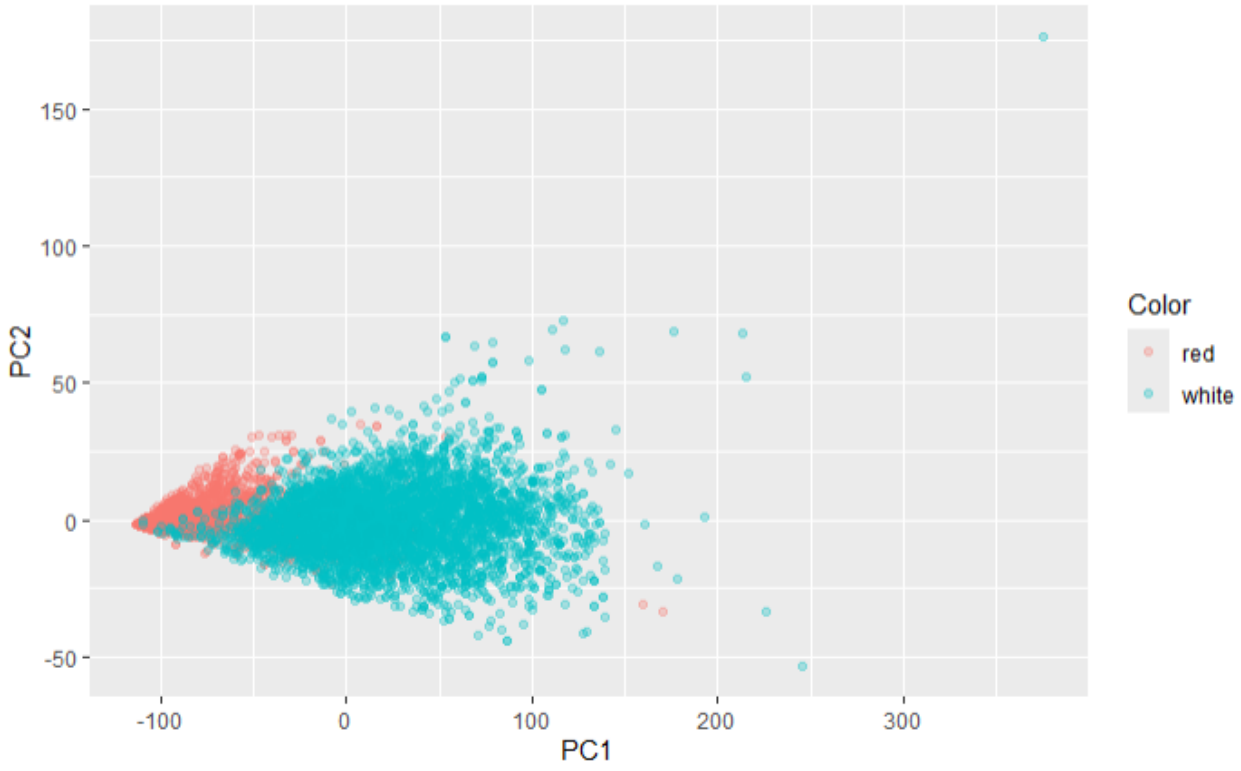
Problem 1

a.

```
[1] 6497   13
 fixed acidity    volatile acidity  citric acid     residual sugar     chlorides
 Min.    : 3.800   Min.    :0.0800   Min.    :0.0000   Min.    : 0.600   Min.    :0.00900
 1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800   1st Qu.:0.03800
 Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000   Median :0.04700
 Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443   Mean   :0.05603
 3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100   3rd Qu.:0.06500
 Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800   Max.   :0.61100
 free sulfur dioxide total sulfur dioxide    density             pH          sulphates
 Min.    :  1.00      Min.    :  6.0        Min.    :0.9871   Min.    :2.720   Min.    :0.2200
 1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300
 Median : 29.00      Median :118.0        Median :0.9949   Median :3.210   Median :0.5100
 Mean   : 30.53      Mean   :115.7        Mean   :0.9947   Mean   :3.219   Mean   :0.5313
 3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000
 Max.   :289.00      Max.   :440.0        Max.   :1.0390   Max.   :4.010   Max.   :2.0000
    alcohol          quality          type
 Min.    : 8.00   Min.    :3.000   Min.    :0.0000
 1st Qu.: 9.50   1st Qu.:5.000   1st Qu.:0.0000
 Median :10.30   Median :6.000   Median :0.0000
 Mean   :10.49   Mean   :5.818   Mean   :0.2461
 3rd Qu.:11.30   3rd Qu.:6.000   3rd Qu.:0.0000
 Max.   :14.90   Max.   :9.000   Max.   :1.0000
```

b.



c. Based on the visualization, I believe decision tree will be the best performing model because of its ability to capture more robust patterns in the data. There is not enough of a distinction for SVM to achieve a high accuracy score and because the data is highly concentrated, KNN will struggle to account for noise.

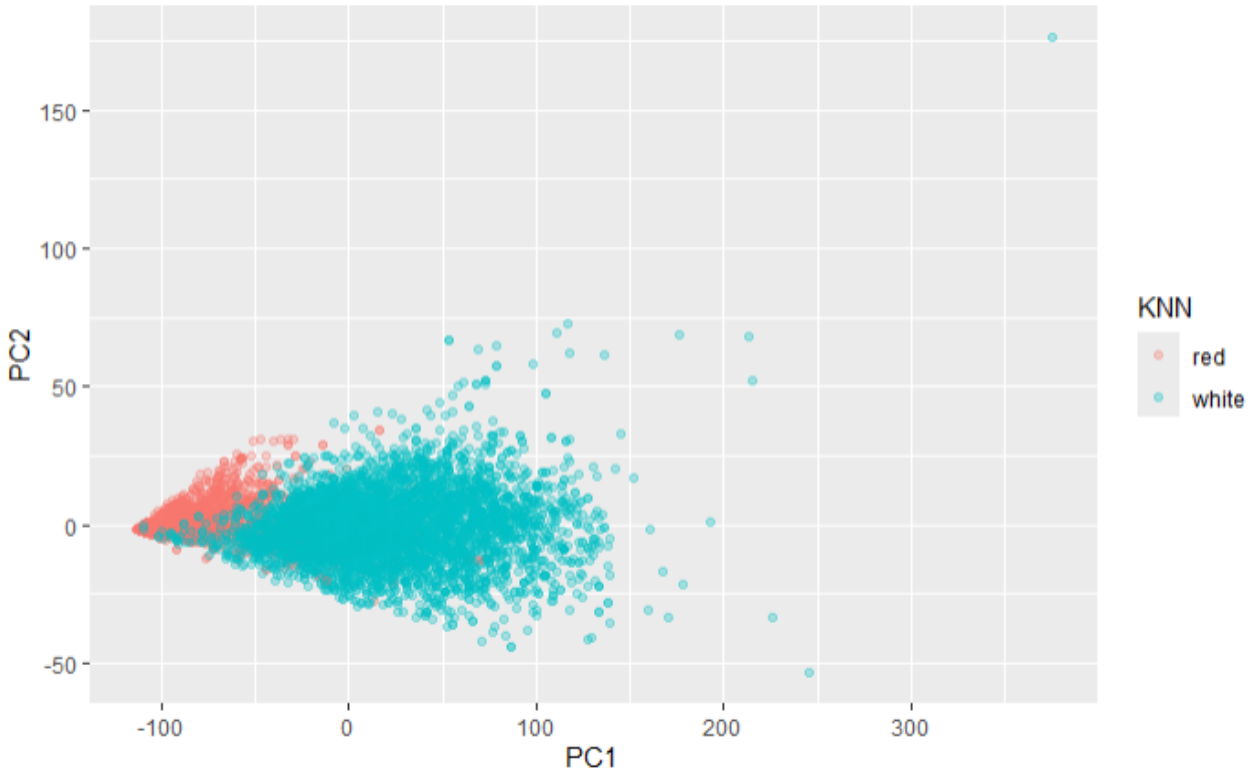
d.

