Arivan Thillaikumaran
CSE 514 HW 3
425096

Problem 1

1. from fancyimpute import KNN
X_filled_knn_ctrl = KNN(k=7).complete(x)
X_filled_knn_case = KNN(k=7).complete(x_case)

These three lines of code were added in order to use k-NN with k=7 to impute the missing values.

2. **Bottom up:**
P = 176
N = 188

The accuracy of each cluster is found by taking the majority classification of the cluster and dividing the number of nodes in the cluster classified as this majority classification by the size of the cluster. (For example, in a cluster of ten nodes, if 7 nodes are patients and 3 are normal, the majority for this cluster is patients and the accuracy is 7/10 = 70%.)

Cluster 1 accuracy = 113/195 = .5795
Cluster 2 accuracy = 95/169 = .5621
Average accuracy = .5705

False Positives = 74 ⇔ True Positives = 95
False Negatives = 82 ⇔ True Negatives = 113
True Positive Rate = 95/176 = .540
Accuracy = (95+113)/364 = .5714

The following patients were mistakenly classified as normal:
WGAAD1, WGAAD3, WGAAD4, WGAAD6, WGAAD7, WGAAD8, WGAAD9, WGAAD10, WGAAD14, WGAAD15, WGAAD16, WGAAD17, WGAAD21, WGAAD26, WGAAD27, WGAAD29, WGAAD30, WGAAD38, WGAAD40, WGAAD41, WGAAD45, WGAAD46, WGAAD47, WGAAD51, WGAAD52, WGAAD53, WGAAD55, WGAAD58, WGAAD59, WGAAD60, WGAAD62, WGAAD70, WGAAD73, WGAAD74, WGAAD75, WGAAD79, WGAAD80, WGAAD87, WGAAD89, WGAAD91, WGAAD94, WGAAD96, WGAAD97, WGAAD99,WGAAD 100, WGAAD101, WGAAD102, WGAAD105, WGAAD106, WGAAD107, WGAAD108, WGAAD110, WGAAD111, WGAAD112, WGAAD113,WGAAD 114, WGAAD115, WGAAD117, WGAAD118, WGAAD119, WGAAD121, WGAAD125, WGAAD131, WGAAD132, WGAAD133, WGAAD134, WGAAD137, WGAAD138, WGAAD139, WGAAD140, WGAAD142, WGAAD144,

WGAAD146, WGAAD149, WGAAD153, WGAAD154, WGAAD156, WGAAD158, 1WGAAD64, WGAAD165, 1WGAAD74, WGAAD176

The following normals were mistakenly classified as patients:
WGACON5, WGACON7, WGACON10, WGACON11, WGACON14, WGACON15, WGACON18, WGACON19, WGACON20, WGACON21, WGACON22, WGACON26, WGACON30, WGACON32, WGACON40, WGACON43, WGACON44, WGACON46, WGACON50, WGACON53, WGACON54, WGACON57, WGACON58, WGACON60, WGACON61, WGACON62, WGACON63, WGACON65, WGACON72, WGACON73, WGACON75, WGACON76, WGACON79, WGACON84, WGACON86,WGACON 87, WGACON90, WGACON93, WGACON94, WGACON95, WGACON97, WGACON98, WGACON103, WGACON104, WGACON106, WGACON109, WGACON110, WGACON120, WGACON124, WGACON125, WGACON126, WGACON127, WGACON132, WGACON139, WGACON144, WGACON148, WGACON150, WGACON151, WGACON157, WGACON158, WGACON160, WGACON163,WGACON 165, WGACON166, WGACON167, WGACON168, WGACON170, WGACON174, WGACON179, WGACON180, WGACON181, WGACON182, WGACON183, WGACON185

3. Top Down

Cluster 1 accuracy = 125/195 = .641
Cluster 2 accuracy = 97/169 = .574
Average cluster accuracy = .6075

False Positives = 62 ⇔ True Positives = 97
False Negatives = 80 ⇔ True Negatives = 125
True Positive Rate = 97/176 = .551
Accuracy = (97+125)/364 = .6101

