Exploring the Evolution and Characteristics of Exoplanets

Mohan Silambarasu Elangkumaran
G01524739 (AIT 580 - 002)
MS Applied Information Technology
George Mason University
Fairfax, VA, USA
melangku@gmu.edu

I. ABSTRACT

This study dives into the exciting field of exoplanet discovery, analyzing a dataset to uncover trends, relationships, and insights into planetary characteristics. The goal is to explore how discovery methods have evolved, identify correlations between planetary attributes like distance, mass, and density, and pinpoint potentially habitable exoplanets based on their equilibrium temperatures. To achieve this, I used tools like Python, R, MySQL, and AWS services to clean, analyze, and visualize the data. By combining SQL for statistical queries with cloud-based databases, the research highlights the importance of using advanced analytics in astronomy. Beyond presenting tables and graphs, the study interprets the findings to provide meaningful insights into how exoplanet discoveries have progressed and what makes certain planets stand out as potentially habitable. This paper not only showcases the power of data analysis but also emphasizes its role in deepening our understanding of the universe.

II. INTRODUCTION

The discovery of exoplanets has transformed our knowledge of planetary systems, expanding our understanding of the universe. From the first exoplanet detection in the 1990s to the thousands discovered since then, advancements in technology have been central to this progress. In this project, I analyzed a detailed dataset of exoplanets to answer key questions about their discovery and characteristics. The study focuses on trends in how exoplanets are discovered, the relationship between their distance and discovery methods, and the identification of potentially habitable planets based on temperature. By using Python, R, MySQL, AWS RDS , AWS Glue Data Brew, I combined data cleaning, analysis, and visualizations to uncover meaningful insights. This research aims to bridge the gap between data analytics and planetary science, offering a new perspective on exoplanet discoveries.

III. LITERATURE REVIEW

The exploration of exoplanets has seen significant advancements through various data-driven and technological approaches, as highlighted in several key studies. The study "The Exploration of Habitable Exoplanets using Data Mining Algorithms and Data Manipulation" [1] focused on applying data mining techniques like deduplication and Circumstellar Habitable Zone (CHZ) calculations, alongside algorithms such as K-Nearest Neighbor (KNN). By using the NASA Exoplanet Archive dataset[4], the researchers identified planets with potential habitability, such as GJ 143b and L 98-59c, emphasizing the importance of mass, density, and radius in determining habitability. This aligns with my third research question regarding the relationship between mass, density, and radius in assessing habitability.

Another study, "Exoplanet Detection Methods" [2], delves into the evolution of discovery methods such as radial velocity, transit photometry, and direct imaging. It discusses notable discoveries, including 51 Pegasi b and the HR 8799 planetary system, highlighting how advancements in detection techniques have enabled exploration across varying distances. This study provides foundational insights into my research questions about the evolution of discovery methods and their correlation with technological advancements and planetary distances.

Finally, "Identifying Exoplanets with Machine Learning Methods: A Preliminary Study" [3] explored supervised and unsupervised machine learning techniques to identify exoplanets using the NASA Exoplanet Archive dataset[4]. The use of models like Naïve Bayes and decision trees achieved high classification accuracy, while unsupervised learning methods examined the relationships between mass, radius, and density. This approach not only demonstrated the efficiency of machine learning over traditional methods but also supported my research question on how technological advancements contribute to exoplanet discovery. Collectively, these studies reinforce the importance of leveraging data-driven methods and evolving technologies to understand exoplanet characteristics and their potential habitability.

IV. DATASET

The selected data set contains details of various exoplanets discovered over a period of 1992 to 2024 such as planetary characteristics, discovery methods, and other characteristics which can be used to detect potential habitable planets. The dataset contains 36557 entries where it also includes multiple fields such as planetary mass, radius, density, distance, discov-

ery year, equilibrium temperature, orbital period, and number of stars and moon associated with each system.

This exoplanet Dataset consists of all data types in NOIR from Nominal , Ordinal , Interval and Ratio. The below Table 1 provides classification of each column in the cleaned dataset as per the four data types. From the above Table 1 it is

TABLE I
DATASET AND NOIR TYPES

Column Name	NOIR	Column Name	NOIR
Serial Number	Ordinal	Discovery Facility	Nominal
Planet Name	Nominal	Discovery Instrument	Nominal
Number of Stars	Ratio	Orbital Period [days]	Ratio
Number of Moons	Ratio	Planet Mass [Earth Mass]	Ratio
Discovery Method	Nominal	Planet Density [g/cm**3]	Ratio
Discovery Year	Interval	Equilibrium Temperature [K]	Interval
Discovery Locale	Ordinal	Distance [pc]	Ratio

evident that Columns like Planet Name , Discovery Method, Discovery Instrument, Discovery Facility , Discovery Locale follows Nominal data type where it represents the data that categorical without any specific order. The ordinal data type such as Serial Number column follows categorical with order. The interval data type represents columns such as Discover Year , Equilibrium Temperature where it includes intervals and has no absolute zero point. The Ratio columns are those in which the data can be classified into ratios such as Distance , Planet Mass , Planet Radius etc.

