Київський національний університет імені Тараса Шевченка
Факультет Комп'ютерних наук і кібернетики
Кафедра Системного аналізу і теорії прийняття рішень

Звіт до Лабораторної роботи №1
На тему: "Попередній аналіз даних"

Студента 3 курсу
Групи САТР-3
Арзамасцева Владислава Олександровича

Київ – 2022

Зміст

Постановка задачі

Обрати довільний набір даних і обробити не менше трьох скалярних змінних для попереднього аналізу обраного набору даних. Для кожної скалярної змінної, що обробляється, необхідно:

1. Дати її класифікацію;
2. Графічно представити Полігоном частот;
3. Побудувати зображення "Скринька з вусами";
4. Підрахувати вибіркові значення: мінімального та максимального спостережень вибірки, медіани, квартилів, децилів;
5. Підрахувати вибіркові значення усіх характеристик положення центру значень: Метематичне сподівання, середнє геометричне, середнє гармонічне, моду, медіану;
6. Підрахувати вибіркові значення усіх характеристик розсіювання значень: Дисперсію, середнє квадратичне відхилення, коефіцієнт варіації, імовірнісне відхилення, розмах вибірки, інтервал концентрації розподілу;
7. Провести аналіз скошеності та гостроверхості розподілу;
8. Провести аналіз отриманих результатів передньої обробки даних та сформулювати відповідні висновки.

Для проведення аналізу візьмемо 3 змінні:

1. age
2. chol
3. trestbps

Опис вхідної інформації

Дані, які використовуються для проведення цього аналізу, взяті з набору даних, зібраного чотирма лікарнями в Клівленді, Угорщині, Швейцарії та Лонг-Біч [1]. Дані називають набором даних про хвороби серця UCI. Цей набір даних складається з 303 осіб з 14 атрибутами, де 138 осіб представлені без серцево-судинних захворювань і 165 осіб, представлених ССЗ. Спочатку було 76 атрибутів, але опубліковані експерименти стосуються використання підмножини лише 14 атрибутів:

1. age: Вік у роках
2. sex: Стать: (1 – чол., 0 – жін.)
3. cp: Тип болю у грудях: 1– типова стенокардія, 2 – нетипова стенокардія, 3 – неангінальний біль, 4 – асимптоматична
4. trestbps: Кров'яний тиск у спокої (в мм рт.ст. при надходженні в лікарню)
5. chol: Сироватковий холестерин в мг/дл
6. fbs: (рівень цукру в крові натще > 120 мг/дл) (1 = так; 0 = ні)
7. restecg: Електрокардіографічні результати у стані спокою — 0: нормальні, 1: наявність аномалії хвиль ST-T (інверсії Т-хвиль та/або висота/спад ST > 0,05 мВ), 2: відображення ймовірної або певної гіпертрофії лівого шлуночка за критеріями Естеса
8. thalach: Найбільша частота серцебиття
9. exang: Біль, що з'являється під час вправ (1 = так; 0 = ні)
10. oldpeak: Спад ST, спричинений вправами, порівняно зі станом спокою
11. slope: Нахил ST-сегмента — 1: підйомний, 2: плоский, 3: нисхідний
12. ca: Кількість основних судин (0-3), забарвлених флюороскопією
13. thal: 3 = у нормі; 6 = фіксоване відхилення; 7 = оборотний дефект
14. target: діагностика серцевих захворювань (стан ангіографічного захворювання) — 0 – звуження діаметра < 50%, 1 – звуження діаметра > 50%

Змінна 1 – скалярна якісна ординальна, змінні 2, 3, 6, 7, 9, 11, 13, 14 – скалярні якісні номінальні, змінні 4, 5, 8, 10, 12 – кількісні. Всі змінні – категоризовані (дискретні).

Результати аналізу даних

Почнемо аналіз даних із візуалізації: побудуємо полігони частот для змінних age, chol і trestbps (рисунки 1, 2 і 3 відповідно):

