Regresión Avanzada 2024 - TP Final

Jorge Nicolau

Table of Contents



# Detección de exoplanetas y sus características

Durante casi tres décadas de investigación, la NASA ha confirmado la existencia de más de 5,600 exoplanetas en 4,151 sistemas planetarios. Estos mundos, conocidos como exoplanetas o planetas extrasolares, orbitan estrellas distintas al Sol. Aunque los astrónomos los imaginaron durante milenios, no fue hasta mediados de la década de 1990 cuando comenzaron a aparecer en los registros científicos.

Aproximadamente dos tercios de los exoplanetas descubiertos hasta ahora provienen del telescopio espacial Kepler, mientras que cientos más han sido identificados gracias a la misión TESS y otros observatorios terrestres. Estas misiones han generado una enorme cantidad de datos, con cerca de 10,000 posibles exoplanetas aún pendientes de confirmación. Un número impresionante si consideramos que hasta principios de los años noventa no se conocía ninguno, pero diminuto en comparación con los cientos de miles de millones que podrían existir solo en la Vía Láctea.

Para detectar estos mundos, los astrónomos emplean diversas técnicas, muchas de ellas llevadas al límite de la tecnología disponible en los observatorios espaciales (Nardi, Luca, n.d.).

## Métodos de detección de exoplanetas

### Astrometría: detectando exoplanetas a través del movimiento estelar

Cuando decimos que los planetas orbitan alrededor de las estrellas, en realidad no estamos contando toda la historia. En términos precisos, tanto la estrella como sus planetas giran en torno al **baricentro** del sistema, el punto en el espacio donde se equilibra la masa total del conjunto. Dado que las estrellas poseen la mayor parte de la masa, este punto suele ubicarse muy cerca de ellas. En el caso del Sol, por ejemplo, el baricentro se encuentra apenas fuera de su superficie.

Esto implica que una estrella con planetas no permanece completamente estática en el espacio, sino que también se desplaza ligeramente mientras orbita el baricentro del sistema. Al observar esta leve oscilación, los astrónomos pueden detectar la presencia de exoplanetas mediante el método de la **astrometría**. Gracias a esta técnica, se han identificado más de 1,000 planetas extrasolares (NeoFronteras, n.d.).

### Velocidades radiales: detectando exoplanetas a través del efecto Doppler

La técnica de velocidad radial se basa en el mismo principio que la astrometría: la rotación de la estrella alrededor del baricentro del sistema debido a la influencia gravitatoria de un planeta. Sin embargo, mientras que la astrometría observa directamente el desplazamiento de la estrella en el cielo, la velocidad radial se centra en sus efectos espectroscópicos.

Cuando la estrella se mueve acercándose a nosotros, la frecuencia de su luz se desplaza hacia el azul, y cuando se aleja, se desplaza hacia el rojo. Este fenómeno, conocido como **efecto Doppler**, permite detectar variaciones periódicas en el espectro estelar. Si se observa un patrón rítmico de estos desplazamientos, se puede inferir la presencia de un exoplaneta que provoca el vaivén gravitacional de la estrella (Marín, Daniel, n.d.b).

### Método del tránsito: detectando exoplanetas a través de su sombra

Más de 4,200 exoplanetas han sido descubiertos mediante el **método del tránsito**, una técnica ampliamente utilizada por el telescopio espacial Kepler. Este método se basa en la observación de la disminución periódica del brillo de una estrella cuando un planeta pasa frente a ella, bloqueando parte de su luz, de manera similar a un eclipse solar.

Al registrar estas variaciones de brillo en múltiples ocasiones, se puede confirmar la presencia de un planeta en órbita. Además, este método proporciona información clave sobre el exoplaneta, como su tamaño y distancia respecto a la estrella. Si el planeta posee una atmósfera, la luz estelar la atraviesa durante el tránsito, permitiendo a los astrónomos analizar su composición química a través de técnicas espectroscópicas. Gracias a esto, es posible estudiar la atmósfera de mundos que ni siquiera podemos observar directamente (Alonso Sobrino, Roi, n.d.).

### Microlente gravitacional: exoplanetas detectados a través de la curvatura del espacio-tiempo

El **método de microlente gravitacional** se basa en un principio fundamental de la **Relatividad General**: la presencia de una masa curva el espacio-tiempo y, como consecuencia, desvía la trayectoria de la luz que pasa cerca de ella. Este fenómeno se conoce como **lente gravitacional** y puede usarse para detectar exoplanetas en circunstancias específicas.

Cuando una estrella lejana se alinea con otra estrella más cercana en nuestra línea de visión, la gravedad de la estrella más cercana actúa como una lente natural, magnificando la luz de la estrella de fondo. Este efecto crea un **aumento temporal del brillo** de la estrella lejana, lo que se conoce como un **evento de microlente gravitacional**.  
Si la estrella más cercana alberga un planeta, la gravedad del planeta también contribuirá al efecto de lente, provocando una **segunda variación en el brillo** de la estrella de fondo. Este cambio adicional en la luz nos permite inferir la presencia del planeta y estimar algunas de sus características, como su masa y distancia con respecto a su estrella (Marín, Daniel, n.d.a).

### Observación directa: fotografiar expplanetas en medio del brillo estelar

Detectar exoplanetas mediante **observación directa** es un desafío, ya que estos son cuerpos pequeños y poco luminosos en comparación con las estrellas que orbitan. La intensa luz de la estrella suele eclipsar por completo la presencia del planeta, dificultando su detección.

Sin embargo, existe una técnica para superar este problema: el uso de **coronógrafos**. Estos dispositivos bloquean la luz directa de la estrella, permitiendo que la tenue luz reflejada por el planeta sea visible. El concepto es similar a tapar el Sol con la mano para ver mejor lo que hay en el horizonte.

Para que este método funcione, el planeta debe estar **lo suficientemente alejado de su estrella**, de modo que el coronógrafo pueda aislar su luz. Cuando esto ocurre, es posible capturar imágenes reales del exoplaneta, lo que se conoce como **observación directa**. Gracias a esta técnica, los astrónomos pueden analizar la luz reflejada por el planeta y obtener información sobre su composición, temperatura y atmósfera (NASA, Equipo de redacción de Ciencia, n.d.).

### Detección de exoplanetas por cómo afectan la rotación de púlsares

Los **púlsares** son un tipo especial de **estrellas de neutrones**, formadas tras la explosión de una supernova y comprimidas por la gravedad hasta alcanzar una densidad extrema. Estas estrellas pueden girar a velocidades sorprendentes, llegando a completar **decenas o incluso cientos de vueltas por segundo**.

Una de las características más notables de los púlsares es la **precisión de su rotación**. Por ejemplo, el **púlsar del Cangrejo**, uno de los más estudiados, gira exactamente **33 veces por segundo**. Sin embargo, si un planeta orbita alrededor de un púlsar, su influencia gravitacional puede provocar **pequeñas variaciones en la periodicidad de los pulsos**.

Midiendo con precisión estas sutiles anomalías en el tiempo de rotación, los astrónomos pueden **inferir la presencia de planetas** en sistemas de púlsares, un método que ha permitido descubrir algunos de los primeros exoplanetas conocidos (Wolszczan, Aleksander & Frail, D. A., n.d.).

## Detección de exoplanetas usando la ciencia de datos

La detección de exoplanetas es un campo de investigación en constante evolución, con nuevos descubrimientos y técnicas emergentes que amplían nuestro conocimiento del universo. La ciencia de datos juega un papel fundamental en este proceso, permitiendo a los astrónomos analizar grandes volúmenes de datos y extraer información relevante sobre los exoplanetas y sus características.

Incluso nada anima a la detección amateur de exoplanetas a traves de su sitio Exoplanet Watch donde ofrece una guía detallada sobre cómo analizar observaciones de tránsitos de exoplanetas para generar curvas de luz, que representan las variaciones en el brillo de una estrella cuando un planeta pasa frente a ella (Brachman, R. Z., n.d.).

Fuente: <https://exoplanets.nasa.gov/exoplanet-watch/how-to-contribute/how-to-analyze-your-data/>

NASA ha recopilado una gran cantidad de datos sobre exoplanetas detectados por misiones como Kepler y TESS, proporcionando una fuente invaluable de información para la investigación científica. Estos datos contienen una variedad de características sobre los exoplanetas y sus sistemas estelares, incluyendo su masa, radio, temperatura, órbita y composición química. Sin embargo esta información no siempre es completa y en muchos casos es necesario realizar análisis avanzados para extraer parámetros faltantes y predecir la habitabilidad de los exoplanetas.

El objetivo de este trabajo es procesar los datos para realizar un análisis exploratorio y utilizar regresión linea y logística para la extracción de paramétros faltantes de los exoplanetas y para predecir la habitabilidad de los mismos.

# El dataset de expolanetas detectados con el telescopio espacial Kepler

El **Archivo de Exoplanetas de la NASA** es un catálogo astronómico en línea y un servicio de datos que recopila, correlaciona y organiza información sobre exoplanetas y sus estrellas anfitrionas. Además, proporciona herramientas para trabajar con estos datos.

Este archivo está dedicado a la recopilación y difusión de **conjuntos de datos públicos clave** utilizados en la búsqueda y caracterización de planetas extrasolares y sus estrellas (IPAC, n.d.b). Entre los datos disponibles se incluyen:

* **Parámetros estelares**: posiciones, magnitudes y temperaturas de las estrellas anfitrionas.
* **Parámetros de exoplanetas**: masas, características orbitales y otros datos fundamentales.
* **Datos de descubrimiento y caracterización**: curvas de velocidad radial publicadas, curvas de luz fotométricas, imágenes y espectros.

Esta base de datos es una herramienta esencial para la investigación en el campo de la **astronomía de exoplanetas** y ha sido utilizada en numerosos estudios científicos y proyectos de investigación. Se encuentra disponible en línea y es de libre acceso para la comunidad científica y el público en general en el sitio del IPAC del Caltech. El **IPAC (Infrared Processing and Analysis Center)** es un centro de investigación y procesamiento de datos astronómicos ubicado en el **Instituto de Tecnología de California (Caltech)**. Se especializa en la gestión y análisis de datos de misiones espaciales y telescopios astronómicos, en colaboración con agencias como la **NASA, la Fundación Nacional de Ciencias (NSF) y el Laboratorio de Propulsión a Chorro (JPL)** (IPAC, n.d.c).

Al momento de la extracción (14 de enero de 2025) el dataset de exoplanetas contiene información sobre 28.217 exoplanetas (entre confirmados, propuestos y en estudio) y sus características. La base de datos de IPAC contiene información sobre los exoplanetas detectados por la misión Kepler y otros telescopios, pero para este trabajo se utilizará solo la información de los exoplanetas detectados por la misión Kepler (Cermak, A. & Cermak, A., n.d.).

Se cargarán los datos, se filtrarán las filas de los explonetas confirmados, las columnas relevantes, se eliminarán los valores faltantes y se realizará un análisis exploratorio de los datos para luego realizar la regresión lineal y logística.

## Preprocesamiento de datos

Se cargan los datos del dataset de exoplanetas de la misión Kepler del archivo **keplerexoplanets.csv**. El mismo se obtuvo de la página del IPAC del Caltech (IPAC, n.d.a). La consulta usada para producir el dataset fue la siguiente:

# This file was produced by the NASA Exoplanet Archive http://exoplanetarchive.ipac.caltech.edu  
# Tue Jan 14 15:20:53 2025  
#  
# User preference: \*  
#  
# CONSTRAINT: where (disc\_facility like '%Kepler%')  
#  
# COLUMN pl\_name: Planet Name  
# COLUMN hostname: Host Name  
# COLUMN pl\_letter: Planet Letter  
# COLUMN hd\_name: HD ID  
# COLUMN hip\_name: HIP ID  
# COLUMN tic\_id: TIC ID  
# COLUMN gaia\_id: GAIA ID  
# COLUMN default\_flag: Default Parameter Set  
# COLUMN sy\_snum: Number of Stars  
# COLUMN sy\_pnum: Number of Planets  
# COLUMN sy\_mnum: Number of Moons  
# COLUMN cb\_flag: Circumbinary Flag  
# COLUMN discoverymethod: Discovery Method  
# COLUMN disc\_year: Discovery Year  
# COLUMN disc\_refname: Discovery Reference  
# COLUMN disc\_pubdate: Discovery Publication Date  
# COLUMN disc\_locale: Discovery Locale  
# COLUMN disc\_facility: Discovery Facility  
# COLUMN disc\_telescope: Discovery Telescope  
# COLUMN disc\_instrument: Discovery Instrument  
# COLUMN rv\_flag: Detected by Radial Velocity Variations  
# COLUMN pul\_flag: Detected by Pulsar Timing Variations  
# COLUMN ptv\_flag: Detected by Pulsation Timing Variations  
# COLUMN tran\_flag: Detected by Transits  
# COLUMN ast\_flag: Detected by Astrometric Variations  
# COLUMN obm\_flag: Detected by Orbital Brightness Modulations  
# COLUMN micro\_flag: Detected by Microlensing  
# COLUMN etv\_flag: Detected by Eclipse Timing Variations  
# COLUMN ima\_flag: Detected by Imaging  
# COLUMN dkin\_flag: Detected by Disk Kinematics  
# COLUMN soltype: Solution Type  
# COLUMN pl\_controv\_flag: Controversial Flag  
# COLUMN pl\_refname: Planetary Parameter Reference  
# COLUMN pl\_orbper: Orbital Period [days]  
# COLUMN pl\_orbsmax: Orbit Semi-Major Axis [au]  
# COLUMN pl\_rade: Planet Radius [Earth Radius]  
# COLUMN pl\_radj: Planet Radius [Jupiter Radius]  
# COLUMN pl\_masse: Planet Mass [Earth Mass]  
# COLUMN pl\_massj: Planet Mass [Jupiter Mass]  
# COLUMN pl\_msinie: Planet Mass\*sin(i) [Earth Mass]  
# COLUMN pl\_msinij: Planet Mass\*sin(i) [Jupiter Mass]  
# COLUMN pl\_cmasse: Planet Mass\*sin(i)/sin(i) [Earth Mass]  
# COLUMN pl\_cmassj: Planet Mass\*sin(i)/sin(i) [Jupiter Mass]  
# COLUMN pl\_bmasse: Planet Mass or Mass\*sin(i) [Earth Mass]  
# COLUMN pl\_bmassj: Planet Mass or Mass\*sin(i) [Jupiter Mass]  
# COLUMN pl\_bmassprov: Planet Mass or Mass\*sin(i) Provenance  
# COLUMN pl\_dens: Planet Density [g/cm\*\*3]  
# COLUMN pl\_orbeccen: Eccentricity  
# COLUMN pl\_insol: Insolation Flux [Earth Flux]  
# COLUMN pl\_eqt: Equilibrium Temperature [K]  
# COLUMN pl\_orbincl: Inclination [deg]  
# COLUMN pl\_tranmid: Transit Midpoint [days]  
# COLUMN pl\_tsystemref: Time Reference Frame and Standard  
# COLUMN ttv\_flag: Data show Transit Timing Variations  
# COLUMN pl\_imppar: Impact Parameter  
# COLUMN pl\_trandep: Transit Depth [%]  
# COLUMN pl\_trandur: Transit Duration [hours]  
# COLUMN pl\_ratdor: Ratio of Semi-Major Axis to Stellar Radius  
# COLUMN pl\_ratror: Ratio of Planet to Stellar Radius  
# COLUMN pl\_occdep: Occultation Depth [%]  
# COLUMN pl\_orbtper: Epoch of Periastron [days]  
# COLUMN pl\_orblper: Argument of Periastron [deg]  
# COLUMN pl\_rvamp: Radial Velocity Amplitude [m/s]  
# COLUMN pl\_projobliq: Projected Obliquity [deg]  
# COLUMN pl\_trueobliq: True Obliquity [deg]  
# COLUMN st\_refname: Stellar Parameter Reference  
# COLUMN st\_spectype: Spectral Type  
# COLUMN st\_teff: Stellar Effective Temperature [K]  
# COLUMN st\_rad: Stellar Radius [Solar Radius]  
# COLUMN st\_mass: Stellar Mass [Solar mass]  
# COLUMN st\_met: Stellar Metallicity [dex]  
# COLUMN st\_metratio: Stellar Metallicity Ratio  
# COLUMN st\_lum: Stellar Luminosity [log(Solar)]  
# COLUMN st\_logg: Stellar Surface Gravity [log10(cm/s\*\*2)]  
# COLUMN st\_age: Stellar Age [Gyr]  
# COLUMN st\_dens: Stellar Density [g/cm\*\*3]  
# COLUMN st\_vsin: Stellar Rotational Velocity [km/s]  
# COLUMN st\_rotp: Stellar Rotational Period [days]  
# COLUMN st\_radv: Systemic Radial Velocity [km/s]  
# COLUMN sy\_refname: System Parameter Reference  
# COLUMN rastr: RA [sexagesimal]  
# COLUMN ra: RA [deg]  
# COLUMN decstr: Dec [sexagesimal]  
# COLUMN dec: Dec [deg]  
# COLUMN glat: Galactic Latitude [deg]  
# COLUMN glon: Galactic Longitude [deg]  
# COLUMN elat: Ecliptic Latitude [deg]  
# COLUMN elon: Ecliptic Longitude [deg]  
# COLUMN sy\_pm: Total Proper Motion [mas/yr]  
# COLUMN sy\_pmra: Proper Motion (RA) [mas/yr]  
# COLUMN sy\_pmdec: Proper Motion (Dec) [mas/yr]  
# COLUMN sy\_dist: Distance [pc]  
# COLUMN sy\_plx: Parallax [mas]  
# COLUMN sy\_bmag: B (Johnson) Magnitude  
# COLUMN sy\_vmag: V (Johnson) Magnitude  
# COLUMN sy\_jmag: J (2MASS) Magnitude  
# COLUMN sy\_hmag: H (2MASS) Magnitude  
# COLUMN sy\_kmag: Ks (2MASS) Magnitude  
# COLUMN sy\_umag: u (Sloan) Magnitude  
# COLUMN sy\_gmag: g (Sloan) Magnitude  
# COLUMN sy\_rmag: r (Sloan) Magnitude  
# COLUMN sy\_imag: i (Sloan) Magnitude  
# COLUMN sy\_zmag: z (Sloan) Magnitude  
# COLUMN sy\_w1mag: W1 (WISE) Magnitude  
# COLUMN sy\_w2mag: W2 (WISE) Magnitude  
# COLUMN sy\_w3mag: W3 (WISE) Magnitude  
# COLUMN sy\_w4mag: W4 (WISE) Magnitude  
# COLUMN sy\_gaiamag: Gaia Magnitude  
# COLUMN sy\_icmag: I (Cousins) Magnitude  
# COLUMN sy\_tmag: TESS Magnitude  
# COLUMN sy\_kepmag: Kepler Magnitude  
# COLUMN rowupdate: Date of Last Update  
# COLUMN pl\_pubdate: Planetary Parameter Reference Publication Date  
# COLUMN releasedate: Release Date  
# COLUMN pl\_nnotes: Number of Notes  
# COLUMN st\_nphot: Number of Photometry Time Series  
# COLUMN st\_nrvc: Number of Radial Velocity Time Series  
# COLUMN st\_nspec: Number of Stellar Spectra Measurements  
# COLUMN pl\_nespec: Number of Eclipse Spectra  
# COLUMN pl\_ntranspec: Number of Transmission Spectra  
# COLUMN pl\_ndispec: Number of Direct Imaging Spectra  
#

El encabezado más arriba estaba incluído en el archivo .CSV generado y se incluye aquí para reproducir la consulta. El dataset contiene información sobre los sistemas exoplanetarios detectados por la misión Kepler y el diccionario de datos se encuentra en línea (IPAC, n.d.d) explicando en más profundidad la semántica de las variables.

Todos los archivos necesarios para este trabajo se encuentran en el repositorio de GitHub del autor (Georgsmeinung, n.d.) y se pueden descargar desde allí para reproducir este análisis, incluyendo el dataset **keplerexoplanets.csv** que a continuación se carga.

