Replicating the Hertzsprung-Russell Diagram

Abstract:

The goal of this project is to replicate the Hertzsprung-Russell Diagram which is a plot that graphically represents the distribution of stars based on specific characteristics. We will be using data from the Gaia database which is part of the European Space agency. By pulling data of stars' absolute magnitude, luminosity, color, and temperature, we are able to use Jupyter Notebook to create fits of the data and eventually replicate the Hertzsprung-Russell Diagram. As you will see through the rest of this paper there are many key factors to take into consideration when pulling data from the Gaia database to be able to get a plot that is not skewed. Our motivation for this project is to gain an appreciation of how one would be able to analyze data sets of millions of stars and begin to get a useful understanding of the stars' roles in our galaxy.

Pulling Data:

To be able to recreate the Hertzsprung-Russell Diagram we needed to find a large reliable data source to pull from. After searching we found the Gaia Archive which has data on nearly 2000 millions and had a recent data update, Gaia Data Release 3 (GDR3). After making an account, their servers had a convenient UI where you can write out queries in Astronomical Data Query Language (ADQL), a similar language to Structured Query Language (SQL), to pull a section of the data that is of interest to you. We were able to find the specific ADQL functions needed in their pdf files for GDR3. For our data queries, we took into consideration the specific parameters we wanted to pull for the purposes of our Hertzsprung-Russell Diagram as well as parameters that would allow for precise and accurate data since some of the data on the stars may not be too accurate due to their positions relative to us and other stars. On average we found that a query for about 10,000 - 100,000 stars gave us enough data for our fits. For our

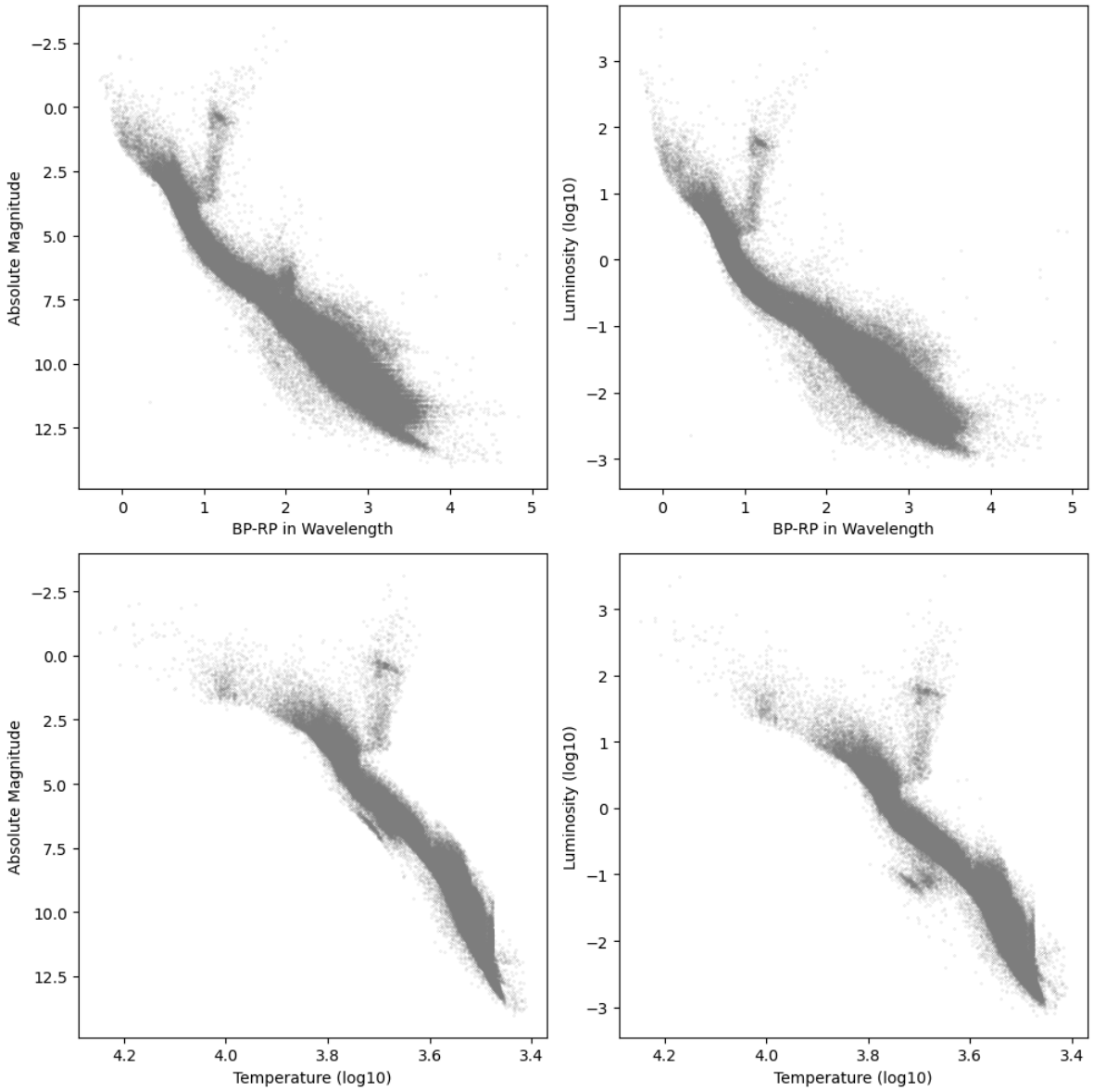
Hertzsprung-Russell Diagram, we were pulling data on the absolute magnitude, luminosity, color, and temperature of each star so we had to make sure that each of those values were not null. In addition, after reading through descriptions and papers of GDR3 we were able to get a better understanding of which stars would give us accurate data. When running our queries we made sure that the parallax over parallax error was greater than or equal to 5, the distance of the star is less than 100 parsecs, and that the star is more than 20 degrees above the galactic plane. The reason we chose these parameters is because it means the data has high confidence in its measurements, the distance is close enough to get accurate data and there are no stars near the start of interest that could affect the accuracy of the data, respectively. To gain a better understanding of what our queries looked like here is an example:

```
select distinct source.source_id, source.bp_rp, params.mg_gspphot,
params.classprob_dsc_combmod_whitedwarf,
params.classprob_dsc_combmod_star
from gaiadr3.gaia_source as source, gaiadr3.astrophysical_parameters as
params
where source.parallax/source.parallax_error >= 5 and
source.distance_gspphot < 100 and
source.b > 20 and
source.bp_rp is not null and
params.mg_gspphot is not null
```

