Lab 5

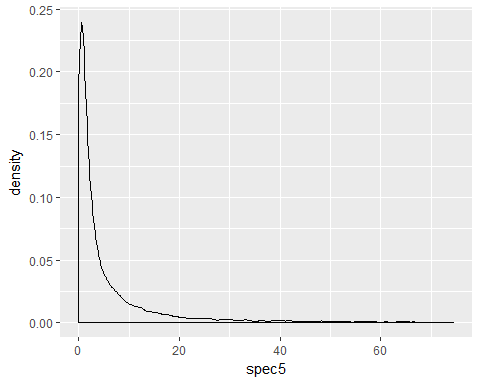
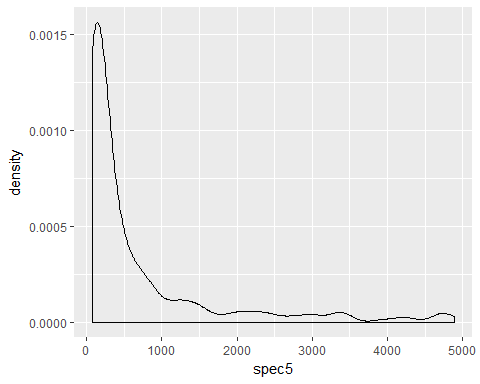
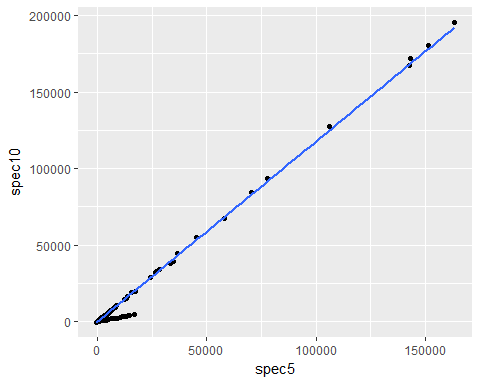
WADDS

February 13, 2018

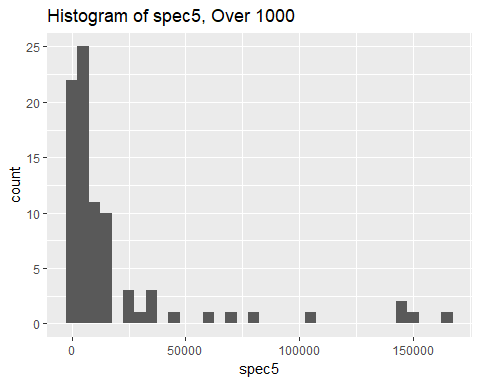
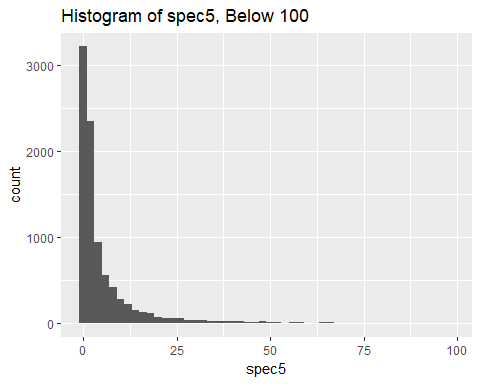
## Team Analysis

To analyze the variation within spec5 we decided to use the denisty plots created by Abby. This shows the distribution of the specs and gives us a closer look into the skewing of the distribution.

## Warning: package 'bindrcpp' was built under R version 3.3.3

 The plot above illustrates that the majority of the data lies between the spec values of 0 and 20. Over a quarter of the data is clustered between values of 0 and 5.And we see that very few spec values reach above 20.  This plot above illustrates that the distribution of the tail looks like. We can see that less than .01 percent of the data is represented at any spec greater than 75. Thus, we see that there are very few large spec values. And, we see that as the spec value increases, the amount of data at that point also decreases.  This plot above displays the relationship between the spec10 and the spec5 data. Since, in our individual plots, we showed that most of the data points are very small, and that the relationship between spec5 and spec10 are very accurately linear as spec5 and spec10 increase, we felt it would be most beneficial to examine the relationship at the small values of spec5 and spec10, specifically, the point where the data points converged into two distinct linear patterns. In our investigations, we have found that spec5 and spec10 have a very strong positive correlation.

