Machine Learning for All: Class 7 Examples for Unsupervised Learning

César E. Montiel-Olea and Anastasiya Yarygina

Friday, March 8, 2019

Practical Excercises

- Clustering:
  - Group observations into similar clusters using K-means algorithm
  - Texts: topic models
- Data:
  - European Protein Consumption, R textbook
  - wine, R
  - we8there, R

K-means

- **Example 1:** Let's cluster Europe by food!

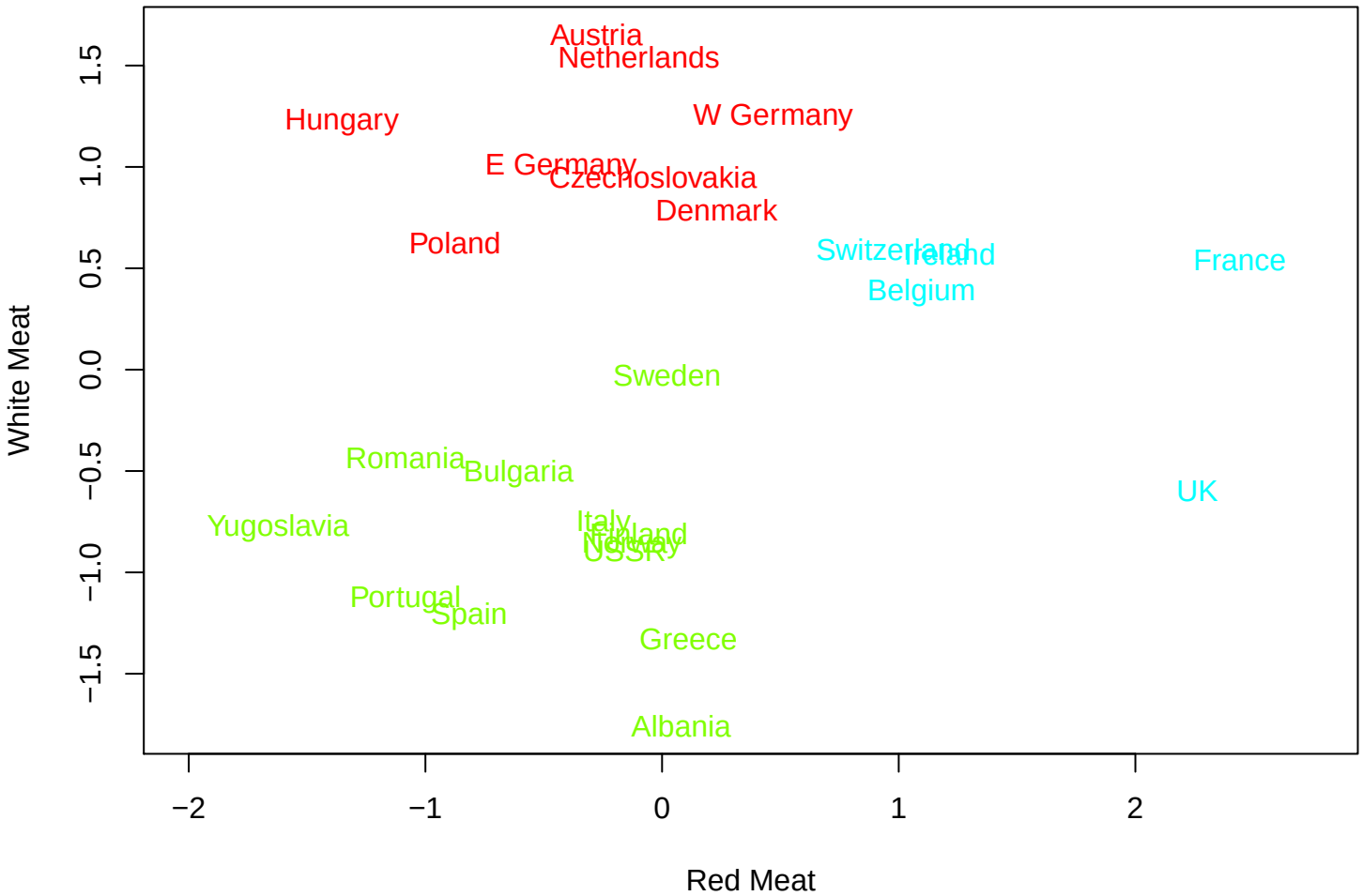
Protein consumption in Europe

- Dataset: consumption of proteins in grams per person per day.

```
##                RedMeat WhiteMeat Eggs Milk Fish Cereals Starch Nuts Fr.Veg
## Albania           10.1       