ST332 Project 2

Group a

2/9/2021

Intro

Data Collection

This report uses data collected in the CAST study (detailed by Lamb et.al., 2009) on 585 patients with ankle sprains. The study analyses different treatments comparatively, with the aim of making clinical recommendations. Encompassing 238 females and 347 males of ages 16-72, the dataset contains variables describing the age, weight, sex and height of each participant, as well as five health scores, which are components of the Foot and Ankle Outcome Score in the initial paper.

```r
summary(cast)
```

```
##       age             sex            height          weight      
##  Min.   :16.00   Min.   :1.000   Min.   :147.3   Min.   : 39.92  
##  1st Qu.:21.00   1st Qu.:1.000   1st Qu.:165.1   1st Qu.: 67.13  
##  Median :28.00   Median :2.000   Median :172.7   Median : 76.20  
##  Mean   :30.03   Mean   :1.582   Mean   :172.9   Mean   : 78.58  
##  3rd Qu.:37.00   3rd Qu.:2.000   3rd Qu.:180.3   3rd Qu.: 87.09  
##  Max.   :72.00   Max.   :2.000   Max.   :200.7   Max.   :133.36  
##                                                                  
##      bsymp           bpain            badl            bsport      
##  Min.   : 0.00   Min.   : 0.00   Min.   :  0.00   Min.   :  0.00  
##  1st Qu.:28.57   1st Qu.:27.78   1st Qu.: 50.00   1st Qu.:  0.00  
##  Median :39.29   Median :38.89   Median : 58.82   Median : 10.00  
##  Mean   :39.16   Mean   :38.24   Mean   : 57.61   Mean   : 13.92  
##  3rd Qu.:50.00   3rd Qu.:50.00   3rd Qu.: 66.18   3rd Qu.: 20.00  
##  Max.   :96.43   Max.   :94.44   Max.   :100.00   Max.   :100.00  
##  NA's   :1       NA's   :1       NA's   :5        NA's   :12      
##      bqual            symp9            pain9             adl9       
##  Min.   :  0.00   Min.   : 14.29   Min.   : 13.89   Min.   : 39.71  
##  1st Qu.:  6.25   1st Qu.: 71.43   1st Qu.: 75.00   1st Qu.: 95.59  
##  Median : 18.75   Median : 85.71   Median : 91.67   Median :100.00  
##  Mean   : 22.82   Mean   : 81.20   Mean   : 84.15   Mean   : 94.22  
##  3rd Qu.: 31.25   3rd Qu.: 96.43   3rd Qu.:100.00   3rd Qu.:100.00  
##  Max.   :100.00   Max.   :100.00   Max.   :102.78   Max.   :100.00  
##  NA's   :2        NA's   :135      NA's   :135      NA's   :214     
##      sport9           qual9            Yscore     
##  Min.   :  0.00   Min.   :  0.00   Min.   :0.000  
##  1st Qu.: 70.00   1st Qu.: 50.00   1st Qu.:2.040  
##  Median : 85.00   Median : 75.00   Median :4.090  
##  Mean   : 79.23   Mean   : 70.89   Mean   :3.913  
##  3rd Qu.:100.00   3rd Qu.: 93.75   3rd Qu.:5.718  
##  Max.   :100.00   Max.   :100.00   Max.   :9.460  
##  NA's   :218      NA's   :135      NA's   :135    
```

Data entry check

The dataset used in the subsequent analysis is imported from a medical trial. We want to find out if the data entry process is robust by checking for variable type (coding) and looking at a summary table of the data.

```
ff_glimpse(cast)
```

```
## $Continuous
##         label var_type   n missing_n missing_percent  mean   sd   min
## age       age    <int> 565         0             0.0  30.0 10.8  16.0
## sex       sex    <int> 565         0             0.0   1.6  0.5   1.0
## height height    <dbl> 565         0             0.0 172.9  9.8 147.3
## weight weight    <dbl> 565         0             0.0  78.6 15.5  39.9
## bsymp   bsymp    <dbl> 564         1             0.2  39.2 16.5   0.0
## bpain   bpain    <dbl> 564         1             0.2  38.2 16.4   0.0
## badl     badl    <dbl> 560         5             0.9  57.6 14.0   0.0
## bsport bsport    <int> 553        12             2.1  13.9 17.8   0.0
## bqual   bqual    <dbl> 563         2             0.4  22.8 21.0   0.0
## symp9   symp9    <dbl> 430       135            23.9  81.2 19.1  14.3
## pain9   pain9    <dbl> 430       135            23.9  84.1 19.9  13.9
## adl9     adl9    <dbl> 351       214            37.9  94.2 11.1  39.7
## sport9 sport9    <dbl> 347       218            38.6  79.2 24.4   0.0
## qual9   qual9    <dbl> 430       135            23.9  70.9 26.6   0.0
## Yscore Yscore    <dbl> 430       135            23.9   3.9  2.4   0.0
##        quartile_25 median quartile_75   max
## age           21.0   28.0        37.0  72.0
## sex            1.0    2.0         2.0   2.0
## height       165.1  172.7       180.3 200.7
## weight        67.1   76.2        87.1 133.4
## bsymp         28.6   39.3        50.0  96.4
## bpain         27.8   38.9        50.0  94.4
## badl          50.0   58.8        66.2 100.0
## bsport         0.0   10.0        20.0 100.0
## bqual          6.2   18.8        31.2 100.0
## symp9         71.4   85.7        96.4 100.0
## pain9         75.0   91.7       100.0 102.8
## adl9          95.6  100.0       100.0 100.0
## sport9        70.0   85.0       100.0 100.0
## qual9         50.0   75.0        93.8 100.0
## Yscore         2.0    4.1         5.7   9.5
##
## $Categorical
## data frame with 0 columns and 565 rows
```

