Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет інформатики та обчислювальної техніки
Кафедра інформаційних систем та технологій

Лабораторна робота №2
з дисципліни
Аналіз даних з використанням мови Python

Виконав:

студент групи ІА-24:
Криворучек В.С.

Перевірила:

ст. викладач
Тимофєєва Ю.С.

Київ 2025

Тема: Статистичний аналіз даних

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

Хід роботи

Завдання:

1.Яка середня кількість дітей в сім’ї і її відхилення?

2.Перевірити чи нормально розподілені доходи.

3.Чи є зв’язок між витратами на пальне та витратами на транспорт?

4.Побудувати лінійну регресійну модель залежності витрат на їжу від доходу.

Код програми:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import shapiro, pearsonr, linregress
import statsmodels.api as sm

# Завантаження даних
file_path = "Budget.csv"
data = pd.read_csv(file_path)

if 'Unnamed: 0' in data.columns:
    data = data.drop(columns=['Unnamed: 0'])

print("\n===== Перші 5 рядків даних =====")
print(data.head().to_string(index=False))

# 1. Розрахунок середньої кількості дітей та стандартного відхилення
mean_children = data['children'].mean()
std_children = data['children'].std()
print("\n===== Кількість дітей у сім'ї =====")
print(f"Середня кількість дітей: {mean_children:.2f}")
print(f"Стандартне відхилення: {std_children:.2f}\n")

# 2. Перевірка нормальності розподілу доходів
stat, p = shapiro(data['income'])
print("===== Перевірка нормальності доходів (Шапіро-Уілк) =====")
print(f"Статистика тесту: {stat:.3f}")
print(f"p-значення: {p:.3f}")
if p > 0.05:
    print("Доходи мають нормальний розподіл. (H0 не відхиляється)\n")
else:
    print("Доходи НЕ мають нормального розподілу. (H0 відхиляється)\n")

# Візуалізація розподілу доходів
plt.figure(figsize=(12, 5))
```

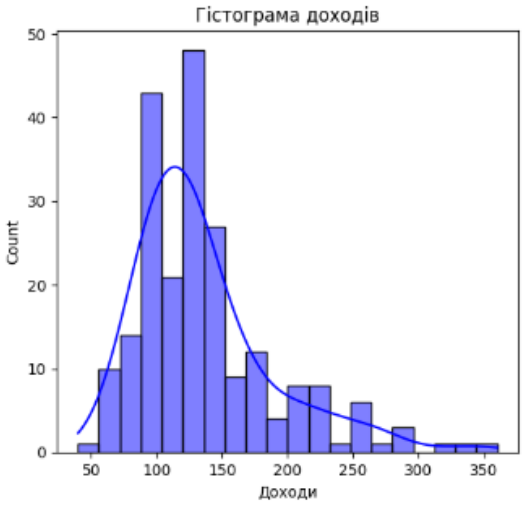
```
plt.subplot(1, 2, 1)
sns.histplot(data['income'], bins=20, kde=True, color='blue')
plt.title("Гістограма доходів")
plt.xlabel("Доходи")

plt.subplot(1, 2, 2)
sns.boxplot(x=data['income'], color='blue')
plt.title("Boxplot доходів")
plt.xlabel("Доходи")

plt.show()

# 3. Кореляція між витратами на пальне та транспорт
corr_coef, p_value = pearsonr(data['wfuel'], data['wtrans'])
print("===== Кореляція між витратами на пальне та транспорт =====")
print(f"Коефіцієнт Пірсона: {corr_coef:.3f}")
print(f"p-значення: {p_value:.3f}")
if p_value < 0.05:
    print("Є статистично значущий зв'язок між витратами на пальне та
транспортом. (H0 відхиляється)\n")
else:
    print("Статистично значущого зв'язку немає. (H0 не
відхиляється)\n")

# 4. Лінійна регресія: витрати на їжу від доходу
X = data['income']
y = data['wfood']
X_const = sm.add_constant(X)
model_sm = sm.OLS(y, X_const).fit()

print("===== Лінійна регресія: витрати на їжу ~ дохід =====")
print(model_sm.summary())