## Sarah

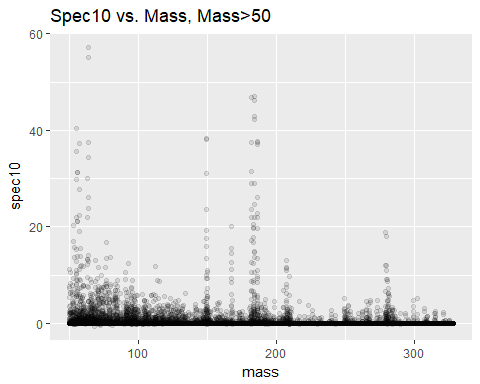
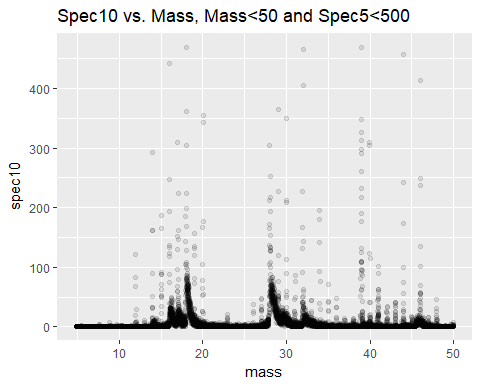
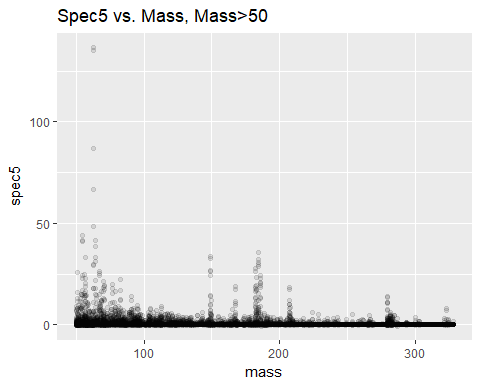
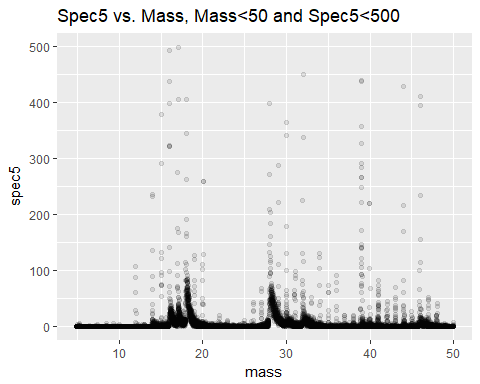
In order to visualize spec5 by itself, we plotted a histogram of it, removing small values so that the histogram would be more revealing. We made two histograms, one showing values under 100 and one showing values over 1000 so we could see the spread of both the highest and lowest values. As one can see, there is a very strong positive skew in the data, and similar can be said for spec10. 

