**AEDA**

Final Project

Raúl García Gómez

Marc Monclús

Table of contents

[Report 3](#_Toc99317521)

[Dataset origin 3](#_Toc99317522)

[Exploratory Data Analysis 3](#_Toc99317523)

[k-median and Partitioning Around Medoids (PAM) algorithms: 4](#_Toc99317524)

[PLOTS & SUMMARY 5](#_Toc99317525)

[R Script 10](#_Toc99317526)

# Report

For the realization of this final project of data analysis, a database from NASA of discovered planets has been used with enough data in columns to meet the specifications of having:

* ID
* Name of the planet
* One binary categorical variable:
  + Controversial Flag (originally an integer class converted to a Boolean)
* 2 polyatomic categorical variables (or more):
  + Discovery method
  + Discovery Year
  + Discovery Facility
  + Spectral Type
* Seven numerical variables.
  + Number of Stars
  + Number of Planets
  + Orbital Period [days]
  + Planet Radius [Earth Radius]
  + Planet Mass or Mass\*sin(i) [Earth Mass]
  + Eccentricity
  + Equilibrium Temperature [K]
  + Distance (parsec)
  + Orbit Semi-Major Axis

## Dataset origin

The database has been obtained from the NASA Exoplanet Archive in the following link:

<https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS>

First of all, the database that contains the link is made up of a large number of columns that we consider prior to downloading in .csv format that are not necessary and it is decided to dispense with them even before downloading the data set itself. Is attached to this project.

Next, so that RStudio can correctly read the data set, the comments are extracted from the file by opening it in notepad and thus saving the csv as a conventional csv (the #comments that correspond to a legend about what each column is are attached in a txt in the present work to understand better the data used).

When importing the database in RStudio, the data that has been considered excessive for the analysis is finished, but. Even so, some more data has been left than is necessarily required for this project due to pure added interest, such as the semi-major axis or some of the variables such as the distance in parsec and the spectral facility.

# Exploratory Data Analysis

To treat the data before doing the analysis, as usual, we dispense with the records in the database that contain Not Available (NA) in any of the variables.

On the other hand, it is important to eliminate the outliers to facilitate the observation of the results without having values that are excessively distant from the group that makes up the records of the same variable.

