Projet 2/5 Statistiques Descriptives

Nicolas SALVAN - Alexandre CORRIOU

2024-05-17

Ce fichier contient le code pour réaliser les *statistiques descriptives* sur les données nettoyées.

Lecture des données nettoyées

Importation du dataset

```
data <- read.csv("data/data_cleaned.csv", header = TRUE, sep = ",")
data$diagnosis <- as.factor(data$diagnosis)
```

Aperçu rapide

```
head(data)
```

```
##   diagnosis radius_mean texture_mean perimeter_mean area_mean smoothness_mean
## 1         M       17.99        10.38         122.80    1001.0         0.11840
## 2         M       20.57        17.77         132.90    1326.0         0.08474
## 3         M       19.69        21.25         130.00    1203.0         0.10960
## 4         M       11.42        20.38          77.58     386.1         0.14250
## 5         M       20.29        14.34         135.10    1297.0         0.10030
## 6         M       12.45        15.70          82.57     477.1         0.12780
##   compactness_mean concavity_mean concave.points_mean symmetry_mean
## 1          0.27760         0.3001             0.14710        0.2419
## 2          0.07864         0.0869             0.07017        0.1812
## 3          0.15990         0.1974             0.12790        0.2069
## 4          0.28390         0.2414             0.10520        0.2597
## 5          0.13280         0.1980             0.10430        0.1809
## 6          0.17000         0.1578             0.08089        0.2087
##   fractal_dimension_mean radius_se texture_se perimeter_se area_se
## 1                0.07871    1.0950     0.9053        8.589  153.40
## 2                0.05667    0.5435     0.7339        3.398   74.08
## 3                0.05999    0.7456     0.7869        4.585   94.03
## 4                0.09744    0.4956     1.1560        3.445   27.23
## 5                0.05883    0.7572     0.7813        5.438   94.44
## 6                0.07613    0.3345     0.8902        2.217   27.19
##   smoothness_se compactness_se concavity_se concave.points_se symmetry_se
## 1      0.006399        0.04904      0.05373           0.01587     0.03003
```

```
## 2      0.005225         0.01308       0.01860           0.01340      0.01389
## 3      0.006150         0.04006       0.03832           0.02058      0.02250
## 4      0.009110         0.07458       0.05661           0.01867      0.05963
## 5      0.011490         0.02461       0.05688           0.01885      0.01756
## 6      0.007510         0.03345       0.03672           0.01137      0.02165
##   fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst
## 1             0.006193        25.38         17.33          184.60     2019.0
## 2             0.003532        24.99         23.41          158.80     1956.0
## 3             0.004571        23.57         25.53          152.50     1709.0
## 4             0.009208        14.91         26.50           98.87      567.7
## 5             0.005115        22.54         16.67          152.20     1575.0
## 6             0.005082        15.47         23.75          103.40      741.6
##   smoothness_worst compactness_worst concavity_worst concave.points_worst
## 1           0.1622            0.6656          0.7119               0.2654
## 2           0.1238            0.1866          0.2416               0.1860
## 3           0.1444            0.4245          0.4504               0.2430
## 4           0.2098            0.8663          0.6869               0.2575
## 5           0.1374            0.2050          0.4000               0.1625
## 6           0.1791            0.5249          0.5355               0.1741
##   symmetry_worst fractal_dimension_worst
## 1         0.4601                 0.11890
## 2         0.2750                 0.08902
## 3         0.3613                 0.08758
## 4         0.6638                 0.17300
## 5         0.2364                 0.07678
## 6         0.3985                 0.12440
```

```
# dim(data)
# str(data)
```

Statistiques descriptives

Nous allons maintenant réaliser des statistiques descriptives sur les données nettoyées.

