

Unveiling the Morphological Characteristics of Seeds: A Comparative Data-Driven Approach to Understanding Domestication and Wild Traits

CAS in Applied Data Science
Module 2



Jared Diamond's questions:

“why were so few species domesticated, in so few locations ?”

Why were some species domesticated and others not is especially hard to answer because only a handful of species were domesticated and these are usually unrelated taxonomically and grow in different locations and climates.

- Domestication events are rare but foundational events of civilisations.
- The two groups of plants that were most often domesticated: the grasses (sugar source), and the legumes (protein source).
- One plant group of 300 species, the pulses (*Viciae*), was domesticated exceptionally [REDACTED] exceptional location: the fertile crescent.

Study System



Vicia faba
(Faba bean)



Lens culinaris ssp. *Culinaris*
(Lentils)



Lathyrus sativus
(Sweet pea)



Pisum sativum ssp. *Sativum*
(Peas)

Data collection



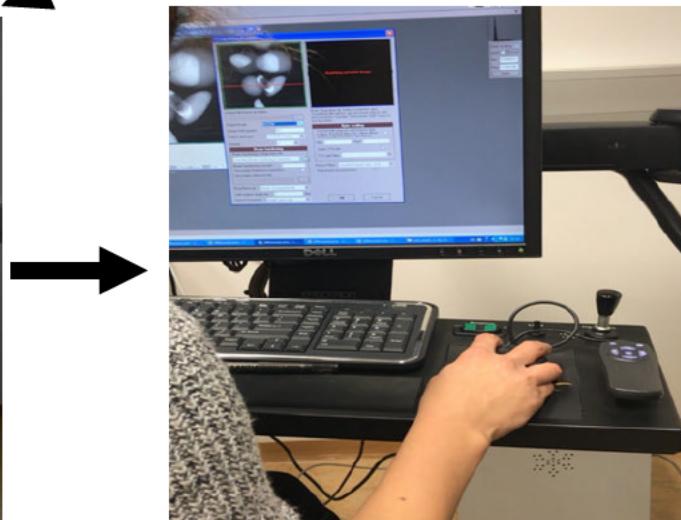
Seeds (mostly from herbarium loan)



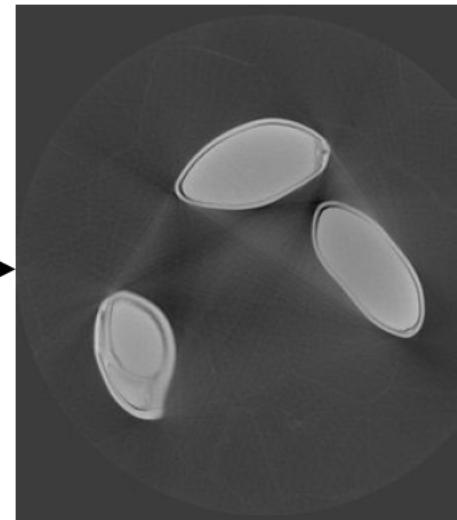
Mounted in Styrofoam
(X-Ray transparent)



Transfer to CT scanner



Scan programming



Scan data: stack of
grayscale pictures

Image and data processing pipeline

- How do we get from

This

to

that

3D data

?
Morphometric
data

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | | | | | |
|----|-------|--------|--------------|-------------|-------------|-------------|---|---------|----------|----------|---------|----------|----------|---------|----------|----------|----------|----------|----------|--------|--|
| 1 | Label | Volume | SurfaceArea | MeanBreadth | Spheriocity | EulerNumber | 1 | 210.228 | 205.819 | 56.798 | 210.229 | 205.819 | 56.799 | 45.983 | 44.12 | 16.515 | 115.46 | 18.253 | | | |
| 2 | 1 | 133500 | 10617.663 | 77.338 | 0.439 | | 1 | 135.304 | 181.588 | 58.951 | 135.304 | 181.588 | 58.951 | 39.976 | 39.217 | 16.406 | -12.554 | -12.251 | | | |
| 3 | 2 | 103594 | 13236.717 | 68.62 | 0.517 | | 1 | 176.228 | 87.011 | 66.02 | 178.228 | 87.011 | 66.02 | 48.986 | 44.519 | 16.861 | 20.401 | -8.632 | | | |
| 4 | 3 | 148842 | 17955.725 | 79.967 | 0.455 | | 1 | 54.414 | 138.485 | 151.022 | 54.414 | 138.485 | 151.022 | 48.269 | 45.108 | 19.606 | 119.845 | 0.531 | | | |
| 5 | 4 | 177679 | 18002.591 | 79.181 | 0.612 | | 1 | 176.511 | 101.358 | 180.18 | 178.511 | 101.358 | 180.18 | 48.491 | 44.447 | 21.343 | 29.405 | -44.691 | | | |
| 6 | 5 | 194782 | 19181.37 | 81.798 | 0.808 | | 1 | 116.481 | 178.264 | 156.542 | 116.481 | 178.264 | 156.542 | 49.579 | 45.595 | 20.734 | 137.133 | 41.459 | | | |
| 7 | 6 | 179162 | 17945.229 | 78.911 | 0.620 | | 1 | 176.511 | 101.358 | 180.18 | 178.511 | 101.358 | 180.18 | 48.491 | 44.447 | 21.343 | 29.405 | -44.691 | | | |
| 8 | 7 | 167023 | 16959.165 | 76.582 | 0.645 | | 1 | 183.281 | 244.047 | 161.538 | 183.281 | 244.047 | 161.538 | 44.736 | 43.625 | 20.954 | 166.432 | -25.064 | | | |
| 9 | 8 | 154202 | 16065.445 | 74.58 | 0.65 | | 1 | 152.842 | 157.125 | 231.923 | 152.842 | 157.125 | 231.923 | 42.305 | 40.153 | 20.153 | 19.977 | -19.457 | | | |
| 10 | 9 | 162629 | 16925.597 | 69.70 | 0.5 | | 1 | 171.181 | 91.312 | 230.229 | 171.181 | 91.312 | 230.229 | 42.459 | 39.371 | 15.153 | 19.893 | -24.457 | | | |
| 11 | 10 | 188542 | 18995.458 | 80.654 | 0.601 | | 1 | 134.353 | 229.853 | 237.72 | 134.353 | 229.853 | 237.72 | 47.637 | 47.489 | 10.021 | 121.238 | -50.154 | | | |
| 12 | 11 | 247139 | 22307.339 | 88.651 | 0.607 | | 1 | 194.816 | 153.655 | 59.413 | 194.816 | 153.655 | 59.413 | 48.655 | 58.413 | 53.307 | 49.789 | 22.525 | -17.484 | 42.469 | |
| 13 | 12 | 241728 | 21573.237 | 86.298 | 0.658 | | 1 | 122.327 | 184.308 | 74.236 | 122.327 | 184.308 | 74.236 | 50.568 | 47.78 | 24.595 | -117.355 | 11.807 | | | |
| 14 | 13 | 241285 | 22204.503 | 87.978 | 0.652 | | 1 | 65.838 | 238.45 | 83.31 | 65.838 | 238.45 | 83.31 | 53.715 | 47.78 | 23.57 | 25.4 | -27.434 | | | |
| 15 | 14 | 100492 | 12385.4 | 67.215 | 0.54 | | 1 | 101.281 | 175.285 | 232.232 | 101.281 | 175.285 | 232.232 | 41.142 | 39.525 | 14.857 | 85.901 | 10.12 | | | |
| 16 | 15 | 231033 | 115 | 64.088 | 0.486 | | 1 | 197.712 | 196.513 | 234.085 | 197.712 | 196.513 | 234.085 | 39.955 | 37.894 | 13.3 | 18.905 | -22.717 | | | |
| 17 | 16 | 100491 | 100 | 61.761 | 0.483 | | 1 | 44.6 | 161.951 | 235.158 | 44.6 | 161.951 | 235.158 | 38.108 | 39.1 | 16.704 | 12.626 | 88.285 | 30.54 | | |
| 18 | 17 | 147570 | 14757.455 | 70.77 | 0.55 | | 1 | 20.701 | 101.1 | 22.241 | 20.701 | 101.1 | 22.241 | 51.1 | 33.9 | 9 | 18.841 | 2.61 | | | |
| 19 | 18 | 20508 | 20508 | 81.96 | 0.54 | | 1 | 7.3 | 10.8 | 14.452 | 10.3 | 10.8 | 14.452 | 48.7 | 48.6 | 20.674 | 2.6 | -4.336 | | | |
| 20 | 19 | 30245 | 30245 | 83.324 | 0.579 | | 1 | 0.5 | 21.5 | 10.375 | 12.3 | 0.5 | 21.5 | 48.1 | 48.3 | 16.444 | 1.1 | 15.728 | | | |
| 21 | 20 | 222474 | 222474 | 10.327 | 0.579 | | 1 | 1.2 | 8.2 | 10.84 | 1.2 | 8.2 | 10.84 | 18.831 | 17.831 | 1.2 | 30.501 | 0.301 | | | |
| 22 | 23 | 170759 | 19407.179 | 88.94 | 0.520 | | 1 | 55.663 | 184.355 | 179.877 | 55.663 | 184.355 | 179.877 | 49.003 | 47.216 | 21.215 | 150.325 | 10.267 | | | |
| 23 | 24 | 223974 | 22203.547 | 88.838 | 0.514 | | 1 | 224.145 | 194.082 | 58.645 | 224.145 | 194.082 | 58.645 | 53.836 | 49.753 | 21.077 | -35.002 | -0.693 | | | |
| 24 | 25 | 118920 | 16949.783 | 78.112 | 0.347 | | 1 | 92.307 | 233.593 | 95.316 | 92.307 | 233.593 | 95.316 | 48.230 | 13.474 | -23.631 | 13.704 | | | | |
| 25 | 26 | 108787 | 15635.057 | 76.096 | 0.35 | | 1 | 168.55 | 84.17 | 73.311 | 168.55 | 84.17 | 73.311 | 48.093 | 44.149 | 12.886 | 51.19 | 0.849 | | | |
| 26 | 27 | 111330 | 16320.129 | 78.073 | 0.322 | | 1 | 59.997 | 118.972 | 80.292 | 59.997 | 118.972 | 80.292 | 49.527 | 41.107 | 13.844 | -30.416 | 3.395 | | | |
| 27 | 28 | 19629 | 15376.749 | 74.589 | 0.445 | | 1 | 106.478 | 238.533 | 190.944 | 106.478 | 238.533 | 190.944 | 44.311 | 42.331 | 15.937 | -157.132 | -2.416 | | | |
| 28 | 29 | 129004 | 10179.971 | 76.578 | 0.448 | | 1 | 149.815 | 102.057 | 204.167 | 149.815 | 102.057 | 204.167 | 45.848 | 42.458 | 16.708 | 38.467 | -2.084 | | | |
| 29 | 30 | 158560 | 14959.753 | 78.733 | 0.522 | | 1 | 49.941 | 178.794 | 80.382 | 49.941 | 178.794 | 80.382 | 48.325 | 43.449 | 18.137 | -37.299 | 38.647 | | | |
| 30 | 31 | 149570 | 14077.927 | 72.145 | 0.378 | | 1 | 105 | 134.449 | 21.111 | 105 | 134.449 | 21.111 | 105.773 | 83.457 | 40.203 | 38.983 | 14.998 | -2.865 | 5.636 | |
| 31 | 32 | 149550 | 14077.258 | 77.764 | 0.518 | | 1 | 47 | 205.662 | 40.465 | 47 | 205.662 | 40.465 | 185.405 | 42.414 | 43.513 | 18.529 | -120.049 | -17.351 | | |
| 32 | 33 | 149549 | 14077.258 | 77.763 | 0.549 | | 1 | 1.2 | 88.111 | 122.111 | 0.1 | 1.2 | 88.111 | 122.111 | 20.78 | 48.074 | 34.54 | 19.529 | 2.6 | | |
| 33 | 34 | 170058 | 17741.177 | 79.149 | 0.581 | | 1 | 5.973 | 156.875 | 87.5 | 5.973 | 156.875 | 87.5 | 844 | 218.971 | 49.953 | 44.394 | 20.129 | -37.724 | 34.048 | |
| 34 | 35 | 1 | 356287.9 | 7065746.82 | 0.5273 | 221.6-6 | | 1 | 1.2 | 88.111 | 122.111 | 0.1 | 1.2 | 88.111 | 122.111 | 3.543 | 0.1 | 0.051 | -11.954 | -0.003 | |
| 35 | 36 | 2 | 344214.196 | 360479.341 | 0.592 | 182.62E-7 | | 1 | 26.121 | 174.551 | 1.445 | 26.121 | 174.551 | 1.445 | 110.969 | 859.287 | 0.059 | -56.769 | 1.1288-4 | | |
| 36 | 37 | 3 | 231039.744 | 2102188.501 | 0.398 | 359.832E-7 | | 1 | 304.430 | 329.803 | 0.177 | 304.430 | 329.803 | 0.177 | 102.269 | 554.083 | 0.177 | 44.238 | 0.1 | | |
| 37 | 38 | 4 | 317315.293 | 2518520.87 | 0.3837 | 126E-7 | | 1 | 597.44 | 474.123 | 0.476 | 597.44 | 474.123 | 0.476 | 113.534 | 592.654 | 0.117 | 37.82 | -0.012 | | |
| 38 | 39 | 5 | 314970.847 | 2727393.643 | 0.3945 | 526E-7 | | 1 | 339.359 | 289.548 | 0.476 | 339.359 | 289.548 | 0.476 | 109.701 | 665.571 | 0.106 | -38.8364 | 0.000E-4 | | |
| 39 | 40 | 6 | 365351.484 | 3125938.081 | 0.4074 | 942E-7 | | 1 | 157.117 | 534.776 | 0.558 | 157.117 | 534.776 | 0.558 | 109.208 | 697.123 | 0.106 | -31.0517 | 620E-4 | | |
| 40 | 41 | 7 | 387851.864 | 4755674.819 | 0.4191 | 582E-7 | | 1 | 383.701 | 708.485 | 0.549 | 383.701 | 708.485 | 0.549 | 131.002 | 495.282 | 0.076 | -79.971 | -0.002 | | |
| 41 | 42 | 8 | 533376.371 | 3575865.765 | 0.4367 | 144E-7 | | 1 | 6208.457 | 3038.986 | 0.727 | 6208.457 | 3038.986 | 0.727 | 133.936 | 699.303 | 0.136 | 66.7273 | 180E-4 | | |
| 42 | 43 | 9 | 512382.578 | 5102903.161 | 0.452 | 180E-7 | | 1 | 4855.481 | 4299.941 | 0.805 | 4855.481 | 4299.941 | 0.805 | 129.007 | 1041.969 | 0.096 | 28.1011 | 789E-4 | | |
| 43 | 44 | 10 | 192300.103 | 192300.262 | 0.487 | 1.5E-7 | | 1 | 6172.544 | 1087.502 | 0.827 | 6172.544 | 1087.502 | 0.827 | 859.133 | 1107.209 | 0.097 | 20.204 | 1.00E-4 | | |
| 44 | 45 | 11 | 1.696304.016 | 422954.77 | 0.4797 | 247E-7 | | 1 | 6172.544 | 1087.502 | 0.228 | 6172.544 | 1087.502 | 0.228 | 222.722 | 769.909 | 0.143 | 78.907 | -0.001 | | |
| 45 | 46 | 12 | 784714.596 | 4333338.346 | 0.5698 | 778E-7 | | 1 | 3587.795 | 182.114 | 0.276 | 3587.795 | 182.114 | 0.276 | 178.114 | 148.496 | 0.743 | 16.789 | 15.712 | | |
| 46 | 47 | 13 | 840459.274 | 101111.421 | 0.5621 | 202E-6 | | 1 | 1929.758 | 5567.758 | 0.256 | 1929.758 | 5567.758 | 0.256 | 236.188 | 494.419 | 0.194 | -38.4751 | 4.119E-4 | | |
| 47 | 48 | 14 | 785757.738 | 4881.652 | 0.6398 | 158E-7 | | 1 | 4172.389 | 759.075 | 0.273 | 4172.389 | 759.075 | 0.273 | 135.405 | 941.591 | 0.149 | -34.9468 | 4.109E-4 | | |
| 48 | 49 | 15 | 856130.235 | 4123937.924 | 0.598 | 842E-7 | | 1 | 6391.54 | 4506.814 | 0.73 | 6391.54 | 4506.814 | 0.73 | 127.327 | 908.109 | 0.144 | -21.742 | 4.47E-4 | | |
| 49 | 50 | 16 | 556843.706 | 4744008.757 | 0.574 | 533E-7 | | 1 | 1293.191 | 570.0 | 0.777 | 1293.191 | 570.0 | 0.777 | 115.339 | 951.783 | 0.129 | 43.403 | 9.441E-4 | | |
| 50 | 51 | 17 | 624439.181 | 6436703.013 | 0.6091 | 854E-7 | | 1 | 4308.839 | 1946.407 | 0.751 | 4308.839 | 1946.407 | 0.751 | 121.021 | 1231.252 | 0.103 | 3.0541 | 328E-4 | | |
| 51 | 52 | 18 | 850883.325 | 0.120139.44 | 0.6032 | 808E-6 | | 1 | 5413.559 | 792.395 | 0.808 | 5413.559 | 792.395 | 0.808 | 1251.277 | 1225.532 | 0.103 | 3.0541 | 328E-4 | | |

