SC1_Proj

Alessio

13 January 2020

Description of the dataset and problem

TODO: describe

```
data <- read_csv("../data/data.csv")
```

```
## Warning: Missing column names filled in: 'X33' [33]
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   diagnosis = col_character(),
##   X33 = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
## Warning: 569 parsing failures.
## row col   expected     actual                file
##   1  -- 33 columns 32 columns '../data/data.csv'
##   2  -- 33 columns 32 columns '../data/data.csv'
##   3  -- 33 columns 32 columns '../data/data.csv'
##   4  -- 33 columns 32 columns '../data/data.csv'
##   5  -- 33 columns 32 columns '../data/data.csv'
## ... ... .......... .......... ..................
## See problems(...) for more details.
```

```
glimpse(data)
```

```
## Observations: 569
## Variables: 33
## $ id                      <dbl> 842302, 842517, 84300903, 84348301, 8435840...
## $ diagnosis               <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M"...
## $ radius_mean             <dbl> 17.990, 20.570, 19.690, 11.420, 20.290, 12....
## $ texture_mean            <dbl> 10.38, 17.77, 21.25, 20.38, 14.34, 15.70, 1...
## $ perimeter_mean          <dbl> 122.80, 132.90, 130.00, 77.58, 135.10, 82.5...
## $ area_mean               <dbl> 1001.0, 1326.0, 1203.0, 386.1, 1297.0, 477....
## $ smoothness_mean         <dbl> 0.11840, 0.08474, 0.10960, 0.14250, 0.10030...
## $ compactness_mean        <dbl> 0.27760, 0.07864, 0.15990, 0.28390, 0.13280...
## $ concavity_mean          <dbl> 0.30010, 0.08690, 0.19740, 0.24140, 0.19800...
## $ `concave points_mean`   <dbl> 0.14710, 0.07017, 0.12790, 0.10520, 0.10430...
## $ symmetry_mean           <dbl> 0.2419, 0.1812, 0.2069, 0.2597, 0.1809, 0.2...
## $ fractal_dimension_mean  <dbl> 0.07871, 0.05667, 0.05999, 0.09744, 0.05883...
## $ radius_se               <dbl> 1.0950, 0.5435, 0.7456, 0.4956, 0.7572, 0.3...
## $ texture_se              <dbl> 0.9053, 0.7339, 0.7869, 1.1560, 0.7813, 0.8...
## $ perimeter_se            <dbl> 8.589, 3.398, 4.585, 3.445, 5.438, 2.217, 3...
## $ area_se                 <dbl> 153.40, 74.08, 94.03, 27.23, 94.44, 27.19, ...
## $ smoothness_se           <dbl> 0.006399, 0.005225, 0.006150, 0.009110, 0.0...
## $ compactness_se          <dbl> 0.049040, 0.013080, 0.040060, 0.074580, 0.0...
## $ concavity_se            <dbl> 0.05373, 0.01860, 0.03832, 0.05661, 0.05688...
## $ `concave points_se`     <dbl> 0.015870, 0.013400, 0.020580, 0.018670, 0.0...
## $ symmetry_se             <dbl> 0.03003, 0.01389, 0.02250, 0.05963, 0.01756...
## $ fractal_dimension_se    <dbl> 0.006193, 0.003532, 0.004571, 0.009208, 0.0...
## $ radius_worst            <dbl> 25.38, 24.99, 23.57, 14.91, 22.54, 15.47, 2...
## $ texture_worst           <dbl> 17.33, 23.41, 25.53, 26.50, 16.67, 23.75, 2...
## $ perimeter_worst         <dbl> 184.60, 158.80, 152.50, 98.87, 152.20, 103....
## $ area_worst              <dbl> 2019.0, 1956.0, 1709.0, 567.7, 1575.0, 741....
## $ smoothness_worst        <dbl> 0.1622, 0.1238, 0.1444, 0.2098, 0.1374, 0.1...
## $ compactness_worst       <dbl> 0.6656, 0.1866, 0.4245, 0.8663, 0.2050, 0.5...
## $ concavity_worst         <dbl> 0.71190, 0.24160, 0.45040, 0.68690, 0.40000...
## $ `concave points_worst`  <dbl> 0.26540, 0.18600, 0.24300, 0.25750, 0.16250...
## $ symmetry_worst          <dbl> 0.4601, 0.2750, 0.3613, 0.6638, 0.2364, 0.3...
## $ fractal_dimension_worst <dbl> 0.11890, 0.08902, 0.08758, 0.17300, 0.07678...
## $ X33                     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
```

```r
colnames(data)[3:32] <- c('radius_m','texture_m', 'perim_m','area_m','smooth_m','compact_m','concav_m',
```

