今回の実験では基礎実験の課題として「都道府県別アルコール類の消費量」「あやめのデータ」「アワビのデータ」を選択した。

課題1 「都道府県別アルコール類の消費量」

　まず散布図行列について調べる。

　図1-1を見ると一部の県では非常に多くの酒類が飲まれていることがわかる。また、全ての散布図において相関係数が高く見える。

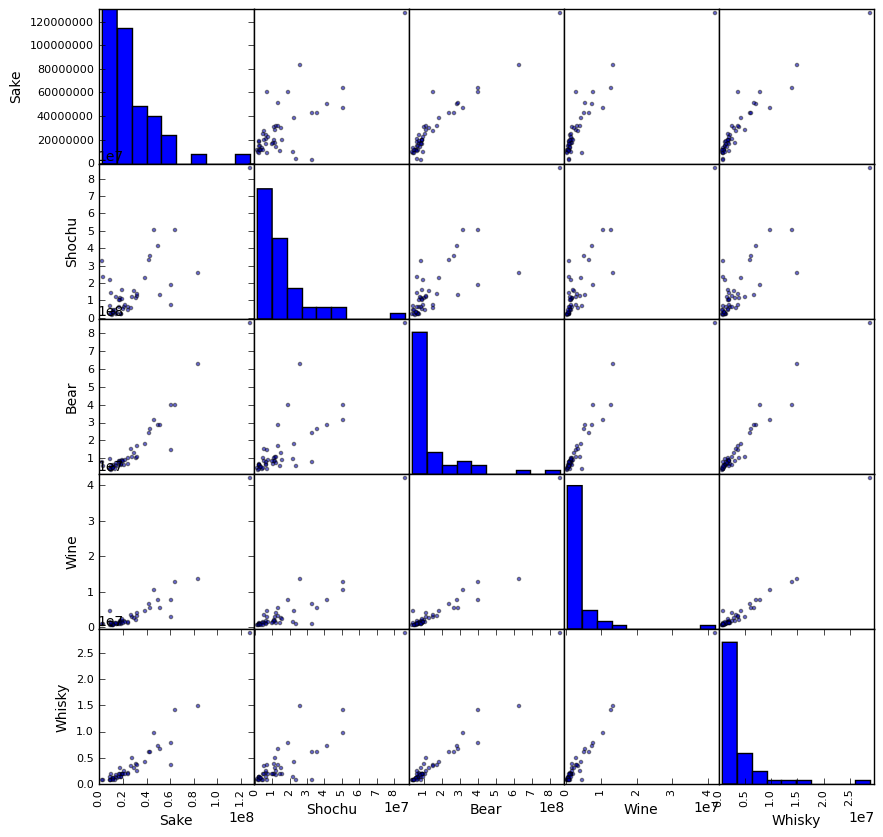


図 -1 アルコール消費の散布図行列

　続いて相関係数について調べる。

Sake Shochu Bear Wine Whisky

Sake 1.000000 0.738865 0.952417 0.883041 0.940801

Shochu 0.738865 1.000000 0.813708 0.844892 0.849852

Bear 0.952417 0.813708 1.000000 0.918850 0.969909

Wine 0.883041 0.844892 0.918850 1.000000 0.965348

Whisky 0.940801 0.849852 0.969909 0.965348 1.000000

図1-2　アルコール消費量の相関係数

　これを見ると全て正の相関を持ち、全て0.7よりも高いことがわかる。この相関係数より、どの県でも酒を飲む人は、これら全ての種類の酒を飲むことがわかる。

　続いてPCAによる分析

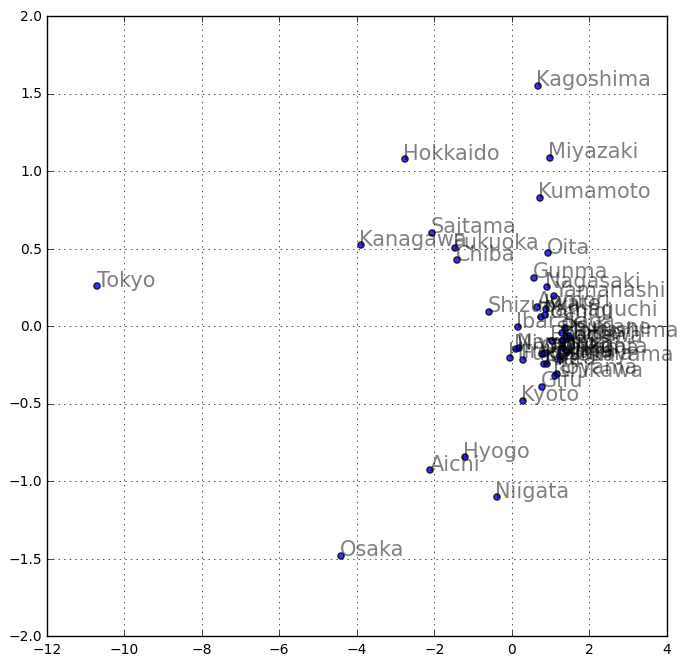


図1-3 主成分分析の表(第一成分,第二成分)

首都圏と大阪福岡などの大都市が塊から外れていることがわかる。累積寄与度を見ると、

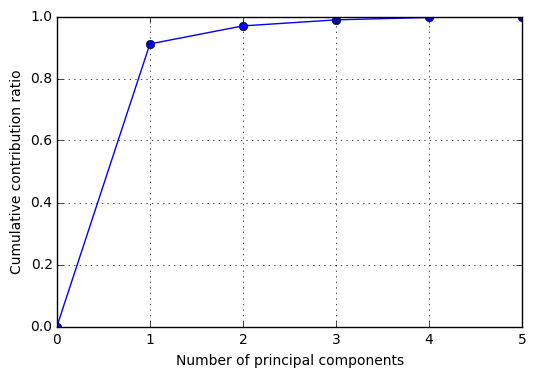


図 1-4 主成分分析の累積寄与度

これを見ると第一成分と第二成分がほぼ全ての割合を占めていることがわかる。第一成分と第二成分がほぼ全ての割合を占めていることがわかる。つまり都道府県別に見たときに都道府県の特徴はだいたい2種類あるということになる。

第一成分は大都市が比較的外れており、東京が大きく外れていることから酒類の総消費量であると予測できる。(E=A+B+C+Dみたいな軸をとっている)

