МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

ОБРОБКА ІНФОРМАЦІЇ В КОМП'ЮТЕРНО-ІНТЕГРОВАНИХ ТА РОБОТОТЕХНІЧНИХ СИСТЕМАХ

Методичні вказівки до проведення практичних занять

для здобувачів третього (освітньо-наукового) рівня вищої освіти за спеціальністю 174, G7 «Автоматизація, комп'ютерно-інтегровані технології та робототехніка» (кваліфікація: PhD доктор філософії)

КИЇВ-2025

УДК 621.391(075.8)

У методичних вказівках представлені методи обробки інформації в комп'ютерно-інтегрованих системах автоматизації та робототехніки. Практичні роботи призначені для аспірантів за спеціальністю 174/G7 «Автоматизація, комп'ютерно-інтегровані технології та робототехніка», а також може бути корисним докторантам і студенетам бакалаврату і магістратури в області автоматизації та інформаційних технологій.

Рекомендовано до друку рішенням вченої ради ННІ енергетики, автоматики і енергозбереження Національного університету біоресурсів і природокористування України (протокол № 3_ від _17_ жовтня 2025 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 ТА РОБОТОТЕХНІЧНИХ СИСТЕМАХ

Методичні вказівки до проведення практичних занять

Укладачі: КІКТЄВ Микола Олександрович,
КОВАЛЬ Валерій Вікторович

Зав. Видавничим центром НУБіП А.П. Колесніков

Підписано до друку 17.10.25 Формат 60x84/16
Ум. друк. арк. 5,9 Обл. вид. арк. 5,2
Наклад 20 прим. Зам. № 2915

Видавничий центр НУБіП
03041, вул. Героїв Оборони, 15

ЗМІСТ

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 (табл. 6).

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 (табл. 7).

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 розшукуються.

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

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 частини 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 лабораторної роботи №2 «Регресійний аналіз» курсу «Попередній аналіз та підготовка даних» системи 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 стандартами 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 медичної діагностичної системи, заснованої на бінарному класифікаторі, отримані наступні результати

Варіант 2. При випробуванні антивіруса, заснованого на бінарному

класифікаторі, отримані наступні результати

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 лабораторної роботи №3 (Бінарна класифікація)» курсу « Аналіз даних» 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 стандартами університету

Приклад виконання роботи

Приклад з лекції

Вводимо вхідні дані:

Доп. столбец	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, :
=МУМНОЖ(A3:C3;$B$13:$B$15)

Отримаємо значення вектору Z.
Вводимо значення вектору h:
=1/(1+EXP(-E3))

Вводимо значення функції штрафу

=ЕСЛИ(D3=0;-LN(1-F3);-LN(F3))

В результаті отримуємо таблицю:

Доп. столбец	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. столбец	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

Хтрансп *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 У

Хтрансп * 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 + 0.01198X1-0.2994X2.

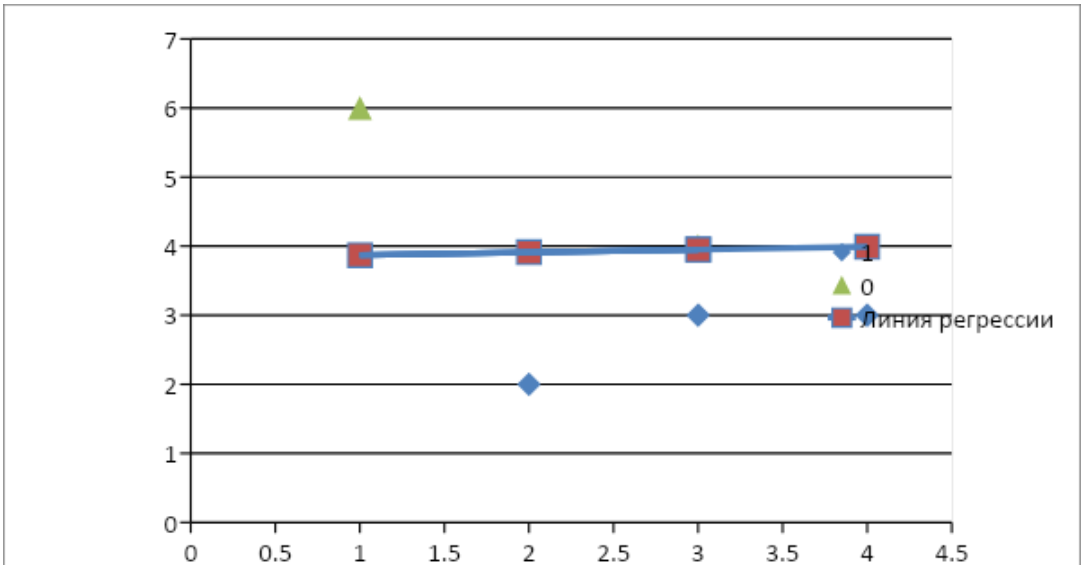
Таким чином, отримуємо таблицю з заданими та розрахунковими значеннями Y:

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

4	3	1	0,79642	3,990047

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 x1-x2, виразимо x2 через x1:
=(0,5-$C$16-$C$17*B3)/$C$18

x2

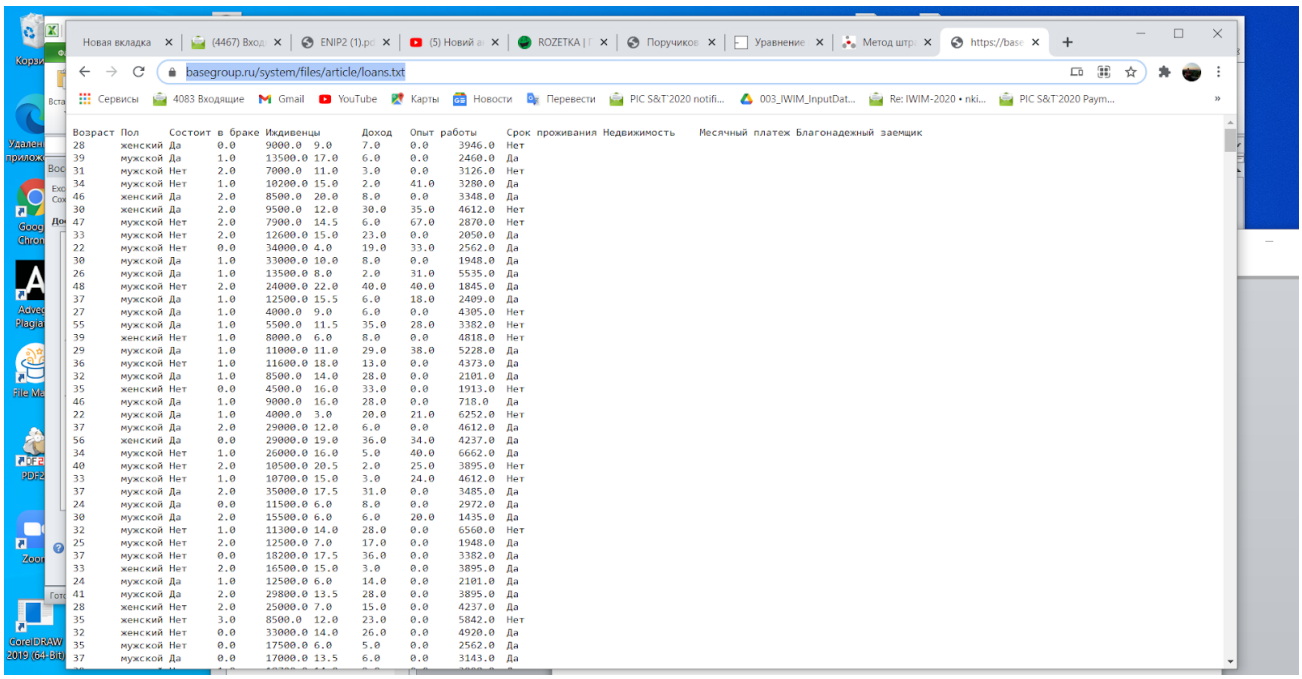


x1

На графіку відображено лінію розподілу між

Приклад 2.Видача банківських кредитів.

Вихідні дані:
https://basegroup.ru/system/files/article/loans.txt



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:

$$Y = -1.7422 + 0.04902X1 + 0.02671X2 + 0.1444X3$$

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).

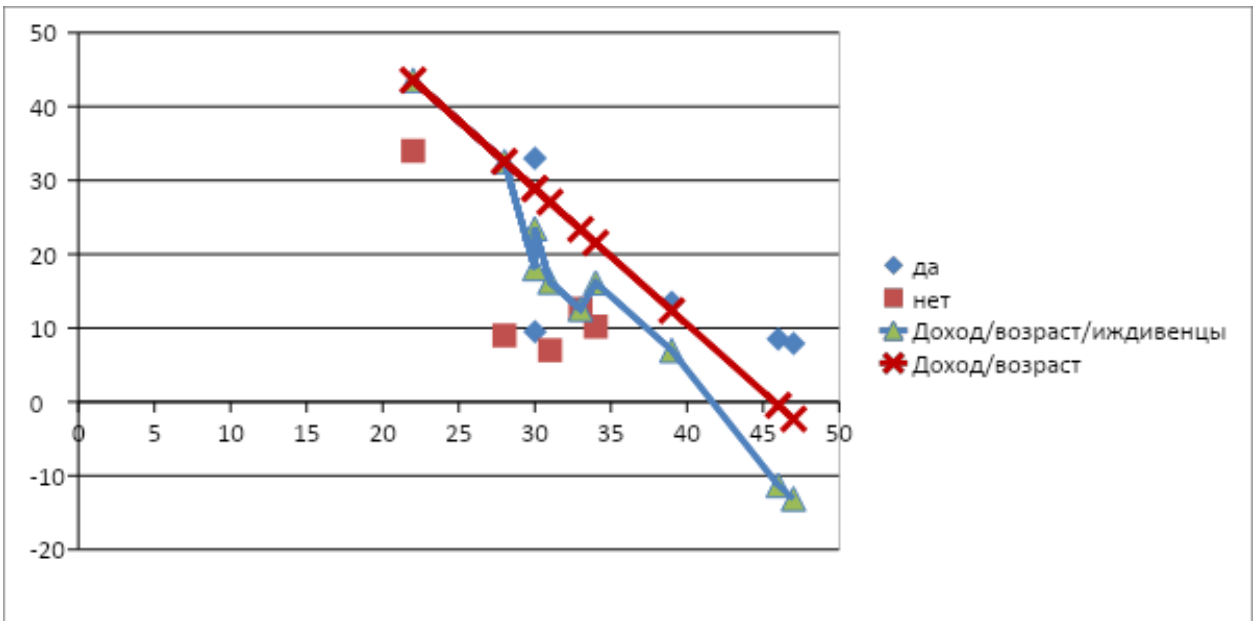
Вік	Дохід	Утриманці	Благонадійний позичальник
39	13,5	1	да
46	8,5	2	да
30	9,5	2	да
47	7,9	2	да
30	33	1	да
28	9	0	нет
31	7	2	нет
34	10,2	1	нет
33	12,6	2	нет
22	34	0	нет

Для нанесення лінії розподілу між класами, виразимо x2 від x1 та x3, та x2 від x1 (не враховуючи x3).
=(0,5-$G$19-$G$20*J32-$G$22*K32)/$G$21
=(0,5-$G$19-$G$20*J32)/$G$21

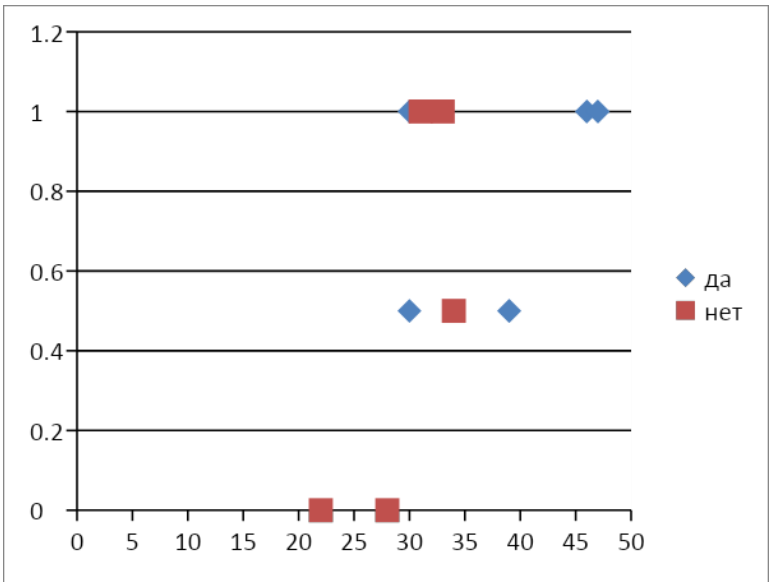
Тепер для побудови графіка впорядкуємо дані за параметром x1 (вік)

Вік	Утриманці	Дохід	Дохід (без x3)
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

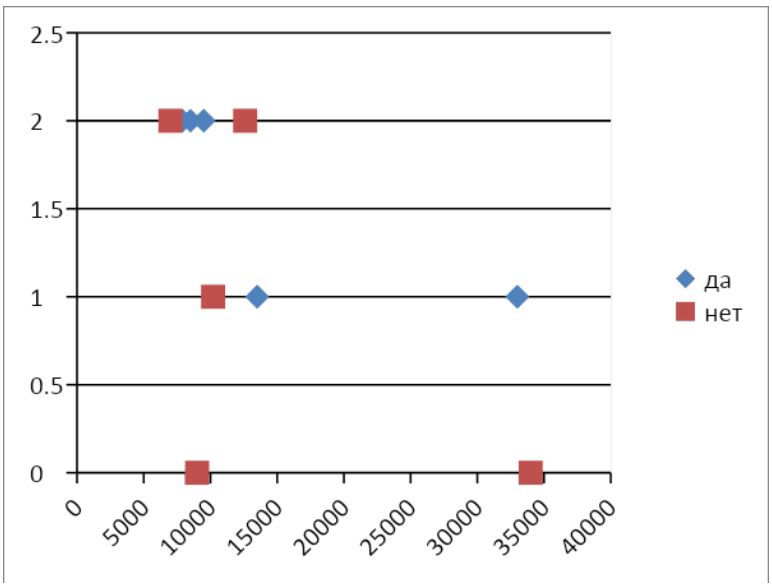
Отримуємо графік.



