Part 1 - Background

The dataset that we used for this assignment is a stellar dataset. It contains information about several stars and their characteristics, such as temperature, luminosity, radius, absolute magnitude, star type, color, and spectral class. This dataset was initially taken from Kaggle [1]. Since the data is already in CSV format, it was easy for us to perform an analysis of this dataset.

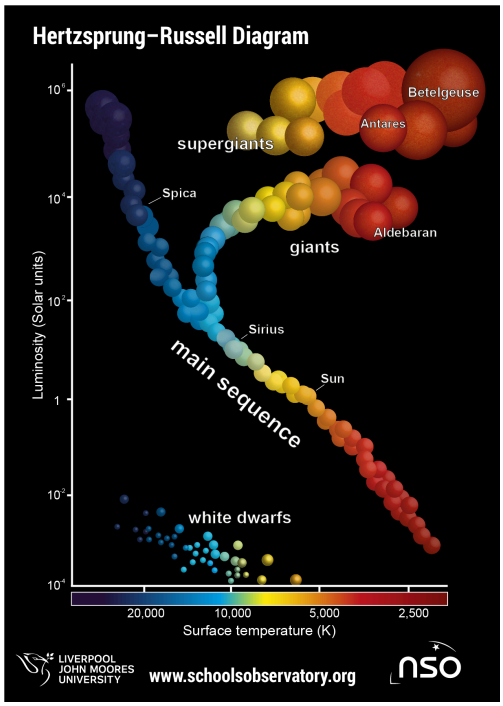
This dataset is interesting because it is detailed and contains many crucial attributes of stars. These traits are crucial for understanding how stars behave and evolve. Analyzing the data and grouping stars helps us learn more about the different stages of a star's life. Moreover, this dataset will help us find connections and groups between stars. It will help us see how they are related or different from each other. We will use this dataset and perform clustering analysis for this assignment to find the different star groups. We will also try to analyze the data and find interesting associations between the color and spectral class of stars using the dataset's categorical attributes.

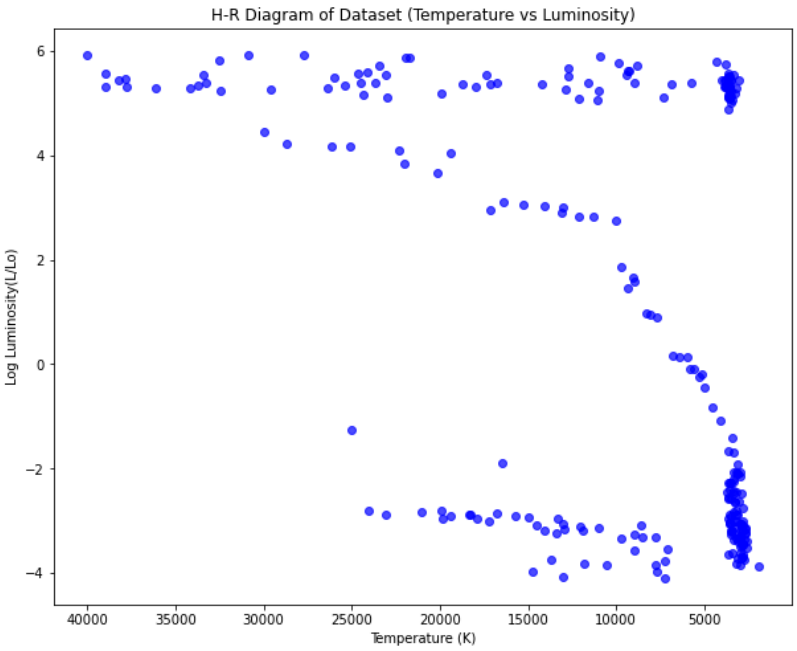
The Hertzsprung-Russell (H-R) diagram is fundamental in astronomy, mainly because we can use it to plot stars based on their luminosity and temperature [2]. The diagram groups stars into four categories: Supergiants, Giants, Main Sequence, and Dwarfs. For the cluster analysis part of this assignment, we will plot the dataset in a way that is similar to the H-R diagram. Then, we will use one of the clustering methods learned in CPS844 to group the data, like how stars are grouped in an H-R diagram. For the association analysis part of this assignment, we will use one of the association methods learned in class to find association rules for star type, spectral class, and star color.

