NASA Exoplanet Archive Exploratory Data Analysis and Conclusions

Ayaan Haque, Saratoga High School

July 2020

1 Introduction

The Exoplanet data is a massive dataset that analyzes exoplanets and lots of different qualities and attributes of these discovered exoplanets. I am not well versed in Astronomy and interplanetary systems, so this dataset originally was very overwhelming, and I even had to research what an exoplanet was to understand this dataset. Most of my machine learning experience is in Computer Vision and deep learning, with a solid understanding of Natural Language Processing. I have very little experience with Exploratory Data Analysis, as I just recently began using Jupyter Notebooks. However, I think this dataset because of how much information it has can be of great use, and is a perfect case study for data science and analysis, but not exactly sure about machine learning. Since there is no need to make new predictions based on this data from what I understand, machine learning doesn't seem like the appropriate application. I think using Exploratory Data Analysis would be the perfect approach to make conclusions of underlying trends in the data. However, I still think this can be used for astronomers who want to analyze exoplanets and make comparisons between different ones.

Figure 1: Universe with Exoplanets

2 Advantages

I think one of the main advantages of this dataset is the sheer size of it. Since there are almost four thousand exoplanets recorded, there are lots of conclusions that can be made. Additionally, there are hundreds of attributes for each planet, meaning that there are lots to learn about specific planets and specific types of planets. Since there are so many different attributes, there can be lots of different correlations to investigate, especially with thousands of permutations of variables. I think grouping the data by planet type and making analysis based on that would be a good first approach. Additionally, I think it would be very important to learn and understand the key(Keck Exoplanet db Columns.pdf) first so that when we go into actually plotting data and finding correlations, it would be much easier. I think another benefit is that the data contains a lot of strings and numbers, which makes it much more comprehensive. A very obvious but strong advantage of the dataset is that it is properly organized and is in a .csv format, meaning that exploring the data using code is very simple. There is no need for me to go through and clean out the data because it already has been cleaned and filtered by previous researchers.

```
In [37]: 1 df.head(10)
```

Out[37]:

	rowid	pl_hostname	pl_letter	pl_name	pl_discmethod	pl_pnum	pl_orbper	pl_orbpererr1	pl_orbpererr2	pl_orbperlim	...	st_bmy	st_bmyerr	st_bmylim
0	1	11 Com	b	11 Com b	Radial Velocity	1	326.03000	0.3200	-0.3200	0.0	...	NaN	NaN	NaN
1	2	11 UMi	b	11 UMi b	Radial Velocity	1	516.21997	3.2000	-3.2000	0.0	...	NaN	NaN	NaN
2	3	14 And	b	14 And b	Radial Velocity	1	185.84000	0.2300	-0.2300	0.0	...	NaN	NaN	NaN
3	4	14 Her	b	14 Her b	Radial Velocity	1	1773.40002	2.5000	-2.5000	0.0	...	0.537	0.001	0.0
4	5	16 Cyg B	b	16 Cyg B b	Radial Velocity	1	798.50000	1.0000	-1.0000	0.0	...	0.418	0.003	0.0
5	6	18 Del	b	18 Del b	Radial Velocity	1	993.30000	3.2000	-3.2000	0.0	...	NaN	NaN	NaN
6	7	1RXS J160929.1-210524	b	1RXS J160929.1-210524 b	Imaging	1	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
7	8	24 Boo	b	24 Boo b	Radial Velocity	1	30.35060	0.0078	-0.0077	0.0	...	NaN	NaN	NaN
8	9	24 Sex	b	24 Sex b	Radial Velocity	2	452.80000	2.1000	-4.5000	0.0	...	0.572	0.003	0.0
9	10	24 Sex	c	24 Sex c	Radial Velocity	2	883.00000	32.4000	-13.8000	0.0	...	0.572	0.003	0.0

10 rows × 355 columns

Figure 2: Overview of Dataset

3 Disadvantages

Just from a brief glance at the data, there are exactly 3838 exoplanets recorded, and a total of 355 attributes that the planet may have. The discovery of these planets ranges back to the 1990s, meaning that the dataset has been expanding for decades. A brief glance also exposes what I would instantly think to be the biggest issue with the dataset. There is a sea of white in the excel sheet, and these missing attributes could be very problematic in terms of properly understanding this data. This undoubtedly is the biggest shortcoming of the dataset.

