СИСТЕМИ КЕРУВАННЯ ДАНИМИ

ЗМІСТ

1. Концепція «Open Data». Життєвий цикл даних. Система управління даними

Визначення відкритої ліцензії

Open Database License (ODbL) — "Share-alike" ліцензійна угода, призначена для того, щоб дозволити користувачам вільно ділитися, змінювати і використовувати відкриті бази даних, зберігаючи при цьому таку ж свободу і для інших. Поточна версія ліцензії — 1.0.

Посилання на джерело: https://uk.wikipedia.org/wiki/Open_Database_License

Концепція відкритих даних

Відкриті дані (або open data) — підхід до зберігання інформації, згідно з яким набір даних має бути у вільному доступі, вільно використовуватися та розповсюджуватися. При цьому такі дані можуть використовувати як неприбуткові організації, так і комерційні установи. Використання відкритих даних із комерційною метою не є забороненим.

До відкритих даних мають застосовуватися в обов'язковому порядку такі параметри:

- постійна доступність онлайн у цілодобовому режимі;
- відсутність паролів чи інших обмежень на рівні доступу;
- безкоштовність та анонімність надання й використання;
- можливість завантажувати у популярних форматах, що є універсальними (.doc, .pdf, .jpg, .jpeg);
- доступність API для поширення й використання на сторонніх платформах;
- отримання виключно з офіційних ресурсів.

Серед джерел, у яких слід шукати відкриті дані:

- онлайн-представництва та електронні сховища державних установ. Там оприлюднюються дані від міністерств. відомств, державних органів та установ на рівні міст. До числа таких ресурсів належить data.gov.ua),
- комерційні підприємства та організації,
- соціальні мережі та платформи.

Посилання на ресурс: https://nachasi.com/2018/06/12/vidkryti-dani/

Ініціатива data.gov

data.gov (англ. *data* – данные; .gov – домен верхнего уровня) — державний сайт США, що надає доступ до відкритих державних даних.

Data.gov - це, перш за все, федеральний відкритий сайт даних уряду. Однак державні, місцеві та племінні уряди можуть також зібрати метадані, що описують їхні відкриті ресурси даних на Data.gov для більшої відкритості. Data.gov не розміщує дані безпосередньо, а скоріше об'єднує метадані про відкриті ресурси даних в одному централізованому місці. Після того, як відкрите джерело даних відповідає необхідним вимогам формату та вимог до метаданих, команда Data.gov може витягнути безпосередньо з нього інформацію, синхронізуючи метадані цього джерела на Data.gov кожні 24 години.

Посилання на ресурс: https://www.data.gov/about

https://ru.m.wikipedia.org/wiki/Data.gov

Відкриті дані держав

Державні дані є одним із ключових інтересів для суспільства і численні некомерційні організації та окремі активісти домагаються відкритості державної інформації у формі, що може підлягати машинній обробці. Багато національних урядів в рамках стратегій «відкритої держави» створили веб-сайти для поширення частини даних, що обробляються в секторі державного управління. Деякі відомі сайти в стратегії відкритих даних: data.gov (портал відкритих даних США), science.gov, data.gov.uk, data.gov.in, data.gov.ua.

Посилання на джерело: https://ru.m.wikipedia.org/wiki/Открытые_данные

Життєвий цикл даних

Основними етапами життєвого циклу даних є виникнення, збереження, застосування та знищення. Знищення, з точки зору життєвого циклу даних, не представляє інтересу, оскільки причиною видалення є втрата інформативності даних. Фаза використання даних включає три етапи:

- пошук;
- обробку;
- аналіз.

Посилання на джерело:
https://pidruchniki.com/11930807/informatika/zhittyeviy_tsikl_danih_zbir_sistematizatsiya_danih

Стадія планування

Ця стадія розбивається на два етапи:

- розробка інформаційно-логічної моделі даних предметної області, який базується на описі предметної області, отриманому в результаті її обстеження. На цьому етапі спочатку визначають склад і структуру даних предметної області, які мають міститись у базі даних та забезпечувати виконання запитів, задач і застосувань користувача.
- визначення логічної структури бази даних.

Стадія збору

Існує декілька методів збору, необхідних для аналізу даних:

1. Облікові системи. Як правило, в облікових системах є механізми побудови звітів і експорту даних, тому отримання потрібної інформації є відносно нескладною операцією.
2. Непрямі дані.

3. Відкриті джерела.
4. Проведення незалежних маркетингових досліджень і аналогічних заходів щодо збору даних.
5. Внутрішні дані. Інформація заноситься в базу за різного роду експертними оцінками працівниками організації.

Стадія попередньої обробки даних

Попередня обробка - розділ аналізу даних що займається отриманням характеристик для подальшого використання у наступних розділах аналізу даних. Цей етап включає:

1. Обчислення базових характеристик (центральні моменти)
2. Перевірка основних гіпотез (симетричності, однорідності)
3. Перевірка стохастичності вибірки
4. Видалення аномальних спостережень
5. Розвідувальний аналіз

Посилання на ресурс: https://uk.wikipedia.org/wiki/Попередня_обробка_даних

Стадія опису

На даному етапі сукупність даних організовують за певною концепцією, яка описує характеристику цих даних і взаємозв'язки між їх елементами.