The following patients were mistakenly classified as normal:
WGAAD1, WGAAD3, WGAAD4, WGAAD6, WGAAD7, WGAAD8, WGAAD9, WGAAD14, WGAAD15, 1WGAAD6, WGAAD17, WGAAD21, WGAAD22, WGAAD26, WGAAD27, WGAAD29, WGAAD30, WGAAD38, WGAAD40, WGAAD41, WGAAD45, WGAAD46, WGAAD47, WGAAD51, WGAAD52, WGAAD53, WGAAD58, WGAAD59, WGAAD60, WGAAD62, WGAAD70, WGAAD71,WGAAD73, WGAAD74, WGAAD77,WGAAD79, WGAAD80, WGAAD87, WGAAD89, WGAAD91, WGAAD94, WGAAD96, WGAAD97, WGAAD99, WGAAD100, WGAAD101, WGAAD102, WGAAD105, WGAAD106, WGAAD107, WGAAD108, WGAAD110, WGAAD111, WGAAD112, WGAAD113, WGAAD115, WGAAD117, WGAAD119, WGAAD121, WGAAD125, WGAAD131, WGAAD132, WGAAD133, 1WGAAD34, WGAAD138, WGAAD139, WGAAD142, WGAAD143, WGAAD144,

WGAAD146, WGAAD149, WGAAD150, WGAAD153, WGAAD156, WGAAD158, WGAAD164, WGAAD165, WGAAD174, WGAAD175, WGAAD176

The following normals were mistakenly classified as patients:
WGACON5, WGACON7, WGACON10, WGACON14, WGACON18, WGACON19, WGACON20, WGACON21, WGACON26, WGACON29, WGACON32, WGACON33, WGACON40, WGACON50, WGACON53, WGACON54, WGACON57, WGACON58, WGACON60, WGACON61, WGACON63, WGACON75, WGACON76, WGACON80, WGACON84, WGACON86, WGACON87, WGACON90, WGACON93, WGACON94, WGACON98, WGACON103, WGACON104, WGACON106, WGACON107, WGACON109, WGACON110, WGACON120, WGACON125, WGACON126, WGACON127, WGACON132, WGACON139, WGACON144, WGACON148, WGACON150, WGACON151, WGACON157, WGACON158,WGACON160, WGACON163, WGACON165, WGACON166, WGACON167, WGACON168, WGACON170, WGACON174, WGACON179, WGACON180, WGACON181, WGACON182, WGACON183

Comparing the first three levels of both Bottom Up and Top Down:

Level 3: This level returns 8 clusters.

Bottom Up**:** Average Cluster Accuracy: 56.49%

Cluster	Size	Majority Classification	Accuracy
1	108	Normal	59.3%
2	32	Normal	62.5%
3	39	Normal	66.7%
4	96	Patient	63.5%
5	30	Normal	70%
6	15	Patient	80%
7	34	Patient	55.9%
8	10	N/A	50%

Top Down: Average Cluster Accuracy: 61.24%

Cluster	Size	Majority Classification	Accuracy
1	37	Normal	64.9%
2	30	Normal	53.3%
3	63	Normal	66.7%
4	75	Normal	53.3%
5	15	Patient	66.7%
6	31	Patient	67.7%
7	40	Patient	62.5%
8	73	Patient	54.8%

Level 2:

Bottom Up: Average Cluster Accuracy = 60.65

Cluster	Size	Majority Classification	Accuracy
1	155	Normal	56.1%
2	73	Normal	54.8%
3	40	Normal	65%
4	96	Patients	66.7%

