

clustering

Stephen Brownsey

25/12/2019

Introduction

The code in this .rmd file covers the analysis undertaken into which clusters methods perform best, given the sparse orchard data. All the Machine learning principles are applied in this section file as well as any extra exploratory data analysis. All the main findings were reported in the final dissertation report, although there are some extra exploratory data analysis and interim steps in this .rmd which are not significant enough to warrant making it to the final report. The process is repeated for both standardised and non-standardised data

Looking at a Markov Approach

This section is going cover two step clustering based on various clustering methods for comparison. In these scenarios each implementation will have a maximum of 8 clusters. Two after pre-bloom, 4 after and 8 after post bloom. Of these 8 final clusters it is expected that some may be sparse with no orchards in

```
load("data")
markov_data <- data_2012 %>%
  select(ends_with(".pre"), ends_with(".blm"), ends_with(".pos"), orchard) %>%
  unique()

#Standardising the data for purpose of clustering

markov_data <- markov_data %>%
  mutate(eiqB11F.pre = (eqB11F.pre - mean(eiqB11F.pre))/sd(eiqB11F.pre)) %>%
  mutate(eiqB11I.pre= (eqB11I.pre - mean(eiqB11I.pre))/sd(eiqB11I.pre)) %>%
  mutate(eiqB11F.blm = (eqB11F.blm - mean(eiqB11F.blm))/sd(eiqB11F.blm)) %>%
  mutate(eiqB11I.blm = (eqB11I.blm - mean(eiqB11I.blm))/sd(eiqB11I.blm)) %>%
  mutate(eiqB11T.blm= (eqB11T.blm - mean(eiqB11T.blm))/sd(eiqB11T.blm)) %>%
  mutate(eiqB11F.pos = (eqB11F.pos - mean(eiqB11F.pos))/sd(eiqB11F.pos)) %>%
  mutate(eiqB11I.pos = (eqB11I.pos - mean(eiqB11I.pos))/sd(eiqB11I.pos)) %>%
  mutate(eiqB11T.pos = (eqB11T.pos - mean(eiqB11T.pos))/sd(eiqB11T.pos))

#function to generate the agglomeraative clustering
aggl <- function(data){
  agnes(dist(data, method = "euclidian"),
        diss=TRUE, method = "ward")
}

#function to reduce copy pasting for each node
aggl_c <- function(data, ends, c_num){
  #define temp dataset of this stage by cluster
  temp <- data %>%
```

```

    filter(cluster == c_num) %>%
    select(ends_with(ends))

#If statement as if 1 element in dataset it'll error
if(nrow(temp) > 1){
  temp$cluster <- cutree(aggl(temp) , k = 2)

  output <- data %>%
    filter(cluster == c_num) %>%
    select(-cluster) %>%
    mutate(cluster = temp$cluster)
  #Else statement for if 1 element in dataset
}else{
  #If only 1 element in the dataset -> set the cluster number to 1
  output <- data %>%
    filter(cluster == c_num) %>%
    select(-cluster) %>%
    mutate(cluster = 1)
}
output
}

####Agglomerative heirachicale clustering
#pre-bloom step
temp <- markov_data %>%
  select(ends_with(".pre"))

temp$cluster <- cutree(aggl(temp) , k = 2)

aggl_node1 <- markov_data %>%
  mutate(cluster = temp$cluster)

#during bloom step 1
aggl_node2 <- aggl_c(aggl_node1, ".blm", 1)

#during bloom step 2
aggl_node3 <- aggl_c(aggl_node1, ".blm", 2)

#post bloom step 1
aggl_node4 <- aggl_c(aggl_node2, ".pos", 1)

#post bloom step 2
aggl_node5 <- aggl_c(aggl_node2, ".pos", 2)

#post bloom step 3
aggl_node6 <- aggl_c(aggl_node3, ".pos", 1)

#post bloom step 4
#Note this node only has 1 element in it so it is automatically set to 1
aggl_node7 <- aggl_c(aggl_node3, ".pos", 2)

```

```

orchard_node_agglom <- tibble(orchard = aggl_node4 %>% filter(cluster == 1) %>% select(orchard) %>% as_v
  bind_rows(tibble(orchard = aggl_node4 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, nod
  bind_rows(tibble(orchard = aggl_node5 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, nod
  bind_rows(tibble(orchard = aggl_node5 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, nod
  bind_rows(tibble(orchard = aggl_node6 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, nod
  bind_rows(tibble(orchard = aggl_node6 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, nod
  bind_rows(tibble(orchard = aggl_node7 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, nod

```

kmeans approach: Even running the 1001 times it is different everytime.

```

#function to generate the kmeans clustering
k_clustering <- function(data, n = 1001){
#creating the dataset to then calculate the optimal cluster from
  for(i in 1:n){
    if(i == 1){
      k_list <- tibble(kmeans(data, 2)$cluster)
    }else{
      k_list <- bind_cols(k_list, tibble(kmeans(data, 2)$cluster))
    }
  }
  apply(k_list[,1:length(k_list)], 1, mfv1) %>%
    enframe(value = "cluster") %>%
    select(cluster)
}

#pre-bloom step
#Getting the most common clustering

kmeans_c <- function(data, ends, c_num){

  data <- data %>%
    filter(cluster == c_num) %>%
    select(-cluster)

  if(nrow(data) > 2){
    temp <- data %>%
      select(ends_with(ends))

    cluster <- k_clustering(temp, 2)
    output <- bind_cols(data, cluster)
  }else{
    #Not more than 2 rows and get the error message as:
    #number of cluster centres must lie between 1 and nrow(x)
    output <- data %>%
      mutate(cluster = 1)
  }

  output
}

temp <- k_clustering(markov_data %>% select(-orchard))

kmeans_node1 <- markov_data %>%

```

```

bind_cols(temp)

#during bloom step 1
kmeans_node2 <- kmeans_c(kmeans_node1, ".blm", 1)

#during bloom step 2
kmeans_node3 <- kmeans_c(kmeans_node1, ".blm", 2)

#post bloom step 1
kmeans_node4 <- kmeans_c(kmeans_node2, ".pos", 1)

#post bloom step 2
kmeans_node5 <- kmeans_c(kmeans_node2, ".pos", 2)

#post bloom step 3
kmeans_node6 <- kmeans_c(kmeans_node3, ".pos", 1)

#post bloom step 4
kmeans_node7 <- kmeans_c(kmeans_node3, ".pos", 2)

#End timepoint nodes
orchard_node_kmeans <- tibble(orchard = kmeans_node4 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node_kmeans = kmeans_node4 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node_km = kmeans_node4 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node = kmeans_node4 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector)
bind_rows(tibble(orchard = kmeans_node4 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node_kmeans = kmeans_node4 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node_km = kmeans_node4 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node = kmeans_node4 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector),
tibble(orchard = kmeans_node5 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node_kmeans = kmeans_node5 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node_km = kmeans_node5 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node = kmeans_node5 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector),
tibble(orchard = kmeans_node5 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node_kmeans = kmeans_node5 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node_km = kmeans_node5 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node = kmeans_node5 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector),
tibble(orchard = kmeans_node6 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node_kmeans = kmeans_node6 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node_km = kmeans_node6 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node = kmeans_node6 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector),
tibble(orchard = kmeans_node6 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node_kmeans = kmeans_node6 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node_km = kmeans_node6 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node = kmeans_node6 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector),
tibble(orchard = kmeans_node7 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node_kmeans = kmeans_node7 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node_km = kmeans_node7 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node = kmeans_node7 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector),
tibble(orchard = kmeans_node7 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node_kmeans = kmeans_node7 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node_km = kmeans_node7 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node = kmeans_node7 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector))

```

Just a look at the difference in the final clusters between the two methods

```

clusterings <- orchard_node_agglom %>%
  arrange(orchard) %>%
  bind_cols(node_km = orchard_node_kmeans %>% arrange(orchard) %>%
    select(node) %>% as_vector)
clusterings

```

```

## # A tibble: 19 x 3
##   orchard node node_km
##   <fct>   <dbl>   <dbl>
## 1 A       8       8
## 2 B       9       9
## 3 C       9       9
## 4 D       9       9
## 5 E       8       8
## 6 F       8       8
## 7 G       9       9
## 8 H.nooil 8       8
## 9 I       8       8
## 10 J      8       8
## 11 K      8       8

```

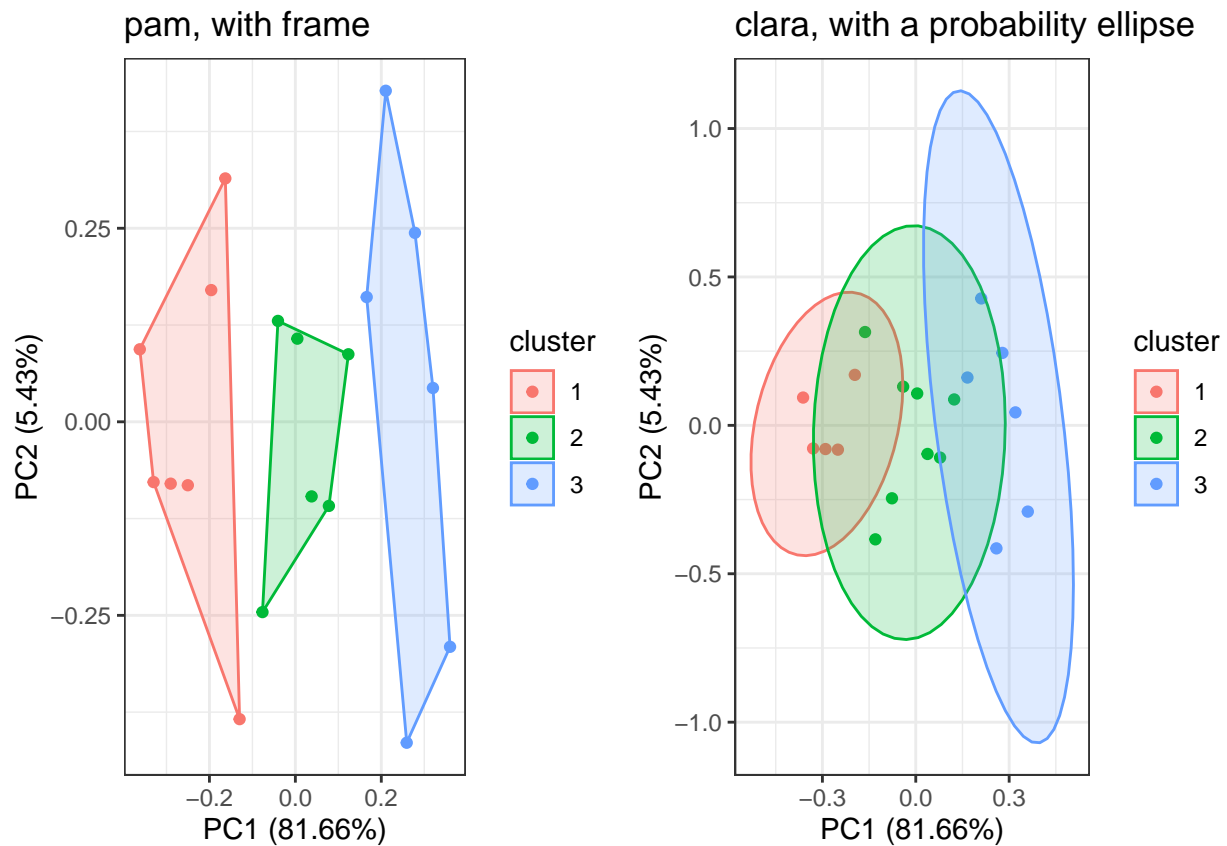
```
## 12 L      8      8
## 13 M      8      8
## 14 N      8      8
## 15 O.nooil 12     8
## 16 P     10     12
## 17 Q      8      8
## 18 R      8      8
## 19 S      8      8
```

Looking at top two PCAs just to see how much variation is absorbed by these

```
p1 <- autoplot(pam(markov_data, 3), frame = TRUE) +
  theme_bw() +
  ggtitle("pam, with frame")

p2 <- autoplot(clara(markov_data, 3), frame = TRUE, frame.type = 'norm') +
  theme_bw() +
  ggtitle("clara, with a probability ellipse")

grid.arrange(p1, p2, nrow = 1)
```



Non-markov structuring

This section will approach looking at the data as one time point with 1 clustering point rather than 3 as above: Method Justification - Maximum linkage function is best for both standardised and unstandardised.

```

maximum_cluster <- data_2012 %>%
  arrange(orchard) %>%
  select(ends_with(".pre"), ends_with(".blm"), ends_with(".pos")) %>%
  filter(row_number() %% 2 == 1) %>%
  mutate(eiqB11F.pre = (eqB11F.pre - mean(eiqB11F.pre))/sd(eiqB11F.pre)) %>%
  mutate(eiqB11I.pre = (eqB11I.pre - mean(eiqB11I.pre))/sd(eiqB11I.pre)) %>%
  mutate(eiqB11F.blm = (eqB11F.blm - mean(eiqB11F.blm))/sd(eiqB11F.blm)) %>%
  mutate(eiqB11I.blm = (eqB11I.blm - mean(eiqB11I.blm))/sd(eiqB11I.blm)) %>%
  mutate(eiqB11T.blm = (eqB11T.blm - mean(eiqB11T.blm))/sd(eiqB11T.blm)) %>%
  mutate(eiqB11F.pos = (eqB11F.pos - mean(eiqB11F.pos))/sd(eiqB11F.pos)) %>%
  mutate(eiqB11I.pos = (eqB11I.pos - mean(eiqB11I.pos))/sd(eiqB11I.pos)) %>%
  mutate(eiqB11T.pos = (eqB11T.pos - mean(eiqB11T.pos))/sd(eiqB11T.pos))

euclidean_cluster <- maximum_cluster

# matrix of methods to compare,
# rerun - without %>%mutate segment for non-standardised values
m <- c("average", "single", "complete", "ward")
names(m) <- c("average", "single", "complete", "ward")
distances <- c("euclidean", "maximum", "manhattan", "canberra",
               "minkowski")
names(distances) <- c("euclidean", "maximum", "manhattan",
                     "canberra", "minkowski")
clust_comps <- matrix(nrow = length(distances), ncol = length(m),
                     dimnames = list(distances, m))

# function to compute coefficient to see which is the best method
ac <- function(distance, linkage) {
  dista <- dist(maximum_cluster, method = distance)
  # Agglomerative Nesting form of Hierarchical Clustering
  agnes(dista, method = linkage)$ac
}

for(i in 1:length(distances)) {
  for(j in 1:length(m)) {
    clust_comps[i,j] <- ac(distances[i], m[j])
  }
}

# In future or in write up, perhaps change, needs more clustering analysis
# Look at literature in more detail to see other benefits/drawbacks and assumptions
# Then decide on actual best method, for now using euclidian as used that before
standardised_clust_comps <- clust_comps

```

Clustering code - Maximum

```

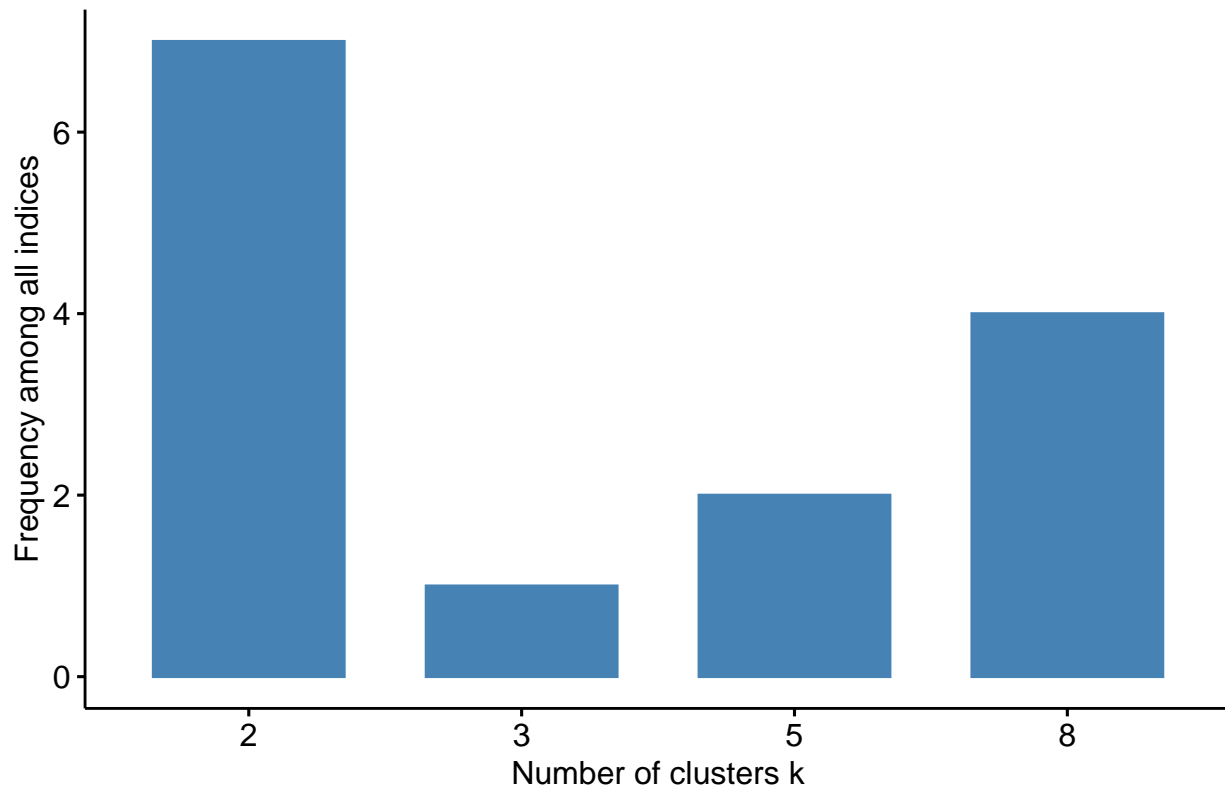
# NBclust function breaks when using index = all, instead using all the working indexes instead
fviz_nbclust(NbClust(maximum_cluster, distance = "maximum",
                     min.nc = 2, max.nc = 8, method = "ward.D2", index = c("kl", "ch", "ccc")

## Among all indices:
## =====
## * 7 proposed 2 as the best number of clusters
## * 1 proposed 3 as the best number of clusters
## * 2 proposed 5 as the best number of clusters
## * 4 proposed 8 as the best number of clusters

```

```
##
## Conclusion
## =====
## * According to the majority rule, the best number of clusters is 2 .
```

Optimal number of clusters – k = 2



```
#Agglomerative clustering
clustered <- agnes(dist(maximum_cluster, method = "maximum"),
                    diss=TRUE, method = "ward")
# add cluster labels to the training data
maximum_cluster$cluster <- cutree(clustered, k = 2)

maximum_cluster
```

```
##      eiqB11F.pre eiqB11I.pre eiqB11F.blm eiqB11I.blm eiqB11T.blm eiqB11F.pos
## 1  0.56294059 -0.3418658 -0.02661776  0.80366758 -0.24200410 -0.83657918
## 2  0.88788914 -0.3418658 -0.19022893 -0.30585457 -0.30620065  0.01058593
## 3  0.88788914 -0.3418658 -0.19022893 -0.30585457 -0.30620065  0.01058593
## 4  0.88788914 -0.3418658 -0.19022893 -0.30585457 -0.30620065  0.01058593
## 5 -1.21724288 -0.3418658 -0.47141495  0.25463750 -0.26826632  0.05673110
## 6 -1.78818873 -0.3418658 -0.71087017  2.78888695  0.36494510 -0.33951546
## 7  0.92973857  1.1014288  1.78834024  0.06780681 -0.26972533  2.88563060
## 8  0.01045773 -0.3418658  2.26526351 -0.36201840  0.01478210  1.10101634
## 9 -0.28688425  0.2579449 -0.21970544  0.87817061 -0.28577447 -0.22916831
## 10 -0.05899826 -0.3418658 -1.93232366 -1.17238220 -0.30620065  0.53623960
## 11  1.50859712 -0.1356809  0.15454685 -0.81247522 -0.29744657 -0.54716872
## 12  1.44670216 -0.3068769  0.15156608 -0.81247522 -0.30167771 -0.56622868
## 13 -0.33664640 -0.3418658 -0.18559218 -0.44683724 -0.30124000 -0.62240541
```

```
## 14 -1.63608890 -0.3418658 0.37048710 0.12282444 -0.30620065 -0.53312454
## 15 -1.58069994 3.8693055 -1.36167172 -0.74943419 -0.30620065 -1.18316953
## 16 -0.14885146 -0.3418658 0.30623494 -0.51446307 4.07083681 -0.08621861
## 17 -0.08097801 -0.3418658 -0.85792151 1.91089732 -0.28285645 -0.80197030
## 18 -0.02629240 -0.3418658 -0.15710926 -0.07432206 -0.05816852 -0.72071468
## 19 0.03876764 -0.3418658 1.45747473 -0.96491990 -0.30620065 1.85488796
##      eiQB11I.pos eiQB11T.pos cluster
## 1 0.07606715 -0.29236499 1
## 2 1.78589159 -0.29236499 1
## 3 1.78589159 -0.29236499 1
## 4 1.78589159 -0.29236499 1
## 5 -0.67656322 -0.29236499 1
## 6 0.61131654 -0.29236499 1
## 7 1.48904351 -0.23224030 1
## 8 -0.90112187 -0.29236499 1
## 9 -0.10337469 -0.17211561 1
## 10 -1.07800026 -0.29236499 1
## 11 -0.07517669 -0.16918270 1
## 12 -0.07517669 0.04491986 1
## 13 -0.78166487 -0.20291118 1
## 14 -1.05339109 -0.29236499 1
## 15 -0.68425358 -0.29236499 2
## 16 -0.61093877 4.10700262 2
## 17 -0.65349212 -0.29236499 1
## 18 -0.28179116 0.03025530 1
## 19 -0.55915698 -0.18971308 1
```

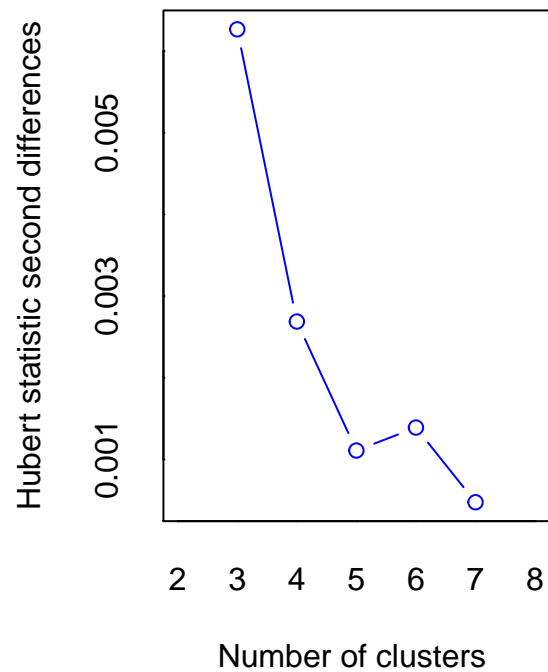
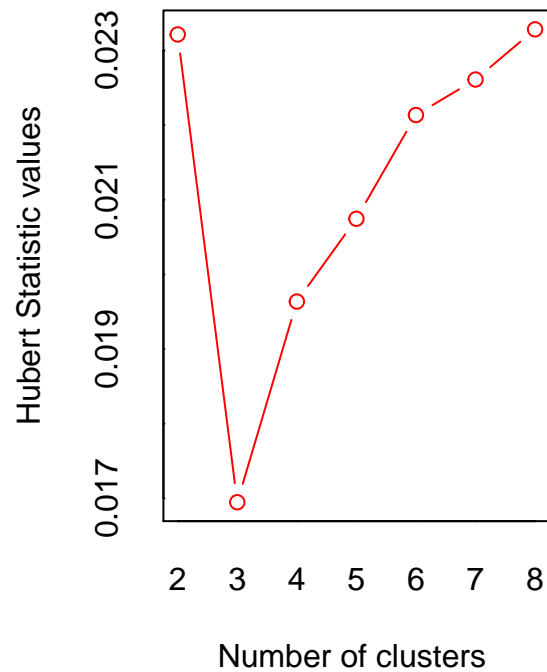
#Clustering summaries based on non-markov approach (Before Adjustment)

```
maximum_non_markov_clusters <- data_2012 %>%
  group_by(orchard) %>%
  summarise(mean_honey_ab = mean(apisAb),
            mean_wild_ab = mean(wildAbF),
            mean_social_rich = mean(socialRichF)) %>%
  mutate(cluster = maximum_cluster$cluster) %>%
  group_by(cluster) %>%
  summarise(mean_honey_ab = mean(mean_honey_ab),
            mean_wild_ab = mean(mean_wild_ab),
            mean_social_rich = mean(mean_social_rich),
            n = n())
```

