Exercici 1

Apliqueu una anàlisi PCA a les dades de la PAC i estudieu els valors propis i les variances resultants. Quantes components principals són necessàries per representar un 95% de la variança de les dades originals?
Decidiu si cal centrar o normalitzar les dades prèviament.

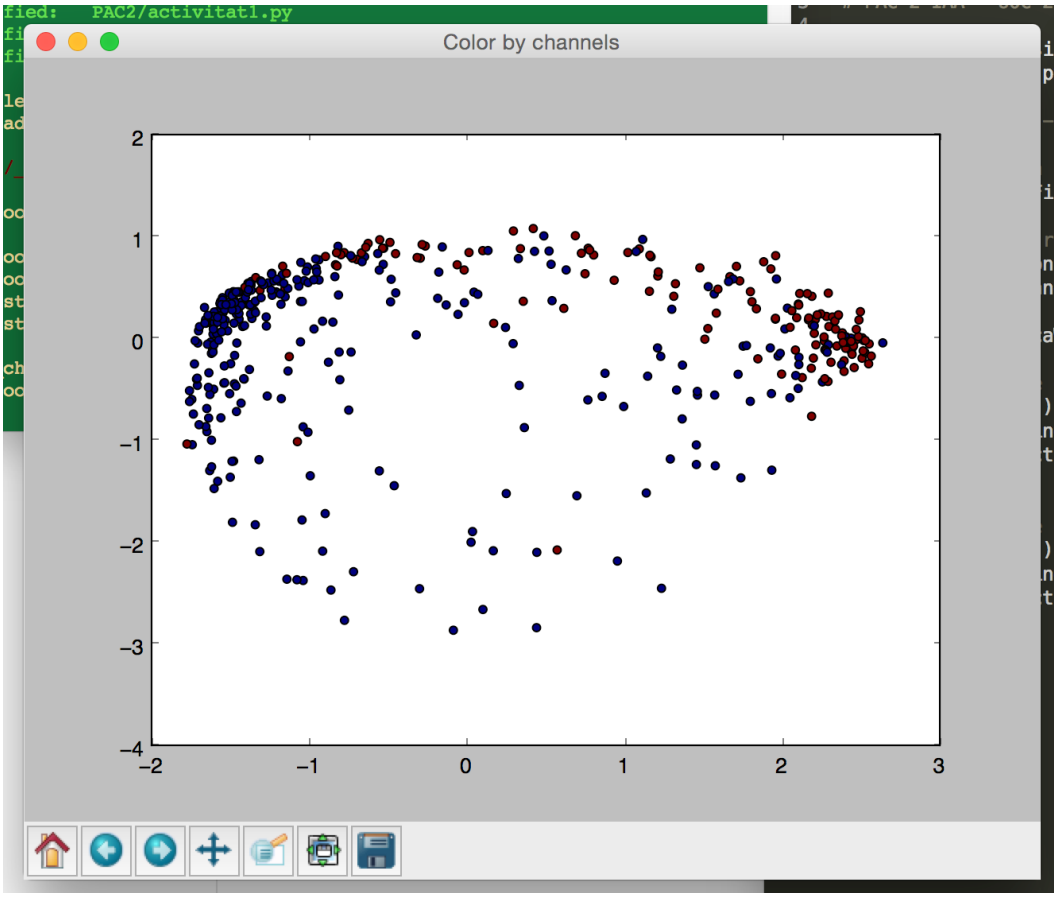
En el fitxer adjunt ***activitat1.py*** hi ha implementada la execució dels mètodes necessàris per al càlcul de l'anàlisi PCA resultant amb el nombre de components principals necessàris per a obtenir el 95% de la variaça. El resultat d'aquest anàlisi és que necessitem **4** components principals per tal d'aconseguir aquest percentatge de variança.
Per tal d'obtenir aquests resultats hem hagut de tractar les dades originals, aconseguint valors comparables i vàlids per a fer l'anàlisi: a cada valor se li resta la mitjana i es divideix entre la desviació estàndard. Per al processament del fitxer de dades, hem assumit que la primera linia d'aquest és un *header* que no conté dades importants.

