STAT 602 Final Project Smokeless Gunpowder Analysis

Divanshu Mittal, Angela Rose, Jacob Liester, Anna Leisa Sauser & Hacene Salmi

4/30/2023

Overview:

In this research, we are doing an analysis of smokeless propellant (powder) used in the creation of small arms. The main purpose of this analysis is to investigate multiple recovered samples from exploded and unexploded IED's to determine the brand of each sample and then compare the results. This analysis consists of two main approaches:

1. Data exploration of the provided train data and the recovered samples using bar & box plots, histograms, Anova test, and Kruskal-Wallis's analysis.
2. Building, comparing, and selecting the best multi-class model to predict the brands. The following classification models are used in the analysis:

- LDA
- QDA
- Random Forest
- MClustDA

In addition to the above, we compared the results of the recovered sample 1 and Recovered sample 2 to find out if they are from the same brand or not. Further, for recovered sample 3 and recovered sample 4, we did a separate analysis to find out if more than one brand is used by the manufacturers for making the IEDs.

Data Description:

The given training data set has twelve variables, which consist of eight numeric variables and three class variables. The numeric variables are Area, Perim., Major, Minor, Circ., AR, Round, and Solidity and the class variables are Distributor, Brand, and Shape. The data has no missing values. The data contains nine unique distributors, 154 unique brands, and four unique shapes. For data cleaning, we removed the index variable (x).

In addition to the train data set, we were provided with four recovered SAP samples from exploded and un-exploded IEDs. These samples contain only the predictor variables Area, Perim., Major, Minor, Circ., AR, Round, and Solidity.

Exploratory Data Analysis:

We started exploratory data analysis by building the bar plot and boxplots. Since there are 154 brands, we used the shape variable as a group to compare the brands across different shapes. We create bar plots for each numeric predictor variable versus Shape to get a better insight into the distribution of data. The boxplots showed that there are numerous outliers in the data and a good amount of separation for the two predictors Major and Minor among shapes categories which means no stronger association with the shape variable while the other variables showed a stronger association.

The bar plot below illustrates the distribution of unique brands among shapes of particles. The cylindrical shape has the highest value followed by flattened_spherical, flake, and spherical shapes.

Data Split:

We split the data into 70/30 split by each brand and created a training and test set. We also created a validation set to hyper-tune the parameters of the classification models. The training set consists of 27957 rows while the test data consists of 11987 rows and the validation set consists of 2393 rows.

Statistical Analysis Techniques:

We used two techniques to analyze the similarities of the recovered datasets.

Analysis of Variance (ANOVA)

Analysis of Variance is a technique that is used to compare the variance in means of the different groups to see if the groups are statistically similar or different. It also analyzes variability within each group. With the null hypothesis being that the recovered samples populations are the same, the Analysis of Variance test returned a very small p-value, indicating that we should reject the null hypothesis, meaning that the difference between the samples is statistically significant. We utilized Tukey's test to validate the results.

Kruskal-Wallis

The Kruskal-Wallis test is a test similar to ANOVA. This test is used when the assumptions for ANOVA, like normality, are not met. Kruskal-Wallis uses ranks to calculate a chi-squared test statistic. Similar to our ANOVA test, with the null hypothesis being that the sample populations are the same, the Kruskal-Wallis test returned a very small p-value, indicating that we should reject the null hypothesis, meaning that the difference between the samples is statistically significant. We utilized Dunn's test to validate the results.

Multi-Classification Techniques:

Linear Discriminant Analysis

In LDA, the goal is to find a straight decision boundary that best separates different classes of data. Since there are more than two classes of Brands in the training set and collinearity in the predictor variables, we started our analysis using the LD model.

We tested three LDA utilizing different transformations of variables and found our best total brand accuracy to be around 29 percent. It predicted nine brands with greater than 80 percent accuracy but does have a 48 percent absolute misclassification rate meaning it was not able to classify some of the brands at all.

Quadratic Discriminant Analysis

In QDA, a quadratic equation is used to find a decision boundary that separates different classes of data. Using the variables from the most accurate linear discriminant analysis model, our best total brand accuracy was about 22 percent. It predicted nineteen brands with greater than 80 percent accuracy but does have a 51 percent absolute misclassification rate, which is higher than the linear discriminant analysis.

Random Forest Classification

Random forest classification uses many decision trees to create a 'forest'. Each decision tree is made up of nodes (or leaves) and branches. When the data travels through a branch and arrives at a node, that leaf categorizes and splits the data based on the question it is fed. These new groups/sets of data are then sent along another branch to another node, where it is again categorized and split into more groups. This process creates a tree-like structure. Each tree outputs its recommended prediction, and an overall vote is taken to choose the classification prediction.

With a minor tweak in the interactions from our linear discriminant analysis, the random forest model resulted in a total brand accuracy of around 32 percent. It predicted nine brands with an accuracy greater than 80 percent and has a 26 percent absolute misclassification rate, which is lower than both the linear and quadratic discriminant analysis.

Model-Based Clustering

Model-based clustering is a statistical approach used to group data points into clusters based on a particular probability distribution, either normal or a mixture of normal. In cluster-based modeling, the number of groups is not predetermined. For this analysis, we used MclustDA model (Discriminant analysis based on Gaussian finite mixture modeling) in which the model is fitted to the data to estimate the parameters of the distribution using maximum likelihood estimation to create clusters.

The MclustDA total brand accuracy was about 24 percent. It predicted sixteen brands with greater than 80 percent accuracy but does have a 55 percent absolute misclassification rate, which is higher than both linear and quadratic discriminant analysis and higher than random forest model.

Comparison of Models

We compared the LDA, QDA, Random Forest, and MclustDA models and found that the Random Forest model has the best overall accuracy and the least amount of misclassification. We then used the Random Forest model for making predictions.

Results validation technique:

To confirm our predictions for recovered samples we also created a Random Forest model for shape using train data and the variables used for predicting the brands in the selected model above. Our accuracy for shape was 90 percent. The model for shape was helpful in validating our results.

Predictions/ Results:

Part 1: Sample 1 & Sample 2 Analysis: Comparing the samples to find out if they are from the same brand or from different brands and then finding the brand name.

Methodology:

We first applied the selected random forest model on Sample 1 & Sample 2 to predict the brands. Then we took the maximum occurrence of a brand in both samples and compared the results. After that, we validated our results by implementing the LDA model on both samples. In addition to this, we took an extra step to validate our results by predicting shapes using a different random forest model we created in the analysis just for predicting the shapes of the samples.

Recovered Sample 1:

Based on the results from the selected Random Forest model, recovered sample 1 has the highest predicted value of 164 and is from Brand Reddot. We validated sample 1 results by implementing the LDA Model and the predicted value of Reddot is the highest with LDA as well. Then we predicted the shape of the recovered sample 1 and the flake has the highest predicted value. We checked the shape of Reddot and it is Flake from the train data, this confirms our results that the recovered sample 1 is from the brand Reddot.

Recovered Sample 2:

According to the results from the selected Random Forest model, recovered sample 2 has the highest predicted value of 65 and is from Brand Reddot. We validated sample 2 results by implementing the LDA Model and the predicted value of Reddot is the highest with LDA as well. Then we predicted the shape of the recovered sample 2 and the flake has the highest predicted value. This confirms our results that the recovered sample 2 is from the brand Reddot.

Part 2: Smokeless Gun Powder presence of multiple Brand Analysis in Sample 3 and Sample 4.

Methodology:

Similarly, like part 1 we first applied the selected random forest model on Sample 1 & Sample 2 to predict the brands. Then we took the five maximum occurrences of the brands in both samples and compared the results. After that, we validated our results by implementing the LDA model on both samples. In addition to this, we took an extra step to validate our results by predicting shapes using a different random forest model we created in the analysis just for predicting the shapes of the samples.

Recovered Sample 3:

Based on the results from the selected Random Forest model, recovered sample 3 has the highest predicted value of 61 and is from Brand RamshotEnforcer. We validated sample 3 results by implementing the LDA Model and the predicted value of RamshotEnforcer is the highest with LDA as well. We checked sample 3 for the presence of other brands and both Random Forest and LDA showed the presence of other brands AmericanSelect, AccurateNo.2 and Accurate4100. Then we predicted the shape of the recovered sample 3 and the sample have a spherical and flake shape. This confirmed our analysis that manufacturers are using multiple brands in the recovered sample 3, but the majority of particles are from the brand RamshotEnforcer.

Recovered Sample 4:

According to the results from the selected Random Forest model, recovered sample 4 has the highest predicted value of 71 and is from Brand RamshotEnforcer. We validated sample 4 results by implementing the LDA Model and the predicted value of RamshotEnforcer is the highest with LDA as well. We checked sample 4 for the presence of other brands and both Random Forest and LDA showed the presence of other brands AccurateNo.2, Accurate4100 & BL-C(2). Then we predicted the shape of the recovered sample 4 and the sample have spherical and flattened_spherical shape. This confirmed our analysis that manufacturers are using multiple brands in the recovered sample 4, but the majority of particles are from the brand RamshotEnforcer.

Conclusion:

The results above indicate that Recovered Sample 1 and Sample 2 are from the same brand ("RedDot"). The table for samples 3 and 4 shows that manufacturers are using multiple brands in the samples. For sample 3, the presence of the following brands Accurate4100, AccurateNo.2, AmericanSelect & RamshotEnforcer are there and in sample 4 Accurate4100, AccurateNo.2, BL-C(2) & RamshotEnforcer brands are found. The majority of the particles are from the brand "RamshotEnforcer" in sample 3 and sample 4 and the common brands in both sample 3 and sample 4 are Accurate4100, AccurateNo.2 & RamshotEnforcer.

Importing Libraries

Step1 : Loading the given train data, Missing values Check, Summary of the data & Vizually checking the data.

```
## 'data.frame':    39944 obs. of  11 variables:
##  $ Distributor: Factor w/ 9 levels "Alliant","Hodgdon",..: 9 9 9 9
9 9 9 9 9 9 ...
##  $ Brand      : Factor w/ 154 levels "0.41","20/28",..: 4 4 4 4 4 4
4 4 4 4 ...
##  $ Shape      : Factor w/ 4 levels "cylindrical",..: 3 3 3 3 3 3 3
3 3 3 ...
##  $ Area       : num  273031 213566 297572 254810 237245 ...
##  $ Perim.     : num  1964 1771 2070 1956 2048 ...
##  $ Major      : num  601 614 627 634 655 ...
##  $ Minor      : num  578 443 604 512 461 ...
##  $ Circ.      : num  0.89 0.856 0.872 0.837 0.711 ...
##  $ AR         : num  1.04 1.38 1.04 1.24 1.42 ...
##  $ Round      : num  0.961 0.722 0.963 0.807 0.704 ...
##  $ Solidity   : num  0.978 0.974 0.98 0.97 0.916 ...

## [1] 0

##      Distributor                Brand                       Shape
##  Hodgdon   :12493   H335           : 2704   cylindrical
:13665
##  Alliant   : 9135   BL-C(2)        : 2389   flake              :
6705
##  Western   : 8524   Bullseye       : 1425
flattened_spherical:16360
##  IMR       : 4123   Reloader23     :  843   spherical          :
3214
##  Winchester: 2417   RamshotEnforcer:  745
##  VihtaVuori: 2387   RedDot         :  613
```

```
##  (Other)    :  865   (Other)        :31225
##       Area             Perim.            Major            Minor
##  Min.   :  10104   Min.   :  399.5   Min.   : 128.9   Min.   :
46.3
##  1st Qu.: 286247   1st Qu.: 2019.1   1st Qu.: 627.7   1st Qu.:
575.8
##  Median : 590779   Median : 3089.6   Median : 949.1   Median :
784.5
##  Mean   : 847612   Mean   : 3414.8   Mean   :1099.5   Mean   :
832.4
##  3rd Qu.:1341502   3rd Qu.: 4833.9   3rd Qu.:1544.1   3rd
Qu.:1088.4
##  Max.   :6145767   Max.   :13114.7   Max.   :5451.8   Max.
:2226.9
##
##      Circ.              AR             Round            Solidity
##  Min.   :0.2520   Min.   :1.000   Min.   :0.1500   Min.   :0.5715
##  1st Qu.:0.7152   1st Qu.:1.046   1st Qu.:0.6979   1st Qu.:0.9580
##  Median :0.8461   Median :1.127   Median :0.8870   Median :0.9730
##  Mean   :0.8012   Mean   :1.289   Mean   :0.8195   Mean   :0.9656
##  3rd Qu.:0.8928   3rd Qu.:1.433   3rd Qu.:0.9562   3rd Qu.:0.9800
##  Max.   :0.9680   Max.   :6.650   Max.   :1.0000   Max.   :0.9928
##
```

Based on the summary output, we can see that the data set consists of 39944 observations and 12 variables distributed as three categorical variables of type factor and eight numeric variables. The data has no missing values, and it contains nine unique distributors, 154 unique brands, and four unique shapes. For data cleaning, we removed the index variable (x).

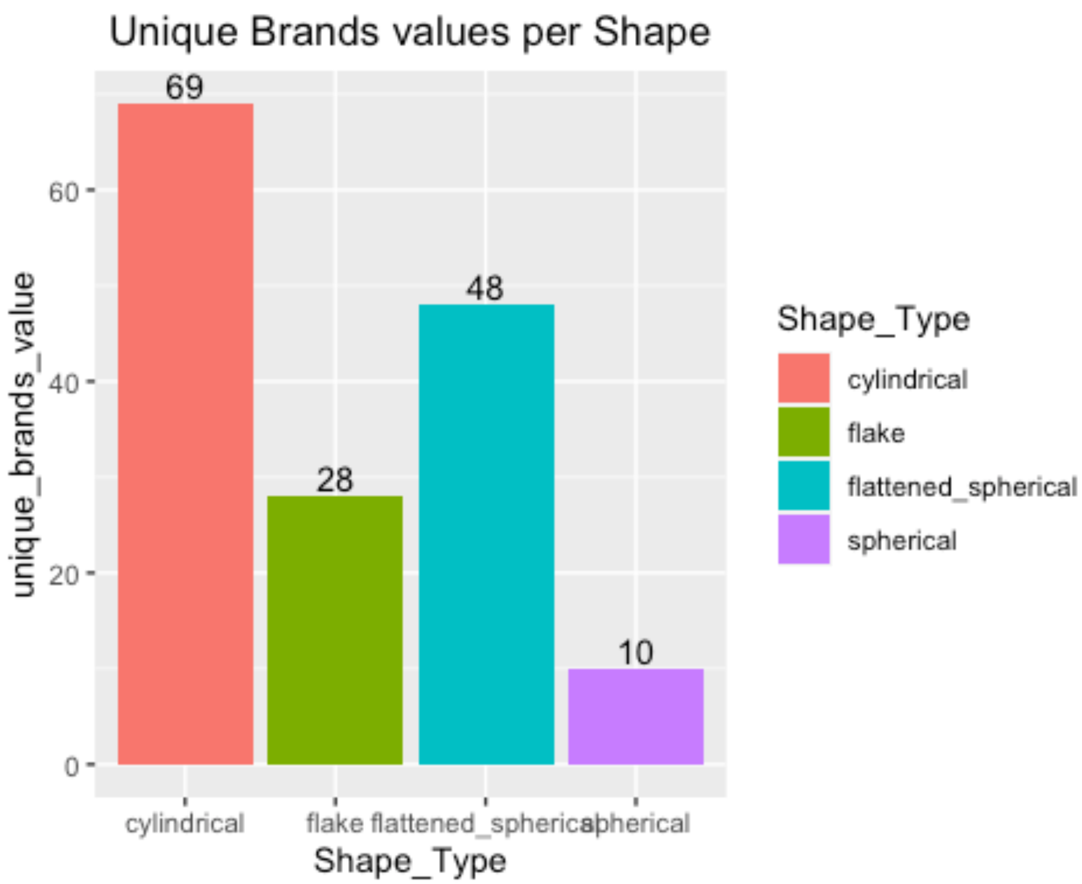
Sub-setting input data based on Shape variable for exploratory data analysis on brands within a Shape. Since there are 154 brands it will be hard to do exploratory data analysis on brands in one step, so we divided the brands by shape and explore the measurements of the brand within a particular Shape.

```
## [1] 13665    11

## [1] 6705   11

## [1] 16360    11

## [1] 3214   11
```

Step2 a: Exploratory Data Analysis for sap.train data

Exploring the data using Box plots.

