

Опис датасету

Першоджерело: <https://www.kaggle.com/sooyoungheer/smoking-drinking-dataset/data>

Column	Description
sex	male, female
age	round up to 5 years
height	round up to 5 cm
weight	kg
sight_left	eyesight (left)
sight_right	eyesight (right)
hear_left	hearing left; 1(normal), 2(abnormal)
hear_right	hearing right; 1(normal), 2(abnormal)
sbp	systolic blood pressure [mm/hg]
dbp	diastolic blood pressure [mm/hg]
blds	blds or fsg (fasting blood glucose) [mg/dl]
tot_chole	total cholesterol [mg/dl]
hdl_chole	hdl cholesterol [mg/dl]
ldl_chole	ldl cholesterol [mg/dl]
triglyceride	triglyceride [mg/dl]
hemoglobin	hemoglobin [g/dl]
urine_protein	protein in urine, 1(-), 2(+/-), 3(+1), 4(+2), 5(+3), 6(+4)
serum_creatinine	serum(blood) creatinine[mg/dL]
SGOT_AST	SGOT(Glutamate-oxaloacetate transaminase) AST(Aspartate transaminase)[IU/L]
SGOT_ALT	ALT(Alanine transaminase)[IU/L]
gamma_GTP	y-glutamyl transpeptidase[IU/L]
SMK_stat_type_cd	Smoking state, 1(never), 2(used to smoke but quit), 3(still smoke)
DRK_YN	Drinker or Not

Гіпотези

1. Рівень гемоглобіну збільшується з вагою

Збільшення ваги = збільшення загальної кількості клітин в організмі. Кожна клітина потребує кисень, отже і загальна потреба у гемоглобіні збільшується.

2. Вживання алкоголю впливає на SBP та DBP

Під дією алкоголю серце починає активніше працювати. Відповідно зростає тиск в судинах.

3. Вживання алкоголю впливає на рівень загального холестерину (tot_chole)

Загальновідомим є свідчення того, що деяким людям із високим рівнем холестерину корисно пити помірну кількість червоного вина.

4. Вживання алкоголю впливає на рівень гемоглобіну

Процес засвоєння етанолу C_2H_5OH вимагає великої кількості кисню, який виступає каталізатором розкладу спирту до CO_2 , H_2O . Таким чином потреба клітин у кисні, у чийх мітохондріях буде відбуватись реакція, збільшується. Отже організм має забезпечити доставку понаднормової кількості кисню до клітин, тому збільшується кількість «транспортів» кисню – гемоглобіну.

5. Значення обох видів тиску людей, що кинули палити (2), є вищим за значення груп 1 і 3

Люди рідко позбавляються шкідливих звичок за власним бажанням, можливо люди групи 2 набули інших хронічних захворювань

6. Куріння збільшує рівень гемоглобіну, Після припинення паління з часом гемоглобін зменшується.

Безпосередньо процес паління призводить до надходження чадного газу CO до легень. Чадний газ зв'язується з гемоглобіном з майже такою самою ефективністю, як і кисень. Таким чином частина гемоглобіну в тілі курця транспортує чадний газ замість кисню, тобто є «зіпсованою». Організм виділяє додатковий гемоглобін, щоб компенсувати нестачу, спричинену «зіпсованими» тілами.

Початкове очищення даних

Датасет має високу якість даних – пропущені значення відсутні, всі спостереження є придатними до використання у EDA:

```
# load and display data
```

```
data <- read_csv('original-data.csv')
```

```
glimpse(data)
```

```
# check if there are any NaN values in whole data
```

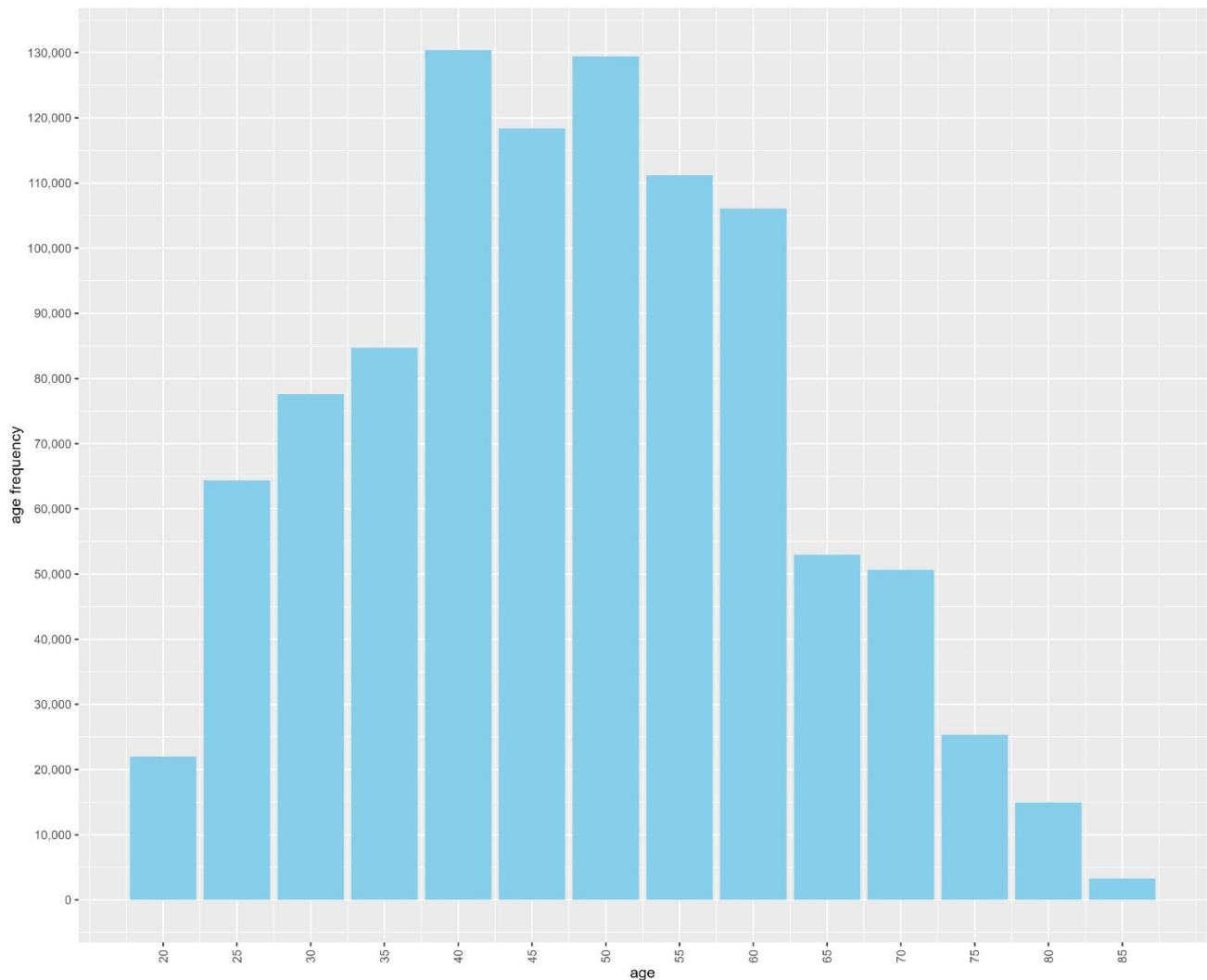
```
nan_check <- data %>% summarise_all(~ any(is.nan(.)))
```

```
print(nan_check)
```

```
> # check if there are any NaN values in whole data
> nan_check <- data %>% summarise_all(~ any(is.nan(.)))
> print(nan_check)
# A tibble: 1 × 24
  sex    age    height weight waistline sight_left sight_right hear_left hear_right SBP    DBP    BLDS    tot_chole HDL_chole
  <lgl> <lgl> <lgl>   <lgl>   <lgl>   <lgl>   <lgl>   <lgl>   <lgl>   <lgl> <lgl> <lgl> <lgl>   <lgl>
1 FALSE FALSE FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE FALSE FALSE FALSE  FALSE
# i 10 more variables: LDL_chole <lgl>, triglyceride <lgl>, hemoglobin <lgl>, urine_protein <lgl>, serum_creatinine <lgl>,
#   SGOT_AST <lgl>, SGOT_ALT <lgl>, gamma_GTP <lgl>, SMK_stat_type_cd <lgl>, DRK_YN <lgl>
```

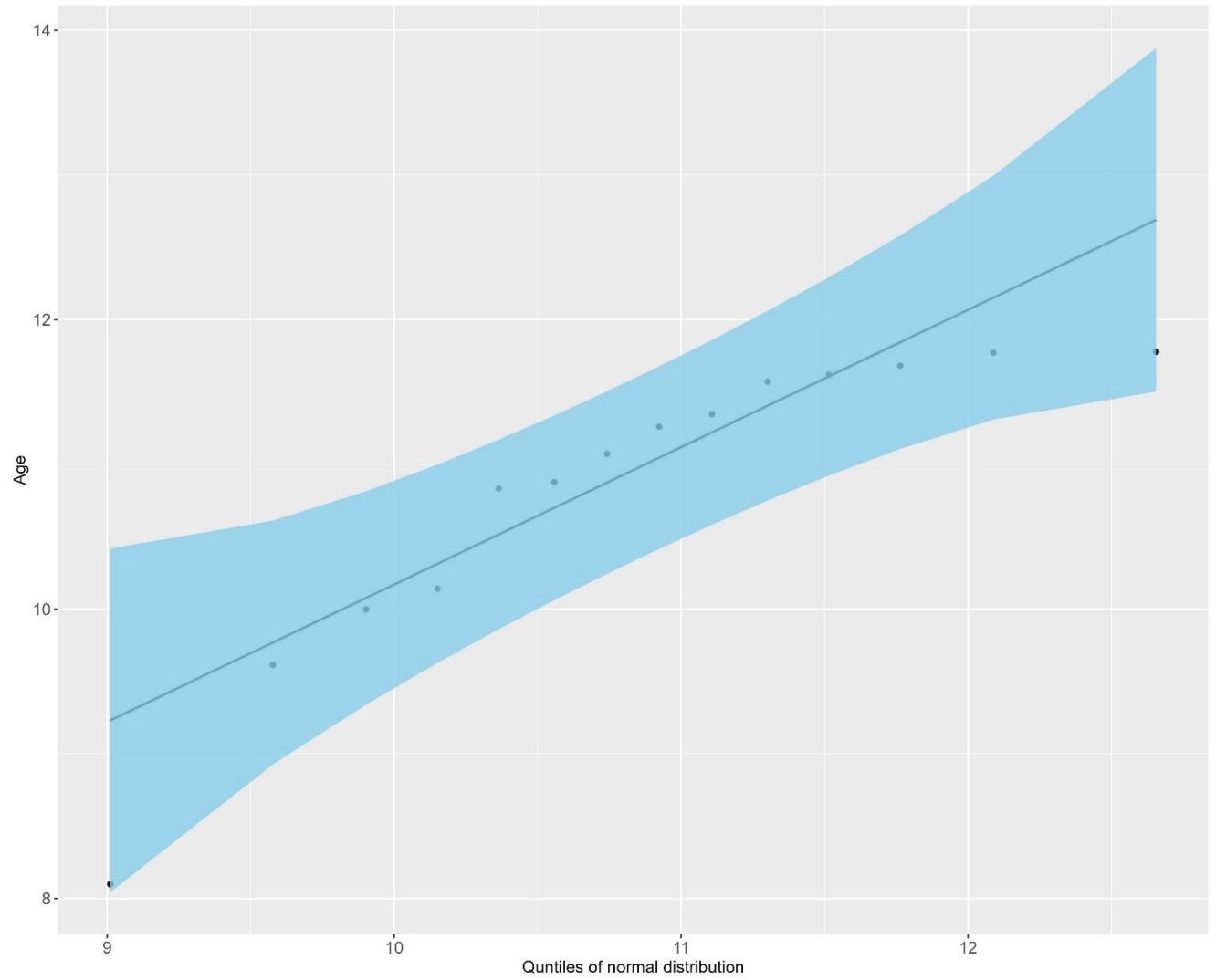
Огляд характеристик

Age (округлений з точністю до 5 років)



