Work sample Snapshot: Dealing with missing data and try different models

Jie Luo

3/14/2021

```
# Load libraries and the data
library(tidyverse)
library(tidymodels)
```

```
## Warning: package 'tidymodels' was built under R version 4.0.3
```

```
## Warning: package 'dials' was built under R version 4.0.3
```

```
## Warning: package 'infer' was built under R version 4.0.3
```

```
## Warning: package 'modeldata' was built under R version 4.0.3
```

```
## Warning: package 'rsample' was built under R version 4.0.3
```

```
## Warning: package 'tune' was built under R version 4.0.3
```

```
## Warning: package 'workflows' was built under R version 4.0.3
```

```
## Warning: package 'yardstick' was built under R version 4.0.3
```

```
library(NHANES)
```

```
## Warning: package 'NHANES' was built under R version 4.0.3
```

```
library(visdat)
```

```
## Warning: package 'visdat' was built under R version 4.0.3
```

```
data("NHANES")
```

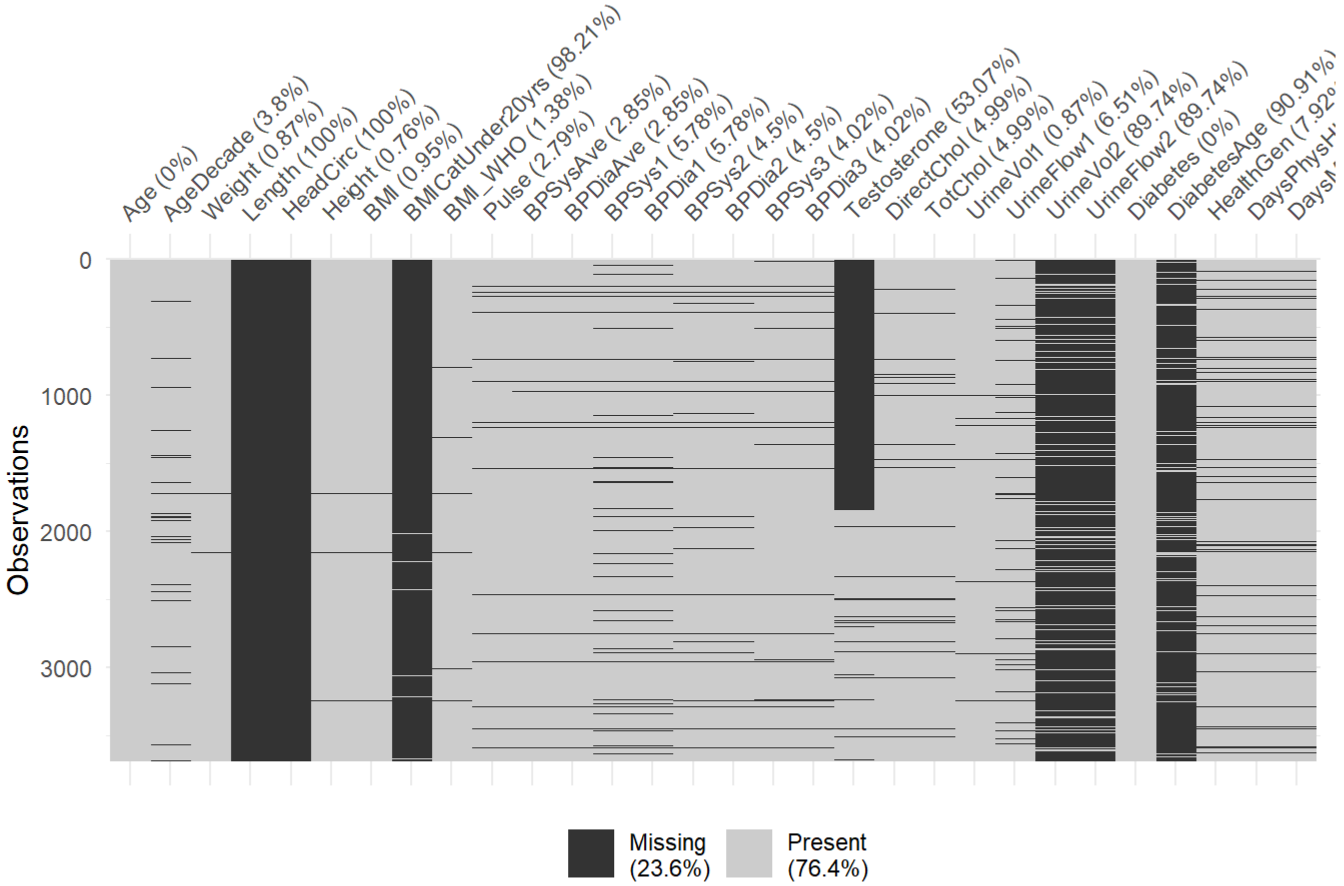
Part 1

```
# Removed observations that are under 18 years old
# Split the data into male and female group and we first focus on male

NHANES_male <- NHANES %>%
  filter(Age >= 18) %>%
  filter(Gender == "male")
```