Стадія збереження

Зібрані дані перетворюються до єдиного формату, наприклад, таблиць Excel, текстових файлів, або компонентів довільної бази даних. Однією із важливих дій при цьому є визначення способу представлення даних. Як правило, вибирають один з наступних видів - число, рядок, дата, логічна змінна.

Стадія розвідки

Суть розвідки полягає у зборі інформації різного характеру.

Стадія інтегрування

Інтегрування даних - діяльність, що має на меті оптимальну організацію бази даних, при якій реалізовано всі необхідні взаємозв'язки між елементами даних, але база не містить повторів і зайвих елементів. Включає об'єднання даних, що знаходяться в різних джерелах, і надання даних користувачам в уніфікованому вигляді.

Посилання на джерело: https://uk.wikipedia.org/wiki/Інтеграція_даних

Стадія аналізу

Аналіз даних включає виконання послідовних, логічних дій з інтерпретації зібраних даних та їх перетворення у статистичні форми, необхідні для ухвалення маркетингових та управлінських рішень.

Визначення системи управління даними

Система управління даними – це спеціальний пакет програм, що забезпечує створення, супроводження і використання даних багатьма користувачами.

Посилання на джерело: https://sites.google.com/site/tehnikakomp/home/samostijne-vivcenna-materialu/sistema-upravlinna-bazami-danih-subd-subd-microsoft-access

Постачальники даних

Постачальники даних є основою для легкого контролю за тим, як дані можуть бути надані з джерела (як правило, вмісту файлу набору даних). Вони призначені для:

- Простого оголошення, налаштування та поєднання з деякими джерелами даних.
- Швидкого та ефективного знаходження даних за конкретними запитами, які надають лише вказані обсяги даних із зазначених місць.

2. Методи та засоби реалізації стадії збору та попереднього опрацювання даних.

Процедура збору даних

Збиранням, аналізом і використанням даних, як правило, займаються групи людей — від самоорганізовних активістських і дослідницьких груп до потужних державних і корпоративних інституцій. Ця робота може бути як спеціально організованим дослідженням, так і результатом обліку в процесі господарчої чи адміністративної діяльності. Її результати можуть як використовуватися всередині груп чи організацій, так і передаватися іншим групам чи організаціям для подальшої обробки і використання, аж до випадку публічно доступних відкритих даних.

Завдяки цьому, маючи намір вивчити і опрацювати дані про певне явище, можна спершу спробувати знайти їх у загальному доступі, потім, якщо їх там немає, звернутися з запитом до відповідних організацій, що мусять їх мати, і лише за недоступності їх у такий спосіб, думати над організацією спеціального дослідження.

Суцільне спостереження передбачає обстеження усіх без винятку одиниць генеральної сукупності.

Суцільні спостереження типові для дослідження відносно малих груп — в невеликій організації простіше опитати всіх працівників, ніж будувати вибірки. Такі спостереження можуть мати тривалий в часі характер, як от ведення традиційного класного журналу, куди методично і щоденно збирається інформація про відвідуваність та оцінки з усіх дисциплін, що вивчаються, для кожного з учнів цього конкретного класу.

Вибіркові спостереження менш ресурсномісткі і в багатьох випадках, єдино можливі. Здійснюючи спостереження, дослідник взаємодіє з досліджуваним об’єктом, чим змінює його властивості. Така зміна може бути незначною, а може бути драстичною.

Формати даних csv xml xlsx txt json

Машиночитаним називають формат стандартною комп’ютерною мовою, що може бути зчитаний автоматично броузером чи іншою комп’ютерною системою. Такі формати як **XML, JSON, CSV** із заголовками колонок є машиночитаними форматами.

Людиночитаними називають такі формати, які можна відкрити за допомогою простого текстового редактора, типу формату **.txt** (звичайний віндівський Блокнот) і щось там розібрати або відредагувати. Тобто, цілком читабельний текст у зображенні не є ані машиночитаним, ані людиночитаним.

Текстові файли мають перед бінарними дві переваги — вони «людиночитані» і менш прив’язані до конкретної програми. З іншого боку, бінарні файли значно компактніші за текстові. Проте і їх можна привести до компактнішого стану, використовуючи алгоритми компресії, що може відбуватися як на рівні одного файла чи групи файлів, так і на рівні файлової системи чи протоколу передачі.

Бінарні файли можуть бути як послідовностями команд процесора — програмами, так і файлами даних програм, наприклад формат Microsoft Excel **.xlsx** (який прийшов на заміну формату .xls) є бінарним. Бінарними є практично всі файли відео, зображень і звуку.

Формат, де поля значень розділено комами, називається **CSV**, Comma Separated Values, і це найрозповсюдженіший формат для збереження табличних даних у текстовий файл. **CSV** дозволяє записати не лише список значень, але й просту таблицю, починаючи кожен рядок таблиці з нового рядка.

