Seeds and Data Quality EDA Project

Jordan Brown

Fairleigh Dickinson University

CSCI_3269_31: Introduction to Data Mining

Dr. Tamraparni Dasu

February 20, 2025

Please randomize your data set by choosing a sample with replacement from the seeds data set. The sample size should be the same as the original data set. The R command **sample()** might be useful. Please submit a report (along with your code) specifically addressing the following questions. You will be required to present your report in class. The examples are based on the penguin data set.

UG (100)

Answer Qs 1-6 for Seeds and Data_Quality data and compare.

1. Data description: number of attributes, number of objects, identify type of each attribute, e.g., quantitative, continuous, ratio. (10)

Seeds Original

Data description: number of attributes = 8, number of objects = 210, identify type of each attribute, e.g., quantitative, continuous, ratio. (10)

Area = quantitative, continuous, ratio; Perimeter = quantitative, continuous, ratio;
Compactness = quantitative, continuous, ratio; LengthKernel = quantitative, continuous, ratio;
WidthKernel = quantitative, continuous, ratio;
AsymmetryCoefficient = quantitative, continuous, ratio;
LengthKernelGroove = quantitative, continuous, ratio; Class = Qualitative, Nominal

Data Quality

Data description: number of attributes = 8, number of objects = 295, identify type of each attribute, e.g., quantitative, continuous, ratio. (10)

Area = quantitative, continuous, ratio; Perimeter = quantitative, continuous, ratio;
Compactness = quantitative, continuous, ratio; LengthKernel = quantitative, continuous, ratio;
WidthKernel = quantitative, continuous, ratio;
AsymmetryCoefficient = quantitative, continuous, ratio;
LengthKernelGroove = quantitative, continuous, ratio; Class = Qualitative, Nominal

2. Data quality issues: For missing values report the following. (10)
    a. number of impacted values, number of impacted objects
    b. % of impacted values, % impacted objects

    (Bonus points for finding other types of data quality issues!)

Seeds Original

```
> total_missing_values <- sum(is.na(seeds_o_sample))
> cat("Total number of missing values:", total_missing_values, "\n"
+ )
Total number of missing values: 0
>  missing_objects <- sum(rowSums(is.na(seeds_o_sample)) > 0)
>  cat("Number of objects (rows) with missing values:", missing_objects, "\n")
Number of objects (rows) with missing values: 0
```

a. number of impacted values = 0, number of impacted objects = 0
b. 0% of impacted values, 0% impacted objects

    (Bonus points for finding other types of data quality issues!)

Data Quality

```
> total_missing_values <- sum(is.na(dq_sample))
> cat("Total number of missing values:", total_missing_values, "\n")
Total number of missing values: 8
> missing_objects <- sum(rowSums(is.na(dq_sample)) > 0)
> cat("Number of objects (rows) with missing values:", missing_objects, "\n")
Number of objects (rows) with missing values: 7
```

a. number of impacted values = 8, number of impacted objects = 7

```
> total_values <- prod(dim(dq_sample))
> percentage_impacted_values <- (total_missing_values / total_values) * 100
> print(paste("Percentage of Impacted Values:", round(percentage_impacted_values, 2), "%"))
[1] "Percentage of Impacted Values: 0.34 %"
> total_rows <- nrow(dq_sample)
> percentage_impacted_objects <- (rows_with_missing / total_rows) * 100
Error: object 'rows_with_missing' not found
> percentage_impacted_objects <- (missing_objects / total_rows) * 100
> print(paste("Percentage of Impacted Objects:", round(percentage_impacted_objects, 2), "%"))
[1] "Percentage of Impacted Objects: 2.37 %"
```

b. 0.34% of impacted values, 2.37% impacted objects

    (Bonus points for finding other types of data quality issues!)

- Another data quality issue found are negative values. For example, there is a negative perimeter value. Perimeter cannot be negative.

3. Five number summary: (20) summary()

```
> summary(input_data[,3])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  32.10   39.23   44.45   43.92   48.50   59.60       2
>
```

   a. Min, Max, median, first and third quartiles of numerical attributes. https://www.geeksforgeeks.org/how-to-calculate-five-number-summary-in-r/

   b. Five number summary grouped by seed type. What do you notice?

Seeds Original

   a. Min, Max, median, first and third quartiles of numerical attributes. https://www.geeksforgeeks.org/how-to-calculate-five-number-summary-in-r/

```
> five_number_summary <- apply(seeds_o_sample[, sapply(seeds_o_sample, is.numeric)], 2, fivenum)
> print(five_number_summary)
      Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,] 10.590     12.41     0.80810        4.8990       2.630               0.7651              4.519
[2,] 12.260     13.45     0.85670        5.2620       2.941               2.5530              5.045
[3,] 14.355     14.32     0.87345        5.5235       3.237               3.5990              5.223
[4,] 17.320     15.73     0.88790        5.9800       3.562               4.7730              5.877
[5,] 21.180     17.25     0.91830        6.6750       4.033               8.4560              6.550
> summary(seeds_o_sample[,1])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.59   12.27   14.36   14.85   17.30   21.18
> summary(seeds_o_sample[,2])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.41   13.45   14.32   14.56   15.71   17.25
> summary(seeds_o_sample[,3])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.8081  0.8569  0.8734  0.8710  0.8878  0.9183
> summary(seeds_o_sample[,4])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.899   5.262   5.524   5.629   5.980   6.675
> summary(seeds_o_sample[,5])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.630   2.944   3.237   3.259   3.562   4.033
> summary(seeds_o_sample[,6])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7651  2.5615  3.5990  3.7002  4.7687  8.4560
> summary(seeds_o_sample[,7])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.519   5.045   5.223   5.408   5.877   6.550
```