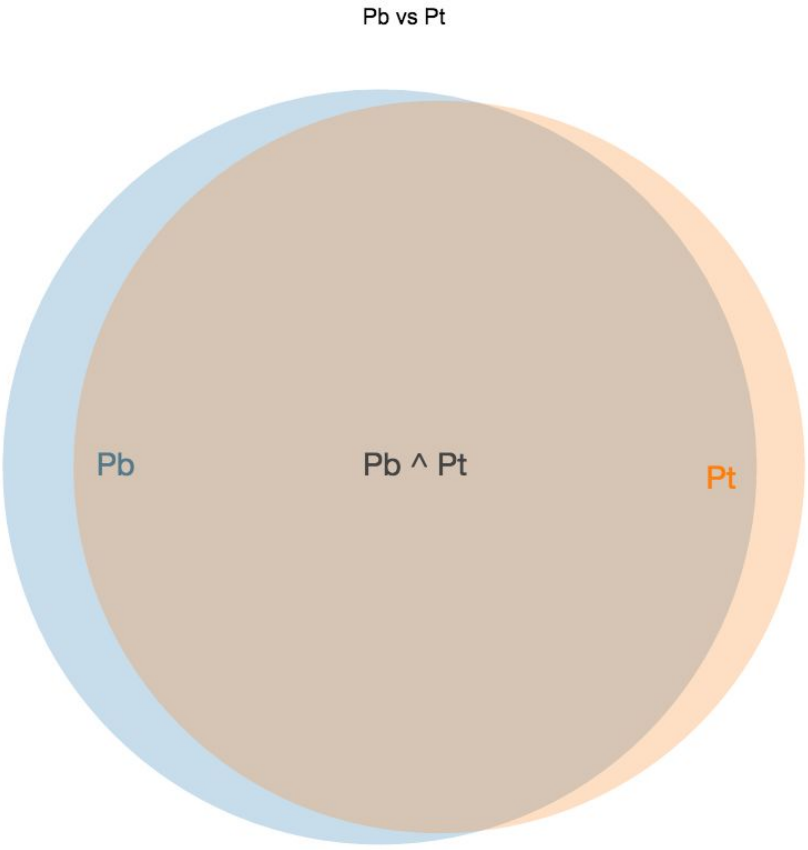
Top Down: Average Cluster Accuracy = 61.83%

Cluster	Size	Majority Classification	Accuracy
1	68	Normal	61.8%
2	137	Normal	60.6%
3	46	Patients	67.4%
4	113	Patients	57.5%

Top Down consistently returns better average accuracies in creating its clusters, both in level 3, level 2, and level 1. I also noted that in level 3, top down created more evenly sized clusters than did bottom up, which may help in performance as it doesn't make the decisive decisions too early. This makes sense as bottom up is focused on making local patterns due to the amount of clusters that you start out with, whereas top down starts from one big cluster, which can focus on a global distribution of points.

4. Venn Diagram comparing Pt and Pb

|Pt| = 159
|Pb| = 169
Pt ^ Pb = 147



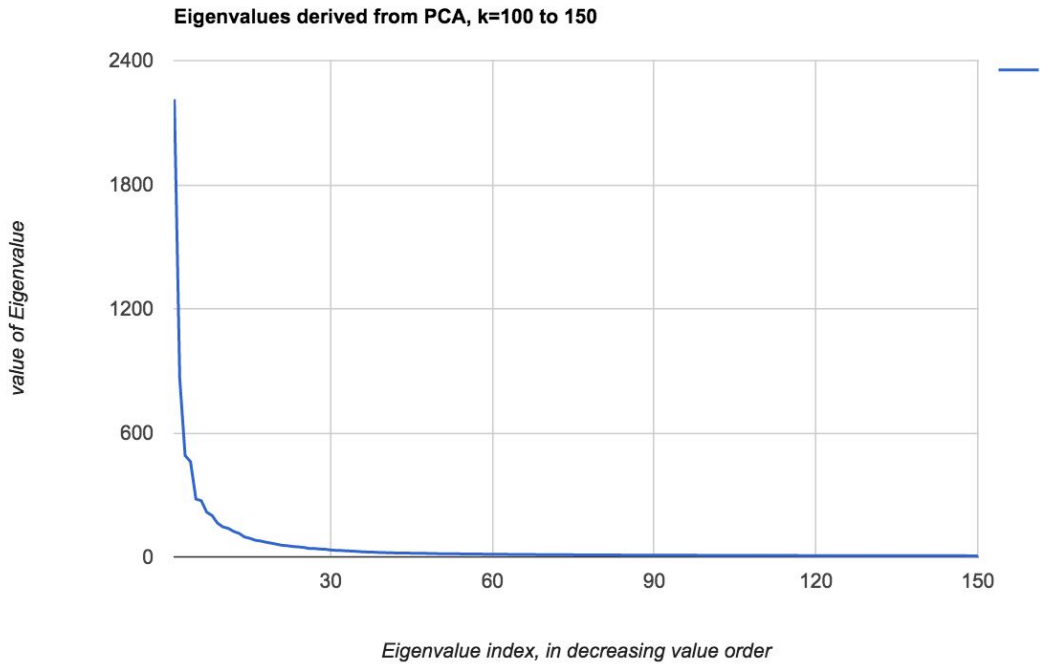
As stated before, Pb has a true positive rate of .54, and Pt has a true positive rate of .551. Therefore, as was analyzed in the three levels of clustering, top down is the higher quality clustering method.