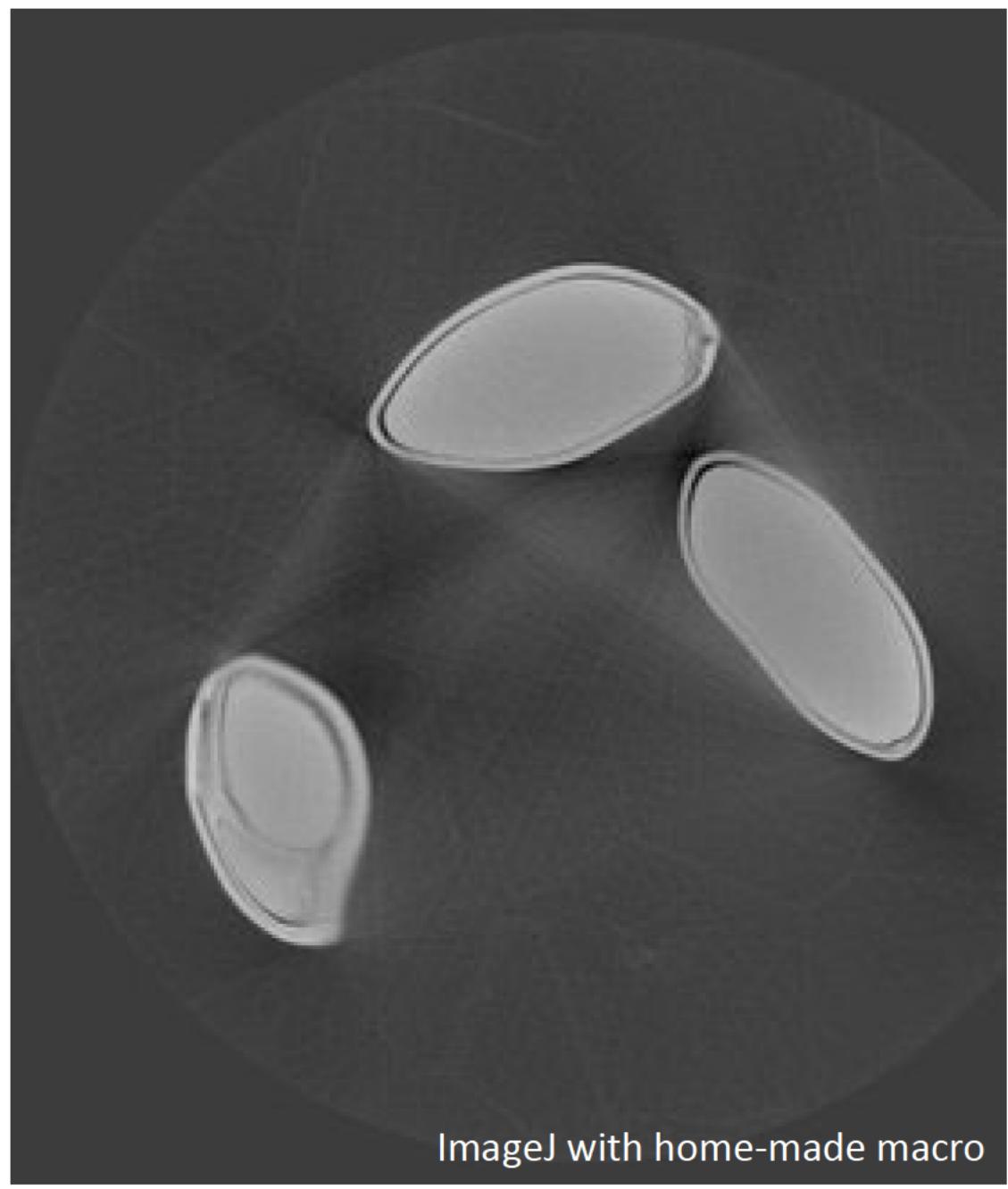
*Processing_macro.ijm

File Edit Language Templates Run Tools Tabs Options

*Processing_macro.ijm

```
1 run("8-bit");
2
3 run("Next Slice [>]");run("Next Slice [>]");
4 run("Next Slice [>]");run("Next Slice [>]");
5 run("Next Slice [>]");run("Next Slice [>]");
6 run("Next Slice [>]");run("Next Slice [>]");
7 run("Next Slice [>]");run("Next Slice [>]");
8 run("Next Slice [>]");run("Next Slice [>]");
9
10 //run("Brightness/Contrast...");
11 setMinAndMax(0, 64000);
12
13
14 run("Convert to Mask",
15 "method=Li background=Dark");
16
17 run("Invert", "stack");
18 run("Erode", "stack");
19 run("Dilate", "stack");
20 run("Dilate", "stack");
21 run("Fill Holes", "stack");
22 run("Erode", "stack");
23 run("Invert", "stack");
24
25
26 run("Connected Components Labeling",
27 "connectivity=26 type=[8 bits]");
28
29 run("Analyze Regions 3D",
30 "volume surface_area mean_breadth sphericity e");
31
32
33 saveAs("Results");
34 saveAs("Tiff");
35 run("Close All");
```

