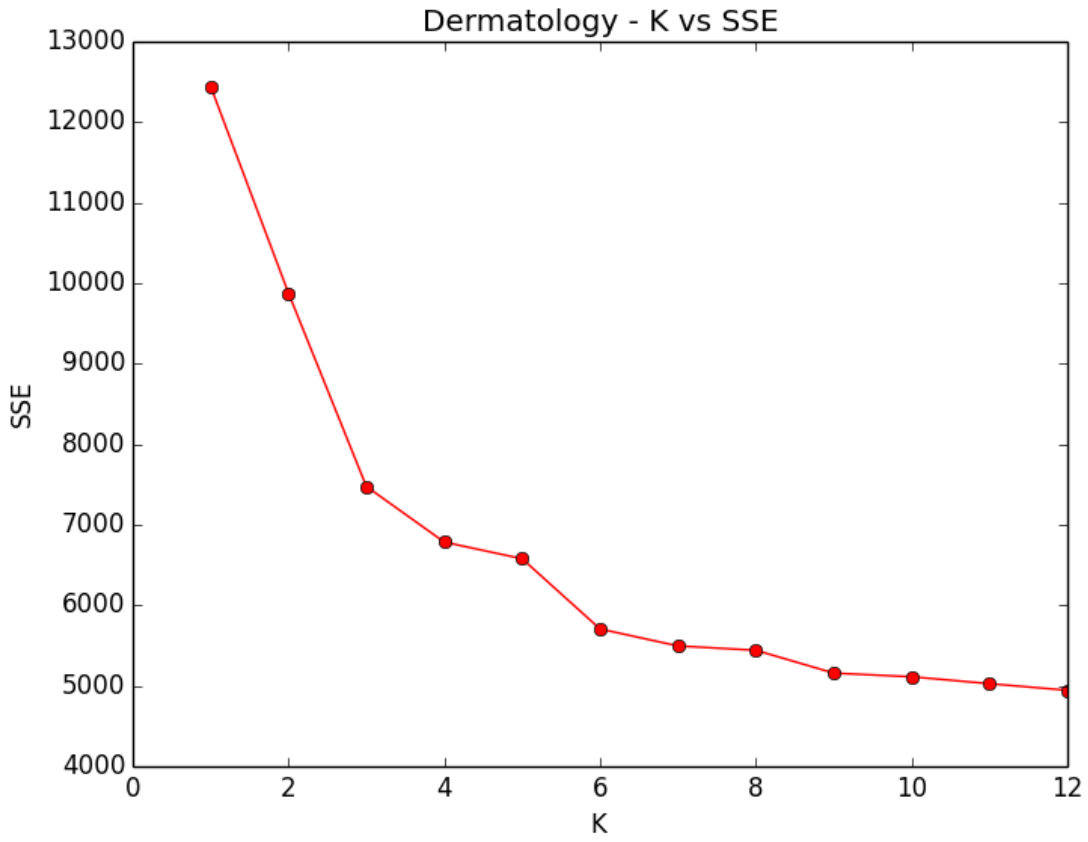
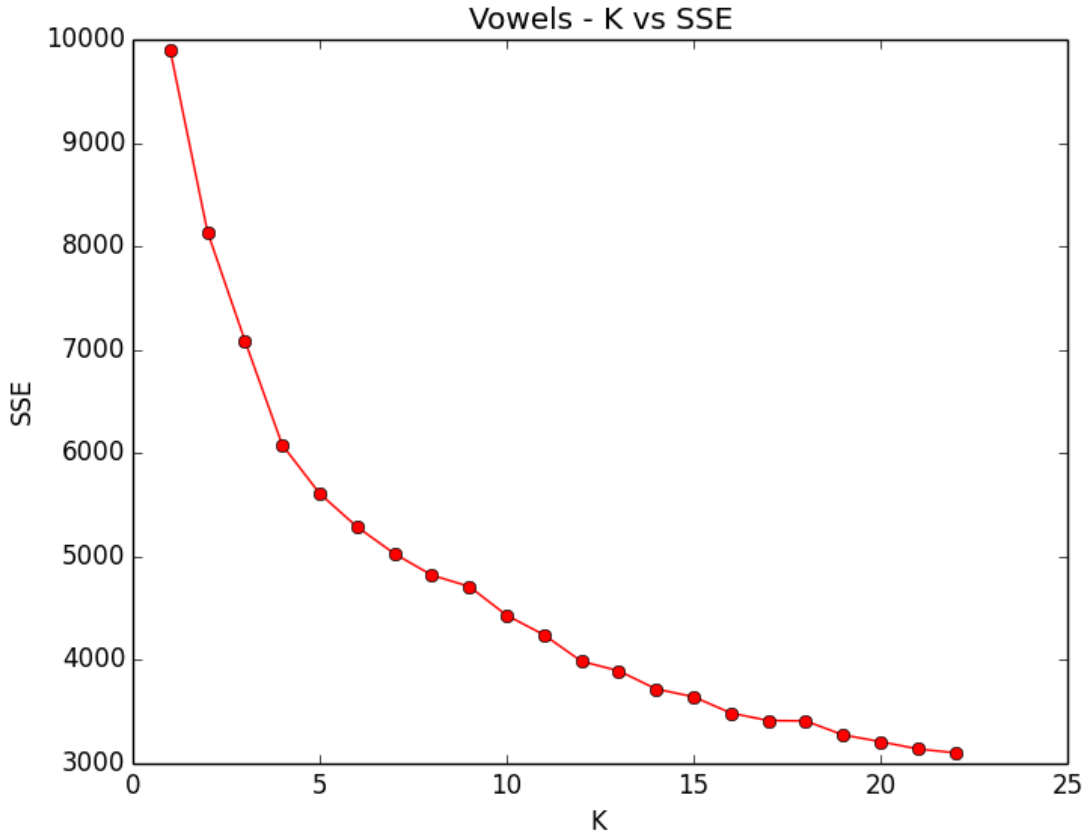
CS 6140: Machine Learning Spring 2018

