HDS502

Matthew Onimus

03Jul2022

Data Set Selection

I choose the Maternal Health Risk Data Set Data Set (https://archive.ics.uci.edu/ml/datasets/Maternal+Health+Risk+Data+Set). The data set contains 1014 observations with 7 variables:

- Age: Any ages in years when a women during pregnant.
- SystolicBP: Upper value of Blood Pressure in mmHg, another significant attribute during pregnancy.
- DiastolicBP: Lower value of Blood Pressure in mmHg, another significant attribute during pregnancy.
- BS: Blood glucose levels is in terms of a molar concentration, mmol/L.
- HeartRate: A normal resting heart rate in beats per minute.
- Risk Level: Predicted Risk Intensity Level during pregnancy considering the previous attribute.

R Set Up and Data Read

I will start by loading the in the packages to be used for the initial data analysis. I will mainly be working in the `tidyverse` ecosystem.

```
library(tidyverse) # used for reading data, data cleaning, visualizations
library(rlang) # fancy data masking functions
library(kableExtra) # used to create tables in the document
library(GGally) # used to create the pairs plot
library(cowplot) # used to combine plots together
library(rstatix) # used for a pairwise t test

sessionInfo()
```

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Linux Mint 20.3
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
```

```
##
## other attached packages:
##  [1] rstatix_0.7.0    cowplot_1.1.1    GGally_2.1.2     kableExtra_1.3.4
##  [5] rlang_1.0.2      forcats_0.5.0    stringr_1.4.0    dplyr_1.0.9
##  [9] purrr_0.3.4      readr_2.1.1      tidyr_1.2.0      tibble_3.1.6
## [13] ggplot2_3.3.6    tidyverse_1.3.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.8.3       svglite_2.0.0      lubridate_1.7.9.2  assertthat_0.2.1
##  [5] digest_0.6.27      utf8_1.2.2         R6_2.5.1           cellranger_1.1.0
##  [9] plyr_1.8.6         backports_1.2.1    reprex_0.3.0       evaluate_0.14
## [13] httr_1.4.2         pillar_1.6.4       curl_4.3.2         readxl_1.3.1
## [17] data.table_1.13.6  rstudioapi_0.13    car_3.0-10         rmarkdown_2.14
## [21] webshot_0.5.2      foreign_0.8-75     munsell_0.5.0      broom_0.8.0
## [25] compiler_3.6.3     modelr_0.1.8       xfun_0.30          pkgconfig_2.0.3
## [29] systemfonts_1.0.1  htmltools_0.5.2    tidyselect_1.1.2   rio_0.5.26
## [33] reshape_0.8.9      fansi_0.5.0        viridisLite_0.4.0  crayon_1.4.2
## [37] tzdb_0.2.0         dbplyr_2.0.0       withr_2.4.3        grid_3.6.3
## [41] jsonlite_1.7.2     gtable_0.3.0       lifecycle_1.0.1    DBI_1.1.0
## [45] magrittr_2.0.1     scales_1.1.1       zip_2.1.1          carData_3.0-4
## [49] cli_3.3.0          stringi_1.7.6      fs_1.5.0           xml2_1.3.2
## [53] ellipsis_0.3.2     generics_0.1.2     vctrs_0.4.1        openxlsx_4.2.3
## [57] RColorBrewer_1.1-2 tools_3.6.3        glue_1.6.2         hms_1.1.1
## [61] abind_1.4-5        fastmap_1.1.0      yaml_2.2.1         colorspace_2.0-1
## [65] rvest_0.3.6        knitr_1.33         haven_2.3.1
```

Next, we need to read the data in to R. It was provided in .csv format from UCI. I will change the outcome variable, 'Risk Level', to a factor as well.

```
maternalHealth <- read_csv("Maternal Health Risk Data Set.csv") %>%
  mutate(RiskLevel = factor(RiskLevel,
                            levels = c('low risk', 'mid risk', 'high risk')))

# kbl(maternalHealth) %>%
#   kable_styling()
```

Data Exploration

I will start my exploration by creating some summary stats for my data and then creating a few plots to visualize the distribution across the outcome variable.

