Міністерство освіти і науки України

Національний університет “Львівська політехніка”

Кафедра інформаційних систем та мереж

**ЗВІТ**

про виконання лабораторної роботи № 11

*“* Аналіз статистичних даних*”*

з дисципліни *“ Методи обчислень та візуалізація даних ”*

Виконала:

Студентка групи СА-33

Солтис М. В., Дудак О. М.

Прийняла к.т.н., доцент

Висоцька В. А.

*Львів 2022*

**Мета роботи:** ознайомлення з основними методами візуалізації, графічного відображення та первинного статистичного опрацювання числових даних, які представлені вибірковою сукупністю або часовим рядом.

**Варіант 13**

**Теоретичні відомості:**

Отримані в різних дослідженнях дані переважно характеризують спостережуваний об’єкт в двох аспектах – статичному й динамічному. Статичний аспект дає характеристику об’єкта значеннями конкретних величин, які можуть бути як реальними, тобто характеризують об’єкт таким як він є (кількість елементів конкретної схеми, значення елемента за даною шкалою) або випадковими – з відомим або невідомим законом розподілу їх ймовірності. Динамічний аспект характеризує дані про об’єкт протягом часу, регулярно або нерегулярно, але так, щоб кожне значення було прив’язане до моменту часу його спостереження чи реєстрації. Ці два аспекти виражаються з точки зору їх опрацювання двома класами: вибірковими сукупностями – вибірками і часовими послідовностями або часовими рядами. Для першого класу результатом опрацювання є визначення закону розподілу випадкових значень елементів вибірки – варіант. Зауважимо, що невипадкові значення не потребують знаходження виду та параметрів закону розподілу, а використовуються безпосередньо в розрахунках або їх перед цим усереднюють. Для другого класу, який є різновидом випадкових процесів, результатом опрацювання є аналітичне подання тенденції розвитку досліджуваного показника в часі. Завершенням таких досліджень є математична модель, в якій ці дві компоненти одного і того ж самого набору даних (вважаємо, що дані прив’язані до часу або до іншого показника) об’єднані аналітично, а сама модель, з точністю до прийнятого критерію адекватності, описує їхню поведінку. Попереднє опрацювання даних фактично дає підстави для побудови такої математичної моделі, проте для цього мають бути реалізовані певні етапи. Для досягнення поставленої мети в цій роботі необхідно забезпечити поетапне виконання низки завдань. Тобто, вирішення цих завдань здійснюється наступними етапами.

Основним типом візуалізації даних у звітах експериментальних та науково-практичних досліджень є графіки. Графік відображає відношення між двома величинами, одна з яких є незалежною змінною і її значення, зазвичай, відкладають вздовж горизонтальної осі – абсциси; друга змінна є залежною і її значення відкладають вздовж вертикальної осі – ординати. Фактично, будь-який графік складається з декількох основних елементів: осей з надписами, експериментальних точок, ліній, що з’єднують ці точки (їх ще називають кривими), пояснюваних написів на рисунку та підпису під рисунком.

Крім табличного та графічного представлення даних, в їх супровід включають загальні числові та статистичні характеристики, які відносять до описової або дескриптивної статистики. Описова статистика дає підстави для формування компетенцій щодо вибору шкали вимірювань, автоматизації опрацювання даних при застосуванні різних форматів на етапі їх збору, подання результатів у різних формах, графічного подання результатів, обчислення статистичних параметрів розподілу та оцінки параметрів генеральної сукупності з використанням інформаційних технологій. Описова (дескриптивна) статистика – це найбільш загальні статистичні показники, що описують розподіл даних, приймаючи за норму нормальний розподіл.

Середнє арифметичне є мірою центральної тенденції, що відображає найбільш характерне для даної вибірки значення.

Мода (позначається «Мо») – це значення, яке найбільш часто зустрічається серед вибірки змінних. Часто застосовується для непараметричних даних і для рангових шкал.

Медіана (позначається «Ме») – значення, яке ділить навпіл впорядковану множину змінних, тобто для визначення медіани необхідно впорядкувати дані, наприклад, за зростанням. Способи визначення значення медіани для парної і непарної кількості даних відрізняються.

Коефіцієнт варіації виражає міру мінливості ознаки у відсотках.

Асиметрія – показник, що відображає перекіс розподілу відносно моди вліво або вправо. Це має місце в тих випадках, коли якісь причини сприяють більш частій появі значень, які є більшими або, навпаки, меншими за середнє арифметичне. Для лівосторонньої або додатної асиметрії в розподілі частіше зустрічаються більш низькі значення, а для правосторонньої або від’ємної – вищі.

Ексцес – показник, що відображає висоту розподілу. У тих випадках, коли якісь причини сприяють появі близьких до середніх значень, утворюється розподіл з додатнім ексцесом. Якщо ж в розподілі переважають крайні значення, причому одночасно і більш низькі, і більш високі, то такий розподіл характеризується від’ємним ексцесом, і в центрі розподілу може утворитися впадина, яка перетворює його в двовершинний.

**Хід роботи**

Форми і методи подання та попереднє статистичне опрацювання числових даних часових послідовностей: вхідні дані, хід роботи, результати:

1. Попереднє опрацювання даних та подання результатів:
   1. Формування файлу даних у формі таблиць.

Формуємо з усіх даних таблицю в середовищі RStudio.

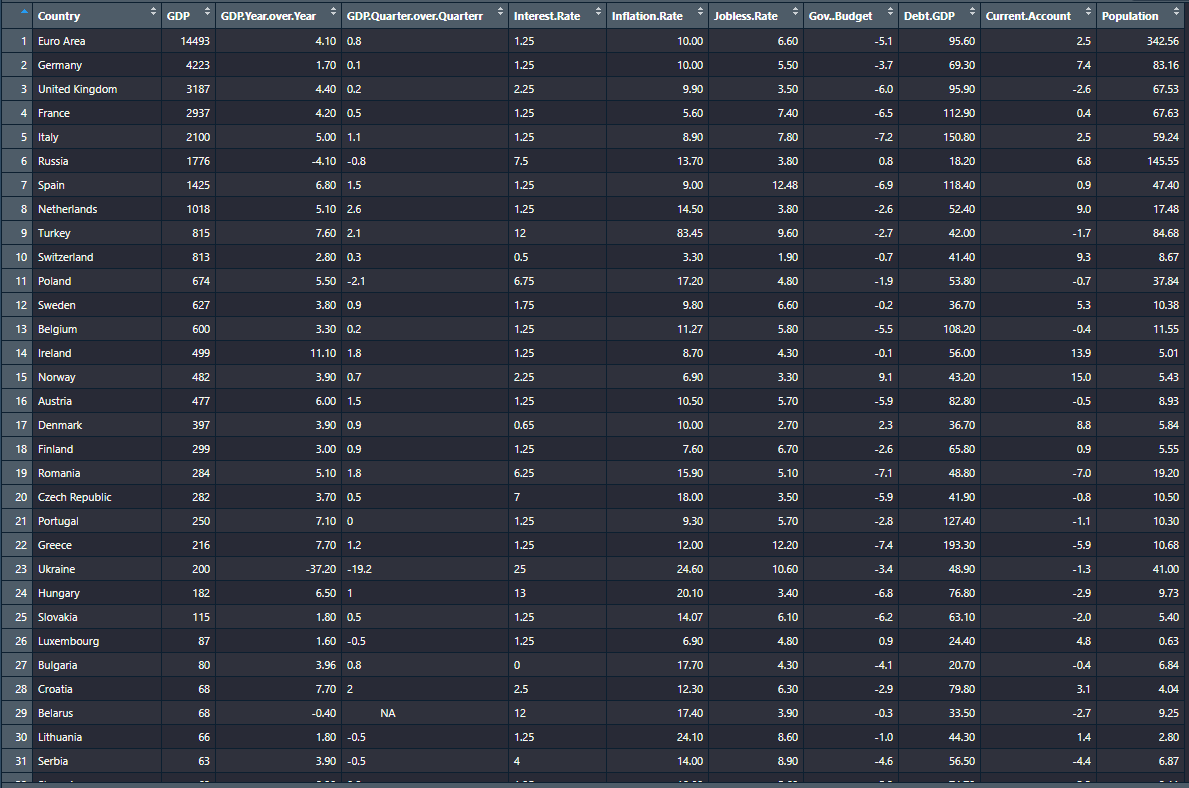


Рис. 1. Таблиця даних

Уся подальша робота буде проводитися саме з цією таблицею.

* 1. Графічне подання даних.