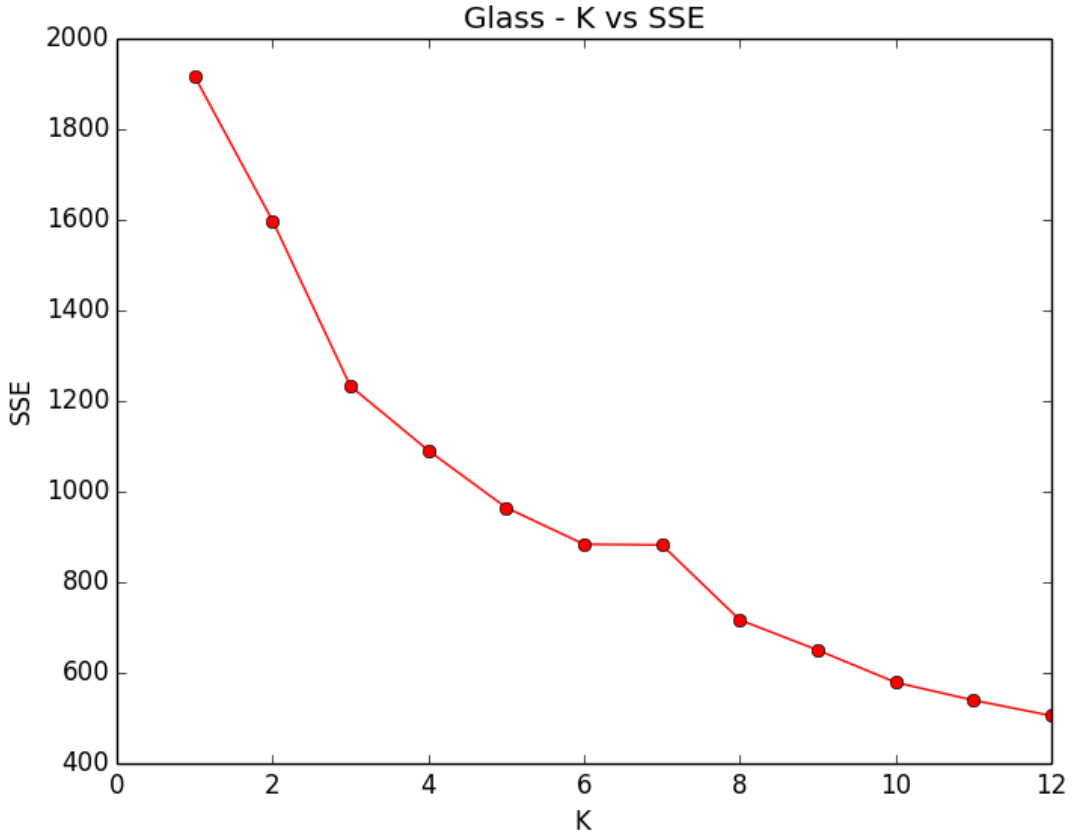
Assignment 5

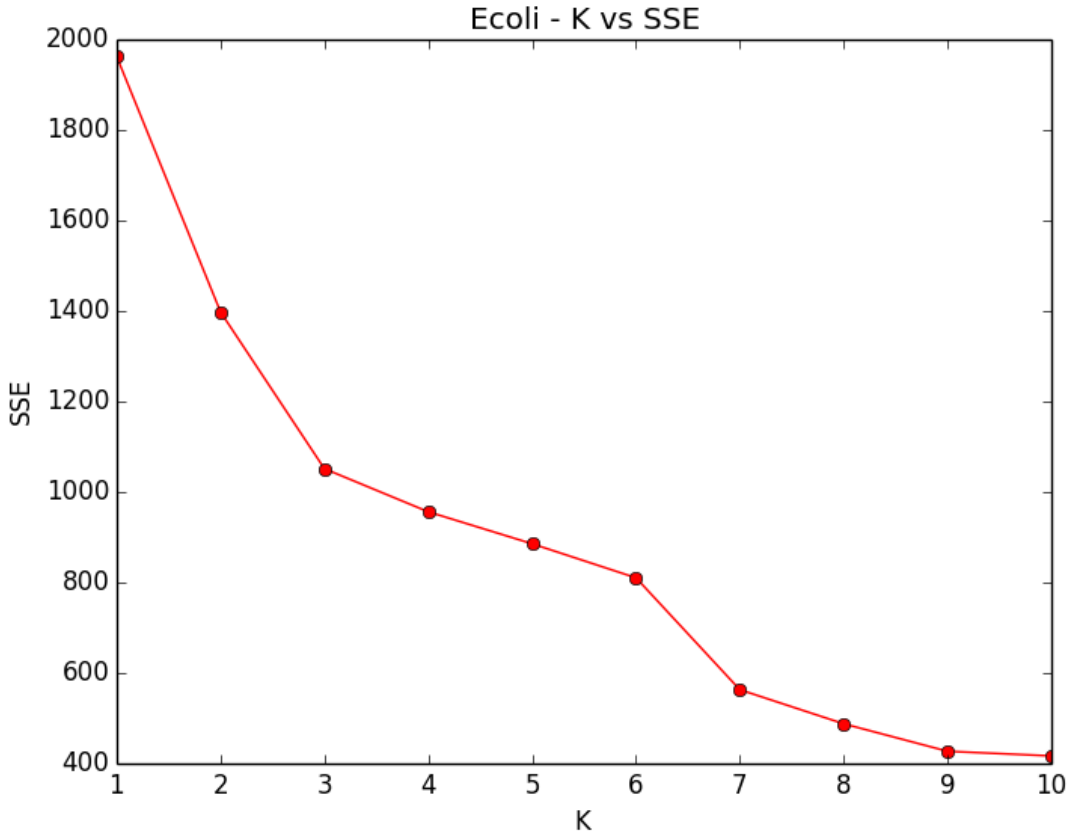
3 K-Means Clustering

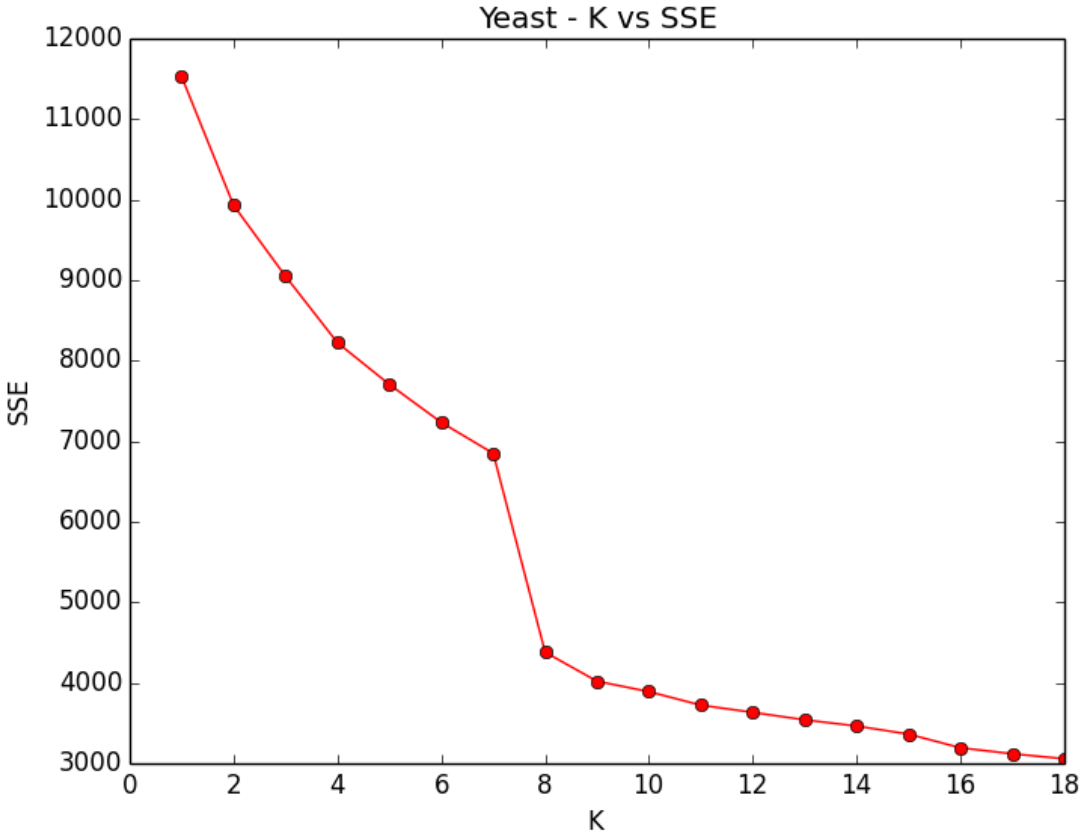
3.2.1

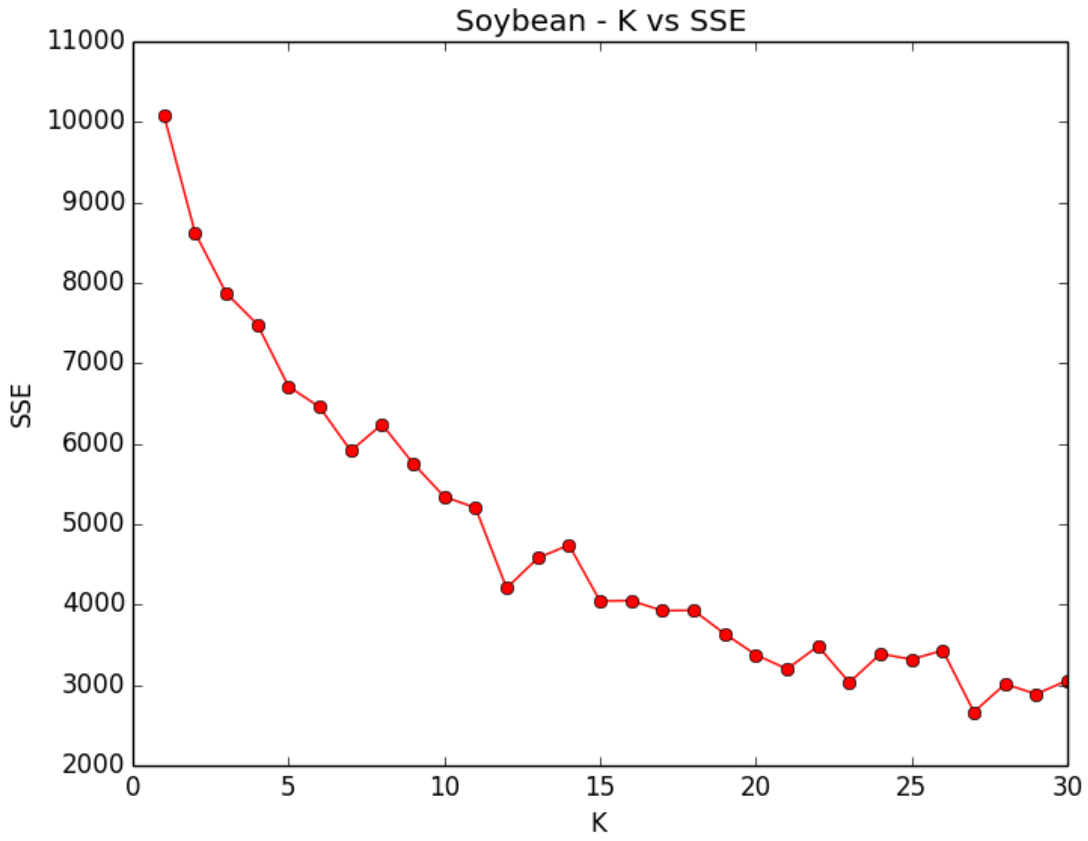



Vowels - K vs SSE
SSE
K
10000
9000
8000
7000
6000
5000
4000
3000
0
5
10
15
20
25


Glass - K vs SSE
SSE
K
2000
1800
1600
1400
1200
1000
800
600
400
0
2
4
6
8
10
12


Ecoli - K vs SSE
SSE
K
2000
1800
1600
1400
1200
1000
800
600
400
1
2
3
4
5
6
7
8
9
10


Yeast - K vs SSE
SSE
K
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
