ІНФОРМАЦІЙНЕ ЗАБЕЗПЕЧЕННЯ СТАТИСТИЧНИХ ДОСЛІДЖЕНЬ. ВСТУП

Статистика
Поняття інформаційного забезпечення

Статистика - це сукупність методів збору, обробки та аналізу даних для виявлення закономірностей

Статистика - це спосіб перетворити масив даних на конкретні факти, на основі яких можна приймати рішення

У контексті статистики та ІТ, **інформаційне забезпечення (ІЗ)** - це сукупність єдиної системи класифікації та кодування інформації, уніфікованих систем документації та потоків інформації, які необхідні для проведення аналізу та прийняття рішень

Складові ІЗ

1. Методологічна складова

– Системи показників. Що саме ми рахуємо? (наприклад, не просто «трафік», а «кількість унікальних сесій»)

– Методики розрахунку. За якими формулами обробляються дані?

– Класифікатори. Як групуються дані? (наприклад, коди країн, категорії товарів, типи помилок у коді)

2. Ресурсна складова (наповнення системи)

– Бази та сховища даних. Де знаходиться інформація

– Джерела. Логи (log files) серверів, результати парсингу, масиви з державних реєстрів

– Документообіг. Вхідні та вихідні форми звітів

3. Технологічна складова (інструментарій). Обладнання та програмне забезпечення:

– **ПЗ для збору.** Скрипти на Python, системи анкетування (Google Forms), лічильники (Google Analytics)

– **ПЗ для обробки.** SQL, Excel, SPSS, MathCAD, мови R/Python

– **Засоби передачі.** Мережеві протоколи, API, хмарні сервіси

Життєвий цикл дослідження

Статистичне дослідження - це послідовний процес перетворення ідеї в обґрунтований звіт. Кожен етап є критично важливим для отримання достовірного результату.

1. Планування. Визначення мети та об’єкта дослідження. *(Наприклад: «Аналіз причин нестабільної роботи мобільного застосунку»).*

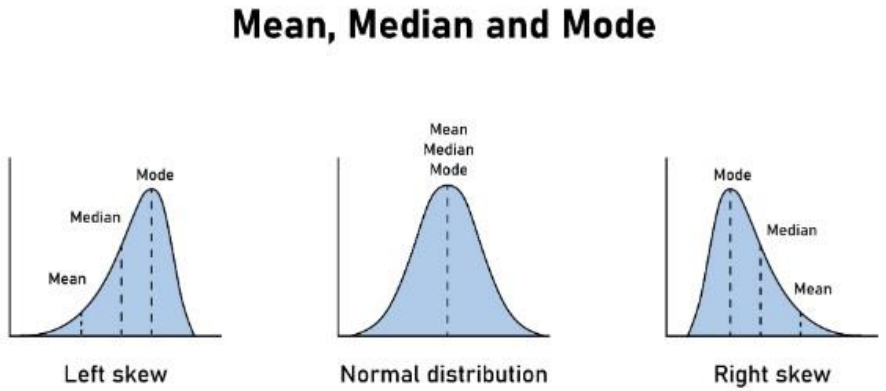
2. Збір даних. Отримання первинної інформації з обраних джерел. *(Приклад: Вивантаження серверних логів за певний період).*

3. Підготовка (очищення): Фільтрація даних, усунення помилок та зайвих записів, що не стосуються мети дослідження. *(Приклад: Відбір лише тих записів, що містять коди помилок, та виключення успішних сесій).*

4. Статистичний аналіз: Обробка даних для виявлення закономірностей. *(Приклад: Встановлення зв’язку між кількістю збоїв та версією операційної системи).*

5. Формування висновків: Інтерпретація результатів та підготовка рекомендацій для прийняття рішень. *(Приклад: Надання звіту розробникам з аргументацією щодо необхідності виправлення вразливостей конкретної версії ОС)*

Методологічний апарат обробки даних



1. Описова статистика

Мета описової статистики - представити великий масив “сирих” даних у вигляді кількох інформативних показників

