Project 2 15th June 2019

Virginia Howarth

June 15, 2019

Are Olives the key to the long Life Expectancy in Northern Italy?

Introduction

The Mediterranean Diet is widely acclaimed for the longevity and lack of chronic disease experienced by people in the Mediterranean region. Italy has one of the highest life expectancies in the world. According to the United Nations in the 2010-2015 List Italy ranked 3rd with a life expectancy of 83.31 (including men 80 and women 86.49). Australia ranked 2nd at 83.42 and Japan ranked first at 83.74. Northern Italy has a higher life expectancy than Southern Italy. In Northern Italy there are a number of regions with life expectancy over 84 including Trentino’s expectancy of 84.3, Bolzano of 84.1, Umbria of 84.1, Lombardy 84.0 and Marche 84.

Olive oil is one of the key ingredients in the Mediterranean diet. There are two purposes of this project. The primary is to predict the region and the area in Italy based on the presence of the various fatty acids present in olive oil. The secondary purpose is to analyse the olive data in conjunction with the life expectancy data by region to see if the combination of the data is able to predict the areas with greater level of accuracy.

The first data set is the olive data set from dslabs. The data is compiled and maintained by Rafael Irizarry. The second data set is the regional life expectancy table of Europe of Wikipedia.

First we analyse the data using summary tables, pairs, various graphical representations for the oils in each region. Then we apply the caret package with the rpart partition function to identify which data can predict the areas and regions in Italy.

```
## -- Attaching packages ---------------------------------------------------------------
---- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0       v purrr   0.3.0
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
## -- Conflicts -------------------------------------------------------------------- t
idyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
##           region         area palmitic palmitoleic stearic oleic linoleic
## 1 Southern Italy North-Apulia    10.75        0.75    2.26 78.23     6.72
## 2 Southern Italy North-Apulia    10.88        0.73    2.24 77.09     7.81
## 3 Southern Italy North-Apulia     9.11        0.54    2.46 81.13     5.49
## 4 Southern Italy North-Apulia     9.66        0.57    2.40 79.52     6.19
## 5 Southern Italy North-Apulia    10.51        0.67    2.59 77.71     6.72
## 6 Southern Italy North-Apulia     9.11        0.49    2.68 79.24     6.78
##   linolenic arachidic eicosenoic
## 1      0.36      0.60       0.29
## 2      0.31      0.61       0.29
## 3      0.31      0.63       0.29
## 4      0.50      0.78       0.35
## 5      0.50      0.80       0.46
## 6      0.51      0.70       0.44
```

Data Analysis

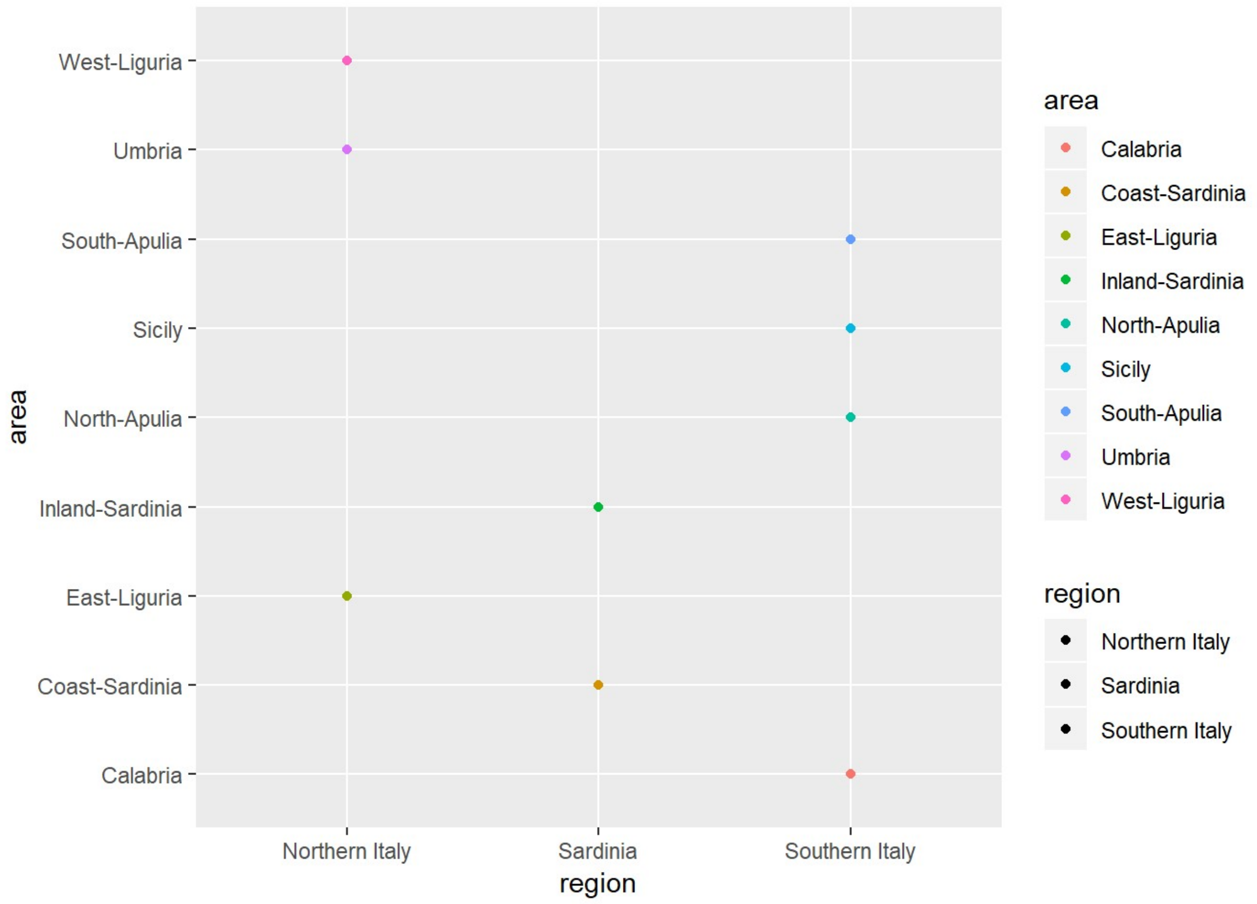
To understand the data we are first going to view the variables in the dataset. We see there is the region, area and the various olive fatty acids.

[Data from the external web address has been cross-checked with the following code. url3 <- “https://raw.githubusercontent.com/rafalab/dslabs/master/inst/extdata/olive.csv (https://raw.githubusercontent.com/rafalab/dslabs/master/inst/extdata/olive.csv)” temp_filename <- tempfile() download.file(url3, temp_filename) dat <- read.csv(temp_filename) However the data omited region and so not used.]

Areas in each region

We view which regions are in each of the areas. Nothern Italy includes West-Liguria, Umbria and East-Liguria. Sardinia includes Inland-Sardinia and Coast-Sardinia. Southern Italy includes South-Apulia, North Apulia, Calabria and Sicily.

```
olive %>% ggplot(aes(region,area,fill = region,color = area)) +
  geom_point()
```



Life Expectancy of European Regions

Next we review the life expectancy by region in the Wikipedia web page.

```
library(rvest)
```

```
## Loading required package: xml2
```

```
##
## Attaching package: 'rvest'
```

```
## The following object is masked from 'package:purrr':
##
##     pluck
```

```
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

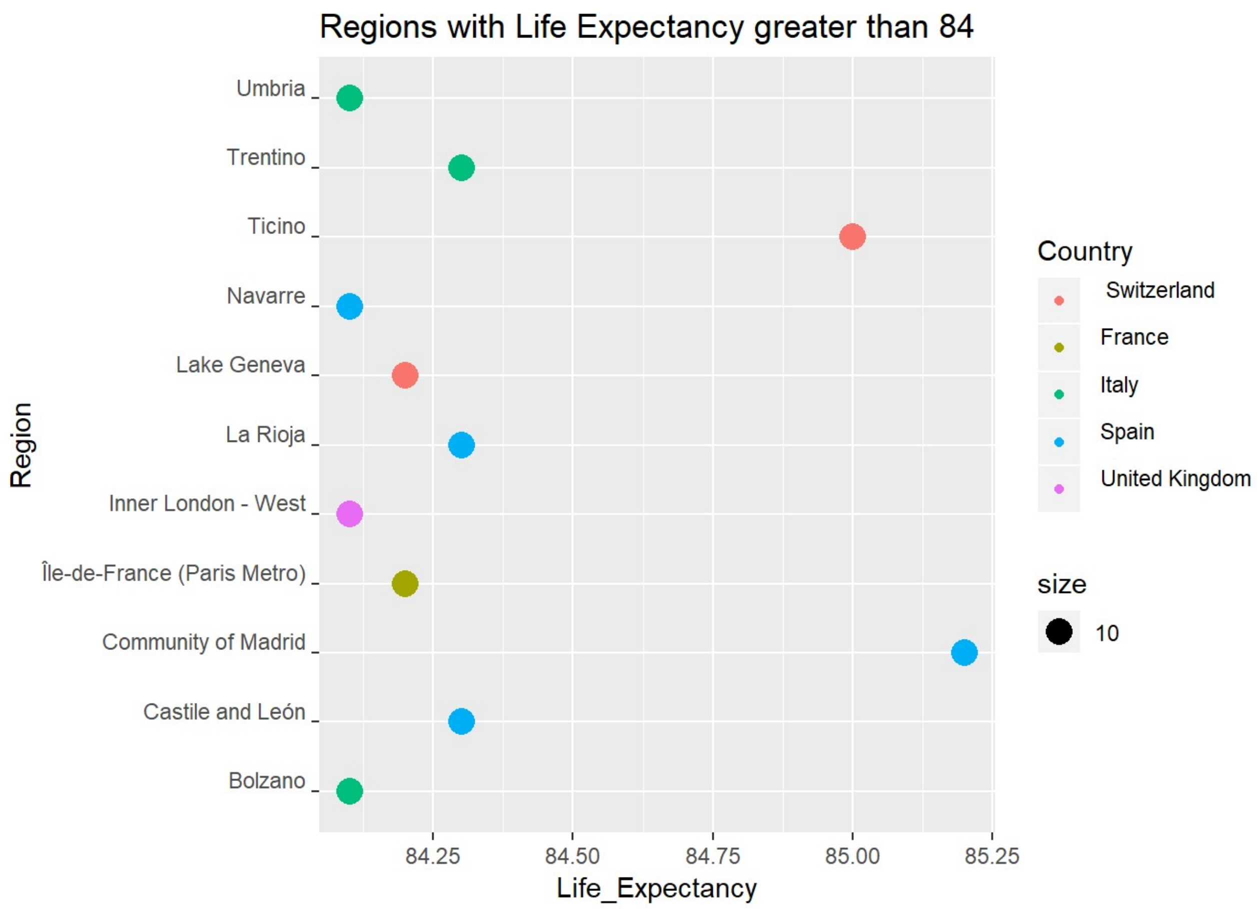
```
library(dplyr)
url <- "https://en.wikipedia.org/wiki/List_of_European_regions_by_life_expectancy"
h <- read_html(url)
tab <- h %>%  html_nodes("table")
tab <- tab[[4]]
tab <- tab %>%  html_table(tab,header = TRUE, fill = TRUE)
names(tab)
```

