Analyse de données hospitalières – EDA

I- Analyse univariée :

I.1- Distribution du DRG Code :

DRG_CODE <chr>	n <int>	% <dbl>
087	1166	42.757609
576	626	22.955629
127	591	21.672167
089	234	8.580858
090	110	4.033737

5 rows

- La majorité des admissions correspond principalement à un DRG 087 : *«Pulmonary edema and respiratory failure »* (en grande partie), 576 : *«Sepsis without medical ventilation» ou* 127 : *«Heart failure and shock»*.
- Celles qui correspondent à un DRG 089 *«Pneumonia and pleuritis with complications »* et 090 *«Pneumonia and pleuritis >17 year old »* constituent chacune des plus petites proportions.

I.2- Distribution du département d'admission (WARD_ADMISSION) :

WARD_ADMISSION <chr>	n <int>	% <dbl>
2605	1133	41.547488
2604	525	19.251925
21	377	13.824716
68	339	12.431243
08	219	8.030803
24	134	4.913825

6 rows

- Les admissions en Ward *« General Medicine »* sont majoritaires : une grande partie en Ward 2605 : *« General Medicine 2 »* et moins conséquente en Ward 2604 : *« General Medicine 1 »*.
- Celles en Wards 21 : *« Geriatrics »*, 68 : *«Respiratory Medicine »*, 08 : *« Cardiology »* et Ward 24 : *«Infection and Immunology»* constituent chacune des plus petites proportions.

I.3- Distribution du département de sortie (WARD_DISCHARG)

WARD_DISCHARG <chr>	n <int>	% <dbl>
2605	1133	41.547488
2604	525	19.251925
21	377	13.824716
68	339	12.431243
08	219	8.030803
24	134	4.913825

6 rows

- Les discharges depuis les Wards *« General Medicine »* sont majoritaires : une grande partie en Ward 2605 : *« General Medicine 2 »* et moins conséquente en Ward 2604 : *« General Medicine 1 »*.
- Celles en Wards 21 : *« Geriatrics »*, 68 : *«Respiratory Medicine »*, 08 : *« Cardiology »* et Ward 24 : «*Infection and Immunology*» constituent chacune des plus petites proportions.

I.4- Distribution du nombre de décès en établissement (DEATH_INHOSP)

DEATH_INHOSP <dbl>	n <int>	% <dbl>
0	2384	87.42208
1	343	12.57792

2 rows

- Les décès surviennent principalement (en très grande partie) en dehors des hôpitaux, ceux qui surviennent en milieux hospitaliers ne constituent qu'une petite proportion.

I.5- Distribution des coûts par département (COST_DAY_WARD)

COST_DAY_WARD <dbl>	n <int>	% <dbl>
285	1658	60.799413
313	377	13.824716
463	339	12.431243
363	219	8.030803
481	134	4.913825

5 rows

- Les coûts hospitaliers journaliers en Ward évalués à 285 € constituent la grande majorité.
- Ceux évalués à 313 €, 463 €, 363 € et à 481 € constituent chacune des plus petites proportions.

I.6- Distribution des sévérités (SEVERITY)

SEVERITY <dbl>	n <int>	% <dbl>
2	1672	61.312798
3	599	21.965530
1	402	14.741474
4	54	1.980198

4 rows

- Les admissions de patients avec des risques de sévérités de niveau 2 : *« Moderate »* sont nettement majoritaires
- S'en suivent les admissions avec des risques de sévérités de niveaux 3 : *« Major»* et 1 : *« Minor »*.
- Celles avec des risques de sévérités de niveau 4 : *« Severe »* sont quant à elles peu nombreuses.

I.7- Distribution du risque de décès en établissement (RISKDEATH)

RISKDEATH <dbl>	n <int>	% <dbl>
2	1672	61.312798
3	599	21.965530
1	402	14.741474
4	54	1.980198

4 rows

- Les admissions de patients avec des risques de décès de niveau 2 : *« Moderate »* sont nettement majoritaires
- S'en suivent les admissions avec des risques de décès de niveau 3 : *« High »* et 1 : *« Low »*
- Celles avec des risques de décès de niveau 4 : *« Extreme »* sont quant à elles peu nombreuses.

I.8- Distribution selon le sexe (SEX)

SEX <dbl>	n <int>	% <dbl>
2	1465	53.72204
1	1262	46.27796

2 rows

- Les femmes admises dans les hôpitaux sont plus nombreuses que les hommes.

I.9- Distribution du nombre d'interventions (N_INTERVENTION)

N_INTERVENTION <dbl>	n <int>	% <dbl>
3	489	17.931793
7	470	17.235057
4	457	16.758343
5	417	15.291529
6	323	11.844518
0	269	9.864320
2	170	6.233957
1	83	3.043638
8	49	1.796846

9 rows

- La grande majorité des admissions nécessite entre 3 à 7 interventions chirurgicales.
- Les admissions ne nécessitant aucune intervention chirurgicale constituent une proportion relativement *intéressante* comparée à celles qui ne nécessitent que 1 à 2 interventions.
- Les admissions nécessitant jusqu'à 8 interventions chirurgicales quant à elles sont peu nombreuses.

I.10- Distribution selon l'indice de comorbidité de Charlson (CCI)

CCI <dbl>	n <int>	% <dbl>
2	993	36.41364136
3	752	27.57609094
1	425	15.58489182
4	385	14.11807847
5	113	4.14374771
0	40	1.46681335
6	17	0.62339567
7	2	0.07334067

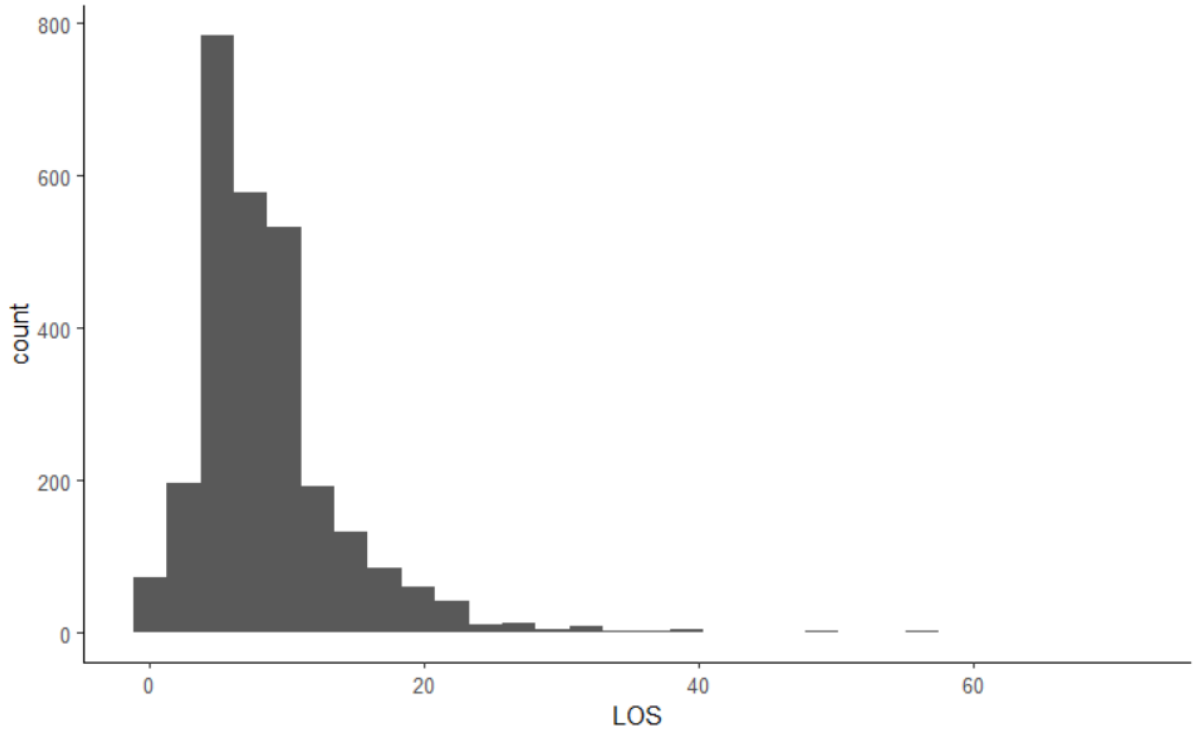
8 rows

- La grande majorité des admissions correspond à un indice de CCI allant généralement de 1 à 4, avec un *intérêt* particulier sur les scores 2 et 3 qui à eux deux concernent nettement plus de la moitié des admissions.
- Certes moins nombreuses, les admissions correspondant à un score de CCI égal à 5 s'avère être relativement importante en termes de proportion.
- Particulièrement, les admissions correspondant à un indice de CCI égal à 0 sont relativement peu nombreuses pour leur part.

- Celles qui correspondent à un indice supérieur ou égal à 6 peuvent être considérés comme étant peu nombreuses, voir même rares pour celles avec un indice de 7.

I.11- Les durées de séjour en établissement (LOS)

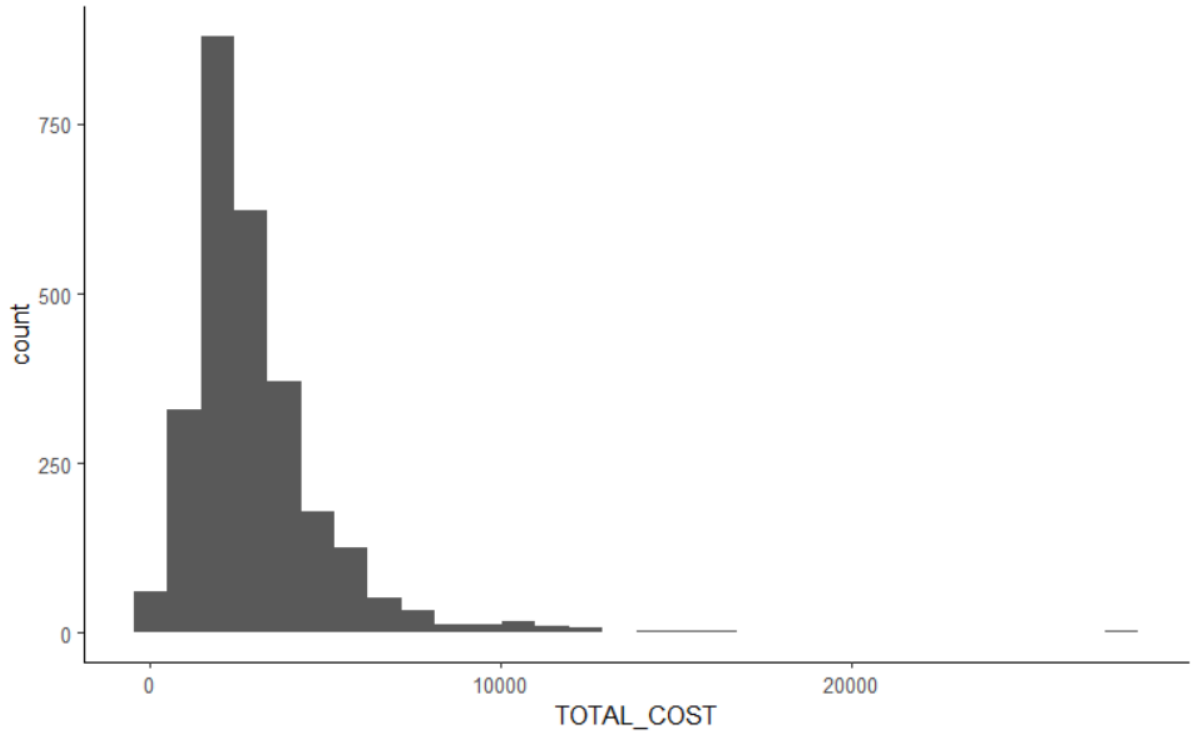
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   5.000   7.000   8.666  11.000  72.000
```



- Le pic correspondant aux LOS se situe clairement entre 5 et 10 jours (histogramme), tandis que la dispersion des données elle, s'étend de 1 à 75 jours environs.
- La valeur de LOS minimum est de 1 jour, tandis que la valeur de LOS maximum est de 72 jours : Une différence importante (de 71 jours) peut être observée entre les deux valeurs, la valeur de LOS maximum est très élevée.
- Le 1er Quartile des valeurs de LOS est de 5 jours, autrement dit, 25% des valeurs des LOS sont inférieures ou égales à 5 jours.
- Le 2nd Quartile (Médiane) des valeurs de LOS est de 7 jours, autrement dit, 50 % des valeurs des LOS sont inférieures ou égales à 7 jours, et les autres 50% sont supérieures ou égales à 7 jours.
- Le 3^{è} Quartile des valeurs de LOS est de 11 jours, autrement dit, 75% des valeurs des LOS sont inférieures ou égales à 11 jours.
- La Moyenne des valeurs de LOS est de 8,666 jours (*des valeurs de LOS aberrantes peuvent donc être présentes parmi les données*), soit supérieure à la Médiane qui elle, est de 7 jours : Nous avons donc une asymétrie vers la droite.

I.12- Les coûts totaux (TOTAL_COST)

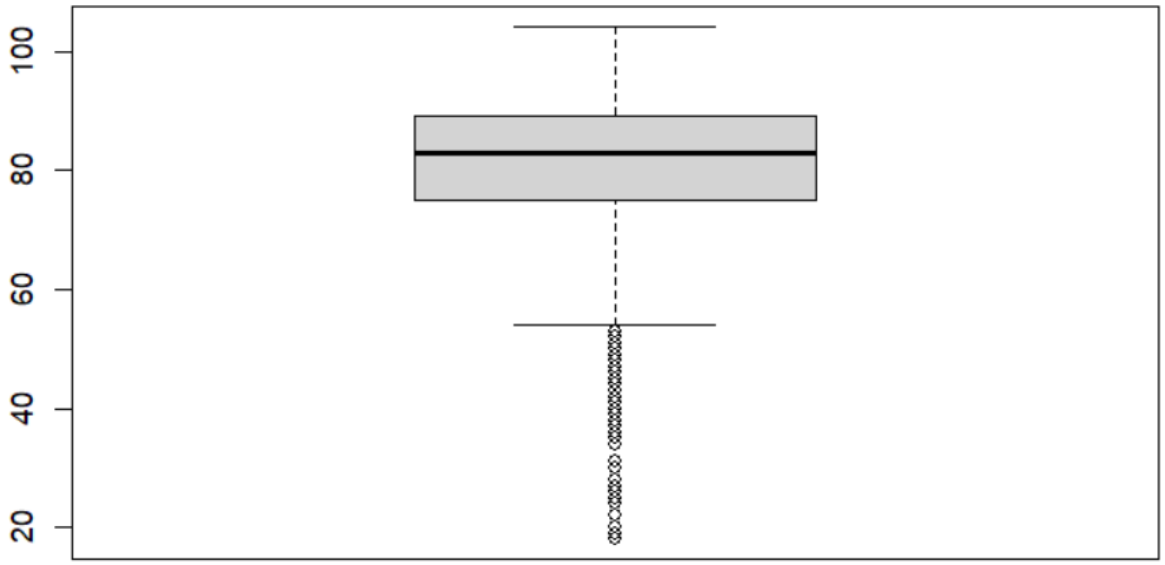
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    285    1738    2508    3038    3718   27985
```

