Glucose project

Van Nhut Ho

16/06/2020

Contents

1 EXECUTIVE SUMMARY The aim of this project is to show if the age and the weight influence on the glycemia (glucose). The Glycemia refers to the concentration of sugar or glucose in the blood. In many of the European countries blood glucose or sugar is also measured as millimol per decilitre (mmol/dl). These measurements are referred to as the SI units. If the blood glucose level above 6.1 mmol/l or 1.10 g/l (hyperglycemia) and less than 3.5 mmol/l or 0.54 g/l (hypoglycemia) When fasting blood glucose is above 7 mmol/l (1.26 g/l), the diagnosis of diabetes is made.

2 METHODS AND ANALYSIS 2.1 Data Analysis 2.1.1 Dataset exploration In this section, I'll present the methods and the analysis of the data. Before continuing the process, it's really important to discover the variables (names, types, counts, lenght etc.)

Getting descriptive statistics This dataset contains 10 variables and 284 observations.

```
## 'data.frame':    284 obs. of  10 variables:
##  $ id      : int  220 54 168 114 159 74 178 22 241 155 ...
##  $ chol    : int  163 173 200 219 215 170 222 228 165 271 ...
##  $ glyhb   : num  4.31 4.44 3.55 5.23 4.66 ...
##  $ location: Factor w/ 2 levels "Buckingham","Louisa": 1 1 1 1 2 2 1 1 2 1 ...
##  $ age     : int  29 76 40 76 78 41 51 24 22 55 ...
##  $ gender  : Factor w/ 2 levels "female","male": 1 1 1 2 2 1 1 1 1 1 ...
##  $ height  : num  157 155 157 163 165 ...
##  $ weight  : num  44.8 46.2 47.6 47.6 49.4 ...
##  $ waist   : int  30 31 26 29 33 29 28 33 28 30 ...
##  $ age_cat : Factor w/ 3 levels "20-39","40-59",..: 1 3 2 3 3 2 2 1 1 2 ...
```

Type of variable : - id "integer" = unique identification value per patient - chol "integer" = unique identification value of cholesterol per patient - glyhb "numeric" = unique identification value of glycemia per patient - location "factor" = specific city per patient - age "integer" = age of patient - gender "factor" = gender female or male - height "numeric" = unique identification value per patient in cm - weight "numeric" = unique identification value per patient in kg - waist "integer" = unique identification value per patient in cm - age_cat "factor" = age groups 20-39/40-59/60+ years

checking for missing variables This data set has no missing vaule.

```
##       id     chol    glyhb location      age   gender   height   weight
##        0        0        0        0        0        0        0        0
##    waist  age_cat
##        0        0
```

looking closer attributes and data values For all variables (except id, location, gender, age_cat because of nonsens), the mean and the median are close (= "egalitarian" distribution). We can see a large range for the variable chol and weight,.

```
##        id              chol           glyhb             location
##  Min.   :  1.00   Min.   : 78.0   Min.   :3.330   Buckingham:137
##  1st Qu.: 78.75   1st Qu.:174.8   1st Qu.:4.270   Louisa    :147
##  Median :157.50   Median :199.0   Median :4.660
##  Mean   :158.92   Mean   :200.5   Mean   :4.688
##  3rd Qu.:238.25   3rd Qu.:223.2   3rd Qu.:5.093
##  Max.   :318.00   Max.   :300.0   Max.   :6.210
##       age           gender        height          weight           waist
##  Min.   :20.00   female:171   Min.   :132.1   Min.   : 44.85   Min.   :26.0
##  1st Qu.:32.00   male  :113   1st Qu.:160.0   1st Qu.: 65.69   1st Qu.:33.0
##  Median :41.00                Median :167.6   Median : 77.01   Median :37.0
```

```
##  Mean   :43.82                Mean   :167.5   Mean   : 77.47   Mean   :36.8
##  3rd Qu.:54.25                3rd Qu.:175.3   3rd Qu.: 86.07   3rd Qu.:40.0
##  Max.   :79.00                Max.   :193.0   Max.   :116.42   Max.   :50.0
##   age_cat
##  20-39:120
##  40-59:109
##  60+  : 55
##
##
##
```

2.1.2 Univariate Plots Section One of the main goals of visualizing the data here is to observe which features are most helpful in predicting of type of glycemia (hyper-hypo glycemia). Now we look closer the attributes and data values for: **glyhb** Mean-Median are close ("egalitarian" distribution) Mean = 4.68 Median = 4.66 Skewness = 0.22 : asymmetry slightly to the right (-1 and +1) (+=to the right and 0=symmetry) coefficient of dyssimetry Kurtosis = -0.3 : flat curve, symmetry relative concentration of observations because < 0 ,flattening coefficient <0=platycurtic; >0=leptocurtic distribution therefore sharper curve=lower flattening Shape of the curve = Skewness+Kurtosis Conclusion: almost symmetric

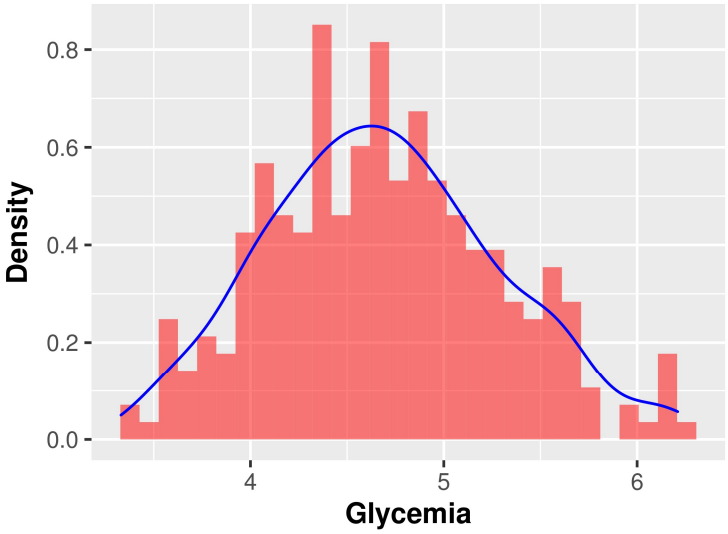
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.330   4.270   4.660   4.688   5.093   6.210
```

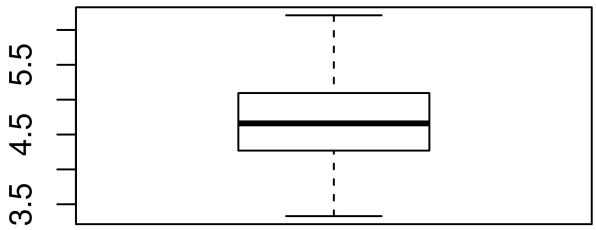
```
## [1] 0.2279742
## attr(,"method")
## [1] "moment"
```