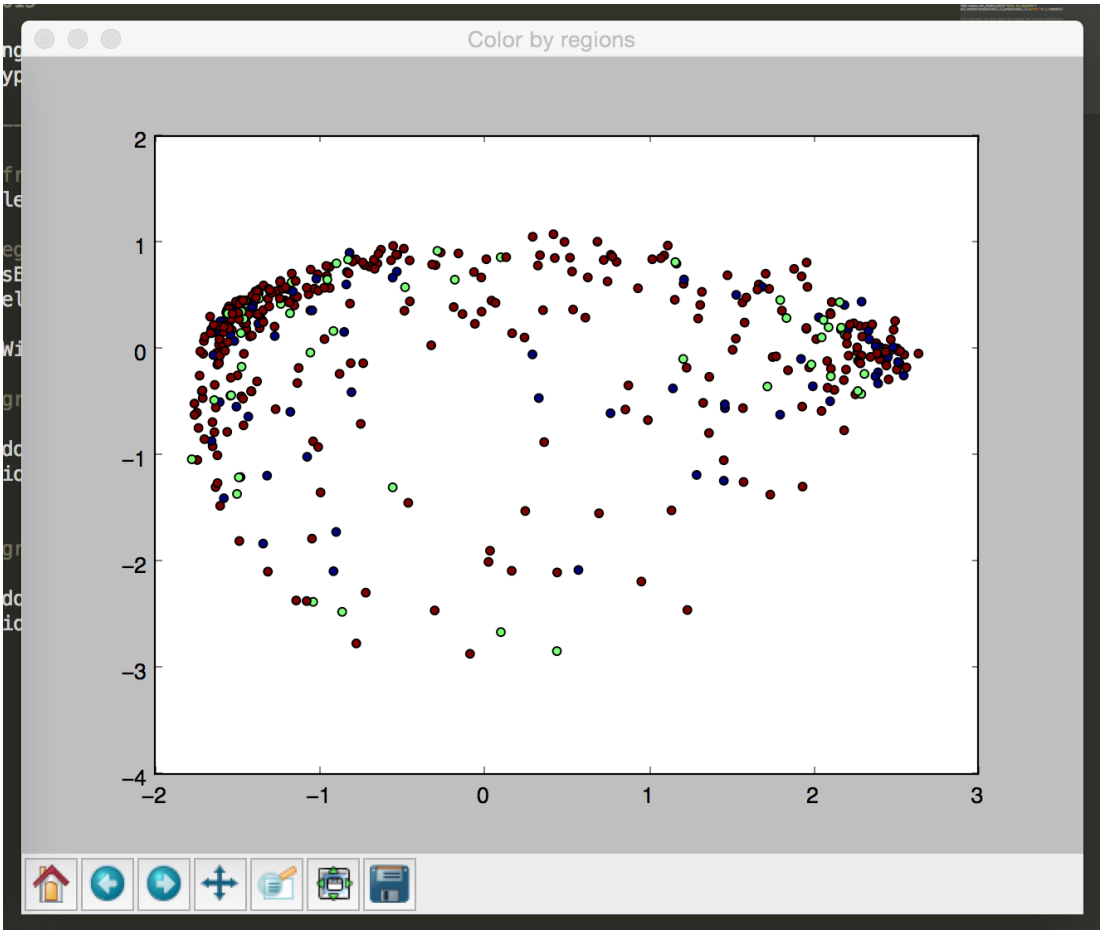
La implementació del càlcul i de tots els mètodes que es crien des dels arxius corresponents a cada activitat la trovem en l'arxiu ***fileProcessing.py***

Exercici 2

Amb les dues components principals obtingudes a l'exercici anterior, representeu gràficament les dades en dues dimensions (cada línia del fitxer haurà de ser un punt de la gràfica). Haureu de dibuixar dues gràfiques. A la primera, el color dels punts serà el tipus de client (engròs/detall). A la segona, la zona de procedència (Lisboa/ Oporto/altres).
Es poden distingir les classes?

