

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І
ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ**

**ОБРОБКА ІНФОРМАЦІЇ В КОМП'ЮТЕРНО-ІНТЕГРОВАНИХ
СИСТЕМАХ АВТОМАТИЗАЦІЇ**

**Методичні вказівки
до проведення практичних занять**
для здобувачів третього (освітньо-наукового) рівня вищої освіти за
спеціальністю 151 «Автоматизація та комп'ютерно-інтегровані технології»
(кваліфікація: PhD доктор філософії)

Київ-2022

УДК 621.391(075.8)

У методичних вказівках представлені методи обробки інформації в комп'ютерно-інтегрованих системах автоматизації. Практичні роботи призначені для аспірантів, які навчаються за спеціальністю 151 «Автоматизація та комп'ютерно-інтегровані технології», галузі знань 15 «Автоматизація та приладобудування» (кваліфікація: PhD доктор філософії), а також може бути корисним докторантам, які проводять дослідження в області автоматизації та інформаційних технологій.

Рекомендовано до друку рішенням вченої ради ННІ енергетики, автоматики і енергозбереження Національного університету біоресурсів і природокористування України (протокол № 11 від 17 грудня 2021 р.).

Укладачі: к.т.н., доц. М.О. Кіктєв, д.т.н., проф. В.В. Коваль

Рецензенти: д.т.н., проф. С.А. Шворов, к.т.н., доц. Є.А. Антипов

НАВЧАЛЬНЕ ВИДАННЯ

ОБРОБКА ІНФОРМАЦІЇ В КОМП'ЮТЕРНО-ІНТЕГРОВАНИХ СИСТЕМАХ АВТОМАТИЗАЦІЇ

Методичні вказівки до проведення практичних занять
для здобувачів третього (освітньо-наукового) рівня вищої освіти за спеціальністю 151 «Автоматизація та комп'ютерно-інтегровані технології» (кваліфікація: PhD доктор філософії)

Укладачі: КІКТЄВ Микола Олександрович,
КОВАЛЬ Валерій Вікторович

Зав. Видавничим центром НУБіП А.П. Колесніков

Підписано до друку 17.10.25
Ум. друк. арк. 5,9
Наклад 20 прим.

Формат 60x84/16
Обл. вид. арк. 5,2
Зам. № 2915

Видавничий центр НУБіП
03041, вул. Героїв Оборони, 15

ЗМІСТ

Практична робота № 1. Збір та підготовка даних.....	4
Практична робота № 2. Регресійний аналіз.....	8
Практична робота № 3. Бінарна класифікація.....	11
Практична робота № 4. Кластерний аналіз.....	21
Практична робота № 5. Порівняння і аналіз двох вибірок з використанням критеріїв Стюдента і хи-квадрат.....	30
Практична робота № 6. Робота в середовищі мови R.....	32
Практична робота № 7. Статистична обробка даних на мові R...	36
СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ.....	42

Практична робота № 1

«Збір та підготовка даних»

Загальні відомості

Цілями роботи є:

- ознайомлення зі структурою джерел відкритих даних, вивчення способів зберігання і представлення даних;
- придбання навички побудови системи збору даних.

Завдання 1. Дослідження наборів даних, представлених на порталі відкритих даних ukrstat.org

1. Держава
- 2 Економіка
- 3 Освіта
- 4 Здоров'я
- 5 Екологія
- 6 Транспорт
- 7 Культура
- 8 Спорт
- 9 Будівництво
- 10 Дозвілля та відпочинок
- 11 Торгівля
- 12 Туризм
- 13 Електроніка
- 14 Картографія
- 15 Безпека
- 16 Метеодані

Завдання 2. Дослідження наборів даних, представлених на порталі data.worldbank.org.

1Agriculture & Rural Development

- 2 Aid Effectiveness
- 3 Climate Change
- 4 Economy & Growth
- 5 Education
- 6 Energy & Mining
- 7 Environment 8 External Debt
- 9 Financial Sector
- 10 Gender
- 11 Health
- 12 Infrastructure
- 13 Poverty
- 14 Private Sector
- 15 Public Sector
- 16 Science & Technology
- 17 Social Development
- 18 Social Protection & Labor
- 19 Trade
- 20 Urban Development

Завдання 3. Побудова автоматизованої системи збору даних.

Як інструментальний засіб використовується програмне забезпечення Microsoft Excel.

Варіанти завдання

- 1 Онлайн-табло будь-якого аеропорту / вокзалу
- 2 Котирування акцій / валют / дорогоцінних металів / корисних копалин на будь-якої біржі
- 3 За пропозицією студента

Порядок виконання

- 1 Дослідження наборів даних на порталі **ukrstat.org**:

1.1 Виберіть варіант завдання.

1.2 Знайдіть довільний набір даних на порталі **ukrstat.org** по тематиці, зазначеної в обраному варіанті завдання. набір повинен бути представлений в форматі csv і кодуванні Windows.

1.3 Завантажте на комп'ютер знайдений набір даних

1.4 Проведіть аналіз набору даних: визначте кількість записів і полів в наборі даних.

2 Дослідження наборів даних на порталі **data.worldbank.org**:

2.1 Виберіть варіант завдання.

2.2 Знайдіть довільний набір даних на порталі **data.worldbank.org** по тематиці, зазначеної в обраному варіанті завдання.

2.3 Завантажте на комп'ютер знайдений набір даних в форматі XLS.

2.4 На основі набору даних підготуйте вибірку, що містить значення показника за всі роки для трьох довільно вибраних країн світу.

2.5 На основі підготовленої вибірки побудуйте графік, який ілюструє зміна показника з часом для трьох країн світу.

2.6 Збережіть файл.

3 Побудова автоматизованої системи збору даних:

3.1 Виберіть варіант завдання.

3.2 Знайдіть інтернет-сайт, що містить зазначені в завданні дані.

3.3 Запустіть Microsoft Excel.

3.4 Виберіть пункт «З Інтернету» в меню «Дані».

3.5 У адресному рядку вікна, що з'явилося «Створення веб-запиту» наберіть адресу знайденої раніше веб-сторінки.

3.6 Виберіть таблицю, яка містить дані, які розшуковуються.

3.7 Натисніть кнопку «Імпорт».

3.8 У вікні «Імпорт даних» натисніть кнопку «Властивості».

3.9 У вікні «Властивості зовнішнього діапазону» задайте параметр «Період оновлення», рівний 1 хвилині, параметр «Оновлення при відкритті файлу» - «Так».

3.10 Натисніть кнопку «ОК».

3.11 У вікні «Імпорт даних» натисніть кнопку «ОК».

3.12 Збережіть файл.

4 Звіт про роботу:

4.1 Складіть звіт про роботу.

4.2 Перетворіть звіт в формат PDF.

4.3 Запакуйте звіт (PDF) і всі використані і створені

в роботі файли в архів формату ZIP.

4.4 Прикріпіть архів в розділ «Звіт з практичної роботи №1 (Збір та підготовка даних)» курсу «Аналіз даних» [2].

Звіт повинен містити:

1 Титульний аркуш: найменування роботи, варіант завдання, ПІБ студента, номер навчальної групи, дата виконання роботи.

2 Реферат.

3 Зміст.

4 Частина 1 «Дослідження наборів даних на порталі ukrstat.org»:

4.1 Завдання.

4.2 Копія екрану з набором даних, відкритому в Microsoft Excel.

4.3 Опис набору даних відповідно до наведеної нижче формі

Таблиця . Форма опису набору даних

Найменування	Значення
Посилання	
формат	
Кількість записів	
кількість полів	
в т.ч. числових	
в т.ч. текстових	

Практична робота № 2.

Регресійний аналіз

Загальні відомості.

Метою роботи є придбання навички регресійного аналізу. Як інструментальний засіб використовується програмне забезпечення Microsoft Excel.

Вхідні данні

Таблиця 1. Варіанти завдання по регресійному аналізу

Варіант	Сфера	Дані	Залежність
	Пасажирські авіаперевезення	Дальність і час перельоту між різними містами	Час від дальності
2	Пасажирські авіаперевезення	Дальність і вартість перельоту між різними містами економічним класом	Вартість від дальності
3	Пасажирські авіаперевезення	Дальність і вартість перельоту між різними містами бізнес-класу	Вартість від дальності
4	Пасажирські залізничні перевезення	Дальність і час поїздки між різними містами	Час від дальності
5	Пасажирські залізничні перевезення	Дальність і вартість поїздки між різними містами в купе	Вартість від дальності
6	Пасажирські залізничні перевезення	Дальність і вартість поїздки між різними містами в плацкарті	Вартість від дальності
7	Ринок нерухомості	Площа і вартість квартир на первинному ринку	Вартість від площі
8	Ринок нерухомості	Площа і вартість квартир на вторинному ринку	Вартість від площі
9	Ринок автотранспорту	Вартість і пробіг автомобілів будь-якої марки на вторинному ринку	Вартість від пробігу
10	Ринок автотранспорту	Вартість і вік автомобілів будь-якої марки на вторинному ринку	Вартість від віку
11	Ринок	Вік і пробіг автомобілів	Пробіг від

	автотранспорту	будь-якої марки на вторинному ринку	віку
12	Світова економіка	Тривалість життя і доходи на душу населення країн світу	Тривалість життя від доходів

Таблиця 2. Перевірочні дані

Варіант	Дані
1	<u>Час польоту на 500, 1000 та 3000 км</u>
2	<u>Вартість перельоту на 500, 1000 та 3000 км</u>
3	<u>Вартість перельоту на 500, 1000 та 3000 км</u>
4	<u>Час поїздки на 400, 800 та 2000 км</u>
5	<u>Вартість поїздки на 400, 800 та 2000 км</u>
6	<u>Вартість поїздки на 400, 800 та 2000 км</u>
7	<u>Вартість для площі 30, 50, 100 кв.м.</u>
8	<u>Вартість для площі 30, 50, 100 кв.м.</u>
9	<u>Вартість для пробігу 20 тис., 50 тис., 150 тис. км.</u>
10	<u>Вартість для віку 2 роки, 5 р, 10 р</u>
11	<u>Пробіг для віку 2 роки, 5 р, 10 р</u>
12	<u>Тривалість життя для доходів 5, 20, 50 тис. \$</u>

Порядок виконання

1 Підготовка до роботи:

1.1 Виберіть варіант завдання.

1.2 Знайдіть джерело даних відповідно до завдання.

1.3 Запустіть Microsoft Excel.

1.4 Створіть лист «Вихідні дані» в документі Excel.

1.5 Підготуйте і розмістіть в листі «Вихідні дані» вибірку даних відповідно до обраного варіанту завдання. вибірка повинна містити не менше 15 записів.

2 Побудова лінійної регресії аналітичним методом:

2.1 Створіть лист «Аналітичне рішення».

2.2 Скопіюйте вибірку даних з листа «Вихідні дані» на лист «Аналітичне рішення».

2.3 Виконайте пошук параметрів функції регресії за допомогою нормального рівняння.

2.4 Побудувати на одному графіку вихідні дані та графік функції регресії.

2.5 Створіть прогноз. Як аргумент використовуйте перевірочні дані .

3 Побудова лінійної регресії чисельним методом:

3.1 Створіть лист «Чисельне рішення».

3.2 Скопіюйте вибірку даних з листа «Вихідні дані» на лист «Чисельне рішення».