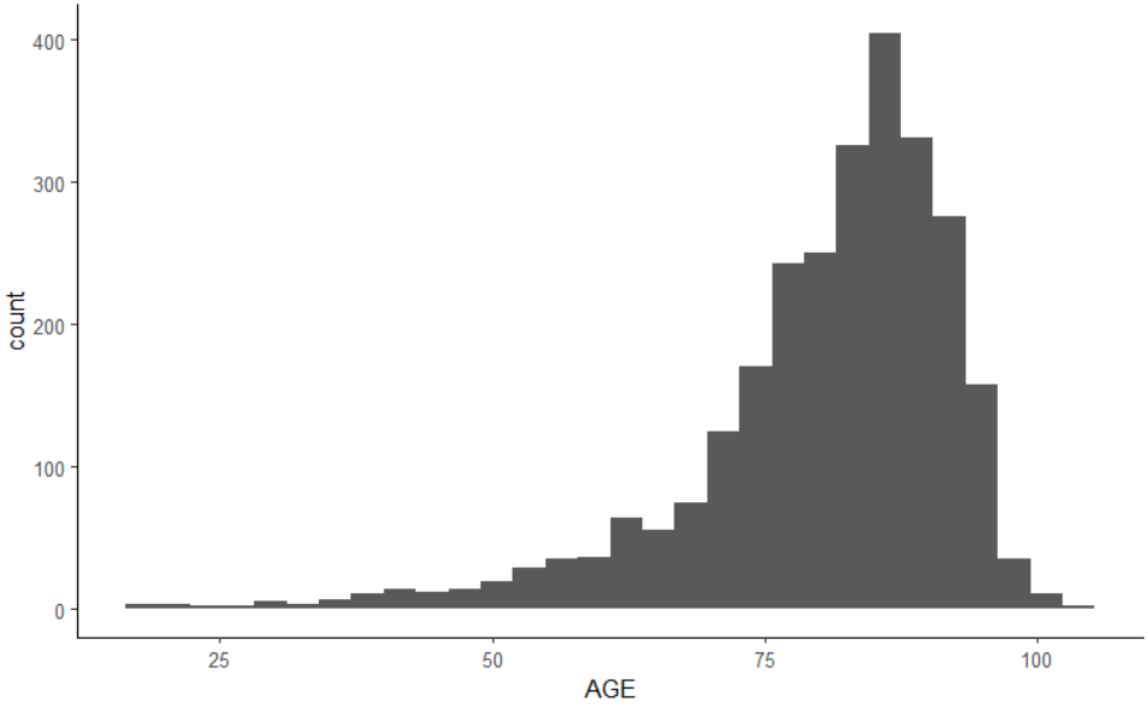


- Le pic correspondant aux coûts totaux se situe clairement entre 1 000£ et 5 000£ (histogramme), tandis que la dispersion des données elle, s'étend de 285£ à 27 985£ environs.
- La valeur de coûts totaux minimum est de 285£, tandis que la valeur de coûts totaux maximum est de 27 985£: Une différence importante (de 27 700£) peut être observée entre les deux valeurs, la valeur de coûts totaux maximum est très élevée.
- Le 1er Quartile des valeurs de coûts totaux est de 1 738£, autrement dit, 25% des valeurs de coûts totaux sont inférieures ou égales à 1 738£.
- Le 2nd Quartile (Médiane) des valeurs de coûts totaux est de 2 508£, autrement dit, 50 % de valeurs des coûts totaux sont inférieures ou égales à 2 508£, et les autres 50% sont supérieures ou égales à 2 508£.
- Le 3è Quartile des valeurs de coûts totaux est de 3 718£, autrement dit, 75% des valeurs de coûts totaux sont inférieures ou égales à 3 718£.
- La Moyenne des valeurs de coûts totaux est de 3 038£ (*des valeurs de* coûts totaux *aberrantes peuvent donc être présentes parmi les données*), soit supérieure à la Médiane qui elle, est de 2 508£ : Nous avons donc une asymétrie vers la droite.

I.13- Analyse des âges (AGE)

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   75.00   83.00   80.29   89.00  104.00
```





- Le pic des données enregistrées sur les âges des patients admis se situe clairement entre 80 ans et 90 ans (histogramme), tandis que la dispersion des données elle, s'étend de 15 ans 105 ans environs.
- L'âge minimum est de 18 ans, tandis que l'âge maximum est de 104 ans. Une différence importante (de 86 ans) peut être déduite entre les deux valeurs.
- Le 1er Quartile des valeurs des âges des patients est de 75 ans, autrement dit, 25% des patients admis sont âgés de 75 ans au maximum.

- Le 2nd Quartile (Médiane) des valeurs des âges est de 83 ans, autrement dit, 50 % des patients admis ont 83 ans au maximum, et les autres 50% ont 83 ans au minimum.
- Le 3è Quartile des valeurs des âges des patients est de 89 ans, autrement dit, 75% des patients admis ont 89 ans au maximum.
- La Moyenne des âges des patients est de 80,29 ans (*des valeurs d'âges aberrantes peuvent donc être présentes parmi les données*), soit inférieure à la Médiane qui elle, est de 83ans : Nous avons donc une asymétrie vers la gauche.
- L'existence de valeurs (en dessous de 60 ans) d'âge aberrantes peut être constatée (Histrogramme & BoxPlot).

I.14- Analyse des coûts des diagnostics (COST_DIAGNOSTIC)

variables	mean	sd	min	q1	median	q3	max
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
COST_DIAGNOSTIC	114.6703	109.6729	0	41.89	88	153.39	1443.66

1 row

- Le coût minimum est de 0£, tandis que le coût maximum est de 1443,66£ (une large différence de 1443,66£).
- Le 1er Quartile des valeurs des coûts de diagnostics est de 41,89£, autrement dit, 25% des valeurs de coûts de diagnostics sont inférieures ou égales à 41,89£.
- Le 2nd Quartile (Médiane) des valeurs des coûts de diagnostics est de 88£, autrement dit, 50% des valeurs des coûts de diagnostics sont inférieures ou égales à 88£, et les autres 50% sont supérieures ou égales à 88£.
- Le 3è Quartile des valeurs des coûts de diagnostics est de 153,39£, autrement dit, 75% des valeurs des coûts de diagnostics sont inférieures ou égales à 153,39£.
- La Moyenne des coûts de diagnostics est de 114,6703£, soit supérieure à la Médiane qui elle, est de 88£ : Nous avons donc une asymétrie vers la droite.
- L'écart type est de 109,6729£, il est clairement élevé. Nous en déduisons donc que les données sur valeurs de coûts de diagnostics sont dispersées et éloignées de la Moyenne qui est de 114,6703£ (Données Hétérogènes).

I.15- Durées de séjour, âges et coûts totaux

variables	mean	sd	min	q1	median	q3	max
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
AGE	80.291896	12.491577	18	75.00	83.00	89.000	104.00
LOS	8.665933	5.677395	1	5.00	7.00	11.000	72.00
TOTAL_COST	3037.975094	2280.428857	285	1737.76	2508.29	3717.725	27985.34

3 rows

- A partir des valeurs de leurs Ecart Type qui sont toutes plus ou moins élevées (respectivement 12.49 ans, 5.68 jours et 2 280.43£), les données sur les âges (I.13), les durées de séjour en établissement (I.11) ainsi que sur les coûts totaux sont toutes dispersées et éloignées de leur Moyennes (respectivement, 80.29 ans, 8.67 jours et 3 037.98£), nous sommes donc en face de Données Hétérogènes.
- Toujours à partir de ces valeurs d'Ecart Type, en les arrangeant dans un ordre croissant, nous pouvons en déduire que les données sur les durées de séjour en établissement sont moins hétérogènes que celles sur les âges, et ces dernières sont moins hétérogènes que les données sur les coûts totaux.

II- Analyse bivariée

II.1- Analyse de la relation entre département d'admission et code DRG

	DRG_CODE				
WARD_ADMISSION	087	089	090	127	576
08	71	1	2	140	5
21	64	83	14	155	61
24	0	21	15	1	97
2604	199	52	36	129	109
2605	566	37	16	163	351
68	266	40	27	3	3

Pourcentage en ligne

	DRG_CODE					
WARD_ADMISSION	087	089	090	127	576	Total
08	32.4	0.5	0.9	63.9	2.3	100.0
21	17.0	22.0	3.7	41.1	16.2	100.0
24	0.0	15.7	11.2	0.7	72.4	100.0
2604	37.9	9.9	6.9	24.6	20.8	100.0
2605	50.0	3.3	1.4	14.4	31.0	100.0
68	78.5	11.8	8.0	0.9	0.9	100.0
Ensemble	42.8	8.6	4.0	21.7	23.0	100.0

Pourcentage en colonne

	DRG_CODE					
WARD_ADMISSION	087	089	090	127	576	Ensemble
08	6.1	0.4	1.8	23.7	0.8	8.0
21	5.5	35.5	12.7	26.2	9.7	13.8
24	0.0	9.0	13.6	0.2	15.5	4.9
2604	17.1	22.2	32.7	21.8	17.4	19.3
2605	48.5	15.8	14.5	27.6	56.1	41.5
68	22.8	17.1	24.5	0.5	0.5	12.4
Total	100.0	100.0	100.0	100.0	100.0	100.0

WARD_ADMISSION <chr>	087 <int>	089 <int>	090 <int>	127 <int>	576 <int>
08	71	1	2	140	5
21	64	83	14	155	61
24	NA	21	15	1	97
2604	199	52	36	129	109
2605	566	37	16	163	351
68	266	40	27	3	3

6 rows

- L'admission dans un Ward dépends du DRG code correspondant, autrement dit, *« WARD_ADMISSION »* correspond à la variable à expliquer (dépendante) tandis que *« DRG_CODE »* correspond à la variable explicative (indépendante) ;
- A partir des données ci-dessus, nous pouvons faire les remarques :

  - Les patients avec un DRG 087 *« Pulmonary edema and respiratory failure »* sont majoritairement admis en Ward *« General Medicine »* (2605 en grande partie), bien plus qu'en Ward 68 *«Respiratory Medicine »* ;

  - Les patients avec un DRG 089 *« Pneumonia and pleuritis with complications »* sont principalement admis dans les Wards 21 *« Geriatrics »*, 2604 *« General Medicine 1 »* & 2605 *« General Medicine 2 »*, 68 *«Respiratory Medicine »* mais rarement en Ward 08 *« Cardiology »*;

  - Les patients avec un DRG 090 *« Pneumonia and pleuritis >17 year old »* sont le plus souvent admis en Ward 2604 *« General Medicine 1 »* qu'en Ward 68 *« Respiratory Medicine »* ;

  - Les patients avec un DRG 127 *« Heart failure and shock »* sont principalement admis en Wards 2605 « General Medicine 2 » & 2604 « General Medicine 1 », 21 « Geriatrics » ou 08 mais rarement en Wards 68 «Respiratory Medicine » ou 24 *« Infection and Immunology »*;

  - les patients avec un DRG 576 *« Sepsis without medical ventilation »* sont majoritairement admis en Wards 2605 *« General Medicine 2 »* (en très grande partie) & 2604 *« General Medicine 1 »* qu'en Ward 24 *« Infection and Immunology »*.

II.2- Analyse de la relation entre département d'admission et décès en établissement