1.4  0.5  8.9  0.2    42.3    0.6  5.5    1.7
## Austria            8.9      14.0  4.3 19.9  2.1    28.0    3.6  1.3    4.3
## Belgium           13.5       9.3  4.1 17.5  4.5    26.6    5.7  2.1    4.0
## Bulgaria           7.8       6.0  1.6  8.3  1.2    56.7    1.1  3.7    4.2
## Czechoslovakia     9.7      11.4  2.8 12.5  2.0    34.3    5.0  1.1    4.0
## Denmark           10.6      10.8  3.7 25.0  9.9    21.9    4.8  0.7    2.4
```

- In K-means scale matters -> **standardize X!**

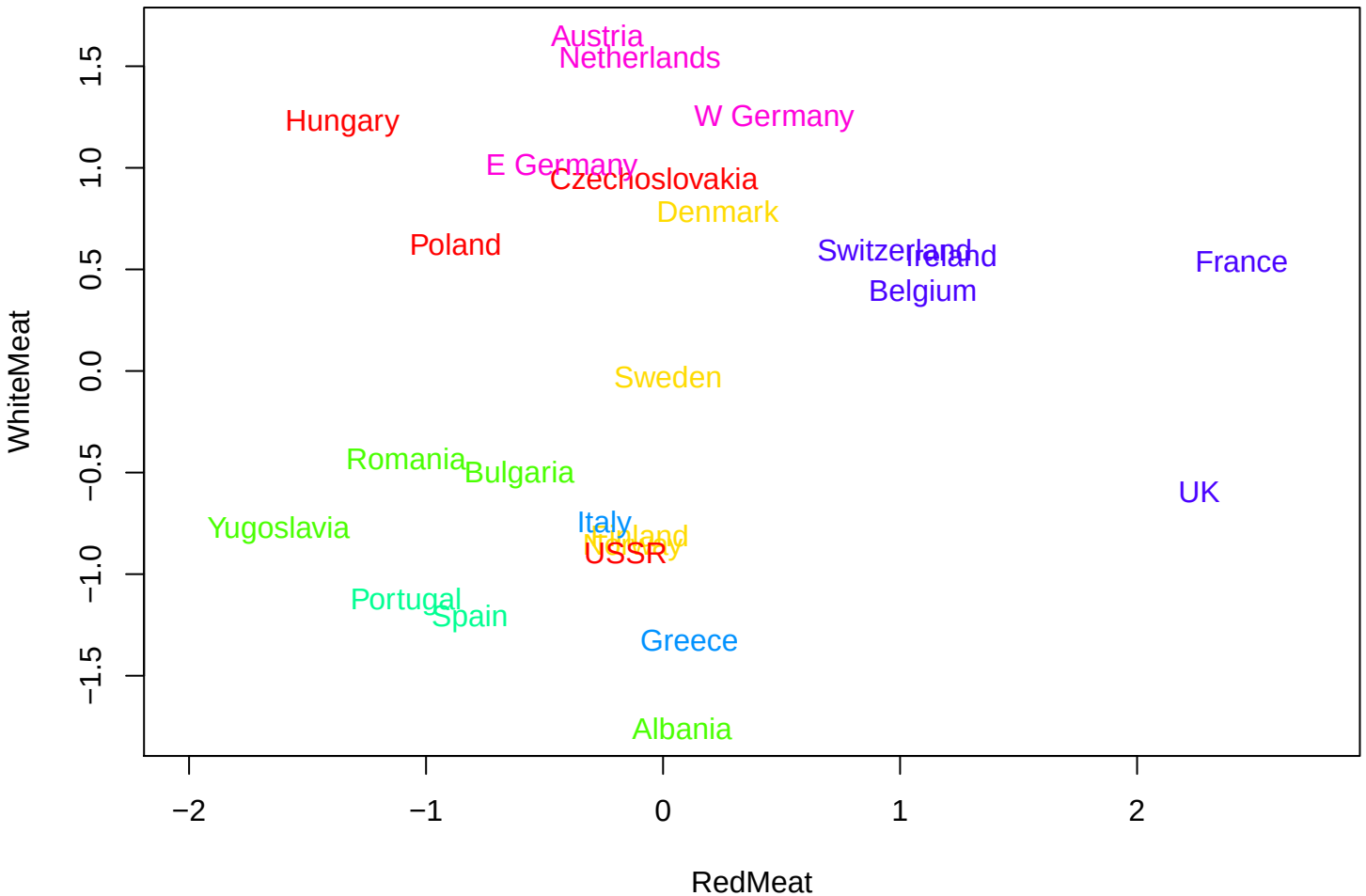
3-means clustering on Red vs. White meat consumption



- Consumption is in units of standard deviation from the mean.

