# Exoplanets Dataset Analysis Report

Hager Othman

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

# EXECUTIVE SUMMARY

While Earth remains our sole known oasis in the vastness of space, the pursuit of habitable planets fuels our scientific curiosity and shapes our understanding of the cosmos …

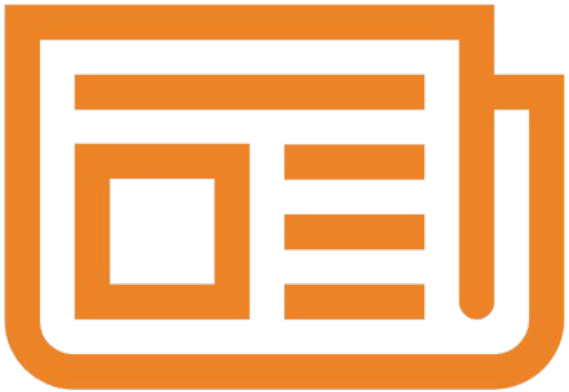
Each data point in our analysis is a clue, guiding us toward answering the age-old question: Are we alone in the universe?

# Introduction

**Dataset Overview**

The dataset comprises information on 1013 celestial bodies , with various features characterising both the planets and their host stars. The key variables such as planet mass , star radius, and star luminosity, orbital period, orbital radius, star mass. while the size is responsible for analysis , the data sounds Imbalanced and the structure poses challenges.

# METHODOLOGY



The methodology involves descriptive statistics, visualization techniques, and correlation analyses.

The examination includes the visualization of distribution, exploration of habitable zones, correlation matrix computation, and outlier detection.

# RESULTS

```
[ ] data.describe()
```

| | planet_mass | planet_radius | orbital_period | orbital_radius | star_mass | star_radius | star_luminosity |
|---|---|---|---|---|---|---|---|
| count | 1013.000000 | 1013.000000 | 1013.000000 | 1013.000000 | 1013.000000 | 1013.000000 | 1013.000000 |
| mean | 343.563980 | 9.033790 | 18.851294 | 0.101647 | 0.992595 | 1.125280 | 3.206746 |
| std | 811.766416 | 5.612988 | 65.343955 | 0.162509 | 0.265571 | 0.497601 | 1.113675 |
| min | 0.889840 | 0.884800 | 0.719573 | 0.015260 | 0.330000 | 0.328000 | 1.020237 |
| 25% | 10.900540 | 2.817920 | 3.118601 | 0.041010 | 0.830000 | 0.794000 | 2.434663 |
| 50% | 133.476000 | 10.673600 | 4.542169 | 0.054900 | 1.000000 | 1.038000 | 3.145093 |
| 75% | 349.580000 | 13.596800 | 11.024540 | 0.096000 | 1.180000 | 1.390000 | 3.902029 |
| max | 11440.800000 | 20.888000 | 1071.232280 | 1.890000 | 1.720000 | 4.230000 | 6.365151 |

```
[ ] sns.heatmap(data, vmin=None, vmax=None, cmap=None, center=None, robust=False, annot=None,  )
```

```
data['is_habitable'] = data.apply(lambda row: 0.8 * row['star_luminosity'] <= row['orbital_radius'] <= 2 * row['star_luminosity'], axis=1)
habitable_planets = data[data['is_habitable']]
```

The provided code define a habitat zone based on the relationship between orbital radius, star luminosity and a predefined criterion.

The number of habitable planets is then displayed.

When dealing with imbalanced data, we can either try to oversampling or undersampling as a way to resampling

1) Perfect Correlation (1.00):
the star_mass and star_luminosity have a perfect positive correlation.

2) High Positive Correlation (0.95 - 0.98):
orbital_period and orbital_radius have a very high positive correlation.

3) Moderate Positive Correlation (0.57 - 0.58):
planet_radiusand orbital_radius have a moderate positive correlation, planet_radius and star_radius have same

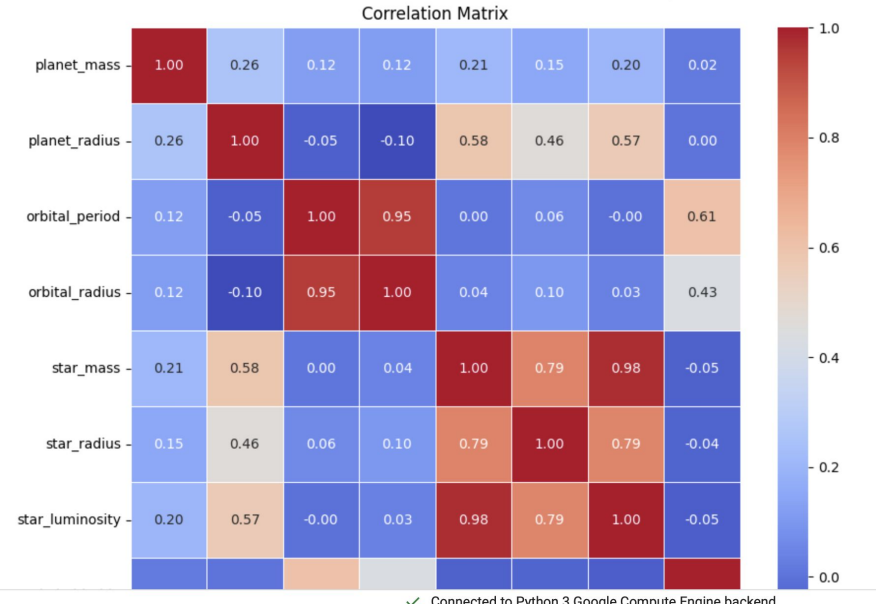
6. Negative Correlation (-0.26):
planet_mass and planet_radius have a negative correlation, planet_mass andis_habitable

4) Weak Correlation (-0.05):
star_mass and is_habitable have a very weak negative correlation.