**Estructura de los Datos**

| Tipo de datos y previsualización |
| --- |
| spc\_tbl\_ [28,217 × 121] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame) |
| $ pl\_name : chr [1:28217] “2MASS J19383260+4603591 b” “2MASS J19383260+4603591 b” “KIC 10001893 b” “KIC 10001893 c” … |
| $ hostname : chr [1:28217] “2MASS J19383260+4603591” “2MASS J19383260+4603591” “KIC 10001893” “KIC 10001893” … |
| $ pl\_letter : chr [1:28217] “b” “b” “b” “c” … |
| $ hd\_name : logi [1:28217] NA NA NA NA NA NA … |
| $ hip\_name : chr [1:28217] NA NA NA NA … |
| $ tic\_id : chr [1:28217] “TIC 271164763” “TIC 271164763” “TIC 158488181” “TIC 158488181” … |
| $ gaia\_id : chr [1:28217] “Gaia DR2 2080063931448749824” “Gaia DR2 2080063931448749824” “Gaia DR2 2130473176626619136” “Gaia DR2 2130473176626619136” … |
| $ default\_flag : num [1:28217] 1 0 1 1 1 1 1 1 0 0 … |
| $ sy\_snum : num [1:28217] 2 2 1 1 1 1 1 1 1 1 … |
| $ sy\_pnum : num [1:28217] 3 3 3 3 3 1 1 1 2 2 … |
| $ sy\_mnum : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |
| $ cb\_flag : num [1:28217] 1 1 0 0 0 0 0 0 0 0 … |
| $ discoverymethod: chr [1:28217] “Eclipse Timing Variations” “Eclipse Timing Variations” “Orbital Brightness Modulation” “Orbital Brightness Modulation” … |
| $ disc\_year : num [1:28217] 2015 2015 2014 2014 2014 … |
| $ disc\_refname : chr [1:28217] “Bara”| **truncated** “Bara”| **truncated** “S”| **truncated** “S”| **truncated** … |
| $ disc\_pubdate : chr [1:28217] “2015-05” “2015-05” “2014-10” “2014-10” … |
| $ disc\_locale : chr [1:28217] “Space” “Space” “Space” “Space” … |
| $ disc\_facility : chr [1:28217] “Kepler” “Kepler” “Kepler” “Kepler” … |
| $ disc\_telescope : chr [1:28217] “0.95 m Kepler Telescope” “0.95 m Kepler Telescope” “0.95 m Kepler Telescope” “0.95 m Kepler Telescope” … |
| $ disc\_instrument: chr [1:28217] “Kepler CCD Array” “Kepler CCD Array” “Kepler CCD Array” “Kepler CCD Array” … |
| $ rv\_flag : num [1:28217] 0 0 0 0 0 1 0 0 0 0 … |
| $ pul\_flag : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |
| $ ptv\_flag : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |
| $ tran\_flag : num [1:28217] 0 0 0 0 0 0 1 1 1 1 … |
| $ ast\_flag : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |
| $ obm\_flag : num [1:28217] 0 0 1 1 1 1 0 0 0 0 … |
| $ micro\_flag : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |
| $ etv\_flag : num [1:28217] 1 1 0 0 0 0 0 0 0 0 … |
| $ ima\_flag : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |
| $ dkin\_flag : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |
| $ soltype : chr [1:28217] “Published Confirmed” “Published Confirmed” “Published Confirmed” “Published Confirmed” … |
| $ pl\_controv\_flag: num [1:28217] 0 0 1 1 1 0 0 0 0 0 … |
| $ pl\_refname : chr [1:28217] “Esme”| **truncated** “Bara”| **truncated** “S”| **truncated** “S”| **truncated** … |
| $ pl\_orbper : num [1:28217] 406 416 0.22 0.325 0.812 … |
| $ pl\_orbsmax : num [1:28217] NA 0.92 NA NA NA … |
| $ pl\_rade : num [1:28217] NA NA NA NA NA NA 5.5 6.9 7.23 6.7 … |
| $ pl\_radj : num [1:28217] NA NA NA NA NA NA 0.491 0.616 0.645 0.598 … |
| $ pl\_masse : num [1:28217] NA 604 NA NA NA … |
| $ pl\_massj : num [1:28217] NA 1.9 NA NA NA NA NA NA NA NA … |
| $ pl\_msinie : num [1:28217] 591 NA NA NA NA … |
| $ pl\_msinij : num [1:28217] 1.86 NA NA NA NA 2 NA NA NA NA … |
| $ pl\_cmasse : logi [1:28217] NA NA NA NA NA NA … |
| $ pl\_cmassj : logi [1:28217] NA NA NA NA NA NA … |
| $ pl\_bmasse : num [1:28217] 591 604 NA NA NA … |
| $ pl\_bmassj : num [1:28217] 1.86 1.9 NA NA NA 2 NA NA NA NA … |
| $ pl\_bmassprov : chr [1:28217] “Msini” “Mass” NA NA … |
| $ pl\_dens : num [1:28217] NA NA NA NA NA NA NA NA NA NA … |
| $ pl\_orbeccen : num [1:28217] 0.33 NA NA NA NA NA NA NA 0 0 … |
| $ pl\_insol : num [1:28217] NA NA NA NA NA NA NA NA 2.14 2.14 … |
| $ pl\_eqt : num [1:28217] NA NA NA NA NA NA NA NA 308 308 … |
| $ pl\_orbincl : num [1:28217] NA NA NA NA NA … |
| $ pl\_tranmid : num [1:28217] NA NA NA NA NA … |
| $ pl\_tsystemref : chr [1:28217] “BJD-TDB” NA NA NA … |
| $ ttv\_flag : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |
| $ pl\_imppar : num [1:28217] NA NA NA NA NA NA NA NA 0.424 0.212 … |
| $ pl\_trandep : num [1:28217] NA NA NA NA NA … |
| $ pl\_trandur : num [1:28217] NA NA NA NA NA … |
| $ pl\_ratdor : num [1:28217] NA NA NA NA NA … |
| $ pl\_ratror : num [1:28217] NA NA NA NA NA … |
| $ pl\_occdep : logi [1:28217] NA NA NA NA NA NA … |
| $ pl\_orbtper : num [1:28217] 2455131 NA NA NA NA … |
| $ pl\_orblper : num [1:28217] 302 NA NA NA NA NA NA NA NA NA … |
| $ pl\_rvamp : num [1:28217] NA NA NA NA NA … |
| $ pl\_projobliq : num [1:28217] NA NA NA NA NA NA NA NA NA NA … |
| $ pl\_trueobliq : logi [1:28217] NA NA NA NA NA NA … |
| $ st\_refname : chr [1:28217] “Esme”| **truncated** “Bara”| **truncated** “S”| **truncated** “S”| **truncated** … |
| $ st\_spectype : chr [1:28217] NA NA NA NA … |
| $ st\_teff : num [1:28217] 29564 NA 27500 27500 27500 … |
| $ st\_rad : num [1:28217] 0.2 NA NA NA NA 1.57 1.01 1.01 1.39 1.39 … |
| $ st\_mass : num [1:28217] 0.48 0.48 NA NA NA 1.22 1.01 0.98 1.03 1.03 … |
| $ st\_met : num [1:28217] NA NA NA NA NA 0.01 -0.04 -0.42 -0.38 -0.38 … |
| $ st\_metratio : chr [1:28217] NA NA NA NA … |
| $ st\_lum : num [1:28217] NA NA NA NA NA NA NA NA NA NA … |
| $ st\_logg : num [1:28217] 5.51 NA 5.35 5.35 5.35 4.09 4.42 4.44 4.16 4.16 … |
| $ st\_age : num [1:28217] NA NA NA NA NA NA NA NA NA NA … |
| $ st\_dens : num [1:28217] NA NA NA NA NA … |
| $ st\_vsin : num [1:28217] NA NA NA NA NA NA NA NA NA NA … |
| $ st\_rotp : num [1:28217] NA NA NA NA NA NA NA NA NA NA … |
| $ st\_radv : num [1:28217] NA NA 25 25 25 NA NA NA NA NA … |
| $ sy\_refname : chr [1:28217] “TICv8” “TICv8” “TICv8” “TICv8” … |
| $ rastr : chr [1:28217] “19h38m32.62s” “19h38m32.62s” “19h09m33.42s” “19h09m33.42s” … |
| $ ra : num [1:28217] 295 295 287 287 287 … |
| $ decstr : chr [1:28217] “+46d03m59.07s” “+46d03m59.07s” “+46d59m04.02s” “+46d59m04.02s” … |
| $ dec : num [1:28217] 46.1 46.1 47 47 47 … |
| $ glat : num [1:28217] 11.7 11.7 16.6 16.6 16.6 … |
| $ glon : num [1:28217] 79 79 77.7 77.7 77.7 … |
| $ elat : num [1:28217] 65.7 65.7 68.4 68.4 68.4 … |
| $ elon : num [1:28217] 315 315 304 304 304 … |
| $ sy\_pm : num [1:28217] 6.97 6.97 6.01 6.01 6.01 … |
| $ sy\_pmra : num [1:28217] 5.2 5.2 1.32 1.32 1.32 … |
| $ sy\_pmdec : num [1:28217] -4.64 -4.64 -5.86 -5.86 -5.86 … |
| $ sy\_dist : num [1:28217] 396 396 1673 1673 1673 … |
| $ sy\_plx : num [1:28217] 2.495 2.495 0.569 0.569 0.569 … |
| $ sy\_bmag : num [1:28217] 12.1 12.1 15.1 15.1 15.1 … |
| $ sy\_vmag : num [1:28217] 12.7 12.7 15.8 15.8 15.8 … |
| $ sy\_jmag : num [1:28217] 12.8 12.8 16.4 16.4 16.4 … |
| $ sy\_hmag : num [1:28217] 12.9 12.9 15.7 15.7 15.7 … |
| $ sy\_kmag : num [1:28217] 13 13 16.6 16.6 16.6 … |
| $ sy\_umag : num [1:28217] NA NA NA NA NA NA NA NA NA NA … |
| $ sy\_gmag : num [1:28217] NA NA NA NA NA NA NA NA NA NA … |
| $ sy\_rmag : num [1:28217] NA NA NA NA NA NA NA NA NA NA … |
| $ sy\_imag : num [1:28217] NA NA NA NA NA NA NA NA NA NA … |
| $ sy\_zmag : num [1:28217] NA NA NA NA NA NA NA NA NA NA … |
| $ sy\_w1mag : num [1:28217] 12.9 12.9 NA NA NA … |
| $ sy\_w2mag : num [1:28217] 12.9 12.9 NA NA NA … |
| $ sy\_w3mag : num [1:28217] 12.7 12.7 NA NA NA … |
| $ sy\_w4mag : num [1:28217] 9.2 9.2 NA NA NA … |
| $ sy\_gaiamag : num [1:28217] 12.1 12.1 15.7 15.7 15.7 … |
| $ sy\_icmag : logi [1:28217] NA NA NA NA NA NA … |
| $ sy\_tmag : num [1:28217] 12.5 12.5 16 16 16 … |
| $ sy\_kepmag : num [1:28217] 12.3 12.3 15.8 15.8 15.8 … |
| $ rowupdate : Date[1:28217], format: “2022-04-19” “2015-06-04” … |
| $ pl\_pubdate : chr [1:28217] “2022-04” “2015-05” “2014-10” “2014-10” … |
| $ releasedate : Date[1:28217], format: “2022-04-19” “2015-06-04” … |
| $ pl\_nnotes : num [1:28217] 1 1 1 1 1 1 2 1 2 2 … |
| $ st\_nphot : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |
| $ st\_nrvc : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |
| $ st\_nspec : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |
| $ pl\_nespec : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |
| $ pl\_ntranspec : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |
| $ pl\_ndispec : num [1:28217] 0 0 0 0 0 0 0 0 0 0 … |

Para el análisis se considerarán solo los exoplanetas confirmados y se eliminarán las columnas no relevantes para el análisis.

Para reducir la dimensionalidad del problema, a continuación se eliminan las columnas con referencias a sitios web no relevantes para el análisis. También se eliminan los datos de referencia de los planetas, estrellas y sistemas no relevantes para el análisis tales como referencias a catálogos externos de estrellas y sistemas, información de publicación del descubrimiento, información de detección, información de fotometría, banderas de organización interna del dataset, información del sistema planetario y detalles técnicos de la detección.

Para reducir la cantidad de registros del dataset y como se busca predecir la habitabilidad de los exoplanetas en sistemas similares al Solar, se filtran los sistemas con estrellas de tamaño similar al Sol (0.9 a 1.1 radios solares). Hay muchas entradas con valores faltantes en las columnas de tipo espectral de las estrellas (la mayoría) que sería un mejor indicador para determinarl la zona habitable, por eso se toma como criterio de filtro el tamaño de las estrellas. Además se consideraran para la muesta solo los exoplanetas publicados en la literatura científica, por lo que se filtra por “Published Confirmed” en la columna **soltype**.

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

# PARTE 1 - FILTRO DE DATOS  
# Seleccionar los sistemas con estrellas de tamaño  
# similar al Sol (0.9 a 1.1 radios solares)  
kepler\_data <- filter(kepler\_data, st\_rad >= 0.9 & st\_rad <= 1.1)  
  
# Seleccionar solo los exoplanetas publicados en la literatura científica  
kepler\_data <- filter(kepler\_data, soltype == "Published Confirmed")  
  
# PARTE 2 - REDUCCION DE LA DIMENSIONALIDAD  
# Eliminar columnas de referencias a sitios web  
kepler\_data <- subset(kepler\_data, select = c(-disc\_refname, -pl\_refname, -st\_refname, -sy\_refname))  
  
# Eliminar columnas de id de catalogos externos de estrellas y sistemas  
kepler\_data <- subset(kepler\_data, select = c(-tic\_id, -gaia\_id))  
  
# Eliminar información referente a la publicación del descubrimiento  
# así como la información del instrumento utilizado, para todos es Kepler  
kepler\_data <- subset(kepler\_data, select = c(-disc\_year, -disc\_pubdate, -disc\_locale, -disc\_facility, -disc\_telescope, -disc\_instrument, -rowupdate, -pl\_pubdate, -releasedate))  
  
# Eliminar columnas con información referente a la detección  
kepler\_data <- subset(kepler\_data, select = c(-rv\_flag, -pul\_flag, -ptv\_flag, -tran\_flag, -ast\_flag, -obm\_flag, -micro\_flag, -etv\_flag, -ima\_flag, -dkin\_flag))  
  
# Eliminar columnas con información de fotometria  
kepler\_data <- subset(kepler\_data, select = c(-sy\_bmag, -sy\_vmag, -sy\_jmag, -sy\_hmag, -sy\_kmag, -sy\_umag, -sy\_gmag, -sy\_rmag, -sy\_imag, -sy\_zmag, -sy\_w1mag, -sy\_w2mag, -sy\_w3mag, -sy\_w4mag, -sy\_gaiamag, -sy\_icmag, -sy\_tmag, -sy\_kepmag))  
  
# Eliminar banderas de organización interna del dataset  
kepler\_data <- subset(kepler\_data, select = c(-pl\_nnotes, -pl\_controv\_flag, -default\_flag))  
  
# Eliminar columnas con información del sistema planetario  
# y detalles técnicos de la detección  
kepler\_data <- subset(kepler\_data, select = c(-sy\_mnum, -pl\_ndispec))  
  
# Eliminar columnas con información adicional de fotometria  
kepler\_data <- subset( kepler\_data, select = c(-st\_nphot, -st\_nrvc, -st\_nspec, -pl\_nespec, -pl\_ntranspec))

El dataset resultante se guarda en un archivo CSV llamado **keplerfiltered.csv**. El dataset resultante contiene 2547 registros y 68 columnas.

Las columnas en el dataset filtrado, y su descripción extraída del diccionario de datos (IPAC, n.d.d), son las siguientes:

| Variable | Descripción |
| --- | --- |
| cb\_flag | Flag de circumbinario (el planeta orbita un sistema binario) |
| dec | Declinación de la estrella (en grados decimales) |
| decstr | Declinación de la estrella (en grados sexagesimales) |
| discoverymethod | Método de descubrimiento del planeta |
| elat | Latitud eclíptica de la estrella (en grados decimales) |
| elon | Longitud eclíptica de la estrella (en grados decimales) |
| glat | Latitud galáctica de la estrella (en grados decimales) |
| glon | Longitud galáctica de la estrella (en grados decimales) |
| hd\_name | Nombre del sistema estelar en el catálogo Henry Draper |
| hip\_name | Nombre del sistema estelar en el catálogo Hipparcos |
| hostname | Nombre del sistema estelar utilizado en la literatura |
| pl\_bmasse | Mejor estimación de la masa del planeta (en masas terrestres) |
| pl\_bmassj | Mejor estimación de la masa del planeta (en masas jovianas) |
| pl\_bmassprov | Proveedor de la mejor estimación de la masa del planeta (mínima o proyectada) |
| pl\_cmasse | Masa mínima proyectada (en masas terrestres) |
| pl\_cmassj | Masa mínima proyectada (en masas jovianas) |
| pl\_dens | Densidad del planeta (en g/cm³) |
| pl\_eqt | Temperatura de equilibrio del planeta considero como un cuerpo oscur (en grados K) |
| pl\_imppar | Parámetro de impacto del planeta como distancia proyectada del centro del planeta al centro de la estrella (en radios estelares) |
| pl\_insol | Insolación del planeta en terminos la insolación de la tierra (en W/m²) |
| pl\_letter | Letra del planeta en el sistema (primer planeta es b) |
| pl\_masse | Masa del planeta (en masas terrestres) |
| pl\_massj | Masa del planeta (en masas jovianas) |
| pl\_msinie | Mínimo de masa medida por el método de velocidad radial (en masas terrestres) |
| pl\_msinij | Mínimo de masa medida por el método de velocidad radial (en masas jovianas) |
| pl\_name | Profundidad del eclipse del planeta en terminos de el flujo relativo decreciente frente a su estrella (en %) |
| pl\_occdep | Excentricidad orbital del planeta |
| pl\_orbeccen | Inclinación orbital del planeta respecto de la línea de visión desde la Tierra (en grados) |
| pl\_orbincl | El argumento del periastro del planeta (en grados) |
| pl\_orblper | Periodo orbital del planeta (en días) |
| pl\_orbper | Distancia orbital semieje mayor (en UA) |
| pl\_orbsmax | El momento del paso del planeta por el periastro (en grados) |
| pl\_orbtper | Inclinación proyectada del planeta respecto de la línea de visión desde la Tierra (en grados) |
| pl\_projobliq | Radio del planeta (en radios terrestres) |
| pl\_rade | Radio del planeta (en radios jovianos) |
| pl\_radj | cociente del semieje orbital mayor sobre el radio estelar |
| pl\_ratdor | cociente del radio del planeta sobre el radio estelar |
| pl\_ratror | Amplitud de la velocidad radial del planeta (en m/s) |
| pl\_rvamp | Profundidad del tránsito del planeta en terminos de el flujo relativo decreciente frente a su estrella (en %) |
| pl\_trandep | Duración del tránsito del planeta (en horas) |
| pl\_trandur | Tiempo de tránsito medio del planeta respecto de su estrella (en días) |
| pl\_tranmid | Inclinación verdadera del planeta respecto de la línea de visión desde la Tierra (en grados) |
| pl\_trueobliq | Sistema de Tiempo de Referencia del sistema planetario |
| pl\_tsystemref | Ascensión recta de la estrella (en grados decimales) |
| ra | Ascensión recta de la estrella (en grados sexagesimales) |
| rastr | Disposición del planeta según el conjunto de parámetros planetarios dado |
| soltype | Edad de la estrella (en Ga, gigaaños, o miles de millones de años) |
| st\_age | Densidad de la estrella (en g/cm³) |
| st\_dens | Gravedad superficial de la estrella (en logaritmo en base 10 de cm/s²) |
| st\_logg | Luminosidad de la estrella (en logartima en base 10 de unidades solares) |
| st\_lum | Masa de la estrella (en masas solares) |
| st\_mass | Medición del contenido de metales en la fotosfera de la estrella en comparación con el contenido de hidrógeno |
| st\_met | Relación de metalicidad de la estrella |
| st\_metratio | Radio de la estrella (en radios solares) |
| st\_rad | Velocidad radial de la estrella (en km/s) |
| st\_radv | Periodo de rotación de la estrella (en días) |
| st\_rotp | Tipo espectral de la estrella según el sistema Morgan-Keenan |
| st\_spectype | Temperatura efectiva de la estrella (en grados K) |
| st\_teff | Velocidad de rotación de la estrella (en km/s) |
| st\_vsin | Distancia del sistema (en parsecs) |
| sy\_dist | Paralaje del sistema (en miliarcosegundos) |
| sy\_plx | Movimiento propio del sistema (en miliarcosegundos/año) |
| sy\_pm | Movimiento propio del sistema en declinación (en miliarcosegundos/año) |
| sy\_pmdec | Movimiento propio del sistema en ascensión recta (en miliarcosegundos/año) |
| sy\_pmra | Número de planetas en el sistema |
| sy\_pnum | Número de estrellas en el sistema |
| sy\_snum | Flag de tránsito de tiempo variante |
| ttv\_flag | Nombre del planeta |

