Data Science for Astronomy

Application Area Project

Rohan Sai Nalla Siddharth Tayi Mohammed Junaid Shaik
Ashwathi Subramanian Puja Kumari Lnu

Since the dawn of time, people look into the sky stirring up questions out curiousity to know the very meanning of our existence. The questions like why are the seasons as they are? Why are we the only conscious people? Why did earth come to be? What is our purpose? Are we alone? came into consideration. These questions are the foundation to modern day astronomy. Considering the current influx of data in every domain astronomy is one of the first fields to leverage on this transition. With the addition of Data the very nature of discoveries changed. To elucidate this transistion this report present key findings of Exoplanet discoveries considering the timeline from the inception to the current epoch.

Table of contents

Exoplanet Archive Analysis

Premise of this analysis is to answer the question of **Are we alone in the Universe?**, the entire expedition of exoplanets itself is founded on this question. The quest to find more planets which closely resemble to us in all the aspects, making it habitable for other lifeforms. This question is also tied closely to the answering another question that stems from the dawn of time, which is the very purpose of our existence. To Exoplore and understand space comprehensively and scientifically, data plays an important role.

What is an exoplanet?

To put it simply, any planet outside of our solar system is an exoplanet. Rogue planets are exoplanets that don't orbit any star and instead orbit the Milky Way's center.

The majority of the identified exoplanets are in a very tiny area of our galaxy, the Milky Way. According to NASA's Kepler Space Telescope, the galaxy has more planets than stars.

We may observe compositions ranging from highly rocky (like Earth and Venus) to very gas-rich by measuring exoplanet sizes (diameters) and masses (weights) (like Jupiter and Saturn). Exoplanets are composed of elements comparable to those found in our solar system, although their compositions may vary. Water or ice may dominate certain planets, whereas iron or carbon may rule others. We've discovered lava worlds with molten oceans, fluffy planets with the solidity of Styrofoam, and dense cores of planets that are still circling their sun.

Exoplanet Discovery

In the quest of finding other planets which are closest to earth like in-terms of habitability, terrain etc. NASA introduced Kepler and TESS Satellites to observe and detect objects in vast amounts of space.

Transit Method

Kepler looks for planets using the *Transit Method*, when a planet passes in front of the star the data in wave form detects a decrease in brightness.

The transit method works only if the orbital plane of the objects is aligned to the line of sight of kepler. Assuming that the geometric probability of the axis is randomly distributed in space then the geometric probability of a transit is 0.5.

So to find planets you need to observe a lot of stars, half a million stars in Kepler field of view around 150k are chosen for observation.

All of the data collected either through Kepler or TESS is openly available to access through the public API sources through the NASA's Exoplanet-archive.

Importing Data

The data used in this report is sourced from mutiple archives. Both the datasets used in the report have similar attributes but consists of various exoplanets. In aggregation, both archives will have exoplanets from Kepler, TESS and CoRot Satellite data.

```
url1 <- "https://raw.githubusercontent.com/r0han99/Computational-Astronomy/master/Exo-Plan
url2 <- "https://raw.githubusercontent.com/r0han99/Computational-Astronomy/master/Exo-Plan
url3 <- "https://raw.githubusercontent.com/r0han99/Computational-Astronomy/master/Exo-Plan

data <- read.csv(url1)
all_exo <- read.csv(url2)
kepler_data <- read.csv(url3)
```

The three datasets `data`, `all_exo`, & `kepler_data`, represent the same attributes but for different exoplanets in each. Following are the list of attributes involved in the data.

```
tail(colnames(data),-1)
```

```
 [1] "name"                       "light_years_from_earth"
 [3] "planet_mass"                "stellar_magnitude"
 [5] "discovery_date"             "planet_type"
 [7] "planet_radius"              "orbital_radius"
 [9] "orbital_period"             "eccentricity"
[11] "solar_system_name"          "planet_discovery_method"
[13] "planet_orbital_inclination" "planet_density"
[15] "right_ascension"            "declination"
[17] "host_temperature"           "host_mass"
```

```
[19] "host_radius"                "orbital_period_value"
[21] "orbital_period_unit"        "planet_radius_value"
[23] "planet_radius_unit"         "orbital_radius_value"
[25] "orbital_radius_unit"
```

