**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**ІМЕНІ ТАРАСА ШЕВЧЕНКА**

**ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**

**Кафедра інтелектуальних технологій**

**Лабораторна робота №1**

з дисципліни «Інтелектуальний аналіз даних»

Тема роботи: «Візуалізація наборів даних. Загальна статистика»

**Варіант №5**

Виконала:

студентка групи АнД-31

Радоманова С. П.

Перевірив:

Чолишкіна О.Г.

**Київ – 2025**

**Мета роботи:** Вивчення можливостей і основних принципів роботи по

виконанню статистичного аналізу даних. Графічний розвідувальний аналіз

даних. Первинний статистичний аналіз даних.

**Завдання:**

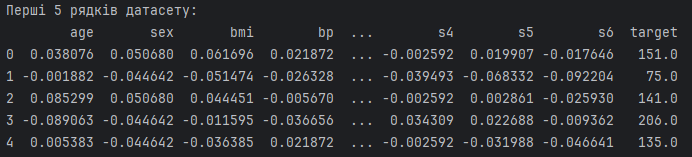
В ході виконання роботи було завантажено датасет датасет Diabetes dataset

З першого переліку і HR Dataset (Multinational Company)з другогою

Посилання на датасети:

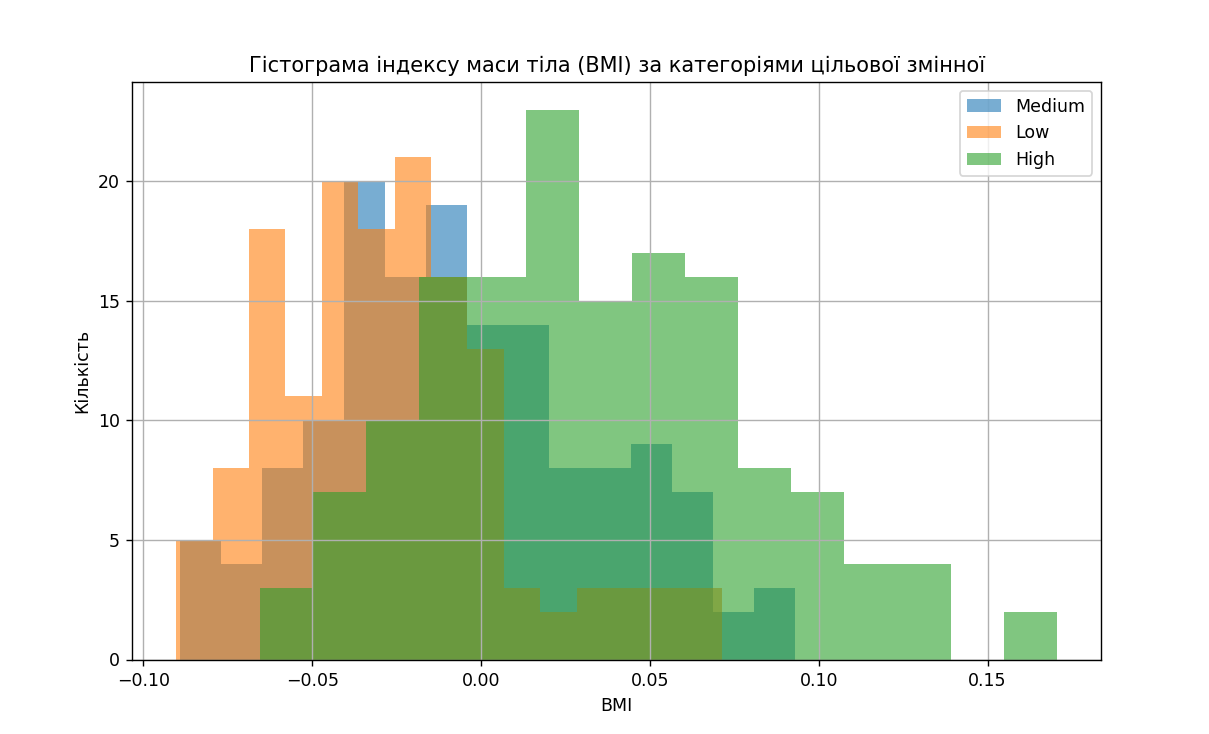
1. [https://scikit-learn.org/1.5/datasets/toy\_dataset.html#diabetes-dataset](https://scikit-learn.org/1.5/datasets/toy_dataset.html%23diabetes-dataset)
2. <https://www.who.int/data/gho/data/themes/topics/global-and-regional-sti-estimates>

Загальна інформація по першому датасету:



Тут значення таргет - це кількісна змінна, яка показує прогнозований рівень прогресування діабету через рік, на основі медичних показників пацієнта. Це числове значення, що вказує на рівень впливу факторів на діабет (в діапазоні приблизно від 25 до 346).

