CIS635-Project

Tyler Reed

4/19/2021

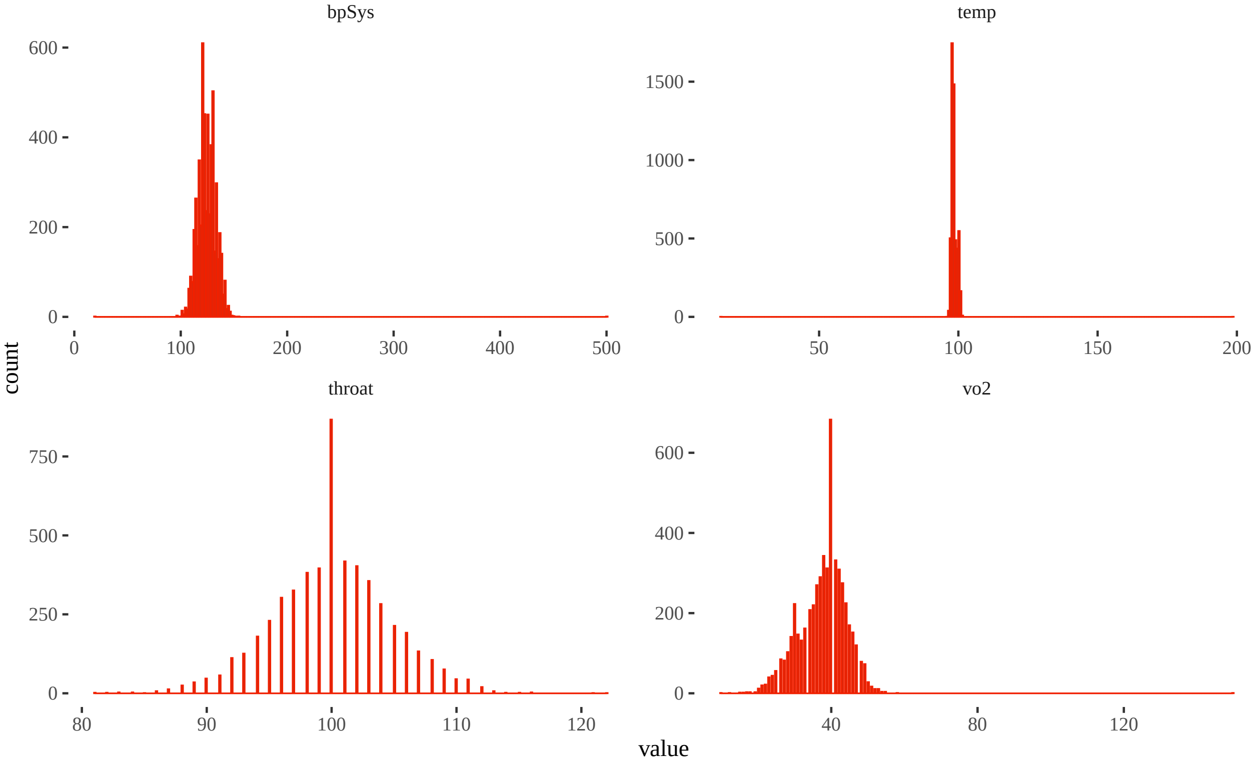
```
knitr::opts_chunk$set(error = TRUE, fig.width = 12, fig.asp = 0.618)
```

```
library(tidyverse)
library(knitr)
library(e1071)
library(rpart)
library(neuralnet)
library(hrbrthemes)
library(readr)
library(purrr)
library(ggthemes)

testA <- read.table("data/dataTestA.txt", header = TRUE)
testB <- read.table("data/dataTestB.txt", header = TRUE)
trainA <- read.table("data/dataTrainA.txt", header = TRUE)
trainB <- read.table("data/dataTrainB.txt", header = TRUE)
```

```
trainA <- as_tibble(trainA)

# Calculate summary statistics and produce visuals to check for outliers/noise/NAs
summary(trainA)
```

```
##        id            temp            bpSys            vo2             throat
##  Min.   :   0   Min.   : 15.00   Min.   : 20.0   Min.   : 10.00   Min.   : 81
##  1st Qu.:1673   1st Qu.: 97.79   1st Qu.:119.0   1st Qu.: 34.00   1st Qu.: 97
##  Median :3352   Median : 98.19   Median :124.0   Median : 39.00   Median :100
##  Mean   :3376   Mean   : 98.47   Mean   :124.6   Mean   : 37.76   Mean   :100
##  3rd Qu.:5084   3rd Qu.: 98.93   3rd Qu.:130.0   3rd Qu.: 42.00   3rd Qu.:103
##  Max.   :6780   Max.   :198.83   Max.   :501.0   Max.   :150.00   Max.   :122
##                 NA's   :1        NA's   :1       NA's   :2        NA's   :1
##      atRisk
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4652
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

```
trainA %>%
  mutate(atRisk = as_factor(atRisk),
         id = as_factor(id)) %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    theme_tufte(base_size = 16) +
    geom_histogram(color = "#F02000", bins = 300)
```



```
# Test for duplicate records
length(unique(trainA$id)) == nrow(trainA)
```

```
## [1] TRUE
```

```
# Results

# id: looks good and no duplicates
# temp: 1 NA, and min and max troublesome, use average
# bbSys: 1 NA, and min and max troublesome, use average
# vo2: 2 NA, max troublesome
# throat: 1 NA, max troublesome
# atRisk: looks good
```

```
trainB <- as_tibble(trainB)

# Calculate summary statistics to check for outliers/noise/NAs
summary(trainB)
```

```
##        id           headA             bodyA            cough
##  Min.   :   0   Min.   : 0.000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:1673   1st Qu.: 3.000   1st Qu.:4.000   1st Qu.:0.0000
##  Median :3352   Median : 3.000   Median :4.000   Median :0.0000
##  Mean   :3376   Mean   : 3.461   Mean   :4.016   Mean   :0.3418
##  3rd Qu.:5084   3rd Qu.: 4.000   3rd Qu.:4.000   3rd Qu.:1.0000
##  Max.   :6780   Max.   :100.000   Max.   :7.000   Max.   :1.0000
##                 NA's   :1
##      runny            nausea          diarrhea         atRisk
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.000   Median :0.0000
##  Mean   :0.1986   Mean   :0.2367   Mean   :0.102   Mean   :0.4652
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :5.0000   Max.   :1.000   Max.   :1.0000
##  NA's   :1                         NA's   :1
```

```r
# Test for duplicate records
length(unique(trainB$id)) == nrow(trainB)
```

```
## [1] TRUE
```

```r
# Results

# id: looks good and no duplicates
# headA: 1 NA, max troublesome
# bodyA: looks good
# cough: looks good
# runny: 1 NA
# nausea: max is troublesome
# diarrhea: 1 NA
# atRisk: looks good
```