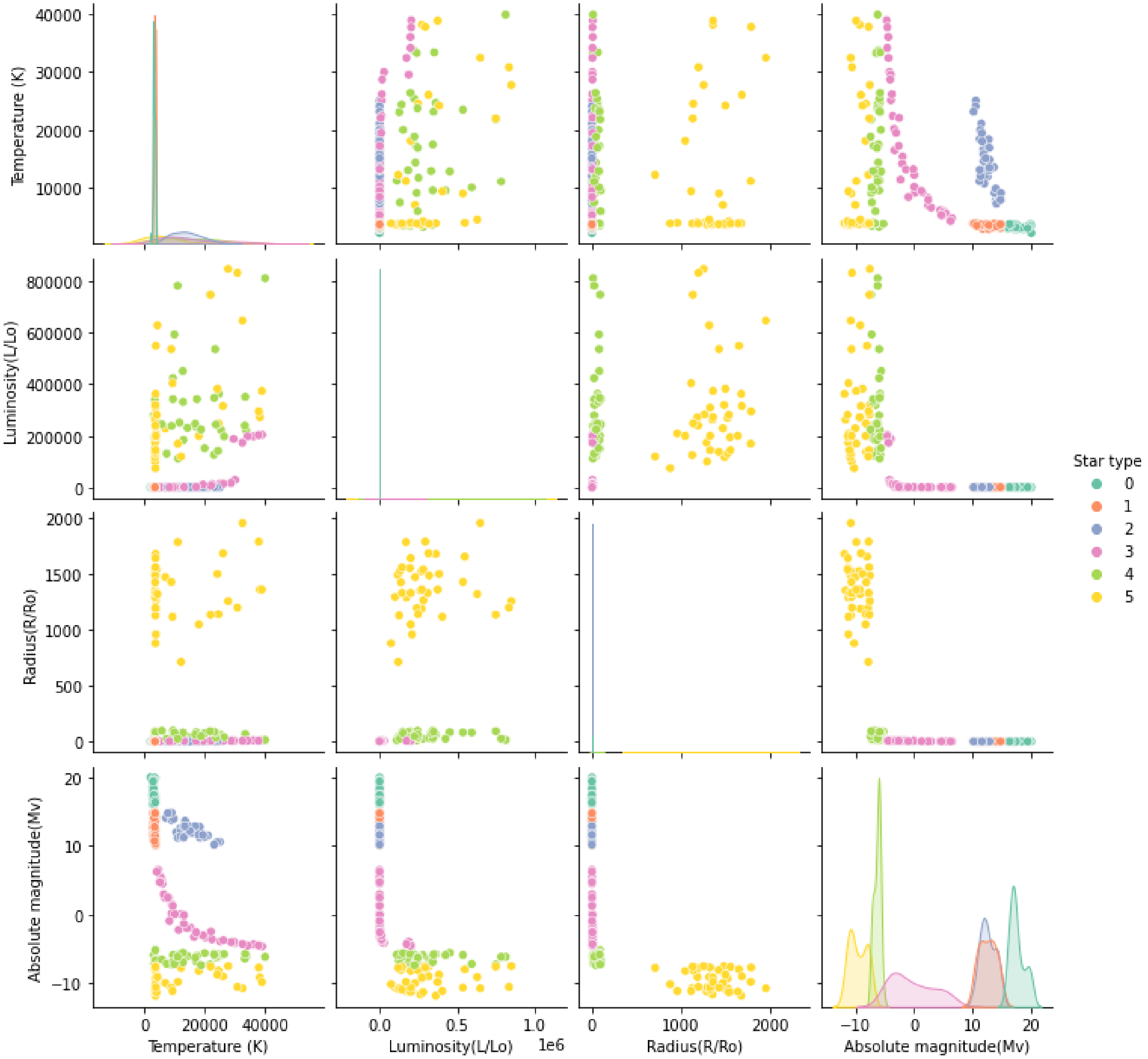
Stellar Classification of stars in Celestial Space

Akhil Chowdary Maddipatla, Deenadayalan Dasarathan, Zijie Wei

Master of Science in Data Science, Khoury College of Computer Sciences,
Northeastern University, Boston, Massachusetts
star_classification

Work InProgress:

The models that we have considered for this phase are K-means and Hierarchical clustering. Before applying the models, the bivariate EDA of the features provide several information about the dataset.

