr>           <int>
##  1      1992 Pulsar Timing       6
##  2      1994 Pulsar Timing       2
##  3      1995 Radial Velocity     5
##  4      1996 Radial Velocity    41
##  5      1997 Radial Velocity     4
```

```
##  6      1998 Radial Velocity    38
##  7      1999 Radial Velocity    77
##  8      2000 Radial Velocity    77
##  9      2001 Radial Velocity    57
## 10      2002 Radial Velocity   134
## # i 21 more rows
```

**Summary**

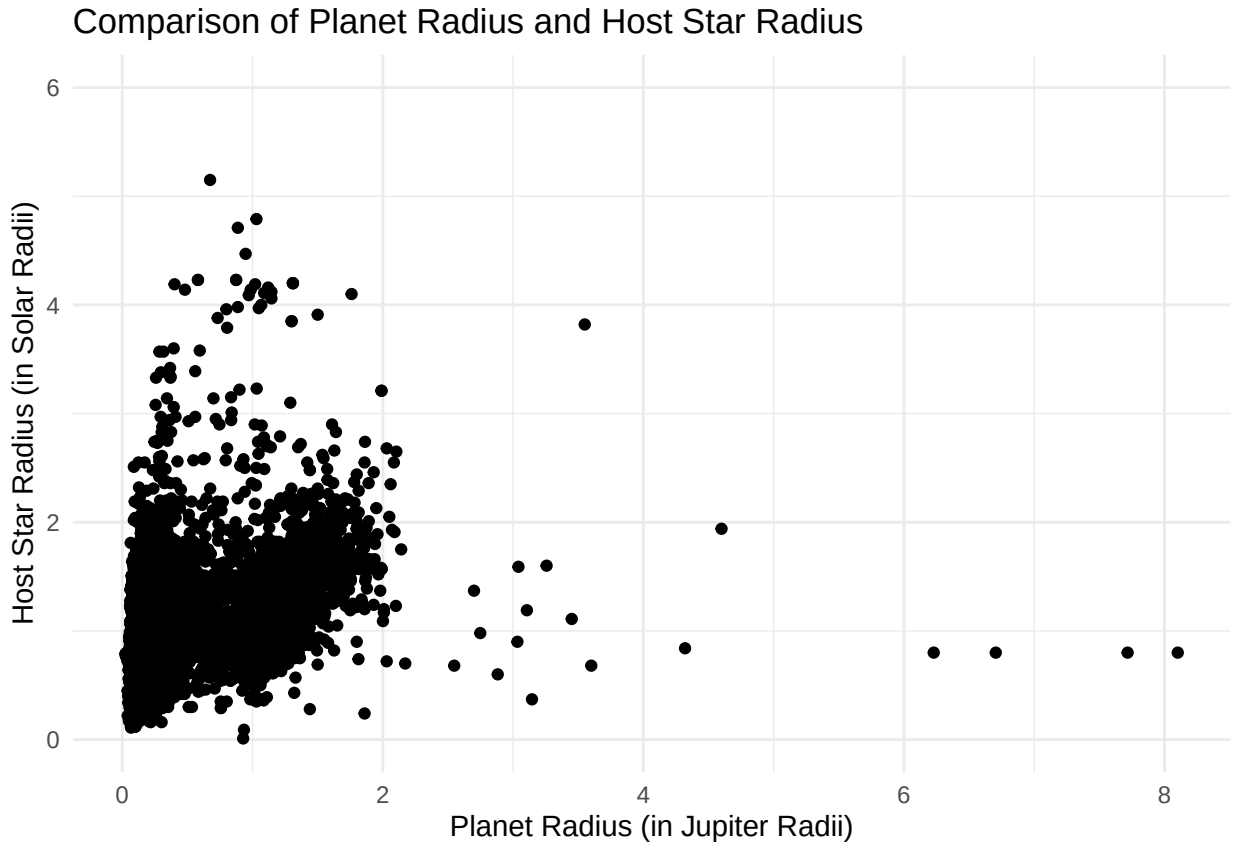
Since 2007, the transit method has become the predominant means of discovering exoplanets, peaking at 12,716 total discoveries in 2016. The years 2014 and 2016 account for 21,759 out of 35,086 total observations, making up 62% of the total exoplanet discoveries from 1992 to 2023. The number of exoplanet discoveries has increased steadily over time, with a particularly sharp increase in 2016, 2014, and somewhat in 2021.

## Exoplanet vs. Host Star Comparisons

### Planet Radius (in Jupiter Radii) vs. Host Star (in Solar Radii)