第二成分は鹿児島大阪が外れており、東京がほぼゼロに近いことから、酒類の総消費量に対する日本酒の割合であることが予測できる。

さらに階層的クラスタリングを行う。

この図より大まかに3つのグループに分けることができた。

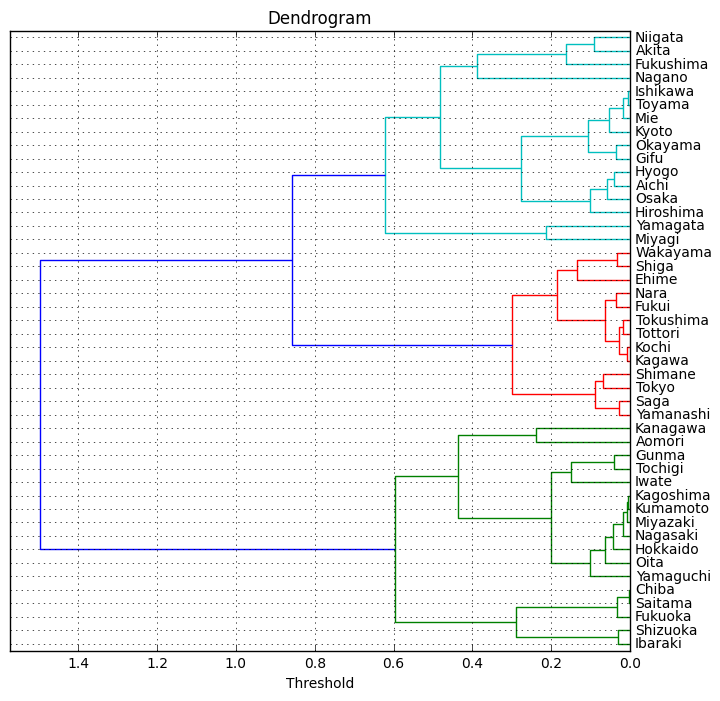


　　　図1-5 階層的クラスタリング

課題2　あやめのデータ

　まず種族ごとに色分けした行列散布図を見る。

これを見ると色分けした3種類は比較的綺麗に分けることができ、Petal.WidthとPetal.Lengthには非常に強い正の相関がありそうということがわかる。

　続いて相関係数について確認する。

Sepal.Length Sepal.Width Petal.Length Petal.Width

Sepal.Length 1.000000 -0.117570 0.871754 0.817941

Sepal.Width -0.117570 1.000000 -0.428440 -0.366126

Petal.Length 0.871754 -0.428440 1.000000 0.962865

Petal.Width 0.817941 -0.366126 0.962865 1.000000

表2-1 あやめの相関係数

Sepal.Widthのみ他との相関係数がマイナスになっており、負の相関があることがわかる。残りの組み合わせの相関係数は非常に高くなっており、正の相関があることがわかる。

