STA130 Capstone Project Final Report: Correlations of Data Related to Stars

Group Members

- Kevin Shao
- Adriano Rymon-Lipinski
- Angus Yeung
- Yutong Han

Contents

Abstract, Introduction, and Data

In this STA130 Capstone Project, our group chose to analyze data based on the *SDSS APOGEE Stellar Spectra*, and we decided to statistically answer three questions based on this dataset. The following is the introduction divided according to each of the three questions.

The first question is “Does the effective temperature of the stars correlate to the wavelength?”. Here, we are going to directly use the *wavelength* variable of the dataset for the “wavelength”, and the teff variable of the dataset for “effective temperature.” Since *wavelength* and *teff* have different sample sizes, we will try to get a random sample of 7514 from the total of 99705 observations for *teff*, and with the same-size population of *wavelength*, since the *wavelength* variable only has 7514 observations in total. As now the two variables have the same number of observations/sample size, we can perform the hypothesis test on a simple linear regression between the two datasets, with the typical p-value of 0.05. Using the hypothesis test results, we can get a statistically significant conclusion to our set Question 1. We also need some background information of some relevant theorems describing the relationship between wavelength and effective temperature for stars. Wien’s Law is the direct relationship between the wavelength and the effective temperature of a star. There are specific equations which are used to calculate different values to measure this correlation, but we will simply use the data sets we have and use linear regression to measure the correlation instead. Using Wien’s Law, we find that there is a wide range of temperatures, which each pertain to different colors of the stars, low being about 3000k and orange, while high being 30,000km and blue (*From Source 2, as indicated in Works Cited*).

The second question is “What is the estimation of the true population mean of the red giants’ surface gravity”, and the large dataset we used is *SDSS APOGEE Stellar*

Spectra. Since this question focuses on the surface gravity, we need to use the variable *logg*, which is the base-10 logarithm of surface gravity g of each star, where g is measured in *cgs units centimeters-grams-seconds*. We select the column *logg* from the overall data by using h5read("logg") and as_tibble() to create a new data called *logg_data*. For the following steps, *logg_data* is the only data that we need to consider. This data has 99,705 observations, 1 variable called "value", which is the base-10 logarithm of surface gravity g of each star. With this data, we can do simulations and find the confidence intervals.

The third question is "How does the range of the amount of iron, correlate to the amount of other elements that can be found on the surface of the stars", and the data we will be analyzing is the abundance of elements found on the surface of the stars. The elements that we are dealing with are iron, aluminum, carbon, calcium, iron, magnesium, nitrogen, and oxygen. Specifically, our question compares iron with the other elements found from the data set. Since iron is found the most, we want to analyze the results of the iron found with the other elements and try to find whether or not there is a correlation between them and what this correlation represents.

Question 1: Methods, Results, and Visualizations

As mentioned in the introduction, since we are interested in whether there is a correlation between the effective temperature of the stars and the wavelength, we will perform a hypothesis test on a simple linear regression between the two data sets.

For this test, we will set the null hypothesis as there is no relationship/correlation between effective temperature and wavelength of stars, while the alternative hypothesis is there is a relationship/correlation between the two variables about stars. In this hypothesis testing, we will set wavelength as the independent variable (y) and effective temperature as the dependent variable (x), but the specific order doesn't really matter.

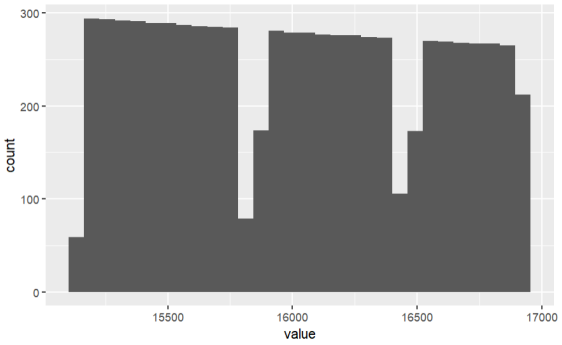
The following is a description of the general approach of how to perform the hypothesis testing using R. First of all, we import the two data sets mentioned above and use the lm() summary() function of R-studio to calculate the p-value. Then, we determine whether the data is statistically significant. If the p-value is less than or equal to 0.05, the test result is statistically significant, and we can reject the null hypothesis, and vice versa. Note that we will still use the most common significance level of 0.05, as mentioned in the introduction section.

