Hospital Data Analysis - Clustering

I- Setup

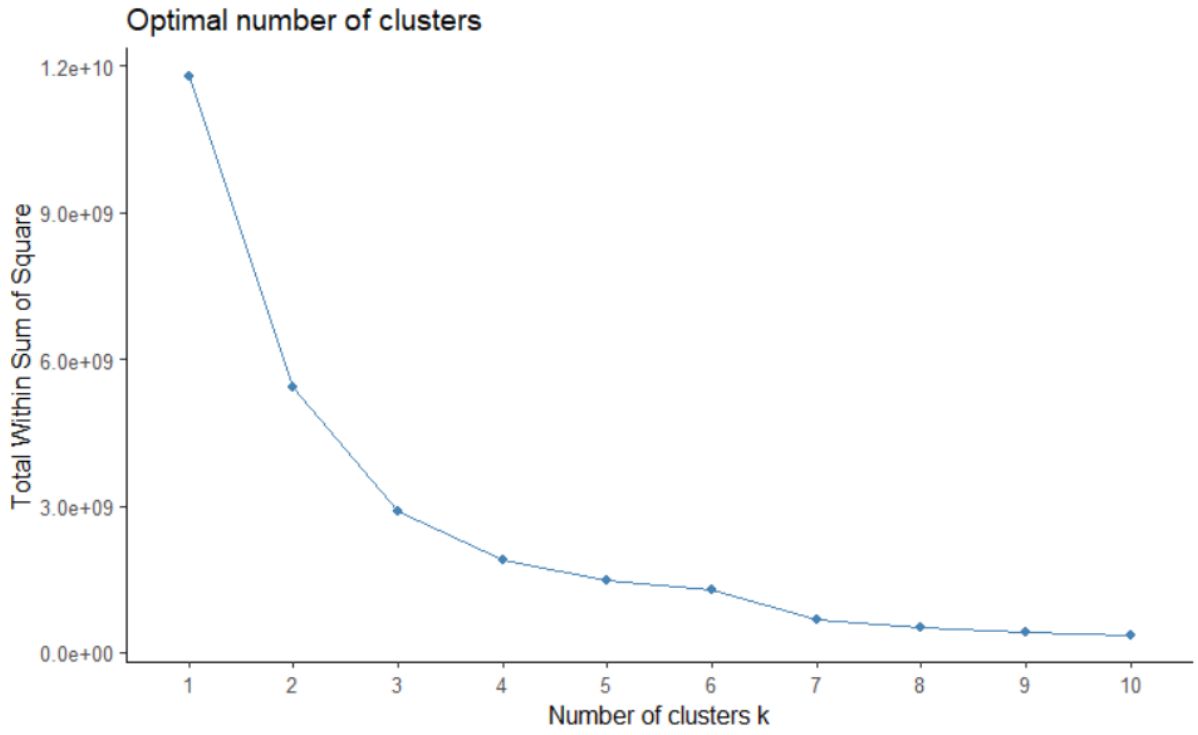
```
df <- df_origine %>%
  filter(AGE > 60)
```

- Select only observations for patients over 60 years old.

II- Model 1: Total costs, age and CCI

```
df_model1 <- df %>%
  select(TOTAL_COST, CCI, AGE)

fviz_nbclust(df_model1, kmeans, method = "wss") + theme_classic()
```



- The data partitioning method that we have chosen was that of the k-means (k-means clustering).
- The method used for estimating the optimal number of Clusters to specify when partitioning data with the k-means method is that of the WSS (Total Within cluster Sums of Squares).
- Based on the curve above, it is clear that we can observe the elbow *"Number of clusters k = 3"*, meaning that 3 is then the optimal number of Clusters to retain.

```
mod1 <- kmeans(df_model1,
               centers = 3,
               nstart = 10,
               iter.max = 200,
)

mod1
```

```
K-means clustering with 3 clusters of sizes 62, 1786, 678

Cluster means:
  TOTAL_COST      CCI      AGE
1  12101.328 2.903226 77.93548
2   2025.263 2.593505 83.52352
3   4739.181 2.507375 81.53687

Clustering vector:
   [1] 2 3 3 2 3 2 2 3 2 2 3 3 2 3 2 1 3 2 2 2 3 3 2 3 3 2 3 2 3 3 2 3 3 3 3 3 3 3 3 3 3 2 3 3 1 2
  [48] 3 2 2 1 1 2 3 2 2 3 3 3 2 2 2 2 2 2 2 2 2 3 2 2 3 3 2 2 3 2 3 2 2 2 2 2 2 3 2 2 2 2 2 2 3 2
  [95] 2 2 2 2 2 2 2 2 3 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 3 2 2 2 2 3 2 3 2 2 2 2 2 3
 [142] 2 2 3 2 2 2 2 3 3 3 2 3 2 2 2 2 2 2 2 2 2 1 2 3 2 2 3 2 2 3 3 3 2 3 2 2 3 3 2 2 2 2 3 2 3 3 3
 [189] 3 2 3 2 2 3 3 2 2 2 2 2 2 2 3 2 2 3 2 3 3 3 3 3 3 3 2 2 2 3 3 3 1 2 2 3 2 2 3 3 3 2 2 3 2 2 2 1
 [236] 2 3 1 2 3 2 3 3 1 2 2 2 2 2 2 3 2 2 3 2 2 2 3 2 2 2 2 2 2 2 2 2 3 2 3 3 2 2 2 2 3 2 2 2 2 3 3
 [283] 1 2 1 3 3 1 2 2 2 2 2 2 2 3 3 2 3 3 3 2 3 3 2 3 2 3 1 3 3 3 2 3 1 2 2 2 2 2 2 2 3 2 2 2 3 2 2
 [330] 3 3 2 2 3 2 2 2 1 3 2 3 3 2 3 3 3 3 3 2 2 2 2 1 2 2 2 2 2 2 2 2 3 2 2 1 1 1 2 2 3 2 2 3 1 1 3 3
 [377] 3 3 2 3 3 3 1 1 2 2 2 2 2 2 2 2 3 3 3 2 1 3 3 3 2 2 3 1 1 2 2 2 2 2 3 3 2 3 3 3 2 2 2 2 2 2 3 2 2 3
 [424] 3 2 2 3 2 3 3 3 3 2 2 2 2 3 3 3 3 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 3 3 2 3 3 2 2 3 3 2 2 3
 [471] 2 2 2 2 2 2 3 2 2 2 3 2 2 3 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 3 2 3 2 2 2 2 2 3 2 2 2 3 3 2
 [518] 2 3 3 3 2 2 2 2 1 2 3 2 3 1 2 3 3 2 2 3 2 3 3 3 2 3 3 3 3 3 2 2 3 3 3 3 2 3 2 3 2 2 3 3 3 3 3
 [565] 1 1 3 2 3 1 1 2 2 3 3 3 2 3 3 2 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 2 2 2 2 3 3 1 3 1 3 1 2 2 2 3
 [612] 2 3 3 2 2 3 3 3 3 3 3 3 3 2 3 3 1 3 3 1 1 3 3 3 2 1 1 2 2 2 2 3 3 2 3 2 2 2 3 2 3 3 2 2 2 2 2
 [659] 2 3 2 2 2 2 2 2 2 3 2 2 2 3 2 3 2 3 2 2 2 2 3 2 2 3 2 2 3 3 3 3 3 2 2 2 3 2 2 3 3 2 3 3 2 3 2
 [706] 3 3 3 2 2 2 2 2 3 2 2 2 2 2 2 2 2 3 2 2 1 2 2 2 2 3 2 2 2 2 2 3 3 3 2 2 3 2 2 3 2 2 2 2 3 2 3
 [753] 2 2 3 3 2 3 2 2 3 2 2 2 2 3 2 2 2 2 3 2 2 2 2 2 2 3 2 3 3 2 3 2 2 2 3 2 3 2 3 2 3 2 2 2 3 2 2 3
 [800] 2 2 2 2 3 2 2 2 3 2 2 2 2 2 2 2 2 3 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 2 2 2 2 2 3
 [847] 2 2 2 2 2 2 3 2 2 2 2 2 3 2 3 2 2 3 2 3 3 3 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 3 2 3 3 3 3 2 3 2 2 2
 [894] 2 3 2 2 2 2 2 3 3 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 3 2 3 2 2 2 2 2 2 2 2 2 3 3 2 2 2 2 2
 [941] 2 3 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 3 2 2 3 2 3 3 2 2 2 2 2 3 2 2 3 2
 [988] 3 2 2 3 2 2 3 2 2 3 3 2 2
 [ reached getOption("max.print") -- omitted 1526 entries ]

Within cluster sum of squares by cluster:
[1]  962615839 1032147453  904193493
 (between_SS / total_SS =  75.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

- The number of observations for each of the 3 clusters will be respectively: 62, 1786 and 678 observations.
- The **coordinates (TOTAL_COST, CCI, AGE)** of each Centroid (position of the Cluster's center) of each Cluster are as follows:
  - **Centroid of Cluster 1 : (12101.328, 2.903226, 77.93548) ;**
  - **Centroid of Cluster 2 : (2025.263, 2.593505, 83.52352) ;**
  - **Centroid of Cluster 3 : (4739.181, 2.507375, 81.53687).**
- Based on the Clustering Vector above, we can consult the Cluster of membership of each line of observations on admissions by just consulting the number of cluster which represents the line of observations on the position of the latter within the vector: *line 1 belongs to Cluster 2, line 2 belongs to Cluster 3, line 988 belongs to Cluster 3…*
- The Sum of the squares of the distances of the points of a Cluster to their Centroid for each of the Clusters 1, 2 and 3 are respectively: 962 615 839, 1 032 147 453 and 904 193 493. From these values, we can deduce that Cluster 3 is the most compact of all, followed by Cluster 1 and, finally, Cluster 2 is the least compact of all.