## Análisis exploratorio de datos

Para realizar un análisis exploratorio de los datos, se cargan el nuevo dataset (puede comenzarse el análisis en este punto porque el dataset preprocesado está almacenado en el archivo **keplerfiltered.csv**) y se realizan algunas visualizaciones y cálculos descriptivos.

| Variable | Media | Des.Est. | Mín. | P25 | Mediana | P75 | Máx. | Hist. |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| cb\_flag | 0.0024 | 0.0485 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | ▇▁▁▁▁ |
| dec | 44.2366 | 3.6695 | 36.5773 | 41.3738 | 44.3155 | 47.1127 | 52.1491 | ▅▇▇▇▃ |
| elat | 64.7746 | 3.6360 | 57.6374 | 61.8480 | 64.7690 | 67.5246 | 72.4913 | ▅▇▇▆▃ |
| elon | 307.9950 | 8.0410 | 288.7515 | 302.3200 | 308.2267 | 314.3187 | 323.9048 | ▂▅▇▇▃ |
| glat | 12.9778 | 3.4193 | 5.9607 | 10.4051 | 12.6586 | 15.7959 | 21.1430 | ▃▇▇▅▂ |
| glon | 76.3718 | 3.6938 | 68.2330 | 73.3769 | 76.4011 | 79.3326 | 84.3822 | ▃▇▇▇▃ |
| pl\_bmasse | 328.4638 | 1965.7298 | 0.7600 | 6.0133 | 18.4000 | 95.1000 | 25426.4000 | ▇▁▁▁▁ |
| pl\_bmassj | 1.0335 | 6.1849 | 0.0024 | 0.0190 | 0.0580 | 0.2992 | 80.0000 | ▇▁▁▁▁ |
| pl\_dens | 4.2419 | 8.2500 | 0.0300 | 0.6950 | 2.1300 | 5.5250 | 77.7000 | ▇▁▁▁▁ |
| pl\_eqt | 938.0333 | 564.4973 | 251.0000 | 438.7500 | 861.5000 | 1146.5000 | 2188.0000 | ▇▆▃▂▂ |
| pl\_imppar | 0.3940 | 0.2883 | 0.0000 | 0.1500 | 0.3300 | 0.6200 | 1.4830 | ▇▅▃▁▁ |
| pl\_insol | 299.3455 | 598.0846 | 0.5900 | 26.9900 | 96.9450 | 304.9000 | 4849.2600 | ▇▁▁▁▁ |
| pl\_masse | 328.4638 | 1965.7298 | 0.7600 | 6.0133 | 18.4000 | 95.1000 | 25426.4000 | ▇▁▁▁▁ |
| pl\_massj | 1.0335 | 6.1849 | 0.0024 | 0.0190 | 0.0580 | 0.2992 | 80.0000 | ▇▁▁▁▁ |
| pl\_msinie | 47.8000 | NA | 47.8000 | 47.8000 | 47.8000 | 47.8000 | 47.8000 | ▁▁▇▁▁ |
| pl\_msinij | 0.1500 | NA | 0.1500 | 0.1500 | 0.1500 | 0.1500 | 0.1500 | ▁▁▇▁▁ |
| pl\_orbeccen | 0.1177 | 0.1778 | 0.0000 | 0.0105 | 0.0420 | 0.1350 | 0.8380 | ▇▁▁▁▁ |
| pl\_orbincl | 88.6474 | 1.7458 | 82.2140 | 87.8060 | 89.1855 | 89.7905 | 93.1500 | ▁▁▃▇▁ |
| pl\_orblper | 146.9116 | 142.8866 | -163.0000 | 49.2755 | 154.7000 | 261.6000 | 357.0300 | ▂▅▆▅▇ |
| pl\_orbper | 31.4997 | 62.6948 | 0.5383 | 5.6992 | 13.0314 | 33.6013 | 1322.3000 | ▇▁▁▁▁ |
| pl\_orbsmax | 0.1834 | 0.2236 | 0.0168 | 0.0655 | 0.1091 | 0.2170 | 2.4200 | ▇▁▁▁▁ |
| pl\_orbtper | 2455003.7584 | 106.5838 | 2454935.8000 | 2454950.2605 | 2454958.2168 | 2455011.7148 | 2455162.8000 | ▇▁▁▁▂ |
| pl\_projobliq | -35.1429 | 60.4109 | -135.0000 | -61.5000 | 0.0000 | 4.0000 | 4.0000 | ▃▁▁▁▇ |
| pl\_rade | 2.9323 | 2.4308 | 0.4000 | 1.6230 | 2.3700 | 3.0240 | 30.8000 | ▇▁▁▁▁ |
| pl\_radj | 0.2616 | 0.2169 | 0.0360 | 0.1450 | 0.2110 | 0.2700 | 2.7480 | ▇▁▁▁▁ |
| pl\_ratdor | 60.9640 | 83.9114 | 3.1000 | 11.5600 | 30.3000 | 78.3000 | 576.7000 | ▇▁▁▁▁ |
| pl\_ratror | 0.0288 | 0.0263 | 0.0056 | 0.0147 | 0.0213 | 0.0286 | 0.2873 | ▇▁▁▁▁ |
| pl\_rvamp | 47.4204 | 88.2590 | 0.2800 | 2.1000 | 3.7200 | 71.9000 | 419.5000 | ▇▂▁▁▁ |
| pl\_trandep | 0.1813 | 0.3624 | 0.0060 | 0.0481 | 0.0705 | 0.1063 | 2.2620 | ▇▁▁▁▁ |
| pl\_trandur | 4.3811 | 2.2576 | 0.8244 | 2.8100 | 3.8212 | 5.4291 | 18.8860 | ▇▅▁▁▁ |
| pl\_tranmid | 2455003.7508 | 121.7668 | 2454832.9010 | 2454967.1167 | 2454973.9640 | 2455004.6853 | 2457959.9661 | ▇▁▁▁▁ |
| ra | 291.3112 | 4.6829 | 280.2066 | 287.6980 | 291.4314 | 295.0027 | 301.5430 | ▂▆▇▇▂ |
| st\_age | 4.5799 | 1.7720 | 0.1050 | 3.8000 | 4.2700 | 4.7900 | 11.9000 | ▁▇▂▁▁ |
| st\_dens | 1.6002 | 1.3618 | 0.0044 | 1.2072 | 1.4497 | 1.6696 | 19.9316 | ▇▁▁▁▁ |
| st\_logg | 4.4442 | 0.0556 | 4.2100 | 4.4100 | 4.4500 | 4.4800 | 5.0000 | ▁▇▁▁▁ |
| st\_lum | -0.0616 | 0.1126 | -0.4090 | -0.1360 | -0.0630 | 0.0280 | 0.2200 | ▁▃▇▆▂ |
| st\_mass | 0.9901 | 0.0642 | 0.6900 | 0.9500 | 0.9900 | 1.0300 | 1.2500 | ▁▁▇▃▁ |
| st\_met | 0.0199 | 0.1570 | -0.8160 | -0.0400 | 0.0200 | 0.1000 | 0.4800 | ▁▁▃▇▁ |
| st\_rad | 0.9922 | 0.0619 | 0.9000 | 0.9400 | 0.9900 | 1.0500 | 1.1000 | ▇▅▅▅▅ |
| st\_radv | -34.3600 | 29.8608 | -98.9300 | -57.1600 | -24.7600 | -20.9300 | 9.9600 | ▂▃▂▇▃ |
| st\_rotp | 12.2225 | 5.6116 | 4.6900 | 10.1775 | 11.8900 | 14.3225 | 22.0500 | ▃▁▇▂▂ |
| st\_teff | 5684.9878 | 236.0809 | 4388.3900 | 5549.5000 | 5688.0000 | 5833.5000 | 6484.0000 | ▁▁▆▇▁ |
| st\_vsin | 2.2671 | 1.9863 | 0.3000 | 0.5000 | 2.0000 | 3.0000 | 10.4000 | ▇▃▁▁▁ |
| sy\_dist | 875.6348 | 389.0261 | 68.1730 | 625.3390 | 860.3910 | 1101.1000 | 2879.8300 | ▃▇▂▁▁ |
| sy\_plx | 1.4880 | 1.2326 | 0.3269 | 0.8774 | 1.1298 | 1.5694 | 14.6396 | ▇▁▁▁▁ |
| sy\_pm | 10.9292 | 9.0598 | 0.1586 | 4.5898 | 8.4847 | 14.5826 | 77.6183 | ▇▂▁▁▁ |
| sy\_pmdec | -3.2419 | 11.7992 | -66.6869 | -9.1002 | -3.1269 | 2.5670 | 48.3025 | ▁▁▇▃▁ |
| sy\_pmra | -0.3364 | 7.1920 | -49.3144 | -3.5431 | -0.4287 | 3.0154 | 32.8344 | ▁▁▇▇▁ |
| sy\_pnum | 2.0628 | 1.2462 | 1.0000 | 1.0000 | 2.0000 | 3.0000 | 6.0000 | ▇▂▁▁▁ |
| sy\_snum | 1.0271 | 0.1624 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 2.0000 | ▇▁▁▁▁ |
| ttv\_flag | 0.1496 | 0.3567 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | ▇▁▁▁▂ |

## Análisis los parámetros numéricos del dataset

En este análisis preliminar se observan varias características en los parámetros del conjunto de datos.

Las variables **cb\_flag** y **ttv\_flag** son binarias y presentan valores mayoritariamente en **0**, ya que su media es baja (~0.0024 y ~0.1496, respectivamente). Esto indica un fuerte **desbalance en los datos**. Además, la variable **sy\_snum** casi siempre toma el valor de **1**, lo que sugiere que la mayoría de los sistemas tienen una sola estrella.

En cuanto a los parámetros de coordenadas celestes, las variables **dec** y **ra** (declinación y ascensión recta), así como **elat, elon, glat** y **glon** (latitudes y longitudes eclípticas y galácticas), muestran distribuciones relativamente centradas. **elon** tiene un rango amplio (288-323), mientras que **glat** y **glon** presentan valores bien concentrados en torno a la media, lo que sugiere la observación de un cúmulo estelar. La longitud eclíptica **elon** (288° a 324°) corresponde a la región de las constelaciones de **Capricornio y Acuario**, extendiéndose hacia **Piscis**, mientras que la latitud eclíptica **elat** (57° a 72°) es alta, indicando que estos objetos están lejos del plano de la eclíptica y ubicados en el **hemisferio norte eclíptico**.

Dado que **elat** presenta valores elevados (muy por encima del plano de la eclíptica), esto indica que los objetos en el conjunto de datos **no están alineados con el disco principal del Sistema Solar**, sino que se encuentran en una región más inclinada del cielo. Esta área corresponde en parte a la región del **Polo Norte Eclíptico**, en la dirección de la constelación del **Dragón (Draco)**.

Las variables **pl\_bmasse** y **pl\_masse** (masa en unidades terrestres), así como **pl\_bmassj** y **pl\_massj** (masa en unidades jovianas), muestran distribuciones **sesgadas a la derecha**, con valores máximos extremadamente grandes (hasta **25,426 masas terrestres**). Hay una gran cantidad de valores pequeños y algunos valores atípicos muy altos. De manera similar, pl\_rade (radio en radios terrestres) y **pl\_radj** (radio en radios jovianos) presentan una **distribución sesgada**, con pocos valores grandes. La mediana de **pl\_rade** es **2.37**, lo que indica que la mayoría de los exoplanetas en el conjunto tienen **más del doble del tamaño de la Tierra**.

En cuanto a la excentricidad orbital, **pl\_orbeccen** es mayoritariamente cercana a **0**, lo que sugiere que la mayoría de los exoplanetas tienen **órbitas casi circulares**. La inclinación orbital, representada por **pl\_orbincl**, muestra una alta concentración en valores cercanos a **90°**, lo que indica que la mayoría de los planetas tienen **órbitas casi perpendiculares a la línea de visión**. Esto es consistente con el método de detección predominante, el **método de tránsito**, que es más efectivo para planetas que pasan frente a su estrella.

El **período orbital** (**pl\_orbper**) presenta una mediana de **~13 días**, aunque existen valores extremos de hasta **1,322 días**, lo que indica una gran variabilidad en la duración de las órbitas. Por otro lado, la **temperatura planetaria de equilibrio** (**pl\_eqt**) tiene una media de **938 K**, con una distribución amplia (mínimo: **251 K**, máximo: **2,188 K**), lo que sugiere una mezcla de **planetas fríos y extremadamente calientes**.

Respecto a la **densidad planetaria** (**pl\_dens**), la mediana es **2.13 g/cm³**, con una gran variabilidad, lo que refleja una diversidad en la **composición de los exoplanetas**.

En cuanto a las características estelares, las variables **st\_mass** y **st\_rad** muestran que la mayoría de las estrellas en el conjunto tienen **masas y radios similares al Sol (~1.0)**, lo que es coherente con el filtro aplicado en el preprocesamiento. La **temperatura estelar efectiva** (**st\_teff**) tiene una media de **5,685 K**, lo que es consistente con **estrellas tipo G (similares al Sol)**. Si bien esto puede deberse al filtro aplicado, también está relacionado con la composición y edad de las estrellas.

La variable **st\_logg** (logaritmo en base 10 de la gravedad estelar superficial) se concentra en valores cercanos a **4.45**, lo que es característico de **estrellas en la secuencia principal** dentro del ciclo de vida estelar. Por su parte, la **luminosidad estelar** (**st\_lum**) tiene una distribución centrada alrededor de **0** en escala logarítmica, lo que indica que muchas de estas estrellas tienen **luminosidades similares a la del Sol**.

El parámetro **sy\_dist** (distancia en parsecs del sistema estelar) varía considerablemente, con valores entre **68 pc y 2,879 pc**. La mediana es **~860 pc**, lo que indica que **la mayoría de estos sistemas están relativamente lejos (más de 2,804 años luz)**.

En cuanto al movimiento propio de las estrellas, las variables **sy\_pm, sy\_pmra** y **sy\_pmdec** (movimiento propio del sistema estelar en milisegundos de arco por año) están **centradas en 0**, aunque algunas estrellas presentan **movimientos más rápidos**.

El **número de planetas por sistema** (**sy\_pnum**) tiene una mediana de **2**, aunque algunos sistemas alcanzan **hasta 6 planetas detectados**.

Las variables **pl\_trandep** y **pl\_trandur** (profundidad y duración del tránsito) muestran una distribución **sesgada**, con la mayoría de los valores en el extremo inferior. Esto es esperable, ya que la mayoría de los **tránsitos son poco profundos y de corta duración**.

Por otro lado, la **oblicuidad del eje de rotación del planeta** (**pl\_projobliq**) tiene una mediana en **0°** (indicando que la mayoría de los planetas tienen un eje de rotación perpendicular a la eclíptica de su sistema), aunque se observan valores extremos de hasta **-135°**.

En general, muchas distribuciones en el conjunto de datos están **sesgadas**, con valores extremos que podrían representar **atípicos (outliers) o simplemente reflejar la diversidad de los exoplanetas**. Los datos sobre **masa y radio** indican que la mayoría de los exoplanetas son **relativamente pequeños**, aunque existen algunos extremadamente grandes. La **temperatura** de los planetas varía ampliamente, lo que sugiere una gran diversidad en los tipos de exoplanetas.

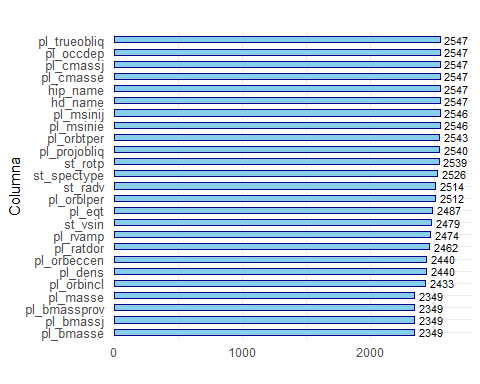
Por último, en cuanto a las estrellas, la mayoría son **similares al Sol**, aunque presentan algunas diferencias en **edad y metalicidad**, lo cual es esperable debido al **filtro aplicado en el preprocesamiento**.

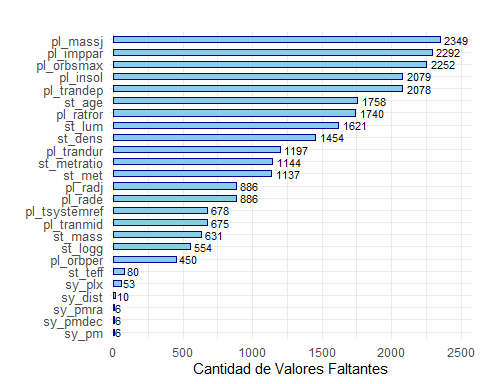
## Detención de valores faltantes

Se analiza la cantidad de valores faltantes en el dataset para identificar posibles problemas de calidad de datos.

library(tidyr)  
  
# Calcular la cantidad de valores faltantes por columna  
missing\_values <- kepler\_data %>%  
 summarise\_all(~sum(is.na(.))) %>%  
 gather() %>%  
 arrange(desc(value))  
  
# Eliminar las columnas sin valores faltantes  
missing\_values <- missing\_values %>%  
 filter(value > 0)  
  
# Total de columnas  
total\_rows <- nrow(missing\_values)  
  
# Calcular el punto medio  
midpoint <- ceiling(total\_rows / 2)

**Valores Faltantes por Columna**





En el análisis del dataset, se detectó una gran cantidad de **valores faltantes** en varias columnas, lo que probablemente se deba a la falta de estimaciones o mediciones para ciertos exoplanetas. Según los prefijos utilizados en las columnas del dataset, las variables pueden clasificarse en tres categorías principales: **parámetros del planeta (con prefijo pl\_)**, **parámetros de la estrella (con prefijo st\_)** y **parámetros del sistema (prefijo sy\_)**.

Una gran parte de los valores faltantes corresponden a **parámetros de los exoplanetas y sus sistemas**, lo que puede representar un desafío para el análisis y la predicción de su habitabilidad. Para abordar este problema, se procederá a **estimar los valores faltantes** en aquellas variables donde exista una cantidad suficiente de datos como para realizar una estimación precisa.

El objetivo principal de este estudio es **predecir la habitabilidad de los exoplanetas**, y para ello es fundamental contar con valores completos en los parámetros clave. Para lograrlo, se emplearán técnicas de **regresión lineal y logística**, que permitirán estimar los valores ausentes y mejorar la precisión del modelo de predicción.