These observations provide insights into the relationships between different variables in the data. See next slide

# Visualising the Distribution :

```
# Visualize the correlation matrix
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)
plt.title('Correlation Matrix')
plt.show()
```




ın the code provided there is some use for libraries to visualise the distribution of planet and star variables.
As you can see star_mass and stat_luminosty have positive high correlations

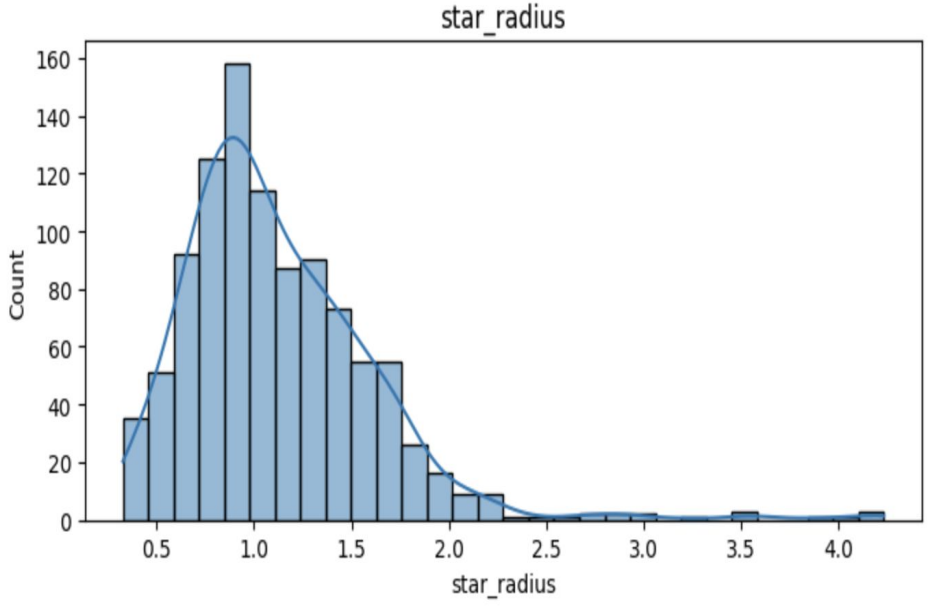
It’s worth mentioning that plotting during the 2 hours exam  posed a challenge due to the restriction on the use of the functions and the substantial volume of data , making it challenging ti generate optimal plots.