Зобразимо залежність безробіття від ВВП країн за допомогою графічного подання на лінійному графіку

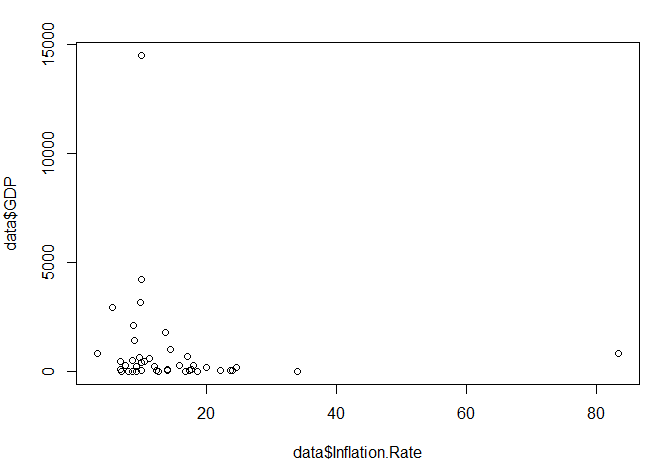


Рис. 2. Графічне подання даних про ВВП країн в декартовій системі координат

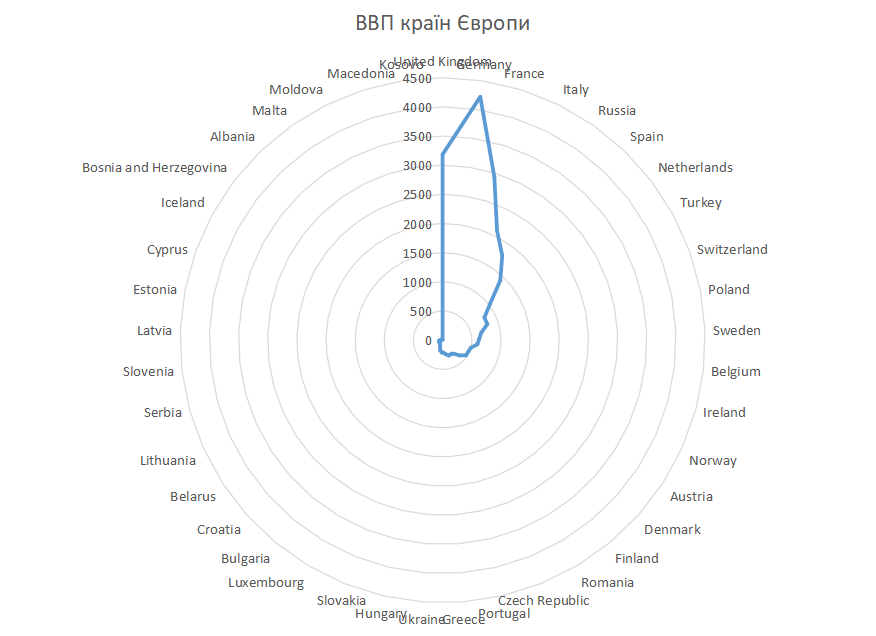


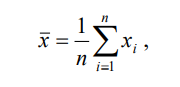
Рис. 3. Графічне подання даних про ВВП країн в полярній системі координат

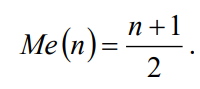
Описова статистика – кількісні характеристики даних.

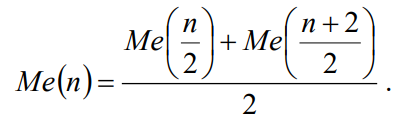
1. Описова статистика – кількісні характеристики даних.

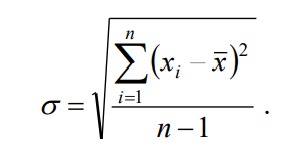
Для виконання цього пункту необхідно знайти наступні дані вибірки: середнє арифметичне (2.1), моду, медіану (2.2, 2.3), розмах вибірки, стандартне відхилення (2.4), коефіцієнт варіації (2.5), а також мінімум та максимум.

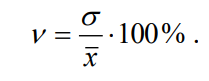
Запишемо формули для розрахунку деяких цих значень:

 (2.1)

 (2.2)

 (2.3)

 (2.4)

 (2.5)

У мові R використаємо команду summary, яка обчислює усі необхідні статистичні дані та видає результат до кожного стовпця.

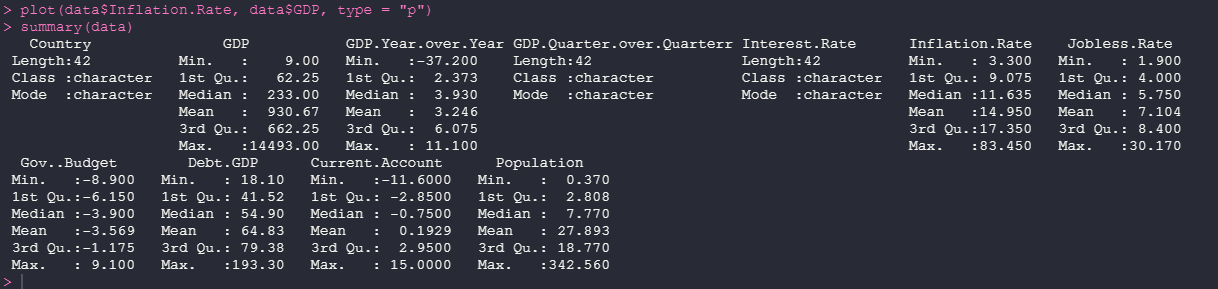
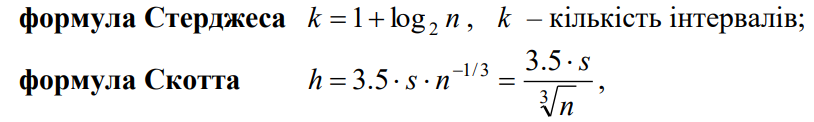


Рис. 4. Готова таблиця даних

* 1. Побудова гістограми

Гістограма є дуже наближеним відображенням графіка функції щільності закону розподілу даних цієї вибірки. Вона також є діаграмою, яка наочно відображає метод групування даних за деякою істотною ознакою. Найбільш проблематичним у побудові гістограми є вибір кількості інтервалів групування, тобто кількості груп, на які розбивається вибірка.

За формулами Стерджеса і Скотта потрібно знайти кількість та ширину інтервалів.



За допомогою команди hist(data$Jobless.Rate) побудуємо гістограму за рейтингом безробіття. За результатами можна побачити, що в більшості країн рейтинг безробіття не перевищує 10-15%



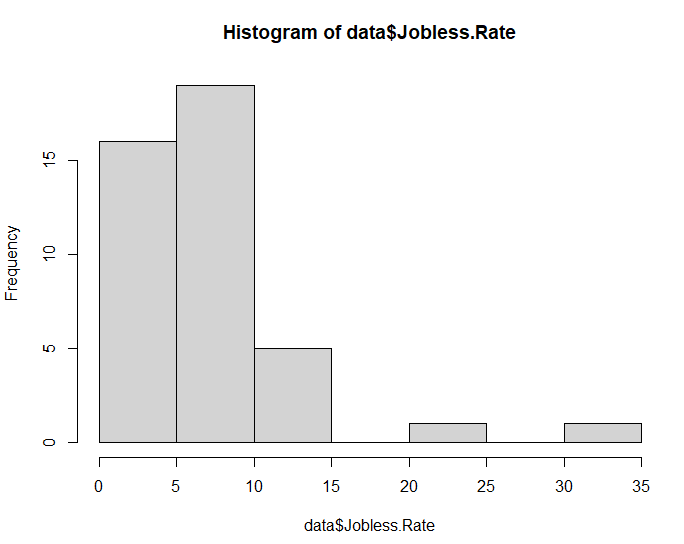


Рис. 5. Таблиця інтервалів та гістограма частот

* 1. Побудова кумуляти.

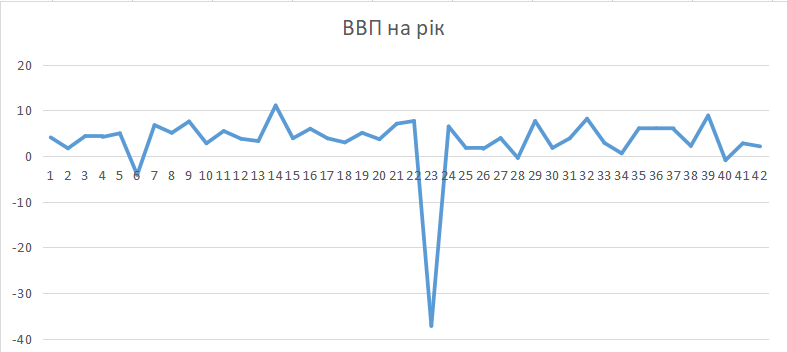


Рис. 6. Кумулята, побудована за гістограмою

Виявлення тенденції часового ряду методами згладжування

При побудові математичних моделей часових рядів, переважно з метою визначення динаміки показника, необхідно в першу чергу виділити тенденцію та відокремити її від випадкових відхилень, зумовлених різними перешкоджальними факторами. При візуальному способі будується графік часового ряду, на основі якого висувається гіпотеза про його структуру, в першу чергу про форму тренду.

