Analyse de données hospitalières - Clustering

I- Setup

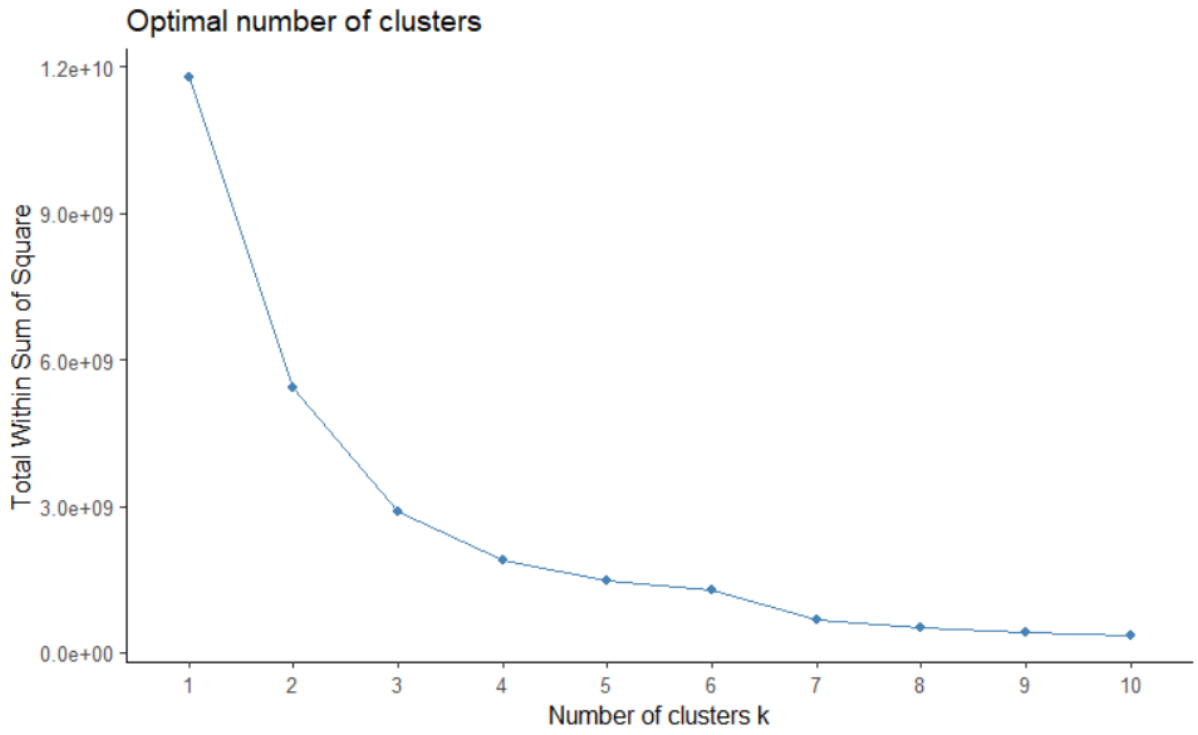
```
df <- df_origine %>%
  filter(AGE > 60)
```

- Sélectionner seulement les observations pour les patients âgés de plus de 60 ans.

II- Modèle 1 : Coûts totaux, âge et CCI

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

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



- La méthode de partitionnement de données choisie est le partitionnement en k-moyennes (k-means clustering).
- La méthode utilisée pour la l'estimation du nombre optimal de Clusters à spécifier lors du partitionnement des données avec la méthode de k-means est celle du WSS (Total Within cluster Sums of Squares).
- En se basant sur le graphique ci-dessus, il est clair que nous pouvons observer le *coude* « Number of clusters k = 3 », signifiant que 3 est alors le nombre optimal de Clusters retenu.

```
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 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 3 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 2 3 3 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 3 3 2 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
 [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 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 2 3 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 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"
```

- Le nombre d'observations pour chacun des 3 clusters sera respectivement de : 62, 1786 et 678 observations.
- Les **coordonnées (TOTAL_COST, CCI, AGE)** de chaque Centroid (position du centre de Cluster) de chaque Cluster sont les suivantes:
  - **Centroid de Cluster 1 : (12101.328, 2.903226, 77.93548) ;**
  - **Centroid de Cluster 2 : (2025.263, 2.593505, 83.52352) ;**
  - **Centroid de Cluster 3 : (4739.181, 2.507375, 81.53687).**
- A partir du Clustering Vector ci-dessus, nous pouvons consulter le Cluster d'appartenance de chaque ligne d'observations sur les admissions en consultant juste le numéro de cluster qui représente la ligne d'observations sur la position de cette dernière dans le vecteur : ligne 1 appartient à Cluster 2, ligne 2 appartient à Cluster 3, ligne 988 appartient à Cluster 3…
- La Somme des carrées des distances des points d'un Cluster à leur Centroid pour chacun des Clusters 1, 2 et 3 sont respectivement de : 962 615 839, 1 032 147 453 et de 904 193 493. A partir de ces valeurs, nous pouvons en déduire que le Cluster 3 est le plus compact de tous, suivi du Cluster 1 et qu'enfin, le Cluster 2 est le moins compact de tous.

- Le ratio ***"between_SS / total_SS = 75.4"*** peut nous indiquer que nous avons là un Clustering *relativement bon*.

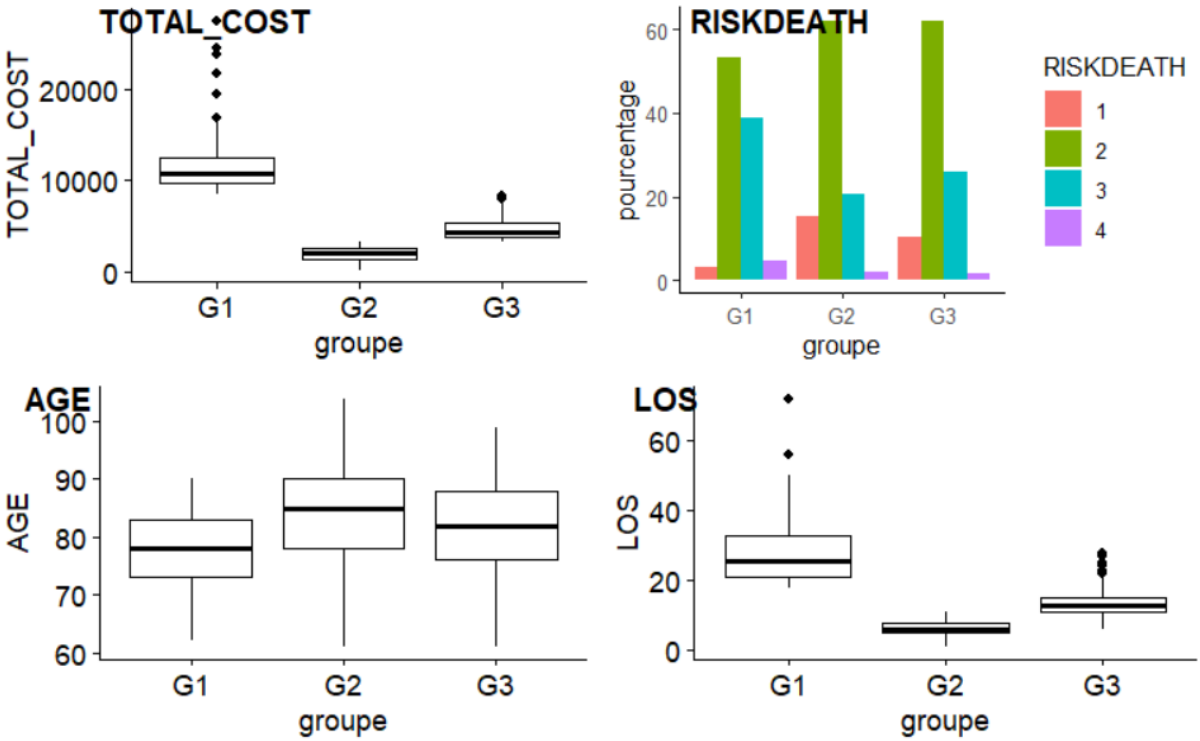
Résidus :

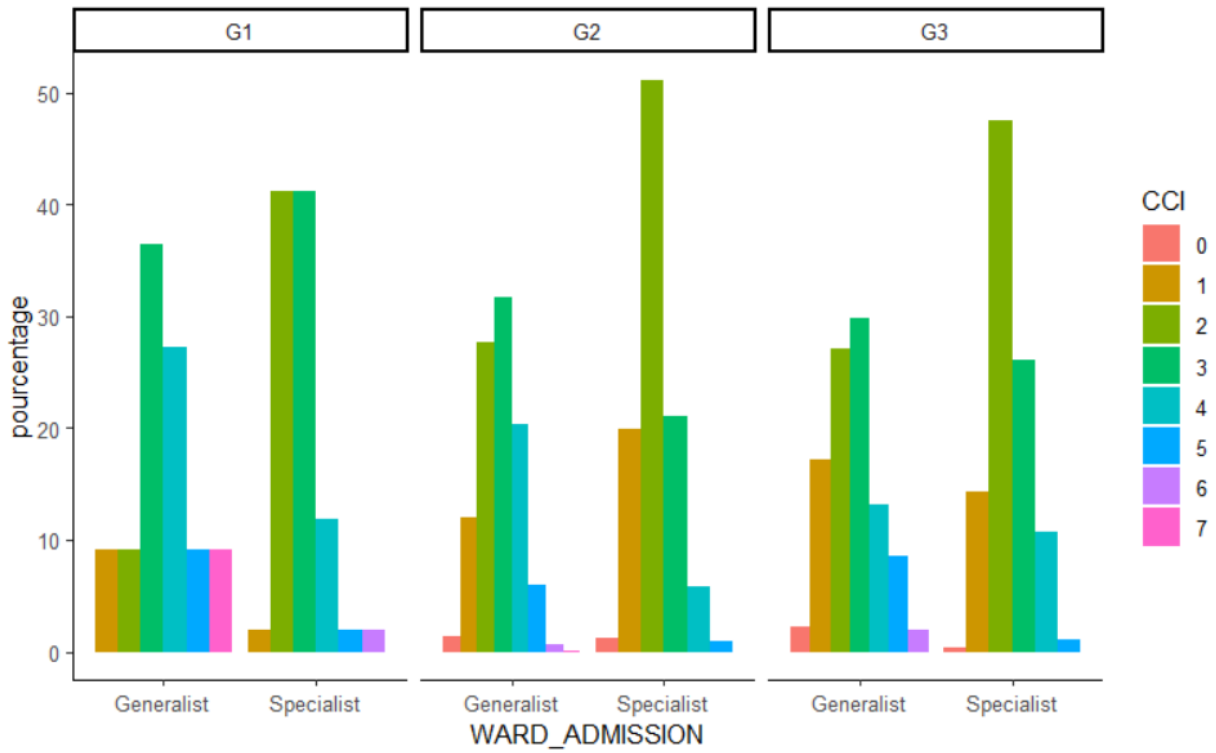
```
res.mod1 <- df %>%
  mutate(groupe= paste0('G',mod1$cluster))
```

