Foetal Health Analysis

Esther Waweru / Miriam Onyango

2024-09-28

Introduction

Fetal health classification is a crucial area of medical research, aimed at early detection of potential health risks during pregnancy. The fetal period is a critical stage of development, where complications such as fetal distress, congenital abnormalities, or other health conditions can have long-lasting impacts on both the mother and child. Monitoring fetal health through various medical techniques—such as cardiotocography (CTG), ultrasound, and other diagnostic tests—enables healthcare providers to intervene early, ensuring better outcomes for both the fetus and mother.

Using features extracted from Cardiotocogram exams, this project focuses on using advanced machine learning techniques to classify fetal health as either, Normal,Suspect or Pathological.

Problem Definition

Our Task is to :

1. Conduct an in-depth analysis of vital fetal indicators, focusing on various features and their interconnected relationships.
2. Classify fetal health to prevent child and maternal mortality (build a predictive model to classify the observations into their respective fetal health category, i.e., Normal or Suspect or Pathological).

Metrics of Success

The success of the fetal health classification model will be measured using various performance metrics such as accuracy, precision, confusion matrix.

Data Description

The attached dataset consists of measurements of fetal heart rate (FHR) and uterine contraction (UC) features on CTGs classified by expert obstetricians. There are a total of 22 Variables in the fetal health dataset. Predictor variables are 21 (20 numeric and 1 categorical). The response variable (fetal_health) has 3 categories: Normal, Suspect and Pathological.

Features:

- baseline value - Baseline Fetal Heart Rate (FHR) (beats per minute)
- accelerations - Number of accelerations per second
- fetal_movement - Number of fetal movements per second
- uterine_contractions - Number of uterine contractions per second

- light_decelerations - Number of light decelerations per second
- severe_decelerations - Number of severe decelerations per second
- prolongued_decelerations - Number of prolonged decelerations per second
- abnormal_short_term_variability - Percentage of time with abnormal short-term variability
- mean_value_of_short_term_variability - Mean value of short-term variability
- percentage_of_time_with_abnormal_long_term_variability - Percentage of time with abnormal long-term variability
- mean_value_of_long_term_variability - Mean value of long-term variability
- histogram_width - Width of FHR histogram (generated from exam)
- histogram_min - Minimum of FHR histogram (generated from exam)
- histogram_max - Maximum of FHR histogram (generated from exam)
- histogram_number_of_peaks - Number of FHR histogram peaks (generated from exam)
- histogram_number_of_zeroes - Number of FHR histogram zeroes (generated from exam)
- histogram_mode - Mode of FHR histogram (generated from exam)
- histogram_mean - Mean of FHR histogram (generated from exam)
- histogram_median - Median of FHR histogram (generated from exam)
- histogram_variance - Variance of FHR histogram (generated from exam)
- histogram_tendency - Tendency of FHR histogram (generated from exam)

Target:

- fetal_health - Fetal health as assessed by expert obstetrician. 1 - Normal, 2 - Suspect, 3 - Pathological

Import libraries and Load data

```r
#import and load libraries
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'ggplot2' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'readr' was built under R version 4.2.3

## Warning: package 'purrr' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.3

## Warning: package 'forcats' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 4.2.3
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.3
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 4.2.3
```

```
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.2.3
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(smotefamily)
```

```
## Warning: package 'smotefamily' was built under R version 4.2.3
```

```r
#load dataset
setwd('C://Users//pc//Documents//Project//Foetal')

data <- read.csv('foetal_health_data.csv')
head(data)
```

```
##   baseline_value accelerations fetal_movement uterine_contractions
## 1            120         0.000              0                0.000
## 2            132         0.006              0                0.006
## 3            133         0.003              0                0.008
## 4            134         0.003              0                0.008
## 5            132         0.007              0                0.008
## 6            134         0.001              0                0.010
##   light_decelerations severe_decelerations prolongued_decelerations
## 1               0.000                    0                    0.000
## 2               0.003                    0                    0.000
## 3               0.003                    0                    0.000
## 4               0.003                    0                    0.000
## 5               0.000                    0                    0.000
## 6               0.009                    0                    0.002
##   abnormal_short_term_variability mean_value_of_short_term_variability
## 1                              73                                  0.5
## 2                              17                                  2.1
## 3                              16                                  2.1
## 4                              16                                  2.4
## 5                              16                                  2.4
## 6                              26                                  5.9
##   percentage_of_time_with_abnormal_long_term_variability
## 1                                                     43
## 2                                                      0
## 3                                                      0
## 4                                                      0
## 5                                                      0
## 6                                                      0
##   mean_value_of_long_term_variability histogram_width histogram_min
## 1                                 2.4              64            62
## 2                                10.4             130            68
## 3                                13.4             130            68
## 4                                23.0             117            53
## 5                                19.9             117            53
## 6                                 0.0             150            50
##   histogram_max histogram_number_of_peaks histogram_number_of_zeroes
## 1           126                         2                          0
## 2           198                         6                          1
## 3           198                         5                          1
## 4           170                        11                          0
## 5           170                         9                          0
## 6           200                         5                          3
##   histogram_mode histogram_mean histogram_median histogram_variance
## 1            120            137              121                 73
## 2            141            136              140                 12
## 3            141            135              138                 13
## 4            137            134              137                 13
## 5            137            136              138                11
## 6             76            107              107               170
##   histogram_tendency fetal_health
## 1                   1            2
## 2                   0            1
## 3                   0            1
## 4                   1            1
## 5                   1            1
## 6                   0            3
```

```
#Shape of our data
dim(data)
```

```
## [1] 2126   22
```

Our data has 2126 rows and 22 columns.

```
#check missing values
sum(is.na(data))
```

```
## [1] 0
```

Our data has no missing values.

```
#check data type

str(data)
```

```
## 'data.frame':    2126 obs. of  22 variables:
##  $ baseline_value                                        : int  120 132 133 134 132 134 134 122 122
##  $ accelerations                                         : num  0 0.006 0.003 0.003 0.007 0.001 0.00
##  $ fetal_movement                                        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ uterine_contractions                                  : num  0 0.006 0.008 0.008 0.008 0.01 0.013
##  $ light_decelerations                                   : num  0 0.003 0.003 0.003 0 0.009 0.008 0
##  $ severe_decelerations                                  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ prolongued_decelerations                              : num  0 0 0 0 0 0.002 0.003 0 0 0 ...
##  $ abnormal_short_term_variability                       : int  73 17 16 16 16 26 29 83 84 86 ...
##  $ mean_value_of_short_term_variability                  : num  0.5 2.1 2.1 2.4 2.4 5.9 6.3 0.5 0.5
##  $ percentage_of_time_with_abnormal_long_term_variability: int  43 0 0 0 0 0 0 6 5 6 ...
##  $ mean_value_of_long_term_variability                   : num  2.4 10.4 13.4 23 19.9 0 0 15.6 13.6
##  $ histogram_width                                       : int  64 130 130 117 117 150 150 68 68 68
##  $ histogram_min                                         : int  62 68 68 53 53 50 50 62 62 62 ...
##  $ histogram_max                                         : int  126 198 198 170 170 200 200 130 130
##  $ histogram_number_of_peaks                             : int  2 6 5 11 9 5 6 0 0 1 ...
##  $ histogram_number_of_zeroes                            : int  0 1 1 0 0 3 3 0 0 0 ...
##  $ histogram_mode                                        : int  120 141 141 137 137 76 71 122 122 12
##  $ histogram_mean                                        : int  137 136 135 134 136 107 107 122 122
##  $ histogram_median                                      : int  121 140 138 137 138 107 106 123 123
##  $ histogram_variance                                    : int  73 12 13 13 11 170 215 3 3 1 ...
##  $ histogram_tendency                                    : int  1 0 0 1 1 0 0 1 1 1 ...
##  $ fetal_health                                          : int  2 1 1 1 1 3 3 3 3 3 ...
```

The variables are mainly numerical. We will change fetal_movement data type to factor.

```
#Change fetal_health data type to factor
data$fetal_health <- as.factor(data$fetal_health)
str(data)
```

```
## 'data.frame':    2126 obs. of  22 variables:
##  $ baseline_value                                        : int  120 132 133 134 132 134 134 122 122
##  $ accelerations                                         : num  0 0.006 0.003 0.003 0.007 0.001 0.00
##  $ fetal_movement                                        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ uterine_contractions                                  : num  0 0.006 0.008 0.008 0.008 0.01 0.013
##  $ light_decelerations                                   : num  0 0.003 0.003 0.003 0 0.009 0.008 0
##  $ severe_decelerations                                  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ prolongued_decelerations                              : num  0 0 0 0 0 0.002 0.003 0 0 0 ...
##  $ abnormal_short_term_variability                       : int  73 17 16 16 16 26 29 83 84 86 ...
##  $ mean_value_of_short_term_variability                  : num  0.5 2.1 2.1 2.4 2.4 5.9 6.3 0.5 0.5
##  $ percentage_of_time_with_abnormal_long_term_variability: int  43 0 0 0 0 0 0 6 5 6 ...
##  $ mean_value_of_long_term_variability                   : num  2.4 10.4 13.4 23 19.9 0 0 15.6 13.6
##  $ histogram_width                                       : int  64 130 130 117 117 150 150 68 68 68
##  $ histogram_min                                         : int  62 68 68 53 53 50 50 62 62 62 ...
##  $ histogram_max                                         : int  126 198 198 170 170 200 200 130 130
##  $ histogram_number_of_peaks                             : int  2 6 5 11 9 5 6 0 0 1 ...
##  $ histogram_number_of_zeroes                            : int  0 1 1 0 0 3 3 0 0 0 ...
##  $ histogram_mode                                        : int  120 141 141 137 137 76 71 122 122 12
##  $ histogram_mean                                        : int  137 136 135 134 136 107 107 122 122
##  $ histogram_median                                      : int  121 140 138 137 138 107 106 123 123
##  $ histogram_variance                                    : int  73 12 13 13 11 170 215 3 3 1 ...
##  $ histogram_tendency                                    : int  1 0 0 1 1 0 0 1 1 1 ...
##  $ fetal_health                                          : Factor w/ 3 levels "1","2","3": 2 1 1 1 1
```

```
#duplicated values

sum(duplicated(data))
```

```
## [1] 13
```

Our data has 13 duplicated rows.

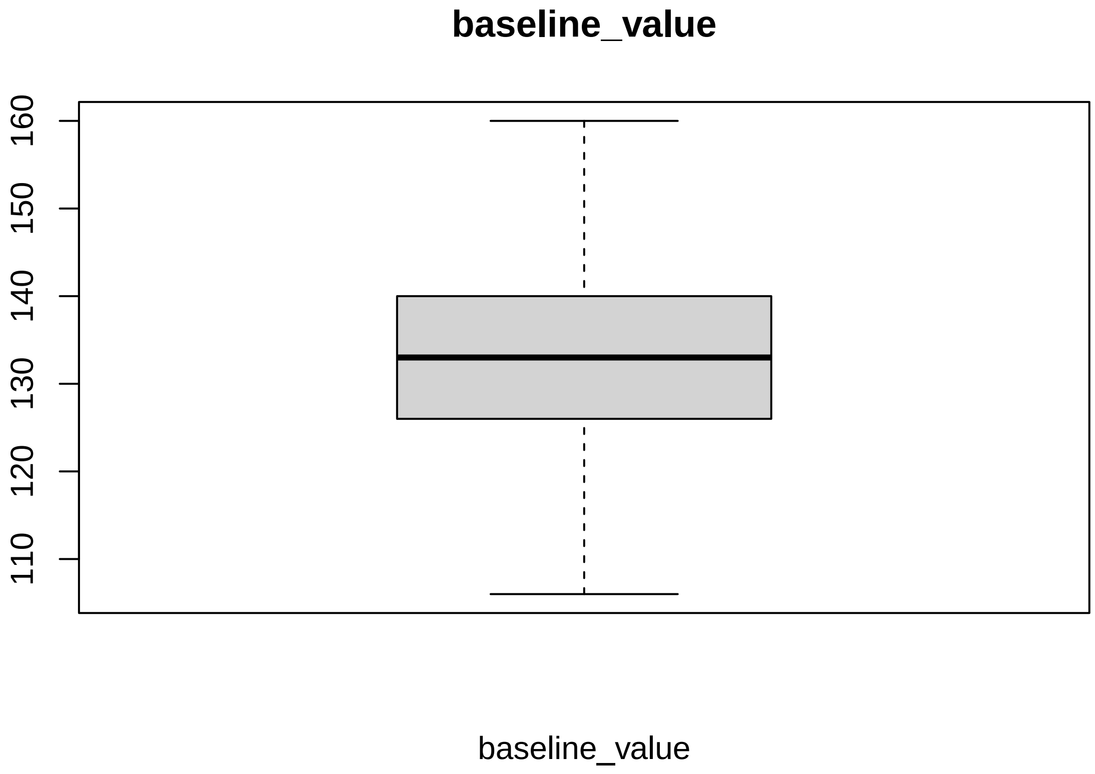
```
#Check rows duplicated
duplicates <- data[duplicated(data), ]
duplicates
```

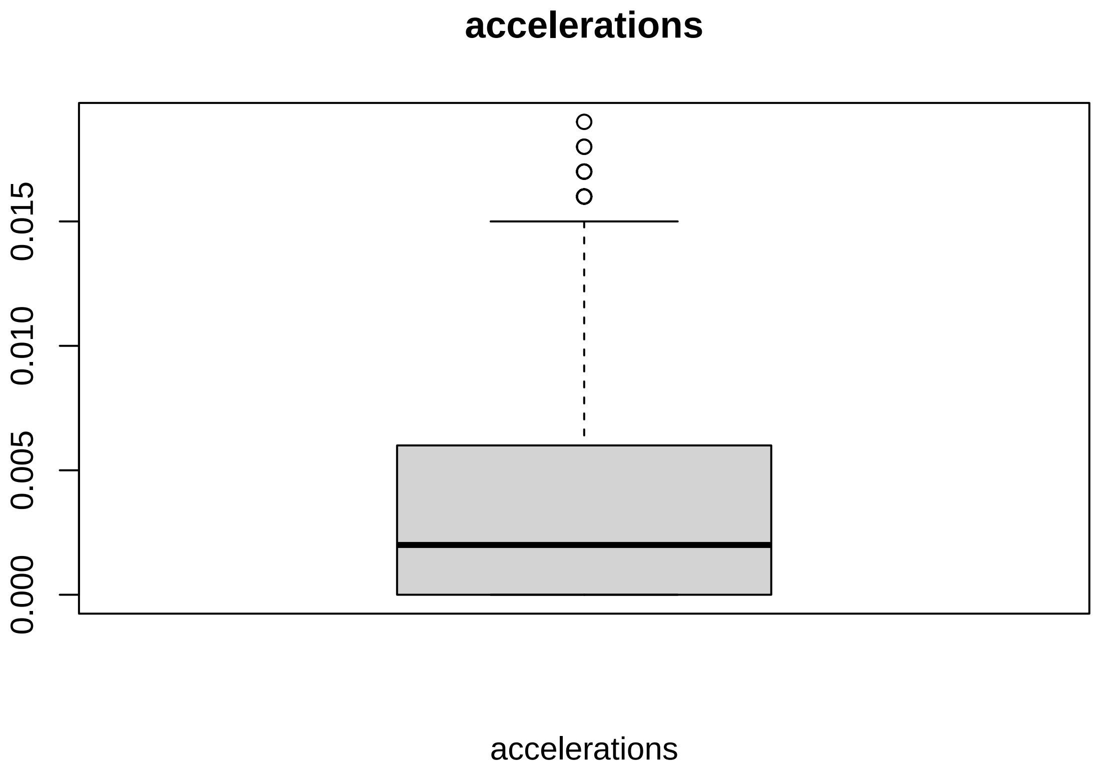
```
##      baseline_value accelerations fetal_movement uterine_contractions
## 69              140         0.007          0.000                0.004
## 235             123         0.000          0.000                0.000
## 307             145         0.000          0.020                0.000
## 325             135         0.000          0.000                0.000
## 334             144         0.000          0.019                0.000
## 788             123         0.003          0.003                0.000
## 792             123         0.003          0.004                0.000
## 799             146         0.000          0.000                0.003
## 850             138         0.002          0.000                0.004
## 1114            122         0.000          0.000                0.000
## 1115            122         0.000          0.000                0.000
## 1116            122         0.000          0.000                0.000
## 1459            148         0.005          0.000                0.002
##      light_decelerations severe_decelerations prolongued_decelerations
## 69                     0                    0                        0
## 235                    0                    0                        0
## 307                    0                    0                        0
## 325                    0                    0                        0
## 334                    0                    0                        0
## 788                    0                    0                        0
## 792                    0                    0                        0
## 799                    0                    0                        0
## 850                    0                    0                        0
## 1114                   0                    0                        0
## 1115                   0                    0                        0
## 1116                   0                    0                        0
## 1459                   0                    0                        0
##      abnormal_short_term_variability mean_value_of_short_term_variability
## 69                                34                                  1.2
## 235                               49                                  0.8
## 307                               77                                  0.2
## 325                               62                                  0.5
## 334                               76                                  0.4
## 788                               52                                  0.8
## 792                               50                                  0.9
## 799                               65                                  0.4
## 850                               41                                  0.8
## 1114                              19                                  1.9
## 1115                              19                                  1.9
## 1116                              19                                  1.9
## 1459                              40                                  0.9
##      percentage_of_time_with_abnormal_long_term_variability
## 69                                                        0
## 235                                                       7
## 307                                                      45
## 325                                                      71
## 334                                                      61
## 788                                                       2
## 792                                                       4
## 799                                                      39
## 850                                                       8
## 1114                                                      0
## 1115                                                      0
## 1116                                                      0
## 1459                                                      0
##      mean_value_of_long_term_variability histogram_width histogram_min
## 69                                  10.3              60           119
## 235                                 13.8              74            63
## 307                                  5.8              21           129
## 325                                  6.9              97            71
## 334                                 10.6              81            71
## 788                                 15.4              90            50
## 792                                 14.8              82            58
## 799                                  7.0              19            137
## 850                                 10.3              51            105
## 1114                                15.1              39            103
## 1115                                15.1              39            103
## 1116                                15.1              39            103
## 1459                                10.6              35            136
##      histogram_max histogram_number_of_peaks histogram_number_of_zeroes
## 69             179                         2                          0
## 235            137                         2                          0
## 307            150                         1                          0
## 325            168                         3                          0
## 334            152                         3                          0
## 788            140                         7                          0
## 792            140                         7                          0
## 799            156                         1                          0
## 850            156                         4                          0
## 1114           142                         1                          0
## 1115           142                         1                          0
## 1116           142                         1                          0
## 1459           171                         1                          0
##      histogram_mode histogram_mean histogram_median histogram_variance
## 69              156            153              155                  5
## 235             129            127              129                  2
## 307             146            145              147                  0
## 325             143            142              144                  1
## 334             145            144              146                  2
## 788             129            128              130                  4
## 792             129            128              130                  5
## 799             150            149              151                  1
## 850             142            142              143                  2
## 1114            120            120              122                  3
## 1115            120            120              122                  3
## 1116            120            120              122                  3
## 1459            153            155              156                  4
##      histogram_tendency fetal_health
## 69                    0            1
## 235                   1            1
## 307                   1            2
## 325                   1            3
## 334                   1            2
## 788                   1            1
## 792                   1            1
## 799                   1            2
## 850                   1            1
## 1114                  0            1
## 1115                  0            1
## 1116                  0            1
## 1459                  0            1
```

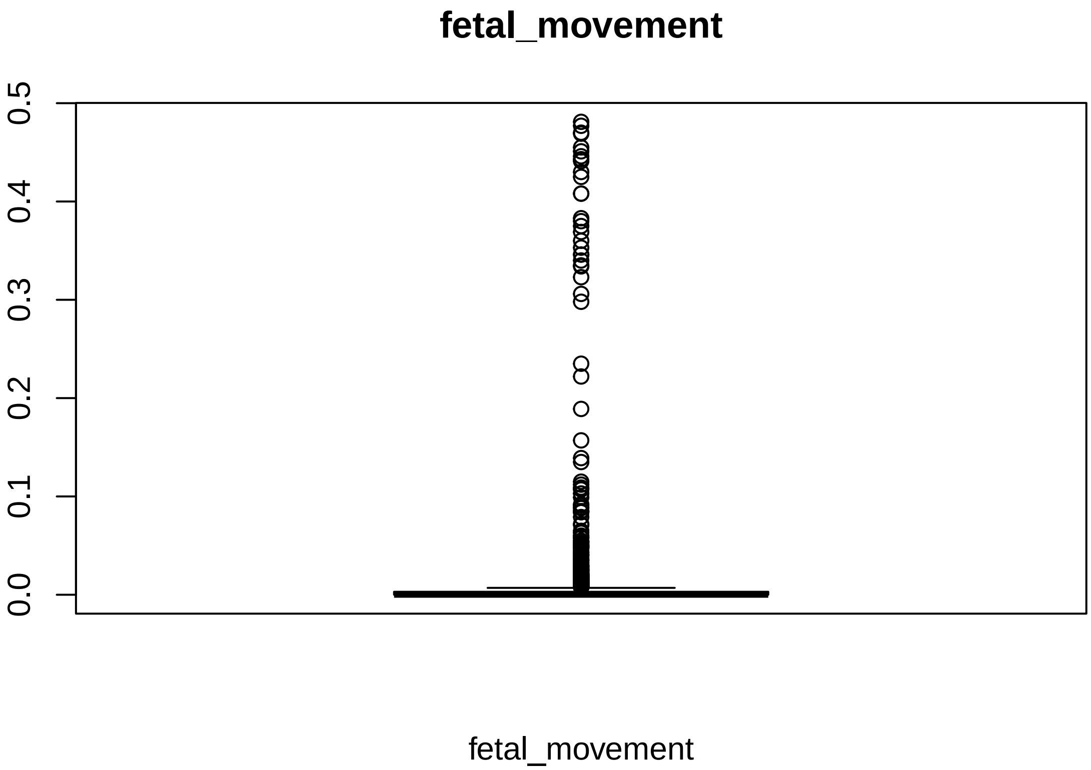
```
#Remove duplicated rows
data_2 <- unique(data)
sum(duplicated(data_2))
```

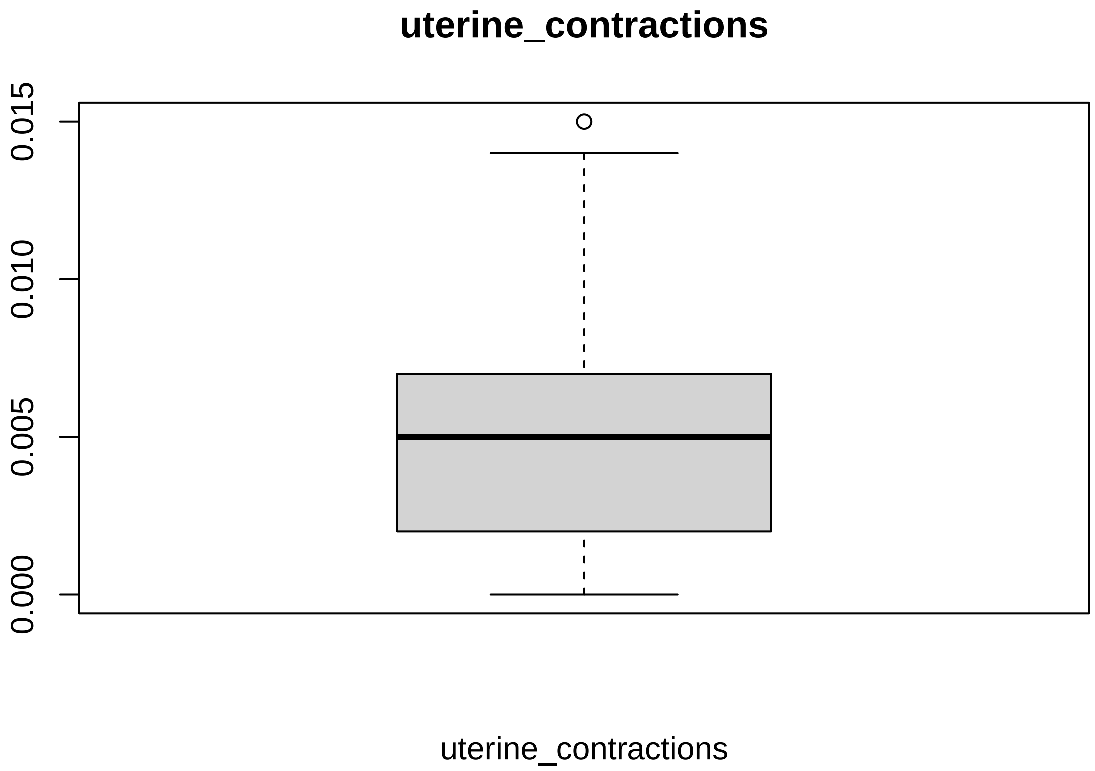
```
## [1] 0
```

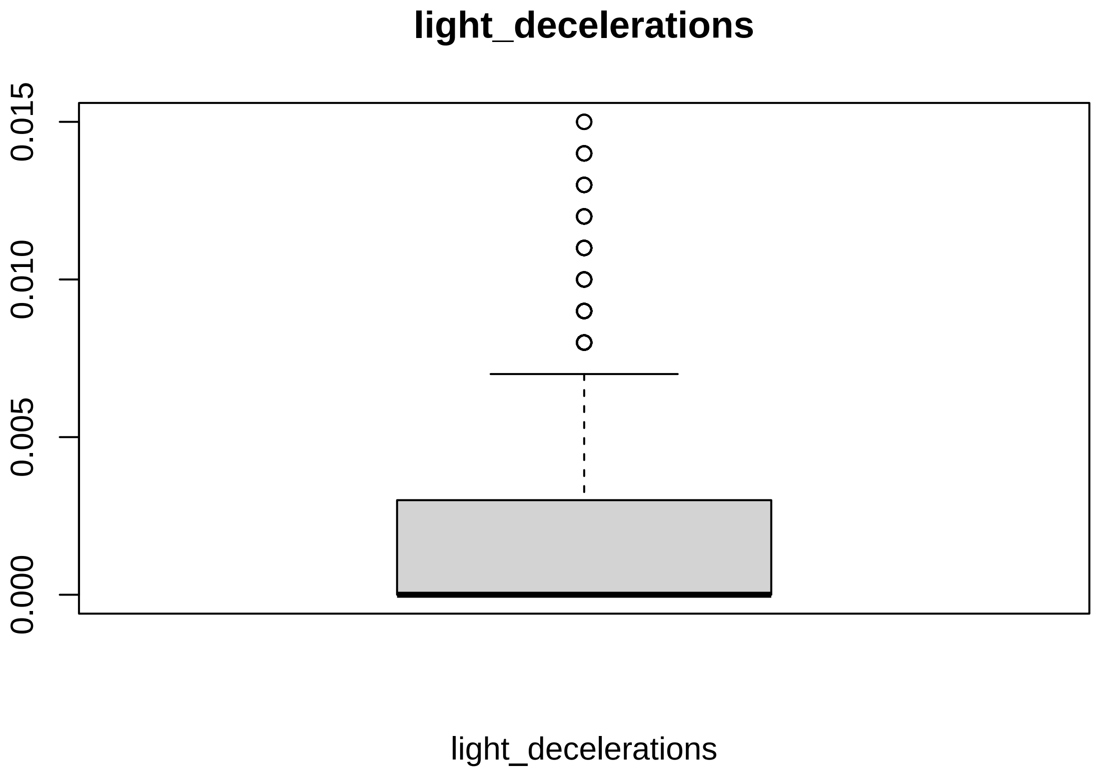
```r
#Check for outliers


# Create boxplots for each variable in the dataset
cols <- colnames(data_2[-22])

for(i in 1 : length(cols)) {
  boxplot(data_2[,i], main = cols[i],
          xlab = cols[i])
}
```