1. Методи згладжування часових рядів

Умовно методи згладжування поділяють на дві різні групи, в основі яких лежать два різні підходи: аналітичний та алгоритмічний

Аналітичний підхід ґрунтується на припущенні, що дослідник може на підставі візуального аналізу задати загальний вигляд функції, вважаючи що її графік відповідає характеру тенденції, тобто описати функцію як експоненту, гіперболу, параболу, степеневу функцію, поліноми вищих степенів чи інші функції.

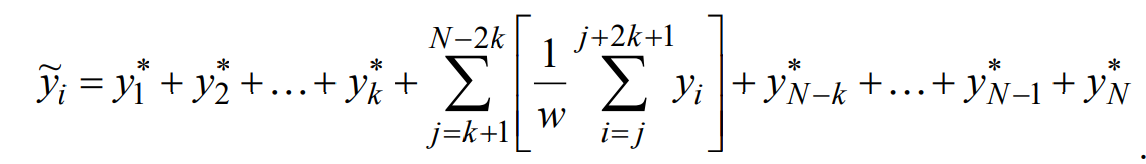
Іншими словами, аналітичний підхід означає заміну значень рівнів часового ряду значеннями, що теоретично розраховані на підставі явного аналітичного вигляду функції, якою апроксимують візуально визначений тренд.

В алгоритмічному підході вигляд тренду отримують за рахунок різних алгоритмів, які практично реалізують процедури згладжування. Ці процедури надають досліднику лише алгоритм розрахунку нового значення часового ряду в будь-який заданий момент часу t. Найбільш вживаними є методи згладжування часових рядів за допомогою ковзних середніх.

1. Метод ковзної середньої.

Метод ковзних (рухомих) середніх є одним із найстаріших відомих способів згладжування часового ряду. Він базується на переході від початкових значень ряду до їх середніх значень на інтервалі часу, довжина якого обрана заздалегідь. При цьому сам вибраний інтервал часу ковзає вздовж ряду. Суть цього методу зводиться до заміни фактичних рівнів ряду послідовностями рівнів, що мають, як правило, значно менші коливання, ніж вихідні дані. Зменшення флуктуації дає можливість наочно виявити основну тенденцію. Для забезпечення часової відповідності інтервал має охоплювати непарну кількість рівнів, тобто .

Таким чином, при обчисленнях середнього рівня межі інтервалу ніби «ковзають» по рівнях часового ряду від його початку до кінця, щоразу відкидаючи один рівень на початку інтервалу і додаючи в кінці наступний. Звідси назва – ковзна середня. Алгоритм для обчислення простої ковзної середньої є такий:



Якщо графічне представлення часового ряду нагадує пряму лінію, то в цьому випадку ковзна середня не спотворює динаміку досліджуваного явища.

