NHANES Data Analysis Project

STAT 420, Summer 2023, Preeti Agrawal, Thimira Bandara, Michael Conlin, Constatin Kappel

2023-07-26

Title

Introduction

Methods

Data Import

```
if (!require(NHANES)) {
  install.packages("NHANES")
}
```

```
library(NHANES)
# head(NHANES)
```

Initial Variable Selection

Ruling out variables by reasoning or by exploratory analysis We have chosen `BMI` (Body mass index (weight/height2 in kg/m2)) as our response variable. In NHANES, this data is reported for participants aged 2 years or older, so that will focus on participants over 2 years old for our analysis. Provided below are all the predictors in NHANES, along with our response variable `BMI`.

```
sort(names(NHANES)) # alphabetic order
```

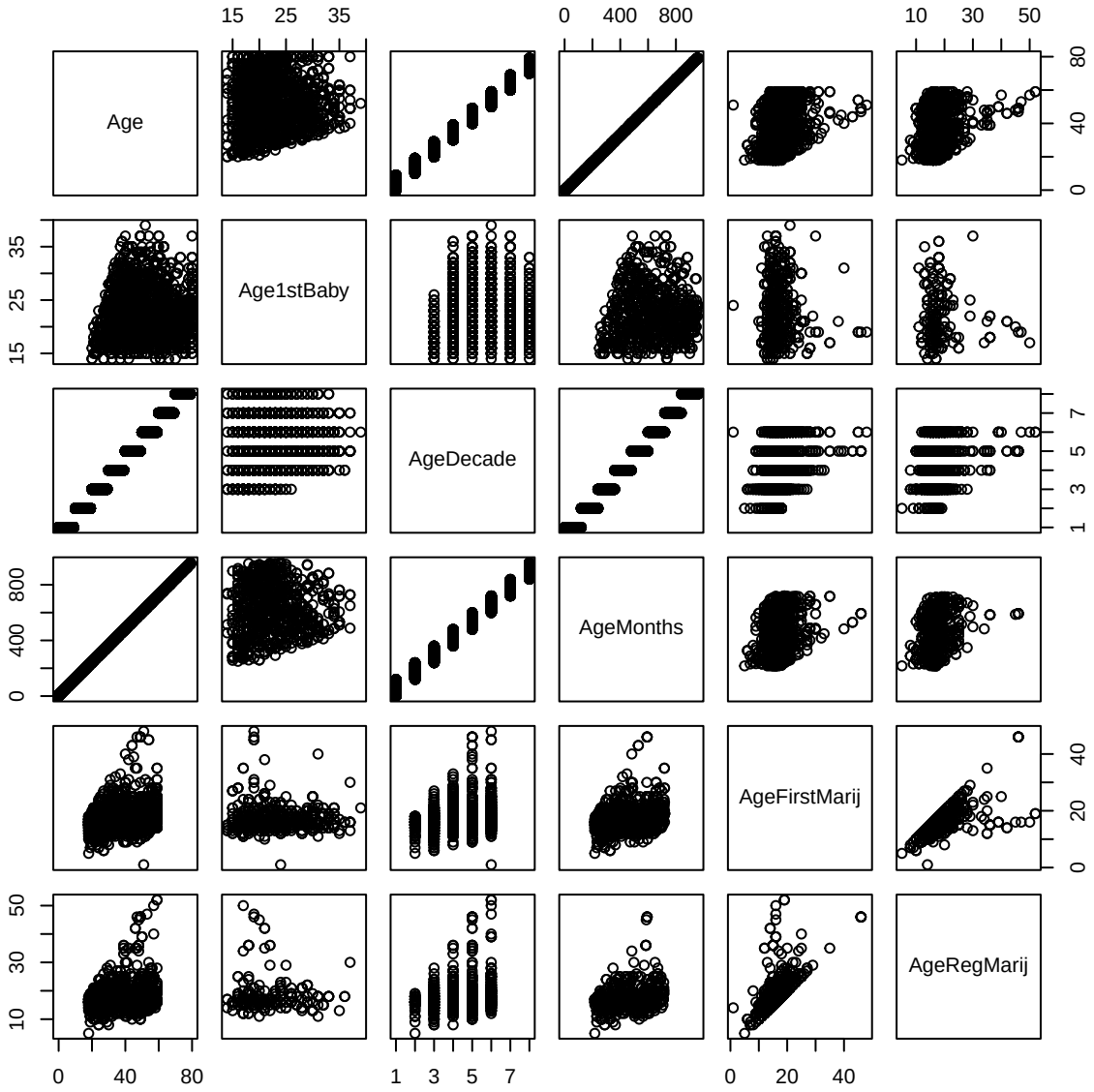
```
##  [1] "Age"              "Age1stBaby"       "AgeDecade"        "AgeFirstMarij"
##  [5] "AgeMonths"        "AgeRegMarij"      "Alcohol12PlusYr"  "AlcoholDay"
##  [9] "AlcoholYear"      "BMI"              "BMI_WHO"          "BMICatUnder20yrs"
## [13] "BPDia1"           "BPDia2"           "BPDia3"           "BPDiaAve"
## [17] "BPSys1"           "BPSys2"           "BPSys3"           "BPSysAve"
## [21] "CompHrsDay"       "CompHrsDayChild"  "DaysMentHlthBad"  "DaysPhysHlthBad"
## [25] "Depressed"        "Diabetes"         "DiabetesAge"      "DirectChol"
## [29] "Education"        "Gender"           "HardDrugs"        "HeadCirc"
## [33] "HealthGen"        "Height"           "HHIncome"         "HHIncomeMid"
## [37] "HomeOwn"          "HomeRooms"        "ID"               "Length"
## [41] "LittleInterest"   "Marijuana"        "MaritalStatus"    "nBabies"
## [45] "nPregnancies"     "PhysActive"       "PhysActiveDays"   "Poverty"
## [49] "PregnantNow"      "Pulse"            "Race1"            "Race3"
## [53] "RegularMarij"     "SameSex"          "SexAge"           "SexEver"
```

```
## [57] "SexNumPartnLife"  "SexNumPartYear"   "SexOrientation"   "SleepHrsNight"
## [61] "SleepTrouble"     "Smoke100"         "Smoke100n"        "SmokeAge"
## [65] "SmokeNow"         "SurveyYr"         "Testosterone"     "TotChol"
## [69] "TVHrsDay"         "TVHrsDayChild"    "UrineFlow1"       "UrineFlow2"
## [73] "UrineVol1"        "UrineVol2"        "Weight"           "Work"
```