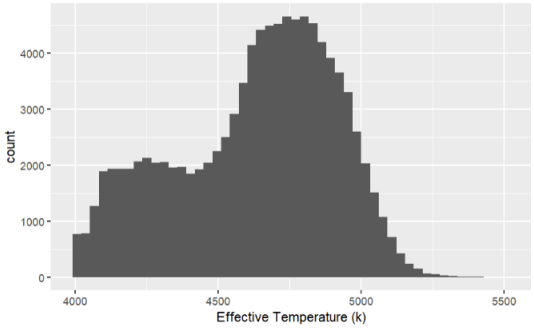
However, when we actually perform this test, we need to process the data, which is basically filtering out all the empty data points in this case. We first select the two columns of data, *wavelength* and *teff* (effective temperature) of stars, out of the large data set. Then, for each column, we use R's omit.na() function to filter out all the data points with NA values. Then, we plug the large data set *SDSS APOGEE Stellar Spectra* and the two variables, and plug them into the lm() function to produce the linear regression model. We can now get a specific summary of all the calculations resulting from the linear regression test using the summary() function.

Then, we proceed into the visualization of the two variables' data. First of all, we need to visualize the general distribution of the two variables' data points. As shown in the two figures below, we have visually represented the distributions of "wavelength" and "teff" as two separate histograms using the geom_histogram() function of R, since

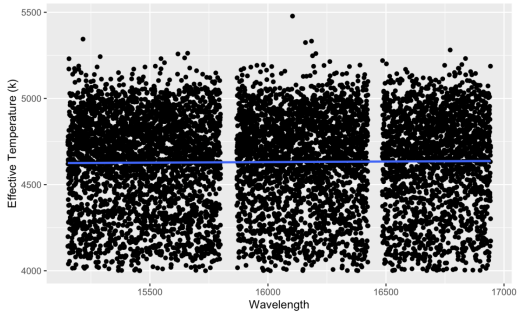
the two are both quantitative variables. Note that the graph with x-axis as “value” and y-axis as “count” is the data for “wavelength” (left figure), while the graph with x-axis as “Effective Temperature (k)” and y-axis as “count” is the data for “teff” (right figure). Here, we can see that *teff* is generally skewed to the right, has a center at roughly around 5000 K, and has a range of values from around 4000 to 5400 K. *Wavelength* is generally symmetrically distributed (with frequencies for each wavelength value to be generally constant), has a center of around 15000 angstroms (1 angstrom = 10^(-10) meter), and has a range of values of generally from 3000 to 16900 angstroms.

However, when performing the hypothesis test for simple linear regression, we realized that the number of observations for each of the two variables must be the same. As shown in the data of R’s environment, *teff* (effective temperature) has 99705 observations, and *wavelength* has 7514 observations. Therefore, in order to make the hypothesis test runnable, we will produce a random sample of 7514 observations for *teff*, and perform this test with the original *wavelength*. As the sample of *teff* is completely random, it can be representative of the entire dataset, and we will also consider this validity using values of residuals. Note that there is major change here from what we designed in the progress report, as we previously designed to perform multiple hypothesis tests containing the entire number of observations of *teff*, which is time-consuming. This method is also inaccurate, since the last test only contains 2023 observations for *teff*, and it is very hard to determine which 2023 of the 7514 observations of *wavelength* is best to use for this last test. Therefore, we decided to use this approach which only contains one test, since 7514 observations of *teff* that are selected purely random is representative of the entire 99705 observations of *teff*, and is also much more convenient for others to reproduce.





Again, just before we actually start the process of performing the hypothesis testing for simple linear regression for *teff* and *wavelength*, we want to first get an idea about what the relationship between wavelength and effective temperature of stars can be. Therefore, we will visualize a scatter plot between the two variables using the ggplot() and geom_point() functions. We will also get a line of best fit using the geom_smooth() function. We get the figure below, which appears that there is no correlation between wavelength and effective temperature. The figure also shows a blue line of best fit which is almost horizontal, with the effective temperature remaining at around 4600 K for all wavelength values, typically 0 to 16800 for the data we used.



