K-means Clustering on Olive Oil Data Set

Jennifer Luu

Sara Rettus

Jelena Segan

Louis Tran

MATH 167PS

Data Set Description

572 samples of olive oil from Italy.

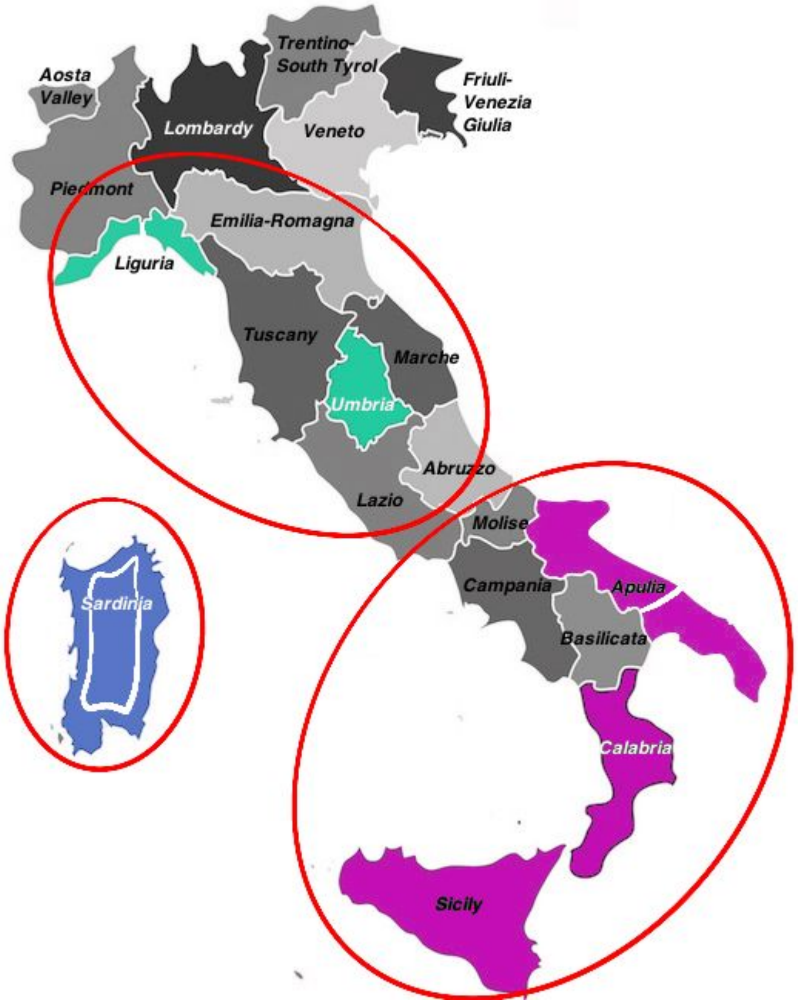
2 categorical variables

- **Macro area:** northern Italy, southern Italy, and Sardinia
- **Region:** north Apulia, south Apulia, Calabria, Sicily, east Liguria, west Liguria, Umbria, Sardinia in-land, Sardinia coastal

8 Continuous Variables

- Acid components of olive oil
- Percentage x 100
- **Components:** palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, and eicosenoic

Geography

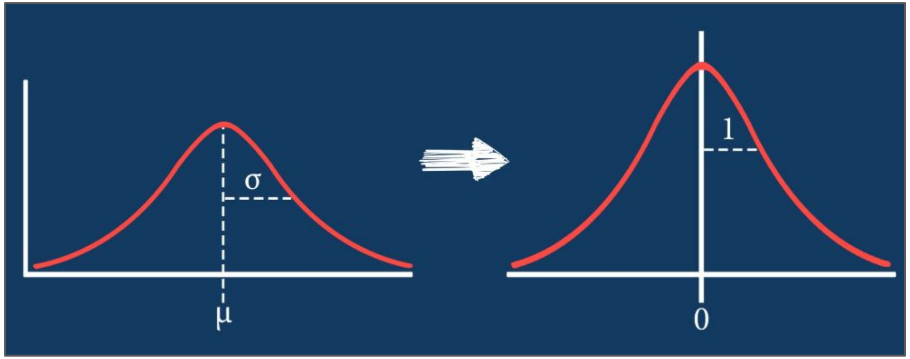

Trentino-South Tyrol
Aosta Valley
Friuli-Venezia Giulia
Lombardy
Veneto
Piedmont
Emilia-Romagna
Liguria
Tuscany
Marche
Umbria
Abruzzo
Lazio
Molise
Sardinia
Campania
Apulia
Basilicata
Calabria
Sicily

