Marea O’Connor

Brooke Lipton

Astron 98 Final Project: Analyzing the Distribution of Spectral Types as Planetary Host Stars

Introduction:

In this project, we analyze the distribution of spectral types as planetary hosts. We will use skills such as data analysis, generating, filtering and fitting to create a model fit. We will explain each step to clarify how we sort a large data set into an accurate model fit to describe the distribution of spectral types.

Choose Phenomenon and Data Source:

The phenomenon we chose for our project was host stars and their spectral types/temperatures. To find this we used data from the NASA Exoplanet Archive, Working with Stellar Hosts. This data contains information about host stars and their stellar effective temperatures (Kelvins) and can be accessed here: exoplanetarchive.stellarhosts

Data Filtering:

After importing our data into jupyter notebook, we filtered our data to make sure that our project produced a reliable and understandable result. Here are the filters we will use:

1. **Dropping Non-Existent Data:**
   We removed any rows with NaN as the Host Star’s stellar effective temperature. This cleaned up our data frame by reducing the amount of rows from 44,395 to 36,348.
2. **Removing outliers:**
   Furthering the filtering of our data, we attempted to remove data points that are significantly out of the typical range through utilizing the interquartile range of said data.

After performing our code, we noticed that no data points were deemed to be outliers, so we continued on with the filtering process.

3. **Averaging Different Temperatures of the Same Host Star:**
   Our last step in filtering our data was to combine all the different temperatures taken at different times for a single host star into one. To do this we took the mean of all temperatures listed under the same hostname. This reduced the data frame down from 36,348 to 4,163 rows.

Equation to Fit Data:

To analyze the distribution of spectral types of our host stars we wanted to compare our data to commonly used models in astronomy research. The model of best fit will be decided based on the similarity of our graph to the models listed below:

Power-law Model: $P(x) = Ax^{-\alpha}$

The power law describes a relationship between two quantities where a change in one of the quantities results in a change in the other quantity that is proportional to the power of the change.

- $P(x)$ represents the probability of spectral types in a parameter. In our research, this parameter is temperatures from 0-40000° Kelvin
- A is the normalization constant
- α is the power-law exponent which determines the shape of the graph
- x is the given temperature ranging from 0-40000°Kelvin

Exponential Model: $P(x) = Ae^{\lambda x}$

The exponential model describes a relationship between two quantities where there is an exponential decrease in one of the quantities and the other one increases.

- $P(x)$ represents the probability of spectral types in a parameter. In our research, this parameter is temperatures from 0-40000° Kelvin
- A is the normalization constant

- e is Euler's number; an irrational number with the first few digits being 2.71828
- λ is the decay constant
- x is the given temperature ranging from 0-40000°Kelvin

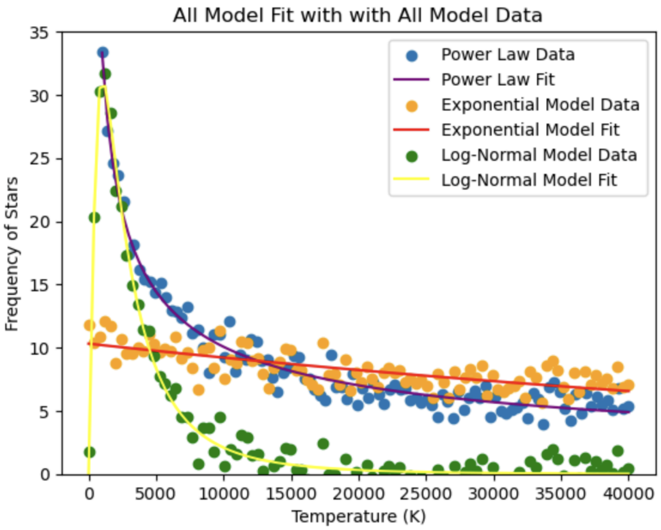
Log-normal Model: $P(x) \ = \ (\frac{1}{x\sigma\sqrt{2\Pi}})e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}}$

The log-normal model describes a situation in which a random variable's logarithm is normally distributed.