# k-median and Partitioning Around Medoids (PAM) algorithms:

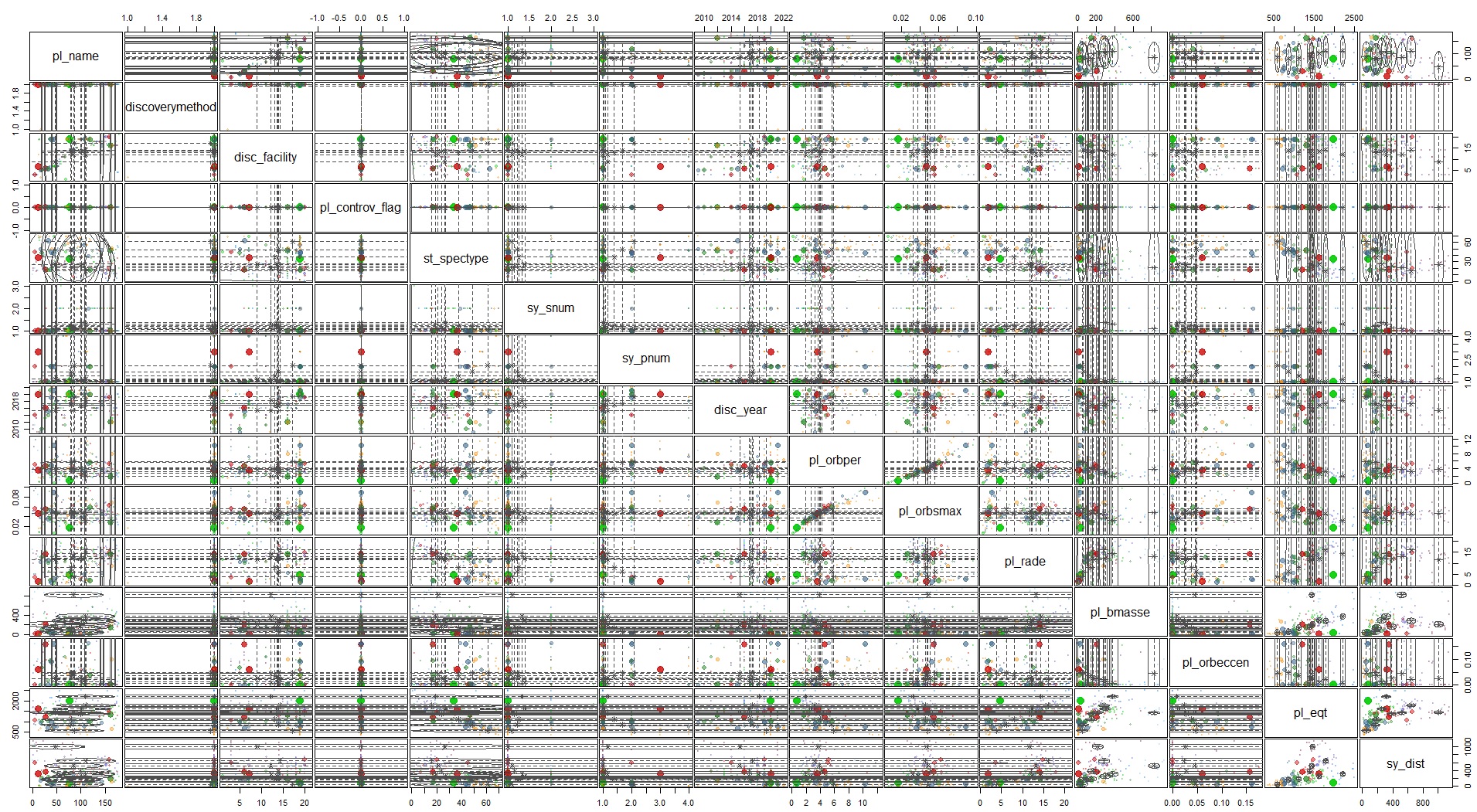
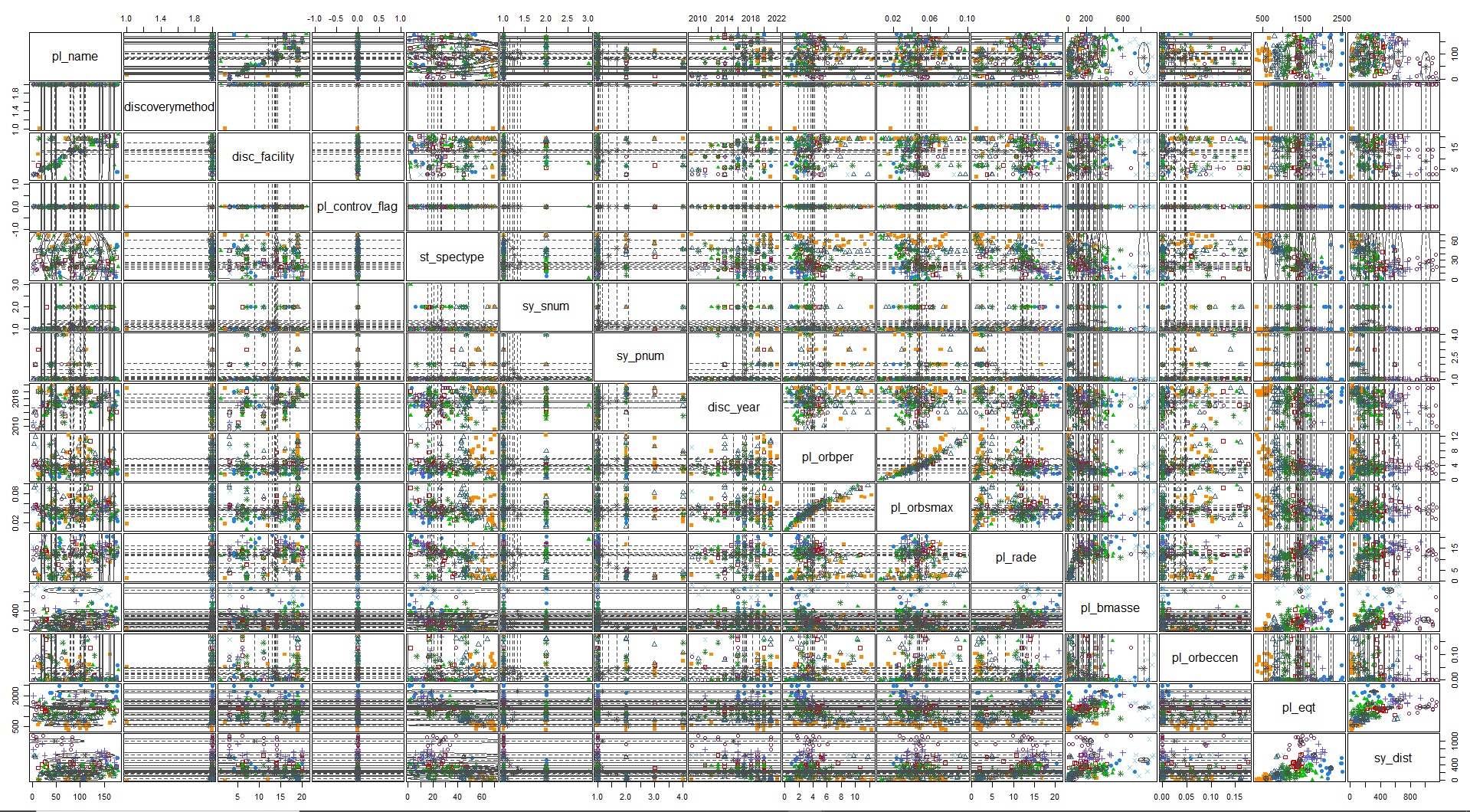
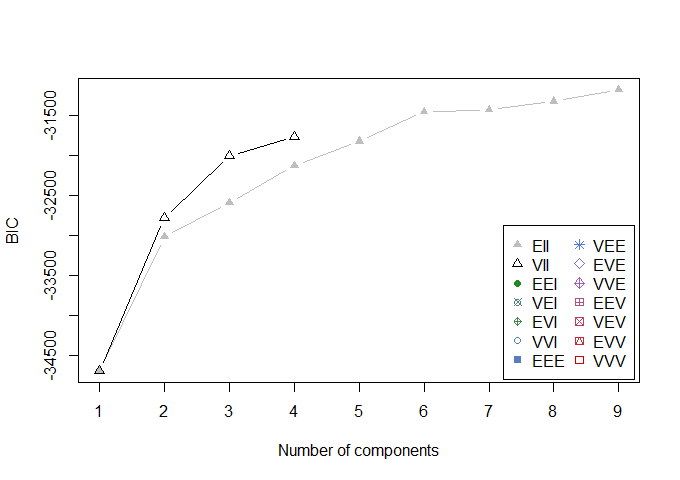
