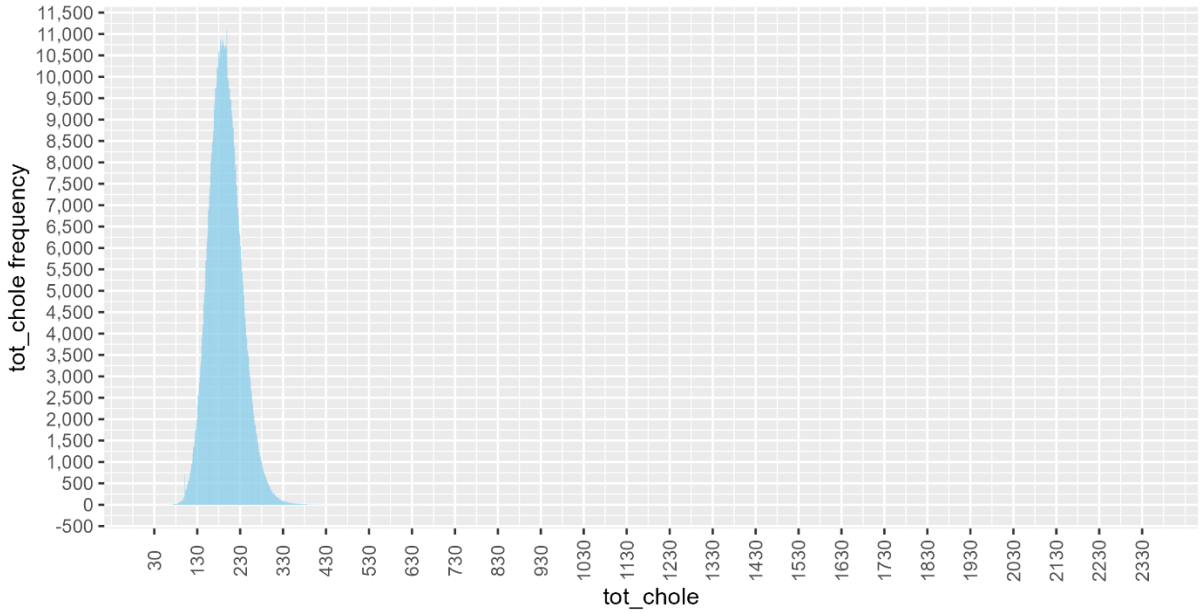
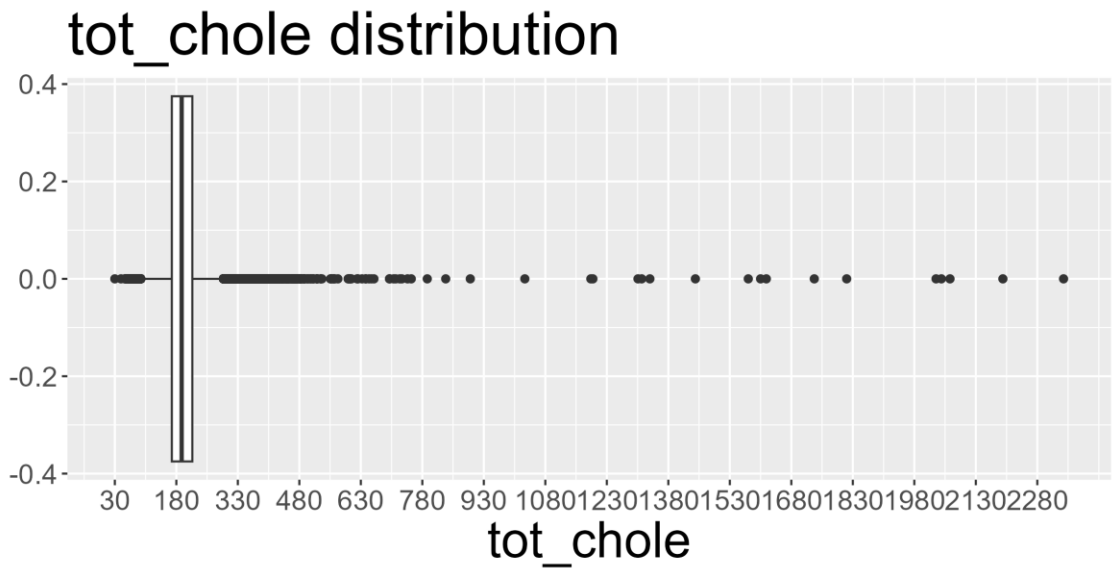
Очищення даних

