



Kx for Wine Tasting

Machine Learning in q/kdb+

Mark Lefevre

Algorithmic Quantitative Analyst

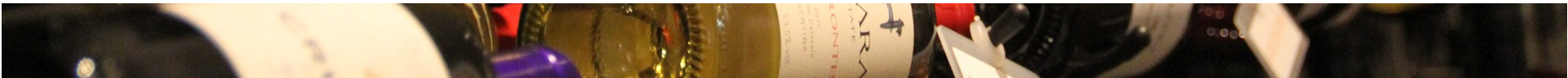






Machine Learning Introduction

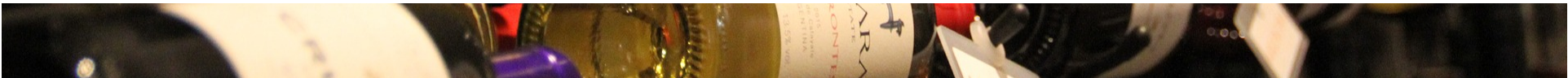
- ML algorithms can be grouped by learning style
 - Supervised Learning
 - Unsupervised Learning
 - Reinforcement Learning
- Or, alternatively, by similarity
 - Regression
 - Clustering
 - Classification
 - Neural Networks
 - Etc.





Unsupervised Learning

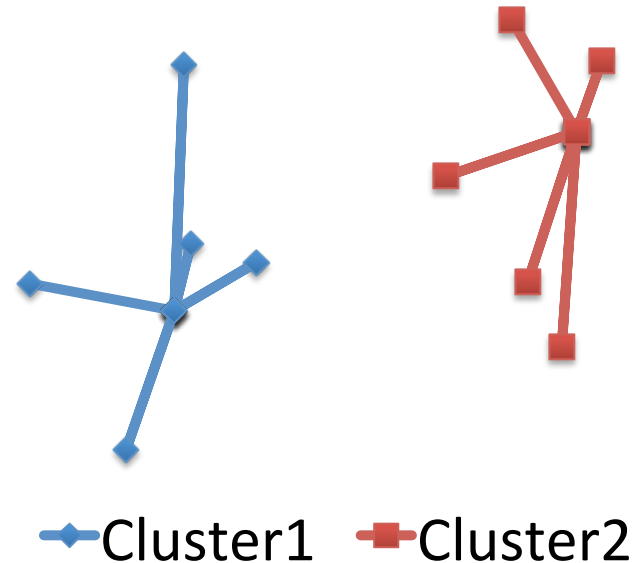
- Uses a dataset with known inputs and unlabeled outputs
 - In a true application, it is impossible to evaluate the accuracy of the algorithm's output
- Infers a function to describe a transformation
- Typical types of problems are classification, **clustering**, anomaly/fraud detection, image processing and topic modeling



K-Means Clustering Algorithm

- Given n d -dimensional data points $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, partition the n observations into k ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, \dots, S_k\}$ that minimize a within-cluster distance measure
- Using a Euclidean distance measure (L^2 -norm)

$$\arg \min_S \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2$$





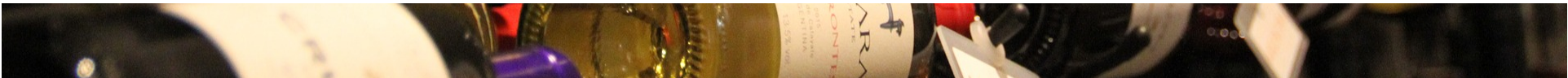
Lloyds Algorithm

- A simple, useful heuristic algorithm is widely used often called Lloyds Algorithm

0. Initialize centroids

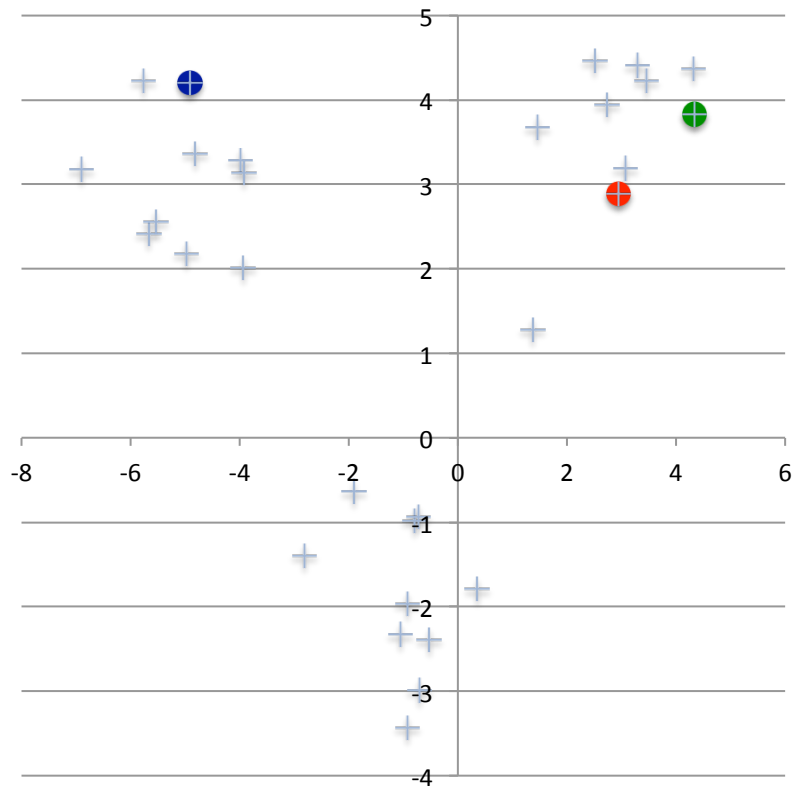
Iterate the following two steps until convergence

