# Лабораторна робота №1

## Первинний аналіз даних* з Pandas

*дані про серцево-судинні захворювання

---

## Завдання

Необхідно дати відповіді (з написанням коду) на запитання щодо набору даних про серцево-судинні захворювання. Дані збережені у файлі *../data/bootcamp5.csv*.

### Проблема

Прогнозування наявності або відсутності серцево-судинних захворювань (ССЗ), використовуючи результати обстеження пацієнта.

### Опис даних

Набір даних сформований на основі реальної інформації про серцево-судинні захворювання пацієнтів і містить ознаки, що можна розбити на 3 групи:

- *Об'єктивні*: фактична інформація;
- *Обстеження*: результати медичного огляду;
- *Суб'єктивні*: інформація, надана пацієнтом.

| Ознака | Група | Назва змінної | Тип значення |
|---|---|---|---|
| Вік | Об'єктивні | age | int (дні) |
| Зріст | Об'єктивні | height | int (см) |
| Вага | Об'єктивні | weight | float (кг) |
| Стать | Об'єктивні | gender | категоріальний код |
| Верхній артеріальний тиск | Обстеження | ap_hi | int |
| Нижній артеріальний тиск | Обстеження | ap_lo | int |
| Холестерин | Обстеження | cholesterol | 1: норма, 2: вище норми, 3: значно вище норми |
| Глюкоза | Обстеження | gluc | 1: норма, 2: вище норми, 3: значно вище норми |

| Ознака | Група | Назва змінної | Тип значення |
|---|---|---|---|
| Куріння | Суб'єктивні | smoke | binary |
| Вживання алкоголю | Суб'єктивні | alco | binary |
| Фізична активність | Суб'єктивні | active | binary |

Цільова ознака (яку цікаво буде прогнозувати): наявність серцево-судинних захворювань за результатами класичного лікарського огляду (**cardio**).

Всі показники отримані на момент огляду.

# Виконання завдання

```
# Імпортуємо необхідні модулі
import pandas as pd
import numpy as np
```

Зчитуємо дані з файлу в память у вигляді об'єкта Pandas.DataFrame

```
df = pd.read_csv('../data/bootcamp5.csv', sep=';')
print('Розмір набору даних: ', df.shape)
df.head()
```

```
Розмір набору даних:  (70000, 13)
```

```
   id    age  gender  height  weight  ap_hi  ap_lo  cholesterol  gluc
smoke  \
0   0  18393       1     168    62.0    110     80            1     1
0
1   1  20228       2     156    85.0    140     90            3     1
0
2   2  18857       2     165    64.0    130     70            3     1
0
3   3  17623       1     169    82.0    150    100            1     1
0
4   4  17474       2     156    56.0    100     60            1     1
0

   alco  active  cardio
0     0       1       0
1     0       1       1
2     0       0       1
3     0       1       1
4     0       0       0
```

Тепер давайте обчислимо деяку статистику для унікальних значень ознак:

```python
for c in df.columns:
    n = df[c].nunique()
    print(c)
    if n <= 3:
        print(n, sorted(df[c].value_counts().to_dict().items()))
    else:
        print(n)
    print(10 * '-')
```

```
id
70000
----------
age
8076
----------
gender
2 [(1, 24470), (2, 45530)]
----------
height
109
----------
weight
287
----------
ap_hi
153
----------
ap_lo
157
----------
cholesterol
3 [(1, 52385), (2, 9549), (3, 8066)]
----------
gluc
3 [(1, 59479), (2, 5190), (3, 5331)]
----------
smoke
2 [(0, 63831), (1, 6169)]
----------
alco
2 [(0, 66236), (1, 3764)]
----------
active
2 [(0, 13739), (1, 56261)]
----------
cardio
2 [(0, 35021), (1, 34979)]
----------
```

Яких типів ознаки, що описують пацієнтів?

## 1. Основні спостереження

**Запитання 1. Скільки чоловіків і жінок представлено в цьому наборі даних? Спосіб кодування для ознаки "Стать" невідомий (1 в `gender` відповідає чоловіку чи жінці?). З'ясувати це можна, проаналізувавши зріст, і, зробивши припущення, що чоловіки в середньому вищі.**

1. 45530 жінок та 24470 чоловіків
2. 45530 чоловіків та 24470 жінок
3. 45470 жінок та 24530 чоловіків
4. 45470 чоловіків та 24530 жінок

```
# Ваш код тут
df_task = df[['gender','height']]
df_task.groupby('gender').mean()
# Гендер 1 по середньому значенню росту, більш схожий на чоловічу
стать, тому відповідь - 45530 жінок та 24470 чоловіків
```

```
            height
gender
1       169.947895
2       161.355612
```