Міри центральної тенденції:

Середнє арифметичне. Базовий показник, що відображає загальний рівень значень

Медіана. Значення, що знаходиться в центрі впорядкованого ряду. На відміну від середнього, вона краще відображає реальний стан справ, якщо в даних є одиничні критичні відхилення (наприклад, поодинокі затримки сервера на фоні стабільної роботи)

Мода. Найбільш частотне значення в наборі даних

Методологічний апарат обробки даних

Показники варіації:

Дисперсія та стандартне відхилення. Показують рівень стабільності даних. Чим вище відхилення, тим менш передбачуваною є система

Стандартне відхилення показує, наскільки окремі результати відрізняються від середнього значення

1. Низьке відхилення

Нехай середня швидкість відповіді сервера - **100 мс**. При низькому відхиленні майже всі запити обробляються за 95 -105 мс.

Передбачуваність. Точно відомо, що користувач отримає відповідь вчасно

2. Високе відхилення

Середня швидкість залишається тією ж - **100 мс**. Але один запит обробляється за 10 мс, а інший - за 500 мс.

Непередбачуваність: Хоча «в середньому» все добре, не можна гарантувати якість сервісу. Для одного користувача сайт «літає», а для іншого - «зависає».

Методологічний апарат обробки даних

2. Статистичне виведення

У більшості випадків отримати дані від кожного користувача неможливо, тому аналіз проводиться на основі вибірки

Перевірка гіпотез. Математичний підхід до підтвердження або спростування припущень.

Приклад: Чи дійсно оновлення алгоритму пришвидшило обробку запитів, чи зміна показників є випадковою?

Рівень значущості (α): Показник, що допомагає відокремити реальну закономірність від статистичної похибки. Традиційно значення $\alpha < 0.05$ вважається достатнім для визнання результату значущим

Методологічний апарат обробки даних

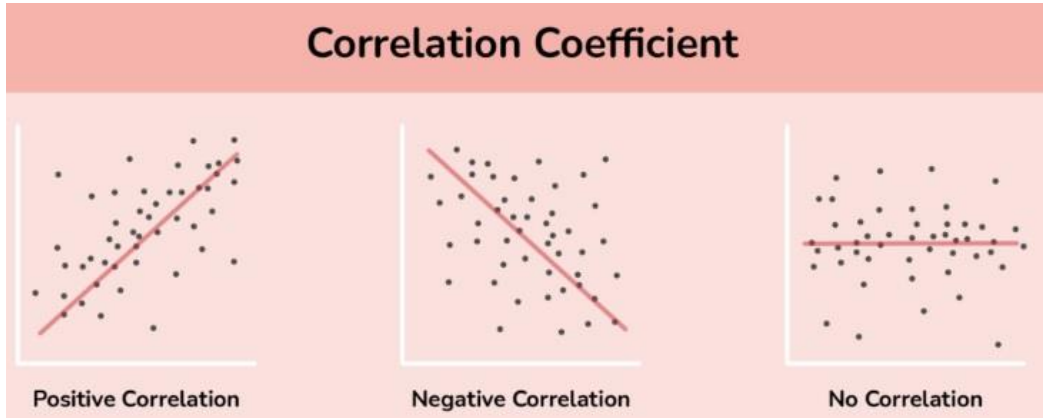
3. Аналіз взаємозв’язків та прогнозування

