EDA - Vital Events

Katie Schilling - 501130072

07/02/2022

Required tools to be loaded

```
library(dlookr)
```

```
##
## Attaching package: 'dlookr'

## The following object is masked from 'package:base':
##
##     transform
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(moments)
```

```
##
## Attaching package: 'moments'
```

```
## The following objects are masked from 'package:dlookr':
##
##     kurtosis, skewness
```

```
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
library(smooth)
```

```
## Loading required package: greybox
```

```
## Package "greybox", v1.0.4 loaded.
```

```
##
## Attaching package: 'greybox'
```

```
## The following object is masked from 'package:lubridate':
##
##     hm
```

```
## This is package "smooth", v3.1.5
```

```
library(greybox)
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
##
## Attaching package: 'forecast'
```

```
## The following object is masked from 'package:greybox':
##
##     forecast
## The following object is masked from 'package:ggpubr':
##
##     gghistogram
```

```
library(funModeling)
```

```
## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following object is masked from 'package:dlookr':
##
##     describe

## The following objects are masked from 'package:base':
##
##     format.pval, units

## funModeling v.1.9.4 :)
## Examples and tutorials at livebook.datascienceheroes.com
##  / Now in Spanish: librovivodecienciadedatos.ai
```

Import the first data set

```
Vital_events<- read.csv("C:/Users/Katie Schilling/Downloads/vital_events_data_by_month_1994-2021_q2 (1)
## Change the column names to cleaner versions
colnames(Vital_events)<- c("Month","Year","Live Births", "Marriages","Deaths","Stillbirths")
```

View the data to see what information is present

```
head(Vital_events)
```

```
##               Month Year Live Births Marriages Deaths Stillbirths
## 1   January/janvier 1994       11631      2078   8094          75
## 2 February/fÃ©vrier 1994       11254      2650   6428          62
## 3        March/mars 1994       13003      2557   6503          73
## 4       April/avril 1994       12576      3967   6224          74
## 5           May/mai 1994       13240      6493   6483          67
## 6         June/juin 1994       13072      7754   6187          66
```

Clean up of the vitals Data event. Removal of the french version of the month, as many of them did not import properly. Makes the data easier to read, view and work with.

```
Vital_events[Vital_events == "January/janvier"] <- "January"
Vital_events[Vital_events == "February/fÃ©vrier"] <- "February"
Vital_events[Vital_events == "March/mars"] <- "March"
Vital_events[Vital_events == "April/avril"] <- "April"
Vital_events[Vital_events == "May/mai"] <- "May"
Vital_events[Vital_events == "June/juin"] <- "June"
Vital_events[Vital_events == "July/juillet"] <- "July"
Vital_events[Vital_events == "August/aoÃ»t"] <- "August"
Vital_events[Vital_events == "September/septembre"] <- "September"
Vital_events[Vital_events == "October/octobre"] <- "October"
Vital_events[Vital_events == "November/novembre"] <- "November"
Vital_events[Vital_events == "December/dÃ©cembre"] <- "December"


## Check data now to see if the changes are sufficent
head(Vital_events)
```

```
##      Month Year Live Births Marriages Deaths Stillbirths
## 1  January 1994       11631      2078   8094          75
## 2 February 1994       11254      2650   6428          62
## 3    March 1994       13003      2557   6503          73
## 4    April 1994       12576      3967   6224          74
## 5      May 1994       13240      6493   6483          67
## 6     June 1994       13072      7754   6187          66
```

Check for any missing values

```
sum(is.na(Vital_events))
```

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

Create function to view basic EDA

```
basic_eda <- function(data)
{
  glimpse(data)
  print(status(data))
  freq(data)
  print(profiling_num(data))
  plot_num(data)
  describe(data)
}
```