We will add all omitted predictors into an overview table `df_exclude`. The variables we would like to keep will be in `df_keep`.

1. Some predictors can be ruled out right away. Our response variable is `BMI`, so we should not use body `Weight` or `Height` as predictors, because `BMI` is calculated by dividing the `Weight` by `Height`.
2. The next group of predictors seems very closely related either by name or logic deduction, for example, `Age`, `AgeDecade`, `AgeMonths`. Let's quickly double-check if they are linearly related:

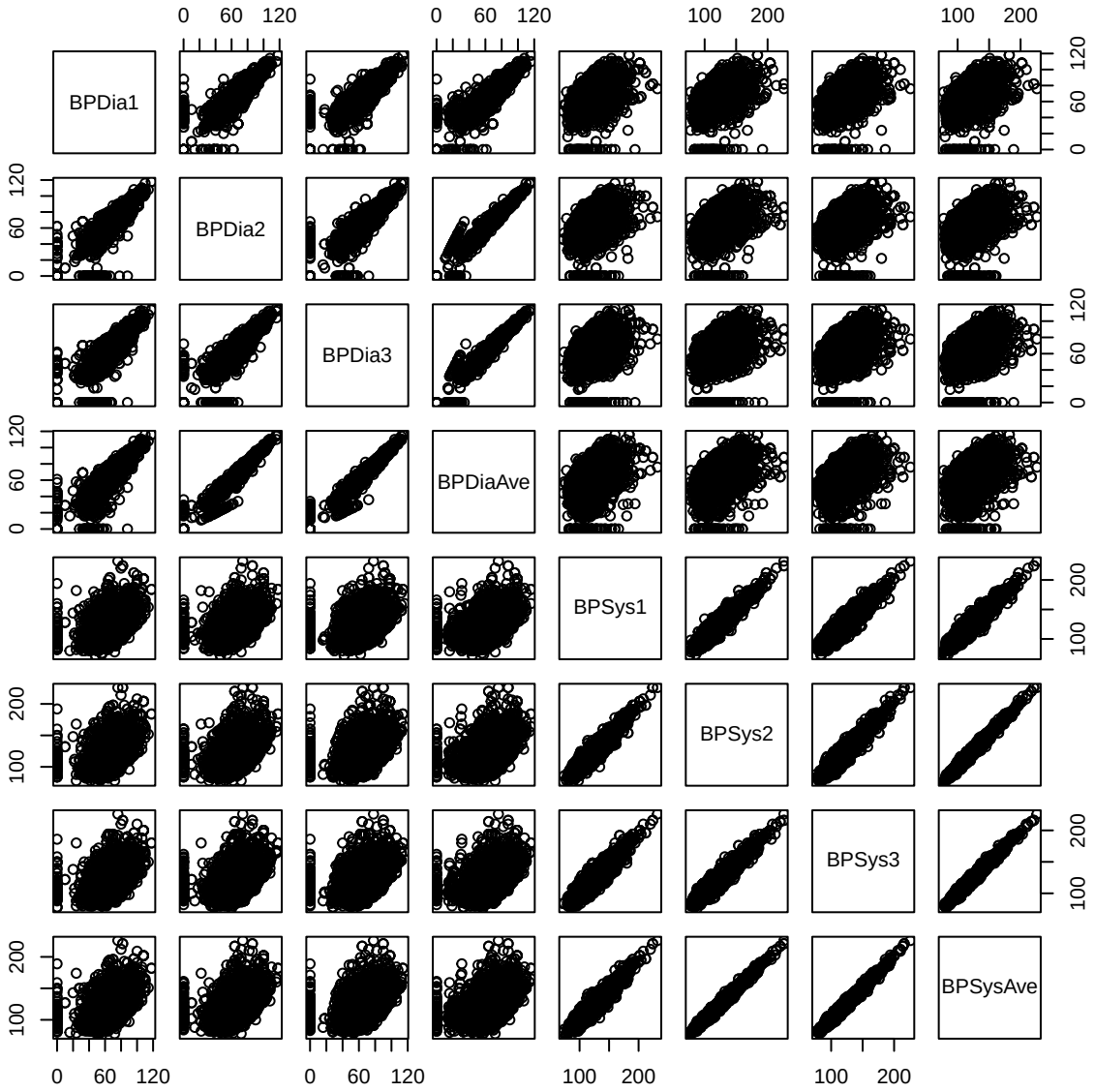
```
pairs(subset(NHANES, select = c('Age', 'Age1stBaby', 'AgeDecade', 'AgeMonths',
                                'AgeFirstMarij', 'AgeRegMarij')))
```