Barplot of unique brands within a Shape (Shape Vs Brand Analysis)

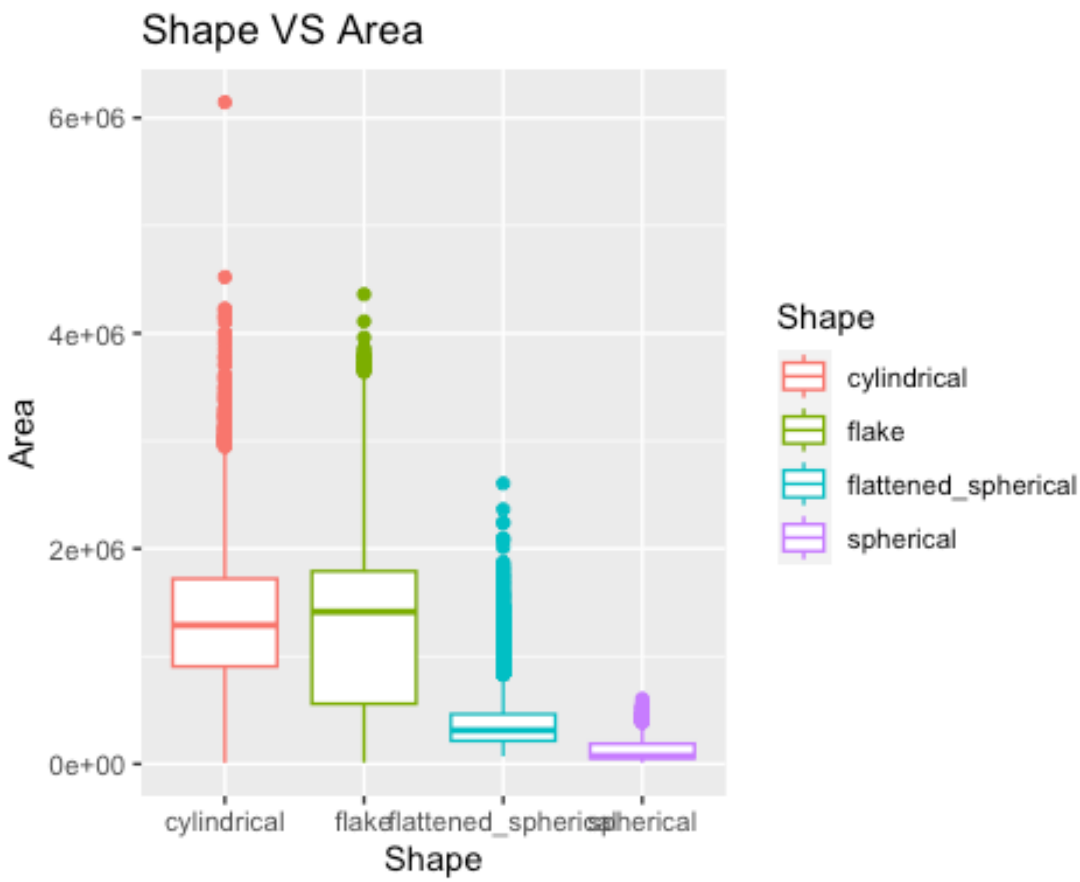


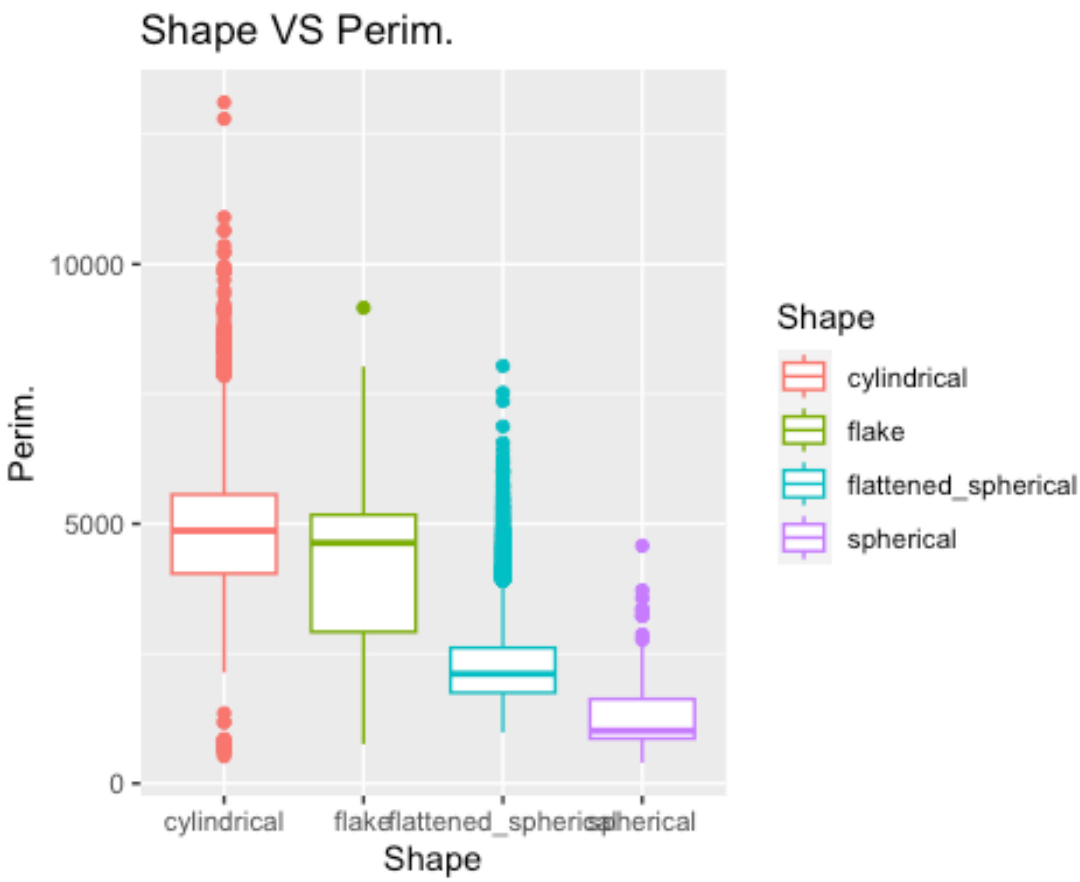
The above bar plot shows that the unique brand values belonging to cylindrical and flattened_spherical shapes are greater than flake and spherical.

Ref:

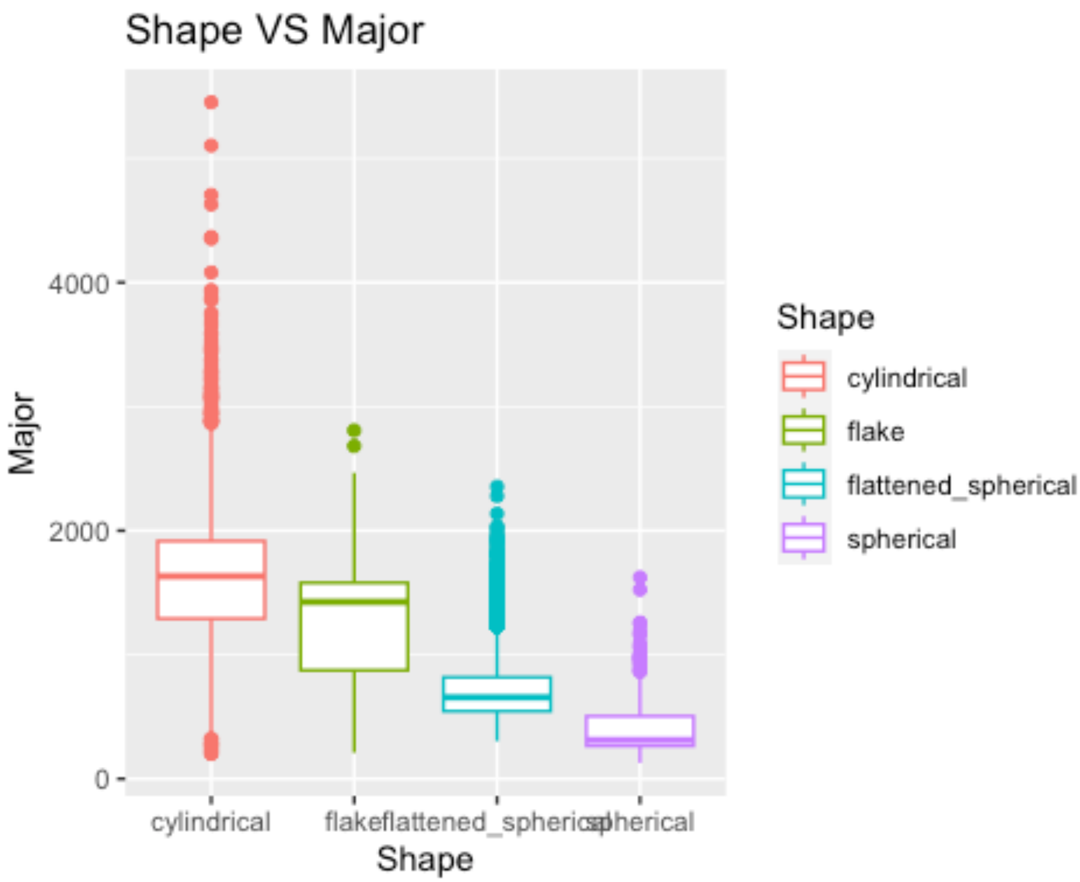
- ggplot2 barplots : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA. (2019). Sthda.com. http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization

Step2b: Box Plots of Measurements of brands within a Shape.

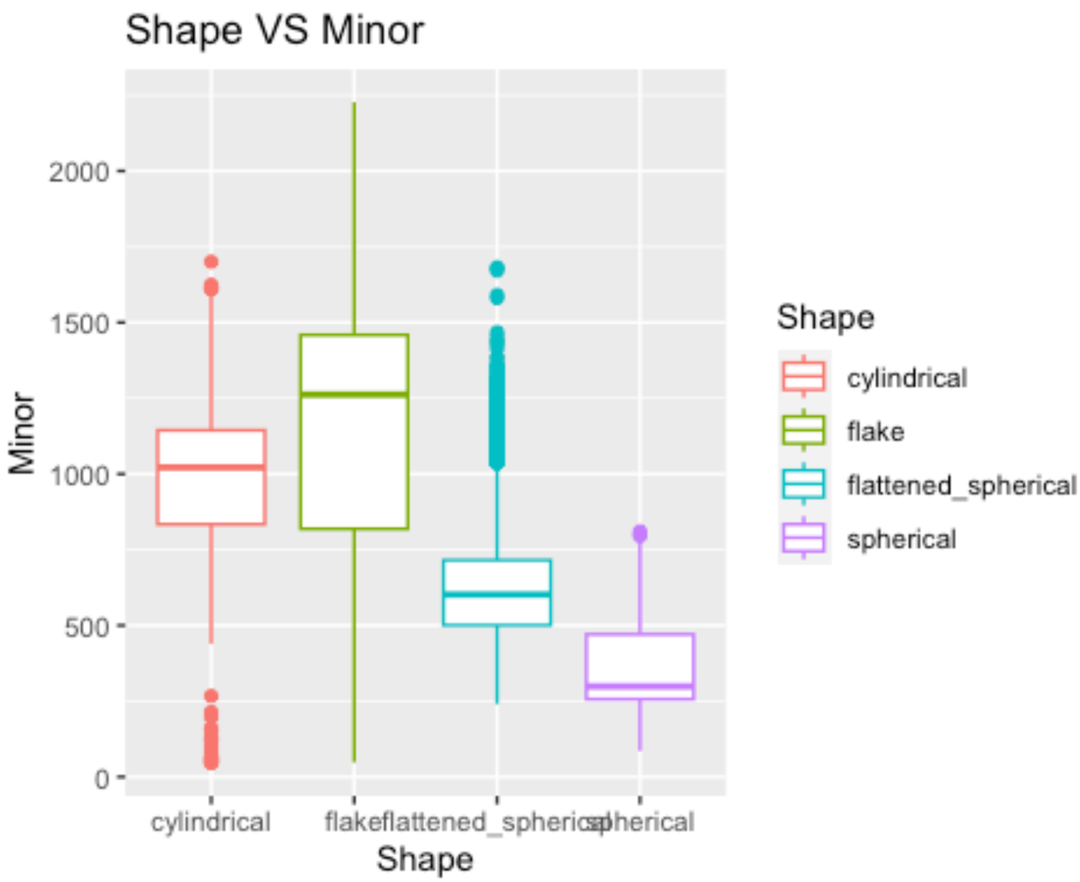

Shape VS Area
6e+06
4e+06
2e+06
0e+00
Area
cylindrical
flake
flattened_spherical
spherical
Shape
Shape
cylindrical
flake
flattened_spherical
spherical



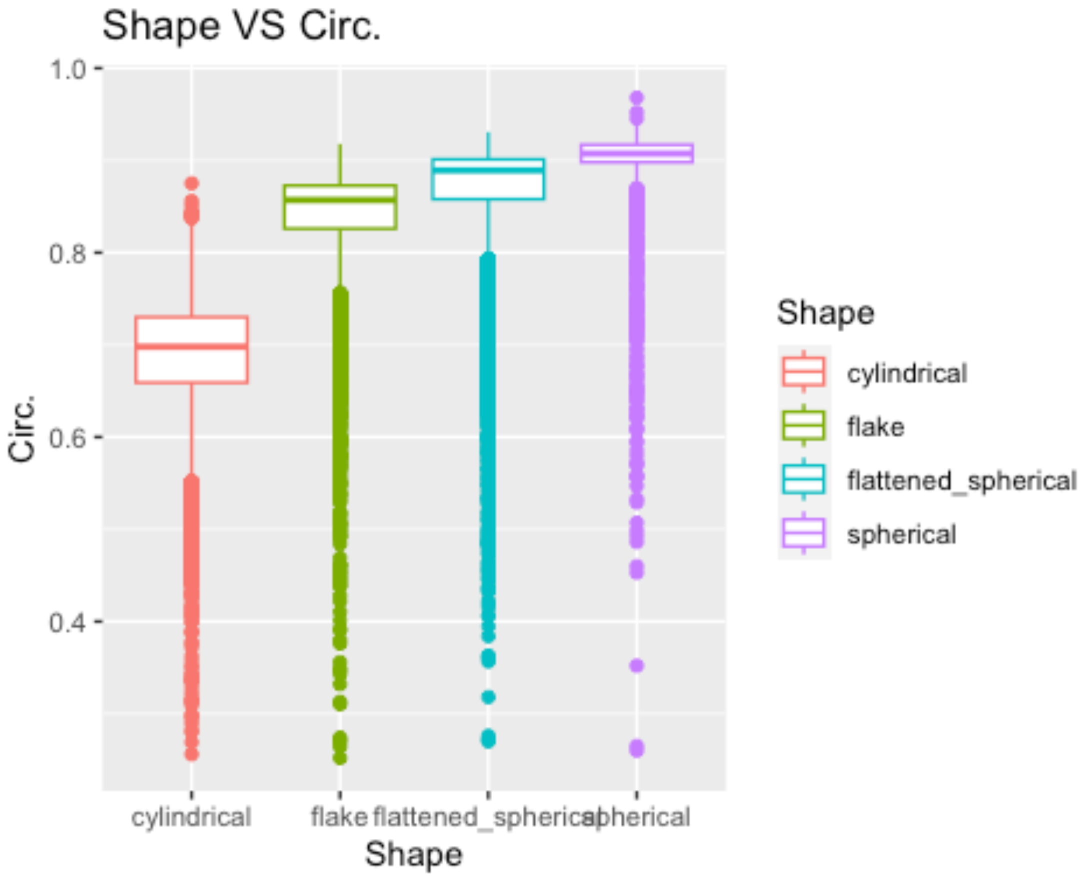
Shape VS Perim.
10000
5000
0
Perim.
cylindrical
flake
flattened_spherical
spherical
Shape
Shape
cylindrical
flake
flattened_spherical
spherical

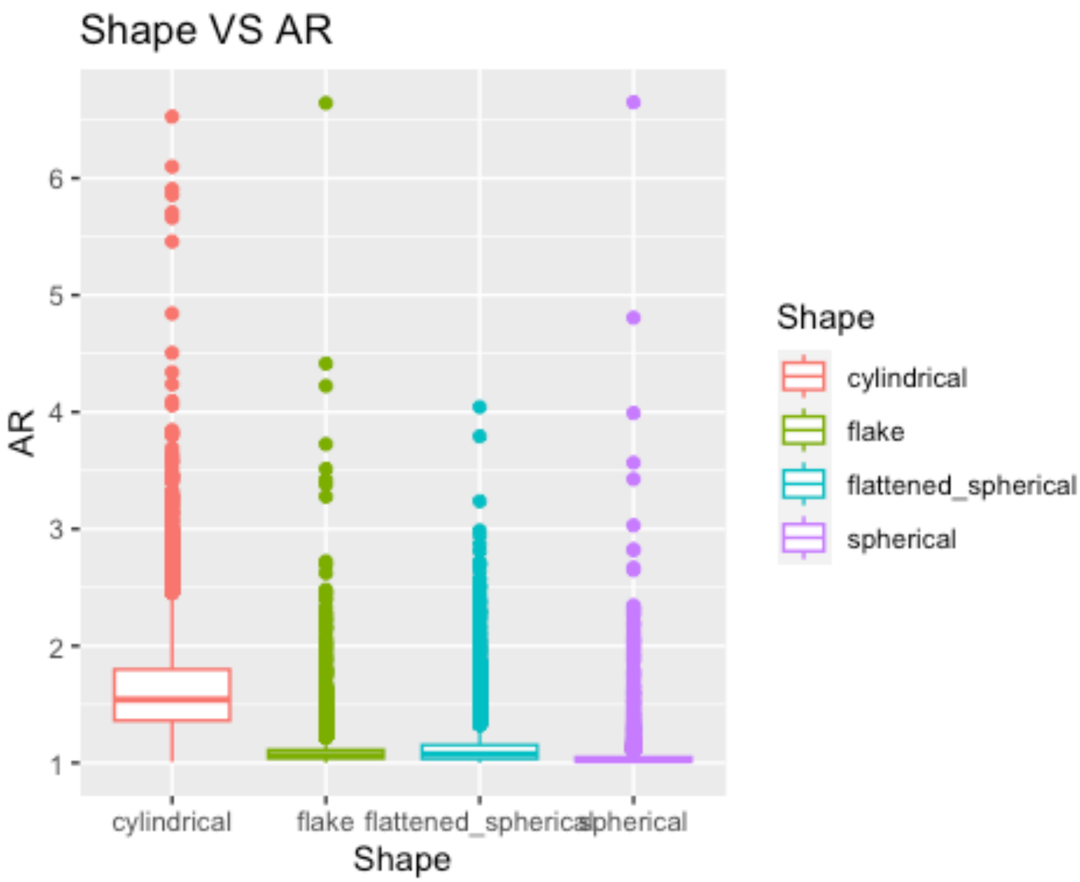


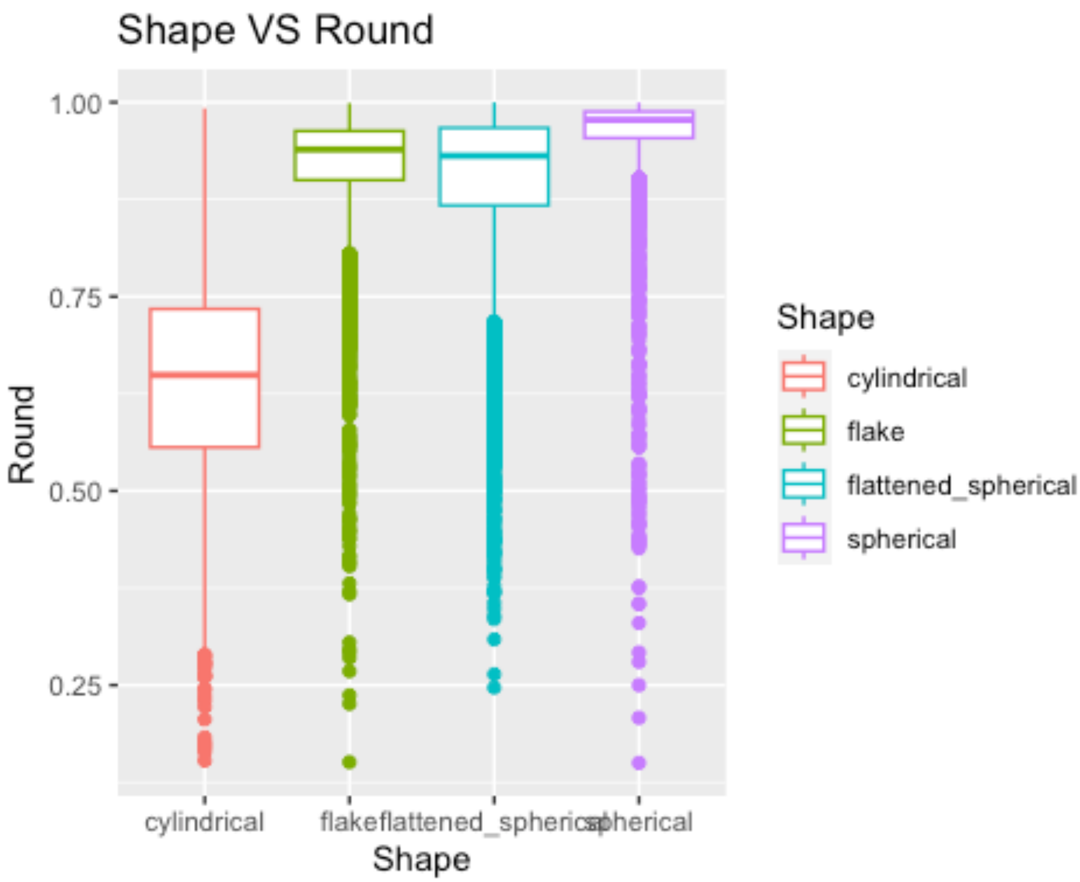
Shape VS Major
4000
2000
0
Major
cylindrical
flake
flattened_spherical
spherical
Shape
Shape
cylindrical
flake
flattened_spherical
spherical