　PCAによる分析を行う。

このPCAの結果を見るとsetosaはよく分かれており、残りのvirginicaとversicolorは少し分かれていることがわかる。

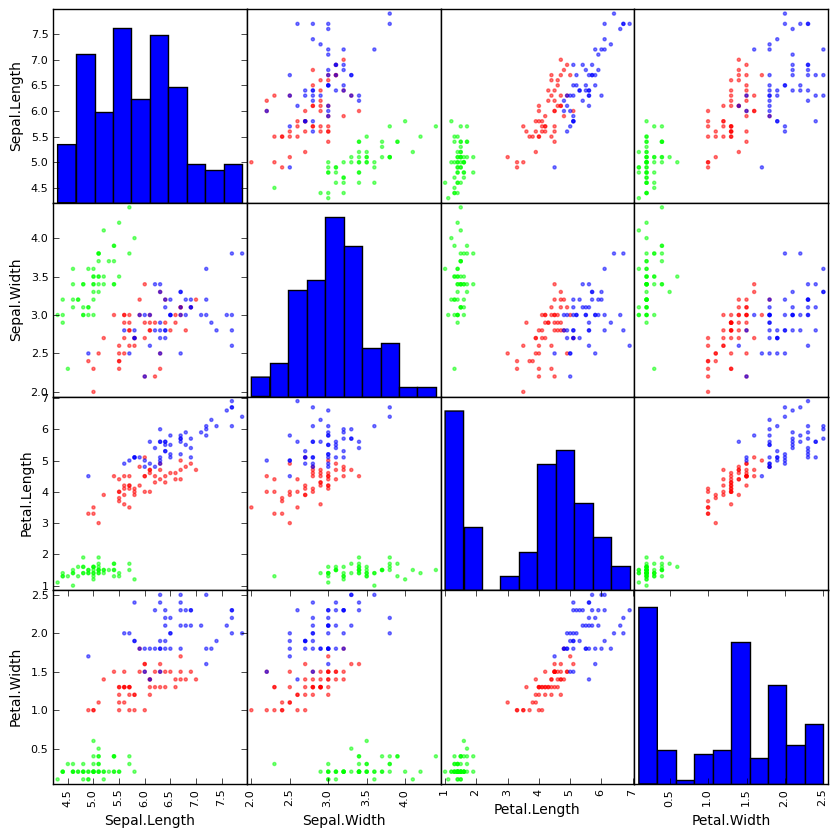


図2-1 あやめの行列散布図

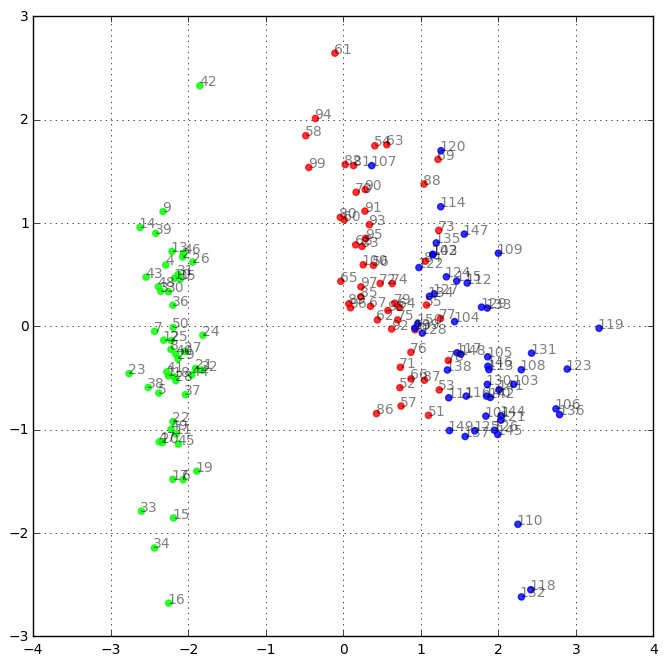


図2-2 PCA解析の図(第一成分、第二成分)

累積寄与度も確認すると、2つの成分であやめをほぼ表現できることがわかる。元データを見ると、明らかにsetosaのみPetalのサイズが小さいことがわかり、これが第一成分であると予測できる。また、versicolorのPetalサイズもvirginicaと比べて少し低めになっていることも第一成分による分離と一致している。