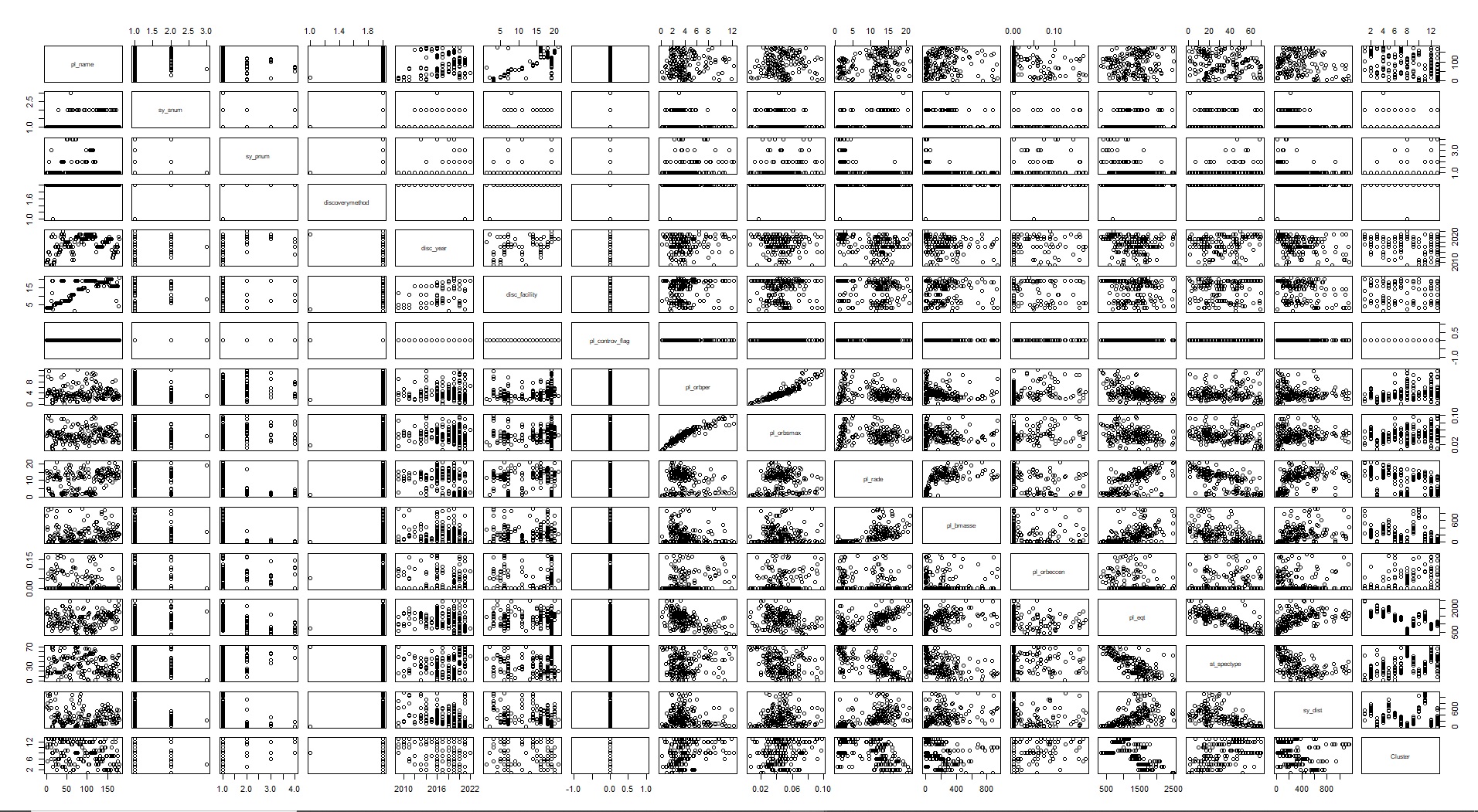
The choice of using the k-median and partitional around medoids (PAM) algorithms is due to the sensitivity of the k-medians with the outliers and to the random selection of centroids.