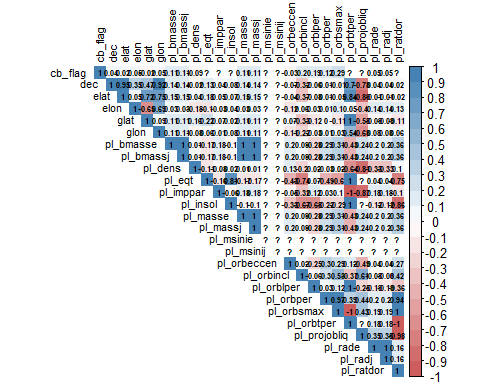
Como parte del proceso, se analizarán las **variables conocidas de los exoplanetas** para identificar aquellas que puedan actuar como **predictoras** de las variables con valores faltantes. Por ejemplo, en el caso del **radio del exoplaneta** (representado por **pl\_radj** en radios jovianos y **pl\_rade** en radios terrestres), que presenta **886 valores faltantes**, es posible estimarlo utilizando otras variables disponibles en el dataset. Algunas de estas incluyen el **período orbital en días** (**pl\_orbper**, con **450 valores faltantes**), el **tiempo de conjunción** (**pl\_tranmid**, con **675 valores faltantes**) y el **logaritmo en base 10 de la gravedad superficial de la estrella del sistema planetario** (**st\_logg**, con **554 valores faltantes**).

Para determinar cuáles de estas variables tienen un mayor potencial como **predictoras**, se utilizará la **matriz de correlación**, que permitirá identificar las relaciones más significativas entre las variables numéricas del dataset. De esta manera, se seleccionarán aquellas que presenten una **correlación positiva o negativa más fuerte** con la variable faltante, lo que facilitará su estimación con un mayor grado de precisión.

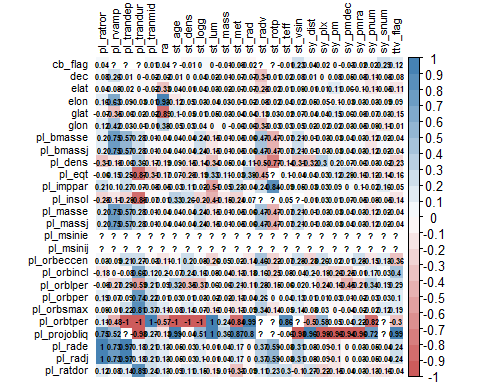
## Buscando correlaciones

Se analiza la correlación entre las variables numéricas del dataset para identificar posibles relaciones entre ellas utilizando los métodos de Pearson (detección de correlaciones lineales).

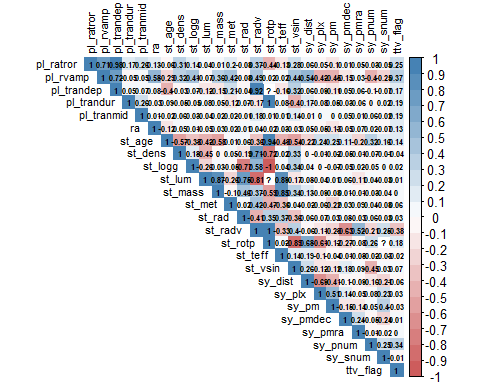
**Matriz de Correlación 1/3 (Pearson Cuadrante Superior)**



**Matriz de Correlación 2/3 (Pearson Cruce)**



**Matriz de Correlación 3/3 (Pearson Cuadrante Inferior)**



La **matriz de correlación** es una herramienta fundamental en el análisis de datos, ya que permite identificar relaciones **lineales** entre las variables numéricas del dataset. A través de esta matriz, se pueden detectar tanto **correlaciones positivas** (cuando un aumento en una variable está asociado con un aumento en otra) como **correlaciones negativas** (cuando un aumento en una variable está relacionado con una disminución en otra). Estas relaciones pueden ser útiles para entender patrones en los datos y seleccionar las mejores variables predictoras para modelos de estimación.

Un ejemplo de estas correlaciones se observa en la relación entre el **cociente del semieje orbital mayor y el período orbital** (**pl\_ratdor**), el cual muestra una **correlación positiva** con el **período orbital** (**pl\_orbper**). Esto sugiere que a medida que el período orbital aumenta, también lo hace este cociente. En contraste, **pl\_ratdor** presenta una **correlación negativa** con la **temperatura de equilibrio del exoplaneta** (**pl\_eqt**), lo que indica que los exoplanetas con períodos orbitales largos tienden a tener temperaturas más bajas, lo cual es esperable, ya que suelen estar más alejados de su estrella anfitriona.

En cuanto a la variable que se seleccionó como **variable objetivo** en el análisis previo, el **radio del exoplaneta** (**pl\_radj**), se observan relaciones significativas con otras variables. Específicamente, tiene una **correlación positiva** con la **profundidad del tránsito** (**pl\_trandep**), lo que indica que los exoplanetas más grandes generan una mayor disminución en el brillo de su estrella al pasar frente a ella. Por otro lado, **pl\_radj** muestra una **correlación negativa** con la **gravedad superficial de la estrella anfitriona** (**st\_logg**), lo que sugiere que los exoplanetas más grandes tienden a encontrarse en sistemas con estrellas de menor gravedad superficial, posiblemente porque estos sistemas albergan planetas de mayor tamaño.

Para visualizar mejor estas relaciones y seleccionar las mejores **variables predictoras**, se realiza un **pair plot** con las variables que presentan una correlación significativa, estableciendo un **umbral del 50% en valor absoluto**. Este gráfico permite examinar cómo se distribuyen los datos y cómo se relacionan las variables entre sí, facilitando la identificación de tendencias y posibles patrones útiles para la estimación de valores faltantes y la construcción de modelos de predicción.

Más allá de la relación lineal evidente entre las medidas del **radio del exoplaneta en diferentes unidades** —es decir, **radios jovianos** y **radios terrestres**—, el análisis de correlación revela la presencia de otras relaciones lineales significativas entre el radio del exoplaneta y otras variables del conjunto de datos. Estas relaciones se ordenan en función de su **fuerza de correlación** en orden descendente, destacando tres principales variables:

La primera variable con una fuerte correlación es el **cociente del radio del planeta sobre el radio de su estrella anfitriona** (**pl\_ratror**). Esta relación tiene sentido desde un punto de vista físico, ya que el tamaño relativo del exoplaneta respecto a su estrella afecta directamente la observación de fenómenos como el tránsito. Cuanto mayor sea este cociente, mayor será el tamaño aparente del exoplaneta durante el tránsito.

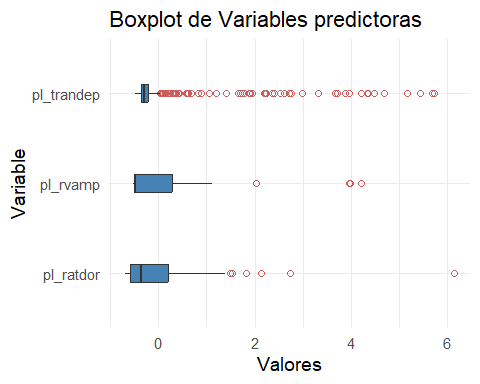
La segunda variable con alta correlación es la **profundidad del tránsito** (**pl\_trandep**). Este valor mide la disminución en el brillo de la estrella cuando el planeta pasa frente a ella. Dado que planetas más grandes bloquean una mayor parte de la luz estelar durante un tránsito, es lógico que exista una relación lineal positiva entre el radio del planeta y la profundidad del tránsito observado.

La tercera variable significativa es la **amplitud de la velocidad radial del planeta** (**pl\_rvamp**). Esta mide el desplazamiento de la estrella debido al tirón gravitacional del exoplaneta en su órbita. Los exoplanetas de mayor tamaño tienden a ejercer una fuerza gravitacional más intensa sobre la estrella, generando mayores variaciones en la velocidad radial detectada mediante el efecto Doppler, lo que explica esta relación.

Con base en estos resultados, se seleccionan estas tres variables como **variables predictoras** para el radio del exoplaneta, ya que presentan una **alta correlación lineal**, positiva o negativa, con la variable faltante. Estas relaciones proporcionan información clave para realizar estimaciones más precisas del radio del exoplaneta mediante técnicas de regresión, basándose en patrones consistentes en los datos.

## Detección de valores atípicos (outliers) en las predictoras

Para las variables predictoras seleccionadas (**pl\_ratdor**,**pl\_trandep**,**pl\_rvamp**), se detectan valores atípicos en el dataset que puedan afectar el análisis y la predicción de la variable faltante (**pl\_radj**).



El **boxplot** revela la presencia de **valores atípicos** en las variables predictoras seleccionadas, lo que puede influir significativamente en el análisis y en la precisión de la estimación de la variable faltante. Estos **outliers** pueden sesgar los resultados y reducir la eficacia de los modelos de predicción, especialmente si se utilizan técnicas sensibles a valores extremos, como la regresión lineal estándar.

Para abordar este problema, es recomendable emplear **métodos de regresión robustos**, diseñados para minimizar el impacto de valores atípicos en el ajuste del modelo. Un ejemplo de este enfoque es la **regresión de Huber**, que combina las ventajas de la regresión lineal y la regresión por mínimos cuadrados, pero con un mecanismo que reduce la influencia de outliers al asignarles un menor peso en el cálculo de los coeficientes.

Otra alternativa es utilizar técnicas como la **regresión de Tukey biweight** o la **regresión de cuantiles**, que modelan la relación entre variables sin verse afectadas por la distribución de los datos extremos. Adicionalmente, se pueden aplicar **métodos de detección y tratamiento de outliers**, como el **recorte de datos (trimming)**, que elimina valores extremos por encima o por debajo de un umbral específico, o la **transformación de variables**, como el uso de logaritmos o escalado robusto, para reducir el impacto de valores extremos en el análisis.

# Analisis de regresión

Para explorar la relación entre las características conocidas de los exoplanetas, se lleva a cabo un **análisis de regresión lineal**. El objetivo es predecir el **radio del exoplaneta (pl\_radj)** utilizando tres variables que podrían estar estrechamente relacionadas con su tamaño:

* **pl\_ratdor**: el cociente entre el radio del exoplaneta y el radio de su estrella, que proporciona una medida relativa del tamaño del planeta.
* **pl\_trandep**: la profundidad del tránsito, que indica cuánto disminuye el brillo de la estrella cuando el exoplaneta pasa frente a ella, lo que está directamente relacionado con su radio.
* **pl\_rvamp**: la amplitud de la velocidad radial del planeta, que refleja su efecto gravitacional sobre la estrella y puede aportar información sobre su masa y, en combinación con otras variables, su tamaño.

A través de este análisis, se busca comprender mejor cómo estas variables influyen en el tamaño de los exoplanetas y determinar cuál de ellas tiene un mayor impacto en la predicción de **pl\_radj**.

## Regresión lineal univariada

Las variablies de características de los sistemas exoplanetarios tales como periodo orbital, radio planetario, temperatura de la estrella central del sistema y masa del planeta pueden inferirse indirectamente a partir de otras variables como excentricidad orbital, distancia a la estrella, caracteristicas de la estrella y otras. Se explorar un análisis de regresión lineal univariada con varias variables independientes evaluando la capacidad de predicción de cada modelo planteado.

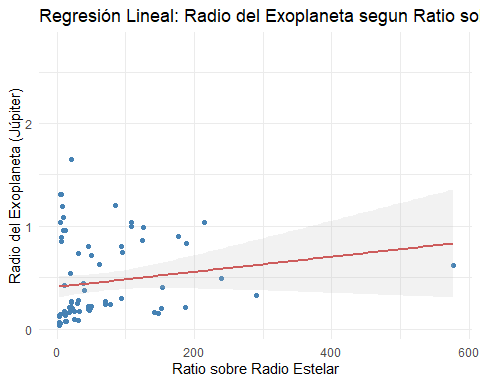
Se realiza un análisis de regresión lineal univariada para predecir el radio del exoplaneta (**pl\_radj**) consecutivamente a partir de las variables independientes seleccionadas.

### Regresión lineal: radio del exoplaneta según ratio del radio sobre el radio estelar (pl\_ratdor)

# Regresion de pl\_radj con pl\_ratdor  
model\_pl\_ratdor <- lm(pl\_radj ~ pl\_ratdor, data = kepler\_data\_num)  
  
model\_pl\_ratdor\_summary <- summary(model\_pl\_ratdor)  
  
model\_pl\_ratdor\_summary

##   
## Call:  
## lm(formula = pl\_radj ~ pl\_ratdor, data = kepler\_data\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.3749 -0.2695 -0.2125 0.3126 1.2264   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.4086542 0.0535610 7.630 0.0000000000489 \*\*\*  
## pl\_ratdor 0.0007341 0.0005069 1.448 0.152   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3872 on 78 degrees of freedom  
## (2467 observations deleted due to missingness)  
## Multiple R-squared: 0.02618, Adjusted R-squared: 0.0137   
## F-statistic: 2.097 on 1 and 78 DF, p-value: 0.1516

## `geom\_smooth()` using formula = 'y ~ x'



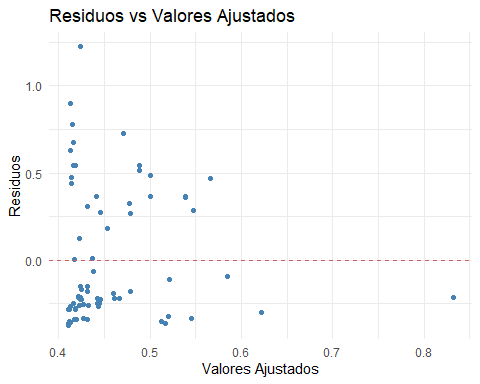
Para el modelo de regresión planteado:

El análisis de regresión sugiere que la variable **pl\_ratdor** no es un buen predictor del radio del exoplaneta (**pl\_radj**). Según los resultados, por cada aumento de 1 unidad en **pl\_ratdor**, se espera un incremento de apenas **0.0007341** en **pl\_radj**. Sin embargo, el valor p asociado a esta relación (**Pr(>|t|) = 0.152**) es mayor que **0.05**, lo que significa que **no hay suficiente evidencia estadística** para afirmar que **pl\_ratdor** influye significativamente en **pl\_radj**. Esto sugiere que el modelo planteado no es el más adecuado para estimar el radio del exoplaneta en función del cociente entre su radio y el radio de su estrella.

Este resultado se refuerza al revisar las estadísticas de ajuste del modelo. El **Error Estándar Residual** (**0.3872**) indica que los valores predichos se desvían bastante de los valores reales, lo que evidencia un ajuste impreciso. Además, el **R² múltiple** (**0.02618**) muestra que solo **el 2.6% de la variabilidad en pl\_radj** es explicada por **pl\_ratdor**, lo que indica un modelo con muy poca capacidad predictiva. De manera similar, el **R² ajustado** (**0.0137**), que corrige por el número de predictores, sigue siendo muy bajo, reforzando la idea de que **pl\_ratdor** no contribuye significativamente a la predicción del radio del exoplaneta.

Por otro lado, el **F-statistic** (2.097) y su valor p (**0.1516**) confirman que el modelo, en su conjunto, **no es estadísticamente significativo**, lo que significa que no es capaz de explicar de manera fiable la relación entre estas variables.

A pesar de estos resultados, se realiza un análisis de los **residuos del modelo** para verificar si cumplen con los supuestos de la regresión lineal, con el objetivo de entender mejor las limitaciones del modelo y explorar posibles ajustes o alternativas.

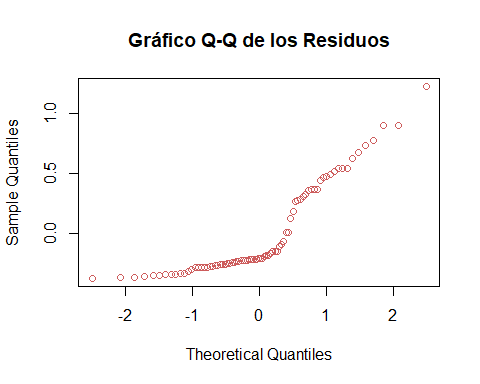


El análisis de los residuos muestra que estos tienden a concentrarse en los valores ajustados más bajos, aproximadamente entre **0.4 y 0.5**. A medida que los valores ajustados aumentan, la dispersión de los residuos también parece **incrementar ligeramente**. Sin embargo, no se observa un patrón claro en forma de parábola o curva, lo que sugiere que la relación entre la variable independiente (**pl\_ratdor**) y la dependiente (**pl\_radj**) podría ser **aproximadamente lineal**.

En general, la mayoría de los residuos se encuentran relativamente **cerca de 0**, aunque hay algunos valores positivos elevados (**mayores a 0.5**), lo que indica que el modelo tiende a **subestimar ciertos valores** de **pl\_radj**. Por otro lado, hay pocos residuos negativos extremos, lo que sugiere que el modelo **no está sobreestimando significativamente** los valores del radio del exoplaneta.

Se detecta una leve **heterocedasticidad**, ya que la dispersión de los residuos parece **mayor en los valores ajustados más altos**. Esto implica que el error no es completamente uniforme a lo largo de la distribución de valores ajustados, lo que podría afectar la **confiabilidad de las predicciones**. En un modelo ideal que cumpliera con la suposición de **homocedasticidad** (es decir, varianza constante de los errores), los residuos deberían distribuirse de manera más uniforme alrededor de la línea de referencia **(residuo = 0)**.

Dado que el modelo tiene un **R² muy bajo (~2.6%)**, era esperable encontrar una dispersión alta en los residuos, lo que confirma que **pl\_ratdor no es un buen predictor de pl\_radj**. Además, la presencia de residuos elevados sugiere que existen **otras variables no consideradas** que podrían explicar mejor la variabilidad en el radio del exoplaneta.



Para que los residuos sigan una **distribución normal**, los puntos en la gráfica deberían alinearse con una **línea diagonal imaginaria**. Sin embargo, se observan **desviaciones** en los extremos, es decir, en los cuantiles más bajos y más altos, donde los puntos se alejan de la línea teórica. Esto indica la presencia de **colas más pesadas** de lo esperado en una distribución normal, lo que sugiere que existen **valores atípicos o que los residuos no siguen completamente una distribución normal**.

Además, se nota una **ligera curvatura en la parte central del gráfico**. Aunque la mayoría de los puntos se alinean con la línea teórica en esta región, la curvatura sugiere que los residuos no son **perfectamente normales**, lo que podría afectar la validez de algunas inferencias estadísticas.

La **regresión lineal** asume que los residuos deben seguir una distribución normal para garantizar la **validez de los intervalos de confianza y de los valores p**. En este caso, la curvatura y las colas pesadas plantean dudas sobre el cumplimiento de esta suposición. Si los residuos no son normales, la inferencia estadística del modelo podría verse afectada, aunque la estimación de los coeficientes seguiría siendo válida.

Para abordar este problema, una posible solución es aplicar **transformaciones en la variable dependiente** (**pl\_radj**), como **logaritmos o raíces cuadradas**, para mejorar la normalidad de los residuos. Para confirmar estas observaciones, se realizan **pruebas de normalidad y homocedasticidad**, con el fin de evaluar si los residuos siguen efectivamente una distribución normal y si la **varianza de los errores se mantiene constante** a lo largo de los valores ajustados.

# Prueba de normalidad de Shapiro-Wilk  
shapiro\_test\_pl\_ratdor <- shapiro.test(res\_pl\_ratdor)  
  
shapiro\_test\_pl\_ratdor

##   
## Shapiro-Wilk normality test  
##   
## data: res\_pl\_ratdor  
## W = 0.81534, p-value = 0.00000001313

La **prueba de Shapiro-Wilk** es una herramienta estadística utilizada para determinar si un conjunto de datos sigue una **distribución normal**. En este análisis, se aplicó a los residuos del modelo para evaluar si cumplen con esta suposición fundamental en la regresión lineal.

Esta prueba parte de la **hipótesis nula (H₀)**, que establece que los residuos siguen una distribución normal. Si el resultado muestra un **p-valor muy bajo**, se rechaza esta hipótesis, indicando que los datos no se ajustan a una distribución normal.

El resultado de la prueba arrojó un **estadístico W de 0.81534**, lo que sugiere una desviación considerable de la normalidad, ya que un valor cercano a **1** indicaría que los datos se distribuyen de manera normal. Además, el **p-valor obtenido fue aproximadamente 0.00000001313**, lo que es significativamente menor a **0.05**. Esto proporciona suficiente evidencia para rechazar la hipótesis nula con un alto nivel de confianza, confirmando que los residuos **no siguen una distribución normal**.

