

Data Scientist Jobs

1.資料合併

(程式碼為 `./src/data_manipulation.py`)

(1) 初始合併檔案位於 `./data/analysis/combined.csv`

2.Feature Engineering

(程式碼為 `./src/feature_engineering.py`)

此專案欲使用這些資料集，進行機器學習模型的建構，依照求職者本身的薪資期望、能力、學歷等資訊，進行分類讓求職者依照自身能力找出適合自己目前狀況的工作。因為對工作能力大部分的要求都在JD之中，因此本專案先採用Regex擷取JD之中的資訊，並整理為不同的特徵值以利後續分析，整理的特徵值如下：

- (1) 工作類別：加上工作類別並將薪資、公司規模拆分為數值特徵以及是否為Senior職缺轉換為特徵值。
- (2) 公司類別：加上是否為近10年來所成立的較年輕公司和公司類別是否為FAANG(Facebook/Apple/Amazon/Netflix/Google)為特徵值
- (3) 程式能力：加上是否要求Python, SQL, R, Java等特徵值
- (4) 其他能力：加上OOP, ML, Modeling, Database, Tableau, Power BI, ETL等特徵
- (5) 學歷：加上碩士要求和博士要求等特徵

整理完成的分析資料位於 `./data/analysis/analysis.csv`

3.EDA & 開放式分析

(程式碼為 `./src/EDA.py`)

本專案想要探討的重點在於職位和薪資與技能、學歷、產業、公司之間的關聯性，並以此進行模型的建構，以資料分析/資料科學相關行業求職者本身的學經歷和技能以及期望薪資進行適合的職位類別推薦。

(1)缺失值

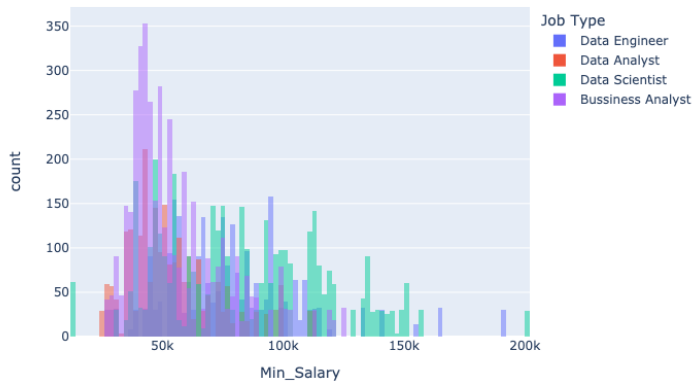
#	Column	Non-Null Count
0	Job Title	12782 non-null
1	Salary Estimate	12781 non-null
2	Job Description	12782 non-null
3	Rating	11829 non-null
4	Company Name	12781 non-null
5	Location	12782 non-null
6	Headquarters	12026 non-null
7	Size	12061 non-null
8	Founded	9508 non-null
9	Type of ownership	12061 non-null
10	Industry	10981 non-null
11	Sector	10984 non-null
12	Revenue	12061 non-null
13	Competitors	3554 non-null
14	Easy Apply	524 non-null
15	Job Type	12782 non-null

由完整初始資料的缺失值中可以發現，競爭者(Competitors)欄位中具有缺失值將近高於全體資料的1/4，且缺失值較難採用統計學或是機器學習方法進行補值，且若直接刪除缺失資料則資料筆數會過少，因此判定直接捨去此欄位。

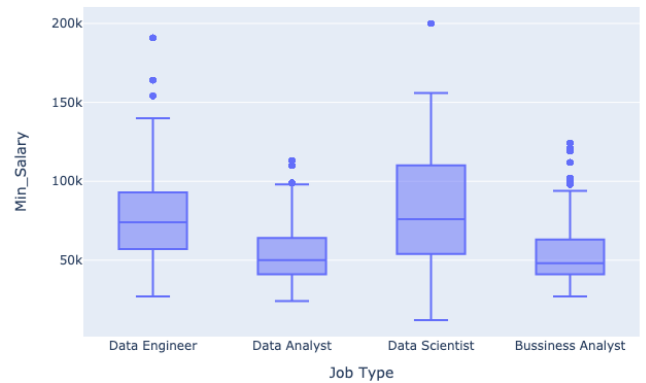
而Easy Apply 欄位中，由於資料集未說明-1的值代表缺失或是false，且僅有524筆具有True值，因此判定直接捨去此欄位。