```
## [1] "Region (NUTS2)\n"
## [2] "Country\n"
## [3] "Life expectancy at birth, total[1]"
## [4] "Life expectancy at birth, men[2]"
## [5] "Life expectancy at birth, woman[3]"
```

```
tab$Country <- tab$`Country`
tab$Life_Expectancy <- tab$`Life expectancy at birth, total[1]`
tab$Region <- tab$`Region (NUTS2)`
tabI <- tab %>% select(Country,Region,Life_Expectancy)
tabI %>% group_by(Country) %>% arrange(desc(Life_Expectancy)) %>%
  filter(Life_Expectancy > 83) %>%
  summarize("Life Expectancy" = mean(Life_Expectancy))
```

```
## # A tibble: 6 x 2
##   Country             `Life Expectancy`
##   <chr>                           <dbl>
## 1 "  Switzerland\n"                83.9
## 2 " France\n"                      83.6
## 3 " Greece\n"                      83.4
## 4 " Italy\n"                       83.7
## 5 " Spain\n"                       84
## 6 " United Kingdom\n"              83.8
```

```
tabI %>% filter(Life_Expectancy > 84) %>% group_by(Country) %>%
  ggplot(aes(Life_Expectancy,Region,size = 10,colour = Country))+
  geom_point()+
  ggtitle("Regions with Life Expectancy greater than 84")
```



```
library(tidyr)
tabItaly <- tabI[133:152,]
tabItaly$area <- tabItaly$Region
tabItaly$area_life <- tabItaly$Life_Expectancy
tabItaly <- tabItaly %>% select(-Region)
tabItaly <- tabItaly %>% select(-Country)
tabItaly <- tabItaly %>% select(-Life_Expectancy)
tabItaly <- separate(tabItaly,area,c("area",NA), sep = "\n")
```

Estimated Life Expectancy in each area

We are also comparing olives in different regions to life expectancy. By viewing the Life Expectancy table in Wikipedia by European region we can see that the areas are not exactly aligned to the olives dataset.

Given that these are incomplete estimates we have created a data frame for the areas and the regions then we will join this data with our olive set. We can also see that the olive data does not encompass every region. For example the town of Acciarola is reputed to have one of the highest number of centenarians and this is in the Campania region which does not appear in the olive area dataset.

```
r1 <- c("Northern Italy","Northern Italy","Northern Italy","Sardinia","Sardinia","Sout
hern Italy","Southern Italy","Southern Italy","Southern Italy")
a1 <- c("West-Liguria","East-Liguria","Umbria","Inland-Sardinia","Coast-Sardinia","Nor
th-Apulia","Calabria","South-Apulia","Sicily")
tabregion <- data.frame(region = r1,area = a1)
tabregion
```

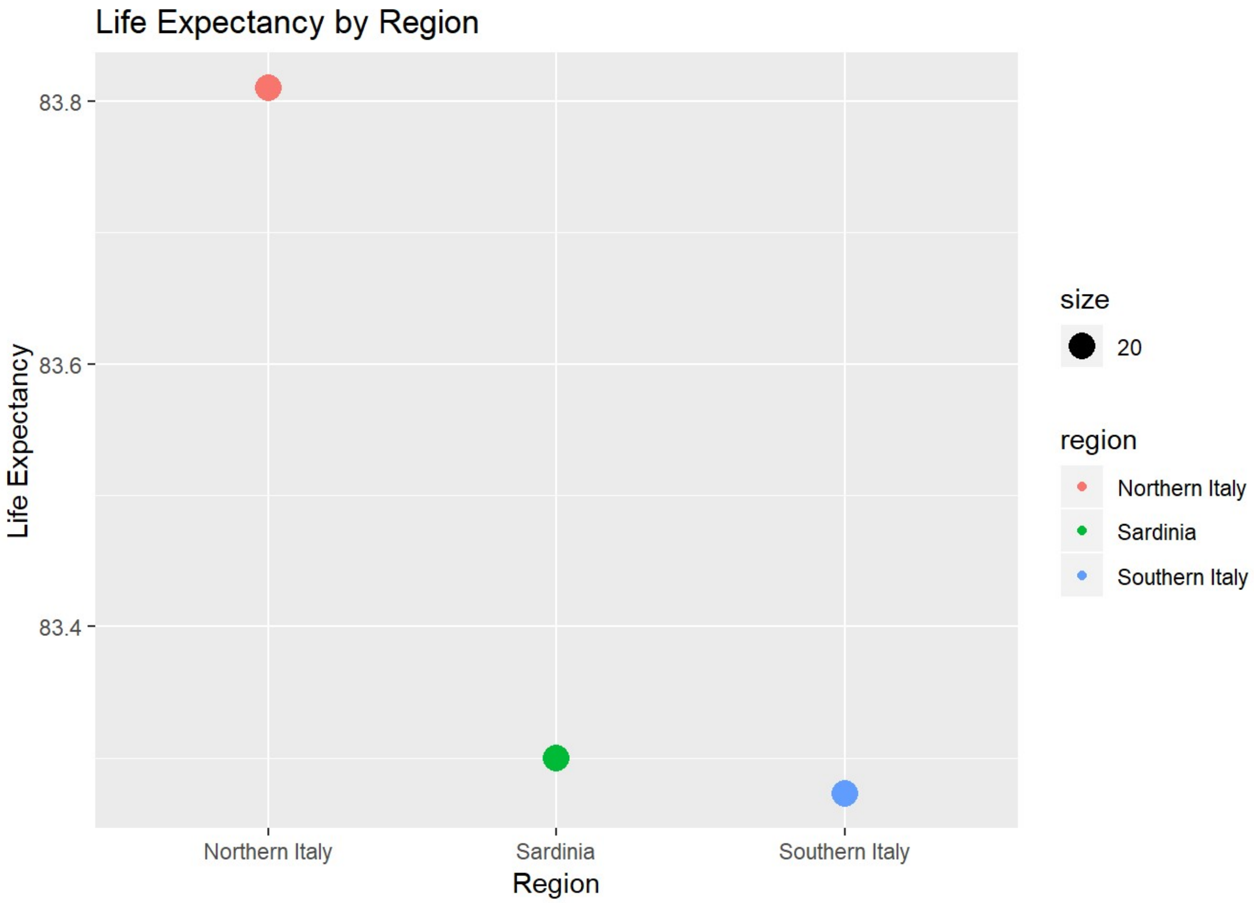
```
##           region            area
## 1 Northern Italy    West-Liguria
## 2 Northern Italy    East-Liguria
## 3 Northern Italy          Umbria
## 4       Sardinia Inland-Sardinia
## 5       Sardinia  Coast-Sardinia
## 6 Southern Italy    North-Apulia
## 7 Southern Italy        Calabria
## 8 Southern Italy    South-Apulia
## 9 Southern Italy          Sicily
```

Next we merge the life expectancy by region from Wikipedia with the regional table just created. The remaining gaps where the names are not aligned are filled. eg. the life expectancy for Sardinia from Wiki is applied to North and South Sardinia. East Liguria is assigned the average life expectancy of Trentino-South Tyrol 84.3,Veneto 83.9,Emilia-Romagna 83.5 and Fruili-Venezia-Guiala 84.3. West Liguria is assigned the average of Liguria 83.6, Lombardy 84, Piedmont 83.3 and Aosta Valley 83.0 of 83.6. The expectancy for Apulia is assigned to North and South Apulia.