# Візуалізація регресії
plt.figure(figsize=(8, 6))
sns.regplot(x='income', y='wfood', data=data, color='blue',
line_kws={'color': 'red'})
plt.xlabel('Дохід')
plt.ylabel('Витрати на їжу')
plt.title('Регресія: витрати на їжу від доходу')
plt.show()
```

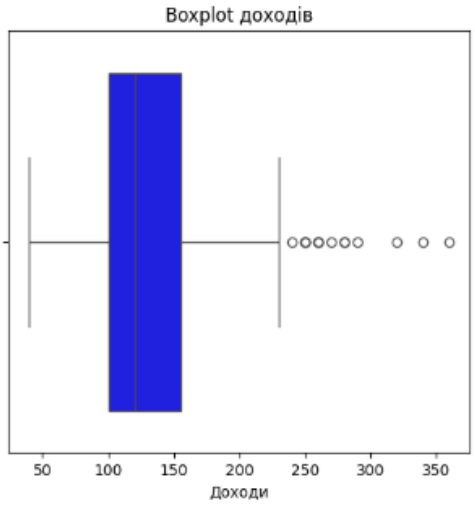
Результат виконання:

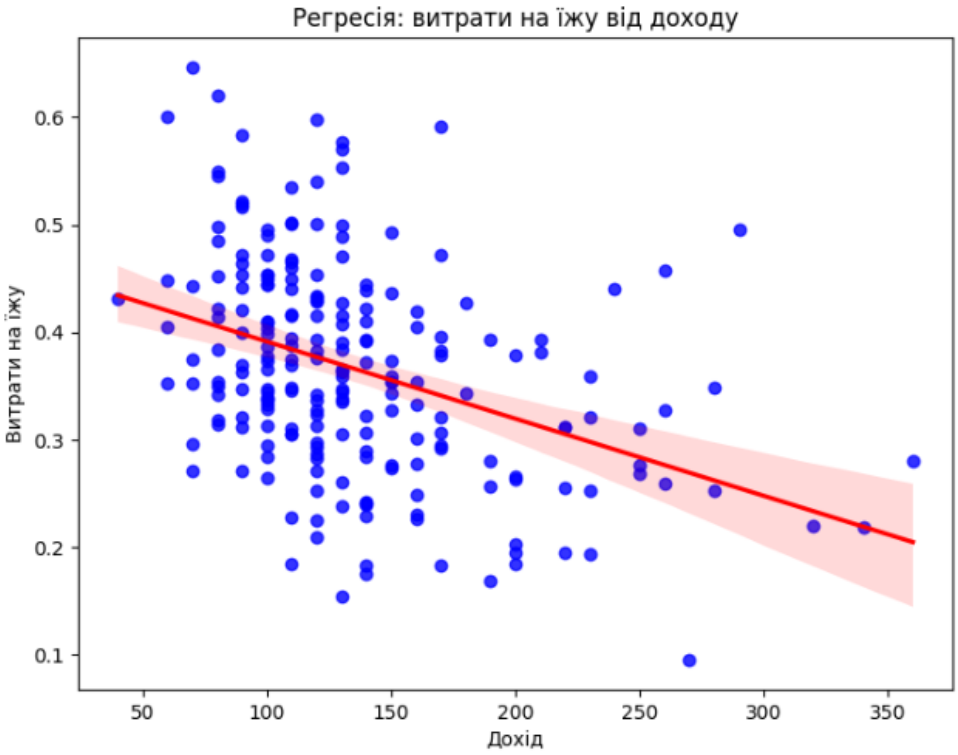
```
===== Перші 5 рядків даних =====
 wfood  wfuel  wcloth   walc  wtrans  wother  totexp  income  age  children
0.4272 0.1342  0.0000 0.0106  0.1458  0.2822      50     130   25         2
0.3739 0.1686  0.0091 0.0825  0.1215  0.2444      90     150   39         2
0.1941 0.4056  0.0012 0.0513  0.2063  0.1415     180     230   47         2
0.4438 0.1258  0.0539 0.0397  0.0652  0.2716      80     100   33         2
0.3331 0.0824  0.0399 0.1571  0.2403  0.1473      90     100   31         1

===== Кількість дітей у сім'ї =====
Середня кількість дітей: 1.60
Стандартне відхилення: 0.49
```

```
===== Перевірка нормальності доходів (Шапіро-Уілк) =====
Статистика тесту: 0.882
p-значення: 0.000
Доходи НЕ мають нормального розподілу. (H0 відхиляється)

===== Кореляція між витратами на пальне та транспорт =====
Коефіцієнт Пірсона: -0.199
p-значення: 0.003
Є статистично значущий зв'язок між витратами на пальне та транспортом. (H0 відхиляється)
```

```
===== Лінійна регресія: витрати на їжу ~ дохід =====
                            OLS Regression Results
==============================================================================
Dep. Variable:                  wfood   R-squared:                       0.151
Model:                            OLS   Adj. R-squared:                  0.147
Method:                 Least Squares   F-statistic:                     38.60
Date:                Wed, 19 Mar 2025   Prob (F-statistic):           2.62e-09
Time:                        12:15:35   Log-Likelihood:                 213.52
No. Observations:                 219   AIC:                            -423.0
Df Residuals:                     217   BIC:                            -416.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4631      0.017     27.484      0.000       0.430       0.496
income        -0.0007      0.000     -6.213      0.000      -0.001      -0.000
==============================================================================
Omnibus:                        3.190   Durbin-Watson:                   2.023
Prob(Omnibus):                  0.203   Jarque-Bera (JB):                3.148
Skew:                           0.292   Prob(JB):                        0.207
Kurtosis:                       2.935   Cond. No.                         397.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Process finished with exit code 0
```







Кількість дітей в середньому ≈ 1.6, невелика варіація між сім'ями. Доходи не мають нормального розподілу, що варто враховувати в майбутньому аналізі. Є слабкий, але значущий негативний зв'язок між витратами на пальне та транспорт. Збільшення доходу зменшує частку витрат на їжу, але модель пояснює лише 15.1% варіації.

Висновок: У ході виконання даної лабораторної роботи я ознайомився з основними функціями бібліотеки NumPy та SciPy для описової статистики, перевірки статистичних гіпотез, кореляційного аналізу та лінійної регресії.