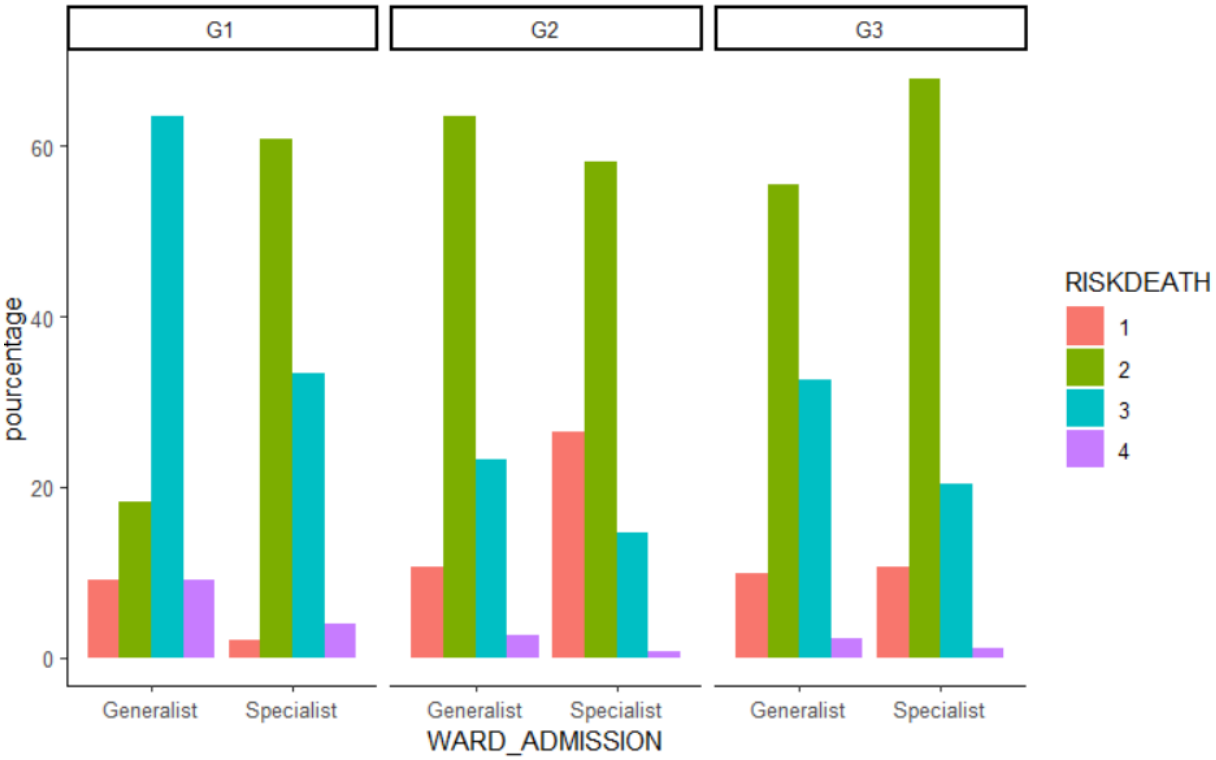
- Nous avons nommé nos Cluster 1, 2 et 3 respectivement par « G1 », « G2 » et « G3 ».

groupe <chr>	variables <chr>	mean <dbl>	sd <dbl>	min <dbl>	q1 <dbl>	median <dbl>	q3 <dbl>	max <dbl>	na <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>Nous pouvons remarquer que pour le cas des TOTAL_COST :</u>
  - Les valeurs les plus élevées sont retrouvées au sein de G1, avec aussi la plus grande variabilité de données caractérisées par une distribution asymétrique en passant ;

- Les valeurs les moins élevées sont clairement retrouvées au sein de G2, avec aussi la plus petite des variabilités de données ;
- Le groupe G3 correspond à des valeurs plus élevées que celles observées en G2, avec une variabilité des données légèrement plus importante et plus ou moins asymétrique en termes de distribution de données ;
- Les données aberrantes sont le plus observées au sein de G1 et largement moins au sein de G3. Elles sont quasi absentes en G2.

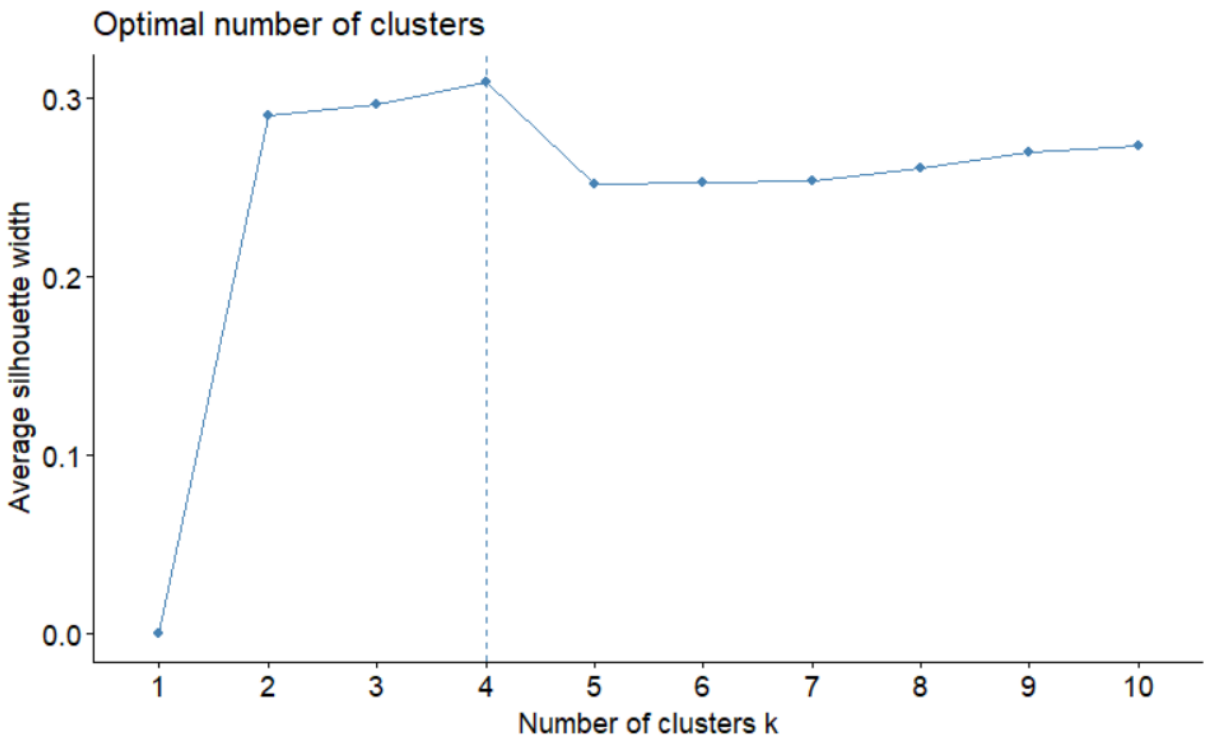
- Pour le cas du RISK_DEATH :
  - Les pics de proportions correspondent tous aux admissions avec un niveau de RISK_DEATH 2 chez les trois groupes, des pics qui sont plus importants en G2 et G3 qu'en G1 ;
  - Le niveau de RISK_DEATH 3 est plus fréquent pour ce qui est des admissions en G1, il l'est légèrement moins pour ceux en G3, et largement moins pour celles en G2 ;
  - Le niveau de RISK_DEATH 1 est plus fréquent pour ce qui est des admissions en G2, l'est légèrement moins pour ceux en G3 et relativement rare pour celles en G1.
  - Le niveau de RISK_DEATH 4 reste relativement rare dans les trois groupes, mais principalement retrouvé dans G1 par contre.
- Pour ce qui est des AGE:
  - Les admissions correspondant aux âges les plus élevés concernent le plus G2, puis G3 et légèrement moins G1 ;
  - La variabilité est assez importante dans chaque groupe, et la distribution des données est quant à elle relativement symétrique dans chaque groupe.
- Pour ce qui est des LOS:
  - Les valeurs de LOS les plus élevées se rencontrent (très) majoritairement au sein de G1, tandis que les plus basses se rencontrent en G2 ;
  - La variabilité est largement moins importante en G2 et en G3, mais plus importante en G1 ;
  - Des données aberrantes se rencontrent en G1 et G3.
- Pour ce qui est des CCI par WARD_ADMISSION:
  - Nous pouvons remarquer que les proportions d'admissions avec un CCI d'indice 2 constituent les pics pour le cas de tous les Wards Specialist de chaque Groupe. Des pics plus importants en G2 (aux environs de 50%) et G3 qu'en G1.
  - Les pics pour le cas des Wards Generalist sont quant à eux constitués par les proportions d'admissions avec un CCI d'indice 3, plus importantes en G1, puis en G2 et finalement moins importantes en G3;
  - Les CCI d'indice 7 sont particulièrement fréquents pour le cas des admissions en G1 : Wards Generalist, et rare, voir même inexistant dans les autres Wards de n'importe quel Groupe.
  - Les CCI d'indice 4 sont toujours plus ou moins fréquents pour le cas des admissions, que ce soit en Wards Specialist ou en Wards Generalist dans tous les Groupes (Plus fréquente en G1 : Wards Generalist).
- Pour ce qui est des RISK_DEATH par WARD_ADMISSION:
  - A l'exception du cas particulier de G1 : Wards generalist dont le pic de proportion en termes d'admissions correspond (nettement) au niveau de RISK_DEATH 3, les pics de proportions correspondent tous au niveau 2 dans tous les autres Wards de n'importe

quel groupe (un niveau cependant moins observé pour ce qui est des admissions en G1 : Wards Generalist) ;

- A l'exception de G1 : Wards Specialist, le niveau 1 semble être similairement fréquent dans les Wards Generalist de chaque groupe. Pour le cas des Wards Specialist, ce niveau 1 est particulièrement plus fréquent en G2, et l'est moins en G3 et G1. ;
- En effectuant notre lecture de G1 vers G3, nous pouvons constater que la fréquence du niveau 4 décroit dans les Wards Generalist, et le même comportement est aussi retrouvé dans les Wards Specialist, avec juste des proportions moins importantes que dans les Wards generalist.

III- Modèle 2 : Coûts totaux, âge et CCI - centré et réduit

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



- La méthode de partitionnement de données choisie est le partitionnement en k-moyennes (k-means clustering).
- La méthode utilisée pour la l'estimation du nombre optimal de Clusters à spécifier lors du partitionnement des données avec la méthode de k-means est celle de l'Average Silhouette.
- En se basant sur le graphique ci-dessus, il est clair que nous avons une valeur maximum de l'Average Silouhette width avec 4 clusters, signifiant que 4 est alors le nombre optimal de Clusters retenu.

```
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 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"
```

- Le nombre d'observations pour chacun des 4 clusters sera respectivement de : 974, 674, 745 et de 133 observations.
- Les **coordonnées (TOTAL_COST, CCI, AGE)** de chaque Centroid (position du centre de Cluster) de chaque Cluster sont les suivantes:
  - **Centroid de Cluster 1 : (-0.19878202, 0.9241646, 0.5033786) ;**
  - **Centroid de Cluster 2 : (-0.04451713, -0.4516112, -1.2275418) ;**
  - **Centroid de Cluster 3 : (-0.23052149, -0.8272487, 0.5551044) ;**
  - **Centroid de Cluster 4 : (2.97260711, 0.1545108, -0.5750399).**
- A partir du Clustering Vector ci-dessus, nous pouvons consulter le Cluster d'appartenance de chaque ligne d'observations sur les admissions en consultant juste le numéro de cluster qui représente la ligne d'observations sur la position de cette dernière dans le vecteur : ligne 1 appartient à Cluster 2, ligne 2 appartient à Cluster 3, ligne 988 appartient à Cluster 1…

- La Somme des carrées des distances des points d'un Cluster à leur Centroid pour chacun des Clusters 1, 2, 3 et 4 sont respectivement de : 1 121.6722, 892.7428, 710.4988 et de 577.1178. A partir de ces valeurs, nous pouvons en déduire que le Cluster 1 est le moins compact de tous, suivi du Cluster 2, puis du Cluster 3 et qu'enfin, le Cluster 4 est le plus compact de tous.
- Cependant, notons tout de même que le ratio ***"between_SS / total_SS = 56.4"*** n'est pas forcément un bon signe que nous avons là un *Clustering suffisamment bon*.