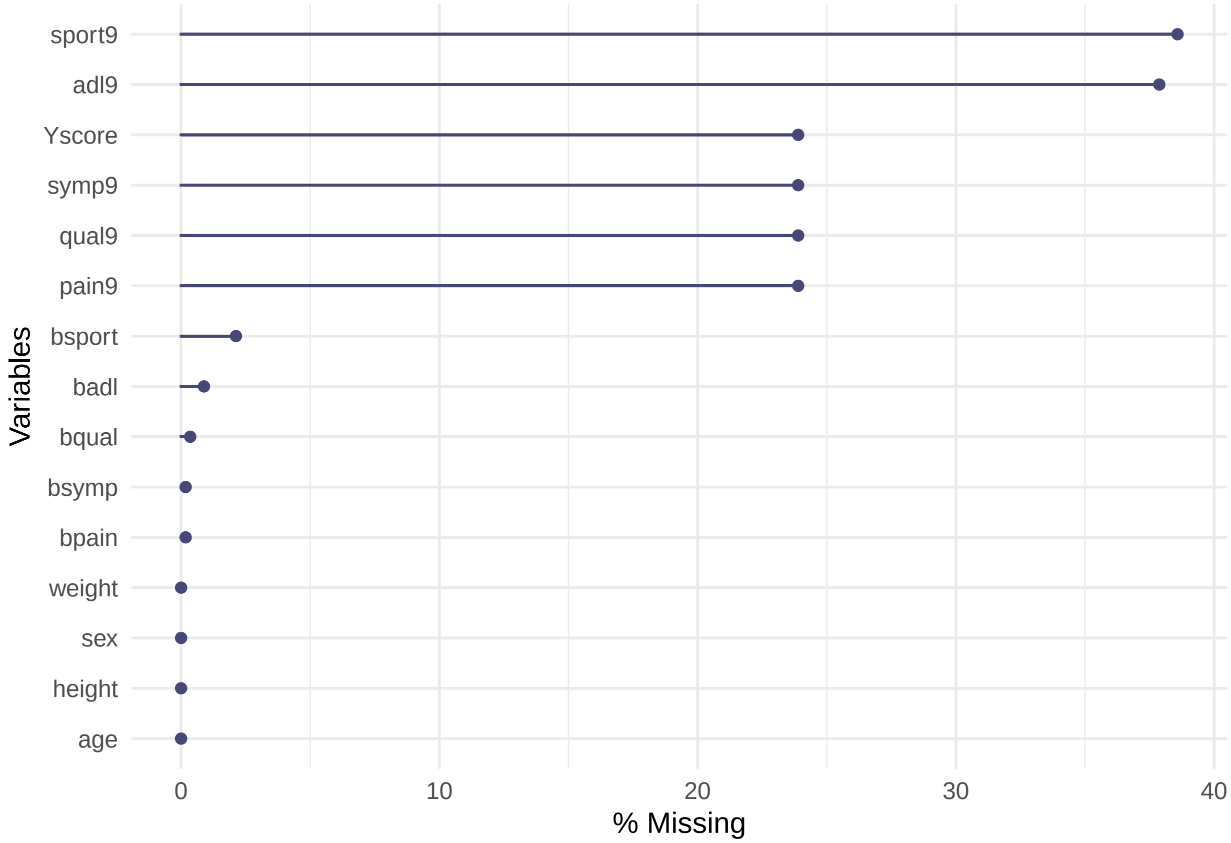
It seems that the CAST dataset encompasses categorical variables and continuous variables. Sex is coded as integer despite being a factor, while the baseline sports variable is also coded as an integer, despite being a continuous variable. We convert sex into a factor with two levels, 0 (female) and 1 (male), and the baseline sports sub-scale into a continuous variable.

The four basic variables (age, sex, height, weight) are recorded for all patients, while the baseline sub-scale measurements are each missing in 1-2 cases, with the exception of the sport&recreation sub-scale which is missing in 12 cases. The subsequent measurements (after 9 months) for pain score, quality of life score and other symptoms score are also missing in lower numbers (all missing in 135 cases) than the daily life score and sports score, which are missing in 214, and, respectively, 218 cases.

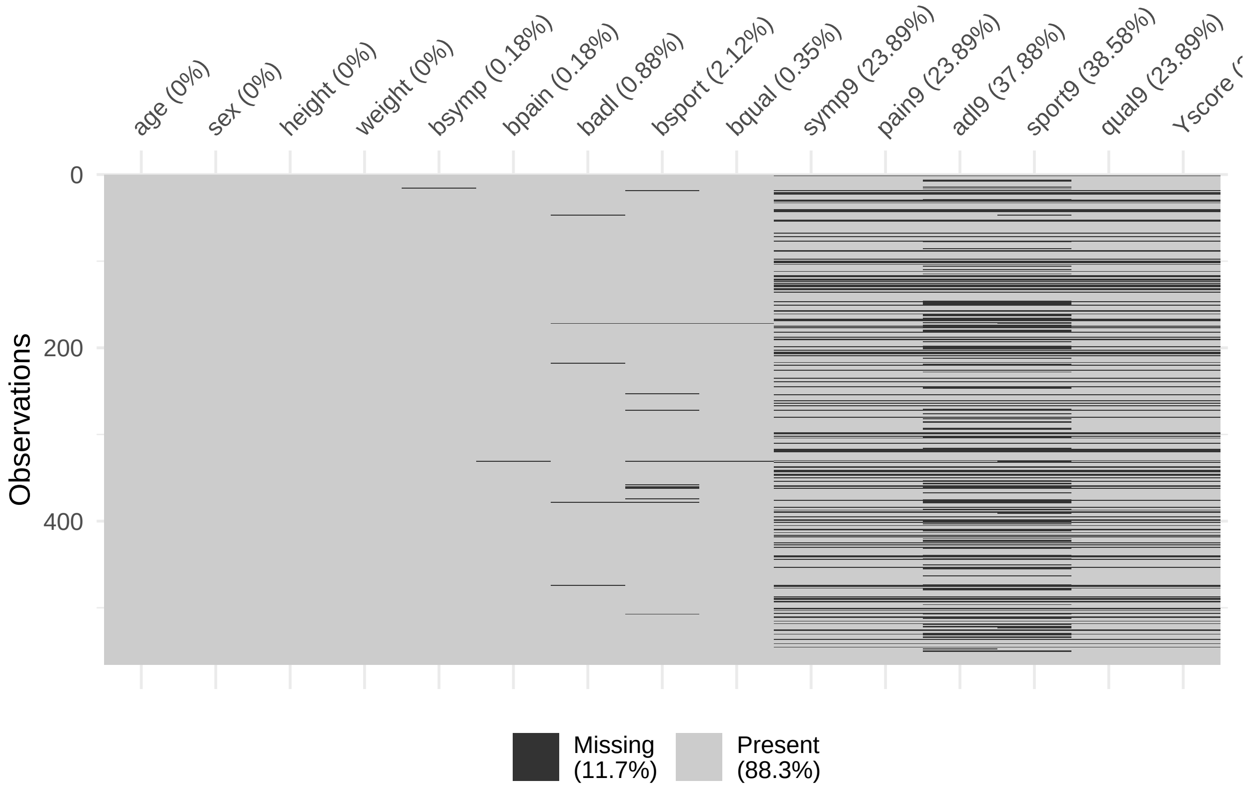
General plots

We begin with some exploratory plots to visualize missingness in the dataset.

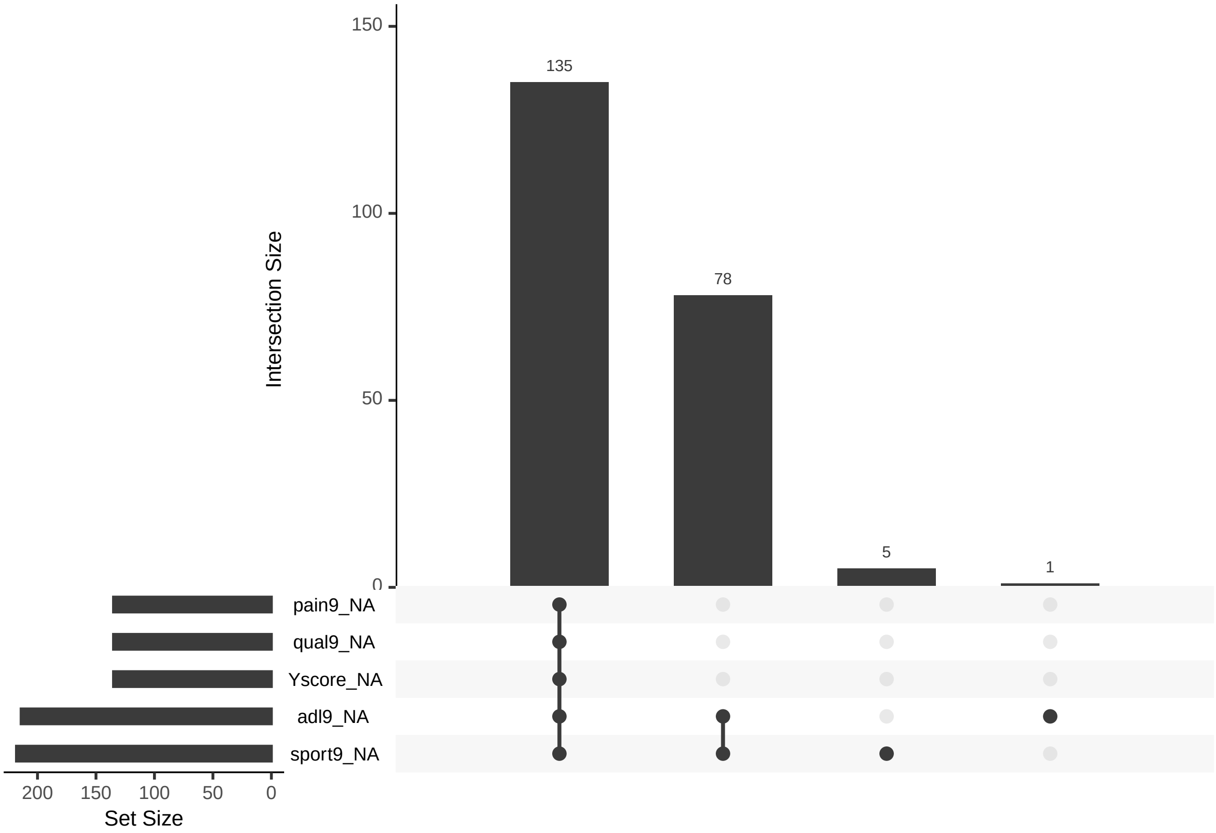
```
gg_miss_var(cast,show_pct=TRUE)
```



```
vis_miss(cast)
```



```
gg_miss_upset(cast)
```