0
2
4
6
8
10
12
14
16
18



3.2.2

Dataset	K	SSE	NMI
Dermatology	6	5706.417484	0.878001156
Vowels	12	3988.34903	0.3640192002
Glass	6	883.7475985	0.3624081679
Ecoli	4	956.2455634	0.6491569767
Yeast	9	4019.829802	0.302999474
Soybean	15	4043.710593	0.7110381082

3.2.3

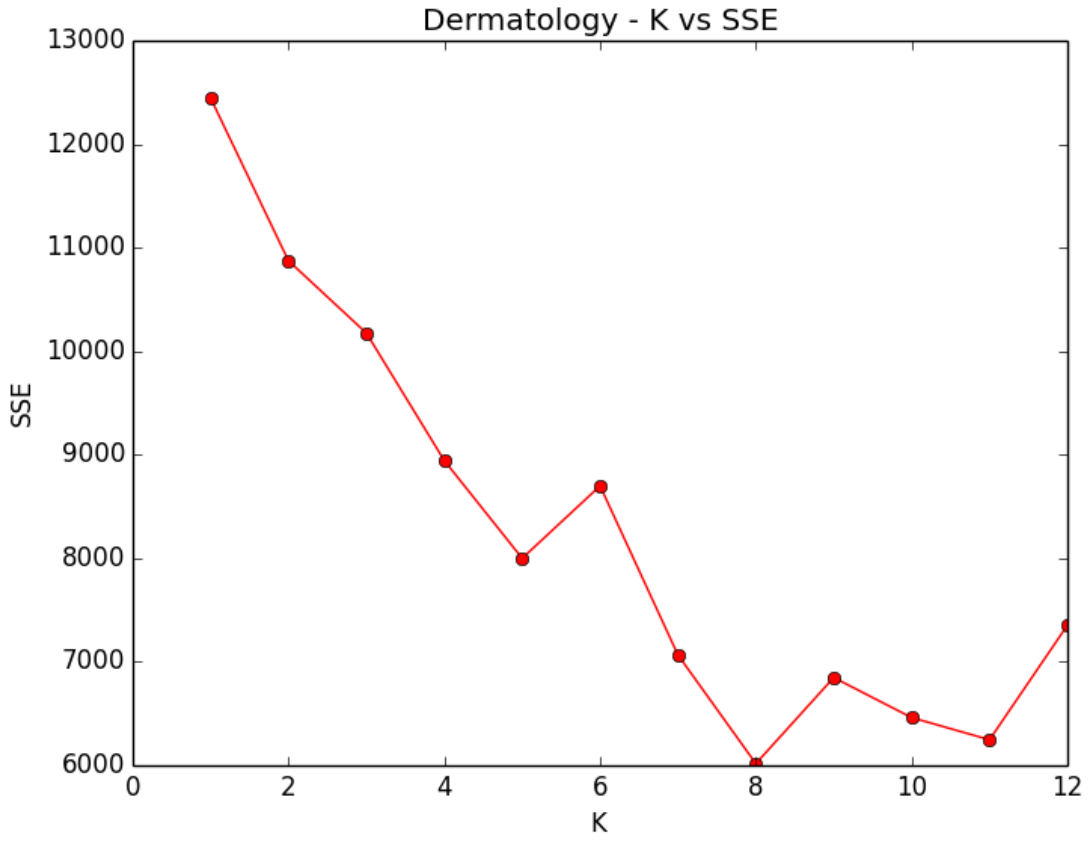
Dataset	K	SSE	NMI
Dermatology	6	5706.417484	0.878001156
Vowels	11	4243.782957	0.3601683762
Glass	6	883.7475985	0.3624081679
Ecoli	5	885.5982258	0.6325266288
Yeast	9	4019.829802	0.302999474
Soybean	15	4043.710593	0.7110381082

