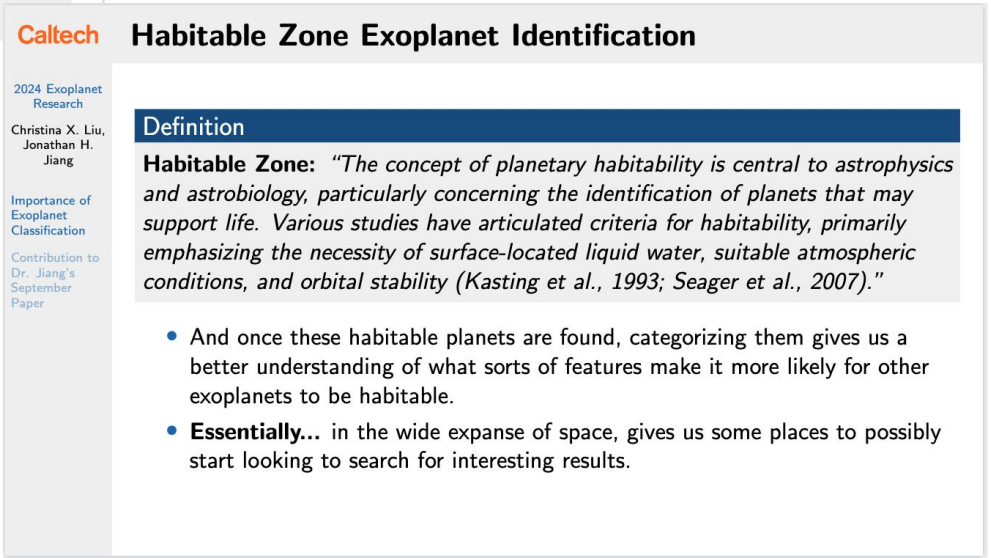
# 2024 Exoplanet Research

Results Under my Internship with Dr. Jiang

**note:** this is a draft for my final presentation, so it's a little easier for me to manage changing elements (since google slides is simpler than latex)

the real thing will look a little like:

# list of things to cover

1. table of contents – split into two main categories: the paper (and graph), and other feature research (including ML)
2. ONE SLIDE quick overview of the importance of exoplanet classification
3. tools and methodology (google colab as IDE, different data sources, python libraries (numPY, matplotlib, pandas) etc.)
4. warning slide: biggest achievements in the ending few weeks, so i've moved them first
5. transition slide into paper (first part) → introduce the paper
6. slide dedicated to explaining different definitions of the habitable zone (while also introducing the paper and pap def)
7. nasa radius-size classification of exoplanets: 6 different types (miniterran, subterran, etc)
8. pull out the graph
9. explain the venus situation (with the rest of the solar system) + assumptions and the TRAPPIST-1 system (echo back to the formula)
10. transition slide into other research (second part) → dataset change: 7203 candidate
11. interesting early findings: stellar age → lead to decision tree classifier, telling us that orbital period was the most interesting
12. orbital period graphs
13. transition slide into the future (third part) → things i might take on research in
    a. stellar age graphs – stars we should be looking at to have highest probability of harboring hz exoplanets
    b. applying decision tree classifier on the planetary system data (7203 candidates, 200-smth habitable)

# A Brief Overview of My Internship

# A Brief Overview of My Internship

- 11 weeks (officially started on **June 3rd, 2024**)

# A Brief Overview of My Internship

- 11 weeks (officially started on **June 3rd, 2024**)
- working on exoplanet classification (especially in habitable zone planets)

# A Brief Overview of My Internship

- 11 weeks (officially started on **June 3rd, 2024**)
- working on exoplanet classification (especially in habitable zone planets)
- two main categories:
  - when i started working on refining my graphs for our paper
  - initial exploration of the datasets from NASA

# Importance of Exoplanet Classification (and Habitable Zone Exoplanets)

# Importance of Exoplanet Classification (and Habitable Zone Exoplanets)

- exoplanet habitability investigation helps advance knowledge of:
  - potential extraterrestrial life

# Importance of Exoplanet Classification (and Habitable Zone Exoplanets)

- exoplanet habitability investigation helps advance knowledge of:
  - potential extraterrestrial life
  - alien environments (and their conditions)

# Importance of Exoplanet Classification (and Habitable Zone Exoplanets)

- exoplanet habitability investigation helps advance knowledge of:
  - potential extraterrestrial life
  - alien environments (and their conditions)
- exoplanet classification

# Importance of Exoplanet Classification (and Habitable Zone Exoplanets)

- exoplanet habitability investigation helps advance knowledge of:
  - potential extraterrestrial life
  - alien environments (and their conditions)
- exoplanet classification
  - compare planetary systems
  - improve detection methods
  - study the diversity of our universe
  - advance theoretical models
  - inform future missions
  - etc.

# Tools and Methodology

# Tools and Methodology

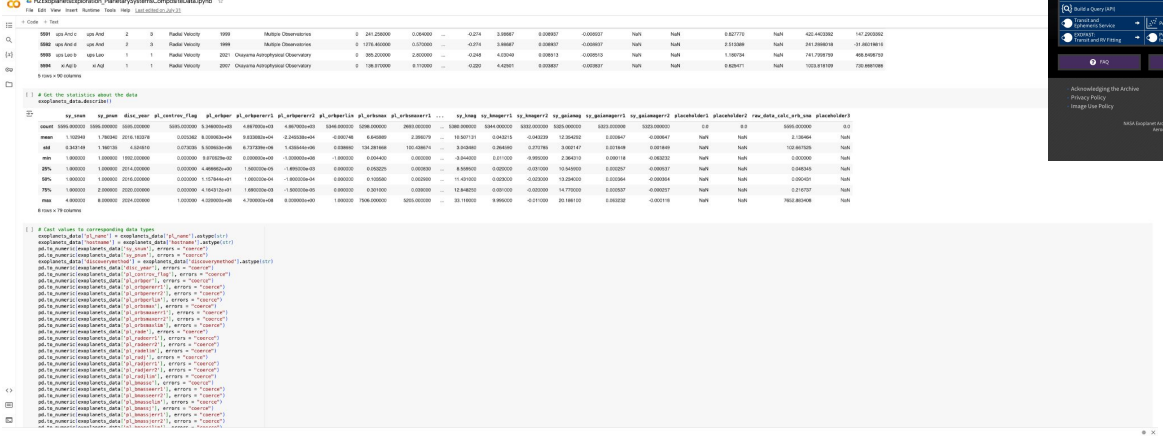
- **dataset:** from the NASA exoplanet archive
  - at different points in research, used different datasets

# Tools and Methodology

- **dataset:** from the NASA exoplanet archive
  - at different points in research, used different datasets
  - free-to-the-public

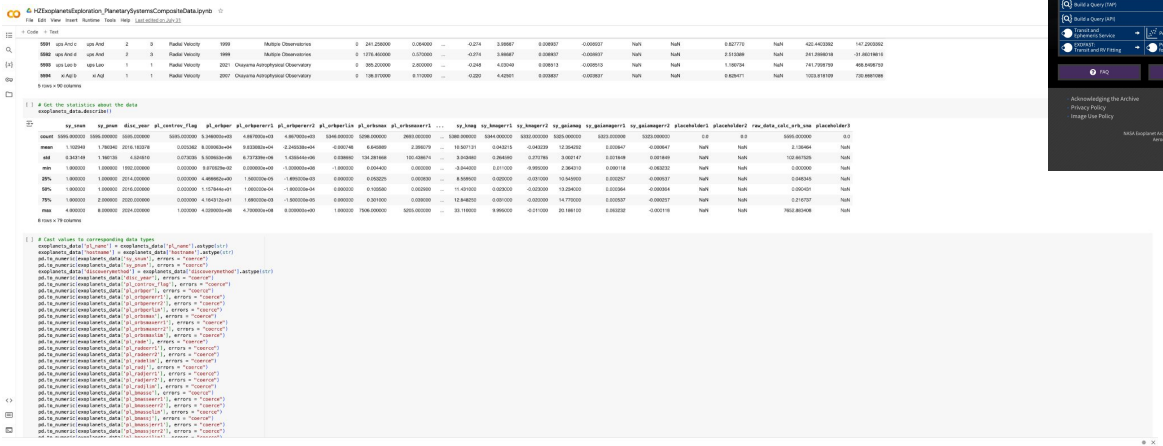
# Tools and Methodology

- **dataset:** from the NASA exoplanet archive
  - at different points in research, used different datasets
  - free-to-the-public
- **IDE:** google colab

# Tools and Methodology

- **dataset:** from the NASA exoplanet archive
  - at different points in research, used different datasets
  - free-to-the-public
- **IDE:** google colab


- **languages + packages:** Python
  - NumPy → for dealing with numbers
  - Matplotlib → for plotting graphs
  - Pandas → for data handling
  - Scikit-learn → for machine learning

# Dr. Jiang’s Paper

# Dr. Jiang’s Paper

- bringing forth different discoveries about HZ exoplanets that we’ve been researching for these past weeks and elucidating patterns in our findings

# Dr. Jiang's Paper

- bringing forth different discoveries about HZ exoplanets that we've been researching for these past weeks and elucidating patterns in our findings
- introduced a specific way of calculating the habitable zone through the **average surface temperature** ($T_{surf,ave}$)

# Dr. Jiang's Paper

- bringing forth different discoveries about HZ exoplanets that we've been researching for these past weeks and elucidating patterns in our findings
- introduced a specific way of calculating the habitable zone through the **average surface temperature** ($T_{surf,ave}$)
- used a **snapshot of the Planetary Systems Composite Data** from the NASA Exoplanet Archive on March 10th, 2024
    - **5595** total data points without any tampering

# The Habitable Zone

# The Habitable Zone

**DEFINITION**
**habitable zone**: the range in which it is possible for an exoplanet to harbor life.

# The Habitable Zone

**DEFINITION**
**habitable zone**: the range in which it is possible for an exoplanet to harbor life.

- the definition used for this paper is as follows:
  - $$T_{surf,ave} = kT_{\odot}(1-A)^{0.25}\left(\frac{R_{\odot}}{2d}\right)^{0.5}$$

where $d$ = the distance from exoplanet to its host star, $T_{\odot}$ = the host star's effective surface temperature, $R_{\odot}$ = the host star's radius, $A$ the exoplanet albedo, and $k$ the additional scaler to account for bulk atmospheric greenhouse gas.

