

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА

Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»
Кафедра системного аналізу та кібербезпеки

ОСВІТНЬО-ПРОФЕСІЙНА ПРОГРАМА «Системний аналіз»

ГАЛУЗЬ ЗНАНЬ 12 «Інформаційні технології»

СПЕЦІАЛЬНІСТЬ 124 «Системний аналіз»

Форма навчання денна

КВАЛІФІКАЦІЙНА МАГІСТЕРСЬКА РОБОТА

на тему:

**«ІНФОРМАЦІЙНА СИСТЕМА АНАЛІЗУ МАКРОЕКОНОМІЧНИХ
ПОКАЗНИКІВ»**

Здобувача вищої освіти Максима БОБРА _____

Науковий керівник: к.т.н., доцент

Іван ПАРХОМЕНКО _____

Робота допущена до захисту перед екзаменаційною
комісією з атестації здобувачів вищої освіти (ЕК)

Завідувачка кафедри:

доктор ф.-м. наук, професор Ірада ДЖАЛЛАДОВА _____

Київ 2024

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА

Навчально-науковий інститут

«Інститут інформаційних технологій в економіці»

Кафедра системного аналізу та кібербезпеки

ОСВІТНЬО-ПРОФЕСІЙНА ПРОГРАМА «Системний аналіз»

СПЕЦІАЛЬНІСТЬ 124 «Системний аналіз»

ПОГОДЖЕНО

Керівник проектної групи (гарант)
освітньо-професійної програми
д.фіз.–мат.н., проф.Джалладова Ірада
Агаверді-кизи

(підпис)

_____2024 р.

ЗАТВЕРДЖУЮ

Завідувач кафедри
д.фіз.–мат.н., проф.Джалладова Ірада
Агаверді-кизи

(підпис)

_____2024 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

здобувача вищої освіти Бобра Максима Олександровича
очної (денної) форми навчання

на підготовку кваліфікаційної магістерської роботи

на тему «Інформаційна система аналізу макроекономічних показників»

Тему затверджено наказом ректора Університету від "_____" _____2024 р. № _____

Кваліфікаційна магістерська робота виконується на основі наукових праць, інформаційних системи підприємства, даних, отриманих на основі інтерв'ю та опитування, даних з відкритих джерел.

Розділ 1	ПОНЯТТЯ ВІДКРИТИХ ДАНИХ
Розділ 2	ІНСТРУМЕНТИ ТА МЕТОДИ АНАЛІЗУ ДАНИХ
Розділ 3	ВИКОРИСТАННЯ МОВИ ПРОГРАМУВАННЯ R ДЛЯ АНАЛІЗУ ВІДКРИТИХ ДАНИХ
Об'єкт дослідження:	Використання математичних методів та інформаційних технологій для аналізу макроекономічних показників на основі відкритих даних.
Предмет дослідження:	Відкриті дані та макроекономічні показники.
Мета кваліфікаційної магістерської роботи:	Визначення можливостей отримання макроекономічних показників серед наборів відкритих даних, та визначення доступних засобів їх аналізу.

Конкретні завдання, які здобувач повинен виконати для досягнення поставленої мети:

Розділ 1

Провести аналіз предметної області. Визначити поняття відкриті дані, встановити їх властивості та відмінності від таких понять як відкрита та публічна інформація. Провести аналіз форматів відкритих даних

Розділ 2

Надати характеристику об'єкту дослідження. Провести аналіз основних методів статистичної обробки даних, та розглянути засоби їх аналізу.

Розділ 3

Використовуючи мову програмування «R», провести аналіз макроекономічних показників, які були оприлюднені у вигляді відкритих даних. Результат візуалізувати.

**Завдання
підготував
науковий
керівник**

І.І. Пархоменко

|
(підпис)

|
(ініціали, прізвище)

«____»_____2024 р.

**Завдання
одержав
здобувач**

М.О. Бобер

|
(підпис)

|
(ініціали, прізвище)

«____»_____2024 р.

РЕФЕРАТ

Кваліфікаційна магістерська робота містить 74 сторінки, 3 таблиці, 2 графіки, 1 гістограму, список літератури з 29 найменувань та 2 додатки.

ІНФОРМАЦІЙНА СИСТЕМА АНАЛІЗУ МАКРОЕКОНОМІЧНИХ ПОКАЗНИКІВ

Дослідження зосереджується на використанні математичних методів та інформаційних технологій для аналізу макроекономічних показників на основі відкритих даних. У сучасному світі об'єм доступної інформації стабільно зростає, що відкриває можливості для застосування аналітичних інструментів для розуміння та прогнозування економічних процесів.

Дослідження включає в себе наступні аспекти:

Відбір та обробка даних - збір та підготовка великих обсягів даних з відкритих джерел, включаючи офіційні статистичні органи, банки даних та інші джерела.

Методи математичного аналізу - використання статистичних методів, моделювання, машинного навчання та інших математичних підходів для аналізу та інтерпретації отриманих даних.

Візуалізація результатів - побудова графіків, діаграм, інфографіки та інших візуальних засобів для наочного представлення аналізу макроекономічних показників.

Прогнозування та прийняття рішень - використання отриманих результатів для прогнозування економічних тенденцій, розробки стратегій управління та прийняття рішень у сфері економіки.

Це дослідження є важливим для розвитку сучасної економічної науки та практики, оскільки воно поєднує у собі передові методи аналізу даних з широким доступом до відкритих джерел, що сприяє підвищенню якості та об'єктивності економічних досліджень.

Метою роботи є визначення можливостей отримання макроекономічних показників серед наборів відкритих даних, та визначення доступних засобів їх аналізу.

Методи досліджень базуються на аналізі поточного законодавства, дослідженні оприлюднених наборів даних та статистичних методах обробки інформації.

Розроблено алгоритм отримання інформації із наборів відкритих даних, рекомендовано програмне забезпечення для їх аналізу.

Результати роботи можуть бути використані у процесі проведення аналізу макроекономічних показників.

Ключові слова: системний аналіз, макроекономічні показники, статистичні методи аналізу, візуалізація даних, мова програмування R.

ABSTRACT

A qualified master's thesis contains 74 pages, 3 tables, 2 graphs, 1 histogram, a list of references with 29 titles and 2 supplements.

INFORMATION SYSTEM FOR ANALYSIS OF MACROECONOMIC INDICATORS

The research focuses on the use of mathematical methods and information technologies to analyze macroeconomic indicators based on open data. In the modern world, the volume of available information is steadily growing, which opens up opportunities for the use of analytical tools for understanding and forecasting economic processes.

The study includes the following aspects:

Data collection and processing - collection and preparation of large volumes of data from open sources, including official statistical bodies, data banks and other sources.

Methods of mathematical analysis - the use of statistical methods, modeling, machine learning and other mathematical approaches to analyze and interpret the obtained data.

Visualization of results - construction of graphs, charts, infographics and other visual means for a visual presentation of the analysis of macroeconomic indicators.

Forecasting and decision-making - using the obtained results to forecast economic trends, develop management strategies and make decisions in the field of economics.

This research is important for the development of modern economic science and practice, as it combines advanced methods of data analysis with wide access to open sources, which contributes to increasing the quality and objectivity of economic research.

The purpose of the work is to determine the possibilities of obtaining macroeconomic indicators among open data sets, and to determine the available means of their analysis.

Research methods are based on the analysis of current legislation, the study of published data sets and statistical methods of information processing.

An algorithm for obtaining information from open data sets has been developed, software for their analysis is recommended.

The results of the work can be used in the process of analyzing macroeconomic indicators.

Keywords: system analysis, macroeconomic indicators, statistical methods of analysis, data visualization, R programming language.

Відгук

про кваліфікаційну магістерську роботу здобувача
навчально-наукового інституту

«Інститут інформаційних технологій в економіці»
освітньо-професійної програми « Системний
аналіз»

Бобра Максима Олександровича

на тему Інформаційна система аналізу
макроекономічних показників

1. Актуальність теми: Актуальність дослідження зумовлена тим, що в сучасних реаліях, концепція відкритих даних набуває все більшої актуальності. Це зумовлено не тільки можливістю залучати до соціальних процесів більшу частину суспільства, але й про зовнішню привабливість держави. Йдеться мова про залучення іноземних інвестицій. Відкритість держави відіграє у цьому процесі важливу роль. Про це свідчить той факт, що у іноземних інвесторів все більшу популярність набувають різноманітні рейтинги, за якими можна порівняти процеси відкриття даних в тих чи інших державах. Не тільки суспільство зацікавлено у відкритті частини інформації, якою володіють державні органи, але й бізнес. Аналіз макроекономічних показників дає можливість приймати більш зважені рішення у процесі стратегічного планування діяльності підприємства. Завдяки відкритим даним кожен може отримувати необхідну інформацію про стан довкілля, економіки тощо. Використовуючи відкриті дані можливо проводити дослідження, та створювати новітні інформаційні продукти.

2. Позитивні риси кваліфікаційної роботи: розглянуто методи статистичного аналізу, приділено увагу методам та засобам аналізу даних, проведено дослідження можливості аналізу макроекономічних показників, використання мови програмування R.

3. Наявність самостійних розробок автора: В роботі представлено теоретичні методи аналізу відкритих даних. Було приділено увагу методам та засобам аналізу даних, проведено дослідження можливості аналізу

макроекономічних показників. Для досліджень автор використав мову програмування R.

4. Цінність теоретичних висновків та практичних рекомендацій: Цінність даної роботи полягає в узагальненні інформації про макроекономічні показники. Використовуючи середовище розробки R-studio був проведений статистичний аналіз деяких макроекономічних показників, а результати візуалізовані.

5. Наявність недоліків: недоліком роботи вважаю недостатня візуалізація результатів аналізу показників. Головна ціль обробки та аналізу даних це отримання нових даних для подальшого користування, таким чином візуалізація сприяє кращому сприйняттю інформації.

6. : Загальна оцінка кваліфікаційної магістерської роботи та її допущення до захисту перед ЕК:

Критерії оцінювання	Шкала, балів	Оцінка, балів
1. Логіко-структурний рівень	0 – 6 – 8 – 10	6
2. Рівень пошукової глибини	0 – 6 – 8 – 10	8
3. Науково-теоретичний рівень (розділ 1)	0 – 6 – 8 – 10	6
4. Аналітико-методичний рівень (розділ 2)	0 – 6 – 8 – 10	8
5. Конструктивний рівень (розділ 3)	0 – 6 – 8 – 10	6
6. Рівень наукової етики	0 – 6 – 8 – 10	8
7. Організаційний рівень	0 – 6 – 8 – 10	8

Загальна оцінка	0 – 70	50
-----------------	--------	----

Науковий керівник _____ К.Т.Н., доцент Пархоменко І.І.
(підпис) (науковий ступінь, учене звання) (прізвище, ініціали)

доцент кафедри системного аналізу та кібербезпеки
(посада)

“24” травня 2024 р.

Резенція
на кваліфікаційну магістерську роботу
здобувача вищої освіти
Бобра Максима Олександровича

Тема

Інформаційна система аналізу макроекономічних показників на основі відкритих даних

Актуальність теми кваліфікаційної роботи і доцільність її розроблення
Актуальність теми дослідження зумовлена необхідністю визначення можливості отримання макроекономічних показників з відкритих даних, та окреслення шляхів їх подальшого аналізу. Макроекономічні показники можуть бути важливими та відображати складні економічні процеси. Використання математичних методів дозволяє розкрити ці зв'язки та зрозуміти їх вплив на економічну динаміку.

Якість проведеного дослідження: використовуючи середовище розробки R-studio був проведений статистичний аналіз деяких макроекономічних показників, а результати візуалізовані. По результатам роботи видно, що необхідно продовжити дослідження теми відкрити даних.

Позитивні риси кваліфікаційної роботи: розглянуто методи статистичного аналізу, приділено увагу методам та засобам аналізу даних, проведено дослідження можливості аналізу макроекономічних показників, використання мови програмування R

Зауваження: недоліком роботи вважаю недостатня візуалізація результатів аналізу показників. Головна ціль обробки та аналізу даних це отримання нових даних для подальшого користування, таким чином візуалізація сприяє кращому сприйняттю інформації.

Практична значимість висновків і рекомендацій: робота виглядає логічною та цілком завершеною, відповідає встановленим вимогам, може бути допущена до захисту, а її автор заслуговує позитивної оцінки та присудження ступеня «магістр» за спеціальністю 124 «Системний аналіз».

к.т.н., доцент, доцент кафедри
кібербезпеки та захисту інформації
факультету
інформаційних технологій
КНУ імені Тараса Шевченка



Микола БРАІЛОВСЬКИЙ

Підпис доцента М.М. Браїловського засвідчую:

заступник декана з наукової роботи
факультету інформаційних технологій
Київського національного університету
імені Тараса Шевченка



Григорій ГНАТІЄНКО

«___» _____ 2024 року

ЗМІСТ

Список умовних позначень	3
ВСТУП.....	4
РОЗДІЛ 1. ПОНЯТТЯ ВІДКРИТИХ ДАНИХ	8
1.1. Визначення терміну відкриті дані	8
1.2. Характеристики відкритих даних.....	14
1.3. Формати відкритих даних	17
РОЗДІЛ 2. ІНСТРУМЕНТИ ТА МЕТОДИ АНАЛІЗУ ДАНИХ.....	23
2.1. Інструменти для обробки даних	23
2.2. Статистичні методи обробки даних.....	27
2.3. Інтелектуальний аналіз даних (DataMining)	33
2.4. Хмарні технології у роботі с даними	44
РОЗДІЛ 3. ВИКОРИСТАННЯ МОВИ ПРОГРАМУВАННЯ R ДЛЯ АНАЛІЗУ ВІДКРИТИХ ДАНИХ	49
3.1. Мова програмування R.....	49
3.2. Аналіз макроекономічних показників.....	57
3.3. Візуалізація даних	66
ВИСНОВКИ	71
ПЕРЕЛІК ПОСИЛАНЬ.....	72

Список умовних позначень

ЗУ – закон України

ІС – інформаційні системи

ІТ – інформаційні технології

ПЗ – програмне забезпечення

API – Application Programming Interface (Інтерфейс прикладного програмування)

CSV – Comma-Separated Values (Файл з розділеними комами значеннями)

GDP – Gross Domestic Product (Валовий внутрішній продукт)

JSON – JavaScript Object Notation (Формат обміну даними між додатками)

XML – eXtensible Markup Language (Розширювана мова розмітки)

БД – База Даних

ВСТУП

Суспільство та держава не існують окремо, а знаходять у постійному контакті, впливаючи один на одне, та формуючи таким чином певне середовище. Кожен громадянин є невід'ємною частиною цього середовища, тому має право отримувати інформацію про його стан. Надати характеристику такому стану, у якому знаходиться середовище, можливо завдяки процесам, які в ньому протікають. Однією із складових таких процесів є процеси, які проходять у сфері економіки.

Актуальність теми дослідження зумовлена необхідністю визначення можливості отримання макроекономічних показників з відкритих даних, та окреслення засобів їх подальшого аналізу.

Зростання обсягів даних. Завдяки цифровій трансформації та розвитку інформаційних технологій обсяги даних про економічні процеси швидко зростають. Величезна кількість цих даних доступна відкрито, що відкриває нові можливості для їх аналізу та використання.

Потреба у точних прогнозах. В умовах швидкої зміни економічного середовища для підприємств, урядових структур та інших учасників економіки важливо мати точні та надійні прогнози макроекономічних показників. Використання математичних методів та аналізу відкритих даних дозволяє покращити якість прогнозів.

Потреба у зрозумінні складних взаємозв'язків. Макроекономічні показники можуть бути взаємопов'язаними та відображати складні економічні процеси. Використання математичних методів дозволяє розкрити ці зв'язки та зрозуміти їх вплив на економічну динаміку.

Підвищення ефективності прийняття рішень. Аналіз макроекономічних показників є важливим етапом у процесі прийняття стратегічних рішень в економіці. Використання інформаційних технологій та математичних методів

дозволяє підвищити ефективність цього процесу та зробити його більш об'єктивним.

Отже, у контексті швидкого розвитку цифрової економіки та зростаючої складності економічних процесів використання математичних методів та інформаційних технологій для аналізу макроекономічних показників у вигляді відкритих даних стає надзвичайно актуальним і важливим напрямком досліджень.

На загальнодержавному рівні стан економіки описується за допомогою макроекономічних показників. Відкриття таких даних вкрай важливо, оскільки може забезпечити захист економічних інтересів громадян. Крім того у цьому зацікавлена і держава. Аналіз макроекономічних показників надає змогу довести до населення об'єктивну інформацію про поточний стан економіки, що у подальшому дасть змогу залучити більшу частину суспільства до прийняття соціально-справедливих рішень.