```
## [1] -0.3428237
## attr(,"method")
## [1] "excess"
```

Histogram

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





Boxplot Outlier : none

weight

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   44.85   65.69   77.01   77.47   86.07  116.42
```

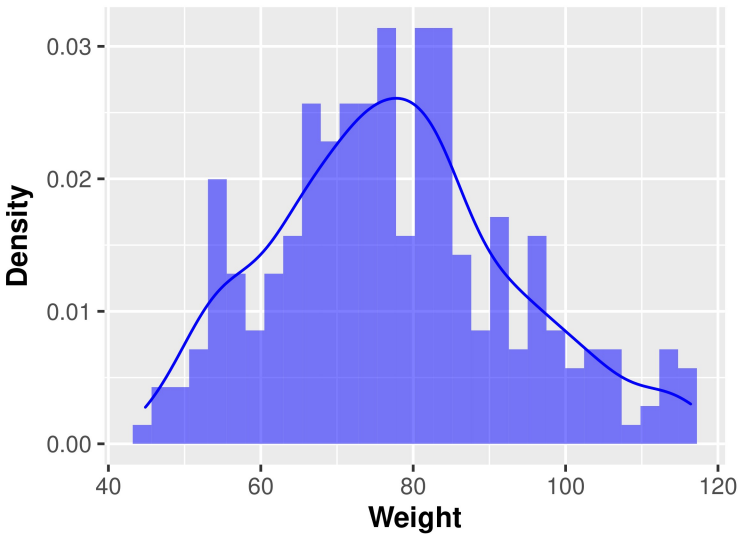
```
## [1] 0.3206615
## attr(,"method")
## [1] "moment"
```

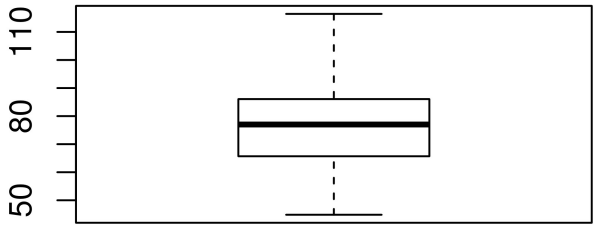
```
## [1] -0.2960106
## attr(,"method")
## [1] "excess"
```

Mean-Median are close (“egalitarian” distribution) Mean = 77.47 Median = 77.01 Skewness = 0.32 : asymmetry slightly to the right (-1 and +1) (+=to the right and 0=symmetry) coefficient of dyssimetry Kurtosis = -0.29 : flat curve, symmetry relative concentration of observations because < 0 ,flattening coefficient <0=platycurtic; >0=leptocurtic distribution therefore sharper curve=lower flattening Shape of the curve = Skewness+Kurtosis Conclusion: almost symmetric

Histogram

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





Boxplot Outlier : none

age

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   32.00   41.00   43.82   54.25   79.00
```

```
## [1] 0.382244
## attr(,"method")
## [1] "moment"
```

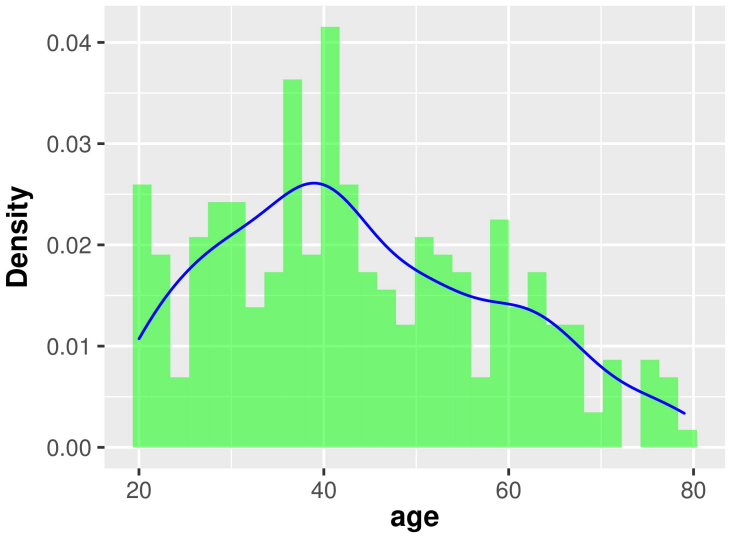
```
## [1] -0.7147143
## attr(,"method")
## [1] "excess"
```

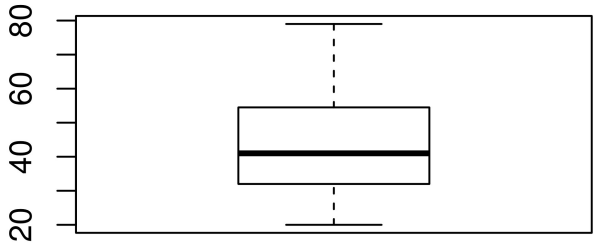
Mean-Median are close ("egalitarian" distribution) Mean = 41 Median = 43.82 Skewness = 0.38 : asymmetry slightly to the right (-1 and +1) (+=to the right and 0=symmetry) coefficient of dyssimetry Kurtosis = -0.71 : flat curve, symmetry relative concentration of observations because < 0 ,flattening coefficient

<0=platycurtic; >0=leptocurtic distribution therefore sharper curve=lower flattening Shape of the curve = Skewness+Kurtosis Conclusion: almost symmetric

Histogram

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





Boxplot Outlier : none

chol

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    78.0   174.8   199.0   200.5   223.2   300.0
```

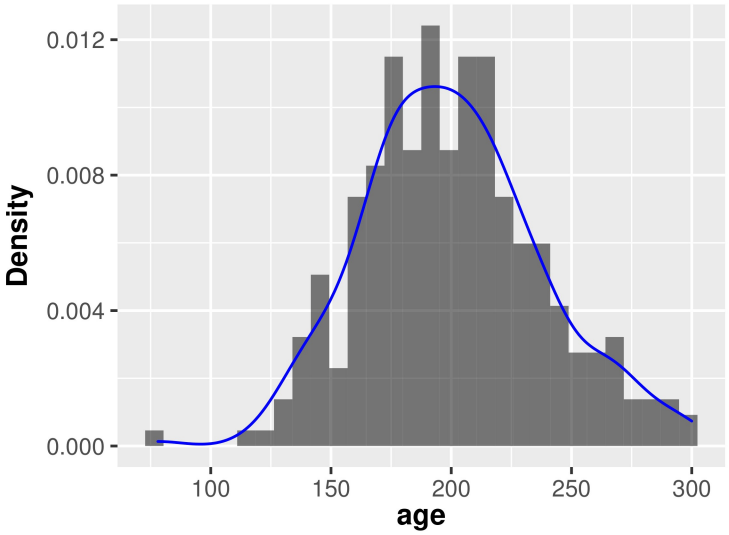
```
## [1] 0.2397935
## attr(,"method")
## [1] "moment"
```