Now, we will perform the actual hypothesis testing for simple linear regression. First of all, we can get a sample of 7514 observations of the effective temperature (*teff*) using the sample_n() function, and then use the lm() summary() function to get the statistical data for the linear regression test. The specific statistical data are shown in the below:

```
Rows: 99,705
Columns: 1
$ value <dbl> 5031.264, 4975.689, 4981.525, 4073.770, 4757.323, 4669.081, 4660.645, 45…

Call:
lm(formula = teff_sample$value ~ wavelength$value)

Residuals:
    Min      1Q  Median      3Q     Max
-634.13 -203.68   44.99  211.14  846.66

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.526e+03  9.763e+01  46.360   <2e-16 ***
wavelength$value 6.539e-03  6.089e-03   1.074    0.283
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 279.2 on 7512 degrees of freedom
Multiple R-squared:  0.0001535, Adjusted R-squared:  2.038e-05
F-statistic: 1.153 on 1 and 7512 DF,  p-value: 0.2829
```

As shown in the figure above, we get a standard value of the residual, denoted as *residual standard error*, of about 279.2 on 7512 degrees of freedom, which is an exceptionally good value for my statistical testing, which proves that this hypothesis testing is relatively accurate with acceptable results and implies that the sample of 7514 observations can indeed give an acceptable representation of the entire dataset. Here, we get a p-value of 0.2829, which is much larger than our set significance level of 0.05.

The p-value is the probability that the test statistic is extreme or more extreme than the observed value, assuming that the null hypothesis is true. In this case, the p-value is not considered to be statistically significant, so we are not able to reject the null hypothesis. Therefore, the results we get is that there is no correlation between the wavelength and effective temperature of stars. Moreover, since the residual standard error is relatively low, this result is also highly acceptable.

Question 2: Methods, Results, and Visualizations

As indicated in the introduction, our question is “What is the estimation of the true population mean of the red giants’ surface gravity.” The large dataset we are going to use is the same for all three questions. For this question, we want to use the bootstrapping method to estimate the true population mean. Bootstrapping is a useful method in general, as it allows us to take a sample of a population, and resample it many times to produce an accurate confidence interval for a set of data. In this instance, we are looking for a confidence interval with regard to the mean of red giants surface gravity. Bootstrapping will be a useful method, since we only have 99, 705 samples, and there are infinite stars in the universe. By bootstrapping these stars, we can have an accurate prediction of the true mean value that we are looking for.

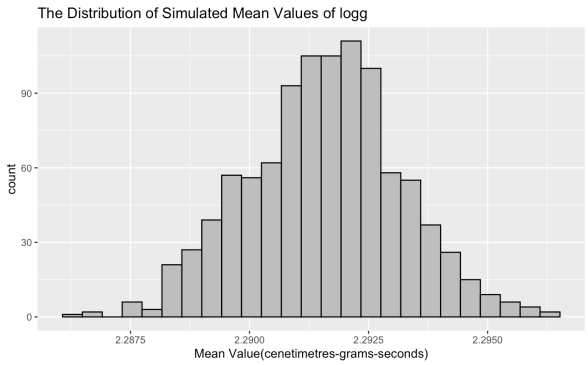
First, we have already made sure the data is cleaned. To check this, we used the “filter(is.na(value))” function on the logg_data and nrow() function to see how many empty data it contains. If it is 0 (no empty data found), we don’t need to clean. Otherwise, we may need to use the function “filter(!is.na(value))” on the logg_data and rename it to the new cleaned “logg_data”. Actually, we found that there is no empty data found in the original logg_data, indicating we could get comfortable with it and move on to the next steps.