1. Assign data points to nearest cluster
2. Calculate new centroids

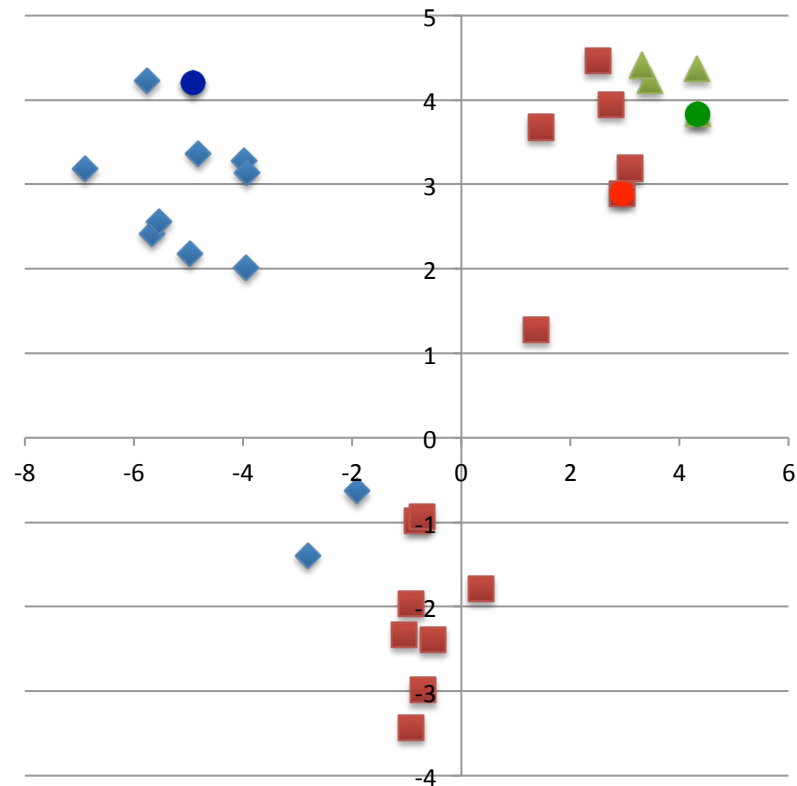


Simple Example (k=3)