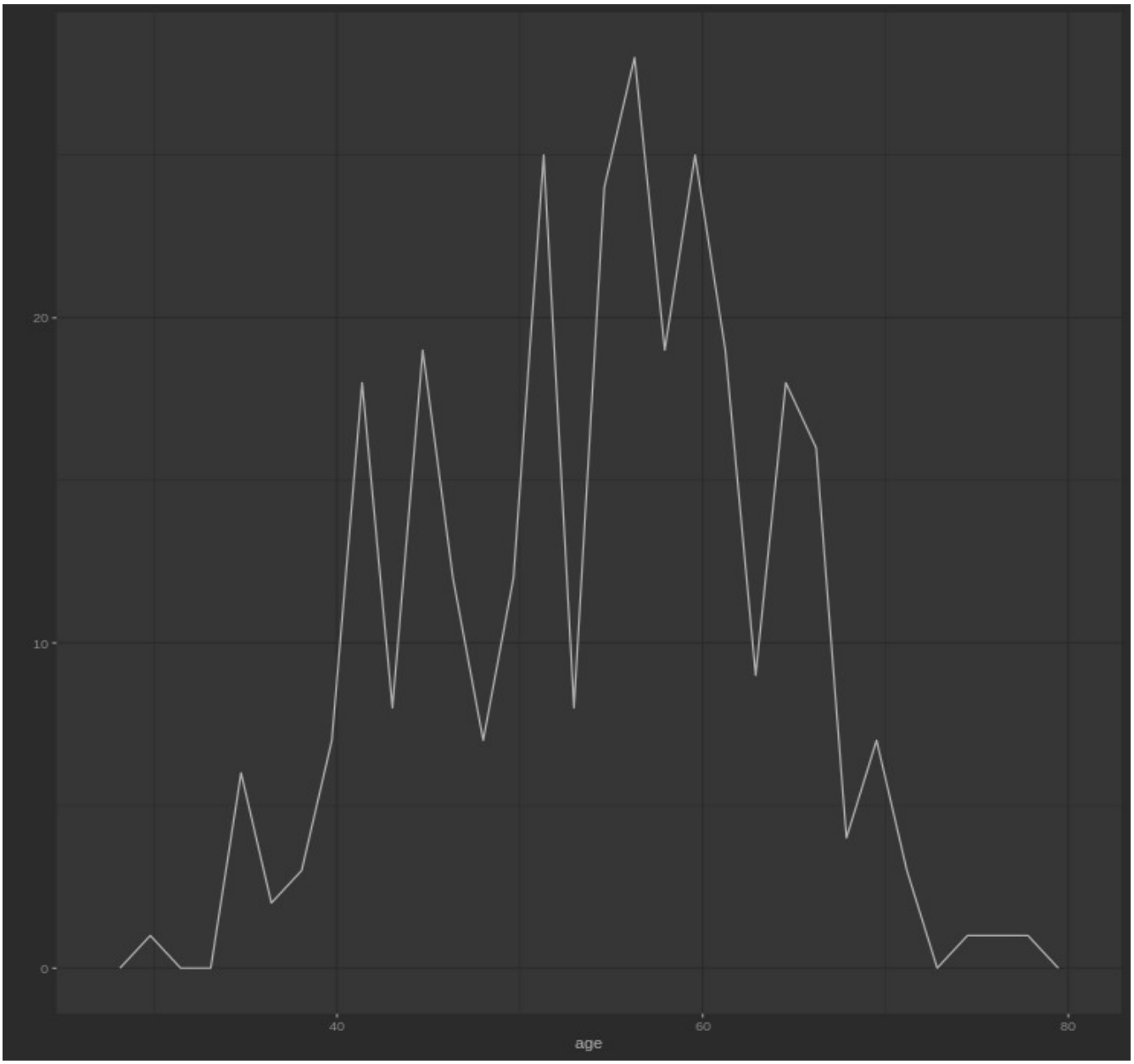

рис.1: Полігон частот для змінної age

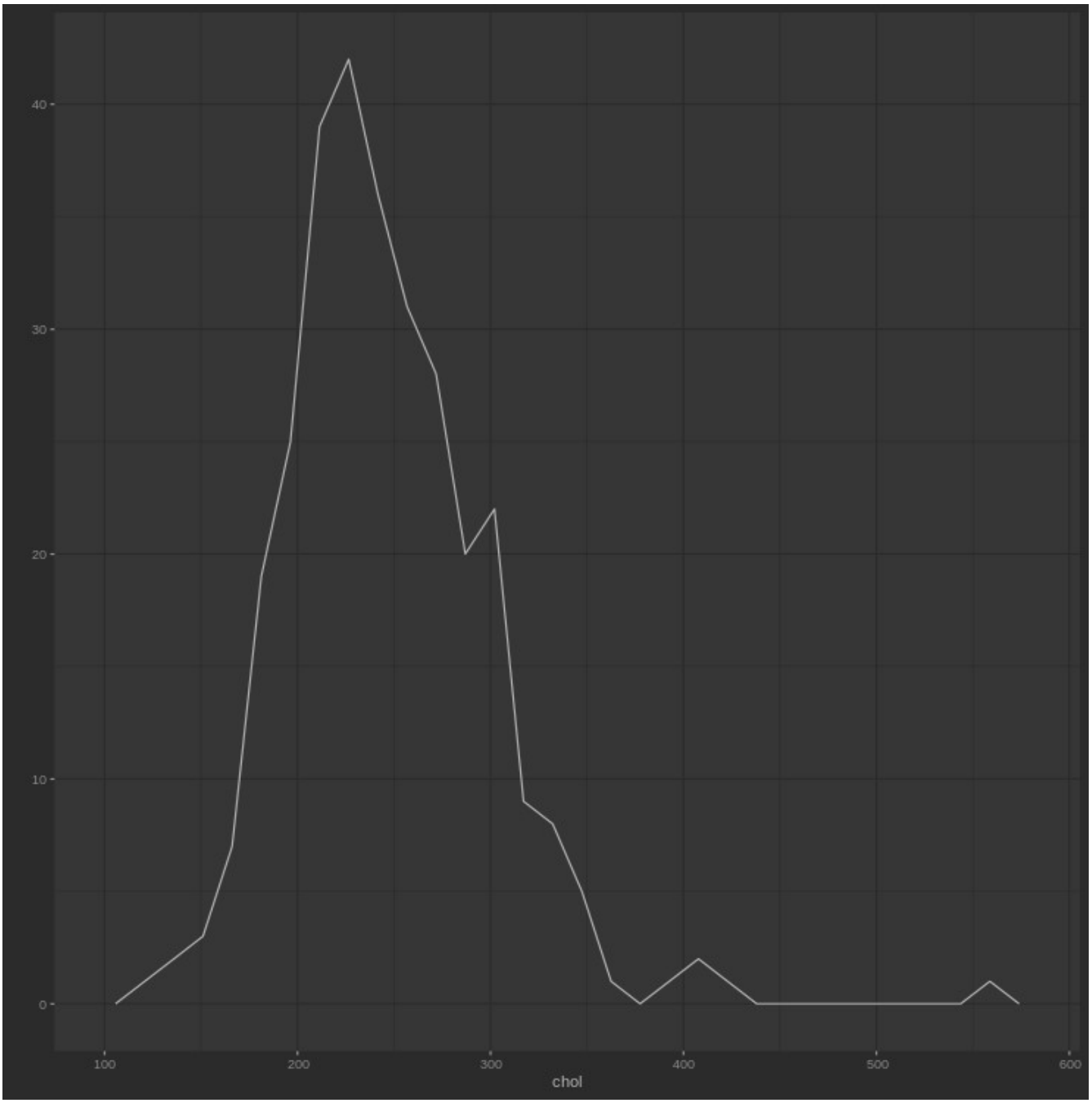

рис.2: Полігон частот для змінної chol