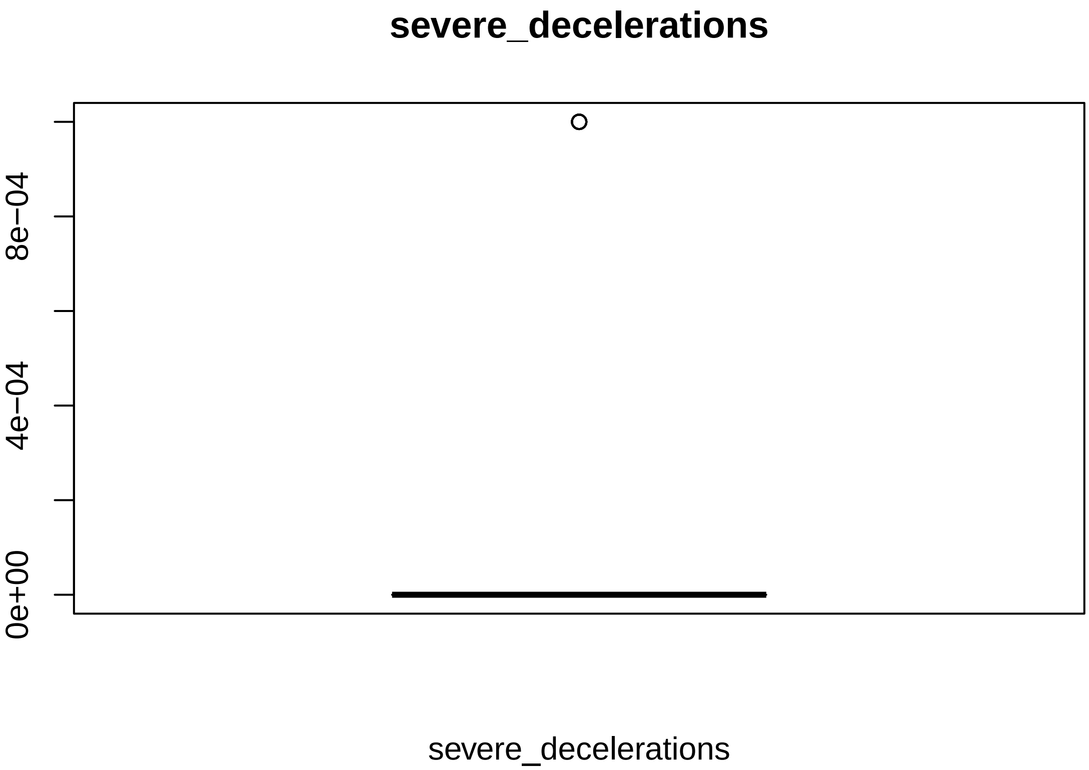


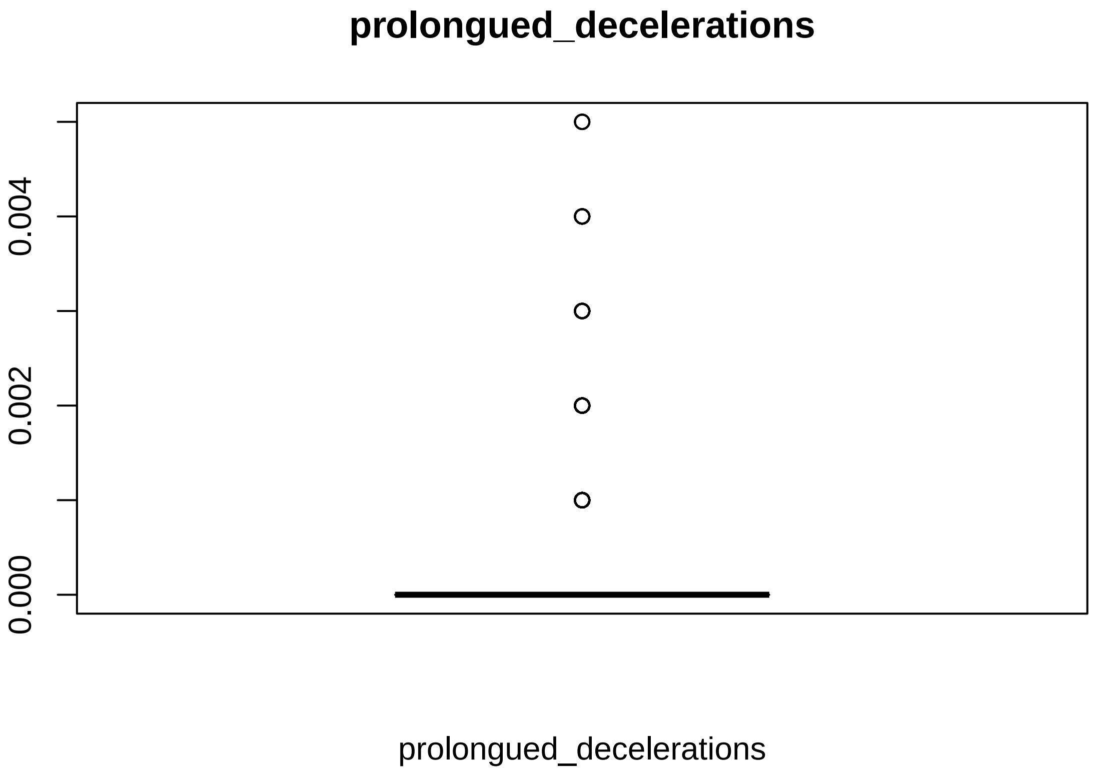

accelerations
0.000
0.005
0.010
0.015
accelerations

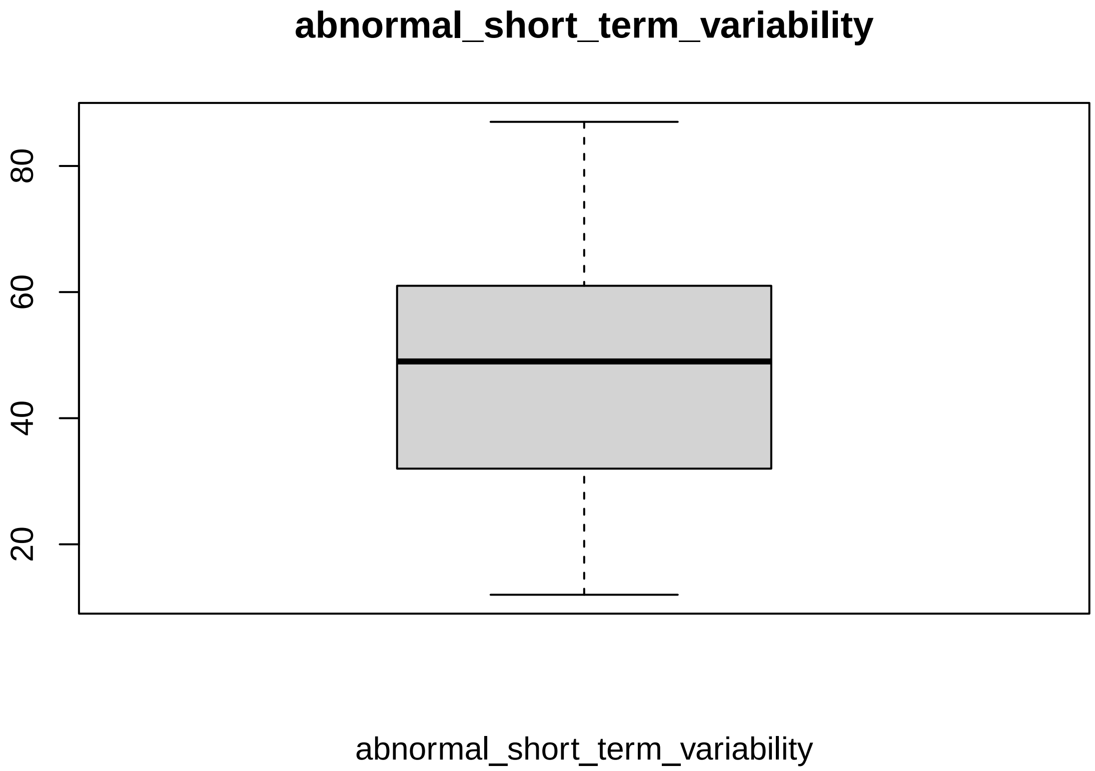

fetal_movement
0.0
0.1
0.2
0.3
0.4
0.5
fetal_movement

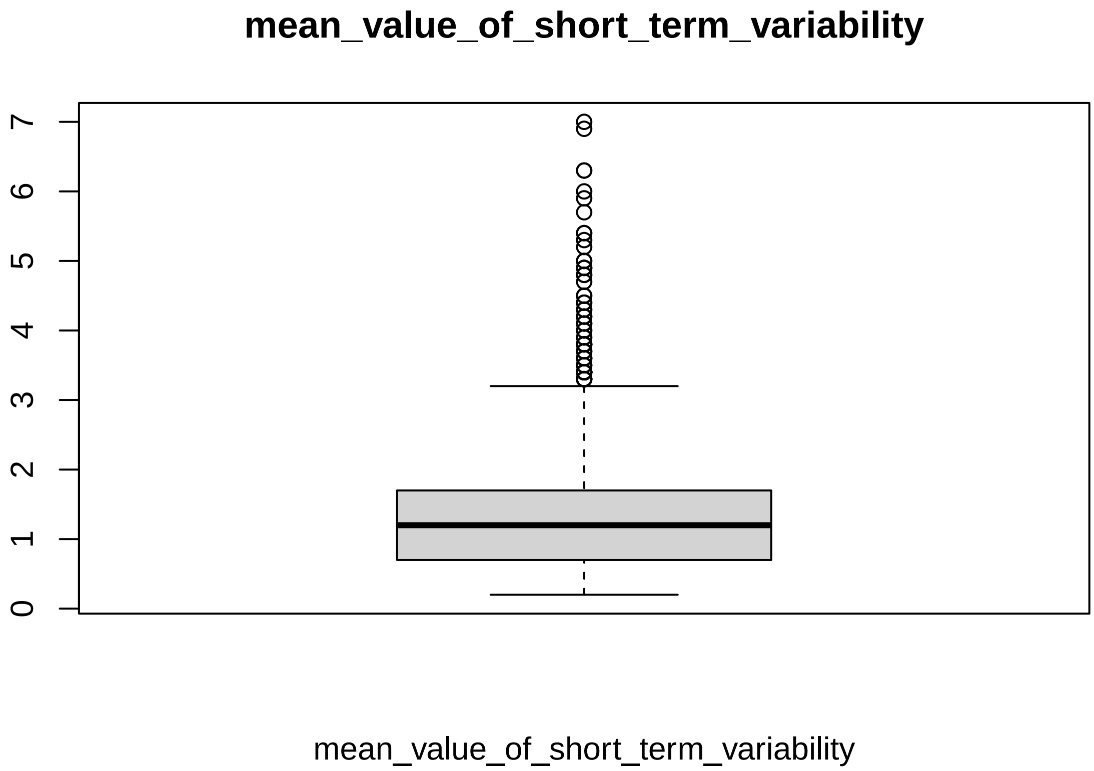

uterine_contractions
0.000
0.005
0.010
0.015
uterine_contractions

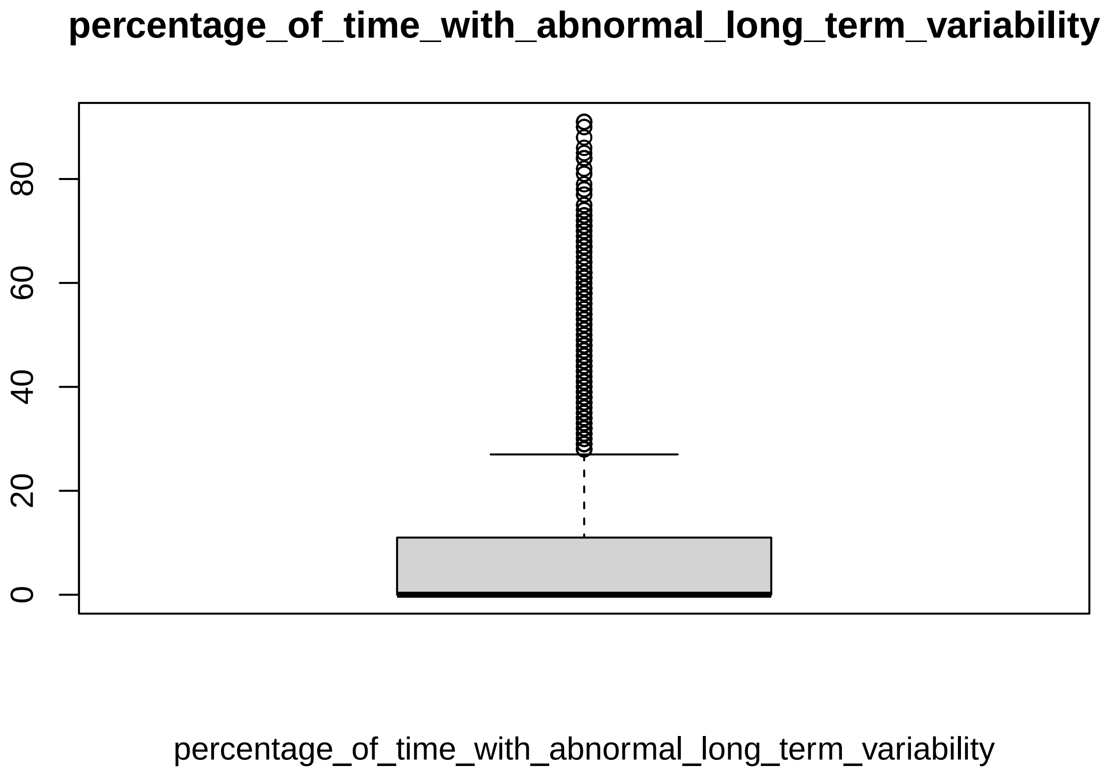

light_decelerations
0.000
0.005
0.010
0.015
light_decelerations

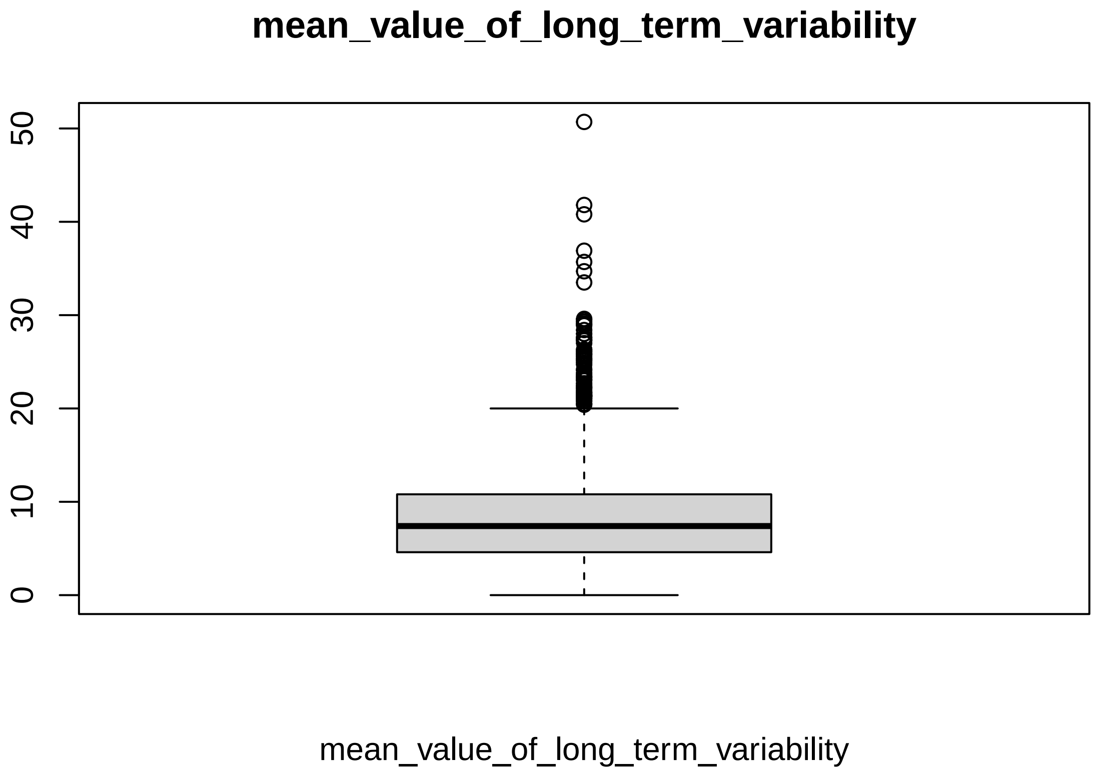

severe_decelerations
0e+00
4e−04
8e−04
severe_decelerations

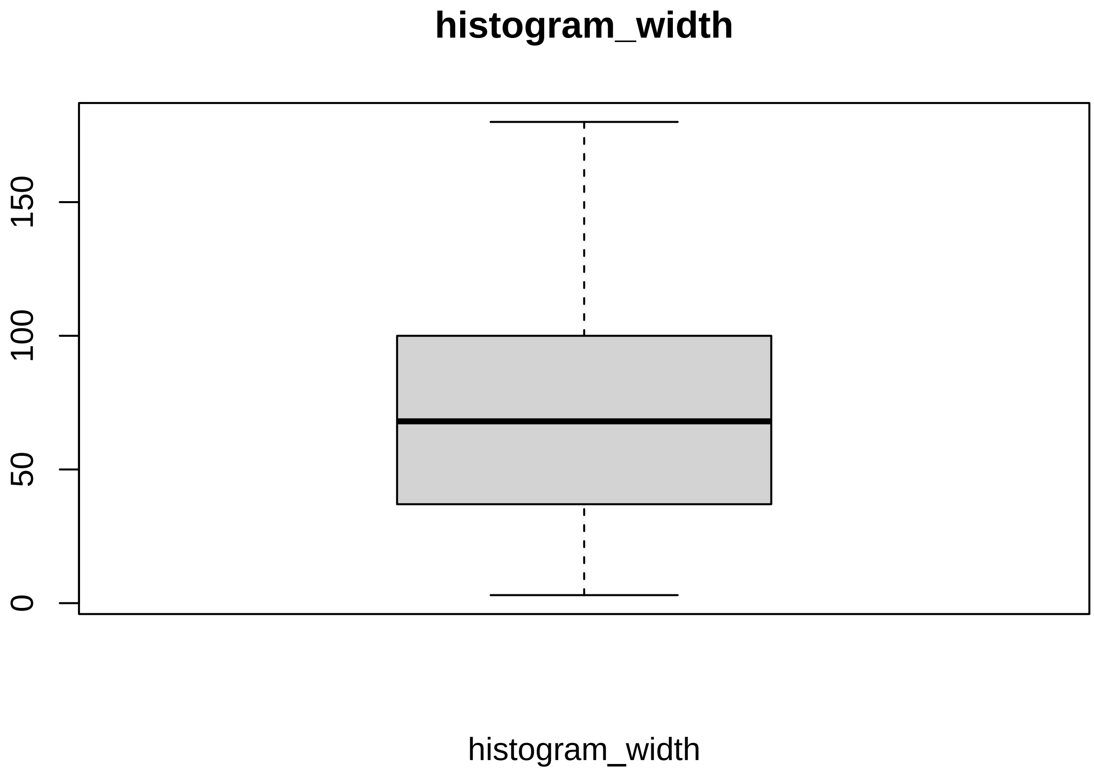

prolongued_decelerations
0.000
0.002
0.004
prolongued_decelerations

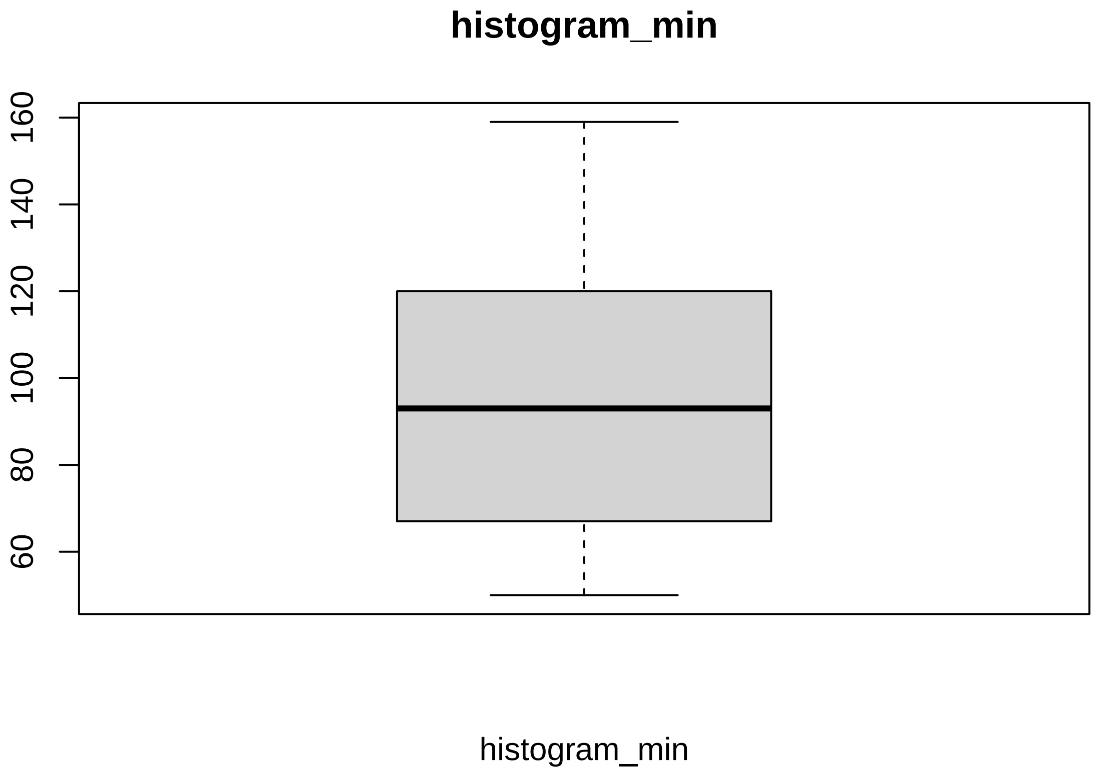

abnormal_short_term_variability
80
60
40
20
abnormal_short_term_variability

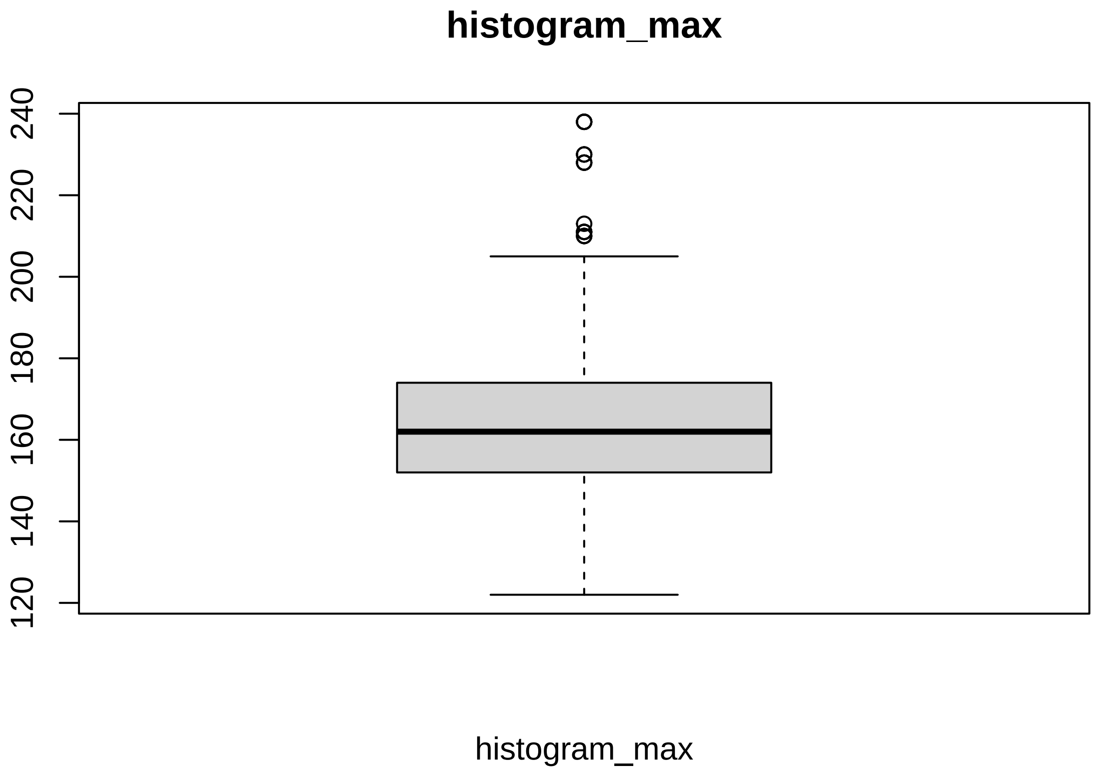

mean_value_of_short_term_variability
0
1
2
3
4
5
6
7
mean_value_of_short_term_variability

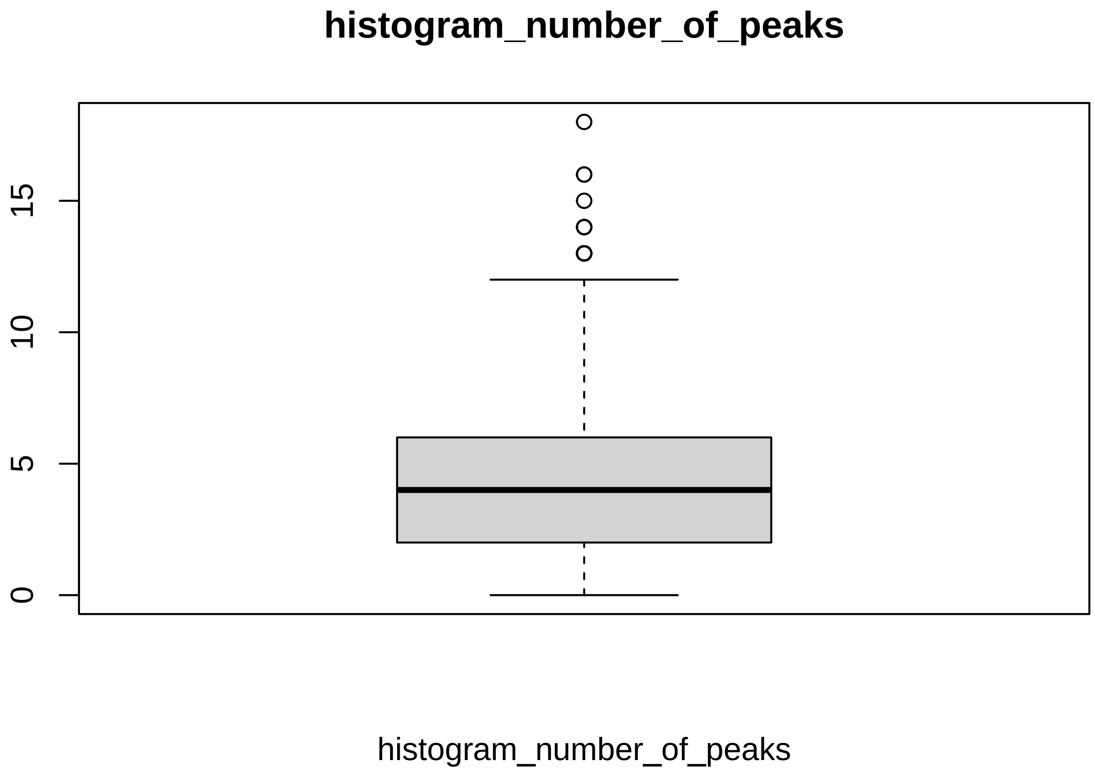

percentage_of_time_with_abnormal_long_term_variability
80
60
40
20
0
percentage_of_time_with_abnormal_long_term_variability

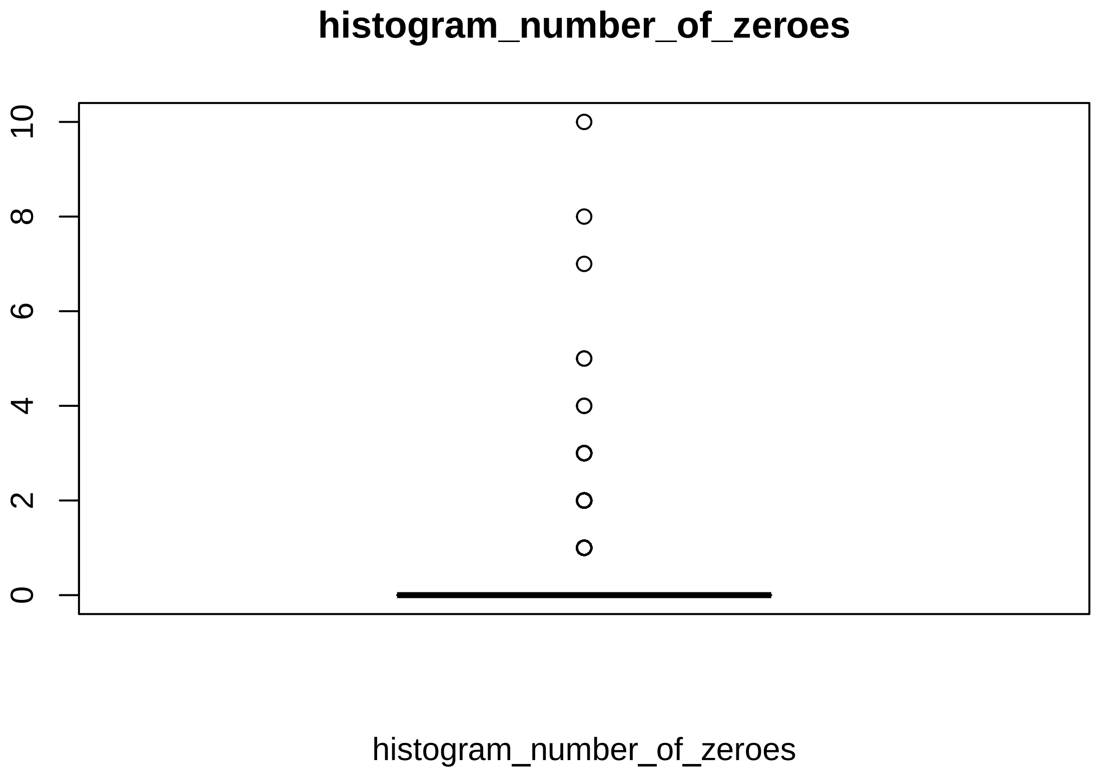

mean_value_of_long_term_variability
0
10
20
30
40
50
mean_value_of_long_term_variability