3.3 Виконайте пошук параметрів функції регресії за допомогою інструменту «Пошук рішення» ПО Microsoft Excel.

- 3.4 Побудувати на одному графіку вихідні дані та графік функції регресії.
- 3.5 Створіть прогноз. Як аргумент використовуйте перевірочні дані.
- 4 Порівняльний аналіз:
 - 4.1 Порівняйте коефіцієнти рівняння регресії, отримані обома методами.
 - 4.2 Порівняйте прогнози, отримані обома методами.
- 5 Підбір функції регресії.
 - 5.1 Розділіть вихідну вибірку на дві частини: навчальну і перевірочну.
 - 5.2 Побудувати регресію за навчальною частиною вибірки для лінійної, квадратичної та кубічної функцій.
 - 5.3 Зобразіть на одному графіку вихідні дані та графіки трьох функцій регресії.
 - 5.4 Зобразіть на одному графіку залежність функції штрафу для навчальної вибірки і функції штрафу для перевірочної вибірки від ступеня полінома функції гіпотези.
 - 5.5 Виберіть найкращу функцію регресії.
- 6 Звіт про роботу:
 - 6.1 Оформіть звіт згідно з вимогами, наведеними нижче.
 - 6.2 Збережіть звіт в форматі PDF.
 - 6.3 Архівуйте звіт і файли Excel, використані в роботі.
 - 6.4 Прикріпіть архів в розділ «Звіт з практичної роботи №2 «Регресійний аналіз» та завантажити в систему дистанційного навчання університету

Вимоги до звіту

Звіт повинен містити:

- 1 Титульний аркуш: найменування роботи, варіант завдання, ПІБ студента, номер навчальної групи, дата виконання роботи.
- 2 Реферат.
- 3 Зміст.
- 4 Завдання.
- 5 Опис виконаної роботи:
 - 5.1 Рішення завдання регресії аналітичним методом.
 - 5.2 Рішення завдання регресії чисельними методами.
 - 5.3 Підбір оптимальної функції регресії.
- 6 Отримані результати.
- 7 Аналіз результатів.
- 8 Список використаних джерел:
 - 8.1 Джерела даних.
 - 8.2 Нормативні документи.
- 9 Додатки.

Звіт повинен бути оформлений відповідно до чинних вимог університету

Практична робота № 3

«Бінарна класифікація»

Загальні відомості

Метою роботи є придбання навички бінарної класифікації даних на основі логістичної регресії. Як інструментальний засіб використовується програмне забезпечення Microsoft Excel.

Завдання (приклад)

Варіант 1. При перевірці діагностичної системи, заснованої на бінарному класифікаторі, отримані наступні результати:

<https://basegroup.u/system/files/article/loans.txt>

Варіант 2. При випробуванні антивіруса, заснованого на бінарному

класифікаторі, отримані наступні результати:

<https://basegroup.u/system/files/article/diabetes.txt>

Тематику для варіанта вибрати самостійно (по прикладу).

Порядок виконання

1 Підготовка:

1.1 Виберіть варіант завдання

1.2 Підготуйте вибірку даних в ПЗ Microsoft Excel.

1.3 Побудуйте діаграму, що відображає вибірку даних.

2 Класифікація:

2.1 Задайте цільову функцію.

2.2 Визначте коефіцієнти функції гіпотези за допомогою інструменту «Пошук рішення».

2.3 Розрахуйте значення точності, чутливості, F-критерію.

3 Зробіть висновок про ефективність цього класифікатора.

4 Звіт про роботу:

4.1 Складіть звіт про роботу.

4.2 Перетворіть звіт в формат PDF.

4.3 Запакуйте звіт (PDF) і файл з даними (XLS) в один архів формату ZIP.

4.4 Прикріпіть архів в розділ «Звіт з практичної роботи №3 (Бінарна класифікація)» та завантажте його в СДО університету.

Зміст звіту

Звіт повинен містити:

1 Титульний аркуш: найменування роботи, варіант завдання, ПІБ студента, номер навчальної групи, дата виконання роботи.

2 Реферат.

3 Зміст.

4 Завдання.

5 Опис виконаної роботи.

6 Отримані результати.

7 Аналіз результатів.

8 Список використаних джерел:

8.1 Джерела даних.

8.2 Нормативні документи.

9 Додатки.

Звіт повинен бути оформлений відповідно до чинних вимог університету

Приклад виконання роботи

Приклад з лекції

Вводимо вхідні дані:

Доп. стовбец	x1	x2	Y
1	1	6	0
1	3	4	0
1	2	2	1
1	3	3	1
1	4	3	1

Вводимо вектор коефіцієнтів рівняння логістичної регресії β :

Beta0	1
Beta1	1
Beta2	1

Вводимо формули для матричних розрахунків, :
 $\text{=МУМНОЖ}(A3:C3;\$B\$13:\$B\$15)$

Отримаємо значення вектору Z.

Вводимо значення вектору h:

=1/(1+EXP(-E3))

Вводимо значення функції штрафу

$\text{=ЕСЛИ}(D3=0;-\text{LN}(1-F3);-\text{LN}(F3))$

В результаті отримуємо таблицю:

Доп. стовбец	x1	x2	Y	Z	h	cost(h,y)
1	1	6	0	8,00	0,9997	8,000335
1	3	4	0	8,00	0,9997	8,000335
1	2	2	1	5,00	0,9933	0,006715
1	3	3	1	7,00	0,9991	0,000911
1	4	3	1	8,00	0,9997	0,000335

Сумма **16,008633**

Beta0	1
Beta1	1
Beta2	1

Після застосування інструменту «Пошук рішень» в результаті отримуємо оптимальні значення коефіцієнтів регресії:

Доп. стовбец	x1	x2	Y	Z	h	cost(h,y)
1	1	6	0	-0,14	0,4656	0,626656
1	3	4	0	0,49	0,6189	0,964791
1	2	2	1	1,07	0,7450	0,294438
1	3	3	1	0,78	0,6866	0,375951
1	4	3	1	0,80	0,6892	0,372212

Сумма **2,634048**

Beta0	1,6467
Beta1	0,012
Beta2	-0,2994

Альтернативою є обчислення коефіцієнтів за допомогою матричних обчислень.

1. Знаходимо транспоновану матрицю до матриці X

X трансп

1	1	1	1	1
1	3	2	3	4
6	4	2	3	3

Умножаємо транспоновану матрицю на вихідну

Xтрансп * X

5	13	18
13	39	43
18	43	74

3. Знаходимо зворотню матрицю до попередньої

Обратная матрица

6,209580838	-1,125748503	-0,85629
-1,125748503	0,275449102	0,113772
-0,856287425	0,113772455	0,155689

4. Знаходимо добуток транспонованої матриці на вектор Y

Xтрансп * Y

3
9
8

5. Знаходимо коефіцієнти рівняння регресії

Вектор оценок коэффициентов регрессии равен

$$(X^T X)^{-1} X^T Y$$

1,646706587
0,011976048
-0,299401198

6. Рівняння регресії:

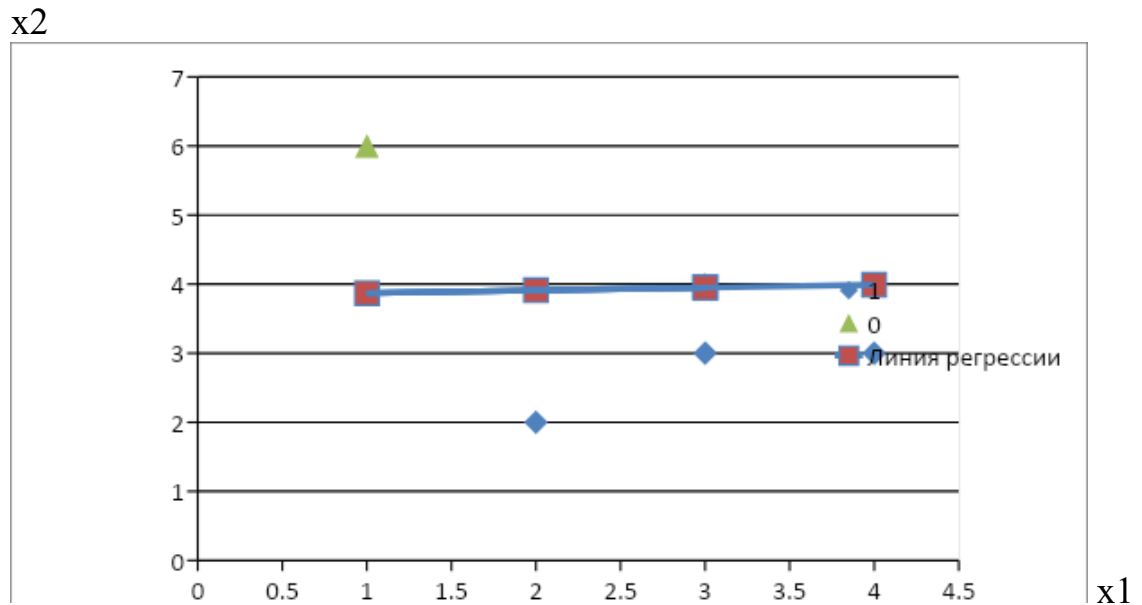
$$Y = 1.6467 + 0.01198X_1 - 0.2994X_2.$$

Таким чином, отримуємо таблицю з заданими та розрахунковими значеннями Y:

x1	x2	Y	Y*	x2=f(x1)
1	6	0	-0,13772	3,870007
3	4	0	0,48504	3,950033
2	2	1	1,07186	3,91002
3	3	1	0,78444	3,950033

Як видно з таблиці, розрахункові значення відносно близько розташовані до класів 0 та 1, тому вважаємо, що задачу вирішено вірно. Для побудови графіку в площині x_1 - x_2 , виразимо x_2 через x_1 :

$$x_2 = (0,5 - x_1 - 0,79642x_1) / 3,990047$$



На графіку відображено лінію розподілу між

Приклад 2. Видача кредитів.