# Categorizing Exoplanets

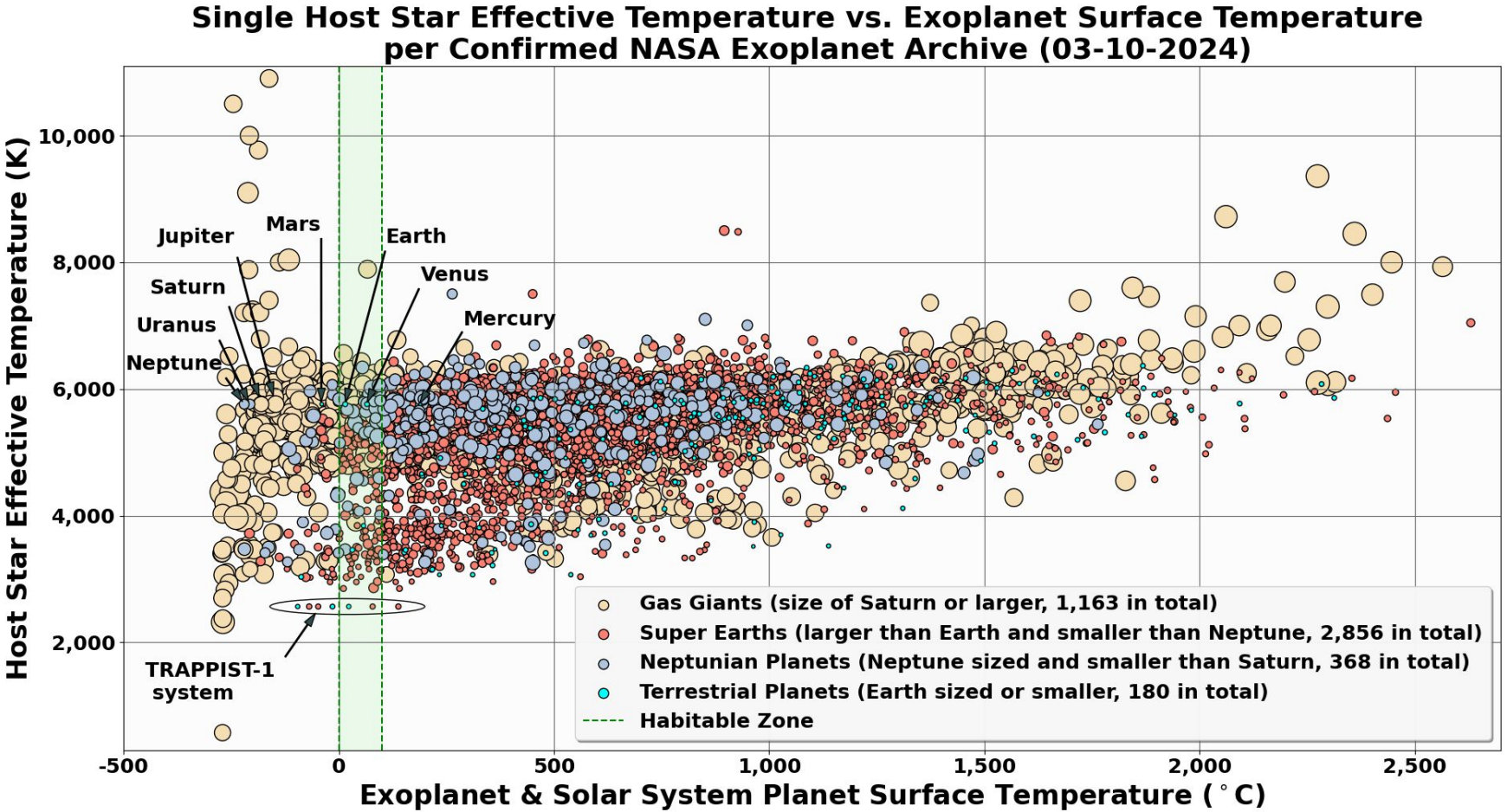
# Categorizing Exoplanets

radius size

# Categorizing Exoplanets

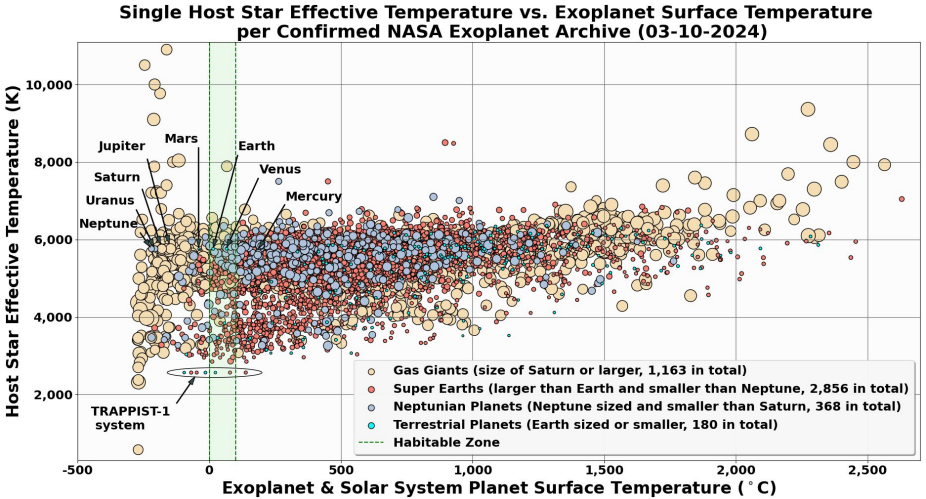
- NASA's proposed different types of exoplanets: https://science.nasa.gov/exoplanets/planet-types/
  - **terrestrial planets**:
    - radius compared to Earth: ≤ 1 times
  - **super-Earths:**
    - radius compared to Earth: 1 to 3.86 times
  - **Neptunian planets:**
    - radius compared to Earth: 3.86 to 9.14 times
  - **gas giants:**
    - radius compared to Earth: ≥ 9.14 times

# The Graph


Single Host Star Effective Temperature vs. Exoplanet Surface Temperature
per Confirmed NASA Exoplanet Archive (03-10-2024)
Host Star Effective Temperature (K)
10,000
8,000
6,000
4,000
2,000
-500
0
500
1,000
1,500
2,000
2,500
Exoplanet & Solar System Planet Surface Temperature (°C)
Jupiter
Mars
Earth
Saturn
Venus
Uranus
Mercury
Neptune
TRAPPIST-1
system
Gas Giants (size of Saturn or larger, 1,163 in total)
Super Earths (larger than Earth and smaller than Neptune, 2,856 in total)
Neptunian Planets (Neptune sized and smaller than Saturn, 368 in total)
Terrestrial Planets (Earth sized or smaller, 180 in total)
Habitable Zone

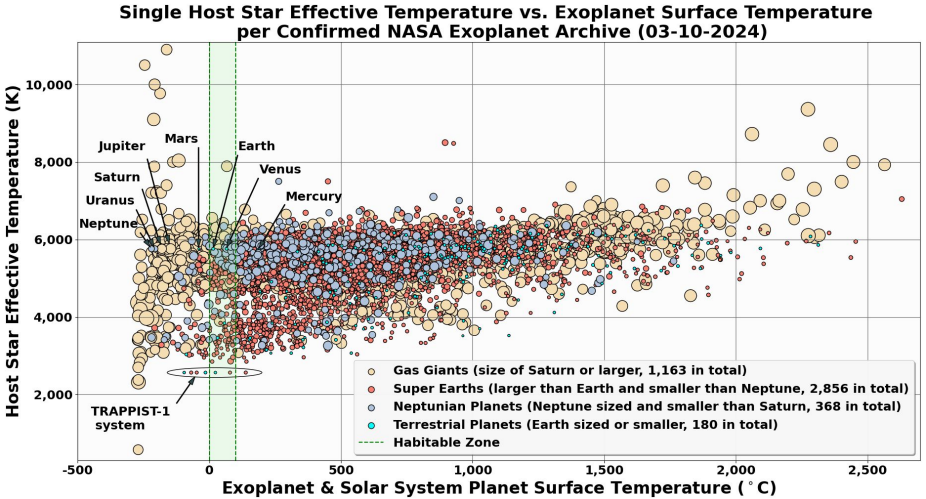
# Things I Included:

- all of the solar system planets:



