ExoPlanetAnalysis

Jasmine Kobayashi & Uday Gadge

# NASA Exoplanet Archive

## Introduction

Exo planets are planets that are outside our solar system. The existence of such planets was formulated in 1917, but they were first discovered in 1992. As of December 2022, 5284 exo planets have been confirmed. There is a bias that the discovery methods favor planets that are near the star.

## Data

We got the data from NASA Exo planet archives. The data is a comprehensive list of the exo planets discovered, the time and method of discovery along with the features associated with the planets like mass, radius, orbit length, orbital period etc.,

[Link for the data](https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PSCompPars)

### Getting the data.

The following loaded dataset was first manually downloaded as a csv file from the previously mentioned NASA planet archives. (Most recent download date: December 7th, 2022). From what we’ve found, most dataset archives for the space sciences or NASA typically provide data to be accessed this way, and we were unable to webscrape the data directly from the link. So, to comply with the project’s reproducibility criteria, the data is loaded from the GitHub repo for this project.

url <- "" # place holder for GitHub data extraction  
file\_name <- "PSCompPars\_2022.12.07\_15.29.01.csv" # remove once GitHub link is added  
planet\_data <- read\_csv(file\_name,skip=88,show\_col\_types = FALSE)  
head(planet\_data)

# A tibble: 6 × 85  
 loc\_rowid pl\_name hostname sy\_snum sy\_pnum discov…¹ disc\_…² disc\_…³ pl\_co…⁴  
 <dbl> <chr> <chr> <dbl> <dbl> <chr> <dbl> <chr> <dbl>  
1 1 11 Com b 11 Com 2 1 Radial … 2007 Xinglo… 0  
2 2 11 UMi b 11 UMi 1 1 Radial … 2009 Thueri… 0  
3 3 14 And b 14 And 1 1 Radial … 2008 Okayam… 0  
4 4 14 Her b 14 Her 1 2 Radial … 2002 W. M. … 0  
5 5 16 Cyg B b 16 Cyg B 3 1 Radial … 1996 Multip… 0  
6 6 17 Sco b 17 Sco 1 1 Radial … 2020 Lick O… 0  
# … with 76 more variables: pl\_orbper <dbl>, pl\_orbpererr1 <dbl>,  
# pl\_orbpererr2 <dbl>, pl\_orbperlim <dbl>, pl\_orbsmax <dbl>,  
# pl\_orbsmaxerr1 <dbl>, pl\_orbsmaxerr2 <dbl>, pl\_orbsmaxlim <dbl>,  
# pl\_rade <dbl>, pl\_radeerr1 <dbl>, pl\_radeerr2 <dbl>, pl\_radelim <dbl>,  
# pl\_radj <dbl>, pl\_radjerr1 <dbl>, pl\_radjerr2 <dbl>, pl\_radjlim <dbl>,  
# pl\_bmasse <dbl>, pl\_bmasseerr1 <dbl>, pl\_bmasseerr2 <dbl>,  
# pl\_bmasselim <dbl>, pl\_bmassj <dbl>, pl\_bmassjerr1 <dbl>, …

# Dataset dimensions  
dim(planet\_data) # n-rows x n-columns

[1] 5220 85

This particular dataset has 5220 exo planets and 84 properties associated with each planet.

colnames(planet\_data)

