如何成為資料分析師：從問題解決到行動方案

https://hahow.in/cr/dajourney

數據交點 | 郭耀仁 yaojenkuo@datainpoint.com

拆解「如何成為資料分析師」問題的三階段

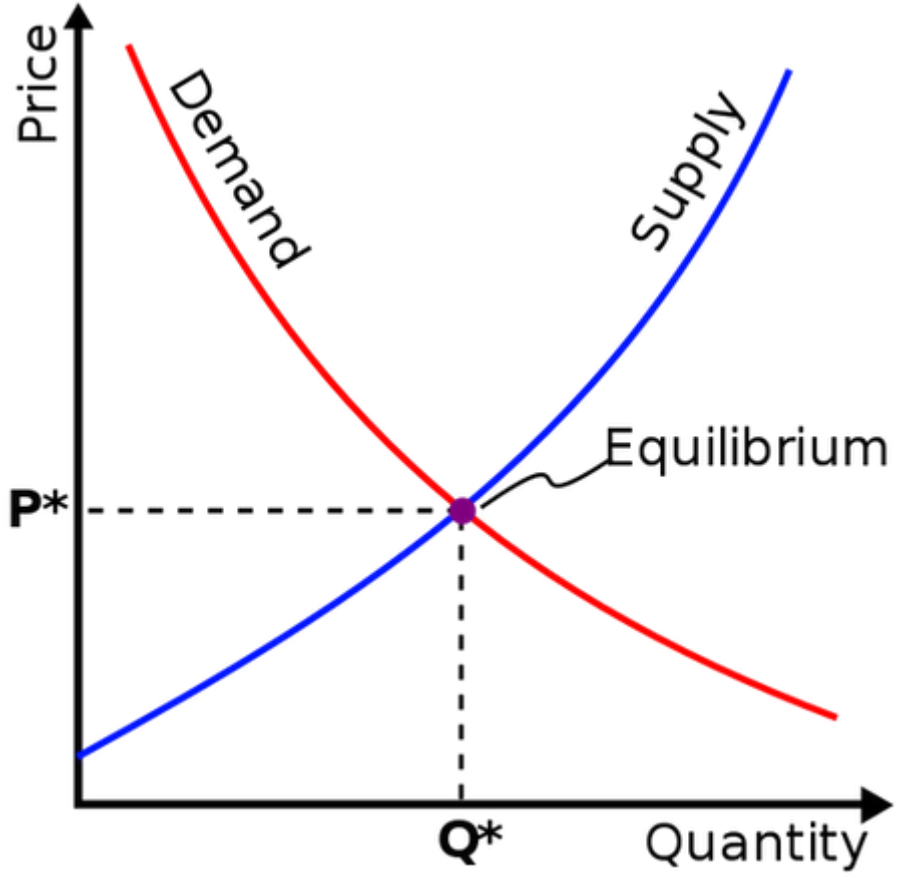
1. 「資料分析師」的部分
2. 驗證
3. 「如何成為」的部分

「資料分析師」的部分

什麼是資料分析師

能夠透過以邏輯、事實以及量化指標為基準來解決問題的人士。

技能供給與職缺需求的交點



來源：Google Search

技能供給與職缺需求

蒐集資料分析師的技能供給

Competition

2021 Kaggle Machine Learning & Data Science Survey

Analytics

in a month • $30,000

2021 Kaggle Machine Learning & Data Science Survey

Competition

2019 Kaggle Machine Learning & Data Science Survey

Analytics

2 years ago • $30,000

2019 Kaggle Machine Learning & Data Science Survey

Competition

2020 Kaggle Machine Learning & Data Science Survey

Analytics

9 months ago • $30,000

2020 Kaggle Machine Learning & Data Science Survey

資料分析師的技能供給

- 主要工作內容
- 使用、推薦程式語言
- 整合開發環境
- 視覺化套件

資料分析師的技能供給（續）

- 商業智慧軟體
- 機器學習框架
- 關聯式資料庫管理系統

In [2]:
```
interested_questions
```

Out[2]:

	question_index	question_desc	question_type
6	Q7	What programming languages do you use on a reg...	multiple selection
7	Q8	What programming language would you recommend ...	multiple choice
8	Q9	Which of the following integrated development ...	multiple selection
13	Q14	What data visualization libraries or tools do ...	multiple selection
15	Q16	Which of the following machine learning framew...	multiple selection
22	Q23	Select any activities that make up an importan...	multiple selection
28	Q29A	Which of the following big data products (rela...	multiple selection
30	Q31A	Which of the following business intelligence t...	multiple selection