From the figures above, it is noteworthy that missingness is most prevalent among the 9-month measurements of the sub-scales except other symptoms of pain variable, and the composite score, as follows: * around 12% of the values are missing from the dataset (Fig 1) * around 1-2% of the baseline sub-score values are missing (Fig 1) * in 135 cases, all sub-scale variables are missing at 9-month point (Fig 2) * in 78 cases, both daily living score at 9 months and sport & recreation score at 9 months are missing, so there might be a correlation between their missingness patterns (Fig 2) * there are 5 cases where only sports & recreation sub-scale values at 9 months are missing (Fig 2) * there is 1 case where only the daily living scale at 9 months (Fig 2).

General patterns and relationships

The function in sport and recreation variables (bsport, sport9) are missing 2.12% and 38.58% of the time. From Fig 1, it is clear that all 11 variables in question have some degree of missingness so we change the default settings of the visualization function to account for all interactions between all the missing variables (nintersects=NA).

```r
gg_miss_upset(cast, nsets=16,nintersects=NA)
```

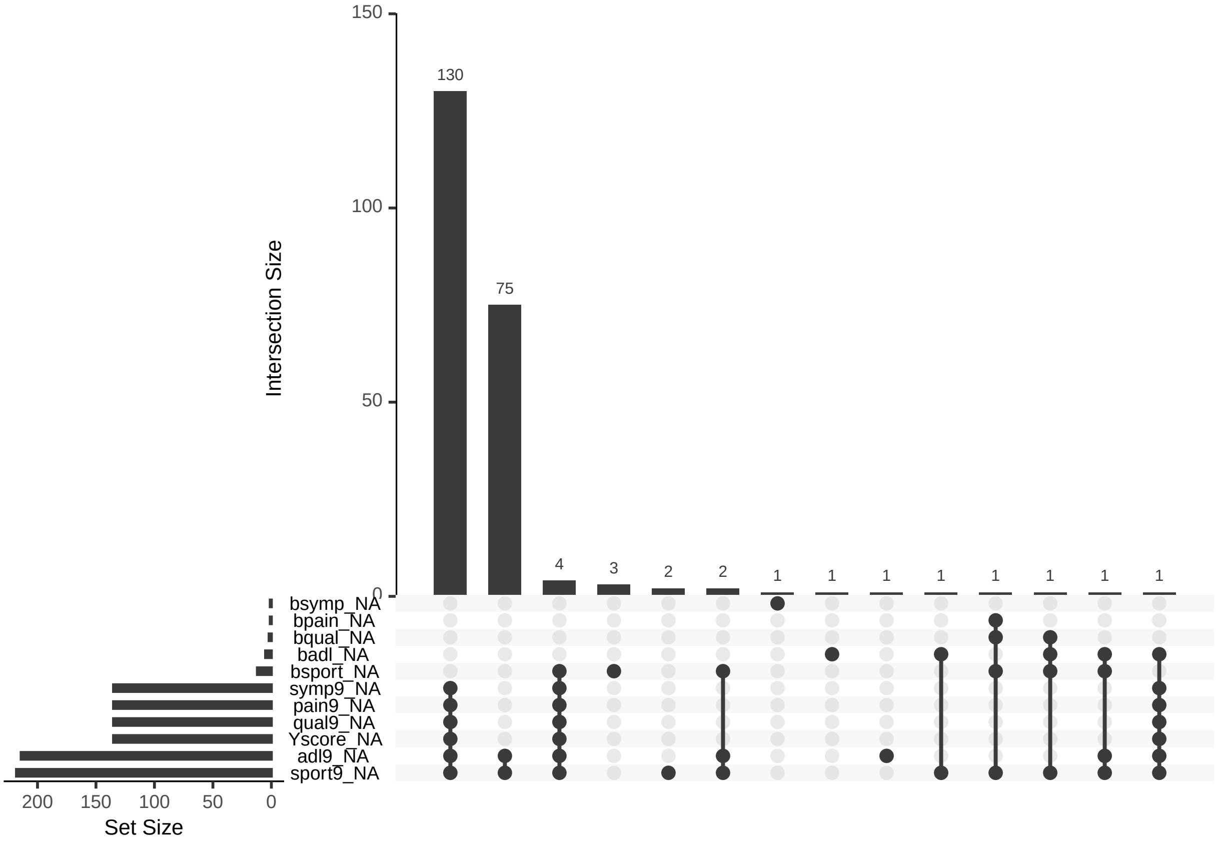
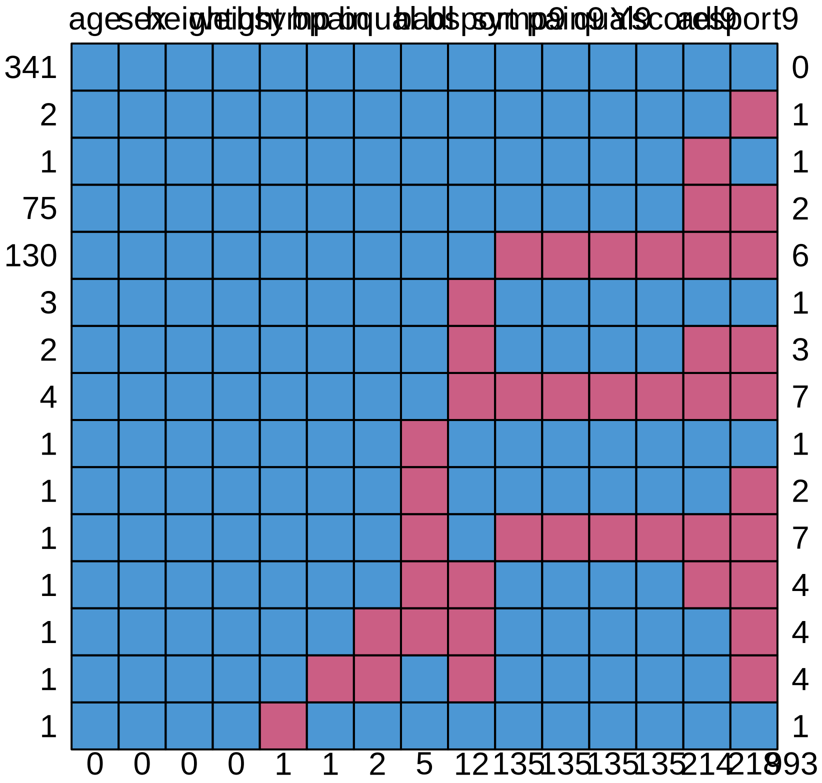