Résidus :

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

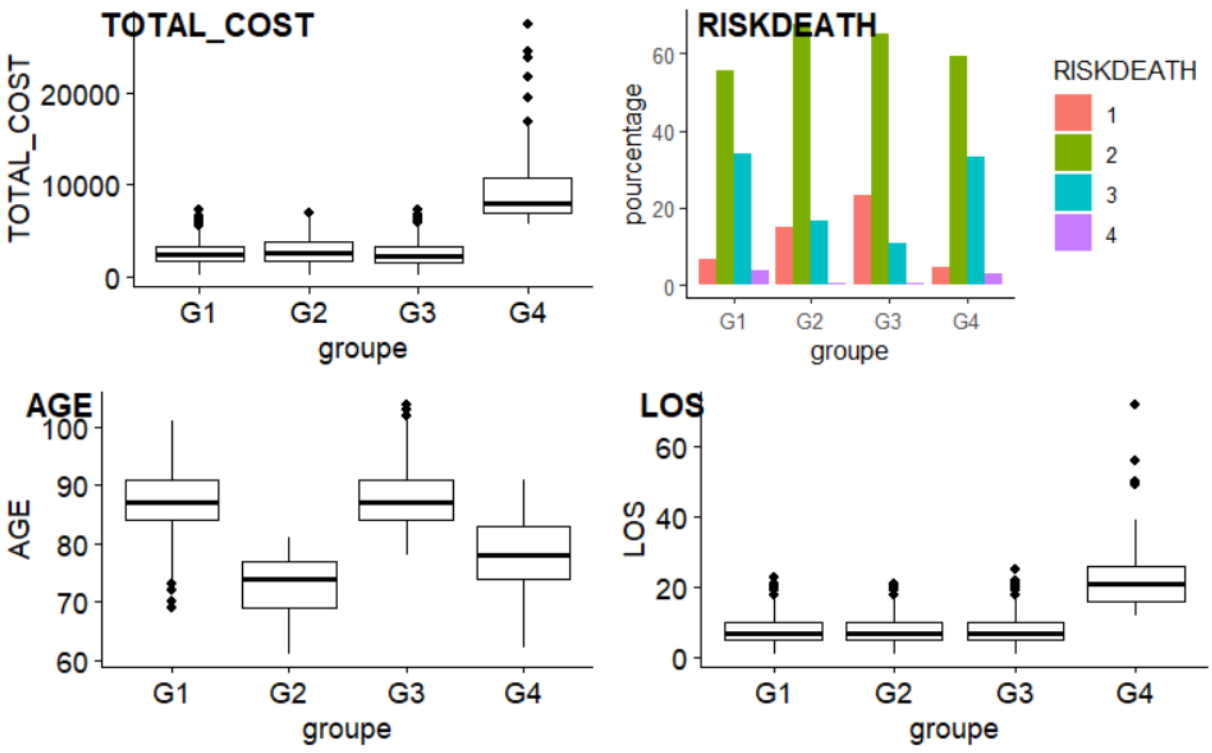
- Nous avons nommé nos Cluster 1, 2, 3 et 4 respectivement par « G1 », « G2 », « G3 » et par « G4 ».

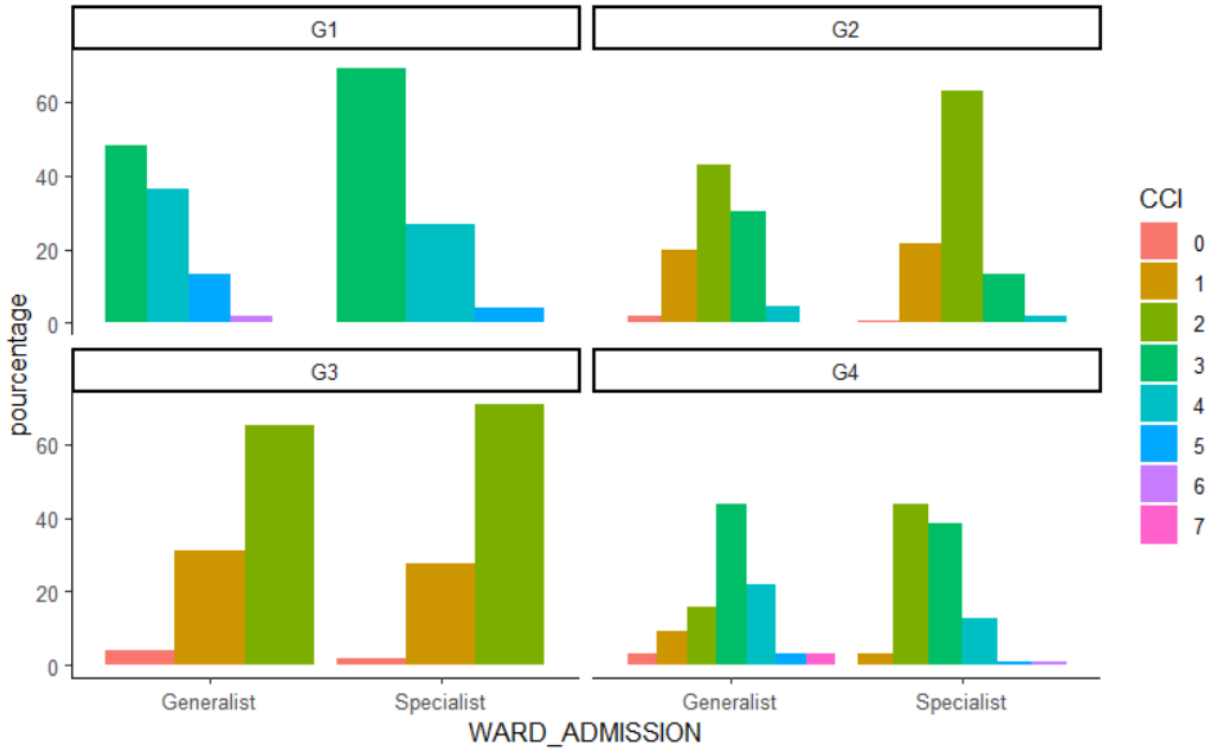
groupe <chr>	variables <chr>	mean <dbl>	sd <dbl>	min <dbl>	q1 <dbl>	median <dbl>	q3 <dbl>	max <dbl>	na <int>
G1	AGE	87.088296	5.3022137	69.00	84.000	87.000	91.000	101.00	0
G1	CCI	3.618070	0.7531516	3.00	3.000	3.000	4.000	7.00	0
G1	TOTAL_COST	2571.619025	1198.4282809	285.00	1734.633	2377.035	3242.677	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

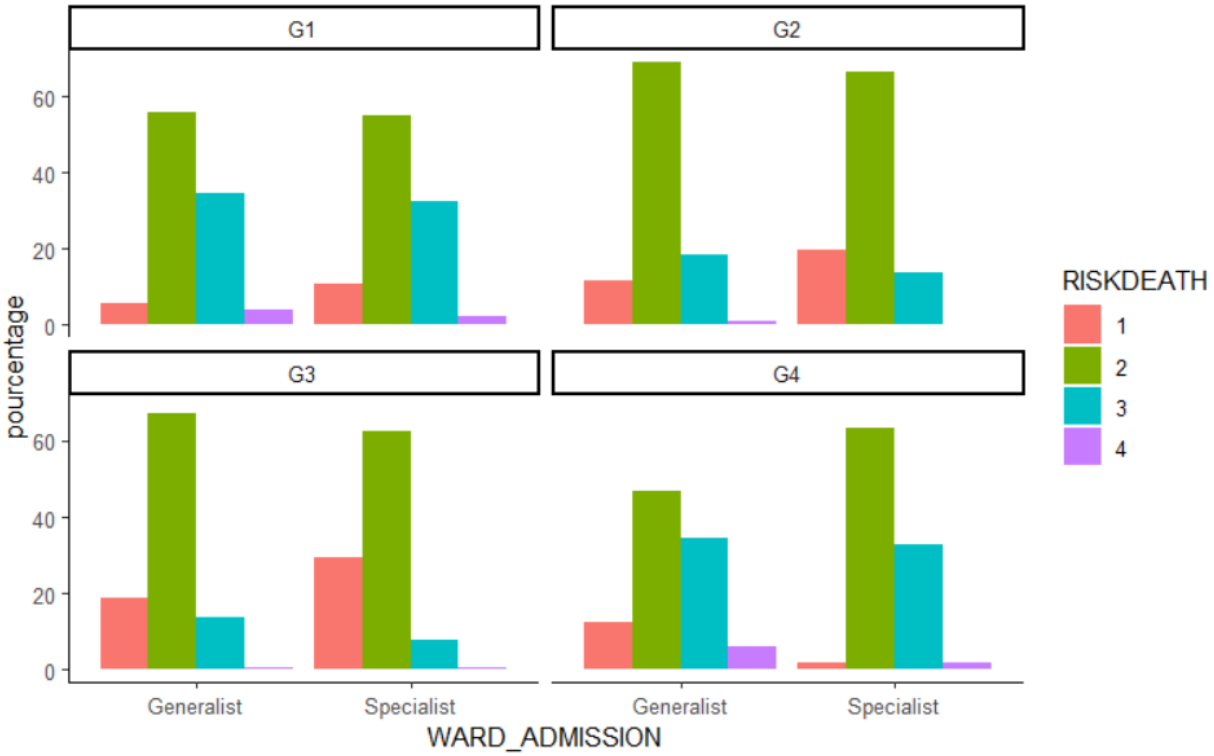
1-10 of 12 rows Previous 1 2 Next

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







- <u>Nous pouvons remarquer que pour le cas des TOTAL_COST :</u>
  - Les valeurs les plus élevées sont uniquement retrouvées au sein de G4, dans un groupe ou la variabilité est, elle aussi, particulièrement élevée et où la distribution des données est nettement asymétrique.

- Les BoxPlots correspondant aux groupes G1, G2 et G3 se trouvent plus ou moins sur le même niveau de TOTAL_COST, avec notamment des variabilités moins importantes comparées à celle de G4 et les distributions des données dans ces groupes sont relativement symétriques.
- Les données aberrantes sont les plus importantes et plus élevées en termes de valeurs de TOTAL_COST pour G4, tandis qu'elles sont similairement moins élevées et moins dispersées pour G1 et G2. Les données aberrantes sont quant à elles peu nombreuses et peu élevées en termes de valeurs de TOTAL_COST.

- Nous pouvons remarquer que pour le cas des RISK_DEATH :
  - Le pic de proportions dans chaque groupe correspond fortement (plus de 50% chacun), aux admissions avec un niveau de RISK_DEATH 2. Dans un ordre décroissant, le pic est plus élevé en G2, puis en G3, puis en G4 et légèrement moins élevé que dans les autres groupes en G1 ;
  - Les proportions d'admissions avec un niveau de RISK_DEATH 3 sont relativement similaires en G1 et G4, tandis qu'elles sont similairement moins élevées en G2 et G3.
  - Les admissions avec un niveau de RISK_DEATH 1 sont les plus fréquentes en G3, puis en G2, tandis qu'elles sont relativement peu fréquentes en G1 et G4 ;
  - Dans tous les groupes, les admissions avec un RISK_DEATH de niveau 4 sont peu fréquentes, voir même relativement rares pour les cas particuliers des groupes G2 et G3.
- Nous pouvons remarquer que pour le cas des AGE :
  - Les admissions correspondant aux patients les plus âgés sont observées au sein de G1 et G3, tandis que celles qui correspondent aux patients les *moins* âgées sont principalement concentrées en G2 qu'en G4.
  - Les variabilités sont plus importantes en G2 et G4, avec une asymétrie de répartition remarquée en G2. Les variabilités sont quant à elles nettement et similairement moins importantes en G1 et G3, avec notamment des distributions plus symétriques.
  - Des données aberrantes inférieures (patents *moins* âgés) sont constatées en G1 et d'autres en G3 (patients plus âgés).
- Nous pouvons remarquer que pour le cas des LOS:
  - Les admissions avec des valeurs de LOS les plus élevées sont concentrées en G4, avec la variabilité la plus importantes de toutes et une distribution de données symétrique ;
  - Les admissions concentrées en G1, G2 et G3 sont quant à elle caractérisées par les valeurs de LOS nettement et surtout similairement (entre ces 3 groupes) moins importantes qu'en G3, avec des variabilités largement moins importantes et des distributions de données relativement asymétriques.
  - Les données aberrantes présentent une similitude en G1, G2 et G3, tandis qu'elles sont , certes moins nombreuses en G4, mais par contre plus dispersées.
- Pour ce qui est des CCI par WARD_ADMISSION:
  - L'indice 3 de CCI correspond au pic des proportions d'admissions en G1 (nettement plus fréquentes dans les Wards Specialist) et en G4 : Wards Generalist.
  - L'indice 2 de CCI correspond au pic des proportions d'admissions en G2 (clairement plus fréquents chez les Wards Specialist), en G3 et en G4 : Wards Specialist.

