James Valles
DSC 441: Homework 3

Part 1: K-means

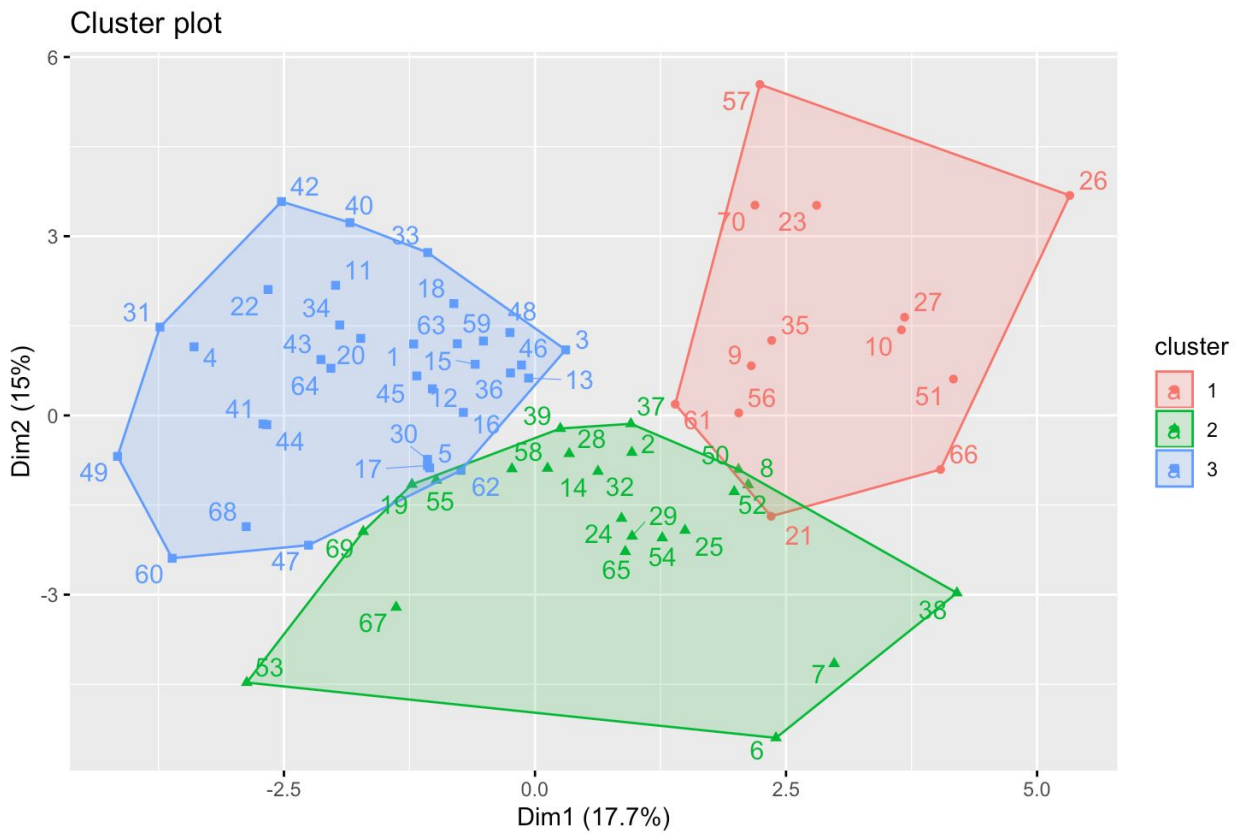
1.

nc <- NbClust(df2, min.nc=3, max.nc=6, method="kmeans")
The optimal value of K would be 3.

```
*******************************************************************
* Among all indices:
* 14 proposed 3 as the best number of clusters
* 6 proposed 4 as the best number of clusters
* 2 proposed 5 as the best number of clusters
* 2 proposed 6 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3


*******************************************************************
> |
```

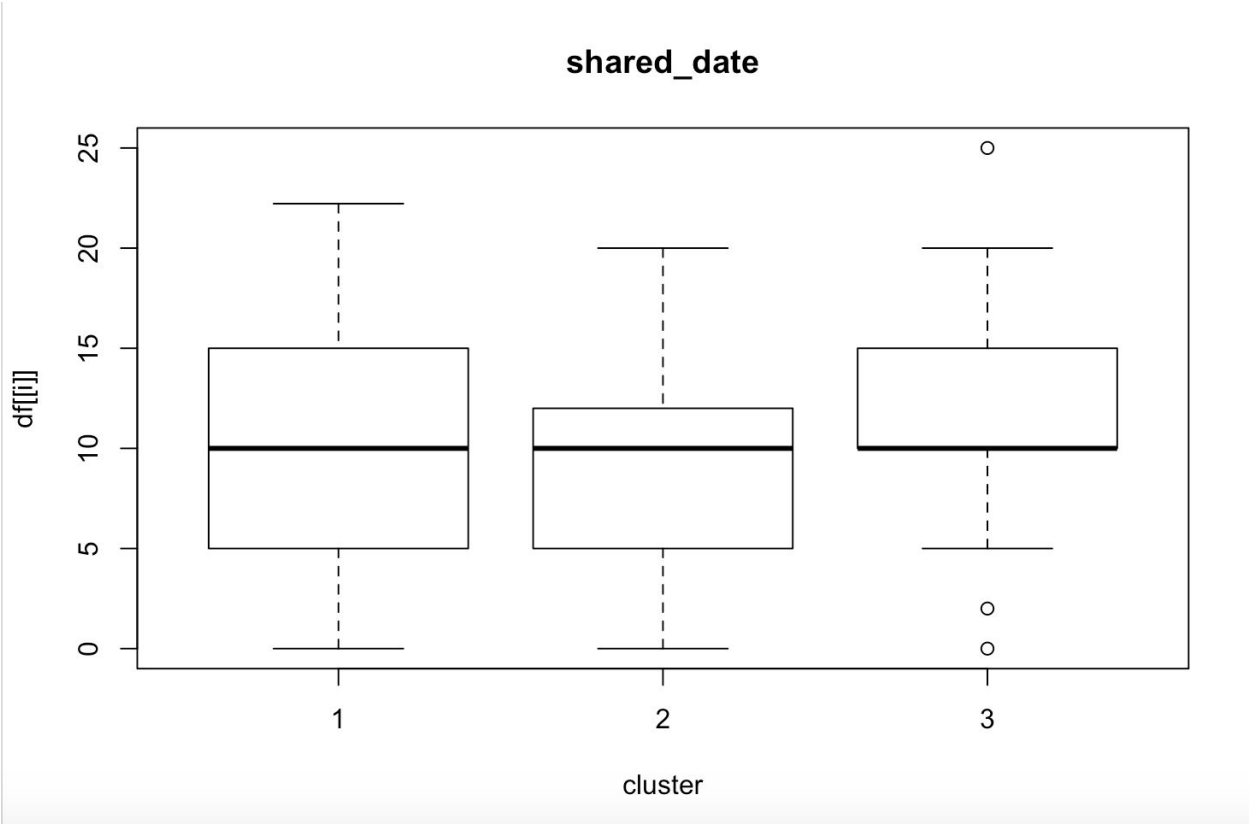
2.