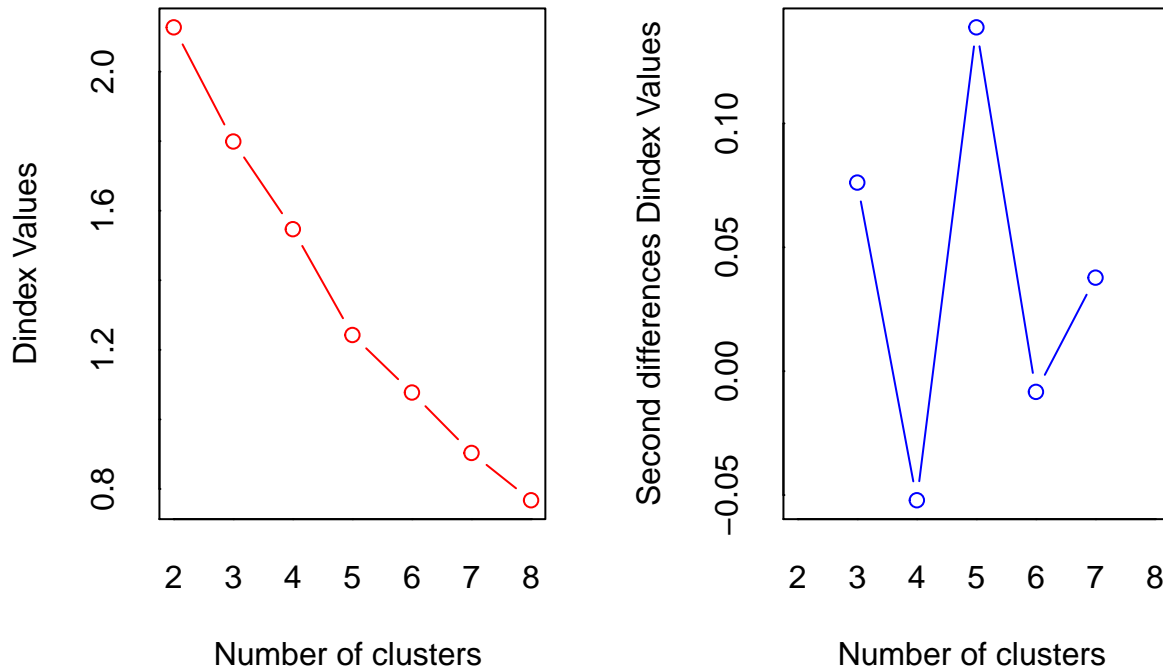
Clustering code - euclidean

#Clusters of whole data combined - 7 - breaks sometimes with euclidean bit weird but maybe another just
#Euclidean clusters.

```
fviz_nbclust(NbClust(euclidean_cluster, distance = "euclidean",
                    min.nc = 2, max.nc = 8, method = "ward.D2", index = "all"))
```

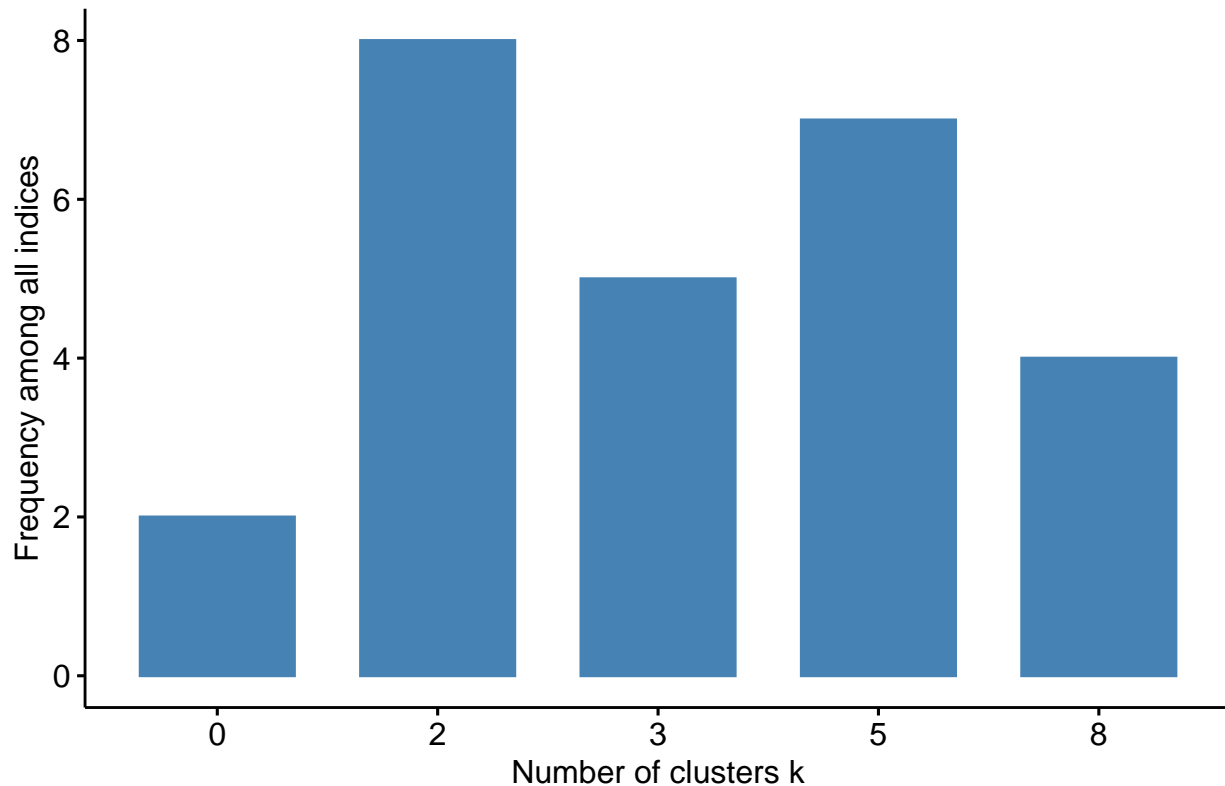
```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##           In the plot of Hubert index, we seek a significant knee that corresponds to a
##           significant increase of the value of the measure i.e the significant peak in Hubert
##           index second differences plot.
##
```



```
## *** : The D index is a graphical method of determining the number of clusters.
##           In the plot of D index, we seek a significant knee (the significant peak in Dindex
##           second differences plot) that corresponds to a significant increase of the value of
##           the measure.
##
## *****
## * Among all indices:
## * 8 proposed 2 as the best number of clusters
## * 5 proposed 3 as the best number of clusters
## * 7 proposed 5 as the best number of clusters
## * 4 proposed 8 as the best number of clusters
##
##           ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is 2
##
## *****
## Among all indices:
## =====
## * 2 proposed 0 as the best number of clusters
## * 8 proposed 2 as the best number of clusters
## * 5 proposed 3 as the best number of clusters
## * 7 proposed 5 as the best number of clusters
## * 4 proposed 8 as the best number of clusters
##
```

```
## Conclusion
## =====
## * According to the majority rule, the best number of clusters is 2 .
```

Optimal number of clusters – k = 2



```
#Agglomerative clustering
clustered <- agnes(dist(euclidean_cluster, method = "euclidean"),
                    diss=TRUE, method = "ward")
# add cluster labels to the training data - 8 proposed2 and 7 proposed 5
euclidean_cluster$cluster <- cutree(clustered, k = 5)
```

```
euclidean_cluster
```

```
##      eiqB11F.pre eiqB11I.pre eiqB11F.blm eiqB11I.blm eiqB11T.blm eiqB11F.pos
## 1  0.56294059 -0.3418658 -0.02661776  0.80366758 -0.24200410 -0.83657918
## 2  0.88788914 -0.3418658 -0.19022893 -0.30585457 -0.30620065  0.01058593
## 3  0.88788914 -0.3418658 -0.19022893 -0.30585457 -0.30620065  0.01058593
## 4  0.88788914 -0.3418658 -0.19022893 -0.30585457 -0.30620065  0.01058593
## 5 -1.21724288 -0.3418658 -0.47141495  0.25463750 -0.26826632  0.05673110
## 6 -1.78818873 -0.3418658 -0.71087017  2.78888695  0.36494510 -0.33951546
## 7  0.92973857  1.1014288  1.78834024  0.06780681 -0.26972533  2.88563060
## 8  0.01045773 -0.3418658  2.26526351 -0.36201840  0.01478210  1.10101634
## 9 -0.28688425  0.2579449 -0.21970544  0.87817061 -0.28577447 -0.22916831
## 10 -0.05899826 -0.3418658 -1.93232366 -1.17238220 -0.30620065  0.53623960
## 11  1.50859712 -0.1356809  0.15454685 -0.81247522 -0.29744657 -0.54716872
## 12  1.44670216 -0.3068769  0.15156608 -0.81247522 -0.30167771 -0.56622868
## 13 -0.33664640 -0.3418658 -0.18559218 -0.44683724 -0.30124000 -0.62240541
## 14 -1.63608890 -0.3418658  0.37048710  0.12282444 -0.30620065 -0.53312454
```

```
## 15 -1.58069994 3.8693055 -1.36167172 -0.74943419 -0.30620065 -1.18316953
## 16 -0.14885146 -0.3418658 0.30623494 -0.51446307 4.07083681 -0.08621861
## 17 -0.08097801 -0.3418658 -0.85792151 1.91089732 -0.28285645 -0.80197030
## 18 -0.02629240 -0.3418658 -0.15710926 -0.07432206 -0.05816852 -0.72071468
## 19 0.03876764 -0.3418658 1.45747473 -0.96491990 -0.30620065 1.85488796
##      eiqB11I.pos eiqB11T.pos cluster
## 1 0.07606715 -0.29236499 1
## 2 1.78589159 -0.29236499 2
## 3 1.78589159 -0.29236499 2
## 4 1.78589159 -0.29236499 2
## 5 -0.67656322 -0.29236499 1
## 6 0.61131654 -0.29236499 1
## 7 1.48904351 -0.23224030 3
## 8 -0.90112187 -0.29236499 3
## 9 -0.10337469 -0.17211561 1
## 10 -1.07800026 -0.29236499 1
## 11 -0.07517669 -0.16918270 2
## 12 -0.07517669 0.04491986 2
## 13 -0.78166487 -0.20291118 1
## 14 -1.05339109 -0.29236499 1
## 15 -0.68425358 -0.29236499 4
## 16 -0.61093877 4.10700262 5
## 17 -0.65349212 -0.29236499 1
## 18 -0.28179116 0.03025530 1
## 19 -0.55915698 -0.18971308 3
```

#Clustering summaries based on non-markov approach (Before Adjustment)

```
euclidean_non_markov_clusters <- data_2012 %>%
  group_by(orchard) %>%
  summarise(mean_honey_ab = mean(apisAb),
            mean_wild_ab = mean(wildAbF),
            mean_social_rich = mean(socialRichF)) %>%
  mutate(cluster = euclidean_cluster$cluster) %>%
  group_by(cluster) %>%
  summarise(mean_honey_ab = mean(mean_honey_ab),
            mean_wild_ab = mean(mean_wild_ab),
            mean_social_rich = mean(mean_social_rich),
            n = n())
```

Variable justification

```
bee_correlations <- data.frame(matrix(0, nrow = 0, ncol = 0))

for(i in 1:7){
  for(j in 1:7){
    bee_correlations[i,j] <- round(cor(data_2012[5+i], data_2012[5 + j]), 2)
  }
}

for(i in 1:7){
  names(bee_correlations)[i] <- colnames(data_2012[5 + i])
}

bee_correlations
```

##	apisAb.1	apisAb.2	apisAb.3	apisAb.4	apisAb.5	apisAb.6	apisAb.7	wildAbF.1
## 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.29
## 2	0.29	0.29	0.29	0.29	0.29	0.29	0.29	1.00
## 3	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.91
## 4	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.96
## 5	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.92
## 6	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.29
## 7	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.31
##	wildAbF.2	wildAbF.3	wildAbF.4	wildAbF.5	wildAbF.6	wildAbF.7	wildRichF.1	
## 1	0.29	0.29	0.29	0.29	0.29	0.29	0.19	
## 2	1.00	1.00	1.00	1.00	1.00	1.00	0.91	
## 3	0.91	0.91	0.91	0.91	0.91	0.91	1.00	
## 4	0.96	0.96	0.96	0.96	0.96	0.96	0.81	
## 5	0.92	0.92	0.92	0.92	0.92	0.92	0.92	
## 6	0.29	0.29	0.29	0.29	0.29	0.29	0.50	
## 7	0.31	0.31	0.31	0.31	0.31	0.31	0.53	
##	wildRichF.2	wildRichF.3	wildRichF.4	wildRichF.5	wildRichF.6	wildRichF.7		
## 1	0.19	0.19	0.19	0.19	0.19	0.19	0.19	
## 2	0.91	0.91	0.91	0.91	0.91	0.91	0.91	
## 3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
## 4	0.81	0.81	0.81	0.81	0.81	0.81	0.81	
## 5	0.92	0.92	0.92	0.92	0.92	0.92	0.92	
## 6	0.50	0.50	0.50	0.50	0.50	0.50	0.50	
## 7	0.53	0.53	0.53	0.53	0.53	0.53	0.53	
##	solitaryAbF.1	solitaryAbF.2	solitaryAbF.3	solitaryAbF.4	solitaryAbF.5			
## 1	0.25	0.25	0.25	0.25	0.25	0.25		
## 2	0.96	0.96	0.96	0.96	0.96	0.96		
## 3	0.81	0.81	0.81	0.81	0.81	0.81		
## 4	1.00	1.00	1.00	1.00	1.00	1.00		
## 5	0.92	0.92	0.92	0.92	0.92	0.92		
## 6	0.02	0.02	0.02	0.02	0.02	0.02		
## 7	0.05	0.05	0.05	0.05	0.05	0.05		
##	solitaryAbF.6	solitaryAbF.7	solitaryRichF.1	solitaryRichF.2				
## 1	0.25	0.25	0.13	0.13				
## 2	0.96	0.96	0.92	0.92				
## 3	0.81	0.81	0.92	0.92				
## 4	1.00	1.00	0.92	0.92				
## 5	0.92	0.92	1.00	1.00				
## 6	0.02	0.02	0.13	0.13				
## 7	0.05	0.05	0.16	0.16				
##	solitaryRichF.3	solitaryRichF.4	solitaryRichF.5	solitaryRichF.6				
## 1	0.13	0.13	0.13	0.13				
## 2	0.92	0.92	0.92	0.92				
## 3	0.92	0.92	0.92	0.92				
## 4	0.92	0.92	0.92	0.92				
## 5	1.00	1.00	1.00	1.00				
## 6	0.13	0.13	0.13	0.13				
## 7	0.16	0.16	0.16	0.16				
##	solitaryRichF.7	socialAbF.1	socialAbF.2	socialAbF.3	socialAbF.4			
## 1	0.13	0.18	0.18	0.18	0.18			
## 2	0.92	0.29	0.29	0.29	0.29			
## 3	0.92	0.50	0.50	0.50	0.50			
## 4	0.92	0.02	0.02	0.02	0.02			
## 5	1.00	0.13	0.13	0.13	0.13			

## 6	0.13	1.00	1.00	1.00	1.00
## 7	0.16	0.99	0.99	0.99	0.99
##	socialAbF.5	socialAbF.6	socialAbF.7	socialRichF.1	socialRichF.2
## 1	0.18	0.18	0.18	0.22	0.22
## 2	0.29	0.29	0.29	0.31	0.31
## 3	0.50	0.50	0.50	0.53	0.53
## 4	0.02	0.02	0.02	0.05	0.05
## 5	0.13	0.13	0.13	0.16	0.16
## 6	1.00	1.00	1.00	0.99	0.99
## 7	0.99	0.99	0.99	1.00	1.00
##	socialRichF.3	socialRichF.4	socialRichF.5	socialRichF.6	socialRichF.7
## 1	0.22	0.22	0.22	0.22	0.22
## 2	0.31	0.31	0.31	0.31	0.31
## 3	0.53	0.53	0.53	0.53	0.53
## 4	0.05	0.05	0.05	0.05	0.05
## 5	0.16	0.16	0.16	0.16	0.16
## 6	0.99	0.99	0.99	0.99	0.99
## 7	1.00	1.00	1.00	1.00	1.00

This demonstrates that to consider the affects on wild bees, just two variables need to be considered as there is very high correlation between the outcomes. WildAbF accounts a .9+ correlation with all wild variables except for the for the solitary bee variable. Then for this SocialRichF could be used.

Updating the bee values to match the environments

In this particular analysis, the interest is in the affects of pesticides on bee population. As such, it is necessary to correct, as best we can, for extra factors (confounding?) to make the bee values representative.

All these graphs were plotted as part of EDA by the use of lapply to the a basic plotting function. Here the relationships between some of the main extra factors (whether they are confounding or not) will be examined.

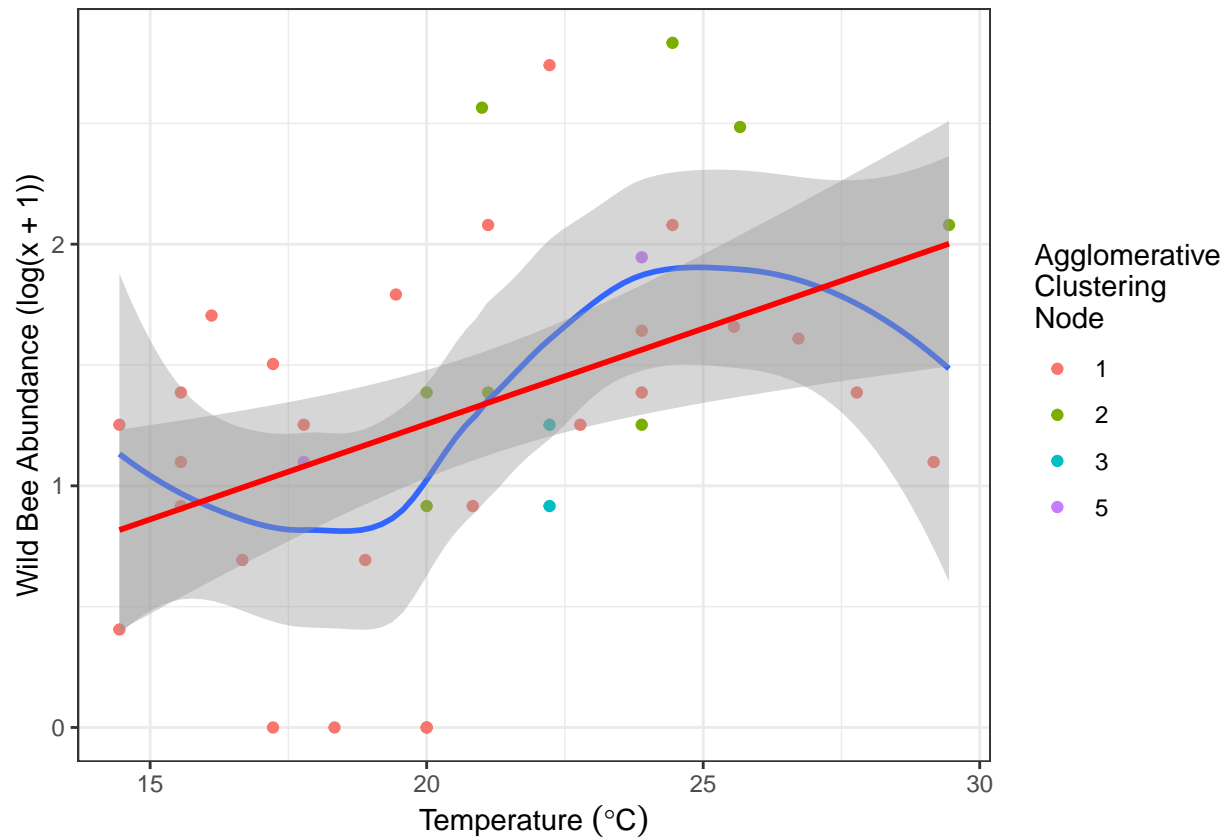
Temperature

It is known that temperature affects the speed at which bees fly, as such this will have an effect on bee abundance and possibly richness although more bees might not necessarily mean more species are observed. From the previous analysis carried out it is known that a $\log(x) + 1$ transposition of bee count allows for a “better” model fit and as such in general this will the case for our observations.

```
##agglom
wild_bee_abundance_agglom <- data_2012 %>%
  inner_join(orchard_node_agglom) %>%
  ggplot(aes(x = temp, y = log(wildAbF + 1))) +
  geom_point(aes(colour = as_factor(node- 7))) +
  geom_smooth() +
  geom_smooth(method = "lm", colour = "red") +
  labs(x = expression("Temperature " ( degree*C)), y = "Wild Bee Abundance (log(x + 1))",
       # title = "Wild Bee Abundance Coloured by Agglomerature Nodes",
       colour = "Agglomerative\nClustering\nNode") +
  theme_bw()

## Joining, by = "orchard"
wild_bee_abundance_agglom

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

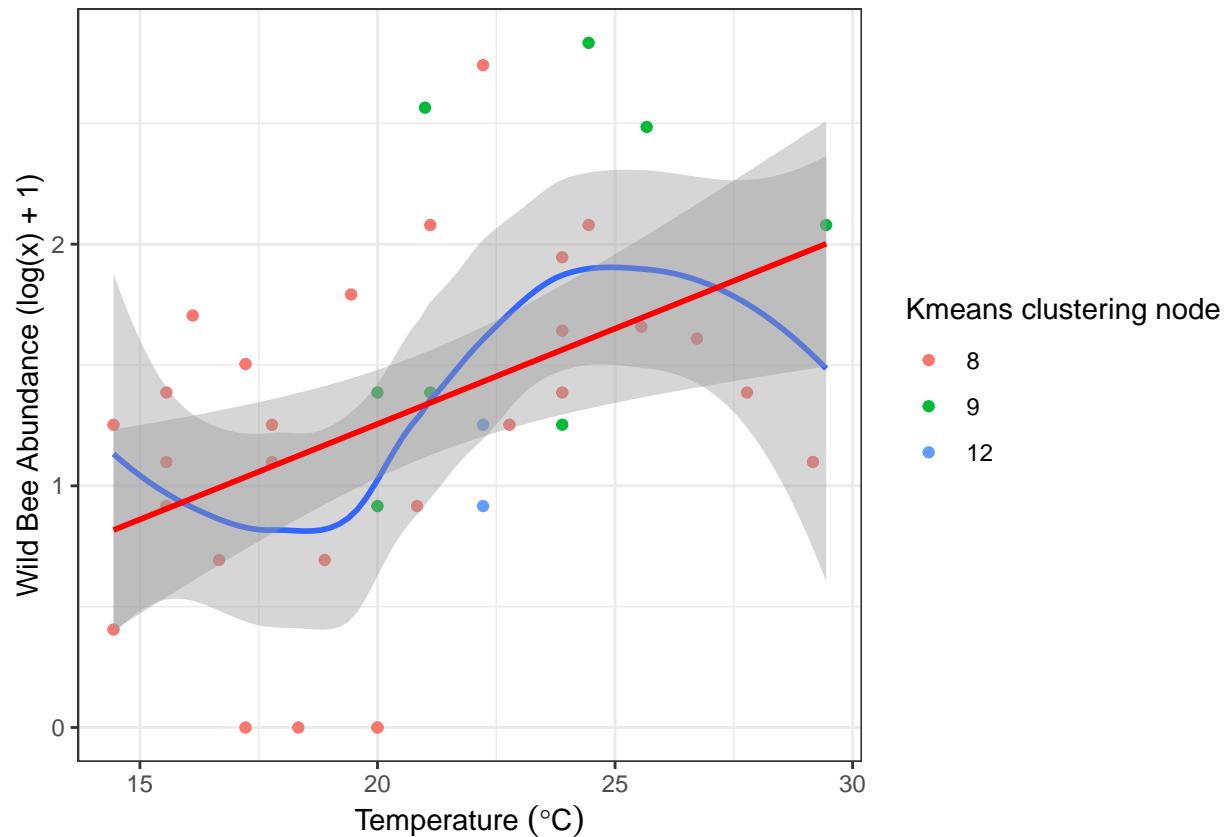


```
##kmeans
wild_bee_abundance_kmeans <- data_2012 %>%
  inner_join(orchard_node_kmeans) %>%
  ggplot(aes(x = temp, y = log(wildAbF + 1))) +
  geom_point(aes(colour = as_factor(node))) +
  geom_smooth() +
  geom_smooth(method = "lm", colour = "red") +
  labs(x = expression("Temperature " ( degree*C)), y = "Wild Bee Abundance ( $\log(x) + 1$ )",
    #title = "Wild Bee Abundance Coloured by Kmeans Nodes",
    colour = "Kmeans clustering node") +
  theme_bw()
```