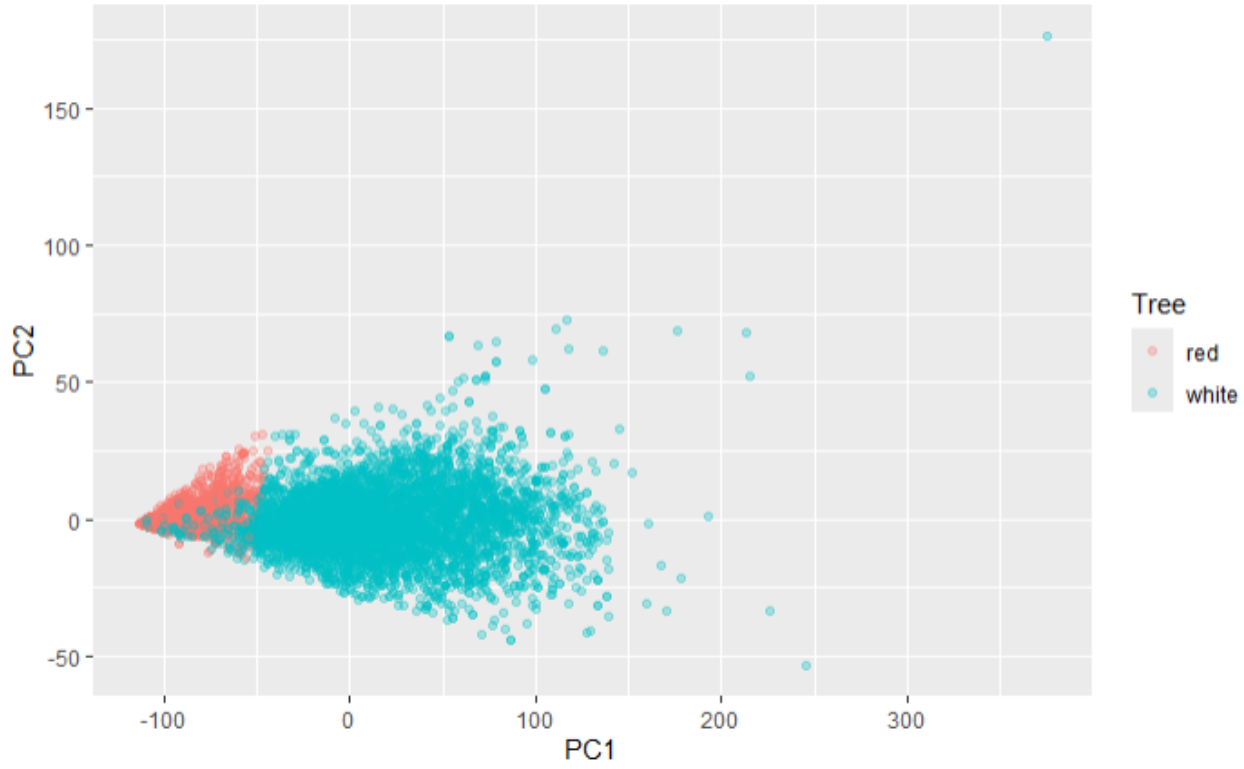
```
k-Nearest Neighbors

6497 samples
  12 predictor
   2 classes: 'red', 'white'

Pre-processing: centered (12), scaled (12)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 5847, 5847, 5847, 5849, 5847, 5847, ...
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  5  0.9924566  0.9796198
  7  0.9930724  0.9813010
  9  0.9929190  0.9809074

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 7.

CART

6497 samples
  12 predictor
   2 classes: 'red', 'white'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 5847, 5847, 5848, 5847, 5847, 5847, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.06253909  0.9538324  0.8697804
  0.06754221  0.9308934  0.8036980
  0.70043777  0.8371435  0.3866620

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.06253909.
```

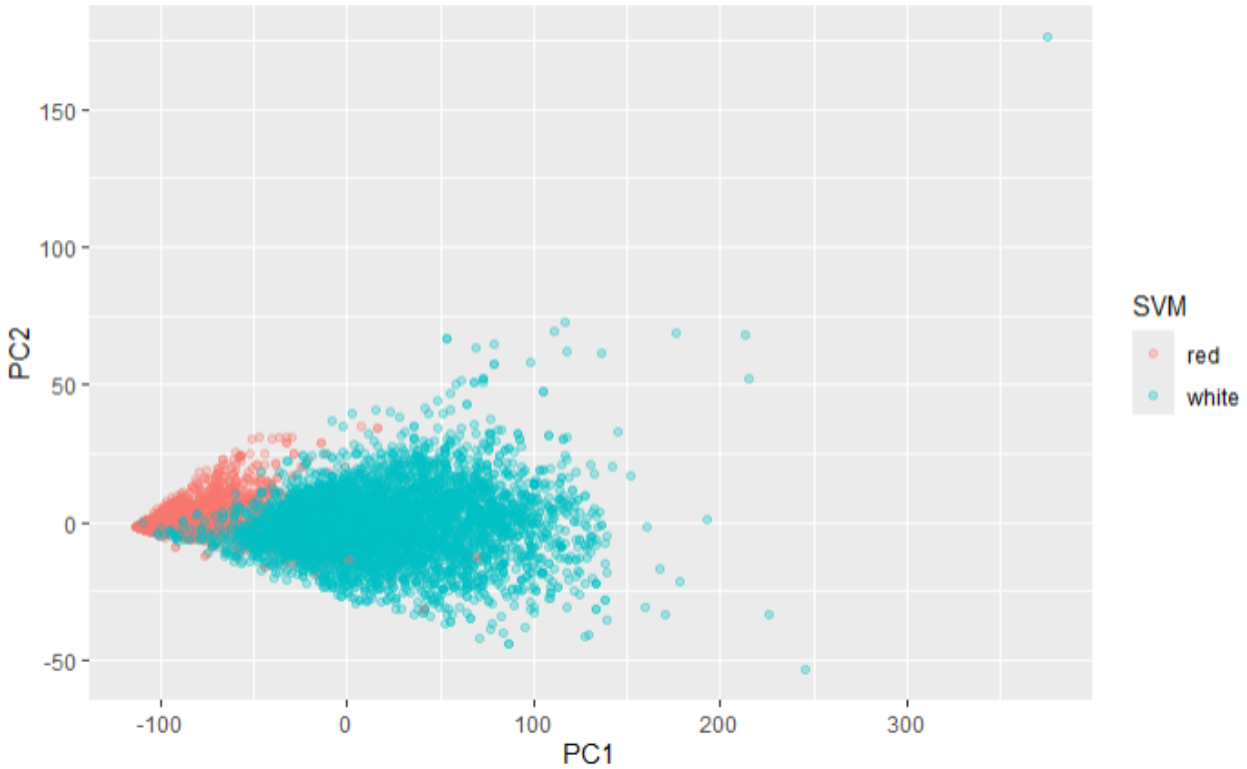
```
Support Vector Machines with Linear Kernel

6497 samples
  12 predictor
   2 classes: 'red', 'white'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 6497, 6497, 6497, 6497, 6497, 6497, ...
Resampling results:

  Accuracy   Kappa
  0.9949203  0.9863384

Tuning parameter 'C' was held constant at a value of 1
```

SVM has the highest accuracy with 99.49%. I did not expect SVM to perform this well, however because the data is highly concentrated, it makes sense that SVM was able to distinguish between the two categories at such a high accuracy. Overall, all the models performed much better than I expected.

e.







The SVM and KNN results are essentially identical, with only a handful of points being different, though there is no obvious reason for these differences. The decision tree plot shows the most

substantial difference, with the model determining -50 PC1 to be the primary split point for predicting red and white wines, with all points > -50 PC1 being white wines. SVM and KNN have a few scattered red wine predictions in the areas that the decision tree determined were all white wines, leading to a better performance of SVM and KNN over the tree.

Problem 2

a.

city.COOL <dbl>	city.DIAMOND_SPRINGS <dbl>	city.EL_DORADO <dbl>	city.EL_DORADO_HILLS <dbl>	city.ELK_GROVE <dbl>
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

6 rows | 7-11 of 111 columns

b. Cosine similarity would be the best choice to deal with the high dimensionality in the data set.

