



IBM Introduction to Data Analytics 筆記 [IBM 數據分析課程]

🔍 Type	🔗 活動專案／作者頁面, , 📄 數據分析／資料科學, 📄 證照測驗／線上課程, 📄 程式設計
# Tag List	🔗 ML 的測驗筆記 Note, 🔗 資料蒐集, 🔗 資料分析, 🔗 「課程」IBM Data Analysis and Visualization Foundations 課程, 🔗 數據分析, 🔗 資料科學
🕒 Update Time	2021年9月29日 上午9:39
⚙️ Status	存檔
⚙️ second input	Not started

現代數據生態系統：



定義

1. 由相互關聯、獨立和不斷發展的實體組成的整個網路
2. 從不同的來源、不同類型的分析和技能中整合以生成見解的數據
3. 最終使用者，如業務利益相關者、分析師和程式師，他們為各種目的使用數據。

開始資料源

可在各種結構化和非結構化數據集中提供

- 包括文本、圖像、視頻、點擊流、使用者對話、社交媒體平臺、物聯網或物聯網設備、流式傳輸數據的即時事件、傳統資料庫以及來自專業數據供應商的數據機構

How to use?

- 將原始來源的數據副本拉入資料儲存庫。
在這裡階段，您只考慮取得需要的數據，以便處理資料格式、源和介面，從而可以提取這些資料。
- 進行組織
- 清理
- 優化
- 最終用戶存取

 資料須符合組織內執行的合規性和標準。

 在組織內堅持主數據表

確保組織所有應用程式和系統中主數據標準化

What's problem?

- 獲得數據的可靠性
 - 安全性
 - 完整性
-
- 數據管理
 - 提供高可用性
 - 靈活性
 - 可訪問性
 - 安全性的數據存儲庫合作。

應用

- 分析人員可能需要原始資料來處理。
- 業務利益相關者可能需要報告和儀錶板。
- 應用程式可能需要自訂 API 來拉取此資料

數據生態系統中的關鍵參與者

資料工程師

將原始資料轉換為可用資料

- 從不同的來源提取、集成和組織資料
- 在資料儲存庫中對資料進行清潔轉換並準備資料設計、儲存和管理資料。

數據分析師

使用此資料產生見解

- 數據分析員將數據和數字翻譯成簡單的語言，以便組織能夠做出決策
- 數據分析員檢查和清理數據以獲得見解、識別相關性、查找模式以及應用統計方法。
- 分析和挖掘資料並視覺化資料，以解釋和呈現資料分析的結果。

數據科學家

數據分析和數據工程來預測未來

- 分析數據以獲得可操作的見解，並構建機器學習或深度學習模型，以訓練過去的數據以創建預測模型
- 下個月我可能會獲得多少新的社交媒體關注者
- 我的客戶在下個季度可能會在競爭中失去多少百分比
- 這種金融交易對於這個客戶來說是不尋常的嗎？
- 需要數學、統計知識，並公平理解程式設計語言、資料庫和構建數據模型。他們還需要領域知識。

商業/智慧分析師

- 利用數據分析師和數據科學家的工作，研究對其業務可能產生的影響，以及他們需要採取或建議的行動。

使用這些見解和預測來推動有利於和發展其業務的決定

BI 分析師

- 重點是市場力量和外部影響，塑造他們的業務。
- 通過組織和監控不同業務功能的數據，並探索這些數據來提取可操作的見解和可操作性，從而提供業務智慧解決方案，從而提高業務績效。