produce and view the results of the EDA function

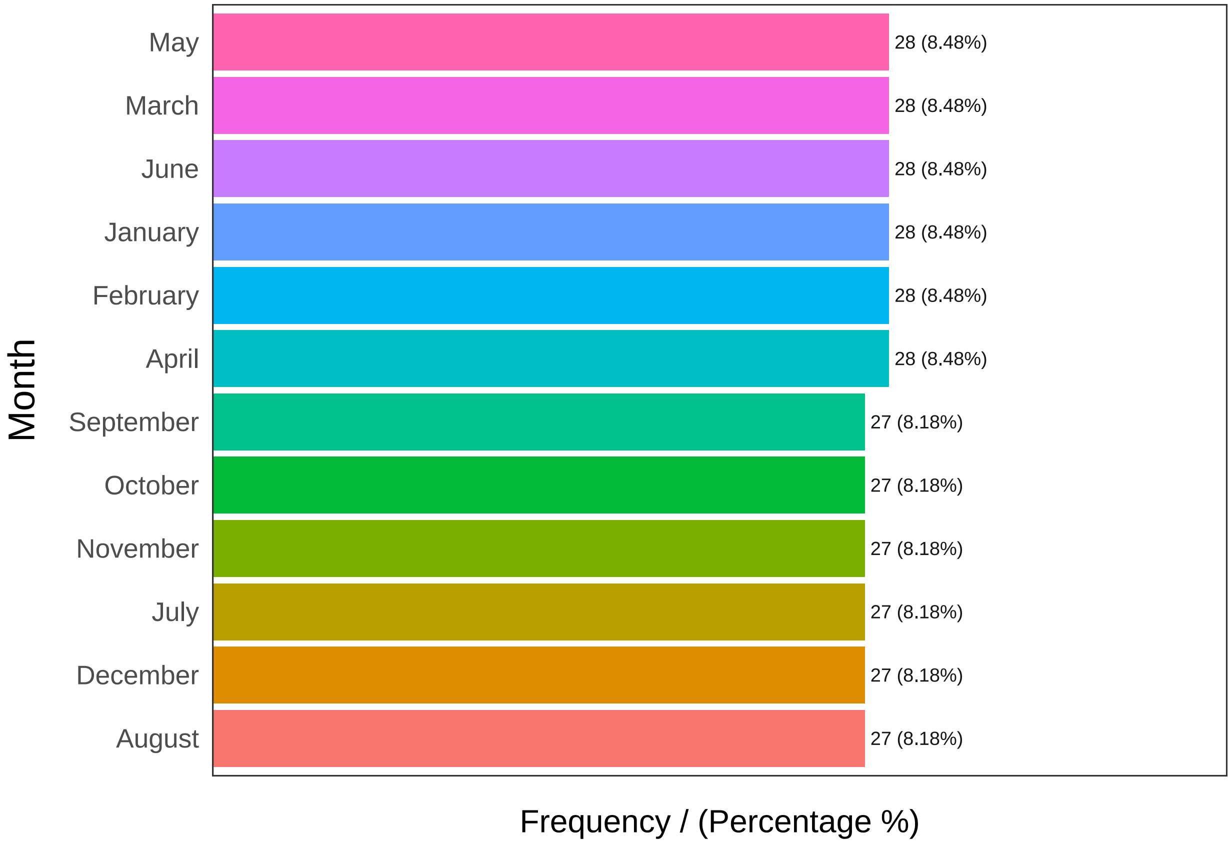
```
basic_eda(Vital_events)
```

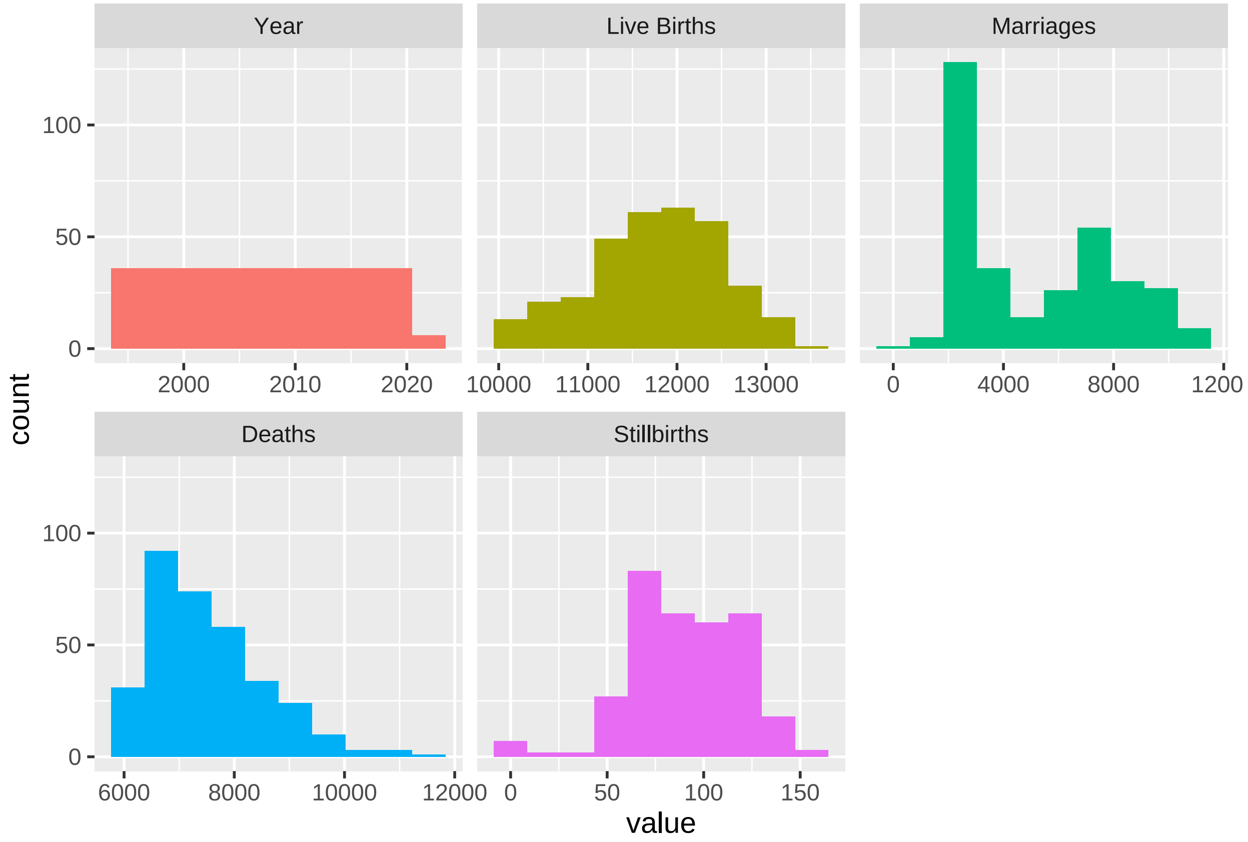
```
## Rows: 330
## Columns: 6
## $ Month         <chr> "January", "February", "March", "April", "May", "June", ~
## $ Year          <int> 1994, 1994, 1994, 1994, 1994, 1994, 1994, 1994, 1994, 19~
## $ ‘Live Births‘ <int> 11631, 11254, 13003, 12576, 13240, 13072, 13045, 12982, ~
## $ Marriages     <int> 2078, 2650, 2557, 3967, 6493, 7754, 9264, 9194, 8540, 74~
## $ Deaths        <int> 8094, 6428, 6503, 6224, 6483, 6187, 6196, 5926, 6062, 65~
## $ Stillbirths   <int> 75, 62, 73, 74, 67, 66, 70, 79, 60, 59, 56, 43, 78, 84, ~
##                 variable q_zeros     p_zeros q_na p_na q_inf p_inf      type
## Month              Month       0 0.000000000    0    0     0     0 character
## Year                Year       0 0.000000000    0    0     0     0   integer
## Live Births Live Births       0 0.000000000    0    0     0     0   integer
## Marriages     Marriages       0 0.000000000    0    0     0     0   integer
## Deaths           Deaths       0 0.000000000    0    0     0     0   integer
## Stillbirths Stillbirths       3 0.009090909    0    0     0     0   integer
##             unique
## Month           12
## Year            28
## Live Births    310
## Marriages      320
## Deaths         315
## Stillbirths    102
```

```
## Warning: ‘guides(<scale> = FALSE)‘ is deprecated. Please use ‘guides(<scale> =
## "none")‘ instead.
```

```
##      variable        mean     std_dev variation_coef     p_01    p_05     p_25
## 1        Year  2007.25455    7.953248    0.003962252  1994.00  1995.0  2000.00
## 2 Live Births 11762.98182  730.720088    0.062120311 10105.90 10441.4 11260.25
## 3   Marriages  5084.97273 2833.875707    0.557304013  1535.98  1939.7  2596.25
## 4      Deaths  7500.17879 1012.726472    0.135026977  6062.58  6202.4  6705.75
## 5 Stillbirths    90.73636   28.294912    0.311836516     2.00    52.0    73.00
##      p_50     p_75    p_95     p_99     skewness kurtosis     iqr
## 1  2007.0  2014.00  2020.0  2021.00  0.003429801 1.801575   14.00
## 2 11818.5 12287.50 12885.7 13131.69 -0.277401692 2.496109 1027.25
## 3  3559.0  7627.25  9803.8 10821.59  0.426691670 1.724241 5031.00
## 4  7325.5  8093.75  9395.0 10552.21  0.956654010 3.842203 1388.00
## 5    90.5   114.00   132.0   146.71 -0.443853296 3.616589   41.00
##             range_98           range_80
## 1       [1994, 2021]       [1996, 2018]
## 2 [10105.9, 13131.69] [10694.1, 12674.2]
## 3 [1535.98, 10821.59]   [2101.4, 9242.4]
## 4 [6062.58, 10552.21]   [6401.8, 8897.4]
## 5         [2, 146.71]        [59.9, 126]
```

```
## Warning: ‘guides(<scale> = FALSE)‘ is deprecated. Please use ‘guides(<scale> =
## "none")‘ instead.
```


Month
May
March
June
January
February
April
September
October
November
July
December
August
28 (8.48%)
28 (8.48%)
28 (8.48%)
28 (8.48%)
28 (8.48%)
28 (8.48%)
27 (8.18%)
27 (8.18%)
27 (8.18%)
27 (8.18%)
27 (8.18%)
27 (8.18%)
Frequency / (Percentage %)

