Лабораторна робота №2

Використання модуля Pandas.
Аналіз даних по серцево-судинних захворюваннях

Мета роботи: набути практичних навичок роботи з модулем Pandas та провести первинний аналіз даних.

Зміст роботи

Завдання 1. Ретельно опрацювати теоретичні відомості:

https://khashtamov.com/ru/pandas-introduction/

http://nbviewer.jupyter.org/github/Yorko/mlcourse_open/blob/master/jupyter_russian/topic01_pandas_data_analysis/topic1_habr_pandas.ipynb

http://nbviewer.jupyter.org/github/Yorko/mlcourse_open/blob/master/jupyter_russian/topic01_pandas_data_analysis/%5Bsolution%5D_lesson1_practice_pandas_titanic.ipynb

https://www.kaggle.com/crawford/python-groupby-tutorial

***Завдання* 2.** Провести аналіз даних за допомогою Pandas

Опис даних.

Вхідні дані знаходяться у csv-файлі за посиланням:

http://nbviewer.jupyter.org/github/Yorko/mlcourse_open/blob/master/data/mlbootcamp5_train.csv

Dataset сформований з реальних даних, і в ньому використовуються ознаки, які можна розбити на 3 групи:

Об'єктивні ознаки:

- Вік (age)
- Зріст (height)
- Вага (weight)
- Пол (gender)

Результати вимірювання:

- Артеріальний тиск верхній і нижній (ap_hi, ap_lo)
- Холестерин (cholesterol)
- Глюкоза (gluc)

Суб'єктивні ознаки (зі слів пацієнта):

- Куріння (smoke)
- Вживання алкоголю (alco)
- Фізична активність (active)

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

Значення показників холестерину і глюкози представлені одним з трьох класів: норма, вище норми, значно вище норми. Значення суб'єктивних ознак - бінарні.

Всі показники отримані на момент огляду.

Необхідно провести первинний аналіз даних навчальної вибірки за допомогою Pandas.

З бібліотек знадобляться тільки *NumPy* і *Pandas*.

```
import numpy as np
import pandas as pd
```

Зчитаємо дані з csv-файлу в об'єкт pandas *DataFrame*.

```
df = pd.read_csv('mlbootcamp5_train.csv', sep=';',index_col='id')
```

Подивимося на перші 5 записів (`df.head()`).

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

Кожну відповідь необхідно проілюструвати фрагментами програмного коду, що відповідають на наступні питання:

Питання 1 (1 бал). Скільки чоловіків і жінок представлено в цьому наборі даних? Не було дано опису ознаки «стать» (якої статі відповідає 1, а якої - 2 в ознаці *gender*) - це можна визначити подивившись на зріст, при розумному припущенні в середньому чоловіки вище.

Питання 2 (1 бал). Хто в середньому рідше вказує, що вживає алкоголь - чоловіки чи жінки?