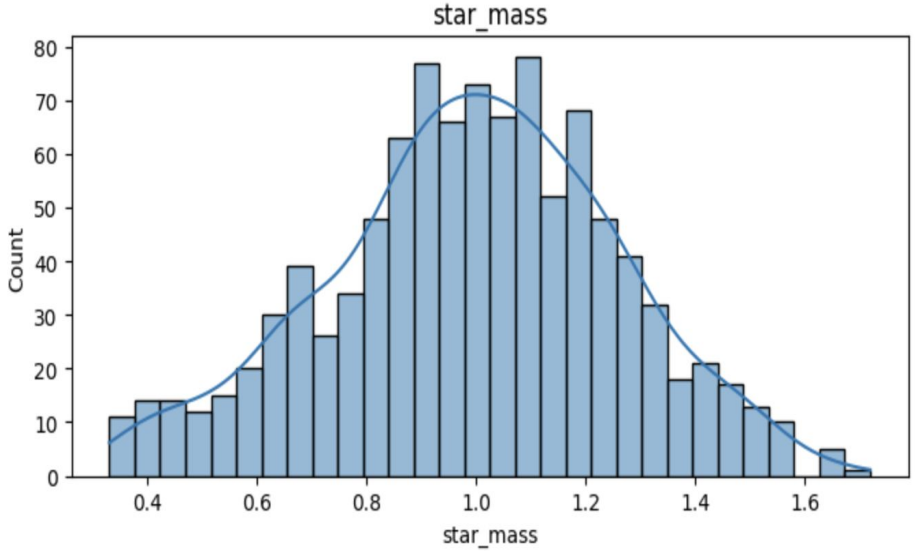
# DashBoard

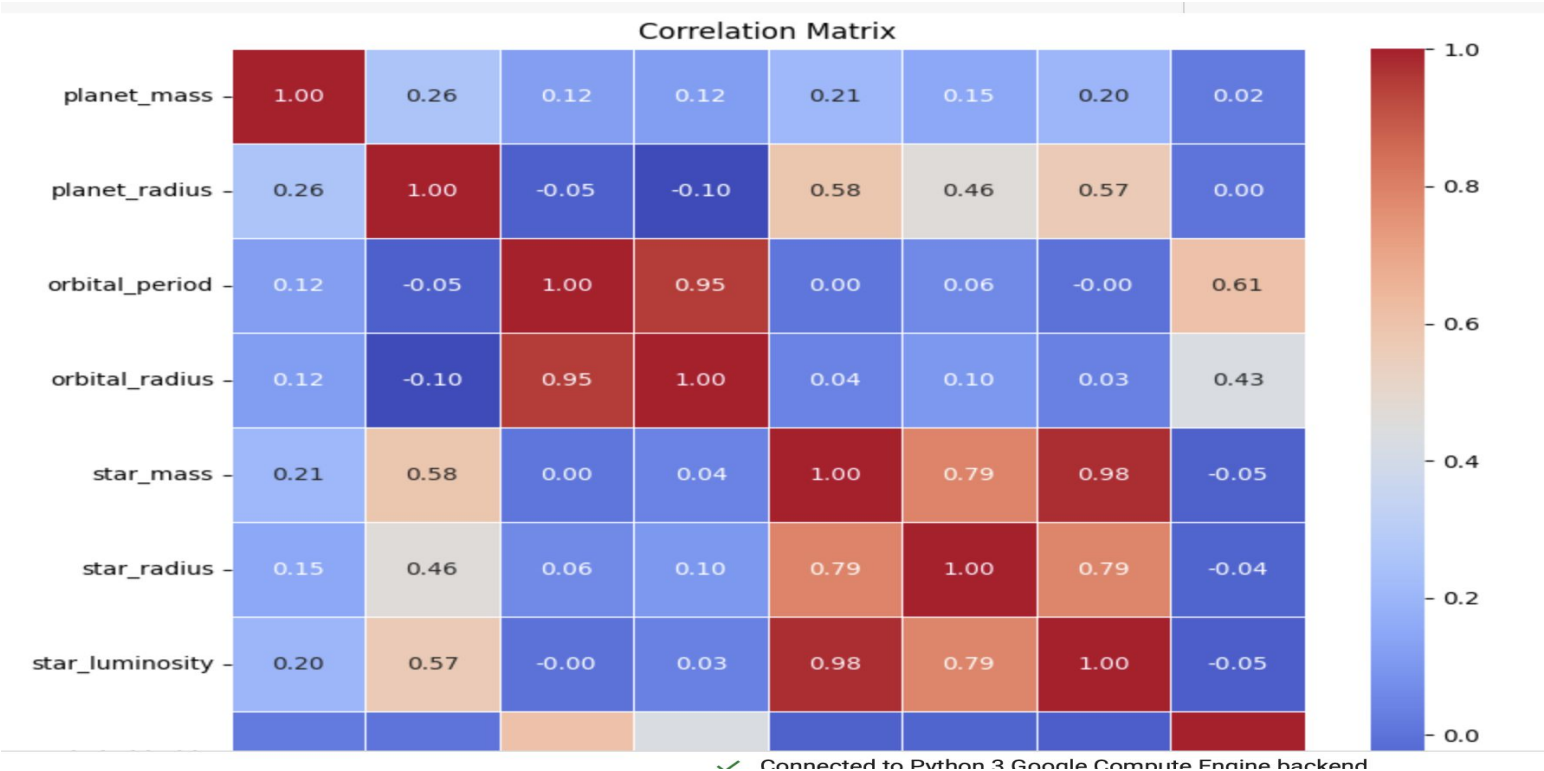
https://colab.research.google.com/drive/11v2XAxrd-uDYd1F3XMjPnct64ibHQZ3

## Habitable Zone Exploration :

The provided code define a habitat zone based on the relationship between orbital radius, star luminosity and a predefined criterion. The number if habitable planets is then displayed.