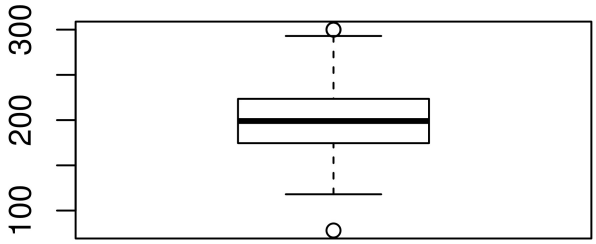
```
## [1] 0.07222744
## attr(,"method")
## [1] "excess"
```

Mean-Median are close (“egalitarian” distribution) Mean = 199.0 Median = 200.5 Skewness = 0.23 : asymmetry slightly to the right (-1 and +1) (+=to the right and 0=symmetry) coefficient of dyssimetry Kurtosis = 0.07 : weak flat curve, symmetry relative concentration of observations because = 0 ,flattening coefficient <0=platycurtic; >0=leptocurtic distribution therefore sharper curve=lower flattening Shape of the curve = Skewness+Kurtosis Conclusion: almost symmetric

Histogram

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





Boxplot Outlier : 1 Intlier : 1

2.1.3 Bivariate/Multivariate Analysis Correlation Plot with numeric variables: We are now interested in how the 6 predictors relate to each other. To see bivariate relationships among these predictors, we

calculate correlations between them. Correlations tell us: - whether this relationship is positive or negative - the strength of the relationship.

Value of r: Strength of relationship: -1.0 to -0.5 or 1.0 to 0.5 Strong -0.5 to -0.3 or 0.3 to 0.5 Moderate -0.3 to -0.1 or 0.1 to 0.3 Weak -0.1 to 0.1 None or very weak

Calculate collinearity Fellowing this corrplot, the variable glyhb has a bivariate relationships with the variable age, weight and waist. Note there is "no" relation between glyhb and height.

```
##      chol glyhb age height  weight waist
## 1     163  4.31  29 157.48  44.847    30
## 2     173  4.44  76 154.94  46.206    31
## 3     200  3.55  40 157.48  47.565    26
## 4     219  5.23  76 162.56  47.565    29
## 5     215  4.66  78 165.10  49.377    33
## 6     170  5.11  41 154.94  49.830    29
## 7     222  4.64  51 167.64  49.830    28
## 8     228  4.61  24 154.94  51.189    33
## 9     165  3.69  22 160.02  51.642    28
## 10    271  4.01  55 160.02  51.642    30
## 11    226  3.88  20 162.56  51.642    31
## 12    149  4.50  20 157.48  52.095    31
## 13    179  4.18  41 182.88  53.454    28
## 14    199  3.67  25 167.64  53.454    32
## 15    168  4.40  33 167.64  53.454    29
## 16    170  3.44  27 160.02  53.907    28
## 17    236  5.24  68 154.94  53.907    29
## 18     78  4.63  67 170.18  53.907    33
## 19    268  4.41  48 177.80  54.360    32
## 20    255  4.29  52 177.80  54.360    30
## 21    163  4.61  31 165.10  54.360    29
## 22    204  4.33  29 162.56  54.360    33
## 23    203  4.31  46 157.48  54.813    29
## 24    263  4.58  66 167.64  54.813    31
## 25    209  4.41  48 160.02  54.813    32
## 26    162  5.56  60 160.02  54.813    32
## 27    135  3.96  29 165.10  55.719    26
## 28    118  4.71  47 162.56  55.719    30
## 29    198  4.44  68 160.02  56.172    32
## 30    182  4.67  30 157.48  56.625    31
## 31    179  4.95  36 160.02  56.625    33
## 32    145  3.99  38 175.26  56.625    31
## 33    194  5.28  36 162.56  57.078    30
## 34    146  4.27  28 162.56  57.078    28
## 35    277  5.03  60 154.94  57.984    33
## 36    194  4.14  54 175.26  58.437    30
## 37    173  4.31  40 139.70  58.890    37
## 38    207  4.82  68 139.70  58.890    29
## 39    151  4.01  28 175.26  58.890    29
## 40    221  5.77  66 162.56  58.890    31
## 41    179  4.20  26 152.40  58.890    32
## 42    235  5.23  79 165.10  60.702    34
## 43    192  4.59  60 157.48  60.702    31
## 44    179  4.99  37 167.64  61.608    33
## 45    197  4.75  36 162.56  61.608    32
```

