Prerequisite

All code and figures are contained in *https://github.com/feedlord18/CSDS313.git*

Problem 1

One plausible explanation is that the population from 1992 to 2002 increased. Therefore the number of high school students who have taken the SAT increased. With an increase in number of SAT test takers, the mean could shift towards the true population mean, with larger variance. We can also notice that the correlation still holds between GPA and SAT scores. Students who perform better in school (indicated by higher GPA) still generally performs better in the SAT in comparison with their respective groups.

Problem 2

We can safely assume that the variance among the Group I population is higher than Group II. Group II contains salary data points more centered around the mean. While Group I have a centered mean as well, however with 20% of the data points lying far away from the mean of 5100. By randomly selecting an individual from each group, we will have a larger chance selecting members who earn 4000 per month from Group I, and equal chance of either selecting a person who earns 4500 or 5500 from Group II.

Let the individual from Group B be B_i and Group A be A_i:

$$P(B_i \geq A_i) = P(B_i = 4500, A_i = 4000) + P(B_i = 5500, A_i = 4000)$$
$$P(B_i \geq A_i) = P(B_i = 4500) \cdot P(A_i = 4000) + P(B_i = 5500) \cdot P(A_i = 4000)$$
$$P(B_i \geq A_i) = 0.5 \cdot 0.8 + 0.5 \cdot 0.8$$
$$P(B_i \geq A_i) = 0.8$$

This is also a valid answer as the entire Group II population earns more than the 80% of the population in Group I. Also the 20% in Group I will always earn more than any individual in Group II.

Problem 3

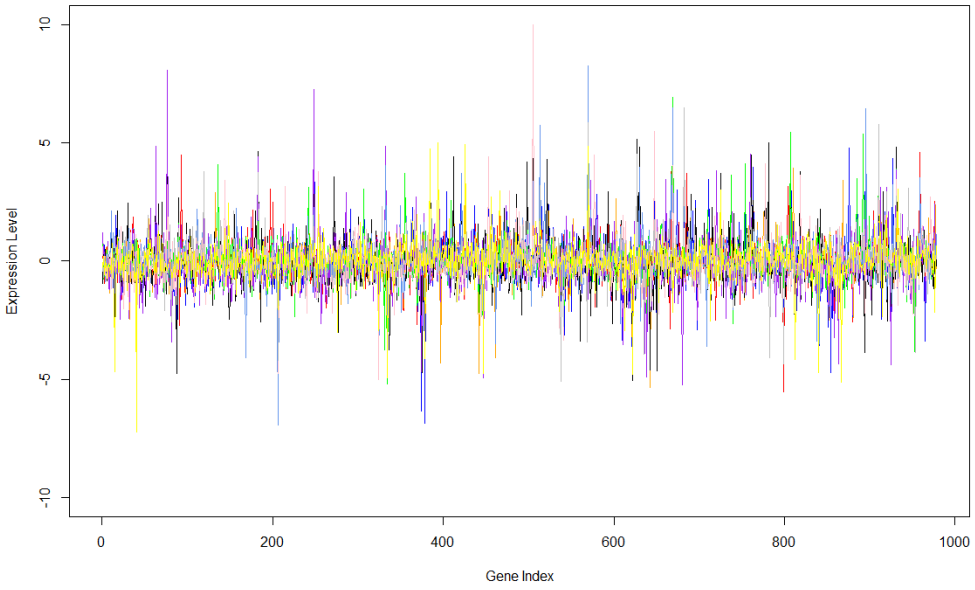
From looking at the bar graph, there is a significant issue. The y-axis begins at 65%, and the increase in prediction accuracy is about 5% using the new algorithm. The bar graph is structured in a way that visually the new method outperforms the old method by 2 folds, however it is simply a 5% in performance. There could be many ways to improve the bar

graph alone. First of all, normalizing the y-axis from 0 to 1 will show the relative performance of each model in a more reliable and accurate manner. Secondly, more metrics and curves could be used to show the performance increase, in terms of convergence time, precision-recall curve, etc.

