STUDY OF EXOPLANETS

2024-03-15

Abstract

Our study conducts a data analysis on around 10,000 exoplanet candidates from NASA's Kepler mission, focusing on their classification, habitability, and the impact of observational and stellar parameters. We explored three main questions: the influence of observational parameters on classification, the habitability of Earth-size planets around different stars, and the correlations between disposition values and flag variables. Using models like GLMM (96.98% accuracy), Random Forest, and KNN, we identified key predictors for planet classification including planetary radius and stellar effective temperature. Our findings reveal only a small fraction of exoplanets could potentially harbor liquid water, with habitability closely linked to star metallicity and planet composition. Additionally, we found significant correlations between disposition scores, commentary, and flag variables, highlighting the challenge of false positives in exoplanet detection. This research enhances our understanding of exoplanetary systems and guides future astronomical endeavors.[]

Keywords Exoplanets, generalized linear/additive mixed models, parametric modeling, non-parametric modeling.

Introduction

The Kepler Space Telescope, a pioneering mission initiated by NASA in 2009, has significantly advanced our understanding of exoplanets orbiting stars beyond the Solar System. Its primary mission was to survey a portion of our galaxy to discover Earth-size and smaller planets in or near the habitable zone and determine how many of the billions of stars in the Milky Way have such planets. By May 2016, Kepler's observations and analyses had led to the verification of 1,284 new exoplanets, marking a significant contribution to the broader field of astrophysics and expanding our knowledge of planetary systems. As of October 2017, the cumulative tally of confirmed exoplanets exceeded 3,000, illustrating the vastness of our galaxy's planetary diversity and the effectiveness of Kepler's observational capabilities (NASA, 2018).

The dataset central to our research represents a thorough aggregation of all "objects of interest" identified by the Kepler mission, which includes approximately 10,000 exoplanet candidates. These candidates are subjects of interest primarily because of the characteristics they share with known exoplanets, including temperature, metallicity, etc. The expansive dataset not only offers insights into the variety and distribution of exoplanets but also serves as a crucial resource for statistical analysis and hypothesis testing within the astrophysical community.

Our research aims to delve into the intricate relationships between the classification of these exoplanets, based on stellar parameters and their inherent properties. By examining these relationships, we intend to uncover the underlying reasons for the classifications of exoplanets. Through this analysis, we aspire to contribute to the broader scientific dialogue on exoplanetary science, offering insights that could guide future missions and observational strategies to understand the cosmos.

Data Description

The Kepler Exoplanet Dataset, a robust compilation of astrophysical data, encompasses observations of 9,564 celestial bodies suspected to be exoplanets. It features a rich array of 50 distinct attributes for each candidate, which includes identifiers like the Kepler ID (KepID) and Kepler Object of Interest Name (KOI

Name). Additionally, the dataset provides a wealth of stellar characteristics, such as effective temperature, along with transit properties such as duration and depth that are crucial for understanding these distant worlds. The central focus of our research is the Exoplanet Archive Disposition (labeled as koi_disposition in the dataset), which classifies each observed object in terms of its candidacy as an exoplanet, based on a set of criteria that includes observational data and validated models. This variable is pivotal for distinguishing between confirmed exoplanets, false positives, and other categories, thereby facilitating a structured approach to exoplanetary studies.

Our research involved meticulous data cleaning to ensure the integrity and usability of the Kepler Exoplanet Dataset. This process included filtering out rows with 'CANDIDATE' status in the koi_disposition variable, refining this variable into a categorical factor with distinct levels for 'CONFIRMED' and 'FALSE POSITIVE' statuses, and eliminating incomplete records. We also streamlined the dataset by removing columns not pertinent to our analysis objectives, such as kepoi_name, koi_comment, koi_vet_stat, and koi_pdisposition. Also, the removal of the koi_score column was the final step in preparing the dataset for in-depth exploration and analysis.

Importing Libraries

```
library(readxl)
library(lattice)
library(data.table)
library(dplyr)
```

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

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

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

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

```
library(robustbase)
library(robust)
```

```
## Loading required package: fit.models
```

```
library(lattice)
library(data.table)
library(dplyr)
library(robustbase)
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-8
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
library(nlme)
```

```
##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
library(mgcv)
```

```
## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(lme4)
```

```
##
## Attaching package: 'lme4'

## The following object is masked from 'package:nlme':
##
##     lmList
```

```
library(glmmTMB)
library(DHARMa)
```

```
## This is DHARMa 0.4.6. For overview type '?DHARMa'. For recent changes, type news(package = 'DHARMa')
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(cluster)
library(randomForest)
```

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(mltools)
library(data.table)
library(ggplot2)
library(class)
library(knitr)
```

DATA CLEANING

```
project_data<-read.csv('main_table.csv')
summary(project_data)
```

```
##    loc_rowid        kepid           kepoi_name        kepler_name
##  Min.   :   1   Min.   :  757450   Length:9564        Length:9564
##  1st Qu.:2392   1st Qu.: 5556034   Class :character   Class :character
##  Median :4782   Median : 7906892   Mode  :character   Mode  :character
##  Mean   :4782   Mean   : 7690628
##  3rd Qu.:7173   3rd Qu.: 9873066
##  Max.   :9564   Max.   :12935144
##
##  koi_disposition    koi_pdisposition     koi_score      koi_fpflag_nt
##  Length:9564        Length:9564        Min.   :0.0000   Min.   :  0.0000
##  Class :character   Class :character   1st Qu.:0.0000   1st Qu.:  0.0000
##  Mode  :character   Mode  :character   Median :0.3340   Median :  0.0000
##                                        Mean   :0.4808   Mean   :  0.2086
##                                        3rd Qu.:0.9980   3rd Qu.:  0.0000
##                                        Max.   :1.0000   Max.   :465.0000
##                                        NA's   :1510
##  koi_fpflag_ss    koi_fpflag_co    koi_fpflag_ec    koi_period
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00   Min.   :     0.24
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:     2.73
##  Median :0.0000   Median :0.0000   Median :0.00   Median :     9.75
##  Mean   :0.2327   Mean   :0.1975   Mean   :0.12   Mean   :    75.67
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00   3rd Qu.:    40.72
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00   Max.   :129995.78
##
##   koi_time0bk       koi_impact        koi_duration       koi_depth
##  Min.   : 120.5   Min.   :  0.0000   Min.   :  0.052   Min.   :      0
##  1st Qu.: 132.8   1st Qu.:  0.1970   1st Qu.:  2.438   1st Qu.:    160
##  Median : 137.2   Median :  0.5370   Median :  3.793   Median :    421
##  Mean   : 166.2   Mean   :  0.7351   Mean   :  5.622   Mean   :  23792
##  3rd Qu.: 170.7   3rd Qu.:  0.8890   3rd Qu.:  6.277   3rd Qu.:   1470
##  Max.   :1472.5   Max.   :100.8060   Max.   :138.540   Max.   :1540000
##                   NA's   :363                          NA's   :363
##     koi_prad            koi_sma           koi_teq        koi_insol
##  Min.   :     0.08   Min.   : 0.0059   Min.   :   25   Min.   :       0
##  1st Qu.:     1.40   1st Qu.: 0.0377   1st Qu.:  539   1st Qu.:      20
##  Median :     2.39   Median : 0.0851   Median :  878   Median :     142
##  Mean   :   102.89   Mean   : 0.2240   Mean   : 1085   Mean   :    7746
##  3rd Qu.:    14.93   3rd Qu.: 0.2144   3rd Qu.: 1379   3rd Qu.:     870
##  Max.   :200346.00   Max.   :44.9892   Max.   :14667   Max.   :10947555
##  NA's   :363         NA's   :363       NA's   :363     NA's   :321
##  koi_model_snr    koi_tce_plnt_num koi_tce_delivname    koi_steff
##  Min.   :   0.0   Min.   :1.000    Length:9564        Min.   : 2661
##  1st Qu.:  12.0   1st Qu.:1.000    Class :character   1st Qu.: 5310
##  Median :  23.0   Median :1.000    Mode  :character   Median : 5767
##  Mean   : 259.9   Mean   :1.244                       Mean   : 5707
##  3rd Qu.:  78.0   3rd Qu.:1.000                       3rd Qu.: 6112
##  Max.   :9054.7   Max.   :8.000                       Max.   :15896
##  NA's   :363      NA's   :346                         NA's   :363
##    koi_slogg        koi_smet          koi_srad         koi_smass
```

```
##  Min.   :0.047   Min.   :-2.5000   Min.   :  0.109   Min.   :0.000
##  1st Qu.:4.218   1st Qu.:-0.2600   1st Qu.:  0.829   1st Qu.:0.845
##  Median :4.438   Median :-0.1000   Median :  1.000   Median :0.974
##  Mean   :4.310   Mean   :-0.1244   Mean   :  1.729   Mean   :1.024
##  3rd Qu.:4.543   3rd Qu.: 0.0700   3rd Qu.:  1.345   3rd Qu.:1.101
##  Max.   :5.364   Max.   : 0.5600   Max.   :229.908   Max.   :3.735
##  NA's   :363     NA's   :386       NA's   :363       NA's   :363
##        ra              dec          koi_kepmag     koi_comment
##  Min.   :279.9   Min.   :36.58   Min.   : 6.966   Length:9564
##  1st Qu.:288.7   1st Qu.:40.78   1st Qu.:13.440   Class :character
##  Median :292.3   Median :43.68   Median :14.520   Mode  :character
##  Mean   :292.1   Mean   :43.81   Mean   :14.265
##  3rd Qu.:295.9   3rd Qu.:46.71   3rd Qu.:15.322
##  Max.   :301.7   Max.   :52.34   Max.   :20.003
##                                  NA's   :1
```

```
data<-project_data

# DATA CLEANING
data <- data[data$koi_disposition != "CANDIDATE",]

data$koi_disposition <- factor(data$koi_disposition, levels=c("CONFIRMED","FALSE POSITIVE"))
# class(data$koi_disposition)

complete_rows <- complete.cases(data)
data <- data[complete_rows, ]


##---- Remove Unwanted columns
# Convert character columns to factors for categorical variables

# Remove unnecessary character columns
data$koi_vet_stat <- NULL
data$koi_pdisposition <- NULL
data$loc_rowid <- NULL
data$kepoi_name <-NULL
data$koi_tce_plnt_num <- NULL
data$koi_tce_delivname <- NULL
data$koi_score <- NULL
data$kepler_name<- NULL


# Removing koi_score form data
data <- data[, !names(data) %in% "koi_score"]
```

CLEANED DATA

```
head(data,n=10)
```

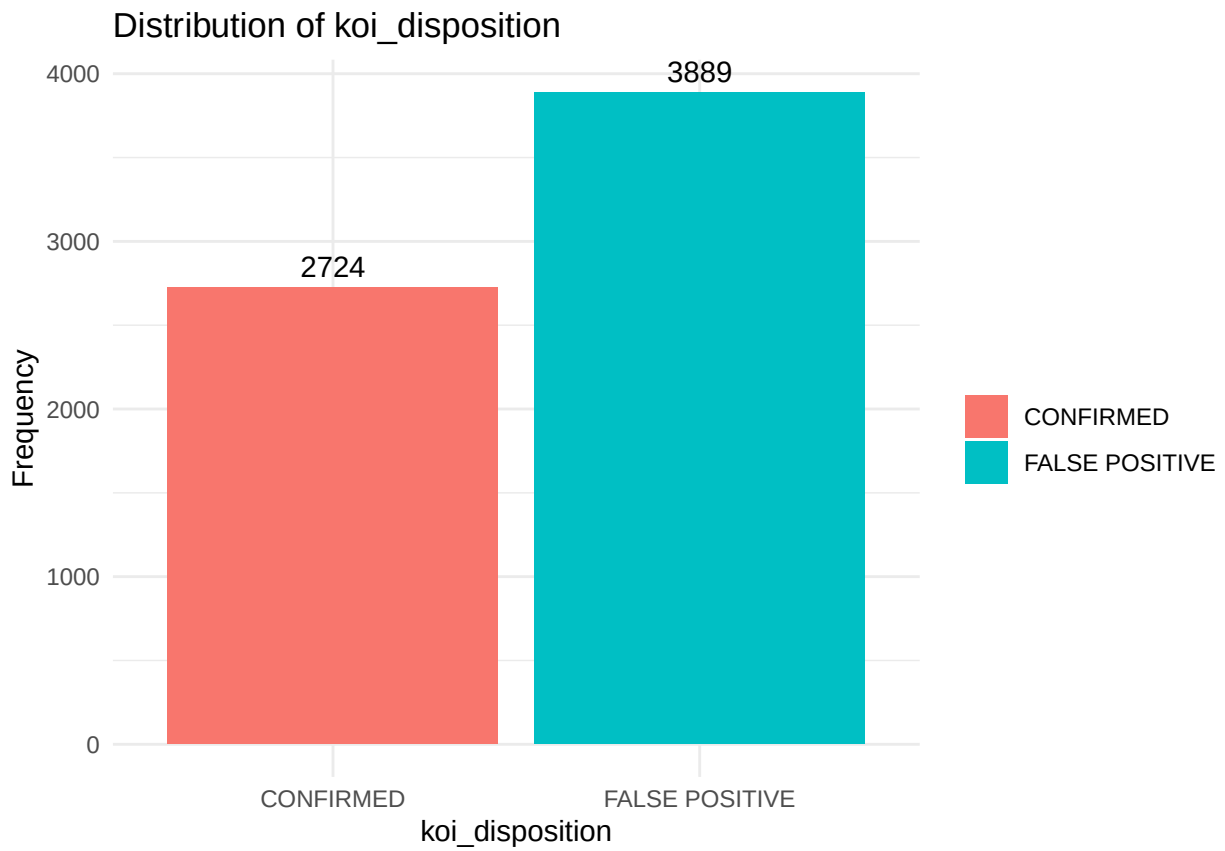
```
##       kepid koi_disposition koi_fpflag_nt koi_fpflag_ss koi_fpflag_co
## 1  10797460       CONFIRMED             0             0             0
## 2  10797460       CONFIRMED             0             0             0
## 4  10848459  FALSE POSITIVE             0             1             0
## 5  10854555       CONFIRMED             0             0             0
```

```
## 6  10872983       CONFIRMED            0            0            0
## 7  10872983       CONFIRMED            0            0            0
## 8  10872983       CONFIRMED            0            0            0
## 9   6721123  FALSE POSITIVE            0            1            1
## 10 10910878       CONFIRMED            0            0            0
## 11 11446443       CONFIRMED            0            0            0
##    koi_fpflag_ec koi_period koi_time0bk koi_impact koi_duration koi_depth
## 1              0   9.488036    170.5387      0.146      2.95750       616
## 2              0  54.418383    162.5138      0.586      4.50700       875
## 4              0   1.736952    170.3076      1.276      2.40641      8080
## 5              0   2.525592    171.5956      0.701      1.65450       603
## 6              0  11.094321    171.2012      0.538      4.59450      1520
## 7              0   4.134435    172.9794      0.762      3.14020       686
## 8              0   2.566589    179.5544      0.755      2.42900       227
## 9              0   7.361790    132.2505      1.169      5.02200       234
## 10             0  16.068647    173.6219      0.052      3.53470      4910
## 11             0   2.470613    122.7633      0.818      1.74319     14200
##    koi_prad koi_sma koi_teq koi_insol koi_model_snr koi_steff koi_slogg
## 1      2.26  0.0853     793     93.59          35.8      5455     4.467
## 2      2.83  0.2734     443      9.11          25.8      5455     4.467
## 4     33.46  0.0267    1395    891.96         505.6      5805     4.564
## 5      2.75  0.0374    1406    926.16          40.9      6031     4.438
## 6      3.90  0.0992     835    114.81          66.5      6046     4.486
## 7      2.77  0.0514    1160    427.65          40.2      6046     4.486
## 8      1.59  0.0374    1360    807.74          15.0      6046     4.486
## 9     39.21  0.0820    1342    767.22          47.7      6227     3.986
## 10     5.76  0.1158     600     30.75         161.9      5031     4.485
## 11    13.04  0.0354    1339    761.46        4304.3      5820     4.457
##    koi_smet koi_srad koi_smass       ra      dec koi_kepmag    koi_comment
## 1      0.14    0.927     0.919 291.9342 48.14165     15.347     NO_COMMENT
## 2      0.14    0.927     0.919 291.9342 48.14165     15.347     NO_COMMENT
## 4     -0.52    0.791     0.836 285.5346 48.28521     15.597 MOD_ODDEVEN_DV
## 5      0.07    1.046     1.095 288.7549 48.22620     15.509     NO_COMMENT
## 6     -0.08    0.972     1.053 296.2861 48.22467     15.714     NO_COMMENT
## 7     -0.08    0.972     1.053 296.2861 48.22467     15.714     NO_COMMENT
## 8     -0.08    0.972     1.053 296.2861 48.22467     15.714     NO_COMMENT
## 9      0.00    1.958     1.358 298.8644 42.15157     12.660     MOD_SEC_DV
## 10     0.16    0.848     0.801 286.9995 48.37579     15.841     NO_COMMENT
## 11    -0.06    0.964     0.971 286.8085 49.31640     11.338 CENT_SATURATED
```

Following the data cleaning phase, we crafted several visualizations to illuminate the characteristics of the dataset.

Distribution for the response variable(koi_disposition).

```r
response_variable <- "koi_disposition"
numeric_data <- data[, sapply(data, is.numeric)]
explanatory_variables <- numeric_data[, !colnames(numeric_data) %in% response_variable]

# Distribution of the response variable
counts <- as.data.frame(table(data$koi_disposition))
ggplot(counts, aes(x = Var1, y = Freq, fill = Var1)) +   geom_bar(stat = "identity", position = "dodge")
```



From the plot, it can be observed that there are 2724 confirmed cases and 3894 false-positive cases. To have a thorough understanding of the distribution, it is neccessary to consider the potential of zero-truncated and zero-inflated distribution. In the realm of probability theory, the Zero-Truncated Poisson (ZTP) distribution is a specific discrete probability distribution limited to positive integers. It represents the conditional probability distribution of a variable that is Poisson-distributed, under the condition that the variable's value is nonzero. Therefore, a ZTP-distributed variable cannot take the value of zero. This distribution is initiated with the premise that it excludes the zero value, focusing solely on the positive integers as its domain ("Zero-truncated Poisson distribution", 2024). The Zero-inflated distribution is when an observed dataset shows a higher frequency of zero counts than what is typically expected under traditional models like the Poisson or negative binomial distributions. This phenomenon, characterized by the surplus of zeros, is referred to as being Zero-inflated ("Zero-inflated model", 2023).

Since our response variable, koi_disposition, is categorical, so it does not have zero-truncated, zero-inflated or other specific distributions.

Next, we explored the relationships between the response variables and explanatory variables through boxplots.

The loc_rowid boxplot illustrates the relationship between koi_disposition and loc_rowid, showing the spread of row IDs for confirmed and false-positive dispositions, indicating a numerical identification of observations rather than a meaningful quantitative relationship.

The kepid plot relates koi_disposition to kepid, depicting the distribution of Kepler IDs across confirmed and false-positive categories. Like loc_rowid, kepid serves as an identifier and does not represent a quantitative relationship.

Then, the following plots explore koi_disposition against flags koi_fpflag_nt, koi_fpflag_ss, koi_fpflag_co, and koi_fpflag_ec. These flags are binary indicators, where a value of 1 typically represents a specific condition being true for a given exoplanet candidate. The plots highlight the frequency of each flag within

the confirmed and false-positive categories, showing a stark contrast in koi_fpflag_ss, where the flag is predominantly set for false positives, indicating a significant secondary eclipse event.

The koi_period plot connects koi_disposition with koi_period, showcasing the orbital period of the exoplanets. It's apparent that confirmed exoplanets have approximate same range of orbital periods as false positives group, and several outliers in both groups indicating unusually long periods.

The koi_time0bk plot is the relationship between koi_disposition and koi_time0bk shows the distribution of the time of the first transit in front of the host star, again comparing confirmed to false-positive exoplanet candidates. There's a wider distribution in the confirmed category, suggesting a broader range of transit times among confirmed exoplanets compared to false positives.

For the koi_depth plot, we see that confirmed exoplanets typically show a lower range of transit depth values, with most clustering near the bottom of the plot, which may suggest smaller planetary sizes or less obstructive transits compared to those flagged as false positives.

The koi_prad plot exhibits the planetary radius with confirmed exoplanets generally having smaller radii, as shown by the compact spread of the box, whereas false positives have a wider spread, indicating a possible misclassification due to larger body sizes or observational errors.

In the koi_teq plot, representing the equilibrium temperature, confirmed exoplanets have a broader interquartile range, suggesting a diverse set of thermal environments. False positives, while showing outliers with very high equilibrium temperatures, mainly cluster in a tighter interquartile range.

For the koi_insol variable, which shows insolation flux, there is a notable difference between confirmed exoplanets, which have lower insolation fluxes, and false positives that show a higher range of insolation values.

The koi_model_snr plot, depicting the signal-to-noise ratio of the transit signal, highlights a distinct pattern with confirmed exoplanets typically presenting lower SNR values compared to false positives, possibly indicating that confirmed signals are more subtle and thus harder to detect.

Finally, the koi_tce_plnt_num and koi_steff plots (stellar effective temperature), and the koi_slogg and koi_srad plots (stellar surface gravity and radius, respectively) together exhibit variations in stellar characteristics associated with confirmed and false-positive exoplanet designations. Confirmed exoplanets are associated with stars that have a narrower range of effective temperatures, surface gravity, and radii compared to the broader ranges seen in false positives. These differences may reflect the varying conditions under which exoplanets are more likely to be accurately detected and confirmed.

Building on our preliminary examination using boxplots to discern relationships between the response variable, koi_disposition, and various explanatory variables, we delved deeper into the interdependencies within explanatory variables via a correlation matrix. This step is critical for uncovering the intricate associations that can inform the development of robust predictive models. Through this matrix, we are able to pinpoint not only isolated pairs of variables with strong correlations but also broader patterns that might influence multiple variables simultaneously. Understanding these connections allows us to better prepare our data for modeling, ensuring that we account for these relationships in our analyses and improve model accuracy.

```
# Correlation

numeric_data <- data[, sapply(data, is.numeric)]
correlation_matrix <- cor(numeric_data); correlation_matrix
```

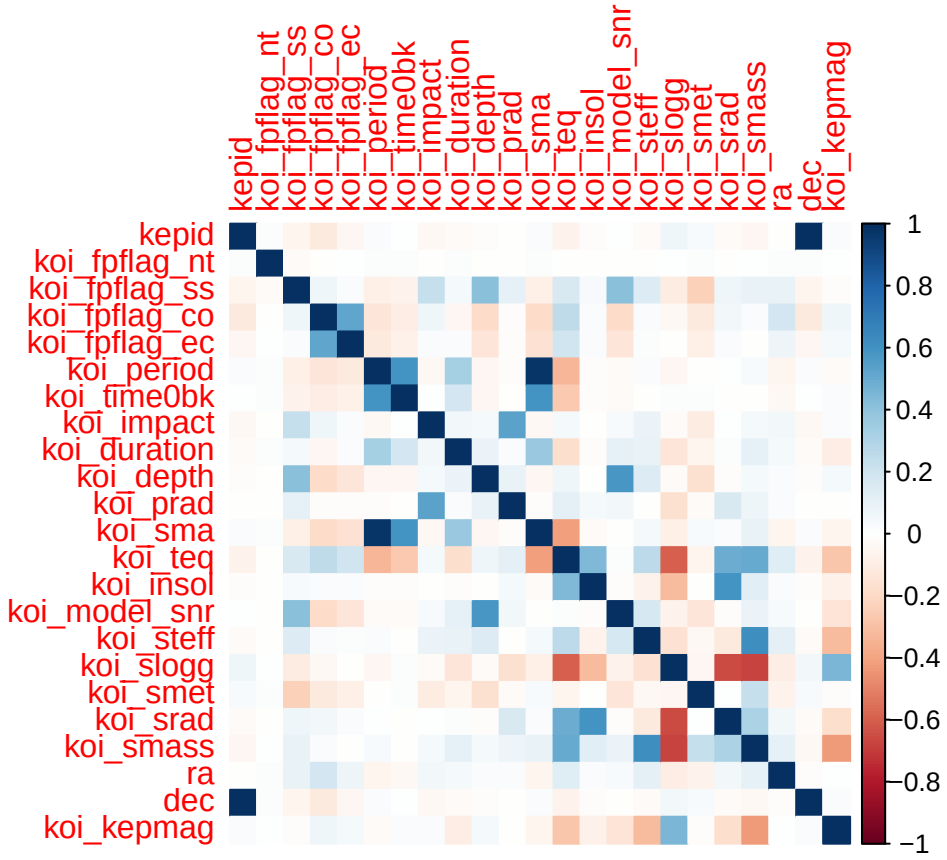
```
##                      kepid koi_fpflag_nt koi_fpflag_ss koi_fpflag_co
## kepid          1.000000000  1.297717e-02   -0.05269523  -0.118636412
## koi_fpflag_nt  0.012977174  1.000000e+00   -0.02322455  -0.004474337
## koi_fpflag_ss -0.052695226 -2.322455e-02    1.00000000   0.066186943
## koi_fpflag_co -0.118636412 -4.474337e-03    0.06618694   1.000000000
## koi_fpflag_ec -0.048811090  1.449281e-03    0.02434176   0.521545095
## koi_period     0.023472232  1.718361e-02   -0.08084287  -0.137119458
```

```
## koi_time0bk    0.003960418  1.312165e-02   -0.06102243  -0.097343352
## koi_impact    -0.030129613 -4.242627e-03    0.23936783   0.067368482
## koi_duration  -0.025809472  1.360843e-02    0.04115877  -0.047487255
## koi_depth     -0.012626512 -5.909600e-03    0.41531425  -0.184698019
## koi_prad      -0.002798876 -1.488524e-03    0.10132469  -0.012779556
## koi_sma        0.027741206  1.673577e-02   -0.08701482  -0.180167320
## koi_teq       -0.069905535 -3.583736e-03    0.16544934   0.259970693
## koi_insol     -0.016231261 -6.368549e-04    0.03710710   0.020468151
## koi_model_snr  0.002196849 -7.702922e-03    0.41394475  -0.188719842
## koi_steff     -0.026410064 -9.330173e-04    0.14491990   0.024168552
## koi_slogg      0.061824442  7.245003e-06   -0.10924365  -0.038594096
## koi_smet       0.035559528  1.253887e-02   -0.23696544  -0.115493534
## koi_srad      -0.025774004 -6.506617e-04    0.06105674   0.054901938
## koi_smass     -0.041971305  4.087841e-03    0.09407866   0.029018359
## ra            -0.004935698  1.567945e-02    0.09160162   0.180927632
## dec            0.993752976  1.303658e-02   -0.05180231  -0.117237934
## koi_kepmag     0.021975643  8.536468e-05   -0.01619494   0.064045542
##               koi_fpflag_ec   koi_period  koi_time0bk   koi_impact koi_duration
## kepid          -0.048811090  0.023472232  0.003960418 -0.030129613  -0.02580947
## koi_fpflag_nt   0.001449281  0.017183613  0.013121645 -0.004242627   0.01360843
## koi_fpflag_ss   0.024341757 -0.080842873 -0.061022427  0.239367835   0.04115877
## koi_fpflag_co   0.521545095 -0.137119458 -0.097343352  0.067368482  -0.04748726
## koi_fpflag_ec   1.000000000 -0.115543424 -0.078713792  0.025839642   0.02612479
## koi_period     -0.115543424  1.000000000  0.597104156 -0.032670370   0.33272373
## koi_time0bk    -0.078713792  0.597104156  1.000000000  0.009299622   0.18766759
## koi_impact      0.025839642 -0.032670370  0.009299622  1.000000000   0.05630680
## koi_duration    0.026124786  0.332723727  0.187667587  0.056306796   1.00000000
## koi_depth      -0.134135249 -0.041974963 -0.045025419  0.044931723   0.08828686
## koi_prad       -0.015787130 -0.011676132 -0.005743323  0.532368052   0.02048066
## koi_sma        -0.153500047  0.970859190  0.598928198 -0.031721823   0.37148193
## koi_teq         0.195354317 -0.336724965 -0.266650532  0.046790456  -0.17781861
## koi_insol       0.026853958 -0.019907496 -0.019906077 -0.009836834  -0.01857727
## koi_model_snr  -0.133161830 -0.022947598 -0.029072300  0.031906726   0.10420380
## koi_steff       0.015738986  0.020083359 -0.006753812  0.080822477   0.09797576
## koi_slogg      -0.001962590 -0.047706836  0.014645432 -0.029085936  -0.13072015
## koi_smet       -0.089513830 -0.003599795  0.017968307 -0.107815259  -0.05012730
## koi_srad        0.023432025  0.008936525 -0.009758209  0.001637597   0.01342643
## koi_smass      -0.003843479  0.039507749 -0.009305857  0.042605118   0.10607515
## ra              0.076990543 -0.053679052 -0.032288562  0.053010343   0.04217229
## dec            -0.049846095  0.023196464  0.001979101 -0.030484159  -0.02628815
## koi_kepmag      0.042747632 -0.028211030  0.028660415  0.023113116  -0.09867295
##                koi_depth     koi_prad      koi_sma      koi_teq     koi_insol
## kepid         -0.01262651 -0.002798876  0.027741206 -0.069905535 -0.0162312606
## koi_fpflag_nt -0.00590960 -0.001488524  0.016735768 -0.003583736 -0.0006368549
## koi_fpflag_ss  0.41531425  0.101324695 -0.087014817  0.165449338  0.0371070959
## koi_fpflag_co -0.18469802 -0.012779556 -0.180167320  0.259970693  0.0204681511
## koi_fpflag_ec -0.13413525 -0.015787130 -0.153500047  0.195354317  0.0268539578
## koi_period    -0.04197496 -0.011676132  0.970859190 -0.336724965 -0.0199074962
## koi_time0bk   -0.04502542 -0.005743323  0.598928198 -0.266650532 -0.0199060770
## koi_impact     0.04493172  0.532368052 -0.031721823  0.046790456 -0.0098368345
## koi_duration   0.08828686  0.020480660  0.371481930 -0.177818606 -0.0185772742
## koi_depth      1.00000000  0.091274489 -0.045349721  0.064874997 -0.0093366697
## koi_prad       0.09127449  1.000000000 -0.010069571  0.118740074  0.0432686722
## koi_sma       -0.04534972 -0.010069571  1.000000000 -0.410605031 -0.0275472892
```