幫助組織挖掘大量數據並將它們轉化為可操作的見解方面所發揮的作用

數據分析

定義

數據分析是收集、清理、分析和挖掘數據、解釋結果和報告結果的過程

- 發現數據中的模式和不同數據點之間的相關性。
- 通過這些模式和相關性，產生了見解，並得出了結論。



描述性分析 發生了什麼

通過總結過去的數據並向利益相關者介紹調查結果，說明回答有關特定時間段內發生的事情的問題

提供對過去事件的基本見解

- 根據組織的關鍵績效指標或現金流分析跟蹤過去表現

診斷分析 為什麼它發生

- 為什麼會這樣？

需要描述性分析的見解來深入挖掘，以找到結果的原因

- 在行銷沒有變化的地區，網站的流量突然發生變化，沒有明顯的原因，或者銷售額增加

預測分析 建議 接下來會發生什麼

- 接下來會發生什麼？

歷史資料和趨勢用於預測未來結果。

- 企業應用預測分析的一些領域是風險評估和銷售預測

注意

預測分析的目的不是 說說將來會發生什麼，目標是預測將來會發生什麼

規範分析 規定 下一步應該做什麼

- 應該對此採取什麼措施？

通過分析過去的決定和事件，不同結果的可能性。估計根據該方針決定行動方針。

資料分析過程關鍵步驟

過程包括：

- 對問題和預期結果有一個瞭解。
- 為評估結果設定一個明確的指標。
- 收集、清潔、分析和挖掘數據以解釋結果。
- 以影響決策的方式傳達調查結果。

1. 瞭解問題和預期結果

數據分析首先瞭解需要解決的問題和需要實現的預期結果。在分析過程開始之前，需要明確定義您所處的位置和想要去的地方。

2. 設定明確的指標

例如

- 在一個區域銷售的產品 X 數量
- 如何測量產品 X。

3. 蒐集數據

- 一旦你知道你要測量什麼，你將如何測量它，你確定你需要的數據，你需要這些數據的數據來源，以及最合適的工作工具。

4. 清潔數據

- 發現缺失或不完整的值和離群值
- 標準化來自多個來源的資料

5. 提取和分析數據

- 瞭解趨勢、識別相關性並查找模式和變化

6. 解釋結果

- 需要評估您的分析是否可以抵禦反對意見，以及您的分析是否存在任何限制或情況可能不成立

7. 介紹您的發現

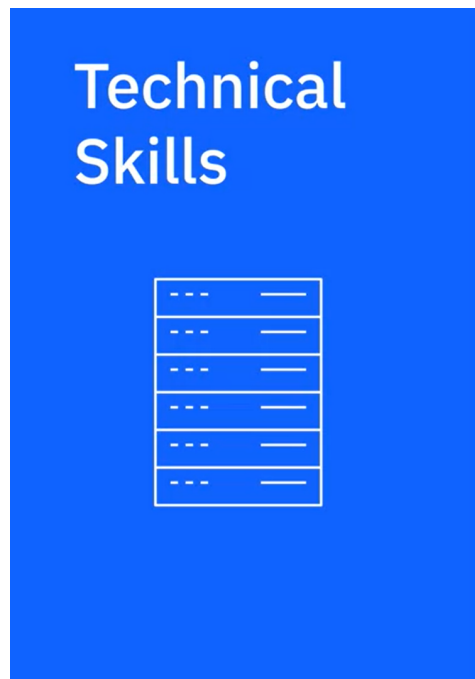
- 以清晰和有影響的方式溝通和展示您的發現的能力與分析本身一樣是數據分析過程的重要組成部分。
 - 報告、儀錶板、圖表、圖表、地圖、案例研究只是您展示數據的一些方式

數據分析員角色

| 職責

- 從初級和次要數據源獲取數據
- 建立查詢
- 從資料庫和其他數據收集系統中提取所需的數據
- 篩選、清理、標準化和重組數據，為數據分析做準備
- 使用統計工具解釋數據集
- 使用統計技術識別數據中的模式和相關性
- 分析複雜資料集中的模式並解釋趨勢

- 編寫有效傳達趨勢和模式的報告和圖表
- 建立適當的文件來定義和演示資料分析過程的步驟



已成為我的習慣

Expertise in using spreadsheets

Microsoft Excel or Google Sheets

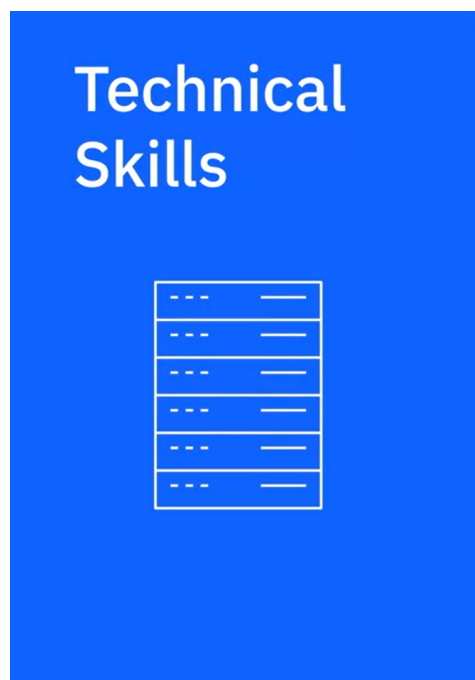
Proficiency in statistical analysis and visualization tools and software