An Overview

```
head(data, n=3)
```

```
  X                 name light_years_from_earth    planet_mass
1 0 11 Comae Berenices b                     305  19.4 Jupiters
2 1   11 Ursae Minoris b                     410 14.74 Jupiters
3 2      14 Andromedae b                     247   4.8 Jupiters
  stellar_magnitude discovery_date planet_type  planet_radius orbital_radius
1             4.740           2007   Gas Giant 1.08 x Jupiter        1.29 AU
2             5.016           2009   Gas Giant 1.09 x Jupiter        1.53 AU
3             5.227           2008   Gas Giant 1.15 x Jupiter        0.83 AU
  orbital_period eccentricity solar_system_name planet_discovery_method
1       326 days         0.23            11 Com         Radial Velocity
2      1.4 years         0.08            11 UMi         Radial Velocity
3     185.8 days            0            14 And         Radial Velocity
  planet_orbital_inclination planet_density right_ascension  declination
1                         NA             NA    12h20m43.03s +17d47m34.3s
2                         NA             NA    15h17m05.89s +71d49m26.0s
3                         NA             NA    23h31m17.42s +39d14m10.3s
  host_temperature host_mass host_radius orbital_period_value
1             4742      2.70       19.00                326.0
2             4213      2.78       29.79                  1.4
3             4813      2.20       11.00                185.8
  orbital_period_unit planet_radius_value planet_radius_unit
1                days                1.08            Jupiter
2               years                1.09            Jupiter
3                days                1.15            Jupiter
  orbital_radius_value orbital_radius_unit
1                 1.29                  AU
2                 1.53                  AU
3                 0.83                  AU
```

```
head(data, n=3)
```

```
  X                 name light_years_from_earth    planet_mass
1 0 11 Comae Berenices b                   305  19.4 Jupiters
2 1   11 Ursae Minoris b                   410 14.74 Jupiters
3 2      14 Andromedae b                   247   4.8 Jupiters
  stellar_magnitude discovery_date planet_type  planet_radius orbital_radius
1             4.740          2007   Gas Giant 1.08 x Jupiter        1.29 AU
2             5.016          2009   Gas Giant 1.09 x Jupiter        1.53 AU
3             5.227          2008   Gas Giant 1.15 x Jupiter        0.83 AU
  orbital_period eccentricity solar_system_name planet_discovery_method
1       326 days         0.23            11 Com         Radial Velocity
2      1.4 years         0.08            11 UMi         Radial Velocity
3     185.8 days            0            14 And         Radial Velocity
  planet_orbital_inclination planet_density right_ascension  declination
1                         NA             NA    12h20m43.03s +17d47m34.3s
2                         NA             NA    15h17m05.89s +71d49m26.0s
3                         NA             NA    23h31m17.42s +39d14m10.3s
  host_temperature host_mass host_radius orbital_period_value
1             4742      2.70       19.00                326.0
2             4213      2.78       29.79                  1.4
3             4813      2.20       11.00                185.8
  orbital_period_unit planet_radius_value planet_radius_unit
1                days                1.08            Jupiter
2               years                1.09            Jupiter
3                days                1.15            Jupiter
  orbital_radius_value orbital_radius_unit
1                 1.29                  AU
2                 1.53                  AU
3                 0.83                  AU
```

Dimensionality - Attributes Involved.

Dimensions of `data`, `kepler_data`,`all_exo`

```
[1] 4251   26
```

```
[1] 2299   14
```

```
[1] 4575   23
```

Data Preprocessing

At this stage, we observed a lot of inconsistencies within the data that needs to be addressed in order to progress further in the analysis. These inconsistencies are omnipresent in any real-world data sets. We initially observed huge magnitude of missing values which are then selectively addressed.

```
gg_miss_upset(data)
```