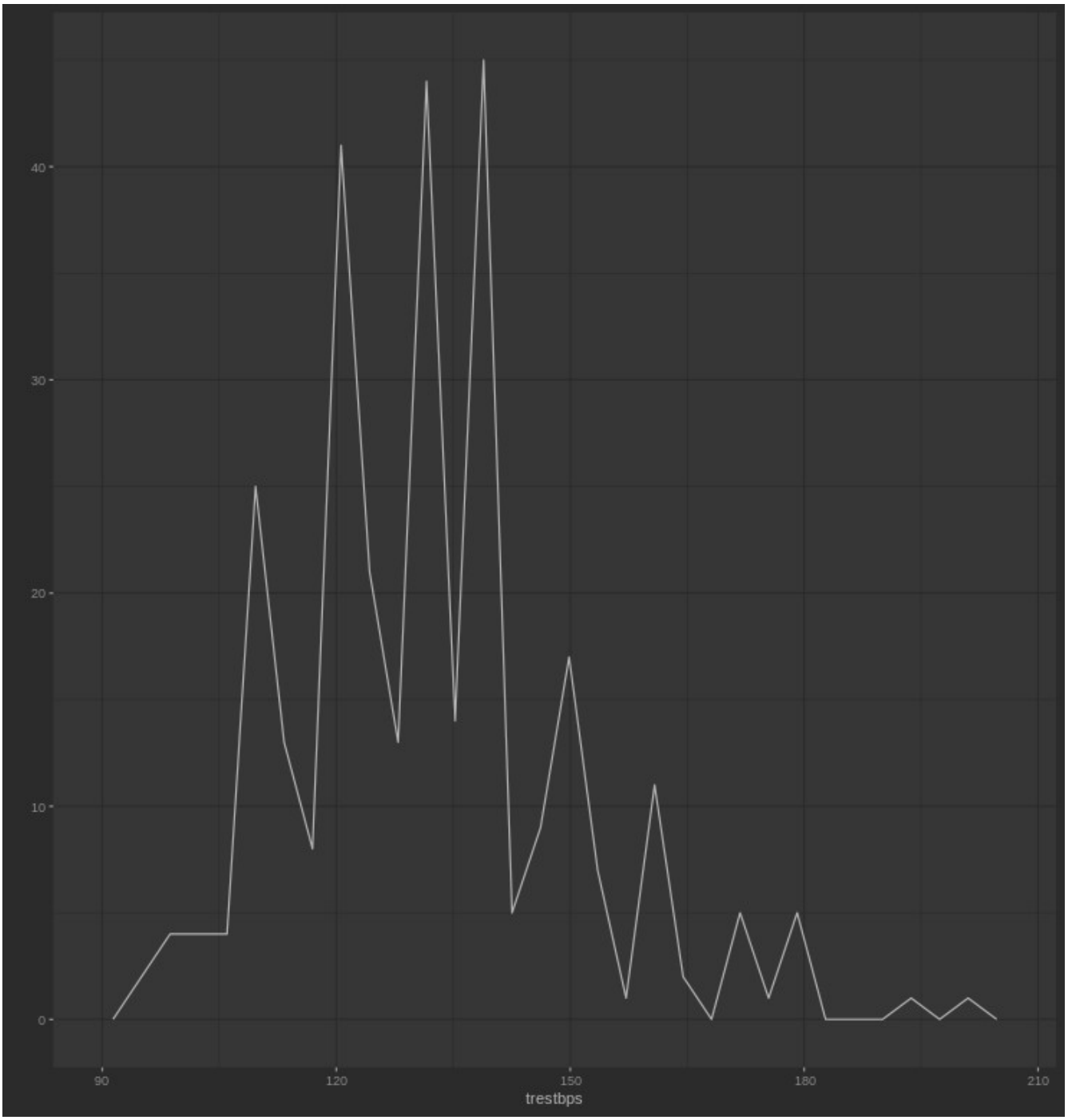

рис.3: Полігон частот для змінної trestbps

Наступним кроком буде побудова зображень "Скринька з вусами" для змінних age, chol, trestbps (рис. 4, 5, 6 відовідно)