b. Five number summary grouped by seed type. What do you notice?

```
> numeric_cols <- names(seeds_o_sample)[sapply(seeds_o_sample, is.numeric)]
> summary_data <- aggregate(seeds_o_sample[numeric_cols], by = list(Class = seeds_o_sample$Class), FUN = fivenum)

> by(seeds_o_sample[numeric_cols], seeds_o_sample$Class, function(x) apply(x, 2, fivenum))
seeds_o_sample$Class: Canadian
       Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,] 10.590     12.41     0.80810        4.899      2.6300                1.661             4.7450
[2,] 11.260     13.00     0.83350        5.136      2.7190                4.048             5.0020
[3,] 11.835     13.25     0.84935        5.224      2.8345                4.839             5.0915
[4,] 12.440     13.47     0.86200        5.325      2.9670                5.469             5.2310
[5,] 13.370     13.95     0.89770        5.541      3.2320                8.456             5.4910
------------------------------------------------------------------------------------------------------
seeds_o_sample$Class: Kama
       Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,] 11.230     12.63      0.8392        4.902      2.8500               0.7651              4.519
[2,] 13.740     13.94      0.8686        5.384      3.1290               1.7910              4.914
[3,] 14.355     14.32      0.8805        5.534      3.2435               2.5455              5.094
[4,] 15.050     14.75      0.8911        5.678      3.3790               3.3280              5.224
[5,] 17.080     15.46      0.9183        6.053      3.6830               6.6850              5.877
------------------------------------------------------------------------------------------------------
seeds_o_sample$Class: Rosa
       Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,] 15.38     14.66      0.8452       5.3630      3.2310               1.4720             5.1440
[2,] 17.32     15.73      0.8722       5.9790      3.5520               2.8430             5.8770
[3,] 18.72     16.21      0.8826       6.1485      3.6935               3.6095             5.9815
[4,] 19.14     16.57      0.8984       6.3150      3.8060               4.4510             6.1880
[5,] 21.18     17.25      0.9108       6.6750      4.0330               6.6820             6.5500
```

- I notice that when comparing the five number summary by seed types you get a lot more information about each of the seed types, like how Rosa, based on its data, seems to be the bigger seed type and Canadian seems to be the smallest. You have more information this way than getting information with all the data of the seed types combined.

Data Quality

1) **Five number summary including all values (even negative and NA)**

a. Min, Max, median, first and third quartiles of numerical attributes. https://www.geeksforgeeks.org/how-to-calculate-five-number-summary-in-r/

```
> five_number_summary <- apply(dq_sample[, sapply(dq_sample, is.numeric)], 2, fivenum)
> print(five_number_summary)
      Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,]  1.890   -13.450      0.8081        4.899     -3.5250               0.7651             4.519
[2,] 12.365    13.470      0.8567        5.314      2.9560               2.6400             5.088
[3,] 14.860    14.585      0.8747        5.597      3.3305               3.6345             5.310
[4,] 18.440    16.175      0.8883        6.111      3.6810               4.8250             5.939
[5,] 21.180    17.250      0.9183        6.675     12.9410               8.4560             6.550
```

```
> summary(dq_sample[,1])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.89   12.37   14.86   15.21   18.44   21.18
> summary(dq_sample[,2])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 -13.45   13.47   14.59   14.67   16.17   17.25       3
> summary(dq_sample[,3])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.8081  0.8567  0.8747  0.8710  0.8883  0.9183       2
> summary(dq_sample[,4])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  4.899   5.314   5.597   5.701   6.110   6.675       1
> summary(dq_sample[,5])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 -3.525   2.957   3.330   3.315   3.679  12.941       1
> summary(dq_sample[,6])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.7651  2.6520  3.6345  3.7382  4.8250  8.4560       1
> summary(dq_sample[,7])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.519   5.088   5.310   5.500   5.939   6.550
```

b. Five number summary grouped by seed type. What do you notice?

```
> numeric_cols <- names(dq_sample)[sapply(dq_sample, is.numeric)]
> summary_data <- aggregate(seeds_o_sample[numeric_cols], by = list(Class = seeds_o_sample$Class)

+ , FUN = fivenum)> summary_data <- aggregate(dq_sample[numeric_cols], by = list(Class = dq_sample$Class), FUN = fivenum)
> by(dq_sample[numeric_cols], dq_sample$Class, function(x) apply(x, 2, fivenum))
dq_sample$Class: Canadian
      Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,] 11.230     12.63     0.83920        4.902      2.8500               0.7651              4.519
[2,] 13.840     14.04     0.87100        5.386      3.1555               1.9510              4.956
[3,] 14.380     14.37     0.88185        5.543      3.2880               2.4620              5.097
[4,] 15.185     14.76     0.89230        5.678      3.3810               3.1780              5.223
[5,] 17.080     15.46     0.91830        6.053      3.6830               6.6850              5.877
------------------------------------------------------------------------------------------------
dq_sample$Class: Candy
      Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,] 12.78        NA      0.8716        5.262       3.026                1.176              4.782
[2,] 12.78        NA      0.8716        5.262       3.026                1.176              4.782
[3,] 12.78        NA      0.8716        5.262       3.026                1.176              4.782
[4,] 12.78        NA      0.8716        5.262       3.026                1.176              4.782
[5,] 12.78        NA      0.8716        5.262       3.026                1.176              4.782
------------------------------------------------------------------------------------------------
dq_sample$Class: Kama
      Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,]  1.89     14.66      0.8452        5.363      -3.525                1.472              5.144
[2,] 17.55     15.86      0.8722        6.017       3.561                2.843              5.879
[3,] 18.76     16.26      0.8823        6.163       3.719                3.600              6.032
[4,] 19.15     16.59      0.8969        6.341       3.806                4.451              6.200
[5,] 21.18     17.25      0.9108        6.675       4.033                6.682              6.550
------------------------------------------------------------------------------------------------
dq_sample$Class: Karma
      Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,] 20.88     17.05      0.8989       6.4500      4.0320                5.016              6.231
[2,] 20.88     17.05      0.8989       6.4500      4.0320                5.016              6.231
[3,] 21.03     17.13      0.9010       6.5115      4.0325                5.398              6.276
[4,] 21.18     17.21      0.9031       6.5730      4.0330                5.780              6.321
[5,] 21.18     17.21      0.9031       6.5730      4.0330                5.780              6.321
------------------------------------------------------------------------------------------------

------------------------------------------------------------------------------------------------
dq_sample$Class: Rosa
       Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,] 10.590    -13.45      0.8081        4.899       2.630                1.661             4.7450
[2,] 11.270     13.02      0.8291        5.160       2.719                4.048             5.0030
[3,] 11.835     13.29      0.8480        5.250       2.821                4.825             5.1605
[4,] 12.440     13.52      0.8596        5.357       2.967                5.469             5.2750
[5,] 13.370     13.95      0.8977        5.541      12.941                8.456             5.4910
------------------------------------------------------------------------------------------------
dq_sample$Class: Rose
    Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
208 13.2     13.66      0.8883        5.236       3.232                8.315             5.056
208 13.2     13.66      0.8883        5.236       3.232                8.315             5.056
208 13.2     13.66      0.8883        5.236       3.232                8.315             5.056
208 13.2     13.66      0.8883        5.236       3.232                8.315             5.056
208 13.2     13.66      0.8883        5.236       3.232                8.315             5.056
>
```