Résumé des données

```
summary(data)
```

```
##  diagnosis  radius_mean      texture_mean   perimeter_mean     area_mean
##  B:357     Min.   : 6.981   Min.   : 9.71   Min.   : 43.79   Min.   : 143.5
##  M:212     1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17   1st Qu.: 420.3
##            Median :13.370   Median :18.84   Median : 86.24   Median : 551.1
##            Mean   :14.127   Mean   :19.29   Mean   : 91.97   Mean   : 654.9
##            3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10   3rd Qu.: 782.7
##            Max.   :28.110   Max.   :39.28   Max.   :188.50   Max.   :2501.0
##  smoothness_mean   compactness_mean  concavity_mean    concave.points_mean
##  Min.   :0.05263   Min.   :0.01938   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.08637   1st Qu.:0.06492   1st Qu.:0.02956   1st Qu.:0.02031
##  Median :0.09587   Median :0.09263   Median :0.06154   Median :0.03350
##  Mean   :0.09636   Mean   :0.10434   Mean   :0.08880   Mean   :0.04892
```

```
##  3rd Qu.:0.10530   3rd Qu.:0.13040   3rd Qu.:0.13070   3rd Qu.:0.07400
##  Max.   :0.16340   Max.   :0.34540   Max.   :0.42680   Max.   :0.20120
##  symmetry_mean    fractal_dimension_mean   radius_se        texture_se
##  Min.   :0.1060   Min.   :0.04996        Min.   :0.1115   Min.   :0.3602
##  1st Qu.:0.1619   1st Qu.:0.05770        1st Qu.:0.2324   1st Qu.:0.8339
##  Median :0.1792   Median :0.06154        Median :0.3242   Median :1.1080
##  Mean   :0.1812   Mean   :0.06280        Mean   :0.4052   Mean   :1.2169
##  3rd Qu.:0.1957   3rd Qu.:0.06612        3rd Qu.:0.4789   3rd Qu.:1.4740
##  Max.   :0.3040   Max.   :0.09744        Max.   :2.8730   Max.   :4.8850
##   perimeter_se       area_se        smoothness_se      compactness_se
##  Min.   : 0.757   Min.   :  6.802   Min.   :0.001713   Min.   :0.002252
##  1st Qu.: 1.606   1st Qu.: 17.850   1st Qu.:0.005169   1st Qu.:0.013080
##  Median : 2.287   Median : 24.530   Median :0.006380   Median :0.020450
##  Mean   : 2.866   Mean   : 40.337   Mean   :0.007041   Mean   :0.025478
##  3rd Qu.: 3.357   3rd Qu.: 45.190   3rd Qu.:0.008146   3rd Qu.:0.032450
##  Max.   :21.980   Max.   :542.200   Max.   :0.031130   Max.   :0.135400
##   concavity_se     concave.points_se   symmetry_se       fractal_dimension_se
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.007882   Min.   :0.0008948
##  1st Qu.:0.01509   1st Qu.:0.007638   1st Qu.:0.015160   1st Qu.:0.0022480
##  Median :0.02589   Median :0.010930   Median :0.018730   Median :0.0031870
##  Mean   :0.03189   Mean   :0.011796   Mean   :0.020542   Mean   :0.0037949
##  3rd Qu.:0.04205   3rd Qu.:0.014710   3rd Qu.:0.023480   3rd Qu.:0.0045580
##  Max.   :0.39600   Max.   :0.052790   Max.   :0.078950   Max.   :0.0298400
##   radius_worst   texture_worst   perimeter_worst    area_worst
##  Min.   : 7.93   Min.   :12.02   Min.   : 50.41   Min.   : 185.2
##  1st Qu.:13.01   1st Qu.:21.08   1st Qu.: 84.11   1st Qu.: 515.3
##  Median :14.97   Median :25.41   Median : 97.66   Median : 686.5
##  Mean   :16.27   Mean   :25.68   Mean   :107.26   Mean   : 880.6
##  3rd Qu.:18.79   3rd Qu.:29.72   3rd Qu.:125.40   3rd Qu.:1084.0
##  Max.   :36.04   Max.   :49.54   Max.   :251.20   Max.   :4254.0
##  smoothness_worst  compactness_worst concavity_worst  concave.points_worst
##  Min.   :0.07117   Min.   :0.02729   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.11660   1st Qu.:0.14720   1st Qu.:0.1145   1st Qu.:0.06493
##  Median :0.13130   Median :0.21190   Median :0.2267   Median :0.09993
##  Mean   :0.13237   Mean   :0.25427   Mean   :0.2722   Mean   :0.11461
##  3rd Qu.:0.14600   3rd Qu.:0.33910   3rd Qu.:0.3829   3rd Qu.:0.16140
##  Max.   :0.22260   Max.   :1.05800   Max.   :1.2520   Max.   :0.29100
##  symmetry_worst   fractal_dimension_worst
##  Min.   :0.1565   Min.   :0.05504
##  1st Qu.:0.2504   1st Qu.:0.07146
##  Median :0.2822   Median :0.08004
##  Mean   :0.2901   Mean   :0.08395
##  3rd Qu.:0.3179   3rd Qu.:0.09208
##  Max.   :0.6638   Max.   :0.20750
```