Вихідні дані:

<https://basegroup.ukr/system/files/article/loans.txt>

Возраст	Пол	Состоит в браке	Иждивенцы	Доход	Опыт работы	Срок проживания	Недвижимость	Исключный платеж	Благонадежный заемщик
28	женский	Да	0,0	9000,0	9,0	0,0	3946,0	Нет	
39	мужской	Да	1,0	13500,0	17,0	6,0	2460,0	Да	
31	мужской	Нет	2,0	7000,0	11,0	3,0	3126,0	Нет	
34	мужской	Нет	1,0	10200,0	15,0	2,0	3280,0	Да	
46	женский	Да	2,0	8500,0	20,0	8,0	3348,0	Да	
30	женский	Да	2,0	9500,0	12,0	30,0	4612,0	Нет	
47	мужской	Нет	2,0	7900,0	14,5	6,0	2870,0	Нет	
33	мужской	Нет	2,0	12600,0	15,0	23,0	2050,0	Да	
22	мужской	Нет	0,0	34000,0	4,0	19,0	2562,0	Да	
30	мужской	Да	1,0	33000,0	10,0	8,0	1948,0	Да	
26	мужской	Да	1,0	13500,0	8,0	2,0	5535,0	Да	
48	мужской	Нет	2,0	24000,0	22,0	40,0	1845,0	Да	
37	мужской	Да	1,0	12500,0	15,5	6,0	2409,0	Да	
27	мужской	Да	1,0	4000,0	9,0	6,0	4305,0	Нет	
55	мужской	Да	1,0	5500,0	11,5	35,0	3382,0	Нет	
39	женский	Нет	1,0	8000,0	6,0	8,0	4818,0	Нет	
29	мужской	Да	1,0	11000,0	11,0	29,0	5228,0	Да	
36	мужской	Нет	1,0	11600,0	18,0	13,0	4373,0	Да	
32	мужской	Да	1,0	8500,0	14,0	28,0	2181,0	Да	
35	женский	Нет	0,0	4500,0	16,0	33,0	1913,0	Нет	
46	мужской	Да	1,0	9000,0	16,0	28,0	718,0	Да	
22	мужской	Да	1,0	4000,0	3,0	20,0	6252,0	Нет	
37	мужской	Да	2,0	29000,0	12,0	6,0	4612,0	Да	
56	женский	Да	0,0	29000,0	19,0	36,0	4237,0	Да	
34	мужской	Нет	1,0	26000,0	16,0	5,0	6662,0	Да	
40	мужской	Нет	2,0	10500,0	20,5	2,0	3895,0	Нет	
33	мужской	Нет	1,0	10700,0	15,0	3,0	4612,0	Нет	
37	мужской	Да	2,0	35000,0	17,5	31,0	3485,0	Да	
24	мужской	Да	0,0	11500,0	6,0	8,0	2972,0	Да	
30	мужской	Да	2,0	15500,0	6,0	6,0	1435,0	Да	
32	мужской	Нет	1,0	11300,0	14,0	28,0	6560,0	Нет	
25	мужской	Нет	2,0	12500,0	7,0	17,0	1948,0	Да	
37	мужской	Нет	0,0	18200,0	17,5	36,0	3382,0	Да	
33	женский	Нет	2,0	16500,0	15,0	3,0	3895,0	Да	
24	мужской	Да	1,0	12500,0	6,0	14,0	2181,0	Да	
41	мужской	Да	2,0	29800,0	13,5	28,0	3895,0	Да	
28	женский	Нет	2,0	25000,0	7,0	15,0	4237,0	Да	
35	женский	Нет	3,0	8500,0	12,0	23,0	5842,0	Нет	
32	женский	Нет	0,0	33000,0	14,0	26,0	4920,0	Да	
35	мужской	Нет	0,0	17500,0	6,0	5,0	2562,0	Да	
37	мужской	Да	0,0	17000,0	13,5	6,0	3143,0	Да	

Вводимо вхідні дані:

Возраст	Доход	Иждивенцы	Благонадежный заёмщик	Y
28	9	0	нет	0
39	13,5	1	да	1
31	7	2	нет	0
34	10,2	1	нет	0
46	8,5	2	да	1
30	9,5	2	да	1
47	7,9	2	да	1
33	12,6	2	нет	0
22	34	0	нет	0
30	33	1	да	1

По аналогії з 1-ю задачею обчислимо коефіцієнти регресії та сформулюємо рівняння регресії:

$$Y = -1.7422 + 0.04902X_1 + 0.02671X_2 + 0.1444X_3$$

Додаємо стовбець з розрахованими значеннями Y

Вік	Дохід	Утриманці	Благонадійний позичальник	Y	Y*
28	9	0	нет	0	-0,12925
39	13,5	1	да	1	0,32041
31	7	2	нет	0	0,06622
34	10,2	1	нет	0	0,072046

46	8,5	2	да	1	0,803004
30	9,5	2	да	1	0,019673
47	7,9	2	да	1	0,85143
33	12,6	2	нет	0	0,1698
22	34	0	нет	0	-0,63705
30	33	1	да	1	-0,10148

Для нанесення значень на площину Вік/Дохід впорядкуємо дані за функцією Y (благонадійність).

Вік	Дохід	Утриманці	Благонадійний позичальник
39	13,5	1	да
46	8,5	2	да
30	9,5	2	да
47	7,9	2	да
30	33	1	да
28	9	0	нет
31	7	2	нет
34	10,2	1	нет
33	12,6	2	нет
22	34	0	нет

Для нанесення лінії розподілу між класами, виразимо x_2 від x_1 та x_3 , та x_2 від x_1 (не враховуючи x_3).

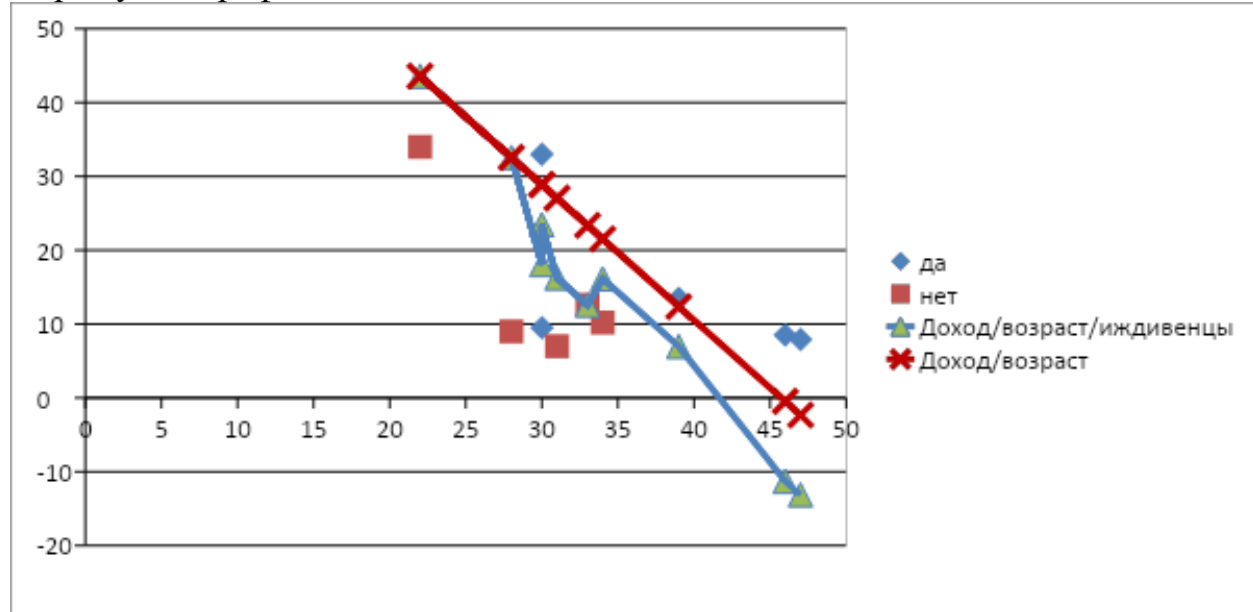
$$=(0,5-\$G\$19-\$G\$20*J32-\$G\$22*K32)/\$G\$21$$

$$=(0,5-\$G\$19-\$G\$20*J32)/\$G\$21$$

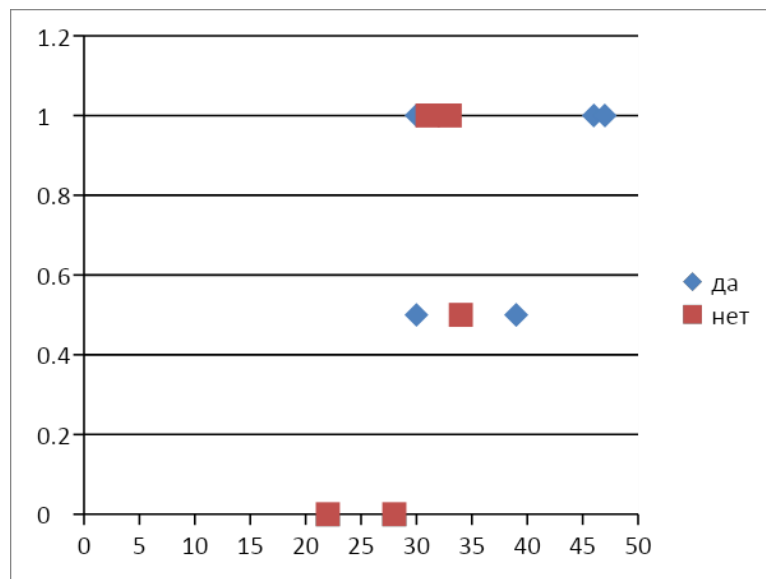
Тепер для побудови графіка впорядкуємо дані за параметром x_1 (вік)

Вік	Утриманці	Дохід	Дохід (без x_3)
22	0	43,5702	43,57019843
28	0	32,55859	32,55859229
30	2	18,07563	28,88805691
30	1	23,48184	28,88805691
31	2	16,24036	27,05278922
33	2	12,56982	23,38225384
34	1	16,14077	21,54698615
39	1	6,964433	12,3706477
46	2	-11,2887	-0,476226133
47	2	-13,1239	-2,311493823

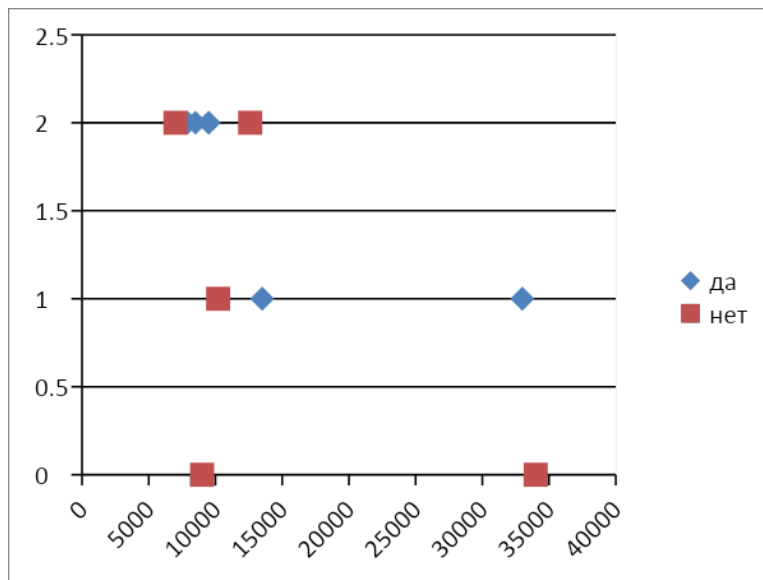
Отримуємо графік.



Аналогічно наносимо на площину в координатах інших пар параметрів:
x1 – вік, x2 – кількість іждивенців



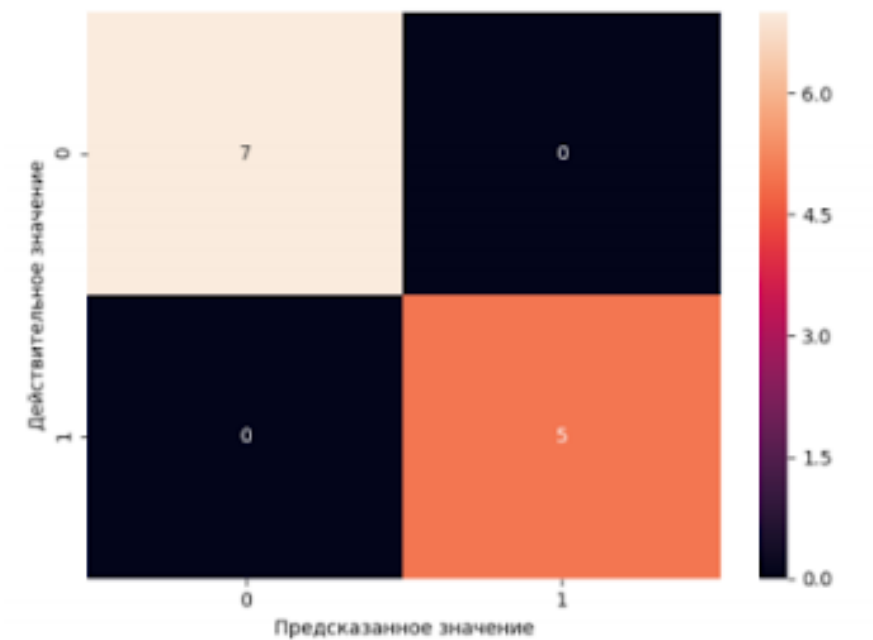
x1 – дохід, x2 – кількість утриманців



Для вирішення задачі данні необхідно поділити на навчальні (70%) та тестові (30%).