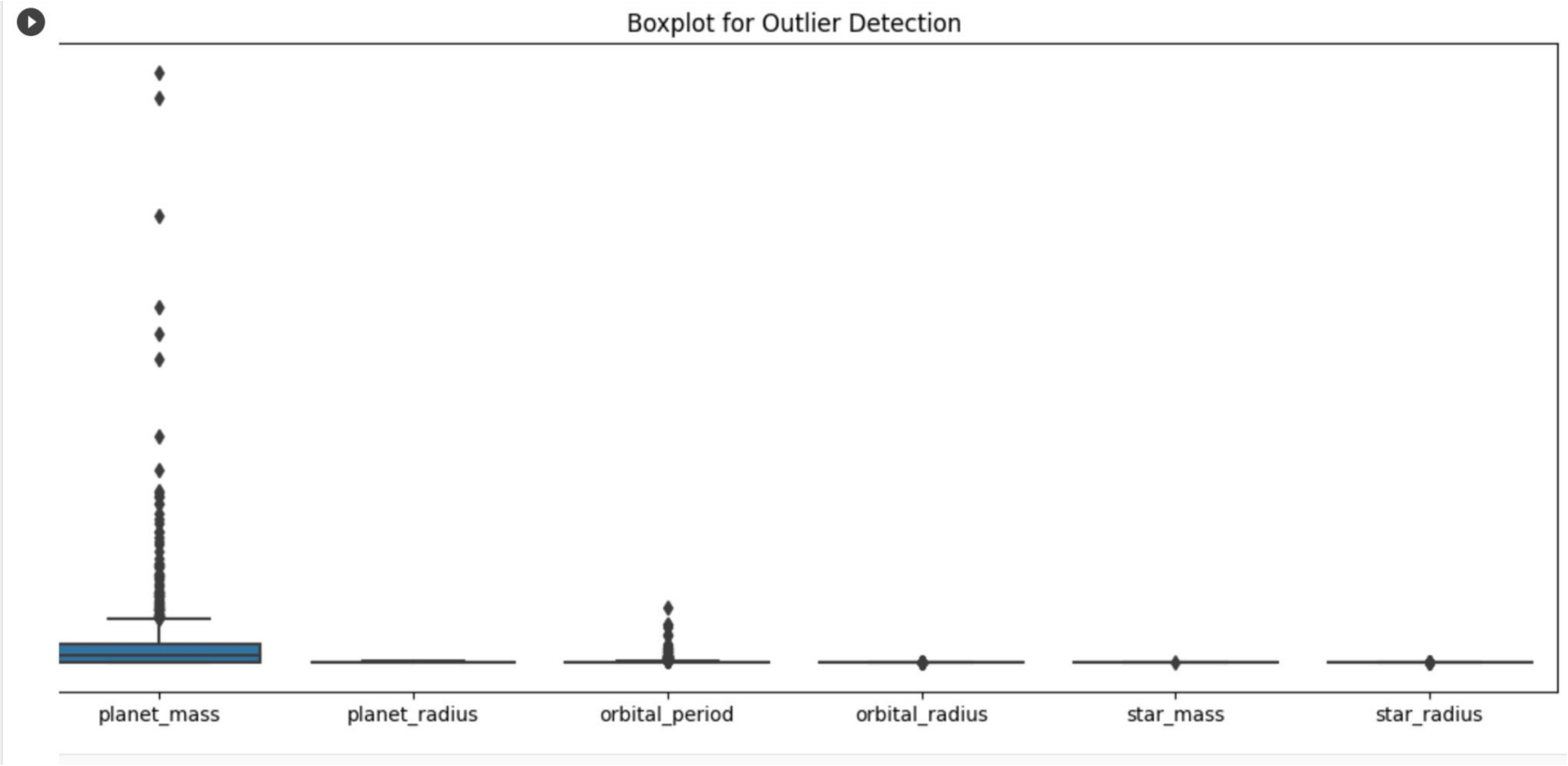






**Correlation Analysis :**

The graph generates a correlation matrix and visualise it using a heat map

This help to understand the relationships  i explained before between different variables in the dataset.