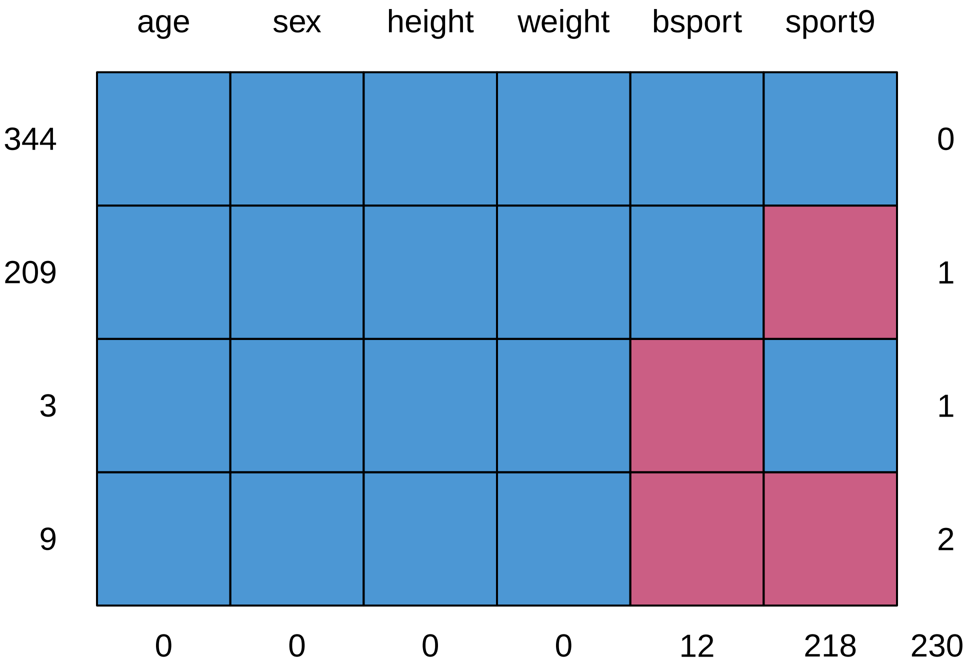
Fig above reveals more patterns of missingness: * there are 130 cases where all sub-scales and the composite score are missing * there are 4 cases when baseline sports & recreation and all other sub-scales at 9 months are missing, but 3 cases when only baseline sports & recreation is missing and 2 cases when only sports & recreation at 9 months is missing * there are 75 cases where sports & recreation and daily living sub-scales are missing * interestingly, across missingness counts, sports variable seems to be missing when daily living variable is missing.



```
##     age sex height weight bsymp bpain bqual badl bsport symp9 pain9 qual9
## 341   1   1      1      1     1     1     1    1      1     1     1     1
## 2     1   1      1      1     1     1     1    1      1     1     1     1
## 1     1   1      1      1     1     1     1    1      1     1     1     1
## 75    1   1      1      1     1     1     1    1      1     1     1     1
## 130   1   1      1      1     1     1     1    1      1     0     0     0
## 3     1   1      1      1     1     1     1    1      0     1     1     1
## 2     1   1      1      1     1     1     1    1      0     1     1     1
## 4     1   1      1      1     1     1     1    1      0     0     0     0
## 1     1   1      1      1     1     1     1    0      1     1     1     1
## 1     1   1      1      1     1     1     1    0      1     1     1     1
## 1     1   1      1      1     1     1     1    0      1     0     0     0
## 1     1   1      1      1     1     1     1    0      0     1     1     1
## 1     1   1      1      1     1     1     0    0      0     1     1     1
## 1     1   1      1      1     1     0     0    1      0     1     1     1
## 1     1   1      1      1     0     1     1    1      1     1     1     1
##         0   0      0      0     1     1     2    5     12   135   135   135
##     Yscore adl9 sport9
## 341      1    1      1   0
## 2        1    1      0   1
## 1        1    0      1   1
## 75       1    0      0   2
## 130      0    0      0   6
## 3        1    1      1   1
## 2        1    0      0   3
## 4        0    0      0   7
## 1         1    1      1   1
## 1         1    1      0   2
## 1         0    0      0   7
## 1         1    0      0   4
## 1         1    1      0   4
## 1         1    1      0   4
## 1         1    1      1   1
##         135  214    218 993
```

Sports & recreation variable

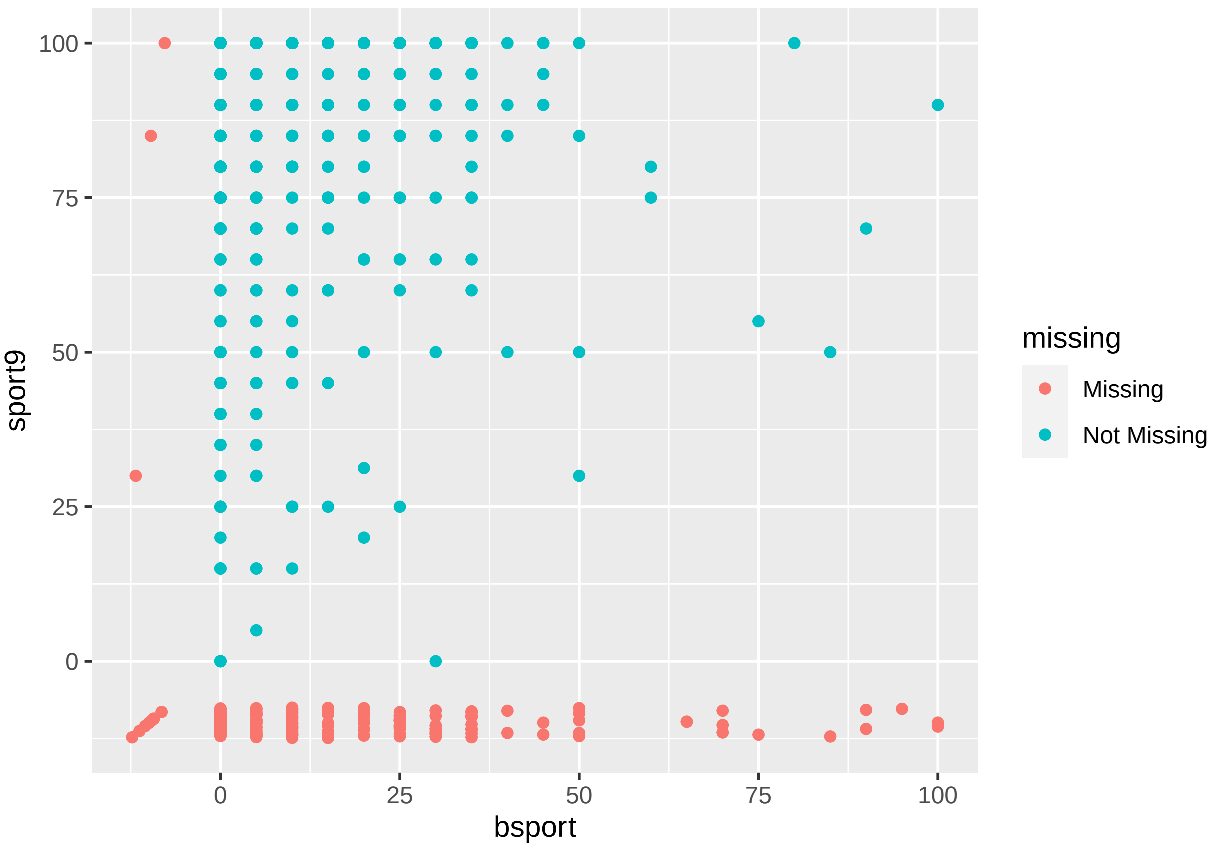
Let's explore a summary of the CAST data with the basic variables and the sports sub-scale.