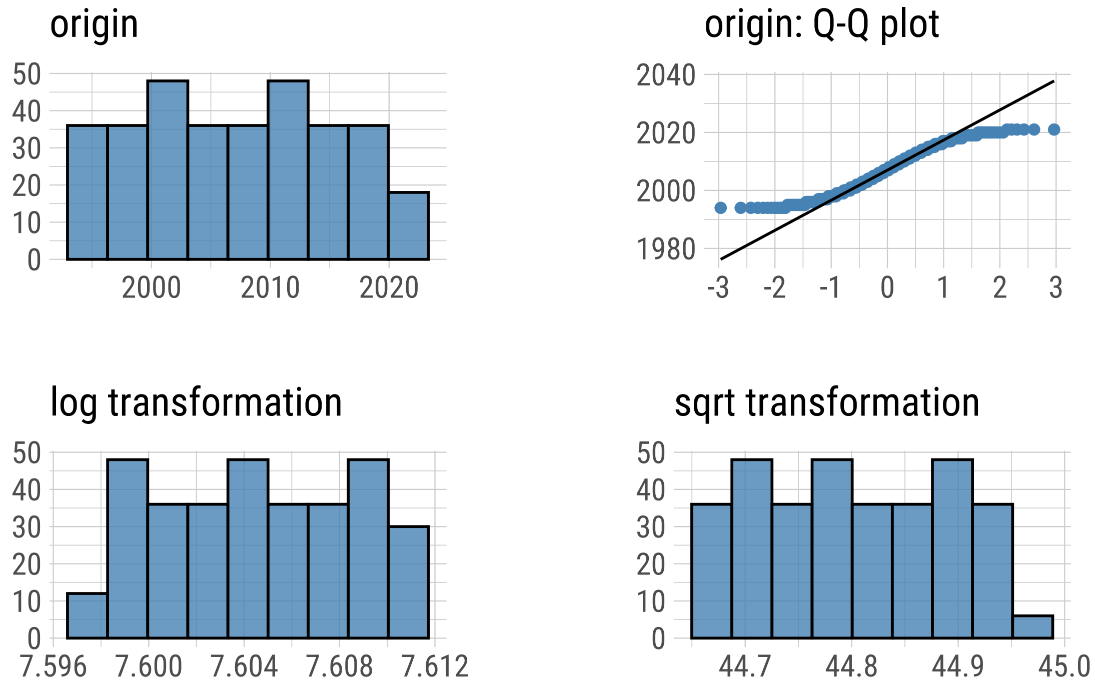


```
## data
##
##  6  Variables      330  Observations
## --------------------------------------------------------------------------------
## Month
##        n  missing distinct
##      330        0       12
##
## lowest : April     August    December  February  January
## highest: March     May       November  October   September
##
## Value          April    August  December  February   January      July
## Frequency         28        27        27        28        28        27
## Proportion     0.085     0.082     0.082     0.085     0.085     0.082
##
## Value           June     March       May  November   October September
## Frequency         28        28        28        27        27        27
## Proportion     0.085     0.085     0.085     0.082     0.082     0.082
## --------------------------------------------------------------------------------
## Year
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      330        0       28    0.999     2007    9.191     1995     1996
##      .25      .50      .75      .90      .95
##     2000     2007     2014     2018     2020
##
## lowest : 1994 1995 1996 1997 1998, highest: 2017 2018 2019 2020 2021
## --------------------------------------------------------------------------------
## Live Births
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      330        0      310        1    11763    832.6    10441    10694
##      .25      .50      .75      .90      .95
##    11260    11818    12288    12674    12886
##
## lowest : 10020 10059 10062 10103 10113, highest: 13104 13143 13195 13240 13398
## --------------------------------------------------------------------------------
## Marriages
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      330        0      320        1     5085     3171     1940     2101
##      .25      .50      .75      .90      .95
##     2596     3559     7627     9242     9804
##
## lowest :   597  1142  1314  1460  1722, highest: 10801 10830 11004 11083 11532
## --------------------------------------------------------------------------------
## Deaths
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      330        0      315        1     7500     1114     6202     6402
##      .25      .50      .75      .90      .95
##     6706     7326     8094     8897     9395
##
## lowest :  5926  6039  6060  6062  6064, highest: 10161 10712 10844 11121 11390
## --------------------------------------------------------------------------------
## Stillbirths
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      330        0      102        1    90.74    31.55     52.0     59.9
##      .25      .50      .75      .90      .95
##     73.0     90.5    114.0    126.0    132.0
##
## lowest :   0   2   8  19  24, highest: 146 147 148 150 156
## --------------------------------------------------------------------------------
```

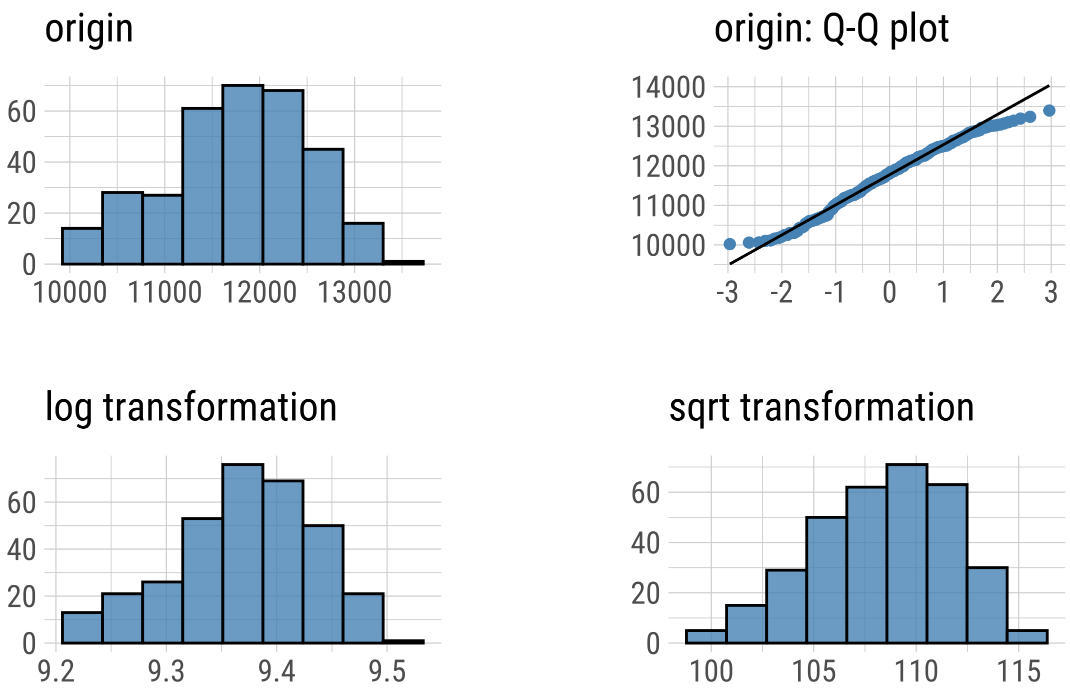
Visual analysis of the normality of the data.

```
plot_normality(Vital_events )
```

