МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ  
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА ПОЛІТЕХНІКА»

Кафедра «Систем штучного інтелекту»



**Лабораторна робота №3**

**З дисципліни:**

«Візуалізація даних»

**Виконала:**

студент групи КН-310

Бурак Марко

**Викладач:**

Бойко Н.І.

Львів2020

Лабораторна робота № 3

Хід роботи

1.Датасети: Анскомбе та diamonds

Це датасет Anscombe

Код програми:  
anscombe = sns.load\_dataset('anscombe')

anscombe.head()

Переглянемо вміст

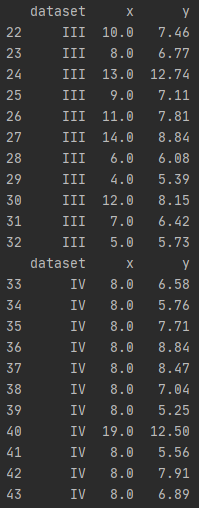
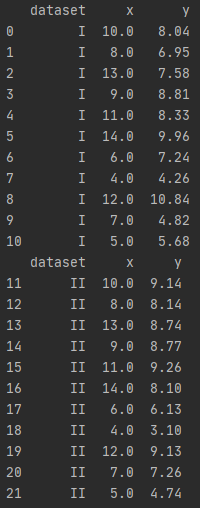


рис 1. Вміст даних у датасеті

Тут зображено датасет Квартет Анскомбе, який складається з чотирьох послідовностей з ідентичними значеннями простих статистичних властивостей, проте графіки цих послідовностей суттєво відрізняються один від іншого.

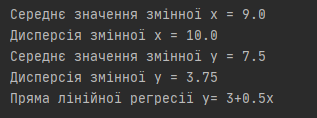


рис 2. Характеристики датасету

Як бачимо з рис2 датасет Анскомбе має одні і ті ж характеристики для кожної послідовності.

Середнє значення у кожній послідовності 9 та 3.75 по х і у відповідно.

Тобто значення коливаються навколо 9 по х та 3.75 по у.

Дисперсія х = 10, а y = 3.75, дисперсія відповідає, насамперед за розсіювання елементів у розподілі, чим більша дисперсія тим більше розсіювання. Отже, х розподіл є більш розсіяний за у.

Також я знайшов, пряму лінійної регресії, y= 3+0.5x, з цієї прямої видно, що кутовий коеіцієнт рівний ½, а вільний член,тобто коефіцієн б рівний 3, це означає, що лінія нахилена до осі Ох, та піднята на 3 пункти по осі ОУ.

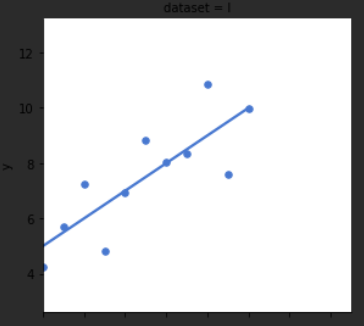
2. Звізуалізуємо для подальшого аналізу наш квартер

def plot\_anscombe(dataset):

sns.lmplot(x="x", y="y", col="dataset", hue="dataset", data=dataset,

col\_wrap=2, ci=None, palette="muted", height=4,

scatter\_kws={"s": 30, "alpha": 1})

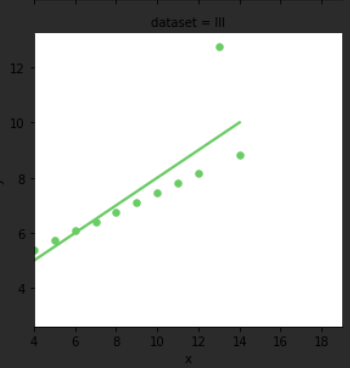
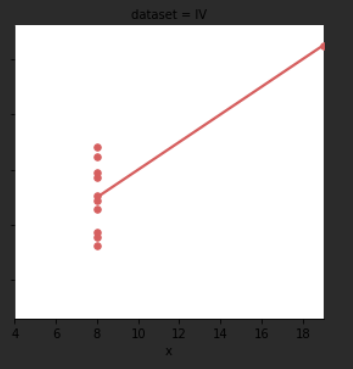
 

рис 3. Візуалізація всіх квартетів

На рис 3 зображено графіки послідовностей.

Можна спостерігати різну поведінку графіків.

На першому графіку бачимо хаотичне розсташування точок.

На другому бачимо точки розсташовано у вигляді дуги.