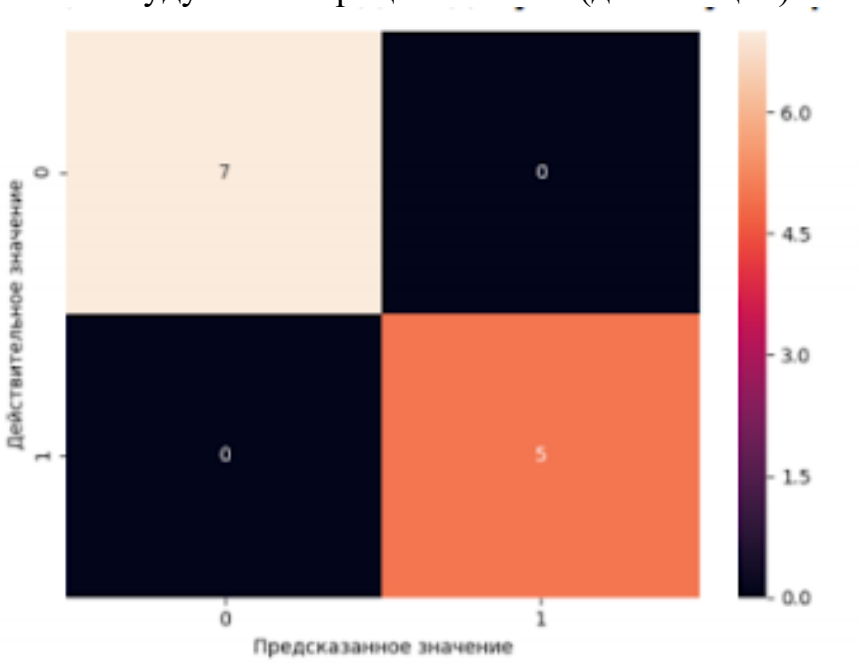
Аналогічно наносимо на площину в координатах інших пар параметрів: x1 – вік, x2 - кількість іждивенців



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 хвороби
https://basegroup.ru/system/files/article/diabetes.txt

Приклад зі вступом до наступного рівня навчання

Таблица 1. Результат классификации данных тестовой выборки

Независимые переменные			Целевая переменная	
gmat	**gpa**	**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:

Т аблиця 1.1

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

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

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»)

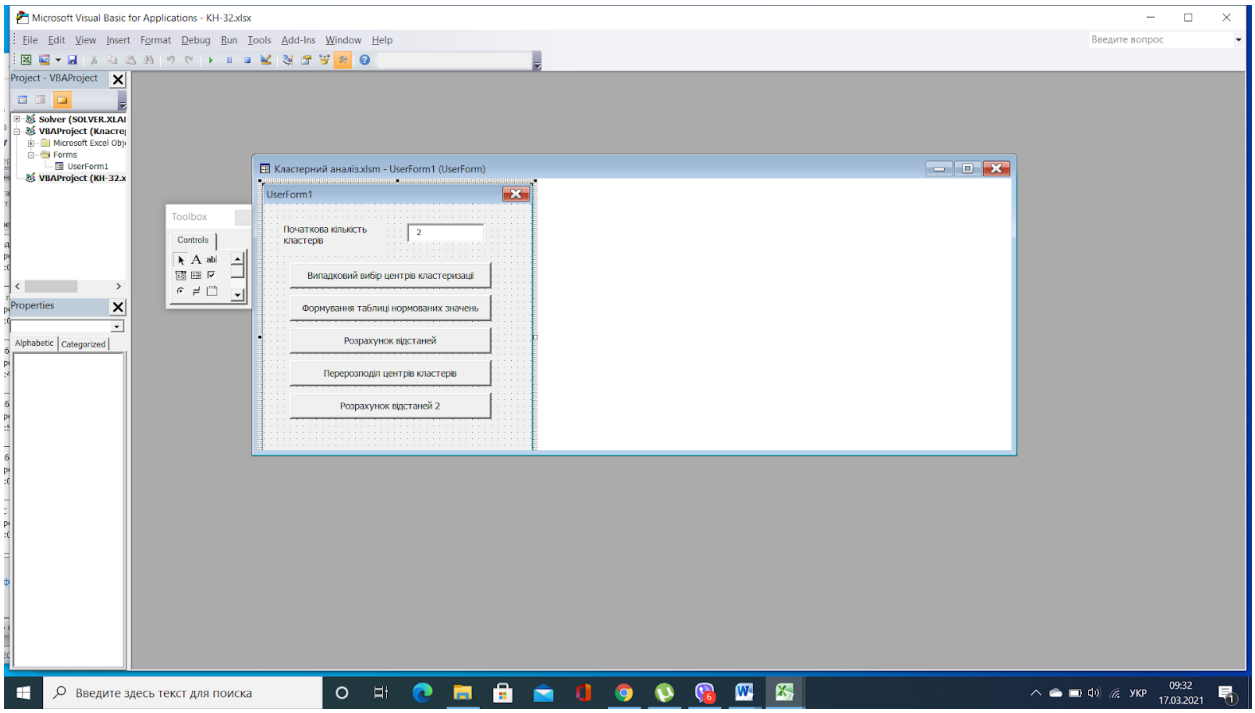
3 Експерименти:

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 організацій 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
Монреаль	1143,6	1865	13,1	13,1	19410	0,3	0,009	22,7	0,02	0,088

Мехико	1130,3	2258,4	12,2	12,2	20510,7	0,312	0,009	22,6	0,026	0,087
Шанхай	1089,9	3127,6	9,9	12,4	21053,8	0,276	0,0044	22,3	0,019	0,078
Токіо	1062,3	1518,4	14	11,9	22089,5	0,306	0,0109	20,5	0,021	0,093
Ріо-де-Жанейро	1021,2	1791,6	10,3	14,2	18294	0,286	0,0082	21,2	0,017	0,081
Буенос-Айрес	991,5	1239,4	12,8	13	22678,8	0,31	0,007	21,3	0,0293	0,093
Ханой	973,9	2754,1	13,5	11	25159	0,291	0,0063	21,7	0,0296	0,086
Сеул	890	1633,1	11	14,6	18178,2	0,327	0,327	24,7	0,014	0,087
Ашхабад	837,8	2192,1	10	15,3	18107	0,3	0,0073	24,7	0,0184	0,155
Ташкент	744,9	991,3	12,3	11,5	22587,8	0,353	0,006	26,3	0,018	0,102

Кількість об'єктів **17** Кількість критеріїв **10**

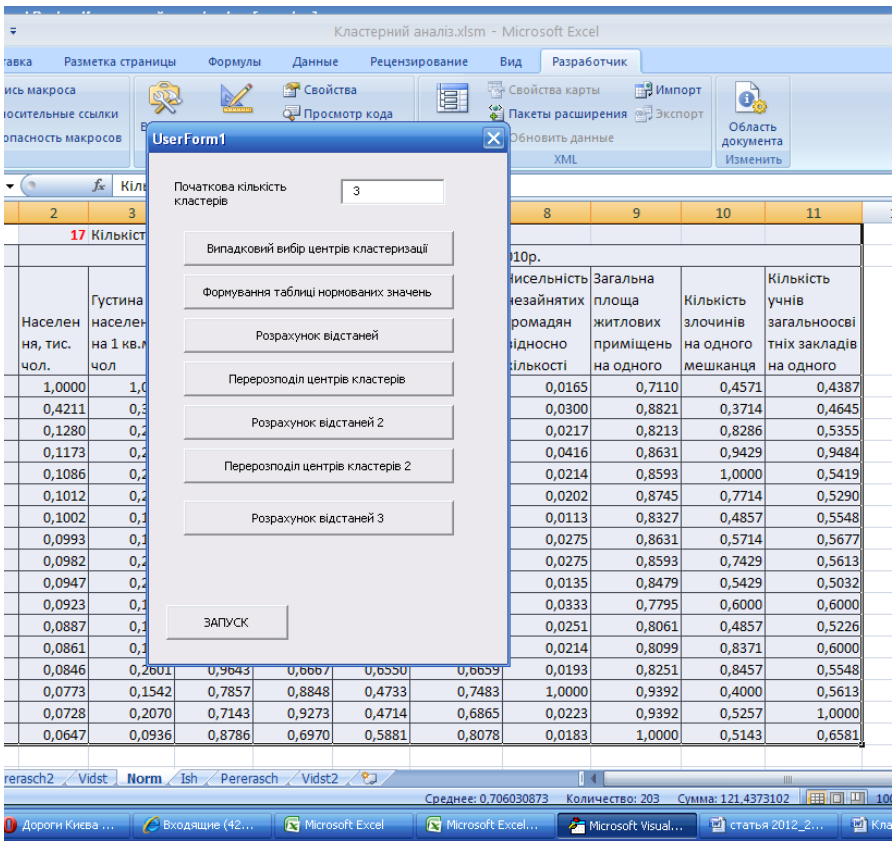
	Показники за жовтень 2010р.									
Місто	Населення, тис. чол.	Густина населення на 1 кв.м., чол	Кількість народившихся на 1000 чол.	Кількість померлих на 1000 чол.	Середньомісячна з/п, руб.	Середньорічна чисельність робітників організацій відносно загального населення, чол.	Чисельність незайнятих громадян 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	1,0000	1,0000	0,7643	0,6606	1,0000	1,0000	0,0165	0,7110	0,4571	0,4387
Лондон	0,4211	0,3287	0,8214	0,8182	0,7079	0,9474	0,0300	0,8821	0,3714	0,4645
Париж	0,1280	0,2784	0,8857	0,7515	0,6086	0,6522	0,0217	0,8213	0,8286	0,5355
Відень	0,1173	0,2351	0,9286	0,7091	0,6044	0,7391	0,0416	0,8631	0,9429	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
Мехико	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
Мехико	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 померлих на 1000 чол.	Середньомісячна 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 померлих на 1000 чол.	Середньомісячна з/п, руб.	Середньорічна чисельність робітників в організації відносно	Чисельність незайнятих громадян відносно кількос	Загальна площа житлових приміщень на одного	Кількість злочинів на одного мешканця	Кількість учнів загальноосвітніх закладів на одного мешканця

					загального населення, чол.	ті насел., чол.	чол., кв.м.		
0,072762	0,103514	0,238095	0,231818	0,094282	0,114416	0,003189	0,117395	0,058413	0,1
0,095787	0,100669	0,286054	0,197727	0,110975	0,118149	0,013287	0,106973	0,075941	0,058848
0,710551	0,332165	0,264286	0,184848	0,170787	0,162281	0,00332	0,099572	0,046032	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 Стьюдента і хи-квадрат

Загальні відомості

https://sites.google.com/site/ktnoscience/Home/lecture/l6

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?

Практична робота № 6-7

Тема роботи: «Мова R»

Завдання 1. 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:

lr6.R

Source on Save

```
conc <- c(9825, 4964, 2480, 1532, 1123, 683, 670, 638)
country <- c("Китай", "США", "Индия", "Россия", "Япония", "Германия", "Иран","Корея")
df <- data.frame(country, conc)
```

lr6.R × df ×

Filter

	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:

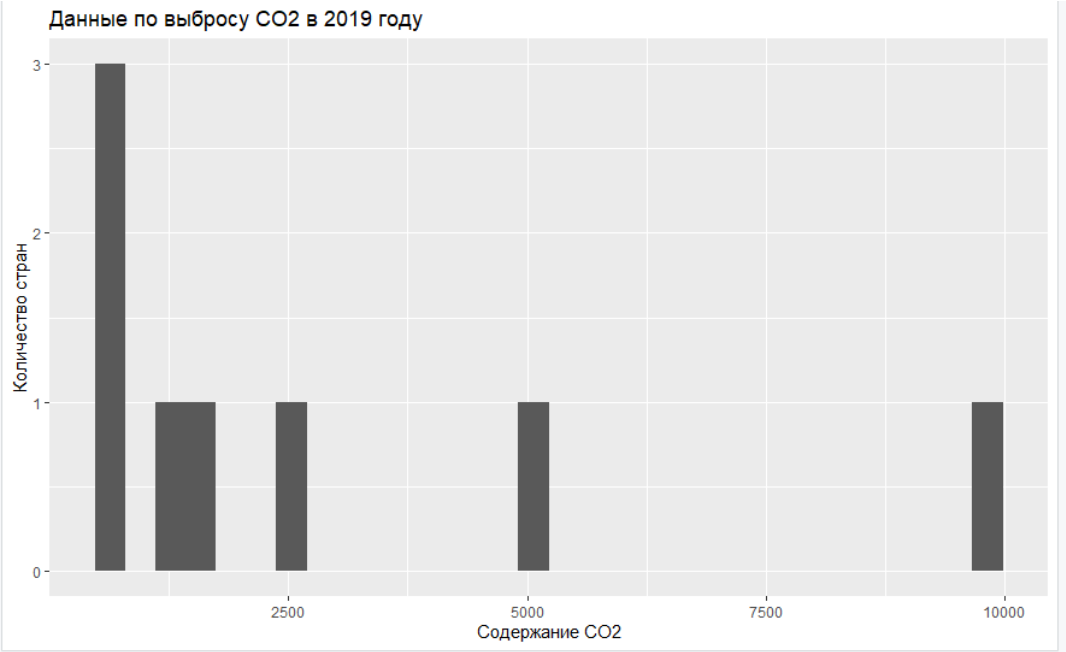
```
11  # новый вектор
12  ser <- mean(df$conc, na.rm = T)
13  s <- conc
14  s <- replace(s, conc == ser, 0)
15  s <- replace(s, conc > ser, 1)
16  s <- replace(s, conc < ser, -1)
```

