# 1 Dataset Information

**Dataset Chosen:** APOGEE2 from SDSS
**Objective:** Identifying Young Stars and Differentiating them from Red Giants?

# 2 Exploratory Data Analysis

On doing EDA we find various trends, the following are the Observations.

## 2.1 Data Density and Galactic Coordinates

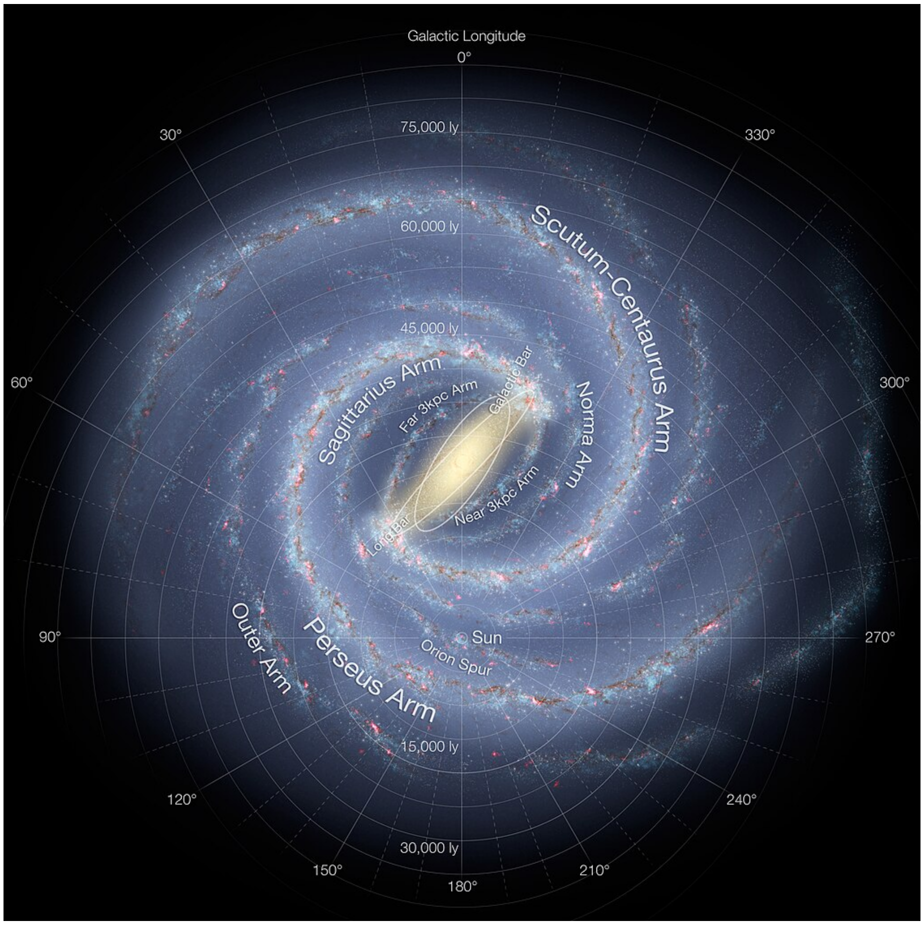

Figure 1: Galactic Coordinate System

The Artist's depiction Above of the Milky Way Galaxy showing the origin and orientation of galactic longitude. The galactic longitude (l) runs from the Sun upwards in the image through the center of the galaxy. The galactic latitude (b) is perpendicular to the image (i.e. coming out of the image) and also centered on the Sun.

From a the APOGEE Targetting Field Map and Density plot of the data points we can see that the data is slighly biased as we physically cannot observe the stars on the opposite side of our galaxy. However, the stars we can observe are well recorded. with only a few peaks in the interesting regions like the Magellanic Clouds.

