Mini Project 1

Jackson Passey, Gavin Hatch, Ty Hawkes

2025-10-20

Introduction

This analysis attempts to answer questions about fatty acids in olives found in two different regions. The data includes response variables palmitic, palmitoleic, stearic, oleic, linoleic, eicosanoic, linolenic, and eicosenoic acids predicted by region (Region 2 and Region 4).

This report assesses whether the olive samples from each region deviate from their historical averages and whether the two regions deviate from each other. We also explore whether linoleic and linolenic acids can be dropped from the list of fatty acids without significantly decreasing the separation of the two samples.

The data analyzed contains several missing values. Part of this report determines whether the data is missing at random and provides a new dataset with imputed values using the Expectation-Maximization (EM) algorithm and multiple imputation.

We will show that... (put conclusions here)

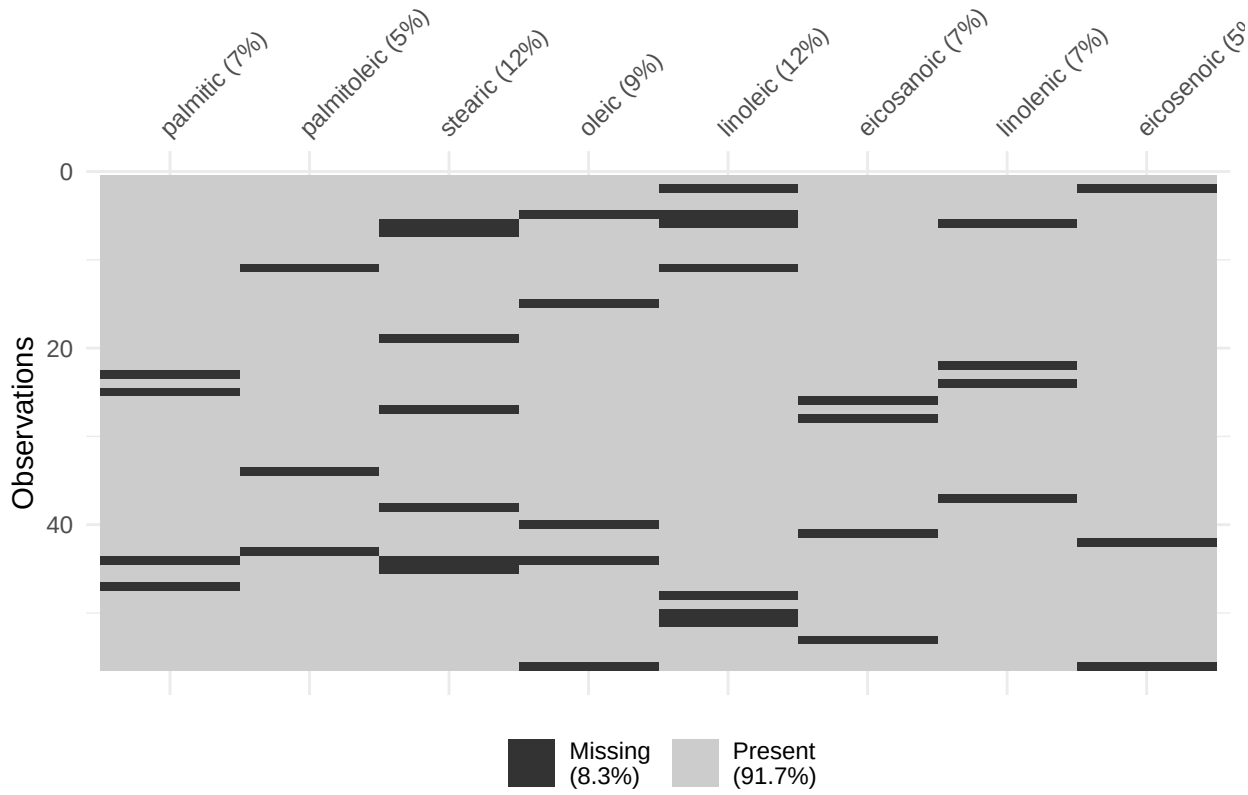
Data Preparation

Due to missing values in the data, we use the EM algorithm combined with multiple imputation to fill in missing values, assuming normality and that the data is missing at random (MAR). We explore whether the data is MAR in the next section.

