Capstone Project Proposal - Stars

Questions

1. **Does the effective temperature of the stars correlate to the wavelength?**

*Method:*

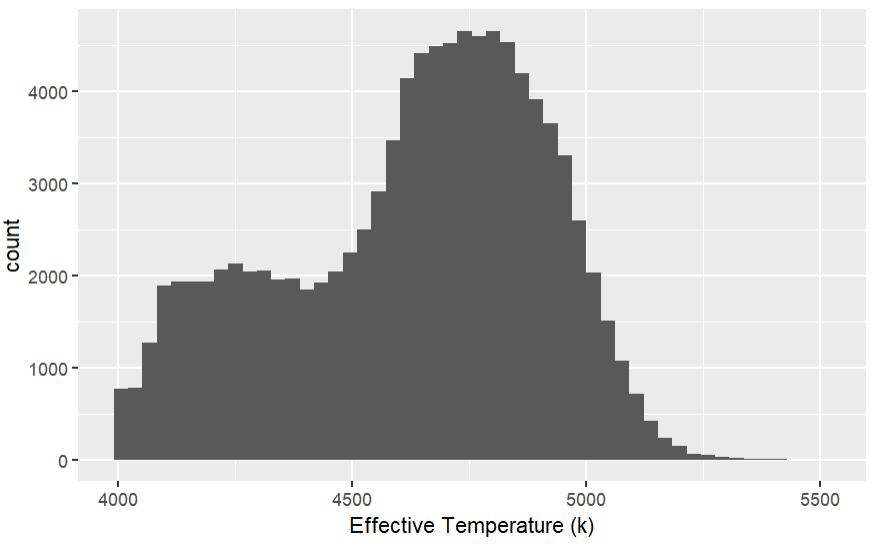
We are trying to determine if the effective temperature of the stars has a relation to the wavelength of the spectra. To do so, we need to find various different values, important ones being the mean and median. We can also observe if the graphs have various characteristics, such as are unimodal, or multimodal, left-skewed or right-skewed. By finding similarities and differences within the graphs and calculations, one can identify if the wavelength is in fact determined, or altered, due to the temperature of the stars, or if the wavelength of the spectra is just a random occurrence.

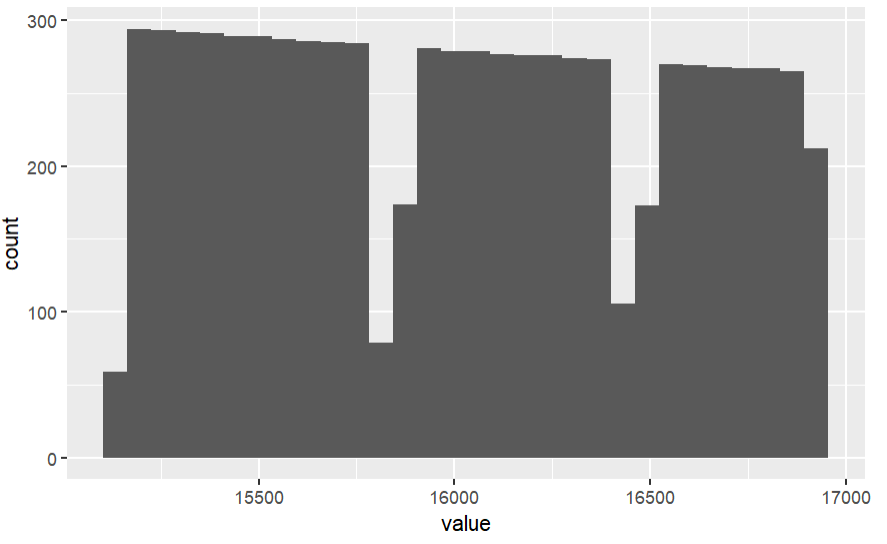
*Data analyzed:*

We will consider the effective temperature values of every star as the data for “the effective temperature of the stars.” For the “wavelength of the spectra”, we will consider a combination of two data sets, including the count of the wavelengths of the stars and the number of stars matching each wavelength. We will be using the three data sets in total to determine the answer to our hypothesis. Specifically, to do this, we would do a linear regression between data for the effective temperature and the data for wavelength. We will set a value of r^2, and the hypothesis testing for this will then be: null hypothesis is r^2 equal to set value, and alternative value is null hypothesis is r^2 bigger than set value. We can consider that value of r^2 as the “acceptable value for the relationship to be highly valid”, so that if our conclusion is that the alternative hypothesis is accepted, then we would assume the relationship to be valid.