Метадані

Змінні в таблиці набору даних варто називати коротко і чітко. Всі розшифровки, уточнення а також застосовані одиниці виміру варто виносити до словника даних — окремої таблиці метаданих (даних про дані). Зрозуміло що і ця таблиця має бути охайно структурованою, наприклад:

Колонка	Назва змінної	Категорія	Опис
id	Ідентифікатор у базі	Службова	Ідентифікатор підприємця в базі даних. Ціле число.
name	Ім'я	Конфіденційна	Ім'я об'єкта. Текст, записується у вигляді «ПРІЗВИЩЕ ІМ'Я ПО-БАТЬКОВІ»
bdate	Дата народження	Демографічна	Дата народження підприємця. Формат – дата, записується у вигляді РРРР-ММ-ДД

Набір даних

Набір даних відповідає змісту однієї таблиці бази даних або статистичній матриці даних, де кожна з колонок таблиці містить однорідні значення, а кожен з рядків таблиці відповідає певному члену набору даних.

Наприклад, набір даних про квіти може містити назву різновиду, розміри пелюсток, яскравість забарвлення тощо.

Посилання на використані джерела:

http://socialdata.org.ua/manual3/

https://uk.wikipedia.org/wiki/Метадані

http://filesreview.com/ru/info/xlsx

http://socialdata.org.ua/manual2/

http://lp.edu.ua/sites/default/files/dissertation/2017/4800/dys_boliubash_y.j.pdf

https://uk.wikipedia.org/wiki/Office_Open_XML

3. Огляд методів та засобів аналітичного опрацювання даних

Загальний процес аналітичного опрацювання великих даних можна розбити на п'ять етапів : перший – отримання та запис, другий – видобування, очищення та анотація, третій – інтеграція, агрегація та репрезентація, четвертий – моделювання та аналітика, п’ятий – інтерпретація.

Ці п'ять етапів складають два основні процеси: управління даними (перших три) та аналітика (останніх два). Управління даними включає в себе процеси та інформаційні технології для отримання та зберігання даних, їх підготовки, опрацювання та аналізу. Аналітика використовує методи, що використовуються для аналізу та видобування знань з інформаційних колекцій великих даних.

Наступні методи представляють собою підмножину інструментів для аналітичного опрацювання великих даних:

- Текстова аналітика – це методи для видобування інформації з текстових даних, таких як соціальні мережі, електронні листи, блоги, онлайн-форуми, опитування, корпоративні документи, новини та журнали викликів, тощо. Текстова аналітика включає статистичний аналіз, обчислювальну лінгвістику та машинне навчання.
- Аудіо аналітика використовується для видобування інформації з неструктурованих аудіоданих. При застосуванні до людської розмовної мови аудіоаналітика також називається аналізом мови.
- Аналіз відео, також відомий як аналіз відео контенту (VCA), включає в себе різні методи моніторингу, аналізу та отримання важливої інформації з відеопотоків.
- Соціальна аналітика використовується для аналізу структурованих та неструктурованих даних з соціокомункаційних каналів та джерел. Соціальні медіа та мережі – це узагальнений термін, який включає різні інформаційно-технологічні платформи, які дозволяють користувачам створювати та обмінюватися вмістом.
- Інтелектуальна аналітика включає в себе різні методи, які отримують результати на основі історичних та поточних даних.

Типи даних, шкали.

Існує чотири типи даних (шкали вимірювання) номінальна, порядкова, інтервальна і відносна. Такий тип класифікації увів Стенлі Стівенс у 1964 році.

Номінальна є найпростішою для розуміння, такі шкали використовуються для маркування змінних без будь-якого кількісного значення . «Номінальні» шкали можуть бути просто названі «мітками». Всі ці шкали є взаємовиключними. Тобто надбання характеристики А унеможливлює наявність характеристики Б. Підтип номінальної шкали тільки з двома категоріями (наприклад, чоловічий / жіночий) називається «дихотомічним».

При використанні порядкових шкал порядок значень є важливим і значущим, але відмінності між кожним з них не відомі. Порядкові шкали зазвичай використовуються для опису нечислових понять, таких як задоволення, щастя, дискомфорт та інші суб'єктивні поняття. Наприклад, ми не можемо точно вказати різницю між “добре” та “дуже добре”.

Інтервальні шкали - це числові шкали, в яких ми знаємо як порядок, так і точні відмінності між значеннями. Класичним прикладом інтервального шкали є температура за

Цельсієм, тому що різниця між значеннями однакова. Наприклад, різниця між 60 і 50 градусами становить 10 градусів, а також різниця між 80 і 70 градусами. Тривалість є ще одним хорошим прикладом інтервального шкали, в якій збільшення відомі і послідовні . Інтервальні шкали хороші тим, що відкривається область статистичного аналізу цих наборів даних. Наприклад, центральна тенденція може бути виміряна модою, медіаною або середнім значенням; стандартне відхилення також можна розрахувати.

Проблемою інтервальних шкал є відсутність «істинного нуля». Наприклад, немає такої речі, як «відсутність температури», принаймні у прикладі з градусами Цельсія. У разі інтервальних шкал нуль не означає відсутність значення, а фактично є іншим числом, використовуваним на шкалі . Без справжнього нуля неможливо обчислити відносини. З інтервальними даними ми можемо скласти і відняти, але не можемо помножити або розділити.

