$c_{XY,\text{Pearson}} = c_{rg(X),rg(Y),\text{Pearson}}$

Beispiel:

i	1	2	3
x_i	5	10	-3
y_i	7	2	1
r_{x_i}	2	3	1
r_{y_i}	3	2	1

5, 10, -3
↓ sortieren
-3,

$c_{rg(X),rg(Y),\text{Pearson}} = \dfrac{(2-2)\cdot(3-2)+(3-2)\cdot(2-2)+(1-2)\cdot(1-2)}{s_{rg(X)}\cdot s_{rg(Y)}} =$

$\bar{x} \approx -25$

$\bar{y} \approx -5$

> 0

$c = \underbrace{(-40-(-25))}_{<0}\cdot\underbrace{(0-(-5))}_{} +$

Beispiel: x_i = Erste im Banner A in Jahr i

y_i = Erste im Banner B in Jahr i

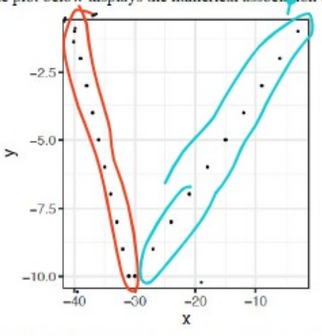
$(x_1,y_1), (x_2,y_2), (x_3,y_3), \ldots, (x_n,y_n)$

$x_i - \bar{x} < 0$, $y_i - \bar{y} \geq 0$

$x_i - \bar{x} > 0$, $y_i - \bar{y} \geq 0$

Problem 5 (2 credits)

The plot below displays the numerical association between x and y.



One of the following options i-v is correct for each correlation coefficient c:

i) c = -1
ii) -1 < c < 0
iii) c = 0
iv) 0 < c < 1
v) c = 1

a) Indicate the correct option for the Pearson correlation coefficient and justify your answer.
b) Indicate the correct option for the Spearman correlation coefficient and justify your answer.
(For each a) and b) 1 point for the correct answer with a correct justification.)

Pearson

$$c_{XY,\text{Pearson}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{s_x \cdot s_y}$$

a)

iv) $-1 \leq c \leq 1$

b) $0 < c < 1$

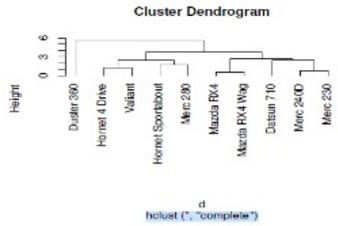
$c = 0$

Hierarchical clustering in R: dendograms

The results of hierarchical clustering can be shown using a dendrogram (i.e., a tree representation)

The height in the dendrogram at which two clusters are merged represents the distance between those two clusters.

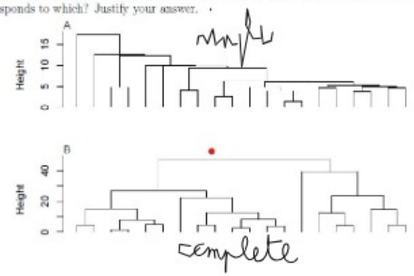
```
plot(hc, hang=-1) # hang=-1 align observation labels at the bottom of the dendogra
```

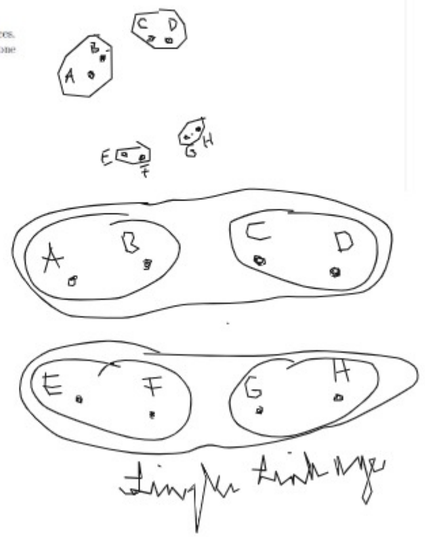


Observations that are determined to be similar by the clustering algorithm are displayed close to each other in the x-axis.

Problem 5 (1 credit)

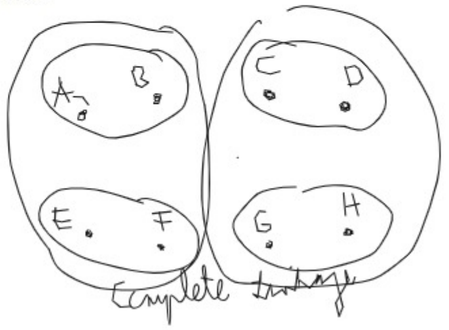
These are the dendrograms after performing hierarchical clustering with Euclidean distances. In one, "single" linkage method was used, while "complete" was used in the other. Which one corresponds to which? Justify your answer.

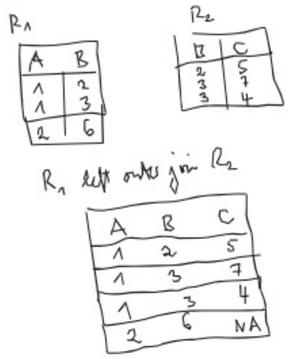


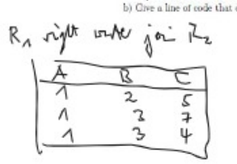


The *linkage rules* define dissimilarity between clusters. Here are four popular linkage rules assuming that A and B are two different clusters:

- **Complete:** The dissimilarity between A and B is the largest dissimilarity between any element of A and any element of B.
- **Single:** The dissimilarity between A and B is the smallest dissimilarity between any element of A and any element of B.
- **Average:** The dissimilarity between A and B is the average dissimilarity between any element of A and any element of B.
- **Centroid:** The dissimilarity between A and B is the dissimilarity between the centroids (mean vector) of A and B.





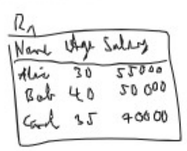


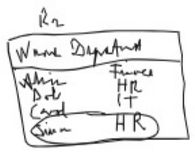
Problem 2 (2 credits)

You are given the following merged data table.

Name	Age	Department	Salary
Alice	30	Finance	55000
Bob	40	HR	50000
Susan	NA	HR	NA
Carol	35	IT	70000

a) Write down the two "minimal tables" that could have been used to generate these tables by a merge operation. These "minimal tables" should only contain the information required to lead to this merged table and should not contain any NA.
b) Give a line of code that creates the shown table from these two tables.





R1 right outer merge

b) merge(R1, R2, by="Name")

Merging by more than one column

We now merge dt1 and dt2 by first name and last name:

```
merge(dt1, dt2, by=c("firstname", "lastname"))
```

```
##    firstname lastname x y
## 1:     Alice     Coop 1 A
## 2:       Bob    Smith 3 C
```

Note that merging by first name only gives a different result (as expected):

```
merge(dt1, dt2, by="firstname")
```

```
##    firstname lastname.x x lastname.y y
## 1:     Alice       Coop 1       Coop A
## 2:     Alice      Smith 2       Coop A
## 3:       Bob      Smith 3     Marley B
## 4:       Bob      Smith 3      Smith C
```