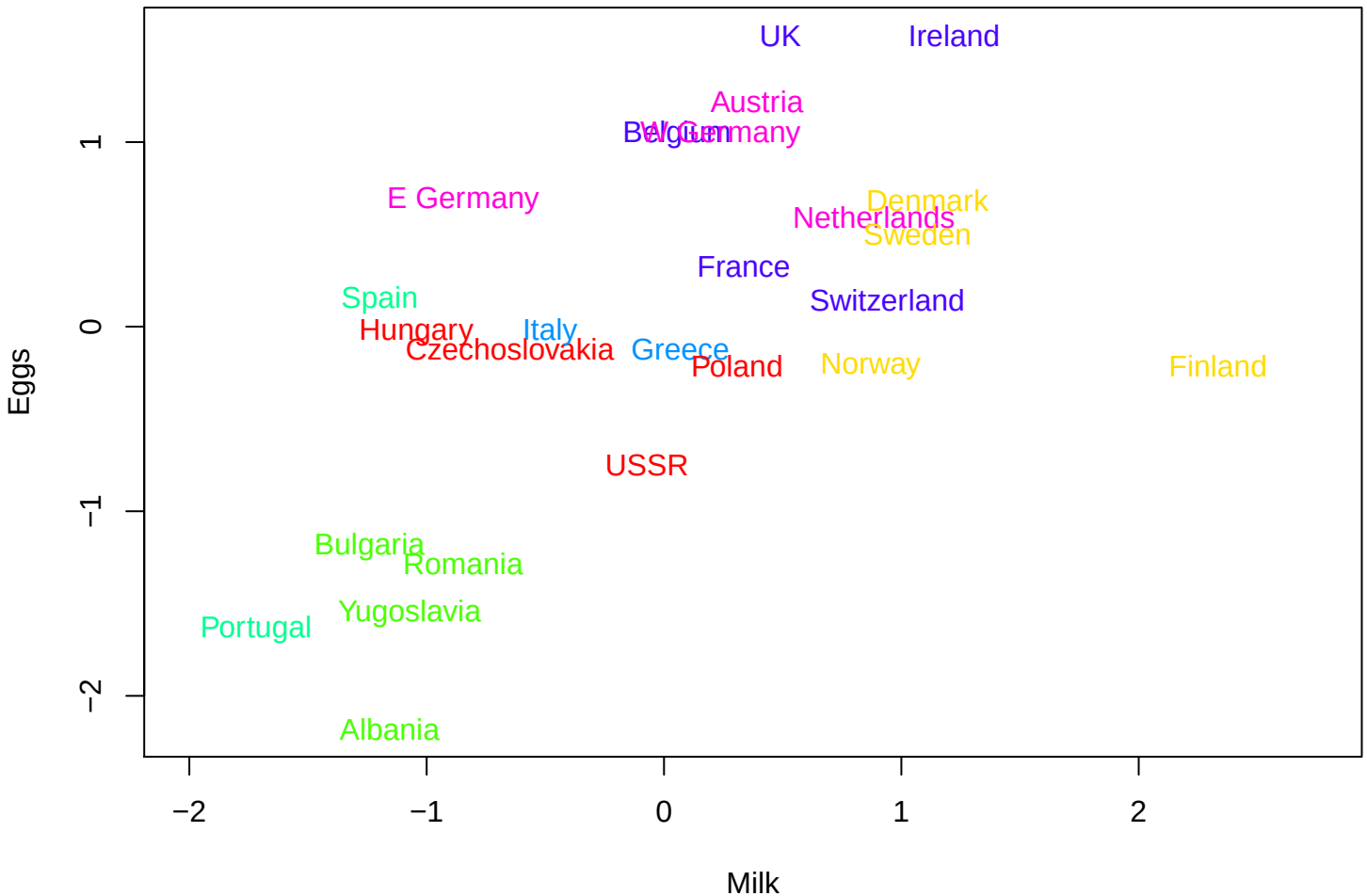
7-means clustering on all nine protein types