Мета цієї роботи полягає в розробці нових технологічних та методологічних рішень для покращення аналізу макроекономічних показників з використанням відкритих даних, що має потенціал значно покращити розуміння та управління економічними процесами.

Об'єктом дослідження є математичні методи та інформаційні технології, що застосовуються для аналізу макроекономічних показників на основі відкритих даних. Це включає в себе широкий спектр методів та технологій, таких як статистичний аналіз, математичне моделювання, машинне навчання, аналіз великих даних (Big Data), штучний інтелект (AI) та інші.

Предметом дослідження є макроекономічні показники з реєстрів відкритих даних, stat.gov.ua зокрема

Таким чином, безпосередньо дослідження проводились на відкритих наборах даних, які містять у собі ряд макроекономічних показників.

Розпорядником такої інформації виступає Державна служба статистики.

Метод дослідження аналітична обробка даних за допомогою пакету аналізу Excel a R-studio.

Наукова новизна отриманих результатів полягає у ряді аспектів:

Інтеграція математичних методів та інформаційних технологій. Дослідження спрямоване на поєднання сучасних математичних методів з інформаційними технологіями для аналізу макроекономічних показників. Це дозволяє використовувати потужність аналітичних методів разом з великими обсягами даних для отримання нових інсайтів та прогнозування економічних тенденцій.

Розробка нових моделей аналізу даних. Дослідження передбачає розробку та вдосконалення математичних моделей для аналізу макроекономічних показників на основі відкритих даних. Це може включати розробку нових методів статистичного аналізу, машинного навчання, та інших алгоритмів для прогнозування економічних процесів.

Використання відкритих даних. Дослідження акцентується на використанні відкритих даних, які є доступними широкому загалу та часто публікуються офіційними джерелами. Це відкриває нові можливості для наукового спілкування та обміну знаннями, а також робить результати дослідження більш доступними для громадськості.

Застосування до реальних економічних проблем. Дослідження спрямоване на застосування математичних методів та інформаційних технологій до реальних економічних проблем, таких як прогнозування економічних криз, аналіз впливу економічних політик, та інші. Це дозволяє дослідникам розробляти практичні рішення для покращення управління економікою.

Для реалізації поставленої мети були використані теоретичні та емпіричні методи. Зокрема, було проведено аналіз та синтез наукової та

спеціальної літератури з програмування, та статистичних методів обробки даних. Крім того було проведено вивчення та аналіз законодавства України в сфері інформації.

Результатом цієї роботи стало визначення засобів для аналізу макроекономічних показників, які були оприлюднені у вигляді відкритих даних. Було встановлено можливість використання для такого аналізу мови програмування R. Як показали дослідження використання програмної оболонки R-studio, має певні переваги, перед аналогічним програмним забезпеченням.

Практичне застосування результатів полягає у тому, що використовуючи засоби, які описані та застосовуються в кваліфікаційній роботі, кожен має можливість самостійно отримувати об'єктивну інформацію про стан економіки, для прийняття обґрунтованих управлінських рішень. Така інформація дозволяє не тільки досвідченому економісту та й пересічному громадянину сформулювати свою неупереджену думку з цього приводу, та дає уявлення про необхідність участі його у тих, чи інших економічних і соціальних процесах. Крім того аналіз макроекономічних показників може статися в нагоді при побудові бізнес-планів, плануванні економічної діяльності.

РОЗДІЛ 1. ПОНЯТТЯ ВІДКРИТИХ ДАНИХ

1.1. Визначення терміну відкриті дані

Поняття «відкриті дані» уперше з'явилося в 1995 році у документах National Oceanic and Atmospheric Administration (Національне управління океанічних та атмосферних досліджень) США.[1] Співробітники цього закладу прийшли до висновку про доцільність оприлюднення інформації про погоду та стан екології. Відкриття вказаної інформації стало першою сходиною у процесі оприлюднення даних.

Останні роки у мережі Internet з'явилося безліч різних наборів даних у різних областях, які оприлюднюються під ліцензією Open Data. Такі набори створюються як фізичними особами, та і різноманітними організаціями. У зв'язку з тим, що доступ до таких наборів є вільним, то всіх їх відносять до відкритих даних. Крім того доволі часто у якості синонімів відкритих даних використовують такі терміни як відкрита та публічна інформація. На перший погляд це дійсно зовсім ідентичні вирази, які характеризують одне і теж явище. Для розуміння що таке є відкриті дані, доцільно розглянути усі ці терміни окремо.

Термін «інформація» походить від латинського слова *informatio*, що означає відомості, роз'яснення, виклад.[2] В даний час у науковці намагаються знайти загальні властивості та закономірності, притаманні такому поняттю як інформація, але на теперішній час багато в чому визначення цього терміну залишається на інтуїтивному рівні, та отримує різні смислові наповнення в залежності від галузі застосування. Наприклад, у техніці під інформацією розуміють повідомлення, які передаються у формі знаків чи сигналів, в теорії інформації інформацією сприймають у якості відомості про об'єкти та явища навколишнього середовища.[3]

Незважаючи на проблематику з визначенням терміну інформація, її все ж наділяють певними властивостями, такими як:

1. Вірогідність. Вірогідність інформації визначається ймовірністю того, що вона відповідає реальному стану речей або вона може бути відтворена без помилок чи спотворень. Ця властивість є критичною для оцінки якості та надійності інформації.

2. Адекватність - відповідність створюваного, за допомогою отриманої інформації, об'єкта реальному процесу.

3. Повнота. Повнота інформації визначається тим, наскільки всі аспекти та деталі певної ситуації, явища або події включені та представлені у ній. Ця властивість інформації важлива для забезпечення її адекватності та використання в прийнятті рішень.

4. Своєчасність інформації вказує на те, наскільки швидко та актуально вона доступна для використання. Ця властивість є критичною для прийняття рішень в умовах змінюючогося середовища та динамічних умов.

5. Об'єктивність. Об'єктивність інформації - це властивість, що вказує на те, наскільки об'єктивно, безпристрасно та неупереджено вона представлена. Це означає, що інформація не має відтінків чи спотворень, вона базується на фактах, даних та об'єктивних спостереженнях, а не на особистих думках, уподобаннях або перекрученнях. [4]

В наші дні поки що немає строгого визначення інформації і вважається, що інформацію слід розглядати сукупність даних, фактів, знань або відомостей, які мають значення та можуть бути сприйняті, зрозумілі та використані людиною або машинами для прийняття рішень, розв'язання проблем, навчання чи розваг. Вона може бути подана у різних формах, таких як текст, зображення, звук, числа тощо, і зазвичай передається за допомогою різних комунікаційних засобів.

У цій роботі невизначеність з поняттям інформація, можливо компенсувати законодавчим визначенням цього терміну. У пункті 3 частині 1

статті 1 Закону України «Про інформацію» зазначено, що «...інформація - будь-які відомості та/або дані, які можуть бути збережені на матеріальних носіях або відображені в електронному вигляді...». [5]

Із тексту витікає, що Законодавець дотримується тієї ідеї, що інформація це все ж таки це відомості, а не знання. Крім того вказаний нормативний акт наділяє інформацію ще одною властивістю, а саме: можливість бути збереженою. У цьому можна побачити деякий сенс, оскільки не має ніякої користі від даних, які неможливо буде використати у майбутньому.

На законодавчому рівні інформацію поділяють на два великих класи, а саме: на відкриту, яку іноді ототожнюють з відкритими даними, та на інформацію з обмеженим доступом. У ст.28 Закону України «Про інформацію» зазначено, що « за порядком доступу інформація поділяється на відкриту інформацію та інформацію з обмеженим доступом. Як можна побачити у основі цього розподілу лежить порядок отримання доступу до інформації». [5]

Іноді такі обмеження мають свій сенс, оскільки відомості, які містить у себе інформація, можуть стосуватися різних сфер діяльності, і їх розголошення може завдати істотної шкоди як інтересам держави, так і окремої людини. Тому в деяких випадках доцільно обмежити коло осіб, які мають до неї доступ. Обмеження доступу до інформації може здійснюватися в інтересах національної безпеки або охорони законних прав фізичних та юридичних осіб.

Повертаючись до відкритої інформації необхідно звернути увагу на те що Закон України «Про інформацію» поділяє інформацію за правилами, за якими її можна отримати. Таким чином відкритість, в аспекті порушеного питання - це стан інформації, коли звичайній людині забезпечено можливість вільно, безперешкодно та своєчасно отримати доступ до неї, для прийняття ефективних та правильних рішень.

Можна дійти висновку, що відкрита інформація - це відомості, оприлюднення яких не можуть нанести істотної шкоди будь чийм інтересам, тому не мають обмеженого доступу і можуть бути оприлюднені. Таким чином потенційно кожен може отримати цю інформацію за умови дотримання деяких правил, наприклад подання заяви на отримання інформації.

Доступність такої інформації забезпечується її систематичним публікуванням в офіційних друкованих виданнях («Офіційний вісник України», «Відомості Верховної Ради України», газета «Урядовий кур'єр»), поширенням її засобами масової комунікації, безпосереднім її наданням зацікавленим громадянам, державним органам та юридичним особам.

У деяких випадках інформацію, яка була оприлюднена, автоматично відносять до публічної інформації, і дійсно у статті 1 Закону України «Про доступ до публічної інформації» вказано, що «...публічна інформація є відкритою, крім випадків, встановлених законом.[6] Для того щоб зрозуміти різницю між цими поняттями необхідно проаналізувати зміст поняття «публічна інформація».

Спочатку доцільно звернутися до етимології слова «публічний». Поняття «публічний» походить від латинського слова *publicus* - суспільний, народний. У латинська-українському словнику слово *publicus* перекладається як «людовий, народний, громадський, державний». [2] Дехто з науковців лексикографів зміст терміну «публічний» визначає як всенародний, оголошений, явний, відомий; організований для публіки, суспільства, народний, загальнонародний, всенародний, такий, що всім належить. [17]

Виходячи з усього вищезгаданого може здатися що публічна інформація це оприлюднені дані які належать всім. Однак не варто поспішати. Звернемося до закону України «Про доступ до публічної інформації», де у статті 1 зазначено наступне «...публічна інформація - це відображена та задокументована будь-якими засобами та на будь-яких носіях

інформація, що була отримана або створена в процесі виконання суб'єктами владних повноважень своїх обов'язків, передбачених чинним законодавством, або яка знаходиться у володінні суб'єктів владних повноважень, інших розпорядників публічної інформації, визначених цим Законом...» [6]

Законодавець чітко вказує на те, що публічна інформація може бути отримана, або створена в процесі виконання суб'єктами владних повноважень своїх обов'язків. На цьому етапі необхідно зрозуміти кого можна віднести до таких суб'єктів. Поняття суб'єктів владних повноважень надано у п.3, ч.1, ст. І Закону України «Про інформацію», а саме, «...суб'єкт владних повноважень - орган державної влади, орган місцевого самоврядування, інший суб'єкт, що здійснює владні управлінські функції відповідно до законодавства, у тому числі на виконання делегованих повноважень...». [5]

Владні управлінські функції насамперед притаманні органам виконавчої влади та органам місцевого самоврядування. Тобто можна побачити, що у зазначених нормативних актах йдеться мова про суб'єктів публічного права.

Отже, термін «публічна» не вказує на належність інформації будь кому, та на то була вона оприлюднена чи ні, і не стосується режиму доступу до неї. У даному випадку інформація поділяється за суб'єктами що її отримали. Таким чином можна говорити про те, що до публічної відносять інформацію яка була отримана суб'єктами публічного права. Ця інформація може бути як ними створена, в процесі реалізації повноважень, так і отримана від інших суб'єктів чи, наприклад, від органу правонаступником якого вони стали.

Важливо звернути увагу на те, що публічна інформація не може бути отримана в інший спосіб як в процесі виконання суб'єктами владних повноважень своїх обов'язків, передбачених чинним законодавством. Звідси

впливає ще одна властивість публічної інформації, а саме: вона повинна бути отримана законним шляхом.

Законодавство України встановлює, що публічна інформація є відкритою. Тому суспільству має право доступу до такої інформації, а держава повинна забезпечити такий доступ. У статті 5 Закону України «Про доступ до публічної інформації» встановлені способи доступу до публічної інформації. У якості одного із шляхів вказано, що «доступ забезпечується шляхом: систематичного та оперативного оприлюднення інформації на єдиному державному веб-порталі відкритих даних. Чи можна у зв'язку з цим відносити публічну інформацію до відкритих даних розглянемо нижче». [6]

В українському законодавстві не встановлено термін «відкриті дані», але використовується термін «публічна інформація у формі відкритих даних». [6]

Перш ніж перейти до розгляду питання про відкриті дані, необхідно визначитись з деякими термінами. Зокрема необхідно визначити термін дані, та встановити його відмінність від терміну інформація.

У літературі термін дані визначається як «відомості, отримані шляхом вимірювання, спостереження, логічних або арифметичних операцій і представлені у формі, придатній для постійного зберігання, передачі та автоматизованої обробки. [8] Таким чином дані це продукт, кінцевий результат процесу обробки інформації. Вважаю що буде коректно вважати, що інформація виступає у якості об'єкта дослідження, а дані є результатом цього дослідження. Такий результат може бути представлений у різноманітних форматах таких як тексти, таблиці, та представлені у буквено-цифровій, числовій, текстовій формах, які можуть зберігатися в комп'ютері, пересилатися і піддаватися обробці».

Раніш у цій роботі термін «відкритий» був визначений як режим, правила доступу до інформації. Вважаю, що термін «відкритий», у розрізі с даними, можна тлумачити як обнародований, вільний для користування.

Відкриті дані - це обнародовані набори результатів обробки інформації, які можна вільно використовувати, переробляти та поширювати. Ці дані зазвичай публікуються в машино читабельному форматі, що спрощує їхній аналіз та використання.

Таким чином принципова відмінність таких термінів як відкрита та публічна інформація від відкритих даних полягає у тому, що перші два терміни позначають відомість про деяке явище, а відкриті дані представляють з себе оприлюднені, в машиночитному форматі, набори результатів обробки такої інформації.

1.2. Характеристики відкритих даних

У 2015 році, в Мехіко, відбувся всесвітній саміт Партнерства «Відкритий уряд», на якому була схвалена Міжнародна хартія відкритих даних. [9] Цей документ містив у собі набір принципів і кращих практик для оприлюднення урядових відкритих даних. Так у якості основних були встановлені такі принципи як: відкритість за замовчуванням, оперативність і вичерпаність, доступність та готовність до використання, порівнюваність та інтероперабельність. [10]

Принципи, які були закладені у основу концепції відкритих даних та закріплені у Хартії, знайшли своє відображення і у відкритих даних, як оприлюднених наборів даних.

Головна риса відкритих даних це їх доступність. У статті 1 Положення «Про набори даних, які підлягають оприлюдненню у формі відкритих даних» зазначено, що «...доступність використання це забезпечення оприлюднення публічної інформації у формі відкритих даних у машиночитному форматі, а також забезпечення їх подальшого вільного та безоплатного використання...» [11].

Через таке визначення доступність має подвійний сенс, на це вказує використання таких термінів як «машиночитний формат» та «вільне використання».

Так доступність можна розглядати з технічної сторони. При оприлюдненні відкритих даних повинен використовуватись відкритий формат.

У статті 2 Положення «Про набори даних, які підлягають оприлюдненню у формі відкритих даних» зазначено, що «відкритий формат - це формат даних, незалежний від платформи та доступний без обмежень, які перешкоджають його повторному використанню» [11].

На законодавчому рівні забезпечується доступ до даних, незалежно від технічної оснащеності користувача. Таким чином держава забезпечує рівні можливості усім громадянам при реалізації ними своїх прав по доступу до інформації.

Другий сенс поняття доступність має юридичну природу. Так в наш час існує багато «оприлюднених наборів даних, які є відкритими на певній правовій підставі. На різноманітних веб-сайтах, у тому числі в соціальних мережах, користувачі добровільно розміщують особисту інформацію про себе, а компанії - інформацію про свою діяльність. Така інформація фактично є відкритою для обробки третіми особами».

Основною рисою вищезазначеної інформації полягає у тому, що «вона може містити об'єкти авторського права і для використання таких об'єктів необхідно мати правову підставу - дозвіл (ліцензію) від власника виключних майнових авторських прав або підставу, визначену відповідним законом. [12] у будь-якому випадку, таке використання може бути здійснене в межах відповідних ліцензій або в межах норм, що визначають умови добросовісного використання щодо конкретних об'єктів та на інших правових підставах, що визначені в законодавстві відповідної держави».

«...Що стосується відкритих даних, то можна побачити, що законодавець надає майже повну свободу дій особі, що використовує таку інформацію.