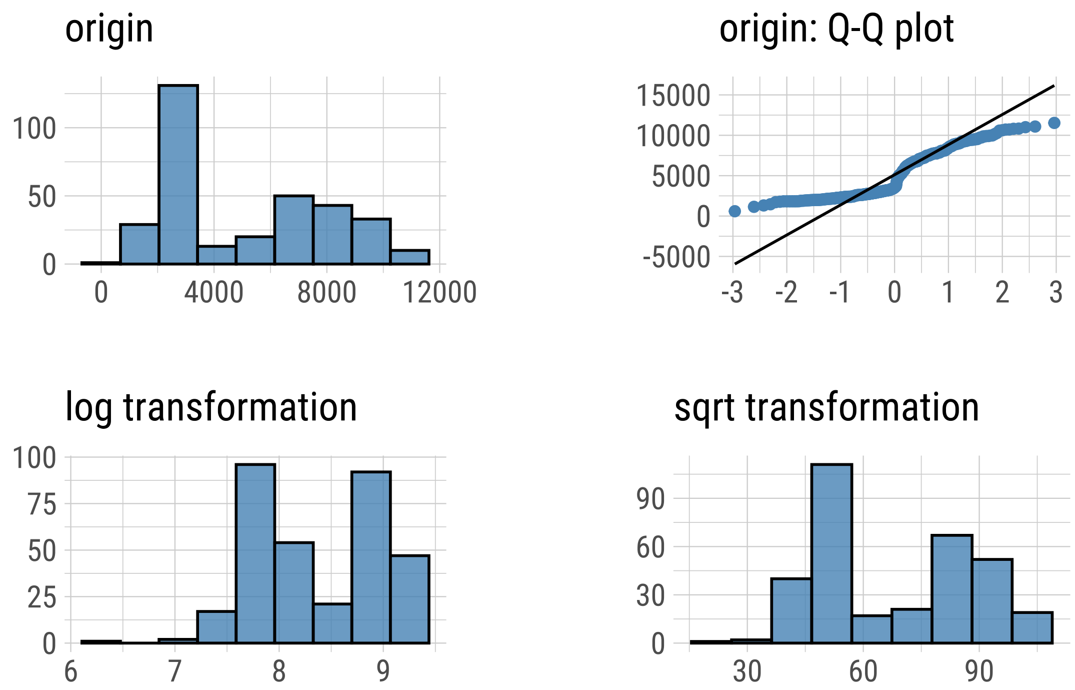
Normality Diagnosis Plot (Year)



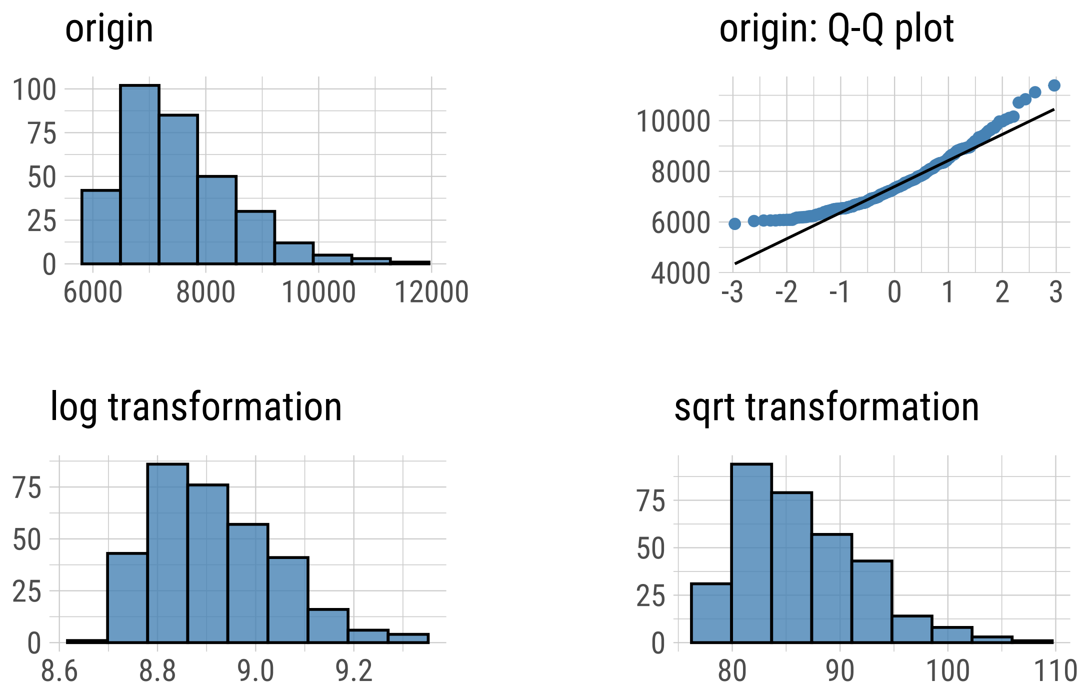
Normality Diagnosis Plot (Live Births)



Normality Diagnosis Plot (Marriages)



Normality Diagnosis Plot (Deaths)