```
## Joining, by = "orchard"
```

```
wild_bee_abundance_kmeans
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

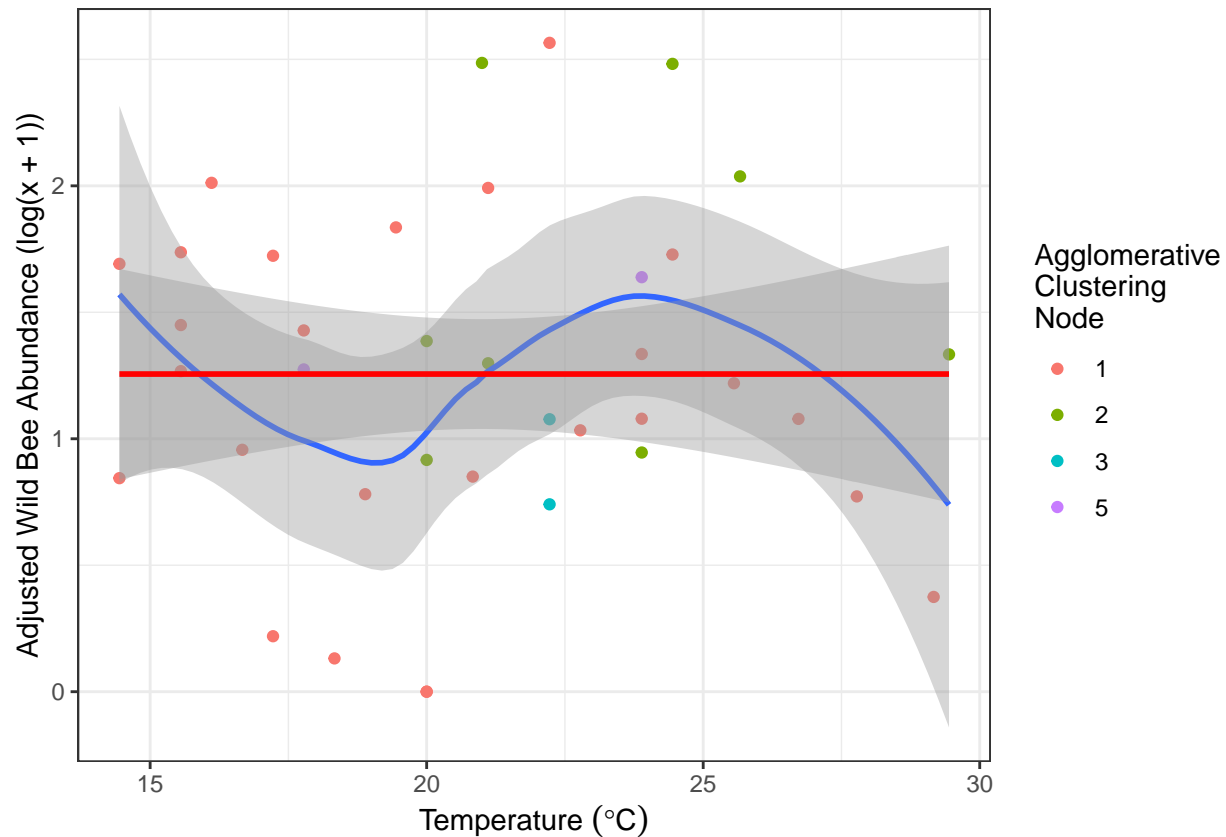


```
##Linear model
lm_temp <- lm(log(data_2012$wildAbF + 1) ~ data_2012$temp)
temp_factor <- tidy(lm_temp) %>%
  slice(2) %>%
  select(estimate) %>%
  as_vector()

#Adjusting bee value as though temp is always 20
adjusted_data <- data_2012 %>%
  inner_join(clusterings) %>%
  mutate(adjusted_beas = (((20 - temp) * temp_factor) + log(wildAbF + 1)))

## Joining, by = "orchard"
wild_bee_abundance_agglom_adjusted <- adjusted_data %>%
  ggplot(aes(x = temp, y = adjusted_beas)) +
  geom_point(aes(colour = as_factor(node - 7))) +
  geom_smooth() +
  geom_smooth(method = "lm", colour = "red") +
  labs(x = expression("Temperature " (degree*C)), y = "Adjusted Wild Bee Abundance (log(x + 1))",
    # title = "Wild Bee Abundance Coloured by Agglomerature Nodes",
    colour = "Agglomerative\nClustering\nNode") +
  theme_bw()
wild_bee_abundance_agglom_adjusted

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

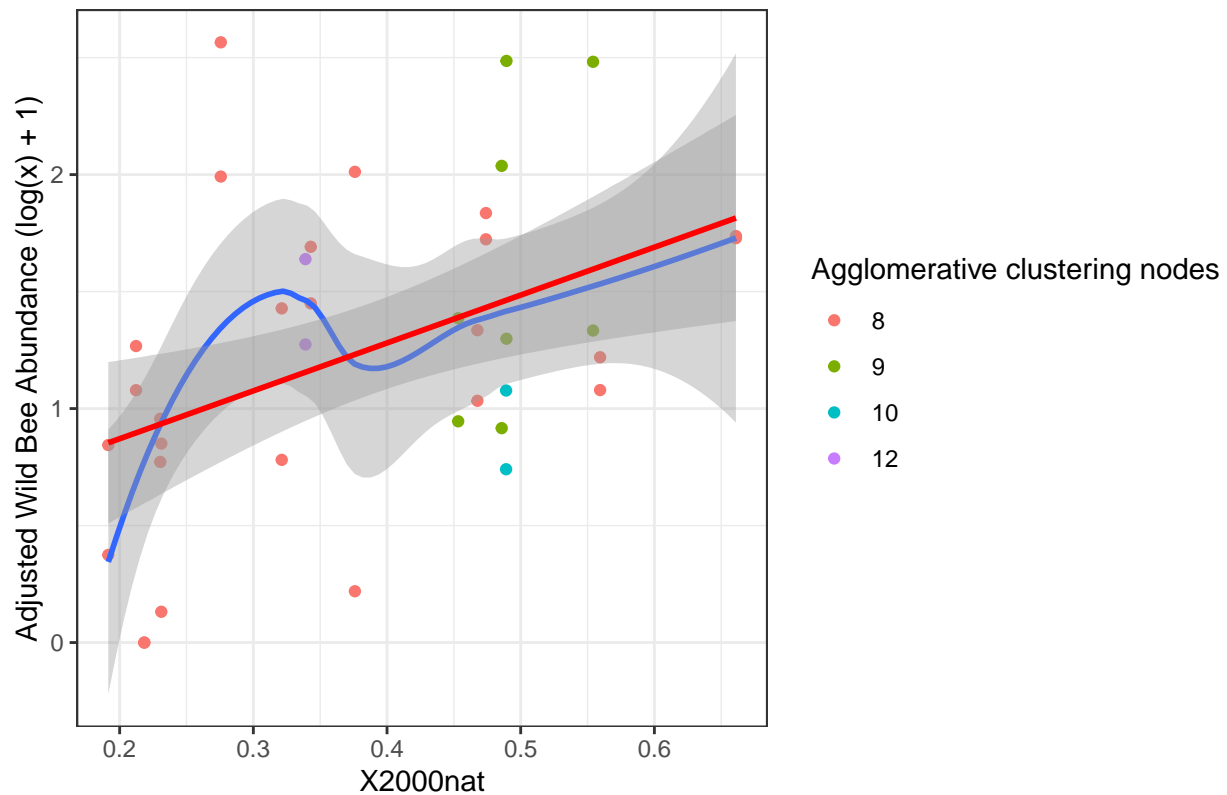
X2000nat

Based on previous research undertaken in the original paper it is known that the X2000nat variable has an effect on the the bee count so it also necessary to adjust for this. Wher Again the value will be adjust as though the X2000nat variable is constant. In this case the mean will be chosen as the constant value and the bee count will be adjusted for this in the same way as above.

```
wild_bee_abundance_agglom_x2000 <- adjusted_data %>%
  ggplot(aes(x = X2000nat, y = adjusted_beas)) +
  geom_point(aes(colour = as_factor(node))) +
  geom_smooth() +
  geom_smooth(method = "lm", colour = "red") +
  labs(x = "X2000nat", y = "Adjusted Wild Bee Abundance (log(x) + 1)",
       title = "Wild Bee Abundance Coloured by Agglomerature Nodes",
       colour = "Agglomerative clustering nodes") +
  theme_bw()
wild_bee_abundance_agglom_x2000
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Wild Bee Abundance Coloured by Agglomerative Nodes



```
#Outcome is 0.388
mean_x2000 <- adjusted_data %>%
  summarise(mean_x2000 = mean(X2000nat)) %>%
  as_vector()

lm_X2000nat <- lm(adjusted_data$adjusted_beas ~ adjusted_data$X2000nat)
X2000_factor <- tidy(lm_X2000nat) %>%
  slice(2) %>%
  select(estimate) %>%
  as_vector()

#Adjusting the data for this variable

adjusted_data <- adjusted_data %>%
  mutate(adjusted_beas = ((mean_x2000 - X2000nat) * X2000_factor) + adjusted_beas)
```

So if you look at them individually, it can only be adjusted for one. So need to combine a linear model of all affecting factors to best adjust at the end.

Analysing other potential confounders

This section will look at finding other parameters to include in the bee variable, dataset used at this point is the original just with the bee variable considered going through a $\log(x + 1)$ transformation. The only bee variable to be considered at this point is the wild bee abundance, in future more could be considered in this

way or it would be expected to be similar for all bee variables and I could just use the variables decided upon in this way for all the bee variables considered.

Local Diversity

The aim here is to whether adding local diversity as a adjustment factor is going to be beneficial or not. This will be achieved by running a F test to see whether the variances are equal, a shapiro test to confirm the bee data can be assumed to be normal. Note, independence assumption is not validated, this is due more than 1 result being used from each orchard, I decided to include these as I thought the extra power gained from twice the observations was worth the penalty of this assumption.

```
#When adding temperature in other variables have an effect
logged_data <- data_2012 %>%
  inner_join(clusterings) %>%
  mutate(adjusted_bees = (((20 - temp) * temp_factor) + log(wildAbF + 1)))

## Joining, by = "orchard"

#Just logged
logged_data <- data_2012 %>%
  inner_join(clusterings) %>%
  mutate(adjusted_bees = log(wildAbF + 1))

## Joining, by = "orchard"

#X2000nat adjustment
logged_data <- data_2012 %>%
  inner_join(clusterings) %>%
  mutate(adjusted_bees = log(wildAbF + 1))%>%
  mutate(adjusted_bees = ((mean_x2000 - X2000nat) * X2000_factor) + adjusted_bees)

## Joining, by = "orchard"

#Testing underlying normal data assumption, again demonstrating why the transformation is neccessary:
#Not normal
shapiro.test(logged_data$wildAbF)

##
## Shapiro-Wilk normality test
##
## data: logged_data$wildAbF
## W = 0.79764, p-value = 9.361e-06

#After transformation -> normal
shapiro.test(logged_data$adjusted_bees)

##
## Shapiro-Wilk normality test
##
## data: logged_data$adjusted_bees
## W = 0.97681, p-value = 0.6046

#Checking if the two subsets have the same variance have the same variance
#Outcome shows that they can be assumed to have the same variance as p.value > 0,05
simple <- logged_data %>% filter(local.diversity == 0) %>% select(adjusted_bees) %>% as_vector()
diverse <- logged_data %>% filter(local.diversity == 1) %>% select(adjusted_bees) %>% as_vector()
tidy(var.test(simple, diverse)) %>%
  select(statistic, p.value)
```

```
## Multiple parameters; naming those columns num.df, denom.df

## # A tibble: 1 x 2
##   statistic p.value
##   <dbl>     <dbl>
## 1      1.17    0.725

#Performing a two sample t-test on the data to see whether they are the same.
#Since H0: Means the same, H1: means are different ~ p.value not significant

#This implies local.diversity does NOT have an effect on the original bee counts.
#I thought this was an interesting point and worth noting, same happens with Region.
tidy(t.test(wildAbF ~ local.diversity, logged_data, var.equal = TRUE))
```

```
## # A tibble: 1 x 9
##   estimate1 estimate2 statistic p.value parameter conf.low conf.high method
##   <dbl>     <dbl>     <dbl>   <dbl>     <dbl>     <dbl>     <dbl> <chr>
## 1      3.5      4.18     -0.532  0.598        36     -3.27      1.91 " Two~
## # ... with 1 more variable: alternative <chr>
```

```
#However the the adjusted version of bee count does list local diversity
#as a factor which has an effect on the log transformed bee counts
tidy(t.test(adjusted_beas ~ local.diversity, logged_data, var.equal = TRUE))
```

```
## # A tibble: 1 x 9
##   estimate1 estimate2 statistic p.value parameter conf.low conf.high method
##   <dbl>     <dbl>     <dbl>   <dbl>     <dbl>     <dbl>     <dbl> <chr>
## 1      1.22      1.41     -0.896  0.376        36     -0.617     0.239 " Two~
## # ... with 1 more variable: alternative <chr>
```

From This analysis, it can be concluded that local.diversity does NOT play a factor in bee count ### Region Region has 3 possible options - as such an ANOVA test will be used to see whether there is a difference between the means in wild abundance based on region.

```
#Checking Anova assumption that the variances are the same
# H0: Variances the same: H1: Atleast one variance is different
leveneTest(adjusted_beas ~ region, data = logged_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 2  0.1954 0.8234
##      35
```

```
#Since pr > 0.05 we can assume variances are the same
```

```
#Running the anova test
anova1 <- aov(adjusted_beas ~ region, data = logged_data)

summary(anova1)
```

```
##      Df Sum Sq Mean Sq F value Pr(>F)
## region 2  0.522  0.2611  0.624  0.542
## Residuals 35 14.652  0.4186
```

```
#probability shows that region is not an affecting factor to wildAbF
```

```
#Checking that the following anova assumption is TRUE:
#Residuals of the response variable are normally distributed is NOT true
shapiro.test(residuals(anova1))
```

```
##
## Shapiro-Wilk normality test
##
## data: residuals(anova1)
## W = 0.96856, p-value = 0.3546
```

Day

Looking at the day to see whether that also has an effect, since the data can be paired by day 1 and day 2 a paired t-test will be carried out on this to test whether there is a difference in bee count. When using just the X2000nat variable, then it becomes very close to significant.

```
day_1 <- logged_data %>% filter(day == 1) %>% select(adjusted_bees) %>% as_vector()
day_2 <- logged_data %>% filter(day == 2) %>% select(adjusted_bees) %>% as_vector()
#This demonstrates that the variances come from the same distribution:
tidy(var.test(day_1, day_2)) %>%
  select(statistic, p.value)
```

```
## Multiple parameters; naming those columns num.df, denom.df
## # A tibble: 1 x 2
##   statistic p.value
##   <dbl>    <dbl>
## 1      0.936    0.891
```

```
#Hence we can run a paired t-test with equal variances
tidy(t.test(adjusted_bees ~ day, logged_data, var.equal = TRUE, paired = TRUE))
```

```
## # A tibble: 1 x 8
##   estimate statistic p.value parameter conf.low conf.high method
##   <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl> <chr>
## 1   -0.342    -2.02  0.0588        18   -0.697    0.0142 Paire~
## # ... with 1 more variable: alternative <chr>
```

```
#Bit surprising but it demonstrates that there is no statistical significance between days
#So combining values to start with is probably appropriate
```

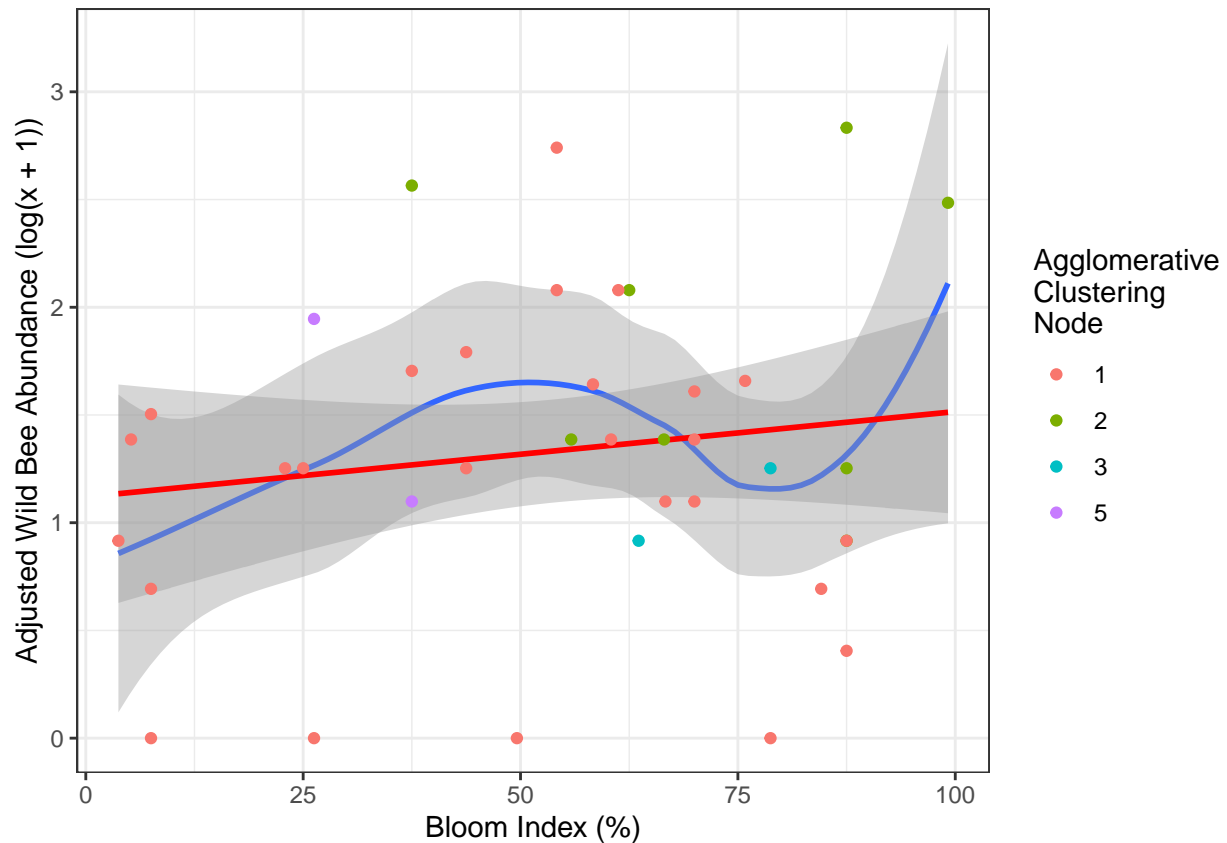
Bloom

First it shall be graphed to see if there is any obvious correlation.

```
data_2012 %>%
  inner_join(orchard_node_agglom) %>%
  ggplot(aes(x = bloom.index, y = log(wildAbF + 1))) +
  geom_smooth() +
  geom_smooth(method = "lm", colour = "red") +
  theme_bw() +
  geom_point(aes(colour = as_factor(node - 7))) +
  labs(x = "Bloom Index (%)", y = "Adjusted Wild Bee Abundance (log(x + 1))",
       # title = "Wild Bee Abundance Coloured by Agglomerature Nodes",
       colour = "Agglomerative\nClustering\nNode")
```

```
## Joining, by = "orchard"
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



There doesn't seem to be any obvious correlation, but just to be sure I'll split the bloom into three categories and test these with an Anova test.

```
logged_data <- logged_data %>%
  mutate(bloom_category = if_else(bloom.index <= 33, 1,
    if_else(bloom.index <= 66, 2, 3)))

#Running the anova test
anova1 <- aov(adjusted_bees ~ bloom_category, data = logged_data)
summary(anova1)
```

	##	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bloom_category	1	0.19	0.1897	0.456	0.504	
Residuals	36	14.98	0.4162			

*#As expected there is no evidence to suggest that bloom index (category)
#Has an effect on the bee counts (although only 1 DF in bloom_category thought it should be 2??)*

Bloom is a continuous variable but for the sakes of deciding whether to include it or not

Final Confounders

Based on the analysis above, the confounders that could be taken into account are: Region and local diversity based on the statistical tests, temperature and X2000nat based on the literature.

```
#Checking the variables to see if any are heavily correlated (binary/categorical ones won't be)
#So checking other two and we get a low correlation so acceptable to use them together
cor(logged_data$temp, logged_data$X2000nat)
```

```
## [1] 0.2065843
f_lm <- lm(adjusted_beets ~ temp + X2000nat + local.diversity + region, data = logged_data)
#Interesting values here seem to suggest that only significant one is X2000nat
summary(f_lm)

##
## Call:
## lm(formula = adjusted_beets ~ temp + X2000nat + local.diversity +
##     region, data = logged_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23257 -0.35949 -0.06686  0.23707  1.30560
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.33682    0.61856  -0.545  0.58985
## temp           0.08343    0.02620   3.184  0.00323 **
## X2000nat       0.09182    0.79392   0.116  0.90865
## local.diversity 0.20663    0.28923   0.714  0.48014
## regionLO      -0.42271    0.36305  -1.164  0.25290
## regionS       -0.08406    0.46728  -0.180  0.85837
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5782 on 32 degrees of freedom
## Multiple R-squared:  0.2949, Adjusted R-squared:  0.1848
## F-statistic: 2.677 on 5 and 32 DF,  p-value: 0.03941
```

Notes, from running the adjustments: When you use temperature as an original adjustment, only significant one in the model is X2000nat and when you use X2000nat in the original data only significant one is temp. When you use neither and just apply the linear model on the logged data then both temp and nat are significant -> suggesting that just these two variables would be best to adjust by. This suggest to me that fitting a joint lm using both temp and X2000nat to adjust the data is probably best to avoid overfitting and keep the model as parsimonious as possible. Despite region and local.diversity being significant for temp adjustment originally, unlikely to be worth implementing but could be considered to cover model changes section. To take this further, could apply lasso regression on the full dataset to see what that comes up with as tends to have a lower MSE than a standard LM.

#Using the logged_data with no adjustments

```
logged_data <- data_2012 %>%
  inner_join(clusterings) %>%
  mutate(adjusted_beets = log(wildAbF + 1)) %>%
  mutate(adjusted_social = log(socialRichF + 1))
```

```
## Joining, by = "orchard"
```

#Looking at the model ~ only X2000nat and temp significant

```
lm(adjusted_beets ~ temp + X2000nat + local.diversity + region + day + bloom.index, data = logged_data)
summary
```

```
##
## Call:
## lm(formula = adjusted_beets ~ temp + X2000nat + local.diversity +
##     region + day + bloom.index, data = logged_data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12852 -0.32604 -0.09243  0.25194  1.30949
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.380739    0.641776  -2.151   0.0396 *
## temp           0.071839    0.027191   2.642   0.0130 *
## X2000nat       2.187169    0.791708   2.763   0.0097 **
## local.diversity 0.220885    0.288239   0.766   0.4495
## regionLO      -0.495222    0.368157  -1.345   0.1887
## regionS       -0.052776    0.469875  -0.112   0.9113
## day           0.164934    0.201587   0.818   0.4197
## bloom.index    0.004444    0.004698   0.946   0.3517
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5756 on 30 degrees of freedom
## Multiple R-squared:  0.4896, Adjusted R-squared:  0.3705
## F-statistic: 4.111 on 7 and 30 DF,  p-value: 0.002834

#Defining linear model with temp and X2000nat as factors
temp_nat_lm <- tidy(lm(adjusted_beets ~ temp + X2000nat, data = logged_data))
#Extracting temp and nat estimate values
temp_factor <- temp_nat_lm %>%
  select(estimate) %>%
  slice(2) %>%
  as_vector()
nat_factor <- temp_nat_lm %>%
  select(estimate) %>%
  slice(3) %>%
  as_vector()

#Now adjustments shall be made for these two variables:
#Temperature will be set to 20 degrees
#X2000 nat will be set to the mean x2000 nat value
temp_nat_adjusted <- logged_data %>%
  mutate(adjusted_beets = (adjusted_beets + ((20 - temp) * temp_factor) + ((mean_x2000 - X2000nat) * X2000nat)))

#Checking that after the adjustment they are not significant at all in the model
tidy(lm(adjusted_beets ~ temp + X2000nat, data = temp_nat_adjusted))

## # A tibble: 3 x 5
##   term      estimate std.error statistic p.value
##   <chr>      <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept) 1.24e+ 0    0.549    2.25e+ 0    0.0309
## 2 temp       -9.42e-18    0.0249  -3.79e-16    1.000
## 3 X2000nat     9.12e- 2    0.735    1.24e- 1    0.902

#As expected now the bee abundance has been adjusted for these two parameters
```

The same process can be applied to the variable SocialRichF, the other bee variable which needs to be considered. Since we know that only temp and nat were statistic on the wildAbF we can go straight to the LM and test that (May need to check this with Julia, if not just cp previous analysis...)