Els gràfics corresponents a aquest càlcul els veiem a continuació:

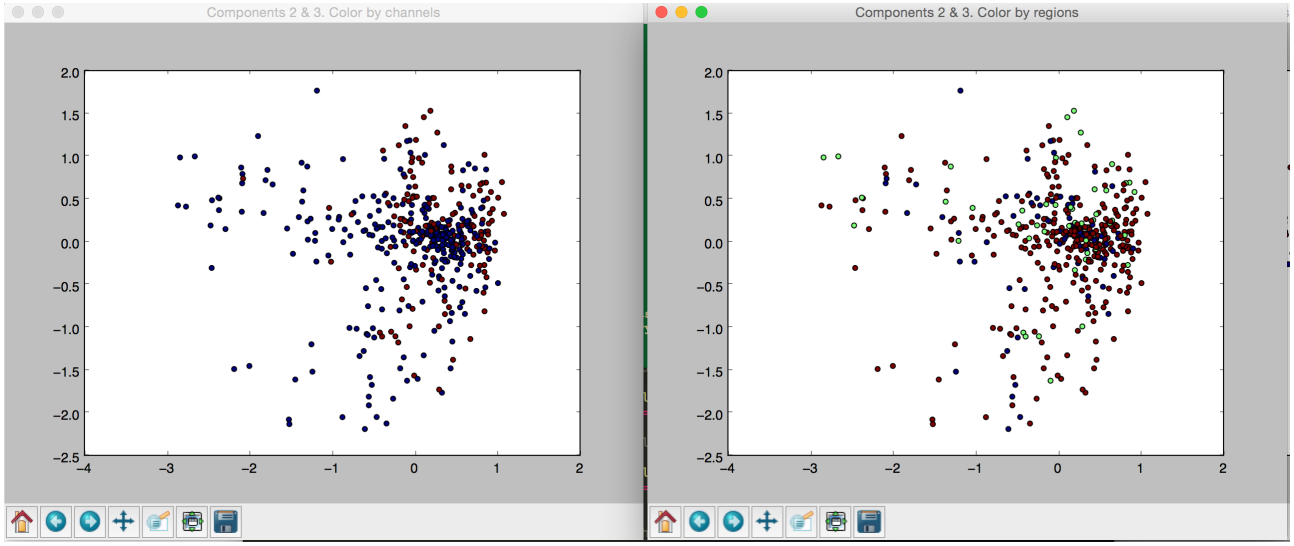


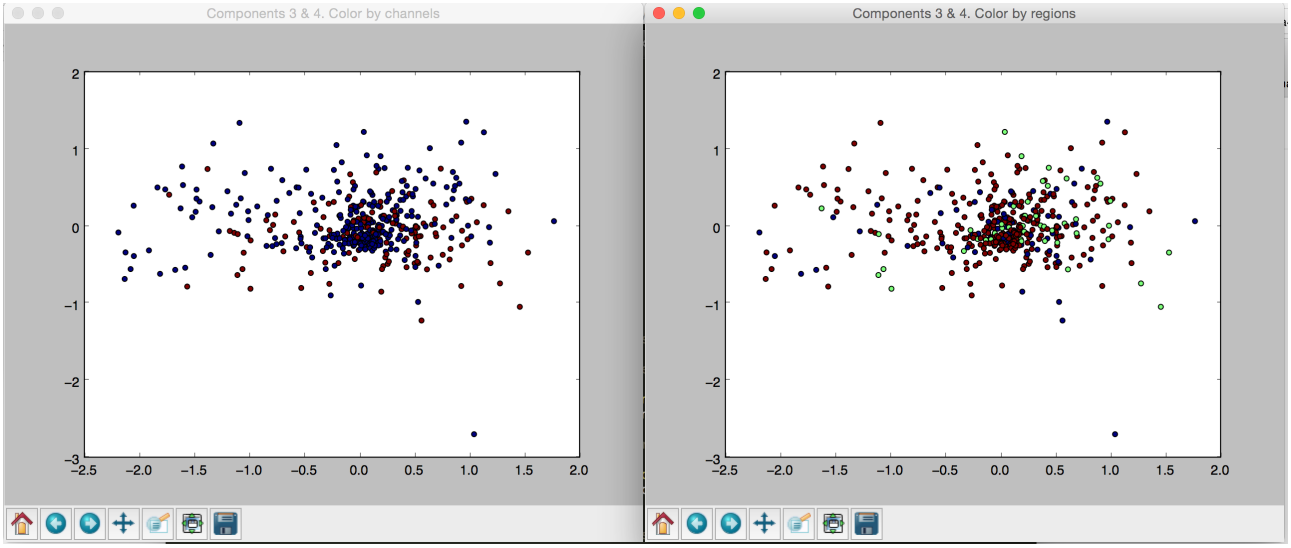


Podem veure una clara diferència en la distribució de les classes, tanta que gairebé podriem fer servir aquesta distribució per a realitzar la classificació de classes. Això no passa, però, amb la distribució de les regions. Podem trobar la implementació en el fitxer ***activitat2.py***

Exercici 3

Repetiu l'exercici anterior però amb les components 2 i 3 del PCA, i després amb les 3 i 4. Analitzeu les diferències amb les gràfiques de l'exercici 2.