`Age`, `AgeDecade` and `AgeMonth` are clearly collinear, so we will only keep `Age`. Likewise, both variables for Marijuana use appear collinear, so we keep only one, say `AgeRegMarij` and may decide to drop it later if it is not useful. We might keep `Age1stBaby`.

3. Now let's check for collinearity between different blood pressure related variables:

```
to_test = c("BPDia1", "BPDia2", "BPDia3", "BPDiaAve", "BPSys1", "BPSys2", "BPSys3", "BPSysAve" )
pairs(subset(NHANES, select=to_test))
```



The blood pressure variables fall into two groups: diastolic and systolic blood pressure readings. We would expect there to be strongly collinearity within each group, which is the case. So, we only keep the average in each group `BPDiaAve` and `BPSysAve`.

For the following variable assessments for collinearity, see the appendix for their `pairs()` plot.

4. Next, let's check all variables related to alcohol: We again performed a `pairs()` graph to visualize possible collinearity, and this graph is in the appendix. Collinearity is not as clear in this case, but we believe one predictor related to alcohol consumption may be sufficient. We will keep `AlcoholYear`.

5. Let's now investigate the collinearity of other drug-related variables (note: `AgeRegMarij` was in the other group as well and we kept it). Most of these predictors are categorical, so collinearity cannot be seen, except for `SmokeAge` and `AgeRegMarij`. The latter makes sense as this drug is usually consumed

via smoking. We can thus use one as a proxy for the other. Let's keep `SmokeNow` and `HardDrugs` as proxies for drug abuse and its potential effect on BMI.

6. Next, let's investigate a few life-style variables related to being physically active or the opposite thereof, screen time: Due to the nature of these variables being categorical, a clear picture of collinearity is not observable. Let's keep half of these parameters for now, which are the ones with a bit denser levels, `PhysActiveDays`, `TVHrsDay`, `CompHrsDay`.

7. Now we should look into some other health related variables. Let's see for cholesterol and diabetes related predictors: `DirectChol` and `TotChol` appear to be collinear, let's keep `TotChol`. Out of the diabetes related ones, we keep `Diabetes`.

8. Now let's analyze more health related variables, such as those related to urine volume and flow below: Urine volume and urine flow appear collinear. Moreover, there might be collinearity between the first and second urine measurement, respectively. Let's keep `UrineVol1` for now.

9. Next up are a somewhat heterogenic group of variables related to health or mental health. For example, somebody who is depressed might show little interest in doing things. Again, collinearity is not easy to spot in categorical variables. Let's pick `LittleInterest` as a mild form of mental health issue (which might lead to little physical activity and obesity) and `HealthGen` as a general health rating.

10. We decided to keep `Poverty` which is a ratio of family income to poverty guidelines, and drop `HHIncomeMid` and `HHIncome`, as they both capture similar information to what the `Poverty` variable captures.

11. Finally, let's add `nPregnancies`, `Poverty`, `SleepHrsNight`, `Gender`, `Race1`, `Education`, and `MartialStatus` as we believe they can have an effect on BMI, and do not suspect collinearity.

```
#Setting up the data frames with the variables we will be excluding and keeping for model building

df_exclude = data.frame(predictor = c('Weight', 'Height', 'Age1stBaby', 'AgeDecade', 'AgeMonth', 'AgeReg
reason_to_omit = c('linear dependence with BMI','linear dependence with BMI', 'redundant with nPregnanci

df_keep = data.frame(predictor = c('SurveyYr', 'Age', 'AlcoholYear', 'Marijuana', 'SmokeNow', 'HardDrugs
knitr::kable(df_keep, caption = "Initial Predictor Selected")
```

Table 1: Initial Predictor Selected

predictor
SurveyYr
Age
AlcoholYear
Marijuana
SmokeNow
HardDrugs
BPDiaAve
BPSysAve
PhysActiveDays
TVHrsDay
CompHrsDay
TotChol
Diabetes
UrineVol1
HealthGen
LittleInterest

predictor
nPregnancies
Poverty
SleepHrsNight
Gender
Race1
Education
MaritalStatus