Problem 4

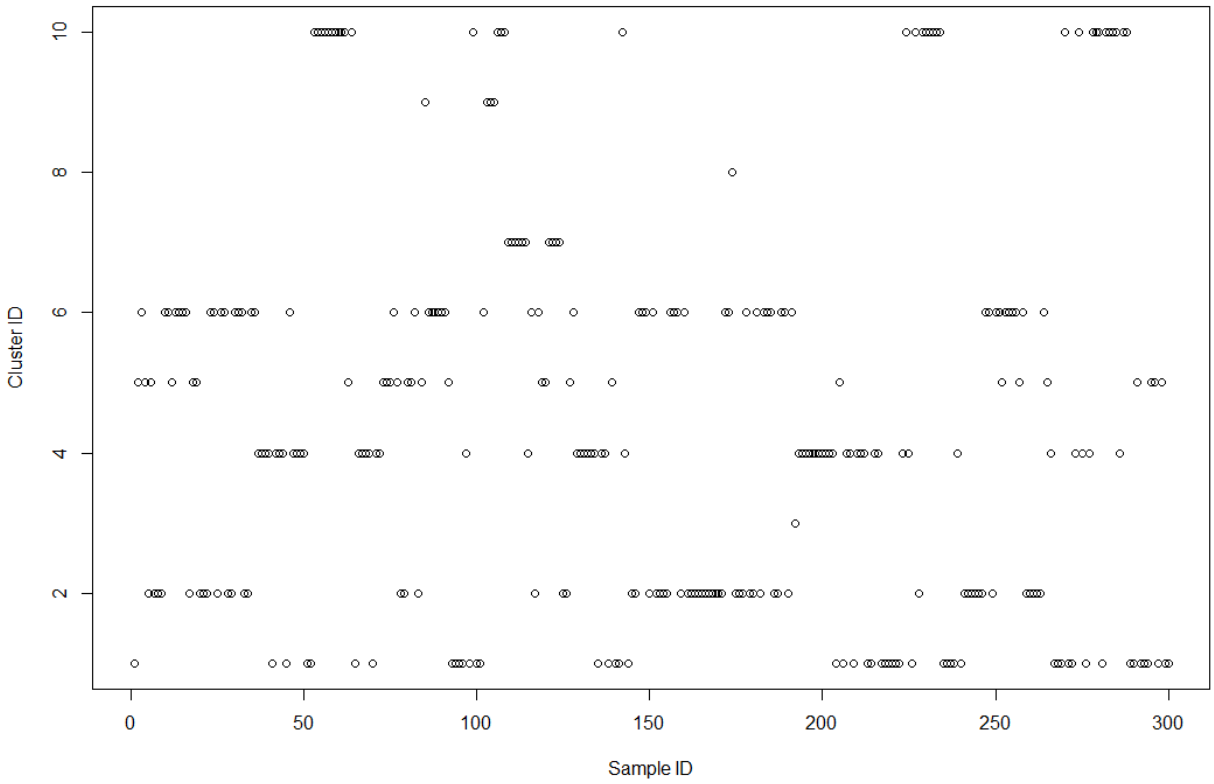
4a

We obtain the first 10 sample and plot the gene expressions with respective to the gene index. All 10 samples are colored differently.



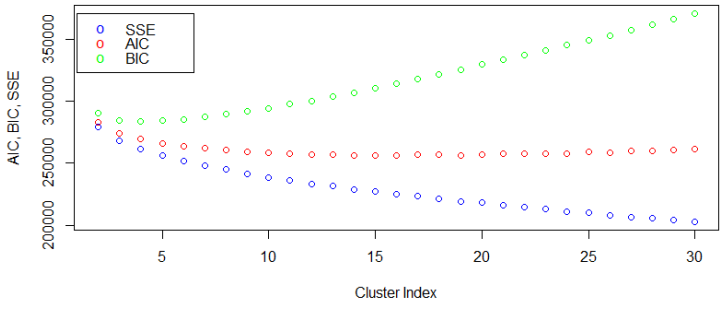
4b

Using R's 'factoextra' library, we can perform k-means clustering algorithm and determine the clustering distribution of all 300 data samples. The clusters and samples can be plotted to demonstrate cluster assignment by the algorithm. The total SSE of the clusters is 239035.4.