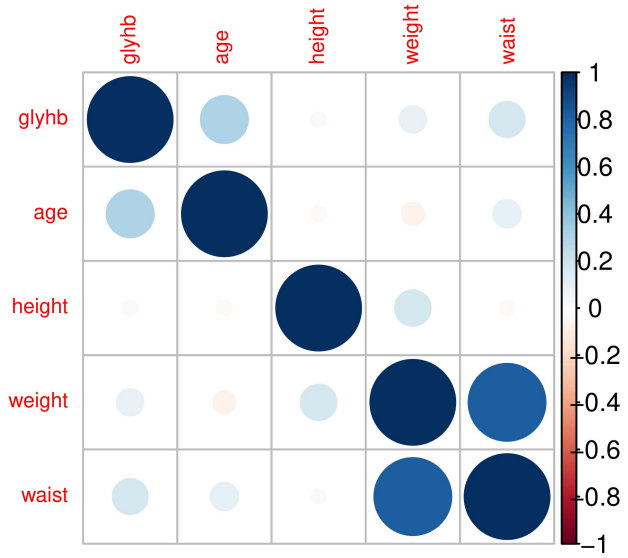
```
## 46   213  5.96  72 149.86  62.061    40
## 47   223  4.25  22 157.48  62.061    28
## 48   138  4.80  38 152.40  62.514    31
## 49   220  5.63  59 167.64  62.514    32
## 50   216  4.41  54 165.10  62.514    33
## 51   241  4.79  41 149.86  62.967    29
## 52   182  4.47  52 172.72  62.967    29
## 53   174  4.95  55 177.80  63.420    32
## 54   160  4.62  43 162.56  63.420    37
## 55   196  4.34  50 170.18  63.420    35
## 56   166  4.95  27 182.88  63.873    33
## 57   164  3.97  20 177.80  63.873    32
## 58   266  5.41  47 172.72  64.326    35
## 59   203  4.10  21 160.02  64.326    28
## 60   204  6.11  52 190.50  64.326    31
## 61   239  5.16  39 152.40  65.232    33
## 62   216  5.91  38 172.72  65.685    34
## 63   176  4.50  31 157.48  65.685    36
## 64   214  4.41  37 162.56  65.685    34
## 65   164  4.51  20 182.88  65.685    29
## 66   173  6.21  57 180.34  65.685    31
## 67   185  5.28  53 154.94  65.685    37
## 68   254  4.52  43 157.48  65.685    31
## 69   179  4.97  31 167.64  65.685    33
## 70   212  5.49  51 165.10  65.685    38
## 71   201  4.10  27 165.10  65.685    32
## 72   210  4.96  78 167.64  65.685    38
## 73   201  4.81  48 172.72  66.138    32
## 74   180  4.43  40 162.56  66.138    37
## 75   192  4.38  51 165.10  66.138    33
## 76   188  5.13  50 154.94  66.591    34
## 77   216  4.40  45 170.18  66.591    32
## 78   172  3.78  22 162.56  67.044    35
## 79   215  4.04  37 149.86  67.044    32
## 80   186  5.17  36 175.26  67.950    31
## 81   191  5.46  45 170.18  68.403    33
## 82   129  6.13  56 187.96  68.403    34
## 83   229  4.86  65 157.48  68.403    37
## 84   198  5.68  61 187.96  68.856    33
## 85   188  3.75  43 167.64  68.856    37
## 86   189  4.36  41 160.02  69.309    32
## 87   219  4.40  40 157.48  69.309    36
## 88   206  4.88  52 175.26  69.309    36
## 89   184  4.16  28 170.18  69.762    35
## 90   196  4.25  76 165.10  69.762    37
## 91   284  4.39  51 160.02  69.762    32
## 92   184  4.05  41 175.26  69.762    34
## 93   169  4.82  25 152.40  69.762    40
## 94   231  3.75  71 160.02  70.215    33
## 95   135  4.21  21 175.26  70.215    31
## 96   242  4.77  60 165.10  70.668    39
## 97   225  4.36  41 180.34  70.668    31
## 98   191  5.63  42 154.94  70.668    36
## 99   213  3.41  33 165.10  71.121    37
```

```
## 100  202  4.17  44 172.72  71.121    33
## 101  187  4.40  21 160.02  71.574    39
## 102  143  4.81  68 170.18  71.574    37
## 103  171  4.04  52 180.34  72.027    33
## 104  260  5.34  44 157.48  72.027    36
## 105  230  4.53  20 170.18  72.027    31
## 106  236  5.63  62 193.04  72.480    35
## 107  136  4.58  22 167.64  72.480    35
## 108  212  5.22  37 162.56  72.480    37
## 109  269  5.37  41 157.48  72.480    39
## 110  168  5.09  44 162.56  72.480    40
## 111  224  5.05  78 160.02  72.480    36
## 112  209  4.85  31 170.18  72.480    30
## 113  300  4.56  34 167.64  72.480    40
## 114  212  4.61  63 177.80  72.933    37
## 115  178  5.23  36 177.80  72.933    34
## 116  170  4.39  20 162.56  72.933    37
## 117  231  4.90  33 175.26  73.839    35
## 118  190  5.55  43 157.48  73.839    40
## 119  198  4.43  60 177.80  73.839    36
## 120  244  4.54  21 180.34  73.839    34
## 121  171  5.10  34 160.02  74.292    34
## 122  185  4.83  23 193.04  74.292    32
## 123  138  4.70  57 185.42  74.292    31
## 124  148  6.14  54 170.18  74.745    42
## 125  172  4.52  42 165.10  74.745    33
## 126  183  4.59  40 149.86  74.745    37
## 127  224  4.92  34 152.40  74.745    34
## 128  144  4.13  30 182.88  74.745    31
## 129  145  4.73  30 165.10  74.745    33
## 130  157  5.70  63 175.26  75.198    39
## 131  177  4.84  45 175.26  75.198    34
## 132  202  5.50  64 157.48  75.651    44
## 133  204  4.69  72 165.10  75.651    45
## 134  206  4.07  38 175.26  75.651    36
## 135  233  4.56  45 162.56  75.651    39
## 136  244  4.36  44 180.34  76.104    36
## 137  159  5.02  38 172.72  76.557    34
## 138  218  4.87  35 175.26  76.557    39
## 139  132  4.01  21 165.10  76.557    39
## 140  300  5.18  61 170.18  76.557    40
## 141  180  3.59  63 175.26  76.557    35
## 142  238  4.47  27 152.40  77.010    35
## 143  179  3.33  29 172.72  77.010    38
## 144  239  4.69  35 187.96  77.010    32
## 145  228  4.11  54 167.64  77.010    36
## 146  179  3.98  34 182.88  77.010    31
## 147  293  4.87  50 180.34  77.010    34
## 148  172  3.59  48 160.02  77.010    35
## 149  262  4.90  33 160.02  77.010    33
## 150  179  4.68  20 147.32  77.010    34
## 151  218  3.89  52 157.48  77.010    40
## 152  227  3.94  37 149.86  77.010    34
## 153  199  4.96  71 175.26  77.463    38
```

```
## 154  195  5.68  78 167.64  77.916    40
## 155  134  4.29  48 177.80  78.369    36
## 156  182  5.66  61 175.26  78.822    49
## 157  155  4.17  26 185.42  78.822    30
## 158  273  3.76  53 162.56  78.822    34
## 159  149  4.09  26 157.48  78.822    38
## 160  177  5.11  42 165.10  78.822    37
## 161  240  5.74  54 165.10  79.275    37
## 162  194  4.61  63 185.42  79.275    34
## 163  215  5.03  59 160.02  79.728    34
## 164  221  5.53  59 157.48  80.181    39
## 165  206  4.82  67 170.18  80.634    37
## 166  219  3.75  53 162.56  81.087    39
## 167  293  4.76  63 162.56  81.087    47
## 168  150  3.97  35 185.42  81.087    32
## 169  197  4.95  43 180.34  81.087    37
## 170  207  5.01  46 160.02  81.087    38
## 171  211  3.55  40 172.72  81.087    37
## 172  241  5.04  27 160.02  81.087    40
## 173  260  4.78  69 149.86  81.087    45
## 174  179  4.05  32 157.48  81.087    37
## 175  181  4.88  29 172.72  81.540    38
## 176  204  5.20  62 172.72  81.540    38
## 177  164  3.80  28 170.18  81.540    39
## 178  158  4.31  50 180.34  81.540    36
## 179  205  4.87  72 154.94  81.540    39
## 180  229  4.73  23 182.88  81.540    34
## 181  169  6.14  40 165.10  81.540    40
## 182  207  5.06  30 182.88  81.540    35
## 183  237  5.35  43 162.56  81.993    36
## 184  194  4.26  63 177.80  81.993    37
## 185  268  5.36  38 160.02  81.993    38
## 186  225  4.66  53 160.02  82.446    38
## 187  179  4.75  23 165.10  82.899    43
## 188  242  5.47  46 157.48  82.899    37
## 189  255  4.33  50 165.10  82.899    37
## 190  234  3.70  41 170.18  82.899    38
## 191  208  5.60  56 172.72  82.899    36
## 192  122  3.98  36 180.34  82.899    41
## 193  191  4.67  36 175.26  82.899    36
## 194  206  4.03  41 157.48  83.352    39
## 195  160  4.64  36 162.56  83.805    39
## 196  270  3.58  42 167.64  83.805    39
## 197  184  4.71  66 187.96  83.805    40
## 198  281  5.56  66 157.48  83.805    48
## 199  147  4.67  23 154.94  83.805    43
## 200  204  4.84  27 170.18  83.805    35
## 201  193  4.74  42 190.50  84.258    37
## 202  217  4.07  33 157.48  84.258    42
## 203  183  4.03  47 167.64  84.258    39
## 204  235  4.90  60 175.26  84.258    40
## 205  152  4.27  40 132.08  84.711    38
## 206  204  4.44  59 185.42  84.711    38
## 207  174  5.53  20 177.80  84.711    37
```

```
## 208  217  3.66  20 170.18  84.711    40
## 209  178  4.10  41 165.10  85.164    35
## 210  159  4.18  76 167.64  85.164    40
## 211  215  4.37  40 177.80  85.617    37
## 212  189  3.62  45 175.26  86.070    39
## 213  183  4.37  31 167.64  86.070    41
## 214  214  5.38  44 177.80  86.070    38
## 215  248  4.81  34 180.34  86.070    36
## 216  269  5.14  41 170.18  86.523    38
## 217  195  4.84  30 175.26  86.523    46
## 218  225  4.38  36 170.18  86.976    40
## 219  192  4.04  37 180.34  88.335    36
## 220  215  4.97  38 147.32  88.335    42
## 221  204  5.02  48 172.72  88.788    38
## 222  199  4.49  41 160.02  89.241    41
## 223  261  5.12  52 162.56  89.694    42
## 224  174  5.07  67 172.72  89.694    36
## 225  206  5.49  61 160.02  90.147    41
## 226  218  5.52  40 185.42  90.600    38
## 227  249  5.12  51 165.10  90.600    43
## 228  189  4.38  28 162.56  90.600    38
## 229  168  4.17  28 160.02  90.600    42
## 230  242  3.97  70 167.64  90.600    41
## 231  293  5.17  31 170.18  90.600    41
## 232  204  4.16  55 167.64  91.506    43
## 233  263  5.78  55 160.02  91.506    45
## 234  199  5.44  37 154.94  91.959    42
## 235  205  4.21  75 175.26  92.412    44
## 236  214  3.89  28 172.72  92.412    40
## 237  147  4.62  38 175.26  92.865    39
## 238  190  3.56  46 182.88  92.865    46
## 239  224  5.26  36 175.26  92.865    37
## 240  189  4.86  49 157.48  92.865    40
## 241  203  4.67  27 170.18  94.677    34
## 242  173  4.40  43 175.26  95.130    44
## 243  188  6.17  66 172.72  95.130    45
## 244  194  4.97  63 147.32  95.130    44
## 245  174  5.35  34 180.34  95.130    37
## 246  190  4.66  27 165.10  95.130    39
## 247  227  4.98  58 177.80  95.583    38
## 248  156  4.55  37 170.18  96.036    48
## 249  244  4.66  32 177.80  96.036    39
## 250  171  4.59  40 180.34  96.942    41
## 251  201  5.35  58 167.64  97.395    46
## 252  158  5.56  50 177.80  97.395    40
## 253  162  4.40  43 170.18  97.848    41
## 254  165  4.44  29 162.56  98.754    46
## 255  217  4.84  34 185.42  99.207    41
## 256  157  5.57  55 167.64  99.207    43
## 257  143  5.15  61 165.10  99.660    40
## 258  193  5.01  21 154.94  99.660    40
## 259  214  4.48  40 182.88 100.566    40
## 260  217  3.93  22 180.34 101.019    46
## 261  277  5.24  63 162.56 101.019    45
```