- I noticed that all the values for the Candy and Rose seed types are the same. The reason for them being the same is because there is only one row with the class Rose and the class Candy.

```
> specific_seed_data <- dq_sample[dq_sample$Class == "Rose", ]
> print(specific_seed_data)
    Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove Class
208 13.2     13.66      0.8883        5.236       3.232                8.315             5.056  Rose
> specific_seed_data <- dq_sample[dq_sample$Class == "Candy", ]
> print(specific_seed_data)
    Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove Class
65 12.78        NA      0.8716        5.262       3.026                1.176             4.782 Candy
> specific_seed_data <- dq_sample[dq_sample$Class == "Karma", ]
> print(specific_seed_data)
    Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove Class
90 20.88     17.05      0.9031        6.450       4.032                5.016             6.321 Karma
89 21.18     17.21      0.8989        6.573       4.033                5.780             6.231 Karma
> specific_seed_data <- dq_sample[dq_sample$Class == "Canadian", ]
> print(specific_seed_data)
     Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove    Class
22  14.11     14.26      0.8722        5.520       3.168                2.6880             5.219 Canadian
38  17.08     15.38      0.9079        5.832       3.683                2.9560             5.484 Canadian
9   16.63     15.46      0.8747        6.053       3.465                2.0400             5.877 Canadian
67  14.34     14.37      0.8726        5.630       3.190                1.3130             5.150 Canadian
26  16.19     15.16      0.8849        5.833       3.421                0.9030             5.307 Canadian
35  15.05     14.68      0.8779        5.712       3.328                2.1290             5.360 Canadian
288 14.29     14.09      0.9050        5.291       3.337                2.6990             4.825 Canadian
```

2) Five number summary after cleaning (removing negative and NA values)

a. Min, Max, median, first and third quartiles of numerical attributes. https://www.geeksforgeeks.org/how-to-calculate-five-number-summary-in-r/

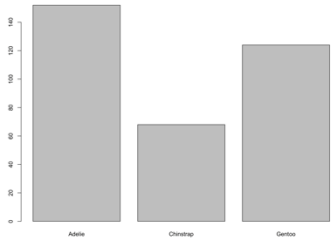
```
> cleaned_data <- dq_sample
> cleaned_data[cleaned_data < 0] <- NA
> cleaned_data <- na.omit(cleaned_data)
> summary(cleaned_data)
      Area          Perimeter      Compactness      LengthKernel    WidthKernel     AsymmetryCoefficient LengthKernelGroove     Class
 Min.   : 1.89   Min.   :12.41   Min.   :0.8081   Min.   :4.899   Min.   : 2.630   Min.   :0.7651       Min.   :4.519      Length:286
 1st Qu.:12.31   1st Qu.:13.48   1st Qu.:0.8565   1st Qu.:5.314   1st Qu.: 2.957   1st Qu.:2.6990       1st Qu.:5.088      Class :character
 Median :14.83   Median :14.56   Median :0.8747   Median :5.614   Median : 3.335   Median :3.6345       Median :5.310      Mode  :character
 Mean   :15.23   Mean   :14.77   Mean   :0.8710   Mean   :5.706   Mean   : 3.341   Mean   :3.7577       Mean   :5.508
 3rd Qu.:18.45   3rd Qu.:16.18   3rd Qu.:0.8883   3rd Qu.:6.113   3rd Qu.: 3.681   3rd Qu.:4.8460       3rd Qu.:5.964
 Max.   :21.18   Max.   :17.25   Max.   :0.9183   Max.   :6.675   Max.   :12.941   Max.   :8.4560       Max.   :6.550
> |
```