```
## koi_teq        0.06487500  0.118740074 -0.410605031  1.000000000  0.4448415438
## koi_insol     -0.00933667  0.043268672 -0.027547289  0.444841544  1.0000000000
## koi_model_snr  0.58557192  0.054504241 -0.009680602  0.014458895 -0.0117194031
## koi_steff      0.14438424 -0.004675729  0.046305179  0.266122547 -0.0644886130
## koi_slogg     -0.02251395 -0.165855714 -0.088099322 -0.593580536 -0.3142703051
## koi_smet      -0.16668410 -0.023804950  0.032351529 -0.059503095 -0.0033250514
## koi_srad      -0.01874847  0.165324788  0.022242722  0.499622879  0.5991950775
## koi_smass      0.04920683  0.075135390  0.097142491  0.508686228  0.1253165917
## ra             0.02519529  0.025358429 -0.059639752  0.131144657  0.0285183855
## dec           -0.01357471 -0.003312871  0.027499526 -0.069645041 -0.0180088907
## koi_kepmag     0.04097200 -0.006698036 -0.053989315 -0.276147250 -0.0724533275
##               koi_model_snr     koi_steff     koi_slogg     koi_smet
## kepid          0.0021968486 -0.0264100644  6.182444e-02  0.035559528
## koi_fpflag_nt -0.0077029220 -0.0009330173  7.245003e-06  0.012538869
## koi_fpflag_ss  0.4139447528  0.1449198967 -1.092436e-01 -0.236965436
## koi_fpflag_co -0.1887198424  0.0241685525 -3.859410e-02 -0.115493534
## koi_fpflag_ec -0.1331618301  0.0157389862 -1.962590e-03 -0.089513830
## koi_period    -0.0229475983  0.0200833589 -4.770684e-02 -0.003599795
## koi_time0bk   -0.0290723001 -0.0067538121  1.464543e-02  0.017968307
## koi_impact     0.0319067256  0.0808224775 -2.908594e-02 -0.107815259
## koi_duration   0.1042038041  0.0979757600 -1.307202e-01 -0.050127299
## koi_depth      0.5855719240  0.1443842418 -2.251395e-02 -0.166684101
## koi_prad       0.0545042410 -0.0046757288 -1.658557e-01 -0.023804950
## koi_sma       -0.0096806016  0.0463051788 -8.809932e-02  0.032351529
## koi_teq        0.0144588946  0.2661225467 -5.935805e-01 -0.059503095
## koi_insol     -0.0117194031 -0.0644886130 -3.142703e-01 -0.003325051
## koi_model_snr  1.0000000000  0.1730236867 -6.255470e-02 -0.137384602
## koi_steff      0.1730236867  1.0000000000 -1.505088e-01 -0.033719660
## koi_slogg     -0.0625547005 -0.1505087889  1.000000e+00 -0.047838039
## koi_smet      -0.1373846021 -0.0337196605 -4.783804e-02  1.000000000
## koi_srad      -0.0112955150 -0.1157500451 -6.566317e-01  0.005810504
## koi_smass      0.0831905898  0.6158715620 -6.779476e-01  0.230346355
## ra             0.0383353836  0.1108858539 -9.343950e-02 -0.065099777
## dec           -0.0009274743 -0.0260987626  5.982561e-02  0.038208011
## koi_kepmag    -0.1414040754 -0.3163070633  4.590933e-01 -0.018071764
##                   koi_srad    koi_smass           ra           dec
## kepid         -0.0257740036 -0.041971305 -0.004935698  0.9937529759
## koi_fpflag_nt -0.0006506617  0.004087841  0.015679448  0.0130365779
## koi_fpflag_ss  0.0610567445  0.094078659  0.091601617 -0.0518023121
## koi_fpflag_co  0.0549019383  0.029018359  0.180927632 -0.1172379340
## koi_fpflag_ec  0.0234320248 -0.003843479  0.076990543 -0.0498460947
## koi_period     0.0089365252  0.039507749 -0.053679052  0.0231964638
## koi_time0bk   -0.0097582089 -0.009305857 -0.032288562  0.0019791012
## koi_impact     0.0016375969  0.042605118  0.053010343 -0.0304841588
## koi_duration   0.0134264254  0.106075150  0.042172288 -0.0262881536
## koi_depth     -0.0187484689  0.049206825  0.025195290 -0.0135747101
## koi_prad       0.1653247883  0.075135390  0.025358429 -0.0033128708
## koi_sma        0.0222427216  0.097142491 -0.059639752  0.0274995256
## koi_teq        0.4996228793  0.508686228  0.131144657 -0.0696450413
## koi_insol      0.5991950775  0.125316592  0.028518385 -0.0180088907
## koi_model_snr -0.0112955150  0.083190590  0.038335384 -0.0009274743
## koi_steff     -0.1157500451  0.615871562  0.110885854 -0.0260987626
## koi_slogg     -0.6566317215 -0.677947642 -0.093439504  0.0598256074
## koi_smet       0.0058105035  0.230346355 -0.065099777  0.0382080109
```

```
## koi_srad       1.0000000000  0.311457655  0.057686727 -0.0270891877
## koi_smass      0.3114576548  1.000000000  0.108717110 -0.0381245453
## ra             0.0576867271  0.108717110  1.000000000 -0.0172320061
## dec           -0.0270891877 -0.038124545 -0.017232006  1.0000000000
## koi_kepmag    -0.1718928778 -0.426056616  0.005640721  0.0204891506
##                  koi_kepmag
## kepid          2.197564e-02
## koi_fpflag_nt  8.536468e-05
## koi_fpflag_ss -1.619494e-02
## koi_fpflag_co  6.404554e-02
## koi_fpflag_ec  4.274763e-02
## koi_period    -2.821103e-02
## koi_time0bk    2.866041e-02
## koi_impact     2.311312e-02
## koi_duration  -9.867295e-02
## koi_depth      4.097200e-02
## koi_prad      -6.698036e-03
## koi_sma       -5.398931e-02
## koi_teq       -2.761473e-01
## koi_insol     -7.245333e-02
## koi_model_snr -1.414041e-01
## koi_steff     -3.163071e-01
## koi_slogg      4.590933e-01
## koi_smet      -1.807176e-02
## koi_srad      -1.718929e-01
## koi_smass     -4.260566e-01
## ra             5.640721e-03
## dec            2.048915e-02
## koi_kepmag     1.000000e+00
```

```
# correlation_matrix
corrplot(correlation_matrix, method='color')
```



This correlation matrix visualizes the strength and direction of the relationships between pairs of variables. Darker shades of blue represen Variables with little to no correlation appear in a lighter shade, near the center of the color scale (closer to 0), indicating a weaker relationship. The diagonal line of dark blue, naturally at 1, shows each variable's perfect positive correlation with itself. In the correlation matrix, we see several variables with notable relationships. For instance, koi_prad (planetary radius) and koi_depth (transit depth) exhibit a strong positive correlation, indicating that larger planets tend to have deeper transits. Similarly, koi_teq (equilibrium temperature) and koi_insol (insolation flux) are positively correlated, suggesting that planets with higher equilibrium temperatures receive more stellar radiation. koi_steff (stellar effective temperature) and koi_srad (stellar radius) show a relationship as well, potentially indicating that larger stars have higher temperatures. These insights could be pivotal in building the future robust predictive models.

After exploring the relationships within the explanatory variables, it is important to consider the effect of random effects in the model building. Random effects allow the model to account for the variability within clusters or groups that is not explained by the fixed effects (the measured variables).

```
# Random Effects
# setwd('/Users/pc/desktop/')

data_old <- project_data
data_old$koi_disposition <- factor(data_old$koi_disposition, levels=c("CONFIRMED","FALSE POSITIVE"))
data_old <- na.omit(data_old, cols = "koi_disposition")
x <-colnames(data_old)
print(x)
```

```
##  [1] "loc_rowid"         "kepid"             "kepoi_name"
##  [4] "kepler_name"       "koi_disposition"   "koi_pdisposition"
##  [7] "koi_score"         "koi_fpflag_nt"     "koi_fpflag_ss"
```

```
## [10] "koi_fpflag_co"     "koi_fpflag_ec"     "koi_period"
## [13] "koi_time0bk"       "koi_impact"        "koi_duration"
## [16] "koi_depth"         "koi_prad"          "koi_sma"
## [19] "koi_teq"           "koi_insol"         "koi_model_snr"
## [22] "koi_tce_plnt_num"  "koi_tce_delivname" "koi_steff"
## [25] "koi_slogg"         "koi_smet"          "koi_srad"
## [28] "koi_smass"         "ra"                "dec"
## [31] "koi_kepmag"        "koi_comment"
```

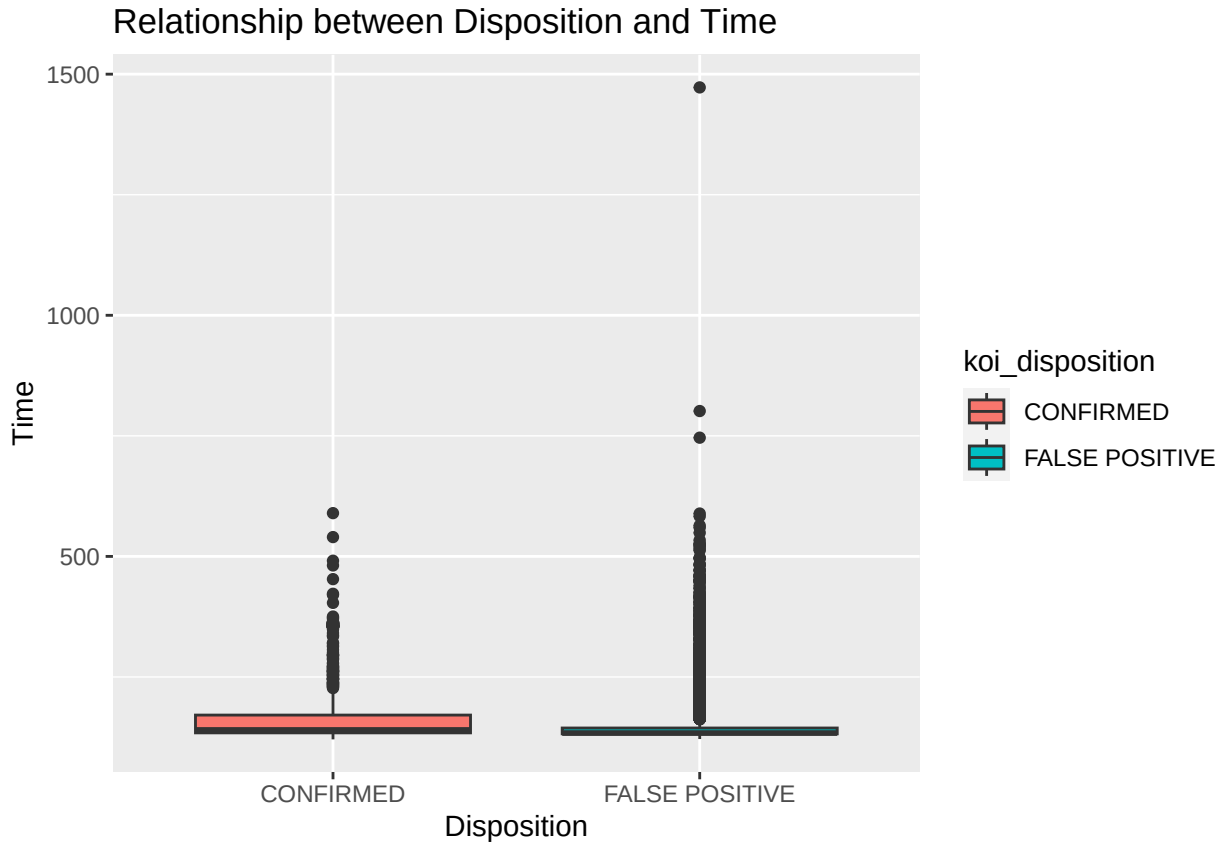
```r
# Subset the data to keep only the rows that satisfy all conditions

data_old$KOI_integer_part <- substr(data_old$kepoi_name, 1, 6)


# Assuming 'data' is your dataset and 'KOI_integer_part' is the column containing the KOI integer parts

# Count the frequency of each KOI_integer_part
frequency <- table(data_old$KOI_integer_part)
frequency_df <- as.data.frame(frequency)

# Filter the rows where the frequency is greater than 1
repeated_measurements <- frequency_df[frequency_df$Freq > 1, ]

# Print the filtered data frame
# print(repeated_measurements)

# 107 exoplanets are repeatedly measured. Therefore, it would be beneficial to consider random effects.


# Are data dependent in time ????

gam_model <- gam(koi_disposition ~ s(koi_time0bk), data = data_old, family = binomial)
summary(gam_model)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## koi_disposition ~ s(koi_time0bk)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.38774    0.04364   8.886   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                  edf Ref.df Chi.sq p-value
## s(koi_time0bk) 7.764  8.004  550.1  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0899   Deviance explained = 6.81%
## UBRE = 0.26542  Scale est. = 1         n = 6613
```

```r
# The GAM model provided a highly significant p-value of <2e-16, which indicates a strong relationship
# The GAM model has a edf of 4.738, which suggests a highly non-linear relationship.

# Plot the response variable against time

# Time variable: koi_time0bk -> The time corresponding to the center of the first detected transit in B

summary(data_old$koi_disposition)
```

```
##      CONFIRMED FALSE POSITIVE
##           2724           3889
```

```r
ggplot(data_old, aes(x = koi_disposition, y = koi_time0bk, fill = koi_disposition)) +   geom_boxplot(pos
```



In our analysis, we observed that certain planets in the dataset were subject to multiple measurements, suggesting the presence of potential random effects stemming from varying measurement times. We identified 107 planets with repeat observations and postulated that the Transit Epoch (koi_time0bk), representing the first detected transit's central time, might be linked with these multiple measurements.

To assess this, we utilized a Generalized Additive Model (GAM), an appropriate choice given the categorical nature of the response variable and the expected non-linear associations. The statistical significance of the relationship between koi_time0bk and koi_disposition was confirmed by an extremely small p-value from the GAM output, indicating that koi_time0bk does indeed influence the likelihood of a planet's confirmed disposition.

Complementary to the GAM findings, the boxplot analysis between koi_time0bk and koi_disposition revealed noticeable differences in transit epochs between the two disposition groups, reinforcing the relevance of koi_time0bk as a significant factor. This insight will be invaluable when we develop more robust predictive

models in the future, ensuring that such random effects are adequately accounted for.

Research Questions

RQ1: How do the various observational parameters of Kepler Objects of Interest (KOIs) influence their classification as actual planets? Impact of stellar parameters such as effective temperature and metallicity Impact of transit properties such as duration and depth

RQ2: To gauge Earth-size+ planets in the habitable zone ("Goldilocks") across various star types

RQ3: To establish the correlation between the different causes for disposition values "FALSE POSITIVE", "CONFIRMED" and "CANDIDATE". Also to establish the correlation between disposition values and flag variables.

How do the various observational parameters of Kepler Objects of Interest (KOIs) influence their classification as actual planets?

Impact of stellar parameters such as effective temperature and metallicity

Impact of transit properties such as duration and depth

Understanding how stellar parameters like effective temperature and metallicity, as well as transit properties such as duration and depth, influence the classification of Kepler Objects of Interest (KOIs) as actual planets is crucial for refining exoplanet detection methods. Higher metallicity stars may favor planet formation, while transit properties offer insights into planetary size and orbital dynamics. Addressing these factors and their impacts enhances our ability to accurately identify and classify exoplanets.

Sampling

Since there is clear class imbalance in our dataset, let's perform resampling of our dataset. There are various approaches to this issue: we can increase the number of instances in the minority class by randomly replicating them, we can reduce the instances of the majority class by randomly removing some of its instances, or sometimes, a combination of oversampling the minority class and undersampling the majority class can yield better results.

We cannot reduce the majority instances because we can't lose the data and oversampling the instances doesn't seem like good approach considering it's a scientific data captured by the keplar telescope. Let's try the third approach of combining undersampling and oversampling.

```
set.seed(123)
library(ROSE)
```

```
## Loaded ROSE 0.0-4
```

```
data <- ovun.sample(koi_disposition ~ ., data = data, method = "both", p = 0.5, N = 1.5*length(data$koi

summary(data)
```

```
##      kepid                koi_disposition koi_fpflag_nt      koi_fpflag_ss
##  Min.   :  757450   FALSE POSITIVE:5007   Min.   :  0.0000   Min.   :0.000
##  1st Qu.: 5563937   CONFIRMED     :4912   1st Qu.:  0.0000   1st Qu.:0.000
##  Median : 8007644                         Median :  0.0000   Median :0.000
##  Mean   : 7739883                         Mean   :  0.1485   Mean   :0.286
##  3rd Qu.: 9936698                         3rd Qu.:  0.0000   3rd Qu.:1.000
##  Max.   :12935144                         Max.   :465.0000   Max.   :1.000
##  koi_fpflag_co    koi_fpflag_ec      koi_period         koi_time0bk
##  Min.   :0.0000   Min.   :0.0000   Min.   :   0.2997   Min.   : 121.1
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:   2.3828   1st Qu.: 132.6
##  Median :0.0000   Median :0.0000   Median :   7.4074   Median : 136.0
```

```
##  Mean   :0.2258   Mean   :0.1429   Mean   :  31.8341   Mean   : 156.5
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:  21.5263   3rd Qu.: 161.9
##  Max.   :1.0000   Max.   :1.0000   Max.   :1071.2326   Max.   :1472.5
##    koi_impact       koi_duration        koi_depth           koi_prad
##  Min.   : 0.0000   Min.   :  0.1046   Min.   :     1.7   Min.   :   0.08
##  1st Qu.: 0.2130   1st Qu.:  2.4479   1st Qu.:   190.0   1st Qu.:   1.49
##  Median : 0.5860   Median :  3.7360   Median :   510.0   Median :   2.54
##  Mean   : 0.6148   Mean   :  5.2971   Mean   : 28492.9   Mean   :  26.52
##  3rd Qu.: 0.8990   3rd Qu.:  5.9210   3rd Qu.:  2140.0   3rd Qu.:  20.84
##  Max.   :23.1270   Max.   :117.5200   Max.   :922000.0   Max.   :4633.66
##     koi_sma          koi_teq        koi_insol        koi_model_snr
##  Min.   :0.0059   Min.   :  129   Min.   :       0   Min.   :   0.8
##  1st Qu.:0.0344   1st Qu.:  621   1st Qu.:      35   1st Qu.:  17.9
##  Median :0.0734   Median :  940   Median :     185   Median :  33.8
##  Mean   :0.1452   Mean   : 1145   Mean   :    9934   Mean   : 314.0
##  3rd Qu.:0.1502   3rd Qu.: 1437   3rd Qu.:    1009   3rd Qu.: 117.2
##  Max.   :2.1525   Max.   :14667   Max.   :10947555   Max.   :8755.2
##    koi_steff       koi_slogg        koi_smet          koi_srad
##  Min.   : 2661   Min.   :0.047   Min.   :-1.9400   Min.   :  0.116
##  1st Qu.: 5297   1st Qu.:4.247   1st Qu.:-0.2400   1st Qu.:  0.826
##  Median : 5743   Median :4.442   Median :-0.0800   Median :  0.992
##  Mean   : 5667   Mean   :4.331   Mean   :-0.1037   Mean   :  1.644
##  3rd Qu.: 6078   3rd Qu.:4.545   3rd Qu.: 0.0700   3rd Qu.:  1.294
##  Max.   :11360   Max.   :5.283   Max.   : 0.5600   Max.   :229.908
##    koi_smass            ra             dec          koi_kepmag
##  Min.   :0.094   Min.   :279.9   Min.   :36.58   Min.   : 6.966
##  1st Qu.:0.837   1st Qu.:288.5   1st Qu.:40.78   1st Qu.:13.496
##  Median :0.966   Median :292.2   Median :43.80   Median :14.548
##  Mean   :1.001   Mean   :291.9   Mean   :43.89   Mean   :14.299
##  3rd Qu.:1.090   3rd Qu.:295.8   3rd Qu.:46.81   3rd Qu.:15.316
##  Max.   :3.686   Max.   :301.7   Max.   :52.34   Max.   :20.003
##  koi_comment
##  Length:9919
##  Class :character
##  Mode  :character
##
##
##
```

Feature Selection

We have selected the stellar and transit properties as described by NASA data categorization to start with modelling the data.

Parametric Modeling

Let's start with the Generalized Linear Model (GLM) family because it offers a comprehensive and flexible framework for analyzing and interpreting diverse types of data. We can model the data more accurately by directly relating the response variable to linear combinations of the predictors through a suitable link function. Assuming that initiating our analysis with GLMs aligns closely with the underlying data structure (since our target variable is binary) and expecting more meaningful and interpretable results, we can fit the data as:

Generalized Linear Model (GLM)

```
set.seed(123)
##---- Fit glm model
model_glm <- glm(koi_disposition ~ koi_period+koi_impact+koi_duration+koi_depth+koi_prad+koi_sma+koi_te
                 family = binomial(link = "logit"))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##---- Model_1 summary
summary(model_glm)
```

```
##
## Call:
## glm(formula = koi_disposition ~ koi_period + koi_impact + koi_duration +
##     koi_depth + koi_prad + koi_sma + koi_teq + koi_insol + koi_model_snr +
##     koi_steff + koi_slogg + koi_smet + koi_srad + koi_smass +
##     koi_kepmag, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.614e+01  2.168e+00  12.057  < 2e-16 ***
## koi_period    -6.143e-03  3.486e-03  -1.762 0.077989 .
## koi_impact    -7.189e-01  1.067e-01  -6.736 1.62e-11 ***
## koi_duration  -1.989e-01  1.201e-02 -16.564  < 2e-16 ***
## koi_depth     -8.340e-05  1.175e-05  -7.098 1.26e-12 ***
## koi_prad      -1.211e-01  7.919e-03 -15.289  < 2e-16 ***
## koi_sma       -2.062e+00  1.394e+00  -1.480 0.138995
## koi_teq       -4.044e-03  1.501e-04 -26.939  < 2e-16 ***
## koi_insol      1.212e-05  2.656e-06   4.563 5.05e-06 ***
## koi_model_snr  1.272e-03  1.623e-04   7.840 4.49e-15 ***
## koi_steff      3.819e-04  1.471e-04   2.597 0.009417 **
## koi_slogg     -4.267e+00  4.105e-01 -10.396  < 2e-16 ***
## koi_smet       3.509e+00  2.078e-01  16.885  < 2e-16 ***
## koi_srad      -6.384e-01  1.917e-01  -3.331 0.000867 ***
## koi_smass     -8.612e-01  5.744e-01  -1.499 0.133799
## koi_kepmag    -4.873e-02  3.229e-02  -1.509 0.131291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13750  on 9918  degrees of freedom
## Residual deviance:  5847  on 9903  degrees of freedom
## AIC: 5879
##
## Number of Fisher Scoring iterations: 12
```

Significant Predictors

Variables such as koi_period, koi_time0bk, koi_impact, koi_duration, koi_depth, koi_prad, koi_teq, koi_model_snr, koi_slogg, koi_srad have very small p-values (< 2e-16 to 2.31e-13), indicating they significantly affect the likelihood of the koi_disposition

Non Significant Predictors

Variables like kepid, koi_insol, koi_steff, and koi_kepmag have large p-values, indicating no significant

evidence of their effect on koi_disposition

We can see a large drop in deviance from Null deviance to Residual deviance indicating a better fit with predictors.

Step Modeling

Let's perform stepwise modeling because it offers a structured, algorithmic approach to go through the multitude of available predictors in our dataset, selecting those that contribute most significantly to our model. Given the complexity and the high dimensionality of our data, stepwise modeling will allow us to efficiently pinpoint the most relevant factors, ensuring that our final model is both robust and manageable.

```
set.seed(123)
suppressWarnings({
  step_model <- step(model_glm, direction = "both")

summary(step_model)
})
```

```
## Start:  AIC=5879.01
## koi_disposition ~ koi_period + koi_impact + koi_duration + koi_depth +
##     koi_prad + koi_sma + koi_teq + koi_insol + koi_model_snr +
##     koi_steff + koi_slogg + koi_smet + koi_srad + koi_smass +
##     koi_kepmag
##
##                 Df Deviance    AIC
## - koi_insol      1     5848   5878
## <none>                 5847   5879
## - koi_sma        1     5849   5879
## - koi_smass      1     5849   5879
## - koi_kepmag     1     5849   5879
## - koi_steff      1     5854   5884
## - koi_srad       1     5861   5891
## - koi_impact     1     5893   5923
## - koi_model_snr  1     5920   5950
## - koi_smet       1     6142   6172
## - koi_teq        1     6176   6206
## - koi_duration   1     6235   6265
## - koi_prad       1     6259   6289
## - koi_period     1   110366 110396
## - koi_slogg      1   114979 115009
## - koi_depth      1   162269 162299
##
## Step:  AIC=5878.15
## koi_disposition ~ koi_period + koi_impact + koi_duration + koi_depth +
##     koi_prad + koi_sma + koi_teq + koi_model_snr + koi_steff +
##     koi_slogg + koi_smet + koi_srad + koi_smass + koi_kepmag
##
##                 Df Deviance    AIC
## - koi_sma        1   5850.0 5878.0
## <none>               5848.1 5878.1
## - koi_kepmag     1   5850.5 5878.5
## - koi_smass      1   5850.5 5878.5
## + koi_insol      1   5847.0 5879.0
## - koi_period     1   5851.7 5879.7
## - koi_steff      1   5854.8 5882.8
```

```
## - koi_srad       1   5861.0 5889.0
## - koi_impact     1   5893.8 5921.8
## - koi_model_snr  1   5921.1 5949.1
## - koi_depth      1   5939.6 5967.6
## - koi_slogg      1   6026.2 6054.2
## - koi_smet       1   6141.9 6169.9
## - koi_duration   1   6235.2 6263.2
## - koi_prad       1   6260.5 6288.5
## - koi_teq        1   6732.5 6760.5
##
## Step:  AIC=5878
## koi_disposition ~ koi_period + koi_impact + koi_duration + koi_depth +
##     koi_prad + koi_teq + koi_model_snr + koi_steff + koi_slogg +
##     koi_smet + koi_srad + koi_smass + koi_kepmag
##
##                 Df Deviance    AIC
## <none>               5850.0 5878.0
## + koi_sma        1   5848.1 5878.1
## - koi_kepmag     1   5852.5 5878.5
## - koi_smass      1   5853.0 5879.0
## + koi_insol      1   5849.2 5879.2
## - koi_steff      1   5856.0 5882.0
## - koi_srad       1   5863.8 5889.8
## - koi_impact     1   5895.0 5921.0
## - koi_model_snr  1   5924.6 5950.6
## - koi_depth      1   5942.5 5968.5
## - koi_slogg      1   6027.7 6053.7
## - koi_smet       1   6142.6 6168.6
## - koi_duration   1   6251.6 6277.6
## - koi_prad       1   6264.4 6290.4
## - koi_period     1   6276.7 6302.7
## - koi_teq        1   8218.3 8244.3
##
## Call:
## glm(formula = koi_disposition ~ koi_period + koi_impact + koi_duration +
##     koi_depth + koi_prad + koi_teq + koi_model_snr + koi_steff +
##     koi_slogg + koi_smet + koi_srad + koi_smass + koi_kepmag,
##     family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.573e+01  2.151e+00  11.963  < 2e-16 ***
## koi_period    -1.119e-02  5.880e-04 -19.028  < 2e-16 ***
## koi_impact    -7.114e-01  1.066e-01  -6.675 2.48e-11 ***
## koi_duration  -2.017e-01  1.189e-02 -16.971  < 2e-16 ***
## koi_depth     -8.390e-05  1.179e-05  -7.118 1.10e-12 ***
## koi_prad      -1.216e-01  7.925e-03 -15.339  < 2e-16 ***
## koi_teq       -3.849e-03  1.010e-04 -38.122  < 2e-16 ***
## koi_model_snr  1.284e-03  1.624e-04   7.906 2.65e-15 ***
## koi_steff      3.647e-04  1.474e-04   2.474 0.013346 *
## koi_slogg     -4.188e+00  4.070e-01 -10.290  < 2e-16 ***
## koi_smet       3.507e+00  2.083e-01  16.835  < 2e-16 ***
## koi_srad      -6.429e-01  1.914e-01  -3.359 0.000781 ***
```

```
## koi_smass     -1.002e+00  5.703e-01  -1.757 0.078921 .
## koi_kepmag    -5.037e-02  3.227e-02  -1.561 0.118516
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13750  on 9918  degrees of freedom
## Residual deviance:  5850  on 9905  degrees of freedom
## AIC: 5878
##
## Number of Fisher Scoring iterations: 11
```

The final model includes several predictors (koi_period, koi_time0bk, koi_impact, koi_duration, koi_depth, koi_prad, koi_teq, koi_model_snr, koi_steff, koi_slogg, koi_srad, and ra) after the stepwise procedure, indicating their significance in relation to koi_disposition.

After reviewing the correlation matrix, and the outputs from the stepwise model selection, we can choose predictors that not only have high predictive power but also maintain a balance between statistical significance, independence (low multicollinearity), and theoretical relevance to the phenomenon being studied. Based on these considerations:

Significant in Stepwise Modeling:

From the stepwise model, variables that were significant and not removed in the process include:

- **koi_period**
- **koi_time0bk**
- **koi_impact**
- **koi_duration**
- **koi_depth**
- **koi_prad**
- **koi_teq**
- **koi_model_snr**
- **koi_slogg**
- **koi_srad**
- **koi_smet**

These predictors are selected based on a combination statistical significance from stepwise modeling, low inter-correlations to avoid multicollinearity, and their relevance to the phenomenon under study.

Back to modeling, we can fit the data to our glm model as:

```
set.seed(123)
model_glm_filtered =  glm(koi_disposition ~ koi_impact+koi_teq+koi_prad+koi_period+
               family = binomial(link = "logit"))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

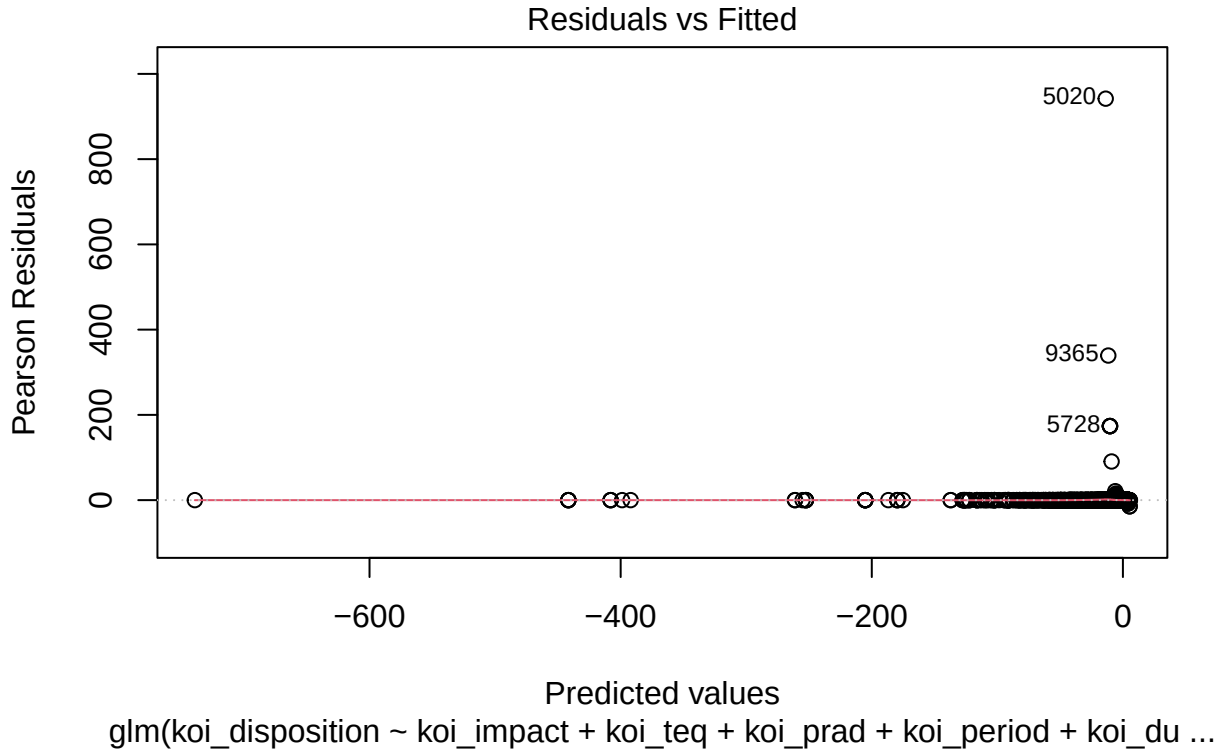
```
summary(model_glm_filtered)
```

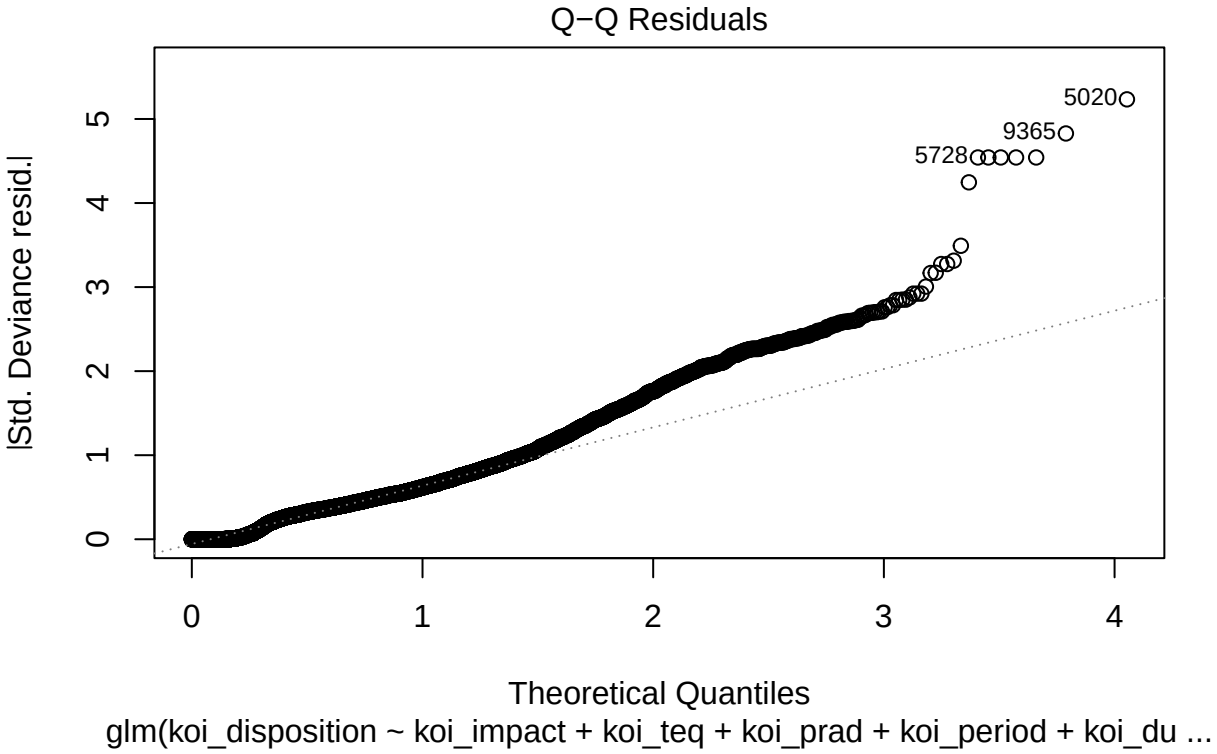
```
##
## Call:
## glm(formula = koi_disposition ~ koi_impact + koi_teq + koi_prad +
##     koi_period + koi_duration + koi_depth + koi_model_snr + koi_slogg +
##     koi_srad + koi_smet, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
```