```
life <- merge(tabregion,tabItaly, by.x = "area",all.x = TRUE)
#Assign life to Sardinian regions of 83.3
life$area_life[2] <- 83.3
life$area_life[4] <- 83.3

#Assign East Liguria and West Liguria the average of their regions
life$area_life[3] <- 83.85
life$area_life[9] <- 83.475

#Assign life expectancy of Apulia to South Apulia and north Apulia
life$area_life[5] <- 83.5
life$area_life[7] <- 83.5

olivelife <- olive %>% merge(life, by = "area", all.x = TRUE,all.y = TRUE)
olivelife$region <- olivelife$region.x
olivelife <- olivelife %>% select(-region.x)
olivelife <- olivelife %>% select(-region.y)
```

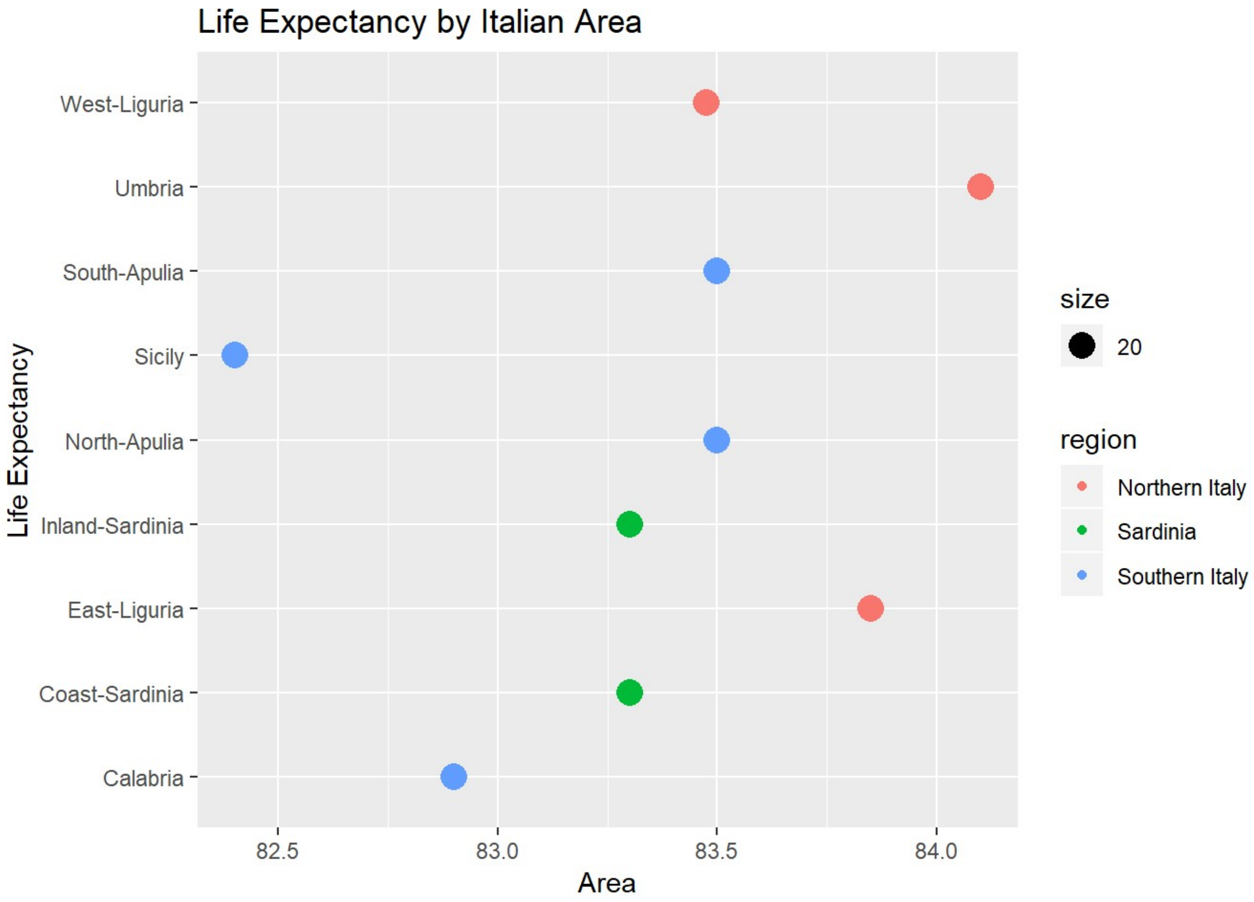
Then we can summarize the life expectancy for the three regions based on the data we have from the areas. We combine the olive and regionlife data and graph the life expectancy by region.

```
regionlife <- olivelife %>% group_by(region) %>%
  summarize(region_life = mean(area_life))
olivelife <- merge(olivelife,regionlife, by = "region", all.x = TRUE)
olivelife %>% group_by(region) %>%
  mutate(region_life = mean(area_life)) %>%
  ggplot(aes(region,region_life,color = region,size = 20)) +
  geom_point() +
  ggtitle("Life Expectancy by Region")+
  xlab("Region")+
  ylab("Life Expectancy")
```



Next we view the life expectancy by the area in Italy.

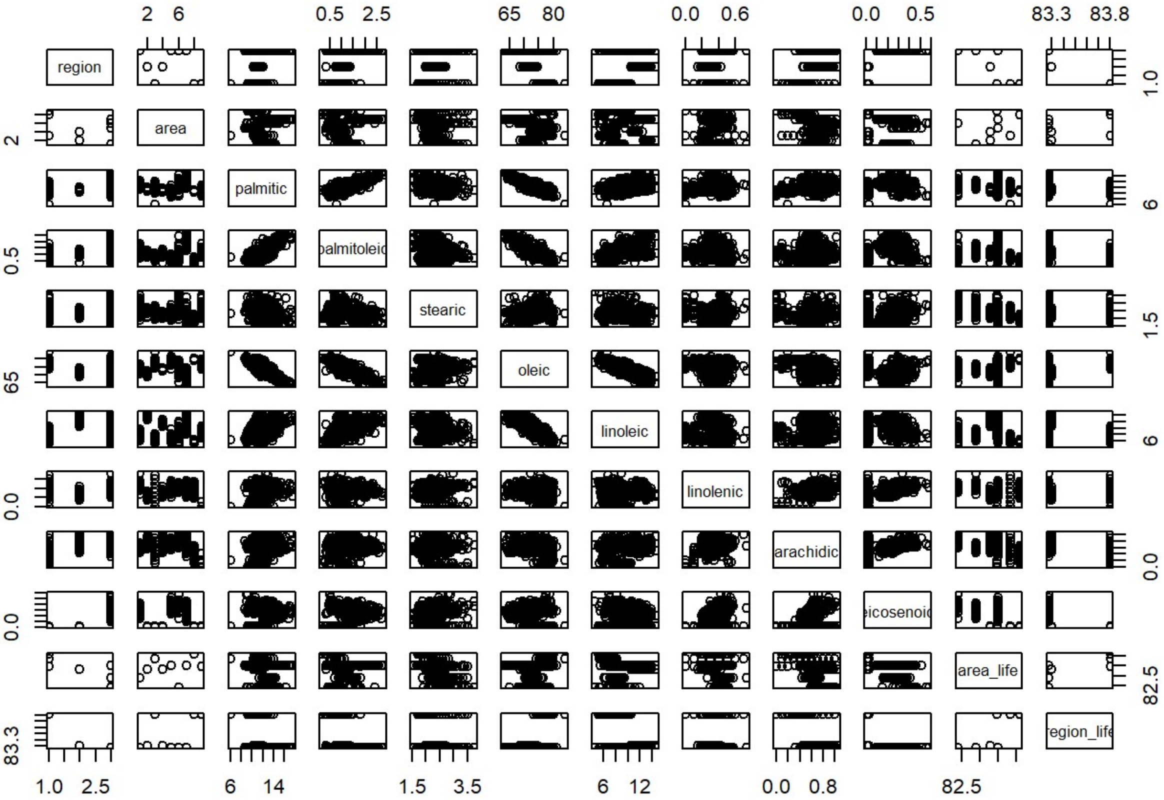
```
olivelife %>% ggplot(aes(area_life,area,color = region,size = 20)) +
  geom_point() +
  ggtitle("Life Expectancy by Italian Area")+
  xlab("Area")+
  ylab("Life Expectancy")
```



Correlation between the olive fatty acids and regions and life expectancy

Next we analyse the correlation between each of the variables using the pairs function for olives across all regions.

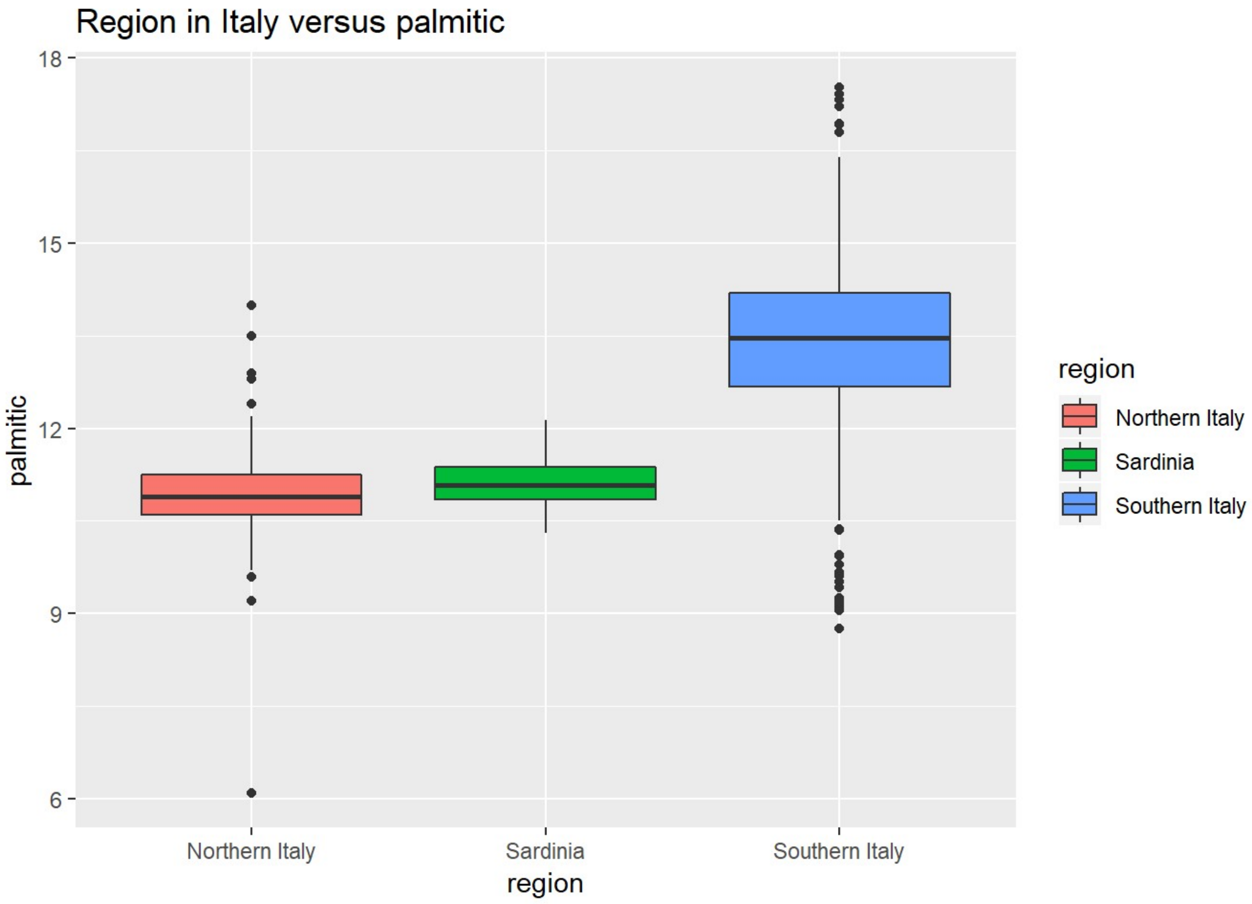
```
pairs(olivelife)
```



Visualisation of the fatty acids in each region

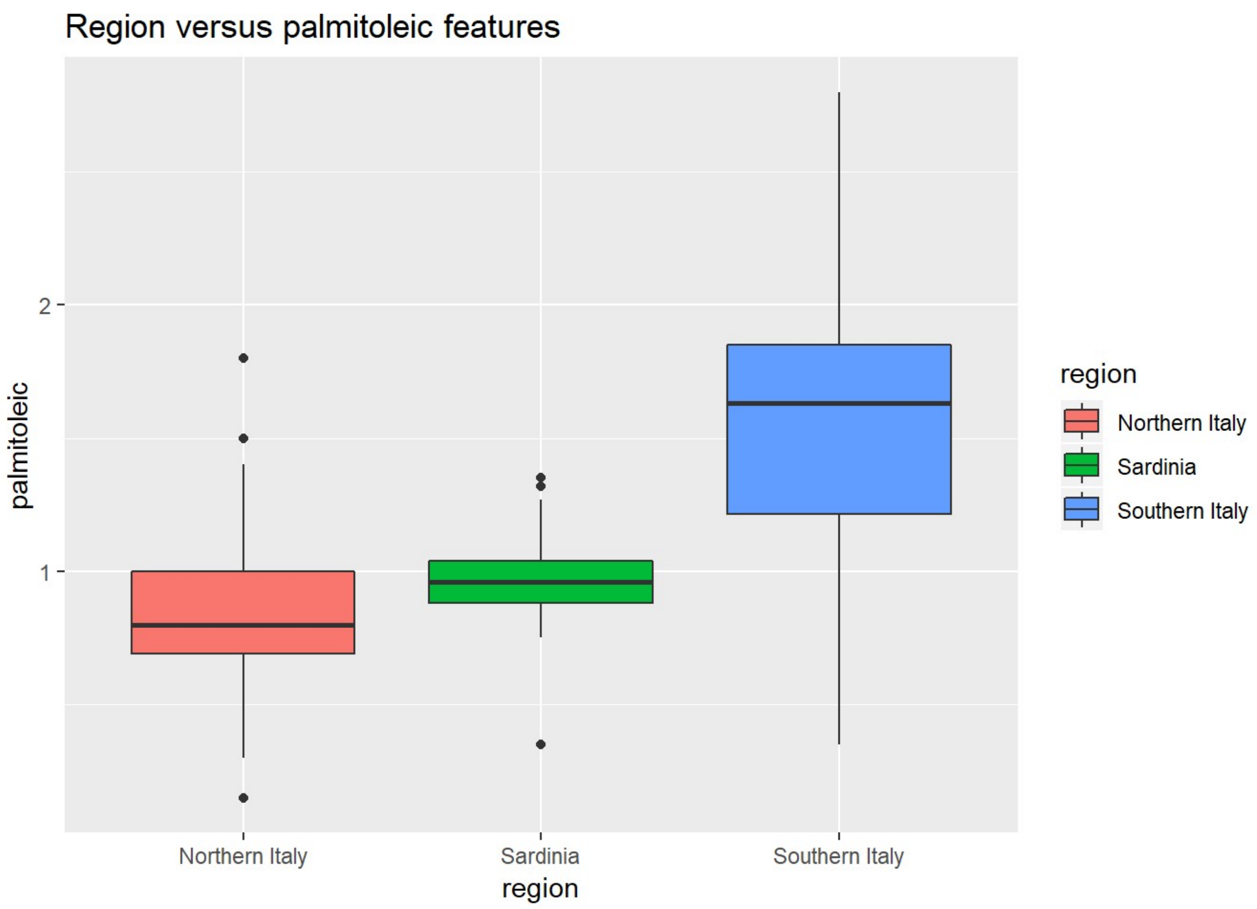
Region versus palmitic