EDA показало наявність викидів в оригінальному датасеті. Особливо виділяється розмах значень tot_chole:



Діаграма частот значень tot_chole оригінального датасету



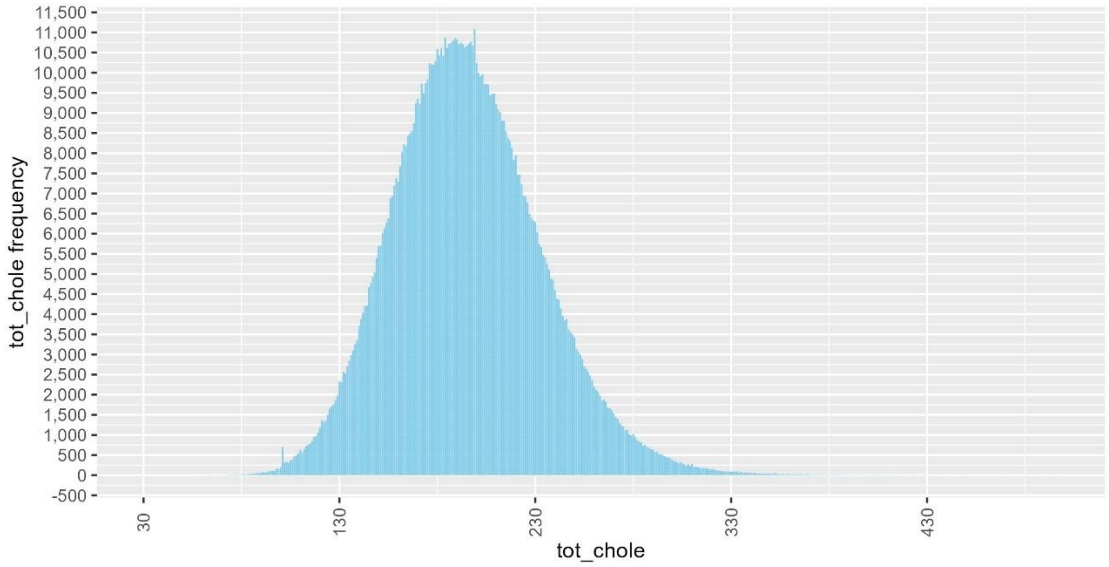
Box plot tot_chole оригінального датасету

Як виявилось, кількість спостережень, для яких tot_chole > 500, є малою – всього лиш 61 спостереження:

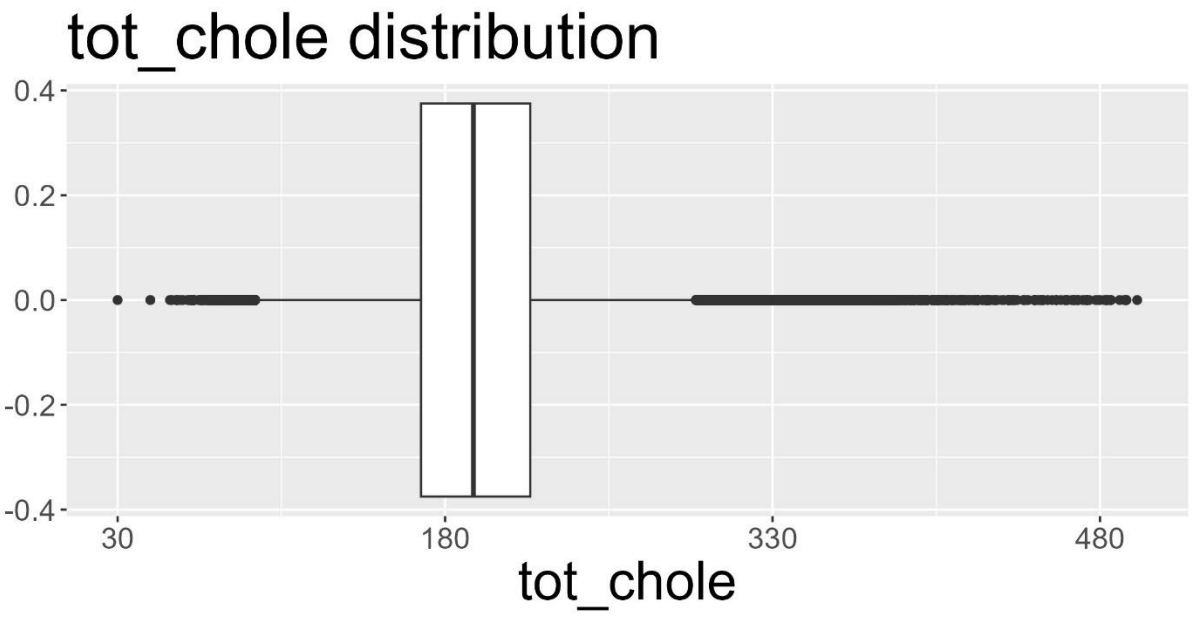
```
big_tot_chole_count <- nrow(data %>%
+   filter(tot_chole > 500))
> big_tot_chole_count
[1] 61
```

Було вирішено видалити ці 61 спостереження.

Графіки tot_chole після видалення:

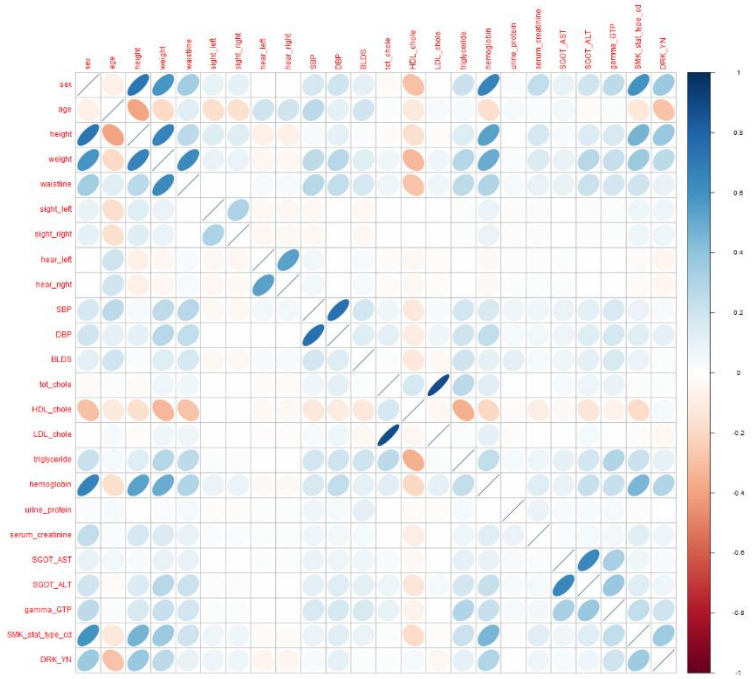


Діаграма частот значень tot_chole чистого датасету

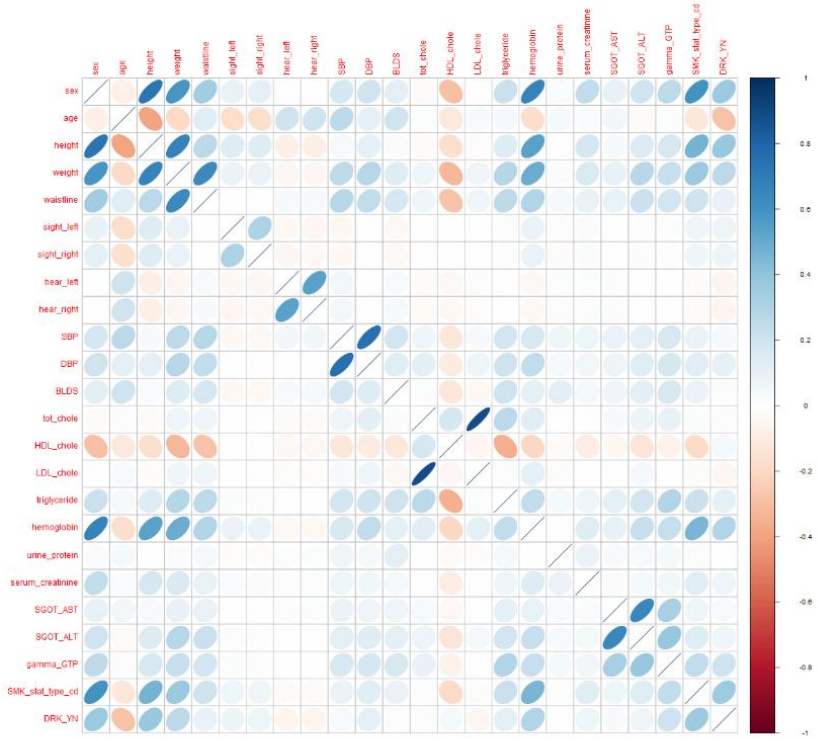


Box plot tot_chole чистого датасету

Таким чином, розподіл вирівнявся у порівнянні з початковим (хоча все ще наявне відхилення вліво). При цьому видалення відповідних спостережень не має значущого впливу на характер даних:



Матриця кореляцій оригінального датасету



Матриця кореляцій очищеного датасету

Довірчі інтервали для середніх значень

name	mean	standard_deviation	confidence_interval_norm_a	confidence_interval_norm_b
age	47.61452822803556	14.18143239760604	47.586611202626315	47.642445253444805
height	162.24050322611885	9.282932834572376	162.22222919959253	162.25877725264516
weight	63.283771351096156	12.51408744869878	63.25913659718105	63.308406105011265
waistline	81.23305652063384	11.850382483398542	81.20972831092816	81.25638473033952
sight_left	0.9808368741954878	0.6059612464003457	0.9796440020625128	0.9820297463284628
sight_right	0.9784308028778836	0.6047867022605942	0.9772402429075784	0.9796213628481889
hear_left	1.0314965237005742	0.17465544213706122	1.0311527036773709	1.0318403437237775
hear_right	1.030475625552314	0.17189209291353677	1.0301372453527	1.030814005751928
SBP	122.43204671920459	14.542805706063477	122.40341830812206	122.46067513028713
DBP	76.05226958167387	9.889232704609686	76.03280201641303	76.07173714693471
BLDS	100.42162589126829	24.170510958360314	100.3740447478647	100.46920703467188
tot_chole	195.5132747023053	38.054265230965484	195.4383625312638	195.58818687334679
HDL_chole	56.92946118367693	15.220790071277209	56.89949811850445	56.95942424884941
LDL_chole	113.00705045173734	34.95017868182186	112.93824886675199	113.0758520367227
triglyceride	132.04616537742967	100.4162789672993	131.8484897313571	132.24384102350226
hemoglobin	14.229770479499317	1.5848706873330576	14.22665056368206	14.232890395316574
urine_protein	1.0941839069328265	0.437596771436844	1.0933224706656928	1.0950453431999603
serum_creatinine	0.8604647104159857	0.48053998001597714	0.8595187377909266	0.8614106830410448
SGOT_AST	25.986995654121323	23.483270964269703	25.94076738508467	26.033223923157976
SGOT_ALT	25.75357314351891	26.306113918022902	25.70178793381105	25.80535835322677
gamma_GTP	37.12669325844057	50.36653709788572	37.02754361991925	37.225842896961886