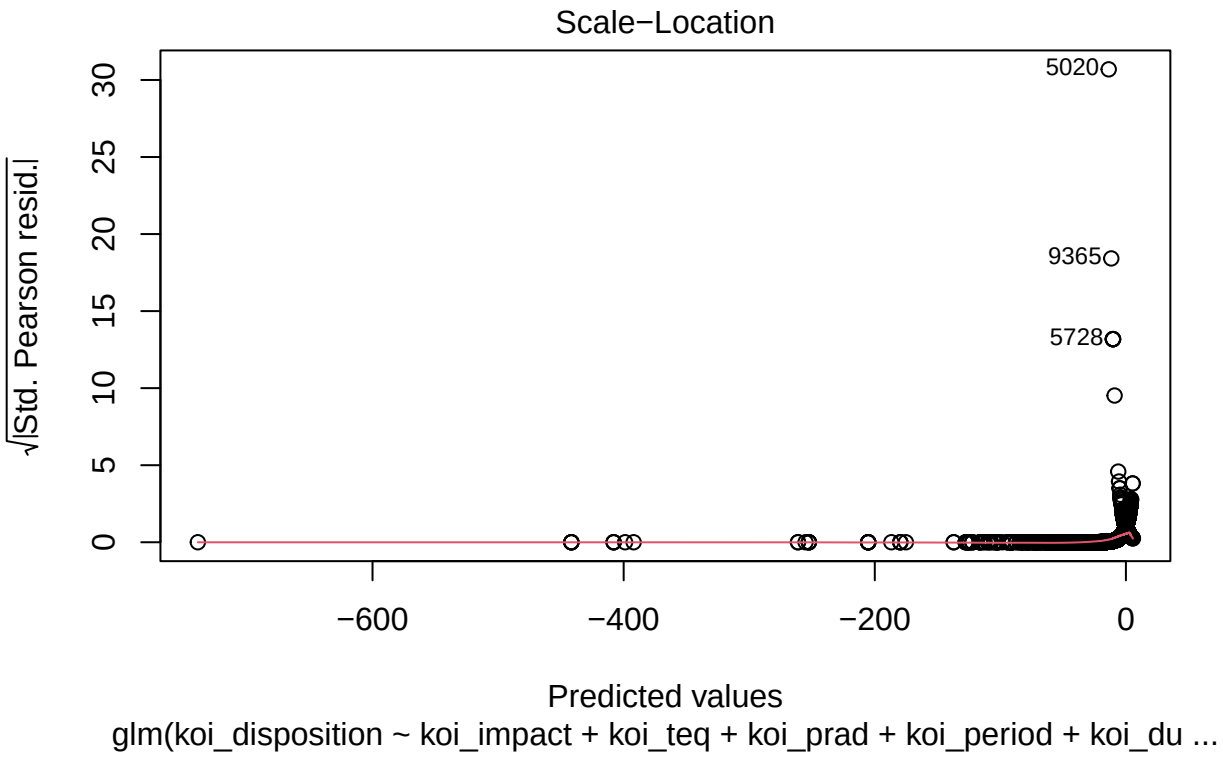
```
## (Intercept)    2.942e+01  1.797e+00  16.365  < 2e-16 ***
## koi_impact    -6.808e-01  1.057e-01  -6.439 1.20e-10 ***
## koi_teq       -3.798e-03  9.737e-05 -39.009  < 2e-16 ***
## koi_prad      -1.230e-01  7.925e-03 -15.517  < 2e-16 ***
## koi_period    -1.103e-02  5.812e-04 -18.976  < 2e-16 ***
## koi_duration  -1.991e-01  1.166e-02 -17.070  < 2e-16 ***
## koi_depth     -8.604e-05  1.187e-05  -7.251 4.15e-13 ***
## koi_model_snr  1.335e-03  1.610e-04   8.294  < 2e-16 ***
## koi_slogg     -4.881e+00  3.639e-01 -13.411  < 2e-16 ***
## koi_srad      -9.814e-01  1.424e-01  -6.891 5.53e-12 ***
## koi_smet       3.232e+00  1.496e-01  21.610  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13749.7  on 9918  degrees of freedom
## Residual deviance:  5860.6  on 9908  degrees of freedom
## AIC: 5882.6
##
## Number of Fisher Scoring iterations: 10
```

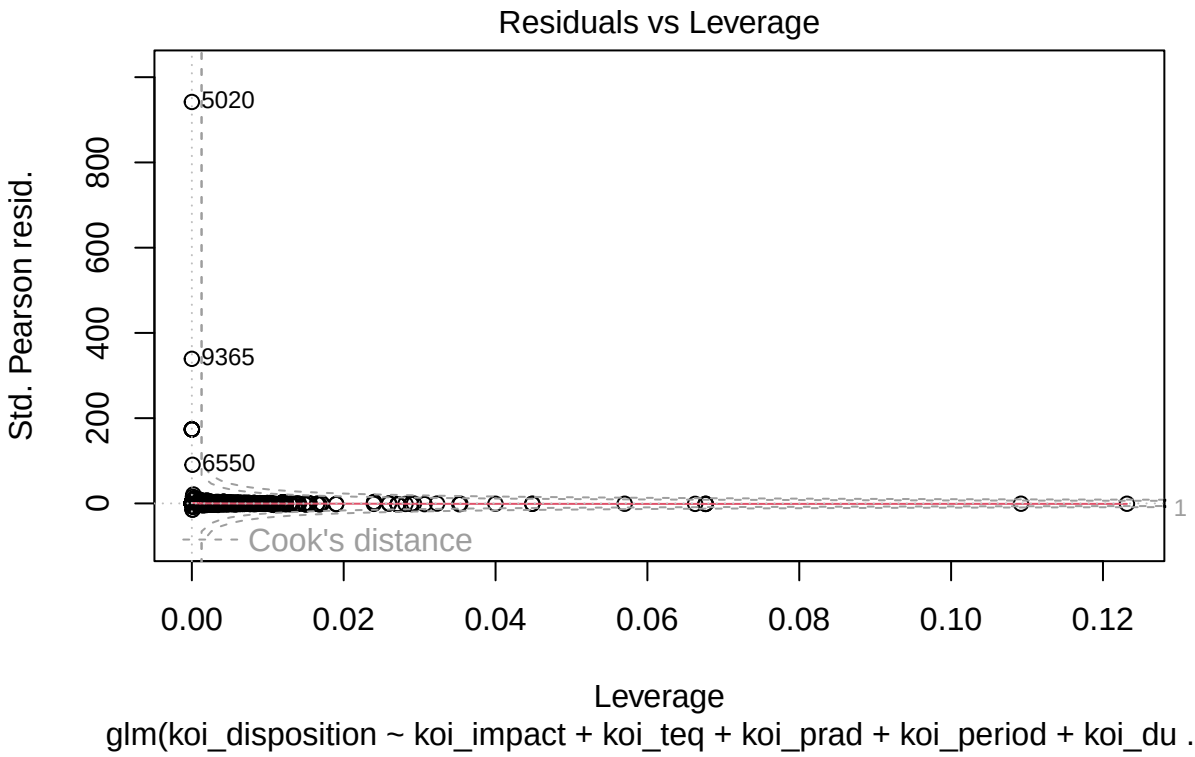
Let's plot the residual plots for better interpretation:

```
set.seed(123)
plot(model_glm_filtered)
```




Q−Q Residuals
|Std. Deviance resid.|
0
1
2
3
4
5
5020
9365
5728
0
1
2
3
4
Theoretical Quantiles
glm(koi_disposition ~ koi_impact + koi_teq + koi_prad + koi_period + koi_du ...


Scale−Location
√|Std. Pearson resid.|
0
5
10
15
20
25
30
5020
9365
5728
−600
−400
−200
0
Predicted values
glm(koi_disposition ~ koi_impact + koi_teq + koi_prad + koi_period + koi_du ...



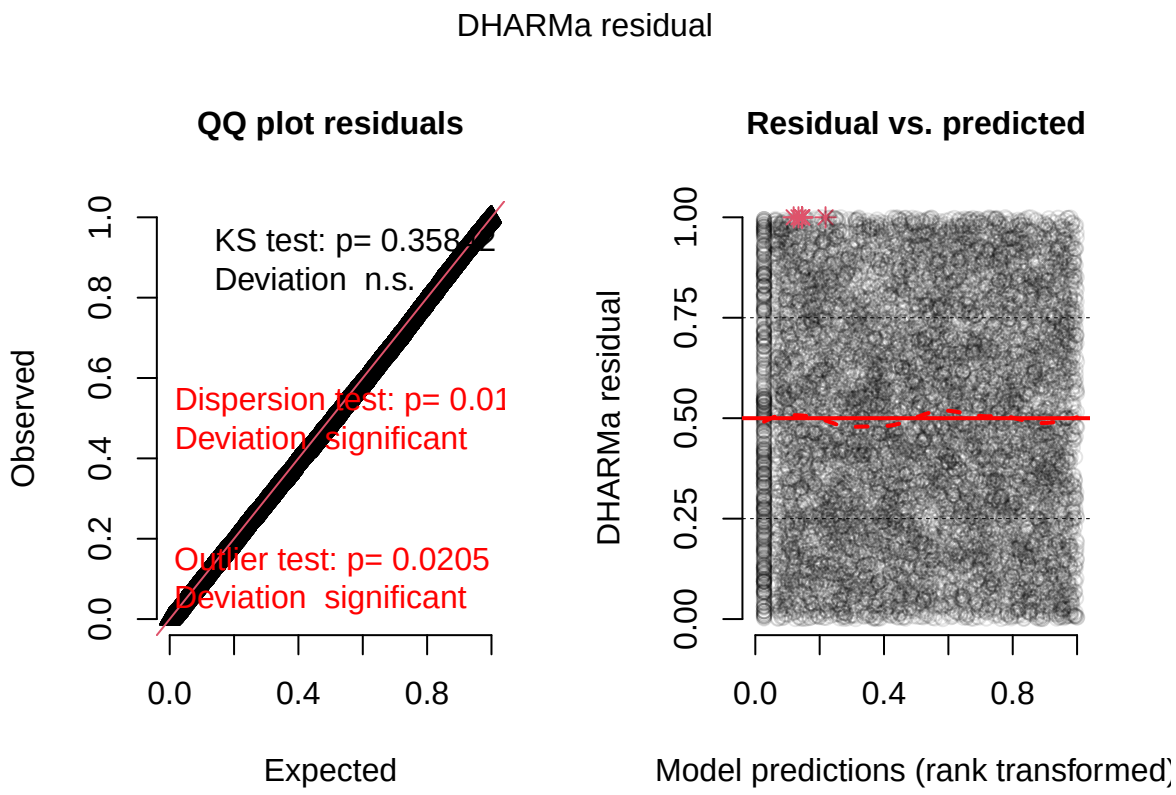
From the residuals vs fitted plot, there's an indication of a potential issue with a few large residuals, but the vast majority of data seems clustered around zero, indicating reasonable performance. Q-Q plot shows residuals may not be normally distributed, which is common for GLM with binomial outcomes. Scale-location plot shows some pattern indicating possible Heteroscedasticity. Residual vs leverage plot indicates a few points with higher leverage, which might be influential and could be potential outliers or have high leverage.

Summary of each model throws a warning called "Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred". This indicates model may be overfitting to the training data, our data may have some variables that perfectly predict the outcome, meaning that for certain values of the predictors, the outcome is always 0 or always 1, there may be outliers or influential observations that are significantly different from the rest of the data.

```
residuals_simulation <- simulateResiduals(fittedModel = model_glm_filtered, n = 250)

# Plotting the simulated residuals
plot(residuals_simulation)
```

Plotting Diagnostic Plots

```
## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```



QQ Plot - The Kolmogorov-Smirnov (KS) test has a p-value of 0.78884, indicating no significant deviation from the expected uniform distribution. This suggests that overall, the residuals are well-distributed. The dispersion test has a p-value of 0.008, which indicates significant overdispersion or underdispersion in the model. In the context of a binomial model, this could suggest that the variance of the response is not adequately captured by the model (i.e., the model could be over-predicting common outcomes or under-predicting rare outcomes). The outlier test has a p-value of 0.0111, suggesting the presence of significant outliers in the residuals. These are observations for which the model's predictions are substantially different from the actual values.

Residuals vs Predicted - The residuals should be randomly scattered around zero, with no discernible pattern. However, there appears to be a pattern in the residuals, with higher deviations occurring at certain ranges of predicted values, which could indicate that the model does not fit all areas of the predictor space equally well.

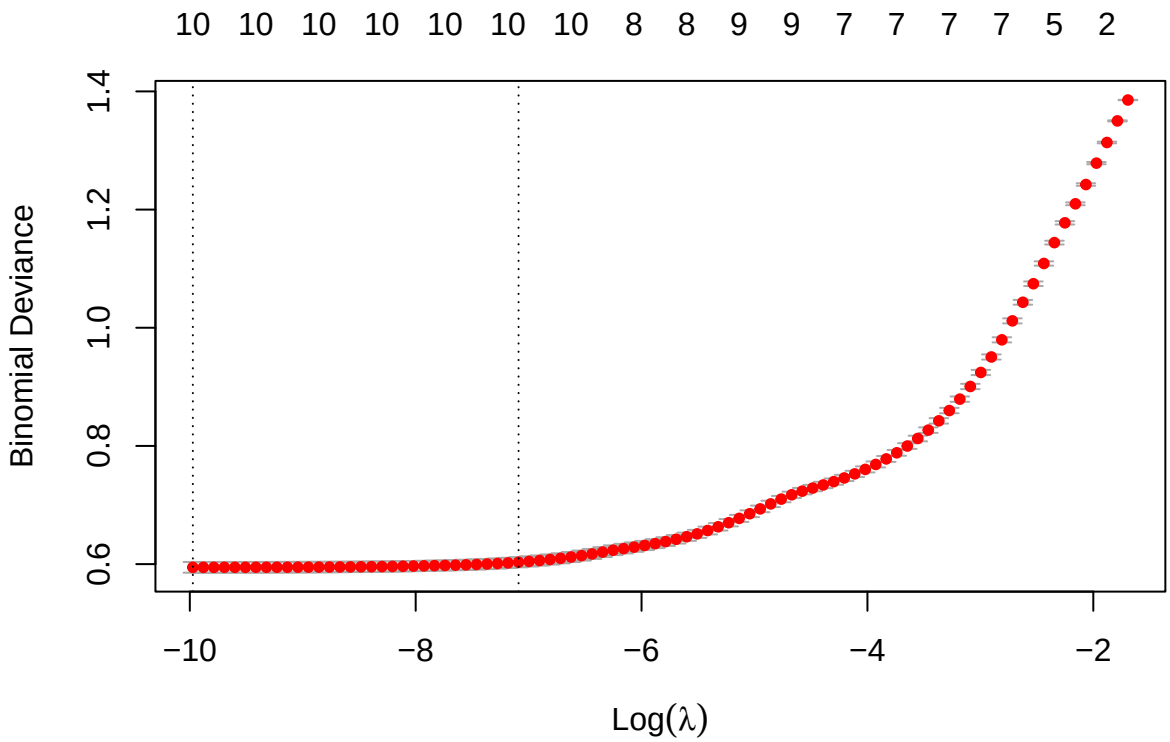
To deal with overfitting, let's use regularization techniques such as ridge regression or lasso, which can help prevent overfitting by penalizing the size of the coefficients.

```
# Perform Lasso Regression
set.seed(123)
# Prepare matrix of Targets and Predictors
x <- model.matrix(koi_disposition ~  koi_impact + koi_teq + koi_prad +
    koi_period + koi_duration + koi_depth + koi_model_snr + koi_slogg +
    koi_srad + koi_smet, data = data)

y <- data$koi_disposition

# Run Cross validated Lasso Regression
cv_fit <- cv.glmnet(x ,y, family="binomial", alpha = 1) #alpha = 1 for lasso
```

```r
# Check the best lambda value - shrinkage parameter in lasso regression
best_lambda <- cv_fit$lambda.min

# Fit the model
model_lasso <- glmnet(x, y, family = "binomial", alpha = 1, lambda = best_lambda)

coef(model_lasso)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                            s0
## (Intercept)     2.859133e+01
## (Intercept)     .
## koi_impact     -6.860607e-01
## koi_teq        -3.773253e-03
## koi_prad       -1.216294e-01
## koi_period     -1.099642e-02
## koi_duration   -1.974377e-01
## koi_depth      -7.894477e-05
## koi_model_snr   1.241599e-03
## koi_slogg      -4.719249e+00
## koi_srad       -9.165994e-01
## koi_smet        3.218619e+00
```

```r
# Plot the cv
plot(cv_fit)
```



The output from coef(model_lasso) displays the coefficients for a Lasso regression model fitted to predict the variable koi_disposition. From the graph, we see the red dots which represent binomial deviance and x-axis we have log(lambda). Dotted vertical lines represent where min deviation occurs and grey line indicates error bars for one standard deviation between mean deviance of each lambda.

It helps us to choose optimal value of regularization parameter(lambda).

For Evaluation, we typically examine metrics such as Accuracy, Area Under the Curve (AUC) from the Receiver Operating Characteristic (ROC) curve, Precision, Recall, and the F1 Score.

```
# Setting a seed for reproducibility
set.seed(123)
trainIndex <- createDataPartition(data$koi_disposition, p = .7, list = FALSE, times = 1)
trainData <- data[trainIndex, ]
testData <- data[-trainIndex, ]

# Create the model matrix for the new data
testdata_matrix <- model.matrix(~  koi_impact + koi_teq + koi_prad +
    koi_period + koi_duration + koi_depth + koi_model_snr + koi_slogg +
    koi_srad + koi_smet, data = testData)

# Create the model matrix for the new data
traindata_matrix <- model.matrix(~  koi_impact + koi_teq + koi_prad +
    koi_period + koi_duration + koi_depth + koi_model_snr + koi_slogg +
    koi_srad + koi_smet,data = trainData)

model_lasso_subset <- glmnet(x, y, family = "binomial", alpha = 1, lambda = best_lambda)

# Make predictions on the testing data
predicted_probabilities <- predict(model_lasso, newx = testdata_matrix, type = "response")

predicted_class <- ifelse(predicted_probabilities > 0.5, "CONFIRMED", "FALSE POSITIVE")
predicted_class <- factor(predicted_class, levels = levels(trainData$koi_disposition))

# Calculate the confusion matrix
confusion_matrix <- confusionMatrix(predicted_class, testData$koi_disposition)

# Convert to a numeric vector
predicted_probabilities <- as.numeric(predicted_probabilities)

# Calculate ROC curve and AUC
roc_result <- roc(testData$koi_disposition, predicted_probabilities)
```

```
## Setting levels: control = FALSE POSITIVE, case = CONFIRMED
```

```
## Setting direction: controls < cases
```

```
auc_value <- auc(roc_result)

# Output the confusion matrix and AUC value
print(confusion_matrix)
```

```
## Confusion Matrix and Statistics
##
##                Reference
## Prediction       FALSE POSITIVE CONFIRMED
##   FALSE POSITIVE           1275       103
```

```
##   CONFIRMED                227      1370
##
##                Accuracy : 0.8891
##                  95% CI : (0.8772, 0.9001)
##     No Information Rate : 0.5049
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7783
##
##  Mcnemar's Test P-Value : 1.28e-11
##
##             Sensitivity : 0.8489
##             Specificity : 0.9301
##          Pos Pred Value : 0.9253
##          Neg Pred Value : 0.8579
##              Prevalence : 0.5049
##          Detection Rate : 0.4286
##    Detection Prevalence : 0.4632
##       Balanced Accuracy : 0.8895
##
##        'Positive' Class : FALSE POSITIVE
##
```

```
print(auc_value)
```

```
## Area under the curve: 0.9451
```

The AUC value is 0.9451, which is a measure of the model's ability to discriminate between the positive and negative classes. An AUC of 1 represents a perfect model, while an AUC of 0.5 represents a worthless model. An AUC of 0.9451 suggests that the model has a very good discriminatory ability.

The Accuracy of the model is 0.88 The Kappa statistic is 0.77, which is a measure of how much better the classifier is performing over the performance of a classifier that simply guesses at random.

Let's find confusion matrix values before performing regularization:

```
set.seed(123)
predicted_probabilities <- predict(model_glm_filtered, newdata = testData, type = "response")

predicted_outcomes <- ifelse(predicted_probabilities > 0.5, "CONFIRMED", "FALSE POSITIVE")
predicted_outcomes_factor <- factor(predicted_outcomes, levels = levels(testData$koi_disposition))

# Use the confusionMatrix function from the caret package
conf_matrix <- confusionMatrix(predicted_outcomes_factor, testData$koi_disposition)

# Print the confusion matrix
print(conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##                 Reference
## Prediction       FALSE POSITIVE CONFIRMED
##   FALSE POSITIVE           1279       104
##   CONFIRMED                 223      1369
##
##                Accuracy : 0.8901
##                  95% CI : (0.8783, 0.9011)
```

```
##     No Information Rate : 0.5049
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7803
##
##  Mcnemar's Test P-Value : 6.781e-11
##
##             Sensitivity : 0.8515
##             Specificity : 0.9294
##          Pos Pred Value : 0.9248
##          Neg Pred Value : 0.8599
##              Prevalence : 0.5049
##          Detection Rate : 0.4299
##    Detection Prevalence : 0.4649
##       Balanced Accuracy : 0.8905
##
##        'Positive' Class : FALSE POSITIVE
##
```

Results are almost similar prior to regularization.

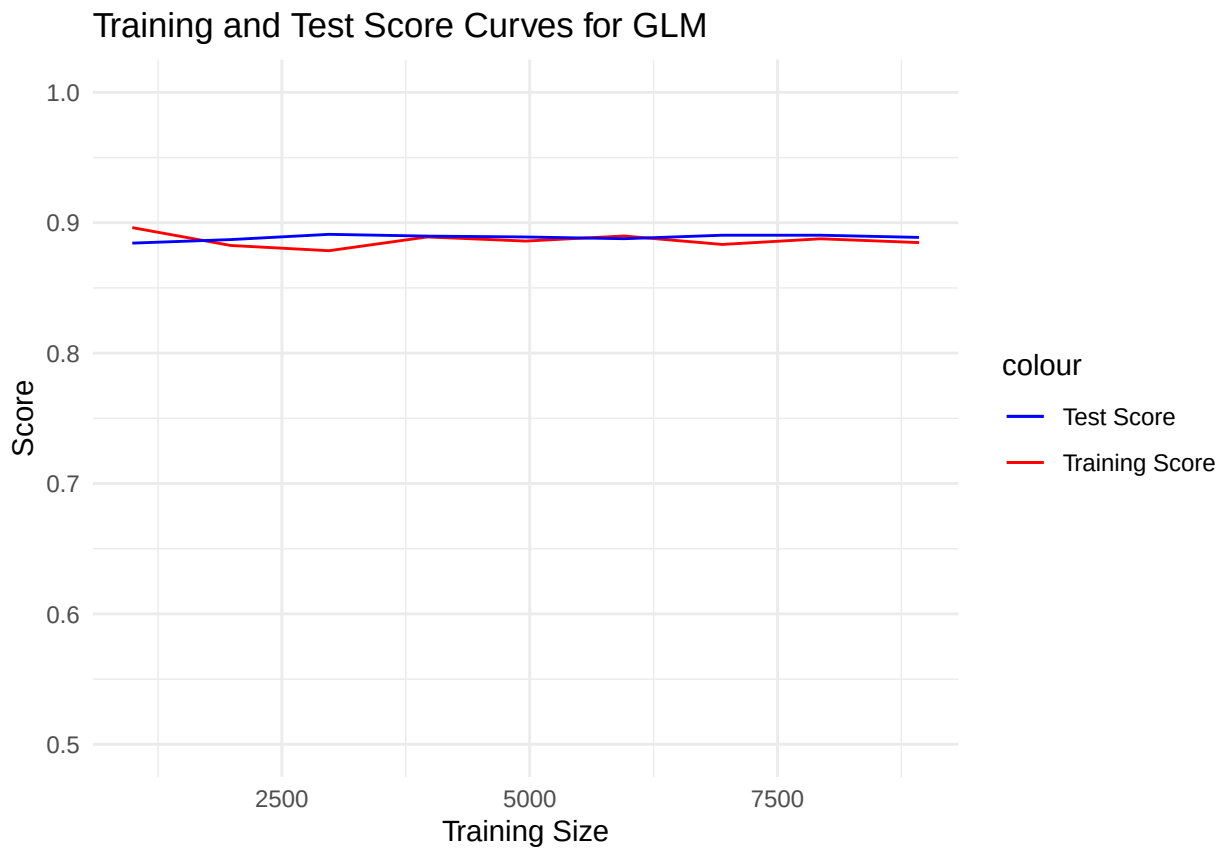
Let's plot the learning curve before regularization

```r
set.seed(123) # For reproducibility

# Split your data into training and testing sets
trainIndex <- createDataPartition(data$koi_disposition, p = .7, list = FALSE, times = 1)
trainData <- data[trainIndex, ]
testData <- data[-trainIndex, ]

# Initialize vectors to store the performance metrics
training_sizes <- seq(0.1, 0.9, by = 0.1) # Adjust this for more or fewer points
train_accuracies <- numeric(length(training_sizes))
test_accuracies <- numeric(length(training_sizes))


scores_1 <- matrix(NA, nrow = length(training_sizes), ncol = 3)
# Loop over the defined training sizes
for (i in seq_along(training_sizes)) {
  subsetIndex <- sample(seq_len(nrow(trainData)), size = floor(training_sizes[i] * nrow(trainData)))
  trainDataSubset <- trainData[subsetIndex, ]

  # Fit the GLM model on the subset of training data
  model_glm_filtered_subset <- glm(koi_disposition ~  koi_impact + koi_teq + koi_prad + koi_period + koi
    koi_srad + koi_smet, data = trainDataSubset, family = binomial(link = "logit"))

  # Make predictions and calculate accuracy for the subset of training data
  train_predictions <- predict(model_glm_filtered_subset, newdata = trainDataSubset, type = "response")
  train_predicted_class <- ifelse(train_predictions > 0.5, "CONFIRMED", "FALSE POSITIVE")
  train_accuracies[i] <- mean(train_predicted_class == trainDataSubset$koi_disposition)

  # Make predictions and calculate accuracy for the testing data
  test_predictions <- predict(model_glm_filtered_subset, newdata = testData, type = "response")
  test_predicted_class <- ifelse(test_predictions > 0.5, "CONFIRMED", "FALSE POSITIVE")
  test_accuracies[i] <- mean(test_predicted_class == testData$koi_disposition)
```

```
  scores_1[i, ] <- c(floor(training_sizes[i] * nrow(data)), train_accuracies[i], test_accuracies[i])
}
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
scores_df_1 <- as.data.frame(scores_1)
names(scores_df_1) <- c("train_size", "train_score", "test_score")

ggplot(scores_df_1, aes(x = train_size)) +
  geom_line(aes(y = train_score, color = "Training Score")) +
  geom_line(aes(y = test_score, color = "Test Score")) +
  scale_color_manual(values = c("blue", "red")) +
  labs(x = "Training Size", y = "Score") +
  ggtitle("Training and Test Score Curves for GLM") +
  theme_minimal()+
  scale_y_continuous(limits = c(0.50, 1))
```



Both the training score (red line) and the test score (blue line) are almost horizontal, suggesting that the performance of the model is quite stable across different training sizes. The model seems to have reached its performance limit with the given features and model complexity.

The training and test scores are very close to each other across the entire range of training sizes. This convergence typically suggests that the model is neither overfitting nor underfitting significantly. It is well-generalized to unseen data.

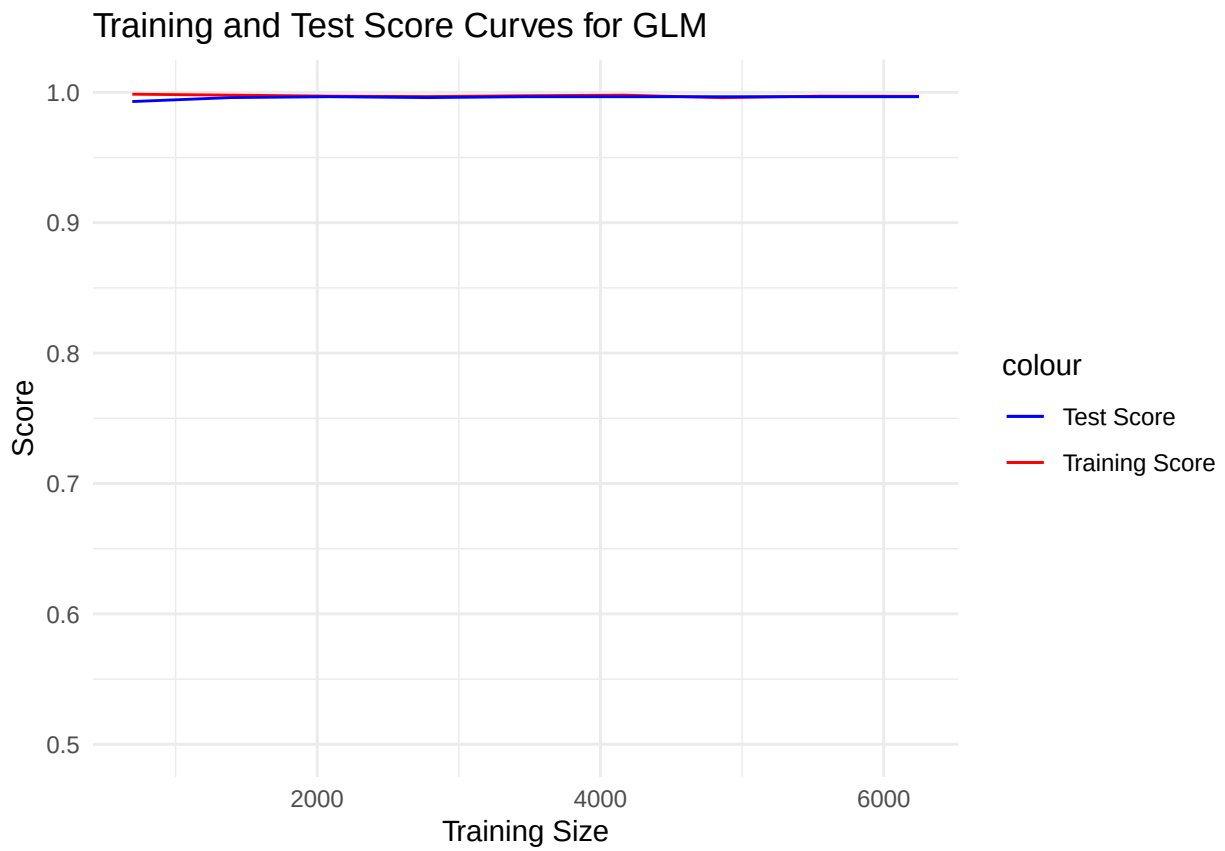
The score appears to be consistently high (just under 0.9), indicating good predictive performance. This level of performance might be satisfactory for the given task, depending on the complexity of the problem and the baseline or threshold for success.

Overall, the learning curve suggests that the model is not learning much even if we add more data i.e there is no change in accuracy as data sizes expand.

Let's plot the Learning curves after Regularization

```
set.seed(123)
suppressWarnings({
trainIndex2 <- createDataPartition(data$koi_disposition, p = .7, list = FALSE, times = 1)
trainData2 <- data[trainIndex2, ]
testData2 <- data[-trainIndex2, ]

# Initialize vectors to store the performance metrics
training_sizes <- seq(0.1, 0.9, by = 0.1) # Adjust this for more or fewer points
train_accuracies_2 <- numeric(length(training_sizes))
test_accuracies_2 <- numeric(length(training_sizes))
```

```
scores_2 <- matrix(NA, nrow = length(training_sizes), ncol = 3)
# Loop over the defined training sizes
for (i in seq_along(training_sizes)) {
  subsetIndex <- sample(seq_len(nrow(trainData2)), size = floor(training_sizes[i] * nrow(trainData2)))
  trainDataSubset <- trainData2[subsetIndex, ]

  # Check if trainDataSubset is empty or missing
  if (nrow(trainDataSubset) == 0) {
    cat("Warning: trainDataSubset is empty for training size", training_sizes[i], "\n")
    next
  }

  # Check if column names are present in trainDataSubset
  required_cols <- c("koi_disposition", "koi_impact", "koi_teq", "koi_prad", "koi_period", "koi_duratio
  if (!all(required_cols %in% colnames(trainDataSubset))) {
    cat("Warning: Required columns are missing in trainDataSubset\n")
    next
  }

  # Fit the GLM model on the subset of training data
  model_lasso_subset <- glmnet(as.matrix(trainDataSubset[, -1]), as.factor(trainDataSubset$koi_dispositi

  # Make predictions and calculate accuracy for the subset of training data
  train_predictions <- predict(model_lasso_subset, newx = as.matrix(trainDataSubset[, -1]), type = "resp
  train_predicted_class <- ifelse(train_predictions > 0.5, "CONFIRMED", "FALSE POSITIVE")
  train_accuracies_2[i] <- mean(train_predicted_class == trainDataSubset$koi_disposition)

  # Make predictions and calculate accuracy for the testing data
  test_predictions <- predict(model_lasso_subset, newx = as.matrix(testData2[, -1]), type = "response")
  test_predicted_class <- ifelse(test_predictions > 0.5, "CONFIRMED", "FALSE POSITIVE")
  test_accuracies_2[i] <- mean(test_predicted_class == testData$koi_disposition)

  scores_2[i, ] <- c(floor(training_sizes[i] * nrow(trainData2)), train_accuracies_2[i], test_accuracies
}
scores_df_2 <- as.data.frame(scores_2)
names(scores_df_2) <- c("train_size", "train_score", "test_score")
# Create a data frame for plotting
# learning_curve_df <- data.frame(
#   TrainingSize = rep(training_sizes, each = 2) * nrow(trainData),
#   Accuracy = c(train_accuracies, test_accuracies),
#   DataSet = factor(rep(c("Training", "Testing"), each = length(train_accuracies)))
# )

ggplot(scores_df_2, aes(x = train_size)) +
  geom_line(aes(y = train_score, color = "Training Score")) +
  geom_line(aes(y = test_score, color = "Test Score")) +
  scale_color_manual(values = c("blue", "red")) +
  labs(x = "Training Size", y = "Score") +
  ggtitle("Training and Test Score Curves for GLM") +
  theme_minimal()+
  scale_y_continuous(limits = c(0.50, 1))

})
```



Both curves are situated high on the y-axis, which suggests that the model performs well on both the training and testing data. This is typically indicative of a good fit to the data. The training and testing curves are very close to each other across the entire range of training sizes. This behavior implies that the model generalizes well to unseen data and there is little to no overfitting occurring.

There is minimal fluctuation in scores as the training size increases. The performance of the model is stable across different amounts of training data, indicating that additional data does not significantly change the model's predictions. Both curves are flat..

Overall, after regularization as well, the learning curve suggests that the model is not learning much even if we add more data i.e there is no change in accuracy as data sizes expand

Generalized Additive Model (GAM)

We can test with GAM since it provides a flexible generalization of ordinary linear or generalized linear models, allowing for the modeling of non-linear relationships between the predictor variables and the response variable.