Missing Values

Variable Selection

Model Building

Stepwise Model Selection

Model Diagnostics

Collinearity

Variance Assessment

Normality Assessment

Data Interactions

Data Transformations

Outlier Assessment

Final Model

Results

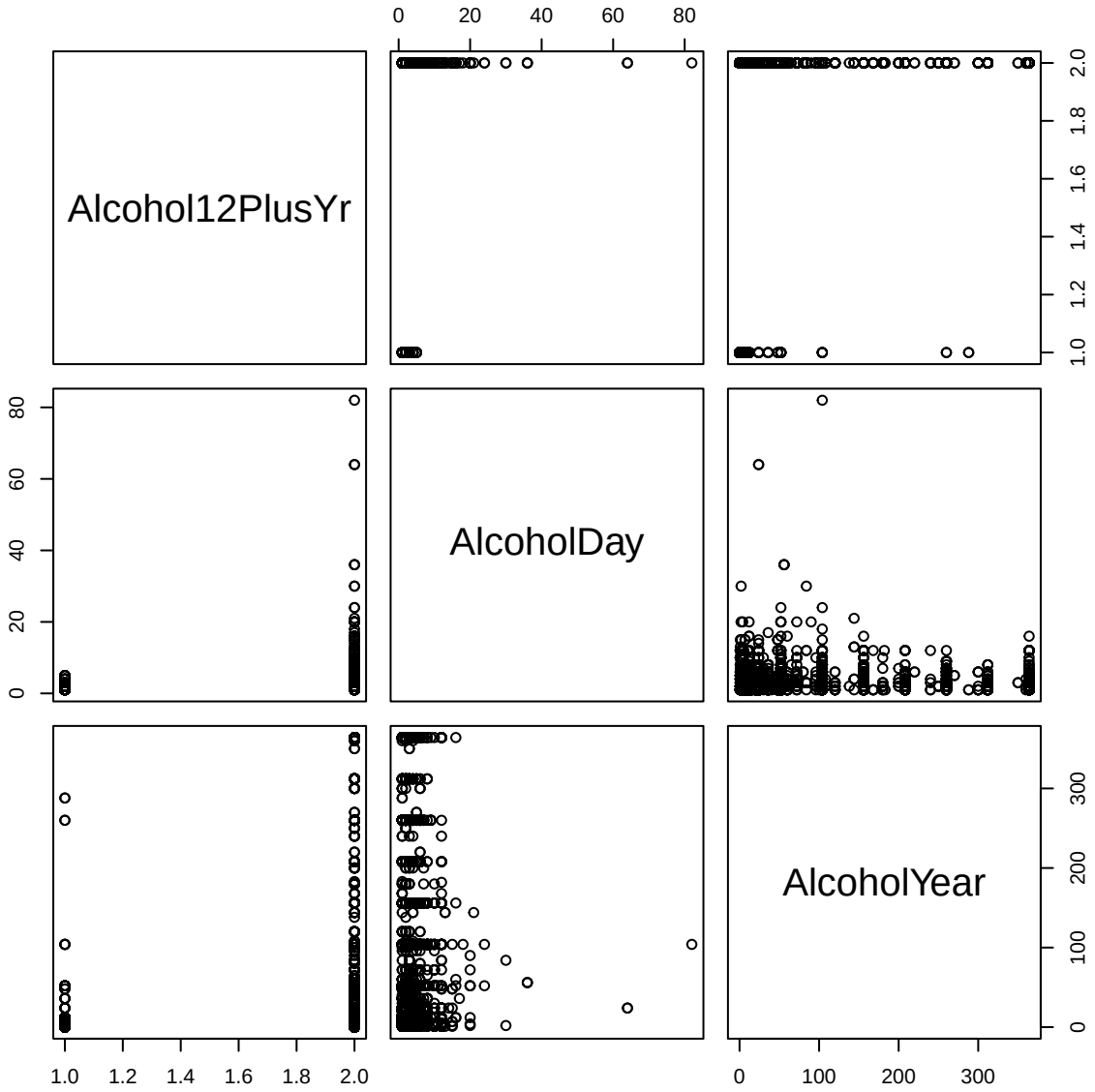