Перевірити класифікацію на тестовій вибірці.

Побудувати матрицю помилок (див. лекцію).



Перевірити модель на новому наборі даних

Приклад з діагностикою вірусів

<https://basegroup.ukr/system/files/article/diabetes.txt>

Приклад зі вступом до наступного рівня навчання

Таблица 1. Результат классификации данных тестовой выборки

Независимые переменные			Целевая переменная	
gmat	gra	work_experience	Действительное значение	Предсказываемое значение
550	3,3	4	0	0
620	4,3	2	0	0
670	4,3	6	1	1
680	4,9	4	1	1
610	3,7	3	0	0
610	4,0	1	0	0
650	4,7	6	1	1
690	4,7	5	1	1
540	3,7	2	0	0
660	4,3	5	1	1
580	3,3	2	0	0
650	3,3	1	0	0

У першому масиві будуть перебувати стовпці, дані яких відповідають незалежних змінних (середня оцінка за час навчання на попереднього ступеня освіти, бали іспиту GMAT і стаж роботи), а в другому – стовпець значень.

Практична робота № 4.

Кластерний аналіз.

Мета роботи: навчитися на основі застосування методу **k-means clustering** (k-середніх) вихідну сукупність об'єктів розділити на кластери або групи (класи) схожих між собою об'єктів.

В якості вихідних даних використовуються статистичні дані Світового банку. Як інструментальний засіб для проведення експериментів використовується програмне забезпечення MS Excel, вбудована мова програмування Visual Basic for Application (VBA)

1. Короткі теоретичні відомості

Зазвичай перед початком класифікації дані стандартизуються (віднімається середнє і здійснюється поділ на корінь квадратний з дисперсії). Отримані в результаті стандартизації змінні мають нульове середнє і одиничну дисперсію. Цю операцію успішно можна проводити в програмі **MS Excel**. Розглянуті нижче дані вже стандартизовані.

Принципово метод **k-середніх** «працює» в такий спосіб:

- спочатку задається деякий розбиття даних на кластери (число кластерів визначається користувачем);
- обчислюються центри тяжкості кластерів;
- відбувається переміщення точок: кожна точка поміщається в найближчий до неї кластер;
- обчислюються центри тяжкості нових кластерів;
- кроки 2, 3 повторюються, поки не буде знайдена стабільна конфігурація (тобто кластери перестануть змінюватися) або число ітерацій не перевищить заданий користувачем. Підсумкова конфігурація і є шуканої.

Методи обчислення відстаней:

Таблиця 1.1

показники	Формули
Для кількісних шкал	
лінійна відстань	$d_{lij} = \sum_{l=1}^m x_i^l - x_j^l $

евклідова відстань	$d_{Eij} = \left(\sum_{l=1}^m (x_i^l - x_j^l)^2 \right)^{\frac{1}{2}}$
Квадрат евклідов відстані	$d^2_{Eij} = \sum_{l=1}^m (x_i^l - x_j^l)^2$
Узагальнене ступовий відстань Мінковського	$d_{Pij} = \left(\sum_{l=1}^m (x_i^l - x_j^l)^P \right)^{\frac{1}{P}}$
відстань Чебишева	$d_{ij} = \max_{1 \leq l \leq m} x_i^l - x_j^l $
Відстань міських кварталів (Манхеттенський відстань)	$d_H(x_i, x_j) = \sum_{l=1}^k x_i^l - x_j^l $

Порядок виконання роботи.

1 Підготовка:

1.1 Виберіть завдання.

1.2 Підготуйте вихідні дані для кластеризації:

1.2.1 На сайті Світового банку знайдіть дані по країнам світу відповідно до завдання.

1.2.2 Завантажте відповідні файли на комп'ютер (Download) data - Excel file).

1.2.3 Зберіть дані з двох завантажених файлів в один файл в форматі CSV. Файл повинен містити три стовпці: назва країни, показник №1, показник №2. Схема підготовки файлу з вихідними даними наведена нижче

2 Скласти програму на мові VBA, що реалізує алгоритм методу k-means

(запуск редактора VBA з MS Excel – Alt-F11, або через меню «Макроси»)

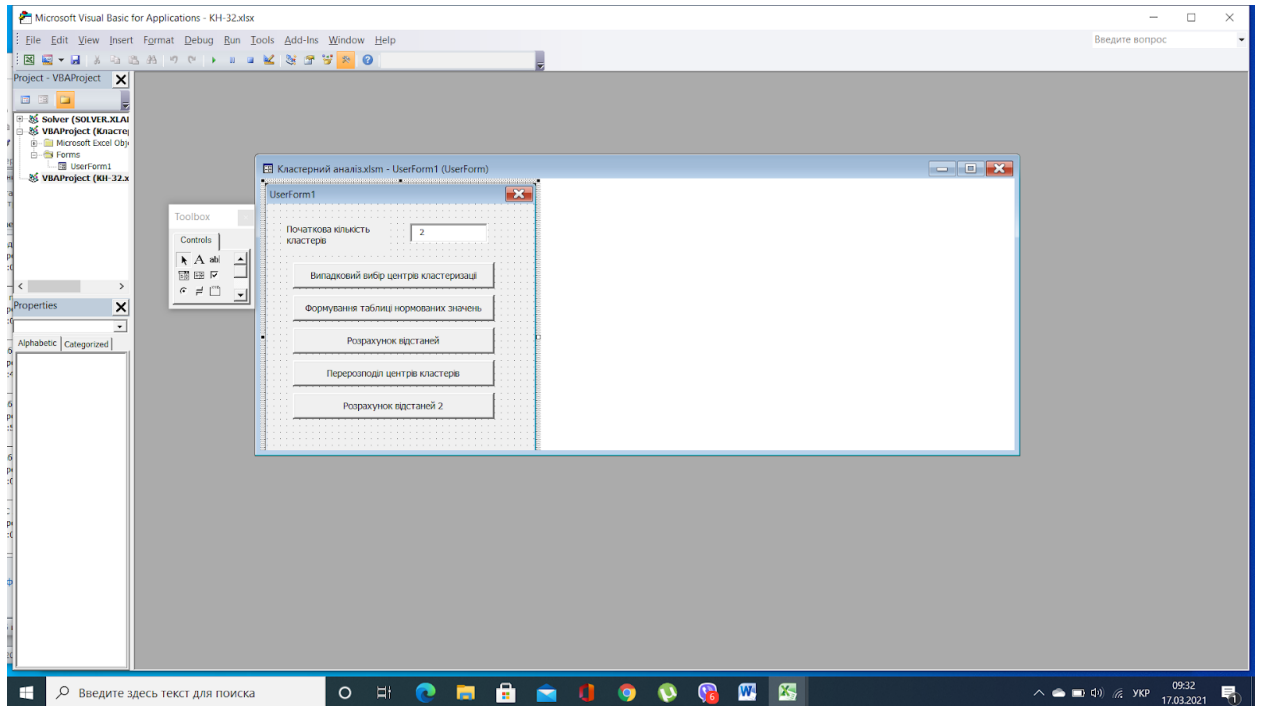
3 Експерименти:

3.1 Запустіть програму

3.2 Встановіть кількість кластерів, що дорівнює одиниці.

3.3 Виберіть файл з вихідними даними

3.4 Натискайте відповідні кнопки



3.5. Побудуйте діаграму з кривою навчання, діаграму кластерів і файл з кластеризованими об'єктами.

3.6 Запишіть номер експерименту і значення функції штрафу в таблицю експериментальних даних

3.7 Повторіть експеримент (кроки 2.2-2.6) п'ять разів. В результаті журнал експериментальних даних буде містити п'ять записів.

3.8 Послідовно збільшуючи число кластерів до восьми, проведіть серії експериментів (кроки 3.2 - 3.7). В результаті журнал експериментальних буде містити 40 записів.

4 Обробка експериментальних даних.

4.1 Виберіть експерименти, в яких досягнуто мінімальне значення функції штрафу для кожного числа кластерів, запишіть ці дані в таблицю оброблених експериментальних даних

4.2 На основі отриманої таблиці оброблених експериментальних даних побудуйте графік залежності мінімального значення функції штрафу від числа кластерів. Приклад такого графіка наведено нижче

4.3 По побудованій графіком, користуючись методом ліктя, визначте оптимальне число кластерів. Для наведеного вище графіка характерний злам відбувається при числі кластерів, що дорівнює трьом, відповідно в даному випадку оптимальне число кластерів дорівнює трьом.

4.4 Зробіть висновки по роботі:

5 Звіт по роботі:

5.1 Складіть звіт про роботу.

Варіанти завдань

Нижче наведений приклад рішення завдання

Кількість об'єктів	17	Кількість критеріїв	10							
	Показники за жовтень 2010р.									
Місто	Населення, тис. чол.	Густина населення на 1 кв.м., чол.	Кількість народившихся на 1000 чол.	Кількість померлих на 1000 чол.	Середньомісячна з/п, руб.	Середньорічна чисельність робітників організації відносно загального населення, чол.	Чисельність незайнятих громадян відносно кількості насел., чол.	Загальна площа житлових приміщень на одного чол., кв.м.	Кількість злочинів на одного мешканця	Кількість учнів загальноосвітніх закладів на одного мешканця
Нью-Йорк	11514,3	10588,4	10,7	10,9	38410,5	0,437	0,0054	18,7	0,016	0,068
Лондон	4848,7	3480	11,5	13,5	27189,5	0,414	0,0098	23,2	0,013	0,072
Париж	1473,7	2947,4	12,4	12,4	23374,9	0,285	0,0071	21,6	0,029	0,083
Відень	1350,1	2489,6	13	11,7	23216,6	0,323	0,0136	22,7	0,033	0,147
Кейптаун	1250,6	3153,9	10,8	16,5	21821,2	0,353	0,007	22,6	0,035	0,084
Веллінгтон	1164,9	2152,9	11,1	15,6	20690,5	0,34	0,0066	23	0,027	0,082
Нью-Делі	1154	1923,3	11,8	12,6	19317,1	0,275	0,0037	21,9	0,017	0,086