This data set is extracted from NASA's Exoplanet Archive [4] as it provides accurate information on exoplanets that were discovered beyond our solar system , making this an ideal data set to address the research questions posed.

V. RESEARCH QUESTIONS

1) **How has the discovery of exoplanets evolved over the years and are methods linked to technical advancements?**

   Exploring the history of exoplanet discovery reveals how technological advancements played a crucial role in detecting planets beyond our solar system which are potential planets for life to sustain. This can also help us in future research and investments in selected types of technologies used in the field of astronomy.

2) **Is there any relation between the distance of exoplanets from Earth and their discovery methods or years?**

   Analyzing the relationship between the distance of exoplanets and various techniques used to detect them can help to understand which methods work best for detecting planets at longer distances. This understanding could streamline the discovery process, making it more efficient and reducing time and effort in future explorations.

3) **How does the mass of an exoplanet influence its density and radius across various planets and Which exoplanets fall within the habitable equilibrium temperature range (175K – 270 K)?**

   Exploring relationship between mass , density and radius of various planets help us to understand more about the planet structure and composition. Identifying planets which fall in the habitable equilibrium temperature range of 175 K – 270 K help us to narrow down our research on these planets which are a potential candidate for second Earth.

VI. TOOLS AND METHODS

In this project, Python and R were used for cleaning, analysis, and visualization of the exoplanets dataset. Python libraries such as `matplotlib`, `pandas`, `numpy`, and `seaborn`, along with R libraries like `dplyr` and `ggplot2`, were used extensively in understanding multiple trends and patterns across various exoplanets.

For Descriptive analysis, I worked with MySQL. I created a database schema in a MySQL database, loaded the dataset, and ran queries to explore and understand the data.

To efficiently handle this large dataset of exoplanets discovered so far, I had used AWS service. With the help of AWS , I had uploaded this large dataset into an S3 bucket, set up an RDS database, and used an EC2 instance to integrate MySQL with AWS. This setup allowed me to run complex queries and explore the data seamlessly. The uploaded S3 bucket dataset is then used for AWS Glue Data Brew analysis which provides the column statistics and summary about columns.The combination of these tools enabled me to analyze this complex dataset and extract meaningful insights to address my research questions.

VII. RESULTS AND ANALYSIS

A. Descriptive Statistics Using SQL

In this project, SQL is used to analyze the exoplanet dataset and few queries are run to explore the planets. Initially , existing database are checked in SQL and then a new database AIT580PROJECT is created for this project. After creating database, A new schema along with column name and data type is defined in the planets table. Once Initial Setup and Schema is done, the csv dataset is loaded using load data local infile command into the planets table. After loading the csv file into planets table , A sample of the whole datafirst few rows are displayed as shown in below Figure 1:

planet_name	number_of_sta...	number_of_moo...	discovery_meth...	discovery_year	discovery_loca...	discovery_facility	discovery_instrument	orbital_peri...
11 Com b	2	0	Radial Velocity	2007	Ground	Xinglong Station	Coude Echelle Spectrograph	0
11 Com b	2	0	Radial Velocity	2007	Ground	Xinglong Station	Coude Echelle Spectrograph	326.03
11 Com b	2	0	Radial Velocity	2007	Ground	Xinglong Station	Coude Echelle Spectrograph	323.21
11 UMi b	1	0	Radial Velocity	2009	Ground	Thueringer Landessternwarte Tautenburg	Coude Echelle Spectrograph	516.22
11 UMi b	1	0	Radial Velocity	2009	Ground	Thueringer Landessternwarte Tautenburg	Coude Echelle Spectrograph	0
11 UMi b	1	0	Radial Velocity	2009	Ground	Thueringer Landessternwarte Tautenburg	Coude Echelle Spectrograph	516.22
14 And b	1	0	Radial Velocity	2008	Ground	Okayama Astrophysical Observatory	HIDES Echelle Spectrograph	186.76
14 And b	1	0	Radial Velocity	2008	Ground	Okayama Astrophysical Observatory	HIDES Echelle Spectrograph	0
14 And b	1	0	Radial Velocity	2008	Ground	Okayama Astrophysical Observatory	HIDES Echelle Spectrograph	185.84
14 Her b	1	0	Radial Velocity	2002	Ground	W. M. Keck Observatory	HIRES Spectrometer	1773.4

Fig. 1. Preview of the loaded dataset in MySQL

1) Total number of exoplanets discovered by each Instrument: A SQL query is run to get the total number of exoplanets discovered so far according to each instrument. From the below Figure 2 , it is clear that Kepler CCD Array is most vital instrument where a total of 28817 planets were discovered using that instrument. This is followed by TESS CCD Array with 1498 discovery of exoplanets , iKon-L CCD Camera with 1148 and other multiple instruments used in in the discovery of 660 planets. The HIRES spectrometer is the instrument used very less in terms of discovering exoplanets.