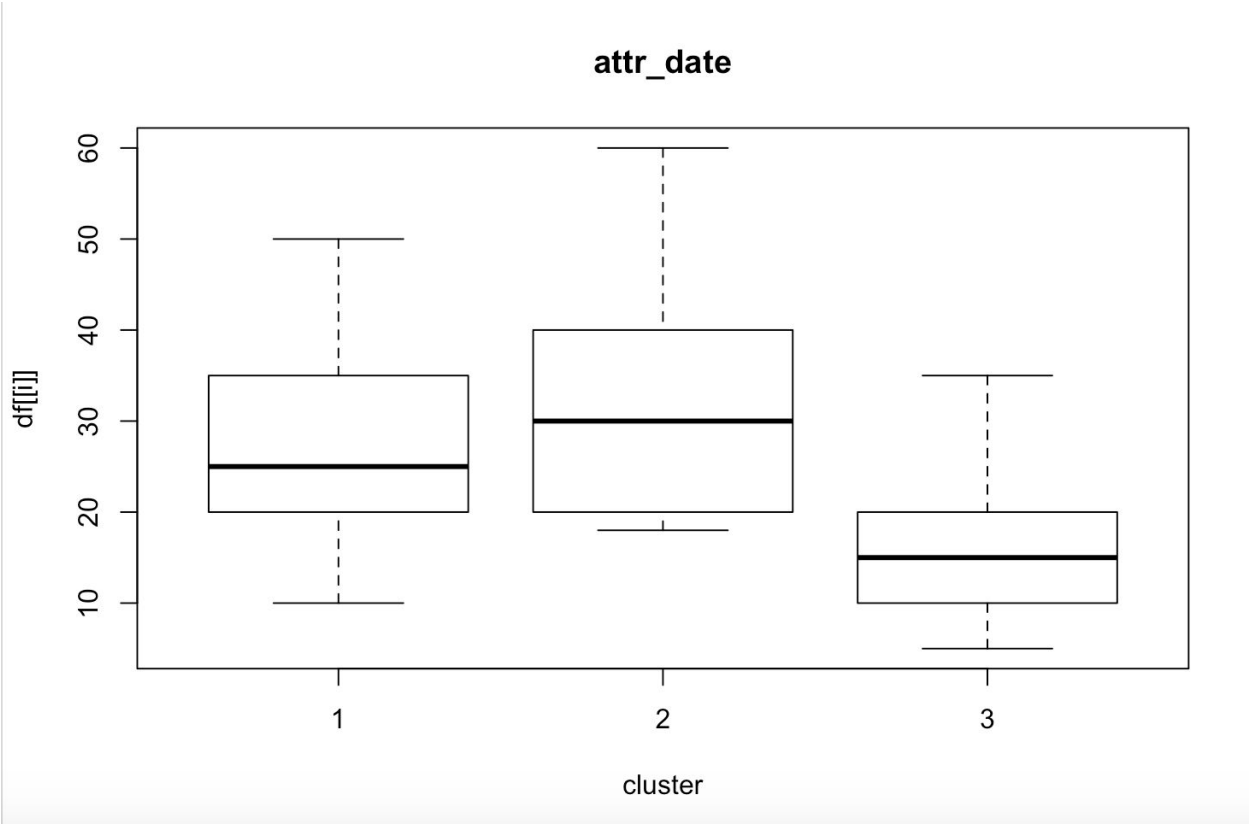
fviz_cluster(km_3, data = df2, repel = TRUE, show.clust.cent = FALSE)