Este resultado es relevante porque la falta de normalidad en los residuos puede afectar la validez de los **intervalos de confianza** y los **valores p** en la regresión. Para corregir este problema, podría ser necesario aplicar transformaciones a los datos o considerar el uso de métodos estadísticos más robustos que no dependan estrictamente de la normalidad de los residuos.

# Prueba de homocedasticidad de Breusch-Pagan  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

# Prueba de Breusch-Pagan  
bp\_test\_pl\_ratdor <- bptest(model\_pl\_ratdor)  
  
bp\_test\_pl\_ratdor

##   
## studentized Breusch-Pagan test  
##   
## data: model\_pl\_ratdor  
## BP = 1.1306, df = 1, p-value = 0.2876

El **test de Breusch-Pagan** es una prueba estadística utilizada para evaluar si los residuos del modelo presentan **homocedasticidad** (varianza constante) o **heterocedasticidad** (varianza no constante). La prueba parte de la **hipótesis nula**, que asume que los residuos tienen una varianza constante. Si se encuentra suficiente evidencia para rechazar esta hipótesis, se concluye que existe heterocedasticidad en el modelo.

En este análisis, el **estadístico de la prueba (BP)** obtuvo un valor de **1.1306**, el cual es relativamente bajo. Un valor alto de este estadístico indicaría la presencia de heterocedasticidad, pero en este caso no se observa una señal clara de ello. Dado que solo se utiliza un predictor, la prueba se basa en **un grado de libertad (df = 1)**. Además, el **p-valor obtenido fue de 0.2876**, lo que es superior al umbral de **0.05**. Esto significa que no hay suficiente evidencia estadística para rechazar la hipótesis nula, lo que sugiere que **no se detecta heterocedasticidad significativa en los residuos**.

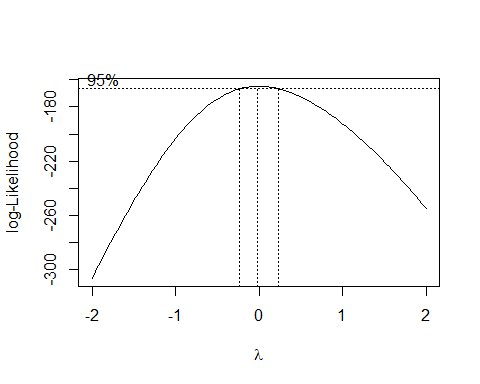
A pesar de esto, el modelo no parece ajustarse bien a los datos, ya que se observa una gran dispersión en los residuos y la posible presencia de heterocedasticidad en ciertos rangos de los valores ajustados. Para mejorar el modelo, se exploran posibles **transformaciones en la variable dependiente**, como la aplicación de logaritmos. Además, se analiza si una **transformación logarítmica en las variables predictoras** podría mejorar la capacidad predictiva del modelo, utilizando para ello la **prueba de Box-Cox**, que permite determinar la mejor transformación para estabilizar la varianza y mejorar el ajuste del modelo.

# Prueba de Box-Cox para transformar la variable st\_rad  
library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

# Prueba de Box-Cox  
boxcox\_pl\_ratdor <- boxcox(model\_pl\_ratdor, lambda = seq(-2, 2, by = 0.1))



La **transformación de Box-Cox** es una técnica utilizada en modelos de regresión para encontrar la mejor manera de transformar la variable dependiente, con el fin de mejorar la **normalidad de los residuos** y garantizar una **varianza constante** (homocedasticidad).

La gráfica generada en este análisis muestra cómo varía la **log-verosimilitud** en función del **parámetro lambda**, el cual determina la forma de la transformación aplicada a la variable dependiente **Y**. Dependiendo del valor de **lambda**, la transformación puede ser una raíz cuadrada, un logaritmo, una inversa, entre otras, ayudando a estabilizar la varianza y mejorar el ajuste del modelo.

# Obtener el lambda óptimo  
lambda\_optimal <- boxcox\_pl\_ratdor$x[which.max(boxcox\_pl\_ratdor$y)]  
  
lambda\_optimal

## [1] -0.02020202

El valor óptimo del parámetro **lambda** obtenido en la transformación de **Box-Cox** es **-0.0202**, lo que indica que la mejor transformación para la variable **pl\_ratdor** es cercana a **0**. Dado que un **lambda de 0** corresponde a una **transformación logarítmica**, aplicar el logaritmo a esta variable podría mejorar la capacidad predictiva del modelo de regresión lineal.

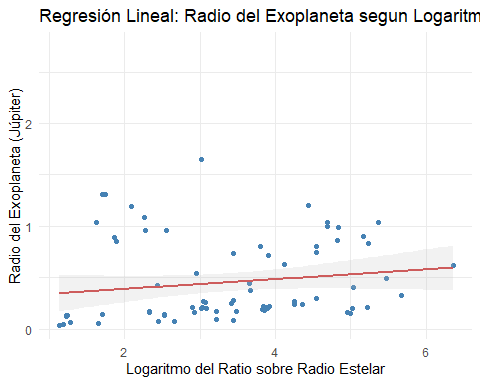
Para verificar si esta transformación realmente mejora el ajuste del modelo, se lleva a cabo una nueva **regresión lineal utilizando la variable transformada**. Luego, se evalúa el desempeño del modelo actualizado, analizando métricas como el **R²**, los **residuos** y la **significancia de los coeficientes** para determinar si la predicción de **pl\_radj** mejora con la nueva transformación.

# Regresión de pl\_radj con log(pl\_ratdor)   
model\_pl\_ratdor\_log <- lm(pl\_radj ~ log(pl\_ratdor), data = kepler\_data\_num)  
  
model\_pl\_ratdor\_log\_summary <- summary(model\_pl\_ratdor\_log)  
  
model\_pl\_ratdor\_log\_summary

##   
## Call:  
## lm(formula = pl\_radj ~ log(pl\_ratdor), data = kepler\_data\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.3776 -0.2720 -0.2215 0.2916 1.2130   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.29440 0.12286 2.396 0.019 \*  
## log(pl\_ratdor) 0.04726 0.03397 1.391 0.168   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3876 on 78 degrees of freedom  
## (2467 observations deleted due to missingness)  
## Multiple R-squared: 0.02421, Adjusted R-squared: 0.0117   
## F-statistic: 1.935 on 1 and 78 DF, p-value: 0.1681

# Scatter plot de pl\_radj vs log(pl\_ratdor) con la recta de regresión en rojo  
# y los margenes de confianza en gris  
ggplot(kepler\_data\_num, aes(x = log(pl\_ratdor), y = pl\_radj)) +  
 geom\_point(color = "steelblue") +  
 geom\_smooth(method = "lm", color = "indianred", fill = "grey", alpha = 0.2) +  
 labs(title = "Regresión Lineal: Radio del Exoplaneta segun Logaritmo del Ratio sobre el Radio Estelar",  
 x = "Logaritmo del Ratio sobre Radio Estelar",  
 y = "Radio del Exoplaneta (Júpiter)") +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'



El modelo ajustado con la transformación logarítmica de **pl\_ratdor** muestra resultados muy similares al modelo original, lo que indica que la transformación **no mejoró significativamente la capacidad predictiva**. El **error estándar de los residuos** es **0.3876**, prácticamente idéntico al del modelo sin transformación (**0.3872**), lo que sugiere que la variabilidad de los residuos sigue siendo la misma.

El coeficiente de determinación **R² múltiple** es **0.02421 (2.4%)**, lo que significa que solo un **2.4% de la variabilidad de pl\_radj** es explicada por **log(pl\_ratdor)**, un valor extremadamente bajo. Incluso al ajustar por el número de predictores, el **R² ajustado** apenas mejora, quedando en **1.2%**, lo que refuerza la idea de que el modelo sigue sin capturar bien la relación entre las variables.

El **estadístico F**, con un valor de **1.935** y un **p-valor de 0.1681**, indica que no hay suficiente evidencia estadística para afirmar que la transformación logarítmica de **pl\_ratdor** mejora significativamente el modelo.

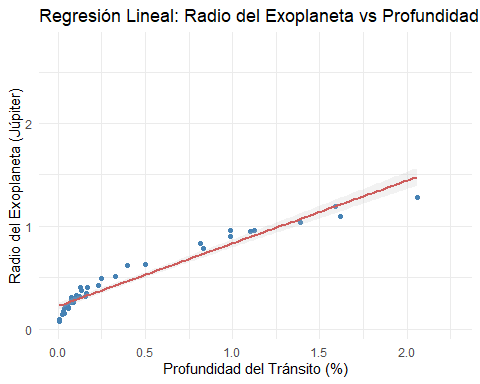
En conclusión, el modelo sigue sin ser adecuado para predecir **pl\_radj** a partir de **pl\_ratdor**, incluso después de aplicar la transformación logarítmica. Ante estos resultados, se procede a probar con **otras variables predictoras o combinaciones de variables** para mejorar la capacidad de ajuste del modelo.

### Regresión lineal: radio del exoplaneta según la profundidad del tránsito (pl\_trandep)

# Regresion del radio del exploplaneta (pl\_radj)   
# con la profundidad del tránsito del planeta (pl\_trandep)  
# como predictor  
model\_pl\_trandep <- lm(pl\_radj ~ pl\_trandep, data = kepler\_data\_num)  
  
model\_pl\_trandep\_summary <- summary(model\_pl\_trandep)  
  
model\_pl\_trandep\_summary

##   
## Call:  
## lm(formula = pl\_radj ~ pl\_trandep, data = kepler\_data\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.203530 -0.056704 0.002386 0.056979 0.150580   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.22308 0.01590 14.03 <0.0000000000000002 \*\*\*  
## pl\_trandep 0.61058 0.02434 25.08 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08315 on 40 degrees of freedom  
## (2505 observations deleted due to missingness)  
## Multiple R-squared: 0.9402, Adjusted R-squared: 0.9387   
## F-statistic: 629.1 on 1 and 40 DF, p-value: < 0.00000000000000022

## `geom\_smooth()` using formula = 'y ~ x'



En el modelo:

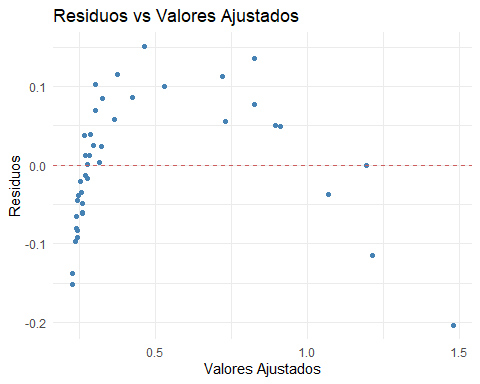
Cada aumento unitario en **pl\_trandep** se asocia con un incremento de 0.61058 en **pl\_radj**. Además es altamente significativo (con p-value < 0.00000000000000022), lo que indica una fuerte relación entre **pl\_trandep** y **pl\_radj**. Además: El modelo de regresión planteado para predecir **pl\_radj** en función de **pl\_trandep** muestra una relación clara y significativa entre ambas variables. En este caso, **por cada aumento unitario en pl\_trandep**, se observa un **incremento de 0.61058 en pl\_radj**. Además, esta relación es altamente significativa, con un **p-valor extremadamente bajo** (< 2.2 × 10⁻¹⁶), lo que indica que hay una fuerte conexión entre ambas variables.

El modelo presenta un **error estándar de los residuos de 0.08315**, lo que sugiere que la variabilidad en las predicciones es muy baja y, por lo tanto, el ajuste del modelo es bastante preciso. Además, el **coeficiente de determinación R²** alcanza un **94.02%**, lo que significa que el modelo explica **casi toda la variabilidad** de **pl\_radj**. Incluso el **R² ajustado**, que corrige por el número de predictores, sigue siendo **muy alto (93.87%)**, lo que confirma la solidez del ajuste.

La prueba de **significancia global del modelo**, evaluada con el **estadístico F**, arroja un valor de **629.1** con un **p-valor extremadamente bajo**, lo que refuerza la validez del modelo y su capacidad para explicar la relación entre las variables.

En comparación con los modelos anteriores (**pl\_radj ~ pl\_ratdor** y **pl\_radj ~ log(pl\_ratdor)**), este nuevo modelo muestra una mejora **drástica** en el ajuste, con una capacidad explicativa mucho mayor. La relación entre **pl\_trandep** y **pl\_radj** no solo es fuerte, sino que también es estadísticamente **muy significativa**, lo que indica que **pl\_trandep** es un excelente predictor del radio del exoplaneta.

# Obtener los residuos y valores ajustados  
res\_pl\_trandep <- residuals(model\_pl\_trandep)  
  
# Calcular los valores ajustados  
fit\_val\_pl\_trandep <- fitted(model\_pl\_trandep)  
  
# Crear un data frame con los residuos y valores ajustados  
res\_pl\_trandep\_df <- data.frame(  
 Residuals = res\_pl\_trandep,  
 Fitted\_Values = fit\_val\_pl\_trandep  
)  
  
# Scatter plot de residuos vs valores ajustados  
ggplot(res\_pl\_trandep\_df, aes(x = Fitted\_Values, y = Residuals)) +  
 geom\_point(color = "steelblue") +  
 geom\_hline(yintercept = 0, color = "indianred", linetype = "dashed") +  
 labs(title = "Residuos vs Valores Ajustados",  
 x = "Valores Ajustados",  
 y = "Residuos") +  
 theme\_minimal()

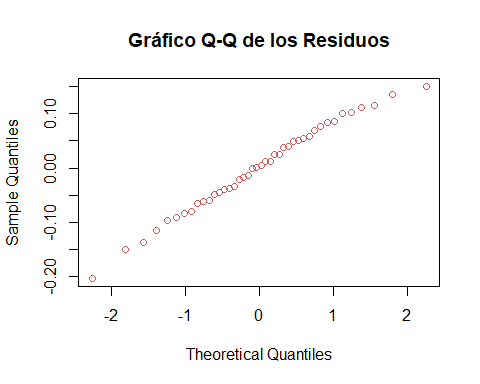


El gráfico de **residuos vs valores ajustados** revela un patrón en la dispersión de los residuos. Para valores ajustados bajos, los residuos tienden a agruparse cerca de **0**, lo que indica una menor variabilidad en esta región. Sin embargo, a medida que los valores ajustados aumentan, los residuos se dispersan más, sugiriendo que el error del modelo no es uniforme en todo el rango de predicciones.

Este comportamiento sugiere la presencia de **heterocedasticidad**, ya que la dispersión de los residuos parece seguir un **patrón de parábola invertida**, con una mayor variabilidad en los valores ajustados más altos. La **heterocedasticidad** ocurre cuando la varianza de los errores no es constante, lo que puede afectar la fiabilidad de las inferencias estadísticas del modelo. Para confirmar si este problema está presente, se puede realizar el **test de Breusch-Pagan**, que evalúa si la varianza de los residuos cambia sistemáticamente con los valores ajustados.

En un modelo bien especificado, los residuos deberían distribuirse **de manera aleatoria** alrededor de la línea de referencia **(y=0)**, sin mostrar patrones claros. Sin embargo, en este caso, se observa una mayor concentración de puntos en la parte izquierda del gráfico, lo que sugiere que el modelo podría no estar captando completamente la relación entre **pl\_radj** y **pl\_trandep**. Este comportamiento indica que el modelo podría beneficiarse de una revisión, como la inclusión de nuevas variables explicativas o la aplicación de transformaciones para mejorar la estabilidad de la varianza de los errores.

# Crear un gráfico Q-Q de los residuos  
qqnorm(res\_pl\_trandep, main = "Gráfico Q-Q de los Residuos", col="indianred")



El **gráfico Q-Q (Quantile-Quantile)** de los residuos permite evaluar si estos siguen una **distribución normal** comparando sus cuantiles con los de una distribución normal teórica. En este caso, los puntos del gráfico se alinean bastante bien con la **línea diagonal**, lo que indica que, en general, los residuos siguen un **comportamiento cercano a la normalidad**.

Sin embargo, se observan **pequeñas desviaciones en los extremos**, tanto en los valores más negativos como en los más positivos. Esto sugiere la presencia de **colas más pesadas** en la distribución de los residuos o la existencia de algunos **valores atípicos leves**. A pesar de estas desviaciones, no parecen ser lo suficientemente grandes como para indicar una **violación grave de la normalidad**.

Para complementar este análisis y confirmar si los residuos cumplen con los **supuestos de normalidad y homocedasticidad**, se realizan pruebas estadísticas específicas. Estas pruebas permiten verificar con mayor precisión si los residuos siguen efectivamente una distribución normal y si la **varianza de los errores es constante** en todo el rango de valores ajustados. Si se detectan problemas significativos, podrían ser necesarias transformaciones en los datos o el uso de modelos más robustos para mejorar la validez del análisis.

# Prueba de normalidad de Shapiro-Wilk  
shapiro\_test\_pl\_trandep <- shapiro.test(res\_pl\_trandep)  
  
shapiro\_test\_pl\_trandep

##   
## Shapiro-Wilk normality test  
##   
## data: res\_pl\_trandep  
## W = 0.98715, p-value = 0.9115

La **prueba de Shapiro-Wilk** se utilizó para evaluar si los residuos del modelo siguen una distribución normal. El resultado arrojó un **estadístico W de 0.98715**, un valor cercano a **1**, lo que indica que los datos se ajustan bien a una distribución normal. Además, el **p-valor obtenido fue de 0.9115**, lo que, al ser mayor que **0.05**, significa que no hay suficiente evidencia estadística para rechazar la **hipótesis nula**. En otras palabras, los residuos **no presentan desviaciones significativas de la normalidad**.

Este resultado confirma que el **supuesto de normalidad de los residuos se cumple**, lo que hace que el modelo sea **estadísticamente válido para realizar inferencias**. Además, es coherente con la observación del **gráfico Q-Q**, que ya sugería una buena alineación de los residuos con la distribución normal. Esto refuerza la confianza en la fiabilidad del modelo y en la precisión de sus estimaciones.

# Prueba de homocedasticidad de Breusch-Pagan  
library(lmtest)  
  
# Prueba de Breusch-Pagan  
bp\_test\_pl\_trandep <- bptest(model\_pl\_trandep)  
  
bp\_test\_pl\_trandep

##   
## studentized Breusch-Pagan test  
##   
## data: model\_pl\_trandep  
## BP = 6.2822, df = 1, p-value = 0.0122

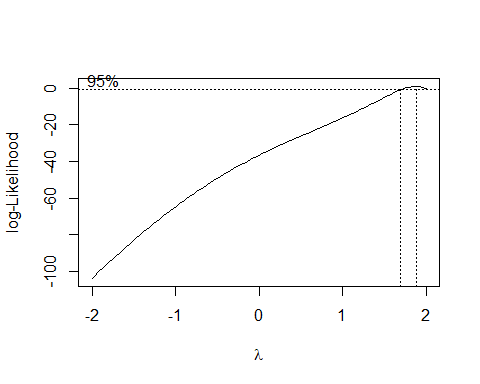
El **test de homocedasticidad de Breusch-Pagan** se utilizó para evaluar si la varianza de los errores en el modelo es constante. Los resultados muestran un **estadístico BP de 6.2822**, lo que sugiere una mayor evidencia de **heterocedasticidad**. Dado que solo se incluye un predictor en el modelo, la prueba se basa en **un grado de libertad**. Además, el **p-valor obtenido fue de 0.0122**, lo que, al ser menor que **0.05**, indica que se debe **rechazar la hipótesis nula** y concluir que los residuos presentan **heterocedasticidad significativa**. En otras palabras, la varianza de los errores no es constante en todo el rango de valores ajustados.

Esta violación de la suposición de **homocedasticidad** puede afectar la **precisión de las estimaciones y pruebas de hipótesis**, haciendo que los **intervalos de confianza y valores p sean menos confiables** debido a la variabilidad desigual en los errores del modelo.