Goals of Our Project

1. Determine groups that capture the relationship between acid components and region
2. Choose the optimal number for K
3. Measure performance of K-means clustering
4. Visualize variation

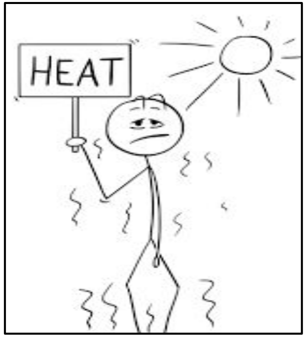
Data Preparation

- Measurements are percentages of each fatty acid composition x 100 for all 8 variables
  - Justifications:
    - Unscaled - same units of measurements
    - Scaled - wide range of numbers
      - Min of eicosenoic - 1
      - Max of oleic - 8410
- Test and train
  - Split 70% of full data on train and 30% to test



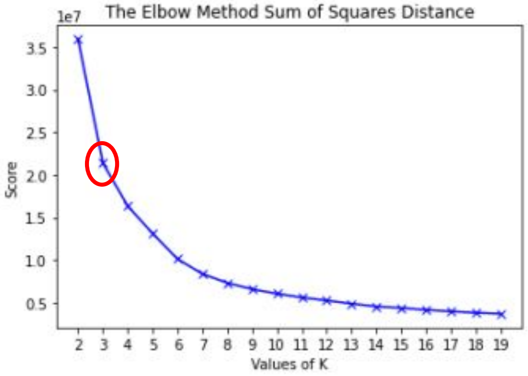
K-means Clustering

- **Goal:** Is there a connection between regions and fatty acid composition?
  - Regional borders are socially constructed
    - Did not assume that clusters would fall along them
- Ran clustering multiple times with random states 10, 11, and 12
  - Want to see if results stabilize



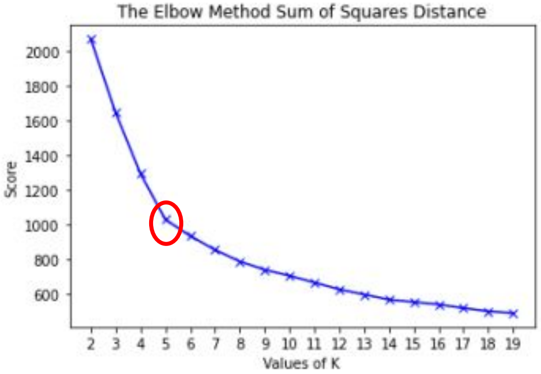
Statistical Methods: Selecting K - Elbow Method

- Visualizes the explained variation as a function of the number of clusters
  - Inertia: sum of squared distances between the observation and its cluster center
- Look for a sharp turning point (elbow)