Batch Kill persistent Show Errors Clear



ImageJ with home-made macro

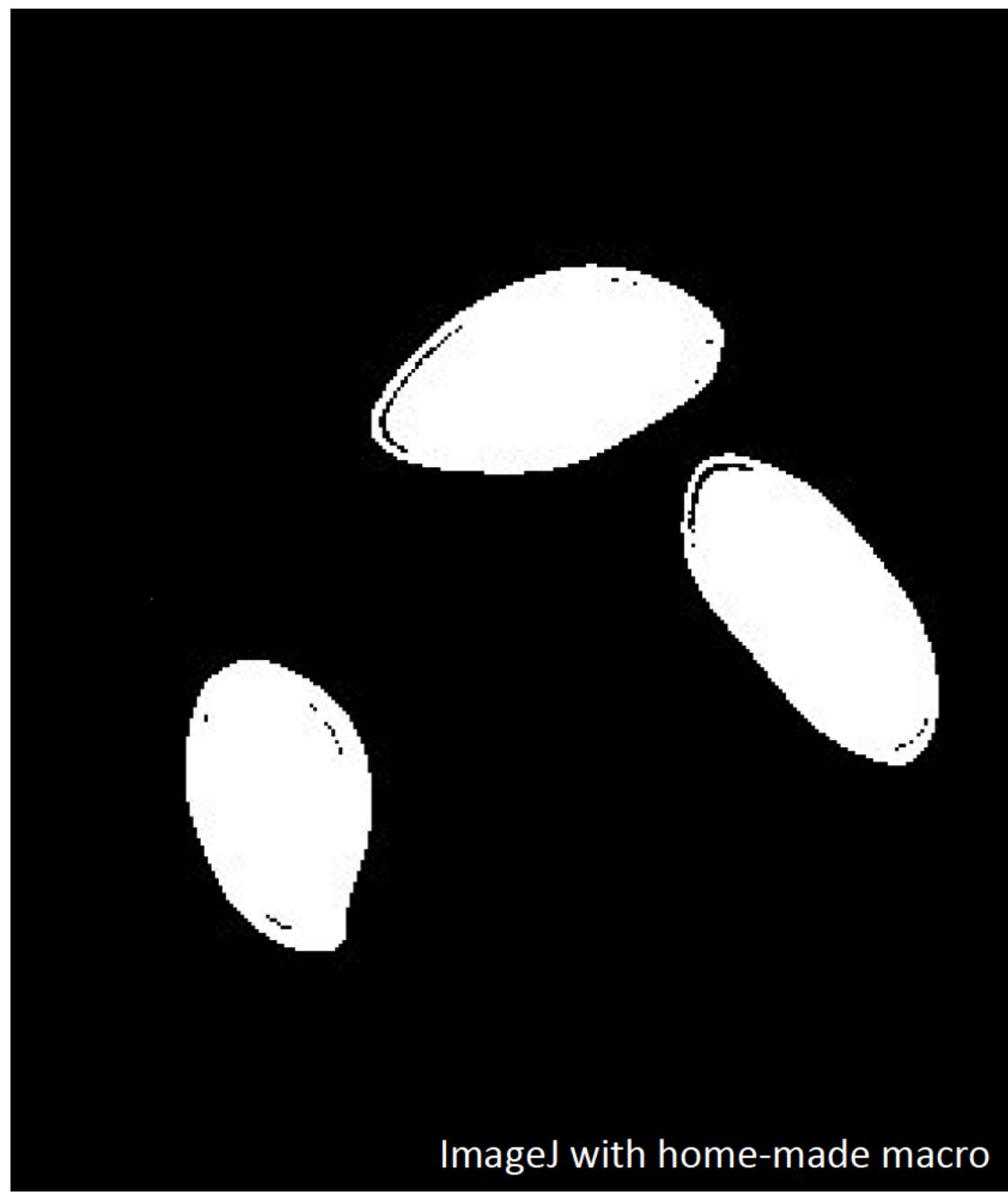
*Processing_macro.ijm

File Edit Language Templates Run Tools Tabs Options

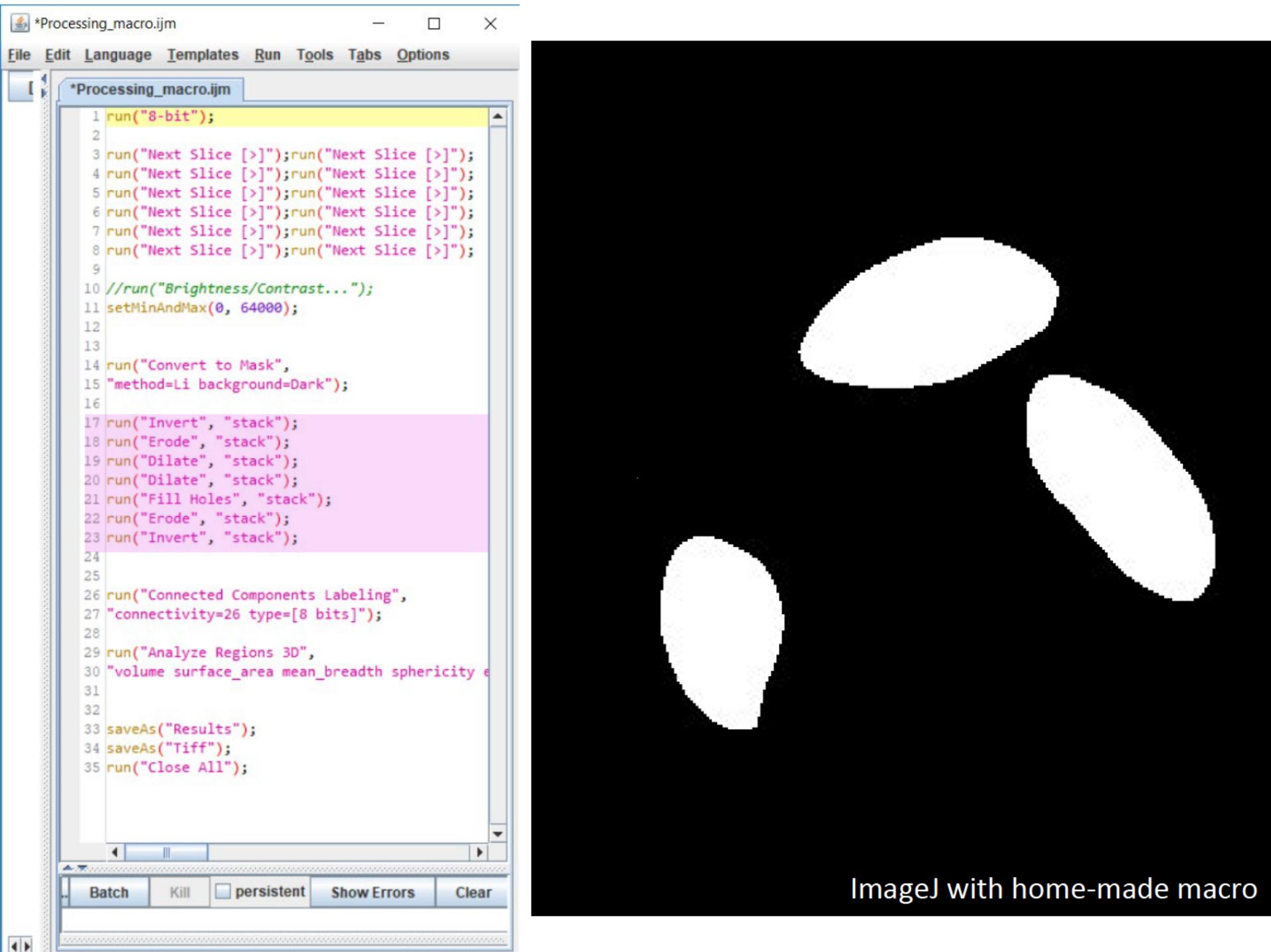
*Processing_macro.ijm

```
1 run("8-bit");
2
3 run("Next Slice [>]");run("Next Slice [>]");
4 run("Next Slice [>]");run("Next Slice [>]");
5 run("Next Slice [>]");run("Next Slice [>]");
6 run("Next Slice [>]");run("Next Slice [>]");
7 run("Next Slice [>]");run("Next Slice [>]");
8 run("Next Slice [>]");run("Next Slice [>]");
9
10 //run("Brightness/Contrast...");
11 setMinAndMax(0, 64000);
12
13
14 run("Convert to Mask",
15 "method=Li background=Dark");
16
17 run("Invert", "stack");
18 run("Erode", "stack");
19 run("Dilate", "stack");
20 run("Dilate", "stack");
21 run("Fill Holes", "stack");
22 run("Erode", "stack");
23 run("Invert", "stack");
24
25
26 run("Connected Components Labeling",
27 "connectivity=26 type=[8 bits]");
28
29 run("Analyze Regions 3D",
30 "volume surface_area mean_breadth sphericity e");
31
32
33 saveAs("Results");
34 saveAs("Tiff");
35 run("Close All");
```

Batch Kill persistent Show Errors Clear



ImageJ with home-made macro



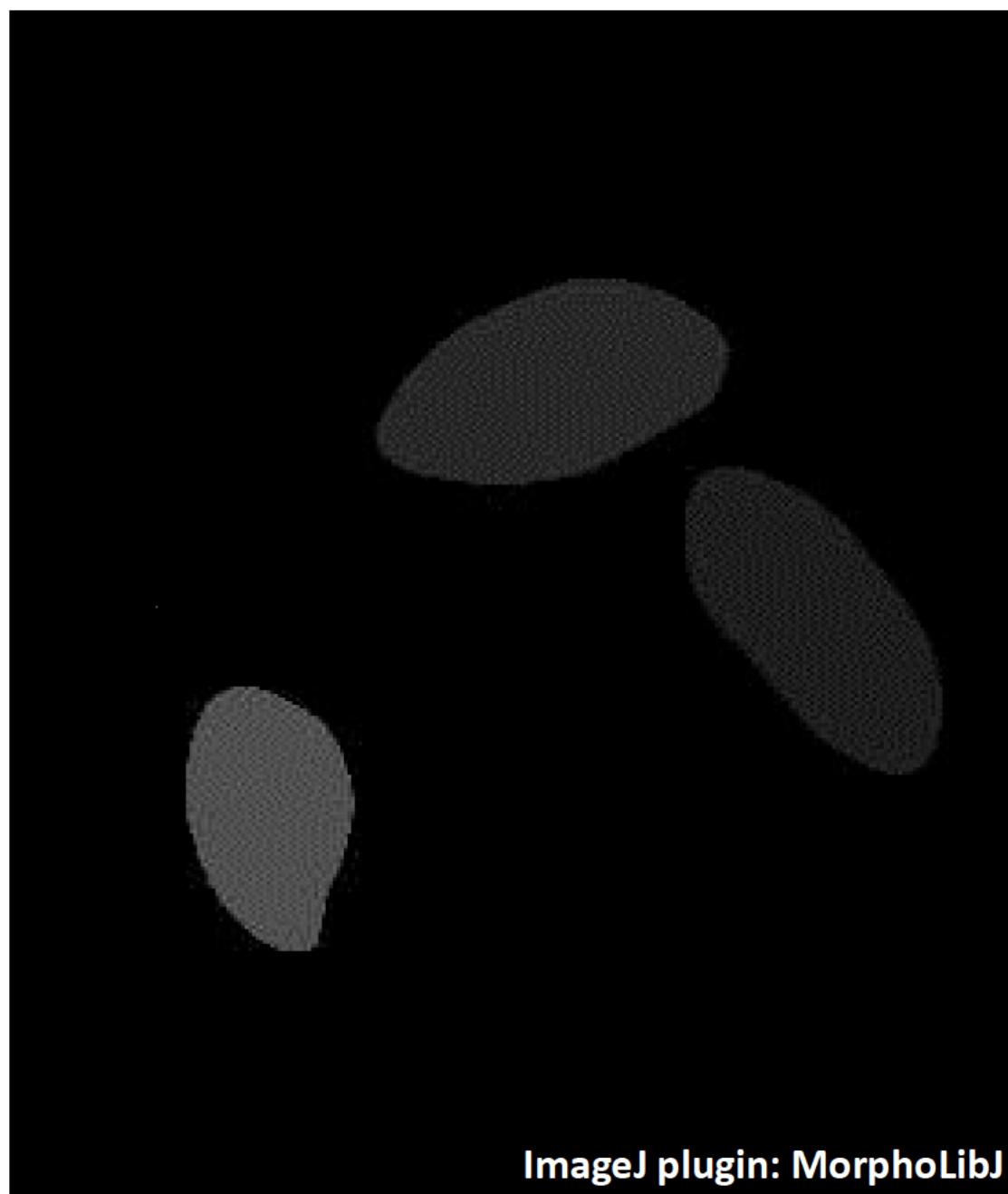
*Processing_macro.ijm

File Edit Language Templates Run Tools Tabs Options

*Processing_macro.ijm

```
1 run("8-bit");
2
3 run("Next Slice [>]");run("Next Slice [>]");
4 run("Next Slice [>]");run("Next Slice [>]");
5 run("Next Slice [>]");run("Next Slice [>]");
6 run("Next Slice [>]");run("Next Slice [>]");
7 run("Next Slice [>]");run("Next Slice [>]");
8 run("Next Slice [>]");run("Next Slice [>]");
9
10 //run("Brightness/Contrast..."); 
11 setMinAndMax(0, 64000);
12
13
14 run("Convert to Mask",
15 "method=Li background=Dark");
16
17 run("Invert", "stack");
18 run("Erode", "stack");
19 run("Dilate", "stack");
20 run("Dilate", "stack");
21 run("Fill Holes", "stack");
22 run("Erode", "stack");
23 run("Invert", "stack");
24
25
26 run("Connected Components Labeling",
27 "connectivity=26 type=[8 bits]");
28
29 run("Analyze Regions 3D",
30 "volume surface_area mean_breadth sphericity");
31
32
33 saveAs("Results");
34 saveAs("Tiff");
35 run("Close All");
```