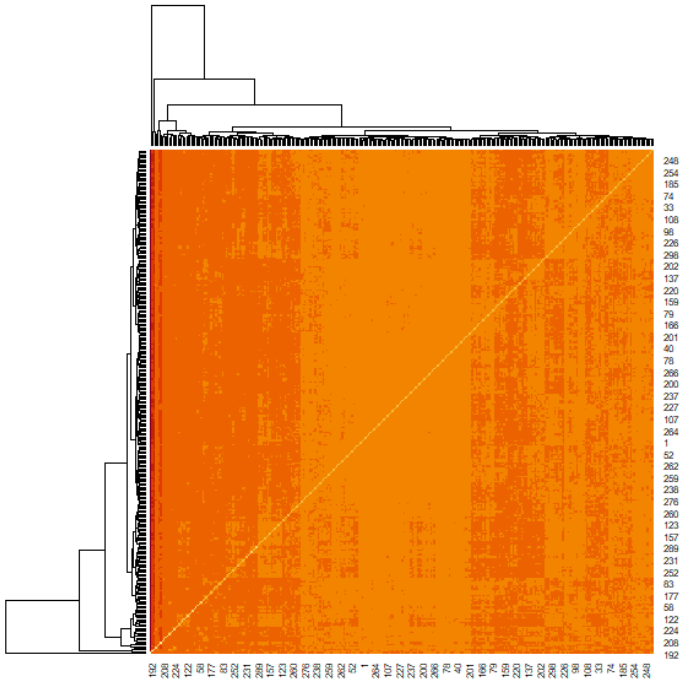


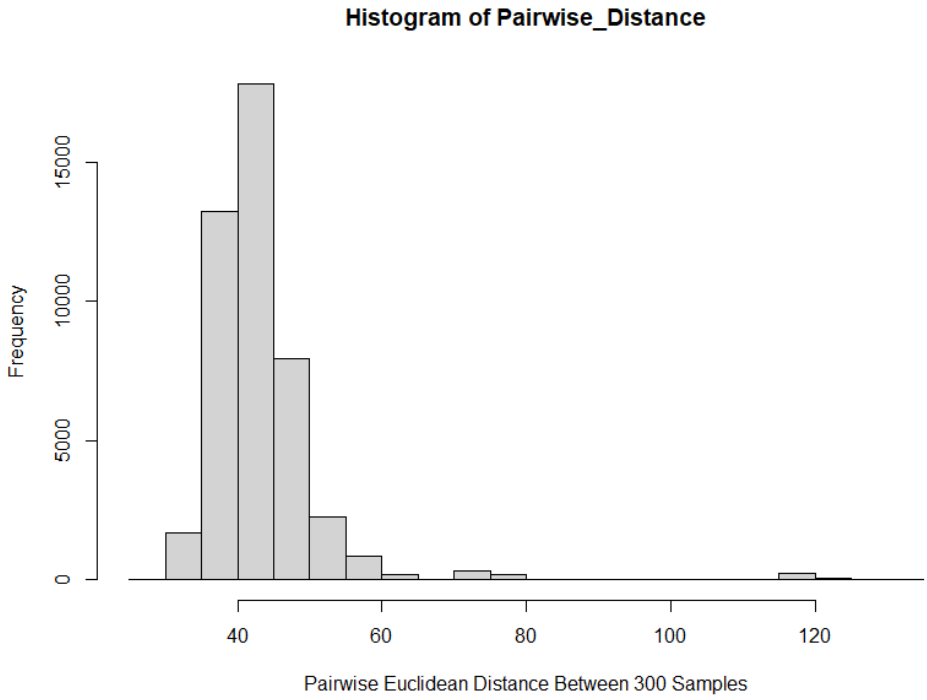
4c

Testing cluster center numbers from 2 to 30, we can see the following trend of SSE, AIC, and BIC values given a certain k. The cluster number which yields minimum AIC is 15 and minimum BIC is 3.