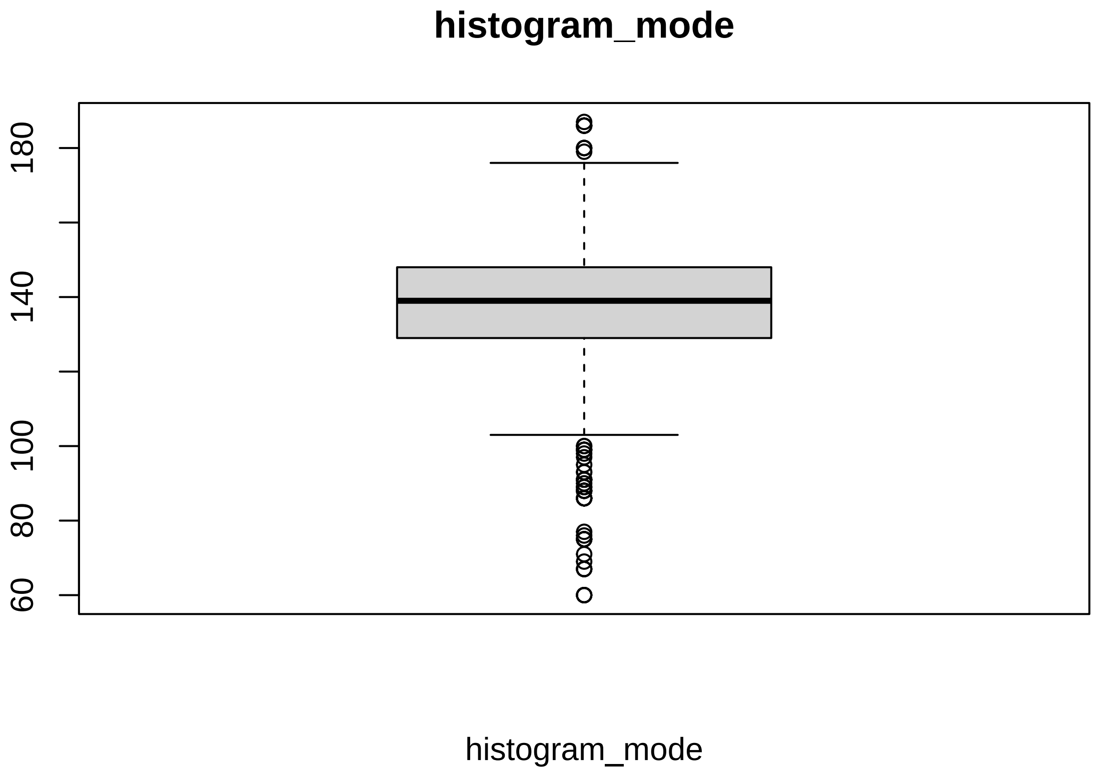

histogram_width
0
50
100
150
histogram_width

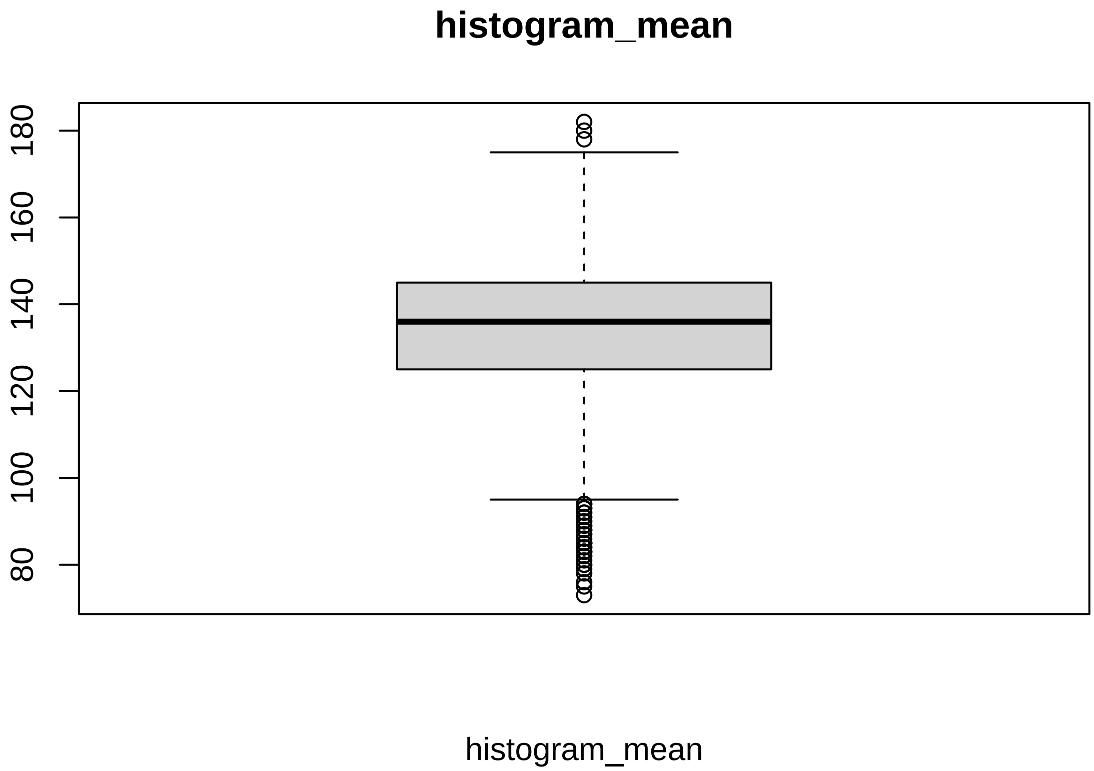

histogram_min
160
140
120
100
80
60
histogram_min

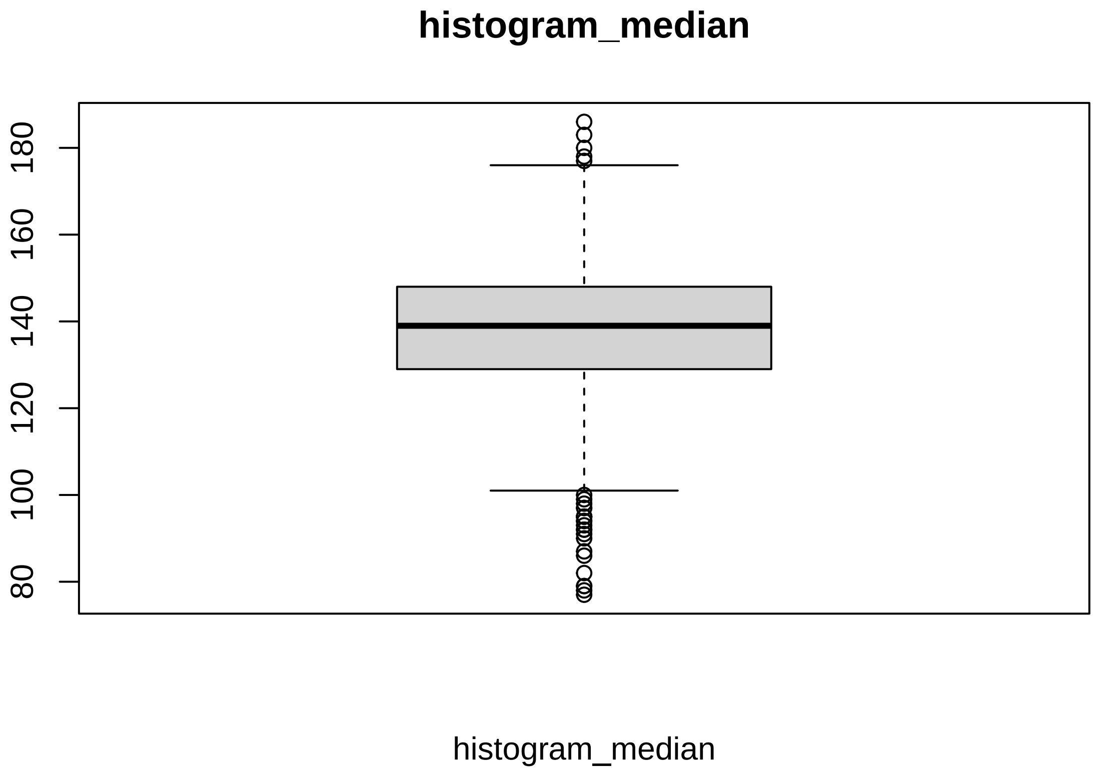

histogram_max
240
220
200
180
160
140
120
histogram_max

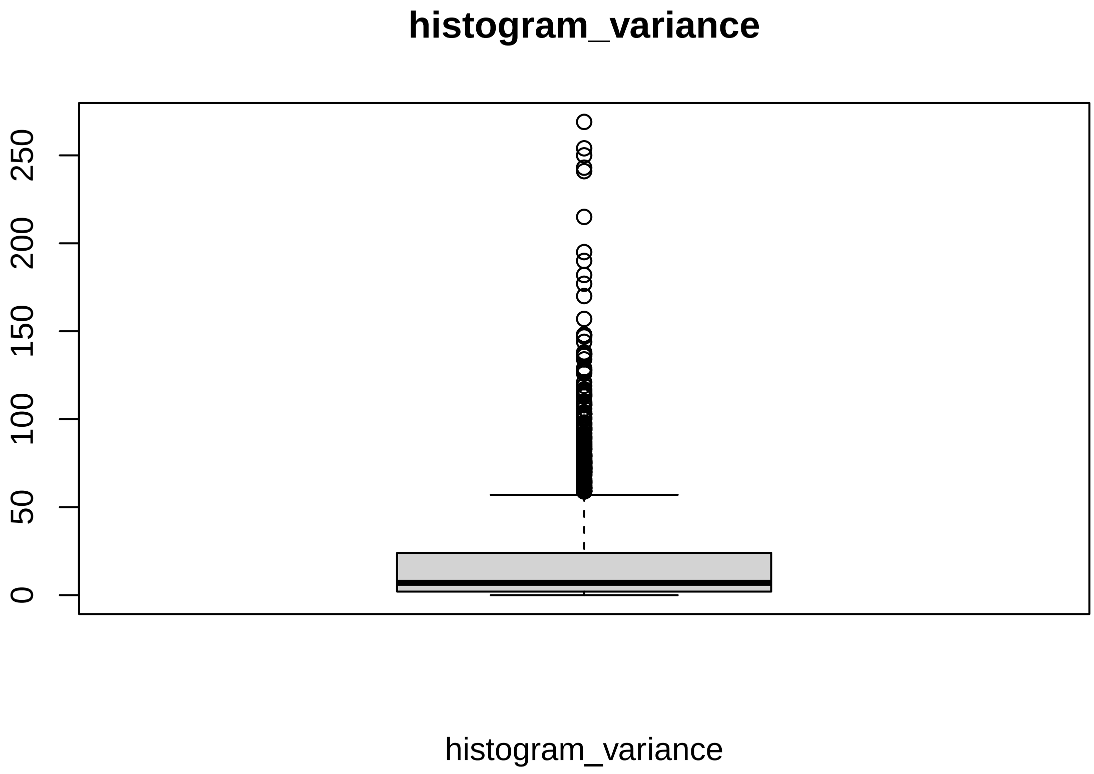

histogram_number_of_peaks
0
5
10
15
histogram_number_of_peaks

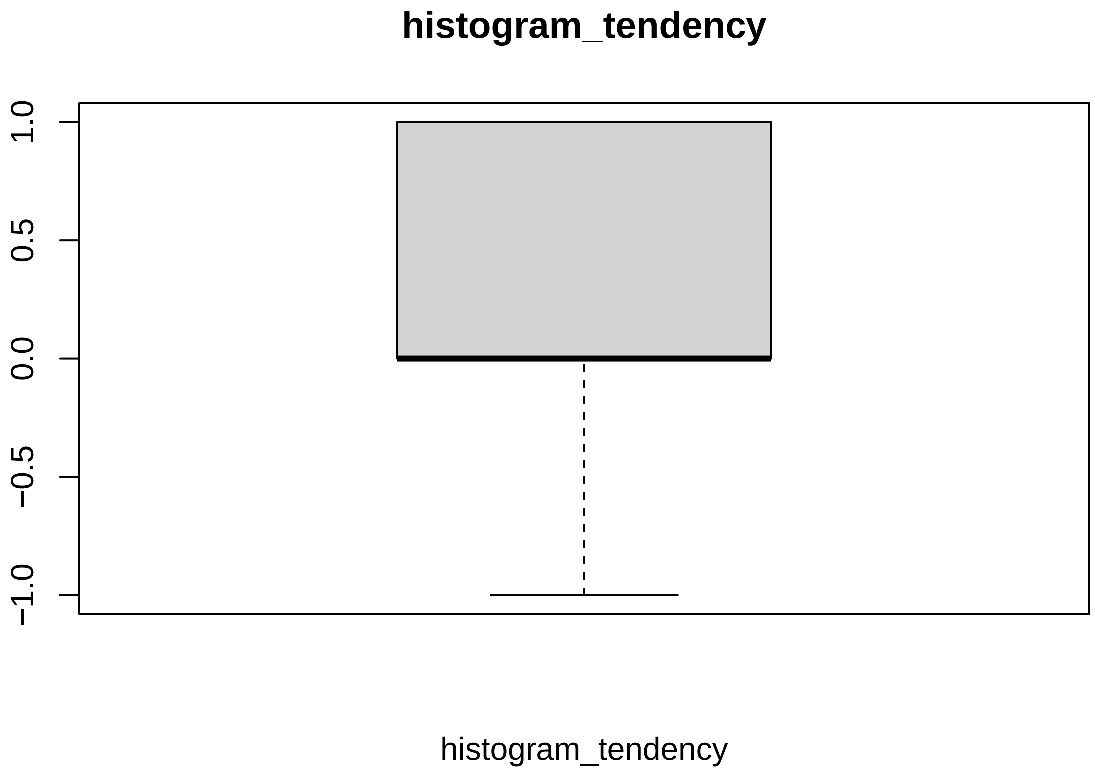

histogram_number_of_zeroes
0
2
4
6
8
10
histogram_number_of_zeroes


histogram_mode
180
140
100
80
60
histogram_mode


histogram_mean
180
160
140
120
100
80
histogram_mean


histogram_median
180
160
140
120
100
80
histogram_median


histogram_variance
0
50
100
150
200
250
histogram_variance



From the boxplot,

- All variables have outliers except for baseline_value, abnormal_short_term_variability, histogram_width,histogram_min and histogram_tendency.
- The variables fetal_movement, severe_decelerations, prolongued_decelerations and histogram_number_of_zeroes appear to have a tight spread meaning data is more concentrated around the median.
- Some of the variables appear to be skewed like accelerations,uterine_contractions,light_decelerations, mean_value_of_short_term_variability, percentage_of_time_with_abnormal_long_term_variability and histogram_variance.

Since measurements are from a cardiotocogram which is a precise medical equipment, outliers are legitimate measurements and not erroneous. Hence we retain.

```
colnames(data_2)
```

```
##  [1] "baseline_value"
##  [2] "accelerations"
##  [3] "fetal_movement"
##  [4] "uterine_contractions"
##  [5] "light_decelerations"
##  [6] "severe_decelerations"
##  [7] "prolongued_decelerations"
##  [8] "abnormal_short_term_variability"
##  [9] "mean_value_of_short_term_variability"
## [10] "percentage_of_time_with_abnormal_long_term_variability"
## [11] "mean_value_of_long_term_variability"
## [12] "histogram_width"
## [13] "histogram_min"
## [14] "histogram_max"
## [15] "histogram_number_of_peaks"
## [16] "histogram_number_of_zeroes"
## [17] "histogram_mode"
## [18] "histogram_mean"
## [19] "histogram_median"
## [20] "histogram_variance"
## [21] "histogram_tendency"
## [22] "fetal_health"
```

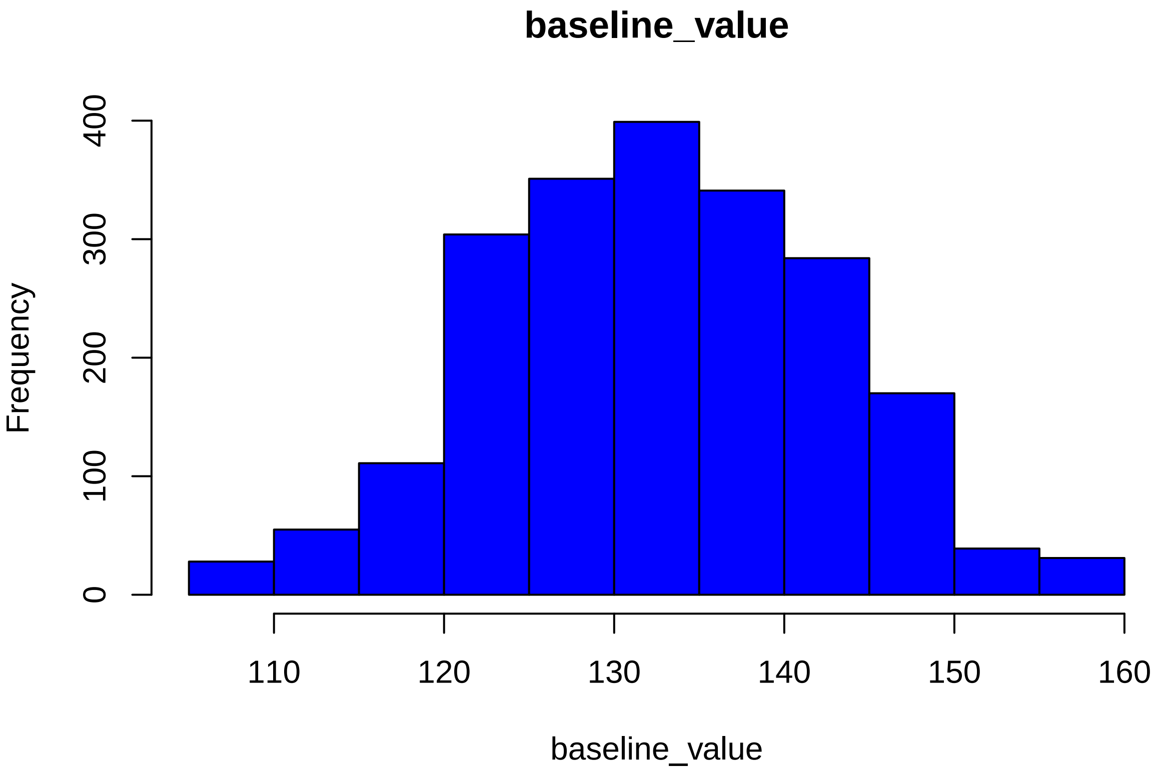
```r
summary(data_2)
```

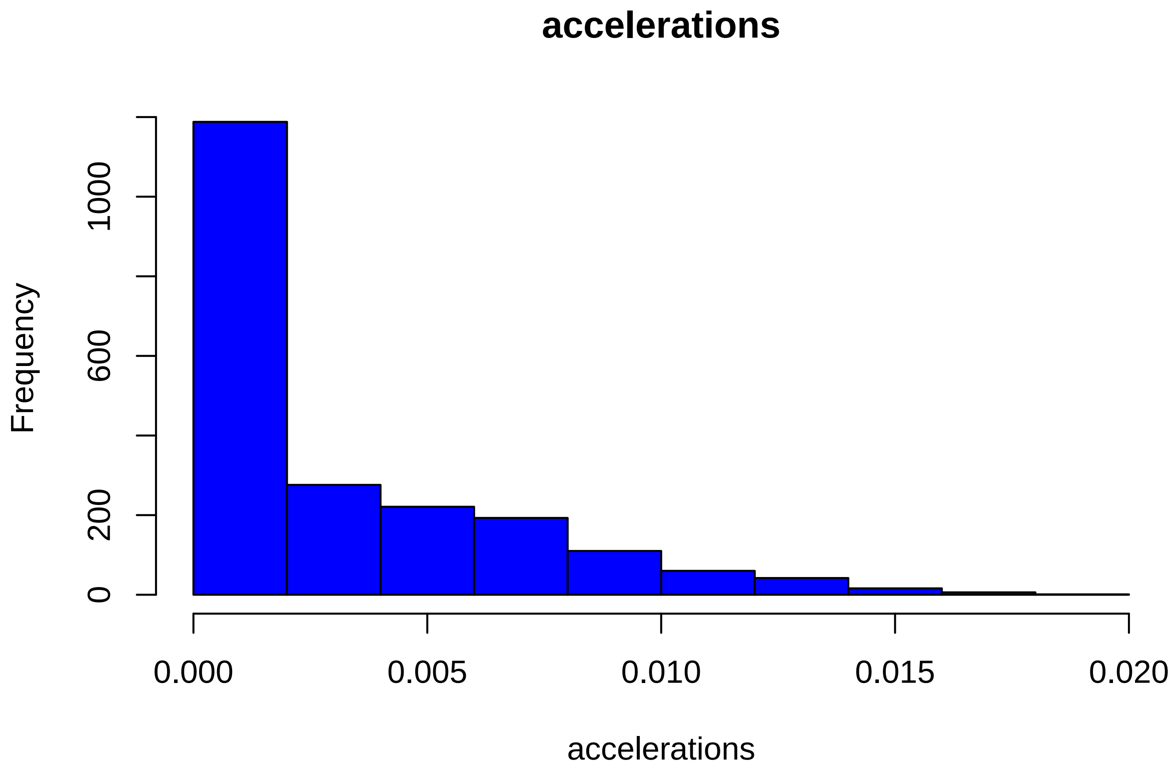
```
##  baseline_value  accelerations      fetal_movement     uterine_contractions
##  Min.   :106.0   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:126.0   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.002000
##  Median :133.0   Median :0.002000   Median :0.000000   Median :0.005000
##  Mean   :133.3   Mean   :0.003188   Mean   :0.009517   Mean   :0.004387
##  3rd Qu.:140.0   3rd Qu.:0.006000   3rd Qu.:0.003000   3rd Qu.:0.007000
##  Max.   :160.0   Max.   :0.019000   Max.   :0.481000   Max.   :0.015000
##  light_decelerations severe_decelerations prolongued_decelerations
##  Min.   :0.000000    Min.   :0.000e+00    Min.   :0.0000000
##  1st Qu.:0.000000    1st Qu.:0.000e+00    1st Qu.:0.0000000
##  Median :0.000000    Median :0.000e+00    Median :0.0000000
##  Mean   :0.001901    Mean   :3.313e-06    Mean   :0.0001595
##  3rd Qu.:0.003000    3rd Qu.:0.000e+00    3rd Qu.:0.0000000
##  Max.   :0.015000    Max.   :1.000e-03    Max.   :0.0050000
##  abnormal_short_term_variability mean_value_of_short_term_variability
##  Min.   :12.00                   Min.   :0.200
##  1st Qu.:32.00                   1st Qu.:0.700
##  Median :49.00                   Median :1.200
##  Mean   :46.99                   Mean   :1.335
##  3rd Qu.:61.00                   3rd Qu.:1.700
##  Max.   :87.00                   Max.   :7.000
##  percentage_of_time_with_abnormal_long_term_variability
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   : 9.795
##  3rd Qu.:11.000
##  Max.   :91.000
##  mean_value_of_long_term_variability histogram_width  histogram_min
##  Min.   : 0.000                      Min.   :  3.00   Min.   : 50.00
##  1st Qu.: 4.600                      1st Qu.: 37.00   1st Qu.: 67.00
##  Median : 7.400                      Median : 68.00   Median : 93.00
##  Mean   : 8.167                      Mean   : 70.54   Mean   : 93.56
##  3rd Qu.:10.800                      3rd Qu.:100.00   3rd Qu.:120.00
##  Max.   :50.700                      Max.   :180.00   Max.   :159.00
##  histogram_max   histogram_number_of_peaks histogram_number_of_zeroes
##  Min.   :122.0   Min.   : 0.000            Min.   : 0.0000
##  1st Qu.:152.0   1st Qu.: 2.000            1st Qu.: 0.0000
##  Median :162.0   Median : 4.000            Median : 0.0000  
##  Mean   :164.1   Mean   : 4.077            Mean   : 0.3256  
##  3rd Qu.:174.0   3rd Qu.: 6.000            3rd Qu.: 0.0000  
##  Max.   :238.0   Max.   :18.000            Max.   :10.0000  
##  histogram_mode  histogram_mean  histogram_median histogram_variance
##  Min.   : 60.0   Min.   : 73.0   Min.   : 77.0    Min.   :  0.00    
##  1st Qu.:129.0   1st Qu.:125.0   1st Qu.:129.0    1st Qu.:  2.00    
##  Median :139.0   Median :136.0   Median :139.0    Median :  7.00    
##  Mean   :137.5   Mean   :134.6   Mean   :138.1    Mean   : 18.91    
##  3rd Qu.:148.0   3rd Qu.:145.0   3rd Qu.:148.0    3rd Qu.: 24.00    
##  Max.   :187.0   Max.   :182.0   Max.   :186.0    Max.   :269.00    
##  histogram_tendency fetal_health
##  Min.   :-1.0000    1:1646      
##  1st Qu.: 0.0000    2: 292      
##  Median : 0.0000    3: 175      
##  Mean   : 0.3185                
##  3rd Qu.: 1.0000                
##  Max.   : 1.0000
```

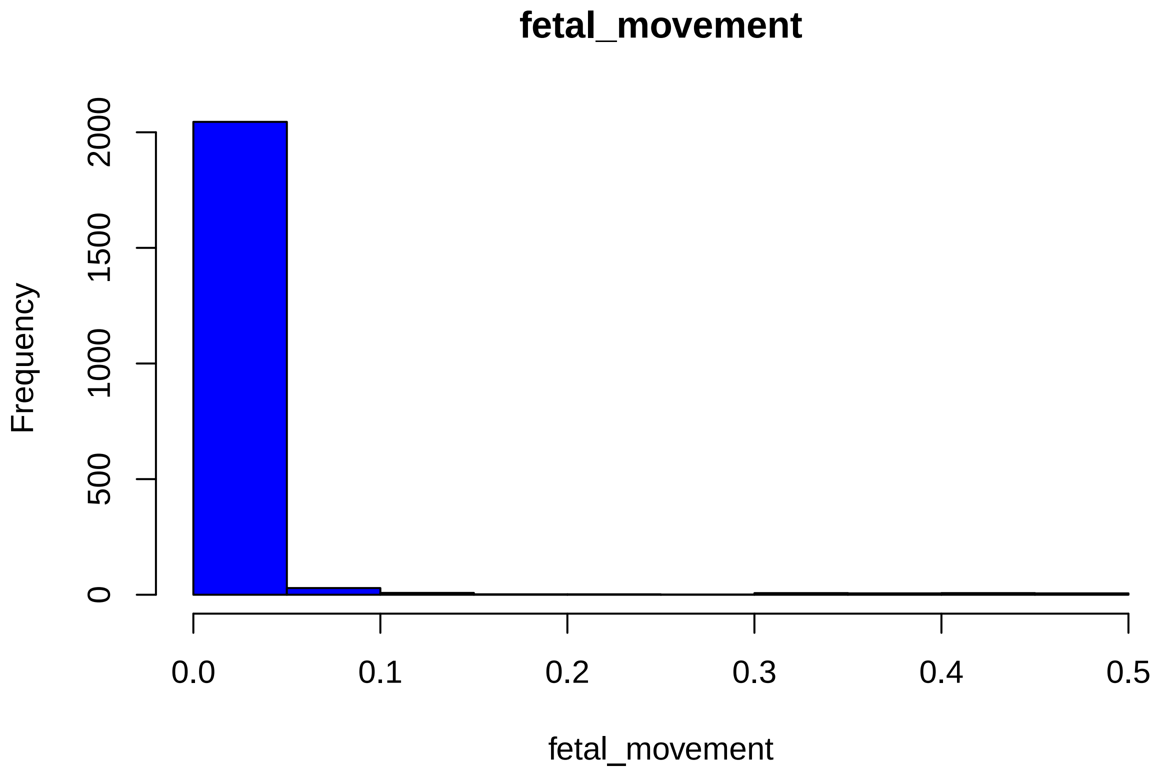
EDA

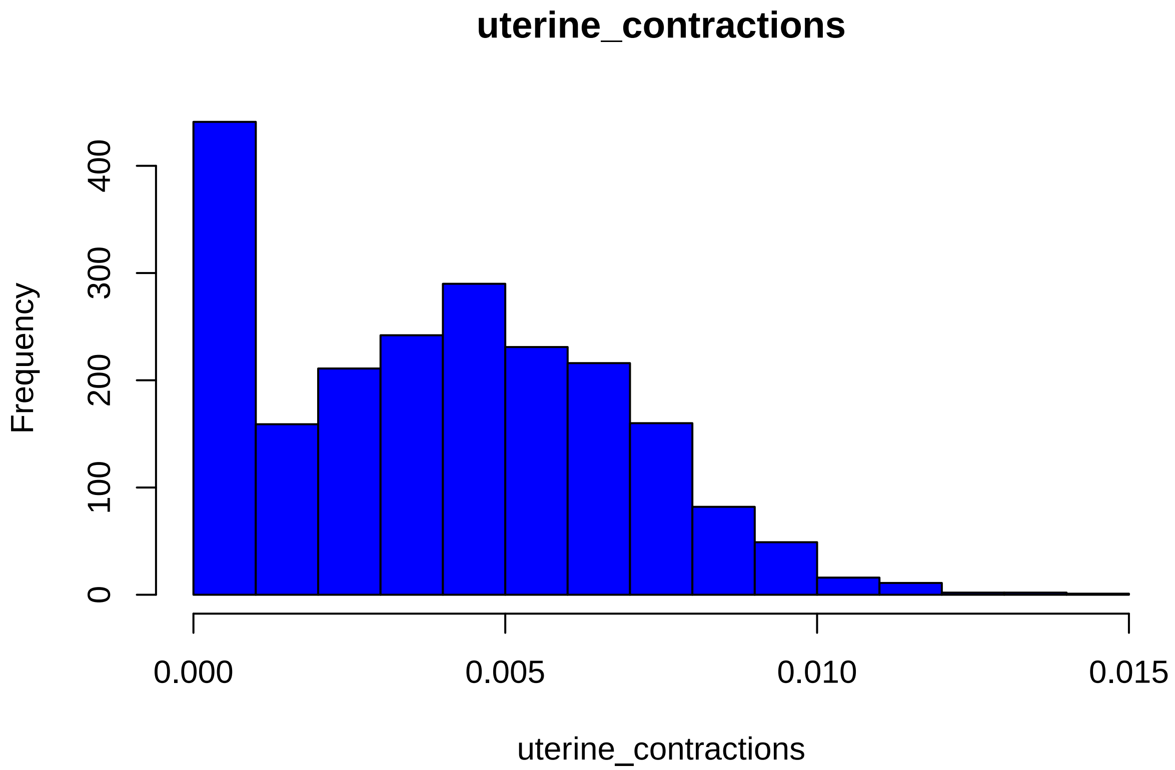
1. **Numerical Variables**

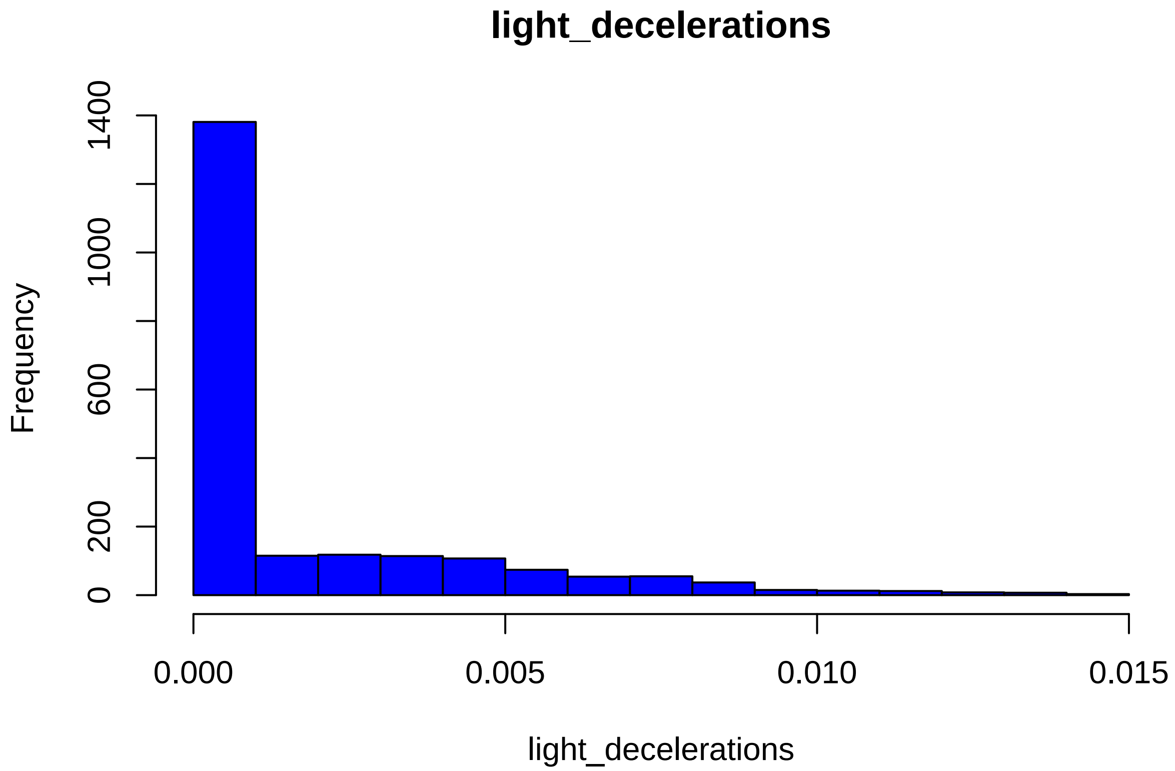
```
#Non-histogram features

cols1 <- colnames(data_2[1:11])

for(i in 1 : length(cols1)){
  hist(data_2[,i], main = cols1[i], xlab = cols1[i], col = 'blue')
}
```

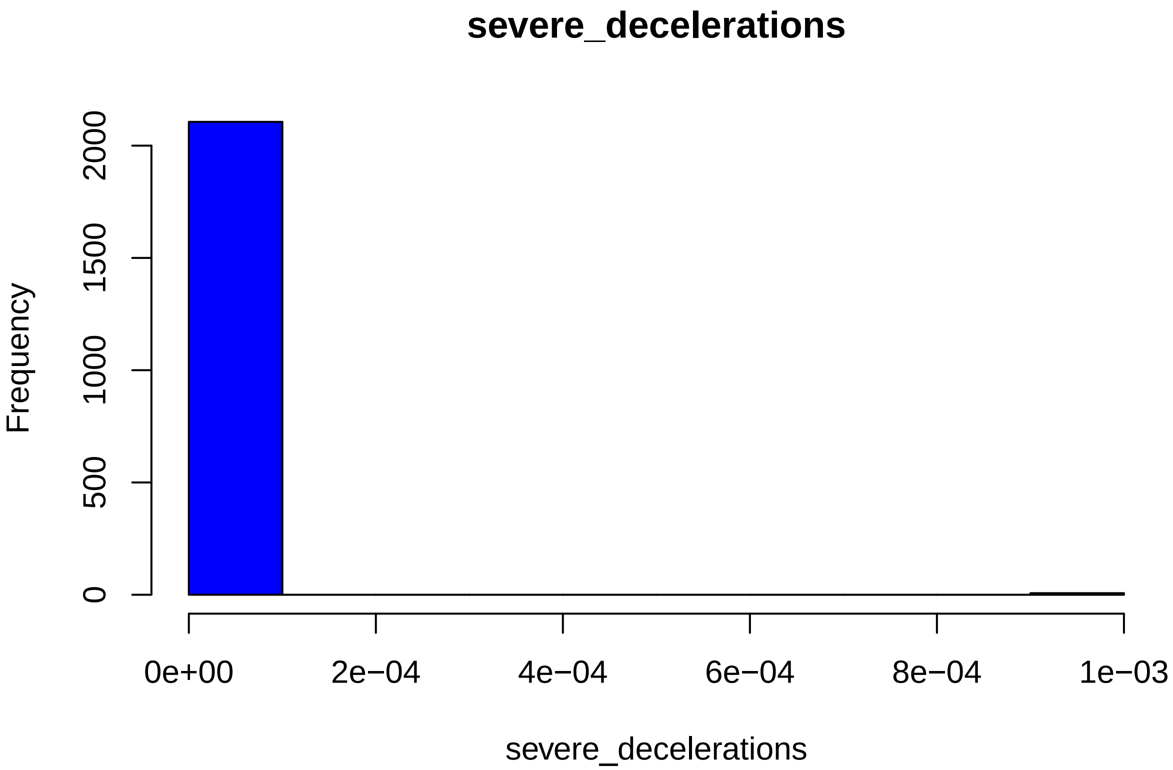

baseline_value
Frequency
0
100
200
300
400
110
120
130
140
150
160
baseline_value