主要工作內容的問卷問題

Select any activities that make up an important part of your role at work: (Select all that apply)

使用、推薦程式語言的問卷問題

- What programming languages do you use on a regular basis? (Select all that apply)
- What programming language would you recommend an aspiring data scientist to learn first?

整合開發環境的問卷問題

Which of the following integrated development environments (IDE's) do you use on a regular basis? (Select all that apply)

視覺化套件的問卷問題

What data visualization libraries or tools do you use on a regular basis? (Select all that apply)

商業智慧軟體的問卷問題

Which of the following business intelligence tools do you use on a regular basis? (Select all that apply)

機器學習框架的問卷問題

Which of the following machine learning frameworks do you use on a regular basis? (Select all that apply)

關聯式資料庫管理系統的問卷問題

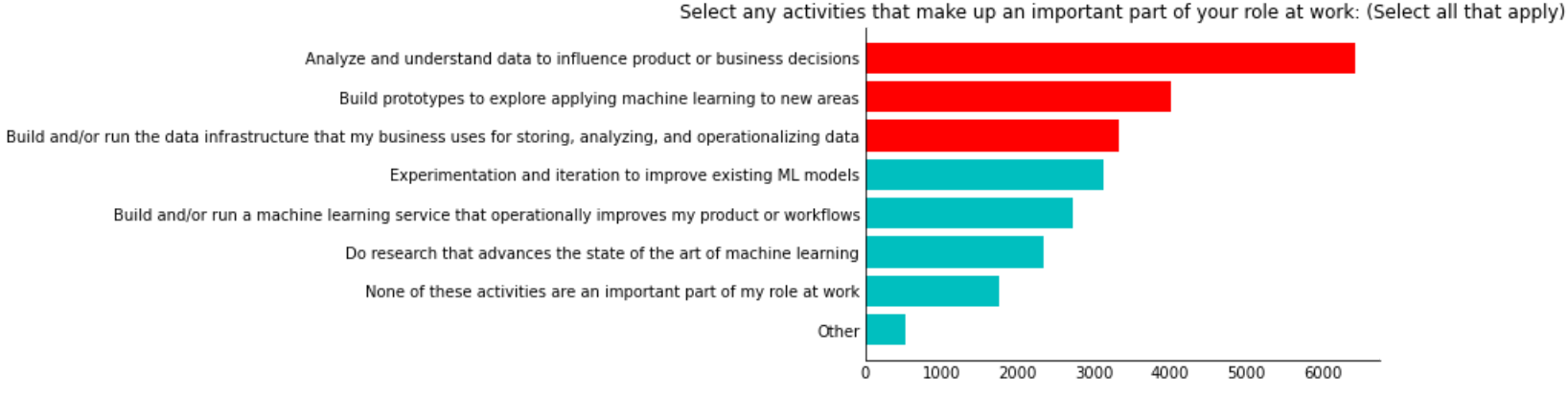
Which of the following big data products (relational databases, data warehouses, data lakes, or similar) do you use on a regular basis? (Select all that apply)

資料分析師的主要工作內容

分析與瞭解資料進而影響產品或商業上的決策。

In [3]:
```
ks.plot_summary("Q23")
```

```
Select any activities that make up an important part of your role at work:
(Select all that apply)
```