While the dataset in incredibly massive, I think in some ways that can also be a downside of trying to explore the dataset. Since it is so large, making

```
In [6]:  1 df.isnull().sum()

Out[6]: rowid                0
        pl_hostname          0
        pl_letter            0
        pl_name              0
        pl_discmethod        0
                          ...
        st_m1lim          3518
        st_c1             3518
        st_c1err          3519
        st_c1lim          3518
        st_colorn            0
        Length: 355, dtype: int64
```

Figure 3: Total Null Values in Dataset

appropriate graphs and visualizations would be challenging, and making proper inferences without real automated algorithms would be difficult. It would also be challenging to understand which of the attributes are important for what a researcher might be looking for. A minor downside of the dataset is that there aren't units or very descriptive labels for each attribute for each planet. However, since there is a PDF that is a legend/key for the dataset, it solves most of that issue, however, if there was a way to incorporate labels and units inside the dataset, that would be ideal. I think having a strong understanding of exoplanets and their details would be important to truly and properly understand this dataset.

4 Methods and Results

4.1 Approach

In terms of dealing with the shortcomings of the dataset, I think there are a few approaches. First, for the attributes which the majority of planets have empty, those can easily be ignored for making predictions and analyzing this dataset. I think it is somewhat unreasonable and unnecessary to understand all the attributes of each exoplanet. I think there are a good amount that are very important, obviously such as the first few attributes like name and other key statistics. However, using the data that is already provided, it could be possible to develop a model that could fill in the missing attributes, through a semi-supervised approach per say. I think creating a model that is trained

```
1 df.pl_discmethod.unique()
```

```
array(['Radial Velocity', 'Imaging', 'Eclipse Timing Variations',
       'Transit', 'Astrometry', 'Orbital Brightness Modulation',
       'Pulsation Timing Variations', 'Microlensing',
       'Transit Timing Variations', 'Pulsar Timing'], dtype=object)
```

```
1 df.groupby('pl_discmethod').size()
```

```
pl_discmethod
Astrometry                          1
Eclipse Timing Variations           9
Imaging                            44
Microlensing                       70
Orbital Brightness Modulation       6
Pulsar Timing                       6
Pulsation Timing Variations         2
Radial Velocity                   687
Transit                          2998
Transit Timing Variations          15
dtype: int64
```

Figure 4: Missing Values Per Column

on the dataset and can then make predictions for other datasets by inputting a baseline amount of parameters would be the best way to adjust for the missing values. However, creating such a model would be very hard, and I am not sure what I could even code the model with. However, if it was possible to create such a model, I think it would be an ideal approach. I wasn't able to develop this model when I was exploring the data, but if such a model hasn't been considered, it should for real researchers working with this dataset.

In terms of using these exceptions within analyzing the data, I think modeling with both the exceptions included and excluded will be important. Using the exceptions would allow us to understand what specific attributes are the hardest to find and figure out how the missing data affects the analysis. I think not having these attributes and values is a key factor for the dataset, and I think when visualization is done, it will be easy to tell what is missing and how it changes the results. Also creating visualizations that are exclusive of the exceptions would also be important because it would allow me to properly understand what can be determined based on what we know about these exoplanets. There would be no speculation about what these values possibly could be. While this might lead to less understanding of the data and exoplanets as a whole, it would be the most accurate and proper approach because it doesn't speculate or predict anything. However, I think being able to predict and fill

in many of these missing attributes could unlock lots of new discoveries and inferences that could be groundbreaking.

4.2 Methods

My first thought about analyzing this data is to use Pandas and Jupyter Notebooks. This is because Jupyter Notebooks are great data analyzing tools because of how easy it is to visualize data and make graphs and charts, which allows for better conclusion making. Pandas is a great tool to read .csv files, which is what the exoplanet dataset is in. I would use excel, but I think that using code and notebooks is a much better approach because of how easy it will be to actually visualize data. Since I don't have excel, I would have to use Google Sheets, which I don't think has the same amount of valuable tools like the Jupyter Notebooks. I also think that using Jupyter Notebooks is effective because of markdown, and markdown will easily allow me to jot down notes and trends that I notice myself, and it allows for a better documentation process. I think using heat maps and data clustering would be a good way to categorize the data and find underlying trends. There is also lots of available documentation for all the different functions in Pandas, making it a very useful tool. Loading data and making adjustments to it would be incredibly easy, and using packages such as Matplotlib will also help me create appropriate visualizations that will help me make conclusions about the dataset.

4.3 Results and Findings

Once I started using Pandas in Jupyter Notebooks, I began by investigating the missing values. I first found that most if not all planets had at least one null value, proving the idea that missing values are a big problem in this dataset. While it is obviously hard to collect this data, if it was possible to fill in many of these missing attributes, that would be ideal. The majority of the missing attributes lay in finding the size attributes of the planets, such as radius and density. The stellar columns have the most amount of missing data points. The first few attributes are completely full for each exoplanet as expected, such as the star hostname, planet letter, and the name. There are 8 different planet types(letters), from b-i, with by far the most planets being letter "b". To understand the sizes and the distribution of sizes of these exoplanets, I used pandas to create a scatter plot of each planet's mass vs the radius. I found that most were on the smaller side and clustered in the bottom corner of the plot. There were a few outliers, but the outliers were accompanied by other large planets.