- The ratio **"between_SS / total_SS = 75.4"** can tell us that we have a relatively good Clustering here.

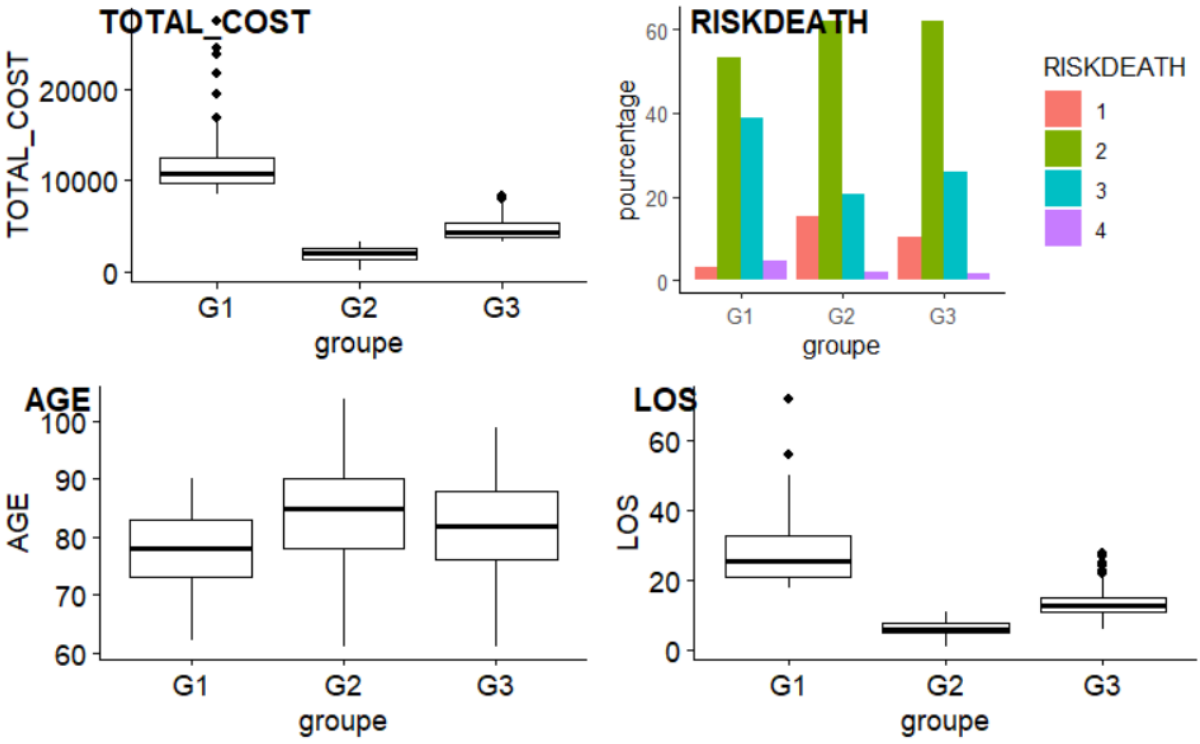
Residuals:

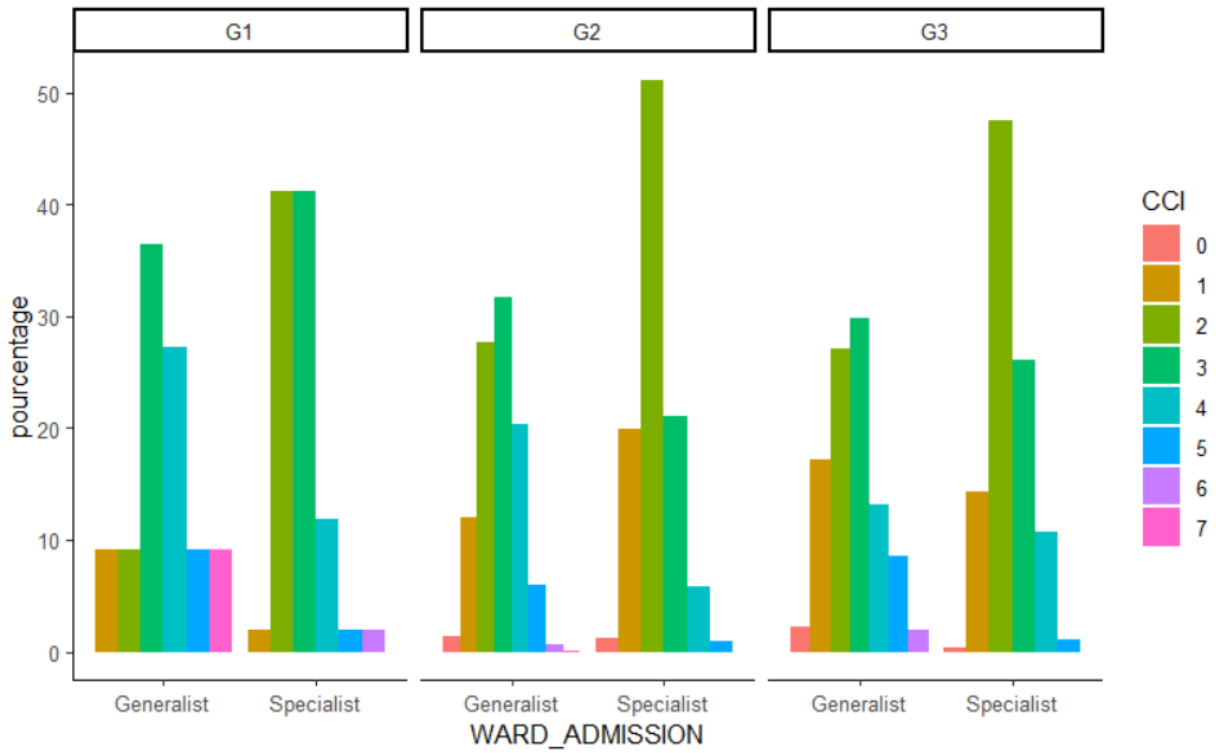
```
res.mod1 <- df %>%
  mutate(groupe= paste0('G',mod1$cluster))
```

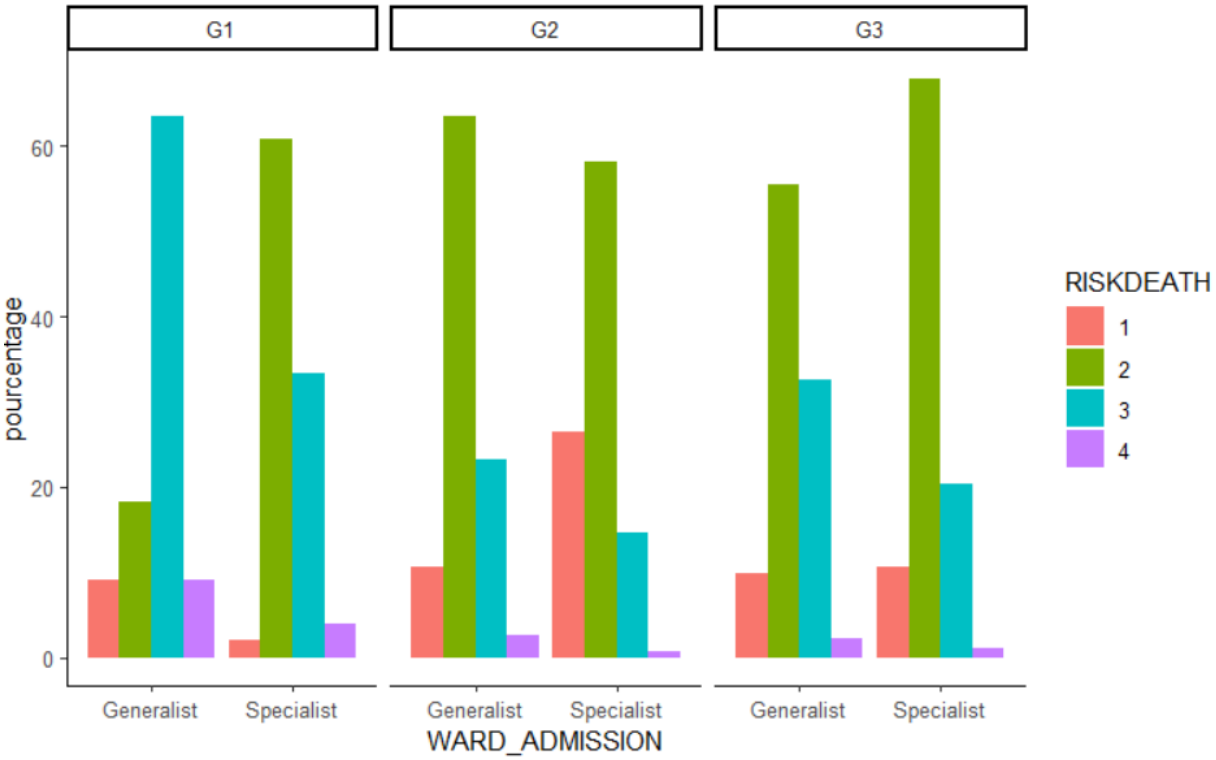
- We have named our Clusters 1, 2 and 3 respectively by « G1 », « G2 » and « G3 ».

groupe	variables	mean	sd	min	q1	median	q3	max	na
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
G1	AGE	77.935484	7.160619	62.00	73.000	78.00	83.000	90.00	0
G1	CCI	2.903226	1.082039	1.00	2.000	3.00	3.000	7.00	0
G1	TOTAL_COST	12101.327742	3972.472162	8498.98	9804.040	10816.94	12534.075	27400.28	0
G2	AGE	83.523516	8.308732	61.00	78.000	85.00	90.000	104.00	0
G2	CCI	2.593505	1.127424	0.00	2.000	2.00	3.000	7.00	0
G2	TOTAL_COST	2025.262906	760.370674	285.00	1496.487	2057.13	2623.028	3381.37	0
G3	AGE	81.536873	8.499246	61.00	76.000	82.00	88.000	99.00	0
G3	CCI	2.507375	1.118835	0.00	2.000	2.00	3.000	6.00	0
G3	TOTAL_COST	4739.181445	1155.644897	3385.74	3817.535	4374.00	5468.855	8397.59	0

9 rows







- <u>We can notice that for the case of the TOTAL_COST :</u>
  - The highest values are found within G1, with also the greatest variability of data characterized by an asymmetric data distribution;

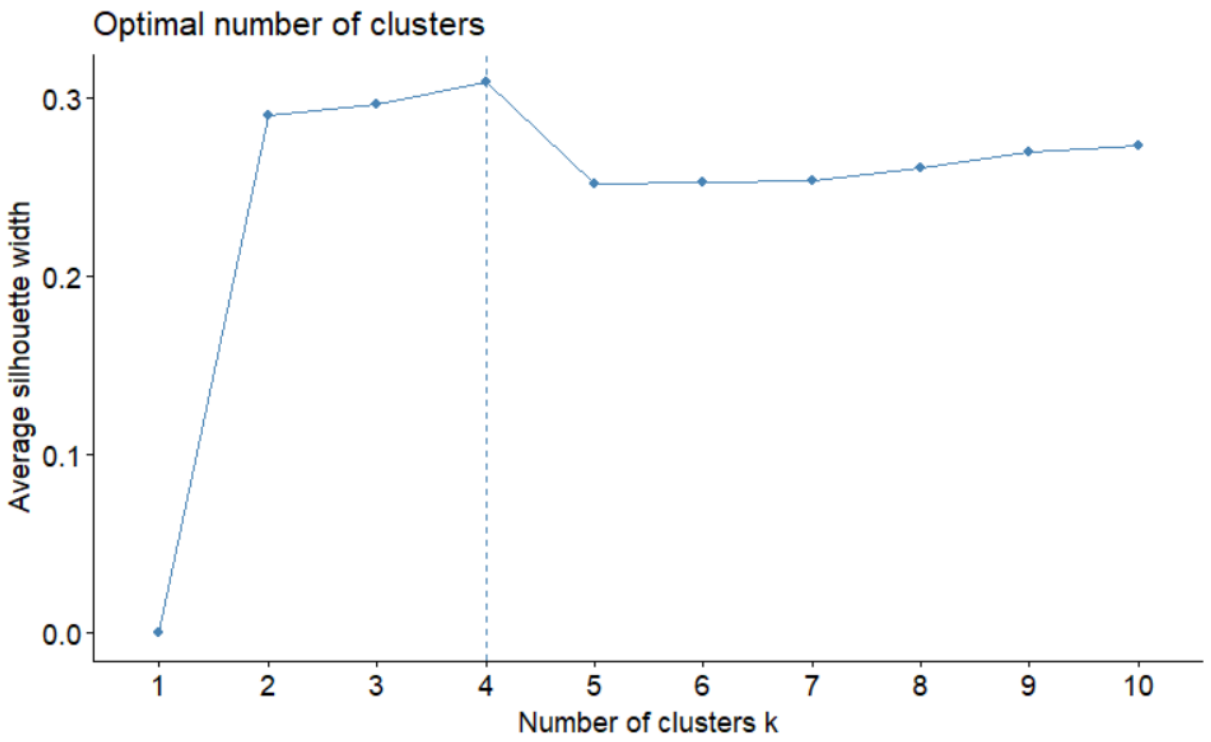
- The lowest values are clearly found within G2, with also the smallest of the data variability;
- The group G3 corresponds to higher values than those observed in G2, with a slightly greater data variability and a more or less asymmetrical data distribution;
- Outliers are observed the most within G1 and much less in G3. They are literally missing in G2.

- For the case of the RISK_DEATH :
  - The peaks of proportions all correspond to the admissions with a RISK_DEATH level of 2 within the 3 groups, peaks that are more important in G2 and G3 than in G1;
  - The RISK_DEATH level 3 is more frequent for the case of admissions in G1, it is slightly less for those in G3, and much less for those in G2;
  - The RISK_DEATH level 1 is more frequent for the case of admissions in G2, slightly less for those in G3 and relatively rare for those in G1;
  - The RISK_DEATH level 4 remains relatively rare in the three groups, however, it is mainly found in G1.
- For the case of the AGE:
  - The admissions corresponding to the oldest ages concern the most the group G2, then G3 and slightly less G1;
  - The variability is quite significant in each group, and the data distribution is relatively symmetrical in each group.
- For the case of the LOS:
  - The highest values of LOS are (very) mainly found within G1, whereas the lowest ones are found in G2;
  - The variability is largely less significant in G2 and G3, but more significant in G1;
  - Outliers are found within G1 and G3.
- For the case of the CCI per WARD_ADMISSION:
  - We can notice that the proportions of admissions with a CCI 2 constitute the peaks for the case of all the *Specialist Wards* of each group. The peaks are more significant in G2 (around 50%) and G3 than in G1;
  - The peaks, for the case of the *Generalist Wards,* are made up of the proportions of admissions with a CCI 3, more significant in G1, then in G2 and finally less significant in G3;
  - CCIs 7 are particularly frequent for the case of admissions in *G1:Generalist Wards*, and rare, even non-existent within the other Wards of any Group;
  - CCIs 4 are always more or less frequent for the case of admissions, both in *Specialist and Generalist Wards,* in all of the Groups (more frequent in *G1: Generalist Wards*)
- For the case of the RISK_DEATH per WARD_ADMISSION:
  - Except for the particular case of *G1:Generalist Wards*, whose peak of proportions in terms of admissions corresponds (obviously) to the RISK_DEATH level 3, all the peaks of proportions correspond to the Level 2 in all the other Wards of any group (a level, however, less observed for the case of admissions in *G1:Generalist Wards*) ;
  - Except for the case of *G1: Specialist Wards*, the level 1 seems to be similarly frequent in the *Generalist Wards* of each group. For the case of the *Specialist Wards,* this level 1 is particularly more frequent in G2, and less frequent in G3 and G1;

- By performing our reading from G1 to G3, we can notice that the frequencies of the level 4 are decreasing in the *Generalist Wards,* with just smaller proportions than in the *Generalist Wards*.

III- Model 2: Total costs, age and CCI – normal distribution