На третьому графіку видно розсташування точок у вигляді прямої, проте одна з точок віддалена.

Четвертий ж графік зображено у вигляді прямої, яка йде по осі ОУ при х =8, з одною віддаленою точкою у х=19.

3. Знайдемо кореляцію 3 наборів

correlation('I')

correlation('II')

correlation('III')

correlation('IV')



рис 4. Кореляція Пірсона для кожного з 4 квартетів

Не зважаючи, на різне розсташування точок, коефіцієнт кореляції однаковий для кожного графа.

Так стається, через віддалені точки,для прикладу на 3 графіку,точки розміщені неподалік від регресії, проте, коефіцієнт Пірсона такий ж як і у першому хаотичному графіку, все черз точку, яка відповідає y=13 х=13.

Також цей коефіцієнт вказує на те, що значення Y лінійної регресії будуть зростати відносно зростання Х.

4. Знайдемо залишки

y1, model\_y1\_predict = linear\_regression(anscombe, 'I')

residuals1 = y1 - model\_y1\_predict

перший елемент з квартету

другогий елемент з квартету



третій елемент з квартету



четвертий елемент з квартету  


Залишки, це відрізки, які проходять між точкою, та лінійною регресією, чим більші залишки, тим більша буде кореляція відповідно.

Прикладом великого залишку, є віддалена точка з 3 квартету. У масиві залишків, видно, що лише одне значення має довжину більшою за 3, усі решта 1 або менше, цей якраз відрізок вказує на віддалену точку, в якої великий залишок.

5. Побудую гістограму для залишків

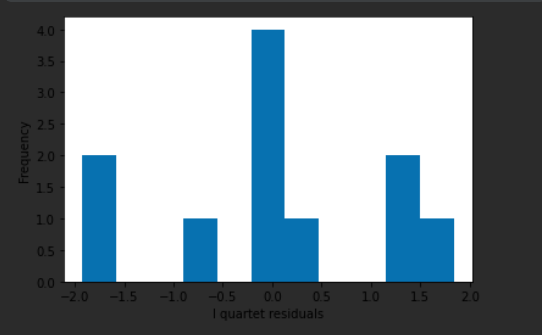


Рис 4. Розподіл залишків для 1 квартету

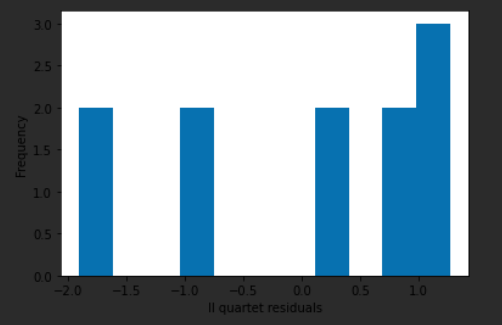


рис 5. Роздподіл залишків для 2 квартету

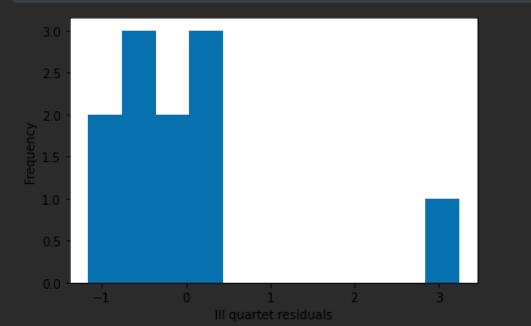


рис 6. Розподіл залишків для 3 квартету

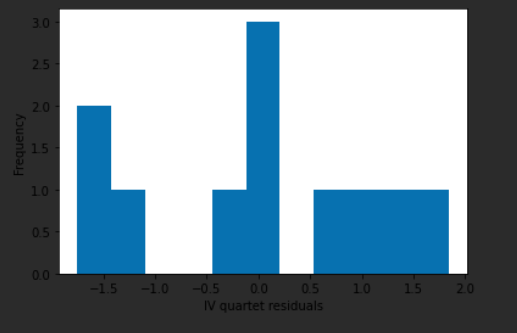


рис 7. Розподіл залишків для 4 квартету

На цих графіках зображено гістограми залишків, тут можна бачити, як розкидані залишки кожного квартету. Знову ж таки, дуже хорошим є приклад 3 квартету для аналізу, адже всі точки мають відносно малі залишки, а віддалена має велику довжину. Тому ж вона стоїть так далеко на гістограмі.