# PLOTS & SUMMARY

#Line 198

#summary(pl\_clean\_df\_2\_scaled)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| pl\_name | sy\_snum | sy\_pnum | discoverymethod | disc\_year | disc\_facility | pl\_controv\_flag | pl\_orbper | pl\_orbsmax | pl\_rade | pl\_bmasse | pl\_orbeccen | pl\_eqt | st\_spectype | sy\_dist |
| Length:179 | Min. :1.000 | Min. :1.00 | Length:179 | Min. :2009 | Length:179 | Mode :logical | Min. : 0.1797 | Min. :0.00580 | Min. : 0.718 | Min. : 0.4 | Min. :0.00000 | Min. : 353 | Length:179 | Min. : 8.074 |
| Class :character | 1st Qu.:1.000 | 1st Qu.:1.00 | Class :character | 1st Qu.:2016 | Class :character | FALSE:179 | 1st Qu.: 2.4218 | 1st Qu.:0.03424 | 1st Qu.: 4.151 | 1st Qu.: 26.0 | 1st Qu.:0.00000 | 1st Qu.: 957 | Class :character | 1st Qu.: 115.698 |
| Mode :character | Median :1.000 | Median :1.00 | Mode :character | Median :2018 | Mode :character | NA | Median : 3.4655 | Median :0.04500 | Median :11.904 | Median :163.4 | Median :0.00000 | Median :1370 | Mode :character | Median : 271.103 |
| NA | Mean :1.179 | Mean :1.33 | NA | Mean :2017 | NA | NA | Mean : 4.0537 | Mean :0.04649 | Mean :10.269 | Mean :215.9 | Mean :0.03091 | Mean :1306 | NA | Mean : 338.286 |
| NA | 3rd Qu.:1.000 | 3rd Qu.:1.00 | NA | 3rd Qu.:2020 | NA | NA | 3rd Qu.: 4.8530 | 3rd Qu.:0.05600 | 3rd Qu.:14.325 | 3rd Qu.:305.8 | 3rd Qu.:0.05850 | 3rd Qu.:1613 | NA | 3rd Qu.: 512.714 |
| NA | Max. :3.000 | Max. :4.00 | NA | Max. :2022 | NA | NA | Max. :12.2519 | Max. :0.09800 | Max. :21.017 | Max. :940.7 | Max. :0.17600 | Max. :2492 | NA | Max. :1148.930 |