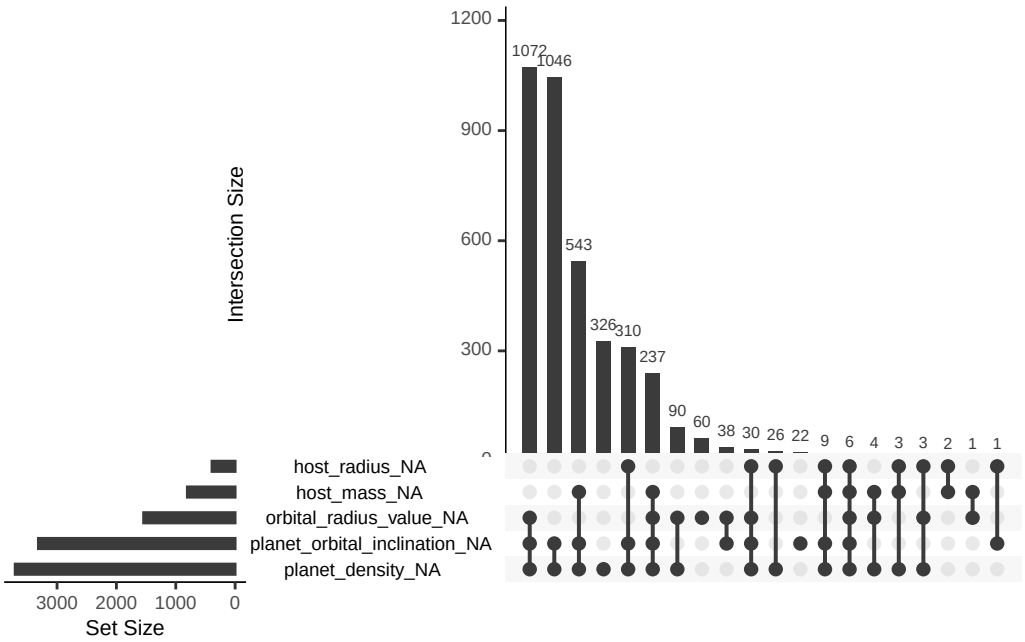


Figure 1: Missing Value Representation

The attributes, `planet_density`, `planet_oribital_inclination` have a significant number of missing values and are also unrelevant to our further analysis.

Dropping 'planet_density', 'planet_orbital_inclination'

```
df <- within(data, rm("planet_density","planet_orbital_inclination"))
```

Counting the number of 'NA' values in the dataset

```
countNas1 <- colSums(is.na(df))
```

```
countNas1
```

```
                      X                    name  light_years_from_earth
                      0                       0                      10
           planet_mass       stellar_magnitude         discovery_date
                      0                     107                       0
           planet_type           planet_radius         orbital_radius
                      0                       0                       0
        orbital_period            eccentricity       solar_system_name
                      0                       0                       0
planet_discovery_method         right_ascension             declination
                      0                       0                       0
       host_temperature               host_mass             host_radius
                    271                     805                     390
   orbital_period_value     orbital_period_unit     planet_radius_value
                      0                       0                       0
     planet_radius_unit    orbital_radius_value     orbital_radius_unit
                      0                    1541                       0
```

Dealing with Infinity Values

```
df <- do.call(data.frame, lapply(df, function(x) {
  replace(x, is.infinite(x) | is.na(x), 0)
})
)
```

Filling 'NA' values as Not Recorded

In a conventional approach, normally the best course of action to deal with numeric missing values is to perform either *Median Imputation* or *Mean Imputation*. In astronomical data, every thing seems to be out of place, the values though seem like outliers are true to nature. So an average representation can skew the analysis. We concurred that imputing the values with `Not Recorded` will be the best course of action.

```
df$light_years_from_earth[is.na(df$light_years_from_earth)] <- "Not Recorded"

df$host_mass[is.na(df$host_mass)] <- "Not Recorded"

df$host_temperature[is.na(df$host_temperature)] <- "Not Recorded"

df$host_radius[is.na(df$host_radius)] <- "Not Recorded"
```

```
df<- df %>% mutate(across(stellar_magnitude, ~replace_na(., median(., na.rm=TRUE))))
```

Exploratory Data Analysis

The Major Mode of Detection & Facility

```
ggplot(all_exo, aes(x=Discovery.Method)) + geom_bar(fill='orange') + coord_flip()
```