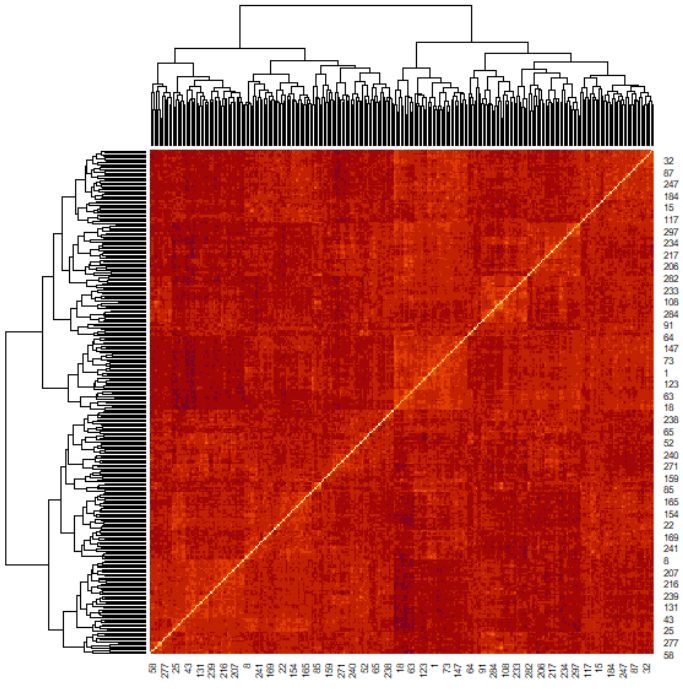


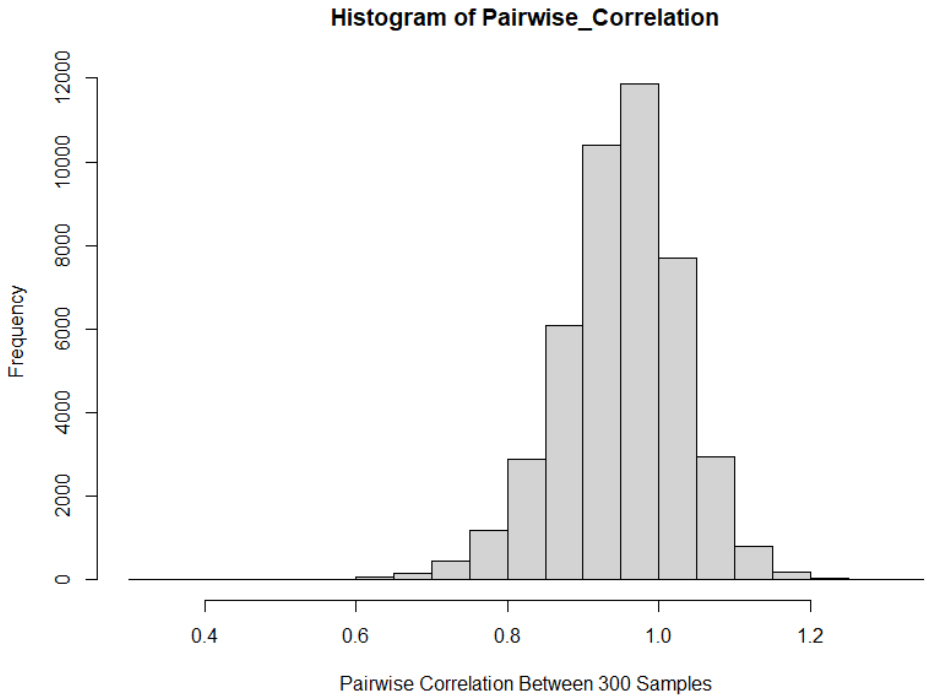
4d



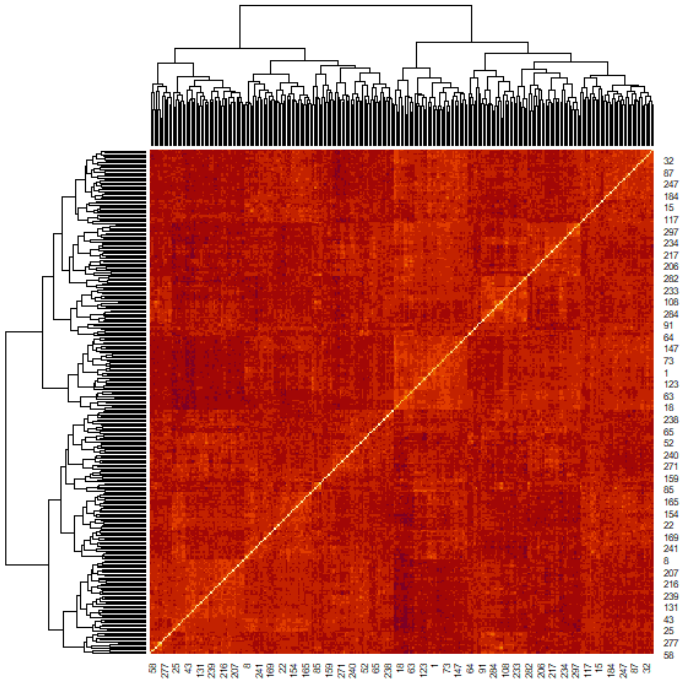

Histogram of Pairwise_Distance
Frequency
0
5000
10000
15000
40
60
80
100
120
Pairwise Euclidean Distance Between 300 Samples