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

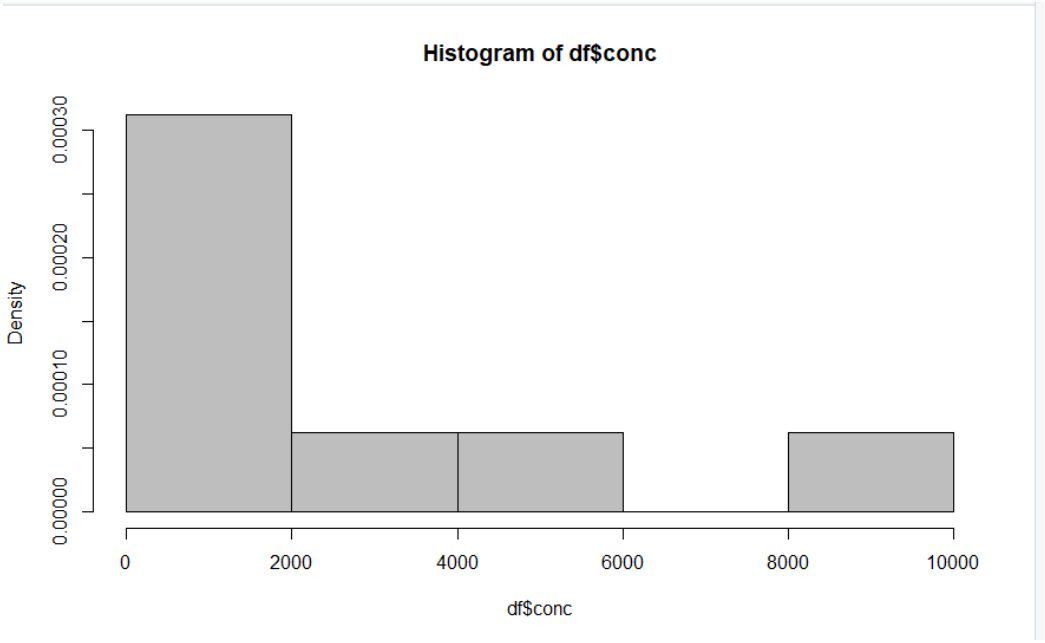
Виведемо описову статистику для змінної conc:

```
> 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:



Результуючий файл lr6.R:

lr6.R df

Source on Save Run Source

```
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)
```

Завдання 2. Лінійна регресія

Завантажимо набір даних для свого варіанту та ознайомимося з ним:

RStudio

File Edit Code View Plots Session Build Debug Profile

Go to file/function | Addins

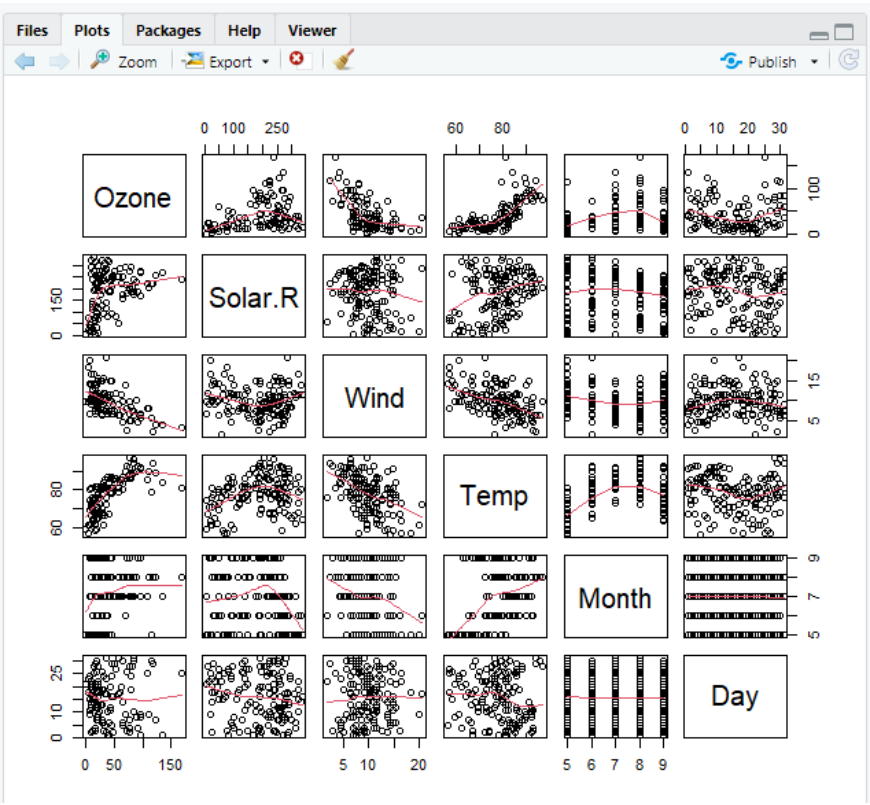
lr6.R | lr6_2.R | d

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

Побудуємо графік кореляційного поля для кожного фактору:



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
>
```

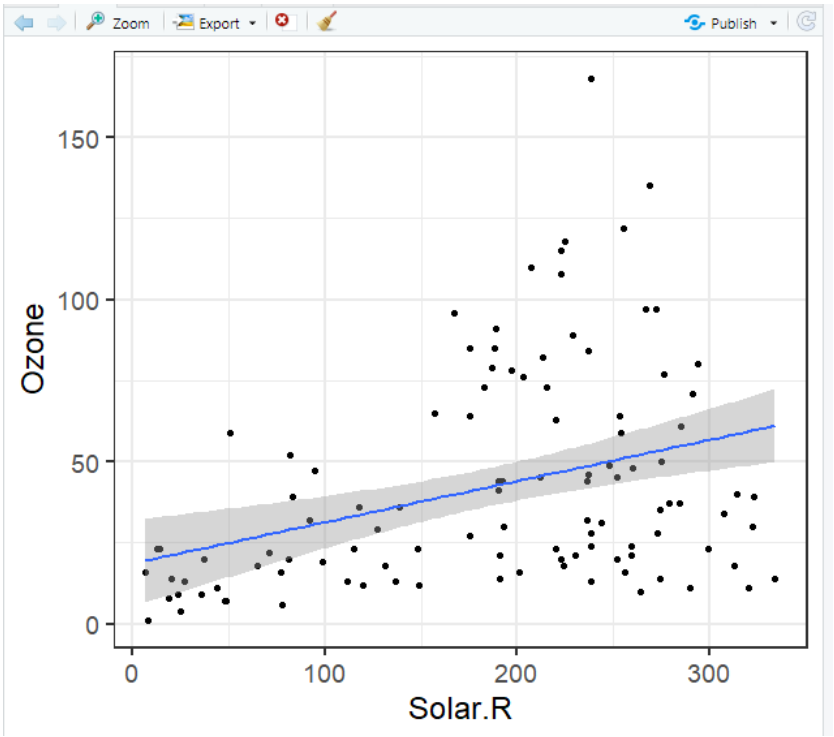
Перевіримо значущість кожного з отриманих рівнянь регресії:

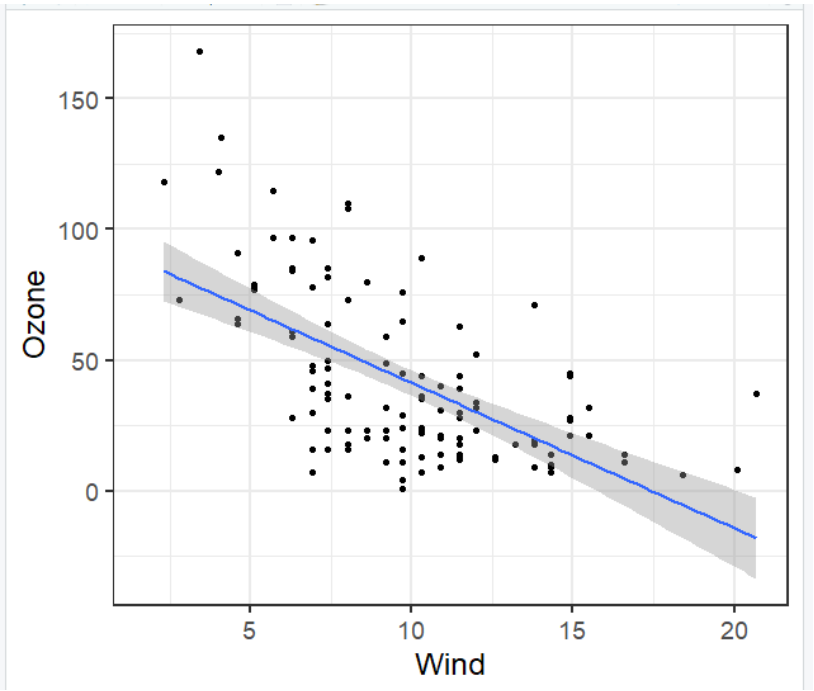
```
> summary(model1)

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
```

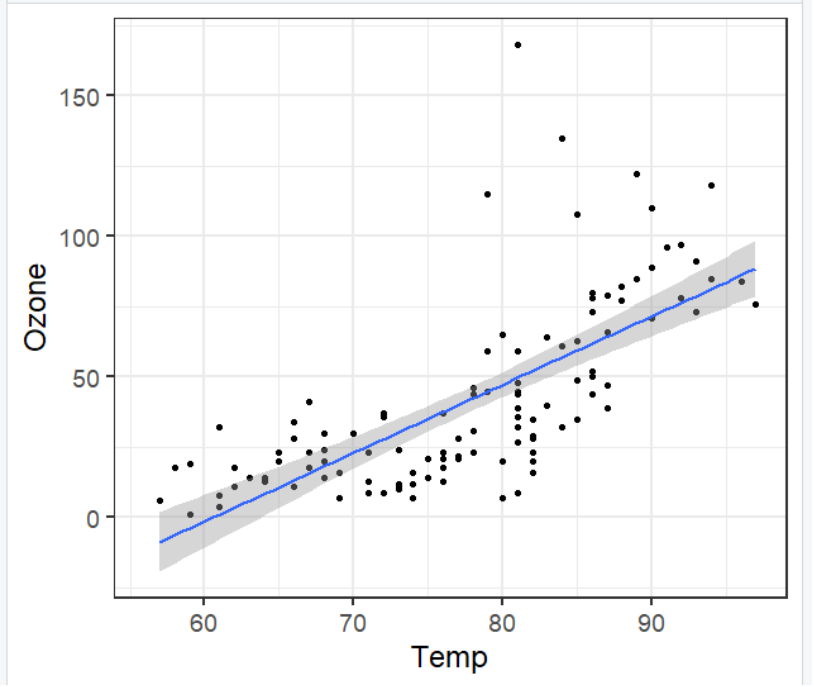
```
> summary(model2)

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
```

Відобразимо рівняння регресії із заданим довіреним інтервалом на графіках:







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

```
> # Построить прогнозы по каждому из уравнений парной регрессии для заданных в варианте значений факторов
> 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
```

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        1.000      NA         NA         NA         NA           NA
Solar.R         NA       1         NA         NA         NA           NA
Wind            NA      NA  1.0000000 -0.4579879 -0.1782926  0.027180903
Temp            NA      NA -0.4579879  1.0000000  0.4209473 -0.130593175
Month           NA      NA -0.1782926  0.4209473  1.0000000 -0.007961763
Day             NA      NA  0.0271809 -0.1305932 -0.0079618  1.000000000
>
```

Побудуємо прогноз по рівнянню множинної регресії для заданих значень факторів:

```
> # Построить прогноз по уравнению множественной регрессии для заданных в варианте значений факторов
> predict(model1, data.frame(Solar.R = 350, Wind = 8.9, Temp = 80))
       1
63.10657
```

Результуючий файл lr6_2.R:

lr6.R | lr6_2.R | d

Source on Save

```
library("psych")
library("lmtest")
library("ggplot2")
library("dplyr")
library("MASS")
library(car)

# Загрузить набор данных для своего варианта, ознакомиться с его содержимым
d <- airquality

# Построить график корреляционного поля для каждого фактора
pairs(d, panel=panel.smooth)

# Построить уравнение парной линейной регрессии для каждого фактора
model1 <- lm(data=d, Ozone~Solar.R)
model2 <- lm(data=d, Ozone~Wind)
model3 <- lm(data=d, Ozone~Temp)

model1$coefficients
model2$coefficients
model3$coefficients

# Проверить значимость каждого из полученных уравнений регрессии
summary(model1)
summary(model2)
summary(model3)
# Показать уравнения регрессии с заданным в варианте доверительным интервалом на графиках
qplot(data = d, Solar.R, Ozone) + stat_smooth(method="lm", level = 0.95) + theme_bw(base_size = 18)
qplot(data = d, Wind, Ozone) + stat_smooth(method="lm", level = 0.95) + theme_bw(base_size = 18)
qplot(data = d, Temp, Ozone) + stat_smooth(method="lm", level = 0.95) + theme_bw(base_size = 18)

# Построить прогнозы по каждому из уравнений парной регрессии для заданных в варианте значений факторов
predict(model1, data.frame(Solar.R = 350))
predict(model2, data.frame(Wind = 8.3))
predict(model3, data.frame(Temp = 80))

# Построить уравнение множественной линейной регрессии и получить корреляционную матрицу
model <- lm(data=d, Ozone~Solar.R+Wind+Temp)
model$coefficients

cor(d)