0. Initialize 3 Centroids

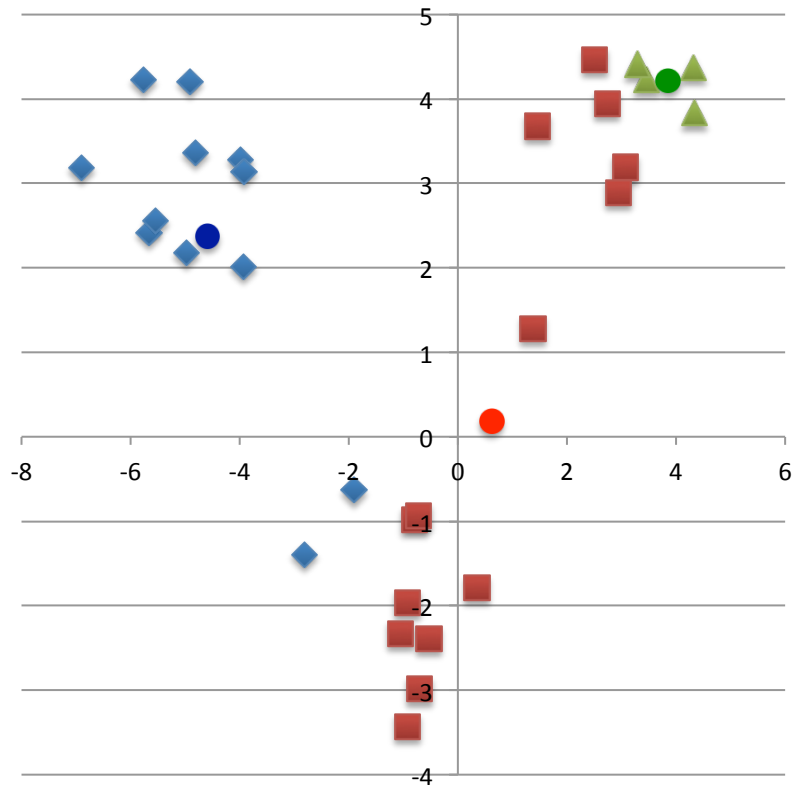


1. Cluster Assignment

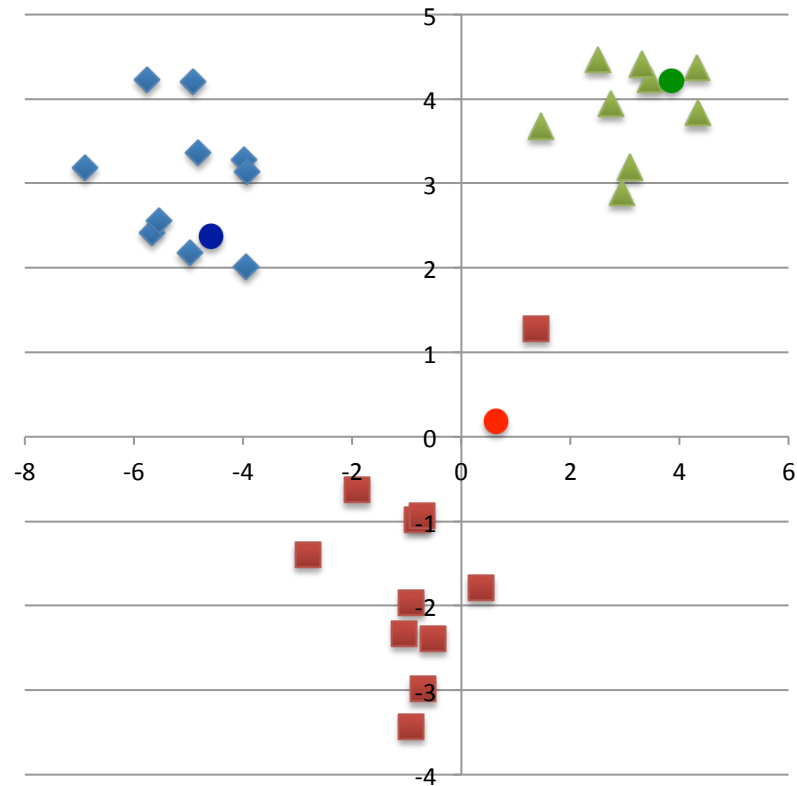


Simple Example

2. Calculate New Centroids

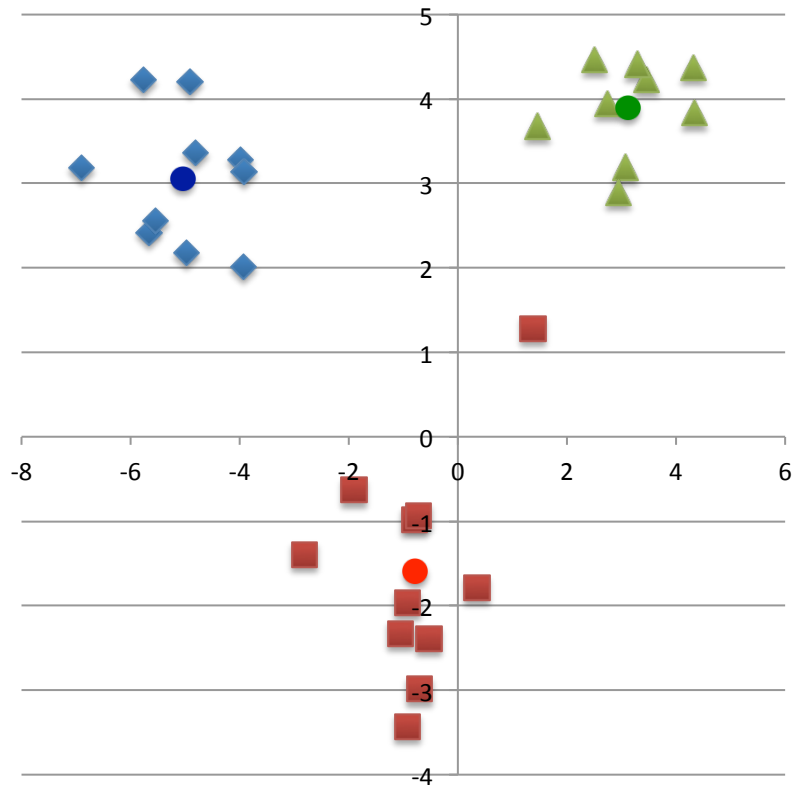


3. Cluster Assignment

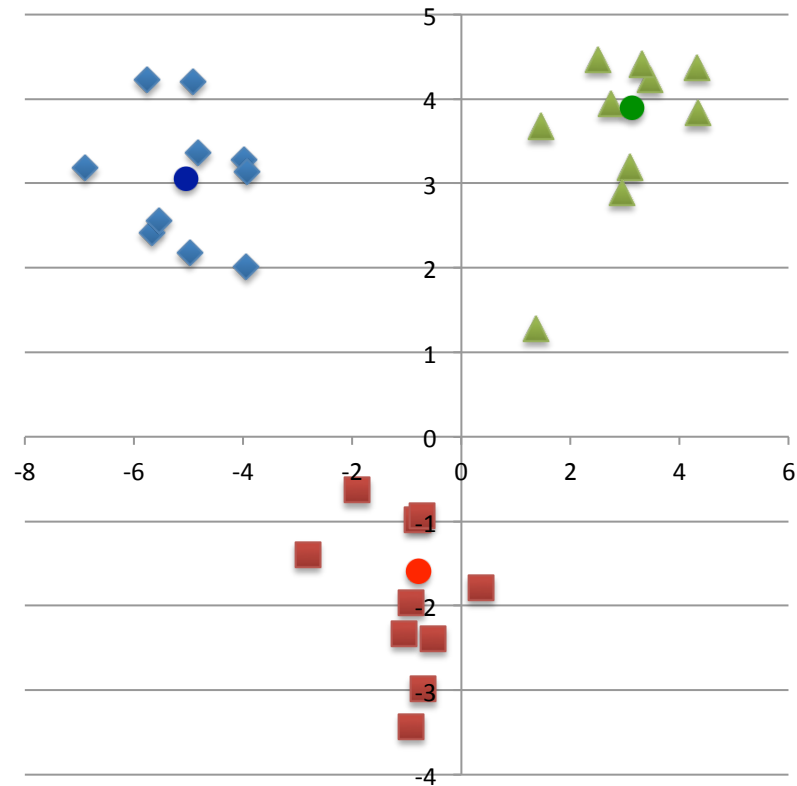


Simple Example

4. Calculate New Centroids



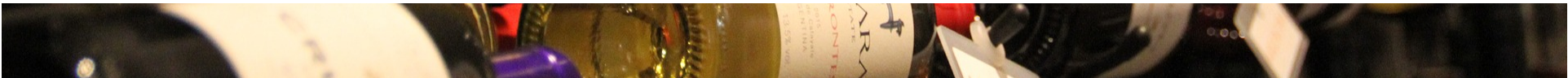
5. Cluster Assignment





Wine Dataset

- UCI Machine Learning Repository
 - <http://archive.ics.uci.edu/ml>
 - Irvine, CA: University of California, School of Information and Computer Science.
 - Consists of 178 instances, 13 chemical analysis attributes and a column indicating the actual class
1. Alcohol
 2. Malic acid
 3. Ash
 4. Alcalinity of ash
 5. Magnesium
 6. Total phenols
 7. Flavanoids
 8. Nonflavanoid phenols
 9. Proanthocyanins
 10. Color intensity
 11. Hue
 12. OD280/OD315
 13. Proline