```
df_model2 <- df %>%
  select(TOTAL_COST, CCI, AGE) %>%
  scale()
fviz_nbclust(df_model2, kmeans, method = "silhouette")
```



- The data partitioning method that we have chosen is that of the k-means (k-means clustering).
- The method used for estimating the optimal number of Clusters to specify when partitioning data with the k-means method is that of the Average Silhouette Width.
- Based on the curve above, it is clear that we have a maximum value of the Average Silhouette width with 4 clusters, which means that 4 is then the optimal number of Clusters to be kept.

```
mod2_nc <- 4

mod2 <- kmeans(df_model2,
               centers = mod2_nc,
               nstart = 10,
               iter.max = 200,
)

mod2
```

```
K-means clustering with 4 clusters of sizes 974, 674, 745, 133


Cluster means:
   TOTAL_COST        CCI        AGE
1 -0.19878202  0.9241646  0.5033786
2 -0.04451713 -0.4516112 -1.2275418
3 -0.23052149 -0.8272487  0.5551044
4  2.97260711  0.1545108 -0.5750399

Clustering vector:
   [1] 2 3 3 1 2 2 3 4 2 2 2 3 3 1 3 4 3 2 2 3 2 3 2 1 3 2 4 3 2 2 3 2 4 2 4 2 3 2 2 3 2 3 1 2 2 4 2
  [48] 2 2 1 4 4 3 1 3 3 3 2 2 2 2 2 2 3 3 1 2 1 3 3 3 2 1 2 3 2 3 3 2 2 3 2 2 3 3 3 3 3 1 3 2 3 3 2
  [95] 3 3 2 1 3 3 2 2 2 3 3 2 3 3 3 3 2 3 2 3 3 2 3 3 2 2 2 3 3 2 3 3 3 2 2 2 3 2 1 2 1 3 3 2 3 2 4
 [142] 2 2 2 2 2 2 2 1 3 3 2 2 2 2 3 3 3 3 2 3 3 4 3 2 3 2 4 2 3 3 1 2 3 2 3 3 2 3 2 3 3 3 2 1 2 2 3
 [189] 1 3 3 3 3 3 3 2 2 3 3 2 3 3 3 2 2 3 4 1 1 2 1 4 1 1 2 2 1 2 2 2 4 2 1 2 2 2 3 2 3 3 2 3 2 2 4
 [236] 3 3 4 3 2 2 3 3 4 3 3 2 3 3 3 1 2 3 1 3 3 3 2 1 3 2 2 3 2 3 2 1 1 2 2 3 3 3 2 3 3 3 3 3 3 2 3
 [283] 4 3 4 1 1 4 2 1 3 3 2 3 2 3 4 2 2 2 2 2 4 2 1 2 1 2 4 1 4 1 1 1 4 1 2 2 2 3 2 3 1 3 2 1 3 2 3
 [330] 1 2 3 2 1 1 1 1 4 3 3 1 2 3 2 2 2 4 3 1 2 3 1 4 1 2 1 1 2 1 1 1 3 1 4 4 4 2 1 3 1 1 1 4 4 4 4
 [377] 4 2 1 3 2 4 4 2 1 1 3 3 3 1 2 1 4 2 2 4 2 1 1 1 2 1 4 4 2 1 1 2 4 1 1 2 1 3 1 1 3 1 1 1 1 2 3
 [424] 4 1 1 1 1 1 2 1 1 2 1 3 3 1 1 2 2 3 3 1 3 1 1 1 3 3 2 2 3 2 2 1 1 2 3 3 4 1 3 2 3 3 1 1 1 1 4
 [471] 3 3 3 2 3 1 3 3 3 3 3 2 2 1 1 1 2 2 1 2 3 1 1 1 1 2 1 1 3 1 2 3 1 3 2 1 1 1 1 2 1 1 2 3 1 3 1
 [518] 3 3 3 1 2 2 1 1 4 2 1 2 2 4 3 4 1 3 3 1 1 2 3 2 3 2 4 2 4 2 1 3 2 2 4 3 2 4 1 2 1 2 4 2 2 3 2
 [565] 4 4 2 2 2 2 4 4 2 2 3 1 4 3 3 2 2 3 2 2 1 2 1 2 4 4 2 2 1 1 2 4 1 3 1 2 2 1 1 4 2 4 4 4 1 2 3 1
 [612] 1 2 4 3 2 4 2 4 2 2 3 2 2 2 2 4 2 4 4 4 4 4 4 2 2 2 4 4 2 2 3 3 3 3 3 1 2 2 3 1 1 3 3 2 1 3 3
 [659] 1 1 1 1 1 2 3 1 3 2 1 1 1 3 1 3 3 1 2 2 3 2 2 3 1 1 3 1 1 3 2 3 4 1 2 2 1 2 1 1 1 2 2 3 1 1 3
 [706] 1 2 3 2 3 1 1 1 1 1 1 3 1 1 3 3 3 1 2 1 4 1 3 3 1 3 3 1 1 2 1 3 3 3 1 2 1 1 1 3 2 1 3 1 1 2 3
 [753] 3 1 1 3 1 2 1 2 4 3 3 1 1 2 2 3 1 1 4 3 1 1 2 1 2 2 1 4 1 1 1 1 3 1 2 1 3 1 3 1 3 3 4 1 1 2
 [800] 1 3 3 3 1 3 1 1 2 1 3 3 1 2 2 3 2 3 2 3 4 2 3 1 1 2 3 3 1 3 1 3 1 1 1 1 2 3 2 1 1 1 2 1 3 1 3
 [847] 2 1 3 1 1 2 2 1 3 1 3 3 2 1 3 2 3 3 3 3 2 3 1 1 3 1 3 1 2 1 3 2 1 1 3 2 1 2 2 2 1 1 3 1 2 1 1
 [894] 1 2 1 3 2 1 3 2 1 2 3 3 2 1 1 1 1 3 3 3 1 4 3 1 2 1 3 2 2 1 3 1 3 1 2 3 1 1 1 2 2 4 2 3 1 1 3
 [941] 1 1 2 3 1 2 1 3 1 3 1 3 3 1 2 3 1 1 3 2 1 1 3 1 1 2 3 2 1 1 1 1 1 1 1 1 3 3 1 2 1 1 1 2 2 2 1
 [988] 1 2 1 2 3 3 4 3 2 1 4 1 2
 [ reached getOption("max.print") -- omitted 1526 entries ]

Within cluster sum of squares by cluster:
[1] 1121.6722  892.7428  710.4988  577.1178
 (between SS / total SS =  56.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

- The number for observations for each of the 4 clusters will be respectively: 974, 674, 745 and de 133 observations;
- The **coordinates (TOTAL_COST, CCI, AGE)** of each Centroid (position of the Cluster's center) of each Cluster will be as follows:
  - **Centroid of Cluster 1 : (-0.19878202, 0.9241646, 0.5033786) ;**
  - **Centroid of Cluster 2 : (-0.04451713, -0.4516112, -1.2275418) ;**
  - **Centroid of Cluster 3 : (-0.23052149, -0.8272487, 0.5551044) ;**
  - **Centroid of Cluster 4 : (2.97260711, 0.1545108, -0.5750399).**
- From the Clustering Vector above, we can consult the Cluster of membership of each line of observations on admissions by just looking at the Cluster number which represents the line of observations on the position of this latter within the vector: *line 1 belongs to Cluster 2, line 2 belongs to Cluster 3, line 988 belongs to Cluster 1…*

- The Sum of the squares of the distances of the points of a Cluster to their Centroid for each of the Clusters 1, 2, 3 and 4 are respectively: 1 121.6722, 892.7428, 710.4988 and 577.1178. From these values, we can deduce that Cluster 1 is the least compact of all, followed by Cluster 2, then Cluster 3 and, finally, Cluster 4 is the most compact of all.
- However, let us notice that the ratio ***"between_SS / total_SS = 56.4"*** is not necessarily a good sign that we have a good enough Clustering here.

Residuals :

```
res.mod2 <- df %>%
  mutate(groupe= paste0('G',mod2$cluster))
```

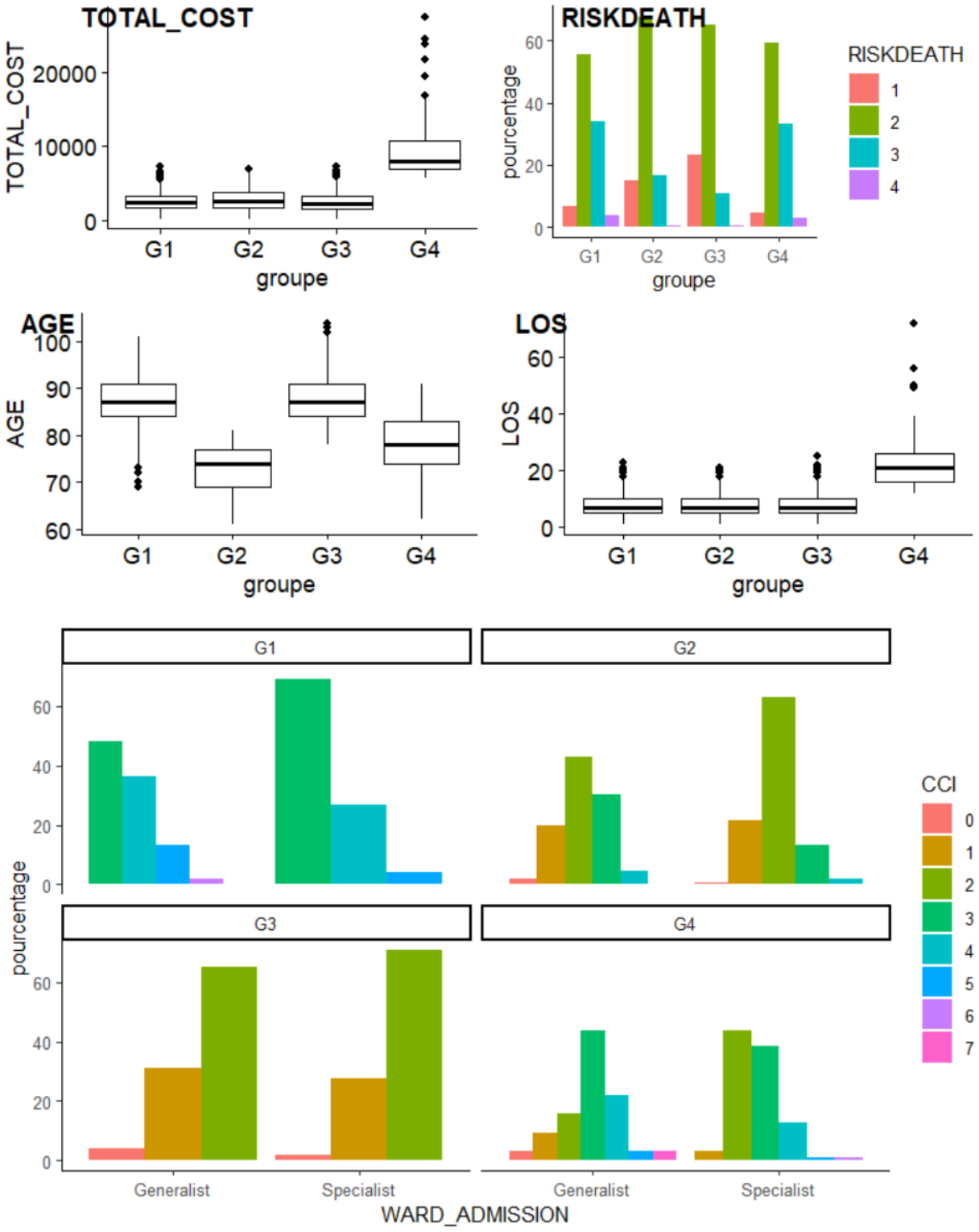
- We have named our Clusters 1, 2, 3 and 4 respectively by « G1 », « G2 », « G3 » and « G4 ».

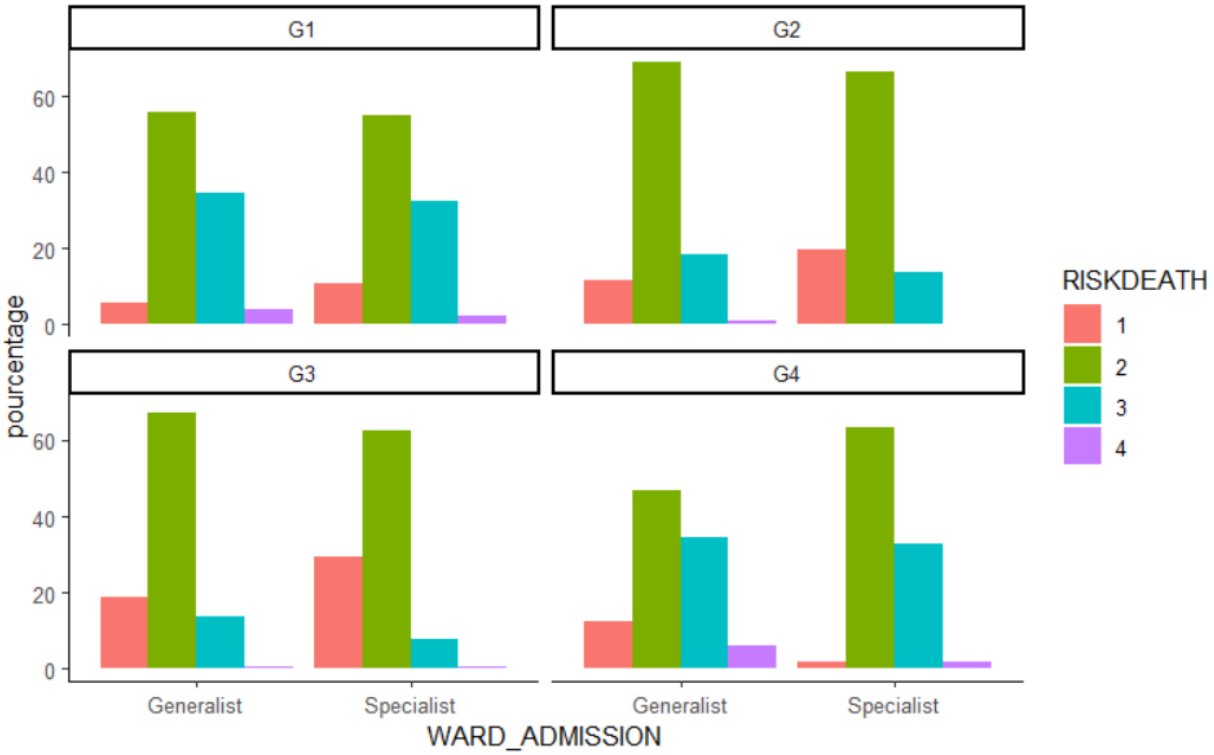
groupe <chr>	variables <chr>	mean <dbl>	sd <dbl>	min <dbl>	q1 <dbl>	median <dbl>	q3 <dbl>	max <dbl>	na <int>
G1	AGE	87.088296	5.3022137	69.00	84.000	87.000	91.000	101.00	0
G1	CCI	3.618070	0.7531516	3.00	3.000	3.000	4.000	7.00	0
G1	TOTAL_COST	2571.619025	1198.4282809	285.00	1734.633	2377.035	3242.677	7344.56	0
G2	AGE	72.525223	5.3094537	61.00	69.000	74.000	77.000	81.00	0
G2	CCI	2.069733	0.7972115	0.00	2.000	2.000	3.000	5.00	0
G2	TOTAL_COST	2904.852908	1410.7090223	285.00	1815.102	2667.760	3850.835	6943.40	0
G3	AGE	87.523490	4.8719958	78.00	84.000	87.000	91.000	104.00	0
G3	CCI	1.646980	0.5365107	0.00	1.000	2.000	2.000	2.00	0
G3	TOTAL_COST	2503.057289	1353.1357352	285.00	1532.640	2298.820	3299.990	7316.94	0
G4	AGE	78.015038	6.7554523	62.00	74.000	78.000	83.000	91.00	0

1-10 of 12 rows Previous 1 2 Next

groupe <chr>	variables <chr>	mean <dbl>	sd <dbl>	min <dbl>	q1 <dbl>	median <dbl>	q3 <dbl>	max <dbl>	na <int>
G4	CCI	2.751880	0.9801516	0.00	2.000	3.000	3.000	7.00	0
G4	TOTAL_COST	9422.266316	3722.0386671	5825.12	6992.140	8073.530	10793.010	27400.28	0

11-12 of 12 rows Previous 1 2 Next


TOTAL_COST
TOTAL_COST
20000
10000
0
G1
G2
G3
G4
groupe
RISKDEATH
pourcentage
60
40
20
0
G1
G2
G3
G4
groupe
RISKDEATH
1
2
3
4
AGE
AGE
100
90
80
70
60
G1
G2
G3
G4
groupe
LOS
LOS
60
40
20
0
G1
G2
G3
G4
groupe
G1
G2
G3
G4
pourcentage
60
40
20
0
Generalist
Specialist
Generalist
Specialist
WARD_ADMISSION
CCI
0
1
2
3
4
5
6
7



- We can notice that for the case of the TOTAL_COST :
  - The highest values are only found within G4, in a group where the variability is particularly high and where the data distribution is clearly asymmetrical;
  - The BoxPlots that correspond to the groups G1, G2 and G3 are more or less on the same level of TOTAL_COST, with a less significant variability compared to that of G4, and the data distribution within those groups are relatively symmetrical;
  - Outliers are more significant and higher in terms of TOTAL_COST values for the case of G4, while they are less high and less dispersed in a similar way for the cases of G1 and G2. However, they are less numerous and less high in terms of TOTAL_COST values.
- We can notice that for the case of the RISK_DEATH :
  - The peak of proportions in each group strongly corresponds (more than 50 % each), to the admissions with a RISK_DEATH level 2. In a descending order, the peak is higher in G2, then in G3, then in G4 and slightly less high than in all the other groups in G1;
  - The proportions of admission with a RISK_DEATH level 3 are relatively similar in G1 and G4, while they are similarly less high in G2 and G3;
  - Admissions with a RISK_DEATH level 1 are more frequent in G3, then in G2, whereas they are relatively less frequent in G1 and G4;
  - In all of the groups, admissions with a RISK_DEATH level 4 are less frequent, even relatively rare for the particular cases of the groups G2 and G3;
- We can notice for the case of the AGE :
  - Admissions related to the oldest patients are observed within G1 and G3, whereas those which are related to the less old ones are mainly more concentrated in G2 than in G4;
  - Variability is more significant in G2 than in G4, with a noticeable distribution asymmetry in G2. However, this variability is clearly and similarly less significant in G1 and G3, but with more symmetrical distributions;