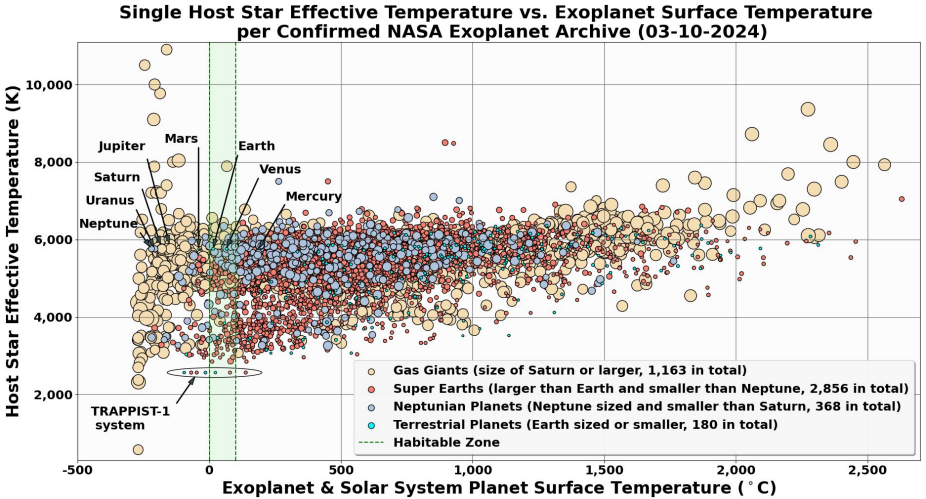
# Things I Included:

- all of the solar system planets:
  - Temperatures, sizes, and general data from official NASA figures



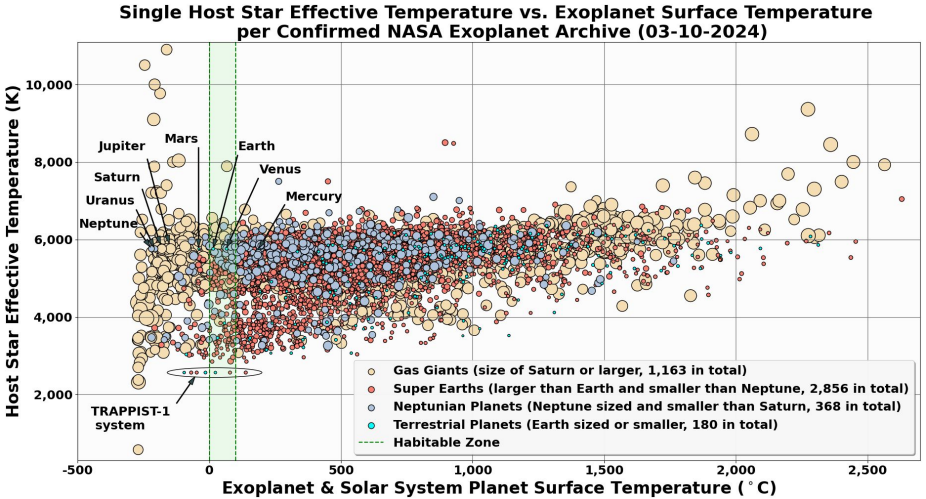
# Things I Included:

- all of the solar system planets:
  - Temperatures, sizes, and general data from official NASA figures
  - Venus.



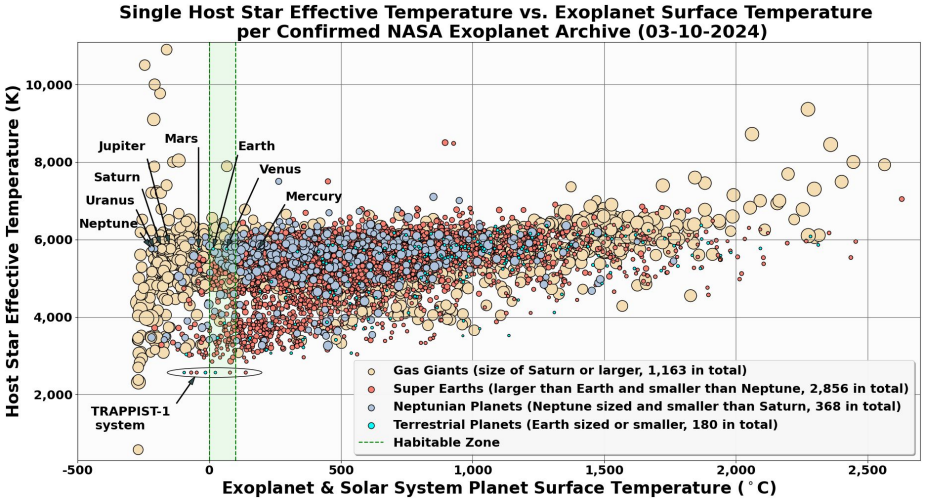
# Things I Included:

- all of the solar system planets:
  - Temperatures, sizes, and general data from official NASA figures
  - Venus.
- TRAPPIST-1 system



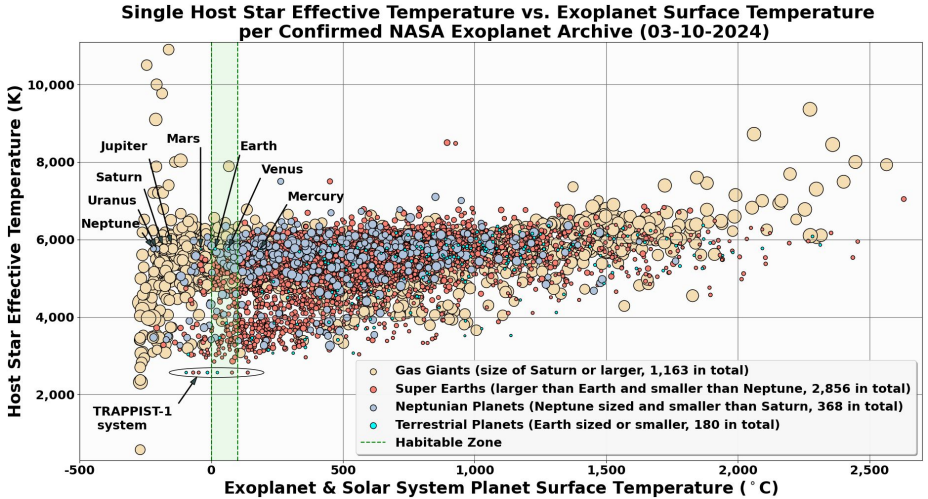
# Things I Included:

- all of the solar system planets:
  - Temperatures, sizes, and general data from official NASA figures
  - Venus.
- TRAPPIST-1 system
  - 7 terrestrial, Earth-sized planets discovered orbiting the red dwarf star TRAPPIST-1
  - 3-4 confirmed habitable by NASA, 2 by us



# Things I Included:

- all of the solar system planets:
  - Temperatures, sizes, and general data from official NASA figures
  - Venus.
- TRAPPIST-1 system
  - 7 terrestrial, Earth-sized planets discovered orbiting the red dwarf star TRAPPIST-1
  - 3-4 confirmed habitable by NASA, 2 by us
- data stats:
  - total data points:




| pl_hz_status | count |
|---|---|
| Too Hot | 3942 |
| Too Cold | 399 |
| In HZ | 226 |

| pl_type | count |
|---|---|
| Super-Earths | 2856 |
| Gas-Giants | 1163 |
| Neptunian-Planets | 368 |
| Terrestrial-Planets | 180 |

**4567** data points

# The Venus Conundrum

# The Venus Conundrum

You’ve probably noticed it too: what’s happening with Venus?

# The Venus Conundrum

You've probably noticed it too: what's happening with Venus?

$$T_{surf,ave} = kT_{\odot}(1-A)^{0.25}\left(\frac{R_{\odot}}{2d}\right)^{0.5}$$

# The Venus Conundrum

You've probably noticed it too: what's happening with Venus?

$$T_{surf,ave} = kT_{\odot}(1-A)^{0.25}\left(\frac{R_{\odot}}{2d}\right)^{0.5}$$

- $T_{\odot}$ is the effective temperature of **the sun** (same across all planets)
- $R_{\odot}$ is the radius of **the sun** (same across all planets)

# The Venus Conundrum

You've probably noticed it too: what's happening with Venus?

$$T_{surf,ave} = kT_{\odot}(1-A)^{0.25}\left(\frac{R_{\odot}}{2d}\right)^{0.5}$$

- $T_{\odot}$ is the effective temperature of **the sun** (same across all planets)
- $R_{\odot}$ is the radius of **the sun** (same across all planets)
- 2.6 Assumptions and Limitations

  Our analysis was underpinned by several assumptions, notably adopting Earth's albedo (A = 0.306) as a baseline for exoplanets as well as accounting for the atmospheric greenhouse gas effect through the bulk temperature factor (k = 1.13), again using Earth as the standard. Recognizing that our empirical relationships, based on a large but nonetheless limited dataset, might introduce certain biases, we were careful to frame our findings within these constraints. Where assumptions were necessary to complete calculations, rational bracketing conditions were applied accordingly.

# The Venus Conundrum

You've probably noticed it too: what's happening with Venus?

$$T_{surf,ave} = kT_{\odot}(1 - A)^{0.25}\left(\frac{R_{\odot}}{2d}\right)^{0.5}$$

- $T_{\odot}$ is the effective temperature of **the sun** (same across all planets)
- $R_{\odot}$ is the radius of **the sun** (same across all planets)
- 2.6 Assumptions and Limitations

  Our analysis was underpinned by several assumptions, notably adopting Earth's albedo (A = 0.306) as a baseline for exoplanets as well as accounting for the atmospheric greenhouse gas effect through the bulk temperature factor (k = 1.13), again using Earth as the standard. Recognizing that our empirical relationships, based on a large but nonetheless limited dataset, might introduce certain biases, we were careful to frame our findings within these constraints. Where assumptions were necessary to complete calculations, rational bracketing conditions were applied accordingly.