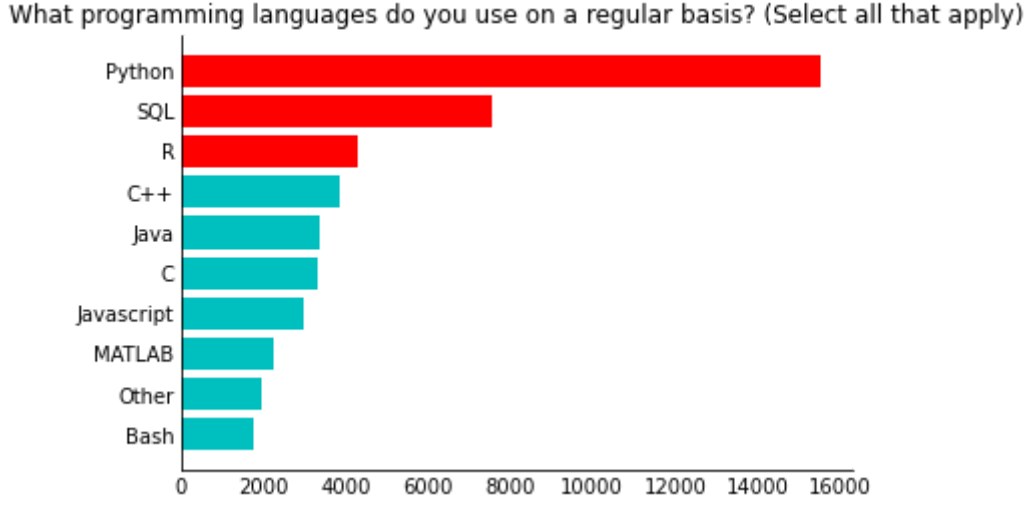
資料分析師使用、推薦的程式語言

主要使用 Python、SQL 與 R，並大力推薦初學者由 Python 入門。

In [4]:

```
ks.plot_summary("Q7")
```

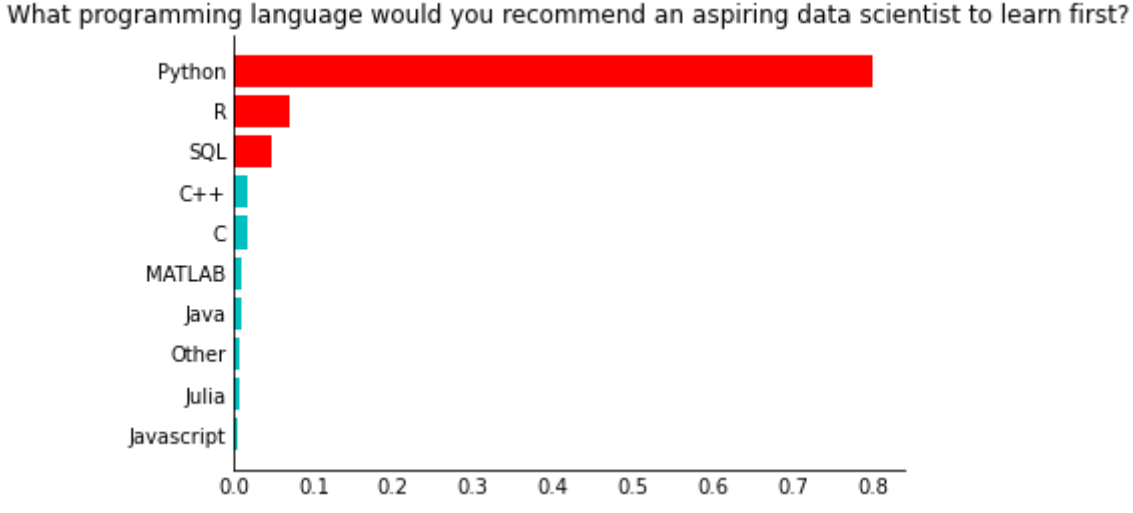
```
What programming languages do you use on a regular basis? (Select all that
apply)
Too many categories, only showing the top 10.
```



In [5]:

```
ks.plot_summary("Q8")
```

```
What programming language would you recommend an aspiring data scientist t
o learn first?
Too many categories, only showing the top 10.
```