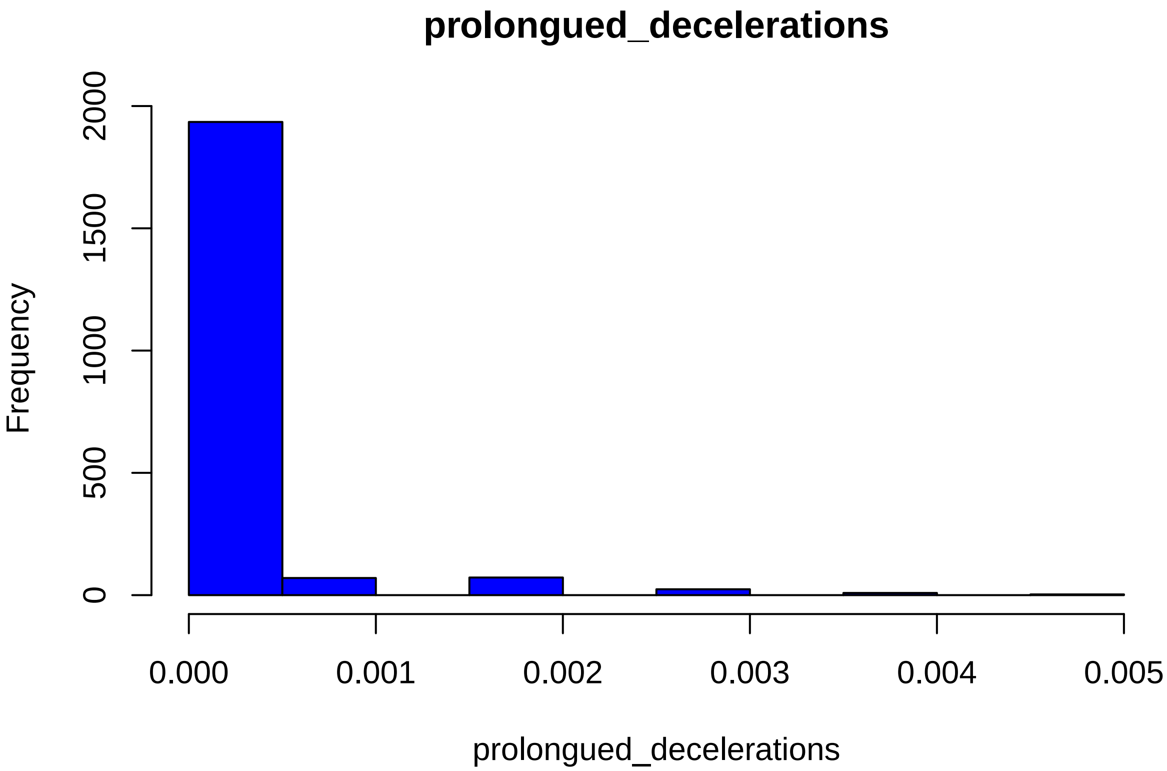

accelerations
Frequency
0
200
600
1000
0.000
0.005
0.010
0.015
0.020
accelerations

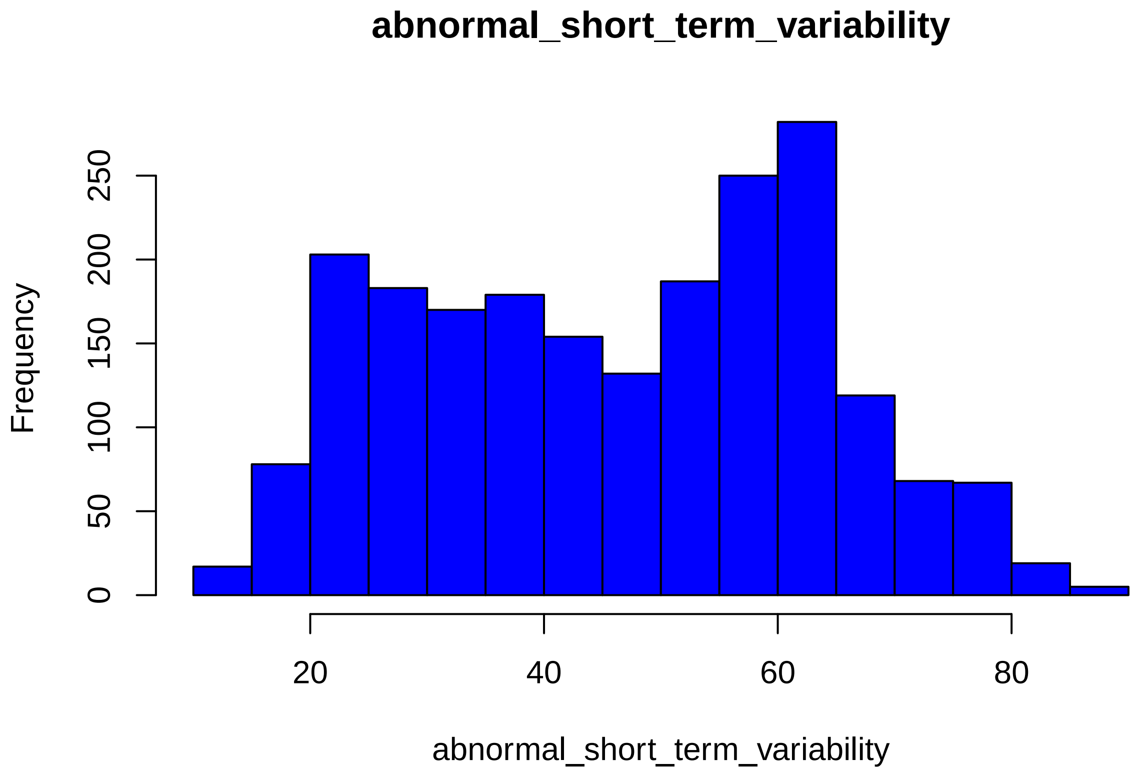

fetal_movement
Frequency
0
500
1000
1500
2000
0.0
0.1
0.2
0.3
0.4
0.5
fetal_movement

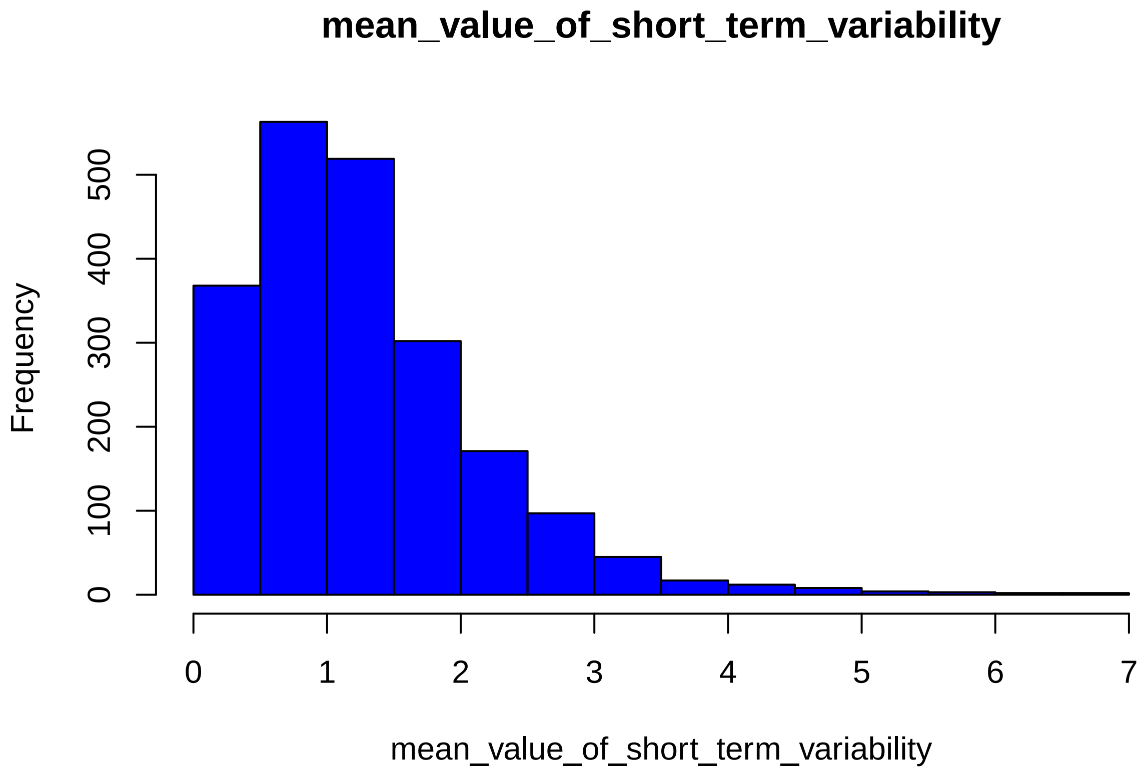

uterine_contractions
Frequency
0
100
200
300
400
0.000
0.005
0.010
0.015
uterine_contractions

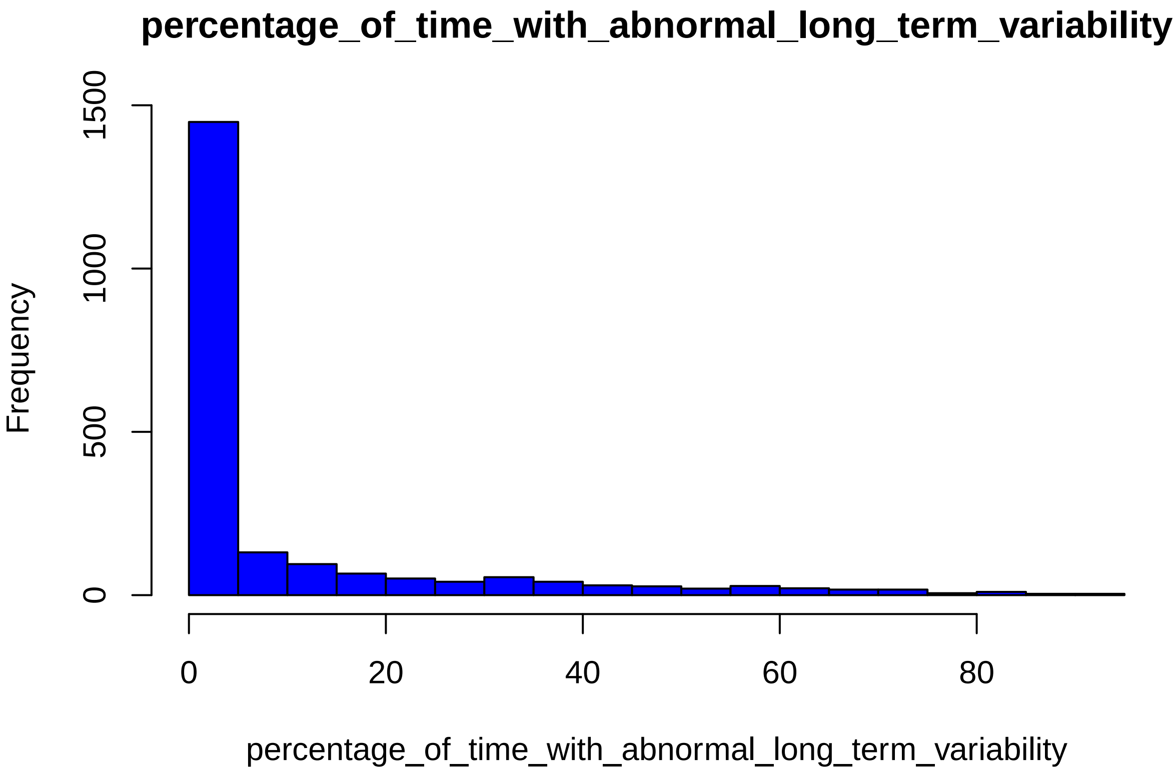

light_decelerations
Frequency
0
200
600
1000
1400
0.000
0.005
0.010
0.015
light_decelerations

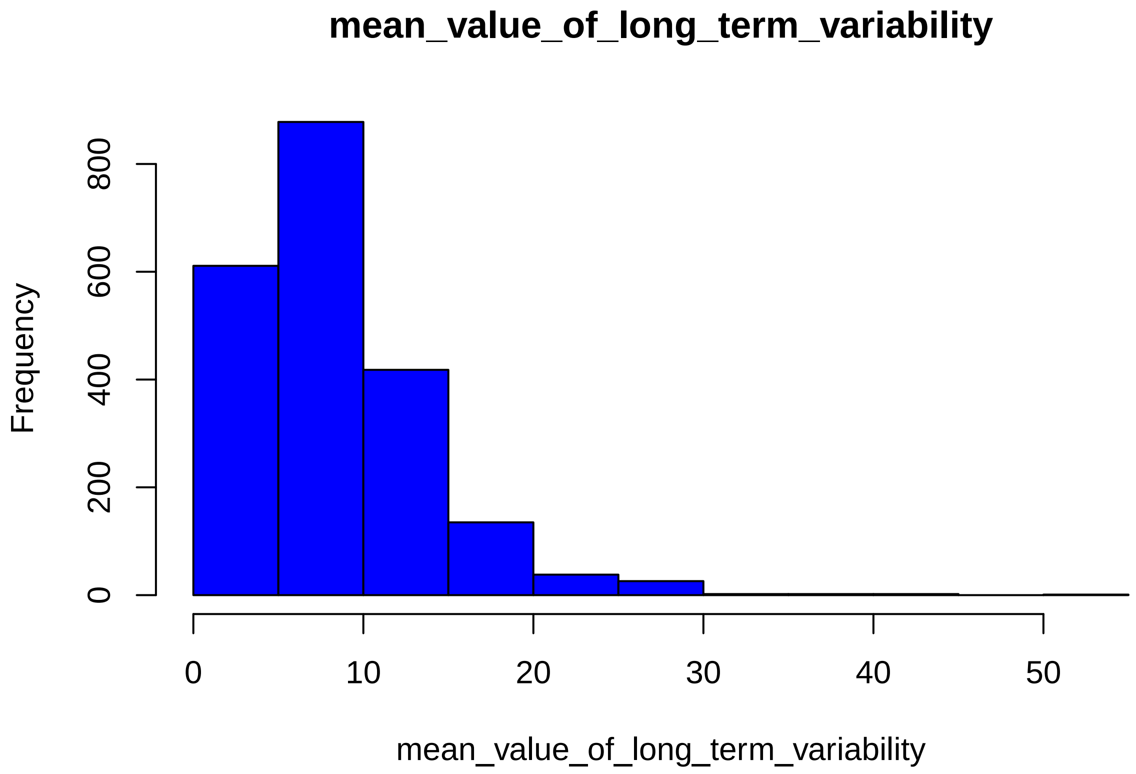

severe_decelerations
Frequency
0
500
1000
1500
2000
0e+00
2e−04
4e−04
6e−04
8e−04
1e−03
severe_decelerations


prolongued_decelerations
Frequency
0
500
1000
1500
2000
0.000
0.001
0.002
0.003
0.004
0.005
prolongued_decelerations


abnormal_short_term_variability
Frequency
0
50
100
150
200
250
20
40
60
80
abnormal_short_term_variability

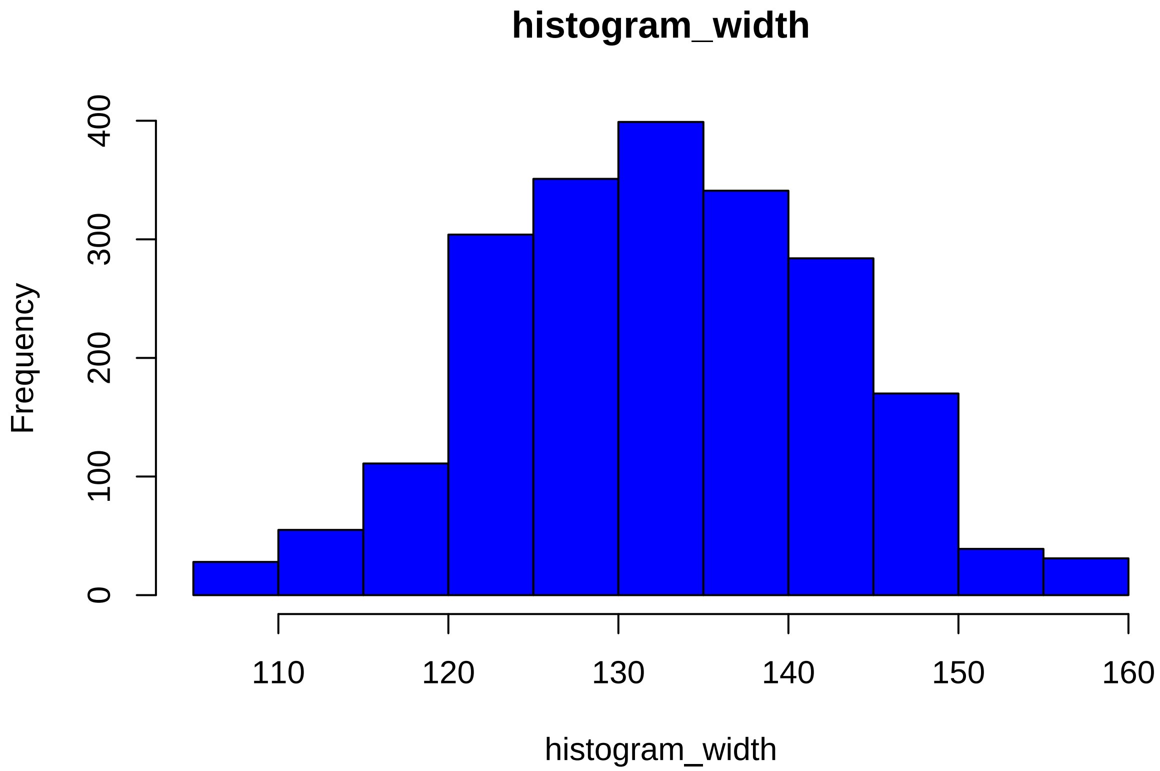

mean_value_of_short_term_variability
Frequency
0
100
200
300
400
500
0
1
2
3
4
5
6
7
mean_value_of_short_term_variability

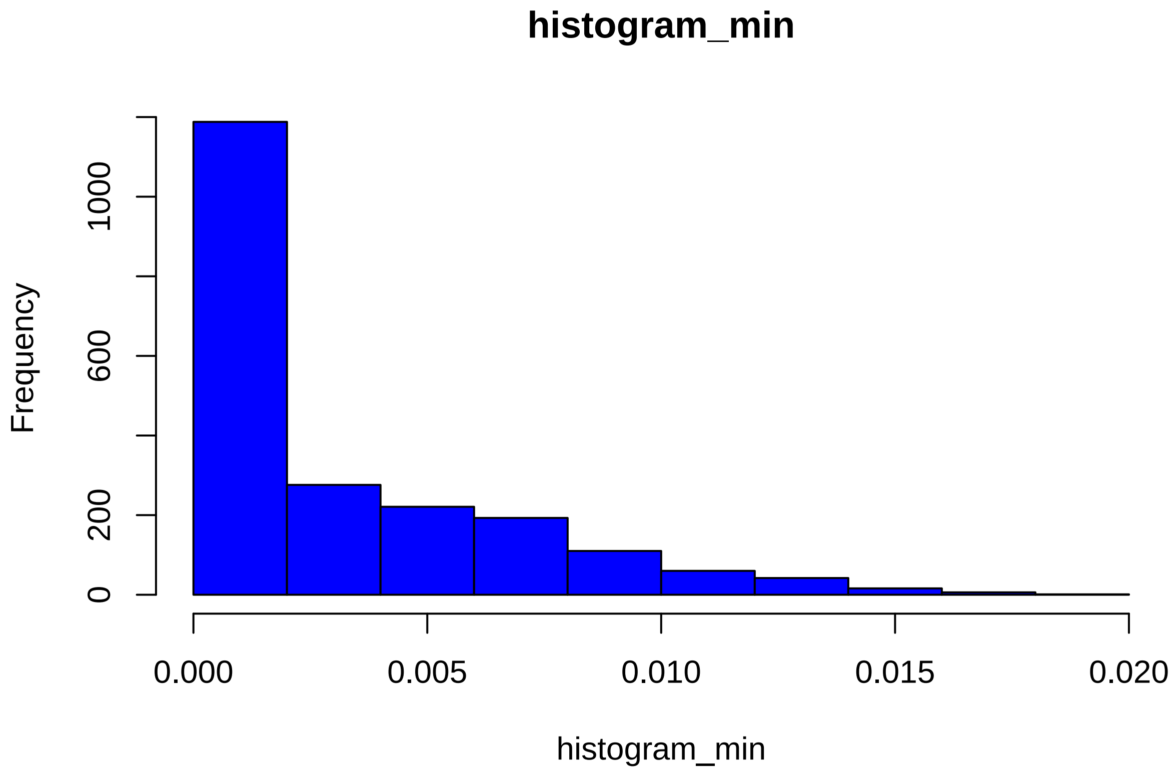

percentage_of_time_with_abnormal_long_term_variability
Frequency
0
500
1000
1500
0
20
40
60
80
percentage_of_time_with_abnormal_long_term_variability

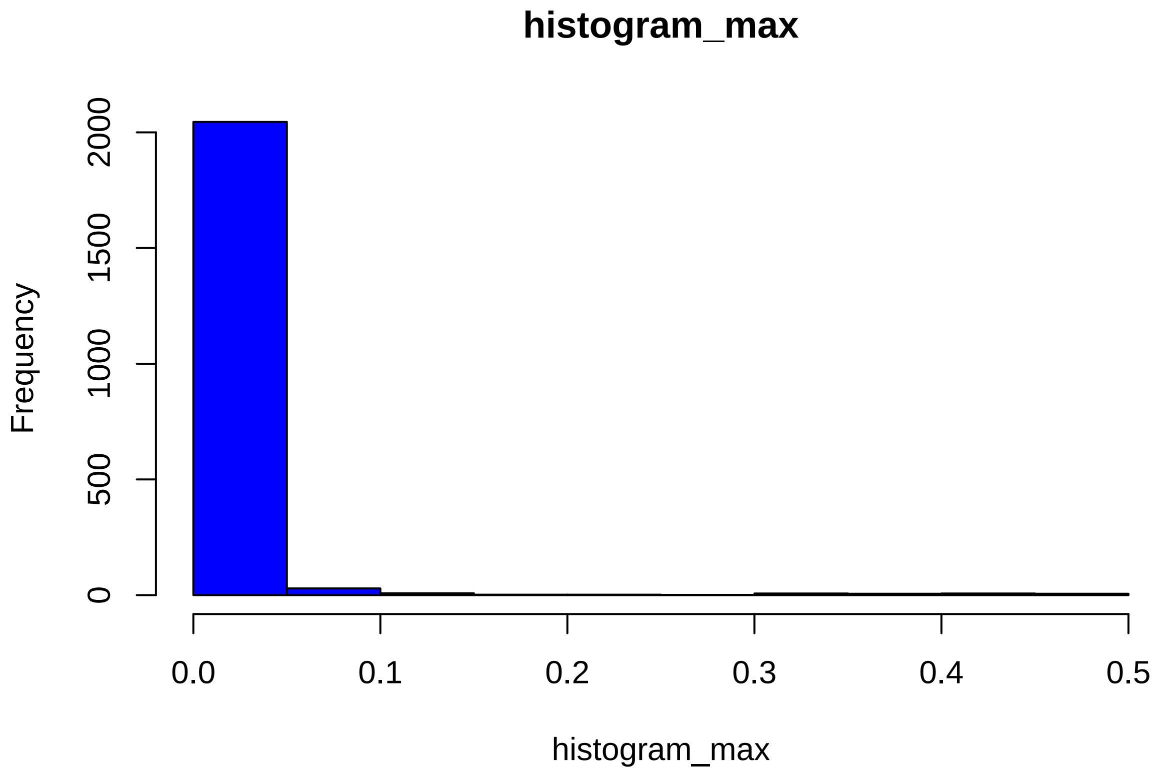


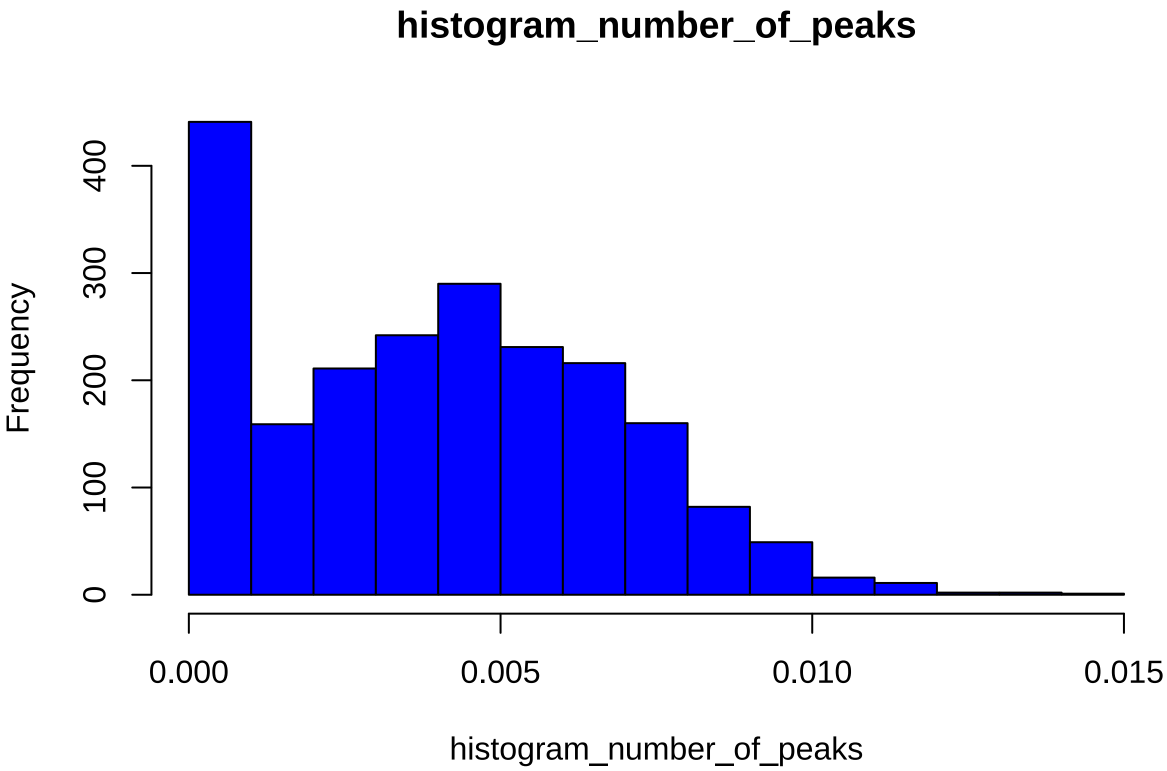
1. Baseline_value appears to be almost normally distributed.
2. Most of the variables like acceleration, light_decelaration , uterine_contractions are right skewed, meaning most of the values of the variables are concentrated on the lower end, and fewer values are at the higher end.