However, when we consider our conclusion, we will also note that the relationship between the two variables may not necessarily be a linear function (first power polynomial), so we will also consider the relationship to be higher-power polynomials (i.e. second, third, fourth-power), and other types (i.e. logarithmic, trigonometric, etc.) to get the function that satisfies the relationship the most.

Finally, also note that we may not necessarily use r^2 as the variable, as any linear regression variable is accepted. In statistics, r-squared can be defined as “the proportion in the variance of the response variable that can be explained by the predictor variable in the regression models”, while r can be defined as “the correlation between the predictor variable and the response variable”, and any of the two can be chosen as the final variable considered. However, as we have currently not proceeded to the linear regression part of STA130H1, we will leave this as uncertain.





1. **How well can the confidence interval predict the median of the wavelength of all the stellar spectra?**

*Method:*

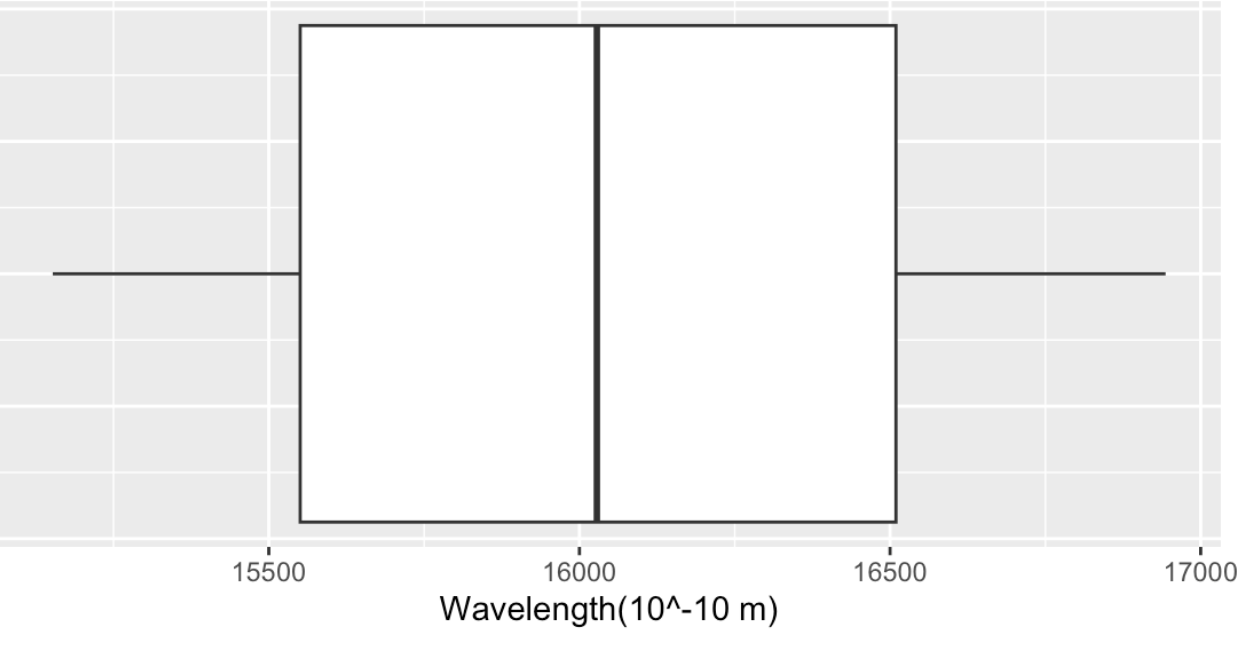
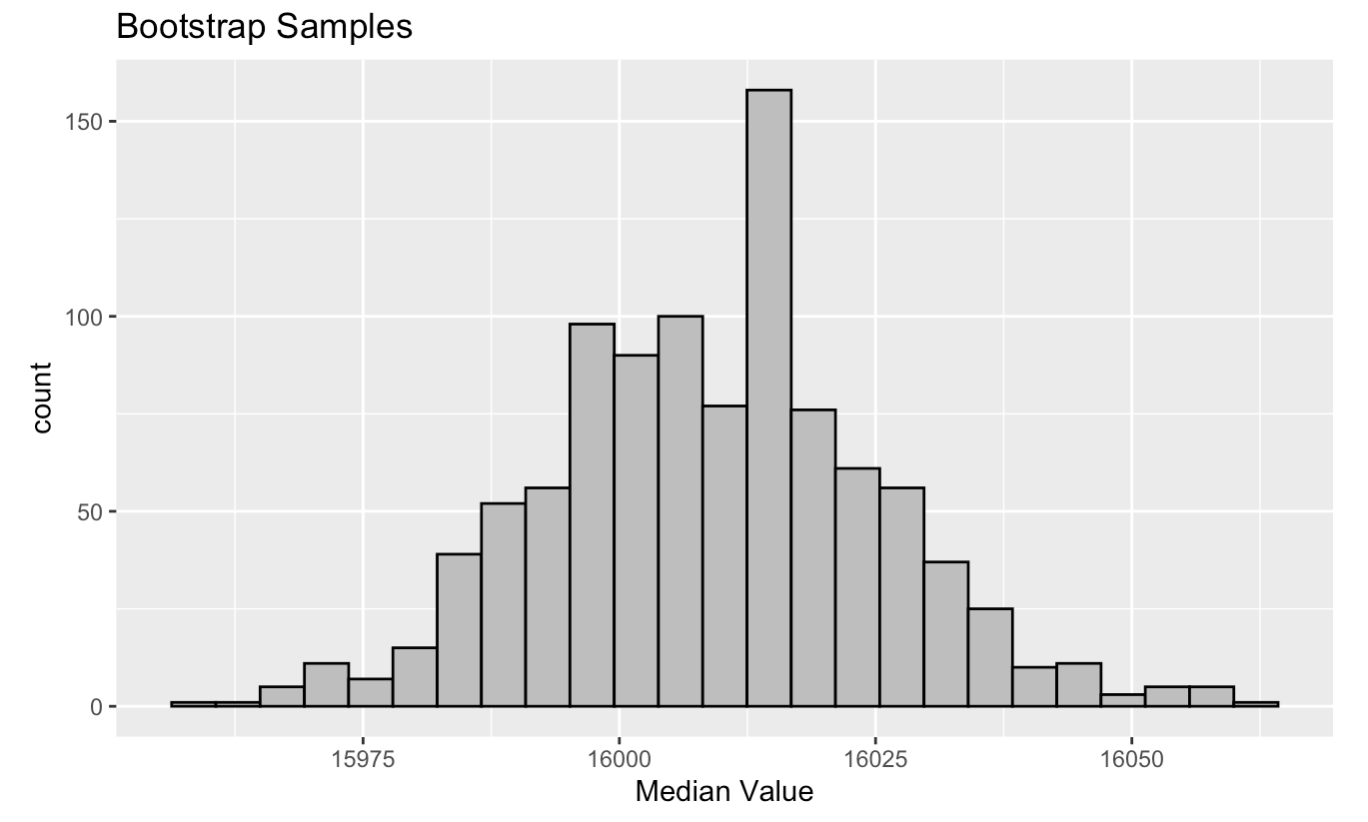
We only need the wavelength data, as mentioned in the “Data analyzed” of our previous question, for getting a final answer for this research question in the actual Capstone Project.