Діаграма частот значень age

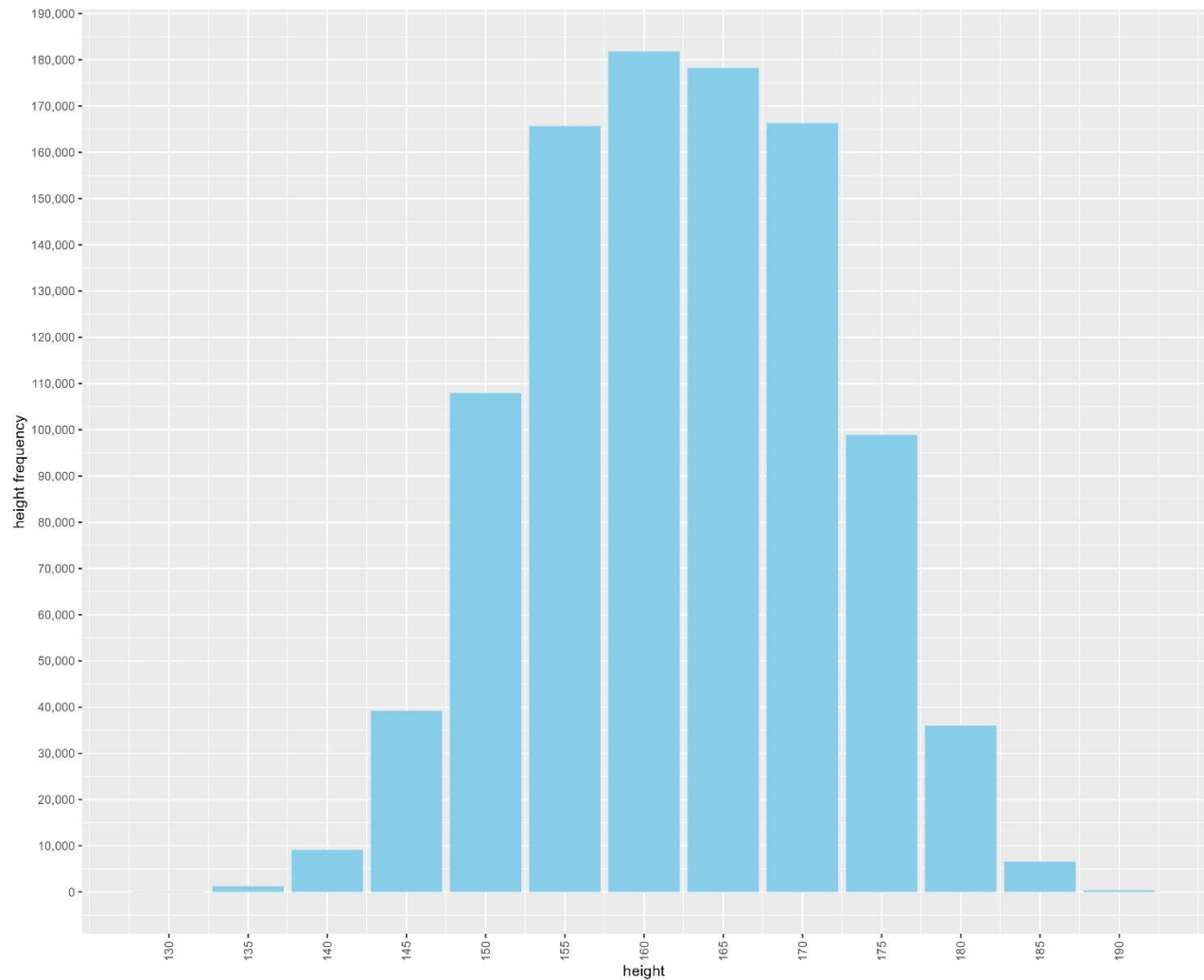
Маємо нормальний розподіл, наявне невелике зміщення вліво. Можливо населення Південної Кореї є не старим. Щоб додатково впевнитись у характері розподілу, побудуємо qq plot для age:



Квантиль-квантильний прологарифмований графік частот age.

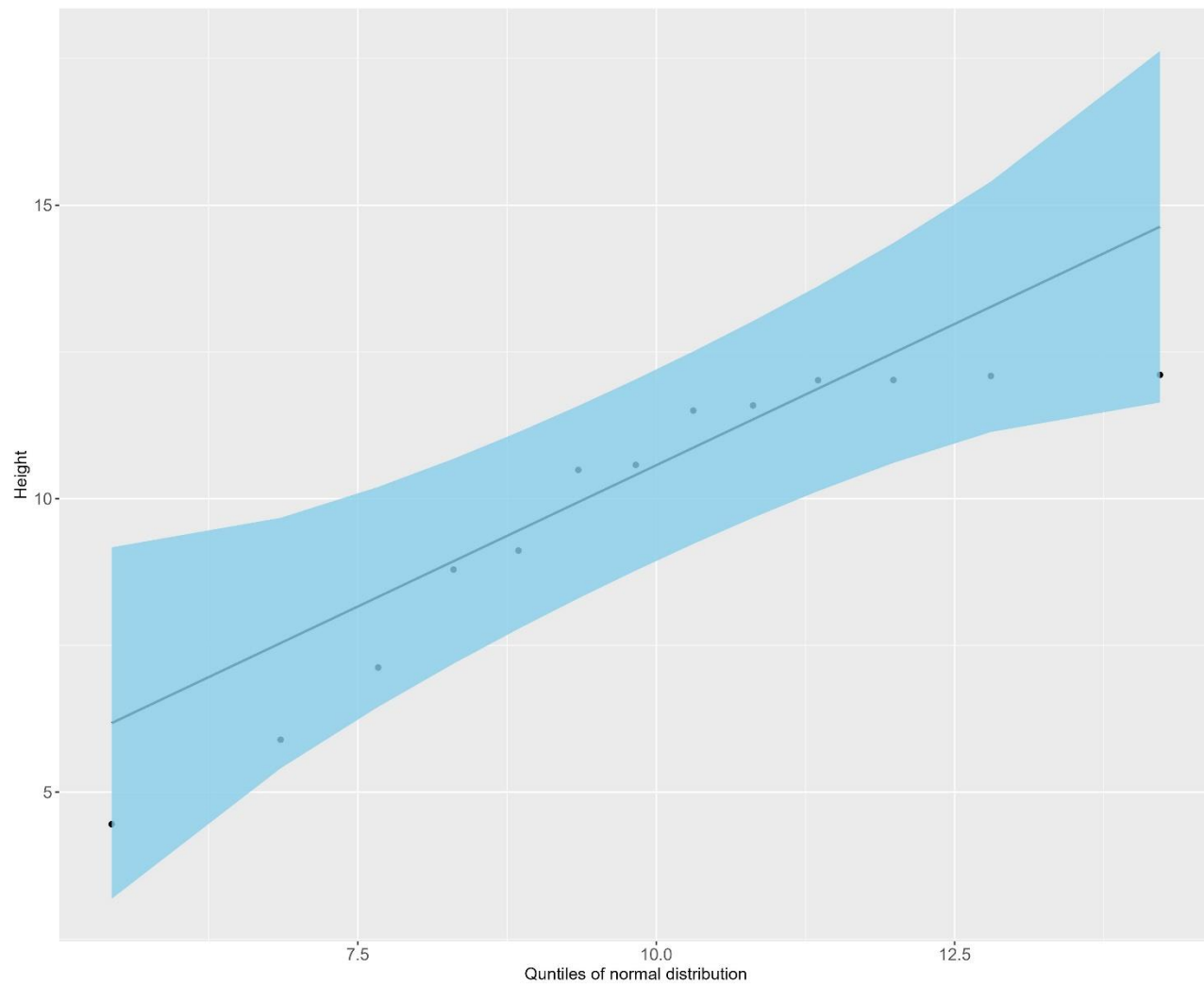
Height (округлений з точністю до 5 см)

Маємо чітко виражений нормальний розподіл:



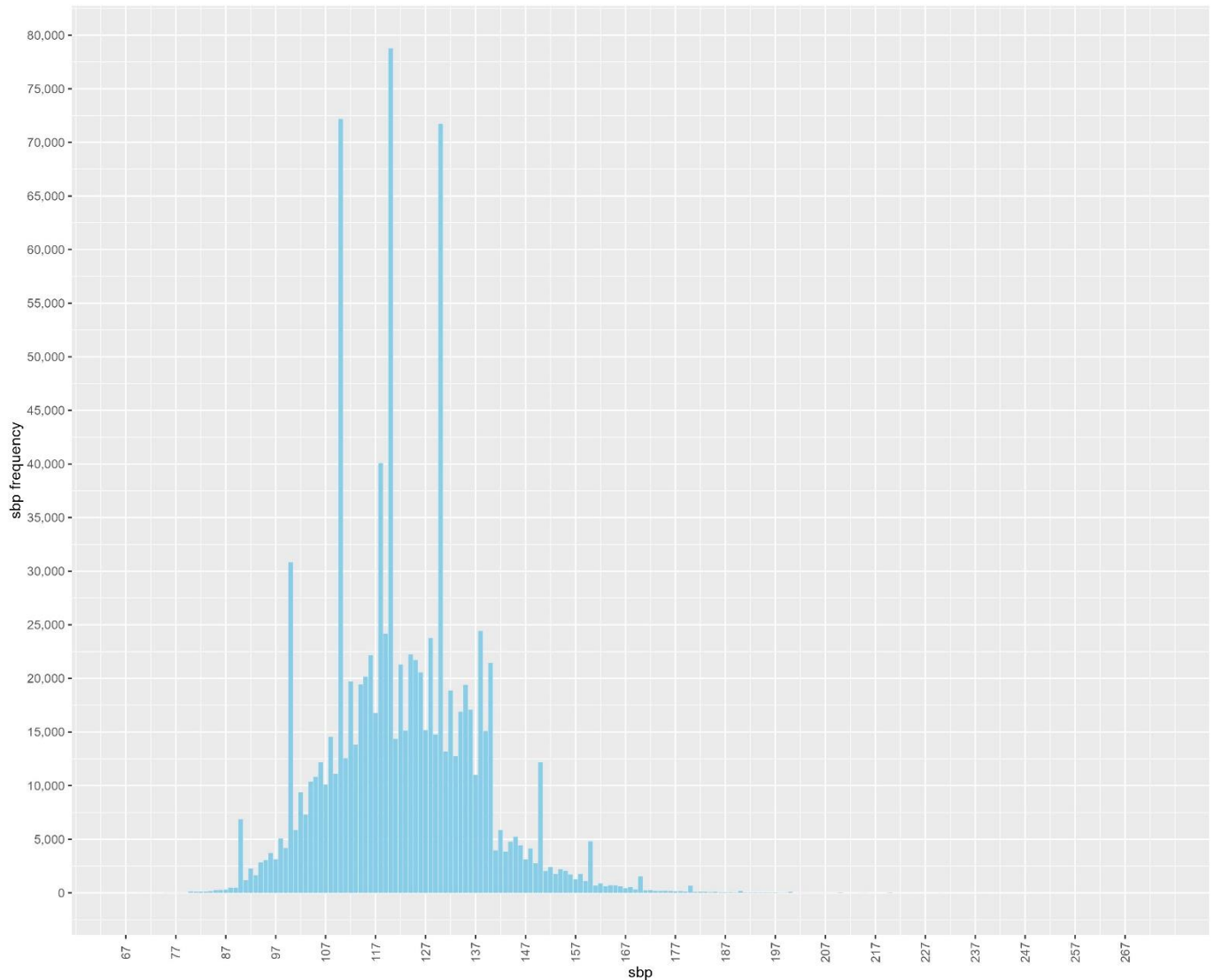
Діаграма частот height

Побудуємо qq plot для height щоб додатково пересвідчитись:



Квантиль-квантильний прологарифмований графік частот Height.