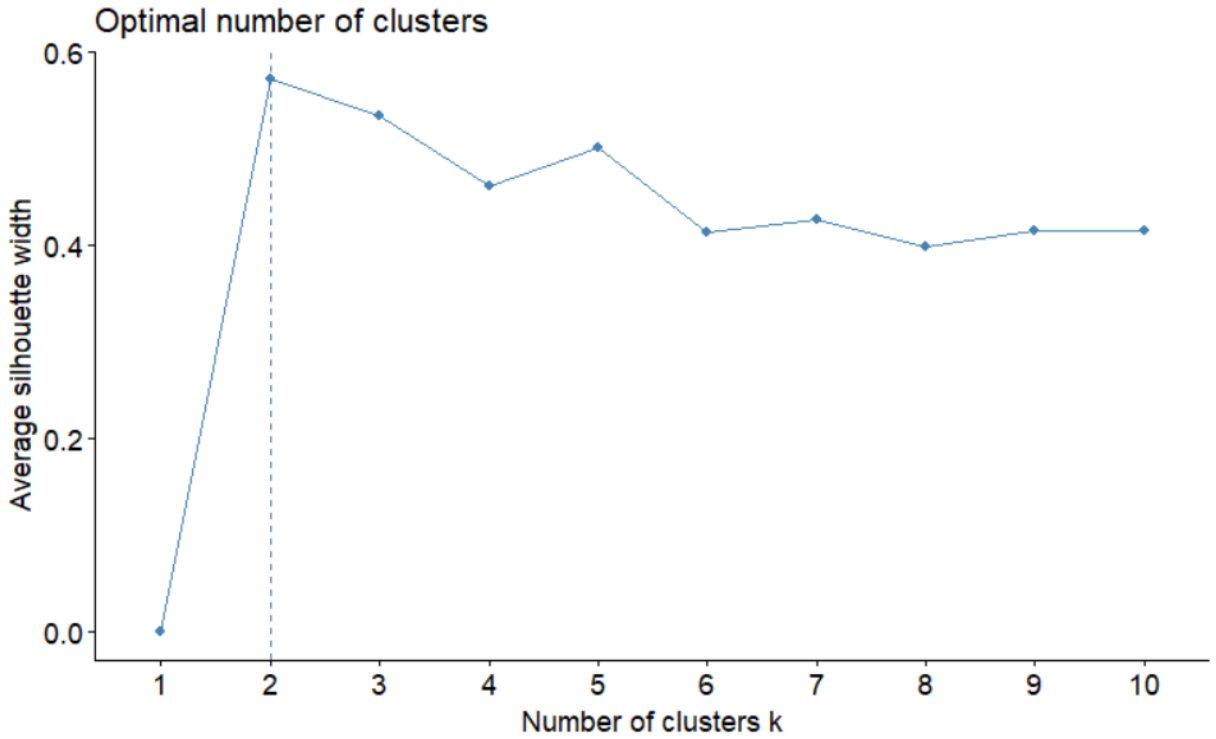
- Some lower outliers (*younger* patients) are found in G1 and others in G3 (older patients).

- We can notice for the case of the LOS:
  - Admissions with the highest LOS values are concentrated in G4, with the most significant variability of them all and a symmetrical data distribution;
  - Admissions concentrated in G1, G2 and G3 are characterized by LOS values that are clearly and above all similarly (between these 3 groups) less significant than in G4, with much lower variability and relatively asymmetrical data distributions;
  - Outliers show some similarities in G1, G2 and G3, whereas they are certainly less numerous in G4, but on the other hand more dispersed.
- For the case of the CCI per WARD_ADMISSION:
  - The Index 3 of CCI corresponds to the peak of admissions proportions in G1 (clearly more frequent within the *Specialist Wards*) and in *G4: Generalist Wards;*
  - The index 2 of CCI corresponds to the peak of admissions proportions in G2 (clearly more frequent in the *Specialist Wards*), in G3 and in *G4: Specialist Wards*;
  - The index 1 of CCI is frequently more observed (in a particularly similar way) within the admissions, in both *Generalist* and *Specialist Wards*, in G2 and in G3. It is less observed in G4 and even literally quasi-unobserved in G1;
  - The index 4 of CCI is the more observed within the admissions in G1 (more within the *Generalist Wards*),a little less in G4 (once again, mainly in *Generalist Wards*), and more or less rare (even inexistent) in the other Wards of the other groups;
  - The index 5 of CCI is not very frequent within the admissions in *G1: Generalist Wards* and very infrequent within those of *G1: Specialist Wards*. It is also not frequent in G4(in both *Specialist* and *Generalist Wards*) and relatively inexistent in the other Wards of the groups that have not been mentioned yet;
  - The indexes 6 and 7 of CCI are observed only very rarely within the admissions: rarely seen in *G1: Generalist Wards* and *G4: Specialist Wards* for the case of index 6 and very little noticed for that of index 7 in *G4: Generalist Wards*.
- For the case of the RISK_DEATH per WARD_ADMISSION:
  - In all of the Wards (both *Generalist* and *Specialist* ones), the peaks of proportions clearly correspond to the admissions with a level of RISK_DEATH 2. Let us just notice that the 2 peaks in G2 and G3 are relatively more significant than those in G1 and G4;
  - Let us notice that admissions with a level of RISK_DEATH 3 are similarly quite frequent both in *Generalist* and *Specialist Wards* for the groups G1 and G4, less frequent in G2 (both Specialist & Generalist) and relatively not frequent in G3 (once again, in both *Generalist* and *Specialist Wards*) ;
  - Admissions with a level of RISK-DEATH 1 are quite frequent in the *Specialist & Generalist Wards* of the group G3, slightly less frequent for the case of G2. This type of admissions is not frequent in *G1: Specialist Wards*, and very infrequent in *G1: Generalist Wards*. The situation is reversed for the case of G4;
  - Admissions with a level of RISK_DEATH 4 are generally not frequent or even rare in all of the Wards of any group, let's just insist on the fact that this *lower frequency* is less significant in *G4: Specialist Wards* and *G3: Generalist Wards*, similarly rare in *G1: Specialist Wards* and *G4: Specialist Wards*, and very rare within the other Wards not yet discussed.

IV- Model 3: Examinations costs/specific analysis

```
df_model3 <- df %>%
  select(COST_RADIOLOGY,
         COST_LAB,
         COST_HAEMATIC,
         COST_CONSULTATIONS,
         COST_CARDIO,
         COST_VAR,
         COST_DIAGNOSTIC
) %>%
  scale()

fviz_nbclust(df_model3, kmeans, method = "silhouette")
```



- The data partitioning method that we have chosen was that of the k-means (k-means clustering);
- The method used for estimating the optimal number of Clusters to specify when partitioning data with the k-means method is that of the Average Silhouette Width;
- Based on the curve above, it is clear that we have a maximum value of the Average Silhouette width with 2 clusters, which means that 2 is then the optimal number of Clusters to be kept.

```
mod3_nc <- 2

mod3 <- kmeans(df_model3,
               centers = mod3_nc,
               nstart = 10,
               iter.max = 200,
)

mod3
```

```
K-means clustering with 2 clusters of sizes 2109, 417


Cluster means:
  COST_RADIOLOGY   COST_LAB COST_HAEMATIC COST_CONSULTATIONS COST_CARDIO   COST_VAR COST_DIAGNOSTIC
1     -0.3516754 -0.1240645   -0.04644016        -0.03621574  -0.3516754 -0.2636288      -0.2992636
2      1.7786175  0.6274629    0.23487361         0.18316308   1.7786175  1.3333169       1.5135417

Clustering vector:
   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1
  [48] 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  [95] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [142] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [189] 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [236] 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 1
 [283] 2 1 2 2 1 2 1 1 1 1 1 1 1 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [330] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [377] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [424] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [471] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [518] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [565] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [612] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [659] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [706] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [753] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [800] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [847] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [894] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [941] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1
 [988] 1 1 1 1 2 1 1 1 1 1 1 1 1
 [ reached getOption("max.print") -- omitted 1526 entries ]

Within cluster sum of squares by cluster:
[1] 4673.694 7568.315
 (between SS / total SS =  30.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

- The number of observations for each of the 2 clusters will be respectively: 2109 and 417 observations;
- The **coordinates (COST_RADIOLOGY, COST_LAB COST_HAEMATIC, COST_CONSULTATIONS, COST_CARDIO, COST_VAR COST_DIAGNOSTIC)** of each Centroid (Position of the Cluster's center) of each Cluster will be as follows:
  - **Centroid of Cluster 1 : (-0.3516754, -0.1240645, -0.04644016, -0.03621574, -0.3516754, -0.2636288, -0.2992636) ;**
  - **Centroid of Cluster 2 : (1.7786175, 0.6274629, 0.23487361, 0.18316308, 1.7786175, 1.3333169, 1.5135417) ;**
- Based on the Clustering Vector above, we can consult the Cluster of membership of each line of observations on admissions by just looking at the number of Cluster representing the line of observations on the latter's position within the vector: *line 1 belongs to Cluster 1, line 2 belongs to Cluster 1, line 988 belongs to Cluster 1… ;*

- The sum of the squares of the distances between the points of a Cluster and their Centroid for each of the Clusters 1 and 2 are respectively 4 673.694 and 7 568.315. From these values, we can deduce that the Cluster 1 is the most compact of the two, whereas the Cluster 2 is the less compact one;
- However, let's notice that the ratio ***" between_SS / total_SS = 30.7 % "*** is not necessarily (perhaps far from being) a good sign that we have a *good enough* Clustering here.

Residuals:

```
res.mod3 <- df %>%
  mutate(groupe= paste0('G',mod3$cluster))
```

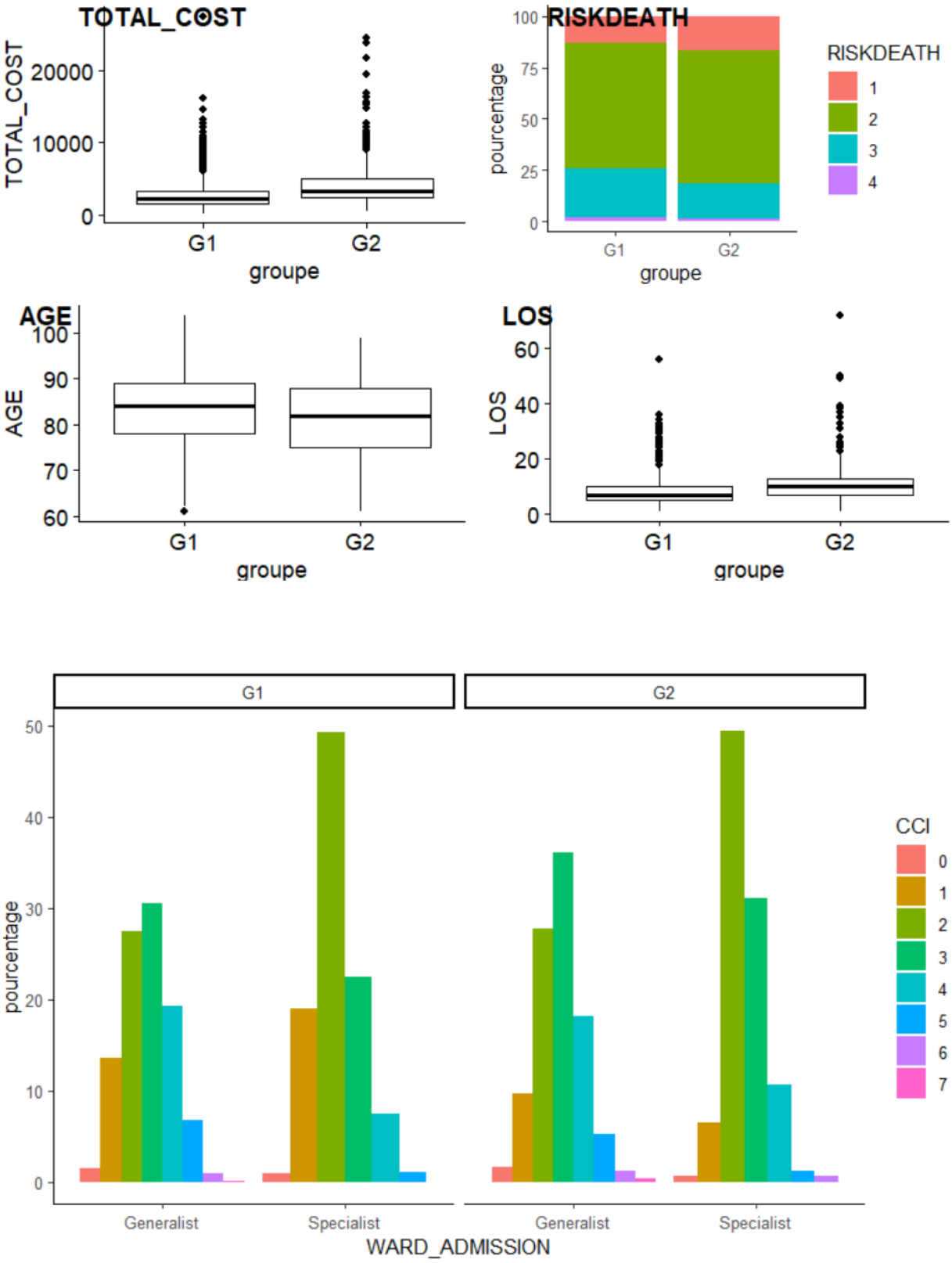
- We have named our Clusters 1 and 2 respectively by « G1 » and « G2 ».

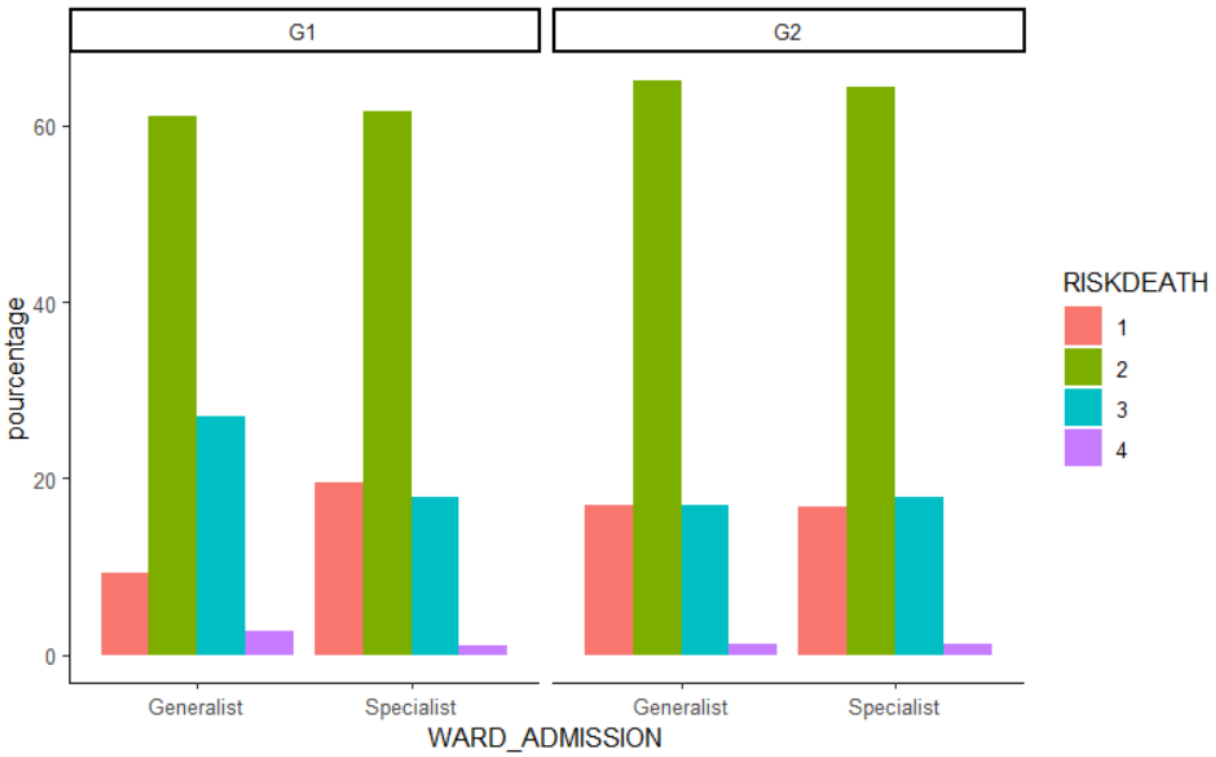
groupe <chr>	variables <chr>	mean <dbl>	sd <dbl>	min <dbl>	q1 <dbl>	median <dbl>	q3 <dbl>	max <dbl>	na <int>
G1	COST_CARDIO	2.024533	5.718183	0.00	0.00	0.00	0.00	46.48	0
G1	COST_CONSULTATIONS	40.315021	67.135490	0.00	0.00	0.00	46.48	583.64	0
G1	COST_DIAGNOSTIC	81.223167	58.749874	0.00	35.33	69.83	115.29	307.20	0
G1	COST_HAEMATIC	1.381105	7.057140	0.00	0.00	0.00	0.00	189.82	0
G1	COST_LAB	75.792997	57.477408	0.00	30.79	63.37	108.58	307.20	0
G1	COST_RADIOLOGY	2.024533	5.718183	0.00	0.00	0.00	0.00	46.48	0
G1	COST_VAR	121.538189	87.888589	0.00	55.37	105.36	165.00	638.36	0
G2	COST_CARDIO	64.144293	42.022785	0.00	60.43	60.43	72.05	319.19	0
G2	COST_CONSULTATIONS	56.087746	91.258732	0.00	0.00	20.66	82.64	743.87	0
G2	COST_DIAGNOSTIC	279.134916	148.571468	120.86	178.76	246.27	330.45	1443.66	0

1-10 of 14 rows Previous 1 2 Next

groupe <chr>	variables <chr>	mean <dbl>	sd <dbl>	min <dbl>	q1 <dbl>	median <dbl>	q3 <dbl>	max <dbl>	na <int>
G2	COST_HAEMATIC	6.503405	41.690301	0.00	0.00	0.00	0.00	556.04	0
G2	COST_LAB	144.342926	172.704129	0.00	39.52	89.07	170.46	1443.66	0
G2	COST_RADIOLOGY	64.144293	42.022785	0.00	60.43	60.43	72.05	319.19	0
G2	COST_VAR	335.222662	176.972470	120.86	221.20	294.48	401.84	1691.60	0

11-14 of 14 rows Previous 1 2 Next


TOTAL_COST
TOTAL_COST
20000
10000
0
G1
G2
groupe
RISKDEATH
pourcentage
100
75
50
25
0
G1
G2
groupe
RISKDEATH
1
2
3
4
AGE
AGE
100
90
80
70
60
G1
G2
groupe
LOS
LOS
60
40
20
0
G1
G2
groupe
G1
G2
pourcentage
50
40
30
20
10
0
Generalist
Specialist
Generalist
Specialist
WARD_ADMISSION
CCI
0
1
2
3
4
5
6
7



- We can notice for the case of the TOTAL_COST :
  - The highest values of TOTAL_COST are found among the admissions in G2, and the lowest ones in G1;
  - The variability is more significant in G2 than in G1 and the data distribution is more asymmetrical in G2 than in G1;
  - Outliers with regard to the TOTAL_COST values are more significant and more dispersed in G2 than in G1.
- We can notice for the case of the RISK_DEATH :
  - Admissions with a level of RISK_DEATH 2 are relatively similar in terms of proportions in the Groups G1 and G2;
  - Admissions with a level of RISK_DEATH 3 are slightly more frequent in G1 than in G2, whereas the situation is reversed when it comes to admissions with a level of RISK_DEATH 1;
  - Admissions with a level of RISK_DEATH 4 are relatively rare within the two groups.
- We can notice for the case of the AGE :
  - Age values are slightly higher in G1 than in G2 (the lowest age values of admitted patients are found within the latter), with however a more significant variability for the case of G2;
  - Data distribution is slightly more symmetrical in G2 than in G1;
  - No indication of the existence of eventual outliers is reported when we look at the two BoxPlots.
- For the case of the LOS :
  - LOS values are slightly higher in G2 than in G1, whereas the variability (not significant) and the symmetry (relatively symmetrical) of the data distribution are relatively similar regarding the two groups;

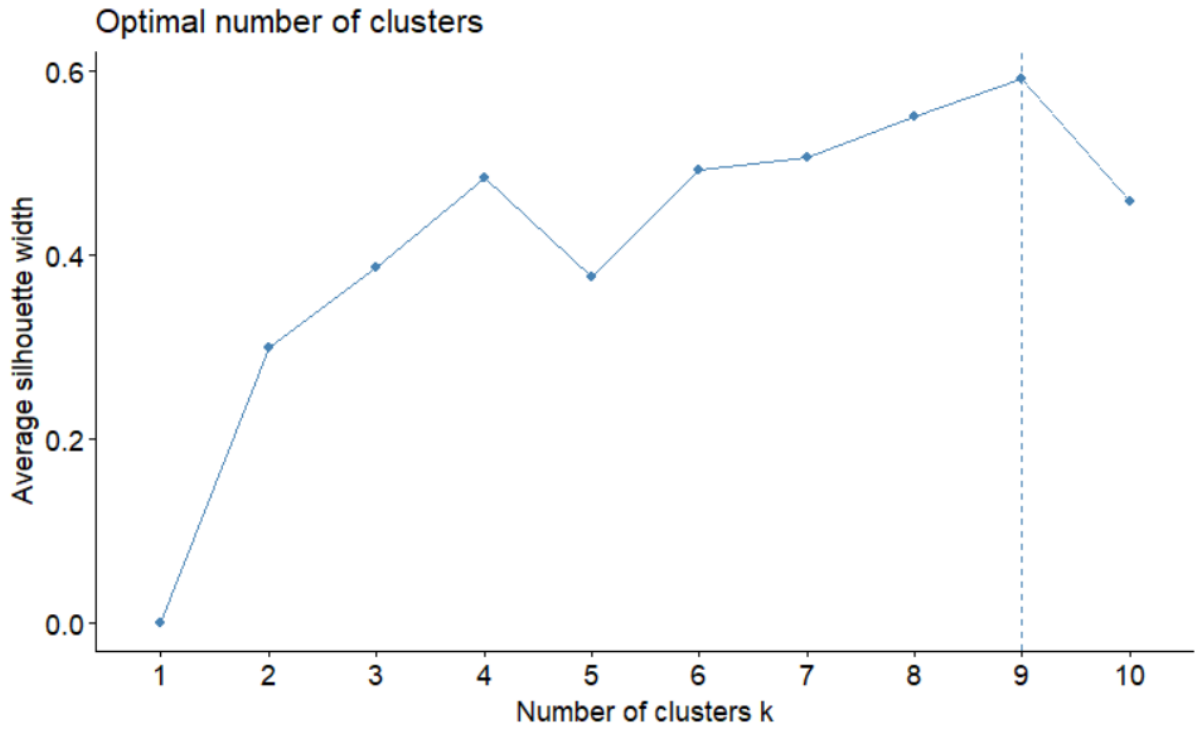
- LOS values are slightly higher in G2 than in G1, whereas the variability (poorly significant) and the symmetry (relatively symmetrical) of the data distribution are relatively similar regarding the two groups;
- Outliers are more scattered in G2 (with the highest values) than in G1;

- For the case of the CCI per WARD_ADMISSION :
  - Within the two groups, the peak of admissions in the *Specialist Wards* (similar peaks) corresponds to that with Indexes 2 of CCI, whereas it is the admissions with the Indexes 3 of CCI that constitute the peaks in the *Generalist Wards* (more significant peak in *G2 : Generalist Wards) ;*
  - Admissions with an index 4 of CCI are relatively similar in terms of proportion in *G1: Generalist Wards* and in *G2: Generalist Wards*, whereas in the *Specialist Wards*, this type of admissions is more frequent in G2 than in G1;
  - Admissions with an index 5 of CCI are not frequent within the two groups for the case of the *Generalist Wards* (just slightly more frequent in G1), whereas in the *Specialist Wards*, this type of admission is similarly very infrequent in both groups;
  - Admissions with indexes 6 or 7 of CCI are very infrequent or rare (even, totally missing for the case of the *Specialist Wards* in G1) within all the Wards of the two groups.
- For the case of the RISK_DEATH per WARD_ADMISSION :
  - Whether in G1 or G2, the peaks within all the Wards correspond to the admissions with a level 2 of RISK_DEATH (generally slightly more significant in G2);
  - Admissions with a level 3 of RISK_DEATH are more significant in terms of proportions in *G1: Generalist Wards* than in *G2: Generalist Wards*, whereas this type of admission is quite similarly frequent in the *Specialist Wards* of the two groups;
  - Admissions with a level 1 of RISK_DEATH are slightly more frequent in *G1: Specialist Wards* than in *G2: Specialist Wards*, whereas this type of admission is clearly less frequent in *G1: Generalist Wards* than in *G2: Generalist Wards*;
  - Admissions with a level 4 of RISK_DEATH are not frequent in all of the Wards of the two groups, with, perhaps, just a particular attention to give to the case of the proportion seen in *G1: Generalist Wards* which seems to be slightly more significant than in the other Wards.

V- Model 4 : Admission department, risk of death and sojourn time

```
df_model4 <- df %>%
  select( IDADMISSION,WARD_ADMISSION, RISKDEATH) %>%
  mutate(RISKDEATH= paste0("Risk_Death_", RISKDEATH),
         WARD_ADMISSION= paste0("WS", WARD_ADMISSION)
  ) %>%
  mutate(val = 1)  %>%
  spread(RISKDEATH, val,  fill = 0 ) %>%
  mutate(val = 1)  %>%
  spread(WARD_ADMISSION, val,  fill = 0 ) %>%
  select(- IDADMISSION)

fviz_nbclust(df_model4, kmeans, method = "silhouette")
```



- The data partitioning method that we have chosen was that of the k-means (k-means clustering);
- The method used for estimating the optimal number of Clusters to specify when partitioning data with the k-means method is that of the Average Silhouette Width;
- Based on the curve above, it is clear that we have a maximum value of the Average Silhouette width with 9 clusters, which means that 9 is then the optimal number of Clusters to be kept.

```
mod4_nc <- 9
mod4 <- kmeans(df_model4,
               centers = mod4_nc,
               nstart = 10,
               iter.max = 200,
)
mod4
```

```
K-means clustering with 9 clusters of sizes 490, 515, 283, 266, 634, 110, 107, 92, 29

Cluster means:
  Risk Death 1 Risk Death 2 Risk Death 3 Risk Death 4        WS08      WS21       WS24 WS2604
1    0.1877551    0.6897959    0.1102041   0.01224490 0.000000000 0.0000000 0.00000000      1
2    0.0000000    0.0000000    1.0000000   0.00000000 0.009708738 0.1825243 0.06601942      0
3    0.2190813    0.7526502    0.0000000   0.02826855 0.000000000 1.0000000 0.00000000      0
4    0.0000000    1.0000000    0.0000000   0.00000000 0.000000000 0.0000000 0.14285714      0
5    0.0000000    1.0000000    0.0000000   0.00000000 0.000000000 0.0000000 0.00000000      0
6    0.0000000    1.0000000    0.0000000   0.00000000 1.000000000 0.0000000 0.00000000      0
7    0.6915888    0.0000000    0.0000000   0.30841121 0.000000000 0.0000000 0.00000000      0
8    1.0000000    0.0000000    0.0000000   0.00000000 1.000000000 0.0000000 0.00000000      0
9    0.9310345    0.0000000    0.0000000   0.06896552 0.000000000 0.0000000 0.58620690      0
     WS2605       WS68
1 0.0000000 0.00000000
2 0.6718447 0.06990291
3 0.0000000 0.00000000
4 0.0000000 0.85714286
5 1.0000000 0.00000000
6 0.0000000 0.00000000
7 1.0000000 0.00000000
8 0.0000000 0.00000000
9 0.0000000 0.41379310
```