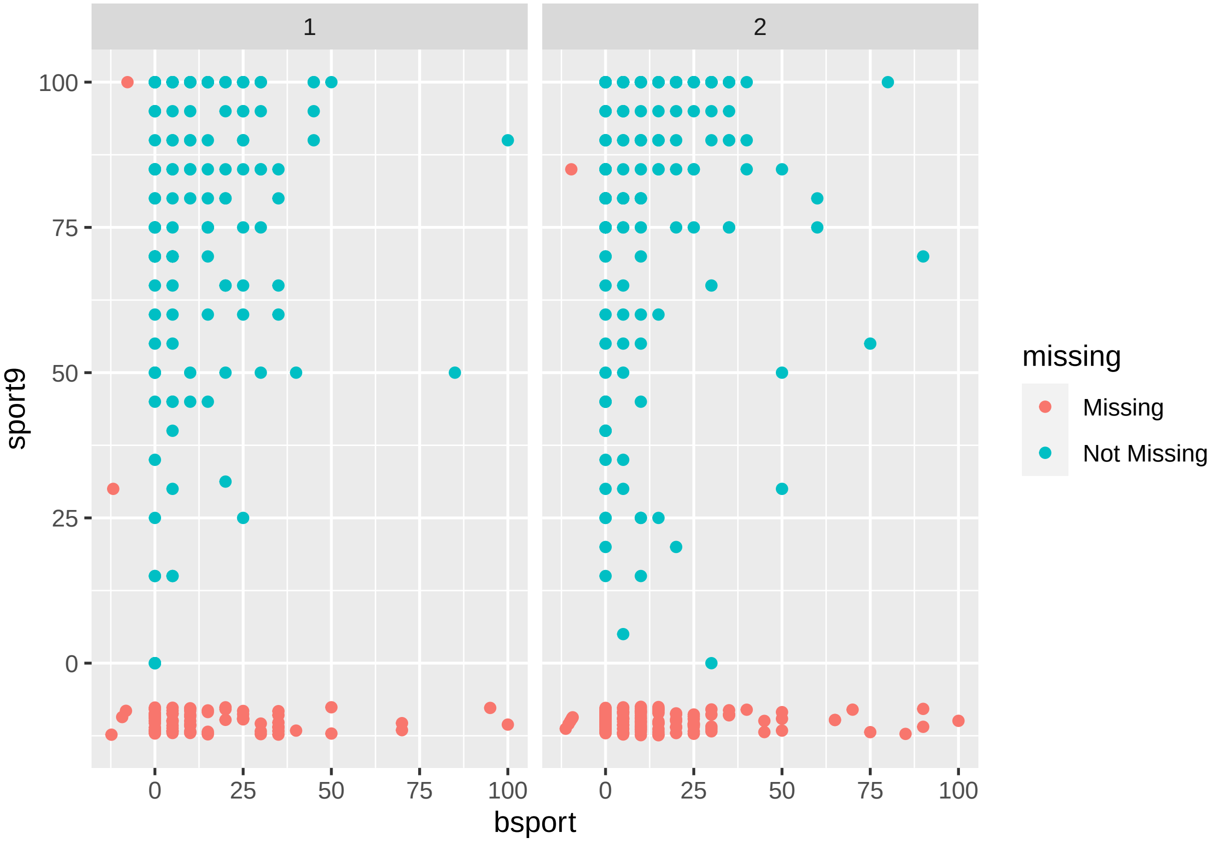
```
##     age sex height weight bsport sport9
## 344   1   1      1      1      1      1   0
## 209   1   1      1      1      1      0   1
## 3     1   1      1      1      0      1   1
## 9     1   1      1      1      0      0   2
##       0   0      0      0     12    218 230
```

There are 304 full observations, 209 observations which are only missing the value at the 9-month time point and 3 observations which are missing the baseline value. The remaining 9 observations are missing both values. The former two categories will be of particular interest to the following analysis.

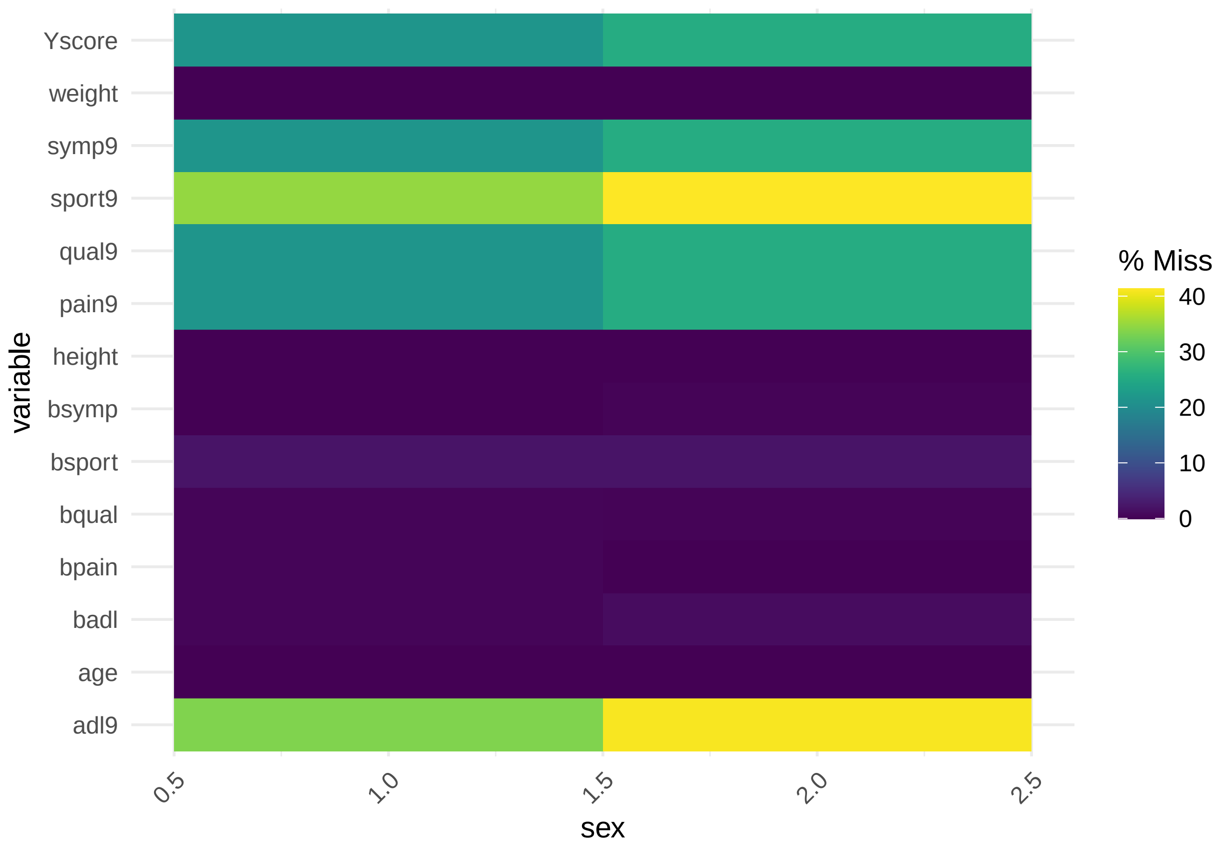
Let's now look at a plot showing the relationship between the prevalence of the sports variable at the two time points- baseline and 9 months.