(2)工作類別與薪資上下限

Histogram of Min Salary

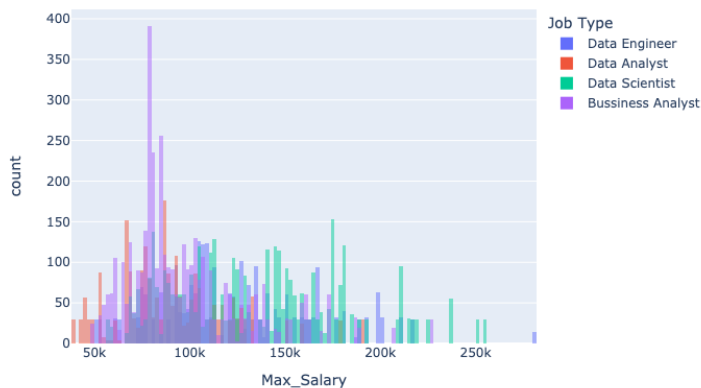


Box Plot of Min Salary

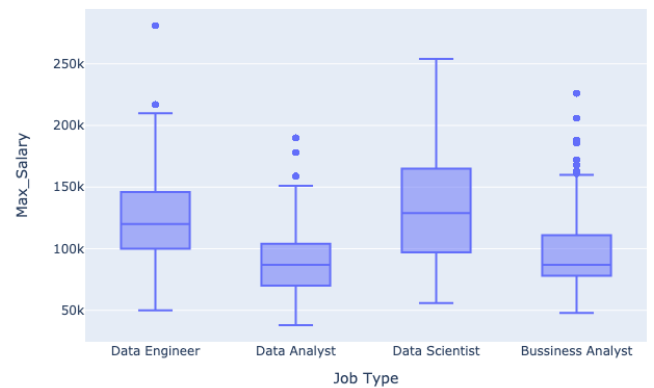


首先由薪資下限的箱型圖及長條圖可以看出，資料科學家及資料工程師相對於資料分析師及商業分析師來說，薪資下限的中位數較高，但資料科學家的薪資下限範圍最廣，而商業分析師及資料分析師的薪資下限範圍相對集中且具有較小的四分位距，可以推斷薪資下限的議價範圍較小且薪資下限集中在較低的薪資水準。

Histogram of Max Salary



Box Plot of Max Salary

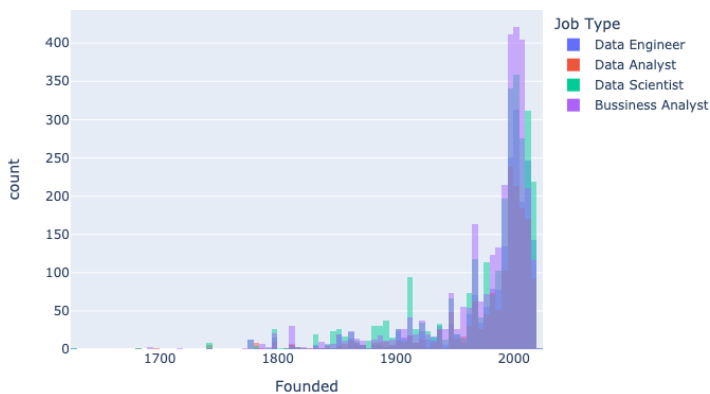


其次，由薪資上限的箱型圖及長條圖可以看出資料工程師及資料科學家及資料工程師在薪資上限的中位數較資料分析師及商業分析師高，且資料科學家的薪資上限範圍也最廣，且分佈較為平均。而商業分析師和資料分析師的薪資上限四分位距仍較窄，可推斷薪資上限的議價範圍亦較小且薪資上限仍集中於較低的薪資水準。但以薪資上限來說資料分析師薪資上限的最小值相較於商業分析師更低，且相更集中於薪資水準較低的區域。