```
Clustering vector:
   [1] 1 2 5 1 1 1 2 1 2 2 2 1 8 6 1 1 1 2 2 2 5 1 5 5 5 2 4 5 1 6 1 5 7 1 3 1 8 4 6 5 1 1 6 2 4 4 1
  [48] 7 5 5 5 4 5 5 7 5 3 5 5 5 3 9 2 3 7 3 5 3 9 6 4 3 7 5 1 4 2 1 1 2 2 2 6 6 1 3 8 5 2 7 5 5 1 2
  [95] 5 3 1 5 1 5 7 4 4 1 7 8 4 5 1 5 7 5 5 1 3 5 4 1 3 3 7 5 1 5 5 5 1 4 5 5 5 2 7 8 4 6 5 5 5 1 1
 [142] 4 5 4 4 4 5 1 6 1 5 5 3 5 1 4 5 3 5 4 1 5 5 5 1 6 4 1 5 5 1 2 3 1 7 2 2 5 1 1 5 2 5 2 1 2 3 2
 [189] 1 2 1 5 1 2 2 4 4 8 5 2 2 3 5 2 5 3 1 7 1 3 8 2 5 1 2 6 5 1 1 1 6 2 1 5 6 5 3 5 2 5 5 3 5 6 2
 [236] 1 3 5 5 5 2 5 8 3 2 7 3 4 5 1 3 5 1 1 5 3 3 1 3 4 5 5 4 2 2 2 2 5 2 3 5 8 7 1 3 2 3 4 1 3 2 5
 [283] 5 2 1 2 5 5 2 2 4 1 2 1 6 4 8 5 4 5 2 4 8 2 2 1 1 5 4 1 7 6 1 2 5 1 3 2 1 1 5 2 7 5 3 8 3 2 5
 [330] 4 4 1 5 5 2 3 5 1 2 1 5 2 7 2 2 5 2 2 5 5 2 1 4 1 2 5 5 1 3 5 5 7 5 5 5 2 1 2 3 1 5 4 5 1 3 5
 [377] 1 8 5 5 2 2 4 1 4 5 2 6 4 6 5 4 3 4 5 3 8 5 1 1 4 2 1 1 7 5 2 1 4 9 3 5 2 2 2 5 1 1 2 1 5 4 3
 [424] 5 1 3 4 7 5 4 5 3 5 2 1 1 4 3 9 4 1 1 8 1 2 5 5 5 2 4 5 9 5 3 4 6 7 4 4 2 5 3 3 2 2 4 5 2 4 2
 [471] 2 1 2 5 2 2 1 8 7 5 2 5 1 1 5 2 6 7 1 2 2 5 1 5 5 5 7 5 5 7 5 2 5 3 2 2 1 5 4 5 2 1 5 4 5 3 5
 [518] 3 1 2 4 4 1 4 2 2 1 5 1 6 2 6 5 5 4 1 1 1 2 5 2 1 5 4 1 4 5 4 1 2 5 2 5 2 1 5 4 3 5 4 3 2 5 1
 [565] 2 5 5 5 5 5 6 1 7 5 5 5 5 5 6 7 6 2 1 2 5 5 5 3 3 6 4 3 2 2 7 2 1 5 7 1 2 2 3 1 1 3 5 1 5 1 2 5
 [612] 6 5 8 2 5 8 7 5 5 2 4 1 6 2 1 1 2 7 4 6 1 2 5 4 5 2 1 3 2 3 1 3 2 1 8 5 2 2 1 4 1 5 5 2 1 2 2
 [659] 5 8 5 5 4 1 5 5 2 6 7 4 3 1 2 2 6 3 6 5 1 5 6 2 7 1 4 8 4 2 4 3 4 5 5 5 1 7 8 2 4 8 3 6 5 5 3
 [706] 3 3 1 6 6 1 5 5 1 5 5 6 4 1 5 3 3 2 3 2 2 5 1 2 2 5 1 4 2 6 1 4 1 4 5 2 2 2 2 5 1 1 5 1 5 5 4
 [753] 5 5 3 3 4 5 2 9 1 5 4 3 1 2 8 4 1 8 3 2 5 1 1 5 5 5 5 5 4 1 1 1 5 2 2 4 5 3 1 3 5 1 1 7 4 3 1 8
 [800] 1 2 1 2 5 1 1 5 1 2 4 5 5 2 2 6 1 3 2 3 3 1 2 7 4 2 5 3 5 2 1 1 8 3 8 2 3 1 1 3 5 2 6 2 5 3 5
 [847] 5 3 5 4 2 5 4 4 1 2 5 3 5 3 4 2 2 1 1 8 5 2 2 5 1 1 2 1 5 5 5 9 6 3 1 1 4 2 5 4 5 2 4 2 4 1 2
 [894] 5 4 5 2 5 4 5 1 1 1 4 5 1 1 5 5 2 1 3 2 5 2 8 4 4 5 7 3 5 3 2 1 2 5 5 5 8 1 2 2 5 2 6 5 2 6 5
 [941] 4 8 4 5 2 5 5 3 8 4 4 2 5 4 1 9 3 3 5 1 7 1 2 1 2 2 3 1 9 2 3 3 5 1 2 1 1 2 2 2 5 9 1 1 2 7 8
 [988] 5 9 2 1 6 4 7 2 1 5 2 4 1
 [ reached getOption("max.print") -- omitted 1526 entries ]

Within cluster sum of squares by cluster:
[1] 233.55102 260.57476 108.87633  65.14286   0.00000   0.00000  45.64486   0.00000  17.79310
 (between_SS / total_SS =  77.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"
```

- The number of observations for each of the 9 Clusters will be respectively: 490, 515, 283, 266, 634, 110, 107, 92 and 29 observations;
- The **coordinates (Risk_Death_1, Risk_Death_2, Risk_Death_3, Risk_Death_4, WS08, WS21, WS24, WS2604, WS2605, WS68)** of each Centroid (position of the Cluster's center) of each Cluster will be as follows:
  - **Centroid of Cluster 1 : (0.1877551, 0.6897959, 0.1102041, 0.01224490, 0.000000000, 0.0000000, 0.00000000, 1, 0.0000000, 0.00000000) ;**
  - **Centroid of Cluster 2 : (0.0000000, 0.0000000, 1.0000000, 0.00000000, 0.009708738, 0.1825243, 0.06601942, 0, 0.6718447, 0.06990291) ;**
  - **Centroid of Cluster 3 : (0.2190813, 0.7526502, 0.0000000, 0.02826855, 0.000000000, 1.0000000, 0.00000000, 0, 0.0000000, 0.00000000) ;**
  - **Centroid of Cluster 4 : (0.0000000, 1.0000000, 0.0000000, 0.00000000, 0.000000000, 0.0000000, 0.14285714 0, 0.0000000, 0.85714286).**
  - **Centroid of Cluster 5 : (0.0000000, 1.0000000, 0.0000000, 0.00000000, 0.000000000, 0.0000000, 0.00000000, 0, 1.0000000, 0.00000000).**
  - **Centroid of Cluster 6 : (0.0000000, 1.0000000, 0.0000000, 0.00000000, 1.000000000, 0.0000000, 0.00000000, 0, 0.0000000, 0.00000000).**
  - **Centroid of Cluster 7 : (0.6915888, 0.0000000, 0.0000000, 0.30841121, 0.000000000, 0.0000000, 0.00000000, 0, 1.0000000, 0.00000000).**
  - **Centroid of Cluster 8 : (1.0000000, 0.0000000, 0.0000000, 0.00000000, 1.000000000, 0.0000000, 0.00000000, 0, 0.0000000, 0.00000000).**
  - **Centroid of Cluster 9 : (0.9310345, 0.0000000, 0.0000000, 0.06896552, 0.000000000, 0.0000000, 0.58620690, 0, 0.0000000, 0.41379310).**
- Based on the Clustering Vector above, we can consult the Cluster of membership of each line of observations on the admissions by just looking at the number of cluster representing the

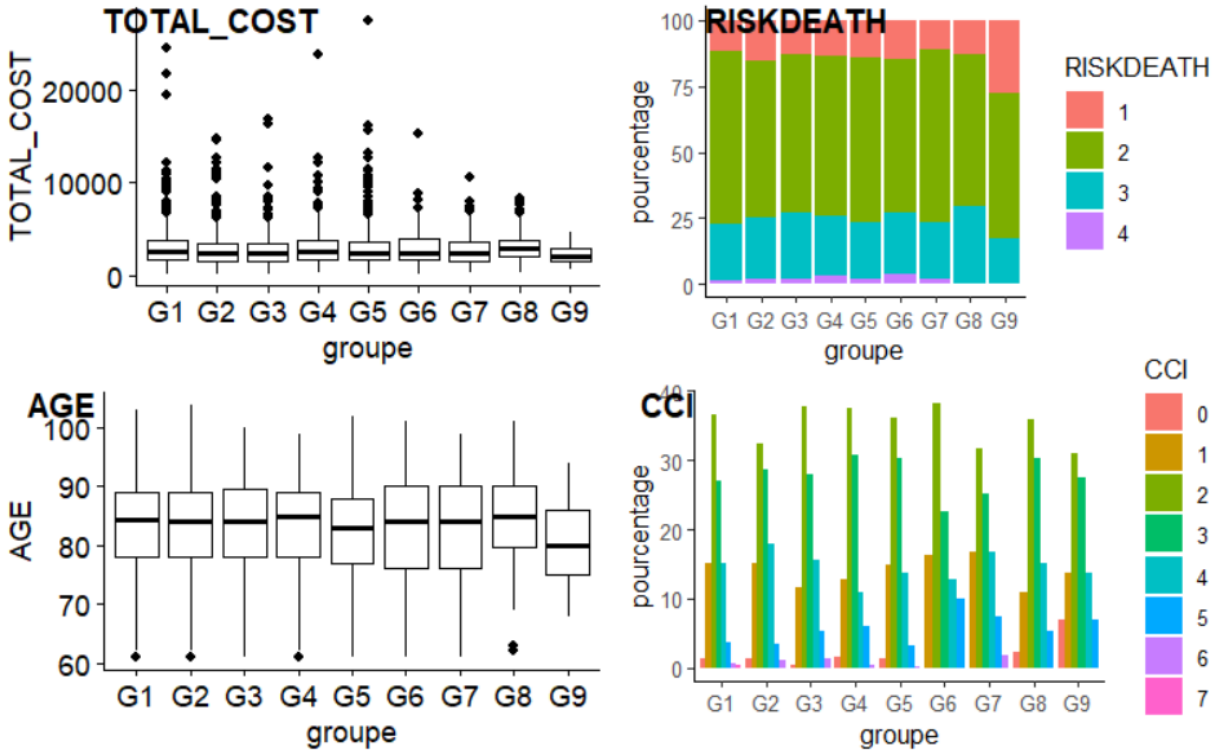
line of observations on the position of this latter within the vector: *line 1 belongs to Cluster 1, line 2 belongs to Cluster 2, line 988 belongs to Cluster 5… ;*

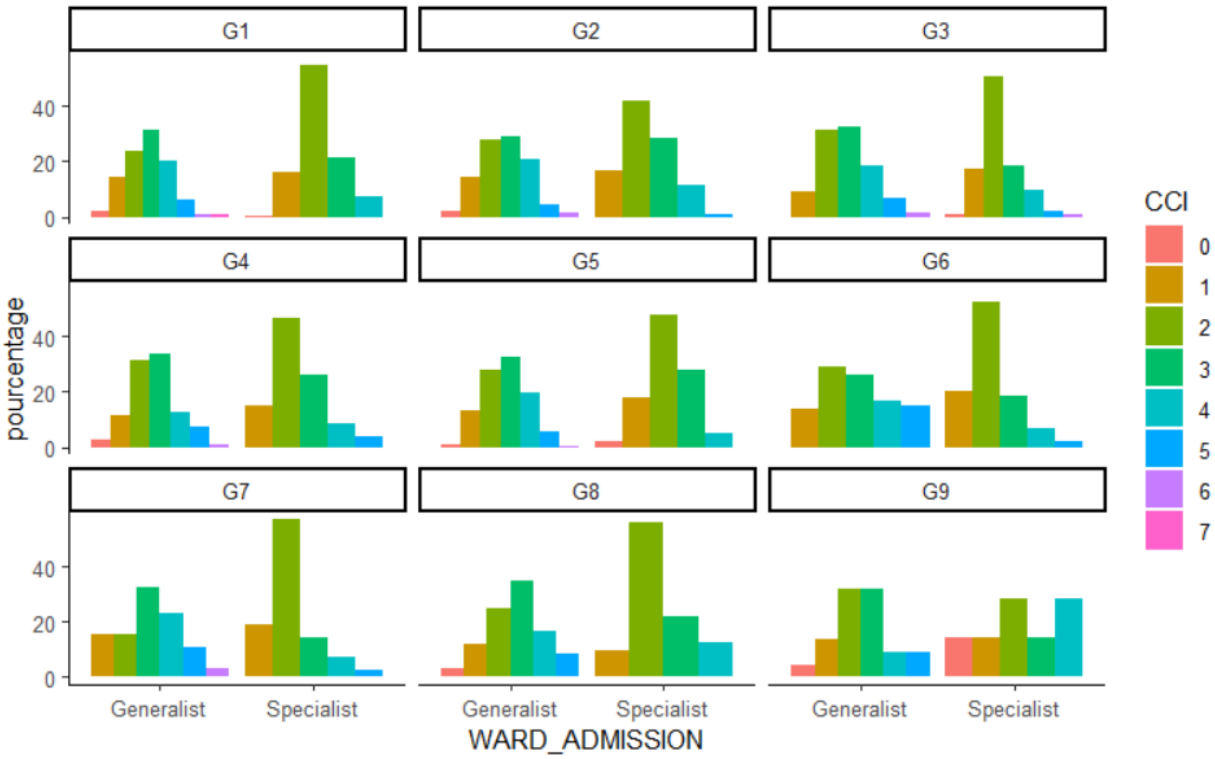
- The sum of the squares of the distances between the points of a Cluster and their Centroid for each of the Clusters 1, 2, 3, 4, 5, 6, 7, 8 and 9 are respectively: 233.55102, 260.57476, 108.87633, 65.14286, 0.00000, 0.00000, 45.64486, 0.00000 and 17.79310. Based on these values, we can sort in an ascending order all the Clusters, from the least compact one to the most compact one: **Cluster 5 – Cluster 6 – Cluster 8 – Cluster 9 – Cluster 7 – Cluster 4 – Cluster 3 – Cluster 1 – Cluster 2** ;
- The ratio ***"between_SS / total_SS = 77.4 %"*** can tell us that we have here a Clustering *relatively* good

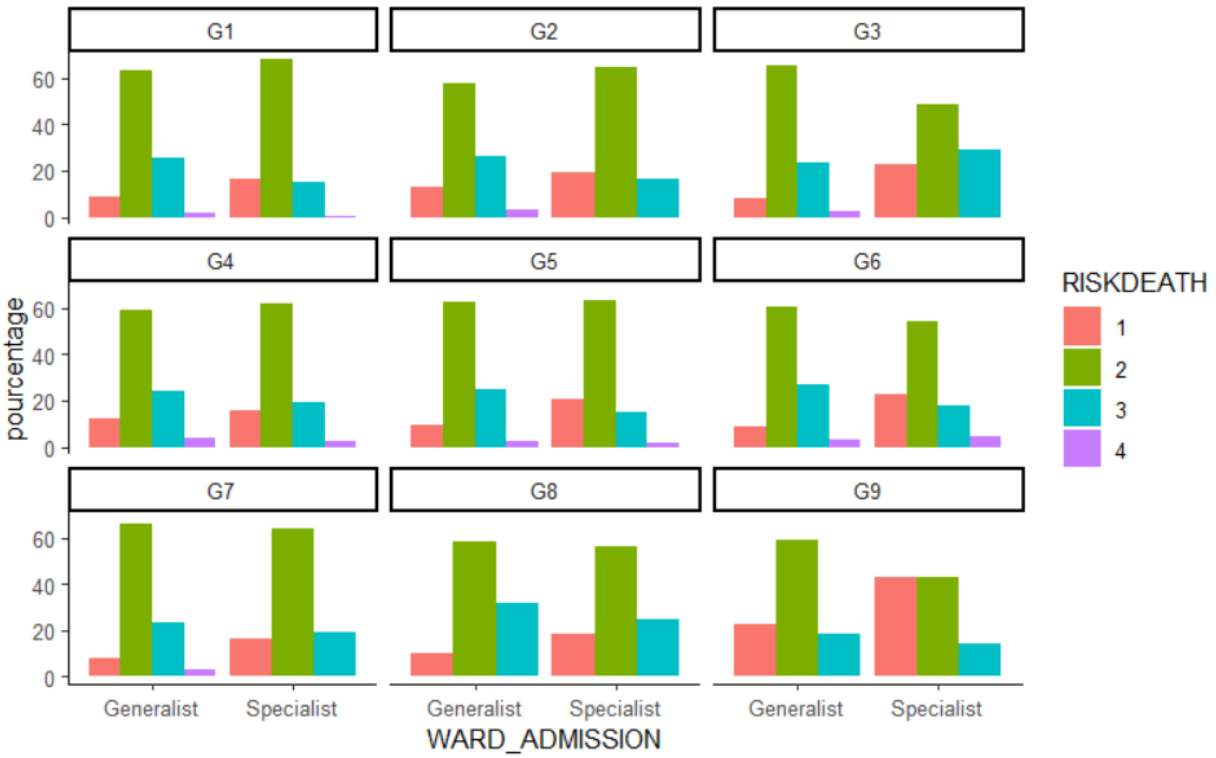
Residuals :