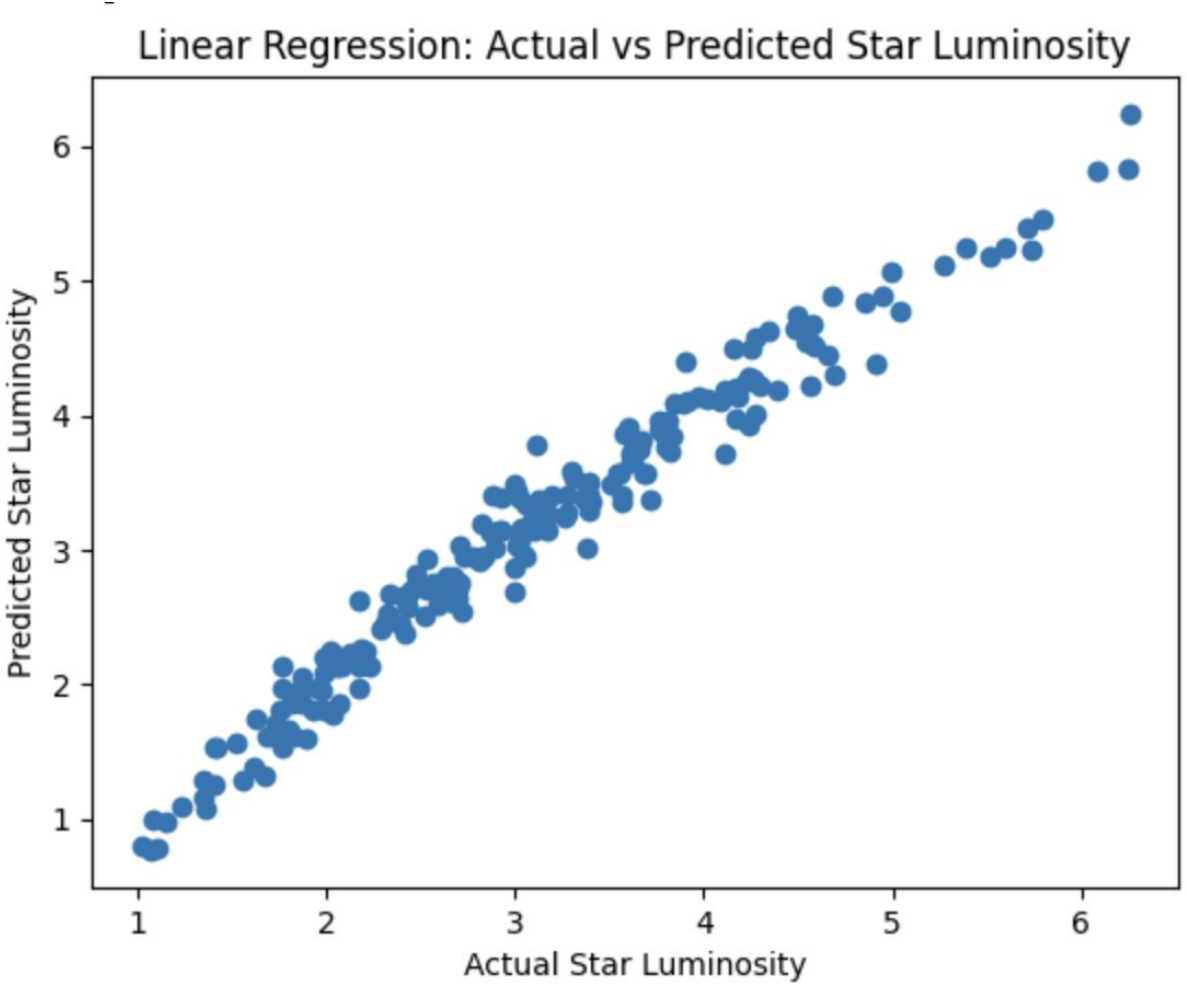

**Outlier Detection :**
Box plots are employed to detect outliers in the dataset. Outliers can be crucial in refining the dataset for more robust analysis.

# Descriptive Statistics :

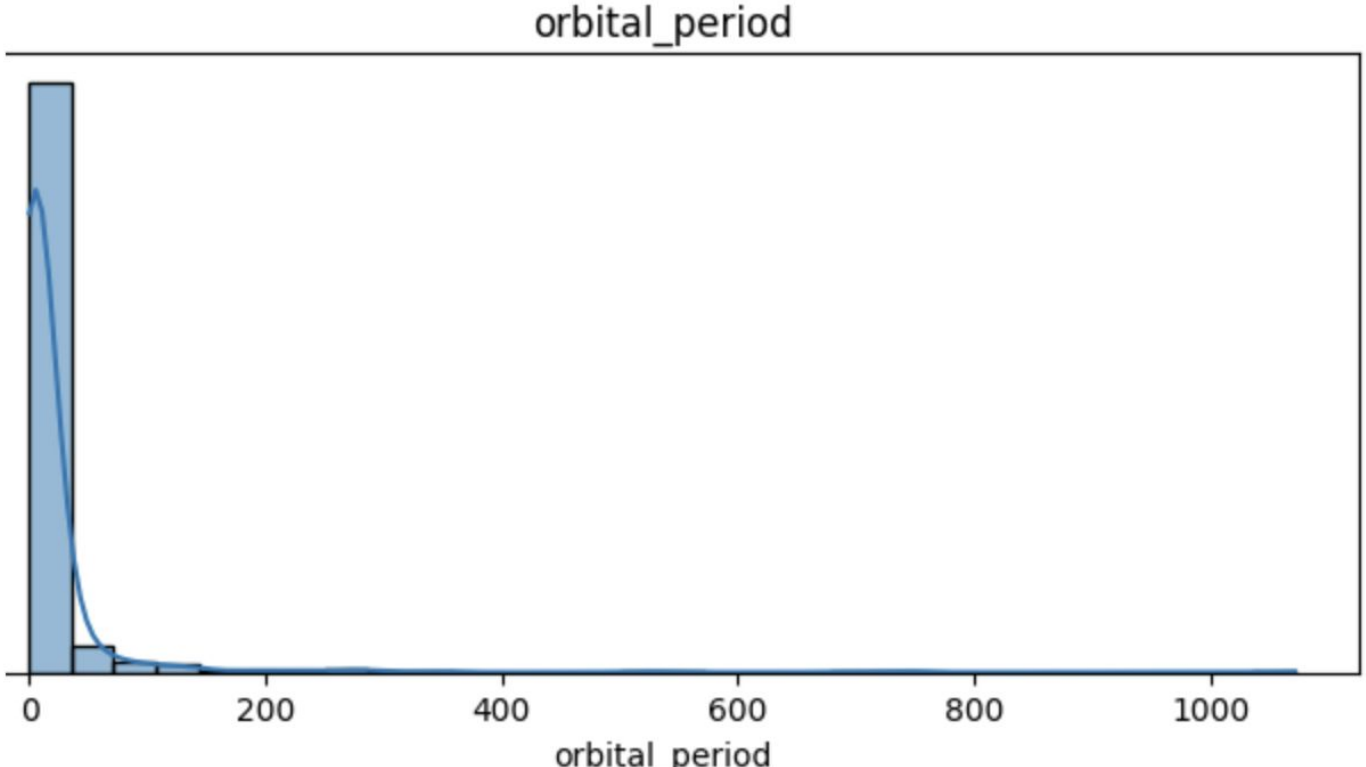
The following are summary statistics for the quantitative variables for example :