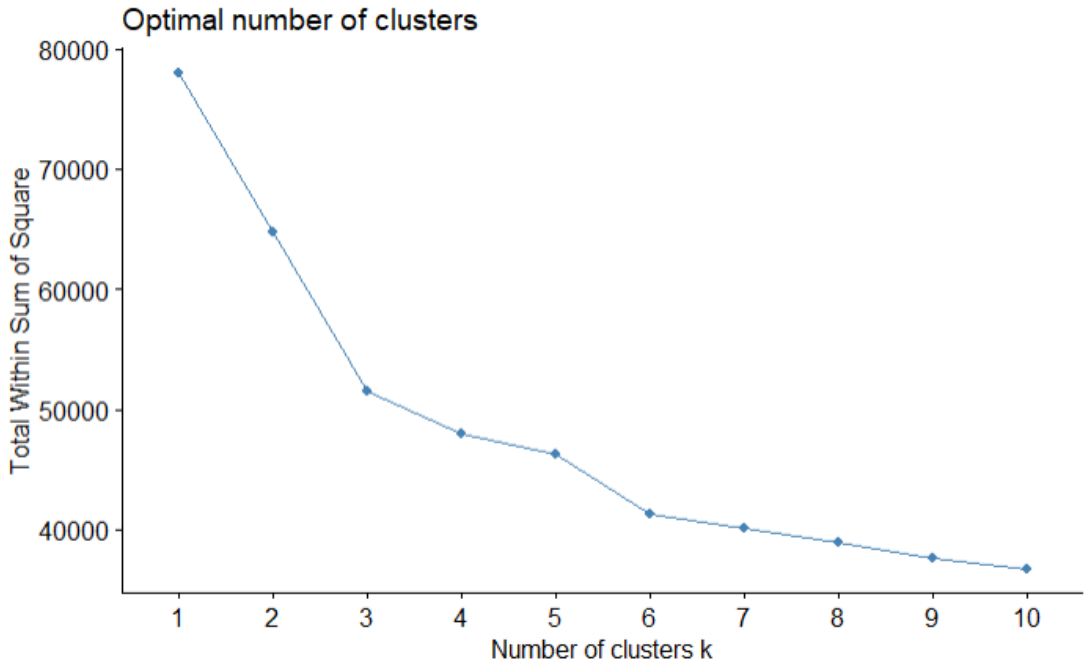
c.

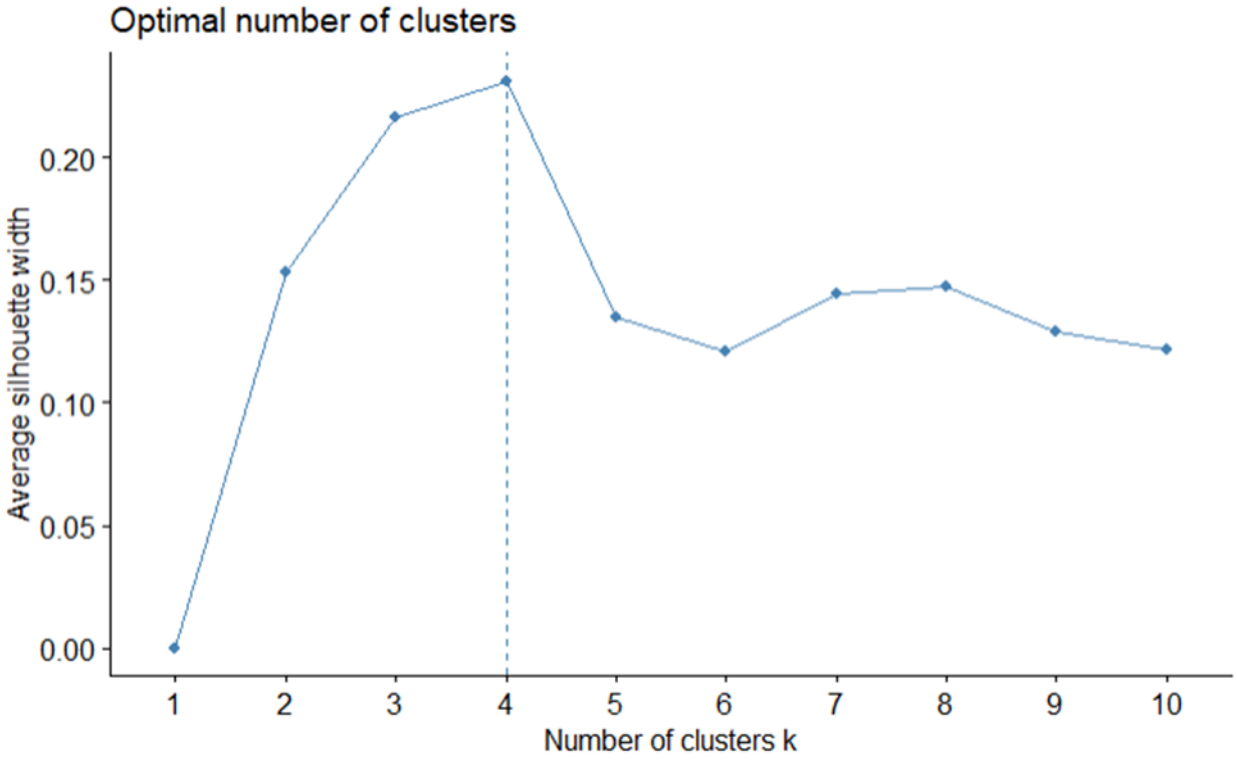
```
932 samples
111 predictors
  3 classes: 'Condo', 'Multi_Family', 'Residential'

Pre-processing: centered (111), scaled (111)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 840, 839, 838, 839, 838, 839, ...
Resampling results across tuning parameters:

  kmax  kernel       distance  Accuracy   Kappa
  3     rectangular  1         0.9452971  0.4366071
  3     rectangular  2         0.9410189  0.4151466
  3     rectangular  3         0.9410189  0.4151466
  3     cos          1         0.9495748  0.5325135
  3     cos          2         0.9485112  0.5396641
  3     cos          3         0.9485112  0.5322278
  4     rectangular  1         0.9452971  0.4366071
  4     rectangular  2         0.9410189  0.4108842
  4     rectangular  3         0.9410189  0.4151466
  4     cos          1         0.9527777  0.5485082
  4     cos          2         0.9474242  0.5175892
  4     cos          3         0.9485112  0.5322278
  5     rectangular  1         0.9399207  0.3410622
  5     rectangular  2         0.9356425  0.3189963
  5     rectangular  3         0.9388683  0.3571331
  5     cos          1         0.9517024  0.5348000
  5     cos          2         0.9474242  0.5175892
  5     cos          3         0.9485112  0.5322278
  6     rectangular  1         0.9399207  0.3410622
  6     rectangular  2         0.9356425  0.3189963
  6     rectangular  3         0.9388683  0.3571331
  6     cos          1         0.9517024  0.5348000
  6     cos          2         0.9474242  0.5175892
  6     cos          3         0.9485112  0.5322278
  7     rectangular  1         0.9399207  0.3410622
  7     rectangular  2         0.9356425  0.3189963
  7     rectangular  3         0.9388683  0.3571331
  7     cos          1         0.9517024  0.5348000
  7     cos          2         0.9474242  0.5175892
  7     cos          3         0.9485112  0.5322278

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were kmax = 4, distance = 1 and kernel = cos.
```

We tuned for k values 3-7, rectangular and cosine-based distance functions, and Minkowski distance h values 1-3. The chosen values are k =4 using cosine distance with Minkowski h = 1, or Manhattan distance. The final accuracy is 95.28%.

Problem 3

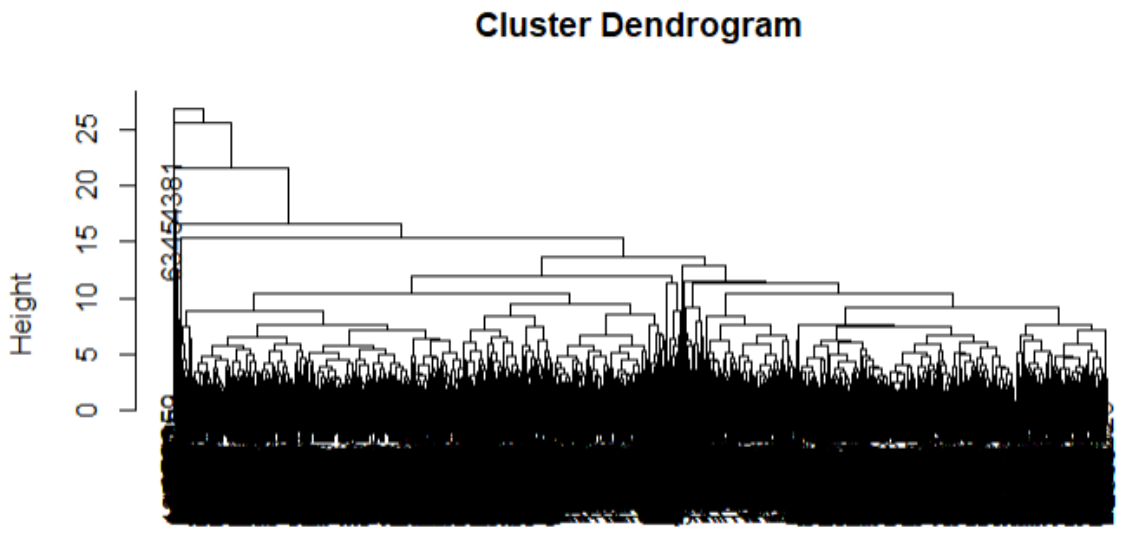
a.

