Hospital Data Analysis – EDA

I- Univariate analysis

I.1- DRG Code distribution :

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

- Most of the admissions mainly correspond to a DRG 087 : *«Pulmonary edema and respiratory failure»* (largely), 576: *«Sepsis without medical ventilation»* or 127: *«Heart failure and shock»*;
- Those which correspond to a DRG 089 *«Pneumonia and pleuritis with complications »* and 090 *«Pneumonia and pleuritis >17 year old »* each constitute smaller proportions.

I.2- Ward Admission distribution (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

- Admissions in Ward « General Medicine » are in majority : a large part in Ward 2605 : *« General Medicine 2 »* and a less significant one in Ward 2604 : *« General Medicine 1 »*;
- Those within the Wards 21: *« Geriatrics »*, 68 : *«Respiratory Medicine »*, 08 : *« Cardiology »* and Ward 24 : *«Infection and Immunology»* each constitute smaller proportions.

I.3- Ward Discharge distribution (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

- Discharges from Wards *« General Medicine »* are in majority: a large part in Ward 2605 : *« General Medicine 2 »* and a less significant one in Ward 2604 : *« General Medicine 1 »*;
- Those in Wards 21 : *« Geriatrics »*, 68 : *«Respiratory Medicine »*, 08 : *« Cardiology »* and Ward 24 : «*Infection and Immunology»* each constitute smaller proportions.

I.4- Number of deaths distribution within/outside Hospitals (DEATH_INHOSP)

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

2 rows

- Deaths occur mainly (in a very large part) outside hospitals, those which occur within hospital settings constitute only a small proportion.

I.5- Costs distribution per Ward (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

- Daily hospital costs valued at €285 constitute the majority;
- Those valued at €313, €463, €363 and at €481 each constitute smaller proportions.

I.6- Severity distribution (SEVERITY)

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

4 rows

- Admissions of patients with a level 2 of severity risk : *« Moderate »* are clearly in majority;
- Then follow the admissions with the levels 3 *« Major»* and 1 : *« Minor »;*
- Those with a level 4 of severity risk *« Severe »* are poorly numerous.

I.7- Risk Death distribution within/outside Hospitals (RISKDEATH)

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

4 rows

- Admissions of patients with a level 2: « Moderate » of Risk Death are clearly in majority;
- Then follow the admissions with a level 3: *« High »* and 1 : *« Low »* of Risk Death;
- Those with a level 4: *« Extreme »* of Risk Death are poorly numerous.

I.8- Sex distribution (SEX)

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

2 rows

- Women admitted to hospitals outnumber Men.

I.9- Number of surgery distribution (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

- The vast majority of admissions requires between 3 to 7 surgeries;
- Admissions that don't require any surgery constitute a proportion relatively *interesting* compared to those which require only 1 to 2 surgeries;
- Admissions requiring up to 8 surgeries are poorly numerous;

I.10- Distribution in function of the Charlson Comirbidity Index (CCI)

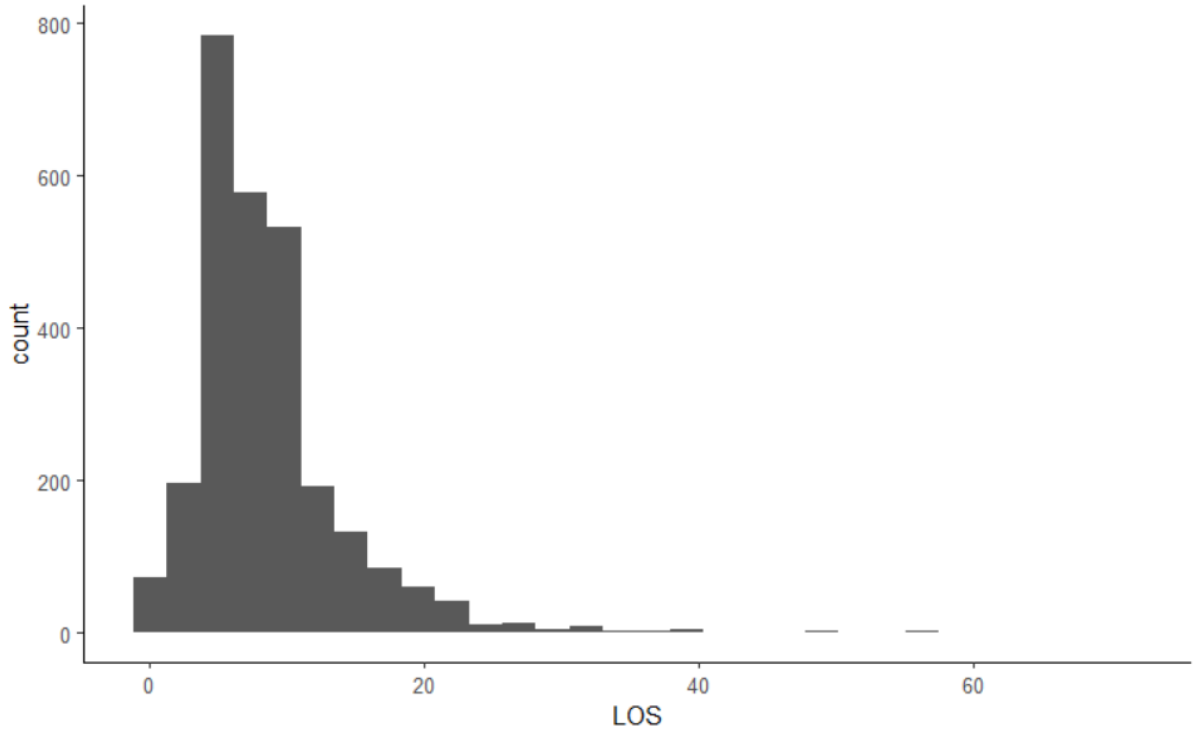
CCI <dbl>	n <int>	% <dbl>
2	993	36.41364136
3	752	27.57609094
1	425	15.58489182
4	385	14.11807847
5	113	4.14374771
0	40	1.46681335
6	17	0.62339567
7	2	0.07334067

8 rows

- The vast majority of admissions corresponds to an index of CCI generally starting from 1 to 4, with a particular *interest* on the scores 2 and 3, which together constitute clearly more than the half of the admissions;
- Although they are less numerous, admissions that correspond to a CCI score equals to 5 seem to be relatively significant in terms of proportions;
- In particular, admissions that correspond to an index of CCI equals to 0 are relatively poorly numerous;
- Those which correspond to an index greater than or equal to 6 can be considered as being poorly numerous, even rare for those with the index 7.

I.11- The Lengths of Stay in Hospital (LOS)

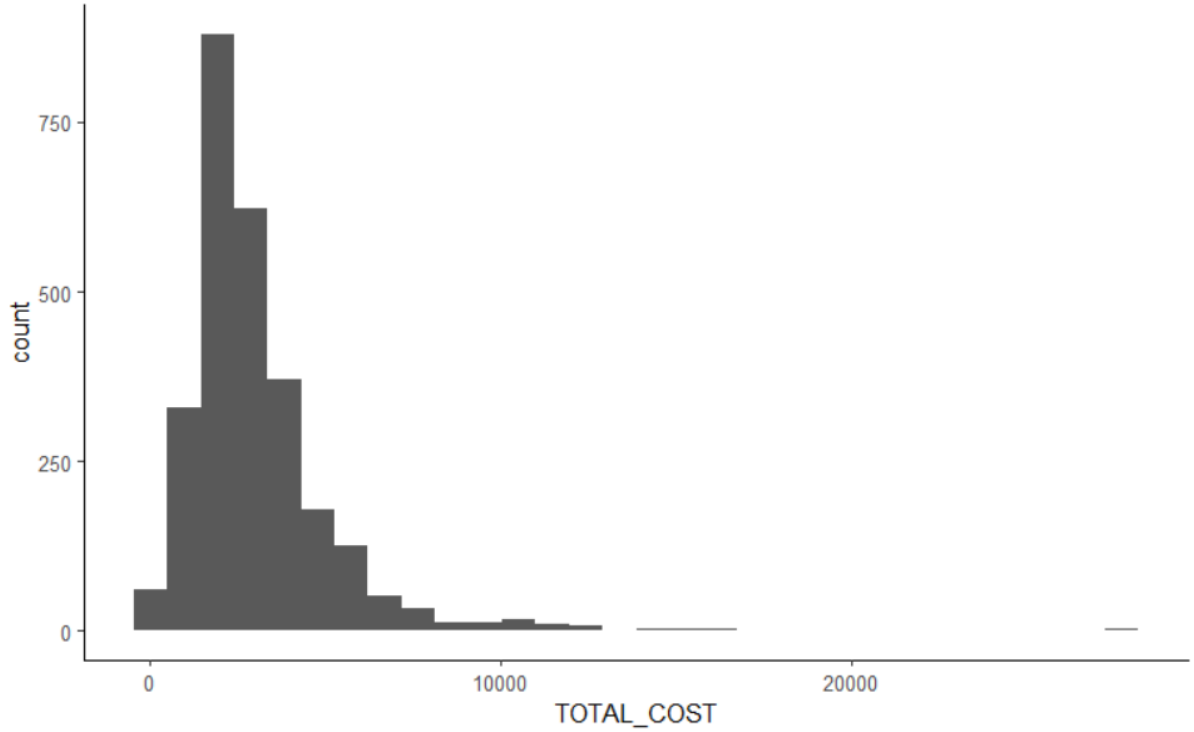
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   5.000   7.000   8.666  11.000  72.000
```



- The peak that corresponds to the LOS is clearly situated between 5 and 10 days (histogram), whereas the data dispersion extends from (around) 1 to 75 days;
- The minimum value of LOS is 1 day, whereas the maximum one is 72 days : a significant difference (71 days) can be observed between the two values, that of the maximum LOS is very high;
- The 1st Quartile of LOS values is 5 days, in other words, 25% of the LOS values are lower than or equal to 5 days;
- The 2nd Quartile (Median) of LOS values is 7 days, in other words, 50% of the LOS values are lower than or equal to 7 days, and the other 50% are greater than or equal to 7 days;
- The 3rd Quartile of LOS values is 11 days, in other words, 75% of the LOS values are lower than or equal to 11 days;
- The Average of LOS values is 8,666 days (outliers might thus be present among the corresponding data), it is greater than the Median which is 7 days: We therefore have an asymmetry to the right here.

I.12- The total costs (TOTAL_COST)

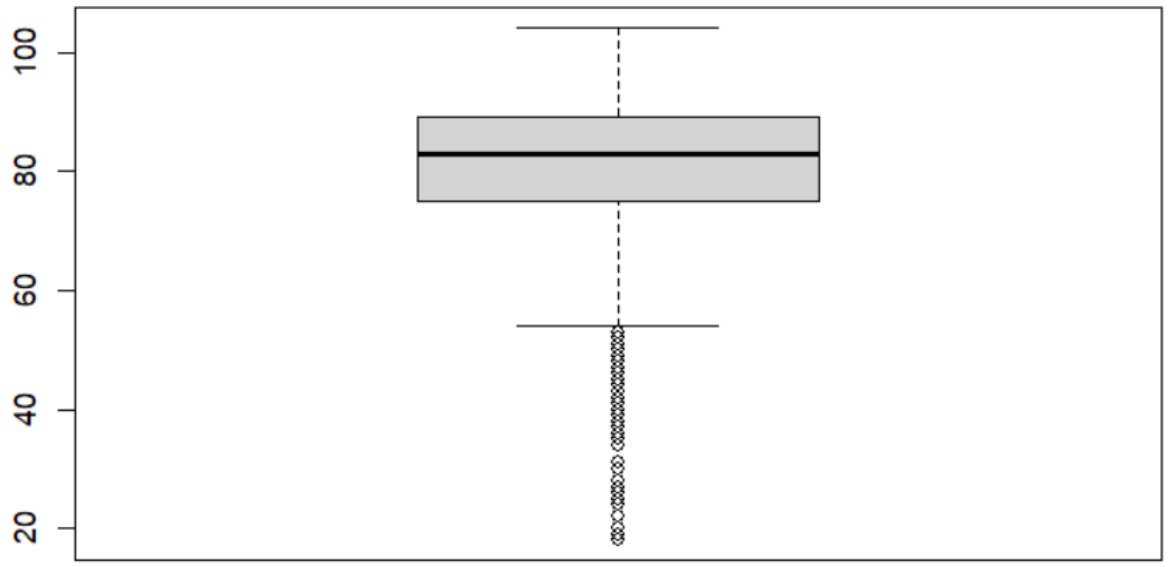
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    285    1738    2508    3038    3718   27985
```

