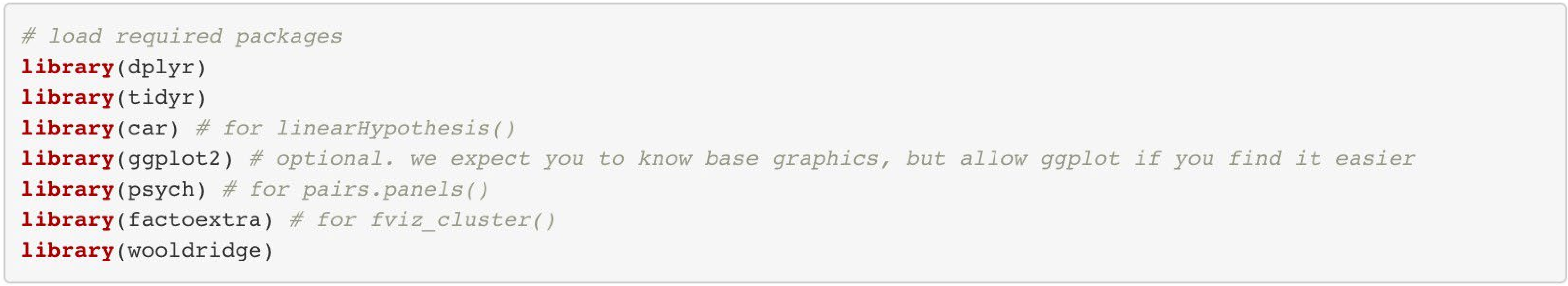
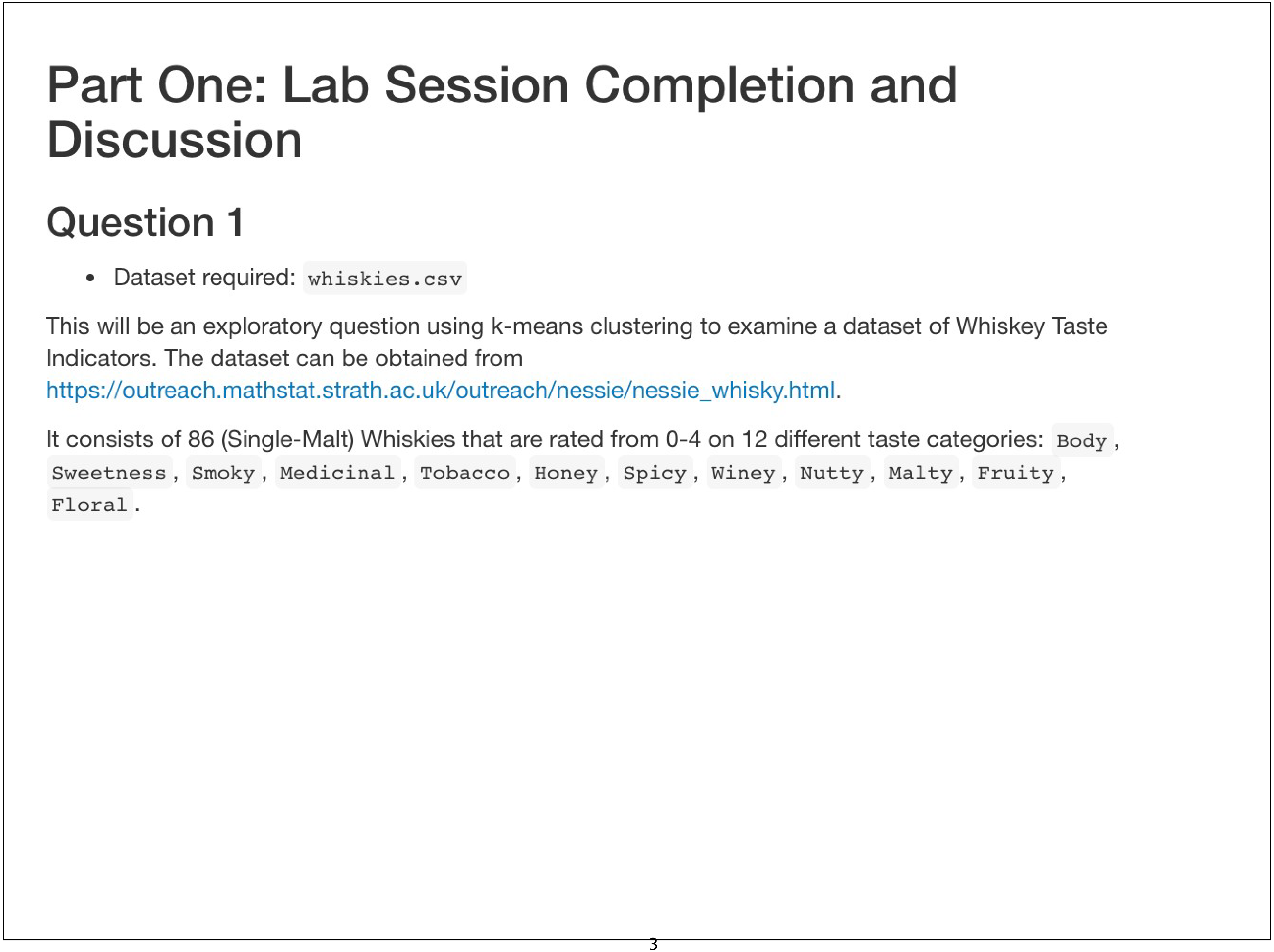
BT1101