Шкали відносин є найбільш повними , коли мова йде про шкали вимірів, тому що вони вказують нам на порядок, на точне значення між одиницями, а також мають абсолютний нуль, що дозволяє використовувати широкий спектр як описової, так і логічної статистики.

Статистичне опрацювання даних

Статистичний аналіз даних - це процедура виконання різних статистичних операцій, це свого роду кількісне дослідження, яке спрямоване на кількісну оцінку даних, і, як правило, застосовується певна форма статистичного аналізу. Кількісні дані в основному включають описові дані, такі як дані обстеження та спостережень.

Аналіз статистичних даних, як правило, включає певну форму статистичних інструментів, які неспеціаліст не може виконувати, не маючи статистичних знань. Існують різні пакети програм для виконання статистичного аналізу даних. Це програмне забезпечення включає в себе систему статистичного аналізу (SAS), статистичний пакет для соціальних наук (SPSS), і т.д.

Дані статистичного аналізу даних складаються з змінних . Дані є одновимірними або багатовимірними. Залежно від кількості змінних, дослідник виконує різні статистичні методи.

Дані в статистичному аналізі даних в основному складаються з 2 типів, а саме безперервних і дискретних даних. Безперервні дані - це те, що не підраховується. Наприклад, інтенсивність світла може бути виміряна, але не може бути підрахована. Дисекретні дані - це ті, які можна підрахувати.

У статистичному аналізі даних є важливим завданням, що включає статистичний висновок. Статистичний висновок в основному складається з двох частин: оцінки та випробувань гіпотези.

Оцінка в статистичному аналізі даних в основному включає параметричні дані - дані, які складаються з параметрів. З іншого боку, тести гіпотези при аналізі статистичних даних в основному включають непараметричні дані - дані, які не містять параметрів.

Композитні індикатори

Інтегральні (композитні) індикатори об’єднують набір змінних за допомогою науково обґрунтованих ваг, щоб відобразити певну сторону складного (латентного явища) в індексі. Їх часто використовують у суспільних науках, щоб показати приховані, недоступні для безпосереднього вимірювання явища чи процеси.

Загалом, для побудови інтегральних індикаторів використовують :

- експертні методи, які включають прямі та непрямі експертні оцінки порівнюваних значень. Якість результатів тут залежить від кваліфікації експертів. Слід зауважити також, що оцінки експертів є суб’єктивні;
- апріорні методи – вид інтегрального показника та його параметри обирають, керуючись теоретичними уявленнями про сутність досліджуваного економічного явища, характер взаємозв’язку вихідних показників, їх значення для зіставлення економічних процесів;
- методи «розпізнавання образів» – це різноманітні методи багатовимірної класифікації об’єктів. Вони є об’єктивніші за вже перелічені методи. Їх можна використовувати для групування часткових показників, з яких потім виділяти в кожній групі найтиповіший показник і розглядати його як інтегральну характеристику для відповідної групи вихідних показників;
- методи факторного і компонентного аналізу застосовують досить часто, оскільки вони дають непогані результати, хоча нерідко виникають певні труднощі: виникнення ваг з негативними значеннями, слабкий кореляційний зв’язок агрегованого показника з деякими з часткових показників тощо.
- непараметричні методи.

Зазначимо, що універсальної методики для створення інтегрального індикатора немає. Необхідно в кожному конкретному випадку керуватись особливостями досліджуваного явища

Нормування

Стандартизація (Нормування) даних є необхідним початковим етапом перетворення даних при використанні багатьох багатовимірних статистичних методів - зниження розмірності простору ознак (факторний, компонентний аналіз, с), класифікації об'єктів (кластерний аналіз,) та ін. , особливо якщо змінні виміряні в шкалах, істотно розрізняються в величинах (мікрони одиниць - мільярди одиниць).

Внаслідок поширеності і необхідності в статистичних пакетах процедура нормування (стандартизації) зазвичай винесена в меню .

Пошук статистичної залежності

Статистична залежність - умова, за якої дві випадкові величини не є незалежними. X і Y позитивно залежні, якщо умовна ймовірність P (X | Y), X задана Y, більше, ніж ймовірність P (X), X, або еквівалентно, якщо P **(X & Y)> P (X). P (Y).**

Якщо зміна однієї з випадкових величин призводить до зміни середнього іншої випадкової величини, то статистичну залежність називають кореляційною. Самі випадкові величини, пов'язані корррреляціонной залежністю, виявляються корельованими.

Часто потрібно визначити, як залежить спостережена випадкова величина від однієї або декількох інших величин. Найбільш загальний випадок такої залежності - залежність статистична.

Для залежних випадкових величин має сенс розглянути математичне очікування однієї з них при фіксованому значенні інший (інших). Таке умовне математичне очікування показує, як впливає на середнє значення першої величини зміна значень другий. Скажімо, вартість квартири залежить від площі, поверху, району та інших параметрів, але не є функцією від них. Зате в широких припущеннях можна вважати її математичне очікування функцією від цих величин.

Зрозуміло, спостерігати це середнє значення ми не можемо - можна лише спостерігати значення першої випадкової величини при різних значеннях інших . Вхідні дані, або «фактори», як правило, відомі. На виході ми спостерігаємо результат перетворення вхідних даних в ящику з будь-яким правилам.