IBM Cognos, IBM SPSS, Oracle Visual Analyzer, Microsoft Power BI, SAS, and Tableau

Proficiency in programming languages

R, Python, C++, Java, and MATLAB

Good knowledge of SQL and ability to work with data in relational and NoSQL databases



如里能夠與本何自適

按下 Esc 結束全螢幕

Proficiency in programming languages

R, Python, C++, Java, and MATLAB

Good knowledge of SQL and ability to work with data in relational and NoSQL databases

The ability to access and extract data from data repositories

Data Marts, Data Warehouses, Data Lakes, and Data Pipelines

Familiarity with Big Data processing tools

Hadoop, Hive, and Spark



- 統計能力
以說明您分析數據，驗證您的分析，並識別謬誤和邏輯錯誤。
- 分析技能
說明您研究和解釋數據，理論化和預測。

- 解決問題的技巧

因為歸根結底，所有數據分析的最終目標是解決問題探究發現過程必不可少的技能，即從不同利益相關者和使用者的角度理解問題，因為數據分析過程真正從明確闡明問題陳述和預期結果開始。

- 資料視覺化技能

說明您根據受眾、資料類型、上下文和分析最終目標，確定有效呈現您的發現的技術和工具。

- 專案管理技能

以管理流程、人員、依賴性和計劃時程表這讓我們瞭解了您作為數據分析師的軟技能

- 有效地溝通，報告和介紹你的發現

→ 講述一個令人信服的故事，為你的工作爭取支持和認同

技能

- Python
- SQL
- R
- Tableau
- Power BI

-
- 知道要使用哪些正確的數據，使用哪些工具
 - 向關鍵利益相關者提供數據
 - 商業頭腦
 - 演示技巧

摘要

資料分析員的角色跨越：

- 獲取最適合使用案例的數據。
- 準備和分析數據以瞭解其代表的內容。
- 向需要根據調查結果採取行動的利益攸關方解釋並有效傳達資訊。

- 確保記錄該過程以供將來參考和重複使用。

為了成功發揮這一作用，數據分析師需要技術、功能和軟技能的組合。

- 技術技能包括不同程度的熟練程度，能夠使用電子錶格、統計工具、可視化工具、程式設計和查詢語言，以及使用不同類型的數據存儲庫和大數據平臺的能力。
- 了解統計、分析技術、解決問題、從多個角度探索情況的能力、數據可視化和專案管理技能 - 所有這些都屬於數據分析師需要的功能技能，以便發揮有效作用。
- 軟技能包括協作、有效溝通、用數據講述引人入勝的故事以及獲得利益相關者的支持和支援的能力。探索不同路徑的好奇心和直覺有助於根據過去的經驗來感受未來，這也是成為一名優秀的數據分析師的基本技能

数据分析生态系统

一般来说，数据包括事实、观察、感知、数字、字符、符号和图像

Types of Data 数据类型

- Structured 结构化

可以存储在数据库中的格式组织良好的数据

- 结构化数据具有定义良好的结构或坚持指定的数据模型
- 可以使用行和列以表格方式表示
- 结构化数据是可以在典型的数据库中收集、导出、存储和调用的客观事实和数字
- 通常可以在关系数据库或 SQL 数据库中存储结构化数据

- Semi-structured 半结构化

部分组织和部分自由形式的数据

- 具有一些组织属性但缺少固定属性的数据或者僵化的模式
- 半结构化数据不能像数据库中那样以行和列的形式存储
- 它包含用于对数据进行分组和组织的标记和元素或元数据
- XML 和 JSON 允许用户定义标记和属性，以便在层次结构中存储数据
- EX. Email, Text

- Unstructured 非结构化的

不能按照常规组织成行和列的数据

- 不能以行和列的形式在主流关系数据库中组织
- 不遵循任何特定的格式、顺序、语义或规则
- 非结构化数据可以处理来源的异质性，并拥有多种商业智能和分析应用程序
- EX. Image, Website page