Batch Kill persistent Show Errors Clear



ImageJ plugin: MorphoLibJ

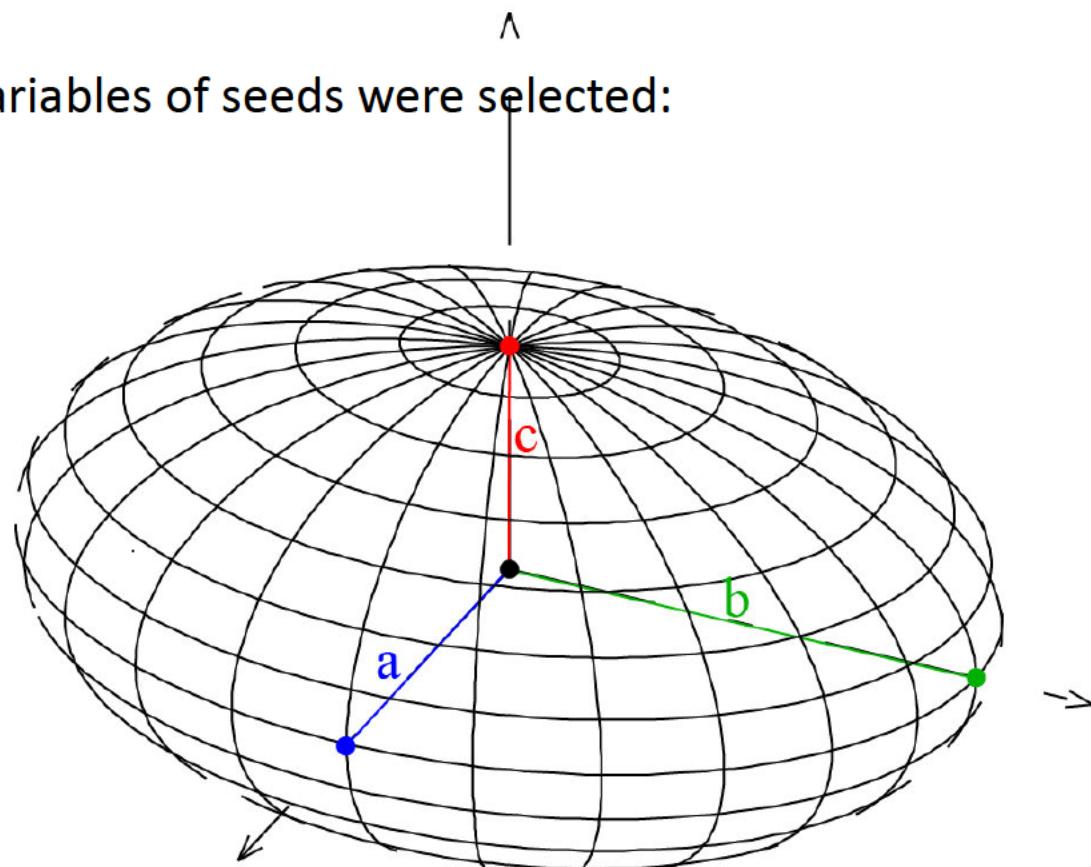
Variables & Sampling

For this project, five morphological variables of seeds were selected:

- 1) Seed volume
- 2) Seed surface
- 3) Fitted ellipsoid radius 1 (**a**)
- 4) Fitted ellipsoid radius 2 (**b**)
- 5) Fitted ellipsoid radius 3 (**c**)

$$\mathbf{a} > \mathbf{b} > \mathbf{c}$$

| Genus | domesticate | wild | wild_relative |
|----------|-------------|------|---------------|
| Lathyrus | 22 | 349 | 42 |
| Lens | 19 | 71 | 97 |
| Pisum | 29 | 0 | 0 |
| Vicia | 35 | 442 | 15 |



**MorphoLibJ can, among other functions
model the best fitting ellipsoid
on the seeds**

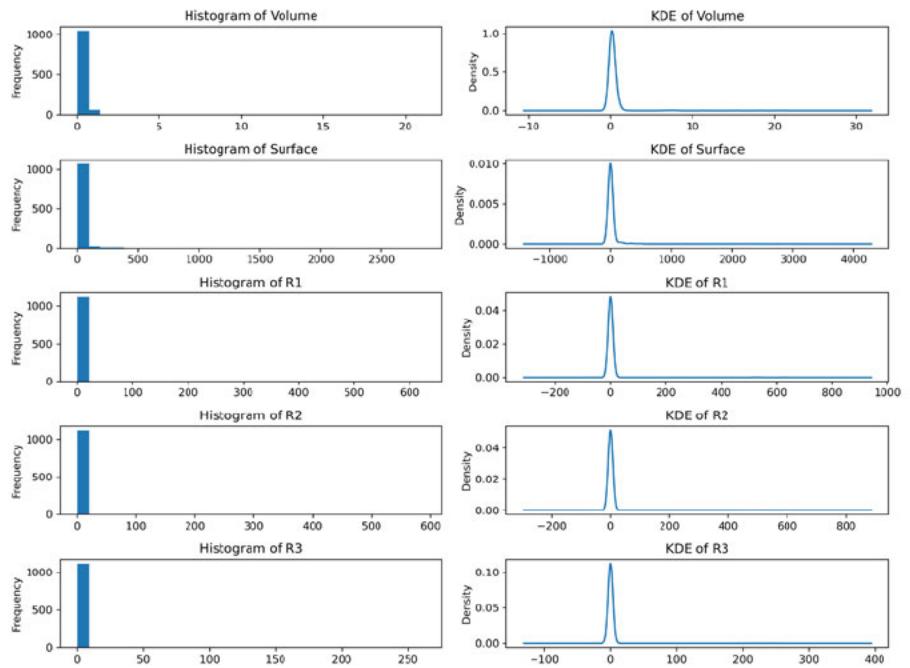
ImageJ plugin: MorphoLibJ

Exploratory & Descriptive Statistics

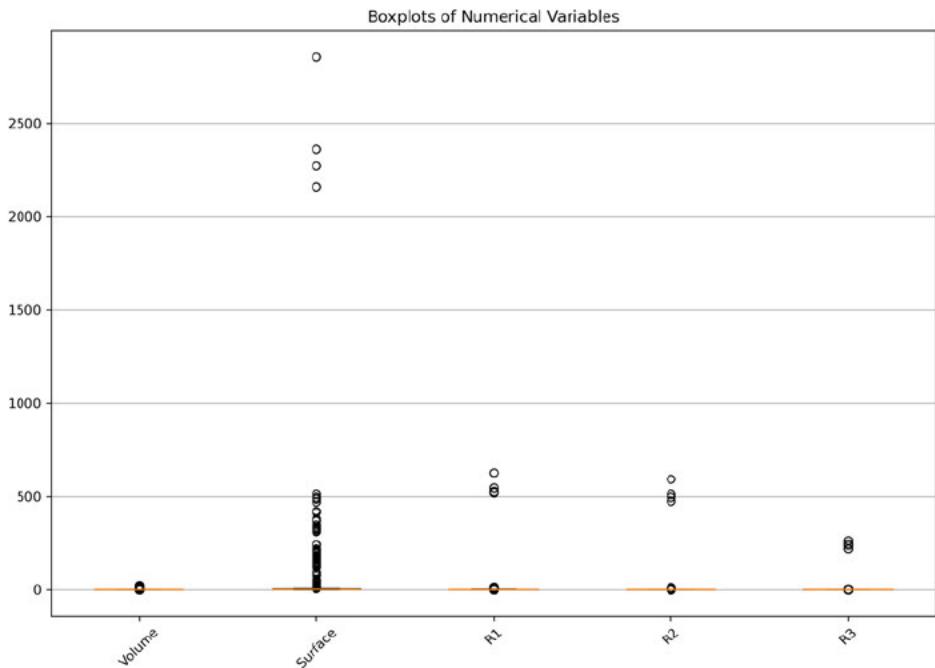
summary_statistics

| | count | mean | std | min | 25 % | 50 % | 75 % | max | skewness | kurtosis |
|---------|--------|---------------------|--------------------|-------------|-------------|-------------|-------------|-------------|--------------------|--------------------|
| Volume | 1121.0 | 0.36717648145115966 | 1.305946527761704 | 1.82e-06 | 0.081070254 | 0.167251036 | 0.357289921 | 21.18948939 | 11.885008134613416 | 159.4513013098811 |
| Surface | 1121.0 | 21.043653425181983 | 153.5104821830469 | 0.000722492 | 1.111470709 | 1.709605634 | 3.07659192 | 2854.976568 | 14.254893636870085 | 225.28185626744795 |
| R1 | 1121.0 | 2.8404765330918824 | 33.159058808582174 | 0.01071862 | 0.36737438 | 0.451246266 | 0.599969944 | 627.677985 | 16.761042843758823 | 282.21865153978246 |
| R2 | 1121.0 | 2.5261981063630685 | 31.047535151458945 | 0.01071862 | 0.327794306 | 0.391379842 | 0.51857484 | 590.701925 | 16.813085565099282 | 283.8840101069201 |
| R3 | 1121.0 | 1.1040347338929528 | 14.378174152964226 | 0.0 | 0.1474451 | 0.241722524 | 0.326241456 | 262.917655 | 16.786339761365596 | 281.586054749796 |