Part 2 - Methods

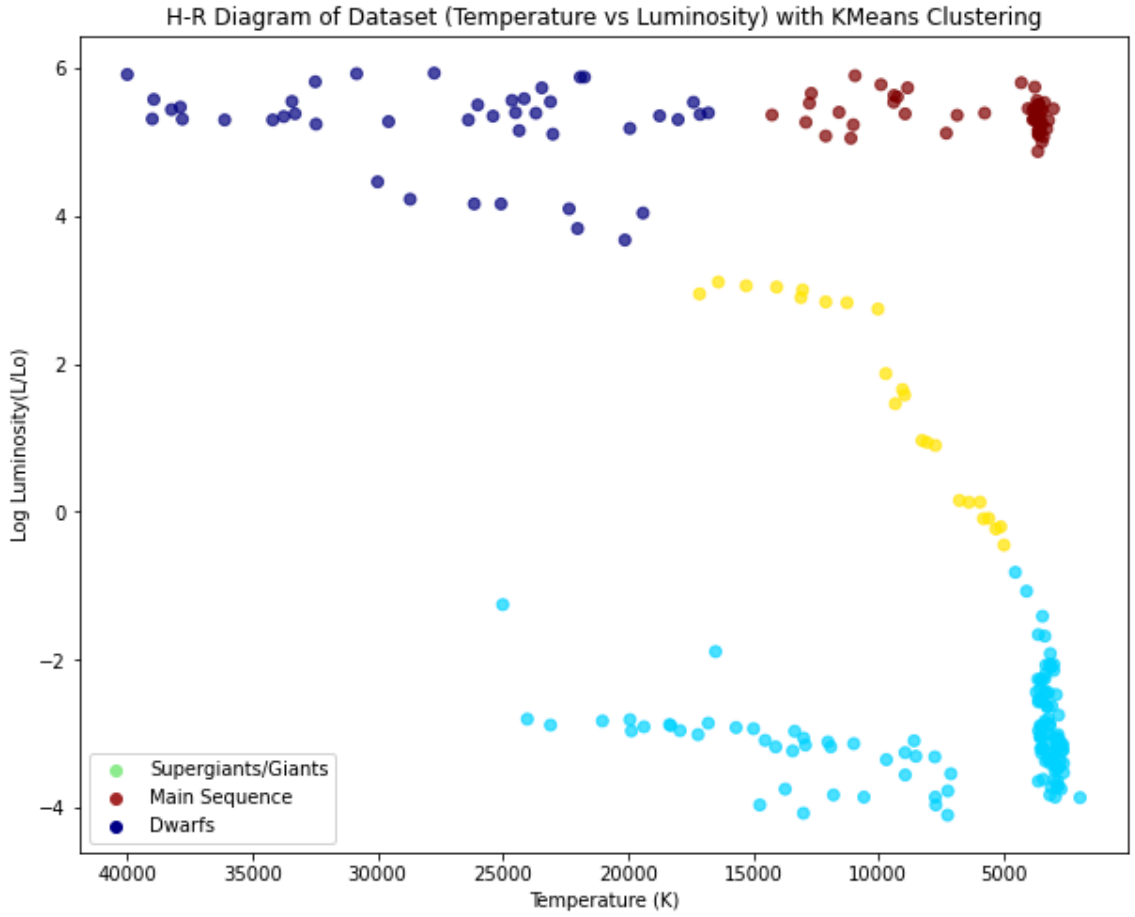
Cluster Analysis

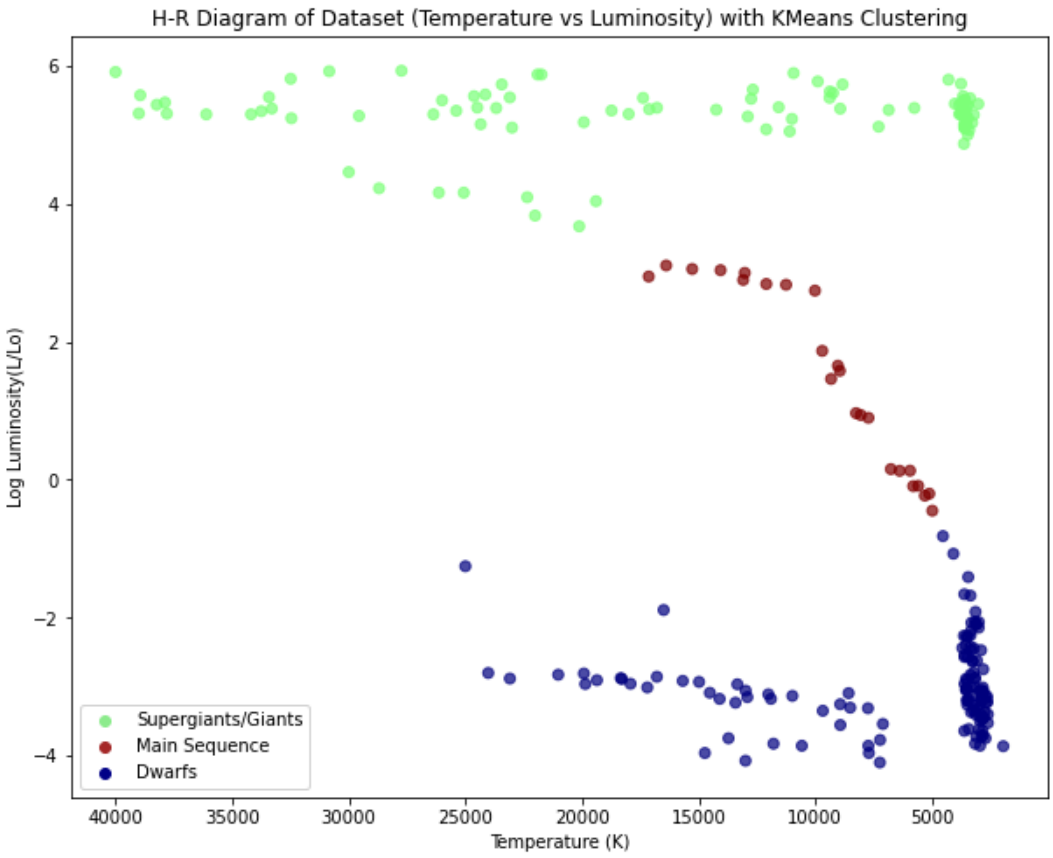
The first part of the assignment was on clustering analysis on the star dataset. Before we performed the clustering, we wanted to plot the dataset first. We decided to plot the data similarly to an H-R diagram. We also decided to use the numerical attributes, particularly luminosity and temperature, as these two are x and y values in an H-R diagram. As described earlier, an H-R diagram uses these two attributes of a star to group them into different star types. Before we began working with the data, we had to preprocess it. We dropped all the remaining ones except for luminosity and temperature attributes as they are not crucial for the cluster analysis. We also transformed the luminosity values using logarithmic transformation to show better the difference between how they are shown in an H-R diagram. The temperature attribute was also standardized using the StandardScaler to make sure both features were measured on the same scale. Before analyzing, this step helps ensure each feature is equally important for the clustering process by making them comparable.





We decided to use the K-Means clustering algorithm to group the stars into clusters. This algorithm divides the data into K cluster numbers while ensuring minor variations within the clusters. We wanted to group the stars similarly to how it is done in the H-R diagram. The H-R diagram groups the stars into four clusters: Supergiants, Giants, Main Sequence, and Dwarfs. We initially tried a k value of 4 in our code to group the instances in our dataset into four groups. However, after trying a k value of 4, we saw that the clustering results were inaccurate. Using the k value of 4, our implementation divided the supergiants group into two separate groups, which was unrealistic. So, we decided to play around with the k value until we found a clustering that made sense. By trying a k value of 3, we got a clustering that somewhat resembled the clustering of an actual H-R diagram. The clusters were then shown on our diagram, revealing different groups that match up with types of stars like supergiants/giants, main sequence, and dwarfs.





Association Analysis

We decided to use only a few selected attributes for the association analysis part of the assignment. We only selected the categorical (non-numerical) attributes from the stellar dataset. For this analysis, we looked at categorical variables like star color, spectral class, and star type. Before diving in, we made sure the data was preprocessed to ensure there were no errors in our analysis results. So first, we dropped all the numerical attributes from the dataset except for star type. The dataset that we downloaded from Kaggle had an attribute star type. However, that attribute star type was discretized by the dataset's author. The star types, such as supergiants, giants, etc, were mapped to numerical values. So, we looked at the author's description and mapped the instances' star-type attributes back to categorical values. Next, we changed all the star color attributes to lowercase values to ensure consistency between values. Similarly, since

most of the values used a dash, we had to replace some missing a '-' between the color names, such as 'blue-white' and 'blue-white.'