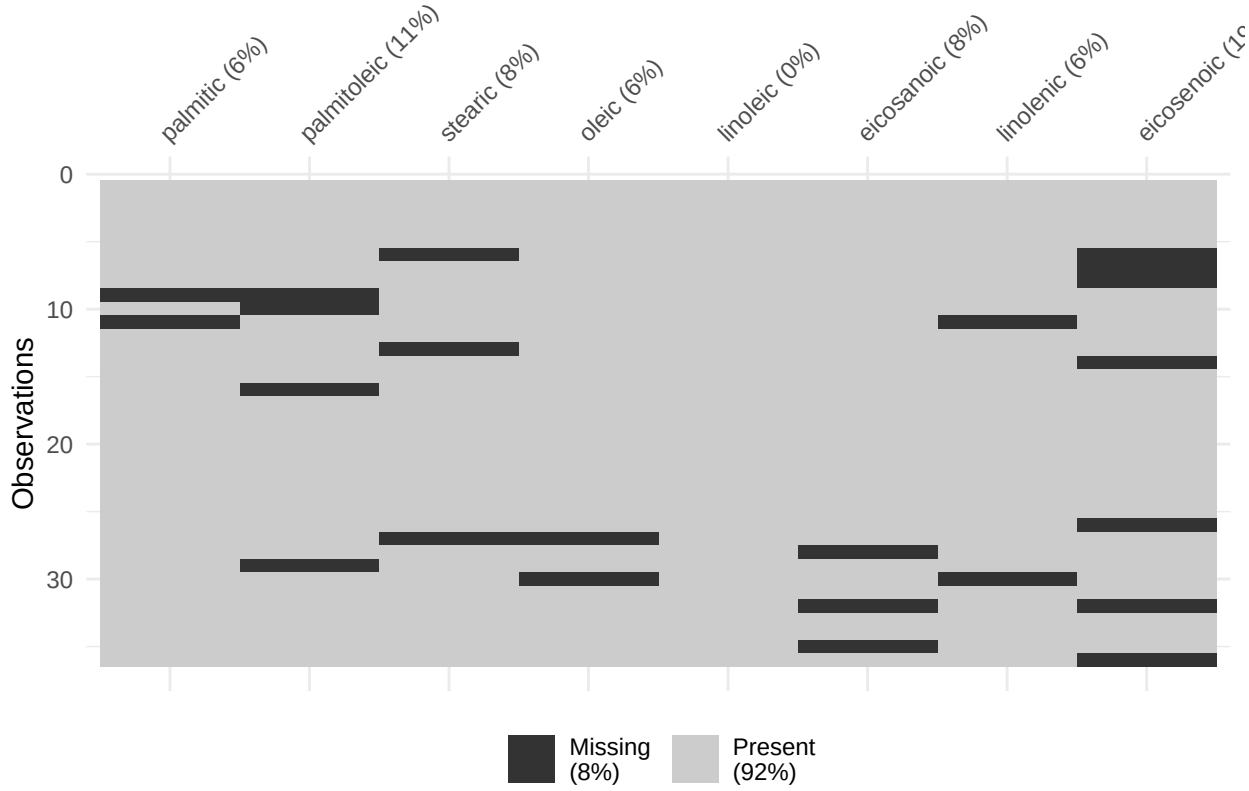
The problem with a simple imputed data set is that it does not take into account random error. To solve this, we created a distribution of values for each missing cell and selected from that distribution at random. (add Jackson's graph with the red line showing this?)

Assessment of Missing at Random (MAR) Assumption

To assess whether the data is missing at random, we used Little's MCAR (Missing Completely At Random) Test. (It looks like that's what Jackson did; Ty did you do something else?)




```
##             palmitic palmitoleic stearic  oleic linoleic eicosanoic linolenic
## palmitic          NA      1.0000  0.4231 0.3196   1.0000          1    1.0000
## palmitoleic   1.0000          NA  1.0000 1.0000   0.3354          1    1.0000
## stearic       0.4231      1.0000      NA 0.5008   1.0000          1    0.4231
## oleic         0.3196      1.0000  0.5008     NA   0.5008          1    1.0000
## linoleic      1.0000      0.3354  1.0000 0.5008       NA          1    0.4231
## eicosanoic    1.0000      1.0000  1.0000 1.0000   1.0000         NA    1.0000
## linolenic     1.0000      1.0000  0.4231 1.0000   0.4231          1        NA
## eicosenoic    1.0000      1.0000  1.0000 0.2487   0.3354          1    1.0000
##             eicosenoic
## palmitic        1.0000
## palmitoleic     1.0000
## stearic         1.0000
## oleic           0.2487
## linoleic        0.3354
## eicosanoic      1.0000
## linolenic       1.0000
## eicosenoic          NA
```




```
##             palmitic palmitoleic stearic  oleic eicosanoic linolenic eicosenoic
## palmitic          NA      0.2127  1.0000 1.0000     1.0000    0.1095     1.0000
## palmitoleic   0.2127          NA  1.0000 1.0000     1.0000    1.0000     0.5658
## stearic       1.0000      1.0000      NA 0.1619     1.0000    1.0000     0.4882
## oleic         1.0000      1.0000  0.1619     NA     1.0000    0.1095     1.0000
## eicosanoic    1.0000      1.0000  1.0000 1.0000         NA    1.0000     0.4882
## linolenic     0.1095      1.0000  1.0000 0.1095     1.0000        NA     1.0000
## eicosenoic    1.0000      0.5658  0.4882 1.0000     0.4882    1.0000         NA
```

