Capstone Project Proposal: STARS

(SDSS APOGEE Stellar Spectra)

Members: Farah Diba, Harnehmat Kaur, Kashish Mittal, Yael Lyshkow

Dataset

Stars (**SDSS APOGEE Stellar Spectra**) dataset. The file contains 99,705 stellar spectra and associated measurements from "red giant" stars collected and pre-processed by Henry Leung. The following columns of the dataset are the ones we'll be using for our project:

- `wavelength`: This is the wavelength that each spectrum is measured at in Angstroms (10^-10 m = 0.1 nm) for the individual stellar spectra. Since all spectra have the same wavelength, there is just one vector of 7514 values that can be applied to all spectra. (The rest of the columns each contain 99,705 entries.)
- `spectra`: Each stellar spectrum contains 7514 normalized intensity measurements (i.e. a vector of 7514 numbers) measured at the wavelength values specified in the wavelength column.
- `snr`: The signal-to-noise ratio (SNR) for the spectrum. This tells us roughly how well each spectrum has been measured (with higher=better).
- `teff`: The effective temperature
- of the star in Kelvin, which roughly tells us how hot it is.
- `logg`: The base-10 logarithm of the surface gravity
- of each star, where g is measured in cgs units (centimeters-grams-seconds), relative to 1 cgs.

Research Question 1

We want to find out whether red giants with higher intensity measurements are harder to measure.

Data for Question 1:
We will use two columns of the dataset for this research question. The first one is the 'spectra' column, which lists 7514 normalized intensity measurements of the star measured at the wavelength specified. The second one we will use is the 'snr' column. This stands for a signal-to-noise ratio for each spectrum, and it tells us roughly how well each spectrum has been measured, with a higher value meaning better.

Data Visualisations for Question 1:

1) **Scatterplot**
   We can plot a scatter plot of average intensity measurements against ‘snr’. A scatter plot is a simple graph that can visually depict the relation between the average intensity measurement and its ‘snr’.

2) **Box Plot**
   After categorizing stars into high-intensity stars and low intensity stars, we will plot a boxplot to show the ‘snr’ values of these two groups. The boxplot will compare the median and interquartile range of ‘snr’ of the two groups so that our audience can visually understand the difference in ‘snr’ of high intensity stars and low-intensity stars.

Methods to analyze data for Question 1:
We will conduct a **hypothesis test** for our claim that higher intensity measurements are harder to measure. Since each star has 7514 normalized intensity measurements, we can find the mean of these and mutate the dataset to contain the average intensity measurement for each star. We can then categorize the stars into high-intensity stars and low-intensity stars by finding the median average intensity of all the stars. After categorizing, we will find the mean ‘snr’ of both groups and conduct our hypothesis test. The null hypothesis will be that the mean ‘snr’ of high intensity stars is equal to the mean ‘snr’ of low intensity stars. The alternate hypothesis will be that the mean ‘snr’ of high intensity stars is lower than the mean ‘snr’ of low intensity stars.

Research Question 2:

We want to find out whether the surface gravity varies by a small factor among the different stars (which we can do by finding out whether the standard deviation is approximately 1).

Data for Question 2:
For this research question, we will be using one column of the dataset. This will be ‘logg’, which contains the base 10 logarithm of the surface gravity of each star.

Data Visualisations for Question 2:

1) **Histogram**
   The first visualization would be a histogram showing the sampling distribution of various standard deviation values taken over a large number of bootstrapped samples.

2) **Scatterplot**
   The second visualization would be a scatterplot of the surface gravity (‘logg’) values with a trend line over the plot to show a rough underlying uniform pattern in the data (geom_smooth()).

Methods to analyze data for Question 2:
We will use **bootstrapping** to resample the dataset multiple times, and get a sampling distribution of standard deviations taken over multiple samples. Then, we can calculate a 90% confidence interval over the distribution and see whether the confidence interval includes the desired value (1).

Research Question 3:

We want to find out if there is a correlation between the wavelength of red giant stars and their corresponding effective temperature.

Data for Question 3:
For this research question, we will be using two columns from the dataset. The first one is ‘wavelength’, which contains the wavelength of each spectrum measured in Angstroms. The second one is ‘teff’, which contains the effective temperature of each star measured in Kelvin.

Data Visualisations for Question 3:

1) **Correlation Heat Map**
   A correlation Heat Map is a way of visualizing the strength of correlation between two numeric variables. Here, we will use a heatmap to find the strength of correlation between the wavelength and effective temperature of red giant stars.

2) **Line Plot**
   A line plot of ‘wavelength’ with the ‘teff’ can also be helpful in showing the trends between the two continuous variables, and visualize the correlation in a unique way.

Methods to analyze data for Question 3:
We will use the corr() function in R to find the correlation between the two variables in the dataset we need to compare (‘wavelength’ and ‘teff’). After we find the correlation, we can plot the correlation to understand it better using a heatmap. If the correlation is positive, we can say that the wavelength increases with increase in ‘teff’, and when it is negative, the opposite is true, i.e., the wavelength decreases with increase in ‘teff’. If the correlation is close to 0, we can say that the values are not correlated. It is also worth mentioning that correlation does not imply causation,

so we will not be assuming any relationship between the two variables except for correlation.

Group Contributions:

- **Farah**: Coding and one data visualization for RQ 1, recording the results in the final report, working on the poster, working on the final project paper
- **Harnehmat**: Coding and one data visualization for RQ 1, Coding and one data visualization for RQ 2, working on the progress report, working on the poster, working on the final project paper
- **Kashish**: Coding and one data visualization for RQ 2, Coding and one data visualization for RQ 3, working on the progress report, working on the poster, working on the final project paper
- **Yael**: Coding and one data visualization for RQ 3, recording the results in the final report, working on the poster, working on the final project paper

Project Timeline:

- **March 10**: Coding for Research Question 1 (Farah and Harnehmat)
- **March 15**: Coding for Research Question 2 + Progress Report (Harnehmat and Kashish)
- **March 20**: Coding for Research Question 3 (Kashish and Yael)
- **March 31**: Record the findings from R in a report (Yael and Farah)
- **April 1**: Prepare the poster for the STA130 poster fair (All of us)
- **April 4**: Complete the paper using the correct structure and specifications (abstract, summary, conclusion etc.) (All of us)