**Запитання 2. Хто в середньому рідше вказує, що вживає алкоголь — чоловіки чи жінки?**

1. жінки
2. чоловіки

```
# Ваш код тут
df_task = df[["gender","alco"]]
df_task.groupby("gender").mean() * 100
# Відповідь чоловіки, вони вказують, що вживають алкоголь в близько
10% випадків
```

```
             alco
gender
1       10.637515
2        2.549967
```

**Запитання 3. У скільки разів (округлити, `round()`) відсоток курців серед чоловіків більший, ніж відсоток курців серед жінок (за цими анкетними даними)?**

1. 4
2. 12
3. 16
4. 20

```
# Ваш код тут
df_task = df[["gender","smoke"]]
a = df_task.groupby("gender").mean() * 100
```

```
male = a.loc[1, "smoke"]
female = a.loc[2, "smoke"]
round(male / female)
# Відповідь 12
```

```
12
```

**Запитання 4. Яка різниця між значеннями медіан (`median()`) віку тих хто не курить і курців (в місяцях, округлити)?**

1. 8
2. 10
3. 16
4. 20

```
# Ваш код тут
df_task = df[["age","smoke"]]
a = df_task.groupby("smoke").median()
round((a.loc[0, "age"] - a.loc[1, "age"]) / 30)
# Відповідь - 20 місяців
```

```
20
```

## 2. Карти ризиків

## Завдання:

На веб-сайті Європейського товариства кардіологів розміщена шкала SCORE. Вона використовується для розрахунку ризику смерті від серцево-судинного захворювання в найближчі 10 років і виглядає наступним чином:

SCORE – це абревіатура англійських слів «систематична оцінка коронарного ризику», тобто ризику захворювань серця і судин. Ця шкала була запропонована групою експертів Європейського товариства кардіологів у 2003 р. і розроблена на підставі результатів досліджень, проведених в 12 європейських країнах із загальною кількістю пацієнтів понад 205 тисяч.

Шкала – це система квадратів, у якій застосовано принцип світлофора – три основні кольори:

- зелений – низький ризик, що відповідає 1% або менше;
- жовтий – увага! – ризик помірний і коливається у межах від 2 до 4%;
- червоний – небезпека! – 5% і більше.

Для більшої диференціації використані відповідні відтінки цих трьох основних кольорів.

Давайте подивимось на верхній правий прямокутник, на якому відображено підмножину чоловіків, що палять, віком від 60 до 64 років включно. (Неочевидно, але значення на рисунку для віку та тиску означають верхню межу, і вона не включається).

Бачимо значення 9 у лівому нижньому куті прямокутника та 47 у правому верхньому. Це означає, що для чоловіків-курців цієї вікової категорії, у яких систолічний (верхній) артеріальний тиск менший за 120 мм рт.ст., а рівень холестерину – 4 ммоль/л, ризик ССЗ оцінюється приблизно в 5 разів нижче, ніж якби значення тиску знаходилось в інтервалі [160, 180), а холестерину було 6 8 ммоль/л.

Розрахуємо це співвідношення, використовуючи наші дані.

Роз'яснення:

- Створіть нову ознаку `age_years` — вік в роках, заокруглений до цілого. Для цього завдання відберіть лише чоловіків, що палять, віком від 60 до 64 років включно.
- Категорії рівня холестерину на рисунку і в наших даних відрізняються. Перетворення значень на рисунку в значення ознаки `cholesterol` наступне: 4 ммоль/л → 1, 5-7 ммоль/л → 2, 8 ммоль/л → 3.
- Цікавлять 2 підвибірки з відібраних чоловіків: перша з верхнім артеріальним тиском строго меншим за 120 мм рт.ст. і концентрацією холестерину – 4 ммоль/л, а друга – з верхнім артеріальним тиском від 160 (включно) до 180 мм рт.ст. (не включно) і концентрацією холестерину – 8 ммоль/л.

```
# Ваш код тут
df["age_years"] = round(df["age"] / 365)
df_task = df[(df["gender"] == 1) & (df["age_years"] >= 60) &
(df["age_years"] <= 64) & (df["smoke"] == 1)]
low_group = df_task[(df_task['cholesterol'] == 1) & (df_task['ap_hi']
< 120)]
high_group = df_task[(df_task['cholesterol'] == 3) & (df_task['ap_hi']
>= 160) & (df_task['ap_hi'] < 180)]
round(low_group.shape[0] / high_group.shape[0])
# Відповідь - 4 ???? Мабуть...
```

```
4
```