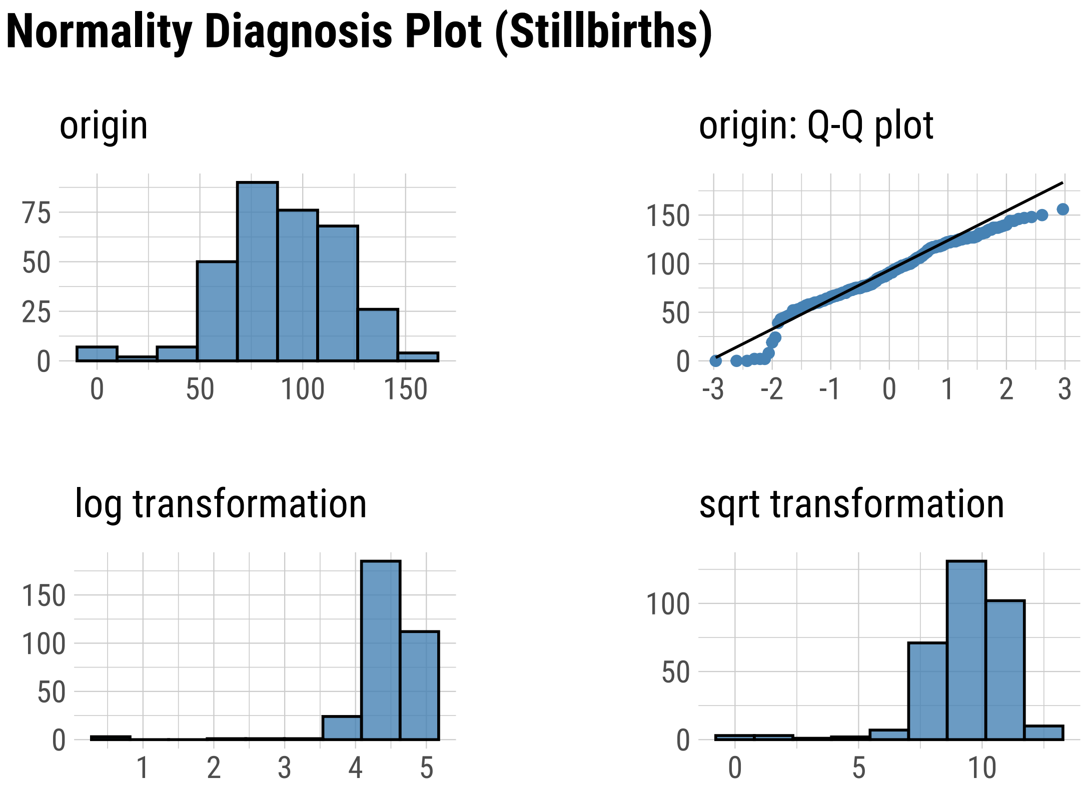


origin

origin: Q-Q plot

log transformation

sqrt transformation



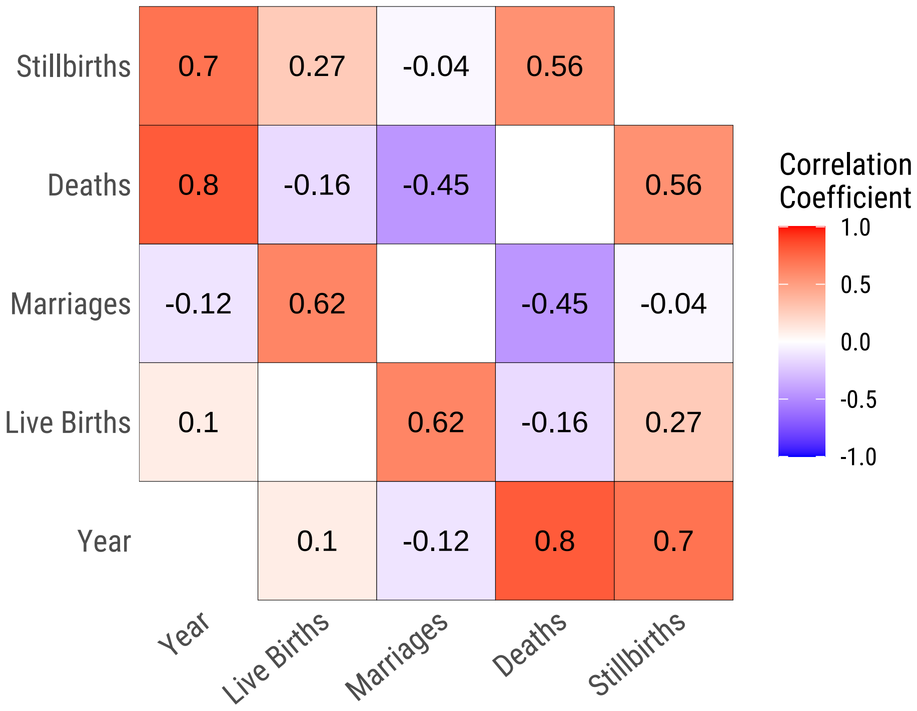
Find the correlation, if any, between the variables in the data

```
correlate(Vital_events)
```

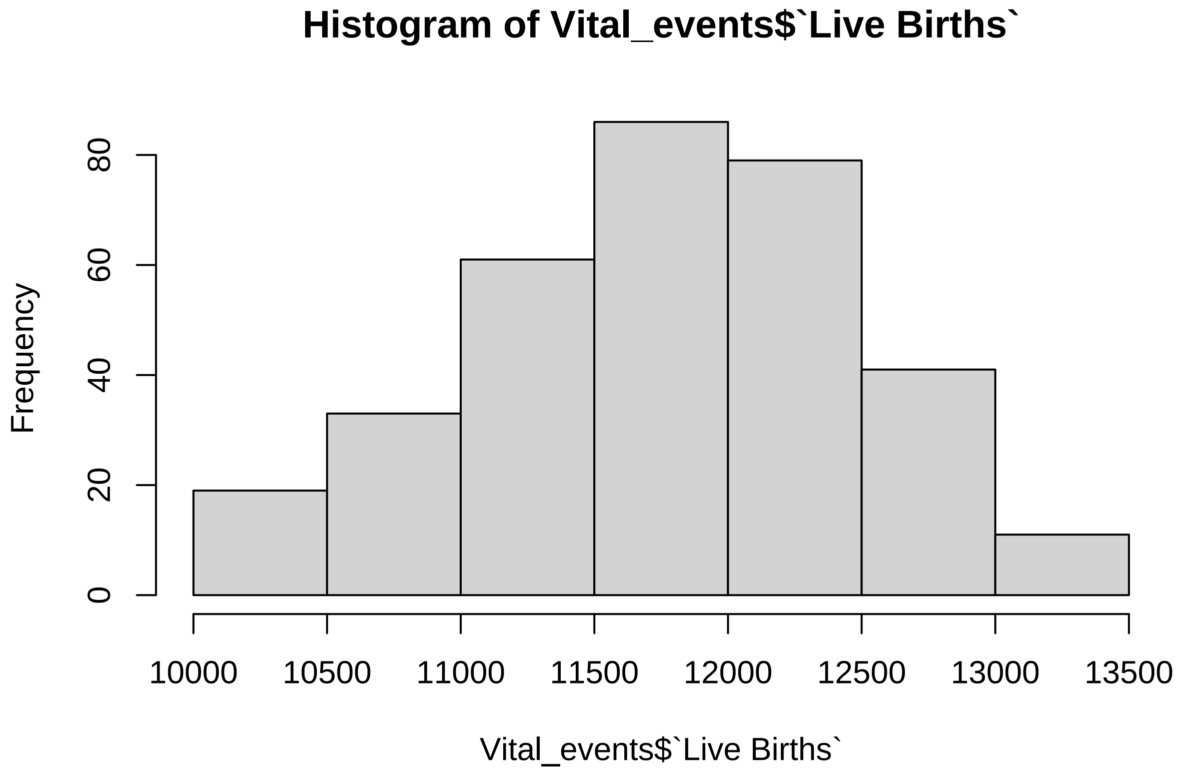
```
## # A tibble: 20 x 3
##    var1        var2        coef_corr
##    <fct>       <fct>           <dbl>
##  1 Live Births Year           0.0954
##  2 Marriages   Year          -0.115
##  3 Deaths      Year           0.799
##  4 Stillbirths Year           0.703
##  5 Year        Live Births    0.0954
##  6 Marriages   Live Births    0.618
##  7 Deaths      Live Births   -0.158
##  8 Stillbirths Live Births    0.273
##  9 Year        Marriages     -0.115
## 10 Live Births Marriages      0.618
## 11 Deaths      Marriages     -0.451
## 12 Stillbirths Marriages     -0.0353
## 13 Year        Deaths         0.799
## 14 Live Births Deaths        -0.158
## 15 Marriages   Deaths        -0.451
## 16 Stillbirths Deaths         0.564
## 17 Year        Stillbirths    0.703
## 18 Live Births Stillbirths    0.273
## 19 Marriages   Stillbirths   -0.0353
## 20 Deaths      Stillbirths    0.564
```