Histograms and KDE of Numerical Variables

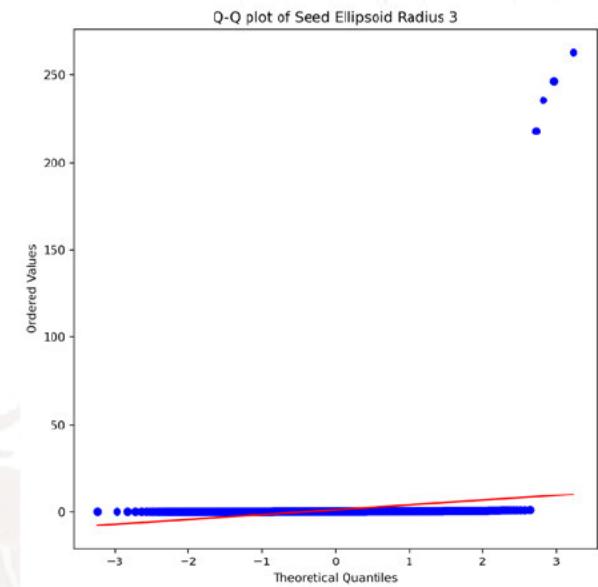
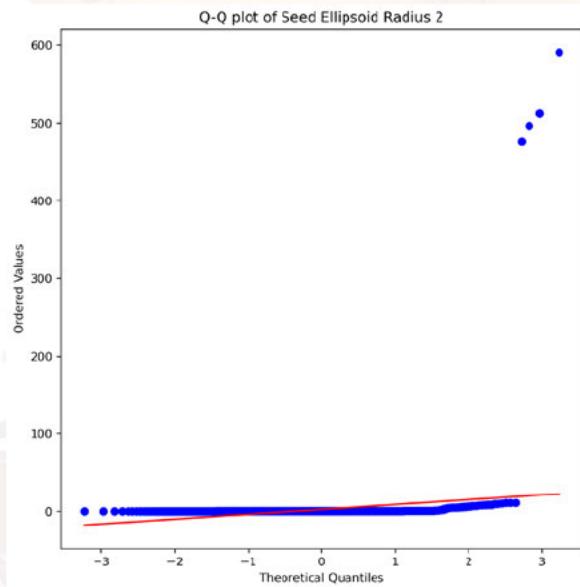
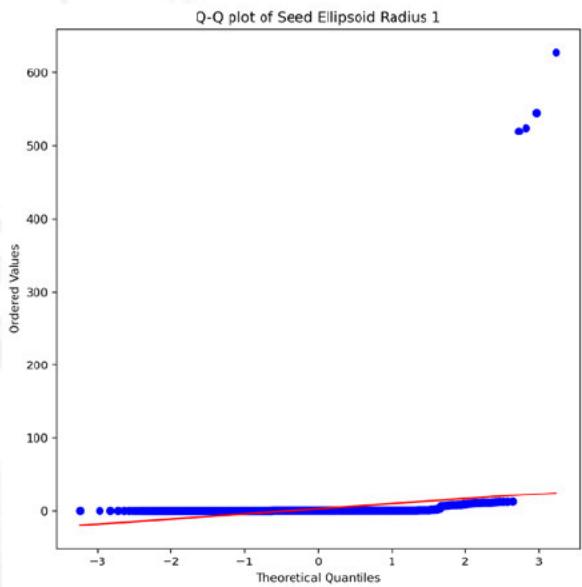
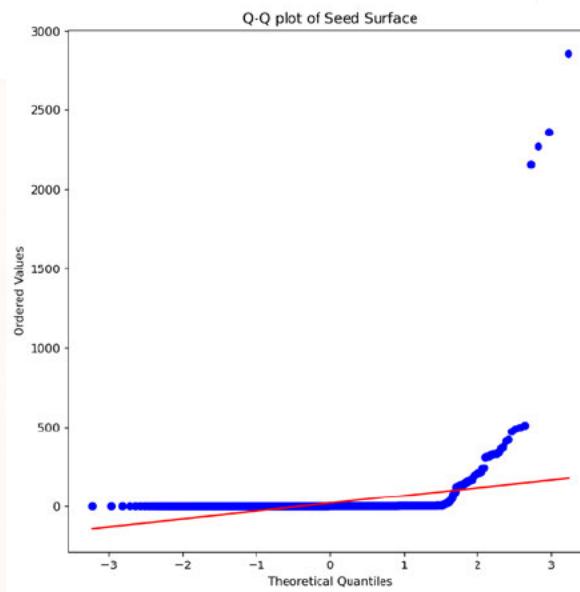
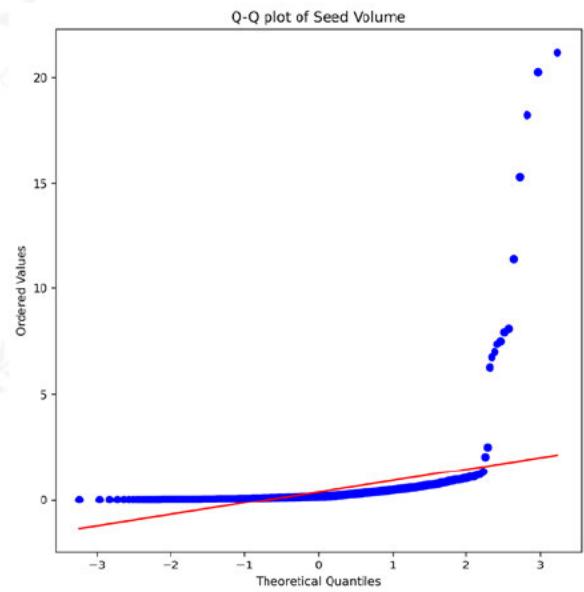


Boxplots of Numerical Variables

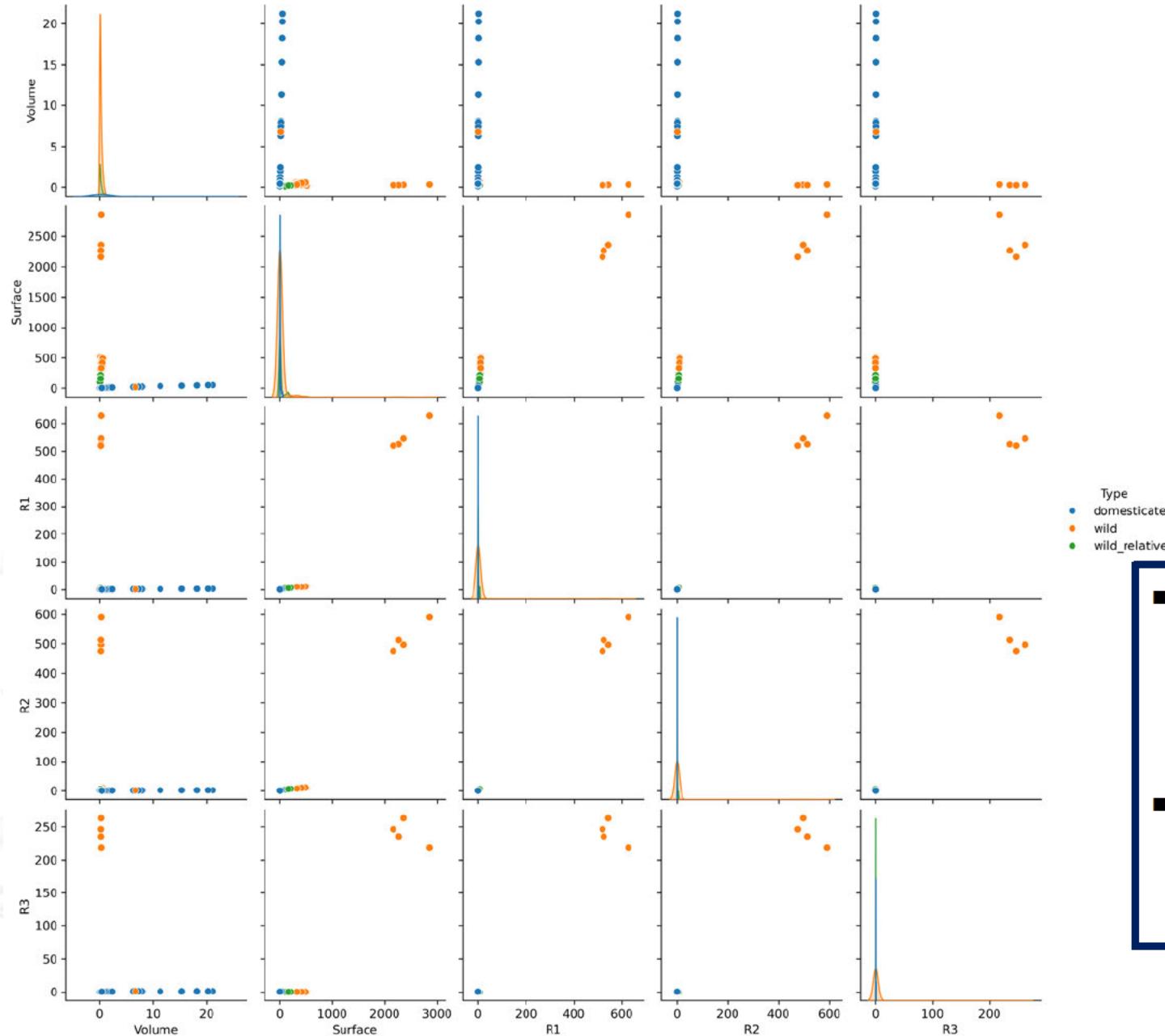


Exploratory & Descriptive Statistics

Q-Q Plot



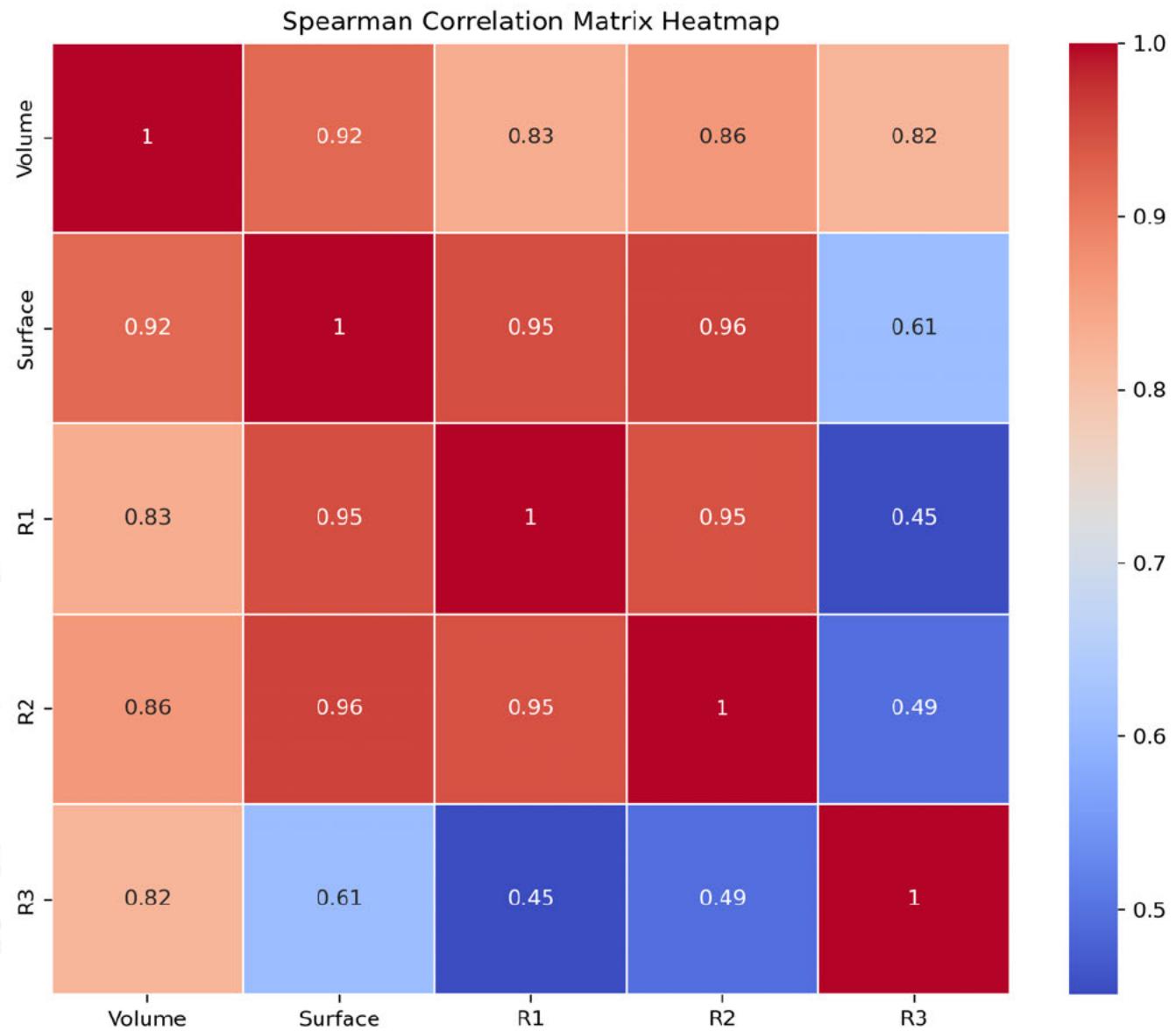
Exploratory & Descriptive Statistics



Pair Plot

- The data contains outliers among wild species
- Volume seems to contain most of the variation

