Практична робота №2

Візуальний аналіз медичних показників

Мета роботи: Опанувати методи дескриптивної статистики, виявлення закономірностей та візуалізації даних для формування гіпотез на етапі *Data Understanding*.

Стек технологій:

Python / Pandas *для обробки структурованих даних.*

Matplotlib / Seaborn для створення візуалізацій.

Plotly *Інтерактивна візуалізація даних.*

Зміст роботи

Завдання 1. Описова статистика та розподіл (Descriptive statistics)

У датасеті 500 Cities CDC містяться агреговані медичні та соціальні показники для міст США. Необхідно провести описовий та візуальний аналіз обраних змінних і сформулювати аналітичні висновки.

1. Обчисліть основні статистичні показники (середнє, медіана, мода, стандартне відхилення, ексцес та асиметрія) для *ACCESS2_CrudePrev* та *ARTHRITIS_CrudePrev*. Показники є агрегованими оцінками поширеності, що необхідно враховувати при інтерпретації результатів.
2. Побудуйте гістограму з накладеною кривою щільності (*KDE*) для рівня артриту. Визначте, чи є розподіл нормальним? Якщо ні, в який бік зміщена асиметрія?
3. Використовуйте *sns.boxplot* для виявлення аномальних міст за чисельністю населення.
4. Побудуйте гістограми показників *Obesity* та *Smoking* та опишіть форму розподілу.
5. Побудуйте діаграму розсіювання для Obesity та Diabetes і зробіть висновок.

Завдання 2. Порівняльний аналіз (Bivariate Analysis)

Дослідження взаємозв'язків між двома змінними.

1. Побудуйте *jointplot* між рівнем відсутності страхування (*ACCESS2*) та рівнем артриту (*ARTHRITIS*).

2. Додайте лінію регресії, щоб візуалізувати тренд.

3. Побудуйте *violinplot* (скрипковий графік), щоб порівняти розподіл артриту в розрізі категорій City_Size (створених у ПР №1). У містах якого розміру спостерігається найбільша варіативність захворюваності? Змінна City_Size є похідною категоріальною ознакою, створеною на основі Population2010 шляхом бінінгу (малі / середні / великі міста).

4. Побудуйте *boxplot* для показників *ожиріння* по штатах і *гіпертонія* в різних містах за штатами та зробіть висновки.

5. Визначте 5 міст з найвищим показником *Smoking* (міста з найвищим рівнем куріння і міста з мінімальною фізичною активністю).

6. Побудуйте графік довірчих інтервалів для Obesity у вибраних містах.

Завдання 3. Багатовимірний аналіз та кореляції (Multivariate Analysis)

Пошук прихованих структур у всьому датасеті.

1. Побудуйте кореляційну матрицю для всіх числових показників, включаючи *Population2010* та створені раніше індекси.

2. Візуалізуйте її за допомогою *sns.heatmap* з кольоровою схемою *coolwarm*.

3. Створіть *pairplot* для підмножини найбільш корельованих ознак, розфарбувавши точки за ознакою приналежності до великих штатів (наприклад, CA, TX, NY, FL).

Завдання 4. Географічний та інтерактивний аналіз (Plotly)

Оскільки датасет містить дані по містах США, важливо побачити географічний контекст.

1. Використовуючи *plotly.express*, побудуйте інтерактивну стовпчикову діаграму 20 міст з найгіршими показниками доступу до медицини.

2. Налаштуйте *hover_data*, щоб при наведенні відображалися назва штату та точне значення 95% довірчого інтервалу.

3. Побудуйте карту поширеності *Diabetes* та опишіть регіональні відмінності.

4. Сформулюйте аналітичні висновки: які фактори (населення, географія чи доступ до лікаря) найсильніше корелюють із хронічними захворюваннями в цьому наборі даних.