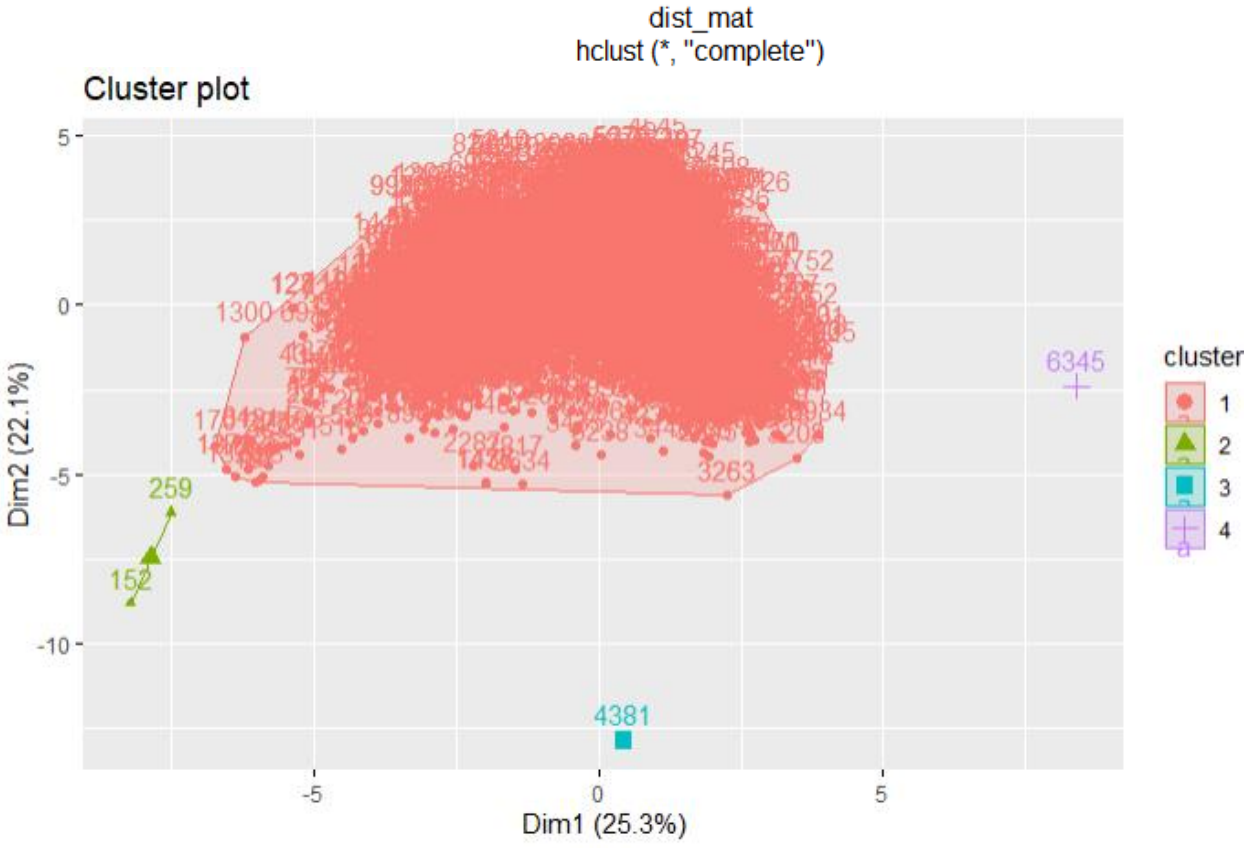




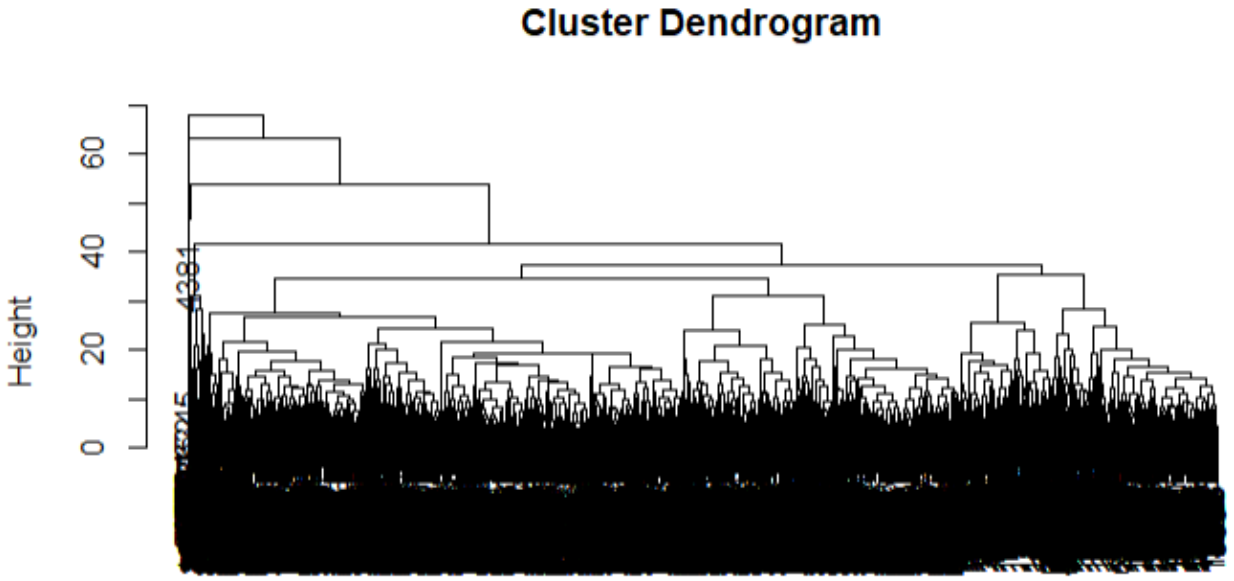
Based on the WSS, the optimal number of clusters is 6 but the silhouette suggests 4 clusters. Looking at both methods, 4 seems to be justifiable in both, whereas 6 has a much lower score in the silhouette method, so we will go with 4 clusters.