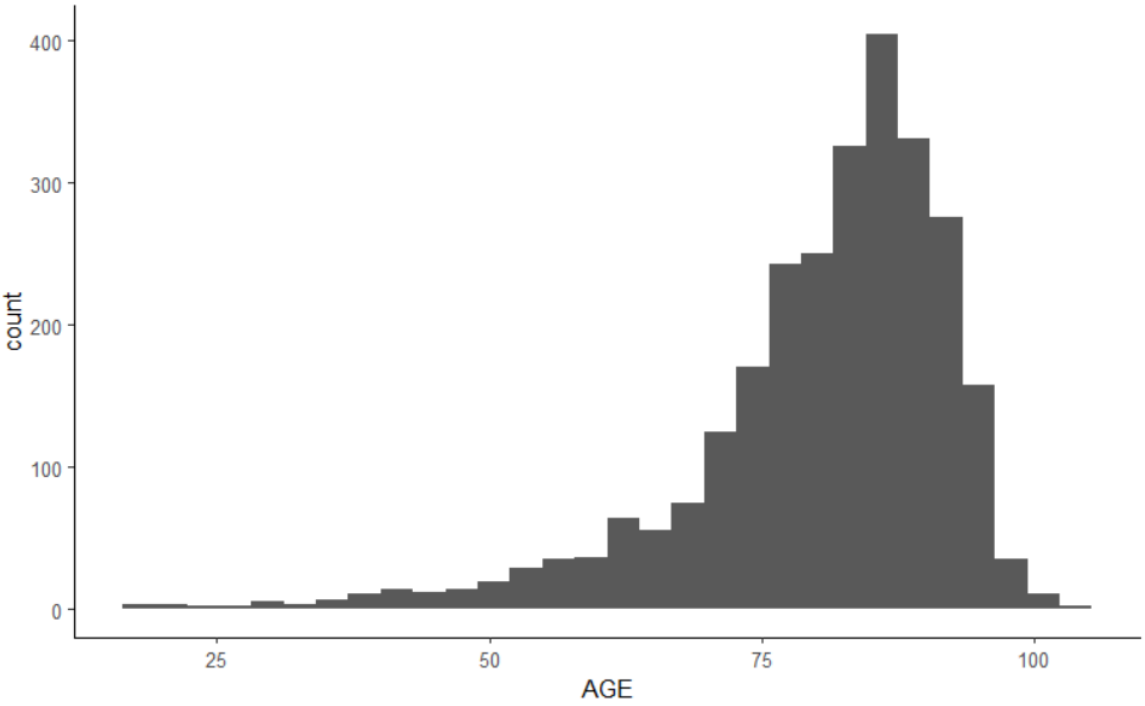


- The peak that corresponds to the total costs is clearly located between €1 000 and €5 000 (Histogram), whereas the data dispersion extends from €285 to €27 985 (around).
- The value of the minimum total costs is €285, whereas that of the maximum total costs is €27 985: A significant difference (€27 700) can be observed between the two values, the value of the maximum total costs is very high;
- The 1st Quartile of the total costs' values is €1 738, in other words, 25% of the values of total costs are lower than or equal to €1 738;
- The 2nd Quartile (Median) of the total costs' values is €2 508, in other words, 50% of the total costs' values are lower than or equal to €2 508, and the remaining 50% are greater than or equal to €2 508;
- The 3rd Quartile of the total costs' values is €3 718, in other words, 75% of the total costs' values are lower than or equal to €3 718;
- The Average of the total costs' values is €3 038 (outliers might be therefore present among the data), this value is greater than the Median which is €2 508: We therefore have an asymmetry to the right here.

I.13- Ages analysis (AGE)

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





- The peak of data recorded on the ages of the admitted patients is clearly situated between 80 years and 90 years (histogram), whereas the data dispersion extends from 15 years old to 105 years old (approximately);
- The minimum age is 18 years old, whereas the maximum one is 104 years old. A significant difference (86 years) can be deduced from these two values;
- The 1st Quartile of age values is 75 years old, in other words, 25% of the patients admitted are aged 75 years old or less;
- The 2nd Quartile (Median) of age values is 83 years old, in other words, 50% of the patients admitted are aged 83 years old or less, and the remaining 50% are at least aged 83 years old;
- The 3rd Quartile of the age values of the patients is 89 years old, in other words, 75% of the patients admitted are aged 89 years old or less;

- The Average of the age of the patients is 80,29 years old (outliers might therefore be present among the data), this value is lower than the Median which is 83 years old: We therefore have an asymmetry to the left here;
- The existence of values (below 60 years old) of outliers can be observed (Histogram & BoxPlot).

I.14- Diagnostic costs analysis (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

- The minimum cost is €0, whereas the maximum one is €1443.66 (a significant difference of € 1443,66);
- The 1st Quartile of the values of the diagnostic cost is €41.89, in other words, 25% of the values of diagnostic costs are lower than or equal to €41.89;
- The 2nd Quartile of the values of diagnostic costs is €88, in other words, 50% of the values of diagnostic cost are lower than or equal to €88, and the remaining 50% are greater than or equal to €88;
- The 3rd Quartile of the values of diagnostic cost is €153.39, in other words, 75% of the values of the diagnostic costs are lower than or equal to €153.39;
- The Average cost of diagnostic is €114.6703, a value that is greater than the Median which is €88: We therefore have an asymmetry to the right;
- The standard deviation is €109.6729, it is clearly high. We therefore deduce that the data on the values of diagnostic cost are dispersed and are situated far from the Average which is €114,6703 (Heterogeneous Data)

I.15- Lengths of stay, ages and total costs

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

- From the values of their Standard Deviation which are all more or less high (respectively 12.49 years old, 5.68 days and €2 280.43), the data on the ages (I.13), on the lengths of stay within the Hospitals (I.11) and on the total costs are all dispersed and are situated far from their Average (respectively: 80.29 years old, 8.67 days and €3 037.98), we therefore have Heterogeneous Data here;

- Always based on these values of Standard Deviation, by sorting them in an ascending order, we can deduce that the data on the lengths of stay within the Hospitals are less heterogeneous than those one the ages, and these latter are less heterogeneous than those on the total costs.

II- Bivariate analysis

II.1- Analysis of the relationship existing between the Admission Ward and the DRG code

WARD_ADMISSION	DRG_CODE 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

Row percentage

WARD_ADMISSION	DRG_CODE 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

Column percentage

WARD_ADMISSION	DRG_CODE 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

- The admission within a Ward depends on the corresponding DRG code, in other words, *« WARD_ADMISSION »* corresponds to the variable to be explained (dependent) whereas the *« DRG_CODE »* corresponds to the explanatory one (independent);
- From the data above, we can make the remarks:
  - Patients with a DRG 087 *« Pulmonary edema and respiratory failure »* are mostly admitted within the Ward *« General Medicine »* (2605 in a large part), more than in Ward 68 *«Respiratory Medicine »* ;
  - Patients with a DRG 089 *« Pneumonia and pleuritis with complications »* are mainly admitted within the Wards 21 *« Geriatrics »*, 2604 *« General Medicine 1 »* & 2605 *« General Medicine 2 »*, 68 *«Respiratory Medicine »* but rarely within the Ward 08 *« Cardiology »*;
  - Patients with a DRG 090 *« Pneumonia and pleuritis >17 year old »* are more often admitted in Ward 2604 *« General Medicine 1 »* than in Ward 68 *« Respiratory Medicine »* ;
  - Patients with a DRG 127 *« Heart failure and shock »* are mainly admitted in Wards 2605 « General Medicine 2 » & 2604 « General Medicine 1 », 21 « Geriatrics » or 08 but rarely in Wards 68 «Respiratory Medicine » or 24 *« Infection and Immunology »*;
  - Patients with a DRG 576 *« Sepsis without medical ventilation »* are mostly admitted in Wards 2605 *« General Medicine 2 »* (in a large part) & 2604 *« General Medicine 1 » than in* Ward 24 *« Infection and Immunology »*.