```

#Looking at the model ~ only X2000nat significant
lm(adjusted_social ~ temp + X2000nat + local.diversity + region + day + bloom.index, data = logged_data)
summary

##
## Call:
## lm(formula = adjusted_social ~ temp + X2000nat + local.diversity +
##     region + day + bloom.index, data = logged_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60609 -0.21944 -0.01005  0.22148  0.59571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.528723   0.378485  -1.397  0.17268
## temp          -0.002537   0.016036  -0.158  0.87534
## X2000nat       1.562216   0.466907   3.346  0.00222 **
## local.diversity -0.155875   0.169988  -0.917  0.36647
## regionLO      -0.137446   0.217119  -0.633  0.53150
## regionS        0.168870   0.277107   0.609  0.54684
## day            0.203878   0.118885   1.715  0.09667 .
## bloom.index    0.002467   0.002771   0.890  0.38045
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3395 on 30 degrees of freedom
## Multiple R-squared:  0.4099, Adjusted R-squared:  0.2723
## F-statistic: 2.978 on 7 and 30 DF,  p-value: 0.01703

social_lm<- tidy(lm(adjusted_social ~ X2000nat, data = logged_data))

social_nat_factor <- social_lm %>%
  select(estimate) %>%
  slice(2) %>%
  as_vector()

#Adjustment for social added
temp_nat_adjusted <- temp_nat_adjusted %>%
  mutate(adjusted_social = (adjusted_social + ((mean_x2000 - X2000nat) * X2000_factor)))

```

Clustering Summarising

Functions involved with summarising the clusters:

```

#Gets the name of a variable as a string ~ now implemented in clusterise but keeping
get_name <- function(x) {
  deparse(substitute(x))
}

#Gets the bee variables that I am summarising for each cluster - makes it easy to change in future
clusterise <- function(bee_data, cluster_data, v_name = deparse(substitute(cluster_data))){
  inner_join(bee_data, cluster_data) %>%
    summarise(mean_honey_ab = mean(mean_honey_ab),

```

```

        mean_wild_ab = mean(mean_wild_ab),
        mean_social_rich = mean(mean_social_rich),
        n = n()) %>%
    mutate(cluster = v_name)
}

#The function to output the summaries of both agglom and kmeans clustering
cluster_summarise <- function(bee_data){
output <- list()

output$agglom_beas <- clusterise(bee_data, aggl_node1) %>%
  bind_rows(clusterise(bee_data, aggl_node2)) %>%
  bind_rows(clusterise(bee_data, aggl_node3)) %>%
  bind_rows(clusterise(bee_data, aggl_node4)) %>%
  bind_rows(clusterise(bee_data, aggl_node5)) %>%
  bind_rows(clusterise(bee_data, aggl_node6)) %>%
  bind_rows(clusterise(bee_data, aggl_node7)) %>%
  bind_rows(clusterise(bee_data, aggl_node4) %>% filter(cluster == 1), "agglom_node8")) %>%
  bind_rows(clusterise(bee_data, aggl_node4) %>% filter(cluster == 2), "agglom_node9")) %>%
  bind_rows(clusterise(bee_data, aggl_node5) %>% filter(cluster == 1), "agglom_node10")) %>%
  bind_rows(clusterise(bee_data, aggl_node5) %>% filter(cluster == 2), "agglom_node11")) %>%
  bind_rows(clusterise(bee_data, aggl_node6) %>% filter(cluster == 1), "agglom_node12")) %>%
  bind_rows(clusterise(bee_data, aggl_node6) %>% filter(cluster == 2), "agglom_node13")) %>%
  #Only 1 result for cluster as none in the latter cluster
  #Unable to to bind as no common variables for an empty node15
  bind_rows(clusterise(bee_data, aggl_node7) %>% filter(cluster == 1), "agglom_node14"))

output$kmeans_beas <- clusterise(bee_data, kmeans_node1) %>%
  bind_rows(clusterise(bee_data, kmeans_node2)) %>%
  bind_rows(clusterise(bee_data, kmeans_node3)) %>%
  bind_rows(clusterise(bee_data, kmeans_node4)) %>%
  bind_rows(clusterise(bee_data, kmeans_node5)) %>%
  bind_rows(clusterise(bee_data, kmeans_node6)) %>%
  bind_rows(clusterise(bee_data, kmeans_node7)) %>%
  bind_rows(clusterise(bee_data, kmeans_node4) %>% filter(cluster == 1), "kmeans_node8")) %>%
  bind_rows(clusterise(bee_data, kmeans_node4) %>% filter(cluster == 2), "kmeans_node9")) %>%
  bind_rows(clusterise(bee_data, kmeans_node5) %>% filter(cluster == 1), "kmeans_node10")) %>%
  bind_rows(clusterise(bee_data, kmeans_node5) %>% filter(cluster == 2), "kmeans_node11")) %>%
  bind_rows(clusterise(bee_data, kmeans_node6) %>% filter(cluster == 1), "kmeans_node12")) %>%
  bind_rows(clusterise(bee_data, kmeans_node6) %>% filter(cluster == 2), "kmeans_node13")) %>%
  bind_rows(clusterise(bee_data, kmeans_node7) %>% filter(cluster == 1), "kmeans_node14")) %>%
  bind_rows(clusterise(bee_data, kmeans_node7) %>% filter(cluster == 2), "kmeans_node15"))

#Return a list containing two tibbles one for each type of clustering algorithm
output
}

```

Original summarising without any variable correction

```

#original Bee data (logged has no changes to it)
bee_values <- logged_data %>%
  group_by(orchard) %>%
  summarise(mean_honey_ab = mean(apisAb),

```

```

    mean_wild_ab = mean(adjusted_bees),
    mean_social_rich = mean(adjusted_social)
  )

original_bee_summary <- cluster_summarise(bee_data = bee_values)

#Agglom
original_bee_summary$agglom_beas

## # A tibble: 14 x 5
##   mean_honey_ab mean_wild_ab mean_social_rich     n cluster
##   <dbl>         <dbl>         <dbl> <int> <chr>
## 1         6.56         1.33         0.379    19 aggl_node1
## 2         6.44         1.32         0.389    18 aggl_node2
## 3         8.75         1.52         0.203     1 aggl_node3
## 4         6.70         1.34         0.391    17 aggl_node4
## 5          2         1.08         0.347     1 aggl_node5
## 6         8.75         1.52         0.203     1 aggl_node6
## 7        NaN         NaN         NaN       0 aggl_node7
## 8         6.01         1.17         0.358    13 agglom_node8
## 9         8.94         1.86         0.499     4 agglom_node9
## 10         2         1.08         0.347     1 agglom_node10
## 11        NaN         NaN         NaN       0 agglom_node11
## 12         8.75         1.52         0.203     1 agglom_node12
## 13        NaN         NaN         NaN       0 agglom_node13
## 14        NaN         NaN         NaN       0 agglom_node14

#Kmeans
original_bee_summary$kmeans_beas

## # A tibble: 15 x 5
##   mean_honey_ab mean_wild_ab mean_social_rich     n cluster
##   <dbl>         <dbl>         <dbl> <int> <chr>
## 1         6.56         1.33         0.379    19 kmeans_node1
## 2         6.81         1.35         0.381    18 kmeans_node2
## 3          2         1.08         0.347     1 kmeans_node3
## 4         6.81         1.35         0.381    18 kmeans_node4
## 5        NaN         NaN         NaN       0 kmeans_node5
## 6          2         1.08         0.347     1 kmeans_node6
## 7        NaN         NaN         NaN       0 kmeans_node7
## 8         6.21         1.20         0.347    14 kmeans_node8
## 9         8.94         1.86         0.499     4 kmeans_node9
## 10        NaN         NaN         NaN       0 kmeans_node10
## 11        NaN         NaN         NaN       0 kmeans_node11
## 12         2         1.08         0.347     1 kmeans_node12
## 13        NaN         NaN         NaN       0 kmeans_node13
## 14        NaN         NaN         NaN       0 kmeans_node14
## 15        NaN         NaN         NaN       0 kmeans_node15

##Original summarising of non-markov approach
maximum_cluster

##   eqiB11F.pre eqiB11I.pre eqiB11F.blm eqiB11I.blm eqiB11T.blm eqiB11F.pos
## 1  0.56294059 -0.3418658 -0.02661776  0.80366758 -0.24200410 -0.83657918
## 2  0.88788914 -0.3418658 -0.19022893 -0.30585457 -0.30620065  0.01058593

```

```
## 3  0.88788914 -0.3418658 -0.19022893 -0.30585457 -0.30620065  0.01058593
## 4  0.88788914 -0.3418658 -0.19022893 -0.30585457 -0.30620065  0.01058593
## 5 -1.21724288 -0.3418658 -0.47141495  0.25463750 -0.26826632  0.05673110
## 6 -1.78818873 -0.3418658 -0.71087017  2.78888695  0.36494510 -0.33951546
## 7  0.92973857  1.1014288  1.78834024  0.06780681 -0.26972533  2.88563060
## 8  0.01045773 -0.3418658  2.26526351 -0.36201840  0.01478210  1.10101634
## 9 -0.28688425  0.2579449 -0.21970544  0.87817061 -0.28577447 -0.22916831
## 10 -0.05899826 -0.3418658 -1.93232366 -1.17238220 -0.30620065  0.53623960
## 11  1.50859712 -0.1356809  0.15454685 -0.81247522 -0.29744657 -0.54716872
## 12  1.44670216 -0.3068769  0.15156608 -0.81247522 -0.30167771 -0.56622868
## 13 -0.33664640 -0.3418658 -0.18559218 -0.44683724 -0.30124000 -0.62240541
## 14 -1.63608890 -0.3418658  0.37048710  0.12282444 -0.30620065 -0.53312454
## 15 -1.58069994  3.8693055 -1.36167172 -0.74943419 -0.30620065 -1.18316953
## 16 -0.14885146 -0.3418658  0.30623494 -0.51446307  4.07083681 -0.08621861
## 17 -0.08097801 -0.3418658 -0.85792151  1.91089732 -0.28285645 -0.80197030
## 18 -0.02629240 -0.3418658 -0.15710926 -0.07432206 -0.05816852 -0.72071468
## 19  0.03876764 -0.3418658  1.45747473 -0.96491990 -0.30620065  1.85488796
##      eiqB11I.pos eiqB11T.pos cluster
## 1  0.07606715 -0.29236499      1
## 2  1.78589159 -0.29236499      1
## 3  1.78589159 -0.29236499      1
## 4  1.78589159 -0.29236499      1
## 5 -0.67656322 -0.29236499      1
## 6  0.61131654 -0.29236499      1
## 7  1.48904351 -0.23224030      1
## 8 -0.90112187 -0.29236499      1
## 9 -0.10337469 -0.17211561      1
## 10 -1.07800026 -0.29236499      1
## 11 -0.07517669 -0.16918270      1
## 12 -0.07517669  0.04491986      1
## 13 -0.78166487 -0.20291118      1
## 14 -1.05339109 -0.29236499      1
## 15 -0.68425358 -0.29236499      2
## 16 -0.61093877  4.10700262      2
## 17 -0.65349212 -0.29236499      1
## 18 -0.28179116  0.03025530      1
## 19 -0.55915698 -0.18971308      1
```

```
#Clustering summaries based on non-markov approach
non_markov_clusters <- logged_data %>%
  group_by(orchard) %>%
  summarise(mean_honey_ab = mean(apisAb),
            mean_wild_ab = mean(adjusted_bees),
            mean_social_rich = mean(adjusted_social)) %>%
  mutate(cluster = maximum_cluster$cluster) %>%
  group_by(cluster) %>%
  summarise(mean_honey_ab = mean(mean_honey_ab),
            mean_wild_ab = mean(mean_wild_ab),
            mean_social_rich = mean(mean_social_rich),
            n = n())
```

After Adjustment:

```
adjusted_bee_values <- temp_nat_adjusted %>%
  group_by(orchard) %>%
```

```

summarise(mean_honey_ab = mean(apisAb),
           mean_wild_ab = mean(adjusted_bees),
           mean_social_rich = mean(adjusted_social)
)

adjusted_bee_summary <- cluster_summarise(bee_data = adjusted_bee_values)

#Agglom
adjusted_bee_summary$agglom_bees

```

```

## # A tibble: 14 x 5
##   mean_honey_ab mean_wild_ab mean_social_rich     n cluster
##   <dbl>         <dbl>         <dbl> <int> <chr>
## 1         6.56         1.27         0.379    19 aggl_node1
## 2         6.44         1.25         0.383    18 aggl_node2
## 3         8.75         1.57         0.303     1 aggl_node3
## 4         6.70         1.28         0.397    17 aggl_node4
## 5          2         0.735         0.140     1 aggl_node5
## 6         8.75         1.57         0.303     1 aggl_node6
## 7        NaN         NaN         NaN       0 aggl_node7
## 8         6.01         1.24         0.434    13 agglom_node8
## 9         8.94         1.44         0.278     4 agglom_node9
## 10          2         0.735         0.140     1 agglom_node10
## 11        NaN         NaN         NaN       0 agglom_node11
## 12         8.75         1.57         0.303     1 agglom_node12
## 13        NaN         NaN         NaN       0 agglom_node13
## 14        NaN         NaN         NaN       0 agglom_node14

```

```

#Kmeans
adjusted_bee_summary$kmeans_bees

```

```

## # A tibble: 15 x 5
##   mean_honey_ab mean_wild_ab mean_social_rich     n cluster
##   <dbl>         <dbl>         <dbl> <int> <chr>
## 1         6.56         1.27         0.379    19 kmeans_node1
## 2         6.81         1.30         0.392    18 kmeans_node2
## 3          2         0.735         0.140     1 kmeans_node3
## 4         6.81         1.30         0.392    18 kmeans_node4
## 5        NaN         NaN         NaN       0 kmeans_node5
## 6          2         0.735         0.140     1 kmeans_node6
## 7        NaN         NaN         NaN       0 kmeans_node7
## 8         6.21         1.26         0.425    14 kmeans_node8
## 9         8.94         1.44         0.278     4 kmeans_node9
## 10        NaN         NaN         NaN       0 kmeans_node10
## 11        NaN         NaN         NaN       0 kmeans_node11
## 12          2         0.735         0.140     1 kmeans_node12
## 13        NaN         NaN         NaN       0 kmeans_node13
## 14        NaN         NaN         NaN       0 kmeans_node14
## 15        NaN         NaN         NaN       0 kmeans_node15

```

```

##Non-Markov
##Maximum
#Clustering summaries based on non-markov approach
maximum_non_markov_clusters <- temp_nat_adjusted %>%
  group_by(orchard) %>%

```

```

summarise(mean_honey_ab = mean(apisAb),
           mean_wild_ab = mean(adjusted_bees),
           mean_social_rich = mean(adjusted_social)) %>%
mutate(cluster = maximum_cluster$cluster) %>%
group_by(cluster) %>%
summarise(mean_honey_ab = mean(mean_honey_ab),
           mean_wild_ab = mean(mean_wild_ab),
           mean_social_rich = mean(mean_social_rich),
           n = n())

##euclidean:
euclidean_non_markov_clusters <- temp_nat_adjusted %>%
group_by(orchard) %>%
summarise(mean_honey_ab = mean(apisAb),
           mean_wild_ab = mean(adjusted_bees),
           mean_social_rich = mean(adjusted_social)) %>%
mutate(cluster = euclidean_cluster$cluster) %>%
group_by(cluster) %>%
summarise(mean_honey_ab = mean(mean_honey_ab),
           mean_wild_ab = mean(mean_wild_ab),
           mean_social_rich = mean(mean_social_rich),
           n = n())

```

Although honey bee abundance is one of the summarised variables, this is likely to be related directly to the `hive.acr` variable.

Comparing cluster values before and after adjustments, for now just looking at the agglomerative node clustering selection as it is, in my opinion, better than Kmeans.

```

##agglom
adjusted_bee_summary$agglom_bees %>%
select(mean_wild_ab, mean_social_rich ,cluster) %>%
rename(adjusted = mean_wild_ab) %>%
rename(adjusted_social = mean_social_rich) %>%
inner_join(original_bee_summary$agglom_bees %>%
           select(mean_wild_ab, mean_social_rich, cluster)) %>%
select(cluster,mean_wild_ab, adjusted, mean_social_rich, adjusted_social)

```

```
## Joining, by = "cluster"
```

```
## # A tibble: 14 x 5
```

##	cluster	mean_wild_ab	adjusted	mean_social_rich	adjusted_social
##	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
##	1 aggl_node1	1.33	1.27	0.379	0.379
##	2 aggl_node2	1.32	1.25	0.389	0.383
##	3 aggl_node3	1.52	1.57	0.203	0.303
##	4 aggl_node4	1.34	1.28	0.391	0.397
##	5 aggl_node5	1.08	0.735	0.347	0.140
##	6 aggl_node6	1.52	1.57	0.203	0.303
##	7 aggl_node7	NaN	NaN	NaN	NaN
##	8 agglom_node8	1.17	1.24	0.358	0.434
##	9 agglom_node9	1.86	1.44	0.499	0.278
##	10 agglom_node10	1.08	0.735	0.347	0.140
##	11 agglom_node11	NaN	NaN	NaN	NaN
##	12 agglom_node12	1.52	1.57	0.203	0.303
##	13 agglom_node13	NaN	NaN	NaN	NaN

```
## 14 agglom_node14      NaN      NaN      NaN      NaN
```

I guess potentially can see two *paths* one with a mean slightly lower than the other if you think of the structure.

Looking at clusters

The aim of this section is to look at the clusterings from a low, medium, high point of view...

```
#Non_markov clustering summary maximum
maximum_final_standardised <- maximum_cluster %>%
  group_by(cluster) %>%
  summarise_all(funs(mean)) %>%
  inner_join(maximum_non_markov_clusters)

## Joining, by = "cluster"

#Non_markov clustering summary maximum
euclidean_final_standardised <- euclidean_cluster %>%
  group_by(cluster) %>%
  summarise_all(funs(mean)) %>%
  inner_join(euclidean_non_markov_clusters)

## Joining, by = "cluster"

#Cluster 1 -> Low insecticide, High Fungicide, low thinner
#Cluster 2 -> High insecticide, Low Fungicide, low (no) thinner
#Cluster 3 -> middle insecticide, High Fungicide, low thinner
#Cluster 4 -> Low insecticide, High Fungicide, High thinner

#Agglom clustering summary (end stage)

bee_summary_agglom <- adjusted_bee_summary$agglom_bees %>%
  slice(8:14) %>%
  select(-cluster) %>%
  mutate(cluster = c(1,2,3,4,5,6,7))

pesticide_summary_agglom <- aggl_c(aggl_node2, ".pos", 1) %>%
  bind_rows(aggl_c(aggl_node2, ".pos", 2) %>%
    mutate(cluster = if_else(cluster == 1, 3, 4))) %>%
  bind_rows(aggl_c(aggl_node3, ".pos", 1) %>%
    mutate(cluster = if_else(cluster == 1, 5, 6))) %>%
  bind_rows(aggl_c(aggl_node3, ".pos", 2) %>%
    mutate(cluster = if_else(cluster == 1, 7, 8))) %>%
  group_by(cluster) %>%
  summarise_all(funs(mean))

##

agglom_summary_standardised <- bee_summary_agglom %>%
  inner_join(pesticide_summary_agglom) %>%
  select(-orchard)

## Joining, by = "cluster"
```

```

# select(-orchard) %>%
# #1 for high, 0 for low, just using bloom
# mutate(pest_rating = c(0,0,1,1,0,0,1)) %>%
# mutate(insect_rating = c(1,0,1,0,0,1,1)) %>%
# mutate(thinner_rating = c(1,0,0,1,0,1,0))

bee_summary_kmeans <- adjusted_bee_summary$kmeans_bees %>%
  slice(8:14) %>%
  select(-cluster) %>%
  mutate(cluster = c(1,2,3,4,5,6,7))

pesticide_summary_kmeans <- aggl_c(kmeans_node2, ".pos", 1) %>%
  bind_rows(aggl_c(kmeans_node2, ".pos", 2) %>%
    mutate(cluster = if_else(cluster == 1, 3, 4))) %>%
  bind_rows(aggl_c(kmeans_node3, ".pos", 1) %>%
    mutate(cluster = if_else(cluster == 1, 5, 6))) %>%
  bind_rows(aggl_c(kmeans_node3, ".pos", 2) %>%
    mutate(cluster = if_else(cluster == 1, 7, 8))) %>%
  group_by(cluster) %>%
  summarise_all(funs(mean))

##
kmeans_summary_standardised <- bee_summary_kmeans %>%
  inner_join(pesticide_summary_kmeans) %>%
  select(-orchard)

## Joining, by = "cluster"

#write.csv(agglom_summary, "agglom_summary.csv")

#Overall
#HLH
#HHL
#HHL
#HLH
#LLL
#LHH
#LLL

#Just Bloom
#LHH
#LLL
#HHL
#HLH
#LLL
#LHH
#HHL

##Just bloom and just fungicide and insecticide

```


Looking at the difference between standardised data and non-standardised data: This section is non-standardised

write up explanation of aggregation before vs after, conditional probabilities along with the tree would be a good idea This section is going cover two step clustering based on various clustering methods for comparison. Visit data will be combined into one bee result for each orchard, note in principle this is fundamentally flawed as the data does depend on previous years and the assumption that independent of the past is false as pesticide last year affects bee population this year as would be expected. In these scenarios each implementation will have a maximum of 8 clusters. Two after pre-bloom, 4 after and 8 after post bloom.

```
load("data")
markov_data <- data_2012 %>%
  select(ends_with(".pre"), ends_with(".blm"), ends_with(".pos"), orchard) %>%
  unique()
#function to generate the agglomeration clustering
aggl <- function(data){
  agnes(dist(data, method = "euclidian"),
        diss=TRUE, method = "ward")
}

#function to reduce copy pasting for each node
aggl_c <- function(data, ends, c_num){
  #define temp dataset of this stage by cluster
  temp <- data %>%
    filter(cluster == c_num) %>%
    select(ends_with(ends))

  #If statement as if 1 element in dataset it'll error
  if(nrow(temp) > 1){
    temp$cluster <- cutree(aggl(temp) , k = 2)

    output <- data %>%
      filter(cluster == c_num) %>%
      select(-cluster) %>%
      mutate(cluster = temp$cluster)
    #Else statement for if 1 element in dataset
  }else{
    #If only 1 element in the dataset -> set the cluster number to 1
    output <- data %>%
      filter(cluster == c_num) %>%
      select(-cluster) %>%
      mutate(cluster = 1)
  }
  output
}

####Agglomerative hierarchical clustering
#pre-bloom step
temp <- markov_data %>%
  select(ends_with(".pre"))

temp$cluster <- cutree(aggl(temp) , k = 2)
```

```

aggl_node1 <- markov_data %>%
  mutate(cluster = temp$cluster)

#during bloom step 1
aggl_node2 <- aggl_c(aggl_node1, ".blm", 1)

#during bloom step 2
aggl_node3 <- aggl_c(aggl_node1, ".blm", 2)

#post bloom step 1
aggl_node4 <- aggl_c(aggl_node2, ".pos", 1)

#post bloom step 2
aggl_node5 <- aggl_c(aggl_node2, ".pos", 2)

#post bloom step 3
aggl_node6 <- aggl_c(aggl_node3, ".pos", 1)

#post bloom step 4
#Note this node only has 1 element in it so it is automatically set to 1
aggl_node7 <- aggl_c(aggl_node3, ".pos", 2)

orchard_node_agglom <- tibble(orchard = aggl_node4 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node_id = 1)
bind_rows(tibble(orchard = aggl_node4 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node_id = 2))
bind_rows(tibble(orchard = aggl_node5 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node_id = 3))
bind_rows(tibble(orchard = aggl_node5 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node_id = 4))
bind_rows(tibble(orchard = aggl_node6 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node_id = 5))
bind_rows(tibble(orchard = aggl_node6 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, node_id = 6))
bind_rows(tibble(orchard = aggl_node7 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, node_id = 7))

```

kmeans approach: Even running the 1001 times it is different everytime: store output later on but run it like 1million times at each stage.