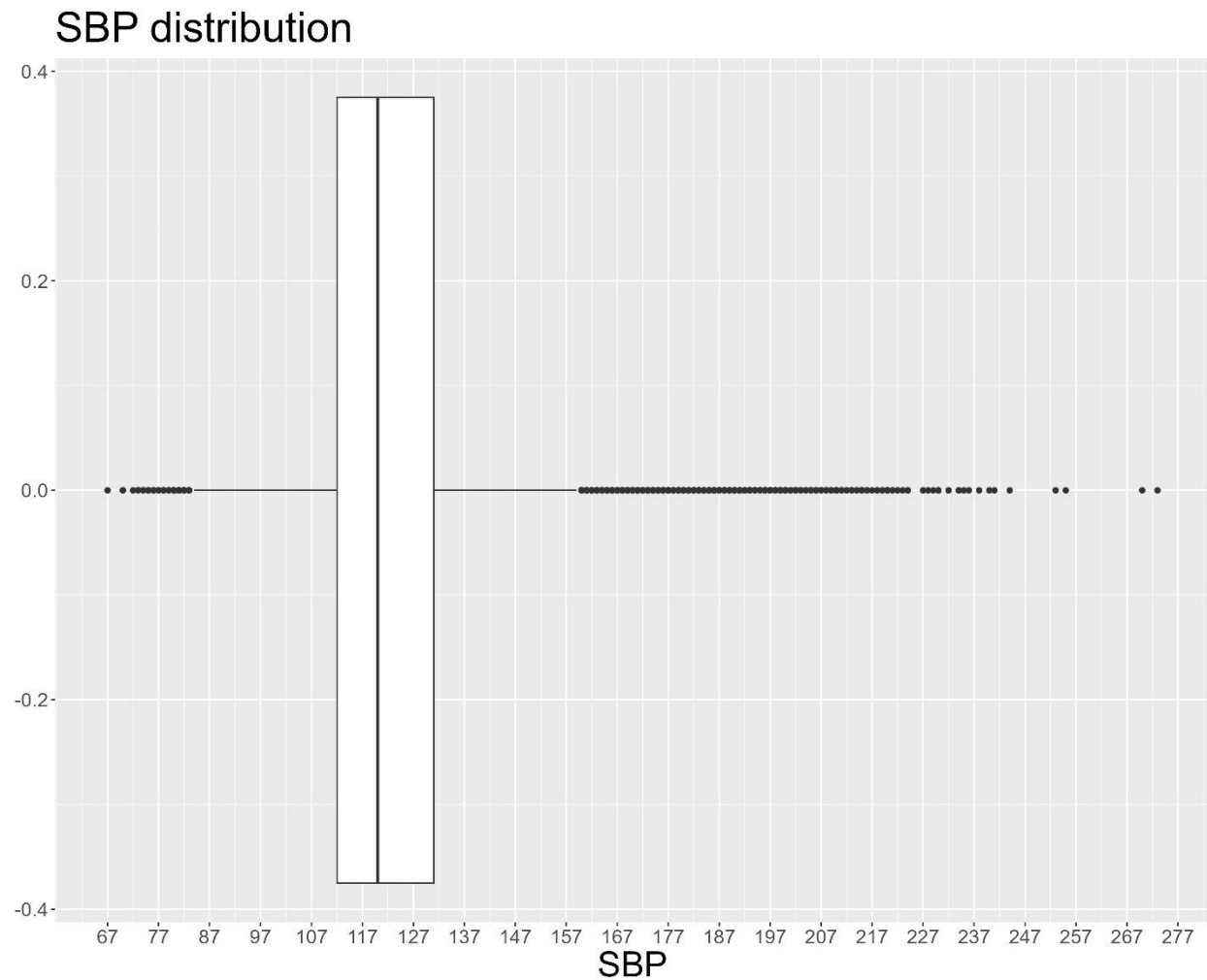
SBP (systolic blood pressure, тиск в судинах під час скорочення серця)



Діаграма частот sbp

Розподіл є віддалено схожим на нормальний, наявні 5 найтипівіших значень. При цьому графік не відображає частоти для багатьох ($sbp \geq 207$) значень sbp.

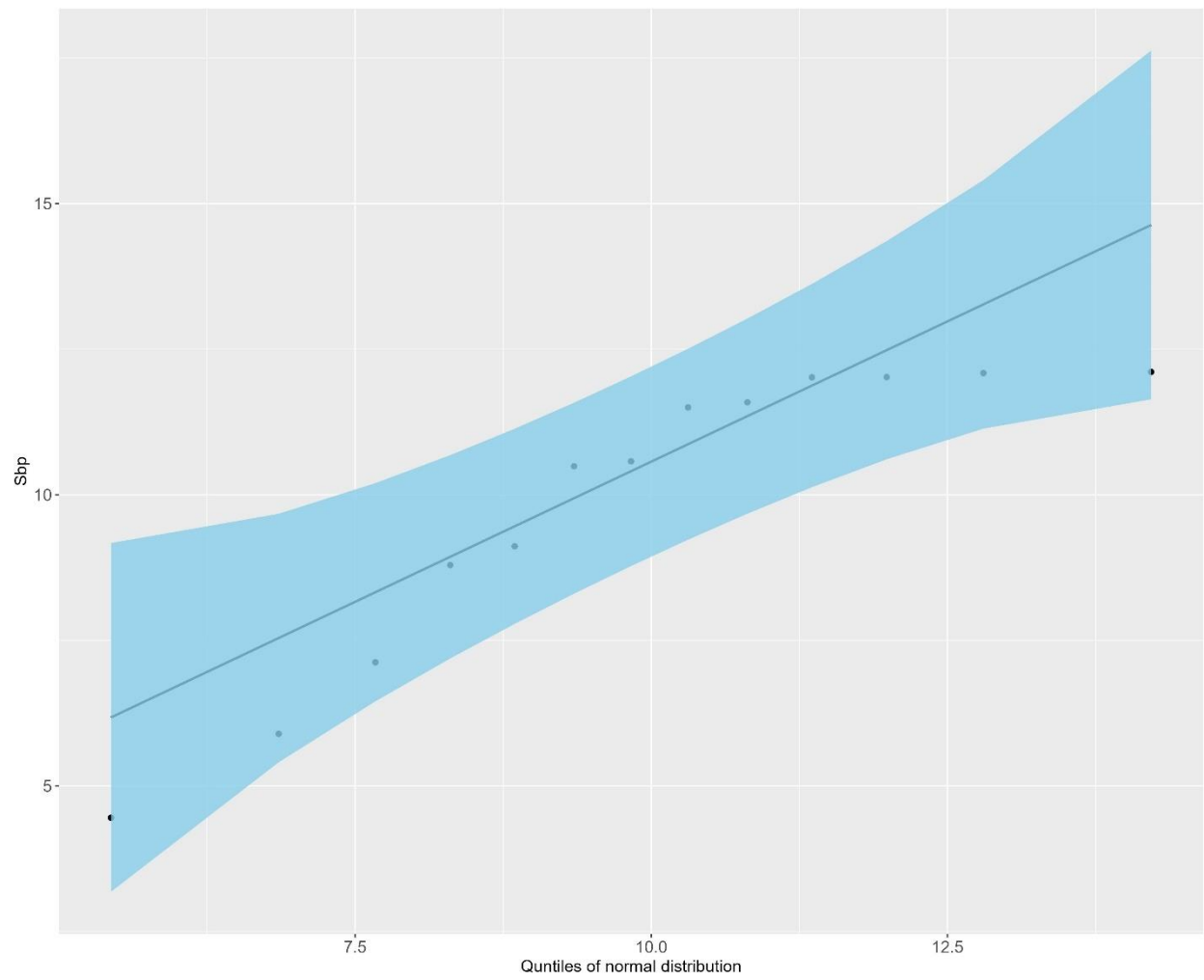
Побудуємо box plot щоб приблизно оцінити частоти всіх значень sbp:



Box plot sbp

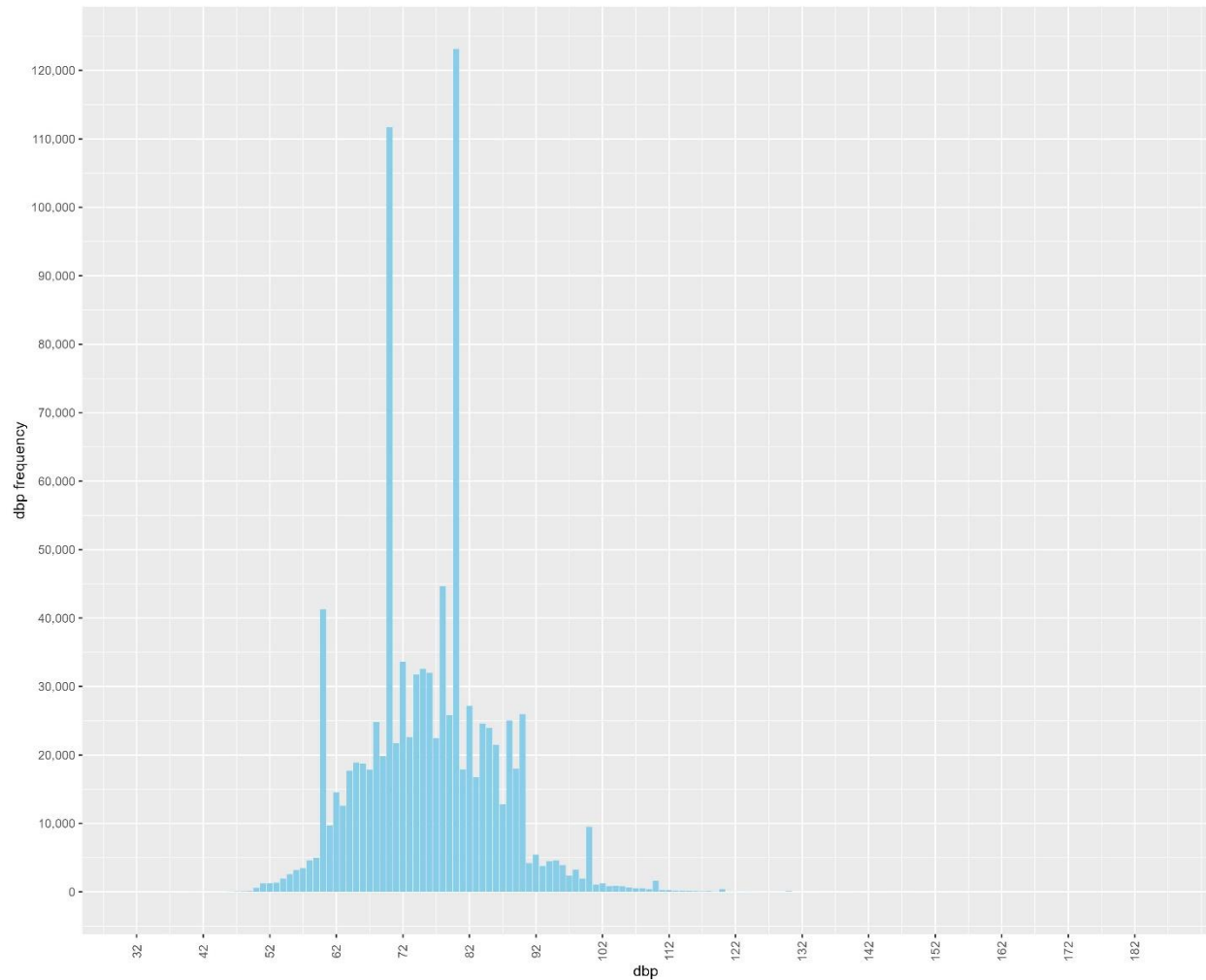
Box plot дозволяє оцінити частоти всіх значень SBP, в тому числі і аномально великих. Наявні викиди.

QQ plot також свідчить про «нормальний» характер розподілу:



Квантиль-квантильний прологарифмований графік частот sbr.