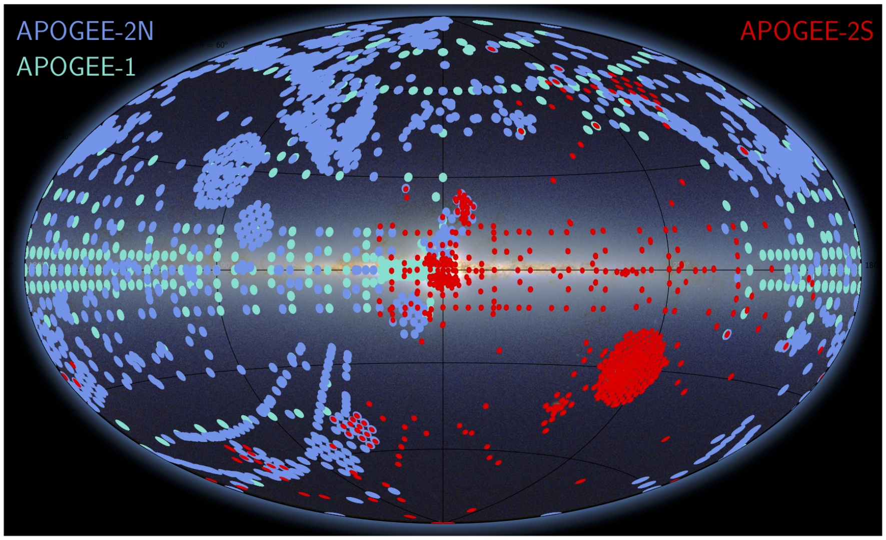

Figure 2: DR17 APOGEE field map where the fields are color-coded by the APOGEE sub-survey. APOGEE-1 in cyan, APOGEE-2N in blue, and APOGEE-2S in red. Figure by C. Hayes. Background image from 2MASS.

