Зображення, що містить текст

Автоматично згенерований опис

Міністерство освіти і науки України

Національний технічний університет України

“Київський політехнічний інститут імені Ігоря Сікорського”

Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

**Лабораторна робота №2**

Аналіз даних з використанням мови Python

**Тема:** Статистичний аналіз даних

**Варіант:** 1

Виконав Перевірив:

студент групи ІП-11: Тимофєєва Ю. С

Панченко С. В.

Київ 2023

Зміст

[1 Мета лабораторної роботи 6](#__RefHeading___Toc131_2937707210)

[2 Завдання 7](#__RefHeading___Toc133_2937707210)

[3 Виконання 8](#__RefHeading___Toc135_2937707210)

[3.1 Знайти середній вік матерів і батьків і порівняти ці середні значення 8](#__RefHeading___Toc137_2937707210)

[3.2 Перевірити чи нормально розподілена вага дітей 9](#__RefHeading___Toc139_2937707210)

[3.3 Перевірити за допомогою статистичних гіпотез чи у матерів, що палять, легші діти 9](#__RefHeading___Toc141_2937707210)

[3.4 Чи є зв’язок між зростом матері та дитини? 10](#__RefHeading___Toc143_2937707210)

[4 Висновок 11](#__RefHeading___Toc145_2937707210)

# Мета лабораторної роботи

Ознайомитись з основними функціями бібліотеки NumPy та SciPy для описової статистики, перевірки статистичних гіпотез, кореляційного аналізу та лінійної регресії.

# Завдання

Варіант 1.

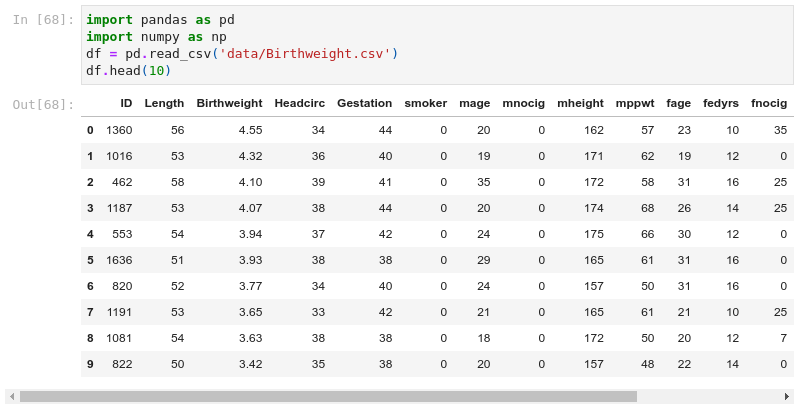
Файл Birthweight.csv.

1. Знайти середній вік матерів і батьків і порівняти ці середні значення.
2. Перевірити чи нормально розподілена вага дітей.
3. Перевірити за допомогою статистичних гіпотез чи у матерів, що палять, легші діти.
4. Чи є зв’язок між зростом матері та дитини?

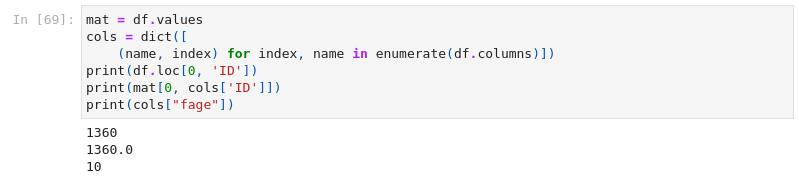
# Виконання

## Знайти середній вік матерів і батьків і порівняти ці середні значення

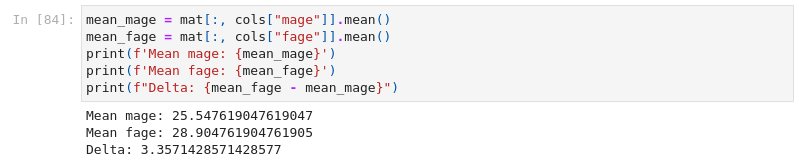
Для початку створимо датафрейм з датасету Birthweight.csv.

  
  
Рисунок 3.1 - Датафрейм Birthweight.csv

Оскільки лабораторна робота вимагає саме використання NumPy, а не Pandas, то дістанемо матрицю даних з датафрейма, а індекси колонок будемо знаходити за допомогою словника.

  
  
Рисунок 3.2 - Робота з матрицею NumPy

Знайдемо середній вік матері та батька по всіх новонароджених, порівняємо їх. mage - колонка віку матері дитини, fage - колонка віку батька дитини. Побачимо, що вік батька в середньому вищий.

  
  
Рисунок 3.3 - Середній вік батьків - матері та батька дитини

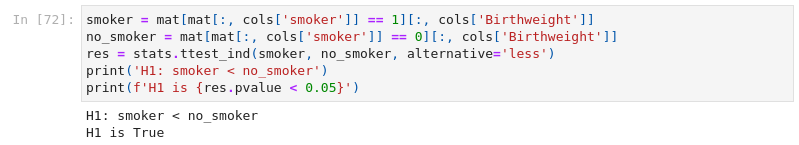
## Перевірити чи нормально розподілена вага дітей

Перевіримо нормальність розподілу ваги дітей за допомогою критерію Андерсона. Імпортуємо з пакету scipy модуль stats, використаємо тест Колмогорова-Смірнова. Побачимо, що розподіл ваги дітей не є нормальним. Якщо pvalue < 0.05, то приймається альтернативна гіпотеза.

  
  
Рисунок 3.4 - Перевірка розподілу ваги на нормальність за допомогою теста Колмогорова-Смірнова

## Перевірити за допомогою статистичних гіпотез чи у матерів, що палять, легші діти

Застосуємо двовибірковий t-критерій з альтернативною гіпотезою того, що у матерів, які палять, легші діти. Як бачимо альтернативна гіпотеза приймається, тому у курців діти мають меншу масу ніж у тих людей, що не палять.

  
  
Рисунок 3.5 - Доведення альтернативної гіпотези, що курці мають новонароджених з меншою масою

## Чи є зв’язок між зростом матері та дитини?

Використаємо тест Пірсона для перевірки даної гіпотези. Функція stats.pearsonr бере за нульову гіпотезу те, що послідовності є некорельованими та нормально розподіленими. 'two-sided' - кореляція є ненульовою, 'less' - кореляція є від'ємною, 'greater' - кореляція є додатною. Бачимо, що поле staticstic показує чітку кореляцію, а оскільки pvalue > 0.05, то кореляція існує точно.

  
  
Рисунок 3.6 - Тест Пірсона для перевірки кореляції між зростом матері та дитини

# Висновок

Під час виконання даної лабораторної роботи я ознайомився з основними функціями бібліотеки NumPy та SciPy для описової статистики, перевірки статистичних гіпотез, кореляційного аналізу та лінійної регресії. У першому завданні було отримано, що середній вік батьків вищий ніж матерів. У другому завданні був використаний тест Колмогорова-Смірнова, далі побачили, що вага немовлят не розподілена нормально. За допомогою двовибіркового t-критерію дізналися, що матері, які палять, мають легших новонароджених дітей. І в кінці за допомогою тесту Пірсона вказали, що існує кореляція між зростом матері та дитини.