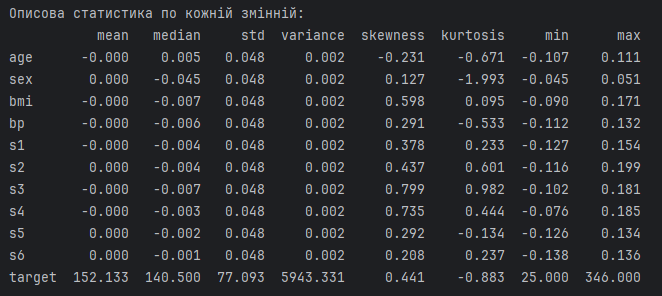
Побудувана категоризована гістограма:

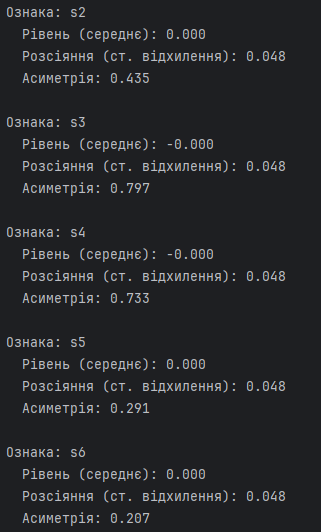
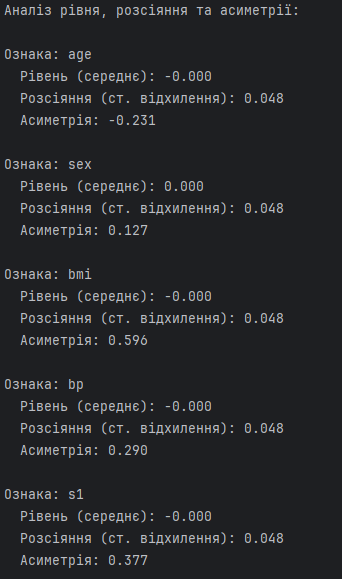


Побудувана радіальна діаграма:



Обчислена описова статистика: медіана, середньоквадратичне відхилення, дисперсія, асиметрія, ексцес, min, max