Будь-хто може копіювати, публікувати, поширювати, використовувати, у тому числі в комерційних цілях, у поєднанні з іншою інформацією, або шляхом включення до складу власного продукту, публічну інформацію у формі відкритих даних з обов'язковим посиланням на джерело отримання такої інформації. [13] Відкриті дані не обмежуються правом інтелектуальної власності, що робить їх привабливими для використання у різноманітних проектах.

Використання відкритих даних у комерційних проектах з правової точки зору можливе та практикується. Візьмемо випадок, коли компанія прагне використати у своєму продукті, який планує розповсюджувати з комерційною метою, відкриті дані різних країн, наприклад, щодо екологічного стану річок в цих країнах. Продуктом такої компанії може бути будь-що, від мобільного додатку, до великої бази даних...»

Як було зазначено вище, у мережі Internet зараз оприлюднена велика кількість різноманітних наборів даних. Всі вони не з'явилися самі по собі, а є результатом праці конкретної людини, або колективу авторів. Таким чином вказані дані, можуть виступати як об'єкти інтелектуальної власності.

Згідно статті 421 цивільного кодексу України «..суб'єктами прав інтелектуальної власності є: творець об'єкта права інтелектуальної власності та інші особи, яким належать особисті немайнові та майнові права інтелектуальної власності. [14] Особисті майнові права полягають у тому, що тільки автор має право користуватися результатами своєї праці. Як правило, такі об'єкти створюються для отримання прибутку від їх використання. Тому зовсім не дивно, що за використання даних вносити плату...»

Що стосується відкритих даних, то тут все навпаки. Основна ідея полягає у тому, що вони є суспільною власністю. В Україні у вигляді

відкритих даних оприлюднюється інформація, яка була створена державними органами, тими хто фінансується за рахунок платників податків. Таким чином кожний громадянин зробив свій внесок у формування даних, і як наслідок має право на їх використання. Звідси випливає ще одна риса відкритих даних - це їх безоплатність.

Ще одною ключовою рисою відкритих даних є їх цифровий формат та можливість автоматизованого оброблення електронними засобами. Тому до «відкритих даних не належать відповіді на запити у форматі паперових листів, скановані документи, таблиці, графіки, дашборди, інтерактивні карти, що розміщені на вебсайтах».

1.3. Формати відкритих даних

Деякі види людської діяльності пов'язані з вивченням оточуючого світу, як природу, так і соціального середовища тощо. Отриманні у процесі такої діяльності знання можуть будуть корисні у подальшій діяльності людей. Для цього знання необхідно зберегти. Для цього може бути використаний текст, цифри та малюнки тощо.

Відповідно до образу, у якому представлені дані, будуть підібрані методи для їх збереження. Кожен метод буде мати свої властивості та характеристики, а саме діапазон можливих значень, види операцій, які можливо над ними проводити та як зберігати. Таки характеристики мають назву тип даних.[15] Для відкритих даних таки типи наведені у Положенні «Про набори даних, які підлягають оприлюдненню у формі відкритих даних» де, зазначено, що дані можуть мати таки типи як текстові, структуровані, графічні, відеодані, аудіодані, дані, розроблені з використанням програми

Macromedia Flash та архіви даних. [11]

У наш час, для збереження будь-якої інформації, необхідно її записати на електронний носій. А як зазначено вище кожен інформація кожного типу

має свої властивості, тому і зберігатися буде відповідно до певних, специфічних правил. Такі правила мають назву формат даних. [16]

Недотримання порядку записування, тобто формату, може зробити дані незрозумілими не лише іншій людині, а й комп'ютеру. Тому під час визначення формату для оприлюднення набору даних необхідно зважати саме на відповідність типу даних файловому формату.

Одним із головних питань стосовно відкритих даних є питання форми їх оприлюднення. Розпорядник може оприлюднити велику кількість наборів даних у такому вигляді, що їх можна буде використовувати тільки за наявності спеціальних, глибоких знань у програмуванні. Формально законодавство порушено не буде, але істотно звужить коло користувачів, що напряду суперечить принципу доступності.

Для того щоби уникнути таких проблем в Положенні «Про набори даних які підлягають оприлюдненню у формі відкритих даних» наведено перелік можливих типів даних, та відповідні їм формати збереження.

В наш час існує дуже багато різноманітних інструментів, завдяки яким можливо відображати та зберігати дані, тому, враховуючи це, для кожного типу даних було встановлено декілька форматів. Взагалі створювати набори даних потрібно у так званих відкритих форматах, тобто таких, що не залежать від платформи, та доступні без обмежень. До таких відкритих форматів належать такі формати як CSV, JSON та XML.[17]

Публікувати текстову інформацію у вигляді відкритих даних потрібно передусім у відкритих форматах таких як RTF та ODT. Дозволяються також формати DOC(X) і PDF (з несканованим зображенням). Категорично не підходять для текстових даних формати JPG, JPEG, PNG, GIF, TIFF, а також PDF зі сканованим зображенням. Публікація текстових даних у цих форматах унеможлиблює їх обробку автоматизованими засобами.

Іноді набір даних являє собою зображення, тобто представляють графічний тип даних. Один із прикладів графічного типу даних - це графік.

Графік може бути використаний для візуалізації числової інформації та залежностей між різними змінними. Основні типи графіків:

Лінійний графік (використовується для відображення залежностей між двома або більше змінними через час або інші фактори. Лінії з'єднують точки даних, відображаючи тенденції та зміни значень змінних).

Стовпчиковий графік (використовується для порівняння значень різних категорій. Він складається з вертикальних стовпців, де висота кожного стовпця представляє значення кожної категорії).

Кругова діаграма (використовується для відображення часток чи відсоткових часток всього. Круг поділений на сегменти, пропорційні відсоткам кожної категорії).

Точкова діаграма (використовується для відображення індивідуальних значень даних та їх розподілу. Кожна точка на графіку представляє одне спостереження).

Гістограма (схожа на стовпчиковий графік, але використовується для відображення частоти різних значень однієї змінної. Стовпці не мають проміжків між собою, оскільки вони представляють діапазони значень).

Ці типи графіків можуть бути використані для візуалізації різних типів даних та залежностей між ними, що допомагає зрозуміти та аналізувати інформацію більш ефективно.

Найбільш частіше для оприлюднення використовують структуровані дані. У Положенні вказано що такі дані можуть бути представлені у таких форматах як RDF, XML, JSON, CSV, XLS(X). Зокрема, некоректною є публікація структурованих даних у форматах DOC(X) чи PDF, призначених для текстових даних, або у форматах JPG чи PNG, призначених для графічних даних.

Структуровані дані умовно поділяють на табличні та ієрархічні. Для збереження табличних даних достатньо використати формат CSV.

Вказаний

формат дозволяє записати не лише список значень, але й просту таблицю, починаючи кожен рядок таблиці з нового рядка. [18]

Поширене бачення CSV як де-факто стандарту запису табличних даних.

Справді, дані в такому форматі імпортують і експортують чи не всі електронні таблиці та бази даних. Не менш важливо, що в CSV можна зберегти лише самі дані - без формул, макросів, тощо, які не є даними, а методами їх обробки, і які тією чи іншою мірою інкорпоруєть файли електронних таблиць.

Для збереження ієрархічних даних використовують такі формати як JSON та XML.[18]

Формат JSON був розроблений для облегшення передачі інформації між додатками. Він має просту структуру: вся інформація зберігається у форматі ключ-значення.

Ключами виступають строки, які поміщені у лапки. Вони відіграють роль ідентифікаторів для значень та допомагають структурувати дані. Ключи є унікальними у рамках одного об'єкту. Значеннями в JSON можуть бути різноманітні типи даних. Завдяки такої простої структури JSON став популярним засобом обміну даними між різноманітними системами.

Ще один формат для зберігання структурованих даних, який передбачено у Положенні, це XML. Це мова розмітки, створена для кодування документів які можуть читати як люди, так і машини. Кодування робиться за допомогою тегів, які визначають структуру документу, а також спосіб зберігання та транспортування. Найпростіше XML можливо порівняти з HTML. Суттєва різниця між ними полягає у тому, що HTML фокусується на презентації контенту, а XML на збереженні даних.

Формати відкритих даних відіграють ключову роль у забезпеченні прозорості, доступності та повторного використання даних. Відкриті дані, представлені у стандартизованих і зрозумілих форматах, мають великий

потенціал для створення нових можливостей у різних галузях, включаючи науку, бізнес, громадське управління та інші. ключові аспекти впливу формату відкритих даних:

- **Покращення доступності та прозорості:** Відкриті дані, представлені у стандартизованих форматах, таких як CSV, JSON, XML, RDF, дозволяють широкому колу користувачів, включаючи громадськість, дослідників та підприємців, легко отримувати і використовувати ці дані. Це сприяє підвищенню прозорості діяльності урядових організацій і приватних компаній.

- **Забезпечення сумісності та інтеграції:** Використання відкритих форматів даних забезпечує їхню сумісність між різними системами і платформами. Це полегшує інтеграцію даних з різних джерел і створення комплексних аналітичних рішень, що базуються на великих обсягах даних.

- **Сприяння інноваціям:** Відкриті дані створюють нові можливості для інновацій. Розробники можуть створювати нові додатки і сервіси, використовуючи ці дані, що сприяє розвитку економіки знань і створенню нових бізнес-моделей.

- **Підвищення ефективності та зменшення витрат:** Відкриті дані допомагають організаціям і державним установам підвищити ефективність своєї роботи, зменшуючи витрати на збір і обробку інформації. Відкриті формати дозволяють автоматизувати процеси обміну даними і зменшити дублювання зусиль.

- **Забезпечення наукових досліджень і освіти:** Відкриті дані у стандартизованих форматах є цінним ресурсом для наукових досліджень і освітніх програм. Вони дозволяють дослідникам проводити нові види аналізу і підтверджувати результати досліджень на основі доступних даних.

- **Підтримка громадських ініціатив:** Відкриті дані сприяють участі громадян у суспільному житті. Громадські організації та активісти можуть використовувати ці дані для моніторингу та оцінки державної політики, а

також для вирішення соціальних і екологічних проблем.

Отже, формати відкритих даних відіграють важливу роль у сучасному світі, сприяючи прозорості, інноваціям, ефективності та участі громадськості. Вони забезпечують можливість повторного використання даних у різних контекстах, що створює значний позитивний вплив на розвиток суспільства та економіки.

РОЗДІЛ 2. ІНСТРУМЕНТИ ТА МЕТОДИ АНАЛІЗУ ДАНИХ

2.1. Інструменти для обробки даних

Стрімкий зріст комп'ютерних технологій обумовив появу нових методів обробки та збереження інформації. Прискорення всіх процесів, які протікають у суспільстві призвело не тільки до пришвидшення скорості інформаційних потоків а й до появи значної кількості даних. Завдяки цьому суттєво зростає попит на інструменти, завдяки котрим можливо провести обробітку великої кількості даних та швидко знайти відповіді на інтересуючи нас питання.

У наш час провести якісний та швидкий аналіз великого об'єму інформації старими методами вже не можливо. Тому зараз для обробки даних використовують різноманітне програмне забезпечення.

В загальні програмне забезпечення - це загальний термін для різних видів програм що використовуються для роботи з комп'ютерами. Воно розробляється багатьма компаніями, для різноманітних цілей, враховуючи специфіку області його застосування. Так були створені спеціалізовані програми для таких дисциплін як хімія, математика тощо.

У мережі Інтернет викладено багато наборів даних які можливо проаналізувати. Ці дані представлені у різноманітних форматах і, іноді, мають досить великі об'єми. Для прискорення процесу аналізу необхідно, щоби програмне забезпечення мало зручні функції імпорту. Крім того набори даних можуть різнитися у стилях представлення даних, а також мати деякі помилки. Тому вкрай важливо щоби програмне забезпечення мало змогу вносити доповнення чи зміни у аналізовані дані. У цій роботі приділено увагу пакетам які призначені для статистичної обробки матеріалу. Враховуючи функціонал таких програм їх умовно можна поділити на три групи: універсальні, професійні, спеціалізовані.

В основу цього розподілу покладено функціонал, який містять у собі такі пакети. Так універсальні пакети мають широкий набір статистичних функцій та зручний інтерфейс користувача. Вони більш доступні широкому колу користувачів, та підійдуть для початківця. У відмінності від попередніх, професійні пакети мають у своєму наборі додаткові методи які дозволяють зробити більше аналізу. Вказані пакети призначені для спеціалістів які мають справу з великими даними, або дуже спеціалізованими методами аналізу. Спеціалізовані ж пакети орієнтовані на конкретну вузьку область аналізу даних, містять у собі один-два розділи статистики, або методи, які використовують у конкретній галузі. Необхідною умовою використання таких програм є обов'язковий високий рівень користувача.

Охарактеризуємо найпопулярніші засоби для статистичної обробки інформації

EViews - це програмне забезпечення для економетричного аналізу та моделювання, яке часто використовується в академічних та дослідницьких установах, а також в сфері бізнесу для аналізу економічних даних. Воно дозволяє користувачам виконувати різні економетричні операції, такі як регресійний аналіз, аналіз часових рядів, моделювання та прогнозування динаміки економічних показників. [20]

GenStat. GenStat - це статистичне програмне забезпечення, яке використовується для аналізу даних, проведення експериментів та статистичного моделювання. Розроблене компанією VSN International, GenStat має довгу історію та використовується у багатьох галузях, включаючи науку, сільське господарство, медицину, екологію, соціологію та інші.[20]

MathCAD. Mathcad - це програмне забезпечення для математичного обчислення та аналізу інженерних та наукових даних. Розроблене компанією PTC, Mathcad використовується для виконання математичних обчислень,

створення та виконання складних алгоритмів, аналізу даних, моделювання, вирішення рівнянь та багато іншого.

MathCAD можна використовувати як клавіатуру так і спеціальні панелі інструментів. Зручно використовувати для навчання, обчислень і інженерних розрахунків. [20]

Mathematica. Mathematica - це комплексне програмне забезпечення для обчислювальної математики, символьних обчислень, візуалізації даних, аналізу та моделювання. Розроблене компанією Wolfram Research, Mathematica є потужним інструментом для вирішення різних завдань у науці, інженерії, комп'ютерних науках, фізиці, економіці та багатьох інших галузях. Інформацію можна подавати як у звичайному математичному вигляді, так і у вигляді функцій. За допомогою Mathematica можливо проводити розв'язання систем поліноміальних, тригонометричних та диференціальних рівнянь. Також має можливість проводити обчислювання різноманітних функцій. [20]

Matlab. Комп'ютерна оболонка для інтерактивних та командних обчислень та візуалізації. Містить у собі пакети прикладних програм для операцій над многочленами та матрицями. Функціонал програми дозволяє вирішувати алгебраїчні та диференціальні рівняння. [20]

Mesosaur. Вказана програма призначена для аналізу часових рядів. У формі часових рядів можуть виступати різноманітні дані з економіки, техніки, медицини тощо. Гранична довжина одного аналізованого ряду дорівнює шістнадцять тисяч значень. Одночасно можна аналізувати до двісті п'ятдесят шість різних змінних, проте їх сумарне значення не повинно перевищувати шістдесят тисяч чисел. У MESOSAUR реалізовані описова статистика, непараметричні тести, регресійний аналіз та інші. [20]

MS Excel. Найбільш поширений додаток з пакету офісних програм VS Office. Причиною тому є наявність російськомовної версії та тісна інтеграція

з MS Word s PowerPoint. Проте MS Excel - це електронна таблиця з досить потужними математичними можливостями, де деякі статистичні

функції є просто додатковими вбудованими формулами. Недоліком MS Excel є те, що за допомогою вказаного інструменту неможливо побудувати якісні наукові графіки. Воно підходить для накопичення даних, проміжного перетворення та попередніх статистичних обчислень. Проте остаточний статистичний аналіз необхідно робити в програмах які спеціально для цього створені. [20]

NCSS. Статистичний аналіз та побудова графіків за будь-якими даними.

Розрахований на непрофесіоналів в області статистичної обробки. Має багатовіковий інтерфейс. Усі дії користувача супроводжуються підказками, що спрощує роботу з цим інструментом. Дозволяє здійснювати статистичний аналіз введених даних та давати їм графічну інтерпретацію.

OriginPro. Потужний пакет для аналізу результатів статистичних та наукових досліджень, надає засоби для програмування та побудови графіків та діаграм. Для виконання операцій можна використовувати інструмент графічного інтерфейсу користувача, або викликати засоби у програмах. У Origin включений власний компілятор C/C++ з підтримкою і оптимізацією векторних і матричних обчислень. Система Origin створена для побудови двовимірної і тривимірної наукової графіки, яка створюється за допомогою готових шаблонів, доступних до редагування користувачем.

