CUNY 621, Discussion 5

Eric Hirsch

3/1/2022

```
library(tidyverse)
library(EHData)
library(faraway)
data(kanga, package="faraway")
library(tidyverse)
library(gridExtra)
library(MASS)
```

Distinguishing between MARS and MCARS

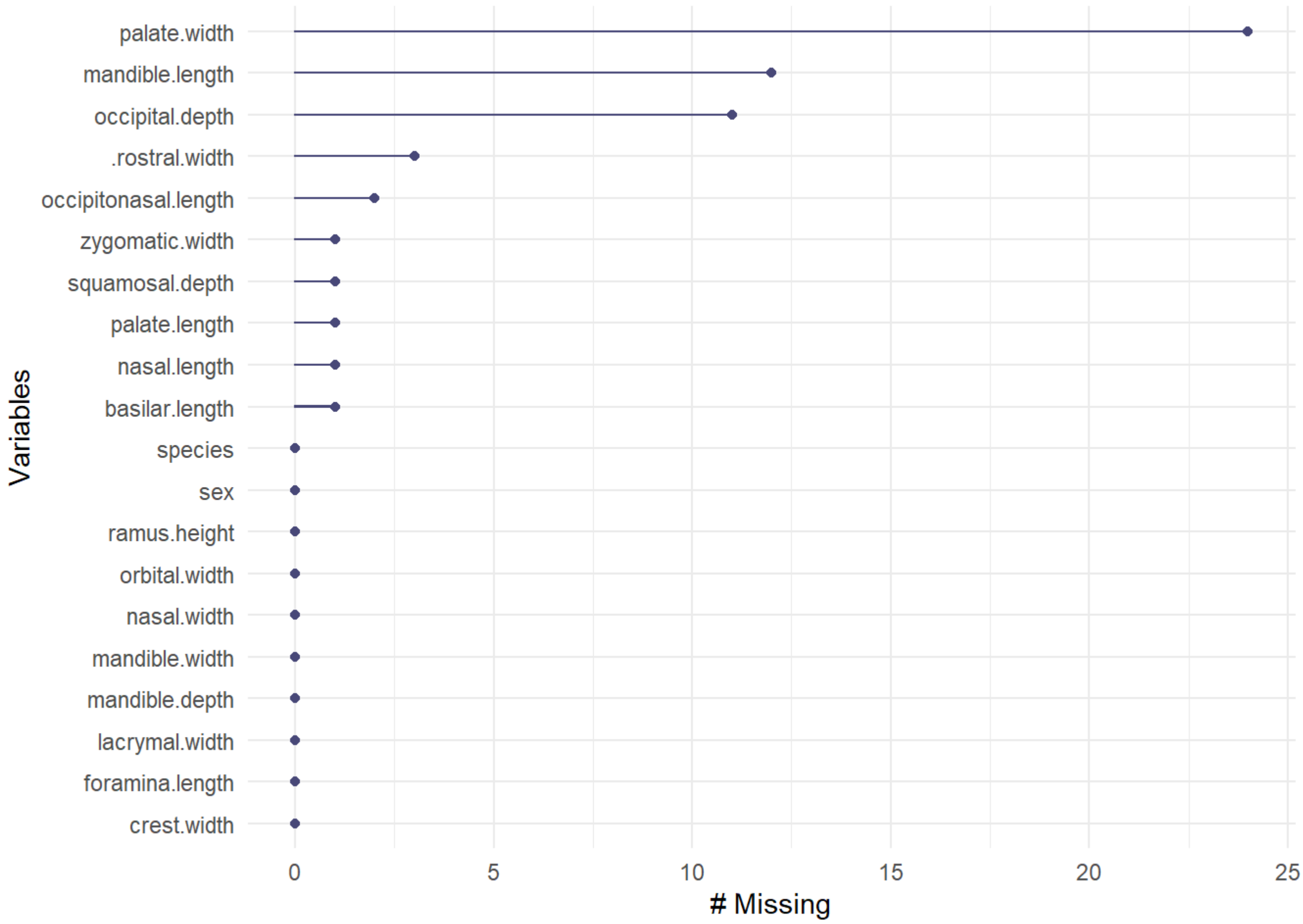
We have more options when our missing data is completely random than when it’s just plain random. How can we tell the difference? Here are a few ways I can think of to test the non-complete randomness of missings - but I would like to know What are the best practices?

I will be using the kangaroo skull dataset from LMR (kanga) to investigate these questions, focusing particularly on palate.width because it has the most missing values.