Understanding Different Types of File Formats 理解不同类型的文件格式

- Delimited text file (.csv)
 - CSV ,
 - TSV Tab
- 微软 Excel 打开 XML 电子表格 (.xlsx)
 - 在.XLSX (也称为工作簿)中，可以有多个工作表
 - 每个工作表被组织成行和列，在其交叉处为单元格
 - 每个单元格包含数据
- Markup Language (.xml)
 - XML 文件格式对于人类和机器都是可读的
 - 自我描述性语言，用于通过互联网发送信息
- Portable Document Format (.pdf)
 - Adobe 开发的一种用于呈现文档的文件格式
 - 在任何设备上以同样的方式显示
 - 经常用于法律和财务文件
- JavaScript Object Notation (JSON)
 - 基于文本的开放式传输标准
 - 通过网络传输结构化数据
 - 易于使用，兼容广泛的浏览器，并被认为是最好的共享工具（任何大小和类型的数据，甚至音频和视频）

Sources of Data 数据来源

- 关系数据库
 - 平面文件
 - CSV
 - XLSX
 - XML 数据集
 - 层次结构
 - 在线调查数据、银行对账单和其他数据非结构化数据
 - Api Web 服务
 - text, XML, HTML, JSON, or media files.
 - APIs
 - Twitter
 - Stock Market APIs
 - 网络信息搜集
 - 提取文本、联系信息、图片、视频
 - 数据流
 - 来自仪器、物联网设备和应用程序、来自汽车的 GPS 数据、计算机程序、网站和社交媒体帖子
 - Tools
 - apachekafka, ApacheSpark Streaming, and Apache Storm
 - 订阅 Feed
 - 通常用于从在线论坛和新闻站点获取更新的数据
-

Languages for Data Professionals

数据专业人员的语言

- 查询语言 访问和操作数据库中的数据
 - SQL
 - 快速有效地检索大量数据
 - 提供一个统一的平台

- 编程语言 开发应用程序和控制应用程序行为的编程语言
 - Python
 - 它的语法允许程序员用更少的代码行表达他们的概念
 - 被认为是最容易学习的语言之一，并拥有大量的开发人员社區
 - 注重简单性和可读性，而且学习曲线很低，所以它是一个理想产品
 - 适合在大量数据中执行高计算量的任务
 - 提供了像 Numpy 和 Pandas 这样的库，它通过使用并行方式简化了这项任务处理
 - R
 - 用于数据分析的开源编程语言和数据可视化
 - 用于开发统计软件和执行数据分析
 - 创造引人注目的可视化效果而闻名
 - 使用嵌入在其中的数据和脚本制作报告，也可以使用交互式 web 应用程序允许用户玩结果和数据
 - Java
 - 被用于数据分析的许多过程中，包括清理数据、导入和导出数据、统计分析和数据可视化
 - 用于大数据的流行框架和工具通常都是其编写
 - 适合于速度至关重要的项目
- shell 脚本 自动执行重复操作任务的 Shell 和脚本语言
 - Unix/linuxshell 和 PowerShell，非常适合重复使用
 - 它是用纯文本文件编写的一系列 UNIX 命令，用于完成特定的任务
 - 对于通过键入来执行可能比较耗时的重复性任务，它最为有用

Overview of Data Repositories

数据储存库概览

数据储存库是一个通用术语，用来指已经收集、组织、存储的数据

- 关系数据库，也称为 rdbms

- 织成表格格式，行和列跟在后面
- 定义良好的结构和模式
- SQL
- 非关系数据库，也称为 NoSQL
 - 为了响应数量、多样性和速度
 - 为了速度、灵活性和规模，非关系数据库使得数据存储成为可能
 - 无模式或自由格式