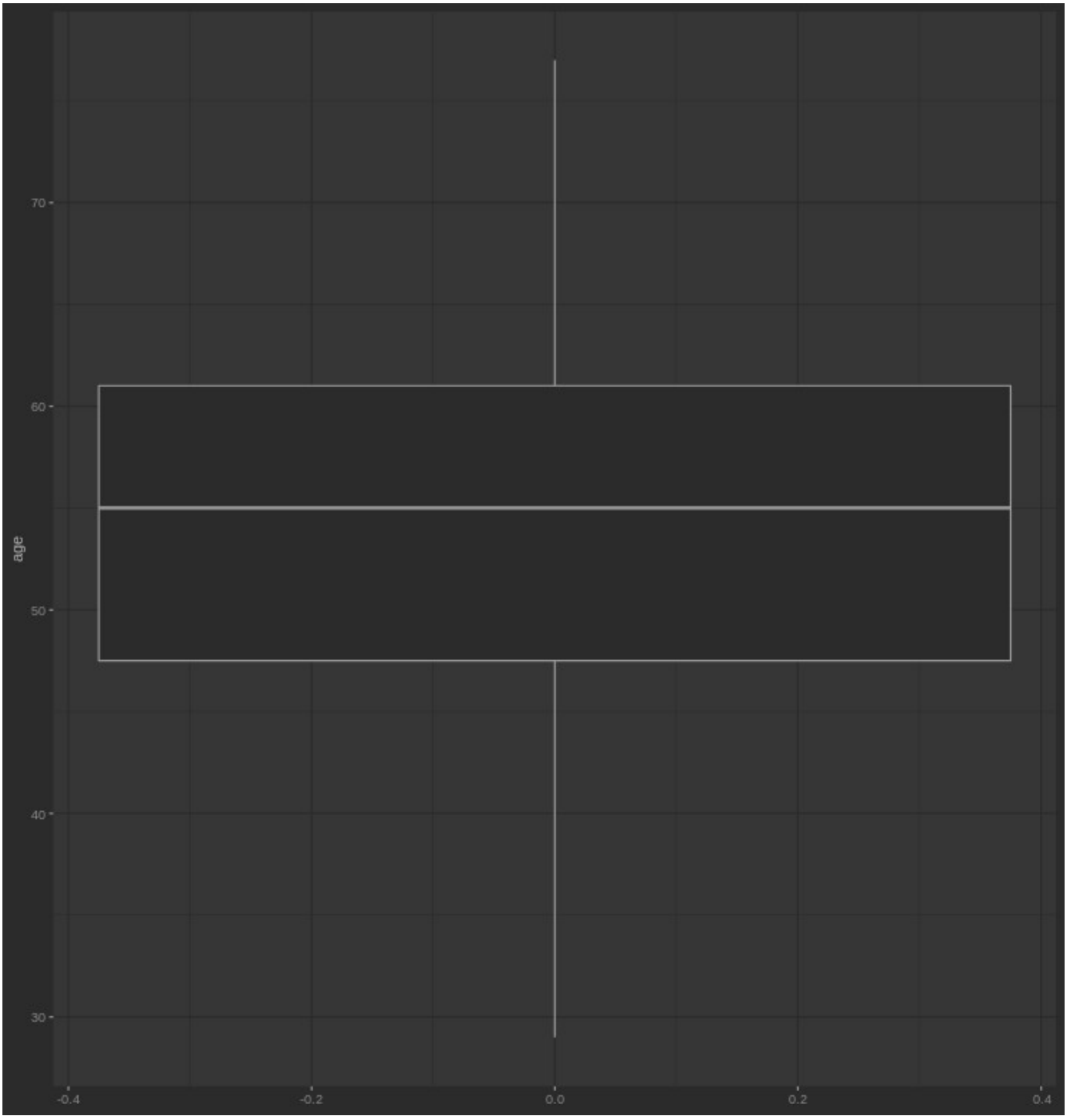

рис.4: Скринька з вусами для змінної age

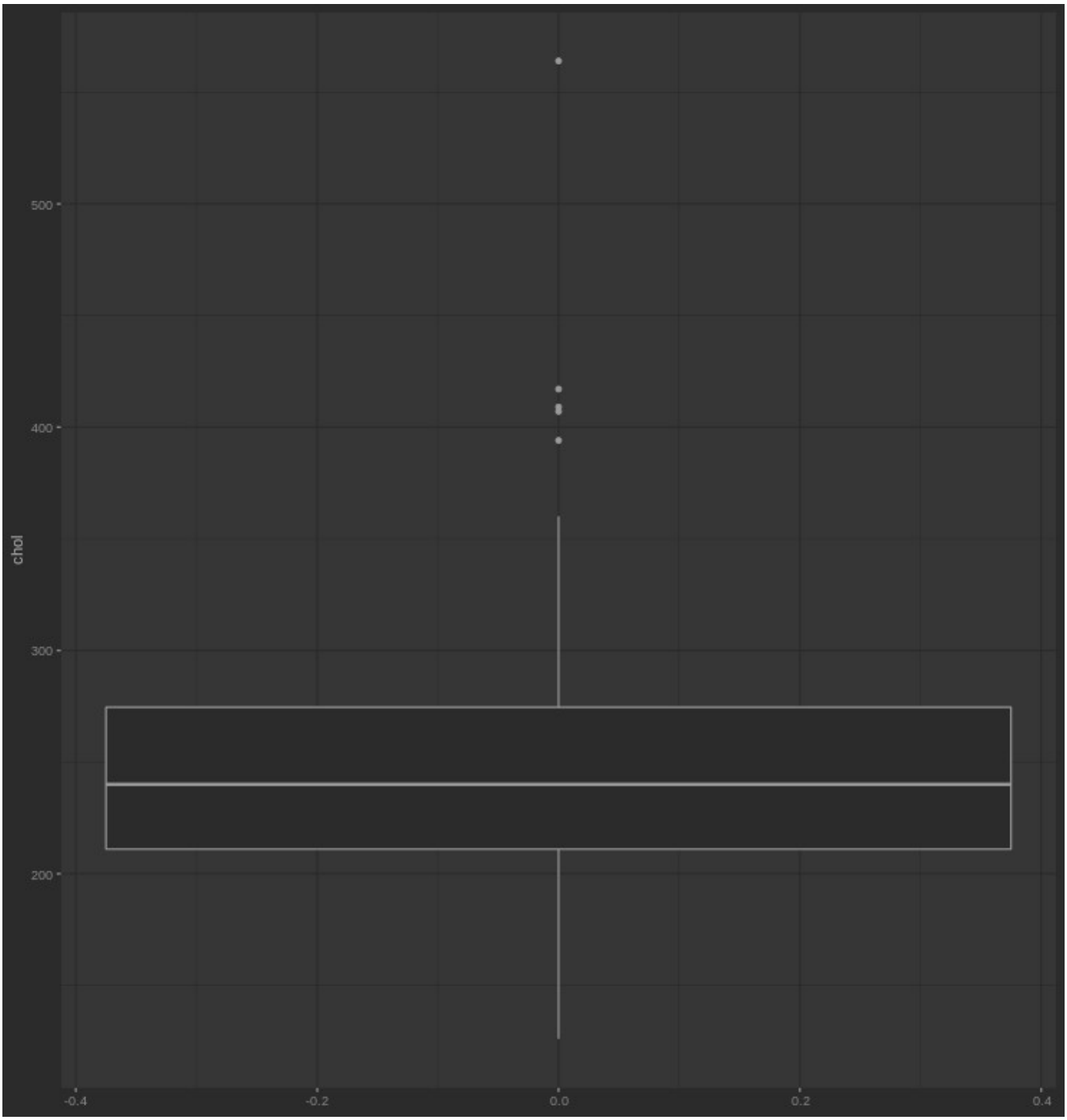

рис.5: Скринька з вусами для змінної chol

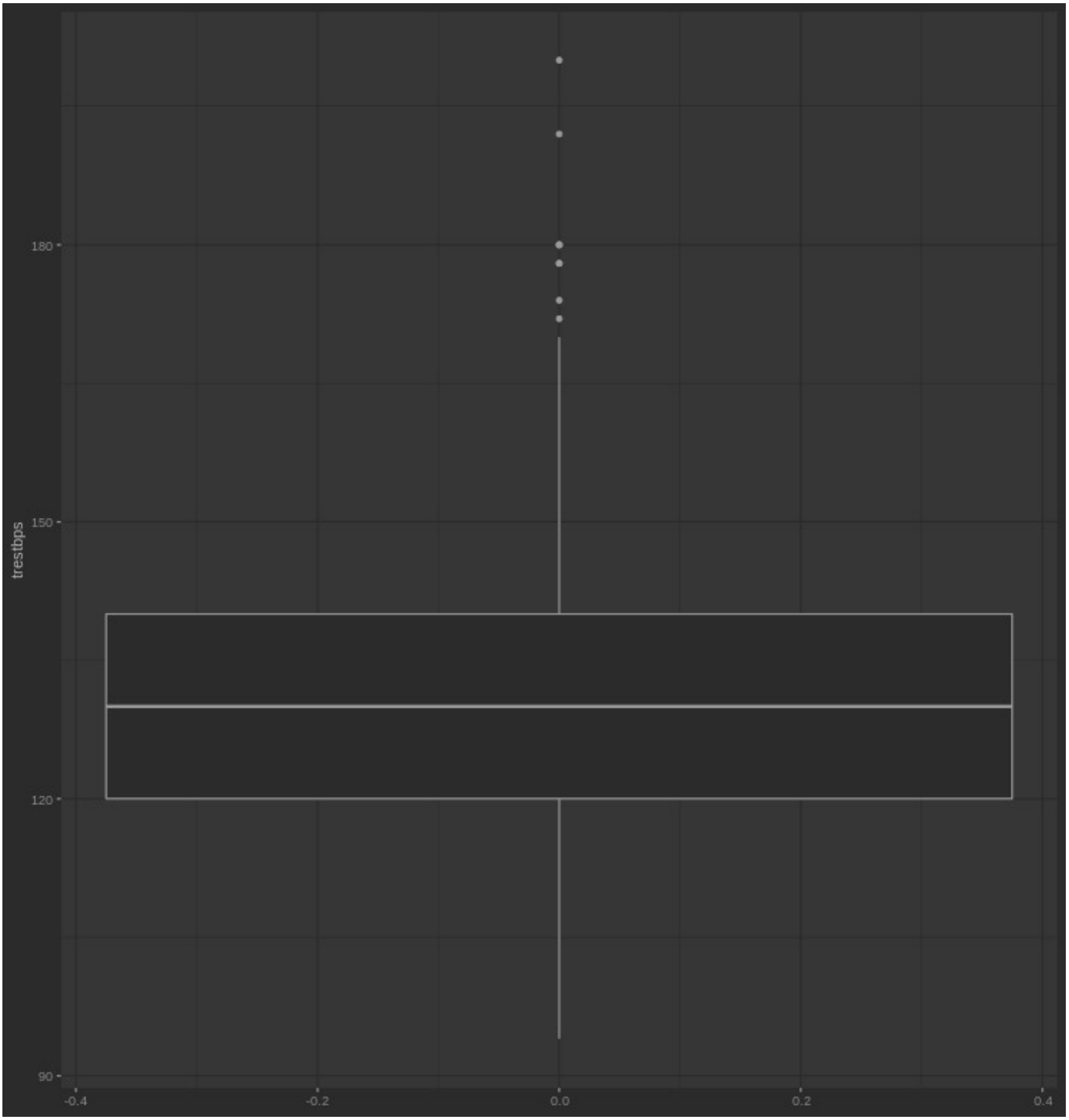

рис.6: Скринька з вусами для змінної trestbps

Проекція горизонтальної лінії скриньки на вісь ординат дає нам значення медіани, нижньої границі – нижнього квартиля, верхньої – верхнього. Проекції

кінців нижнього і верхнього вусів - найменше і найбільше значення. При наявності викидів на зображенні вони з'являються у вигляді окремих точок, відображених нижче і вище кінців вищевказаних ліній.

Знайдемо для кожної змінної найменше значення (min), найбільше (max), квартилі (Qu.), медіану(Median, 2nd Qu.), матсподівання (Mean), децилі (Deciles), геометричне середнє(Geaometric mean), середнє гармонічне (Harmonic mean), моду (Mode), дисперсію (Dispersion), стандартне відхилення (Standard Deviation), коефіцієнт варіації (Coefficient of variation), імовірнісне відхилення (Probabilistic Deviation), розмах