Dado que esta condición puede comprometer la validez de los resultados, es importante considerar estrategias para corregir la heterocedasticidad. Una opción es aplicar una **transformación en la variable dependiente**, como el **logaritmo (log(Y)) o la raíz cuadrada (sqrt(Y))**, para estabilizar la varianza de los errores. Otra alternativa es utilizar una **regresión robusta con errores estándar de White**, que ajusta los errores para hacerlos menos sensibles a la heterocedasticidad. También puede aplicarse una **regresión ponderada (WLS, Weighted Least Squares)**, en la que se asignan pesos a las observaciones para compensar la variabilidad desigual en los errores, especialmente si la heterocedasticidad está relacionada con una variable específica.

Como primer intento de corrección, se aplicará una **transformación a la variable dependiente pl\_radj**, con el objetivo de reducir la heterocedasticidad y mejorar la estabilidad del modelo.

# Prueba de Box-Cox para transformar la variable pl\_trandep  
library(MASS)  
  
# Prueba de Box-Cox  
boxcox\_pl\_trandep <- boxcox(model\_pl\_trandep, lambda = seq(-2, 2, by = 0.1))



El análisis de la **transformación de Box-Cox** muestra que el **pico de la curva** ocurre aproximadamente en **lambda ≈ 2**, lo que indica que la mejor transformación para la variable dependiente se encuentra en esa región. Como este valor está dentro del intervalo de confianza, se considera una opción adecuada para transformar los datos y mejorar el ajuste del modelo.

Dado que el valor óptimo de lambda es cercano a **2**, la transformación sugerida para la variable dependiente es **elevarla al cuadrado (Y²)**. Este tipo de transformación puede ayudar a estabilizar la varianza de los residuos y mejorar la linealidad de la relación entre las variables.

Por otro lado, no se recomienda utilizar una **transformación logarítmica (log(Y))** correspondiente a **lambda = 0**, ni la **raíz cuadrada (sqrt(Y))**, que correspondería a **lambda = 0.5**. Ambas opciones quedan fuera del intervalo óptimo definido por la prueba y, por lo tanto, no ofrecerían una mejora significativa en la normalidad de los residuos ni en la homocedasticidad del modelo.

# Regresión de pl\_radj^2 con pl\_trandep  
model\_pl\_trandep\_square <- lm(I(pl\_radj^2) ~ pl\_trandep, data = kepler\_data\_num)  
  
model\_pl\_trandep\_square\_summary <- summary(model\_pl\_trandep\_square)  
  
model\_pl\_trandep\_square\_summary

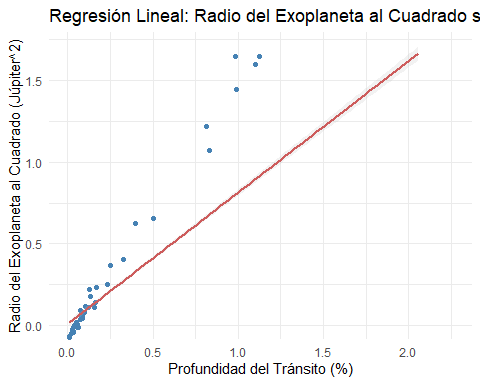
##   
## Call:  
## lm(formula = I(pl\_radj^2) ~ pl\_trandep, data = kepler\_data\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.108946 -0.015184 -0.004665 0.010102 0.134871   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.010364 0.007744 1.338 0.188   
## pl\_trandep 0.803698 0.011855 67.795 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04049 on 40 degrees of freedom  
## (2505 observations deleted due to missingness)  
## Multiple R-squared: 0.9914, Adjusted R-squared: 0.9912   
## F-statistic: 4596 on 1 and 40 DF, p-value: < 0.00000000000000022

# Scatter plot de pl\_radj^2 vs pl\_trandep con la recta de regresión en rojo  
# y los margenes de confianza en gris  
ggplot(kepler\_data\_num, aes(x = pl\_trandep, y = I(pl\_radj^2)))+  
 geom\_point(color = "steelblue") +  
 geom\_smooth(method = "lm", color = "indianred", fill = "grey", alpha = 0.2) +  
 labs(title = "Regresión Lineal: Radio del Exoplaneta al Cuadrado segun Profundidad del Tránsito",  
 x = "Profundidad del Tránsito (%)",  
 y = "Radio del Exoplaneta al Cuadrado (Júpiter^2)") +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'

## Warning: Removed 2505 rows containing non-finite outside the scale range  
## (`stat\_smooth()`).

## Warning: Removed 2505 rows containing missing values or values outside the scale range  
## (`geom\_point()`).



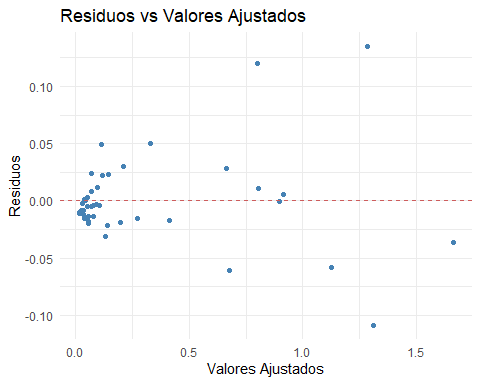
El análisis del modelo después de aplicar la transformación cuadrática muestra una mejora significativa en su precisión y capacidad predictiva. El **error estándar de los residuos** se redujo a **0.04049**, un valor menor en comparación con el modelo sin transformación, lo que indica una mayor precisión en las predicciones.

El modelo ahora explica **el 99.14% de la variabilidad en pl\_radj²**, con un **R² de 0.9914**, lo que refleja un ajuste excepcional. Incluso al corregir por el número de predictores, el **R² ajustado** sigue siendo extremadamente alto, con un valor de **0.9912**, lo que confirma que la relación entre las variables es muy sólida.

Además, el **estadístico F**, con un valor de **4,596** y un **p-valor extremadamente bajo (< 2.2 × 10⁻¹⁶)**, demuestra que el modelo es altamente significativo.

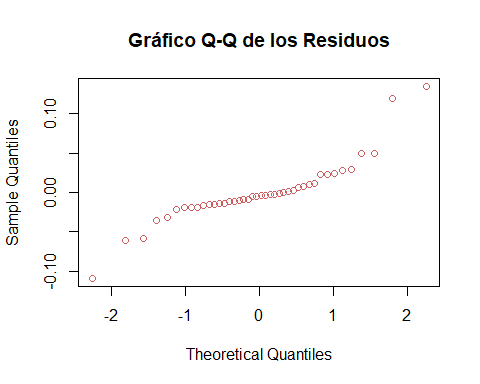
En general, la transformación cuadrática mejoró notablemente el ajuste en comparación con el modelo original. La relación entre **pl\_trandep** y **pl\_radj²** no solo es fuerte, sino también estadísticamente significativa, lo que sugiere que este modelo es una representación más precisa de la relación entre las variables.

# Obtener los residuos y valores ajustados  
res\_pl\_trandep\_square <- residuals(model\_pl\_trandep\_square)  
  
# Calcular los valores ajustados  
fit\_val\_pl\_trandep\_square <- fitted(model\_pl\_trandep\_square)  
  
# Crear un data frame con los residuos y valores ajustados  
res\_pl\_trandep\_square\_df <- data.frame(  
 Residuals = res\_pl\_trandep\_square,  
 Fitted\_Values = fit\_val\_pl\_trandep\_square  
)  
  
# Scatter plot de residuos vs valores ajustados  
ggplot(res\_pl\_trandep\_square\_df, aes(x = Fitted\_Values, y = Residuals)) +  
 geom\_point(color = "steelblue") +  
 geom\_hline(yintercept = 0, color = "indianred", linetype = "dashed") +  
 labs(title = "Residuos vs Valores Ajustados",  
 x = "Valores Ajustados",  
 y = "Residuos") +  
 theme\_minimal()



Los residuos tienen menor dispersión, lo que confirma que la heterocedasticidad detectada anteriormente fue reducida con la transformación.

# Crear un gráfico Q-Q de los residuos  
qqnorm(res\_pl\_trandep\_square, main = "Gráfico Q-Q de los Residuos", col="indianred")



En el gráfico Q-Q de los residuos, los puntos se alinean bastante bien con la línea diagonal, lo que sugiere que los residuos siguen una distribución normal. Aunque hay una pequeña desviación en los extremos (valores muy negativos y positivos), la desviación no es severa.

# Prueba de normalidad de Shapiro-Wilk  
shapiro\_test\_pl\_trandep\_square <- shapiro.test(res\_pl\_trandep\_square)  
  
shapiro\_test\_pl\_trandep\_square

##   
## Shapiro-Wilk normality test  
##   
## data: res\_pl\_trandep\_square  
## W = 0.8507, p-value = 0.00006504

El **test de normalidad de Shapiro-Wilk** arroja un **estadístico W de 0.8507** y un **p-valor de 0.00006504**, lo que indica que los residuos **no siguen una distribución normal**. A pesar de que en el **gráfico Q-Q** los puntos parecían alinearse de manera razonable con la línea diagonal, la prueba estadística proporciona evidencia suficiente para rechazar la hipótesis de normalidad.

library(lmtest)  
# Prueba de homocedasticidad de Breusch-Pagan  
bp\_test\_pl\_trandep\_square <- bptest(model\_pl\_trandep\_square)  
  
bp\_test\_pl\_trandep\_square

##   
## studentized Breusch-Pagan test  
##   
## data: model\_pl\_trandep\_square  
## BP = 14.87, df = 1, p-value = 0.0001152

El **test de Breusch-Pagan** arroja un **estadístico BP de 14.87**, con **1 grado de libertad** y un **p-valor de 0.0001152**, lo que indica que los residuos presentan **heterocedasticidad significativa**. Esto significa que la varianza de los errores no es constante en todo el modelo, lo que puede afectar la fiabilidad de las inferencias estadísticas.

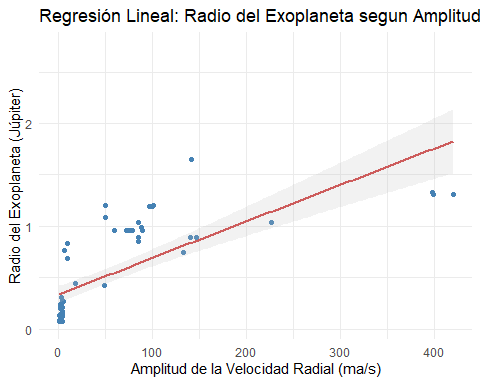
Aunque la **transformación cuadrática** aplicada a la variable dependiente ayudó a **reducir la heterocedasticidad**, los resultados muestran que aún persiste en cierta medida. Esto sugiere que, aunque el modelo ha mejorado, puede ser necesario aplicar **otras estrategias**, como regresión ponderada o el uso de errores estándar robustos, para garantizar una estimación más estable y confiable.

### Regresión lineal: radio del exoplaneta según la amplitud de la velocidad radial (pl\_rvamp)

# Regresion del radio del exoplaneta (pl\_radj)   
# utilizando la amplitude de la velocidad radial del planeta (pl\_rvamp)  
# como predictor  
model\_pl\_rvamp <- lm(pl\_radj ~ pl\_rvamp, data = kepler\_data\_num)  
  
model\_pl\_rvamp\_summary <- summary(model\_pl\_rvamp)  
  
model\_pl\_rvamp\_summary

##   
## Call:  
## lm(formula = pl\_radj ~ pl\_rvamp, data = kepler\_data\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5172 -0.2119 -0.1206 0.2644 0.8116   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.337785 0.041804 8.080 0.00000000001903 \*\*\*  
## pl\_rvamp 0.003550 0.000406 8.744 0.00000000000124 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3026 on 66 degrees of freedom  
## (2479 observations deleted due to missingness)  
## Multiple R-squared: 0.5367, Adjusted R-squared: 0.5297   
## F-statistic: 76.46 on 1 and 66 DF, p-value: 0.000000000001239

## `geom\_smooth()` using formula = 'y ~ x'



Para el modelo de regresión planteado:

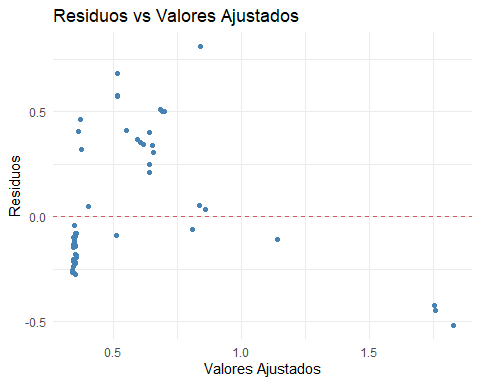
Las estadísticas del modelo muestran que:

El modelo presenta un **R² de 0.5367**, lo que indica que **explica el 53.67% de la variabilidad en pl\_radj**. Esto sugiere una relación **moderada a fuerte** entre las variables, aunque todavía queda un **46.33% de variabilidad sin explicar**, lo que implica que otros factores podrían estar influyendo en la variable dependiente.

El **R² ajustado** es **0.5297**, lo que significa que, incluso al corregir por el número de predictores, el modelo sigue manteniendo un **buen nivel de ajuste**. Además, el **estadístico F** tiene un valor de **76.46**, con un **p-valor extremadamente bajo**, lo que confirma que el modelo es **altamente significativo** y que la relación entre las variables no es producto del azar.

A pesar de que el modelo tiene un ajuste razonable, no es tan preciso como el obtenido en la regresión de **pl\_radj ~ pl\_trandep**. Sin embargo, la relación entre **pl\_rvamp y pl\_radj** sigue siendo **estadísticamente significativa** y muestra un nivel de asociación importante, lo que sugiere que pl\_rvamp es un predictor relevante, aunque no el más fuerte disponible.

# Obtener los residuos y valores ajustados  
res\_pl\_rvamp <- residuals(model\_pl\_rvamp)  
  
# Calcular los valores ajustados  
fit\_val\_pl\_rvamp <- fitted(model\_pl\_rvamp)  
  
# Crear un data frame con los residuos y valores ajustados  
res\_pl\_rvamp\_df <- data.frame(  
 Residuals = res\_pl\_rvamp,  
 Fitted\_Values = fit\_val\_pl\_rvamp  
)  
  
# Scatter plot de residuos vs valores ajustados  
ggplot(res\_pl\_rvamp\_df, aes(x = Fitted\_Values, y = Residuals)) +  
 geom\_point(color = "steelblue") +  
 geom\_hline(yintercept = 0, color = "indianred", linetype = "dashed") +  
 labs(title = "Residuos vs Valores Ajustados",  
 x = "Valores Ajustados",  
 y = "Residuos") +  
 theme\_minimal()

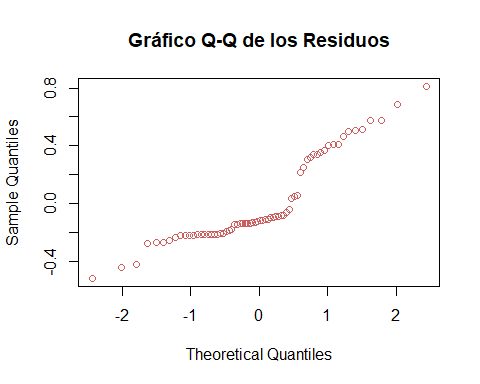


El análisis de los residuos revela que su **dispersión no es constante** a lo largo de los valores ajustados, lo que sugiere la posible presencia de **heterocedasticidad**. En los valores ajustados más pequeños (**menores a 0.5**), los residuos están **muy concentrados**, lo que indica una menor variabilidad en esta región. Sin embargo, a medida que los valores ajustados aumentan (**mayores a 1.0**), los residuos muestran una dispersión mucho mayor, lo que sugiere que la **varianza de los errores no es constante en todo el modelo**.

La presencia de **heterocedasticidad** puede indicar que el modelo **no está capturando adecuadamente la variabilidad de los datos**, lo que podría afectar la precisión de sus predicciones. Para corregir este problema, se pueden aplicar **transformaciones en las variables**, como el uso de logaritmos o raíces cuadradas, o utilizar **métodos de regresión robustos** que ajusten los errores estándar para hacerlos menos sensibles a la variabilidad desigual.

Además, si los residuos muestran una **tendencia curva o patrones de agrupación**, esto podría ser una señal de que el **modelo lineal no es la mejor opción**. En estos casos, podría ser necesario recurrir a un **modelo no lineal** o incluir **términos polinomiales** en la regresión para capturar mejor la relación entre las variables y mejorar la precisión del ajuste.

# Crear un gráfico Q-Q de los residuos  
qqnorm(res\_pl\_rvamp, main = "Gráfico Q-Q de los Residuos", col="indianred")



El análisis del **gráfico Q-Q de los residuos** muestra **desviaciones significativas en los extremos**, especialmente en los cuantiles más bajos y más altos. Esto sugiere que la distribución de los residuos **no sigue completamente una distribución normal** y que **las colas de la distribución son más pesadas de lo esperado**. En otras palabras, hay más valores extremos en los residuos de lo que se esperaría en una distribución normal ideal.

Los residuos más alejados de la media, tanto en la dirección **negativa como positiva**, se desvían notablemente de la **línea teórica**, lo que indica la presencia de valores atípicos que podrían estar afectando el modelo. A pesar de esto, la mayor parte de los puntos en la **zona central** del gráfico parecen alinearse mejor con la distribución normal, aunque los extremos muestran una curvatura evidente.

La falta de normalidad en los residuos pone en duda la **validez de los intervalos de confianza y las pruebas de hipótesis del modelo**, ya que muchos procedimientos estadísticos dependen de esta suposición. La presencia de valores extremos también sugiere que **outliers o una distribución sesgada** podrían estar influyendo en el comportamiento del modelo. Para abordar este problema, se pueden considerar **transformaciones de las variables, la eliminación de valores atípicos** o el uso de **modelos más robustos** que no dependan estrictamente de la normalidad de los errores.

# Prueba de normalidad de Shapiro-Wilk  
shapiro\_test\_pl\_rvamp <- shapiro.test(res\_pl\_rvamp)  
  
shapiro\_test\_pl\_rvamp

##   
## Shapiro-Wilk normality test  
##   
## data: res\_pl\_rvamp  
## W = 0.86708, p-value = 0.000003207

El resultado de la **prueba de normalidad de Shapiro-Wilk** indica que los residuos **no siguen una distribución normal**, con un **estadístico W de 0.86708** y un **p-valor de 0.000003207**. Dado que este p-valor es extremadamente bajo, hay suficiente evidencia para rechazar la hipótesis de normalidad.

Aunque en el **gráfico Q-Q** los residuos parecían alinearse en cierta medida con la línea diagonal, la prueba estadística confirma que **existen desviaciones significativas**. Esto sugiere que los residuos podrían estar influenciados por la **presencia de valores atípicos** o que su distribución tiene características diferentes a las de una normal ideal.

La falta de normalidad en los residuos podría afectar la **validez de los intervalos de confianza y las pruebas de hipótesis**, lo que hace necesario considerar posibles soluciones, como aplicar **transformaciones a los datos**, eliminar valores extremos o utilizar **métodos estadísticos más robustos** que no dependan estrictamente de la normalidad de los errores.

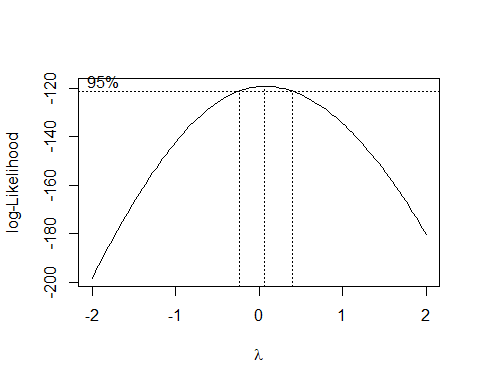
# Prueba de homocedasticidad de Breusch-Pagan  
library(lmtest)  
  
# Prueba de Breusch-Pagan  
bp\_test\_pl\_rvamp <- bptest(model\_pl\_rvamp)  
  
bp\_test\_pl\_rvamp

##   
## studentized Breusch-Pagan test  
##   
## data: model\_pl\_rvamp  
## BP = 10.096, df = 1, p-value = 0.001486

El resultado del **test de Breusch-Pagan** indica la presencia de **heterocedasticidad significativa**, con un **estadístico BP de 10.096** y un **p-valor de 0.001486**. Esto significa que la **varianza de los errores no es constante**, lo que puede comprometer la precisión de las **estimaciones del modelo y las pruebas de hipótesis**.