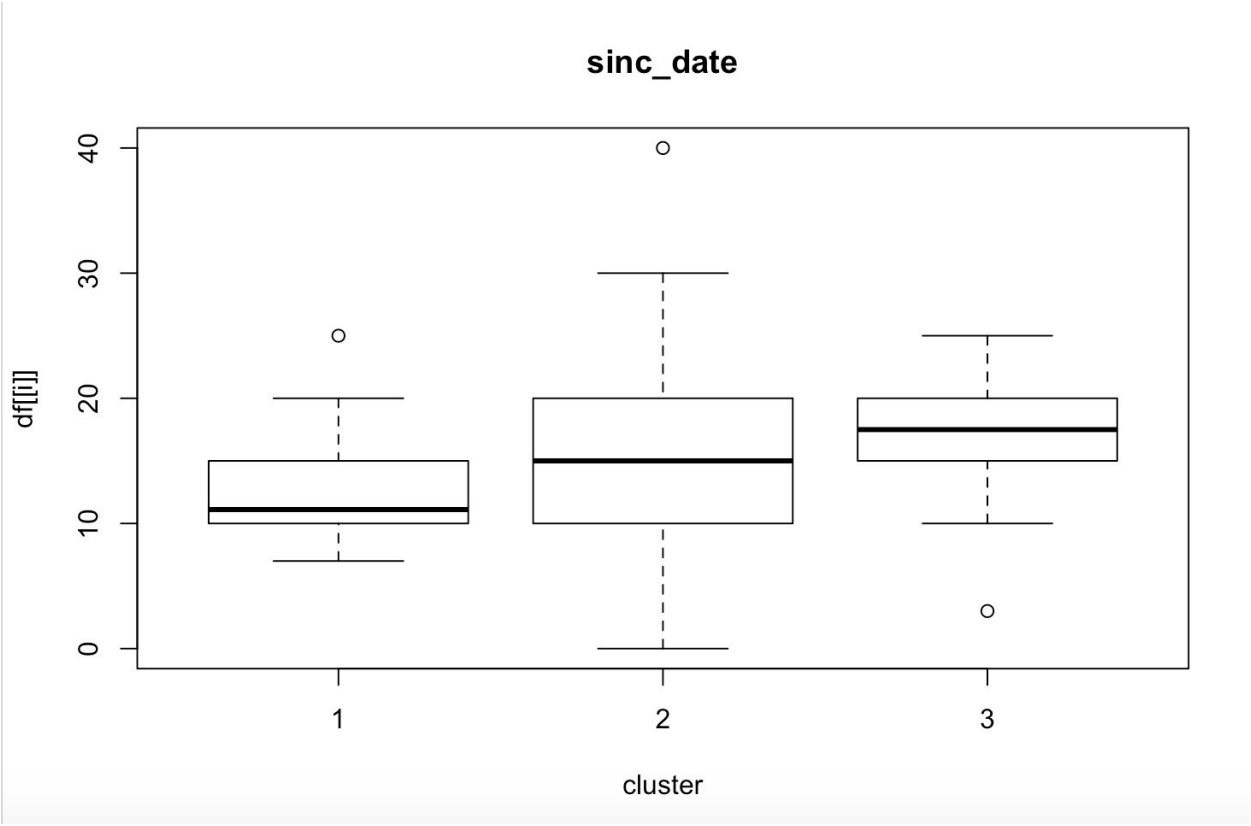


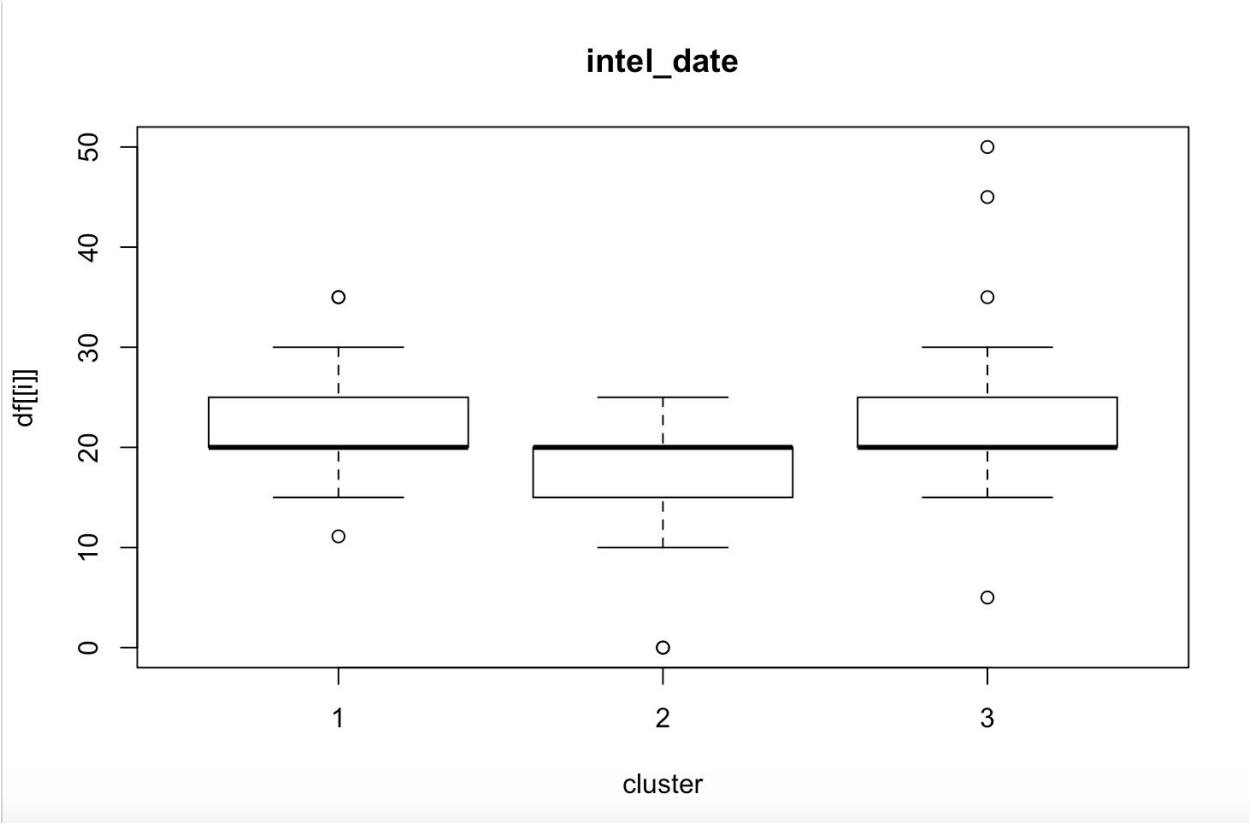
3. Generate boxplots for variables: attr_date, sinc_date, intel_date, amb_date, shared_date