вибірки (Sampling span), інтервал концентрації (Concentration interval), коефіцієнт асиметрії (Kurtosis) β_1 і показник ексцесу (Skewness) β_2 для змінних age, chol, trestbps (рис. 7, 8, 9 відовідно)

```
[1] "Summary of ' Age ': "
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29.00   47.50   55.00   54.37   61.00   77.00

[1] "Deciles of ' Age ': "
10% 20% 30% 40% 50% 60% 70% 80% 90%
 42  45  50  53  55  58  59  62  66

[1] "Geometric Mean of ' Age ':  53.5699673020373"

[1] "Harmonic Mean of ' Age ':  52.7308835533732"

[1] "Mode of ' Age ':  58"

[1] "Dispersion of ' Age ':  82.4845583896138"

[1] "Standard Deviation of ' Age ':  9.08210098983786"

[1] "Coefficient of Variation of ' Age ':  16.7053760694523"

[1] "Probabilistic Deviation of ' Age ':  6.75"

[1] "Sampling Span of ' Age ':  48"

[1] "Concentration Interval of ' Age ': ( 27.1200336641498 ,  81.6126396031769 )"

[1] "Kurtosis of ' Age ':  -0.56912374270958"

[1] "Skewness of ' Age ':  -0.200463188242731"
```

рис.7: аналіз змінної age

```
[1] "Summary of ' Serum cholesterol ': "
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  126.0   211.0   240.0   246.3   274.5   564.0

[1] "Deciles of ' Serum cholesterol ': "
  10%   20%   30%   40%   50%   60%   70%   80%   90%
188.0 204.0 217.6 230.0 240.0 254.0 268.0 285.2 308.8

[1] "Geometric Mean of ' Serum cholesterol ':  241.172757655525"

[1] "Harmonic Mean of ' Serum cholesterol ':  236.246442176603"

[1] "Mode of ' Serum cholesterol ':  197, 204, 234"

[1] "Dispersion of ' Serum cholesterol ':  2686.42674797281"

[1] "Standard Deviation of ' Serum cholesterol ':  51.83075098793"

[1] "Coefficient of Variation of ' Serum cholesterol ':  21.0468218785585"

[1] "Probabilistic Deviation of ' Serum cholesterol ':  31.75"

[1] "Sampling Span of ' Serum cholesterol ':  438"

[1] "Concentration Interval of ' Serum cholesterol ': ( 90.7717734388501 ,  401.75627936643 )"

[1] "Kurtosis of ' Serum cholesterol ':  4.36284085490358"

[1] "Skewness of ' Serum cholesterol ':  1.13210492871597"
```