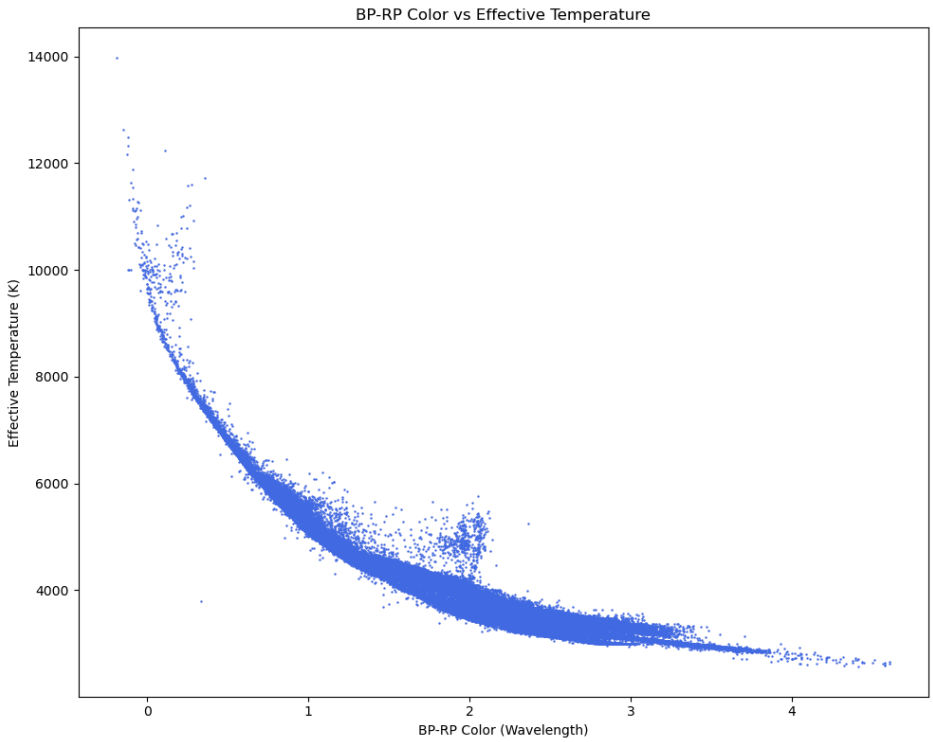
In this query we are pulling each star based on its unique source id and making sure that the data we want is not null for each star and that the stars have accurate data based on the parameters we have laid out previously.