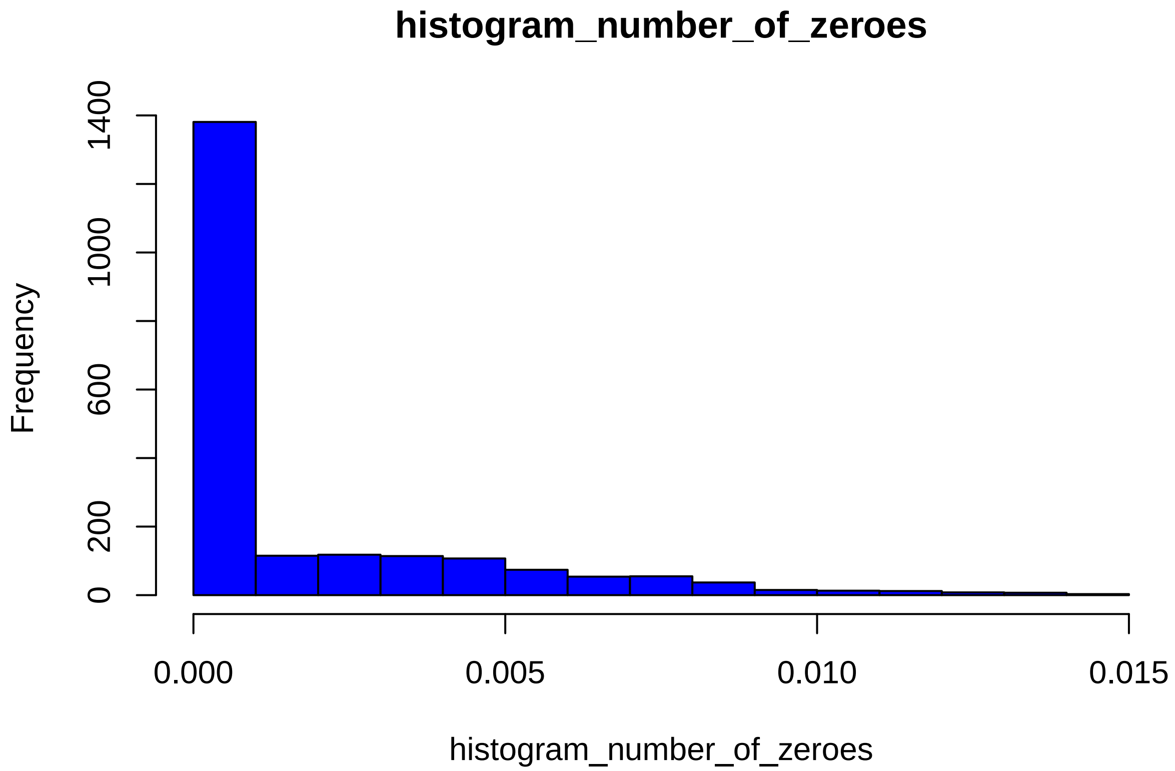
```r
#Histogram features
cols2 <- colnames(data[12:21])

for(i in 1 : length(cols2)){
  hist(data_2[,i], main = cols2[i], xlab = cols2[i], col = 'blue')
}
```

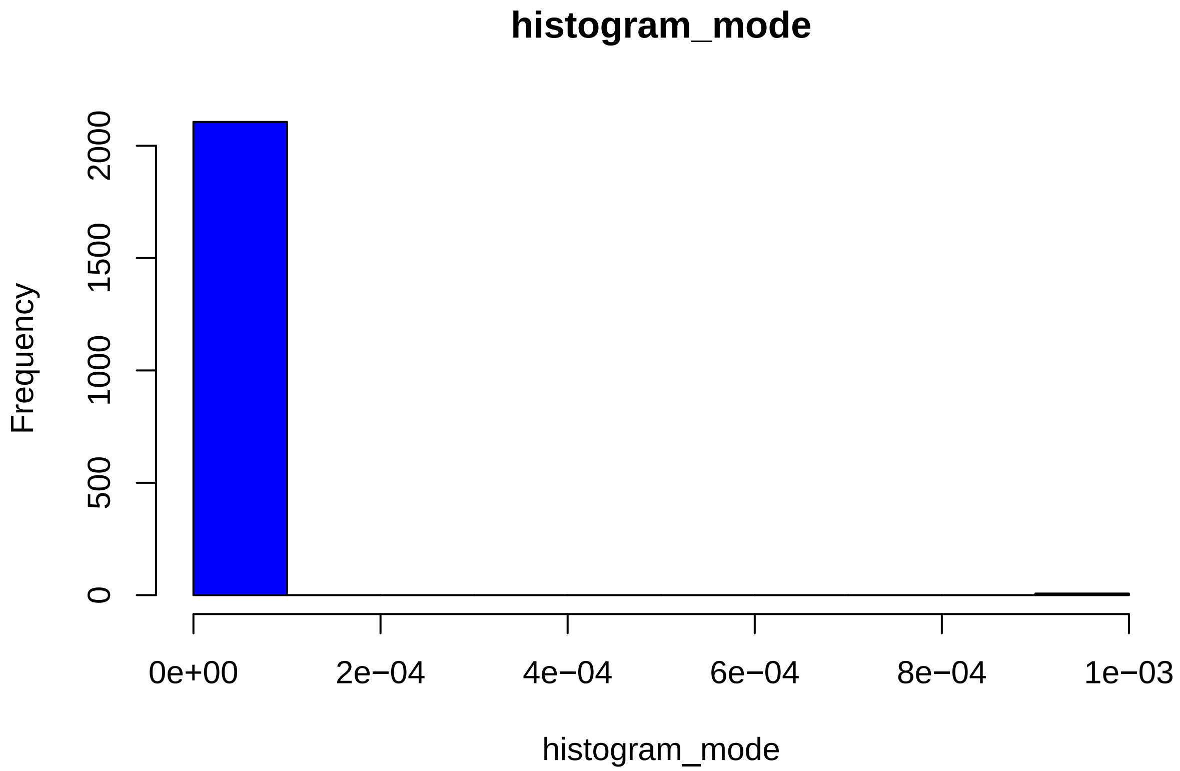

histogram_width
Frequency
0
100
200
300
400
110
120
130
140
150
160
histogram_width

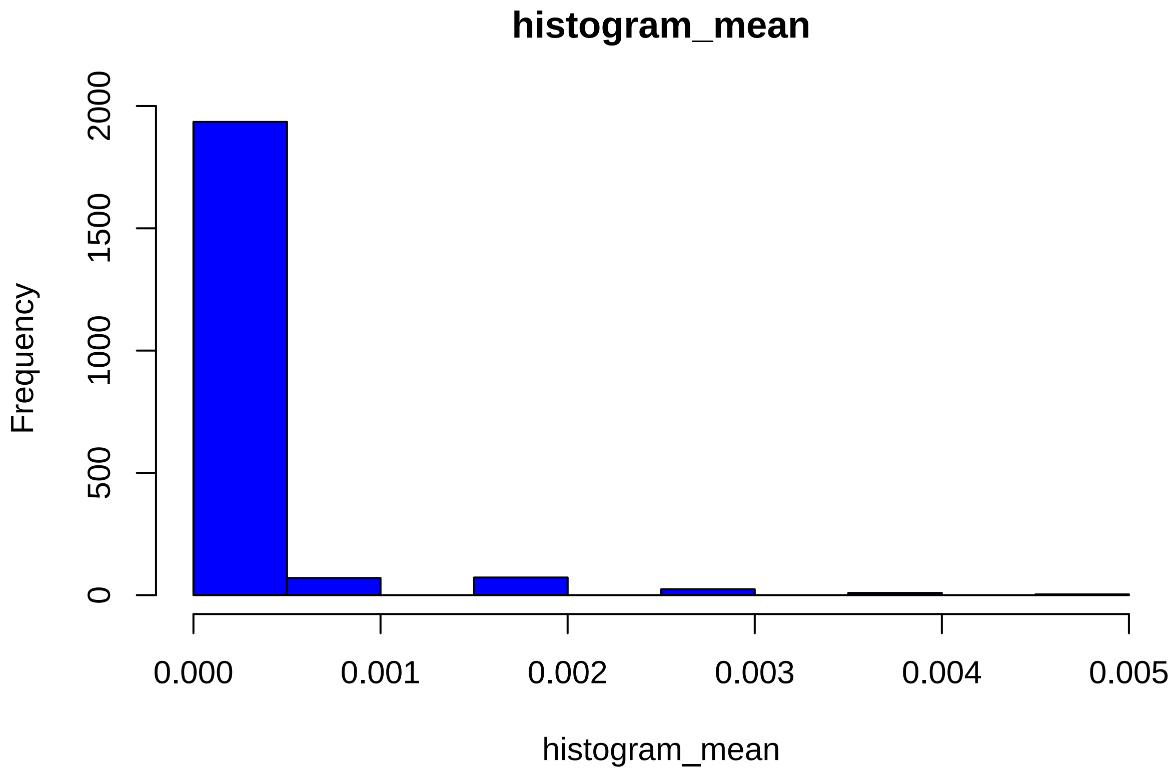

histogram_min
Frequency
0
200
600
1000
0.000
0.005
0.010
0.015
0.020
histogram_min

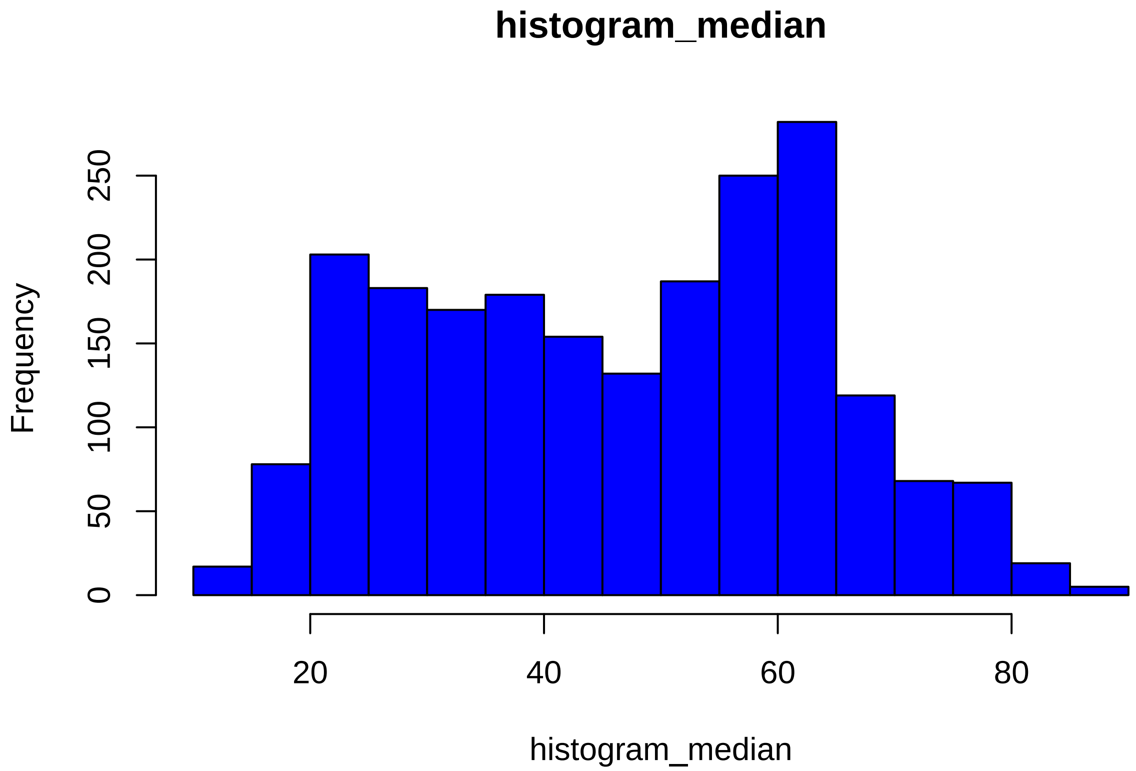

histogram_max
Frequency
0
500
1000
1500
2000
0.0
0.1
0.2
0.3
0.4
0.5
histogram_max

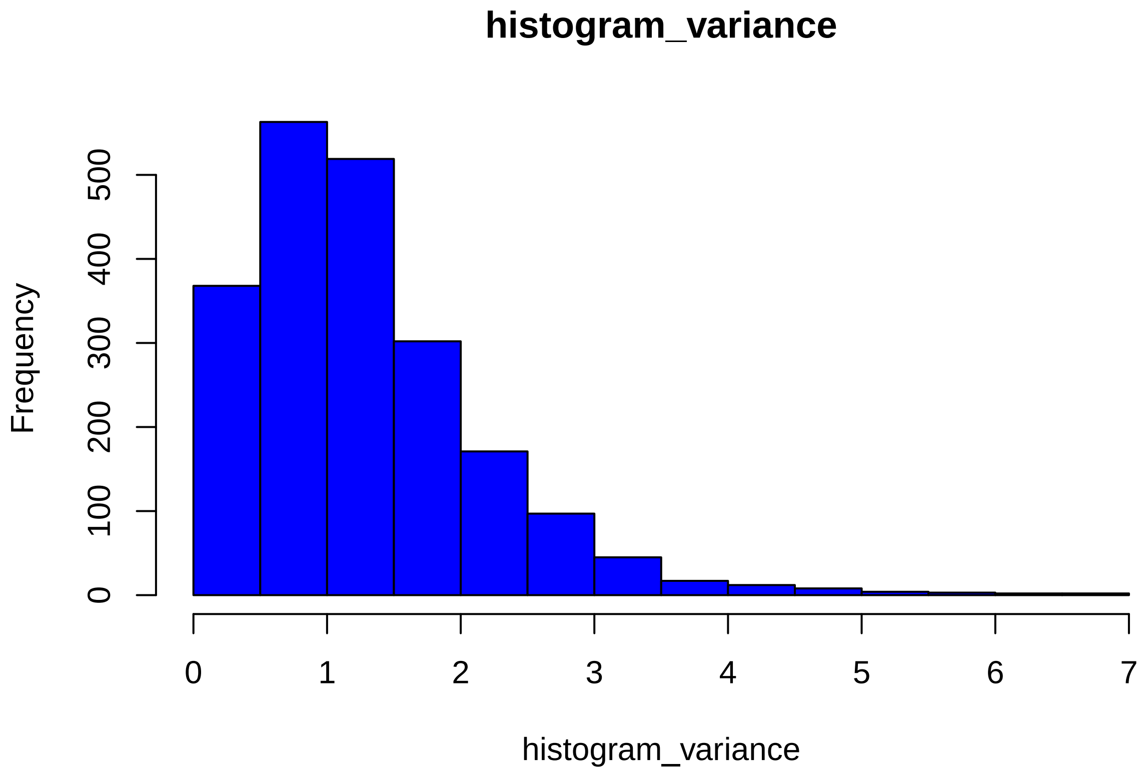

histogram_number_of_peaks
Frequency
0
100
200
300
400
0.000
0.005
0.010
0.015
histogram_number_of_peaks

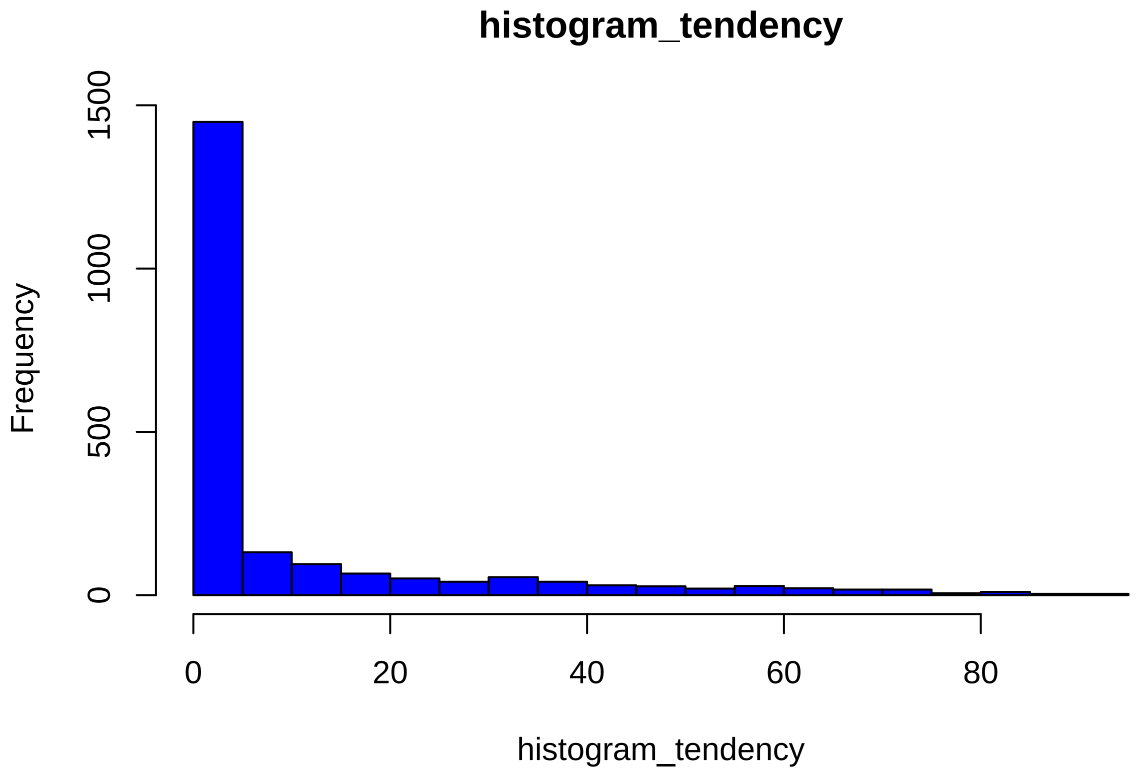

histogram_number_of_zeroes
Frequency
0
200
600
1000
1400
0.000
0.005
0.010
0.015
histogram_number_of_zeroes


histogram_mode
Frequency
0
500
1000
1500
2000
0e+00
2e−04
4e−04
6e−04
8e−04
1e−03
histogram_mode


histogram_mean
Frequency
0
500
1000
1500
2000
0.000
0.001
0.002
0.003
0.004
0.005
histogram_mean


histogram_median
Frequency
0
50
100
150
200
250
20
40
60
80
histogram_median


histogram_variance
Frequency
0
100
200
300
400
500
0
1
2
3
4
5
6
7
histogram_variance

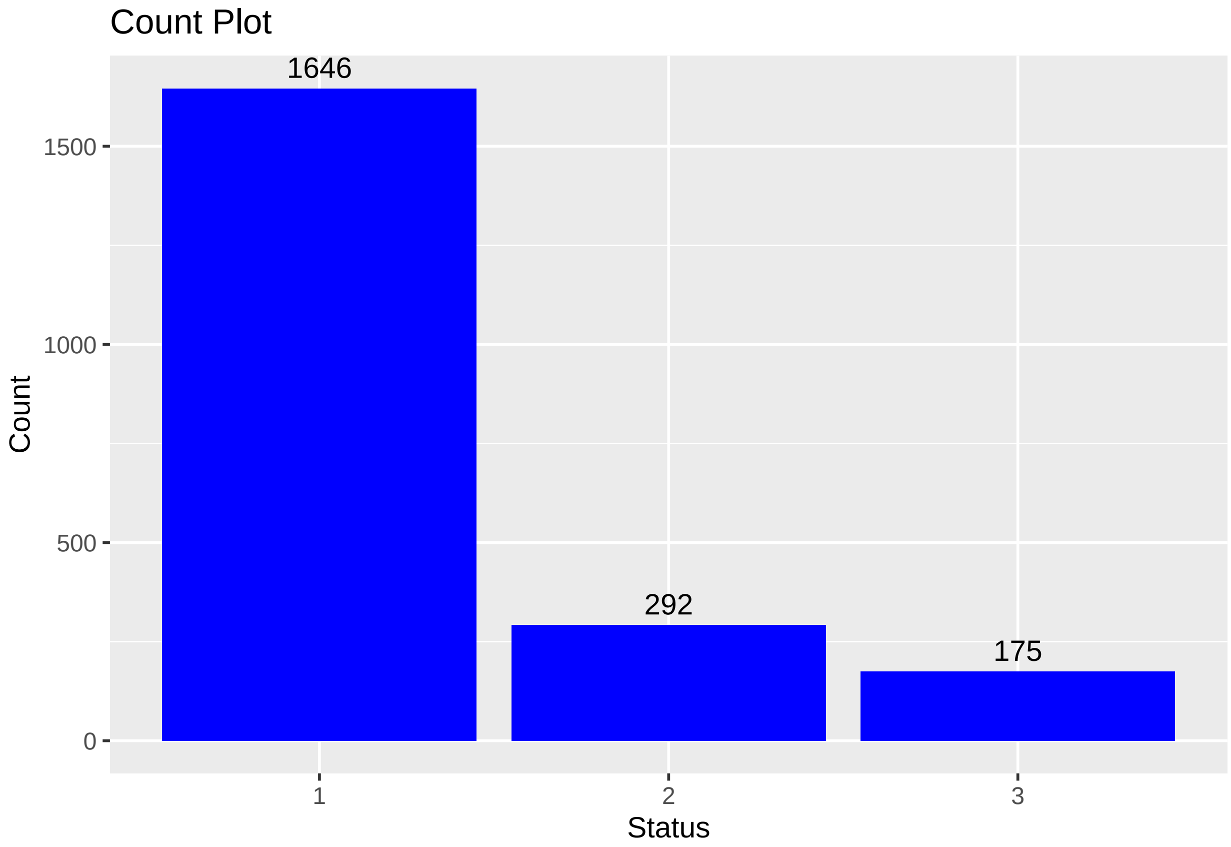


Most of the variables are right skewed except histogram_width which appears normally distributed and histogram_median which shows a somewhat bimodal distribution, with most values concentrated between 20 and 70, peaking near 60. There are fewer extreme values beyond this central range.

2. **Target Variable - Fetal Movement**

```
ggplot(data_2, aes(x = fetal_health)) +
  geom_bar(fill = 'blue') +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
  labs(title = "Count Plot", x = "Status", y = "Count")
```

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



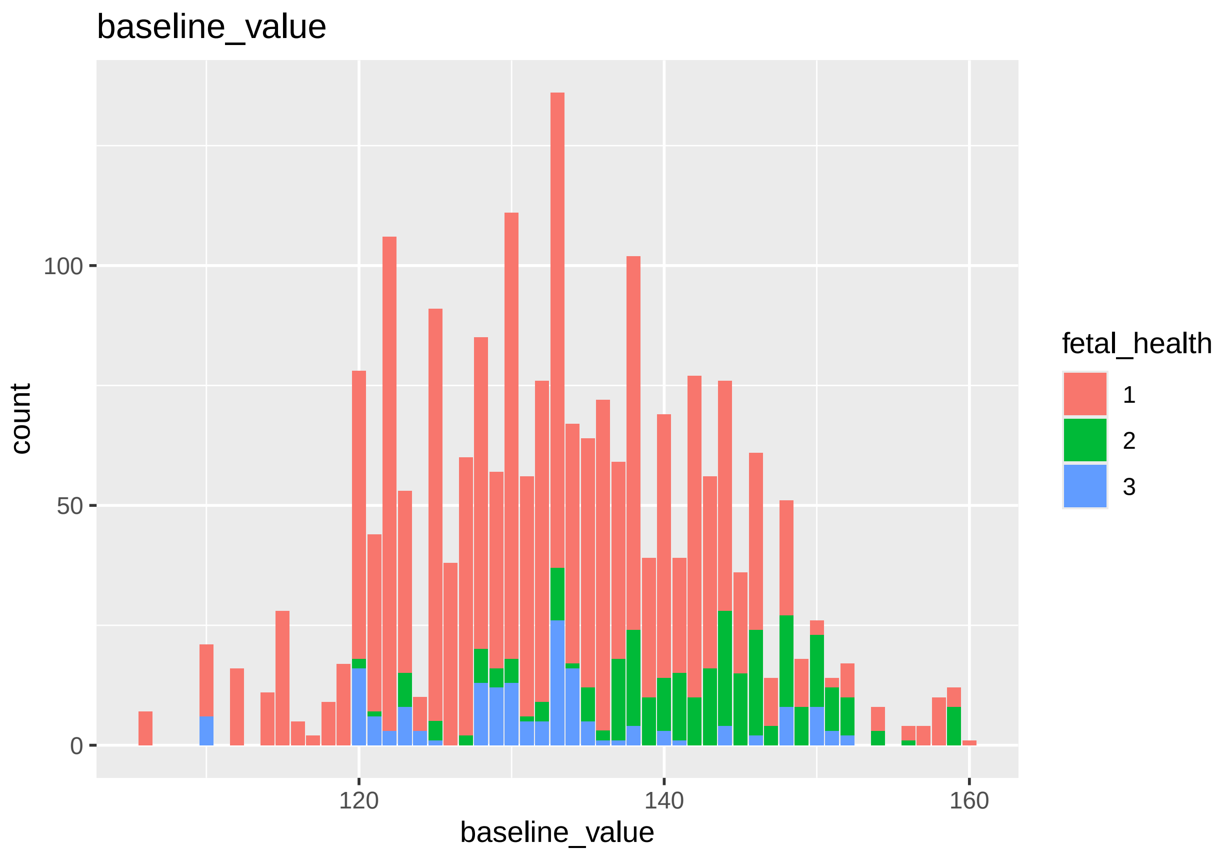
Our data is highly imbalanced with most of our subject's having the fetal_health status as Normal.

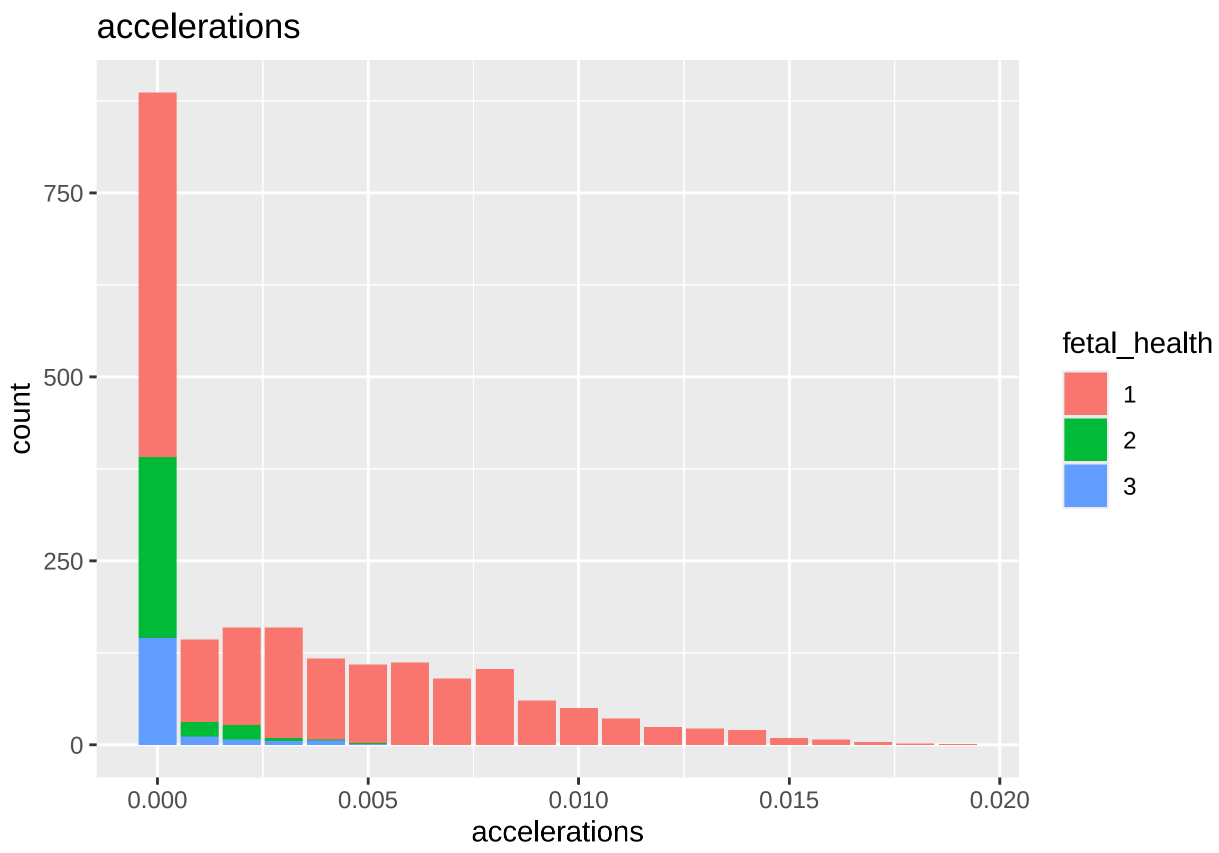
3. **Numerical Variables vs Fetal_Health**

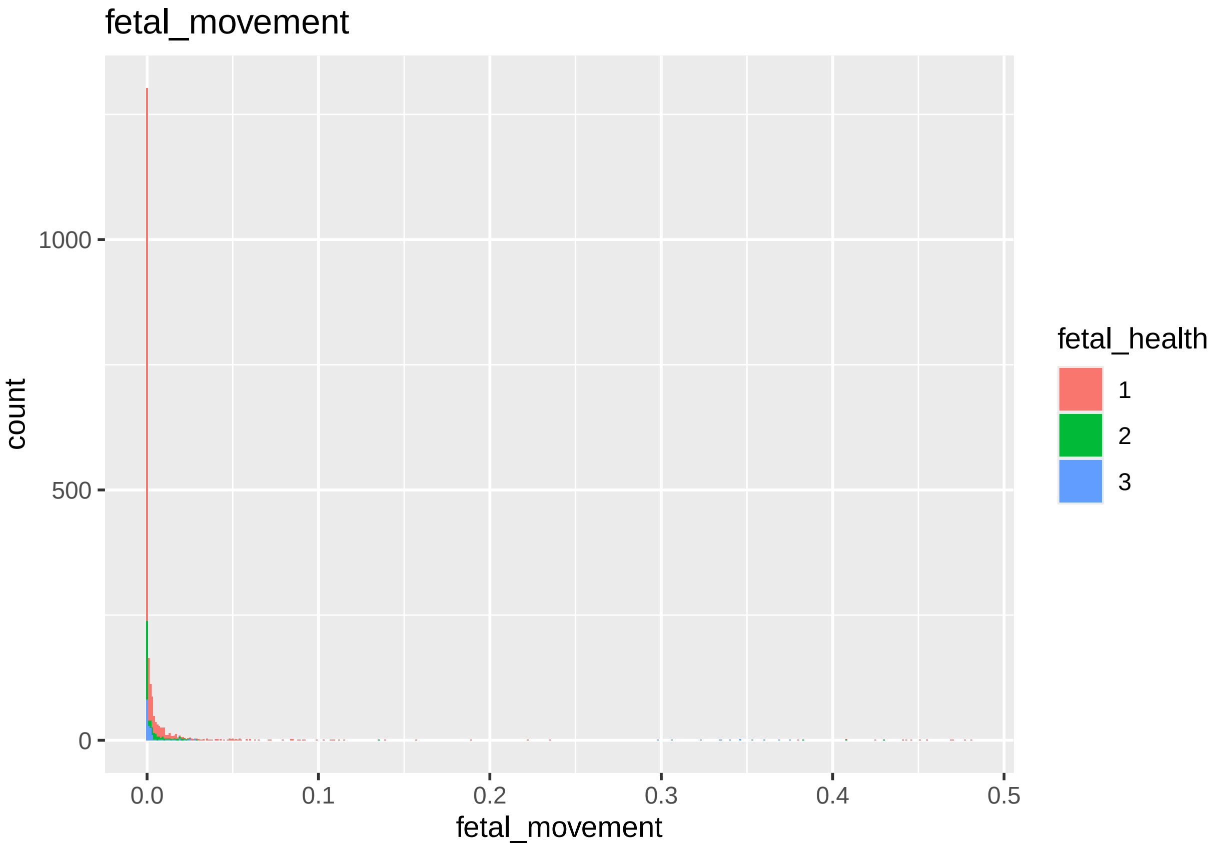
- Non-histogram Features

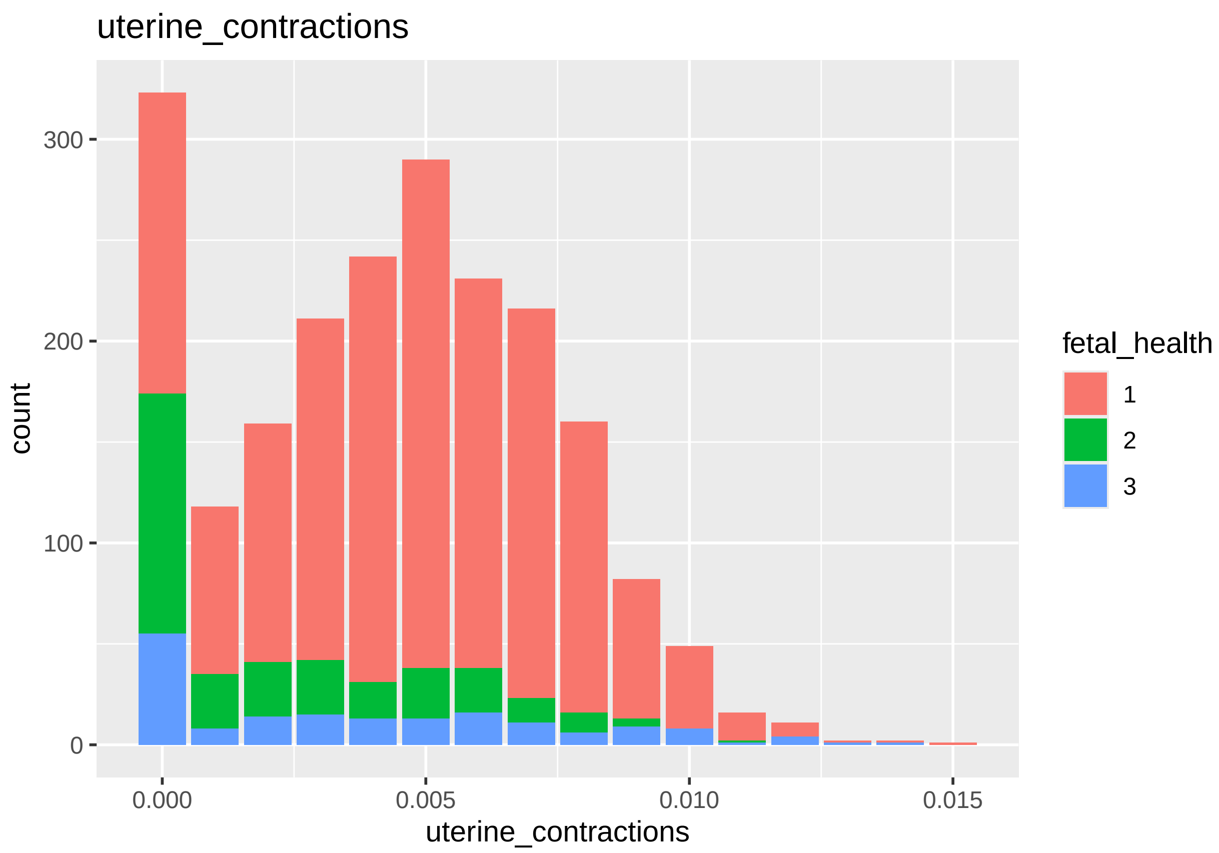
```
#Non-histogram Features

for(i in 1 : length(cols1)){
    plot <- ggplot(data_2, aes_string(x = cols1[i], fill = "fetal_health")) +
    geom_bar() +
    ggtitle(cols1[i])

    print(plot)
  #hist(data_2[,i], main = cols1[i], xlab = cols1[i], col = 'blue')
}
```