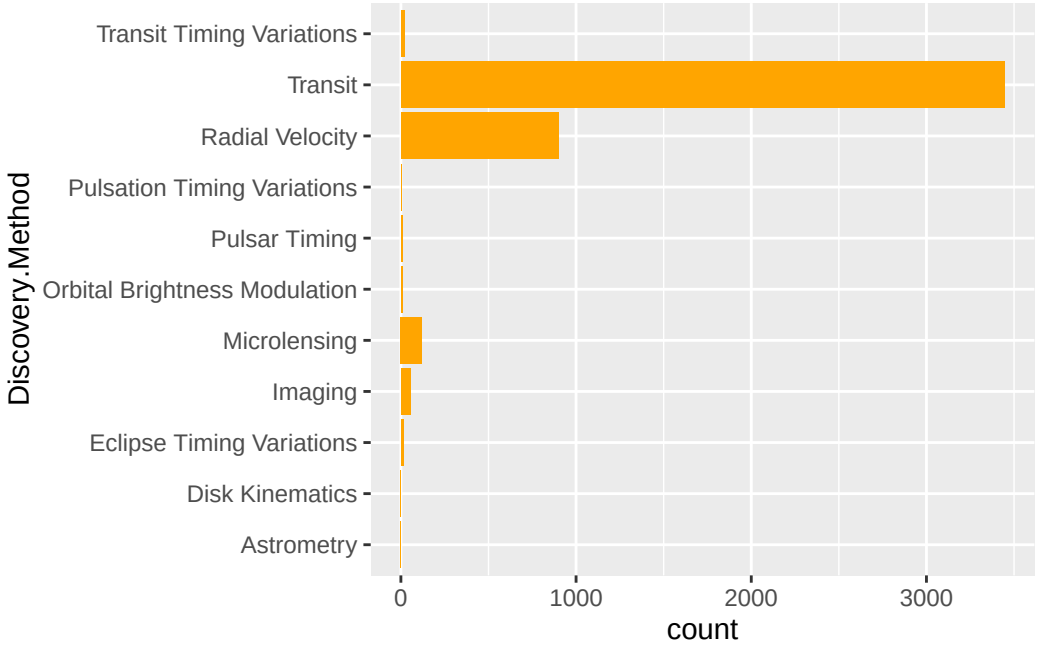


Figure 2: Methods used to detect a Exoplanet Signature

According to the data, the majority of the discoveries are by the Kepler Space Telescope, but there are significant other observatories that have contributed to the exploration. Citing from the Mode of detection chart, the go-to methodology for detecting an exoplanet is through the transit method, which seem quite plausible only through observing the minute change in luminosity in relation to time.

Exoplanet Discoveries Over the years

Although the science of exoplanets kicked of in the 1990's but the success of the space telescopes (x-ray and gamma ray detection) CoRoT and kepler. In 90s the discovery rate was 1 or 2 of exoplanets a year now with kepler it is about 100's of new planets a year are discovered

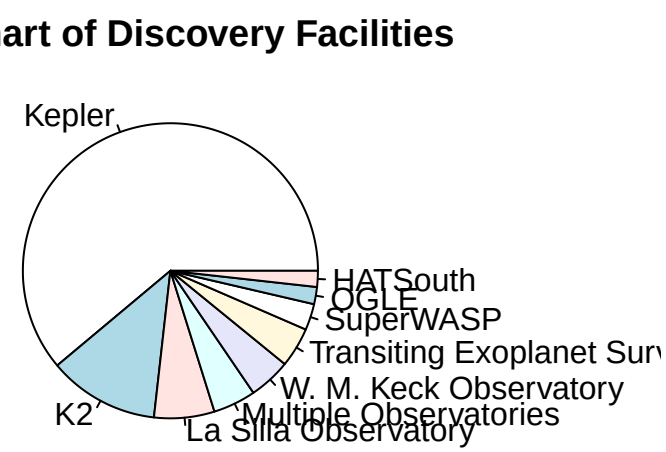


Figure 3: Facilities Used for Exoplanet Exoploration

In 1999, David Charbonneau and Greg Henry lead different research teams that detect a planet moving across the front of the star HD 209458 in the constellation Pegasus. This discovery enables astronomers to examine the planet's atmosphere, which they think includes water, oxygen, nitrogen, and carbon. Because of the planet's tight orbit to its star, its atmosphere is being stripped away, leaving a comet-like tail behind it. Artist's rendering of the HD 209458 system.

Orbital Period Vs Planet Radii

The orbital period (The time it takes to complete a single orbit around its host) in relation with the planet radii, illustrates the key information of its magnitude. This plot reveals key information about the oribital period of planets in relation to their size, it is evident from the plot that earth like planets (Super Earths) are having orbital periods closer to the that of earth days. Whereas the planets which are in the magnitudes of Jupiter are seen to have larger orbital periods which is a natural trend even in our solar system.