b. 1. Distance = Euclidean, Linkage = Complete

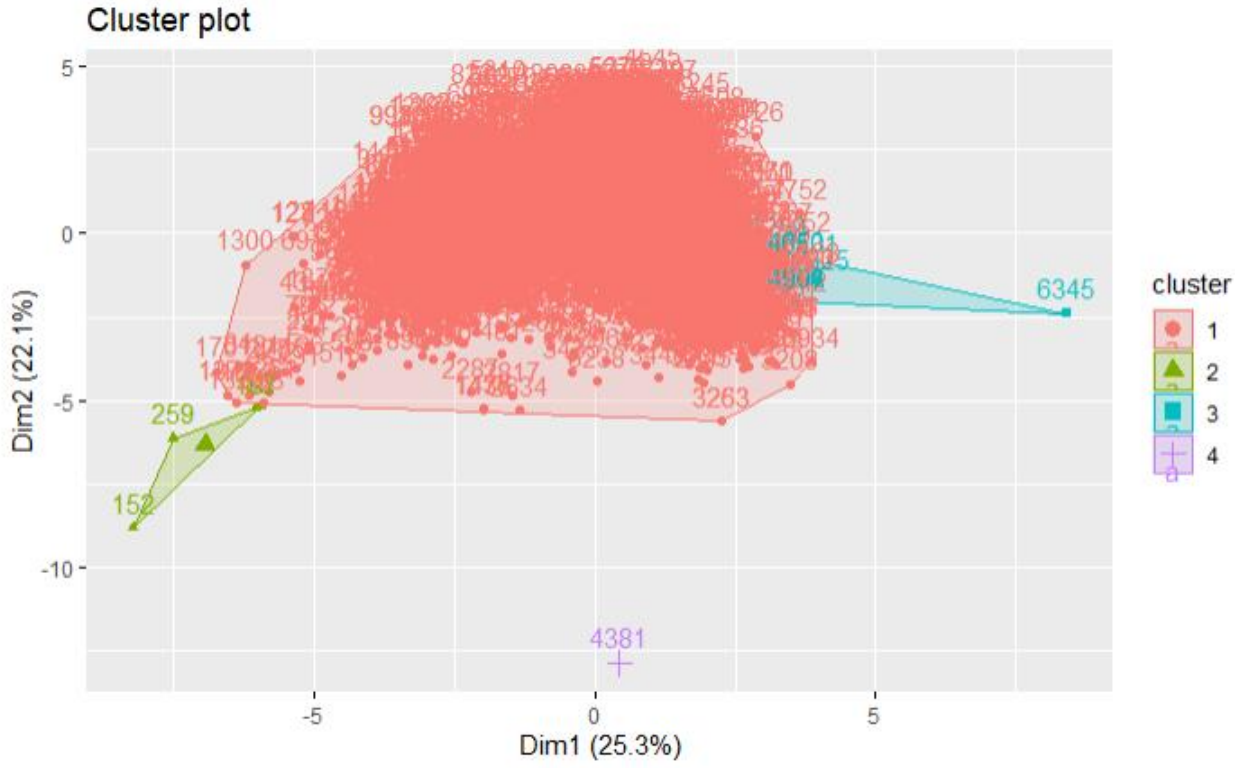




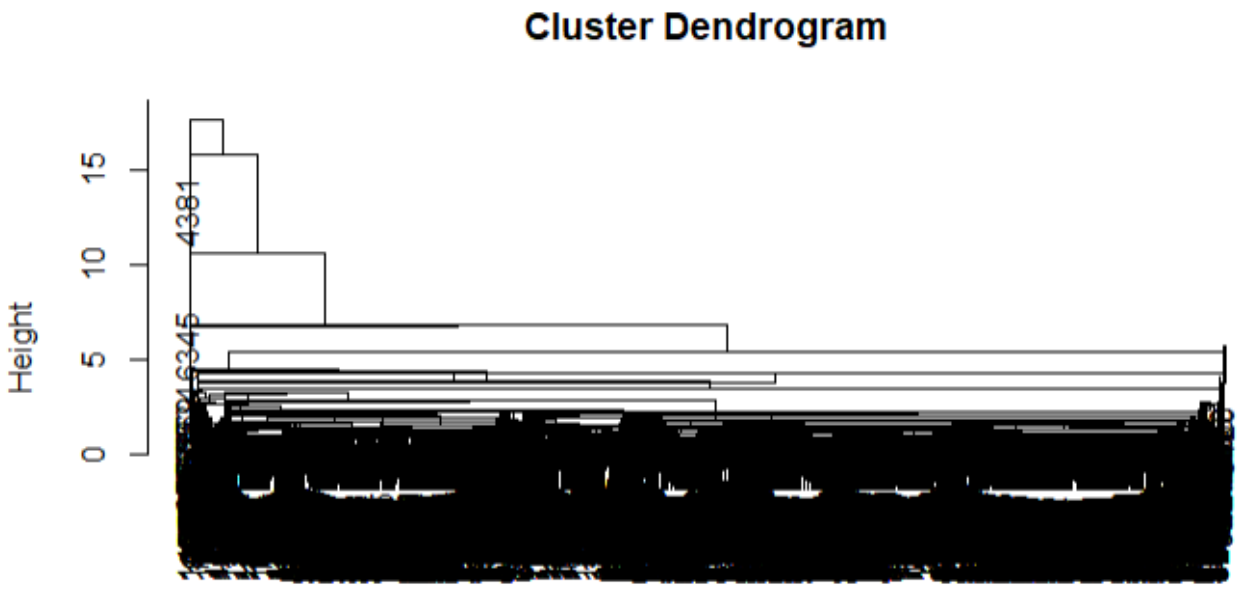
2. Distance = Manhattan, Linkage = Median

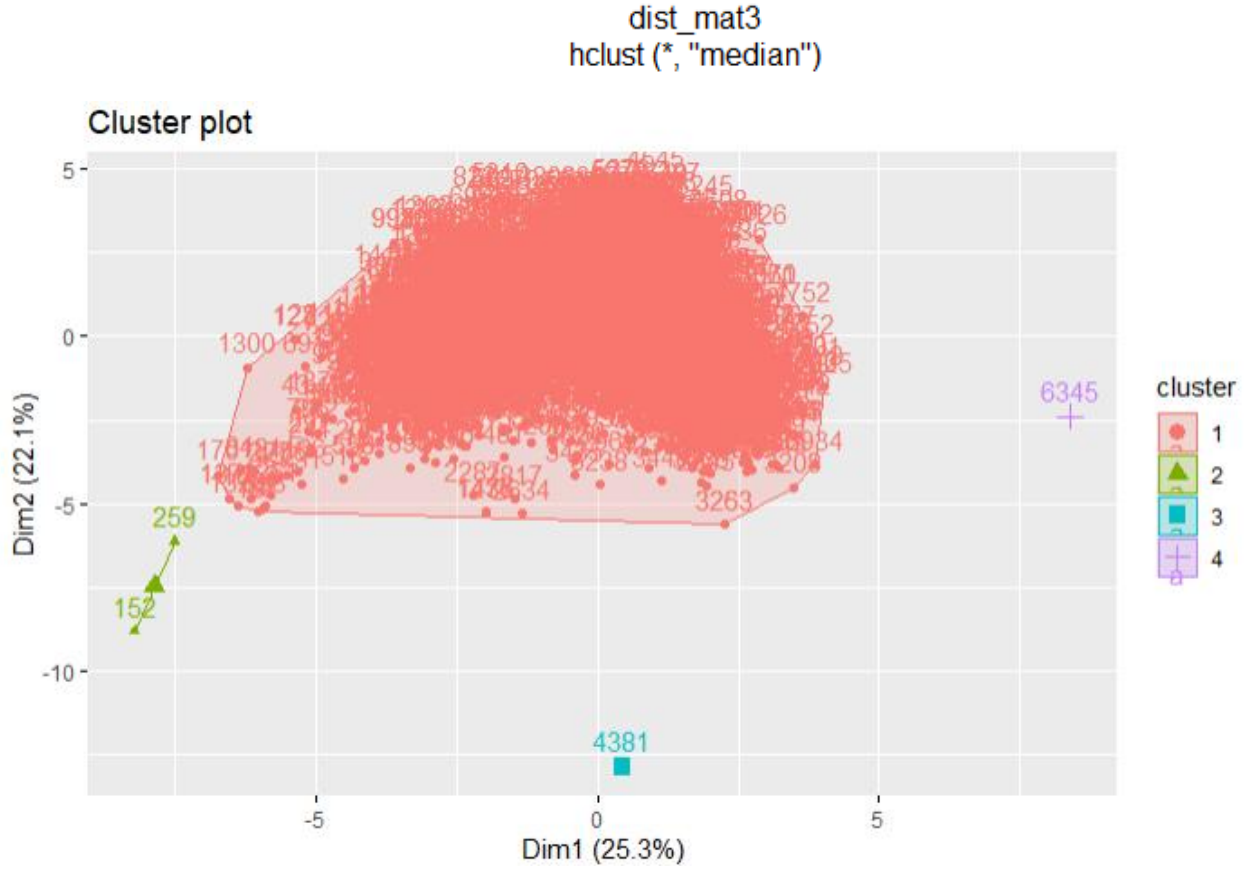


dist_mat2
hclust (*, "complete")