Методичні рекомендації

Датасет:

Розшифровка полів датасету 500 Cities CDC

Поле	Тип даних	Одиниці виміру	Опис / Інтерпретація
StateAbbr	категоріальний (string)	—	Двобуквена абревіатура штату США (CA, TX, NY тощо). Використовується для регіонального аналізу.
PlaceName	категоріальний (string)	—	Назва міста або населеного пункту (місто + штат). Основна одиниця аналізу.
PlaceFIPS	код (string / int)	—	Унікальний FIPS-код міста для геопросторової ідентифікації.
Population2010	числовий (int)	осіб	Чисельність населення міста за переписом 2010 року.
ACCESS2_CrudePrev	числовий (float)	%	Частка дорослого населення без медичного страхування (сирий показник).
ACCESS2_Crude95CI	інтервальний (string)	%	95% довірчий інтервал для ACCESS2_CrudePrev.
ACCESS2_AdjPrev	числовий (float)	%	Віково-стандартизований рівень відсутності медичного страхування.
ACCESS2_Adj95CI	інтервальний (string)	%	95% довірчий інтервал для ACCESS2_AdjPrev.
ARTHRITIS_CrudePrev	числовий (float)	%	Поширеність діагностованого артриту (Crude).
ARTHRITIS_Crude95CI	інтервальний (string)	%	95% CI для сирого рівня артриту.
ARTHRITIS_AdjPrev	числовий (float)	%	Віково-стандартизована поширеність артриту.
ARTHRITIS_Adj95CI	інтервальний (string)	%	95% CI для віково-стандартизованого артриту.
BINGE_CrudePrev	числовий (float)	%	Частка осіб з епізодичним надмірним вживанням алкоголю.

BINGE_Crude95CI	інтервальний (string)	%	95% CI для BINGE_CrudePrev.
BINGE_AdjPrev	числовий (float)	%	Віково-стандартизований рівень binge drinking.
BINGE_Adj95CI	інтервальний (string)	%	95% CI для BINGE_AdjPrev.
BPHIGH_CrudePrev	числовий (float)	%	Поширеність підвищеного артеріального тиску.
BPHIGH_Crude95CI	інтервальний (string)	%	95% CI для гіпертонії (Crude).
BPHIGH_AdjPrev	числовий (float)	%	Віково-стандартизована поширеність гіпертонії.
BPHIGH_Adj95CI	інтервальний (string)	%	95% CI для BPHIGH_AdjPrev.
BPMED_CrudePrev	числовий (float)	%	Частка осіб з гіпертонією, які приймають ліки.
BPMED_Crude95CI	інтервальний (string)	%	95% CI для прийому ліків від гіпертонії.
BPMED_AdjPrev	числовий (float)	%	Віково-стандартизований показник BPMED.
BPMED_Adj95CI	інтервальний (string)	%	95% CI для BPMED_AdjPrev.
DIABETES_CrudePrev	числовий (float)	%	Поширеність діабету серед дорослого населення.
DIABETES_Crude95CI	інтервальний (string)	%	95% CI для сирого рівня діабету.
DIABETES_AdjPrev	числовий (float)	%	Віково-стандартизована поширеність діабету.
DIABETES_Adj95CI	інтервальний (string)	%	95% CI для DIABETES_AdjPrev.
OBESITY_CrudePrev	числовий (float)	%	Поширеність ожиріння (BMI ≥ 30).
OBESITY_Crude95CI	інтервальний (string)	%	95% CI для ожиріння (Crude).
OBESITY_AdjPrev	числовий (float)	%	Віково-стандартизована поширеність ожиріння.
OBESITY_Adj95CI	інтервальний (string)	%	95% CI для OBESITY_AdjPrev.
MHLTH_CrudePrev	числовий (float)	%	Частка осіб з поганим ментальним здоров’ям (≥14 днів/міс).
MHLTH_Crude95CI	інтервальний (string)	%	95% CI для MHLTH_CrudePrev.
MHLTH_AdjPrev	числовий (float)	%	Віково-стандартизований показник поганого ментального здоров’я.
MHLTH_Adj95CI	інтервальний (string)	%	95% CI для MHLTH_AdjPrev.
Geolocation	просторовий (string)	lat, lon	Географічні координати міста.
Uninsured_Population_Count	числовий (int)	осіб	Абсолютна кількість мешканців без медичного страхування.
City_Size	категоріальний	—	Категорія розміру міста,

	(створений)		сформована на основі Population2010.

Описова статистика та аналіз розподілу

При аналізі показників A*CCESS2_CrudePrev та ARTHRITIS_CrudePrev* необхідно:

– використовувати *df.describe()* для первинного огляду;

– окремо обчислити:

- середнє (mean),
- медіану (median),
- моду (mode),
- стандартне відхилення (std),
- асиметрію (skew),
- ексцес (kurtosis).

📌 *Інтерпретація:*

– велика різниця між середнім і медіаною свідчить про асиметрію;

– позитивна асиметрія характерна для соціально-медичних показників (Наприклад, більшість міст мають рівень артриту в діапазоні 15–25%, невелика кількість міст (переважно міста з високою часткою людей похилого віку або регіони з нижчим доступом до медицини мають значення 30–40% і вище).