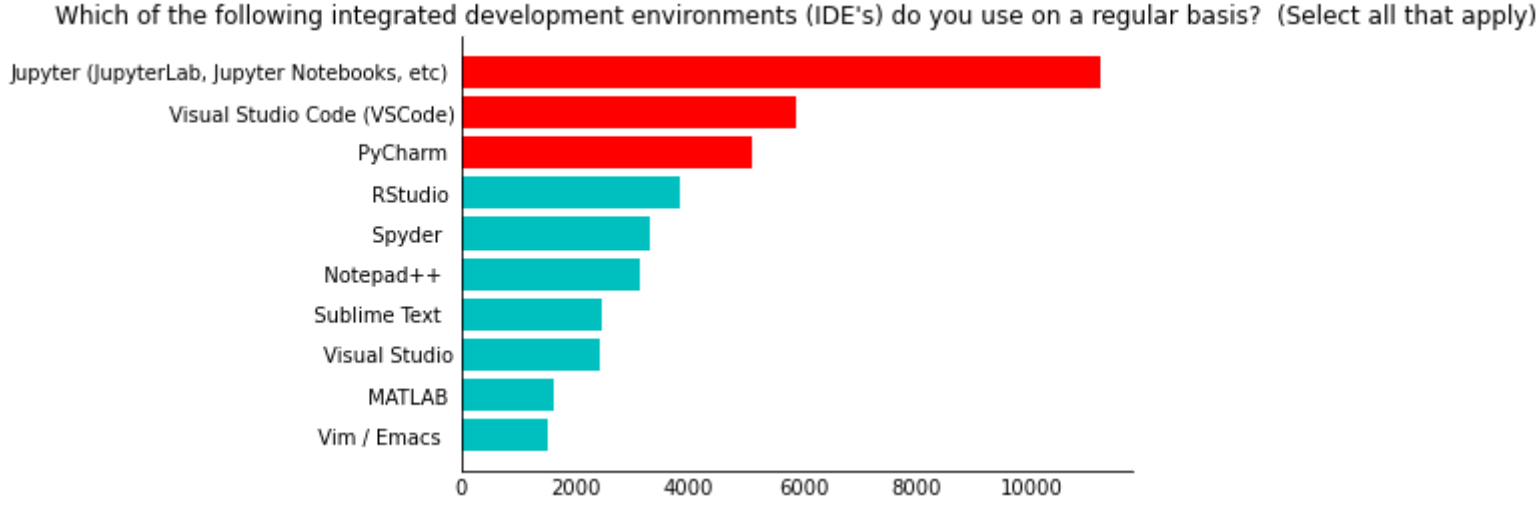


資料分析師的整合開發環境

以 Jupyter、VS Code 與 PyCharm 撰寫 Python、以 RStudio 撰寫 R、以文字編輯器寫作其他語言。

In [6]:
```
ks.plot_summary("Q9")
```

```
Which of the following integrated development environments (IDE's) do you
use on a regular basis?  (Select all that apply)
Too many categories, only showing the top 10.
```