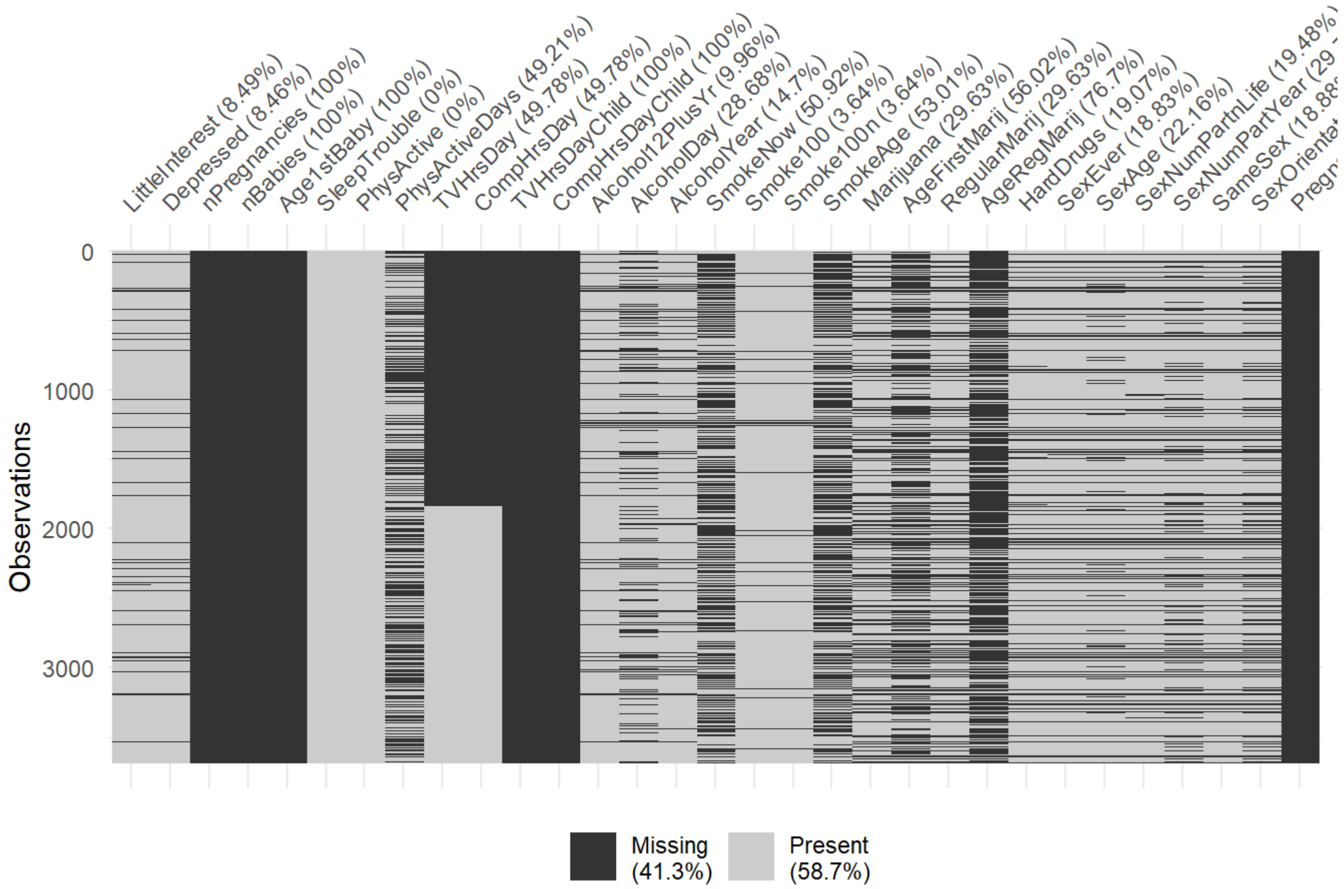
```
# Drop study variables, partial demographic variables, and sleepHrsNight
# The removed variables include: SurveyYr, ID, Gender, from agemonth to homeown, and SleepHrsNig
ht
drop <- c(1:3,6:16,50)
NHANES_male1 <- NHANES_male[, -drop]
head(NHANES_male1)
```

```
## # A tibble: 6 x 61
##     Age AgeDecade Weight Length HeadCirc Height   BMI BMICatUnder20yrs BMI_WHO
##   <int> <fct>      <dbl>  <dbl>    <dbl>  <dbl> <dbl> <fct>            <fct>
## 1    34 " 30-39"    87.4     NA       NA   165.  32.2 <NA>             30.0_p~
## 2    34 " 30-39"    87.4     NA       NA   165.  32.2 <NA>             30.0_p~
## 3    34 " 30-39"    87.4     NA       NA   165.  32.2 <NA>             30.0_p~
## 4    66 " 60-69"    68       NA       NA   170.  23.7 <NA>             18.5_t~
## 5    58 " 50-59"    78.4     NA       NA   182.  23.7 <NA>             18.5_t~
## 6    54 " 50-59"    74.7     NA       NA   169.  26.0 <NA>             25.0_t~
## # ... with 52 more variables: Pulse <int>, BPSysAve <int>, BPDiaAve <int>,
## #   BPSys1 <int>, BPDia1 <int>, BPSys2 <int>, BPDia2 <int>, BPSys3 <int>,
## #   BPDia3 <int>, Testosterone <dbl>, DirectChol <dbl>, TotChol <dbl>,
## #   UrineVol1 <int>, UrineFlow1 <dbl>, UrineVol2 <int>, UrineFlow2 <dbl>,
## #   Diabetes <fct>, DiabetesAge <int>, HealthGen <fct>, DaysPhysHlthBad <int>,
## #   DaysMentHlthBad <int>, LittleInterest <fct>, Depressed <fct>,
## #   nPregnancies <int>, nBabies <int>, Age1stBaby <int>, SleepTrouble <fct>,
## #   PhysActive <fct>, PhysActiveDays <int>, TVHrsDay <fct>, CompHrsDay <fct>,
## #   TVHrsDayChild <int>, CompHrsDayChild <int>, Alcohol12PlusYr <fct>,
## #   AlcoholDay <int>, AlcoholYear <int>, SmokeNow <fct>, Smoke100 <fct>,
## #   Smoke100n <fct>, SmokeAge <int>, Marijuana <fct>, AgeFirstMarij <int>,
## #   RegularMarij <fct>, AgeRegMarij <int>, HardDrugs <fct>, SexEver <fct>,
## #   SexAge <int>, SexNumPartnLife <int>, SexNumPartYear <int>, SameSex <fct>,
## #   SexOrientation <fct>, PregnantNow <fct>
```

```
# visualize the missing data
par(mfrow=c(1,2))
vis_miss(NHANES_male1[,1:30])
```



```
vis_miss(NHANES_male1[,31:61])
```