Consumption of White and Red meat



7-means clustering on all nine protein types

Consumption of Milk and Eggs



First takeaways

- K-means assigns in homogeneous groups
- Number of clusters **K** is **chosen by the analist**. How?
  - There are some methods and tools
  - But remember: **Clustering is an exploration** exercise
  - So, choose **K that makes sence**
- In K-means centroids are chosen by a random guess
  - Consequence: **Results are sensitive to the initial guess**
  - Solution: **run k-means several times**. **R will choose** the selection of **centroids** that yields the **lowest within cluster variation**
- **Example 2**: Wine dataset

The wine dataset

- Contains the results of chemical analysis of wines.

```
## 'data.frame':    6497 obs. of  13 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
##  $ color               : Factor w/ 2 levels "red","white": 1 1 1 1 1 1 1 1 1 1 ...
```

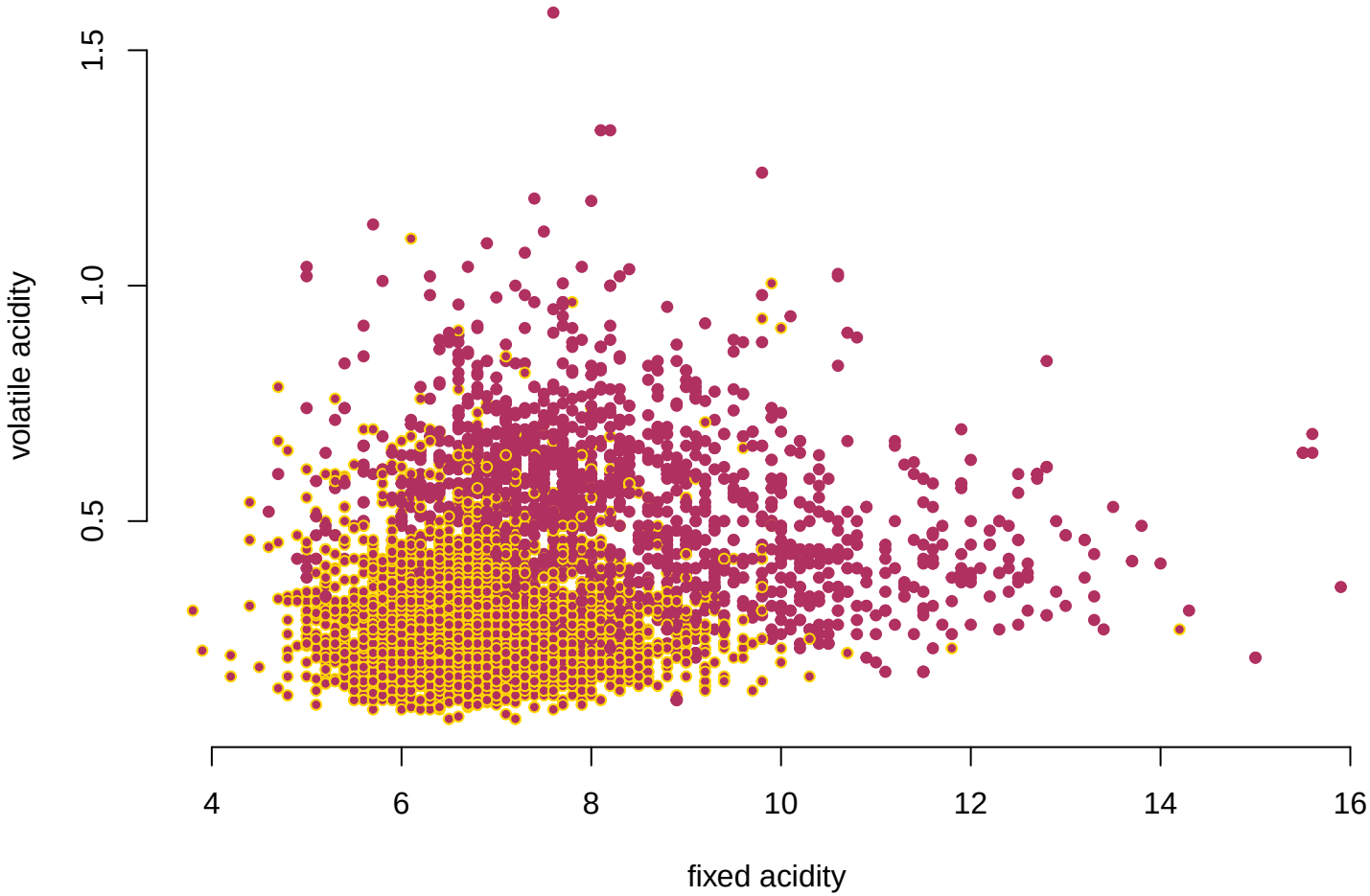
- In K-means scale matters -> **standardize X!**