Correlations of Variables

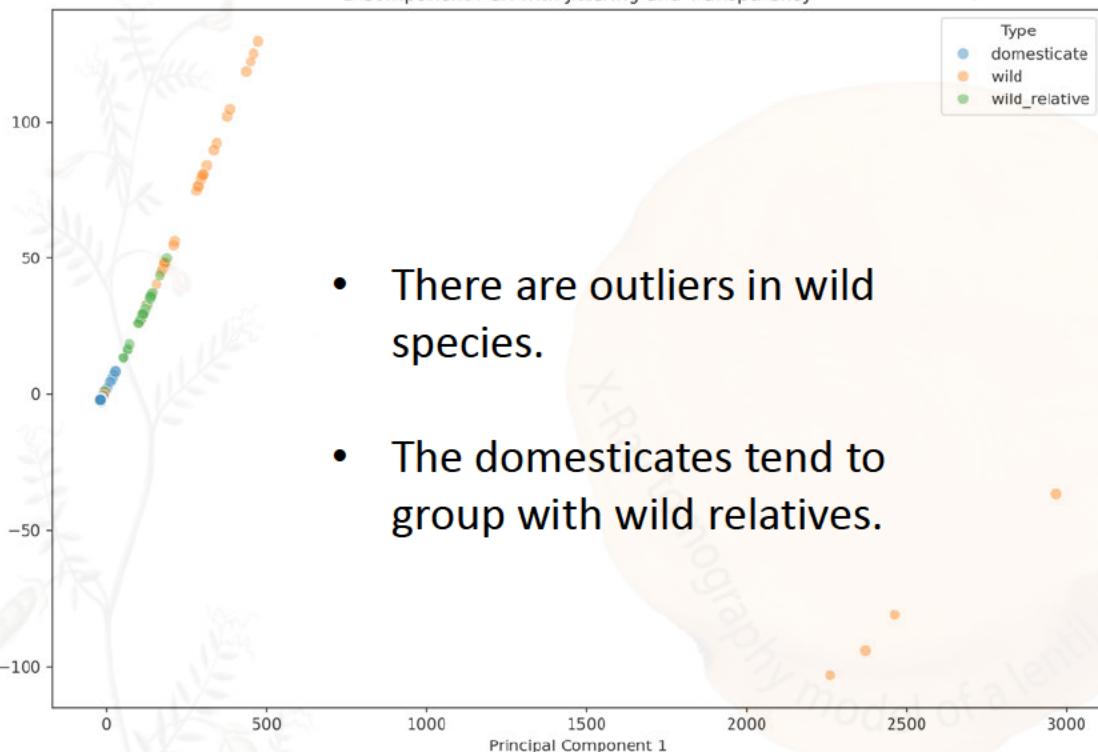


Most variables are strongly correlated with each other, except **R3** which quantifies the thickness of the seed.

Dimensionality Reduction

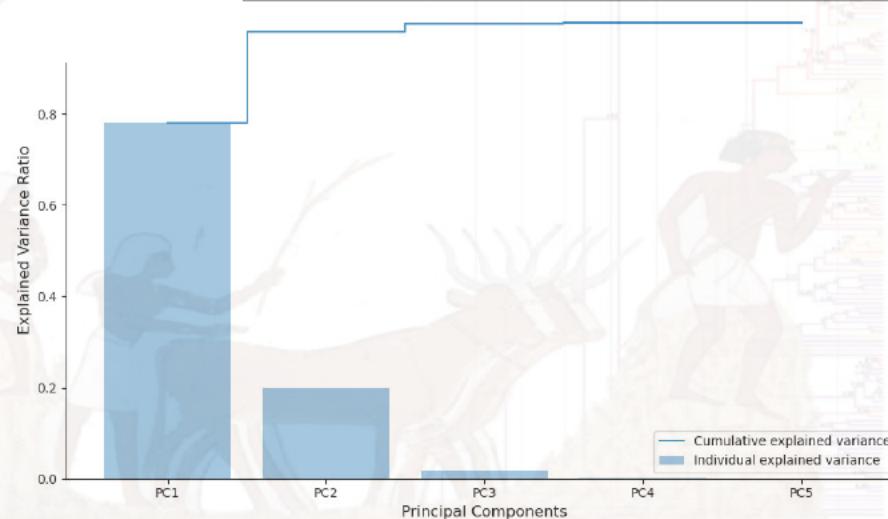
2 Component PCA with Jittering and Transparency

- There are outliers in wild species.
- The domesticates tend to group with wild relatives.



| | Volume | Surface | R1 | R2 | R3 |
|-----|-----------|-----------|-----------|-----------|-----------|
| PC1 | 0.003078 | 0.490235 | 0.505095 | 0.504593 | 0.499925 |
| PC2 | 0.999901 | 0.010402 | -0.004430 | -0.006337 | -0.005485 |
| PC3 | -0.013313 | 0.845547 | -0.164571 | -0.190315 | -0.470712 |
| PC4 | 0.003269 | -0.208335 | 0.395399 | 0.524535 | -0.724644 |
| PC5 | 0.000812 | 0.034755 | -0.749290 | 0.658779 | 0.058022 |

- The loading on PC1 are the surface, R1, R2, and R3 (ca. 80% variation).
- The loading on PC2 is the volume (ca. 20% of variation).



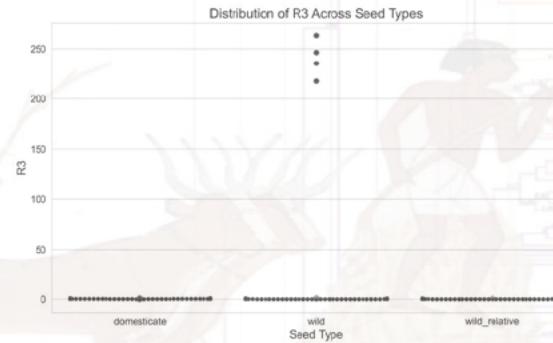
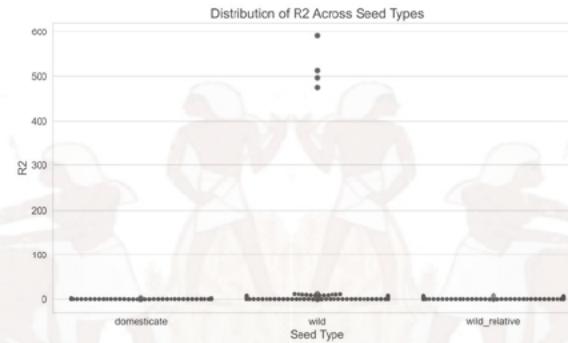
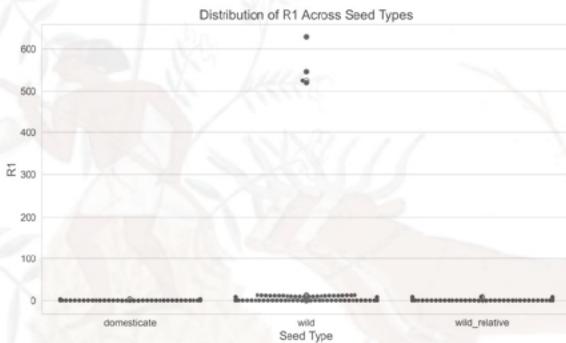
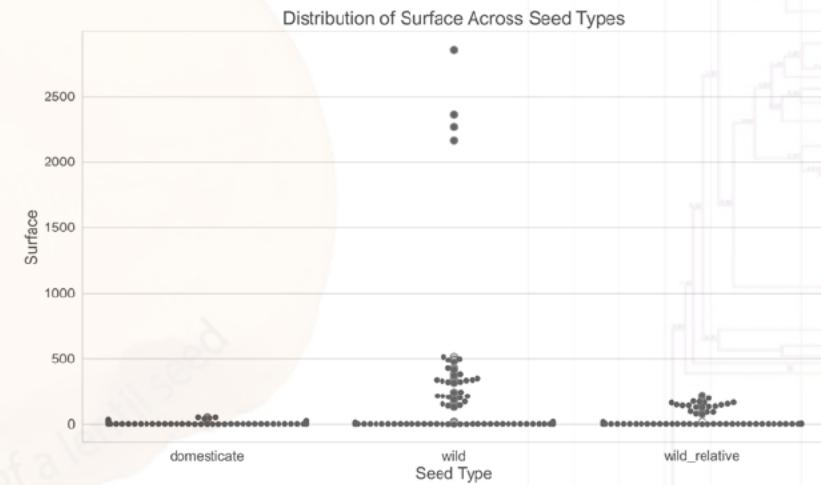
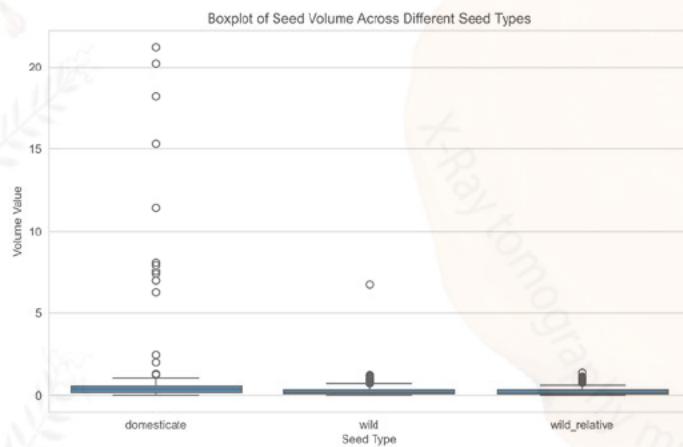
Hypotheses Testing

To investigate the hypotheses considering the non-normal distribution of the data, non-parametric statistical tests were carried out:

- 1) Kruskal-Wallis H-test which is used to compare three or more independent samples.
- 2) Mann-Whitney U test which is used for determining whether two independent samples were drawn from a population with the same distribution.

Hypotheses Testing

Q1: Do the morphological variables differ significantly among the three seed types (**domesticate**, **wild relative**, and **wild** species)?



Hypotheses Testing

Q1: Do the morphological variables differ significantly among the three seed types (**domesticate**, **wild relative**, and **wild** species)?

- H_0 : There is no difference in [morphological variable] across different seed types.
- H_a : There is a difference in [morphological variable] across different seed types.

kruskal_results

| | Statistic | p-value |
|----------------|--------------------|------------------------|
| Volume | 35.609249371729085 | 1.851610976155955e-08 |
| Surface | 25.628994166410322 | 2.721038134967955e-06 |
| R1 | 23.63835721240094 | 7.36200233233331e-06 |
| R2 | 37.37842816608115 | 7.644999346542194e-09 |
| R3 | 44.723763318846565 | 1.9424873101891612e-10 |

Our analyses reject the null hypothesis and accept that there are differences across groups for all variables.

Given that $p < 0.05$, I proceed to post-hoc analysis to know which groups significantly differ from each other.

Hypotheses Testing

Q1: Do the morphological variables differ significantly among the three seed types (**domesticate**, **wild relative**, and **wild** species)?

posthoc_results_Volume

| | domesticate | wild | wild_relative |
|----------------------|------------------------|------------------------|-----------------------|
| domesticate | 1.0 | 1.5972274441724896e-07 | 5.048144349941972e-08 |
| wild | 1.5972274441724896e-07 | 1.0 | 0.24864524937954008 |
| wild_relative | 5.048144349941972e-08 | 0.24864524937954008 | 1.0 |

posthoc_results_Surface

| | domesticate | wild | wild_relative |
|----------------------|-----------------------|-----------------------|----------------------|
| domesticate | 1.0 | 2.320328254633047e-06 | 0.013322566436874968 |
| wild | 2.320328254633047e-06 | 1.0 | 0.2545792590074899 |
| wild_relative | 0.013322566436874968 | 0.2545792590074899 | 1.0 |

Summary:

- The "domesticate" seed type is significantly different from both the "wild" and "wild relative" seed types, except for R1 & R2 of wild relatives.
- No significant difference is observed between the "wild" and "wild relative" seed types except for R2 & R3.

posthoc_results_R1

| | domesticate | wild | wild_relative |
|----------------------|-----------------------|-----------------------|----------------------|
| domesticate | 1.0 | 1.144130111716078e-05 | 0.05960613620866917 |
| wild | 1.144130111716078e-05 | 1.0 | 0.1095706071396374 |
| wild_relative | 0.05960613620866917 | 0.1095706071396374 | 1.0 |

posthoc_results_R2

| | domesticate | wild | wild_relative |
|----------------------|------------------------|------------------------|-----------------------|
| domesticate | 1.0 | 2.6504097181519644e-07 | 0.18661154147056103 |
| wild | 2.6504097181519644e-07 | 1.0 | 0.0008760299449781818 |
| wild_relative | 0.18661154147056103 | 0.0008760299449781818 | 1.0 |

posthoc_results_R3

| | domesticate | wild | wild_relative |
|----------------------|------------------------|------------------------|------------------------|
| domesticate | 1.0 | 0.00010592241814652835 | 1.1233064148379653e-10 |
| wild | 0.00010592241814652835 | 1.0 | 8.565281770459626e-06 |
| wild_relative | 1.1233064148379653e-10 | 8.565281770459626e-06 | 1.0 |

Hypotheses Testing

Q2-1: Are the distribution of seed variables the same between **domesticate** and **wild relative** Types?