visualize the correlation, if any, of the data

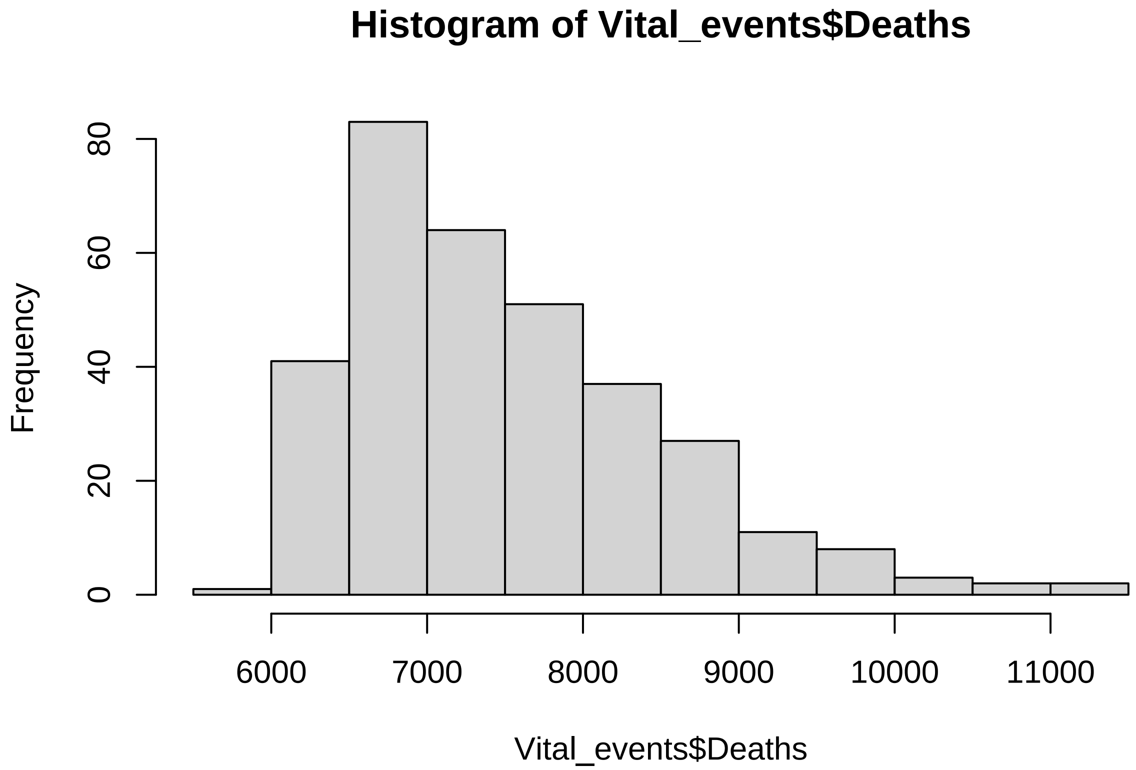
```
plot_correlate(Vital_events)
```



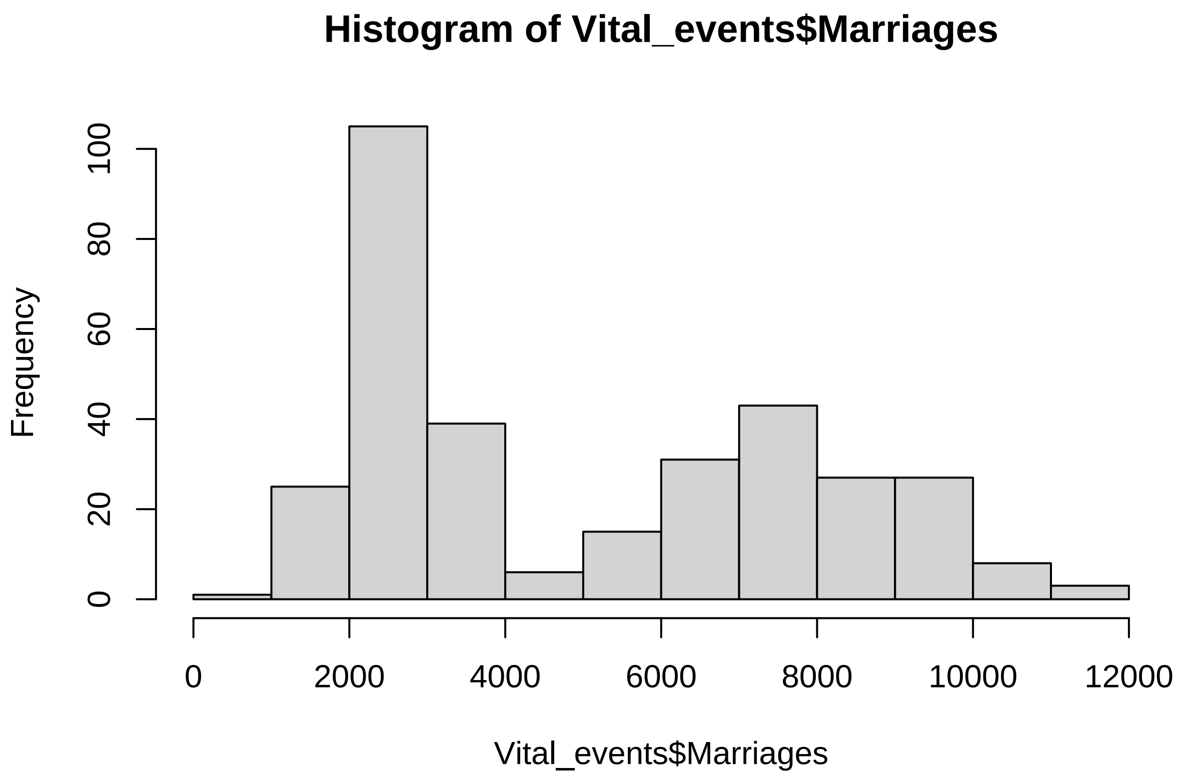
```
hist(Vital_events$`Live Births`)
```