II.2- Analysis of the relationship existing between the Admission Ward and the Death within/outside Hospitals

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

Row percentage

```
        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
```

Column percentage

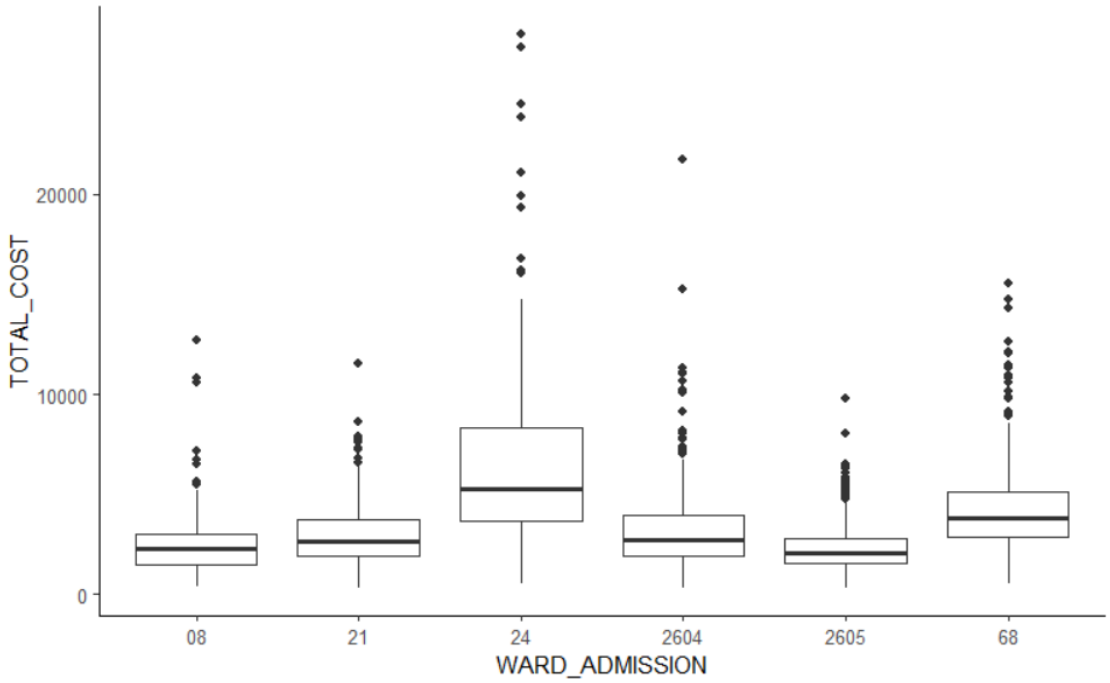
```
                DEATH_INHOSP
WARD_ADMISSION 0     1     Ensemble
         08      8.5   5.0   8.0
         21     12.9  20.1  13.8
         24      4.8   5.5   4.9
         2604   19.0  21.0  19.3
         2605   41.2  43.7  41.5
         68     13.5   4.7  12.4
         Total 100.0 100.0 100.0
```

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

- The table on the Row Percentage provides data from which we can notice that the proportion of deaths occurring outside Hospitals or during the stays within the Hospitals depends on the Ward of Admission of the patient;
- From this table on the Row Percentage, we can notice that the proportions of deaths occurring outside the Hospitals for the patients that have been admitted within the Wards 08 and 68 tend to be higher than those within the other Wards, and that, logically, those of the deaths occurring within the Hospitals tend to be less high than within the other Wards.

II.3- Total Costs in function of the department



- Here, we can identify the *« WARD_ADMISSION »* as being the independent variable (explanatory variable) and the *« TOTAL_COST »* as the dependent one (to be explained from the WARD_ADMISSION);
- We can notice from those BoxPlots:
  - The admissions in Ward 24 globally correspond to the highest total costs observed. With the most significant variability of costs, however less constant, we notice that the corresponding data are asymmetrical;
  - After the admissions in Ward 24 come those in Ward 68 in terms of high costs, this time with a less significant variability and a better data constancy when it comes to the costs;
  - The admissions within the other Wards correspond to total costs that are less significant, but the corresponding data are generally constant and relatively symmetrical compared to those which correspond to the ward 24.

II.4- Total Costs in function of the type of admission department (generalist VS specialist)

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: The variances of the two groups « Generalist » and « Specialist » are equal.

```
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
```

- There is a significant difference between the costs when the patients are directed to the *generalist* wards and when they are directed to the *specialist* ones.

```
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
```

- The costs distribution related to the admissions within the departments of the *Generalist* type doesn’t follow a normal distribution.

```
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
```

- The costs distribution related to the admissions within the departments of the Specialist type doesn’t follow a normal distribution.

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

 Shapiro-Wilk normality test

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

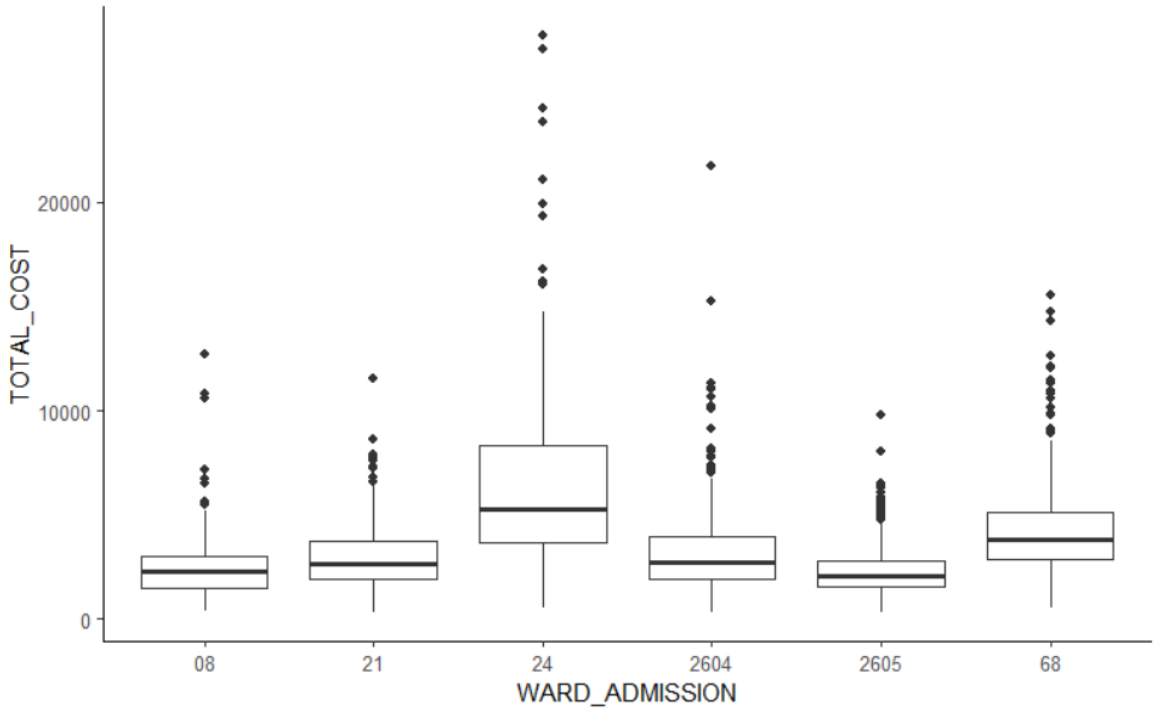
- The distribution of the total costs doesn't follow a normal distribution.

```
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
```

- The total costs within each department of admissions aren't identical populations

```
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
```

- We have noticed a significant difference of costs between:
  - department 08 vs department 21 ;
  - department 08 vs department 24 ;
  - department 08 vs department 2604 ;
  - department 08 vs department 68 ;
  - department 21 vs department 24 ;
  - department 21 vs department 2605 ;
  - department 24 vs department 2604 ;
  - department 24 vs department 2605 ;
  - department 24 vs department 68 ;
  - department 2604 vs department 2605 ;
  - department 2604 vs department 68 ;
  - department 2605 vs department 68 ;

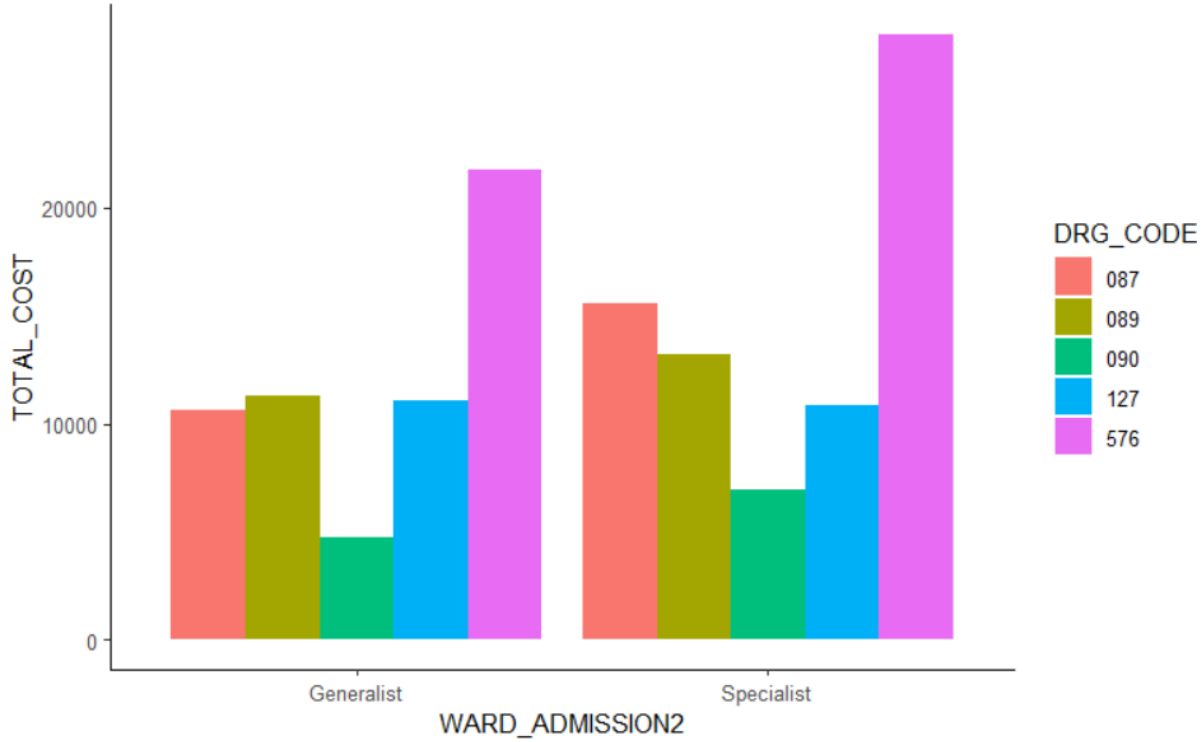


II.5- Hospital cost of a patient within the department (per day) (in Euros) in function of department of admission