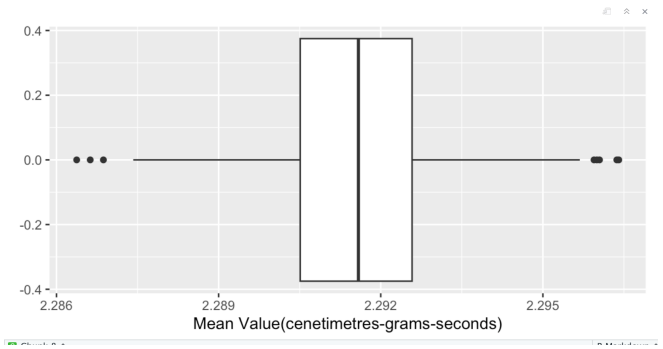
The second step we have done was to set some variables needed. We want the simulation to produce 1000 simulated bootstrap samples, so we set a variable called “M” and assigned 1000 to it.(M<-1000). Another variable was called the sample_size which means the number of elements in each simulated bootstrap sample. This number was the same as the size of the original data “logg_data” which has 99,705 observations. So we set sample_size = 99705. The last variable we created was a vector with 1000 “NA”s, the variable name is “sim_mean_logg” . We do this because it will finally be substituted by 1000 simulated mean values. Then we can see the distribution of these simulated mean values from a histogram or box plot.

The third step we have done was to start the simulation. We use set.seed() and plugged a random number in it to make sure the random sample is reproducible. In the “for loop” range from 1 to M, each time we shuffled the “value” in the original data “logg_data” 99705 times with replacement to get one of the bootstrap samples with the same size as the original data called “bootstrap_sample”. We can reach this by using the function “slice_sample(n=sample_size, replace = TRUE)”. Then we used the “summaries()” function to get the mean value “sim_mean” for each “bootstrap_sample” Here we have a “as.numeric” applied on the “sim_mean” since we need to use coercion to change the type from “list” to “dbl” . The last step in the for loop was to save the mean value to the right index of the vector “sim_mean_logg”. Thus all the NAs in

“sim_mean_logg” were substituted by those mean values of 1000 simulated bootstrap samples(sim_mean).

The fourth step we have done was to create a visualization and see the distribution of the simulated mean values. But here we changed the “sim_mean_logg” from a vector to a tibble. Otherwise the plot will fail to be created since it is a numeric vector rather than a <data.frame>. So we used the tibble() function and successfully got a data form of the “sim_mean_logg” called “sim_mean_logg_data”. After that we can use “ggplot(data=sim_mean_logg_data, aes(x=sim_mean_logg)) + geom_histogram()” to get a visualization.





From the visualization, we discovered that the histogram(below) has a symmetric unimodal distribution range from 2.286 to 2.296.

The last step we have done was to calculate the (X% confidence intervals). Here we need to use quantile() function and a formula

q1 = (1 - X / 100) / 2), q2 = (1 + X / 100) / 2)

We got the confidence interval by:

quantile(sim_mean_logg, c(q1,q2))

And we did four confidence intervals in total.

Currently we have X% (99%, 95%, 90%, 85%) CIs that can estimate the true population mean of the base-10 logarithm of red giants’ surface gravity.

```r
```{r}
samp_size <- 99705
n_trial <- 1000

set.seed(911)
sim_mean_logg <- rep(NA,n_trial)
for (i in 1:n_trial){
  bootstrap_sample <- logg %>% slice_sample(n=samp_size,replace = TRUE)
  sim_mean = bootstrap_sample %>% summarise(mean(value)) %>% as.numeric()
  sim_mean_logg[i] = sim_mean
}
sim_mean_logg_data<- tibble(sim_mean_logg)
#sim_mean_logg_data %>% glimpse()

sim_mean_logg_data %>% ggplot(aes(x=sim_mean_logg)) +
  geom_histogram(color = "black",fill ="grey",bins = 25) +
  labs(title="The Distribution of Simulated Mean Values of logg",
       x="Mean Value(cenetimetres-grams-seconds)")


X1=c((1-0.99)/2,(1+0.99)/2)
X2=c((1-0.95)/2,(1+0.95)/2)
X3=c((1-0.90)/2,(1+0.90)/2)
X4=c((1-0.85)/2,(1+0.85)/2)

quantile(sim_mean_logg,X1)
quantile(sim_mean_logg,X2)
quantile(sim_mean_logg,X3)
quantile(sim_mean_logg,X4)
```

```
    0.5%    99.5% 
2.287432 2.295685 
    2.5%    97.5% 
2.288447 2.294669 
      5%      95% 
2.288829 2.294173 
    7.5%    92.5% 
2.289210 2.293809 
```
```

Question 3: Methods, Results, and Visualizations

"How does the range of the amount of iron correlate to the amount of other elements that can be found on the surface of the stars?" was our initial question that we came up with from the data given to us. After receiving feedback, we have realized that this question seems too broad and vague and we decided to change it up to "How does the range of the amount of iron, correlate to the amount of other elements that can be found on the surface of the stars". This new question will allow us to gain deeper insight on the correlation between these elements and allow us to use more strategies to answer our question.