```
set.seed(123)
suppressWarnings({
# Fit a GAM model
gam_model <- gam(koi_disposition ~ s(koi_impact) + s(koi_teq) + s(koi_prad) +
                 s(koi_period) + s(koi_duration) + s(koi_depth) +
                 s(koi_model_snr) + s(koi_slogg) + s(koi_srad) + s(koi_smet),
                 family = binomial(link = "logit"), data = trainData)

# Summary of the GAM model
summary(gam_model)
```

```
# Make predictions
predicted_probabilities_gam <- predict(gam_model, newdata = testData, type = "response")

# ROC analysis
roc_result_gam <- roc(testData$koi_disposition, predicted_probabilities_gam)

# AUC value
auc_value_gam <- auc(roc_result_gam)
print(auc_value_gam)
})
```

```
## Setting levels: control = FALSE POSITIVE, case = CONFIRMED
```

```
## Setting direction: controls < cases
```

```
## Area under the curve: 0.9536
```

```
set.seed(123)
print("AIC for gam_model")
```

```
## [1] "AIC for gam_model"
```

```
print(AIC(gam_model))
```

```
## [1] 3786.538
```

The adjusted R-squared is 68%, which means that about 68% of the variability in the data is explained by the model. Similar to R-squared, 61% of the deviance (a measure of goodness of fit) is explained by the model. The area under the curve (AUC) for your model is 0.95, which is a measure of the model's ability to discriminate between the two classes ("FALSE POSITIVE" and "CONFIRMED"). An AUC close to 1 indicates a very good discriminative ability.

The edf (estimated degrees of freedom) values for the smooth terms of the predictors range from close to 1 (which would suggest a linear relationship) to around 9, which suggests a more complex, non-linear relationship. A higher edf value indicates a more complex shape.

s(koi_impact), s(koi_teq), s(koi_prad), s(koi_period), s(koi_duration), s(koi_depth), s(koi_model_snr), s(koi_slogg), and s(koi_srad) are highly significant ($p < 0.001$), indicating strong non-linear effects of these variables on the response variable.

With an AUC of 0.95, our GAM model shows a strong discriminatory ability to differentiate between "CONFIRMED" and "FALSE POSITIVE" instances of koi_disposition.

In summary, the GAM model shows a strong performance in classifying koi_disposition with significant non-linear effects from several predictors. The high AUC value suggests that the model has a high predictive accuracy.

In summary, the GAM model shows a strong performance in classifying koi_disposition with significant non-linear effects from several predictors. The high AUC value suggests that the model has a high predictive accuracy.

Let's evaluate using confusion matrix:

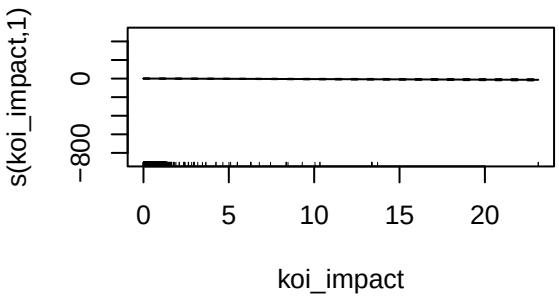
```
set.seed(123)
# Make predictions on the testing data
predicted_probabilities_gam <- predict(gam_model, newdata = testData, type = "response")

predicted_class_gam <- ifelse(predicted_probabilities_gam > 0.5, "CONFIRMED", "FALSE POSITIVE")
predicted_class_gam <- factor(predicted_class, levels = levels(trainData$koi_disposition))
```

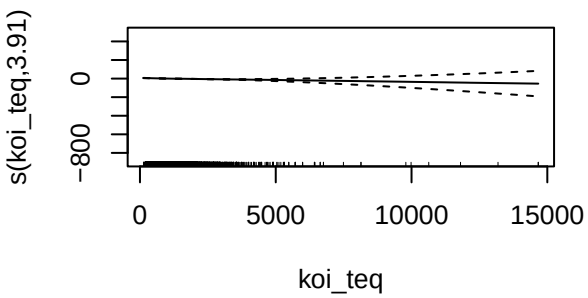
```r
# Calculate the confusion matrix
confusion_matrix <- confusionMatrix(predicted_class_gam, testData$koi_disposition)


print(confusion_matrix)
```

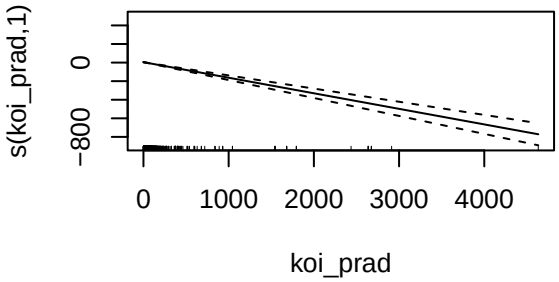
```
## Confusion Matrix and Statistics
##
##                 Reference
## Prediction       FALSE POSITIVE CONFIRMED
##   FALSE POSITIVE           1275       103
##   CONFIRMED                 227      1370
##
##                Accuracy : 0.8891
##                  95% CI : (0.8772, 0.9001)
##     No Information Rate : 0.5049
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7783
##
##  Mcnemar's Test P-Value : 1.28e-11
##
##             Sensitivity : 0.8489
##             Specificity : 0.9301
##          Pos Pred Value : 0.9253
##          Neg Pred Value : 0.8579
##              Prevalence : 0.5049
##          Detection Rate : 0.4286
##    Detection Prevalence : 0.4632
##       Balanced Accuracy : 0.8895
##
##        'Positive' Class : FALSE POSITIVE
##
```

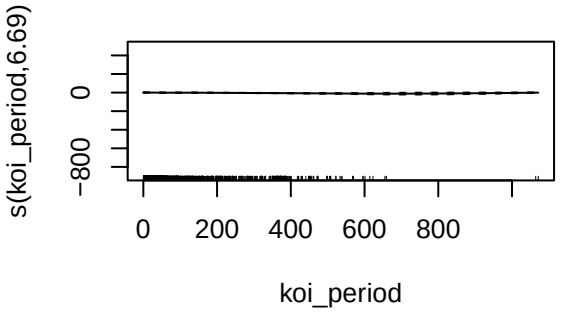
The model correctly predicted 88.59% of the cases. 95% confidence interval for the accuracy, ranging from 87.7% to 89.5% suggests that the model's accuracy is reliably high. The Cohen's Kappa statistic of 0.77 indicates a substantial agreement between the predicted and actual values, correcting for chance agreement.

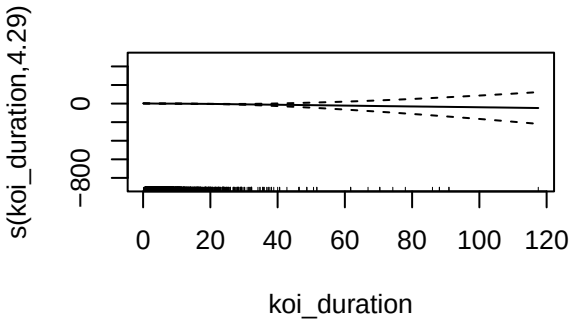
```r
set.seed(123)
par(mfrow = c(2, 2))
plot(gam_model)
```

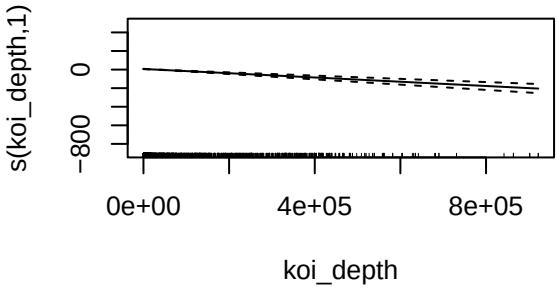

s(koi_impact,1)
0
−800
0
5
10
15
20
koi_impact

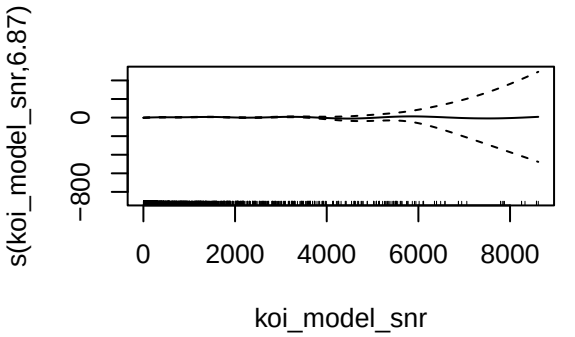

s(koi_teq,3.91)
0
−800
0
5000
10000
15000
koi_teq

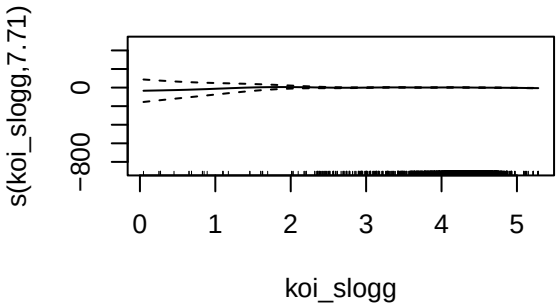

s(koi_prad,1)
0
−800
0
1000
2000
3000
4000
koi_prad

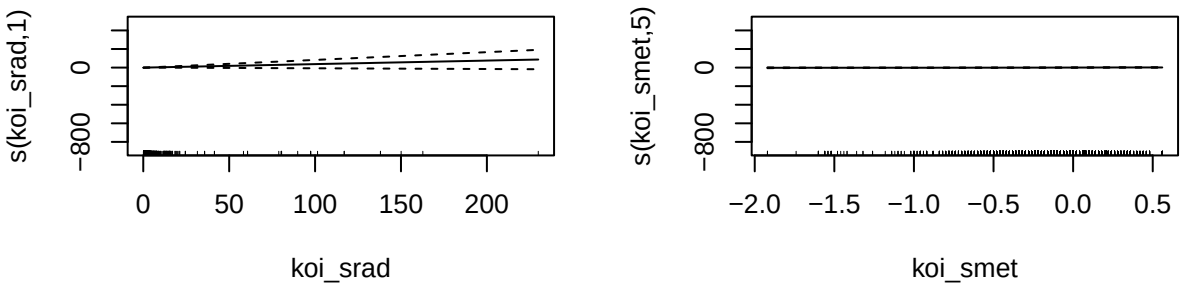

s(koi_period,6.69)
0
−800
0
200
400
600
800
koi_period


s(koi_duration,4.29)
−800
0
0
20
40
60
80
100
120
koi_duration


s(koi_depth,1)
−800
0
0e+00
4e+05
8e+05
koi_depth


s(koi_model_snr,6.87)
−800
0
0
2000
4000
6000
8000
koi_model_snr


s(koi_slogg,7.71)
−800
0
0
1
2
3
4
5
koi_slogg



Variables like koi_impact show an almost linear relationship, whereas variables such as koi_teq, koi_prad, koi_period, and koi_srad demonstrate more complex, non-linear effects on the response. The narrow confidence intervals around the smooth terms for variables like koi_teq imply a higher confidence in these estimates. In contrast, the non-significance of the koi_time0bk term and the relatively linear effect of koi_depth indicate that their relationships with the response are less pronounced.

In the plots, we do not see dramatic curves or fluctuations, which might suggest that while non-linear relationships are present, they are relatively smooth and gradual changes rather than abrupt shifts. Overall, the GAM has identified both linear and non-linear relationships between the predictors and the response variable.

Plotting Diagnostic Plots

```
# If you don't have devtools installed, you need to install it first
options(repos = c(CRAN = "https://cloud.r-project.org"))
install.packages("devtools")
```

```
##
## The downloaded binary packages are in
##  /var/folders/h8/f0jh18c50250m_6y7l0w5hqr0000gn/T//RtmpzhwaO4/downloaded_packages
```

```
# Then you can install mgcViz from GitHub
devtools::install_github("mfasiolo/mgcViz")
```

```
## Using GitHub PAT from the git credential store.
```

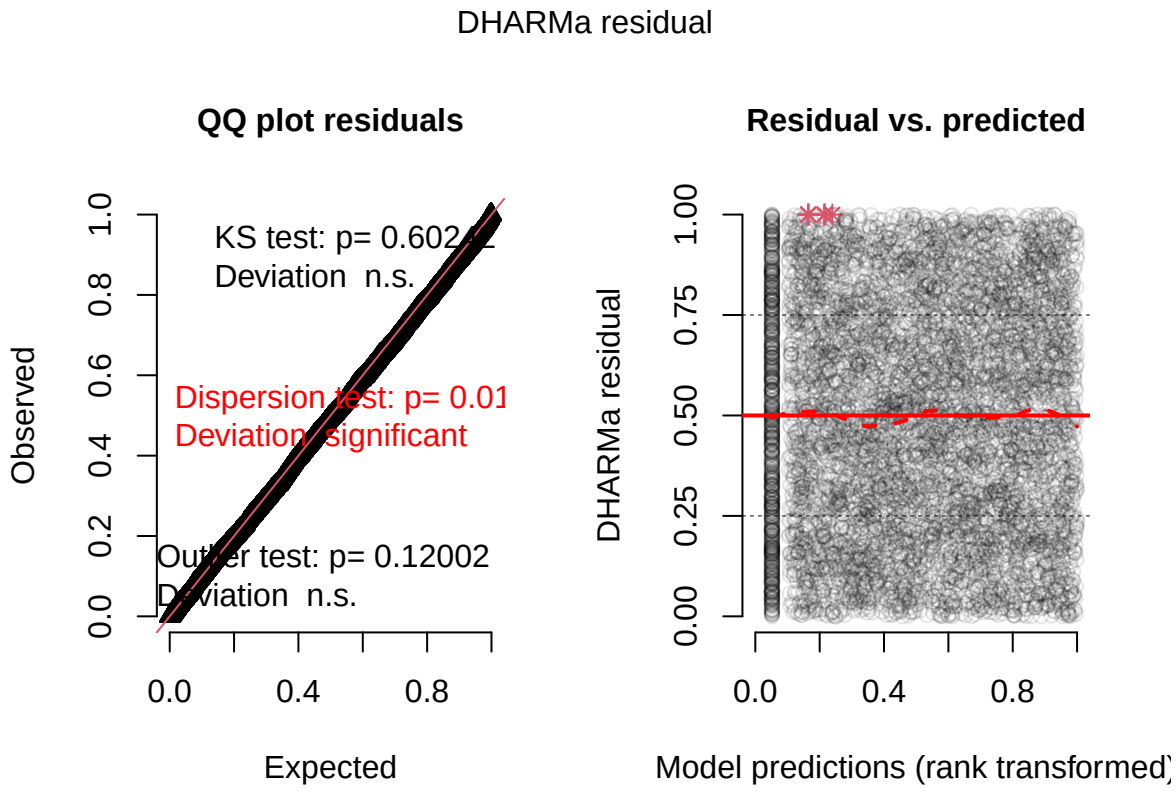
```
## Skipping install of 'mgcViz' from a github remote, the SHA1 (ad85487b) has not changed since last ins
##   Use `force = TRUE` to force installation
```

```
residuals_simulation_gam <- simulateResiduals(fittedModel = gam_model, n = 250)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## Registered S3 method overwritten by 'mgcViz':
##   method from
##   +.gg   GGally
```

```
# Plotting the simulated residuals
plot(residuals_simulation_gam)
```



Q-Q Plot The Kolmogorov-Smirnov (KS) test yields a p-value of 0.57946, suggesting there is no significant deviation from the expected uniform distribution, which means the model's residuals are distributed as expected. The dispersion test has a p-value of 0.056, which is just above the traditional alpha level of 0.05. This suggests that there might be a slight issue with the model's dispersion, but it's not statistically significant at the 0.05 level.

Residual vs Predicted The residuals should be randomly scattered around the centerline (0). Here, we see a wide spread of residuals, which is common for binomial outcomes.The red curve represents the trend of the residuals. In a perfect scenario, this would be a flat line at 0. However, in this plot, the red curve indicates there might be a pattern in the residuals that the model is not capturing. Specifically, the residuals appear to fan out as the predicted values increase, which might indicate heteroscedasticity (non-constant variance).

The model does not show significant issues with the overall distribution of residuals, as per the KS test, but

there is evidence of outliers as indicated by the outlier test. The pattern in the residuals vs. predicted plot suggests potential model misspecification or that the variance of the residuals may not be constant across the range of predictions.

Generalized Linear Mixed Models (GLMM)

Our data have multiple measurements (transit and stellar effects) taken from the same stars, which are nested within each other (i.e., measurements within stars).GLMMs are suitable when the data has a hierarchical or nested structure. Let's choose fixed and random effects for our model.

Fixed Effects:

koi_period: The interval between consecutive planetary transits, a characteristic of the planet's orbit. koi_impact: The impact parameter of the transit, relating to the path the planet takes across the stellar disc. koi_duration: The duration of the observed transits, indicating how long the transit event lasts. koi_depth: The fraction of stellar flux lost at the minimum of the planetary transit, indicative of the size of the planet relative to the star. koi_prad: The radius of the planet, likely a key determinant of the transit depth. koi_teq: The equilibrium temperature of the planet, which could relate to the likelihood of the planet being confirmed. koi_insol: Insolation flux, similar to koi_teq, could affect planet confirmation. koi_model_snr: The signal-to-noise ratio of the transit detection.

Random Effects:

koi_steff, koi_slogg, koi_smet, koi_srad, koi_smass - (these parameters are measured with error), these could be treated as random effects.

Let's incorporate statistically significant variables in our model

```
set.seed(123)
glmm_model <- glmmTMB(koi_disposition ~ koi_impact+koi_teq+koi_prad+koi_period+koi_time0bk+
                          koi_duration+koi_depth+koi_model_snr+
                      (1 | koi_slogg) +
                      (1 | koi_srad),
                    family = binomial(link = "logit"), data = data)

summary(glmm_model)
```

```
##  Family: binomial  ( logit )
## Formula:
## koi_disposition ~ koi_impact + koi_teq + koi_prad + koi_period +
##     koi_time0bk + koi_duration + koi_depth + koi_model_snr +
##     (1 | koi_slogg) + (1 | koi_srad)
## Data: data
##
##      AIC      BIC   logLik deviance df.resid
##   5137.4   5216.6  -2557.7   5115.4     9908
##
## Random effects:
##
## Conditional model:
##  Groups    Name        Variance Std.Dev.
##  koi_slogg (Intercept) 10.25    3.202
##  koi_srad  (Intercept) 18.43    4.293
## Number of obs: 9919, groups:  koi_slogg, 1209; koi_srad, 1670
##
## Conditional model:
```

```
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.383e+01  6.719e-01  20.587  < 2e-16 ***
## koi_impact    -1.865e+00  2.410e-01  -7.741 9.88e-15 ***
## koi_teq       -8.088e-03  3.331e-04 -24.282  < 2e-16 ***
## koi_prad      -2.335e-01  1.828e-02 -12.773  < 2e-16 ***
## koi_period    -2.483e-02  1.734e-03 -14.316  < 2e-16 ***
## koi_time0bk    7.233e-03  1.893e-03   3.820 0.000133 ***
## koi_duration  -4.798e-01  3.343e-02 -14.354  < 2e-16 ***
## koi_depth     -2.113e-04  2.658e-05  -7.949 1.88e-15 ***
## koi_model_snr  3.151e-03  4.008e-04   7.862 3.78e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(c("AIC GLMM " = AIC(glmm_model), "AIC GAM" = AIC(gam_model)))
```

```
## AIC GLMM    AIC GAM
##  5137.390  3786.538
```

```
# Assuming testData is your hold-out dataset
predicted_probabilities_glmm <- predict(glmm_model, newdata = testData, type = "response")
# Convert probabilities to binary outcomes, if necessary
predicted_outcomes_glmm <- ifelse(predicted_probabilities_glmm > 0.5, "CONFIRMED", "FALSE POSITIVE")
predicted_outcomes_glmm <- factor(predicted_outcomes_glmm, levels = levels(trainData$koi_disposition))
# Compare predicted outcomes with actual outcomes
confusionMatrix(data = predicted_outcomes_glmm, reference = testData$koi_disposition)
```

```
## Confusion Matrix and Statistics
##
##                Reference
## Prediction       FALSE POSITIVE CONFIRMED
##   FALSE POSITIVE           1451        24
##   CONFIRMED                  51      1449
##
##                Accuracy : 0.9748
##                  95% CI : (0.9685, 0.9801)
##     No Information Rate : 0.5049
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.9496
##
##  Mcnemar's Test P-Value : 0.00268
##
##             Sensitivity : 0.9660
##             Specificity : 0.9837
##          Pos Pred Value : 0.9837
##          Neg Pred Value : 0.9660
##              Prevalence : 0.5049
##          Detection Rate : 0.4877
##    Detection Prevalence : 0.4958
##       Balanced Accuracy : 0.9749
##
##        'Positive' Class : FALSE POSITIVE
##
```

The summary of our Generalized Linear Mixed Model (GLMM) with a binomial family (using logit link function) reveals several insights into the relationship between the predictors (like koi_impact, koi_teq,

koi_prad, etc.) and the binary outcome variable koi_disposition.

An accuracy of 0.97, indicating a high level of agreement between the predicted and actual koi_disposition. The Kappa statistic of 0.94 suggests that the agreement between the predicted and actual outcomes is much higher than what would be expected by chance.

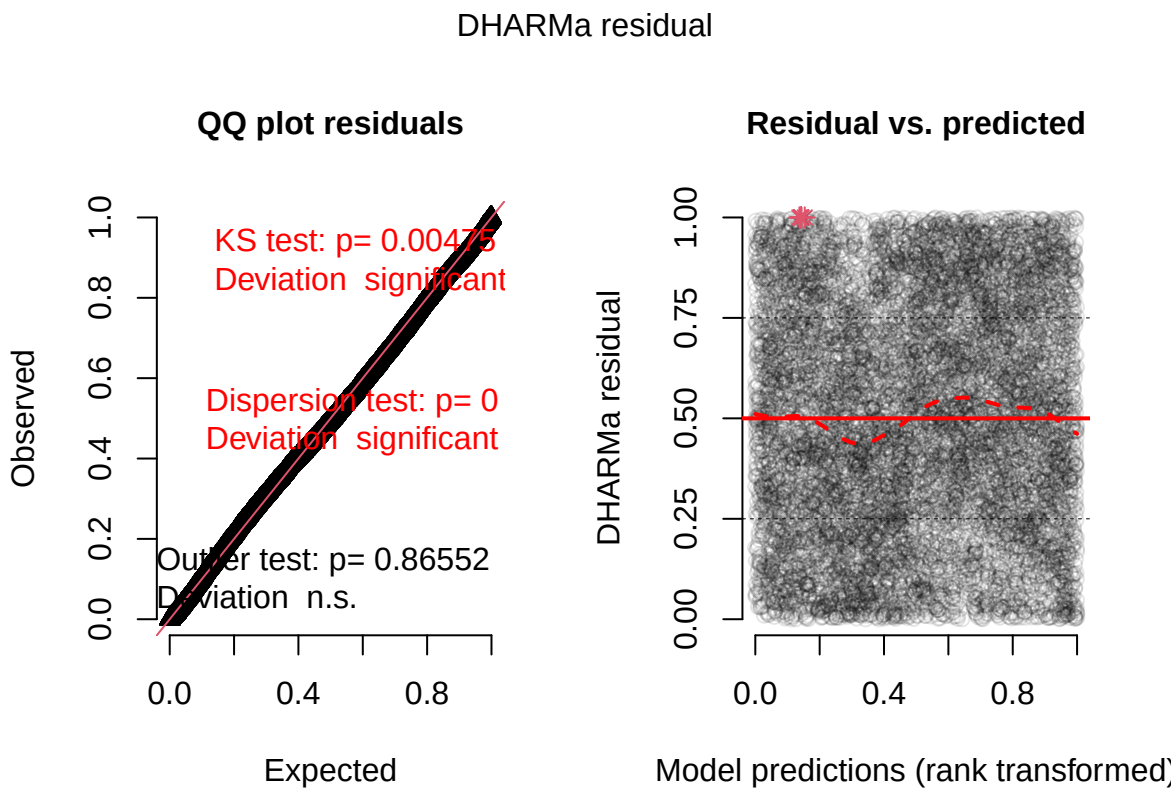
Comparing the AIC of the GLMM model (5137) with that of a Generalized Additive Model (GAM, 3903.473) suggests that the GAM might provide a better fit to the data given its lower AIC value.

Predictors like koi_impact, koi_teq, and koi_prad have significant effects on koi_disposition ($p < 0.001$ for each), indicating strong evidence against the null hypothesis of no effect.

Overall, our GLMM shows a strong predictive performance with high accuracy, sensitivity, and specificity, making it effective for predicting koi_disposition. The random effects included for koi_slogg and koi_srad seem to account for significant variability in the data, improving model fit and prediction accuracy.

Plotting Diagnostic Plots

```
set.seed(123)
residuals_simulation <- simulateResiduals(fittedModel = glmm_model, n = 250)

# Plotting the simulated residuals
plot(residuals_simulation)
```



Q-Q Plot The Kolmogorov-Smirnov (KS) test has a p-value of 0.00047, which indicates a significant deviation from the expected uniform distribution. This suggests that the residuals do not follow the expected uniform pattern, indicating potential issues with the model fit. The Dispersion Test has a p-value of 0, which is highly significant and indicates that the variability of the residuals (dispersion) is not consistent with the model assumptions. This could be a sign of overdispersion or underdispersion in the data.

he Outlier Test has a p-value of 0.03782, which also indicates significant deviation, meaning there are likely outliers in the residuals that the model does not account for well.

Residuals vs Fitted Ideally, the residuals should scatter randomly around the centerline with no pattern. However, the plot suggests there might be a pattern in the residuals, indicated by the red trend line that deviates from the center. This indicates that the residuals are not randomly distributed, which can point to issues such as misspecification of the model. The presence of possible outliers (points far from the centerline) is particularly noticeable in the top region of the plot, corroborating the result from the outlier test on the left.

The diagnostic plots suggest that the model may not fit the data adequately. There are significant deviations in the residuals' distribution from the expected uniform distribution, signs of dispersion issues, and the presence of outliers.

Let's try with choosing koi_time0bk as random effect because our data has multiple instances which belong to same category.