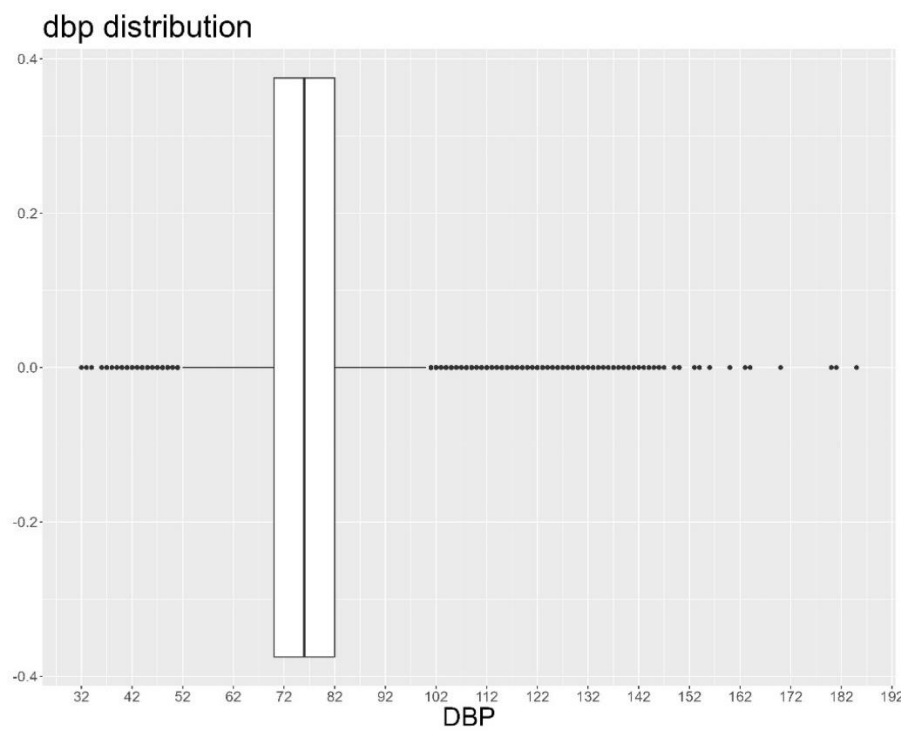
DBP (diastolic blood pressure, тиск в судинах між серцевими скороченнями)



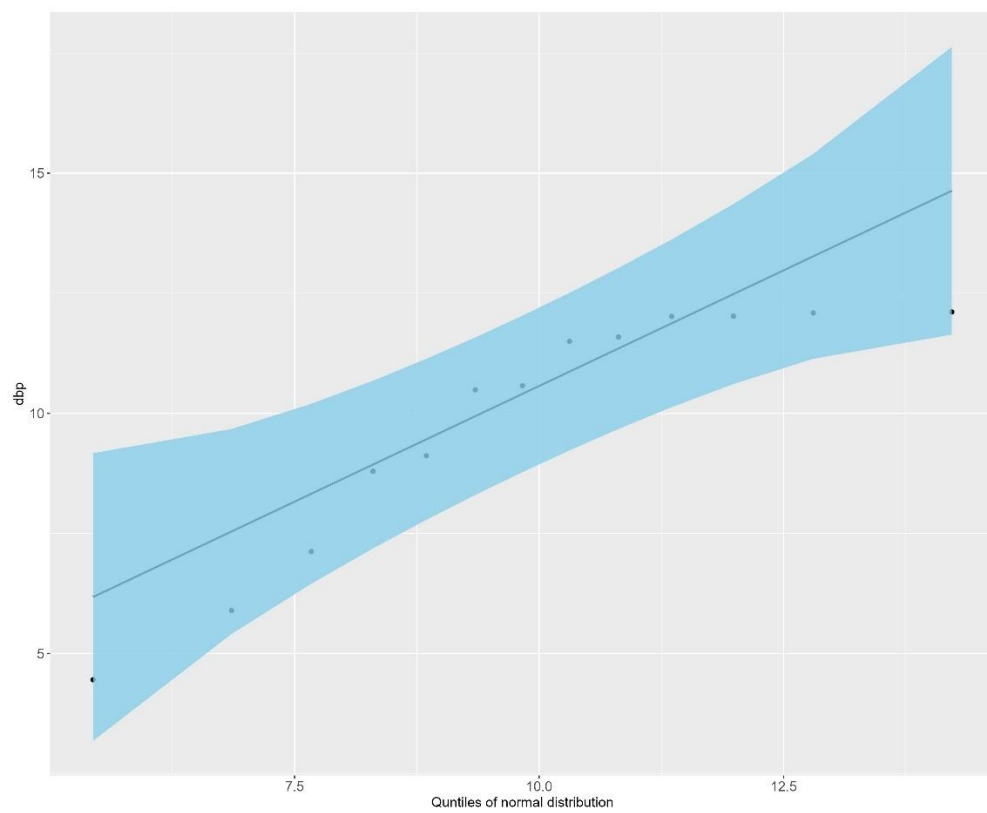
Діаграма частот dbp

Подібно до sbp, розподіл є віддалено схожим на нормальний, наявні 3 найтипівіших значень.

Box plot демонструє наявність викидів:

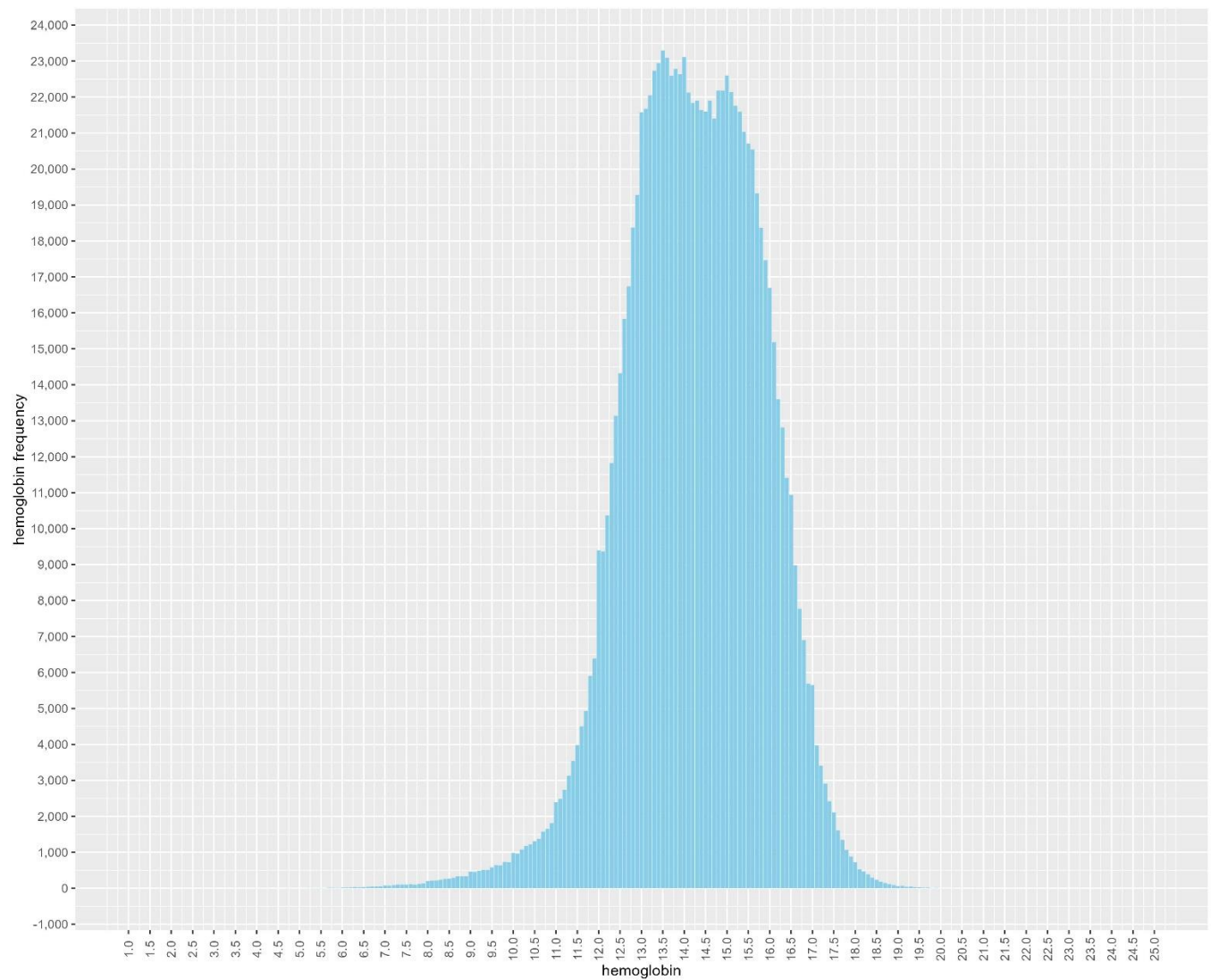


DBP Box plot

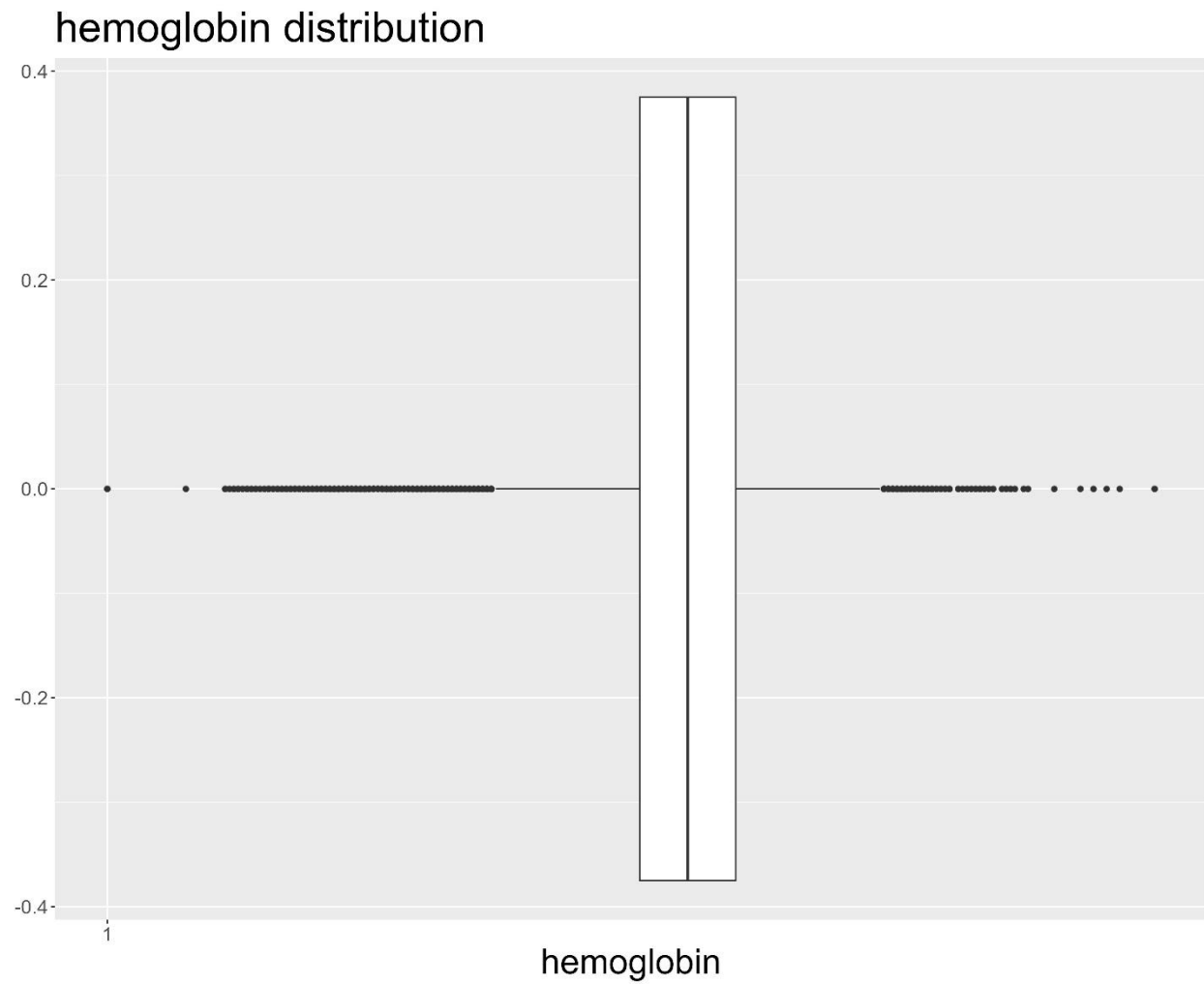


Квантиль-квантильний прологарифмований графік частот dbp.