Let's explore the missingness relationship between baseline and 9-month measurements of the sports sub-scale, dependent on sex.



Let’s look at a plot showing the missingness of the CAST dataset variables by sex.



There seems to be a distinct correlation between missingness of sub-scale values (9-month measurement) and sex, with scores for females missing more sparsely than those for males. At baseline, there does not seem to be a significant difference in willingness to complete the survey between sexes.

Missingness mechanism

We have to decide whether the missing 9-month values of the sports sub-scale are MCAR, MAR or MNAR.

MCAR vs MAR

If the sub-scale variables are MCAR, list-wise deletion would not introduce bias in our models and subsequent inferences. Let's use Little's test [1] to diagnose whether there are any variables missing completely at random in our dataset. We will also create dummy variables for missingness (1 = missing, 0 = observed).

If the sub-scale variables are MAR, their missingness is conditional on other variables, and should therefore be analyzed further.

```
## Iterations of EM:
## 1...2...3...4...5...6...7...8...9...10...11...12...13...14...15...16...17...18...19...20...
## this could take a while
```

```
r[["p.value"]]
```

```
## [1] 9.187207e-12
```

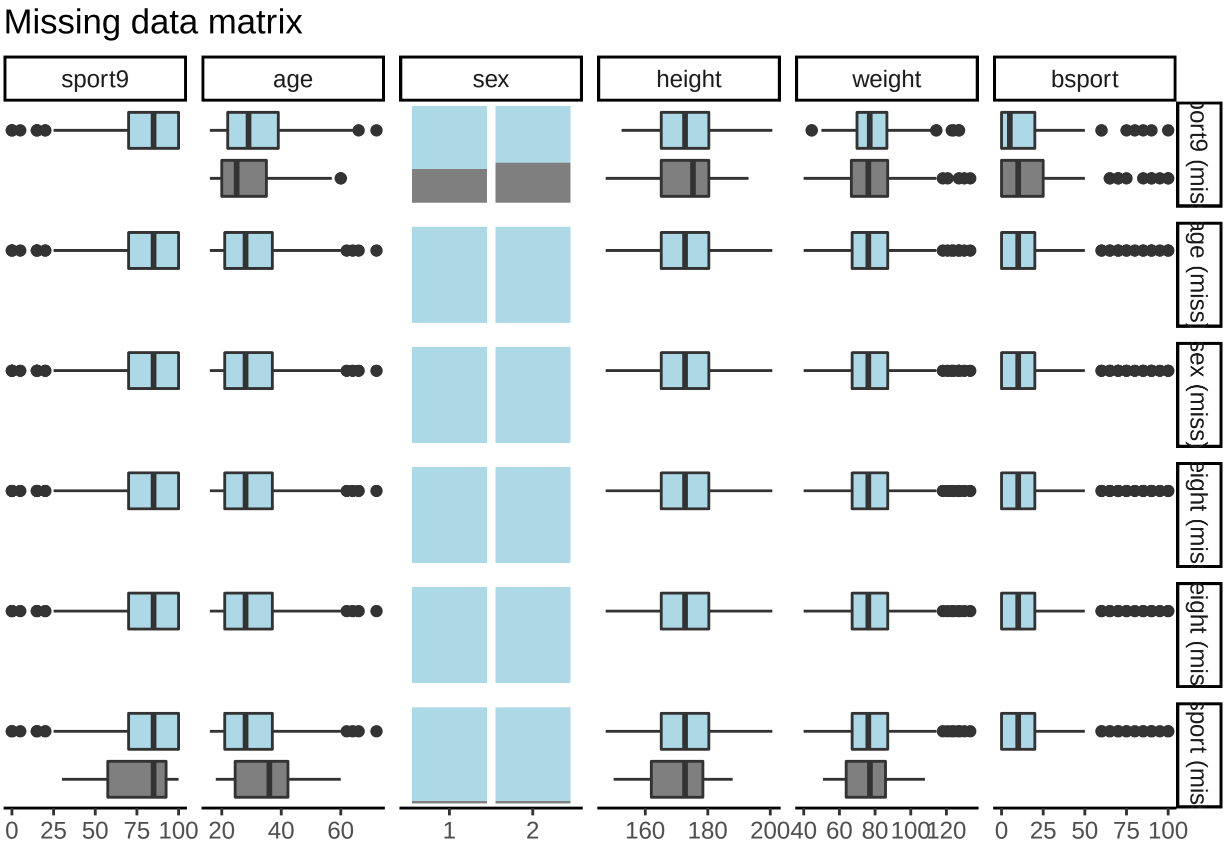
```
r[["missing patterns"]]
```

```
## NULL
```

```
r[["amount.missing"]]
```

```
##                  age sex height weight       bsymp       bpain        badl
## Number Missing     0   0      0      0 1.000000000 1.000000000 5.000000000
## Percent Missing    0   0      0      0 0.001769912 0.001769912 0.008849558
##                      bsport       bqual       symp9       pain9        adl9
## Number Missing   12.00000000 2.000000000 135.0000000 135.0000000 214.0000000
## Percent Missing   0.02123894 0.003539823   0.2389381   0.2389381   0.3787611
##                      sport9       qual9      Yscore
## Number Missing   218.0000000 135.0000000 135.0000000
## Percent Missing    0.3858407   0.2389381   0.2389381
```

The output of Little's MCAR test indicates that no variables are MCAR.



```
##  Missing data analysis: sport9           Not missing      Missing     p
##                             age Mean (SD) 31.1 (11.2)   28.2 (9.9) 0.002
##                             sex         1  154 (65.3)    82 (34.7) 0.134
##                                         2  193 (58.7)   136 (41.3)
##                          height Mean (SD) 172.7 (9.4) 173.1 (10.4) 0.635
##                          weight Mean (SD) 78.9 (14.6)  78.1 (16.7) 0.553
##                          bsport Mean (SD) 12.6 (15.7)  16.1 (20.7) 0.022
```

```
##  Missing data analysis: bsport           Not missing      Missing     p
##                             age Mean (SD) 29.9 (10.8)  35.6 (12.3) 0.072
##                             sex         1  231 (97.9)      5 (2.1) 1.000
##                                         2  322 (97.9)      7 (2.1)
##                          height Mean (SD) 172.9 (9.7) 170.6 (13.2) 0.413
##                         weight Mean (SD) 78.6 (15.4)  75.7 (17.1) 0.515
```

The output indicates a signficant relationship between age and the missingness of the 9-month sport sub-scale measurement, and a weak relationship between the two sport sub-scale variables. This result, coupled with Little's test, allows us to infer that bsport seems to be MCAR with respect to the basic variables, while sport9 is at least MAR.