Lab 8: Data Mining Basics

Installing and loading packages



```
# load required packages
library(dplyr)
library(tidyr)
library(car) # for linearHypothesis()
library(ggplot2) # optional. we expect you to know base graphics, but allow ggplot if you find it easier
library(psych) # for pairs.panels()
library(factoextra) # for fviz_cluster()
library(wooldridge)
```



Part One: Lab Session Completion and Discussion

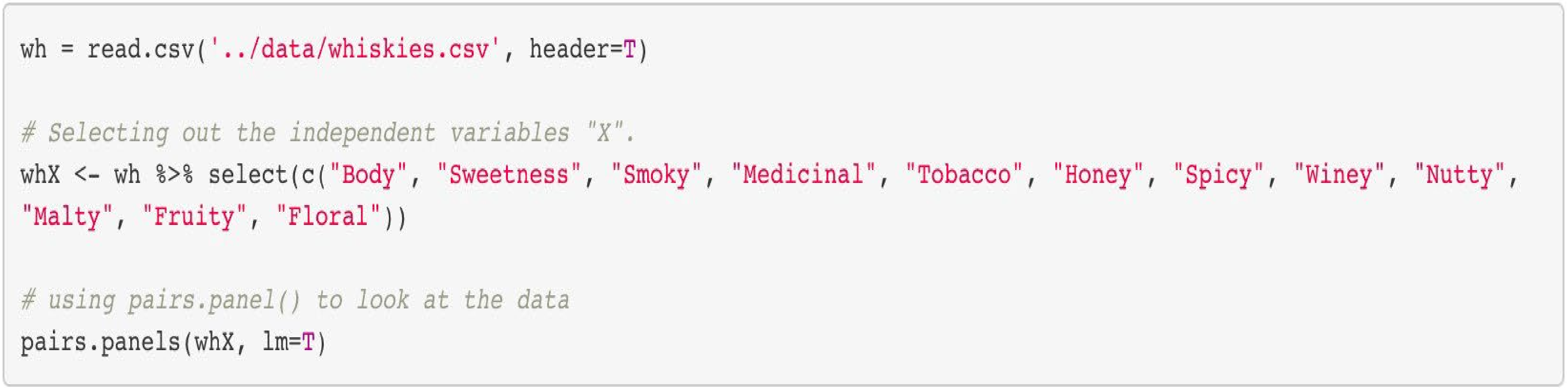
Question 1

- Dataset required: `whiskies.csv`