Мон реал ь	1143, 6	1865	13,1	13,1	19410	0,3	0,009	22,7	0,02	0,088
Мех ико	1130, 3	2258 ,4	12,2	12,2	20510,7	0,312	0,009	22,6	0,026	0,087
Шан хай	1089, 9	3127 ,6	9,9	12,4	21053,8	0,276	0,0044	22,3	0,019	0,078
Токі о	1062, 3	1518 ,4	14	11,9	22089,5	0,306	0,0109	20,5	0,021	0,093
Ріо- де- Жан ейро	1021, 2	1791 ,6	10,3	14,2	18294	0,286	0,0082	21,2	0,017	0,081
Буен ос- Айре с	991,5	1239 ,4	12,8	13	22678,8	0,31	0,007	21,3	0,029 3	0,093
Хано й	973,9	2754 ,1	13,5	11	25159	0,291	0,0063	21,7	0,029 6	0,086
Сеул	890	1633 ,1	11	14,6	18178,2	0,327	0,327	24,7	0,014	0,087
Ашх абад	837,8	2192 ,1	10	15,3	18107	0,3	0,0073	24,7	0,018 4	0,155
Ташк ент	744,9	991, 3	12,3	11,5	22587,8	0,353	0,006	26,3	0,018	0,102

Кількі
сть
об'єкт
ів

Кількість
критеріїв

17

10

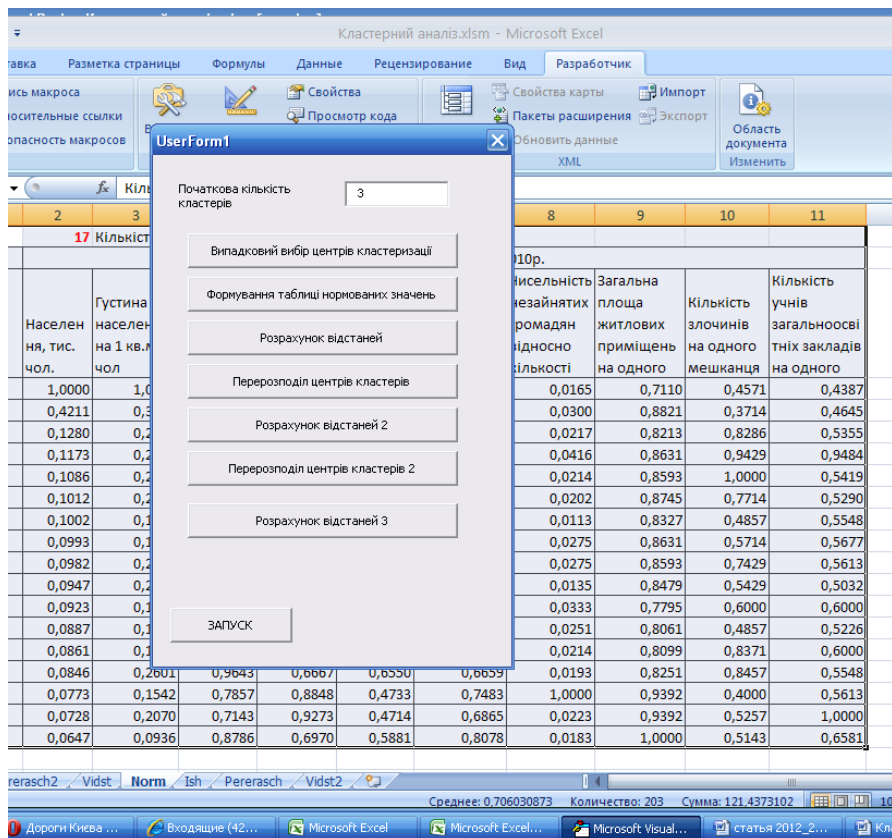
Показники за жовтень 2010р.										
Місто	Насел ення, тис. чол.	Густи на насел ення на 1 кв.м., чол.	Кількіст ь народи вшихся на 1000 чол.	Кільк ість поме рлих на 1000 чол.	Середньо місячна з/п, руб.	Середнь орічна чисельн ість робітни ків організа цій відносно загальн ого населен ня, чол.	Чисел ьність незай нятих грома дян віднос но кілько сті насел. , чол.	Загал ьна площ а житло вих примі щень на одног о чол., кв.м.	Кількі сть злочи нів на одног о мешк анця	Кількість учнів загально освітніх закладів на одного мешканц я
Нью- Йорк	1,000 0	1,000 0	0,7643	0,660 6	1,0000	1,0000	0,0165	0,711 0	0,457 1	0,4387
Лонд он	0,421 1	0,328 7	0,8214	0,818 2	0,7079	0,9474	0,0300	0,882 1	0,371 4	0,4645
Пари ж	0,128 0	0,278 4	0,8857	0,751 5	0,6086	0,6522	0,0217	0,821 3	0,828 6	0,5355
Віден ь	0,117 3	0,235 1	0,9286	0,709 1	0,6044	0,7391	0,0416	0,863 1	0,942 9	0,9484

Кейптаун	0,1086	0,2979	0,7714	1,0000	0,5681	0,8078	0,0214	0,8593	1,0000	0,5419
Веллінгтон	0,1012	0,2033	0,7929	0,9455	0,5387	0,7780	0,0202	0,8745	0,7714	0,5290
Нью-Делі	0,1002	0,1816	0,8429	0,7636	0,5029	0,6293	0,0113	0,8327	0,4857	0,5548
Монреаль	0,0993	0,1761	0,9357	0,7939	0,5053	0,6865	0,0275	0,8631	0,5714	0,5677
Мехіко	0,0982	0,2133	0,8714	0,7394	0,5340	0,7140	0,0275	0,8593	0,7429	0,5613
Шанхай	0,0947	0,2954	0,7071	0,7515	0,5481	0,6316	0,0135	0,8479	0,5429	0,5032
Токіо	0,0923	0,1434	1,0000	0,7212	0,5751	0,7002	0,0333	0,7795	0,6000	0,6000
Ріо-де-Жанейро	0,0887	0,1692	0,7357	0,8606	0,4763	0,6545	0,0251	0,8061	0,4857	0,5226
Буенос-Айрес	0,0861	0,1171	0,9143	0,7879	0,5904	0,7094	0,0214	0,8099	0,8371	0,6000
Ханой	0,0846	0,2601	0,9643	0,6667	0,6550	0,6659	0,0193	0,8251	0,8457	0,5548
Сеул	0,0773	0,1542	0,7857	0,8848	0,4733	0,7483	1,0000	0,9392	0,4000	0,5613
Ашхабад	0,0728	0,2070	0,7143	0,9273	0,4714	0,6865	0,0223	0,9392	0,5257	1,0000
Ташкент	0,0647	0,0936	0,8786	0,6970	0,5881	0,8078	0,0183	1,0000	0,5143	0,6581

Перша ітерація реалізації алгоритму

Місто	Відстань1	Відстань2	Відстань3	Мінімальна відстань	Кластер
Нью-Йорк	1,52043	1,444266	1,390355	1,39035463	3
Лондон	0,772878	0,691017	0,548228	0,54822763	3
Париж	0,640153	0,194628	0,310405	0,19462817	2
Відень	0,547342	0,397736	0,553797	0,39773582	2
Кейптаун	0,695196	0,37566	0,536093	0,3756601	2
Веллінгтон	0,553711	0,261097	0,308674	0,26109726	2
Нью-Делі	0,510094	0,387684	0,146806	0,14680561	3
Монреаль	0,514165	0,294012	0	0	3
Мехіко	0,558003	0,172251	0,198768	0,17225109	2
Шанхай	0,551011	0,425941	0,280252	0,28025176	3
Токіо	0,574981	0,264853	0,156481	0,1564807	3
Ріо-де-Жанейро	0,504861	0,430668	0,243149	0,24314889	3
Буенос-Айрес	0,596738	0	0,294012	0	2
Ханой	0,693609	0,214633	0,352167	0,21463336	2
Сеул	1,085196	1,099812	1,008712	1,00871155	3
Ашхабад	0	0,596738	0,514165	0	1
Ташкент	0,491976	0,404965	0,269211	0,26921061	3

Форма для введення кількості кластерів та запуску окремих процедур алгоритму



Населення, тис. чол.	Густина населення на 1 кв.м., чол.	Кількість народившихся на 1000 чол.	Кількість помрлих на 1000 чол.	Середньо місячна з/п, руб.	Середньорічна чисельність робітників в організації відносно загального населення, чол.	Чисельність незайнятих громадян відносно кількості насел., чол.	Загальна площа житлових приміщень на одного чол., кв.м.	Кількість злочинів на одного мешканця	Кількість учнів загальноосвітніх закладів на одного мешканця
0,0728	0,103514	0,2381	0,231818	0,09428	0,114416	0,003189	0,117395	0,058413	0,1
0,1034	0,114652	0,2918	0,2	0,11712	0,120628	0,003532	0,105581	0,094739	0,061014
0,2265	0,141238	0,2767	0,193098	0,11949	0,126028	0,018659	0,106411	0,054674	0,054122

Населення, тис. чол.	Густина населення на 1 кв.м., чол.	Кількість народившихся на 1000 чол.	Кількість помрлих на 1000 чол.	Середньо місячна з/п, руб.	Середньорічна чисельність робітників в організації відносно загального населення, чол.	Чисельність незайнятих громадян відносно кількості насел., чол.	Загальна площа житлових приміщень на одного чол., кв.м.	Кількість злочинів на одного мешканця	Кількість учнів загальноосвітніх закладів на одного мешканця
----------------------	------------------------------------	-------------------------------------	--------------------------------	----------------------------	--	---	---	---------------------------------------	--

	кв.м., чол		1000 чол.		в організац ій відносно загально го населенн я, чол.	ян віднос но кількос ті насел., чол.	примі щень на одного чол., кв.м.	о мешк анця	на одного мешканця
0,0727 62	0,1035 14	0,238095	0,231 818	0,094282	0,114416	0,0031 89	0,1173 95	0,0584 13	0,1
0,0957 87	0,1006 69	0,286054	0,197 727	0,110975	0,118149	0,0132 87	0,1069 73	0,0759 41	0,058848
0,7105 51	0,3321 65	0,264286	0,184 848	0,170787	0,162281	0,0033 2	0,0995 72	0,0460 32	0,045161

Друга ітерація реалізації алгоритму

Місто	Відстань1	Відстань2	Відстань3	Мінімальна відстань	Кластер
Нью-Йорк	2,086879	2,05343	2,001671	2,00167141	3
Лондон	1,656144	1,638573	1,628454	1,62845373	3
Париж	1,598556	1,576236	1,602784	1,5762359	2
Відень	1,853092	1,838852	1,86673	1,83885192	2
Кейптаун	1,803202	1,788403	1,816252	1,78840333	2
Веллінгтон	1,652775	1,641029	1,667961	1,64102916	2
Нью-Делі	1,41021	1,399595	1,42406	1,39959544	2
Монреаль	1,53086	1,51674	1,542465	1,51674021	2
Мехико	1,56733	1,549669	1,577658	1,54966906	2
Шанхай	1,389909	1,379759	1,402642	1,37975908	2
Токіо	1,542367	1,523348	1,551104	1,52334779	2
Ріо-де-Жанейро	1,386026	1,381183	1,406195	1,38118294	2
Буенос-Айрес	1,644154	1,625435	1,657636	1,62543533	2
Ханой	1,639351	1,613659	1,643925	1,6136592	2
Сеул	1,809073	1,80755	1,818063	1,80755034	2
Ашхабад	1,701373	1,710333	1,733915	1,70137343	1
Ташкент	1,627817	1,618591	1,6442	1,61859131	2