Planets Discovered After the introduction of TESS & Kepler

A Delta II rocket carrying NASA's Kepler space telescope lifts off from Florida's Cape Canaveral Air Force Station. Kepler will spend four years staring at a region of sky with

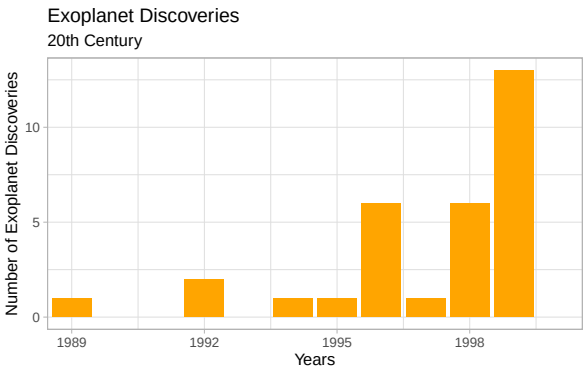


Figure 4: 20th Century

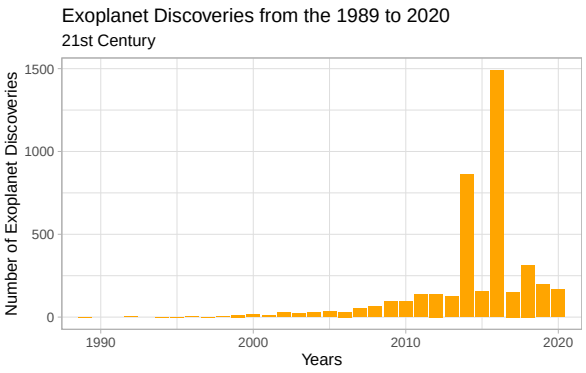


Figure 5: 21st Century

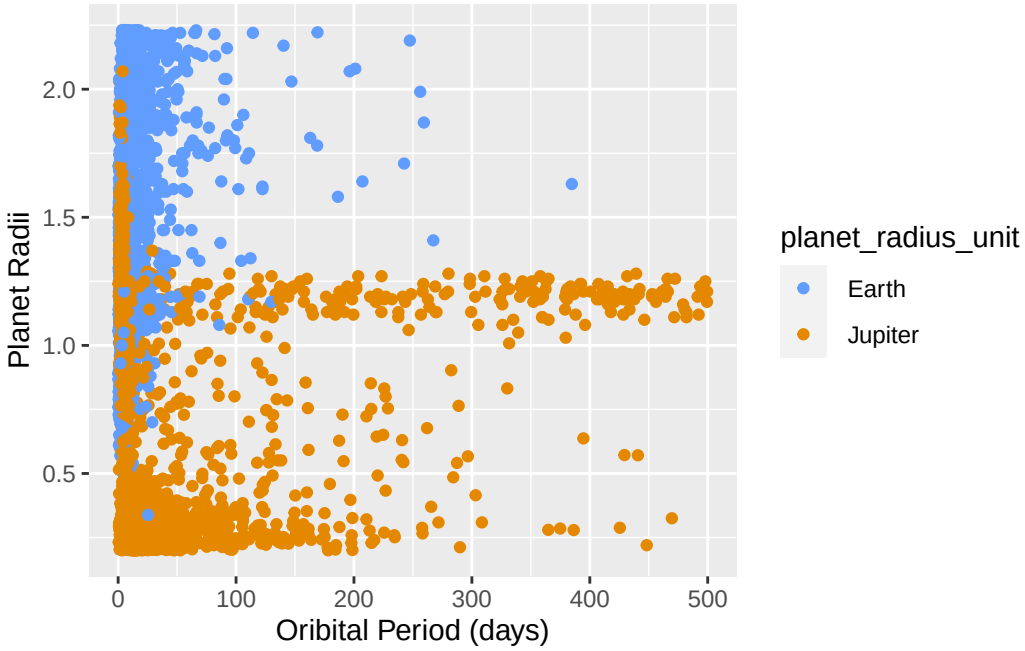


Figure 6: Exoplanets Orbital Period Vs Radius Graph

150,000 stars, looking for small dips in illumination when planets pass in front of some of them. Before a malfunction terminates its main mission in 2013, the pioneering spacecraft will have discovered over 1,000 verified exoplanets—a gold rush of discovery.

