Exploratory Data Analysis and Regression of Wheat Seeds

Assignement 2

Rebecca Leu

2020

**Introduction**

The following is an exploratory data analysis and regression of the Wheat Seeds dataset found in Kaggle. The dataset consists of 199 observations each with 8 variables. I chose this dataset because it was easy to work with the measurements provided to hopefully be able to predict what type of seeds they were based off these measurements. There were also several observations and variables to look at.

**The initial look at the data**

After loading in the dataset from the .xlsx which was downloaded off Kaggle, we begin by looking at a summary of the data and see what category names there are. The variables are as follows:

> names(seeds)

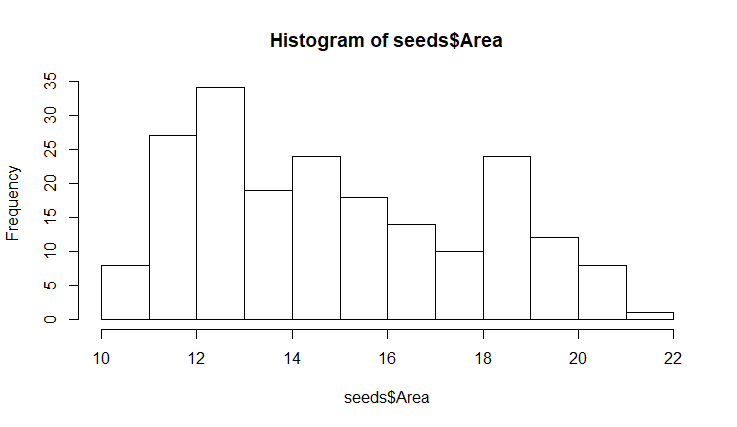
[1] "Area" "Perimeter" "Compactness" "Kernel.Length"

[5] "Kernel.Width" "Asymmetry.Coeff" "Kernel.Groove" "Type"

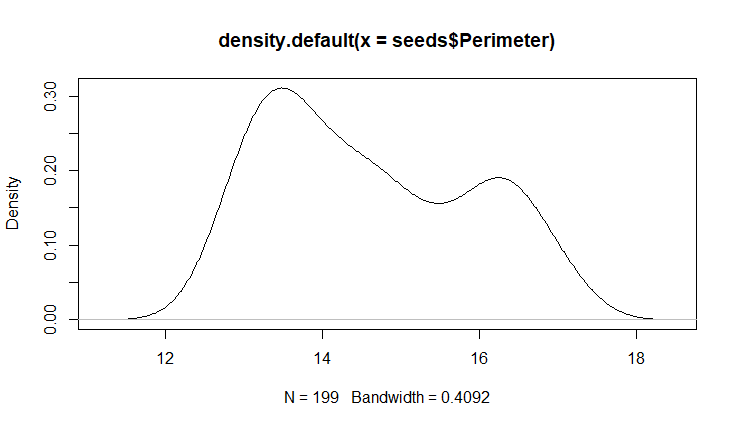
I also completed some additional explorations using the summary command.

After that, I moved on to histograms and density plots of some of the variables.

hist(seeds$Area)

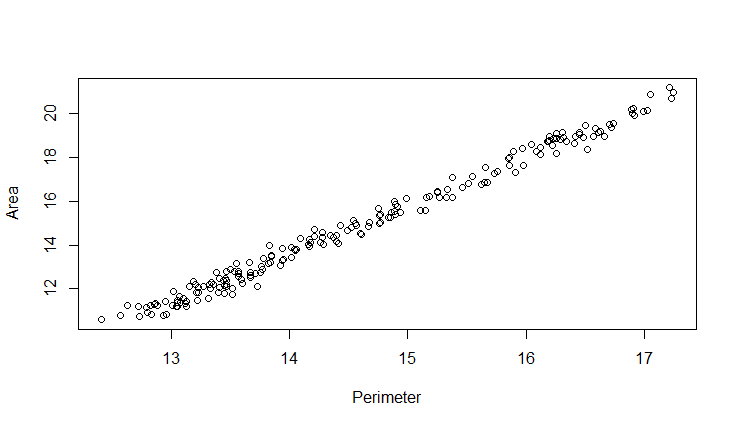


plot(density(seeds$Perimeter))



I also looked at plots comparing two attributes. I was able to see that some of these attributes were highly correlated, like the perimeter and area of the seeds.

plot(Area~Perimeter, data=seeds)