Метод головних компонентів

Метод головних компонентів (МГК) базується на ідеї обліку однієї ознаки на підставі другої . Слід відзначити, що мова не йде тільки про дві ознаки. У такому випадку метод головних компонент малоефективний. Його використовують, як правило, при десятках взаєпов'язаних ознак. При цьому ставиться мета " набрати" певну частину загальної варіації результативної ознаки мінімальною кількістю змінних. Останні підбирають до тих пір, поки сума їх дисперсій не сягатиме заданої частки у дисперсії досліджуваного явища (наприклад, 60 %, 80 %, 90 % і т.д.).

Метод головних компонент розв'язує такі завдання:

- 1. Відшкодування скритих, об'єктивно існуючих закономірностей у зміні явищ.
- 2. Характеристика явища, що вивчається, числом ознак, значно меншим взятих, на початковому етапі. Число головних компонент, виділених в процесі дослідження, буде вміщувати (у компактній формі) більше інформації, ніж початково виміряні ознаки.
- 3. Виявлення ознак, найбільш тісно пов'язаних з головною компонентою. Інакше кажучи, вивчення стохастичного зв'язку між ними (зв'язок, при якому зі зміною однієї змінної змінюється закон розподілу другої).
- 4. Прогнозування рівней досліджуваних явищ на підставі рівняння регресії, яке одержане по інформації головних компонент.

Переваги такого методу прогнозування на відміну від класичного регресійного аналізу можна пояснити тим, що при останньому в модель намагаються включити максимально можливу кількість факторів, які в економічних явищах часто характеризуються істотною корельованістю (мультилінеарністю). Прогноз за такими змінними, як правило, буває не точним. Тому виникає завдання про заміну вихідних взаємопов'язаних змінних сукупністю некорельованих параметрів. Це завдання вирішується математичним апаратом - методом головних компонент, який являє собою характеристики, побудовані на підставі первинно виміряних ознак.

Як негативну сторону методу головних компонент слід назвати складність математичного апарату, зумовленого абсолютністю знань теорії ймовірностей, математичної статистики, лінійної алгебри, а також математичного забезпечення ПЕОМ. Формальне використання стандартних програм без розуміння математичної суті обчислювальних процедур може призвести до необгрунтованих висновків. Слід також пам'ятати про професіональні знання суті досліджуваних економічних явищ. Тільки за таких умов метод головних компонент може стати могутнім математичним засобом пізнання існуючих реалей у галузі соціально - економічних явищ.

Кластерування даних

Кластерний аналіз (кластерування даних) — задача розбиття заданої вибірки об'єктів (ситуацій) на підмножини, які називаються кластерами, так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися. Завдання кластеризації відноситься до статистичної обробки, а також до широкого класу завдань навчання без вчителя.

Кластерний аналіз — це не якийсь один алгоритм, а загальна задача, для розв'язання якої використовуються різні підходи. Зокрема, алгоритми побудови кластерів можуть суттєво відрізнятись у розумінні того, що відносити в один кластер і як їх ефективно шукати. Серед популярних концепцій кластерів є групи з елементами, які утворюються ґрунтуючись на відстані між ними, на щільності ділянок у просторі даних, інтервалах або на конкретних статистичних розподілах.

Тому кластеризація може бути сформульована як задача багатокритеріальної оптимізації. Відповідний алгоритм кластеризації та вибору параметрів (включаючи такі параметри, як функція відстані, порогове значення щільності або кількість очікуваних кластерів) залежать від конкретного набору даних та мети використання результатів.

Кластерний аналіз як такий є не автоматизованим завданням, а ітераційним процесом виявлення знань або інтерактивної багатокритеріальної оптимізації, який містить спроби та невдачі. Часто доводиться змінювати процес опрацювання даних та параметри моделі поки не буде отримано з результат з заданими властивостями.

Модулі для виконання аналітичного опрацювання даних.

В OLAP (або Online Analytical Processing) зростає популярність завдяки збільшенню обсягів даних і визнанню ділової цінності аналітики. До середини 90-х років виконання OLAP-аналізу було надзвичайно дорогим процесом, який обмежувався переважно великими організаціями.

Основним постачальником OLAP є Hyperion, Cognos, Business Objects, MicroStrategy. Вартість одного місця становила від 1500 до 5000 доларів на рік. Створення середовища для проведення аналізу OLAP також вимагатиме значних інвестицій у час та грошові ресурси.

Це змінилося, оскільки основний постачальник бази даних почав впроваджувати модулі OLAP у свою базу даних - Microsoft SQL Server 2000 з службами аналізу, Oracle з Express та Darwin, а також IBM з DB2.