數據集市、數據湖、ETL 和數據管道

數據倉庫

將傳入的數據整合到一個綜合倉庫中

唯一的真實來源

存儲當前和歷史已清理、整合和分類的數據

數據集市

本質上是數據倉庫的子部分，用於隔離特定業務功能或用例的數據

數據倉庫的一個子部分

專門為特定業務構建 功能、目的或用戶社區

數據集市為數據倉庫的受限區域提供分析能力，它提供隔離的安全性和隔離的性能。數據集市最重要的作用是特定於業務的報告和分析

數據湖

以原生格式存儲大量結構化、半結構化和非結構化數據的存儲庫

一個存儲庫

存儲大量結構化、半結構化、和原始格式的非結構化數據，使用元數據進行分類和標記

每個數據元素都被賦予一個唯一標識符並用元標籤進行標記 供進一步使用

與數據倉庫不同，數據湖將保留所有源數據，沒有任何排除

數據湖有時也用作數據倉庫的暫存區

數據湖最重要的作用是在預測和高級分析中

大數據存儲

提供分佈式計算和存儲基礎設施來存儲、擴展和處理非常大的數據集

ETL

將原始數據轉換為分析就緒數據的方式

一個自動化過程，您可以在其中從已確定的來源收集原始數據，提取符合您的報告和分析需求的信息，清理、標準化數據並將其轉換為可在上下文中使用的格式

提取

從源位置收集數據以進行轉換的步驟。

數據提取可以通過：批處理，即源數據，以大塊的形式從源移動到以預定的時間間隔為目標系統。

批處理工具包括 Stitch 和 Blendo。

流處理

源數據是實時從源和在傳輸過程中和加載到數據存儲庫之前進行轉換。

流處理工具包括 Apache Samza、Apache Storm 和 Apache Kafka。

數據管道

一個高性能系統，支持長時間運行的批處理查詢和較小的交互式查詢。

數據管道的目的地通常是數據湖，儘管數據也可能是加載到不同的目標目的地，例如另一個應用程序或可視化工具。

加載

將處理後的數據傳輸到目標系統或數據存儲庫的步驟。

它可能是：

初始加載

填充存儲庫中的所有數據

增量加載

即根據需要定期應用持續更新和修改

完全刷新

即擦除一個或多個表的內容並重新加載新的數據

負載驗證

包括缺失值或空值的數據檢查、服務器性能、和監控負載故障，是此流程步驟的重要部分。

密切關注負載故障並確保正確的恢復機制是至關重要的

ETL 歷來被用於大規模的批處理工作負載。然而，隨著流式 ETL 工具的出現，它們越來越多地被用

大數據

5V

速度，數量、種類、真實性和價值

大數據處理工具

- Apache Hadoop
 - 蒐集分佈式存儲和處理大數據的工具
 - 允許分佈式存儲和處理 跨計算機集群的大型數據集
 - Hadoop 分佈式文件系統 **HDFS**
 - 運行在多個連接的商品硬件上的大數據存儲系統
 - 通過對多個文件進行分區來提供可擴展且可靠的大數據存儲 節點
 - 將大文件拆分到多台計算機上，允許並行訪問它們。因此，計算可以在存儲數據的每個節點上並行運行
 - 移動計算的過程 更靠近數據所在的節點
- Apache Hive
 - 建立在 Hadoop 之上的用於數據查詢和分析的數據倉庫
 - 用於讀取、寫入和管理大數據 設置直接存儲在 HDFS 或其他數據存儲系統中的文件
 - 用於長順序掃描，並且由於 Hive 基於 Hadoop，因此查詢 具有非常高的延遲
 - Hive 不適合用於涉及很高比例的寫操作
 - 更適合於 ETL、報告和數據分析等數據倉庫任務 並包括通過 SQL 輕鬆訪問數據的工具
- Apache Spark
 - 分佈式數據分析框架，旨在實時執行複雜的數據分析
 - 為各種應用程序處理大量數據，包括交互式 分析、流處理、機器學習、數據集成和 ETL

- 利用內存處理顯著提高計算速度 並且僅在內存受限時才溢出到磁盤
- 能夠快速處理流數據並實時執行複雜的分析

識別數據進行分析

方法

1. 確定您要收集的信息
 - a. 您需要的特定信息做出決定
 - b. 數據的可能來源
 - c. 取決於您尋求實現的目標
2. 定義提取所需數據的來源並定義收集數據的計劃
 - a. 為收集已確定的數據建立一個時間表
 - b. 定義足以達到的數據量以進行可信的分析
3. 確定您的數據收集方法
 - a. 定義如何從已確定的數據源收集數據
 - b. 方法將取決於
 - i. 數據類型
 - ii. 您需要數據的時間範圍
 - iii. 數據量
4. 實施您的數據
 - a. 收集策略並開始收集數據
 - b. 定義質量特徵、指標和檢查點，以確保 您的分析將基於質量數據
 - c. 注意與數據治理相關的問題，例如安全、監管、 和合規性
 - d. 數據隱私。

您收集的數據需要勾選保密性、使用許可和 遵守規定的法規。
 - e. 需要計劃檢查、驗證和可審計的跟踪。

數據源

- 主要

- 來自內部來源，例如來自 組織、CRM、HR 或工作流應用程序。它可以還包括您通過調查直接收集的數據，訪談、討論、觀察和焦點小組。
- 次要
 - 從 現有資源，例如外部數據庫、研究 文章、出版物、培訓材料和互聯網搜索，或作為公共數據提供的財務記錄。這也可以包括通過外部調查收集的數據，訪談、討論、觀察和焦點小組。
- 第三方來源
 - 從聚合商處購買的數據 從各種來源收集數據並將其組合成 全面的數據集純粹是為了銷售 數據。