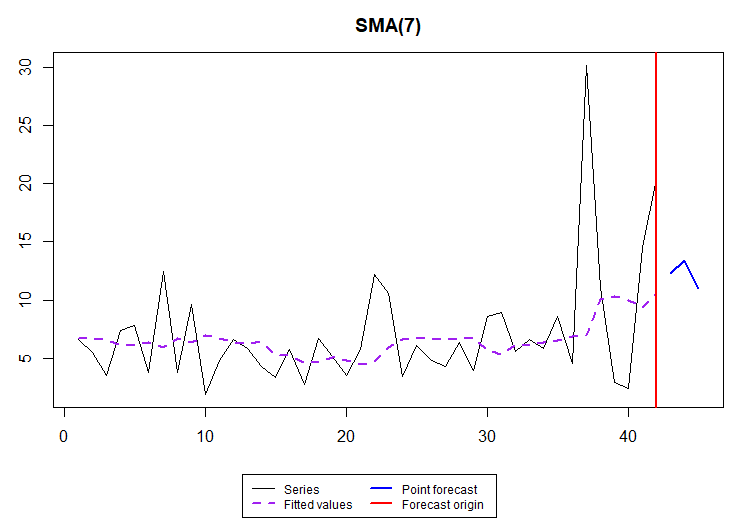


Рис. 7. Графік згладжування при w = 3

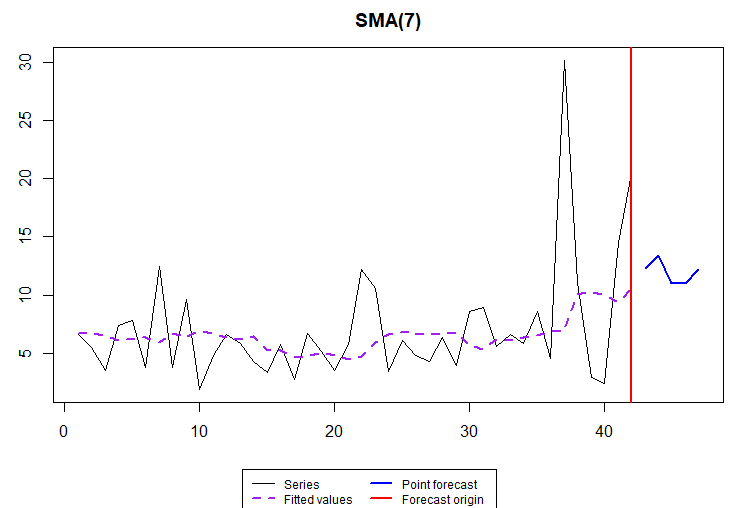


Рис. 8. Графік згладжування при w = 5

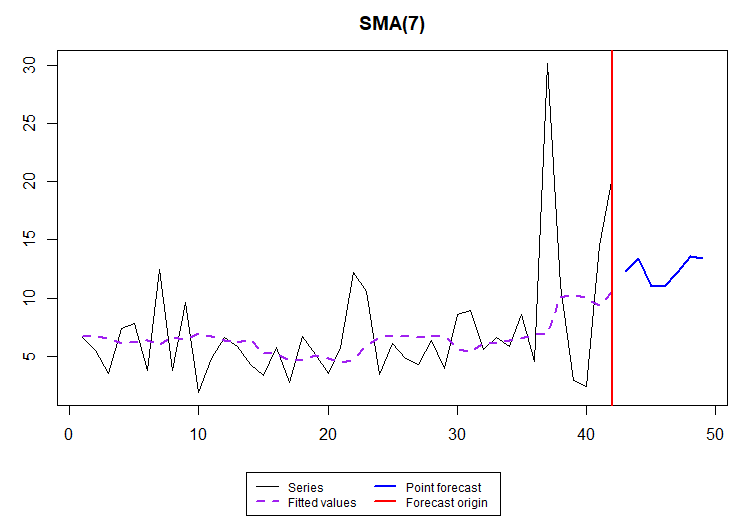


Рис. 9. Графік згладжування при w = 7

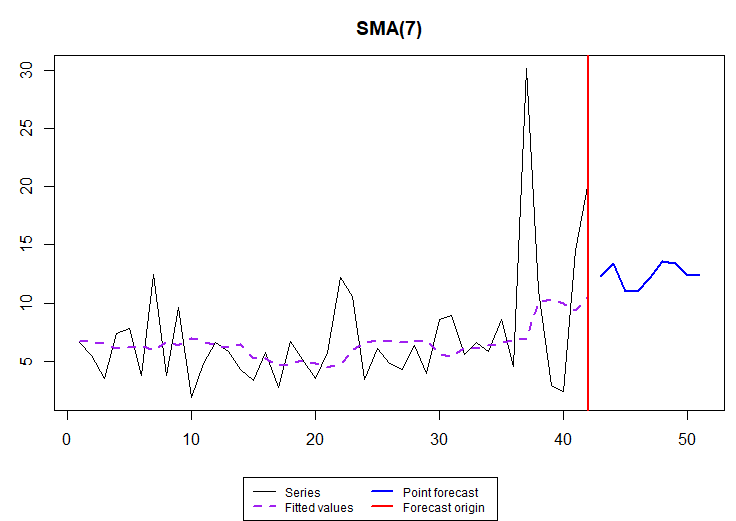


Рис. 10. Графік згладжування при w = 9

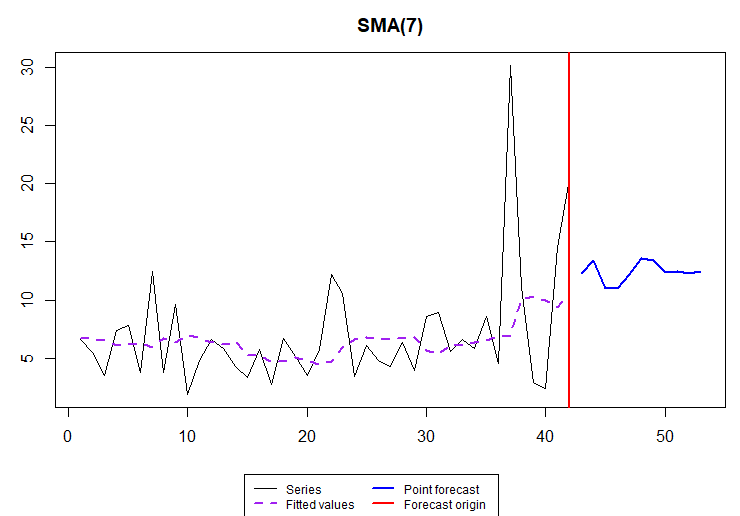
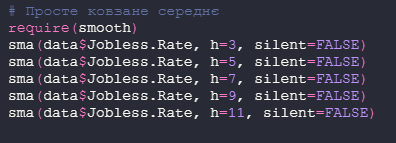


Рис. 11. Графік згладжування при w = 11

Код до завдання:



Експотенціальне згладжування

Суттєвою ознакою застосування експоненціальних середніх є обґрунтування величини параметра згладжування альфа. Чим він менший, тим більше згладжуються рівні в ряді, який аналізується. Це означає зростання питомої альфа.

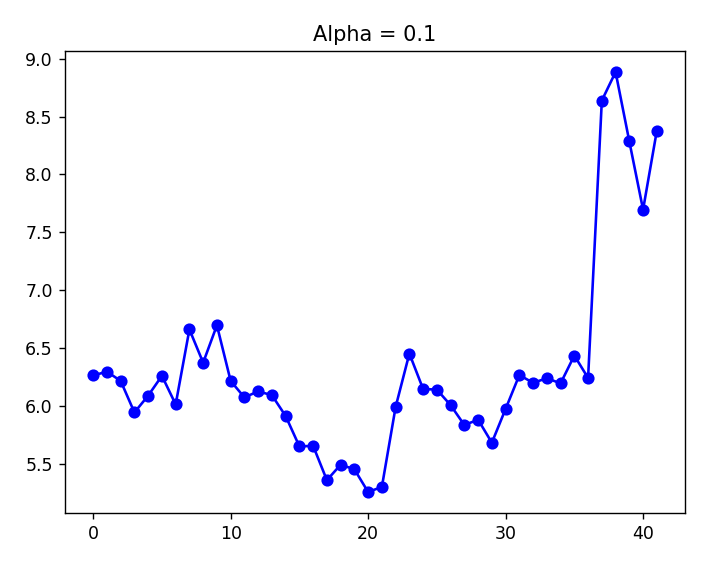


Рис. 12. Експотенціальне згладжування при альфа = 0,1

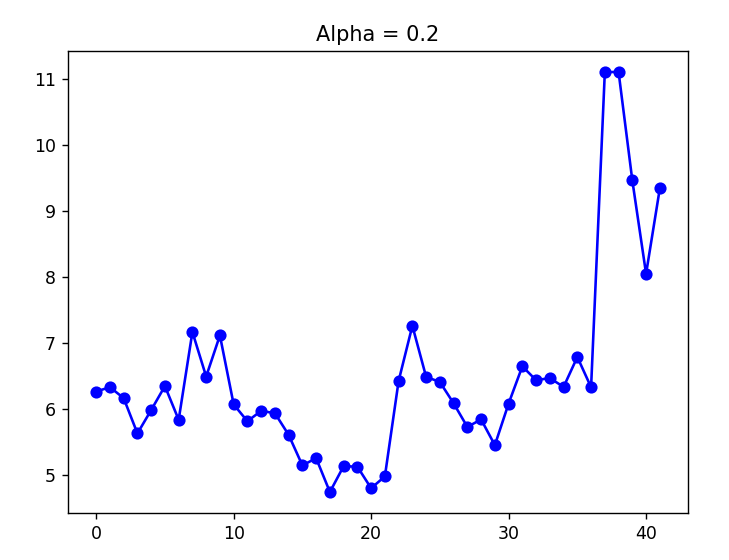


Рис. 13. Експотенціальне згладжування при альфа = 0,2

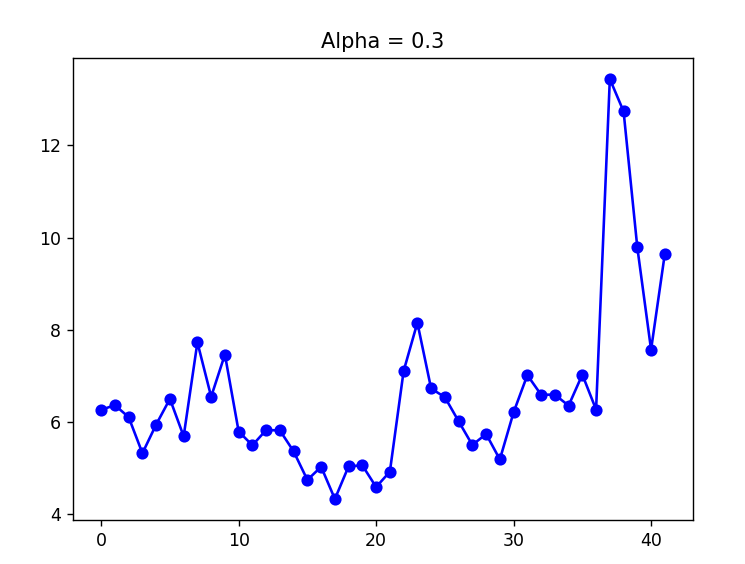
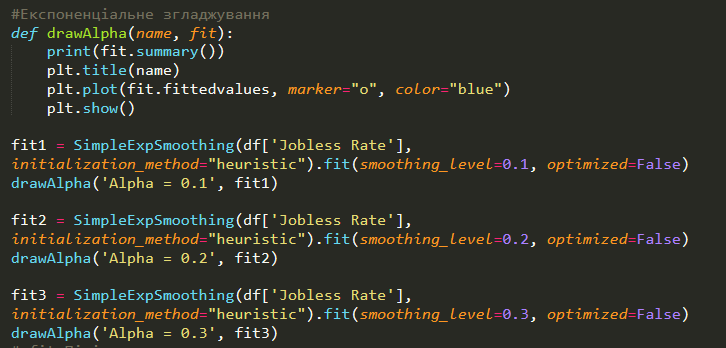


Рис. 14. Експотенціальне згладжування при альфа = 0,3

Код до завдання:



1. Метод зваженої ковзної середньої.

Поряд з простими ковзними середніми застосовуються також поліноміальні або зважені середні. Вони дозволяють точніше описати початок форми основну тенденцію ряду, оскільки при обчисленні зваженої середньої кожному рівню ряду в межах інтервалу згладжування приписується певна вагу, що залежить від відстані до середини інтервалу. Вагові коефіцієнти визначаються за допомогою методу найменших квадратів (МНК), причому немає необхідності щоразу обчислювати їх заново для рівнів ряду, що входять в інтервал згладжування, так як вони будуть однаковими для кожного його положення

1. Медіанна фільтрація.

Зміст алгоритму медіанного згладжування часових рядів полягає у визначені значення медіани для рівнів інтервалу згладжування. Далі значенням медіани заміняють значення того рівня часового ряду, якому відповідає середина інтервалу згладжування. Іншими словами, медіана відповідає моменту часу t за умови, що межі інтервалу згладжування відповідають моментам часу [t − k, t + k] , тобто, коли інтервал згладжування є рівним w = 2k +1. Медіанне згладжування повністю усуває поодинокі екстремальні або аномальні значення рівнів, які віддалені один від одного як мінімум на половину величини інтервалу згладжування; зберігає різкі перепади в тенденції (ковзне середнє та експоненціальне згладжування їх змазує); ефективно усуває поодинокі рівні з дуже великими або дуже малими значеннями, які мають випадковий характер і різко виділяються серед інших рівнів.