[1] "loc\_rowid" "pl\_name" "hostname" "sy\_snum"   
 [5] "sy\_pnum" "discoverymethod" "disc\_year" "disc\_facility"   
 [9] "pl\_controv\_flag" "pl\_orbper" "pl\_orbpererr1" "pl\_orbpererr2"   
[13] "pl\_orbperlim" "pl\_orbsmax" "pl\_orbsmaxerr1" "pl\_orbsmaxerr2"   
[17] "pl\_orbsmaxlim" "pl\_rade" "pl\_radeerr1" "pl\_radeerr2"   
[21] "pl\_radelim" "pl\_radj" "pl\_radjerr1" "pl\_radjerr2"   
[25] "pl\_radjlim" "pl\_bmasse" "pl\_bmasseerr1" "pl\_bmasseerr2"   
[29] "pl\_bmasselim" "pl\_bmassj" "pl\_bmassjerr1" "pl\_bmassjerr2"   
[33] "pl\_bmassjlim" "pl\_bmassprov" "pl\_orbeccen" "pl\_orbeccenerr1"  
[37] "pl\_orbeccenerr2" "pl\_orbeccenlim" "pl\_insol" "pl\_insolerr1"   
[41] "pl\_insolerr2" "pl\_insollim" "pl\_eqt" "pl\_eqterr1"   
[45] "pl\_eqterr2" "pl\_eqtlim" "ttv\_flag" "st\_spectype"   
[49] "st\_teff" "st\_tefferr1" "st\_tefferr2" "st\_tefflim"   
[53] "st\_rad" "st\_raderr1" "st\_raderr2" "st\_radlim"   
[57] "st\_mass" "st\_masserr1" "st\_masserr2" "st\_masslim"   
[61] "st\_met" "st\_meterr1" "st\_meterr2" "st\_metlim"   
[65] "st\_metratio" "st\_logg" "st\_loggerr1" "st\_loggerr2"   
[69] "st\_logglim" "rastr" "ra" "decstr"   
[73] "dec" "sy\_dist" "sy\_disterr1" "sy\_disterr2"   
[77] "sy\_vmag" "sy\_vmagerr1" "sy\_vmagerr2" "sy\_kmag"   
[81] "sy\_kmagerr1" "sy\_kmagerr2" "sy\_gaiamag" "sy\_gaiamagerr1"   
[85] "sy\_gaiamagerr2"

The detailed description of the columns in the data is available [here](https://exoplanetarchive.ipac.caltech.edu/docs/API_PS_columns.html)

### Data cleaning

Selecting columns that are of interest:

* pl\_name - planet name
* hostname - stellar name
* sy\_snum - number of stars in the planetary system
* sy\_pnum - number of planets in the planetary system
* discoverymethod - method through which it was discovered
* disc\_year - year in which it was discovered
* disc\_facility - facility at which it was discovered
* pl\_controv\_flag - If the existence of it is controversial
* pl\_orbper - Estimated period of orbiting (in days)
* pl\_orbsmax - Length of the semi major axis of the orbit
* pl\_bmasse - Best estimated mass of the planet in units of earth’s mass
* pl\_bmassj - Best estimated mass of the planet in units of earth’s jupiter

planet\_data <- planet\_data %>% select(pl\_name, hostname, sy\_snum, sy\_pnum, discoverymethod, disc\_year, disc\_facility, pl\_controv\_flag, pl\_orbper, pl\_orbsmax, pl\_bmasse, pl\_bmassj)  
  
head(planet\_data)

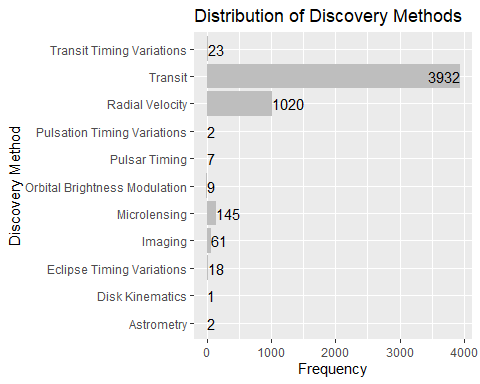
# A tibble: 6 × 12  
 pl\_name hostname sy\_snum sy\_pnum discover…¹ disc\_…² disc\_…³ pl\_co…⁴ pl\_or…⁵  
 <chr> <chr> <dbl> <dbl> <chr> <dbl> <chr> <dbl> <dbl>  
1 11 Com b 11 Com 2 1 Radial Ve… 2007 Xinglo… 0 326.  
2 11 UMi b 11 UMi 1 1 Radial Ve… 2009 Thueri… 0 516.  
3 14 And b 14 And 1 1 Radial Ve… 2008 Okayam… 0 186.  
4 14 Her b 14 Her 1 2 Radial Ve… 2002 W. M. … 0 1765.  
5 16 Cyg B b 16 Cyg B 3 1 Radial Ve… 1996 Multip… 0 798.  
6 17 Sco b 17 Sco 1 1 Radial Ve… 2020 Lick O… 0 578.  
# … with 3 more variables: pl\_orbsmax <dbl>, pl\_bmasse <dbl>, pl\_bmassj <dbl>,  
# and abbreviated variable names ¹​discoverymethod, ²​disc\_year,  
# ³​disc\_facility, ⁴​pl\_controv\_flag, ⁵​pl\_orbper