Третя ітерація реалізації алгоритму

Місто	Відстань1	Відстань2	Відстань3	Мінімальна відстань	Кластер
Нью-Йорк	2,086879	2,071687	1,758818	1,758817943	3
Лондон	1,656144	1,651154	1,610893	1,610893151	3
Париж	1,598556	1,59236	1,689733	1,592360219	2
Відень	1,853092	1,854012	1,950105	1,85309192	1
Кейптаун	1,803202	1,80449	1,895806	1,803202143	1
Веллінгтон	1,652775	1,654944	1,766779	1,652775169	1
Нью-Делі	1,41021	1,411504	1,544	1,410209517	1

Монреаль	1,53086	1,529121	1,655053	1,529121141	2
Мехико	1,56733	1,564011	1,682648	1,564011143	2
Шанхай	1,389909	1,39323	1,509075	1,389908973	1
Токіо	1,542367	1,536212	1,666617	1,536212121	2
Ріо-де-Жанейро	1,386026	1,392582	1,532517	1,386026174	1
Буенос-Айрес	1,644154	1,639871	1,770972	1,639871342	2
Ханой	1,639351	1,629645	1,741603	1,629644868	2
Сеул	1,809073	1,81023	1,928579	1,809073493	1
Ашхабад	1,701373	1,720779	1,840759	1,701373433	1
Ташкент	1,627817	1,628886	1,763618	1,627817318	1

Практична робота № 5

Порівняння і аналіз двох вибірок

з використанням критеріїв Стюдента і хи-квадрат

Загальні відомості

<https://sites.google.com/site/ktnoscience/Home/lecture/16>

1. Задані результати бігу на дистанцію 100 м в секундах в двох групах студентів. Студенти першої групи протягом року відвідували факультативні заняття з фізкультури. Визначити, чи достовірні відмінності за результатами бігу в цих групах.

Які відвідували факультет	Які не відвідували факультет
12,6	12,8
12,3	13,2
11,9	13,0
12,2	12,9
13,0	13,5
12,4	13,1

- 2) В ході соціологічного опитування на питання про перенесений в дитинстві захворюванні відповіді розподілилися наступним чином:

	Так	Ні	Не пам'ятаю
Чоловіки	58	10	11
Жінки	35	25	23

Чи є достовірні відмінності у відповідях жінок і чоловіків?

3. Наведено дані щомісячної результативності (кількість голів) футбольної команди в двох сезонах

Місяць	3	4	5	6	7	8	9	10	11
2008 р	3	4	5	8	9	1	2	4	5
2009 р	6	19	3	2	14	4	5	17	1

Визначити, чи є статистичні відмінності в щомісячній результативності команди в розглянутих сезонах?

4. Визначити, чи достовірні відмінності в кількості придбаних туристських путівок сімейними парами і окремими туристами.

	Кількість придбаних путівок					
Місяць	1	2	3	4	5	6
Пари	67	75	58	89	96	94
Одиночки	43	56	78	87	85	90

5. У таблиці наведено результати групи студентів зі швидкісного читання до і після спеціального курсу по швидкому читанню.

Студент	1	2	3	4	5	6	7	8	9	10
До курсу	86	83	86	70	66	90	70	85	77	86
Після курсу	82	79	91	77	68	86	81	90	85	94

Чи відбулися статистично значущі зміни швидкості читання у студентів?

6. Розглядається заробітна плата обслуговуючого персоналу і працівників ресторану готелю.

Персонал	Ресторан
2100	2100
2100	2100
2000	2000
2000	2000
2000	2000
1900	1900
1800	1800

Чи можна за цими даними зробити висновок про більшу зарплату працівників ресторану?

Вимоги до звіту

Звіт повинен містити:

- 1 Титульний аркуш: найменування роботи, варіант завдання, ПІБ студента, номер навчальної групи, дата виконання роботи.
- 2 Реферат.
- 3 Зміст.
- 4 Завдання.
- 5 Опис виконаної роботи:
- 6 Отримані результати.
- 7 Аналіз результатів.
- 8 Список використаних джерел:

Практична робота № 6

Тема: Робота в середовищі мови R

Мета: Навчитись працювати в середовищі мови R

Завдання. Встановлення ПО та робота з мовою R у середовищі

Встановимо необхідні пакети:

psych - містить функції для розрахунку описових статистик

dplyr - містить функції для роботи з data.frame;

ggplot2 - найпотужніший пакет для побудови красивих графіків, діаграм, карт і т. д.

```
> install.packages(c("psych","dplyr","ggplot2"))
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing packages into 'C:/Users/RAZER/Documents/R/win-library/4.0'
(as 'lib' is unspecified)
also installing the dependencies 'colorspace', 'tmvnsim', 'fansi', 'pkgconfig', 'purrr', 'cli', 'crayon', 'utf8', 'farver',
'labeling', 'munsell', 'RColorBrewer', 'viridisLite', 'mnormt', 'ellipsis', 'generics', 'glue', 'lifecycle', 'magrittr', 'R
6', 'rlang', 'tibble', 'tidyselect', 'vctrs', 'pillar', 'digest', 'gtable', 'isoband', 'scales', 'withr'

There is a binary version available but the source version is later:
  binary source needs_compilation
dplyr  1.0.5  1.0.6             TRUE

Binaries will be installed
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/colorspace_2.0-1.zip'
Content type 'application/zip' length 2641019 bytes (2.5 MB)
downloaded 2.5 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/tmvnsim_1.0-2.zip'
```

Дані для свого варіанту:

Список країн по викидам CO2 за 2019 рік

№	Страна	2019 млн т/год
1	 Китай	9825,8
2	 США	4964,7
3	 Индия	2480,4
4	 Россия	1532,6
5	 Япония	1123,1
6	 Германия	683,8
7	 Иран	670,7
8	 Республика Корея	638,6

Дані, представлені у вигляді змінної-вектор:

```
1 conc <- c(9825, 4964, 2480, 1532, 1123, 683, 670, 638)
2 country <- c("китай", "США", "индия", "Россия", "япония", "Германия", "иран", "корей")
3 df <- data.frame(country, conc)
```


	country	conc
1	Китай	9825
2	США	4964
3	Индия	2480
4	Россия	1532
5	Япония	1123
6	Германия	683
7	Иран	670
8	Корея	638

Серед даних пропусків немає (нема значень NA). Можемо це перевірити, використавши функцію `sum()`. Якщо є пропуски, то і сума буде дорівнювати NA:

```
> view(df)
> sum(conc)
[1] 21915
> |
```

Створення нової змінної-вектору, в якій будуть 1, якщо значення у вихідному векторі більше середнього, і -1, якщо значення змінної менше середнього, і 0, якщо значення дорівнює середньому:

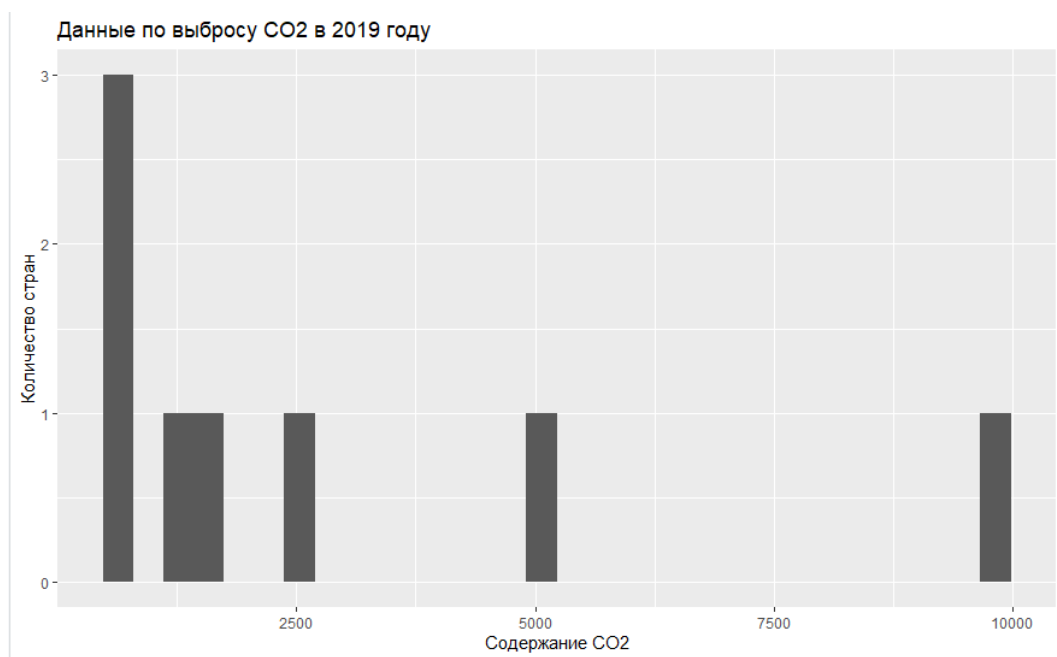
```
10
11 # НОВЫЙ ВЕКТОР
12 ser <- mean(df$conc, na.rm = T)
13 s <- conc
14 s <- replace(s, conc == ser, 0)
15 s <- replace(s, conc > ser, 1)
16 s <- replace(s, conc < ser, -1)

> s
[1] 1 1 -1 -1 -1 -1 -1 -1
> |
```

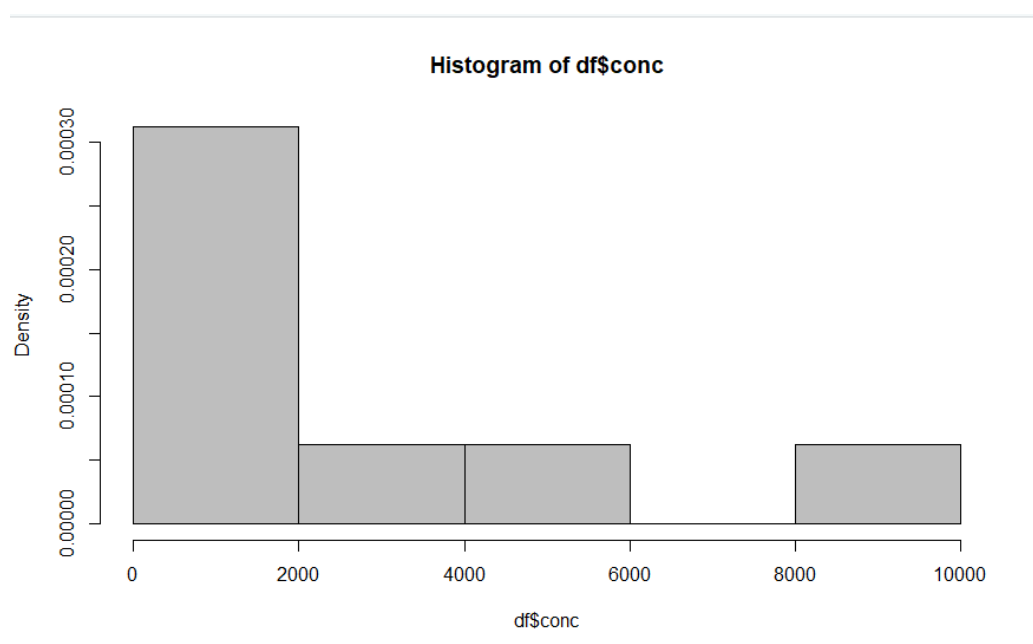
Виведемо описову статистику для змінної `conc`:

```
> describe(df)
      vars n   mean    sd median trimmed   mad min  max range skew kurtosis   se
country*  1 8    4.50   2.45     4.5    4.50   2.97  1    8     7  0.00   -1.65   0.87
conc      2 8 2739.38 3212.02 1327.5 2739.38 998.53 638 9825 9187 1.26    0.06 1135.62
>
```

Графік абсолютних частот:



Графік щільності розподілу:



Результующий файл lr6.R:

```
1 conc <- c(9825, 4964, 2480, 1532, 1123, 683, 670, 638)
2 country <- c("Китай", "США", "Индия", "Россия", "Япония", "Германия", "Иран", "Корея")
3 df <- data.frame(country, conc)
4 # описательная статистика
5 describe(df)
6 # график абсолютных частот
7 qplot(data=df, conc, xlab="Содержание CO2", ylab="Количество стран", main="Данные по выбросу CO2 в 2019 году")
8 # гистограмма плотности распределения
9 hist(df$conc, probability = TRUE, col="grey")
10
11 # новый вектор
12 ser <- mean(df$conc, na.rm = T)
13 s <- conc
14 s <- replace(s, conc == ser, 0)
15 s <- replace(s, conc > ser, 1)
16 s <- replace(s, conc < ser, -1)
```

Вимоги до звіту

Звіт повинен містити:

1 Титульний аркуш: найменування роботи, варіант завдання, ПІБ студента, номер навчальної групи, дата виконання роботи.