Unscaled

K = 3

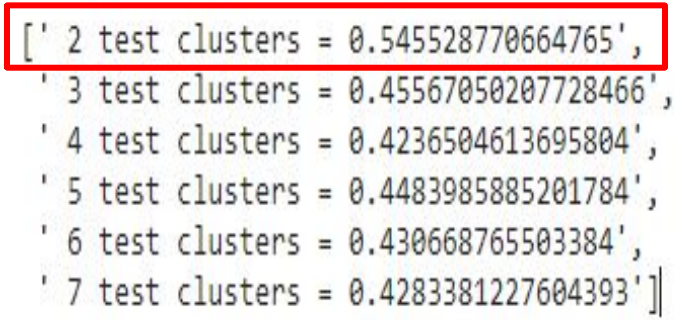


Scaled

K = 5

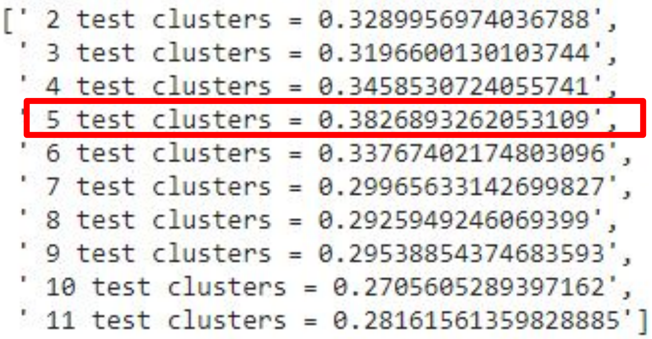
Statistical Methods: Selecting K - Silhouette Method

- Measures on average how close each data point is to its own cluster compared to other clusters
- Calculated with Euclidean distance
- Scores from [-1, 1]
  - 1 = points are well matched to their clusters, -1 = poorly matched



```
[' 2 test clusters = 0.545528770664765',
 ' 3 test clusters = 0.45567050207728466',
 ' 4 test clusters = 0.4236504613695804',
 ' 5 test clusters = 0.4483985885201784',
 ' 6 test clusters = 0.430668765503384',
 ' 7 test clusters = 0.4283381227604393']
```

Unscaled data
K = 2

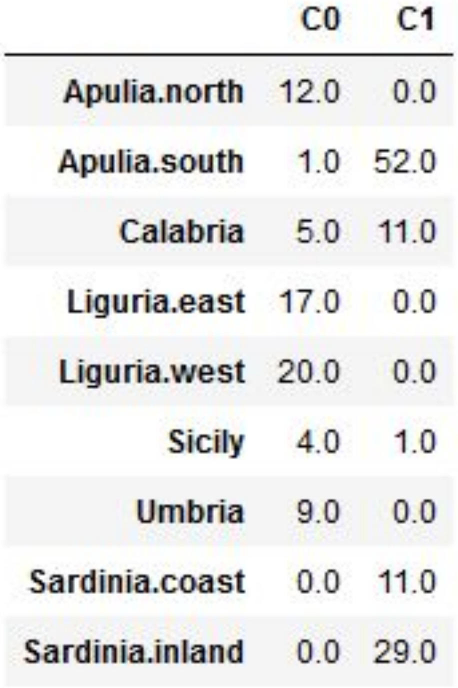


```
[' 2 test clusters = 0.3289956974036788',
 ' 3 test clusters = 0.3196600130103744',
 ' 4 test clusters = 0.345853072405574 1',
 ' 5 test clusters = 0.3826893262053109',
 ' 6 test clusters = 0.33767402174803096',
 ' 7 test clusters = 0.29965633142699827',
 ' 8 test clusters = 0.292594924606939 9',
 ' 9 test clusters = 0.29538854374683593',
 ' 10 test clusters = 0.2705605289397162',
 ' 11 test clusters = 0.2816156135982888 5']
```

Scaled data
K = 5

Creating *Pseudo True* Labels

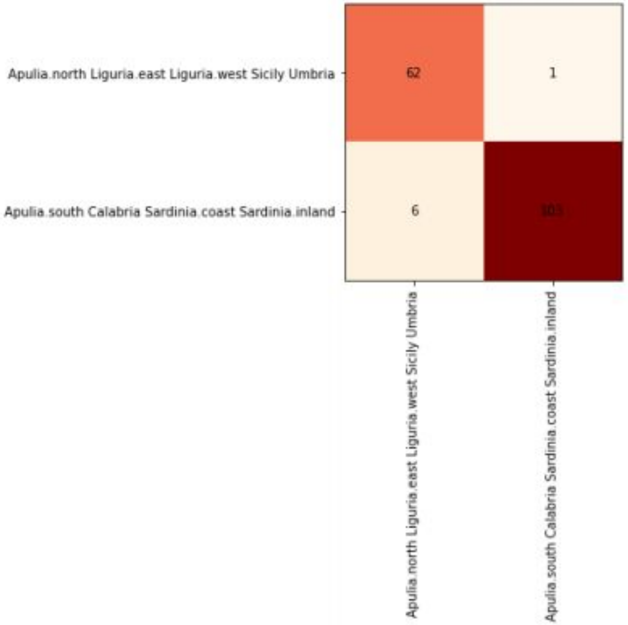
- **Problem:** Data set only provided the macro area and region to which each sample belongs.
    - We do not know to which areas clusters truly belong (if any).
    - Determined the labels (consists of regions) based on highest counts.