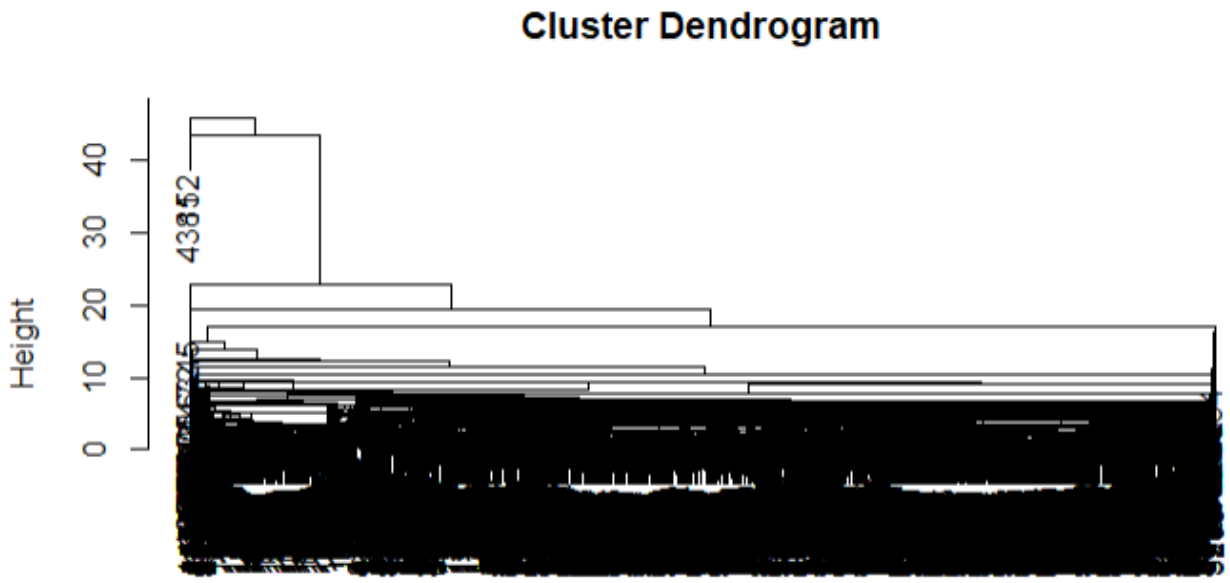


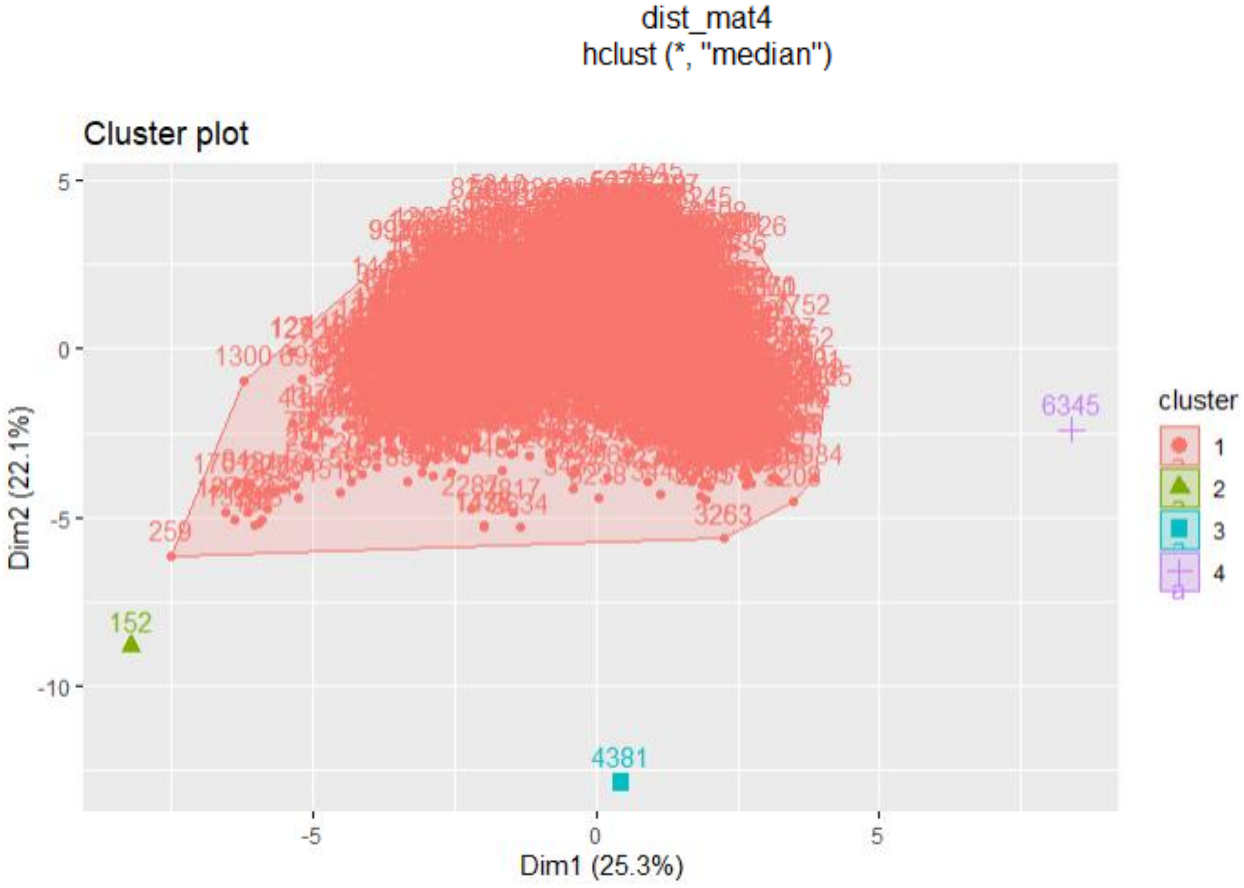
3. Distance = Euclidean, Linkage = Complete





4. Distance = Manhattan, Linkage = Median





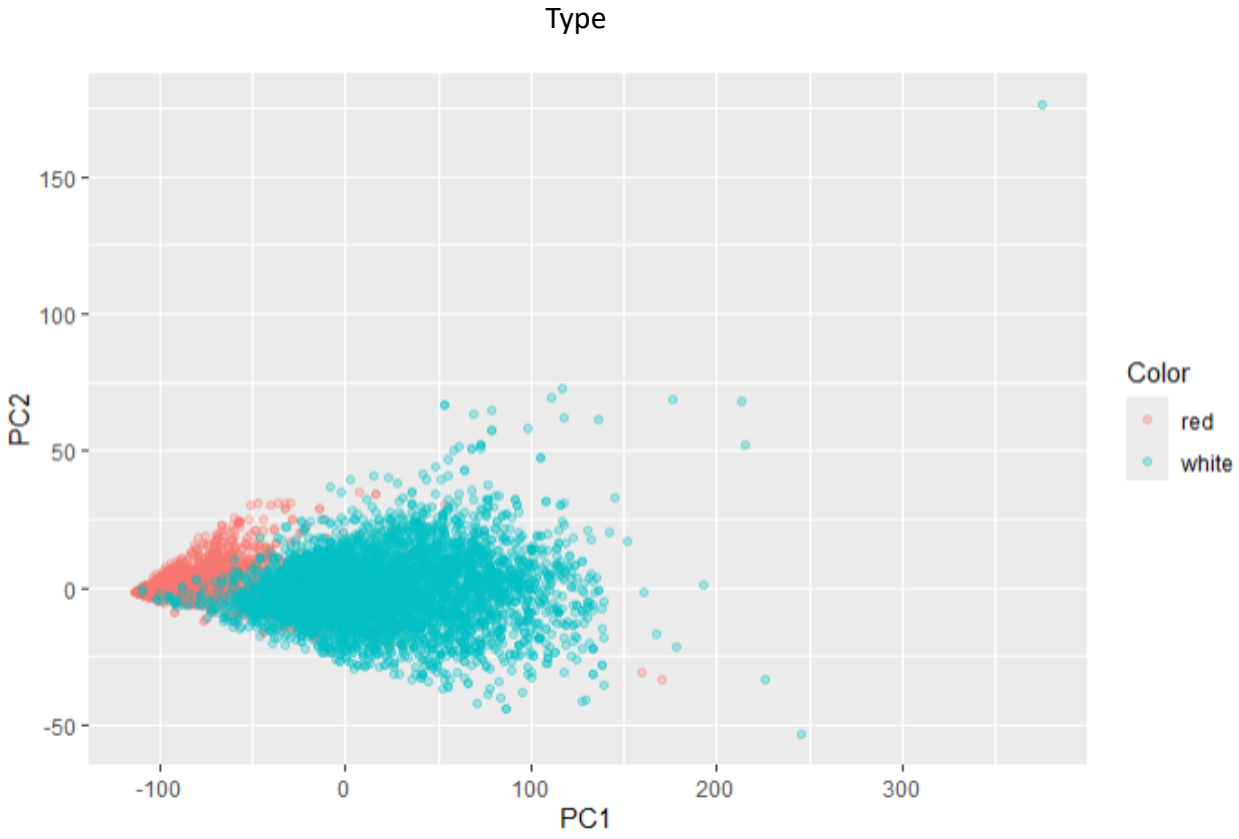
c.