```
                DEATH_INHOSP
WARD_ADMISSION    0   1
          08    202  17
          21    308  69
          24    115  19
          2604  453  72
          2605  983 150
          68    323  16
```

Pourcentage en ligne

```
        DEATH_INHOSP
WARD_ADMISSION 0     1     Total
      08        92.2   7.8 100.0
      21        81.7  18.3 100.0
      24        85.8  14.2 100.0
      2604      86.3  13.7 100.0
      2605      86.8  13.2 100.0
      68        95.3   4.7 100.0
      Ensemble  87.4  12.6 100.0
```

Pourcentage en colonne

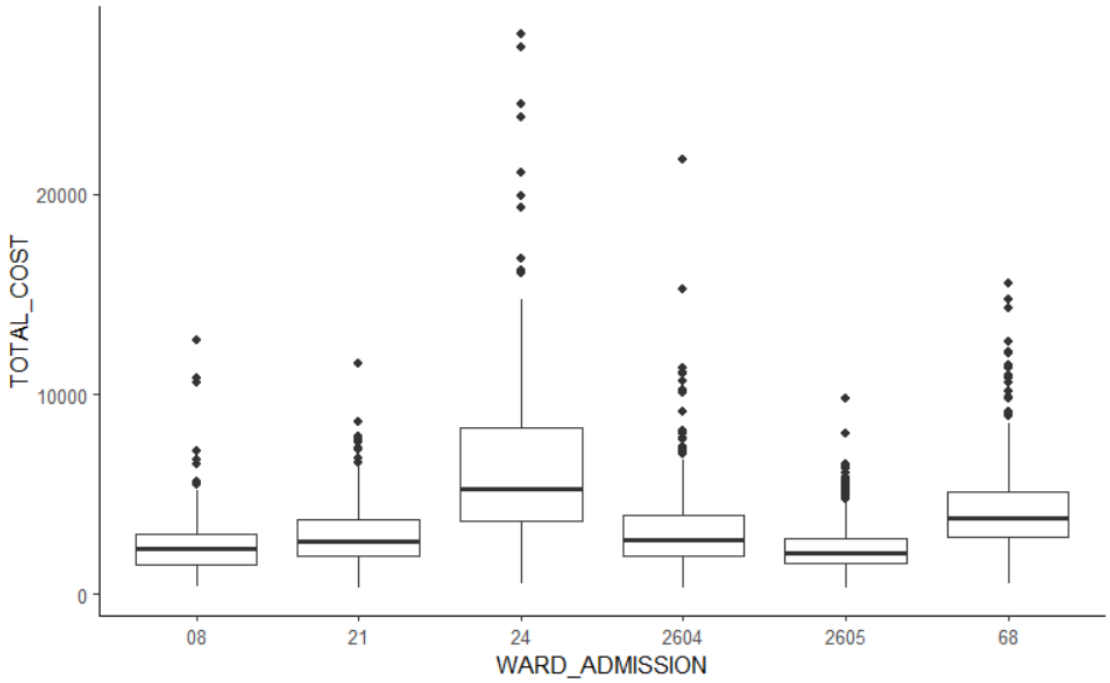
```
                DEATH_INHOSP
WARD_ADMISSION 0     1     Ensemble
         08      8.5   5.0   8.0
         21     12.9  20.1  13.8
         24      4.8   5.5   4.9
         2604   19.0  21.0  19.3
         2605   41.2  43.7  41.5
         68     13.5   4.7  12.4
         Total 100.0 100.0 100.0
```

WARD_ADMISSION <chr>	0 <int>	1 <int>
08	202	17
21	308	69
24	115	19
2604	453	72
2605	983	150
68	323	16

6 rows

- Le tableau de Poucentage en ligne nous fournit les données à partir desquelles nous pouvons constater que la proportion des décès survenus hors des hôpitaux ou lors des séjours dans les hôpitaux dépend du Ward d'Admission du patient.
- A partir de ce tableau de Pourcentage en ligne, nous pouvons remarquer que les proportions de décès survenus hors des hôpitaux pour les patients ayant été admis en Wards 08 et 68 tendent à être plus élevées que celles des autres Wards, et que logiquement, celles des décès survenus lors des séjours dans les hôpitaux quant à eux tendent à être moins élevées que dans les autres Wards.

II.3- Coûts totaux selon le département



- Nous pouvons ici identifier les *« WARD_ADMISSION »* comme étant la variable indépendante (explicative) et les « TOTAL_COST » comme étant la variable dépendante (à expliquer à partir des WARD_ADMISSION ».
- Nous pouvons remarquer à partir des BoxPlots :
  - Les admissions en Ward 24 correspondent globalement aux coûts totaux les plus élevés constatés. Avec la plus grande variabilité des coûts, pourtant moins constantes, nous constatons que les données correspondantes sont asymétriques ;
  - Après les admissions en Ward 24 viennent ensuite les admissions en Ward 68 en termes de coûts élevés, cette fois-ci avec une plus petite variabilité et une meilleure constance des données sur les coûts.
  - Les admissions dans les autres Wards correspondent à des coûts totaux moins conséquentes, mais les données correspondantes sont généralement constantes et relativement symétriques comparées à celles qui correspondent au Ward 24.

II.4- Coûts totaux selon le type de département d'admission (généraliste vs. spécialiste)

WARD_ADMISSION2 <chr>	n <int>	sum <dbl>	mean <dbl>	sd <dbl>
Generalist	1658	4214559	2541.954	1522.633
Specialist	1069	4069999	3807.295	2949.948

2 rows

```
car::leveneTest(TOTAL_COST ~ WARD_ADMISSION2, df_WA_TTC)
```

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    1  129.61 < 2.2e-16 ***
      2725
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- H0 : Les variances des deux groupes « Generalist » et « Specialist » sont égales.

```
wilcox.test(TOTAL_COST ~ WARD_ADMISSION2 , df_WA_TTC, conf.int = T)
```

```
Wilcoxon rank sum test with continuity correction

data:  TOTAL_COST by WARD_ADMISSION2
W = 603617, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -950.10 -718.36
sample estimates:
difference in location
                -829.9
```

- Il y a une différence significative entre les coûts selon que les patients sont dirigés vers les généralistes ou vers les spécialistes.

```
df_WA_TTC %>%
  filter(WARD_ADMISSION2 == 'Generalist') %>%
  select(TOTAL_COST ) %>% unlist()  %>%  as.numeric() %>%  shapiro.test()
```

```
  Shapiro-Wilk normality test

data:  .
W = 0.82541, p-value < 2.2e-16
```

- La distribution des coûts pour les admissions dans les départements de type généraliste ne suit pas une loi normale.

```
df_WA_TTC %>%
  filter(WARD_ADMISSION2 != 'Generalist') %>%
  select(TOTAL_COST ) %>% unlist()  %>%  as.numeric() %>%  shapiro.test()
```

```
 Shapiro-Wilk normality test

data:  .
W = 0.74452, p-value < 2.2e-16
```

- La distribution des coûts pour les admissions dans les départements de type spécialiste ne suit pas une loi normale.

```
   df_WA_TTC %>%
  select(TOTAL_COST ) %>% unlist()  %>%  as.numeric() %>%  shapiro.test()
```

```
 Shapiro-Wilk normality test

data:  .
W = 0.7264, p-value < 2.2e-16
```

- La distribution des coûts totaux ne suit pas une loi normale.

```
kruskal.TTC_WA<- kruskal.test(TOTAL_COST ~ WARD_ADMISSION, data = df_WA_TTC
)
kruskal.TTC_WA
```

```
 Kruskal-Wallis rank sum test

data:  TOTAL_COST by WARD_ADMISSION
Kruskal-Wallis chi-squared = 522.87, df = 5, p-value < 2.2e-16
```

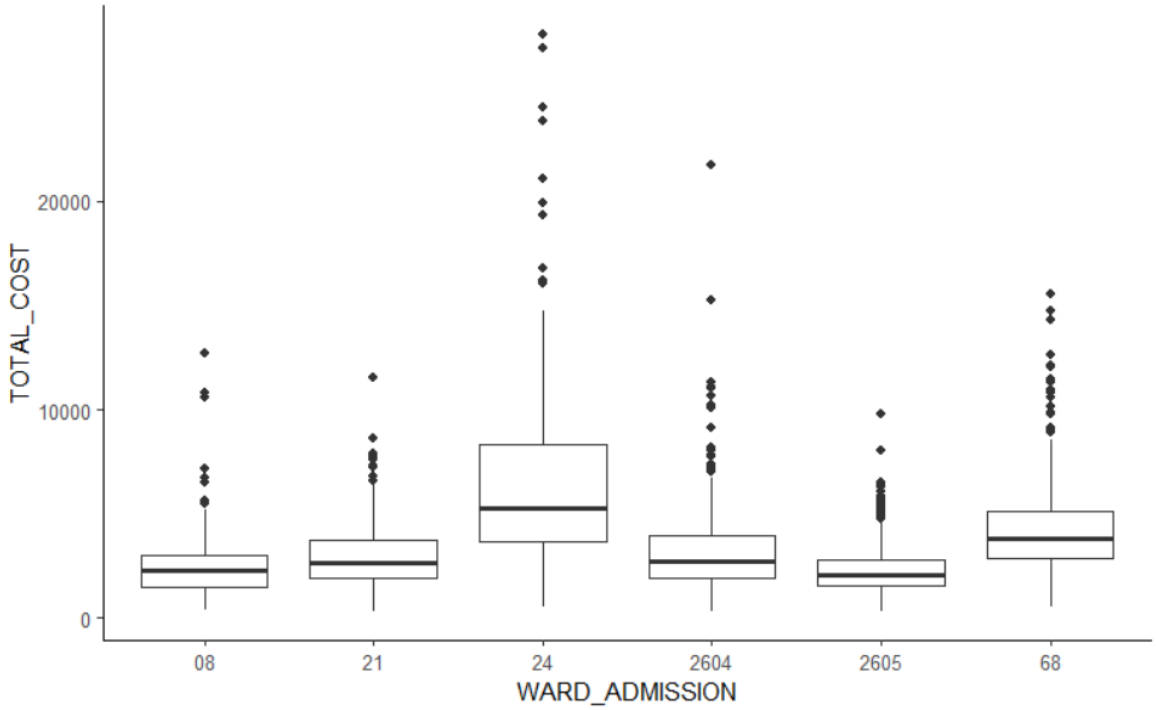
- Les coûts totaux dans chaque département d'admission ne sont pas des populations identiques.