# Построить прогноз по уравнению множественной регрессии для заданных в варианте значений факторов
predict(model1, data.frame(Solar.R = 350, Wind = 8.9, Temp = 80))
```

Практична робота № 8

Побудова траєкторії руху робота. Алгоритми навігації.

Аналіз існуючих датчиків для побудови карт перешкод, необхідних для розрахунку траєкторій роботизованих платформ, показав, що датчик LiDAR (Light Detection and Ranging) має значну перевагу завдяки огляду на 360 градусів, що дозволяє виявляти перешкоди навколо роботизованої платформи. платформи і наносять їх на карту з точністю до кількох сантиметрів (рис. 6). Датчик LiDAR дозволяє роботизованій платформі виявляти перешкоди, генеруючи великі обсяги даних, включаючи дані на основі точок або пікселів. Обробка отриманих даних зосереджена на сегментації, класифікації точкових і піксельних зображень, що дозволяє кластеризувати та позиціонувати.

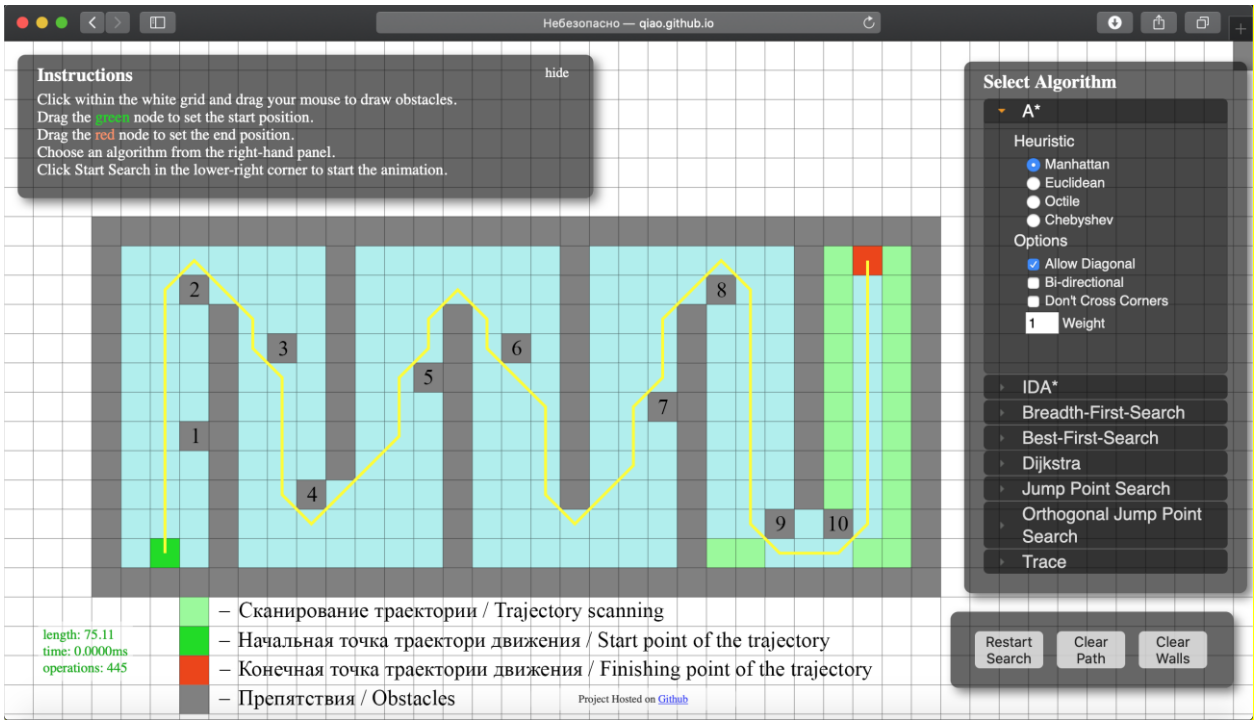
Аналіз існуючих систем керування автономними мобільними енергетичними об’єктами показав, що для навігації та керування в польових умовах можна використовувати алгоритми, наведені в табл. 1.

Таблиця 1 - Алгоритми навігації та управління

No.	**Назва алгоритму**	**Суть методу**
1	Алгоритм помилок	Обхід перешкоди, обхід перешкоди до тих пір, поки не стане можливим досягти наміченої мети
2	Алгоритм навігації	Використання глобальних супутникових систем
3	Гібридний алгоритм навігації	Використання глобальних супутникових систем у поєднанні з активними або пасивними датчиками
4	Планування шляху та алгоритм проходження шляху	Виконання маневрів через визначені маршрутні точки, перетинаючи графік і знаходячи оптимальний шлях
5	Алгоритм Дейкстри	Уніфікований пошук вартості
6	Алгоритм векторного поля гістограми	Гістограма векторного поля, представлення перешкод за допомогою сітки гістограми
7	Алгоритм потенційного поля	Застосування штучного потенційного поля для позиціонування
8	Візуальний алгоритм	Порівняння різниці між двома послідовними зображеннями, отриманими в заданий інтервал часу, з використанням камер комп'ютерного зору
9	Алгоритм переслідування	Розрахунок найкоротшої відстані, побудова траєкторії з урахуванням «кривини»
10	Векторний алгоритм переслідування	Метод слідування шляху з використанням теорії гвинта

11	Алгоритм розпізнавання ознак за допомогою обробки зображень	Застосування нейронних мереж для обробки зображень
12	Алгоритми SLAM	Одночасна локалізація та картографування, побудова карти в невідомому просторі

Приклад пошуку шляху в графі з перешкодами (1-10) за допомогою алгоритму A* в веб-сервісі Github PathFindings показано на рис. 1.



Завдання

1) Обрати промислового або сільськогосподарського робота та вивчити його траєкторію руху.
2) Запустити ресурс http://qiao.github.io/PathFinding.js/visual/
3) Нарисувати начальну, кінцеву точки та перешкоди
4) Обрати алгоритм та встановити потрібні параметри.
5) Виконати пошук шляху в графі з перешкодами.
6) Порівняти результати пошуку різними 3-ма алгоритмами
7) Зробити висновок.

Практична робота № 9

Підготовка даних при формуванні датасету (набору даних) для навчання згорткової нейронної мережі

Завдання

1) Вивчити технологію розмітки зображень для підготовки даних (приклад на відео за посиланням https://drive.google.com/file/d/15LH_fKzTRLpETw1IcIvnlpSmM2VfAWqg/view?usp=share_link
2) Запустити ресурс http://cvat.ai
3) Обрати папку з зображеннями в архіві відповідно до своєї групи і свого варіанту посиланням: https://drive.google.com/drive/folders/1rwTgNvajngQ3ew-ZKN1ZLL5i_THLqhaD?usp=share_link
4) Виконати розмітку зображень
5) Зберегти розмічені зображення в папці, а також файл json і відправити архів
6) Навчити згорткову нейромережу за допомогою ресурсу https://roboflow.com/
7) Розпізнати реальний об’єкт за допомогою камери, порахувати кількість розпізнаних об’єктів і порівняти з реальною кількістю

Практична робота № 10-11.

Формування масиву даних для аналітичної обробки

Мета практичної роботи, що відповідає розділу, – дослідження та збір «великих даних» у визначених форматах із соціальних мереж (RSS-фідів), збереження та підготовка їх для подальшої аналітичної обробки.

Розвиток напрямку обробки надвеликих масивів даних (Big Data) насам перед пов‘язаний з розвитком інформаційної складової мережі Інтернет, а саме, веб-ресурсів і соціальних мереж. Саме такий контент розглядається як своєрідний полігон для проведення практичних робіт в рамках курсу, що вивчається. Для подальшої інтелектуальної обробки таких інформаційних ресурсів необхідно на першому етапі створити систему її збирання, але вирішення цієї цікавої задачі виходить за межі навчального курсу.

Саме тому, для первинного збирання даних для її подальшої інтелектуальної обробки обмежимось лише одним видом неструктурованої текстової інформації, яка вільно доступна в мережі Інтернет і є, мабуть, однією із найбільш стандартизованих, а саме інформацією, наведеною в форматі RSS.

У кінці XX століття для вирішення завдання уніфікації контенту мережі Інтернет з метою її подальшого оброблення програмними застосунками, узагальнення і подальшого розподілу (синдикації) було створено декілька форматів опису даних на основі стандарту XML. Найпоширеніший формат отримав назву RSS, що означає Really Simple Syndication, Rich Site Summary, хоча спочатку він називався RDF Site Summary. Зміст всіх цих абревіатур полягає в простому способі узагальнення та розподілу інформаційного наповнення веб-сайтів – синдикації контенту.

Розробка RSS, почалася 1997 року. Визнання ця технологія здобула,

коли компанія Netscape використовувала її для наповнення каналів свого порталу Netcenter. Незабаром технологія RSS стала використовуватися для

трансляції контенту на багатьох новинних сайтах – у тому числі таких як BBC, CNET, CNN, Disney, Forbes, Wired, Red Herring, Slashdot, ZDNet та багатьох інших. Першою відкритою офіційною версією RSS стала версія

0.90. Формат був заснований на RDF (Resource Description Framework – стандарт схеми опису ресурсів).

Сьогодні практично всі провідні веб-сайти, блоги, деякі соціальні мережі, що працюють в Інтернет, використовують RSS як інструмент оперативного представлення своїх оновлень. Специфікації окремих версій формату RSS наведені на таких веб-сторінках:

RSS 0.90: http://www.purplepages.ie/RSS/netscape/rss0.90.html

RSS 0.91: http://my.netscape.com/publish/formats/rss-spec-0.91.html

RSS 1.0: http://web.resource.org/rss/1.0/

RSS 2.0: http://backend.userland.com/rss/

Термін «RSS» (Really Simple Syndication) часто використовується як загальна назва усіх веб-стрічок, включаючи ті, що мають формат, відмінний від RSS (наприклад, для стрічок у форматі Atom).

Формат даних

RSS базується на стандарті XML. Нижче наведено основні відомості щодо RSS 2.0. Перший тег в RSS-документі обов‘язково вказує на формат XML, що застосовується. Після нього йде тег <rss> з обов‘язковим атрибутом version, який вказує на версію документа.

У обов‘язковому порядку RSS-документ містить 2 теги: <channel> і <item>.

Основну інформацію про цей RSS-канал містить тег <channel>. Він зустрічається лише 1 раз.

У обов‘язковому порядку він містить такі три теги:

<title> – ім'я каналу. Може збігатися із ім'ям сайту.

<link> – Посилання на сайт, з яким пов'язаний канал.

<description> – опис каналу.

Опціональні теги:

<language> – мова стрічки.

<copyright> – авторські права.

<managingEditor> – електронна пошта редактора вмісту каналу

<webMaster> – електронна пошта веб-майстра каналу

<pubDate> – дата публікації каналу.

<lastBuildDate> – дата останньої зміни вмісту каналу.

<category> – категорії контенту каналу.

<generator> – програма, за допомогою якої було згенеровано канал.

<docs> – посилання на документацію використовуваного формату RSS.

<ttl> – час актуальності каналу в хвилинах.

<image> – зображення, яке відображається з каналом. У свою чергу, тег має такі параметри:

<Title> – заголовок.

<Description> – опис (аналог тега ALT в HTML).

<Link> – посилання на сайт, з яким пов'язаний канал.

<URL> – адреса зображення.

<Width> – ширина зображення.

<Height> – висота зображення.

<skipHours> – скільки годин не вимагати оновлення з каналу

<skipDays> – скільки днів не вимагати оновлення з каналу

Тег <item> містить інформацію про публікацію.

Обов'язкові вкладені теги:

<title> – назва публікації.

<link> – посилання на сторінку з повним текстом публікації.

<description> – короткий текст публікації.

Необов'язкові вкладені теги:

<Author> – електронна пошта автора

Приклад:

<Category> – категорія повідомлення

<Comments> – посилання на сторінку з коментарями до повідомлення

<Enclosure> – приєднаний мультимедійний об'єкт. Його параметри:

<URL> – адреса об‘єкта

<Length> – розмір об‘єкта в байтах

<Type> – MIME-тип файлу

<Guid> – ідентифікатор повідомлення

<PubDate> – дата публікації.

```
<?xml version="1.0" encoding="windows-1251" ?>
<rss version="2.0">
<channel>
<title>Мої новини</title>
<link>http://site.ua</link>
<description>Опис моїх новин</description>
<image>
<url>http://site.ua /moiLogotip.gif</url>
<link>http://site.ua </link>
<title>Мої новини</title>
</image>
<lastBuiltDate>26 feb 2021 11:11:11 +0300</lastBuildDate>
<item>
<title>З‘явилась нова публікація!</title>
<link>http://site.ua /?str=news</link>
<description>Сьогодні з‘явилась нова стаття щодо «великих
даних»</description>
<pubDate>26 feb 2021 11:11:11 +0300</pubDate>
<guid> http://site.ua /?str=news</guid>
</item>
```

</channel>

</rss>

RSS-фіди

Масиви документів, які надаються на ресурсах веб-серверів у форматі RSS мають назви «RSS-фіди» (від англ. Feed – годувати, постачати), канали інформації. Адреси цих фідів задаються у явному вигляді на сторінках веб сайтів (стандартне позначення –), або їх можна знайти у вихідному коді HTML сторінок, наприклад, як у фрагменті першої сторінки веб-сайту агентства новин УНІАН:

```
<a class="footer-menu_icon " target="_blank" rel="noopener" aria-label="rss" href="https://rss.unian.net/site/news_rus.rss">
<i class="unian-rss"></i></a>
```

Процедури завантаження даних

Для автоматичного завантаження RSS-фіди зазвичай можна використовувати стандартні утиліти завантаження інтернет-контенту, яких існує дуже багато, наприклад програми cURL або wget.

cURL — це утиліта для організації вибірки даних з веб-сайтів, яка надає можливість оперувати з файлами на боці веб-серверу за допомогою параметрів, що можуть бути переданими в рядку URL. За допомогою cURL можна отримувати веб-сторінки, не використовуючи для цього браузера. Крім HTTP-запитів, cURL підтримує SMTP, IMAP, Telnet, FTP та інші мережні протоколи. Базове використання cURL полягає у простому наборі у командній консолі команди cURL, за якою іде URL для завантаження. cURL за замовчуванням відображає вивід отриманого у стандартний потік виводу системи, тобто виклик cURL покаже програмний код сторінки в вікні терміналу.

Програма cURL може записати вивід до файлу при завданні опції «-o»:

```
curl -o example.html www.example.com
```

Такий виклик забезпечує збереження коду титульної сторінки www.example.com у файлі example.html.

Wget – це консольна утиліта для завантаження файлів за протоколами HTTP, HTTPS та FTP. wget дає змогу рекурсивно завантажувати мережеві файли, копіювати як окремі сторінки, так і цілі веб- сайти, конвертувати посилання тощо. Ця утиліта портована й запускається на

багатьох UNIX-подібних системах, Microsoft Windows, MacOS X тощо. wget — не інтерактивна програма, тобто, після того, як її запущено з певними параметрами, вона виконує всі необхідні дії і не потребує додадкового втручання у свою роботу. Wget може працювати як пошуковий робот, тобто отримувати ресурси, на які посилаються елементи HTML сторінки, й рекурсивно просуватися web-деревом, доки всі необхідні файли не будуть завантаженні.

Процедури скачування RSS-фідів (каналів інформації) мають вигляд:
Або

curl –o habr https://habr.com/ru/rss/all/all/

wget –O habr https://habr.com/ru/rss/all/all/

Нижче наведено список RSS-фідів, які можна використовувати для формування масиву даних для подальших практичних робіт.

http://economictimes.indiatimes.com/tech/software/rssfeeds/13357555.cms

http://nypost.com/tech/feed
http://www.smh.com.au/rssheadlines/technology-news/article/rss.xml http://zeenews.india.com/rss/science-technology-news.xml

http://www.washingtontimes.com/rss/headlines/culture/technology/
https://www.thehindu.com/sci-tech/technology/?service=rss
https://www.economist.com/science-and-technology/rss.xml

https://feeds.skynews.com/feeds/rss/technology.xml
http://www.innertemplelibrary.com/category/secure-hospitals/feed/
http://e-news.com.ua/export/news.xml http://fakty.ua/rss_feed/science

Приклад фіду:

<rss version="2.0">

<channel>

<title>Наука</title>

<link>https://fakty.ua/rss_feed/science</link>

<description>Новини науки та технологій</description>

<language>uk</language>

<image>

<link>https://fakty.ua/images/</link>

<url>https://fakty.ua/images/logo_uk.png</url>

<title>Наука</title>

</image>

<item>

<title>

<![CDATA[

У додатку «Дія» з‘явилася нова функція, яка допоможе українцям у боротьбі із шахраями

]]>

</title>

<link>

<![CDATA[

https://science.fakty.ua/393252-v-prilozhenii-diya-poyavilas-novaya-funkciya-kotoraya-pomozhet-ukraincam-v-borbe-s-moshennikami

]]>

</link>

<description>

<![CDATA[

Вона працює за замовчуванням, оновлювати «Дію» не потрібно

]]>

</description>

<category>

<![CDATA[Наука та технології]]>

</category>

<guid>

https://science.fakty.ua/393252-v-prilozhenii-diya-poyavilas-novaya-funkciya-kotoraya-pomozhet-ukraincam-v-borbe-s-moshennikami
</guid>

<pubDate>Tue, 04 Jan 2022 22:30:00 +0200</pubDate>

</item>

<item>

<title>

<![CDATA[

Спойлери, реакції, QR-коди: Telegram випустив передноворічне оновлення

]]>

</title>

<link>

<![CDATA[

https://science.fakty.ua/392950-spojlery-reakcii-qr-kody-telegram-vypustil-prednovogodnee-obnovlenie
]]>

</link>

<description>

<![CDATA[З‘явилися нові корисні функції]]>

</description>

<category>

<![CDATA[Наука та технології]]>

</category>

<guid>

https://science.fakty.ua/392950-spojlery-reakcii-qr-kody-telegram-vypustil-prednovogodnee-obnovlenie
</guid>

<pubDate>Thu, 30 Dec 2021 16:58:00 +0200</pubDate>

</item>

>Tue, 28 Dec 2021 16:47:00 +0200</pubDate>

</item>

<item>

<title>

<![CDATA[

Недоступні повідомлення та коментарі: у Telegram стався масштабний збій
]]>

</title>

<link>

<![CDATA[

https://science.fakty.ua/392723-nedostupny-soobcsheniya-i-kommentarii- v-telegram-proizoshel-masshtabnyj-sboj
]]>

</link>

<description>

<![CDATA[Користувачі помітили збій увечері 27 грудня]]>

</description>

<category>

<![CDATA[Наука та технології]]>

</category>

<guid>

https://science.fakty.ua/392723-nedostupny-soobcsheniya-i-kommentarii- v-telegram-proizoshel-masshtabnyj-sboj

</guid>

<pubDate>Mon, 27 Dec 2021 20:10:00 +0200</pubDate>

</item>

<item>

…

</channel>

</rss>

Утиліти cURL або wget можна викликати за розкладом, наприклад, використовуючи засоби crontab. При цьому у подальшому на етапі обробки скачаних фалів треба передбачити заборону урахування документів, що мають в мережі однакові адреси (тег <link> у RSS-фіді).

Питання до практичної роботи:

1. Що таке RSS і як він використовується?
2. Що таке RSS канал?
3. З чого складається RSS?
4. У чому різниця між веб-стрічками та RSS?

Завдання на самостійну роботу

1. Розширити список RSS-фідів каналами комп‘ютерної і телекомунікаційної спрямованості (але не торговельними майданчиками).
2. Створити процедуру (скрипт) для періодичного скачування інформації із створеного переліку RSS-фідів.
3. Реалізувати процедуру створення файлу, в якому об‘єднуються всі скачані RSS-фіди, і підключити її до скрипту скачування.

Література

1. Dave Johnson. RSS and Atom in action: web 2.0 building blocks. – Manning Publications, 2006. – 398 pp. ISBN 9781932394498, 1932394494.
2. Michael Schrenk. Webbots, Spiders, and Screen Scrapers: A Guide to Developing Internet Agents with PHP/CURL. – No Starch Press, 2007. – 328 pp. ISBN 9781593271206, 1593271204.
3. https://www.gnu.org/software/wget/

Практична робота № 12

Формування JSON-файлу для завантаження в Elasticsearch

***Мета практичної роботи**, що відповідає розділу, – дослідження та трансформація RSS-форматів «великих даних» у стандартні текстові JSON-файли для подальшого їх завантаження у пошукову систему Elasticsearch для індексування та пошуку будь-яких типів документів.*

У попередньому розділі було сформовано інформаційну основу наступних практичних занять, файли за тематикою, що визначається інформаційними джерелами, завантажені в форматі RSS, який можна вважати обмінним, комунікаційним форматом.

Разом з цим, для подальшої роботи із сучасними пошуковими системами отриманні дані мають бути перетворені до вхідних форматів таких систем. Зокрема, для подальшої роботи із системою Elasticsearch, дані, що мають завантажуватись, повинні бути надані у форматі JSON.

Формат JSON

JSON (англ. JavaScript Object Notation – запис об‘єктів JavaScript) – це текстовий формат обміну даними між комп'ютерами. JSON базується на тексті, може бути прочитаним людиною. Формат дає змогу описувати об‘єкти та інші структури даних. Цей формат використовується переважно для передачі структурованої інформації через мережу.

JSON з‘явився через необхідність обміну даними із сервером у реальному часі без використання плагінів для браузерів, flash-застосунків або Java-аплетів. За рахунок своєї лаконічності в порівнянні з XML, формат JSON є більш придатним для представлення складних структур.

JSON будується на двох структурах:

- Набір пар назва/значення. У різних мовах програмування це

реалізовано як об'єкт, запис, структура, словник, хеш-таблиця, список із ключем або асоціативним масивом.

Впорядкований список значень. У багатьох мовах це реалізовано як масив, вектор, список або послідовність.

Структури даних, що використовуються JSON, підтримуються будь-якою сучасною мовою програмування, що дозволяє застосовувати JSON для обміну даними між різними мовами програмування і програмними системами.

Як значення в JSON можуть бути використані:

- запис – це невпорядкована множина пар «ключ-значення», укладене у фігурні дужки «{ }». Ключ описується рядком, між ним та значенням стоїть символ «:». Пари ключ-значення відокремлюються один від одного комами.
- масив (одномірний) – це впорядкована множина значень. Масив записується у квадратних дужках «[ ]». Значення поділяються комами. Масив може бути порожнім, тобто не містити жодного значення. Значення не більше одного масиву можуть мати різний тип.
- число (ціле або дійсне).
- літерали true (логічне значення «істина»), false (логічне значення «хибне») та null.
- рядок — це впорядкована множина з нуля або більше символів юнікоду, укладена в подвійні лапки. Символи можуть бути вказані з використанням escape-послідовностей, що починаються зі зворотної косої риси "\" (підтримуються варіанти \", \\, \/, \t, \n, \r, \f і \b), або записані шістнадцятковим кодом у кодуванні Unicode у вигляді \uFFFF.

«Рядок» дуже схожий на літерал однойменного типу даних у мові Javascript. «Число» теж дуже схоже на Javascript-число, за винятком того, що використовується лише десятковий формат (з точкою як роздільник).

Пробіли можуть бути вставлені між будь-якими двома синтаксичними

елементами.

Наступний приклад показує JSON-подання даних про об'єкт, що описує людину. У даних присутні рядкові поля імені та прізвища, інформація про адресу та масив, що містить список телефонів. Як видно з прикладу, значення може бути вкладеною структурою.

```
{
  "firstName": "Іван", "lastName":
  "Іванов", "address": {
    "streetAddress": "пров. П.Мирного, 10, кв.101", "city":
    "Суми",
    "postalCode": 10101
  },
  "phoneNumbers": [ "067 123-
    1234",
    "050 123-4567"
  ]
}
```

Треба звернути увагу на пару "postalCode": 10101. В якості значень JSON можуть бути використані як числа, так і рядки. Тому запис "postalCode": "10101" містить рядок, а "postalCode": 10101 – числове значення. Через слабку типізацію в Javascript і PHP рядок може бути приведений до числа і не впливати на логіку програми. Тим не менш, рекомендується акуратно поводитися з типом значення, оскільки JSON служить для міжсистемного обміну.

Система Elasticsearch не нав'язує чітку структуру даних; можна зберігати будь-які документи JSON, які, зокрема, підходять для Elasticsearch на відміну від рядків та стовпців у реляційних базах даних. Такий документ можна порівняти із записом у таблиці БД. У традиційних реляційних базах даних таблиці структуровані: мають фіксовану кількість стовпців, кожен має свій тип і розмір. Але дані

можуть динамічно змінюватися, що вимагатиме підтримки нових або динамічних стовпців. Документи JSON спочатку підтримують такий тип даних.

Наведений нижче ще один приклад показує JSON представлення об'єкта, що описує клієнта:

```
{
"name": "John Smith",
"address": "121 John Street, NY, 10010",
"age": 40
}
```

Так може виглядати запис клієнта. У ній вказано ім'я, адресу та вік клієнта. А інший запис може виглядати так:

```
{
"name": "John Doe", "age": 38,
"email": "john.doe@company.org"
}
```

Треба звернути увагу, що другий клієнт не має поля для адреси, але замість нього є поле для електронної пошти. А в інших клієнтів може бути зовсім інший набір полів. Отже, реалізована гнучкість щодо того, що може зберігатися в таблиці.

Для завантаження в систему Elasticsearch необхідно отриманий на попередній практичній роботі файл перетворити з формату RSS, а саме створити формат JSON. Пропонується такий спрощений перелік полів документа, які надалі мають завантажуватися до системи Elasticsearch:

"title" – назва повідомлення; "textBody" –

текст повідомлення; "source" – джерело

повідомлення;

"pubDate" – дата і час у форматі YYYYMMDD HH:MM

"url" – адреса повідомлення в мережі Інтернет.

Нижче наведено фрагмент файлу в форматі JSON, який має бути отримано:

{

"title":" У додатку «Дія» з'явилася нова функція, яка допоможе українцям у боротьбі із шахраями",

"textBody":" Вона працює за замовчуванням, оновлювати «Дію» не потрібно",

"source":"Факти", "PubDate":"2021-12-28T10:04:00Z",

"URL":"https://science.fakty.ua/393252-v-prilozhenii-diya-poyavilas- novaya-funkciya-kotoraya-pomozhet-ukraincam-v-borbe-s-moshennikami"

}

,

{

"title":"Суд у смартфоні: «Дія» тестує новий сервіс", "textBody":"Через додаток хочуть надсилати повідомлення про судові засідання",

"source":"Факти", "PubDate":"2021-12-28T10:54:00Z",

"URL":"https://science.fakty.ua/392788-sud-v-smartfone-diya-testiruet-novyj-servis"

}

{

"title":"James Webb може переписати космічну історію: відео запуску новітнього телескопу від NASA",

"textBody":"Роботу апарат розпочне влітку 2022 року", "source":"Факти",

"PubDate":"2021-12-25T10:14:00Z",

"URL":" https://science.fakty.ua/392589-james-webb-mozhet-perepisat-kosmicheskuyu-istoriyu-video-zapuska-novogo-teleskopa-ot-nasa"

}

Програма перетворення RSS в JSON

Практичним завданням, що вирішується в рамках цього розділу, є розробка програми мовою Python, яка перетворює (конвертує) інформацію, надану у форматі RSS, до формату JSON, не змінюючи зміст цієї інформації. Нижче наведено приклад такої програми (пояснення надані у вигляді коментарів у тексті програми, після знаку #):

```
#Програма конвертації даних із формату RSS до формату JSON #Підключення модулів для роботи із регулярними виразами і часом import re
import datetime

#Відкриття файлу rss.xml у режимі «читання» f
= open("rss.xml","r")

#Зчитування вмісту файлу rss.xml у змінну t t =
f.read()

#Закриття файлу rss.xml f.close()

#Розбиття файлу по рядкам і склеювання рядків через пропуск rss =t.split('\n')
t=""
for i in range(len(rss)): t=t+" "+rss[i]
```

```
#Видалення перших пропусків
t=re.sub('^\s','',t)

#Формування масиву заголовків
title = re.findall('<title>(.+?)<\/title>', t)

#Перший заголовок – назва фіду, далі – його специфічна обробка
source=title[0]
source=re.sub('[\s\-]*$','',source)
source=re.sub('"','\"',source)

#Розмірність масиву заголовків
x=range(len(title))

#Формування масиву описів
text = re.findall('<description>(.*?)<\/description>', t)

#Формування масиву гіперпосилань
link = re.findall('<link>(.+?)<\/link>', t)

#Формування дати і часу в форматі "YY-MM-
MMTHH:MM:00Z" now = datetime.datetime.now()
tim=now.strftime("%Y-%m-%dT%H:%M:00Z")

#Виведення результатів
for i in range(1, len(title)):
    print "{\n\"title\":\""+title[i]+"\","

    #Специфічна обробка тексту
    text[i]=re.sub('[\s\-]*$','',text[i])
    text[i]=re.sub('"','\"',text[i])
    text[i]=re.sub('\'','&amp;',text[i])
```