- $P(x)$ represents the probability of spectral types in a parameter. In our research, this parameter is temperatures from 0-40000° Kelvin
- e is Euler's number; an irrational number with the first few digits being 2.71828
- x is the given temperature ranging from 0-40000°Kelvin
- μ is the mean of the distribution
- σ is the standard deviation

Random Data Generation:

We generated random data points to test the accuracy of our different models. This test data will be created using Python's numpy library to imitate the real values of temperatures of host stars between 0 and 40000° Kelvin.

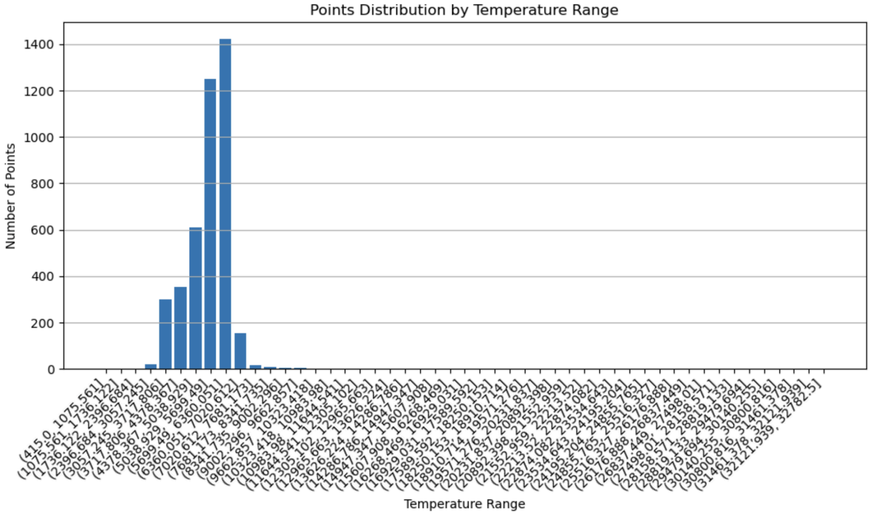


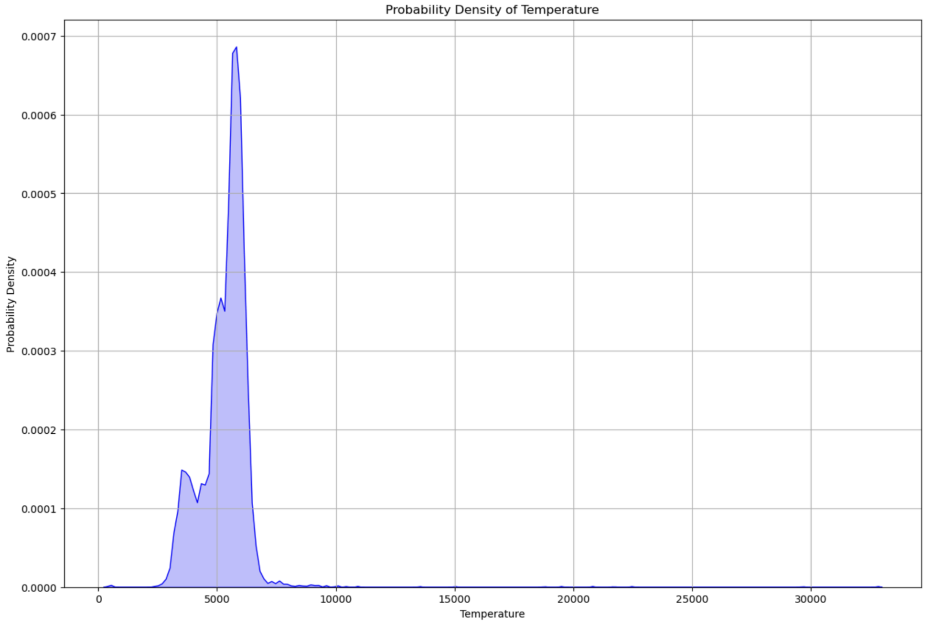
Data Fitting:

To analyze the distribution of spectral types of host stars we will fit the data using a variety of mathematical models as well as consider the errors associated with them.