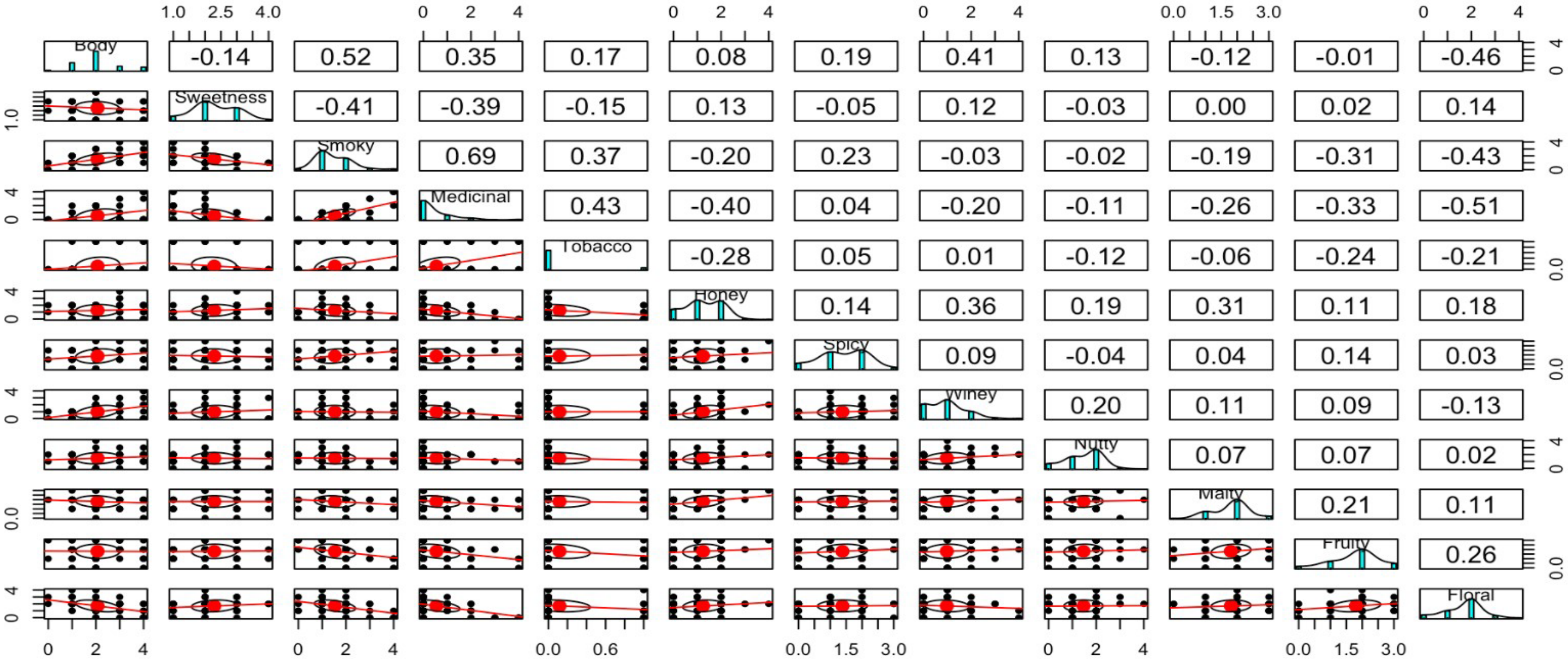
This will be an exploratory question using k-means clustering to examine a dataset of Whiskey Taste Indicators. The dataset can be obtained from https://outreach.mathstat.strath.ac.uk/outreach/nessie/nessie_whisky.html.

It consists of 86 (Single-Malt) Whiskies that are rated from 0-4 on 12 different taste categories: `Body`, `Sweetness`, `Smoky`, `Medicinal`, `Tobacco`, `Honey`, `Spicy`, `Winey`, `Nutty`, `Malty`, `Fruity`, `Floral`.