```
summary(maternalHealth)
```

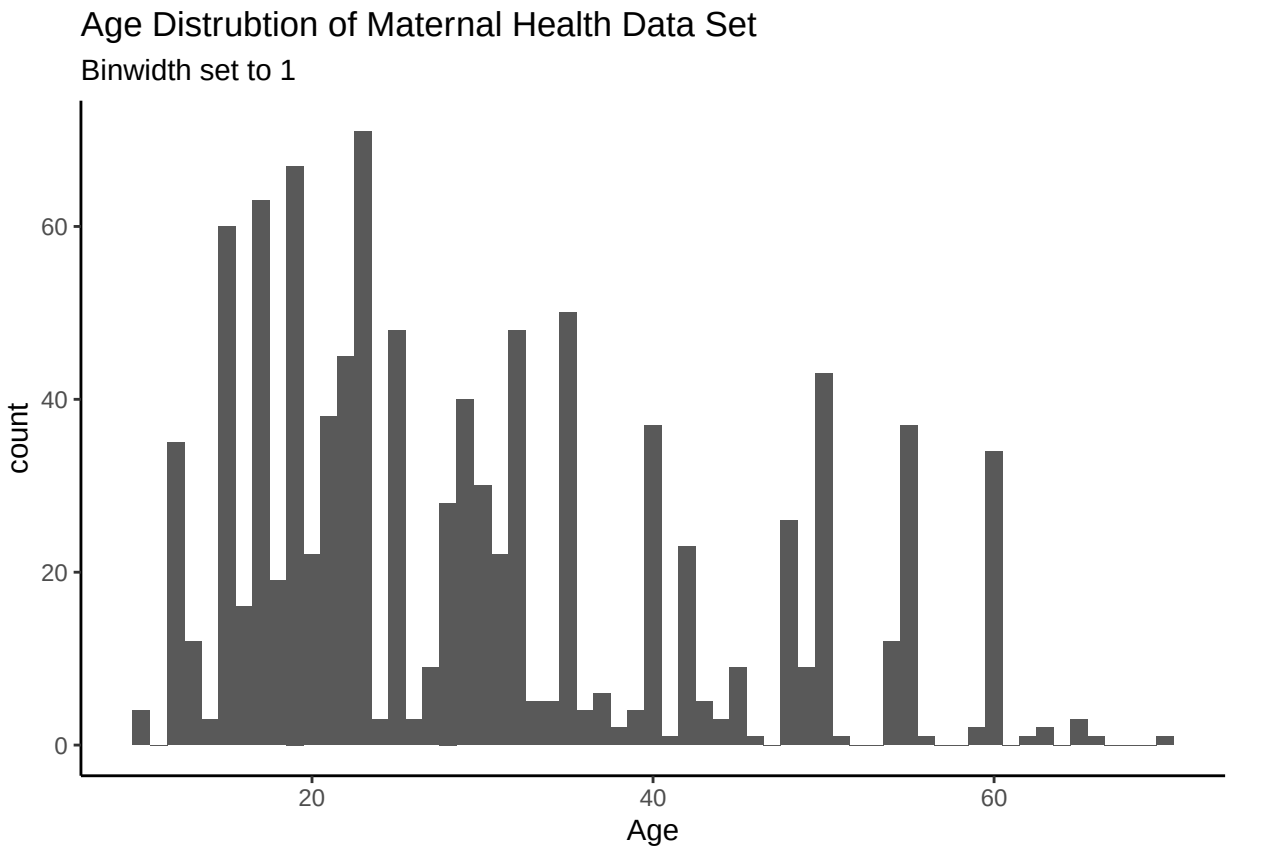
```
##       Age          SystolicBP     DiastolicBP          BS
##  Min.   :10.00   Min.   : 70.0   Min.   : 49.00   Min.   : 6.000
##  1st Qu.:19.00   1st Qu.:100.0   1st Qu.: 65.00   1st Qu.: 6.900
##  Median :26.00   Median :120.0   Median : 80.00   Median : 7.500
##  Mean   :29.87   Mean   :113.2   Mean   : 76.46   Mean   : 8.726
##  3rd Qu.:39.00   3rd Qu.:120.0   3rd Qu.: 90.00   3rd Qu.: 8.000
##  Max.   :70.00   Max.   :160.0   Max.   :100.00   Max.   :19.000
##     BodyTemp        HeartRate        RiskLevel
##  Min.   : 98.00   Min.   : 7.0   low risk :406
##  1st Qu.: 98.00   1st Qu.:70.0   mid risk :336
##  Median : 98.00   Median :76.0   high risk:272
##  Mean   : 98.67   Mean   :74.3
##  3rd Qu.: 98.00   3rd Qu.:80.0
```