Довірчі інтервали для медіан (bootstrap)

```
# boot median confidence interval fail
> calc_median <- function(x, i) {
+   return(median(x[i]))
+ }
> boot_medians_by_column <- bind_rows(lapply(names(data_all_numeric), function(column_name) {
+     boot_result <- boot(data_all_numeric[[column_name]], statistic = calc_median, R = 500)
+     boot_ci_result <- boot.ci(boot_result, type = "basic", conf = 0.95)
+     tibble(name = column_name,
+            median_t0 = boot_result$t0,
+            median_ci = boot_ci_result$t0)
+   })
+ )
[1] "All values of t are equal to  45 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  160 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  60 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  81 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  1 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  1 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  1 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  1 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  120 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  76 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  96 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  193 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  55 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  111 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  14.3 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  1 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  0.8 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  23 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  20 \n Cannot calculate confidence intervals"
[1] "All values of t are equal to  23 \n Cannot calculate confidence intervals"
```

```
boot_samples <- boot(data_all_numeric$age, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  45 \n Cannot calculate confidence intervals"
> boot_ci
NULL


boot_samples <- boot(data_all_numeric$height, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  160 \n Cannot calculate confidence intervals"
> boot_ci
NULL


boot_samples <- boot(data_all_numeric$weight, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  60 \n Cannot calculate confidence intervals"
> boot_ci
NULL


boot_samples <- boot(data_all_numeric$waistline, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  81 \n Cannot calculate confidence intervals"
> boot_ci
NULL


boot_samples <- boot(data_all_numeric$sight_left, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  1 \n Cannot calculate confidence intervals"
> boot_ci
NULL


boot_samples <- boot(data_all_numeric$sight_right, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  1 \n Cannot calculate confidence intervals"
> boot_ci
NULL


boot_samples <- boot(data_all_numeric$hear_left, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  1 \n Cannot calculate confidence intervals"
> boot_ci
NULL


boot_samples <- boot(data_all_numeric$hear_right, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  1 \n Cannot calculate confidence intervals"
> boot_ci
NULL


boot_samples <- boot(data_all_numeric$SBP, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  120 \n Cannot calculate confidence intervals"
> boot_ci
NULL
```

```
boot_samples <- boot(data_all_numeric$DBP, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  76 \n Cannot calculate confidence intervals"
> boot_ci
NULL

boot_samples <- boot(data_all_numeric$BLDS, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  96 \n Cannot calculate confidence intervals"
> boot_ci
NULL

boot_samples <- boot(data_all_numeric$tot_chole, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  193 \n Cannot calculate confidence intervals"
> boot_ci
NULL

boot_samples <- boot(data_all_numeric$HDL_chole, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  55 \n Cannot calculate confidence intervals"
> boot_ci
NULL

boot_samples <- boot(data_all_numeric$LDL_chole, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  111 \n Cannot calculate confidence intervals"
> boot_ci
NULL

boot_samples <- boot(data_all_numeric$triglyceride, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
> boot_ci
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 100 bootstrap replicates

CALL :
boot.ci(boot.out = boot_samples, conf = 0.95, type = "basic")

Intervals :
Level      Basic
95%   (105, 106 )
Calculations and Intervals on Original Scale
Some basic intervals may be unstable

boot_samples <- boot(data_all_numeric$hemoglobin, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  14.3 \n Cannot calculate confidence intervals"
> boot_ci
NULL

boot_samples <- boot(data_all_numeric$urine_protein, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  1 \n Cannot calculate confidence intervals"
> boot_ci
NULL

boot_samples <- boot(data_all_numeric$serum_creatinine, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  0.8 \n Cannot calculate confidence intervals"
> boot_ci
NULL

boot_samples <- boot(data_all_numeric$SGOT_AST, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  23 \n Cannot calculate confidence intervals"
> boot_ci
NULL
```

```
boot_samples <- boot(data_all_numeric$SGOT_ALT, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  20 \n Cannot calculate confidence intervals"
> boot_ci
NULL

boot_samples <- boot(data_all_numeric$gamma_GTP, statistic = calc_median, R = 100)
> boot_ci <- boot.ci(boot_samples, type = "basic", conf = 0.95)
[1] "All values of t are equal to  23 \n Cannot calculate confidence intervals"
> boot_ci
NULL
```

За допомогою bootstrap вдалось обчислити довірчий інтервал медіани лише для однієї ха рактеристики – triglyceride. Це пов'язано з наступними факторами:

1. Велика кількість спостережень – 991284 спостереження
2. Відносна одноманітність даних – значення характеристик в датасеті округлені до 1 знаку після коми.