Dataset Preprocessing Visualisation and Exploration

```r
colSums(is.na(data))
```

```
##           id    diagnosis     radius_m    texture_m      perim_m       area_m
##            0            0            0            0            0            0
##     smooth_m    compact_m     concav_m  concav_pt_m   symmetry_m   frac_dim_m
##            0            0            0            0            0            0
##    radius_se   texture_se     perim_se      area_se    smooth_se   compact_se
##            0            0            0            0            0            0
##    concav_se concav_pt_se  symmetry_se  frac_dim_se     radius_w    texture_w
##            0            0            0            0            0            0
##      perim_w       area_w     smooth_w    compact_w     concav_w  concav_pt_w
##            0            0            0            0            0            0
##   symmetry_w   frac_dim_w          X33
##            0            0          569
```

```r
data %<>% mutate_at(vars(diagnosis), factor)
```

```r
train <- data %>% sample_frac(0.8)
test <- anti_join(data,train, by='id')
```

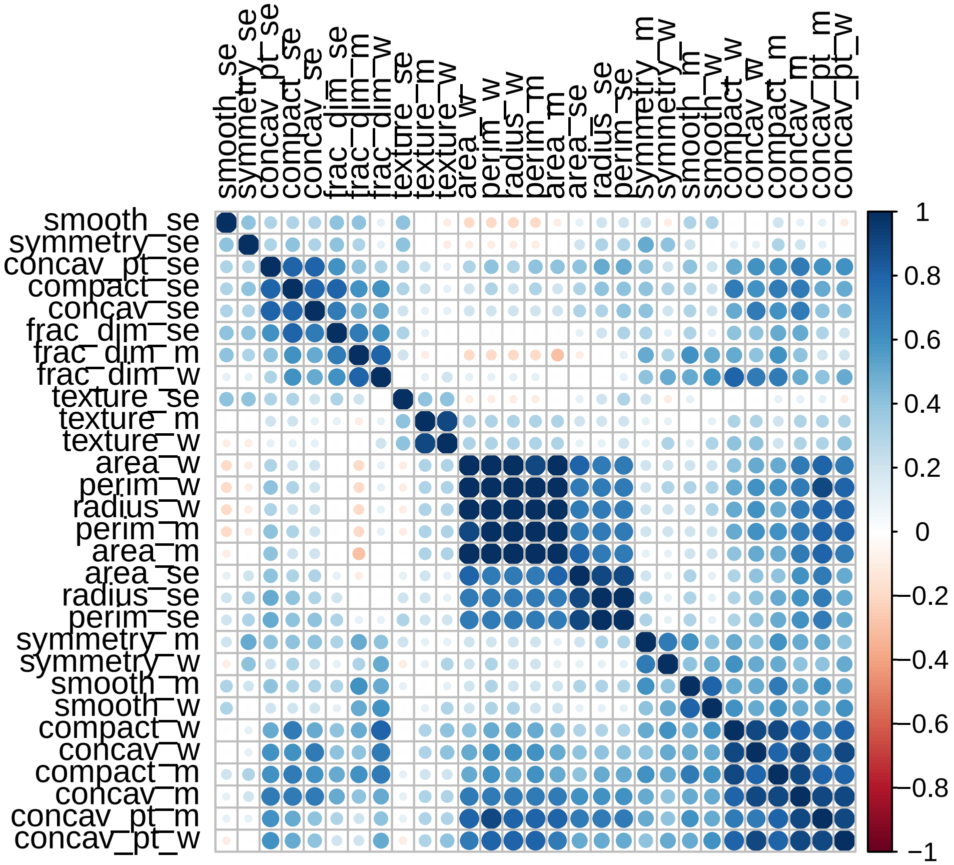
```r
data %<>%
  dplyr::select(-c(id, X33))
train %<>%
  dplyr::select(-c(id, X33))
test %<>%
  dplyr::select(-c(id, X33))
```