	C0	C1
Apulia.north	12.0	0.0
Apulia.south	1.0	52.0
Calabria	5.0	11.0
Liguria.east	17.0	0.0
Liguria.west	20.0	0.0
Sicily	4.0	1.0
Umbria	9.0	0.0
Sardinia.coast	0.0	11.0
Sardinia.inland	0.0	29.0

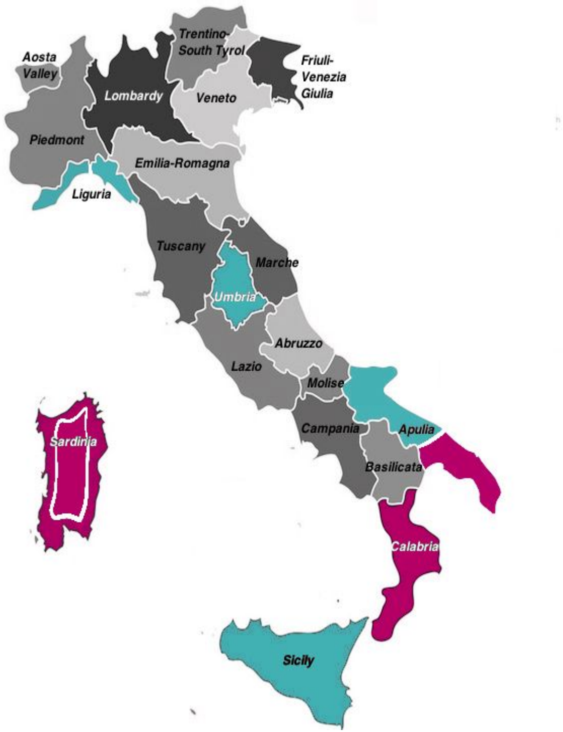
Checking for Accuracy: 2-Cluster Model (unscaled)

- **Cluster 0:** North Apulia, East/West Liguria, Sicily, Umbria
- **Cluster 1:** South Apulia, Calabria, Sardinia coast/inland
- **Accuracy score:** 95.93%



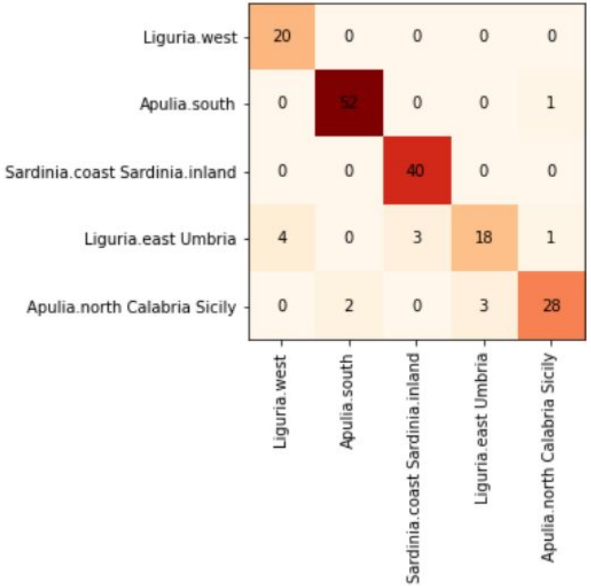
99%

91%



Checking for Accuracy: 5-Cluster Model (scaled)

- **Cluster 0:** West Liguria
- **Cluster 1:** North Apulia, Calabria, and Sicily
- **Cluster 2:** South Apulia
- **Cluster 3:** Sardina (coastal and inland)
- **Cluster 4:** East Liguria and Umbria
- **Accuracy score:** 91.86%