```
olive %>% ggplot(aes(region, y = palmitic, fill = region))+
  geom_boxplot()+
  ggtitle("Region in Italy versus palmitic")
```



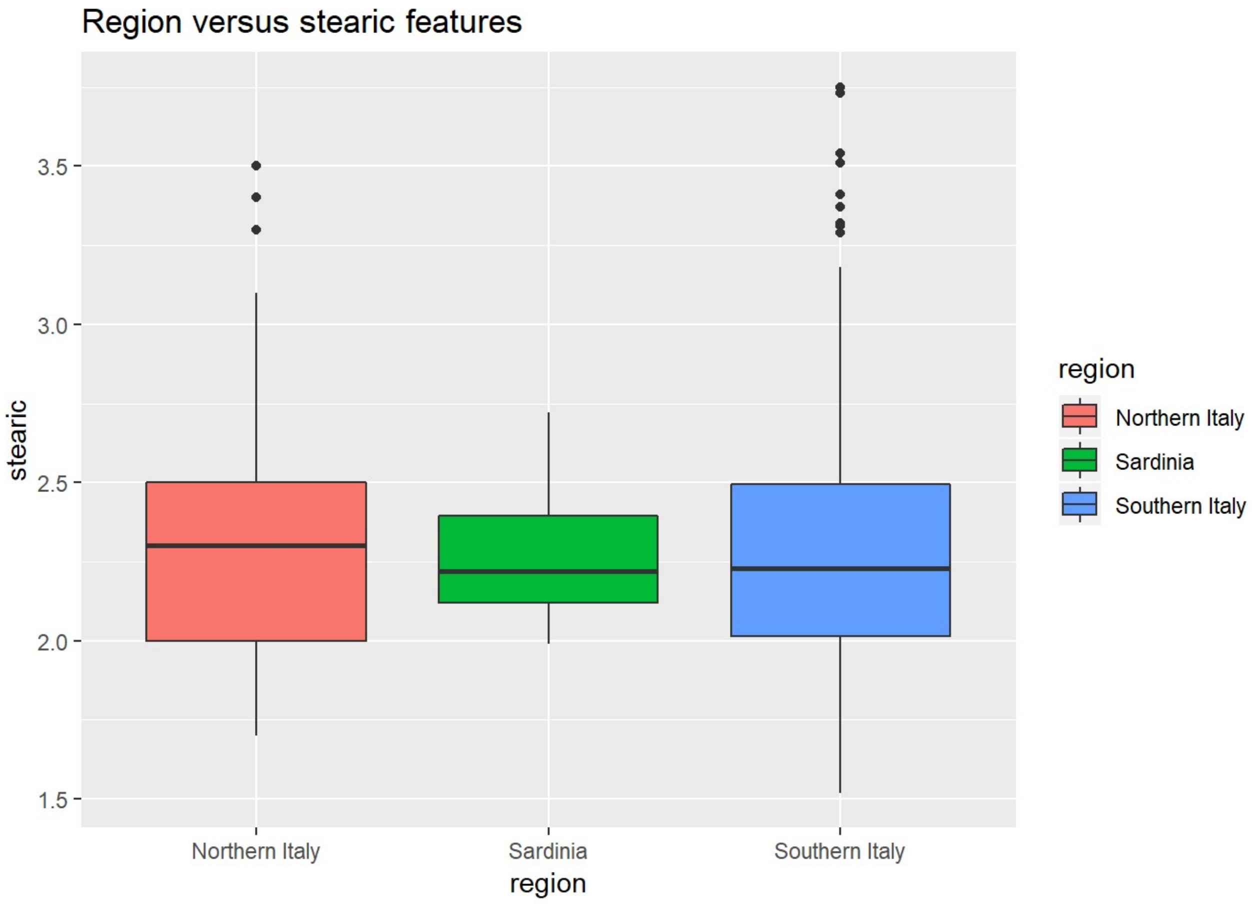
Region versus palmitoleic features