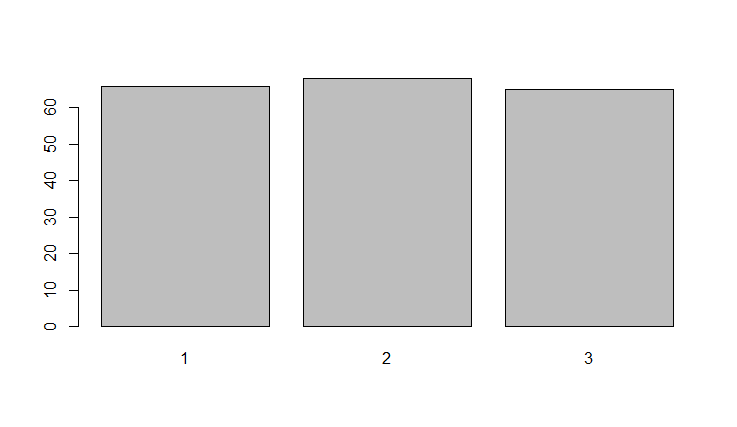


Before beginning the regression analysis, I did want to note about how many of each type of seed we were looking at. There are 3 types in this dataset and a pretty even split of each

1 2 3

66 68 65

> barplot(table(seeds$Type))



One way to begin to look at correlation is to run a correlation test. I already knew that area and perimeter would be highly correlated so I tried that first. Sure enough, the values were very close to 1.

> cor(seeds$Area, seeds$Perimeter)

[1] 0.9944373

I also looked at two other attributes that I thought might not be highly correlated and tried this:

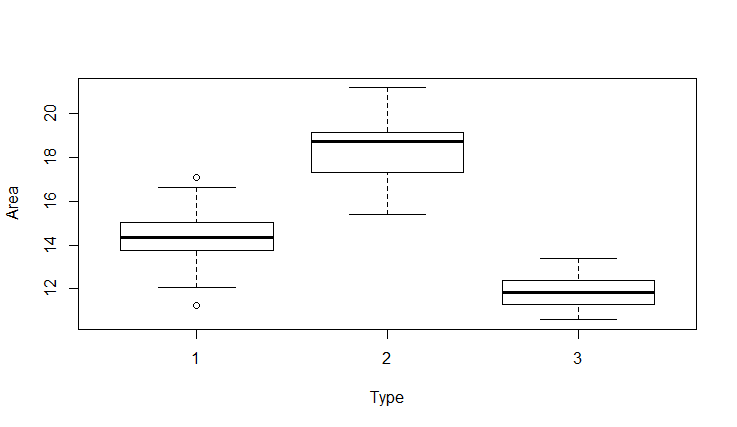
> cor(seeds$Compactness, seeds$Kernel.Groove)

[1] 0.2270165

Again, I was correct in my assumptions and these two attributes were not highly correlated.

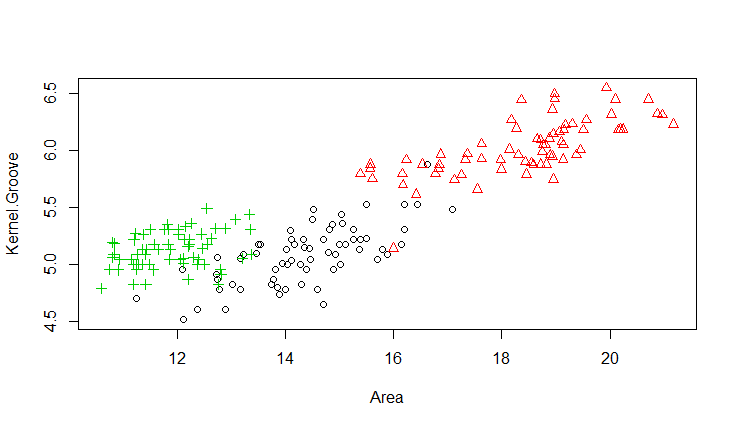
Another interesting plot I found while completing my exploratory analysis was this box plot of area and type of seed. The seed types are evenly split based on area alone. This tells me that we might be able to predict the seed type based off this one attribute pretty well.

boxplot(Area ~ Type, data = seeds, xlab="Type", ylab = "Area")