```
pairwise.wilcox.test(df_WA_TTC$TOTAL_COST, df_WA_TTC$WARD_ADMISSION,p.adjus
t.method = "BH")
```

```
Pairwise comparisons using Wilcoxon rank sum test with continuity correctio
n

data:  df_WA_TTC$TOTAL_COST and df_WA_TTC$WARD_ADMISSION

     08      21      24      2604    2605
21   3.1e-05 -       -       -       -
24   < 2e-16 < 2e-16 -       -       -
2604 1.3e-06 0.38    < 2e-16 -       -
2605 0.33    3.0e-15 < 2e-16 < 2e-16 -
68   < 2e-16 < 2e-16 2.9e-09 < 2e-16 < 2e-16

P value adjustment method: BH
```

- Nous constatons une différence significative des coûts entre :
  - département 08 vs département 21 ;
  - département 08 vs département 24 ;
  - département 08 vs département 2604 ;
  - département 08 vs département 68 ;
  - département 21 vs département 24 ;
  - département 21 vs département 2605 ;
  - département 24 vs département 2604 ;
  - département 24 vs département 2605 ;

- département 24 vs département 68 ;
- département 2604 vs département 2605 ;
- département 2604 vs département 68 ;
- département 2605 vs département 68 ;

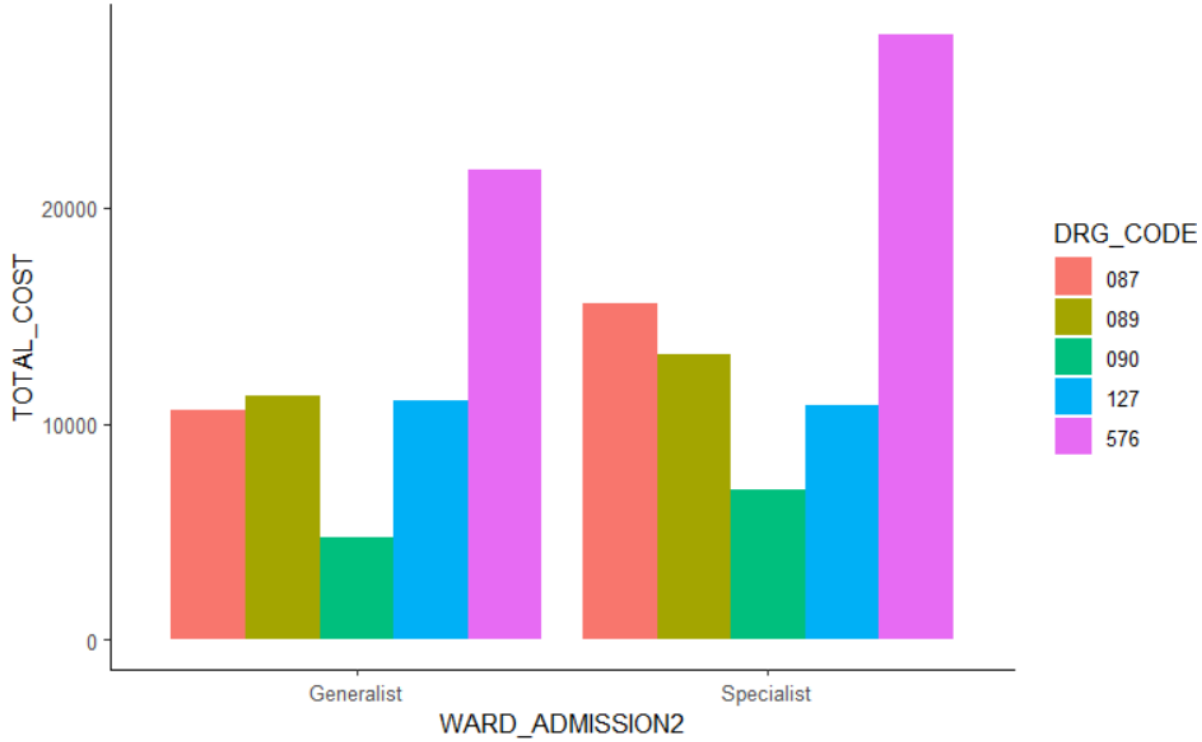


II.5- Coût hospitalier d'un patient dans le département (par jour) (en Euros) selon le département d'admission

	COST_DAY_WARD				
WARD_ADMISSION	285	313	363	463	481
08	0	0	219	0	0
21	0	377	0	0	0
24	0	0	0	0	134
2604	525	0	0	0	0
2605	1133	0	0	0	0
68	0	0	0	339	0

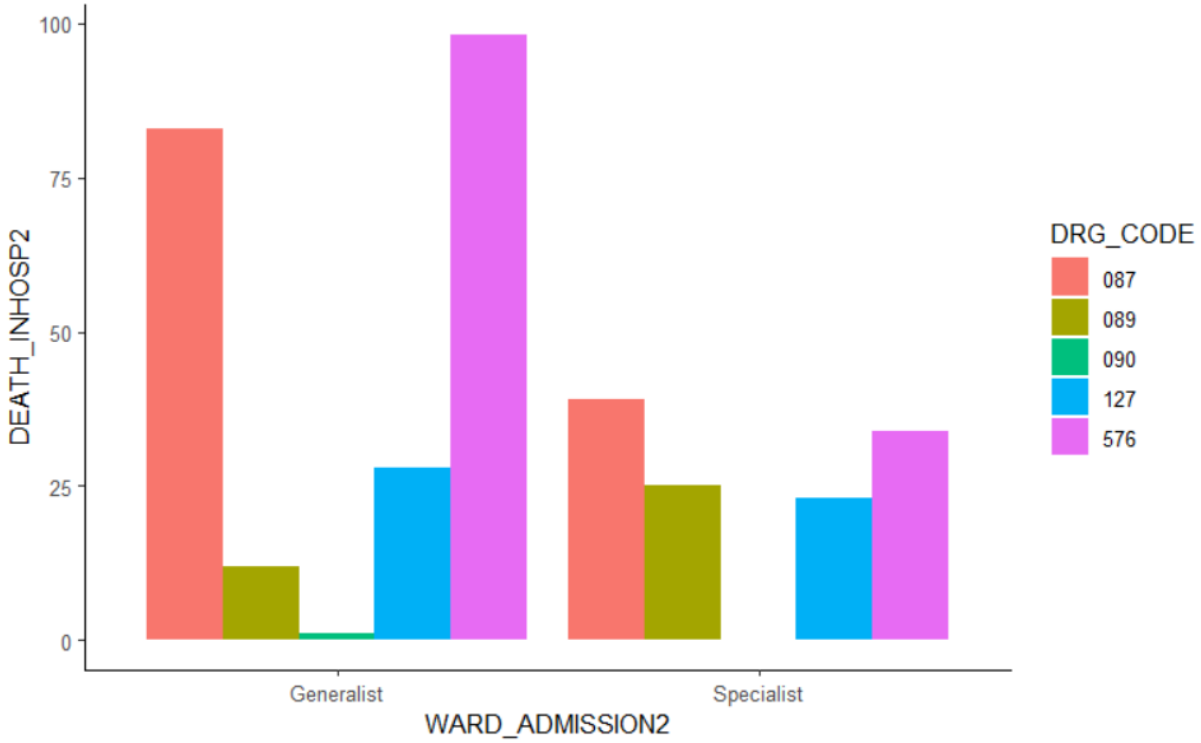
- Nous pouvons clairement comprendre que:
  - Les coûts à 285£ concernent uniquement les Wards « General Medicine » : Très majoritairement pour le Ward 2605, et une autre plus petite partie pour le Ward 2604.
  - Les coûts à 313£ concernent uniquement le Ward 21.
  - Les coûts à 363£ concernent uniquement le Ward 08.
  - Les coûts à 463£ concernent uniquement le Ward 68.
  - Les coûts à 481£ concernent uniquement le Ward 24.

II.6- Distribution des coûts totaux selon le type de département d'admission (généraliste vs spécialiste) et le code DRG



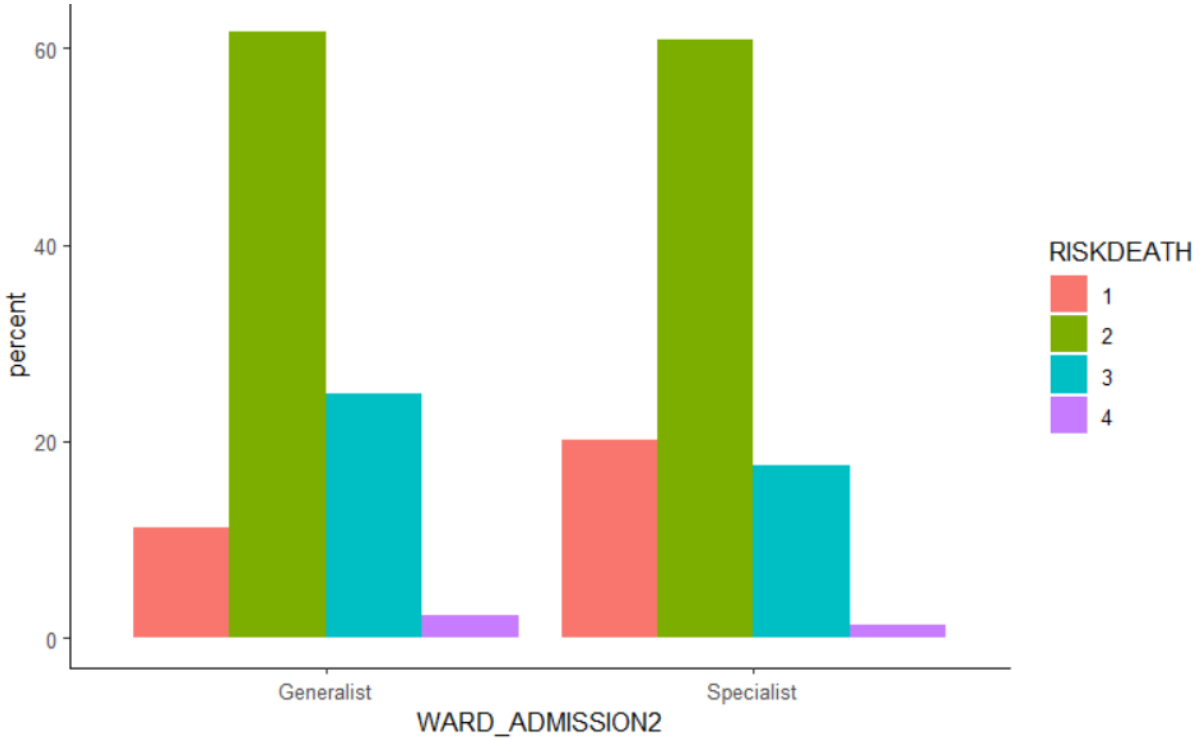
- Les pics de valeurs de l'ensemble des coûts totaux correspondent tous les deux aux admissions avec un DRG 376 dans les deux groupes d'amissions selon le type de département, mais le pic est nettement plus significatif chez les admissions en Specialist qu'en Generalist.
- Si les coûts totaux semblent être aussi élevés pour les admissions en Generalist que pour les admissions en Specialist lorsqu'il s'agit d'admissions avec un DRG 127, ceux-ci sont moins importants pour les admissions en départements Specialist que pour celles en Generalist lorsqu'il s'agit des admissions avec un DRG 090.
- Cependant, les coûts totaux sont moins élevés pour les admissions dans les départements Generalist que pour celles dans les départements Specialist lorsqu'il s'agit des admissions avec un DRG 089 ou 087.

II.7- Distribution des décès en établissement selon le type de département d'admission (généraliste vs spécialiste) et le code DRG

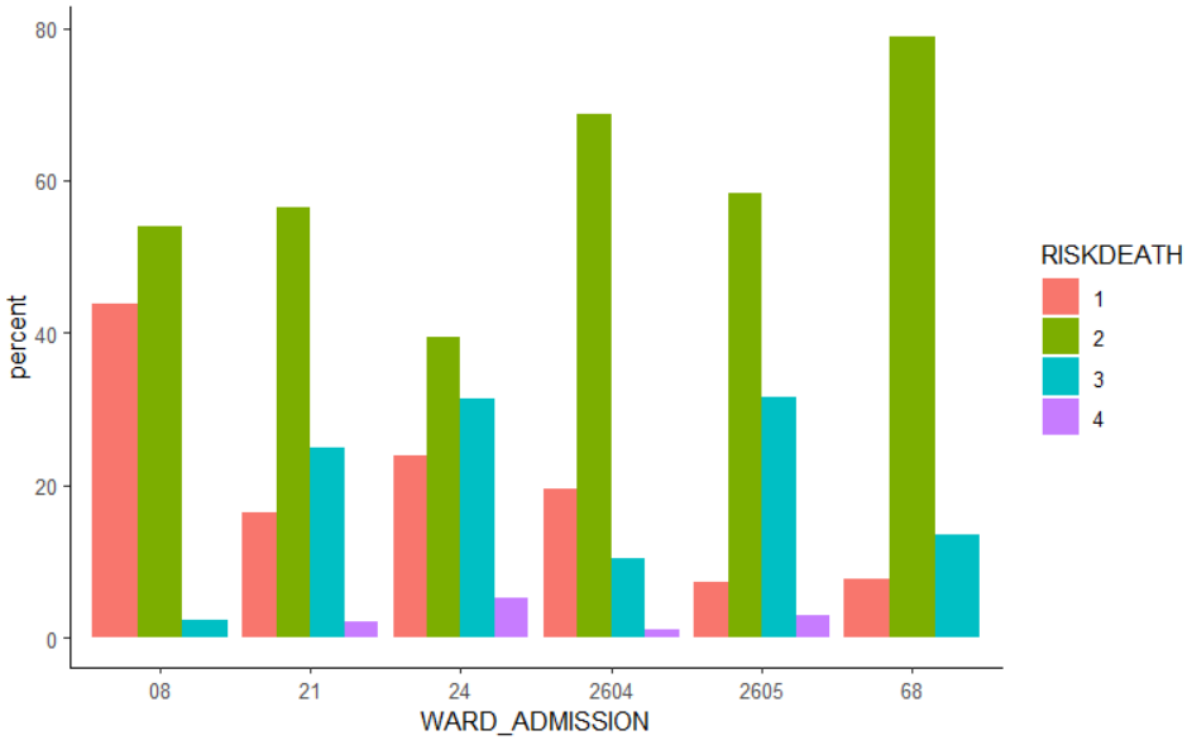


- Le nombre de décès survenus en milieu hospitalier et liés à un DRG 576 constitue le pic pour le cas des départements Generalist, ce qui n'est pas le cas de son homologue des départements Specialist, largement moins important pour sa part.
- Certes largement moins important que celui des départements Generalist, le nombre de décès survenus en milieu hospitalier et liés à un DRG 087 constitue un *timide* pic pour le cas des départements Specialist.
- Le nombre de décès survenus en milieu hospitalier et liés à un DRG 089 en départements Specialist est plutôt supérieur à celui des départements Generalist, tandis que la situation s'inverse pour ceux qui sont liés à un DRG 127.
- Les décès survenus en milieux hospitaliers et liés à un DRG 090 sont relativement rares pour les départements Generalist, et quasi inexistants en ce qui concerne les départements Specialist.

II.8- Distribution des risques de décès en établissement selon le département d'admission

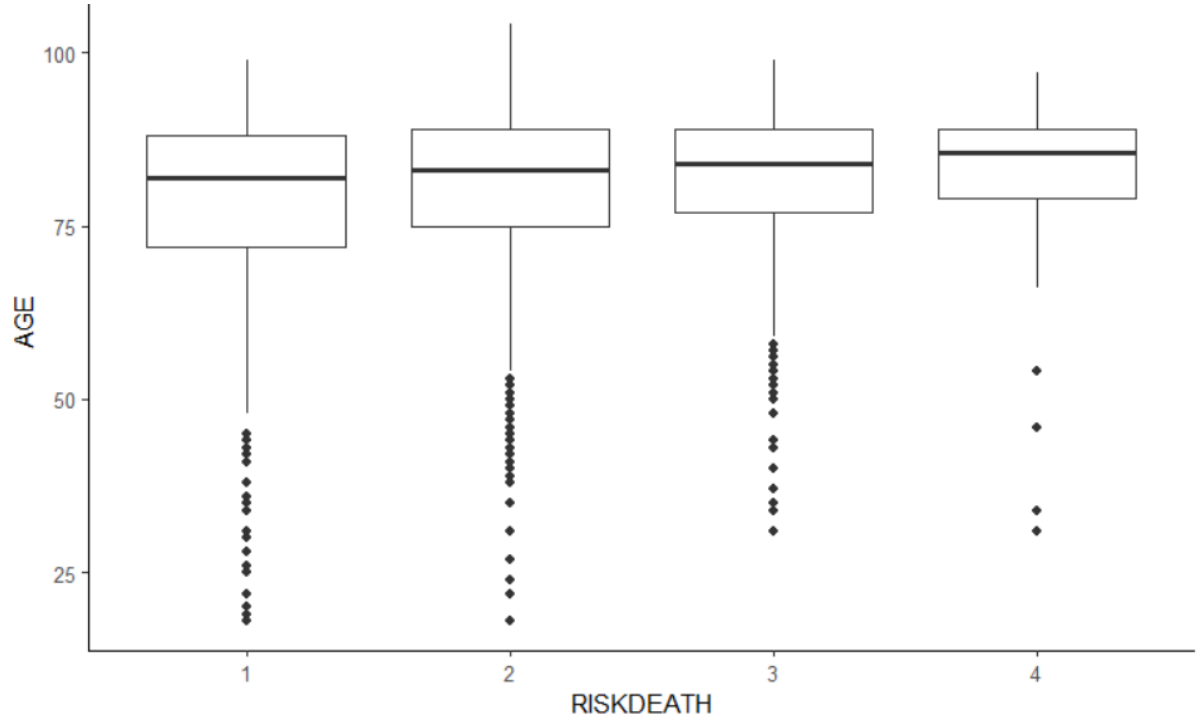


- Les proportions d'admissions avec un RISKDEATH de niveau 2 constituent les pics (des pics relativement similaires) correspondant respectivement aux départements Generalist et Specialist (comportement retrouvé après re-décomposition les groupes de département).
- Les admissions avec un RISKDEATH de niveau 1 sont plus fréquentes dans les départements Specialist que dans les départements Generalist, tandis que la situation est inversée dans le cas des admissions cette fois-ci avec un RISKDEATH de niveau 3.
- Les admissions avec un RISKDEATH de niveau 4 sont relativement peu fréquentes que ce soit du côté des départements Generalist ou celle des départements Specialist.



- Les pics d'admissions avec un RISKDEATH de niveau 2 des groupes Generalist et Specialist sont retrouvés après re-décomposition de ces derniers.
- Les admissions avec un RISKDEATH de niveau 1 sont particulièrement plus fréquentes pour le cas du Ward 08, en contrepartie, celles avec un RISKDEATH de niveau 3 sont particulièrement peu fréquentes.
- Les admissions avec un RISKDEATH de niveau 4 sont plutôt rares, seulement constatés en très petites proportions aux niveaux des Wards 21, 24, 2604 et 2605.
- Les admissions avec un RISKDEATH de niveau 3 sont plus ou moins fréquentes dans chaque Ward, le Ward 08 faisant une petite exception à ce comportement avec notamment ce type d'admission qui est particulièrement peu constaté pour son cas.

II.9- Distribution des âges selon les risques de décès



- A partir de ces BoxPlots, nous pouvons remarquer :
  - Très clairement, les RISKDEATH concernent principalement les patients âgées (70 ans et plus), avec néanmoins quelques donnés aberrantes pour chaque Boxplot.
  - Les 4 Boxplots se trouvent relativement au même niveau (d'age), ont relativement les même valeurs de mediane, mais des distributions légèrement différente en fonction des niveaux de RISKDEATH ;
  - En commençant notre lecture du niveau 1 jusqu'au niveau 4 sur l'axe consacré au RISKDEATH, on constate clairement que la variabilité des données sur les âges diminuent continuellement, ainsi que les quantités de données aberrantes (les risques plus élevés de décès concernent le plus les personnes âgées que les plus jeunes).