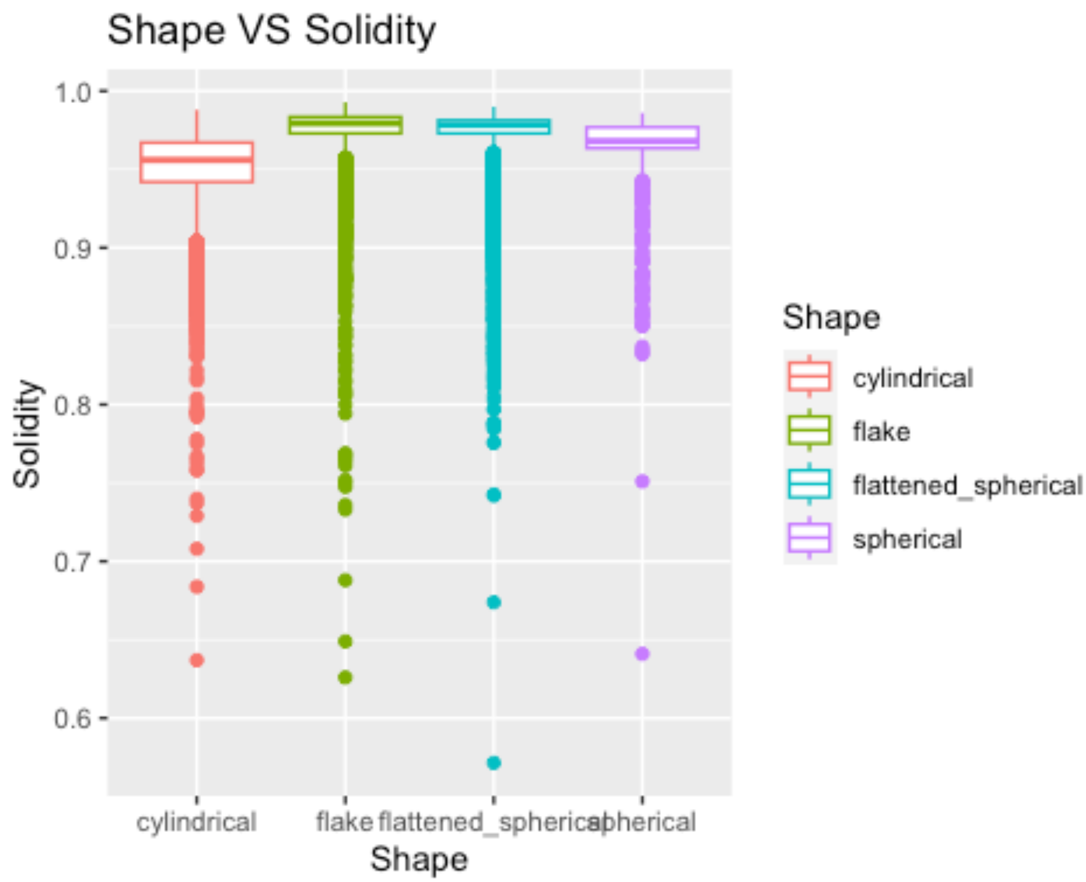


Shape VS Minor
Minor
2000
1500
1000
500
0
cylindrical
flake
flattened_spherical
spherical
Shape
Shape
cylindrical
flake
flattened_spherical
spherical


Shape VS Circ.
1.0
0.8
0.6
0.4
Circ.
cylindrical
flake
flattened_spherical
spherical
Shape
Shape
cylindrical
flake
flattened_spherical
spherical


Shape VS AR
AR
6
5
4
3
2
1
cylindrical
flake
flattened_spherical
spherical
Shape
Shape
cylindrical
flake
flattened_spherical
spherical


Shape VS Round
1.00
0.75
0.50
0.25
Round
cylindrical
flake
flattened_spherical
spherical
Shape
Shape
cylindrical
flake
flattened_spherical
spherical



From the above box plots, we can see there are numerous outliers in the data. For the most part, there is not much separation in the box plots which shows that there is a stronger association with the class variable Shape except for the two predictors Major and Minor which show a good amount of separation among shapes categories.

Ref:

- Wickham, H., Chang, W., & Henry, L. (n.d.). A box and whiskers plot (in the style of Tukey). ggplot2.tidyverse. Retrieved 10 March 2023, from https://ggplot2.tidyverse.org/reference/geom_boxplot.html
- Statistics Globe (n.d.). Draw multiple Boxplots in one graph.statisticsglobe .Retrieved 10 March 2023, from https://statisticsglobe.com/draw-multiple-boxplots-in-one-graph-in-r

Step 3- Hypothesis Testing

Importing the four recovered samples.

Analyzing Samples 1 & 2

Summary of Exploration of Area Variable

Running both ANOVA and Kruskal-Wallis, and having low values for both in this case, helps strengthen the case that there is evidence of a difference between groups. The evidence against the null hypothesis is strong.

ANOVA p-value: 3.39e-15 Kruskal-Wallis p-value: 2.2e-16

Differences between the p-values exist because of the different mathematical components of ANOVA v. Kruskal-Wallis, but similar p-values with both tests help strengthen our case.

We used Tukey Honestly Significant Difference test after ANOVA to compare the means of all the possible pairs of groups to determine which are significantly different from one another. It controls the family-wise error rate, which is the probability of making at least one type 1 error across the pairwise comparisons.

For Area, the Tukey results tell us that there is a statistically significant difference between the means of Group 2 and Group 1, with Group 2 having a higher mean than Group 1. We can see this echoed in the box plot.

We also used Dunn's test, as a follow-up to our Kruskal-Wallis analysis. Dunn's test has a similar goal of comparing all possible pairs of groups and determining which pairs are significantly different from each other. We used the Holm-Bonferroni correction to adjust the p-values. Again, a very significant difference is shown between Group 1 and Group 2, and the test is highly significant. (See line 15)

Summary of Exploration of Perimeter Variable

Similar to Area, Perimeter has a low p-value and our subsequent analyses show significant differences between the two groups.

ANOVA p-value: 3.39e-15 Kruskal-Wallis p-value: 2.2e-16

The difference between those two groups is clearly shown in the boxplot, as well.

Summary of Exploration of Major Variable

In the case of Major, the difference between the two groups is less, but still statistically significant. ANOVA p-value: 9.09e-11 Kruskal Wallis p-value: p-value = 1.297e-15

In this case, the Turkey results confirm that the mean in sample 2 is higher than in sample 1, with a difference of 56.26 units and a 95% confidence interval.

And our Dunn's test results confirm sample 2's mean is different from sample 1, with a small p-value and evidence against the null hypothesis.

Summary of Minor Variable

Again, both tests show extremely low p-values.

ANOVA p-value: 5.95e-09 Kruskal-Wallis: 1.197e-13

Our Tukey and Dunn's tests tell us there are statistical differences between the means. Though we can see on the boxplot there are more outliers in this variable.

Summary of Circ. Variable

Similar to all our tests so far, we find low p-values (in general, values below 0.05 are considered statistically significant) and statistical differences between the means of the two samples. So far, all of these variables are usable and useful.

ANOVA: 0.0294 Kruskal: 0.0003183

Summary of AR Variable

In the case of AR, there is not enough evidence to reject the null hypothesis that the means of the groups are equal. We cannot find significant enough differences between samples 1 and 2 to consider them statistically significant.

Summary of Round Variable

As with AR, the differences between the two sample groups are not significant enough to be statistically meaningful.

Summary of Solidity Variable

As with AR and Round, the Solidity variable between the two groups is too similar to be meaningful or useful for our analysis.

Analysis Step 1: Visual inspection of one dataset, getting a "lay of the land," so to speak.

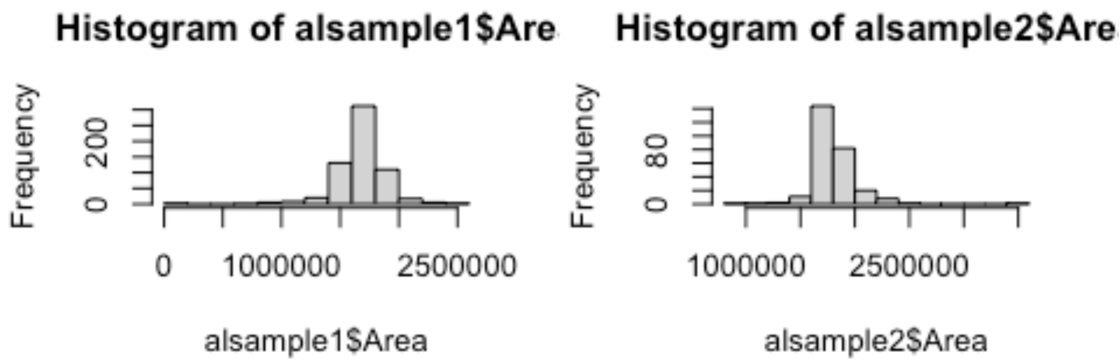
Creating own samples to not interfere with analysis already done.

First step:

Explore and compare samples 1 and 2. Visual inspection of histograms to understand the basic shape of the data. Use of both ANOVA for normal data, and use of Kruskal-Wallis for skewed data.

Looking at Area in samples 1 and 2.

Histograms of Area



Creating a Group Variable

Grouping samples 1 and 2 to compare the two datasets to one another.

Checking to validate that bind worked for samples 1 and 2.

Anova Test

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)    
## sample        1 2.998e+12 2.998e+12    64.3 3.39e-15 ***
## Residuals   882 4.112e+13 4.662e+10                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA shows us the p-value is less than .001, which means there is a significant statistical difference in the average area between the samples. Running Kruskal-Wallis below to verify the findings of ANOVA in the case that the data might be skewed.

Kruskal Test

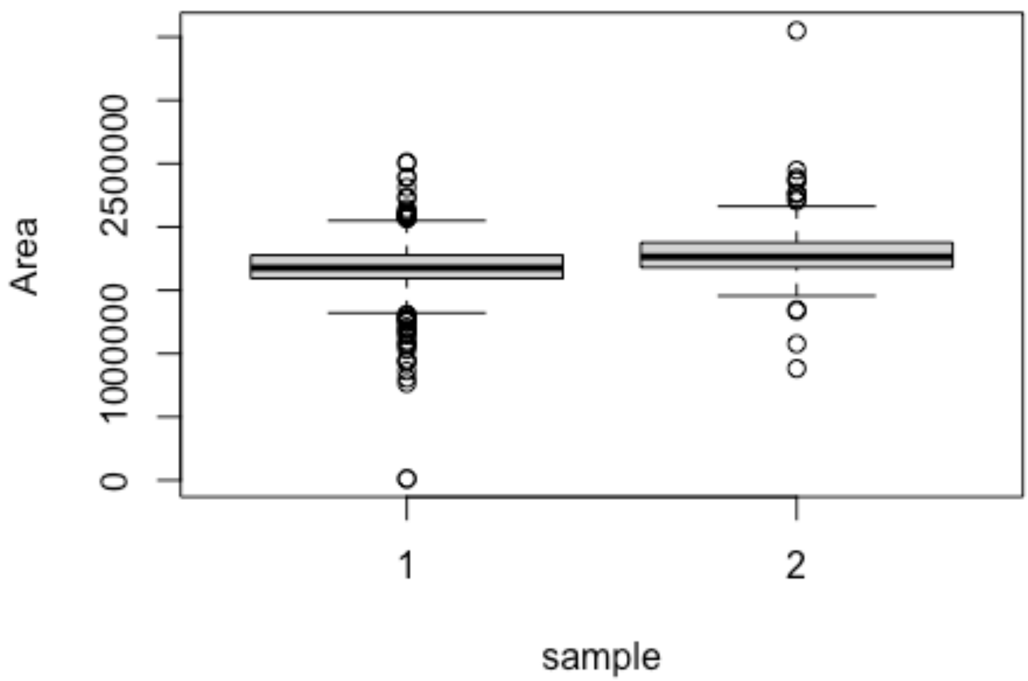
```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  Area by sample
## Kruskal-Wallis chi-squared = 81.809, df = 1, p-value < 2.2e-16
```

Both ANOVA and Kruskal show very low p-values.

Next, we are running pair-wise analysis and then for Kruskal-Wallis for ANOVA to see which sample areas are significantly different.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = Area ~ sample, data = groupeddata1)
## 
## $sample
##         diff      lwr      upr p adj
## 2-1 126297.6 95383.87 157211.4     0
```

Box Plot of Area Vs Sample

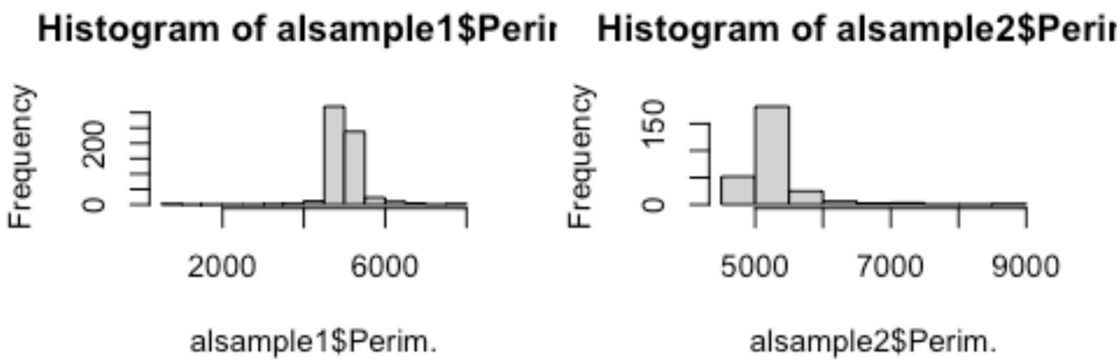



```
##   Comparison        Z      P.unadj        P.adj
## 1      1 - 2 -9.04485 1.498713e-19 1.498713e-19
```

Above, p.adj is the adjusted value for the Bonferroni method, which multiplies the p-value by the number of tests you're doing. This helps when you're doing multiple tests because they're more likely to make a Type 1 error and reject a null when we should not.

Repeating the above analysis for Perimeter in samples 1 and 2.

Histograms of Perim.



Anova test

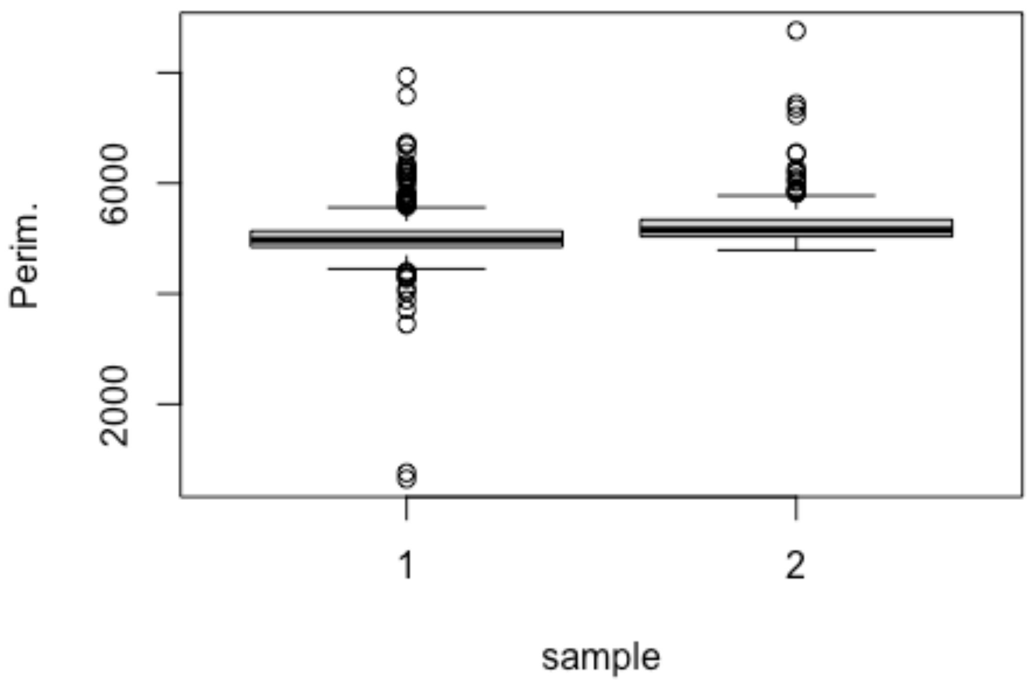
```
##                Df    Sum Sq  Mean Sq F value   Pr(>F)
## sample          1  11091354 11091354   59.34 3.57e-14 ***
## Residuals     882 164864091   186921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, very low p-value. Check with Kruskal-Wallis in case the data is skewed, which is visually apparent in sample 2 for Perimeter.