b. Five number summary grouped by seed type. What do you notice?

```
> numeric_cols <- names(cleaned_data)[sapply(cleaned_data, is.numeric)]
> summary_data <- aggregate(cleaned_data[numeric_cols], by = list(Class = cleaned_data$Class)

+ summary(summary_data)> summary_data <- aggregate(cleaned_data[numeric_cols], by = list(Class = cleaned_data$Class), FUN = fivenum)
> by(cleaned_data[numeric_cols], cleaned_data$Class, function(x) apply(x, 2, fivenum))
cleaned_data$Class: Canadian
      Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,] 11.230    12.630     0.83920        4.902       2.850               0.7651              4.519
[2,] 13.840    14.030     0.87160        5.386       3.154               1.9690              4.935
[3,] 14.380    14.350     0.88190        5.541       3.291               2.5040              5.097
[4,] 15.185    14.755     0.89335        5.676       3.388               3.1780              5.223
[5,] 17.080    15.460     0.91830        6.053       3.683               6.6850              5.877
------------------------------------------------------------------------------------------------------------------------------------------------------
cleaned_data$Class: Kama
      Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,]  1.89     14.66      0.8452       5.3630       3.231               1.472              5.144
[2,] 17.59     15.86      0.8722       6.0170       3.562               2.843              5.879
[3,] 18.76     16.26      0.8823       6.1720       3.719               3.563              6.053
[4,] 19.15     16.59      0.8973       6.3535       3.808               4.451              6.200
[5,] 21.18     17.25      0.9108       6.6750       4.033               6.682              6.550
------------------------------------------------------------------------------------------------------------------------------------------------------
cleaned_data$Class: Karma
      Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,] 20.88     17.05      0.8989       6.4500      4.0320               5.016              6.231
[2,] 20.88     17.05      0.8989       6.4500      4.0320               5.016              6.231
[3,] 21.03     17.13      0.9010       6.5115      4.0325               5.398              6.276
[4,] 21.18     17.21      0.9031       6.5730      4.0330               5.780              6.321
[5,] 21.18     17.21      0.9031       6.5730      4.0330               5.780              6.321
------------------------------------------------------------------------------------------------------------------------------------------------------
cleaned_data$Class: Rosa
      Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
[1,] 10.590     12.41      0.8081       4.8990       2.630               1.661             4.7450
[2,] 11.265     13.03      0.8310       5.1525       2.719               4.048             5.0075
[3,] 11.835     13.29      0.8480       5.2465       2.827               4.825             5.1605
[4,] 12.410     13.52      0.8596       5.3535       2.967               5.469             5.2750
[5,] 13.370     13.95      0.8977       5.5410      12.941               8.456             5.4910
------------------------------------------------------------------------------------------------------------------------------------------------------

------------------------------------------------------------------------------------------------------
cleaned_data$Class: Rose
    Area Perimeter Compactness LengthKernel WidthKernel AsymmetryCoefficient LengthKernelGroove
208 13.2     13.66      0.8883        5.236       3.232               8.315              5.056
208 13.2     13.66      0.8883        5.236       3.232               8.315              5.056
208 13.2     13.66      0.8883        5.236       3.232               8.315              5.056
208 13.2     13.66      0.8883        5.236       3.232               8.315              5.056
208 13.2     13.66      0.8883        5.236       3.232               8.315              5.056
```

- I've noticed that now with the data cleaned, candy is no longer included in the data set. This is because before there was only one row with the candy class, that row had NA as its value for perimeter. Because it was NA that row got removed and there are no other rows with the candy class. Also, since the data quality issues are removed, the data is more accurate.

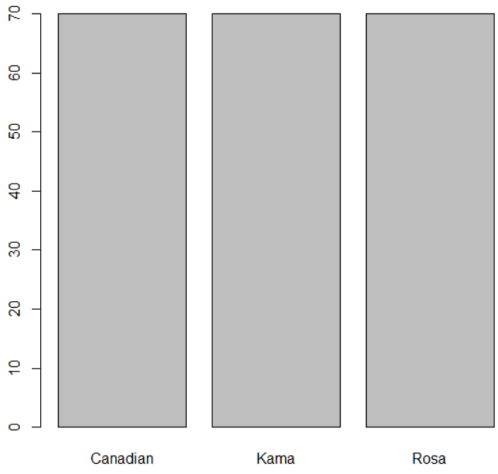
4. Frequency distributions of nominal variables (10) barplot(table())



https://www.geeksforgeeks.org/frequency-table-in-r/

Seeds Original

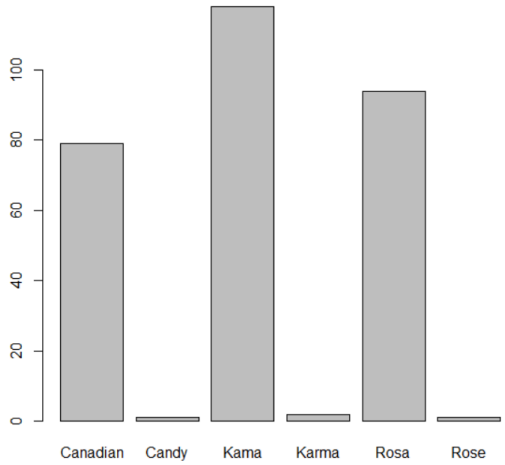
```
> table <- table(seeds_o_sample[,8])
> print(table)

Canadian     Kama     Rosa
      70       70       70
> barplot(table)
```



Data Quality

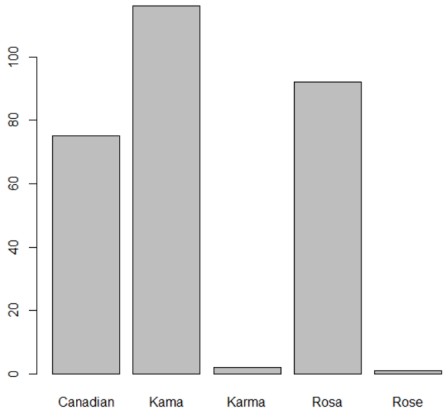
1) All values (even negative and NA)

```
> table <- table(dq_sample[,8])
> print(table)

Canadian    Candy     Kama    Karma     Rosa     Rose
      79        1      118        2       94        1
> barplot(table)
```



2) After cleaning (removing negative and NA values)

```
> table <- table(cleaned_data[,8])
> print(table)

Canadian     Kama    Karma     Rosa     Rose
      75      116        2       92        1
> barplot(table)
```