	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

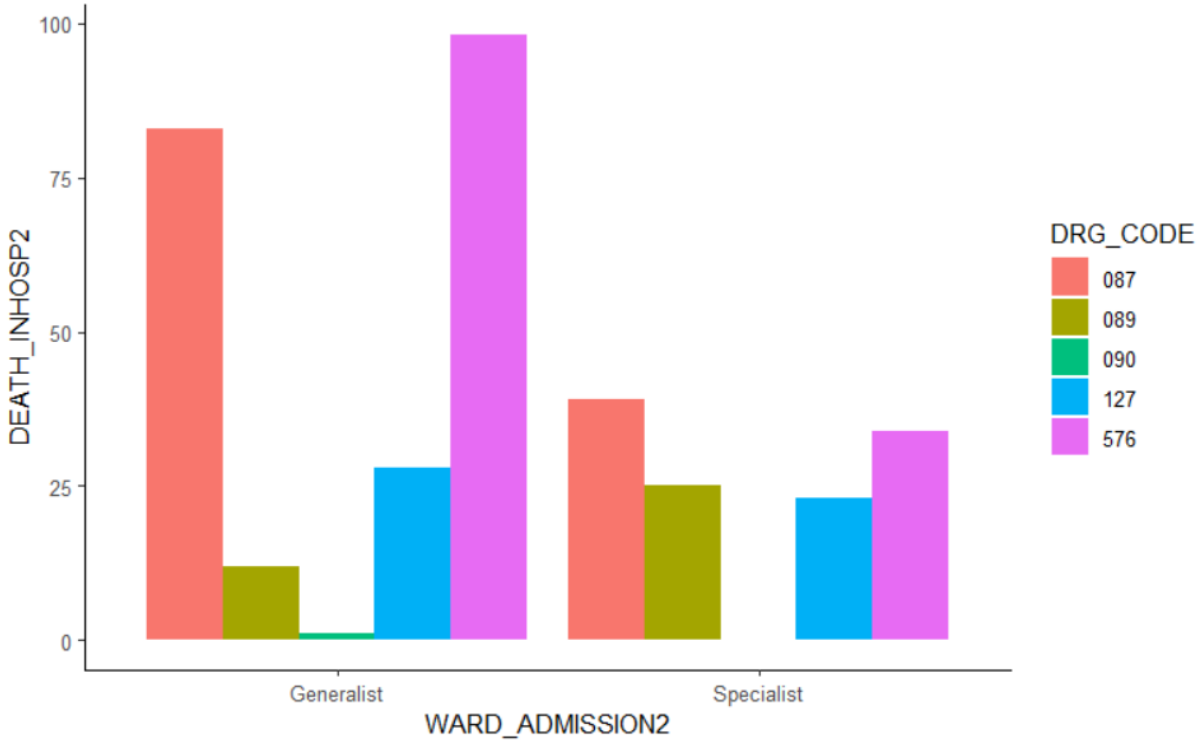
- We can obviously understand that :
  - The costs at €285 concern only the Wards « General Medicine » : Mostly for the Ward 2605, and a smaller part for the Ward 2604;
  - The costs at €313 concern only the Ward 21;
  - The costs at €363 concern only the Ward 08;
  - The costs at €463 concern only the Word 68;
  - The costs at €481 concern only the Ward 24.

II.6- Total Costs distribution in function of the Admission Ward (Generalist and Specialist) and the DRG code



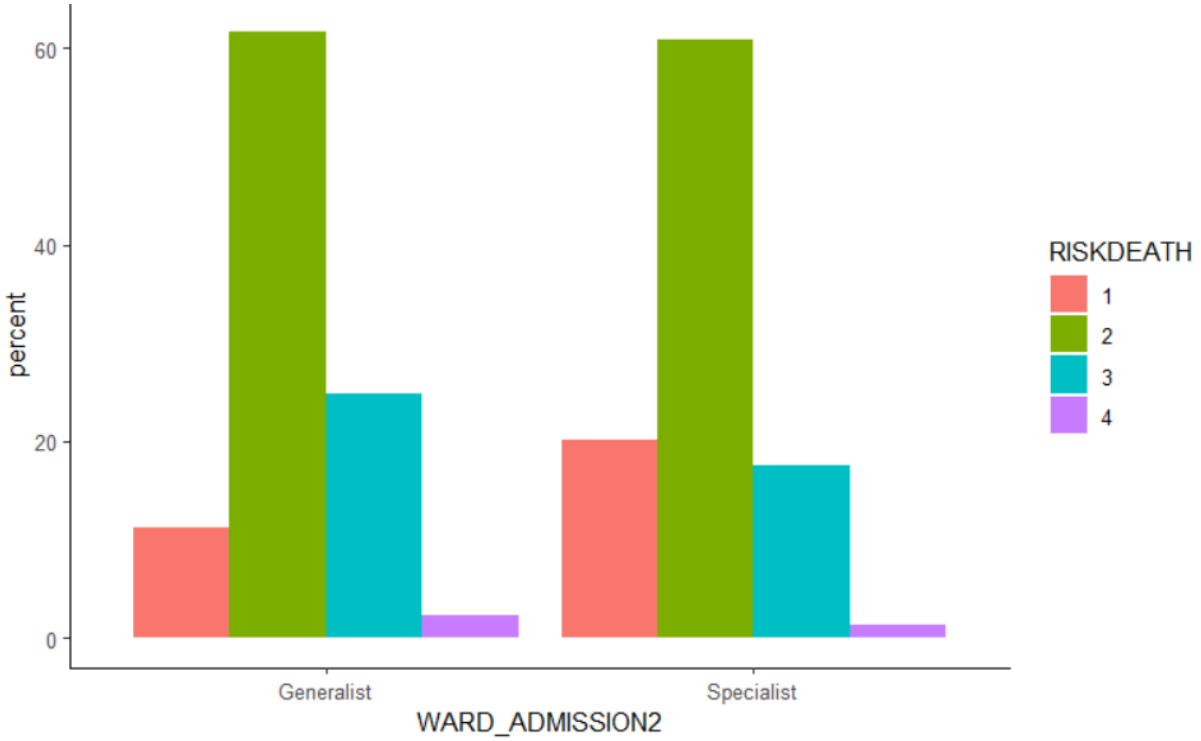
- The peaks of values of the set of Total Costs both correspond to the admissions with a DRG 376 within the two groups of admissions regarding the type of ward, but the peak is clearly more significant for the admissions within the *Specialist* Wards than within the *Generalist* ones;
- If the Total Costs seem to be as high for the admissions in the Generalist Wards as for those in the Specialist ones when it comes to the admissions with a DRG 127, they are less significant for the admissions within the Specialist Wards than those within the Generalist ones when it comes to the DRG 090;
- However, Total Costs are less high for the admissions within the Generalist Wards than those within the Specialist Wards when it comes to the admissions with a DRG 089 or 087.

II.7- Death within/outside Hospitals distribution in function of the Ward of Admission (Generalist vs Specialist) and the DRG Code

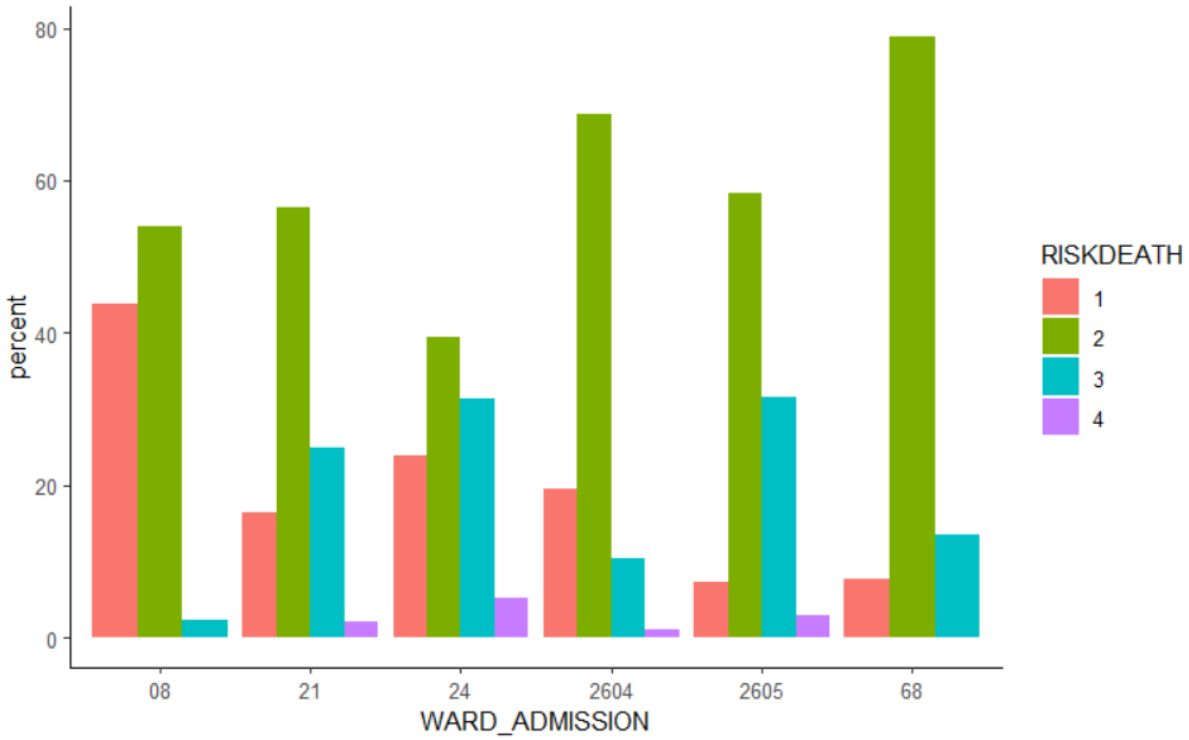


- The number of deaths occurred within Hospitals and linked to a DRG 576 constitutes the peaks for the case of the *Generalist* Wards, which is not the case for its counterpart in the *Specialist* Wards, largely less significant;
- Although it is largely less significant than that of the *Generalist* Wards, the number of deaths occurred within Hospitals and linked to DRG 087 constitutes a *timid* peak for the case of the Specialist Wards;
- The number of deaths occurred within Hospitals and linked to a DRG 089 within the *Specialist* Wards is greater than that of the *Generalist* ones, whereas for those linked to a DRG 127, the situation is reversed;
- Deaths occurred within Hospitals and linked to a DRG 090 are relatively rare for the *Generalist* Wards and almost non-existent for the case of the *Specialist* ones.