  ***A*** = 0.306 (same across all planets)

  ***k*** = 1.13 (same across all planets)

# The Venus Conundrum

You've probably noticed it too: what's happening with Venus?

$$T_{surf,ave} = kT_{\odot}(1 - A)^{0.25}\left(\frac{R_{\odot}}{2d}\right)^{0.5}$$

- $T_{\odot}$ is the effective temperature of **the sun** (same across all planets)
- $R_{\odot}$ is the radius of **the sun** (same across all planets)
- 2.6 Assumptions and Limitations

  Our analysis was underpinned by several assumptions, notably adopting Earth's albedo (A = 0.306) as a baseline for exoplanets as well as accounting for the atmospheric greenhouse gas effect through the bulk temperature factor (k = 1.13), again using Earth as the standard. Recognizing that our empirical relationships, based on a large but nonetheless limited dataset, might introduce certain biases, we were careful to frame our findings within these constraints. Where assumptions were necessary to complete calculations, rational bracketing conditions were applied accordingly.

***A*** = 0.306 (same across all planets)

***k*** = 1.13 (same across all planets)

**so the only thing that's different is d (distance from the sun).**

# Venus in the Habitable Zone…

# Venus in the Habitable Zone…

- In the case of venus, venus's atmosphere is extremely heavy, which means that its *k* value should have been much higher than the Earth standard we used across the solar system.

# Venus in the Habitable Zone…

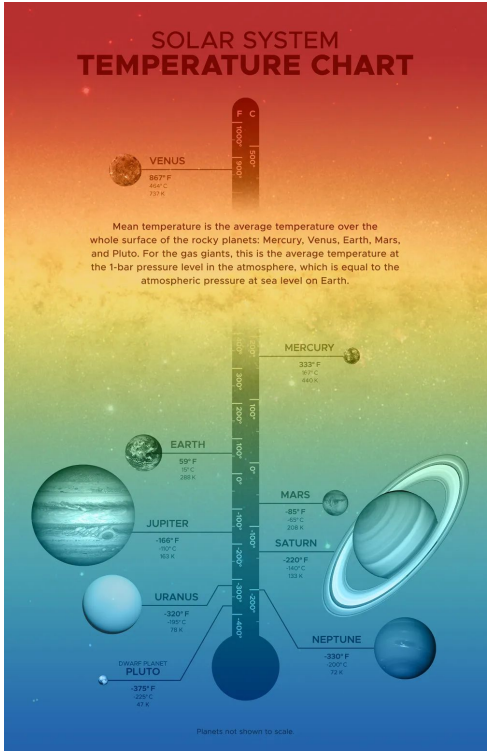
- In the case of venus, venus's atmosphere is extremely heavy, which means that its *k* value should have been much higher than the Earth standard we used across the solar system.
- But in general, we weren't too off:

| | | |
|---|---|---|
| **MERCURY** | 0 | 188.396908 |
| **VENUS** | 1 | 64.527543 |
| **EARTH** | 2 | 13.975212 |
| **MARS** | 3 | −40.566530 |
| **JUPITER** | 4 | −147.285665 |
| **SATURN** | 5 | −180.398576 |
| **URANUS** | 6 | −207.607572 |
| **NEPTUNE** | 7 | −220.789400 |

^^ FORMULA RESULTS ^^

- Mercury: 333°F (167°C)
- Venus: 867°F (464°C)
- Earth: 59°F (15°C)
- Mars: Minus 85°F (-65°C)
- Jupiter: Minus 166°F (-110°C)
- Saturn: Minus 220°F (-140°C)
- Uranus: Minus 320°F (-195°C)
- Neptune: Minus 330°F (-200°C)
- Dwarf Planet Pluto: Minus 375°F (-225°C)

^^ NASA DATA ^^



# The Venus Paragraph

This figure also brings attention to assumption limitations introduced earlier in this paper. For the planets in our solar system, surface temperatures calculated with Equation (1) generally align with known mean temperatures (https://science.nasa.gov/resource/solar-system-temperatures/) with a 6-37% difference. However, Venus is an exception. Although its surface is too hot for life as we know it, Equation (1) flags the planet as within the habitable zone. This discrepancy arises from the assumption of a standardized bulk temperature factor (k=1.13k) based on Earth's values when accounting for the atmospheric greenhouse effect. In reality, Venus has a very thick atmosphere composed primarily of $CO_2$, trapping heat and resulting in a much higher bulk temperature factor (k=3.17). This limitation is discussed earlier in section 2.6. Bracketing the inner Solar System-based atmospheric greenhouse assumption, this on the cooler end, is Mars. While the atmosphere of Mars is also predominantly composed of $CO_2$, it is far less dense and accordingly much less capable of trapping solar radiation. The particular exception of Venus indicates that variations in the atmospheric greenhouse effect will need to be further considered to better determine exoplanet surface temperatures.

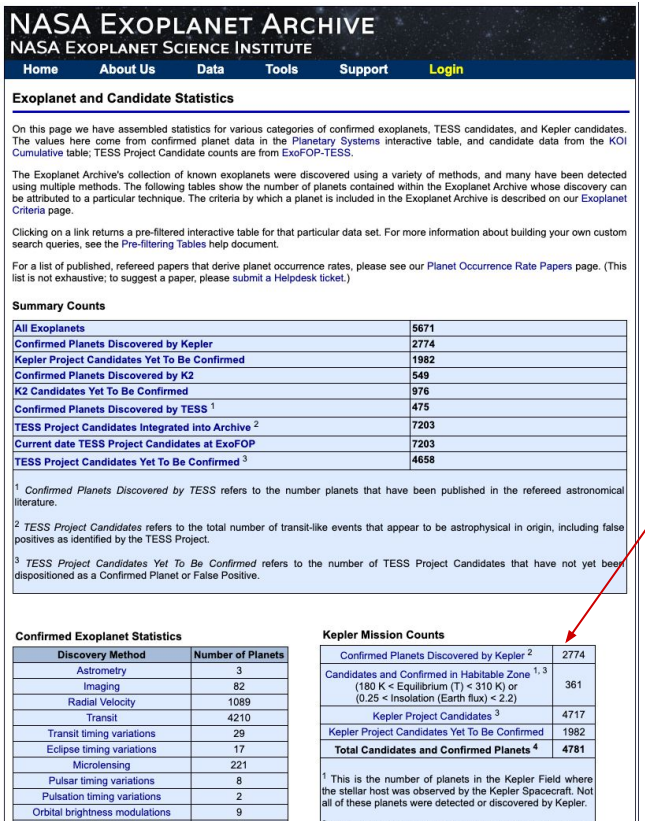
# Section 2

Initial Exploration

# Stellar Age and Orbital Period v. Number of Exoplanets

**How much do different features of the dataset affect the number of (habitable) exoplanets that share that feature?**

# Data Source



NASA EXOPLANET ARCHIVE
NASA EXOPLANET SCIENCE INSTITUTE

Home | About Us | Data | Tools | Support | Login

**Exoplanet and Candidate Statistics**

On this page we have assembled statistics for various categories of confirmed exoplanets, TESS candidates, and Kepler candidates. The values here come from confirmed planet data in the Planetary Systems interactive table, and candidate data from the KOI Cumulative table; TESS Project Candidate counts are from ExoFOP-TESS.

The Exoplanet Archive's collection of known exoplanets were discovered using a variety of methods, and many have been detected using multiple methods. The following tables show the number of planets contained within the Exoplanet Archive whose discovery can be attributed to a particular technique. The criteria by which a planet is included in the Exoplanet Archive is described on our Exoplanet Criteria page.

Clicking on a link returns a pre-filtered interactive table for that particular data set. For more information about building your own custom search queries, see the Pre-filtering Tables help document.

For a list of published, refereed papers that derive planet occurrence rates, please see our Planet Occurrence Rate Papers page. (This list is not exhaustive; to suggest a paper, please submit a Helpdesk ticket.)

**Summary Counts**

| | |
|---|---|
| **All Exoplanets** | **5671** |
| **Confirmed Planets Discovered by Kepler** | **2774** |
| **Kepler Project Candidates Yet To Be Confirmed** | **1982** |
| **Confirmed Planets Discovered by K2** | **549** |
| **K2 Candidates Yet To Be Confirmed** | **976** |
| **Confirmed Planets Discovered by TESS** [1] | **475** |
| **TESS Project Candidates Integrated into Archive** [2] | **7203** |
| **Current date TESS Project Candidates at ExoFOP** | **7203** |
| **TESS Project Candidates Yet To Be Confirmed** [3] | **4658** |

[1] *Confirmed Planets Discovered by TESS* refers to the number planets that have been published in the refereed astronomical literature.

[2] *TESS Project Candidates* refers to the total number of transit-like events that appear to be astrophysical in origin, including false positives as identified by the TESS Project.

[3] *TESS Project Candidates Yet To Be Confirmed* refers to the number of TESS Project Candidates that have not yet been dispositioned as a Confirmed Planet or False Positive.