Kruskal Test

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Perim. by sample
## Kruskal-Wallis chi-squared = 127.67, df = 1, p-value < 2.2e-16

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Perim. ~ sample, data = groupeddata1)
##
## $sample
##         diff      lwr      upr p adj
## 2-1 242.9425 181.0434 304.8416     0
```

Box Plot of Perim. Vs Sample

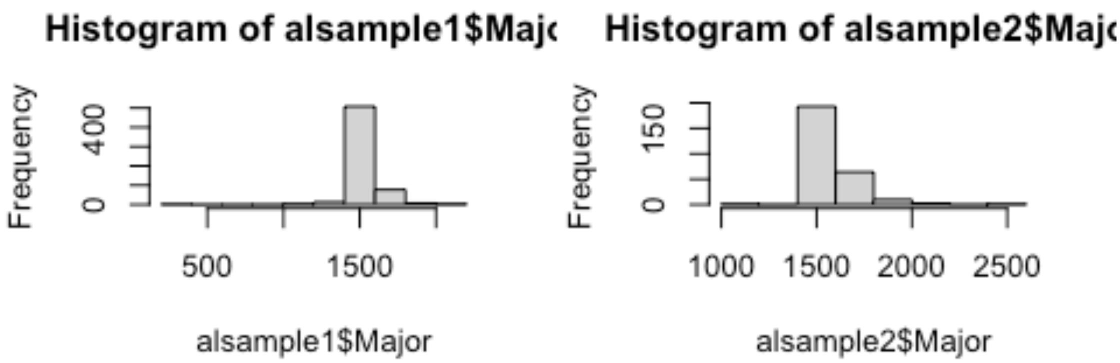



```
##   Comparison       Z      P.unadj        P.adj
## 1      1 - 2 -11.299 1.327274e-29 1.327274e-29
```

Again, p-values are low, meaning each has a different perimeter and is individually significant.

Looking at Major variable next.

Histograms of Major



Anova Test

```
##               Df   Sum Sq Mean Sq F value   Pr(>F)    
## sample         1   594852  594852   43.05 9.09e-11 ***
## Residuals    882 12187168   13818                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

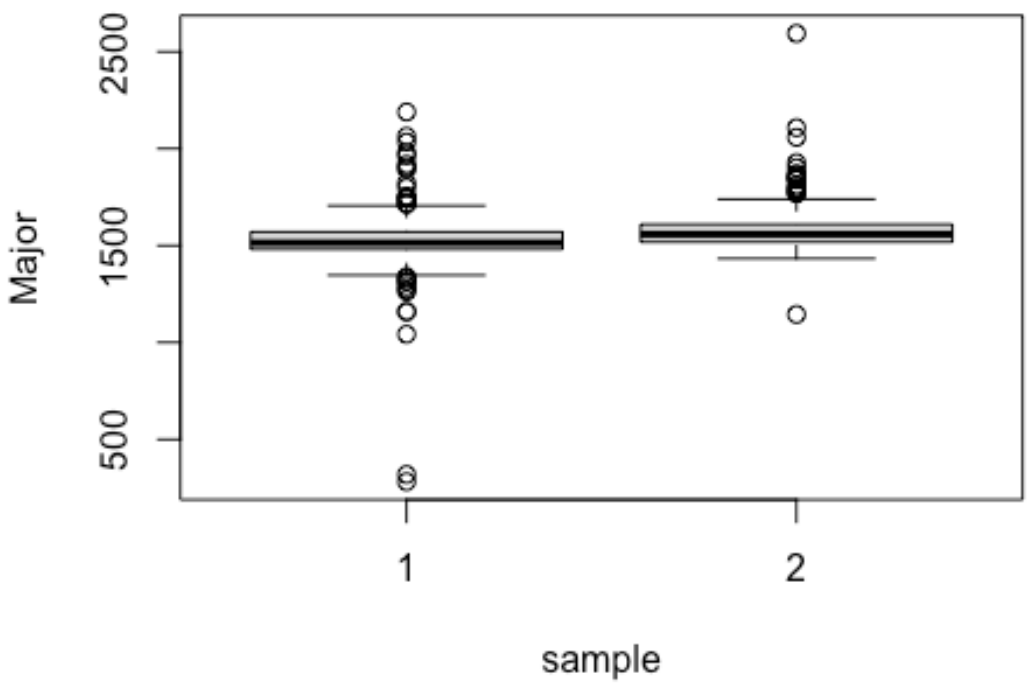
Kruskal Test

```
## 
##  Kruskal-Wallis rank sum test
## 
```

```
## data:  Major by sample
## Kruskal-Wallis chi-squared = 63.919, df = 1, p-value = 1.297e-15
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Major ~ sample, data = groupeddata1)
##
## $sample
##         diff      lwr      upr p adj
## 2-1 56.26209 39.43253 73.09165     0
```

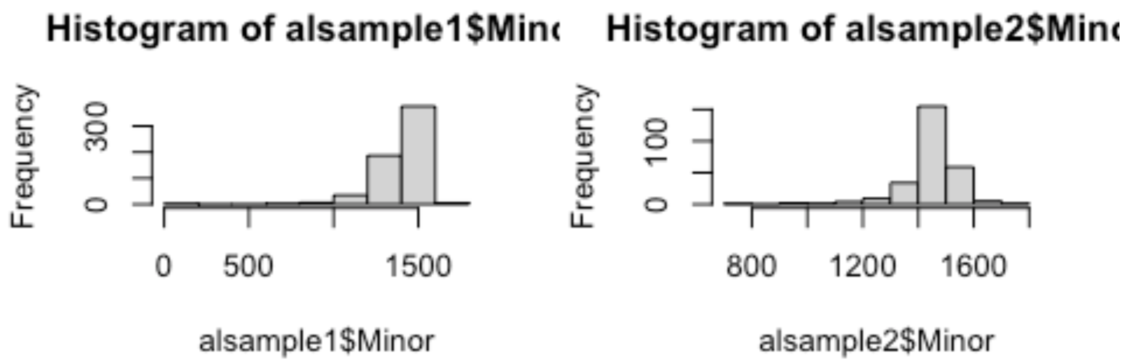
Box Plot of Major Vs Sample




```
##   Comparison         Z      P.unadj        P.adj
## 1      1 - 2 -7.994916 1.296624e-15 1.296624e-15
```

Next, looking at the Minor variable in samples 1 and 2.

Histograms of Minor



Anova Test

```
##               Df   Sum Sq Mean Sq F value   Pr(>F)    
## sample         1   596393  596393   34.53 5.95e-09 ***
## Residuals    882 15234571   17273                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

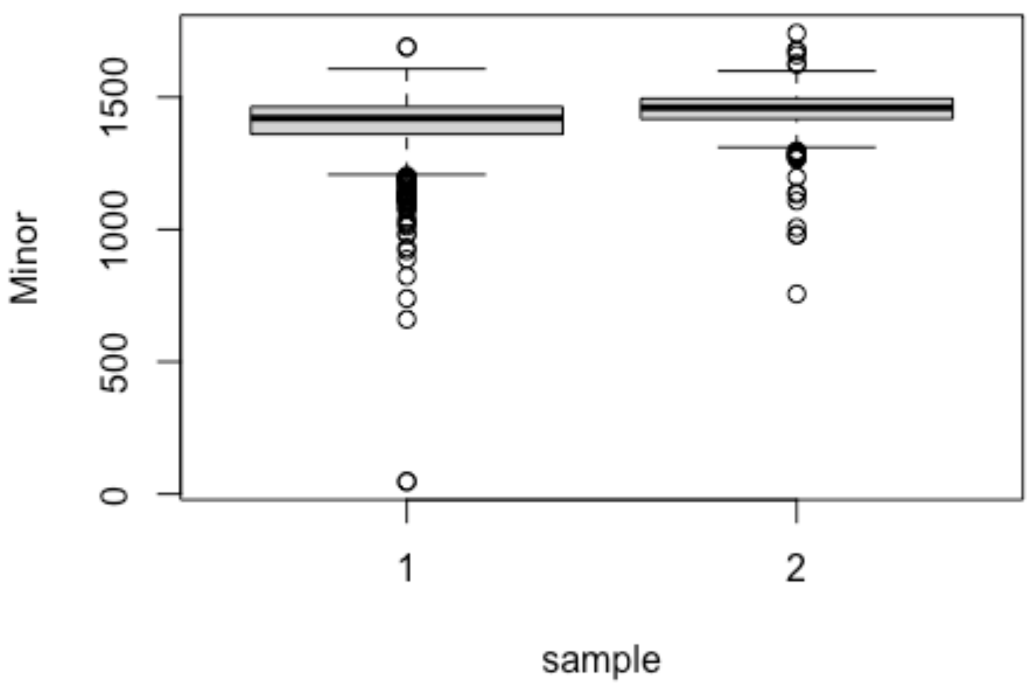
kruskal Test

```
## 
##  Kruskal-Wallis rank sum test
## 
```

```
## data:  Minor by sample
## Kruskal-Wallis chi-squared = 55.013, df = 1, p-value = 1.197e-13
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Minor ~ sample, data = groupeddata1)
##
## $sample
##         diff      lwr      upr p adj
## 2-1 56.33491 37.51851 75.15131     0
```

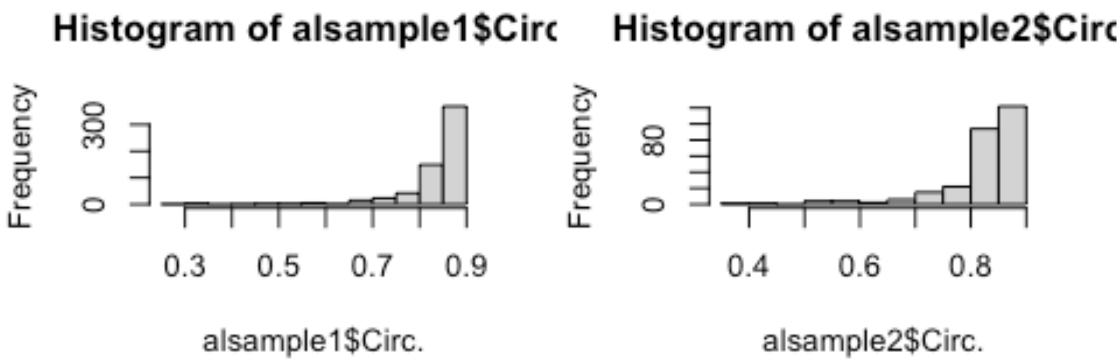
Box plot of Minor Vs Sample




```
##   Comparison         Z      P.unadj        P.adj
## 1      1 - 2 -7.417096 1.197164e-13 1.197164e-13
```

Exploring the Circumference variable for samples 1 and 2.

Histograms of Circ.



Anova Test

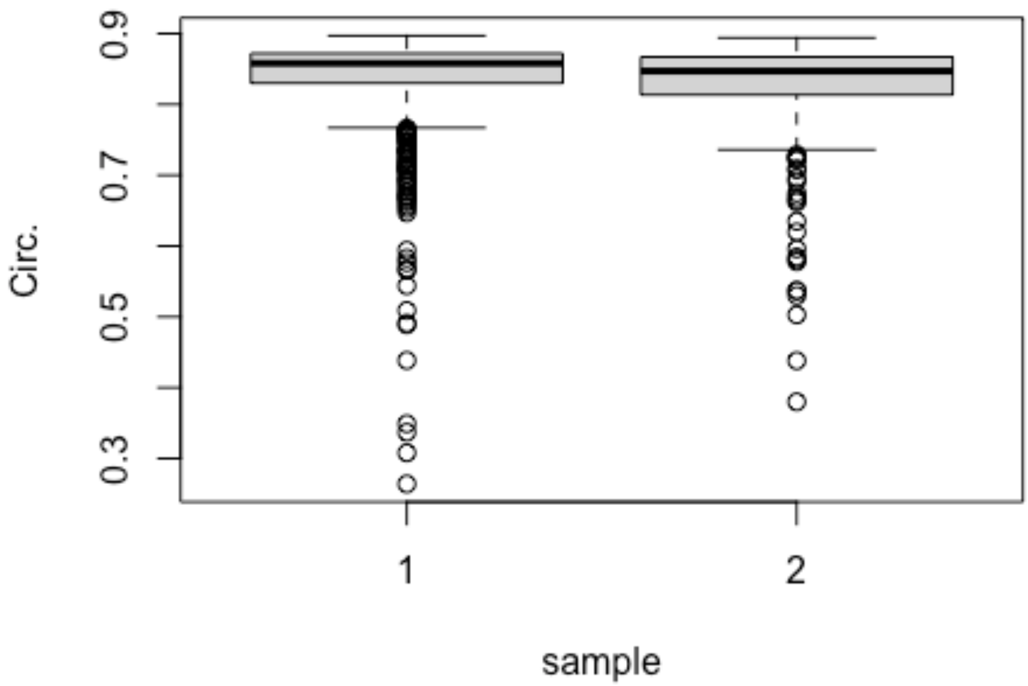
```
##               Df Sum Sq  Mean Sq F value Pr(>F)
## sample         1  0.027 0.026791    4.76 0.0294 *
## Residuals    882  4.964 0.005628
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kruskal test

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  Circ. by sample
## Kruskal-Wallis chi-squared = 12.959, df = 1, p-value = 0.0003183
```

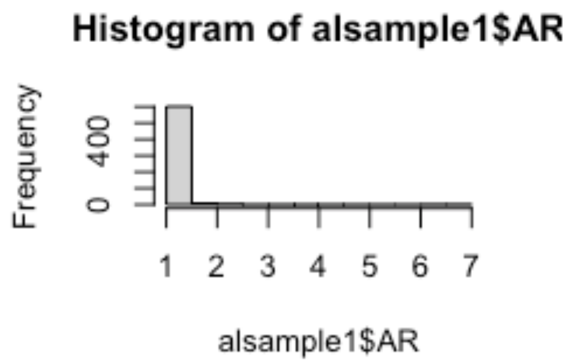
Box Plot of Circ. Vs Sample

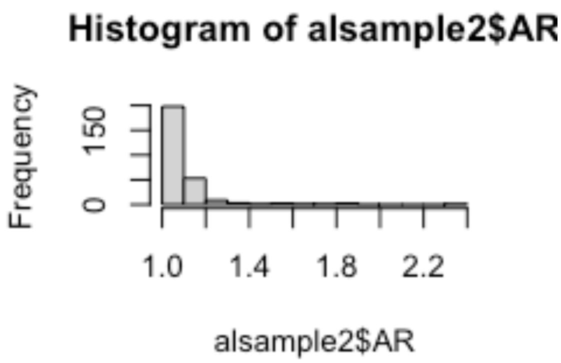



```
##   Comparison        Z    P.unadj      P.adj
## 1      1 - 2 3.599924 0.00031831 0.00031831
```

Looking at AR variable in samples 1 and 2.

Histograms of AR





Anova Test

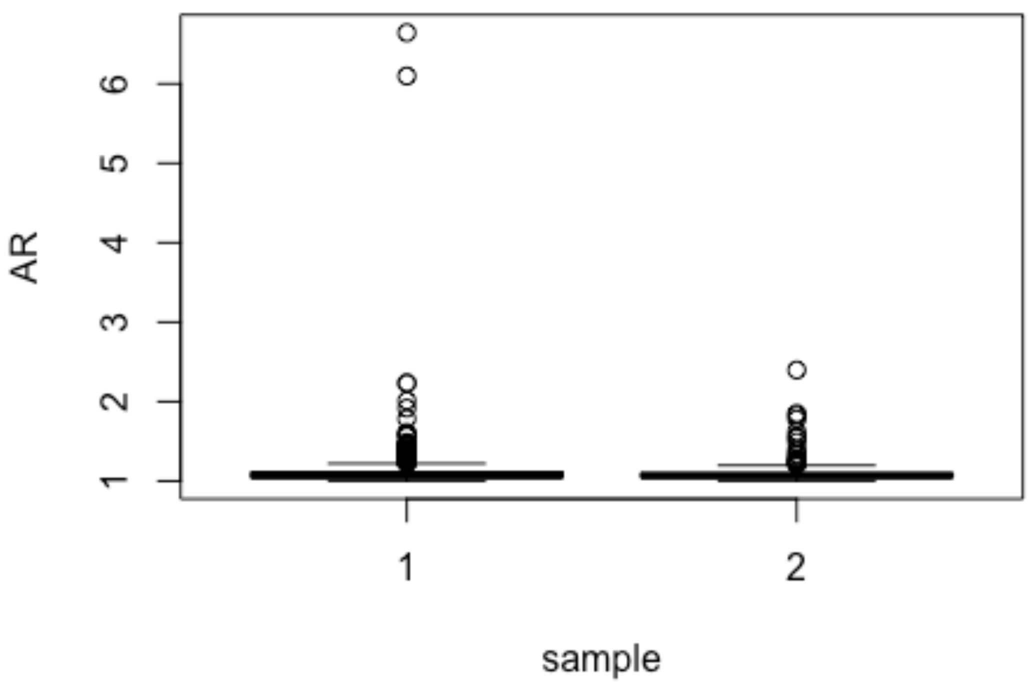
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## sample         1   0.09 0.09224   1.151  0.284
## Residuals    882  70.71 0.08017
```

Kruskal Test

```
##
##  Kruskal-Wallis rank sum test
##
## data:  AR by sample
## Kruskal-Wallis chi-squared = 0.4221, df = 1, p-value = 0.5159
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = AR ~ sample, data = groupeddata1)
##
## $sample
##            diff         lwr        upr     p adj
## 2-1 -0.02215495 -0.06269274 0.01838285 0.2837249
```

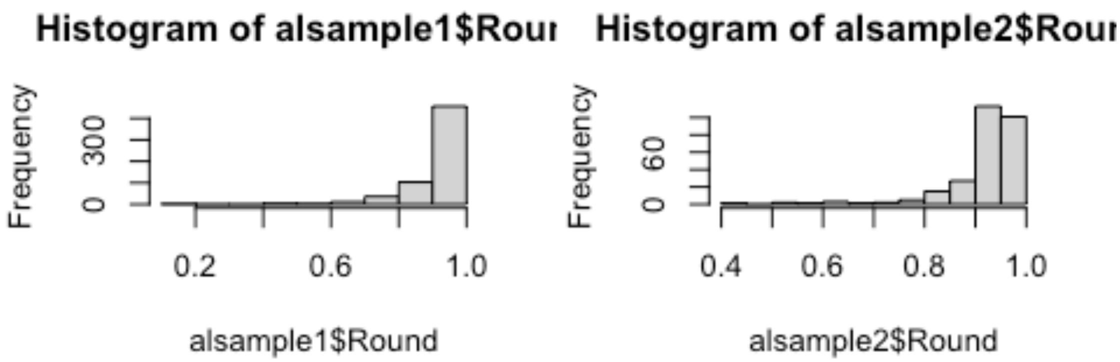
Box Plot of AR vs Sample




```
##   Comparison       Z   P.unadj     P.adj
## 1      1 - 2 0.64969 0.5158925 0.5158925
```

Looking at Round variable in samples 1 and 2.