R. R - є безкоштовним програмним середовищем з відкритим кодом, яке орієнтовано в першу чергу на статистичні розрахунки, аналізу даних та представлення результатів в графічному вигляді. R містить у собі реалізації багатьох статистичних методів і функцій, а також графічних інструментів. Перевагою R, в порівнянні з іншими програмами, призначеними для статистичного аналізу, є доступність. Іншою сильною стороною є можливість створення високоякісних і інформативних графіків для публікації в наукових статтях та звітах.

Stata. Це потужний пакет для обробки статистичних і графічних даних.

Він позиціонується як інструмент аналізу, що призначений для фахівців, що займаються науковими дослідженнями. Завдяки гнучкій модульній структурі пакет можливо застосовувати для аналізу даних із різних областей знань: суспільні науки, економіка, медицина та інші. Stata надає користувачеві наступні можливості статистичного аналізу як: динамічні факторні моделі, лінійні моделі, моделі вибуття, узагальнені оцінки рівнянь, перевірка гіпотез і робота з оціненими моделями та інше.[20]

Statistica. Це добре збалансоване за співвідношенням «потужність/зручність» програмне забезпечення. Має широкий спектр функціональних алгоритмів і розвинену графіку. Містить у собі більш 250 статистичних функцій, які об'єднані в спеціальні модулі. Користувач має знати статистичну термінологію. Нескладні в засвоєнні пакети можна рекомендувати для статистичних досліджень будь-якої складності. У вказаному програмному забезпеченні реалізовано обмін між додатками Windows. Будь-яка інформація із Statistica може бути виведена у файл формату RTF, який редагується у Microsoft Office Word.[20]

Умовою використання більшості із згаданого програмного забезпечення є придбання ліцензії. Але в мережі деяких пакети, поряд з комерційними версіями, представлені безкоштовними альтернативними продуктами. Такі продукти мають значно менший функціонал у порівнянні з комерційними аналогами. Треба зазначити, що у багатьох випадках його досить для проведення якісної обробки даних та отримання необхідних результатів.

2.2. Статистичні методи обробки даних

Метою аналізу даних завжди є отримання знання про об'єкт дослідження, виявлення його характеристик та властивостей. Аналізувати дані можна за допомогою різних інструментів таких як математичні та

статистичні тощо. У цьому розділі будуть розглянуті базові статистичні методи аналізу даних.

В більшості випадків завдання статистичного аналізу зводиться до опису групи чи груп даних з розрахунком параметрів розподілу.

Найпростішим способом це зробити - застосувати методи описової статистики.

Описова статистика забезпечує швидке розуміння про вибірку. Основна ідея полягає у виборі одного числа, яке найкращим чином описує всі значення з набору даних. Отримавши таке число дослідник отримує інформацію про розподіл в стислій формі. При цьому за допомогою цього числа можна порівнювати між собою різні набори даних.

Основні статистичні показники, які використовує описова статистика, можна поділити на дві групи, а саме: міри центральної тенденції та міри мінливості. До мір центральної тенденції включають такі показники як середнє значення, моду та медіану, до мір мінливості відносять дисперсію, середнє квадратичне відхилення, розмах, коефіцієнт асиметрії, а також мінімальне та максимальне значення.

Центром статистичного розподілу є середнє значення. Воно визначається як середнє арифметичне вибірки, тобто, як сума всіх значень поділена на її обсяг. Для дискретних величин середнє значення розраховується за формулою:

$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

Де μ - середнє значення;

x_i - елемент вибірки;

n - кількість елементів. [21]

Мода - значення, що найбільш часто зустрічається у вибірці або наборі даних. У разі, якщо дані згруповані і побудовано розподіл частот, модою є значення, що має найбільшу частоту.[21]

Медіану можна визначити як точку на ряді розподілу. Її властивостями є то, що до цієї точки розташовано половина всіх значень, і після неї теж половина всіх значень. Тобто медіана, це значення, яке поділяє впорядкований ряд навпіл. [21]

У якості найбільш значущого елементу серед мір мінливості можна виділити дисперсію. Дисперсія - це міра розподілу даних відносно їх середнього значення. В статистиці дисперсія використовується для визначення ступеня варіації чи розмаху значень в наборі даних.

Формально, дисперсія обчислюється як середнє квадратичне відхилення кожного значення від середнього значення, з урахуванням кількості спостережень у вибірці.

Вона підраховується як середнє квадратів різниці значень розподілу і середнього розподілу, та вираховується за формулою :

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

Де σ^2 - дисперсія;

x_i - елемент вибірки;

$\bar{x}$ - середнє значення;

n - кількість елементів.[21]

Для того щоб переділити наскільки конкретні значенні у середньому відхиляються від середньої величини використовують такий показник як стандартне відхилення, яке розраховується наступним чином:

$$\sigma = \sqrt{\sigma^2}$$

Де σ - середнє квадратичне відхилення;

σ^2 - дисперсія. [21]

Міри розкиду, а саме дисперсія та середнє квадратичне відхилення, дають уяву наскільки добре середнє представляє весь набір даних. Великий

розкид даних означає що між даними є великі відмінності, тому середнє значення не є репрезентативним. [21]

Значення дисперсія використовується також і в других методах аналізу.

На практиці особливо поширеними методами є кореляційний, регресійний та факторний аналіз.

Кореляційним аналізом називають сукупність методів виявлення кореляційного зв'язку між досліджуваними ознаками. Його сутність полягає у визначенні ступеня тісноти зв'язку між двома величинами Кореляційний аналіз здійснюють на початковому етапі статистичного дослідження, так як він дає змогу встановити сам факт існування залежності між двома признаками, та оцінити тісноту їх зв'язку.[21]

Кореляцією, чи кореляційним зв'язком, між двома величинами називають наявність статистичного або імовірнісного зв'язку між ними.

Необхідно вказати на те, що наявність такого зв'язку не означає існування причинно-наслідкового зв'язку. [21]

У кореляції є дві властивості, а саме: сила і напрям. Сила кореляції визначається числовим значенням, а напрям тим, чи має кореляції позитивне значення або негативне. Позитивна кореляція вказує на те, що змінні рухаються у одному напрямку, так при зростанні однієї величини зростає і друга. Негативна кореляція навпаки вказує на протилежний напрямок двох змінних. [21]

Для отримання числової характеристики сили кореляції використовують коефіцієнт кореляції. Він показує ступінь, до якого дві змінні пов'язані між собою. Коефіцієнт кореляції може приймати значення від -1 до +1. Чим ближче значення коефіцієнта наближується до одиниці, тим міцніше зв'язок між двома факторами. Значення коефіцієнту одиниця вказує на абсолютну (функціональну) залежність між ними, а знак вказує на напрямок зв'язку: прямий чи зворотній.

На практиці, для оцінки тісноти зв'язку найбільш частіше використовують коефіцієнт Пірсона. Вказаний коефіцієнт вимірює ступінь лінійного кореляційного зв'язку між кількісними ознаками. Його розраховують за формулою:

$$r = \frac{cov_{xy}}{\sigma_x + \sigma_y}$$

Де σ - коефіцієнт кореляції;

cov_{xy} - коваріація признаку x та y;

σ_x, σ_y - дисперсії. [21]

Використання коефіцієнту Пірсона як міри зв'язку є обґрунтованим за умови що сумісний розподіл ознак є нормальним. [22]

При проведенні дослідження складних систем і процесів не досить встановити існує між декількома параметрами зв'язок чи ні. Також необхідно встановити яку він має форму. Розв'язання таких завдань є предметом регресійного аналізу. Його методи дозволяють встановити структуру і параметри моделі, яка зв'язує кількісні показники системи.

Класичний регресійний аналіз включає методи побудови математичних моделей досліджуваних систем, методи визначення параметрів цих моделей та перевірки їх на адекватність. У регресійному аналізі припускається, що зв'язок між параметрами є лінійний. До цього прийшли тому, що зв'язки між реальними показниками, частіш за все, мають лінійну форму, крім того якщо зв'язок таким не є, його можливо трансформувати таким чином, що зв'язки будуть носити лінійний характер. [23]

Звичайна процедура регресійного аналізу починається з обирання теоретичної моделі та типу залежності. Встановлюють наскільки змінні залежні одне від одного, та проводять розрахунки параметрів. Наприкінці, за допомогою статистичних критеріїв, перевіряють адекватність моделі. У

цілому доцільно першим кроком візуалізувати дані, щоб зрозуміти чи взагалі має сенс робити регресійний аналіз чи ні.

Одним з основних припущень у регресійному аналізі є то, що найкращі оцінки параметрів моделі можна отримати використовуючи метод найменших квадратів. [23] Сутність цього припущення полягає у тому, що вважається, найкращі значення невідомих параметрів можуть будуть ті, для яких сума квадратів відхилень емпіричних значень має мінімальне значення.

Найбільший інтерес у аналізі будь якої системи представляє не тільки загальна тенденція, але і можливість прогнозування поведінки цієї системи.

Для таких цілей визначають функцію регресії у вигляді математичного рівняння. Найпростішим для аналізу і найбільш дослідженим є випадок лінійної залежності між досліджуваними величинами та Y , яке має вигляд:

$$y = k \cdot x + b$$

Де y - прогнозоване значення;

k - кутовий коефіцієнт.

x - перемінне значення;

b - вільний член рівняння. [21]

Процес знаходження функції регресії називається вирівнюванням окремих значень залежної змінної. Знаючи функцію регресії можна відтворити значення залежної змінної в середині інтервалів факторів що впливають, або оцінити можливість протікання процесів поза заданим інтервалом. Ці задачі розв'язуються шляхом підстановки у рівняння регресії відповідних факторів. [23]

Ще один з видів аналізу - факторний. У цієї роботі він не застосовується, тому зупинимося на ньому кратко. Суть його полягає у тому що у випадку коли невідомий фактор впливає на декілька вимірюваних ознак останні починають корелювати між собою. Тому загальна кількість факторів, у випадку сильної кореляції між собою, може бути значно зменшеною, ніж

кількість вимірюваних. Факторний аналіз застосовують для зменшення кількості вимірюваних факторів.

2.3. Інтелектуальний аналіз даних (DataMining)

Інтелектуальний аналіз даних (англ. "Data Mining") - це процес виявлення цінної інформації з великих обсягів даних, що зазвичай включають широкий спектр даних із різних джерел і форматів. Історія його виникнення відображає еволюцію підходів до аналізу даних і розвиток технологій.

Початок 1980-х років. Термін "Data Mining" виник у ранні 1980-ті роки. На початку свого існування інтелектуальний аналіз даних в основному зосереджувався на статистичних методах, машинному навчанні та штучному інтелекті.

1990-ті роки. У цей період відбувся значний прорив в області інтелектуального аналізу даних завдяки зростанню потужності обчислювальних систем та розвитку нових алгоритмів, таких як дерева прийняття рішень, штучні нейронні мережі, генетичні алгоритми тощо.

Початок 21 століття. З впровадженням Інтернету та збільшенням кількості доступної інформації інтелектуальний аналіз даних став надзвичайно актуальним. З'явилися нові методи та інструменти для аналізу великих обсягів даних, такі як обробка текстів, аналіз соціальних мереж, візуалізація даних тощо.

Сьогодення. Інтелектуальний аналіз даних залишається однією з найбільш важливих областей в сфері аналізу даних. Він застосовується в різних галузях, включаючи науку, медицину, фінанси, маркетинг, бізнес-аналітику та інші, де великі обсяги даних потребують аналізу для виявлення закономірностей, трендів та можливостей [21].

Інтелектуальний аналіз даних включає в себе різноманітні методи та техніки для виявлення цінної інформації з великих обсягів даних.

Кластерний аналіз. Цей метод використовується для групування схожих об'єктів в одні кластери на основі їхніх характеристик. Кластерний аналіз може допомогти виявити структуру даних і розуміти взаємозв'язки між об'єктами.

Класифікація. Цей метод використовується для призначення об'єктів до певних категорій або класів на основі їхніх характеристик. Класифікація використовується для побудови моделей, які можуть передбачати клас об'єкта на основі вхідних даних.

Регресійний аналіз. Цей метод використовується для виявлення зв'язку між залежною змінною (вихідними даними) та однією або декількома незалежними змінними (вхідними даними). Регресійний аналіз допомагає встановити функціональні залежності між змінними та прогнозувати значення вихідних даних на основі вхідних.

Асоціативний аналіз. Цей метод використовується для виявлення зв'язків та асоціацій між різними елементами в наборі даних. Найбільш відомим прикладом асоціативного аналізу є виявлення асоціативних правил в магазинних даних, таких як "якщо клієнт купує продукт А, то він ймовірно також купить продукт В".

Методи машинного навчання. Це широкий клас методів, які використовуються для навчання комп'ютера виконувати певні завдання, не програмуючи його явно. Методи машинного навчання включають в себе навчання з учителем (наприклад, класифікація та регресія) та навчання без учителя (наприклад, кластерний аналіз та зменшення розмірності даних).

Так як після застосування інтелектуального аналізу даних можуть з'являтися не тільки корисні результати, а й хибні, то в такому випадку можна користуватися методами машинного навчання. За визначенням,

машинне навчання - це наука, що вивчає комп'ютерні алгоритми, які автоматично покращуються під час роботи.

Інтелектуальний аналіз даних (англ. Data Mining) - це складний процес, який включає кілька етапів для виявлення цінної інформації з великих обсягів даних.

Основні етапи:

Збір даних. Перший етап полягає в зборі необхідних даних для подальшого аналізу. Це може бути внутрішня або зовнішня інформація, яка зазвичай збирається з різних джерел і форматів.

Очищення даних. Після збору даних їх потрібно очистити від помилок, дублікатів, пропущених значень та інших аномалій, які можуть вплинути на результати аналізу.

Вибір даних. На цьому етапі вибираються підмножини даних, які найбільш важливі для аналізу, або які відповідають певним критеріям або цілям дослідження.

Трансформація даних. Дані можуть бути трансформовані для підготовки до подальшого аналізу. Це може включати шкалювання, нормалізацію, кодування категорійних змінних та інші операції.

Моделювання даних. На цьому етапі будуються математичні моделі або алгоритми для аналізу даних та виявлення закономірностей або патернів.

Оцінка та валідація моделей. Одержані моделі або алгоритми оцінюються на їхню ефективність та точність за допомогою тестових даних або крос-валідації.

Інтерпретація та використання результатів. На останньому етапі результати аналізу даних інтерпретуються та використовуються для прийняття рішень, розробки стратегій або планування майбутніх дій.

Ці етапи можуть варіювати в залежності від конкретного завдання або проекту, а також використовувати різні методи та техніки аналізу даних.

Однак, вони загалом відображають загальний процес інтелектуального аналізу даних від збору і до використання інформації для прийняття рішень.

Прогнозоване моделювання - це процес розробки математичних моделей або алгоритмів для прогнозування майбутніх значень на основі наявних даних. Цей процес включає в себе декілька етапів, які допомагають створити ефективні та точні моделі прогнозування.

Етапи прогнозованого моделювання:

Збір даних. Перший етап полягає у зборі всієї необхідної інформації, яка буде використовуватися для прогнозування. Це може бути числова або категоріальна інформація, що відображається в різних форматах і джерелах.

Очищення та підготовка даних. На цьому етапі дані перевіряються на наявність помилок, пропущених значень, дублікатів та інших аномалій. Дані можуть бути очищені, виправлені та підготовлені для подальшого аналізу.

Розділення даних на навчальний та тестовий набори. Для оцінки ефективності моделі потрібно розділити наявні дані на дві частини: навчальний набір, на якому модель буде навчатися, і тестовий набір, на якому буде оцінюватися її ефективність.

Вибір моделі. На цьому етапі вибирається тип моделі, яка найкраще підходить для конкретного завдання прогнозування. Це може бути лінійна регресія, дерева рішень, нейронні мережі, методи ансамблювання тощо.

Навчання моделі. Обрана модель навчається на навчальному наборі даних з метою встановлення зв'язків між вхідними та вихідними змінними.

Оцінка моделі. Після навчання моделі її ефективність оцінюється на тестовому наборі даних. Це може включати оцінку точності, чутливості, специфічності та інших метрик ефективності моделі.

Налаштування параметрів моделі. У деяких випадках параметри моделі можуть бути налаштовані для покращення її ефективності.

Прогнозування майбутніх значень. За допомогою навченої моделі прогнозуються майбутні значення на основі нових даних або об'єктів.

Оцінка та вдосконалення. Результати прогнозування оцінюються і використовуються для вдосконалення моделі або прийняття рішень в певній галузі. В разі необхідності процес може повторюватися для покращення результатів.

Ці етапи допомагають побудувати ефективні та точні прогнозні моделі на основі наявних даних, які можуть бути використані для прийняття рішень та планування майбутніх дій.