```
# A scatter plot of planet radius vs its host star radius
create_scatterplot(exoplanets, pl_radj, st_rad, "Comparison of Planet Radius and Host Star Radius", "Pla
```

```
## Warning: Removed 24251 rows containing missing values (`geom_point()`).
```



```r
# Calculating correlation coefficient and linear regression for planet radius and host star radius
calculate_correlation_and_regression(exoplanets, "pl_radj", "st_rad")
```

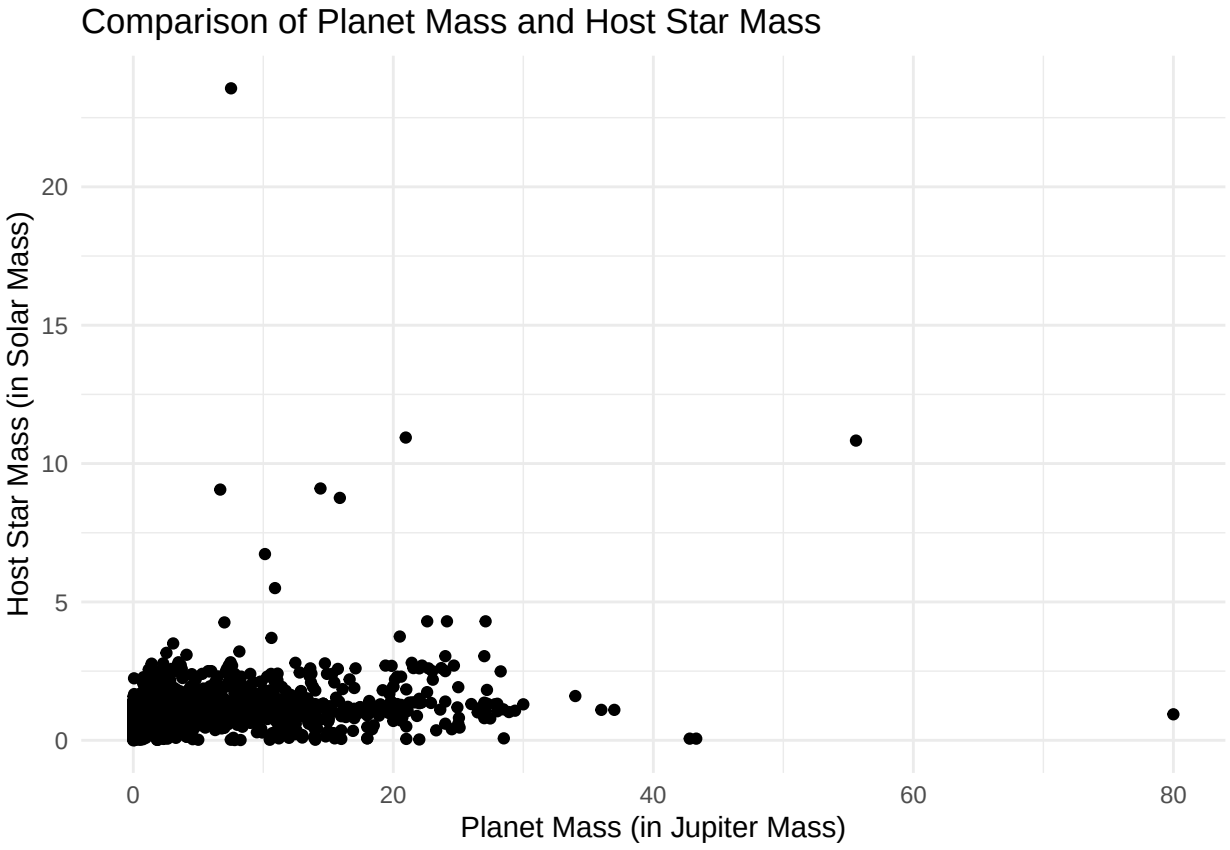
```
## Correlation coefficient: 0.37
##
## Call:
## lm(formula = data[[y_var]] ~ data[[x_var]])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0754 -0.2452 -0.0670  0.1790  5.0295
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.867980   0.005604   154.9   <2e-16 ***
## data[[x_var]] 0.371194   0.008859    41.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.429 on 10840 degrees of freedom
##   (24244 observations deleted due to missingness)
## Multiple R-squared:  0.1394, Adjusted R-squared:  0.1393
## F-statistic:  1756 on 1 and 10840 DF,  p-value: < 2.2e-16
```

```
## $correlation
## [1] 0.3733327
##
## $regression_model
##
## Call:
## lm(formula = data[[y_var]] ~ data[[x_var]])
##
## Coefficients:
##   (Intercept)  data[[x_var]]
##        0.8680         0.3712
```

**Planet Mass (in Jupiter Mass) vs. Host Star Mass (in Solar Mass)**