We also can get a binned estimate of the data for both spec5 and spec10

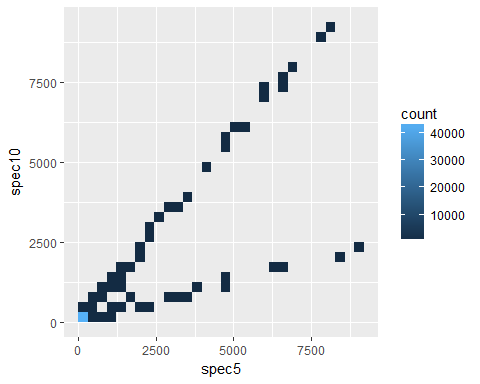
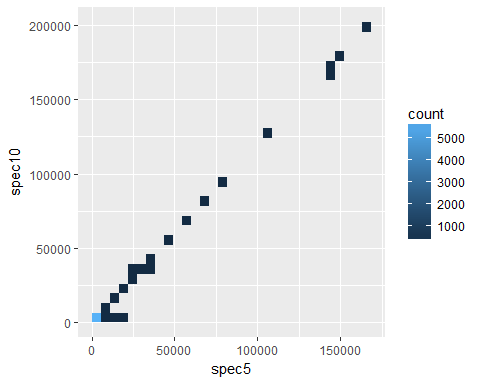
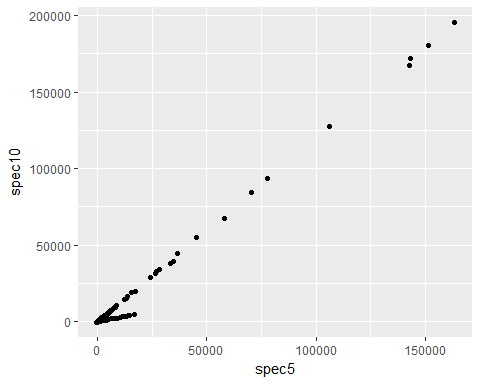
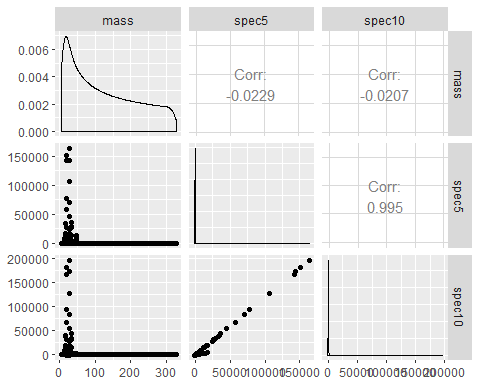
## total numNAs neg zero small med large realbig  
## 1 45312 0 893 35006 9206 123 54 25

## total numNAs neg zero small med large realbig  
## 1 45312 0 838 35599 8670 129 53 18

Now we will visualize the relationships between mass with spec5 and spec10, separately. In order to generate a more helpful visualization, I will again separate the spec10 and spec5 data into high and low values. Note that for data points with a mass of less than 50, I only plotted the spec5 and spec10 values below 500 so that the spread of the data can be seen more accurately. However, the spec5 and spec10 data points above 500 match the general trend of the points below 500.



The spec5 and spec10 variables have a correlation of 0.9953. This is a very strong positive correlation. A visualization of this correlation is included below.

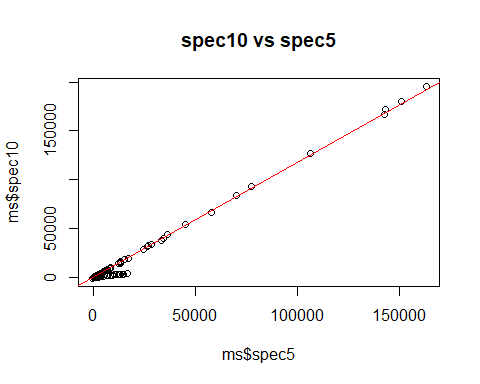


## David

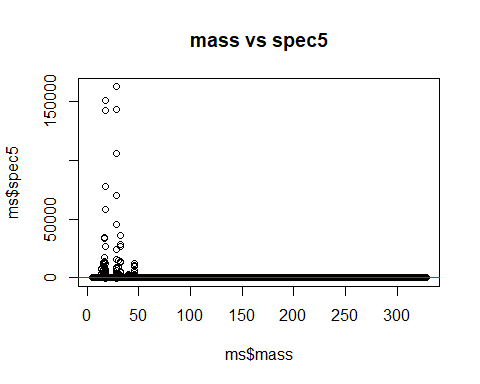
Here I have used linear regression to predict the relationship between spec10 vs spec5, and we can see that for the red line and plots in the graph. These plots are closely related to the red line I have drawn. I have calculated the covariance value betwwen spec 10 and spec 5, which is 0.9953

##   
## Call:  
## lm(formula = spec10 ~ spec5, data = ms)  
##   
## Coefficients:  
## (Intercept) spec5   
## -3.971 1.178

##   
## Call:  
## lm(formula = spec10 ~ spec5, data = ms)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15550 4 4 4 3262   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.9705345 0.9080004 -4.373 1.23e-05 \*\*\*  
## spec5 1.1776264 0.0005395 2182.634 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 193.2 on 45310 degrees of freedom  
## Multiple R-squared: 0.9906, Adjusted R-squared: 0.9906   
## F-statistic: 4.764e+06 on 1 and 45310 DF, p-value: < 2.2e-16