```
olive %>% ggplot(aes(region, palmitoleic,fill = region))+
  geom_boxplot()+
  ggtitle("Region versus palmitoleic features")
```



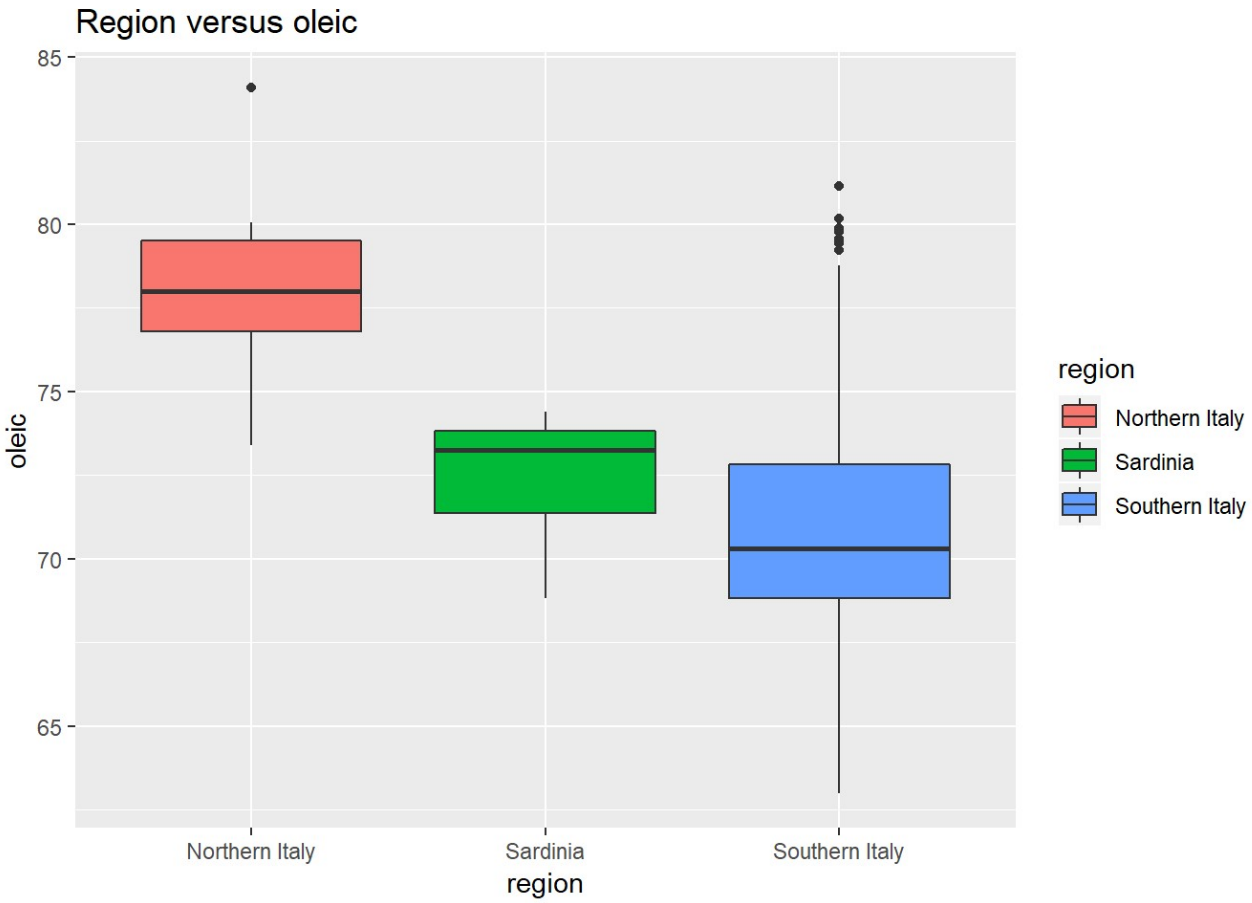
Region versus stearic features

```
olive %>% ggplot(aes(region, stearic,fill = region))+
  geom_boxplot()+
  ggtitle("Region versus stearic features")
```



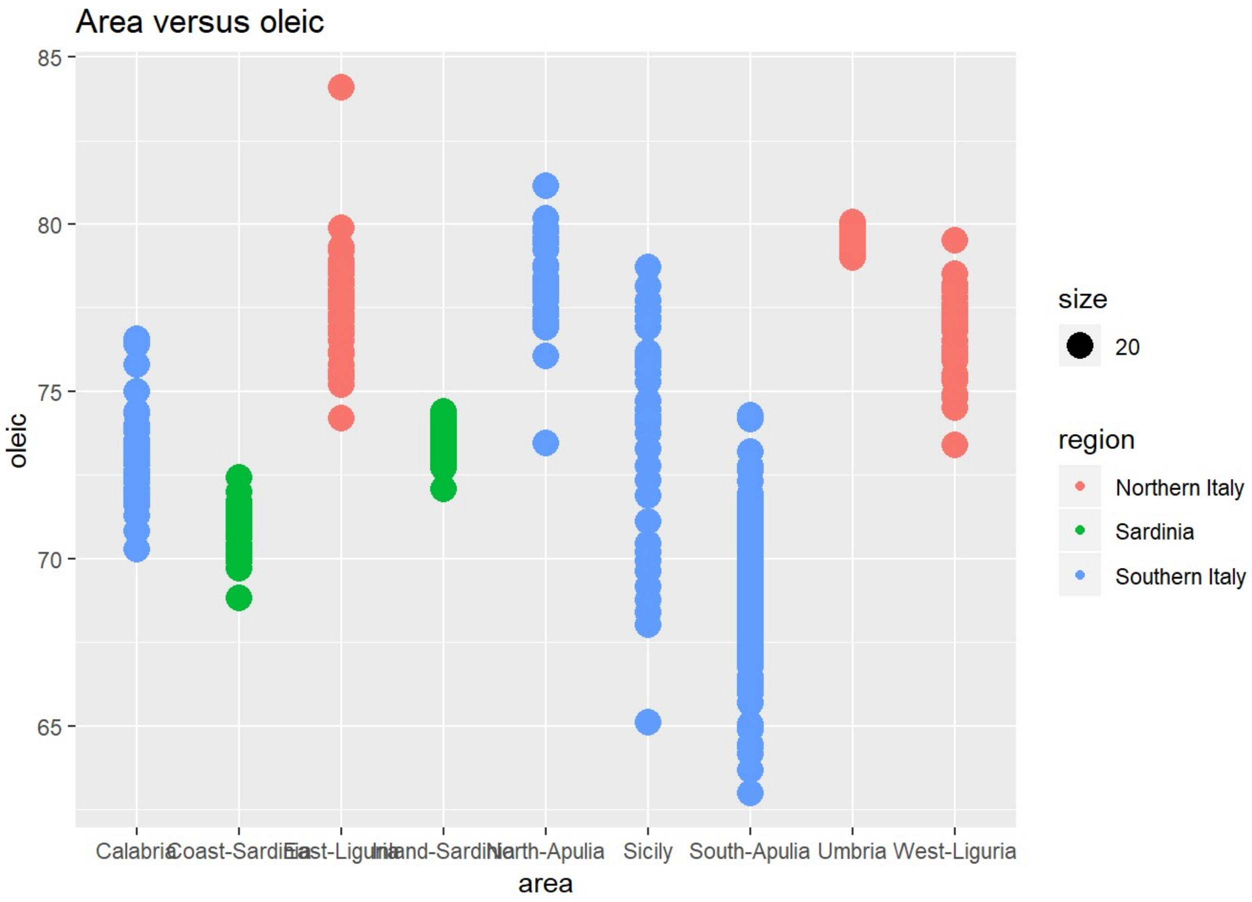
Region versus oleic

```
olive %>% ggplot(aes(region, y = oleic, fill = region))+
  geom_boxplot()+
  ggtitle("Region versus oleic")
```

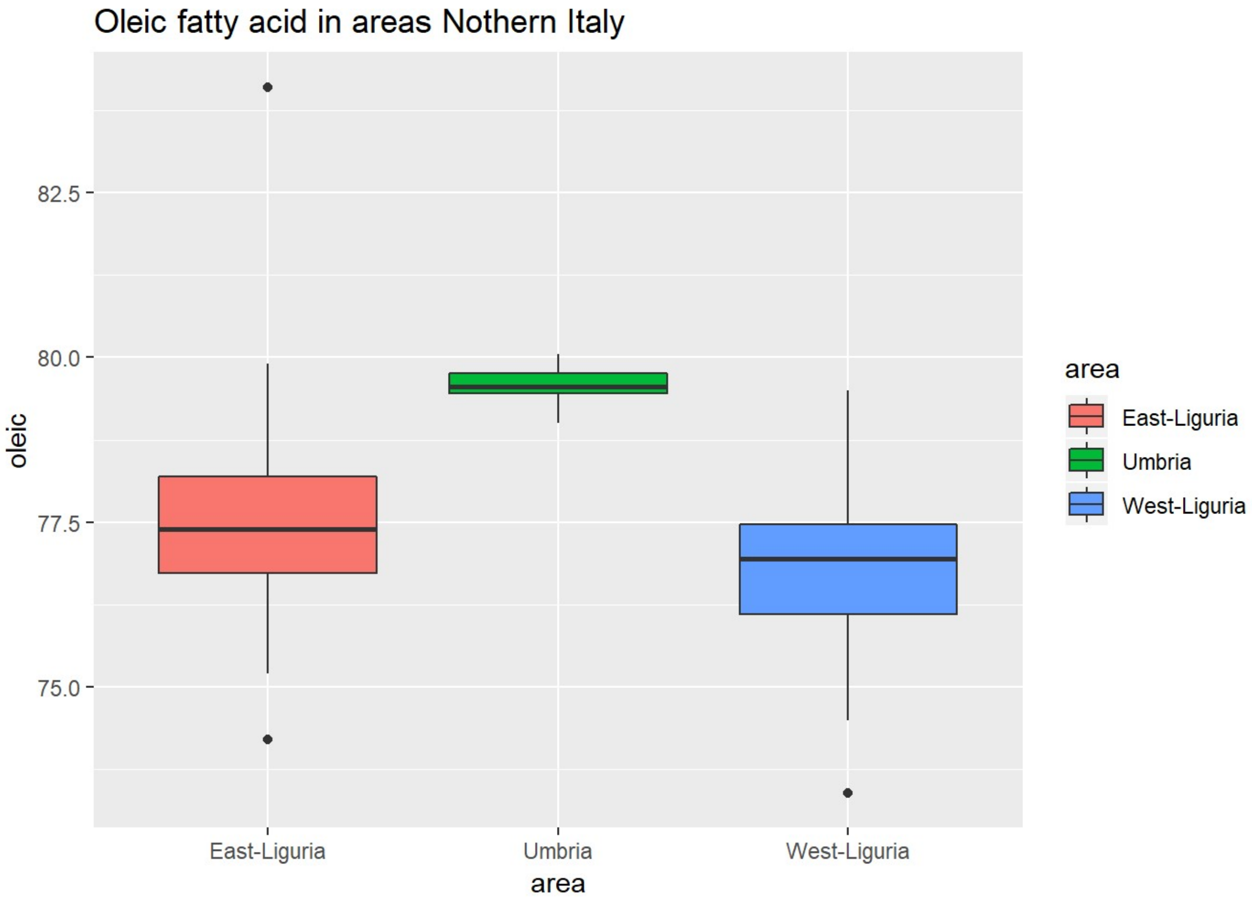


Area versus oleic