# % of missing values for each column  
colMeans(is.na(planet\_data))\*100

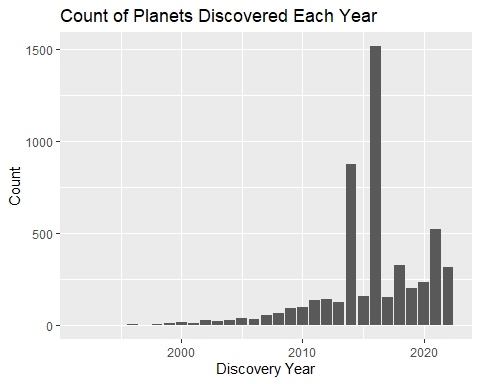
pl\_name hostname sy\_snum sy\_pnum discoverymethod   
 0.000000 0.000000 0.000000 0.000000 0.000000   
 disc\_year disc\_facility pl\_controv\_flag pl\_orbper pl\_orbsmax   
 0.000000 0.000000 0.000000 3.524904 5.287356   
 pl\_bmasse pl\_bmassj   
 0.440613 0.440613

The above chunk displays the percentage of missing values in each column. Seeing that no column is missing more than 6% of their data, we found it reasonable to conclude that there is not a significant amount of missing information (for the selected columns we chose).

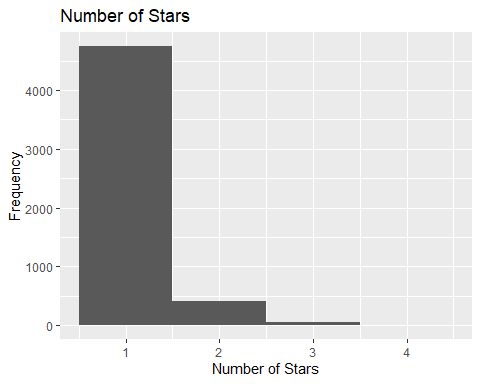
## Exploratory Data Analysis



There is a clear imbalance in the number of planets discovered by different methods. Transit and Radial velocity are the methods through which a lot of planets were discovered.



There’s a clear indication of the increase in discoveries over time, although 2014 and 2016 were particularly successful years.



Most planetary systems are like solar systems with a single star.

# Number of "raised" controversy flags   
sum(planet\_data$pl\_controv\_flag == 1)

[1] 28

The dataset includes information about potential controversy over a planet’s existence, in which the column pl\_controv\_flag consists of binary values (0 = no or 1 = yes) that indicate if there have been published papers that challenge the existence of the planet.

There are only 28 exo planets whose existence are in question so far.

## Hypothesis testing

As mentioned earlier, there is a bias that the discovery methods tend to favor planets near the sun. We can do a hypothesis testing to check if this is significant bias.

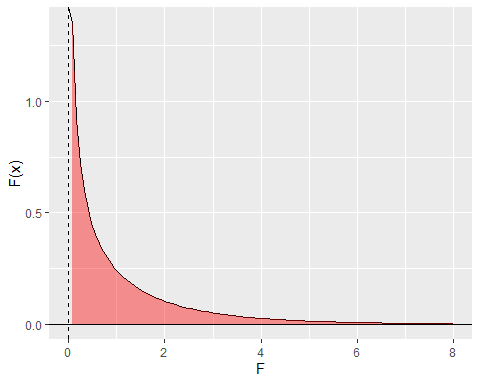
To perform this analysis, we can take our own solar system as a sample (Let us assume that all the planets in our solar system are close to the sun). If there is no bias, the variance of the exo planets orbital semi major axis should be significantly different than the variance of the solar system planets orbital semi major axis.

For the solar system planets orbital semi major axis,we got the data from [here](https://windows2universe.org/our_solar_system/planets_orbits_table.html#:~:text=The%20semimajor%20axis%20the%20average,period%2C%20which%20is%20a%20year.)

solar\_planets <- c("Mercury","Venus","Earth","Mars","Jupiter","Saturn","Uranus","Neptune")  
axis\_lengths <- c(0.3871,0.7233,1.000 ,1.5273 ,5.2028 ,9.5388 ,19.1914 ,30.0611)  
  
solar\_planets <- data.frame(solar\_planets,axis\_lengths)  
colnames(solar\_planets) <- c("pl\_name","pl\_orbsmax")  
solar\_planets$System <- "Solar System"  
  
exo\_planets <- planet\_data %>% select(pl\_name,pl\_orbsmax) %>%   
 mutate(System = "Exo Planets")  
  
planets <- rbind(solar\_planets,exo\_planets) %>% filter(!is.na(pl\_orbsmax))  
  
out <- aov(formula = pl\_orbsmax ~ System, data = planets)  
summary(out)