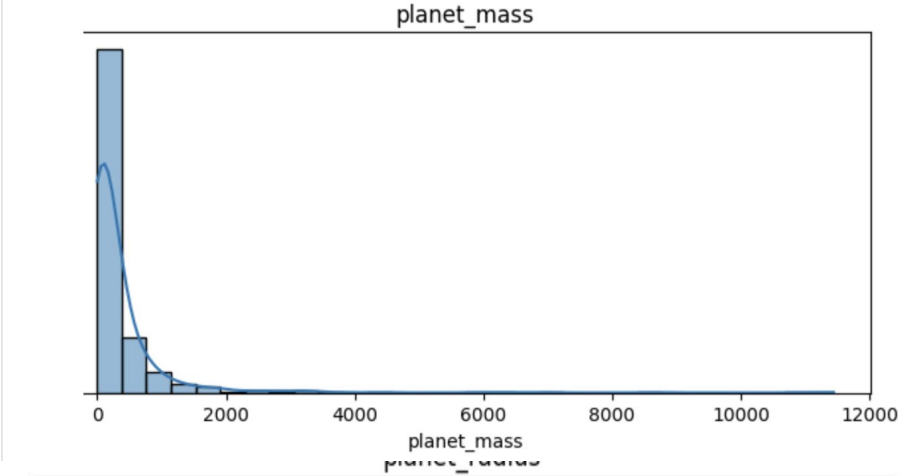
- **The Planet Mass:**
  - Mean: 343.56 Earth masses
  - Standard Deviation: 811.77 Earth masses
  - Minimum: 0.89 Earth masses
  - Maximum: 11440.80 Earth masses
- **The Planet Radius:**
  - Mean: 9.03 Earth radii
  - Standard Deviation: 5.61 Earth radii
  - Minimum: 0.88 Earth radii
  - Maximum: 20.89 Earth radii

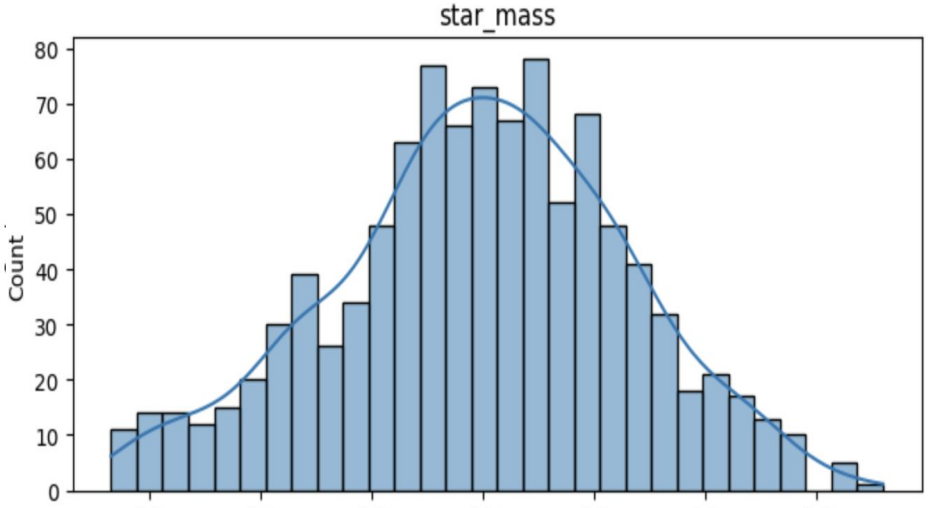
```
[ ] data.corr

<bound method DataFrame.corr of      planet_mass  planet_radius  orbital_period  orbital_radius  star_mass  \
0       8.590134       1.94544        0.736548        0.015439      1.015
1       4.512760       1.57920        3.537960        0.043600      0.910
2      36.864800       5.11952        8.463035        0.083050      0.500
3      32.097800       3.13600       18.859014        0.141700      0.500
4     146.188000      11.53600        3.487800        0.046000      0.990
...          ...           ...             ...             ...        ...
1008  197.036000      10.89760        2.615838        0.036900      0.980
1009 3746.862000      13.63040        3.191524        0.045400      1.410
1010  513.564800      14.75040        4.124730        0.054850      1.320
1011  342.270600      11.53600        4.187754        0.048700      0.880
1012  230.722800      15.07520        2.864133        0.043290      1.405

      star_radius  star_luminosity
0           0.980         3.310754
1           1.000         2.582873
2           0.750         1.508021
3           0.750         1.339036
4           1.710         2.893016
...           ...              ...
1008        0.964         2.917152
1009        1.490         5.564275
1010        1.550         4.544627
1011        1.060         2.537156
1012        1.480         5.308418

[1013 rows x 7 columns]>
```