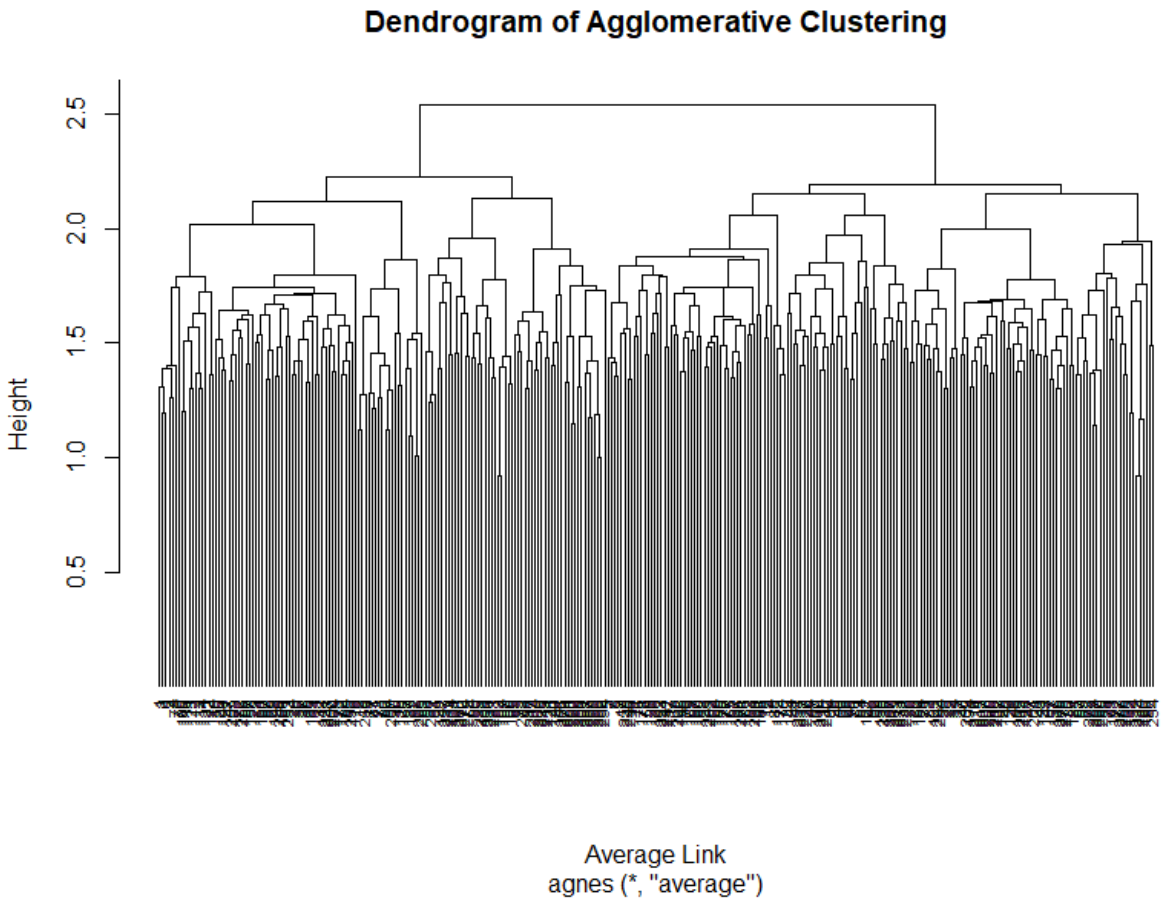
4e

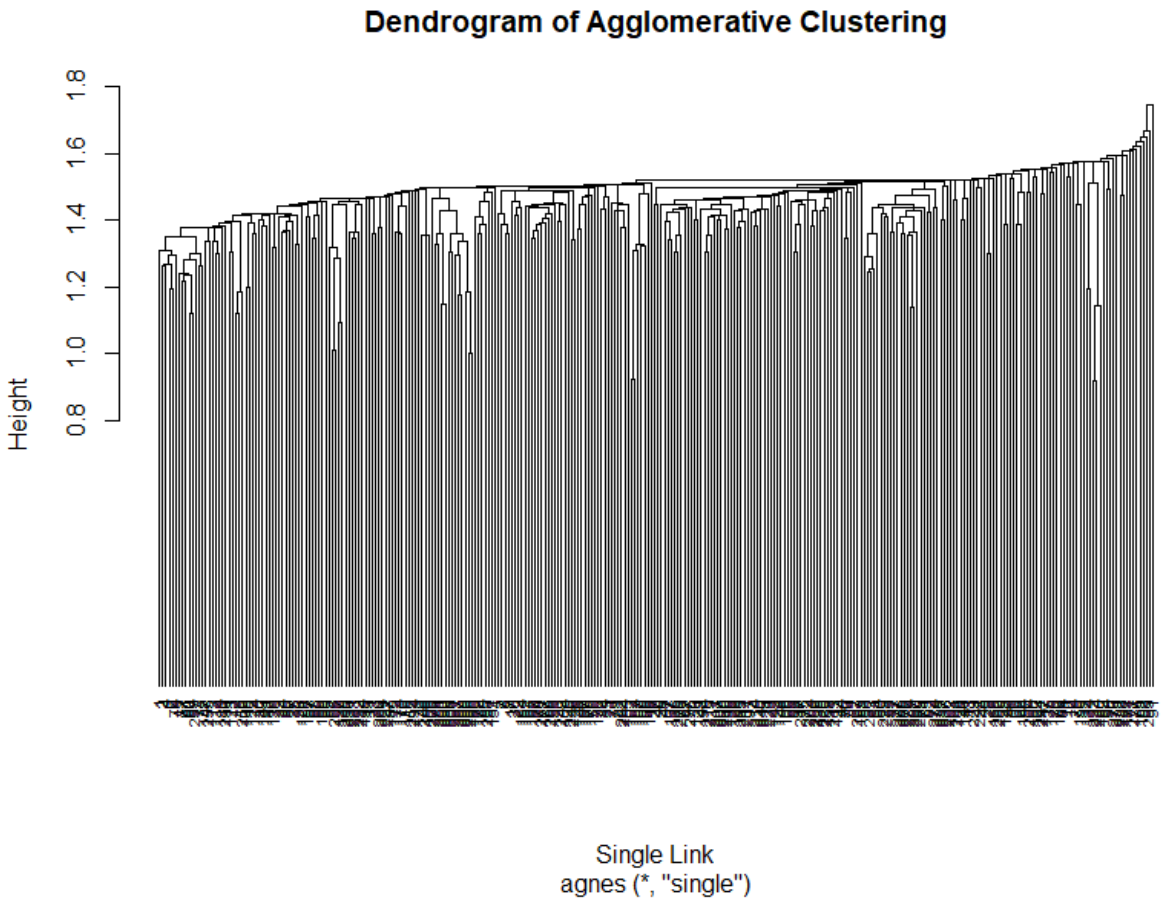


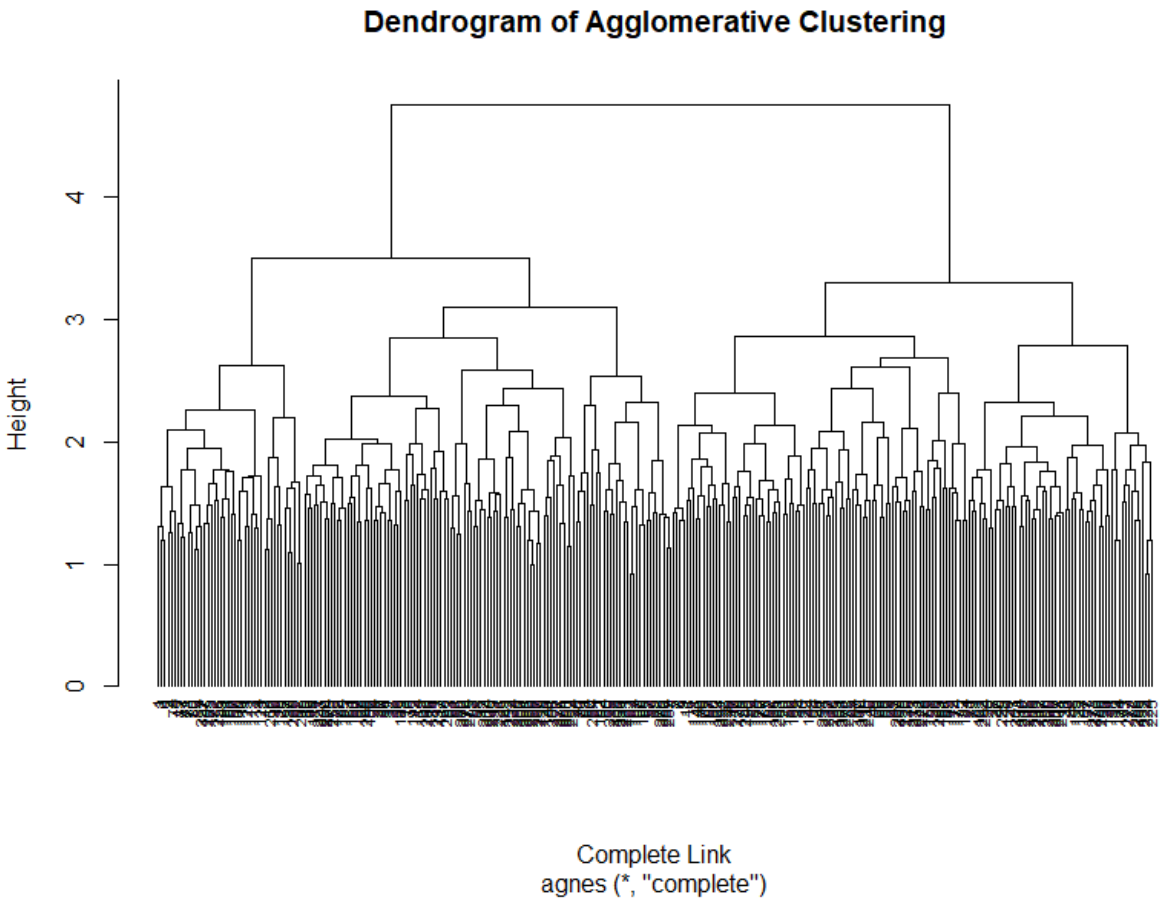

Histogram of Pairwise_Correlation
Frequency
0
2000
4000
6000
8000
10000
12000
0.4
0.6
0.8
1.0
1.2
Pairwise Correlation Between 300 Samples

4f


32
67
247
184
15
117
297
234
217
206
262
233
108
284
91
64
147
73
1
123
63
18
238
65
52
240
271
159
85
165
154
22
169
241
6
207
216
239
131
43
25
277
58
58 277 25 43 131 239 216 207 6 241 169 22 154 165 85 159 271 240 52 65 238 18 63 123 1 73 147 64 91 284 108 233 262 206 217 234 297 117 15 184 247 67 32


Dendrogram of Agglomerative Clustering
Height
0.5
1.0
1.5
2.0
2.5
Average Link
agnes (*, "average")


Dendrogram of Agglomerative Clustering
Height
0.8
1.0
1.2
1.4
1.6
1.8
Single Link
agnes (*, "single")


Dendrogram of Agglomerative Clustering
Height
0
1
2
3
4
Complete Link
agnes (*, "complete")