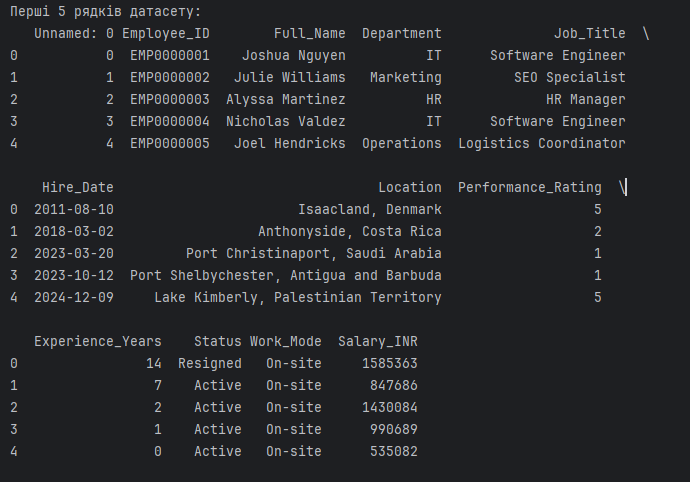


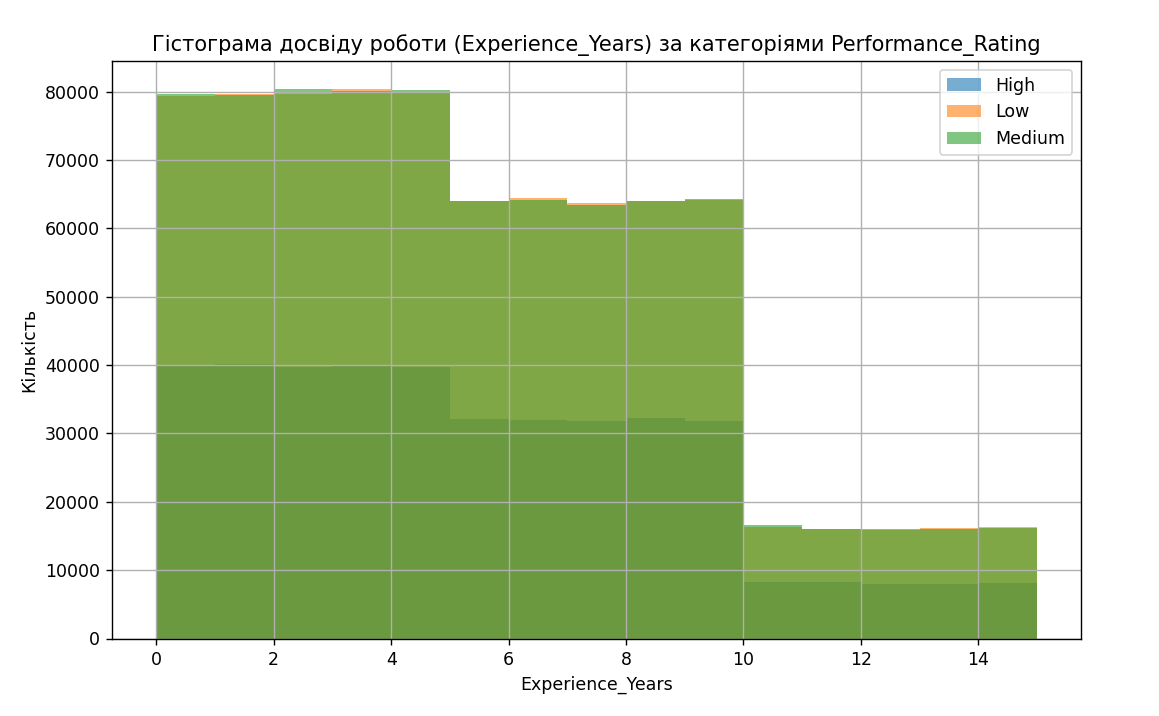
Проведений аналіз рівня, розсіяння та асиметрії

Оскільки всі значення в цій бібліотеці нормалізовані, тому середні значення всіх ознак в цій бібліотеці =0.

Повторюємо все те саме для другого датасету, який представляє кадрову інформацію про працівників мультинаціональних компаній.

Загальна інформація по другому датасету:

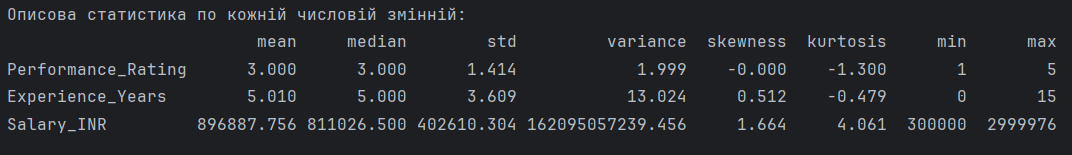


Тут в якості таргету виступає стовпчик Performance\_Rating з оцінками роботи від 1 до 5. 

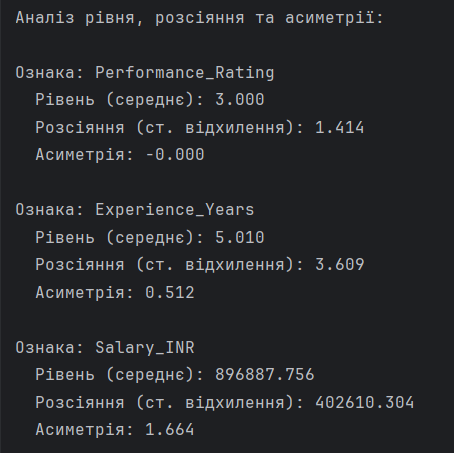
Оскільки обраний датасет містить два числові показники окрім оцінок роботи, а інша показники неможливо конвертувати у числові змінні, радіальна діаграма виходить зовсім незмістовна



Обчислена описова статистика(по числовим змінним): медіана, середньоквадратичне відхилення, дисперсія, асиметрія, ексцес, min, max



Проведений аналіз рівня, розсіяння та асиметрії



**Висновки:** У ході виконання лабораторної роботи було проведено комплексний статистичний аналіз двох різнотипних наборів даних: Diabetes dataset та HR датасету мультинаціональної компанії.