Discussion

Appendix

Variable Selection

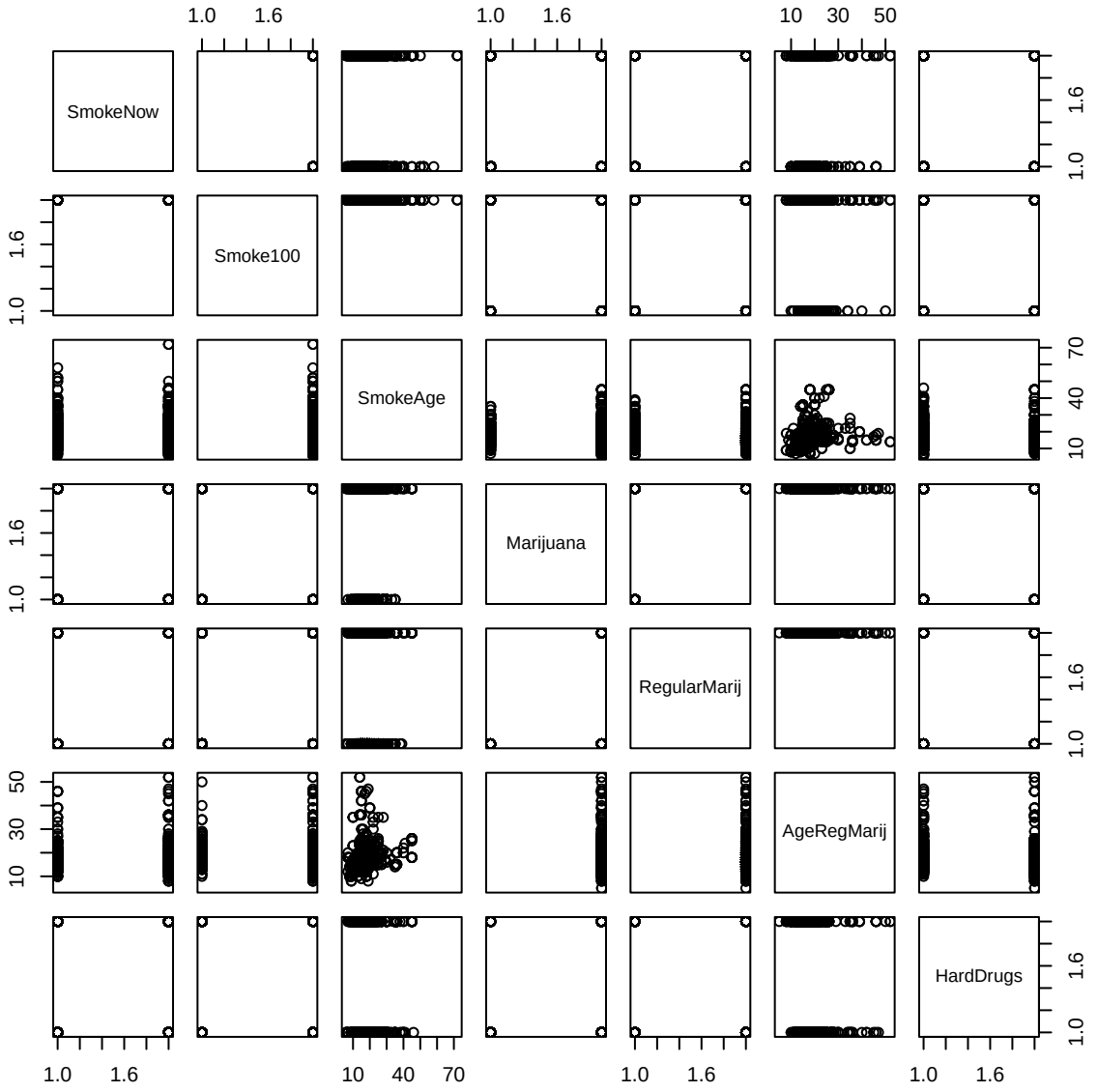
Additional `pairs()` Plots for Collinearity Assessment Alcohol related variables:

```
to_test = c("Alcohol12PlusYr", "AlcoholDay", "AlcoholYear")
pairs(subset(NHANES, select = to_test))
```