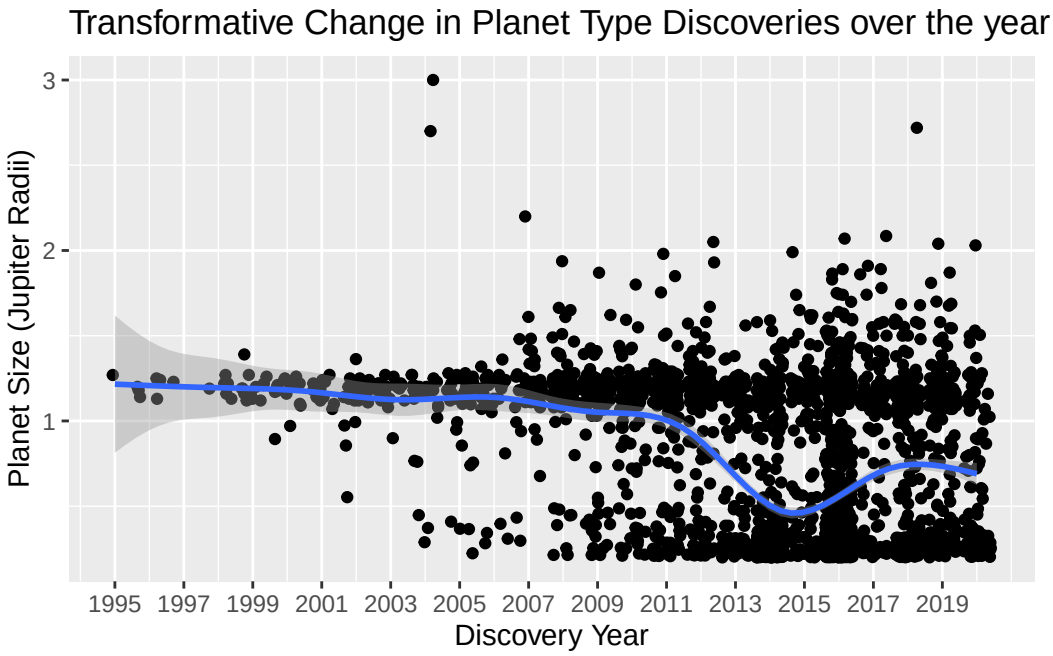
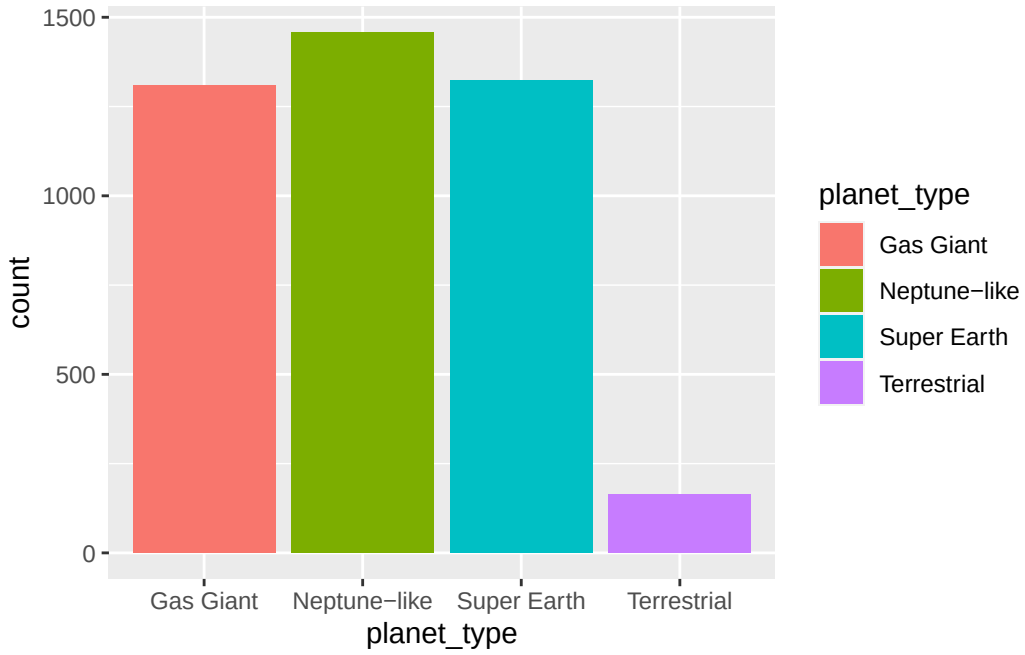