Diabetes datasetмістить нормалізовані медичні показники пацієнтів з цільовою змінною - рівнем прогресування діабету. Усі ознаки мають нульове математичне сподівання через попередню нормалізацію. Категоризовані гістограми та радіальні діаграми ефективно візуалізують взаємозв'язки між ознаками та цільовою змінною

HR датасет містить обмежену кількість числових змінних (лише три), що обмежило можливості багатовимірного аналізу. Радіальна діаграма виявилася малоінформативною через недостатню кількість числових ознак. Аналіз розсіяння показав значну дисперсію у зарплатах (Salary\_INR), тоді як досвід роботи (Experience\_Years) має відносно рівномірний розподіл

**Лістинг програми:**

import numpy as np  
import pandas as pd  
from scipy.stats import kurtosis, skew  
import matplotlib.pyplot as plt  
from sklearn.datasets import load\_diabetes  
  
# Завантаження  
data = load\_diabetes()  
df = pd.DataFrame(data.data, columns=data.feature\_names)  
df['target'] = data.target  
  
  
print("Перші 5 рядків датасету:")  
print(df.head())  
  
#Категоризація цільової змінної  
df['target\_group'] = pd.qcut(df['target'], q=3, labels=['Low', 'Medium', 'High'])  
  
#Категоризована гістограма  
plt.figure(figsize=(10, 6))  
for label in df['target\_group'].unique():  
 subset = df[df['target\_group'] == label]  
 plt.hist(subset['bmi'], bins=15, alpha=0.6, label=label)  
plt.title('Гістограма індексу маси тіла (BMI) за категоріями цільової змінної')  
plt.xlabel('BMI')  
plt.ylabel('Кількість')  
plt.legend()  
plt.grid(True)  
plt.show()  
  
#Радіальна діаграма  
# Групування по категоріях  
grouped = df.groupby('target\_group').mean(numeric\_only=True)  
  
features = data.feature\_names  
angles = np.linspace(0, 2 \* np.pi, len(features), endpoint=False).tolist()  
angles += angles[:1]  
  
fig, ax = plt.subplots(figsize=(8, 8), subplot\_kw=dict(polar=True))  
  
for group\_name, row in grouped.iterrows():  
 values = row[features].tolist()  
 values += values[:1]  
 ax.plot(angles, values, label=f'{group\_name}')  
 ax.fill(angles, values, alpha=0.2)  
  
ax.set\_xticks(angles[:-1])  
ax.set\_xticklabels(features)  
ax.set\_title("Радіальна діаграма середніх значень ознак за групами")  
ax.legend()  
plt.show()  
  
# Описова статистика по всіх змінних  
desc\_stats = pd.DataFrame({  
 'mean': df.mean(numeric\_only=True),  
 'median': df.median(numeric\_only=True),  
 'std': df.std(numeric\_only=True),  
 'variance': df.var(numeric\_only=True),  
 'skewness': df.skew(numeric\_only=True),  
 'kurtosis': df.kurtosis(numeric\_only=True),  
 'min': df.min(numeric\_only=True),  
 'max': df.max(numeric\_only=True)  
})  
  
print("\nОписова статистика по кожній змінній:")  
print(desc\_stats.round(3))  
  
#Висновок по рівню, розсіянню, асиметрії  
print("\nАналіз рівня, розсіяння та асиметрії:")  
for column in data.feature\_names:  
 print(f"\nОзнака: {column}")  
 print(f" Рівень (середнє): {df[column].mean():.3f}")  
 print(f" Розсіяння (ст. відхилення): {df[column].std():.3f}")  
 print(f" Асиметрія: {skew(df[column]):.3f}")

import numpy as np  
import pandas as pd  
from scipy.stats import kurtosis, skew  
import matplotlib.pyplot as plt  
  