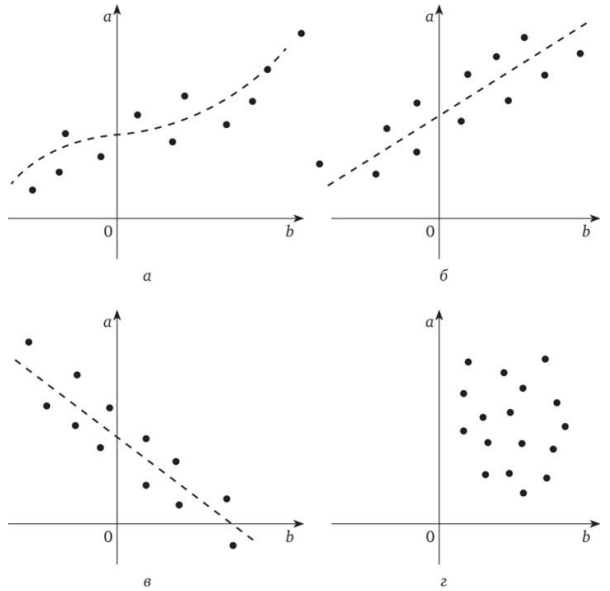
Статистика поєднує описовий аналіз минулих подій та прогнозування майбутніх показників

Кореляційний аналіз. Визначення наявності та сили зв’язку між двома змінними. Кореляція вказує на взаємозв'язок, але не обов'язково на причинно-наслідковий зв'язок.

Регресійний аналіз. Побудова математичної моделі, яка дозволяє прогнозувати одну величину на основі іншої.

Приклад: Прогнозування навантаження на мережу залежно від кількості активних сесій





Програмне забезпечення для статистичного аналізу

Вибір інструменту залежить від обсягу даних, необхідної глибини аналізу та автоматизації процесів

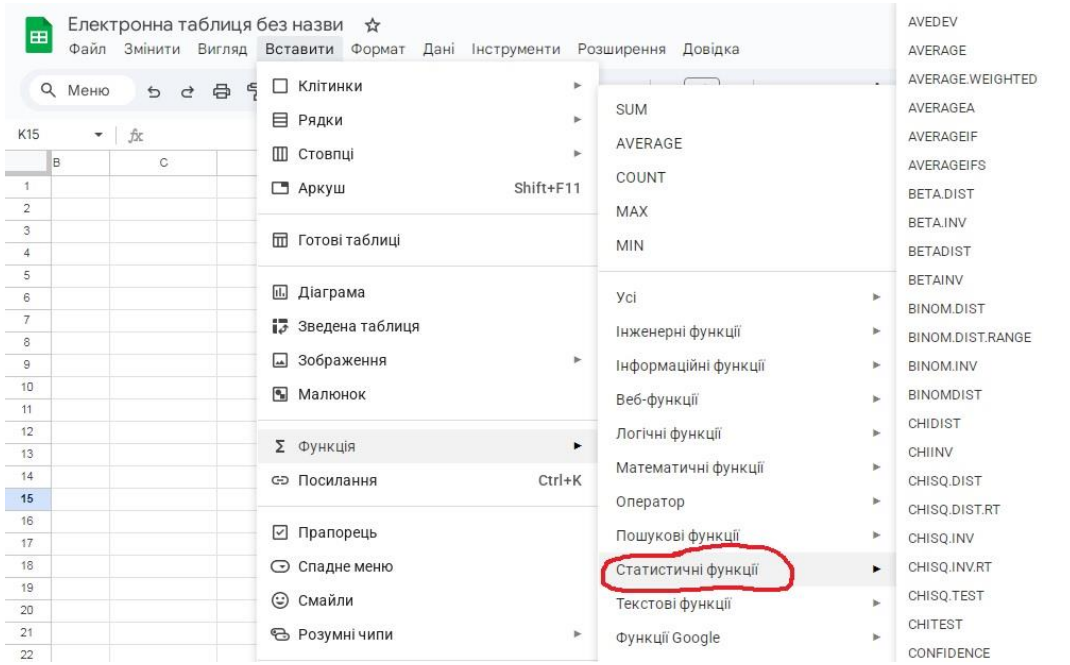
Електронні таблиці (Microsoft Excel, Google Sheets)