Аналіз розподілу

При побудові гістограми з KDE:

– зверніть увагу на форму розподілу (дзвоноподібна чи скошена);

– оцініть симетричність та наявність «довгого хвоста».

Візуальна нормальність ≠ статистична нормальність. Остаточний висновок робиться після тестів (Shapiro–Wilk).

Виявлення викидів

Boxplot для Population2010 дозволяє:

– ідентифікувати мегаполіси як природні, але статистичні викиди;

– зрозуміти, чому масштабування є критично важливим.

Порівняльний (двовимірний) аналіз

Jointplot (ACCESS2 & ARTHRITIS)

– оцініть напрям зв’язку (позитивний/негативний);

– зверніть увагу на щільність точок.

Лінія регресії

– використовуйте її виключно для візуалізації тренду**,** а не як доказ причинно-наслідкового зв’язку.

Violinplot

При інтерпретації:

– ширина «скрипки» відображає щільність значень;

– більша ширина → більша варіативність;

– порівняйте стабільність показників між малими та великими містами.

Багатовимірний аналіз і кореляції

Кореляційна матриця

- аналізуйте лише числові змінні;
- пам’ятайте: кореляція ≠ причинність.

Heatmap(кольорова гама coolwarm)

- червоні зони → сильна позитивна кореляція;
- сині → негативна;
- білі/світлі → слабкий або відсутній зв’язок.

Pairplot

- дозволяє виявити нелінійні залежності;
- розфарбування за штатами допомагає побачити регіональні ефекти.

Географічний та інтерактивний аналіз

Інтерактивна стовпчикова діаграма

- ранжування міст дозволяє швидко ідентифікувати проблемні регіони;
- інтерактивність покращує аналітичне сприйняття.

Hover-дані

- використовуйте довірчі інтервали для демонстрації статистичної невизначеності це особливо важливо для медичних даних.

Довірчий інтервал - це діапазон значень, у якому з певною ймовірністю (рівнем довіри) знаходиться істинне значення параметра генеральної сукупності.

У контексті медико-соціальних даних CDC довірчі інтервали дозволяють:

– кількісно оцінити невизначеність оцінок;

– уникнути помилкових категоричних висновків;

– порівнювати регіони, міста або групи з урахуванням статистичної похибки;

– відрізняти реальні відмінності від випадкових флуктуацій.

📌 *Інтерпретація:*

Якщо побудувати 95% довірчі інтервали для великої кількості випадкових вибірок, то приблизно у 95% випадків істинне значення параметра потрапить у відповідний інтервал.

Візуалізація довірчих інтервалів:

– error bars (стовпчики з «вусами»);

– shading (затінені області);

– hover-дані в інтерактивних графіках (Plotly).

Приклад. Середній рівень артриту в місті Mean = 23.4% , 95% CI = [21.8%; 25.0%]

📌 *Інтерпретація:*

– істинний рівень артриту ймовірно лежить у цьому діапазоні;

– якщо довірчий інтервал двох міст перекриваються, відмінність може бути статистично незначущою;

– вузький довірчий інтервал → висока надійність оцінки.

2. *Аналітичні висновки*

У висновках необхідно:

- відокремлювати факти від припущень;
- формулювати гіпотези для подальшої статистичної перевірки.

Типові помилки:

– механічне використання графіків без інтерпретації;

– ігнорування асиметрії розподілу;

– ототожнення кореляції з причинністю;

– відсутність текстових висновків.

Контрольні питання

1. Чим *StandardScaler* відрізняється від *MinMaxScaler* і як вони впливають на візуалізацію розподілу?
2. Яка різниця між кореляцією Пірсона та Спірмена, і яку краще використовувати для медичних показників?
3. Чому важливо перевіряти дані на асиметрію (skewness) перед початком моделювання?
4. Чому медичні дані часто не мають нормального розподілу?
5. Як асиметрія впливає на вибір статистичного тесту?
6. Чому великі міста часто вважаються статистичними викидами?
7. Які переваги *KDE* порівняно зі звичайною гістограмою?
8. Чому *violinplot* інформативніший за *boxplot* у деяких випадках?
9. Що таке мультиколінеарність і як її побачити на *heatmap*?
10. Які ризики виникають при аналізі агрегованих медичних даних?
11. Чому інтерактивна візуалізація є важливою для *Data Understanding*?