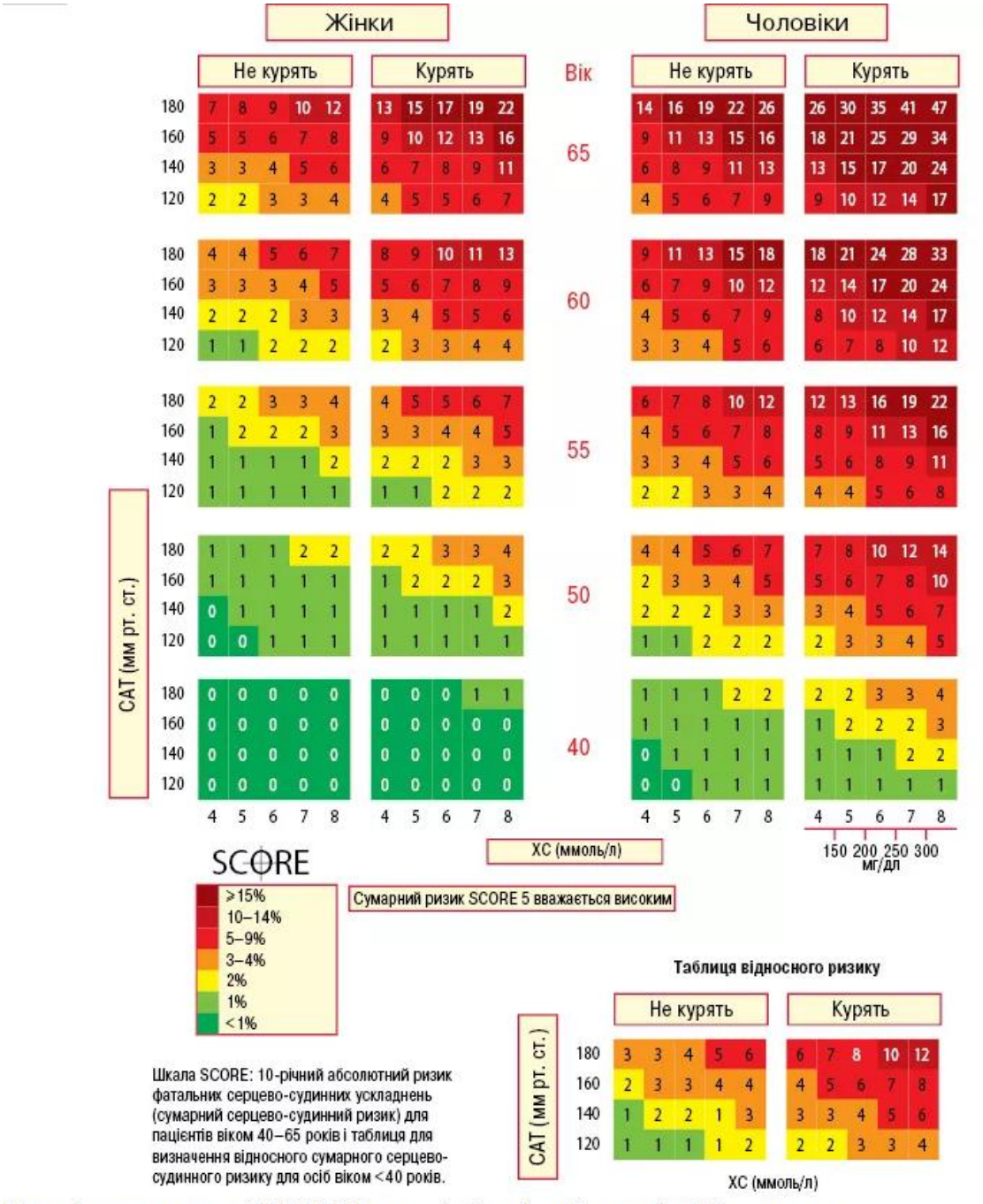
Питання 3 (1 бал). У скільки разів (округлити, *round*) відсоток курців серед чоловіків більше, ніж відсоток курців серед жінок (принаймні, за цими анкетними даними)?

Питання 4 (1 бал). Ви напевно помітили, що значення віку якісь дивні. Здогадайтеся, в чому тут вимірюється вік, і дайте відповідь, на скільки місяців (приблизно) відрізняються медіанне значення віку курців і тих хто не курить.

Питання 5 (2 бали). У статті Wikipedia про серцево-судинний ризик показана шкала SCORE для розрахунку ризику смерті від серцево-судинного захворювання в найближчі 10 років. На рис.1. представлена оцінка ризику серцево-судинних захворювань.

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, в якій застосовано принцип світлофора – три основні кольори: зелений – це низький ризик, що відповідає 1% або менше, жовтий колір – увага! ризик помірний і коливається в межах 2–4%, червоний колір – небезпека! 5% і більше. Для більшої диференціації застосовані відповідні відтінки цих трьох основних кольорів.



Шкала SCORE: 10-річний абсолютний ризик фатальних серцево-судинних ускладнень (сумарний серцево-судинний ризик) для пацієнтів віком 40–65 років і таблиця для визначення відносного сумарного серцево-судинного ризику для осіб віком <40 років.