Instrument	Number_of_Planets
Kepler CCD Array	28817
TESS CCD Array	1498
iKon-L CCD Camera	1148
Multiple Instruments	660
HIRES Spectrometer	630

Fig. 2. Count of planets discovered by various discovery instruments

2) Finding the earliest discovered Exoplanets: The SQL query here is used to get the most earliest discovered exo planet along with the method used to detect planet. The below Figure 3 indicates the list of earliest discovered planets. From the above figure 3 , it is seen that the earliest discovered exoplanets are discovered in the year of 1992. The planets are PSR B1257+12 c and PSR B1257+12 d . Both these exoplanets were discovered with the help Pulsar Timing Method back in 1992.

discovery_year	planet_name	discovery_method
1992	PSR B1257+12 c	Pulsar Timing
1992	PSR B1257+12 d	Pulsar Timing

Fig. 3. Earliest discovered exoplanets

3) Finding the latest discovered Exoplanets: An SQL is run to get the list of latest discovered Exoplanets. Figure 4 displays the list of such exoplanets.

From the above figure 4 , it is clear that the most recent exo planets discovered were this year 2024. The total number of exoplanets discovered this year are around 428. This query returns the name of the planet and the year and the method used to detect such exoplanets. The Methods used this year are Imaging, Radial Velocity, Transit , Imaging , Microlensing , Pulsar Timing and Transit Timing variations.

B. Exploratory Analysis and Visualization using Python

1) Data Cleaning in Python: After Setting up the initial librarys , the dataset is then cleaned and transformed using Python. Now, The dataset loaded in the pandas data frame is then cleaned by inserting serial number , dropping duplicated , unwanted columns. Now the dataset is cleaned , ensures accurate dataset with appropriate columns , without duplicates and filtering out redundant entries.

discovery_year	planet_name	discovery_method
2024	2MASS J03590...	Imaging
2024	2MASS J110119...	Imaging
2024	2MASS J11550...	Imaging
2024	2MASS J21252...	Imaging
2024	Barnard b	Radial Velocity
2024	BD-14 3065 b	Transit
2024	CI Tau c	Radial Velocity
2024	G 196-3 b	Imaging
2024	Gaia22dkvL b	Microlensing
2024	GJ 238 b	Transit
2024	GJ 900 b	Imaging
2024	Gliese 12 b	Transit
2024	HD 104067 c	Radial Velocity
2024	HD 118203 c	Radial Velocity
2024	HD 134606 b	Radial Velocity
2024	HD 134606 c	Radial Velocity
2024	HD 134606 d	Radial Velocity
2024	HD 134606 e	Radial Velocity
2024	HD 134606 f	Radial Velocity
2024	HD 21520 b	Transit
2024	HD 222237 b	Radial Velocity
2024	HD 48948 b	Radial Velocity

Fig. 4. Earliest discovered exoplanets

2) Summary Statistics of the Exoplanet Dataset: Using python, a summary statistics is provided for all the 36,557 exoplanets which is composed of multiple columns. From the below summary it is clear that most of the exoplanets discovered fall between 1992 and 2024 which peaked in the year 2015. The planets in general have short orbital period less than 10 median days and size of planets are usually around 2.29 Earth Radius. The average equilibrium temperature of exoplanets are around 876 K indicating most of them have a high equilibrium temperature. Along with this, there are extreme planets in the dataset with are very huge and high orbital periods.

```
       Serial Number  Number of Stars  Number of Moons  Discovery Year  \
count   36557.000000     36557.000000          36557.0   36557.000000
mean    18279.000000         1.086194              0.0    2015.412643
std     10553.241232         0.311337              0.0       3.901443
min         1.000000         1.000000              0.0    1992.000000
25%      9140.000000         1.000000              0.0    2014.000000
50%     18279.000000         1.000000              0.0    2016.000000
75%     27418.000000         1.000000              0.0    2016.000000
max     36557.000000         4.000000              0.0    2024.000000

       Orbital Period [days]  Planet Radius [Earth Radius]  \
count           3.337900e+04                  24923.000000
mean            1.308128e+04                      5.290766
std             2.201616e+06                     69.262793
min             9.070629e-02                      0.270000
25%             4.360707e+00                      1.550000
50%             1.033808e+01                      2.290000
75%             2.665656e+01                      3.250000
max             4.020000e+08                   4282.980000

       Planet Mass [Earth Mass]  Planet Density [g/cm**3]  \
count               4072.000000                2259.000000
mean                724.754938                   3.800168
std                1579.667577                  42.372213
min                   0.020000                   0.010000
25%                  13.200000                   0.569000
50%                 171.628200                   1.300000
75%                 603.637870                   3.565485
max               25426.400000                2000.000000

       Equilibrium Temperature [K]  Distance [pc]
count                 16625.000000   35723.000000
mean                    876.683068     730.277528
std                     426.049111     805.363940
min                      34.000000       1.301190
25%                     568.000000     288.763000
50%                     795.000000     594.447000
---                     ----------     ----------
```