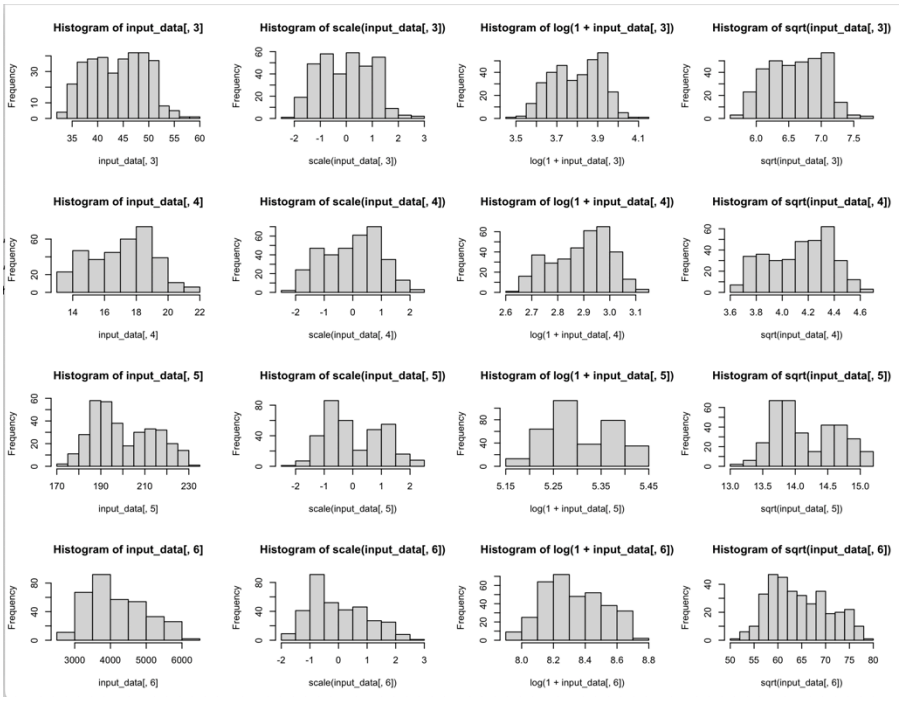


5. Histograms of the 7 quantitative variables, comment on their shapes. hist() https://cran.r-project.org/web/packages/lessR/vignettes/Histogram.html Cool tip: You can put multiple plots in a grid by using the command par(mfrow=c(nrows,ncols)) before using the plot command. I used par(mfrow=c(4,4)) for the chart below.

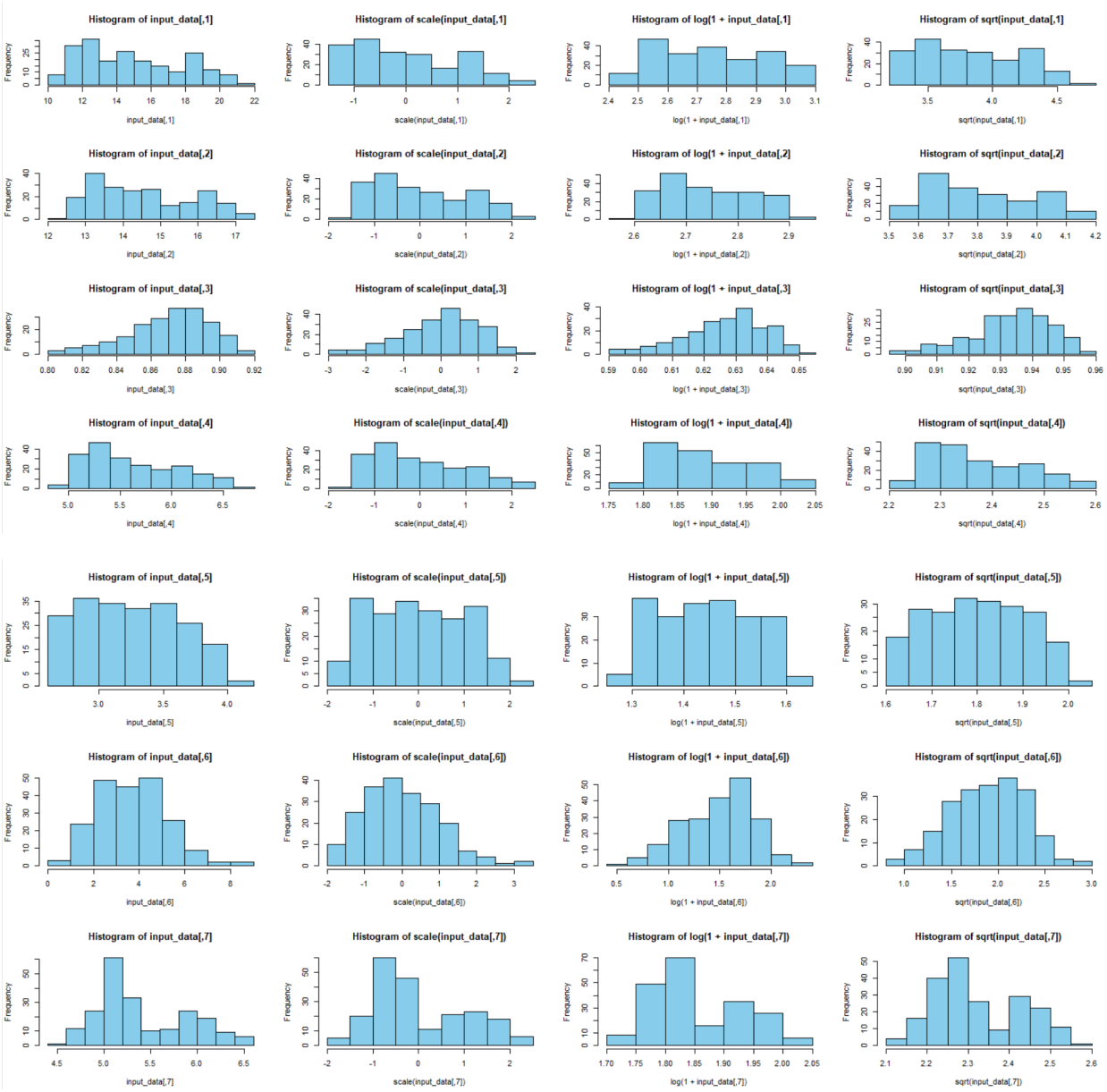
a. Raw values (10)

b. Standardized values, i.e. (x-(mean(x))/sd(x), scale() function in R. Which variable's histogram changes the least after the transformation? Why? (10)

c. Log of raw values, i.e., log(1+x) in R. Which variable's histogram changes the least after the transformation? Why? (10)



Seeds Original



a. Raw values (10)

b. Standardized values, i.e. (x-(mean(x))/sd(x), scale() function in R. Which variable's histogram changes the least after the transformation? Why? (10)