```
olivelife %>% ggplot(aes(area, oleic, fill = region, color = region, size = 20))+
  geom_point()+
  ggtitle("Area versus oleic")
```



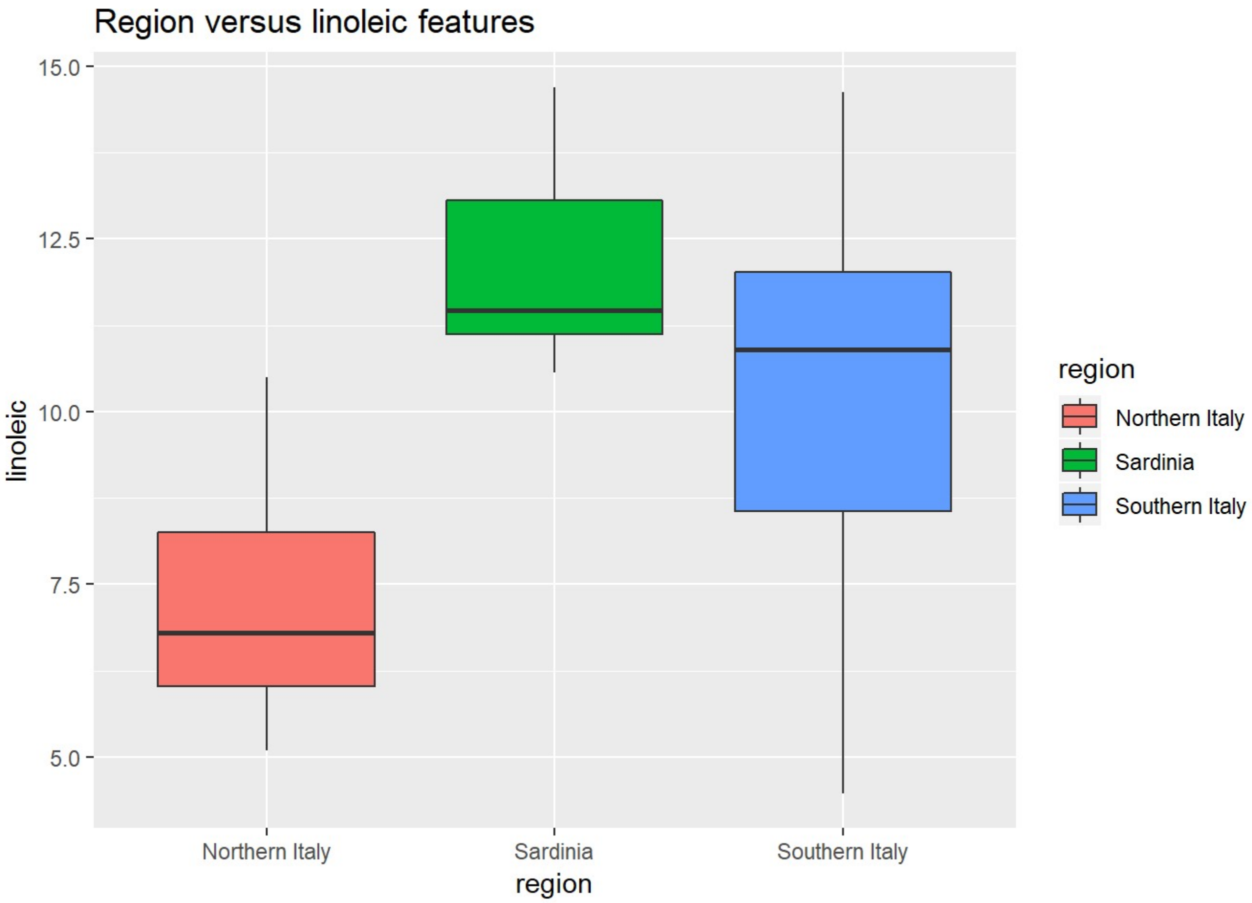
```
olive %>% filter(region =="Northern Italy") %>%
  group_by(area) %>%
  ggplot(aes(area,oleic,fill = area)) +
  geom_boxplot() +
  ggtitle("Oleic fatty acid in areas Nothern Italy")
```



The oleic fatty acid is the highest on average in Umbria follwed by East Liguria and then West Liguria.

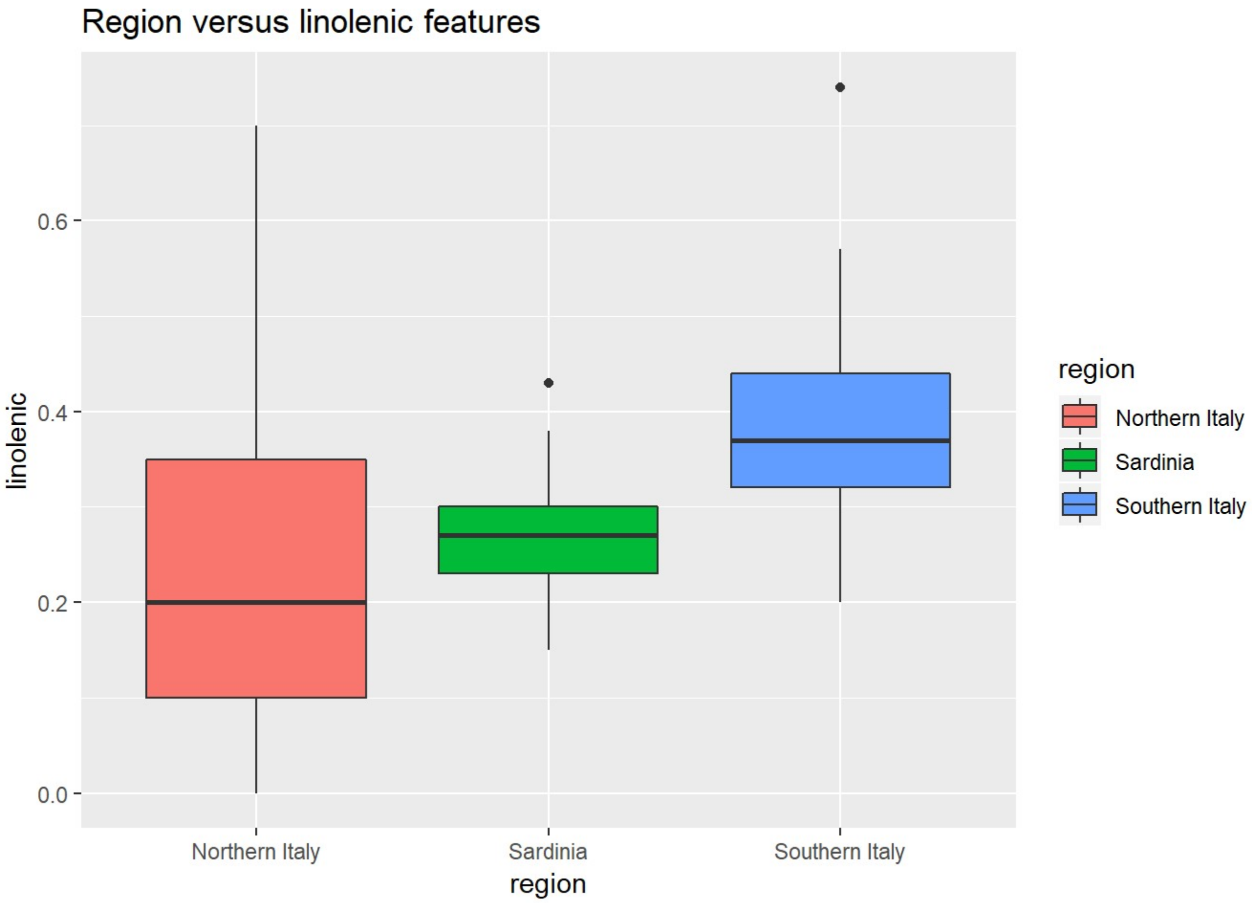
Region versus linoleic features