The goal is to see whether the true median of wavelength applied to all spectra is included in the CI. The true median is available for us, which we can get from a box plot of the wavelength data. Therefore, all we have to do is to compute different confidence intervals (70%, 80%, 90%, 99%) based on boot-strapped data, and then compare each of them to the median displayed in the box plot. The bootstrap sample can be addressed by slicing a random subset of the wavelength data. We get the confidence intervals based on the formula.

After the comparison of the median and CI, we can give a conclusion. If the median in the box plot is a value around the center of the interval,then we say this confidence interval predicts the median very well, if the value of the median is inside the interval but near the edge, we say it predicts not very well. Otherwise (out of the interval), we consider that this prediction failed. If all the confidence intervals correctly contain the true median, we consider that they all predict well.

*Visualizations:*

This histogram is a distribution of the median value of 1000 bootstrap samples. Each bootstrap sample has a size of 2000 wavelengths. Further we can get those CIs from this plot and compare it to the wavelength median in our boxplot.



1. **How does the range of the amount of calcium on stars relate to the amount of iron that is found on the stars?**

*Method:*

In order to find the amount of calcium located within the stars, we first have to call on the data. By using the fe\_h and ca\_h, we can create a tibble with these two types of data. With visualizations such as plots, we are able to briefly compare between the two. By narrowing down these plots to histograms and boxplots, the histogram shows us more about the distribution and spread of the “count” of iron and calcium while the boxplot gives a more summarized result such as the median. Using these two tibbles, we can narrow down information further by specifically finding attributes such as the distribution, standard deviation, median and so on. After finding these attributes, we are able to finallize results and make a conclusion of how the range of the two separate dataframes relate to each other or not.

*Visualizations:*

A histogram of the level of calcium in stars. A boxplot of levels of iron in the stars. We will be using more visualizations to determine a better representation of the levels of calcium and irons in the stars such as another boxplot or histogram.