```
#Подальша виведення результатів
print "\"textBody\":\""+text[i]+"\","
print "\"source\":\""+source+"\","
print "\"PubDate\":\""+tim+"\","
print "\"URL\":\"",link[i],"\"\n}"

if i<len(title)-1: print ","
```

У результаті виконання наведеної програми має сформуватися JSON- файл, придатний для завантаження у середовище інформаційно-пошукової системи Elasticsearch.

Питання до практичної роботи:

1. Що таке JSON–формат?
2. Які основні переваги JSON–формату?
3. Які ви знаєте правила створення структури JSON–файлу в об‘єкті, масиві і при присвоєнні значення?
4. Що таке пошукова система Elasticsearch та її призначення?

Завдання на самостійну роботу

1. Написати програму формування пакетного файлу в форматі JSON.
2. Встановити на комп‘ютері бібліотеку для роботи із регулярними виразами у середовищі мови програмування (Python).
3. Ознайомитися із основними можливостями мови програмування Python щодо роботи із строковими даними.

Література

1. Основы Data Science и Big Data. Python и наука о данных / Дэви Силен, Арно Мейсман. – Питер, 2017. – 336 с.
2. Пранав Шукла, Шарат Кумар. Elasticsearch, Kibana, Logstash и поисковые системы нового поколения. – Питер, 2019. – 363 с.

3. Бонцанини М. Анализ социальных медиа на Python / пер. с анг. А. В. Логунова. – М.: ДМК Пресс, 2018. – 288 с.
4. С. Шапошникова. Основы программирования на Python. Вводный курс. – Лаборатория юного линуксоида, 2011. – 44 с.

Практична робота № 14

Побудова регресійної моделі та визначення точності апроксимації і коефіцієнта кореляції

Мета роботи: навчитися апроксимувати експериментальні дані однією з функціональних залежностей із знаходженням рівняння цієї функції, визначати коефіцієнт детермінації, кореляції та здійснити прогнозування.

Теоретичні відомості

Регресійна модель

Одним із поширених видів досліджень є визначення типу функціонального зв'язку між ознаками, які знаходяться у певній причинно-наслідковій залежності між собою. Іншими словами, при наявності експериментальних значень показника (y) та фактора (x), який викликає зміну показника, необхідно визначити тип функціональної залежності $y = f(x)$, яка б відображала закономірності розвитку явища.

Послідовність знаходження рівняння (математичної моделі), яка б описувала експериментальні точки передбачає такі етапи: виявлення факторів, що суттєво впливають на показник, вибір типу рівняння регресії і знаходження параметрів даного виду рівняння, аналіз адекватності побудованої моделі.

Прикладом припущень щодо зв'язку між ознаками можуть бути рівняння:

$y = a + b \cdot x$ - лінійна залежність;

$y = a + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3 + ... + b_6 \cdot x^6$ - поліноміальна залежність;

$y = a + b \cdot \ln x$ - логарифмічна залежність;

$y = a \cdot x^b$ - степенева залежність;

$y = a \cdot e^{b \cdot x}$ - експоненціальна залежність.

Вид відповідного рівняння обирається шляхом побудови графічної залежності між експериментальними даними і проведення аналізу їх розміщення. Тип графічної залежності у даному випадку вибирається *"Точечная"!*.

Довідкова інформація за технологією роботи з режимом "Регрессия" пакету аналізу MS Excel

Режим роботи "Регрессия" служить для розрахунку параметрів рівняння лінійної регресії і перевірки його адекватності досліджуваному процесу.

Для вирішення завдання регресійного аналізу в MS Excel вибираємо в меню **Сервис** команду **Анализ данных** і інструмент аналізу "**Регрессия".**

У діалоговому вікні, що з'явилося, задаємо наступні параметри:

1. *Входной интервал* *Y* - це діапазон даних за результативною ознакою. Він повинен складатися з одного стовпця.

2. *Входной интервал* *X* - це діапазон комірок, що містять значення факторів (незалежних змінних). Число вхідних діапазонів (стовпців) має бути не більше 16.

3. Прапорець *Метки*, встановлюється втом випадку, якщо в першому рядку діапазону міститься заголовок.

4. Прапорець *Уровень надежности* активізується, якщо в поле, що знаходиться поряд з ним необхідно ввести рівень надійності, відмінний від встановленого за замовчуванням. Використовується для перевірки значущості коефіцієнта детерміації R^2 і коефіцієнтів регресії.

5. *Константа ноль.* Даний прапорець необхідно встановити, якщо лінія регресії повинна пройти через початок координат (a_0=0).

6. *Выходной интервал/ Новый рабочий лист/ Новая рабочая книга* вказати адресу верхнього лівого осередку вихідного діапазону.

7. Прапорці в групі *Остатки* встановлюються, якщо необхідно включити у вихідний діапазон відповідні стовпці або графіки.

8. Прапорець *График нормальной вероятности* необхідно зробити активним, якщо потрібно вивести на лист точковий графік залежності спостережуваних значень Y від автоматично відформатованих інтервалів.

Після натиснення кнопки ОК у вихідному діапазоні отримуємо звіт.

Знаходження параметрів рівняння

Для знаходження параметрів рівняння (математичної моделі), яка б описала дану залежність необхідно:

- виділити побудований графік (один раз клацнути лівою клавішею миші по одній з точок графіка);
- клацнути правою клавішею миші по виділеній лінії графіка;
- у меню, що з'явиться, вибрати пункт **"Добавить линию тренда"**;
- у діалоговому вікні, що з'явиться, у закладці **"Тип"** вибрати одну із запропонованих ліній функціональних залежностей, якою планується апроксимація експериментальних точок;
- у закладці **"Параметри"** відмітити пункт **"показывать уравнение на диаграме"**;
- натиснути кнопку **"Ок"**;
- на графіку з'явиться апроксимуюча лінія з рівнянням, за яким вона побудована.

Аналіз адекватності побудованої моделі

Графічне зіставлення експериментальних точок з лінією, побудованою на основі вибраного рівняння не завжди забезпечує правильний вибір функції для відображення основної тенденції зв'язку між ознаками внаслідок схожості зовнішнього вигляду ліній рівнянь функціональних залежностей.

Для вибору найбільш придатного виду рівняння, тобто для аналізу адекватності побудованої моделі використовується коефіцієнт детермінації R^2, що також називається квадратом коефіцієнта множинної кореляції R. R^2 (міра визначеності) завжди знаходиться в межах інтервалу [0;1].

Якщо значення R^2 близьке до одиниці, це означає, що побудована модель пояснює майже всю множину відповідних змінних. І навпаки, значення R-квадрата, близьке до нуля, означає погану якість побудованої моделі.

Коефіцієнт детерміації R^2 показує, на скільки відсотків ($R^2 \cdot 100\%$) знайдена функція регресії описує зв'язок між початковими значеннями чинників X і Y

$$R^2 = \frac{\sum_{i=1}^{n}\left(y_i^{\delta} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

де $\left(y_i^p - \bar{y}\right)^2$ - пояснена варіація; $\left(y_i - \bar{y}\right)^2$ - загальна варіація.

Відповідно, величина $(1-R^2)\cdot 100\%$ показує, скільки відсотків варіації параметра Y обумовлено чинниками, не включеними в регресійну модель. При високому ($R^2 \geq 75\%$) значенні коефіцієнта детерміації можна робити прогноз $y^* = f(x^*)$ для конкретного значення x^*.

Зіставлення коефіцієнтів детермінації для різних типів рівнянь дає змогу визначити найпридатніше для апроксимації експериментальних даних. Проте не завжди вибирається той вид рівняння, для якого R^2 буде більшим.

Для визначення коефіцієнта детермінації а також зміни рівняння необхідно:

- виділити апроксимуючу лінію;
- натиснути праву клавішу миші на лінії, в меню, що з'явиться, вибрати пункт **"Формат линии тренда"**;
- у діалоговому вікні, що з'явиться, у закладці **"Тип"** вибрати іншу функціональну залежність;
- у закладці **"Параметри"** відмітити пункт **"поместить на диаграмму величину достоверности аппроксимации"**;
- натиснути кнопку **"Ок"**.

Для проведення регресійного аналізу і прогнозування необхідно:

1) ***побудувати графік*** початкових даних і спробувати, приблизно визначити характер залежності;

2) ***вибрати вид функції*** регресії, яка може описувати зв'язок початкових даних;

3) ***визначити чисельні коефіцієнти*** функції регресії методом найменших квадратів;

4) оцінити силу знайденої регресійної залежності на основі коефіцієнта детермінації R^2;

5) зробити прогноз (при $R^2 \geq 75\%$) або зробити висновок про неможливість прогнозування за допомогою знайденої регресійної залежності. При цьому не рекомендується використовувати модель регресії для тих значень незалежного параметра X, які не належать інтервалу, заданому в початкових даних.

Приклад виконання практичної роботи

Завдання: Деяка фірма займається постачаннями різних вантажів на короткі відстані усередині міста. Оцінити вартість таких послуг в залежності від часу, що витрачається на постачання. Як найбільш важливий чинник, що

впливає на час постачання, вибрана пройдена відстань. Були зібрані початкові дані про десять постачань (таблиця 1)

Таблица 1

Відстань, км	3,5	2,4	4,9	4,2	3,0	1,3	1,0	3,0	1,5	4,1
Час, хв	16	13	19	18	12	11	8	14	9	16

Визначите характер залежності між відстанню і витраченим часом, використовуючи майстер діаграм MS Excel, проаналізуйте якість застосування методу найменших квадратів (МНК), побудуйте рівняння регресії, використовуючи МНК, проаналізуйте силу регресійного зв'язку. Проведіть регресійний аналіз, використовуючи режим роботи "**Регрессия**" в MS Excel і порівняєте з результатами, отриманими раніше. Зробіть прогноз часу поїздки на 2 км. Порахувати і побудувати графічно міру помилки регресійної моделі використовуючи табличний процесор Excel.

Рішення

На графіку рис.1 будуємо початкові дані по десяти поїздках.

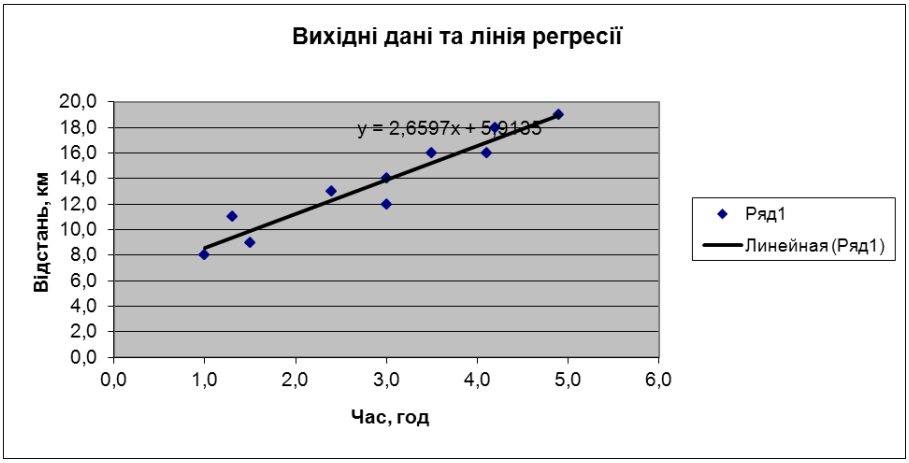


Рис.1. Вихідні дані та лінія регресії

Окрім відстані на час постачання впливають «пробки» на дорогах, час доби, дорожні роботи, погода, кваліфікація водія, вид транспорту. Побудовані точки не знаходяться точно на лінії, що обумовлене описаними вище чинниками. Але ці крапки зібрані навколо прямої лінії, тому можна припустити лінійний зв'язок між параметрами. Всі початкові точки рівномірно

розподілені уздовж передбачуваної прямої лінії, що дозволяє застосувати метод найменших квадратів.

Проведемо регресійний аналіз з використанням режиму **Регрессия** MS Excel. Значення параметрів, встановлених в однойменному діалоговому вікні, представлені на рис.2.

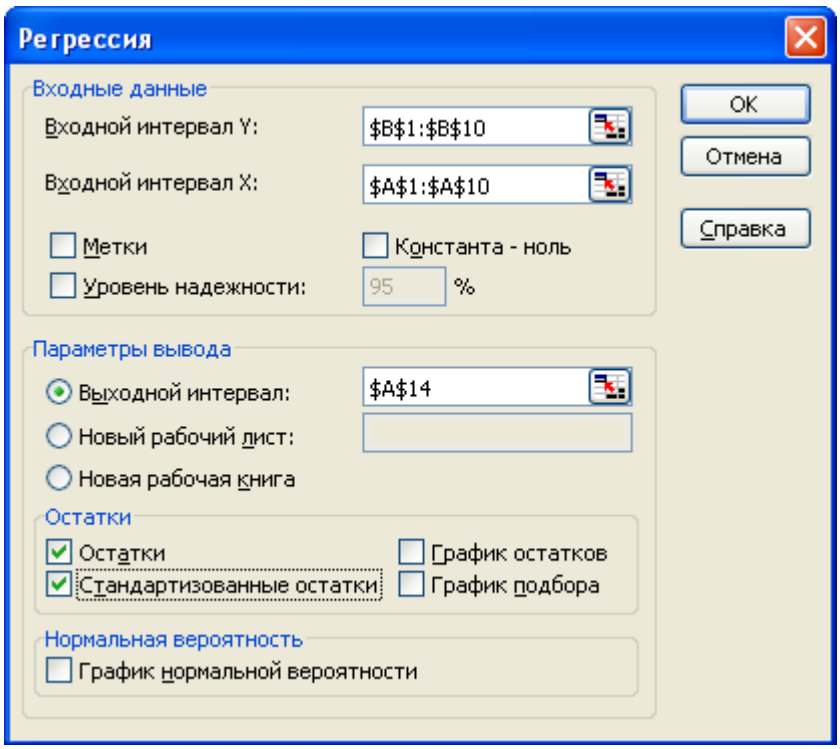


Рис. 2. Значення параметрів, встановлених в діалоговому вікні

Сформовані результати за регресійною статистикою, представлені в таблиці 2.

Таблиця 2

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,958275757
R-квадрат	0,918292427
Нормированный R-квадрат	0,90807898
Стандартная ошибка	1,11809028
Наблюдения	10

Розглянемо представлену в таблиці 3 регресійну статистику. Величина *R*-квадрат, яка також називається мірою визначеності, характеризує якість отриманої регресійної прямої. Ця якість виражається ступенем відповідності між початковими даними і регресійною моделлю (розрахунковими даними). Міра визначеності завжди знаходиться в межах інтервалу [0;1]. У нашому прикладі міра визначеності дорівнює 0,91829, що говорить про дуже хорошу підгонку регресійної прямої до початкових даних і збігається з коефіцієнтом детермінації R^2.