```
mod4Ordre <- paste0('G',1:mod4_nc)
```

- We have named our Clusters respectively by « G1 », « G2 », « G3 », « G4 », « G5 », « G6 », « G7 », « G8 » and « G9 ».







- <u>We can notice for the case of the TOTAL_COST:</u>
  - The BoxPlots corresponding to the Groups are more or less on the same level regarding the TOTAL_COST, with a few exceptions for the cases of G1, G6 and G8, having the highest values, and for those of G7 and G9, this time, having the lowest values;

- Concerning the symmetry of the data distribution, all of the Groups' BoxPlots seem to show a relative symmetry, except for G7 which presents a quit significant asymmetry;
- Outliers are more significant within G1, G2 and G5 (corresponding in particular to the highest outlier observed), whereas they are clearly less significant within G6, G7 and G8, even inexistent within G9.

- We can notice for the case of the RISK_DEATH:
  - For all of the groups, admissions with a level 2 of RISK_DEATH constitute the peak of proportions;
  - Admissions with a level 1 of RISK_DEATH are particularly frequent in G9, whereas this type of admissions is less frequent within G1 and G7;
  - Admissions with a level 3 of RISK_DEATH are particularly frequent within G8, whereas all the other groups all present a proportion relatively similar when it comes to the admissions linked to this level of risk;
  - Admissions with a level of RISK_DEATH 4 are not generally frequent (G4 and G6), rare (G1, G2, G3, G5, G7) or even inexistent (G8 and G9) within all of the groups.
- For the case of the AGE:
  - For the cases of the highest age values, based on the BoxPlots above, we can notice that the Groups G1, G2, G3, G4, G6, G7 and G8 are relatively concerned by those values, whereas when it comes to the lowest ones, it is the groups G6 (with a high variability and a data distribution slightly asymmetrical), G7 (with a high variability and a data distribution slightly asymmetrical) and G9 that are the most concerned;
  - Apart from the groups G6 and G7 already mentioned previously, the variability and the distributions of data within the other groups all seem to be roughly similar to each other;
  - A few outliers can be noticed within G1, G2, G4 and G8.
- For the case of the CCI:
  - In all of the groups, admissions with an index of CCI 2 constitute the peak of proportion (peak slightly more significant in G6, G3 and G4, and less significant in G9, G7 and G2);
  - Admissions with an index of CCI 3 are generally quite frequent within the groups (however, less significant in G6 and G7 compared to the proportions within the other groups);
  - Admissions with the indexes of CCI 5 and 4 are generally less frequent within the groups (especially within G1 and G2, but also within G5 concerning the index of CCI 5);
  - Admissions with the indexes of CCI 6 and 7 remain more or less frequent within all of the groups, even inexistent within some of them (G9, G8, G6).
- For the case of the CCI per WARD_ADMISSION:
  - For all of the Specialist Wards of each group, we clearly notice that admissions with an index of CCI 2 constitute the peak of proportions (peak particularly less significant in *G9: Specialist Wards*);
  - For all of the Generalist Wards of each group, it is the admissions with and index of CCI 3 that make up the peak of proportions (peak exceptionally less significant for the case of G9);
  - Except in *G9: Specialist Wards* (exceptionally more frequent), admissions with CCI indexes 4 and 5 are relatively less frequent within the Specialist Wards of the groups;

Whereas for the case of the Generalist ones, except for G6, these types of admissions are less frequent within the groups;

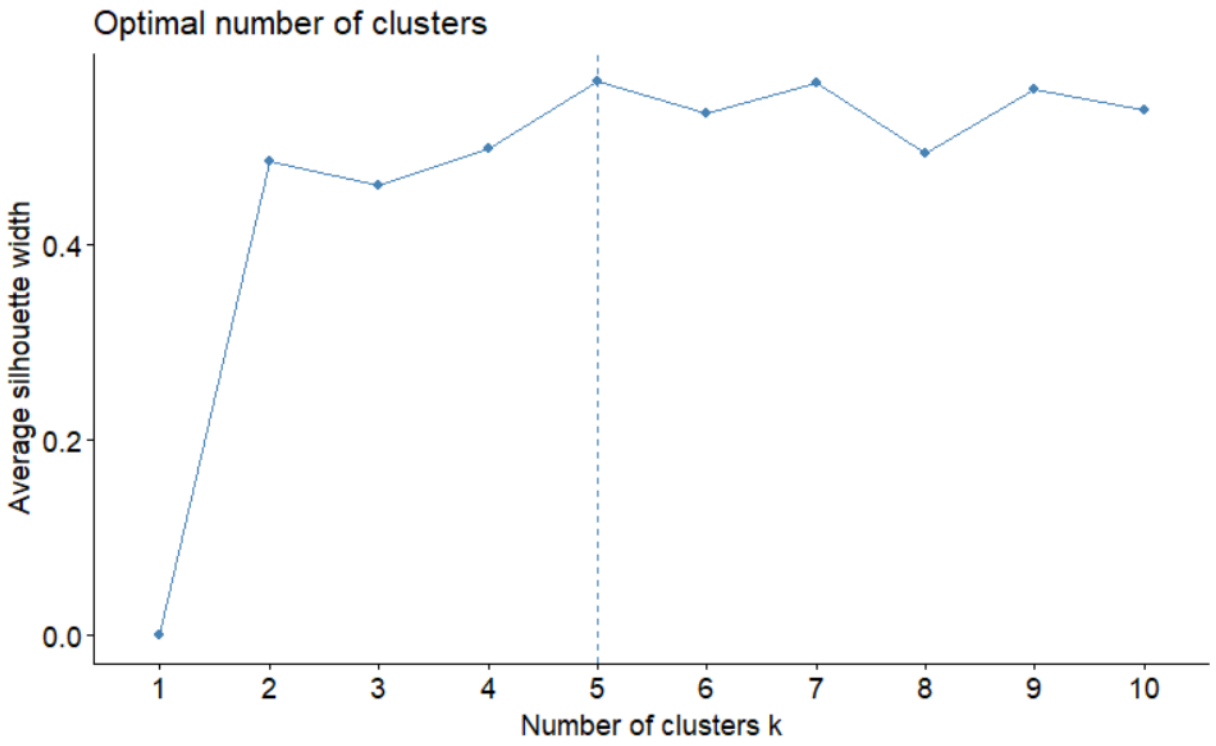
- Admissions corresponding to the indexes 6 and 7 of CCI are almost rare within the majority of the Wards of each group, even rare sometimes.

- For the case of the RISK_DEATH per WARD_ADMISSION:
  - Whether in the Generalist Wards or in the Specialist Wards, admissions with a level of RISK_DEATH 2 always constitute the peaks of proportions;
  - Except for the case of G3, admissions with a level 3 of RISK_DEATH are more frequent in the Generalist Wards than in the Specialist ones within all the groups;
  - Admissions with a level 1 of RISK_DEATH are generally more frequent in the Specialist Wards than in the Generalist ones, with a special attention given to the group G9, where this type of admissions is exceptionally more frequent in the Specialist Wards;
  - Admissions with a level 4 of RISK_DEATH are not frequent, rare, even inexistent (in G8 and G9) within the Wards of the groups.

VI-Model 5: Sojourn Time and Risk of Death

```
df_model5 <- df %>%
  select(LOS, RISKDEATH) %>%
  scale()

fviz_nbclust(df_model5, kmeans, method = "silhouette")
```



- The data partitioning method that we have chosen was that of the k-means (k-means clustering);
- The method used for estimating the optimal number of Clusters to specify when partitioning data with the k-means method is that of the Average Silhouette Width;
- Based on the curve above, it is clear that we have a maximum value of the Average Silhouette width with 5 clusters, which means that it is then the optimal number of Clusters to be kept.

```
mod5_nc <- 5

mod5 <- kmeans(df_model5,
               centers = mod5_nc,
               nstart = 10,
               iter.max = 200,
)

mod5
```

```
K-means clustering with 5 clusters of sizes 568, 337, 913, 553, 155

Cluster means:
         LOS  RISKDEATH
1  0.5084901 -0.2079862
2 -0.3391334 -1.7295931
3 -0.5555552 -0.1944729
4 -0.1548816  1.4572387
5  2.6989485  0.4690952

Clustering vector:
   [1] 1 1 1 3 2 1 2 1 2 3 4 1 1 2 2 5 3 4 2 3 1 1 3 1 3 2 4 3 3 1 1 1 1 1 1 1 1 1 1 2 1 1 3 1 5 5 3
  [48] 1 3 3 5 5 4 5 3 2 3 4 4 3 3 2 2 1 2 3 3 3 3 2 1 1 4 3 3 1 3 1 3 3 2 3 3 3 3 1 3 2 3 3 2 3 3 3
  [95] 3 2 3 3 3 2 3 2 1 2 3 3 3 3 3 2 2 3 3 3 2 3 2 3 3 4 1 2 3 3 2 3 1 3 3 2 2 3 1 3 1 3 3 2 3 3 1
 [142] 3 2 1 4 3 1 3 4 1 1 3 3 3 4 4 3 2 2 2 3 2 5 3 3 2 4 1 2 2 3 3 4 3 3 3 2 1 1 3 2 3 3 3 3 4 1 2
 [189] 3 3 2 2 1 1 2 3 3 2 2 3 2 2 3 1 3 2 5 4 2 1 1 5 1 5 2 2 3 1 1 4 5 3 1 1 1 1 2 3 3 2 1 2 3 3 1
 [236] 2 1 5 3 1 2 1 1 1 1 3 2 2 2 2 1 2 2 1 3 3 2 2 4 3 2 3 3 2 2 2 3 1 2 1 1 3 2 3 2 1 3 2 3 2 2 1
 [283] 1 3 5 4 4 5 2 4 2 2 3 2 3 1 1 2 2 2 3 2 1 2 4 4 1 1 5 1 4 1 3 1 5 3 4 1 1 2 4 1 1 3 1 3 1 3 3
 [330] 4 5 1 1 1 3 3 4 5 5 2 1 1 3 4 1 1 1 2 4 3 2 4 5 3 3 3 4 1 3 4 1 3 3 5 5 5 3 2 5 1 1 4 5 5 5 5
 [377] 1 5 1 1 1 5 5 4 4 1 3 3 3 3 3 5 4 4 2 5 4 1 5 1 3 1 5 1 1 3 3 3 5 1 3 4 1 3 3 3 2 4 4 1 3 2 4
 [424] 5 3 4 4 3 1 4 1 1 2 1 1 3 3 5 2 2 3 3 2 1 3 3 1 1 2 2 2 3 1 3 2 4 1 2 2 5 1 3 1 2 3 1 4 1 1 5
 [471] 1 3 2 3 1 4 1 3 2 3 2 2 3 4 2 2 3 3 4 3 3 2 1 3 3 2 2 1 2 2 3 3 1 2 2 3 2 2 3 2 1 3 3 1 4 5 3
 [518] 3 1 1 1 1 1 1 3 5 3 2 2 1 5 2 4 5 2 3 1 1 4 3 1 4 1 1 1 5 3 2 3 3 1 1 1 3 1 3 3 2 2 1 3 1 1 1
 [565] 5 5 3 3 2 5 5 3 3 2 1 1 2 2 3 3 1 1 1 1 1 1 4 5 1 1 4 1 1 3 1 1 2 4 3 3 4 1 5 5 5 4 5 2 3 3 4
 [612] 4 1 5 2 2 5 4 1 2 1 2 3 2 1 1 1 1 5 5 4 5 5 1 1 1 2 5 5 3 1 4 4 5 4 2 4 4 1 3 1 4 5 5 3 1 4 3
 [659] 3 4 1 3 3 3 1 1 3 4 4 4 1 1 4 5 4 4 1 3 2 4 1 4 4 1 1 1 5 5 5 5 4 1 1 3 4 2 4 4 4 3 1 1 4 1 3
 [706] 4 1 1 3 3 3 4 3 1 1 1 3 4 1 4 2 3 1 3 4 5 3 3 3 3 4 3 3 3 2 4 1 1 1 3 4 4 3 1 4 3 3 3 4 4 3 1
 [753] 1 4 1 1 3 4 3 4 5 1 1 3 4 5 2 3 3 3 1 1 4 4 2 4 3 4 3 5 4 4 2 3 4 1 4 2 3 3 3 1 4 1 1 5 3 3 1
 [800] 2 3 2 3 3 3 3 4 1 4 1 4 3 3 2 1 3 2 3 2 5 3 1 1 3 3 3 1 4 1 3 3 4 3 4 3 1 1 3 1 4 4 3 3 3 4 1
 [847] 4 4 3 4 3 3 1 3 1 1 3 4 1 4 1 1 1 2 4 1 5 1 4 3 1 3 4 4 3 2 2 1 4 4 3 3 4 3 4 1 4 4 3 4 3 4 4
 [894] 3 1 1 4 4 4 1 4 4 1 3 3 4 4 1 4 4 3 1 2 3 5 1 4 1 4 3 3 1 4 2 3 2 4 4 3 4 4 1 4 5 5 3 3 4 3 3
 [941] 3 1 4 2 3 1 3 4 1 1 1 3 1 4 3 2 3 4 3 3 1 2 3 4 4 3 2 3 2 4 5 1 4 1 2 1 1 1 4 3 1 2 1 2 2 1 4
 [988] 4 1 3 3 3 4 1 3 3 2 2 1 3
 [ reached getOption("max.print") -- omitted 1526 entries ]