Логічний метод - це метод аналізу та розуміння інформації, що ґрунтується на застосуванні принципів логіки та розумового мислення. Основна ідея полягає в тому, щоб використовувати правила логіки для виведення висновків з вихідних premisів або фактів.

Метод крос-табуляції є ефективним інструментом для аналізу зв'язків між двома або більше категоріальними змінними. Цей метод дозволяє візуалізувати та аналізувати взаємозв'язки між різними факторами у вигляді таблиць, які називаються крос-таблицями. Основні методи крос-табуляції включають:

Стандартна крос-таблиця. Цей метод використовується для аналізу зв'язку між двома категоріальними змінними. Дані представляються у вигляді таблиці, де одна змінна розміщена по вертикалі, інша - по горизонталі, і кожна комірка таблиці містить кількість спостережень, що відповідають кожній комбінації значень цих змінних.

Крос-таблиця з відсотками. Цей метод використовується для порівняння відсоткових співвідношень між різними категоріями. Замість абсолютних значень кількостей спостережень використовуються відсотки від загальної кількості спостережень у кожній комірці таблиці.

Крос-таблиця зі зведеною статистикою. Цей метод включає в себе розрахунок додаткових статистичних метрик, таких як середні, середньоквадратичні відхилення або інші міри центральної та дисперсійної тенденції для кожної комбінації значень змінних.

Крос-таблиця з ефектами. Цей метод використовується для виявлення взаємодії між різними змінними та визначення наявності статистично значущих ефектів. Для цього можуть використовуватися різноманітні статистичні тести, такі як χ^2 -квадрат, Фішера, статистика залежності Крамера тощо.

Метод візуалізації даних - це процес представлення інформації у вигляді графіків, діаграм, карт або інших візуальних елементів з метою кращого розуміння даних та виявлення патернів, тенденцій і взаємозв'язків. Візуалізація даних дозволяє швидко виявити ключові особливості даних, спростити складність інформації та робити зрозумілі висновки.

Лінійні графіки. Вони використовуються для візуалізації змін певної змінної в залежності від іншої змінної або часу. Наприклад, графік може показати динаміку змін цін на акції у певний період часу.

Стовпчасті та кругові діаграми. Вони використовуються для порівняння величин різних категорій або частин до загальної суми. Наприклад, кругова діаграма може показати розподіл доходів на різні види витрат.

Гістограми. Вони використовуються для візуалізації розподілу числових даних у вигляді стовпців. Наприклад, гістограма може показати розподіл висоти людей у певній групі.

Кореляційні матриці та теплові карти. Вони використовуються для візуалізації зв'язків між числовими змінними у вигляді матриці або теплової карти. Це дозволяє виявити кореляції між різними змінними.

Графіки розсіювання. Вони використовуються для візуалізації взаємозв'язків між двома або більше змінними. Наприклад, графік розсіювання може показати залежність між витратами на рекламу та продажами.

Карти і географічні візуалізації. Вони використовуються для візуалізації географічних даних та взаємозв'язків на карті. Наприклад, карта може показати географічний розподіл клієнтів компанії або локації її філій.

Методи, засновані на рівняннях, широко використовуються для аналізу даних та вирішення різних завдань у науці, інженерії, фінансах, економіці та інших галузях.

Лінійна регресія. Цей метод використовується для встановлення лінійного зв'язку між залежною та однією або декількома незалежними змінними. Лінійна регресія дозволяє прогнозувати значення залежної змінної на основі значень незалежних змінних.

Логістична регресія. Цей метод використовується для прогнозування ймовірності виникнення події на основі значень деяких вхідних змінних. Логістична регресія застосовується у задачах класифікації, де потрібно визначити, до якої категорії належить об'єкт.

Аналіз часових рядів. Цей метод використовується для аналізу змін в часі. Він дозволяє виявити тенденції, сезонні варіації, циклічність та інші закономірності в часових даних.

Методи оптимізації. Ці методи використовуються для пошуку оптимального рішення в умовах обмежень. Вони дозволяють максимізувати або мінімізувати деякі цільову функцію за умови деяких обмежень.

Методи чисельного розв'язку рівнянь. Ці методи використовуються для знаходження чисельних розв'язків складних математичних рівнянь, які не можна вирішити аналітично.

Методи регресії з використанням машинного навчання. Ці методи використовуються для побудови складних моделей регресії з використанням алгоритмів машинного навчання, які можуть автоматично виявляти складні взаємозв'язки в даних.

Логічні методи, або методи логічної індукції, включають:

Нечіткі запити і аналізи. Цей метод використовується для роботи з нечіткими даними або змінними, які можуть мати різні ступені приналежності до різних категорій. Він дозволяє враховувати нечіткість та неоднозначність у визначенні категорій і виробляти відповідні аналізи.

Дерева рішень. Цей метод використовується для моделювання рішень у вигляді дерева, де кожен вузол представляє можливий вибір або рішення, а гілки відображають альтернативні шляхи. Дерева рішень використовуються в прийнятті рішень, класифікації даних та інших задачах.

Символьні правила. Цей метод використовується для формалізації знань у вигляді правил, які можна використовувати для прийняття рішень або аналізу даних. Символьні правила зазвичай мають вигляд "ЯКЩО/ТО, ТО/ІНАКШЕ", де вказується умова і відповідний висновок.

Генетичні алгоритми. Цей метод базується на еволюційних принципах та використовує механізми відбору, мутації та схрещування для пошуку оптимальних рішень. Вони часто застосовуються для оптимізації функцій, вирішення задач комбінаторної оптимізації та інших завдань.

Методи цієї групи, мабуть, найбільше інтерпретуються, адже вони оформляють знайдені закономірності в досить прозорому вигляді з точки зору користувача. Отримані правила можуть включати безперервні і дискретні змінні. Слід зауважити, що дерева рішень можуть бути легко перетворені в набори символьних правил шляхом генерації одного правила по шляху від кореня дерева до його термінальної вершини. Дерева рішень і правила фактично є різними способами вирішення однієї задачі і відрізняються лише за своїми можливостями. Крім того, реалізація правил здійснюється більш повільними алгоритмами, ніж індукція дерев рішень.

Крос-таблиці. Цей метод включає побудову таблиць, де одна змінна розташована в рядках, а інша - в стовпцях, і кожна клітина таблиці показує кількість спостережень, які відповідають певній комбінації значень цих

змінних. Крос-таблиці дозволяють виявляти взаємозв'язки між категоріальними змінними та робити порівняльний аналіз.

Агенти (Agent-based modeling). Цей метод використовується для моделювання поведінки і взаємодії індивідуальних агентів у системі. Ці агенти можуть мати різні характеристики та стратегії, і їх взаємодія може призводити до складних емерджентних явищ.

Баєсові мережі (Bayesian Networks). Цей метод використовується для моделювання ймовірностних зв'язків між змінними за допомогою графічної моделі. Баєсові мережі дозволяють виявляти ймовірність виникнення певних подій на основі відомих фактів або даних.

Крос-табуляція - це потужний метод аналізу даних, який допомагає виявляти зв'язки між змінними та робити висновки на основі цих взаємозв'язків. Агенти та Баєсові мережі - це інші методи аналізу даних, які можуть використовуватися у різних контекстах, але вони зазвичай не включаються в методи крос-табуляції.

Методи на основі рівнянь висловлюють виявлені закономірності у вигляді математичних виразів - рівнянь. Отже, вони можуть працювати лише з чисельними змінними, і змінні інших типів повинні бути закодовані відповідним чином. Це дещо обмежує застосування методів даної групи, проте вони широко використовуються при вирішенні різних завдань, особливо завдань прогнозування.

Основні методи даної групи: статистичні методи та нейронні мережі.

Статистичні методи найбільш часто застосовуються для вирішення завдань прогнозування. Існує безліч методів статистичного аналізу даних, наприклад:

Гармонійний аналіз. Цей метод використовується для аналізу періодичних або циклічних змін у даних, наприклад, коливань частоти, амплітуди або фази. Він допомагає виявляти складові гармоніки у динаміці даних і розкривати їх структуру.

Кореляційно-регресійний аналіз. Цей метод використовується для вивчення взаємозв'язків між змінними і прогнозування значень однієї змінної на основі значень інших змінних. Він дозволяє виявляти кореляційні зв'язки між змінними та будувати прогностичні моделі.

Кореляція рядів динаміки. Цей метод використовується для вивчення кореляційних зв'язків між динамічними рядами даних. Він дозволяє виявляти схожі тенденції або зв'язки між різними часовими рядами.

Виявлення тенденцій динамічних рядів: Цей метод використовується для аналізу та виявлення тенденцій у динаміці даних. Він дозволяє виявляти зростаючі, спадні або стаціонарні тенденції у часових рядах і робити прогнози на їх основі.

Статистичні методи представляють собою чотири взаємопов'язані розділи:

- Попередній аналіз природи статистичних даних - полягає в перевірці гіпотез стаціонарності, незалежності, нормальності, однорідності, а також оцінці виду функції розподілу, її параметрів і т.д.
- Виявлення зв'язків та закономірностей - кореляційний аналіз, лінійний та нелінійний регресійний аналіз і т.д.
- Багатовимірний статистичний аналіз - кластерний аналіз, компонентний аналіз, лінійний та нелінійний дискримінантний аналіз і т.д.

Динамічні моделі та прогноз на основі часових рядів.

Кібернетичні методи інтелектуального аналізу даних - це набір підходів, що об'єднані ідеєю комп'ютерної математики та використанням теорії штучного інтелекту. До цієї групи належать такі методи:

- еволюційне програмування;
- штучні нейронні мережі (розпізнавання, кластеризація);
- генетичні алгоритми (оптимізація);
- асоціативні правила (пошук аналогів, прототипів);
- дерева рішень;

- нечітка логіка;
- системи обробки експертних знань.

Методи інтелектуального аналізу даних також можна класифікувати за завданнями інтелектуального аналізу даних.

Відповідно до такої класифікації можна виділити дві групи. Перша група – це підрозділ методів, що займається вирішенням завдань сегментації (тобто завдання класифікації і кластеризації) та прогнозування.

Відповідно до другої класифікації методи інтелектуального аналізу даних можуть бути спрямовані на отримання описових і прогнозованих результатів.

Описові методи служать для знаходження шаблонів або зразків, що описують дані, які піддаються інтерпретації з точки зору аналітика.

До методів, спрямованих на отримання описових результатів, відносяться ітеративні методи кластерного аналізу, в тому числі:

- алгоритм k-медіани;
- алгоритм k-середніх;
- методи крос-табличної візуалізації;
- ієрархічні методи кластерного аналізу;
- карти Кохонена, що само організуються;
- різні методи візуалізації.

Прогнозовані методи використовують значення одних змінних для передбачення та прогнозування невідомих (пропущених) або майбутніх значень інших змінних.

До методів, що спрямовані на отримання прогнозованих результатів, відносяться такі методи:

- дерева рішень;
- лінійна регресія;
- нейронні мережі;
- метод найближчого сусіда;

- метод опорних векторів;
- та ін.

Різні методи інтелектуального аналізу даних характеризуються певними властивостями, які можуть бути визначальними при виборі методу аналізу даних. Методи можна порівнювати між собою, оцінюючи характеристики їх властивостей.

Серед основних властивостей і характеристик методів інтелектуального аналізу даних існують наступні:

- масштабованість;
- точність;
- гнучкість;
- швидкість;
- інтерпретованість;
- трудомісткість;
- популярність.

Оцінка кожної з характеристик проведена за наступними категоріями (в порядку зростання):

- надзвичайно низька;
- дуже низька;
- низька/нейтральна;
- нейтральна/низька;
- нейтральна;
- нейтральна/висока;
- висока;
- дуже висока [23].

2.4. Хмарні технології у роботі с даними

У наш час практично не можливо уявити світовий ринок ІТ послуг без «хмарних» технологій. Хмарні сервіси дозволяють є не тільки гнучкими

інструментами зберігання та керувати інформаційною інфраструктурою, економити на обладнанні, резервному копіювання та технічній підтримки, але також отримувати доступ до величезної кількості послуг та даних з будь-якої точки земної кулі. Можливість підключення - це основний принцип хмарної технології.

Хмарні технології у широкому розумінні - це спосіб надання обчислювальних потужностей як послугу через мережу Інтернет. Однією з переваг таких технологій це те що не потрібно самостійно налаштовувати

обладнання, це покладено на постачальника послуг. Користувач отримає майданчик який називають «хмара» де може робити будь-що зберігати дані, або проводити наукові дослідження. По своїй суті це сервер, який може бути розташований в одному з центрів обробки даних по всьому світу. Доступ до такого серверу здійснюється за допомогою мережі Інтернет.

Хмарні обчислення стали можливими завдяки віртуалізації, яка дозволяє створити змодельований цифровий віртуальний комп'ютер, який подібний до фізичного, з власним апаратним забезпеченням. Таку симуляцію називають віртуальна машина. Сервер може одночасно запускати безліч таких машин.

Віртуальні машини ізольовані одна від одної тому файли та програми однієї машини не видно іншим, навіть якщо вони знаходяться на одному фізичному сервері. Користувачі отримують доступ до сервісів через браузер, або додаток за допомогою Інтернету. [24]

Хмарні технології надають широкий спектр послуг де можна підібрати те що потрібно користувачеві. Існує безліч сервісів, серед котрих є і такі, що надають змогу проводити статистичне оброблення даних. Про деякі з них буде вказано нижче.

Одним з багатьох інструментом для аналізу даних є Google Colaboratory, це один з хмарних сервісів від Google Research. Вказаний сервіс

дозволяє будь-кому писати код в своєму редакторі та запускати його з браузера. Зокрема він підтримує мову програмування Python. [25]

Ця послуга заснована на Jupyter Notebook та розміщена абсолютно безкоштовно. Доступ до цієї послуги здійснюється за допомогою облікового запису Gmail, і не вимагає завантаження чи установки Jupyter.

Коли користувач отримує доступ до Google Colaboratory за допомогою свого облікового запису, він отримує віртуальну машину, на якій може запускати свій код, ізольований від інших користувачів і ресурсів. Отримані результати можна зберегти у блокноті GDrive.

MapleCloud. Це потужний математичний web-інструмент, що призначений для вирішення низки математичних задач. Даний сервіс спрямований не тільки на окремі розрахунки але й на розв'язання задач з математичного аналізу та інших дисциплін пов'язаних з ним. Крім того він

обладнаний потужною системою онлайн допомоги, яка структурована за розділами дисциплін, серед яких наявний розділ «Статистика та аналіз даних». [26]

MapleCloud дає змогу поширювати власні документи іншим користувачам, навіть якщо в них не встановлено жодного з додатків Map.

Сервіс доступний через довільний веб-браузер як зі стаціонарних комп'ютерів так і з інших мобільних пристроїв.

У MapleCloud наявні численні математичні додатки, що дають змогу проводити різної складності обчислення в таких галузях як статистика, фізика, інженерія та ін. Варто зазначити, що цей інструмент можна застосовувати для онлайн навчання без придбання ліцензійного програмного забезпечення.

Scilab - пакет прикладних математичних програм, що являє собою потужне середовище для технічних та наукових розрахунків, а саме для розв'язання нелінійних рівнянь та систем, розв'язання задач лінійної алгебри, розв'язання задач оптимізації; диференціювання та інтегрування; обробка

експериментальних даних (інтерполяція, апроксимація, метод найменших квадратів).

Scilab був спроектований як відкрита система, у яку користувачі можуть додавати свої типи даних та операції.[27]

Scilab має хмаро орієнтований додаток такий як ScilabCloud. В «хмарі» цей додаток поширюється як сервіс з відкритим інтерфейсом, що може вільно використовуватись для інтерпретації коду в хмарах у якості SaaS (програмне забезпечення як послуга).

ScilabCloud використовує централізацію даних, а також приховування коду від кінцевого користувача, забезпечуючи таким чином захист інтелектуальної власності.

Переваги, які надають хмарні технології, дають підставу вважати, що попит на хмарні сервіси у майбутньому буде тільки зростати.

Хмарні технології стали невід'ємною частиною сучасної обробки даних, пропонуючи численні переваги для організацій різних розмірів та галузей. Ось кілька ключових аспектів їхнього впливу на роботу з даними:

- Гнучкість і масштабованість: Хмарні платформи дозволяють легко масштабувати обчислювальні ресурси відповідно до потреб бізнесу. Це означає, що компанії можуть швидко реагувати на зміну обсягів даних без необхідності значних інвестицій в апаратне забезпечення.

- Висока доступність і надійність: Хмарні провайдери забезпечують високу доступність і відмовостійкість своїх сервісів. Це означає, що дані та додатки доступні користувачам практично завжди, що є критично важливим для безперервної бізнес-діяльності.