Figure 7: Transformative Change in Planets Discovered

Types of Exoplanets

In the sample of data we have, the majority of the dataset is populated with the planets which are in the magnitudes of Jupiter and Neptune, there are also significant others which are barely less than Neptune called as the Super Earths. Discovery of these types significantly increased after the introduction of Kepler which had a primary mission to observe planets that produce only a minor signature. This can be observed in the above figures 6 and 7 where the number of planets after 2009 fall in between the jupiter radii of 0-1.

```
ggplot(df, aes(x=planet_type,fill=planet_type)) +
  geom_bar()
```



From the sample of data, there are main 4 categories with respect to the size of the planets: 'Gas Giant'/Jupiter, 'Neptune-like', 'Super Earth', and 'Terrestrial'. Most of the planets that are explored are the size of 'Neptune' and the least number of planets are the size of the 'Earth'/Terrestrial Analysis before kepler (pre kepler) concluded that the planets observed were much larger than earth and Neptune, but was later proven wrong after Kepler's discovery. In current era, most of the planets discovered are closer to the size of 'Earth'. This is due to the fact that, kepler was tuned to observe the transits from planets which are smaller.

Exoploring the Habitable Exoplanets

The habitable zone is the area of space around a star where temperatures are neither too hot nor too cold for liquid water to exist on neighbouring planets' surfaces. The habitable zone is essentially influenced by two factors: the mass and period of the star. The spectral type and brightness of a star change as it evolves. The habitable zone is defined as the range of orbital distances from a star at which a planet may sustain liquid water. This implies that water is necessary for life to exist, which is not always the case.

Properties of a Habitable Planet

A star's output has been constant for billions of years; orbit it. Moreover it should be far enough away from the star for its surface water to be liquid rather than solid at the proper temperature. Also, the exoplanet should have a circular orbit, so that during the course of its "year," the same conditions prevail. In a quantified sense, the planets in between 0.95 AU and

1.67 AU are deemed to be habitable, considering the Star is in the same classification as our Sun which is the K-type classification.

The star classification in this context is imperative to host life on a planet, if the Star type os "O" or "B" which are 'super massive stars' though the exoplanet is in habitable zone, the planet's temperature will be too hot to hold any liquid water or even life.

Throughout the kepler's expedition there have been certain planets that are deemed to be habitable by scientists for the unique properties they hold. Through our research we found out the code names for these **Potentially Habitable** Planets. The following section includes the detailed look into *three* such planets which are scientifically agreed to be habitable. To perform this we use the `kepler_data` which is a sample from the entire cummulative dataset from NASA's Kepler Exoplanet Archive.

NASA's **Exoplanet Catalog** has a really good illustrated representation of the planet and its properties on it website, we used this report to cross-validate our findings and learn more about the object of interest.

- Kepler 442 b
- Kepler 22 b
- Kepler 62 f

The Kepler Naming convention is informative in itself. The number and alphabet, represents the Code number and the sequence of planet. The alphabet, a-z represents the number of planets in the entire system of that KOI (Kepler Object of Interest). For instace, kepler 62 f is a 5 planet system with "f" being in the habitable zone. All of these can be visually seen through *NASA's Exoplanet Eyes* by selecting the system field.

Kepler's Habitable Planets

```
k442b <- filter(kepler_data, kepler_name=='Kepler-442 b')
k22b <- filter(kepler_data, kepler_name=='Kepler-22 b')
k62f <- filter(kepler_data, kepler_name=='Kepler-62 f')
```

Kepler 442 b ~ Illustration

Kepler-442 b is a super Earth exoplanet in the habitable zone of a K-type star (similar to that of our sun). It has a mass of 2.36 Earths, takes 112.3 days to complete one circle around its star, and is 0.409 AU away from it. Its finding was revealed in the year 2001.

Planet Name	Disposition	Host-Star Temperature(k)	Host-Star Radii(sun)	Planet Radii(earth)	Orbital Period(days)
Kepler-442 b	CONFIRMED	4401	0.595	1.3	112.3031

Kepler 22 b - Illustration

Kepler-22b is a super-Earth that might be completely covered by a mega ocean. The judgment is still out on Kepler-22b's real composition; at 2.4 times the radius of the Earth, it may potentially be gaseous. According to current computer simulation, an ocean planet turned on its side - similar to our solar system's ice giant, Uranus - turns out to be reasonably livable. An exoplanet of Earth's size, at a similar distance from its sun and covered in water, might have an average surface temperature of roughly 60 degrees Fahrenheit, according to researchers (15.5 Celsius). Because of its extreme tilt, the planet's north and south poles would alternately be drenched in sunshine and darkness for half a year as it circled its star.