#налаштування для відображення в пандасі  
pd.set\_option("display.max\_columns", None) # показувати всі колонки  
pd.set\_option("display.max\_rows", None) # показувати всі рядки (обережно з великими датафреймами)  
pd.set\_option("display.width", None) # прибирає перенесення рядків  
pd.set\_option("display.max\_colwidth", None) # повна довжина назв колонок  
pd.set\_option("display.float\_format", '{:.3f}'.format) # завжди показувати float у нормальному форматі  
  
  
  
  
# Завантаження  
file\_path = r"E:\3 курс\1 сем\IDA\datasets\HR\_Data\_MNC\_Data Science Lovers.csv"  
df = pd.read\_csv(file\_path)  
  
# Видаляємо непотрібні колонки  
df = df.drop(columns=['Unnamed: 0'], errors='ignore')  
  
print("Перші 5 рядків датасету:")  
print(df.head())  
  
# Категоризація цільової змінної Performance\_Rating  
df['target\_group'] = pd.qcut(df['Performance\_Rating'], q=3, labels=['Low', 'Medium', 'High'])  
  
# Категоризована гістограма для Experience\_Years  
plt.figure(figsize=(10, 6))  
for label in df['target\_group'].unique():  
 subset = df[df['target\_group'] == label]  
 plt.hist(subset['Experience\_Years'], bins=15, alpha=0.6, label=label)  
plt.title('Гістограма досвіду роботи (Experience\_Years) за категоріями Performance\_Rating')  
plt.xlabel('Experience\_Years')  
plt.ylabel('Кількість')  
plt.legend()  
plt.grid(True)  
plt.show()  
  
from sklearn.preprocessing import MinMaxScaler  
  
# Радіальна діаграма середніх значень ознак за групами  
numeric\_columns = df.select\_dtypes(include=[np.number]).columns  
features = numeric\_columns.drop('Performance\_Rating', errors='ignore')  
  
# Обчислюємо середні значення ознак для кожної групи  
grouped = df.groupby('target\_group')[features].mean()  
  
# Нормалізуємо (0-1), щоб усі ознаки були порівнянні  
scaler = MinMaxScaler()  
grouped\_scaled = pd.DataFrame(scaler.fit\_transform(grouped),  
 columns=grouped.columns,  
 index=grouped.index)  
  
# Кути для діаграми  
angles = np.linspace(0, 2 \* np.pi, len(features), endpoint=False).tolist()  
angles += angles[:1]  
  
# Побудова графіка  
fig, ax = plt.subplots(figsize=(8, 8), subplot\_kw=dict(polar=True))  
  
for group\_name, row in grouped\_scaled.iterrows():  
 values = row.tolist()  
 values += values[:1] # замикаємо контур  
 ax.plot(angles, values, label=f'{group\_name}')  
 ax.fill(angles, values, alpha=0.25)  
  
ax.set\_xticks(angles[:-1])  
ax.set\_xticklabels(features, fontsize=9)  
ax.set\_title("Радіальна діаграма середніх значень ознак за групами Performance\_Rating")  
ax.legend(loc="upper right", bbox\_to\_anchor=(1.2, 1.1))  
plt.show()  
  
# Описова статистика по всіх числових змінних  
desc\_stats = pd.DataFrame({  
 'mean': df.mean(numeric\_only=True),  
 'median': df.median(numeric\_only=True),  
 'std': df.std(numeric\_only=True),  
 'variance': df.var(numeric\_only=True),  
 'skewness': df.skew(numeric\_only=True),  
 'kurtosis': df.kurtosis(numeric\_only=True),  
 'min': df.min(numeric\_only=True),  
 'max': df.max(numeric\_only=True)  
})  
  
print("\nОписова статистика по кожній числовій змінній:")  
print(desc\_stats.round(3))  
  
# Висновок по рівню розсіянню, асиметрії  
print("\nАналіз рівня, розсіяння та асиметрії:")  
for column in numeric\_columns:  
 print(f"\nОзнака: {column}")  
 print(f" Рівень (середнє): {df[column].mean():.3f}")  
 print(f" Розсіяння (ст. відхилення): {df[column].std():.3f}")  
 print(f" Асиметрія: {skew(df[column]):.3f}")