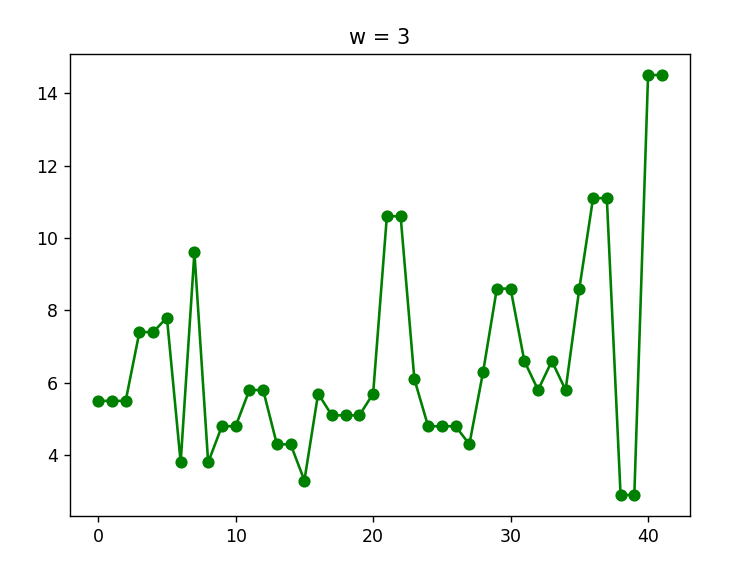


Рис. 15. Медіанне згладжування при w = 3

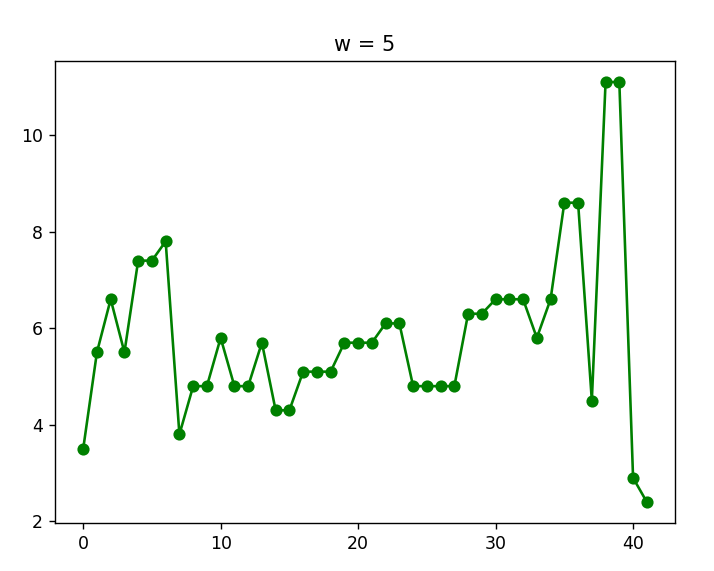


Рис. 16. Медіанне згладжування при w = 5

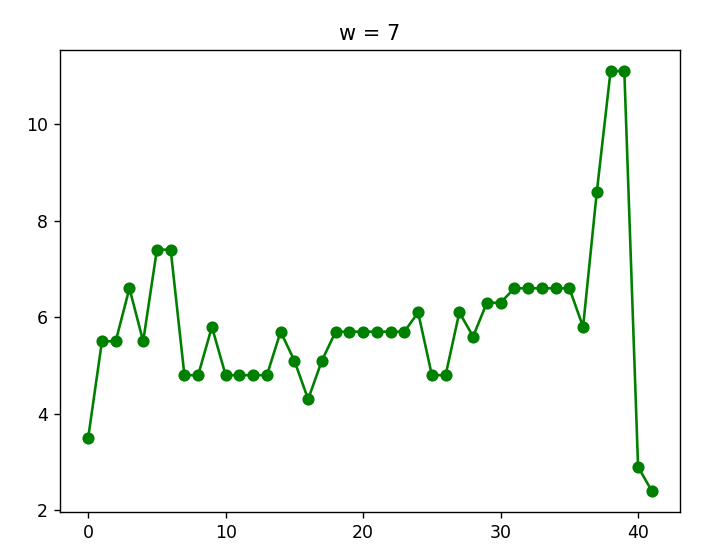


Рис. 17. Медіанне згладжування при w = 7

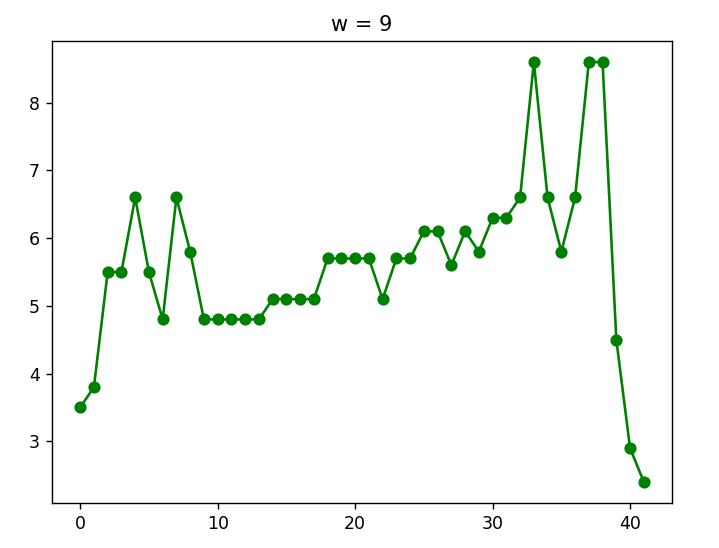


Рис. 18. Медіанне згладжування при w = 9

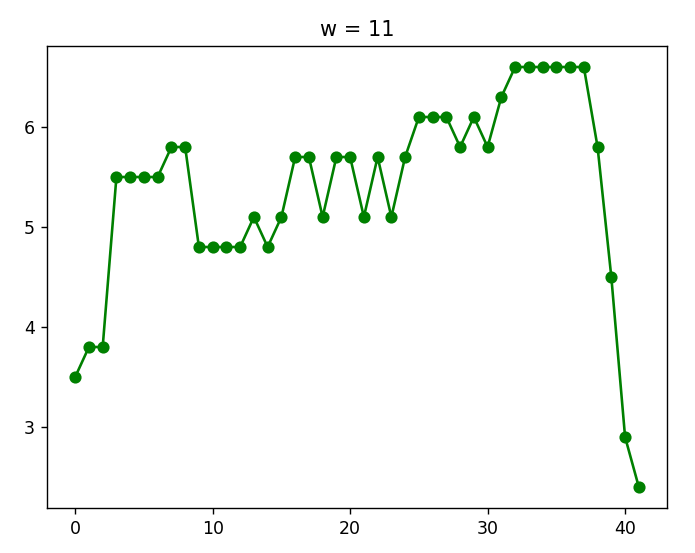
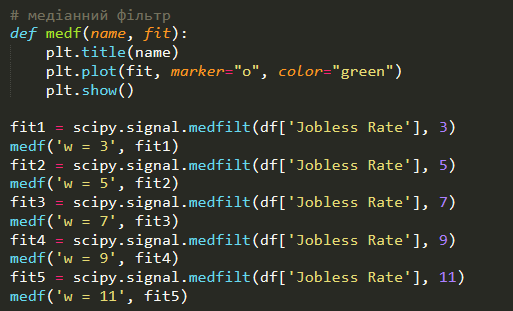


Рис. 19. Медіанне згладжування при w = 11

Код до завдання:



Кореляція даних

Графічне подання взаємозв’язку між двома досліджуваними послідовностями називається кореляційним полем або полем кореляції або діаграмою розсіювання. Графічний метод забезпечує наочне зображення форми зв'язку між цими послідовностями. Для цього, в прямокутній системі координат будують графік – по осі ординат відкладають індивідуальні значення однієї послідовності, вибраної в якості результативної ознаки Y , а по осі абсцис – індивідуальні значення іншої – факторної ознаки X . Саме сукупність точок результативної і факторної ознак називається полем кореляції.

1. Побудова кореляційного поля

Побудуємо кореляцію рейтингу інфляції до рейтингу безробіття

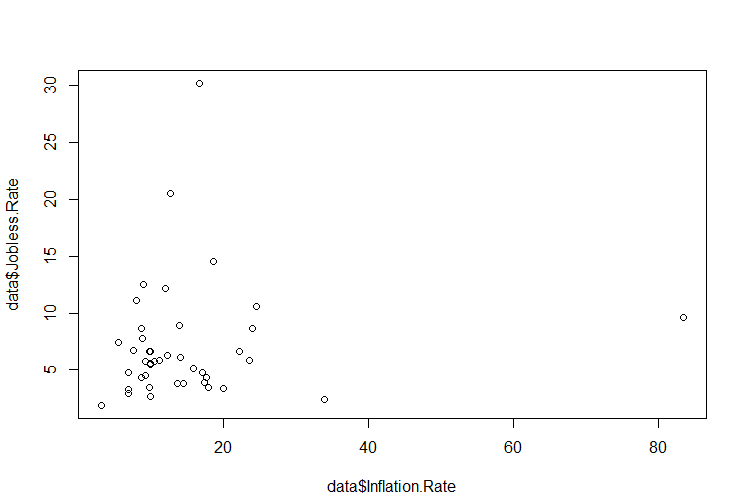


Рис. 20. Поле кореляції

Код завдання:



1. Коефіцієнт кореляції

Для кількісної оцінки тісноти зв'язку служить вибірковий коефіцієнт кореляції. Вибірковий коефіцієнт кореляції r за абсолютною величиною не перевищує одиниці. Для незалежних випадкових величин коефіцієнт кореляції дорівнює нулю, але він може бути рівний нулю для деяких залежних величин, які при цьому називаються некорельованими.