Посилання на використані джерела:

https://www.statisticssolutions.com/statistical-data-analysis/

https://www.tandfonline.com/doi/abs/10.1080/00182494.1969.10593890?journalCode=vzhm20

https://www.bigskyassociates.com/blog/bid/356764/5-Most-Important-Methods-For-Statistical-Data-Analysis

https://www.jinfonet.com/resources/bi-defined/olap/

4. Огляд методів та засобів візуального опрацювання даних

Візуалізація інформації — це інтерактивне вивчення візуального представлення абстрактних даних для посилення людського пізнання. Абстрактні дані включають в себе як числові так і нечислові, такі як текст і географічна інформація. Тим не менш, візуалізація інформації відрізняється від наукової візуалізації: «існує infovis, коли просторове уявлення обране, і це scivis, коли просторове уявлення дається»

Кращі практики візуалізації даних

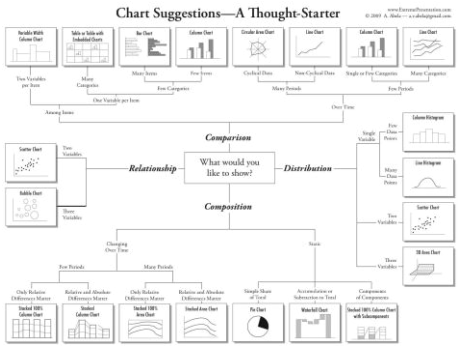
Існує чотири основні типи презентацій, які можна використовувати для подання даних:

- Порівняння
- Композиція
- Розподіл
- Відносини

Вибір правильної діаграми

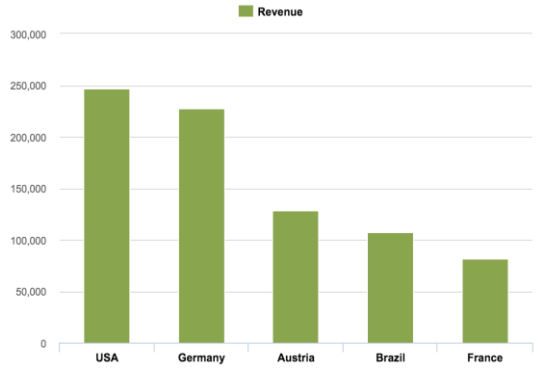
Щоб визначити, яка діаграма найкраще підходить для кожного з цих типів презентацій, спочатку потрібно відповісти на кілька запитань:

- Скільки змінних ви хочете показати на одному графіку? Один, два, три, багато?
- Скільки елементів (точок даних) ви будете показувати для кожної змінної? Лише кілька або багато?
- Чи відображатимете значення протягом певного періоду часу або серед елементів або груп?



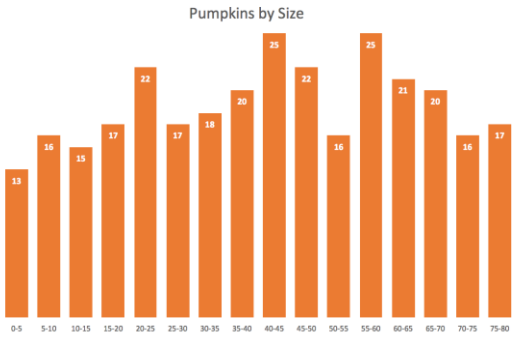
Варто зауважити, що **таблиці** є по суті джерелом для всіх графіків. Вони найкраще використовуються для порівняння, аналізу або аналізу відносин, коли існує лише кілька змінних і точок даних. Не варто мати сенс створювати діаграму, якщо дані можна легко інтерпретувати з таблиці.

Типи діаграм та їх використання



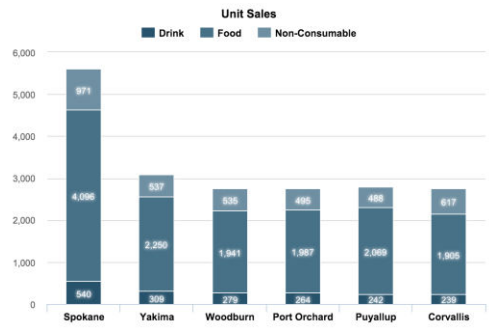
Стовпчасті діаграми

Стовпчасті діаграми, ймовірно, найбільш часто використовуваний тип діаграми. Ця діаграма найкраще використовується для порівняння різних значень, коли важливі певні значення, і очікується, що користувачі будуть шукати та порівнювати окремі значення між кожним стовпцем. З графіками стовпців можна порівняти значення для різних категорій або порівняти зміни значень за певний період часу для однієї категорії.



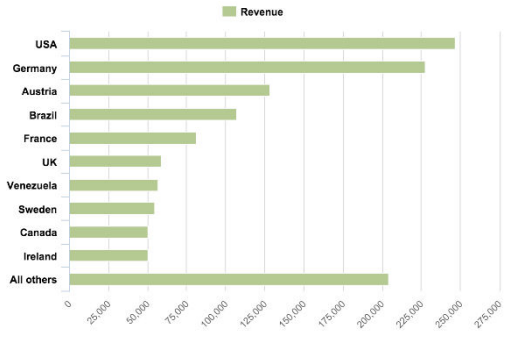
Гістограма

Гістограма - це звичайна варіація стовпчастіої діаграми, які використовуються для представлення розподілу та співвідношень однієї змінної над набором категорій. Гарним прикладом гістограми може бути розподіл оцінок на шкільному іспиті або розмір гарбуза, розділених за групами розмірів, на фестивалі гарбуза.