如何收集和導入數據

關係數據庫 SQL

一些非關係型數據庫自帶查詢工具，例如 Cassandra 的 CQL

API 是從需要數據並訪問包含端點的應用程序調用的 數據

端點可以包括數據庫、Web 服務和數據市場。

網頁抓取

也稱為屏幕抓取或網頁收集，用於下載基於定義的參數來自網頁的特定數據

RSS 提要

從在線論壇獲取更新數據的來源 以及持續更新數據的新聞網站

數據流

聚合來自源的恆定數據流的流行來源

例如儀器、物聯網設備和應用程序以及來自汽車的 GPS 數據

總結

- 識別數據的過程首先確定需要收集的信息，而這又取決於您尋求實現的目標。
- 確定數據後，下一步是確定要從中提取所需數據的來源並定義數據收集計劃。關於您需要數據集的時間範圍的決定，以及有多少數據足以得出可信的分析，在這個階段也很重要。
- 數據源可以是組織內部的或外部的，它們可以是主要的、次要的或第三方的，具體取決於您是直接從原始來源獲取數據、從外部可用數據源檢索數據還是從數據聚合器。

- 您可以從中收集數據的一些數據源包括數據庫、網絡、社交媒體、互動平台、傳感器設備、數據交換、調查和觀察研究。
- 從各種數據源中識別和收集的數據使用各種工具和方法進行組合，以提供可以查詢和操作數據的單一界面。
- 您識別的數據、該數據的來源以及您用於收集數據的做法會對質量、安全性和隱私產生影響，這些都需要在此階段加以考慮。

數據整理

What is Data Wrangling? 什麼是數據整理？

涉及數據的迭代過程 探索、轉化、驗證，並使其可用於可信和有意義的分析

- 發現
 - 更好地理解您的數據
 - 具體弄清楚如何最好地清潔、構建、組織、並為您的用例映射數據
- 轉換

為轉換數據而承擔的任務

例如結構化、規範化、對數據進行反規範化、清理和豐富

- 結構化
 - 更改數據形式和架構的操作
 - 為了合併多元數據，您需要更改數據的形式或架構
- 數據的規範化和非規範化
 - 範化側重於清理數據庫中未使用的數據並減少冗餘和 不一致
 - 反規範化用於將來自多個表的數據合併到一個表中，以便 它可以更快地查詢
- 清潔
 - 修復數據中的違規行為以產生可信和 準確的分析
 - 不準確、缺失或不完整的數據可能會扭曲您的分析結果和 需要考慮
 - 分析數據以發現質量問題，可視化數據以發現異常值，以及修復缺失值、重複數據、不相關數據、格式不一致、語法錯誤和異常值等問題

- 豐富數據
 - 在信息分散在各個系統的大型組織中，您 可能需要用其他系統提供的信息來豐富一個系統提供的數據集 系統，甚至公共數據集
- 驗證
 - 驗證一致性、質量、 以及您擁有的數據的安全性
- 發布
 - 為下游項目提供整理數據的輸出

數據整理工具

- Excel Power Query / Spreadsheets
 - 具有許多功能和內置 可以幫助您識別問題、清理和轉換數據的公式
 - 插件允許您從多種不同類型的來源導入數據並根據需要清理和轉換數據
- OpenRefine
 - 允許您導入和導出各種數據 格式，例如 TSV、CSV、XLS、XML 和 JSON
 - 清理數據，將其從一種格式轉換為另一種格式，並擴展 數據與 Web 服務和外部數據
 - 基於菜單的操作，這意味著您無需記住命令或語法
- Google DataPrep
 - 智能雲數據服務，可讓您直觀地探索、 清理並準備結構化和非結構化數據以進行分析
 - 完全託管的服務，這意味著您無需安裝或管理 軟件或基礎設施
 - 可以自動檢測模式、數據類型和異常
- Watson Studio Refinery
 - 允許您發現、清理、 並使用內置操作轉換數據。
 - 將大量原始數據轉換為可用的、準備好的質量信息 用於分析
 - 自動檢測數據類型和分類，並強制執行適用的 自動數據治理策略
- Trifacta Wrangler
 - 基於雲的交互式服務，用於清理和轉換數據。 它獲取雜亂的真實數據，並將其清理並重新排列到數據表