Fig. 5. Summary Statistics

3) Research Question 1: How has the discovery of exoplanets evolved over the years - ***Uni variate Analysis of Discovery Year*** *:* The distribution of planets discovered over each year is a example of univariate analysis. The below Figure displays, the bar graph plotted between the number of planets discovered and the year they were discovered.

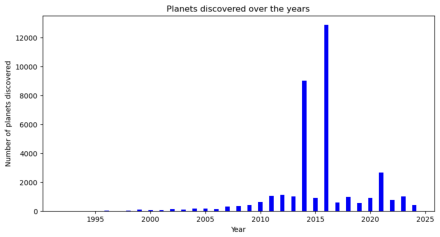

Fig. 6. Distribution of Planets discovery over the year

From the above Figure 6, it is evident that the number of planets discovered over the years has always been increasing where it peaked in the year of 2015 and 2016. This is due to advancements in the latest technology where more advanced methods like Transit have been more commonly used. However, there is a decline in discoveries since 2016 as we could see there is shift towards to those planets which are classified as potential habitable planets.

4) How are the methods used for discovery linked to technical advancements over the period?: The below figure 7 displays the distribution of discovery of planets by each discovery method over the period. From the graph, it is evident that the transit method is most predominantly used in detecting exoplanets especially during 2013 to 2016. However, Methods like Radial Velocity and Micro lensing are being used consistently but in lesser compared to transit method.

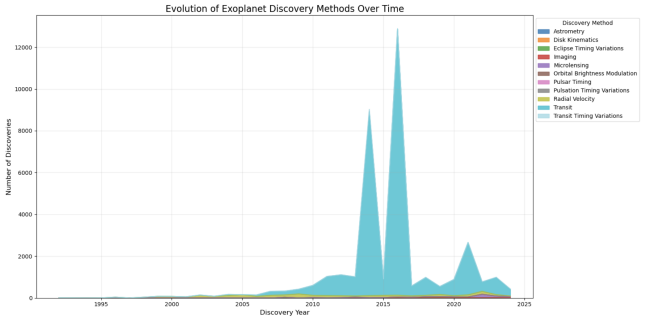

Fig. 7. Distribution of Discovery Methods over the year

The least common methods used in discovery of exoplanets over time are Imaging and Pulsar Timing. Overall the discoveries were less before 2015 and then it attained peak in the year around 2015 mainly due to performance transit methods.

5) Research Question 2: Is there any relation between the distance of exoplanets from Earth and their discovery methods or year?: The below describes the relationship between Average Distance of Exoplanets over the year. It is evident from the below Figure that the average distance of exoplanets discovered during the initial years were around 600 parsecs. After 1994, it the average distance started to see upward trend due to highly advanced methods like microlensing and transits. The year 2022 marked the year where the average distant exoplanet discovered at a distance around 1200 parsecs. The

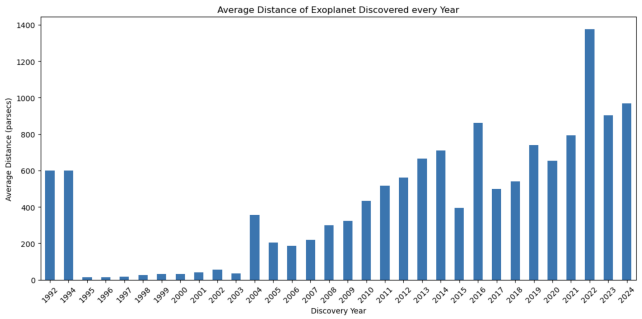

Fig. 8. Distribution of Average Distance of Exoplanets over time

graph continues to go in upward trend even now with the average distance around 800 parsecs.