2-means clustering on all features

What is the color distribution in each cluster?

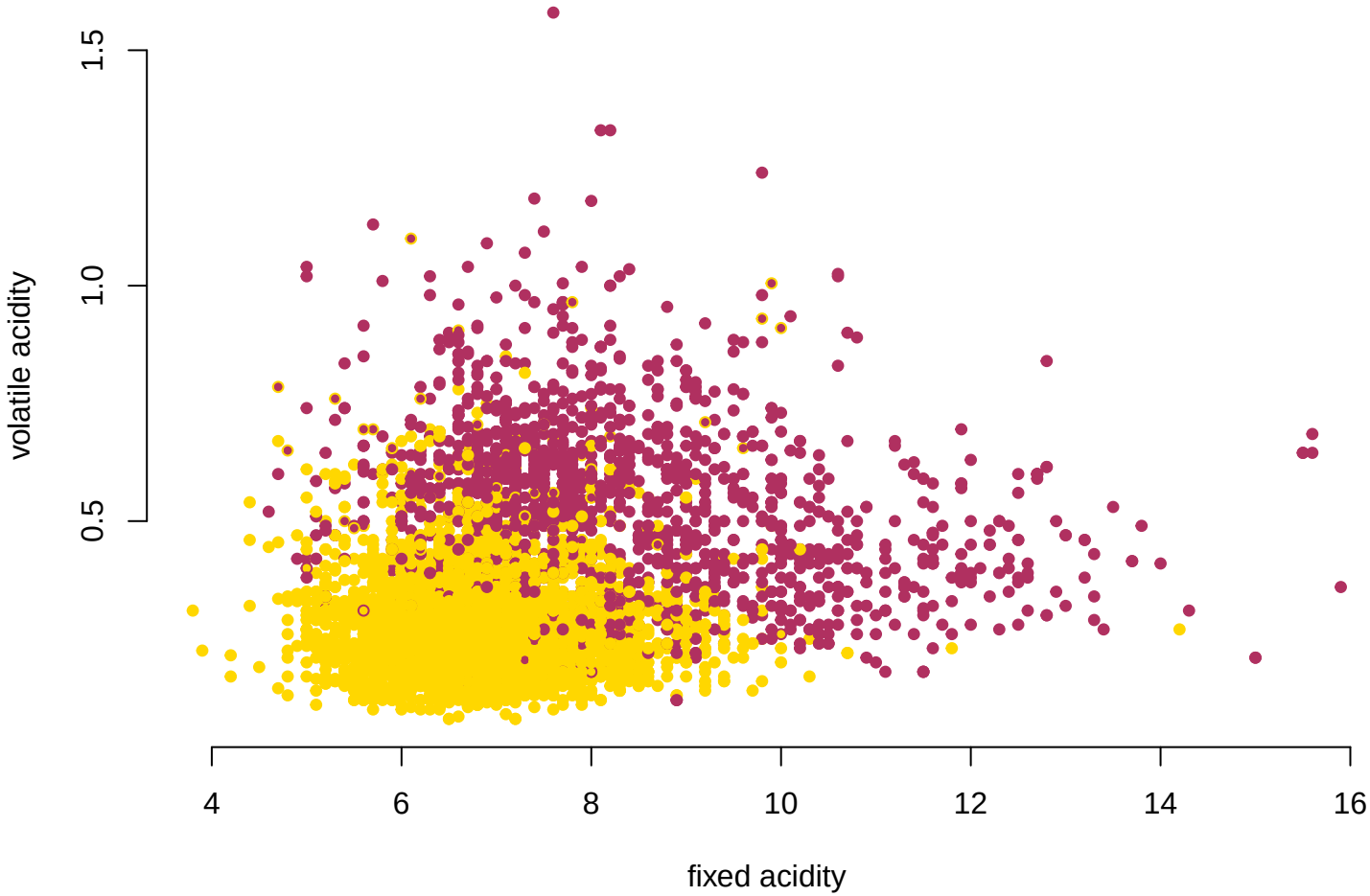
```
## $`1`
##
##   red white
##    24  4830
##
## $`2`
##
##   red white
##  1575    68
```

2-means clustering overlayed on wine color: **Oops**!