Le jeu de données contient 569 observations et 31 variables. On observe qu'il y a une variable qualitative "diagnosis" qui correspond au diagnostic de la patiente. Toutes les autres variables sont quantitatives, et décrivent les caractéristiques du cancer détecté.

Cette sortie nous donne quelques statistiques descriptives sur nos données, notamment les moyennes, les médianes, les minimums et maximums, et les quartiles.

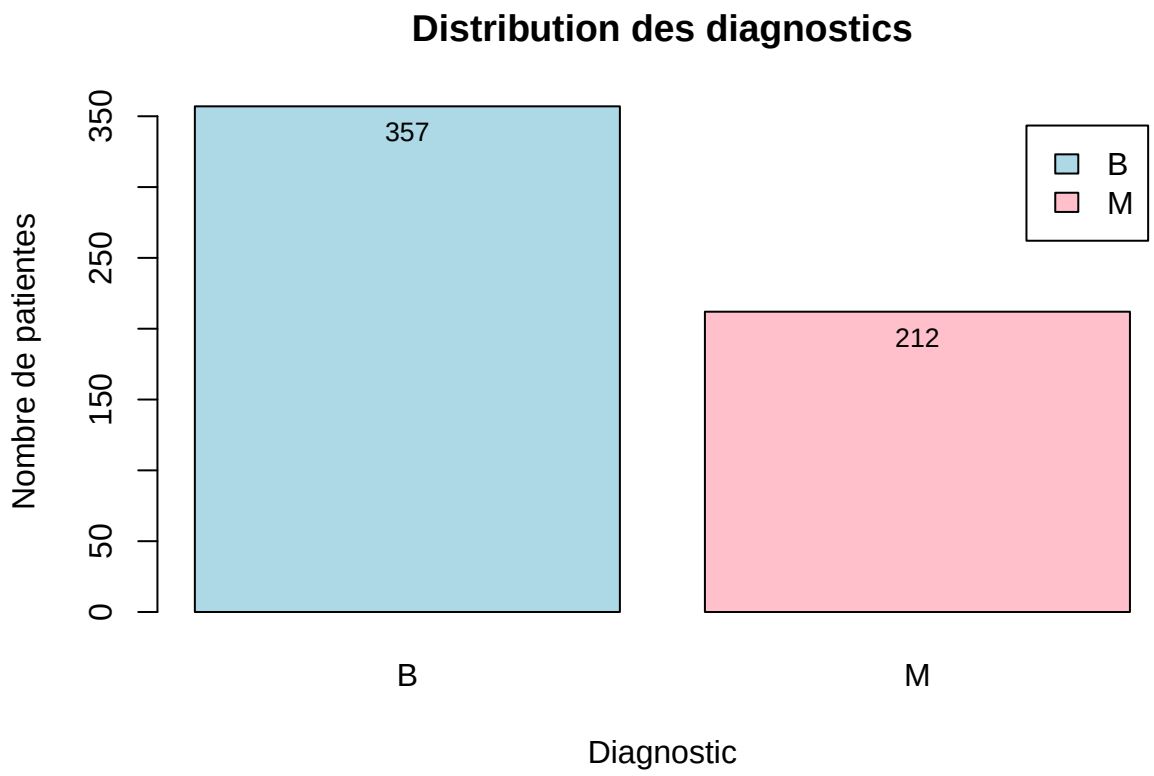
Distribution des données

Observons la distribution des différentes variables.