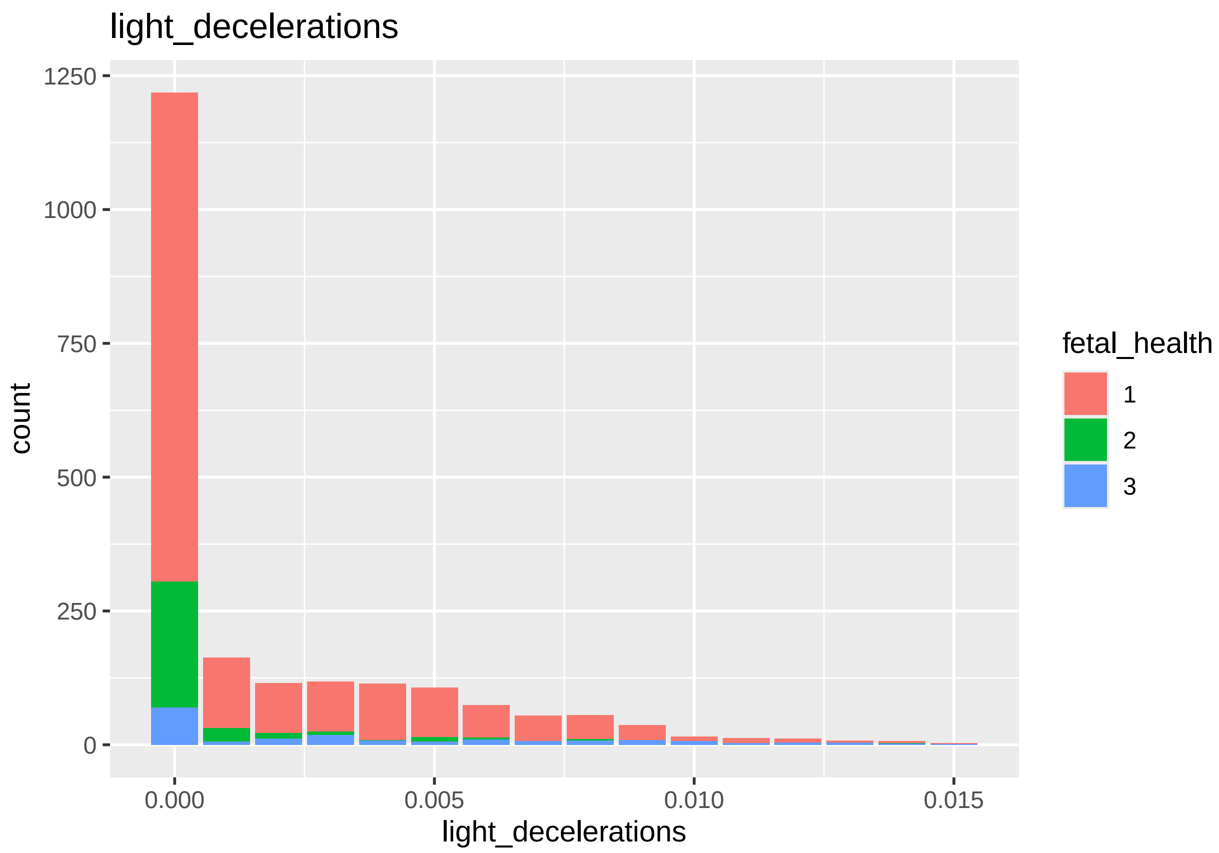
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

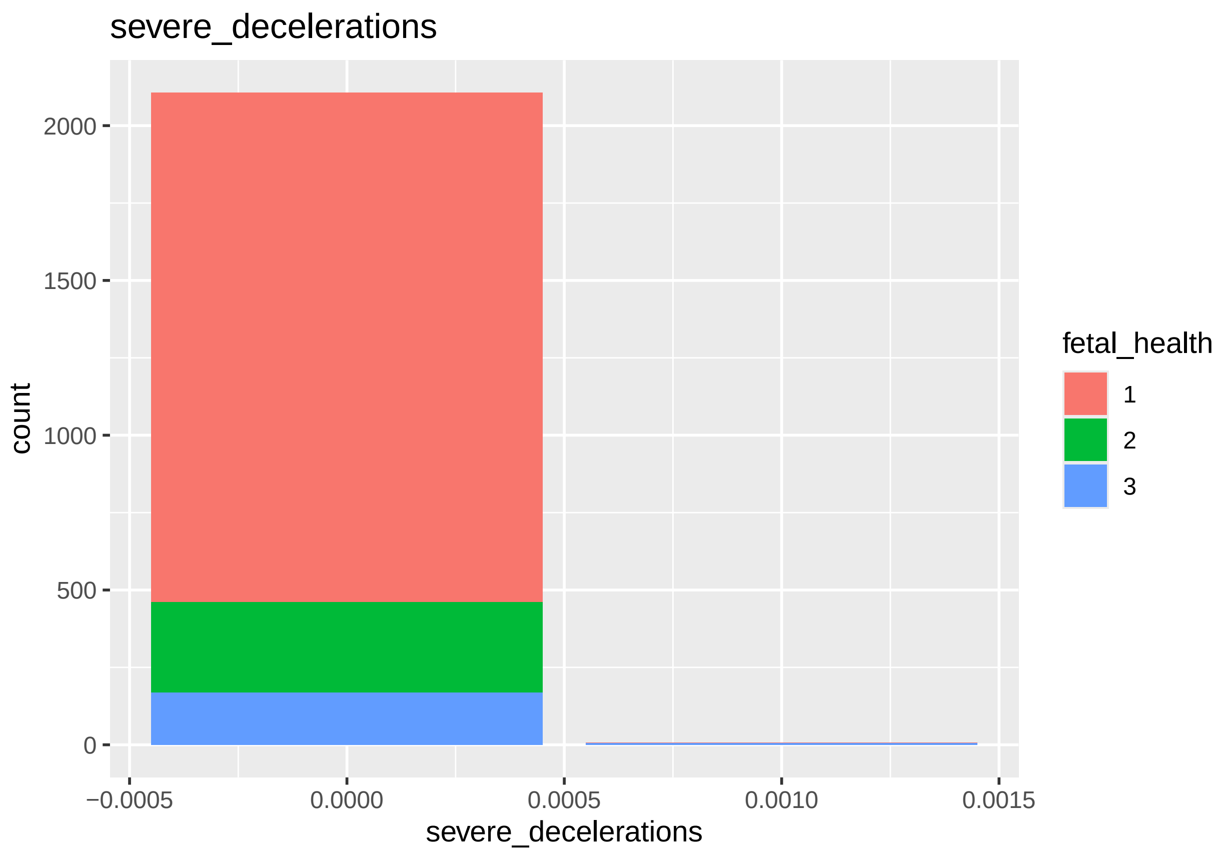

baseline_value
count
100
50
0
120
140
160
baseline_value
fetal_health
1
2
3

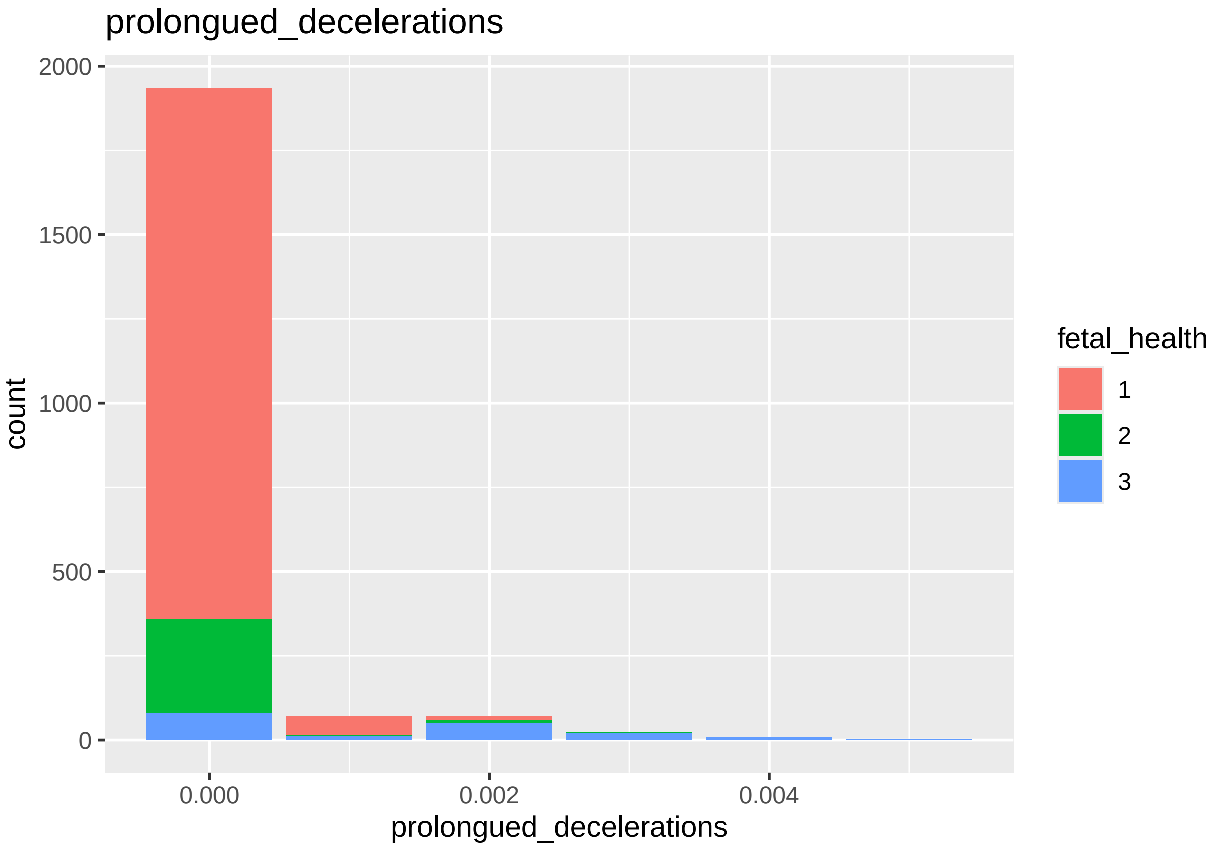

accelerations
750
500
250
0
count
0.000
0.005
0.010
0.015
0.020
accelerations
fetal_health
1
2
3

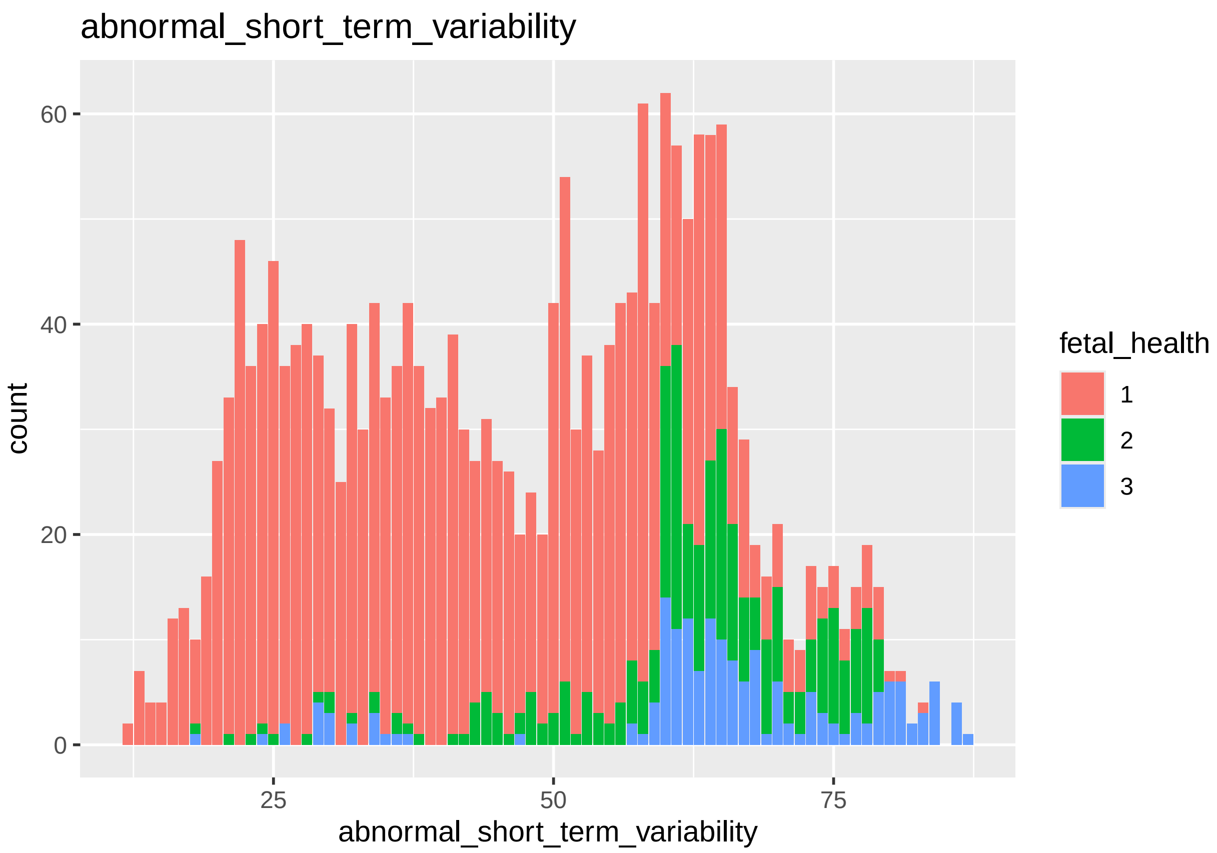

fetal_movement
count
1000
500
0
0.0
0.1
0.2
0.3
0.4
0.5
fetal_movement
fetal_health
1
2
3

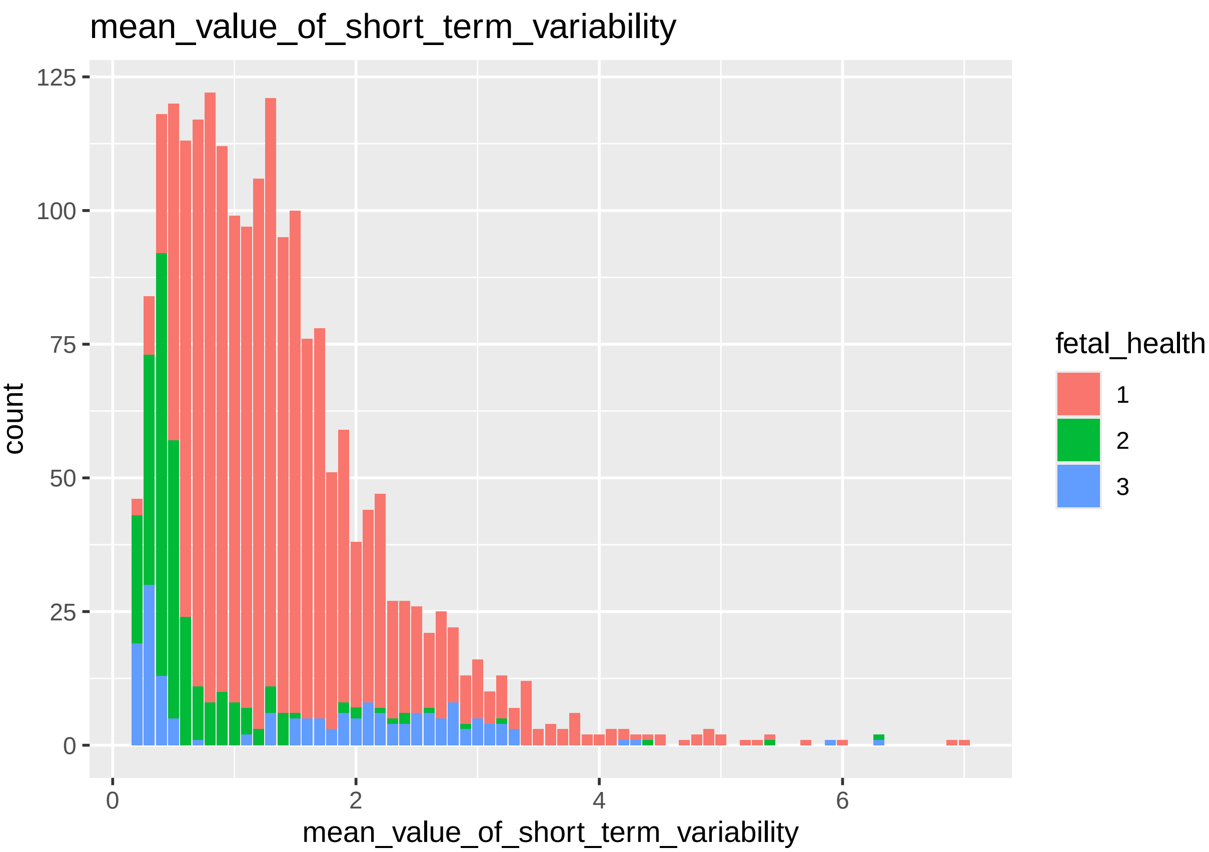

uterine_contractions
count
300
200
100
0
0.000
0.005
0.010
0.015
uterine_contractions
fetal_health
1
2
3

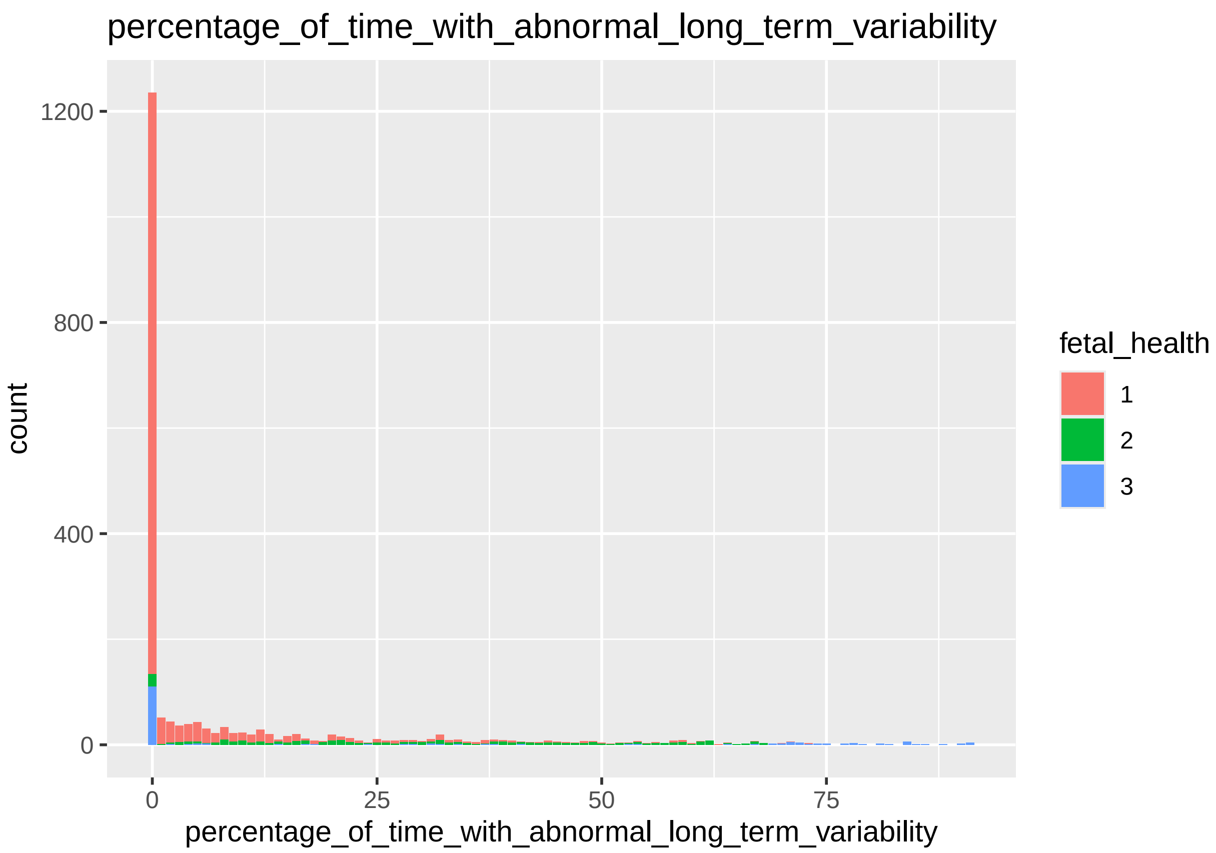

light_decelerations
1250
1000
750
500
250
0
count
0.000
0.005
0.010
0.015
light_decelerations
fetal_health
1
2
3

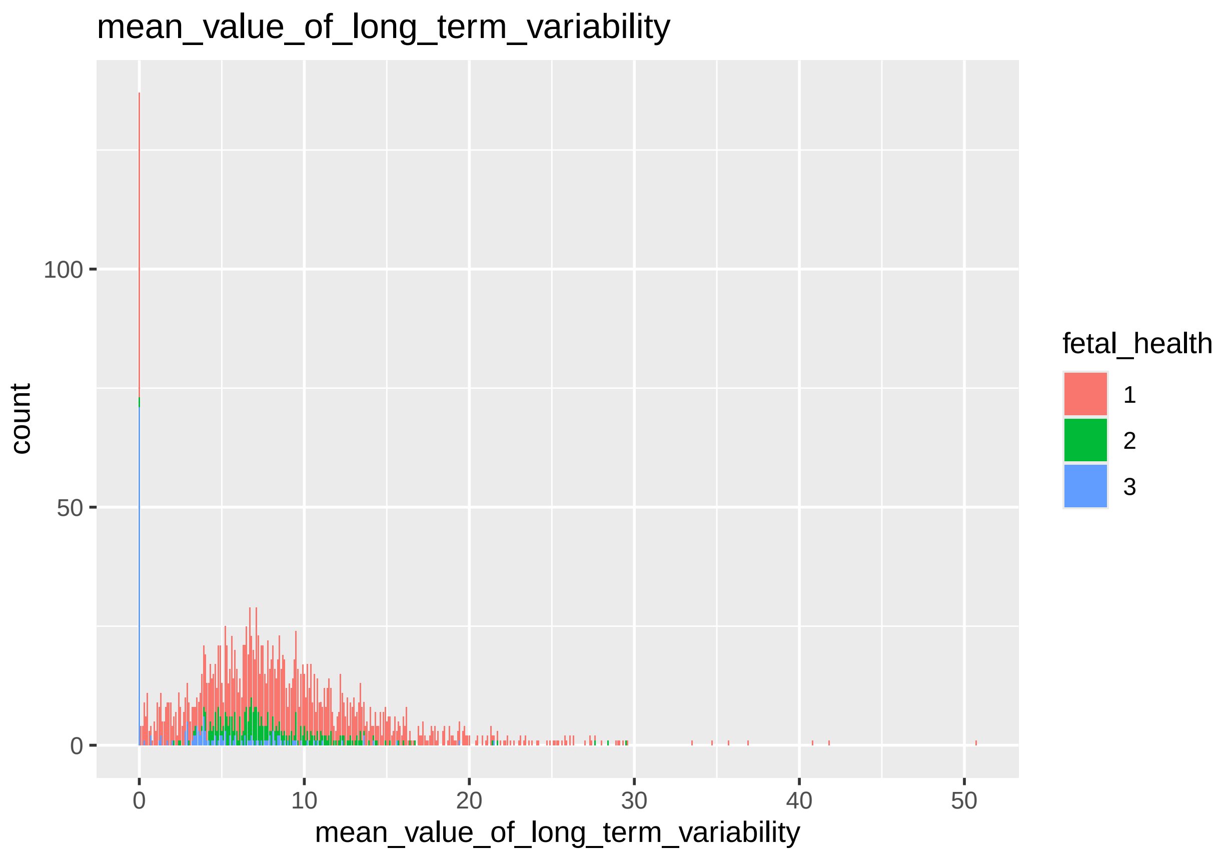

severe_decelerations
count
2000
1500
1000
500
0
−0.0005
0.0000
0.0005
0.0010
0.0015
severe_decelerations
fetal_health
1
2
3


prolongued_decelerations
2000
1500
1000
500
0
count
0.000
0.002
0.004
prolongued_decelerations
fetal_health
1
2
3


abnormal_short_term_variability
count
0
20
40
60
25
50
75
abnormal_short_term_variability
fetal_health
1
2
3


mean_value_of_short_term_variability
count
125
100
75
50
25
0
0
2
4
6
mean_value_of_short_term_variability
fetal_health
1
2
3


percentage_of_time_with_abnormal_long_term_variability
count
1200
800
400
0
0
25
50
75
percentage_of_time_with_abnormal_long_term_variability
fetal_health
1
2
3

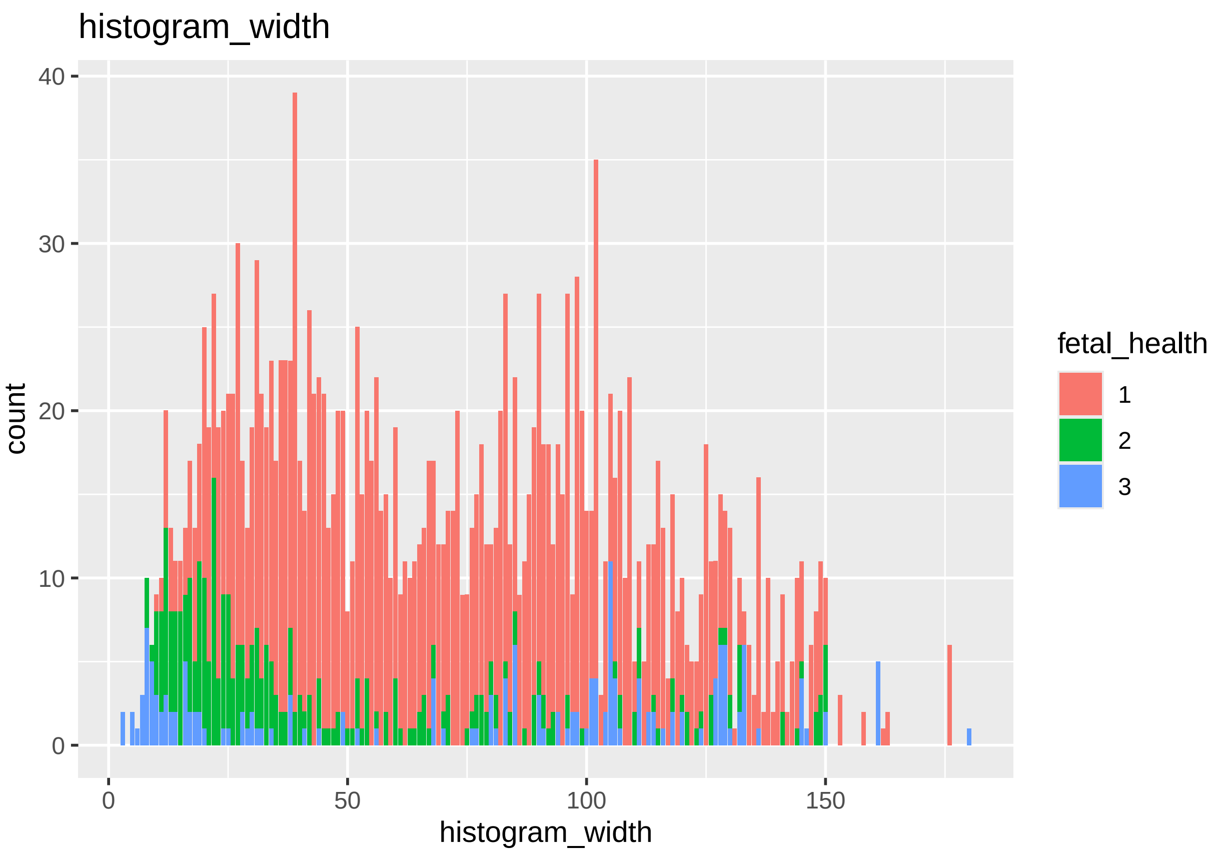


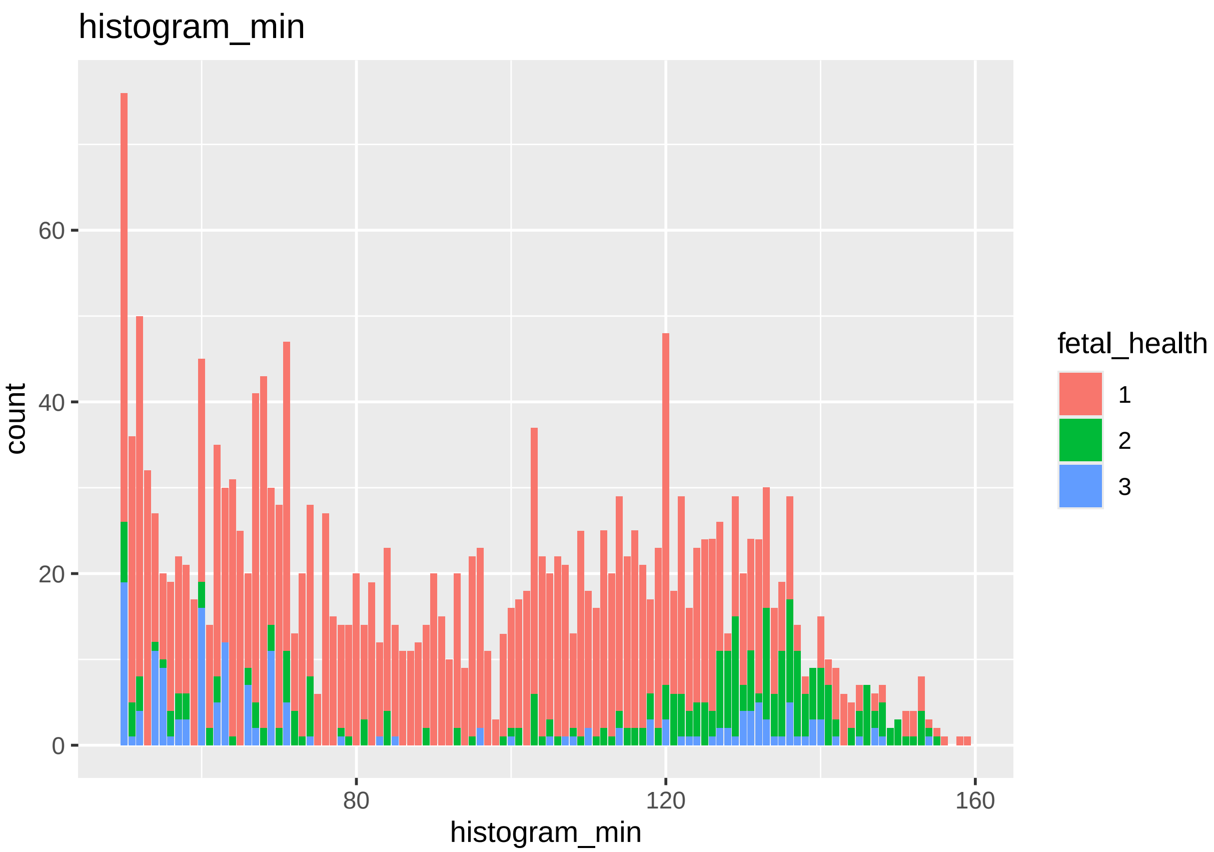
There are clear differences across the different status of fetal health. Majority of the values fall within the 'Normal' status.