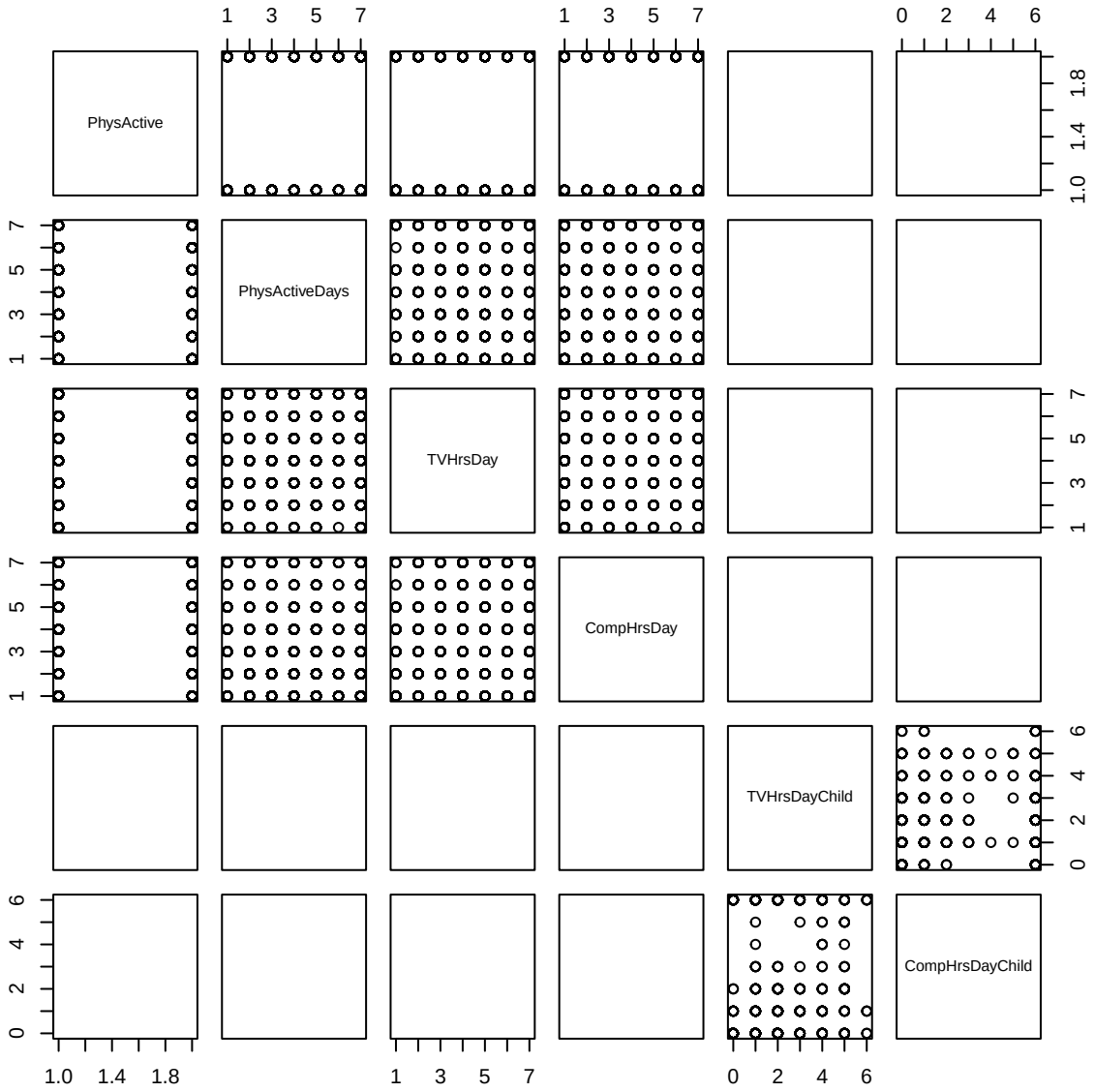
Smoking and Drug related variables:

```
to_test = c("SmokeNow", "Smoke100", "SmokeAge", "Marijuana", "RegularMarij", "AgeRegMarij", "HardDrugs"
pairs(subset(NHANES, select = to_test))
```



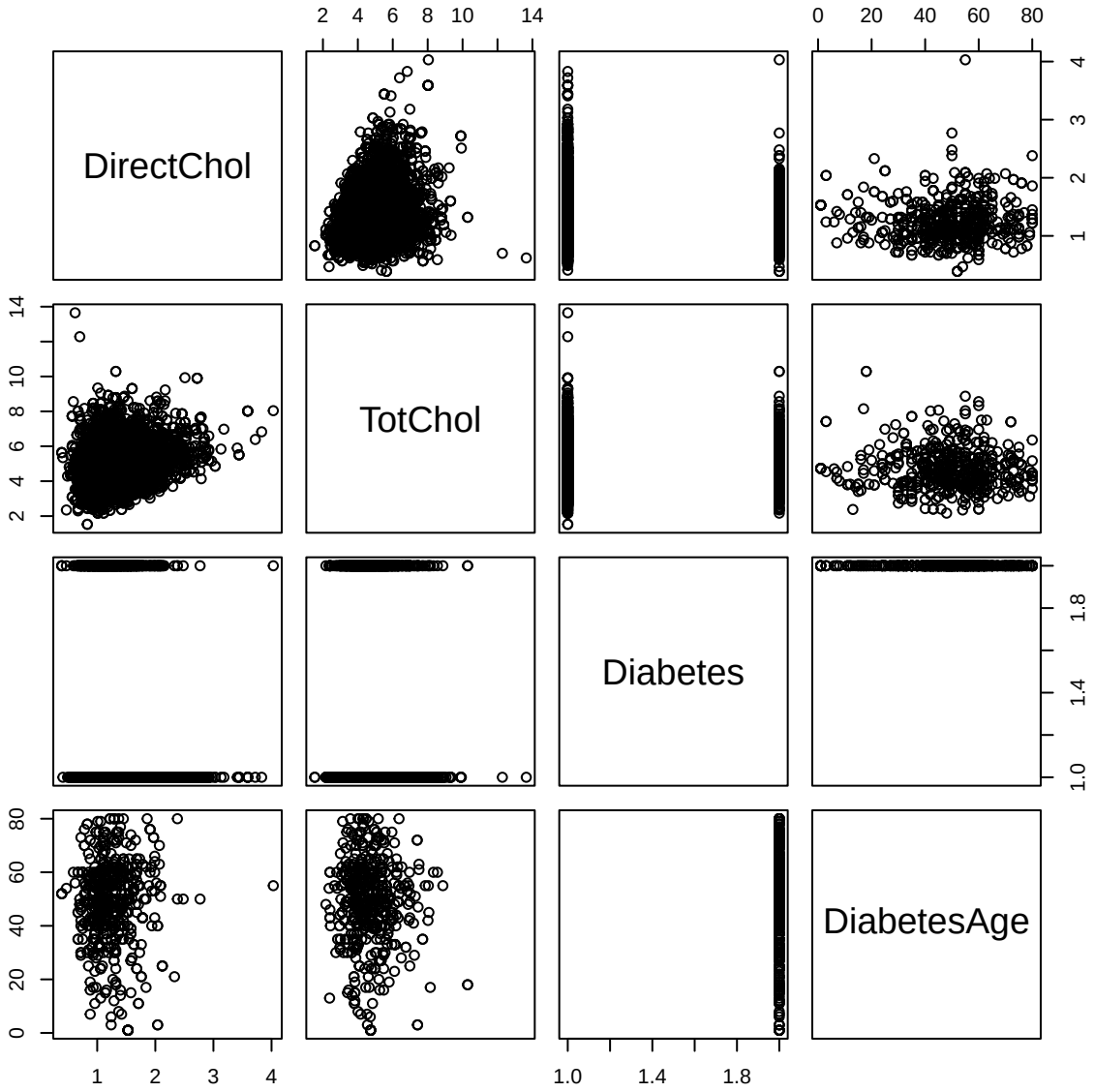
Lifestyle related variables:

```
to_test = c("PhysActive", "PhysActiveDays", "TVHrsDay", "CompHrsDay", "TVHrsDayChild", "CompHrsDayChild
pairs(subset(NHANES, select=to_test))
```