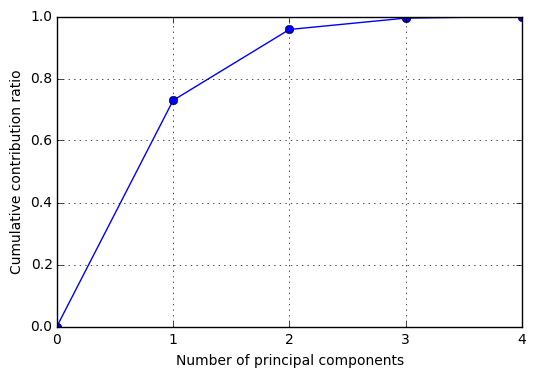


図2-3 累積寄与度の図

最後に階層的クラスタリングをする。

この図を見ると、階層的クラスタリングではversicolorとvirginicaの判別はできなさそうだが、setosaはある程度綺麗に分けられている。

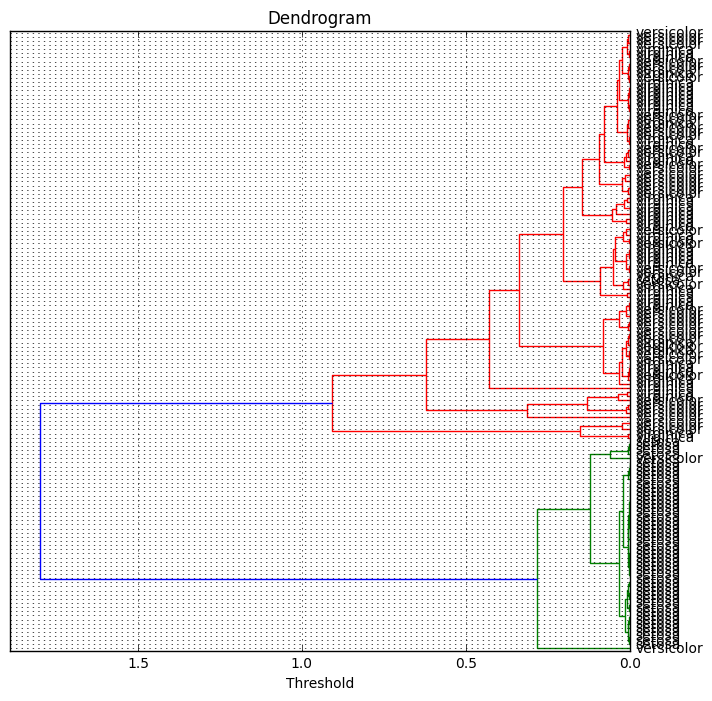


図2-4 階層的クラスタリング

課題3 アワビのデータ

　まず散布図行列を用いて全体の傾向を把握する。

Ring以外の要素が正の相関を持っていることがわかる。また、Ringにはノイズが入っているように見えるが、これはRingに整数しか入っていないためこのような散布図になっている。I(緑)は年齢が低く、それ以外の値もM(赤)F(青)よりも低いことが読み取れる。