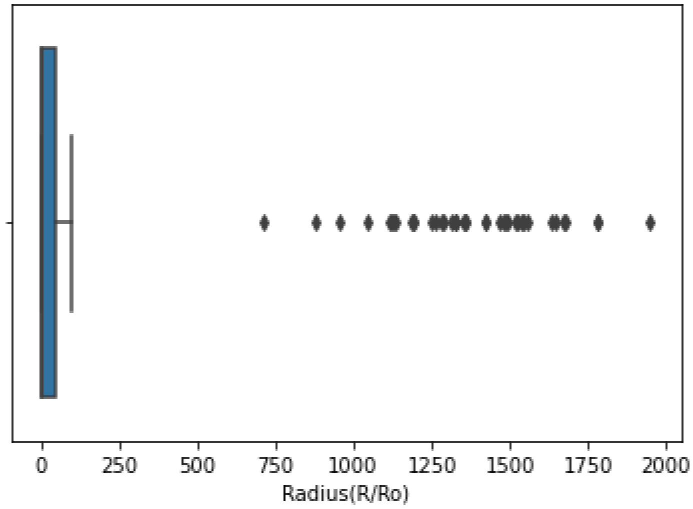


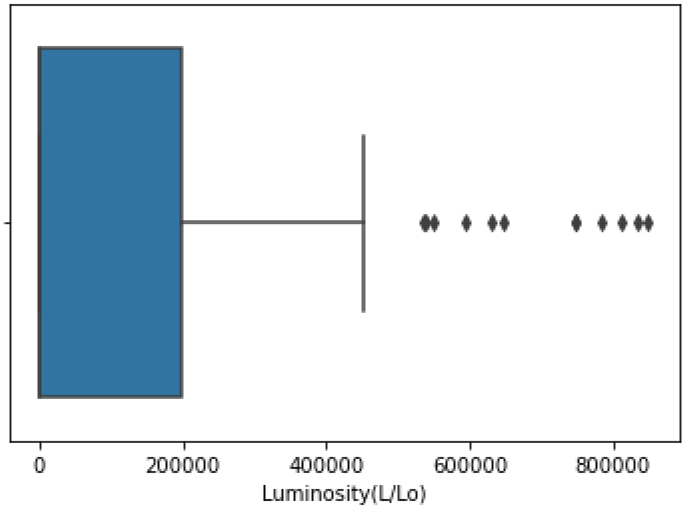
From the above pair plot, we can see start type 5, 3 and 0 are easily separable, but the other start types have much overlapping features. Therefore, we expect K-means and Hierarchical models to exhibit average performance as there are chances of misclassifying the stars. But we believe PCA before applying clustering algorithms would improve the classification results. We will discuss the performance of the models, before and after applying PCA in the final report.

New Ideas:

For the new idea, we have planned to use dimensionality reduction techniques like PCA and t-SNE before applying the clustering models and a new algorithm which is called Isolation Forest to classify the stars better.

The reason why we are using this is that we find there are some isolation observations in this dataset. For example, when we plotted the histogram of radius, we can see that most of the stars' radius are at low values. When we are doing the count of luminosity, we can see that there are very low amounts of observations that have the value of luminosity which are larger than 400000. Below are the plots.





By looking at the images above, we can see that there are some values which located in outliers. So, we are thinking about using isolation forest to isolate these values. We think this clustering algorithm may help us cluster the stars.

Scope of the work for the Final report:

First, we are going to introduce the background of the project. The project will be based on classifying stars in celestial space. We will explain our understanding about the dataset though EDA of the data. We will introduce three different clustering algorithms K-means, Hierarchical clustering, and Isolation forest. For K-means we use

Elbow-method to find optimal number of clusters(K) into which data may be clustered and Silhouette analysis to determine the degree of separation between clusters. We will study the model performance before and after applying dimensionality reduction techniques like PCA and t-SNE. Also, we will compare the results of the three clustering algorithms to show the best model.