Visualizing the Data:

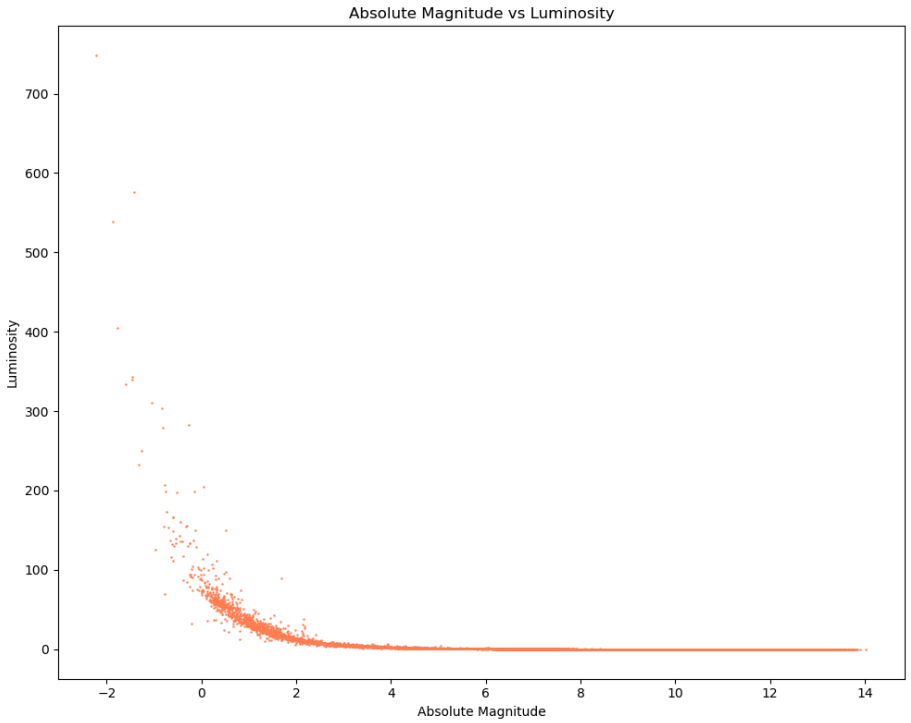
With the data gathered, we planned on plotting the information in a two-dimensional histogram with BP-RP color magnitude and Effective Surface Temperature against Absolute Magnitude and Stellar Luminosity in a quad-axis plot, as is practice in most HR Diagrams. However, the first trouble that came with the visualization of our data was the inconsistency in plotting different axes against each other. Below is one of our rudimentary tests in attempting to plot multiple x-axes against multiple y-axes for a dataset of 110,000 stars.



While the graphs are related, there exists slight variations between each pair of graphs, so we sought to remedy these variations. The final decision was a compromise: we would plot along only two axes, but to include the other data we would develop an equation that would estimate the relationship between pairs of axes. We decided to plot BP-RP magnitude against Absolute Magnitude, and therefore would need two equations to describe how BP-RP color is related to Surface Temperature and how Absolute Magnitude is related to Stellar Luminosity.