Аби визначити коефіцієнт кореляції, застосуємо три різні методи: метод Кендела, Пірсона та Спірмена.

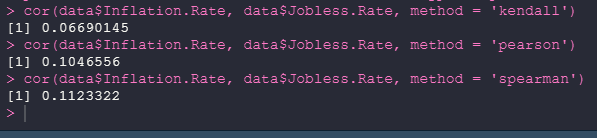


Рис. 21. Коефіцієнти кореляції

1. Кореляційне відношення

При відхиленні парної статистичної залежності від лінійної коефіцієнт кореляції втрачає свій сенс як характеристика ступеня тісноти зв'язку. У цьому випадку користуються таким виміром зв'язку як кореляційне відношення.

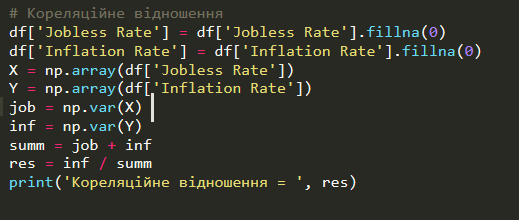




Рис. 22. Кореляційне відношення та код до завдання

Розіб’ємо послідовність на три рівні частини аби побудувати для них кореляційну матрицю.

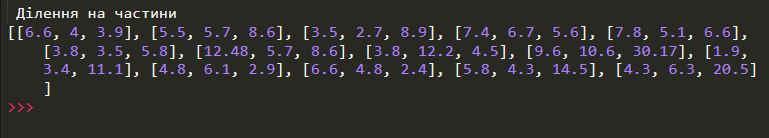
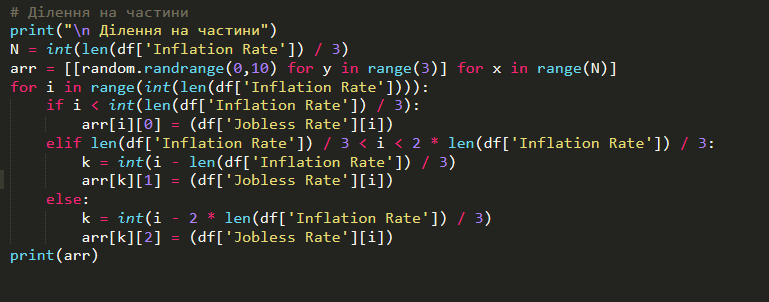


Рис. 24. Результат розбиття послідовності

Код програми:



1. Кореляційна матриця

У випадку великої кількості спостережень, коли коефіцієнти кореляції необхідно послідовно обчислювати для декількох вибірок, для зручності отримані коефіцієнти зводять в таблиці, які називають кореляційними матрицями.

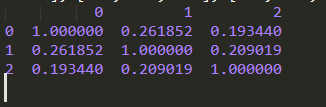
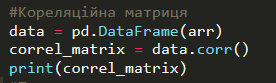


Рис. 25. Виведення кореляційної матриці

Код завдання:



Для кращого представлення кореляції було створено теплову карту.

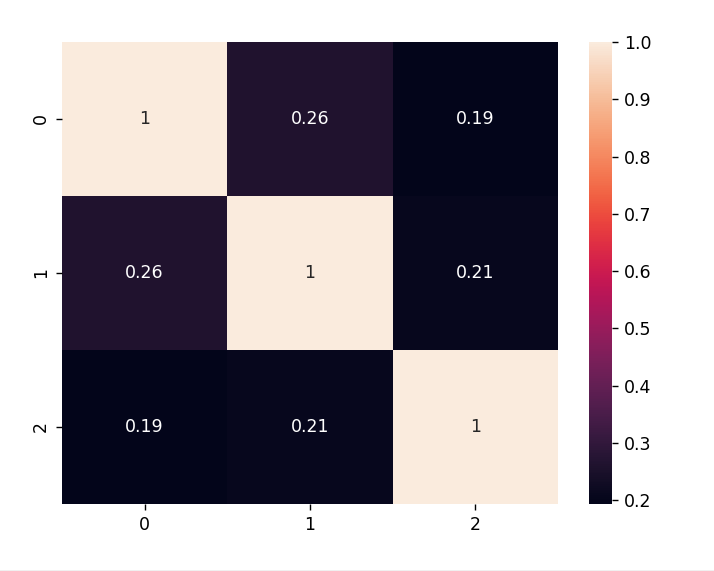
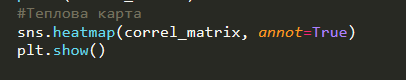


Рис. 26. Теплова карта

Код завдання:



1. Автокореляція в часових рядах.

Явище автокореляції має місце у тих випадках, коли кореляційний аналіз проводиться за даними за певні періоди, може виявитися явище автокореляції, тобто зв'язок між даними за попередні і подальші періоди. За наявності тенденції і циклічних коливань значення кожного наступного рівня ряду залежать від попередніх значень. Кореляційну залежність між послідовними рівнями часового ряду називають автокореляцією рівнів ряду.

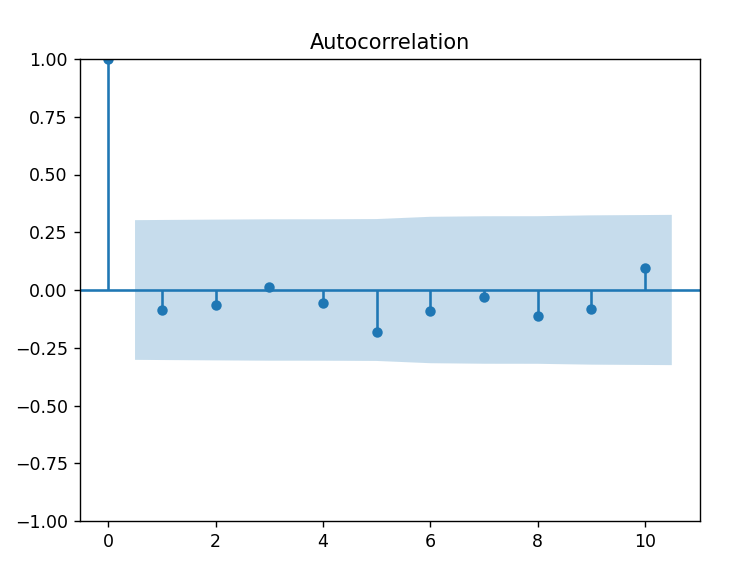


Рис. 27. Графік автокореляції

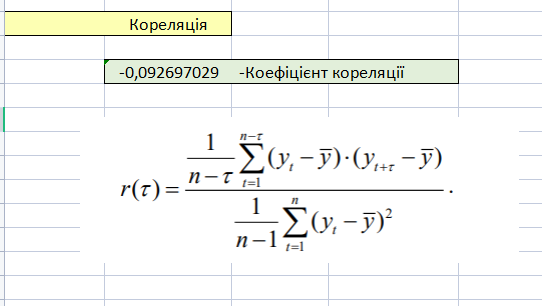
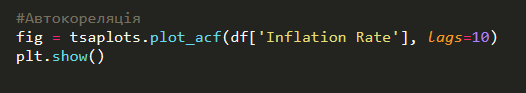


Рис. 28. Коефіцієнт кореляції та його формула

Код завдання:



Кластерний аналіз даних

Кластерний аналіз є одним із методів багатовимірного статистичного аналізу, тобто аналізу даних, коли кожне спостереження подається не одним деяким показником, а сукупністю значень різних показників. Він включає низку алгоритмів, за допомогою яких здійснюється і формування самих кластерів і розподіл об’єктів за кластерами. Кластерний аналіз, перш за все, розв’язує задачу внесення структури в дані, тобто їх групової однорідності, а також забезпечує виділення компактних, віддалених одна від одної груп об’єктів, тобто шукає «природне» розбиття сукупності на області скупчення об’єктів. Кластерний аналіз дозволяє розглядати досить значні обсяги даних, різко скорочувати і стискати їх, робити їх компактними та наочними.