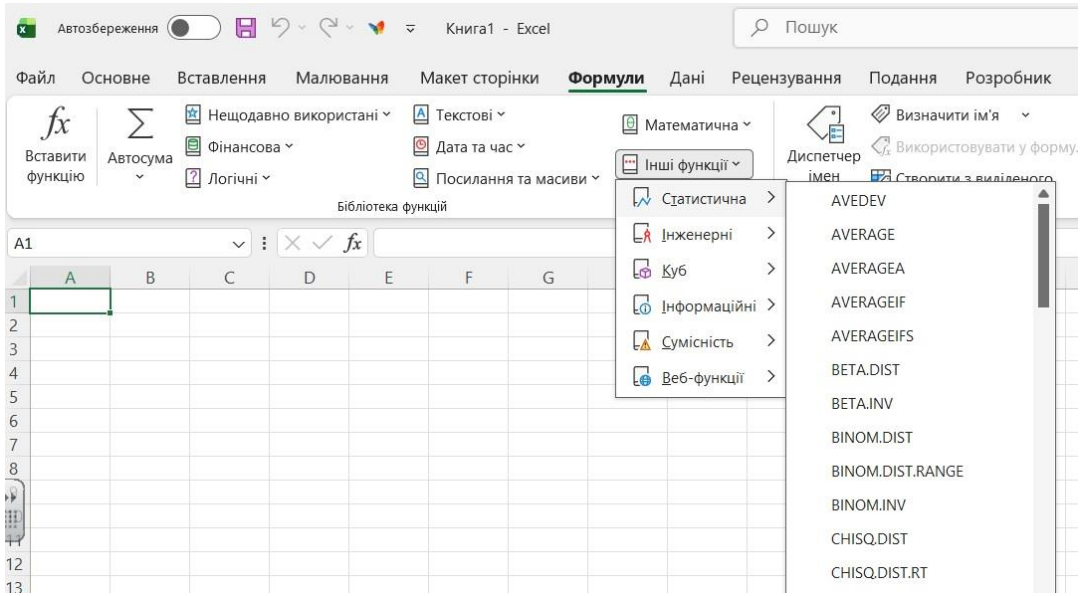
Це базовий інструментарій для швидкого аналізу невеликих масивів даних.

Переваги. Візуалізація «на льоту», доступність, вбудовані фінансові та статистичні функції

Сфера застосування. Оперативна звітність, прості розрахунки, побудова діаграм для презентацій

Обмеження. Складність роботи з великими даними (понад 1 млн. рядків), відсутність автоматизації складних моделей