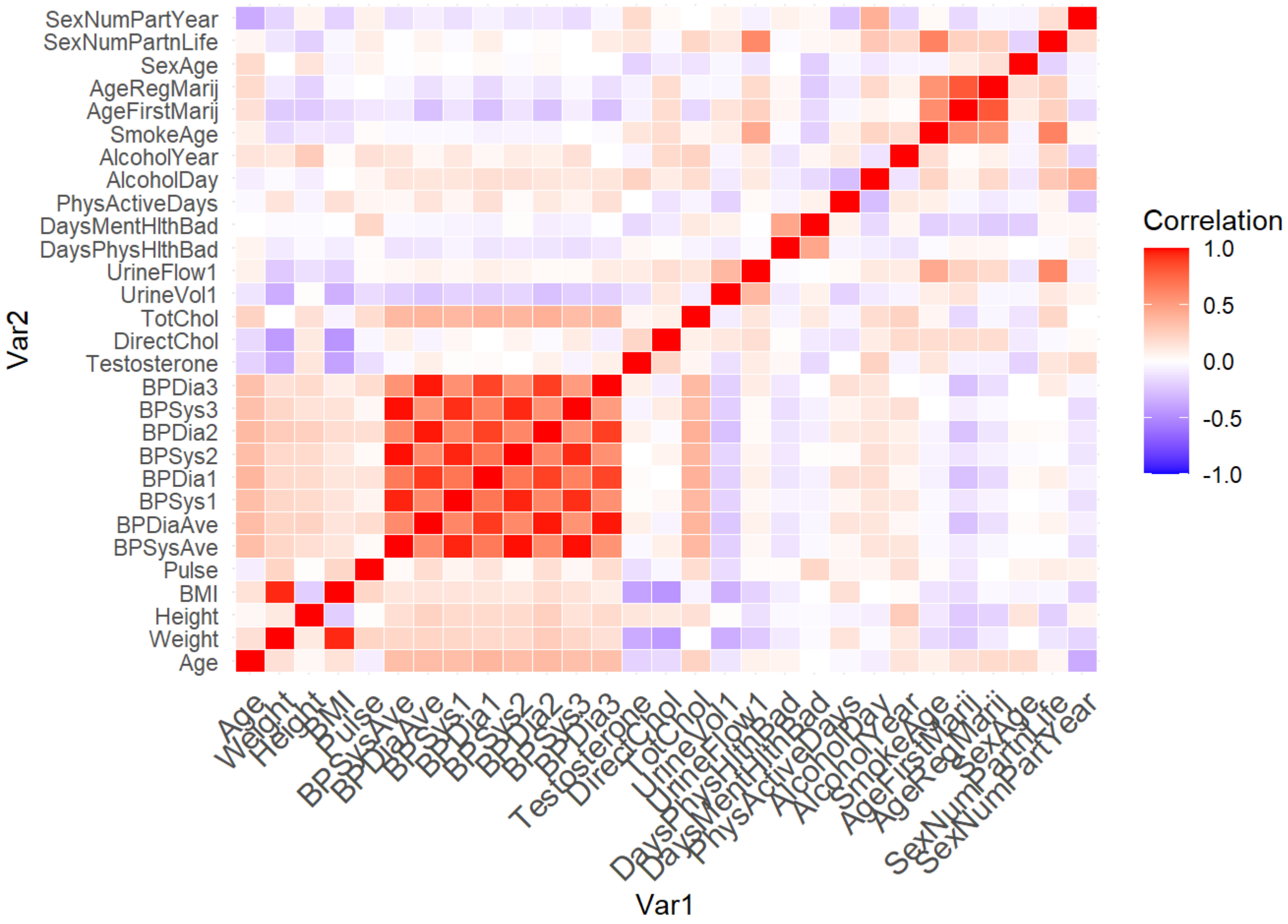
```
# Remove variables with over 90% missing data and variables that are not related to male partici
pants
# Removed varaibels includes: length, headcirc, BMICatUnder20yrs, UrineVol2, UrineFlow2,
# DiabetesAge, Npregnant, nBabies, Age1stBaby, TVHrdDayChild, CompHrsDaychild, and pregnant

drop2 <- c(4,5,8,24,25,27,33:35,41,42,61)
NHANES_male2 <- NHANES_male1[, -drop2]
head(NHANES_male2)
```

```
## # A tibble: 6 x 49
##     Age AgeDecade Weight Height   BMI BMI_WHO Pulse BPSysAve BPDiaAve BPSys1
##   <int> <fct>      <dbl>  <dbl> <dbl> <fct>   <int>    <int>    <int>  <int>
## 1    34 " 30-39"    87.4   165.  32.2 30.0_p~    70      113       85    114
## 2    34 " 30-39"    87.4   165.  32.2 30.0_p~    70      113       85    114
## 3    34 " 30-39"    87.4   165.  32.2 30.0_p~    70      113       85    114
## 4    66 " 60-69"    68     170.  23.7 18.5_t~    60      111       63    124
## 5    58 " 50-59"    78.4   182.  23.7 18.5_t~    62      104       74    108
## 6    54 " 50-59"    74.7   169.  26.0 25.0_t~    76      134       85    136
## # ... with 39 more variables: BPDia1 <int>, BPSys2 <int>, BPDia2 <int>,
## #   BPSys3 <int>, BPDia3 <int>, Testosterone <dbl>, DirectChol <dbl>,
## #   TotChol <dbl>, UrineVol1 <int>, UrineFlow1 <dbl>, Diabetes <fct>,
## #   HealthGen <fct>, DaysPhysHlthBad <int>, DaysMentHlthBad <int>,
## #   LittleInterest <fct>, Depressed <fct>, SleepTrouble <fct>,
## #   PhysActive <fct>, PhysActiveDays <int>, TVHrsDay <fct>, CompHrsDay <fct>,
## #   Alcohol12PlusYr <fct>, AlcoholDay <int>, AlcoholYear <int>, SmokeNow <fct>,
## #   Smoke100 <fct>, Smoke100n <fct>, SmokeAge <int>, Marijuana <fct>,
## #   AgeFirstMarij <int>, RegularMarij <fct>, AgeRegMarij <int>,
## #   HardDrugs <fct>, SexEver <fct>, SexAge <int>, SexNumPartnLife <int>,
## #   SexNumPartYear <int>, SameSex <fct>, SexOrientation <fct>
```

```