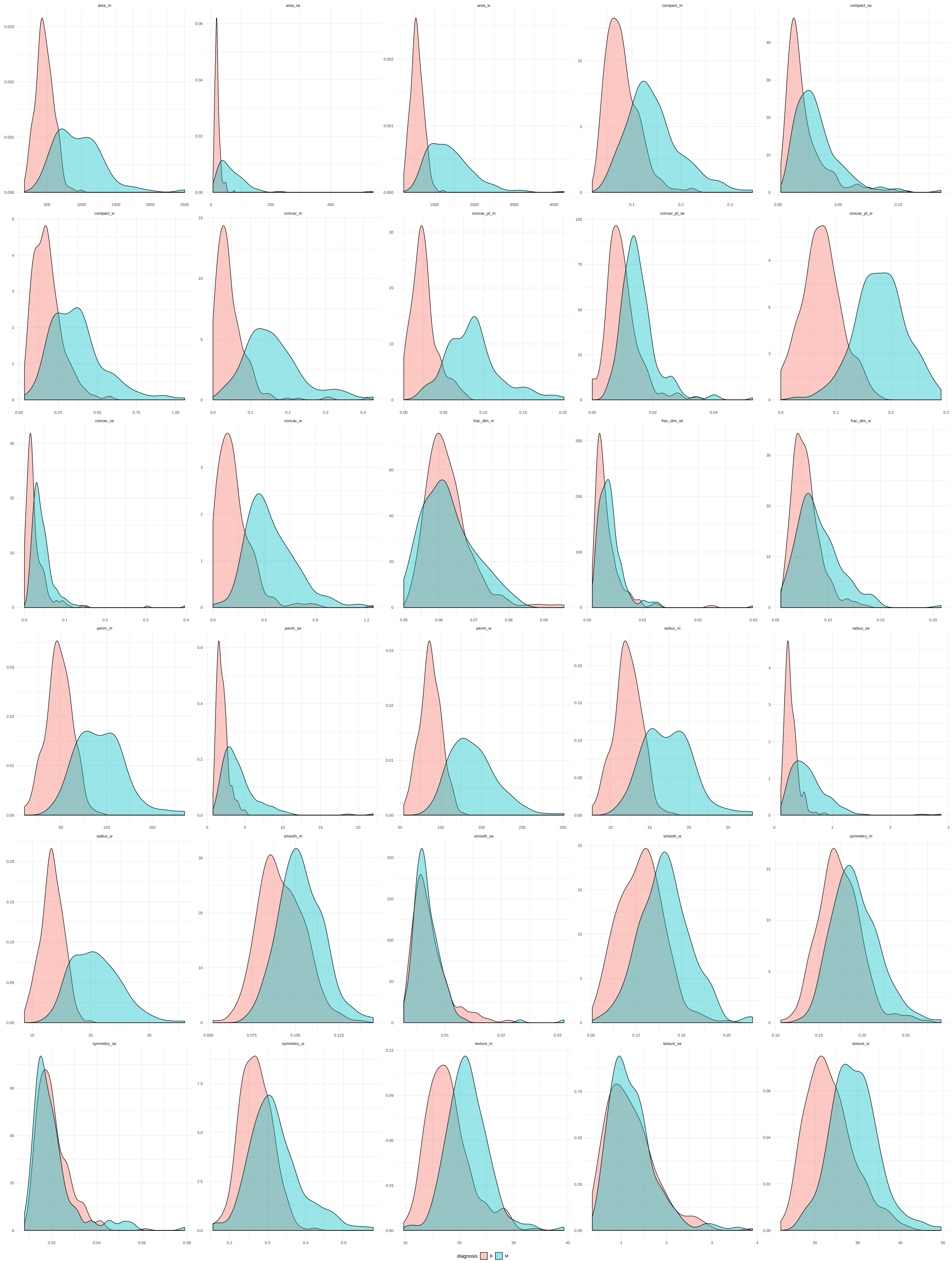
```r
sum(is.na(data))
```

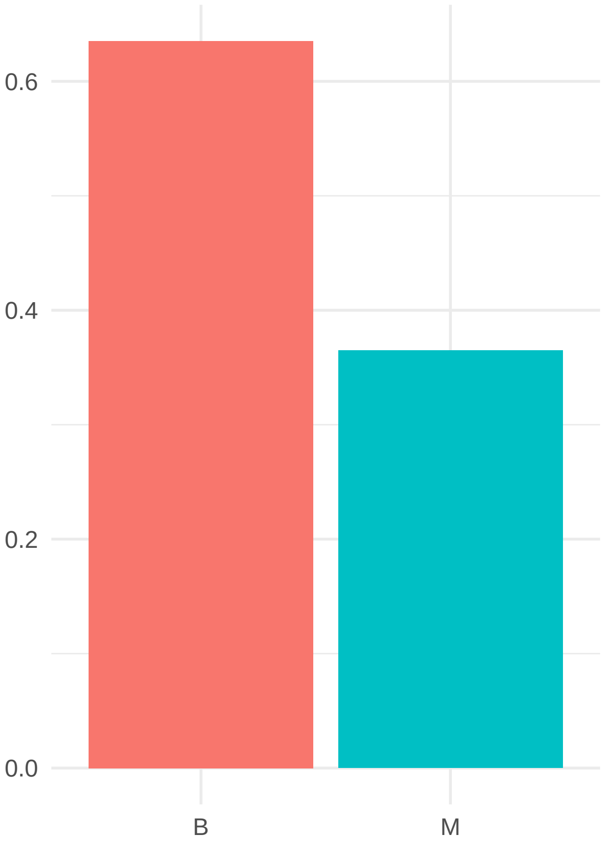
```
## [1] 0
```

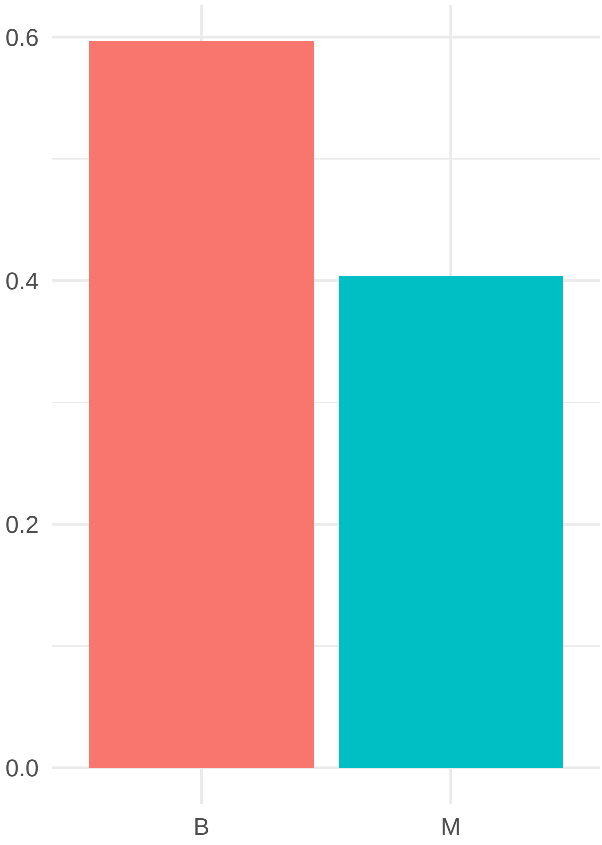
```
training_data <- train[2:dim(train)[2]]
training_classes <- train[1]

test_data <- test[2:dim(test)[2]]
test_classes <- test[1]
```




area_m
0.000
0.001
0.002
0.003
500
1000
1500
2000
2500
area_se
0.00
0.02
0.04
0.06
0
200
400
area_w
0.000
0.001
0.002
1000
2000
3000
4000
compact_m
0
5
10
0.1
0.2
0.3
compact_se
0
10
20
30
40
0.00
0.05
0.10
compact_w
0
1
2
3
4
5
0.00
0.25
0.50
0.75
1.00
concav_m
0
5
10
15
0.0
0.1
0.2
0.3
0.4
concav_pt_m
0
10
20
30
0.00
0.05
0.10
0.15
0.20
concav_pt_se
0
25
50
75
100
0.00
0.02
0.04
concav_pt_w
0
3
6
9
0.0
0.1
0.2
0.3
concav_se
0
10
20
30
0.0
0.1
0.2
0.3
0.4
concav_w
0
1
2
3
0.0
0.4
0.8
1.2
frac_dim_m