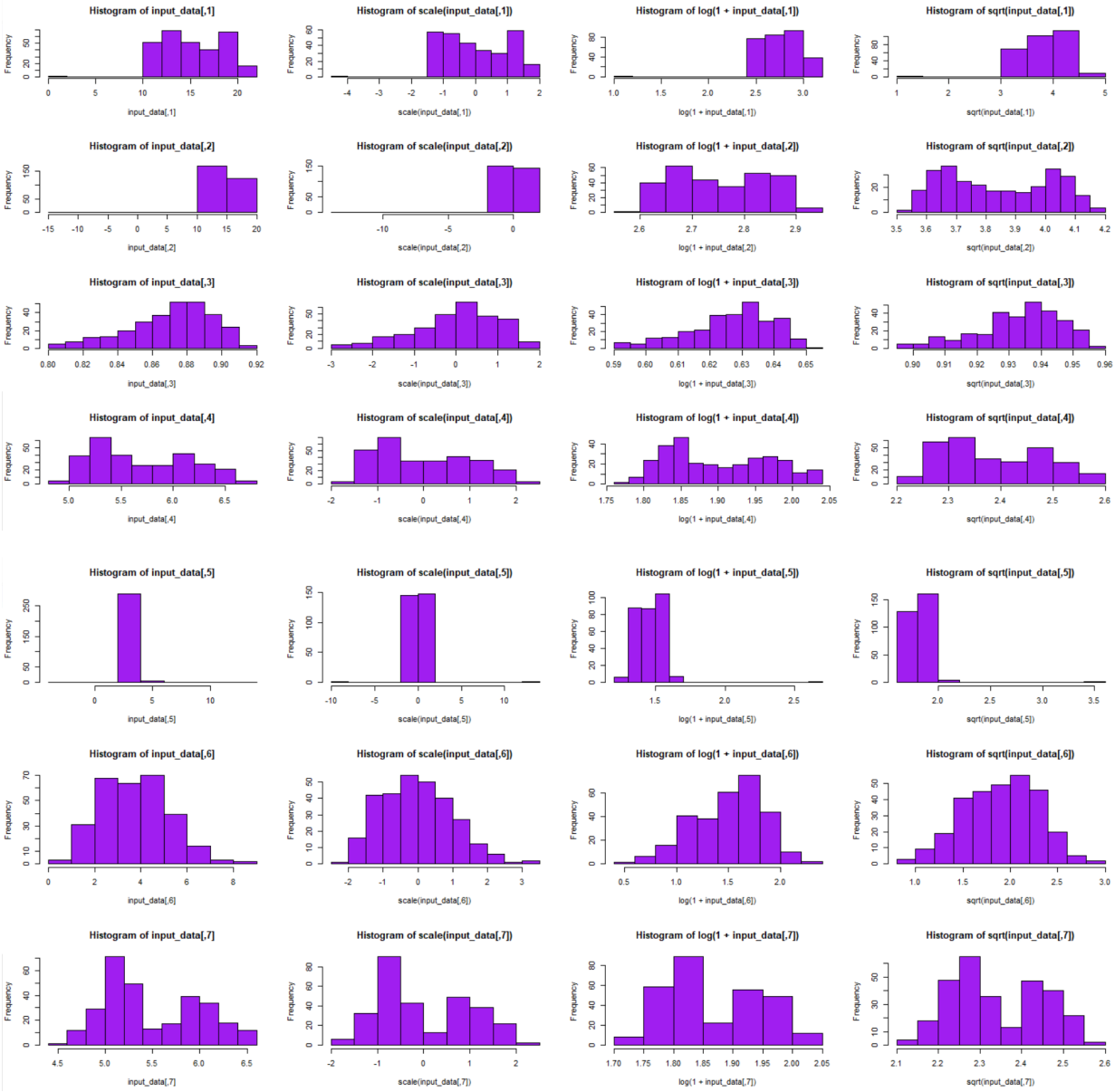
   i. LengthKernel because its shape is very similar to the shape of the raw values. There is a big increase at the start and then as the values increase, the frequency decreases.

c. Log of raw values, i.e., log(1+x) in R. Which variable's histogram changes the least after the transformation? Why? (10)

i. Compactness because for both the raw values and the log of the raw values, it slowly rises at the start and then towards the end it drops a bit quicker.