**Confirmed Exoplanet Statistics**

| Discovery Method | Number of Planets |
|---|---|
| Astrometry | 3 |
| Imaging | 82 |
| Radial Velocity | 1089 |
| Transit | 4210 |
| Transit timing variations | 29 |
| Eclipse timing variations | 17 |
| Microlensing | 221 |
| Pulsar timing variations | 8 |
| Pulsation timing variations | 2 |
| Orbital brightness modulations | 9 |

**Kepler Mission Counts**

| | |
|---|---|
| Confirmed Planets Discovered by Kepler [2] | 2774 |
| Candidates and Confirmed in Habitable Zone [1, 3] (180 K < Equilibrium (T) < 310 K) or (0.25 < Insolation (Earth flux) < 2.2) | 361 |
| Kepler Project Candidates [3] | 4717 |
| Kepler Project Candidates Yet To Be Confirmed | 1982 |
| **Total Candidates and Confirmed Planets** [4] | **4781** |

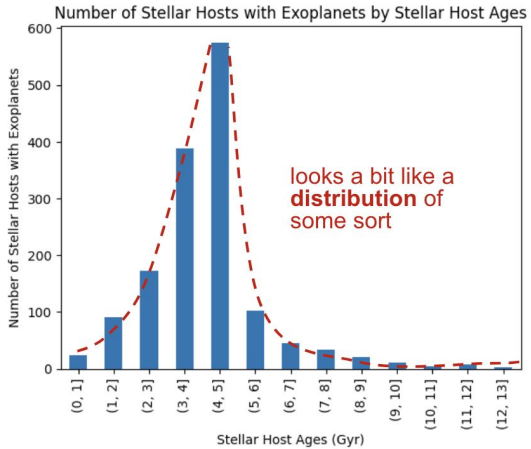
[1] This is the number of planets in the Kepler Field where the stellar host was observed by the Kepler Spacecraft. Not all of these planets were detected or discovered by Kepler.

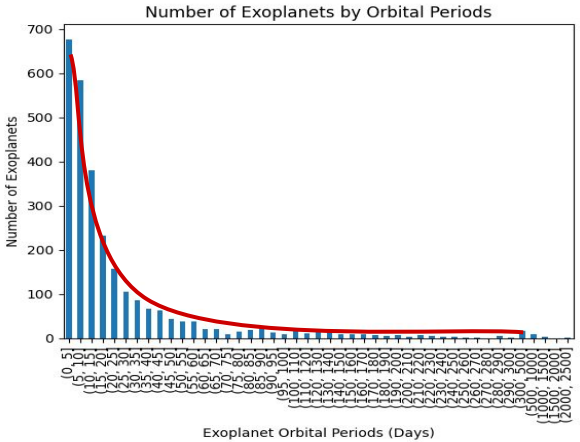
- Data from: Confirmed Planets Discovered by Kepler
  - 2774 confirmed

# Stellar Age and Orbital Period v. Number of Exoplanets

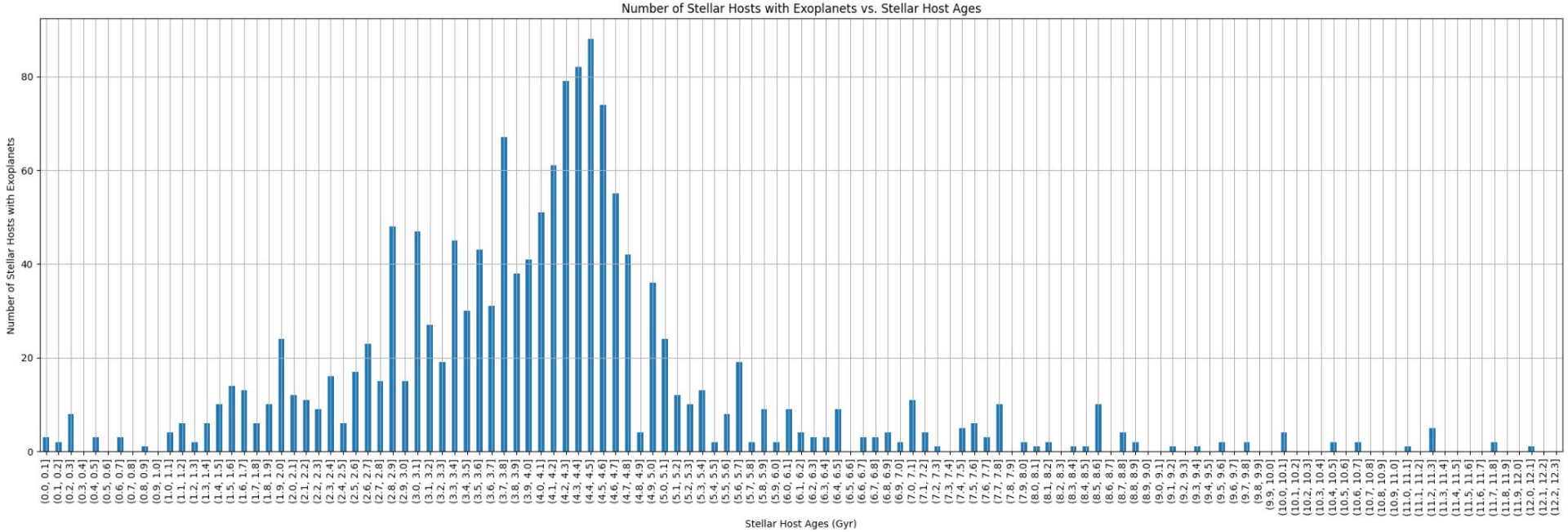
## How much do different features of the dataset affect the number of (habitable) exoplanets that share that feature?

stellar age and orbital periods came up very early on:

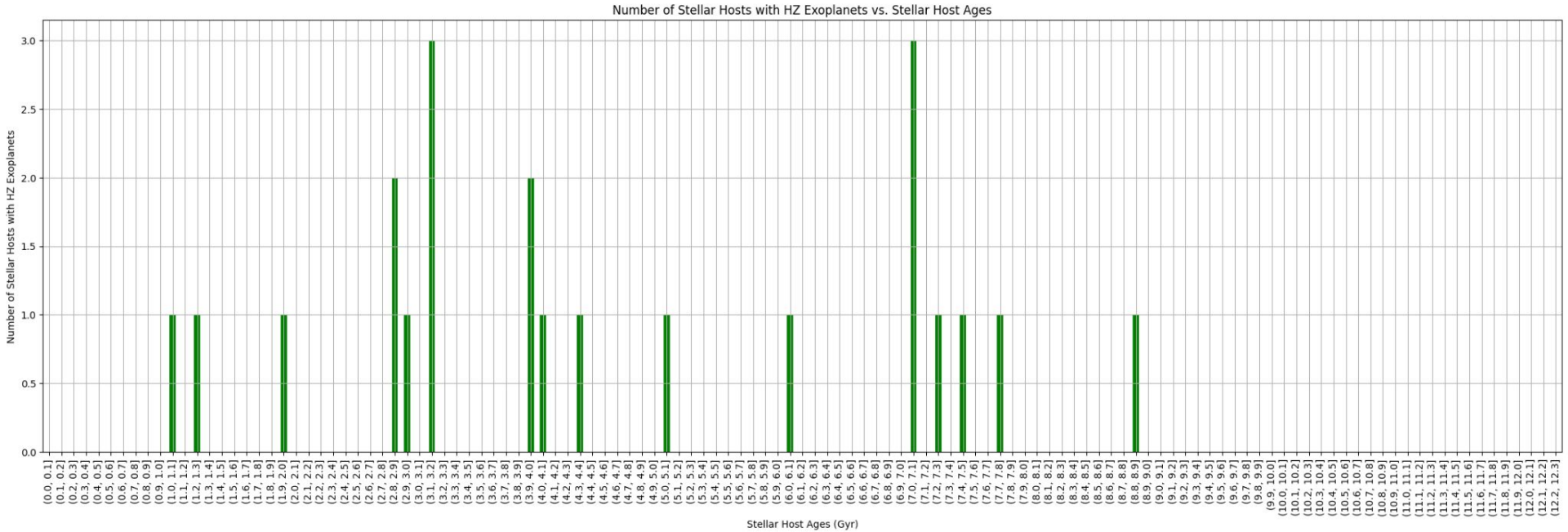


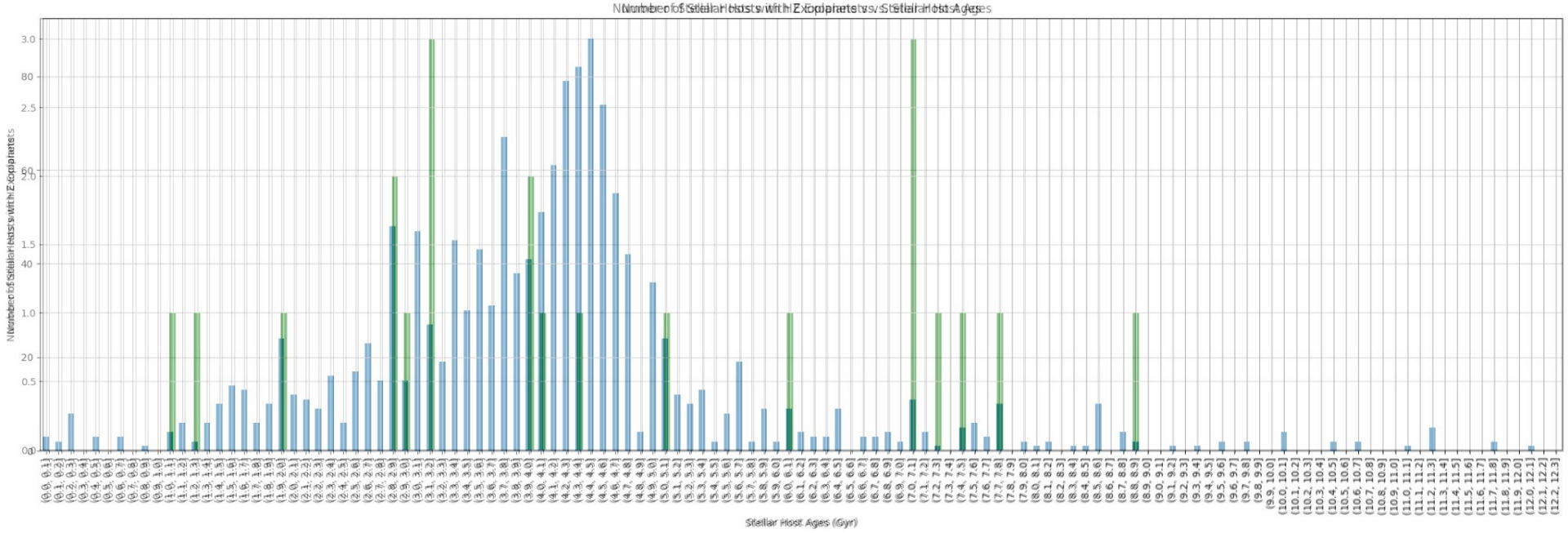
# Stellar Age v. Number of Exoplanets



# Stellar Age v. Number of HZ Exoplanets



# Stellar Age



# What other relationships could have been missed?

# What other relationships could have been missed?

we can look into this with a **decision tree machine learning model!**

# Habitable Zone Determinations

data from [kepler confirmed exoplanet archive]