Df Sum Sq Mean Sq F value Pr(>F)  
System 1 17 17 0.001 0.976  
Residuals 4950 95382874 19269

df1 <- 1  
df2 <- 4938  
F\_cr <- qf(0.05,1,4938)  
f <- 0.001  
  
rejection\_f\_test <- function(x){  
 y <- df(x,df1,df2)  
 y[x<F\_cr] <- NA  
 y  
}  
  
ggplot(data.frame(x = c(0,8)), aes(x)) +  
 stat\_function(fun = function(x){df(x,1,4938)},geom = "line")+  
 stat\_function(fun = function(x){rejection\_f\_test(x)}, geom = "area", fill = "red",alpha = 0.4)+  
 geom\_vline(xintercept = f, linetype = 'dashed') +   
 geom\_hline(yintercept = 0) + xlab('F')+ylab('F(x)')



The obtained F-Value is 0.01 and the corresponding p\_value is 0.976 which is significantly higher than , we fail to reject the null hypothesis.

## Conclusion

The discovery methods are biased towards exo planets that are closer to the star if we define the planets in the solar system to be closer to the sun.

## Session info

R version 4.2.2 (2022-10-31 ucrt)  
Platform: x86\_64-w64-mingw32/x64 (64-bit)  
Running under: Windows 10 x64 (build 19044)  
  
Matrix products: default  
  
locale:  
[1] LC\_COLLATE=English\_United States.utf8   
[2] LC\_CTYPE=English\_United States.utf8   
[3] LC\_MONETARY=English\_United States.utf8  
[4] LC\_NUMERIC=C   
[5] LC\_TIME=English\_United States.utf8   
  
attached base packages:  
[1] stats graphics grDevices utils datasets methods base   
  
other attached packages:  
 [1] rvest\_1.0.3 forcats\_0.5.2 stringr\_1.5.0 dplyr\_1.0.10   
 [5] purrr\_0.3.5 readr\_2.1.3 tidyr\_1.2.1 tibble\_3.1.8   
 [9] ggplot2\_3.4.0 tidyverse\_1.3.2  
  
loaded via a namespace (and not attached):  
 [1] lubridate\_1.9.0 assertthat\_0.2.1 digest\_0.6.30   
 [4] utf8\_1.2.2 R6\_2.5.1 cellranger\_1.1.0   
 [7] backports\_1.4.1 reprex\_2.0.2 evaluate\_0.18   
[10] httr\_1.4.4 pillar\_1.8.1 rlang\_1.0.6   
[13] googlesheets4\_1.0.1 readxl\_1.4.1 rstudioapi\_0.14   
[16] rmarkdown\_2.18 labeling\_0.4.2 googledrive\_2.0.0   
[19] bit\_4.0.5 munsell\_0.5.0 broom\_1.0.1   
[22] compiler\_4.2.2 modelr\_0.1.10 xfun\_0.35   
[25] pkgconfig\_2.0.3 htmltools\_0.5.3 tidyselect\_1.2.0   
[28] fansi\_1.0.3 crayon\_1.5.2 tzdb\_0.3.0   
[31] dbplyr\_2.2.1 withr\_2.5.0 grid\_4.2.2   
[34] jsonlite\_1.8.3 gtable\_0.3.1 lifecycle\_1.0.3   
[37] DBI\_1.1.3 magrittr\_2.0.3 scales\_1.2.1   
[40] cli\_3.4.1 stringi\_1.7.8 vroom\_1.6.0   
[43] farver\_2.1.1 fs\_1.5.2 xml2\_1.3.3   
[46] ellipsis\_0.3.2 generics\_0.1.3 vctrs\_0.5.1   
[49] tools\_4.2.2 bit64\_4.0.5 glue\_1.6.2   
[52] hms\_1.1.2 parallel\_4.2.2 fastmap\_1.1.0   
[55] yaml\_2.3.6 timechange\_0.1.1 colorspace\_2.0-3   
[58] gargle\_1.2.1 knitr\_1.41 haven\_2.5.1