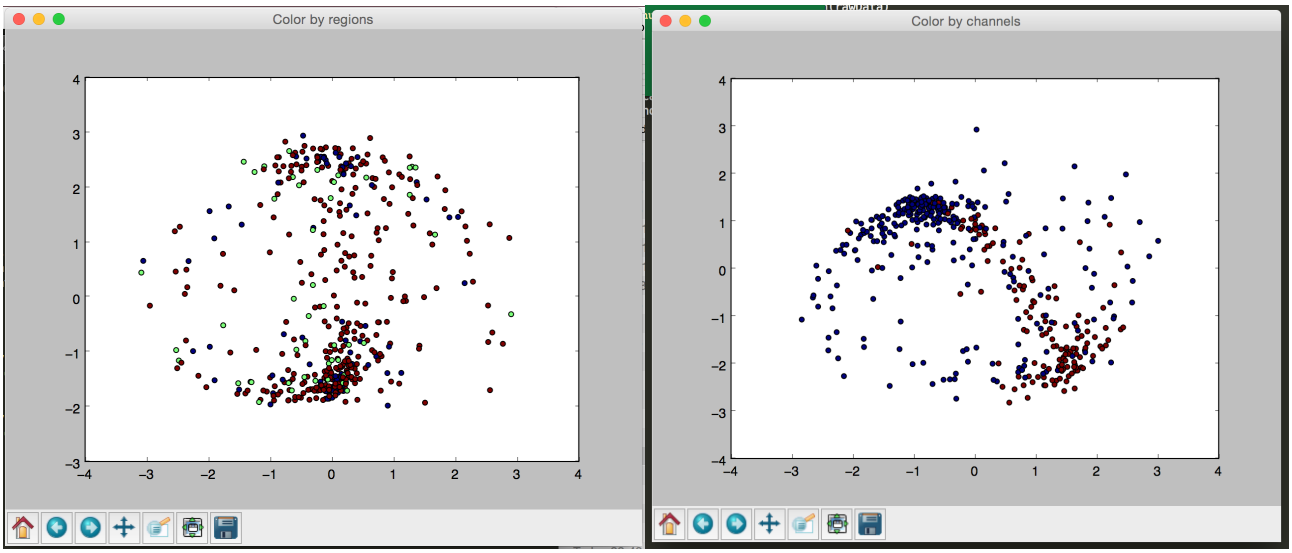


Tal com podem veure en les distribucions corresponents als components 2 i 3, la distribució de classes no està tant clara comparat amb la distribució dels components 1 i 2 (la que hem realitzat en l'exercici anterior). El mateix ens passa amb la distribució corresponent als components 3 i 4, que encara que es pot veure una concentració més gran d'un dels components, no es podria arribar a fer una classificació només amb aquesta informació aïllada.

La implementació del càlcul i representació el tenim implemntat en el fitxer ***activitat3.py***

Exercici 4

Apliqueu el mètode "multidimensional scaling" (MDS) a les dades de vendes. Dibuixeu una gràfica en dues dimensions colorejant amb el tipus de client i una altra amb la seva procedència. Compareu-les amb les gràfiques de l'exercici 2.
Podreu trobar la implementació i molta informació sobre com fer-la servir al web de scikit-learn.



Amb aquesta mètode també obtenim una clara separació en la distribució dels punts separats per canals, però no tant pel que fa a la distribució per regions. Si comparem amb els resultats de l'exercici 2 veiem que aquest mètode també ens permet observar una separació clara per canals, Comparant amb els resultats obtinguts en l'exercici 2 però amb valors diferents.
La implementació dels càlculs la podem trobar en l'arxiu ***activitat4.py***