Histograms of Round



Anova Test

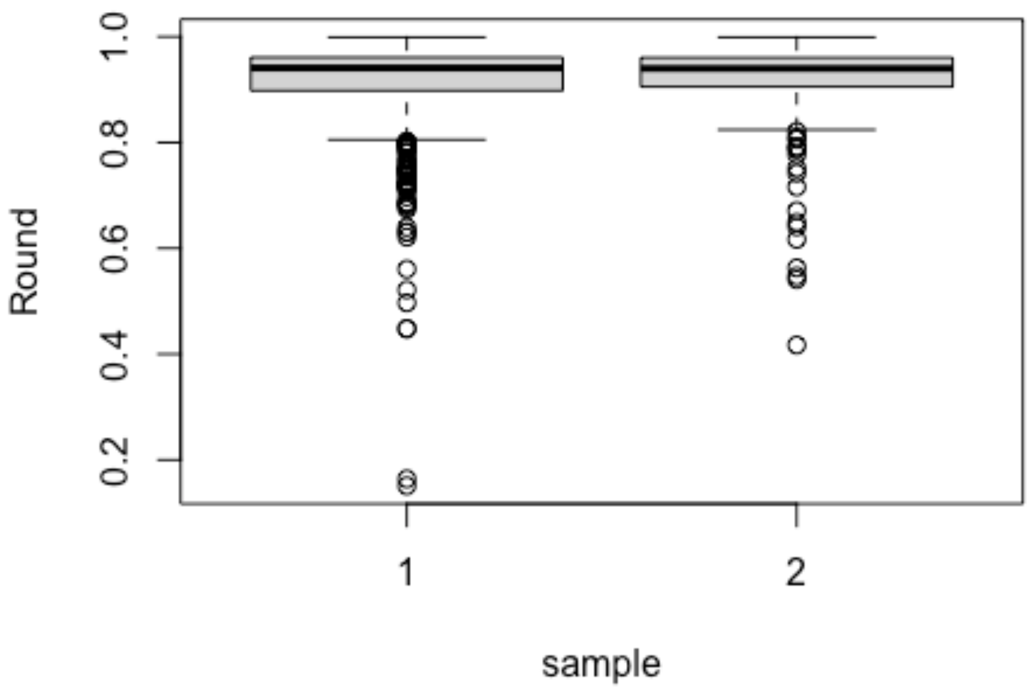
```
##              Df Sum Sq  Mean Sq F value Pr(>F)
## sample        1  0.009 0.009334   1.287  0.257
## Residuals   882  6.398 0.007254
```

Kruskal Test

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Round by sample
## Kruskal-Wallis chi-squared = 0.41655, df = 1, p-value = 0.5187
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Round ~ sample, data = groupeddata1)
##
## $sample
##            diff          lwr        upr     p adj
## 2-1 0.007047712 -0.005145958 0.01924138 0.2569438
```

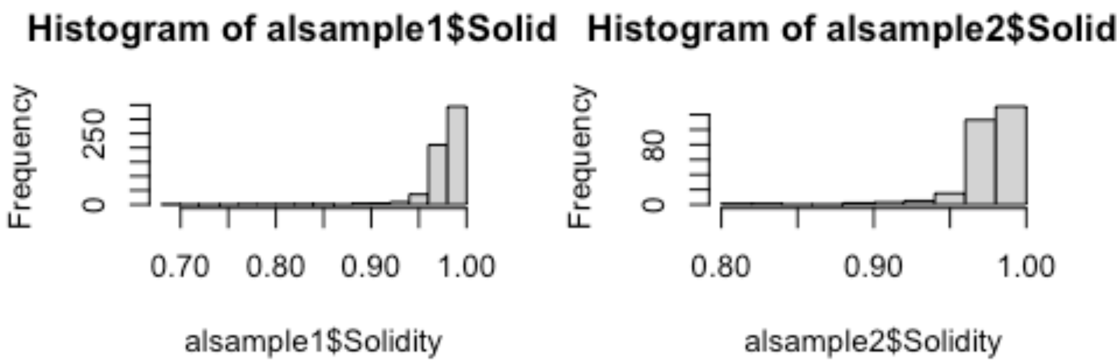
Box plot of Round Vs Sample




```
##   Comparison         Z   P.unadj     P.adj
## 1      1 - 2 -0.645406 0.5186641 0.5186641
```

Looking at Solidity in samples 1 and 2.

Histograms of Solidity



Anova Test

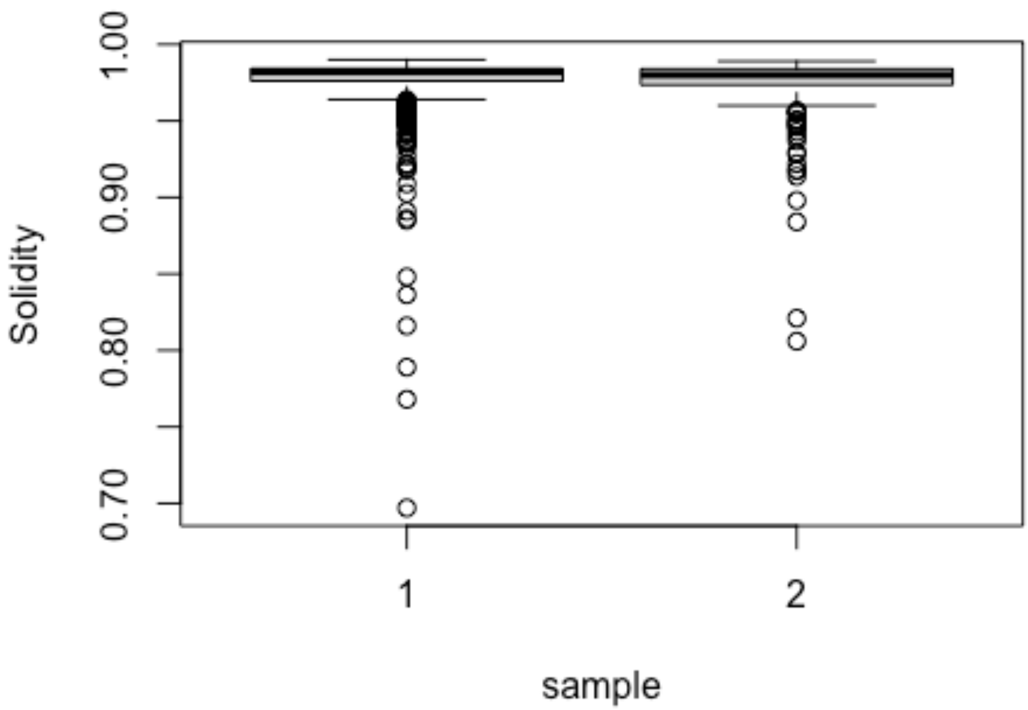
```
##               Df Sum Sq   Mean Sq F value Pr(>F)
## sample         1 0.0002 0.0001887   0.381  0.537
## Residuals    882 0.4373 0.0004958
```

Kruskal Test

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Solidity by sample
## Kruskal-Wallis chi-squared = 5.7378, df = 1, p-value = 0.0166
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Solidity ~ sample, data = groupeddata1)
##
## $sample
##             diff          lwr         upr     p adj
## 2-1 -0.001002196 -0.004190225 0.002185833 0.5374036
```

Box Plot of Solidity Vs Sample




```
##   Comparison        Z    P.unadj      P.adj
## 1      1 - 2 2.395368 0.01660371 0.01660371
```

Lastly in the exploratory process, repeat the same procedures (histograms, ANOVA and Kruskal-Wallis) for variables in samples 3 and 4.

Grouping data in samples 3 and 4.

Checking to validate that bind worked for samples 3 and 4.

Samples 3 & 4

Summary of Exploration of Area Variable

Our tests show statistically significant differences between the two groups, and low p-values for Area. Data is more skewed, so relying on Kruskal more than ANOVA.

ANOVA p-value: 2e-16 Kruskal p-value: 6.689e-05

Summary of Exploration of Perim. Variable

Low p-values and statistically significant differences in the Perim. variable confirms these are valuable for analysis between samples 3 and 4. Data is skewed, so looking at Kruskal value here.

ANOVA p-value: 2e-16 Kruskal p-value: 4.06e-05

Summary of Exploration of Major Variable

As with our other values in samples 3 and 4 groups so far, there are low p-values and significant differences between the two groups.

ANOVA p-value: 2e-16 Kruskal p-value: 3.09e-05

Summary of Exploration of Minor Variable

Low p-values and significant differences between the means of each group when looking at the minor variable.

ANOVA p-value: 3.38e-16 Kruskal p-value: 0.0003214

Summary of Exploration of Circ. Variable

The data looks skewed in both samples, so we will lean on Kruskal and Dunn's analysis here. Again, low p-values and significant differences between the means make this variable useful for analysis.

ANOVA p-value: 2e-16 Kruskal: 3.388e-07

Summary of Exploration of AR Variable

Low p-values and differences between the means make this variable significant.

ANOVA p-value: 2e-16 Kruskal: 2e-16

Summary of Exploration of Round Variable

Both sample group variables look skewed, but our analysis gives us low p-values and statistically different means between the samples.

ANOVA p-value: 2e-16 Kruskal: 2e-16

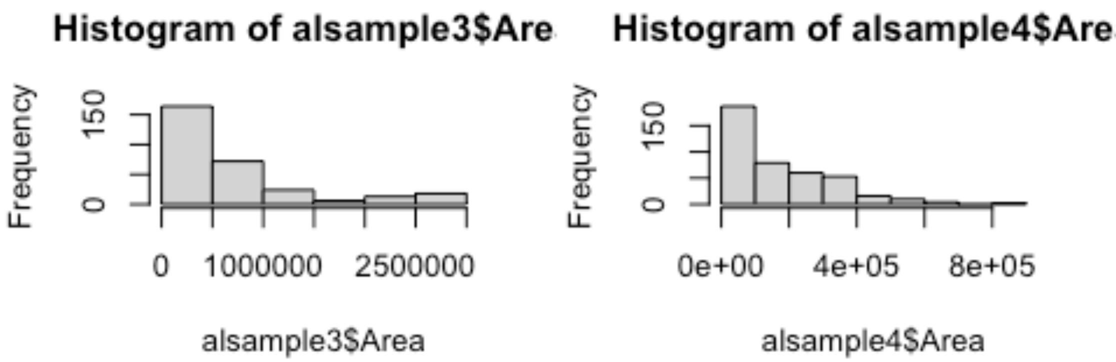
Summary of Exploration of Solidity Variable

As with the other values, significant differences between the groups and low p-values.

ANOVA p-value: 6.05e-14 Kruskal: 2.2e-16

Area in samples 3 and 4.

Histograms of Area



Anova Test

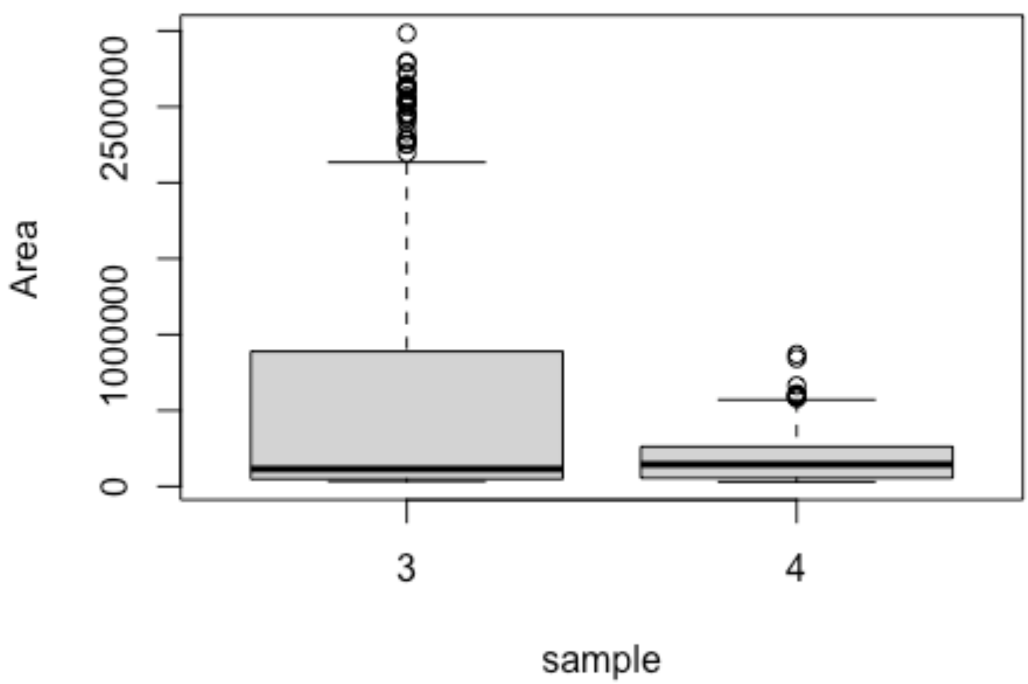
```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## sample         1 3.503e+13 3.503e+13   130.5 <2e-16 ***
## Residuals    706 1.895e+14 2.684e+11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kruskal Test

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  Area by sample
## Kruskal-Wallis chi-squared = 15.897, df = 1, p-value = 6.689e-05

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Area ~ sample, data = groupeddata2)
##
## $sample
##          diff       lwr       upr p adj
## 4-3 -450757.7 -528223.9 -373291.4     0
```

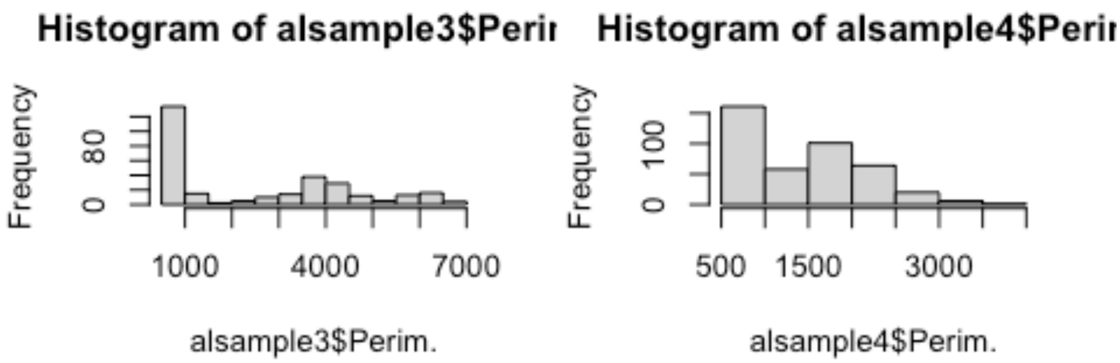
Box Plot of Area Vs Sample




```
##   Comparison        Z      P.unadj        P.adj
## 1      3 - 4 3.987095 6.688713e-05 6.688713e-05
```

Perim. in samples 3 and 4.

Histograms of Perim.



Anova Test

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## sample        1 2.143e+08 214336148   118.8 <2e-16 ***
## Residuals   706 1.273e+09   1803494
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

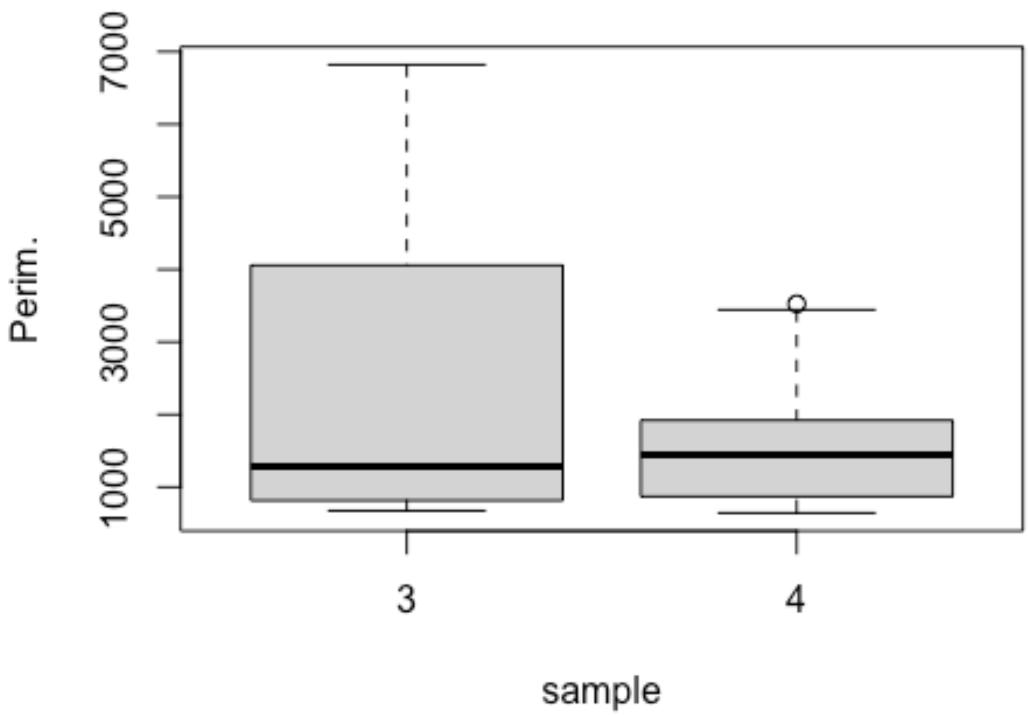
Kruskal Test

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  Perim. by sample
## Kruskal-Wallis chi-squared = 16.843, df = 1, p-value = 4.06e-05
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Perim. ~ sample, data = groupeddata2)
##
## $sample
##          diff       lwr       upr p adj
## 4-3 -1114.975 -1315.777 -914.1732     0
```

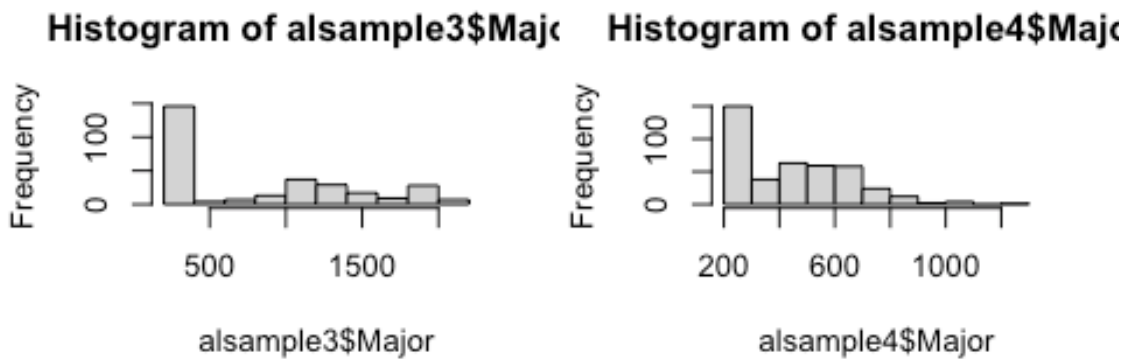
Box Plot of Perim. Vs Sample