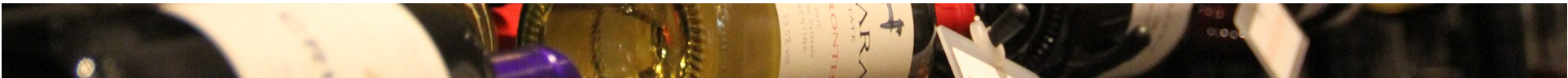




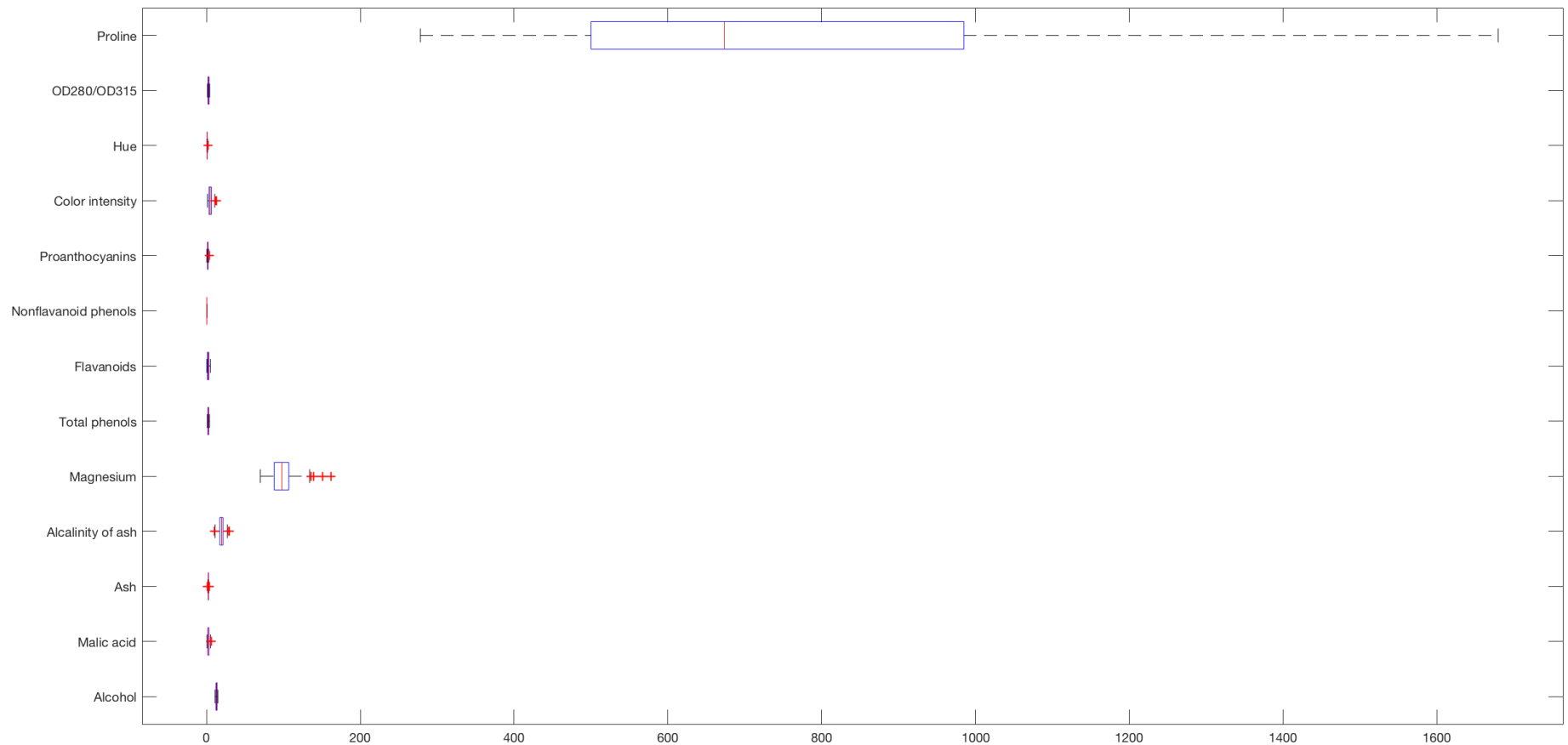
Quick Look at Raw Wine Dataset

- Here are 9 samples, 3 from each class
- What do you notice about the data?
- Could you find a pattern to distinguish the 3 cultivars from each other?

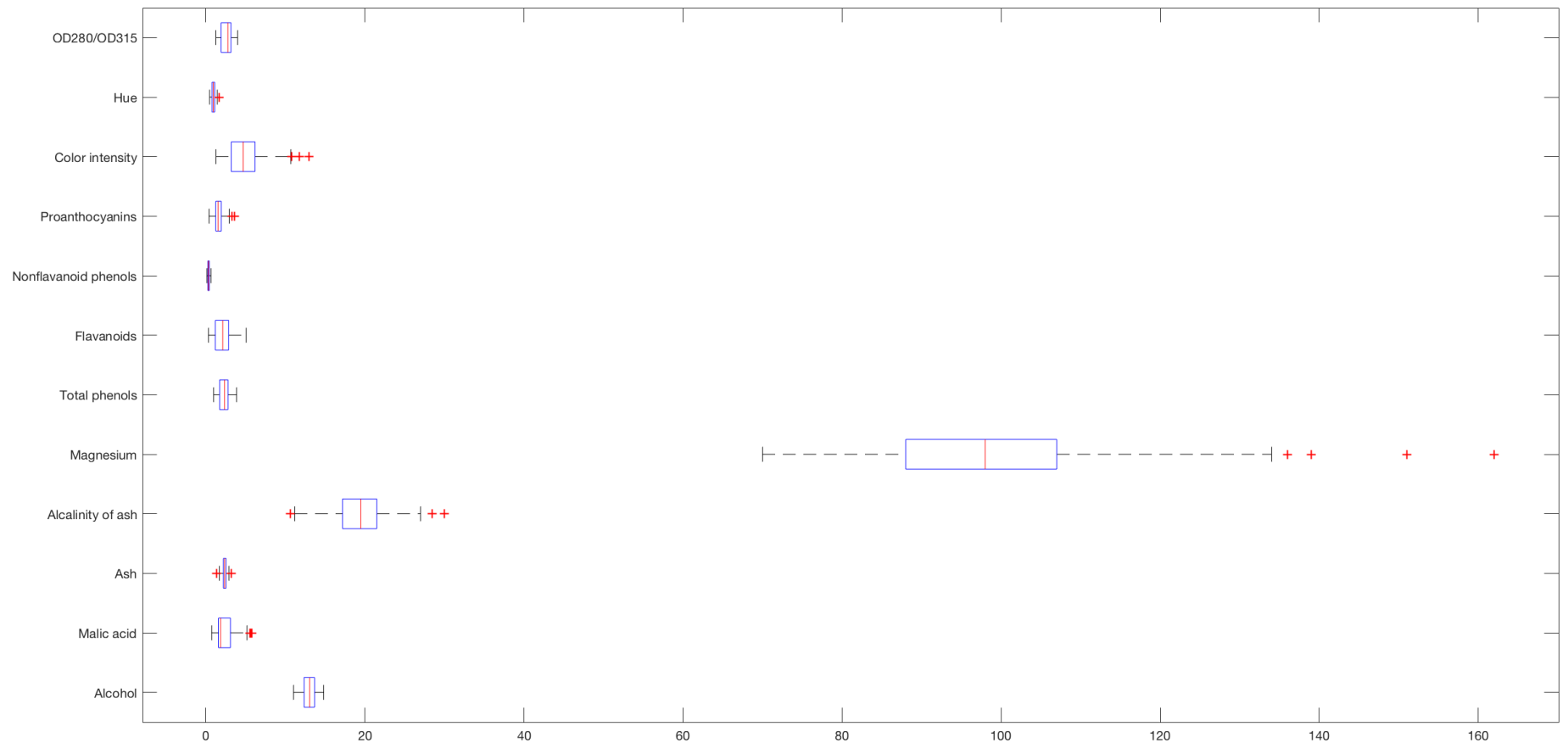
1	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065
1	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050
1	13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185
2	11.79	2.13	2.78	28.5	92	2.13	2.24	0.58	1.76	3	0.97	2.44	466
2	12.37	1.63	2.3	24.5	88	2.22	2.45	0.4	1.9	2.12	0.89	2.78	342
2	12.04	4.3	2.38	22	80	2.1	1.75	0.42	1.35	2.6	0.79	2.57	580
3	12.86	1.35	2.32	18	122	1.51	1.25	0.21	0.94	4.1	0.76	1.29	630
3	12.88	2.99	2.4	20	104	1.3	1.22	0.24	0.83	5.4	0.74	1.42	530
3	12.81	2.31	2.4	24	98	1.15	1.09	0.27	0.83	5.7	0.66	1.36	560