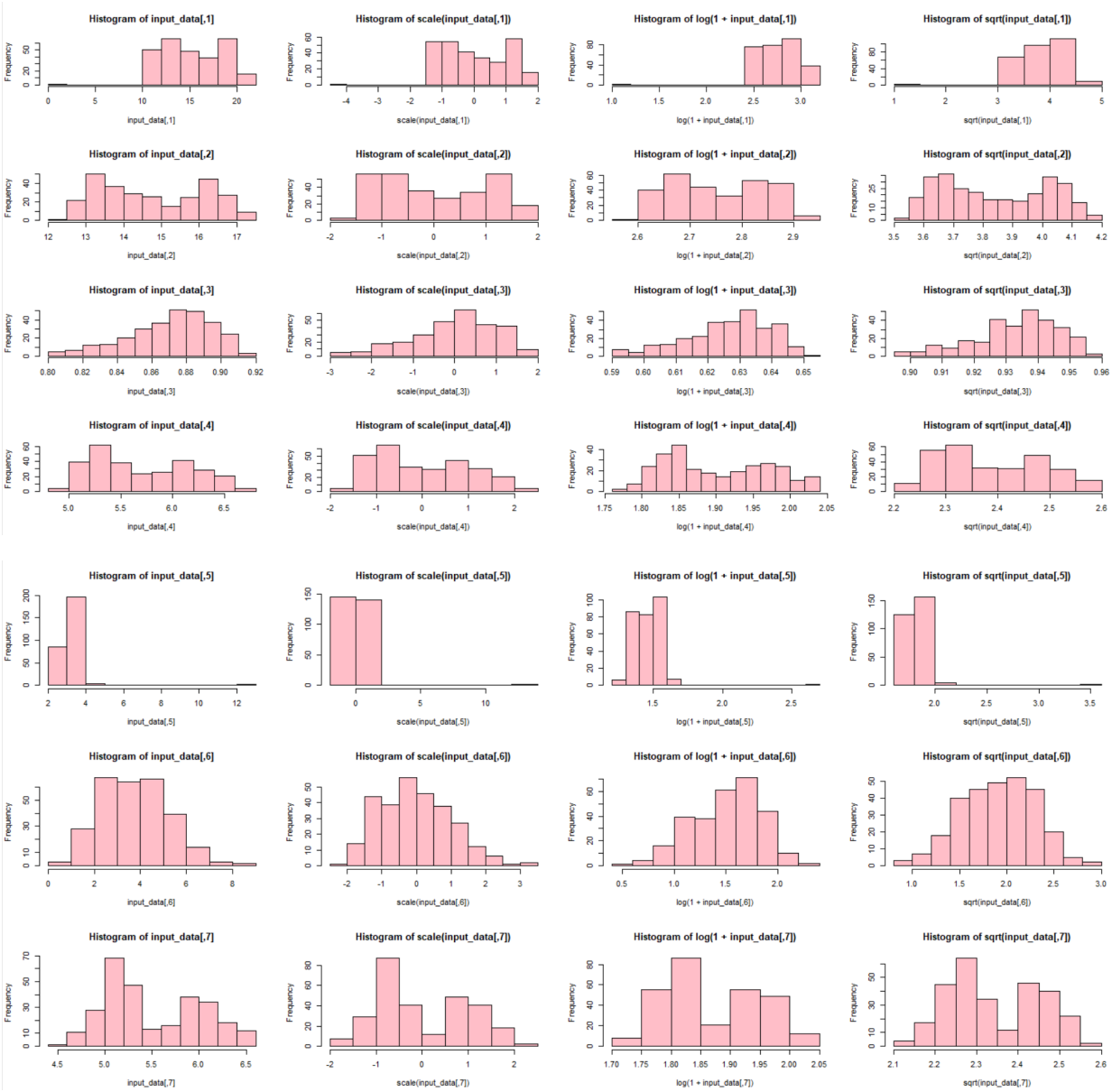
Data Quality

1) All values (even negative and NA)



d. Raw values (10)
e. Standardized values, i.e. (x-(mean(x))/sd(x), scale() function in R. Which variable's histogram changes the least after the transformation? Why? (10)
    i. Perimeter because there weren't many bins to compare, so there weren't many differences to spot.
f. Log of raw values, i.e., log(1+x) in R. Which variable's histogram changes the least after the transformation? Why? (10)
    i. Compactness because it follows the same pattern slowly increasing and then decreasing at the end.

2) After cleaning (removing negative and NA values)



g. Raw values (10)

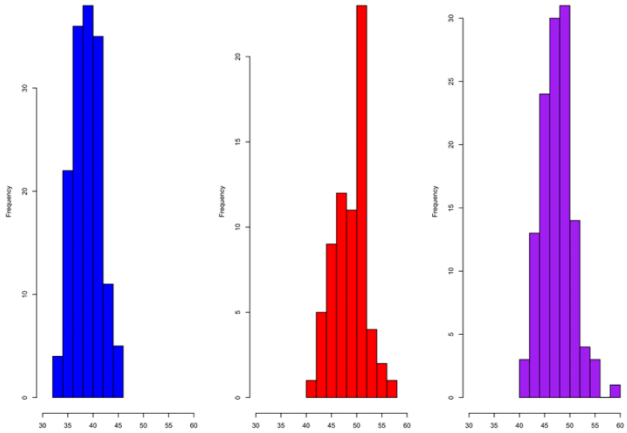
h. Standardized values, i.e. (x-(mean(x))/sd(x), scale() function in R. Which variable's histogram changes the least after the transformation? Why? (10)

   i. LengthKernelGroove because it has the same shape, big increase, then big decrease, then slight increase, and lastly , slight decrease.

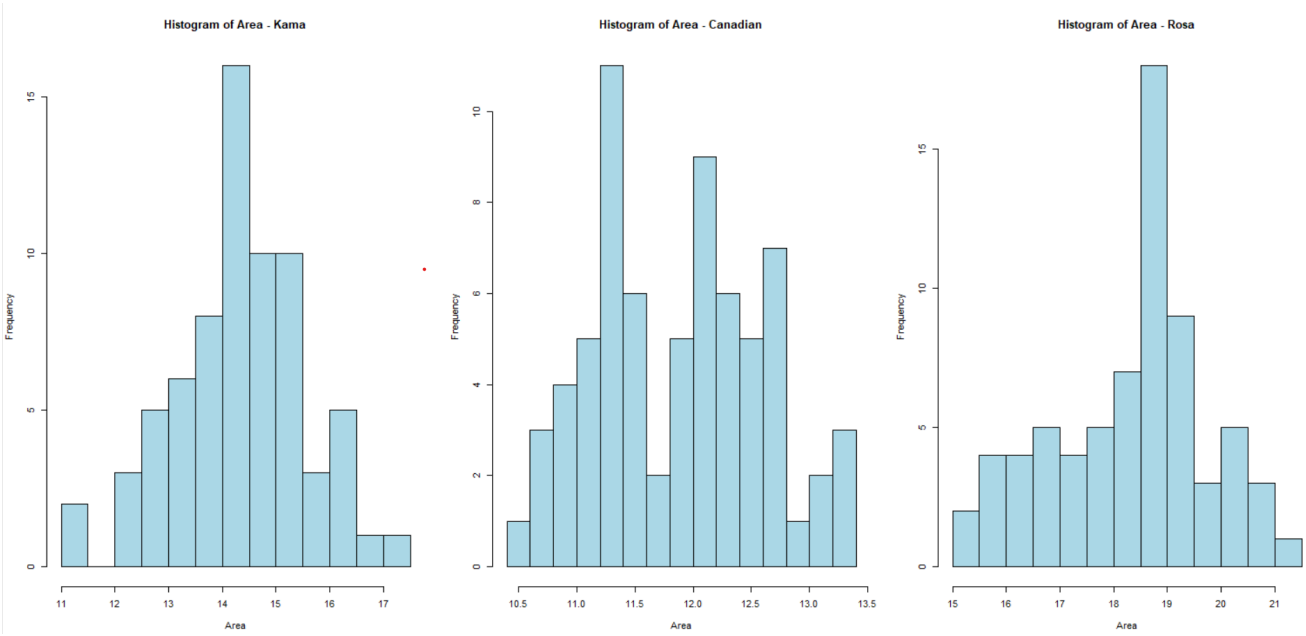
i. Log of raw values, i.e., log(1+x) in R. Which variable's histogram changes the least after the transformation? Why? (10)

i. LengthKernelGroove because it has the same shape, big increase, then big decrease, then slight increase, and lastly, a decrease.