Here we see the columns in the dataset and their missing values. In all, 31% of the records have missing values so we are unlikely to discard all recoprds with missing, but it is still useful to know how non-random the randomness is.

```
data(kanga, package="faraway")
library(dplyr)

qqq <- EHData::EHSummarize_MissingValues(kanga)
print(qqq[[1]])
```



Here are some methods I’ve considered:

1. Run a regression with the data mean-imputed and and compare with a regression with the na’s omitted. There should be some difference because the variance in the imputed database is lower - but is the difference significant? In our example, the difference between the r-squared for mean imputed, median imputed and na omitted is relatively small:

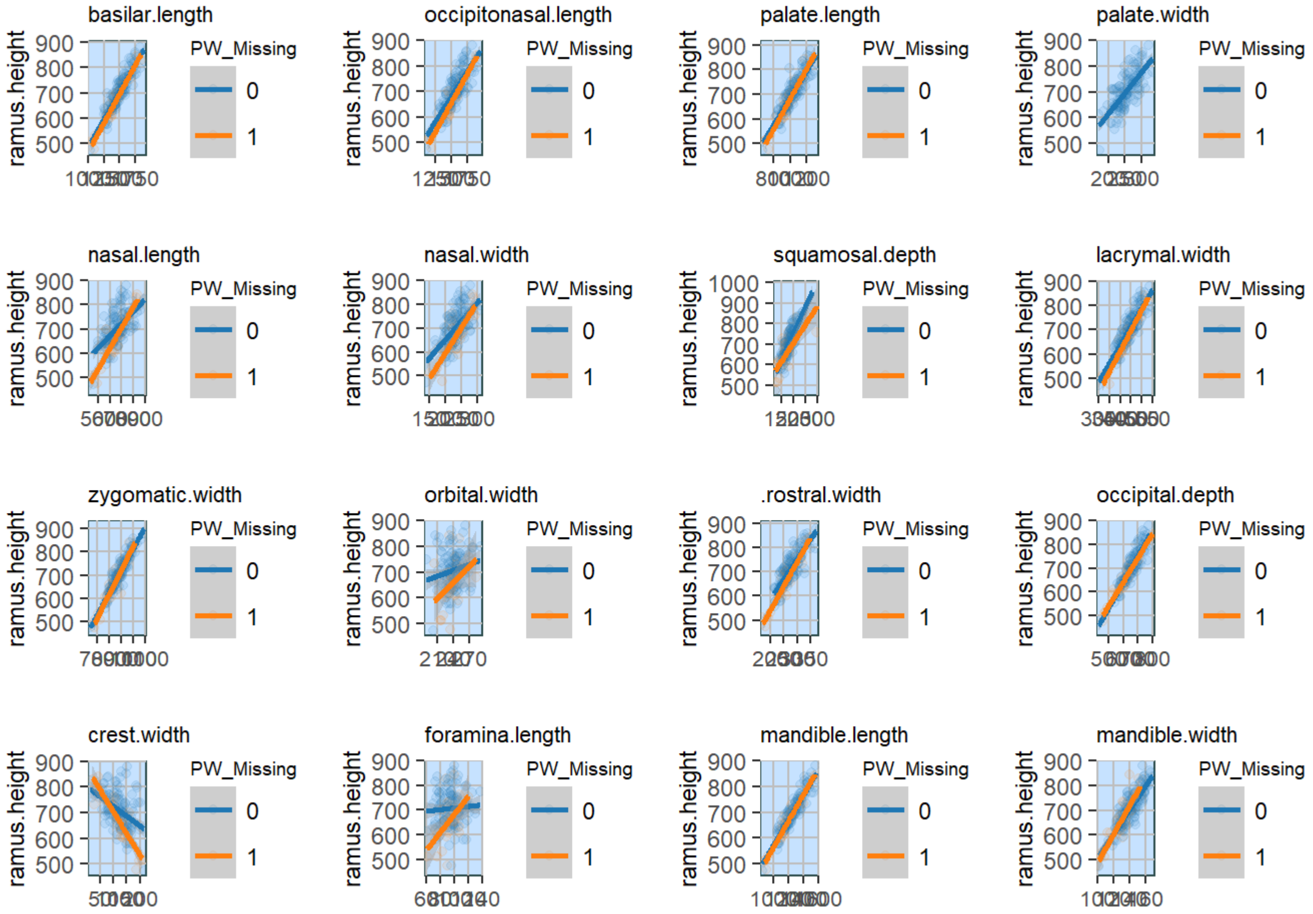
```
dfImputed <- EHPrepare_MissingValues_Imputation(kanga, "ramus.height")
```

```
## [1] "type:" "mean"
## [1] "r2mean:" "0.9492"
## [1] "r2median:" "0.9491"
## [1] "r2omit" "0.9427"
```