6. Оцінка нормальності розподілу 1(qqplot)

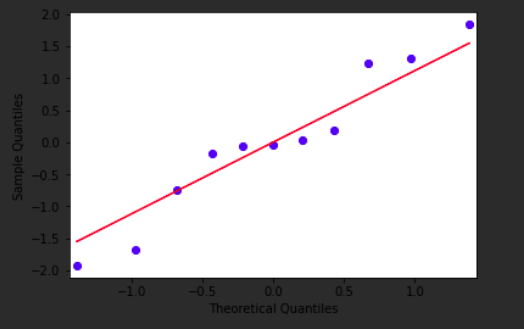


рис 8. Перевірка нормальності розподілу 1 з Qqplot

7. Оцінка нормальності розподілу 2

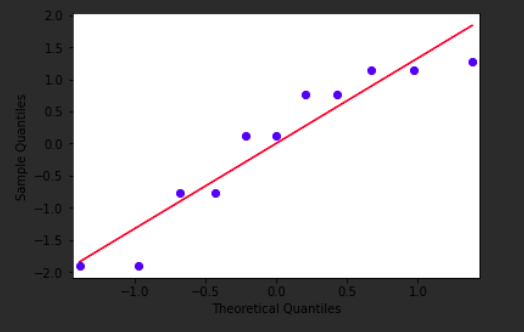


рис 9. Перевірка нормальності розподілу 2 з QQplot

8. Оцінка нормальності розподілу 3

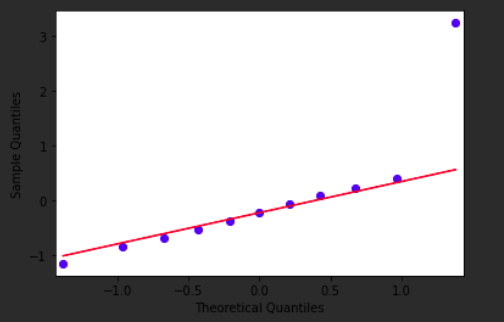


рис 10. Перевірка нормальності розподілу 3 з QQplot

9. оцінка нормальності розподілу 4

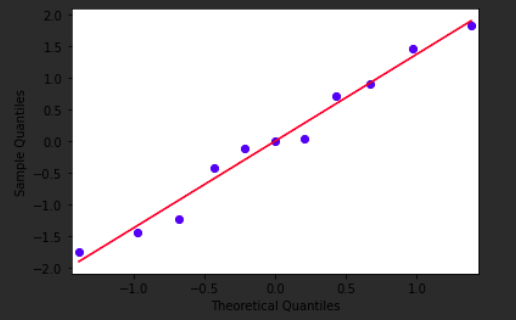


рис 11. Перевірка нормальності розподілу 4 з QQplot

відхилення є у 3, адже, якщо б всі точки лежали на прямій то це був би нормальний розподіл, тут спостерігаємо, що одна точка випадає і робить цей графік менш схожим до нормального.

Інші ж графіки, теж далекі від нормального розподілу, адже жоден з розподілів не лежить на прямій.

10. Оцінюємо варіативність залишків 1, по суті просто малюємо скеттер плот плот для кожних залишків