Distribution des diagnostics (variable qualitative)

```
counts <- table(data$diagnosis)
bp <- barplot(counts,
              main = "Distribution des diagnostics",
              xlab = "Diagnostic",
              ylab = "Nombre de patientes",
              col = c("lightblue", "pink"),
              legend = rownames(counts))

# Ajouter les labels au-dessus des barres
text(x = bp, y = counts, labels = counts, pos = 1, cex = 0.8, col = "black")
```



Ici, on peut voir les effectifs des deux diagnostics possibles : “M” pour “Malignant” ou Malin en français, et “B” pour “Benign” ou bénin.

```
proportions <- prop.table(table(data$diagnosis))
proportions
```

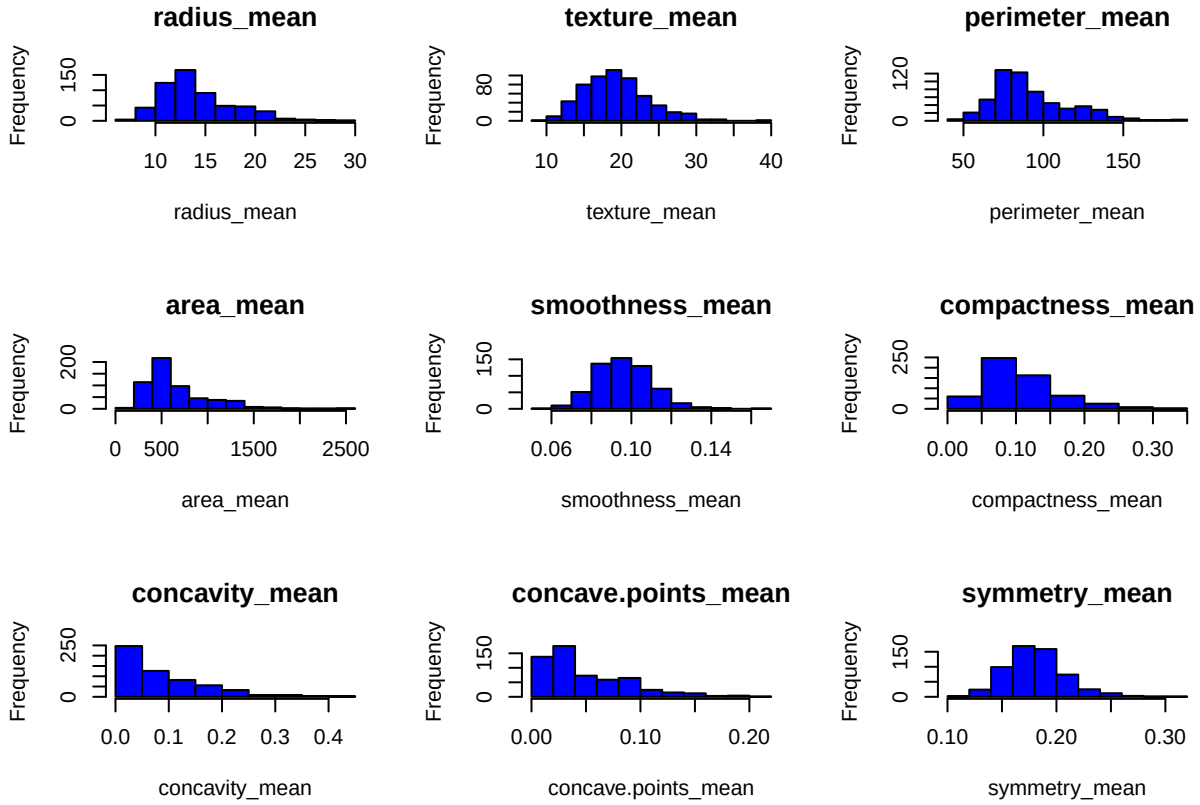
```
##
```

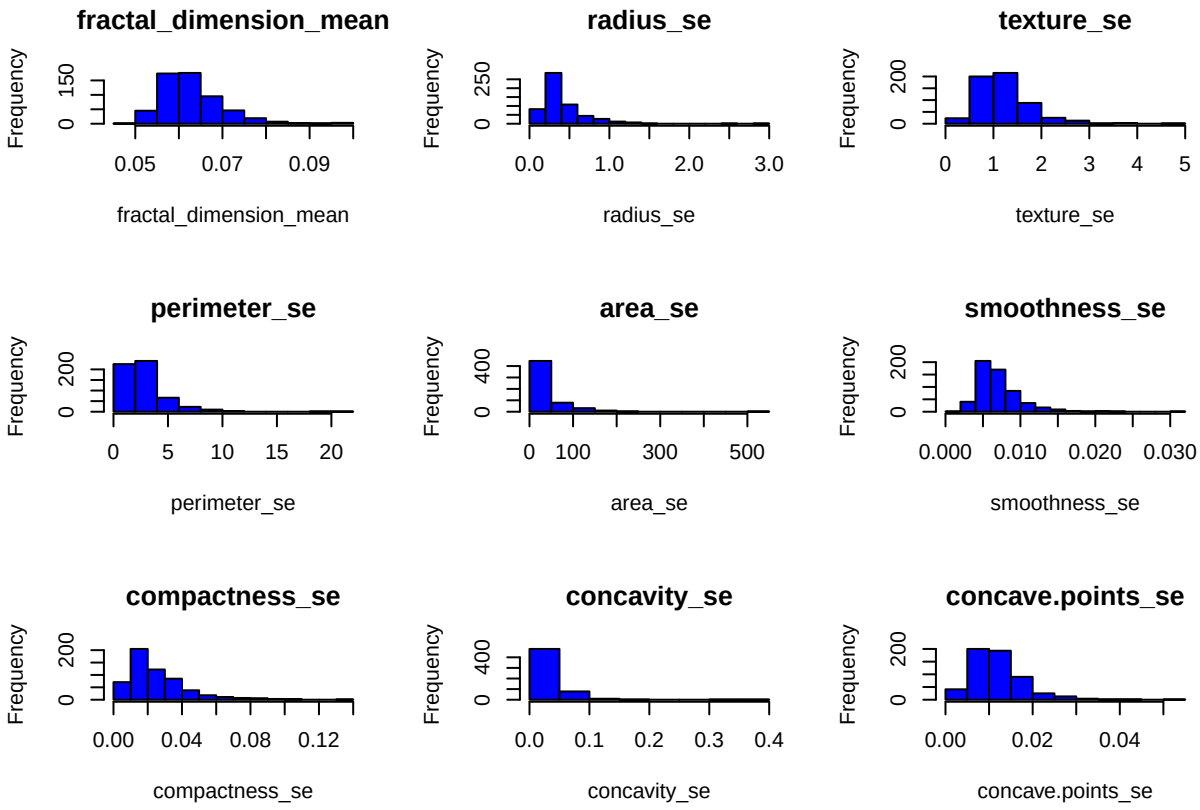
```
##         B         M
## 0.6274165 0.3725835
```

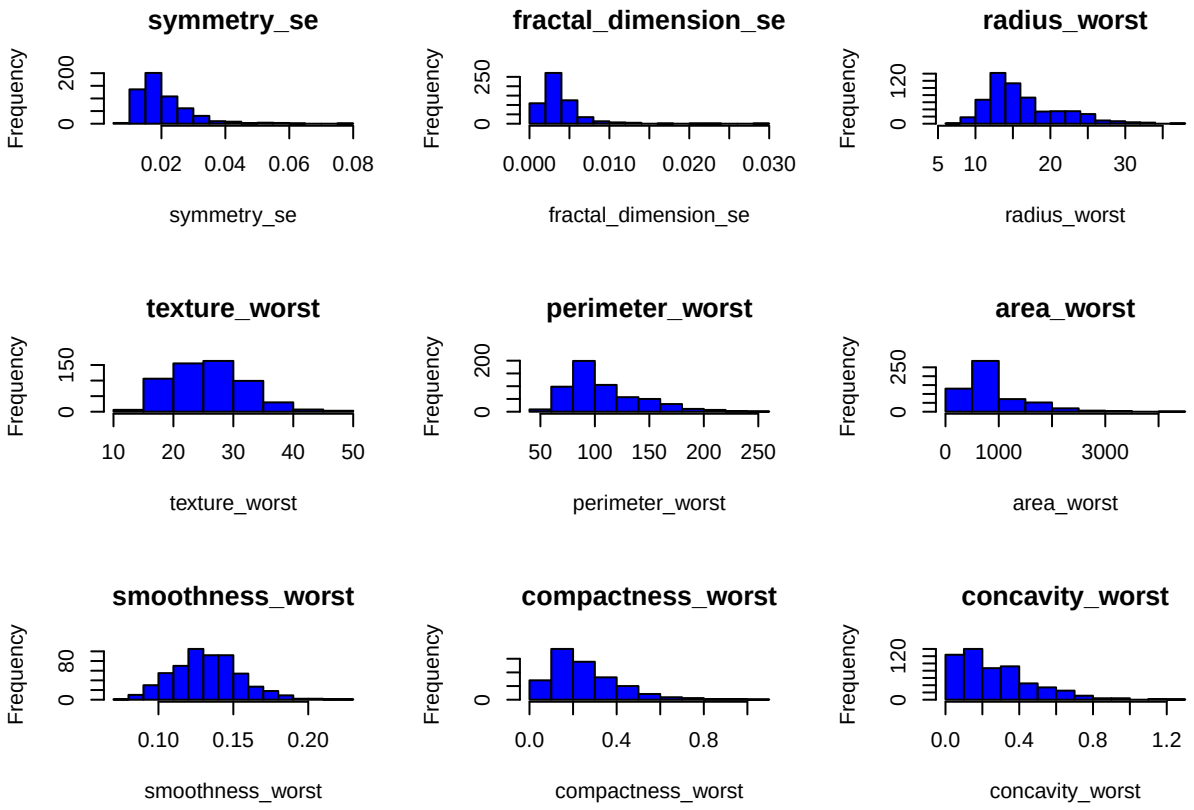
On observe que 63% des patientes ont un diagnostic bénin, et 37% un diagnostic malin. Il faut avoir cela en tête lorsque l'on étudiera notre jeu de données.