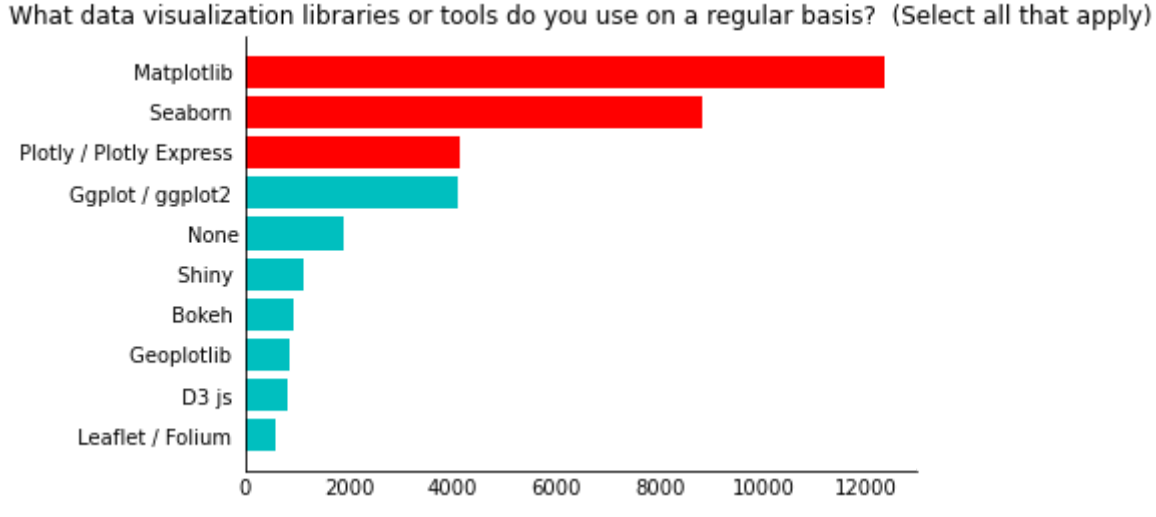


資料分析師的視覺化套件

主要使用的視覺化套件是在 Python 使用 matplotlib、seaborn 與 plotly，在 R 使用 ggplot2 與 plotly。

In [7]:
```
ks.plot_summary("Q14")
```

```
What data visualization libraries or tools do you use on a regular basis?
(Select all that apply)
Too many categories, only showing the top 10.
```

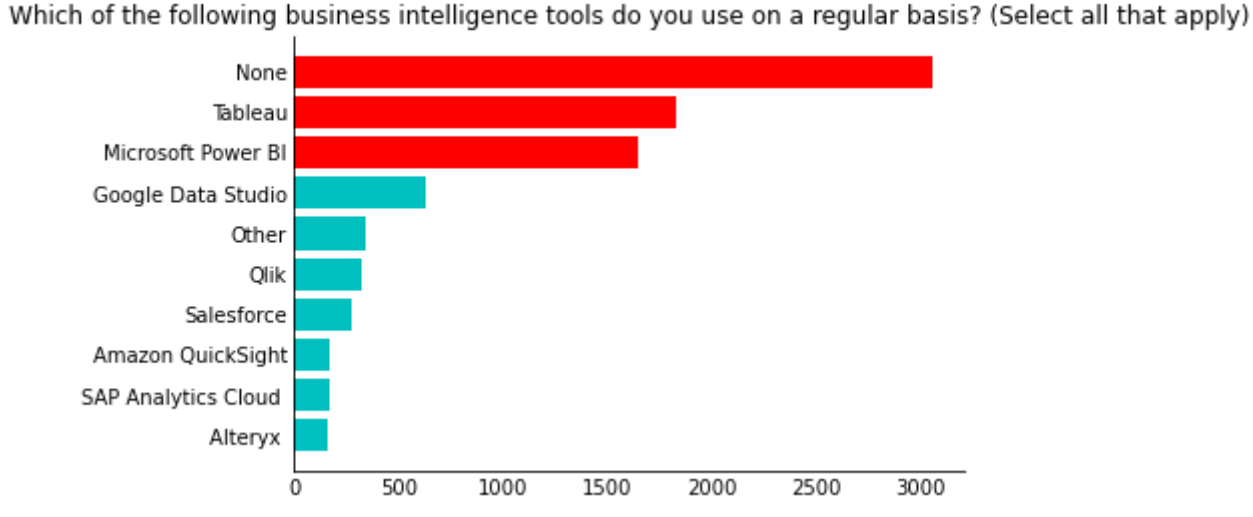


資料分析師的商業智能軟體

主要使用的商業智能軟體是 Tableau 與 PowerBI。

In [8]:
```
ks.plot_summary('Q31A')
```

```
Which of the following business intelligence tools do you use on a regular
basis? (Select all that apply)
Too many categories, only showing the top 10.
```