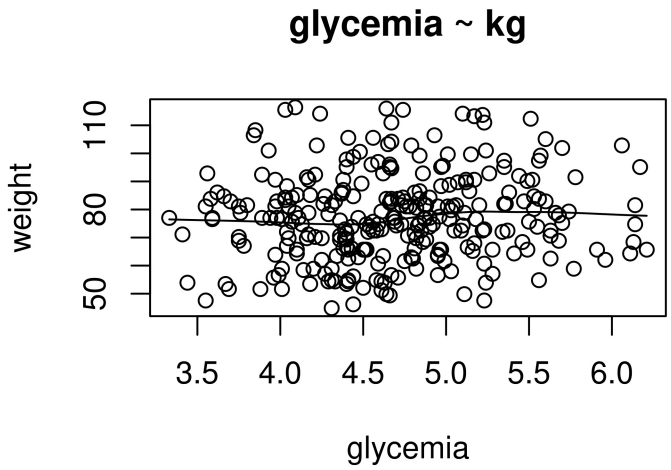
```
## 262  132  5.70  28 172.72 101.925    41
## 263  283  4.22  26 182.88 102.831    41
## 264  188  4.79  31 170.18 102.831    47
## 265  255  6.06  64 172.72 102.831    44
## 266  185  4.65  50 162.56 103.284    42
## 267  234  4.67  47 170.18 104.190    45
## 268  223  5.60  47 165.10 105.096    46
## 269  219  4.56  65 160.02 105.549    40
## 270  243  4.41  37 162.56 105.549    49
## 271  175  3.84  23 165.10 106.455    44
## 272  199  4.93  42 170.18 106.455    47
## 273  243  3.85  43 162.56 108.267    48
## 274  164  5.23  23 175.26 110.985    44
## 275  134  4.67  25 160.02 110.985    47
## 276  251  5.51  38 162.56 112.344    49
## 277  192  5.17  30 182.88 113.250    43
## 278  169  5.22  62 167.64 113.703    50
## 279  232  5.10  37 172.72 114.156    43
## 280  176  4.24  32 160.02 114.156    45
## 281  199  4.74  36 167.64 115.515    47
## 282  181  4.03  39 167.64 115.515    46
## 283  228  4.64  58 154.94 115.968    49
## 284  181  4.09  30 167.64 116.421    47
```



Linear regression Now let's have a quantitative score throught a scatter plots and calculation that can help visualize any linear relationships between the dependent (response) variable and independent (predictor) variables.

