Лабораторна робота №4

Візуалізація даних за допомогою matplotlib та Seaborn

Мета роботи: Ознайомитись з основними діаграмами та графіками, що використовуються при аналізі даних. Навчитись будувати їх за допомогою бібліотек matplotlib та Seaborn.

Короткі теоретичні відомості

Для візуалізації даних використовується бібліотека Matplotlib, яка працює з ndarray. Існують бібліотеки, що надають API для роботи з Matplotlib, які дозволяють розширювати функціональні можливості. Наприклад, бібліотека Seaborn, зокрема, працює з об'єктами бібліотеки Pandas.

```
import matplotlib.pyplot as plt
import seaborn as sns
```

Стовпчикова діаграма створюється шляхом позначення всіх категорій даних на одній осі та частоти кожної категорії даних по іншій осі. Висота стовпчика (якщо діаграма вертикальна) або його довжина (якщо діаграма горизонтальна) показує частоту кожної категорії. Категорії не впорядковані, а між стовпчиками зазвичай присутні проміжки.

Стовпчикова діаграма використовується, коли потрібно порівняти значення показників у різних підгрупах даних.

Часто по осі x відкладається ознака з якісними значеннями.

Кількісні ознаки набувають впорядкованих числових значень. Ці значення можуть бути дискретними, як цілі числа, або безперервними, як дійсні числа, і зазвичай виражають підрахунок або вимірювання.

Найпростіший спосіб подивитися на розподіл кількісної ознаки — побудувати її гістограму

Діаграма розмаху або коробкова діаграма — це схематичне представлення положення даних, включаючи найменші та найбільші значення, нижню та верхню чверть вибірки (нижній та верхній квартилі), медіану та статистичні викиди.

В багатьох випадках доводиться працювати з даними, що мають багатовимірний характер. Тобто кожне спостереження складається з вимірювань декількох змінних.

Діаграма розсіювання або точкова діаграма часто використовується для графічного відображення потенційного зв'язку між парою змінних. Значення однієї змінної відкладаються на осі X, другою – на осі Y. Якщо спостереження має більше двох змінних, то використовується декілька діаграм розсіювання, що зображають зв'язок між кожною парою цих змінних.

В загальному випадку теплокарти — це графічне представлення даних різними кольорами в залежності від значень. Нерідко їх використовують для представлення кореляційної матриці, візуалізуючи ступінь зв'язку між різними змінними.

Завдання до лабораторної роботи

Створити програму, яка виконує завдання відповідно до варіанту з використанням matplotlib та/або Seaborn.

Оформити звіт. Звіт повинен містити:

— титульний лист;
— код програми;
— результати виконання коду.

Продемонструвати роботу програми та відповісти на питання стосовно теоретичних відомостей та роботи програми.

Варіант 1.
Файл diamonds.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість діамантів кожного з класів якості; б) максимальну ціну діамантів кожного класу якості; в) середню глибину діамантів різного класу якості з різною якістю кольору.
2. Побудувати гістограму глибини діамантів у відсотках (depth), загальну і для кожного класу якості.
3. Побудувати діаграму розмаху параметру table (загальну і в залежності від якості кольору), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) довжиною і шириною; б) глибиною у % і глибиною у мм. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 2.
Файл penguins.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість пінгвінів кожного виду; б) мінімальну довжину дзьобу пінгвінів кожного виду; в) середню вагу пінгвінів кожного виду з розподілом за статтю.
2. Побудувати гістограму глибини дзьобу, загальну і для кожного виду.
3. Побудувати діаграму розмаху довжини ласт (загальну і в залежності від виду), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) довжиною і глибиною дзьобу; б) вагою і довжиною ласт. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 3.
Файл merc.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість мерседесів кожної моделі; б) медіанну ціну мерседесів кожної

моделі; в) середню ціну мерседесів кожної моделі з розподілом за типом коробки передач.
2. Побудувати гістограму розподілу ціни, загальну і для кожного виду коробки передач.
3. Побудувати діаграму розмаху витрат палива (загальну і в залежності від типу палива), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) ціною та пробігом; б) витратами на паливо та об'ємом двигуна. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 4.

Файл insurance.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість людей з кожного регіону; б) мінімальний вік людей кожного регіону; в) середню кількість дітей у людей з кожного регіону з розподілом за статтю.
2. Побудувати гістограму індексу маси тіла, загальну і для курців і не курців.
3. Побудувати діаграму розмаху витрат (загальну і в залежності від регіону), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) віком та витратами; б) індексом маси тіла та витратами. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 5.

Файл StudentsPerformance.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість учнів кожної раси/етносу; б) максимальні бали за математику у учнів кожної раси/етносу; в) середні бали за письмо у учнів кожної раси/етносу з розподілом за статтю.
2. Побудувати гістограму балів за читання, загальну і в залежності від проходження підготовчого курсу.
3. Побудувати діаграму розмаху балів за математику (загальну і в залежності від рівня освіти батьків), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) балами за читання і письмо; б) балами за математику і читання. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 6.