```
##   Comparison        Z      P.unadj        P.adj
## 1      3 - 4 4.104039 4.059995e-05 4.059995e-05
```

Major variable in samples 3 and 4.

Histograms of Major



Anova Test

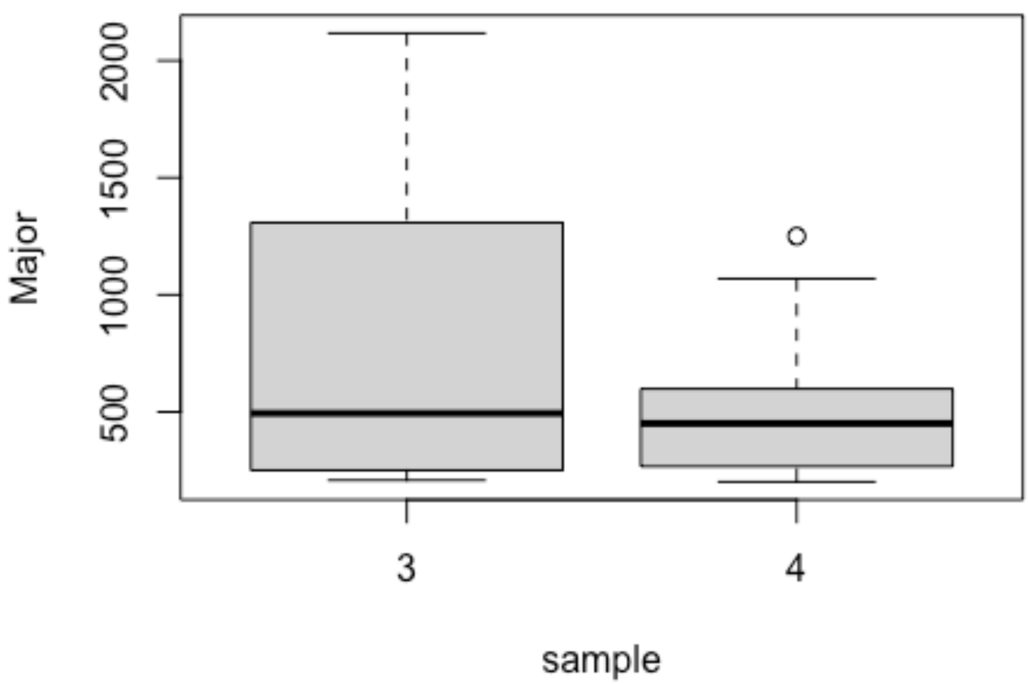
```
##               Df    Sum Sq  Mean Sq F value Pr(>F)    
## sample         1  23615683 23615683   126.5 <2e-16 ***
## Residuals    706 131821657   186716                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kruskal Test

```
## 
##  Kruskal-Wallis rank sum test
## 
```

```
## data:  Major by sample
## Kruskal-Wallis chi-squared = 17.362, df = 1, p-value = 3.09e-05

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Major ~ sample, data = groupeddata2)
##
## $sample
##          diff       lwr       upr p adj
## 4-3 -370.0989 -434.7092 -305.4886     0
```

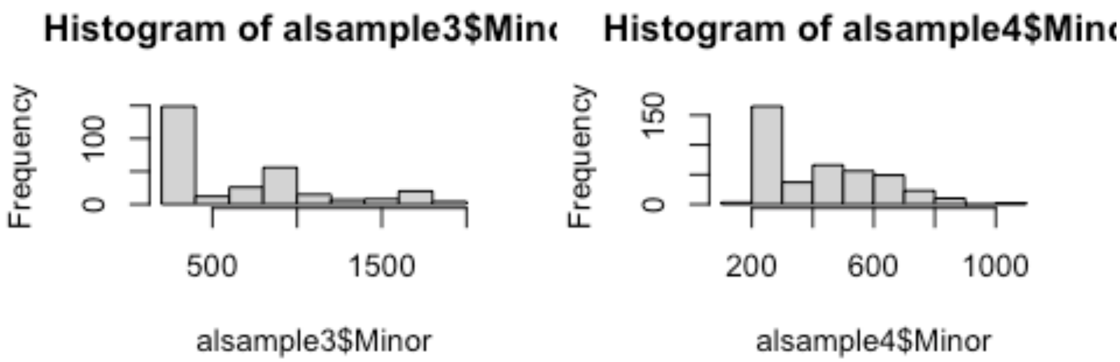
Box Plot of Major Vs Sample




```
##   Comparison       Z      P.unadj        P.adj
## 1      3 - 4 4.16676 3.089593e-05 3.089593e-05
```

Minor variable in samples 3 and 4.

Histograms of Minor



Anova Test

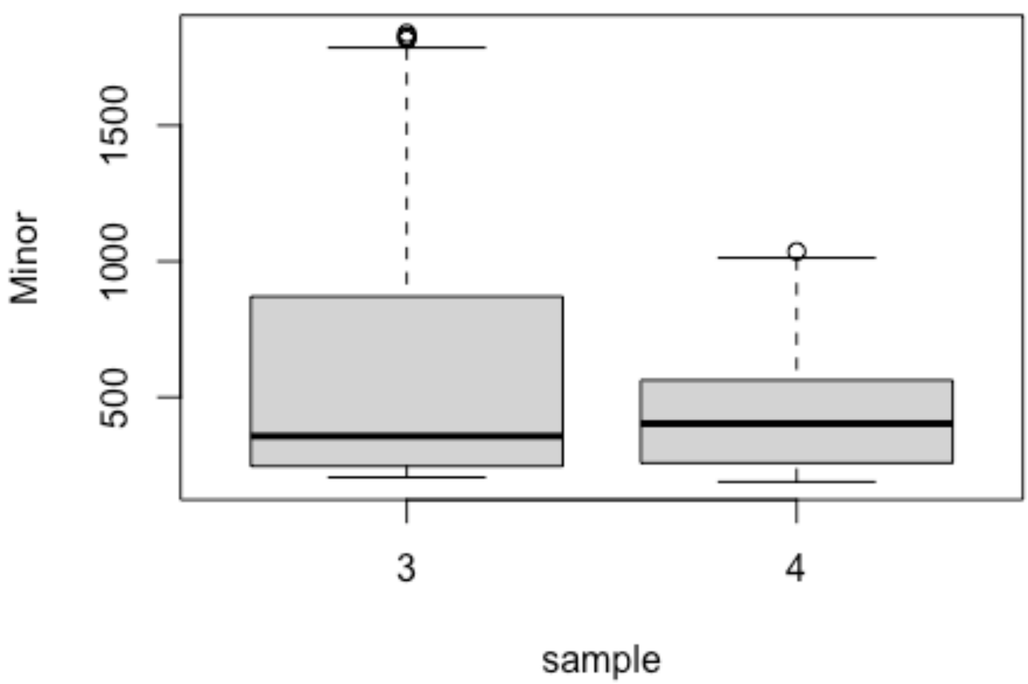
```
##                Df   Sum Sq Mean Sq F value   Pr(>F)    
## sample          1  7851981 7851981   69.86 3.38e-16 ***
## Residuals     706 79352856  112398                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kruskal Test

```
## 
##  Kruskal-Wallis rank sum test
## 
```

```
## data:  Minor by sample
## Kruskal-Wallis chi-squared = 12.941, df = 1, p-value = 0.0003214

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Minor ~ sample, data = groupeddata2)
##
## $sample
##           diff       lwr       upr p adj
## 4-3 -213.4062 -263.5353 -163.2772     0
```

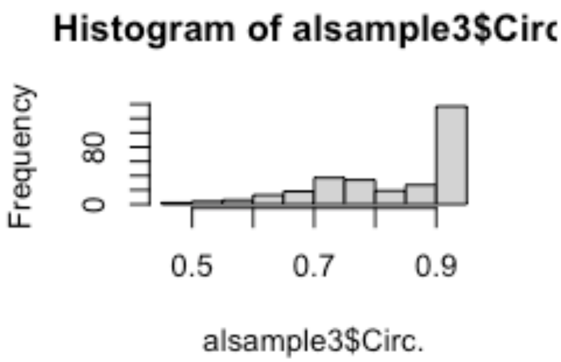
Box Plot of Minor Vs Sample

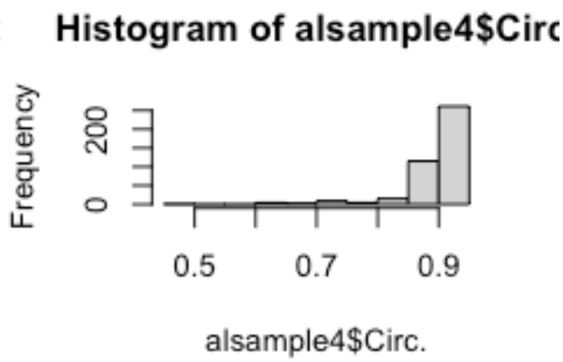



```
##   Comparison        Z     P.unadj       P.adj
## 1      3 - 4 3.597415 0.000321396 0.000321396
```

Circumference variable in samples 3 and 4.

Histograms of Circumference





Anova Test

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## sample        1  0.758  0.7585   113.8 <2e-16 ***
## Residuals   706  4.706  0.0067
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

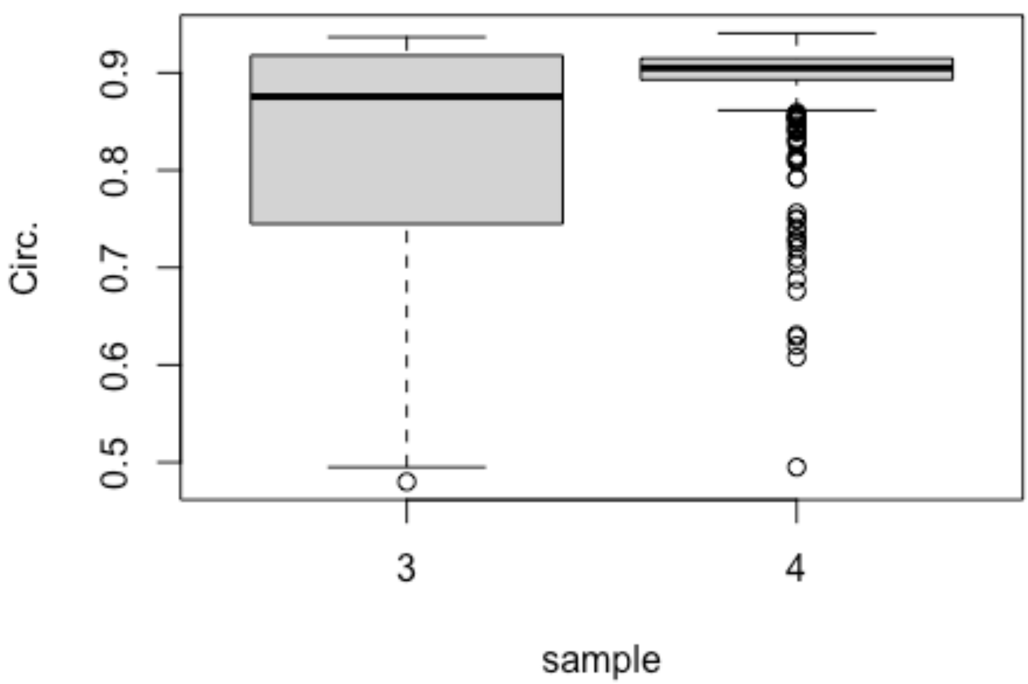
Kruskal Test

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  Circ. by sample
## Kruskal-Wallis chi-squared = 26.015, df = 1, p-value = 3.388e-07
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Circ. ~ sample, data = groupeddata2)
##
## $sample
##           diff        lwr        upr p adj
## 4-3 0.06632738 0.05412019 0.07853456     0
```

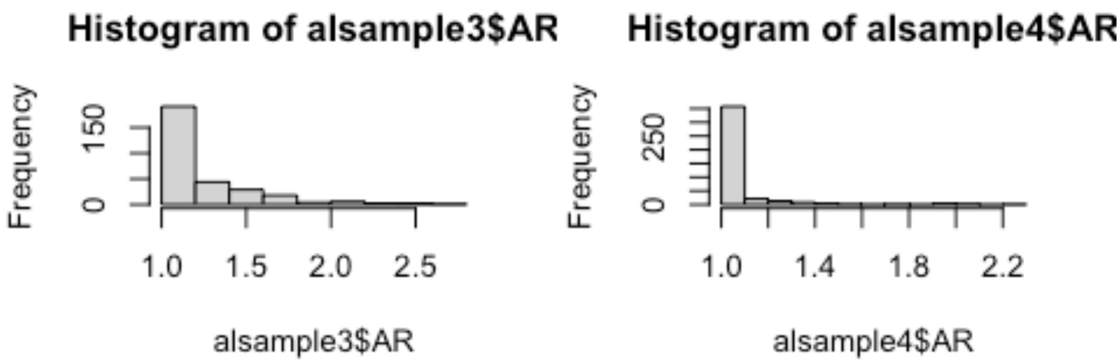
Box Plot of Circ. Vs Sample




```
##   Comparison         Z      P.unadj        P.adj
## 1      3 - 4 -5.100489 3.387762e-07 3.387762e-07
```

AR variable in samples 3 and 4.

Histograms of AR



Anova Test

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## sample        1   4.25   4.247   83.89 <2e-16 ***
## Residuals   706  35.75   0.051
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

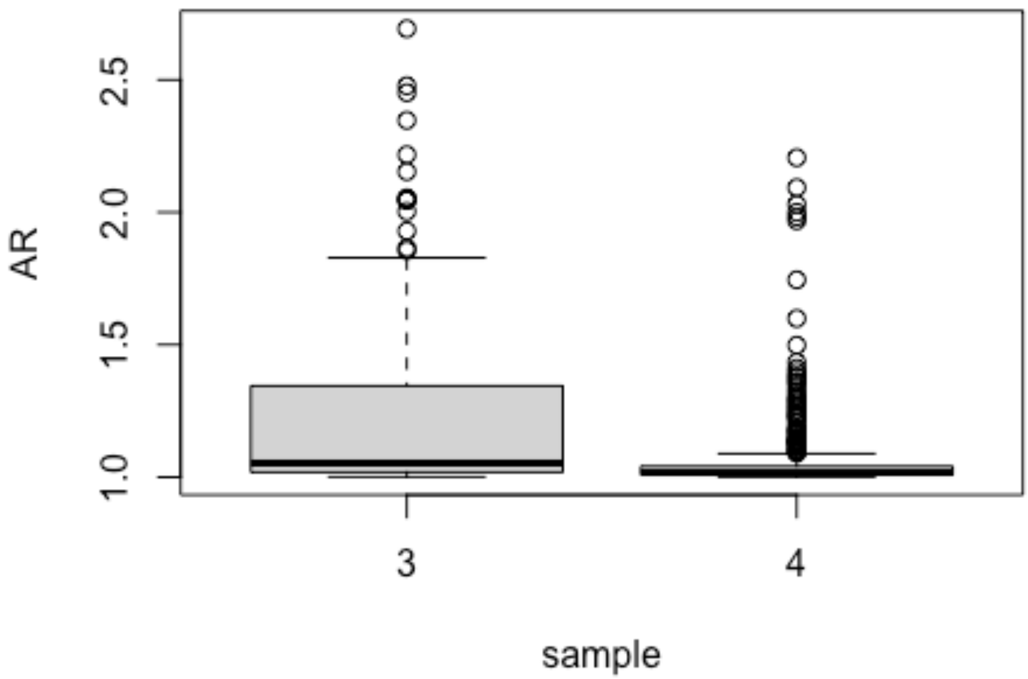
Kruskal Test

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  AR by sample
## Kruskal-Wallis chi-squared = 86.418, df = 1, p-value < 2.2e-16
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = AR ~ sample, data = groupeddata2)
##
## $sample
##           diff        lwr        upr p adj
## 4-3 -0.1569554 -0.1906002 -0.1233105     0
```

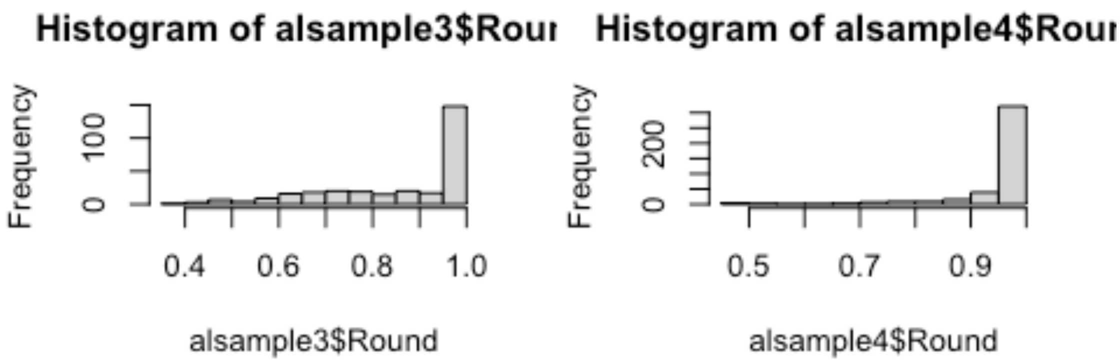
Box Plot of AR Vs Sample