- Зниження витрат: Використання хмарних технологій дозволяє знизити капітальні витрати на ІТ-інфраструктуру та зменшити операційні витрати за рахунок моделі "плати за використання". Це робить хмарні рішення економічно вигідними для багатьох компаній.

- Покращена безпека і відповідність вимогам: Хмарні провайдери

інвестують значні ресурси в забезпечення безпеки своїх платформ, включаючи засоби шифрування, управління доступом та моніторинг. Багато хмарних сервісів також допомагають компаніям відповідати різноманітним регуляторним вимогам, таким як GDPR, HIPAA тощо.

- Інновації і швидкість розгортання: Хмарні технології сприяють швидшому впровадженню нових технологій і інновацій, таких як машинне навчання, аналітика великих даних, інтернет речей (IoT). Це дозволяє компаніям швидко отримувати конкурентні переваги.

- Колаборація і доступність даних: сприяє більшій колаборації серед співробітників, забезпечуючи доступ до даних і додатків з будь-якої точки світу. Це особливо актуально в умовах зростаючої популярності віддаленої роботи.

Отже, хмарні технології значно підвищують ефективність роботи з даними, надаючи інструменти для обробки великих обсягів інформації, забезпечуючи надійність та безпеку, а також сприяючи інноваціям і зниженню витрат. Вони стають критично важливим елементом сучасної ІТ-інфраструктури, дозволяючи організаціям більш ефективно досягати своїх бізнес-цілей.

РОЗДІЛ 3. ВИКОРИСТАННЯ МОВИ ПРОГРАМУВАННЯ R ДЛЯ АНАЛІЗУ ВІДКРИТИХ ДАНИХ

3.1. Мова програмування R

В умовах сучасних реалій людині постійно необхідно отримувати інформацію, яка надає їй змогу приймати правильні та зважені рішення. При цьому не важливо будуть ці рішення стосуватися побуту, або планування бізнес процесів. Одним із основних інструментів одержання такої інформації це аналіз даних, які поступають з різноманітних джерел.

Поява Інтернету, хмарних технологій та новітніх методів зберігання інформації призвело до зростання обсягу даних. Для характеристики деяких масивів навіть з'явився такий термін як «великі дані». Даним терміном позначають набори інформації настільки великих розмірів, що до них не можуть бути застосовані традиційні способи обробки. [28] Отримання інформації з таких масивів зараз перетворюється в окрему галузь діяльності.

Збільшенні масивів даних оказало вплив і на сам підхід до аналізу даних. Крім того сучасні методи обробки даних все частіше спираються на графічний спосіб представлення результатів досліджень, тобто на візуалізацію. У зв'язку з цим з'явилося багато засобів для аналізу. Одним з таких засоби це мова програмування R.

Свій початок мова програмування R бере у 1993 року. В цьому році Росс Ихак та Роберт Джентльмен, новозеландські вчені презентували свою нову розробку, яку вони назвали R. За основу було взято мову програмування розвиненої комерційної системи статистичної обробки даних S-PLUS і на її основі створено безплатну, вільну версію. Незабаром виникла розподілена система зберігання й поширення пакетів до R, відома під аббревіатурою «CRAN»(Comprehensive R Archive Network). Основна ідея

організації якої - постійне розширення, колективне тестування й оперативне поширення прикладних засобів обробки даних. [29]

У цілому R - це середовище програмування для статистичного аналізу даних. Система статистичного аналізу та візуалізації в R складається з таких основних частин як: мови програмування високого рівня R, що дозволяє одним рядком реалізовувати різні операції з об'єктами, векторами матрицями, списками; великого набору функцій обробки даних, зібраних в окремі пакети та розвиненою системою підтримки, що включає оновлення компонентів середовища, інтерактивну допомогу й різні освітні ресурси, призначені як для початкового вивчення R, так і для подальших консультацій щодо виникаючих ускладнень.

При роботі з R можна використовувати одразу багато команд, що записані у окремому файлі. Якщо довга команда не вміщається у одному рядочку її можна розбити на декілька рядочків, при цьому при переході до нового рядочка R автоматично виводить символ продовження.

Програми, що складаються з команд R, називають скриптами. Вони мають стандартне розширення г. У базовій програмі є можливість відкрити вікно редактора для створення нового скрипту, або завантажити файл зі скриптом.

За потреби всі технології статистичної обробки можна реалізувати, використовуючи лише базовий пакет R. Але для більш зручного користування R - технологіями можна використовувати спеціальні надбудови-оболонки над R, які надають більш можливостей для програмування, перегляду змінних та користуванню графікою.

Такою оболонкою інтерпретатором є система R - studio. Вона також розповсюджується безкоштовно. Перш ніж інстальовати цю програму, на комп'ютері треба встановити базовий R.

Основне вікно користувацького інтерфейсу R - studio розділене на чотири дочірні вікна. У лівому верхньому вікні виведено script - програму,

яка редагується. Це вікно має багато вкладок, тому дає змогу працювати з декількома файлами одночасно. Нижче розташована консоль, у якій виконуються команди. Тут можна запускати скрипти або їх частини. Праве верхнє вікно надає змогу побачити активні змінні за якими працює програма, а також можна побачити історію роботи. Найбільш навантажене вікно внизу праворуч. Сюди виводяться рисунки, які робить програма, можна подивитися які додаткові пакети було завантажено. За потреби користувач може змінювати ці вікна місцями.

Інструмент R -studio корисний також підказками, які він робить під час набору команд на консолі та у вікні скриптів. Ще одна додаткова його зручність, так це можливість генерації текстових звітів, які виконуються з поєднанням системи програмування R та системи форматування текстів LaTeX. Найбільш схожою за можливостями аналізу макроєкономічних показників мовою програмування є Python. R та Python - це дві популярні мови програмування та середовища для аналізу даних та наукових обчислень. Ось короткий огляд кожного з них:

R:

Призначення: R - це мова програмування та середовище для статистичного аналізу даних, візуалізації та машинного навчання.

Особливості: В R існує велика кількість пакетів для різних статистичних аналізів та візуалізації даних. Це дозволяє швидко та ефективно виконувати різні аналізи безпосередньо у R-середовищі.

Спільнота: R має активну спільноту користувачів та розробників, яка постійно розвиває нові пакети та рішення для аналізу даних.

Використання: R широко використовується в наукових дослідженнях, академічних дослідженнях, фінансах, біоінформатиці та інших галузях для аналізу та візуалізації даних.

Python:

Призначення: Python - це загальне призначення мова програмування, яка також широко використовується для аналізу даних, штучного інтелекту, веб-розробки та інших цілей.

Особливості: Python має багато бібліотек та фреймворків для аналізу даних, таких як Pandas, NumPy, SciPy, Matplotlib, Seaborn, Scikit-learn і багато інших. Це робить Python потужним інструментом для аналізу даних та машинного навчання.

Спільнота: Python має одну з найбільших та найактивніших спільнот у світі програмування. Це означає, що ви знайдете безліч ресурсів, документації, пакетів та допомоги у вирішенні будь-яких проблем.

Використання: Python використовується у багатьох галузях, включаючи науку про дані, інженерію даних, веб-розробку, ігрову розробку, наукові дослідження та інші сфери.

Обидва інструменти мають свої переваги та недоліки, і вибір між ними залежить від вашого контексту, потреб та особистих вподобань.

Мова програмування R є однією з найпопулярніших мов для статистичного аналізу даних, візуалізації та машинного навчання. Ось кілька ключових особливостей мови програмування R:

Спеціалізація на аналізі даних: R була спеціально розроблена для статистичного аналізу даних. Вона має велику кількість вбудованих функцій та пакетів, що полегшують аналіз даних, включаючи стандартні статистичні методи, візуалізацію та машинне навчання.

Безкоштовність та відкритість: R є вільним програмним забезпеченням з відкритим кодом, що означає, що ви можете вільно використовувати його, модифікувати та розповсюджувати. Це сприяє розвитку активної та інноваційної спільноти користувачів та розробників.

Широкий спектр пакетів: R має велику кількість пакетів для різноманітних задач, включаючи статистичний аналіз, візуалізацію даних, машинне навчання, байесівську статистику та багато іншого. Ці пакети

розробляються спільнотою користувачів та дозволяють швидко та ефективно вирішувати різноманітні завдання.

Можливості візуалізації: R має потужні засоби для візуалізації даних, включаючи графіки, діаграми, графи та інші. Ви можете створювати професійно виглядаючі візуалізації з допомогою пакетів, таких як ggplot2 та lattice.

Інтеграція з іншими мовами та інструментами: R може взаємодіяти з іншими мовами програмування, такими як Python, Java, C++ та іншими. Це дозволяє вам використовувати R як частину більшого програмного рішення.

В цілому, мова програмування R є потужним інструментом для аналізу даних та досліджень, особливо у наукових та статистичних дослідженнях, фінансовому аналізі, біоінформатиці та інших галузях, де необхідний високий рівень аналізу даних.

Python зарекомендував себе як першокласний інструмент для наукових обчислювальних завдань. Він може виконувати ті ж функції що і R, однак, він краще підходить для реалізації машинного навчання. Коди Python легше підтримувати і вони є більш надійними ніж R. Більшість завдань з наукових досліджень можливо виконати за допомогою п'яти бібліотек Python:

- NumPy для маніпулювання однорідними масивами даних;
- Pandas для маніпулювання гетерогенними і маркованими даними;
- SciPy для загальних наукових обчислювальних завдань;
- Matplotlib для візуалізації якості публікації;
- IPython для інтерактивного виконання та спільного використання коду;
- Scikit-Learn для машинного навчання.

Проаналізувавши основні характеристики мов програмування R та Python можливо навести різницю між ними у вигляді Таблиці 3.1.

Таблиця 3.1- Порівняльний аналіз R та Python

Параметр	R	Python
Завдання	Аналіз даних та статистика	Розгортання та виробництво
Основні користувачі	Вчені, дослідники та програмісти	Програмісти та розробники
Гнучкість	Проста у використанні доступних бібліотек	Легко побудувати нові моделі з нуля
Крива навчання	Важко на початку	Лінійна
Популярність мови програмування на 2018, (%)	4,23	21,69
Інтеграція	Використовується локально	Добре інтегрується з додатками
Завдання	Легко отримати первинні результати	Добре розгортати алгоритм
Важливі пакети та	tidyverse, ggplot2, caret, zoo	pandas, scipy, scikit-learn,

Параметр	R	Python
бібліотеки		TensorFlow, caret
Недоліки	Повільна крива навчання Залежність від бібліотек	Не така велика кількість бібліотек ніж у R
Переваги	- Швидкий результат - Великий каталог для аналізу даних - Інтерфейс GitHub - RMarkdown - Shiny	- Математичні обчислення - Математичні обчислення - Розгортання - Читання кодів - Швидкість - Функції в Python

Візуалізація даних відіграє важливу роль у кібернетиці, як у моделюванні певних економічних показників та і в галузі Data Science. Для подальшого дослідження моделей та складних даних їх краще візуалізувати, аніж намагатись їх розбирати в електронних таблицях.

У R існує велика кількість бібліотек, що допомагає візуалізовувати економетричні залежності, однак, якісно виділяється пакет Shiny.

Shiny – це відкритий пакет від RStudio, який використовується для побудови інтерактивних веб-сторінок з R. Він забезпечує потужні методи для розробки, розповсюдження та подальшої взаємодії з моделями чи потоками даних. Shiny забезпечує функціональність для інтерактивних веб-

програм. Також він надає нові можливості для взаємодії з досліджуваними економетричними моделями як дослідникам, так і викладачам.

За допомогою пакету Shiny та його додатків можна розміщувати автономні програми на веб-сторінці, вставляти їх у документи R Markdown, створювати інформаційні панелі. Також є можливість вбудовування додатку в електронний підручник зі збереженням інтерактивної складової. Також можна розширити додатки Shiny за допомогою таблиць стилів CSS, HTML-віджетів та скриптів JavaScript.

У Python існує ряд бібліотек, що дозволяють візуалізовувати досліджувані дані, такі як наприклад Matplotlib. Matplotlib - це мультиплатформенна бібліотека візуалізації даних, побудована на масивах NumPy і призначена для роботи з ширшим стеком SciPy. Він був задуманий як патч для IPython для вмикання інтерактивного MATLAB-стилю з допомогою gnuplot з командного рядка IPython, однак, згодом він був розширений та випущений як окремий пакет.

Однак, головною перевагою Python є можливість інтеграції алгоритмів у додатки. Таким чином можливо створити базову модель веб-сайту, де основні обчислювальні функції виконуватиме сервер на базі Python, а дані для візуалізації передаватимуться на сторінку користувача та відображатимуться за допомогою JavaScript та його бібліотек.

Отже, сучасні методи дослідження економетричних залежностей за допомогою мов програмування R та Python надають великий спектр можливостей для аналізу даних.

Використання того чи іншого інструменту залежить від необхідності отримання певного результату. R та велика кількість його бібліотек дозволяє здійснити первинний аналіз досліджуваних даних. Python, загальна мова програмування, яку краще використовувати для реалізації значних алгоритмів та машинного навчання.

В загальні R -studio - це потужний інструмент для аналізу даних. Він має певні переваги перед іншими аналогічними продуктами. Більшість комерційних статистичних програм можливо використовувати тільки на платній основі, тоді як R-studio є безоплатною. Вказане програмне забезпечення має можливість імпортувати дані з різних джерел. Крім того в R вже реалізовані складні статистичні процедури, які можуть бути ще недоступними в інших програмах. R-studio є платформою для простого написання програм, що реалізують нові статистичні методи. Тому це, та означені вище переваги, робить цей інструмент привабливим для використання у різних областях діяльності.

3.2. Аналіз макроекономічних показників

У цьому розділі буде виконано аналіз деяких макроекономічних показників, які були оприлюднені у формі відкритих даних. Необхідна для цього інформація була отримана з офіційного порталу відкритих даних в Україні, а саме з <https://data.gov.ua>

Для проведення аналізу пропонується використати програмну оболонку R- studio. Перед началом обробки даних, для більш коректної роботи програмного забезпечення, необхідно встановити робочу директорію, тобто те місце, з якого R- studio почине свою роботу. Зробити це досить легко. У верхньому меню необхідно зайти пункт Session, у випадіючому списку є значення Set Working Directory, у ньому знаходимо пункт Choose Directory, що дасть змогу обрати необхідне місце. Так як усі отримані дані зберігаються на робочому столі у папці «Data», то вона і була встановлена у якості робочої директорії. [29]

Аналізовані дані розміщені по різних наборах, які збережені у форматі xIs. Для роботи з xls- файлами у R-studio передбачена функція read.xlsx.

Завдяки якій можна отримати дані з файлу xls- файлу, та записати їх у будь-яку змінну. [129]

Необхідно вказати на те, що перед началом аналізу треба вивчити сам набір даних, встановити які дані він у собі зберігає, їх тип, та зрозуміти чи мають вони будь-яке значення для подальшого аналізу.

При вивченні набору даних «Капітальні інвестиції за регіонами» стало ясно, що він містить у собі розділи про Автономну республіку Крим та м. Севастополь, що добре видно з таблиці 3.2.

Починаючи з 2014 року вказані території не фінансуються з бюджету України, таким чином інформація про них вказана як NA (not available). Надалі це може ускладнити процес аналізу, тому доцільно провести «очищення» набору від таких даних, тобто видалити небажану інформацію.

Таблиця 3.2- Частина набору «Капітальні інвестиції за регіонами»

code	attributes	period	data
КОДИ	атрибути	періоди	дані
0000000000	Україна	2010	180575.5
0100000000	Автономна Республіка Крим	2010	NA
0500000000	Вінницька	2010	3544.7
0700000000	Волинська	2010	1767.0
1200000000	Дніпропетровська	2010	16016.9
1400000000	Донецька	2010	14993.9

1800000000	Житомирська	2010	2011.5
2100000000	Закарпатська	2010	2205.4
2300000000	Запорізька	2010	7963.8
2600000000	Івано- Франківська	2010	4378.7
3200000000	Київська	2010	11331.4

Для подальшого аналізу набір даних «Капітальні інвестиції за регіонами» був записаний у змінну `preliminary Data`. З'ясувалось, що він містить у собі два зайві стовпчика, тому вони були видалені за допомогою звичайного «зрізу даних»:

```
install.packages("xlsx")
library("xlsx")
preliminaryData<- read.xlsx("capitalInvest.xlsx",2)
clearedData<-preliminaryData [-1, 1:4]
invesbyRegions<-subset
(clearedData, clearedData$attributes
!= "Автономна Республіка Крим" & clearedData$attributes
!="м. Севастополь" & clearedData$attributes !="Україна" )
```

Після реалізації вказаного коду масив був підготовлений для обробки та аналізу.