```

#function to generate the kmeans clustering
k_clustering <- function(data, n = 1001){
  #creating the dataset to then calculate the optimal cluster from
  for(i in 1:n){
    if(i == 1){
      k_list <- tibble(kmeans(data, 2)$cluster)
    }else{
      k_list <- bind_cols(k_list, tibble(kmeans(data, 2)$cluster))
    }
  }
  apply(k_list[, 1:length(k_list)], 1, mfv1) %>%
    enframe(value = "cluster") %>%
    select(cluster)
}

#pre-bloom step
#Getting the most common clustering

kmeans_c <- function(data, ends, c_num){

```

```

data <- data %>%
  filter(cluster == c_num) %>%
  select(-cluster)

if(nrow(data) > 2){
  temp <- data %>%
    select(ends_with(ends))

  cluster <- k_clustering(temp, 2)
  output <- bind_cols(data, cluster)
}else{
  #Not more than 2 rows and get the error message as:
  #number of cluster centres must lie between 1 and nrow(x)
  output <- data %>%
    mutate(cluster = 1)
}

output
}

temp <- k_clustering(markov_data %>% select(-orchard))

kmeans_node1 <- markov_data %>%
  bind_cols(temp)

#during bloom step 1
kmeans_node2 <- kmeans_c(kmeans_node1, ".blm", 1)

#during bloom step 2
kmeans_node3 <- kmeans_c(kmeans_node1, ".blm", 2)

#post bloom step 1
kmeans_node4 <- kmeans_c(kmeans_node2, ".pos", 1)

#post bloom step 2
kmeans_node5 <- kmeans_c(kmeans_node2, ".pos", 2)

#post bloom step 3
kmeans_node6 <- kmeans_c(kmeans_node3, ".pos", 1)

#post bloom step 4
kmeans_node7 <- kmeans_c(kmeans_node3, ".pos", 2)

#End timepoint nodes
orchard_node_kmeans <- tibble(orchard = kmeans_node4 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector,
  bind_rows(tibble(orchard = kmeans_node4 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector,
  bind_rows(tibble(orchard = kmeans_node5 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector,
  bind_rows(tibble(orchard = kmeans_node5 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector,
  bind_rows(tibble(orchard = kmeans_node6 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector,
  bind_rows(tibble(orchard = kmeans_node6 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector,

```

```
bind_rows(tibble(orchard = kmeans_node7 %>% filter(cluster == 1) %>% select(orchard) %>% as_vector, n
bind_rows(tibble(orchard = kmeans_node7 %>% filter(cluster == 2) %>% select(orchard) %>% as_vector, n
```

Just a look at the difference in the final clusters between the two methods

```
clusterings <- orchard_node_agglom %>%
  arrange(orchard) %>%
  bind_cols(node_km = orchard_node_kmeans %>% arrange(orchard) %>%
    select(node) %>% as_vector)
clusterings
```

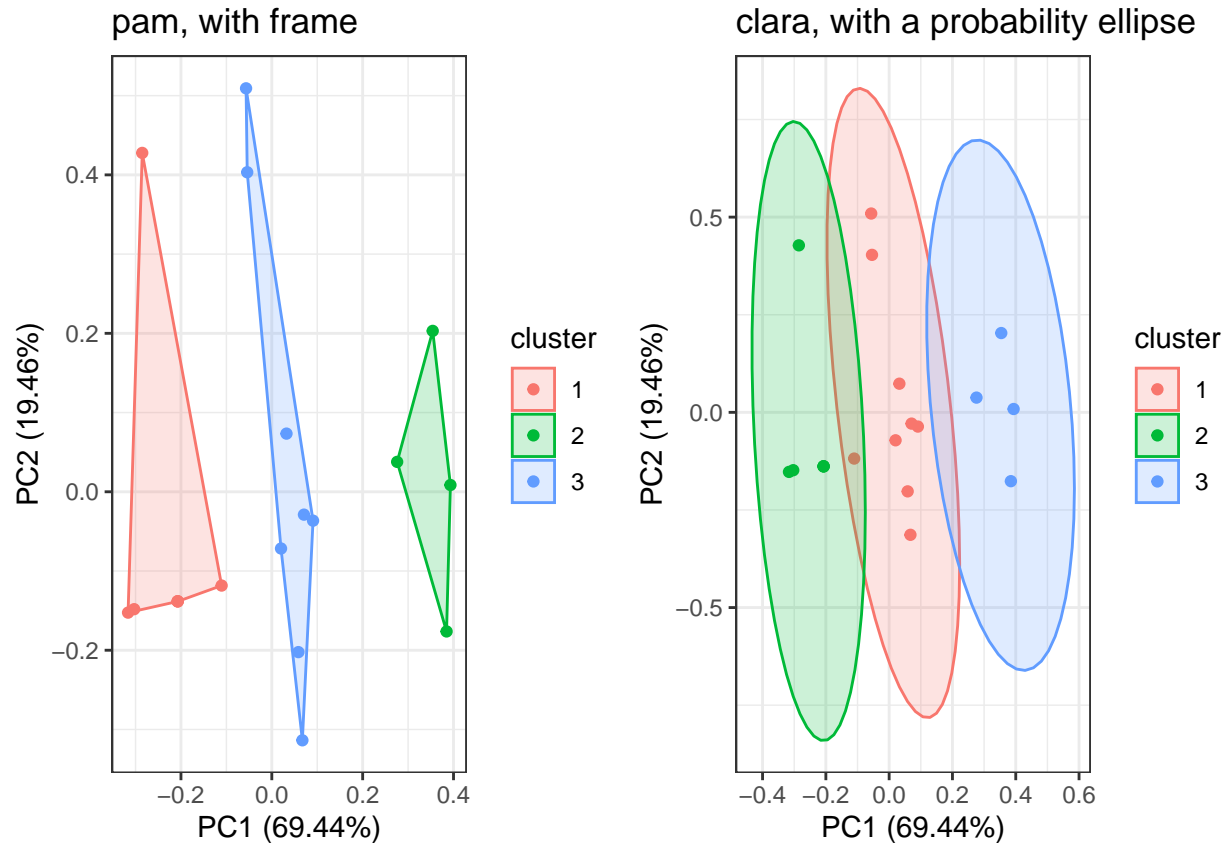
```
## # A tibble: 19 x 3
##   orchard node node_km
##   <fct>   <dbl>   <dbl>
## 1 A      8      10
## 2 B      9      10
## 3 C      9      10
## 4 D      9      10
## 5 E     12      12
## 6 F     13      13
## 7 G     10      8
## 8 H.nooil 11      9
## 9 I      8      10
## 10 J      8      9
## 11 K      8      10
## 12 L      8      10
## 13 M      8      10
## 14 N     14      12
## 15 O.nooil 12      12
## 16 P      8      10
## 17 Q      8      9
## 18 R      8      10
## 19 S     11      9
```

Looking at top two PCAs just to see how much variation is absorbed by these

```
p1 <- autoplot(pam(markov_data, 3), frame = TRUE) +
  theme_bw() +
  ggtitle("pam, with frame")

p2 <- autoplot(clara(markov_data, 3), frame = TRUE, frame.type = 'norm') +
  theme_bw() +
  ggtitle("clara, with a probability ellipse")

grid.arrange(p1, p2 ,nrow = 1)
```



Non-markov structuring

This section will approach looking at the data as one time point with 1 clustering point rather than 3 as above: Method Justification - Maximum linkage function is best for both standardised and unstandardised.

```
maximum_cluster <- data_2012 %>%
  arrange(orchard) %>%
  select(ends_with(".pre"), ends_with(".blm"), ends_with(".pos")) %>%
  filter(row_number() %% 2 == 1)

euclidean_cluster <- maximum_cluster

# matrix of methods to compare,
# rerun - without %>%mutate segment for non-standardised values
m <- c("average", "single", "complete", "ward")
names(m) <- c("average", "single", "complete", "ward")
distances <- c("euclidean", "maximum", "manhattan", "canberra",
               "minkowski")
names(distances) <- c("euclidean", "maximum", "manhattan",
                     "canberra", "minkowski")
clust_comps <- matrix(nrow = length(distances), ncol = length(m),
                     dimnames = list(distances, m))

# function to compute coefficient to see which is the best method
ac <- function(distance, linkage) {
  dista <- dist(maximum_cluster, method = distance)
```

```

#Agglomerative Nesting form of Hierarchical Clustering
agnes(dista, method = linkage)$ac
}
for(i in 1:length(distances)) {
  for(j in 1:length(m)) {
    clust_comps[i,j] <- ac(distances[i], m[j])
  }
}

#In future or in write up, perhaps change, needs more clustering analysis
#Look at literature in more detail to see other benefits/drawbacks and assumptions
#Then decide on actual best method, for now using euclidian as used that before
clust_comps = clust_comps

```

Clustering code - Maximum

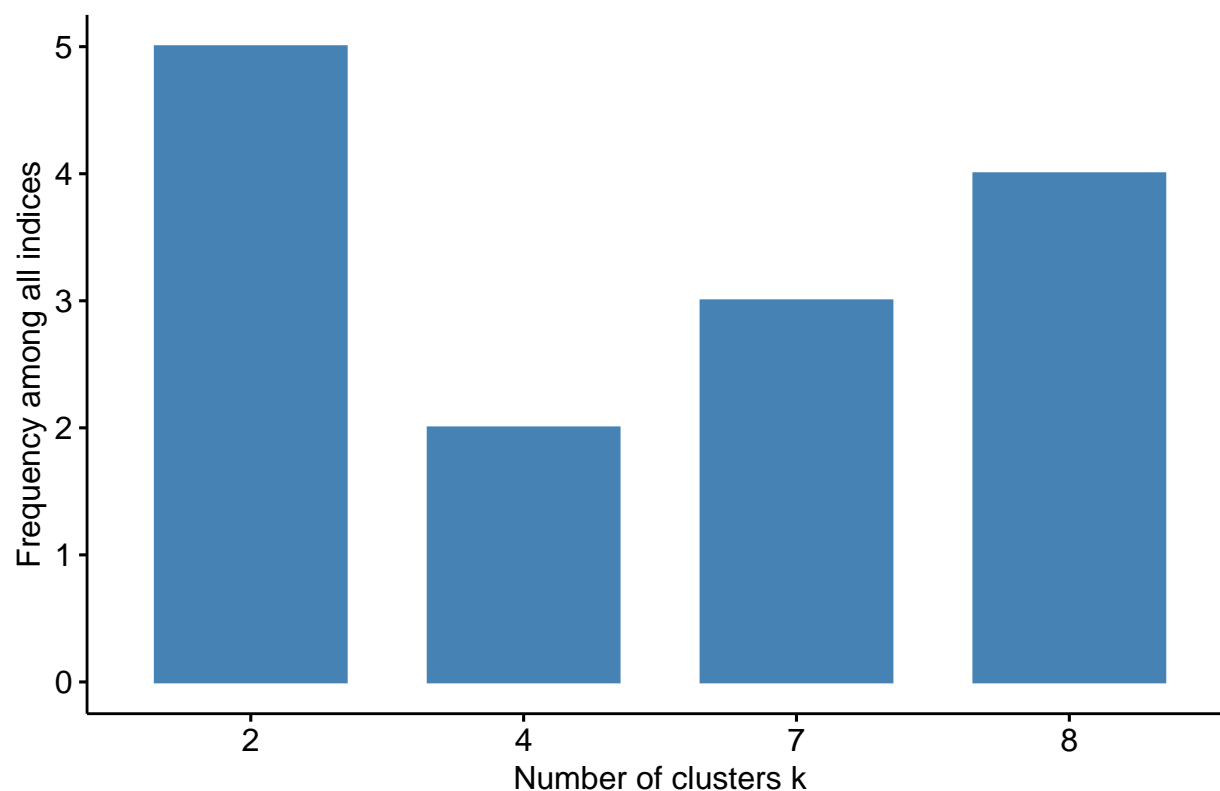
```

#NBclust function breaks when using index = all, instead using all the working indexes instead
fviz_nbclust(NbClust(maximum_cluster, distance = "maximum",
                     min.nc = 2, max.nc = 8, method = "ward.D2", index = c("kl", "ch", "ccc"

## Among all indices:
## =====
## * 5 proposed 2 as the best number of clusters
## * 2 proposed 4 as the best number of clusters
## * 3 proposed 7 as the best number of clusters
## * 4 proposed 8 as the best number of clusters
##
## Conclusion
## =====
## * According to the majority rule, the best number of clusters is 2 .

```

Optimal number of clusters – k = 2



```
#Agglomerative clustering
clustered <- agnes(dist(maximum_cluster, method = "maximum"),
                    diss=TRUE, method = "ward")
# add cluster labels to the training data - 5 say 2 and 4 say 8... needs more thinking about
maximum_cluster$cluster <- cutree(clustered, k = 8)
```

maximum_cluster

	eiqB11F.pre	eiqB11I.pre	eiqB11F.blm	eiqB11I.blm	eiqB11T.blm	eiqB11F.pos
## 1	143.09	0.000	58.10	19.04	0.0440	14.48
## 2	161.57	0.000	53.16	9.36	0.0000	31.37
## 3	161.57	0.000	53.16	9.36	0.0000	31.37
## 4	161.57	0.000	53.16	9.36	0.0000	31.37
## 5	41.85	0.000	44.67	14.25	0.0260	32.29
## 6	9.38	0.000	37.44	36.36	0.4600	24.39
## 7	163.95	2.310	112.90	12.62	0.0250	88.69
## 8	111.67	0.000	127.30	8.87	0.2200	53.11
## 9	94.76	0.960	52.27	19.69	0.0140	26.59
## 10	107.72	0.000	0.56	1.80	0.0000	41.85
## 11	196.87	0.330	63.57	4.94	0.0060	20.25
## 12	193.35	0.056	63.48	4.94	0.0031	19.87
## 13	91.93	0.000	53.30	8.13	0.0034	18.75
## 14	18.03	0.000	70.09	13.10	0.0000	20.53
## 15	21.18	6.740	17.79	5.49	0.0000	7.57
## 16	102.61	0.000	68.15	7.54	3.0000	29.44
## 17	106.47	0.000	33.00	28.70	0.0160	15.17

```
## 18      109.58      0.000      54.16      11.38      0.1700      16.79
## 19      113.28      0.000      102.91      3.61      0.0000      68.14
##      eiQB11I.pos eiQB11T.pos cluster
## 1         24.31      0.000      1
## 2         57.66      0.000      1
## 3         57.66      0.000      1
## 4         57.66      0.000      1
## 5          9.63      0.000      2
## 6         34.75      0.000      3
## 7         51.87      0.041      4
## 8          5.25      0.000      5
## 9         20.81      0.082      6
## 10        1.80      0.000      7
## 11        21.36      0.084      8
## 12        21.36      0.230      8
## 13         7.58      0.061      6
## 14         2.28      0.000      2
## 15         9.48      0.000      3
## 16        10.91      3.000      6
## 17        10.08      0.000      6
## 18        17.33      0.220      6
## 19        11.92      0.070      5
```

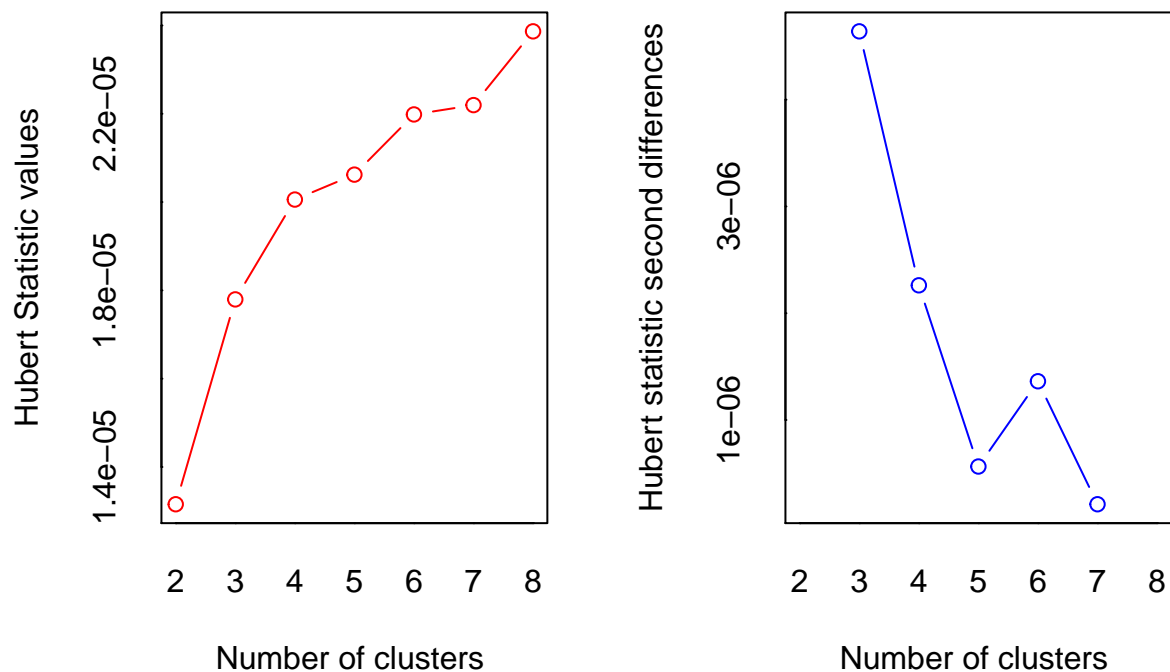
#Clustering summaries based on non-markov approach (Before Adjustment)

```
maximum_non_markov_clusters <- data_2012 %>%
  group_by(orchard) %>%
  summarise(mean_honey_ab = mean(apisAb),
            mean_wild_ab = mean(wildAbF),
            mean_social_rich = mean(socialRichF)) %>%
  mutate(cluster = maximum_cluster$cluster) %>%
  group_by(cluster) %>%
  summarise(mean_honey_ab = mean(mean_honey_ab),
            mean_wild_ab = mean(mean_wild_ab),
            mean_social_rich = mean(mean_social_rich),
            n = n())
```

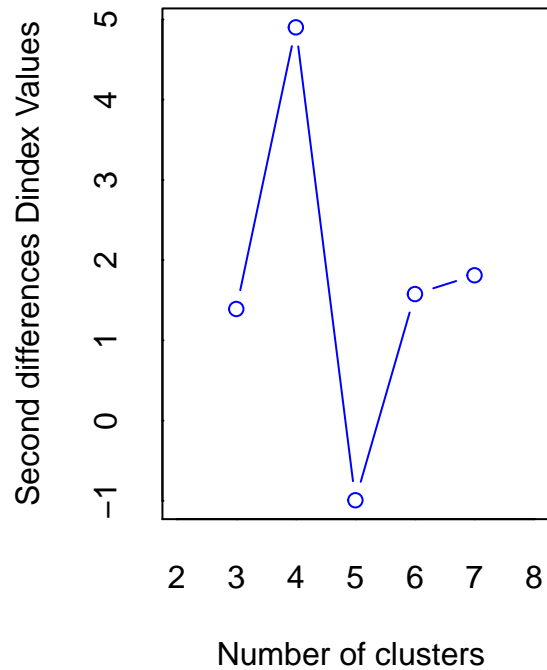
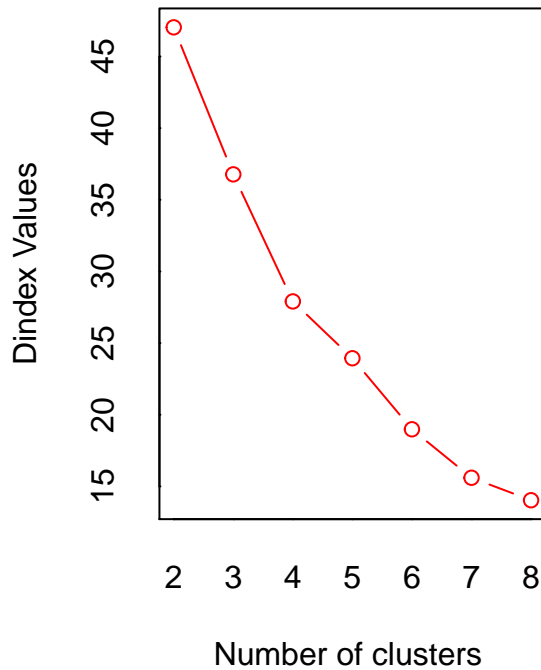
Clustering code - euclidean

*#Clusters of whole data combined - 7 - breaks sometimes with euclidean bit weird but maybe another just
#Euclidean clusters.*

```
fviz_nbclust(NbClust(euclidean_cluster, distance = "euclidean",
                    min.nc = 2, max.nc = 8, method = "ward.D2", index = "all"))
```

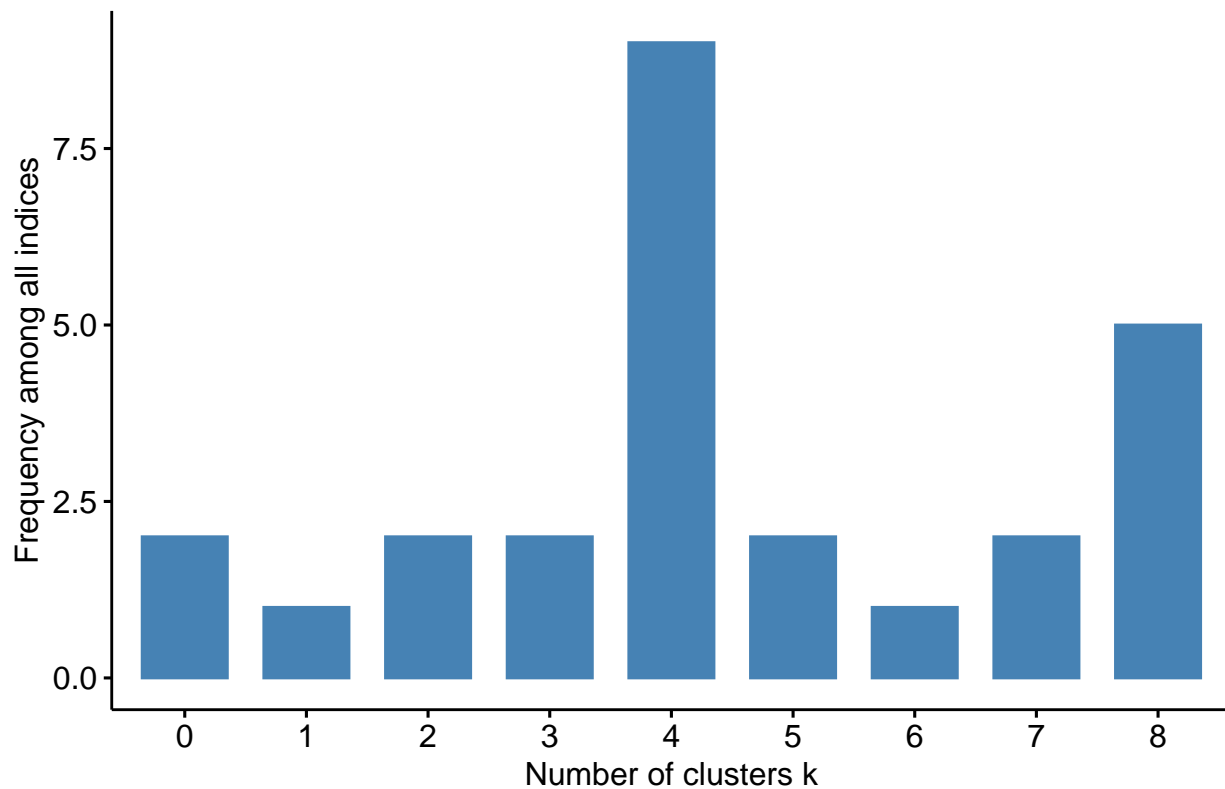
```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##       In the plot of Hubert index, we seek a significant knee that corresponds to a
##       significant increase of the value of the measure i.e the significant peak in Hubert
##       index second differences plot.
##
```



```
## *** : The D index is a graphical method of determining the number of clusters.
##           In the plot of D index, we seek a significant knee (the significant peak in Dindex
##           second differences plot) that corresponds to a significant increase of the value of
##           the measure.
##
## *****
## * Among all indices:
## * 2 proposed 2 as the best number of clusters
## * 2 proposed 3 as the best number of clusters
## * 9 proposed 4 as the best number of clusters
## * 2 proposed 5 as the best number of clusters
## * 1 proposed 6 as the best number of clusters
## * 2 proposed 7 as the best number of clusters
## * 5 proposed 8 as the best number of clusters
##
##           ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is 4
##
## *****
## Among all indices:
## =====
## * 2 proposed 0 as the best number of clusters
## * 1 proposed 1 as the best number of clusters
## * 2 proposed 2 as the best number of clusters
```

```
## * 2 proposed 3 as the best number of clusters
## * 9 proposed 4 as the best number of clusters
## * 2 proposed 5 as the best number of clusters
## * 1 proposed 6 as the best number of clusters
## * 2 proposed 7 as the best number of clusters
## * 5 proposed 8 as the best number of clusters
##
## Conclusion
## =====
## * According to the majority rule, the best number of clusters is 4 .
```

Optimal number of clusters – k = 4