```
##   Comparison        Z      P.unadj        P.adj
## 1      3 - 4 9.296145 1.456296e-20 1.456296e-20
```

Round variable in samples 3 and 4.

Histograms of Round



Anova Test

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## sample          1  1.559  1.5594   107.5 <2e-16 ***
## Residuals     706 10.239  0.0145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

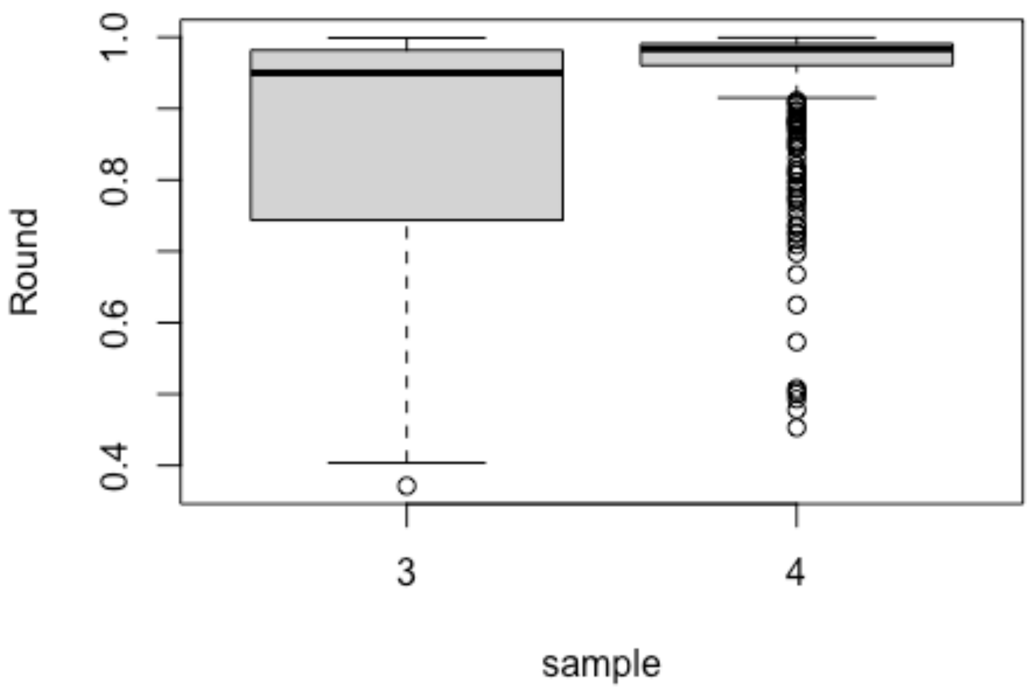
Kruskal Test

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  Round by sample
## Kruskal-Wallis chi-squared = 86.55, df = 1, p-value < 2.2e-16
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Round ~ sample, data = groupeddata2)
##
## $sample
##          diff        lwr       upr p adj
## 4-3 0.0951048 0.07709834 0.1131113     0
```

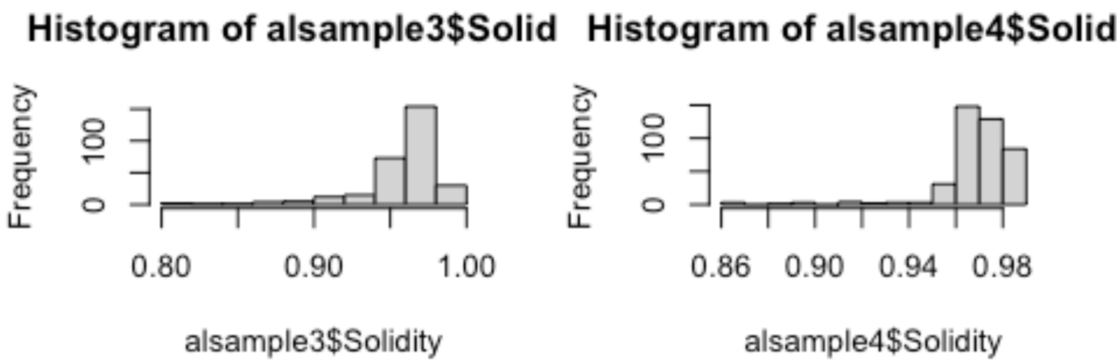
Box Plot of Round Vs Sample




```
##   Comparison         Z      P.unadj        P.adj
## 1      3 - 4 -9.303217 1.362587e-20 1.362587e-20
```

Solidity variable in samples 3 and 4.

Histograms of Solidity



Anova Test

```
##              Df  Sum Sq  Mean Sq F value   Pr(>F)    
## sample        1 0.02564 0.025636   58.71 6.05e-14 ***
## Residuals   706 0.30829 0.000437                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

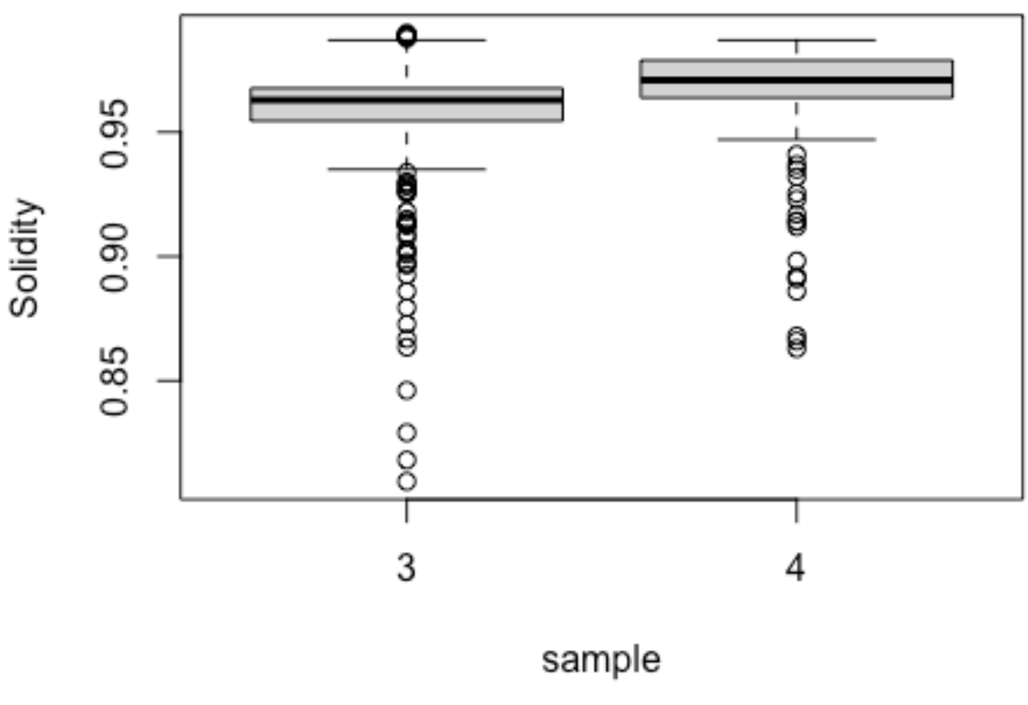
Kruskal Test

```
## 
##  Kruskal-Wallis rank sum test
## 
```

```
## data:  Solidity by sample
## Kruskal-Wallis chi-squared = 91.59, df = 1, p-value < 2.2e-16
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Solidity ~ sample, data = groupeddata2)
##
## $sample
##           diff         lwr        upr p adj
## 4-3 0.01219388 0.009069339 0.01531841     0
```

Box Plot of Solidity Vs Sample




```
##   Comparison         Z      P.unadj        P.adj
## 1      3 - 4 -9.570269 1.066282e-21 1.066282e-21
```

Ref:

- Bevans, R. (2022). ANOVA in R | A Complete Step-by-Step Guide with Examples. Scribbr.Retrieved April 22, 2023, from https://www.scribbr.com/statistics/anova-in-r/
- Kruskal-Wallis Test | R Tutorial. (n.d.). Chi Yau. scribbr. Retrieved April 22, 2023, from https://www.r-tutor.com/elementary-statistics/non-parametric-methods/kruskal-wallis-test

Step4 Splitting the data set(70,30) based on the Brand Variable in to training and testing data.

```
## [1] 27957    11

## [1] 11987    11
```

The training data set has 27957 records, and the testing data has 11987.

Step4: Creating a validation data set to hyper-tune the parameters to select the best classifier.

```
## [1] 2393   11
```

The validation data is a subset of test data and has 2393 records.

Ref:

- Z. (2022, April 12). How to Split Data into Training & Test Sets in R. Statology. Retrieved March 11, 2023, from https://www.statology.org/train-test-split-r/
- R: How to split a data frame into training, validation, and test sets?, Stack Overflow. Retrieved March 11, 2023, from https://stackoverflow.com/questions/36068963/r-how-to-split-a-data-frame-into-training-validation-and-test-sets.

Step 5: Creating the Model to Predict Brands

Approach taken : Since the response variable has more than two classes and collinearity in the predictor variables, we chose LDA for variable selection based on the best accuracy observed. Since the predictors have lot of similar characters, we took a approach to add variable interactions and transformations to LDA model to increase the accuracy and find the best model. We tested three LDA model with different scenarios.

a: Variable selection using the box plots and testing accuracy of LDA models with different predictors

```
##   Accuracy_Above_80
## 1                10
```

```
##   Missing_pred
## 1           63
```

```
## LDA Accuracy mdl1 for Brand is: 0.2875052
```

*b: LDA Model2 with Brand as the response variables and log(Area) + log(Perim.) + log(Major) + Minor + Circ. + AR + Round + Solidity + Area * Perim. * Major + Solidity * Circ. * Round predictors.*

```
##   Accuracy_Above_80
## 1                 9
```

```
##   Missing_pred
## 1           48
```

```
## LDA Accuracy mdl2 for Brand is: 0.2933556
```

*c: LDA Model3 with Shape as the response variables and log(Area) + log(Perim.) + Major + Minor + Circ. + AR + Round + Solidity + Area * Major * Circ. + Perim. * Major * Solidity predictors..*

```
##   Accuracy_Above_80
## 1                 5
```

```
##   Missing_pred
## 1           54
```

```
## LDA Accuracy mdl3 for Brand is: 0.2804012
```

From the above LDA model: LDA model2 has an overall accuracy of 29 percent. It predicted 9 brands with an accuracy above 80 percent. It classified 52% of the unique brands, but has 48% absolute mis-classification, in comparison to other LDA models the absolute mis-classification is the least for LDA model2.

Ref:

- Z. (2020, October 30). Linear Discriminant Analysis in R (Step-by-Step). Statology. Retrieved March 11, 2023, from https://www.statology.org/linear-discriminant-analysis-in-r/
- Sarkar, Priyankur. “What Is LDA: Linear Discriminant Analysis for Machine Learning.” What Is Linear Discriminant Analysis (LDA)?, Knowledgehut, 27 Dec. 2022, https://www.knowledgehut.com/blog/data-science/linear-discriminant-analysis-for-machine-learning.

*Step 6: Beginning QDA Analysis with Brand as the response variables and log(Area) + log(Perim.) + log(Major) + Minor + Circ. + AR + Round + Solidity + Area * Perim. * Major + Solidity * Circ. * Round predictors..*

QDA model

```
##   Accuracy_Above_80
## 1                19

##   Missing_pred
## 1           51

## QDA Accuracy mdl for Brand is: 0.2264939
```

From the above QDA model: QDA model has an overall accuracy of 22 percent. It predicted 19 brands with an accuracy above 80 percent which is better than the LDA model. It classified 49 % of the unique brands, but has 51% absolute mis-classification, in comparison to the LDA model the absolute mis-classification is higher.

Ref:

- Z. (2020, November 2). Quadratic Discriminant Analysis in R (Step-by-Step). Statology. Retrieved March 12, 2023, from https://www.statology.org/quadratic-discriminant-analysis-in-r/
- Saunders, C. (2023, February 10). Classification Part 2 LDA and QDA [Lecture]. D2l.Sdbor.edu. https://d2l.sdbor.edu/d2l/le/content/1781558/viewContent/11116132/View

*Step 7: Beginning Random forest with Brand as the response variables and log(Area) + log(Perim.) + log(Major) + Circ. + AR + Round + Solidity + Area * Major * Circ. + Perim. * Major * Round interactions.*

```
##   Accuracy_Above_80
## 1                 8

##   Missing_pred
## 1           24

## Random Forest Accuracy mdl for Brand is: 0.3297117
```

From the above Random Forest model: Random Forest model has an overall accuracy of 32 percent. It predicted 9 brands with an accuracy above 80 percent, which is very close to LDA and QDA models. It classified 74% of the unique brands, but have 26% absolute mis-classification, in comparison to other models the absolute mis-classification is the least for Randomforest model.

Ref:

- Finnstats. (2021, April 13). Random Forest in R: R-bloggers. R. Retrieved April 22, 2023, from https://www.r-bloggers.com/2021/04/random-forest-in-r/

Step 8: Beginning MclustDa model with Brand as the response variable.

```
## [1] "Set seed"

##   Accuracy_Above_80
## 1                16

##   Missing_pred
## 1           55

## MclustDAAccuracy mdl for Brand is: 0.2356874
```

From the above MclustDA model: MclustDA model has an overall accuracy of 24 percent. It predicted 16 brands with an accuracy above 80 percent. It classified 45% of the unique brands, but has 55% absolute mis-classification, in comparison to Randomforest model the mis-classification is high.

Ref:

- Fraley, C., Raftery, A. E., & Scrucca, L. (n.d.). MclustDA discriminant analysis. Mclust-Org.Github. Retrieved March 13, 2023, from https://mclust-org.github.io/mclust/reference/MclustDA.html

- shanem@mtu.edu, Shane T. Mueller. Model-Based Clustering and Mclust, 28 Mar. 2021, https://pages.mtu.edu/~shanem/psy5220/daily/Day19/modelbasedclustering.html#content.

- Saunders, C. (2023, February 24). MclustDA part1 [Lecture]. D2l.Sdbor.edu. https://d2l.sdbor.edu/d2l/le/content/1781558/viewContent/11116023/View

- Saunders, C. (2023, February 24). MclustDA part2 [Lecture]. D2l.Sdbor.edu. https://d2l.sdbor.edu/d2l/le/content/1781558/viewContent/11116022/View

- Saunders, C. (2023, February 24). MclustDA part3 Cross validation [Lecture]. D2l.Sdbor.edu. https://d2l.sdbor.edu/d2l/le/content/1781558/viewContent/11116024/View

- Saunders, C. (2023, February 24). Mclust Play Part2 R file [Lecture]. D2l.Sdbor.edu. https://d2l.sdbor.edu/d2l/le/content/1781558/viewContent/11116026/View

- Saunders, C. (2023, February 24). Mclust Play Part3 R file [Lecture]. D2l.Sdbor.edu. https://d2l.sdbor.edu/d2l/le/content/1781558/viewContent/11116025/View

Step 9: Creating table for final prediction:

Smokeless Gun Powder data results in the form of a table

Model Results Analysis Table

Models	Overall_Accuracy	Brand_Count_with_Acc_80	Brand_Count_with_Pred_Mis
LDA	0.2934	9	48
QDA	0.2265	19	51
Randomforest	0.3297	8	24
MclustDA	0.2357	16	55

Based on the above table we can see that Random Forest has the best overall accuracy of 32%. It predicted 9 brands with an accuracy above 80 percent, which is very close to LDA and QDA models. It classified 74% of the unique brands, but have 26% absolute mis-classification, in comparison to other models the absolute mis-classification is the least for the Randomforest model. This is the best model for the analysis.

Ref:

- kTable: Make Nicely Formatted Tables in Kmisc: Kevin Miscellaneous. (n.d.). Rdrr.io. Retrieved April 25, 2023, from https://rdrr.io/cran/Kmisc/man/kTable.html

Step 10: Creating table for final prediction by each model

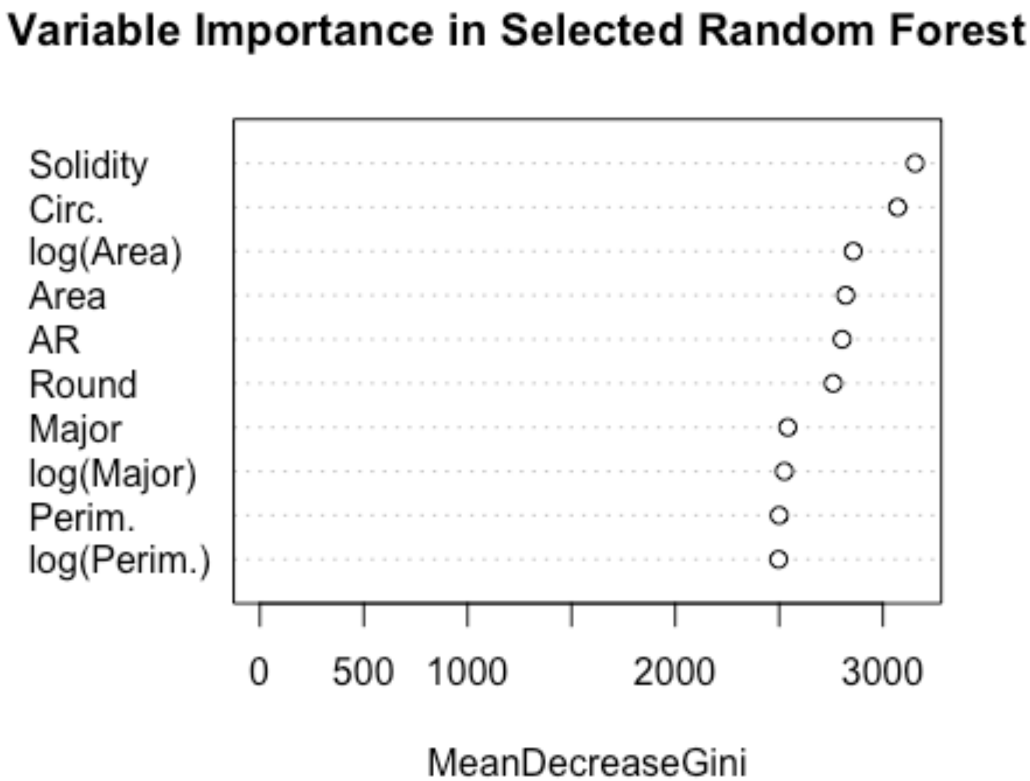
```
##            Brand LDA_accuracy_percent QDA_accuracy_percent
## 147    TrailBoss            0.6666667            0.7777778
## 138    Reloader7            0.4736842            0.3684211
## 81        H50BMG            0.8000000            0.9000000
## 151        US869            0.5454545            0.8181818
## 84  HI-Skor800-X            0.9166667            0.9166667
## 144       Target            0.8571429            0.9285714
## 76         H4198            0.7777778            0.8888889
## 80         H4895            0.9000000            1.0000000
## 12          4198            0.9230769            0.9230769
## 78         H4831            1.0000000            1.0000000
```

```
##     RandomForest_accuracy_percent MclustDAt_accuracy_percent
## 147                     0.7777778                  0.7777778
## 138                     0.7894737                  0.4210526
## 81                      0.8000000                  0.9000000
## 151                     0.8181818                  0.3636364
## 84                      0.8333333                  0.9166667
## 144                     0.8571429                  0.8571429
## 76                      0.8888889                  0.8888889
## 80                      0.9000000                  1.0000000
## 12                      1.0000000                  0.9230769
## 78                      1.0000000                  1.0000000
```

Step 11a: Summary and importance of the Random forest Model.