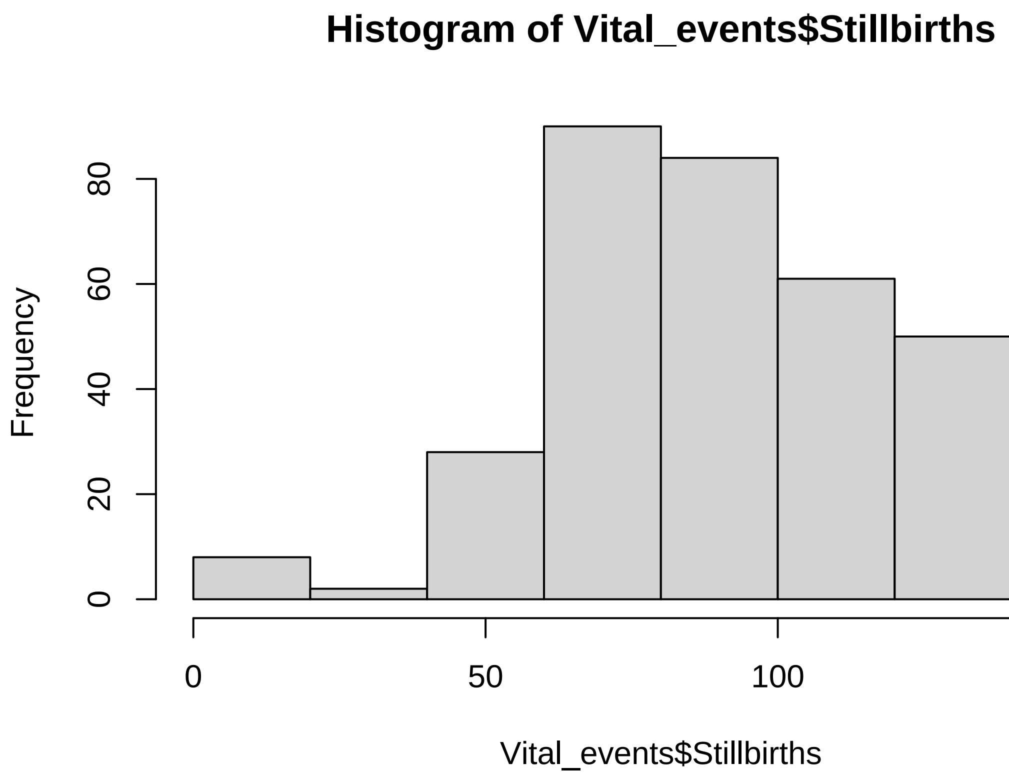
```
hist(Vital_events$Deaths)
```



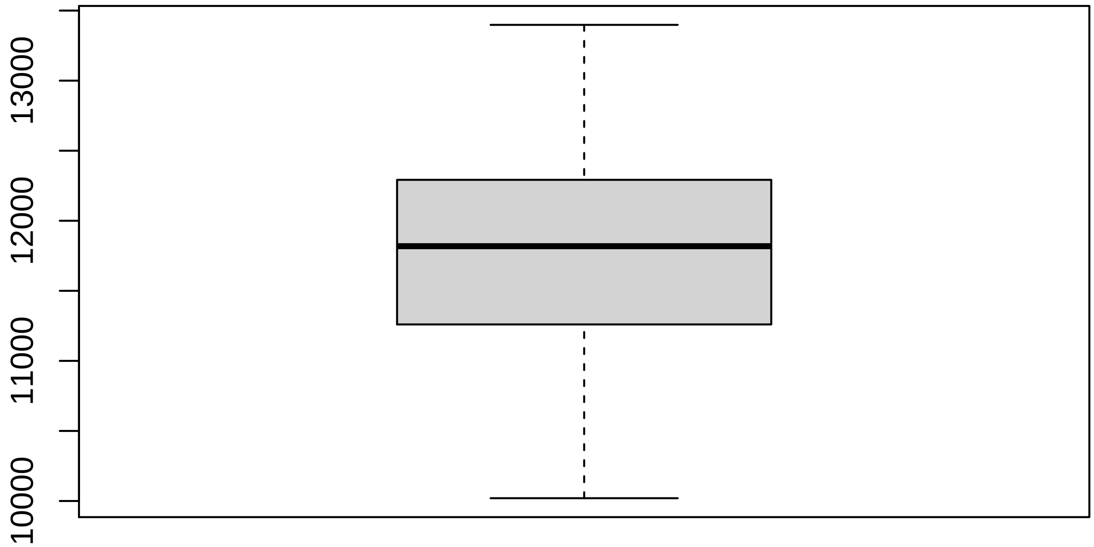
```
hist(Vital_events$Marriages)
```



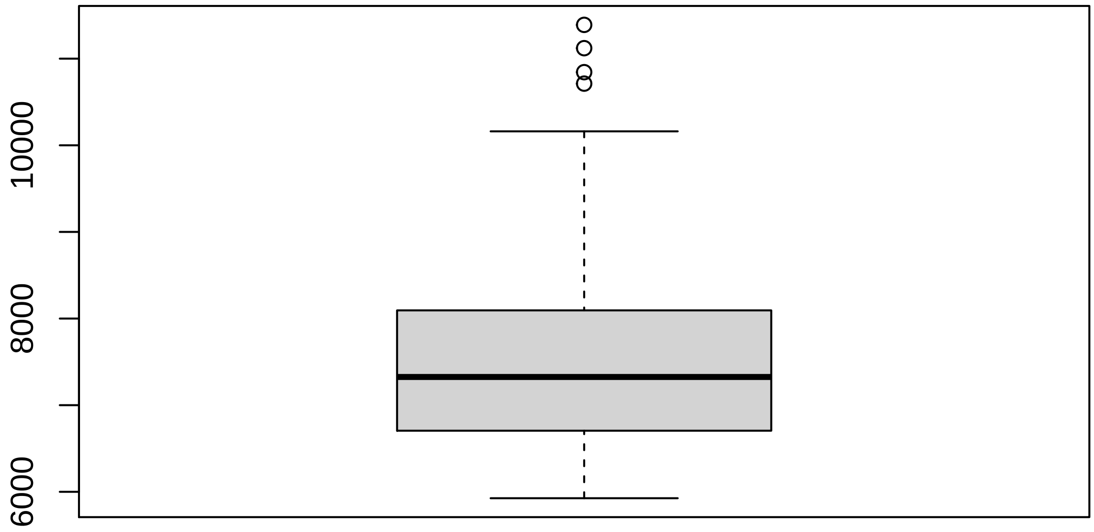
```
hist(Vital_events$Stillbirths)
```



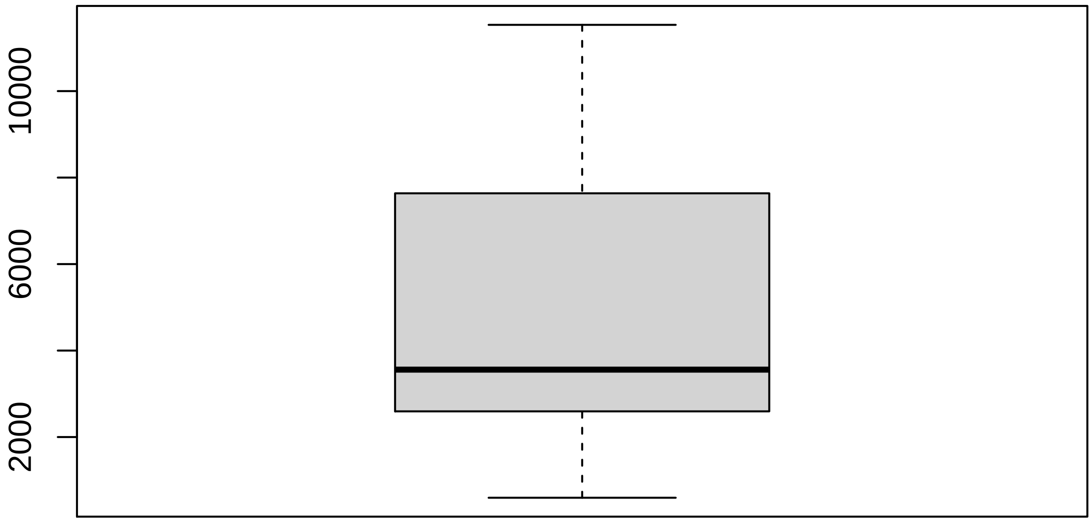
```
boxplot(Vital_events$`Live Births`)
```