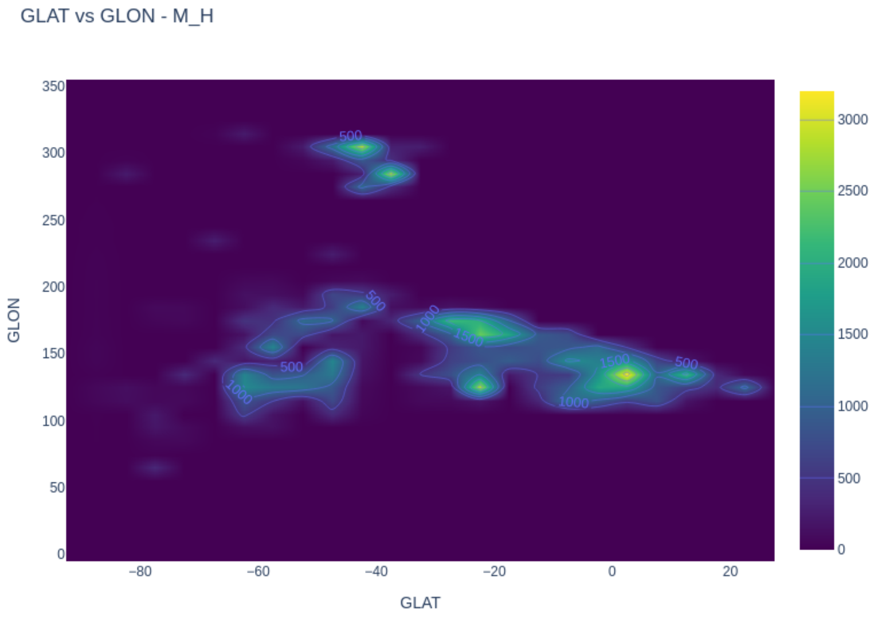

Figure 3: Density Plot

## 2.2 HR Diagram

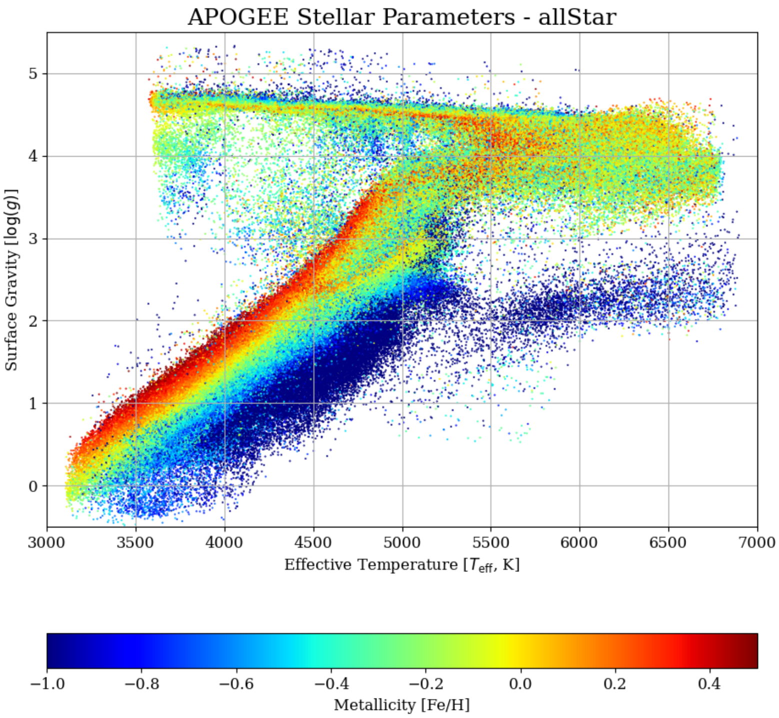

Figure 4: HR - Diagram

From a scatterplot between Effective Temperature at Surface $T_{eff}$ and the Surface Gravity $log(g)$ we can see that:

- the Relationship is mostly linear upto $T_{eff} < 5000K$ and $log(g) < 3$, this is known as the main sequence of stars, which is most of the lifetime of the star. Any given star moves from a lower Metallicity to a Higher Metallicity as it grows Older.
- The star moves above $T_{eff} > 5000K$ and $log(g) > 3$ when its dying, converting to various types of Dead Stars such as Red Giants, Red Supergiants, White Dwarfs, Neutron Stars and Black holes depending on a

mixture of various variables.

### 2.3 Star Formation Regions

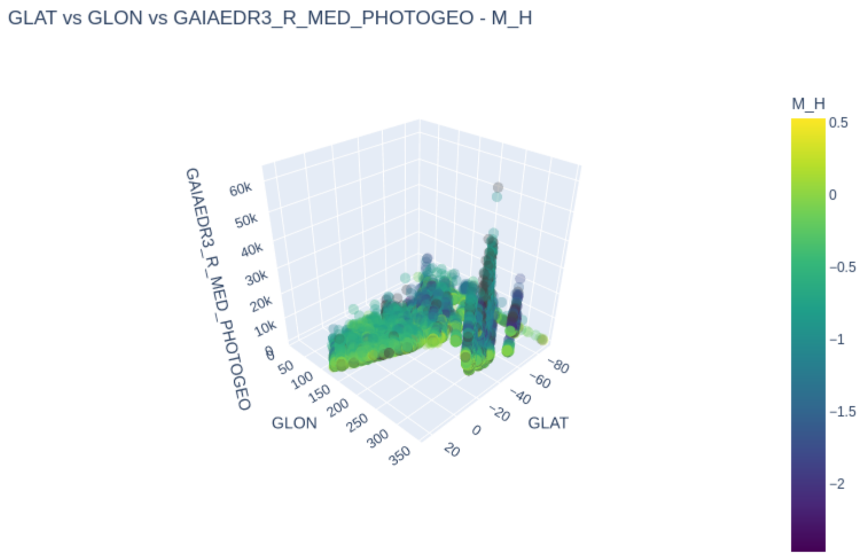

Figure 5: GLAT vs GLON vs Distance Colored by $M_H$

In the above Scatterplot we can Identify the various regions with a high star formation rate. They can be identified by the clumps of datapoints which are a deep blue.

The Deep-Blue Signifies a Low Metallicity which means that the particular star is Young.

## 3 Feature Selection

Techniques Applied: - Correlation Matrix

### 3.1 Correlation Matrix

We create the Covariance Matrix Heatmap and Remove Highly Correlated Features with $|r| > 0.85$:

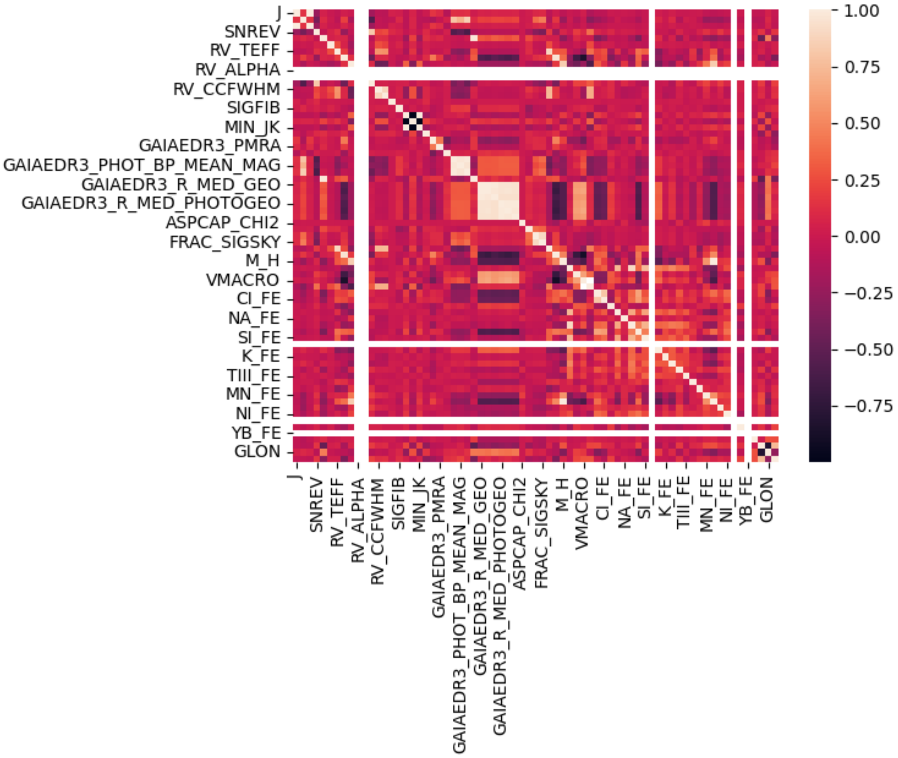

Figure 6: HR - Diagram

After Dropping we get the following columns left:

[‘J’, ‘H’, ‘K’, ‘SNREV’, ‘VHELIO_AVG’, ‘VSCATTER’, ‘RV_TEFF’, ‘RV_LOGG’, ‘RV_FEH’, ‘RV_ALPHA’, ‘RV_CARB’, ‘RV_CHI2’, ‘RV_CCFWHM’, ‘RV_AUTOFWHM’, ‘MEANFIB’, ‘SIGFIB’, ‘MIN_H’, ‘MAX_H’, ‘MIN_JK’, ‘MAX_JK’, ‘GAIAEDR3_PARALLAX’, ‘GAIAEDR3_PMRA’, ‘GAIAEDR3_PMDEC’, ‘GAIAEDR3_PHOT_G_MEAN_MAG’, ‘GAIAEDR3_PHOT_BP_MEAN_MAG’, ‘GAIAEDR3_PHOT_RP_MEAN_MAG’, ‘GAIAEDR3_DR2_RADIAL_VELOCITY’, ‘GAIAEDR3_R_MED_GEO’, ‘GAIAEDR3_R_LO_GEO’, ‘GAIAEDR3_R_HI_GEO’, ‘GAIAEDR3_R_MED_PHOTOGEO’, ‘GAIAEDR3_R_LO_PHOTOGEO’, ‘GAIAEDR3_R_HI_PHOTOGEO’, ‘ASPCAP_CHI2’, ‘FRAC_BADPIX’, ‘FRAC_LOWSNR’, ‘FRAC_SIGSKY’, ‘TEFF’, ‘LOGG’, ‘M_H’, ‘ALPHA_M’, ‘VMICRO’, ‘VMACRO’, ‘VSINI’, ‘C_FE’, ‘CI_FE’, ‘N_FE’, ‘O_FE’, ‘NA_FE’, ‘MG_FE’, ‘AL_FE’, ‘SI_FE’, ‘P_FE’, ‘S_FE’, ‘K_FE’, ‘CA_FE’, ‘TI_FE’, ‘TIII_FE’, ‘V_FE’, ‘CR_FE’, ‘MN_FE’, ‘FE_H’, ‘CO_FE’, ‘NI_FE’, ‘CU_FE’, ‘CE_FE’, ‘YB_FE’, ‘RA’, ‘DEC’, ‘GLON’, ‘GLAT’]