Як правило масиви, які піддають аналізу, містять у собі велику кількість даних. Тому часто виникає потреба описати особливості даних

однією, або декількома характеристиками. Техніка такого опису зветься описовою статистикою. При цьому аналіз даних проводиться не на основі, наперед заданої теоретичної моделі, а виходячи із структури самих даних.

Мова програмування R має достатньо методів, завдяки котрим можна отримати показники описової статистики.

Так, наприклад, для розрахунку середньої величини можливо використати функцію `mean()`, а для отримання максимального чи мінімального значення медіани такі функції як `max()` та `min()` відносно. [29]

Використання таких функцій доцільно коли отриманні значення, у подальшому, будуть використовуватись як якості змінної. Для прикладу нижче наведено код, завдяки якому із набору даних можна отримати інформацію про найбільший та найменший розмір інвестиції по регіонах України. Значення, які повертають функції `max()` та `min()`, використовуються у якості умови пошуку інформації, що добре видно у коді:

```
minMaxData<-invesByRegion [invesByRegion$data== max (as.numeric  
(invesByRegion$data) )  
+ 1 invesByRegion$data== min(as.numeric(invesByRegion$data) ), ]
```

Також, іноді, для вказаних цілей використовують функцію `summary`, яка розраховує деякі показники описової статистики, такі як квантілі та медіана. Як можна побачити усі методи або наляють не повну інформацію, або мають складний синтаксис. [29]

R-studio має в своєму розпорядженні пакет для аналізу «psych». Використовуючи даний пакет одразу можливо отримати усю необхідну інформацію про масив даних. Найкраще для цього підійде функція `describe()`.

Вона має дуже простий синтаксис який наведено далі:

```
install.packages ("xlsx")  
library ("psych")
```

describe (as. numeric (invesByRegion\$data))

Функція describe) розраховує та повертає дванадцять статистичних показники, серед яких стандартне відхилення, розмах, медіана, стандартна похибка. [29] Результат використання вказаної функції наведено нижче.

```
n mean sd median mad min max se
275 14385.92 24939.33 7722 6423.07 1686.9 213247.8 1503.9
```

Рис 3.1 - Результат використання функції describe()

У наведеному прикладі видно наскільки зручно використовувати пакет «psych» у зрівнянні іншими способами.

По результатам аналізу набору даних, щодо розміру інвестицій по регіонах, було отримано такі результати, а саме: середній розмір інвестиції по регіонах складає 14385,92 мільйонів гривень, мінімальне розмір інвестованих грошей складає 1686,9 мільйонів гривень, а найбільший 213247,8 мільйонів гривень. При цьому слід зазначити, що медіанне значення складає 7722 мільйонів гривень, що значно менше за середній показник. Це вказує на те, що частіше регіони отримують, у якості інвестиції, суми, які більш приближені до мінімального значення.

Вище було писані можливості вивчення лише одного об'єкту. Але самі по себе вони не існують, а знаходяться у постійному контакті з іншими об'єктами, оказуючи вплив один на одне. Такі явища можливо спостерігати скрізь: в природі, економіці та інші.

В таких випадках поширеною задачею статистичного аналізу постає дослідження зв'язків між різними об'єктами. У процесі таких досліджень встановлюють наскільки сильним є зв'язок між двома явищами та встановлюють тип такого зв'язку.

З економічною точки зору доцільно проаналізувати чи впливає рівень інвестиції на економічний стан регіонів, якщо до як сильно. Для цього використаємо набір відкритих даних «Валовий регіональний продукт», виписка з якого представлена в таблиці 3.3. Як і у попередньому наборі, цей набір містить дані які треба видалити, для спрощення процесу аналізу.

Вказаний набір також містить у собі деякі небажані типи даних та інше.

Таблиця. 3.3 - Частина набору даних «Валовий регіональний продукт»

	2 004	2 005	2 006	2 007	2 008	2 009
УКРАЇНА	3 E+05	4 E+05	5 E+05	7 E+05	9 E+05	9 E+05
АРК	9 901	1 2848	1 6044	2 0874	2 7365	2 7396
Вінницька	8 123	1 0207	1 2414	1 5381	2 0094	2 0104
Волинська	4 994	6 553	7 687	1 0072	1 2784	1 2225
Дніпропетр овська	3 0040	4 1227	5 2347	7 1173	1 04687	9 3331
Донецька	4 5617	5 8044	7 2361	9 2093	1 17646	1 03739
Житомирсь ка	5 947	7 430	8 784	1 1127	1 5008	1 4731

Закарпатськ	5	6	8	1	1	1
a	297	700	185	0508	3208	2542
Запорізька	1	1	2	3	4	3
	5255	9968	4787	3158	2445	7446
Івано- Франківська	7	9	1	1	1	1
	311	622	1316	3916	7883	7241
Київська	1	1	1	2	3	3
	1883	5362	9188	6221	5687	7548

У попередній главі було розказано, що для оцінки міцності зав'язків використовують коефіцієнт кореляції. Найпоширеніший у використанні це коефіцієнт кореляції Пірсона, який розраховується як значення поділу коваріації на дисперсію.

Для розрахунку такого коефіцієнту, мова програмування R має у своєму розпорядженні таку функцію як `cor()`. Синтаксис функції передбачає три аргументу, перші два - це вектори, між аргументами котрих буде проводитись розрахунок тісноти зв'язку. Обов'язковою умовою є те, що вони повинні бути представлені таким типом даних як `numeric`. Третій аргумент вказує, який саме коефіцієнт необхідно розрахувати. У функції `cor()` передбачено їх три : «pearson», «kendall» та «кальєносець»(непараметрична кореляція). 29]

У цій роботі було проведено дослідження того як впливають капітальні інвестиції на рівень виробництва валового регіонального продукту. Для аналізу було взято два набори даних які містять у собі необхідну інформацію по Україні за період з 2010 по 2020 рік. Код розрахунку наведено нижче:

```
cor (totalInvest, productsTotal, method = "pearson" )
```

Змінні total Invest та products Total містять у собі сумарні значення по капітальним інвестиціям та виробництву продукції по кожному року. У результаті розрахунку було отримано значення коефіцієнту кореляції Пірсона на рівні 0,95, Це дуже велике значення, яке вказує на те що розмір інвестицій дуже сильно впливає на обсяг виробництва.

Одним із важливіших завдань аналізу - це можливість передбачити поведінку системи у майбутньому. Для цього використовують регресійний аналіз. Тому наступним шагом у вивченні зв'язку між виробництвом та інвестиціями стало побудова регресії. Для цього була використана така функція як lm. Синтаксис функції дуже простий, вона приймає два значення.

Знаком тильда вказується яка змінна буде вважатись залежною. Нижче наведено повний код використання функції lm():

```
relation<-lm (productsTotal~totalInvest)
summary (relation)
```

В результаті використання функції lm) були отримані коефіцієнти рівняння лінійної регресії. Так коефіцієнт регресії дорівнює 7,11, а незалежна змінна прийняла негативне значення у розмірі - 205288,33.

Заключним етапом аналізу є перевірка якості побудованої лінійної регресії. Для оцінки статистичної значимості коефіцієнтів регресії та коефіцієнту кореляції розраховується - критерій Стюдента. [29] Перевірка проводиться шляхом порівняння критичного значення з розрахунковим.

Критичне значення - критерія Стюдента можна знайти у таблиці, або розрахувати у електронній таблиці XL. Якщо фактичне значення критерію по модулю більше за критичне то вважають, що між двома параметрами є тісний лінійний зв'язок. Для оцінки якості моделі у цілому, використовують F-критерій Фішера.[29] Принцип перевірки також є порівняння критичного та розрахункового значення критерію.

В арсеналі R - studio є така функція як `summary`), яка може розрахувати усі необхідні коефіцієнти одночасно.

Передавши усі необхідні дані у функцію ми отримали дані для перевірки статистичної значасті рівняння, усі дані можна побачити на рисунку 3.3.

```
summary (relation)
```

```
##
```

```
## Call:
```

```
## lm (formula = summaryData$Production ~ summaryData$Investments)
```

```
##
```

```
## Residuals:
```

```
## Min 1Q Median 3Q Max
```

```
## -348922 -254705 -36173 134494 814096
```

```
##
```

```
## Coefficients:
```

```
##
```

```
##              Estimate Std. Error t value Pr (>|t|
```

```
## (Intercept) -2.053e+05 2.842e+05 -0.722
```

```
8
```

```
## summaryData$ Investments 7.110e+00 7.308e-01 9.728 4.5-0
```

```
6 ***
```

```
##---
```

```
## Signif. codes: 0 * 0.001 •*** 0.01 •** 0.05 ' ' 0.1 ' '
```

```
1
```

```
##
```

```
## Residual standard error: 358400 on 9 degrees of freedom
```

```
## Multiple R-squared: 0.9132, Adjusted R-squared: 0.9035
```

```
## F-statistic: 94.64 on 1 and 9 DE, p-value: 4.498e-06
```

Рисунок 3.3 - Реалізація функції summary

У результаті розрахунків - критерій Стюдента набув значення 9,72. Критичне значення цього показнику, для відповідного ступеня свободи, дорівнює 2,26. Що стосується критерію Фішера, його розрахункове значення дорівнює 94,64, при критичному значенні 2,98. Таким чином обидва критерію перевищують відповідні критичні значення. Це свідчить про існування істотної лінійної залежності між об'ємом виробництва та рівнем інвестицій. Крім того вказує на статистичну значимість регресійної моделі.

3.3. Візуалізація даних

У сучасному аналізі графічне надання даних має велике значення. Подання даних у візуальному образі спрощує їх сприйняття. Візуалізація покликана донести до користувача те, що він не має може побачити у наборі цифр, та інших об'єктів. Завдяки такому методу вже на початковому етапі, не вдаючись до статистичних розрахунків, виявити певні тенденції у великих масивах даних.

У широкому сенсі візуалізація - це графічна презентація інформації, завдяки якій можна лаконічно представити у вигляді зображенні ті об'єкти, які у других формах не будуть мати достатньо інформативний характер, або для опису яких знадобиться інформація великих розмірів.

Існує багато способів візуалізації, дані можна представити за допомогою різноманітних форм графіків, діаграми та інше. Той чи інший спосіб мають як свої переваги, так і недоліки. Так, якщо потрібно показати структуру, то доцільно використовувати кругову діаграму. Але така візуалізація буде недоцільна для презентації великої кількості змінних. Загальні рекомендації вказують, що для візуалізації набору даних, якій містить у собі більш ніж десять змінних, краще підійде горизонтальна

стовпчикова діаграма. Графік найкращим чином підходить для порівняння, або відображення будь-якої тенденції.

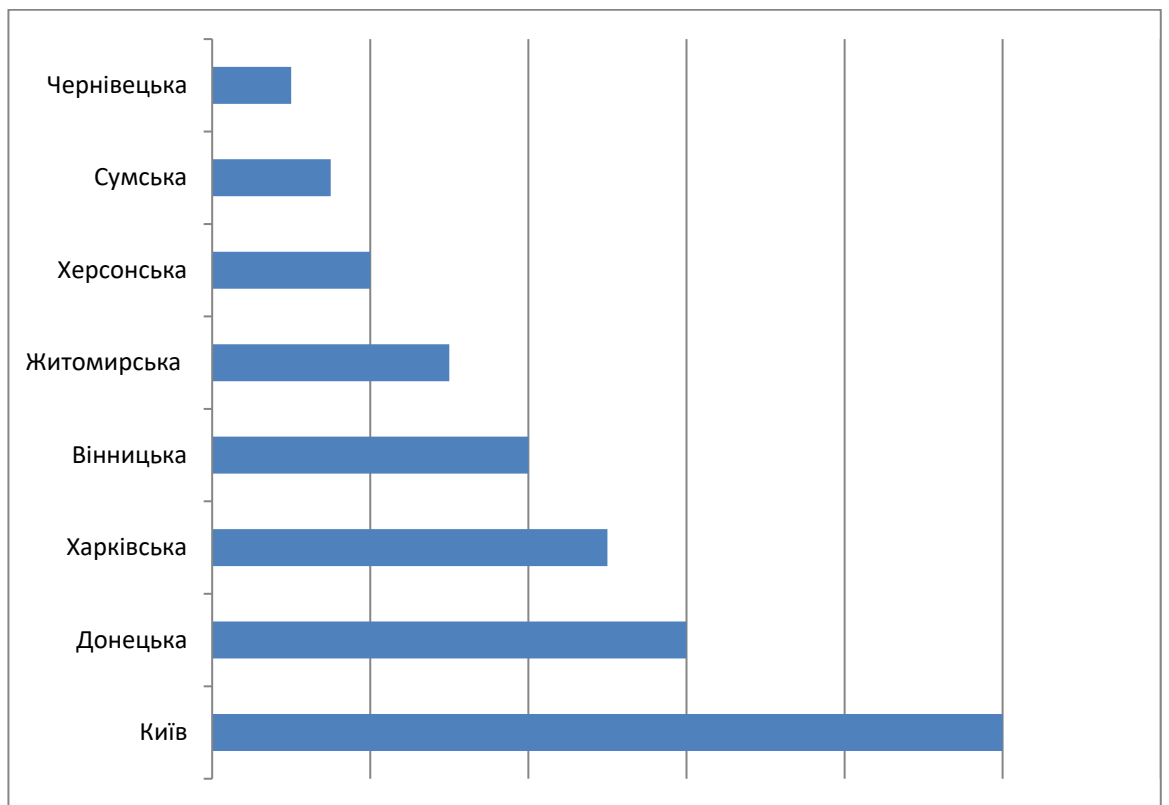
Візуалізувати дані можливо за допомогою мови програмування R. Вказаний інструмент має у своєму арсеналі достатню кількість інструментів, використовуючи які, дані можливо представити в тому, чи іншому образі.

Так, у якості прикладу, можливо вказати таку функцію як `plotbar`). Вказана функція використовується для побудови стовпчикової діаграми. Вона приймає у себе, у якості змінних, базу даних, на основі якої буде проведено побудова діаграми. Додаткові параметри вказують на просторову орієнтацію діаграми, розміщення назви діаграми, колір та розмір стовбців.[29]

У цієї роботи була проведена візуалізації інформації стосовно розподілу інвестиції по регіонах України. Кількість змінних, які необхідно відобразити, дорівнює двадцяти п'яти. Тому, наприклад, використовуючи кругову діаграму, можна отримати об'єкт, який важко буде використати для візуальної оцінки інформації. Для візуалізації цих даних була використана горизонтальна стовпчикова діаграма, яка була побудована з використання функції `plotbar`). Частина коду програми наведена нижче.

```
barplot (summInvest, width = (50, 139), horiz = TRUE, las=1,  
+ col ="lightblue" )
```

Як можна побачити функція має дуже простий синтаксис. Змінні `summInvest`- це база даних, `width` - ширина стовпчиків, `horiz`- розташування по горизонтальній площині, `las`- розташуванні підписів.[29]



Діаграма 3.4 - Розподіл інвестицій за регіонами

Як було зазначено вище, візуалізація дає змогу отримати певну інформацію не використовуючи математичні методи. З огляду на діаграму одразу помітно, що найбільш інвестований регіон це місто Київ, найменше Чернівецька область.

Незважаючи на простоту використання окремих функцій графічного аналізу, останні роки найбільшою популярністю користується пакет «ggplot2».

Цей пакет призначений для будування двомірних графіків та діаграм. Графік, або будь-яка діаграма, у цьому пакеті, будуються пошарово накладаючи один шар на інший.

Однією із базових функцій пакету є функція `ggplot`, завдяки якій створюється сама візуалізація, тобто створює площину, на якій буде побудовано графік.

Друга базова функція це функція `geom`. Якщо завдяки `ggplot` створюється поверхня, то `geom` — це функція нанесення того чи іншого виду графіки. Вибір конкретного об'єкту відбувається за таким синтаксисом `geom`. Вказана функція містить у собі достатньо велику кількість об'єктів, від звичайної точки до складних геометричних фігур. [29]

Пакет `ggplot` має ще додаткові функції, завдяки яким можливо покращити візуалізацію, та надати їй додаткових якостей, вони додаються до основного пакету за допомогою знака плюс, як показано нижче:

```
ggplot (data=investProgress, mapping = aes (x=Pik)) +  
geom line (aes (y=Інвестиції), color="red" )
```

У якості приклада можна навести таку функцію як `color()`, яка відповідає за зміну кольору графіку чи діаграми, а також `ggtitle()` - функція зміну підписів. [29]

Деякі можливості пакету `ggplot2` були використані у цієї роботі при візуалізації таких показників як розмір інвестованих коштів в період з 2010 по 2020 роки. Для цього найбільшим чином підійде графік, так як сенс візуалізації цього процесу - вивчення динаміки розмірів інвестицій за певний період.