5. Optional
How to identify outliers?

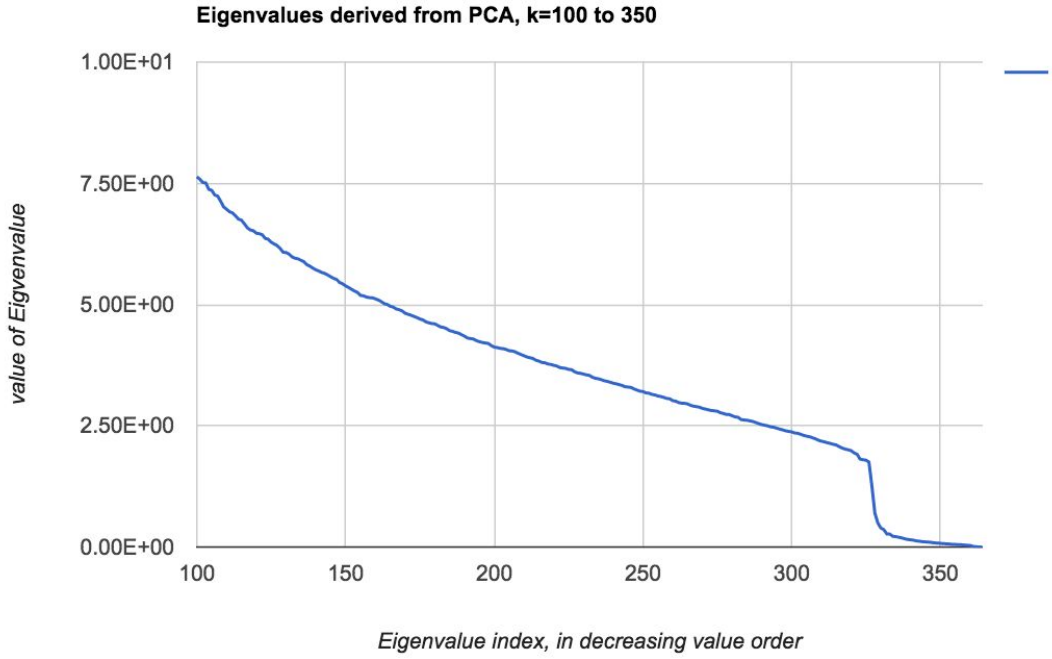
I think that during bottom up traversal, during the proess, there will be clusters made that are very large in size, and there will be some clusters that are isolated and by themselves. When your reach a level that put everything in a cluster except for maybe one or two, that are still in their own cluster, you can see that these are outliers.