Data Exploration



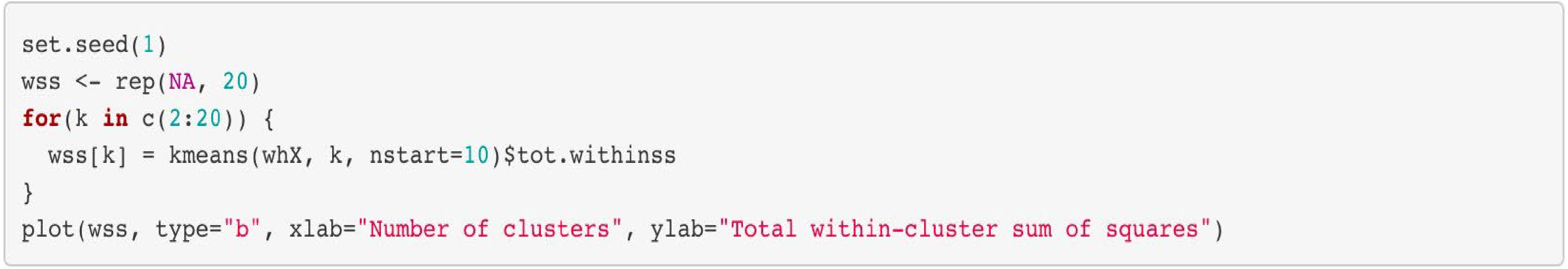
```
wh = read.csv('../data/whiskies.csv', header=T)

# Selecting out the independent variables "X".
whX <- wh %>% select(c("Body", "Sweetness", "Smoky", "Medicinal", "Tobacco", "Honey", "Spicy", "Winey", "Nutty",
"Malty", "Fruity", "Floral"))

# using pairs.panel() to look at the data
pairs.panels(whX, lm=T)
```



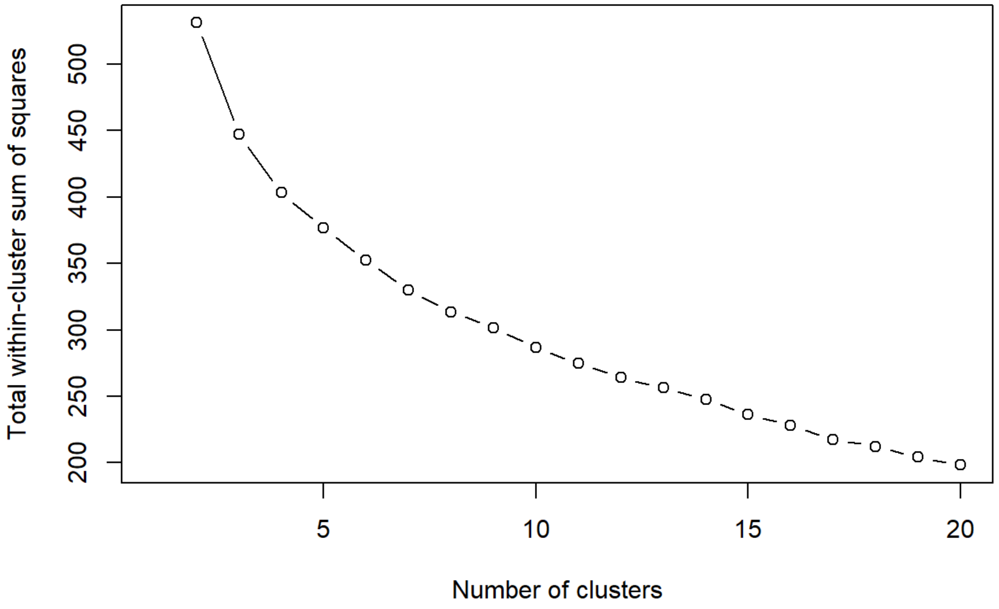
Q1(a)

(1a) Let’s try clustering the different whiskies based on their taste profile. First, let’s use the Elbow method to pick the best number of clusters.