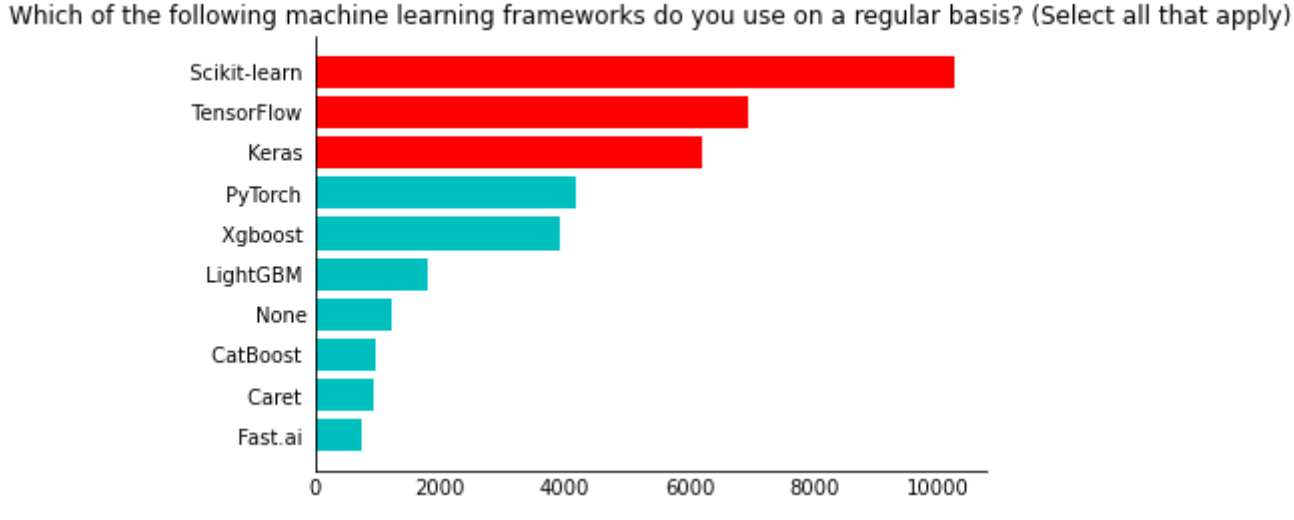


資料分析師的機器學習框架

主要使用的機器學習框架是 Scikit-Learn 以及深度學習 TensorFlow / Keras。

In [9]:
```
ks.plot_summary('Q16')
```

```
Which of the following machine learning frameworks do you use on a regular
basis? (Select all that apply)
Too many categories, only showing the top 10.
```

