ДЕРЖАВНИЙ УНІВЕРСИТЕТ “ЖИТОМИРСЬКА ПОЛІТЕХНІКА”

Факультет інформаційно-комп’ютерних технологій

Кафедра інженерії програмного забезпечення

**Звіт**

з лабораторної роботи №2

з дисципліни

**«Мови інтелектуального аналізу даних»**

Виконав:

Перевірила: Марчук Г.В.

Житомир - 2020

**Лабораторна робота №2.**

**Використання модуля Pandas. Аналіз даних по серцево-судинних захворюваннях**

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

**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

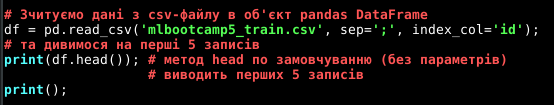
http://nbviewer.jupyter.org/github/Yorko/mlcourse\_open/blob/master/jupyter\_russian/topic01\_pandas\_data\_analysis/%5Bsolution%5D\_lesson1\_practice\_pandas\_titanic.ipynb

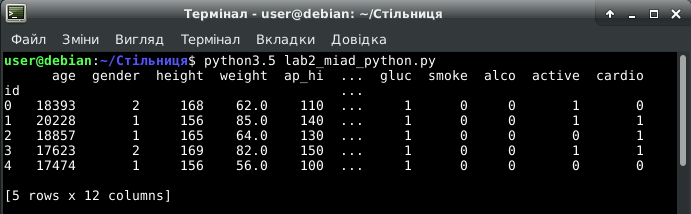
https://www.kaggle.com/crawford/python-groupby-tutorial

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

Зчитуємо дані з CSV-файлу в об'єкт pandas DataFrame та дивимося на перші 5 записів:

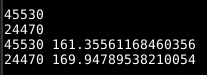






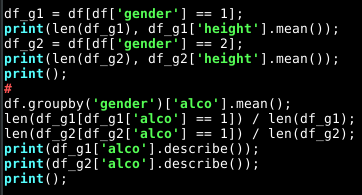
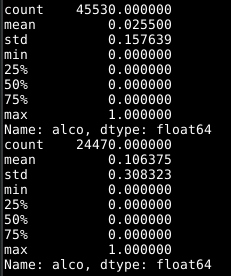
**Питання 1**. Скільки чоловіків і жінок представлено в цьому наборі даних?

Відповідно до зросту, в ознаці «стать» 1 відповідає жінкам, а 2 – чоловікам.

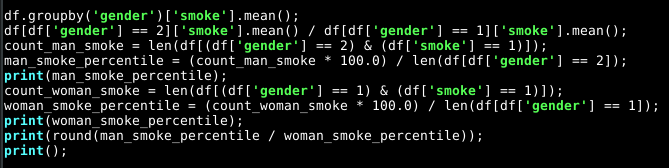
Отримуємо 45530 жінок і 24470 чоловіків.

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

Жінок, які вказали, що вживають алкоголь 1161, натомість чоловіків – 2603. У відсотковому відношенні від загальної кількості виходить, що приблизно 2,5% жінок вказали, що вживають алкоголь, тоді як таких чоловіків 10,63%. Тому в середньому жінки рідше вказують, що вживають алкоголь.

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





Відповідь: в 12 разів.

**Питання 4**. Здогадайтеся, в чому тут вимірюється вік, і дайте відповідь, на скільки місяців (приблизно) відрізняються медіанне значення віку курців і тих хто не курить.

Вік вимірюється в днях.

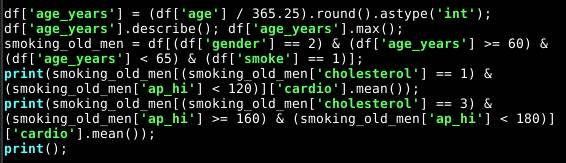




Різниця між медіанним значенням віку курців і тих, хто не курить складає приблизно 20 місяців.

**Питання 5**. У статті на Вікіпедії про серцево-судинний ризик показана шкала SCORE для розрахунку ризику смерті від серцево-судинного захворювання в найближчі 10 років. Порахувати аналогічне значення на наших даних.

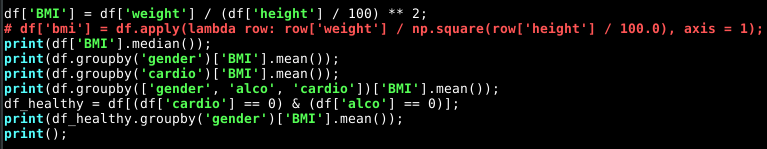
У скільки разів (round) відрізняються частки хворих людей (відповідно до цільової ознаки, cardio) в цих двох підвибірках?

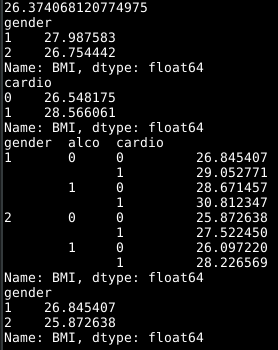




Відповідь: частки хворих людей (відповідно до цільової ознаки, cardio) в цих двох підвибірках відрізняються приблизно в 3 рази.

**Питання 6**. 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.





Відповідь:

1) Перше твердження вірне, медіанний BMI по вибірці перевищує норму в 25 одиниць.

2) Друге твердження не вірне, у жінок в середньому BMI вище, ніж у чоловіків.

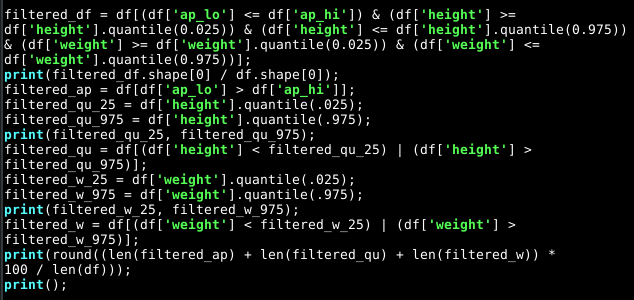
3) Третє твердження не вірне. У здорових в середньому BMI нижче, ніж у хворих.

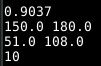
4) Четверту твердження вірне. У сегменті здорових і тих що не вживають алкоголь чоловіків в середньому BMI ближче до норми, ніж в сегменті здорових і тих що не вживають алкоголь жінок.

**Питання 7.**

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





В результаті викинули 10% даних.