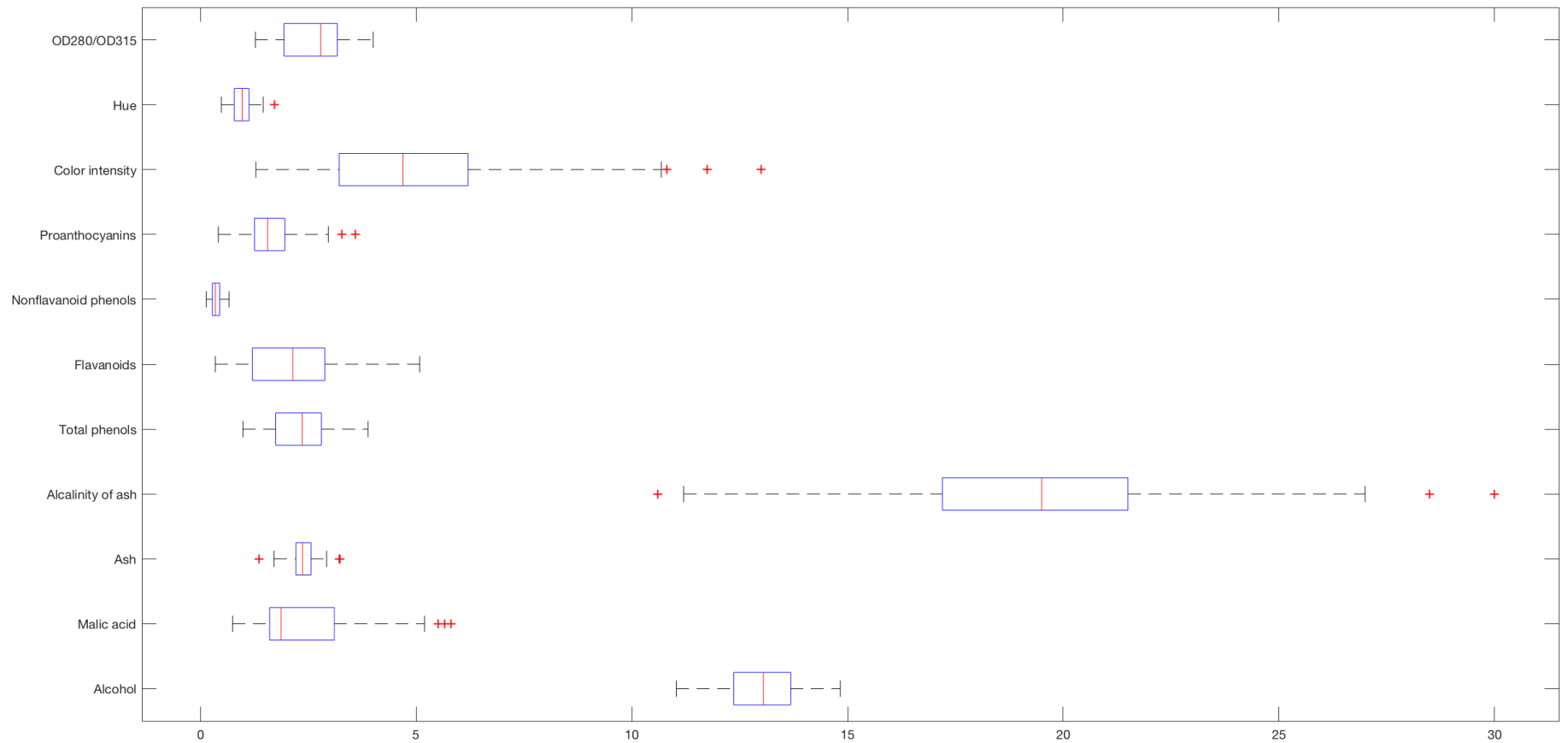
Wine Dataset Boxplots (All features)



Wine Dataset Boxplots (-Proline)

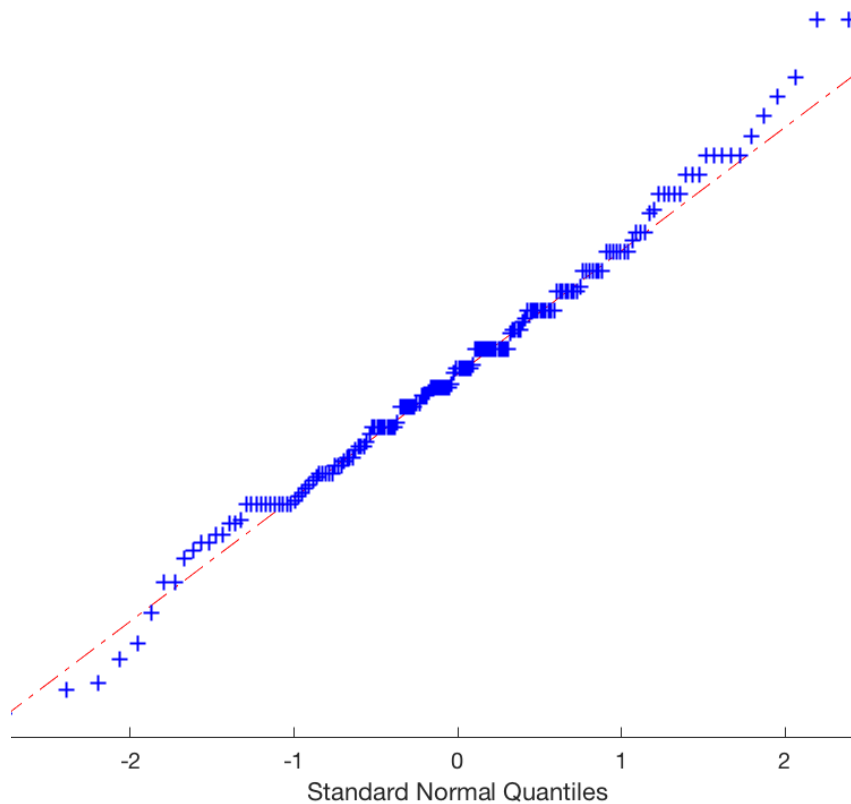


Wine Dataset Boxplots (-Proline, -Magnesium)