6) Is there any relation between the distance of exoplanets from Earth and their discovery methods?: The below graph displays the average distance detected by each discovery Method. From the below Figure 9 it is evident that Micro lensing is used for detecting planets with longer distances. Methods like Astrometry , Imaging and Radial velocity helps to identify closer planets. However, Transit method and Pulsation Timing methods can be used to identify planets that are at moderate distance

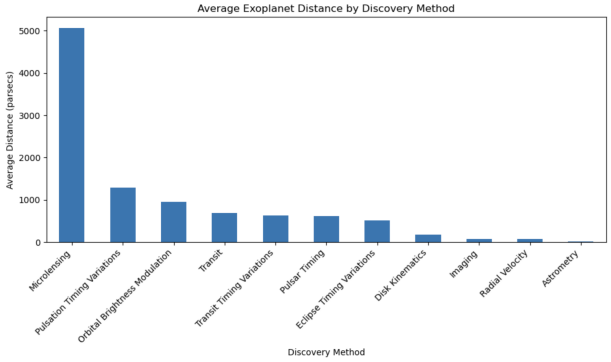

Fig. 9. Distribution of Average Distance of Exoplanets by each Method

C. Exploratory Analysis and Visualization using R

1) Data Cleaning in R: Just like Python , the dataset here is also cleaned and transformed using R before performing any operations. Now, The dataset loaded in the R and librares such dplyr , ggplot2 , maps are installed before to help in visualizations and manipulations. Now the CSV file is read into R and then clean by inserting serial number to uniquely identify row. The Unwanted columns are dropped and duplicate rows are removed as well to get the best set of data which result in optimal analysis and visualization.

2) Summary Statistics in R: Using R, a summary function is used which provides summary statistics for all the 36,557 exoplanets. From the summary it is evident that the statistics work similar using Python as well. The median year of

```
> summary(exoplanetsData)
 Serial.number   Planet.Name         Number.of.Stars Number.of.Moons Discovery.Method
 Min.   :    1   Length:36557       Min.   :1.000   Min.   :0       Length:36557
 1st Qu.: 9140   Class :character    1st Qu.:1.000   1st Qu.:0       Class :character
 Median :18279   Mode  :character    Median :1.000   Median :0       Mode  :character
 Mean   :18279                      Mean   :1.086   Mean   :0
 3rd Qu.:27418                      3rd Qu.:1.000   3rd Qu.:0
 Max.   :36557                      Max.   :4.000   Max.   :0

 Discovery.Year Discovery.Locale   Discovery.Facility Discovery.Instrument
 Min.   :1992   Length:36557       Length:36557       Length:36557
 1st Qu.:2014   Class :character   Class :character   Class :character
 Median :2016   Mode  :character   Mode  :character   Mode  :character
 Mean   :2015
 3rd Qu.:2016
 Max.   :2024

 Orbital.Period..days. Planet.Radius..Earth.Radius. Planet.Mass..Earth.Mass.
 Min.   :        0     Min.   :  0.270              Min.   :    0.02
 1st Qu.:        4     1st Qu.:  1.550              1st Qu.:   13.20
 Median :       10     Median :  2.290              Median :  171.63
 Mean   :    13881     Mean   :  5.291              Mean   :  724.75
 3rd Qu.:       27     3rd Qu.:  3.250              3rd Qu.:  685.64
 Max.   :402800000     Max.   :4282.980             Max.   :25426.40
 NA's   :3178          NA's   :11634                NA's   :32485
 Planet.Density..g.cm.3. Equilibrium.Temperature..K. Distance..pc.
 Min.   :  0.01          Min.   :  34.0              Min.   :   1.301
 1st Qu.:  0.57          1st Qu.: 568.0              1st Qu.: 288.763
 Median :  1.30          Median : 795.0              Median : 594.447
 Mean   :  3.80          Mean   : 876.7              Mean   : 730.278
 3rd Qu.:  3.57          3rd Qu.:1097.0              3rd Qu.: 942.173
 Max.   :2800.00         Max.   :4050.0              Max.   :8800.000
 NA's   :34298           NA's   :19932               NA's   :834
```

Fig. 10. Summary Statistics in R

discoveries are around 2016. The maximum value of orbital period is 4,020,000 days which is around 11000 years. The average mass of planet is 171.63 Earth masses. The closest planet discovered is 1.3 parsecs and the longest is around 8800 parsecs.