Q2-2: Are the distribution of seed variables the same between **wild relative** and **wild** Types?

Hypotheses Testing

Q2-1: Are the distribution of seed variables the same between **domesticate** and **wild relative** Types?

mannwhitney_results_domesticate_vs_wi

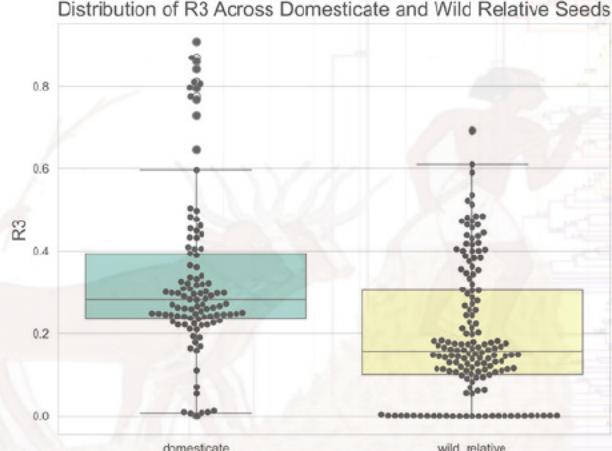
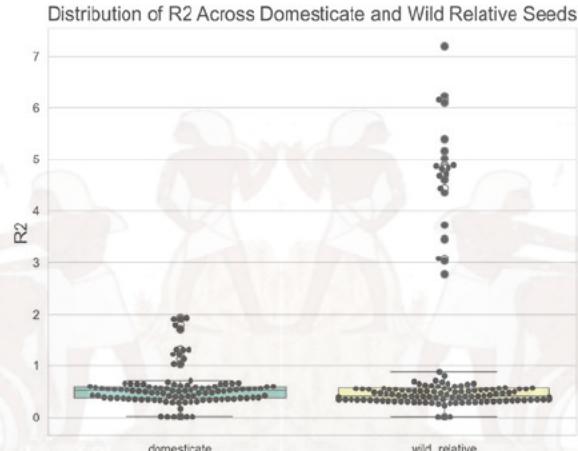
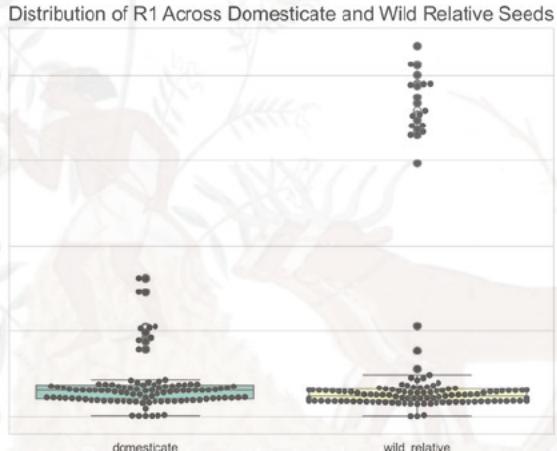
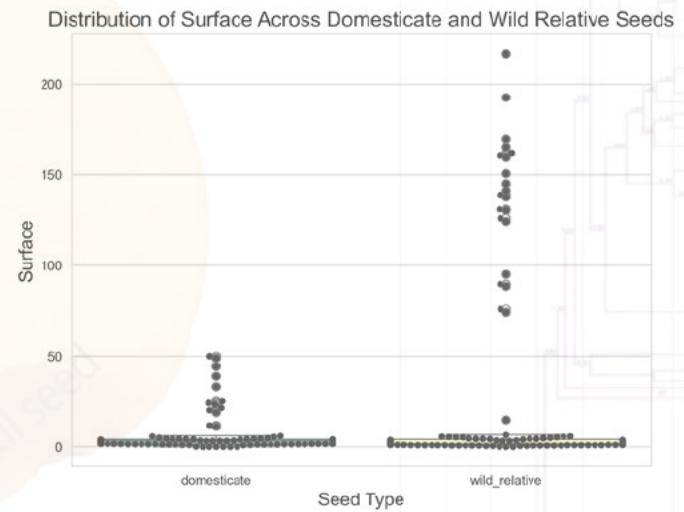
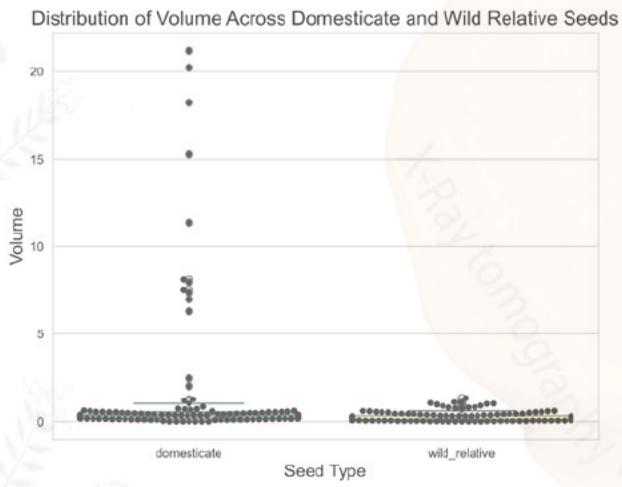
| Variable | Statistic | p_value |
|----------------|-----------|------------------------|
| Volume | 11343.0 | 3.7247067746726724e-08 |
| Surface | 9509.0 | 0.016174553896296563 |
| R1 | 9236.0 | 0.05192801929564013 |
| R2 | 9062.0 | 0.09899135895590888 |
| R3 | 11689.0 | 1.143251527473281e-09 |

Summary:

- **Volume** and **R3** have very small p-values, suggesting strong evidence of differing distributions between the two seed types for these variables.
- **Surface** also shows a significant difference but with a larger (yet still below 0.05) p-value, suggesting moderate evidence of differing distributions.
- **R1** and **R2** do not show statistically significant differences in the distributions between the two seed types.

Hypotheses Testing

Q2-1: Are the distribution of seed variables the same between **domesticate** and **wild relative** Types?



Hypotheses Testing

Q2-2: Are the distribution of seed variables the same between **wild relative** and **wild** Types?

mannwhitney_results_wild_vs_wild_relati

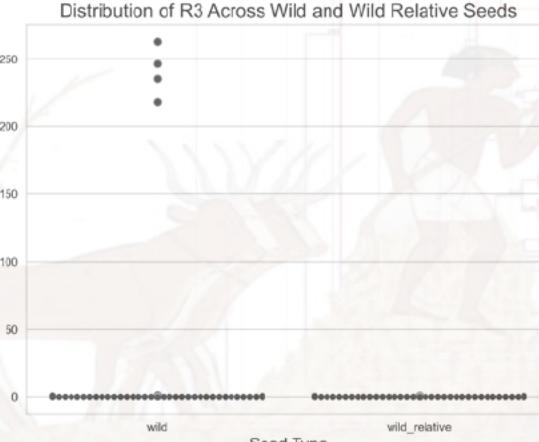
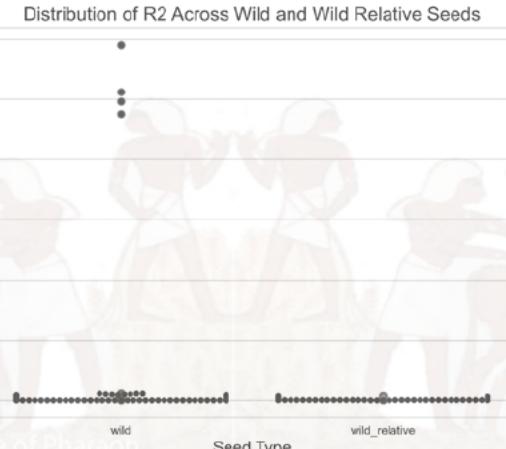
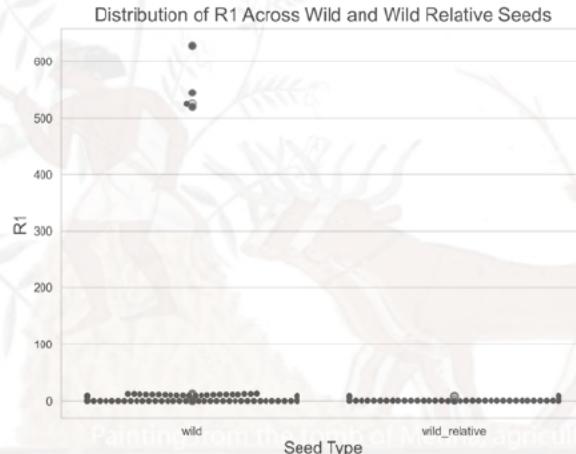
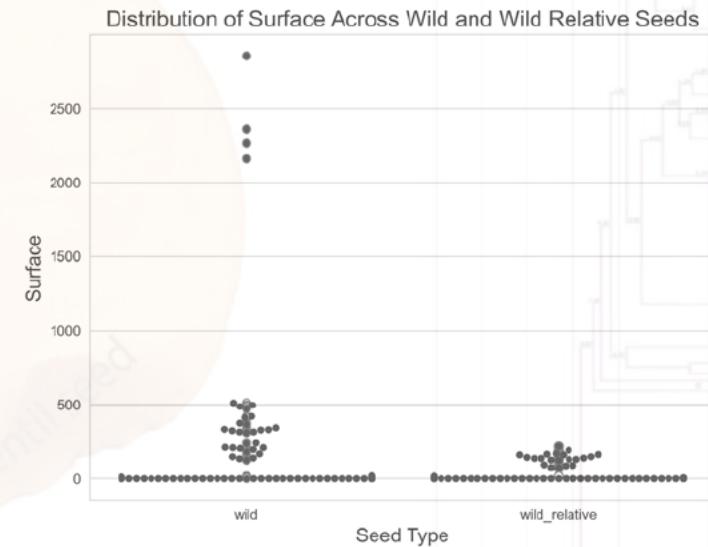
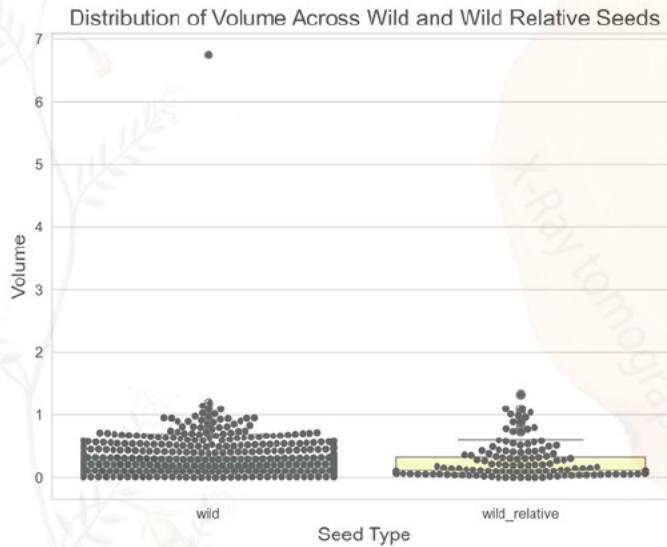
| Variable | Statistic | p_value |
|----------|-----------|------------------------|
| Volume | 72267.0 | 0.07895644081806023 |
| Surface | 60852.0 | 0.0997298965376448 |
| R1 | 59586.0 | 0.04301068865786264 |
| R2 | 54350.0 | 0.00033752726385039055 |
| R3 | 82379.0 | 1.8282534430370777e-06 |

Summary:

- Variables **R1**, **R2**, and **R3** demonstrate statistically significant differences between the two groups being compared.
- Variables **Volume** and **Surface** do not show statistically significant differences based on the conventional 0.05 threshold.