Hemoglobin



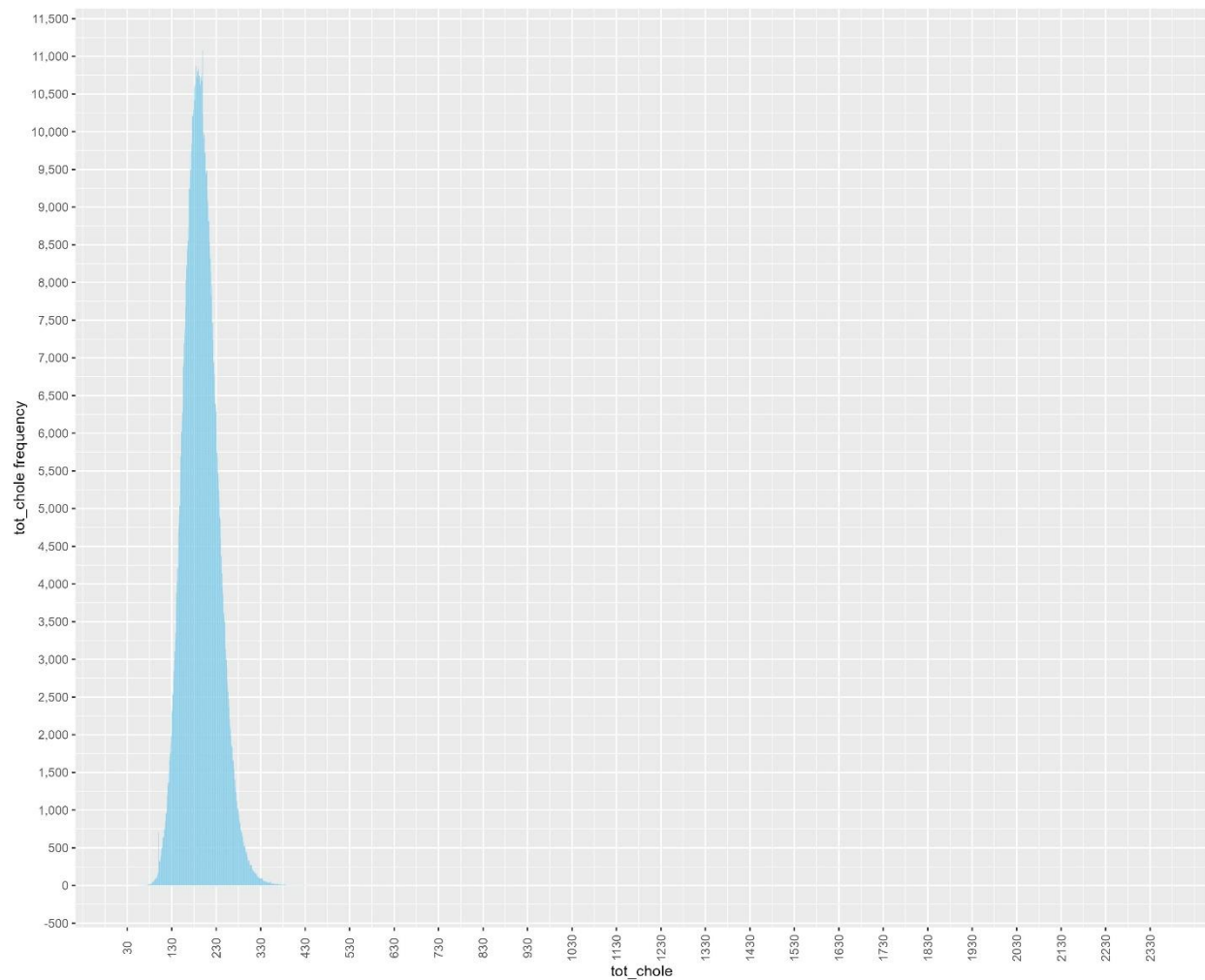
Діаграма частот hemoglobin



Hemoglobin Box plot

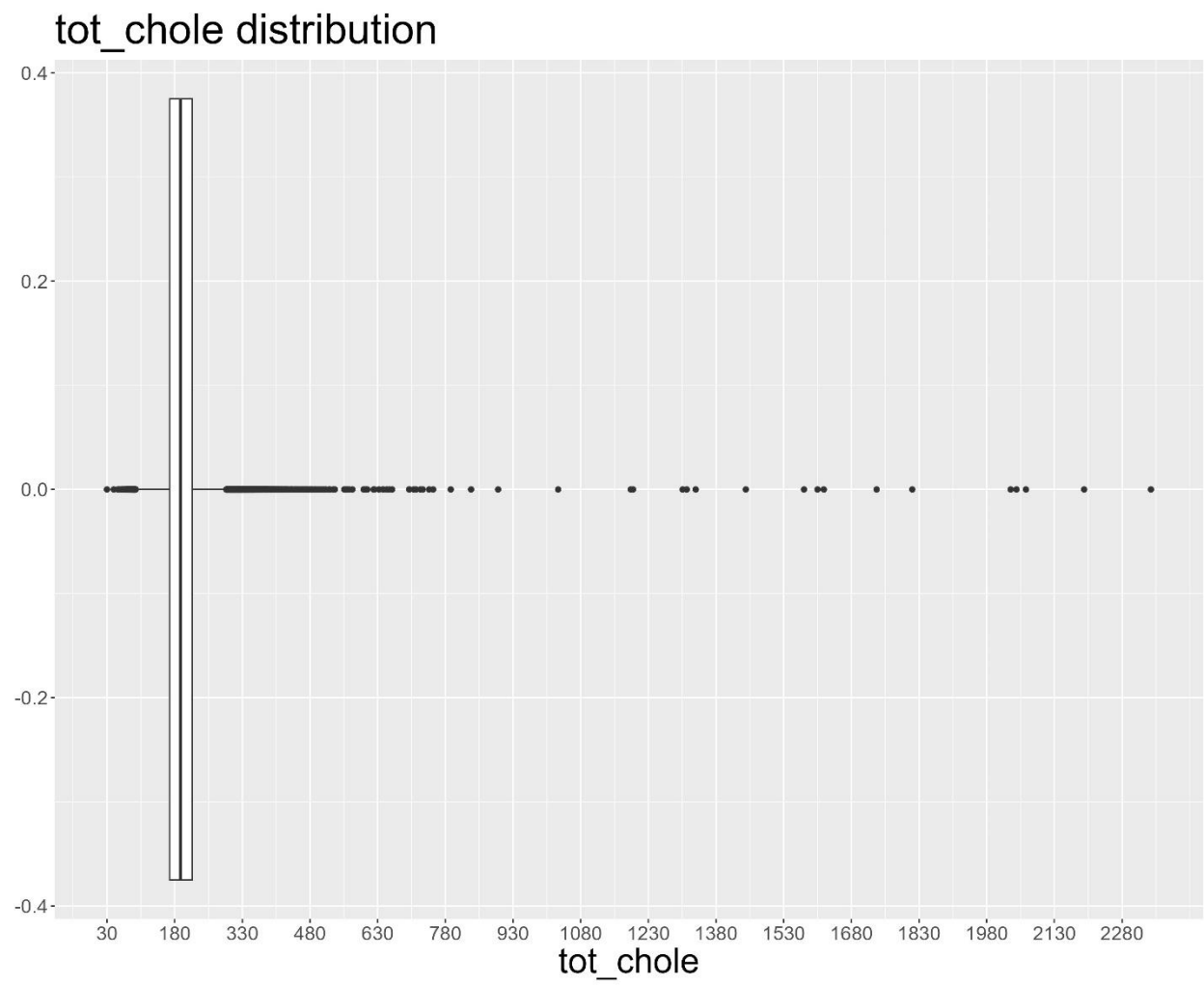
Розподіл значень гемоглобіну є близьким до бімодального. Про це свідчить типові 2 піки на діаграмі частот та 2 скупчення по різні сторони медіани на box plot.

Tot_chole (загальний холестерин)

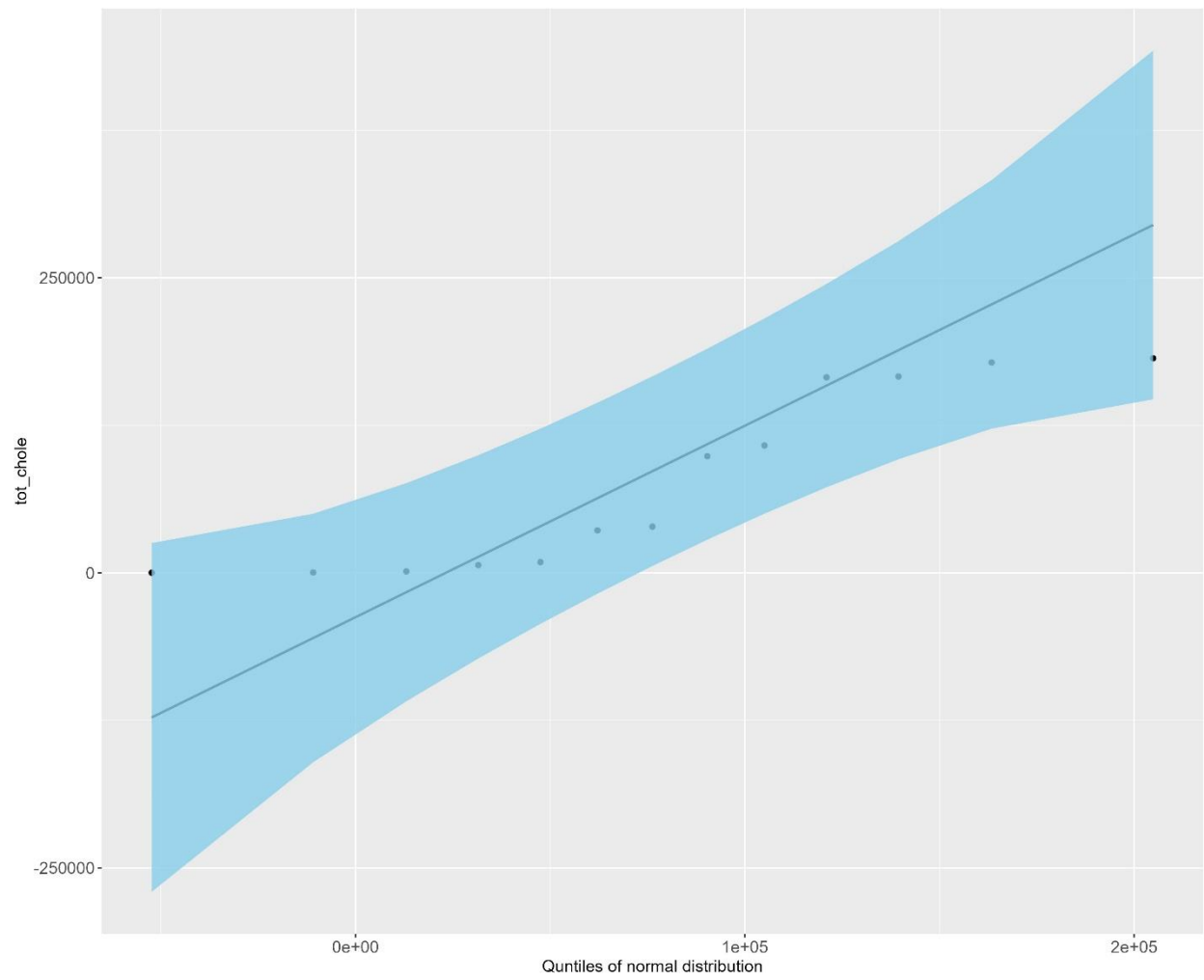


Діаграма частот холестерину

Спостерігаємо нормальний розподіл. На графіку не відображається частота значень в діапазоні (400;...). Побудуємо box plot:

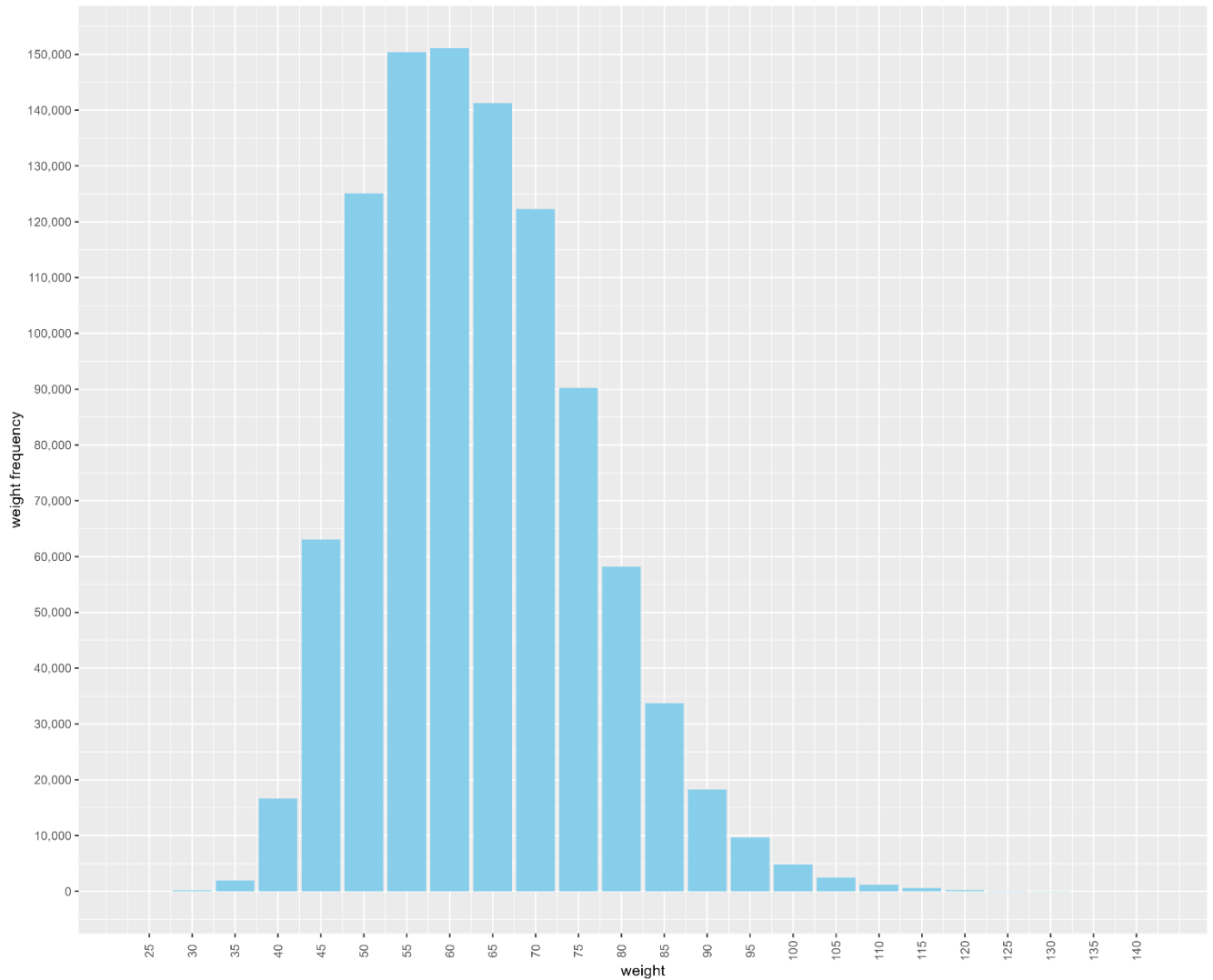


Дійсно, серед значень загального холестерину спостерігається багато викидів.



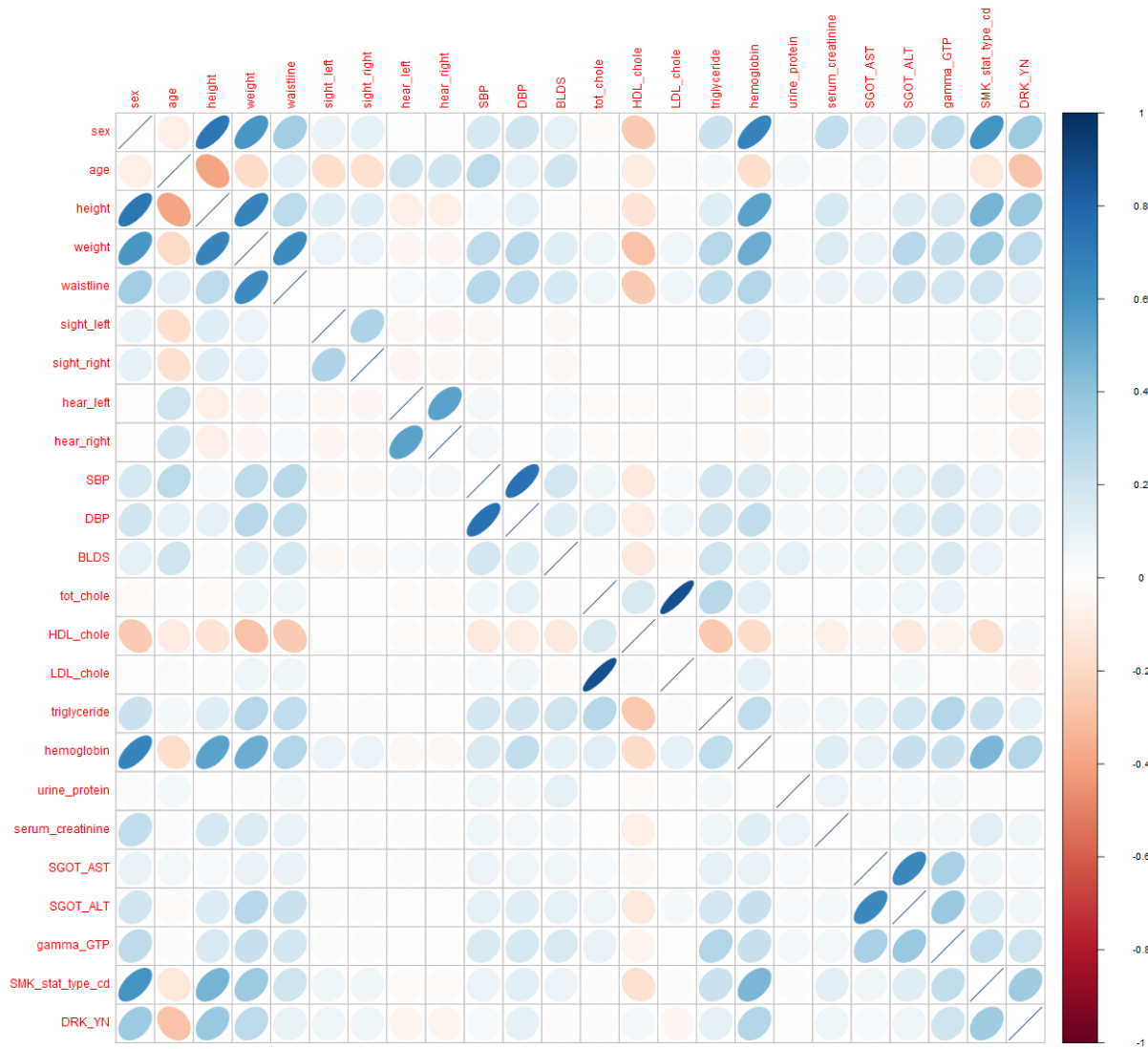
Квантиль-квантильний графік графік частот tot_chole

Weight



З першого погляду розподіл схожий на експоненційний. Якщо прибрати значення, частоти яких менші за 10000 (95кг і більше) – отримаємо нормальний розподіл.

Перевірка гіпотез

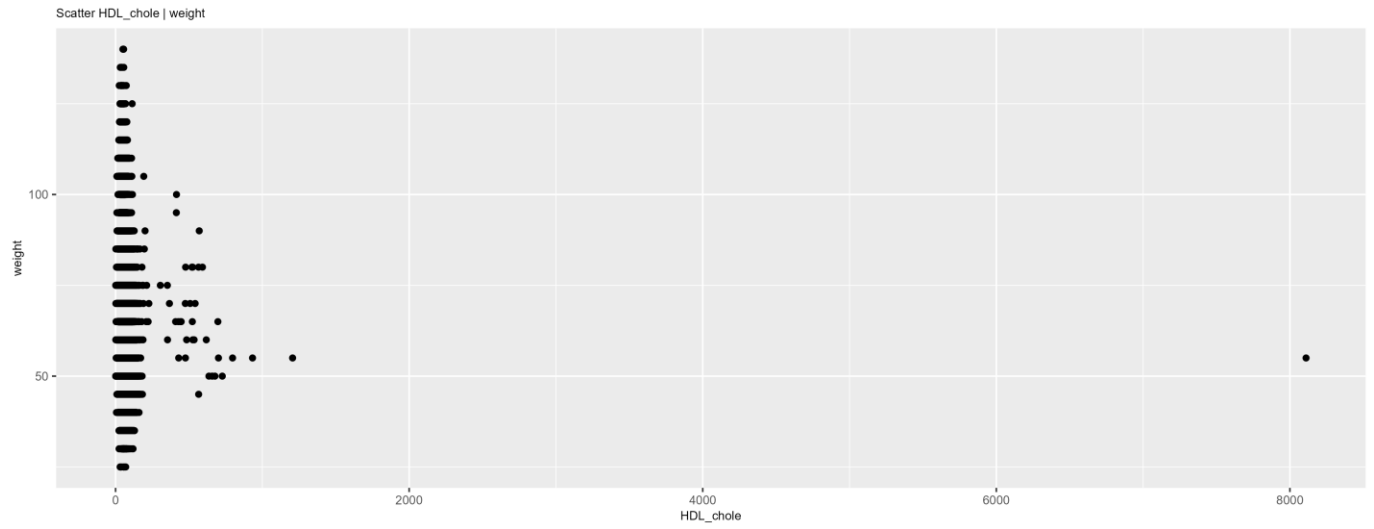


Матриця кореляцій для всіх змінних на всіх зразках датасету

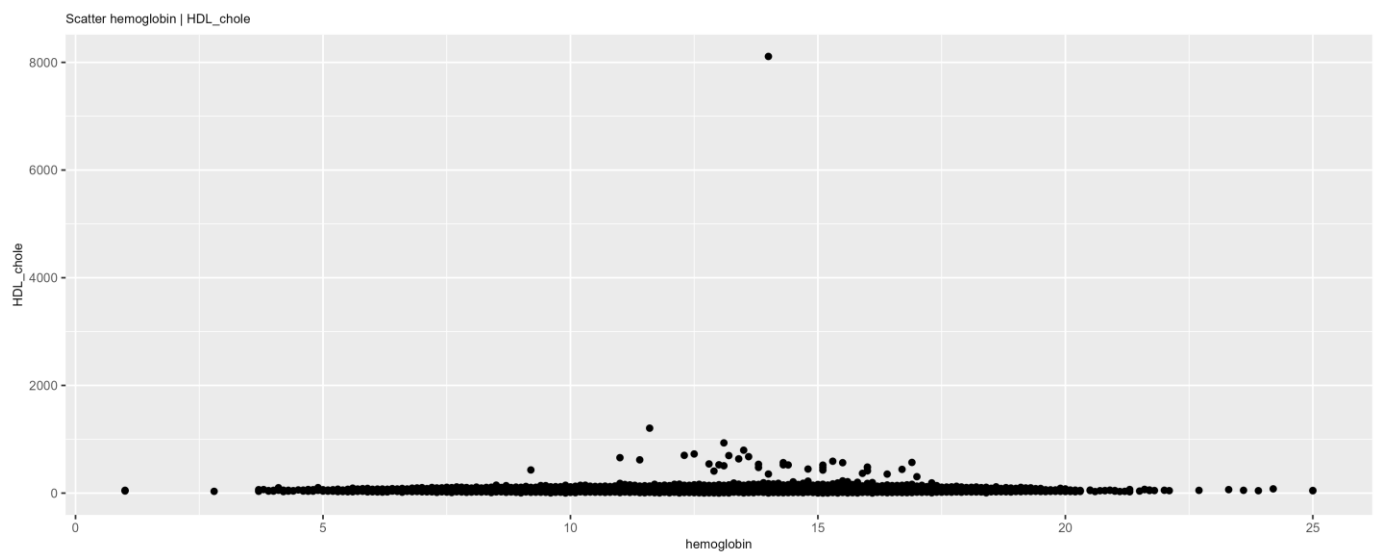
Маємо 4 значні кореляції:

1. Hemoglobin | weight
2. HDL_chole | weight
3. Triglyceride | HDL_chole
4. Hemoglobin | HDL_chole

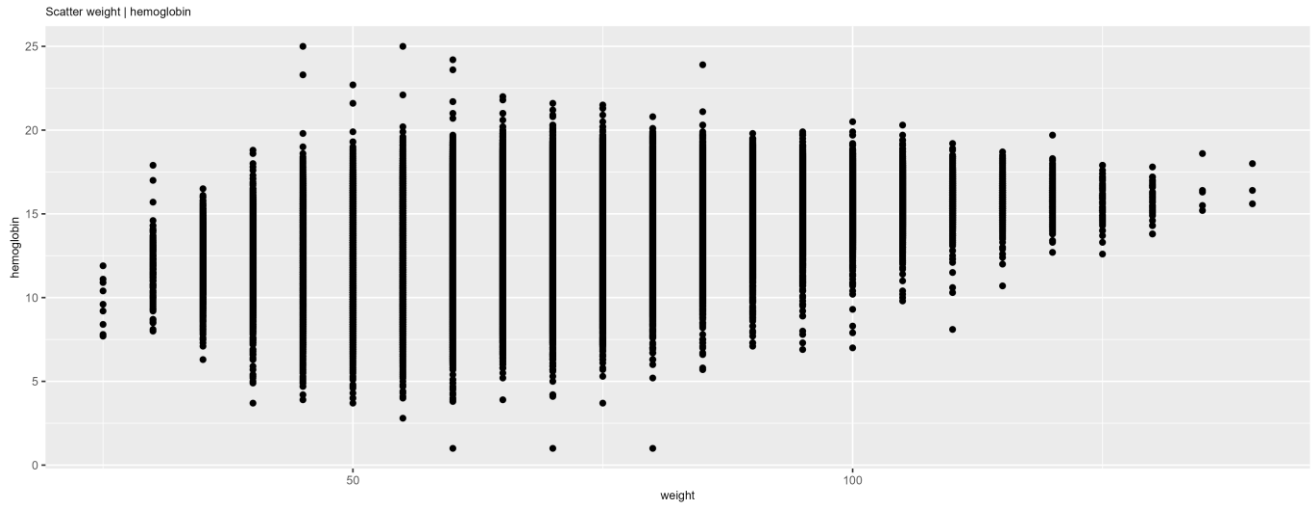
Побудуємо діаграми розсіювання для кожної кореляції:



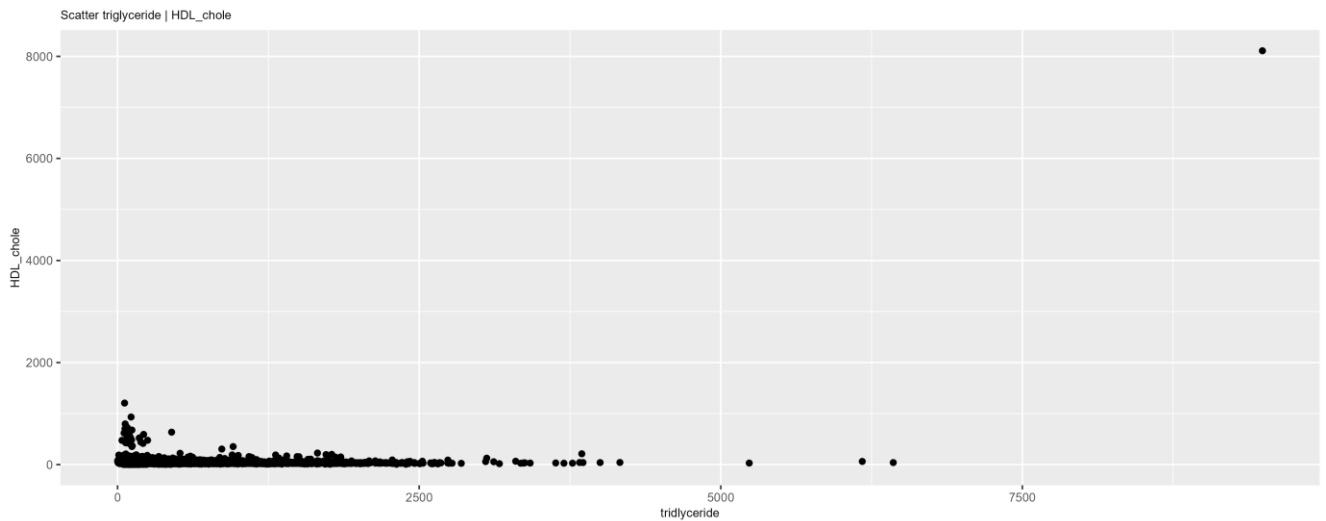
HDL_chole | weight



Hemoglobin | HDL_chole



Weight | hemoglobin



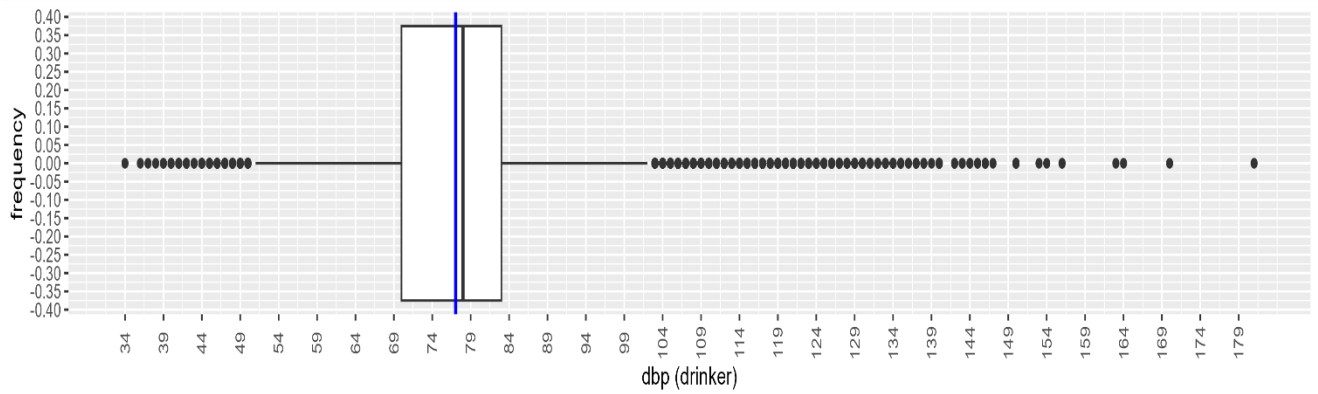
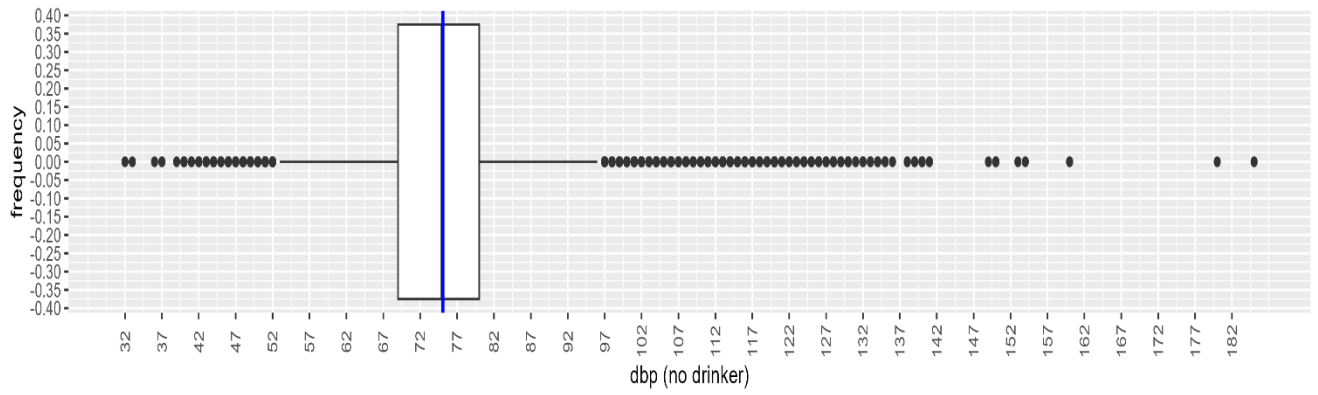
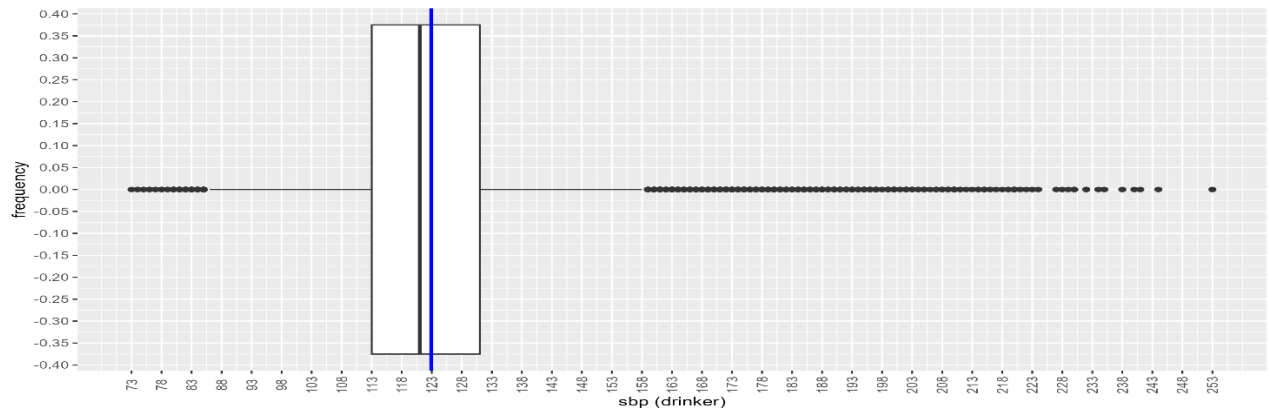
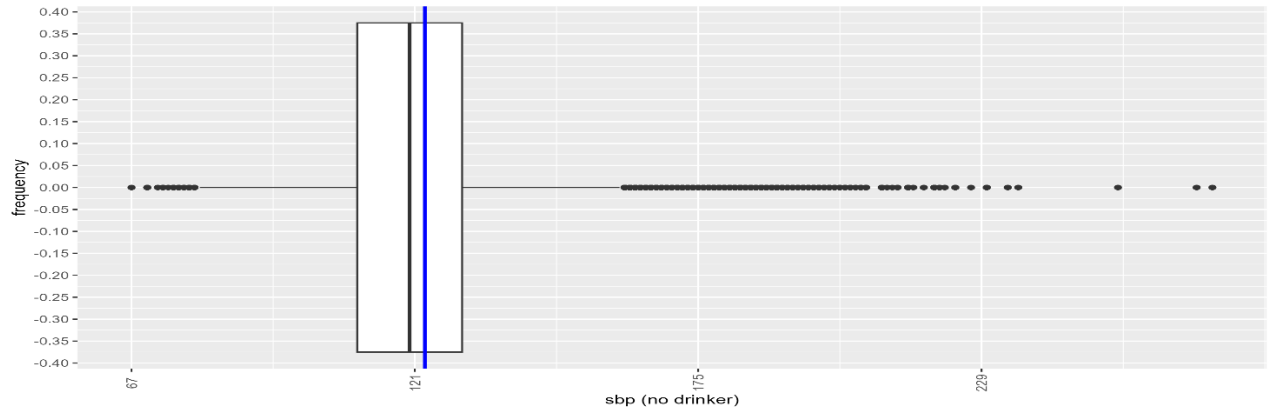
Triglyceride | HDL_chole

Гіпотеза 1: рівень гемоглобіну збільшується з вагою

Гіпотеза попередньо підтверджена наявністю кореляції між гемоглобіном та вагою.

Гіпотеза 2: вживання алкоголю впливає на SBP та DBP

Для перевірки цієї гіпотези порахуємо середні значення SBP та DBP окремо для людей, що вживають алкоголь, та для людей, що не вживають алкоголь (на графіках середнє позначено синім кольором):



"mean SBP (no drinker): 121.950590693303"

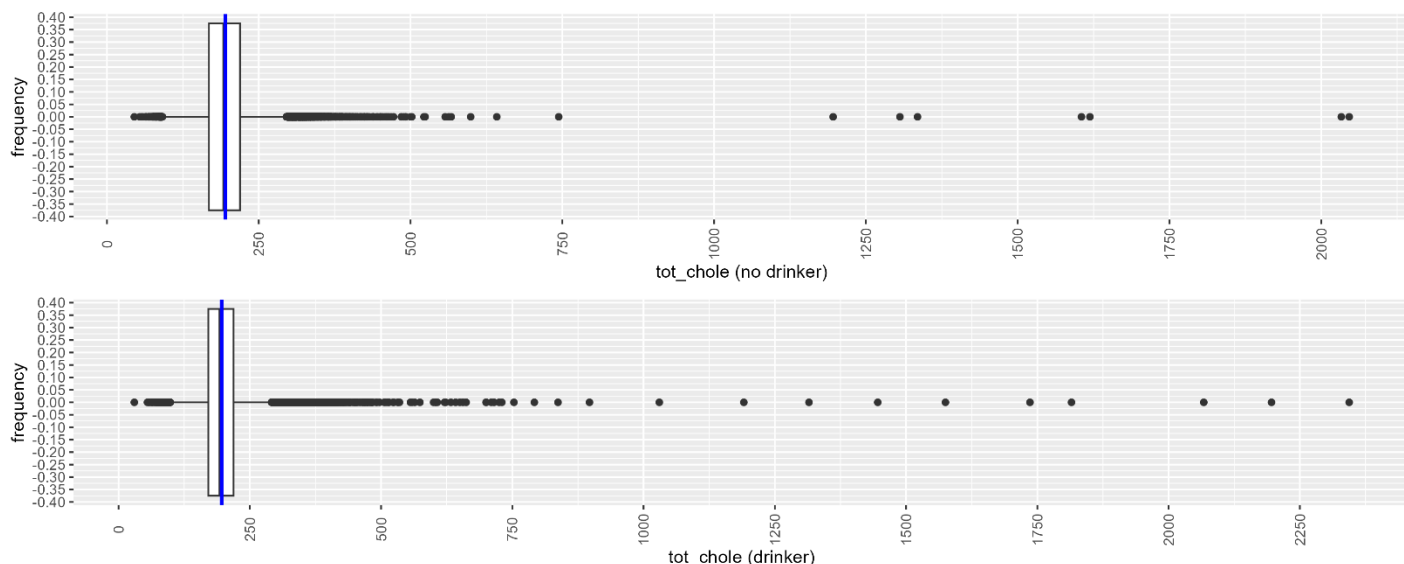
"mean SBP (drinker): 122.914764837897"

"mean DBP (no drinker): 75.0554029581049"

"mean DBP (drinker): 77.0505965835701"

Внаслідок вживання алкоголю помітно збільшилось середнє значення DBP. Є незначне зростання SBP. Попередньо гіпотеза підтверджена, подальше дослідження гіпотези потребує побудови довірчих інтервалів.