	Liguria.west	Apulia.south	Sardinia.coast Sardinia.inland	Liguria.east Umbria	Apulia.north Calabria Sicily
Liguria.west	20	0	0	0	0
Apulia.south	0	52	0	0	1
Sardinia.coast Sardinia.inland	0	0	40	0	0
Liguria.east Umbria	4	0	3	18	1
Apulia.north Calabria Sicily	0	2	0	3	28

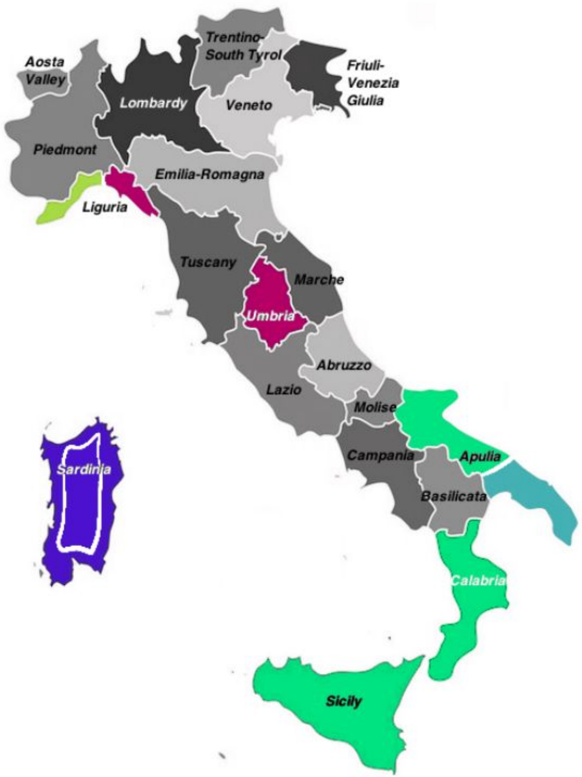
83%

96%

93%

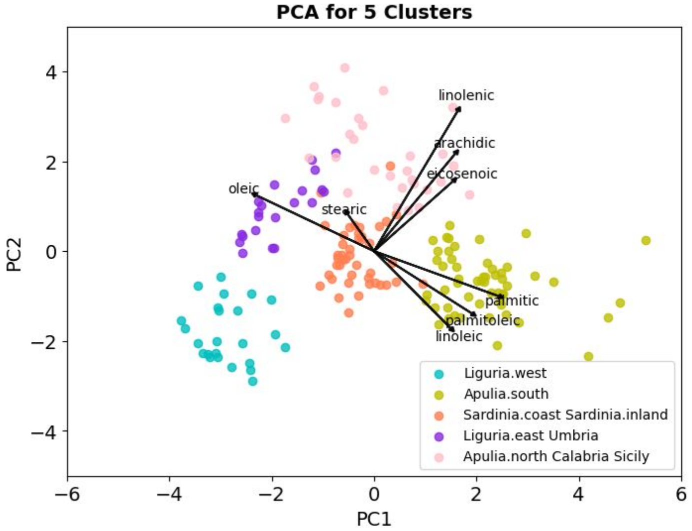
86%

93%



PCA Analysis on 5-Cluster Model

- Goal: Visualize the separation among the 5 clusters based on fatty acid components
    - Problem: Difficult to visualize clusters in 8 dimension
    - Solution: Principal Component Analysis (PCA)
        - Reduce dimension of data...8 → 2
        - Preserve as much variation in the data through PC1 and PC2



Conclusion

- Did not decide which model is the best
  - Had 2 “clusters of conclusion”
- Oleic is one of the most prominent fatty acid components
  - Influences the separation among the clusters the most

- Provides us insights on next steps
  - *García-Inza et al.* found that oleic acid concentration decreased linearly when the temperature increased from 16 to 32 degrees Celsius
  - Is there a connection between other environmental factors in each region and the chemical components in the olive oil?
    - Further investigation: season, weather, etc

Thank You!