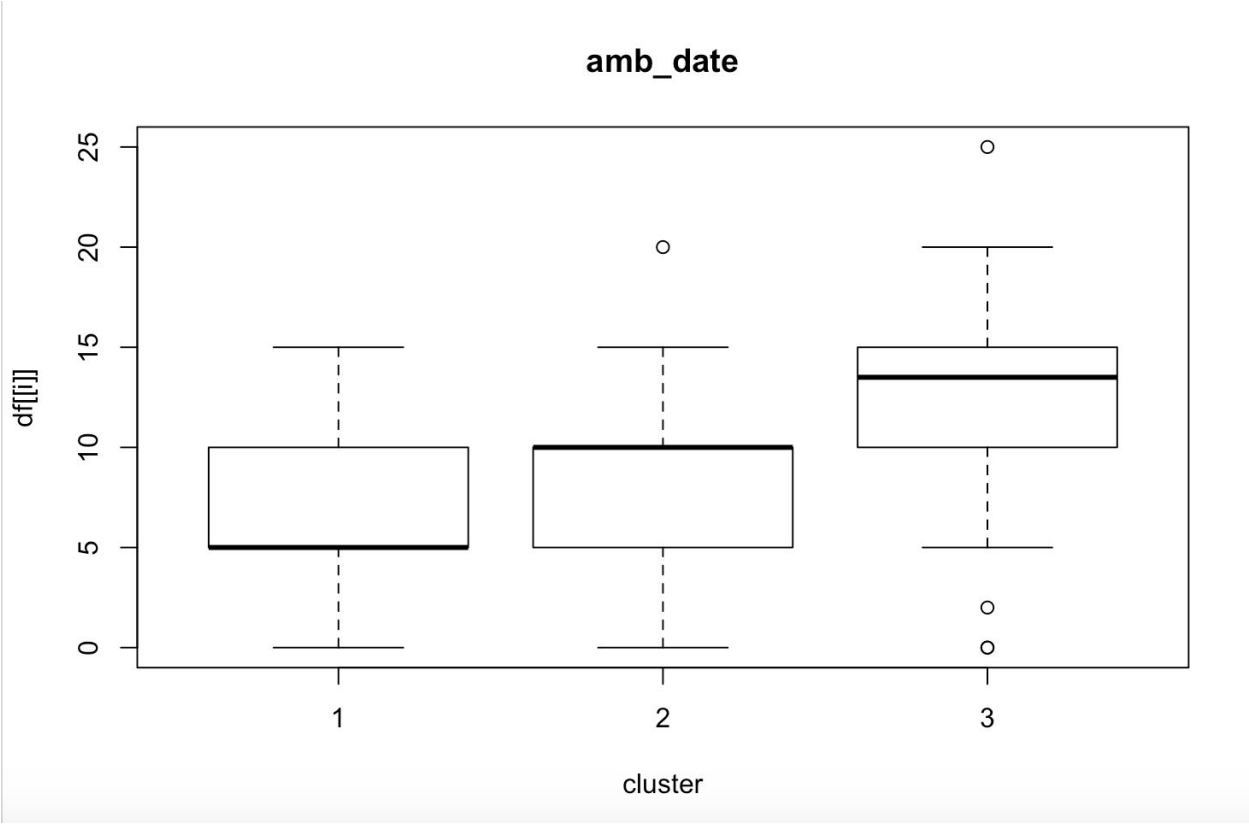
```
par(mfrow-c(1,5))
for (i in 18:22){
  boxplot(df[[i]] ~ km_3$cluster, xlab="cluster", main=names(df)[i])
}
par(mfrow=c(1,1)) # Reset to default plot settings
```






sinc_date
df[[i]]
0
10
20
30
40
1
2
3
cluster


intel_date
df[[i]]
0
10
20
30
40
50
1
2
3
cluster



4.

Analyzing the Boxplots

Attractiveness
In interpreting the cluster boxplots for each of the variables, cluster 2 appears to be most interested in attractiveness and has the highest variability, followed by cluster 1. Cluster 3 seems to be least interested in attractiveness out of all three clusters.

Sincerity
When looking at participants in the three groups, Cluster 3 appears to be the most interested in a prospective date who is sincere, followed by Cluster 2, which appears to have the largest variance when looking at this trait. Cluster 1 is also interested in sincerity, but appears to be the least interested when compared to the other groups.

Intelligence
Intelligence is equally as important to participants in cluster 1 and cluster 3. While, most people in cluster 2 also find intelligence to be important, but there are some who find it less important than those in groups 1 and 3.

Ambition
Those participants in cluster 3 are most interested in a prospective date who is ambitious followed by those people in group 2. However, it appears those people in group 1 are the least interested in ambition when compared to the other two clusters.

Shared Interest
It appears that participants in all three clusters are looking for a prospective date who shares the same interests. However, when digging deeper, cluster 1 appears to have a greater variance in opinions on this. While participants in all three groups want someone with shared interest, those

in cluster 3 appear to be more interested in this factor, followed by those in group 1, and then those participants in group 2.

Overall Narrative:

Group 1 - Participants are most interested in dating someone who is intelligent and shares their interests. They also care about finding an attractive person, but aren't as necessarily concerned with this quality when compared to participants in the other two groups. Last, participants in this group do not necessarily care about looking for someone who is ambitious.

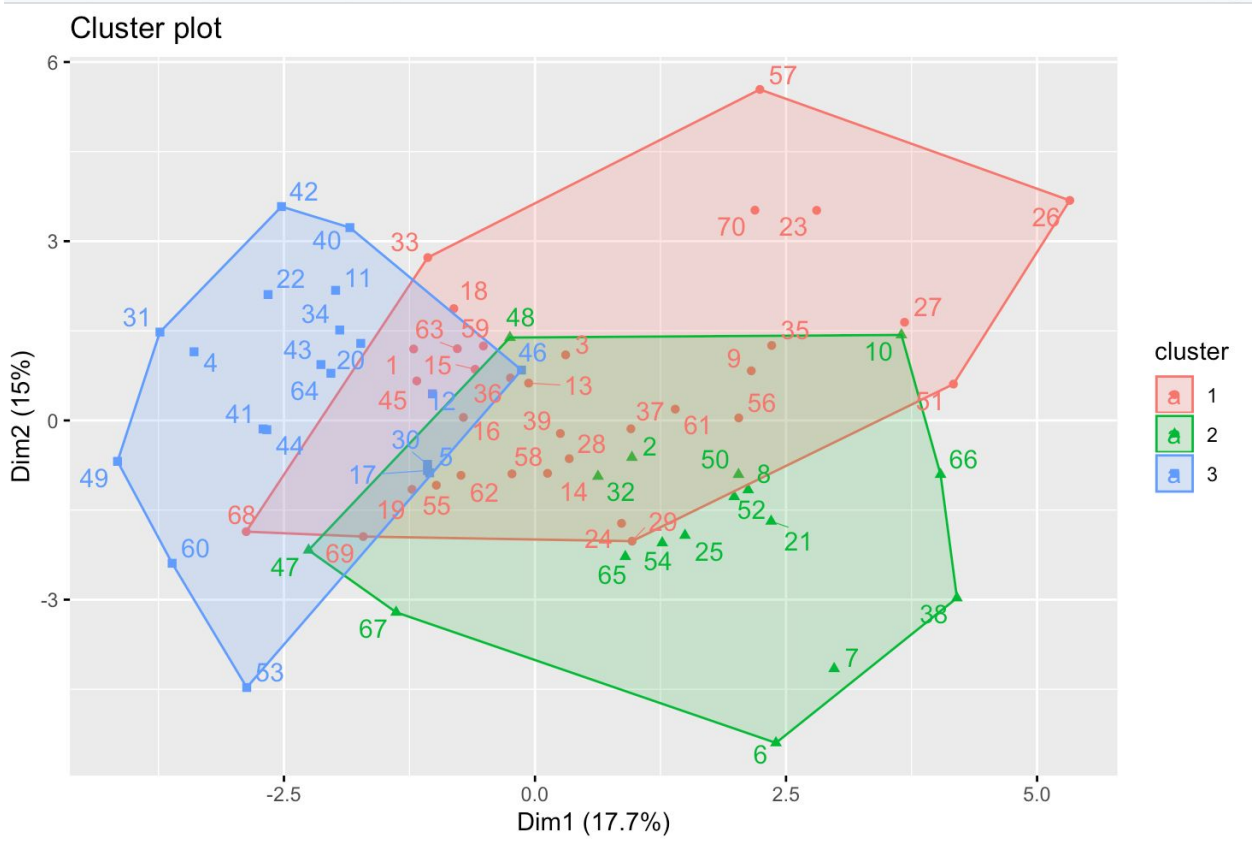
Group 2 - Participants are most interested in dating someone who is attractive and shares their interests. They also care about finding an ambitious, sincere, and intelligent person, but aren't as necessarily concerned with these qualities when compared to participants in the other two groups.