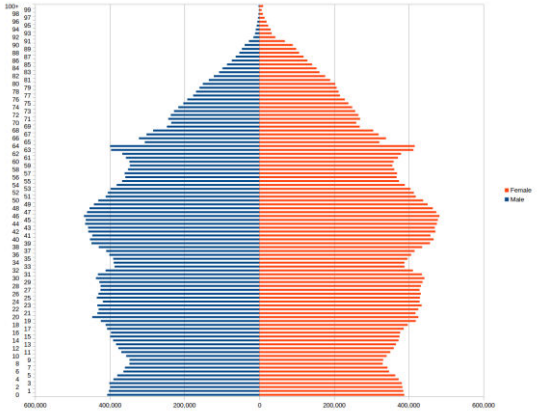
Складені стовпчасті діаграми

Використовуйте складені стовпчасті діаграми для показу композиції. Не використовуйте занадто багато елементів композиції (не більше трьох або чотирьох) і переконайтеся, що складові частини відносно подібні за розміром.



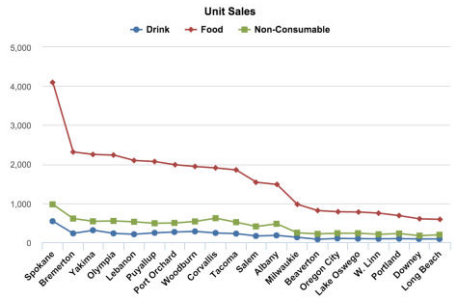
Лінійні діаграми

Штрихові діаграми є по суті горизонтальними графіками стовпців. Якщо ви маєте довгі назви категорій, краще використовувати бар-діаграми, оскільки вони дають більше місця для довгого тексту. Ви повинні також використовувати діаграми, замість стовпчикових, коли кількість категорій перевищує сім (але не більше п'ятнадцяти) або для відображення набору з негативними числами.



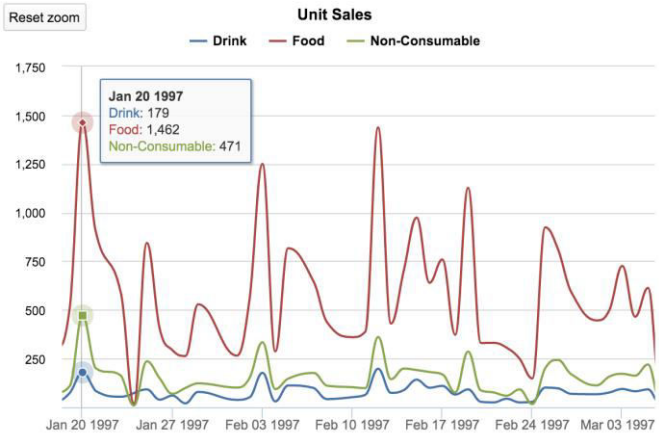
Лінійні гістограми

Так само, як і стовпчасті діаграми, для представлення гістограм можна використовувати лінійні гістограми. Гарний приклад лінійної гістограми - це розподіл населення за віком.



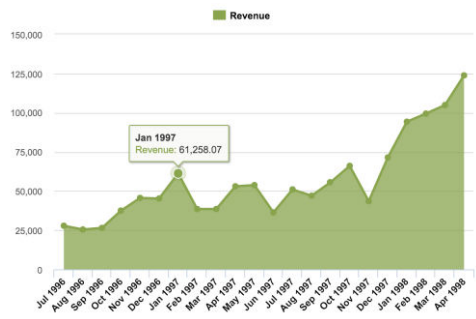
Лінійні графіки

Лінійні діаграми є одними з найбільш часто використовуваних типів діаграм. Використовуйте рядки, якщо у вас є безперервний набір даних. Вони найкраще підходять для візуалізації даних на основі тренду протягом певного періоду часу, коли кількість точок даних дуже висока (більше 20). З лінійними діаграмами акцент робиться на продовження або потік значень (тенденція), але все ще існує певна підтримка для порівняння одиничних значень, використовуючи маркери даних (тільки з менш ніж 20 точками даних).



Графіки з часовою шкалою

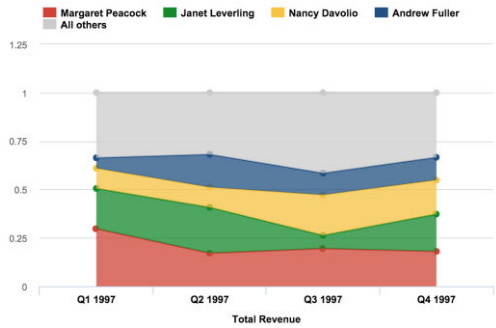
Часова шкала - це варіація лінійних діаграм. Очевидно, що будь-яка діаграма ліній, яка показує значення за певний період часу, є графіком часової шкали. Єдина відмінність полягає в тому, що функціональність - більшість графіків тимчасової шкали дозволяють збільшувати та зменшувати масштаб і стискати або розтягувати осі часу, щоб побачити більш детальну інформацію або загальні тенденції.