```
df_AGE_RISKDEATH <- df  %>%
  select(AGE, RISKDEATH) %>%
  mutate(RISKDEATH = as.character(RISKDEATH) )
```

```
car::leveneTest(AGE ~ RISKDEATH, df_AGE_RISKDEATH)
```

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    3  13.696 7.253e-09 ***
      2723
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- H0 : Les variances sont bien égales

```
aov.RiskDeathAge <- aov(AGE ~ RISKDEATH, data = df_AGE_RISKDEATH)

summary(aov.RiskDeathAge)
```

```
               Df Sum Sq Mean Sq F value  Pr(>F)
RISKDEATH      3   6019    2006   13.03 1.9e-08 ***
Residuals   2723 419345     154
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
shapiro.test(x = residuals(object = aov.RiskDeathAge ))
```

```
Shapiro-Wilk normality test

data:  residuals(object = aov.RiskDeathAge)
W = 0.89206, p-value < 2.2e-16
```

```
kruskal.RiskDeathAge <- kruskal.test(AGE ~ RISKDEATH, data = df_AGE_RISKDEA
TH)

kruskal.RiskDeathAge
```

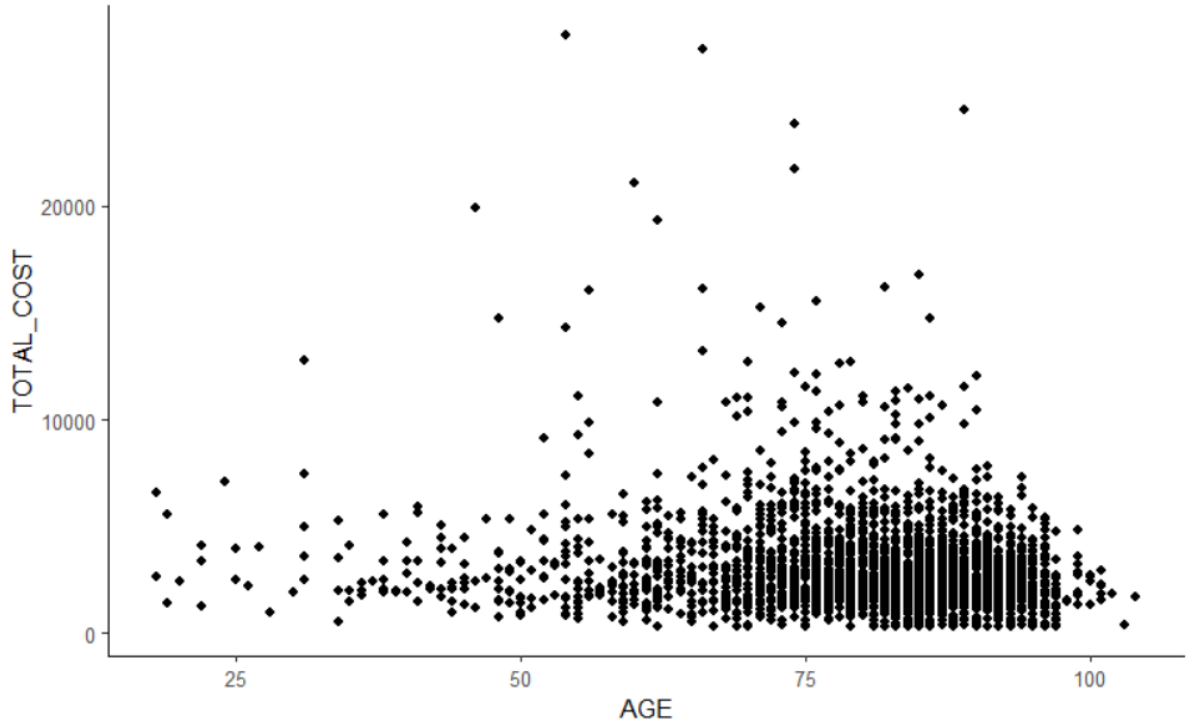
```
Kruskal-Wallis rank sum test

data:  AGE by RISKDEATH
Kruskal-Wallis chi-squared = 17.433, df = 3, p-value = 0.0005756
```

```
pairwise.wilcox.test(df_AGE_RISKDEATH$AGE, df_AGE_RISKDEATH$RISKDEATH,p.adj
ust.method = "BH")
```

```
Pairwise comparisons using Wilcoxon rank sum test with continuity correctio
n

data:  df_AGE_RISKDEATH$AGE and df_AGE_RISKDEATH$RISKDEATH

  1       2       3
2 0.04798 -       -
3 0.00063 0.01487 -
4 0.08924 0.29624 0.78815