# R Script

## Advanced Engineering Data Analysis FINAL PROJECT

## PLANETARY SYSTEMS DATA ANALYSIS

#Authors:

# Marc Monclús Montalvez Master's Degree in Space and Aeronautical Engineering

# Raúl García Gómez Master's Degree in Industrial Engineering

#Professor:

# Daniel Fernandez Martinez

################################################################################

##### LEGEND

# This file was produced by the NASA Exoplanet Archive http://exoplanetarchive.ipac.caltech.edu

# Fri Mar 25 12:15:35 2022

#

# See also: https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS

#

# User preference: \*

#

# CONSTRAINT: where (default\_flag = 1)

# CONSTRAINT: order by pl\_rade desc

#

# COLUMN pl\_name: Planet Name

# COLUMN sy\_snum: Number of Stars

# COLUMN sy\_pnum: Number of Planets

# COLUMN discoverymethod: Discovery Method

# COLUMN disc\_year: Discovery Year

# COLUMN disc\_facility: Discovery Facility

# COLUMN pl\_controv\_flag: Controversial Flag

# COLUMN pl\_orbper: Orbital Period [days]

# COLUMN pl\_orbsmax: Orbit Semi-Major Axis [au])

# COLUMN pl\_rade: Planet Radius [Earth Radius]

# COLUMN pl\_bmasse: Planet Mass or Mass\*sin(i) [Earth Mass]

# COLUMN pl\_orbeccen: Eccentricity

# COLUMN pl\_eqt: Equilibrium Temperature [K]

# COLUMN st\_spectype: Spectral Type

# COLUMN sy\_dist: Distance [parsec]

################################################################################

## Cleaning Data

rm(list = ls())

#Library - Packages import

library(tidyr)

library(dplyr)

library(cluster)

library(ggplot2)

library(factoextra)

library(NbClust)

library(mclust)

## Work Directory (change if needed)

getwd()

setwd("D:/ESEIAAT/Data\_Analysis/Final\_Project")

plantetary\_df <- read.csv("PS\_2022.03.25\_12.15.35.csv")

# Here we only want specific data, therefore we need to remove the excess.

data\_del <- c("default\_flag","pl\_radj", "pl\_bmassj", "pl\_bmassprov", "pl\_insol",

"ttv\_flag", "pl\_insol", "st\_teff",

"st\_rad", "st\_mass", "st\_met",

"st\_logg", "sy\_vmag", "sy\_kmag", "sy\_gaiamag")

planetary\_short\_df <- plantetary\_df[ , !(names(plantetary\_df)%in%data\_del)]