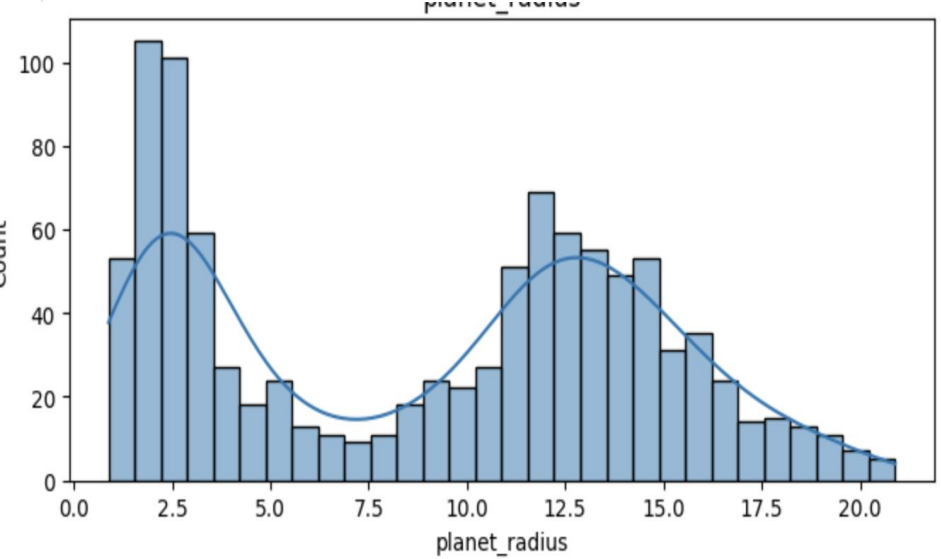

Linear Regression: Actual vs Predicted Star Luminosity
Predicted Star Luminosity
Actual Star Luminosity
1
2
3
4
5
6

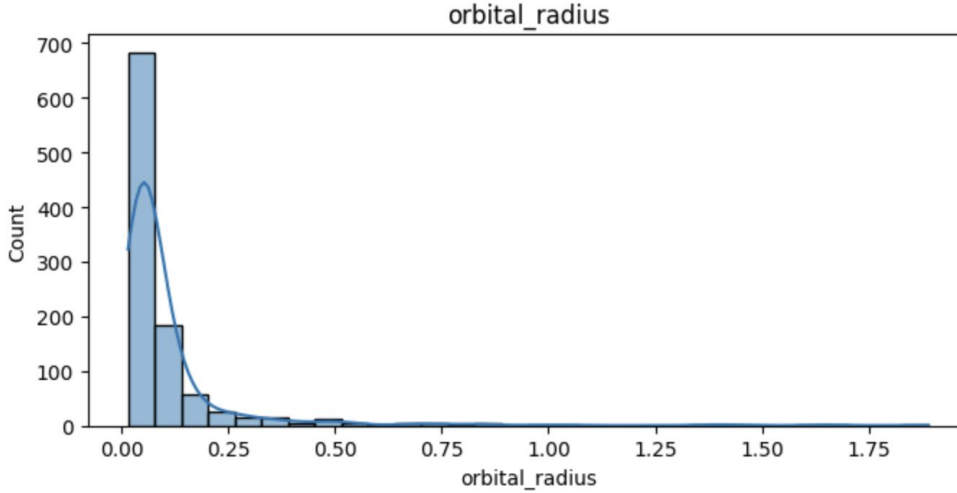
## Visualize the distribution of plane and star variables