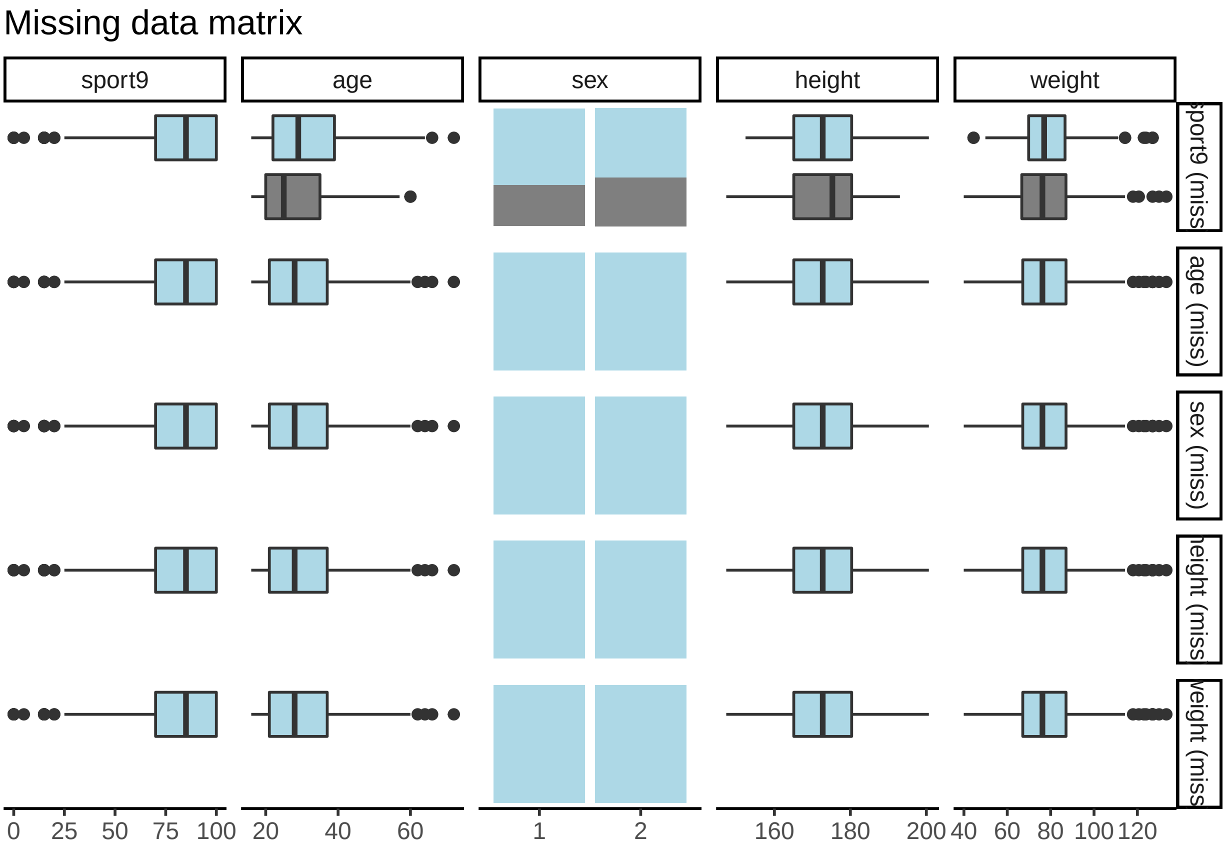
Missingness Solution for Baseline Values

Because "bsport" is MCAR, the missing values can be completed by using multiple imputation. Prior to this, we perform sensitivity analysis because this variable is very important to the subsequent statistical modelling. There are only 12 observations missing the bsport value, which accounts for approximately 2% of the observations.

We complete the missing values of bsport by performing multiple imputation.

MNAR

After concluding that sport9 is at least MAR, let's investigate if it can be MNAR. Understanding the missingness mechanism guides the statistical modelling technique.



Modelling missingness

Let's start by plotting the correlations between the dummy sport9na variable and the basic variables (age, sex, weight, height).

```
## corrplot 0.84 loaded
```

We split the dataset into three subsets: a training subset (10% of the data), which is used for fitting the models, a testing subset(80% of the data), which is used for comparing the estimation power of our models, and a prediction subset(10% of the data), which is used for assessing the predictive power of our final model.

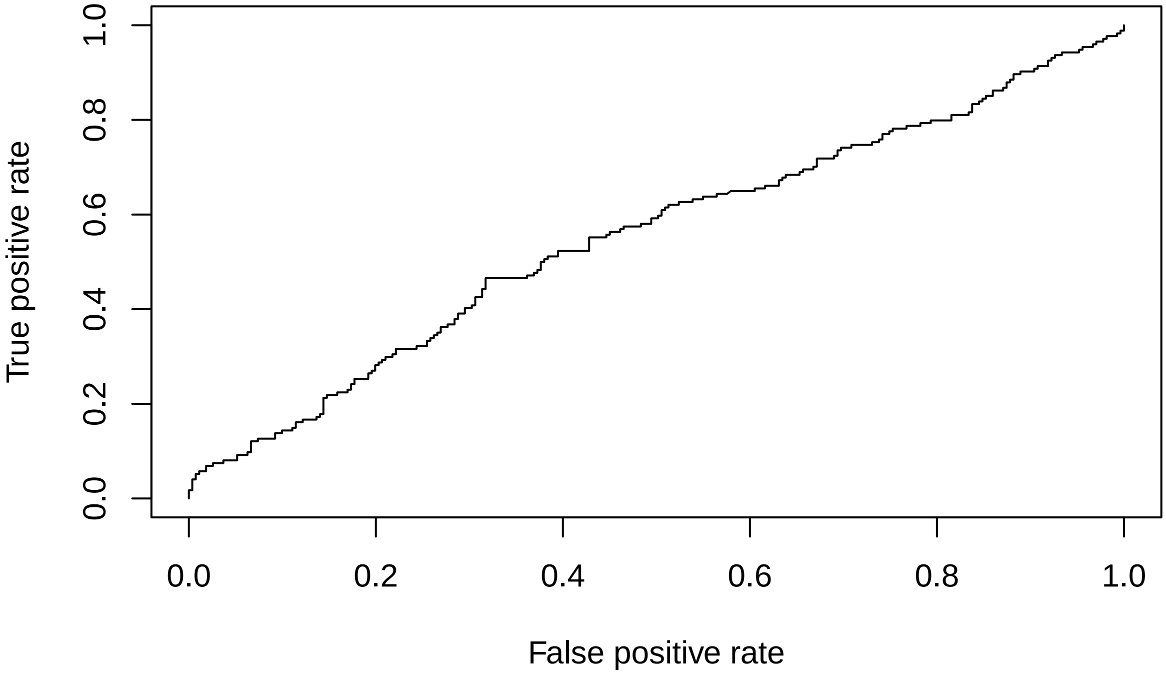
In an additive model including all the variables (age, sex, weight, height, bsport), there seems to be no statistically significant variable.

```
##
## Call:
## glm(formula = sport9na ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2562  -0.9255  -0.6976   1.2317   2.0375
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.206791   8.351333  -1.342    0.180
## age          -0.001745   0.027755  -0.063    0.950
## sex1         -0.625055   0.981110  -0.637    0.524
## height        0.081797   0.054725   1.495    0.135
## weight       -0.045094   0.028507  -1.582    0.114
## bsport1       0.024542   0.016110   1.523    0.128
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 71.743  on 55  degrees of freedom
## Residual deviance: 67.041  on 50  degrees of freedom
## AIC: 79.041
##
## Number of Fisher Scoring iterations: 4
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: sport9na
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                       55     71.743
## age      1  0.41643        54     71.326   0.5187
## sex      1  0.00728        53     71.319   0.9320
## height   1  0.13387        52     71.185   0.7144
## weight   1  1.82404        51     69.361   0.1768
## bsport1  1  2.31993        50     67.041   0.1277
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
## fitting null model for pseudo-r2

##          llh      llhNull           G2     McFadden         r2ML         r2CU
## -33.52060762 -35.87139131   4.70156737   0.06553366   0.08052880   0.11149339
```



```
## [1] 0.5537494
```

Because sport9na is a binary variable, it has a binomial-family distribution. Therefore we perform logistic regression by constructing and analyzing GLMs

First, we try a simple regression model, with age as the explanatory variable. It seems to be statistically significant with a p-value of 0.0021. Then we try out an additive model of all the variables and a model containing all variables and one model which additionally contains all their interactions, without the age variable. In the former, age and bsport seem to be statistically significant, while the latter seems to have bsport as the only statistically relevant explanatory variable.