Group 3 - Participants are most interested in dating someone who is sincere, ambitious, and shares their interests. They also care about finding an intelligent person. Last, participants in this group do not necessarily care about looking for someone who is attractive.

Part 2: K-medoids (PAM)

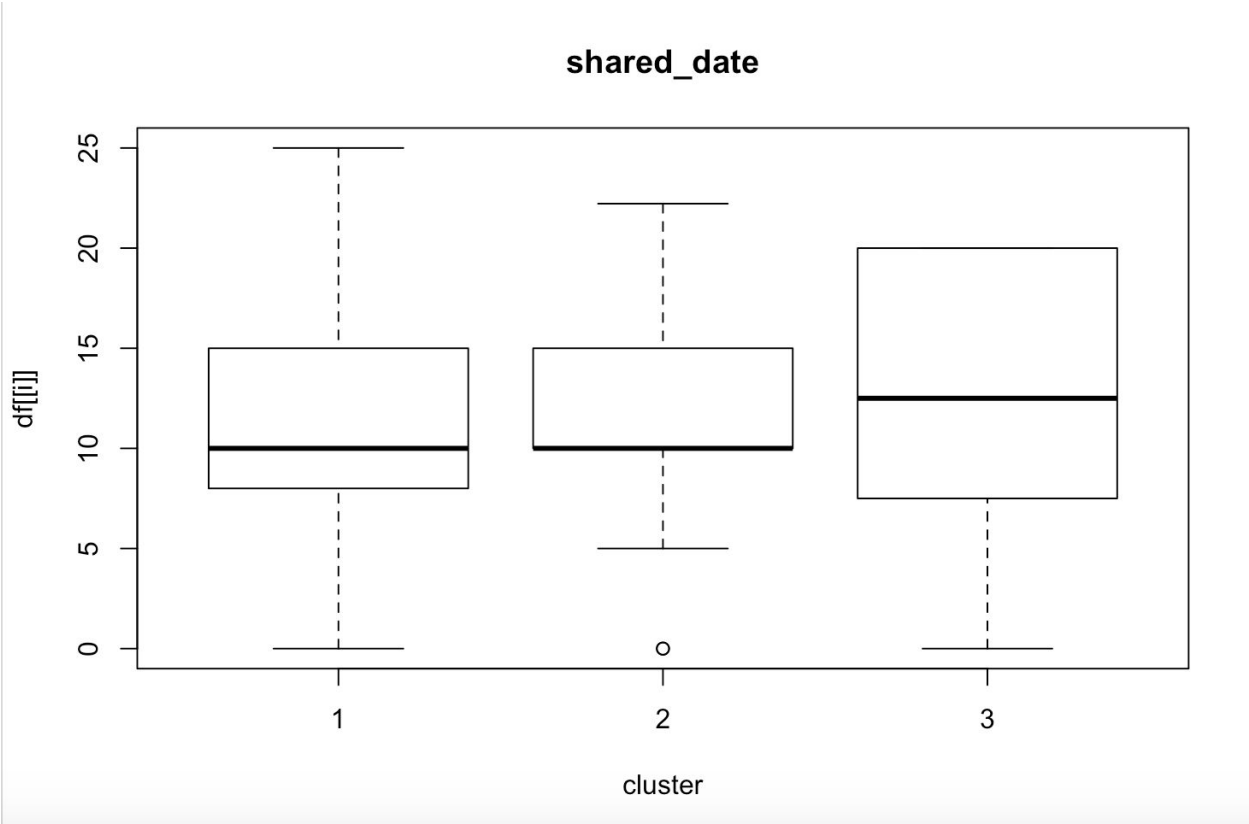
1.