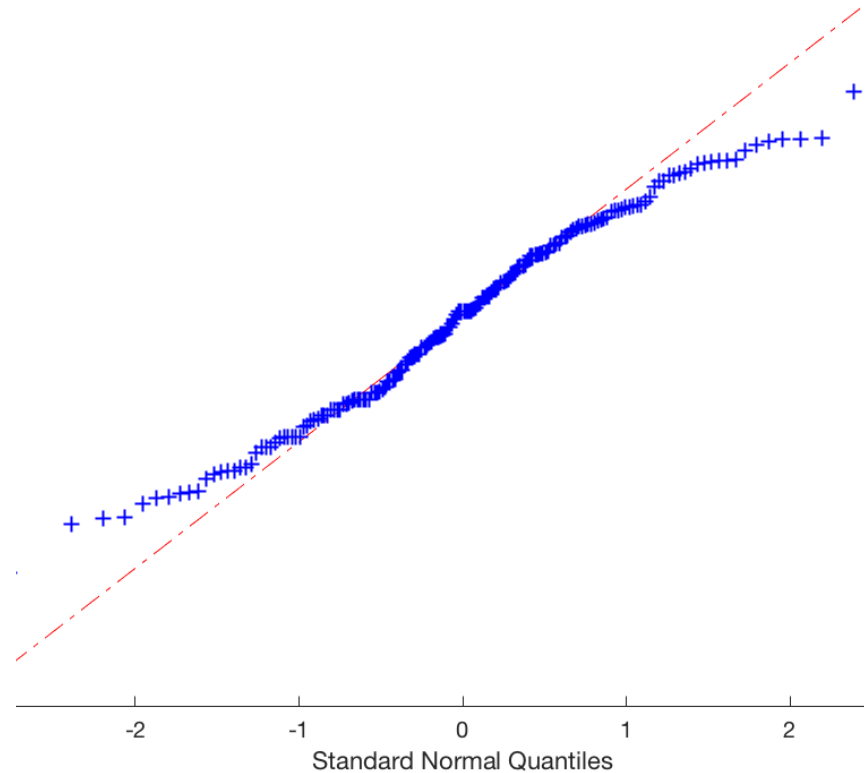


Alcohol and Malic Acid QQ Plots

QQ Plot of Sample Data versus Standard Normal



QQ Plot of Sample Data versus Standard Normal

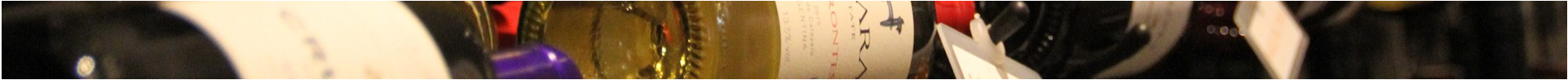




Q Code

```
// Demonstration implementing k-means algorithm/Lloyds algorithm
wds:flip (`$'14#.Q.A)!("J",13#"F";",") 0: `:wine.csv;
actualGroup:wds[`A];
/X:delete A from X;
wds:update g:178?3 from wds;

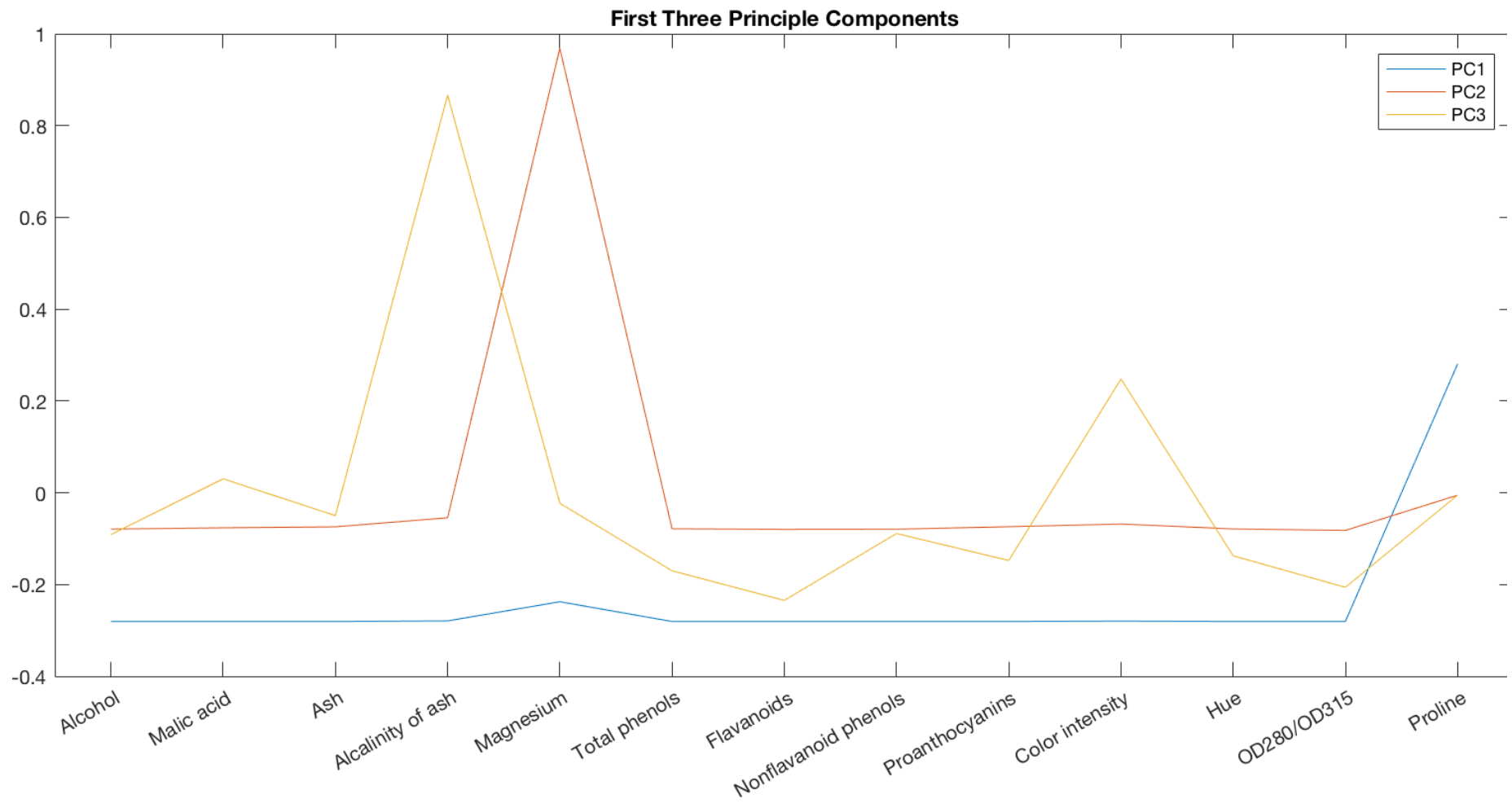
f:[X]
  // Lambda Function to find centroids by group (column name=g)
  C:{{t;b;ac;f} ?[t();b;ac!f,/:ac]} [X;{x!x} raze `g;(cols X) except `g;avg];
  // Group assignments
  newg:{{x?min x}x$x} each
    (raze each delete g from X)-/:\\:(raze each value C);
  update g:newg from X
};
wds:(f/)wds;
```