Using the code discussed in lecture, calculate the Within-Cluster Sum of Squares from k=2 to k=20 clusters using whX, and plot the Within-Cluster Sum of Squares against number of clusters.



```r
set.seed(1)
wss <- rep(NA, 20)
for(k in c(2:20)) {
  wss[k] = kmeans(whX, k, nstart=10)$tot.withinss
}
plot(wss, type="b", xlab="Number of clusters", ylab="Total within-cluster sum of squares")
```



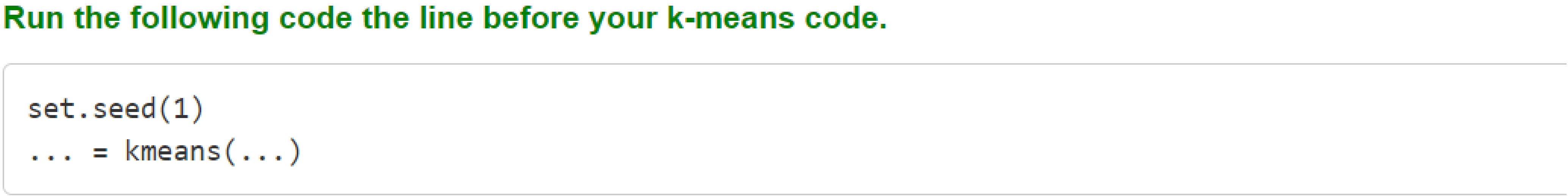
Note, normally if the variables are on very different scales, we should standardize the variables (to have mean 0 and sd 1)

In this case, all the variables are on the same scale (0-4). Hence it is ok to NOT scale the variables.

Q1(b)

(1b) let’s say our local business partner applies his expert intuition, and tells us that k=3 seems a good starting point.?

Use the fviz_cluster() function from the factoextra package to plot the results of this clustering.

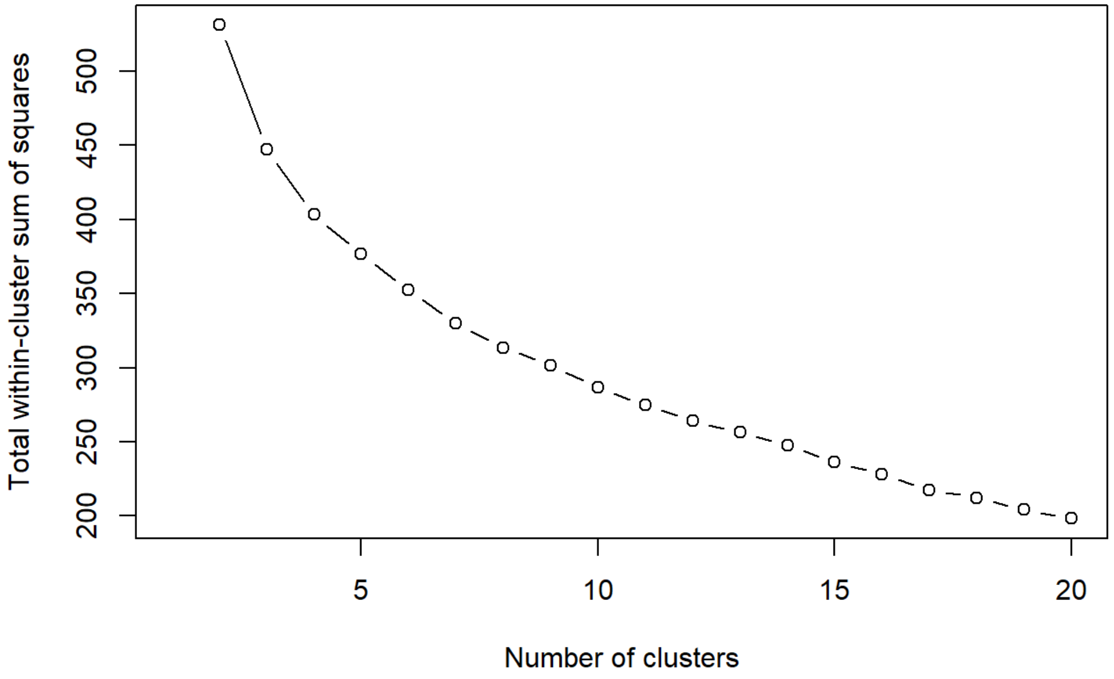


Run the following code the line before your k-means code.

```
set.seed(1)
... = kmeans(...)
```

What do you notice from the graph?

Discuss this with your TA and fellow students.



From the earlier plot, it does not seem like there is a clear Elbow. The Within-Cluster Sum of Squares seems to keep decreasing, and there doesn’t seem to be a clear stopping point.