1. **Model Selection:**

   The mathematical model we selected was a temperature versus frequency graph illustrating which spectral types are more abundant as host stars. We then converted it to a PDF, so we could visualize the probability of certain temperature stars hosting an exo-planetary system.

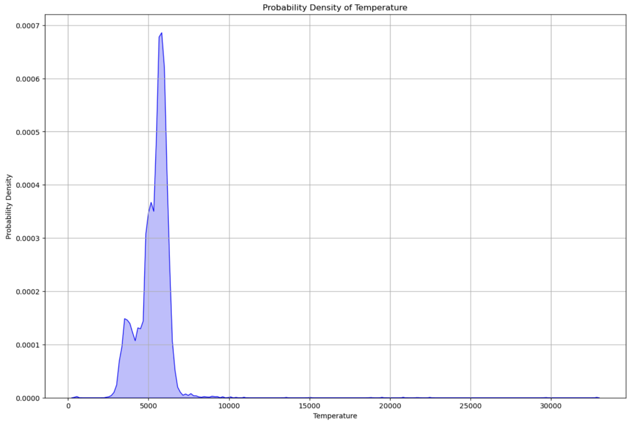


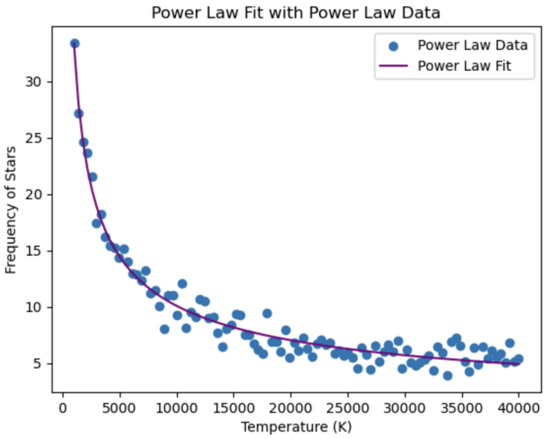


2. **Model Fitting:**

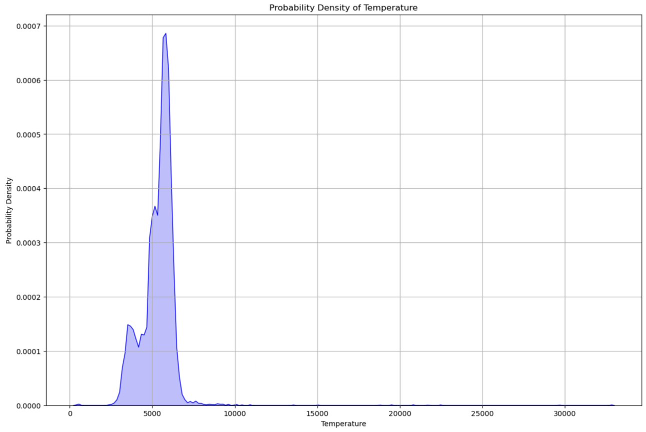
   To fit our data with a particular equation, we did side by side comparisons with each of their models (power law, exponential, and log normal) and our PDF graph in order to visualize which one would be the most reliable in predicting temperature distributions of host stars.