Враховуючи ці фактори, доходимо до висновку, що використання bootstrap для визначен ня довірчих інтервалів медіан є недоцільним.

Довірчі інтервали для медіан (no bootstrap)

name	median	standard_deviation	confidence_interval_norm_a	confidence_interval_norm_b
age	45	14.18143239760604	44.972082974590755	45.027917025409245
height	160	9.282932834572376	159.9817259734737	160.0182740265263
weight	60	12.51408744869878	59.97536524608489	60.02463475391511
waistline	81	11.850382483398542	80.97667179029432	81.02332820970568
sight_left	1	0.6059612464003457	0.998807127867025	1.001192872132975
sight_right	1	0.6047867022605942	0.9988094400296947	1.0011905599703053
hear_left	1	0.17465544213706122	0.9996561799767966	1.0003438200232033
hear_right	1	0.17189209291353677	0.9996616198003859	1.000338380199614
SBP	120	14.542805706063477	119.97137158891746	120.02862841108254
DBP	76	9.889232704609686	75.98053243473916	76.01946756526084
BLDS	96	24.170510958360314	95.95241885659641	96.04758114340359
tot_chole	193	38.054265230965484	192.9250878289585	193.0749121710415
HDL_chole	55	15.220790071277209	54.97003693482752	55.02996306517248
LDL_chole	111	34.95017868182186	110.93119841501465	111.06880158498535
triglyceride	106	100.4162789672993	105.8023243539274	106.1976756460726
hemoglobin	14.3	1.5848706873330576	14.296880084182744	14.303119915817257
urine_protein	1	0.437596771436844	0.9991385637328661	1.0008614362671338
serum_creatinine	0.8	0.48053998001597714	0.7990540273749409	0.8009459726250592
SGOT_AST	23	23.483270964269703	22.953771730963346	23.046228269036654
SGOT_ALT	20	26.306113918022902	19.94821479029214	20.05178520970786
gamma_GTP	23	50.36653709788572	22.90085036147868	23.09914963852132

Варто зауважити, що довірчий інтервал значення медіани для triglyceride, обчислений за допомогою bootstrap, співпадає із інтервалом, обчисленим без bootstrap.

Доведення гіпотез

Для перевірки всіх гіпотез вважатимемо достатньою достовірність 95%.

Обчислення статистики критерія для перевірки гіпотези:

```
test_hypo <- function(x, y) {
   mean_x <- mean(x)
   mean_y <- mean(y)
   mean_d <- mean_x - mean_y
   sd_x <- sd(x)
   sd_y <- sd(y)
   nx <- length(x)
   ny <- length(y)
   se = sqrt(((sd_x * sd_x)/nx) + ((sd_y * sd_y)/ny))
   result = mean_d / se
   return(c(statistics_criteria = result, standard_error = se))
}
```

Гіпотеза 1: рівень гемоглобіну збільшується з вагою

H0: hemoglobin(obese) = hemoglobin(not_obese)

Ha: hemoglobin(obese) > hemoglobin(not_obese)

```
# hypo: weight (body mass index) influence on hemoglobin
> # h0: hemoglobin(obese) = hemoglobin(not_obese)
> # ha: hemoglobin(obese) > hemoglobin(not_obese)
> data <- data %>% mutate(mbi = data$weight / ((0.01 * data$height)^2))
> x <- (data %>% filter(mbi > 25))$hemoglobin
> y <- (data %>% filter(mbi <= 25))$hemoglobin
> test_hypo(x, y)
statistics_criteria      standard_error
       2.180586e+02        3.276411e-03
> c(mean_hemoglobin_not_obese = mean(y), mean_hemoglobin_obese = mean(x))
mean_hemoglobin_not_obese     mean_hemoglobin_obese
                  13.98845                  14.70290
```

Статистика критерія значно виходить за межі інтервалу [-1.96, 1.96] – відхиляємо нульову гіпотезу.