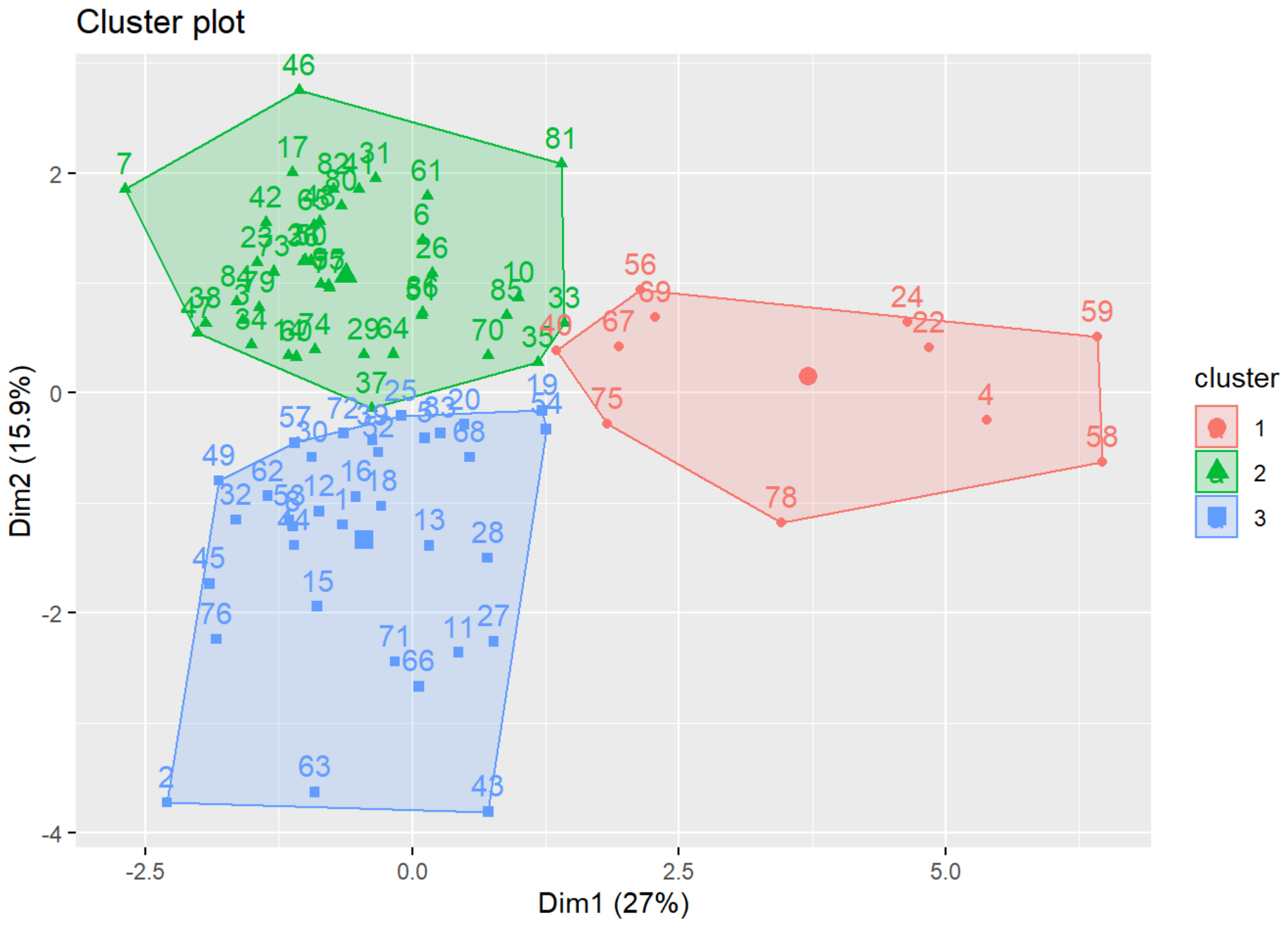
This may happen in real datasets so we may have to use our own judgment to decide on the number of clusters.