Гіпотеза 3: вживання алкоголю впливає на рівень загального холестерину



"mean tot_chole (no drinker): 194.794921126613"

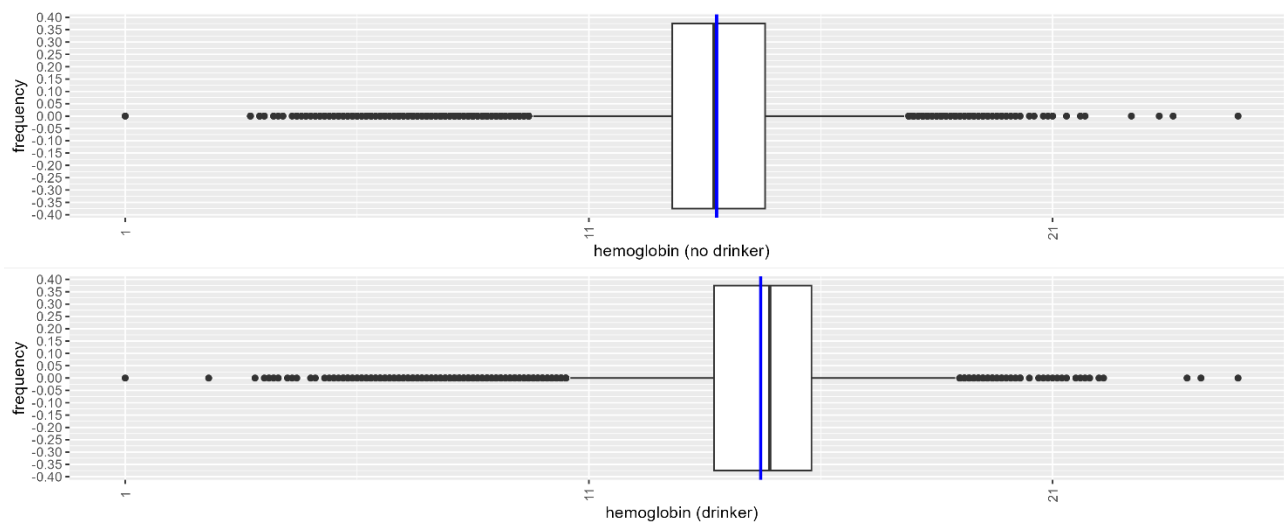
"mean tot_chole (drinker): 196.319688872384"

Зміна tot_chole не є суттєвою, гіпотеза попередньо не підтверджена.

Гіпотеза 4: Вживання алкоголю впливає на рівень гемоглобіну

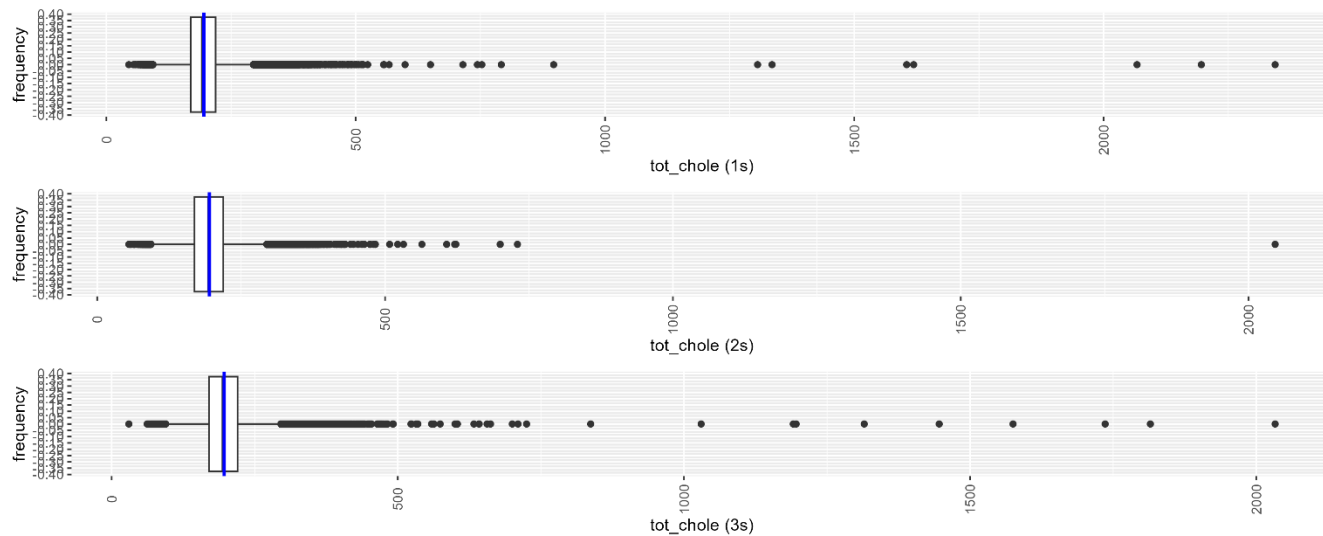
"mean hemoglobin (no drinker): 13.7555671180056"

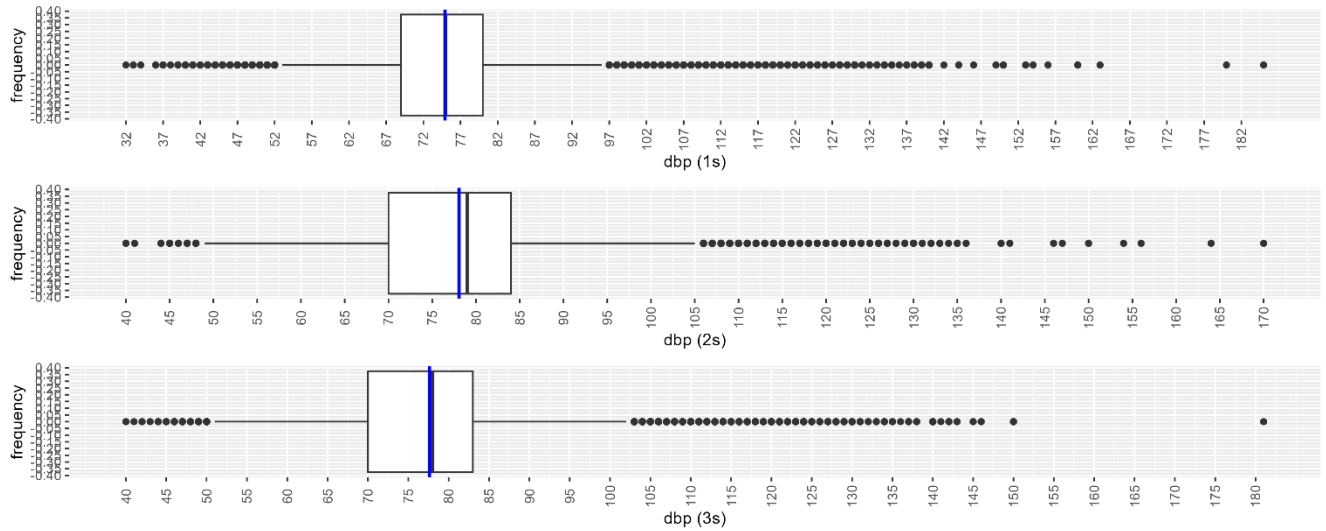
"mean hemoglobin (drinker): 14.7044344161715"



Різниця між середніми значеннями рівня гемоглобіну суттєва, гіпотеза попередньо підтверджена.

Гіпотеза 5: Значення обох видів тиску людей, що кинули палити (2), є вищим за значення груп 1 і 3





"mean sbp (1s): 121.177911197943"

"mean sbp (2s): 125.345010888763"

"mean sbp (3s): 123.583527300261"

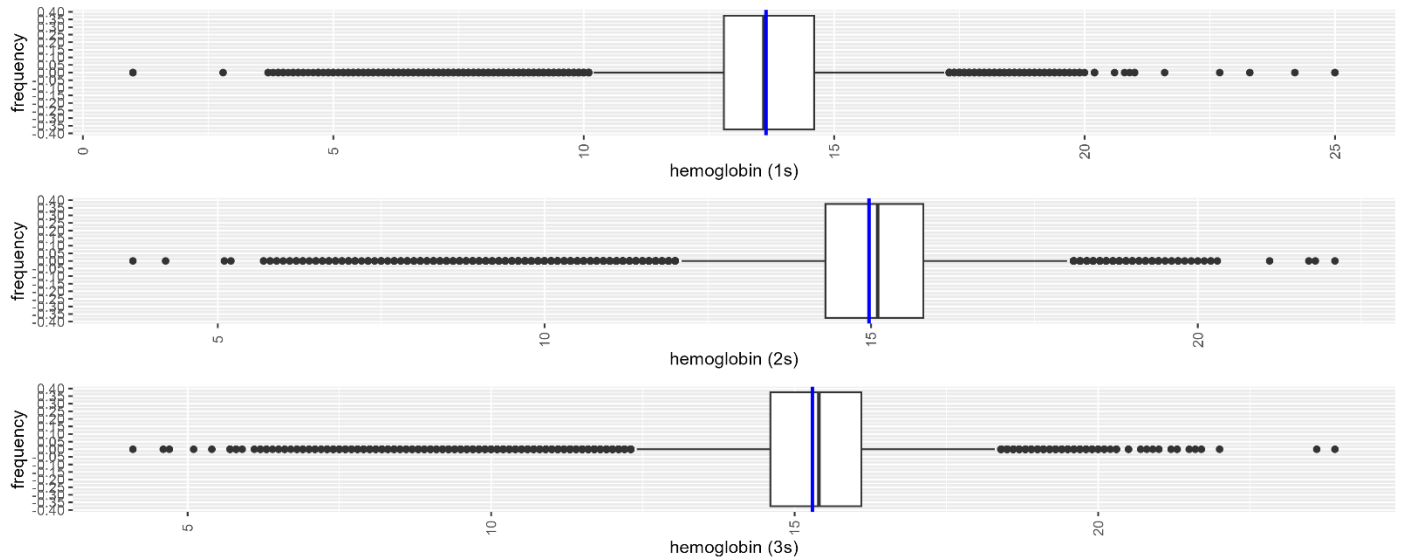
"mean dbp (1s): 74.9164067518645"

"mean dbp (2s): 78.0595881132431"

"mean dbp (3s): 77.6108415827701"

Середнє значення обох тисків для групи (2) є більшим за значення груп (1) і (3) – вважаємо гіпотезу попередньо підтвердженою.

Гіпотеза 6: Куріння збільшує рівень гемоглобіну, Після припинення паління з часом гемоглобін зменшується.



"mean hemoglobin (1s): 13.6387173183764"

"mean hemoglobin (2s): 14.9679693171231"

"mean hemoglobin (3s): 15.2906470549744"

Спостерігаємо менше середнє значення гемоглобіну у людей, що кинули палити, в порівнянні з середнім значенням людей, що палять ($\text{hemo2} < \text{hemo3}$). При цьому середній рівень гемоглобіну людей, що ніколи не палили, є меншим за рівень гемоглобіну тих осіб, хто кинув палити ($\text{hemo1} < \text{hemo2}$). Гіпотеза попередньо підтверджена.