Таким чином, лінійна модель пояснює 91,8% варіацій часу доставки, що означає правильність вибору чинника (відстані). Не пояснюється $100\% - 91{,}8\% = 8{,}2\%$ варіації часу поїздки, які обумовлені рештою чинників, що впливають на час постачання, але не включені в лінійну модель регресії.

Розрахований рівень значущості α_p=1,26e-05 < 0,05 (показник значущість F в таблиці **Дисперсионный анализ**) підтверджує значущість R^2.

Множественный R - коефіцієнт множинної кореляції R - виражає ступінь залежності незалежних змінних (X) і залежної змінної (Y) і дорівнює квадратному кореню з коефіцієнта детерміації, ця величина набуває значень в інтервалі від нуля до одиниці. У простому лінійному регресійному аналізі **Множественный R** дорівнює коефіцієнту кореляції. Дійсно, **Множественный R** в нашому випадку дорівнює коефіцієнту кореляції (0,95827), який обчислюється за формулою:

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2] \cdot [n\sum y^2 - (\sum y)^2]}}$$

Тепер розглянемо середню частину розрахунків, представлену в таблиці 3 (приведена в скороченому варіанті). Тут представлено коефіцієнт регресії a_1 (2,65970168) і зміщення по осі ординат, тобто константа a_0 (5,913462144).

Таблиця 3

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	5,913462144	0,884389599	6,686489927	0,00015485
Переменная X 1	2,65970168	0,280497238	9,482095791	1,26072E-05

Виходячи з розрахунків, можемо записати рівняння регресії таким чином:

$$y^{\partial} = 5{,}91346 + 2{,}65970\,x. \quad (*)$$

Напрям зв'язку між змінними визначається на підставі знаків (негативний або позитивний) коефіцієнта регресії (коефіцієнта a_1). У нашому випадку знак коефіцієнта регресії позитивний, отже, зв'язок також є позитивним.

Далі перевіримо значущість коефіцієнтів регресії: a_0 і a_1. Порівнюючи попарно значення стовпців ***Коэффициенты*** і ***Стандартная ошибка*** в таблиці 4, бачимо, що абсолютні значення коефіцієнтів більш ніж їх стандартні помилки. До того ж ці коефіцієнти є значущими, про що можна судити по значеннях показника **Р-значение** в таблиці 4, які менше заданого рівня значущості α=0,05.

Рішимо задачу прогнозування. Оскільки коефіцієнт детерміації R^2 має достатньо високе значення і відстань 2 кілометри, для якої треба зробити прогноз, знаходиться в межах діапазону початкових даних (таблиця 1), то ми можемо використовувати отримане рівняння лінійної регресії для прогнозування

$$y^*(2\,км) = 5{,}913 + 2{,}660 \cdot 2 = 11{,}2 \text{ хвилин.}$$

При прогнозах на відстані, що не входять в діапазон початкових даних, не можна гарантувати справедливість отриманої моделі. Це пояснюється тим, що зв'язок між часом і відстанню може змінюватися в міру збільшення відстані. На час дальніх перевезень можуть впливати нові чинники такі, як використання швидкісних шосе, зупинки на відпочинок, обід і тому подібне

Таким чином, в результаті використання регресійного аналізу в пакеті Microsoft Excel ми:

- побудували рівняння регресії;
- встановили форму залежності і напрям зв'язку між змінними - позитивна лінійна регресія, яка виражається в рівномірному зростанні функції;
- встановили напрям зв'язку між змінними;
- оцінили якість отриманої регресійної прямої;
- змогли побачити відхилення розрахункових даних від даних початкового набору;
- передбачили майбутнє значення залежної змінної.

Коефіцієнт кореляції

Найчастіше для з'ясування наявності лінійної залежності між показником і фактором можна використовувати коефіцієнт кореляції ***r*** (множинний коефіцієнт кореляції). Це статистична характеристика, яка описує ступінь щільності лінійної залежності між випадковими величинами *x, y* і змінюється в межах від –1 до 1, причому, якщо $r > 0$, то можна вважати, що між *x* і *y* існує пряма залежність, якщо $r < 0$ – зворотна.

Для визначення коефіцієнта кореляції на основі вбудованої функції необхідно:

- активізувати комірку, в яку буде записано результат обчислення ***r***;
- у пункті меню **"Вставка"** вибрати підпункт **"Функция"**;
- у діалоговому вікні, що з'явиться, у полі **"Категория"** вибрати **"Статистические"**, у полі **"Функция"** вибрати **"КОРРЕЛ"**;
- натиснути кнопку **"ОК"**;
- у діалоговому вікні, що з'явиться, у рядку **"Массив1"** записати діапазон комірок однієї змінної, а у рядку **"Массив2"** – іншої, або клацнувши по кнопці "**Свертывание поля**" виділити діапазони комірок із змінними безпосередньо в робочій таблиці. Після виділення потрібного діапазону кнопкою повертаються до розгорнутого вигляду діалогового вікна;
- натиснути кнопку **"ОК"**.

Програма виконання роботи

1. Завантажити табличний процесор Excel.
2. Створити наведену у завданні таблицю.
3. Побудувати графічну залежність між вмістом сирої клейковини в зерні озимої пшениці та силою борошна із цього зерна.
4. Апроксимувати експериментальні дані лінійною функцією.
5. Визначити параметри вказаної функції та точність апроксимації.
6. Визначити коефіцієнт кореляції між вказаними показниками та визначити його значимість.
7. Зберегти створений документ у власній папці.
8. Завершити роботу з Excel.

Завдання на лабораторну роботу

Побудуйте регресійну модель (лінійну) для вихідних даних приведених в таблиці. Для полегшення розрахунків вихідні дані містять тільки чотири пари значень (x_i, y_i).

Вихідні дані

№ варіанта (список журналу)	Координати	Точки				x*
1	X	1	2	3	4	1.6
	Y	30	7	8	1	?
2	X	1	2	3	4	2.3
	Y	25	7	7	2	?
3	X	9	5	2	3	2.9
	Y	25	7	7	2	?
4	X	1	2	3	4	2.6
	Y	15	10	7	0.5	?
5	X	10	3	6	4	8
	Y	25	7	7	2	?
6	X	9	5	2	3	2.5
	Y	15	8.5	7.5	5	?
7	X	2	3	7	8	7.5
	Y	11	8.5	6.5	5	?
8	X	10	3	6	4	9
	Y	15	7	8	6	?
9	X	2	3	4	5	4.5
	Y	13	9	8	7	?
10	X	1	2	3	4	1.5
	Y	7.5	7	5	3.5	?
11	X	1	2	3	4	3.6
	Y	13	9	8	7	?
12	X	3	4	6	10	8
	Y	7.5	7	6.5	3.5	?
13	X	3	4	5	6	7.8
	Y	9	7	5	3	?
14	X	7	5.6	13	14.7	15
	Y	7.5	7	5	3.5	?
15	X	9	5	2	3	5.7

	Y	13	9	8	7	?
16	X	3	4	6	8	5
	Y	7.5	7	6.5	5	?
17	X	2	3	7	8	7.5
	Y	9	9	8	7	?
18	X	9	10	11	12	10.5
	Y	13	9	8	7	?
19	X	1	2	3	4	3.5
	Y	5	4.5	3	3	?
20	X	11	12	13	16	13.6
	Y	7.6	8	6.5	4.2	?
21	X	5	6	7	8	6.5
	Y	5	4.5	3	3	?
22	X	9	10	12	14	12.5
	Y	8	7	6.5	4.2	?
23	X	7	8	9	10	9.6
	Y	8	7	6	4.2	?
24	X	1.5	2.5	3.5	4.5	3.9
	Y	5	4.5	3	3	?
25	X	1	2	5	6	3.9
	Y	5	4	3	3	?
26	X	1.5	2.4	3.8	6.9	4.1
	Y	5.5	5.5	4.8	1.1	?
27	X	1	2	3	4	3.6
	Y	12	3	9	5	?
28	X	1	2	3	7	2.8
	Y	5	5.5	4.8	1.1	?
29	X	11	12	13	16	14.1
	Y	0.25	0.19	5.2	8	?
30	X	1	2	3	4	3.4
	Y	13	4	10	6	?

Дослідіть модель за допомогою режиму Регресія в MS Excel і зробіть прогноз для x^*.

Запитання для самоперевірки

1. Яка послідовність знаходження рівняння апроксимуючої лінії?
2. Як змінити тип апроксимуючої лінії?
3. Як провести аналіз адекватності побудованої моделі?
4. Як вивести на робоче поле рівняння апроксимуючої лінії?
5. Як визначити коефіцієнт кореляції та його значимість?

Практична робота № 14

Перевірка статистичних гіпотез про однорідність дисперсій відхилення часових інтервалів синхроінформації в комп'ютерно-інтегрованих та робототехнічних системах

Мета роботи: отримання практичних навичок використання статистичних критеріїв для перевірки статистичних гіпотез, зокрема критерію Фішера для перевірки гіпотези про однорідність дисперсій за допомогою розрахунків по вбудованим функціям табличного процесора MS Excel

Теоретичні відомості

Поняття про статистичні гіпотези та критерії

Нагальна потреба в перевірці статистичних гіпотез зустрічається в багатьох сферах наукової і професійної діяльності спеціалістів, що займаються рішенням задач з використанням комп'ютерно-інтегрованих та робототехнічних систем. Наприклад при оцінці якості нової технології: показники, одержані при застосуванні нової технології, порівнюються з показниками одержаними за традиційними технологіями. При оцінці якості модернізації обладнання: показники, одержані при застосуванні удосконаленої техніки, порівнюються з показниками одержаними цією технікою до переобладнання.

Будь-яка інформація, отримана в результаті обробки статистичних даних, носить імовірнісний характер. Зокрема, оцінка генеральної дисперсії є величиною випадковою. Тому будь-який висновок, заснований на статистичних даних, є науковим припущенням і називається статистичною гіпотезою. Статистичні гіпотези підлягають перевірці, ціль якої - визначити,

чи не суперечить висунута гіпотеза вихідному статистичному матеріалу (вибірці).

Основну гіпотезу, сформульовану в результаті обробки статистичного матеріалу, називають **нульовою гіпотезою** й позначають $\mathbf{H_0}$. На противагу нульовій гіпотезі призначають одну або декілька **альтернативних** (конкуруючих) гіпотез. Їх позначають H_1, H_2, … і т.д.

До завдань перевірки статистичних гіпотез відносять: порівняння емпіричних законів розподілу, або емпіричного з теоретичним, а також параметрів цих розподілів (у випадку нормального розподілу – середніх значень і дисперсій).

Наприклад, якщо перевіряється гіпотеза про рівність параметра **a** деякому заданому значенню $\mathbf{a_0}$, то як альтернативні гіпотези можна розглянути гіпотези, що а більше або менше $\mathbf{a_0}$:

$$H_0: \mathbf{a} = \mathbf{a_0};$$

$$H_1: \mathbf{a} > \mathbf{a_0};$$

$$H_2: \mathbf{a} < \mathbf{a_0};$$

$$H_3: \mathbf{a} \neq \mathbf{a_0}.$$

Вибір альтернативної гіпотези обумовлюється формулюванням задачі.

Перевірка гіпотези дозволяє зробити висновок щодо відповідності або протиріччя висунутої гіпотези емпіричним даним.

В якості критеріїв для перевірки статистичних гіпотез використовують випадкові величини (статистики), особливість яких полягає в тому, що кожна з них має свій закон розподілу, що не залежить від закону розподілу генеральної сукупності й вибірки, а залежить від умов обробки вибіркових даних. Значення цих випадкових величин, позначимо їх Z, з відповідними їм ймовірностями приводяться в довідкових таблицях.

На підставі вибіркових даних визначають значення критерію Z (його називають *значенням критерію, що спостерігається* або *розрахункове значення критерію*) і порівнюють його з табличним значенням, що відповідає умовам обробки даних. Перевірка статистичної гіпотези заснована на принципі, відповідно до якого малоймовірні події вважаються неможливими, а події, що мають більшу ймовірність, - достовірними. Якщо імовірність *розрахункового значення критерію* досить велика, тобто факт цілком імовірний, то говорять, що гіпотеза не суперечить даним спостереження. Якщо ж ця ймовірність мала, тобто подія практично неможлива, то говорять, що нульова гіпотеза суперечить даним спостереження, і її відхиляють.

Питання про те, яку ймовірність варто вважати досить великою або малою, вирішується не з математичних міркувань, а залежить від наслідків того, що прийнята гіпотеза виявиться невірною. Мала ймовірність, при якій значення критерію вважається практично неможливим, позначається α і називається **рівнем значущості**. У практичних задачах звичайно призначають рівень значущості α = 0,05-0,15. Область значень критерію Z, що відповідає рівню значущості α, називають **критичною областю**. Область значень критерію Z, що відповідають імовірності 1-α, називають **областю прийняття гіпотези**. Значення критерію, що відокремлює область прийняття гіпотези від критичної області називається **критичною точкою z_k** (критичним значенням критерію). Це значення критерію визначається за спеціальними формулами для вибраного рівня надійності і розрахованого числа степенів свободи.