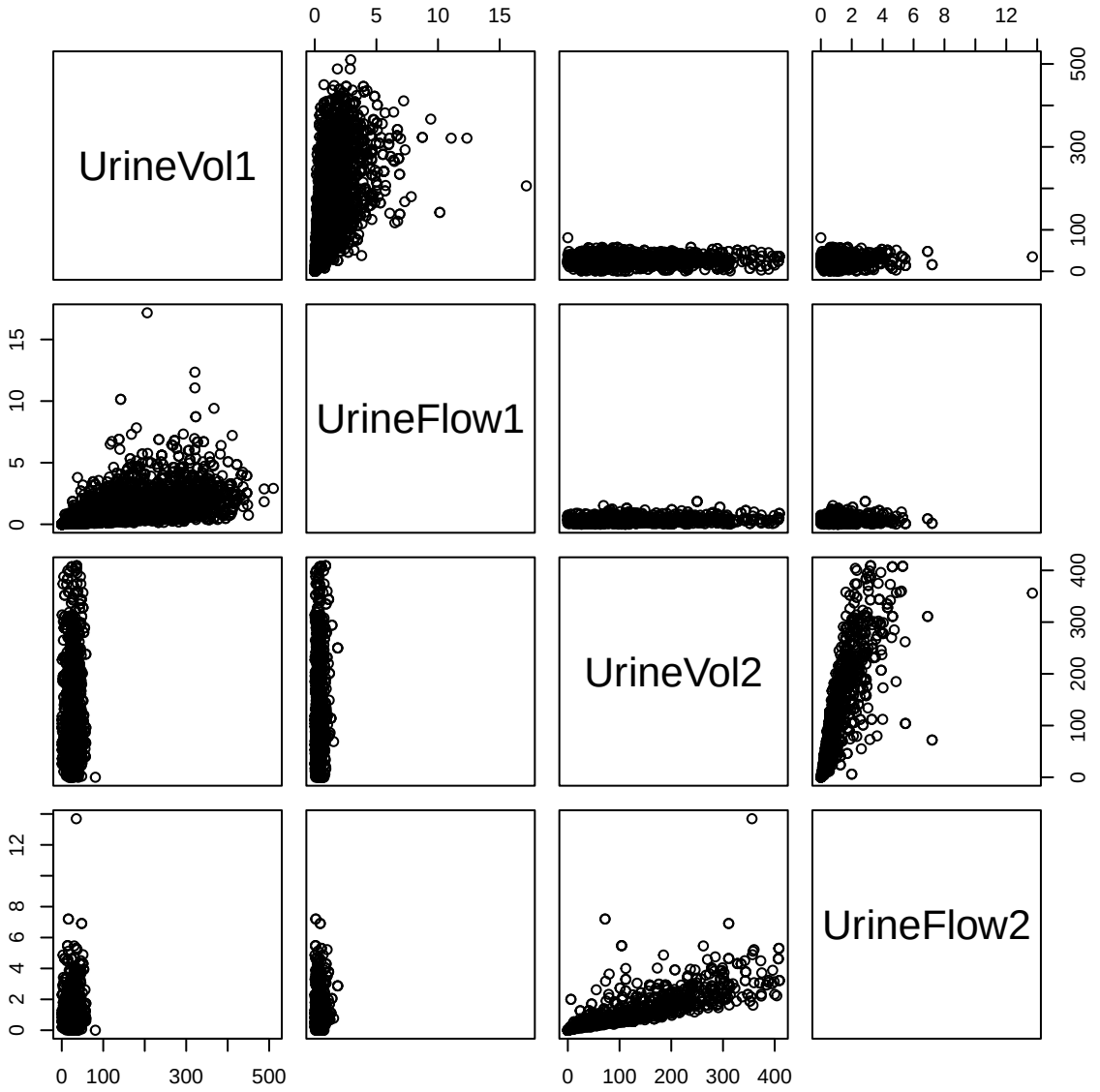
Cholesterol related variables:

```
to_test = c("DirectChol", "TotChol", "Diabetes", "DiabetesAge")
pairs(subset(NHANES, select = to_test))
```