II.8- Risk of Death within/outside Hospitals distribution in function of the Ward of Admission

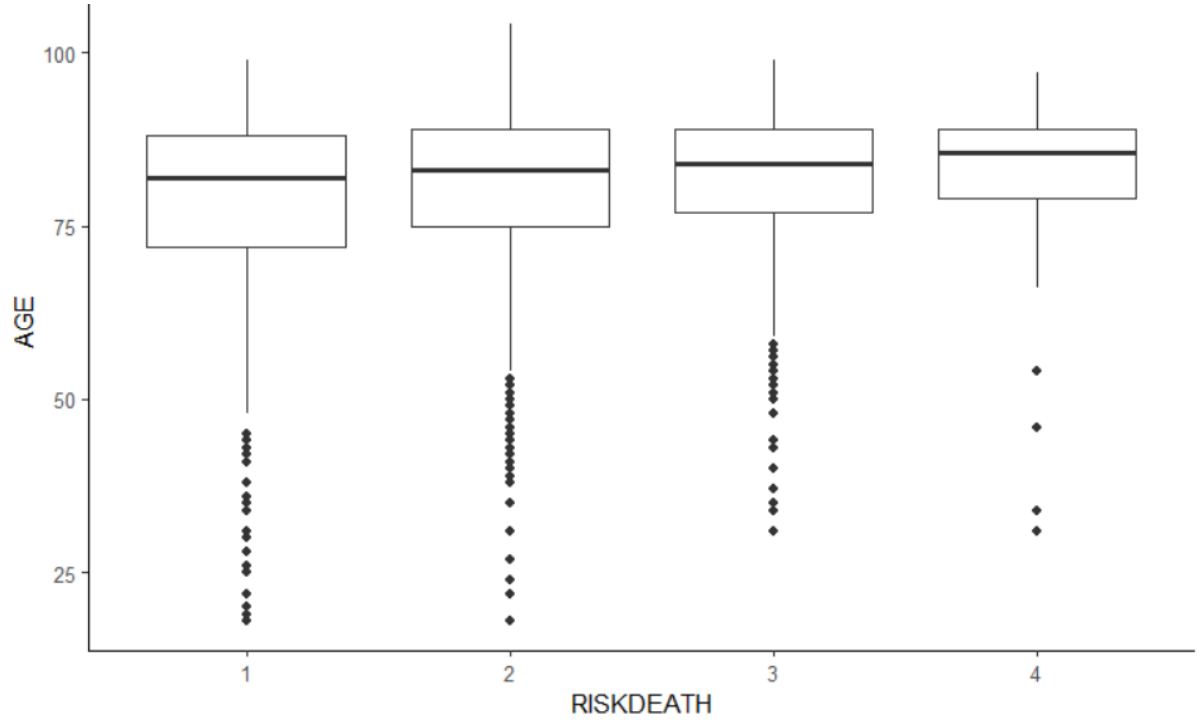


- The proportions of admissions with a level 2 of RISKDEATH constitute the peaks (peaks that are relatively similar) respectively corresponding to the *Generalist* and *Specialist* Wards (behavior found after re-decomposing the groups of wards);
- Admissions with a level 1 of RISKDEATH are more frequent within the *Specialist* Wards than within the *Generalist* ones, whereas for the case of the admissions with a level 3 of RISKDEATH, the situation is reversed;
- Admissions with a level 4 of RISKDEATH are relatively infrequent, whether within the *Generalist* Wards or the *Specialist* ones.



- The admissions' Peaks with a level 2 of RISKDEATH of the *Generalist* and *Specialist* groups are once again (re-)found after the re-decomposition of the latters;
- Admissions with a level 1 of RISKDEATH are particularly more frequent for the case of the Ward 08, whereas, those with a level 3 of RISKDEATH are particularly infrequent;
- Admissions with a level 4 of RISKDEATH are somewhat rare, only observed in very small proportions within the Wards 21, 24, 2604 and 2605;
- Admissions with a level 3 of RISKDEATH are more or less frequent within all the Wards, the Ward 08 making a small exception to this behavior, with this type of admission particularly poorly observed for its case.

II.9- Ages distribution in function of the Risk of Death



- From these BoxPlots, we can notice :
  - Very clearly, RISKDEATHs concern mainly the older patients (older than 70 years old), with however a few outliers for each BoxPlot;
  - The 4 BoxPlots are relatively at the same level of age, have relatively the same values of Median, but their distributions are slightly different in function of the levels of RISKDEATH;
  - By starting our reading from the level 1 to the level 4 on the axis dedicated to the RISKDEATH, we clearly observe that the variability of the data on the ages continually decrease, as well as the quantities of outliers (the highest levels of RISKDEATH concern mostly the older patients) .

```
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 : Variances are equal

```
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
```

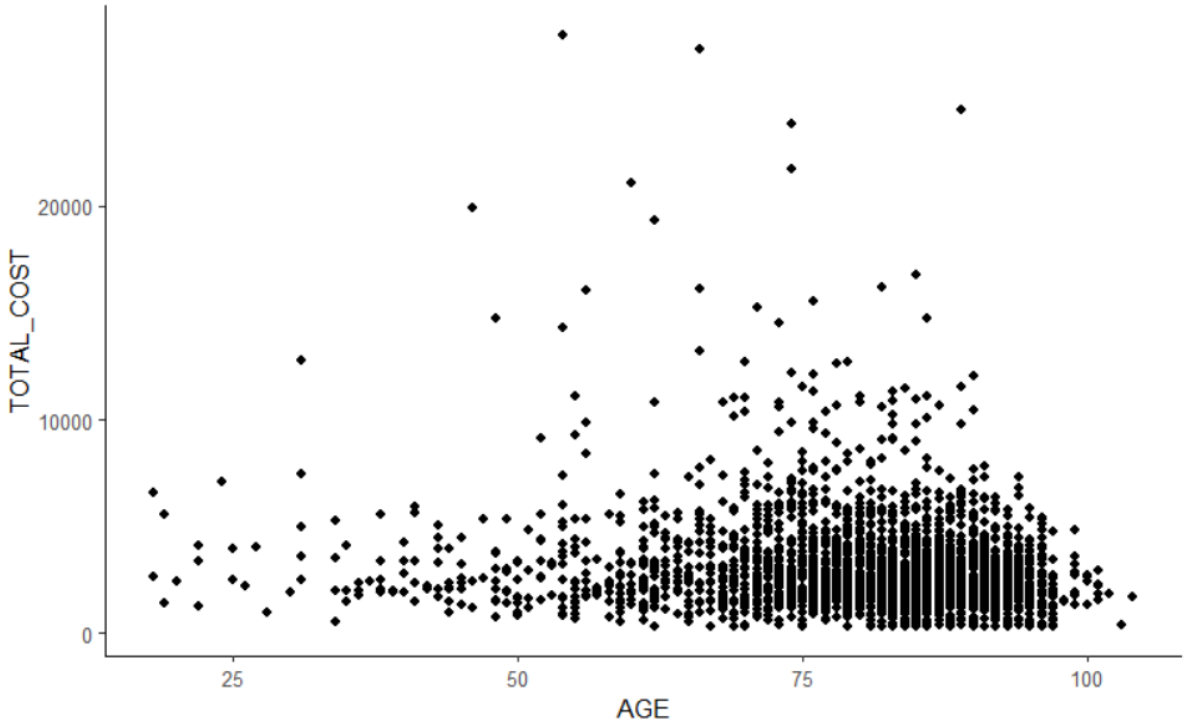
```
Kruskal-Wallis rank sum test