In a model with age, bsport and their interaction, it seems that only age is statistically significant.

```
##
## Call:
## glm(formula = sport9na ~ age, family = "binomial", data = castsp2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1302  -1.0246  -0.8671   1.2928   1.7331
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.30181    0.26082   1.157   0.2472
## age         -0.02586    0.00841  -3.075   0.0021 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 753.54  on 564  degrees of freedom
## Residual deviance: 743.67  on 563  degrees of freedom
## AIC: 747.67
##
## Number of Fisher Scoring iterations: 4
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   0.302    0.261        1.16 0.247
## 2 age          -0.0259   0.00841     -3.08 0.00210
```

```
##
## Call:
## glm(formula = sport9na ~ age + factor(sex) + weight + height +
##     bsport1, family = "binomial", data = castsp2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4607  -1.0000  -0.8278   1.3113   1.8444
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.2507656  2.1885136   1.028   0.3037
## age          -0.0239399  0.0089391  -2.678   0.0074 **
## factor(sex)1  0.3721223  0.2496539   1.491   0.1361
## weight       -0.0007814  0.0062547  -0.125   0.9006
## height       -0.0134742  0.0135063  -0.998   0.3185
## bsport1       0.0114380  0.0047180   2.424   0.0153 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 753.54  on 564  degrees of freedom
## Residual deviance: 735.00  on 559  degrees of freedom
## AIC: 747
##
## Number of Fisher Scoring iterations: 4
```

```
## # A tibble: 6 x 5
##   term          estimate std.error statistic p.value
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   2.25       2.19        1.03  0.304
## 2 age          -0.0239     0.00894    -2.68  0.00740
## 3 factor(sex)1  0.372      0.250       1.49  0.136
## 4 weight       -0.000781   0.00625    -0.125 0.901
## 5 height       -0.0135     0.0135     -0.998 0.318
## 6 bsport1       0.0114     0.00472     2.42  0.0153
```

```
##
## Call:
## glm(formula = sport9na ~ (factor(sex) + weight + height + bsport1) *
##     (factor(sex) + weight + height + bsport1), family = "binomial",
##     data = castsp2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6984  -0.9765  -0.8545   1.3252   1.9867
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.8999504 10.8351319   0.268   0.7890
## factor(sex)1         -4.5086211  4.7686169  -0.945   0.3444
## weight               -0.0450649  0.1317839  -0.342   0.7324
## height               -0.0258704  0.0663204  -0.390   0.6965
## bsport1               0.2888534  0.1307145   2.210   0.0271 *
## factor(sex)1:weight  -0.0154542  0.0176996  -0.873   0.3826
## factor(sex)1:height   0.0337968  0.0277510   1.218   0.2233
## factor(sex)1:bsport1  0.0223396  0.0167674   1.332   0.1828
## weight:height         0.0003158  0.0008008   0.394   0.6933
## weight:bsport1       -0.0004100  0.0003809  -1.076   0.2817
## height:bsport1       -0.0014945  0.0008454  -1.768   0.0771 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 753.54  on 564  degrees of freedom
## Residual deviance: 731.43  on 554  degrees of freedom
## AIC: 753.43
##
## Number of Fisher Scoring iterations: 4
```

```
## # A tibble: 11 x 5
##    term                  estimate std.error statistic p.value
##    <chr>                    <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)           2.90     10.8          0.268  0.789
##  2 factor(sex)1         -4.51      4.77        -0.945  0.344
##  3 weight               -0.0451    0.132       -0.342  0.732
##  4 height               -0.0259    0.0663      -0.390  0.696
##  5 bsport1               0.289     0.131        2.21   0.0271
##  6 factor(sex)1:weight  -0.0155    0.0177      -0.873  0.383
##  7 factor(sex)1:height   0.0338    0.0278       1.22   0.223
##  8 factor(sex)1:bsport1  0.0223    0.0168       1.33   0.183
##  9 weight:height         0.000316  0.000801     0.394  0.693
## 10 weight:bsport1       -0.000410  0.000381    -1.08   0.282
## 11 height:bsport1       -0.00149   0.000845    -1.77   0.0771
```

```
##
## Call:
## glm(formula = sport9na ~ factor(sex) + weight + height + bsport1,
##     family = "binomial", data = castsp2)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max  
## -1.4185  -0.9900  -0.8837   1.3453   1.6585  
## 
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)   
## (Intercept)    0.393576   2.070905   0.190  0.84927   
## factor(sex)1   0.385703   0.248391   1.553  0.12047   
## weight        -0.003864   0.006128  -0.631  0.52832   
## height        -0.005586   0.013155  -0.425  0.67112   
## bsport1        0.012624   0.004677   2.699  0.00695 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 753.54  on 564  degrees of freedom
## Residual deviance: 742.42  on 560  degrees of freedom
## AIC: 752.42
## 
## Number of Fisher Scoring iterations: 4
```

```
## # A tibble: 5 x 5
##   term         estimate std.error statistic p.value
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   0.394     2.07        0.190 0.849  
## 2 factor(sex)1  0.386     0.248       1.55  0.120  
## 3 weight       -0.00386   0.00613    -0.631 0.528  
## 4 height       -0.00559   0.0132     -0.425 0.671  
## 5 bsport1       0.0126    0.00468     2.70  0.00695
```

```
## # A tibble: 4 x 5
##   term         estimate std.error statistic p.value
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  0.282     0.341        0.827 0.409  
## 2 bsport1     -0.00194   0.0146      -0.133 0.894  
## 3 age         -0.0310    0.0110      -2.83  0.00467
## 4 bsport1:age  0.000481  0.000491     0.980 0.327  
```

We perform stepwise regression on the sport9na variable, with the maximum model containing all variables and their interactions, and the minimum model being the identical model.

Modelling Yscore

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  -0.420    0.317       -1.33 0.185  
## 2 bsport        0.0100   0.00495      2.03 0.0427 
## 3 age          -0.0307   0.0101      -3.02 0.00252
```

Limitations

- (bsport) List-wise deletion of the observations in which bsport is missing introduces a bias in the inferences. There is a weak relationship between bsport and age, which means that list-wise deletion implies keeping the observations from younger people, on average. This may have consequences for our conclusions if age is associated with sport9.

- (Yscore) List-wise deletion of observations leads to a bigger drop in male than female participants (as the score is proportionately missing), which might affect the inferences
- (reducing to no na's) Reduction of the dataset to participants with no missing data introduces a higher bias in the results. Apart from subsetting participants with no missing observations for the baseline scores, which account for a fairly small percentage of the dataset (approx. 12%), we would also drop all participants with at least one missing sub-scale, which account for 40% of the dataset.

References

1. Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83(404), 1198–1202.
2. Jamshidian, M. Jalal, S., and Jansen, C. (2014). "MissMech: An R Package for Testing Homoscedasticity, Multivariate Normality, and Missing Completely at Random (MCAR)," Journal of Statistical Software, 56(6), 1-31. URL http://www.jstatsoft.org/v56/i06/.