Analysis for Region 2

words

```
## $mean
##    palmitic palmitoleic     stearic       oleic    linoleic  eicosanoic
##  1299.62401   121.09219   264.46005  7311.74026   817.85339    45.68766
##   linolenic  eicosenoic
##    63.55326    28.26358
##
## $total_variance
##                palmitic  palmitoleic      stearic       oleic    linoleic
```

```
## palmitic      77.1611779  3.636514038   0.62693838 -111.353001   28.165001
## palmitoleic    3.6365140  8.181870042  -4.34185041   -1.961598   -5.515118
## stearic        0.6269384 -4.341850411  23.87359866  -21.244717    1.716108
## oleic       -111.3530007 -1.961598321 -21.24471659  317.543997 -169.350768
## linoleic      28.1650010 -5.515117910   1.71610813 -169.350768  140.573371
## eicosanoic    -0.3036901 -0.622914296   0.42047044   -3.140478    2.545714
## linolenic      1.7397218 -0.003074716  -0.80686608   -5.522402    2.811557
## eicosenoic    -1.0475553 -0.147657826   0.04814726   -2.060295    2.116408
##             eicosanoic    linolenic  eicosenoic
## palmitic    -0.3036901  1.739721801 -1.04755534
## palmitoleic -0.6229143 -0.003074716 -0.14765783
## stearic      0.4204704 -0.806866075  0.04814726
## oleic       -3.1404776 -5.522401651 -2.06029489
## linoleic     2.5457145  2.811557155  2.11640804
## eicosanoic   0.8549886  0.273614050  0.21798168
## linolenic    0.2736140  0.908534000  0.29848845
## eicosenoic   0.2179817  0.298488451  0.81557543
##
## $T2
## [1] 352.5722
##
## $F_stat
## [1] 38.46242
##
## $p_value
## [1] 2.975398e-14
```

Analysis for Region 4

words

```
## $mean
##    palmitic palmitoleic     stearic       oleic    linoleic  eicosanoic
##  1229.00485   105.29473   273.31700  7359.15499   834.72222    42.28743
##   linolenic  eicosenoic
##    75.89850    37.76324
##
## $total_variance
##                palmitic palmitoleic      stearic       oleic     linoleic
## palmitic      897.74220  209.816373  -83.2051369 -1526.41935   426.403920
## palmitoleic   209.81637   65.569374  -29.7454279  -385.56689   115.806335
## stearic       -83.20514  -29.745428   64.8806452   164.27609  -111.775525
## oleic       -1526.41935 -385.566891  164.2760923  3181.05148 -1210.624434
## linoleic      426.40392  115.806335 -111.7755246 -1210.62443   694.083510
## eicosanoic    -11.40085   -3.901047    5.2309958    17.32790    -7.648859
## linolenic      35.45592    9.195726   -6.2975653   -71.11265    26.030260
```

```
## eicosenoic    -12.40019   -2.237210    0.3988072    10.70765     1.110763
##              eicosanoic   linolenic  eicosenoic
## palmitic    -11.4008500  35.4559206 -12.4001937
## palmitoleic  -3.9010470   9.1957263  -2.2372099
## stearic       5.2309958  -6.2975653   0.3988072
## oleic        17.3279020 -71.1126536  10.7076525
## linoleic     -7.6488586  26.0302596   1.1107631
## eicosanoic    0.6767097  -0.3921222   0.2263221
## linolenic    -0.3921222   3.4648483   0.4131747
## eicosenoic    0.2263221   0.4131747   2.0505849
##
## $T2
## [1] 530.1991
##
## $F_stat
## [1] 53.01991
##
## $p_value
## [1] 1.487699e-14
```