　続いて相関係数も調べる

相関係数の表は載せるのが難しかったため代わりにヒートマップを載せる。これを見ると、全て正であり、RingsとHeightのそれぞれの相関係数が低いことがわかる。

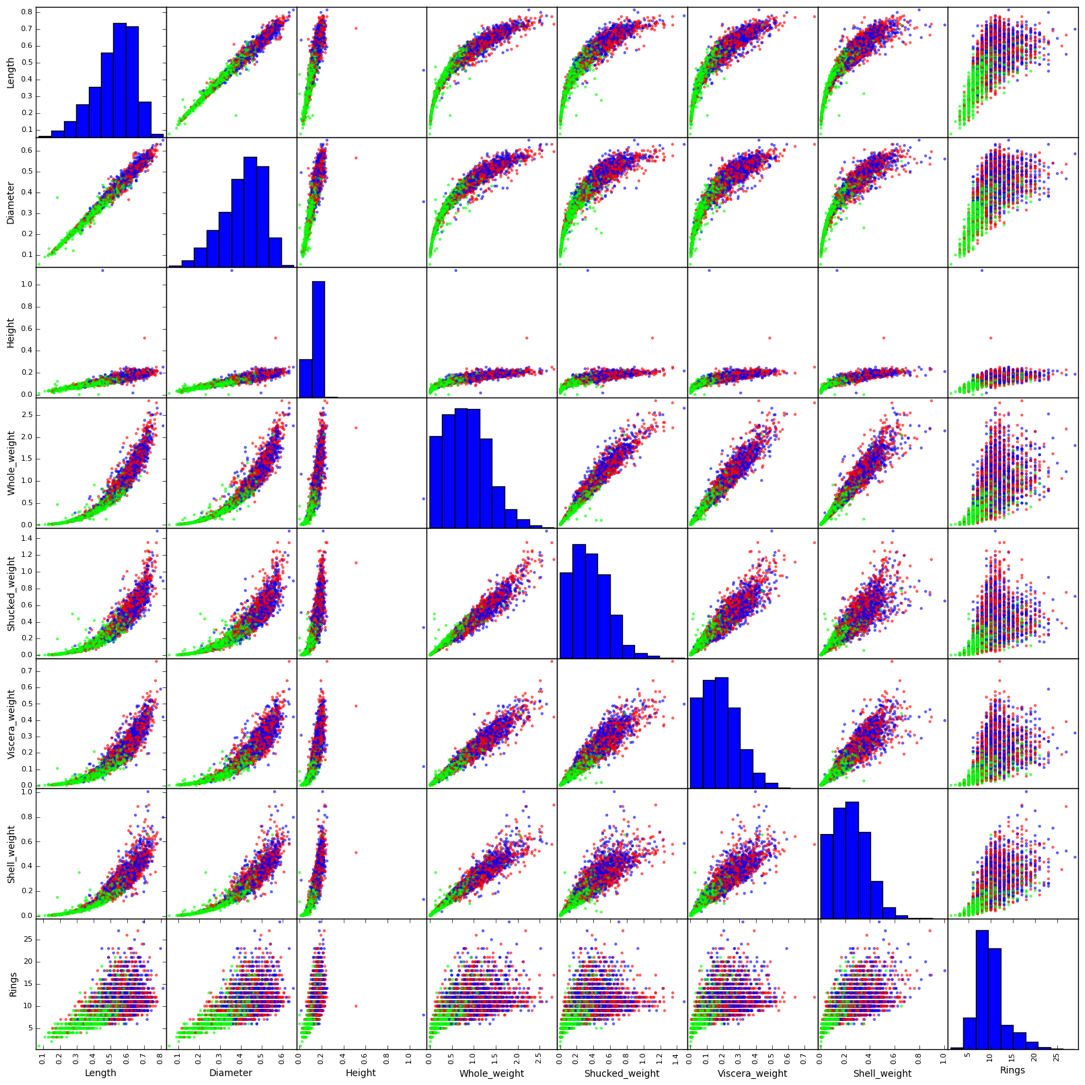


図3-1 アワビの行列散布図

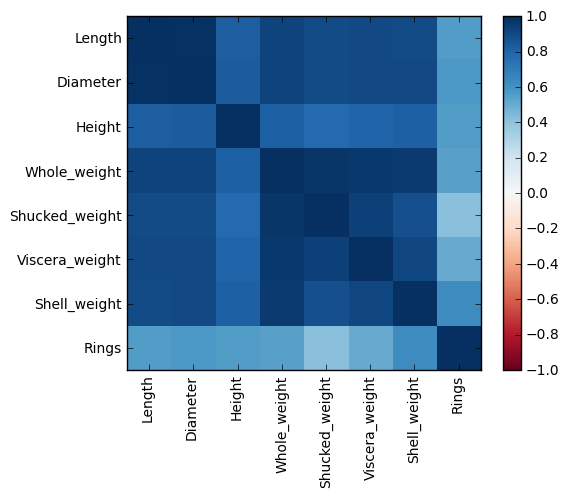


図3-2 アワビの相関係数のヒートマップ