Dado que el modelo no cumple con la **suposición de homocedasticidad** y, además, los residuos **no siguen una distribución normal**, se busca corregir la heterocedasticidad mediante una **transformación de la variable dependiente**. Para determinar cuál es la mejor transformación, se lleva a cabo una **prueba de Box-Cox**, que permite identificar la forma más adecuada para estabilizar la varianza de los errores y mejorar el ajuste del modelo.

# Prueba de Box-Cox para transformar la variable pl\_rvamp  
library(MASS)  
  
# Prueba de Box-Cox  
boxcox\_pl\_rvamp <- boxcox(model\_pl\_rvamp, lambda = seq(-2, 2, by = 0.1))



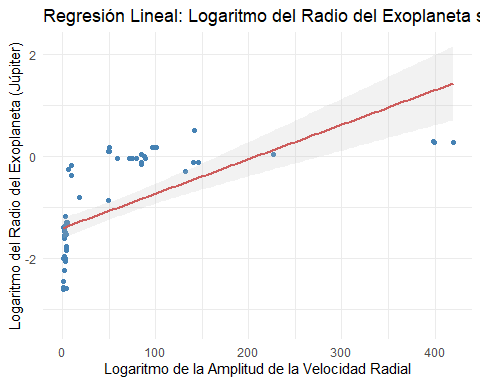
El **pico de la curva** en la prueba de **Box-Cox** se encuentra cerca de **lambda = 0**, lo que indica que la mejor transformación para la variable dependiente se encuentra en esa zona. En términos prácticos, cuando el valor óptimo de **lambda (λ)** es **0**, la transformación recomendada es el **logaritmo natural** de la variable dependiente.

Dado que **λ = 0 está dentro del intervalo de confianza**, esto confirma que aplicar una **transformación logarítmica (log(Y))** es la opción más adecuada. Este tipo de transformación es útil porque puede **reducir la heterocedasticidad**, mejorar la **normalidad de los residuos** y hacer que la relación entre las variables sea **más lineal**, lo que permite que el modelo de regresión tenga un mejor ajuste y sea más preciso en sus estimaciones.

# Regresión de log(pl\_radj) con logpl\_rvamp  
model\_pl\_rvamp\_log <- lm(log(pl\_radj) ~ log(pl\_rvamp), data = kepler\_data\_num)  
  
model\_pl\_rvamp\_log\_summary <- summary(model\_pl\_rvamp\_log)  
  
model\_pl\_rvamp\_log\_summary

##   
## Call:  
## lm(formula = log(pl\_radj) ~ log(pl\_rvamp), data = kepler\_data\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.09181 -0.30754 0.04893 0.21513 0.98300   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.06303 0.06935 -29.75 <0.0000000000000002 \*\*\*  
## log(pl\_rvamp) 0.44267 0.02356 18.79 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3753 on 66 degrees of freedom  
## (2479 observations deleted due to missingness)  
## Multiple R-squared: 0.8425, Adjusted R-squared: 0.8401   
## F-statistic: 353 on 1 and 66 DF, p-value: < 0.00000000000000022

## `geom\_smooth()` using formula = 'y ~ x'



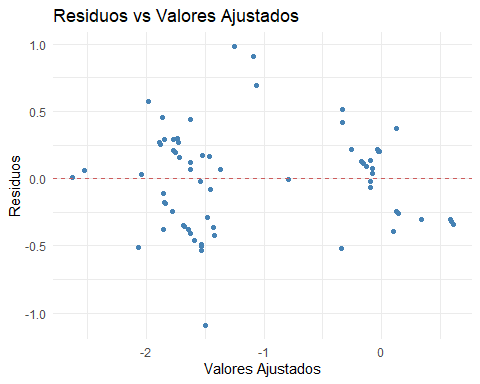
Para el modelo de regresión planteado:

El análisis del modelo después de aplicar la **transformación logarítmica** muestra una mejora significativa en su capacidad predictiva. El **R² obtenido es de 84.25%**, lo que indica que el modelo explica una gran parte de la variabilidad en **log(pl\_radj)**, representando un ajuste excelente. Este resultado es considerablemente mejor que el del modelo sin transformación, donde el **R² era aproximadamente 53.67%**.

Incluso al corregir por el número de predictores, el **R² ajustado** sigue siendo alto, con un valor de **84.01%**, lo que confirma que el modelo mantiene su solidez y no se ve afectado por un exceso de predictores innecesarios. Además, el **estadístico F**, con un valor de **353** y un **p-valor extremadamente bajo (< 2.2 × 10⁻¹⁶)**, confirma que el modelo es altamente significativo y que la relación entre las variables no es producto del azar.

En conclusión, la **transformación logarítmica ha mejorado significativamente el ajuste del modelo** en comparación con el modelo sin transformación. La relación entre **pl\_rvamp y log(pl\_radj)** es fuerte y altamente significativa, lo que sugiere que este modelo es mucho más adecuado para describir la relación entre estas variables y ofrece estimaciones más precisas.

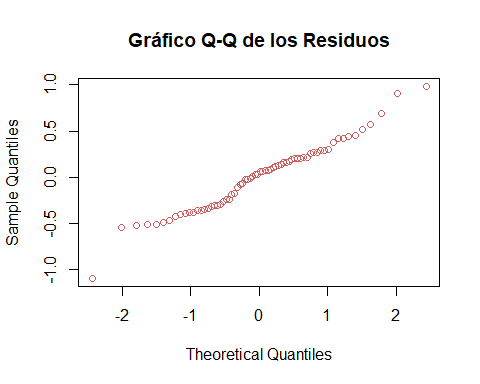
# Obtener los residuos y valores ajustados  
res\_pl\_rvamp\_log <- residuals(model\_pl\_rvamp\_log)  
  
# Calcular los valores ajustados  
fit\_val\_pl\_rvamp\_log <- fitted(model\_pl\_rvamp\_log)  
  
# Crear un data frame con los residuos y valores ajustados  
res\_pl\_rvamp\_log\_df <- data.frame(  
 Residuals = res\_pl\_rvamp\_log,  
 Fitted\_Values = fit\_val\_pl\_rvamp\_log  
)  
  
# Scatter plot de residuos vs valores ajustados  
ggplot(res\_pl\_rvamp\_log\_df, aes(x = Fitted\_Values, y = Residuals)) +  
 geom\_point(color = "steelblue") +  
 geom\_hline(yintercept = 0, color = "indianred", linetype = "dashed") +  
 labs(title = "Residuos vs Valores Ajustados",  
 x = "Valores Ajustados",  
 y = "Residuos") +  
 theme\_minimal()



Después de aplicar la transformación, se observa una **reducción en la dispersión de los residuos**, lo que indica que la **heterocedasticidad detectada previamente ha disminuido considerablemente**. Esto significa que la **varianza de los errores es ahora más constante** a lo largo de los valores ajustados, cumpliendo mejor con la suposición de **homocedasticidad** en la regresión lineal.

La disminución en la variabilidad de los residuos sugiere que el modelo ha logrado un ajuste más estable y confiable. Esto no solo mejora la **precisión de las estimaciones**, sino que también aumenta la **validez de los intervalos de confianza y las pruebas de hipótesis**, ya que las predicciones del modelo son menos sensibles a valores extremos o patrones no deseados en los errores.

# Crear un gráfico Q-Q de los residuos  
qqnorm(res\_pl\_rvamp\_log, main = "Gráfico Q-Q de los Residuos", col="indianred")



La **alineación de los puntos con la línea diagonal** en el gráfico sugiere que los residuos siguen **aproximadamente una distribución normal**. La mayor parte de los puntos se ajusta bien a la línea teórica, lo que indica que el supuesto de normalidad en los residuos se cumple en gran medida.

Sin embargo, en los **extremos del gráfico**, correspondientes a los valores más negativos y más positivos de los residuos, se observa una **ligera desviación**. Esto sugiere que la distribución de los residuos tiene **colas ligeramente más pesadas**, lo que implica la posible presencia de algunos valores atípicos o extremos.

A pesar de esta leve desviación en los valores extremos, el patrón general indica que **los residuos siguen una distribución normal de manera aceptable**. Esto significa que el modelo sigue siendo **estadísticamente válido** para realizar inferencias, aunque la presencia de valores extremos podría justificar un análisis más detallado si se requiere mayor precisión en las estimaciones.

# Prueba de normalidad de Shapiro-Wilk  
shapiro\_test\_pl\_rvamp\_log <- shapiro.test(res\_pl\_rvamp\_log)  
  
shapiro\_test\_pl\_rvamp\_log

##   
## Shapiro-Wilk normality test  
##   
## data: res\_pl\_rvamp\_log  
## W = 0.97622, p-value = 0.2188

El **test de normalidad de Shapiro-Wilk** arroja un **estadístico W de 0.97622** y un **p-valor de 0.2188**, lo que indica que no hay evidencia suficiente para rechazar la hipótesis de normalidad. Esto significa que los **residuos siguen una distribución normal**, cumpliendo con una de las principales suposiciones del modelo.

Si bien el **gráfico Q-Q** ya mostraba una buena alineación de los puntos con la línea diagonal, la prueba estadística **confirma de manera más rigurosa** que los residuos no presentan desviaciones significativas respecto a una distribución normal. Esto refuerza la validez del modelo y la fiabilidad de sus estimaciones.

# Prueba de homocedasticidad de Breusch-Pagan  
library(lmtest)  
  
# Prueba de Breusch-Pagan  
bp\_test\_pl\_rvamp\_log <- bptest(model\_pl\_rvamp\_log)  
  
bp\_test\_pl\_rvamp\_log

##   
## studentized Breusch-Pagan test  
##   
## data: model\_pl\_rvamp\_log  
## BP = 0.48865, df = 1, p-value = 0.4845

El **test de homocedasticidad de Breusch-Pagan** muestra un **estadístico BP de 0.48865** y un **p-valor de 0.4845**, lo que indica que no hay evidencia de **heterocedasticidad significativa** en el modelo. Esto significa que la **varianza de los errores es constante**, cumpliendo con una de las suposiciones clave de la regresión lineal. Dado que los modelos anteriores presentaban problemas de heterocedasticidad, estos resultados confirman que **la transformación logarítmica aplicada a la variable dependiente corrigió eficazmente este problema**.

Tras evaluar los **tres modelos de regresión lineal univariada**, se concluye que el modelo **pl\_radj ~ pl\_rvamp con transformación logarítmica** es el que ofrece el **mejor ajuste**.

Este modelo explica el **84.25% de la variabilidad en log(pl\_radj)**, lo que indica una capacidad predictiva excelente. Además, la relación entre **pl\_rvamp y log(pl\_radj)** es **fuerte y altamente significativa**, con residuos que siguen una **distribución normal** y **no presentan heterocedasticidad significativa**.

### Conclusión sobre los modelos lineales univariados planteados

A lo largo del análisis, se probaron diferentes enfoques para ajustar la relación entre las variables, considerando tanto modelos sin transformación como con transformaciones matemáticas para mejorar su estabilidad. Inicialmente, los modelos sin transformación mostraron **problemas de heterocedasticidad y normalidad en los residuos**, lo que afectaba la validez de sus inferencias. Sin embargo, la **aplicación de una transformación logarítmica a la variable dependiente (pl\_radj) permitió corregir estos problemas**, logrando un modelo más estable y confiable.

# Resumen de los modelos de regresión lineal univariada  
model\_summaries <- list(  
 model\_pl\_ratdor\_summary,  
 model\_pl\_trandep\_summary,  
 model\_pl\_trandep\_square\_summary,  
 model\_pl\_rvamp\_summary,  
 model\_pl\_rvamp\_log\_summary  
)  
  
model\_summaries

## [[1]]  
##   
## Call:  
## lm(formula = pl\_radj ~ pl\_ratdor, data = kepler\_data\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.3749 -0.2695 -0.2125 0.3126 1.2264   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.4086542 0.0535610 7.630 0.0000000000489 \*\*\*  
## pl\_ratdor 0.0007341 0.0005069 1.448 0.152   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3872 on 78 degrees of freedom  
## (2467 observations deleted due to missingness)  
## Multiple R-squared: 0.02618, Adjusted R-squared: 0.0137   
## F-statistic: 2.097 on 1 and 78 DF, p-value: 0.1516  
##   
##   
## [[2]]  
##   
## Call:  
## lm(formula = pl\_radj ~ pl\_trandep, data = kepler\_data\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.203530 -0.056704 0.002386 0.056979 0.150580   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.22308 0.01590 14.03 <0.0000000000000002 \*\*\*  
## pl\_trandep 0.61058 0.02434 25.08 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08315 on 40 degrees of freedom  
## (2505 observations deleted due to missingness)  
## Multiple R-squared: 0.9402, Adjusted R-squared: 0.9387   
## F-statistic: 629.1 on 1 and 40 DF, p-value: < 0.00000000000000022  
##   
##   
## [[3]]  
##   
## Call:  
## lm(formula = I(pl\_radj^2) ~ pl\_trandep, data = kepler\_data\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.108946 -0.015184 -0.004665 0.010102 0.134871   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.010364 0.007744 1.338 0.188   
## pl\_trandep 0.803698 0.011855 67.795 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04049 on 40 degrees of freedom  
## (2505 observations deleted due to missingness)  
## Multiple R-squared: 0.9914, Adjusted R-squared: 0.9912   
## F-statistic: 4596 on 1 and 40 DF, p-value: < 0.00000000000000022  
##   
##   
## [[4]]  
##   
## Call:  
## lm(formula = pl\_radj ~ pl\_rvamp, data = kepler\_data\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5172 -0.2119 -0.1206 0.2644 0.8116   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.337785 0.041804 8.080 0.00000000001903 \*\*\*  
## pl\_rvamp 0.003550 0.000406 8.744 0.00000000000124 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3026 on 66 degrees of freedom  
## (2479 observations deleted due to missingness)  
## Multiple R-squared: 0.5367, Adjusted R-squared: 0.5297   
## F-statistic: 76.46 on 1 and 66 DF, p-value: 0.000000000001239  
##   
##   
## [[5]]  
##   
## Call:  
## lm(formula = log(pl\_radj) ~ log(pl\_rvamp), data = kepler\_data\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.09181 -0.30754 0.04893 0.21513 0.98300   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.06303 0.06935 -29.75 <0.0000000000000002 \*\*\*  
## log(pl\_rvamp) 0.44267 0.02356 18.79 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3753 on 66 degrees of freedom  
## (2479 observations deleted due to missingness)  
## Multiple R-squared: 0.8425, Adjusted R-squared: 0.8401   
## F-statistic: 353 on 1 and 66 DF, p-value: < 0.00000000000000022

El **modelo final con transformación logarítmica** no solo ofrece un mejor ajuste y mayor capacidad explicativa, sino que también cumple con los **supuestos clave de la regresión lineal**, lo que lo convierte en la mejor opción para modelar la relación entre **pl\_radj** y **pl\_rvamp**. Este resultado subraya la importancia de evaluar la calidad del ajuste y de realizar **transformaciones adecuadas** cuando se detectan desviaciones en los supuestos del modelo, garantizando así estimaciones más precisas y robustas.

## Regresión lineal multivariada

Para mejorar la precisión en la predicción del **radio del exoplaneta (pl\_radj)**, se lleva a cabo un **análisis de regresión lineal multivariada**. En este enfoque, se consideran simultáneamente tres variables explicativas que podrían influir en el radio del exoplaneta:

1. **El cociente del radio planetario sobre el radio estelar (pl\_ratdor)**, que representa el tamaño relativo del exoplaneta en comparación con su estrella.
2. **La profundidad del tránsito (pl\_trandep)**, que mide la disminución en el brillo de la estrella cuando el exoplaneta pasa frente a ella, lo que está directamente relacionado con su tamaño.
3. **La amplitud de la velocidad radial del planeta (pl\_rvamp)**, que refleja el impacto gravitacional del exoplaneta sobre su estrella y puede proporcionar información sobre su masa y, en combinación con otras variables, su radio.

Este modelo de regresión multivariada permite analizar **cómo estas tres variables influyen en conjunto** en la predicción del radio del exoplaneta, en lugar de evaluar cada una por separado. Al considerar múltiples factores al mismo tiempo, se espera obtener un **modelo más preciso y robusto**, capaz de capturar mejor la variabilidad en los datos y proporcionar estimaciones más confiables del tamaño de los exoplanetas.

# Regresion de pl\_radj con pl\_ratdor, pl\_trandep y pl\_rvamp  
model\_multivar <- lm(pl\_radj ~ pl\_ratdor + pl\_trandep + pl\_rvamp, data = kepler\_data\_num)  
  
model\_multivar\_summary <- summary(model\_multivar)  
  
model\_multivar\_summary

##   
## Call:  
## lm(formula = pl\_radj ~ pl\_ratdor + pl\_trandep + pl\_rvamp, data = kepler\_data\_num)  
##   
## Residuals:  
## 611 1785 1822 1946 1947 2251 2465 2541   
## 0.00903 -0.07158 -0.03219 0.08415 -0.01535 0.09444 -0.08736 0.01885   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.2110174 0.0531204 3.972 0.016505 \*   
## pl\_ratdor 0.0001583 0.0004277 0.370 0.730111   
## pl\_trandep 0.7120088 0.0781077 9.116 0.000803 \*\*\*  
## pl\_rvamp -0.0008151 0.0006926 -1.177 0.304472   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08727 on 4 degrees of freedom  
## (2539 observations deleted due to missingness)  
## Multiple R-squared: 0.9731, Adjusted R-squared: 0.953   
## F-statistic: 48.28 on 3 and 4 DF, p-value: 0.001342

El modelo de regresión planteado:

Presenta buenas estadísticas por cuanto:

El modelo de regresión muestra un **excelente ajuste**, explicando el **97.31% de la variabilidad en pl\_radj**. Incluso después de ajustar por el número de predictores, el **R² ajustado sigue siendo alto, con un 95.3%**, lo que indica que el modelo generaliza bien y mantiene una fuerte capacidad predictiva. Además, el **estadístico F de 48.28** y un **p-valor de 0.0013** confirman que el modelo en su conjunto es **altamente significativo**, lo que sugiere que al menos una de las variables predictoras tiene un impacto relevante en la estimación de **pl\_radj**.

Sin embargo, existen algunas limitaciones importantes. Después de eliminar los datos con valores faltantes, **solo quedan 4 observaciones efectivas**, lo que es **extremadamente bajo** para un modelo con tres predictores. Esta reducción drástica en la cantidad de datos puede hacer que el modelo sea poco confiable y que las estimaciones de los coeficientes sean **inestables**, ya que se basan en una muestra demasiado pequeña.

Además, se han descartado **2,539 observaciones** debido a valores faltantes, lo que plantea la necesidad de revisar la estructura del dataset. Evaluar si se pueden recuperar más observaciones permitiría mejorar la estabilidad del modelo y evitar conclusiones basadas en un número tan reducido de datos.

Dado que la variable **pl\_trandep** es la más significativa dentro del modelo, una posible alternativa sería **simplificar la regresión eliminando pl\_ratdor y pl\_rvamp**, dejando solo **pl\_trandep** como predictor. Esto llevaría al segundo modelo evaluado, que podría ofrecer un ajuste igualmente bueno pero con una estructura más sencilla y posiblemente más robusta frente a la escasez de datos.

## Regresión logística

Para predecir la habitabilidad de los exoplanetas, se realiza un análisis de regresión logística utilizando las variables conocidas de los exoplanetas y las estrellas del sistema. Se seleccionan las variables conocidas de los exoplanetas y las estrellas como variables independientes y la habitabilidad como variable dependiente.

Existen varios indicadores de habitabilidad de los exoplanetas. Considerando que estamos estudiando estrellas con radios similiares al solar (+/- 10%) los parámetros planetarios más relevantes se describen a continuación.