Отже різниця:

“mean hemoglobin (no obese): 13.98845”

“mean hemoglobin (obese): 14.70290”

Є статистично значущою, **гіпотезу підтверджено.**

Гіпотеза 2.1: Вживання алкоголю впливає на SBP

H0: SBP(Alc) = SBP(NoAlc)

Ha: SBP(Alc) > SBP(NoAlc)

```
> # hypo: Alcohol influences on SBP
> # h0: SBP(Alc) = SBP(NoAlc)
> # ha: SBP(Alc) > SBP(NoAlc)
> x <- (data %>% filter(DRK_YN == "Y"))$SBP
> y <- (data %>% filter(DRK_YN == "N"))$SBP
```

```
> test_hypo(x, y)
statistics_criteria      standard_error
        33.00968331          0.02919659
```
Статистика критерія значно виходить за межі інтервалу [-1.96, 1.96] – відхиляємо нульову гіпотезу.

Отже різниця:

"mean SBP (no drinker): 121.950355053334"

"mean SBP (drinker): 122.914125338334"
є статистично значущою, **гіпотеза підтверджена.**

Гіпотеза 2.2: Вживання алкоголю впливає на DBP

H0: DBP(Alc) = DBP(NoAlc)

Ha: DBP(Alc) > DBP(NoAlc)

```
> # hypo: Alcohol influence on DBP
> # h0: DBP(Alc) = DBP(NoAlc)
> # ha: DBP(Alc) > DBP(NoAlc)
> x <- (data %>% filter(DRK_YN == "Y"))$DBP
> y <- (data %>% filter(DRK_YN == "N"))$DBP
> test_hypo(x, y)
statistics_criteria      standard_error
       100.93139320          0.01976421
```

Статистика критерія значно виходить за межі інтервалу [-1.96, 1.96] – відхиляємо нульов у гіпотезу.

Отже різниця:

"mean DBP (no drinker): 75.0552556162157"

"mean DBP (drinker): 77.0500844698583"
є статистично значущою, **гіпотеза підтверджена**

Гіпотеза 3: вживання алкоголю впливає на загальний рівень холестерину.

H0: tot_chole(Alc) = tot_chole(NoAlc)

Ha: tot_chole(Alc) > tot_chole(NoAlc)

```
> # hypo: Alcohol influence on tot_chole
> # h0: tot_chole(Alc) = tot_chole(NoAlc)
> # ha: tot_chole(Alc) > tot_chole(NoAlc)
> x <- (data %>% filter(DRK_YN == "Y"))$tot_chole
> y <- (data %>% filter(DRK_YN == "N"))$tot_chole
> test_hypo(x, y)
statistics_criteria      standard_error
        19.52149038          0.07642627
```

Статистика критерія значно виходить за межі інтервалу [-1.96, 1.96] – відхиляємо нульову гіпотезу.

Отже різниця:

"mean tot_chole (no drinker): 194.767596870771"

"mean tot_chole (drinker): 196.259551552853"

Є статистично значущою, **гіпотезу підтверджено.**

Гіпотеза 4: вживання алкоголю впливає на рівень гемоглобіну.

H0: hemoglobin(Alc) = hemoglobin(NoAlc)

Ha: hemoglobin(Alc) > hemoglobin(NoAlc)

```
> # hypo: Alcohol influence on hemoglobin
> # h0: hemoglobin(Alc) = hemoglobin(NoAlc)
> # ha: hemoglobin(Alc) > hemoglobin(NoAlc)
> x <- (data %>% filter(DRK_YN == "Y"))$hemoglobin
> y <- (data %>% filter(DRK_YN == "N"))$hemoglobin
> test_hypo(x, y)
statistics_criteria      standard_error
       3.123530e+02        3.037635e-03
```

Статистика критерія значно виходить за межі інтервалу [-1.96, 1.96] – відхиляємо нульову гіпотезу.