```
##  Max.   :103.00   Max.   :90.0
```

I will end up creating visualizations for the variables in the data set but there are 2 variables I am going to be particular interested in. The first is age, I see I have a min of 10 and a max of 70, both of those seem pretty atypical for pregnancies. The second is heart rate; the min of 7 seems much too low to be an accurate measurement.

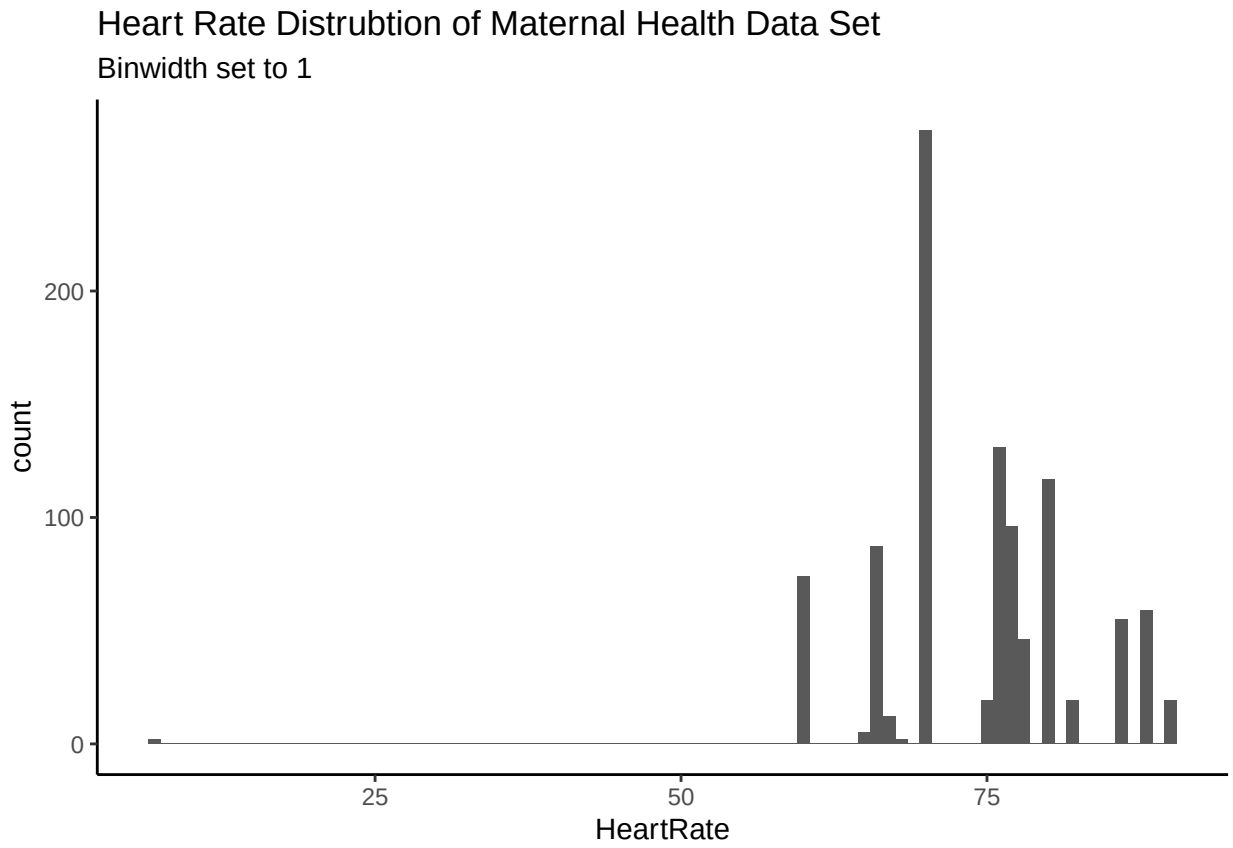
I will start by creating histogram plots for age and heart rate to see how skewed this distributions may be. For the age plot, I will be using a bin width of 1.

```
ggplot(maternalHealth, aes(x = Age)) +
  geom_histogram(binwidth = 1) +
  theme_classic() +
  labs(
    title = "Age Distrubtion of Maternal Health Data Set",
    subtitle = "Binwidth set to 1"
  )
```



As expected, the data is slightly skewed to a younger population but the distrubtion does not look as bad as I thought it might.