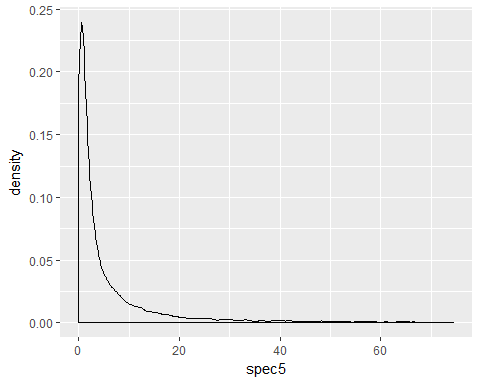


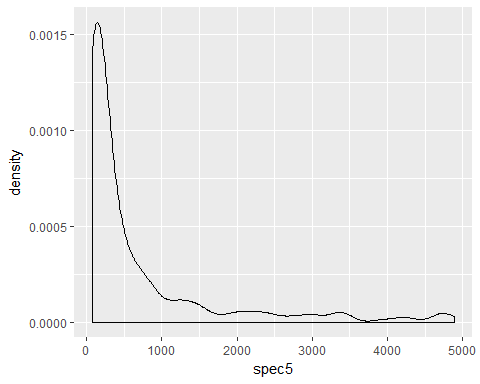
I have used scatter plot and linear regression to show the variance between mass and spec5, and I have found that the variance between mass and spec 5 is pretty small. It has only 0.1757178.



## Abby

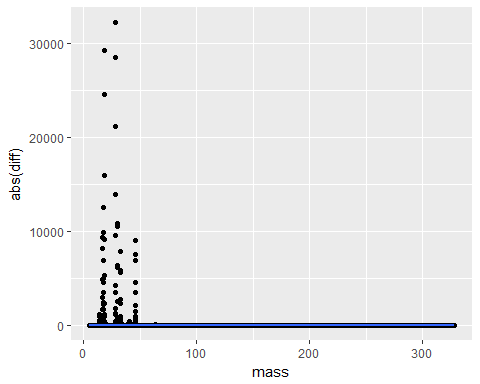
To see the variation within Spec 5 it will be helpful to see the density plot of the spectra. This will allow us to see the shape of the distribution of the data points within Spec 5. However, we will filter by values no greater than 75. Because there are so few values after the spectra reaches 75, it makes the data very hard to read where the majority of the values lie. As we can see from Sarah's plots above, there are not only very few values greater than 75, the data is positively skewed. Taking this into account we can filter out many of the values to get an idea of what the majority of the data looks like.

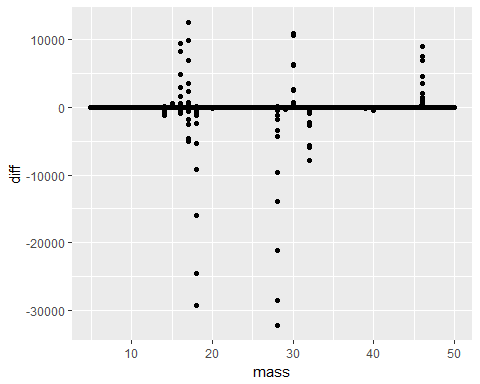
 As we can see from the plot above, most of the data lies between the spectra of 0 and 10. The data is heavily skewed to the right, yet most of the data does not have spectra values greater than 20.

 This density plot of the values of spec5 greater than 75 and less than 5000 show us what the values skewing the distribution look like in greater detail. From the plot we can see that most values in the tail are those of larger masses.And we also see how little observations there are of spectra greater than 75.The density plot peaks at around 0.0015, meaning less than 0.15 percent of the data is represented at that point. This indicates that the data is extremely clustered between the spectra of 0 and 20, while the rest of the data above 75 is very minimal and skews the data to the right.