```
olive %>% ggplot(aes(region, linoleic,fill = region))+
  geom_boxplot()+
  ggtitle("Region versus linoleic features")
```



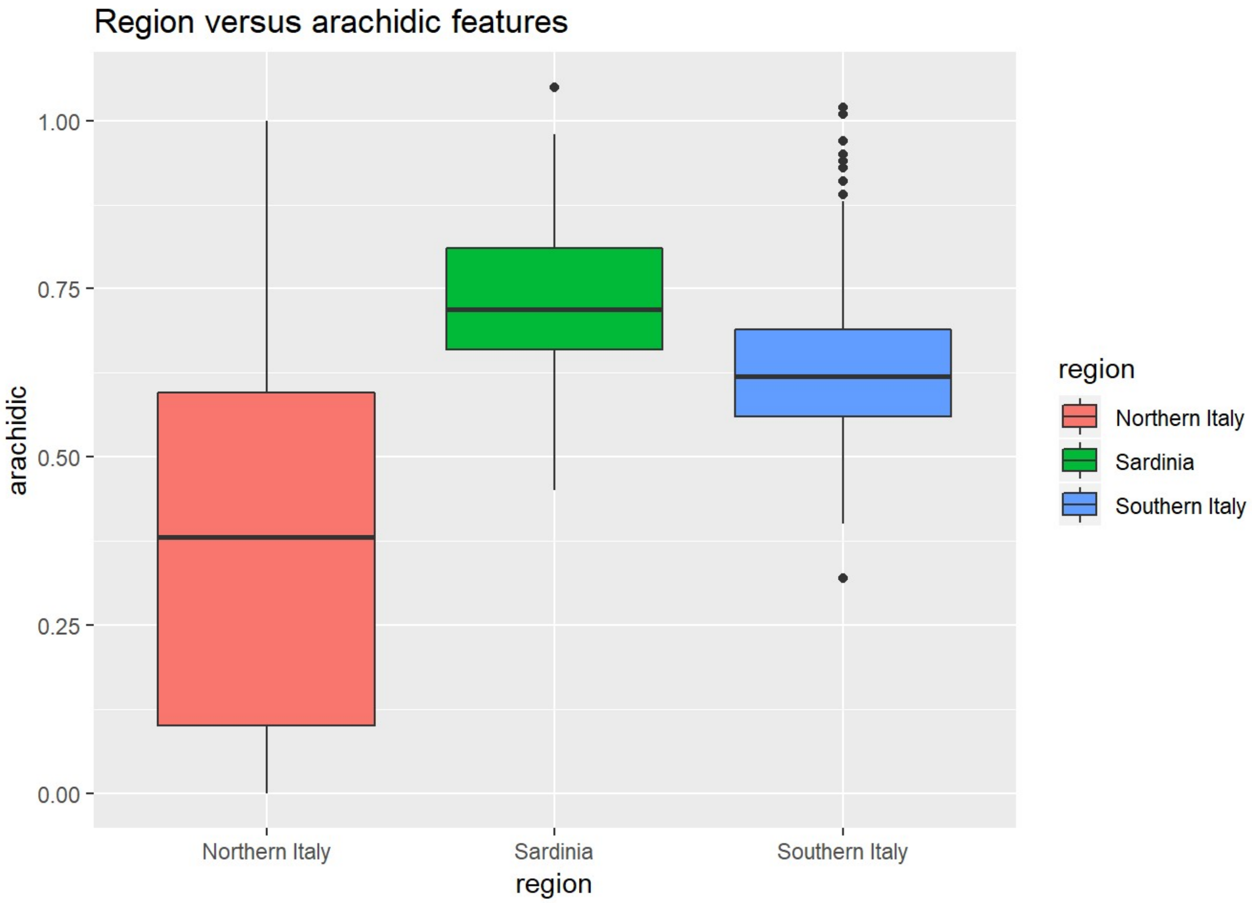
Region versus linolenic features

```
olive %>% ggplot(aes(region, linolenic,fill = region))+
  geom_boxplot()+
  ggtitle("Region versus linolenic features")
```



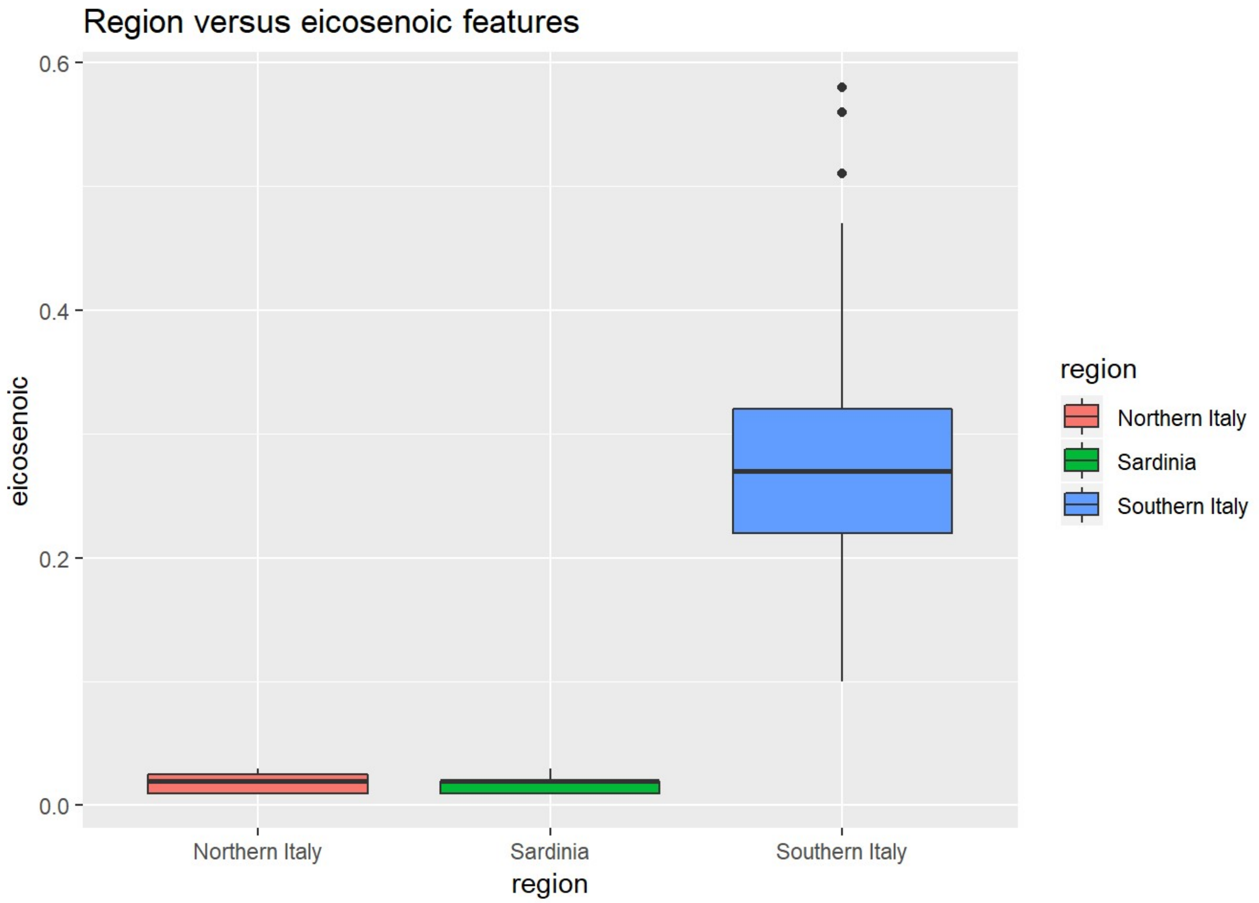
Region versus arachidic features

```
olive %>% ggplot(aes(region, arachidic,fill = region))+
  geom_boxplot()+
  ggtitle("Region versus arachidic features")
```



Region versus eicosenoic features

```
olive %>% ggplot(aes(region, eicosenoic,fill = region))+
  geom_boxplot()+
  ggtitle("Region versus eicosenoic features")
```


Region versus eicosenoic features

Create the model and print the partitioned table.

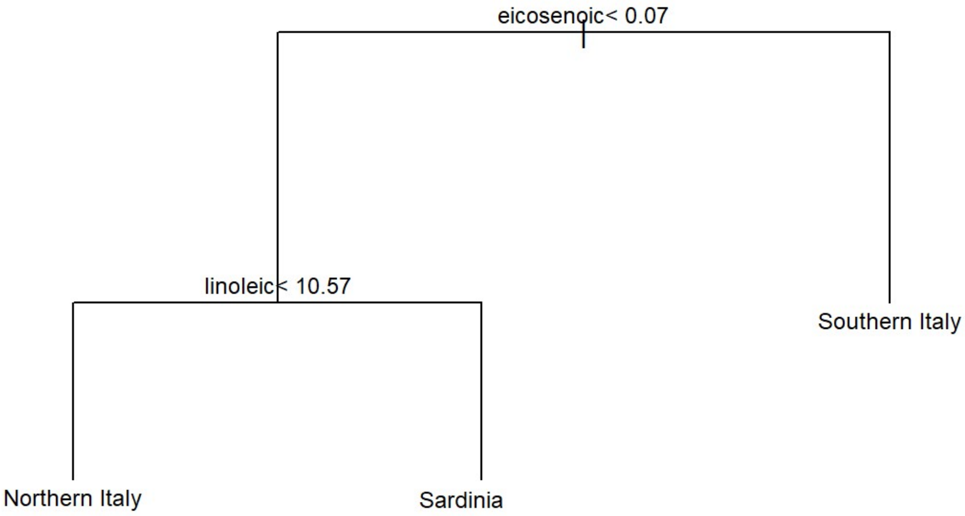
```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(rpart)

set.seed(1)
test_index <- createDataPartition(olive$region, times = 1, p = 0.5, list = FALSE)

test_set <- olive[test_index, ]
train_set <- olive[-test_index, ]
train <- train(region ~ ., method = "rpart",
               tuneGrid = data.frame(cp = seq(0, 0.1, 15)),
               data = train_set)