Для початку, необхідно пронормувати таблицю статистичних даних, побудовану на початку роботи, аби отримати таблицю «об’єкт-властивість».



Рис. 29. Початкова таблиця

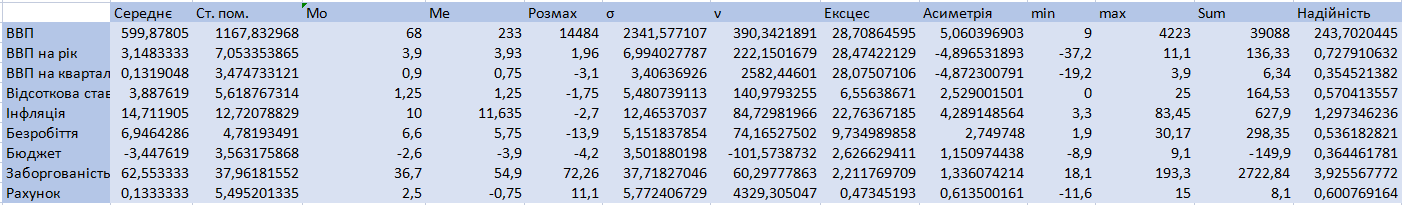
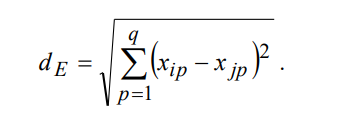


Рис. 30. Таблиця «об’єкт-властивість»

Далі будуємо матрицю близькості. Для цього розміщуємо на листі таблицю-оригінал та таблицю-копію і обчислюємо близькість між об’єктами за допомогою наступної формули:



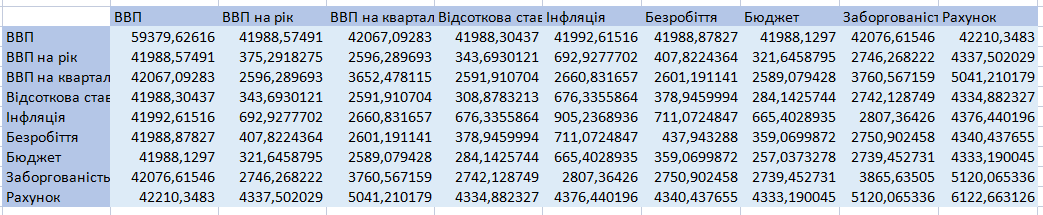
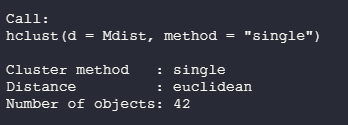


Рис. 31. Таблиця близькості

Вибір стратегії об’єднання об’єктів у групи

Процедура кластерного аналізу базується на перерахунку значень матриці близькості і, в результаті, кожного такого кроку обчислень об’єднуються об’єкти, об’єкт з групою або дві групи. Після кожного такого об’єднання розмірність матриці зменшується на одиницю, а кількість кластерів або кількість об’єктів в конкретному кластері збільшується на одиницю.

Для дослідження даного датасету було обрано стратегію найближчого сусіда. Відстань між двома групами визначається як відстань між двома найближчими елементами з цих груп. Ця стратегія монотонна і сильно стискає простір ознак, а її параметрами є аi = аj = 0.5 , в = 0 , г = −0.5.