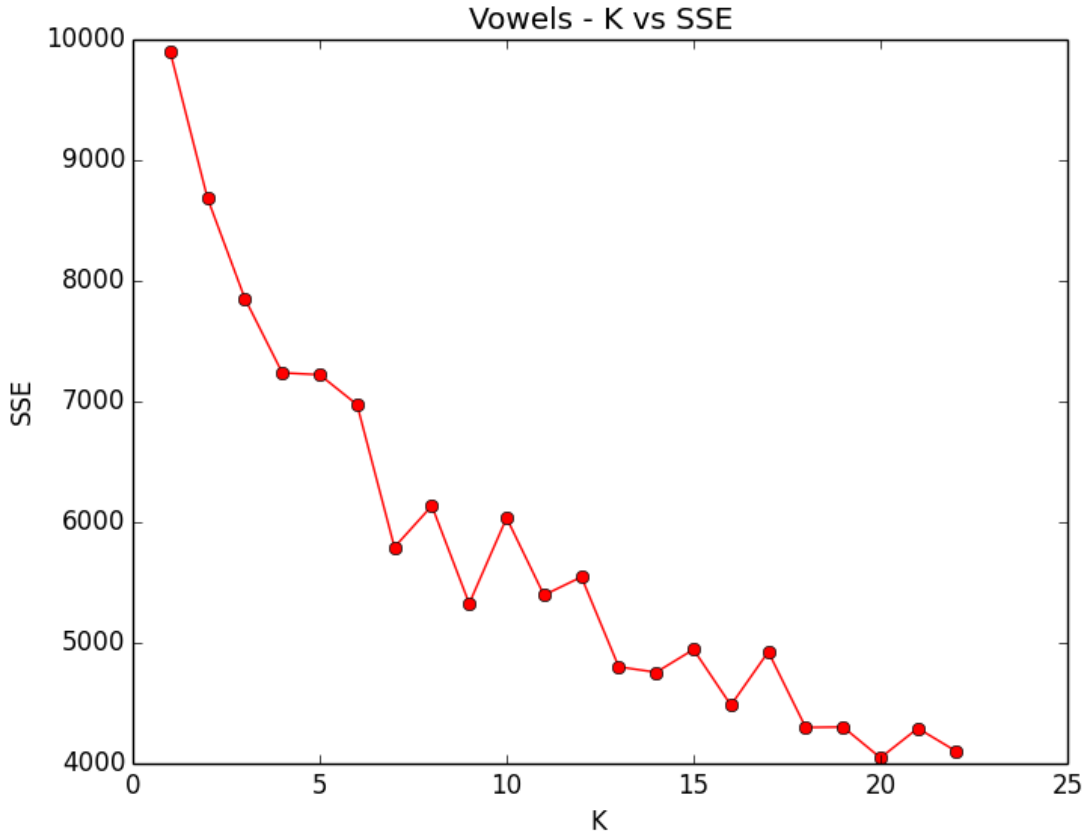
4 Gaussian Mixture Models

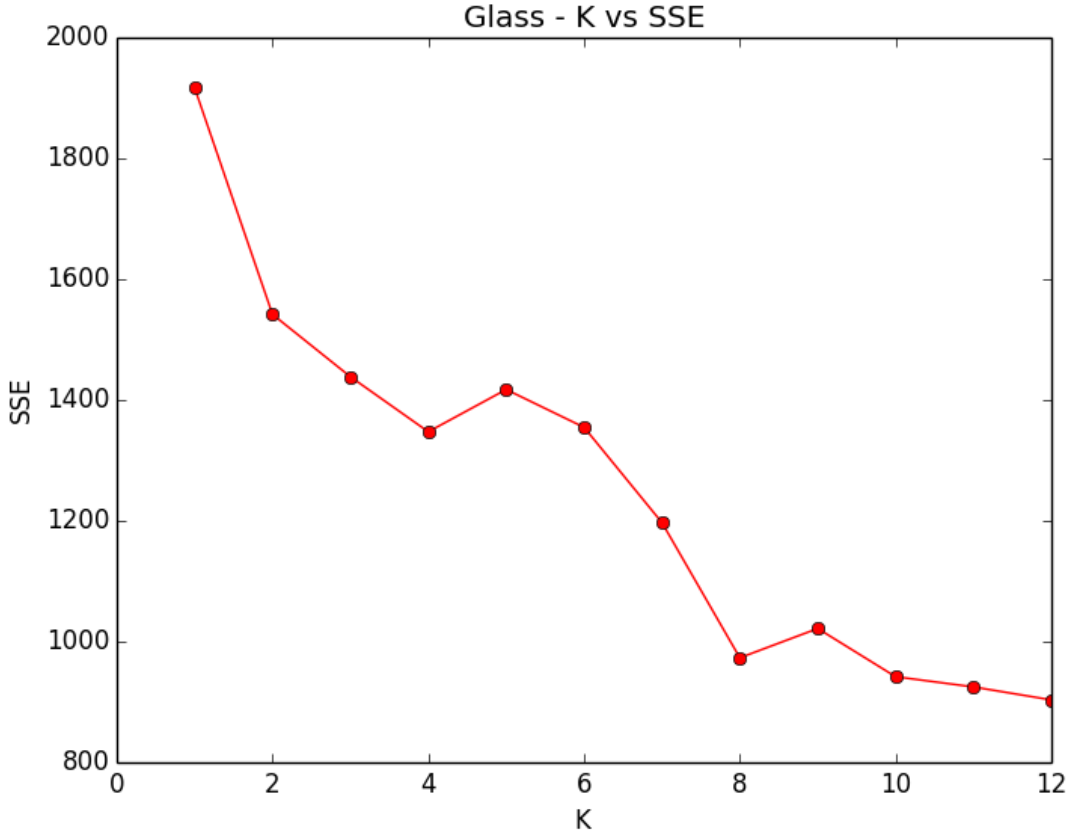
4.1.4

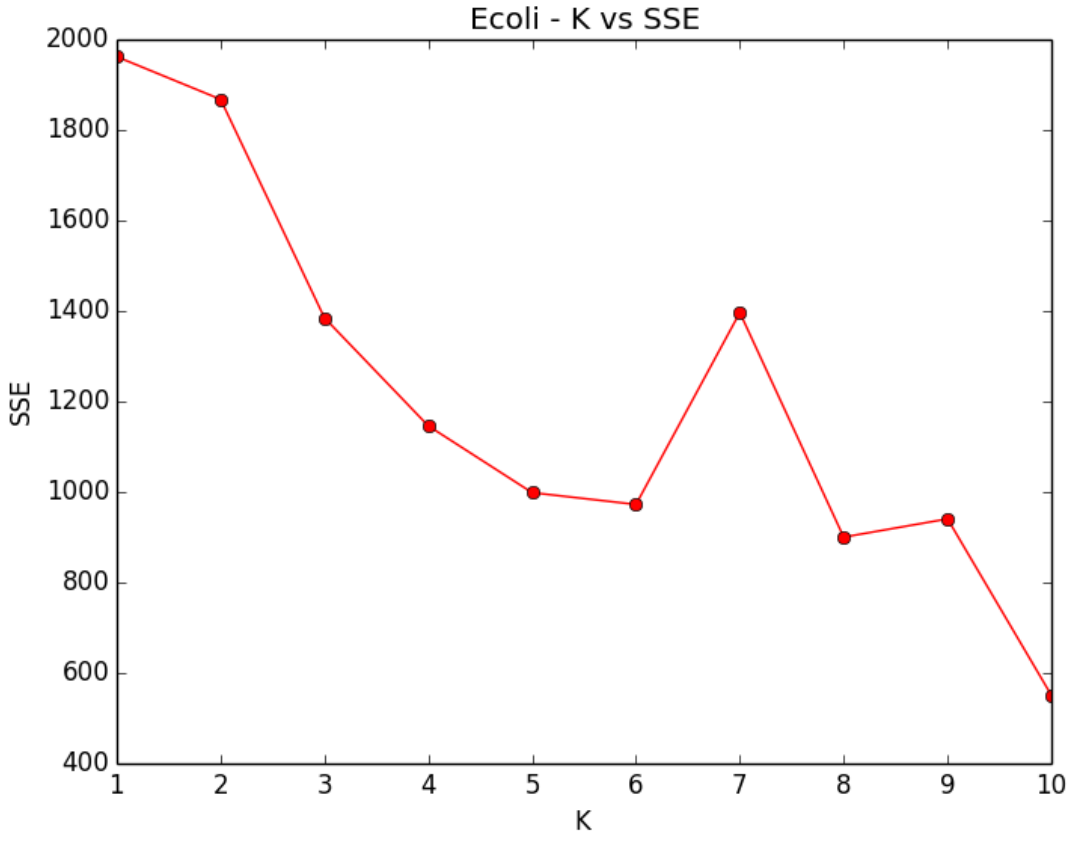
I believe that the NMI criterion is a better measure to evaluate the GMM model as each entry is the dataset has a probability measure associated to it describing how feasible it is to be a part of certain cluster. NMI compares the clusters by assigning a class to each entry and compares it to the actual class of the data.