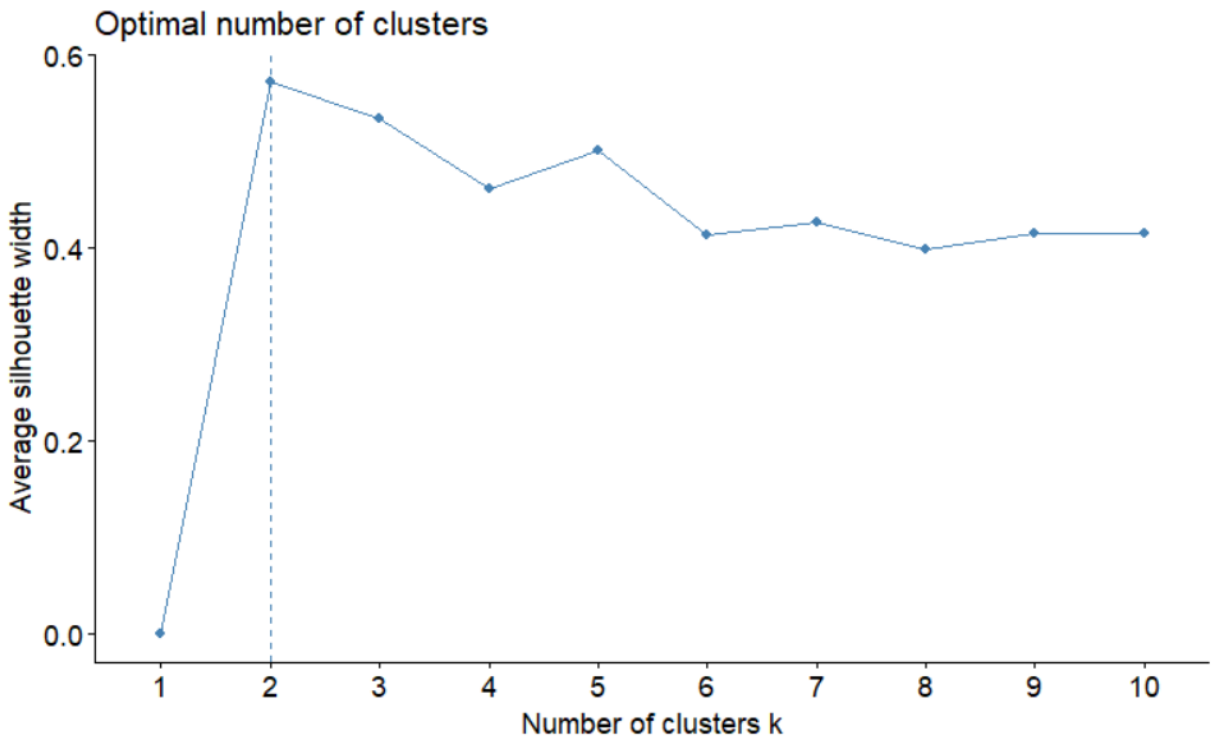
- L'indice 1 de CCI est plus fréquemment observé (d'une manière particulièrement similaire) chez les admissions, que ce soit en Wards Generalist ou Specialist, en G2 et G3. Il l'est moins en G4 et même quasi non-observé en G1 ;
- L'indice 4 de CCI s'observe le plus chez les admissions en G1 (plus chez les Generalist), un peu moins en G4 (encore une fois, principalement en Wards Generalist), et plus ou moins rare (ou même inexistant) dans les autres Wards des autres groupes.
- L'indice 5 de CCI est peu fréquent chez les admissions en G1 : Wards Generalist et très peu fréquent chez celles de G1 : Wards Specialist, très peu fréquent en G4 (Specialist & Generalist) et relativement inexistant dans les autres Wards de Groupe non-mentionnés.
- Les indices 6 et 7 de CCI ne s'observent que très rarement chez les admissions dans Wards : rarement constaté en G1 : Wards Generalist et G4 : Wards Specialist pour l'indice 6, et très peu constaté pour l'indice 7 en G4 : Wards Generalist.

- Pour ce qui est des RISK_DEATH par WARD_ADMISSION:
  - Dans tous les Wards (Generalist & Specialist), les pics de proportions correspondent nettement aux admissions avec un RISK_DEATH de niveau 2. Remarquons juste que les deux pics en G2 et G3 sont relativement plus importants que ceux en G1 et G4 ;
  - Remarquons que les admissions avec un RISK_DEATH de niveau 3 sont similairement assez fréquentes que ce soit dans les Wards Generalist ou Wards Specialist pour les Groupes G1 et G4, le sont moins pour G2 (Wards Generalist & Wards Specialist) et relativement peu fréquentes pour G3 (Wards Generalist & Wards Specialist) ;
  - Les admissions avec un RISK_DEATH 1 sont assez fréquentes dans les Wards Spécialist et Generalist du Groupe G3, légèrement moins le cas en G2. Ce même type d'admissions est peu fréquent en G1 : Wards Specialist, et très peu fréquents en G1 : Wards Generalist, une situation qui s'inverse pour le cas de G4.
  - Les admissions avec un RISK_DEATH 4 sont généralement peu fréquentes ou mêmes rares dans tous les Wards de chaque groupe, insistons juste un peu sur le fait que cette *moindre fréquence* l'est *moins* en G4 : Wards Generalist et G3 : Wards Generalist, similairement rare en G1 : Wards Specialist et G4 : Wards Specialist, et très rare dans les autres Wards non cités.

IV- Modèle 3 : Coûts des examens/analyses spécifiques

```
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")
```



- La méthode de partitionnement de données choisie est le partitionnement en k-moyennes (k-means clustering).
- La méthode utilisée pour la l'estimation du nombre optimal de Clusters à spécifier lors du partitionnement des données avec la méthode de k-means est celle de l'Average Silhouette.
- En se basant sur le graphique ci-dessus, il est clair que nous avons une valeur maximum de l'Average Silouhette width avec 2 clusters, signifiant que 2 est alors le nombre optimal de Clusters retenu.

```
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 2 2 2 1 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 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 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 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 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 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 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 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 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
 [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 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
 [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 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 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 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 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 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 1 1 1 1 2 2 2 1 1 1 1 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"
```

- Le nombre d'observations pour chacun des 2 clusters sera respectivement de : 2109 et 417 observations.
- Les **coordonnées (COST_RADIOLOGY, COST_LAB COST_HAEMATIC, COST_CONSULTATIONS, COST_CARDIO, COST_VAR COST_DIAGNOSTIC)** de chaque Centroid (position du centre de Cluster) de chaque Cluster sont les suivantes:
  - **Centroid de Cluster 1 : (-0.3516754, -0.1240645, -0.04644016, -0.03621574, -0.3516754, -0.2636288, -0.2992636) ;**
  - **Centroid de Cluster 2 : (1.7786175, 0.6274629, 0.23487361, 0.18316308, 1.7786175, 1.3333169, 1.5135417) ;**
- A partir du Clustering Vector ci-dessus, nous pouvons consulter le Cluster d'appartenance de chaque ligne d'observations sur les admissions en consultant juste le numéro de cluster qui représente la ligne d'observations sur la position de cette dernière dans le vecteur : ligne 1 appartient à Cluster 1, ligne 2 appartient à Cluster 1, ligne 988 appartient à Cluster 1…
- La Somme des carrées des distances des points d'un Cluster à leur Centroid pour chacun des Clusters 1 et 2sont respectivement de 4 673.694 et de 7 568.315. A partir de ces valeurs, nous pouvons en déduire que le Cluster 1 est le plus compact des deux, tandis que le Cluster 2 est le moins compact.
- Cependant, notons tout de même que le ratio ***" between_SS / total_SS = 30.7 % "*** n'est pas forcément (peut-être même loin de l'être) un bon signe que nous avons là un *Clustering suffisamment bon*.

Résidus :

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

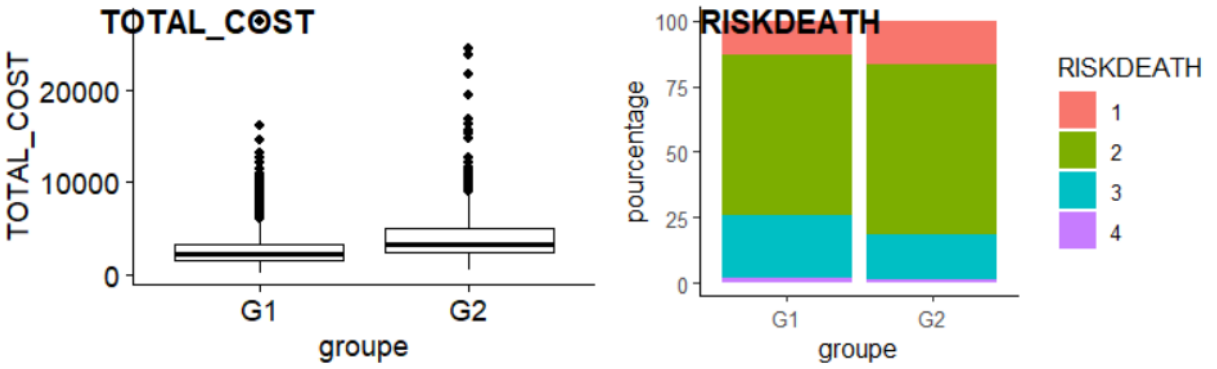
- Nous avons nommé nos Cluster 1 et 2 respectivement par « G1 » et « G2 ».

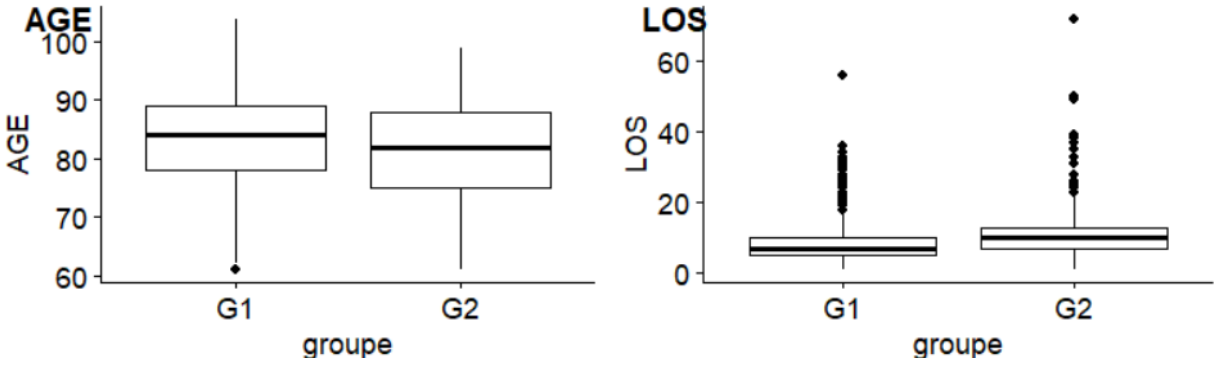
groupe <chr>	variables <chr>	mean <dbl>	sd <dbl>	min <dbl>	q1 <dbl>	median <dbl>	q3 <dbl>	max <dbl>	na <int>
G1	COST_CARDIO	2.024533	5.718183	0.00	0.00	0.00	0.00	46.48	0
G1	COST_CONSULTATIONS	40.315021	67.135490	0.00	0.00	0.00	46.48	583.64	0
G1	COST_DIAGNOSTIC	81.223167	58.749874	0.00	35.33	69.83	115.29	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