```
set.seed(123)
glmm_model <- glmmTMB(koi_disposition ~ koi_impact+koi_teq+koi_prad+koi_period+koi_slogg+koi_srad+
                            koi_duration+koi_depth+koi_model_snr+ (1 | koi_time0bk),
                      family = binomial(link = "logit"), data = data)

summary(glmm_model)
```

```
##  Family: binomial  ( logit )
## Formula:
## koi_disposition ~ koi_impact + koi_teq + koi_prad + koi_period +
##     koi_slogg + koi_srad + koi_duration + koi_depth + koi_model_snr +
##     (1 | koi_time0bk)
## Data: data
##
##      AIC      BIC   logLik deviance df.resid
##   2637.0   2716.3  -1307.5   2615.0     9908
##
## Random effects:
##
## Conditional model:
##  Groups      Name        Variance Std.Dev.
##  koi_time0bk (Intercept) 5293     72.75
## Number of obs: 9919, groups:  koi_time0bk, 5086
##
## Conditional model:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    4.843e+02  5.324e+01   9.097  < 2e-16 ***
## koi_impact    -1.115e+01  2.170e+00  -5.139 2.76e-07 ***
## koi_teq       -5.769e-02  4.836e-03 -11.930  < 2e-16 ***
## koi_prad      -2.083e+00  2.752e-01  -7.570 3.73e-14 ***
## koi_period    -1.511e-01  1.333e-02 -11.336  < 2e-16 ***
## koi_slogg     -8.471e+01  1.009e+01  -8.399  < 2e-16 ***
## koi_srad      -1.059e+01  1.555e+00  -6.809 9.86e-12 ***
## koi_duration  -3.640e+00  4.163e-01  -8.745  < 2e-16 ***
## koi_depth     -1.114e-03  1.357e-04  -8.213  < 2e-16 ***
## koi_model_snr  2.503e-02  2.499e-03  10.017  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
print(c("AIC GLMM " = AIC(glmm_model), "AIC GAM" = AIC(gam_model)))
```

```
## AIC GLMM    AIC GAM
##  2637.041  3786.538
```

```r
# Assuming testData is your hold-out dataset
predicted_probabilities_glmm <- predict(glmm_model, newdata = testData, type = "response")
# Convert probabilities to binary outcomes, if necessary
predicted_outcomes_glmm <- ifelse(predicted_probabilities_glmm > 0.5, "CONFIRMED", "FALSE POSITIVE")
predicted_outcomes_glmm <- factor(predicted_outcomes_glmm, levels = levels(trainData$koi_disposition))
# Compare predicted outcomes with actual outcomes
confusionMatrix(data = predicted_outcomes_glmm, reference = testData$koi_disposition)
```

```
## Confusion Matrix and Statistics
##
##                 Reference
## Prediction       FALSE POSITIVE CONFIRMED
##   FALSE POSITIVE           1502         0
##   CONFIRMED                   0      1473
##
##                Accuracy : 1
##                  95% CI : (0.9988, 1)
##     No Information Rate : 0.5049
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.0000
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 1.0000
##              Prevalence : 0.5049
##          Detection Rate : 0.5049
##    Detection Prevalence : 0.5049
##       Balanced Accuracy : 1.0000
##
##        'Positive' Class : FALSE POSITIVE
##
```

Confusion Matrix shows perfect classification with an accuracy of 1. This is highly unusual and raises suspicion, as perfect classification is rare in practice. The perfect confusion matrix and the significant coefficients for all predictors suggest that the model may be overfitting the data, or there might be issues with the data itself.

Non-Parametric Modelling

Because the distribution of the response variable remains uncertain, non-parametric modeling can be a valuable tool due to its robustness and flexibility. Additionally, the high number of predictor variables in our dataset poses a challenge to data modeling. Non-parametric models like random forest can offer feature selection during the modeling process.

K-Nearest Neighbors (KNN)

For the context of non-parametric modelling, we've used K-nearest neighbors (KNN) algorithm to predict the disposition of exoplanets.

The KNN algorithm is a fundamental algorithm in the study of machine learning, known for its simplicity and effectiveness in classification and regression efforts. In our study, we included the KNN algorithm to classify the disposition of exo planets in our data set.

As a non-parametric model, unlike parametric models such as generalized additive models, the KNN algorithm does not assume any functional form for our data set, which provides a flexible approach to our predictions.

The KNN algorithm works as follows:

1. Choosing the value of K, represents the number of nearest neighbors to make the prediction
2. Calculating the distance between each data point and it's neighbors using approaches such as Euclidean distance.
3. Find the K nearest neighbors with the predefined value of K
4. Assigning the category that is the most common in the K-nearest neighbors (for classification), or calculate the averages for K-nearest neighbors (for regression)
5. Make predictions based on the K-nearest neighbors from step 5.
6. Iterative this process for each data points

```
# Extracting the stellar and transit parameters
data_np <- data[, !names(data) %in% "koi_score"]
data_np <- data_np[, c("koi_disposition", "koi_steff", "koi_slogg", "koi_srad",
                       "koi_period", "koi_time0bk", "koi_impact",
                       "koi_duration", "koi_depth", "koi_prad",
                       "koi_teq", "koi_insol")]



# Convert 'koi_disposition' to a factor
data_np$koi_disposition <- as.factor(data_np$koi_disposition)

# Split the data into predictors (X) and the target variable (Y)
X <- data_np[, -which(names(data_np) == "koi_disposition")]  # Exclude the target variable
Y <- data_np$koi_disposition

# Set up the parameter grid for k
k_values <- seq(1, 40, by = 1)  # Adjust the range and step size as needed

# Perform grid search
accuracy_scores <- numeric(length(k_values))
for (i in 1:length(k_values)) {
  set.seed(123)  # for reproducibility
  model_knn <- knn.cv(train = X, cl = Y, k = k_values[i])
  accuracy_scores[i] <- mean(model_knn == Y)
}

# Find the best k value
best_k <- k_values[which.max(accuracy_scores)]
best_accuracy <- max(accuracy_scores)
```

```r
cat("Best k value:", best_k, "\n")
```

```
## Best k value: 1
```

```r
cat("Best accuracy:", best_accuracy, "\n")
```

```
## Best accuracy: 0.9478778
```

In our model, the best k value is 13, which resulted in an accuracy of 0.8090057.

We split the data set into training and testing in a 70:30 percent ratio. Trained the data using the KNN algorithm, and used the model to predict the testing data.

```r
data_np$koi_disposition <- as.factor(data_np$koi_disposition)

# Split the data into predictors (X) and the target variable (Y)
X <- data_np[, -which(names(data_np) == "koi_disposition")]

# Exclude the target variable
Y <- data_np$koi_disposition

# Split the data into training and testing sets
set.seed(123)
train_indices <- sample(1:nrow(data_np), 0.7 * nrow(data_np))
X_train <- X[train_indices, ]
Y_train <- Y[train_indices]
X_test <- X[-train_indices, ]
Y_test <- Y[-train_indices]

# Train the KNN model
k <- 1
model_knn <- knn(train = X_train, test = X_test, cl = Y_train, k = k)

# Predictions
predictions <- as.factor(model_knn)

# Evaluate the model
conf_matrix <- table(predictions, Y_test)
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0.9153226
```

```r
cat("Confusion Matrix:\n")
```

```
## Confusion Matrix:
```

```r
print(conf_matrix)
```

```
##                Y_test
## predictions     FALSE POSITIVE CONFIRMED
##   FALSE POSITIVE           1310        83
##   CONFIRMED                 169      1414
```

The model resulted in an accuracy of 0.7598187 This means that about 75.98% of the predictions made by the model match the actual labels in the test set.

The confusion matrix is added. It provides detailed information about the model's performance:

True Positives: The model correctly predicted 575 instances of the "CONFIRMED" class.

False Positives: The model incorrectly predicted 253 instances as "CONFIRMED" when they actually belonged to the "FALSE POSITIVE" class.

False Negatives: The model incorrectly predicted 224 instances as "FALSE POSITIVE" when they actually belonged to the "CONFIRMED" class.

True Negatives: The model correctly predicted 934 instances of the "FALSE POSITIVE" class.

From the confusion matrix, the model tends to perform better at predicting the "CONFIRMED" class, compared to the "FALSE POSITIVE" class, as the higher number of true positives and false negatives suggest.

The model demonstrated a moderate accuracy. The accuracy could be used to compare model fits.

Random Forest

We used another non-parametric method of random forest. Random forest is also a useful yet versatile machine learning algorithm, that is applicable for both classification and regression tasks.

Similar to KNN, the random forest does not make explicit assumptions about the data's distribution. Instead, it builds multiple decision trees during the training process, and combine their predictions to make a final prediction that is more robust and accurate.

The random forest algorithm is known for it's randomness and bootstrap bagging. First, the random forest randomly selects a subset of attributes for each decision trees, which introduces randomness in the feature selection and allows diversity among them. Also, the random forest algorithm uses the bootstrap technique to build multiple decision trees, and each decision tree is trained on a bootstrap sample, which is sampled from the original data set with replacement. And the out-of-bag samples could be used for model evaluation.

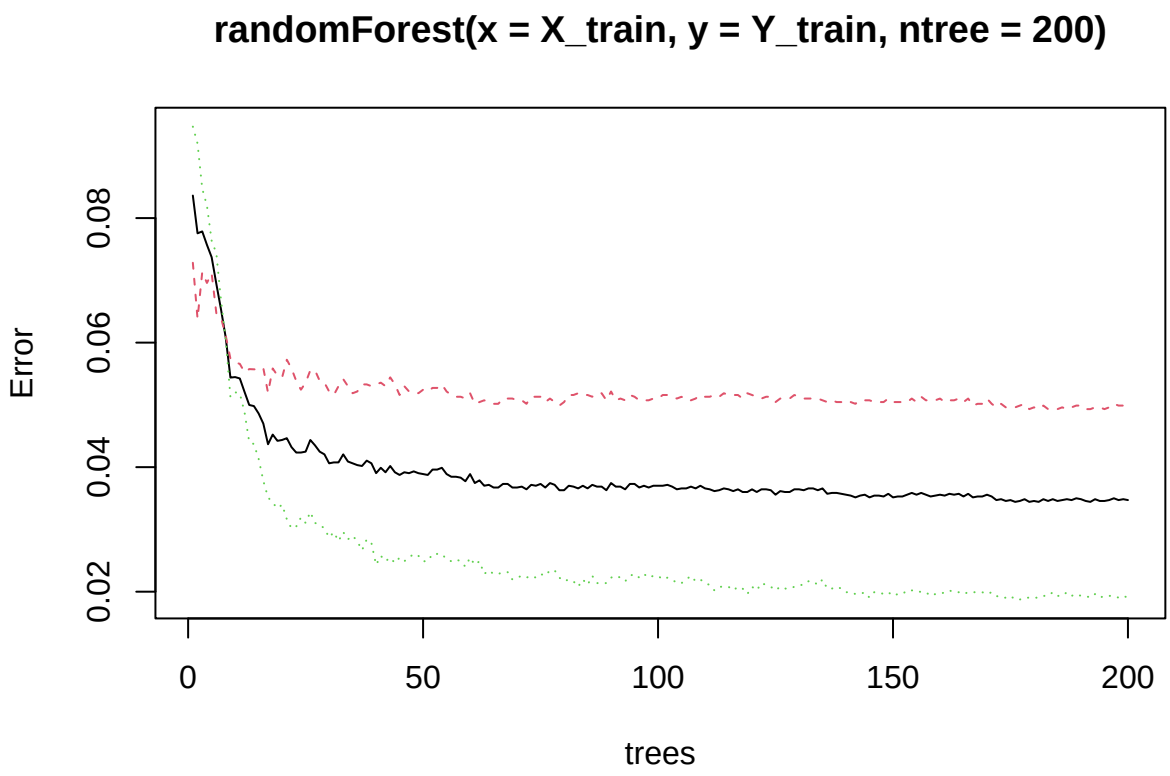
The main procedure of the random forest algorithm summarizes as follows:

1. Randomly select a subset of features from the data set
2. Randomly sample the data set (with replacement) to create bootstrap data sets
3. Train decision trees using each bootstrap sample and selected features
4. Combine the predictions from decision trees to make final regression or classifications. For regression, the predictions of all trees are averaged to make the final prediction. For classification, the prediction with the most votes is chosen as the final prediction.

The significance of random forest lies on it's robustness to over fitting, as the predictions of trees are averaged or voted, which reduces the risk of over fitting, compared to individual decision trees. Random forest also produce accurate decisions on large data sets due to it's efficiency.

```r
# Load the required library

data_np$koi_disposition <- as.factor(data_np$koi_disposition)

# Split the data into predictors (X) and the target variable (Y)
X <- data_np[, -which(names(data_np) == "koi_disposition")]
Y <- data_np$koi_disposition

# Train-Test Split
set.seed(123)
train_indices <- sample(1:nrow(data_np), 0.7 * nrow(data_np))
X_train <- X[train_indices, ]
Y_train <- Y[train_indices]
X_test <- X[-train_indices, ]
```

```r
Y_test <- Y[-train_indices]

plot(randomForest(x = X_train, y = Y_train, ntree = 200))
```



Based on the plot, the error rate decreases significantly at the beginning, suggesting that adding trees initially improves the model's performance significantly. After about 50 trees, the decrease on error rate slows down, suggests that additional trees does not return significantly on model improvement.

Random Forest Model - Performance

We've chosen a tree of 100 in our random forest model.

```r
# Train the Random Forest model
model_rf <- randomForest(x = X_train, y = Y_train, ntree = 100)

# Predictions
predictions <- predict(model_rf, X_test)

# Evaluate the model
accuracy <- mean(predictions == Y_test)
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0.9670699
```

According to the results, the random forest with number of trees equal to 100 resulted in a very well performing model with an accuracy of 0.9103726

The random forest model's accuracy of 0.9103726 is significantly higher than 0.7598187 from KNN. Therefore,

the random forest model is the more applicable non-parametric model for our data set.

Feature Importance with Random Forest

Additionally, the feature importance scores, measured by the mean decrease GINI, provides information on the relative importance of each feature in making predictions. Features with higher mean decrease GINI values contribute more to the overall predictability of our model.

```
# Feature Importance
importance <- importance(model_rf)
print(importance)
```

```
##              MeanDecreaseGini
## koi_steff            134.6872
## koi_slogg            134.1370
## koi_srad             140.7333
## koi_period           309.6917
## koi_time0bk          154.4097
## koi_impact           392.9560
## koi_duration         312.0604
## koi_depth            438.1274
## koi_prad             793.9936
## koi_teq              328.1100
## koi_insol            331.2913
```

According to the results, the variables' importance for both transit and stellar parameters (from highest to lowest) are:

Transit Parameters:

koi_prad: 527.89169 koi_depth: 260.06632 koi_impact: 241.32654 koi_period: 216.23951 koi_insol: 209.61572 koi_duration: 197.02471 koi_teq: 183.30217 koi_time0bk: 107.59483

Stellar Parameters:

koi_steff: 109.09338 koi_slogg: 101.48828 koi_srad: 95.89204

Distribution of Predictor Variables in Random Forest Model

We are interested in the distribution of predictor variables in our random forest model. We stored the predictions as a dataframe to investigate the minimum and maximum values that we are interested.

```
predictions_df <- data.frame(Predicted_Class = predictions, X_test)
str(predictions_df)
```

```
## 'data.frame':    2976 obs. of  12 variables:
##  $ Predicted_Class: Factor w/ 2 levels "FALSE POSITIVE",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ koi_steff      : int  5003 5539 5423 6335 4491 6017 6071 4254 6381 6656 ...
##  $ koi_slogg      : num  4.57 4.39 4.5 4.43 4.58 ...
##  $ koi_srad       : num  0.707 1.027 0.862 1.072 0.711 ...
##  $ koi_period     : num  2.2 25.76 1.15 20.88 2.97 ...
##  $ koi_time0bk    : num  133 140 132 147 133 ...
##  $ koi_impact     : num  0.489 0.445 0.405 0.743 1.207 ...
##  $ koi_duration   : num  3.2 14.87 1.72 8.03 1.2 ...
##  $ koi_depth      : num  345 448000 256 233000 5950 31.1 182000 910 125000 32100 ...
##  $ koi_prad       : num  1.33 74.87 1.39 66.27 24.74 ...
##  $ koi_teq        : int  1087 604 1554 737 881 1323 470 317 1851 1861 ...
##  $ koi_insol      : num  331 31.5 1382 69.7 142 ...
```

```r
#write.csv(predictions_df, file = "prediction.csv", row.names = FALSE)

summary(predictions_df)
```

```
##        Predicted_Class   koi_steff       koi_slogg        koi_srad
##  FALSE POSITIVE:1443   Min.   : 2661   Min.   :0.269   Min.   :  0.116
##  CONFIRMED     :1533   1st Qu.: 5289   1st Qu.:4.244   1st Qu.:  0.827
##                        Median : 5740   Median :4.442   Median :  0.992
##                        Mean   : 5663   Mean   :4.331   Mean   :  1.707
##                        3rd Qu.: 6077   3rd Qu.:4.548   3rd Qu.:  1.298
##                        Max.   :11076   Max.   :5.283   Max.   :150.091
##    koi_period         koi_time0bk      koi_impact       koi_duration
##  Min.   :   0.3069   Min.   :121.1   Min.   : 0.0000   Min.   : 0.3969
##  1st Qu.:   2.3667   1st Qu.:132.6   1st Qu.: 0.2260   1st Qu.: 2.4290
##  Median :   7.8890   Median :136.1   Median : 0.5910   Median : 3.7698
##  Mean   :  31.7531   Mean   :155.0   Mean   : 0.6226   Mean   : 5.4321
##  3rd Qu.:  22.2442   3rd Qu.:161.0   3rd Qu.: 0.8930   3rd Qu.: 5.9321
##  Max.   :1071.2326   Max.   :589.7   Max.   :13.7150   Max.   :90.9500
##    koi_depth           koi_prad           koi_teq        koi_insol
##  Min.   :     8.1   Min.   :   0.140   Min.   :  129   Min.   :       0
##  1st Qu.:   189.8   1st Qu.:   1.498   1st Qu.:  616   1st Qu.:      34
##  Median :   520.0   Median :   2.520   Median :  927   Median :     175
##  Mean   : 28111.5   Mean   :  27.895   Mean   : 1151   Mean   :   17239
##  3rd Qu.:  1902.5   3rd Qu.:  17.805   3rd Qu.: 1450   3rd Qu.:    1044
##  Max.   :922000.0   Max.   :2912.480   Max.   :14667   Max.   :10947555
```

Training and Testing Scores with Random Forest

We are interested in investigating the training and testing scores with random forest model with different training sizes.

```r
# Specify different training sizes
training_sizes <- seq(0.1, 0.9, by = 0.1)  # Example: Training sizes from 10% to 90%

scores <- matrix(NA, nrow = length(training_sizes), ncol = 3)  # 3 columns for train_size, train_score,
seq_along(training_sizes)
```

```
## [1] 1 2 3 4 5 6 7 8 9
```

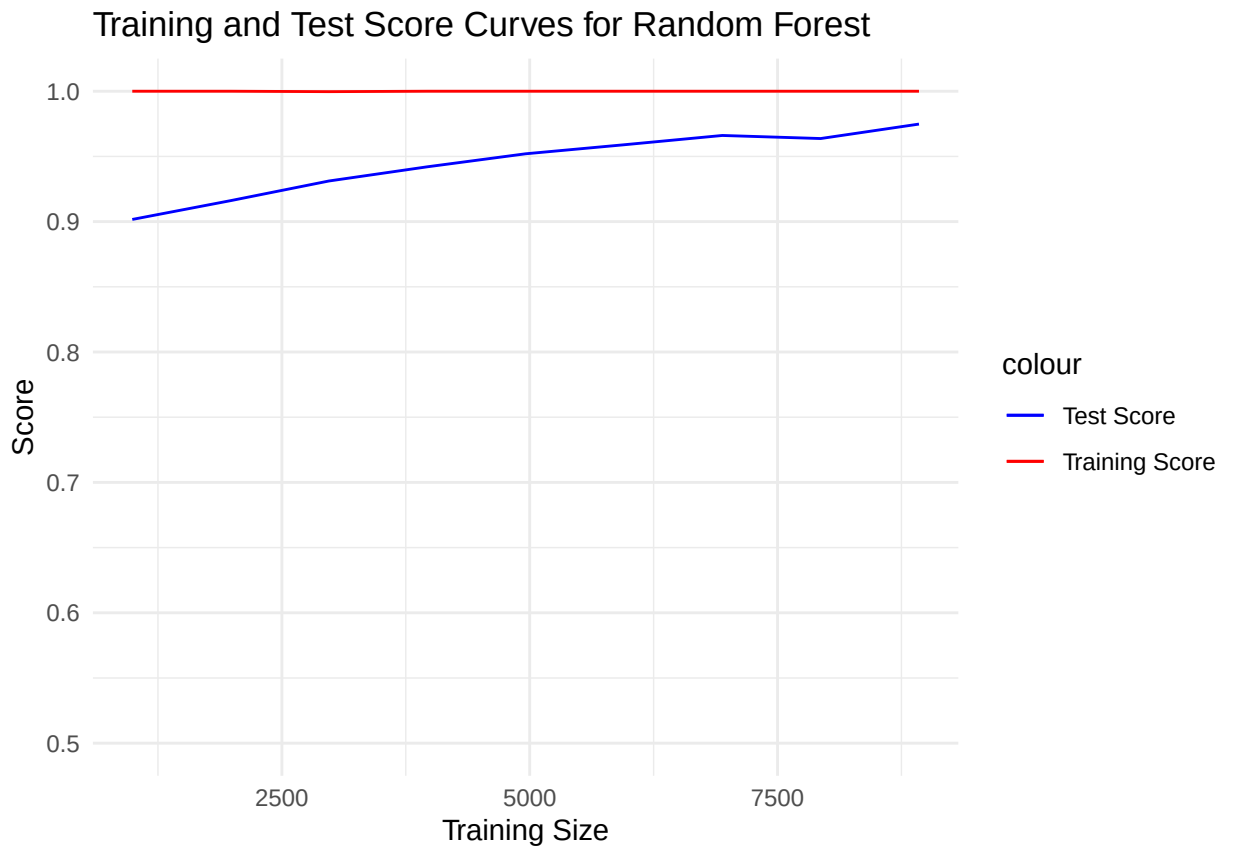
```r
# Loop through different training sizes
for (i in seq_along(training_sizes)) {

  # Train model
  X <- data_np[, -which(names(data_np) == "koi_disposition")]
  Y <- data_np$koi_disposition

  # Train-Test Split
  set.seed(123)
  train_size <- sample(1:nrow(data_np), size = floor(training_sizes[i] * nrow(data_np)))
  X_train <- X[train_size, ]
  Y_train <- Y[train_size]
  X_test <- X[-train_size, ]
  Y_test <- Y[-train_size]

  # Train the Random Forest model
```

```r
  model_rf <- randomForest(x = X_train, y = Y_train, ntree = 100)

  # Calculate training and test scores
  train_pred <- predict(model_rf, newdata = X_train)

  # Calculate accuracy on the training set
  train_accuracy <- mean(train_pred == Y_train)

  test_pred <- predict(model_rf, newdata = X_test)
  test_score <- mean(test_pred == Y_test)

  # Store scores in the matrix
  print(floor(training_sizes[i] * nrow(data_np)))
  print(train_accuracy)
  print(test_score)
  scores[i, ] <- c(floor(training_sizes[i] * nrow(data_np)), train_accuracy, test_score)
}
```

```
## [1] 991
## [1] 1
## [1] 0.9016577
## [1] 1983
## [1] 1
## [1] 0.9160786
## [1] 2975
## [1] 0.9996639
## [1] 0.9311636
## [1] 3967
## [1] 1
## [1] 0.9420363
## [1] 4959
## [1] 1
## [1] 0.9520161
## [1] 5951
## [1] 1
## [1] 0.9589214
## [1] 6943
## [1] 1
## [1] 0.9660618
## [1] 7935
## [1] 1
## [1] 0.9637097
## [1] 8927
## [1] 1
## [1] 0.9747984
```

```r
# Convert scores to a data frame
scores_df <- as.data.frame(scores)
names(scores_df) <- c("train_size", "train_score", "test_score")

# Plot training and test scores
ggplot(scores_df, aes(x = train_size)) +
  geom_line(aes(y = train_score, color = "Training Score")) +
  geom_line(aes(y = test_score, color = "Test Score")) +
```

```
  scale_color_manual(values = c("blue", "red")) +
  labs(x = "Training Size", y = "Score") +
  ggtitle("Training and Test Score Curves for Random Forest") +
  theme_minimal()+
  scale_y_continuous(limits = c(0.50, 1))
```



On the training and test score curve, the x-axis shows the training size, which ranges from 10% to 90% of the dataset.The training score is set as 1 for comparison, and the test Score represents the model's accuracy on the test set. It's slightly lower than the training score, which suggests that the model is generalizing a relatively well predictability, and the overall predictability increases over introducing more data points (training size).

Model Comparison

We've trained and accessed several parametric and non-parametric models.

Generalized Additive Model (GAM): Achieved an accuracy of approximately 89.59%. Generalized Linear Mixed Model (GLMM): Achieved a high accuracy of approximately 96.98%. K-Nearest Neighbors (KNN): Achieved an accuracy of approximately 75.98%. Random Forest: Achieved the highest accuracy among non-parametric models, approximately 91.04%.

```
metadata <- read.csv("modelcomparison.csv", head = T, check.names = F, fileEncoding = "UTF-8")
```

```
## Warning in read.table(file = file, header = header, sep = sep, quote = quote, :
## invalid input found on input connection 'modelcomparison.csv'
```

```
## Warning in read.table(file = file, header = header, sep = sep, quote = quote, :
```

```
## incomplete final line found by readTableHeader on 'modelcomparison.csv'
```

```
kable(metadata, caption = "Model Accuracy")
```

Table 1: Model Accuracy

Model	Accuracy	95% CI	Kappa	Sensitivity	Specificity	PPV	NPV	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy	AUC
Regularization	0.8866	(0.877, 0.8958)	0.7734	0.8492	0.9246	0.9196	0.8580	0.5037	0.4277	0.4652	0.8869	0.9415
GLM	0.8871	(0.8774, 0.8962)	0.7743	0.8496	0.9251	0.9200	0.8584	0.5037	0.4280	0.4652	0.8874	0.9415
GAM	0.8866	(0.877, 0.8958)	0.7734	0.8492	0.9246	0.9196	0.8580	0.5037	0.4277	0.4652	0.8869	0.9487
GLMM	0.9814	(0.977, 0.9852)	0.9628	0.9764	0.9865	0.9865	0.9763	0.5037	0.4918	0.4985	0.9814	NA

Comparing these accuracy scores and plots above, we can observe that Random Forest gives the best fit for the dataset. The GAM also performed well, while the KNN model had the lowest accuracy among the models evaluated.

RESULTS FOR RQ 1

The Random Forest models gives best fit for predicting the disposition of exoplanets, suggesting that they could potentially be the most effective models for this task based on our data.

Our findings indicate that both the characteristics of the observed parameters and stellar characteristics serve as crucial factors in classifying Kepler Objects of Interest as actual planets. Among the observed parameters, koi_prad (Planetary Radius), koi_depth (Transit Depth), and koi_impact (Impact Parameter) emerge as the most influential variables. Particularly, koi_prad (Planetary Radius) stands out as the most significant predictor for distinguishing actual planets.

According to the result, with the random forest model, the minimum and maximum values of predictor variables are as follows:

Stellar Parameters:

Variable Min. Max.

koi_steff	2661	15896	koi_slogg 0.269
5.283	koi_srad	0.116	138.056
Variable	Min.	Max.	

koi_period 0.329 670.646 koi_time0bk 120.6 746.2 koi_impact 0.000 15.329 koi_duration 0.105 138.540 koi_depth 4.5 922000.0 koi_prad 0.180 2674.690 koi_teq 134 14667 koi_insol 0 10947555

Similarly, when considering stellar parameters, koi_steff (Stellar Effective Temperature), koi_slogg (Stellar Surface Gravity), and koi_srad (Stellar Radius) demonstrate comparable effectiveness in modeling efforts, with koi_steff being the most influential among them.

To gauge Earth-size+ planets in the habitable zone ("Goldilocks") across various star types

"Goldilocks planets," also known as "Goldilocks zone planets" or "habitable planets," refer to exoplanets that orbit within the habitable zone of a star. The habitable zone is the region around a star where conditions are just right for liquid water to exist on the surface, which is considered a crucial ingredient for life as we know it.[2]

Checking habitability based on the presence of water The primary criterion for an Earth-like habitable planet is the presence of liquid water. This necessitates an equilibrium temperature (represented as koi_teq in astronomical terms) within the range of 273.2 Kelvin to 373.2 Kelvin. This step is performed just for confirmed exoplanets.

```
#this part can be removed when compiling
data_r2 <- project_data[project_data$koi_disposition != "CANDIDATE",]
data_r2$koi_disposition <- factor(data_r2$koi_disposition, levels=c("CONFIRMED","FALSE POSITIVE"))
##---- Discard any rows which has null value or NA value
complete_rows <- complete.cases(data_r2)

##---- Select the subset which only includes these complete rows
data_r2 <- data_r2[complete_rows, ]
```

```
confirmed_data <- data_r2[data_r2$koi_disposition == "CONFIRMED",]
```

```
# Create a new column 'goldilocks_temp' based on conditions
confirmed_data$goldilocks_temp <- (273.2 <= confirmed_data$koi_teq) & (confirmed_data$koi_teq <= 373.2)

# Initialize a list to store goldilocks counts
goldilocks_counts <- list(temp = list())

# Calculate counts for each temperature range
goldilocks_counts$temp$too_cold <- sum(confirmed_data$koi_teq <= 273.2)
goldilocks_counts$temp$just_right <- sum(confirmed_data$goldilocks_temp == TRUE)
goldilocks_counts$temp$too_hot <- sum(confirmed_data$koi_teq >= 373.2)

# Print the counts
for (key in names(goldilocks_counts$temp)) {
  value <- goldilocks_counts$temp[[key]]
  percentage <- 100 * value / nrow(confirmed_data)
  cat(sprintf("Exoplanets that are %-10s: %4d (%5.2f%%)\n", key, value, percentage))
}
```

```
## Exoplanets that are too_cold  :   52 ( 1.91%)
## Exoplanets that are just_right:  135 ( 4.96%)
## Exoplanets that are too_hot   : 2537 (93.14%)
```

```
counts_df <- data.frame(
  Type = c("too_cold", "just_right", "too_hot"),
  Count = c(goldilocks_counts$temp$too_cold, goldilocks_counts$temp$just_right, goldilocks_counts$temp$
)

# Print the data frame
#print(counts_df)

percentages <- round((counts_df$Count / sum(counts_df$Count)) * 100, 2)
```

```
barplot(counts_df$Count,names.arg = counts_df$Type , col = "skyblue",
        main = "Count of Planets", x_lab = "temperature range", y_lab = "count")
```

```
## Warning in plot.window(xlim, ylim, log = log, ...): "x_lab" is not a graphical
## parameter
```

```
## Warning in plot.window(xlim, ylim, log = log, ...): "y_lab" is not a graphical
## parameter
```

```
## Warning in axis(if (horiz) 2 else 1, at = at.l, labels = names.arg, lty =
## axis.lty, : "x_lab" is not a graphical parameter
```

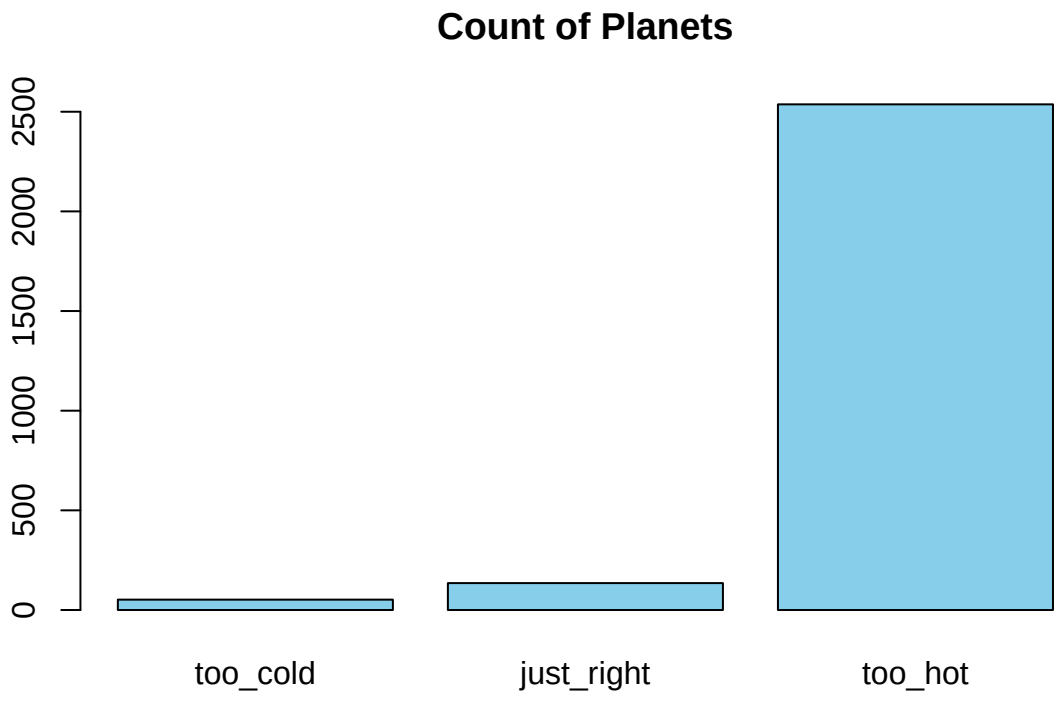
```
## Warning in axis(if (horiz) 2 else 1, at = at.l, labels = names.arg, lty =
## axis.lty, : "y_lab" is not a graphical parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "x_lab" is not a graphical parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "y_lab" is not a graphical parameter
```

```
## Warning in axis(if (horiz) 1 else 2, cex.axis = cex.axis, ...): "x_lab" is not
## a graphical parameter
```

```
## Warning in axis(if (horiz) 1 else 2, cex.axis = cex.axis, ...): "y_lab" is not
## a graphical parameter
```