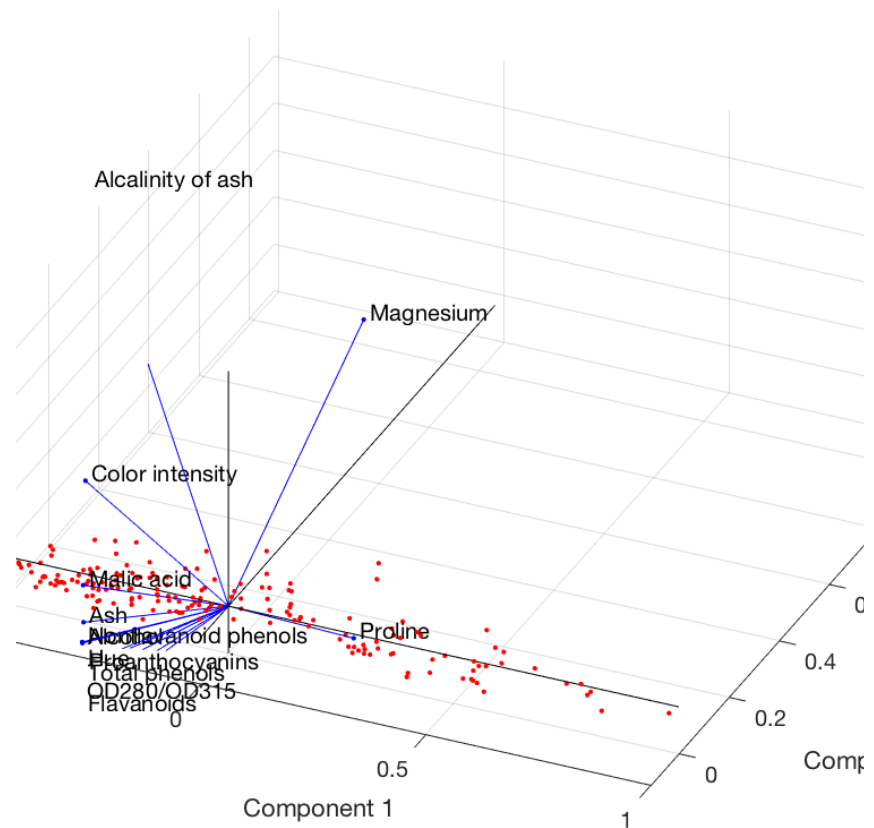
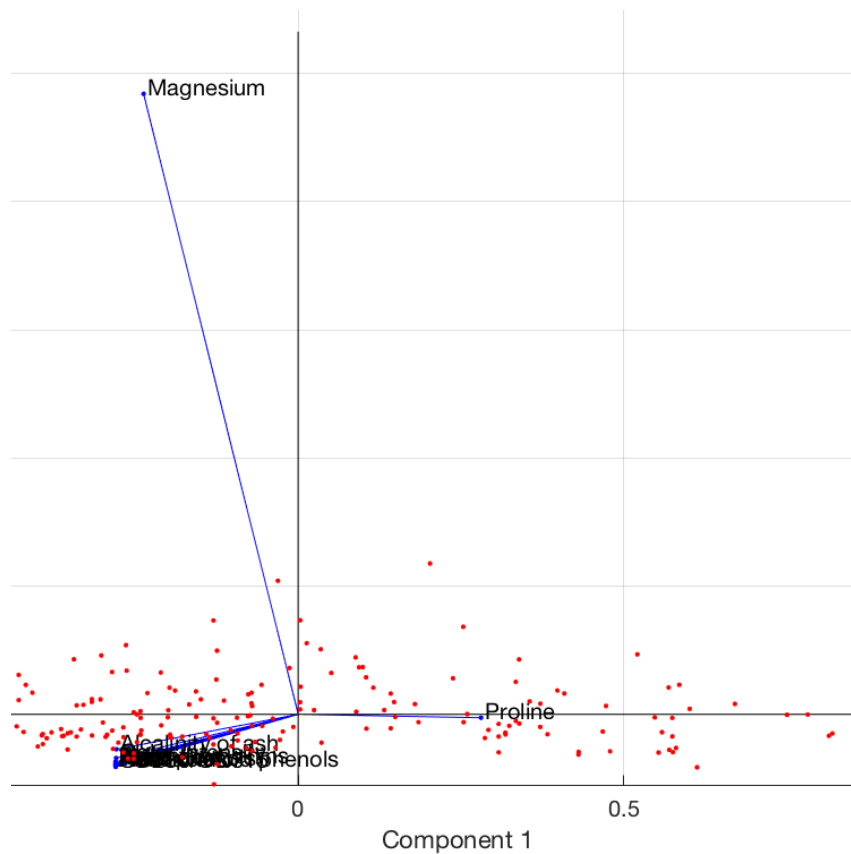
Principal Component Analysis (PCA)

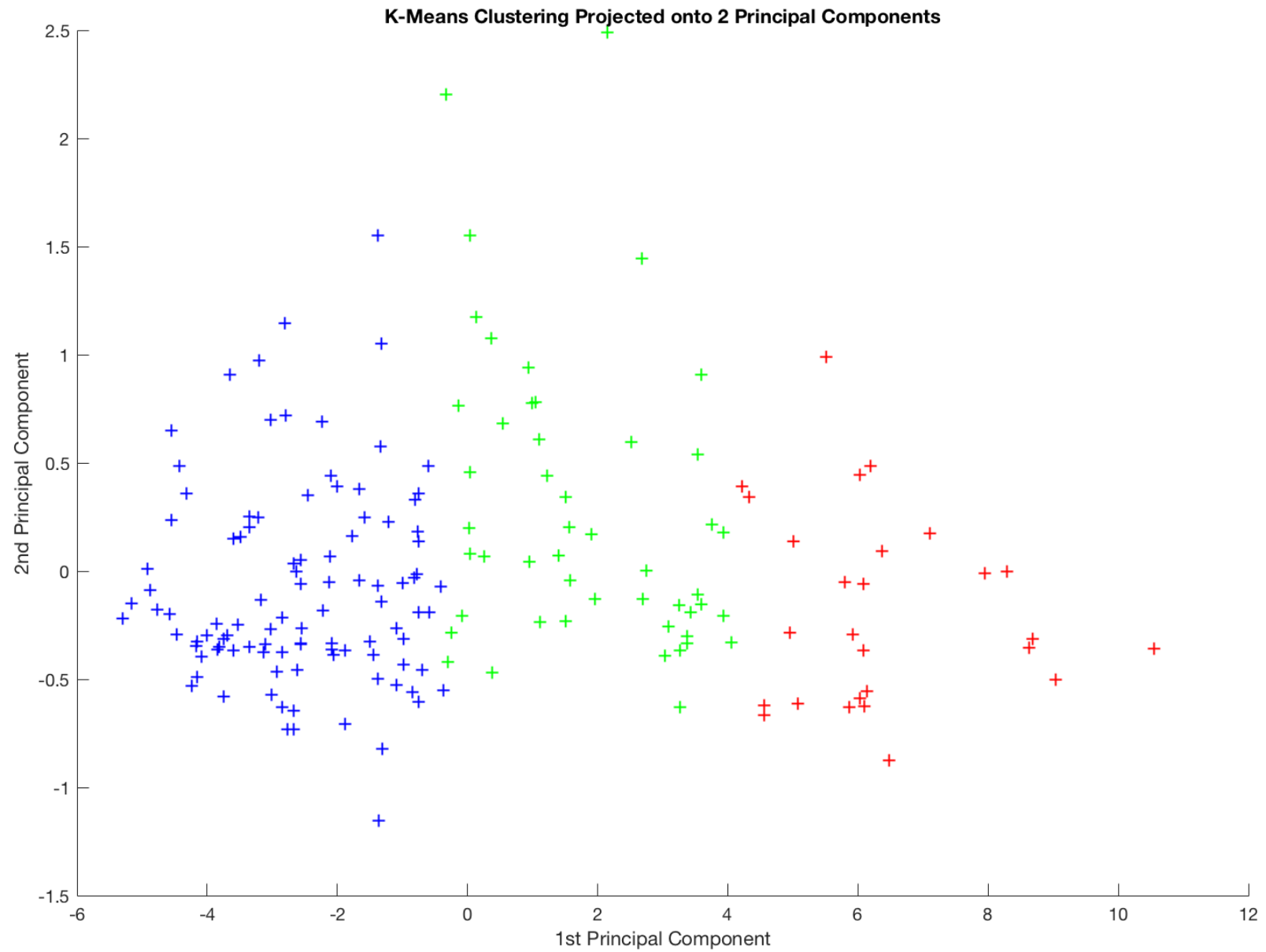
- PCA is a statistical procedure that utilizes orthogonal transformations to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called **principal components**.
- In a word, ***decorrelation***
- The principal components are the eigenvectors of a symmetric variance-covariance matrix
- Eigenvectors are ordered by their corresponding eigenvalues
 - Amount of variance explained by the component
- Taking a few of principal components, we can achieve
 - **dimensionality reduction**
- This is very useful for high dimensionality problems, such as ~~instantaneous forward curve evolutions~~ analyzing wine