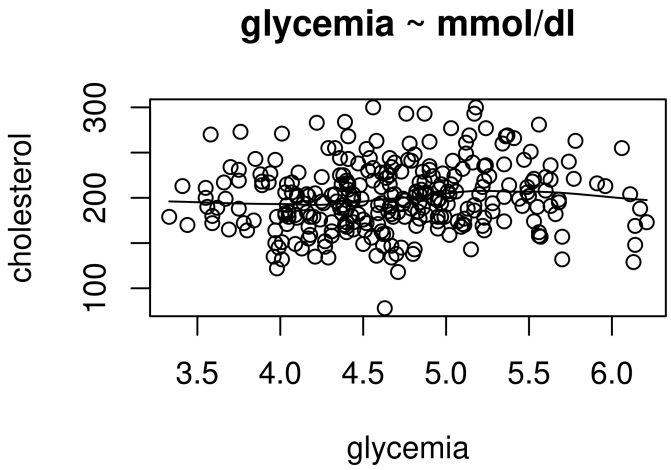
Linear regression between glycemia and weight: Following this score and the scatter plot, there is a weak positive linear relationships between these variables.

```
## [1] 0.09843643
```



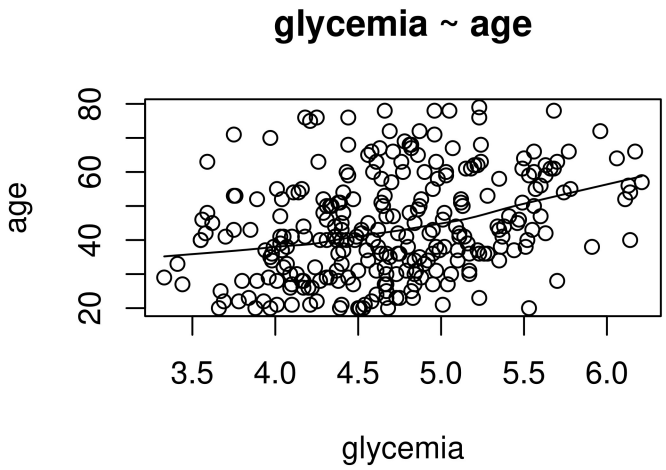
Linear regression between glycemia and cholesterol: As we can see, there is a weak positive linear regression.

```
## [1] 0.1131798
```



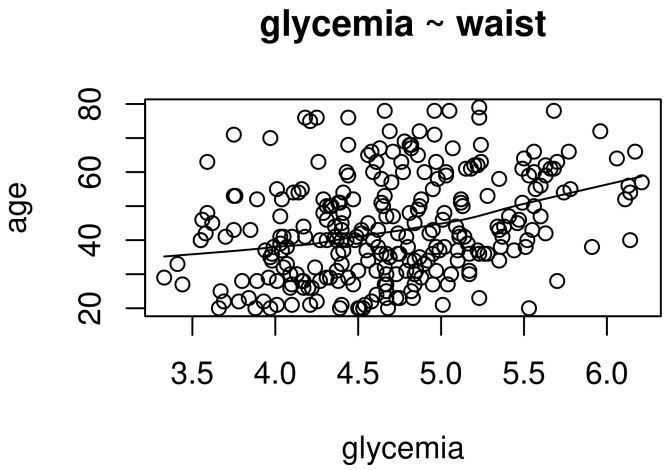
Linear regression between glyemia and age: Here we see a slight positive correlation between these variables.

```
## [1] 0.307663
```



Linear regression between glyemia and wast: The score and the scatter plot show us a slight positive correlation between these variables.

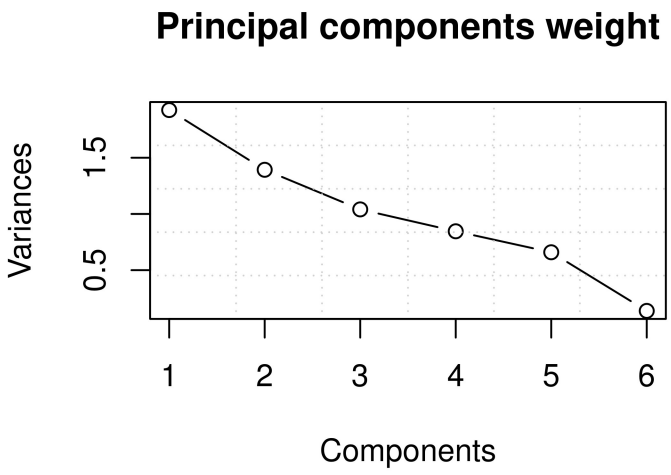
```
## [1] 0.1715594
```



3 RESULTS*

3.1 Principal Components Analysis (PCA) transform Principal component analysis is a statistical technique that is used to analyze the interrelationships among a large number of variables and to explain these variables in terms of a smaller number of variables (called principal components) with a minimum loss of information. It constructs a set of orthogonal (non-collinear, uncorrelated, independent) variables and is used for making predictive models