```
# text(x = barplot(counts_df$Count) - 0.2, y = counts_df$Count + 0.5,
#
```

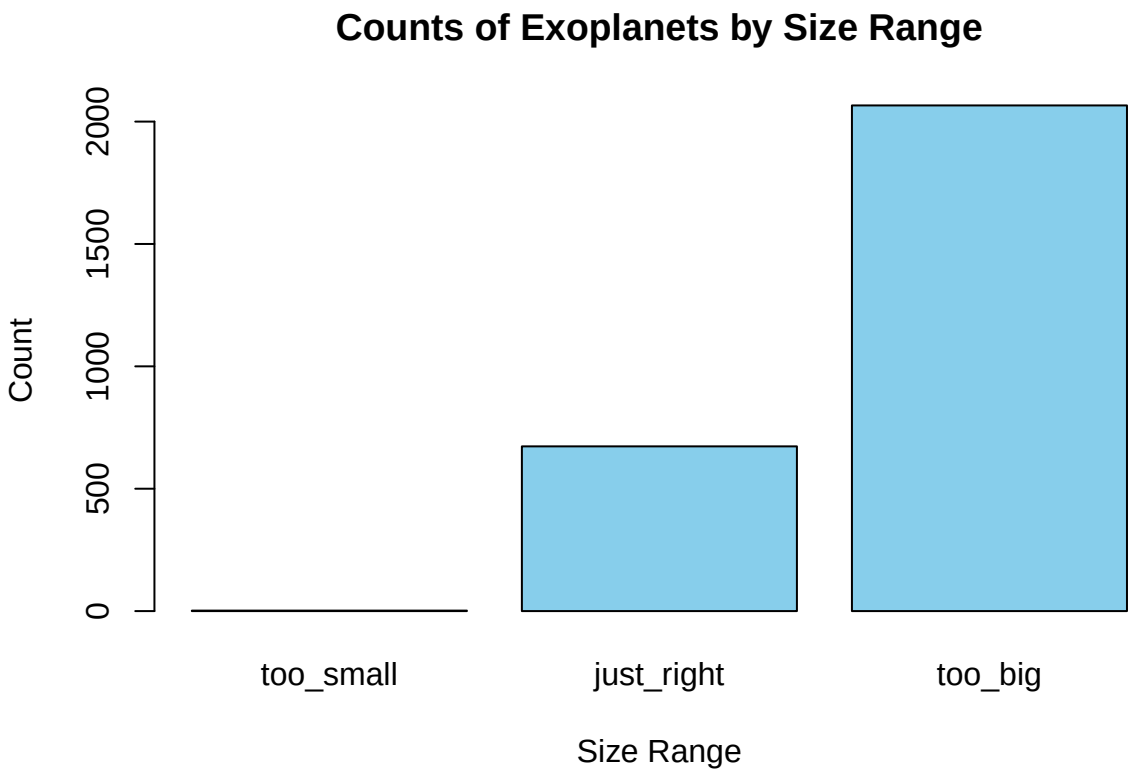
Based on the above results, we can infer that only 4.98% of exoplanents have habitable temperature.

Checking habitability based on the planetary radius Upon researching, we have identified that if the range of radii of exoplanet range is between 0.5 and 1.5 Earth radii, then it can be considered a potential habitable planet. The koi_prad prad column is used for this step.[3]

```
# Create a new column 'goldilocks_size' based on conditions
confirmed_data$goldilocks_size <- (0.5 <= confirmed_data$koi_prad) & (confirmed_data$koi_prad <= 1.5)

# Initialize a list to store goldilocks counts
goldilocks_counts <- list(size = list())

# Calculate counts for each size range
goldilocks_counts$size$too_small <- sum(confirmed_data$koi_prad <= 0.5)
goldilocks_counts$size$just_right <- sum(confirmed_data$goldilocks_size == TRUE)
goldilocks_counts$size$too_big <- sum(confirmed_data$koi_prad >= 1.5)

# Print the counts
for (key in names(goldilocks_counts$size)) {
  value <- goldilocks_counts$size[[key]]
  percentage <- 100 * value / nrow(confirmed_data)
  cat(sprintf("Exoplanets that are %-10s: %4d (%5.2f%%)\n", key, value, percentage))
}
```

```
## Exoplanets that are too_small :    2 ( 0.07%)
## Exoplanets that are just_right:  673 (24.71%)
## Exoplanets that are too_big   : 2066 (75.84%)
```

```
# Visualize the results
# Extract counts and labels for plotting
counts <- unlist(goldilocks_counts$size)
labels <- names(counts)
barplot(counts, names.arg = labels, xlab = "Size Range", ylab = "Count", col = "skyblue", main = "Counts
```



Based on the above results, we can infer that only 24.73% of exoplanents have habitable planetary radii. Now we can check the combination of two parameters for determing exoplanet's habitability.

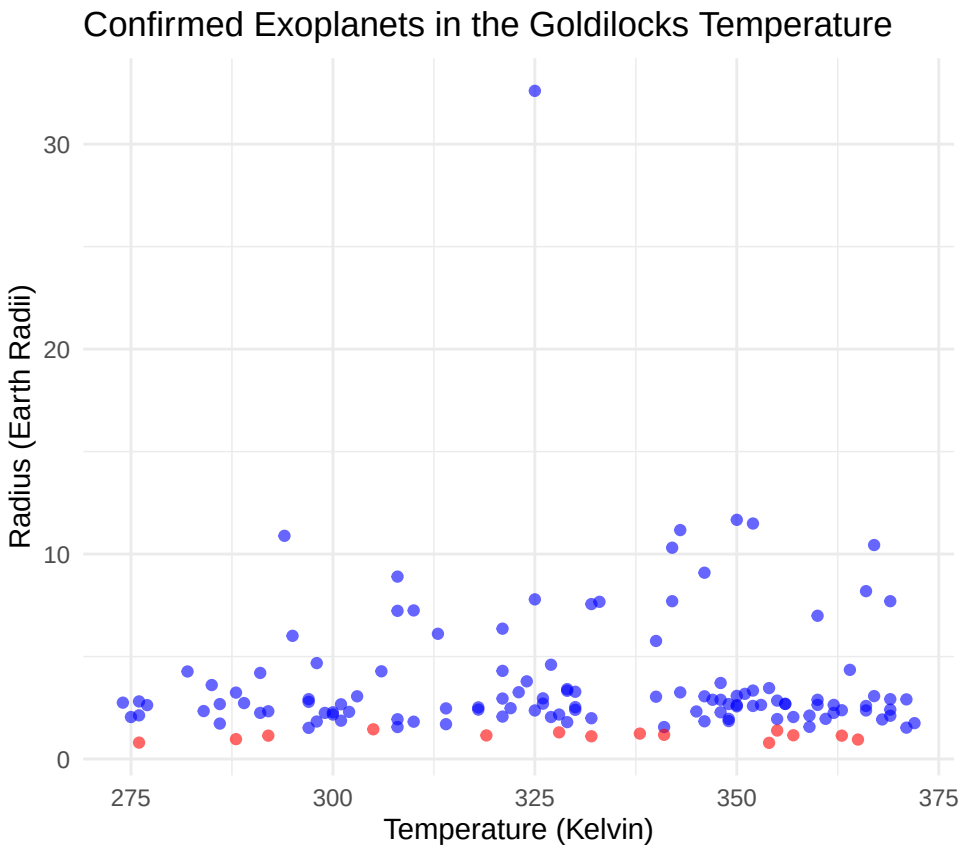
```
# Assuming you have a dataframe named 'dataset' in R similar to the one used in the Python code

# Create a new column 'goldilocks' based on conditions
confirmed_data$goldilocks <- (confirmed_data$goldilocks_temp == TRUE) & (confirmed_data$goldilocks_size

# Initialize a list to store goldilocks counts
goldilocks_counts <- list(combined = list())

# Calculate counts for "just right" exoplanets
goldilocks_counts$combined$just_right <- sum(confirmed_data$goldilocks == TRUE)
# Print the counts
for (key in names(goldilocks_counts)) {
  value <- goldilocks_counts[[key]]$just_right
  percentage <- 100 * value / nrow(confirmed_data)
  cat(sprintf("Exoplanets that are 'just right' %-10s: %4d (%5.2f%%)\n", key, value, percentage))
}
```

```
## Exoplanets that are 'just right' combined  :   14 ( 0.51%)
```

Out of 2730, 14 planets are considered to be potential habitable exoplanets based equilibrium temperature and planetary radii.

```
# Set the plot size
```

```r
options(repr.plot.width=20, repr.plot.height=10)
potentially_habitable <- subset(confirmed_data, goldilocks_temp == TRUE )
# Create the scatter plot
ggplot(data = potentially_habitable , aes(x = koi_teq, y = koi_prad, color = goldilocks)) +
  geom_point(alpha = 0.6) +
  scale_size(range = c(20, 400)) +
  scale_color_manual(values = c("blue", "red"), name = "Goldilocks Temp",
                     labels = c("False", "True")) +
  labs(title = "Confirmed Exoplanets in the Goldilocks Temperature",
       x = "Temperature (Kelvin)",
       y = "Radius (Earth Radii)") +
  theme_minimal()
```



Visualise the confirmed exoplanets

```r
# Set options to display all rows and columns
options(repr.matrix.max.rows=Inf, repr.matrix.max.cols=Inf)

# Calculate and print the number of potentially habitable exoplanets

potentially_habitable <- subset(confirmed_data, goldilocks == TRUE )
cat("Number of potentially habitable exoplanets: ", nrow(potentially_habitable), "\n")
```

```
## Number of potentially habitable exoplanets:  14
```

```r
# Print the names of potentially habitable exoplanets
cat("Names of potentially habitable exoplanets: ", paste(potentially_habitable$kepler_name, collapse =
```

```
## Names of potentially habitable exoplanets:  Kepler-54 d, Kepler-249 d, Kepler-296 b, Kepler-367 c, K
```

```
# Display the potentially habitable exoplanets
#print(potentially_habitable)
```

All the red dots in the scatter plot represent confirmed exoplanets against radius and temperature.

```
# Set options to display all rows and columns
options(repr.matrix.max.rows=Inf, repr.matrix.max.cols=Inf)

# Calculate and print the number of potentially habitable exoplanets
potentially_habitable <- subset(confirmed_data, goldilocks == TRUE)
cat("Number of potentially habitable exoplanets: ", nrow(potentially_habitable), "\n")
```

```
## Number of potentially habitable exoplanets:  14
```

```
# Print the names of potentially habitable exoplanets
cat("Names of potentially habitable exoplanets: ", paste(potentially_habitable$kepler_name, collapse = "
```

```
## Names of potentially habitable exoplanets:  Kepler-54 d, Kepler-249 d, Kepler-296 b, Kepler-367 c, Ke
```

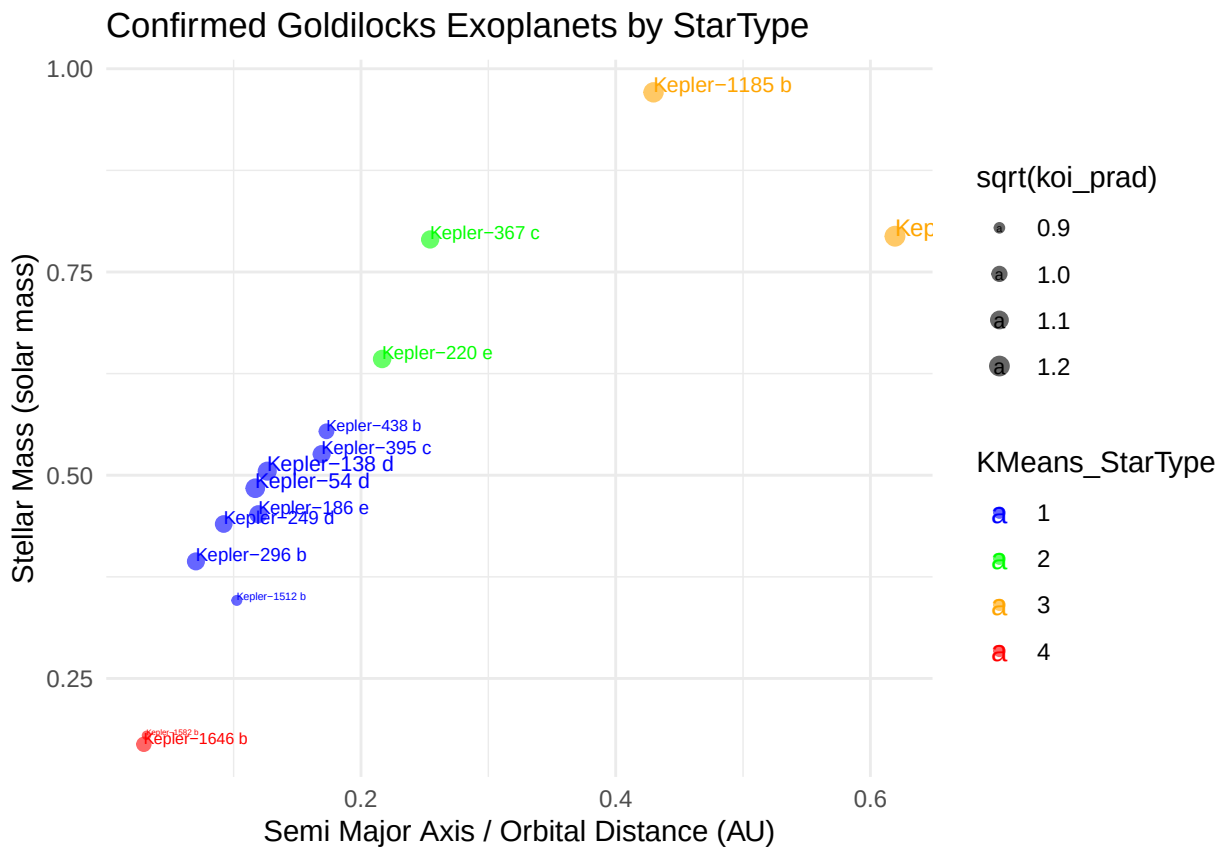
Names of potentially habitable exoplanets: Kepler-54 d, Kepler-249 d, Kepler-296 b, Kepler-367 c, Kepler-395 c, Kepler-138 d, Kepler-186 e, Kepler-220 e, Kepler-1582 b, Kepler-438 b, Kepler-1512 b, Kepler-1126 c, Kepler-1185 b, Kepler-1646 b

```
# Filter the dataset to include only rows where 'goldilocks' is True
df <- subset(confirmed_data, goldilocks == TRUE)

# Perform KMeans clustering
df$KMeans_StarType <- kmeans(df[, c('koi_sma', 'koi_smass')], centers = 4)$cluster
print(df$KMeans_StarType)
```

Visualise the confirmed exoplanets against Stellar Mass and Orbital Distance

```
##  [1] 1 1 1 2 1 1 1 2 4 1 1 3 3 4
```

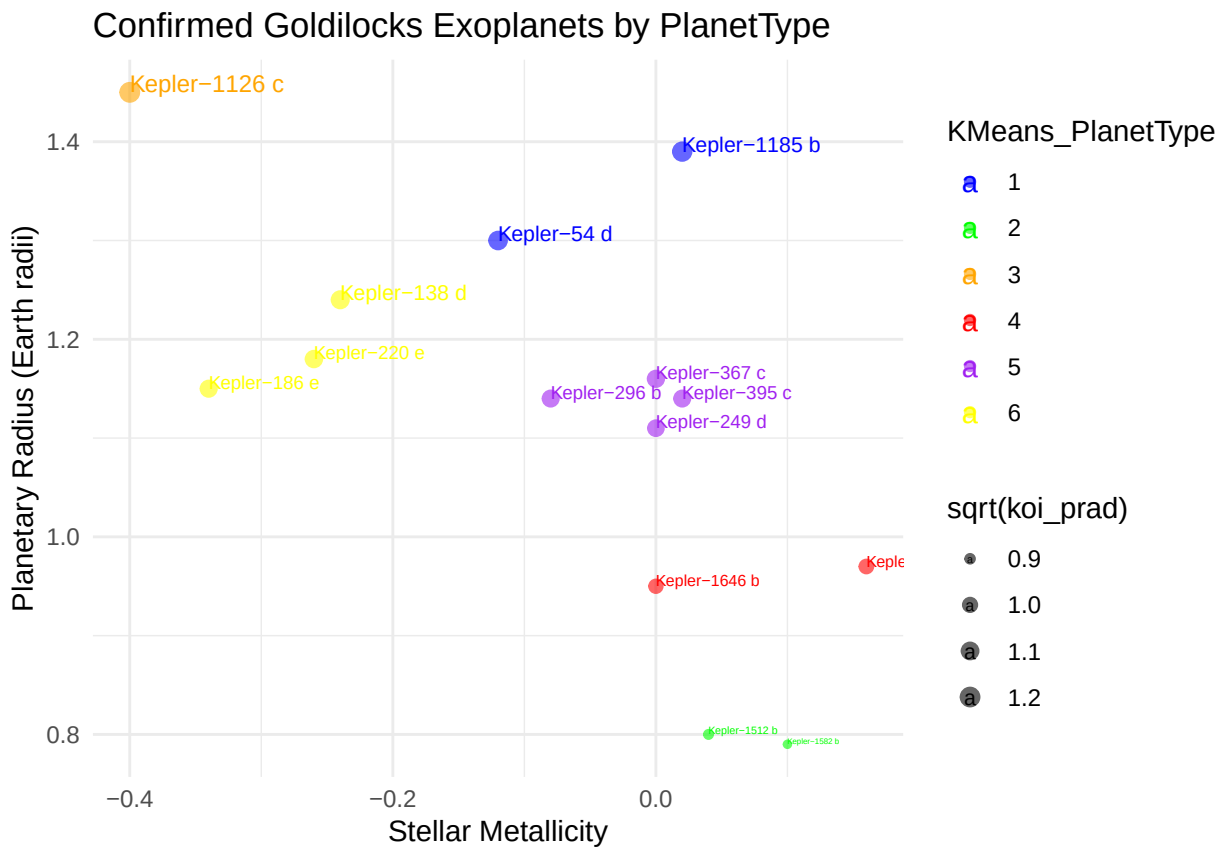
```
# Create the scatter plot
p <- ggplot(data = df, aes(x = koi_sma, y = koi_smass, size = sqrt(koi_prad), color = factor(KMeans_Sta
  geom_point(alpha = 0.6) +
  scale_size(range = c(1, 3)) +
  scale_color_manual(values = c("blue", "green", "orange", "red"), name = "KMeans_StarType") +
  geom_text(aes(label = kepler_name), hjust = 0, vjust = 0) +
  labs(title = "Confirmed Goldilocks Exoplanets by StarType",
       x = "Semi Major Axis / Orbital Distance (AU)",
       y = "Stellar Mass (solar mass)") +
  theme_minimal()

# Print the plot
print(p)
```



Based on the above plot, we can there are majorly four clusters of exoplanets. The largest cluster has been shown above in red. Stellar Mass ranges from 0.3 to 0.62 approx and Orbital Distance ranges from 0.05 -0.3 for the largest cluster.

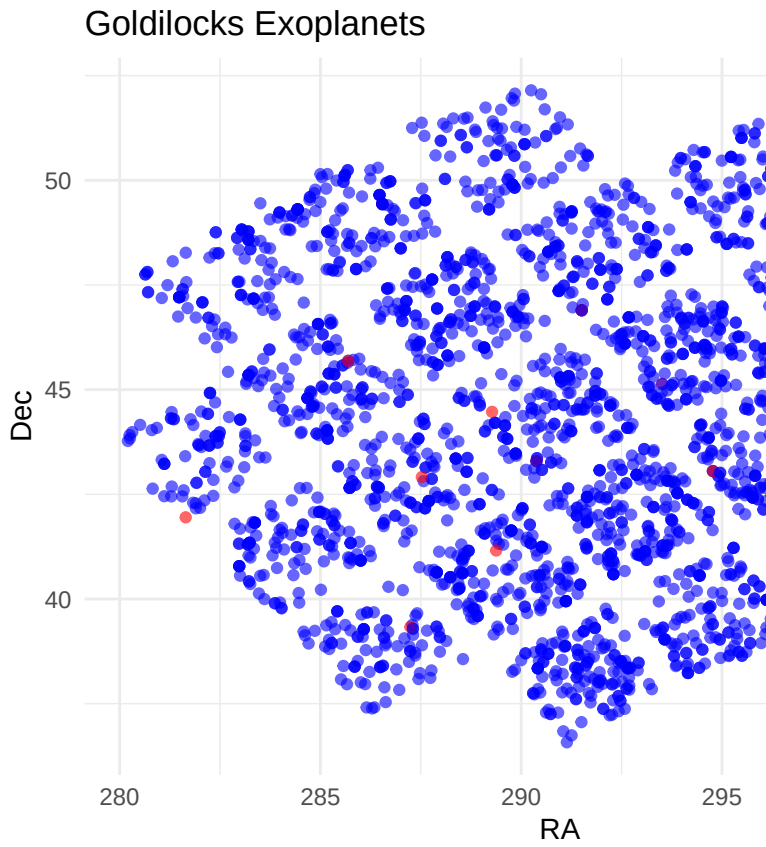
Visualise the confirmed exoplanets against Planetary Radius vs Stellar Metallicity Plotting the planetary radius against stellar metallicity can offer insights into the composition of planets. High-metallicity stars are more inclined to form rocky planets, as opposed to water/ice worlds or carbon planets.

```
df$KMeans_PlanetType <- kmeans(df[, c('koi_smet', 'koi_prad')], centers = 6)$cluster

# Create the scatter plot
p <- ggplot(data = df, aes(x = koi_smet, y = koi_prad, size = sqrt(koi_prad), color = factor(KMeans_Pla
  geom_point(alpha = 0.6) +
  scale_size(range = c(1, 3)) +  # Adjust the range of sizes as needed
  scale_color_manual(values = c("blue", "green", "orange", "red", "purple", "yellow"), name = "KMeans_P
  geom_text(aes(label = kepler_name), hjust = 0, vjust = 0) +
  labs(title = "Confirmed Goldilocks Exoplanets by PlanetType",
       x = "Stellar Metallicity",
       y = "Planetary Radius (Earth radii)") +
  theme_minimal()

# Print the plot
print(p)
```



Based on the above plot, we can there are majorly 6 clusters of exoplanets based on Planetary Radius vs Stellar Metallicity. The largest cluster has been shown above in green.

```
# Convert goldilocks to a factor
confirmed_data$goldilocks <- factor(confirmed_data$goldilocks)

# Create the scatter plot
p <- ggplot(data = confirmed_data, aes(x = ra, y = dec, color = goldilocks)) +
  geom_point(alpha = 0.6) +
  scale_size_manual(values = c(20, 200)) +  # Adjust the size range as needed
  scale_color_manual(values = c("blue", "red"), name = "Goldilocks") +  # Discrete color scale
  labs(title = "Goldilocks Exoplanets",
       x = "RA",
       y = "Dec") +
  theme_minimal()

# Print the plot
print(p)
```



Visualise Starmap to locate exoplanents in the sky

In the above, all red ones are confirmed exoplanets.

RESULTS FOR RQ 2

The analysis of confirmed exoplanets reveals intriguing insights into their potential habitability and composition. By examining the presence of liquid water, a fundamental criterion for habitability, it was found that only 4.98% of exoplanets possess temperatures within the optimal range, crucial for sustaining liquid water on their surfaces. Furthermore, considering the planetary radii, approximately 24.73% of confirmed exoplanets fall within the potential habitable range, indicating their suitability for supporting life based on size criteria alone. Interestingly, when both temperature and radius criteria are combined, only 14 out of 2730 confirmed exoplanets are identified as potential habitable candidates, underscoring the stringent conditions necessary for habitability. Notable among these are Kepler-54 d, Kepler-249 d, and Kepler-296 b, among others. Visualizations of confirmed exoplanets against stellar mass and orbital distance reveal distinct clusters, with the largest cluster characterized by a stellar mass ranging from 0.3 to 0.62 and an orbital distance from 0.05 to 0.3. Additionally, plotting planetary radius against stellar metallicity highlights six clusters, with the largest cluster suggesting a correlation between high-metallicity stars and the prevalence of rocky planets. Lastly, a starmap depicting confirmed exoplanets in the sky showcases their spatial distribution, aiding astronomers in locating and studying these intriguing celestial bodies. These findings shed light on the potential habitability and diverse compositions of confirmed exoplanets, opening avenues for further exploration and research in the field of exoplanetary science.

To establish the correlation between the different causes for disposition values "FALSE POSITIVE", "CONFIRMED" and "CANDIDATE". Also to establish the correlation between disposition values and flag variables.

Understanding the correlations between disposition values ("FALSE POSITIVE," "CONFIRMED," and "CANDIDATE") and associated flag variables in exoplanet data allows for improved classification accuracy, enabling more reliable identification of genuine planets and false positives.[7]

```
library(dplyr)
data_set <- project_data
head(data_set,n=10)
```

```
##    loc_rowid    kepid kepoi_name  kepler_name koi_disposition koi_pdisposition
## 1          1 10797460  K00752.01 Kepler-227 b       CONFIRMED        CANDIDATE
## 2          2 10797460  K00752.02 Kepler-227 c       CONFIRMED        CANDIDATE
## 3          3 10811496  K00753.01                    CANDIDATE        CANDIDATE
## 4          4 10848459  K00754.01               FALSE POSITIVE   FALSE POSITIVE
## 5          5 10854555  K00755.01 Kepler-664 b       CONFIRMED        CANDIDATE
## 6          6 10872983  K00756.01 Kepler-228 d       CONFIRMED        CANDIDATE
## 7          7 10872983  K00756.02 Kepler-228 c       CONFIRMED        CANDIDATE
## 8          8 10872983  K00756.03 Kepler-228 b       CONFIRMED        CANDIDATE
## 9          9  6721123  K00114.01               FALSE POSITIVE   FALSE POSITIVE
## 10        10 10910878  K00757.01 Kepler-229 c       CONFIRMED        CANDIDATE
##    koi_score koi_fpflag_nt koi_fpflag_ss koi_fpflag_co koi_fpflag_ec koi_period
## 1      1.000             0             0             0             0   9.488036
## 2      0.969             0             0             0             0  54.418383
## 3      0.000             0             0             0             0  19.899140
## 4      0.000             0             1             0             0   1.736952
## 5      1.000             0             0             0             0   2.525592
## 6      1.000             0             0             0             0  11.094321
## 7      1.000             0             0             0             0   4.134435
## 8      0.992             0             0             0             0   2.566589
## 9      0.000             0             1             1             0   7.361790
## 10     1.000             0             0             0             0  16.068647
##    koi_time0bk koi_impact koi_duration koi_depth koi_prad koi_sma koi_teq
## 1     170.5387      0.146      2.95750       616     2.26  0.0853     793
## 2     162.5138      0.586      4.50700       875     2.83  0.2734     443
## 3     175.8503      0.969      1.78220     10800    14.60  0.1419     638
## 4     170.3076      1.276      2.40641      8080    33.46  0.0267    1395
## 5     171.5956      0.701      1.65450       603     2.75  0.0374    1406
## 6     171.2012      0.538      4.59450      1520     3.90  0.0992     835
## 7     172.9794      0.762      3.14020       686     2.77  0.0514    1160
## 8     179.5544      0.755      2.42900       227     1.59  0.0374    1360
## 9     132.2505      1.169      5.02200       234    39.21  0.0820    1342
## 10    173.6219      0.052      3.53470      4910     5.76  0.1158     600
##    koi_insol koi_model_snr koi_tce_plnt_num koi_tce_delivname koi_steff
## 1      93.59          35.8                1   q1_q17_dr25_tce      5455
## 2       9.11          25.8                2   q1_q17_dr25_tce      5455
## 3      39.30          76.3                1   q1_q17_dr25_tce      5853
## 4     891.96         505.6                1   q1_q17_dr25_tce      5805
## 5     926.16          40.9                1   q1_q17_dr25_tce      6031
## 6     114.81          66.5                1   q1_q17_dr25_tce      6046
## 7     427.65          40.2                2   q1_q17_dr25_tce      6046
## 8     807.74          15.0                3   q1_q17_dr25_tce      6046
## 9     767.22          47.7                1   q1_q17_dr25_tce      6227
```

```
## 10     30.75         161.9                1   q1_q17_dr25_tce      5031
##    koi_slogg koi_smet koi_srad koi_smass       ra      dec koi_kepmag
## 1      4.467     0.14    0.927     0.919 291.9342 48.14165     15.347
## 2      4.467     0.14    0.927     0.919 291.9342 48.14165     15.347
## 3      4.544    -0.18    0.868     0.961 297.0048 48.13413     15.436
## 4      4.564    -0.52    0.791     0.836 285.5346 48.28521     15.597
## 5      4.438     0.07    1.046     1.095 288.7549 48.22620     15.509
## 6      4.486    -0.08    0.972     1.053 296.2861 48.22467     15.714
## 7      4.486    -0.08    0.972     1.053 296.2861 48.22467     15.714
## 8      4.486    -0.08    0.972     1.053 296.2861 48.22467     15.714
## 9      3.986     0.00    1.958     1.358 298.8644 42.15157     12.660
## 10     4.485     0.16    0.848     0.801 286.9995 48.37579     15.841
##       koi_comment
## 1      NO_COMMENT
## 2      NO_COMMENT
## 3   DEEP_V_SHAPED
## 4  MOD_ODDEVEN_DV
## 5      NO_COMMENT
## 6      NO_COMMENT
## 7      NO_COMMENT
## 8      NO_COMMENT
## 9      MOD_SEC_DV
## 10     NO_COMMENT
```

The variable 'koi_comment' gives the reason for disposition values. Initial analysis is showing that there are 52 different comments assigned for each planet to indicating the reason for 'koi_disposition' value.The description for each comment and the flag values are given in the Figures and References section[4]

```
print(unique(data_set$koi_comment))
```

```
##  [1] "NO_COMMENT"
##  [2] "DEEP_V_SHAPED"
##  [3] "MOD_ODDEVEN_DV"
##  [4] "MOD_SEC_DV"
##  [5] "CENT_SATURATED"
##  [6] "CENT_KIC_POS"
##  [7] "CENT_UNRESOLVED_OFFSET"
##  [8] "INDIV_TRANS_CHASES_MARSHALL"
##  [9] "DEPTH_ODDEVEN_ALT"
## [10] "SAME_NTL_PERIOD"
## [11] "CENT_RESOLVED_OFFSET"
## [12] "CENT_CROWDED"
## [13] "LPP_DV"
## [14] "MOD_SEC_ALT"
## [15] "SEASONAL_DEPTH_DV"
## [16] "PLANET_IN_STAR"
## [17] "LPP_ALT"
## [18] "HAS_SEC_TCE"
## [19] "MOD_NONUNIQ_ALT"
## [20] "MOD_ODDEVEN_ALT"
## [21] "DEPTH_ODDEVEN_DV"
## [22] "CENT_FEW_MEAS"
## [23] "CENT_FEW_DIFFS"
## [24] "INCONSISTENT_TRANS"
## [25] "INDIV_TRANS_RUBBLE_SKYE_ZUMA_TRACKER"
```

```
## [26] "SWEET_EB"
## [27] "TRANS_GAPPED"
## [28] "INDIV_TRANS_MARSHALL_SKYE"
## [29] "SWEET_NTL"
## [30] "EPHEM_MATCH"
## [31] "CENT_UNCERTAIN"
## [32] "HALO_GHOST"
## [33] "INDIV_TRANS_SKYE"
## [34] "INDIV_TRANS_RUBBLE_SKYE_ZUMA"
## [35] "INDIV_TRANS_RUBBLE"
## [36] "ALL_TRANS_CHASES"
## [37] "INDIV_TRANS_CHASES_MARSHALL_ZUMA"
## [38] "INDIV_TRANS_MARSHALL"
## [39] "MOD_POS_DV"
## [40] "INDIV_TRANS_MARSHALL_ZUMA"
## [41] "INDIV_TRANS_RUBBLE_SKYE"
## [42] "INDIV_TRANS_CHASES_SKYE"
## [43] "CENT_NOFITS"
## [44] "IS_SEC_TCE"
## [45] "INDIV_TRANS_SKYE_ZUMA_TRACKER"
## [46] "INDIV_TRANS_CHASES_MARSHALL_SKYE"
## [47] "INDIV_TRANS_ZUMA"
## [48] "INDIV_TRANS_RUBBLE_MARSHALL_SKYE"
## [49] "RESIDUAL_TCE"
## [50] "INDIV_TRANS_SKYE_ZUMA"
## [51] "INDIV_TRANS_CHASES"
## [52] "MOD_NONUNIQ_DV"
## [53] ""
```

One hot encoding the 'koi_comment' variable and 'koi_disposition' classes to establish the importance of each comment over the different values of 'koi_disposition'.