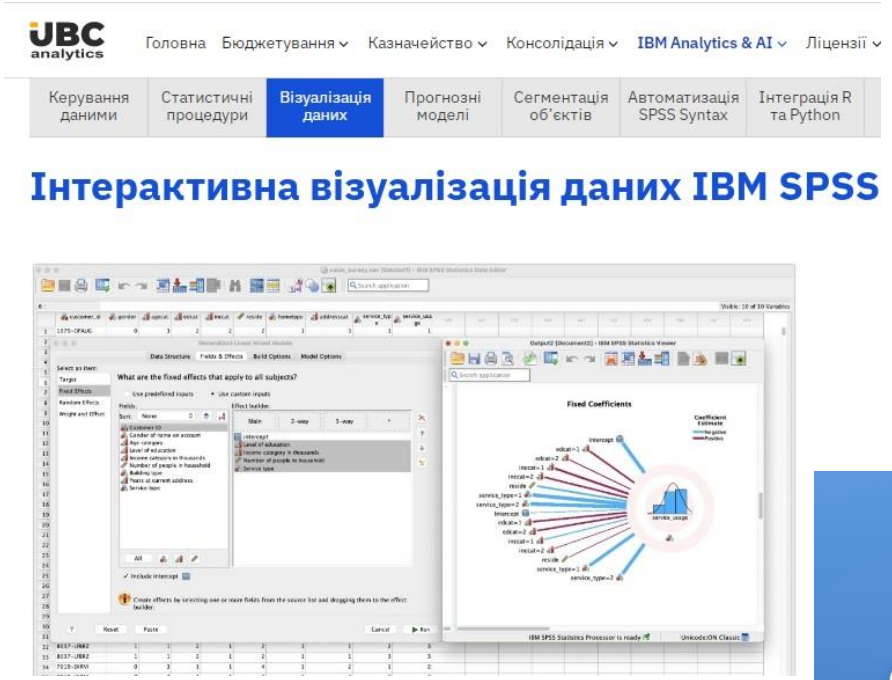


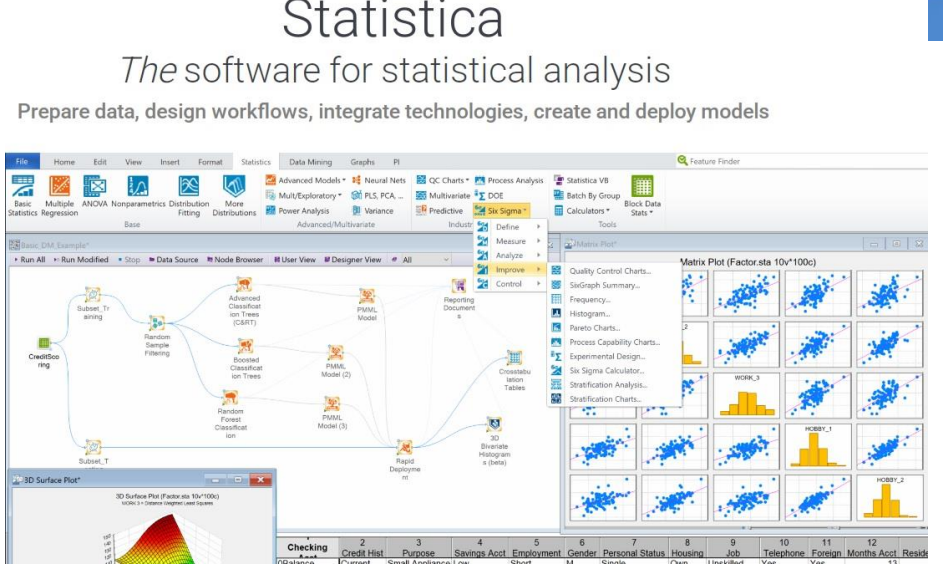
Спеціалізовані пакети (SPSS, PSPP, Statistica)

Професійне ПЗ для поглибленого статистичного аналізу, що часто використовується в соціологічних та медичних дослідженнях.

Переваги: Потужний математичний апарат, зручний інтерфейс для роботи зі складними тестами (ANOVA, факторний аналіз).

Сфера застосування: Академічні дослідження, складне анкетування, глибокий аналіз взаємозв'язків.



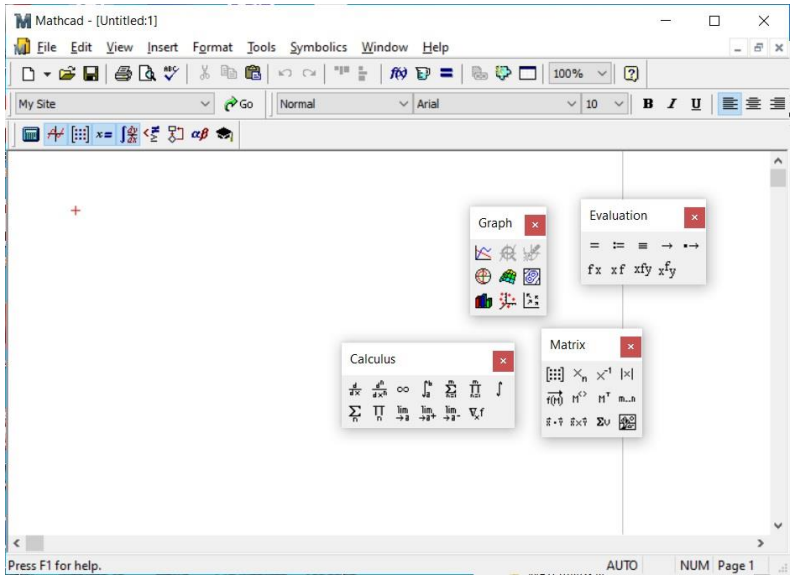


Спеціалізовані пакети (MathCAD, Maple)

MathCAD - це спеціалізоване середовище для автоматизації інженерних та математичних розрахунків. Програмний комплекс поєднує в собі обчислювальний рушій із текстовим редактором, що дозволяє створювати інтерактивні технічні документи

Виконання символьних та чисельних розрахунків, розв'язання систем диференціальних рівнянь, проведення ітераційних обчислень.