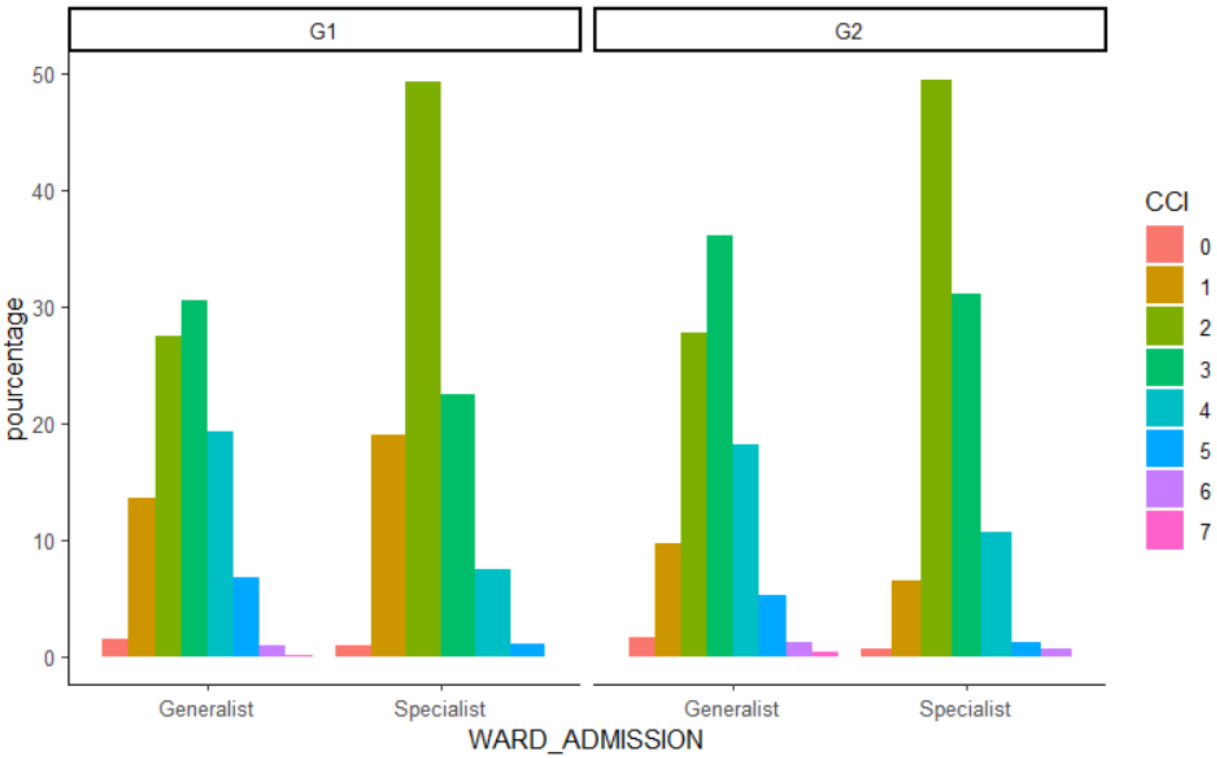
1-10 of 14 rows Previous 1 2 Next

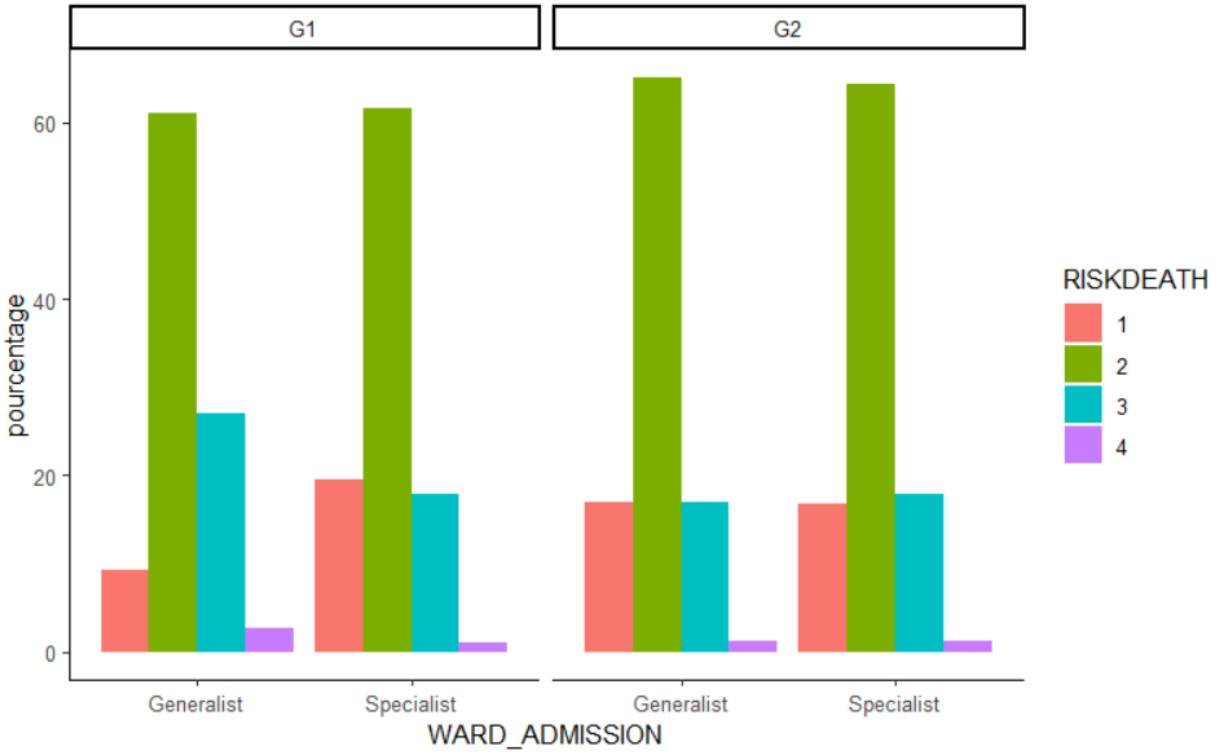
groupe <chr>	variables <chr>	mean <dbl>	sd <dbl>	min <dbl>	q1 <dbl>	median <dbl>	q3 <dbl>	max <dbl>	na <int>
G2	COST_HAEMATIC	6.503405	41.690301	0.00	0.00	0.00	0.00	556.04	0
G2	COST_LAB	144.342926	172.704129	0.00	39.52	89.07	170.46	1443.66	0
G2	COST_RADIOLOGY	64.144293	42.022785	0.00	60.43	60.43	72.05	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
100
75
50
25
0
pourcentage
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



- <u>Nous pouvons remarquer que pour le cas des TOTAL_COST :</u>
  - Les TOTAL_COST les plus élevés sont rencontrés parmi les admissions en G2, et les moins élevés en G1. ;
  - La variabilité est plus importante en G2 qu'en G1 et la distribution des données est plus asymétrique en G2 qu'en G1 ;
  - Les données aberrantes par rapport aux valeurs de TOTAL_COST sont plus importantes et dispersées en G2 qu'en G1.
- <u>Nous pouvons remarquer que pour le cas des RISK_DEATH :</u>
  - Les admissions avec un niveau de RISK_DEATH 2 sont relativement similaires en termes de proportion dans les Groupes G1 et G2 ;
  - Les admissions avec un niveau de RISK_DEATH 3, quant à elles, sont légèrement plus fréquentes en G1 qu'en G2, tandis que la situation s'inverse lorsqu'il s'agit des admissions avec un niveau de RISK_DEATH 1 ;
  - Les admissions avec un niveau de RISK_DEATH 4 sont relativement rares dans les deux groupes.
- <u>Nous pouvons remarquer que pour le cas des AGE :</u>
  - Les valeurs d'âges sont légèrement plus élevées en G1 qu'en G2 (les âges des patients admis les moins élevés y sont retrouvés dans ce dernier), avec cependant une variabilité plus importante pour G2 ;
  - La distribution de données est légèrement plus symétrique en G2 qu'en G1.
  - Aucune indication sur l'existence d'éventuelles données aberrantes n'est signalée en se basant sur les deux BoxPlots.
- <u>En ce qui concerne les LOS :</u>
  - Les valeurs de LOS sont légèrement plus élevées en G2 qu'en G1, tandis que la variabilité (peu importante) et la symétrie (relativement symétrique) de la distribution des données sont relativement similaires en ce qui concerne les deux groupes ;

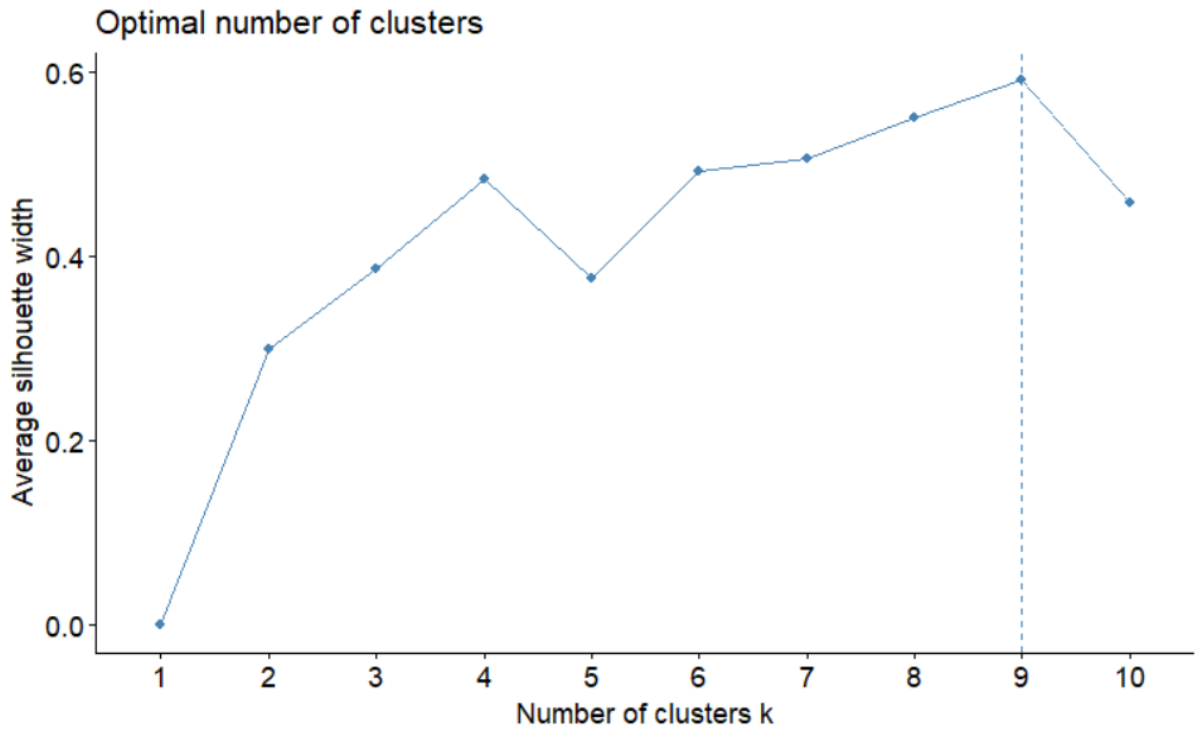
- Les données aberrantes sont plus éparpillées en G2 (avec les plus élevées en termes de valeurs) qu'en G1.

- En ce qui concerne les CCI par WARD_ADMISSION :
  - Dans les deux groupes, le pic des admissions en Wards Specialist (pics similaires) correspond à celle avec des indices 2 de CCI, tandis que ce sont les admissions avec les indices 3 qui constituent les pics en Wards Generalist (pic plus important en G2 : Wards Generalist) ;
  - Les admissions avec un indice de CCI 4 sont relativement similaires en terme de proportion en G1 : Wards Generalist et en G2 : Wards Generalist, tandis que dans les Wards Specialist, ce type d'admission est plus fréquent en G2 qu'en G1 ;
  - Les admissions avec un indice de CCI 5 sont peu fréquentes dans les deux groupes pour le cas des Wards Generalist (juste légèrement plus fréquentes en Groupe G1), tandis que dans les Wards Specialist, ce type d'admission est similairement très peu fréquent dans les deux groupes ;
  - Les admissions avec des indices de CCI 6 et 7 sont très peu fréquentes ou rares (voir même totalement absentes pour le cas des Wards Specialist en G1) dans tous les Wards des deux groupes.
- En ce qui concerne les RISK_DEATH par WARD_ADMISSION :
  - Que ce soit en G1 ou en G2, les pics dans tous les Wards correspondent aux admissions avec un RISK_DEATH 2 (généralement légèrement plus importants dans G2) ;
  - Les admissions avec un niveau RISK_DEATH de niveau 3 sont plus importantes en termes de proportions en G1 : Wards Generalist qu'en G2 : Wards Generalist, tandis que ce type d'admissions est assez similairement fréquent dans les Wards Specialist des deux groupes ;
  - Les admissions avec un niveau RISK_DEATH 1 sont légèrement plus fréquentes en G1 : Wards Specialist qu'en G2 : Wards Specialist, tandis que ce type d'admission est clairement moins fréquent en G1 : Wards Generalist qu'en G2 : Wards Generalist ;
  - Les admissions avec un niveau de RISK_DEATH 4 sont peu fréquentes dans tous les Wards des deux groupes, avec peut-être juste une attention particulière portée au cas de la proportion constatée en G1 : Wards Generalist qui semble y être légèrement plus importante que dans les autres Wards.