Q1(b)

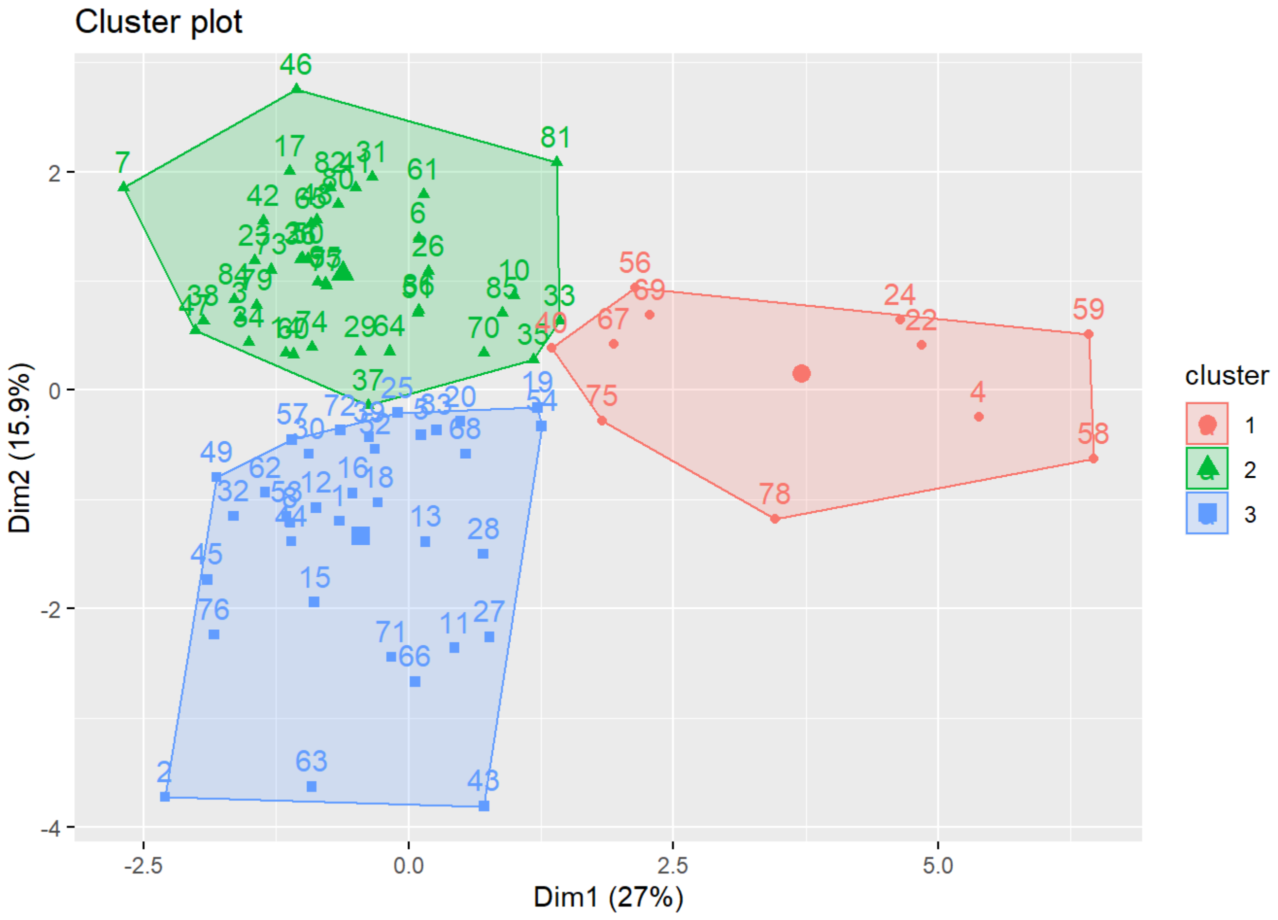
(1b) let’s say our local business partner applies his expert intuition, and tells us that k=3 seems a good starting point.?