Comparison between Regions 2 and 4

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   palmitic = col_double(),
##   palmitoleic = col_double(),
##   stearic = col_double(),
##   oleic = col_double(),
##   linoleic = col_double(),
##   eicosanoic = col_double(),
##   linolenic = col_double(),
##   eicosenoic = col_double()
## )
##
##
## -- Column specification --------------------------------------------------------
## cols(
##   palmitic = col_double(),
##   palmitoleic = col_double(),
##   stearic = col_double(),
##   oleic = col_double(),
##   linoleic = col_double(),
##   eicosanoic = col_double(),
##   linolenic = col_double(),
##   eicosenoic = col_double()
## )
```

```
## Hotelling's T^2 Statistic: 189.7534

## F-statistic: 21.87435

## Numerator degrees of freedom: 8

## Denominator degrees of freedom: 83

## P-value: 0
```

To compare whether Regions 2 and 4 are the same, we used Hotelling's 2-Sample T test. We calculated a T^2 value of 189.7534332 which converts to an F-Statistic of 21.8743541. Using the F distribution with numerator degrees of freedom of 8 and a denominator degrees of freedom of 83 results in a P-Value of $<.0001$.

This is statistically significant and provides overwhelming evidence that the overall fatty-acid profiles of regions 2 and 4 differ. The agronomist's belief that region 2 and region 4 olives have evolved to have essentially the same profile in terms of the eight fatty acids is not supported by the data. The observed differences are unlikely to have come from random sampling variation alone.

It is important to note that the two samples were drawn by different organizations with potentially different data collection procedures, chemical analysis tools, and data censoring mechanisms. This combined with missing data that was imputed using MVI exposes limitations of this analysis. We do not know if significant variation comes from the different data collection procedures.

Looking at the discriminant function, we can see that the largest coefficients are for Linolenic, Eicosanoic, and Eicosenoic acids (in that order). This suggests that these fatty acids contribute the most to the difference between regions 2 and 4. Further investigation into these specific fatty acids may provide more insights into the differences in olive profiles between the two regions.

Assessment of Linoleic and Linolenic Acids

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   palmitic = col_double(),
##   palmitoleic = col_double(),
##   stearic = col_double(),
##   oleic = col_double(),
##   linoleic = col_double(),
##   eicosanoic = col_double(),
##   linolenic = col_double(),
##   eicosenoic = col_double()
## )
##
##
## -- Column specification --------------------------------------------------------
```

```
## cols(
##   palmitic = col_double(),
##   palmitoleic = col_double(),
##   stearic = col_double(),
##   oleic = col_double(),
##   linoleic = col_double(),
##   eicosanoic = col_double(),
##   linolenic = col_double(),
##   eicosenoic = col_double()
## )
```

```
## [1] "F-stat for region effect: 28.2676278919721"
```

```
## [1] "df1: 2 df2: 83"
```

```
## [1] "P-value: 4.33867164417734e-10"
```

```
##                    [,1]
## palmitic     0.04397878
## palmitoleic  0.07029822
## stearic      0.10629253
## oleic        0.06156712
## linoleic     0.06583051
## eicosanoic  -0.29744787
## linolenic    0.42605030
## eicosenoic   0.18167029
```

```
##                   [,1]
## palmitic     5.339851
## palmitoleic  2.453354
## stearic      4.319048
## oleic       14.405667
## linoleic     7.832097
## eicosanoic  -1.765838
## linolenic    3.781679
## eicosenoic   1.241544
```

Using a conditional Wilk's Λ test, we assessed whether linoleic acid and linolenic acid were important in contributing to the significant difference observed between regions 2 and 4. We set up the following hypothesis:

CHECK THESE HYPOTHESES

$$H_0 : \boldsymbol{\mu}^{(full)} = \boldsymbol{\mu}^{(reduced)}$$

$$H_a : \boldsymbol{\mu}^{(full)} \neq \boldsymbol{\mu}^{(reduced)}$$

Where the full model includes all 8 fatty acids and the reduced model excludes linoleic and linolenic acids.

With a Wilk's Λ of 0.5948317 converted to an F statistic, we calculated a P-value of $< .05$ and reject the null hypothesis. To support this rejection, we found that the discriminant functions separating regions also indicate some form of separation for linoleic and linolenic acids. In this case, we see that the standardized discriminant scores for each of them have high positive influence indicating that observations with higher values of linoleic and linolenic acids more likely reside in Region 2.

We conclude that linoleic and linoeic acids are important in contributing to the separation of the two regions beyond the information available from the other 6 acids.

Conclusions and Recommendations

words

Appendix

words