**Kepler Mission Counts**

| | |
|---|---|
| Confirmed Planets Discovered by Kepler [2] | 2774 |
| Candidates and Confirmed in Habitable Zone [1, 3]<br>(180 K < Equilibrium (T) < 310 K) or<br>(0.25 < Insolation (Earth flux) < 2.2) | 361 |

**two ways** to identify habitable zone exoplanets (according to NASA exopl. archive):

**TEMPERATURE**

**INSOLATION**

any one of them being met is enough (?)

# Decision Tree Classifier – Data Labeling

- Data source: [Confirmed Planets Discovered By Kepler](#) (2773 items)
- Labeling data as below (HZ criteria according to NASA exopl. archive)

```
[13] exoplanets_data.loc[((~np.isnan(exoplanets_data['pl_eqt'])) & (exoplanets_data['pl_eqt'] > 180) & (exoplanets_data['pl_eqt'] < 310)), 'hz_label_by_eqt'] = 1
     exoplanets_data.loc[((~np.isnan(exoplanets_data['pl_eqt'])) & (exoplanets_data['pl_eqt'] <= 180) | (exoplanets_data['pl_eqt'] >= 310)), 'hz_label_by_eqt'] = 0
     exoplanets_data['hz_label_by_eqt'].value_counts()

     hz_label_by_eqt
     0.0    193
     1.0     20
     Name: count, dtype: int64
```

**equilibrium temperature fiddling** – if within range, set hz_label_by_eqt 1, otherwise 0. there are **20** in the correct zone

```
[14] exoplanets_data.loc[((~np.isnan(exoplanets_data['pl_insol'])) & (exoplanets_data['pl_insol'] > 0.25) & (exoplanets_data['pl_insol'] < 2.2)), 'hz_label_by_insol'] =
     exoplanets_data.loc[((~np.isnan(exoplanets_data['pl_insol'])) & (exoplanets_data['pl_insol'] <= 0.25) | (exoplanets_data['pl_insol'] >= 2.2)), 'hz_label_by_insol']
     exoplanets_data['hz_label_by_insol'].value_counts()

     hz_label_by_insol
     0.0    135
     1.0     22
     Name: count, dtype: int64
```

**earth flux insolation fiddling** – if within range, set hz_label_by_insol 1, otherwise 0. there are **22** in the correct zone

```
[15] exoplanets_data.loc[(((~np.isnan(exoplanets_data['hz_label_by_eqt'])) & (exoplanets_data['hz_label_by_eqt'] == 1)) | ((~np.isnan(exoplanets_data['hz_label_by_insol'
     exoplanets_data.loc[(((~np.isnan(exoplanets_data['hz_label_by_eqt'])) & (exoplanets_data['hz_label_by_eqt'] == 0)) & ((~np.isnan(exoplanets_data['hz_label_by_insol'
     exoplanets_data.loc[(((~np.isnan(exoplanets_data['hz_label_by_eqt'])) & (exoplanets_data['hz_label_by_eqt'] == 0)) & (np.isnan(exoplanets_data['hz_label_by_insol'])
     exoplanets_data.loc[((np.isnan(exoplanets_data['hz_label_by_eqt'])) & ((~np.isnan(exoplanets_data['hz_label_by_insol'])) & (exoplanets_data['hz_label_by_insol'] ==
     exoplanets_data['hz_label'].value_counts()

     hz_label
     0.0    213
     1.0     31
     Name: count, dtype: int64
```

**both of them together** – if EITHER of the above are set to 1, the “big” label - hz_label is set to 1, otherwise 0.

# Decision Tree Classifier – Data Cleaning & preparation

- casting all numerical values to floats + setting empty fields to NaN
- cleaning data:
  - dropping rows without labels,
  - dropping irrelevant data fields (e.g., exoplanet name, stellar host name, etc.)
  - dropping data fields with too many missing values
  - dropping redundant data fields (e.g. exoplanet radius in Jupiter scales - there is an exoplanet radius data field in Earth scale)
  - dropping data fields that are used for labeling (e.g. exoplanet equilibrium temperature, exoplanet insolution flux)
- imputation - filling in missing values with the mean
- standard scalar – scale features to standardized values

# Decision Tree Classifier – Training Data Stats

Using the labels we have created, training the model using **213 negative and 31 positive samples,** which will be **oversampled** later to account for skew. we have **19 features** in the training.

```
[ ]  # all columns except "hz_lable" become features in the data for model training and testing.
     # "hz_label" is the label in the data for model training and testing
     features = training_data.drop(['hz_label'], axis = 1)
     labels = training_data.hz_label
```

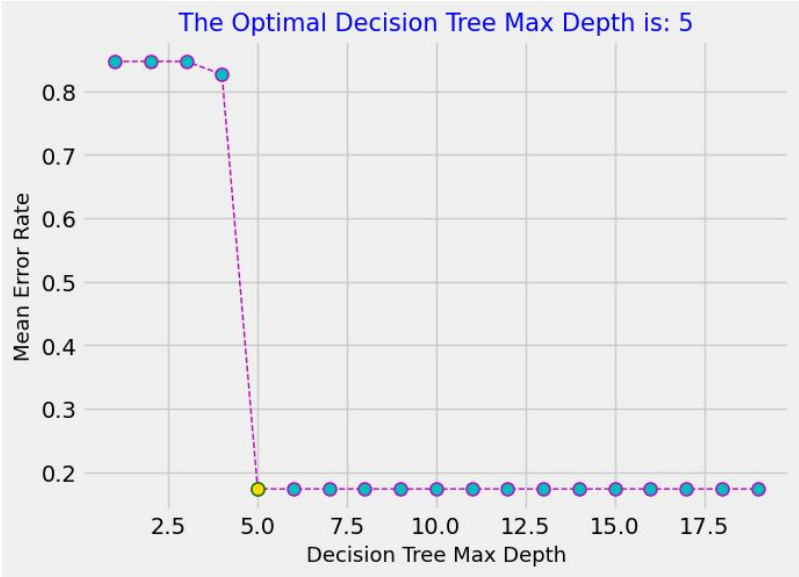
```
[ ]  # we have 213 negative and 31 positive samples in the training data
     # the training data is inbalanced; we'll use oversampling to handle this later on
     labels.value_counts()
```

```
hz_label
0.0    213
1.0     31
Name: count, dtype: int64
```

# Decision Tree Classifier – Optimal Max Tree Depth

I then found the **optimal max tree depth** by finding the lowest average error rate after training the models (for each max tree depth in the range of [1,19], trained models 100 times, each time with data randomly shuffled and then split with 60% as training data and 40% as testing data), and the optimal max tree depth turned out to be **5.**

```
Searching the optimal decision tree max depth in between 1 ~ 19

    In progress...
    For decision tree with max depth 1, mean prediction error rate = 0.847
    For decision tree with max depth 2, mean prediction error rate = 0.847
    For decision tree with max depth 3, mean prediction error rate = 0.847
    For decision tree with max depth 4, mean prediction error rate = 0.827
    For decision tree with max depth 5, mean prediction error rate = 0.173
    For decision tree with max depth 6, mean prediction error rate = 0.173
    For decision tree with max depth 7, mean prediction error rate = 0.173
    For decision tree with max depth 8, mean prediction error rate = 0.173
    For decision tree with max depth 9, mean prediction error rate = 0.173
    For decision tree with max depth 10, mean prediction error rate = 0.173
    For decision tree with max depth 11, mean prediction error rate = 0.173
    For decision tree with max depth 12, mean prediction error rate = 0.173
    For decision tree with max depth 13, mean prediction error rate = 0.173
    For decision tree with max depth 14, mean prediction error rate = 0.173
    For decision tree with max depth 15, mean prediction error rate = 0.173
    For decision tree with max depth 16, mean prediction error rate = 0.173
    For decision tree with max depth 17, mean prediction error rate = 0.173
    For decision tree with max depth 18, mean prediction error rate = 0.173
    For decision tree with max depth 19, mean prediction error rate = 0.173

    Done! The Optimal Decision Tree Max Depth is: 5
```



# Decision Tree Classifier – Set Up Training

```
[ ]  # split data with into training and testing sets
     features_train, features_test, labels_train, labels_test = train_test_split(features,
                                                                                  labels,
                                                                                  test_size=split_test_data_percentage,
                                                                                  random_state=0,
                                                                                  shuffle=True)

     # standadize the scales of features
     features_train_sc = stand_scaler.fit_transform(features_train)
     features_test_sc = stand_scaler.transform(features_test)

     # randomly oversample the positive samples to balance training data
     ros = RandomOverSampler()
     features_train_sc_ros, labels_train_ros = ros.fit_resample(features_train_sc, labels_train)
```