HAC	Type red	white
1	1597	4896
2	2	0
3	0	1
4	0	1

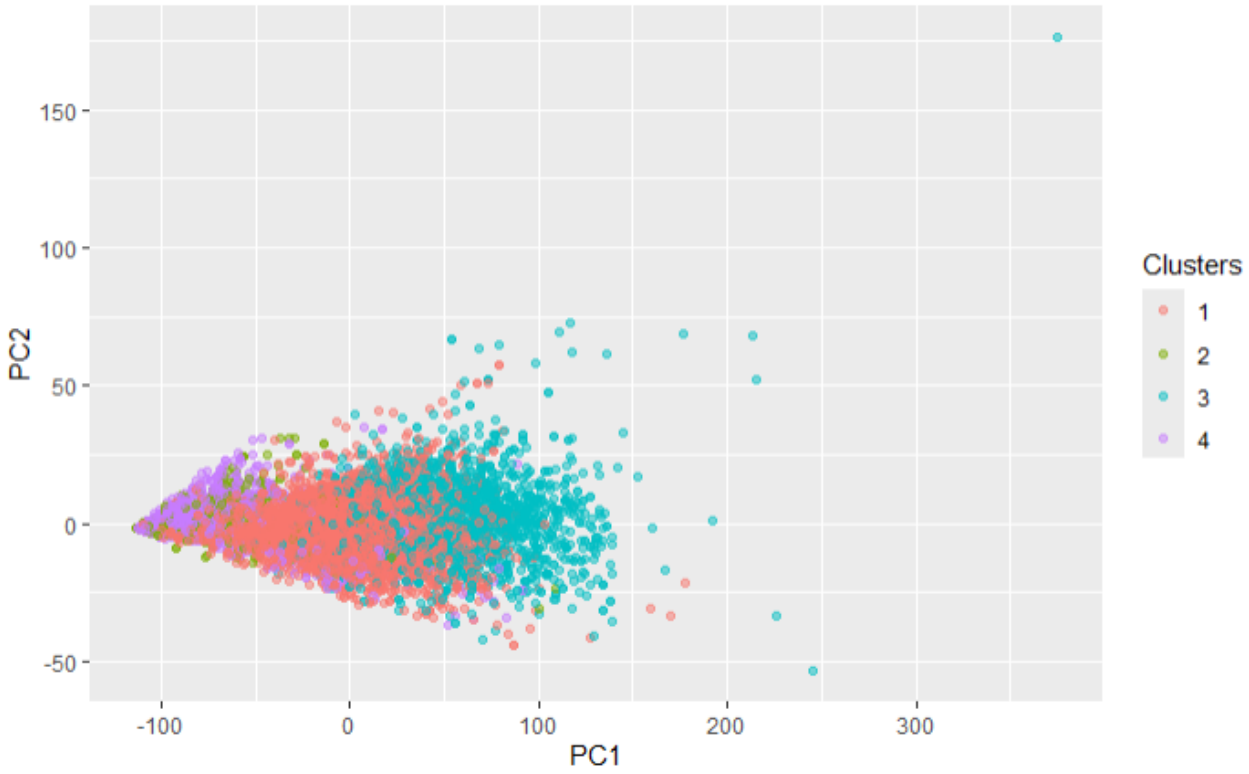
Kmeans	Type red	white
1	60	2775
2	609	52
3	3	1932
4	927	139

The HAC clustering was unable to make any distinctions between the red and white wines and instead placed nearly all of them in one cluster. K-means was much more effective at making meaningful clusters, with clusters 1 and 3 being primarily white wines and clusters 2 and 4 being mostly red wines.

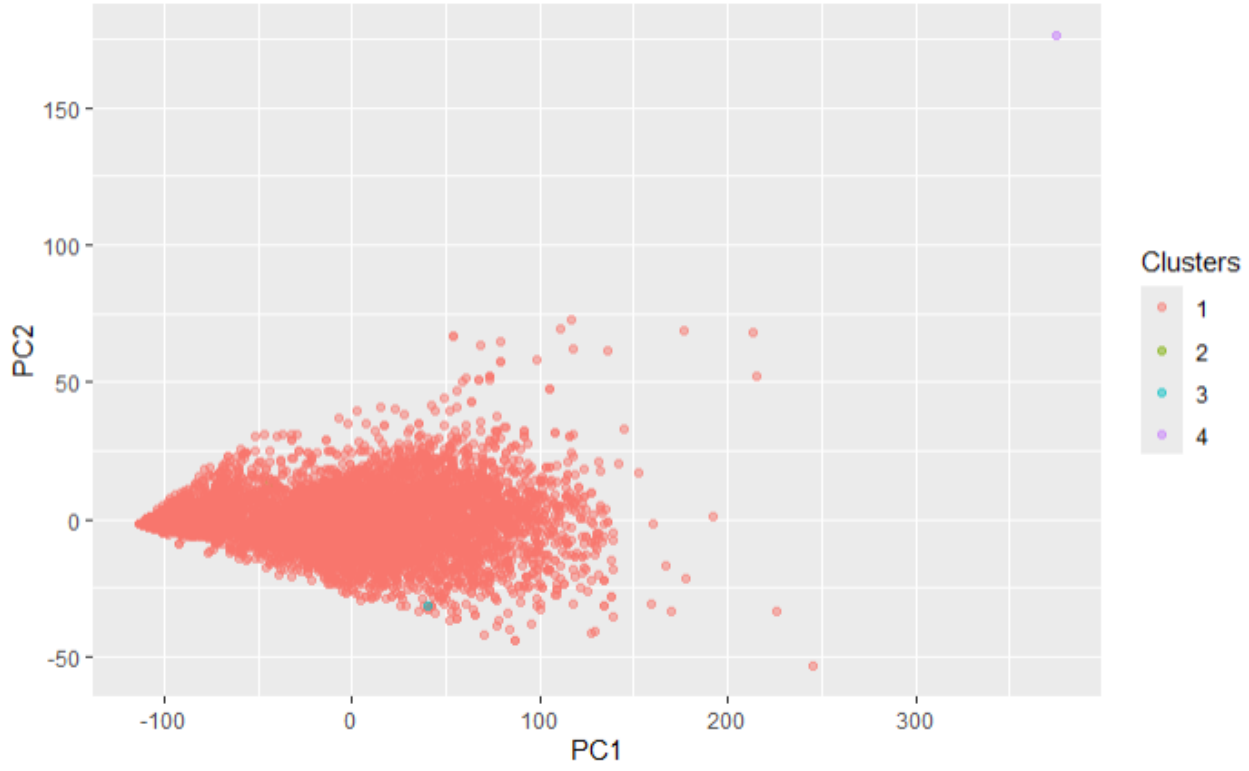
d.



K-Means


150
100
50
0
-50
PC2
-100
0
100
200
300
PC1
Clusters
1
2
3
4

HAC

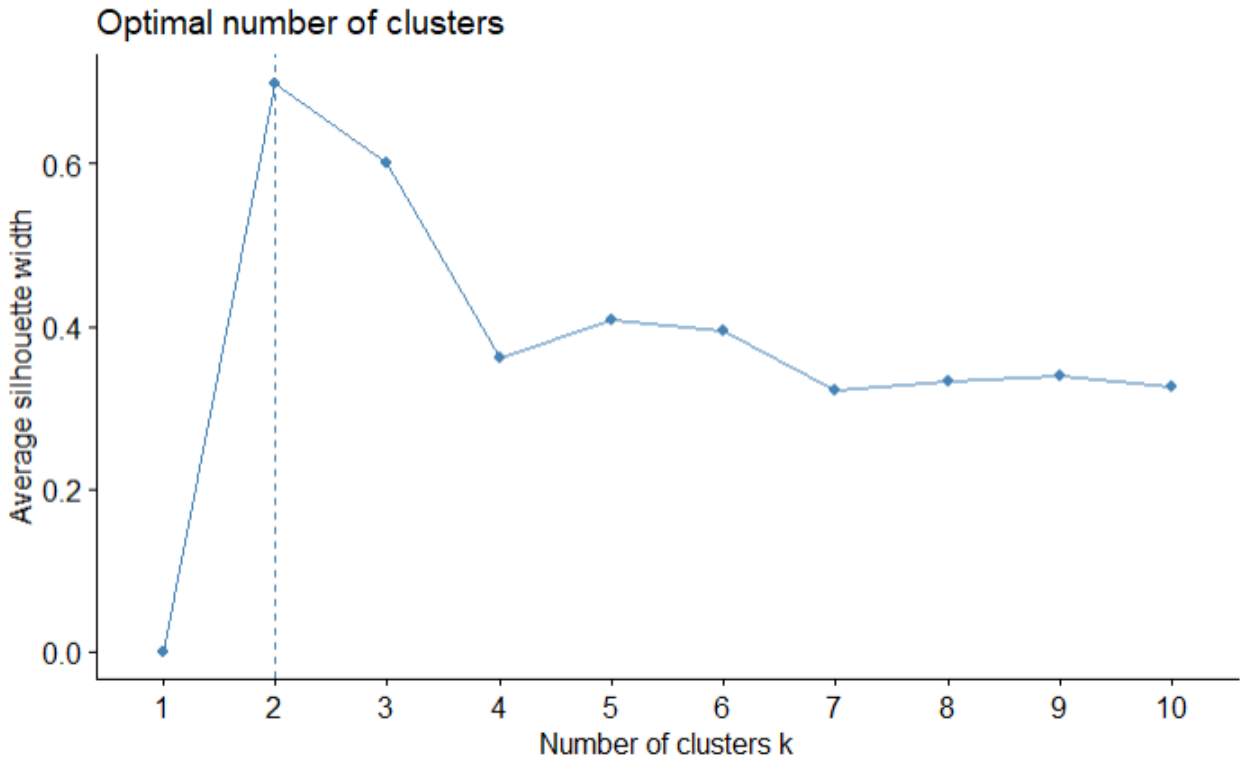

150
100
50
0
-50
PC2
-100
0
100
200
300
PC1
Clusters
1
2
3
4

e. The K-means was able to perform much better in this scenario because by giving the number of clusters and the randomized start, it's going to have more equal cluster sizes where its able to capture a pattern using its convex-shaped clusters. HAC, because it uses a distance measure, is subject to the chaining effect seen in this data, where it ends up creating a large cluster when there should be multiple.