Нижче наведено частина коду, яка відповідає за побудову графіку.

```
2020.0
```

Аналізуючи вищевказаний графік, можна одразу помітити, що за період з 2010 по 2020 року рівень інвестицій у регіони суттєво виріс. При цьому звертає на себе увагу той факт, що до 2014 року динаміка росту не

мала такої ярко вираженої тенденції, у період з 2011 по 2013 взагалі спостерігався спад.

Лінійний графік зручно використовувати не тільки при візуалізації динамічних процесів, а також для отримання поняття про взаємозв'язок між двома явищами.

Так нижче, на графіку показана залежність обсягу виробництва від розміру інвестиції. Одразу помітно, що усі значення досить близько групуються вздовж уявної лінії. Цей факт може вказувати не тільки на лінійний характер зв'язку між вказаними показниками, але і на його міцність.

ВИСНОВКИ

У цій роботі було проведено дослідження можливості отримання макроекономічних показників з наборів відкритих даних, для їх подальшого аналізу. За для досягнення мети було проведено аналіз поточного законодавства у сфері інформації, задіяні ресурси офіційного порталу відкритих даних в Україні, а саме: data.gov.ua.

Вказаний портал містить у собі велику кількість наборів даних з різних сфер. Інформацію, яка стосується макроекономічних показників, можна отримати з наборів даних, розпорядником яких є Державна служба статистики України. Всі ці набори оприлюднені з дотриманням вимог

Положення «Про набори даних, які підлягають оприлюдненню у формі відкритих даних». Вони можуть бути легко завантажені на персональний комп'ютер у таких форматах як XL та CSV.

Формати, у яких оприлюднені дані, дозволяють проводити обробку наборів будь-якими засобами, такими як XL, або за використанням мови програмування python. Обидва вказаних засоби мають як переваги, так і недоліки. Наприклад, використання формат XL обмежує користувача у візуалізації, а застосування мови програмування python вимагає наявності специфічних знань.

Хороші результати, у ході виконання аналізу, показав язик програмування R. Використовуючи середовище розробки R-studio був

проведений статистичний аналіз деяких макроекономічних показників, а результати візуалізовані. Переваги використання R-studio порівняно з іншими засобами полягають у легкості освоєння, широкого спектру інструментів аналізу та візуалізації, наявності у відкритому доступі безкоштовної версії продукту.

По результатам роботи видно, що необхідно продовжити досліджені теми відкрити даних. Для проведення аналізу макроекономічних показників можна рекомендувати використання R-studio.

ПЕРЕЛІК ПОСИЛАНЬ

1. Бідзіля Ю., Кравець Д. Розвиток відкритих даних як нові перспективи й можливості для сучасної української журналістики //DOI. - 2019. - Т. 10. - С. 2524-0331.
2. Ухаль М. М. Латинсько-український та українсько-латинський словник юридичної фразеології. - 2008.
3. Ашанін В. С. Стан і перспективи інформатизації освіти у вузах фізичної культури //Педагогіка, психологія та медико-біологічні проблеми фізичного виховання і спорту.-Харків. - 2000. - №. 19. - С. 20-26.
4. Пожуєв В. І. Осмислення місця і ролі інформації у сучасному суспільстві /Гуманітарний вісник Запорізької державної інженерної академії. - 2010. - №. 42. - С. 4-13.
5. Про інформацію: Закон України від 02.10.1992р. №2658-XII. Поточна редакція від 13.12.2022р. №2849-IX : <http://zakon.rada.gov.ua>.
6. Про доступ до публічної інформації: Закон України від від 13.01.2011р. №2658-XII. Поточна редакція від 01.12.2022р. №2811-IX : <http://zakon.rada.gov.ua>.
7. Семко М.І. Підготовка фахівця з публічного управління в умовах магістратури: Практика та досвід.- 2020.
8. Харенко О. В. Поняття інформація в юридичній науці та законодавстві України //Часопис Київського університету права. - 2014. - №. 3. - С.119-124.
9. Лоленко О. Відкриті дані як інструмент народовладдя// Рада Почесних Членів Honoris Causa- 2019. - С. 241.
10. Корж І. Ф. Право на відкриті дані-як право приватного характеру //Інформація і право, - 2019. - №. 1. - С. 19-28.
11. Про затвердження Положення про набори даних, які підлягають оприлюдненню у формі відкритих даних»: Постанова Кабінету міністрів

України від 21.10.2015 р. № 835. Поточна редакція від 01.01.2024р.
: <http://zakon.rada.gov.ua>.

12. Тарасюк А. В. Відкриті дані та інші дані у публічному доступ: правові аспекти /Інформація і право. - 2017. - №. 2 (21). - С. 59-65.

13. Твердохліб О. С. Методологія та практика реалізації концепту відкритих даних в органах публічного управління України /Публічне адміністрування: теорія та практика. - 2017. - №. 2.

14. Цивільний кодекс: Закон України від 16.01.2003р. №435-IV /10фіц. вісн. України. Поточна редакція від 30.01.2024р. №3450-IX - 2003. : <http://zakon.rada.gov.ua>.

15. Руденко В. Д. Кодування текстових даних //Комп'ютер у школі та сім'ї. - 2014. - №. 7. - С. 30-34.

16. Кавин Я. М., Кавин С. Я., Кавин Б. Я. Методи представлення цифрових зображень в інтерактивних програмних середовищах //Editorial Board. - 2021. - С. 263.

17. Демяненко А. Є. Методи і моделі первинної обробки інформації з елементами самоподібності. - 2020.

18. Золотько К. Простий текстовий формат реляційної бази даних для розробки, опису та обміну даними через мережу /Інформаційні технології та суспільство. - 2022. - №. 4 (6). - С. 34-40.

19. Андрущенко Р. Б., Зайцев С. В., Солдатов А. Ю. Аналіз методів серіалізації структурованих даних для передачі в протоколах прикладного рівня моделі OSI //Математичне моделювання в економіці. -2018.

20.Роїк М. В., Присяжнюк О. І., Денисюк В. О. Огляд програмних засобів статистичного аналізу даних //Ефективна економіка.-2017,-№ 7.-

21. Режим доступу до журналу: <http://www.economy.nauka.com.ua>. - 2017.

22. Герасименко С. С., Головач А. В., Сріна А. М. Статистика: підручник //К.: КНЕУ. -2000. - Т. 467. - С. 10.

23. Батурінич А.В., Антоненко С.И. Визначення схожих рядів даних з використанням коефіцієнтів кореляції //System technologies. - 2021. - Т. 5. - №. 136. - С. 98-109.

24. Пріщенко О., Черемська Н., Черногор Т. Побудова математичних моделей за допомогою методів кореляційного і регресійного аналізу //Вісник Національного технічного університету «ХП». Серія: Інноваційні дослідження у наукових роботах студентів. - 2021. - №. 2. - С. 29-36.

25. Бідніченко О. Г. Дослідження етапів розвитку та особливостей хмарних технологій -2023.

26. Чуб О. І., Новожилова М. В. Створення хмарного середовища вивчення математичного інструментарію Data Science /Амідж сучасного педагога. - 2023. - №. 6 (213). - С. 18-24.

27. Гаврилюк О. Д. Порівняння наявних хмаро орієнтованих технологій навчання для підготовки бакалаврів статистики //Наукові записки Центральноукраїнського державного педагогічного університету імені Володимира Винниченка]. Серія: Педагогічні науки. - 2019. - №. 177 (1). – С.104-107.

28. Мосіюк О. О. Огляд хмарних технологій систем комп'ютерної алгебри. - 2018.

29. Дзямулич М. І., Шматковська І. О., Борисюк О. В. Великі дані та їх роль у формуванні цифрової економіки //Галицький економічний вісник. - 2021. - Т. 70. - №. 3. - С. 16-21.

ДОДАТКИ

Додаток А. Лістинг команд язику R (аналіз)

```
library ("xlsx")
preliminaryData<- read.xlsx ("capitalInvest.xlsx", 2)
clearedData<-preliminaryData[1, 1:41]
str (preliminaryData)
## 'data. frame' : 309 obs. of 6 variables:
## $ code : chr "коди" "00000000000" "01000000000" "05000000
000" ...
## $ attributes: chr "атрибути" "Україна" "Автономна
Республіка Крим" "Вінницька" ...
## $ period : chr "період" "2010" "2010" "2010" ...
## $ data : chr "дані" "180575.5" "NA" "3544.7" ...
## $ NA. : logi NA NA NA NA NA NA NA ...
## $ NA.. 1: logi NA NA NA NA NA NA NA ...
investbyRegions<-subset (clearedData, clearedData$attributes
!="Автономна Республіка Крим" & clearedData$attributes !="™.
Сева стополь" & clearedData$attributes !"Україна" )
str (investbyRegions)

## 'data. frame': 275 obs. Of 4 variables:
## $ code : chr "05000000000" "07000000000" "12000000000" "1
4000000000"
## $ attributes: chr "Вінницька" "Волинська"
"Дніпропетровська" "Донецька"...
## $ period: chr "2010" "2010" "2010" "2010" ...
## $ data: chr "3544.7" "1767.0" "16016.9" "14993. 9" ...

library ("psych")
describe (as. numeric (investbyRegions$data) )
## vars n mean n max sd median trimmed mad mi range
## x1 1 275 14385.92 24939.33 7722 9121.35 6423.07 1686. 9
2 13247.8 211560.9
```

```

## skew kurtosis se
## X1 5.09 31.19 1503.9

initialDataProduct<-read.xlsx ("produktRegin.xls", 1)
dataProductFilter<-initialDataProduct [4:33, 1:18]
dataProductClear<-subset (dataProductFilter,
dataProductFilter$NA
    • !="Україна" & dataProductFilter$NA. !="Крим" &
dataProductFilte
    r$NA. !="області" & dataProductFilter$NA. !="М.
Севастополь") str (dataProductClear)
## 'data. frame': 25 obs. of 18 variables:
## $ NA. :chr "Вінницька" "Волинська" "Дніпропетровська"
"Донецька" ...
## $ NA..1 : chr "8123" "4994" "30040" "45617" ...
## $ NA..2 : num 10207 6553 41227 58044 7430 ...
## $ NA..3 : num 12414 7687 52347 72361 8784 ...
## $ NA..4 : num 15381 10072 71173 92093 11127 ...
## $ NA..5 : num 20094 12784 104687 117646 15008 ...
## $ NA..6 : num 20104 12225 93331 103739 14731 ...
## $ NA..7 : num 23589 14429 116136 128986 18743 ...
## $ NA..8 : num 29099 17637 140020 161021 21928 ...
## $ NA.. 9 : num 33024 20005 147970 170775 24849 ...
## $ NA.. 10: num 36191 20622 152905 164926 25676 ...
## $ NA.. 11: chr "43990" "24195" "176540" "119983" ...
## $ NA.. 12: chr "59871" "31688" "215206" "115012" ...
## $ NA.. 13: chr "74411" "35744" "244478" "137500" ...
## $ NA.. 14: chr "92288" "51919" "313527" "165932" ...
## $ NA.. 15: chr "111489" "60445" "369356" "192161" ...
## $ NA.. 16: chr "129097" "75637" "390325" "204893" ...
## $ NA.. 17: chr "135867" "77404" "398732" "206309" ...
dataProduct<-subset(dataProductClear, select = - c (2:7))
dataProduct$NA.. 11<-as.numeric (dataProduct$NA.. 11)

```

```

dataProduct$NA.. 12<-as .numeric (dataProduct$NA.. 12)

dataProduct$NA.. 13<-as .numeric (dataProduct$NA.. 13)

dataProduct$NA.. 14<-as.numeric(dataProduct$NA.. 14)

dataProduct$NA.. 15<-as .numeric (dataProduct$NA.. 15)

dataProduct$NA.. 16<-as .numeric (dataProduct$NA.. 16)

dataProduct$NA..17<-as .numeric(dataProduct$NA..17)

str (dataProduct)
## 'data. frame': 25 obs. of 12 variables:
## $ NA. : chr "Вінницька" "Волинська" "Дніпропетровська"
"Донецька" ...
## $ NA..7 : num 23589 14429 116136 128986 18743 ...
## $ NA.. 8 : num 29099 17637 140020 161021 21928 ...
## $ NA..9 : num 33024 20005 147970 170775 24849 ...
## $ NA.. 10 : num 36191 20622 152905 164926 25676 ...
## $ NA..11 : num 43990 24195 176540 119983 29815 ...
## $ NA.. 12 : num 59871 31688 215206 115012 38425 ...
## $ NA..13 : num 74411 35744 244478 137500 47919 ...
## $ NA.. 14 : num 92288 51919 313527 165932 61424 ...
## $ NA..15 : num 111489 60445 369356 192161 77103

## $ NA.. 16 : num 129097 75637 390325 204893 85267 ...
## $ NA.. 17 : num 135867 77404 398732 206309 91365 ...

productsPerYears<-colSums (dataProduct I, c (2, 3, 4, 5, 6,
7, 8, 9, 10,1
1,12)], na. rm = TRUE)
productsPerYears

```

```

## NA..7 NA.. 8 NA.. 9 NA.. 10 NA..11 NA..12 NA.. 13 N A..14
NA..15 NA.. 16 NA. .17
## 1042358 1254500 1404669 1465198 1586915 1988544 2385367
29812
27 3560302 3977198
## NA. .17
##4222026
invest<tapply(as.numericinvestbyRegionsSdata),as.numeric(inv
estby RegionsSperiod),
sum)
invest
## 1 2010 1042358 180575.5
## 2 2011 1254500 241285.8
## 3 2012 1404669 578726.4
## 4 2013 1586915 623978.9
## 5 2014 1988544 508216.9
## 6 2015 2385367 273256.1
## 7 2016 2981227 219419.9
## 8 2017 3560302 273116.4
## 9 2018 2385367 273256.1
## 10 2019 2981227 219419.9
##11 2020 3560302 273116.4

```

Додаток Б. Лістинг команд мови R (візуалізація)

```
library ("xlsx")
preliminaryData<- read. xlsx ("capitalInvest.xlsx", 4)
clearedData<-preliminaryData [-1, 1:41
investbyRegions<-subset (clearedData, clearedData$attributes
!=" Autonomous Republic of Crimea" & clearedData$attributes
!="City of Sevastopol" & clearedData$attributes !="Ukraine" )
regionsName<-as. vector (unique (investbyRegions$attributes) )
summInvest<-tapply (as .numeric (investbyRegions$data) ,
investbyRegions$attributes, sum)
reginInvest<-data. frame (Region=regionName,
Totalinvestment=summ Invest)
reginInvest$Totalinvestment<-as.numeric
(reginInvest$Totalinvestment)
reginInvest$Totalinvestment
## [1] 67217.0 54377.1 29269.6 1249706.9 390127.3 236232.6
77738.4
## [8] 167011.3 54828.8 76558.156494.0 305755.3 5678 8.2
191285.3
## [15] 81247.6 2.0 100605.8
## [22] 64380.8 23 170253.2 148987.8 47308.6 52657.2 5435
49254.9 116535.4 57153.5
barplot (summInvest, width= c(50, 140) ,horiz= TRUE,
+las=1, col="lightblue")
library ("ggplot2")
ggplot (data=investProgress, mapping = aes (x=Pi) )
+geom_line (aes (y=Інвестиції), color="red" )
ggplot (data=summaryDatas, mapping = aes (x=Investments) )
+geom_line (aes (y=Production) , color="red")
```

Короткий звіт подібності



Ім'я користувача:
адміністратор НБ НУ ОЮА Librarian

ID перевірки:
1016303672

Дата перевірки:
31.05.2024 14:19:57 EEST

Тип перевірки:
Doc vs Internet

Дата звіту:
31.05.2024 14:50:29 EEST

ID користувача:
84377

Назва документа: Диплом Бобер 2305

Кількість сторінок: 82 Кількість слів: 16347 Кількість символів: 126106 Розмір файлу: 139.49 KB ID файлу: 10160996

15.7% Схожість

Найбільша схожість: 3.39% з Інтернет-джерелом (<https://ela.kpi.ua/handle/123456789/23186>)

15.7% Джерела з Інтернету

657

Сторінка 84

Пошук збігів з Бібліотекою не проводився

2.47% Цитат

Цитати

14

Сторінка 85

Посилання

1

Сторінка 86

0.02% Вилучень

Деякі джерела вилучено автоматично (фільтри вилучення: кількість знайдених слів є меншою за 8 слів та 0%)

0.02% Вилучення з Інтернету

43

Сторінка 87

Немає вилучених бібліотечних джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи

5