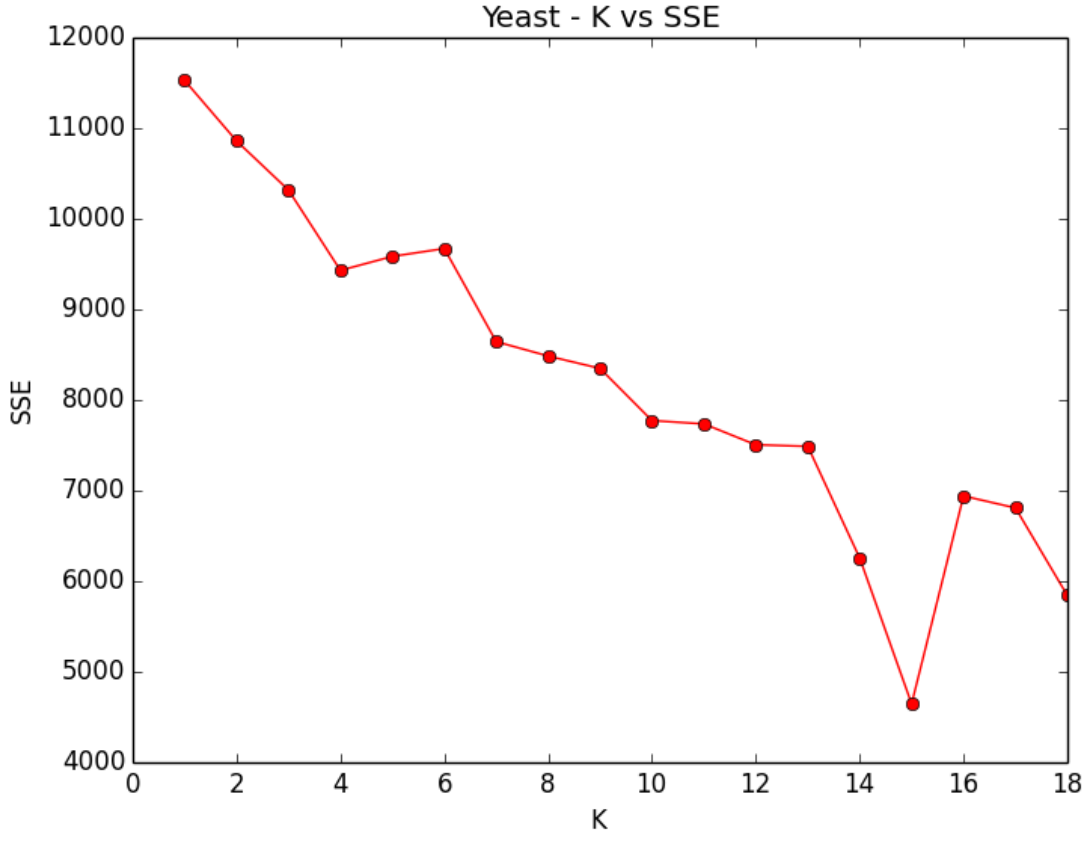
4.2.1

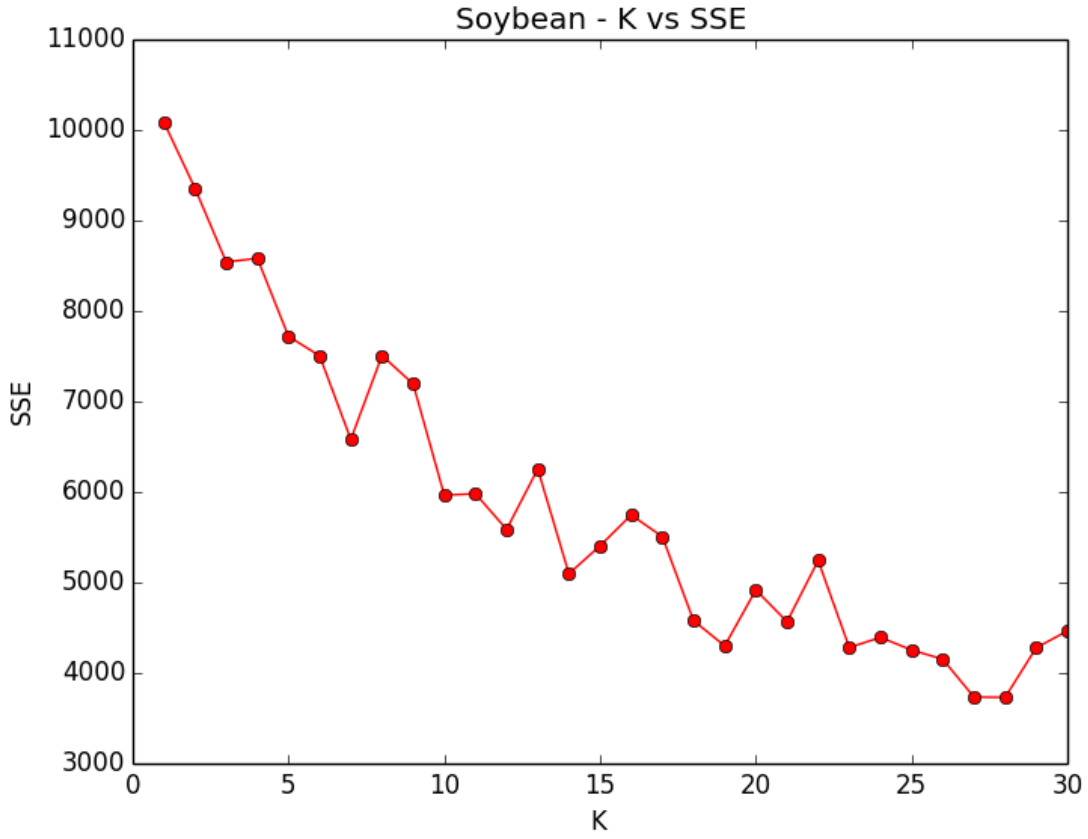



Vowels - K vs SSE
SSE
K
10000
9000
8000
7000
6000
5000
4000
0
5
10
15
20
25

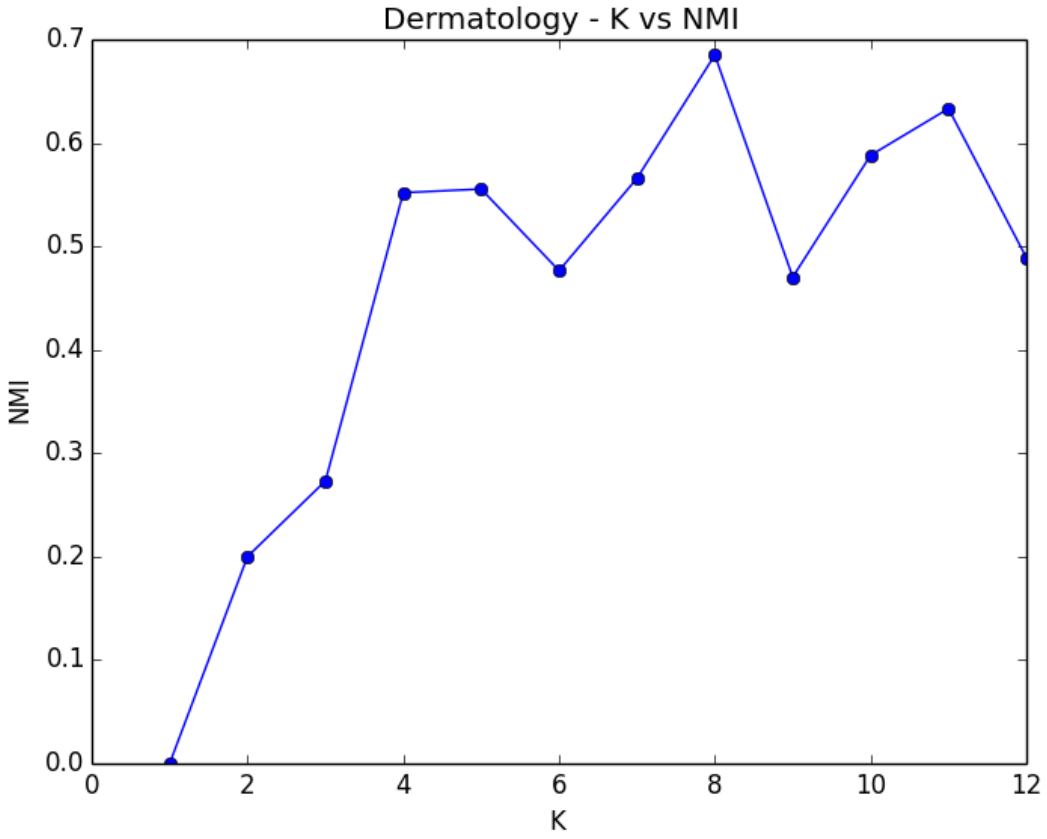