Problem 4

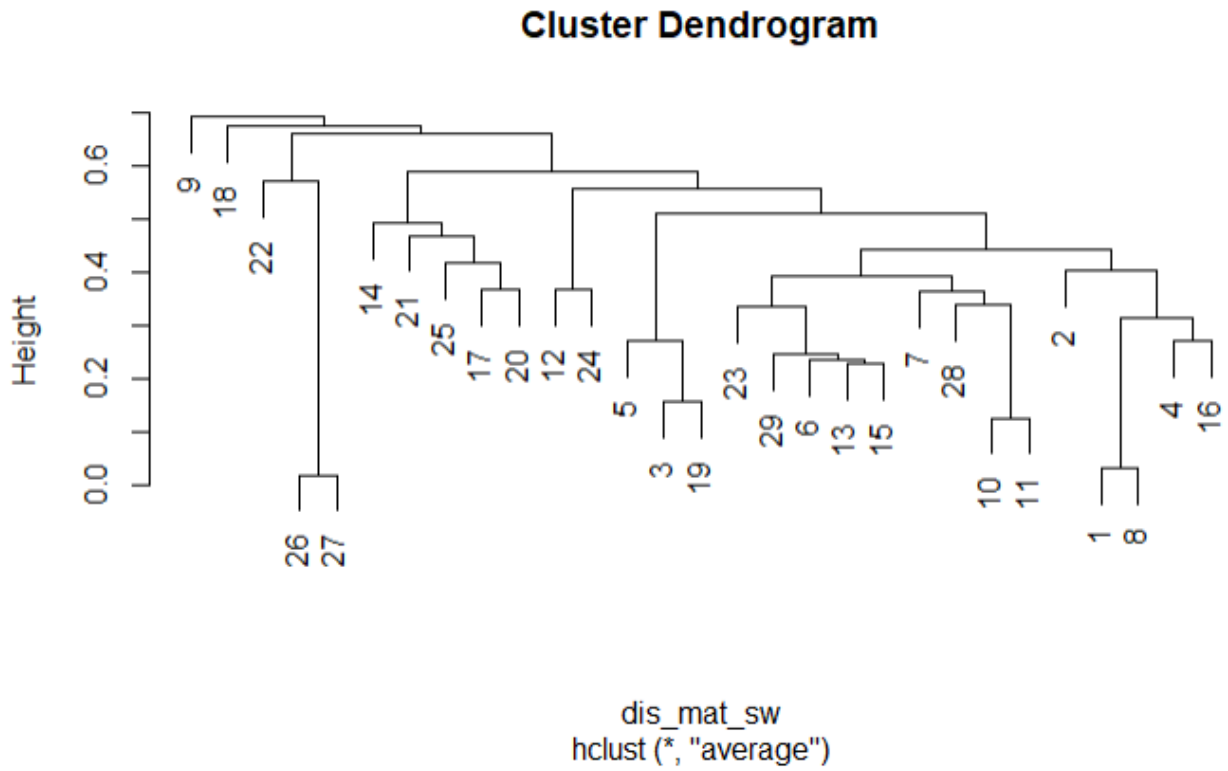
a.

```
406 dissimilarities, summarized :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01953 0.47186 0.57853 0.55149 0.62922 0.84891
Metric :  mixed ;  Types = I, I, N, N, N, I, N, N, N
Number of objects : 29
```



The optimal number of clusters based on the silhouette score is 2.

b.



An anomaly in a dendrogram would be a point that is in a cluster of its own. In this case, 9 is in its own cluster, while every other character is in the other cluster. We can identify 9 as an anomaly but we don't have any information as to what distinguishes that character from the rest. As opposed to using standard deviations to identify outliers which only works with normally distributed data, clustering with distance functions can incorporate categorical variables in our calculations along with numeric values to cluster data and identify anomalies.

c.

```
K-means clustering with 2 clusters of sizes 26, 3

Cluster means:
    height     mass hair_color.auburn, white hair_color.black hair_color.blond hair_color.brown
1 175.4615 76.20769               0.03846154        0.2307692       0.07692308        0.2307692
2 206.3333 91.33333               0.00000000        0.0000000       0.00000000        0.3333333
  hair_color.brown, grey hair_color.grey hair_color.none hair_color.white skin_color.blue
1             0.03846154      0.03846154       0.3461538        0.0000000      0.03846154
2             0.00000000      0.00000000       0.0000000        0.6666667      0.00000000
  skin_color.brown skin_color.brown mottle skin_color.dark skin_color.fair skin_color.green
1       0.03846154               0.03846154      0.07692308       0.2307692       0.03846154
2       0.00000000               0.00000000      0.00000000       0.3333333       0.00000000
  skin_color.light skin_color.orange skin_color.pale skin_color.red skin_color.tan skin_color.unknown
1        0.2307692        0.07692308      0.03846154     0.03846154     0.03846154          0.0000000
2        0.0000000        0.00000000      0.33333333     0.00000000     0.00000000          0.3333333
  skin_color.white skin_color.yellow eye_color.black eye_color.blue eye_color.blue-gray eye_color.brown
1       0.03846154        0.07692308      0.03846154      0.2692308          0.03846154       0.3461538
2       0.00000000        0.00000000      0.00000000      0.3333333          0.00000000       0.3333333
  eye_color.hazel eye_color.orange eye_color.red eye_color.yellow birth_year sex.female  sex.male
1      0.07692308       0.07692308    0.03846154        0.1153846    42.0500  0.2307692 0.7692308
2      0.00000000       0.00000000    0.00000000        0.3333333   131.3333  0.0000000 1.0000000
  homeworld.Alderaan homeworld.Bespin homeworld.Cerea homeworld.Concord Dawn homeworld.Corellia
1         0.03846154       0.03846154       0.0000000             0.03846154          0.07692308
2         0.00000000       0.00000000       0.3333333             0.00000000          0.00000000
  homeworld.Dathomir homeworld.Dorin homeworld.Endor homeworld.Haruun Kal homeworld.Kamino
1         0.03846154       0.03846154       0.03846154           0.03846154       0.03846154
2         0.00000000       0.00000000       0.00000000           0.00000000       0.00000000
  homeworld.Kashyyyk homeworld.Mirial homeworld.Mon Cala homeworld.Naboo homeworld.Ryloth
1          0.0000000        0.07692308         0.03846154       0.1153846       0.03846154
2          0.3333333        0.00000000         0.00000000       0.0000000       0.00000000
  homeworld.Serenno homeworld.Socorro homeworld.Stewjon homeworld.Tatooine homeworld.Trandosha
1         0.0000000        0.03846154        0.03846154          0.2307692          0.03846154
2         0.3333333        0.00000000        0.00000000          0.0000000          0.00000000
  species.Cerean species.Ewok species.Gungan species.Human species.Kel Dor species.Mirialan
1      0.0000000    0.03846154     0.03846154     0.6538462      0.03846154       0.07692308
2      0.3333333    0.00000000     0.00000000     0.3333333      0.00000000       0.00000000
  species.Mon Calamari species.Trandoshan species.Twi'lek species.Wookiee species.Zabrak
1           0.03846154          0.03846154       0.03846154       0.0000000     0.03846154
2           0.00000000          0.00000000       0.00000000       0.3333333     0.00000000

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
 1  1  1  1  1  1  1  1  2  1  1  1  1  1  1  1  1  1  1  1  1  1  2  1  1  1  2  1

Within cluster sum of squares by cluster:
[1] 32451.117  8491.333
 (between_SS / total_SS =  37.6 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

d.

HAC	Gender	
	feminine	masculine
1	6	22
2	0	1

Kmeans	Gender	
	feminine	masculine
1	6	20
2	0	3

Neither clustering method were able to distinguish feminine and masculine characters from this data. As identified earlier, the HAC method only selected one character for the second cluster, while the k-means method selected three. Overall, the imbalance in the gender data along with the small number of objects meant that neither model was able to correctly cluster feminine and masculine characters.