рис.8: аналіз змінної chol

```
[1] "Summary of ' Resting blood pressure ': "
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   94.0   120.0   130.0   131.6   140.0   200.0

[1] "Deciles of ' Resting blood pressure ': "
10% 20% 30% 40% 50% 60% 70% 80% 90%
110 120 120 126 130 134 140 144 152

[1] "Geometric Mean of ' Resting blood pressure ':  130.502946680198"

[1] "Harmonic Mean of ' Resting blood pressure ':  129.418226133205"

[1] "Mode of ' Resting blood pressure ':  120"

[1] "Dispersion of ' Resting blood pressure ':  307.586453347321"

[1] "Standard Deviation of ' Resting blood pressure ':  17.5381428135171"

[1] "Coefficient of Variation of ' Resting blood pressure ':  13.3244503096527"

[1] "Probabilistic Deviation of ' Resting blood pressure ':  10"

[1] "Sampling Span of ' Resting blood pressure ':  106"

[1] "Concentration Interval of ' Resting blood pressure ': ( 79.0093339356863 ,  184.238190816789 )"

[1] "Kurtosis of ' Resting blood pressure ':  0.868396024028717"

[1] "Skewness of ' Resting blood pressure ':  0.706716972667553"
```

рис.9: аналіз змінної trestbps

Дисперсія, стандартне відхилення, коефіцієнт варіації, імовірнісне відхилення, розмах вибірки та інтервал концентрації є мірами відхилення спостережень випадкової величини від її характеристики положення центра значень. Вони вказують, наскільки суттєво можуть відрізнятися значення випадкової величини від центра зосередження значень.

Якщо $\beta_1<0$, то розподіл буде скошеним праворуч, якщо $\beta_1>0$ - ліворуч, а для нормальних розподілів $\beta_1=0$.

Якщо $\beta_2>0$, то розподіл, який досліджується, більш гостроверхий ніж нормальний з відповідними параметрами. Якщо $\beta_2<0$, то розподіл менш

гостроверхий, ніж нормальний з відповідними параметрами. Для нормального розподілу $\beta_2=0$. [3]

Висновки

З фізіологічної точки зору, вік є визначальним фактором ризику серцево-судинних захворювань. З віком податливість аорти і сонних артерій знижується. Це означає, що наша аорта і сонні артерії стають жорсткішими, і, таким чином, у літніх людей артеріальний тиск вищий, ніж у нормі, що є фактором ризику ССЗ та атеросклерозу. Більше того, вікова група 55 років і старше схильна до розвитку ССЗ, як це і було показано на частотному полігоні віку (рис. 1). Крім того, з частотного полігону холестерину (рис. 2) можна побачити, що ССЗ спостерігається в-основному у людей із показником холестерину більше норми.

Виходячи зі значень коефіцієнту асиметрії і показника ексцесу (рис. 7, 8, 9), можна сказати, що розподіл віку скошений праворуч і менш гостроверхий, ніж відповідний нормальний; розподіл холестерину скошений ліворуч і більш гостроверхий, ніж відповідний нормальний, як і розподіл тиску у стані спокою.

Список використаних джерел

1. Датасет: https://www.kaggle.com/datasets/amdirfan/predicte-heart-disease?resource=download
2. Опис змінних: https://towardsdatascience.com/heart-disease-classification-8359c26c7d83
3. Аналіз Даних, 2001, Слабоспицький

Додатки

Додаток 1: Код програми

```r
library(ggplot2)
library( package: "psych")
library( package: "DescTools")
library( package: "e1071")

dataHospital <- read.csv( file: 'data_hospital.csv')

printDelimiterWithNewLines <- function () {
  cat("\n\n================================================================================\n\n")
}

printEmptyLine <- function () {
  cat("\n")
}

buildFrequencyPolygons <- function() {
  age <- qplot(
    age,
    data = dataHospital,
    geom = 'freqpoly'
  )
  chol <- qplot(
    chol,
    data = dataHospital,
    geom = 'freqpoly'
  )
  trestbps <- qplot(
    trestbps,
    data = dataHospital,
    geom = 'freqpoly',
  )
  return(
    list(
```

main.R › analyze

```r
  )
  return(
    list(
      age = age,
      chol = chol,
      trestbps = trestbps
    )
  ) ^buildFrequencyPolygons
}

bultWhiskersAndBoxes <- function() {
  age <- ggplot(
    data = dataHospital,
    aes(y = age)
  ) + geom_boxplot()
  trestbps <- ggplot(
    data = dataHospital,
    aes(y = trestbps)
  ) +
    geom_boxplot()
  chol <- ggplot(
    data = dataHospital,
    aes(y = chol)
  ) +
    geom_boxplot()
  return(
    list(
      age = age,
      trestbps = trestbps,
      chol = chol
    )
  ) ^bultWhiskersAndBoxes
}
```