Отже різниця:

"mean hemoglobin (no drinker): 13.7555536956403"

"mean hemoglobin (drinker): 14.7043682118831"

Є статистично значущою, **гіпотезу підтверджено.**

Гіпотеза 5.1: Значення SBP людей, що кинули палити (2), є вищим за значення груп 1 і 3.

H0: SBP(2) = SBP(1|3)

Ha: SBP(2) > SBP(1|3)

```
> # hypo: Smoking influence on SBP
> # h0: SBP(2) = SBP(1|3)
> # ha: SBP(2) < SBP(1|3)
> x <- (data %>% filter(SMK_stat_type_cd == 2))$SBP
> y <- (data %>% filter(SMK_stat_type_cd != 2))$SBP
> test_hypo(x, y)
statistics_criteria      standard_error
        96.32792468          0.03671778
```

Статистика критерія значно виходить за межі інтервалу [-1.96, 1.96] – відхиляємо нульову гіпотезу.

Отже різниця:

"mean sbp (1s): 121.177524870646"

"mean sbp (2s): 125.344796245591"

"mean sbp (3s): 123.582846404832"

Є статистично значущою, **гіпотезу підтверджено.**

Гіпотеза 5.2: Значення DBP людей, що кинули палити (2), є вищим за значення груп 1 і 3.

H0: DBP(2) = DBP(1|3)

Ha: DBP(2) > DBP(1|3)

```
> # hypo: Smoking influence on DBP
> # h0: DBP(2) = DBP(1|3)
> # ha: DBP(2) < DBP(1|3)
> x <- (data %>% filter(SMK_stat_type_cd == 2))$DBP
> y <- (data %>% filter(SMK_stat_type_cd != 2))$DBP
> test_hypo(x, y)
statistics_criteria      standard_error
        95.21126989          0.02559697
```

Статистика критерія значно виходить за межі інтервалу [-1.96, 1.96] – відхиляємо нульову гіпотезу.

Отже різниця:

"mean dbp (1s): 74.9162075439438"

"mean dbp (2s): 78.0592885601431"

"mean dbp (3s): 77.6101595878949"

Є статистично значущою, **гіпотезу підтверджено.**

Гіпотеза 6.1: Куріння збільшує рівень гемоглобіну.

H0: hemoglobin(3) = hemoglobin(1)

Ha: hemoglobin(3) > hemoglobin(1)

```
> # hypo: Smoking influence on hemoglobin 1
> # h0: hemoglobin(3) = hemoglobin(1)
> # ha: hemoglobin(3) > hemoglobin(1)
> x <- (data %>% filter(SMK_stat_type_cd == 3))$hemoglobin
> y <- (data %>% filter(SMK_stat_type_cd == 1))$hemoglobin
> test_hypo(x, y)
statistics_criteria      standard_error
       4.974940e+02        3.320384e-03
```

Статистика критерія значно виходить за межі інтервалу [-1.96, 1.96] – відхиляємо нульову гіпотезу.

Отже різниця:

"mean hemoglobin (1s): 13.6387173183764"

"mean hemoglobin (3s): 15.2905752456457"

є статистично значущою, **гіпотезу підтверджено.**

Гіпотеза 6.2 Після припинення паління з часом гемоглобін зменшується

H0: hemoglobin(3) = hemoglobin(2)
Ha: hemoglobin(3) > hemoglobin(2)

```
> # hypo: Smoking influence on hemoglobin 2
> # h0: hemoglobin(3) = hemoglobin(2)
> # ha: hemoglobin(3) > hemoglobin(2)
> x <- (data %>% filter(SMK_stat_type_cd == 3))$hemoglobin
> y <- (data %>% filter(SMK_stat_type_cd == 2))$hemoglobin
> test_hypo(x, y)
statistics_criteria      standard_error
       79.456856716         0.004060616
```

Статистика критерія значно виходить за межі інтервалу [-1.96, 1.96] – відхиляємо нульову гіпотезу.

Отже різниця:

"mean hemoglobin (3s): 15.2905752456457"

“mean hemoglobin (2s): 14.9679314740398”

є статично значущою, **гіпотезу підтверджено.**