I also wanted to see if a planet's distance was correlated to the size of the planet, but I wasn't able to see a correlation. All the data just clustered at the bottom of the graph, being relatively close to the sun and smaller in size. There were outliers of course, but not many. Then I learned about the .corr() function that is built into Pandas, and I used that to find the correlations between different variables. I plotted the entire dataset in one correlation matrix, and

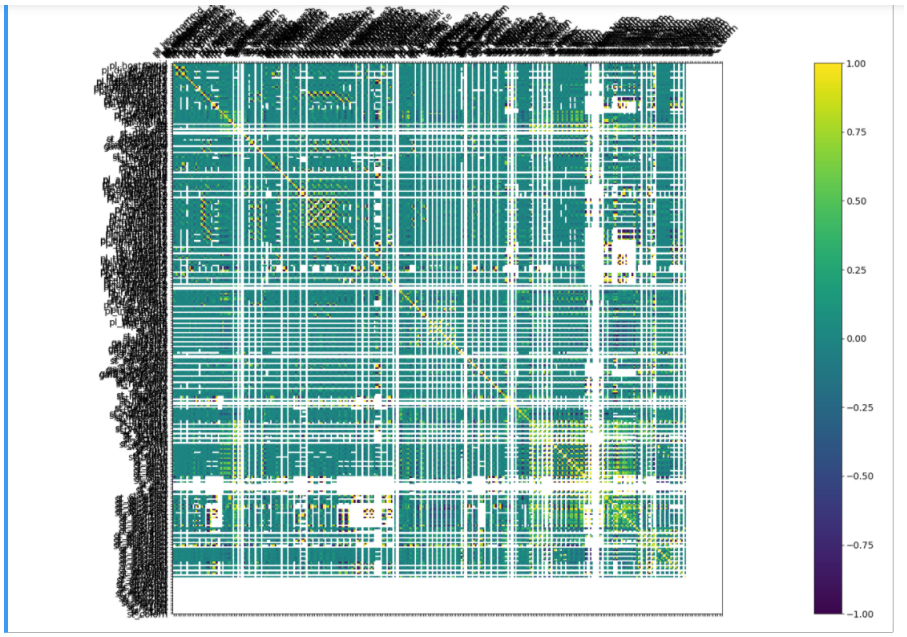

Figure 5: Correlation Matrix of all Attributes

I found that there were a few correlations between certain variables. However, overall there wasn't a strong correlation between a majority of the variables, making it harder for me to make confident conclusions.

5 Conclusion

Lastly, the inputs and outputs of this dataset and project are simple in my opinion. The inputs would simply be the data in its csv format. I would obviously have to clean out the dataset and deal with missing values and make sure that all the data is in a readable format. Next, the transformations done to the data would be through creating graphs and charts and other visualizations. Through Pandas, I can group, sort, rearrange, and add and remove data from the dataset to change how I can understand and visualize the dataset. The outputs would be lots of diverse and in-depth visualizations that can be used to determine correlations and other conclusions between the data and variables. There are many different things to learn from this dataset, and the investigations are endless. Some things to learn are which planets have the possibility of hosting life based on how far they are from other planets and stars. Another possibility is to discover which exoplanets are most similar to one another or most similar to the ones in our solar system, and find out how they compare. This could allow for comparisons of entire solar systems, not just planets. Overall, this was a fun experience and I am glad that I was able to explore a dataset in a field I

```
In [13]: 1 df.plot(x ='pl_bmassj', y='pl_radj', kind = 'scatter')
Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x11f853850>
```

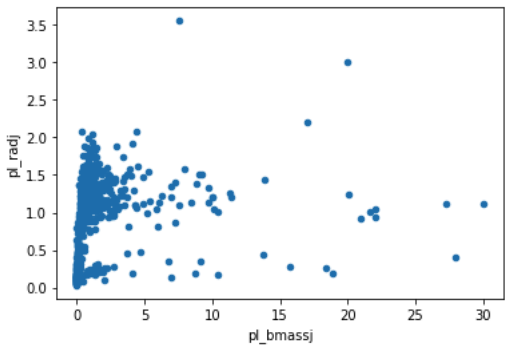

Figure 6: Plot of Mass vs Radius of Exoplanets

```
In [18]: 1 df.plot(x ='pl_radj', y='st_dist', kind = 'scatter')
Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x11e352350>
```

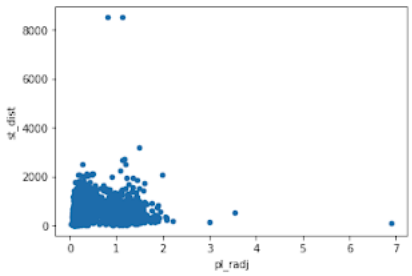

Figure 7: Plot of Radius vs Distance of Exoplanet

have never had experience in before and learning better how to visualize data through notebooks and other libraries was a great experience.