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
```

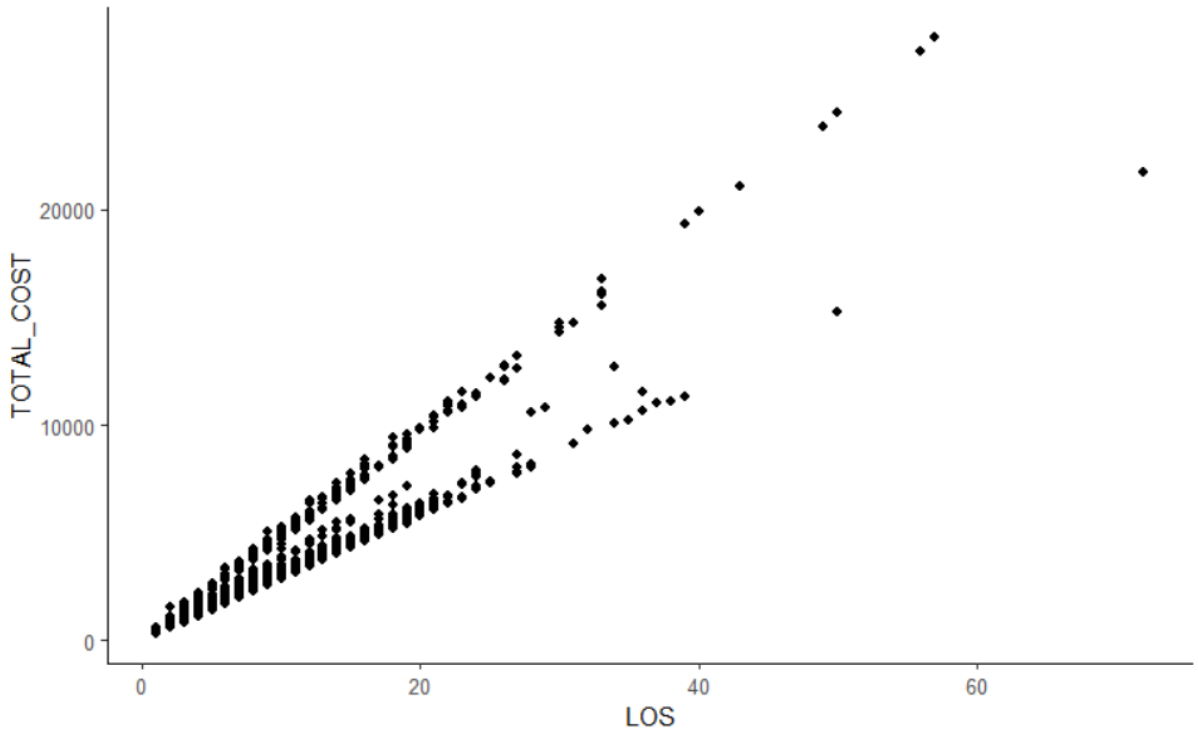
- There is a difference of the age between :
  - Level 1 of RISKDEATH vs Level 2 of RISKDEATH ;
  - Level 1 of RISKDEATH vs Level 3 of RISKDEATH ;
  - Level 2 of RISKDEATH vs Level 3 of RISKDEATH.

II.10- Ages Distribution in function of Total Costs



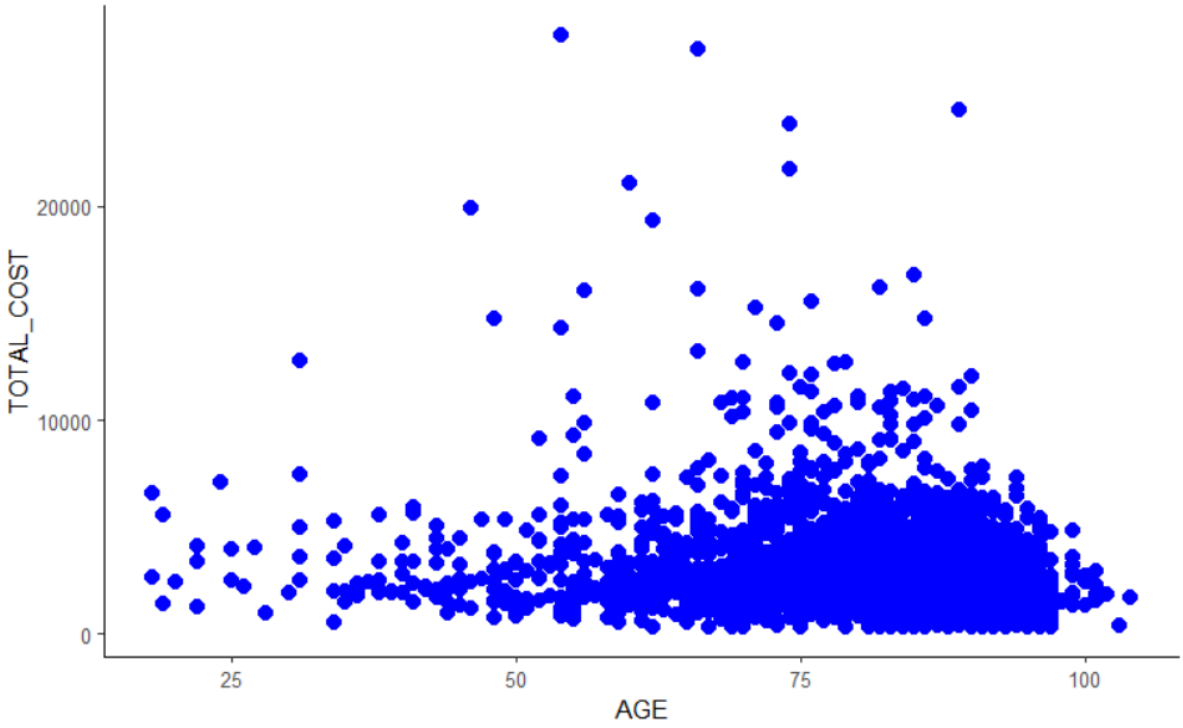
- We can observe that the Total Costs mainly concern elderly patients, and less the younger ones;
- However, we notice that the relationship tends to be *negative* between the TOTAL_COST and the AGE when it comes to the highest costs, these latters concern the most the patients admittedly the elderly ones (i.e. >60 years old), but at the same time not among the oldest ones (i.e. <90 years old), and constitute a significant proportion of the patients.

II.11- Total Costs distribution in function of the Lengths of Stay



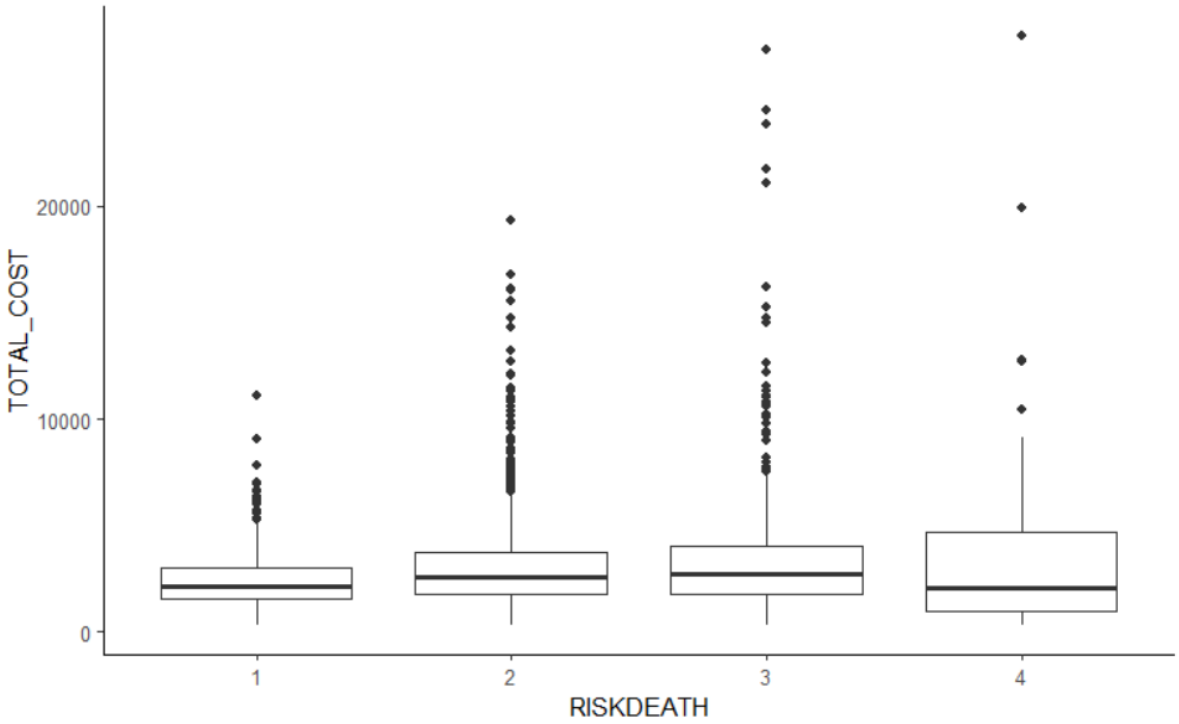
- We can observe that the relationship existing between the LOS and the TOTAL_COST is a Positive one;
- The higher the length of stay is, the higher the total costs are (which is quite logical).

II.12- Total Costs distribution in function of the Age



- We can observe that the elderly patients are the most concerned by the Total Costs than the younger ones;
- However, we can notice that the relationship between the AGE and the TOTAL_COST tends to be negative: elderly patients who don't belong to the group of the oldest patients (i.e. mainly sexagenarians, septuagenarians and some octogenarians) are the main concerned by the highest Total Costs.

II.13- Total Costs distribution in function of the Risks of Death:



- From these BoxPlots, we can notice that :
  - The 4 BoxPolts are at the same level (of TOTAL_COST), relatively have the same values of Median, but with slightly different data distributions in function of the levels of RISKDEATH 1, 2 and 3, and a particular significant variability of data is observed for the case of the level 4 of RISKDEATH;
  - By starting our reading from the level 1 to the level 4 on the axis dedicated to the RISKDEATH, we clearly observe that the variability of data on the TOTAL_COST increases continuously;
  - Outliers in Total Costs are more significant for the levels of RISK_DEATH 2 and 3, and less significant for the levels 1 and 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
```

- Equality of Variance verified.

```
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
```

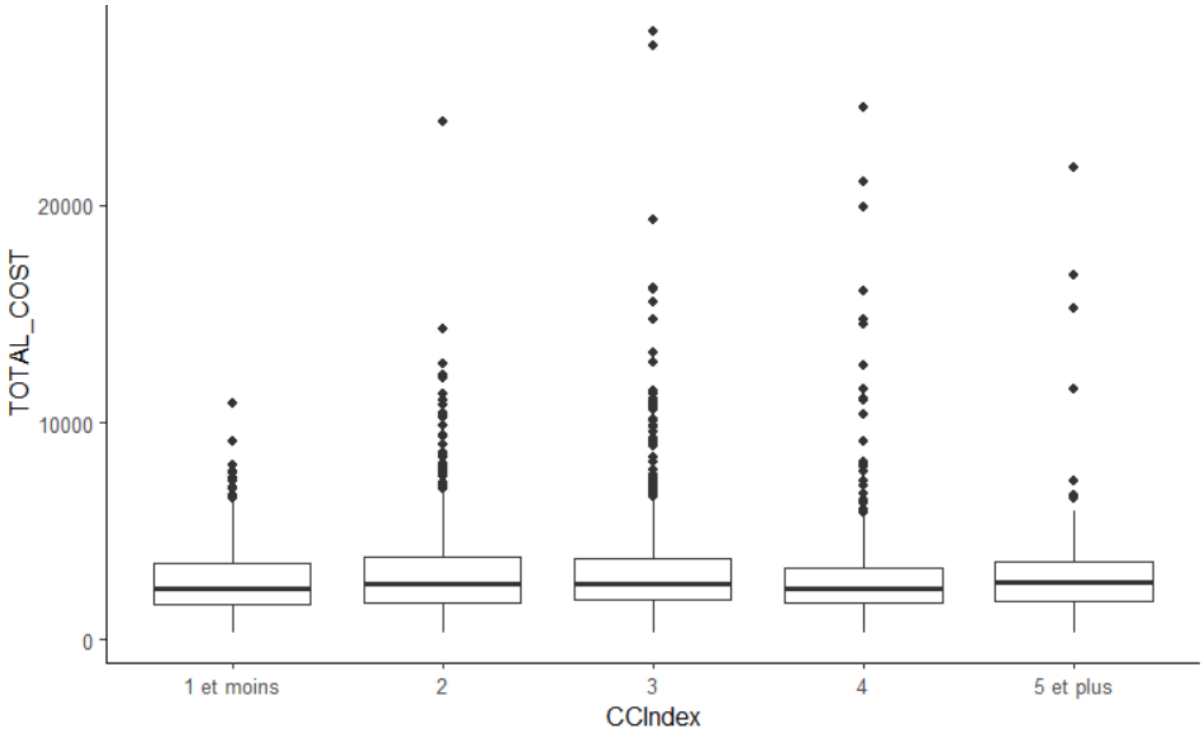
- There is a difference of Total Costs in function of the Risks of Death.

```
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
```