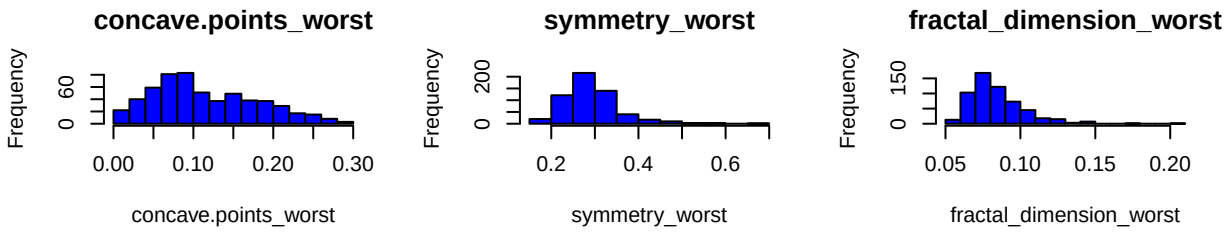
Distribution des variables

```
par(mfrow = c(3,3))
for(i in 2:31){
  hist(data[,i], main = colnames(data)[i], xlab = colnames(data)[i], col = "blue")
}
```




fractal_dimension_mean
Frequency
0
150
0.05
0.07
0.09
fractal_dimension_mean
radius_se
Frequency
0
250
0.0
1.0
2.0
3.0
radius_se
texture_se
Frequency
0
200
0
1
2
3
4
5
texture_se
perimeter_se
Frequency
0
200
0
5
10
15
20
perimeter_se
area_se
Frequency
0
400
0
100
300
500
area_se
smoothness_se
Frequency
0
200
0.000
0.010
0.020
0.030
smoothness_se
compactness_se
Frequency
0
200
0.00
0.04
0.08
0.12
compactness_se
concavity_se
Frequency
0
400
0.0
0.1
0.2
0.3
0.4
concavity_se
concave.points_se
Frequency
0
200
0.00
0.02
0.04
concave.points_se


symmetry_se
Frequency
0
200
0.02
0.04
0.06
0.08
symmetry_se
fractal_dimension_se
Frequency
0
250
0.000
0.010
0.020
0.030
fractal_dimension_se
radius_worst
Frequency
0
120
5
10
20
30
radius_worst
texture_worst
Frequency
0
150
10
20
30
40
50
texture_worst
perimeter_worst
Frequency
0
200
50
100
150
200
250
perimeter_worst
area_worst
Frequency
0
250
0
1000
3000
area_worst
smoothness_worst
Frequency
0
80
0.10
0.15
0.20
smoothness_worst
compactness_worst
Frequency
0
0.0
0.4
0.8
compactness_worst
concavity_worst
Frequency
0
120
0.0
0.4
0.8
1.2
concavity_worst