Діаграми з ділянками

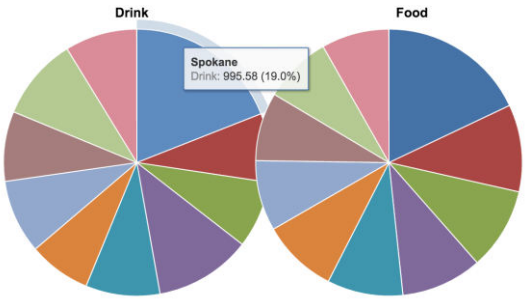
Діаграма по суті є лінійною діаграмою - добре для тенденцій і деяких порівнянь. Діаграми діапазонів будуть поповнювати область нижче лінії, тому найкращим чином використовувати цей тип діаграм для представлення накопичувальних змін у часі, таких як

запас елементів, кількість працівників або ощадний рахунок.



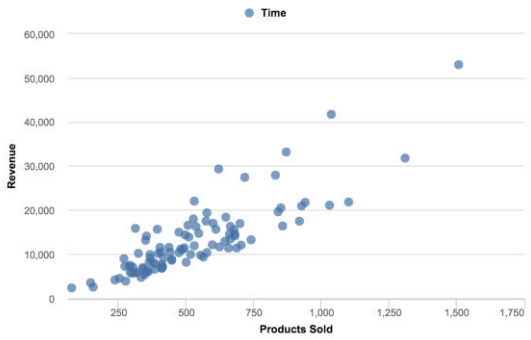
Складені діаграми з ділянками

Графіки стекових областей найкраще використовуються для показу змін у складі в часі. Хорошим прикладом може служити зміна частки ринку серед провідних гравців або частки доходу за виробничою лінійкою протягом певного періоду часу.



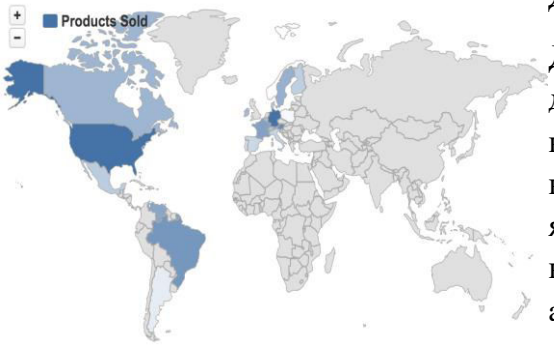
Кругові діаграми

Кругова діаграма зазвичай представляє числа у відсотках, що використовуються для візуалізації частини до цілого відношення або композиції. Кругові діаграми не призначені для порівняння окремих розділів один з одним або для представлення точних значень (для цього слід використовувати діаграму).



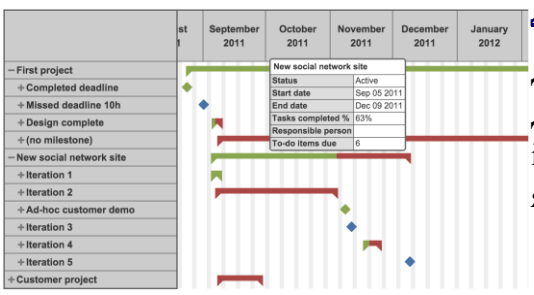
Точкова діаграма

Точкова діаграма використовуються в основному для аналізу кореляції та розподілу. Добре показувати взаємозв'язок між двома різними змінними, коли один з них співвідноситься з іншим.



Діаграми з використанням карт

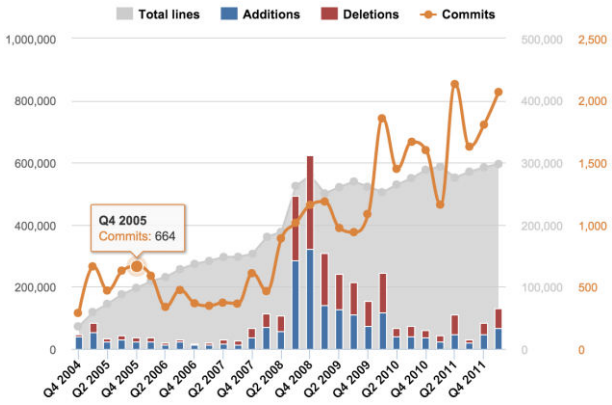
Діаграми з використанням карт добре підходять для того, щоб надати своїм цифрам географічний контекст, щоб швидко визначити найкращі та найгірші місця, тенденції та викиди. Якщо у вас є якісь дані про місцезнаходження, такі як координати, назви країн, назви штатів або абревіатури або адреси, ви можете побудувати пов'язані дані на карті.



Діаграми Ганта

Діаграми Ганта підходять для планування проектів. Діаграми Ганта є по суті проектними картами, що ілюструють, що потрібно зробити, в якому порядку і в який термін. Ви можете візуалізувати загальний час

виконання проекту, залучені ресурси, а також порядок і залежності завдань.



Діаграми багатьох осей

Якщо ви хочете показати зв'язки та порівняти змінні на дуже різних масштабах, найкращим варіантом може бути наявність декількох осей. Діаграма з кількома осями дасть змогу складати дані, використовуючи дві або більше у-осей і одну спільну вісь x.

Посилання на ресурс: https://eazybi.com/blog/data_visualization_and_chart_types/