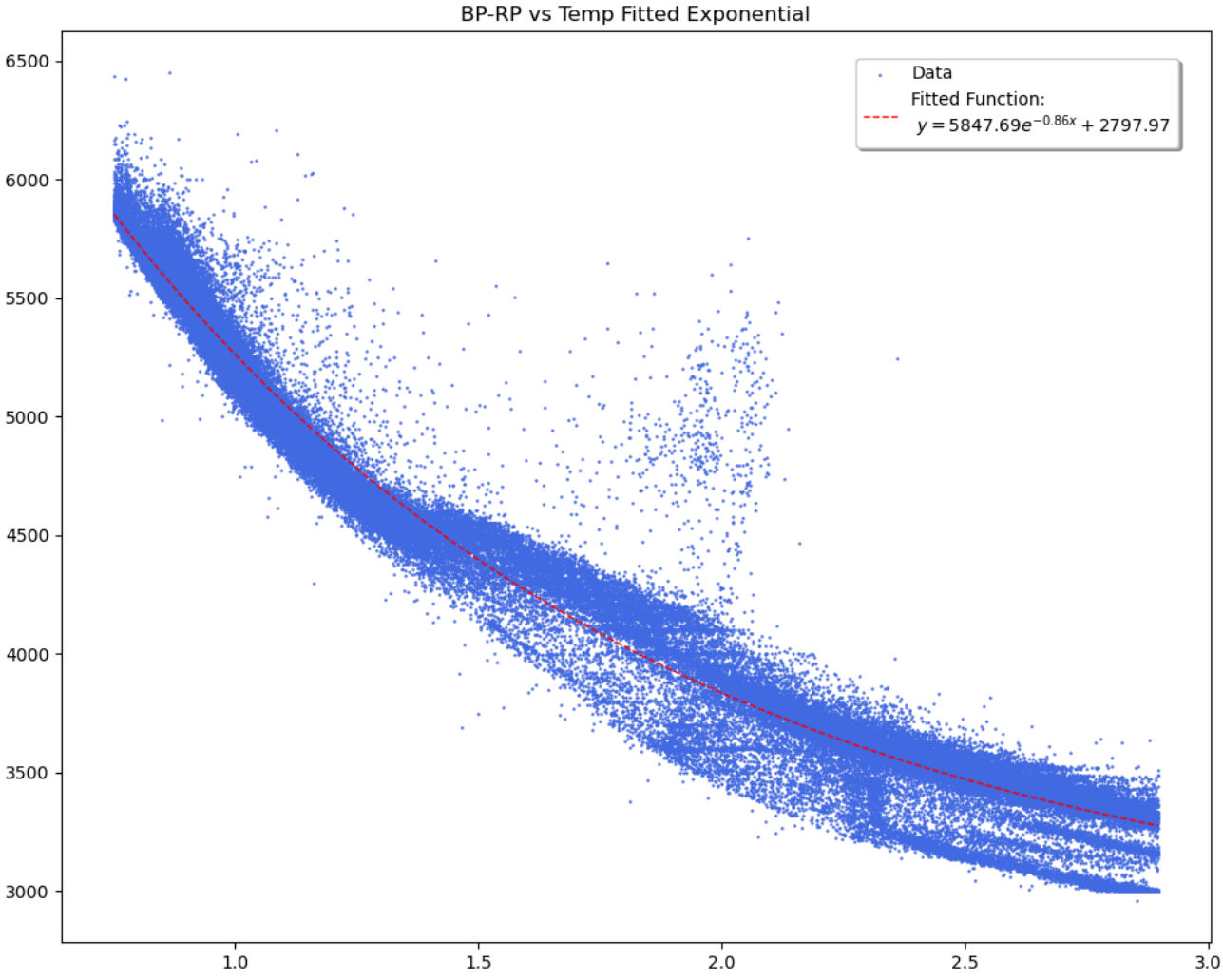


The best way to achieve these equations is to plot the data and determine a line of best fit. To the right are the plots for BP-RP vs Temperature and Absolute Magnitude vs Luminosity. We can clearly determine that there exists a correlation. For BP-RP vs Temperature, we decided to eliminate the first and last 10% of points because we