V- Modèle 4 : Département d'admission, risque de décès et durée de séjour

```
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")
```



- La méthode de partitionnement de données choisie est le partitionnement en k-moyennes (k-means clustering).
- La méthode utilisée pour la l'estimation du nombre optimal de Clusters à spécifier lors du partitionnement des données avec la méthode de k-means est celle de l'Average Silhouette.
- En se basant sur le graphique ci-dessus, il est clair que nous avons une valeur maximum de l'Average Silouhette width avec 9 clusters, signifiant que 9 est alors le nombre optimal de Clusters retenu.

```
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 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 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"
```

- Le nombre d'observations pour chacun des 9 clusters sera respectivement de : 490, 515, 283, 266, 634, 110, 107, 92 et de 29 observations.
- Les **coordonnées (Risk_Death_1, Risk_Death_2, Risk_Death_3, Risk_Death_4, WS08, WS21, WS24, WS2604, WS2605, WS68)** de chaque Centroid (position du centre de Cluster) de chaque Cluster sont les suivantes:
  - **Centroid de Cluster 1 : (0.1877551, 0.6897959, 0.1102041, 0.01224490, 0.000000000, 0.0000000, 0.00000000, 1, 0.0000000, 0.00000000) ;**
  - **Centroid de Cluster 2 : (0.0000000, 0.0000000, 1.0000000, 0.00000000, 0.009708738, 0.1825243, 0.06601942, 0, 0.6718447, 0.06990291) ;**
  - **Centroid de Cluster 3 : (0.2190813, 0.7526502, 0.0000000, 0.02826855, 0.000000000, 1.0000000, 0.00000000, 0, 0.0000000, 0.00000000) ;**
  - **Centroid de Cluster 4 : (0.0000000, 1.0000000, 0.0000000, 0.00000000, 0.000000000, 0.0000000, 0.14285714 0, 0.0000000, 0.85714286).**
  - **Centroid de Cluster 5 : (0.0000000, 1.0000000, 0.0000000, 0.00000000, 0.000000000, 0.0000000, 0.00000000, 0, 1.0000000, 0.00000000).**
  - **Centroid de Cluster 6 : (0.0000000, 1.0000000, 0.0000000, 0.00000000, 1.000000000, 0.0000000, 0.00000000, 0, 0.0000000, 0.00000000).**
  - **Centroid de Cluster 7 : (0.6915888, 0.0000000, 0.0000000, 0.30841121, 0.000000000, 0.0000000, 0.00000000, 0, 1.0000000, 0.00000000).**
  - **Centroid de Cluster 8 : (1.0000000, 0.0000000, 0.0000000, 0.00000000, 1.000000000, 0.0000000, 0.00000000, 0, 0.0000000, 0.00000000).**

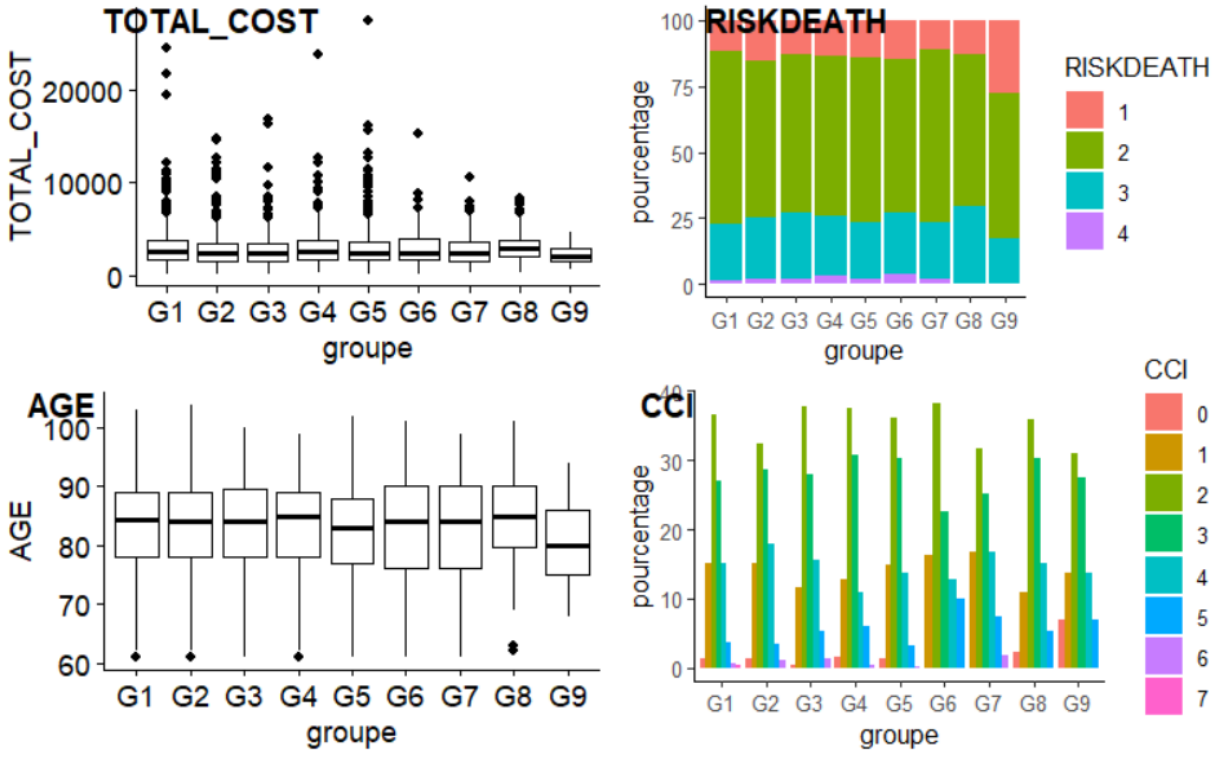
- **Centroid de Cluster 9 : (0.9310345, 0.0000000, 0.0000000, 0.06896552, 0.000000000, 0.0000000, 0.58620690, 0, 0.0000000, 0.41379310).**

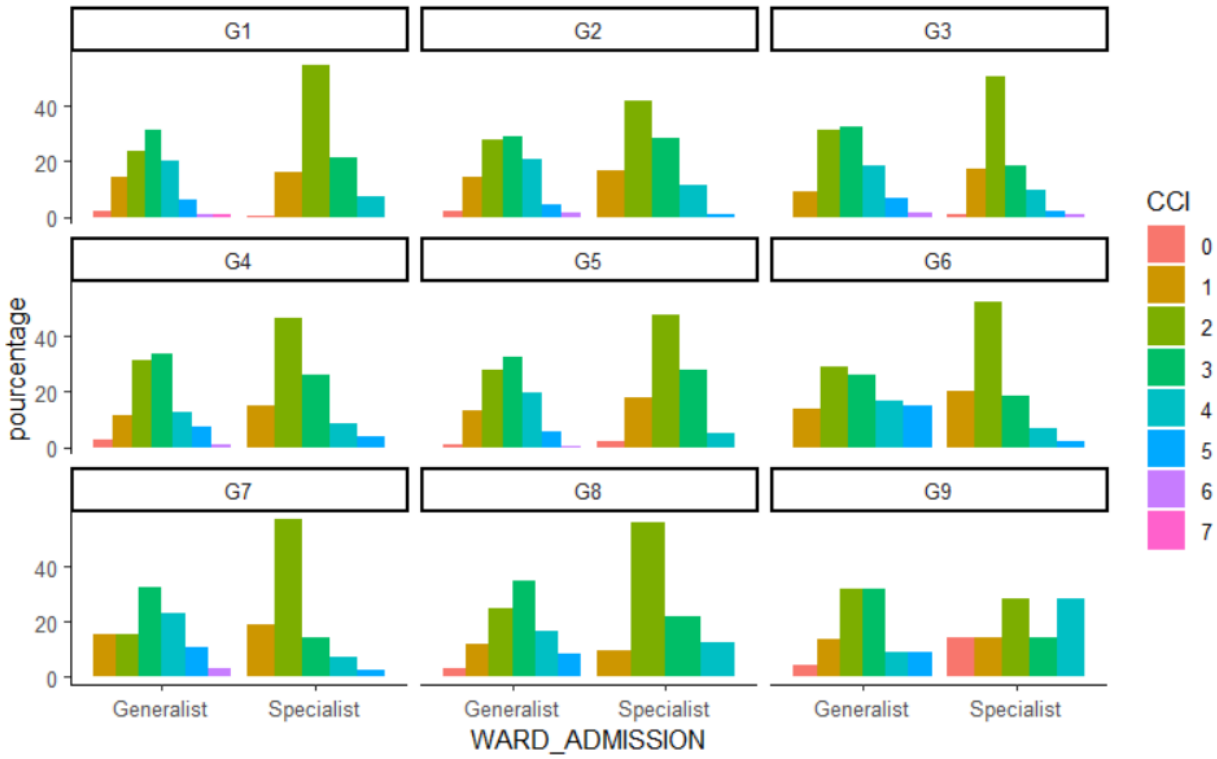
- A partir du Clustering Vector ci-dessus, nous pouvons consulter le Cluster d'appartenance de chaque ligne d'observations sur les admissions en consultant juste le numéro de cluster qui représente la ligne d'observations sur la position de cette dernière dans le vecteur : ligne 1 appartient à Cluster 1, ligne 2 appartient à Cluster 2, ligne 988 appartient à Cluster 5…
- La Somme des carrées des distances des points d'un Cluster à leur Centroid pour chacun des Clusters 1, 2, 3, 4, 5, 6, 7, 8 et 9 sont respectivement de : 233.55102, 260.57476, 108.87633, 65.14286, 0.00000, 0.00000, 45.64486, 0.00000 et de 17.79310. A partir de ces valeurs, nous pouvons en classer dans l'ordre décroissant des Clusters, du moins compact au plus compact : **Cluster 5 – Cluster 6 – Cluster 8 – Cluster 9 – Cluster 7 – Cluster 4 – Cluster 3 – Cluster 1 – Cluster 2** .
- Le ratio ***"between_SS / total_SS = 77.4 %"*** peut nous indiquer que nous avons là un Clustering *relativement bon*.

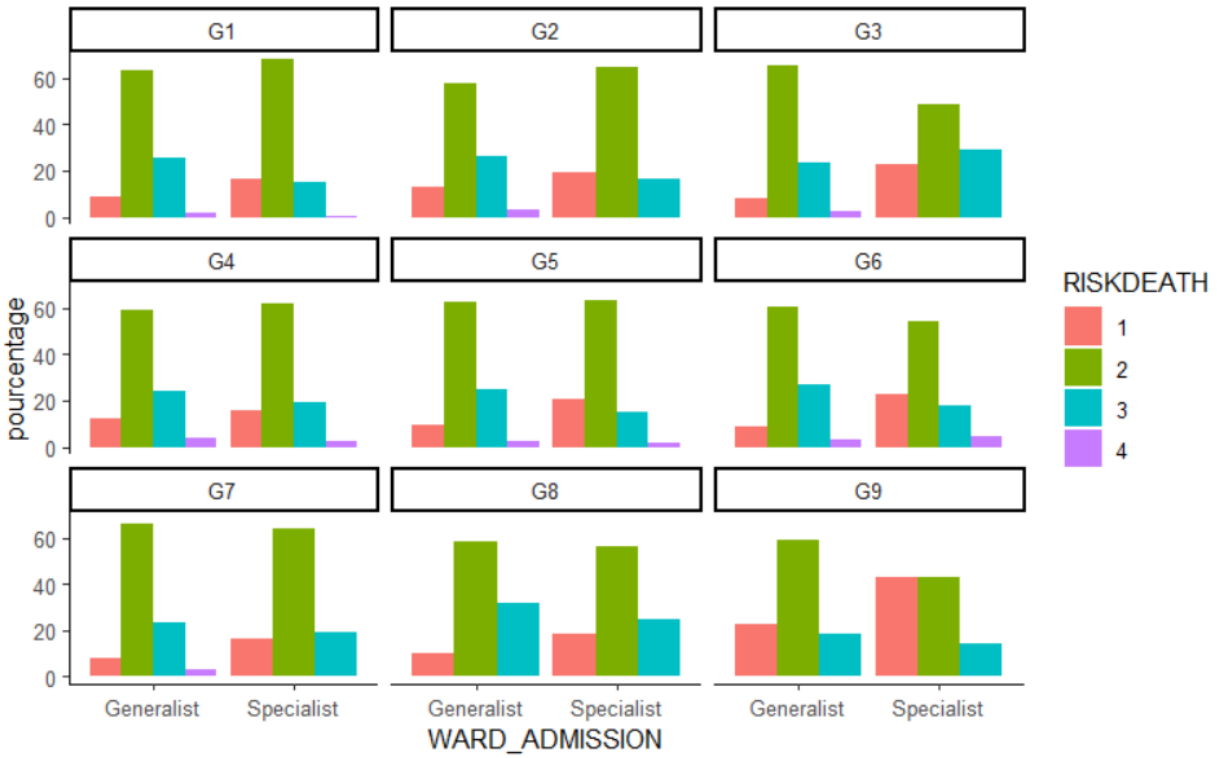
Résidus :

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

- Nous avons nommé nos Clusters respectivement par « G1 », « G2 », « G3 », « G4 », « G5 », « G6 », « G7 », « G8 » et « G9 ».