```
data_set$koi_comment <- as.factor(data_set$koi_comment)
newdata <- one_hot(as.data.table(data_set))

print(newdata)
```

```
##       loc_rowid    kepid kepoi_name  kepler_name koi_disposition
##           <int>    <int>     <char>       <char>          <char>
##    1:         1 10797460  K00752.01 Kepler-227 b       CONFIRMED
##    2:         2 10797460  K00752.02 Kepler-227 c       CONFIRMED
##    3:         3 10811496  K00753.01                    CANDIDATE
##    4:         4 10848459  K00754.01               FALSE POSITIVE
##    5:         5 10854555  K00755.01 Kepler-664 b       CONFIRMED
##   ---
## 9560:      9560 10090151  K07985.01               FALSE POSITIVE
## 9561:      9561 10128825  K07986.01                    CANDIDATE
## 9562:      9562 10147276  K07987.01               FALSE POSITIVE
## 9563:      9563 10155286  K07988.01                    CANDIDATE
## 9564:      9564 10156110  K07989.01               FALSE POSITIVE
##       koi_pdisposition koi_score koi_fpflag_nt koi_fpflag_ss koi_fpflag_co
##                  <char>     <num>         <int>         <int>         <int>
##    1:        CANDIDATE     1.000             0             0             0
##    2:        CANDIDATE     0.969             0             0             0
##    3:        CANDIDATE     0.000             0             0             0
```

```
##    4:   FALSE POSITIVE     0.000             0             1             0
##    5:        CANDIDATE     1.000             0             0             0
##   ---
## 9560:   FALSE POSITIVE     0.000             0             1             1
## 9561:        CANDIDATE     0.497             0             0             0
## 9562:   FALSE POSITIVE     0.021             0             0             1
## 9563:        CANDIDATE     0.092             0             0             0
## 9564:   FALSE POSITIVE     0.000             0             0             1
##       koi_fpflag_ec  koi_period koi_time0bk koi_impact koi_duration koi_depth
##               <int>       <num>       <num>      <num>        <num>     <num>
##    1:             0   9.4880356    170.5387      0.146      2.95750     616.0
##    2:             0  54.4183827    162.5138      0.586      4.50700     875.0
##    3:             0  19.8991399    175.8503      0.969      1.78220   10800.0
##    4:             0   1.7369525    170.3076      1.276      2.40641    8080.0
##    5:             0   2.5255918    171.5956      0.701      1.65450     603.0
##   ---
## 9560:             0   0.5276985    131.7051      1.252      3.22210    1580.0
## 9561:             0   1.7398494    133.0013      0.043      3.11400      48.5
## 9562:             0   0.6814016    132.1817      0.147      0.86500     104.0
## 9563:             0 333.4861690    153.6150      0.214      3.19900     639.0
## 9564:             1   4.8560348    135.9933      0.134      3.07800      76.7
##       koi_prad koi_sma koi_teq koi_insol koi_model_snr koi_tce_plnt_num
##          <num>   <num>   <int>     <num>         <num>            <int>
##    1:     2.26  0.0853     793     93.59          35.8                1
##    2:     2.83  0.2734     443      9.11          25.8                2
##    3:    14.60  0.1419     638     39.30          76.3                1
##    4:    33.46  0.0267    1395    891.96         505.6                1
##    5:     2.75  0.0374    1406    926.16          40.9                1
##   ---
## 9560:    29.35  0.0128    2088   4500.53         453.3                1
## 9561:     0.72  0.0290    1608   1585.81          10.6                1
## 9562:     1.07  0.0157    2218   5713.41          12.3                1
## 9563:    19.30  1.2233     557     22.68          14.0                1
## 9564:     1.05  0.0606    1266    607.42           8.2                1
##       koi_tce_delivname koi_steff koi_slogg koi_smet koi_srad koi_smass
##                  <char>     <int>     <num>    <num>    <num>     <num>
##    1:   q1_q17_dr25_tce      5455     4.467     0.14    0.927     0.919
##    2:   q1_q17_dr25_tce      5455     4.467     0.14    0.927     0.919
##    3:   q1_q17_dr25_tce      5853     4.544    -0.18    0.868     0.961
##    4:   q1_q17_dr25_tce      5805     4.564    -0.52    0.791     0.836
##    5:   q1_q17_dr25_tce      6031     4.438     0.07    1.046     1.095
##   ---
## 9560:   q1_q17_dr25_tce      5638     4.529     0.14    0.903     1.005
## 9561:   q1_q17_dr25_tce      6119     4.444    -0.04    1.031     1.075
## 9562:   q1_q17_dr25_tce      6173     4.447    -0.04    1.041     1.104
## 9563:   q1_q17_dr25_tce      4989     2.992     0.07    7.824     2.190
## 9564:   q1_q17_dr25_tce      6469     4.385     0.07    1.193     1.260
##             ra      dec koi_kepmag koi_comment_ koi_comment_ALL_TRANS_CHASES
##          <num>    <num>      <num>        <int>                        <int>
##    1: 291.9342 48.14165     15.347            0                            0
##    2: 291.9342 48.14165     15.347            0                            0
##    3: 297.0048 48.13413     15.436            0                            0
##    4: 285.5346 48.28521     15.597            0                            0
##    5: 288.7549 48.22620     15.509            0                            0
```

```
##   ---
## 9560: 297.1888 47.09382     14.082            1                            0
## 9561: 286.5094 47.16322     14.757            1                            0
## 9562: 294.1649 47.17628     15.385            1                            0
## 9563: 296.7629 47.14514     10.998            1                            0
## 9564: 297.0098 47.12102     14.826            1                            0
##       koi_comment_CENT_CROWDED koi_comment_CENT_FEW_DIFFS
##                          <int>                      <int>
##    1:                        0                          0
##    2:                        0                          0
##    3:                        0                          0
##    4:                        0                          0
##    5:                        0                          0
##   ---
## 9560:                        0                          0
## 9561:                        0                          0
## 9562:                        0                          0
## 9563:                        0                          0
## 9564:                        0                          0
##       koi_comment_CENT_FEW_MEAS koi_comment_CENT_KIC_POS
##                           <int>                    <int>
##    1:                         0                        0
##    2:                         0                        0
##    3:                         0                        0
##    4:                         0                        0
##    5:                         0                        0
##   ---
## 9560:                         0                        0
## 9561:                         0                        0
## 9562:                         0                        0
## 9563:                         0                        0
## 9564:                         0                        0
##       koi_comment_CENT_NOFITS koi_comment_CENT_RESOLVED_OFFSET
##                         <int>                             <int>
##    1:                       0                                 0
##    2:                       0                                 0
##    3:                       0                                 0
##    4:                       0                                 0
##    5:                       0                                 0
##   ---
## 9560:                       0                                 0
## 9561:                       0                                 0
## 9562:                       0                                 0
## 9563:                       0                                 0
## 9564:                       0                                 0
##       koi_comment_CENT_SATURATED koi_comment_CENT_UNCERTAIN
##                            <int>                     <int>
##    1:                          0                         0
##    2:                          0                         0
##    3:                          0                         0
##    4:                          0                         0
##    5:                          0                         0
##   ---
## 9560:                          0                         0
```

```
## 9561:                          0                          0
## 9562:                          0                          0
## 9563:                          0                          0
## 9564:                          0                          0
##       koi_comment_CENT_UNRESOLVED_OFFSET koi_comment_DEEP_V_SHAPED
##                                    <int>                     <int>
##    1:                                  0                         0
##    2:                                  0                         0
##    3:                                  0                         1
##    4:                                  0                         0
##    5:                                  0                         0
##   ---
## 9560:                                  0                         0
## 9561:                                  0                         0
## 9562:                                  0                         0
## 9563:                                  0                         0
## 9564:                                  0                         0
##       koi_comment_DEPTH_ODDEVEN_ALT koi_comment_DEPTH_ODDEVEN_DV
##                               <int>                        <int>
##    1:                             0                            0
##    2:                             0                            0
##    3:                             0                            0
##    4:                             0                            0
##    5:                             0                            0
##   ---
## 9560:                             0                            0
## 9561:                             0                            0
## 9562:                             0                            0
## 9563:                             0                            0
## 9564:                             0                            0
##       koi_comment_EPHEM_MATCH koi_comment_HALO_GHOST koi_comment_HAS_SEC_TCE
##                         <int>                  <int>                    <int>
##    1:                       0                      0                        0
##    2:                       0                      0                        0
##    3:                       0                      0                        0
##    4:                       0                      0                        0
##    5:                       0                      0                        0
##   ---
## 9560:                       0                      0                        0
## 9561:                       0                      0                        0
## 9562:                       0                      0                        0
## 9563:                       0                      0                        0
## 9564:                       0                      0                        0
##       koi_comment_INCONSISTENT_TRANS koi_comment_INDIV_TRANS_CHASES
##                                <int>                           <int>
##    1:                              0                               0
##    2:                              0                               0
##    3:                              0                               0
##    4:                              0                               0
##    5:                              0                               0
##   ---
## 9560:                              0                               0
## 9561:                              0                               0
## 9562:                              0                               0
```

```
## 9563:                                0                                0
## 9564:                                0                                0
##       koi_comment_INDIV_TRANS_CHASES_MARSHALL
##                                        <int>
##    1:                                      0
##    2:                                      0
##    3:                                      0
##    4:                                      0
##    5:                                      0
##   ---
## 9560:                                      0
## 9561:                                      0
## 9562:                                      0
## 9563:                                      0
## 9564:                                      0
##       koi_comment_INDIV_TRANS_CHASES_MARSHALL_SKYE
##                                             <int>
##    1:                                           0
##    2:                                           0
##    3:                                           0
##    4:                                           0
##    5:                                           0
##   ---
## 9560:                                           0
## 9561:                                           0
## 9562:                                           0
## 9563:                                           0
## 9564:                                           0
##       koi_comment_INDIV_TRANS_CHASES_MARSHALL_ZUMA
##                                             <int>
##    1:                                           0
##    2:                                           0
##    3:                                           0
##    4:                                           0
##    5:                                           0
##   ---
## 9560:                                           0
## 9561:                                           0
## 9562:                                           0
## 9563:                                           0
## 9564:                                           0
##       koi_comment_INDIV_TRANS_CHASES_SKYE koi_comment_INDIV_TRANS_MARSHALL
##                                     <int>                            <int>
##    1:                                   0                                0
##    2:                                   0                                0
##    3:                                   0                                0
##    4:                                   0                                0
##    5:                                   0                                0
##   ---
## 9560:                                   0                                0
## 9561:                                   0                                0
## 9562:                                   0                                0
## 9563:                                   0                                0
## 9564:                                   0                                0
```

```
##       koi_comment_INDIV_TRANS_MARSHALL_SKYE
##                                       <int>
##    1:                                     0
##    2:                                     0
##    3:                                     0
##    4:                                     0
##    5:                                     0
##   ---
## 9560:                                     0
## 9561:                                     0
## 9562:                                     0
## 9563:                                     0
## 9564:                                     0
##       koi_comment_INDIV_TRANS_MARSHALL_ZUMA koi_comment_INDIV_TRANS_RUBBLE
##                                       <int>                           <int>
##    1:                                     0                               0
##    2:                                     0                               0
##    3:                                     0                               0
##    4:                                     0                               0
##    5:                                     0                               0
##   ---
## 9560:                                     0                               0
## 9561:                                     0                               0
## 9562:                                     0                               0
## 9563:                                     0                               0
## 9564:                                     0                               0
##       koi_comment_INDIV_TRANS_RUBBLE_MARSHALL_SKYE
##                                             <int>
##    1:                                           0
##    2:                                           0
##    3:                                           0
##    4:                                           0
##    5:                                           0
##   ---
## 9560:                                           0
## 9561:                                           0
## 9562:                                           0
## 9563:                                           0
## 9564:                                           0
##       koi_comment_INDIV_TRANS_RUBBLE_SKYE
##                                     <int>
##    1:                                   0
##    2:                                   0
##    3:                                   0
##    4:                                   0
##    5:                                   0
##   ---
## 9560:                                   0
## 9561:                                   0
## 9562:                                   0
## 9563:                                   0
## 9564:                                   0
##       koi_comment_INDIV_TRANS_RUBBLE_SKYE_ZUMA
##                                          <int>
```

```
##    1:                                       0
##    2:                                       0
##    3:                                       0
##    4:                                       0
##    5:                                       0
##   ---
## 9560:                                       0
## 9561:                                       0
## 9562:                                       0
## 9563:                                       0
## 9564:                                       0
##       koi_comment_INDIV_TRANS_RUBBLE_SKYE_ZUMA_TRACKER
##                                                 <int>
##    1:                                               0
##    2:                                               0
##    3:                                               0
##    4:                                               0
##    5:                                               0
##   ---
## 9560:                                               0
## 9561:                                               0
## 9562:                                               0
## 9563:                                               0
## 9564:                                               0
##       koi_comment_INDIV_TRANS_SKYE koi_comment_INDIV_TRANS_SKYE_ZUMA
##                              <int>                              <int>
##    1:                            0                                  0
##    2:                            0                                  0
##    3:                            0                                  0
##    4:                            0                                  0
##    5:                            0                                  0
##   ---
## 9560:                            0                                  0
## 9561:                            0                                  0
## 9562:                            0                                  0
## 9563:                            0                                  0
## 9564:                            0                                  0
##       koi_comment_INDIV_TRANS_SKYE_ZUMA_TRACKER koi_comment_INDIV_TRANS_ZUMA
##                                           <int>                        <int>
##    1:                                         0                            0
##    2:                                         0                            0
##    3:                                         0                            0
##    4:                                         0                            0
##    5:                                         0                            0
##   ---
## 9560:                                         0                            0
## 9561:                                         0                            0
## 9562:                                         0                            0
## 9563:                                         0                            0
## 9564:                                         0                            0
##       koi_comment_IS_SEC_TCE koi_comment_LPP_ALT koi_comment_LPP_DV
##                        <int>               <int>              <int>
##    1:                      0                   0                  0
##    2:                      0                   0                  0
```

```
##    3:                      0                   0                   0
##    4:                      0                   0                   0
##    5:                      0                   0                   0
##   ---
## 9560:                      0                   0                   0
## 9561:                      0                   0                   0
## 9562:                      0                   0                   0
## 9563:                      0                   0                   0
## 9564:                      0                   0                   0
##       koi_comment_MOD_NONUNIQ_ALT koi_comment_MOD_NONUNIQ_DV
##                             <int>                       <int>
##    1:                           0                           0
##    2:                           0                           0
##    3:                           0                           0
##    4:                           0                           0
##    5:                           0                           0
##   ---
## 9560:                           0                           0
## 9561:                           0                           0
## 9562:                           0                           0
## 9563:                           0                           0
## 9564:                           0                           0
##       koi_comment_MOD_ODDEVEN_ALT koi_comment_MOD_ODDEVEN_DV
##                             <int>                       <int>
##    1:                           0                           0
##    2:                           0                           0
##    3:                           0                           0
##    4:                           0                           1
##    5:                           0                           0
##   ---
## 9560:                           0                           0
## 9561:                           0                           0
## 9562:                           0                           0
## 9563:                           0                           0
## 9564:                           0                           0
##       koi_comment_MOD_POS_DV koi_comment_MOD_SEC_ALT koi_comment_MOD_SEC_DV
##                        <int>                   <int>                   <int>
##    1:                      0                       0                       0
##    2:                      0                       0                       0
##    3:                      0                       0                       0
##    4:                      0                       0                       0
##    5:                      0                       0                       0
##   ---
## 9560:                      0                       0                       0
## 9561:                      0                       0                       0
## 9562:                      0                       0                       0
## 9563:                      0                       0                       0
## 9564:                      0                       0                       0
##       koi_comment_NO_COMMENT koi_comment_PLANET_IN_STAR
##                        <int>                      <int>
##    1:                      1                          0
##    2:                      1                          0
##    3:                      0                          0
##    4:                      0                          0
```

```
##    5:                       1                           0
##   ---
## 9560:                       0                           0
## 9561:                       0                           0
## 9562:                       0                           0
## 9563:                       0                           0
## 9564:                       0                           0
##       koi_comment_RESIDUAL_TCE koi_comment_SAME_NTL_PERIOD
##                          <int>                        <int>
##    1:                        0                            0
##    2:                        0                            0
##    3:                        0                            0
##    4:                        0                            0
##    5:                        0                            0
##   ---
## 9560:                        0                            0
## 9561:                        0                            0
## 9562:                        0                            0
## 9563:                        0                            0
## 9564:                        0                            0
##       koi_comment_SEASONAL_DEPTH_DV koi_comment_SWEET_EB koi_comment_SWEET_NTL
##                               <int>                <int>                   <int>
##    1:                             0                    0                       0
##    2:                             0                    0                       0
##    3:                             0                    0                       0
##    4:                             0                    0                       0
##    5:                             0                    0                       0
##   ---
## 9560:                             0                    0                       0
## 9561:                             0                    0                       0
## 9562:                             0                    0                       0
## 9563:                             0                    0                       0
## 9564:                             0                    0                       0
##       koi_comment_TRANS_GAPPED
##                          <int>
##    1:                        0
##    2:                        0
##    3:                        0
##    4:                        0
##    5:                        0
##   ---
## 9560:                        0
## 9561:                        0
## 9562:                        0
## 9563:                        0
## 9564:                        0
```

```
newdata$koi_disposition <-as.factor(newdata$koi_disposition)
final_dataset <- one_hot(as.data.table(newdata))

print(final_dataset)
```

```
##       loc_rowid    kepid kepoi_name  kepler_name koi_disposition_CANDIDATE
##           <int>    <int>     <char>       <char>                     <int>
##    1:         1 10797460  K00752.01 Kepler-227 b                         0
```

```
##    2:         2 10797460  K00752.02 Kepler-227 c                         0
##    3:         3 10811496  K00753.01                                       1
##    4:         4 10848459  K00754.01                                       0
##    5:         5 10854555  K00755.01 Kepler-664 b                         0
##   ---
## 9560:      9560 10090151  K07985.01                                       0
## 9561:      9561 10128825  K07986.01                                       1
## 9562:      9562 10147276  K07987.01                                       0
## 9563:      9563 10155286  K07988.01                                       1
## 9564:      9564 10156110  K07989.01                                       0
##       koi_disposition_CONFIRMED koi_disposition_FALSE POSITIVE koi_pdisposition
##                           <int>                          <int>           <char>
##    1:                         1                              0        CANDIDATE
##    2:                         1                              0        CANDIDATE
##    3:                         0                              0        CANDIDATE
##    4:                         0                              1   FALSE POSITIVE
##    5:                         1                              0        CANDIDATE
##   ---
## 9560:                         0                              1   FALSE POSITIVE
## 9561:                         0                              0        CANDIDATE
## 9562:                         0                              1   FALSE POSITIVE
## 9563:                         0                              0        CANDIDATE
## 9564:                         0                              1   FALSE POSITIVE
##       koi_score koi_fpflag_nt koi_fpflag_ss koi_fpflag_co koi_fpflag_ec
##           <num>         <int>         <int>         <int>         <int>
##    1:     1.000             0             0             0             0
##    2:     0.969             0             0             0             0
##    3:     0.000             0             0             0             0
##    4:     0.000             0             1             0             0
##    5:     1.000             0             0             0             0
##   ---
## 9560:     0.000             0             1             1             0
## 9561:     0.497             0             0             0             0
## 9562:     0.021             0             0             1             0
## 9563:     0.092             0             0             0             0
## 9564:     0.000             0             0             1             1
##        koi_period koi_time0bk koi_impact koi_duration koi_depth koi_prad
##             <num>       <num>      <num>        <num>     <num>    <num>
##    1:  9.4880356    170.5387      0.146      2.95750     616.0     2.26
##    2: 54.4183827    162.5138      0.586      4.50700     875.0     2.83
##    3: 19.8991399    175.8503      0.969      1.78220   10800.0    14.60
##    4:  1.7369525    170.3076      1.276      2.40641    8080.0    33.46
##    5:  2.5255918    171.5956      0.701      1.65450     603.0     2.75
##   ---
## 9560:  0.5276985    131.7051      1.252      3.22210    1580.0    29.35
## 9561:  1.7398494    133.0013      0.043      3.11400      48.5     0.72
## 9562:  0.6814016    132.1817      0.147      0.86500     104.0     1.07
## 9563: 333.4861690    153.6150      0.214      3.19900     639.0    19.30
## 9564:  4.8560348    135.9933      0.134      3.07800      76.7     1.05
##       koi_sma koi_teq koi_insol koi_model_snr koi_tce_plnt_num
##         <num>   <int>     <num>         <num>            <int>
##    1:  0.0853     793     93.59          35.8                1
##    2:  0.2734     443      9.11          25.8                2
##    3:  0.1419     638     39.30          76.3                1
```

```
##    4:  0.0267    1395    891.96         505.6                1
##    5:  0.0374    1406    926.16          40.9                1
##   ---
## 9560:  0.0128    2088   4500.53         453.3                1
## 9561:  0.0290    1608   1585.81          10.6                1
## 9562:  0.0157    2218   5713.41          12.3                1
## 9563:  1.2233     557     22.68          14.0                1
## 9564:  0.0606    1266    607.42           8.2                1
##       koi_tce_delivname koi_steff koi_slogg koi_smet koi_srad koi_smass
##                  <char>     <int>     <num>    <num>    <num>     <num>
##    1:   q1_q17_dr25_tce      5455     4.467     0.14    0.927     0.919
##    2:   q1_q17_dr25_tce      5455     4.467     0.14    0.927     0.919
##    3:   q1_q17_dr25_tce      5853     4.544    -0.18    0.868     0.961
##    4:   q1_q17_dr25_tce      5805     4.564    -0.52    0.791     0.836
##    5:   q1_q17_dr25_tce      6031     4.438     0.07    1.046     1.095
##   ---
## 9560:   q1_q17_dr25_tce      5638     4.529     0.14    0.903     1.005
## 9561:   q1_q17_dr25_tce      6119     4.444    -0.04    1.031     1.075
## 9562:   q1_q17_dr25_tce      6173     4.447    -0.04    1.041     1.104
## 9563:   q1_q17_dr25_tce      4989     2.992     0.07    7.824     2.190
## 9564:   q1_q17_dr25_tce      6469     4.385     0.07    1.193     1.260
##            ra      dec koi_kepmag koi_comment_ koi_comment_ALL_TRANS_CHASES
##          <num>    <num>      <num>        <int>                        <int>
##    1: 291.9342 48.14165     15.347            0                            0
##    2: 291.9342 48.14165     15.347            0                            0
##    3: 297.0048 48.13413     15.436            0                            0
##    4: 285.5346 48.28521     15.597            0                            0
##    5: 288.7549 48.22620     15.509            0                            0
##   ---
## 9560: 297.1888 47.09382     14.082            1                            0
## 9561: 286.5094 47.16322     14.757            1                            0
## 9562: 294.1649 47.17628     15.385            1                            0
## 9563: 296.7629 47.14514     10.998            1                            0
## 9564: 297.0098 47.12102     14.826            1                            0
##       koi_comment_CENT_CROWDED koi_comment_CENT_FEW_DIFFS
##                          <int>                       <int>
##    1:                        0                           0
##    2:                        0                           0
##    3:                        0                           0
##    4:                        0                           0
##    5:                        0                           0
##   ---
## 9560:                        0                           0
## 9561:                        0                           0
## 9562:                        0                           0
## 9563:                        0                           0
## 9564:                        0                           0
##       koi_comment_CENT_FEW_MEAS koi_comment_CENT_KIC_POS
##                           <int>                     <int>
##    1:                         0                         0
##    2:                         0                         0
##    3:                         0                         0
##    4:                         0                         0
##    5:                         0                         0
```

```
##   ---
## 9560:                          0                          0
## 9561:                          0                          0
## 9562:                          0                          0
## 9563:                          0                          0
## 9564:                          0                          0
##       koi_comment_CENT_NOFITS koi_comment_CENT_RESOLVED_OFFSET
##                         <int>                            <int>
##    1:                       0                                0
##    2:                       0                                0
##    3:                       0                                0
##    4:                       0                                0
##    5:                       0                                0
##   ---
## 9560:                       0                                0
## 9561:                       0                                0
## 9562:                       0                                0
## 9563:                       0                                0
## 9564:                       0                                0
##       koi_comment_CENT_SATURATED koi_comment_CENT_UNCERTAIN
##                            <int>                       <int>
##    1:                          0                           0
##    2:                          0                           0
##    3:                          0                           0
##    4:                          0                           0
##    5:                          0                           0
##   ---
## 9560:                          0                           0
## 9561:                          0                           0
## 9562:                          0                           0
## 9563:                          0                           0
## 9564:                          0                           0
##       koi_comment_CENT_UNRESOLVED_OFFSET koi_comment_DEEP_V_SHAPED
##                                    <int>                      <int>
##    1:                                  0                          0
##    2:                                  0                          0
##    3:                                  0                          1
##    4:                                  0                          0
##    5:                                  0                          0
##   ---
## 9560:                                  0                          0
## 9561:                                  0                          0
## 9562:                                  0                          0
## 9563:                                  0                          0
## 9564:                                  0                          0
##       koi_comment_DEPTH_ODDEVEN_ALT koi_comment_DEPTH_ODDEVEN_DV
##                               <int>                         <int>
##    1:                             0                             0
##    2:                             0                             0
##    3:                             0                             0
##    4:                             0                             0
##    5:                             0                             0
##   ---
## 9560:                             0                             0
```

```
## 9561:                            0                            0
## 9562:                            0                            0
## 9563:                            0                            0
## 9564:                            0                            0
##       koi_comment_EPHEM_MATCH koi_comment_HALO_GHOST koi_comment_HAS_SEC_TCE
##                         <int>                 <int>                   <int>
##    1:                       0                     0                       0
##    2:                       0                     0                       0
##    3:                       0                     0                       0
##    4:                       0                     0                       0
##    5:                       0                     0                       0
##   ---
## 9560:                       0                     0                       0
## 9561:                       0                     0                       0
## 9562:                       0                     0                       0
## 9563:                       0                     0                       0
## 9564:                       0                     0                       0
##       koi_comment_INCONSISTENT_TRANS koi_comment_INDIV_TRANS_CHASES
##                                <int>                          <int>
##    1:                              0                              0
##    2:                              0                              0
##    3:                              0                              0
##    4:                              0                              0
##    5:                              0                              0
##   ---
## 9560:                              0                              0
## 9561:                              0                              0
## 9562:                              0                              0
## 9563:                              0                              0
## 9564:                              0                              0
##       koi_comment_INDIV_TRANS_CHASES_MARSHALL
##                                         <int>
##    1:                                       0
##    2:                                       0
##    3:                                       0
##    4:                                       0
##    5:                                       0
##   ---
## 9560:                                       0
## 9561:                                       0
## 9562:                                       0
## 9563:                                       0
## 9564:                                       0
##       koi_comment_INDIV_TRANS_CHASES_MARSHALL_SKYE
##                                              <int>
##    1:                                            0
##    2:                                            0
##    3:                                            0
##    4:                                            0
##    5:                                            0
##   ---
## 9560:                                            0
## 9561:                                            0
## 9562:                                            0
```

```
## 9563:                                          0
## 9564:                                          0
##       koi_comment_INDIV_TRANS_CHASES_MARSHALL_ZUMA
##                                             <int>
##    1:                                          0
##    2:                                          0
##    3:                                          0
##    4:                                          0
##    5:                                          0
##   ---
## 9560:                                          0
## 9561:                                          0
## 9562:                                          0
## 9563:                                          0
## 9564:                                          0
##       koi_comment_INDIV_TRANS_CHASES_SKYE koi_comment_INDIV_TRANS_MARSHALL
##                                   <int>                               <int>
##    1:                                 0                                   0
##    2:                                 0                                   0
##    3:                                 0                                   0
##    4:                                 0                                   0
##    5:                                 0                                   0
##   ---
## 9560:                                 0                                   0
## 9561:                                 0                                   0
## 9562:                                 0                                   0
## 9563:                                 0                                   0
## 9564:                                 0                                   0
##       koi_comment_INDIV_TRANS_MARSHALL_SKYE
##                                       <int>
##    1:                                    0
##    2:                                    0
##    3:                                    0
##    4:                                    0
##    5:                                    0
##   ---
## 9560:                                    0
## 9561:                                    0
## 9562:                                    0
## 9563:                                    0
## 9564:                                    0
##       koi_comment_INDIV_TRANS_MARSHALL_ZUMA koi_comment_INDIV_TRANS_RUBBLE
##                                       <int>                           <int>
##    1:                                    0                               0
##    2:                                    0                               0
##    3:                                    0                               0
##    4:                                    0                               0
##    5:                                    0                               0
##   ---
## 9560:                                    0                               0
## 9561:                                    0                               0
## 9562:                                    0                               0
## 9563:                                    0                               0
## 9564:                                    0                               0
```

```
##       koi_comment_INDIV_TRANS_RUBBLE_MARSHALL_SKYE
##                                              <int>
##    1:                                            0
##    2:                                            0
##    3:                                            0
##    4:                                            0
##    5:                                            0
##   ---
## 9560:                                            0
## 9561:                                            0
## 9562:                                            0
## 9563:                                            0
## 9564:                                            0
##       koi_comment_INDIV_TRANS_RUBBLE_SKYE
##                                    <int>
##    1:                                  0
##    2:                                  0
##    3:                                  0
##    4:                                  0
##    5:                                  0
##   ---
## 9560:                                  0
## 9561:                                  0
## 9562:                                  0
## 9563:                                  0
## 9564:                                  0
##       koi_comment_INDIV_TRANS_RUBBLE_SKYE_ZUMA
##                                         <int>
##    1:                                       0
##    2:                                       0
##    3:                                       0
##    4:                                       0
##    5:                                       0
##   ---
## 9560:                                       0
## 9561:                                       0
## 9562:                                       0
## 9563:                                       0
## 9564:                                       0
##       koi_comment_INDIV_TRANS_RUBBLE_SKYE_ZUMA_TRACKER
##                                                 <int>
##    1:                                               0
##    2:                                               0
##    3:                                               0
##    4:                                               0
##    5:                                               0
##   ---
## 9560:                                               0
## 9561:                                               0
## 9562:                                               0
## 9563:                                               0
## 9564:                                               0
##       koi_comment_INDIV_TRANS_SKYE koi_comment_INDIV_TRANS_SKYE_ZUMA
##                              <int>                               <int>
```

```
##    1:                            0                                0
##    2:                            0                                0
##    3:                            0                                0
##    4:                            0                                0
##    5:                            0                                0
##   ---
## 9560:                            0                                0
## 9561:                            0                                0
## 9562:                            0                                0
## 9563:                            0                                0
## 9564:                            0                                0
##       koi_comment_INDIV_TRANS_SKYE_ZUMA_TRACKER koi_comment_INDIV_TRANS_ZUMA
##                                          <int>                        <int>
##    1:                                       0                            0
##    2:                                       0                            0
##    3:                                       0                            0
##    4:                                       0                            0
##    5:                                       0                            0
##   ---
## 9560:                                       0                            0
## 9561:                                       0                            0
## 9562:                                       0                            0
## 9563:                                       0                            0
## 9564:                                       0                            0
##       koi_comment_IS_SEC_TCE koi_comment_LPP_ALT koi_comment_LPP_DV
##                        <int>                <int>               <int>
##    1:                      0                    0                   0
##    2:                      0                    0                   0
##    3:                      0                    0                   0
##    4:                      0                    0                   0
##    5:                      0                    0                   0
##   ---
## 9560:                      0                    0                   0
## 9561:                      0                    0                   0
## 9562:                      0                    0                   0
## 9563:                      0                    0                   0
## 9564:                      0                    0                   0
##       koi_comment_MOD_NONUNIQ_ALT koi_comment_MOD_NONUNIQ_DV
##                             <int>                       <int>
##    1:                           0                           0
##    2:                           0                           0
##    3:                           0                           0
##    4:                           0                           0
##    5:                           0                           0
##   ---
## 9560:                           0                           0
## 9561:                           0                           0
## 9562:                           0                           0
## 9563:                           0                           0
## 9564:                           0                           0
##       koi_comment_MOD_ODDEVEN_ALT koi_comment_MOD_ODDEVEN_DV
##                             <int>                       <int>
##    1:                           0                           0
##    2:                           0                           0
```

