**Опис датасету**

Першоджерело: <https://www.kaggle.com/datasets/sooyoungher/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.

1. **Вживання алкоголю впливає на SBP та DBP**

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

1. **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.

1. **Вживання алкоголю впливає на рівень гемоглобіну**

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

1. **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

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

Безпосередньо процес паління призводить до надходження чадного газу 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)

A computer code with text

Description automatically generated with medium confidence

**Огляд характеристик**

*Age (округлений з точністю до 5 років)*

A graph of a number of blue bars

Description automatically generated

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:

A graph showing a curve

Description automatically generated

Квантиль-квантильний прологарифмований графік частот age.

*Height (округлений з точністю до 5 см)*

Маємо чітко виражений нормальний розподіл:

A graph of a number of blue bars

Description automatically generated

Діаграма частот height

Побудуємо qq plot для height щоб додатково пересвідчитись:

A graph showing a curve

Description automatically generated

Квантиль-квантильний прологарифмований графік частот Height.

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

A graph showing a number of blue lines

Description automatically generated

Діаграма частот sbp

Розподіл є віддалено схожим на нормальний, наявні 5 найтиповіших значень. При цьому графік не відображає частоти для багатьох (sbp >= 207) значень sbp.

Побудуємо box plot щоб приблизно оцінити частоти всіх значень sbp:

A graph of a number of data

Description automatically generated with medium confidence

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:

A graph showing a curve

Description automatically generated

Квантиль-квантильний прологарифмований графік частот sbp.

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

A graph of a number of blue lines

Description automatically generated with medium confidence

Діаграма частот dbp

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

Box plot демонструє наявність викидів:

A graph of a distribution graph

Description automatically generated with medium confidence

DBP Box plot

A graph showing a curve

Description automatically generated

Квантиль-квантильний прологарифмований графік частот dbp.

*Hemoglobin*

A graph showing a blue line

Description automatically generated with medium confidence

Діаграма частот hemoglobin

A graph of a graph showing a number of lines

Description automatically generated with medium confidence

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)

A graph showing a tall tower

Description automatically generated

Діаграма частот холестерину

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

A graph of a number of numbers

Description automatically generated

Дійсно, серед значень загального холестерину спостерігається багато викидів.

A blue curve with black lines

Description automatically generated with medium confidence

Квантиль-квантильний графік графік частот tot\_chole

*Weight*

A graph of a graph

Description automatically generated with medium confidence

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

**Перевірка гіпотез**

A grid with dots and lines

Description automatically generated

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:

A graph of a sound wave

Description automatically generated

HDL\_chole | weight

A graph with a black oval

Description automatically generated

Hemoglobin | HDL\_chole

A graph with lines in the center

Description automatically generated with medium confidence

Weight | hemoglobin

A graph of a graph showing a number of dots

Description automatically generated with medium confidence

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):

A screenshot of a graph

Description automatically generated

A graph with lines and numbers

Description automatically generated with medium confidence

"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**

A screenshot of a graph

Description automatically generated

"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"

A screenshot of a graph

Description automatically generated

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**

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

"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.**

A screenshot of a computer

Description automatically generated

"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 (hemo2 < hemo3).

При цьому середній рівень гемоглобіну людей, що ніколи не палили, є меншим за рівень гемоглобіну тих осіб, хто кинув палити (hemo1 < hemo2). Гіпотеза попередньо підтверджена.