6. Choose a numeric attribute of your choice and draw side-by-side histograms for each of the three seed types and compare (penguins example below). What do you notice? (20)



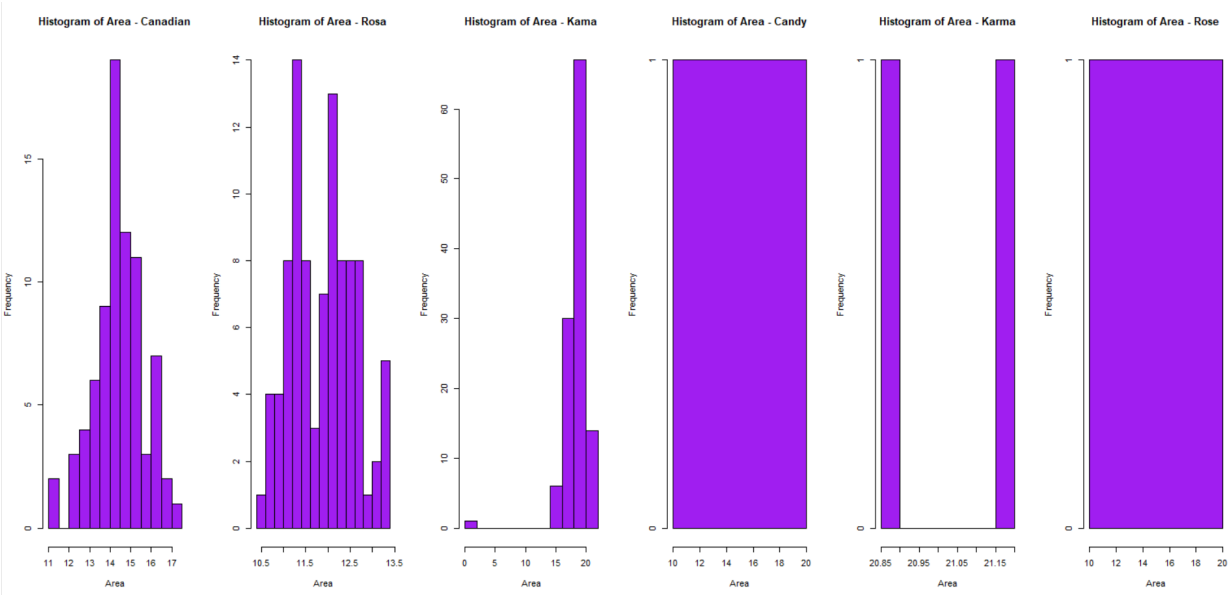
Seeds Original



- All of their areas are different from each other. The graphs are very different from each other. Their peak areas are all different, for Kama, it's around 14/15, for Canadian, it's around 11, and for Rosa, it's around 18/19.

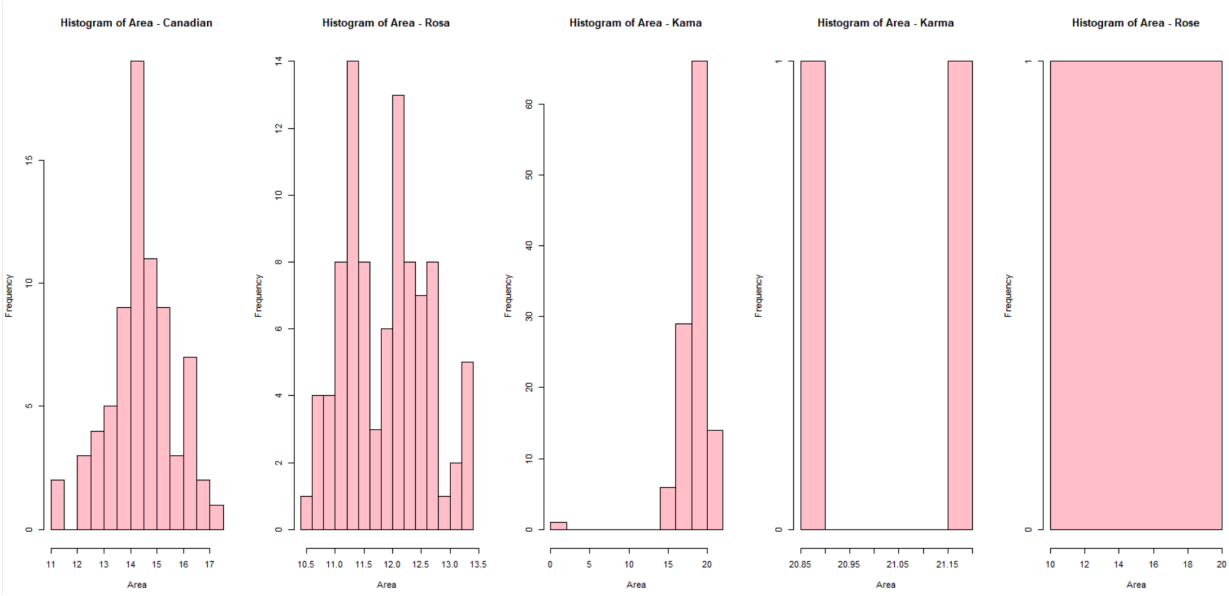
Data Quality

1) All values (even negative and NA)



- Very different graphs from each other. For Candy and Rose, there is one big bar. This is due to there being only one instance of each of those classes.

2) After cleaning (removing negative and NA values)



- Very different graphs from each other, but aren't very different from before the data was cleaned. One noticeable difference is there being one less histogram, Candy.