La temperatura de equilibrio (**pl\_eqt**) planetaria es una variable fundamental en la definición de la zona habitable, ya que está directamente relacionada con la posibilidad de agua líquida en la superficie de un planeta. En (Kopparapu, R. K., Ramirez, R., Kasting, J. F., Eymet, V., Robinson, T. D., Mahadevan, S., Terrien, R. C., Domagal-Goldman, S., Meadows, V., & Deshpande, R., n.d.) se analizan factores para ajustar los límites de la zona habitable y define los rangos ideales de temperatura para habitabilidad. - La insolación (**pl\_insol**), esto es flujo de radiación recibido por el planeta en comparación con la Tierra, es otro predictor clave, ya que determina el balance energético del planeta. En (Shields, Aomawa L., Ballard, Sarah, & Johnson, John Asher, n.d.) se estudian los efectos del flujo de radiación en la zona habitable y su impacto en la capacidad de un planeta para retener agua. - El semieje mayor orbital (pl\_orbsmax) es otro indicador importante, ya que determina la distancia del planeta a su estrella anfitriona y, por lo tanto, su temperatura superficial. En (Kane, Stephen R. & Gelino, Dawn M., n.d.) se demuestra cómo la distancia orbital y la luminosidad de la estrella afectan la habitabilidad.

La determinación de la habitabilidad de un exoplaneta es un tema complejo y multifactorial que depende de varios indicadores físicos y químicos. Si bien no existe una fórmula única universalmente aceptada para calcular la “habitabilidad”, se pueden combinar los indicadores antes mencionados para aproximar el potencial de habitabilidad de un exoplaneta. Una de las aproximaciones más comunes es usar una combinación ponderada de las condiciones adecuadas para la presencia de agua líquida, ya que esta es fundamental para la vida tal como la conocemos. El uso de una combinación lineal ponderada de variables tales como las consideradas es una técnica válida en la modelización de habitabilidad planetaria. En (Méndez, Abel, Rivera-Valentin, Edgard, Schulze-Makuch, Dirk, Filiberto, Justin, Ramirez, Ramses, Wood, Tana, Dávila, Alfonso, McKay, Chris, Ceballos, Kevin, Jusino-Maldonado, Marcos, Torres-Santiago, Nicole, Gomez, Guillermo Nery, Heller, René, Byrne, Paul, Malaska, Michael, Nathan, Erica, Simões, Marta, Antunes, André, Martínez-Frías, Jesús, & Haqq-Misra, Jacob., n.d.) se propone el Earth Similarity Index (ESI), que usa una fórmula normalizada combinando temperatura, flujo estelar y parámetros orbitales.

Así con los indicadores seleccionados definimos un Índice de Habitabilidad Planetaria (PHI - Planetary Habitability Index) como una función normalizada que combine los indicadores clave:

Donde:

* son los pesos normalizados asociados a cada indicador sumando 1.
* son funciones de transformación o normalización que convierten las medidas en valores entre 0 y 1, según el rango de habitabilidad conocido.

Estas variables se consideran como posibles predictores de la habitabilidad de los exoplanetas cuando están en determinados rangos.

### Temperatura de equilibrio (pl\_eqt)

Idealmente la temperatua de equilibrio debe estar en un rango compatible con agua líquida, aproximadamente entre 0°C y 100°C. Así la normalización resultan en:

Donde y son los límites inferiores y superiores. El agua líquida, esencial para la vida, puede existir en un rango de temperaturas superficiales de aproximadamente 273K (0 °C) a 373K (100 °C). Sin embargo, debido a factores atmosféricos y de presión, este rango puede extenderse ligeramente: puede ser el umbral inferior, considerando atmósferas densas como la de Marte y puede considerarse el umbral superior, considerando atmósferas con alta presión como la de Venus.

Según la literatura especializada el peso de la temperatura de equilibrio en el índice de habitabilidad es de cuando se considera la temperatura de equilibrio como un indicador preponderante de habitabilidad y no se tienen muchos más datos del planeta.

### La insolación (pl\_insol)

La insolación es la cantidad de energía recibida por el planeta y es un indicador importante de la habitabilidad. La insolación se mide en unidades de flujo de energía recibida por unidad de área y se puede normalizar en un rango de 0 a 1. La insolación ideal para la vida es aquella que permite la presencia de agua líquida en la superficie del planeta (generalmente valores cercanos a la Tierra son ideales). La insolación se puede normalizar considerando que La Tierra tiene una insolación de 1S (donde es la insolación solar a la distancia de la Tierra) y se puede considerar que la vida puede existir en un rango de insolación de (aproximadamente el límite interior de la zona habitable para una estrella como el Sol, más allá de este valor el agua podría congelarse) a (aproximadamente el límite exterior de la zona habitable, más allá de este valor el agua podría evaporarse). Así la normalización resulta en:

La insolación es un factor crítico para mantener agua líquida en la superficie del planeta, pero su impacto depende de otros factores como la atmósfera y el albedo, según la literatura especializada el peso de la insolación en el índice de habitabilidad es de . Es importante destacar que El rango de e puede variar según la luminosidad de la estrella, ya que la zona habitable cambia en función del tipo estelar por lo que esta estimación sirve para estrellas similares al Sol.

### La distancia orbital del semieje mayor (pl\_orbsmax)

La distancia orbital semieje mayor es un indicador de la zona habitable de un sistema planetario. La distancia orbital semieje mayor se mide en unidades astronómicas (UA) y se puede normalizar en un rango de 0 a 1. La distancia orbital semieje mayor ideal para la vida es aquella que permite la presencia de agua líquida en la superficie del planeta. La distancia orbital semieje mayor se puede normalizar considerando que La Tierra tiene una distancia orbital semieje mayor de 1 UA y se puede considerar que la vida puede existir en un rango de distancias de UA (aproximadamente el límite interior de la zona habitable para una estrella como el Sol, más allá de este valor el agua podría congelarse) a UA (aproximadamente el límite exterior de la zona habitable, más allá de este valor el agua podría evaporarse). Estos rangos pueden variar según el tipo de estrella. Así la normalización resulta en:

En la mayoría de los modelos, el peso para el valor normalizado de **pl\_orbsmax** oscila entre 15% y 25%. Esto se debe a que **pl\_orbsmax** complementa indicadores más directos como **pl\_eqt** (temperatura de equilibrio) y **pl\_insol** (insolación). Se tomará un peso de .

### Índice de habitabilidad planetaria (PHI)

Por lo expuesto anteriormente, el **Índice de Habitabilidad Planetaria (PHI)** se calcula utilizando **variables normalizadas** y sus respectivos **pesos**, lo que permite evaluar la habitabilidad de un exoplaneta de manera más equilibrada y comparable entre diferentes sistemas.

La fórmula para calcular el **PHI** que se utilizará en el modelo de regresión logística es:

library(dplyr)  
  
# Crear un data frame con las variables seleccionadas  
kepler\_data\_logistic <- kepler\_data %>%  
 dplyr::select(pl\_eqt, pl\_insol, pl\_orbsmax)  
  
# Normalizar las variables  
kepler\_data\_logistic\_normalized <- kepler\_data\_logistic %>%  
 mutate(  
 pl\_eqt\_normalized = pmin(1, pmax(0, (pl\_eqt - 200) / (400 - 200))),  
 pl\_insol\_normalized = pmin(1, pmax(0, (pl\_insol - 0.3) / (1.7 - 0.3))),  
 pl\_orbsmax\_normalized = pmin(1, pmax(0, (pl\_orbsmax - 0.05) / (2 - 1.05)))  
 )  
  
# Calcular el índice de habitabilidad planetaria (PHI)  
kepler\_data\_logistic\_normalized <- kepler\_data\_logistic\_normalized %>%  
 mutate(  
 phi = 0.4 \* pl\_eqt\_normalized + 0.3 \* pl\_insol\_normalized + 0.3 \* pl\_orbsmax\_normalized   
 )  
  
# Remover columnas no normalizadas  
kepler\_data\_logistic\_normalized <- subset(kepler\_data\_logistic\_normalized, select = -c(pl\_eqt, pl\_insol, pl\_orbsmax))

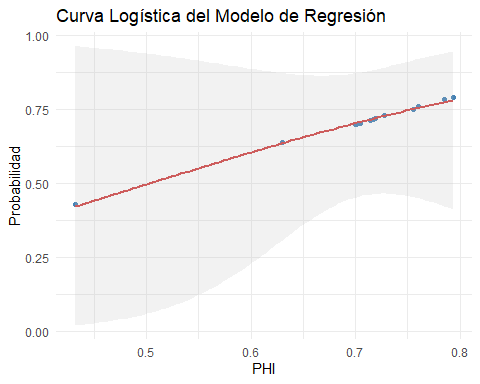
# Crear un modelo de regresión logística  
model\_logistic <- glm(phi ~ ., data = kepler\_data\_logistic\_normalized, family = binomial)

## Warning in eval(family$initialize): non-integer #successes in a binomial glm!

model\_logistic\_summary <- summary(model\_logistic)  
  
model\_logistic\_summary

##   
## Call:  
## glm(formula = phi ~ ., family = binomial, data = kepler\_data\_logistic\_normalized)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -2.488 7.878 -0.316 0.752  
## pl\_eqt\_normalized 2.228 11.813 0.189 0.850  
## pl\_insol\_normalized 1.105 6.760 0.163 0.870  
## pl\_orbsmax\_normalized 1.569 6.123 0.256 0.798  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 0.45354364 on 16 degrees of freedom  
## Residual deviance: 0.00063461 on 13 degrees of freedom  
## (2530 observations deleted due to missingness)  
## AIC: 19.641  
##   
## Number of Fisher Scoring iterations: 4

## `geom\_smooth()` using formula = 'y ~ x'



El **modelo de regresión logística** obtenido no es particularmente sólido debido a la **escasez de datos disponibles** para calcular el **PHI (Planetary Habitability Index)**. En este caso, la cantidad de **exoplanetas con información completa** sobre los factores clave para la predicción, como **temperatura de equilibrio, insolación y semieje mayor orbital**, es **muy limitada**, con solo **17 observaciones** disponibles.

La baja cantidad de datos afecta significativamente la **precisión y confiabilidad** del modelo, ya que un tamaño de muestra tan reducido **limita la capacidad de generalización** y hace que los coeficientes del modelo sean **menos estables**. Para obtener un modelo más robusto y representativo, sería necesario contar con **un mayor número de exoplanetas con datos completos**, lo que permitiría entrenar la regresión con más información y mejorar la validez de las predicciones sobre habitabilidad.

El **modelo de regresión logística** planteado presenta varias limitaciones que afectan su utilidad y confiabilidad.

Uno de los principales problemas es la **drástica reducción en la deviance residual**, que pasó de **0.4535** (sin predictores) a **0.0006** (con predictores). Si bien una reducción en la deviance suele indicar una mejora en el modelo, en este caso el valor **extremadamente bajo** sugiere un **sobreajuste**, lo que significa que el modelo está ajustándose demasiado a los pocos datos disponibles y podría no generalizar bien a nuevas observaciones.

El **criterio de información de Akaike (AIC)**, con un valor de **19.641**, indica que el modelo penaliza adecuadamente la complejidad, pero la **falta de significancia en los coeficientes** sugiere que **no es útil para la predicción**, ya que no logra establecer una relación clara entre las variables.

Otro problema grave es la **eliminación de datos faltantes**, que dejó **solo 17 observaciones disponibles** después de descartar **2,530 registros incompletos**. Con tan pocos datos, los **grados de libertad (df=13) son extremadamente bajos**, lo que **afecta la estabilidad del modelo** y hace que las estimaciones sean poco confiables.

En su estado actual, el modelo no es útil, ya que **no cuenta con variables predictoras significativas** y enfrenta **una pérdida masiva de datos**, lo que limita su capacidad para hacer inferencias válidas. Antes de sacar conclusiones sobre la relación entre **PHI** y las variables predictoras, sería recomendable **limpiar los datos y probar un modelo más simple**. Sin embargo, dada la **escasez de observaciones**, es posible que no sea factible construir un modelo confiable para la **predicción de habitabilidad** en estas condiciones.

# Posibles mejoras para futuros análisis

Para mejorar el **modelo de regresión logística**, una posible solución es abordar el problema de los **valores faltantes** en las variables clave utilizadas para calcular el **PHI**: **temperatura de equilibrio (pl\_eqt), insolación (pl\_insol) y semieje mayor orbital (pl\_orbsmax)**. En lugar de descartar todas las observaciones con datos faltantes, se puede aplicar **regresión lineal** para estimar estos valores en función de otras variables relacionadas dentro del conjunto de datos. Al hacer esto, se logra recuperar una mayor cantidad de observaciones, lo que aumentaría el tamaño de la muestra utilizada en la regresión logística y mejoraría la estabilidad del modelo.

Sin embargo, utilizar variables estimadas mediante regresión como **input en otro modelo predictivo** conlleva ciertos **riesgos e implicaciones estadísticas**. Cuando una variable ha sido generada mediante una predicción previa, **su error de estimación se transfiere al modelo final**, en este caso, a la regresión logística. Como resultado, el modelo puede **subestimar la incertidumbre real** y producir **intervalos de confianza demasiado optimistas**, ya que no está considerando los errores asociados a las predicciones previas.

Además, el uso de **valores regresionados** puede introducir **colinealidad artificial**, especialmente si la variable imputada depende de otras variables que ya están incluidas en la regresión logística. Esto puede distorsionar los coeficientes y hacer que el modelo sea menos interpretable. También es posible que la **estructura de los errores en las variables imputadas** no sea idéntica a la de los valores observados originalmente, lo que podría generar un sesgo en las estimaciones del modelo.

Para mitigar estos riesgos, se pueden aplicar estrategias como la **validación cruzada** para evaluar el impacto de las variables imputadas en el desempeño del modelo y realizar un **análisis de sensibilidad**, probando el modelo con y sin las variables imputadas para verificar la estabilidad de los resultados. Además, una opción más robusta sería utilizar **métodos de imputación múltiple**, que generan varias versiones de los datos imputados y permiten incorporar la incertidumbre en las predicciones finales.

En conclusión, aunque la regresión lineal de las variables faltantes puede ayudar a recuperar observaciones y mejorar la estabilidad del modelo logístico, es importante ser consciente de los **errores de predicción acumulados** y del impacto que pueden tener en la validez de las estimaciones. Evaluar cuidadosamente estos aspectos garantizará que el modelo final sea más confiable y que sus conclusiones sean interpretadas con la debida cautela.

# Conclusiones

El análisis realizado permitió evaluar diferentes modelos de regresión para predecir el radio de los exoplanetas y explorar la relación entre sus características físicas. Se probaron modelos de regresión lineal y logística, incluyendo transformaciones para mejorar la normalidad de los residuos y corregir problemas de heterocedasticidad. Los resultados indicaron que el modelo **pl\_radj ~ pl\_rvamp con transformación logarítmica** fue el más efectivo, explicando el **84.25% de la variabilidad** en el radio del exoplaneta y cumpliendo con los principales supuestos estadísticos. Sin embargo, el modelo de **regresión logística** para predecir la habitabilidad planetaria presentó limitaciones debido a la escasez de datos completos, lo que afectó su estabilidad y capacidad predictiva.

Para mejorar la calidad de los modelos, es fundamental abordar el problema de los valores faltantes, ya sea mediante **imputación de datos** o la recolección de más observaciones. Además, al utilizar variables estimadas en modelos posteriores, se debe considerar el impacto de los **errores de predicción acumulados** en los resultados finales. Si bien las transformaciones de datos ayudaron a mejorar la precisión de algunos modelos, la falta de datos sigue siendo una limitación clave para la predicción de la **habitabilidad planetaria**. Futuras investigaciones deberían centrarse en ampliar el conjunto de datos y explorar enfoques más robustos, como modelos de aprendizaje automático, que podrían mejorar la capacidad predictiva de estos análisis.

# Referencias

Alonso Sobrino, Roi. n.d. “Detección y Caracterización de Exoplanetas Mediante El Método de Los Tránsitos.” Instituto de Astrofísica de Canarias (IAC). <https://www.iac.es/es/ciencia-y-tecnologia/publicaciones/deteccion-y-caracterizacion-de-exoplanetas-mediante-el-metodo-de-los-transitos>.

Brachman, R. Z. n.d. “How to Analyze Your Data | How to Get Started – Exoplanet Exploration: Planets Beyond Our Solar System.” Exoplanet Exploration: Planets Beyond Our Solar System. <https://exoplanets.nasa.gov/exoplanet-watch/how-to-contribute/how-to-analyze-your-data/>.

Cermak, A. & Cermak, A. n.d. “Kepler / K2 - NASA Science.” NASA Science. <https://science.nasa.gov/mission/kepler/>.

Georgsmeinung. n.d. “GitHub - Georgsmeinung/Ra-Tp-Final.” GitHub. <https://github.com/georgsmeinung/ra-tp-final/>.

IPAC. n.d.a. “NASA Exoplanet Archive.” NASA Exoplanet Science Institute. <https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS>.

———. n.d.b. “NASA Exoplanet Archive Overview and Holdings.” <https://exoplanetarchive.ipac.caltech.edu/docs/intro.html>.

———. n.d.c. “Our Mission.” <https://www.ipac.caltech.edu/page/mission>.

———. n.d.d. “Planetary Systems and Planetary Systems Composite Parameters Data Column Definitions.” <https://exoplanetarchive.ipac.caltech.edu/docs/API_PS_columns.html>.

Kane, Stephen R. & Gelino, Dawn M. n.d. “The Habitable Zone and Extreme Planetary Orbits.” <https://doi.org/10.1089/ast.2011.0798>.

Kopparapu, R. K., Ramirez, R., Kasting, J. F., Eymet, V., Robinson, T. D., Mahadevan, S., Terrien, R. C., Domagal-Goldman, S., Meadows, V., & Deshpande, R. n.d. “Habitable Zones Around Main-Sequence Stars: New Estimates.” <https://doi.org/10.1088/0004-637x/765/2/131>.

Marín, Daniel. n.d.a. “Detectando Planetas Desde El Espacio Gracias a Einstein - Eureka.” Eureka. <https://danielmarin.naukas.com/2011/05/23/detectando-planetas-desde-el-espacio-gracias-a-einstein/>.

———. n.d.b. “EarthFinder: Un Telescopio Espacial Para Buscar Exotierras Por El Método de La Velocidad Radial - Eureka.” <https://danielmarin.naukas.com/2018/03/14/earthfinder/>.

Méndez, Abel, Rivera-Valentin, Edgard, Schulze-Makuch, Dirk, Filiberto, Justin, Ramirez, Ramses, Wood, Tana, Dávila, Alfonso, McKay, Chris, Ceballos, Kevin, Jusino-Maldonado, Marcos, Torres-Santiago, Nicole, Gomez, Guillermo Nery, Heller, René, Byrne, Paul, Malaska, Michael, Nathan, Erica, Simões, Marta, Antunes, André, Martínez-Frías, Jesús, & Haqq-Misra, Jacob. n.d. “Habitability Models for Astrobiology.” <https://doi.org/10.1089/ast.2020.2342>.

Nardi, Luca. n.d. “¿Cuáles Son Las Técnicas Utilizadas Para Descubrir Exoplanetas?” WIRED. <https://es.wired.com/articulos/cuales-son-las-tecnicas-utilizadas-para-descubrir-exoplanetas>.

NASA, Equipo de redacción de Ciencia. n.d. “Webb de La NASA Obtiene Su Primera Imagen Directa de Un Mundo Distante.” NASA Ciencia. <https://ciencia.nasa.gov/universo/webb-de-la-nasa-obtiene-su-primera-imagen-directa-de-un-mundo-distante/>.

NeoFronteras. n.d. “Detección de Exoplanetas Por Astrometría.” NeoFronteras. <https://neofronteras.com/?p=8093>.

Shields, Aomawa L., Ballard, Sarah, & Johnson, John Asher. n.d. “The Habitability of Planets Orbiting M-Dwarf Stars.” <https://doi.org/10.1016/j.physrep.2016.10.003>.

Wolszczan, Aleksander & Frail, D. A. n.d. “A Planetary System Around the Millisecond Pulsar PSR1257 + 12.” <https://doi.org/10.1038/355145a0>.