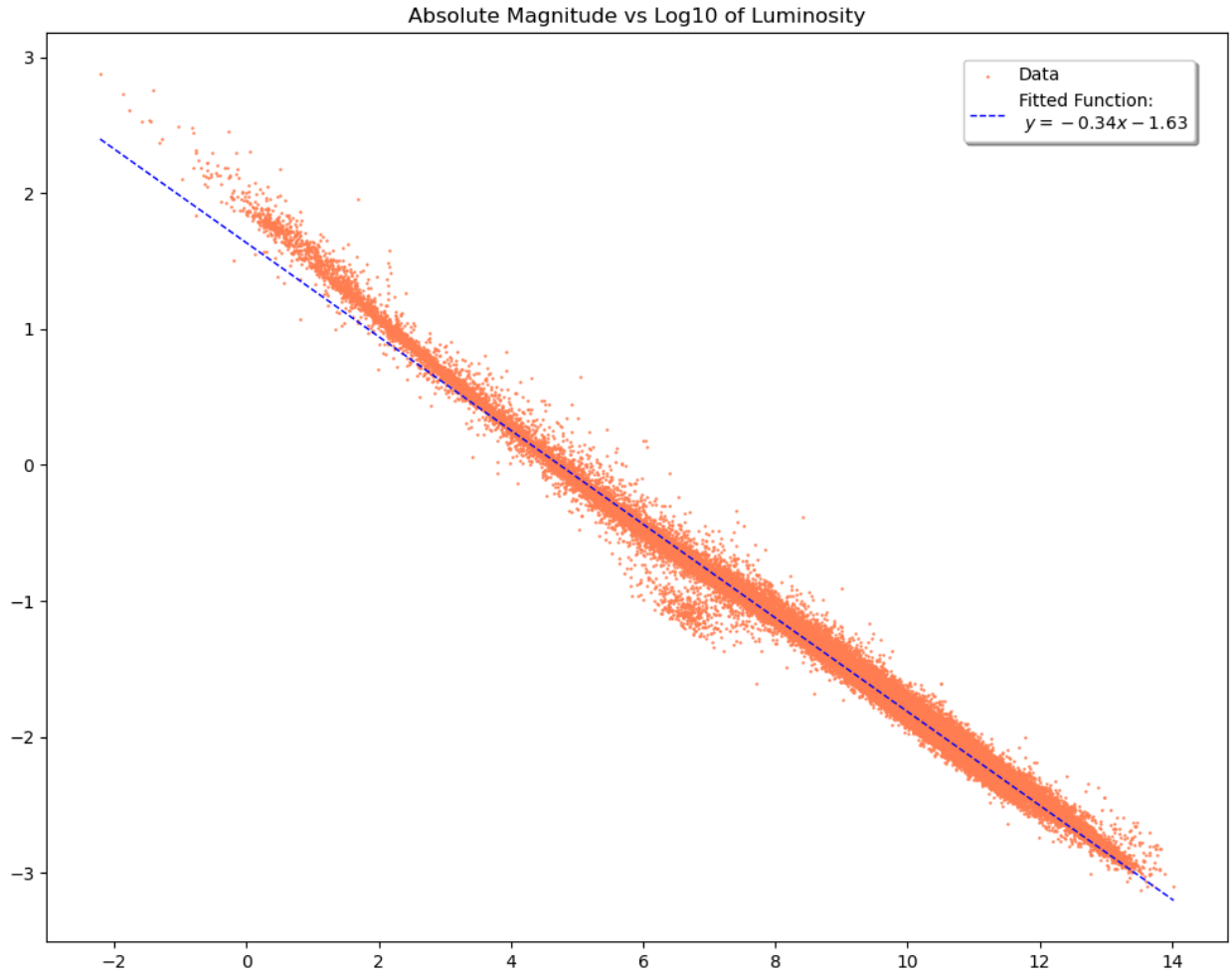


noticed a strong density of points near the center of the plot. We also noticed the relationship is of exponential decay. Here are the results of fitting an equation to this graph. The resulting equation was as follows:

$$y = 5847.69e^{-0.86x} + 2797.97$$

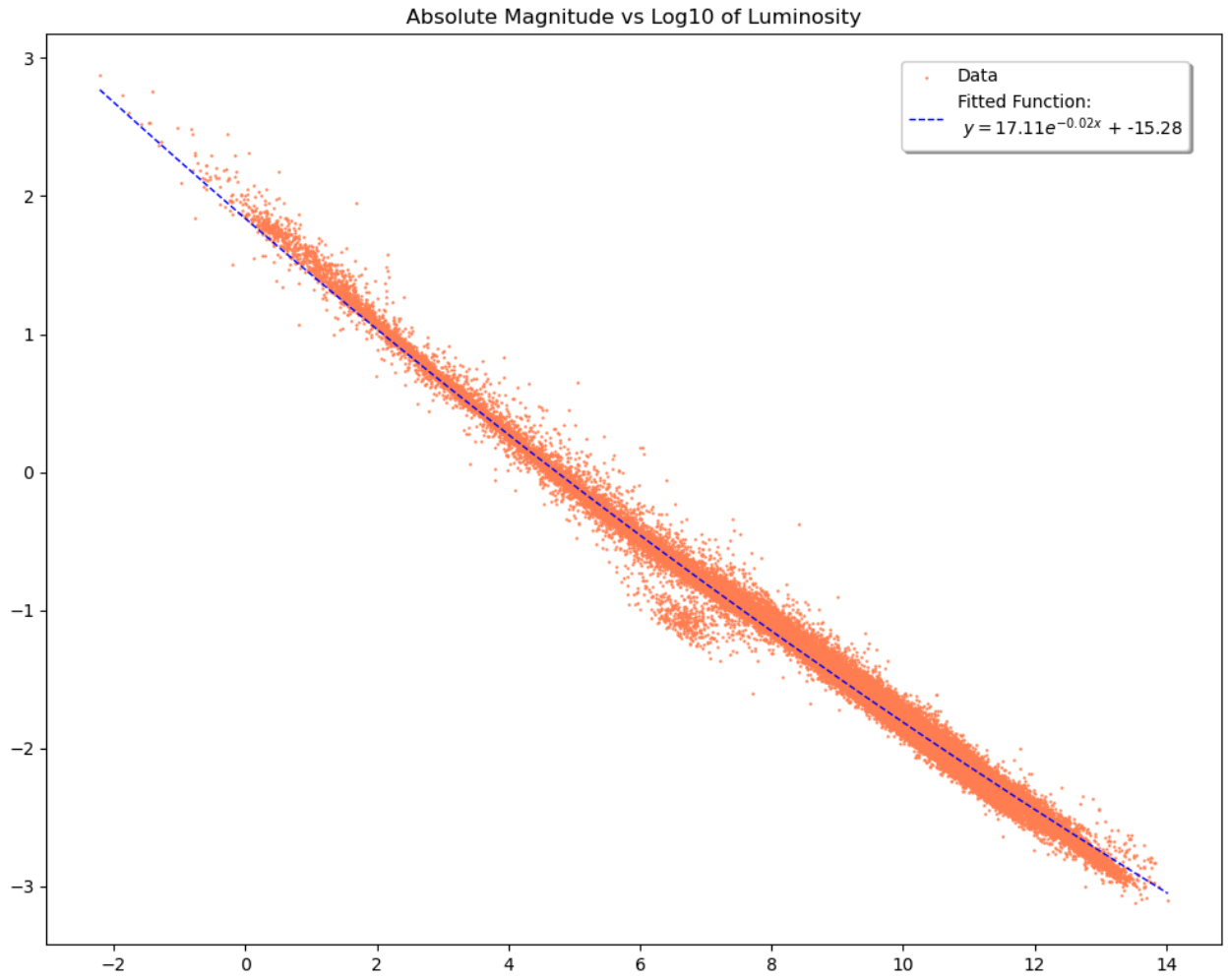


We proceeded to do the same with the Absolute Magnitude vs Luminosity plot. However, HR diagrams often relate the logarithm of Luminosity to Absolute Magnitude. Therefore, we first decided to plot a linear relationship after taking the logarithm.