P value adjustment method: BH
```

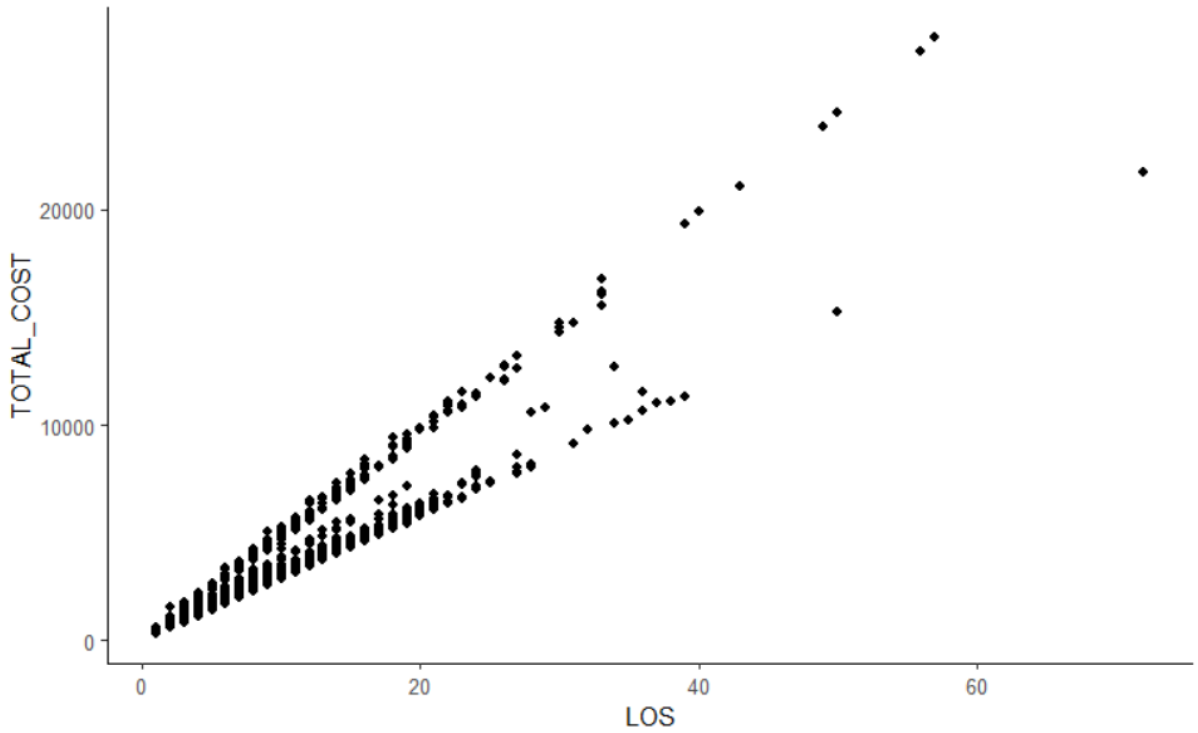
- Il y a une différence de l’âge entre :
  - Risque de décès 1 vs Risque de décès 2 ;
  - Risque de décès 1 vs Risque de décès 3 ;
  - Risque de décès 2 vs Risque de décès 3.

II.10- Distribution des âges selon les coûts totaux



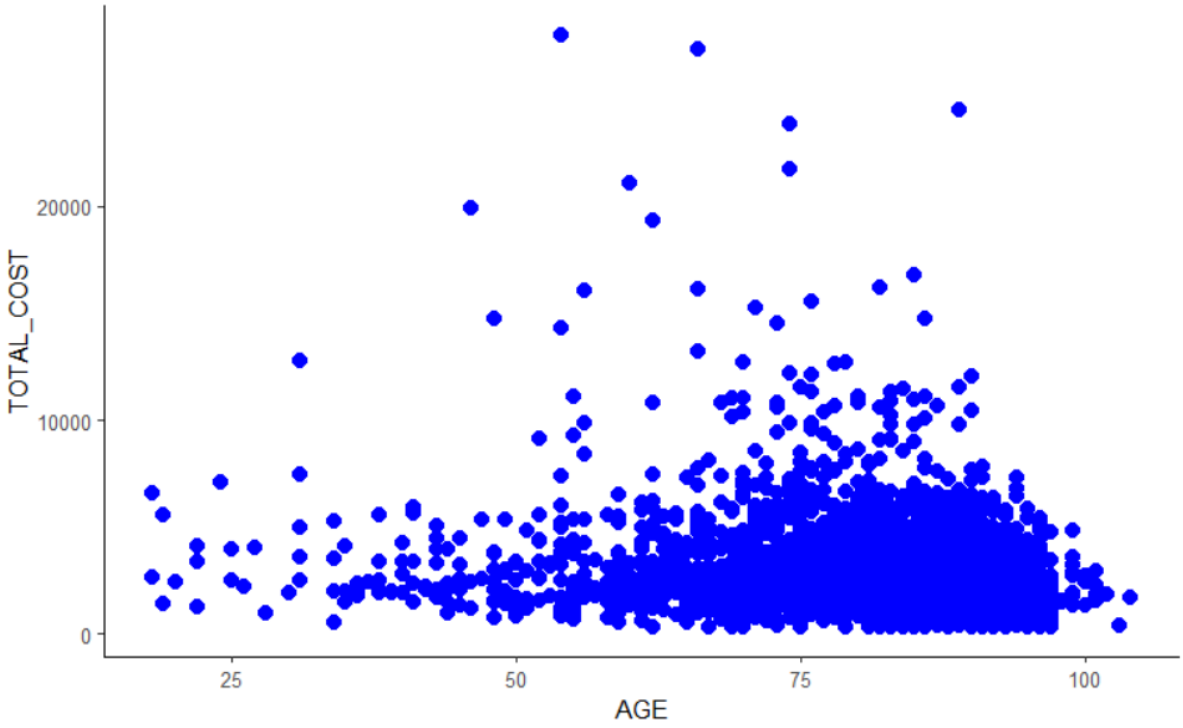
- Nous pouvons remarquer les coûts totaux concernent principalement les patients âgés, et concerne relativement moins les plus jeunes ;
- Cependant, on remarque que la relation tend à être *négative* entre le TOTAL_COST et l'AGE car lorsque les coûts sont les plus élevés, ces derniers concernent le plus les patients certes âgés (soit >60 ans), mais ne faisant pas en même temps partie des plus âgés (soit <90 ans), et constituant une proportion importante des patients.

II.11- Distribution des coûts totaux selon des durées de séjour



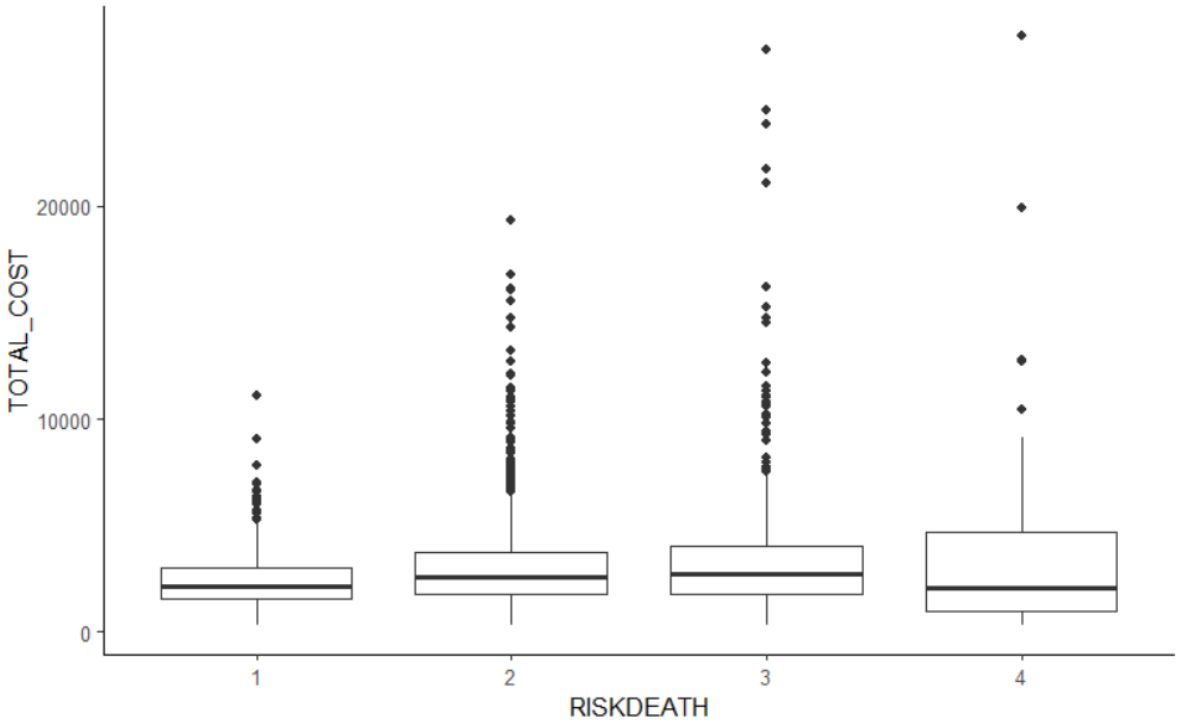
- Nous pouvons remarquer que la relation existante entre le LOS et TOTAL_COST est Positive.
- Plus le temps de séjour est élevé, plus les coûts totaux, eux aussi, sont élevés (ce qui bien logique).

II.12- Distribution des coûts totaux selon l'âge



- Nous pouvons remarquer que les patients âgés sont les plus concernés par les coûts totaux que les plus jeunes.
- Cependant, on remarque que la relation tend à être plutôt négative entre le AGE et les TOTAL_COST : les patients âgés mais ne faisant pas non plus partie de ceux des plus âgés (soit principalement les sexagénaires, septuagénaires et une partie des octogénaires) sont les principaux concernés par les coûts totaux les plus élevés.

II.13- Distribution des coûts totaux selon les risques de décès



- A partir de ces BoxPlots, nous pouvons remarquer :
  - Les 4 Boxplots se trouvent relativement au même niveau (de TOTAL_COST), ont relativement les même valeurs de mediane, mais avec des distributions de données légèrement différentes en fonction des niveaux de RISKDEATH 1, 2 et 3, et une plus grande variabilité des données est constatée pour le niveau de RISKDEATH 4 ;
  - En commençant notre lecture du niveau 1 jusqu'au niveau 4 sur l'axe consacré au RISKDEATH, on constate clairement que la variabilité des données sur les TOTAL_COST augmentent continuellement.
  - Les données sur les coûts totaux aberrantes sont plus importantes pour les niveaux de RISK_DEATH 2 et 3, et moins importantes pour les niveaux 1 et 4.

```
car::leveneTest(TOTAL_COST ~ RISKDEATH, df_TTC_RISKDEATH)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    3   24.35 1.515e-15 ***
      2723
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Egalité de variance vérifiée

```
kruskal.RiskDeathTTC <- kruskal.test(TOTAL_COST ~ RISKDEATH, data = df_TTC_
RISKDEATH)
kruskal.RiskDeathTTC

 Kruskal-Wallis rank sum test
```

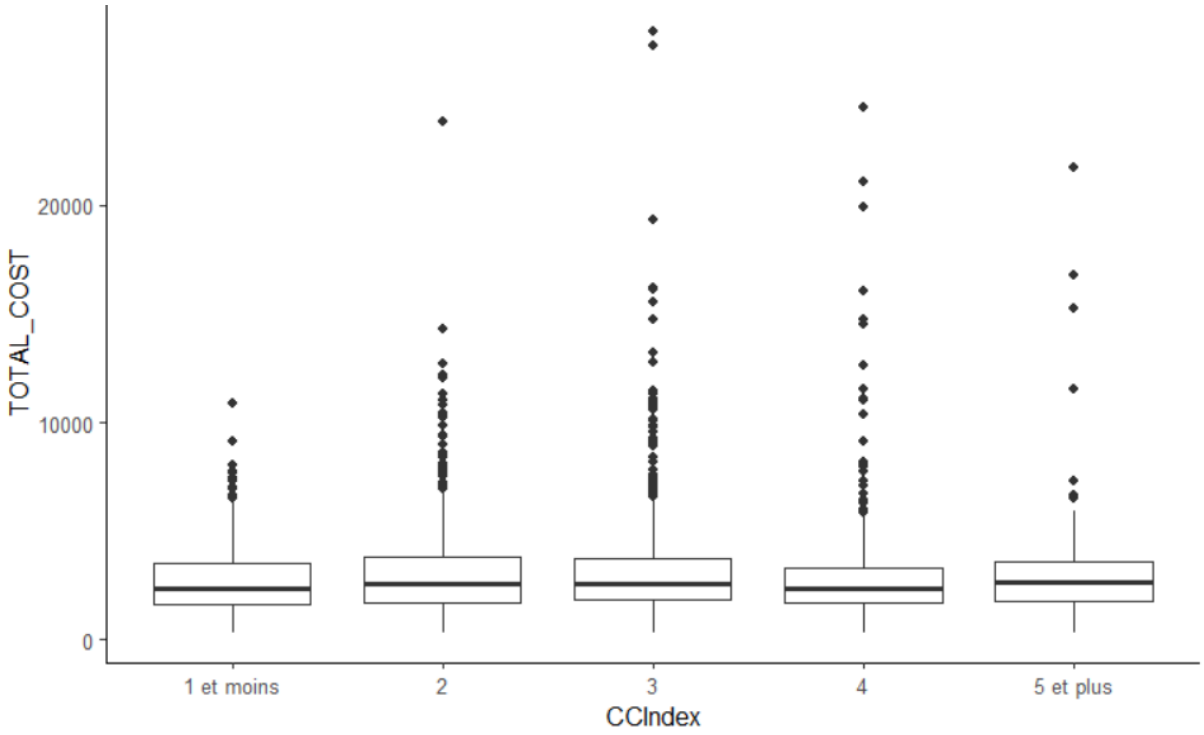
```
data:  TOTAL_COST by RISKDEATH
Kruskal-Wallis chi-squared = 37.997, df = 3, p-value = 2.831e-08
```

- Il y a une différence des coûts totaux selon les risques de décès.

```
pairwise.wilcox.test(df_TTC_RISKDEATH$TOTAL_COST, df_TTC_RISKDEATH$RISKDEAT
H,p.adjust.method = "BH")


    Pairwise comparisons using Wilcoxon rank sum test with continuity corre
ction

data:  df_TTC_RISKDEATH$TOTAL_COST and df_TTC_RISKDEATH$RISKDEATH

  1       2    3
2 1.8e-07 -    -
3 2.1e-07 0.18 -
4 0.71    0.16 0.16

P value adjustment method: BH
```

- Nous constatons une différence de coût entre :
  - Risque de décès 1 vs Risque de décès 2 ;
  - Risque de décès 1 vs Risque de décès 3.