```
#Agglomerative clustering
clustered <- agnes(dist(euclidean_cluster, method = "euclidean"),
                    diss=TRUE, method = "ward")
# add cluster labels to the training data
euclidean_cluster$cluster <- cutree(clustered, k = 4)

euclidean_cluster
```

```
##      eiqB11F.pre eiqB11I.pre eiqB11F.blm eiqB11I.blm eiqB11T.blm eiqB11F.pos
## 1      143.09      0.000      58.10      19.04      0.0440      14.48
## 2      161.57      0.000      53.16      9.36      0.0000      31.37
## 3      161.57      0.000      53.16      9.36      0.0000      31.37
## 4      161.57      0.000      53.16      9.36      0.0000      31.37
## 5       41.85      0.000      44.67      14.25      0.0260      32.29
## 6        9.38      0.000      37.44      36.36      0.4600      24.39
## 7      163.95      2.310     112.90      12.62      0.0250      88.69
```

```
## 8      111.67      0.000      127.30      8.87      0.2200      53.11
## 9       94.76      0.960      52.27      19.69      0.0140      26.59
## 10      107.72      0.000       0.56       1.80      0.0000      41.85
## 11      196.87      0.330      63.57       4.94      0.0060      20.25
## 12      193.35      0.056      63.48       4.94      0.0031      19.87
## 13       91.93      0.000      53.30       8.13      0.0034      18.75
## 14       18.03      0.000      70.09      13.10      0.0000      20.53
## 15       21.18      6.740      17.79       5.49      0.0000       7.57
## 16      102.61      0.000      68.15       7.54      3.0000      29.44
## 17      106.47      0.000      33.00      28.70      0.0160      15.17
## 18      109.58      0.000      54.16      11.38      0.1700      16.79
## 19      113.28      0.000     102.91       3.61      0.0000      68.14
##      eiqB11I.pos eiqB11T.pos cluster
## 1         24.31      0.000         1
## 2         57.66      0.000         1
## 3         57.66      0.000         1
## 4         57.66      0.000         1
## 5          9.63      0.000         2
## 6         34.75      0.000         2
## 7         51.87      0.041         3
## 8          5.25      0.000         3
## 9         20.81      0.082         4
## 10         1.80      0.000         4
## 11         21.36      0.084         1
## 12         21.36      0.230         1
## 13          7.58      0.061         4
## 14          2.28      0.000         2
## 15          9.48      0.000         2
## 16         10.91      3.000         4
## 17         10.08      0.000         4
## 18         17.33      0.220         4
## 19         11.92      0.070         3
```

#Clustering summaries based on non-markov approach (Before Adjustment)

#Cluster numbers are 6 ,4 ,3 ,6

```
euclidean_non_markov_clusters <- data_2012 %>%
  group_by(orchard) %>%
  summarise(mean_honey_ab = mean(apisAb),
            mean_wild_ab = mean(wildAbF),
            mean_social_rich = mean(socialRichF)) %>%
  mutate(cluster = euclidean_cluster$cluster) %>%
  group_by(cluster) %>%
  summarise(mean_honey_ab = mean(mean_honey_ab),
            mean_wild_ab = mean(mean_wild_ab),
            mean_social_rich = mean(mean_social_rich),
            n= n())
euclidean_non_markov_clusters
```

A tibble: 4 x 5

```
##   cluster mean_honey_ab mean_wild_ab mean_social_rich      n
##   <int>      <dbl>      <dbl>      <dbl> <int>
## 1     1         8.31         4.62         0.542     6
## 2     2         7.94         5.5          0.688     4
## 3     3         3.75         2.88         0.625     3
## 4     4         5.29         2.60         0.569     6
```

Variable justification

```
bee_correlations <- data.frame(matrix(0, nrow = 0, ncol = 0))

for(i in 1:7){
  for(j in 1:7){
    bee_correlations[i,j] <- round(cor(data_2012[5+i], data_2012[5 + j]), 2)
  }
}
for(i in 1:7){
  names(bee_correlations)[i] <- colnames(data_2012[5 + i])
}
bee_correlations
```

	apisAb.1	apisAb.2	apisAb.3	apisAb.4	apisAb.5	apisAb.6	apisAb.7	wildAbF.1
## 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.29
## 2	0.29	0.29	0.29	0.29	0.29	0.29	0.29	1.00
## 3	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.91
## 4	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.96
## 5	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.92
## 6	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.29
## 7	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.31

	wildAbF.2	wildAbF.3	wildAbF.4	wildAbF.5	wildAbF.6	wildAbF.7	wildRichF.1
## 1	0.29	0.29	0.29	0.29	0.29	0.29	0.19
## 2	1.00	1.00	1.00	1.00	1.00	1.00	0.91
## 3	0.91	0.91	0.91	0.91	0.91	0.91	1.00
## 4	0.96	0.96	0.96	0.96	0.96	0.96	0.81
## 5	0.92	0.92	0.92	0.92	0.92	0.92	0.92
## 6	0.29	0.29	0.29	0.29	0.29	0.29	0.50
## 7	0.31	0.31	0.31	0.31	0.31	0.31	0.53

	wildRichF.2	wildRichF.3	wildRichF.4	wildRichF.5	wildRichF.6	wildRichF.7
## 1	0.19	0.19	0.19	0.19	0.19	0.19
## 2	0.91	0.91	0.91	0.91	0.91	0.91
## 3	1.00	1.00	1.00	1.00	1.00	1.00
## 4	0.81	0.81	0.81	0.81	0.81	0.81
## 5	0.92	0.92	0.92	0.92	0.92	0.92
## 6	0.50	0.50	0.50	0.50	0.50	0.50
## 7	0.53	0.53	0.53	0.53	0.53	0.53

	solitaryAbF.1	solitaryAbF.2	solitaryAbF.3	solitaryAbF.4	solitaryAbF.5
## 1	0.25	0.25	0.25	0.25	0.25
## 2	0.96	0.96	0.96	0.96	0.96
## 3	0.81	0.81	0.81	0.81	0.81
## 4	1.00	1.00	1.00	1.00	1.00
## 5	0.92	0.92	0.92	0.92	0.92
## 6	0.02	0.02	0.02	0.02	0.02
## 7	0.05	0.05	0.05	0.05	0.05

	solitaryAbF.6	solitaryAbF.7	solitaryRichF.1	solitaryRichF.2
## 1	0.25	0.25	0.13	0.13
## 2	0.96	0.96	0.92	0.92
## 3	0.81	0.81	0.92	0.92
## 4	1.00	1.00	0.92	0.92
## 5	0.92	0.92	1.00	1.00
## 6	0.02	0.02	0.13	0.13
## 7	0.05	0.05	0.16	0.16

```
## solitaryRichF.3 solitaryRichF.4 solitaryRichF.5 solitaryRichF.6
## 1 0.13 0.13 0.13 0.13
## 2 0.92 0.92 0.92 0.92
## 3 0.92 0.92 0.92 0.92
## 4 0.92 0.92 0.92 0.92
## 5 1.00 1.00 1.00 1.00
## 6 0.13 0.13 0.13 0.13
## 7 0.16 0.16 0.16 0.16
## solitaryRichF.7 socialAbF.1 socialAbF.2 socialAbF.3 socialAbF.4
## 1 0.13 0.18 0.18 0.18 0.18
## 2 0.92 0.29 0.29 0.29 0.29
## 3 0.92 0.50 0.50 0.50 0.50
## 4 0.92 0.02 0.02 0.02 0.02
## 5 1.00 0.13 0.13 0.13 0.13
## 6 0.13 1.00 1.00 1.00 1.00
## 7 0.16 0.99 0.99 0.99 0.99
## socialAbF.5 socialAbF.6 socialAbF.7 socialRichF.1 socialRichF.2
## 1 0.18 0.18 0.18 0.22 0.22
## 2 0.29 0.29 0.29 0.31 0.31
## 3 0.50 0.50 0.50 0.53 0.53
## 4 0.02 0.02 0.02 0.05 0.05
## 5 0.13 0.13 0.13 0.16 0.16
## 6 1.00 1.00 1.00 0.99 0.99
## 7 0.99 0.99 0.99 1.00 1.00
## socialRichF.3 socialRichF.4 socialRichF.5 socialRichF.6 socialRichF.7
## 1 0.22 0.22 0.22 0.22 0.22
## 2 0.31 0.31 0.31 0.31 0.31
## 3 0.53 0.53 0.53 0.53 0.53
## 4 0.05 0.05 0.05 0.05 0.05
## 5 0.16 0.16 0.16 0.16 0.16
## 6 0.99 0.99 0.99 0.99 0.99
## 7 1.00 1.00 1.00 1.00 1.00
```

This demonstrates that to consider the affects on wild bees, just two variables need to be considered as there is very high correlation between the outcomes. WildAbF accounts a .9+ correlation with all wild variables except for the for the solitary bee variable. Then for this SocialRichF could be used.

Updating the bee values to match the environments

In this particular analysis, the interest is in the affects of pesticides on bee population. As such, it is necessary to correct, as best we can, for extra factors (confounding?) to make the bee values representative.

All these graphs were plotted as part of EDA by the use of lapply to the a basic plotting function. Here the relationships between some of the main extra factors (are they confounding?) will be examined.

Temperature

It is known that temperature affects the speed at which bees fly, as such this will have an effect on bee abundance and possibly richness although more bees might not necessarily mean more species are observed. From the previous analysis carried out it is known that a $\log(x) + 1$ transposition of bee count allows for a “better” model fit and as such in general this will the case for our observations.

```
##agglom
wild_bee_abundance_agglom <- data_2012 %>%
```

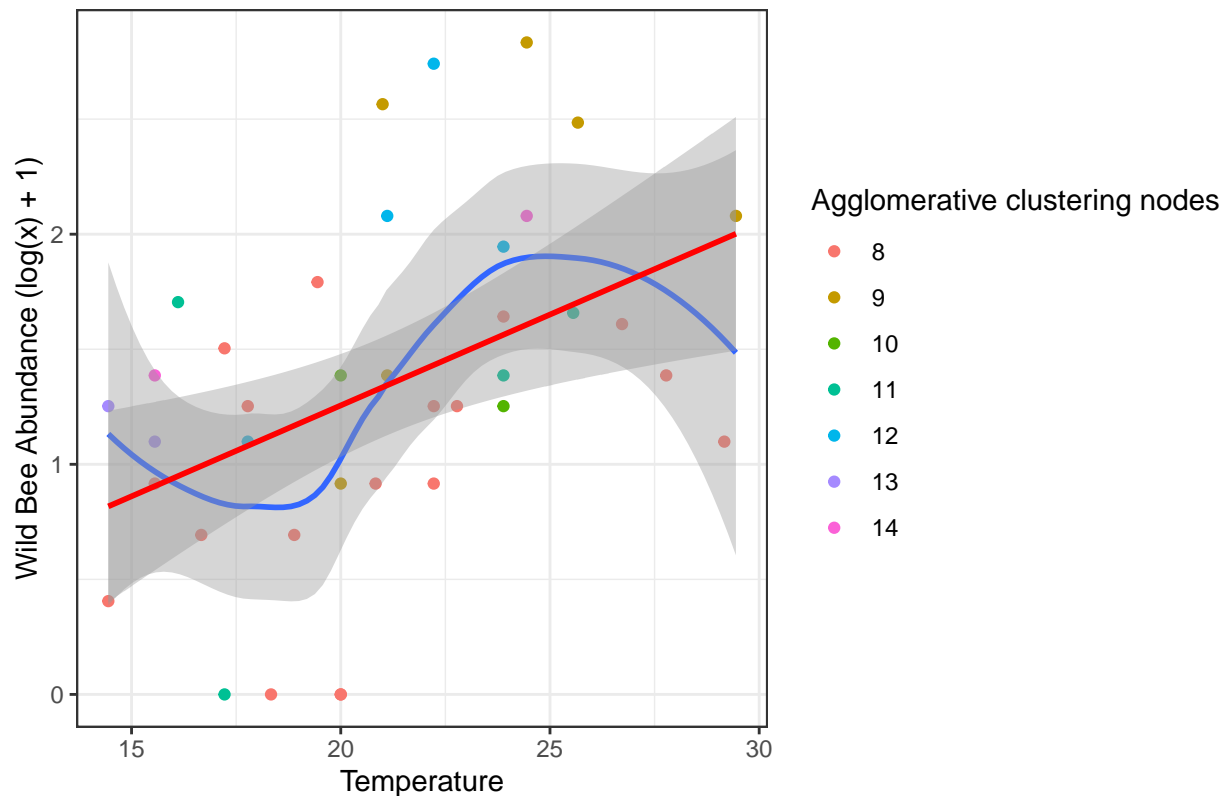
```
inner_join(orchard_node_agglom) %>%
ggplot(aes(x = temp, y = log(wildAbF + 1))) +
geom_point(aes(colour = as_factor(node))) +
geom_smooth() +
geom_smooth(method = "lm", colour = "red") +
labs(x = "Temperature", y = "Wild Bee Abundance (log(x) + 1)",
      title = "Wild Bee Abundance Coloured by Agglomerature Nodes",
      colour = "Agglomerative clustering nodes") +
theme_bw()
```

```
## Joining, by = "orchard"
```

```
wild_bee_abundance_agglom
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Wild Bee Abundance Coloured by Agglomerature Nodes



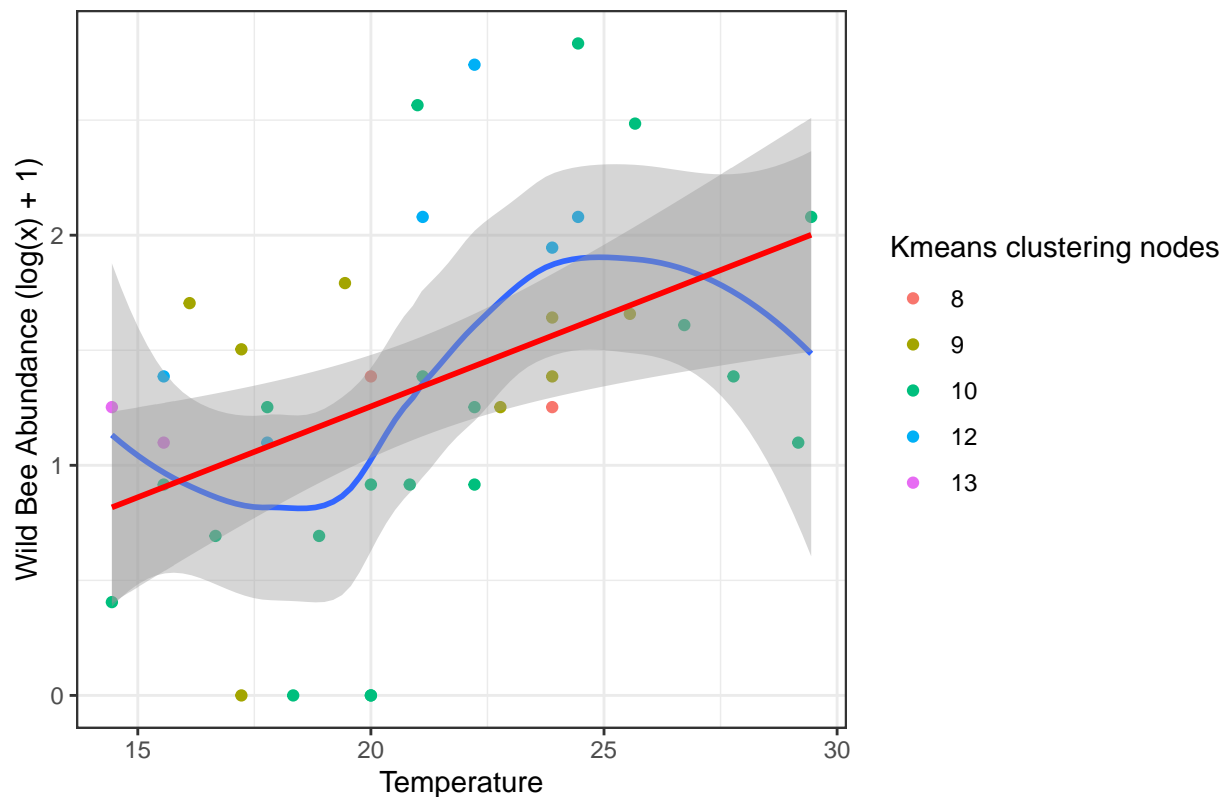
```
##kmeans
wild_bee_abundance_kmeans <- data_2012 %>%
inner_join(orchard_node_kmeans) %>%
ggplot(aes(x = temp, y = log(wildAbF + 1))) +
geom_point(aes(colour = as_factor(node))) +
geom_smooth() +
geom_smooth(method = "lm", colour = "red") +
labs(x = "Temperature", y = "Wild Bee Abundance (log(x) + 1)",
      title = "Wild Bee Abundance Coloured by Kmeans Nodes",
      colour = "Kmeans clustering nodes") +
theme_bw()
```

```
## Joining, by = "orchard"
```

```
wild_bee_abundance_kmeans
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Wild Bee Abundance Coloured by Kmeans Nodes



```
##Linear model
```

```
lm_temp <- lm(log(data_2012$wildAbF + 1) ~ data_2012$temp)
```

```
temp_factor <- tidy(lm_temp) %>%
```

```
  slice(2) %>%
```

```
  select(estimate) %>%
```

```
  as_vector()
```

```
#Adjusting bee value as though temp is always 20
```

```
adjusted_data <- data_2012 %>%
```

```
  inner_join(clusterings) %>%
```

```
  mutate(adjusted_beas = (((20 - temp) * temp_factor) + log(wildAbF + 1)))
```

```
## Joining, by = "orchard"
```

```
wild_bee_abundance_agglom_adjusted <- adjusted_data %>%
```

```
  ggplot(aes(x = temp, y = adjusted_beas)) +
```

```
  geom_point(aes(colour = as_factor(node))) +
```

```
  geom_smooth() +
```

```
  geom_smooth(method = "lm", colour = "red") +
```

```
  labs(x = "Temperature", y = "Adjusted Wild Bee Abundance (log(x) + 1)",
```

```
        title = "Wild Bee Abundance Coloured by Agglomerate Nodes",
```

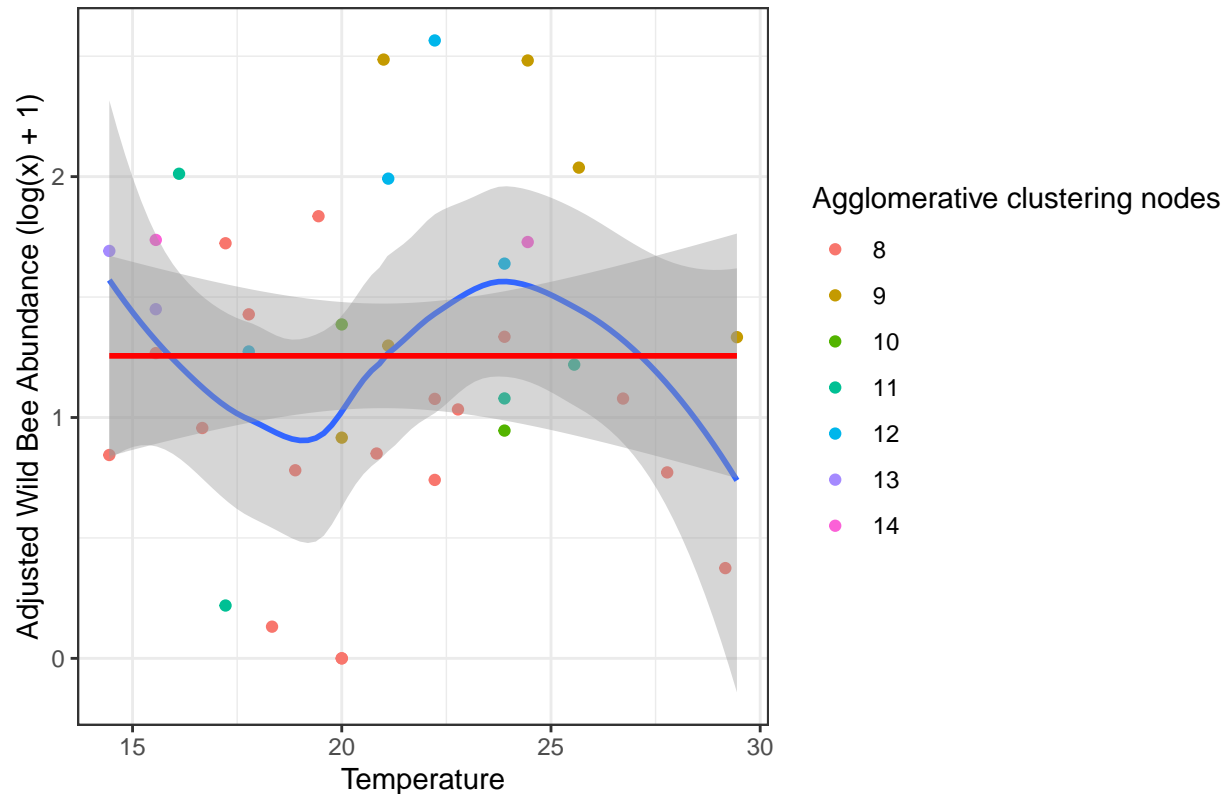
```
        colour = "Agglomerative clustering nodes") +
```



```
theme_bw()
wild_bee_abundance_agglom_adjusted
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Wild Bee Abundance Coloured by Agglomerature Nodes



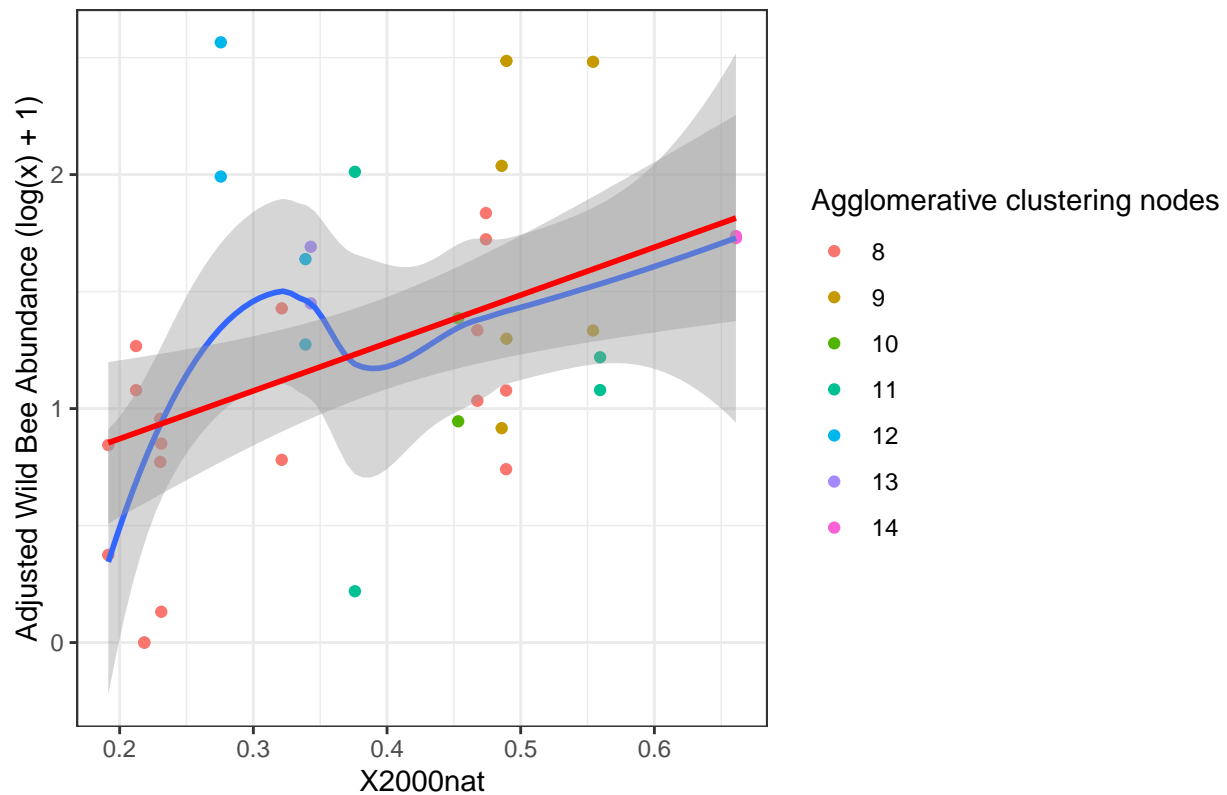
X2000nat

Based on previous research undertaken in the white paper it is known that the X2000nat variable has an effect on the the bee count so it also necessary to adjust for this. Wher Again the value will be adjust as though the X2000nat variable is constant. In this case the mean will be chosen as the constant value and the bee count will be adjusted for this in the same way as above.

```
wild_bee_abundance_agglom_x2000 <- adjusted_data %>%
  ggplot(aes(x = X2000nat, y = adjusted_beas)) +
  geom_point(aes(colour = as_factor(node))) +
  geom_smooth() +
  geom_smooth(method = "lm", colour = "red") +
  labs(x = "X2000nat", y = "Adjusted Wild Bee Abundance (log(x) + 1)",
       title = "Wild Bee Abundance Coloured by Agglomerature Nodes",
       colour = "Agglomerative clustering nodes") +
  theme_bw()
wild_bee_abundance_agglom_x2000
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Wild Bee Abundance Coloured by Agglomerate Nodes



```
#Outcome is 0.388
mean_x2000 <- adjusted_data %>%
  summarise(mean_x2000 = mean(X2000nat)) %>%
  as_vector()

lm_X2000nat <- lm(adjusted_data$adjusted_beas ~ adjusted_data$X2000nat)
X2000_factor <- tidy(lm_X2000nat) %>%
  slice(2) %>%
  select(estimate) %>%
  as_vector()

#Adjusting the data for this variable

adjusted_data <- adjusted_data %>%
  mutate(adjusted_beas = ((mean_x2000 - X2000nat) * X2000_factor) + adjusted_beas)
```

So if you look at them individually, it can only be adjusted for one. So need to combine a linear model of all affecting factors to best adjust at the end.

Analysing other potential confounders

This section will look at finding other parameters to include in the bee variable, dataset used at this point is the original just with the bee variable considered going through a $\log(x + 1)$ transformation. The only bee variable to be considered at this point is the wild bee abundance, in future more could be considered in this

way or it would be expected to be similar for all bee variables and I could just use the variables decided upon in this way for all the bee variables considered.

Local Diversity

The aim here is to whether adding local diversity as a adjustment factor is going to be beneficial or not. This will be achieved by running a F test to see whether the variances are equal, a shapiro test to confirm the bee data can be assumed to be normal. Note, independence assumption is not validated, this is due more than 1 result being used from each orchard, I decided to include these as I thought the extra power gained from twice the observations was worth the penalty of this assumption.