3) Research Question 3: How does the mass of an exoplanet influence its density and radius across various planets?: The below both Figures displays the relationship between the mass of an exoplanet with density and Radius. In the above Mass Vs

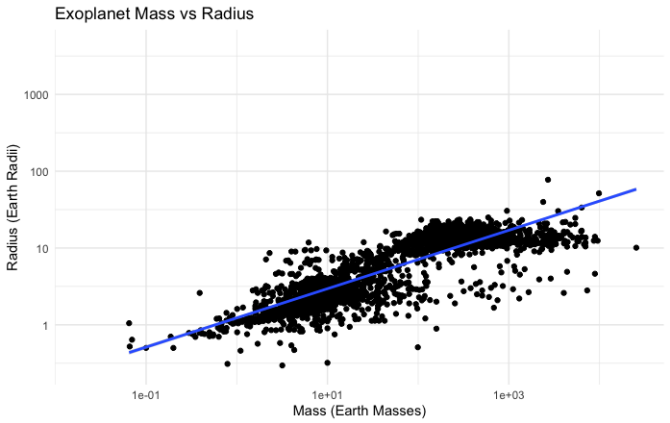

Fig. 11. Mass vs Radius

Radius graph, where the relationship in scatter plot is shown between Planet Mass and Radius in a log scale revealing a positive correlation . The data points show a lot of variability which means the planets composes of all kinds such as gas , rocky etc. The trend line in between confirms the relationship between Mass and Radius. Hence, Large Planets tend to have larger Radii. In the above scatter plot, the relationship is explored between Planet Mass and Density . We could see a negative correlation between Mass and Density which means as Mass increases, Density decreases. This maybe due to higher gas planets indicate the higher mass which makes it less dense. whereas most of planets discovered with lower mass are rocky planets which has higher density.

4) Find planets that are having equilibrium temperature in habitable range.: The below histogram displays the distributions of equilibrium temperature in Kelvin for exoplanets. From the below graph , it is evident that Most of the planets

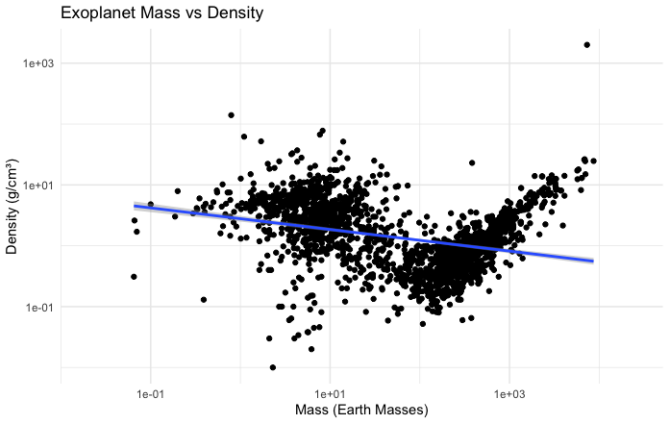

Fig. 12. Mass vs Density

have temperature below 1000k where the peak graph attains at 500 to 800K. This indicates that most of exoplanets are cooler than Earth like temperature.

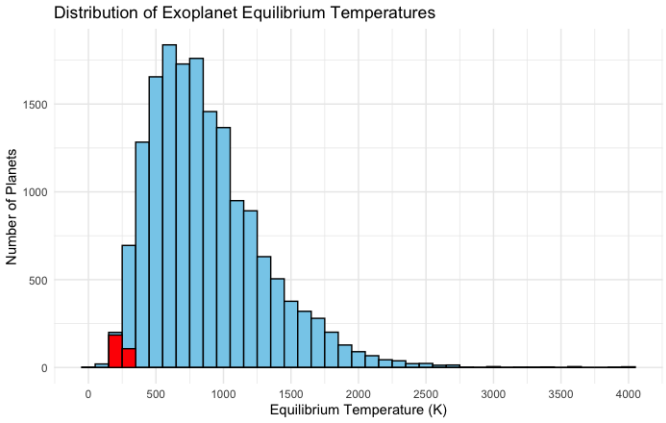

Fig. 13. Equilibrium Temperature Distribution

The Red color portion highlighted in the above graph indicates the habitable temperature which falls between 175K to 270K. This is the temperature where liquid water could potentially exist. However, beyond 1000k, the graph declines reflecting there are very less number of high temperature planets. Overall, the data shows a diverse range in planetary temperatures, with a more focus on cooler planets. The above

```
> head(planetList, 10)
 [1] "Kepler-1981 b" "Kepler-511 b"  "Kepler-62 e"   "Kepler-1544 b" "Kepler-315 c"
 [6] "TOI-4633 c"    "Kepler-1552 b" "Kepler-1652 b" "Kepler-1746 b" "Kepler-1097 b"
```

Fig. 14. List of top 10 planets that are in habitable range

figure shows the list of top ten planets fall in habitable range between 175k and 270k. These planets are potential candidate for liquid water to exist out of which Kepler-62 e and Kepler-1652 b are famous ones . Although the temperature does not guarantee any habitability , while this provides a significant list of planets that can be further explored and helps scientist to narrows their research in such planets.