- <u>Nous pouvons remarquer que pour le cas des TOTAL_COST:</u>
  - Les BoxPlots correspondant aux Groupes se trouvent à peu près aux mêmes niveaux de valeurs de TOTAL_COST, avec quelques exceptions près pour les cas du G1, G6 et G8 ayant les valeurs les plus élevées et pour ceux de G7 et G9 avec les valeurs les moins élevées ;

- Les variabilités sont plus importantes pour G6 et G7, tandis qu'elles le sont peu avec G9 ;
- En ce qui concerne la symétrie de la distribution des données, tous les BoxPlots des Groupes semblent montrer une relative symétrie, à l'exception de celui du G7, présentant une asymétrie assez importante.
- Les données aberrantes sont plus importantes au sein de G1, G2 et de G5 (correspondant notamment à la valeur aberrante la plus élevée constaté), tandis qu'elles le sont nettement moins au sein de G6, G7 et de G8, voir même inexistantes au sein de G9.

- Nous pouvons remarquer que pour le cas des RISK_DEATH:
  - Pour tous les groupes, les admissions avec le niveau de RISK_DEATH 2 constituent le pic des proportions.
  - Les admissions avec un niveau de RISK_DEATH 1 sont particulièrement fréquentes en G9, tandis que ce type d'admissions est moins fréquent au sein de G1 et G7 ;
  - Les admissions avec un niveau de RISK_DEATH 3 sont particulièrement fréquentes au sein de G8, tandis que les autres groupes présentent chacune une proportion relativement similaire lorsqu'il s'agit des admissions liées à ce risque.
  - Les admissions avec un RISK_DEATH de niveau 4 sont généralement peu fréquentes (G4 et G6), rares (G1, G2, G3, G5, G7) ou même inexistantes (G8 et G9) dans tous les groupes.
- Pour ce qui est du cas des AGE:
  - Pour les cas des valeurs d'âges les plus élevées, nous pouvons constater à partir des BoxPlots correspondants que les Groupes G1, G2, G3, G4, G6, G7 et G8 sont relativement concernés par ces valeurs, tandis que pour le cas des valeurs moins élevées, ce sont les groupes G6 (avec une grande variabilité, distribution de données légèrement asymétrique), G7 (avec une grande variabilité, distribution de données légèrement asymétrique) et G9 qui sont les plus concernés.
  - A part les groupes G6 et G7 déjà mentionnés précédemment, les variabilités et les distributions de données dans les autres groupes semblent toutes être à peu près similaires les unes aux autres.
  - Quelques données aberrantes peuvent être remarquées au sein de G1, G2, G4 et de G8.
- Pour ce qui est du cas des CCI:
  - Dans quasiment tous les groupes, les admissions avec un indice de CCI 2 constituent le pic de proportion (pics légèrement plus importants en G6, G3 et G4, et moins importants en G9, G7 et G2) ;
  - Les admissions avec un CCI 3 sont généralement assez fréquentes au sein de groupes (moins importantes tout de même en G6 et G7 comparés aux proportions des autres groupes) ;
  - Les admissions avec des CCI 5 et 4 sont quant à elles généralement moins fréquentes dans les groupes (particulièrement au sein de G1 et de G2, mais aussi au sein de G5 pour le cas du CCI 5) ;
  - Les admissions avec des CCI 6 et 7 demeurent plus ou moins peu fréquentes dans tous les groupes, voir même inexistantes dans quelque uns (G9, G8, G6).
- Pour ce qui est du cas des CCI par WARD_ADMISSION:

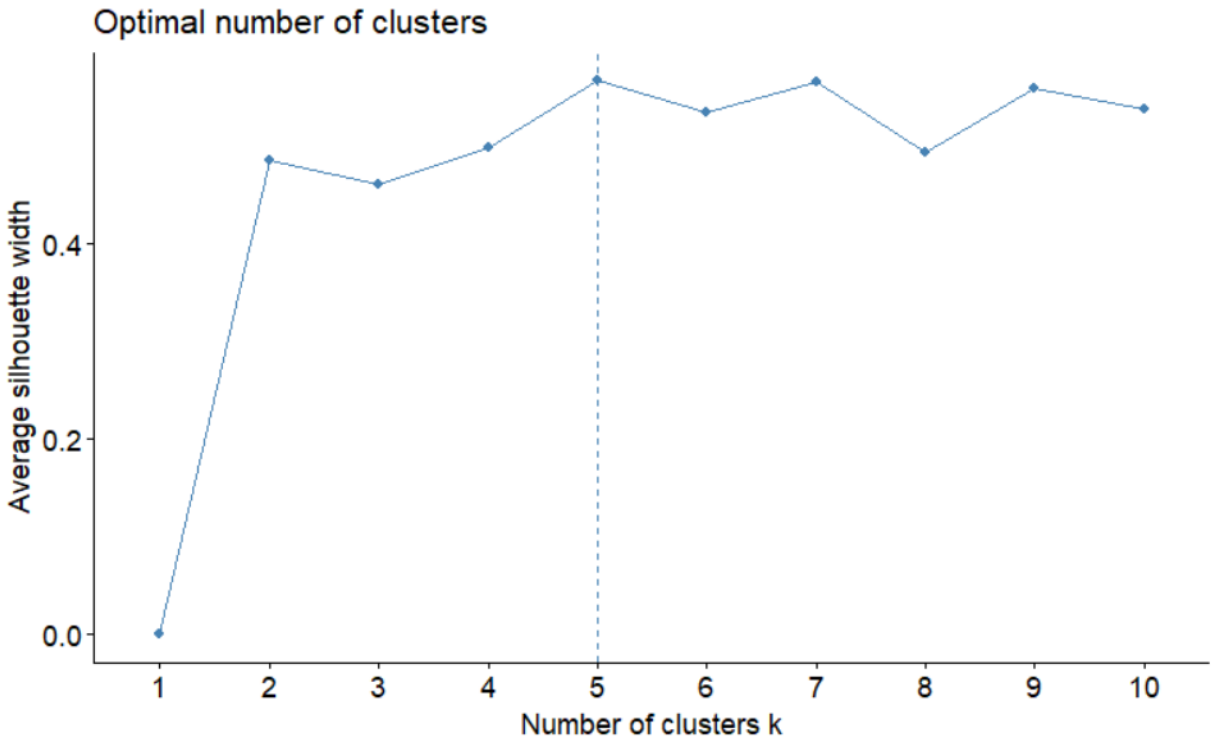
- Pour tous les Wards Specialist de chaque groupe, nous remarquons clairement que les admissions avec un indice CCI 2 constituent les pics des proportions (pic particulièrement moins important en G9 : Wards Specialist) ;
- Pour tous les Wards Generalist de chaque groupe, ce sont les admissions avec un indice de CCI 3 qui forment les pics de proportions (pic exceptionnellement moins important pour le cas de G9) ;
- Excepté en G9 : Wards Specialist (exceptionnellement plus fréquentes), les admissions avec des CCI 4 et 5 sont relativement moins fréquentes au sein des Wards Specialist des groupes ; Tandis que pour le cas des Wards Generalist, à l'exception de G6, ces types d'admissions sont moins fréquents au sein des groupes.
- Les admissions correspondant à des CCI 6 et 7 sont plutôt rares quant à leur part dans la majorité des Wards de chaque groupe, voir même inexistantes dès fois.

- Pour ce qui est du cas des RISK_DEATH par WARD_ADMISSION:
  - Que ce soit dans les Wards Generalist ou Specialist, les admissions avec un RISK_DEATH 2 constituent toujours les pics de proportions ;
  - A l'exception de G3, les admissions avec un RISK_DEATH 3 sont plus fréquentes en Wards Generalist qu'en Wards Specialist dans tous les groupes ;
  - Les admissions avec un RISK_DEATH 1 sont généralement plus fréquentes en Wards Specialist qu'en Ward Generalist, avec une attention particulière portée au groupe G9, là où ce type d'admission est exceptionnellement plus fréquent en Wards Specialist.
  - Les admissions avec un RISK_DEATH 4 sont peu fréquentes, rares, voir même inexistantes (G8 et G9) dans les Wards des groupes.

VI- Modèle 5 : Durée de séjour et risque de décès

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

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



- La méthode de partitionnement de données choisie est le partitionnement en k-moyennes (k-means clustering).
- La méthode utilisée pour l'estimation du nombre optimal de Clusters à spécifier lors du partitionnement des données avec la méthode de k-means est celle de l'Average Silhouette.
- En se basant sur le graphique ci-dessus, il est clair que nous avons une valeur maximum de l'Average Silouhette width avec 5 clusters, signifiant que 5 est alors le nombre optimal de Clusters retenu.

```
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 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"
```

- Le nombre d'observations pour chacun des 5 clusters sera respectivement de : 568, 337, 913, 553 et de 155 observations.
- Les **coordonnées (LOS, RISKDEATH)** de chaque Centroid (position du centre de Cluster) de chaque Cluster sont les suivantes:
  - **Centroid de Cluster 1 : 0.5084901, -0.2079862) ;**
  - **Centroid de Cluster 2 : (-0.3391334, -1.7295931) ;**
  - **Centroid de Cluster 3 : (-0.5555552, -0.1944729) ;**
  - **Centroid de Cluster 4 : (-0.1548816, 1.4572387) ;**
  - **Centroid de Cluster 5 : (2.6989485, 0.4690952) ;**
- A partir du Clustering Vector ci-dessus, nous pouvons consulter le Cluster d'appartenance de chaque ligne d'observations sur les admissions en consultant juste le numéro de cluster qui représente la ligne d'observations sur la position de cette dernière dans le vecteur : ligne 1 appartient à Cluster 1, ligne 2 appartient à Cluster 1, ligne 988 appartient à Cluster 4…
- La Somme des carrées des distances des points d'un Cluster à leur Centroid pour chacun des Clusters 1, 2, 3, 4 et 5 sont respectivement de 126.5125, 114.9211, 103.1526, 369.1019 et de 450.9033. A partir de ces valeurs, nous pouvons classer dans l'ordre décroissant les Clusters, du plus compact au moins compact : **Cluster 3 – Cluster 2 – Cluster 1 – Cluster 4 – Cluster 5**.
- **Le ratio *" between_SS / total_SS = 76.9 %"*** peut nous indiquer que nous avons là un Clustering relativement bon.