# Run a correlation matrix for numerical data

cor1 <- round(cor(NHANES_male2 %>% select(where(is.numeric)) %>% na.omit()),2)
```

```
# Visualize the results in the heat map

library(reshape2)
mel_cor1 <- melt(cor1)
```

```
hm <- ggplot(data = mel_cor1, aes(x=Var1, y=Var2, fill=value))+
        geom_tile(color = "white")+
        scale_fill_gradient2(low = "blue", high = "red", mid = "white",
          midpoint = 0, limit = c(-1,1), space = "Lab",
          name="Correlation") +
        theme_minimal()+ # minimal theme
        theme(axis.text.x = element_text(angle = 45, vjust = 1,
        size = 12, hjust = 1))

print(hm)
```



```
drop3 <- c(2:4,6,10:15,38,40,42)

NHANES_male3 <- NHANES_male2[,-drop3]

head(NHANES_male3)
```

```
## # A tibble: 6 x 36
##     Age   BMI Pulse BPSysAve BPDiaAve Testosterone DirectChol TotChol UrineVol1
##   <int> <dbl> <int>    <int>    <int>        <dbl>      <dbl>   <dbl>     <int>
## 1    34  32.2    70      113       85           NA       1.29    3.49       352
## 2    34  32.2    70      113       85           NA       1.29    3.49       352
## 3    34  32.2    70      113       85           NA       1.29    3.49       352
## 4    66  23.7    60      111       63           NA       0.67    4.99       113
## 5    58  23.7    62      104       74           NA       0.96    4.24       163
## 6    54  26.0    76      134       85           NA       1.16    6.41       215
## # ... with 27 more variables: UrineFlow1 <dbl>, Diabetes <fct>,
## #   HealthGen <fct>, DaysPhysHlthBad <int>, DaysMentHlthBad <int>,
## #   LittleInterest <fct>, Depressed <fct>, SleepTrouble <fct>,
## #   PhysActive <fct>, PhysActiveDays <int>, TVHrsDay <fct>, CompHrsDay <fct>,
## #   Alcohol12PlusYr <fct>, AlcoholDay <int>, AlcoholYear <int>, SmokeNow <fct>,
## #   Smoke100 <fct>, Smoke100n <fct>, Marijuana <fct>, RegularMarij <fct>,
## #   HardDrugs <fct>, SexEver <fct>, SexAge <int>, SexNumPartnLife <int>,
## #   SexNumPartYear <int>, SameSex <fct>, SexOrientation <fct>
```