D. Integration of AWS Cloud service with SQL using RDS

This Analysis uses AWS services to explore and analyze the dataset focusing on data storage and integrating with MySQL solutions. Initially S3 bucket is created with name ait580exoplanetprojectdataset. The chosen dataset is now uploaded to Amazon S3 bucket, ensuring access to cloud storage. After uploading, the S3 URI and Object URL are generated which can be used to access from other AWS services or systems.

Fig. 15. AWS EC2 Instance Setup

Now, In the local EC2 instance needs to be setup. This is done using EC2 dashboard in the AWS service where EC2 can be launched and connected with the local terminal. The below figure displays the terminal after a connection has been established.

After Setting up the EC2 instance, A RDS database needs to be created to integrate MySQL solutions on the dataset which is in s3 bucket. After RDS database is created, the MySQL client is installed locally using the endpoint and port provided in RDS dashboard. After MySQL is setup, A new database and table needs to be created similar to that of data exploration in SQL performed earlier. Now, SQL queries can be executed as per research questions.

1) Count of Planets Discovered Each Year:: The below SQL query displays the count of planets that are discovered over time period. It is evident that 2020 had most number of planets discovered which is followed by 2024.

Fig. 16. Displaying Count of planets over time period

2) Count of Planets Discovered in each Method:: The below SQL query displays the count of planets that are discovered as per each method. It is seen that transit is the most used discovery method with 32942 followed by radial velocity of 2571 counts. The least used discovery method is Disk Kinematics.

3) Average Distance of planets discovered in each discovery year:: The below SQL query displays the average distance of planets discovered across each year. In the year 1992, 1994 the average distance is around 600 parsecs.

Fig. 17. Displaying Count of planets across each method

Fig. 18. Average Distance of planets discovered in each discovery year:

4) Average Distance of planets discovered in each discovery method:: The below SQL query displays the average distance of planets discovered across each method. The maximum average distance is from Micro lensing , Pulsation timing and Orbital Brightness. The Astromety is usally used for detecting planets are close by.

Fig. 19. Average Distance of planets discovered in each discovery method:

5) Get list of first 10 planets that fall in the habitable zone of equilibrium temperature: The below SQL results show that the top 10 planets that are having a habitable zone of equilibrium temperature. The Kepler 159 C, Kepler 1545 b are such planets that fall in this range.

Fig. 20. ExoPlanets in habitable zone of equilibrium temperature

E. UNI-VARIATE ANALYSIS USING AWS GLUE DATA BREW

The dataset which was uploaded to S3 bucket is used for analysis in a AWS tool called AWS Glue Data Brew. Once the

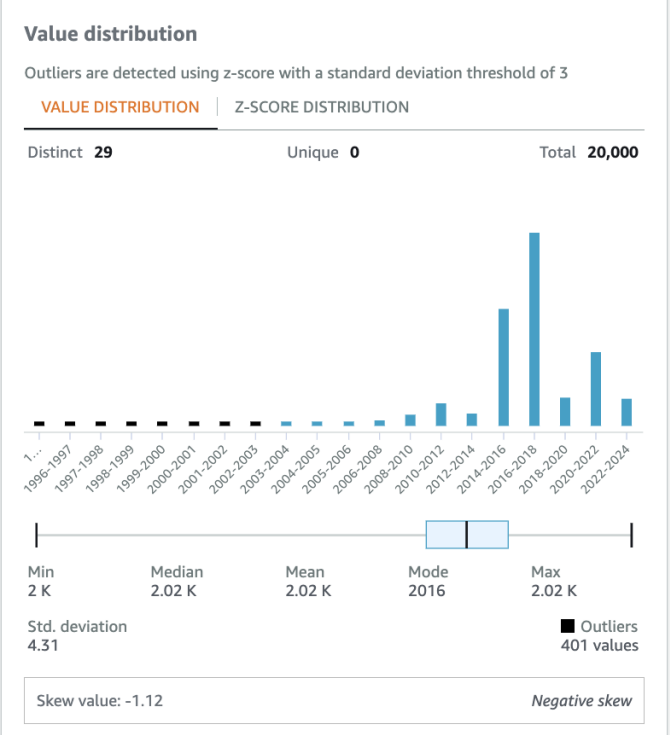

Fig. 21. AWS GLUE DATA BREW DISCOVERY YEAR DISTRIBUTION

data brew project is created, a profile job was run with recipe which provides analysis for each columns in the dataset.

The below figure displays the snippet of advanced summaries , which indicates the count of 29 rows, the median of 2.02 K and third quartile of 2.02K. The discovery year column consists 29 distinct values ranging from 1992 to 2024. The skewness is in negative which indicates the distribution of discovery year is longer on the end of the tail. The variance is around 18.55 and low standard deviation is closely distributed around the median year 2016.