Glass - K vs SSE
SSE
K
2000
1800
1600
1400
1200
1000
800
0
2
4
6
8
10
12

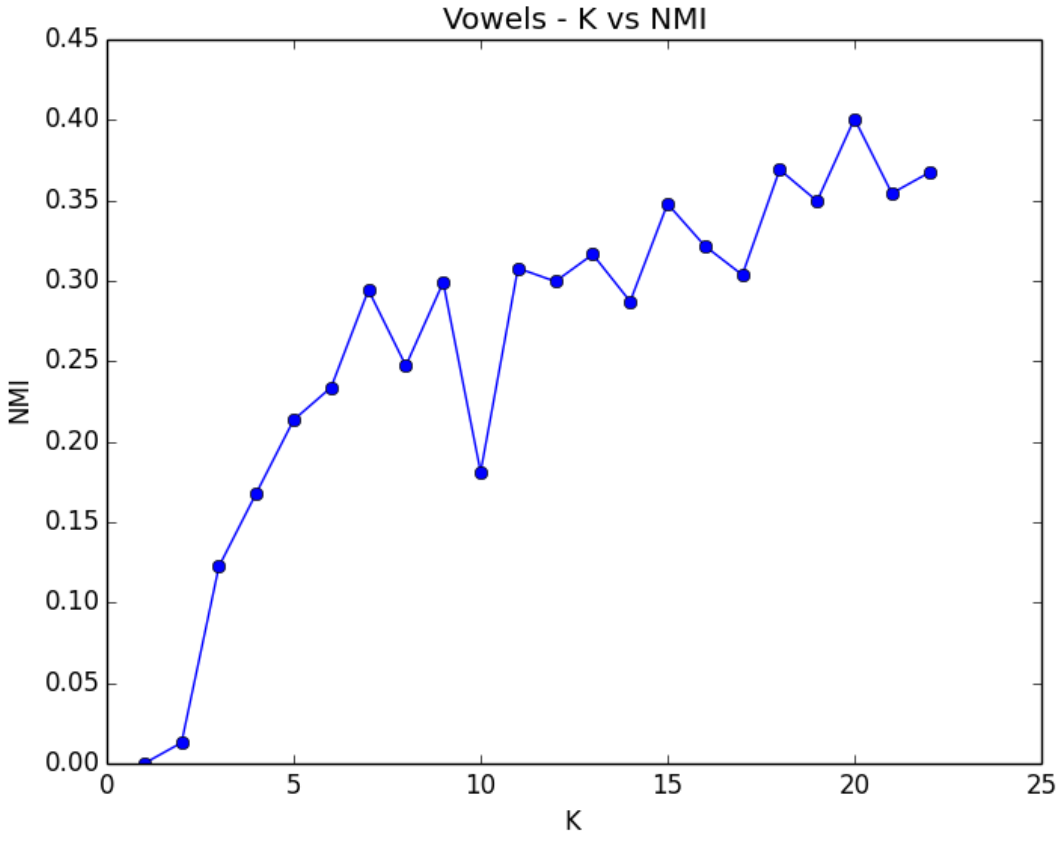

Ecoli - K vs SSE
SSE
K
2000
1800
1600
1400
1200
1000
800
600
400
1
2
3
4
5
6
7
8
9
10