II.14- Distribution des coûts totaux selon l'indice CCI



- A partir de ces BoxPlots, nous pouvons remarquer :
  - Les 5 Boxplots se trouvent relativement au même niveau (de TOTAL_COST), ont relativement les même valeurs de médiane de coûts totaux, et ont des distributions de données sur les coûts totaux légèrement différentes en fonction des index de CCI.
  - Les données sur les coûts totaux aberrantes (élevés) sont plus importantes pour les index de CCI 2, 3 et 4, et relativement moins importantes pour les index de CCI « 1 et moins » et « 5 et plus ».

TOTAL_COST <dbl>	CCI <dbl>	CCIndex <chr>
2675.10	3	3
4963.76	0	1 et moins
5272.50	1	1 et moins
1955.84	1	1 et moins
969.90	3	3
4223.96	1	1 et moins

6 rows

```
car::leveneTest(TOTAL_COST ~ CCIndex, df_TTC_CCI)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    4  1.7774 0.1305
      2722
```

```
aov.CCITTC <- aov(TOTAL_COST ~ CCIndex, data = df_TTC_CCI)
summary(aov.CCITTC)
```

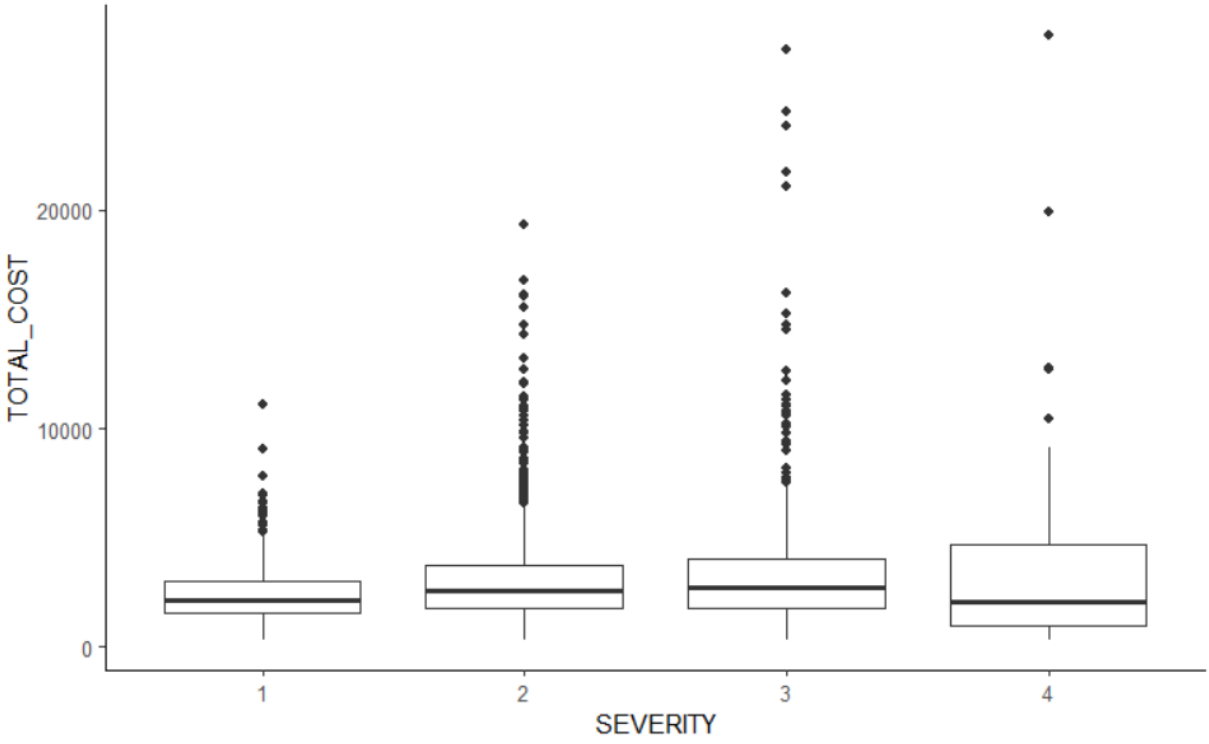
```
Df    Sum Sq  Mean Sq F value  Pr(>F)

CCIndex        4 7.434e+07 18585200   3.587 0.00635 **
Residuals   2722 1.410e+10  5180687
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
shapiro.test(x = residuals(object = aov.CCITTC ) )
    Shapiro-Wilk normality test

data:  residuals(object = aov.CCITTC)
W = 0.73005, p-value < 2.2e-16
```

```
kruskal.CCITTC <- kruskal.test(TOTAL_COST ~ CCIndex, data = df_TTC_CCI)
kruskal.CCITTC

Kruskal-Wallis rank sum test

data:  TOTAL_COST by CCIndex
Kruskal-Wallis chi-squared = 9.0362, df = 4, p-value = 0.0602
```

- Il n’y a pas de différence significative du coût selon CCI.

II.15- Distribution des coûts totaux selon la sévérité



- A partir de ces BoxPlots, nous pouvons remarquer :

- Les 4 Boxplots ont relativement les même valeurs de mediane de TOTAL_COST, mais avec des distributions de données légèrement différentes en fonction des indices 1, 2 et 3, et des coûts totaux particulièrement plus élevés ainsi qu' une plus grande variabilité des données sont constatés pour ce qui est du cas de l'indice 4 ;
- En commençant notre lecture de l'indice 1 jusqu'au niveau 3 sur l'axe consacré au SEVERITY, on constate clairement que la variabilité des données sur les TOTAL_COST augmentent légèrement et continuellement, tandis que l'augmentation de la variabilité est plutôt *brusque* comparée aux précédentes pour le cas du Boxplot correspondant à l'indice de SEVERITY 4.
- Les données sur les coûts totaux aberrantes sont plus importantes pour les indices de SEVERITY 2 et 3, mais moins importantes pour les niveaux 1 et 4 (par contre plus dispersées pour le cas de ce dernier).

TOTAL_COST <dbl>	SEVERITY <chr>
2675.10	2
4963.76	1
5272.50	2
1955.84	1
969.90	2
4223.96	1

6 rows

```
car::leveneTest(TOTAL_COST ~ SEVERITY, df_TTC_SEV)

Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    3   24.35 1.515e-15 ***
      2723
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Egalité de variance vérifiée.

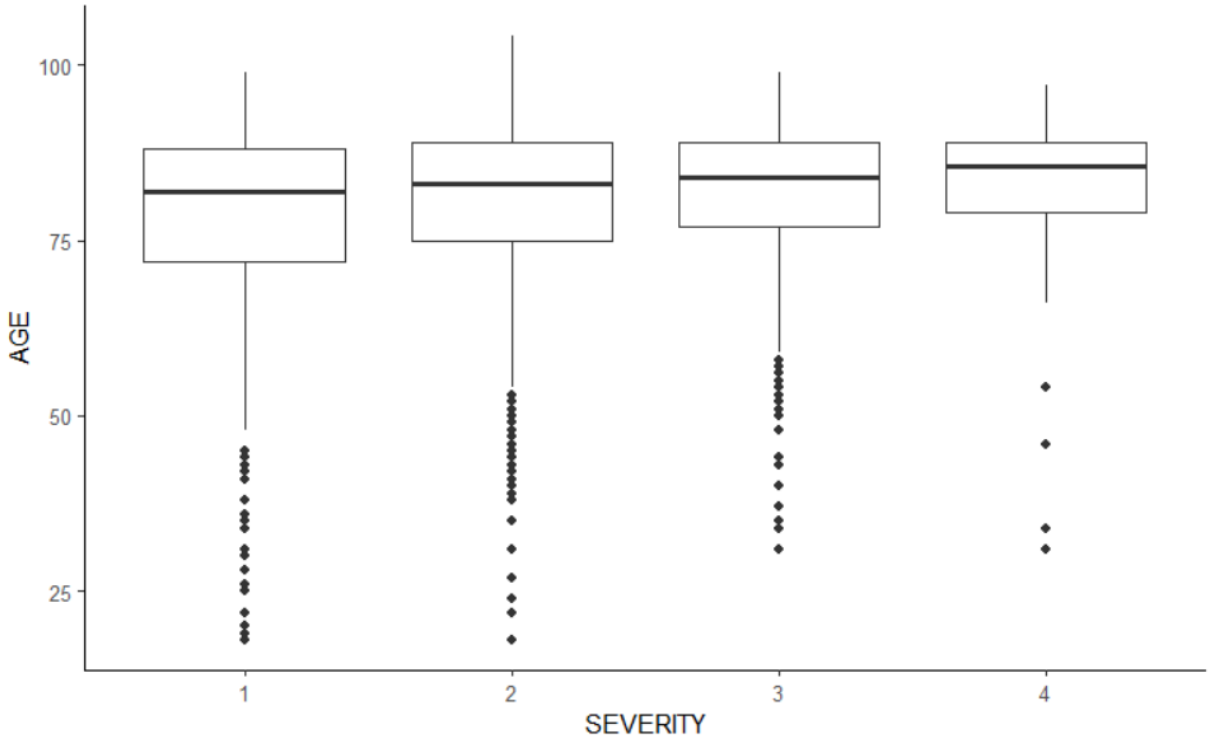
```
kruskal.SEVTTC <- kruskal.test(TOTAL_COST ~ SEVERITY, data = df_TTC_SEV)
kruskal.SEVTTC

    Kruskal-Wallis rank sum test

data:  TOTAL_COST by SEVERITY
Kruskal-Wallis chi-squared = 37.997, df = 3, p-value = 2.831e-08
```

- Nous constatons une différence du coût selon la sévérité.