From the heat map, we can clearly see that some of the numerical variables highly correlated with one of another. \ For example, BMI is higher correlated with weigh and height (BMI = weight/height^2).Thus, I remove height and weight\ Also, several measures of blood pressures are also correlated. I only consider keeping BPDiaAve and BPSDiaAve. \ On the other hand, the age of first use of smoke is highly correlated with the age of regularly using marijuana and the age of first use of marijuana. Consider big chunks of missing data for the three variables(53.1%, 56.02%, 76.7%), here I chose to removed them at the first place because I think others variables such as smokeNow or RegularMarij are enough to represent the situations of the participants on the habit of using drugs or smoke. If interested, I will include in the future step.\

Part 2

SmokeNow might be the variable that should be fixed. According to the data description, the SmokeNow is conditional response when the answer to Smoke100 is yes. In other words, SmokeNow will be NA if a participant has not smoked 100 or more cigarettes in their life. Hence, if the answer for smoke100 is no, then SmokeNow should also be no since they have not even smoked 100 or more cigarettes.

```
NHANES_male4 <- NHANES_male3 %>%
  replace_na(list(SmokeNow = "No"))
```

```
library(VIM)
```

```
## Warning: package 'VIM' was built under R version 4.0.3
```

```
library(mice)
```

```
## Warning: package 'mice' was built under R version 4.0.3
```

```
library(naniar)
```

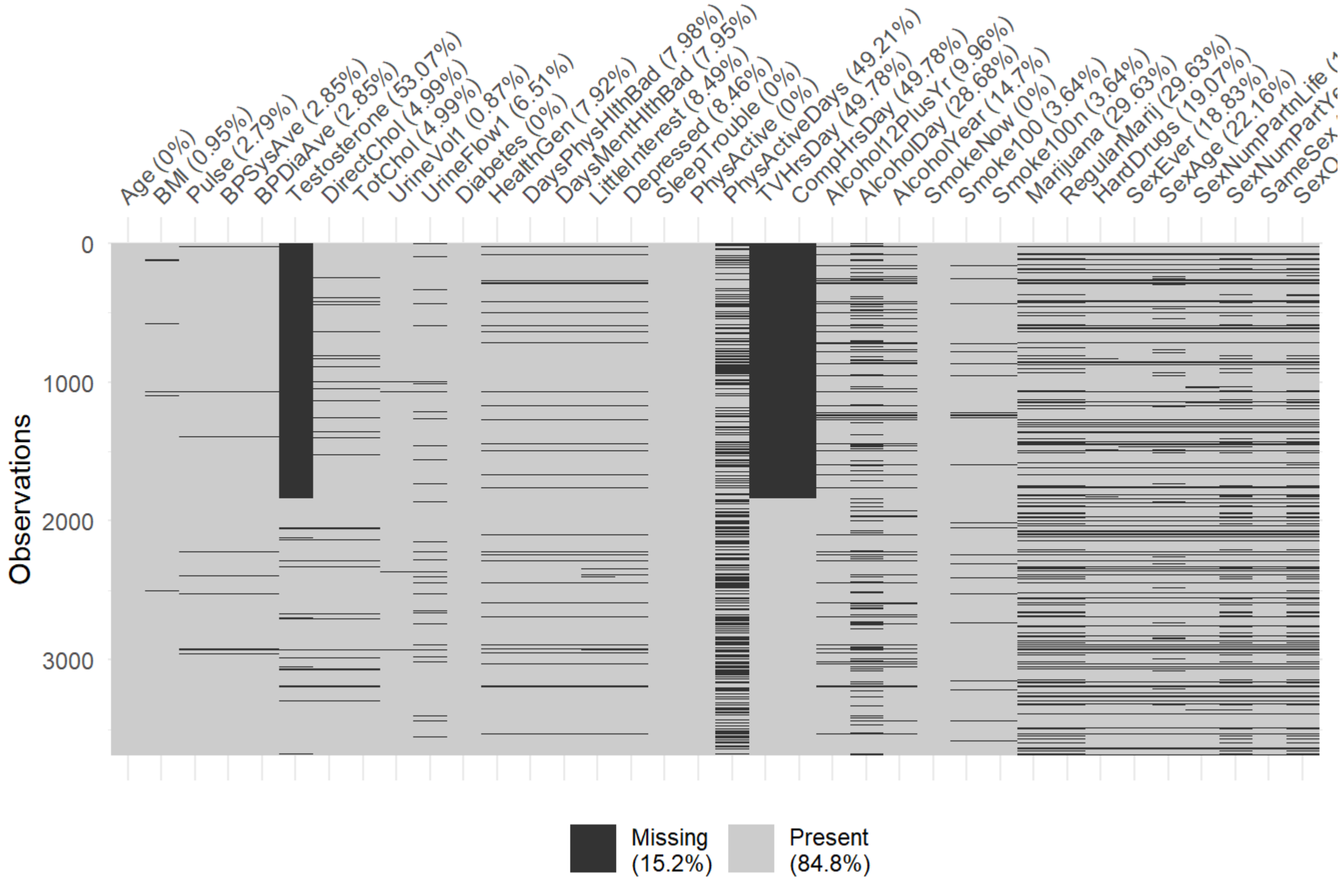
```
## Warning: package 'naniar' was built under R version 4.0.3
```

```
library(simputation)
```