plot(train$finalModel, margin = 0.1)
text(train$finalModel, cex = 0.75)
```

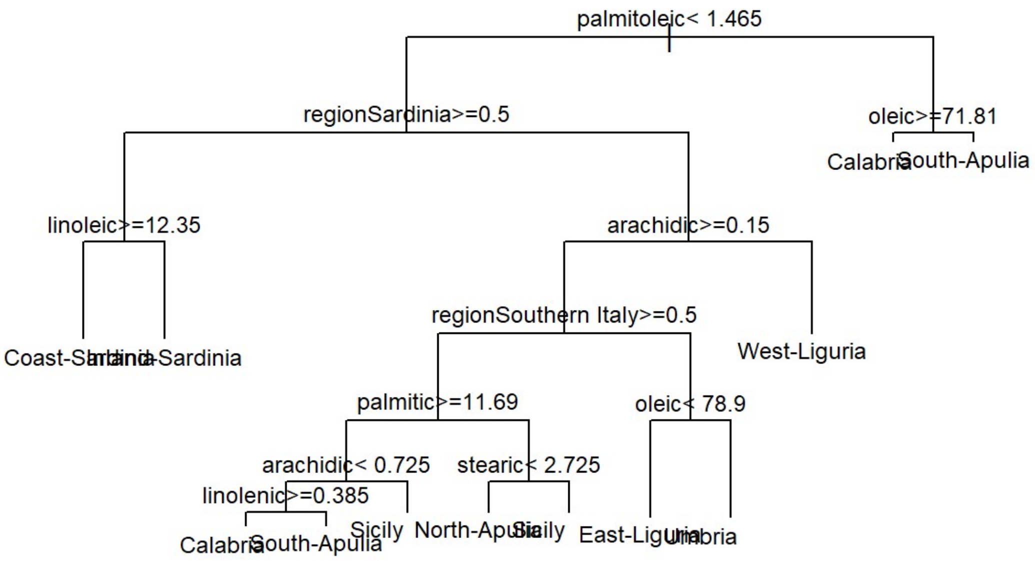


Test the Accuracy of the Partitioned Prediction on the test set.

```
confusionMatrix(predict(train, test_set), test_set$region)$overall["Accuracy"]
```

```
## Accuracy
##        1
```

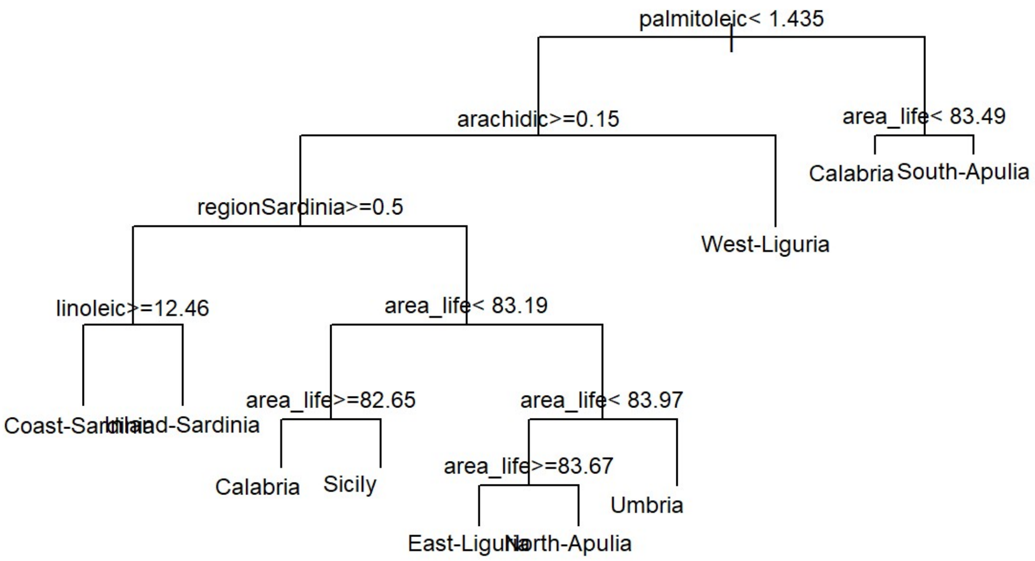
```
set.seed(1)
test_index2 <- createDataPartition(olive$area, times = 1, p = 0.5, list = FALSE)

test_set2<- olive[test_index, ]
train_set2 <- olive[-test_index, ]
train2 <- train(area ~ ., method = "rpart",
               tuneGrid = data.frame(cp = seq(0, 0.1, 15)),
               data = train_set2)

plot(train2$finalModel, margin = 0.05)
text(train2$finalModel, cex = 0.75)
```



```
confusionMatrix(predict(train2, test_set2), test_set2$area)$overall["Accuracy"]
```

```
## Accuracy
## 0.902439
```

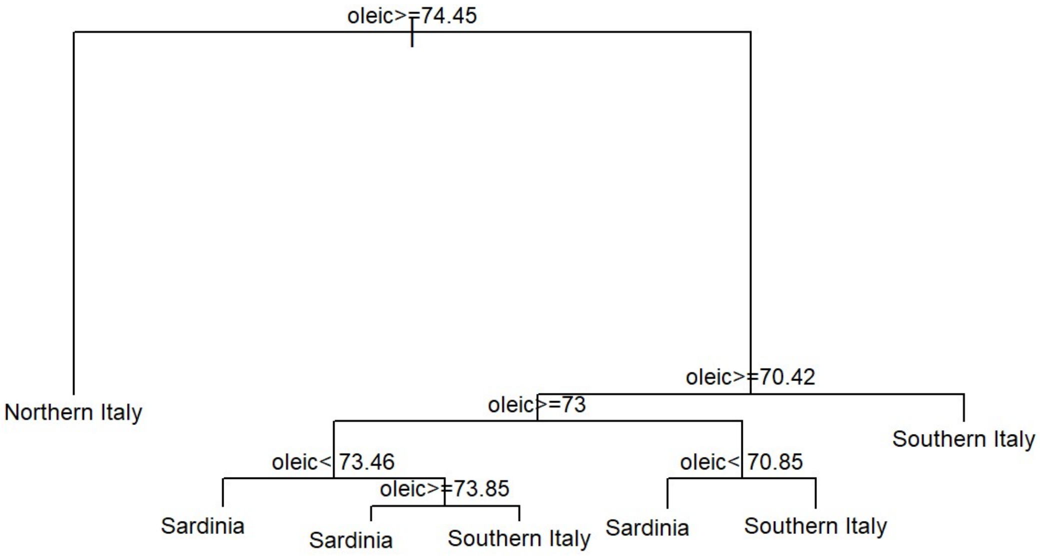
```
set.seed(1)
test_index3 <- createDataPartition(olivelife$area, times = 1, p = 0.5, list = FALSE)

test_set3<- olivelife[test_index, ]
train_set3 <- olivelife[-test_index, ]
train3 <- train(area ~ ., method = "rpart",
               tuneGrid = data.frame(cp = seq(0, 0.1, 10)),
               data = train_set3)

plot(train3$finalModel, margin = 0.05)
text(train3$finalModel, cex = 0.75)
```



```
confusionMatrix(predict(train3, test_set3), test_set3$area)$overall["Accuracy"]
```

```
##  Accuracy
## 0.9372822
```

```
set.seed(1)
test_index4 <- createDataPartition(olivelife$area, times = 1, p = 0.5, list = FALSE)

test_set4<- olivelife[test_index, ]
train_set4 <- olivelife[-test_index, ]
train4 <- train(region ~ oleic, method = "rpart",
                tuneGrid = data.frame(cp = seq(0, 0.1, 10)),
              data = train_set4)

plot(train4$finalModel, margin = 0.05)
text(train4$finalModel, cex = 0.75)
```



```
confusionMatrix(predict(train4, test_set4), test_set4$region)$overall["Accuracy"]
```

```
##  Accuracy
## 0.7491289
```

Results

Model 1: Firstly, the Italian region with the higest life expectancy is Northern Italy. This region can be identified with 100% accuracy by the presence of the eicosenoic fatty acid being less than 0.07% of the composition of the olives and the linoleic fatty acid being less than 10.57%.

However what is more interesting is that Northern Italian olives have the highest percentage of oleic fatty acids of the three regions. Is this driving the life expectancy? The high percentage of oleic fatty acid can predict Northern Italy with accuracy but is less successful at discriminating between Sardinia and Southern Italy.

Model 2: Using the olive data set the area is predicted with an accuracy of 90.2%.

Model 3: If we include the life expectancy data in the combined data set our ability to product an area increases to Using the combined life and olive data set, areas can be predicted with a 93.7% overall accuracy by the combination of fatty acids and the life expectancy. The one area with the life expectancy of greater than 83.82 is East Liguria in Northern Italy. This is the area including Trentino-South Tyrol, Emilia-Romagna, Fruilli-Venezia Guilia and Veneto.

Model 4: This model is used to predict the area based on the oleic fatty acid only. This model's accuracy drops to a poor 74.9% level of accuracy.

Conclusion

The olive oil fatty acids data can predict the region in Italy with 100% accuracy. When we consider the olive oil of all of the areas in Italy, we can predict the area with a 90.2% accuracy. When we add the life expectancy data we can predict the area with a 93.7% level of accuracy. This is despite tha fact that the coverage of the olive oil data in every area is not complete and the alignment between the regions in the olive data set and the life expectancy data from wikipedia is not exactly aligned.

We can see that the region in Italy with the highest life expectancy is Northern Italy and within that East Liguria. This region is characterised by olive oil with the highest percentage of the oleic fatty acid. Can we conclude that this is driving the long life expectancy? Our prediction of the region based on the oleic fatty acid alone achieves a much lower level of accuracy of 74.9% suggesting that this alone is not sufficient as a sole predictor. Other factors may contribute to a high life expectancy in Northern Italy such as the level of exercise, stress, other dietarty factors and genetics of the population. The findings here suggest that olive oil is a contributing factor to the long life expectancy in Italy and Northern Italy in particular.