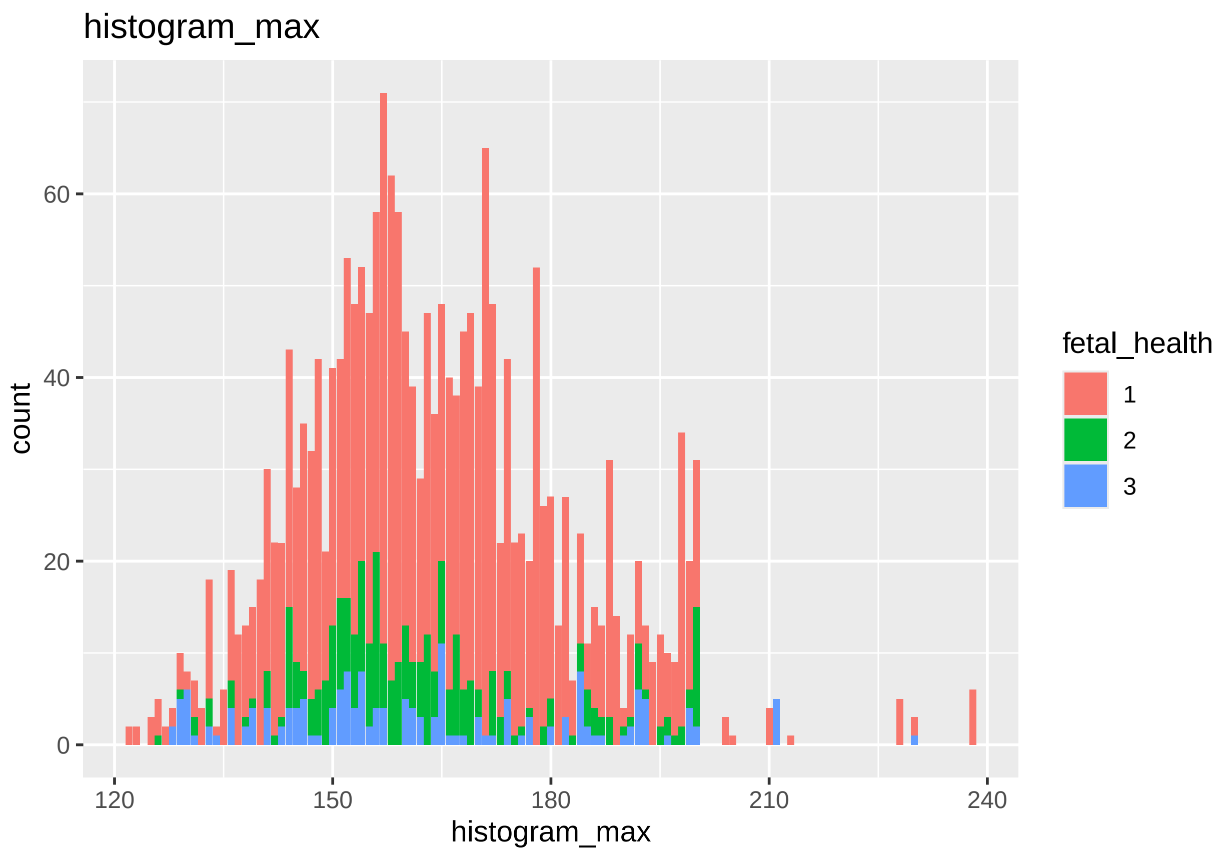
Skewness is common across features, particularly for accelerations and light_decelerations, though level of skewness tends to vary by status within features.

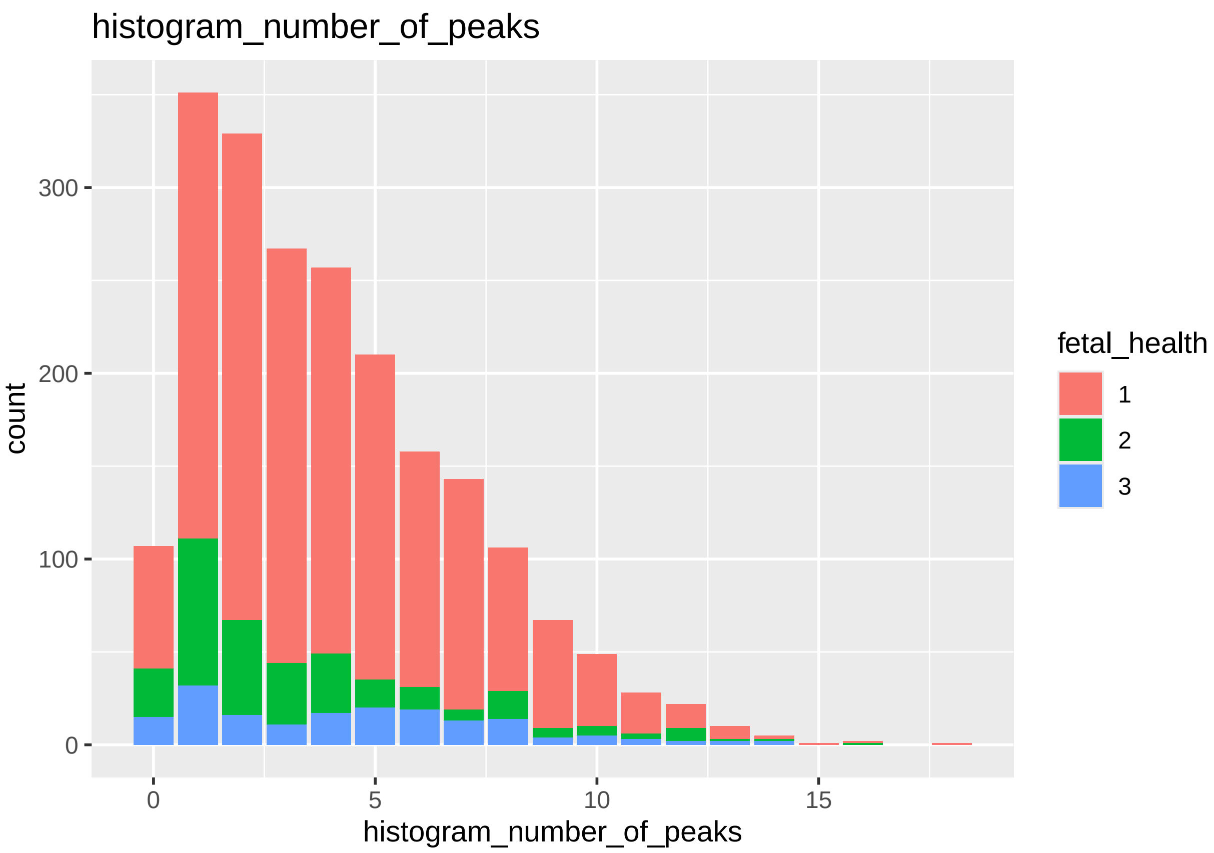
- Histogram features

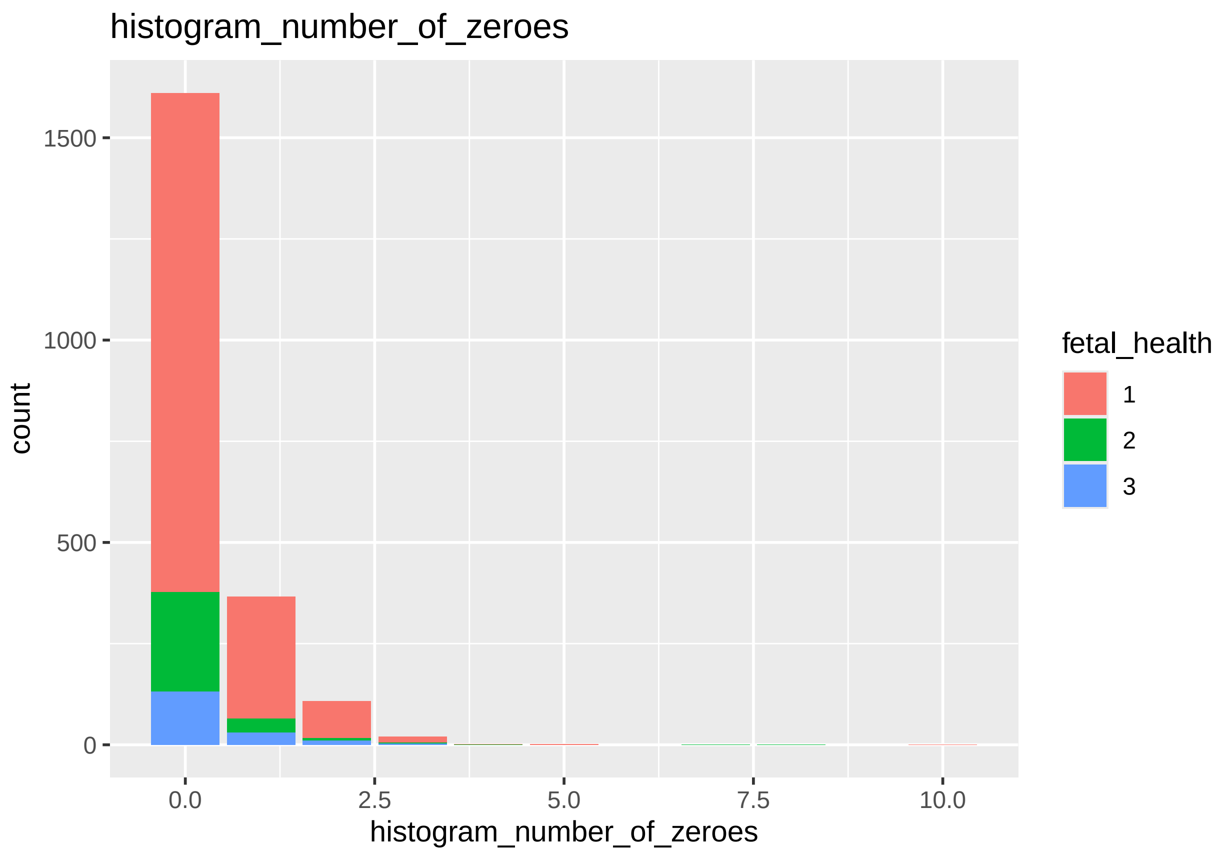
```
for(i in 1 : length(cols2)){
    plot2 <- ggplot(data_2, aes_string(x = cols2[i], fill = "fetal_health")) +
    geom_bar() +
    ggtitle(cols2[i])

    print(plot2)
  #hist(data_2[,i], main = cols1[i], xlab = cols1[i], col = 'blue')
}
```

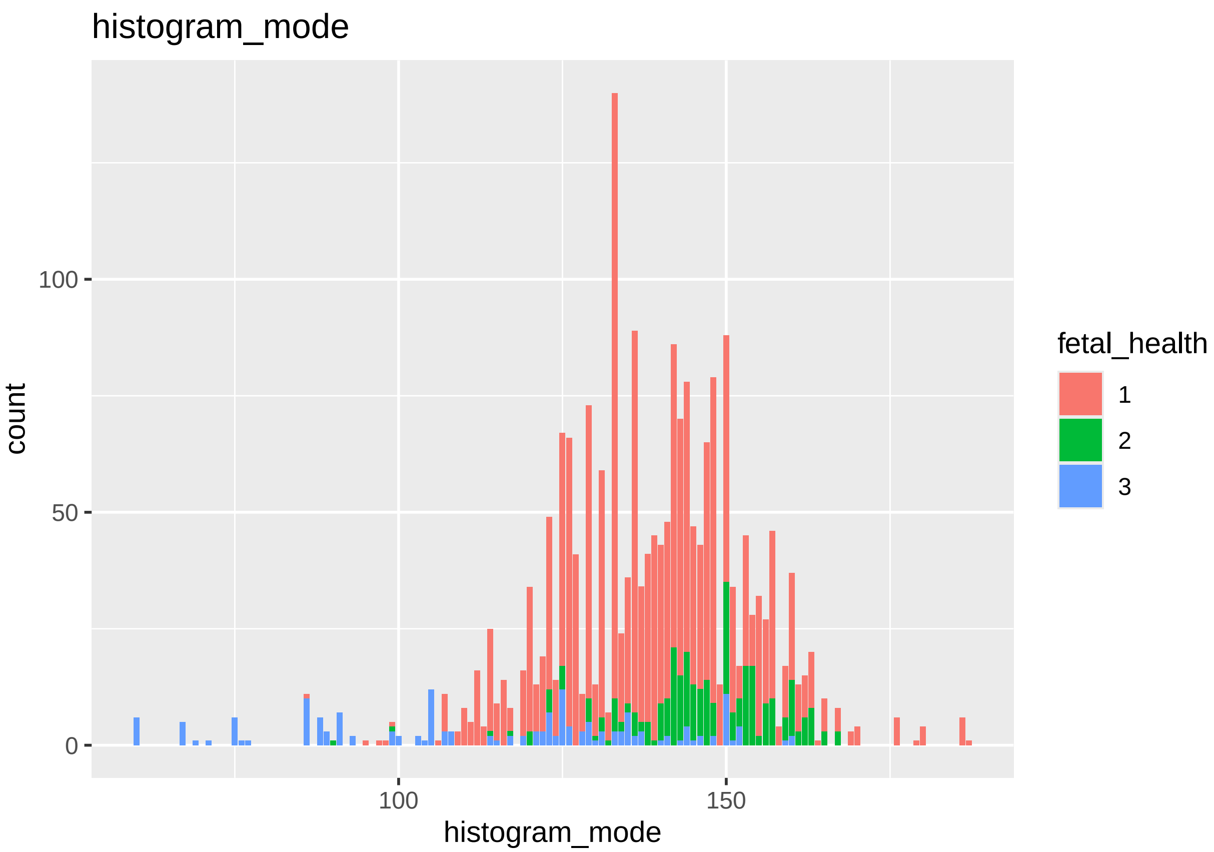

histogram_width
count
40
30
20
10
0
0
50
100
150
histogram_width
fetal_health
1
2
3

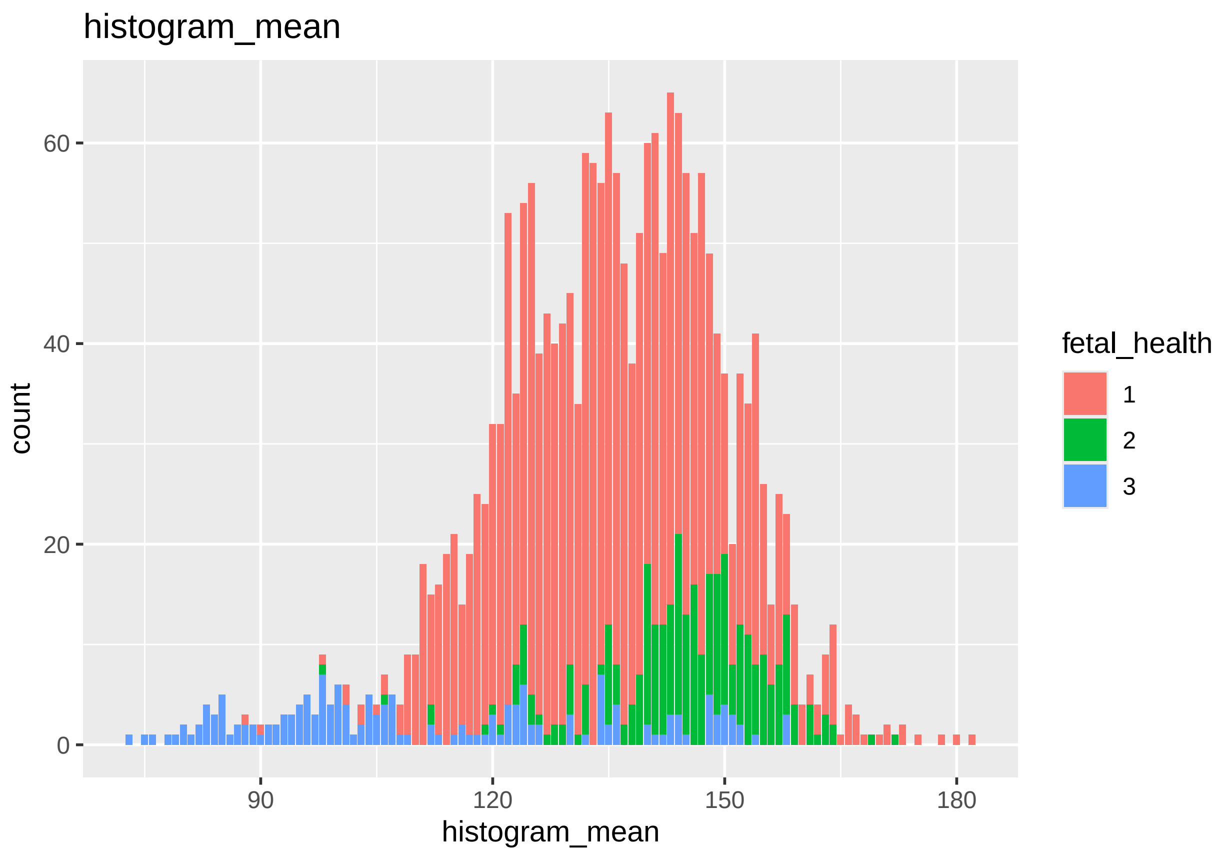

histogram_min
60
40
20
0
count
80
120
160
histogram_min
fetal_health
1
2
3

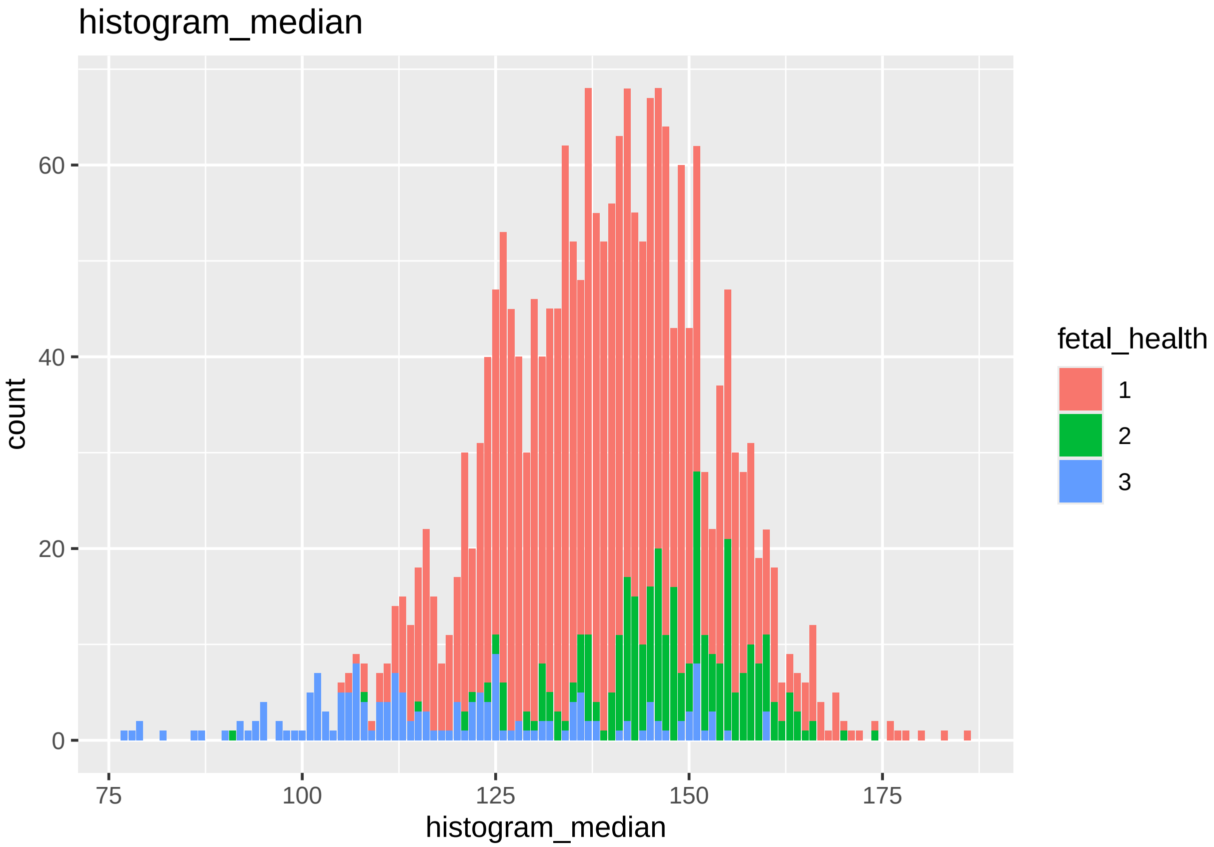

histogram_max
count
0
20
40
60
120
150
180
210
240
histogram_max
fetal_health
1
2
3

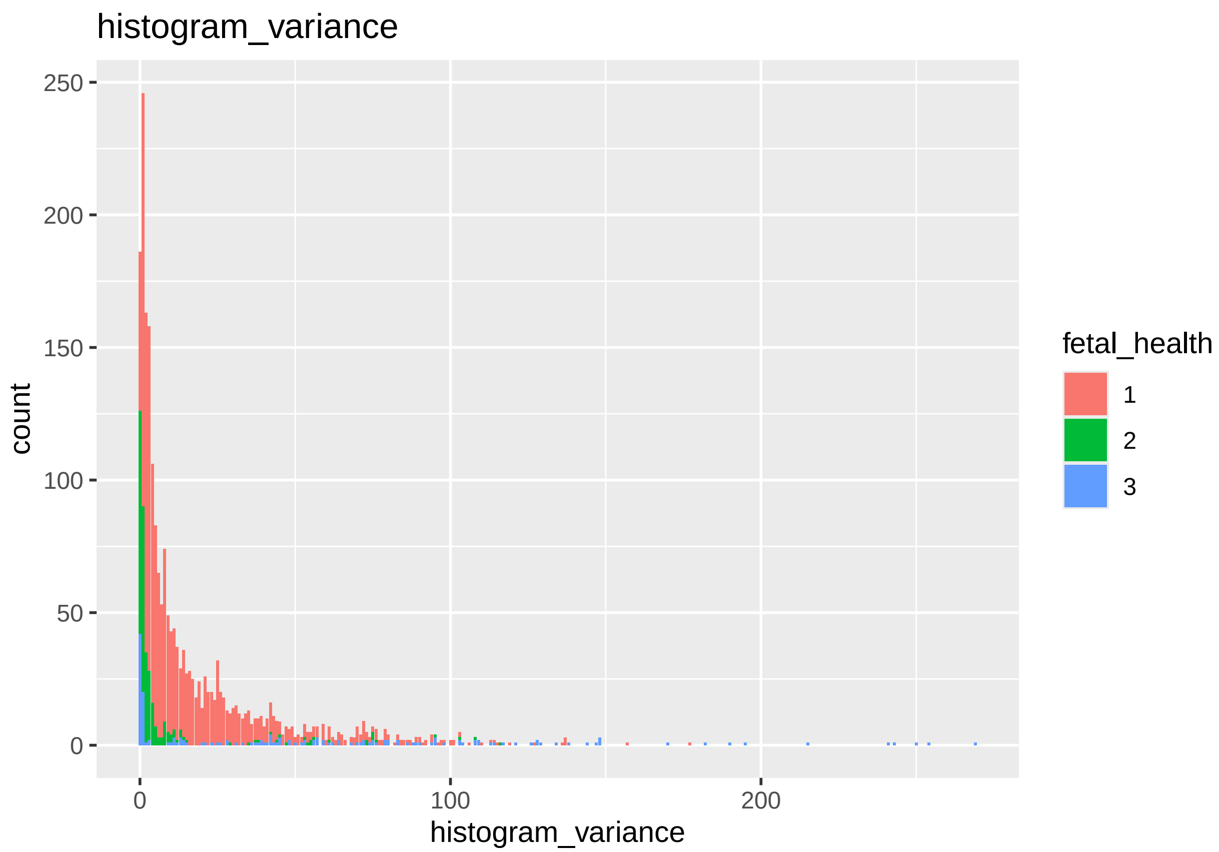

histogram_number_of_peaks
count
300
200
100
0
0
5
10
15
histogram_number_of_peaks
fetal_health
1
2
3

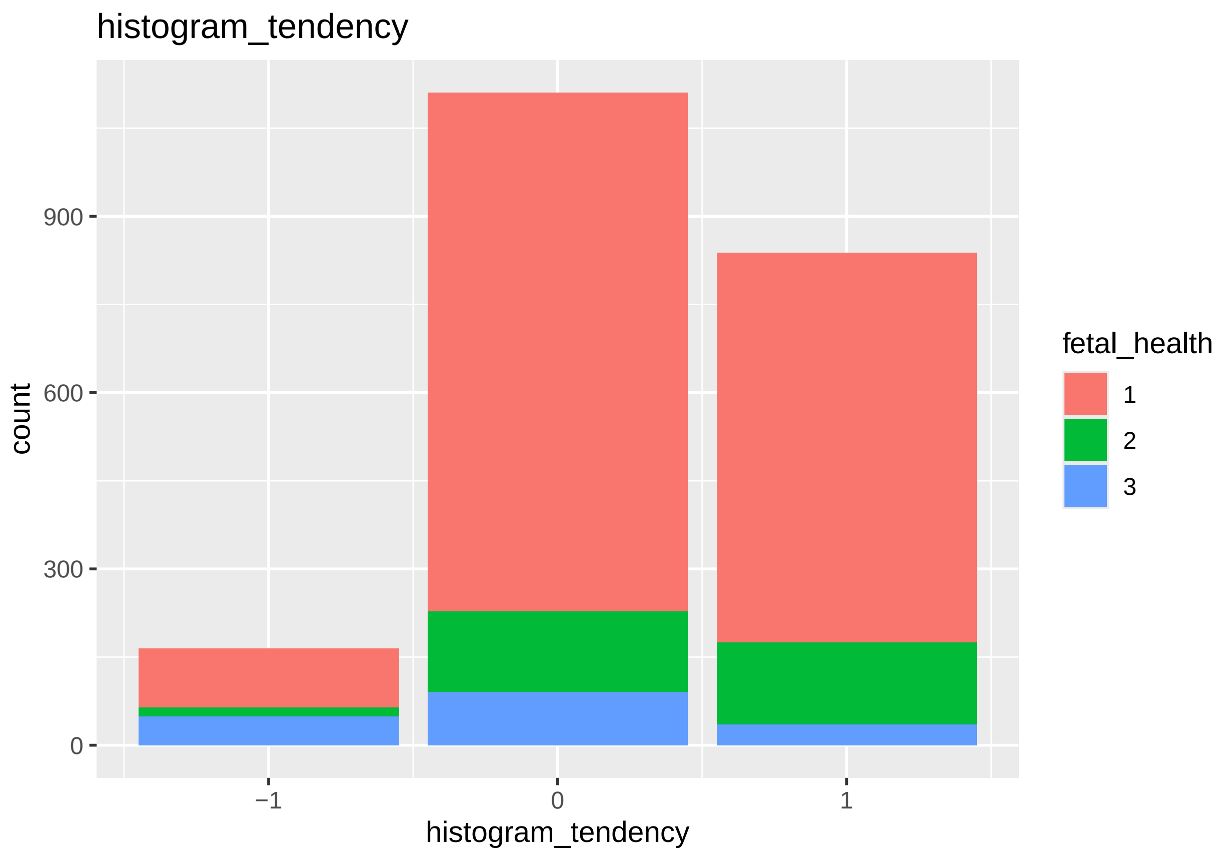

histogram_number_of_zeroes
1500
1000
500
0
count
0.0
2.5
5.0
7.5
10.0
histogram_number_of_zeroes
fetal_health
1
2
3


histogram_mode
count
100
50
0
100
150
histogram_mode
fetal_health
1
2
3


histogram_mean
count
60
40
20
0
90
120
150
180
histogram_mean
fetal_health
1
2
3


histogram_median
count
60
40
20
0
75
100
125
150
175
histogram_median
fetal_health
1
2
3


histogram_variance
250
200
150
100
50
0
count
0
100
200
histogram_variance
fetal_health
1
2
3

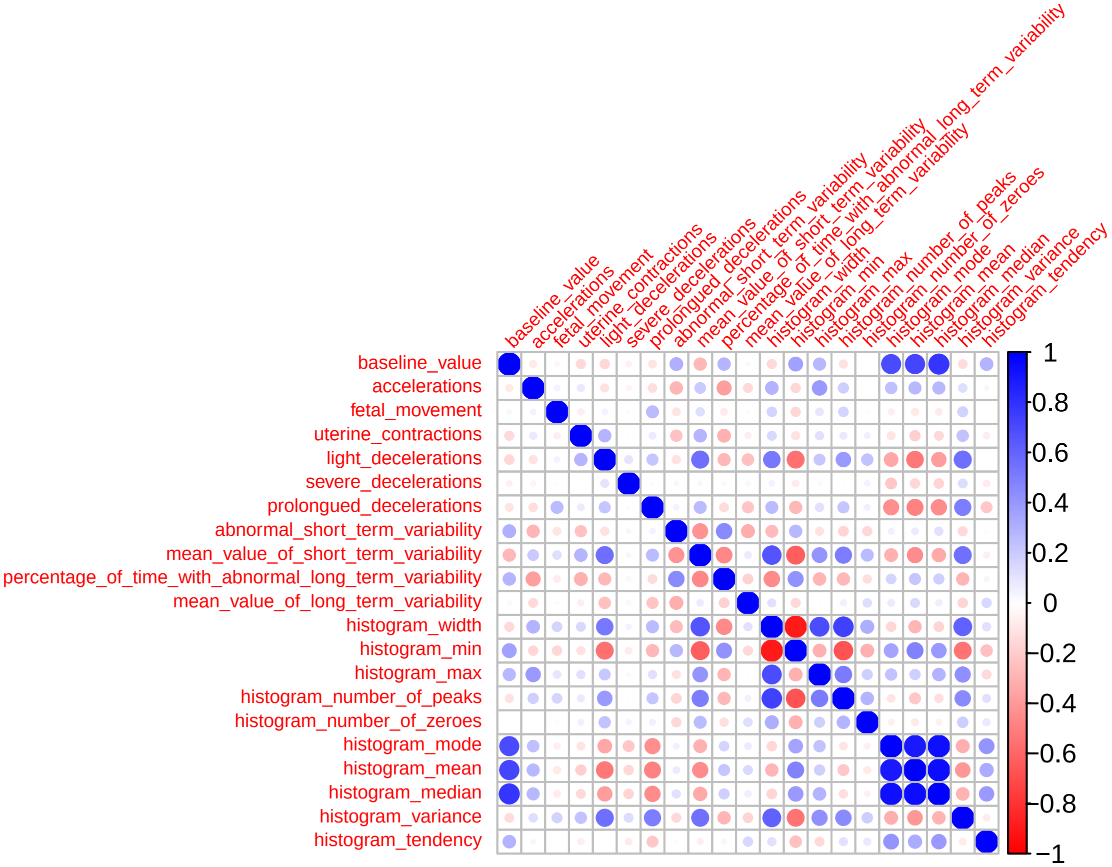


The three measure of central tendency features(mean,mode,median) show different distributions across classes of fetal_health, but those distributions are similar regardless of measure.

