

## Exercise 3 – R in Practice

Info. Display of diagrams and code follows in the directories “code” and “diagrams”

### Part 1

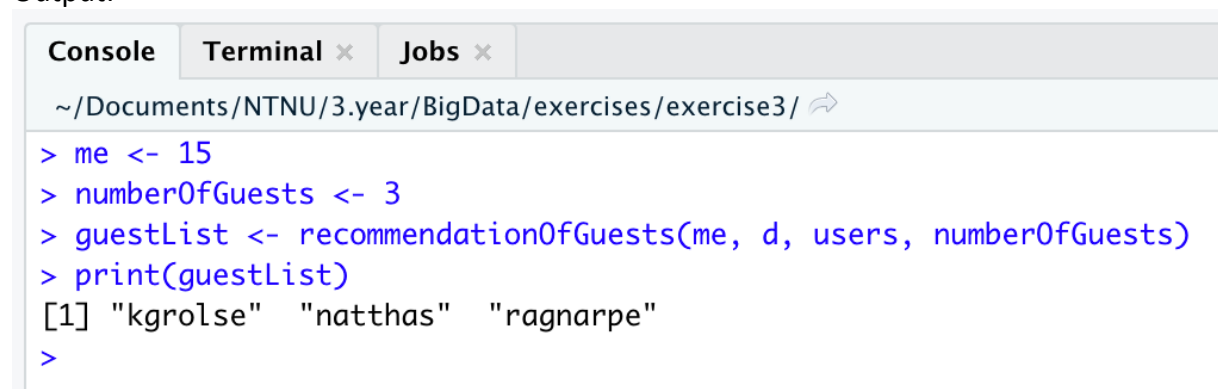
Aggregated the dishes using the aggregate function and plotted the result in a scatterplot diagram. The dishes average score follows along the x – axis and the dishes names follow the y – axis.

### Part 2

Used the cmdscale() function to plot the my neighborhood. Also made the function recommendationOfGuests(), who finds the k closest friends to anyone in the neighborhood based on Euclidian distance. The function returns the k closest friend’s usernames.

Example. I am user number 15 and want to find my three friends with most similar taste in food as myself.

Output:



```
Console Terminal x Jobs x
~/Documents/NTNU/3.year/BigData/exercises/exercise3/ ↗
> me <- 15
> numberOfGuests <- 3
> guestList <- recommendationOfGuests(me, d, users, numberOfGuests)
> print(guestList)
[1] "kgrolse" "natthas" "ragnarpe"
>
```

### Part 3

Both the classification tree and regression tree are applicable since the quality of the wine is based on integers and the range between the lowest quality, four, to the best quality, seven, is so small. I will try to explain this with an example.

A classification tree’s outcome of a decision is either true or false. Example we compare ten different types of wines quality. The wine types quality are either good or bad. Seven of the wine types are good and three wine types are bad. In a classification tree seven of the wines will be classified with good quality = true and three wine types will be classified with good quality = false.

A regression tree on the other hand has a numeric continuous outcome. First, we have to do a regression analysis of the wine qualities. The regression result will be some place between average and good. Then we have to compare this result with the ten wine types qualities. Seven of the wine types will have wine quality better than the regression result = true and three of the wine types will have wine qualities better than the regression result = false.

If the qualities ratings were more spread, example from one to twenty, we would have seen a bigger difference between the classification tree and the regression tree.

I choose to use the variables; total sulfur dioxide, density, pH value, sulphates and alcohol when I created a classification tree for finding the quality, because I believe that with more arguments than these, the tree becomes too large and difficult to read.

I choose to use  $cp \text{ value} = 0.0071$  because this value resulted in a readable classification tree with good depth. I used the function `plotcp()` and `printcp()` to experiment with the  $cp$  – value.