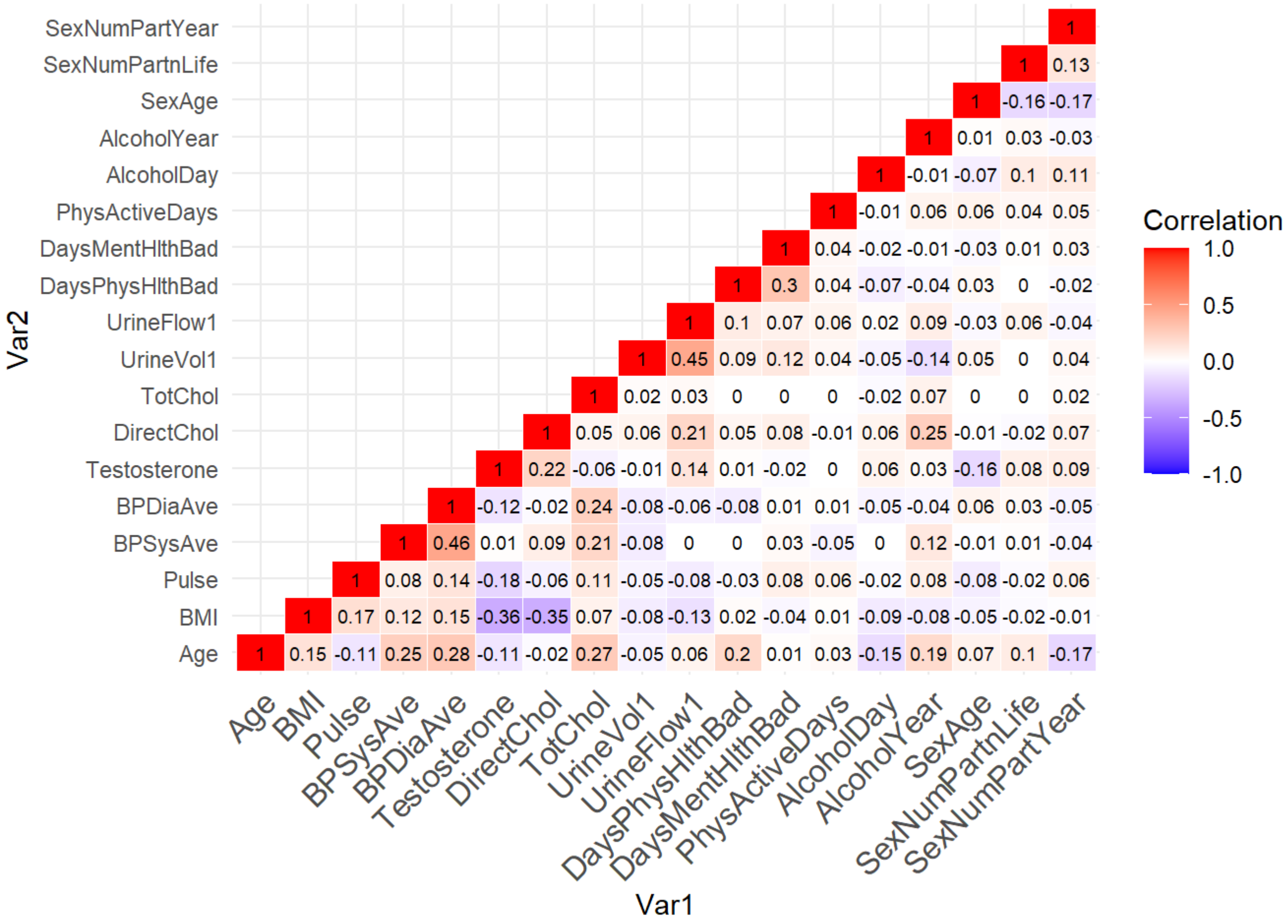
```
## Warning: package 'simputation' was built under R version 4.0.3
```

Part 3

```
# Using the same missing value data shown as part 1
vis_miss(NHANES_male4)
```



```
# create the correlation matrix for numerical data for our final selected dataset

cor2 <- round(cor(NHANES_male4 %>% select(where(is.numeric)) %>% na.omit()),2)
cor2[upper.tri(cor2)] =NA
mel_cor2 <- melt(cor2, na.rm=TRUE)

hm2 <- ggplot(data = mel_cor2, aes(x=Var1, y=Var2, fill=value))+
        geom_tile(color = "white")+
        scale_fill_gradient2(low = "blue", high = "red", mid = "white",
          midpoint = 0, limit = c(-1,1), space = "Lab",
          name="Correlation") +
        theme_minimal()+
        theme(axis.text.x = element_text(angle = 45, vjust = 1,
        size = 12, hjust = 1)) +
        geom_text(aes(x=Var1, y=Var2,label=value),color="black", size=2.5)

print(hm2)
```