```
##    3:                          0                         0
##    4:                          0                         1
##    5:                          0                         0
##   ---
## 9560:                          0                         0
## 9561:                          0                         0
## 9562:                          0                         0
## 9563:                          0                         0
## 9564:                          0                         0
##       koi_comment_MOD_POS_DV koi_comment_MOD_SEC_ALT koi_comment_MOD_SEC_DV
##                        <int>                   <int>                   <int>
##    1:                      0                       0                       0
##    2:                      0                       0                       0
##    3:                      0                       0                       0
##    4:                      0                       0                       0
##    5:                      0                       0                       0
##   ---
## 9560:                      0                       0                       0
## 9561:                      0                       0                       0
## 9562:                      0                       0                       0
## 9563:                      0                       0                       0
## 9564:                      0                       0                       0
##       koi_comment_NO_COMMENT koi_comment_PLANET_IN_STAR
##                        <int>                      <int>
##    1:                      1                          0
##    2:                      1                          0
##    3:                      0                          0
##    4:                      0                          0
##    5:                      1                          0
##   ---
## 9560:                      0                          0
## 9561:                      0                          0
## 9562:                      0                          0
## 9563:                      0                          0
## 9564:                      0                          0
##       koi_comment_RESIDUAL_TCE koi_comment_SAME_NTL_PERIOD
##                          <int>                       <int>
##    1:                        0                           0
##    2:                        0                           0
##    3:                        0                           0
##    4:                        0                           0
##    5:                        0                           0
##   ---
## 9560:                        0                           0
## 9561:                        0                           0
## 9562:                        0                           0
## 9563:                        0                           0
## 9564:                        0                           0
##       koi_comment_SEASONAL_DEPTH_DV koi_comment_SWEET_EB koi_comment_SWEET_NTL
##                               <int>                <int>                   <int>
##    1:                             0                    0                       0
##    2:                             0                    0                       0
##    3:                             0                    0                       0
##    4:                             0                    0                       0
```

```
##    5:                             0                   0                   0
##   ---
## 9560:                             0                   0                   0
## 9561:                             0                   0                   0
## 9562:                             0                   0                   0
## 9563:                             0                   0                   0
## 9564:                             0                   0                   0
##       koi_comment_TRANS_GAPPED
##                          <int>
##    1:                        0
##    2:                        0
##    3:                        0
##    4:                        0
##    5:                        0
##   ---
## 9560:                        0
## 9561:                        0
## 9562:                        0
## 9563:                        0
## 9564:                        0
```

```
str(final_dataset)
```

```
## Classes 'data.table' and 'data.frame':   9564 obs. of  86 variables:
##  $ loc_rowid                                    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ kepid                                        : int  10797460 10797460 10811496 10848459 108545
##  $ kepoi_name                                   : chr  "K00752.01" "K00752.02" "K00753.01" "K0075
##  $ kepler_name                                  : chr  "Kepler-227 b" "Kepler-227 c" "" "" ...
##  $ koi_disposition_CANDIDATE                    : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ koi_disposition_CONFIRMED                    : int  1 1 0 0 1 1 1 1 0 1 ...
##  $ koi_disposition_FALSE POSITIVE               : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ koi_pdisposition                             : chr  "CANDIDATE" "CANDIDATE" "CANDIDATE" "FALSE
##  $ koi_score                                    : num  1 0.969 0 0 1 1 1 0.992 0 1 ...
##  $ koi_fpflag_nt                                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_fpflag_ss                                : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ koi_fpflag_co                                : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ koi_fpflag_ec                                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_period                                   : num  9.49 54.42 19.9 1.74 2.53 ...
##  $ koi_time0bk                                  : num  171 163 176 170 172 ...
##  $ koi_impact                                   : num  0.146 0.586 0.969 1.276 0.701 ...
##  $ koi_duration                                 : num  2.96 4.51 1.78 2.41 1.65 ...
##  $ koi_depth                                    : num  616 875 10800 8080 603 1520 686 227 234 49
##  $ koi_prad                                     : num  2.26 2.83 14.6 33.46 2.75 ...
##  $ koi_sma                                      : num  0.0853 0.2734 0.1419 0.0267 0.0374 ...
##  $ koi_teq                                      : int  793 443 638 1395 1406 835 1160 1360 1342 6
##  $ koi_insol                                    : num  93.59 9.11 39.3 891.96 926.16 ...
##  $ koi_model_snr                                : num  35.8 25.8 76.3 505.6 40.9 ...
##  $ koi_tce_plnt_num                             : int  1 2 1 1 1 1 2 3 1 1 ...
##  $ koi_tce_delivname                            : chr  "q1_q17_dr25_tce" "q1_q17_dr25_tce" "q1_q1
##  $ koi_steff                                    : int  5455 5455 5853 5805 6031 6046 6046 6046 62
##  $ koi_slogg                                    : num  4.47 4.47 4.54 4.56 4.44 ...
##  $ koi_smet                                     : num  0.14 0.14 -0.18 -0.52 0.07 -0.08 -0.08 -0.
##  $ koi_srad                                     : num  0.927 0.927 0.868 0.791 1.046 ...
##  $ koi_smass                                    : num  0.919 0.919 0.961 0.836 1.095 ...
##  $ ra                                           : num  292 292 297 286 289 ...
```

```
##  $ dec                                          : num  48.1 48.1 48.1 48.3 48.2 ...
##  $ koi_kepmag                                   : num  15.3 15.3 15.4 15.6 15.5 ...
##  $ koi_comment_                                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_ALL_TRANS_CHASES                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_CENT_CROWDED                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_CENT_FEW_DIFFS                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_CENT_FEW_MEAS                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_CENT_KIC_POS                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_CENT_NOFITS                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_CENT_RESOLVED_OFFSET             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_CENT_SATURATED                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_CENT_UNCERTAIN                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_CENT_UNRESOLVED_OFFSET           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_DEEP_V_SHAPED                    : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ koi_comment_DEPTH_ODDEVEN_ALT                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_DEPTH_ODDEVEN_DV                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_EPHEM_MATCH                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_HALO_GHOST                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_HAS_SEC_TCE                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INCONSISTENT_TRANS               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_CHASES               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_CHASES_MARSHALL      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_CHASES_MARSHALL_SKYE : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_CHASES_MARSHALL_ZUMA : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_CHASES_SKYE          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_MARSHALL             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_MARSHALL_SKYE        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_MARSHALL_ZUMA        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_RUBBLE               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_RUBBLE_MARSHALL_SKYE : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_RUBBLE_SKYE          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_RUBBLE_SKYE_ZUMA     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_RUBBLE_SKYE_ZUMA_TRACKER: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_SKYE                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_SKYE_ZUMA            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_SKYE_ZUMA_TRACKER    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_INDIV_TRANS_ZUMA                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_IS_SEC_TCE                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_LPP_ALT                          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_LPP_DV                           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_MOD_NONUNIQ_ALT                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_MOD_NONUNIQ_DV                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_MOD_ODDEVEN_ALT                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_MOD_ODDEVEN_DV                   : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ koi_comment_MOD_POS_DV                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_MOD_SEC_ALT                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_MOD_SEC_DV                       : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ koi_comment_NO_COMMENT                       : int  1 1 0 0 1 1 1 1 0 1 ...
##  $ koi_comment_PLANET_IN_STAR                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_RESIDUAL_TCE                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_SAME_NTL_PERIOD                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_SEASONAL_DEPTH_DV                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_SWEET_EB                         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ koi_comment_SWEET_NTL                        : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ koi_comment_TRANS_GAPPED                          : int  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
correlation_mat_flags <- data.frame(
  koi_disposition_CONFIRMED = numeric(0),
  koi_disposition_CANDIDATE = numeric(0),
  koi_disposition_FALSE_POSITIVE = numeric(0)
)
```

```
comments <- grep('koi_comment', names(final_dataset), value = TRUE)
# print(comments)


target_vars <- c("koi_disposition_CONFIRMED", "koi_disposition_CANDIDATE", "koi_disposition_FALSE POSIT
flag_vars <- c("koi_fpflag_nt","koi_fpflag_ss","koi_fpflag_co","koi_fpflag_ec")

# Get the column names containing 'koi_comment'
comment_vars <- grep("^koi_comment_", names(final_dataset), value = TRUE)

# Subset the final_dataset to include only the descriptor and target variables
subset_1 <- final_dataset[, c(comment_vars), with = FALSE]
subset_2 <- final_dataset[, c(target_vars), with = FALSE]
subset_3 <- final_dataset[, c(flag_vars), with = FALSE]

# dim(subset_data)
# dim(subset1)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:randomForest':
##
##     outlier
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
correlation_matrix1 <- cor(subset_1,subset_2)
correlation_matrix2 <- cor(subset_3,subset_2)
```

```
corr_df1 <- as.data.frame(correlation_matrix1)
corr_df1
```

CORRELATION TABLE FOR COMMENTS VS DISPOSITION VALUES

```
##                                                koi_disposition_CONFIRMED
## koi_comment_                                               -0.276616894
## koi_comment_ALL_TRANS_CHASES                               -0.036832784
## koi_comment_CENT_CROWDED                                    0.024107000
## koi_comment_CENT_FEW_DIFFS                                 -0.044532169
## koi_comment_CENT_FEW_MEAS                                  -0.026811855
## koi_comment_CENT_KIC_POS                                    0.111111642
## koi_comment_CENT_NOFITS                                    -0.013441508
## koi_comment_CENT_RESOLVED_OFFSET                            0.010647071
```

```
## koi_comment_CENT_SATURATED                                  0.027639833
## koi_comment_CENT_UNCERTAIN                                 -0.010838565
## koi_comment_CENT_UNRESOLVED_OFFSET                          0.028454236
## koi_comment_DEEP_V_SHAPED                                   0.004963257
## koi_comment_DEPTH_ODDEVEN_ALT                              -0.027642696
## koi_comment_DEPTH_ODDEVEN_DV                               -0.045822305
## koi_comment_EPHEM_MATCH                                    -0.010806301
## koi_comment_HALO_GHOST                                     -0.017679707
## koi_comment_HAS_SEC_TCE                                    -0.013647039
## koi_comment_INCONSISTENT_TRANS                             -0.031615291
## koi_comment_INDIV_TRANS_CHASES                             -0.011230183
## koi_comment_INDIV_TRANS_CHASES_MARSHALL                    -0.018343609
## koi_comment_INDIV_TRANS_CHASES_MARSHALL_SKYE               -0.009168927
## koi_comment_INDIV_TRANS_CHASES_MARSHALL_ZUMA                0.016129654
## koi_comment_INDIV_TRANS_CHASES_SKYE                         0.001826645
## koi_comment_INDIV_TRANS_MARSHALL                           -0.017900477
## koi_comment_INDIV_TRANS_MARSHALL_SKYE                      -0.010638788
## koi_comment_INDIV_TRANS_MARSHALL_ZUMA                      -0.006483071
## koi_comment_INDIV_TRANS_RUBBLE                             -0.006650349
## koi_comment_INDIV_TRANS_RUBBLE_MARSHALL_SKYE               -0.006483071
## koi_comment_INDIV_TRANS_RUBBLE_SKYE                        -0.004384787
## koi_comment_INDIV_TRANS_RUBBLE_SKYE_ZUMA                    0.014883474
## koi_comment_INDIV_TRANS_RUBBLE_SKYE_ZUMA_TRACKER            0.013647324
## koi_comment_INDIV_TRANS_SKYE                               -0.022054626
## koi_comment_INDIV_TRANS_SKYE_ZUMA                          -0.011230183
## koi_comment_INDIV_TRANS_SKYE_ZUMA_TRACKER                   0.001826645
## koi_comment_INDIV_TRANS_ZUMA                               -0.009168927
## koi_comment_IS_SEC_TCE                                     -0.008608488
## koi_comment_LPP_ALT                                         0.001712596
## koi_comment_LPP_DV                                         -0.058477491
## koi_comment_MOD_NONUNIQ_ALT                                -0.028906625
## koi_comment_MOD_NONUNIQ_DV                                 -0.022470945
## koi_comment_MOD_ODDEVEN_ALT                                 0.004161321
## koi_comment_MOD_ODDEVEN_DV                                  0.001346714
## koi_comment_MOD_POS_DV                                     -0.006483071
## koi_comment_MOD_SEC_ALT                                    -0.033773588
## koi_comment_MOD_SEC_DV                                     -0.033375143
## koi_comment_NO_COMMENT                                      0.298638317
## koi_comment_PLANET_IN_STAR                                 -0.011404468
## koi_comment_RESIDUAL_TCE                                   -0.006483071
## koi_comment_SAME_NTL_PERIOD                                -0.017934603
## koi_comment_SEASONAL_DEPTH_DV                              -0.004384787
## koi_comment_SWEET_EB                                       -0.062209914
## koi_comment_SWEET_NTL                                      -0.028725580
## koi_comment_TRANS_GAPPED                                   -0.001048483
##                                              koi_disposition_CANDIDATE
## koi_comment_                                                0.0161512700
## koi_comment_ALL_TRANS_CHASES                               -0.0125515265
## koi_comment_CENT_CROWDED                                   -0.0128361077
## koi_comment_CENT_FEW_DIFFS                                 -0.0120028158
## koi_comment_CENT_FEW_MEAS                                  -0.0047878542
## koi_comment_CENT_KIC_POS                                   -0.0273791625
## koi_comment_CENT_NOFITS                                     0.0188353261
## koi_comment_CENT_RESOLVED_OFFSET                           -0.0174703055
```

```
## koi_comment_CENT_SATURATED                                0.0003689022
## koi_comment_CENT_UNCERTAIN                                0.0091324956
## koi_comment_CENT_UNRESOLVED_OFFSET                       -0.0024356321
## koi_comment_DEEP_V_SHAPED                                -0.0141393399
## koi_comment_DEPTH_ODDEVEN_ALT                             0.0106799896
## koi_comment_DEPTH_ODDEVEN_DV                              0.0416825760
## koi_comment_EPHEM_MATCH                                  -0.0034074658
## koi_comment_HALO_GHOST                                    0.0123227511
## koi_comment_HAS_SEC_TCE                                   0.0138889851
## koi_comment_INCONSISTENT_TRANS                            0.0089389721
## koi_comment_INDIV_TRANS_CHASES                            0.0055051786
## koi_comment_INDIV_TRANS_CHASES_MARSHALL                  -0.0058766913
## koi_comment_INDIV_TRANS_CHASES_MARSHALL_SKYE              0.0104404429
## koi_comment_INDIV_TRANS_CHASES_MARSHALL_ZUMA             -0.0052299921
## koi_comment_INDIV_TRANS_CHASES_SKYE                      -0.0090595594
## koi_comment_INDIV_TRANS_MARSHALL                          0.0061429510
## koi_comment_INDIV_TRANS_MARSHALL_SKYE                    -0.0159029907
## koi_comment_INDIV_TRANS_MARSHALL_ZUMA                    -0.0052299921
## koi_comment_INDIV_TRANS_RUBBLE                            0.0180871610
## koi_comment_INDIV_TRANS_RUBBLE_MARSHALL_SKYE              0.0199942360
## koi_comment_INDIV_TRANS_RUBBLE_SKYE                      -0.0116970644
## koi_comment_INDIV_TRANS_RUBBLE_SKYE_ZUMA                 -0.0090595594
## koi_comment_INDIV_TRANS_RUBBLE_SKYE_ZUMA_TRACKER         -0.0058766913
## koi_comment_INDIV_TRANS_SKYE                              0.0006085777
## koi_comment_INDIV_TRANS_SKYE_ZUMA                         0.0200699165
## koi_comment_INDIV_TRANS_SKYE_ZUMA_TRACKER                 0.0055051786
## koi_comment_INDIV_TRANS_ZUMA                              0.0104404429
## koi_comment_IS_SEC_TCE                                   -0.0043047470
## koi_comment_LPP_ALT                                       0.0091324956
## koi_comment_LPP_DV                                        0.0343131370
## koi_comment_MOD_NONUNIQ_ALT                               0.0111031440
## koi_comment_MOD_NONUNIQ_DV                               -0.0181276528
## koi_comment_MOD_ODDEVEN_ALT                              -0.0086189686
## koi_comment_MOD_ODDEVEN_DV                               -0.0147272407
## koi_comment_MOD_POS_DV                                    0.0199942360
## koi_comment_MOD_SEC_ALT                                   0.0346560342
## koi_comment_MOD_SEC_DV                                    0.0330313077
## koi_comment_NO_COMMENT                                   -0.0632278649
## koi_comment_PLANET_IN_STAR                                0.0031932427
## koi_comment_RESIDUAL_TCE                                 -0.0052299921
## koi_comment_SAME_NTL_PERIOD                               0.0258982788
## koi_comment_SEASONAL_DEPTH_DV                            -0.0004140867
## koi_comment_SWEET_EB                                      0.0342669943
## koi_comment_SWEET_NTL                                     0.0284285361
## koi_comment_TRANS_GAPPED                                 -0.0021362568
##                                          koi_disposition_FALSE POSITIVE
## koi_comment_                                                  0.2371048829
## koi_comment_ALL_TRANS_CHASES                                  0.0434929603
## koi_comment_CENT_CROWDED                                     -0.0113966105
## koi_comment_CENT_FEW_DIFFS                                    0.0500121634
## koi_comment_CENT_FEW_MEAS                                     0.0281337307
## koi_comment_CENT_KIC_POS                                     -0.0783003428
## koi_comment_CENT_NOFITS                                      -0.0031149045
## koi_comment_CENT_RESOLVED_OFFSET                              0.0045356456
```

```
## koi_comment_CENT_SATURATED                                -0.0252994887
## koi_comment_CENT_UNCERTAIN                                 0.0023983581
## koi_comment_CENT_UNRESOLVED_OFFSET                        -0.0237620355
## koi_comment_DEEP_V_SHAPED                                  0.0069757357
## koi_comment_DEPTH_ODDEVEN_ALT                              0.0163429741
## koi_comment_DEPTH_ODDEVEN_DV                               0.0076477590
## koi_comment_EPHEM_MATCH                                    0.0125373284
## koi_comment_HALO_GHOST                                     0.0059993532
## koi_comment_HAS_SEC_TCE                                    0.0010817896
## koi_comment_INCONSISTENT_TRANS                             0.0213479269
## koi_comment_INDIV_TRANS_CHASES                             0.0056938250
## koi_comment_INDIV_TRANS_CHASES_MARSHALL                    0.0213570563
## koi_comment_INDIV_TRANS_CHASES_MARSHALL_SKYE              -0.0001723998
## koi_comment_INDIV_TRANS_CHASES_MARSHALL_ZUMA              -0.0103485577
## koi_comment_INDIV_TRANS_CHASES_SKYE                        0.0056938250
## koi_comment_INDIV_TRANS_MARSHALL                           0.0112099933
## koi_comment_INDIV_TRANS_MARSHALL_SKYE                      0.0225179343
## koi_comment_INDIV_TRANS_MARSHALL_ZUMA                      0.0101047603
## koi_comment_INDIV_TRANS_RUBBLE                            -0.0086508836
## koi_comment_INDIV_TRANS_RUBBLE_MARSHALL_SKYE              -0.0103485577
## koi_comment_INDIV_TRANS_RUBBLE_SKYE                        0.0134507425
## koi_comment_INDIV_TRANS_RUBBLE_SKYE_ZUMA                  -0.0061161387
## koi_comment_INDIV_TRANS_RUBBLE_SKYE_ZUMA_TRACKER          -0.0075788959
## koi_comment_INDIV_TRANS_SKYE                               0.0194550433
## koi_comment_INDIV_TRANS_SKYE_ZUMA                         -0.0061161387
## koi_comment_INDIV_TRANS_SKYE_ZUMA_TRACKER                 -0.0061161387
## koi_comment_INDIV_TRANS_ZUMA                              -0.0001723998
## koi_comment_IS_SEC_TCE                                     0.0112769660
## koi_comment_LPP_ALT                                       -0.0089542265
## koi_comment_LPP_DV                                         0.0250700188
## koi_comment_MOD_NONUNIQ_ALT                                0.0171430845
## koi_comment_MOD_NONUNIQ_DV                                 0.0350240659
## koi_comment_MOD_ODDEVEN_ALT                                0.0032248422
## koi_comment_MOD_ODDEVEN_DV                                 0.0107236215
## koi_comment_MOD_POS_DV                                    -0.0103485577
## koi_comment_MOD_SEC_ALT                                    0.0024471790
## koi_comment_MOD_SEC_DV                                     0.0034042098
## koi_comment_NO_COMMENT                                    -0.2188508369
## koi_comment_PLANET_IN_STAR                                 0.0077261232
## koi_comment_RESIDUAL_TCE                                   0.0101047603
## koi_comment_SAME_NTL_PERIOD                               -0.0047779445
## koi_comment_SEASONAL_DEPTH_DV                              0.0043018270
## koi_comment_SWEET_EB                                       0.0284834292
## koi_comment_SWEET_NTL                                      0.0029308611
## koi_comment_TRANS_GAPPED                                   0.0026805634
```

Inference

One major inference that can be made from the correlation results is that whenever the correlation is negative for 'comment' and 'False Positive', 'Confirmed' has a positive value there and vice versa. Majority of the comment vaues are posotively correlated with FALSE POSITIVE and anti-correlated with CONFIRMED or CANDIDATE. Also, 'NO_COMMENT' has the highest negative correlation with 'Confirmed' disposition meaning that the exoplanets confirmed doesn't have any justification mentioned for their disposition value.

CONFIRMED	CANDIDATE	FALSE POSITIVE
koi_comment_NO_COMMENT	koi_comment_MOD_SEC_ALT	koi_comment_LPP_DV
koi_comment_CENT_KIC_POS	koi_comment_CENT_RESOLVED_OFFSET	koi_comment_CENT_FEW_MEAS
koi_comment_CENT_SATURATED	koi_comment_DEPTH_ODDEVEN_DV	koi_comment_SWEET_EB
koi_comment_CENT_UNRESOLVED_OFFSET	koi_comment_MOD_NONUNIQ_DV	koi_comment_MOD_NONUNIQ_DV
koi_comment_CENT_CROWDED	koi_comment_MOD_SEC_DV	koi_comment_ALL_TRANS_CHASES

```
corr_df2 <- as.data.frame(correlation_matrix2)
corr_df2
```

Correlation between FLAG variables and disposition values

```
##               koi_disposition_CONFIRMED koi_disposition_CANDIDATE
## koi_fpflag_nt               -0.005042617               -0.02232551
## koi_fpflag_ss               -0.341521834               -0.28046961
## koi_fpflag_co               -0.314524626               -0.25373180
## koi_fpflag_ec               -0.234150963               -0.18889314
##               koi_disposition_FALSE POSITIVE
## koi_fpflag_nt                     0.02266393
## koi_fpflag_ss                     0.53632972
## koi_fpflag_co                     0.49023000
## koi_fpflag_ec                     0.36495656
```

The correlation values show that 'fp_flag_ss' has the highest correlation with False positive meaning that secondary signals has the highest impact on exoplanet being falsely identified. Every flag value has positive correlation with 'FALSE POSITIVE' disposition and anti-correlation with 'CoNFIRMED'.

RESULTS FOR RQ 3

Overall it can be concluded that False Positive values can be majorly due to A KOI that is observed to have a significant secondary event, transit shape, or out-of-eclipse variability, which indicates that the transit-like event is most likely caused by an eclipsing binary. However, self-luminous, hot Jupiters with a visible secondary eclipse will also have this flag set, but with a disposition of PC. Since the flag 'fp_flag_ss' has the highest correlation including all the different Comment values and flag values.[8]

Figures

Comment	Description
"NO_COMMENT"	No specific comment or observation provided
"MOD_ODDEVEN_DV"	there are discernible fluctuations in the brightness of the celestial object during alternate occurrences, with the variations being of moderate strength.
"CENT_SATURATED"	the central regions of the observed target exhibit extremely high brightness levels that have maxed out the detector's capacity to accurately capture the intensity, potentially impacting the quality of the data obtained from those pixels.
"CENT_UNRESOLVED_OFFSET"	there are small, undetermined shifts or deviations present in the central regions of the observed data that cannot be definitively identified or resolved due to limitations in the measurement or observational techniques.
"DEPTH_ODDEVEN_ALT"	there are periodic changes in the depth of transit events, occurring alternately between different instances of transits. These variations could be indicative of various factors, such as differences in the properties of the transiting object or the geometry of the transit events themselves.
"CENT_RESOLVED_OFFSET"	any observed shifts or deviations in the central pixels are clearly identifiable and can be accurately characterized, indicating a higher level of clarity and resolution in the observations.
"LPP_DV"	periodic variations in stellar brightness after removing long-term trends, often indicative of stellar activity or intrinsic variability.
"SEASONAL_DEPTH_DV"	changes in exoplanet transit depths fluctuate periodically over different seasons, possibly due to orbital or observational effects.
"LPP_ALT"	Low-frequency phenomena observed in the alternate transits
"MOD_NONUNIQ_ALT"	there are several plausible but differing interpretations of the data collected during alternate transit events, and these interpretations are moderately uncertain.
"DEPTH_ODDEVEN_DV"	brightness changes in stars after accounting for long-term trends, possibly caused by exoplanet transits.
"CENT_FEW_DIFFS"	minimal changes in brightness captured by the telescope's central pixels during the observation of a celestial object.
"INDIV_TRANS_RUBBLE_SKYE_ZUMA_TRACKER"	to a single occurrence of a celestial body passing in front of its host star, detected by the specified instrument.

Figure 1: Alt text

"TRANS_GAPPED"	interruptions or missing data segments within the recorded occurrences of celestial bodies passing in front of their host stars.
"SWEET_NTL"	a favorable balance between useful data (the "look") and background noise, enhancing the quality of astronomical observations.
"CENT_UNCERTAIN"	ambiguity or lack of clarity in the data recorded by the central pixels of the telescope during observations.
"INDIV_TRANS_SKYE"	detection of a single instance of a celestial body passing in front of its host star, as recorded by the instrument.
"INDIV_TRANS_RUBBLE"	the identification of a single occurrence of a celestial body passing in front of its host star amid noisy or obscured data.
"INDIV_TRANS_CHASES_MARSHALL_ZUMA"	the detection of a single instance of a celestial body passing in front of its host star across multiple observation points or instruments. Chases, Marshall, and Zuma are the names of specific regions or sectors within the Kepler field of view where observations are conducted.
"MOD_POS_DV"	the presence of observable events with slightly increased brightness levels after removing systematic trends from the data.
"INDIV_TRANS_RUBBLE_SKYE"	detection of a single passage of a celestial object in front of its host star amidst noisy or obscured data, likely from the Skye observation region.
"CENT_NOFITS"	the data captured by the central pixels of the telescope do not conform to the expected pattern or model.
"INDIV_TRANS_SKYE_ZUMA_TRACKER"	detection of a singular event where a celestial body passed in front of its host star, likely recorded using the Zuma tracking system in the Skye observation region.
"INDIV_TRANS_ZUMA"	detection of a single instance where a celestial body passed in front of its host star, specifically observed within the Zuma observation region.
"RESIDUAL_TCE"	the detection of a lingering presence of a celestial body passing in front of its host star, evident in the data after initial observations.
"INDIV_TRANS_CHASES"	the detection of a single occurrence of a planet crossing in front of its parent star.
"DEEP_V_SHAPED"	a significant and distinct dip in brightness recorded by NASA's Kepler telescope, indicating a possible planetary transit with a unique profile.
"MOD_SEC_DV"	the presence of additional effects beyond the primary transit signal in the processed brightness data from NASA's Kepler telescope.

Figure 2: Alt text

"CENT_KIC_POS"	a positive position' implies that the central region of the observed object is brighter or more pronounced.
"INDIV_TRANS_CHASES_MARSHALL"	'individual transit observed in Chases and Marshall' signifies the detection of a single instance of a planet passing in front of its host star at both Chases and Marshall observatories."
"SAME_NTL_PERIOD"	same noise-to-look period observed' indicates consistent patterns of noise relative to observation periods, suggesting stable observational conditions.
"CENT_CROWDED"	central pixels are crowded' suggests a high density of objects or sources in the central region of the observed field.
"MOD_SEC_ALT"	moderate secondary phenomena observed in the alternate transits' indicates the presence of additional effects during alternate occurrences of planetary transits.
"PLANET_IN_STAR"	"a planet is observed within a star" suggests the detection of a planet passing in front of its host star, causing a temporary dimming in the star's brightness.
"HAS_SEC_TCE"	"secondary transiting celestial object observed" indicates the detection of an additional celestial body passing in front of its host star besides the primary planet being observed.
"MOD_ODDEVEN_ALT"	moderate odd-even variations observed in the alternate transits" suggests periodic differences in brightness between alternate occurrences of planetary transits, indicating potential atmospheric or orbital variations.
"CENT_FEW_MEAS"	central pixels exhibit few measurements" indicates a scarcity of recorded data points within the central region of the observed field.
"INCONSISTENT_TRANS"	"inconsistent transits observed" suggests irregular occurrences of planetary transits, possibly indicating variability in the orbit or properties of the observed celestial bodies.
"SWEET_EB"	the detection of a binary star system exhibiting eclipsing behavior, where one star periodically passes in front of the other.
"INDIV_TRANS_MARSHALL_SKYE"	the detection of a single instance of a planet passing in front of its host star at both the Marshall and Skye observatories.
"EPHEM_MATCH"	the observed timing of celestial events aligns with the predicted schedule based on their calculated ephemerides, indicating accurate orbital tracking.

Figure 3: Alt text

"HALO_GHOST"	the detection of a faint, halo-like artifact in the data, possibly caused by scattered light or instrumental effects.
"INDIV_TRANS_RUBBLE_SKYE_ZUMA"	Detection of a single instance of a planet passing in front of its star, observed from the Skye and Zuma observatories.
"ALL_TRANS_CHASES"	Every occurrence of a planet passing in front of its star was detected at the Chases observatory.
"INDIV_TRANS_MARSHALL"	Detection of a single planetary transit at the Marshall observatory.
"INDIV_TRANS_MARSHALL_ZUMA"	Single instance of a planet passing in front of its star, observed at both the Marshall and Zuma observatories.
"INDIV_TRANS_CHASES_SKYE"	Detection of a single planetary transit at both the Chases and Skye observatories.
"IS_SEC_TCE"	Detection of another celestial body, besides the primary planet, passing in front of its star.
"INDIV_TRANS_CHASES_MARSHALL_SKYE"	Detection of a single planetary transit at all three observatories: Chases, Marshall, and Skye.
"INDIV_TRANS_RUBBLE_MARSHALL_SKYE"	Detection of a single planetary transit observed from the Marshall and Skye observatories, possibly with noise or interference.
"INDIV_TRANS_SKYE_ZUMA"	Detection of a single planetary transit observed at the Skye and Zuma observatories.
"MOD_NONUNIQ_DV"	Presence of moderate ambiguity in processed brightness data, indicating challenges in accurately interpreting transit signals.

Figure 4: Alt text

Flag	Significance
koi_fpflag_nt = Not Transit-Like Flag	A KOI whose light curve is not consistent with that of a transiting planet. This includes, but is not limited to, instrumental artifacts, non-eclipsing variable stars, and spurious (very low SNR) detections.
koi_fpflag_ss = Stellar Eclipse Flag	A KOI that is observed to have a significant secondary event, transit shape, or out-of-eclipse variability, which indicates that the transit-like event is most likely caused by an eclipsing binary. However, self-luminous, hot Jupiters with a visible secondary eclipse will also have this flag set, but with a disposition of PC.
koi_fpflag_co = Centroid Offset Flag	The source of the signal is from a nearby star, as inferred by measuring the centroid location of the image both in and out of transit, or by the strength of the transit signal in the target's outer (halo) pixels as compared to the transit signal from the pixels in the optimal (or core) aperture.
koi_fpflag_ec = Ephemeris Match Indicates Contamination Flag	The KOI shares the same period and epoch as another object and is judged to be the result of flux contamination in the aperture or electronic crosstalk

Figure 5: Alt text

References:

[1] NASA Exoplanet Archive. KOI (Kepler Objects of Interest). Exoplanet Archive. https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=koi

[2] Phys.org. (2011, September). Heavy metal stars look to be the perfect environment for Earth-like planets. https://phys.org/news/2011-09-heavy-metal-stars-earth-like-planets.html

[3] Wikipedia contributors. Planetary habitability. Wikipedia. https://en.wikipedia.org/wiki/Planetary_habitability

[4] NASA Exoplanet Science Institute. Kepler Candidate Overview Table Column Definitions. Exoplanet Archive. https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html

[5] Matteo Fasiolo, (Year). mgcViz: An R Package for Visual Inference and Diagnostics with Additive Models. GitHub repository. Available at: https://github.com/mfasiolo/mgcViz

[6] Exoplanet exploration: Planets beyond our solar system (2015) NASA. Available at: https://exoplanets.nasa.gov/ (Accessed: 10 April 2024).

[7] NASA. (2022, June). Hubble Focus: Exoplanets. Retrieved from https://www.nasa.gov/wp-content/uploads/2022/06/hubble_focus_exoplanets_june2022.pdf

[8] Howell, E. and Harvey, A. (2022) The 10 most Earth-like exoplanets, Space.com. Available at: https://www.space.com/30172-six-most-earth-like-alien-planets.html.