In our initial method, we were comparing calcium and iron using histograms but now since we know how to use multivariable linear regression models, we can compare multiple materials with accuracy.

For us to understand the correlation of the elements, we first must know a little more about the material first so that our conclusions will make sense. We know that iron has the widest range found on stars compared to the other elements. This means that we can compare the other elements of iron. By using iron to compare with the other elements found on the stars, we are able to visualize and find a correlation between the data found between iron and the other elements.

In order to visualize this we must separate the iron data with the data of the other elements. We can use a variable "not_iron" to represent all the elements that aren't iron by adding the other elements up. This way, instead of comparing each of the elements with iron, we now have a group representing all the elements which can be used to compare with iron.

```
iron_elements_correlation <- cor(elements$feh, elements$not_iron)
iron_elements_correlation
```

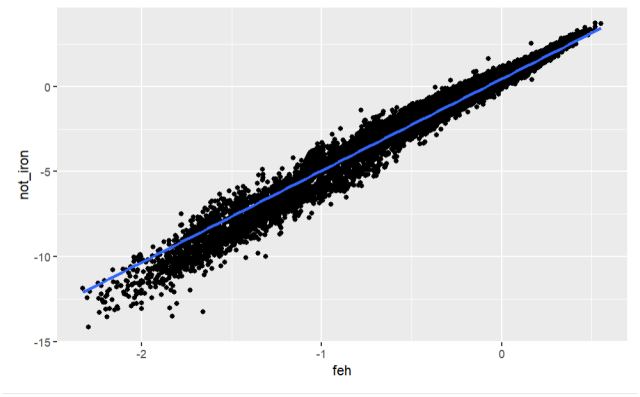
After combining the other elements except iron, we come up with a tibble that we can summarize. These values represent the abundance of the element produced by the stars.One way that we can find correlation between these materials would be using the cor() function in R. This should give us a number between -1 to 1 revealing whether they are positively or negatively correlated. Furthermore when the absolute value of the output is closer to one, it represents the strength of correlations

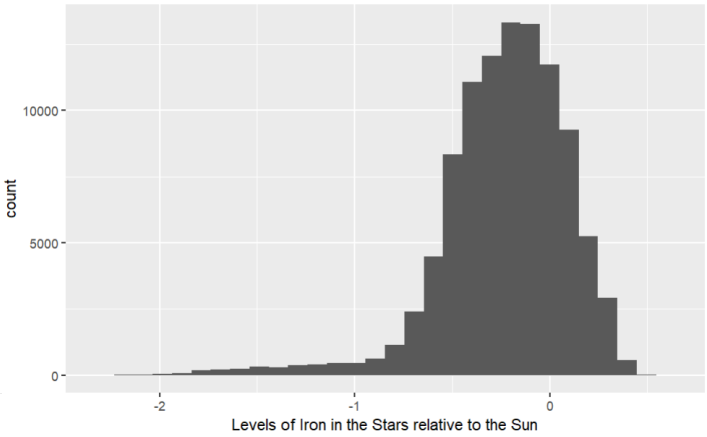
```
[1] 0.9814156
```

After running the cor() function with arguments of iron and non-iron elements, we find that the correlation number is around 0.98 which is close to 1. This means that the correlation between the iron and elements that are not iron are positively correlated. Since 0.98 is very close to 1, as it is 0.02 off from 1, this means that these two variables are strongly correlated.

This means that once we plot out our line of best fit, the slope of the line will be positive and as the x value becomes higher, the y value should as well.

Another method would be plotting these materials on scatter plots and then comparing. With scatter plots instead of histograms, we can draw a line of best fit which can tell us the growth or decrease in the other materials compared to iron. Furthermore, we can use our existing histogram of the levels of iron in the stars to give us a better understanding when analyzing the scatterplot.