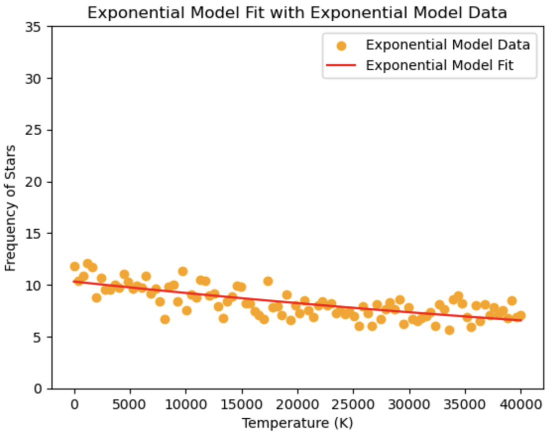
Power Law :



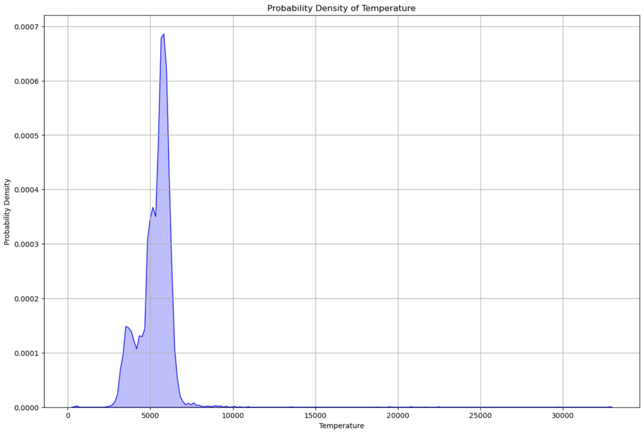


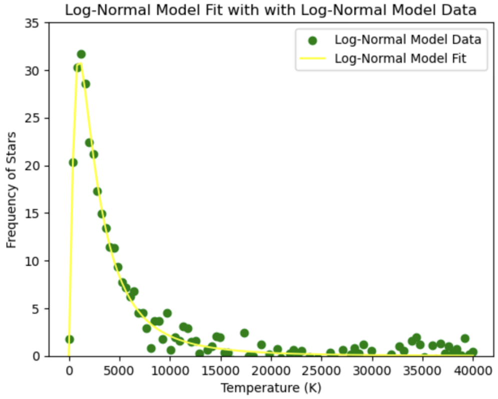
Exponential Model :





Log-Normal Model :





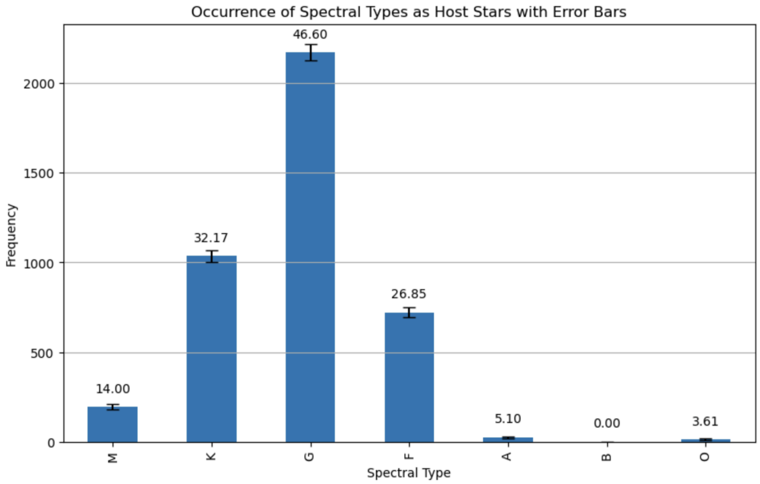
3. **Explanation of model fit:**

   From comparing our data to the commonly used models it is obvious that the model best fit is a log-normal model. The log normal model best fits situations where there is a peak. There is a peak in our data around 5,000-6,000° Kelvins which is similar to the peak seen in the log-normal model. Therefore, the log-normal model is the most accurate model to fit our data!