Файл bike.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість покупців різних професій; б) медіанний дохід покупців різних

професій; в) середній вік покупців різних професій з розподілом на тих, хто купив велосипед і хто — ні.
2. Побудувати гістограму кількості дітей, загальну і в залежності від покупки велосипеду.
3. Побудувати діаграму розмаху доходу (загальну і в залежності від рівня освіти), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) доходами і віком; б) кількістю дітей і машин. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 7.

Файл telecom.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість лояльних та нелояльних клієнтів; б) максимальну кількість хвилин в денний час для лояльних та нелояльних клієнтів; в) середню кількість дзвінків в денний час з врахуванням, чи підключена голосова пошта, і з врахуванням лояльності.
2. Побудувати гістограму оплати в денний час, загальну і в залежності від лояльності.
3. Побудувати діаграму розмаху дзвінків в вечірній час (загальну і в залежності від підключеного роумінгу), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) хвилинами і дзвінками в денний час; б) хвилинами і оплатою в денний час. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 8.

Файл diamonds.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість діамантів кожного з класів якості кольору; б) мінімальну вагу діамантів кожного класу якості кольору; в) середню ціну діамантів різного класу якості з різною якістю кольору.
2. Побудувати гістограму довжини діамантів, загальну і для кожної якості кольору.
3. Побудувати діаграму розмаху параметру depth (загальну і в залежності від якості) визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) довжиною і вагою; б) глибиною у % і вагою. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 9.

Файл merc.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість мерседесів кожного року реєстрації; б) максимальний пробіг мерседесів кожного року реєстрації; в) середній пробіг мерседесів кожного року реєстрації з розподілом за типом палива.
2. Побудувати гістограму розподілу витрат палива, загальну і для кожного типу палива.
3. Побудувати діаграму розмаху пробігу (загальну і в залежності від типу коробки передач), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) пробігом та роком реєстрації; б) витратами на паливо та ціною. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 10.

Файл penguins.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість пінгвінів на кожному острові; б) медіанну вагу пінгвінів на кожному острові; в) середню довжину ласт пінгвінів на кожному острові з розподілом за статтю.
2. Побудувати гістограму довжини ласт пінгвінів, загальну і для кожного виду.
3. Побудувати діаграму розмаху ваги пінгвінів (загальну і в залежності від острову), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) довжиною дзьобу і ласт; б) вагою і довжиною дзьобу. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 11

Файл insurance.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість людей, що мають певну кількість дітей; б) максимальний дохід людей з різною кількістю дітей; в) середній вік людей з різною кількістю дітей з розподілом за статтю.
2. Побудувати гістограму витрат, загальну і за статтю.
3. Побудувати діаграму індексу маси тіла (загальну і в залежності від того, палить людина чи ні), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) віком та індексом маси тіла; б) кількістю дітей та віком. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 12.

Файл StudentsPerformance.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість учнів з батьками, що мають різні рівні освіти; б) мінімальні бали за

письмо з батьками, що мають різні рівні освіти; в) середні бали за читання у учнів з батьками, що мають різні рівні освіти з розподілом за тим, чи пройдено підготовчий курс.
2. Побудувати гістограму балів за математику, загальну і для різних типів обіду.
3. Побудувати діаграму розмаху балів за читання (загальну і в залежності від раси/етносу), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) балами за письмо і математику; б) балами за читання і математику. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 13.

Файл telecom.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість клієнтів з підключеним роумінгом та без; б) максимальну кількість хвилин в нічний час для клієнтів з підключеним роумінгом та без; в) середню кількість міжнародних дзвінків для клієнтів з підключеним роумінгом та без і з врахуванням лояльності.
2. Побудувати гістограму оплати в вечірній час, загальну і в залежності від лояльності.
3. Побудувати діаграму хвилин на міжнародні дзвінки (загальну і в залежності від підключеного роумінгу), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) хвилинами і дзвінками в нічний час; б) хвилинами і оплатою в нічний час. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 14.

Файл Stars.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість зірок різного кольору; б) медіанну температуру зірок різного кольору; в) середній радіус зірок різного кольору з розподілом за типом зірки.
2. Побудувати гістограму температури, загальну і в залежності від спектрального класу.
3. Побудувати діаграму розмаху світності (загальну і в залежності від кольору), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) температурою і абсолютною величиною; б) радіусом та температурою. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.

Варіант 15.

Файл telecom.csv.

1. Побудувати стовпчикові діаграми, на яких відобразити а) кількість клієнтів з підключеною голосовою поштою та без; б) мінімальну кількість хвилин в вечірній час для клієнтів з підключеною голосовою поштою і без; в) середню оплату міжнародних дзвінків для клієнтів з підключеним роумінгом та без і з врахуванням лояльності.
2. Побудувати гістограму хвилин в нічний час, загальну і в залежності від підключеного роумінгу.
3. Побудувати діаграму розмаху дзвінків в денний час (загальну і в залежності від підключеної голосової пошти), визначити чи присутні викиди.
4. За допомогою діаграм розсіювання зробити висновки щодо залежності між а) міжнародними хвилинами і дзвінками; б) міжнародними хвилинами і оплатою. Порахувати коефіцієнт кореляції за допомогою відповідних функцій.