Hypotheses Testing

Q2-2: Are the distribution of seed variables the same between wild relative and wild Types?



Hypotheses Testing

a **logistic regression model** was carried out to understand if some traits made some species suitable to be the target of domestication (domesticate Type were removed from the data).

Can Morphological Variables Predict domestication?

Logit Regression Results

| Dep. Variable: | IsWildRelative | No. Observations: | | | |
|------------------|------------------|-------------------|---------|-------|--------|
| 1016 | | | | | |
| Model: | Logit | Df Residuals: | | | |
| 1010 | | | | | |
| Method: | MLE | Df Model: | | | |
| 5 | | | | | |
| Date: | Sun, 08 Oct 2023 | Pseudo R-squ.: | | | |
| 0.09378 | | | | | |
| Time: | 14:32:48 | Log-Likelihood: | | | |
| -391.70 | | | | | |
| converged: | True | LL-Null: | | | |
| -432.24 | | | | | |
| Covariance Type: | nonrobust | LLR p-value: | | | |
| 5.017e-16 | | | | | |
| | coef | std err | z | P> z | [0.025 |
| 0.975] | | | | | |
| const | -1.5713 | 0.126 | -12.495 | 0.000 | -1.818 |
| -1.325 | | | | | |
| Volume | 1.3184 | 0.465 | 2.837 | 0.005 | 0.407 |
| 2.229 | | | | | |
| Surface | -0.0834 | 0.016 | -5.363 | 0.000 | -0.114 |
| -0.053 | | | | | |
| R1 | 0.2983 | 0.257 | 1.159 | 0.246 | -0.206 |
| 0.803 | | | | | |
| R2 | 2.4684 | 0.547 | 4.513 | 0.000 | 1.396 |
| 3.540 | | | | | |
| R3 | -6.4733 | 1.062 | -6.094 | 0.000 | -8.555 |
| -4.391 | | | | | |

Summary:

- **Model Significance:** The LLR p-value is quite low, suggesting the model as a whole is significant.
- **Effect of Variables:**
 - ✓ **Volume, Surface, R2, and R3** are statistically significant predictors of IsWildRelative.
- These results are preliminary and model needs more validation etc.

General Conclusion

- ❑ The difference in volume between domesticates and wild species and wild relatives is the expected outcome of selective breeding for yield.
- ❑ The fact that the significant differences between wild relatives and other wild species involve ellipsoid radii suggests a role in seed shape, especially thickness (R3).
- ❑ Taken together, our results suggest that stone age humans had a preference for some types of seeds, but that these preferences were not simply limited to volume, but also included other factors such as seed shape, hinting at the role of other factors such as seed handling and storage.

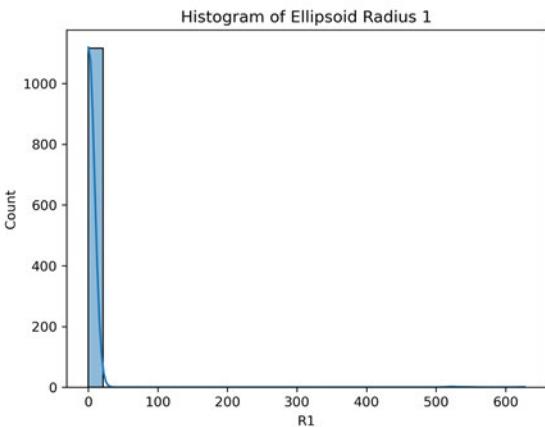
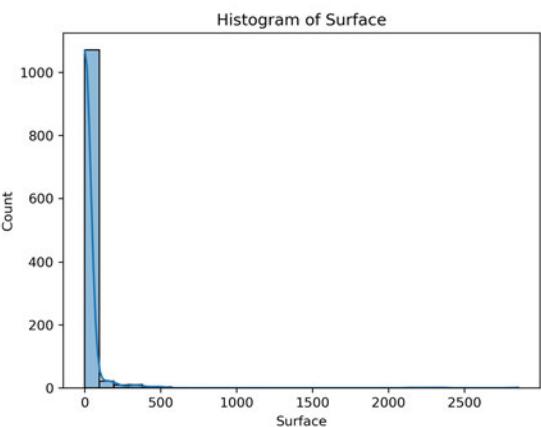
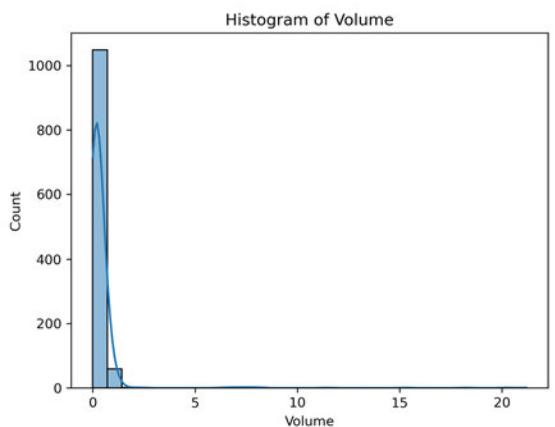
Thank you very much for your attention!



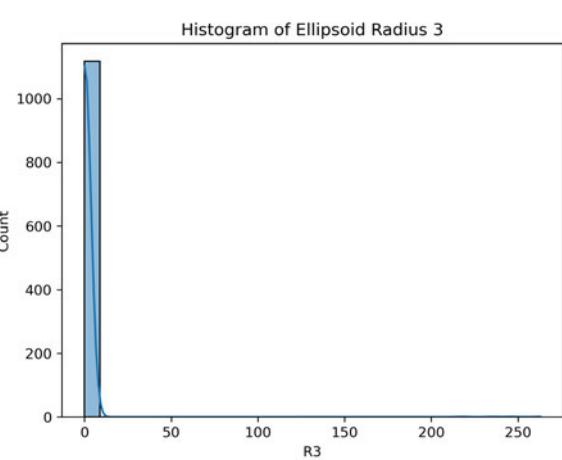
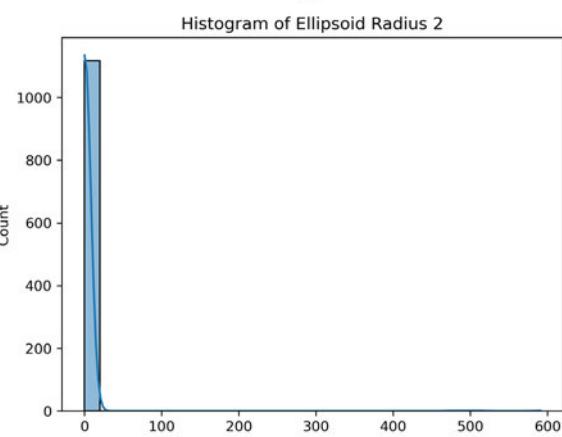
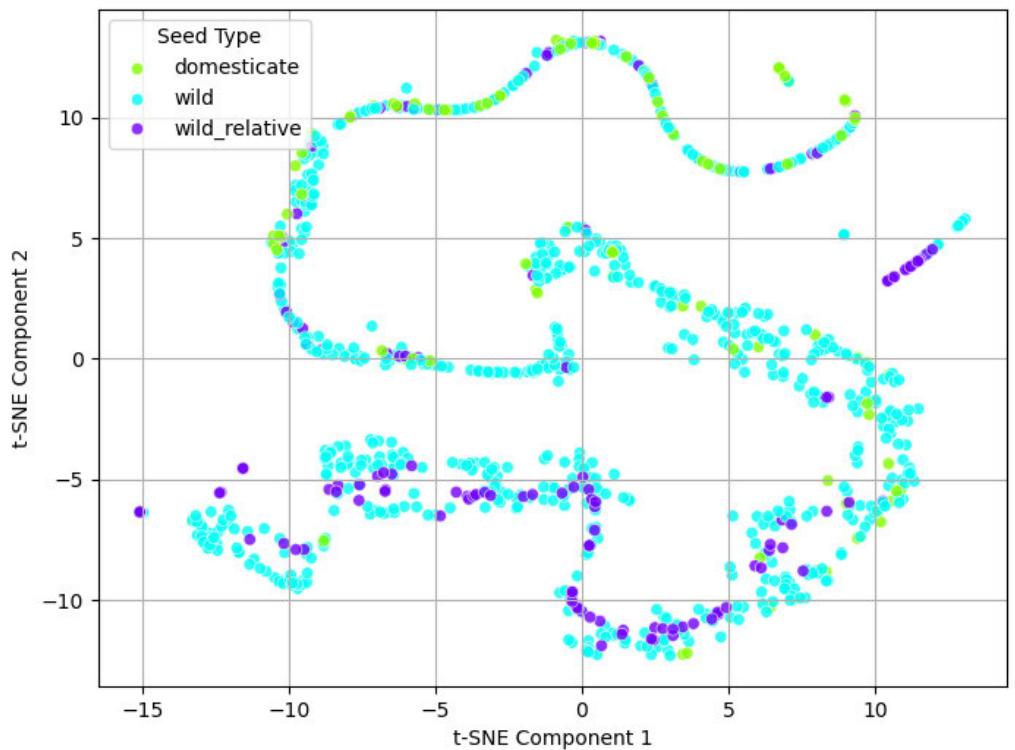
Painting from the tomb of Menna, agricultural scribe of Pharaoh

Various libraries were used for this project:

```
import pandas as pd  
import numpy as np  
import seaborn as sns  
import matplotlib.pyplot as plt  
from sklearn.decomposition import PCA  
from sklearn.manifold import TSNE  
from sklearn.preprocessing import StandardScaler  
from scipy.stats import Kruskal  
from scikit_posthocs import posthoc_dunn  
import scipy.stats as stats  
import scikit_posthocs as sp  
import statsmodels.api as sm
```



t-SNE 2D Visualization of Seed Data



Hypotheses Testing

To conduct a hypothesis test with the goal of determining which variables predict domestication in our seed dataset, a **logistic regression model** was used. This approach is widely used for binary outcome variables.

Can Morphological Variables Predict domestication?

| Logit Regression Results | | | | | |
|--------------------------|---------|------------------|---------|-------------------|--------|
| | | | | | |
| Dep. Variable: | | IsDomesticate | | No. Observations: | |
| 1121 | | | | | |
| Model: | | Logit | | Df Residuals: | |
| 1115 | | | | | |
| Method: | | MLE | | Df Model: | |
| 5 | | | | | |
| Date: | | Sun, 08 Oct 2023 | | Pseudo R-squ.: | |
| 0.1028 | | | | | |
| Time: | | 14:36:20 | | Log-Likelihood: | |
| -312.72 | | | | | |
| converged: | | True | | LL-Null: | |
| -348.56 | | | | | |
| Covariance Type: | | nonrobust | | LLR p-value: | |
| 4.548e-14 | | | | | |
| <hr/> | | | | | |
| 0.975] | coef | std err | z | P> z | [0.025 |
| const | -2.5001 | 0.169 | -14.788 | 0.000 | -2.831 |
| -2.169 | | | | | |
| Volume | 1.2369 | 0.413 | 2.998 | 0.003 | 0.428 |
| 2.045 | | | | | |
| Surface | -0.0502 | 0.042 | -1.187 | 0.235 | -0.133 |
| 0.033 | | | | | |
| R1 | -0.1162 | 1.145 | -0.101 | 0.919 | -2.361 |
| 2.128 | | | | | |
| R2 | 0.9125 | 1.363 | 0.669 | 0.503 | -1.759 |
| 3.584 | | | | | |
| R3 | -1.4980 | 1.300 | -1.152 | 0.249 | -4.046 |
| 1.050 | | | | | |

Summary:

- **Model Significance:** The LLR p-value is quite low, suggesting the model as a whole is significant.

Effect of Variables:

- ✓ **Volume** is the only predictor that provides a significant contribution to explaining the variance in IsDomesticate.
- ✓ **Surface, R1, R2, and R3** do not appear to significantly contribute to the model.