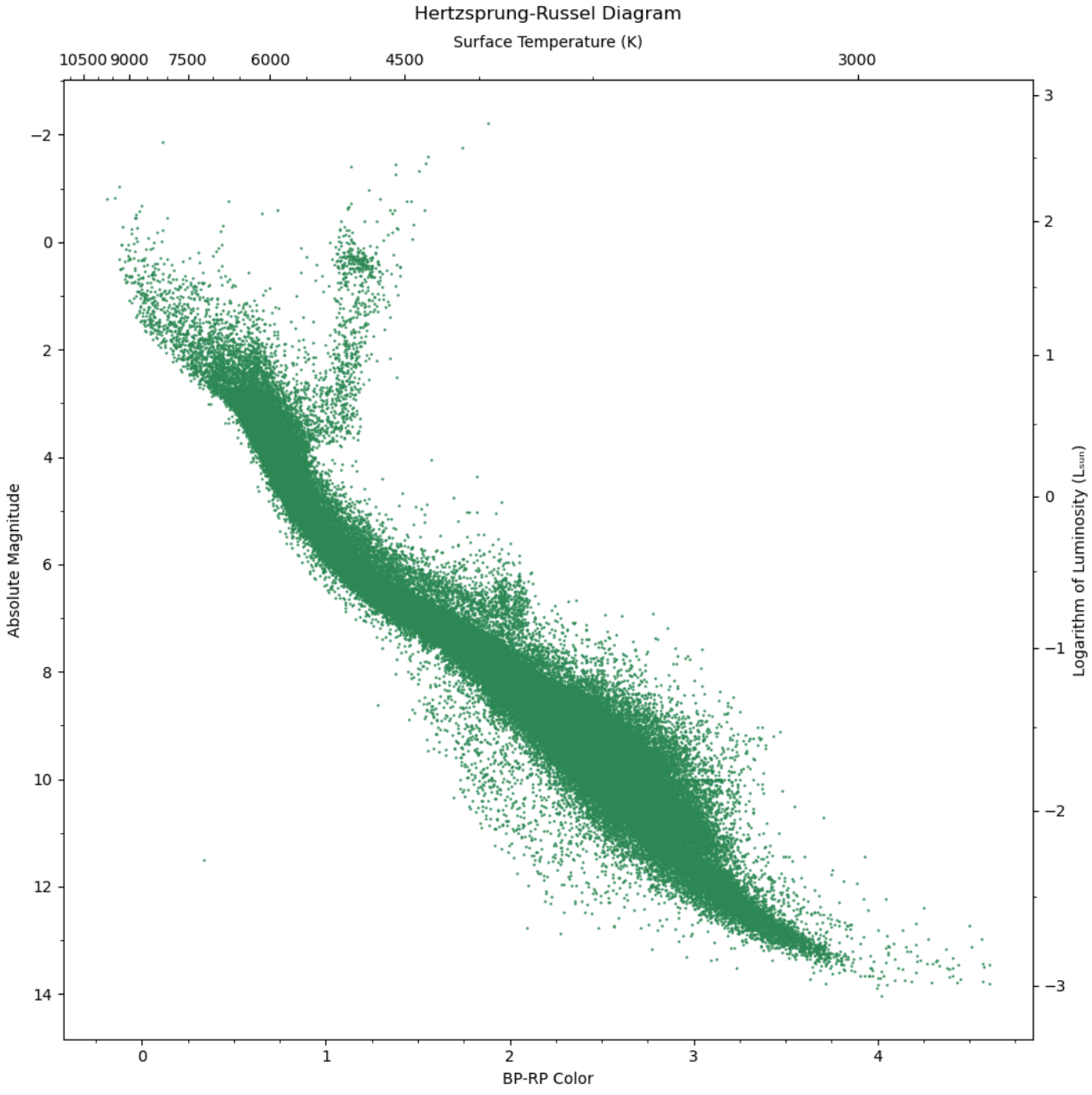


While the relationship may look linear, we actually see the line diverge near values of low Absolute Magnitude and high Luminosity. Therefore, we determined that the relationship was actually slightly exponential, even after taking the logarithm. Here is that relationship plotted, and the resulting line of best fit:

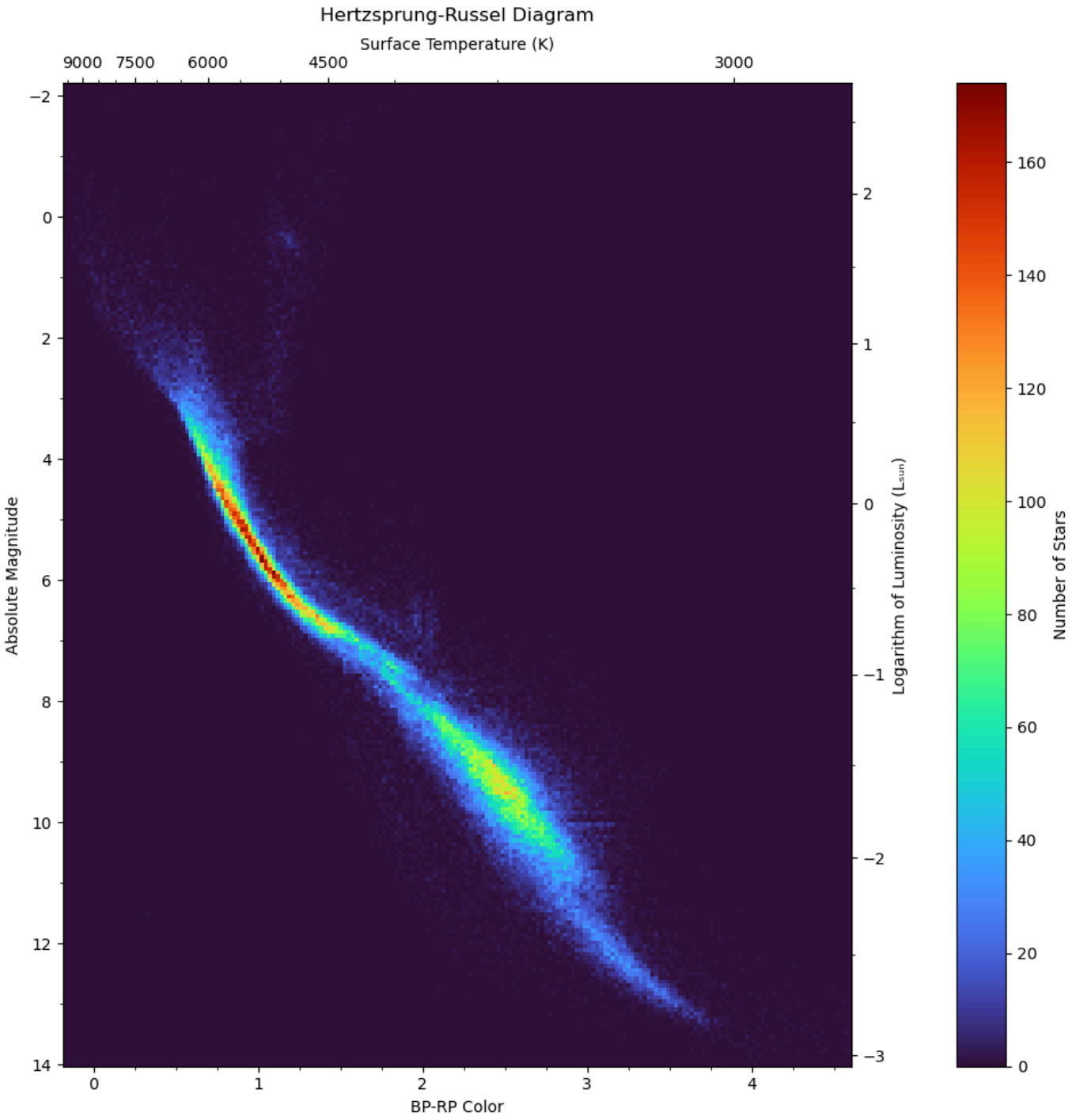
$$y = 17.11e^{-0.02x} - 15.28$$



Now with these relationships in hand, we can plot the data on all four axes. Here is the visualization of the BP-RP color and Absolute Magnitude of 110,000 stars.



Although we can see that there exists a relationship between all of this data, this form of visualization tells very little about the density of stars in certain locations. Therefore, we decided to replicate this plot but instead include a visualization of that density of stars in the form of a two-dimensional histogram plot. Below is that visualization.



Lastly, we noticed that with the stars mostly located near the center of the plot follow a slightly cubic shape. With this equation, we would be able to estimate the Absolute Magnitude, Surface Temperature, and Stellar Luminosity of a star given only just its BP-RP color magnitude. Therefore, by again eliminating the outside 10% of the data set, we arrive at this equation that fits a cubic relationship to the HR Diagram:

$$y = 0.50x^3 - 3.03x^2 - 0.82x - 0.82$$

Interpreting our Result:

While we achieved our end goal of determining an equation to estimate information about stars, our visualization is lacking a few key characteristics found in other Hertzsprung-Russell Diagrams.

Firstly, after months of querying and tweaking conditions and parameters, we were unable to reliably find certain sets of stars that fall into other categories on the HR diagram besides the main branch – notably, we were unable to replicate the band of white dwarfs that often sits in the low BP-RP color and high Absolute Magnitude portion of the diagram. Clearly, we were able to find an abundance of stellar spectra that landed in the "main branch" portion of the diagram, and even some stars in the "giant branch", but very few of these stars were classified as white dwarf stars. We hypothesize that this may be a result of our restrictions during the querying process, but when querying times were growing longer and a deadline for this project was approaching, it became clear that we would have to work with the data that we had already acquired.

Secondly, it should be noted that this paper barely scratches the surface on the sheer number of visualizations we made prior to this final result. Located in the "various tests" and "gaia queries" directories in our GitHub submission are dozens of tests and failures that document our process towards this final result. We also spent a large portion of our time scouring catalogs upon catalogs within the Gaia Archive in order to prune our data for appropriate visualizations.

However, with the limitations and obstacles that we encountered, we were still able to develop a stunning representation of the lifecycle of a star in the form of the Hertzsprung-Russell Diagram and we were able to acquire an equation to estimate various characteristics of a star given an initial condition.

Sources:

1. https://www.esa.int/ESA_Multimedia/Images/2018/04/Gaia_s_Hertzsprung-Russell_diagram
2. https://astronomy.stackexchange.com/questions/39610/is-there-a-formula-for-absolute-magnitude-that-does-not-contain-an-apparent-magn
3. https://academic.oup.com/mnras/article/508/3/3877/6373953
4. http://csep10.phys.utk.edu/OJTA2dev/ojta/c2c/ordinary_stars/magnitudes/absolute_tl.html
5. https://www.gaia.ac.uk/mission/blue-and-red-photometers
6. https://www.phys.ksu.edu/personal/wysin/astro/magnitudes.html
7. https://gea.esac.esa.int/archive/
8. https://gea.esac.esa.int/archive/documentation/GDR3/Gaia_archive/chap_datamodel/sec_dm_main_source_catalogue/ssec_dm_gaia_source.html
9. https://gea.esac.esa.int/archive/documentation/GDR3/Gaia_archive/chap_datamodel/sec_dm_astrophysical_parameter_tables/ssec_dm_astrophysical_parameters.html
10. https://www.cosmos.esa.int/web/gaia/iow_20220609
11. https://python-graph-gallery.com/83-basic-2d-histograms-with-matplotlib/
12. https://www.omnicalculator.com/physics/luminosity
13. https://matplotlib.org/3.4.3/gallery/ticks_and_spines/major_minor_demo.html
14. https://peps.python.org/pep-3101/