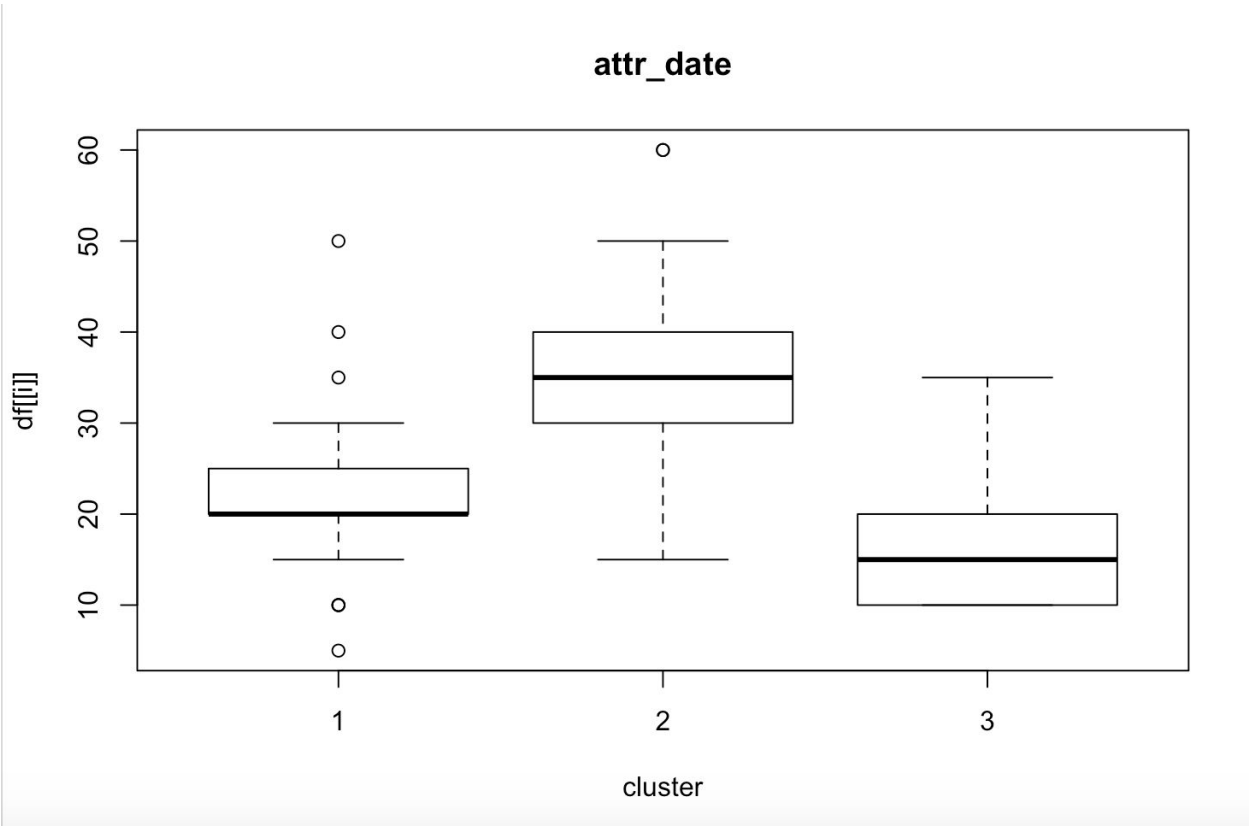
fviz_cluster(pam_3, data = df2, repel = TRUE, show.clust.cent = FALSE)