**Запитання 5. Обчисліть частки людей із ССЗ в двох описаних вище підвибірках. Яке відношення цих часток (округлити)?**

1. 1
2. 2
3. 3
4. 4

## 3. Аналіз BMI

## Завдання:

Створіть нову ознаку – BMI (Body Mass Index, Індекс маси тіла). Для цього треба вагу у кілограмах поділити на квадрат зросту в метрах. Вважається, що нормальні значення BMI в діапазоні від 18.5 до 25.

```python
# Ваш код тут
df["bmi"] = df["weight"] / ((df["height"] / 100) * (df["height"] / 100))
df.describe()
# Відповідь 1, тут медіана (50%) = 26.37, коли максимальна межа норми 25
```

```
                  id           age        gender        height        weight  \
count  70000.000000  70000.000000  70000.000000  70000.000000  70000.000000
mean   49972.419900  19468.865814      1.650429    164.359229     74.205690
std    28851.302323   2467.251667      0.476838      8.210126     14.395757
min        0.000000  10798.000000      1.000000     55.000000     10.000000
25%    25006.750000  17664.000000      1.000000    159.000000     65.000000
50%    50001.500000  19703.000000      2.000000    165.000000     72.000000
75%    74889.250000  21327.000000      2.000000    170.000000     82.000000
max    99999.000000  23713.000000      2.000000    250.000000    200.000000

               ap_hi         ap_lo   cholesterol          gluc         smoke  \
count  70000.000000  70000.000000  70000.000000  70000.000000  70000.000000
mean     128.817286     96.630414      1.366871      1.226457      0.088129
std      154.011419    188.472530      0.680250      0.572270      0.283484
min     -150.000000    -70.000000      1.000000      1.000000      0.000000
25%      120.000000     80.000000      1.000000      1.000000      0.000000
50%      120.000000     80.000000      1.000000      1.000000      0.000000
75%      140.000000     90.000000      2.000000      1.000000      0.000000
max    16020.000000  11000.000000      3.000000      3.000000      1.000000

                alco        active        cardio     age_years           bmi
count  70000.000000  70000.000000  70000.000000  70000.000000  70000.000000
mean       0.053771      0.803729      0.499700     53.338686
```

```
27.556513
std         0.225568      0.397179      0.500003      6.765294
6.091511
min         0.000000      0.000000      0.000000     30.000000
3.471784
25%         0.000000      1.000000      0.000000     48.000000
23.875115
50%         0.000000      1.000000      0.000000     54.000000
26.374068
75%         0.000000      1.000000      1.000000     58.000000
30.222222
max         1.000000      1.000000      1.000000     65.000000
298.666667
```

**Запитання 6. Виберіть правильні твердження:**

1. Медіана BMI перевищує норму.
2. BMI для жінок в середньому нижчий, ніж для чоловіків.
3. У здорових людей в середньому BMI вищий, ніж у людей із ССЗ.
4. Для здорових чоловіків, що не вживають алкоголь в середньому BMI ближче до норми, ніж для здорових жінок, що не вживають алкоголь.

## 4. Очищення даних

## Завдання:

Можна помітити, що дані не є досконалими. В них багато «бруду» і неточностей. Ще краще це видно на візуалізації даних.

Видаліть наступних пацієнтів (вважаємо це помилками в даних):

- вказане нижнє значення артеріального тиску більше верхнього;
- зріст менший за 2.5-й процентиль (Для обчислення цього значення використовуйте `pd.Series.quantile`. Якщо ви не знайомі з функцією, будь ласка, прочитайте документацію.);
- зріст більший за 97.5-й процентиль;
- вага менша за 2.5-й процентиль;
- вага більша за 97.5-й процентиль.

Це не все, що можна було зробити для очищення даних, але поки зупинимося на цьому.

```
# Ваш код тут
height_lq = df['height'].quantile(0.025)
height_uq = df['height'].quantile(0.975)
weight_lq = df['weight'].quantile(0.025)
weight_uq = df['weight'].quantile(0.975)
brand_new_df = df[(df["ap_hi"] > df["ap_lo"]) & (df["height"] >
height_lq) & (df["height"] < height_uq) & (df["weight"] > weight_lq) &
(df["weight"] < weight_uq)]
```

```
round((1 - (brand_new_df.shape[0] / df.shape[0])) * 100)
# Відповідь - 11%
```

```
11
```