0
20
40
60
0.05
0.06
0.07
0.08
0.09
frac_dim_se
0
100
200
300
0.00
0.01
0.02
0.03
frac_dim_w
0
10
20
30
0.05
0.10
0.15
0.20
perim_m
0.00
0.01
0.02
0.03
80
120
160
perim_se
0.0
0.2
0.4
0.6
0
5
10
15
20
perim_w
0.00
0.01
0.02
0.03
50
100
150
200
250
radius_m
0.00
0.05
0.10
0.15
0.20
10
15
20
25
radius_se
0
1
2
3
4
0
1
2
3
radius_w
0.00
0.05
0.10
0.15
0.20
10
20
30
smooth_m
0
10
20
30
0.050
0.075
0.100
0.125
smooth_se
0
50
100
150
200
0.01
0.02
0.03
smooth_w
0
5
10
15
20
0.08
0.12
0.16
0.20
symmetry_m
0
5
10
15
0.10
0.15
0.20
0.25
symmetry_se
0
20
40
60
0.02
0.04
0.06
0.08
symmetry_w
0.0
2.5
5.0
7.5
0.2
0.3
0.4
0.5
texture_m
0.00
0.03
0.06
0.09
0.12
10
20
30
40
texture_se
0.00
0.25
0.50
0.75
1
2
3
4
texture_w
0.00
0.02
0.04
0.06
20
30
40
50
diagnosis
B
M


0.6
0.4
0.2
0.0
B
M


0.6
0.4
0.2
0.0
B
M