```r
# A scatter plot of planet mass vs its host star mass
create_scatterplot(exoplanets, pl_bmassj, st_mass, "Comparison of Planet Mass and Host Star Mass", "Pla
```

```
## Warning: Removed 29615 rows containing missing values (`geom_point()`).
```




```
# Calculating correlation coefficient and linear regression for planet mass and host star mass
calculate_correlation_and_regression(exoplanets, "pl_bmassj", "st_mass")
```

```
## Correlation coefficient: 0.27
##
## Call:
## lm(formula = data[[y_var]] ~ data[[x_var]])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8618 -0.2285  0.0121  0.2019 22.3854
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.902059   0.008794   102.6   <2e-16 ***
## data[[x_var]] 0.036247   0.001759    20.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5762 on 5469 degrees of freedom
##   (29615 observations deleted due to missingness)
## Multiple R-squared:  0.07202,    Adjusted R-squared:  0.07185
## F-statistic: 424.5 on 1 and 5469 DF,  p-value: < 2.2e-16

## $correlation
```

```
## [1] 0.2683708
##
## $regression_model
##
## Call:
## lm(formula = data[[y_var]] ~ data[[x_var]])
##
## Coefficients:
##   (Intercept)  data[[x_var]]
##       0.90206        0.03625
```

**Summary**

Correlation coefficient:

- There is a 0.37 correlation coefficient for the relationship between a planet's radius and its host star radius. This indicates a positive correlation, meaning that as one increases, the other tends to increase, but the relationship is not very strong. The value of 0.37 suggests a moderately weak positive linear relationship.
- There is a 0.27 correlation coefficient for the relationship between a planet's mass and its host star mass. This also indicates a positive correlation, but the relationship is weaker than in the previous case. The value of 0.27 suggests a relatively weak positive linear relationship.

In both cases, the correlation coefficients suggest that there is some positive linear relationship between the two variables, but it's not very strong.

Linear Regression Model:

- Model 1: Planet Radius vs Host Star Radius -
  - Significance: Both the intercept and the coefficient for host star radius are highly significant ($p < 0.001$), indicating a strong association with planet radius.
  - R-squared: The multiple R-squared value is 0.1394, suggesting that about 13.94% of the variability in planet radius is explained by host star radius.
- Model 2: Planet Mass vs. Host Star Mass -
  - Significance: Both the intercept and the coefficient for host star mass are highly significant ($p < 0.001$), indicating a strong association with planet mass.
  - R-squared: The multiple R-squared value is 0.07202, suggesting that about 7.20% of the variability in planet mass is explained by host star mass.

In both cases, the R-squared values are relatively low, indicating that only a small portion of the variability is explained. Overall, these models suggest that there are statistically significant positive relationships between the planet properties (radius or mass) and their host star properties (radius or mass). However, the explanatory power of these relationships is limited, as indicated by the relatively low R-squared values. Other factors not included in the models may also influence planet/star properties.

# Analysis/Data Considerations

## Limitations

As initially noted, two data sets pertaining to the Planetary Systems are available:

- Planetary Systems, and
- Planetary Systems Composite Data

For the analyses in this project I have elected to use the Planetary Systems data. The choice between the Planetary Systems and the Planetary Systems Composite Data stems from the way data is organized and presented within the archive.

In the Planetary Systems data set, each row corresponds to a single reference and contains a self-contained set of parameters for each planetary system. For any given reference, the table captures all available data for the planets and their host stars. However, not all references provide a comprehensive set of stellar and planetary parameters. A single row may be missing values for certain parameters, even though those values might be available in other rows.

In contrast, the Planetary Systems Composite Data is designed to address the needs of users who prefer a more complete data set, with only one row per planet (removing duplicate rows). This table aggregates and compiles data from various references, aiming to provide a more comprehensive view of each planet. However, this completeness comes at the cost of potential inconsistencies, as the data may be drawn from multiple sources. Thus, while this table offers a more filled-in set of parameters, it may not maintain the same level of self-consistency found in the Planetary Systems data set.

The choice of data set for analysis ultimately has a minimal impact on the overarching insights drawn from the study. While the specific numerical values, calculations, and resulting graphical representations may vary, the fundamental information remains consistent. As an example, 2016 consistently emerges as the year with the highest number of exoplanet discoveries. The selection of one data set over the other influences quantitative aspects, such as the exact count of discoveries, but it does not alter the qualitative patterns or overarching trends in the data.