## here we delete rows with empty or NA values (tidyr library)

drop\_na\_df <- planetary\_short\_df

drop\_na\_df[drop\_na\_df == ""] <- NA

drop\_na\_df <- drop\_na\_df %>% drop\_na(pl\_name)

drop\_na\_df <- drop\_na\_df %>% drop\_na(sy\_snum)

drop\_na\_df <- drop\_na\_df %>% drop\_na(sy\_pnum)

drop\_na\_df <- drop\_na\_df %>% drop\_na(discoverymethod)

drop\_na\_df <- drop\_na\_df %>% drop\_na(disc\_year)

drop\_na\_df <- drop\_na\_df %>% drop\_na(disc\_facility)

drop\_na\_df <- drop\_na\_df %>% drop\_na(pl\_controv\_flag)

drop\_na\_df <- drop\_na\_df %>% drop\_na(pl\_orbper)

drop\_na\_df <- drop\_na\_df %>% drop\_na(pl\_orbsmax)

drop\_na\_df <- drop\_na\_df %>% drop\_na(pl\_rade)

drop\_na\_df <- drop\_na\_df %>% drop\_na(pl\_bmasse)

drop\_na\_df <- drop\_na\_df %>% drop\_na(pl\_orbeccen)

drop\_na\_df <- drop\_na\_df %>% drop\_na(pl\_eqt)

drop\_na\_df <- drop\_na\_df %>% drop\_na(st\_spectype)

drop\_na\_df <- drop\_na\_df %>% drop\_na(sy\_dist)

###Other methods that can be used:

#planetary\_short2\_df<-planetary\_short1\_df[complete.cases(planetary\_short1\_df$pl\_orbper),]

#planetary\_short\_df[!is.na(planetary\_short\_df$pl\_orbper)]

#after removing NA's we get a dataframe of 466 rows from 5005 observations

pl\_clean\_df<-drop\_na\_df

################################################################################

#class of "plantetary\_df$default\_flag"

class(pl\_clean\_df$pl\_controv\_flag)

################################################################################

#we need to convert this column to a binary

pl\_clean\_df$pl\_controv\_flag = as.logical(pl\_clean\_df$pl\_controv\_flag)

pl\_clean\_df\_2<-pl\_clean\_df

#Remove outliers

#Some are not checked because the outliers are interesting (like the Controversial flag)

#or because thy cannot be outliers (like discovery center)

#outlierx2<-boxplot.stats(pl\_clean\_df\_2$sy\_snum)$out

#outlierx2rows<-which(pl\_clean\_df\_2$sy\_snum %in% c(outlierx2))

#outlierx2rows

#pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$sy\_snum %in% outlierx2),]

#outlierx3<-boxplot.stats(pl\_clean\_df\_2$sy\_pnum)$out

#outlierx3rows<-which(pl\_clean\_df\_2$sy\_pnum %in% c(outlierx3))

#outlierx3rows

#pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$sy\_pnum %in% outlierx3),]

#outlierx4<-boxplot.stats(pl\_clean\_df\_2$discoverymethod)$out

#outlierx4rows<-which(pl\_clean\_df\_2$discoverymethod %in% c(outlierx4))

#outlierx4rows

#pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$discoverymethod %in% outlierx4),]

#outlierx5<-boxplot.stats(pl\_clean\_df\_2$disc\_year)$out

#outlierx5rows<-which(pl\_clean\_df\_2$disc\_year %in% c(outlierx5))

#outlierx5rows

#pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$disc\_year %in% outlierx5),]

#outlierx6<-boxplot.stats(pl\_clean\_df\_2$disc\_facility)$out

#outlierx6rows<-which(pl\_clean\_df\_2$disc\_facility %in% c(outlierx6))

#outlierx6rows

#pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$disc\_facility %in% outlierx6),]

#outlierx7<-boxplot.stats(pl\_clean\_df\_2$pl\_controv\_flag)$out

#outlierx7rows<-which(pl\_clean\_df\_2$pl\_controv\_flag %in% c(outlierx7))

#outlierx7rows

#pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$pl\_controv\_flag %in% outlierx7),]

outlierx8<-boxplot.stats(pl\_clean\_df\_2$pl\_orbper)$out

outlierx8rows<-which(pl\_clean\_df\_2$pl\_orbper %in% c(outlierx8))

outlierx8rows

pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$pl\_orbper %in% outlierx8),]

outlierx9<-boxplot.stats(pl\_clean\_df\_2$pl\_orbsmax)$out

outlierx9rows<-which(pl\_clean\_df\_2$pl\_orbsmax %in% c(outlierx9))

outlierx9rows

pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$pl\_orbsmax %in% outlierx9),]