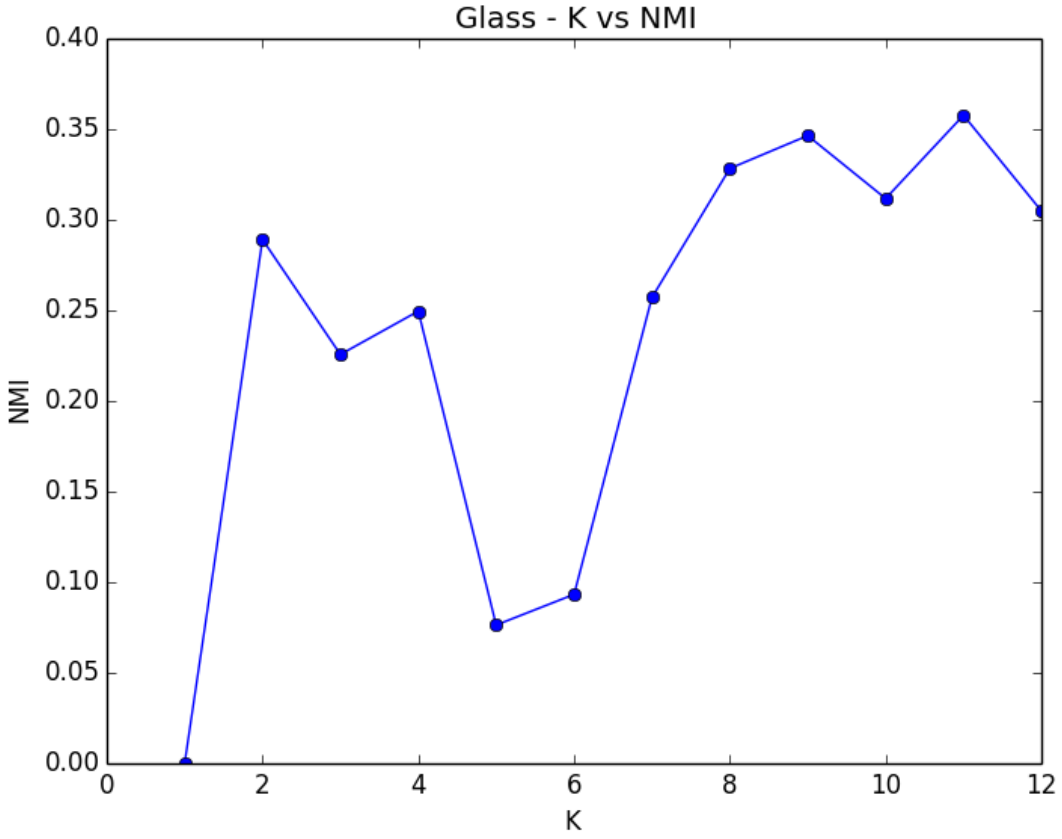

Yeast - K vs SSE
SSE
K
12000
11000
10000
9000
8000
7000
6000
5000
4000
0
2
4
6
8
10
12
14
16
18

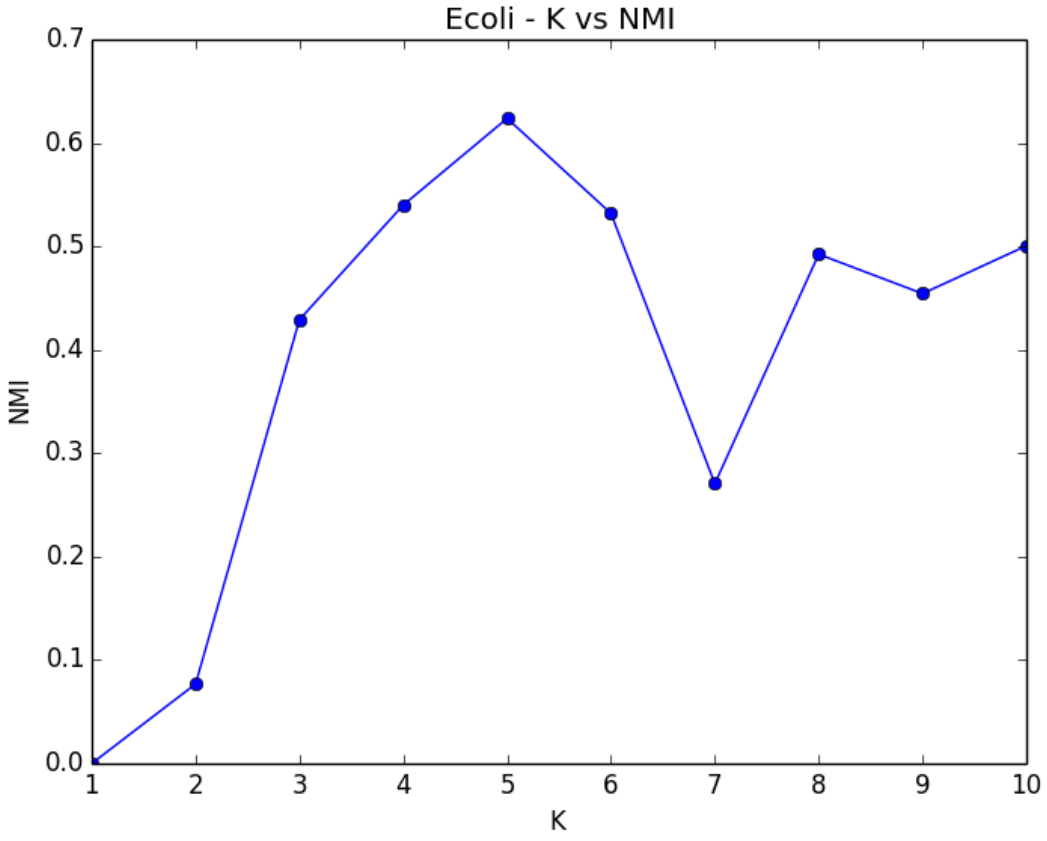

Soybean - K vs SSE
11000
10000
9000
8000
7000
6000
5000
4000
3000
SSE
0
5
10
15
20
25
30
K