```
ggplot(maternalHealth, aes(x = HeartRate)) +
  geom_histogram(binwidth = 1) +
  theme_classic() +
  labs(
    title = "Heart Rate Distrubtion of Maternal Health Data Set",
    subtitle = "Binwidth set to 1"
  )
```

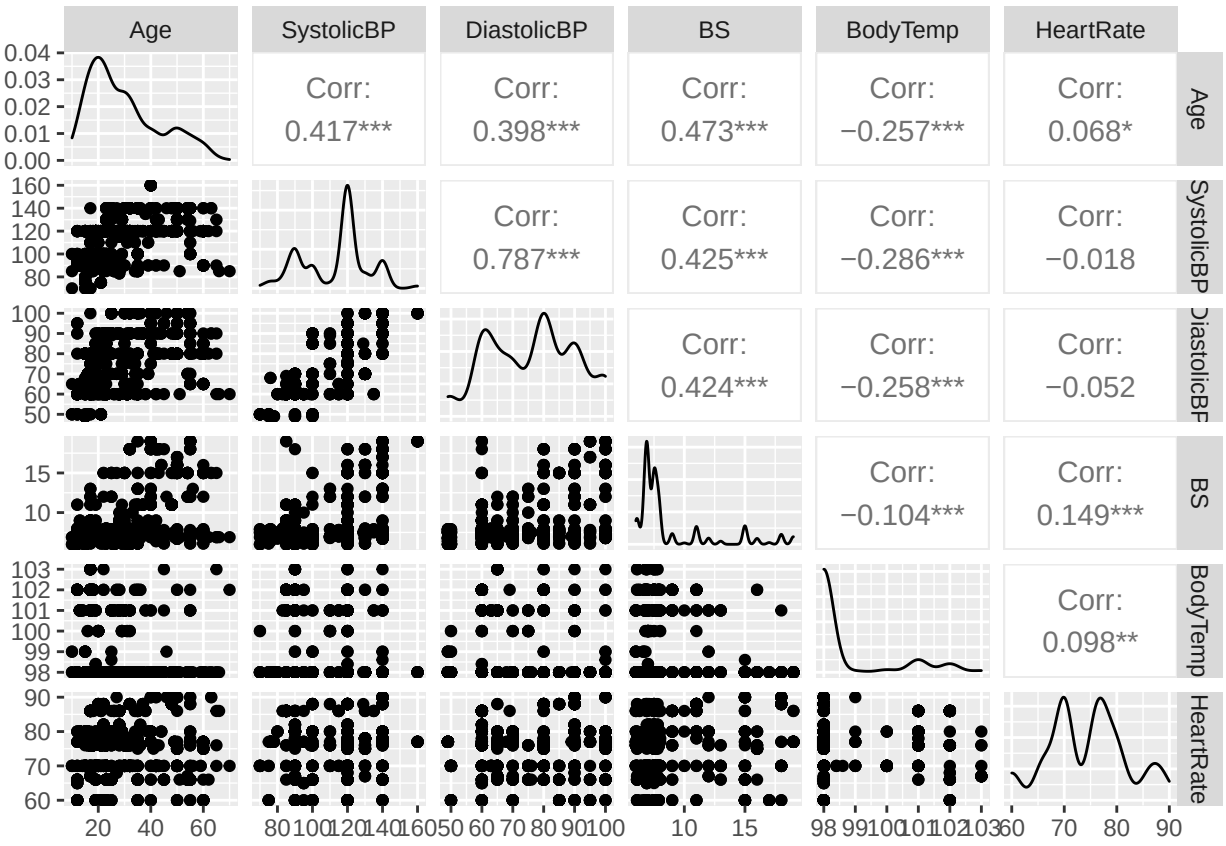


Based upon the heart rate histogram, I am going to remove the heart rate of 7, that measurement is an outlier.

```
maternalHealthClean1 <- maternalHealth %>%
  filter(HeartRate > 50)
```

Next, we are going to see if any our of the variables have any colinearity using a `pairs` plot.