Point border is true color, body is cluster membership

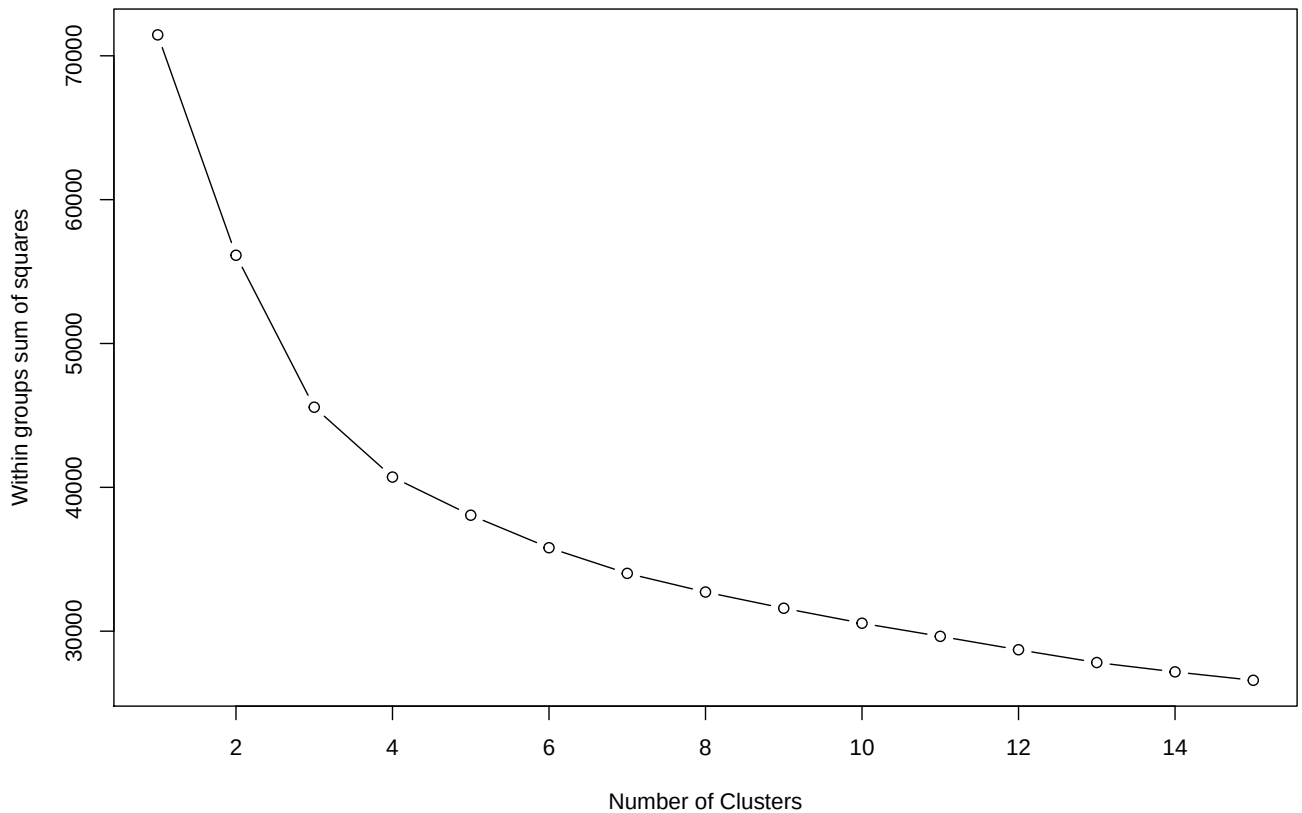
2-means clustering overlayed on wine color: **Nailed it**!