```
pairwise.wilcox.test(df_TTC_SEV$TOTAL_COST, df_TTC_SEV$SEVERITY,p.adjust.method = "BH")

Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  df_TTC_SEV$TOTAL_COST and df_TTC_SEV$SEVERITY

  1       2    3
2 1.8e-07 -    -
3 2.1e-07 0.18 -
4 0.71    0.16 0.16

P value adjustment method: BH
```

- Il y a une difference de coût entre :
  - Séverité 1 vs Séverité 2 ;
  - Séverité 1 vs Séverité 3.

II.16- Distribution des âges selon la sévérité



- A partir de ces BoxPlots, nous pouvons remarquer :
  - Très clairement, les SEVERITY concernent principalement les patients âgées (70 ans et plus), avec néanmoins quelques donnés aberrantes sur les âges pour chaque Boxplot.
  - Les 4 Boxplots se trouvent relativement au même niveau (d'âge), ont relativement les même valeurs de médiane d'âge, mais avec des distributions différentes en fonction de l'indice de sévérité;

- En commençant notre lecture de l'indice de SEVERITY 1 jusqu'à celui du 4 sur l'axe consacré au SEVERITY, on constate clairement que la variabilité des données sur les âges diminuent continuellement, ainsi que les quantités de données aberrantes (les indices plus élevés de SEVERITY concernent le plus les patients âgées que les plus jeunes).

AGE <dbl>	SEVERITY <chr>
74	2
93	1
88	2
36	1
86	2
71	1

6 rows

```
car::leveneTest(AGE ~ SEVERITY, df_AGE_SEV)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    3  13.696 7.253e-09 ***
      2723
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Egalité de variance vérifiée.

```
kruskal.SEVAGE <- kruskal.test(AGE ~ SEVERITY, data = df_AGE_SEV)
kruskal.SEVAGE


    Kruskal-Wallis rank sum test

data:  AGE by SEVERITY
Kruskal-Wallis chi-squared = 17.433, df = 3, p-value = 0.0005756
```

- Il y a une différence de l'âge selon la sévérité.

```
pairwise.wilcox.test(df_AGE_SEV$AGE, df_AGE_SEV$SEVERITY,p.adjust.method =
"BH")

    Pairwise comparisons using Wilcoxon rank sum test with continuity corre
ction

data:  df_AGE_SEV$AGE and df_AGE_SEV$SEVERITY

  1       2       3
```

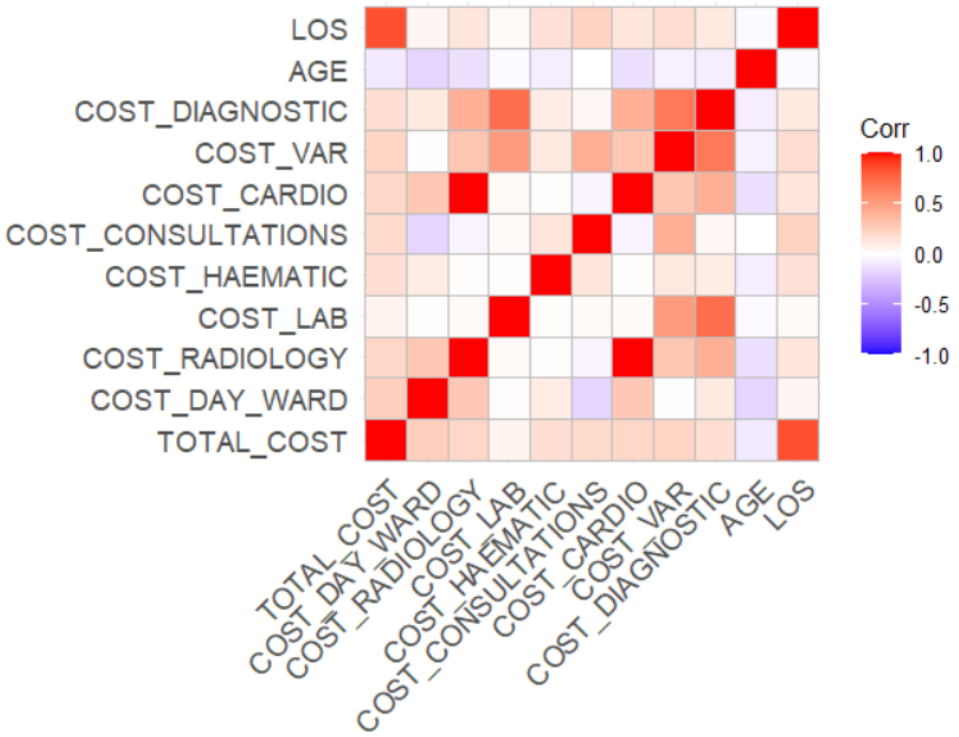
```
2 0.04798 -       -
3 0.00063 0.01487 -
4 0.08924 0.29624 0.78815

P value adjustment method: BH
```

- Il y a une différence de AGE entre:
  - Sévérité 1 vs Sévérité 2 ;
  - Sévérité 2 vs Sévérité 3 ;
  - Sévérité 1 vs Sévérité 2.

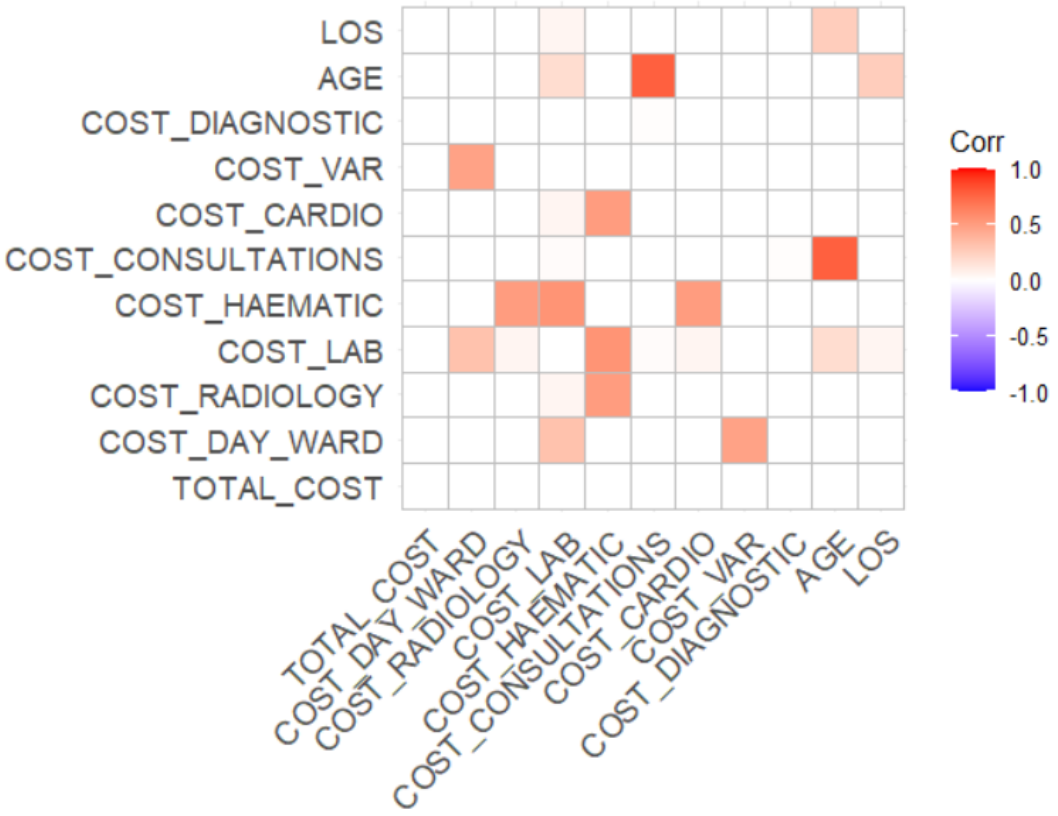
II.17- Corrélations des variables quantitatives (Correlation forte entre durées de séjours et coûts totaux)

```
cor.test(df$LOS, df$TOTAL_COST)

Pearson's product-moment correlation

data:  df$LOS and df$TOTAL_COST
t = 133.1, df = 2725, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9257757 0.9357958
sample estimates:
      cor
0.9309608
```

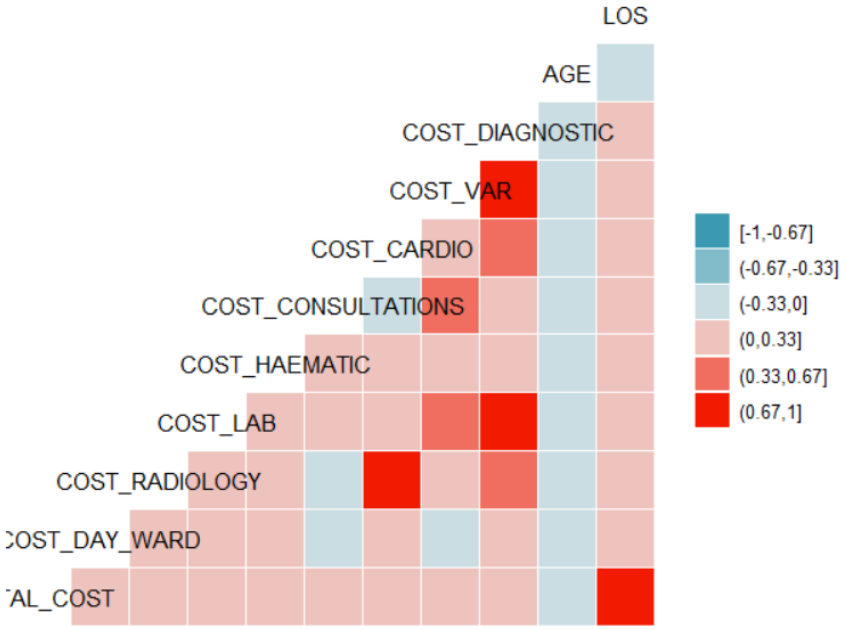
```
df_corr_quantVar <-  df %>%
  select(TOTAL_COST, starts_with("COST_"), AGE, LOS) %>%
  cor(., method = 'kendall')
ggcorrplot(df_corr_quantVar)
```



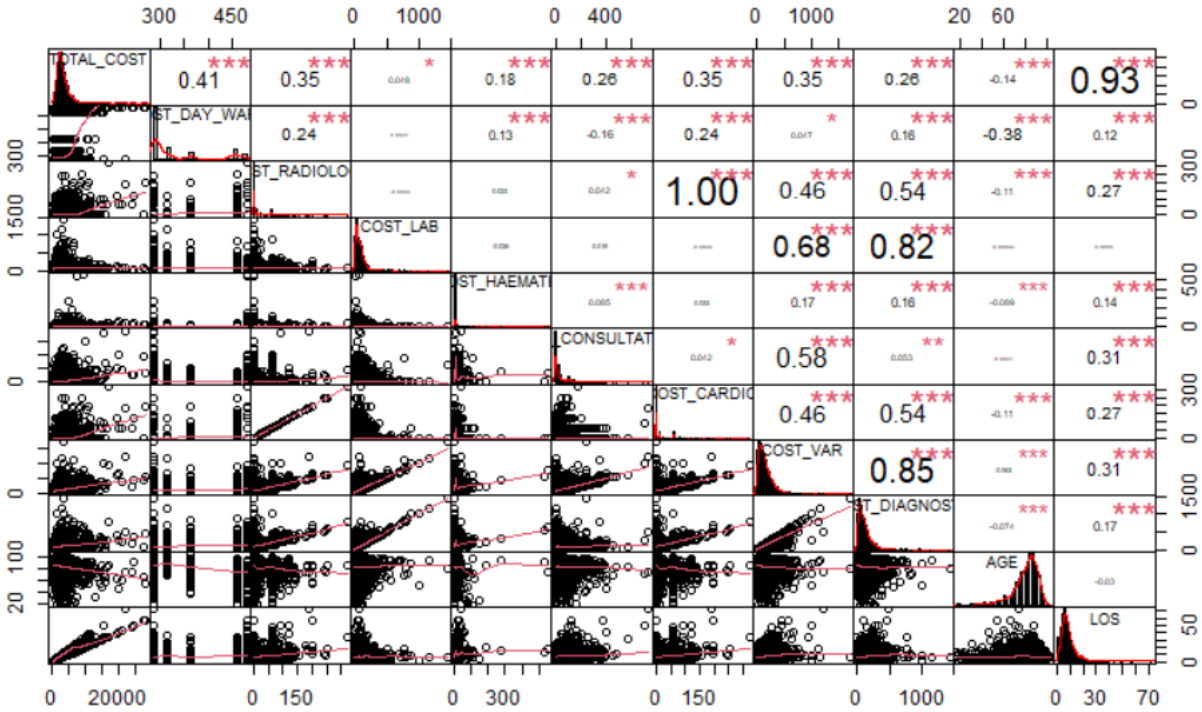

```
df_pmat <-  df %>%
  select(TOTAL_COST, starts_with("COST_"), AGE, LOS) %>%
  ggcorrplot::cor_pmat(., method = 'kendall')
ggcorrplot(df_pmat)
```



```
df %>%
  select(TOTAL_COST, starts_with("COST_"), AGE, LOS) -> df_For_Corr
GGally::ggcorr(df_For_Corr, method = c("everything", "kendall"), nbreaks =
6)
```




```
PerformanceAnalytics::chart.Correlation(df_For_Corr, histogram = TRUE, meth
od = "pearson" ,pch = 19)
```



- Coefficient de corrélation entre LOS et TOTAL_COST égale à 0.93, très proche de 1.
- La corrélation entre LOS et TOTAL_COST est donc Forte.