　続いてPCAによる解析を行う。

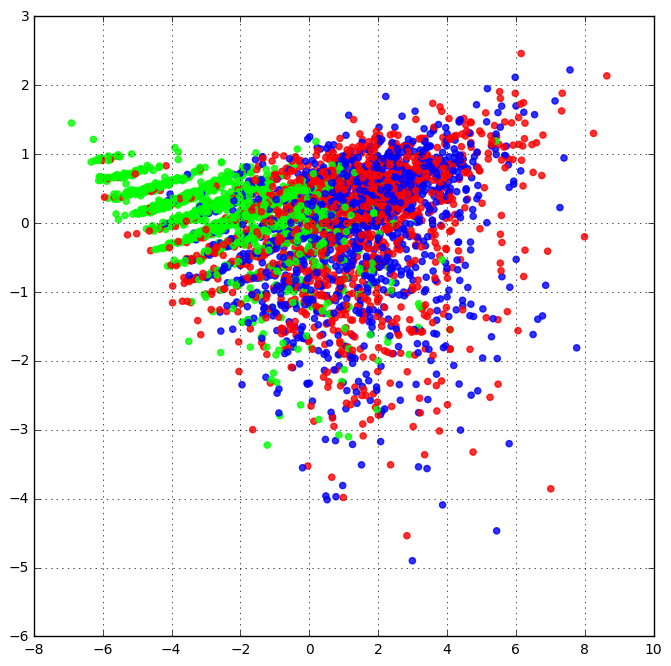


図3-3 PCA解析(第一成分、第二成分)

累積寄与度を見ると第二成分までで9割を占めているためこれが全ての点を最も広く分離できる図であると思われるが、あまり分離できていないように思う。このことからどの成分を用いてもこれ以上MFIを分離できるとは思えない。

今回のデータセットでは、アワビのオスメス子供はうまく分離できないことがわかった。