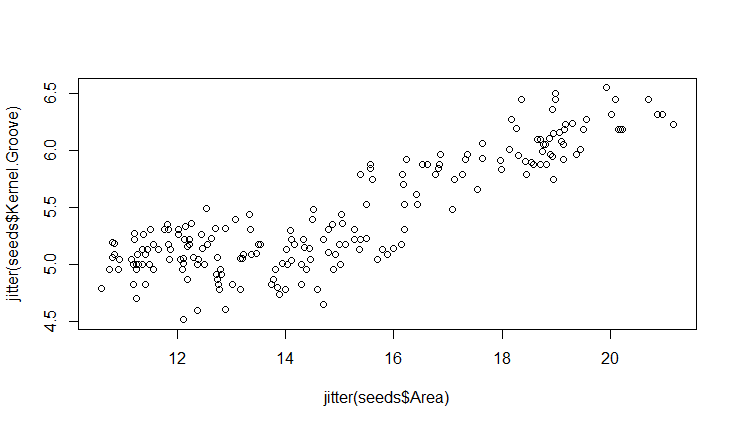


I also looked at scatter plots, adding jitter and smoothing the scatters

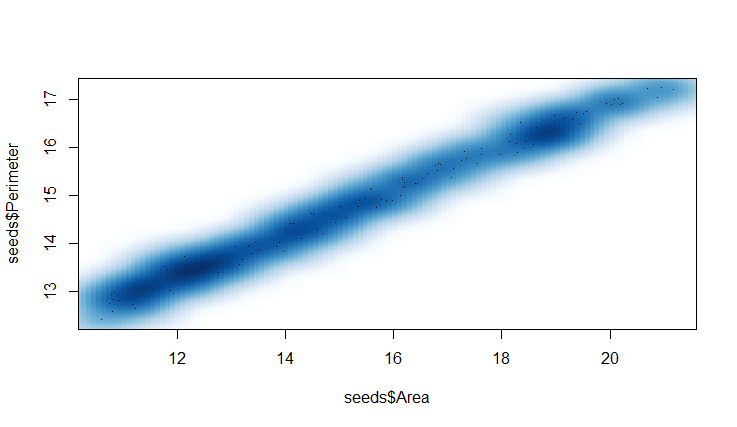
with(seeds, plot(Area, Kernel.Groove, col=Type, pch=as.numeric(Type)))



plot(jitter(seeds$Area), jitter(seeds$Kernel.Groove))



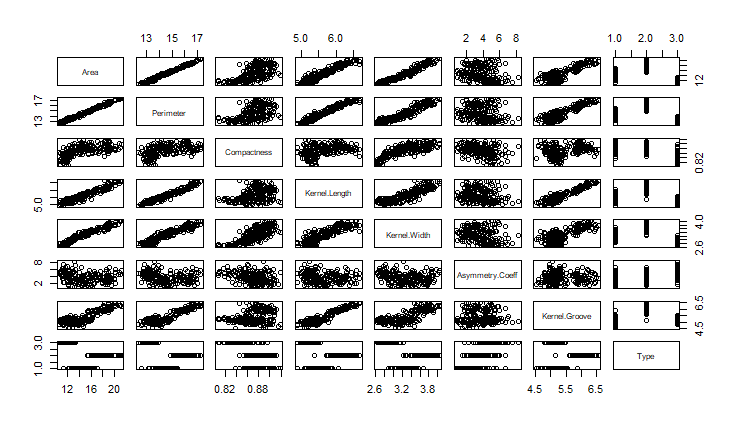
smoothScatter(seeds$Area, seeds$Perimeter)



Although it might be nice to look at these plots separately, I find that looking at all the attributes and how they interact with each other all in one matrix gives us the best snapshot of each of the attributes all at once.

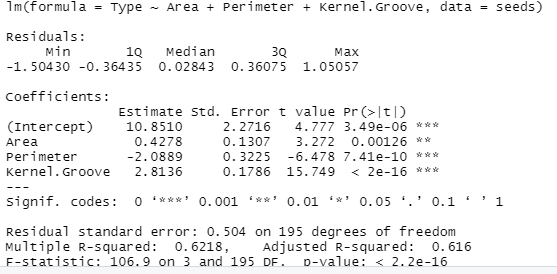
pairs(seeds)

Looking at this plot matrix, I can tell that Compactness and Asymmetry Coefficient seem to be the least correlated with each other, or with the seed types. Based off of these plots alone, it is my assumption that Area, Perimeter, and Kernel Groove might be able to predict the type of seed most.

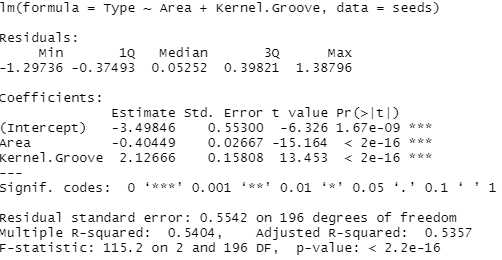


**Regression**

I have decided to start the regression by looking at the 3 attributes I thought might tell us the seed type best based off the above graphs. Here is the summary of that regression:



The p values are all incredibly significant from this regression. I do know that perimeter and area are highly correlated. For that reason, I am going to remove the perimeter.



With that, I am confident we have a good regression model, and we will get a great prediction of type of seed based off the area of the seed and the kernel groove alone.

References:

Caro, J. (2018, October 09). Wheat-Seeds. Retrieved November 07, 2020, from https://www.kaggle.com/jmcaro/wheat-seedsuci?select=seeds.csv