#No outliers in here

outlierx10<-boxplot.stats(pl\_clean\_df\_2$pl\_rade)$out

outlierx10rows<-which(pl\_clean\_df\_2$pl\_rade %in% c(outlierx10))

outlierx10rows

#pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$pl\_rade %in% outlierx10),]

outlierx11<-boxplot.stats(pl\_clean\_df\_2$pl\_bmasse)$out

outlierx11rows<-which(pl\_clean\_df\_2$pl\_bmasse %in% c(outlierx11))

outlierx11rows

pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$pl\_bmasse %in% outlierx11),]

outlierx12<-boxplot.stats(pl\_clean\_df\_2$pl\_orbeccen)$out

outlierx12rows<-which(pl\_clean\_df\_2$pl\_orbeccen %in% c(outlierx12))

outlierx12rows

pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$pl\_orbeccen %in% outlierx12),]

outlierx13<-boxplot.stats(pl\_clean\_df\_2$pl\_eqt)$out

outlierx13rows<-which(pl\_clean\_df\_2$pl\_eqt %in% c(outlierx13))

outlierx13rows

pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$pl\_eqt %in% outlierx13),]

#outlierx14<-boxplot.stats(pl\_clean\_df\_2$st\_spectype)$out

#outlierx14rows<-which(pl\_clean\_df\_2$st\_spectype %in% c(outlierx14))

#outlierx14rows

#pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$st\_spectype %in% outlierx14),]

outlierx15<-boxplot.stats(pl\_clean\_df\_2$sy\_dist)$out

outlierx15rows<-which(pl\_clean\_df\_2$sy\_dist %in% c(outlierx15))

outlierx15rows

pl\_clean\_df\_2<- pl\_clean\_df\_2[-which(pl\_clean\_df\_2$sy\_dist %in% outlierx15),]

#Scaling for clustering

pl\_clean\_df\_2\_scaled <- pl\_clean\_df\_2

pl\_clean\_df\_2\_scaling <- pl\_clean\_df\_2

pl\_clean\_df\_2\_noscaling <- pl\_clean\_df\_2

pl\_clean\_df\_2\_scaling <- pl\_clean\_df\_2[c(2,3,5,8,9,10,11,12,13,15)]

pl\_clean\_df\_2\_noscaling <- pl\_clean\_df\_2[c(1,4,6,7,14)]

pl\_clean\_df\_2\_scaled <- scale(pl\_clean\_df\_2\_scaling)

pl\_clean\_df\_2\_scaled<- cbind.data.frame(pl\_clean\_df\_2\_noscaling, pl\_clean\_df\_2\_scaling)

#PAM looking for 13 clusters because an heuristic says choosing square root of number

#of rows is a good choice

pl\_clean\_df\_2\_scaled\_PAM <- pam(pl\_clean\_df\_2\_scaled, 13)

plot(pl\_clean\_df\_2\_scaled\_PAM, which.plots=2, main="")

plotPAMcluster <- cbind(pl\_clean\_df\_2, Cluster = pl\_clean\_df\_2\_scaled\_PAM$clustering)

plot(plotPAMcluster)

#Model based clustering

pl\_clean\_df\_2\_scaled\_Mcluster <- Mclust(pl\_clean\_df\_2\_scaled)

summary(pl\_clean\_df\_2\_scaled\_Mcluster)

plot(pl\_clean\_df\_2\_scaled\_Mcluster)

# Legend of variables

* pl\_name: Planet Name
* sy\_snum: Number of Stars
* sy\_pnum: Number of Planets
* discoverymethod: Discovery Method
* disc\_year: Discovery Year
* disc\_facility: Discovery Facility
* pl\_controv\_flag: Controversial Flag
* pl\_orbper: Orbital Period [days]
* pl\_orbsmax: Orbit Semi-Major Axis [au])
* pl\_rade: Planet Radius [Earth Radius]
* pl\_bmasse: Planet Mass or Mass\*sin(i) [Earth Mass]
* pl\_orbeccen: Eccentricity
* pl\_eqt: Equilibrium Temperature [K]
* st\_spectype: Spectral Type
* sy\_dist: Distance [parsec]