The above visualization is the scatter plot representation of the iron and non-iron elements. From this plot, the x is the iron found on the stars while the y is the non-iron elements found on the stars. As we can see from the visualization, it starts off at the bottom with a large spread of iron and other elements. Then, as more iron is produced, the value of the non-iron elements becomes less spread. From this graph, this tells us that there are less of other elements being produced the more that iron is being produced. Not only are there less, the concentration of other elements compared to iron are also being increased. We can see that the points at the tip of this plot are highly condensed compared to the bottom which is spread out.

The second visualization of the histogram shown above represents the levels of iron in the stars relative to the sun. We can see that there is a left skew in the histogram with a median of a little bit under 0. By understanding the shape of this visualization, we are able to understand some of the secondary questions we get when analyzing the scatter plot such as how the points are plotted on the scatterplot when dealing with the x axis.

The line of best fit from this graph explains more about the correlation of iron and the other elements. Since the x is the iron produced and y is the other elements produced, the line of best fit would have a slope and since it is positive in the visualization, it can conclude a fact that as iron is being produced, the other elements are not decreasing but rather increasing as well. Furthermore, we can see that this slope is fairly consistent as it increases and is not too vertical. This can tell us that the

proportion of iron to other elements are quite similar. From just the visualizations, we are unable to make a confident conclusion with the abundance of iron compared to other elements.

If we want numbers instead of visualizations, we can use the lm() method to compare the slopes produced by the material and their levels on the stars. This way, the numbers can give us a more accurate representation which can help to make a clearer observation.

```
Call:
lm(formula = feh ~ not_iron, data = elements)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44484 -0.02565  0.01615  0.03883  0.79455

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0879051  0.0002193  -400.9   <2e-16 ***
not_iron     0.1786974  0.0001107  1614.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06323 on 99703 degrees of freedom
Multiple R-squared:  0.9632,    Adjusted R-squared:  0.9632
F-statistic: 2.608e+06 on 1 and 99703 DF,  p-value: < 2.2e-16
```

Based on these results from the lm() function, we can see that our estimated linear regression equation gives us an intercept of -0.08 and a slope of 0.17. From our previous visualization of the scatter plot, the line of best fit is now represented by a slope of 0.17 which tells us the relationship between iron and the other elements. As the abundance of iron rises, the other elements rise as well; however, with a slope of 0.17, it means that iron is indeed found more on the stars and as the amount increases, the other elements slowly increase along with the iron.

Represented below here are some of the codes and functions we have utilized to generate these visualizations for this question.

```r
```{r}



oh <- "STA130_APOGEE.h5" %>% h5read("o_h") %>% as_tibble()
ch <- "STA130_APOGEE.h5" %>% h5read("c_h") %>% as_tibble()
mgh <- "STA130_APOGEE.h5" %>% h5read("mg_h") %>% as_tibble()
alh <- "STA130_APOGEE.h5" %>% h5read("al_h") %>% as_tibble()
nh <- "STA130_APOGEE.h5" %>% h5read("n_h") %>% as_tibble()
cah <- "STA130_APOGEE.h5" %>% h5read("ca_h") %>% as_tibble()
feh <- "STA130_APOGEE.h5" %>% h5read("fe_h") %>% as_tibble()

not_iron <- oh + ch +mgh + alh + nh + cah


elements <- bind_cols(not_iron, feh) %>% rename(not_iron = value...1, feh =
value...2)
elements


iron_elements_regression <- lm(feh ~ not_iron, data = elements)
iron_elements_regression

iron_elements_correlation <- cor(elements$feh, elements$not_iron)
iron_elements_correlation

ggplot(data = elements) + aes(x = feh, y = not_iron) + geom_point() +
geom_smooth(method=lm, se=FALSE)
```
```

Conclusions and Discussions

The following is the conclusions and reflections of our Capstone Project, and we will divide it for each of the three questions.