Підтримка природного математичного запису (WYSIWYG) та вбудована система контролю розмірностей фізичних величин, що мінімізує ризик помилок у складних інженерних проєктах.



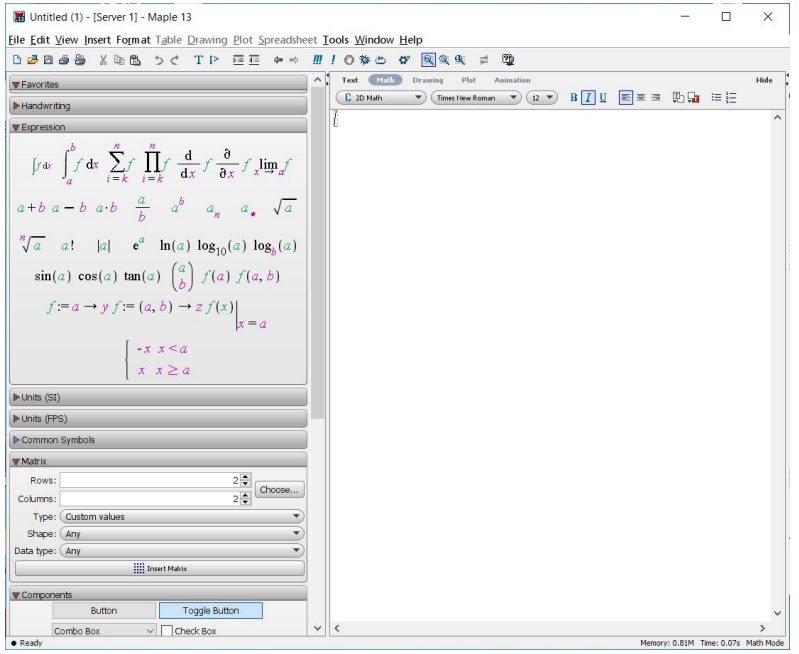
Сфера застосування. Проєктування, будівництво, прикладна фізика та підготовка технічної документації.

Maple - це універсальна система комп'ютерної алгебри (Computer Algebra System, CAS), призначена для виконання складних символьних і чисельних обчислень, а також для створення інтерактивних технічних застосунків

Програма спеціалізується на аналітичних перетвореннях, диференціюванні та інтегруванні, розв'язанні диференціальних рівнянь у символьному вигляді та лінійній алгебрі

Інструменти для створення високоякісної 2D та 3D графіки, анімації математичних процесів та розробки інтерактивних елементів керування.

Сфера застосування. Фундаментальна наука (математика, теоретична фізика), криптографія, освіта та розробка складних математичних алгоритмів