- 導出到 Excel、Tableau 和 R
- 以協作功能而聞名
- Python
 - 有一個龐大的庫和一組包，可提供強大的數據操作功能
 - Jupyter Notebook 是一個廣泛用於數據清理和轉換的開源 Web 應用程序，統計建模，還有數據可視化
 - Numpy 或 Numerical Python，是 Python 提供的最基本的包。它快速、通用、可互操作且易於使用。它提供對大型、多維數組和矩陣以及高級數學的支持 對這些數組進行操作的函數。
 - Pandas 專為快速簡便的數據分析操作而設計。它允許進行複雜的操作，例如合併、連接和轉換大量數據，使用簡單的單行命令執行。
- R
 - 提供了一系列為整理而明確創建的庫和包
 - Dplyr 強大的數據整理庫。它具有精確而直接的語法
 - Data.table 有助於快速聚合大型數據集
 - Jsonlite 強大的 JSON 解析工具，非常適合與 Web API 交互

數據清洗

- 檢查
 - 使用允許您定義特定規則和約束的腳本和工具以及 根據這些規則和約束驗證您的數據
 - 使用數據分析和數據可視化工具進行檢查
- 清理
 - 缺失值
 - 過濾掉具有缺失值的記錄或找到一種方法來獲取該記錄 信息
 - 重複數據
 - 即數據集中重複的數據點。這些都需要移除
 - 不相關的數據
 - 不適合您的用例上下文的數據可以被認為是不相關的 數據
 - 語法錯誤

- 字符串開頭或結尾的空格或額外空格是一種需要糾正的錯誤。這還可以包括修正拼寫錯誤或格式
- 清理也可能涉及數據類型轉換
- 清理數據以使其標準化。
 - 對於字符串，您可能希望所有值都為小寫。
 - 日期格式和測量單位也需要標準化。
- 驗證
 - 檢查結果以確定結果的有效性和準確性
 - 在您進行更正後重新檢查數據以確保規則和適用於約束

統計分析

統計學是數學的一個分支，處理數值或定量數據的收集、分析、解釋和表示。

統計分析涉及使用統計方法，以了解數據代表的含義。

統計分析可以是：

- 描述性的；提供數據所代表內容的摘要。常見的度量包括中心趨勢、分散和偏度。
- 推論；涉及對數據進行推斷或概括的內容。常見的度量包括假設檢驗、置信區間和回歸分析。

簡單地說，數據挖掘就是從數據中提取知識的過程。它涉及使用模式識別技術、統計分析和數學技術，以識別數據中的相關性、模式、變化和趨勢。

有幾種技術可以幫助挖掘數據

數據挖掘

- 聚類類似於分類，但涉及 將數據分組到集群中，以便將它們視為組
- 關聯規則挖掘是一種有助於建立 我們兩個數據事件之間的關係
- 親和分組是一種用於發現 Co 發生的技術 在關係中
 - 交叉銷售和向上銷售他們的產品 根據其他人的購買歷史向人們提供產品 誰購買了相同的物品。
- 決策樹有助於構建分類模型 具有多個分支的樹結構形式，其中每個 分支代表可能發生的事件。這種技術有助於建立對之間關係的清晰理解 輸入和輸出。

- 回歸是一種技術，可以幫助識別 兩個變量之間的關係，這可能是因果關係 或相關

數據挖掘工具

- Microsoft Excel 和Google 表格通常用於執行基本的數據挖掘任務
 - 電子表格可以用於以易於訪問的方式託管從其他系統導出的數據 和易於閱讀的格式
- R
 - 打包了數百個庫專為數據挖掘操作而構建，例如回歸、分類、數據聚類、關聯規則挖掘、文本挖掘、異常值檢測和社交網絡分析
- SPSS
 - 有一個易於使用的界面需要最少的代碼來完成複雜的任務。它包括高效的數據管理工具，因其深入的分析能力和準確的數據結果而廣受歡迎
- IBM Watson Studio
 - 包含在 IBM Cloud Pak for Data 中，利用一系列開放的 源工具，例如 Jupyter notebook，並使用閉源 IBM 工具對其進行擴展 這使其成為數據分析和數據科學的強大環境。可以使用通過公共雲、私有云上的 Web 瀏覽器以及作為桌面應用程序。Watson Studio 使團隊成員能夠在項目上進行協作
 - 使您能夠為您的業務數據快速開發預測模型。