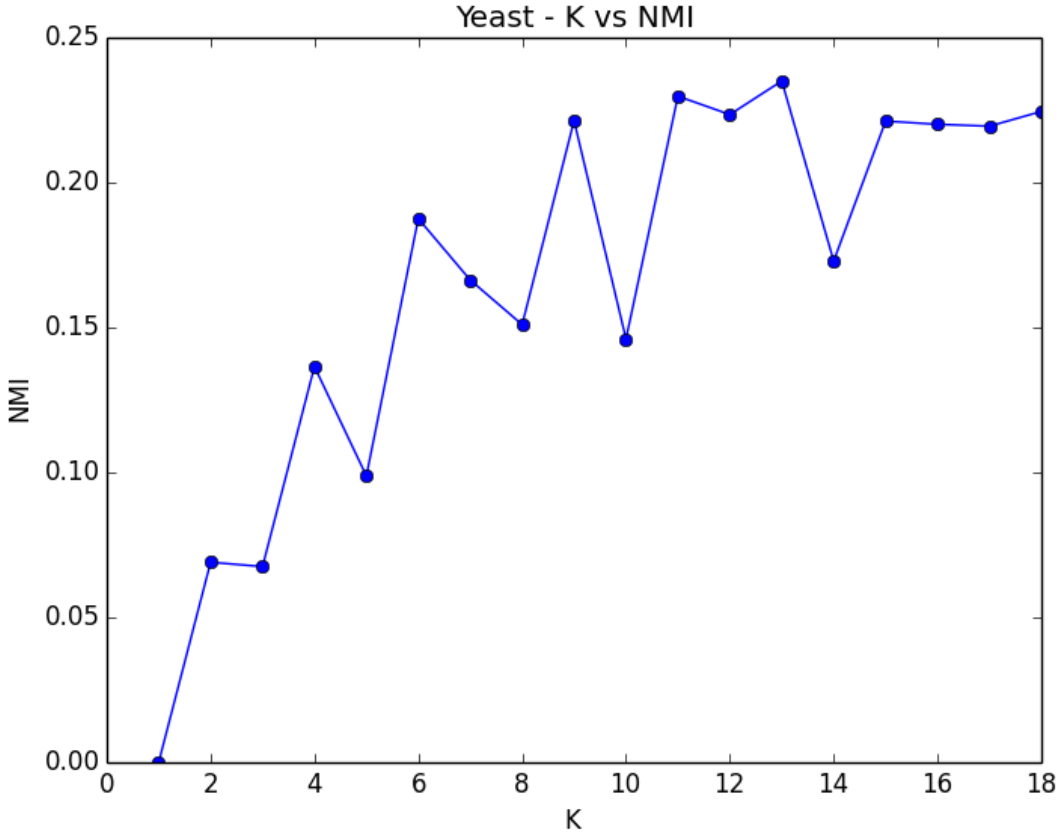
4.2.2

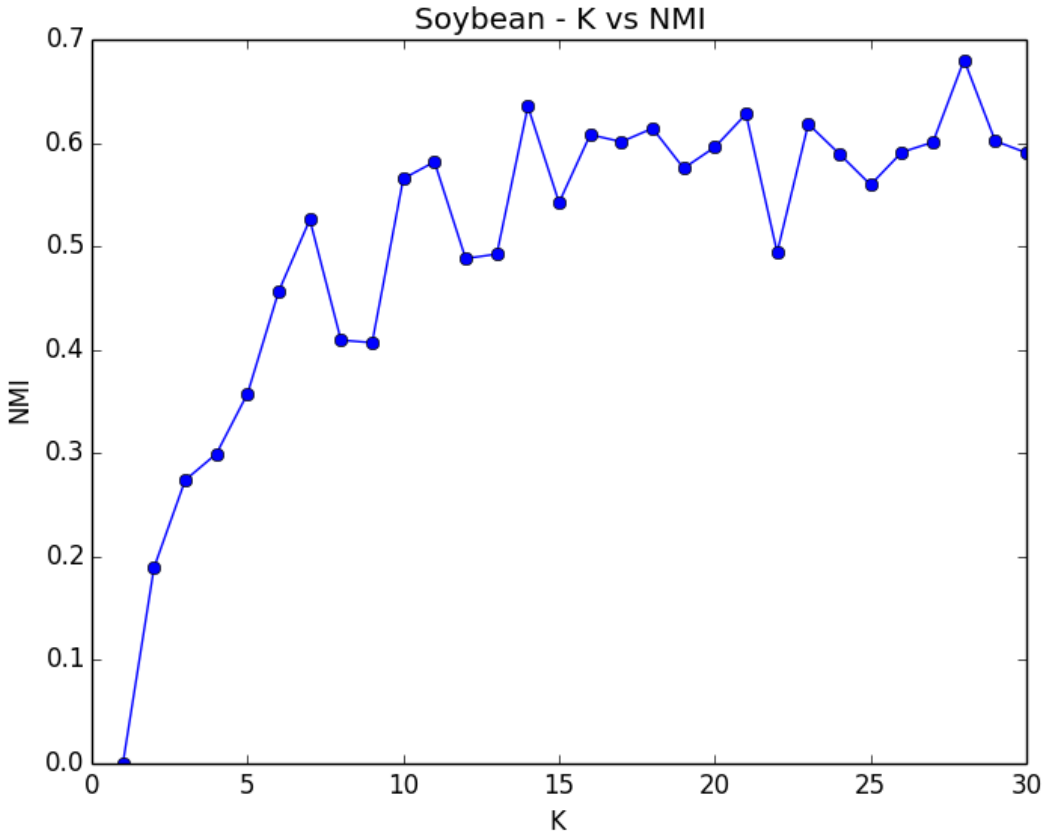



Vowels - K vs NMI
NMI
K
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0
5
10
15
20
25


Glass - K vs NMI
NMI
K
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
2
4
6
8
10
12


Ecoli - K vs NMI
NMI
K
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
3
4
5
6
7
8
9
10


Yeast - K vs NMI
NMI
K
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
12
14
16
18



4.2.3

Dataset	K	SSE	NMI
Dermatology	5	7999.950168	0.5553901139
Vowels	13	4801.037196	0.31644556
Glass	8	973.1674291	0.328329076
Ecoli	5	998.5598042	0.6237376276
Yeast	10	7772.727149	0.1460581299
Soybean	14	5097.019705	0.6352379054

4.2.4

Dataset	K	SSE	NMI
Dermatology	8	6017.416556	0.6850201991
Vowels	13	4801.037196	0.31644556
Glass	9	1021.669864	0.3464536446
Ecoli	5	998.5598042	0.6237376276
Yeast	9	8348.352787	0.2213501123
Soybean	14	5097.019705	0.6352379054

4.2.5

Dataset	K	SSE	NMI
Dermatology	6	8699.204659	0.4764883825
Vowels	11	5395.643102	0.3078697918
Glass	6	1354.748444	0.09326688599
Ecoli	5	998.5598042	0.6237376276
Yeast	9	8348.352787	0.2213501123
Soybean	15	5405.400575	0.5423750842

5 Comparing k-Means and GMM

5.1

Dataset	Algorithm
Dermatology	GMM
Vowels	k-Means
Glass	k-Means
Ecoli	GMM
Yeast	k-Means
Soybean	GMM

5.2

Yes clustering each dataset using the k-means and GMM gives us an idea on the type of data.