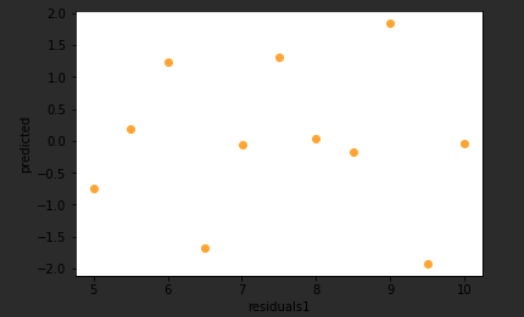


рис 12. Варіантивність залишків 1

11. Оцінюємо варіативність залишків 2

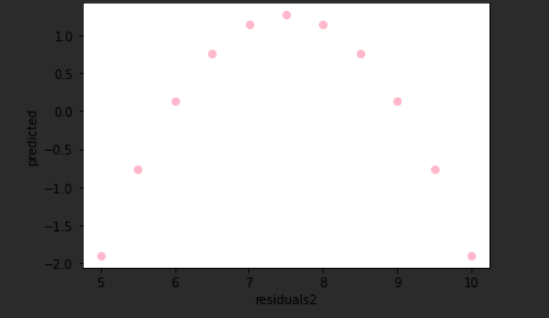


рис 13. Варіантивінсть залишків 2

12. Оцінюємо варіатиність залишків 3

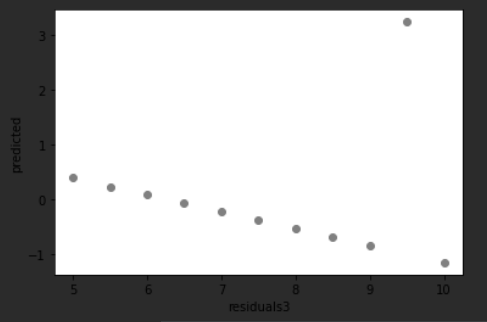


рис 14. Варіативність залишків 3

13. Оцінюємо варіативність залишків 4

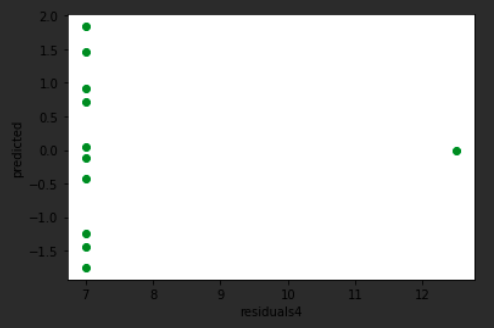


рис 15. варіативність залишків 4

На цих графіках, ми вирішили побудувати варіативність залишків, перевіряємо,наскільки точки віддалені один від одного. Якщо значення віддалене у розподілі від попереднього то можна зробити висновок, що дані ненормалізовані.

Кожен з графіків, має хоча б одну точку, яка знаходиться далеко від попередніх, одже можна зробити висновок, що вони всі ненормалізовані.

14. Побудую графік карат до ціни

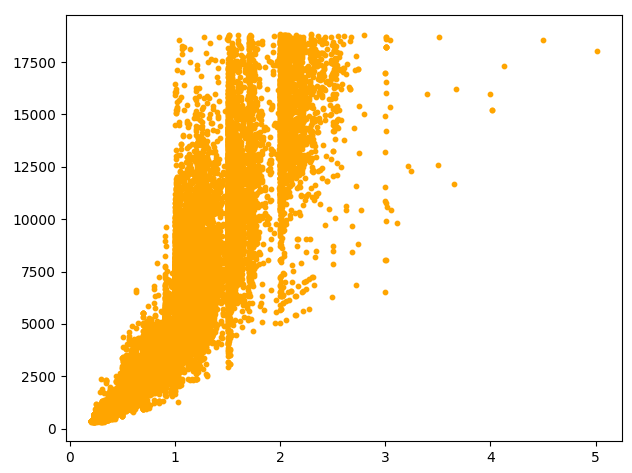


рис 17. графік розсіювання карат до ціни

На цьому графіку зображено розподіл відносно ціни до капат за діамант. На графіку видно ,схожу , до лінійного зростання залежність для діамантів від 0 до 1 карат. У цьому відрізку, видно, що чим більше має діамант карат, тим, здебільшого збільшується його ціна. Проте з відрізка 1 і більше, графік поводиться більш хаотично, можемо спостерігати, що ціна за діамант 1 карату може бути такою ж як і діамант 5 карат, або ж навіть меншою. Починаючи з 1 карату і більше,ціни за карат влаштовані хаотично, проте, зберігають залежність.

16. Знайдемо Коефіцієнт кореляції

print(diamonds['carat'].corr(diamonds['price']).round(4))

0.9216,отже лінійний графік лінійної регресії буде зростатиY при зростанні X

При кореляції 1 , точки лежать на одній прямій лінійною регресією. При 0, точки хаотичні. У нашому ж випадку, коефіцієнт кореляції 0.9, з чого можна зробити висновок, що точки на графіку не є сильно відхилені від лінійної регресії.

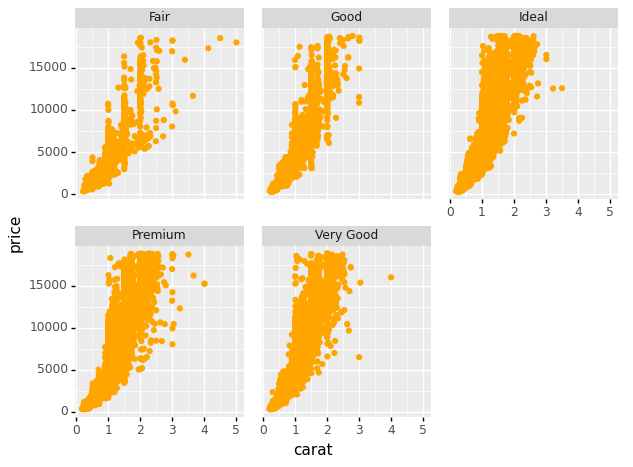
17.Давайте подивимось, як розподілені вага та ціна в залежності від ступеня обробки діамантів: 

рис 18 розподіл ваги та ціни в залежності від ступеня обробки

Як ми знаємо з датасету, буває 5 ступенів обробки діамантів, Fair , Good, Ideal, Premium та very good. З вигляду видно, що при обробці Fair можна купити діаманти більшої каратності за ту ж ціну, яка є у діамантів кращої обробки меншої ваги.