Var2	Age	BMI	Pulse	BPSysAve	BPDiaAve	Testosterone	DirectChol	TotChol	UrineVol1	UrineFlow1	DaysPhysHlthBad	DaysMentHlthBad	PhysActiveDays	AlcoholDay	AlcoholYear	SexAge	SexNumPartnLife	SexNumPartYear
SexNumPartYear																		1
SexNumPartnLife																	1	0.13
SexAge																1	-0.16	-0.17
AlcoholYear															1	0.01	0.03	-0.03
AlcoholDay														1	-0.01	-0.07	0.1	0.11
PhysActiveDays													1	-0.01	0.06	0.06	0.04	0.05
DaysMentHlthBad												1	0.04	-0.02	-0.01	-0.03	0.01	0.03
DaysPhysHlthBad											1	0.3	0.04	-0.07	-0.04	0.03	0	-0.02
UrineFlow1										1	0.1	0.07	0.06	0.02	0.09	-0.03	0.06	-0.04
UrineVol1									1	0.45	0.09	0.12	0.04	-0.05	-0.14	0.05	0	0.04
TotChol								1	0.02	0.03	0	0	0	-0.02	0.07	0	0	0.02
DirectChol							1	0.05	0.06	0.21	0.05	0.08	-0.01	0.06	0.25	-0.01	-0.02	0.07
Testosterone						1	0.22	-0.06	-0.01	0.14	0.01	-0.02	0	0.06	0.03	-0.16	0.08	0.09
BPDiaAve					1	-0.12	-0.02	0.24	-0.08	-0.06	-0.08	0.01	0.01	-0.05	-0.04	0.06	0.03	-0.05
BPSysAve				1	0.46	0.01	0.09	0.21	-0.08	0	0	0.03	-0.05	0	0.12	-0.01	0.01	-0.04
Pulse			1	0.08	0.14	-0.18	-0.06	0.11	-0.05	-0.08	-0.03	0.08	0.06	-0.02	0.08	-0.08	-0.02	0.06
BMI		1	0.17	0.12	0.15	-0.36	-0.35	0.07	-0.08	-0.13	0.02	-0.04	0.01	-0.09	-0.08	-0.05	-0.02	-0.01
Age	1	0.15	-0.11	0.25	0.28	-0.11	-0.02	0.27	-0.05	0.06	0.2	0.01	0.03	-0.15	0.19	0.07	0.1	-0.17

Part 4

To deal with missing data, first, I notice that the majority of missting data come from Testosterone, TVHrsDays, and CompHrsDay. Among there three variables, Only Testosterone is numerical. I also see that it is correlated with Age, BMI, DirectChol, and Pulse. So here I hypothsize that there is a linear relationship among them and I want to run a linear model to test it out.

```
lm1 <- lm(Testosterone~ BMI + DirectChol + Pulse + Age, data=NHANES_male4)
summary(lm1)
```

```
##
## Call:
## lm(formula = Testosterone ~ BMI + DirectChol + Pulse + Age, data = NHANES_male4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -473.67 -110.49  -15.03   90.46 1401.87
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 799.0073    39.0452  20.464  < 2e-16 ***
## BMI         -10.1527     0.7598 -13.362  < 2e-16 ***
## DirectChol   44.5593    13.6075   3.275  0.00108 **
## Pulse        -1.4781     0.3398  -4.350 1.44e-05 ***
## Age          -1.0319     0.2408  -4.286 1.92e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 163.1 on 1680 degrees of freedom
##   (2001 observations deleted due to missingness)
## Multiple R-squared:  0.1654, Adjusted R-squared:  0.1634
## F-statistic: 83.25 on 4 and 1680 DF,  p-value: < 2.2e-16
```

The linear model suggests that all of the selected variables are statistically significant at determining the Testosterone level. Consider the run time of using mice to impute all of the data, here I will use linear model to impute the Testoterone, and use the mice to impute the rest.

```
NHANES_male5 <- NHANES_male4 %>%
  impute_lm(Testosterone~ BMI + DirectChol + Pulse + Age)
```

Next, I will use mice to impute the rest of the missing data

```
NHANES_mice <- mice(NHANES_male5, seed=256, print=FALSE)
```

```
## Warning: Number of logged events: 26
```

Part 5

```
# Extract the imputed data
mice_imputed <- complete(NHANES_mice)
```

```
# splitting the data into train and test set
set.seed(123)

num<-nrow(mice_imputed)

split <- mice_imputed %>%
  initial_split(prop=0.8)

train <- split %>%
  training()

test <- split %>%
  testing()

list(train, test) %>%
  map_int(nrow)
```

```
## [1] 2949  737
```

```
# fit a logistice regression

library(yardstick)

fit1 <- logistic_reg(mode="classification") %>%
  set_engine("glm") %>%
  fit(SleepTrouble ~ Testosterone + DaysMentHlthBad + PhysActiveDays, data=train)

pred1_train <- train %>%
  select(SleepTrouble) %>%
  bind_cols(
    predict(fit1, new_data = train, type = "class")
  ) %>%
  rename(sleep_log = .pred_class)

pred1_test <- test %>%
  select(SleepTrouble) %>%
  bind_cols(
    predict(fit1, new_data = test, type = "class")
  ) %>%
  rename(sleep_log = .pred_class)
```

```
# Creating a confusion matrix to visualize the result

confusion_log1 <- pred1_train %>%
  conf_mat(truth = SleepTrouble, estimate = sleep_log)

confusion_log2 <- pred1_test %>%
  conf_mat(truth = SleepTrouble, estimate = sleep_log)

confusion_log1
```

```
##           Truth
## Prediction   No  Yes
##        No  2240  552
##        Yes   70   87
```

```
confusion_log2
```

```
##           Truth
## Prediction  No Yes
##        No  556 140
##        Yes  18  23
```

```
accuracy(pred1_train, SleepTrouble, sleep_log)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.789
```

```
accuracy(pred1_test, SleepTrouble, sleep_log)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.786
```

The accuracy on train set is 78.91% and it is 78.56% on test set. Overall the result is not too bad considering only three variables are chosen. Next, we will look for a better model.