Point border is true color, body is cluster membership

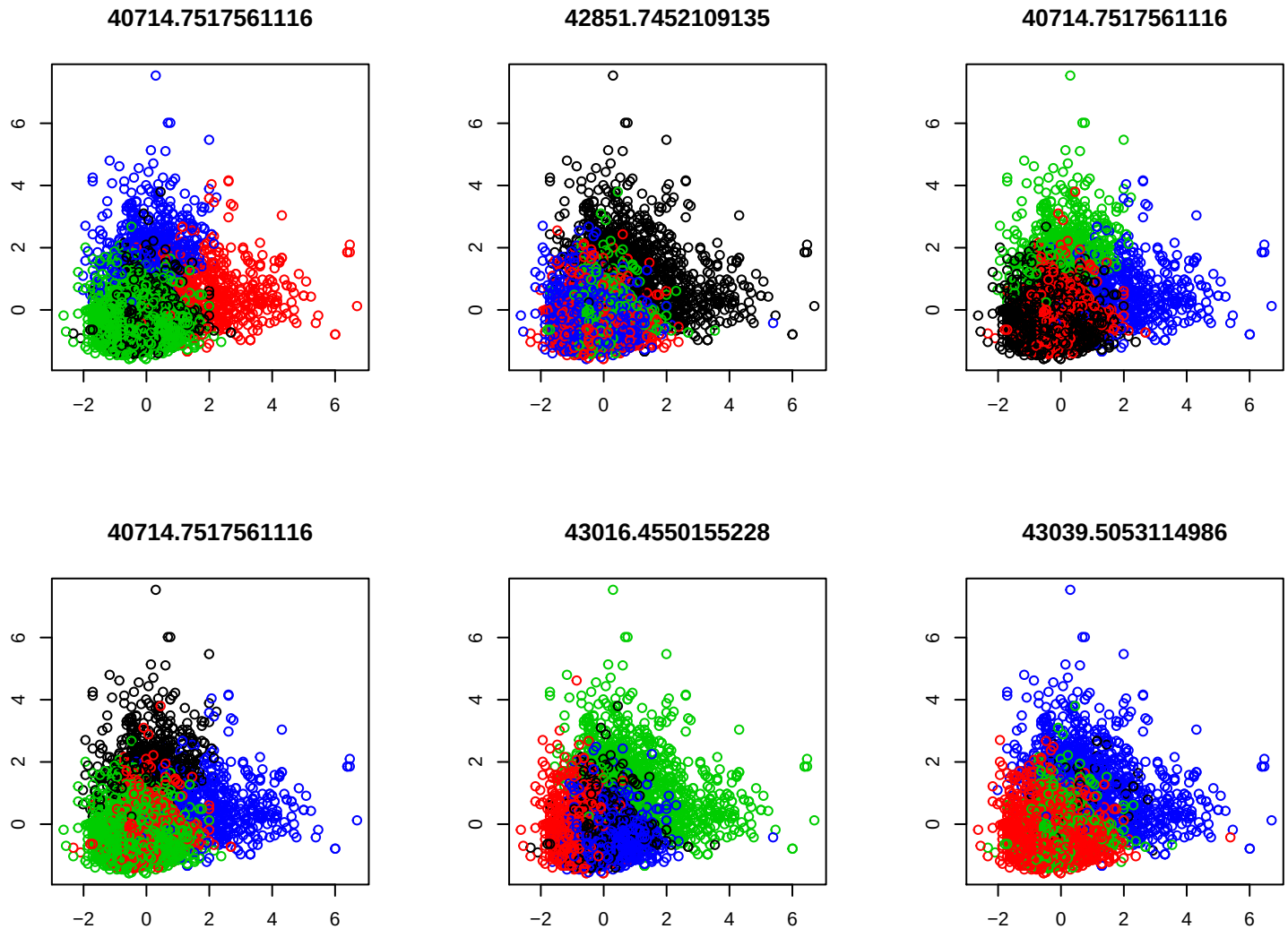
Choose the number of clusters: **Elbow method, scree plot**

- Choose the number of clusters so that adding another cluster does not improve within-group Sum of Squares much



Does the initial guess matter?