Усі розподіли мають схожий розподіл від 0 до 1 карату. Проте графік fair

ступеню обробки суттєво відрізняється. Він зростає майже лінійно відносно ціни та ваги, коли інші ростуть майже по експоненті.

Інші ступені крім Fair зростають у ціні до ваги швидше, аніж діаманти fair.

Щоб переконатися у правильності наших припущень, побудуємо графіки лінійної регресії.

18. Додамо до графіка ще лінію лінійної регресії

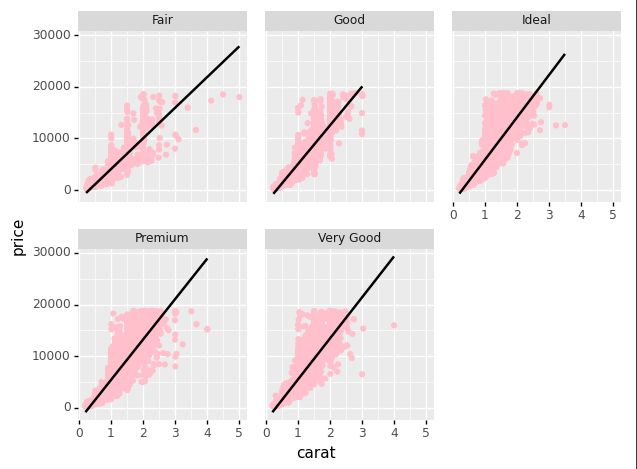


рис 19. розподіл і лінійна регресія

З цих графіків і дійсно видно, що лінійна регресія ступеня fair проходить приблизно під кутом 45 градусів, коли всі інші регресії похилені до осі ОУ, що підтверджує припущення з минулого пункту.

19. Побудуємо моделі лінійної регресії lin.diamond.idealта lin.diamond.fair для обробки Ideal Fair ціни від ваги.

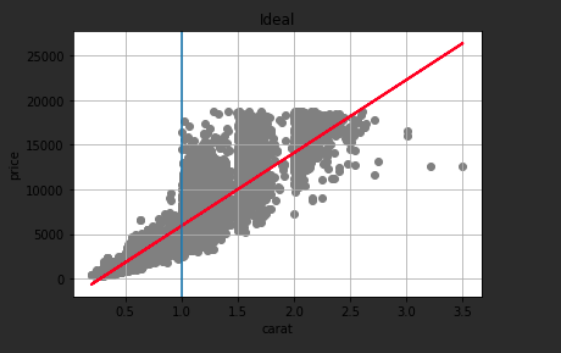


Рис 20 Для обробки ‘Ideal’

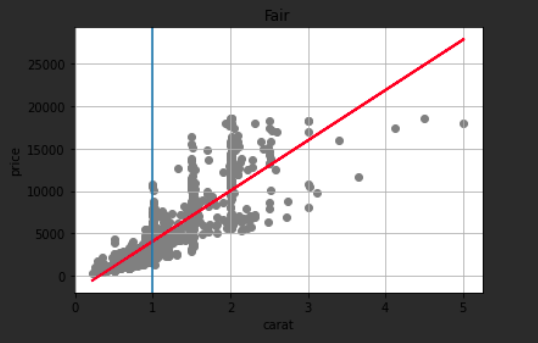


рис 21. для обробки Fair

З графіків вище можна дізнатися ціну за діаманти різних обробок, використовуючи лінійну регресію.

Після апробації програми ціна за один карат обробки idial була 5892.

В той же час, ціна за карат обробки fair склала 4085, що є суттєво менше

**Висновок**: на цій лабораторній роботі навчились працювати з лінійною регресією також навчилися аналізувати дані датасету. Зрозуміли сутність кореляції, та від яких факторів вона залежить. Зрозуміли, що таке залишки. Навчилися аналізувати дані реальних продажей діамантів.