2 Реферат.

3 Зміст.

4 Завдання.

5 Опис виконаної роботи:

6 Отримані результати.

7 Аналіз результатів.

8 Список використаних джерел:

8.1 Джерела даних.

8.2 Нормативні документи.

9 Додатки.

Звіт повинен бути оформлений відповідно до чинних вимог університету

Практична робота № 7

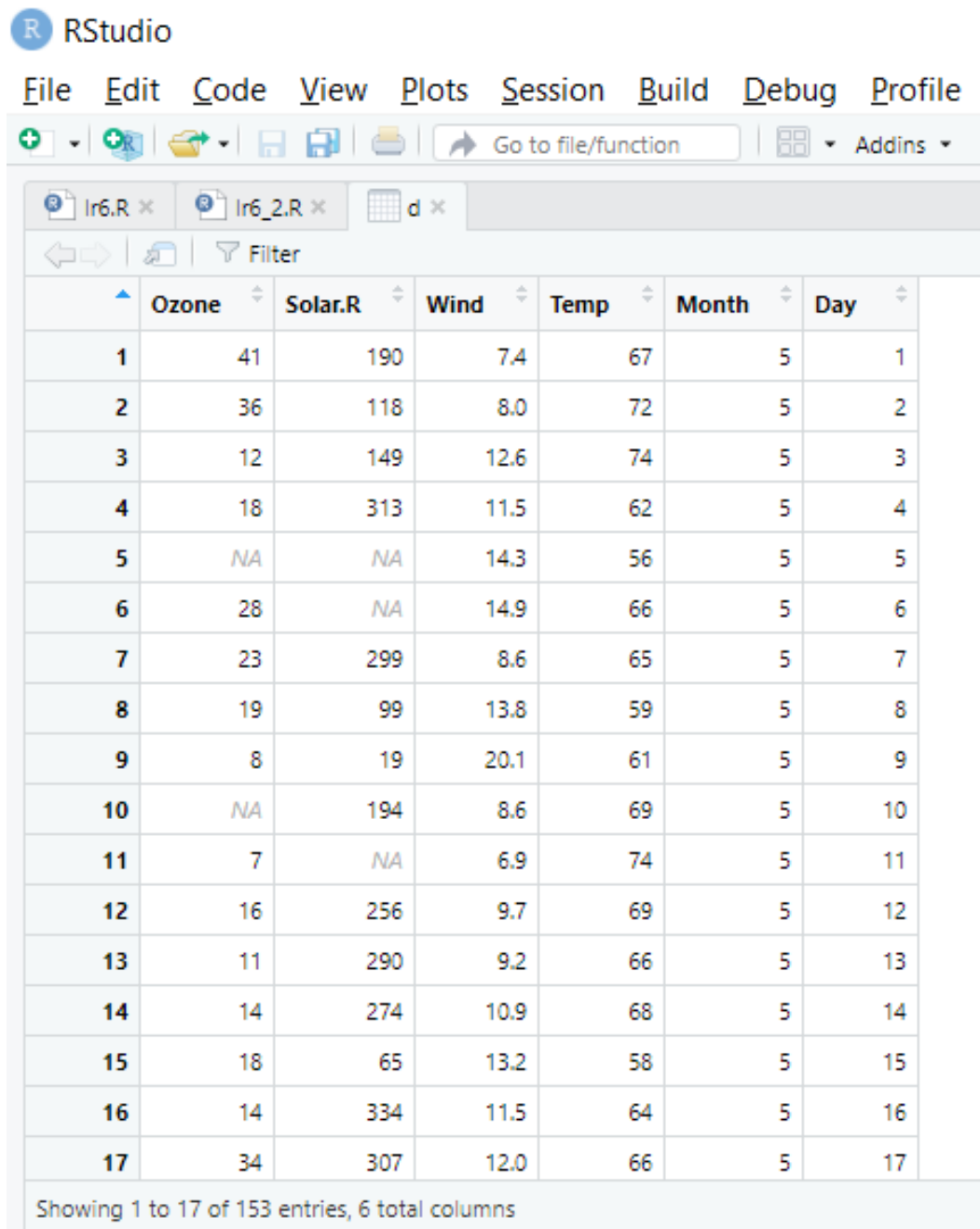
Тема: Статистична обробка даних на мові R

Мета: Навчитись працювати в середовищі мови R та здійснювати статистичну обробку даних

Завдання: Отримання лінійної регресії в середовищі мови R

Послідовність виконання роботи:

1. Завантажимо набір даних для свого варіанту та ознайомимося з ним:



RStudio

File Edit Code View Plots Session Build Debug Profile

Go to file/function

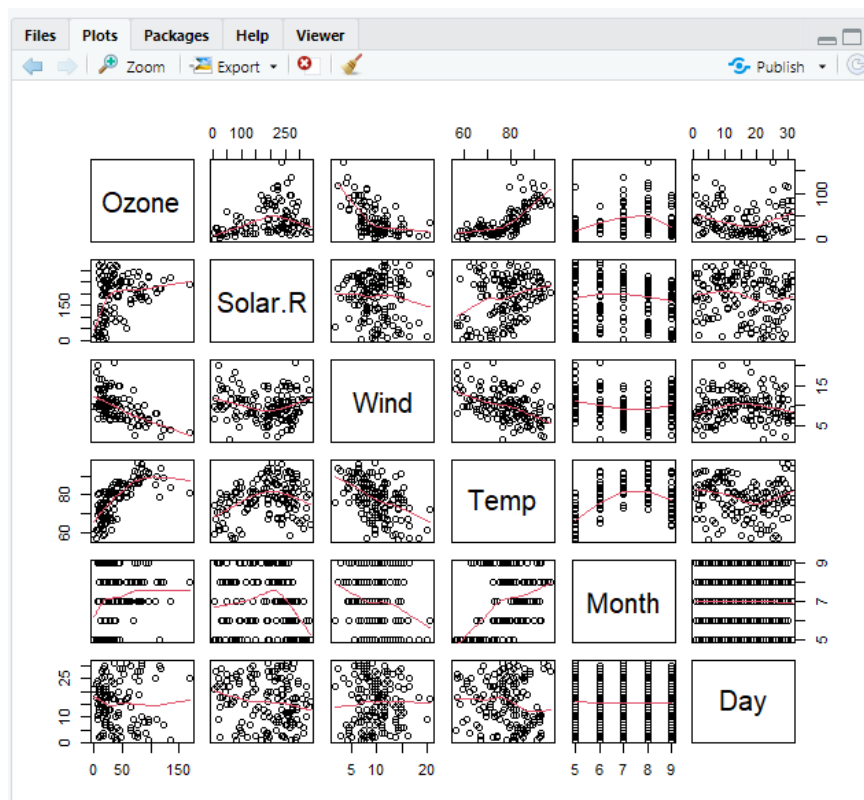
Ir6.R x Ir6_2.R x d x

Filter

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
10	NA	194	8.6	69	5	10
11	7	NA	6.9	74	5	11
12	16	256	9.7	69	5	12
13	11	290	9.2	66	5	13
14	14	274	10.9	68	5	14
15	18	65	13.2	58	5	15
16	14	334	11.5	64	5	16
17	34	307	12.0	66	5	17

Showing 1 to 17 of 153 entries, 6 total columns

2. Побудуємо графік кореляційного поля для кожного фактору:



3. Побудуємо рівняння парної лінійної регресії для кожного фактору:

```
<
> model1$coefficients
(Intercept)      Solar.R
18.5987278      0.1271653
> model2$coefficients
(Intercept)      Wind
96.872895      -5.550923
> model3$coefficients
(Intercept)      Temp
-146.995491      2.428703
> |
```