There is less skewness among this set of features, though it is still present and is substantial for histogram_number_of_peaks, histogram_variance, and histogram_number_of_zeroes.

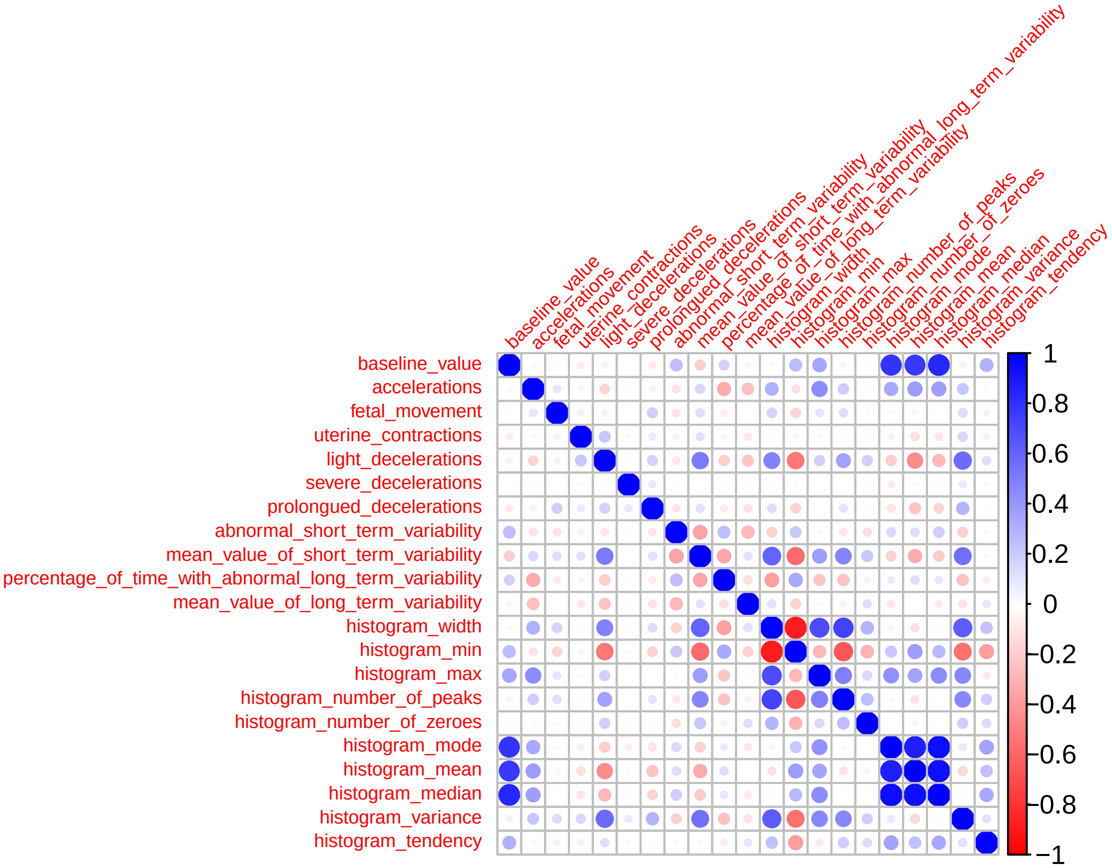
Correlation of Features

```r
par(mar = c(1, 1, 1, 1))

corrplot(cor(data_2 %>%
              dplyr::select(-fetal_health)),
              col = colorRampPalette(c("red", "white", "blue"))(200),
              tl.srt = 45,tl.cex = 0.6)
```



There is high correlation between baseline__value and histogram__mode,histogram__mean,histogram__median; between histogram__mean and histogram__mode and between histogram__median and histogram__mean,histogram__mode.

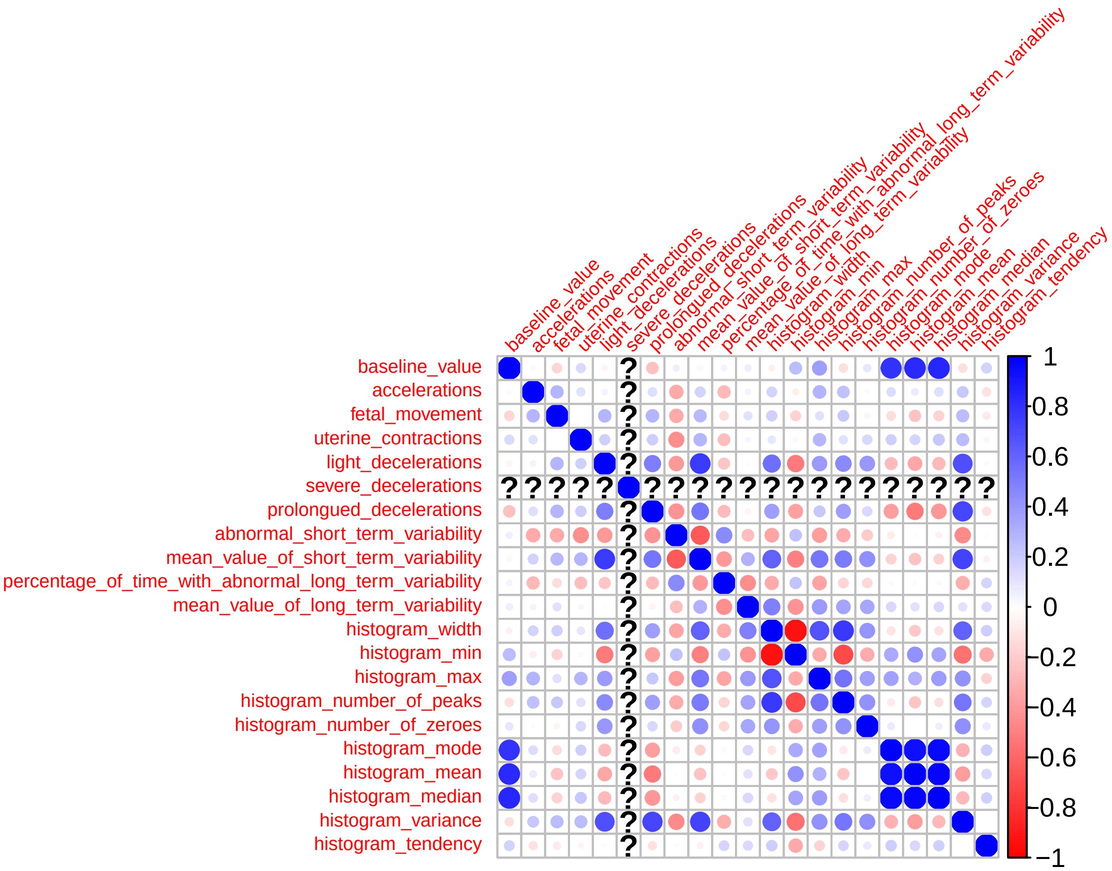
```
#When foetal health is Normal
par(mar = c(1, 1, 1, 1))

corrplot(cor(data_2 %>%
               filter(fetal_health == 1) %>%
               dplyr::select(-fetal_health)),
               col = colorRampPalette(c("red",   "white", "blue"))(200),
               tl.srt = 45,tl.cex = 0.6)
```



The heat map for fetal_health status ‘Normal’ shares similar pattern with one that includes all status. This is understandable since majority of our values are of ‘Normal’ status.

```
#When foetal health is Suspect
par(mar = c(1, 1, 1, 1))

corrplot(cor(data_2 %>%
               filter(fetal_health == 2) %>%
               dplyr::select(-fetal_health)),
               col = colorRampPalette(c("red", "white", "blue"))(200),
               tl.srt = 45,tl.cex = 0.6)
```

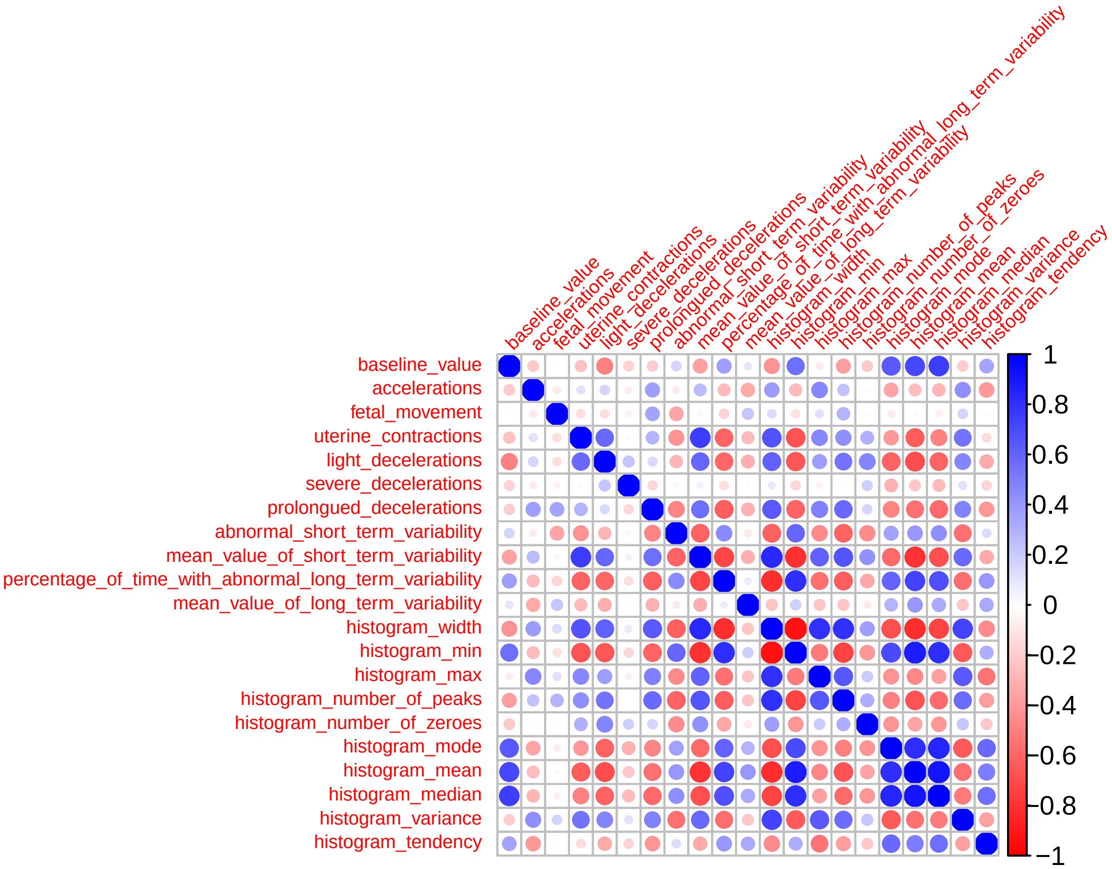
```
## Warning in cor(data_2 %>% filter(fetal_health == 2) %>%
## dplyr::select(-fetal_health)): the standard deviation is zero
```



The question marks for severe_decelerations indicate that there are zero values for that feature for this class.

Compared to the fetal_heath ‘Normal’, there are stronger relationships between the histogram and non-histogram features.

```
#When foetal health is Pathological
par(mar = c(1, 1, 1, 1))

corrplot(cor(data_2 %>%
               filter(fetal_health == 3) %>%
               dplyr::select(-fetal_health)),
               col = colorRampPalette(c("red", "white", "blue"))(200),
               tl.srt = 45,tl.cex = 0.6)
```



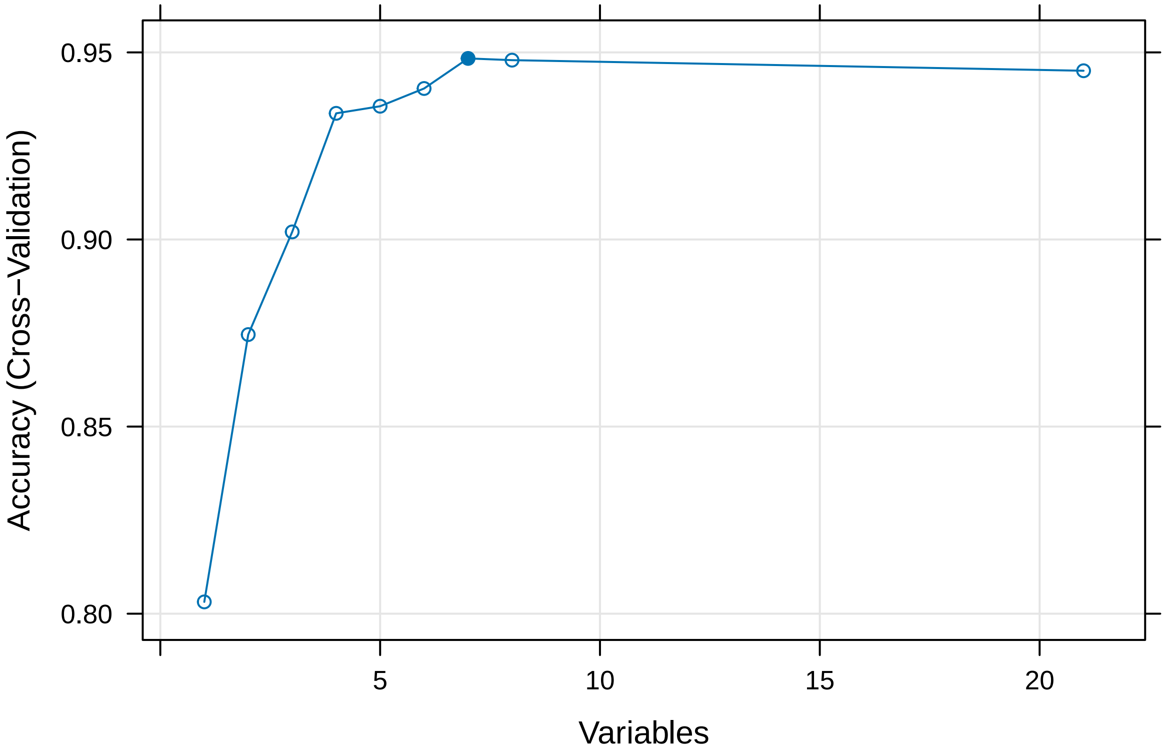
'Pathological' status shows even stronger correlations, both positive and negative.

However, the variables baseline_value,accelerations,fetal_movement, severe_decelerations, mean_value_of_long_term_vari and histogram_number_of_zeroes show relatively weaker correlations.

Feature Selection

Process to select the relevant and useful variables for predicted model.

```
#Use RFE (Recursive Feature Elimination)

set.seed(123)

# define the control using a random forest selection function
control <- rfeControl(functions=rfFuncs, method="cv", number=10)

# run the RFE algorithm
results <- rfe(data_2[,1:21], data_2[,22], sizes=c(1:8), rfeControl=control)

# plot the results
plot(results, type=c("g", "o"))
```



```
# Display the selected features
selected_features <- predictors(results)
print(selected_features)
```

```
## [1] "abnormal_short_term_variability"
## [2] "percentage_of_time_with_abnormal_long_term_variability"
## [3] "mean_value_of_short_term_variability"
## [4] "histogram_mean"
## [5] "accelerations"
## [6] "uterine_contractions"
## [7] "histogram_mode"
```

```
# Subset the original data with the selected features
f_data <- data_2[,c(selected_features,"fetal_health")]
head(f_data)
```

```
##   abnormal_short_term_variability
## 1                              73
## 2                              17
## 3                              16
## 4                              16
## 5                              16
## 6                              26
##   percentage_of_time_with_abnormal_long_term_variability
## 1                                                     43
## 2                                                      0
## 3                                                      0
## 4                                                      0
## 5                                                      0
## 6                                                      0
##   mean_value_of_short_term_variability histogram_mean accelerations
## 1                                  0.5            137         0.000
## 2                                  2.1            136         0.006
## 3                                  2.1            135         0.003
## 4                                  2.4            134         0.003
## 5                                  2.4            136         0.007
## 6                                  5.9            107         0.001
##   uterine_contractions histogram_mode fetal_health
## 1                0.000            120            2
## 2                0.006            141            1
## 3                0.008            141            1
## 4                0.008            137            1
## 5                0.008            137            1
## 6                0.010             76            3
```

```
nrow(f_data)
```

```
## [1] 2113
```

```
str(f_data)
```

```
## ’data.frame’:    2113 obs. of  8 variables:
##  $ abnormal_short_term_variability                       : int  73 17 16 16 16 26 29 83 84 86 ...
##  $ percentage_of_time_with_abnormal_long_term_variability: int  43 0 0 0 0 0 0 6 5 6 ...
##  $ mean_value_of_short_term_variability                  : num  0.5 2.1 2.1 2.4 2.4 5.9 6.3 0.5 0.5
##  $ histogram_mean                                        : int  137 136 135 134 136 107 107 122 122
##  $ accelerations                                         : num  0 0.006 0.003 0.003 0.007 0.001 0.00
##  $ uterine_contractions                                  : num  0 0.006 0.008 0.008 0.008 0.01 0.013
##  $ histogram_mode                                        : int  120 141 141 137 137 76 71 122 122 12
##  $ fetal_health                                          : Factor w/ 3 levels "1","2","3": 2 1 1 1 1
```

```
#write.csv(f_data,"C://Users//pc//Documents//Project//Foetal//fdata.csv", row.names = T)
```

Dealing with imbalanced data

SMOTE(Synthetic Minority Oversampling Technique)

Involves creating new dataset by oversampling observations from minority class.

```
# Separate the majority class (class 1) and the minority classes (class 2 and 3)
majority_class <- subset(f_data, fetal_health == '1')
minority_class_2 <- subset(f_data, fetal_health == '2')
minority_class_3 <- subset(f_data, fetal_health == '3')
```

```r
#Split data into train and Test set
#For normal class

set.seed(1)

indexesN <- sample(1:nrow(majority_class), size = 0.2*nrow(majority_class))
testN<- majority_class[indexesN,]
trainN<- majority_class[-indexesN,]
testN.Y<- majority_class$fetal_health [indexesN]
trainN.Y<- majority_class$fetal_health [-indexesN]

#For Suspect class
set.seed(1)

indexesS <- sample(1:nrow(minority_class_2), size = 0.2*nrow(minority_class_2))
testS<- minority_class_2[indexesS,]
trainS<- minority_class_2[-indexesS,]
testS.Y<- minority_class_2$fetal_health [indexesS]
trainS.Y<- minority_class_2$fetal_health [-indexesS]

#For Pathological class
set.seed(1)

indexesP <- sample(1:nrow(minority_class_3), size = 0.2*nrow(minority_class_3))
testP<- minority_class_3[indexesP,]
trainP<- minority_class_3[-indexesP,]
testP.Y<- minority_class_3$fetal_health [indexesP]
trainP.Y<- minority_class_3$fetal_health [-indexesP]

#create TESTSET and TRAINSET set
test<- rbind(testN,testS,testP)
train <- rbind(trainN,trainS,trainP)
```

```r
dim(test)
```

```
## [1] 422   8
```

```r
dim(train)
```

```
## [1] 1691    8
```

```r
unique(train$fetal_health)
```

```
## [1] 1 2 3
## Levels: 1 2 3
```

- **Train data**

```r
#balancing class using SMOTE for TRAINSET
library(smotefamily)
set.seed(123)
```

```r
#oversampling class S = 2

for (i in 1:nrow(train)){
  train$sus[i] <- ifelse(train$fetal_health[i] == 2,2,0)
}
train.2 <- train[,-8]

smote_result22 = SMOTE(train.2[,-8],target = train.2$sus, K = 3, dup_size = 3)

oversampled22 = smote_result22$data


library(dplyr)
BS2<- filter(oversampled22, oversampled22$class == 2)
BS2$fetal_health <- BS2$class
BS2 <- BS2[-8]
nrow(BS2)
```

```
## [1] 936
```

```r
str(BS2)
```

```
## ’data.frame’:    936 obs. of  8 variables:
##  $ abnormal_short_term_variability                       : num  25 59 64 63 44 74 77 61 66 79 ...
##  $ percentage_of_time_with_abnormal_long_term_variability: num  0 32 31 30 61 42 45 8 38 25 ...
##  $ mean_value_of_short_term_variability                  : num  1.9 0.4 0.4 0.4 0.6 0.4 0.2 0.4 0.5
##  $ histogram_mean                                        : num  125 153 147 149 140 128 145 147 146
##  $ accelerations                                         : num  0.001 0 0 0 0 0 0 0 0 0 ...
##  $ uterine_contractions                                  : num  0.004 0.006 0.004 0 0.003 0 0 0 0
##  $ histogram_mode                                        : num  129 155 150 150 141 131 146 148 147
##  $ fetal_health                                          : chr  "2" "2" "2" "2" ...
```

```r
#oversampling class P = 3
set.seed(123)
for (i in 1:nrow(train)){
  train$path[i] <- ifelse(train$fetal_health[i] == '3','3',0)
}
train.3<- train[, -c(8,9)]
smote_result33 = SMOTE(train.3[,- 8],target = train.3$path, K = 4, dup_size = 4)

oversampled33 = smote_result33$data
BP3 <- filter(oversampled33, oversampled33$class == 3)
BP3$fetal_health <- BP3$class
BP3 <- BP3[-8]
nrow(BP3)
```

```
## [1] 700
```

```r
str(BP3)
```

```
## ’data.frame’:    700 obs. of  8 variables:
```

```
##  $ abnormal_short_term_variability                       : num  26 34 60 70 63 65 63 64 67 64 ...
##  $ percentage_of_time_with_abnormal_long_term_variability: num  0 0 0 54 0 0 0 0 0 0 ...
##  $ mean_value_of_short_term_variability                  : num  4.3 2.2 3.2 0.3 4.2 2.5 1.3 1.3 3.2
##  $ histogram_mean                                        : num  105 99 94 121 73 94 100 98 80 92 ...
##  $ accelerations                                         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ uterine_contractions                                  : num  0.007 0.003 0.007 0 0.008 0.007 0.00
##  $ histogram_mode                                        : num  126 75 93 123 69 104 107 86 105 86 .
##  $ fetal_health                                          : chr  "3" "3" "3" "3" ...
```

```r
dim(trainN)
```

```
## [1] 1317    8
```

```r
str(trainN)
```

```
## 'data.frame':    1317 obs. of  8 variables:
##  $ abnormal_short_term_variability                       : int  17 16 16 16 28 28 21 19 24 23 ...
##  $ percentage_of_time_with_abnormal_long_term_variability: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mean_value_of_short_term_variability                  : num  2.1 2.1 2.4 2.4 1.4 1.5 2.3 2.3 2.1
##  $ histogram_mean                                        : int  136 135 134 136 134 137 125 127 128
##  $ accelerations                                         : num  0.006 0.003 0.003 0.007 0.005 0.009
##  $ uterine_contractions                                  : num  0.006 0.008 0.008 0.008 0.008 0.006
##  $ histogram_mode                                        : int  141 141 137 137 135 141 143 134 143
##  $ fetal_health                                          : Factor w/ 3 levels "1","2","3": 1 1 1 1 1
```

```r
#create NEWTRAIN SET
newTR.df <- rbind(trainN,BS2,BP3)
newTR <- newTR.df[,-8]
newTR.LABEL <- newTR.df$fetal_health
unique(newTR.LABEL)
```

```
## [1] 1 2 3
## Levels: 1 2 3
```

```r
dim(newTR)
```

```
## [1] 2953    7
```

- **Test data**

```r
set.seed(123)

#oversampling class S = 2

for (i in 1:nrow(test)){
  test$sus[i] <- ifelse(test$fetal_health[i] == 2,2,0)
}
test.2 <- test[,-8]

smote_result_2 = SMOTE(test.2[,-8],target = test.2$sus, K = 3, dup_size = 3)
```

```
oversampled_2 = smote_result_2$data


library(dplyr)
TS2<- filter(oversampled_2, oversampled_2$class == 2)
TS2$fetal_health <- TS2$class
TS2 <- TS2[-8]
nrow(TS2)
```

```
## [1] 232
```

```
str(TS2)
```

```
## 'data.frame':    232 obs. of  8 variables:
##  $ abnormal_short_term_variability                       : num  62 70 64 78 50 76 66 65 62 69 ...
##  $ percentage_of_time_with_abnormal_long_term_variability: num  6 29 12 59 62 62 20 41 67 39 ...
##  $ mean_value_of_short_term_variability                  : num  0.5 0.4 0.5 0.2 0.5 0.2 0.4 0.4 0.4
##  $ histogram_mean                                        : num  163 123 142 139 159 140 141 150 158
##  $ accelerations                                         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ uterine_contractions                                  : num  0.003 0.001 0 0 0.005 0 0 0.005 0.004
##  $ histogram_mode                                        : num  163 125 143 140 160 142 144 152 160
##  $ fetal_health                                          : chr  "2" "2" "2" "2" ...
```

```
#oversampling class P = 3
set.seed(123)
for (i in 1:nrow(test)){
  test$path[i] <- ifelse(test$fetal_health[i] == '3','3',0)
}
test.3<- train[, -c(8,9)]
smote_result_3 = SMOTE(test.3[,- 8],target = test.3$path, K = 4, dup_size = 4)

oversampled_3 = smote_result_3$data
TP3 <- filter(oversampled_3, oversampled_3$class == 3)
TP3$fetal_health <- TP3$class
TP3 <- TP3[-8]
nrow(TP3)
```

```
## [1] 700
```

```
str(TP3)
```

```
## 'data.frame':    700 obs. of  8 variables:
##  $ abnormal_short_term_variability                       : num  26 34 60 70 63 65 63 64 67 64 ...
##  $ percentage_of_time_with_abnormal_long_term_variability: num  0 0 0 54 0 0 0 0 0 0 ...
##  $ mean_value_of_short_term_variability                  : num  4.3 2.2 3.2 0.3 4.2 2.5 1.3 1.3 3.2
##  $ histogram_mean                                        : num  105 99 94 121 73 94 100 98 80 92 ...
##  $ accelerations                                         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ uterine_contractions                                  : num  0.007 0.003 0.007 0 0.008 0.007 0.00
##  $ histogram_mode                                        : num  126 75 93 123 69 104 107 86 105 86 .
##  $ fetal_health                                          : chr  "3" "3" "3" "3" ...
```

```
#create NEWTRAIN SET
newTS.df <- rbind(testN,TS2,TP3)
newTS <- newTS.df[,-8]
newTS.LABEL <- newTS.df$fetal_health
unique(newTS.LABEL)
```

```
## [1] 1 2 3
## Levels: 1 2 3
```

```
dim(newTS)
```

```
## [1] 1261    7
```

Model Building

```
#scale data
s_train <- as.data.frame(scale(newTR))
train_s <- cbind(s_train, fetal_health = newTR.df$fetal_health)

s_test <- as.data.frame(scale(newTS))
test_s <- cbind(s_test, fetal_health = newTS.df$fetal_health)
```

1. **Decision Tree**

```
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 4.2.3
```

```
## Fit decision tree model
model_tree <- rpart(fetal_health ~ ., data = train_s, method = "class")

# Make predictions
predictions_tree <- predict(model_tree, test_s, type = "class")

# Evaluate the model
caret::confusionMatrix(predictions_tree, as.factor(test_s$fetal_health))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2   3
##          1 296  30 223
##          2  31 202 116
##          3   2   0 361
##
## Overall Statistics
##
##                Accuracy : 0.6812
##                    95% CI : (0.6547, 0.7069)
##     No Information Rate : 0.5551
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                     Kappa : 0.5282
##
##  Mcnemar’s Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity            0.8997   0.8707   0.5157
## Specificity            0.7285   0.8571   0.9964
## Pos Pred Value         0.5392   0.5788   0.9945
## Neg Pred Value         0.9537   0.9671   0.6225
## Prevalence             0.2609   0.1840   0.5551
## Detection Rate         0.2347   0.1602   0.2863
## Detection Prevalence   0.4354   0.2768   0.2879
## Balanced Accuracy      0.8141   0.8639   0.7561
```

- The overall performance (accuracy of 64.16%) is moderate.
- The model is good at detecting class 1 and 2, but struggles with class 3 (lower sensitivity).
- Precision for class 1 is relatively low (many false positives), while class 3 has high precision but low sensitivity.

2. **Random Forest**

```
# Install and load necessary packages

library(randomForest)
```

```
## Warning: package ’randomForest’ was built under R version 4.2.3
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: ’randomForest’
```

```
## The following object is masked from ’package:dplyr’:
##
##     combine
```

```
## The following object is masked from ’package:ggplot2’:
##
##     margin
```

```
library(caret)

# Fit random forest model
model_rf <- randomForest(fetal_health ~ ., data = train_s, ntree = 100)

# Make predictions
predictions_rf <- predict(model_rf, test_s)

# Evaluate the model
confusionMatrix(predictions_rf, as.factor(test_s$fetal_health))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2   3
##          1 309  23 211
##          2  18 209 119
##          3   2   0 370
##
## Overall Statistics
##
##                Accuracy : 0.7042
##                  95% CI : (0.6782, 0.7293)
##     No Information Rate : 0.5551
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5607
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity            0.9392   0.9009   0.5286
## Specificity            0.7489   0.8669   0.9964
## Pos Pred Value         0.5691   0.6040   0.9946
## Neg Pred Value         0.9721   0.9749   0.6288
## Prevalence             0.2609   0.1840   0.5551
## Detection Rate         0.2450   0.1657   0.2934
## Detection Prevalence   0.4306   0.2744   0.2950
## Balanced Accuracy      0.8441   0.8839   0.7625
```

- Overall Accuracy has improved to 73.99%
- The model shows strong results for class 1 and class 2, though it still struggles somewhat with class 3 in terms of sensitivity (ability to detect all class 3 instances).
- Precision is highest for class 3, which means when class 3 is predicted, it's usually correct.

3. **Support Vector Machine**

```r
# Install and load necessary packages
#install.packages("e1071")
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.2.3
```

```r
library(caret)


# Fit SVM model
model_svm <- svm(fetal_health ~ ., data = train_s)

# Make predictions
predictions_svm <- predict(model_svm, test_s)

# Evaluate the model
confusionMatrix(predictions_svm, as.factor(test_s$fetal_health))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2   3
##          1 303  14 163
##          2  25 218 141
##          3   1   0 396
##
## Overall Statistics
##
##                Accuracy : 0.7272
##                  95% CI : (0.7017, 0.7516)
##     No Information Rate : 0.5551
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5928
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity            0.9210   0.9397   0.5657
## Specificity            0.8101   0.8387   0.9982
## Pos Pred Value         0.6312   0.5677   0.9975
## Neg Pred Value         0.9667   0.9840   0.6481
## Prevalence             0.2609   0.1840   0.5551
## Detection Rate         0.2403   0.1729   0.3140
## Detection Prevalence   0.3807   0.3045   0.3148
## Balanced Accuracy      0.8655   0.8892   0.7820
```

- The overall accuracy of 73.2% remains strong, and the model continues to perform well in identifying class 1 and class 2.
- Class 3 sensitivity is still an area for improvement, though its precision is very high.
- Precision for class 2 is a bit low, indicating many predicted class 2 instances are misclassified from class 3.

Conclusion

- Random Forest is the best model for this classification task due to its high accuracy, balanced performance across all classes, and ability to generalize well on unseen data.