```
#When adding temperature in other variables have an effect
logged_data <- data_2012 %>%
  inner_join(clusterings) %>%
  mutate(adjusted_bees = (((20 - temp) * temp_factor) + log(wildAbF + 1)))

## Joining, by = "orchard"

#Just logged
logged_data <- data_2012 %>%
  inner_join(clusterings) %>%
  mutate(adjusted_bees = log(wildAbF + 1))

## Joining, by = "orchard"

#X2000nat adjustment
logged_data <- data_2012 %>%
  inner_join(clusterings) %>%
  mutate(adjusted_bees = log(wildAbF + 1))%>%
  mutate(adjusted_bees = ((mean_x2000 - X2000nat) * X2000_factor) + adjusted_bees)

## Joining, by = "orchard"

#Testing underlying normal data assumption, again demonstrating why the transformation is neccessary:
#Not normal
shapiro.test(logged_data$wildAbF)

##
## Shapiro-Wilk normality test
##
## data: logged_data$wildAbF
## W = 0.79764, p-value = 9.361e-06

#After transformation -> normal
shapiro.test(logged_data$adjusted_bees)

##
## Shapiro-Wilk normality test
##
## data: logged_data$adjusted_bees
## W = 0.97681, p-value = 0.6046

#Checking if the two subsets have the same variance have the same variance
#Outcome shows that they can be assumed to have the same variance as p.value > 0,05
simple <- logged_data %>% filter(local.diversity == 0) %>% select(adjusted_bees) %>% as_vector()
diverse <- logged_data %>% filter(local.diversity == 1) %>% select(adjusted_bees) %>% as_vector()
tidy(var.test(simple, diverse)) %>%
  select(statistic, p.value)
```

```
## Multiple parameters; naming those columns num.df, denom.df

## # A tibble: 1 x 2
##   statistic p.value
##   <dbl>     <dbl>
## 1      1.17    0.725

#Performing a two sample t-test on the data to see whether they are the same.
#Since H0: Means the same, H1: means are different ~ p.value not significant

#This implies local.diversity does NOT have an effect on the original bee counts.
#I thought this was an interesting point and worth noting, same happens with Region.
tidy(t.test(wildAbF ~ local.diversity, logged_data, var.equal = TRUE))
```

```
## # A tibble: 1 x 9
##   estimate1 estimate2 statistic p.value parameter conf.low conf.high method
##   <dbl>     <dbl>     <dbl>  <dbl>     <dbl>     <dbl>     <dbl> <chr>
## 1      3.5      4.18     -0.532  0.598        36     -3.27      1.91 " Two~
## # ... with 1 more variable: alternative <chr>
```

```
#However the the adjusted version of bee count does list local diversity
#as a factor which has an effect on the log transformed bee counts
tidy(t.test(adjusted_beas ~ local.diversity, logged_data, var.equal = TRUE))
```

```
## # A tibble: 1 x 9
##   estimate1 estimate2 statistic p.value parameter conf.low conf.high method
##   <dbl>     <dbl>     <dbl>  <dbl>     <dbl>     <dbl>     <dbl> <chr>
## 1      1.22      1.41     -0.896  0.376        36     -0.617     0.239 " Two~
## # ... with 1 more variable: alternative <chr>
```

From This analysis, it can be concluded that local.diversity does NOT play a factor in bee count ### Region Region has 3 possible options - as such an ANOVA test will be used to see whether there is a difference between the means in wild abundance based on region.

```
#Checking Anova assumption that the variances are the same
# H0: Variances the same: H1: Atleast one variance is different
leveneTest(adjusted_beas ~ region, data = logged_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 2  0.1954 0.8234
##      35
```

```
#Since pr > 0.05 we can assume variances are the same
```

```
#Running the anova test
anova1 <- aov(adjusted_beas ~ region, data = logged_data)

summary(anova1)
```

```
##      Df Sum Sq Mean Sq F value Pr(>F)
## region 2  0.522  0.2611  0.624  0.542
## Residuals 35 14.652  0.4186
```

```
#probability shows that region is not an affecting factor to wildAbF
```

```
#Checking that the following anova assumption is TRUE:
#Residuals of the response variable are normally distributed is NOT true
shapiro.test(residuals(anova1))
```

```
##
## Shapiro-Wilk normality test
##
## data: residuals(anova1)
## W = 0.96856, p-value = 0.3546
```

Day

Looking at the day to see whether that also has an effect, since the data can be paired by day 1 and day 2 a paired t-test will be carried out on this to test whether there is a difference in bee count. When using just the X2000nat variable, then it becomes very close to significant.

```
day_1 <- logged_data %>% filter(day == 1) %>% select(adjusted_bees) %>% as_vector()
day_2 <- logged_data %>% filter(day == 2) %>% select(adjusted_bees) %>% as_vector()
#This demonstrates that the variances come from the same distribution:
tidy(var.test(day_1, day_2)) %>%
  select(statistic, p.value)
```

```
## Multiple parameters; naming those columns num.df, denom.df
## # A tibble: 1 x 2
##   statistic p.value
##   <dbl>    <dbl>
## 1      0.936    0.891
```

```
#Hence we can run a paired t-test with equal variances
tidy(t.test(adjusted_bees ~ day, logged_data, var.equal = TRUE, paired = TRUE))
```

```
## # A tibble: 1 x 8
##   estimate statistic p.value parameter conf.low conf.high method
##   <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl> <chr>
## 1   -0.342    -2.02  0.0588        18   -0.697    0.0142 Paire~
## # ... with 1 more variable: alternative <chr>
```

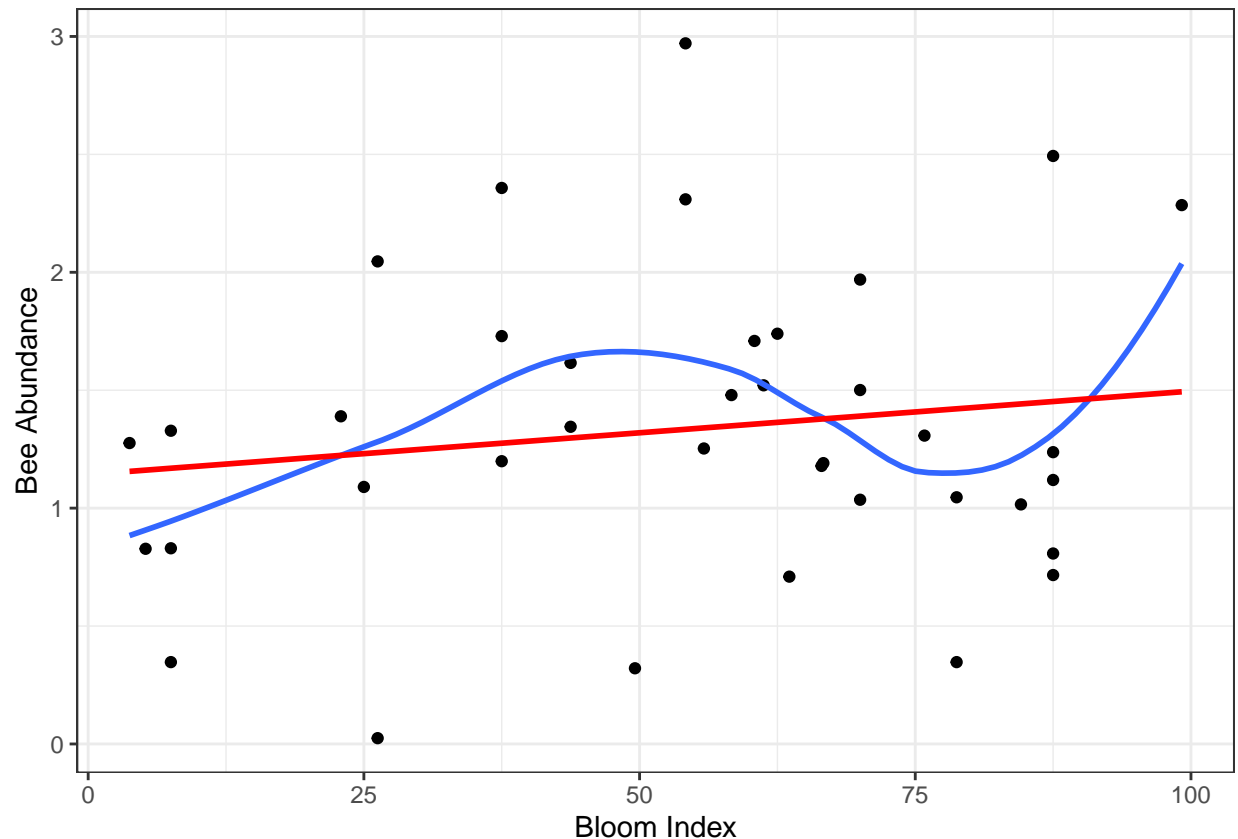
```
#Bit surprising but it demonstrates that there is no statistical significance between days
#So combining values to start with is probably appropriate
```

Bloom

First it shall be graphed to see if there is any obvious correlation.

```
logged_data %>%
  ggplot(aes(x = bloom.index, y = adjusted_bees)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  geom_smooth(method = "lm", colour = "red", se = FALSE) +
  theme_bw() +
  geom_point() %>%
  labs(x = "Bloom Index", y = "Bee Abundance")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



There doesn't seem to be any obvious correlation, but just to be sure I'll split the bloom into three categories and test these with an Anova test.

```
logged_data <- logged_data %>%
  mutate(bloom_category = if_else(bloom.index <= 33, 1,
    if_else(bloom.index <= 66, 2, 3)))

#Running the anova test
anova1 <- aov(adjusted_bees ~ bloom_category, data = logged_data)
summary(anova1)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## bloom_category 1    0.19   0.1897    0.456  0.504
## Residuals    36   14.98   0.4162
```

```
#As expected there is no evidence to suggest that bloom index (category)
#Has an effect on the bee counts (although only 1 DF in bloom_category thought it should be 2??)
```

Bloom is a continuous variable but for the sakes of deciding whether to include it or not

Final Confounders

Based on the analysis above, the confounders that could be taken into account are: Region and local diversity based on the statistical tests, temperature and X2000nat based on the literature.

```
#Checking the variables to see if any are heavily correlated (binary/categorical ones won't be)
#So checking other two and we get a low correlation so acceptable to use them together
cor(logged_data$temp, logged_data$X2000nat)
```

```
## [1] 0.2065843
f_lm <- lm(adjusted_beets ~ temp + X2000nat + local.diversity + region, data = logged_data)
#Interesting values here seem to suggest that only significant one is X2000nat
summary(f_lm)

##
## Call:
## lm(formula = adjusted_beets ~ temp + X2000nat + local.diversity +
##     region, data = logged_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23257 -0.35949 -0.06686  0.23707  1.30560
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.33682    0.61856  -0.545  0.58985
## temp           0.08343    0.02620   3.184  0.00323 **
## X2000nat       0.09182    0.79392   0.116  0.90865
## local.diversity 0.20663    0.28923   0.714  0.48014
## regionLO      -0.42271    0.36305  -1.164  0.25290
## regionS       -0.08406    0.46728  -0.180  0.85837
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5782 on 32 degrees of freedom
## Multiple R-squared:  0.2949, Adjusted R-squared:  0.1848
## F-statistic: 2.677 on 5 and 32 DF,  p-value: 0.03941
```

Notes, from running the adjustments: When you use temperature as an original adjustment, only significant one in the model is X2000nat and when you use X2000nat in the original data only significant one is temp. When you use neither and just apply the linear model on the logged data then both temp and nat are significant -> suggesting that just these two variables would be best to adjust by. This suggest to me that fitting a joint lm using both temp and X2000nat to adjust the data is probably best to avoid overfitting and keep the model as parsimonious as possible. Despite region and local.diversity being significant for temp adjustment originally, unlikely to be worth implementing but could be considered to cover model changes section. To take this further, could apply lasso regression on the full dataset to see what that comes up with as tends to have a lower MSE than a standard LM.

```
#Using the logged_data with no adjustments
logged_data <- data_2012 %>%
  inner_join(clusterings) %>%
  mutate(adjusted_beets = log(wildAbF + 1)) %>%
  mutate(adjusted_social = log(socialRichF + 1))
```

```
## Joining, by = "orchard"
```

```
#Defining linear model with temp and X2000nat as factors
temp_nat_lm <- tidy(lm(adjusted_beets ~ temp + X2000nat, data = logged_data))
#Extracting temp and nat estimate values
temp_factor <- temp_nat_lm %>%
  select(estimate) %>%
  slice(2) %>%
  as_vector()
nat_factor <- temp_nat_lm %>%
```

```

select(estimate) %>%
slice(3) %>%
as_vector()

#Now adjustments shall be made for these two variables:
#Temperature will be set to 20 degrees
#X2000 nat will be set to the mean x2000 nat value
temp_nat_adjusted <- logged_data %>%
  mutate(adjusted_bees = (adjusted_bees + ((20 - temp) * temp_factor) + ((mean_x2000 - X2000nat) * X2000nat)))

#Checking that after the adjustment they are not significant at all in the model
tidy(lm(adjusted_bees ~ temp + X2000nat, data = temp_nat_adjusted))

```

```

## # A tibble: 3 x 5
##   term          estimate std.error statistic p.value
##   <chr>          <dbl>    <dbl>    <dbl>   <dbl>
## 1 (Intercept)  1.24e+ 0    0.549    2.25e+ 0  0.0309
## 2 temp        -9.42e-18    0.0249  -3.79e-16  1.000
## 3 X2000nat      9.12e- 2    0.735    1.24e- 1  0.902

```

#As expected now the bee abundance has been adjusted for these two parameters

The same process can be applied to the variable SocialRichF, the other bee variable which needs to be considered. Since we know that only temp and nat were statistic on the wildAbF we can go straight to the LM and test that (May need to check this with Julia, if not just cp previous analysis...)

```

#Looking at the model ~ only X2000nat significant
lm(adjusted_social ~ temp + X2000nat + local.diversity + region, data = logged_data) %>%
summary

```

```

##
## Call:
## lm(formula = adjusted_social ~ temp + X2000nat + local.diversity +
##   region, data = logged_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51626 -0.27605 -0.06653  0.19262  0.72315
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.287012   0.381099  -0.753   0.45689
## temp          0.007198   0.016144   0.446   0.65869
## X2000nat      1.507687   0.489142   3.082   0.00421 **
## local.diversity -0.164463   0.178196  -0.923   0.36295
## regionL0      -0.094068   0.223680  -0.421   0.67690
## regionS       0.165221   0.287892   0.574   0.57005
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3562 on 32 degrees of freedom
## Multiple R-squared:  0.3068, Adjusted R-squared:  0.1985
## F-statistic: 2.833 on 5 and 32 DF,  p-value: 0.03153
social_lm<- tidy(lm(adjusted_social ~ X2000nat, data = logged_data))

```



```

social_nat_factor <- social_lm %>%
  select(estimate) %>%
  slice(2) %>%
  as_vector()

#Adjustment for social added
temp_nat_adjusted <- temp_nat_adjusted %>%
  mutate(adjusted_social = (adjusted_social + ((mean_x2000 - X2000nat) * X2000_factor)))

```

Clustering Summarising

Functions involved with summarising the clusters:

```

#Gets the name of a variable as a string ~ now implemented in clusterise but keeping
get_name <- function(x) {
  deparse(substitute(x))
}

#Gets the bee variables that I am summarising for each cluster - makes it easy to change in future
clusterise <- function(bee_data, cluster_data, v_name = deparse(substitute(cluster_data))){
  inner_join(bee_data, cluster_data) %>%
    summarise(mean_honey_ab = mean(mean_honey_ab),
              mean_wild_ab = mean(mean_wild_ab),
              mean_social_rich = mean(mean_social_rich),
              n = n()) %>%
    mutate(cluster = v_name)
}

#The function to output the summaries of both agglom and kmeans clustering
cluster_summarise <- function(bee_data){
  output <- list()

  output$agglom_beas <- clusterise(bee_data, aggl_node1) %>%
    bind_rows(clusterise(bee_data, aggl_node2)) %>%
    bind_rows(clusterise(bee_data, aggl_node3)) %>%
    bind_rows(clusterise(bee_data, aggl_node4)) %>%
    bind_rows(clusterise(bee_data, aggl_node5)) %>%
    bind_rows(clusterise(bee_data, aggl_node6)) %>%
    bind_rows(clusterise(bee_data, aggl_node7)) %>%
    bind_rows(clusterise(bee_data, aggl_node4 %>% filter(cluster == 1), "agglom_node8")) %>%
    bind_rows(clusterise(bee_data, aggl_node4 %>% filter(cluster == 2), "agglom_node9")) %>%
    bind_rows(clusterise(bee_data, aggl_node5 %>% filter(cluster == 1), "agglom_node10")) %>%
    bind_rows(clusterise(bee_data, aggl_node5 %>% filter(cluster == 2), "agglom_node11")) %>%
    bind_rows(clusterise(bee_data, aggl_node6 %>% filter(cluster == 1), "agglom_node12")) %>%
    bind_rows(clusterise(bee_data, aggl_node6 %>% filter(cluster == 2), "agglom_node13")) %>%
    #Only 1 result for cluster as none in the latter cluster
    #Unable to to bind as no common variables for an empty node15
    bind_rows(clusterise(bee_data, aggl_node7 %>% filter(cluster == 1), "agglom_node14"))

  output$kmeans_beas <- clusterise(bee_data, kmeans_node1) %>%
    bind_rows(clusterise(bee_data, kmeans_node2)) %>%
    bind_rows(clusterise(bee_data, kmeans_node3)) %>%

```

```

bind_rows(clusterise(bee_data, kmeans_node4)) %>%
bind_rows(clusterise(bee_data, kmeans_node5)) %>%
bind_rows(clusterise(bee_data, kmeans_node6)) %>%
bind_rows(clusterise(bee_data, kmeans_node7)) %>%
bind_rows(clusterise(bee_data, kmeans_node4) %>% filter(cluster == 1), "kmeans_node8")) %>%
bind_rows(clusterise(bee_data, kmeans_node4) %>% filter(cluster == 2), "kmeans_node9")) %>%
bind_rows(clusterise(bee_data, kmeans_node5) %>% filter(cluster == 1), "kmeans_node10")) %>%
bind_rows(clusterise(bee_data, kmeans_node5) %>% filter(cluster == 2), "kmeans_node11")) %>%
bind_rows(clusterise(bee_data, kmeans_node6) %>% filter(cluster == 1), "kmeans_node12")) %>%
bind_rows(clusterise(bee_data, kmeans_node6) %>% filter(cluster == 2), "kmeans_node13")) %>%
bind_rows(clusterise(bee_data, kmeans_node7) %>% filter(cluster == 1), "kmeans_node14")) %>%
bind_rows(clusterise(bee_data, kmeans_node7) %>% filter(cluster == 2), "kmeans_node15"))

#Return a list containing two tibbles one for each type of clustering algorithm
output
}

```

Original summarising without any variable correction

```

#original Bee data (logged has no changes to it)
bee_values <- logged_data %>%
  group_by(orchard) %>%
  summarise(mean_honey_ab = mean(apisAb),
            mean_wild_ab = mean(adjusted_bees),
            mean_social_rich = mean(adjusted_social)
            )

original_bee_summary <- cluster_summarise(bee_data = bee_values)

#Agglom
original_bee_summary$agglom_beas

## # A tibble: 14 x 5
##   mean_honey_ab mean_wild_ab mean_social_rich      n cluster
##   <dbl>         <dbl>         <dbl> <int> <chr>
## 1         6.56         1.33         0.379    19 aggl_node1
## 2         6.19         1.23         0.386    15 aggl_node2
## 3         7.94         1.71         0.353     4 aggl_node3
## 4         6.80         1.23         0.376    12 aggl_node4
## 5         3.75         1.23         0.423     3 aggl_node5
## 6         8.75         1.70         0.135     3 aggl_node6
## 7         5.5          1.73         1.01     1 aggl_node7
## 8         5.46         0.963        0.325     9 agglom_node8
## 9        10.8          2.04         0.530     3 agglom_node9
## 10        3.25         1.32         0.405     1 agglom_node10
## 11         4           1.19         0.432     2 agglom_node11
## 12         8           1.97         0.101     2 agglom_node12
## 13        10.2         1.18         0.203     1 agglom_node13
## 14         5.5         1.73         1.01     1 agglom_node14

#Kmeans
original_bee_summary$kmeans_beas

## # A tibble: 15 x 5
##   mean_honey_ab mean_wild_ab mean_social_rich      n cluster

```

##	<dbl>	<dbl>	<dbl>	<int>	<chr>
## 1	6.56	1.33	0.379	19	kmeans_node1
## 2	6.19	1.23	0.386	15	kmeans_node2
## 3	7.94	1.71	0.353	4	kmeans_node3
## 4	6.5	1.36	0.516	5	kmeans_node4
## 5	6.04	1.17	0.320	10	kmeans_node5
## 6	7.94	1.71	0.353	4	kmeans_node6
## 7	NaN	NaN	NaN	0	kmeans_node7
## 8	3.25	1.32	0.405	1	kmeans_node8
## 9	7.31	1.37	0.544	4	kmeans_node9
## 10	6.04	1.17	0.320	10	kmeans_node10
## 11	NaN	NaN	NaN	0	kmeans_node11
## 12	7.17	1.89	0.403	3	kmeans_node12
## 13	10.2	1.18	0.203	1	kmeans_node13
## 14	NaN	NaN	NaN	0	kmeans_node14
## 15	NaN	NaN	NaN	0	kmeans_node15

##Original summarising of non-markov approach
maximum_cluster

##	eiqB11F.pre	eiqB11I.pre	eiqB11F.blm	eiqB11I.blm	eiqB11T.blm	eiqB11F.pos
## 1	143.09	0.000	58.10	19.04	0.0440	14.48
## 2	161.57	0.000	53.16	9.36	0.0000	31.37
## 3	161.57	0.000	53.16	9.36	0.0000	31.37
## 4	161.57	0.000	53.16	9.36	0.0000	31.37
## 5	41.85	0.000	44.67	14.25	0.0260	32.29
## 6	9.38	0.000	37.44	36.36	0.4600	24.39
## 7	163.95	2.310	112.90	12.62	0.0250	88.69
## 8	111.67	0.000	127.30	8.87	0.2200	53.11
## 9	94.76	0.960	52.27	19.69	0.0140	26.59
## 10	107.72	0.000	0.56	1.80	0.0000	41.85
## 11	196.87	0.330	63.57	4.94	0.0060	20.25
## 12	193.35	0.056	63.48	4.94	0.0031	19.87
## 13	91.93	0.000	53.30	8.13	0.0034	18.75
## 14	18.03	0.000	70.09	13.10	0.0000	20.53
## 15	21.18	6.740	17.79	5.49	0.0000	7.57
## 16	102.61	0.000	68.15	7.54	3.0000	29.44
## 17	106.47	0.000	33.00	28.70	0.0160	15.17
## 18	109.58	0.000	54.16	11.38	0.1700	16.79
## 19	113.28	0.000	102.91	3.61	0.0000	68.14