Now that our data was preprocessed for analysis, we began implementing our association analysis. We decided to use the Apriori algorithm, which is found in the apyori library. This is the algorithm we used in class and the labs of CPS844. However, since this algorithm uses a list of lists, we had to do one additional step of preprocessing, which is to store all the data values we had for this dataset into a list of lists such as the following: ['Brown Dwarf', 'red', 'M']. Initially, we set a confidence value of 70% as we are only interested in rules often found to be true. For the support value, however, we decided to start with a value of 5%, which we know we had to adjust later depending on the results we found. We ended up with many rules that were not correct. For example, a few of the rules we found were {Hypergiant, M} -> {red} and {red, Hypergiant} -> {M}, which are not accurate because stars in class M are usually the coolest and the smallest [3].

{Hypergiant, M} -> {red}	0.0875	0.954545
{red, Hypergiant} -> {M}	0.0875	0.913043

We decided to increase the support value up to 10% to make sure the rules we got made sense. The results we got with a support value of 10% were significantly better. We have several rules relating star-type dwarfs to spectral class M [3]. We also got a few rules stating that blue is associated with spectral class O, which is also true [4]. We found a unique rule that associated blue-white with spectral class B, which is also true [4]. Last but not least, we found a rule associating Supergiants with the color blue, and some of the brightest and most prominent stars in the galaxy are blue stars [5].

Rule	Support	Confidence
{blue-white} -> {B}	0.116667	0.7
{Brown Dwarf} -> {M}	0.166667	1
{Brown Dwarf} -> {red}	0.166667	1
{Red Dwarf} -> {M}	0.166667	1
{M} -> {red}	0.458333	0.990991
{red} -> {M}	0.458333	0.982143
{O} -> {Supergiant}	0.120833	0.725
{Supergiant} -> {O}	0.120833	0.725

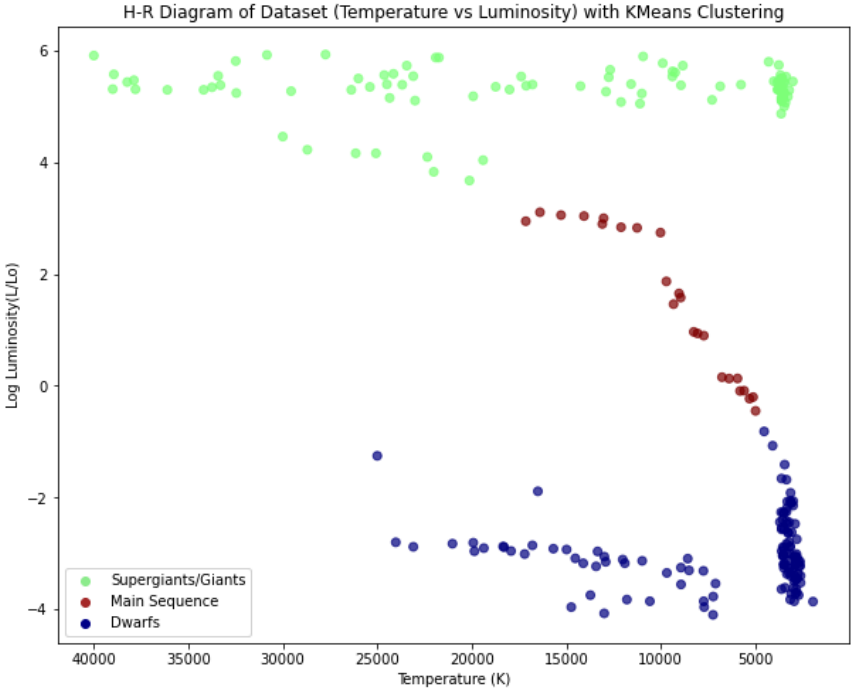
Evaluation Strategy

The evaluation of clustering effectiveness was primarily visual, utilizing the H-R diagram to assess the coherence and separation of the clusters. The evaluation focused on the support and confidence metrics for association rule mining, ensuring that the derived rules are statistically significant and reliable.