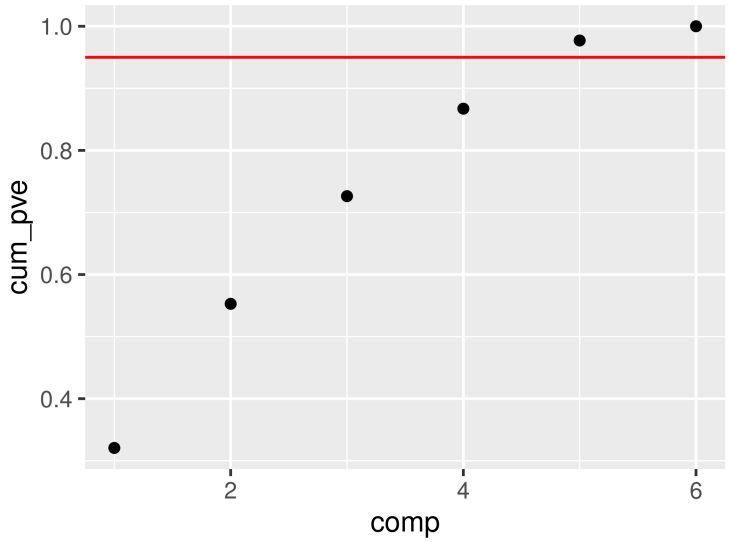
This plot show us the increase of components(variables) the decrease of variances:



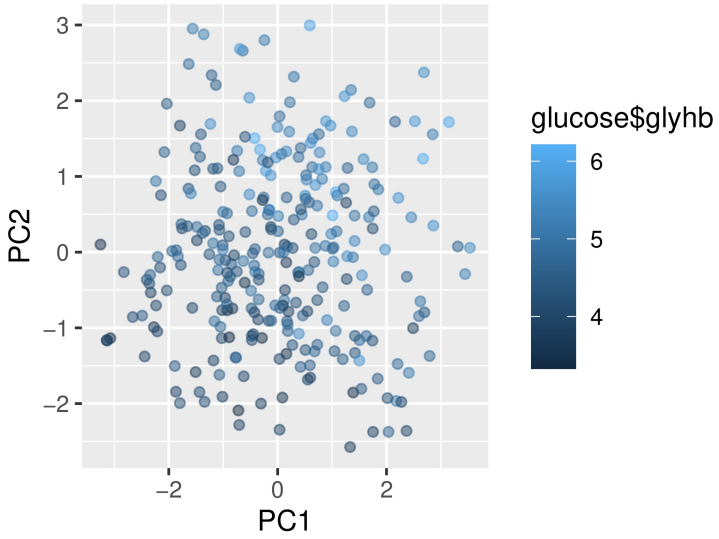

```
## Importance of components:
##                            PC1    PC2    PC3    PC4    PC5     PC6
## Standard deviation     1.3871 1.1802 1.0201 0.9196 0.8118 0.37086
## Proportion of Variance 0.3207 0.2322 0.1734 0.1410 0.1098 0.02292
## Cumulative Proportion  0.3207 0.5528 0.7263 0.8672 0.9771 1.00000
```

Following these results, the standard deviation of variables are low and indicate that the values tend to be close to the mean (also called the expected value) of the set.

Proportion of variance explained To explain more than 0.95 of the variance, it's required 5 principal components and 5 for 0.99.



The features with highest dimmensions or aligned with the leading principal component are the ones with



highest variance.

3.2 Machine learning Machine learning algorithms build a mathematical model based on sample data (training data), in order to make predictions or decisions making without being explicitly programmed to do so. **Split data into train and test sets** The division of the dataset into two parts makes it possible to check the performance of the learning machine. I will Split the available data into a train set (65%) and a test set (35).

train_set = 187 rows test_set = 97 rows

```
## [1] 187
```

```
## [1] 97
```

Overview train_set and test_set

	id	chol	glyhb	location	age	gender	height	weight	waist	age_cat
1	220	163	4.31	Buckingham	29	female	157.48	44.847	30	20-39
3	168	200	3.55	Buckingham	40	female	157.48	47.565	26	40-59
6	74	170	5.11	Louisa	41	female	154.94	49.830	29	40-59
7	178	222	4.64	Buckingham	51	female	167.64	49.830	28	40-59
8	22	228	4.61	Buckingham	24	female	154.94	51.189	33	20-39
9	241	165	3.69	Louisa	22	female	160.02	51.642	28	20-39

	id	chol	glyhb	location	age	gender	height	weight	waist	age_cat
2	54	173	4.44	Buckingham	76	female	154.94	46.206	31	60+
4	114	219	5.23	Buckingham	76	male	162.56	47.565	29	60+
5	159	215	4.66	Louisa	78	male	165.10	49.377	33	60+
11	245	226	3.88	Louisa	20	female	162.56	51.642	31	20-39
13	23	179	4.18	Buckingham	41	female	182.88	53.454	28	40-59
15	128	168	4.40	Buckingham	33	female	167.64	53.454	29	20-39

3.2.1 Naive Bayes In machine learning, naïve Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong independence assumptions between the features

Average of all glycemia

```
## [1] 4.687746
```

Predict the RMSE on the validation set

Finally a dataframe of result

model	RMSE
Naive Mean-Baseline Model	0.5481574

3.2.2 Model building on training data to predict the glycemia on test data From the model summary, the model p value and predictor's p value are less than the significance level = statistically significant model.

Calculate Akaike Information Criterion (AIC) The AIC is essentially an estimated measure of the quality of each of the available econometric models as they relate to one another for a certain set of data, making it an ideal method for model selection.

AIC between glyhb and weight:

```
##
## Call:
## lm(formula = glyhb ~ weight, data = glucose)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35602 -0.39505 -0.02877  0.37948  1.56648
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.397030   0.178551  24.626   <2e-16 ***
## weight      0.003753   0.002259   1.661   0.0978 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5959 on 282 degrees of freedom
## Multiple R-squared:  0.00969,    Adjusted R-squared:  0.006178
## F-statistic: 2.759 on 1 and 282 DF,  p-value: 0.09781
```

```
##
## Call:
## lm(formula = glyhb ~ age, data = glucose)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33254 -0.39033 -0.04089  0.36364  1.49902
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.151304   0.104419   39.76  < 2e-16 ***
## age         0.012242   0.002255    5.43 1.22e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5698 on 282 degrees of freedom
## Multiple R-squared:  0.09466,    Adjusted R-squared:  0.09145
## F-statistic: 29.48 on 1 and 282 DF,  p-value: 1.218e-07
```

```
##        df      AIC
## lmMod   3 515.8976
## lmMod2  3 490.4217
```

Calculate Bayesian Information Criterion (BIC) The BIC is a variant of AIC with a stronger penalty for including additional variables to the model. BIC between glyhb and weight:

```
##        df      BIC
## lmMod   3 526.8445
## lmMod2  3 501.3686
```

AIC and BIC are small meaning that the econometric models is powerful.

Average prediction error rate Dividing the RSE by the average value of the outcome variable will give us the prediction error rate, which should be as small as possible:

```
## [1] 0.1271167
```

The average prediction error rate is 12.7%.