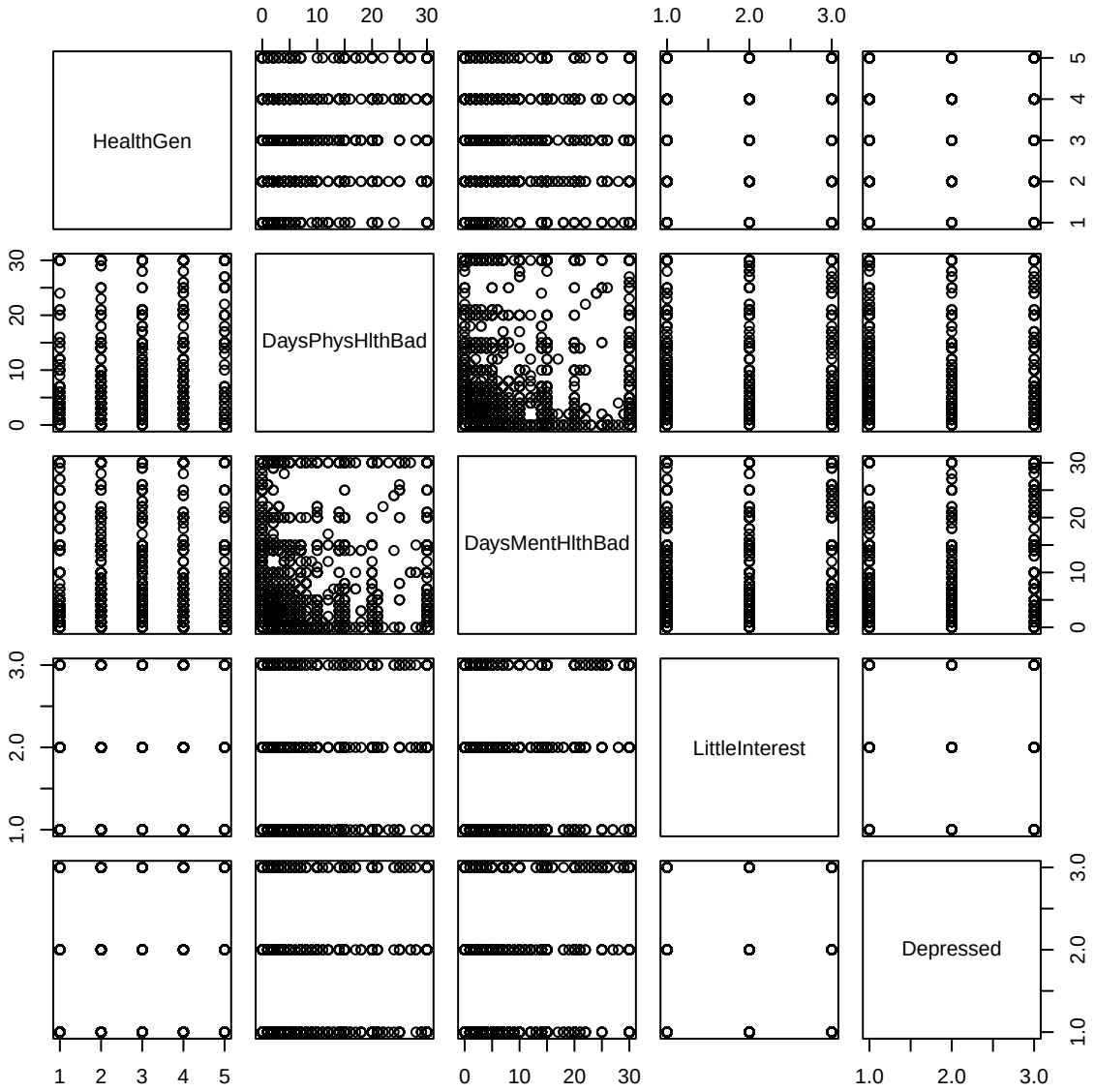
Urine related variables:

```
to_test = c("UrineVol1", "UrineFlow1", "UrineVol2", "UrineFlow2")
pairs(subset(NHANES, select = to_test))
```



Mental health related variables:

```
to_test = c("HealthGen", "DaysPhysHlthBad", "DaysMentHlthBad", "LittleInterest", "Depressed" )
pairs(subset(NHANES, select = to_test))
```


HealthGen
DaysPhysHlthBad
DaysMentHlthBad
LittleInterest
Depressed