Within cluster sum of squares by cluster:
[1] 126.5125 114.9211 103.1526 369.1019 450.9033
 (between_SS / total_SS =  76.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.within
ss" "betweenss"
[7] "size"         "iter"         "ifault"
```

- The number of observations for each of the 5 Clusters will be respectively: 568, 337, 913, 553 and 155 observations;

- The **coordinates (LOS, RISKDEATH)** of each Centroid (position of the Cluster's center) of each Cluster are as follows:
  - **Centroid of Cluster 1 : 0.5084901, -0.2079862) ;**
  - **Centroid of Cluster 2 : (-0.3391334, -1.7295931) ;**
  - **Centroid of Cluster 3 : (-0.5555552, -0.1944729) ;**
  - **Centroid of Cluster 4 : (-0.1548816, 1.4572387) ;**
  - **Centroid of Cluster 5 : (2.6989485, 0.4690952) ;**
- From the Clustering Vector above, we can consult the Cluster of membership of each line of observations on admissions by just looking at the number of cluster that represents the line of observations on the position of the latter within the vector: *line 1 belongs to Cluster 1, line 2 belongs to Cluster 1, line 988 belongs to Cluster 4…*
- The sum of the square of the distances between the points of a Cluster and their Centroid for each of the Clusters 1, 2, 3, 4 and 5 are respectively 126.5125, 114.9211, 103.1526, 369.1019 and 450.9033. From these values, we can rank the Clusters in a descending order, from the most compact one to the least compact one: **Cluster 3 – Cluster 2 – Cluster 1 – Cluster 4 – Cluster 5**;
- The ratio ***"between_SS / total_SS = 76.9 %"*** can tell us that we have here a Clustering relatively good.

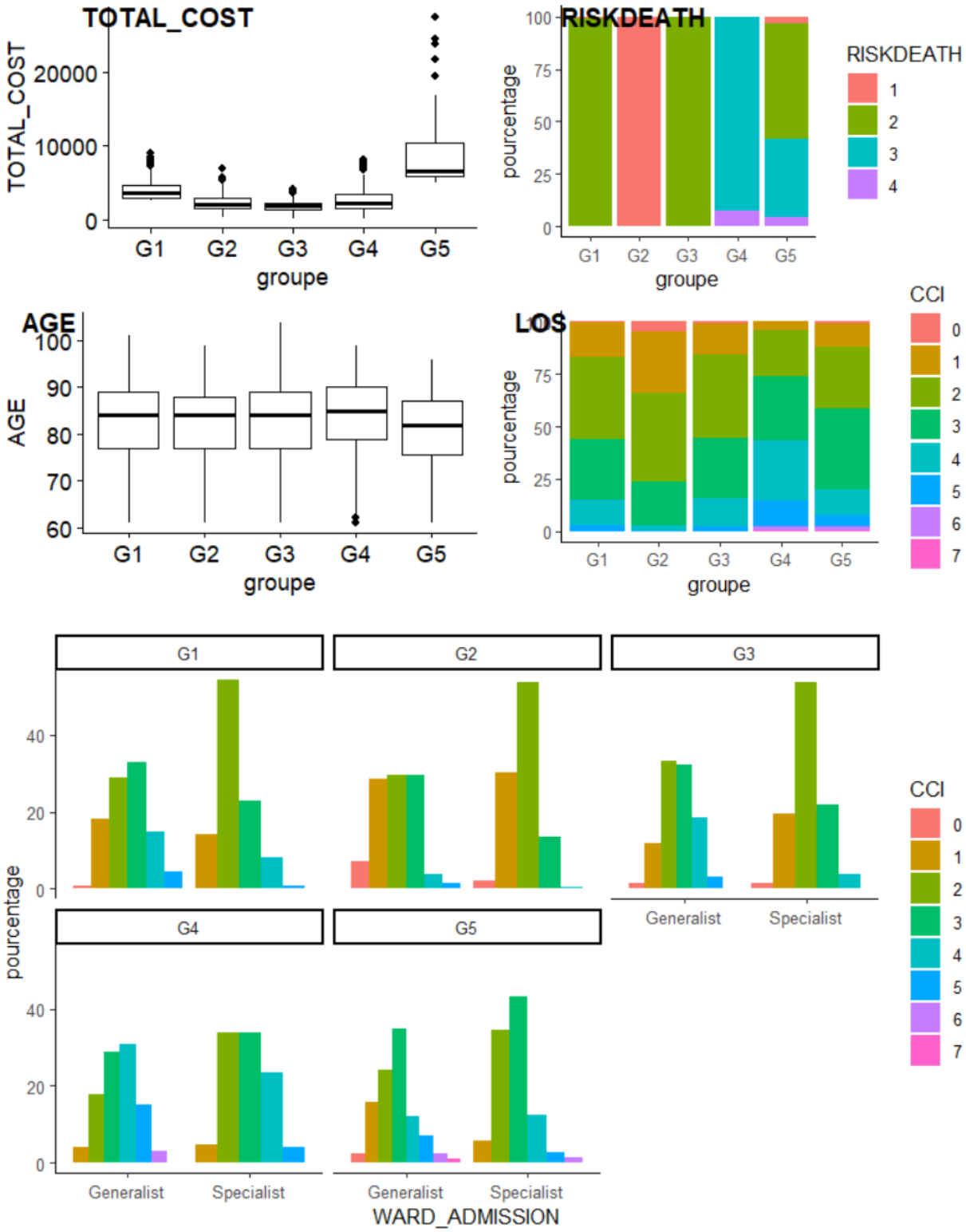
Residuals :

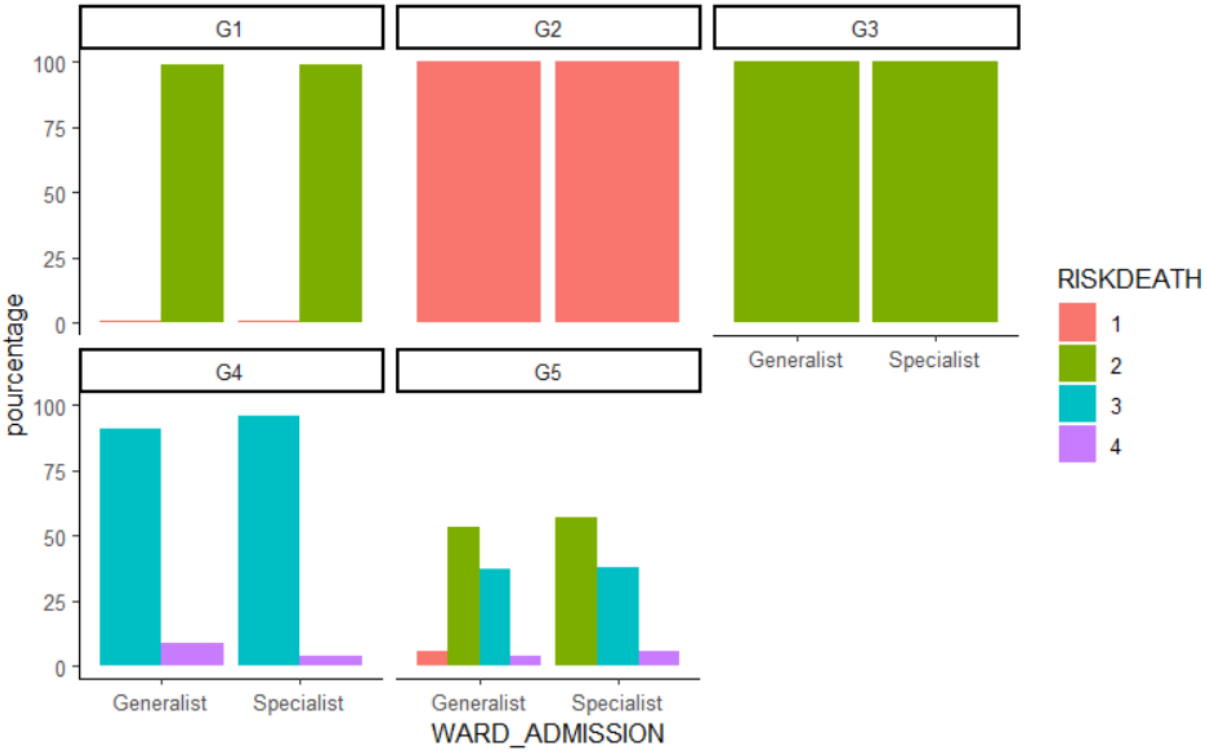
```
res.mod5 <- df %>%
  mutate(groupe= paste0('G',mod5$cluster))
```

- We have named our Clusters respectively by: « G1 », « G2 », « G3 », « G4 » and « G5 ».

groupe <chr>	variables <chr>	mean <dbl>	sd <dbl>	min <dbl>	q1 <dbl>	median <dbl>	q3 <dbl>	max <dbl>	na <int>
G1	LOS	11.521127	2.5028190	9	9	11	13.0	19	0
G1	RISKDEATH	1.991197	0.0934918	1	2	2	2.0	2	0
G2	LOS	6.807122	3.2525004	1	5	6	9.0	16	0
G2	RISKDEATH	1.000000	0.0000000	1	1	1	1.0	1	0
G3	LOS	5.603505	1.8703799	1	4	6	7.0	8	0
G3	RISKDEATH	2.000000	0.0000000	2	2	2	2.0	2	0
G4	LOS	7.831826	3.9442124	1	5	7	11.0	19	0
G4	RISKDEATH	3.075949	0.2651569	3	3	3	3.0	4	0
G5	LOS	23.703226	7.8228555	17	19	21	24.5	72	0
G5	RISKDEATH	2.432258	0.6347037	1	2	2	3.0	4	0

1-10 of 10 rows


TOTAL_COST
TOTAL_COST
20000
10000
0
G1
G2
G3
G4
G5
groupe
RISKDEATH
pourcentage
100
75
50
25
0
RISKDEATH
1
2
3
4
AGE
AGE
100
90
80
70
60
LOS
CCI
0
1
2
3
4
5
6
7
G1
G2
G3
G4
G5
Generalist
Specialist
WARD_ADMISSION



- We can notice for the case of the TOTAL_COST:
  - The group G5 clearly groups the highest values of TOTAL_COST, with a variability particularly significant and a very asymmetrical data distribution;
  - With the least significant variability, the Group G3 seems to contain the lowest values of TOTAL_COST;
  - The Groups G2 and G4 are relatively at the same levels in terms of TOTAL_COST values, with a variability slightly less significant in G4;
  - Outliers are found within all the groups, relatively less significant and lower in G1, G2, G3 and G4, and more significant and more dispersed in G5.
- We can notice for the case of the RISK_DEATH:
  - The Groups G1 and G3 are mainly and largely made up of admissions with a level 2 of RISK_DEATH, an admission type which is however less frequent in G5;
  - The Group G2 is exceptionally constituted by admissions with a level 1 of RISK_DEATH, an admission type which is however in a very small quantity in G5;
  - Admissions with a level 3 of RISK_DEATH are very frequent in G4, and less frequent in G5;
  - Admissions with a level 4 of RISK_DEATH are very rarely observed in G4 and G5.
- We can notice for the case of the AGE:
  - The BoxPlots corresponding to the five groups are more or less at the same level of AGE values, with roughly the same variability and relatively asymmetrical data distributions;
  - However, we can notice that the highest values of ages are observed in G4, G1 and G3, and the lowest ones are roughly found in G5;
  - A few outliers can be found within G4.
- For the case of the CCI:
  - Generally, each group is made up mainly of admissions with an index 3 of CCI, then of admissions with and index 2 of CCI;

- Admissions with and index 1 of CCI are exceptionally more frequent in G2 than in the other groups, whereas for those with the index 4 of CCI, they are more frequent in G4 than in the other groups;
- Admissions with and index 5 of CCI are seen only in small proportions within the Groups G5, G4 an G1;
- Admissions that correspond to an index 0 of CCI are very rarely seen in G2;
- Admissions that correspond to an index 6 or 7 of CCI are very rarely observed in G4 and G5.

- For the case of CCI per WARD_ADMISSION:
  - Concerning the Specialist Wards, the peak of proportions are made up of admissions with an index 2 of CCI in G1, G2 and G3, whereas in G4 and G5, those with the index 3 seem to take over;
  - Concerning the Generalist Wards, the peaks are made up of admissions with an index 3 of CCI in G1, G2 (slightly) and G5. In G3, it is those with the index 2 that take over, whereas, exceptionally, in G4, it is those with the index 4 that take over;
  - Admissions with an index 4 or 5 of CCI constitute proportions that are more or less significant within the Generalist Wards of the Groups G1, G3, G4 and G5; This same type of admissions is exceptionally found in a good proportion within *G4: Specialist Wards*, which is not however a behavior observed at the level of the Specialist Wards of the other groups;
  - Admissions with the indexes 6 and 7 of CCI are found only rarely within the groups;
  - Admissions with an index 0 of CCI are also rarely found within the groups, mainly in *G2: Generalist Wards*.
- For the case of the RISK_DEATH per WARD_ADMISSION:
  - Strangely, Wards in G2 only contain admissions with a level 1 of RISK_DEATH, whereas this type of admissions concerns only a small proportion of *G5: Generalist Wards* and tiny proportions in *G1: Generalist Wards* and *G1: Specialist Wards*;
  - Once again, strangely, Wards in G3 only contain admissions with a level 2 of RISK_DEATH. A type of admission that also constitutes the peaks of proportions in *G1: Generalist Wards* and *G1: Specialist Wards*, but also in *G5: Generalist Wards* and *G5: Specialist Wards*;
  - Admissions with a level 3 of RISK_DEATH constitute the peaks of proportions in *G4: Generalist Wards* and *G4: Specialist Wards*, and also constitute some good proportions of admissions in *G5: Generalist Wards* and *G5: Specialist Wards*;
  - Admissions with a level 4 of RISK_DEATH are found only very infrequently (*G4: Generalist Wards* & *G5: Specialist Wards*), rarely (*G4: Specialist Wards* & *G5: Generalist Wards),* or even never found (G1, G2 and G3) within the groups.