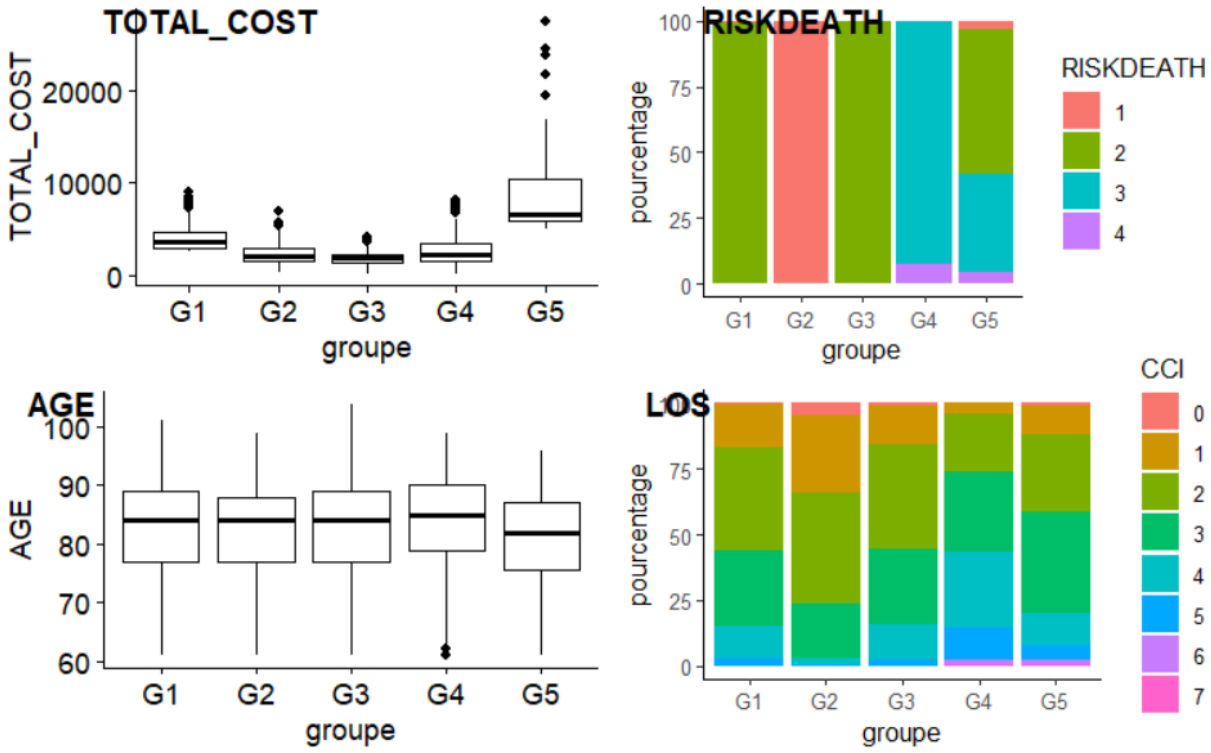
Résidus :

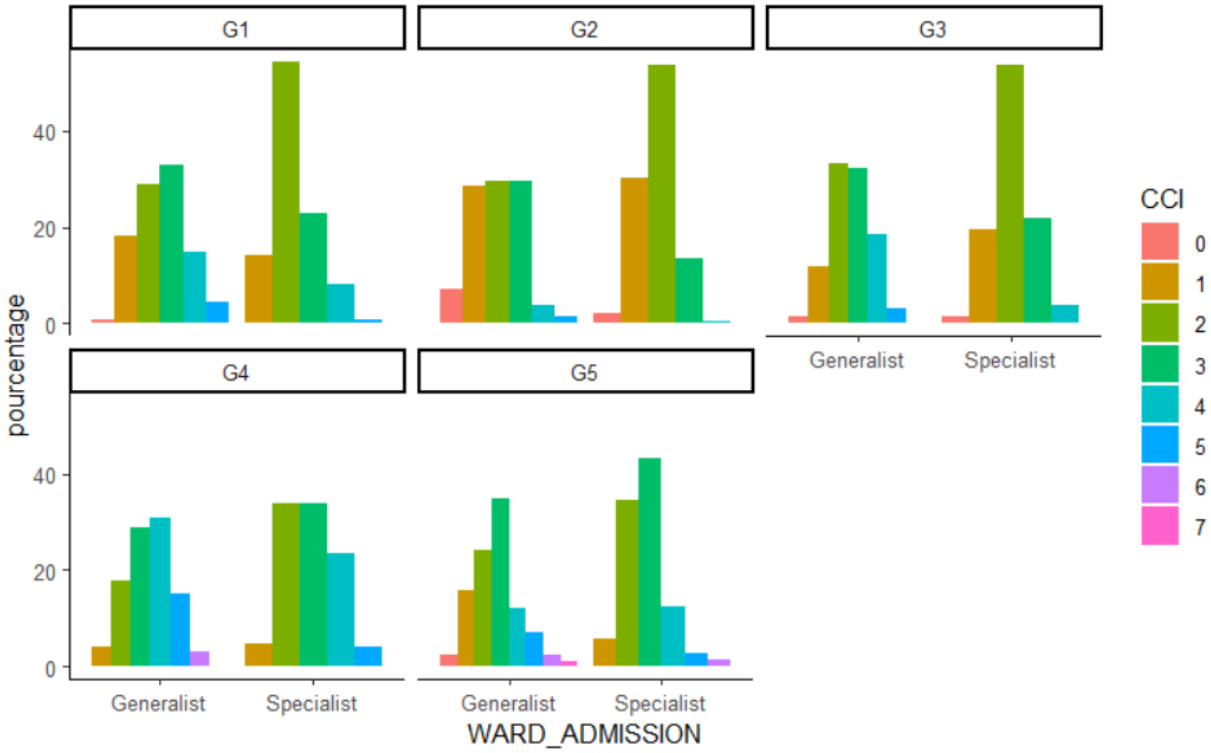
```
res.mod5 <- df %>%
  mutate(groupe= paste0('G',mod5$cluster))
```

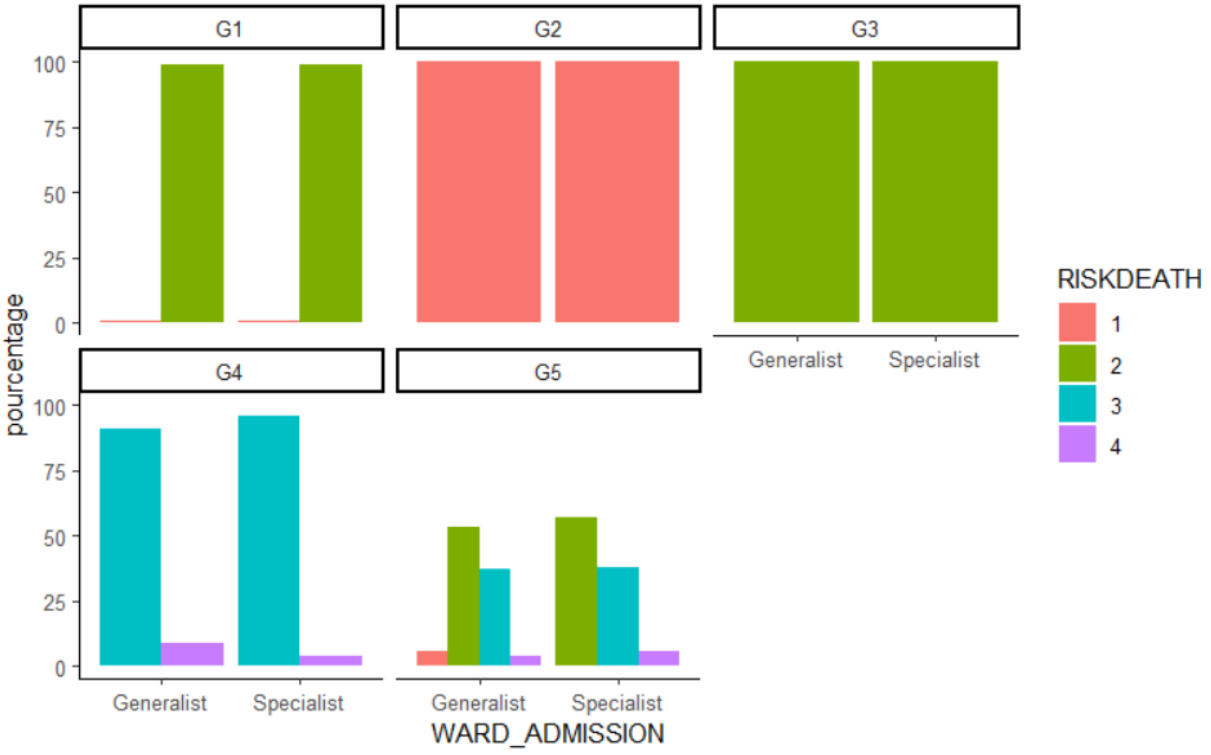
- Nous avons nommé nos Clusters respectivement par « G1 », « G2 », « G3 », « G4 » et « G5 ».

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







- <u>Nous pouvons remarquer que pour le cas des TOTAL_COST:</u>
  - Le groupe G5 regroupe nettement les valeurs les plus élevées de TOTAL_COST, avec une variabilité particulièrement importante et une distribution très asymétrique des données ;

- Avec la plus petite des variabilités, le Groupe G3 semble contenir les valeurs les moins élevées de TOTAL_COST ;
- Les Groupes G2 et G4 sont relativement aux mêmes niveaux en termes de valeurs de TOTAL_COAST, avec une variabilité légèrement plus importante chez G4 ;
- Des valeurs aberrantes peuvent être constatées dans tous les groupes, relativement moins conséquentes et moins élevées en G1, G2, G3 et G4, et plutôt importantes et plus dispersées en G5.

- Nous pouvons remarquer que pour le cas des RISK_DEATH:
  - Les groupes G1 et G3 sont principalement et largement constitués d'amissions avec un RISK_DEATH de niveau 2, un type d'admission qui est cependant moins fréquent en G5 ;
  - Le Groupe G2 est exceptionnellement constitué d'admission avec un RISK_DEATH 1, un type d'admission qui n'est cependant qu'en très petite quantité en G5;
  - Les admissions avec un RISK_DEATH 3 sont très fréquentes en G4, et sont moins fréquentes en G5 ;
  - Les admissions avec un RISK_DEATH 4 ne sont que très peu observées en G4 et G5.
- Nous pouvons remarquer que pour le cas des AGE:
  - Les Boxplots correspondant aux cinq groupes se trouvent à peu près au même niveau de valeurs d'AGE, avec à peu près une même variabilité et des distributions de données relativement asymétriques.
  - Cependant, nous pouvons quand même remarquer que les plus grandes valeurs d'âges sont observées en G4, G1 et G3, et les plus petites à peu près en G5.
  - Quelques données aberrantes peuvent être observées en G4.
- Pour le cas des CCI:
  - Globalement, chaque groupe est constitué principalement d'admissions avec un indice de CCI 3, puis d'admissions avec un indice de CCI 2.
  - Les admissions avec un CCI 1 sont exceptionnellement plus fréquentes en G2 que dans les autres groupes, tandis que pour celles avec un CCI 4, elles sont plus fréquentes en G4 que dans les autres groupes.
  - Les admissions avec un CCI 5 ne sont constatées qu'en petites proportions dans les Groupes G5, G4 et G1.
  - Les admissions correspondant à un CCI 0 ne sont que très peu rencontrées en G2 ;
  - Les admissions correspondant à un CCI 6 ou à un CCI 7 ne sont que très rarement observées en G4 et G5.
- Pour le cas des CCI par rapport aux WARD_ADMISSION:
  - En ce qui concerne les Wards Specialist, les pics de proportions sont constituées par les admissions avec un CCI 2 dans G1, G2 et G3, alors que dans G4 et G5, celles avec un CCI 3 semblent prendre le dessus ;
  - En ce qui concerne les Wards Generalist, les pics sont constitués par les admissions avec un CCI 3 dans G1, G2 (légèrement) et G5. Dans G3, ce sont celles avec un CCI 2 qui prennent le dessus, tandis que, exceptionnellement, dans G4, ce sont celles avec un CCI 4 qui sont dominantes.
  - Les admissions avec un CCI 4 ou 5 constituent des proportions plus ou moins importantes dans les Wards Generalist des Groupe G1, G3, G4 et G5 ; Ce même type d'admissions se rencontre exceptionnellement en bonne proportion dans G4 : Wards

Specialists, ce qui n'est pas pourtant un comportement observé au niveau des Wards Specialists des autres groupes ;

- Les admissions avec des CCI 6 et CCI 7 ne se rencontrent que plus ou moins rarement dans les groupes ;
- Les admissions avec un CCI 0 aussi ne se rencontrent que rarement dans les groupes, principalement en G2 : Wards Generalist.

- Pour le cas des RISK_DEATH par rapport aux WARD_ADMISSION:
  - Etonnement, les Wards en G2 ne contiennent que des admissions avec un RISK_DEATH de niveau 1, alors que ce type d'admission ne concerne qu'une petite proportion des G5 : Wards Generalist et d'infimes proportions en G1 : Wards Generalists et G1 : Wards Specialists ;
  - A nouveau, étonnement, les Wards en G3 ne contiennent que des admissions avec un RISK_DEATH de niveau 2. Un type d'admission qui constitue aussi les pics de proportions en G1 : Wards Generalist et G1 : Wards Specialist, mais aussi en G5 : Wards Generalist et G5 : Wards Specialist
  - Les admissions avec un RISK_DEATH de niveau 3 constituent les pics de proportions en G4 : Wards Generalist et G4 : Wards Specialist, et constituent aussi de bonnes proportions des admissions en G5 : Wards Generalist et G5 : Wards Specialist ;
  - Les admissions avec un RISK_DEATH de niveau 4 ne se rencontrent que peu fréquemment (G4 : Wards Generalist & G5 : Wards Specialist), rarement (G4 : Wards Specialist & G5 : Wards Generalist), ou même jamais (G1, G2 et G3) au sein des groupes.