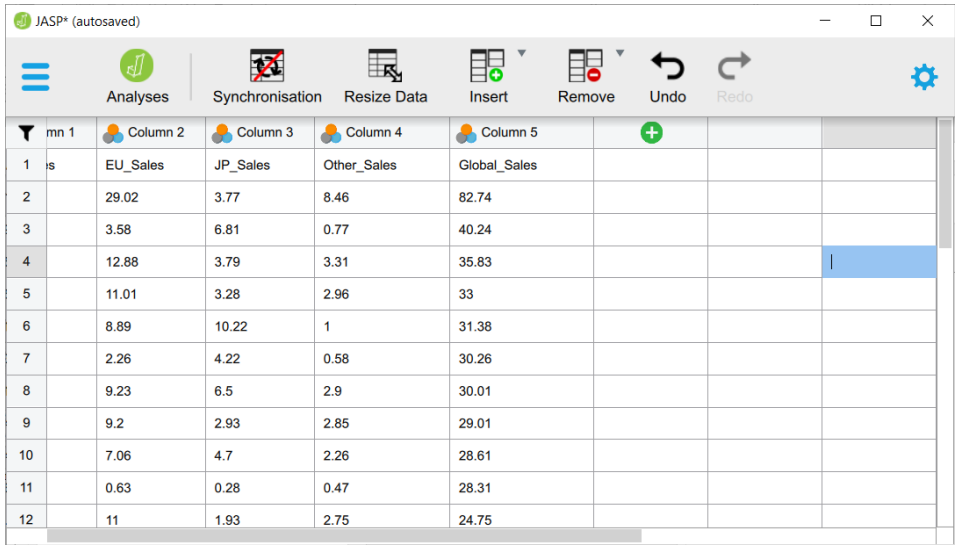
Спеціалізовані пакети (JASP)

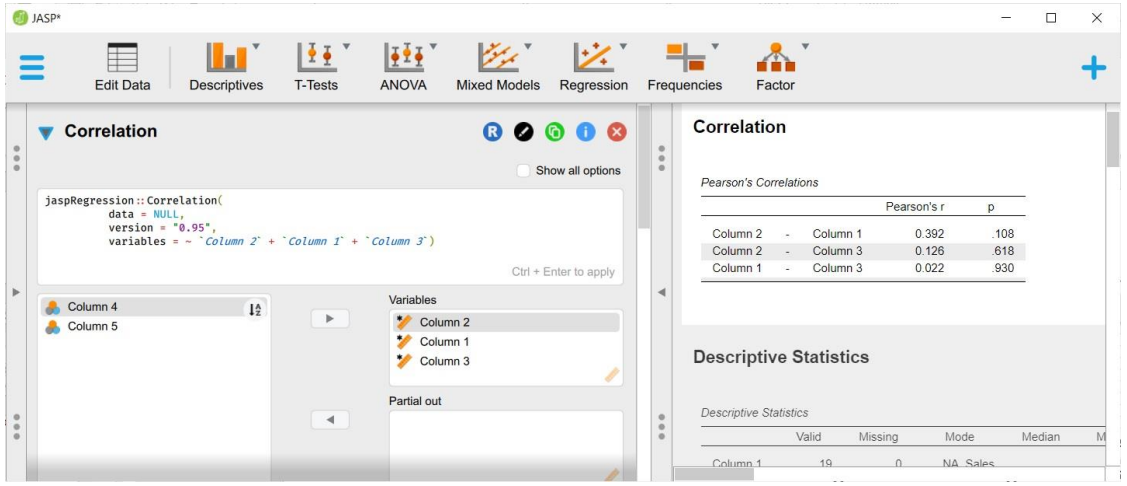
JASP - це відкрите (Open Source) програмне забезпечення, розроблене Університетом Амстердама для проведення сучасного статистичного аналізу. Пакет позиціонується як функціональна та зручна альтернатива комерційним продуктам (наприклад, SPSS).

Призначення - статистичне моделювання, перевірка гіпотез, кореляційний та факторний аналізи.

Ключові переваги - безкоштовність, сучасний графічний інтерфейс та унікальна можливість проведення як класичного (частотного), так і Байєсівського статистичного аналізу в межах однієї платформи

Сфери застосування - соціальні та поведінкові науки, медицина, маркетинг та академічна освіта





Мови програмування та бібліотеки (Python, R)

Стандарт для сучасної IT-індустрії та Data Science

Python. Використовує бібліотеки **Pandas** (обробка таблиць), **NumPy** (математика), **SciPy** (статистика) та **Matplotlib/Seaborn** (візуалізація)

Мова R. Мова, створена статистиками для статистиків. Має найширший вибір специфічних пакетів для аналізу

Переваги. Робота з Big Data, повна автоматизація (пайплайни - послідовні процеси обробки даних), можливість інтеграції в програмні продукти