交流和共享數據分析結果

1. 誰是我的聽眾？
什麼對他們很重要？
什麼會幫助他們信任我？
2. 圍繞您的聽眾的信息水平進行構建。
根據您對受眾的了解，您將決定信息的內容和數量，對於更好地理解您的發現至關重要。
3. 數據可視化——圖形、圖表、圖表——是讓數據栩栩如生的好方法。
 - 確保您的聽眾能夠信任您、理解您並與您的發現和見解相關聯。
 - 建立您的發現的可信度。
 - 在結構化的敘述中呈現數據。

- 通過強大的可視化支持您的溝通，使信息清晰簡潔，並促使您的受眾採取行動。

數據可視化

數據可視化是通過使用圖形、圖表和地圖等視覺元素來傳達信息的學科。可視化數據的目標是使信息易於理解、解釋和保留。

要使數據可視化有價值，您需要：

- 想想你的聽眾的關鍵要點。
- 預測他們的信息需求和問題，然後規劃可視化，以清晰而有影響力地傳達您的信息。

有多種類型的圖形和圖表可供您繪製任何類型的數據，例如條形圖、柱形圖、餅圖和折線圖。

數據可視化軟件和工具

- 電子表格
 - 例如 Microsoft Excel 和 Google Sheets，可能是最常用的 使用軟件來製作數據集的圖形表示。
 - 電子表格易於學習，並提供大量文檔和視頻教程 網上準備參考。
 - Excel 提供了多種圖表類型，
 - 基本的條形圖、折線圖、餅圖和數據透視圖，
 - 更高級的選項，如散點圖、趨勢線、甘特圖、瀑布圖 圖表和組合圖表（使用它可以組合多種類型的圖表）。
 - Excel 還提供有關數據集的最佳可視化表示的建議
- Jupyter Notebook
 - Matplotlib
 - 是一個廣泛使用的 Python 數據可視化庫。它提供了不同種類的 2D 和 3D 繪圖，並且可以靈活地以多種方式創建繪圖。不同的方式。
 - 使用 Matplotlib，您可以創建高質量的交互式圖形和繪圖
 - Bokeh
 - 提供交互式圖表和繪圖，以提供高性能而聞名 大型或流數據集上的交互性。
 - Bokeh 為應用交互、佈局和不同的樣式選項提供了靈活性 可視化。

- 轉換用其他一些 Python 庫編寫的可視化，例如如 Matplotlib、Seaborn 和 Ggplot。
- Dash
 - 用於創建基於 Web 的交互式可視化
 - Dash 易於維護、跨平台和移動就緒。
- R-Studio
 - 您可以創建基本的可視化效果，例如直方圖、條形圖、折線圖、箱線圖和散點圖；和高級可視化，如熱圖、馬賽克圖、3D 圖形和相關圖。
 - Shiny 是一個 R 包
 - 構建可作為獨立應用程序託管的交互式 Web 應用程序
- IBM Cognos Analytics 是一種端到端分析解決方案
 - 可提供基於數據的時間序列數據建模和預測 在相應的可視化中呈現
 - 於可視化的建議您的數據；條件格式，可讓您查看數據的分佈和突出顯示異常數據點
- Tableau
 - 創建儀錶盤形式的交互式圖形和圖表 和工作表，帶有拖放手勢
 - 導入 R 和 Python 腳本並利用其可視化功能
 - 兼容 Excel 文件、文本文件、關係數據庫和雲數據庫 來源，例如 Google Analytics 和 Amazon Redshift。
- Power BI
 - 創建報告和儀表板
 - 與多種來源兼容，包括 Excel、SQL Server 和基於雲的 數據存儲庫

數據分析的職業機會

數據分析師的工作角色可以大致分類如下：

- 數據分析師專家角色 - 在這條道路上，您從初級數據分析師開始，並通過不斷將您的技術、統計和分析技能從基礎級別提升到專家級別，升至首席分析師級別。
- 領域專家角色 - 如果您已獲得特定領域的專業知識並希望通過自己的努力被視為您所在領域的權威，則這些角色適合您。

- 支持分析的工作角色 - 這些角色包括具有分析技能可以提升您的績效並使您與同齡人區分開來的工作。
- 其他數據職業——現代數據生態系統中還有其他幾個角色，例如數據工程師、大數據工程師、數據科學家、業務分析師或商業智能分析師。如果您根據所需技能提升自己的技能，則可以過渡到這些角色。