main.R › analyze

```
      trestbps = trestbps,
      chol = chol
    )
  ) ^bultWhiskersAndBoxes
}

printSummary <- function (vector, name) {
  summary <- summary(vector)
  print(paste("Summary of '", name, "': "))
  print(summary) ^printSummary
}

printDeciles <- function (vector, name) {
  deciles <- quantile(
    vector,
    probs = seq(.1, .9, by = .1)
  )
  print(paste("Deciles of '", name, "': "))
  print(deciles) ^printDeciles
}

printGeometricalMeanWithoutZeroes <- function (vector, name) {
  print(
    paste(
      "Geometric Mean of '", name, "': ", exp(mean(log(vector[vector>0])))
    )
  )
}

printHarmonicMeanWithoutZeroes <- function (vector, name) {
  print(
    paste(
      "Harmonic Mean of '", name, "': ", harmonic.mean(vector, zero = FALSE)
```

main.R › analyze

```r
printHarmonicMeanWithoutZeroes <- function (vector, name) {
  print(
    paste(
      "Harmonic Mean of '", name, "': ", harmonic.mean(vector, zero = FALSE)
    )
  )
}

printMode <- function (vector, name) {
  vectorMode <- Mode(vector)
  print(
    paste(
      "Mode of '", name, "': ", toString(vectorMode)
    )
  ) ^printMode
}

printDispersion <- function (vector, name) {
  dispersion <- var(vector)
  print(
    paste(
      "Dispersion of '", name, "': ", dispersion
    )
  ) ^printDispersion
}

printStandardDeviation <- function (vector, name) {
  sd <- sd(vector)
  print(
    paste(
      "Standard Deviation of '", name, "': ", sd
    )
  ) ^printStandardDeviation
```

```r
printStandardDeviation <- function (vector, name) {
  sd <- sd(vector)
  print(
    paste(
      "Standard Deviation of '", name, "': ", sd
    )
  ) ^printStandardDeviation
}

printCoefficientOfVariation <- function (vector, name) {
  cv <- sd(vector) / mean(vector) * 100
  print(
    paste(
      "Coefficient of Variation of '", name, "': ", cv
    )
  ) ^printCoefficientOfVariation
}

printProbabilisticDeviation <- function (vector, name) {
  pd <- IQR(vector) / 2
  print(
    paste(
      "Probabilistic Deviation of '", name, "': ", pd
    )
  ) ^printProbabilisticDeviation
}

printSamplingSpan <- function (vector, name) {
  max <- max(vector)
  min <- min(vector)
  print(
    paste(
      "Sampling Span of '", name, "': ", max - min
```

```r
printSamplingSpan <- function (vector, name) {
  max <- max(vector)
  min <- min(vector)
  print(
    paste(
      "Sampling Span of '", name, "': ", max - min
    )
  ) ^printSamplingSpan
}

printConcentrationInterval <- function (vector, name) {
  mean <- mean(vector)
  sd <- sd(vector)
  print(
    paste(
      "Concentration Interval of '", name, "': (", mean - 3 * sd, ", ", mean + 3 * sd, ")"
    )
  ) ^printConcentrationInterval
}

printKurtosis <- function (vector, name) {
  print(
    paste(
      "Kurtosis of '", name, "': ", kurtosis(vector)
    )
  )
}

printSkewness <- function (vector, name) {
  print(
    paste(
      "Skewness of '", name, "': ", skewness(vector)
```

```r
printSkewness <- function (vector, name) {
  print(
    paste(
      "Skewness of '", name, "': ", skewness(vector)
    )
  )
}

buildFrequencyPolygons()
bultWhiskersAndBoxes()

printDelimiterWithNewLines()

analyze <- function (vector, vectorName) {
  printSummary(vector, vectorName)
  printEmptyLine()
  printDeciles(vector, vectorName)
  printEmptyLine()
  printGeometricalMeanWithoutZeroes(vector, vectorName)
  printEmptyLine()
  printHarmonicMeanWithoutZeroes(vector, vectorName)
  printEmptyLine()
  printMode(vector, vectorName)
  printEmptyLine()
  printDispersion(vector, vectorName)
  printEmptyLine()
  printStandardDeviation(vector, vectorName)
  printEmptyLine()
  printCoefficientOfVariation(vector, vectorName)
  printEmptyLine()
  printProbabilisticDeviation(vector, vectorName)
  printEmptyLine()
```

main.R › analyze

```r
analyze <- function (vector, vectorName) {
  printSummary(vector, vectorName)
  printEmptyLine()
  printDeciles(vector, vectorName)
  printEmptyLine()
  printGeometricalMeanWithoutZeroes(vector, vectorName)
  printEmptyLine()
  printHarmonicMeanWithoutZeroes(vector, vectorName)
  printEmptyLine()
  printMode(vector, vectorName)
  printEmptyLine()
  printDispersion(vector, vectorName)
  printEmptyLine()
  printStandardDeviation(vector, vectorName)
  printEmptyLine()
  printCoefficientOfVariation(vector, vectorName)
  printEmptyLine()
  printProbabilisticDeviation(vector, vectorName)
  printEmptyLine()
  printSamplingSpan(vector, vectorName)
  printEmptyLine()
  printConcentrationInterval(vector, vectorName)
  printEmptyLine()
  printKurtosis(vector, vectorName)
  printEmptyLine()
  printSkewness(vector, vectorName)
  printDelimiterWithNewLines() ^analyze
}

analyze(dataHospital$age, vectorName: "Age")
analyze(dataHospital$trestbps, vectorName: "Resting blood pressure")
analyze(dataHospital$chol, vectorName: "Serum cholesterol")
```

main.R › analyze