2. Create flags for missing values and look for interactions. If there are significant interactions we might discard the column but retain the flag. Here we test for interactions with one variable and find only very minor interactions:

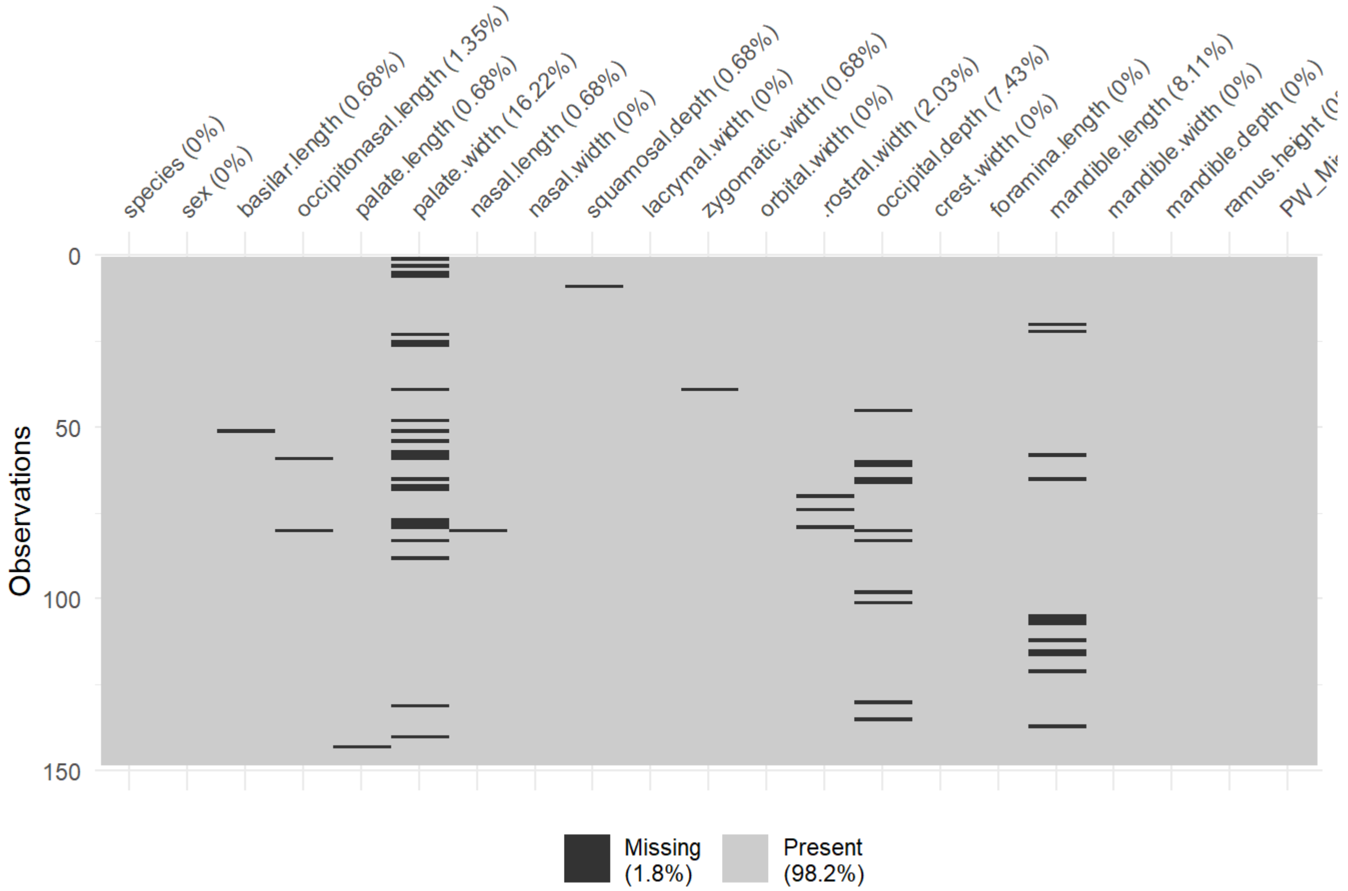
```
kanga2 <- kanga %>%
  dplyr::mutate(PW_Missing = ifelse(is.na(palate.width), 1, 0))

q <- EHData::EHExplore_Interactions_Scatterplots(kanga2, "ramus.height", "PW_Missing")
grid.arrange(grobs=q[c(1:16)], ncol=4)
```



3. Look for overlap among missing values. Here we see little overlap.

```
qqq <- EHData::EHSummarize_MissingValues(kanga2)
print(qqq[[2]])
```



```
kanga3 <- na.omit(kanga2)
```

I would onclude from this anlaysis that the missing values are likely to be MCARS. However, I’d like to know how experts approach it.