Part 3 - Results

Cluster Analysis

The steps we took before analyzing the data were necessary for the clustering analysis since we wanted to group the data, like how stars are grouped in the H-R diagram. This preprocessing ensured that the traits we used for calculations were measured fairly so that the stars were grouped into the groups of the H-R diagram. Using the K-Means clustering approach, we successfully recreated most of the groups of the H-R diagram. At first, we tried to match up with the usual way stars are classified on the H-R diagram, which has four groups: Supergiants, Giants, Main Sequence, and Dwarfs. However, after trying it out, we found that using three groups gave us more realistic results. This adjustment helped us make a more straightforward but good way of putting stars into groups: supergiants/giants, main sequence, and dwarfs. When we plotted the clusters in the diagram, the 3-means clustering closely resembled the actual H-R diagram.



Association Analysis

For the association analysis part, we looked at certain traits of stars, such as color, spectral class, and type. Before we began analyzing the data for association rules, we preprocessed the data to make sure everything was consistent. Then, we used the Apriori algorithm to find interesting connections between these traits. We adjusted the support threshold to 10% to make sure we were getting meaningful results. Some interesting things we found were that dwarf stars often have spectral class M and that blue stars are often classified as spectral class O, which matches what we know about stars. Similarly, we found many rules that accurately match what we know about stars.

Index	Rule	Support	Confidence
0	{blue-white} -> {B}	0.116667	0.7
1	{Brown Dwarf} -> {M}	0.166667	1
2	{Brown Dwarf} -> {red}	0.166667	1
3	{Red Dwarf} -> {M}	0.166667	1
4	{M} -> {red}	0.458333	0.990991
5	{red} -> {M}	0.458333	0.982143
6	{O} -> {Supergiant}	0.120833	0.725
7	{Supergiant} -> {O}	0.120833	0.725
8	{O} -> {blue}	0.166667	1
9	{blue} -> {O}	0.166667	0.727273
10	{Red Dwarf} -> {red}	0.166667	1
11	{Supergiant} -> {blue}	0.129167	0.775
12	{Brown Dwarf} -> {red, M}	0.166667	1
13	{M, Brown Dwarf} -> {red}	0.166667	1
14	{red, Brown Dwarf} -> {M}	0.166667	1
15	{Red Dwarf} -> {red, M}	0.166667	1
16	{M, Red Dwarf} -> {red}	0.166667	1
17	{red, Red Dwarf} -> {M}	0.166667	1
18	{O} -> {blue, Supergiant}	0.120833	0.725
19	{Supergiant} -> {blue, O}	0.120833	0.725
20	{Supergiant, O} -> {blue}	0.120833	1
21	{blue, O} -> {Supergiant}	0.120833	0.725
22	{blue, Supergiant} -> {O}	0.120833	0.935484

Part 4 - Conclusion

With the help of K-Means clustering, we could mimic the groups similar to what we see in the traditional H-R diagram. Although we did not get an accurate copy of the clusters in the H-R diagram, we got a result that closely resembles the H-R diagram clusters. With association rule mining, we discovered some interesting connections between different traits of stars, like their spectral class and color. These findings confirmed what we knew about stars and showed us how powerful algorithms can find hidden patterns in complex datasets.

Overall, this project was a good experience where we could practically apply the data mining theories and approaches that we have learned in class and laboratories. The results we got solidify the fact that we can apply data mining techniques to real-life situations and find meaningful results in the process of doing so.

References

[1] Star dataset to predict star types. (n.d.). www.kaggle.com. https://www.kaggle.com/datasets/deepu1109/star-dataset

[2] *The Schools' Observatory*. Hertzsprung-Russell Diagram | The Schools' Observatory. (n.d.). https://www.schoolsobservatory.org/learn/astro/stars/class/hrdiagram#:~:text=The%20Hertzsprung%2DRussell%20diagram%20shows,stars%20in%20clusters%20or%20galaxies.

[3] *The Schools' Observatory*. Stellar Classification | The Schools' Observatory. (n.d.). https://www.schoolsobservatory.org/learn/astro/stars/class#:~:text=The%20classes%20are%20called%20O,are%20the%20smallest%20and%20coolest.&text=If%20you%20look%20closely%20at,not%20all%20the%20same%20colour.

[4] Types of stars. Las Cumbres Observatory. (n.d.). https://lco.global/spacebook/stars/types-stars/

[5] May, A. (2022, September 26). *Blue stars: The biggest and brightest stars in the galaxy*. Space.com. https://www.space.com/blue-stars