# Decision Tree Classifier – Training the Model

- using the optimal max tree depth
- using entropy as the criterion to select features and thresholds to split and build the tree hierarchy

```
[ ] # train a Decision Tree classifier with the training data
    decision_tree_classifier = DecisionTreeClassifier(criterion='entropy', max_depth=optimal_max_depth, random_state=0)
    decision_tree_classifier.fit(features_train_sc_ros, labels_train_ros)
```

```
DecisionTreeClassifier
DecisionTreeClassifier(criterion='entropy', max_depth=5, random_state=0)
```

# Decision Tree Classifier – Accuracy, Precision, Recall, etc.

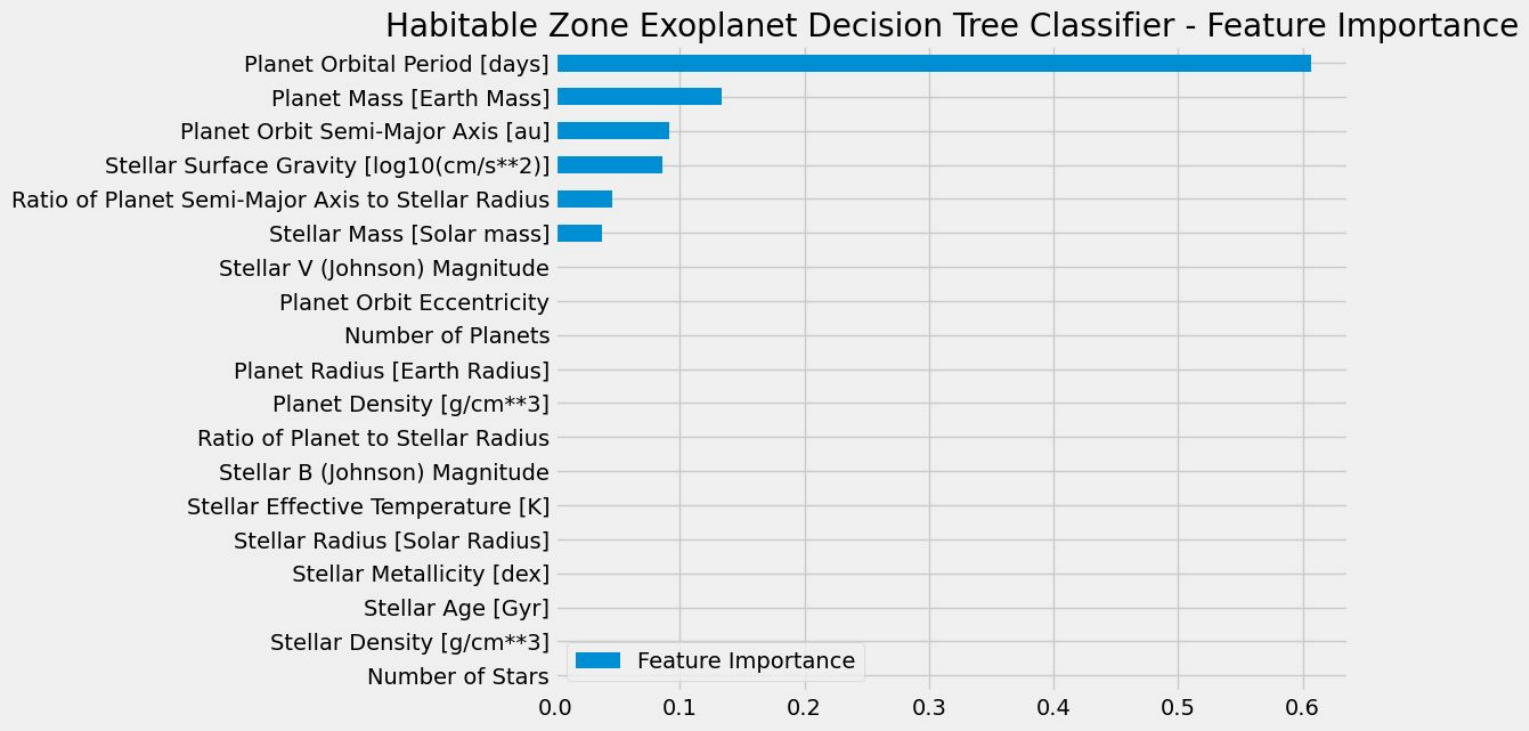
```
[ ] # calculate accuracy, precision, recall, and F-1 scores for the Decision Tree classifier
    print("Decision Tree Classifier Accuracy: ", accuracy_score(labels_test, labels_pred))
    print()
    print("Decision Tree Classification Report :\n",(classification_report(labels_test,labels_pred)))
```

```
Decision Tree Classifier Accuracy:  0.8877551020408163

Decision Tree Classification Report :
              precision    recall  f1-score   support

         0.0       0.91      0.96      0.94        83
         1.0       0.70      0.47      0.56        15

    accuracy                           0.89        98
   macro avg       0.80      0.72      0.75        98
weighted avg       0.88      0.89      0.88        98
```

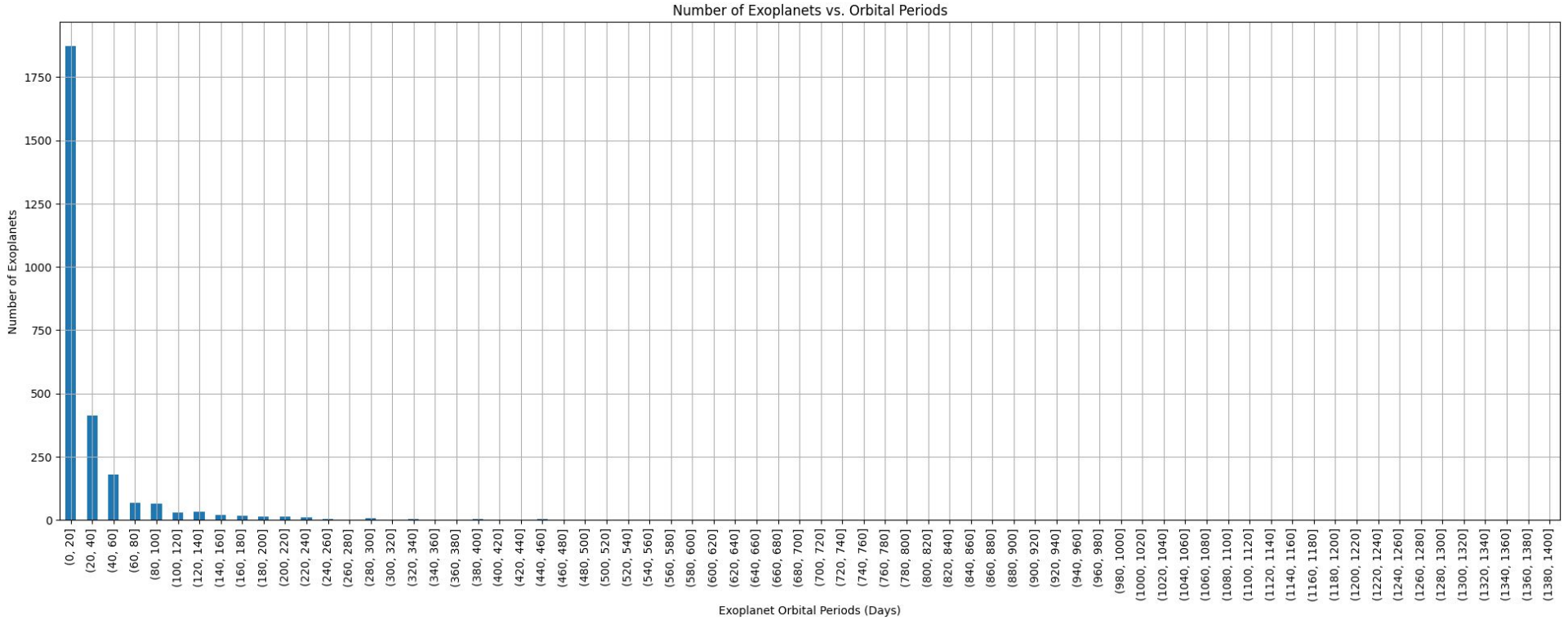
compare with: **KNN HZ exoplanet classifier**

```
KNN Classifier Accuracy:  0.8648648648648649

KNN Classifier Classification Report :
              precision    recall  f1-score   support

         0.0       0.89      0.95      0.92        62
         1.0       0.62      0.42      0.50        12

    accuracy                           0.86        74
   macro avg       0.76      0.68      0.71        74
weighted avg       0.85      0.86      0.85        74
```