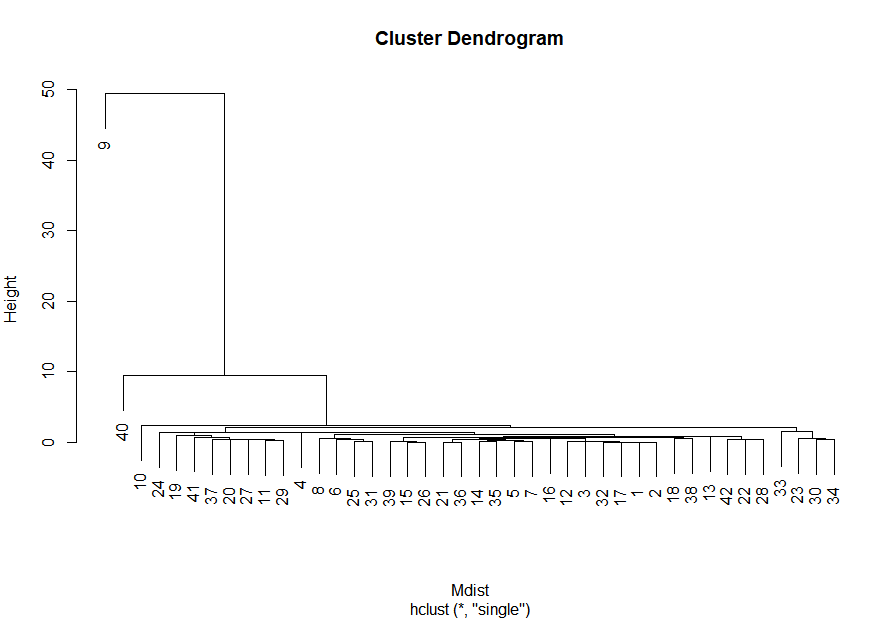


Рис. 32. Кластерна дендограма

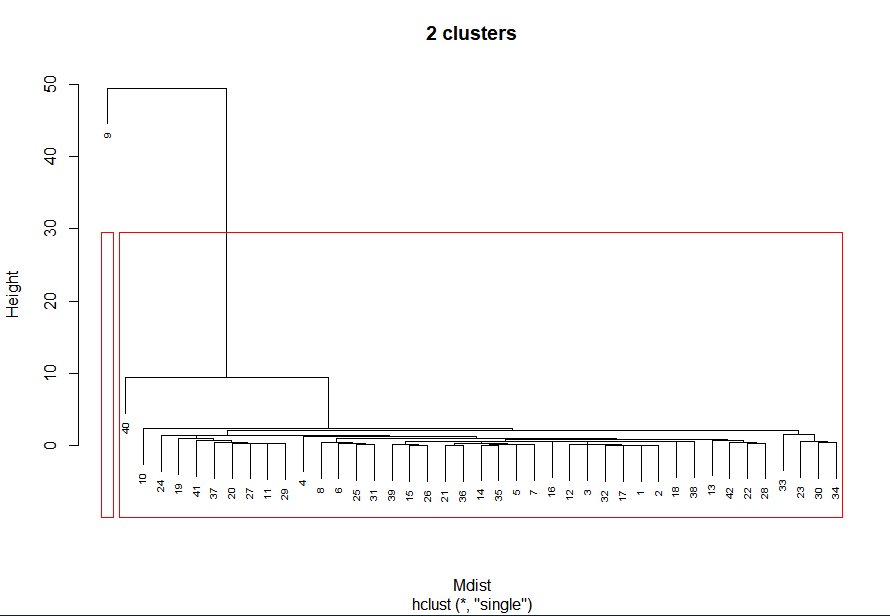


Рис. 33. Дендограма для двох кластерів

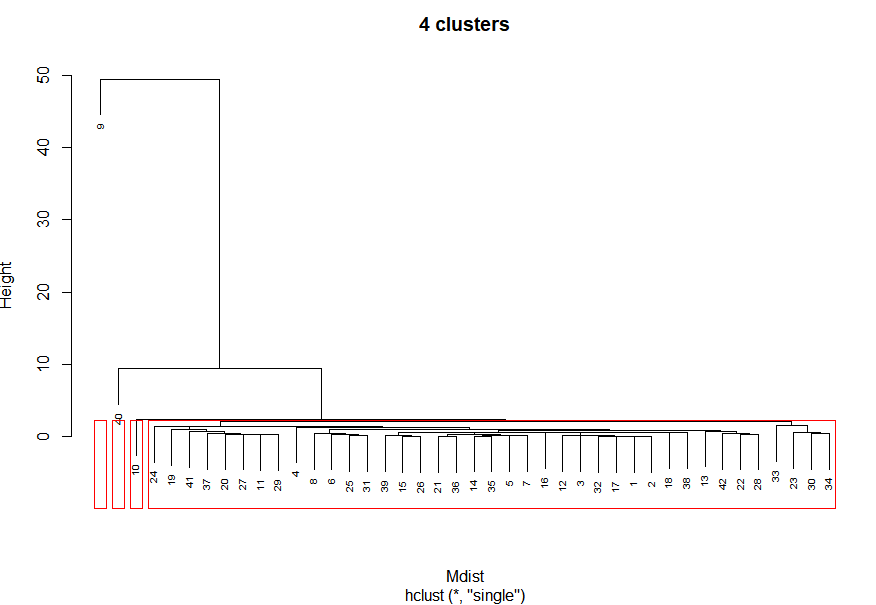


Рис. 34. Дендограма для чотирьох кластерів

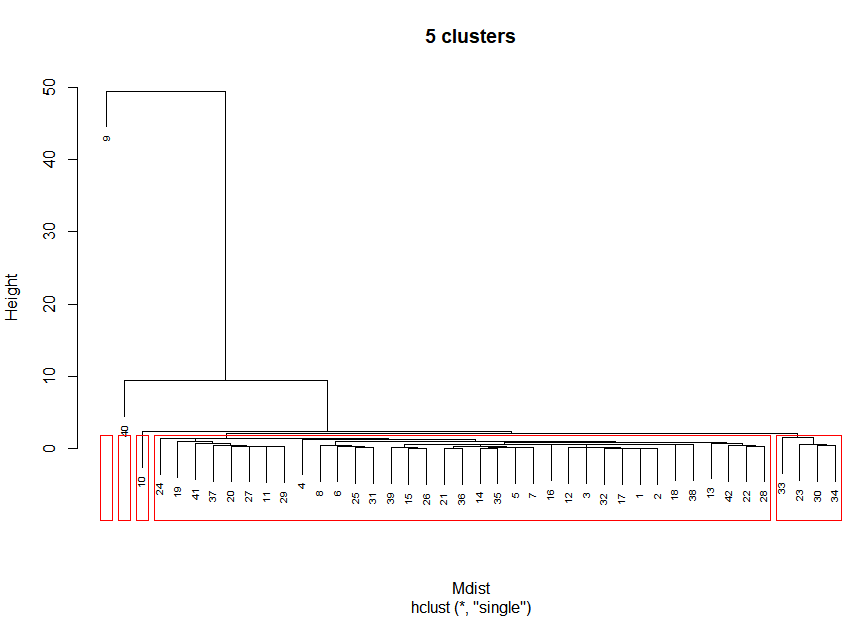
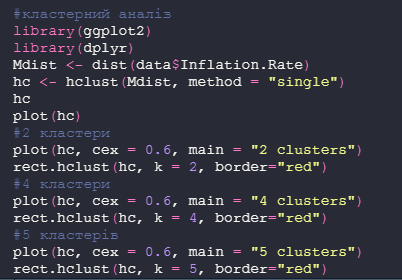


Рис. 35. Дендограма для п’яти кластерів

Код програми:



Висновок:

При виконанні даної лабораторної роботи було повністю проаналізовано заданий датасет. Спочатку було сформовано таблицю даних, відображено її графічно та обчислено основні статистичні дані, за таблицею було побудовано гістограму та кумуляту, що допомогло провести аналіз частот значень. Було вивчено основні методи висвітлення тенденції поведінки досліджуваної вибірки за допомогою згладжування часових рядів. Аби це реалізувати, було виконано згладжування за формулами Кендела, Полларда, а також експоненціальне та медіанне згладжування. Згладжування проводилось з інтервалами згладжування від w = 3 до w = 11 та параметрами α (де 0,1 ≤ α ≤ 0,3). Було проведено кореляцію даних, в процесі було побудовано кореляційне поле, визначено коефіцієнт кореляції та обчислено кореляційне відношення. Окрім того, було побудовано графіки автокореляційних функцій та кореляційну матрицю, по якій можна зрозуміти, що залежність між даними відсутня. Було проведено кластерний аналіз даних, в процесі чого було побудовано таблицю типу “об’єкт-властивість”, за допомогою якої було побудовано матрицю близькості. Останнім кроком було побудовано дендрограми для різної кількості кластерів.

Код завдання:

Код на мові R:

library(Hmisc)

#data forming

data <- read.csv("C:/Users/Ciri/Desktop/Навчання/Economy\_Indicators.csv", sep=",")

View(data)

#Graphical view

plot(data$Inflation.Rate, data$GDP, type = "p")

summary(data)

#Histogram

hist(data$Jobless.Rate)

# Просте ковзане середнє

require(smooth)

sma(data$Jobless.Rate, h=3, silent=FALSE)

sma(data$Jobless.Rate, h=5, silent=FALSE)

sma(data$Jobless.Rate, h=7, silent=FALSE)

sma(data$Jobless.Rate, h=9, silent=FALSE)

sma(data$Jobless.Rate, h=11, silent=FALSE)

plot(data$Inflation.Rate, data$Jobless.Rate, type = "p")

#коефіцієнти кореляції

cor(data$Inflation.Rate, data$Jobless.Rate, method = 'kendall')

cor(data$Inflation.Rate, data$Jobless.Rate, method = 'pearson')

cor(data$Inflation.Rate, data$Jobless.Rate, method = 'spearman')

#кластерний аналіз

library(ggplot2)

library(dplyr)

Mdist <- dist(data$Inflation.Rate)

hc <- hclust(Mdist, method = "single")

hc

plot(hc)

#2 кластери

plot(hc, cex = 0.6, main = "2 clusters")

rect.hclust(hc, k = 2, border="red")

#4 кластери

plot(hc, cex = 0.6, main = "4 clusters")

rect.hclust(hc, k = 4, border="red")

#5 кластерів

plot(hc, cex = 0.6, main = "5 clusters")

rect.hclust(hc, k = 5, border="red")

Python:

import pandas as pd

from statsmodels.tsa.api import SimpleExpSmoothing

from statsmodels.graphics import tsaplots

import matplotlib.pyplot as plt

import random

import scipy.signal

import seaborn as sns

import numpy as np

import statsmodels.api as sm

from math import \*

df = pd.read\_csv('C:/Users/Ciri/Desktop/Навчання/Economy\_Indicators.csv')

#Експоненціальне згладжування

def drawAlpha(name, fit):

print(fit.summary())

plt.title(name)

plt.plot(fit.fittedvalues, marker="o", color="blue")

plt.show()

fit1 = SimpleExpSmoothing(df['Jobless Rate'],

initialization\_method="heuristic").fit(smoothing\_level=0.1, optimized=False)

drawAlpha('Alpha = 0.1', fit1)

fit2 = SimpleExpSmoothing(df['Jobless Rate'],

initialization\_method="heuristic").fit(smoothing\_level=0.2, optimized=False)

drawAlpha('Alpha = 0.2', fit2)

fit3 = SimpleExpSmoothing(df['Jobless Rate'],

initialization\_method="heuristic").fit(smoothing\_level=0.3, optimized=False)

drawAlpha('Alpha = 0.3', fit3)

# fit Підігнати модель

# smoothing\_level - Значення простого експоненційного згладжування

# медіанний фільтр

def medf(name, fit):

plt.title(name)

plt.plot(fit, marker="o", color="green")

plt.show()

fit1 = scipy.signal.medfilt(df['Jobless Rate'], 3)

medf('w = 3', fit1)

fit2 = scipy.signal.medfilt(df['Jobless Rate'], 5)

medf('w = 5', fit2)

fit3 = scipy.signal.medfilt(df['Jobless Rate'], 7)

medf('w = 7', fit3)

fit4 = scipy.signal.medfilt(df['Jobless Rate'], 9)

medf('w = 9', fit4)

fit5 = scipy.signal.medfilt(df['Jobless Rate'], 11)

medf('w = 11', fit5)

# Кореляційне відношення

df['Jobless Rate'] = df['Jobless Rate'].fillna(0)

df['Inflation Rate'] = df['Inflation Rate'].fillna(0)

X = np.array(df['Jobless Rate'])

Y = np.array(df['Inflation Rate'])

job = np.var(X)

inf = np.var(Y)

summ = job + inf

res = inf / summ

print('Кореляційне відношення = ', res)

# Ділення на частини

print("\n Ділення на частини")

N = int(len(df['Inflation Rate']) / 3)

arr = [[random.randrange(0,10) for y in range(3)] for x in range(N)]

for i in range(int(len(df['Inflation Rate']))):

if i < int(len(df['Inflation Rate']) / 3):

arr[i][0] = (df['Jobless Rate'][i])

elif len(df['Inflation Rate']) / 3 < i < 2 \* len(df['Inflation Rate']) / 3:

k = int(i - len(df['Inflation Rate']) / 3)

arr[k][1] = (df['Jobless Rate'][i])

else:

k = int(i - 2 \* len(df['Inflation Rate']) / 3)

arr[k][2] = (df['Jobless Rate'][i])

print(arr)

#Кореляційна матриця

data = pd.DataFrame(arr)

correl\_matrix = data.corr()

print(correl\_matrix)

#sm.tsa.acf(df['Inflation Rate'])

#Теплова карта

sns.heatmap(correl\_matrix, annot=True)

plt.show()

#Автокореляція

fig = tsaplots.plot\_acf(df['Inflation Rate'], lags=10)

plt.show()