Problem 2

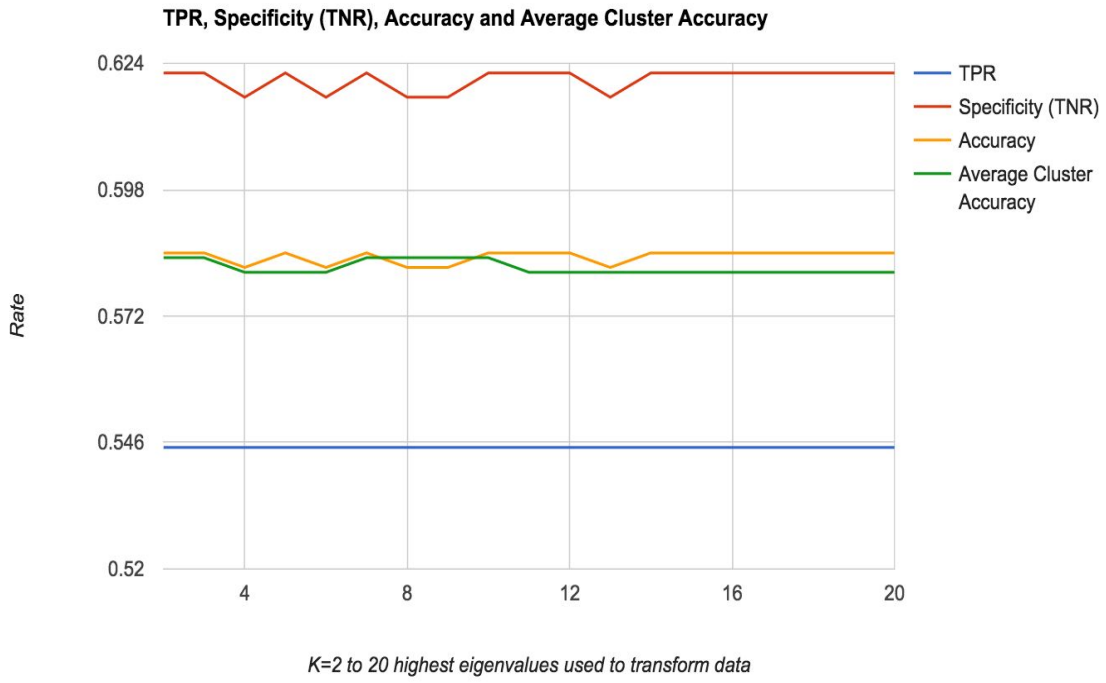
1. Eigenvalues in nonincreasing order, plotted to see the values decrease



Zooming in on the tail...

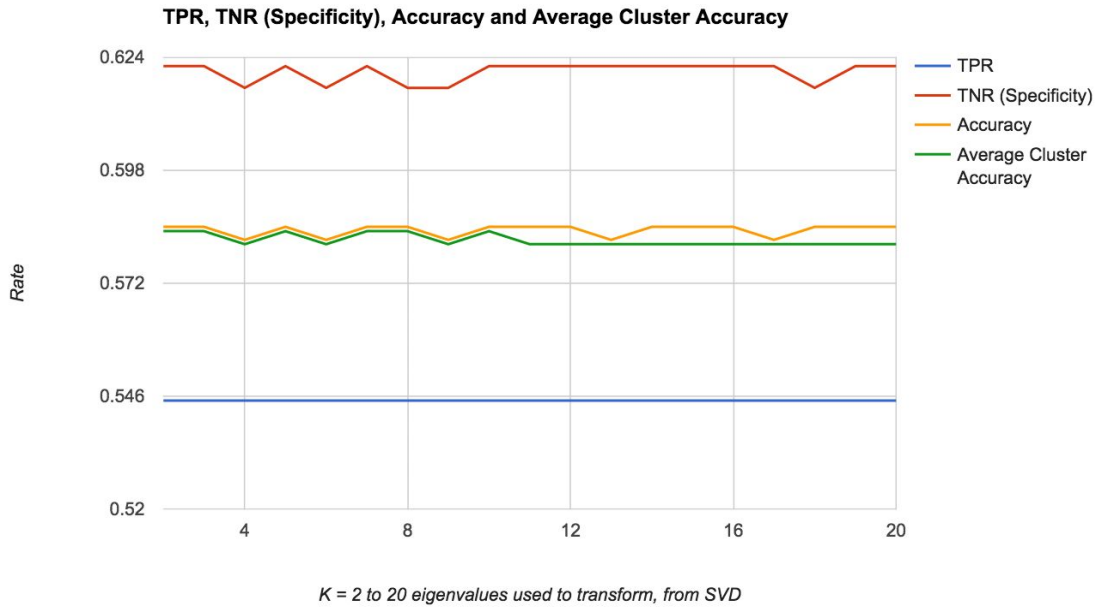


2.



TPR = True Positives / Real Positives
TNR = True Negatives / Real Negatives
Accuracy = (True Positives+True Negatives) / (Real Positives+Real Negatives)
Average Cluster Accuracy described previously.

Problem 3 (optional - using Singular Value Decomposition)



Using singular value decomposition to transform our dataset yielded almost identical results to using Principal Component Analysis. This makes sense to me because they are very similar methods as they are both forms of eigendecomposition.