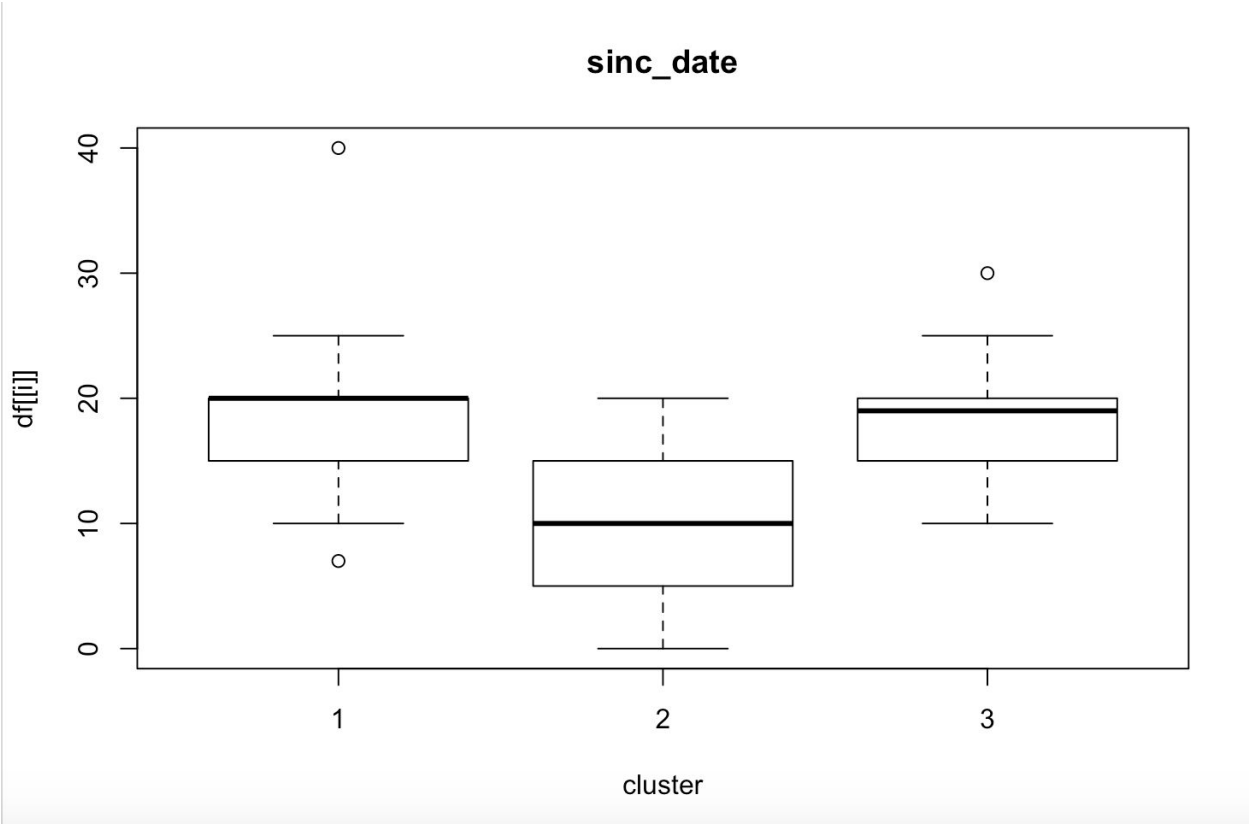


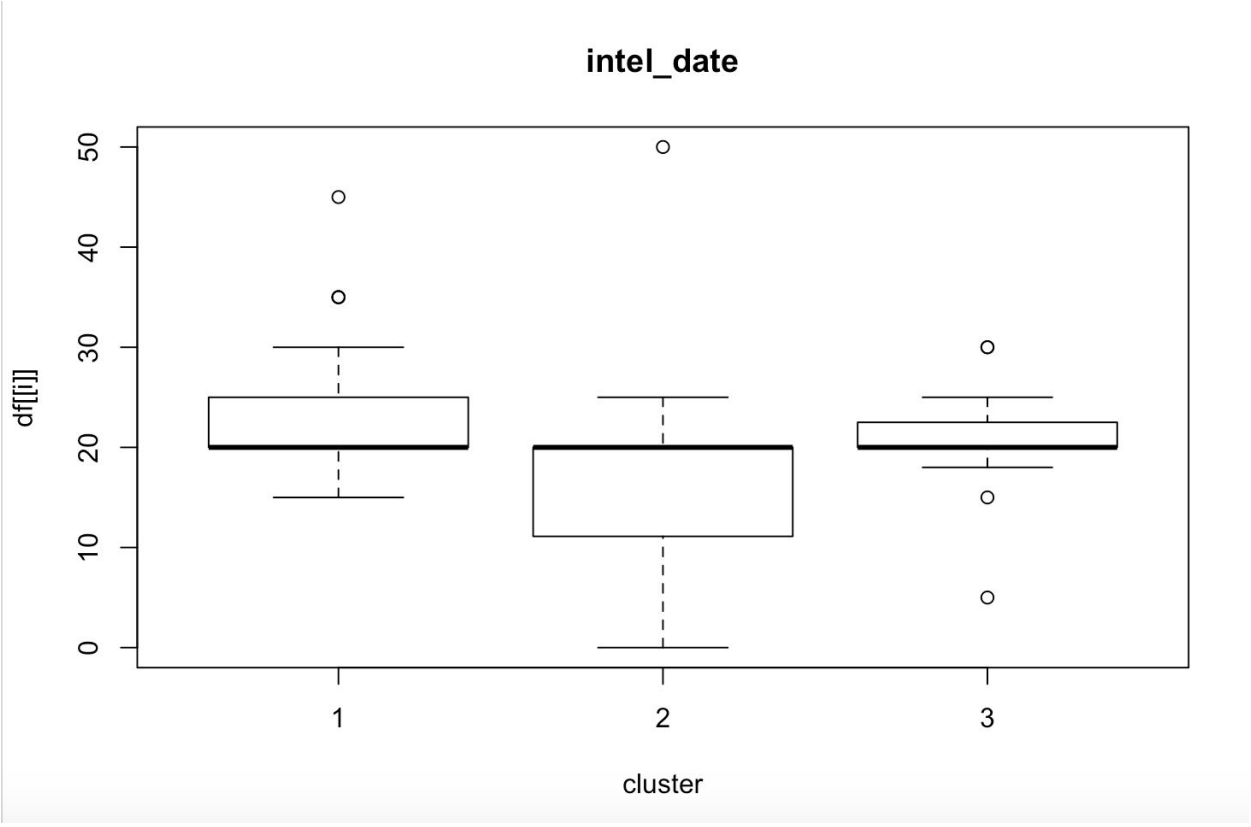
2. Generate boxplots for variables: attr_date, sinc_date, intel_date, amb_date, shared_date

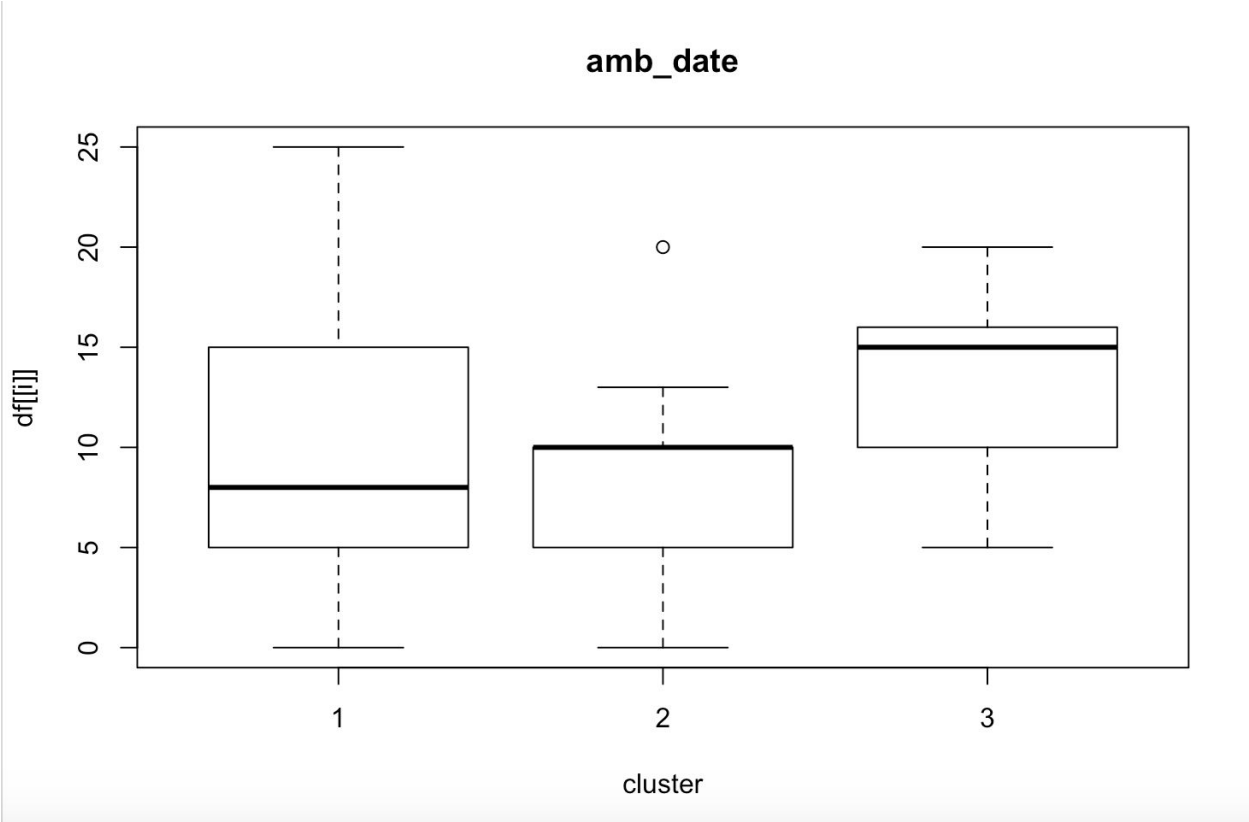
```
par(mfrow-c(1,5))
for (i in 18:22){
  boxplot(df[[i]] ~ pam_3$cluster, xlab="cluster", main=names(df)[i])
}
par(mfrow=c(1,1)) # Reset to default plot settings
```






sinc_date
df[[i]]
0
10
20
30
40
1
2
3
cluster


intel_date
df[[i]]
0
10
20
30
40
50
1
2
3
cluster



How does the dating preferences for k-medoids clusters compare to the k-mean clusters?