- We observe differences of cost between:
  - Risk of Death level 1 vs Risk of Death level 2;
  - Risk of Death level 1 vs Risk of Death level 3;

II.14- Total Costs distribution in function of the CCI index



- From these BoxPlots, we can notice that:
  - The 5 BoxPlots are at the same level (of TOTAL_COST), relatively have the same values of median of Total Costs, and have slightly different data distributions in function of the CCI index;
  - Outliers (high ones) are more significant for the indexes of CCI 2, 3 and 4, and relatively less significant for the indexes *« 1 and less »* et *« 5 and more »;*

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
```

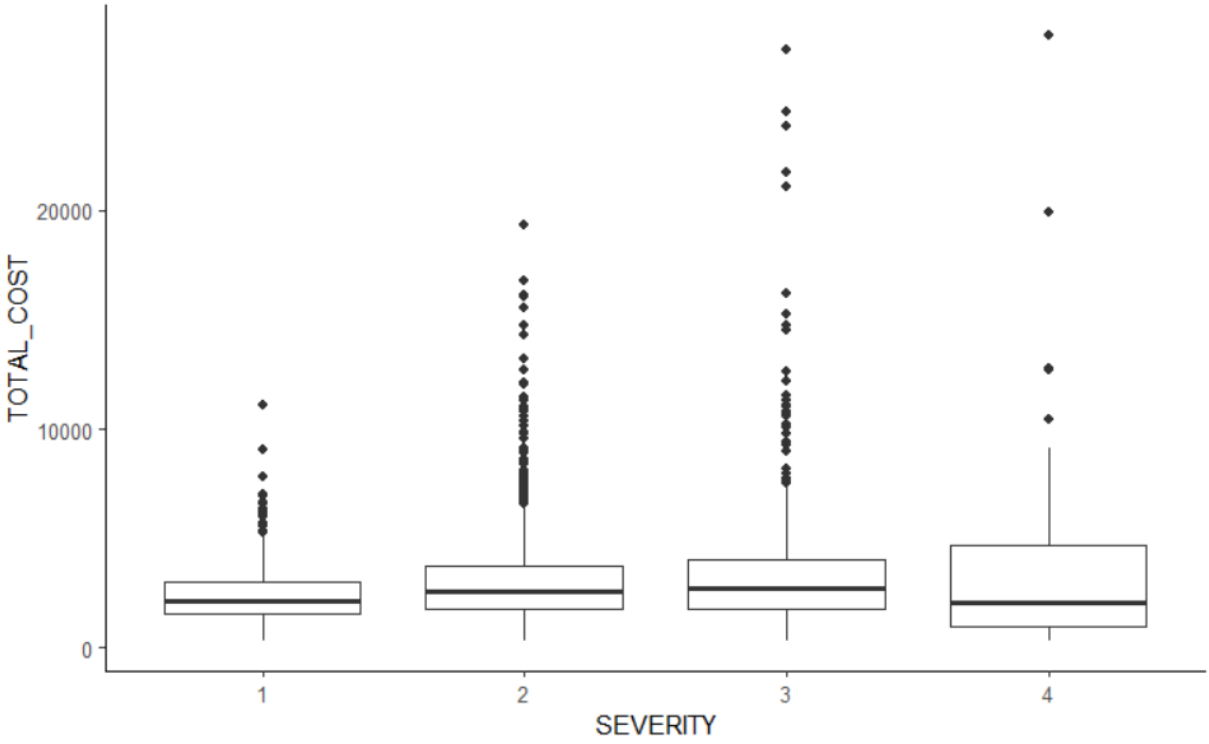
```
aov.CCITTC <- aov(TOTAL_COST ~ CCIndex, data = df_TTC_CCI)
summary(aov.CCITTC)

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
```

- There is no significant difference of cost in function of the CCI.

II.15- Total Costs distribution in function of the severity



- From these BoxPlots, we can observe :
  - The 4 BoxPlots have relatively the same values of Median of TOTAL_COST, but with slightly different data distributions in function of the indexes 1, 2 and 3. Particularly, higher Total Costs values and also a higher variability of data are observed for the case of the index 4;
  - By starting our reading from the index 1 to the index 3 on the axis dedicated to the SEVERITY, we clearly see that the variability of data on the TOTAL_COST increases slightly and continuously, whereas this increase of the variability is particularly *abrupt* compared to that of the previous ones for the case of the BoxPlot which corresponds to the index of SEVERITY 4;
  - Outliers on Total Costs are more significant for the indexes of SEVERITY 2 and 3, but less significant for the levels 1 and 4 (however, they are more dispersed for the particular case of this latter).

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
```

- Equality of Variance verified.

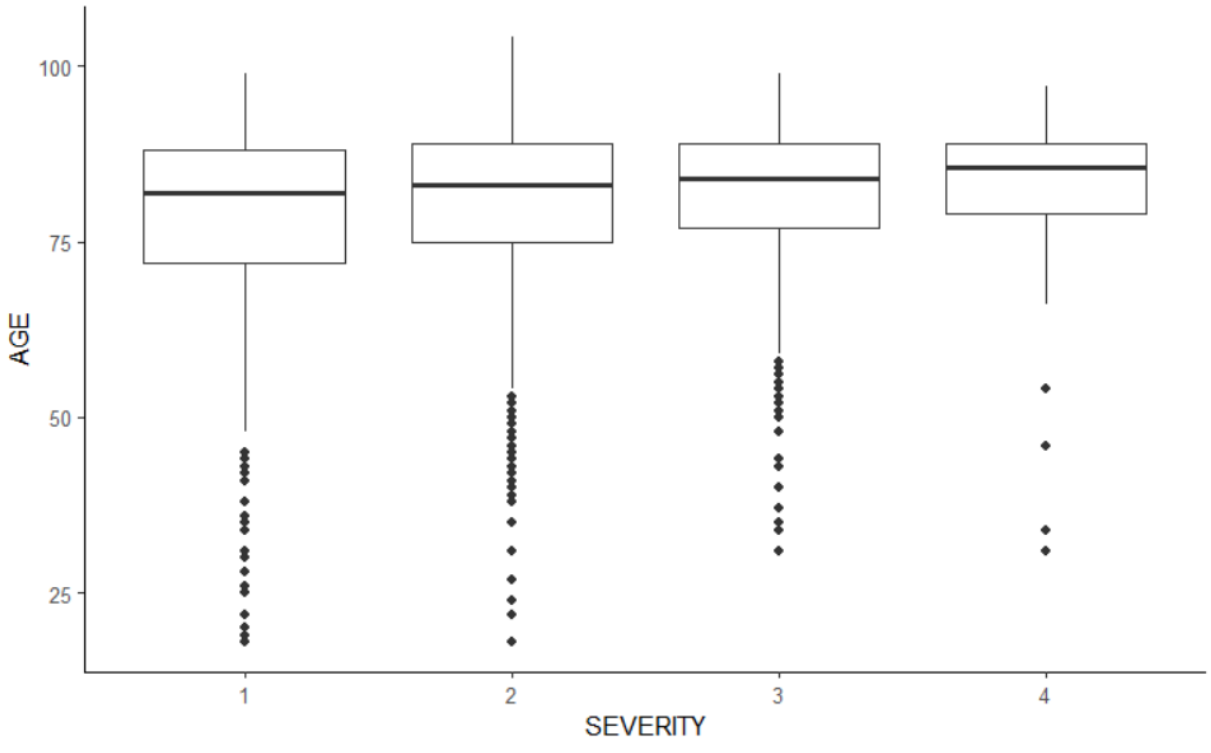
```
kruskal.SEVTTC <- kruskal.test(TOTAL_COST ~ SEVERITY, data = df_TTC_SEV)
kruskal.SEVTTC

    Kruskal-Wallis rank sum test

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

- We observe a difference of cost in function of the severity.

```
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
```

- There is a difference of cost between:
  - Severity 1 vs Severity 2 ;
  - Severity 1 vs Severity 3.

II.16- Age distribution in function of the Severity



- From these BoxPlots, we can observe:
  - Very clearly, SEVERITYs concern mainly elderly patients (70 years old and more), with nevertheless some outliers on the ages of each BoxPlot;
  - The 4 Boxplots are relatively at the same level (of Age), have relatively the same values of Median of Age, but with different data distributions in function of the Index of Severity;
  - By starting our reading from the index of SEVERITY 1 to the index 4 on the axis dedicated to the SEVERITY, we clearly see that the data variability on the Ages decreases continuously, as well as the quantity of outliers (the highest indexes of SEVERITY concern more elderly patients than the younger ones).

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
```

- Equality of variance verified.

```
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
```

- There is a difference of Age in function of the severity.