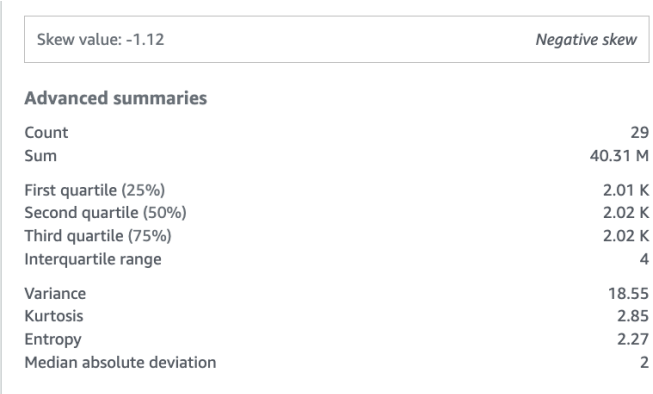

Fig. 22. AWS DATABREW Summary

VIII. LIMITATIONS

The above data exploration and analysis of exoplanets dataset using Python, R , SQL and AWS has been very insightful in learning about the scale of discoveries, the most common and least methods in detecting planets and also finding planets that fall in habitable zone of equilibrium temperature.

Although the dataset is extracted from NASA's Exoplanet Archive [4] , the dataset does contain missing values of some crucial attributes. While dataset contains planets are from long distances, there can be a slight imbalance in dataset regarding number of moons. Another limitation is that, the static nature of dataset might vary with the actual characteristics of planets. The potential habitable planets are solely based on Equilibrium temperature, However this helps to narrow down further research , a lot more data is required to get a list of truly potential planets. The research currently relies on Statistical and visualization techniques using Python , R and SQL. While this works well for exploratory analysis, More advanced Machine learning techniques could provide more wide range of capable analysis. The dependency of public data also limits the comprehensiveness of the analysis. Addressing these limitation in this research will lead to more robust findings.

IX. CONCLUSIONS

This research explores various discovered exoplanets using tools like Python , R , SQL and AWS to analyze a dataset of 36000 entries. This analysis highlights how various technological advancements have helped in discovering exoplanets. It is evident that transit method is most widely used discovery method over the period of time. While Micro lensing Method is predominantly used of discovering planets that are at longer distance. The Research also provides relationship between Exoplanet's Mass over Density and Radius. One of the most important point noted is that , list of planets that fall in equilibrium range is being extracted using the analysis which can be potential candidate for sustaining life in the future.

However, there are limitations to study such as reliance on public and static data. The study of limitations are equally important for future explorations. By Incorporating Advanced Machine learning Techniques, it would help in deep exploration of exoplanets in the future. In the end, This project primarily focuses on data analysis in astronomy , analyzing and exploring various trends in the planets outside our system. This analysis helps to find planets that are similar to earth like exoplanets with respect to equilibrium temperature.

X. REFERENCES

REFERENCES

[1] H. Pai, S. Dornala, A. Nathoo, S. Mayya, W. Regan, O. Upadhyay, S. K. Rajan, A. Diwan, P. Soni, Aspiring Scholars Directed Research Program, and R. Downing, "The Exploration of Habitable Exoplanets using Data Mining Algorithms and Data Manipulation," [Online]. Available: https://www.academia.edu/49090548/The_Exploration_of_Habitable_Exoplanets_using_Data_Mining_Algorithms_and_Data_Manipulation. [Accessed: Dec. 7, 2024].

[2] J. T. Wright and B. S. Gaudi, "Exoplanet Detection Methods," *arXiv*, vol. 1210.2471, Oct. 2012. [Online]. Available: http://arxiv.org/abs/1210.2471v2. [Accessed: Dec. 7, 2024].

[3] Y. Jin, L. Yang, and C. Chiang, "Identifying Exoplanets with Machine Learning Methods: A Preliminary Study," *International Journal on Cybernetics & Informatics*, vol. 11, no. 2, pp. 31–42, 2022, doi: 10.5121/ijci.2022.110203. [Accessed: Dec. 7, 2024].

[4] NASA Exoplanet Archive, "NASA Exoplanet Archive," NASA Exoplanet Science Institute, California Institute of Technology. [Online]. Available: https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS. [Accessed: Oct. 20, 2024].

[5] L. Kaltenegger and D. Sasselov, "Exploring The Habitable Zone for Kepler Planetary Candidates," *The Astrophysical Journal*, vol. 736, no. 2, 2011, doi: 10.1088/2041-8205/736/2/l25.