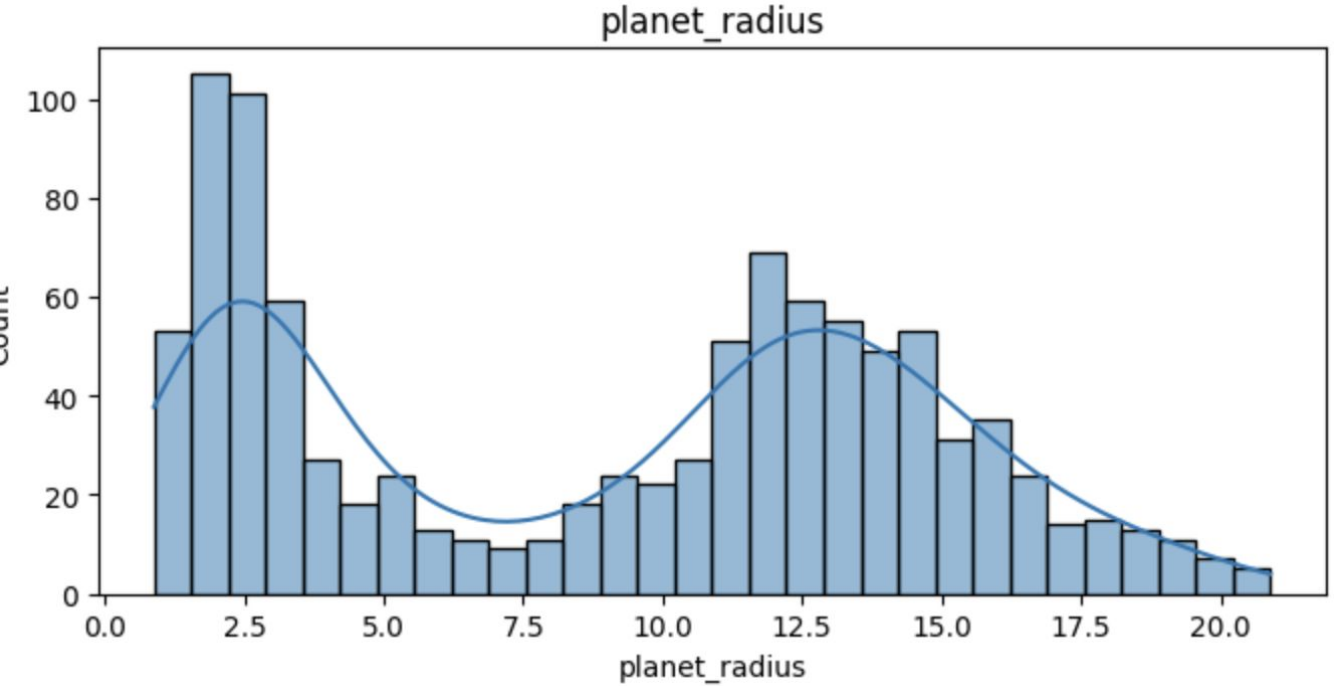

planet_mass
0
2000
4000
6000
8000
10000
12000
planet_mass


star_mass
Count
0
10
20
30
40
50
60
70
80

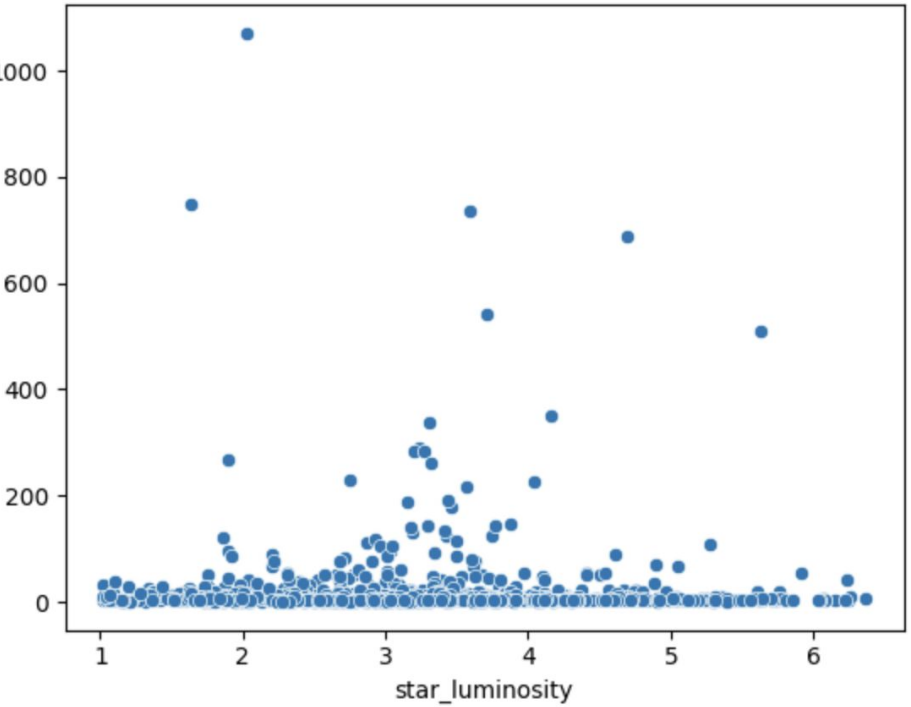

0
20
40
60
80
100
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
planet_radius

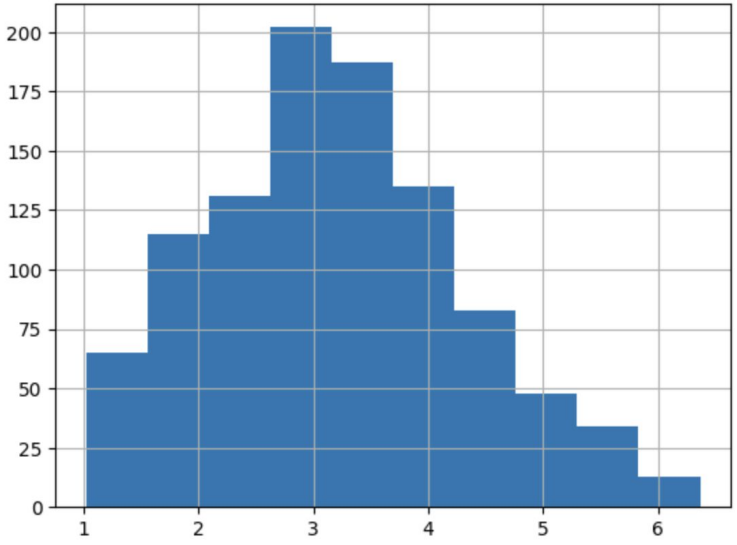

orbital_radius
Count
0
100
200
300
400
500
600
700
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
orbital_radius

The histograms and kernel density estimation offers a comprehensive view of the dataset’s characteristics .



```
data['star_luminosity'].hist()
plt.show()
```



# DISCUSSION

The accuracy is the overall correctness of the model, and in this case, it is quite high (0.995).

```
print("Classification Report:\n", classification_report(y_test, predictions))
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_split.py:700: UserWarning: The least populated class in y has onl
  warnings.warn(
Best Hyperparameters: {'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 50}
Accuracy: 0.9950738916256158
Classification Report:
              precision    recall  f1-score   support

       False       1.00      1.00      1.00       202
        True       0.00      0.00      0.00         1

    accuracy                           1.00       203
   macro avg       0.50      0.50      0.50       203
```

# OVERALL FINDINGS & IMPLICATIONS

The findings reveal 2 of habitable planets, correlations between variables, and potential outliers. These insights have implications for further exploration and refinement of dataset.

# CONCLUSION

The analysis contributes to the understanding of celestial body characteristics, emphasizing the potential habitability of planets.

The identified challenges provide opportunities for future refinement and exploration.

Thanks