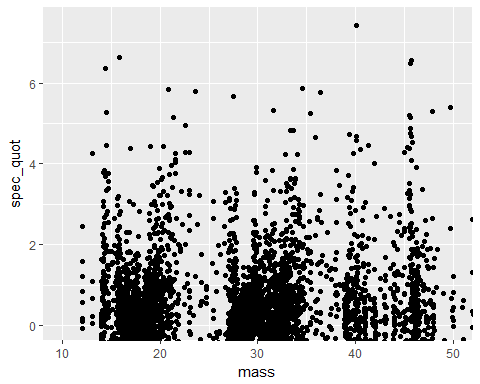
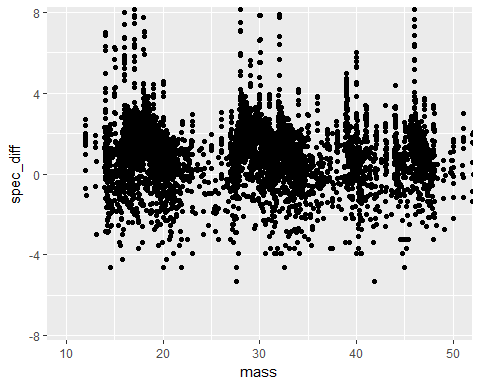
We can see how the variables spec5 and spec10 interact by looking at the difference between them. This will allow us to see how these variables may be related, as a huge difference between the two could indicate no correlation, while a small difference could indicate a strong correlation between the two variables.

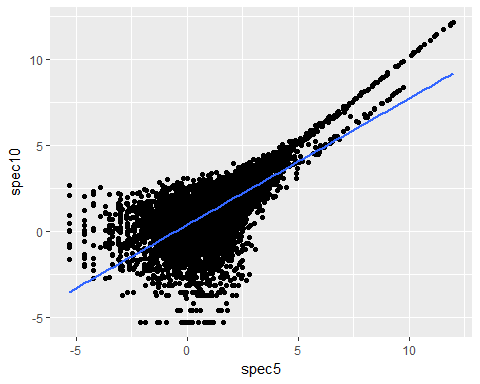
## Warning: Ignoring unknown parameters: re

 Plotting the absolute value of the differences in spectra between spec5 and spec10 shows us there are many more values clustered at the lower end of the masses. Generally, we can see that most data falls between the masses of 0 to 50, and there is still a right skew in the data, with a few data points from higher masses. We can also see that most of the differences in spectra fall under 10000, with a few differences reaching as high as 30000. We can zoom in to see what the data looks like between the masses of 0 and 50, where most of the data lies.

 As we can see from the graph above, There appear to only be certain masses that yeild spectra results in either spec10 or spec5. Since the difference is calculated by spec5 minus spec10, we can tell that at some mass levels, spec5 is greater than spec10, such as around the masses of 30 and 46. Whereas, at the masses of around 18 and 28, spec10 is greater than spec5. Since there are such big differences between the spectra readings, we can conclude that interacting the variables of spec10 and spec5 will produce different results, and change the values of the variables significantly.

## Derek





## Will

## Individual Contributions

David: I have created the convariance graph between spec5 and spec10, it could clearly see that for majority of points is very close to the line I have drawn, and the value of covariance between spec5 and spec10 is pretty large. Also I have found the interesting point that the variance value between mass and spec5 is pretty small, which means that these value are very close to each other and close to mean value.

Abby: I created density plots to show the variation within the spec5 data. At first I created a density plot with the majority of the data, to get an idea of the spread without skewing. I think created a density plot to look at what the distribution within the tail looks like. This gave me a good idea of what the data tends toward, regardless of skewing. I then looked at the covariation between spec5 and spec10 by looking at their differences. Through a general picture I saw that the majority of data points were again concentrated at the lower end of the masses. Zooming in on these points gave me a good idea of how the values differed at given masses, and how these variables may change when interacted with each other.

Sarah: I made jitter plots to analyze the relations between mass and the spec5 and spec10 data. I sectioned the data points between higher and lower values of mass, along with lower values of spec5 and spec10 in order to give a fuller view of the spread. I also made the data points translucent so that it could be more easily seen where there are high build ups of data points. In order to plot the covariation between the spec5 and spec10 data, I made scatter plots similar to in the outline of the assignment, along with a plot that utilized bin2d to show the density of points in their correlation.