From the process described in Question 1: The Entire Process, we can get that there is no correlation between wavelength and effective temperature of the stars, as derived from the p-value of about 0.2829 and the set significance level of 0.05. However, we must also consider some factors that may affect the result of our question. First of all, we randomly chose 7514 observations from the effective temperature, with a total of more than 50,000 observations, so there is a probability that all 7514 observations happen to fit the correlation by coincidence, so this does negatively affect our results for Question 2, but as shown in the low residuals calculated using lm() summary(), the validity of our results still seem to relatively high. Secondly, as shown in the scatterplot in Question 1: The Entire Process section, the line of best fit is a horizontal line, which appears to also support that there is no correlation between the two variables, but it is also possible that a very weak correlation exists, so the specific answer is still unsure by using the scatterplot. Moreover, there are also other methods to get results, including the use of bootstrapping, but this random sample method is much more convenient, simple, and easy to reproduce, but also has an acceptable relatively-high accuracy. We can also use the r-squared values to make our hypothesis testing more valid, which we did not use in the process. However, as shown in the figure of statistical data, the r-squared value is also very close to zero, meaning that a very low proportion of the dependent variable (effective temperature) can be explained by the independent variable (wavelength), which also implies that there is no correlation between the two variables. Therefore, as our validity still remains high, and few additional factors can fundamentally change our conclusion of the hypothesis test, we firmly believe that the methodologies and conclusions we used are still acceptable for Question 1 of this Capstone Project.

From our second question. We got 4 confidence intervals of 4 confidence levels(X% = 99%, 95%, 90%, 85%). So we can better predict the range of real population mean value of the red giants' surface gravity. We can draw conclusions: We have 99% confidence that the real population mean base-10 logarithm of surface gravity g of red giants is between 2.287432 to 2.295685. We have 95% confidence that the real population mean base-10 logarithm of surface gravity g of red giants is between 2.288447 to 2.294669. We have 90% confidence that the real population mean base-10 logarithm of surface gravity g of red giants is between 2.288829 to 2.294669. And we have 85% confidence that the real population mean base-10 logarithm of surface gravity g of red giants is between 2.289210 to 2.293809. In addition, we have discovered: as the confidence level decreases, the size of the confidence interval also decreases. The confidence interval of a lower confidence level is a subinterval of the confidence interval of a higher confidence level.

From our third question, we can conclude that our answer to our question has been explained through the process of analyzing the data. We were able to get a strong positive correlation between the iron and non- iron elements found on the stars. Furthermore, we were able to see this correlation by visualizing using a scatter plot and looking at the line of best fit within the points. Then by using a linear regression model, it was made clear that the correlation between iron and the non-iron elements could be

represented by a slope. Also note that based on the results that we obtained in the third question, we can see that our results are statistically significant. Some of the information may be extra and can be used to tell us more about the original question we had. For example, the linear regression model gave us p-values. If we had a null and alternate hypothesis, we could have incorporated it into getting more of our findings. However, some small supplementary questions were answered by some of the residuals from the summary of the lm(). Something learned from the methods would be to filter out the information we don't need and only input the data to get the result we want.

Works Cited

1. Requirements of Capstone Project Given by STA130, The Dataset *SDSS APOGEE Stellar Spectra* As Given
2. https://www.atnf.csiro.au/outreach/education/senior/cosmicengine/stars_colour.html#:~:text=Wien%27s%20law%20is%20name%20of%20the%20direct%20relationship,and%20is%20a%20measure%20of%20the%20surface%20temperature. (Related to Wien's Law, Used for Question 1's Background Information)

Individual Contributions

The following are the individual contributions written by each group member:

1. Kevin Shao (Student ID: 1008781908): In the Capstone Project, I was responsible for setting the question, designing the entire process of the statistical approach, and writing the conclusions for Question 1. Note that I did not complete the visualizations by myself.
2. Adriano Rymon-Lipinski (Student ID: 1008717453): In the Capstone Project, I was responsible for the visualizations and other R-code parts for Question 1, 2, and 3, and also gave many inspirations and ideas to Kevin and Angus (main producers for analysis of Question 1 and 2).
3. Angus Yeung (Student ID: 1008724089): I was responsible for finishing all of question 3 except the visualizations (which was done by Adriano). I was also responsible for researching more about the elements and stars so that I was able to describe question 3 with more precision. Furthermore, using the visualizations that Adriano was able to provide me with along with the guidance of myself, I was able to make conclusions about some small supplementary questions I had dealing with question 3. The data, discussion, and conclusion were also updated throughout the process of doing question 3.
4. Yutong Han (Student ID: 1009038783): In this final report, I was responsible for making an introduction to the second question. Based on the visualizations and the CI results, I made a conclusion for answering the second research question and I wrote a new discovery from the result that was not considered at the beginning.