Шкала оцінки загального ризику ССЗ SCORE. САТ — систолічний артеріальний тиск; тут і далі: ХС — холестерин

Рис.1. Оцінка ризику серцево-судинних захворювань

Правий верхній прямокутник відображає сегмент чоловіків, які палять у віці від 60 до 64 років включно. (Неочевидно, але тут для віку і тиску цифри означають верхню межу, і вона не включається).

Бачимо 9-ку в лівому нижньому кутку цього прямокутника і 47 - в правому верхньому. Тобто якщо при цьому систолічний (тобто верхнє) артеріальний тиск - менше 120 мм рт.ст., а рівень холестерину - 4 ммоль/л, то ризик ССЗ оцінюється приблизно в 5 разів нижче, ніж якби тиск знаходився в інтервалі [160, 180), а холестерину було б 8 ммоль/л.

Порахуємо аналогічне значення на заданих даних.

Уточнення:

– Створіть нову ознаку *age_years* - вік в роках, округливши до цілих (*round*). Для даного прикладу відберіть кращих чоловіків від 60 до 64 років включно.

– Категорії рівня холестерину на малюнку і в наших даних відрізняються. Відображення значень на зображенні в значення ознаки *cholesterol* наступне: 4 ммоль/л →→ 1, 5-7 ммоль/л →→ 2, 8 ммоль/л →→ 3.

– Цікавлять 2 підгрупи чоловіків, які палять вік від 60 до 64 років включно: перша з верхнім артеріальним тиском строго менше 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.

У скільки разів (*round*) відрізняються частки хворих людей (відповідно до цільової ознаки, *cardio*) в цих двох підвибірках? **Порахуйте на представлених даних.**

Питання 6 (2 бали). Побудуйте нову ознаку - BMI (*Body Mass Index*). Для цього треба вагу у кілограмах поділити на квадрат зросту в метрах. Нормальними вважаються значення BMI від 18.5 до 25. Виберіть вірні твердження.

Твердження:

- Медіанний BMI по вибірці перевищує норму.
- У жінок в середньому BMI нижче, ніж у чоловіків.
- У здорових в середньому BMI вище, ніж у хворих.
- У сегменті здорових і тих що не вживають алкоголь чоловіків в середньому BMI ближче до норми, ніж в сегменті здорових і тих що не вживають алкоголь жінок.

Питання 7 (2 бали). Можна помітити, що дані не чисті, багато в них «бруду» і неточностей. Краще це можна побачити на візуалізації даних.

Відфільтруйте наступні сегменти пацієнтів (вважаємо це помилками в даних):

– вказане нижнє значення артеріального тиску строго вище верхнього;

– зріст строго менше 2.5% - перцентілі або строго більше 97.5% - перцентілі (використовуйте *pd.Series.quantile*, якщо не знаєте, що це таке - прочитайте)

– вага строго менше 2.5% - перцентілі або строго більше 97.5% - перцентілі

Це зовсім не вся чистка даних, яку можна було виконати, але поки зупинимося на цьому.

Скільки відсотків даних (round) було видкінуто?

Методичні рекомендації

Python - відмінна мова для аналізу даних і в першу чергу завдяки фантастичній екосистемі пакетів, орієнтованих на дані. Pandas є одним з таких пакетів і значно спрощує імпорт і аналіз даних.

Функція Pandas *dataframe.groupby ()* використовується для розділення даних на групи за деякими критеріями.

Наприклад

```
import numpy as np
import pandas as pd

df = pd.DataFrame({'Animal': ['Falcon', 'Falcon',
                              'Parrot', 'Parrot'],
                   'Max Speed': [380., 370., 24., 26.]})
df
```

	Animal	Max Speed
0	Falcon	380.0
1	Falcon	370.0
2	Parrot	24.0
3	Parrot	26.0

```
df.groupby('Animal')['Max Speed'].mean()
```

```
Animal
Falcon    375.0
Parrot     25.0
Name: Max Speed, dtype: float64
```

```
df['Animal'].value_counts()
```

```
Falcon    2
Parrot    2
Name: Animal, dtype: int64
```