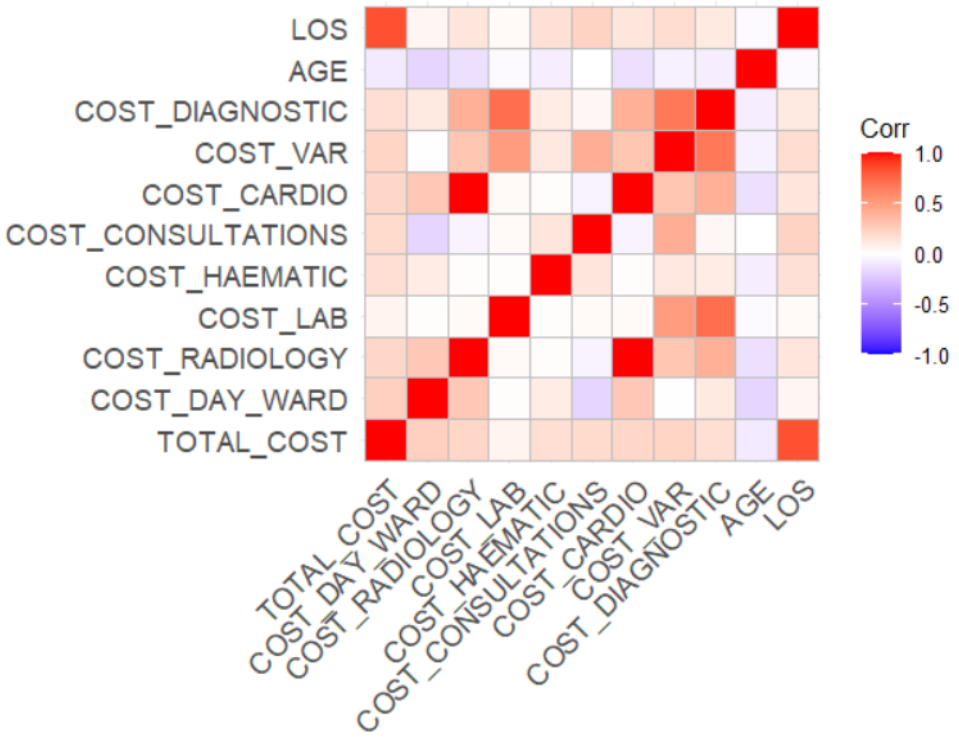
```
pairwise.wilcox.test(df_AGE_SEV$AGE, df_AGE_SEV$SEVERITY,p.adjust.method =
"BH")

    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
```

- There is a difference of AGE between :
  - Severity 1 vs Severity 2 ;
  - Severity 2 vs Severity 3 ;
  - Severity 1 vs Severity 2.

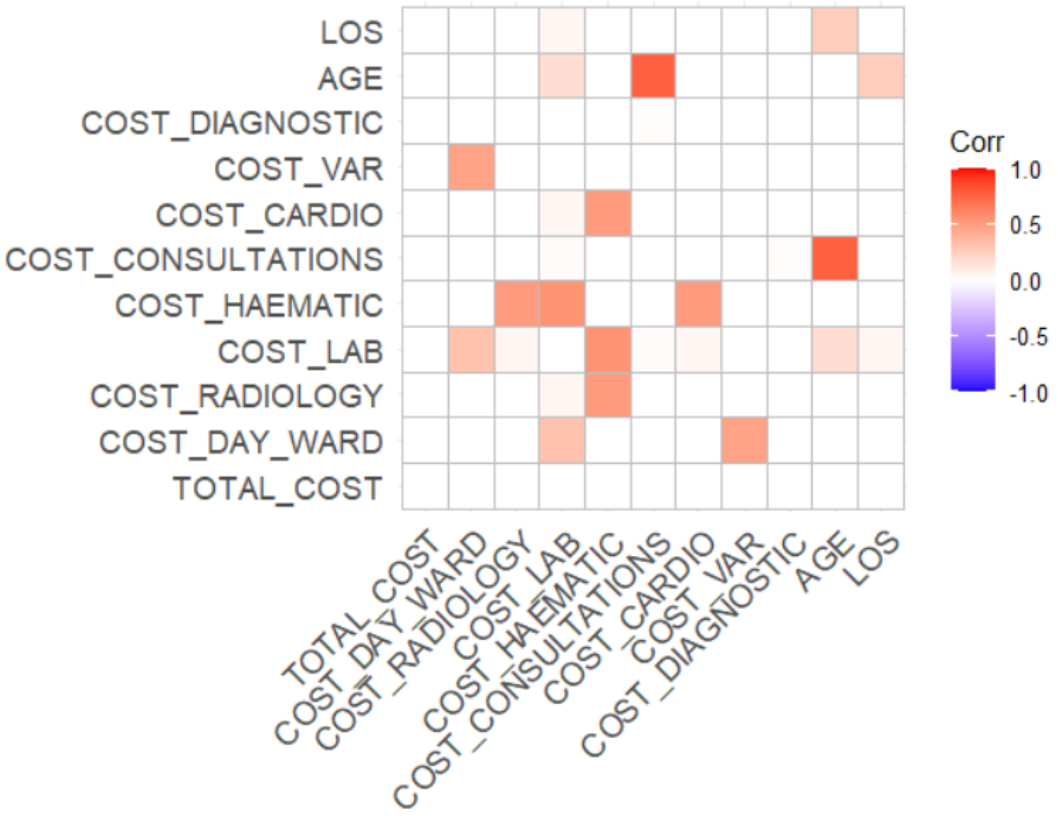
II.17- Correlations of quantitative variables (Strong correlation between the lengths of stay and the total costs)

```
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
```

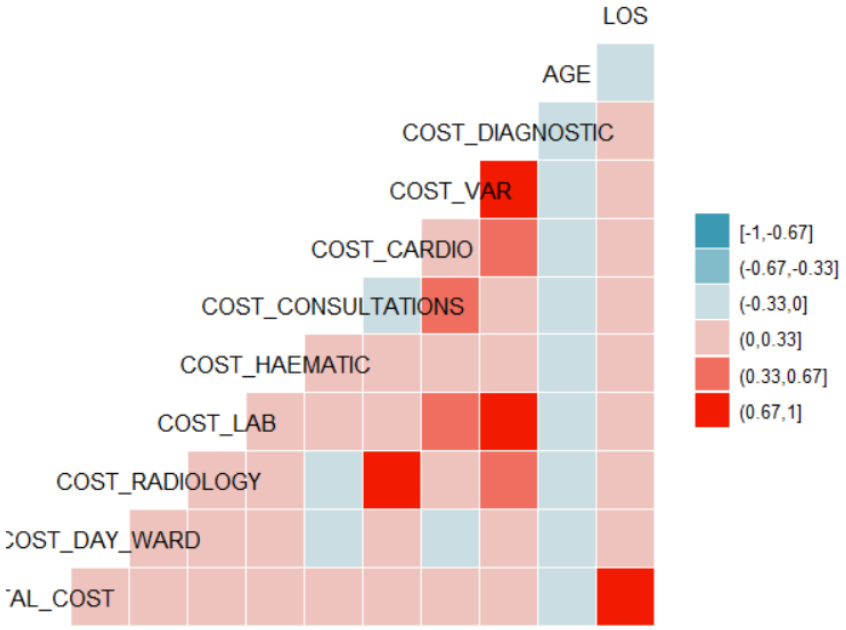
```
df_corr_quantVar <-  df %>%
  select(TOTAL_COST, starts_with("COST_"), AGE, LOS) %>%
  cor(., method = 'kendall')
ggcorrplot(df_corr_quantVar)
```



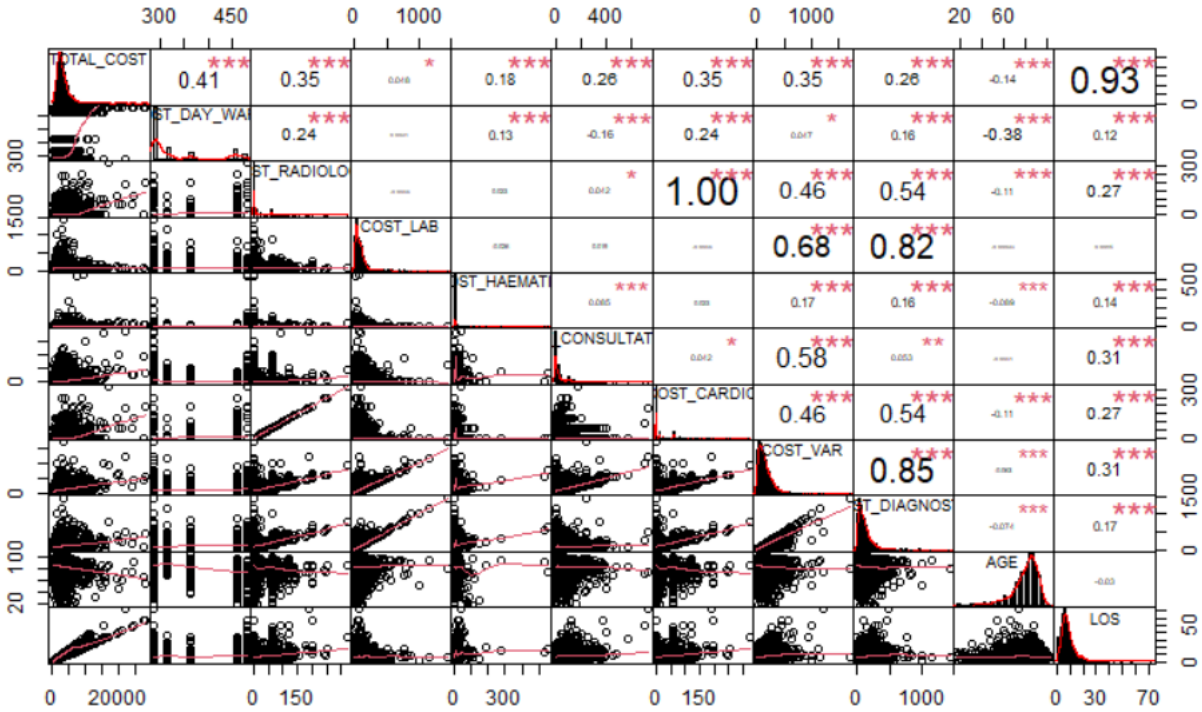

```
df_pmat <-  df %>%
  select(TOTAL_COST, starts_with("COST_"), AGE, LOS) %>%
  ggcorrplot::cor_pmat(., method = 'kendall')
ggcorrplot(df_pmat)
```



```
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)
```



- The correlation coefficient between the LOS and the TOTAL_COST is equal to 0.93, very close to 1;
- The correlation between the LOS and TOTAL_COST is therefore a Strong one.