**Запитання 7. Скільки відсотків вихідних даних (округлити) ми викинули?**

1. 8
2. 9
3. 10
4. 11

## 5. Різне

**Завдання 8. Відбиріть дані за заданим критерієм та відсортуйте згідно з варіантом. Яку кількість записів було відібрано та який це складає відсоток від загальної кількості?**

| Варіант | Критерій та сортування |
|---|---|
| 1 | Зріст не перевищує 160, за зростанням ваги |
| 2 | Вага менша за 50, за спаданням нижнього тиску |
| 3 | Верхній тиск не менший за 120, за зростанням зросту |
| 4 | Нижній тиск більший за 90, за спаданням ваги |
| 5 | Зріст більший за 180, за зростанням верхнього тиску |
| 6 | Вага не менша за 80, за спаданням зросту |
| 7 | Верхній тиск не перевищує 140, за зростанням ваги |
| 8 | Нижній тиск не менший за 70, за спаданням зросту |
| 9 | Зріст менший за 170, за зростанням нижнього тиску |
| 10 | Вага не перевищує 70, за спаданням верхнього тиску |

```
# Ваш код тут
brand_new_df[(brand_new_df["height"] < 170) & (brand_new_df["ap_lo"] >
40)].sort_values(by="ap_lo")
```

```
          id    age  gender  height  weight  ap_hi  ap_lo  cholesterol
gluc  \
51158  72974  17371       2     160    68.0    160     45            1
1
65010  92788  15611       2     155    60.0     95     49            1
1
67600  96520  21993       2     153    77.0    115     49            1
1
56322  80374  19783       2     157    95.0    100     50            1
1
59044  84283  14608       2     160    58.0     80     50            2
1
...      ...    ...     ...     ...     ...    ...    ...          ...
...
```

```
19732  28172  23332       2     165    78.0    170    150            3
1
54987  78442  19491       2     169    80.0    200    160            3
3
4781    6769  18961       2     158    74.0    200    170            1
1
43998  62861  22652       1     163    70.0    200    180            1
1
38022  54282  21770       2     161    84.0    196    182            2
2
```

| | smoke | alco | active | cardio | age_years | bmi |
|---|---|---|---|---|---|---|
| 51158 | 0 | 0 | 1 | 1 | 48.0 | 26.562500 |
| 65010 | 0 | 0 | 1 | 0 | 43.0 | 24.973985 |
| 67600 | 0 | 0 | 1 | 1 | 60.0 | 32.893332 |
| 56322 | 0 | 0 | 0 | 0 | 54.0 | 38.541117 |
| 59044 | 0 | 0 | 1 | 0 | 40.0 | 22.656250 |
| ... | ... | ... | ... | ... | ... | ... |
| 19732 | 0 | 0 | 1 | 1 | 64.0 | 28.650138 |
| 54987 | 0 | 0 | 1 | 0 | 53.0 | 28.010224 |
| 4781 | 0 | 0 | 1 | 1 | 52.0 | 29.642685 |
| 43998 | 0 | 0 | 0 | 1 | 62.0 | 26.346494 |
| 38022 | 0 | 0 | 1 | 1 | 60.0 | 32.406157 |

```
[41941 rows x 15 columns]
```

**Завдання 9. Створіть бінарну ознаку згідно з критерієм з завдання 8 та побудуйте таблицю спряженості для неї та цільової ознаки.**

```
# Ваш код тут
brand_new_df["8_task"] = brand_new_df["height"] < 170
pd.crosstab(df["8_task"], df["cholesterol"])
```

```
C:\Users\Danik\AppData\Local\Temp\ipykernel_13352\211005490.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  brand_new_df["8_task"] = brand_new_df["height"] < 170
```

| cholesterol | 1 | 2 | 3 |
|---|---|---|---|
| 8_task | | | |
| False | 9656 | 1493 | 1117 |
| True | 31584 | 5533 | 4860 |

**Завдання 10. Побудуйте зведену таблицю обчислення статистики для заданих ознак згідно з варіантом у розрізі цільової ознаки та створеної у завданні 9.**

| Варіант | Статистика та ознаки |
|---|---|
| 1 | Медіана для нижнього та верхнього тисків |
| 2 | Середнє значення для нижнього та верхнього тисків |
| 3 | Середнє значення для зросту та ваги |
| 4 | Медіана для зросту та ваги |
| 5 | Медіана для нижнього та верхнього тисків |
| 6 | Середнє значення для нижнього та верхнього тисків |
| 7 | Середнє значення для зросту та ваги |
| 8 | Медіана для зросту та ваги |
| 9 | Медіана для нижнього та верхнього тисків |
| 10 | Середнє значення для нижнього та верхнього тисків |

```
# Ваш код тут
brand_new_df.pivot_table(["ap_lo", "ap_hi", "cholesterol"],
["8_task"], aggfunc='mean')
```

```
             ap_hi      ap_lo  cholesterol
8_task
False   128.283385  81.466656     1.303848
True    128.341759  81.064512     1.363366
```