Cluster in 4 clusters 6 times, each time with different initial guess of centroids



Takeaways K-means clustering

- K-means clustering is a very simple and fast algorithm
- It deals well with Big Data
  - Exploratory analysis
  - Dimension reduction
- Key disadvantages:
  - It requires to pre-specify the number of clusters
  - Starting point affects the results
  - Sensitive to outliers
- Does it always make sense to cluster on distance to centroid?
  - Text analysis: counts of words

Restaurant review from *we8there*

- Counts of 2640 bigrams (pairs of words) from 6166 reviews
- with 5-star ratings on atmosphere, food, service, value and overall rating.

```
## 'data.frame':    6166 obs. of  5 variables:
##  $ Food      : num  5 5 5 5 5 5 5 5 5 5 ...
##  $ Service   : num  5 5 5 5 5 5 5 4 5 5 ...
##  $ Value     : num  5 5 4 5 5 5 5 5 5 5 ...
##  $ Atmosphere: num  5 5 4 5 4 5 5 5 5 5 ...
##  $ Overall   : num  5 5 5 5 5 5 5 5 5 5 ...
```

```
## Formal class 'dgCMatrix' [package "Matrix"] with 6 slots
##   ..@ i       : int [1:66459] 10 19 42 62 79 86 87 96 140 141 ...
##   ..@ p       : int [1:2641] 0 431 841 1100 1356 1623 1881 2084 2224 2457 ...
##   ..@ Dim     : int [1:2] 6166 2640
##   ..@ Dimnames:List of 2
##   .. ..$ Docs : chr [1:6166] "1" "2" "5" "11" ...
##   .. ..$ Terms: chr [1:2640] "veri good" "go back" "dine room" "dine experi" ...
##   ..@ x       : num [1:66459] 1 1 1 1 1 1 1 1 1 2 ...
##   ..@ factors : list()
```

Topic Modelling[1] using *maptpx* package

Idea: each **bigram is from a different topic**, and the **document is a mixture of topics**.

Exploratory analysis: fit a model with 5 topics.

```
##
## Estimating on a 6166 document collection.
## Fitting the 5 topic model.
## log posterior increase: 3259.5, 270.5, done.
```

```
##
## Top 5 phrases by topic-over-null term lift (and usage %):
##
## [1] 'came chip', 'toast bun', 'wasn whole', 'got littl', 'fri noth' (23.3)
## [2] 'good work', 'staff veri', 'food veri', 'excel place', 'restaur anyon' (22.7)
## [3] 'never bad', 'japanes restaur', 'wait go', 'alway great', 'out world' (18.5)
## [4] 'pm friday', 'select includ', 'seafood entre', 'highlight menu', 'enough share' (17.8)
## [5] 'mexican food', 'list extens', 'dine experi', 'italian food', 'great wine' (17.7)
##
## Dispersion = 7.53
```

[1] Also called LDA (Latent Dirichlet allocation)

Interpreting topics

Rank bigrams by probability within topics

```
##
## Estimating on a 6166 document collection.
## Fit and Bayes Factor Estimation for K = 5 ... 25
## log posterior increase: 3259.5, 270.5, done.
## log BF( 5 ) = 86639.7
## log posterior increase: 4936.9, 222.8, 65.5, done.
## log BF( 10 ) = 98207.27
## log posterior increase: 3633.6, 185.6, 53.7, done.
## log BF( 15 ) = 14897.81
## log posterior increase: 2216.4, 167.7, 48.9, 21.5, done.
## log BF( 20 ) = -59687.04


##
## Top 5 phrases by topic-over-null term lift (and usage %):
##
## [1] 'food veri', 'veri good', 'food excel', 'staff veri', 'veri nice' (13.2)
## [2] 'over minut', 'flag down', 'wait over', 'least minut', 'arriv after' (12.2)
## [3] 'great servic', 'alway great', 'servic alway', 'wait go', 'never bad' (11.4)
## [4] 'enough share', 'highlight menu', 'until pm', 'select includ', 'open daili' (10.5)
## [5] 'mexican food', 'italian food', 'authent mexican', 'list extens', 'food wonder' (10.4)
## [6] 'veri pleasant', 'indian food', 'thai food', 'again again', 'food delici' (9.3)
## [7] 'francisco bay', 'best kept', 'kept secret', 'just right', 'best steak' (9.1)
## [8] 'chicago style', 'carri out', 'great pizza', 'best bbq', 'onion ring' (8.4)
## [9] 'chees steak', 'food place', 'drive thru', 'york style', 'just anoth' (8.2)
## [10] 'over drink', 'wasn whole', 'got littl', 'took seat', 'took bite' (7.4)
##
## Log Bayes factor and estimated dispersion, by number of topics:
##
##              5       10       15        20
## logBF 86639.70 98207.27 14897.81 -59687.04
## Disp      7.53     5.22     4.17      3.47
##
```

Visualization using ***wordcloud*** package