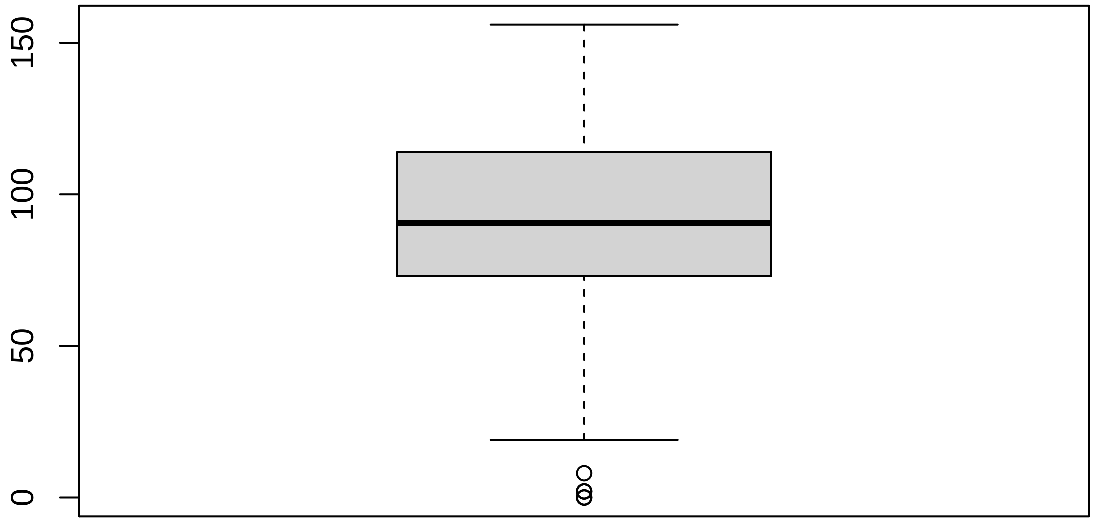
```
boxplot(Vital_events$Deaths)
```



```
boxplot(Vital_events$Marriages)
```



```
boxplot(Vital_events$Stillbirths)
```



```
eda_web_report(Vital_events)
```

```
##
##
## processing file: eda_temp.Rmd

##   |                                                                            |
##   ordinary text without R code
##
##   |                                                                            |..
## label: setup (with options)
## List of 3
##  $ echo   : logi FALSE
##  $ warning: logi FALSE
##  $ message: logi FALSE
##
##   |                                                                            |...
##   ordinary text without R code
##
##   |                                                                            |....
## label: load_packages
##   |                                                                            |......
##   ordinary text without R code
##
##   |                                                                            |.......
## label: unnamed-chunk-14 (with options)
## List of 2
##  $ echo  : logi FALSE
##  $ engine: chr "css"
##
##   |                                                                              |........
##   ordinary text without R code
##
##   |                                                                              |.........
## label: udf (with options)
## List of 3
##  $ echo   : logi FALSE
##  $ warning: logi FALSE
##  $ message: logi FALSE
##
##   |                                                                              |..........
##   ordinary text without R code
##
##   |                                                                              |...........
## label: check_variables (with options)
## List of 4
##  $ echo   : logi FALSE
##  $ warning: logi FALSE
##  $ message: logi FALSE
##  $ comment: chr ""
##
##   |                                                                              |............
##   ordinary text without R code
##
##   |                                                                              |.............
## label: create-overview
##   |                                                                              |..............
##   ordinary text without R code
##
##   |                                                                              |................
## label: overview (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                              |.................
##   ordinary text without R code
##
##   |                                                                              |..................
## label: overview-pre (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                              |..................
##   ordinary text without R code
##
##   |                                                                              |..................
## label: unnamed-chunk-15 (with options)
## List of 1
##  $ results: chr "asis"
```

```
##
##   |                                                                               |.................
##   ordinary text without R code
##
##   |                                                                               |.................
## label: unnamed-chunk-16 (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                               |.................
##   ordinary text without R code
##
##   |                                                                               |.................
## label: variables (with options)
## List of 1
##  $ results: chr "asis"

##   |                                                                               |.................
##   ordinary text without R code
##
##   |                                                                               |.................
## label: normality (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                               |.................
##   ordinary text without R code
##
##   |                                                                               |.................
## label: normality-list (with options)
## List of 2
##  $ comment: chr ""
##  $ results: chr "asis"

##   |                                                                               |.................
##   ordinary text without R code
##
##   |                                                                               |.................
## label: unnamed-chunk-17 (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                               |.................
##   ordinary text without R code
##
##   |                                                                               |.................
## label: unnamed-chunk-18 (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                               |.................
##   ordinary text without R code
##
##   |                                                                               |.................
## label: compare_numerical (with options)
## List of 1
##  $ results: chr "asis"

##   |                                                                               |.................
##   ordinary text without R code
##
##   |                                                                               |.................
## label: unnamed-chunk-19 (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                               |.................
##   ordinary text without R code
##
##   |                                                                               |.................
## label: compare-category (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                               |.................
##   ordinary text without R code
##
##   |                                                                               |.................
## label: unnamed-chunk-20 (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                               |.................
##   ordinary text without R code
##
##   |                                                                               |.................
## label: unnamed-chunk-21 (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                               |.................
##   ordinary text without R code
##
##   |                                                                               |.................
## label: unnamed-chunk-22 (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                               |.................
##   ordinary text without R code
##
##   |                                                                               |.................
## label: correlation (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                               |.................
##   ordinary text without R code
```

```
##
##   |                                                                              |..................
## label: unnamed-chunk-23 (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                              |..................
##   ordinary text without R code
##
##   |                                                                              |..................
## label: plot-correlation (with options)
## List of 1
##  $ results: chr "asis"

##   |                                                                              |..................
##   ordinary text without R code
##
##   |                                                                              |..................
## label: unnamed-chunk-24 (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                              |..................
##   ordinary text without R code
##
##   |                                                                              |..................
## label: unnamed-chunk-25 (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                              |..................
##   ordinary text without R code
##
##   |                                                                              |..................
## label: group-numerical (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                              |..................
##   ordinary text without R code
##
##   |                                                                              |..................
## label: unnamed-chunk-26 (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                              |..................
##   ordinary text without R code
##
##   |                                                                              |..................
## label: group-categorical (with options)
## List of 1
##  $ results: chr "asis"
##
```

```
##   |                                                                              |..................
##   ordinary text without R code
##
##   |                                                                              |..................
## label: unnamed-chunk-27 (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                              |..................
##   ordinary text without R code
##
##   |                                                                              |..................
## label: group-correlation (with options)
## List of 1
##  $ results: chr "asis"
##
##   |                                                                              |..................
##   ordinary text without R code
```

```
## output file: eda_temp.knit.md
```

```
## "C:/Program Files/RStudio/bin/pandoc/pandoc" +RTS -K512m -RTS eda_temp.knit.md --to html4 --from marl
```

```
##
## Output created: C:\Users\KATIES~1\AppData\Local\Temp\RtmpkX4pZh/EDA_Report.html
```