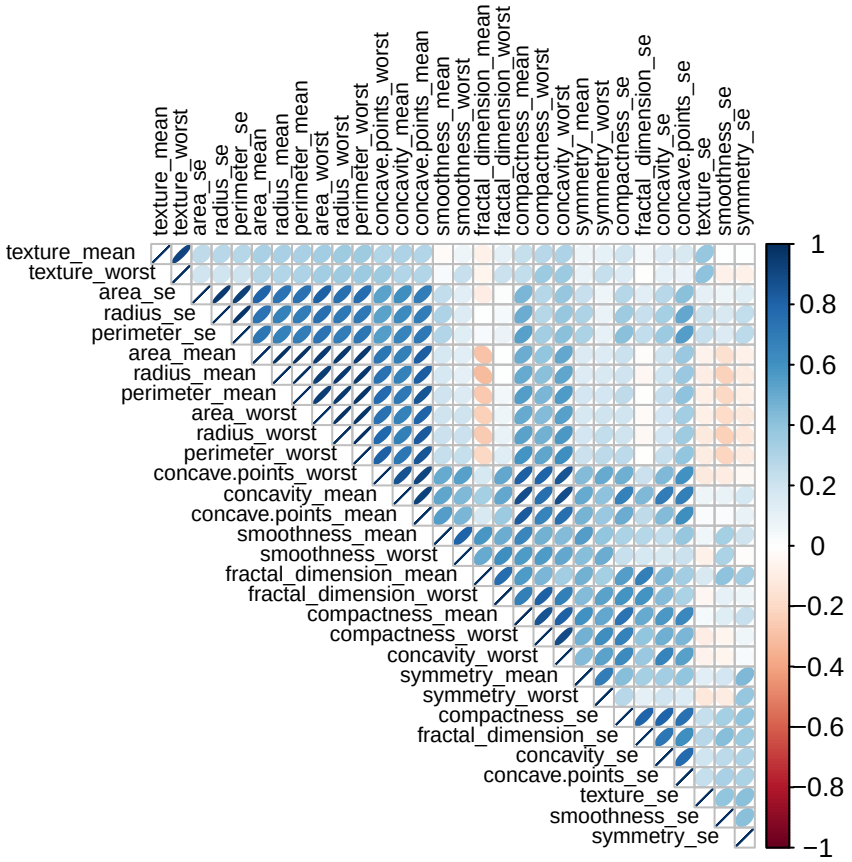
Matrice de corrélation

```
# install.packages("corrplot")
library("corrplot")
```

```
## corrplot 0.92 loaded
```

```
correlation <- cor(data[2:31])
```

```
corrplot(correlation, method = "ellipse", type = "upper", order = "hclust", tl.col = "black", tl.srt = 9
```



On remarque que les variables sont très corrélées entre elles. Il faudra faire attention à la multicollinéarité lors de la modélisation.

On affiche les variables avec un coefficient de corrélation supérieur à 0.98.

```
# Fonction carrément volée sur internet https://rpubs.com/sediaz/Correlations
corr_check <- function(Dataset, threshold){
  matriz_cor <- cor(Dataset)
  matriz_cor

  for (i in 1:nrow(matriz_cor)){
    correlations <-  which((abs(matriz_cor[i,i:ncol(matriz_cor)]) > threshold) & (matriz_cor[i,i:ncol(ma

    if(length(correlations)> 0){
      lapply(correlations,FUN =  function(x) (cat(paste(colnames(Dataset)[i], "with",colnames(Dataset)[

    }
  }
}
```

```
corr_check(data[2:31], 0.98)
```

```
## radius_mean with perimeter_mean
## radius_mean with area_mean
## perimeter_mean with texture_mean
## radius_worst with perimeter_mean
## radius_worst with area_mean
```

On remarque que les colonnes liées sont le rayon, le périmètre, l'aire. On va supprimer le périmètre car, de part la forme circulaire des cancers, il peut être calculé comme étant 2 * pi * rayon. On devrait pouvoir l'observer dans les prochaines étapes de notre analyse.

Conclusion

Nous avons réalisé des statistiques descriptives sur notre jeu de données nettoyé. Nous avons pu observer la distribution des diagnostics, et des différentes variables. Nous avons également étudié la corrélation entre les variables, et avons identifié des variables fortement corrélées, qui devraient être prises en compte lors de la modélisation.