4. Перевіримо значущість кожного з отриманих рівнянь регресії:

```
> summary(model1)

Call:
lm(formula = ozone ~ Solar.R, data = d)

Residuals:
    Min       1Q   Median       3Q      Max
-48.292 -21.361  -8.864  16.373 119.136

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.59873    6.74790   2.756 0.006856 **
Solar.R       0.12717    0.03278   3.880 0.000179 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.33 on 109 degrees of freedom
(42 observations deleted due to missingness)
Multiple R-squared:  0.1213,    Adjusted R-squared:  0.1133
F-statistic: 15.05 on 1 and 109 DF,  p-value: 0.0001793
```

```
> summary(model2)

Call:
lm(formula = ozone ~ wind, data = d)

Residuals:
    Min       1Q   Median       3Q      Max
-51.572 -18.854  -4.868  15.234  90.000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  96.8729     7.2387   13.38  < 2e-16 ***
wind        -5.5509     0.6904   -8.04  9.27e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.47 on 114 degrees of freedom
(37 observations deleted due to missingness)
Multiple R-squared:  0.3619,    Adjusted R-squared:  0.3563
F-statistic: 64.64 on 1 and 114 DF,  p-value: 9.272e-13

> summary(model3)

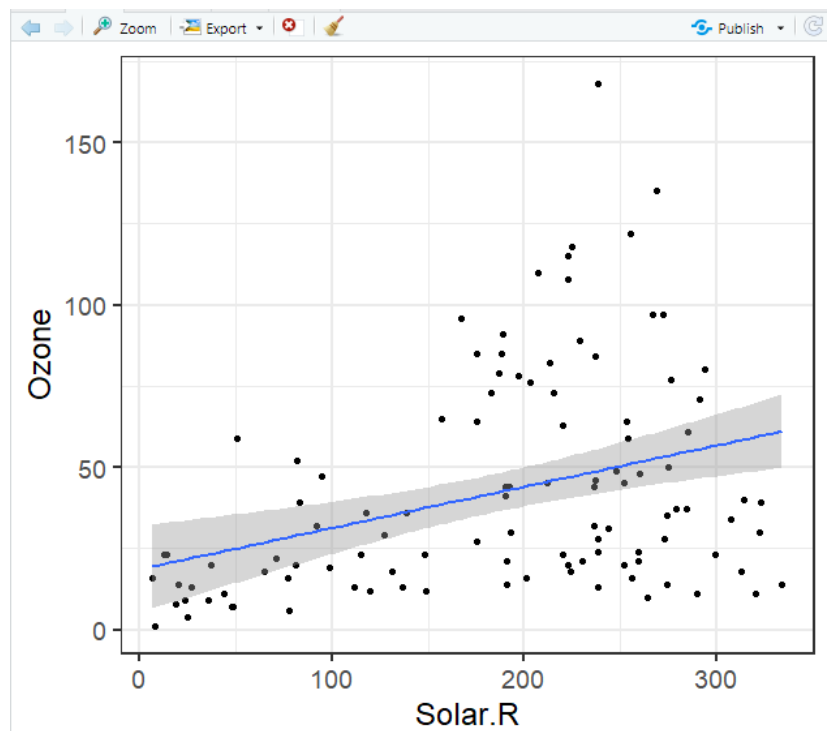
Call:
lm(formula = ozone ~ Temp, data = d)

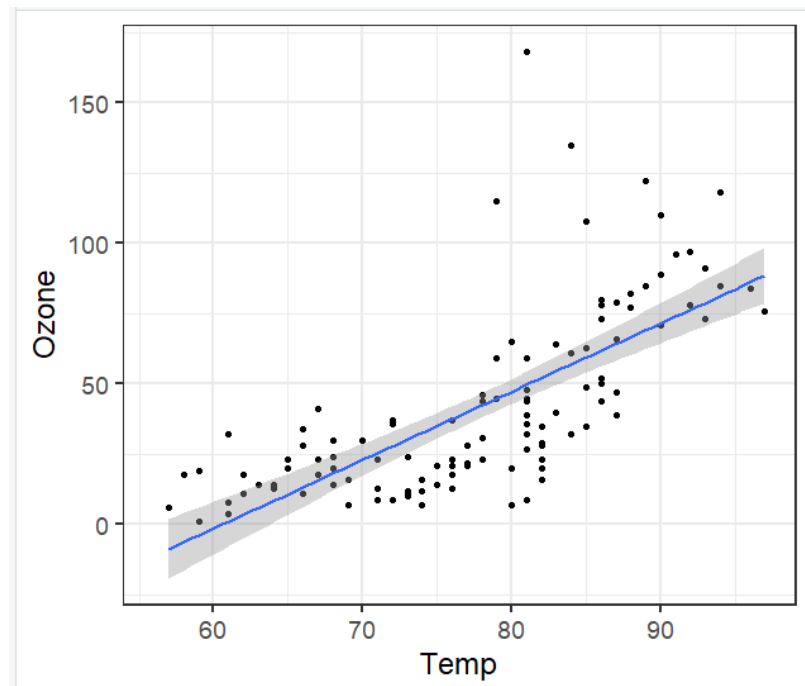
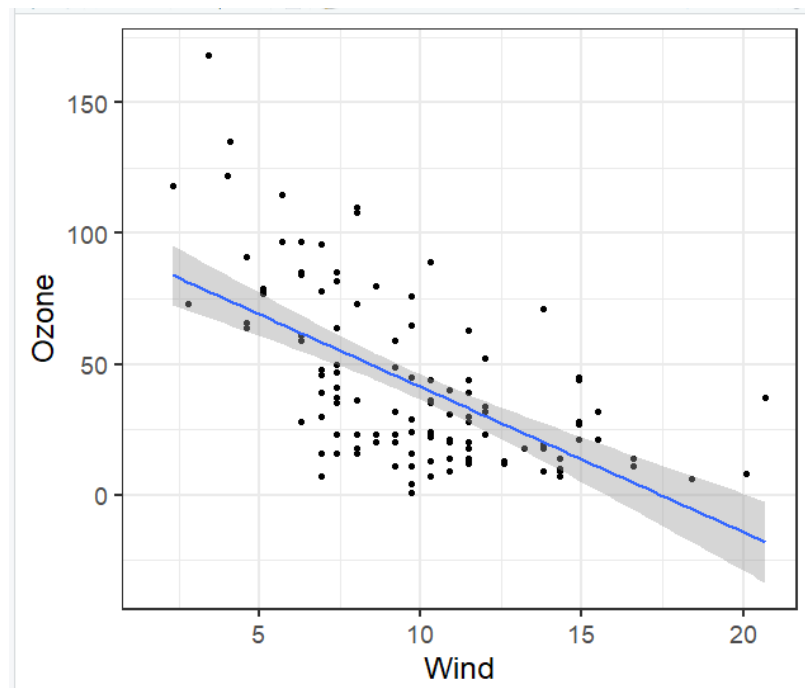
Residuals:
    Min       1Q   Median       3Q      Max
-40.729 -17.409  -0.587  11.306 118.271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -146.9955     18.2872  -8.038  9.37e-13 ***
Temp          2.4287      0.2331  10.418  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.71 on 114 degrees of freedom
(37 observations deleted due to missingness)
Multiple R-squared:  0.4877,    Adjusted R-squared:  0.4832
F-statistic: 108.5 on 1 and 114 DF,  p-value: < 2.2e-16
```

5. Відобразимо рівняння регресії із заданим довіреним інтервалом на графіках





6. Побудуємо прогнози по кожному з рівнянь парної регресії згідно заданих значень факторів:

```

> # Побудуємо 31 рядок, що містить тестові дані (згенеровано)
> # Побудуємо прогнози по кожному з рівнянь парної регресії для заданих в варіанте значень факторів
> predict(model1, data.frame(Solar.R = 350))
1
63.10657
> predict(model2, data.frame(wind = 8.3))
1
50.80023
> predict(model3, data.frame(Temp = 80))
1
47.30077

```

7. Побудуємо рівняння множинної лінійної регресії та отримаємо кореляційну матрицю:

```

> model <- lm(data=d, Ozone~Solar.R+wind+Temp)
> model$coefficients
      (Intercept)      Solar.R      wind      Temp
-64.34207893    0.05982059   -3.33359131    1.65209291
>
> cor(d)
      Ozone Solar.R      wind      Temp      Month      Day
Ozone      1      NA      NA      NA      NA      NA
Solar.R    NA      1      NA      NA      NA      NA
wind      NA      NA      1.0000000 -0.4579879  0.420947252  0.027180903
Temp      NA      NA      -0.4579879  1.0000000  0.420947252 -0.130593175
Month     NA      NA      -0.1782926  0.4209473  1.000000000 -0.007961763
Day       NA      NA      0.0271809 -0.1305932 -0.007961763  1.000000000
>

```

8. Побудуємо прогноз по рівнянню множинної регресії для заданих значень факторів:

```

> # Побудуємо прогноз по рівнянню множинної регресії для заданих значень факторів
> predict(model1, data.frame(Solar.R = 350, wind = 8.9, Temp = 80))
      1
63.10657

```

9. Результуючий файл lr6_2.R:

```

1 library("psych")
2 library("lmtest")
3 library("ggplot2")
4 library("dplyr")
5 library("MASS")
6 library("car")
7
8 # Загрузить набор данных для своего варианта, ознакомиться с его содержанием
9 d <- airquality
10
11 # Построить график корреляционного поля для каждого фактора
12 pairs(d, panel=panel.smooth)
13
14 # Построить уравнение парной линейной регрессии для каждого фактора
15 model1 <- lm(data=d, Ozone~Solar.R)
16 model2 <- lm(data=d, Ozone~wind)
17 model3 <- lm(data=d, Ozone~Temp)
18
19 model1$coefficients
20 model2$coefficients
21 model3$coefficients
22
23 # Проверить значимость каждого из полученных уравнений регрессии
24 summary(model1)
25 summary(model2)
26 summary(model3)
27
28 # Показать уравнения регрессии с заданным в варианте доверительным интервалом на графиках
29 qplot(data = d, Solar.R, Ozone) + stat_smooth(method="lm", level = 0.95) + theme_bw(base_size = 18)
30 qplot(data = d, wind, Ozone) + stat_smooth(method="lm", level = 0.95) + theme_bw(base_size = 18)
31 qplot(data = d, Temp, Ozone) + stat_smooth(method="lm", level = 0.95) + theme_bw(base_size = 18)
32
33 # Построить прогнозы по каждому из уравнений парной регрессии для заданных в варианте значений факторов
34 predict(model1, data.frame(Solar.R = 350))
35 predict(model2, data.frame(wind = 8.3))
36 predict(model3, data.frame(Temp = 80))
37
38 # Построить уравнение множественной линейной регрессии и получить корреляционную матрицу
39 model <- lm(data=d, Ozone~Solar.R+wind+Temp)
40 model$coefficients
41
42 cor(d)
43
44 # Построить прогноз по уравнению множественной регрессии для заданных в варианте значений факторов
45 predict(model1, data.frame(Solar.R = 350, wind = 8.9, Temp = 80))

```


Вимоги до звіту

Звіт повинен містити:

1 Титульний аркуш: найменування роботи, варіант завдання, ПІБ студента, номер навчальної групи, дата виконання роботи.

2 Реферат.

3 Зміст.

4 Завдання.

5 Опис виконаної роботи:

6 Отримані результати.

7 Аналіз результатів.

8 Список використаних джерел:

8.1 Джерела даних.

8.2 Нормативні документи.

9 Додатки.

Звіт повинен бути оформлений відповідно до чинних вимог університету

СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

1. Цибульник С.О., Павловський О.М. Сучасні методи обробки інформації. Лекції [Електронний ресурс]: навч. посіб. для здобувачів ступеня доктора філософії спеціальності 151 «Автоматизація та комп'ютерно-інтегровані технології» / С.О. Цибульник, О.М. Павловський; КПІ ім. Ігоря Сікорського. Київ: КПІ ім. Ігоря Сікорського, 2021. 111 с.
2. Матеріали VI Міжнародної науково-технічної конференції молодих учених. Актуальні задачі сучасних технологій – Тернопіль 16-17 листопада 2017. 141 с.
3. Бахрушин В.Є. Методи аналізу даних : навчальний посібник. Запоріжжя : КПУ, 2011. 268 с.
4. Ланде Д.В., Субач І.Ю., Бояринова Ю.Є. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навчальний посібник. К.: ІСЗЗІ КПІ ім. Ігоря Сікорського», 2018. 300 с.

Кіктєв Микола Олександрович

Коваль Валерій Вікторович

НАВЧАЛЬНЕ ВИДАННЯ

**ОБРОБКА ІНФОРМАЦІЇ В КОМП'ЮТЕРНО-ІНТЕГРОВАНИХ
СИСТЕМАХ АВТОМАТИЗАЦІЇ**

Методичні вказівки до проведення практичних занять
для здобувачів третього (освітньо-наукового) рівня вищої освіти за
спеціальністю 151 «Автоматизація та комп'ютерно-інтегровані технології»
(кваліфікація: PhD доктор філософії)

Укладачі: КІКТЄВ Микола Олександрович,
КОВАЛЬ Валерій Вікторович

Зав. Видавничим центром НУБіП А.П. Колесніков

Підписано до друку 17.10.25
Ум. друк. арк. 5,9
Наклад 20 прим.

Формат 60x84/16
Обл. вид. арк. 5,2
Зам. № 2915

Видавничий центр НУБіП
03041, вул. Героїв Оборони, 15