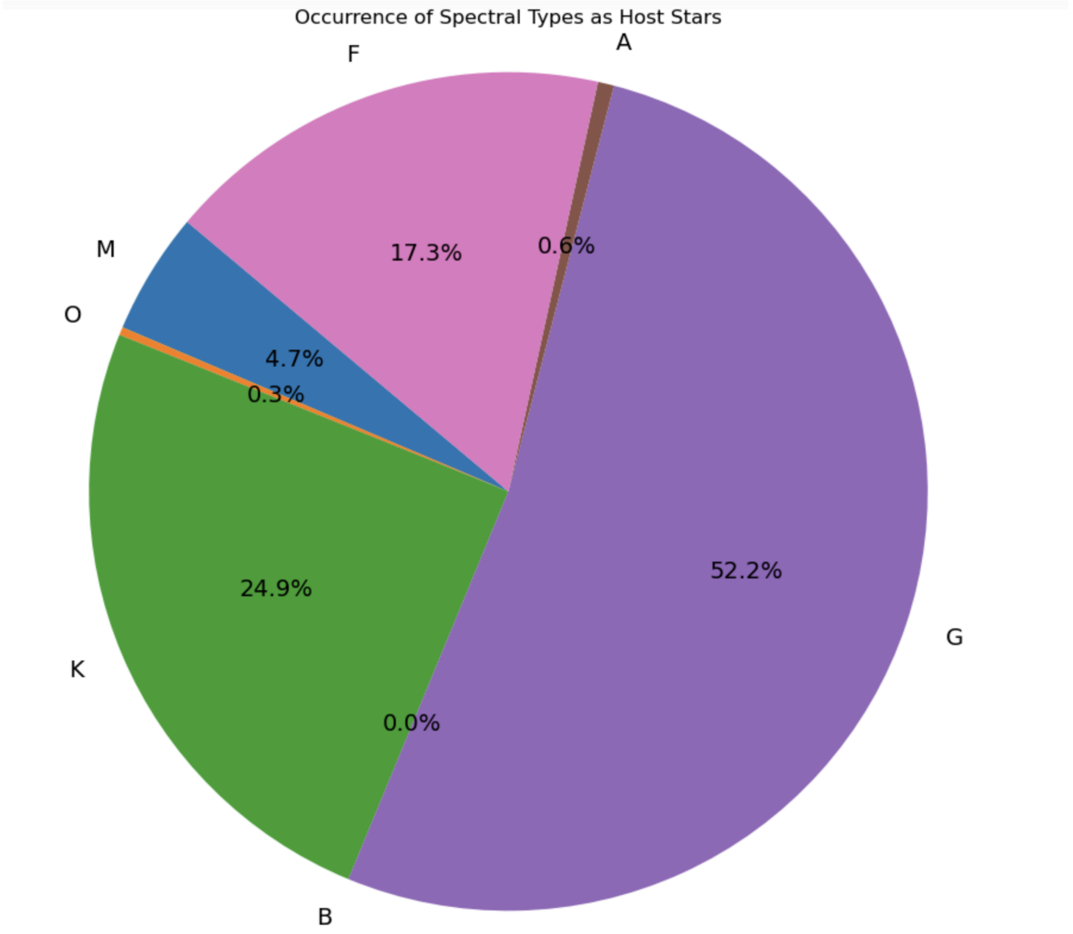
Conclusion:

Through the process of data fitting, we can see that the log-normal model is the most accurate in representing our data. There are slight differences between the model and our actual data representation, but overall we can see that most host stars are around 5,000 degrees Kelvin.

To wrap up our project we wanted to categorize the temperatures of the host stars to their associated spectral type. When we did this, we created another bar graph with frequency of occurrence on the y-axis, but instead of temperature on the x-axis, we looked at spectral type. Through this, we got a consistent representation with our previous models, still maintaining a log-normal distribution shape.



Lastly, we created a pie chart to represent the percentage of host stars that fall into certain spectral types.



In both models utilizing the spectral type categorization, the most abundant temperatures of host stars are within the G spectral type (5,000 - 6,000). This matches our pdf perfectly, and matches the log-normal model fairly well with slight error.