Use the fviz_cluster() function from the factoextra package to plot the results of this clustering.

```
set.seed(1)
km_obj <- kmeans(whX, 3)
fviz_cluster(km_obj, whX)
```



Q1(b)



Discussion points

There seems to be three clearly separated clusters:

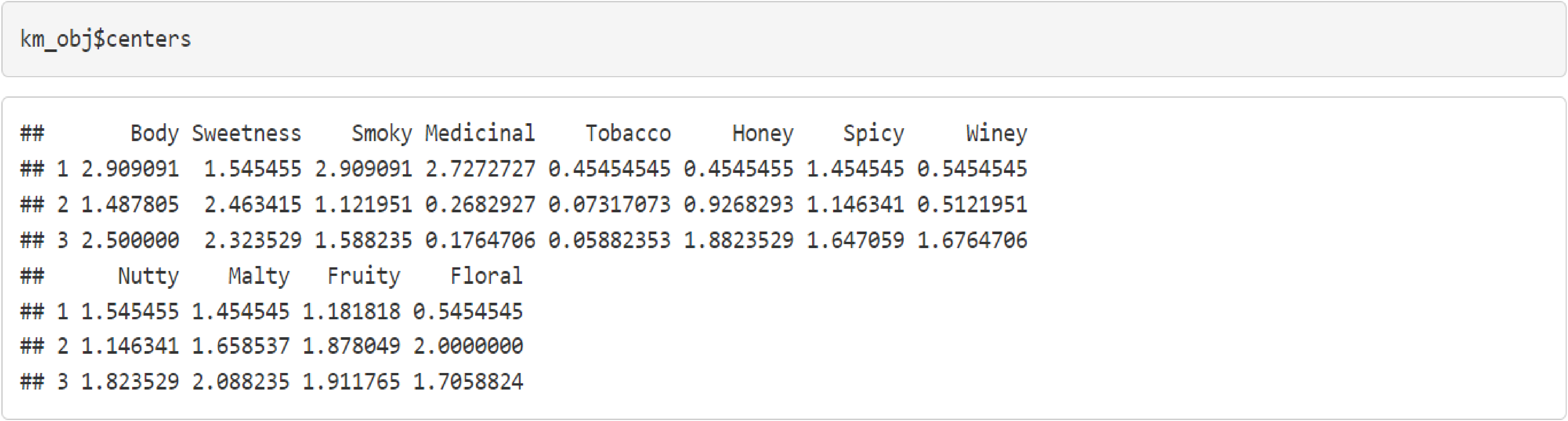
One (In the graph above, Cluster 1) that's much higher on PC1 than the rest (i.e., on the right of the graph).

The other two (Clusters 2 and 3) are on the left side of the graph, but they are separated by PC2, such that Cluster 2 is higher on PC2 and Cluster 3 is lower on PC2.

Q1(c)

Use <kmeans_object_name> $center (where <kmeans_object_name> is the name of the kmeans model you fit above) to extract the centers of the 3 clusters.

Try to interpret the clusters. If your client really likes very rich, Smoky, Medicinal Whiskies, which cluster would you recommend to him? (e.g., if this were a real client, you could go back and look at the Distilleries in wh and generate a list of those in the same cluster.)



```
km_obj$centers
```

```
##       Body Sweetness    Smoky Medicinal    Tobacco     Honey    Spicy     Winey
## 1 2.909091  1.545455 2.909091 2.7272727 0.45454545 0.4545455 1.454545 0.5454545
## 2 1.487805  2.463415 1.121951 0.2682927 0.07317073 0.9268293 1.146341 0.5121951
## 3 2.500000  2.323529 1.588235 0.1764706 0.05882353 1.8823529 1.647059 1.6764706
##      Nutty    Malty   Fruity    Floral
## 1 1.545455 1.454545 1.181818 0.5454545
## 2 1.146341 1.658537 1.878049 2.0000000
## 3 1.823529 2.088235 1.911765 1.7058824
```

Q1(c)

Discussion points

Cluster 1 will be fuller bodied, less sweet, more smoky, more medicinal, more tobaccoy, less honey, less fruity and less floral than the rest. This is probably what PC1 is picking up on (what we saw in Q1b).

Cluster 2 and 3 are relatively more similar to each other, but compared to Cluster 3, Cluster 2 has: less body, less honey, less "winey" and nutty tastes.