Calculate prediction accuracy and error rates A simple correlation between the actuals and predicted values can be used as a form of accuracy measure. A higher correlation accuracy implies that the actuals and predicted values have similar directional movement, i.e. when the actuals values increase the predicteds also increase and vice-versa.

```
##    actuals predicteds
## 2     4.44   4.570426
```

```
## 4     5.23   4.575526
## 5     4.66   4.582325
## 11    3.88   4.590825
## 13    4.18   4.597625
## 15    4.40   4.597625
```

min_max_accuracy The result of min max accuracy is hight:

```
## [1] 0.9103795
```

mean absolute percentage deviation The mean absolute percentage deviation is low:

```
## [1] 0.09756926
```

3.2.3 Cross-validation methods

The Validation set Approach When comparing two models, the one that produces the lowest test sample RMSE is the preferred model. Making predictions and computing the R2, RMSE and MAE:

```
##           R2      RMSE       MAE
## 1 0.01978701 0.5791717 0.4689291
```

The RMSE and the MAE are measured in the same scale as the outcome variable. Dividing the RMSE by the average value of the outcome variable will give us the prediction error rate, which should be as small as possible:

```
## [1] 0.1238938
```

Note that, the validation set method is only useful when you have a large data set that can be partitioned. Therefore, the test error rate can be highly variable, depending on which observations are included in the training set and which observations are included in the validation set.

K-fold cross-validation The k-fold cross-validation method evaluates the model performance on different subset of the training data and then calculate the average prediction error rate. The algorithm is as follow: - Randomly split the data set into k-subsets (or k-fold) (for example 5 or 10 subsets) - Reserve one subset and train the model on all other subsets - Test the model on the reserved subset and record the prediction error - Repeat this process until each of the k subsets has served as the test set. - Compute the average of the k recorded errors (cross-validation error) serving as the performance metric for the model. - K-fold cross-validation (CV) is a robust method for estimating the accuracy of a model.

Define training control and train the model

```
## Error in train(glyhb ~ ., data = glucose, method = "lm", trControl = train.control): arguments inuti
```

Summarize the results

```
##
## Call:
## lm(formula = glyhb ~ ., data = train_set)
##
## Coefficients:
##    (Intercept)             id           chol  locationLouisa            age
```

```
##      2.2745534      -0.0008731      -0.0001066       0.0932791       0.0222294
##     gendermale          height          weight           waist    age_cat40-59
##     -0.2044325       0.0077879       0.0003019       0.0118329      -0.2077641
##     age_cat60+
##     -0.2373397
```

Repeated K-fold cross-validation The process of splitting the data into k-folds could be repeatedas many times as wished and called repeated k-fold cross validation. The final model error is taken as the mean error from the number of repeats.

This following example uses 10-fold cross validation with 4 repeats: Define training control and model

```
## Error in train(glyhb ~ ., data = glucose, method = "lm", trControl = train.control2): arguments inut
```

Summarize the results

```
## Error in print(model2): objet 'model2' introuvable
```

Summarize RMSE table This table shows the RMSE results of models builted and trained. As we can see, applying K-fold cross-validation model reduce the RMSE score.

model	RMSE
Naive Mean-Baseline Model	0.5481574
K-fold cross-validation	0.5791717

Conclusion - The glycemia has a slight positive correlation with waist (0.17) and with age (0.30). - The performant model metrics are powerfull (low AIC,BIC). - The Naive Mean-Baseline Model has a low RMSE (prediction errors) score. - The K-fold Cross-Validation has a low RMSE score.

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 18362)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=French_Switzerland.1252  LC_CTYPE=French_Switzerland.1252
## [3] LC_MONETARY=French_Switzerland.1252 LC_NUMERIC=C
## [5] LC_TIME=French_Switzerland.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] mlr_2.17.1        ParamHelpers_1.14 corrplot_0.84     timeDate_3043.102
##  [5] kableExtra_1.1.0  data.table_1.12.8 caret_6.0-86      lattice_0.20-41
##  [9] forcats_0.5.0     stringr_1.4.0     dplyr_1.0.0       purrr_0.3.4
## [13] readr_1.3.1       tidyr_1.1.0       tibble_3.0.1      ggplot2_3.3.1
## [17] tidyverse_1.3.0   foreign_0.8-72
##
## loaded via a namespace (and not attached):
##  [1] nlme_3.1-148         fs_1.4.1             lubridate_1.7.9
##  [4] webshot_0.5.2        httr_1.4.1           tools_3.6.2
##  [7] backports_1.1.7      R6_2.4.1             rpart_4.1-15
## [10] DBI_1.1.0            colorspace_1.4-1     nnet_7.3-14
## [13] withr_2.2.0          tidyselect_1.1.0     compiler_3.6.2
## [16] parallelMap_1.5.0    cli_2.0.2            rvest_0.3.5
## [19] xml2_1.3.2           labeling_0.3         scales_1.1.1
## [22] checkmate_2.0.0      digest_0.6.25        rmarkdown_2.2
## [25] pkgconfig_2.0.3      htmltools_0.4.0      dbplyr_1.4.4
## [28] rlang_0.4.6          readxl_1.3.1         rstudioapi_0.11
## [31] BBmisc_1.11          farver_2.0.3         generics_0.0.2
## [34] jsonlite_1.6.1       ModelMetrics_1.2.2.2 magrittr_1.5
## [37] Matrix_1.2-18        Rcpp_1.0.4.6         munsell_0.5.0
## [40] fansi_0.4.1          lifecycle_0.2.0      stringi_1.4.6
## [43] pROC_1.16.2          yaml_2.2.1           MASS_7.3-51.6
## [46] plyr_1.8.6           recipes_0.1.12       grid_3.6.2
## [49] blob_1.2.1           parallel_3.6.2       crayon_1.3.4
## [52] haven_2.3.1          splines_3.6.2        hms_0.5.3
## [55] knitr_1.28           pillar_1.4.4         reshape2_1.4.4
## [58] codetools_0.2-16     stats4_3.6.2         fastmatch_1.1-0
## [61] reprex_0.3.0         glue_1.4.1           evaluate_0.14
## [64] modelr_0.1.8         vctrs_0.3.1          foreach_1.5.0
## [67] cellranger_1.1.0     gtable_0.3.0         assertthat_0.2.1
## [70] xfun_0.14            gower_0.2.1          prodlim_2019.11.13
## [73] broom_0.5.6          class_7.3-17         survival_3.1-12
## [76] viridisLite_0.3.0    iterators_1.0.12     lava_1.6.7
## [79] ellipsis_0.3.1       ipred_0.9-9
```