##	eiqB11I.pos	eiqB11T.pos	cluster
## 1	24.31	0.000	1
## 2	57.66	0.000	1
## 3	57.66	0.000	1
## 4	57.66	0.000	1
## 5	9.63	0.000	2
## 6	34.75	0.000	3
## 7	51.87	0.041	4
## 8	5.25	0.000	5
## 9	20.81	0.082	6
## 10	1.80	0.000	7
## 11	21.36	0.084	8
## 12	21.36	0.230	8
## 13	7.58	0.061	6
## 14	2.28	0.000	2

```
## 15      9.48      0.000      3
## 16     10.91      3.000      6
## 17     10.08      0.000      6
## 18     17.33      0.220      6
## 19     11.92      0.070      5
```

#Clustering summaries based on non-markov approach

```
non_markov_clusters <- logged_data %>%
  group_by(orchard) %>%
  summarise(mean_honey_ab = mean(apisAb),
            mean_wild_ab = mean(adjusted_bees),
            mean_social_rich = mean(adjusted_social)) %>%
  mutate(cluster = maximum_cluster$cluster) %>%
  group_by(cluster) %>%
  summarise(mean_honey_ab = mean(mean_honey_ab),
            mean_wild_ab = mean(mean_wild_ab),
            mean_social_rich = mean(mean_social_rich),
            n = n())
```

After Adjustment:

```
adjusted_bee_values <- temp_nat_adjusted %>%
  group_by(orchard) %>%
  summarise(mean_honey_ab = mean(apisAb),
            mean_wild_ab = mean(adjusted_bees),
            mean_social_rich = mean(adjusted_social)
            )

adjusted_bee_summary <- cluster_summarise(bee_data = adjusted_bee_values)

#Agglom
adjusted_bee_summary$agglom_beas
```

```
## # A tibble: 14 x 5
##   mean_honey_ab mean_wild_ab mean_social_rich     n cluster
##   <dbl>         <dbl>         <dbl> <int> <chr>
## 1      6.56      1.27      0.379    19 aggl_node1
## 2      6.19      1.15      0.395    15 aggl_node2
## 3      7.94      1.72      0.319     4 aggl_node3
## 4      6.80      1.19      0.426    12 aggl_node4
## 5      3.75      1.01      0.270     3 aggl_node5
## 6      8.75      1.90      0.276     3 aggl_node6
## 7      5.5       1.17      0.449     1 aggl_node7
## 8      5.46      1.06      0.474     9 agglom_node8
## 9     10.8       1.56      0.281     3 agglom_node9
## 10     3.25      1.06      0.272     1 agglom_node10
## 11      4        0.980      0.269     2 agglom_node11
## 12      8        2.05      0.266     2 agglom_node12
## 13     10.2       1.59      0.295     1 agglom_node13
## 14      5.5       1.17      0.449     1 agglom_node14
```

#Kmeans

```
adjusted_bee_summary$kmeans_beas
```

```
## # A tibble: 15 x 5
##   mean_honey_ab mean_wild_ab mean_social_rich     n cluster
```

```
##           <dbl>           <dbl>           <dbl> <int> <chr>
##  1         6.56           1.27           0.379   19 kmeans_node1
##  2         6.19           1.15           0.395   15 kmeans_node2
##  3         7.94           1.72           0.319    4 kmeans_node3
##  4         6.5            1.13           0.356    5 kmeans_node4
##  5         6.04           1.16           0.414   10 kmeans_node5
##  6         7.94           1.72           0.319    4 kmeans_node6
##  7         NaN           NaN           NaN      0 kmeans_node7
##  8         3.25           1.06           0.272    1 kmeans_node8
##  9         7.31           1.15           0.377    4 kmeans_node9
## 10         6.04           1.16           0.414   10 kmeans_node10
## 11         NaN           NaN           NaN      0 kmeans_node11
## 12         7.17           1.76           0.327    3 kmeans_node12
## 13         10.2           1.59           0.295    1 kmeans_node13
## 14         NaN           NaN           NaN      0 kmeans_node14
## 15         NaN           NaN           NaN      0 kmeans_node15
```

```
##Non-Markov
##Maximum
#Clustering summaries based on non-markov approach
maximum_non_markov_clusters <- temp_nat_adjusted %>%
  group_by(orchard) %>%
  summarise(mean_honey_ab = mean(apisAb),
            mean_wild_ab = mean(adjusted_bees),
            mean_social_rich = mean(adjusted_social)) %>%
  mutate(cluster = maximum_cluster$cluster) %>%
  group_by(cluster) %>%
  summarise(mean_honey_ab = mean(mean_honey_ab),
            mean_wild_ab = mean(mean_wild_ab),
            mean_social_rich = mean(mean_social_rich),
            n= n())

##euclidean:
euclidean_non_markov_clusters <- temp_nat_adjusted %>%
  group_by(orchard) %>%
  summarise(mean_honey_ab = mean(apisAb),
            mean_wild_ab = mean(adjusted_bees),
            mean_social_rich = mean(adjusted_social)) %>%
  mutate(cluster = euclidean_cluster$cluster) %>%
  group_by(cluster) %>%
  summarise(mean_honey_ab = mean(mean_honey_ab),
            mean_wild_ab = mean(mean_wild_ab),
            mean_social_rich = mean(mean_social_rich),
            n= n())
```

Although honey bee abundance is one of the summarised variables, this is likely to be related directly to the `hive.acr` variable.

Comparing cluster values before and after adjustments, for now just looking at the agglomerative node clustering selection as it is, in my opinion, better than Kmeans.

```
##agglom
adjusted_bee_summary$agglom_bees %>%
  select(mean_wild_ab, mean_social_rich ,cluster) %>%
  rename(adjusted = mean_wild_ab) %>%
  rename(adjusted_social = mean_social_rich) %>%
```

```
inner_join(original_bee_summary$agglom_bees %>%
  select(mean_wild_ab, mean_social_rich, cluster)) %>%
select(cluster, mean_wild_ab, adjusted, mean_social_rich, adjusted_social)

## Joining, by = "cluster"

## # A tibble: 14 x 5
##   cluster      mean_wild_ab adjusted mean_social_rich adjusted_social
##   <chr>          <dbl>    <dbl>          <dbl>          <dbl>
## 1 aggl_node1      1.33      1.27            0.379          0.379
## 2 aggl_node2      1.23      1.15            0.386          0.395
## 3 aggl_node3      1.71      1.72            0.353          0.319
## 4 aggl_node4      1.23      1.19            0.376          0.426
## 5 aggl_node5      1.23      1.01            0.423          0.270
## 6 aggl_node6      1.70      1.90            0.135          0.276
## 7 aggl_node7      1.73      1.17            1.01           0.449
## 8 agglom_node8     0.963     1.06            0.325          0.474
## 9 agglom_node9     2.04      1.56            0.530          0.281
## 10 agglom_node10    1.32      1.06            0.405          0.272
## 11 agglom_node11    1.19      0.980           0.432          0.269
## 12 agglom_node12    1.97      2.05            0.101          0.266
## 13 agglom_node13    1.18      1.59            0.203          0.295
## 14 agglom_node14    1.73      1.17            1.01           0.449

##Kmeans
adjusted_bee_summary$kmeans_bees %>%
  select(mean_wild_ab, mean_social_rich ,cluster) %>%
  rename(adjusted = mean_wild_ab) %>%
  rename(adjusted_social = mean_social_rich) %>%
  inner_join(original_bee_summary$kmeans_bees %>%
    select(mean_wild_ab, mean_social_rich, cluster)) %>%
  select(cluster, mean_wild_ab, adjusted, mean_social_rich, adjusted_social)

## Joining, by = "cluster"

## # A tibble: 15 x 5
##   cluster      mean_wild_ab adjusted mean_social_rich adjusted_social
##   <chr>          <dbl>    <dbl>          <dbl>          <dbl>
## 1 kmeans_node1      1.33      1.27            0.379          0.379
## 2 kmeans_node2      1.23      1.15            0.386          0.395
## 3 kmeans_node3      1.71      1.72            0.353          0.319
## 4 kmeans_node4      1.36      1.13            0.516          0.356
## 5 kmeans_node5      1.17      1.16            0.320          0.414
## 6 kmeans_node6      1.71      1.72            0.353          0.319
## 7 kmeans_node7      NaN       NaN             NaN            NaN
## 8 kmeans_node8      1.32      1.06            0.405          0.272
## 9 kmeans_node9      1.37      1.15            0.544          0.377
## 10 kmeans_node10     1.17      1.16            0.320          0.414
## 11 kmeans_node11     NaN       NaN             NaN            NaN
## 12 kmeans_node12     1.89      1.76            0.403          0.327
## 13 kmeans_node13     1.18      1.59            0.203          0.295
## 14 kmeans_node14     NaN       NaN             NaN            NaN
## 15 kmeans_node15     NaN       NaN             NaN            NaN
```

Potentially can see two *paths* one with a mean slightly lower than the other if you think of the structure.

Looking at clusters

The aim of this section is to look at the clusterings from a low, medium, high point of view, so that an end user doesn't need to understand the numerical implications.

```
#Non_markov clustering summary maximum
maximum_final <- maximum_cluster %>%
  group_by(cluster) %>%
  summarise_all(funs(mean)) %>%
  inner_join(maximum_non_markov_clusters)

## Joining, by = "cluster"

#Non_markov clustering summary euclidean
euclidean_final <- euclidean_cluster %>%
  group_by(cluster) %>%
  summarise_all(funs(mean)) %>%
  inner_join(euclidean_non_markov_clusters)

## Joining, by = "cluster"

#Cluster 1 -> Low insecticide, High Fungicide, low thinner
#Cluster 2 -> High insecticide, Low Fungicide, low (no) thinner
#Cluster 3 -> middle insecticide, High Fungicide, low thinner
#Cluster 4 -> Low insecticide, High Fungicide, High thinner

#Agglom clustering summary (end stage)
bee_summary_agglom <- adjusted_bee_summary$agglom_bees %>%
  slice(8:14) %>%
  select(-cluster) %>%
  mutate(cluster = c(1,2,3,4,5,6,7))

pesticide_summary_agglom <- aggl_c(aggl_node2, ".pos", 1) %>%
  bind_rows(aggl_c(aggl_node2, ".pos", 2) %>%
    mutate(cluster = if_else(cluster == 1, 3, 4))) %>%
  bind_rows(aggl_c(aggl_node3, ".pos", 1) %>%
    mutate(cluster = if_else(cluster == 1, 5, 6))) %>%
  bind_rows(aggl_c(aggl_node3, ".pos", 2) %>%
    mutate(cluster = if_else(cluster == 1, 7, 8))) %>%
  group_by(cluster) %>%
  summarise_all(funs(mean))

##final agglom
agglom_summary <- bee_summary_agglom %>%
  inner_join(pesticide_summary_agglom) %>%
  select(-orchard) #>%

## Joining, by = "cluster"

#1 for high, 0 for low, just using bloom
# mutate(pest_rating = c(0,0,1,1,0,0,1)) %>%
# mutate(insect_rating = c(1,0,1,0,0,1,1)) %>%
# mutate(thinner_rating = c(1,0,0,1,0,1,0))
```

```

#Kmeans
bee_summary_kmeans <- adjusted_bee_summary$kmeans_bees %>%
  slice(8:14) %>%
  select(-cluster) %>%
  mutate(cluster = c(1,2,3,4,5,6,7))

pesticide_summary_kmeans <- aggl_c(kmeans_node2, ".pos", 1) %>%
  bind_rows(aggl_c(kmeans_node2, ".pos", 2) %>%
    mutate(cluster = if_else(cluster == 1, 3, 4))) %>%
  bind_rows(aggl_c(kmeans_node3, ".pos", 1) %>%
    mutate(cluster = if_else(cluster == 1, 5, 6))) %>%
  bind_rows(aggl_c(kmeans_node3, ".pos", 2) %>%
    mutate(cluster = if_else(cluster == 1, 7, 8))) %>%
  group_by(cluster) %>%
  summarise_all(funs(mean))

##
kmeans_summary <- bee_summary_kmeans %>%
  inner_join(pesticide_summary_kmeans) %>%
  select(-orchard)

## Joining, by = "cluster"

#write.csv(agglom_summary, "agglom_summary.csv")

#Overall
#HLH
#HHL
#HHL
#HLH
#LLL
#LHH
#LLL

#Just Bloom
#LHH
#LLL
#HHL
#HLH
#LLL
#LHH
#HHL

##Just bloom and just fungicide and insecticide

##Final summaries between models This section summarises all of the cluster modelled scenarios
round_df <- function(df, digits = 2) {
  #Using a temp variable to not round cluster and n as these are integers
  temp <- df %>%
    mutate(cluster = as.character(cluster)) %>%
    mutate(n = as.character(n))
  nums <- vapply(temp, is.numeric, FUN.VALUE = logical(1))

```



```

df[,nums] <- round(df[,nums], digits = digits)

df
}

#Need a different rounding function when cluster/n not present
round_df_2 <- function(df, digits = 2) {
  nums <- vapply(df, is.numeric, FUN.VALUE = logical(1))
  df[,nums] <- round(df[,nums], digits = digits)

  df
}

apple_order <- function(data){
  data %>%
    select(cluster, n, mean_wild_ab, mean_social_rich, contains("pre"), contains("blm"), contains("pos"))
}

#Rounding data for shiny app display
maximum_final_standardised <- round_df(maximum_final_standardised) %>%
  apple_order()
euclidean_final_standardised <- round_df(euclidean_final_standardised) %>%
  apple_order()
#no n
agglom_summary_standardised <- round_df(agglom_summary_standardised) %>%
  apple_order()
#no n
kmeans_summary_standardised <- round_df(kmeans_summary_standardised) %>%
  apple_order()
# 8 clusters in this method - 6 different options so this is the cluster that will used - missing options
maximum_final <- round_df(maximum_final) %>%
  apple_order()

maximum_final_app <- maximum_final %>%
  mutate(fung_level = c("high", "low", "low", "high", "high", "low", "low", "high")) %>%
  mutate(insect_level = c("low", "high", "high", "high", "low", "high", "low", "low")) %>%
  mutate(thinner_level = c("low", "low", "high", "low", "high", "high", "low", "high"))

protocol_summary <- maximum_final_app %>%
  group_by(fung_level, insect_level, thinner_level) %>%
  summarise(mean_wild_ab = mean(mean_wild_ab),
            mean_social_rich = mean(mean_social_rich)) %>%
  ungroup() %>%
  mutate(unlogged_ab = exp(mean_wild_ab) + 1) %>%
  mutate(unlogged_rich = exp(mean_social_rich) + 1) %>%
  round_df_2()

euclidean_final <- round_df(euclidean_final) %>%
  apple_order()

```

```

#7 Clusters in this method - 5 diff options
agglom_summary <- round_df(agglom_summary) %>%
  mutate(fung_level = c(1,1,1,1,0,0,0)) %>%
  mutate(insect_level = c(1,1,1,0,0,1,0)) %>%
  mutate(thinner_level = c(1,0,1,1,0,1,0)) %>%
  apple_order()
#Removes previous NA inserted due to joining.
kmeans_summary <- round_df(kmeans_summary) %>%
  na.omit() %>%
  apple_order()

#Displaying table outputs
maximum_final_standardised

## # A tibble: 2 x 12
##   cluster      n mean_wild_ab mean_social_rich eqB11F.pre eqB11I.pre
##   <int> <int>      <dbl>          <dbl>      <dbl>      <dbl>
## 1     1     17      1.28            0.4        0.1       -0.21
## 2     2      2      1.15            0.22       -0.86      1.76
## # ... with 6 more variables: eqB11F.blm <dbl>, eqB11I.blm <dbl>,
## #   eqB11T.blm <dbl>, eqB11F.pos <dbl>, eqB11I.pos <dbl>,
## #   eqB11T.pos <dbl>
euclidean_final_standardised

## # A tibble: 5 x 12
##   cluster      n mean_wild_ab mean_social_rich eqB11F.pre eqB11I.pre
##   <int> <int>      <dbl>          <dbl>      <dbl>      <dbl>
## 1     1      9      1.44            0.49       -0.54      -0.28
## 2     2      5      1.17            0.3        1.12      -0.290
## 3     3      3      1.01            0.27        0.33       0.14
## 4     4      1      1.57            0.3       -1.58       3.87
## 5     5      1      0.74            0.14       -0.15      -0.34
## # ... with 6 more variables: eqB11F.blm <dbl>, eqB11I.blm <dbl>,
## #   eqB11T.blm <dbl>, eqB11F.pos <dbl>, eqB11I.pos <dbl>,
## #   eqB11T.pos <dbl>
agglom_summary_standardised

## # A tibble: 4 x 12
##   cluster      n mean_wild_ab mean_social_rich eqB11F.pre eqB11I.pre
##   <dbl> <int>      <dbl>          <dbl>      <dbl>      <dbl>
## 1     1     13      1.24            0.43       -0.14      -0.28
## 2     2      4      1.44            0.28        0.9       0.02
## 3     3      1      0.74            0.14       -0.15      -0.34
## 4     5      1      1.57            0.3       -1.58       3.87
## # ... with 6 more variables: eqB11F.blm <dbl>, eqB11I.blm <dbl>,
## #   eqB11T.blm <dbl>, eqB11F.pos <dbl>, eqB11I.pos <dbl>,
## #   eqB11T.pos <dbl>
kmeans_summary_standardised

## # A tibble: 3 x 12
##   cluster      n mean_wild_ab mean_social_rich eqB11F.pre eqB11I.pre
##   <dbl> <int>      <dbl>          <dbl>      <dbl>      <dbl>

```

```
## 1      1      14      1.26      0.42      -0.25      0.02
## 2      2       4      1.44      0.28       0.9      0.02
## 3      5       1      0.74      0.14     -0.15     -0.34
## # ... with 6 more variables: eqB11F.blm <dbl>, eqB11I.blm <dbl>,
## #   eqB11T.blm <dbl>, eqB11F.pos <dbl>, eqB11I.pos <dbl>,
## #   eqB11T.pos <dbl>
```

maximum_final

```
## # A tibble: 8 x 12
##   cluster      n mean_wild_ab mean_social_rich eqB11F.pre eqB11I.pre
##   <int> <int>      <dbl>      <dbl>      <dbl>      <dbl>
## 1      1      4      1.48      0.36     157.        0
## 2      2      2      1.85      0.34     29.9        0
## 3      3      2      1.58      0.3      15.3        3.37
## 4      4      1      1.06      0.27     164.        2.31
## 5      5      2      0.98      0.27     112.        0
## 6      6      5      1.12      0.5      101.        0.19
## 7      7      1      1.58      0.48     108.        0
## 8      8      2      0.580     0.33     195.        0.19
## # ... with 6 more variables: eqB11F.blm <dbl>, eqB11I.blm <dbl>,
## #   eqB11T.blm <dbl>, eqB11F.pos <dbl>, eqB11I.pos <dbl>,
## #   eqB11T.pos <dbl>
```

euclidean_final

```
## # A tibble: 4 x 12
##   cluster      n mean_wild_ab mean_social_rich eqB11F.pre eqB11I.pre
##   <int> <int>      <dbl>      <dbl>      <dbl>      <dbl>
## 1      1      6      1.18      0.35     170.        0.06
## 2      2      4      1.72      0.32     22.6        1.68
## 3      3      3      1.01      0.27     130.        0.77
## 4      4      6      1.2       0.5      102.        0.16
## # ... with 6 more variables: eqB11F.blm <dbl>, eqB11I.blm <dbl>,
## #   eqB11T.blm <dbl>, eqB11F.pos <dbl>, eqB11I.pos <dbl>,
## #   eqB11T.pos <dbl>
```

agglom_summary

```
## # A tibble: 7 x 12
##   cluster      n mean_wild_ab mean_social_rich eqB11F.pre eqB11I.pre
##   <dbl> <int>      <dbl>      <dbl>      <dbl>      <dbl>
## 1      1      9      1.06      0.47     127.        0.15
## 2      2      3      1.56      0.28     162.        0
## 3      3      1      1.06      0.27     164.        2.31
## 4      4      2      0.98      0.27     112.        0
## 5      5      2      2.05      0.27     31.5        3.37
## 6      6      1      1.59      0.290     9.38        0
## 7      7      1      1.17      0.45     18.0        0
## # ... with 6 more variables: eqB11F.blm <dbl>, eqB11I.blm <dbl>,
## #   eqB11T.blm <dbl>, eqB11F.pos <dbl>, eqB11I.pos <dbl>,
## #   eqB11T.pos <dbl>
```

kmeans_summary

```
## # A tibble: 5 x 12
##   cluster      n mean_wild_ab mean_social_rich eqB11F.pre eqB11I.pre
```

```
##      <dbl> <int>      <dbl>      <dbl>      <dbl>      <dbl>
## 1      1      1      1.06      0.27      164.      2.31
## 2      2      4      1.15      0.38      110.      0
## 3      3     10      1.16      0.41      133.      0.19
## 4      5      3      1.76      0.33       27.0     2.25
## 5      6      1      1.59      0.290      9.38      0
## # ... with 6 more variables: eiqB11F.blm <dbl>, eiqB11I.blm <dbl>,
## #   eiqB11T.blm <dbl>, eiqB11F.pos <dbl>, eiqB11I.pos <dbl>,
## #   eiqB11T.pos <dbl>
```

#Could do something with this and plots coloured by cluster or something like that perhaps?

```
temp_nat_adjusted %>%
  group_by(orchard) %>%
  summarise(mean_honey_ab = mean(apisAb),
            mean_wild_ab = mean(adjusted_bees),
            mean_social_rich = mean(adjusted_social))
```

```
## # A tibble: 19 x 4
##   orchard mean_honey_ab mean_wild_ab mean_social_rich
##   <fct>      <dbl>      <dbl>      <dbl>
## 1 A          5.12      1.22      0.595
## 2 B          8        1.70     -0.207
## 3 C          5.75      1.67      0.556
## 4 D         18.8      1.32      0.493
## 5 E          7.25      2.53      0.230
## 6 F         10.2      1.59      0.295
## 7 G          3.25      1.06      0.272
## 8 H.nooil    3.75      1.09      0.574
## 9 I          1.75      1.04      0.808
## 10 J         6.25      1.58      0.485
## 11 K         6.25      0.347     0.347
## 12 L          6        0.806     0.321
## 13 M         2.75      1.55      0.562
## 14 N          5.5      1.17      0.449
## 15 O.nooil   8.75      1.57      0.303
## 16 P          2        0.735     0.140
## 17 Q         15        1.07      0.487
## 18 R          4        1.22      0.525
## 19 S         4.25      0.869    -0.0364
```

##Looking at total amount for raw data max and euclid

```
maximum_final %>%
  mutate(sum = rowSums(.[5:12])) %>%
  select(cluster, n , mean_wild_ab, mean_social_rich,sum)
```

```
## # A tibble: 8 x 5
##   cluster      n mean_wild_ab mean_social_rich  sum
##   <int> <int>      <dbl>      <dbl> <dbl>
## 1      1      4      1.48      0.36  300.
## 2      2      2      1.85      0.34  133.
## 3      3      2      1.58      0.3   106.
## 4      4      1      1.06      0.27  432.
## 5      5      2      0.98      0.27  303.
## 6      6      5      1.12      0.5   205.
## 7      7      1      1.58      0.48  154.
## 8      8      2      0.580     0.33  305.
```

```
euclidean_final %>%
  mutate(sum = rowSums(.[5:12])) %>%
  select(cluster, n , mean_wild_ab, mean_social_rich,sum)
```

```
## # A tibble: 4 x 5
##   cluster      n mean_wild_ab mean_social_rich    sum
##   <int> <int>      <dbl>          <dbl> <dbl>
## 1     1     6      1.18            0.35  302.
## 2     2     4      1.72            0.32  119.
## 3     3     3      1.01            0.27  346.
## 4     4     6      1.2             0.5   196.
```

#Saving the final clusters for use in the shiny app

```
save(data_2012, maximum_final, maximum_final_app, euclidean_final, agglom_summary, kmeans_summary,
      maximum_final_standardised, euclidean_final_standardised,
      agglom_summary_standardised, kmeans_summary_standardised, clust_comps, standardised_clus
```