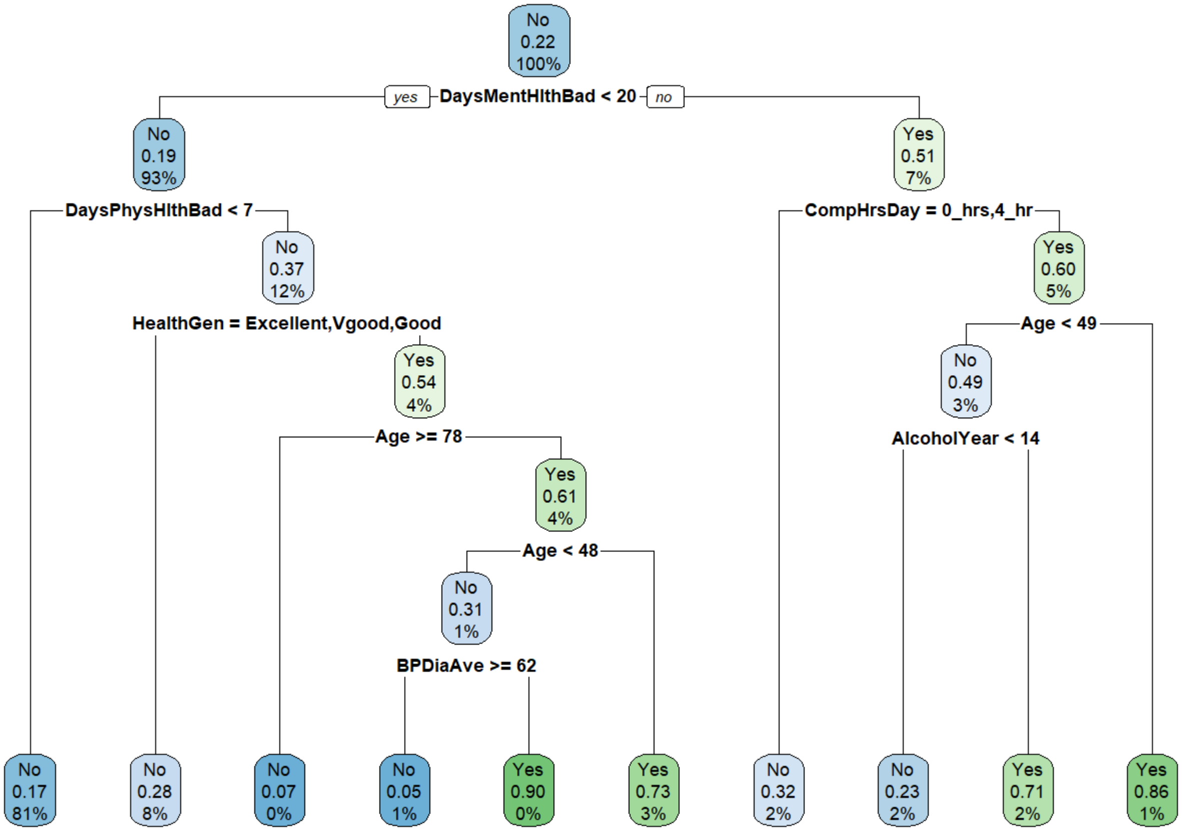
Part 6

The second model I choose is the decision tree.

```
library(rpart)
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.0.3
```

```
tree.fit <- rpart(SleepTrouble ~ ., data = train)
rpart.plot(tree.fit)
```



```
# predicting using train set

pred2.train <- predict(tree.fit, train, type="class")

table_mat1 <- table(pred2.train,train$SleepTrouble)
table_mat1
```

```
##
## pred2.train   No  Yes
##          No  2266  499
##          Yes   44  140
```

The accuracy is 81.59% on training set

```
# predicting using train set

pred2.test <- predict(tree.fit, test, type="class")

table_mat2 <- table(pred2.test,test$SleepTrouble)
table_mat2
```

```
##
## pred2.test  No Yes
##        No  553 136
##        Yes  21  27
```

The accuracy is (553+27)/737 = 78.70% on test set

Part 7

I prefer the decision tree. First of all, the decision tree model gives better accuracy for both train set and test set. Second, the variables that decision tree has chosen make more sense than the three variables I choose for the logistic model. Specifically, it select Age, HealthGen, and DaysPhyhethbad, which I did not consider before.

Part 8

1. I chose to build different model for male and female, and the previous analysis is for male only. If given more time, I will try to fit another model for female.
2. For the decision tree model, we see that there is fairly big difference between the accuracy of test set and train set. One thing I will consider is to tune the hyper-parameter, which might include tuning the maximum depth and number of sample a leaf node should have, etc.
3. For logistic regression, I will consider refit the model with other variables.