Principle Components



Visualization of Wine Data Using Principal Component Analysis



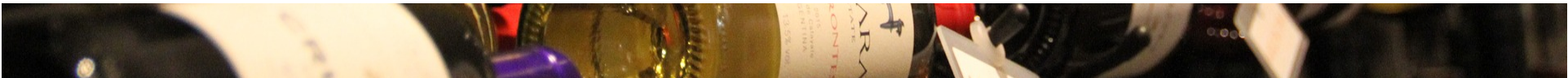




Another Look at Raw Wine Dataset

- Here are 9 samples, 3 from each class
- Can you find a pattern to distinguish the 3 cultivars, *if you knew the principle components?*

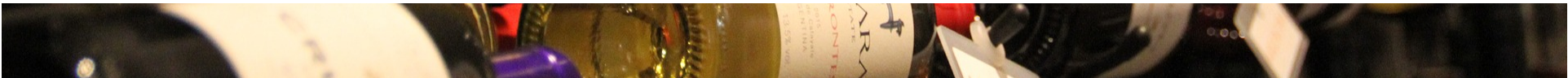
1	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065
1	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050
1	13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185
2	11.79	2.13	2.78	28.5	92	2.13	2.24	0.58	1.76	3	0.97	2.44	466
2	12.37	1.63	2.3	24.5	88	2.22	2.45	0.4	1.9	2.12	0.89	2.78	342
2	12.04	4.3	2.38	22	80	2.1	1.75	0.42	1.35	2.6	0.79	2.57	580
3	12.86	1.35	2.32	18	122	1.51	1.25	0.21	0.94	4.1	0.76	1.29	630
3	12.88	2.99	2.4	20	104	1.3	1.22	0.24	0.83	5.4	0.74	1.42	530
3	12.81	2.31	2.4	24	98	1.15	1.09	0.27	0.83	5.7	0.66	1.36	560





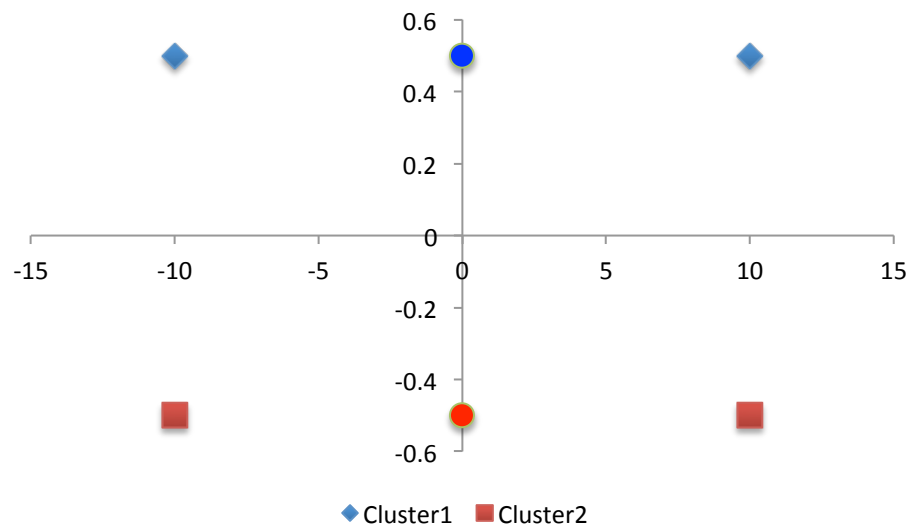
Weaknesses of K-Means

- K is an input
- Sensitivity to initialization
 - Multiple runs with different random initializations
 - Kmeans++
- Empty clusters
 - Delete cluster
 - Randomly chose another centroid
- Hyperspherical clusters
 - Cannot handle non globular clusters well
- Outliers
 - K-medians algorithm
- No guarantee it will converge to global optimum
 - NP-hard



K-means++

- Improved initialization algorithm
- Addresses potentially bad initial guesses

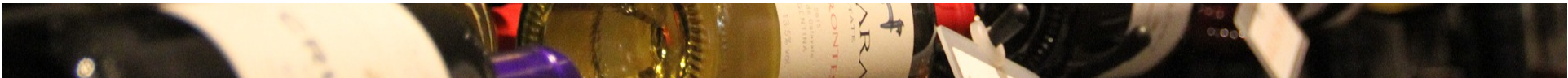


1. Choose random data point as first centroid
2. Compute distance, $D(x)$, from initial random point to all other data points
3. Choose a new centroid from those data points using a weighted probability distribution proportional to $D(x)^2$
4. Repeat Steps 2 and 3 until k centers have been chosen



Conclusion

- Briefly introduced machine learning, the concept of unsupervised learning and the k-means algorithm
- Showed how this algorithm can be easily written in q and can be used to learn how to categorize wine cultivars
- Hopefully, this has provide an interesting look at the opportunities to utilize q/kdb+ in machine learning





About Me

- Mark is currently consulting at one of the largest banks in Tokyo as an algorithmic quantitative analyst developing high-performance algorithmic trading systems on the e-FX desk. Prior to moving to Japan, he worked in London for Unicredit on the Equity-Linked Origination desk creating convertible bonds for European corporates, consulted in the US on e-commerce analytics and worked for several high-tech software companies.
- Earlier in his career, he worked for Mitsubishi Semiconductor America designing semiconductors and a startup developing a DSP. He then moved into applications engineering for an Electronic Design Automation (EDA) company and, subsequently, internet software companies in CA and Europe.
- Mark has a bachelors degree in Electrical Engineering and Computer Science from Duke University, a masters degree in Computer Engineering from North Carolina State University and an MBA in Quantitative Finance from the Wharton School of Business. He recently completed a Certificate in Quantitative Finance (CQF).
- He dreams of the day when he can create software without encountering a single type error