Overall, I am noticing a reduced distribution spread and less Q1-Q3 overlapping for the boxplots using k-medoids compared to k-mean. When comparing the specific dating preferences, there definitely appears to be differences, they are as follows.

Attractiveness
For attractiveness, the preferences seem to be the same for both k-medoids and k-means clusters. Cluster 2 seems to be more drawn to a person who is attractive, followed by Cluster 1, and then Cluster 3. The noticeable difference is there is less of a Q1 - Q3 spread when looking at k-medoids Clusters 1 & 2 compared to k-means.

Sincerity
The participants' preferences for the k-medoids clusters definitely changes compared to k-means. In the k-means cluster, the order of clusters most interested in someone who is sincere is Cluster 3 (most interested), 2, 1 (least interested). However, when looking at k-medoids the order changes, Clusters 1 & 3 seem to be most interested, with Cluster 2 being the least interested in dating someone who is sincere.

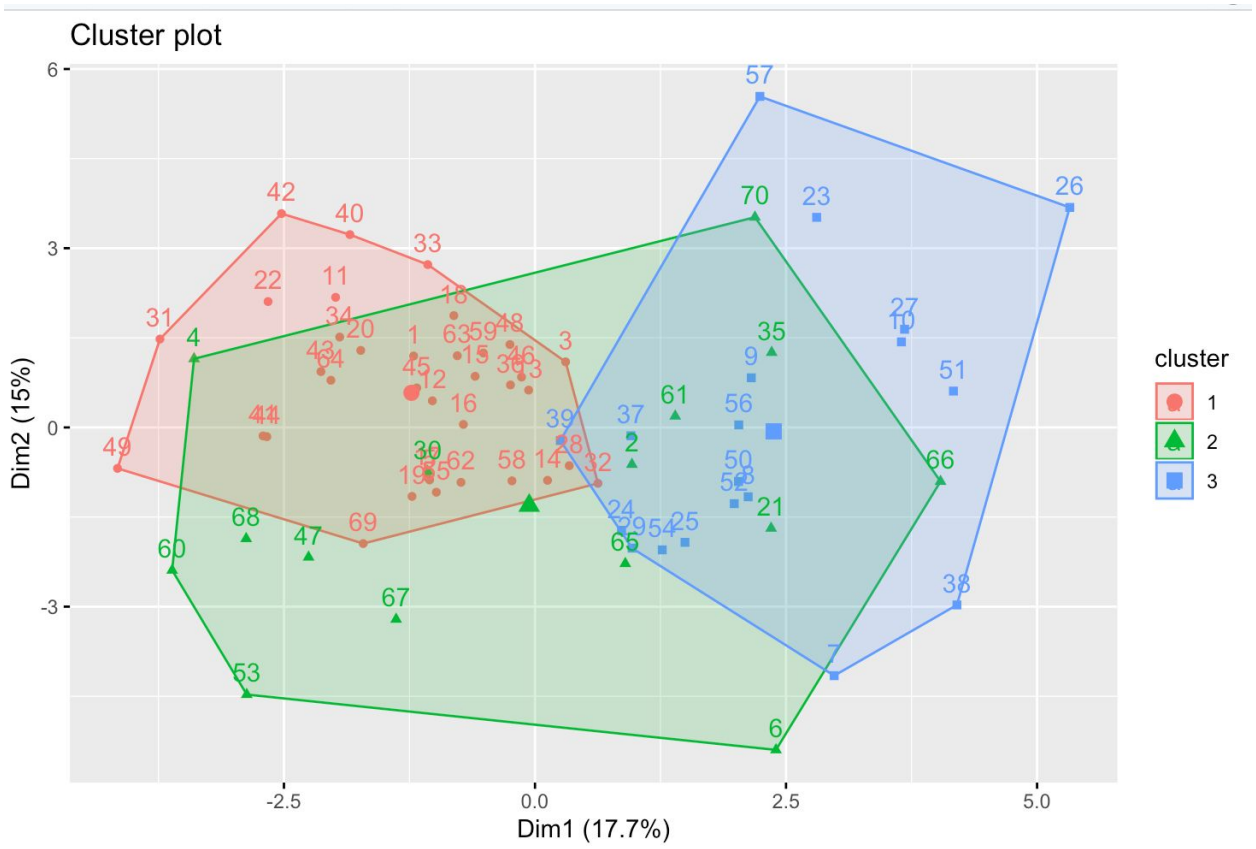
Intelligence
For intelligence, the preferences didn't seem to change much when looking at the k-medoids clusters and k-means clusters. Participants in all three groups appear to be equally interested in dating someone who is intelligent.

Ambitious
The clusters preference in wanting an ambitious person didn't change much when comparing k-medoids to k-means clusters. One difference, k-medoids Cluster 1 had a larger distribution when compared to k-means Cluster 1.

Shared Interests
k-medoids cluster 3 appears to be slightly more interested in finding someone with shared interest compared to k-means cluster 3. Other than that, there wasn't much a shift in preference when comparing k-medoids and k-means clusters.

Part 3: Hierarchical Clustering

```
hc <- hclust(dist(df2), method="complete")

hc$cluster <- cutree(hc, k = 3)

fviz_cluster(list(data = df2, cluster = hc$cluster))
```



2.

Based solely on the plots, Hierarchical clustering seems worse than k-means. Because in k-means there is more cohesion and the clusters are well separated from one another, whereas, in hierarchical clustering there is less separation, and a lot of overlapping making it hard to distinguish the three clusters.

3.

```
km3_stats <- cluster.stats(dist(df2),  km_3$cluster, silhouette = TRUE)
km3_stats$avg.silwidth
```

K-means:
[1] 0.09880885

```
pam3_stats <- cluster.stats(dist(df2),  pam_3$cluster, silhouette = TRUE)
pam3_stats$avg.silwidth
```

K-medoids:
[1] 0.04747687

```
hc_stats <- cluster.stats(dist(df2),  hc$cluster, silhouette = TRUE)
hc_stats$avg.silwidth
```

Hierarchical Clustering:
[1] 0.07579023