Порівняння двох дисперсій

Порівняння дисперсій використовують для оцінки степені розсіяння двох випадкових величин, оцінки точності визначення певних показників тощо. Наприклад, при порівнянні двох технологій, що використовуються в комп'ютерно-інтегрованих та робототехнічних системах кращою буде та, яка забезпечує, окрім найкращих значень множини певних показників, мінімальні

відхилення окремих вимірювань цих показників від своїх середніх значень. Середнє квадратичне відхилення (корінь квадратний з дисперсії) використовують для оцінки точності визначення середніх значень показників, що є необхідною умовою при визначенні статистичної значущості різниці між ними. Порівняння дисперсій передує порівнянню середніх, оскільки від його результату залежить вибір інструменту аналізу для порівняння середніх значень.

Нехай є дві незалежні вибірки й варто визначити, чи взяті вони з нормальних генеральних сукупностей X і Y з однаковою дисперсією. Запишемо нульову гіпотезу:

$$H_0: D[X] = D[Y], \quad (1)$$

і сформулюємо альтернативну гіпотезу:

$$H_1: D[X] \neq D[Y]. \quad (2)$$

Як критерій для перевірки нульової гіпотези про рівність дисперсій нормальних генеральних сукупностей приймають випадкову величину, що являє собою відношення більшої дисперсії до меншої:

$$F_{\text{îáñ}} = \frac{S_1^2}{S_2^2}, \quad (3)$$

де $S_1 > S_2$.

Нульова гіпотеза припускає, що дві вибірки незалежні й узяті з нормальних генеральних сукупностей з однаковими дисперсіями, у цьому випадку F=1. Однак, навіть якщо гіпотеза вірна, то малоймовірно, що S_1 прийме точно таке ж значення, як S_2 через вплив випадковості. Таким чином, завдання полягає в тому, щоб перевірити, чи буде випадкова величина F досить близька до одиниці.

Випадкова величина F має розподіл Фішера, що залежить тільки від ступенів свободи $k_1=n_1-1$ і $k_2=n_2-1$, де n_1 і n_2 об'єм вибірки з більшою й з

меншою вибірковими дисперсіями відповідно, і не залежить від інших параметрів.

Для перевірки нульової гіпотези (1) при конкуруючій гіпотезі (2) будують двосторонню критичну область, що відповідає рівню значимості α. Двостороння критична область визначається двома нерівностями:

$$F < F_{кр1};$$

$$F > F_{кр2},$$

а область прийняття гіпотези:

$$F_{кр1} < F < F_{кр2}$$

причому, імовірність влучення критерію в кожний із двох інтервалів критичної області дорівнює $\alpha/2$.

Розрахункове значення критерію порівнюють з критичним, обчисленим з прийнятим рівнем значущості. Якщо розрахункове значення критерію $F_{ðîçð}$ перевищує критичне $F_{êðèò}$, то нульова гіпотеза про рівність двох дисперсій відхиляється. У цьому випадку різниця між дисперсіями, що порівнюються, не є випадковою, а є статистично значущою, тобто дисперсії двох генеральних сукупностей неоднорідні. Якщо розрахункове значення критерію Фішера менше критичного, то висунута нульова гіпотеза приймається, тобто дисперсії однорідні.

Порівняння дисперсій двох вибірок за допомогою вбудованих функцій табличного процесора MS Excel

Для порівняння дисперсій в MS Excel використовується засіб під назвою **Двухвыборочный F-тест для дисперсии.** Для його використання виконують таку послідовність дій: **Сервис → Анализ данных → Двухвыборочный F-тест для дисперсии.**

В діалоговому вікні **Двухвыборочный F-тест для дисперсии** вводять такі данні:

- у групі **Входные данные** у полі **Интервал переменный 1** вводять адресу інтервалу комірок, що містять дані першої вибірки (дисперсія якої має бути більшою), а в полі **Интервал переменный 2** вводять адресу інтервалу комірок, що містять дані другої вибірки;
- у полі **Альфа** встановлюють рівень значущості (за замовчення встановлено a=0,05);
- у групі **Параметры ввода** для виведення результатів обчислень на поточному робочому аркуші активізують перемикач **Выходной интервал** і вказують у полі справа від перемикача адресу комірки для виведення даних (верхню ліву);
- для виведення результатів обчислень на новий аркуш активізують перемикач **Новый рабочий**;
- для виведення результатів обчислень у новий файл активізують перемикач **Новая рабочая книга**;
- після встановлення всіх необхідних параметрів натискують кнопку **Ок**.

У результаті виконання дій з'явиться таблиця, що містить обчислені середні значення, дисперсії, число степенів свободи для кожної вибірки (у рядку df), значення критерію Фішера, що спостерігається (у рядку F) та інші. Якщо дисперсія першої змінної виявиться меншою за дисперсію другої змінної, то одержані результати вилучають і обчислення повторюють, причому в якості першої змінної використовують ту, дисперсія якої більше. Тобто, при заповненні полів F-тесту змінюють послідовність введення діапазонів комірок з даними.

Для прийняття рішення порівнюють розрахункове значення критерію Фішера у рядку F даної таблиці, з критичним значенням розподілу Фішера $F_{крит}$ із останнього рядка таблиці. Якщо $F > F_{крит}$, то дисперсії, що порівнюються, неоднорідні (нульова гіпотеза про їх однорідність відхиляється). Якщо $F < F_{крит}$, то дисперсії, що порівнюються, однорідні (нульова гіпотеза про їх однорідність приймається).

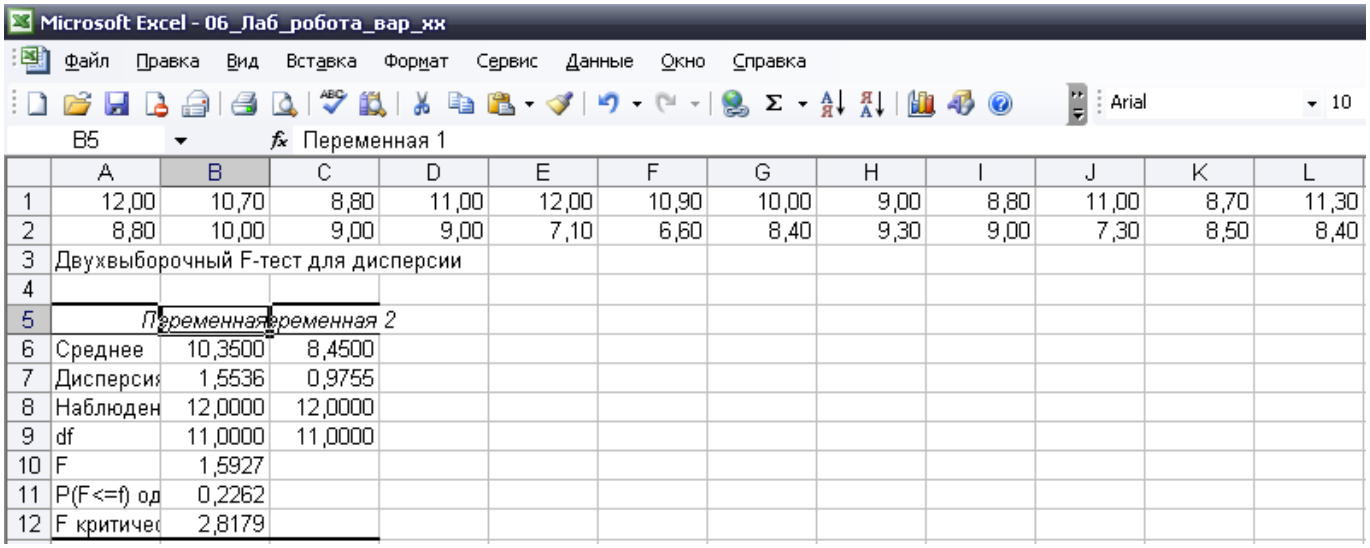
Приклад. Досліджували відхилення часових інтервалів, що задаються пристроєм формування синхроінформації (ПФС) в комп'ютерно-інтегрованих та робототехнічних системах до його модернізації та після модернізації, при якій впроваджено заходи по температурній оптимізації. Параметр, що досліджувався $\left(\Delta\Theta_t / \Theta\right)\cdot 10^{-6}$ - відносне відхилення часових інтервалів синхроінформації; y_i – значення відносного відхилення часових інтервалів $\left(\Delta\Theta_t / \Theta\right)\cdot 10^{-6}$, отримане в результаті дослідження ПФС в комп'ютерно-інтегрованих та робототехнічних системах до модернізації; x_i – після модернізації. Результати дослідження наведено в табл. 1.

Таблиця 1

Результати дослідження ПФС в комп'ютерно-інтегрованих та робототехнічних системах

x_i	8,8	10	9	9	7,1	6,6	8,4	9,3	9	7,3	8,5	8,4
y_i	12	10,7	8,8	11	12	10,9	10	9	8,8	11	8,7	11,3

Необхідно встановити за рівнем значущості a=0,05 чи однорідні дисперсії двох вибіркових сукупностей. Результати розрахунку наведено на рис. 1.



	A	B	C	D	E	F	G	H	I	J	K	L
1	12,00	10,70	8,80	11,00	12,00	10,90	10,00	9,00	8,80	11,00	8,70	11,30
2	8,80	10,00	9,00	9,00	7,10	6,60	8,40	9,30	9,00	7,30	8,50	8,40
3	Двухвыборочный F-тест для дисперсии											
4												
5		Переменная 1	Переменная 2									
6	Среднее	10,3500	8,4500									
7	Дисперсия	1,5536	0,9755									
8	Наблюден	12,0000	12,0000									
9	df	11,0000	11,0000									
10	F	1,5927										
11	P(F<=f) од	0,2262										
12	F критичес	2,8179										

Рис. 1. Результати розрахунку

Оскільки $F=1,5927<F_{крит}=2,8179$, то дисперсії, що порівнюються, не мають статистично значущої відмінності, тобто вони однорідні.

Програма виконання роботи

1. Завантажити табличний процесор MS Excel.
2. Вибрати один варіант даних дослідження ПФС в комп'ютерно-інтегрованих та робототехнічних системах з табл.2 (вихідні данні для виконання роботи), відповідно до останніх двох цифр номера в списку журналу та оформити дані у вигляді таблиці на аркуші MS Excel.
3. Перевірити однорідність дисперсій двох вибірок за критерієм Фішера з рівнем значущості a=0,05.
4. Проаналізувати одержані результати.
5. У випадку неоднорідності дисперсій повторити перевірку з рівнем значущості a=0,1 та a=0,01.
6. Порівняти одержані результати.

Завдання для виконання роботи

Вихідні данні для виконання роботи наведено в табл.2.

Таблиця 2

№ вар.	Вхідні дані	
01, 02, 03, 04	x_i	8,9; 8,2; 8,8; 8,8; 8,6; 8,1; 9,2; 9,3; 9,1; 8,3; 9,0; 9,2
	y_i	10,3; 5,5; 7,3; 8,3; 7,6; 8,3; 9,0; 8,3; 7,4; 7,4; 10; 6,6

05, 06, 07, 08	x_i	6,8; 7,6; 7,1; 7,5; 8,9; 7,7; 6,4; 8,6; 5,6; 7,8; 6,8; 7,3
	y_i	9,8; 10,3; 9,9; 9,1; 9,5; 8,4; 10,6; 8,4; 9,1; 10,1; 9,1; 10,1
09, 10, 11, 12	x_i	7,1; 5,9; 6,0; 7,7; 8,0; 6,7; 9,0; 8,0; 8,7; 7,9; 6,4; 8,4
	y_i	9,4; 9,0; 9,5; 12,0; 9,4; 8,6; 11,5; 12,0; 6,7; 11,9; 7,3; 6,9
12, 14, 15, 16	x_i	10,3; 5,5; 7,3; 8,3; 7,6; 8,3; 9,0; 8,3; 7,4, 7,4; 10; 6,6
	y_i	9; 8,7; 12,0; 10,2; 9,6; 7,9; 10,7; 7,6; 6,7; 11,6; 8,9; 10,5
17, 18, 19, 20	x_i	6,6; 9,8; 9,1; 10,1; 10,6; 9,2; 10,6; 7,6; 11,7; 11; 9,6; 9,6
	y_i	8,7; 9,9; 7,4; 12,0; 8,9; 7,4; 8,8; 11,3; 11,1; 9,6; 9,5; 10,8
21, 22, 23, 24	x_i	10; 8,5; 10,7; 8,3; 10,5; 9,4; 10,9; 9,2; 10,8; 10,4; 10,6; 9,6
	y_i	9,7; 10,9; 10,6; 12; 9,1; 11,7; 10,7; 11; 11,6; 9,8; 10,5; 11
25, 25, 27, 28	x_i	8,9; 8,2; 8,8; 8,8; 8,6; 8,1; 9,2; 9,3; 9,1; 8,3; 9,0; 9,2
	y_i	11,4; 11,3; 10,3; 9,4; 11,9; 9,7; 12; 8; 12; 10; 9; 11
29, 30, 31, 32	x_i	9,5; 8; 7; 6,6; 7,3; 5,3; 7,7; 7; 6,2; 6,5; 5,8; 6,5
	y_i	7; 11,7; 11; 12,6; 12; 11,4; 8,8; 10,2; 11, 10,2; 12; 11,9
33, 34, 35, 36	x_i	8,7; 9,9; 7,4; 12,0; 8,9; 7,4; 8,8; 11,3; 11,1; 9,6; 9,5; 10,8
	y_i	10; 8,5; 10,7; 8,3; 10,5; 9,4; 10,9; 9,2; 10,8; 10,4; 10,6; 9,6
37, 38, 39, 40	x_i	7,1; 5,9; 6,0; 7,7; 8,0; 6,7; 9,0; 8,0; 8,7; 7,9; 6,4; 8,4
	y_i	10; 8,5; 10,7; 8,3; 10,5; 9,4; 10,9; 9,2; 10,8; 10,4; 10,6; 9,6

Запитання для самоперевірки

1. Що називається статистичною гіпотезою?
2. Які бувають статистичні гіпотези?
3. Що називається статистичним критерієм?
4. Для чого використовується порівняння двох дисперсій?

5. За яких умов дисперсії вважаються однорідними і за яких - ні?

6. Про що свідчить неоднорідність дисперсій?

7. За яким критерієм порівнюють дві дисперсії і який порядок його застосування?

8. Які засоби MS Excel використовуються для порівняння дисперсій?

9. За якими показниками, виведеними у таблиці F-тесту MS Excel, визначають однорідність дисперсій?

10. Як впливає значення рівня значущості на результати?

Література

1. Ачкасов А.Є., Плакіда В.Т., Воронков О.О., Воронкова Т.Б. Теорія імовірностей і математична статистика: Навчальний посібник. - Харків, ХНАМГ, 2008. - 247 с.
2. Гнеденко Б.В. Курс теории вероятностей: Уч. для студ. математ. спец. ун-в. – 6-е изд. - М.: Наука, 1988. – 448с.
3. Севастьянов Б.А. Курс теории вероятности и математической статистики. М.: Наука, 1982. – 255с.
4. Кремер Н.Ш. Теория вероятностей и математическая статистика: учебник для вузов, обучающихся по экономич. спец. – 3-е изд. перераб. и доп. – М.:ЮНИТИ-ДАНА, 2007. - 551 с.
5. Гмурман В. Е. Теория вероятностей и математическая статистика/В.Е.Гмурман.-7-е изд. - М.: Высшая школа, 2000.- 479с.
6. Гмурман В.Е. Руководство по решению задач по теории вероятностей и математической статистике: Уч. пособие для студентов вузов. 11-е изд. перераб. и доп. М.: Высш. образование, 2007. - 404с.
7. Методичні вказівки до виконання лабораторних робіт з дисципліни «Математичне моделювання та планування експерименту» / Гаріна С.М., Тарасенко Р.О. – К.: НАУ, 2008. – 102с.
8. Баврин И.И., Матросов В.Л. Краткий курс теории вероятностей и математическая статистика. - М.: Прометей, 1989. - 136с.

Кіктєв Микола Олександрович

Коваль Валерій Вікторович

ОБРОБКА ІНФОРМАЦІЇ В КОМП'ЮТЕРНО-ІНТЕГРОВАНИХ СИСТЕМАХ АВТОМАТИЗАЦІЇ ТА РОБОТОТЕХНІКИ

Методичні вказівки до виконання практичних робіт

(для аспірантів спеціальності 174/G7)

Підписано до видання 12.10.2025 г. Умов. друк. арк. - 19,2.