Planet Name	Disposition	Host-Star Temperature(k)	Host-Star Radii(sun)	Planet Radii(earth)	Orbital Period(days)
Kepler-22 b	CONFIRMED	5516	0.886	2.34	289.8641

Kepler 62 f - Illustration

Kepler-62 f is a super Earth exoplanet in the habitable zone of a K-type star. It has a mass of 35 Earths, takes 267.3 days to complete one circle around its star, and is 0.718 AU away from it. It was discovered in 2013 and was publicized.

Planet Name	Disposition	Host-Star Temperature(k)	Host-Star Radii(sun)	Planet Radii(earth)	Orbital Period(days)
Kepler-62 f	CONFIRMED	4926	0.662	1.43	267.2825

Conclusion

This report consists of base level analysis on just a small sample of enourmous exoplanet archive that is openly available. The objective of this project is is to elucidate how the influx of data in the current age is influencing the scientific discoveries. Pre-kepler era, where the observations are done from the ground base systems used to rely on sluggish processes which take a lot of time and effort to validate only a few objects. But, in the age of data,

the information obtained from Kepler is enormous with hundreds of parameters related to a particular exoplanet to be ready at disposal to categorise with Data Mining. The programmatic illustrations and statistical analysis help scientists to grasp huge amount of information with some simple code.

Identifying our Bias — There are multiple segments in the report where we can observe an induced bias.For the analysis of habitable exoplanets we could only use a miniature sample of the entire kepler archive, narrowing down our analysis to only three *Earth Like* objects. Same applies to the original data which includes all Exoplanets from different observation facilities and methods. The sheer lack of data accounts for not having a good analytical representation.

Session Info

```
sessionInfo()
```

```
R version 4.2.2 (2022-10-31)
Platform: aarch64-apple-darwin20 (64-bit)
Running under: macOS Monterey 12.6.1

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] UpSetR_1.4.0    naniar_0.6.1    forcats_0.5.2   stringr_1.4.1
 [5] purrr_0.3.5     readr_2.1.3     tidyr_1.2.1     tibble_3.1.8
 [9] ggplot2_3.4.0   tidyverse_1.3.2 dplyr_1.0.10

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.9          lattice_0.20-45     lubridate_1.9.0
 [4] assertthat_0.2.1    digest_0.6.30       utf8_1.2.2
 [7] R6_2.5.1            cellranger_1.1.0    plyr_1.8.8
```

```
[10] backports_1.4.1     reprex_2.0.2        visdat_0.5.3
[13] evaluate_0.18       highr_0.9           httr_1.4.4
[16] pillar_1.8.1        rlang_1.0.6         googlesheets4_1.0.1
[19] readxl_1.4.1        rstudioapi_0.14     Matrix_1.5-1
[22] rmarkdown_2.18      splines_4.2.2       labeling_0.4.2
[25] googledrive_2.0.0   munsell_0.5.0       broom_1.0.1
[28] compiler_4.2.2      modelr_0.1.10       xfun_0.35
[31] pkgconfig_2.0.3     mgcv_1.8-41         htmltools_0.5.3
[34] tidyselect_1.2.0    gridExtra_2.3       fansi_1.0.3
[37] crayon_1.5.2        tzdb_0.3.0          dbplyr_2.2.1
[40] withr_2.5.0         grid_4.2.2          nlme_3.1-160
[43] jsonlite_1.8.3      gtable_0.3.1        lifecycle_1.0.3
[46] DBI_1.1.3           magrittr_2.0.3      scales_1.2.1
[49] cli_3.4.1           stringi_1.7.8       farver_2.1.1
[52] fs_1.5.2            xml2_1.3.3          ellipsis_0.3.2
[55] generics_0.1.3      vctrs_0.5.1         tools_4.2.2
[58] glue_1.6.2          hms_1.1.2           fastmap_1.1.0
[61] yaml_2.3.6          timechange_0.1.1    colorspace_2.0-3
[64] gargle_1.2.1        rvest_1.0.3         knitr_1.41
[67] haven_2.5.1
```