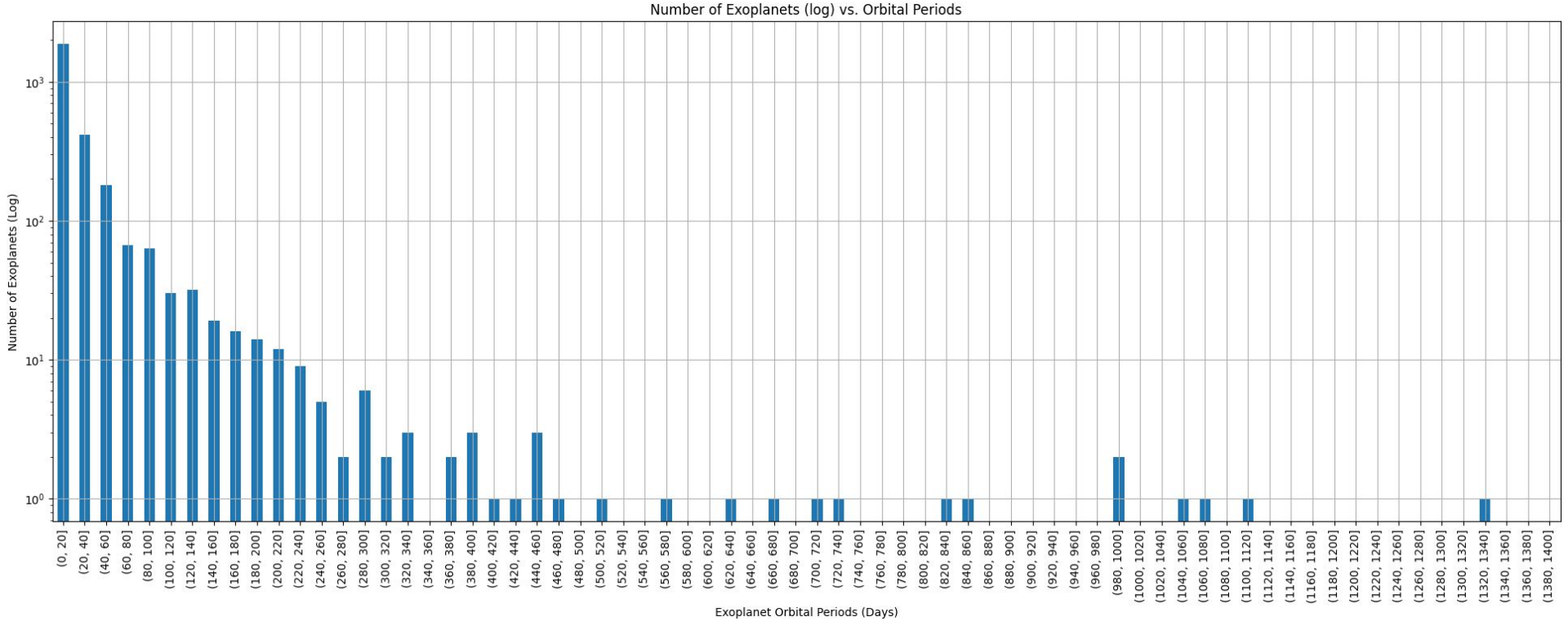
# Decision Tree Classifier – Feature Importance



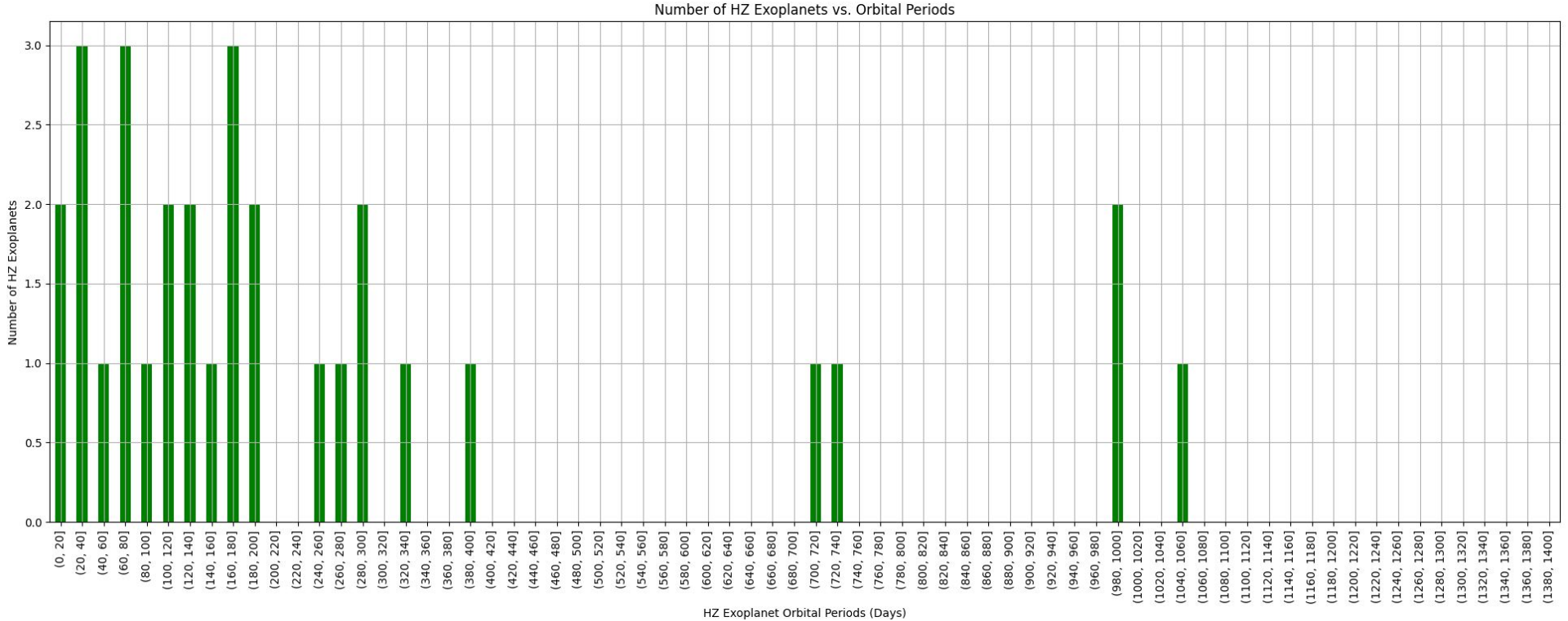
# Planet Orbital Period (Days) – Number of Exoplanets



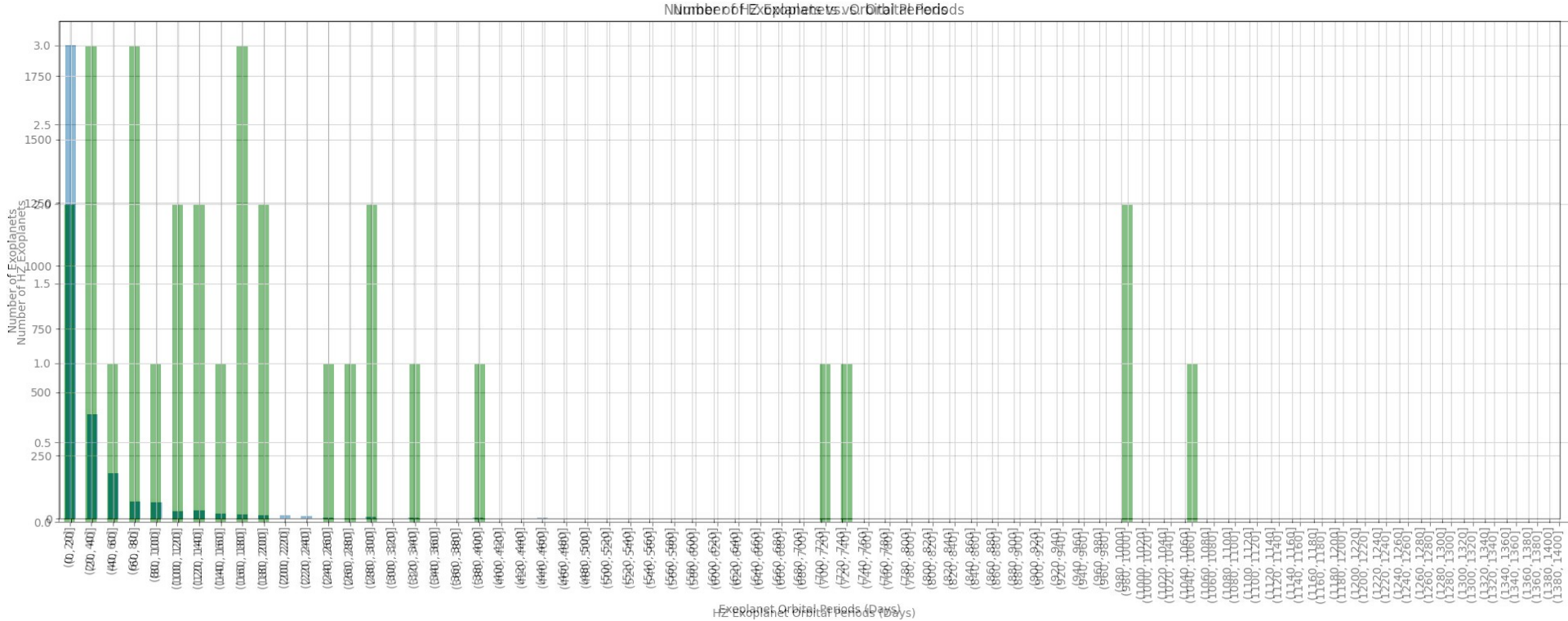
# Planet Orbital Period (Days) – Number of Exoplanets (log Scale)



# Planet Orbital Period (Days) – Number of HZ Exoplanets



# Planet Orbital Period (Days)



# Future Possibilities

- look further into stellar age and orbital period (days) and their effects on exoplanet habitability

# Future Possibilities

- look further into stellar age and orbital period (days) and their effects on exoplanet habitability
- experiment/explore with machine learning models with a bigger dataset

# Future Possibilities

- look further into stellar age and orbital period (days) and their effects on exoplanet habitability
- experiment/explore with machine learning models with a bigger dataset

thank you!