```
##                 Length  Class  Mode
## call                  3 -none- call
## type                  1 -none- character
## predicted         27957 factor numeric
## err.rate          77500 -none- numeric
## confusion         23870 -none- numeric
## votes           4305378 matrix numeric
## oob.times         27957 -none- numeric
## classes             154 -none- character
## importance           10 -none- numeric
## importanceSD          0 -none- NULL
## localImportance       0 -none- NULL
## proximity             0 -none- NULL
## ntree                 1 -none- numeric
## mtry                  1 -none- numeric
## forest               14 -none- list
## y                 27957 factor numeric
## test                  0 -none- NULL
## inbag                 0 -none- NULL
## terms                 3 terms  call

##             MeanDecreaseGini
## log(Area)           2857.012
## log(Perim.)         2497.622
## log(Major)          2525.247
## Circ.               3070.622
## AR                  2803.538
## Round               2760.394
```

```
## Solidity             3155.026
## Area                 2820.907
## Major                2541.627
## Perim.               2500.330
```



Step 11b: Testing the Random forest Model.

```
##                  Brand accuracy_percent
## 1                 0.41       0.54545455
## 2                20/28       0.44680851
## 3                  201       0.27083333
## 4                  231       0.03658537
## 5                 2400       0.14516129
## 6                  244       0.02127660
## 7                  296       0.14765101
## 8                 3031       0.54901961
## 9              4007SSC       0.26000000
```

```
## 10                4064       0.65486726
## 11                4166       0.22807018
## 12                4198       0.93650794
## 13                4227       0.40243902
## 14                4320       0.49295775
## 15                4350       0.52000000
## 16                4451       0.25862069
## 17                4831       0.60377358
## 18                4895       0.52000000
## 19                4955       0.15789474
## 20                 572       0.06521739
## 21                 748       0.06382979
## 22                 760       0.05172414
## 23                7828       0.38000000
## 24             7828SSC       0.35416667
## 25                7977       0.28846154
## 26                8133       0.22448980
## 27             8208XBR       0.25000000
## 28     AASuper-Handicap       0.02173913
## 29         Accurate1680       0.17431193
## 30         Accurate2015       0.27659574
## 31         Accurate2200       0.09589041
## 32         Accurate2230       0.05263158
## 33         Accurate2460       0.05882353
## 34         Accurate2495       0.46428571
## 35         Accurate2520               NA
## 36         Accurate2700       0.26415094
## 37         Accurate4064       0.30188679
## 38         Accurate4100       0.36363636
## 39         Accurate4350       0.26086957
## 40         Accurate5744       0.53191489
## 41        AccurateLT-30       0.42307692
## 42        AccurateLT-32       0.54782609
## 43       AccurateMagPro       0.18965517
## 44   AccurateNitro100NF       0.06000000
## 45         AccurateNo.2       0.70394737
## 46         AccurateNo.5       0.06306306
## 47         AccurateNo.7       0.02020202
## 48         AccurateNo.9       0.14062500
## 49          AccurateTCM       0.04761905
```

```
## 50      AmericanSelect      0.67346939
## 51             AR-Comp      0.21666667
## 52            AutoComp      0.05454545
## 53               BE-86      0.21839080
## 54           Benchmark      0.41666667
## 55             BL-C(2)      0.39191074
## 56                BLUE      0.36842105
## 57             BlueDot      0.28846154
## 58            Bullseye      0.74065421
## 59              CFE223              NA
## 60              CFEBLK      0.20792079
## 61           CFEPistol      0.06896552
## 62             ClayDot      0.05454545
## 63               Clays      0.21153846
## 64             D032-03      0.30000000
## 65             D073-08      0.15384615
## 66                  E3      0.11666667
## 67          Extra-Lite      0.14583333
## 68               GREEN      0.37777778
## 69            GreenDot      0.03508772
## 70               H1000      0.58823529
## 71                H110      0.06015038
## 72                H322      0.22807018
## 73                H335      0.47595561
## 74                H380      0.04285714
## 75                H414      0.06349206
## 76               H4198      0.80000000
## 77               H4350      0.59183673
## 78               H4831      0.92000000
## 79             H4831SC      0.63207547
## 80               H4895      0.81632653
## 81              H50BMG      0.89583333
## 82               Herco      0.10638298
## 83         HI-Skor700-X      0.11764706
## 84         HI-Skor800-X      0.76271186
## 85               HP-38      0.04000000
## 86                HS-6      0.12345679
## 87           Hybrid100V      0.32142857
## 88        International      0.18000000
## 89       Leverevolution              NA
```

```
## 90             Lil_Gun       0.08695652
## 91            Longshot       0.16176471
## 92                N110       0.28301887
## 93                N133       0.15217391
## 94                N135       0.26666667
## 95                N140       0.06382979
## 96                N150       0.13043478
## 97                N160       0.14545455
## 98                N165       0.02173913
## 99                N310       0.56521739
## 100               N320       0.19230769
## 101               N340       0.30909091
## 102               N350       0.21153846
## 103               N540       0.15789474
## 104               N550       0.36363636
## 105               N560       0.52173913
## 106            No.11FS       0.07608696
## 107           norma200       0.28000000
## 108           norma202       0.24590164
## 109          norma203B       0.11320755
## 110        PowerPistol       0.02083333
## 111     PowerPro1200-R               NA
## 112    PowerPro2000-MR       0.01754386
## 113     PowerPro300-MP       0.06185567
## 114    PowerPro4000-MR       0.40350877
## 115    PowerProVarmint               NA
## 116           ProReach       0.73913043
## 117      RamshotBigGame       0.07246377
## 118     RamshotEnforcer       0.60538117
## 119      RamshotHunter       0.06382979
## 120      RamshotMagnum       0.17910448
## 121  RamshotSilhouette       0.07407407
## 122         RamshotTAC       0.01428571
## 123     RamshotTrueBlue       0.42857143
## 124 RamshotX-Terminator       0.02816901
## 125         RamshotZip       0.09523810
## 126                RED       0.21739130
## 127             RedDot       0.39130435
## 128         Reloader10x       0.10000000
## 129         Reloader15       0.13207547
```

```
## 130          Reloader16       0.08333333
## 131          Reloader17       0.16363636
## 132          Reloader19       0.22105263
## 133          Reloader22       0.13095238
## 134          Reloader23       0.65612648
## 135          Reloader25       0.10638298
## 136          Reloader33       0.70833333
## 137          Reloader50       0.66666667
## 138           Reloader7       0.57291667
## 139             Retumbo       0.53061224
## 140         SportPistol       0.53061224
## 141          StaBall6.5       0.20833333
## 142               Steel       0.66666667
## 143       Superformance       0.09722222
## 144              Target       0.83333333
## 145           Titegroup       0.07500000
## 146             Titewad       0.08235294
## 147           TrailBoss       0.80851064
## 148              Unique       0.18000000
## 149           Universal       0.49180328
## 150                 URP       0.10416667
## 151               US869       0.69811321
## 152              Varget       0.34545455
## 153                 WSF       0.02127660
## 154                 WST       0.01851852

## randomForest Accuracy mdl for Brand is: 0.3225995
```

The Final testing confirms that the random forest is the best model for predicting the Brand.

Step 12: Creating a Model to predict the Shape. We are creating this to validate our results based on the Shape.

```
## Randomforest Accuracy for Shape is: 0.8967214
```

The Overall Accuracy for shape is 90%.

Step 13: Prediction of Recovered Samples using the RandomForest Model.

Part1: Sample 1 and Sample 2 Analysis (Comparing the brands and finding the brand name).

Predictions of recovered sample 1

```
## # A tibble: 1 × 2
##   Brand  Predicted_value
##   <fct>            <dbl>
## 1 RedDot            158.

## # A tibble: 1 × 2
##   LDA_sample1.class Predicted_value
##   <fct>                       <int>
## 1 RedDot                        242

## # A tibble: 1 × 2
##   predict.Shape_randomforest_MDL..recovered.sample1.
Predicted_value
##   <fct>
<int>
## 1 flake
570
```

Based on the results from the selected Random Forest model recovered sample 1 has the highest predicted value of 164 and is from Brand Reddot. We validated our results by implementing the LDA Model and the predicted value of Reddot is the highest with LDA as well. To further confirm our results, we predicted the shape of the recovered sample 1 and the flake has the highest predicted value. We checked the shape of Reddot is Flake from the train data, this confirms our results that the recovered sample 1 is from the brand Reddot.

Predictions of recovered sample 2

```
## # A tibble: 1 × 2
##   Brand  Predicted_value
##   <fct>            <dbl>
## 1 RedDot            62.6

## # A tibble: 1 × 2
##   LDA_sample2.class Predicted_value
##   <fct>                       <int>
## 1 RedDot                         87

## # A tibble: 1 × 2
##   predict.Shape_randomforest_MDL..recovered.sample2.
```

```
Predicted_value
##   <fct>
<int>
## 1 flake
255
```

Based on the results from the selected Random Forest model recovered sample 2 has the highest predicted value of 65 and is from Brand Reddot. We validated our results by implementing the LDA Model and the predicted value of Reddot is the highest with LDA as well. To further confirm our results, we predicted the shape of the recovered sample 2 and the flake has the highest predicted value. This confirms our results that the recovered sample 2 is from the brand Reddot.

Part 2: Smokeless Gun Powder presence of multiple Brand Analysis in Sample 3 and Sample 4.

Predictions of recovered sample 3

```
## # A tibble: 5 × 2
##   Brand           Predicted_value
##   <fct>                     <dbl>
## 1 N133                       2.74
## 2 AccurateNo.2               4.93
## 3 Accurate4100              14.2
## 4 AmericanSelect            19.5
## 5 RamshotEnforcer           61.1

## # A tibble: 5 × 2
##   LDA_sample3.class Predicted_value
##   <fct>                       <int>
## 1 8208XBR                         9
## 2 AmericanSelect                 16
## 3 AccurateNo.2                   22
## 4 Accurate4100                   32
## 5 RamshotEnforcer                91

## # A tibble: 4 × 2
##   predict.Shape_randomforest_MDL..recovered.sample3.
Predicted_value
##   <fct>
<int>
## 1 cylindrical
58
## 2 flake
36
```

```
## 3 flattened_spherical
54
## 4 spherical
149
```

Based on the results from the selected Random Forest model recovered sample 3 has the highest predicted value of 61 and is from Brand RamshotEnforcer. We validated our results by implementing the LDA Model and the predicted value of RamshotEnforcer is the highest with LDA as well. We checked sample 3 for the presence of other brands and both Random Forest and LDA showed the presence of other brands AmericanSelect, AccurateNo.2, and Accurate4100.To further confirm our results we predicted the shape of the recovered sample 3 and the sample have a spherical and flake shape. This confirmed our analysis that manufacturers are using multiple brands in the recovered sample 3, but the majority of particles are from the brand RamshotEnforcer.

Predictions of recovered sample 4

```
## # A tibble: 5 × 2
##   Brand           Predicted_value
##   <fct>                     <dbl>
## 1 H335                       7.62
## 2 RamshotTrueBlue            9
## 3 BL-C(2)                   11.8
## 4 Accurate4100              13.5
## 5 RamshotEnforcer           72.6

## # A tibble: 5 × 2
##   LDA_sample4.class Predicted_value
##   <fct>                       <int>
## 1 AccurateNo.9                   21
## 2 Accurate4100                   40
## 3 AccurateNo.2                   41
## 4 BL-C(2)                        71
## 5 RamshotEnforcer                94

## # A tibble: 3 × 2
##   predict.Shape_randomforest_MDL..recovered.sample4.
Predicted_value
##   <fct>
<int>
## 1 cylindrical
2
## 2 flattened_spherical
```

```
187
## 3 spherical
222
```

Based on the results from the selected Random Forest model recovered sample 4 has the highest predicted value of 71 and is from Brand RamshotEnforcer. We validated our results by implementing the LDA Model and the predicted value of RamshotEnforcer is the highest with LDA as well. We checked sample 4 for the presence of other brands and both Random Forest and LDA showed the presence of other brands AccurateNo.2, Accurate4100 & BL-C(2).To further confirm our results we predicted the shape of the recovered sample 4 and the sample have spherical and flattened_spherical shape. This confirmed our analysis that manufacturers are using multiple brands in the recovered sample 4, but the majority of the particles are from the brand RamshotEnforcer.

Creating a table of final results of prediction of all recovered samples:

Recovered Sample Brand Prediction Results Sample1 & sample2

Models	Recovered.Sample1	Recovered.Sample2
RandomForest	RedDot	RedDot
LDA	RedDot	RedDot

```
##   Ranforest.Sample3Brand LDA_sample3.class Ranforest.Sample4Brand
## 1            Accurate4100      Accurate4100           Accurate4100
## 2            AccurateNo.2      AccurateNo.2                   <NA>
## 3          AmericanSelect    AmericanSelect                   <NA>
## 4         RamshotEnforcer   RamshotEnforcer        RamshotEnforcer
## 5                    <NA>              <NA>                BL-C(2)
##   LDA_sample4.class
## 1      Accurate4100
## 2              <NA>
## 3              <NA>
## 4   RamshotEnforcer
## 5           BL-C(2)
```

Recovered Sample Shape Prediction Results

Models	Recovered.Sample1	Recovered.Sample2	Recovered.Sample3	Recovered.Sample4
Random Forest	Flake	Flake	Spherical, Flake	spherical, flattened_spherical

Conclusion: The results above indicate that Recovered Sample 1 and Sample 2 are from the same brand ("RedDot"). The table for samples 3 and 4 shows that manufacturers are using multiple brands in the samples. For sample 3, the presence of the following brands Accurate4100, AccurateNo.2, AmericanSelect & RamshotEnforcer are there and in sample 4 Accurate4100, AccurateNo.2, BL-C(2) & RamshotEnforcer brands are found. The majority of the particles are from the brand "RamshotEnforcer" in sample 3 and sample 4 and the common brands in both sample 3 and sample 4 are Accurate4100, AccurateNo.2 & RamshotEnforcer.

Reference:

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R (2nd ed., p. 612). Springer. https://hastie.su.domains/ISLR2/ISLRv2_website.pdf
2. Wickham, H., Chang, W., & Henry, L. (n.d.). A box and whiskers plot (in the style of Tukey). ggplot2.tidyverse. Retrieved 10 March 2023, from https://ggplot2.tidyverse.org/reference/geom_boxplot.html
3. Statistics Globe (n.d.). Draw multiple Boxplots in one graph.statisticsglobe .Retrieved 10 March 2023, from https://statisticsglobe.com/draw-multiple-boxplots-in-one-graph-in-r
4. GGPLOT2 histogram plot : Quick Start Guide - R Software and Data Visualization. STHDA. (n.d.). Retrieved April 22, 2023, from http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization
5. ggplot2 barplots : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA. (2019). Sthda.com. http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization
6. Zach. (2022, August 3). How to Perform a Two Sample T-Test in R. Statology. Retrieved April 22, 2023, from https://www.statology.org/two-sample-t-test-in-r/
7. Bevans, R. (2022). ANOVA in R | A Complete Step-by-Step Guide with Examples. Scribbr.Retrieved April 22, 2023, from https://www.scribbr.com/statistics/anova-in-r/

8. Kruskal-Wallis Test | R Tutorial. (n.d.). Chi Yau. scribbr. Retrieved April 22, 2023, from https://www.r-tutor.com/elementary-statistics/non-parametric-methods/kruskal-wallis-test
9. Z. (2022, April 12). How to Split Data into Training & Test Sets in R. Statology. Retrieved March 11, 2023, from https://www.statology.org/train-test-split-r/
10. R: How to split a data frame into training, validation, and test sets?, Stack Overflow. Retrieved March 11, 2023, from https://stackoverflow.com/questions/36068963/r-how-to-split-a-data-frame-into-training-validation-and-test-sets.
11. (n.d.). Associations between Variables. Codecademy. Retrieved March 11, 2023, from https://www.codecademy.com/learn/stats-associations-between-variables/modules/stats-associations-between-variables/cheatsheet
12. Z. (2020, October 30). Linear Discriminant Analysis in R (Step-by-Step). Statology. Retrieved March 11, 2023, from https://www.statology.org/linear-discriminant-analysis-in-r/
13. Sarkar, Priyankur. "What Is LDA: Linear Discriminant Analysis for Machine Learning." What Is Linear Discriminant Analysis (LDA)?, Knowledgehut, 27 Dec. 2022, https://www.knowledgehut.com/blog/data-science/linear-discriminant-analysis-for-machine-learning.
14. Z. (2020, November 2). Quadratic Discriminant Analysis in R (Step-by-Step). Statology. Retrieved March 12, 2023, from https://www.statology.org/quadratic-discriminant-analysis-in-r/
15. Saunders, C. (2023, February 10). Classification Part 2 LDA and QDA [Lecture]. D2l.Sdbor.edu. https://d2l.sdbor.edu/d2l/le/content/1781558/viewContent/11116132/View
16. Finnstats. (2021, April 13). Random Forest in R: R-bloggers. R. Retrieved April 22, 2023, from https://www.r-bloggers.com/2021/04/random-forest-in-r/
17. Fraley, C., Raftery, A. E., & Scrucca, L. (n.d.). MclustDA discriminant analysis. Mclust-Org.Github. Retrieved March 13, 2023, from https://mclust-org.github.io/mclust/reference/MclustDA.html
18. shanem@mtu.edu, Shane T. Mueller. Model-Based Clustering and Mclust, 28 Mar. 2021, https://pages.mtu.edu/~shanem/psy5220/daily/Day19/modelbasedclustering.html#content.
19. Saunders, C. (2023, February 24). MclustDA part1 [Lecture]. D2l.Sdbor.edu. https://d2l.sdbor.edu/d2l/le/content/1781558/viewContent/11116023/View
20. Saunders, C. (2023, February 24). MclustDA part2 [Lecture]. D2l.Sdbor.edu. https://d2l.sdbor.edu/d2l/le/content/1781558/viewContent/11116022/View

21. Saunders, C. (2023, February 24). MclustDA part3 Cross validation [Lecture]. D2l.Sdbor.edu. https://d2l.sdbor.edu/d2l/le/content/1781558/viewContent/11116024/View
22. Saunders, C. (2023, February 24). Mclust Play Part2 R file [Lecture]. D2l.Sdbor.edu. https://d2l.sdbor.edu/d2l/le/content/1781558/viewContent/11116026/View
23. Saunders, C. (2023, February 24). Mclust Play Part3 R file [Lecture]. D2l.Sdbor.edu. https://d2l.sdbor.edu/d2l/le/content/1781558/viewContent/11116025/View
24. kTable: Make Nicely Formatted Tables in Kmisc: Kevin Miscellaneous. (n.d.). Rdrr.io. Retrieved April 25, 2023, from https://rdrr.io/cran/Kmisc/man/kTable.html
25. Chat.openai.com, https://chat.openai.com/.