```
ggpairs(select(maternalHealthClean1, -RiskLevel))
```



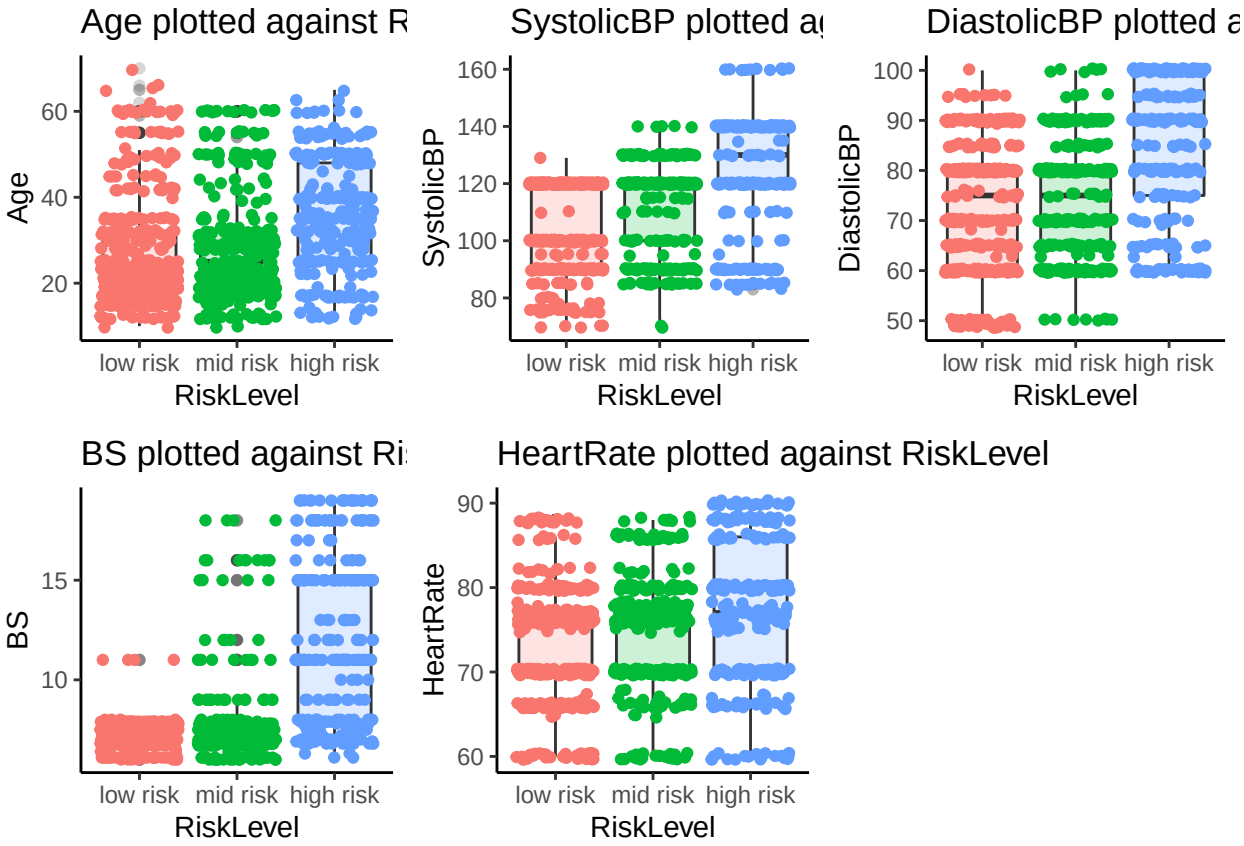
Fortunately, it appears like the only highly correlated variables are systolic and diastolic blood pressure which is not surprising.

Finally, we will create some visualizations to assess the variablity in relation to our outcome variable. I will not be looking at the body temperature variable since those summary stats indicate the data has little spread. I will be creating a scatter/boxplot combo, a plot I use pretty extensively.

```
makeMyPlot <- function(data, outcome, var) {
  ggplot(data, aes(x = {{outcome}}, y = {{var}})) +
    geom_boxplot(aes(fill = {{outcome}}), alpha = 0.2) +
    geom_jitter(aes(color = {{outcome}}))  +
    theme_classic() +
    labs(
      title = paste(as_string(ensym(var)), "plotted against", as_string(ensym(outcome)))
    ) +
    theme(legend.position = 'none')
}

p1 <- makeMyPlot(maternalHealthClean1, RiskLevel, Age)
p2 <- makeMyPlot(maternalHealthClean1, RiskLevel, SystolicBP)
p3 <- makeMyPlot(maternalHealthClean1, RiskLevel, DiastolicBP)
p4 <- makeMyPlot(maternalHealthClean1, RiskLevel, BS)
p5 <- makeMyPlot(maternalHealthClean1, RiskLevel, HeartRate)

plot_grid(p1, p2, p3, p4, p5)
```



Based upon the plots, it appears like there may be a number of indicator for a high risk pregnancy including age, systolic blood pressure, and blood sugar. It may be harder to identify a low vs mid risk pregnancy.

Statistical Test

Since I have 3 possible outcomes, a simple t test will not work. I will need to use a pairwise t-test to compare across the three groups. I will choose the systolic blood pressure variable and hypothesis that I should see a difference in means between the outcome groups.

```
pwt <- maternalHealthClean1 %>%
  pairwise_t_test(SystolicBP ~ RiskLevel, p.adjust.method = "bonferroni")

kbl(pwt) %>%
  kable_styling(latex_options = "HOLD_position")
```

.y.	group1	group2	n1	n2	p	p.signif	p.adj	p.adj.signif
SystolicBP	low risk	mid risk	404	336	0	****	0	****
SystolicBP	low risk	high risk	404	272	0	****	0	****
SystolicBP	mid risk	high risk	336	272	0	****	0	****

Based upon the results of the pairwise t test, each level is significantly different from each other with the low and high being the most(?).