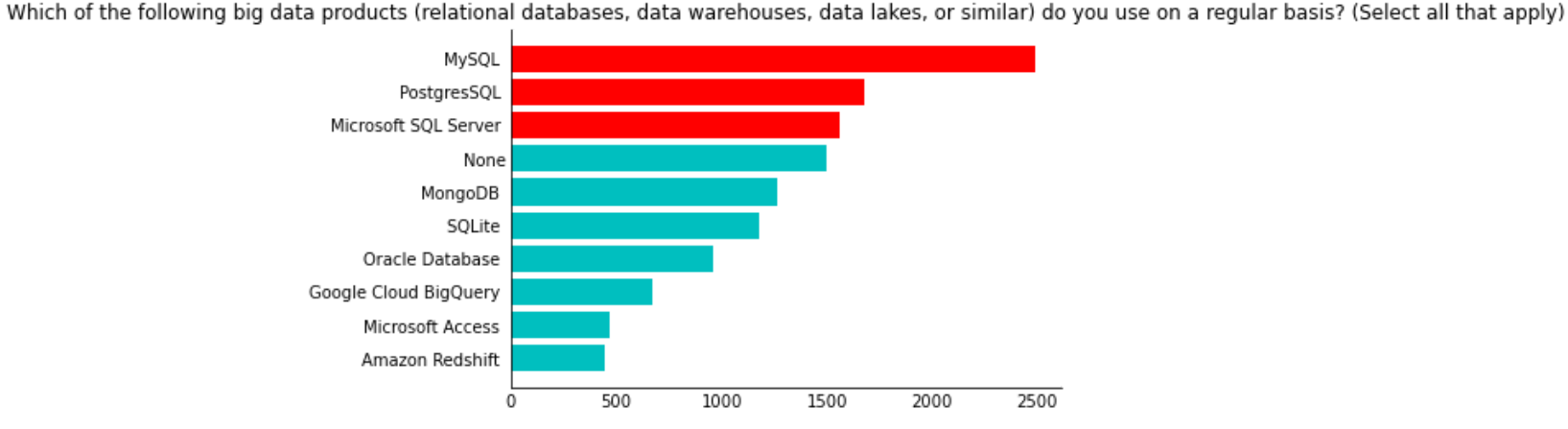


資料分析師的關聯式資料庫管理系統

主要使用的關聯式資料庫為 MySQL、PostgreSQL 與 MS SQL Server。

In [10]:
```
ks.plot_summary('Q29A')
```

```
Which of the following big data products (relational databases, data wareh
ouses, data lakes, or similar) do you use on a regular basis? (Select all
that apply)
Too many categories, only showing the top 10.
```



蒐集資料分析師的職缺需求

- Indeed
- CakeResume
- 104 人力銀行

在瀏覽器啟動分析環境

- kaggle-ml-ds-survey-2020.ipynb
- indeed-cakeresume-104.ipynb

驗證

資料科學團隊面試官訪談

- 資料分析師/資料科學/約會交友 APP
- 資深講師/教育訓練/外商軟體公司
- 副理/信用模型/銀行
- 資料工程師/資料科學/外送 APP

資料科學團隊面試官訪談（續）

- 經理/資料科學/電信公司
- 經理/資料科學/保險公司
- 副理/整合行銷/銀行
- 副理/商業智能/電商

訪談 8 位面試官三個問題

- 您希望初階資料分析師分別在程式語言、統計、機器學習、視覺化與資料庫五個面向所具備的程度為何？若以 1 非常陌生、5 為非常熟稔，您的期待是?
- 在進行初階資料分析師的面試時，請和我們分享您通常會問的三個「技術問題」。
- 在進行初階資料分析師的面試時，請和我們分享您通常會問的三個「非技術問題」。

「如何成為」的部分

目標、現況與學習地圖

- 目標描述
- 現況評估
- 學習地圖

第零站：跟電腦變成好朋友

寫程式是理解統計知識與機器學習理論不可或缺的工具與手段，因此學習的起點必須從程式設計起步，而在開始寫程式之前我們應該跟電腦變成好朋友！學習操作終端機、文字編輯器還有 Git/GitHub。

第一站：使用 SQL 查詢資料

從課程中所分析的 Kaggle 問卷調查我們得知資料分析師使用最多、最推薦的語言是 Python > SQL > R，我會推薦從 SQL 開始，原因是 SQL 作為由資料庫中查詢資料的語言，與人類語言的相似度高、語法單純並且是即戰力的工具，CP 值很高。

第二站：使用視覺化軟體作探索性分析

懂得如何使用 SQL 之後，可以跟試算表、視覺化軟體搭配開始進行探索性分析，從課程中所分析的 Kaggle 問卷調查我們得知 Tableau 與 PowerBI 是現在最普遍的視覺化軟體，而這兩個都有免費的版本可供個人電腦使用，如果是使用 Windows 作業系統的學員可以都試用，使用 macOS 的學員就使用 Tableau。

第三站：學習 Python 程式設計

想要勝任工程導向的資料科學團隊職缺，必須學會一個在專案的各個應用都能介接、並且能夠自己整併清理資料的程式語言，在 2021 年首選是 Python 程式語言。

第四站：機器學習應用

先使用第三方套件實作 Kaggle 專案，至於為何所引用的函式與類別能夠完成機器學習應用，則留待後續在研讀理論後融會貫通。

第六站：統計與機器學習理論

從課程中面試官訪談的部分我們得知，在技術面試時除了會測驗實作，也會詢問第三方套件中相關函式、類別的參數設定以及其背後的理論對應。