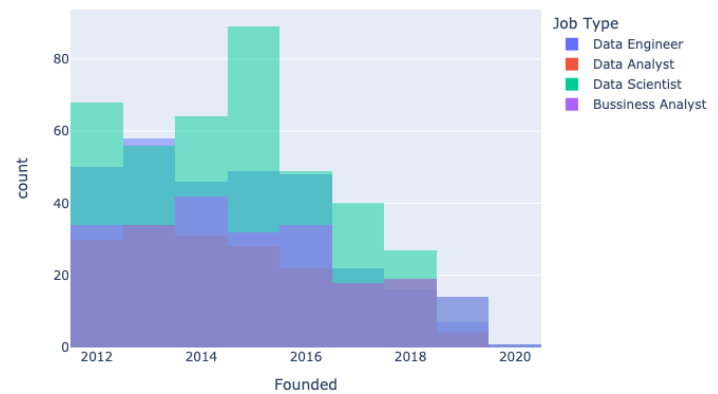
綜上所述可以初步推斷，在不同的工作類別中，薪資下限與上限具有不同的代表性，在後續建立模型時也將薪資上下限作為特徵值。

(3) 工作類別與公司成立時間

Histogram of new Founded



Histogram of Founded



上圖左圖為公司創立時間與工作類別的長條圖，而右圖為近十年所創立的公司與職業類別的長條圖，可以發現，在2000年以後所創立的公司，對於商業分析師和資料分析師的需求達到高峰，但近年來卻逐步下降。而對近十年來所成立的公司而言，資料科學家的需求幾乎都是四種工作類別中最大的，可以推斷出以近年來說，資料科學家對於較晚創立的公司相對更加重要，對於想要進入較年輕公司的求職者來說，可能可以將自己在職涯和技能上的規劃朝資料科學家的方向準備，且不僅在較年輕的公司，對於創立時間較久的成熟企業，資料科學家的需求也大部分相較其他三者多，由此可推斷或許資料科學家所涵蓋的能力較為大多公司所看重，後續建立模型時也將公司成立時間10年作為新舊公司的分水嶺，並依此建立特徵值。

(4) 公司規模與薪資下限

Scatter Plot of size and min salary of Data Engineer



Scatter Plot of size and min salary of Data Scientist



Scatter Plot of size and min salary of Data Analyst

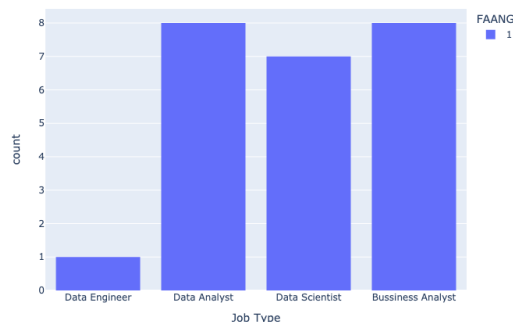
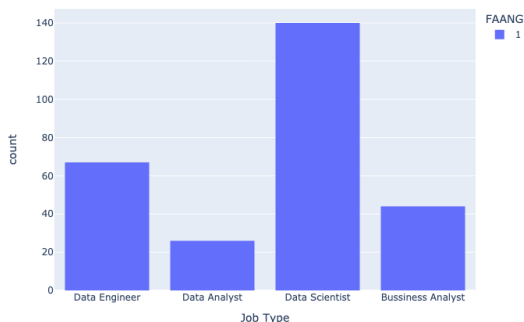


Scatter Plot of size and min salary of Bussiness Analyst



由四種職業的薪資下限與公司規模的散佈圖可以發現，不論公司規模大小，薪資下限的分布並沒有太大的差異，以此和相關係數進一步推斷公司的規模對不同職位薪資下限並沒有太大的影響，不同的薪資上下限在規模不同的公司中都各有分布，並不隨公司規模擴大小而有明顯差異。

(5)工作類別與公司名稱

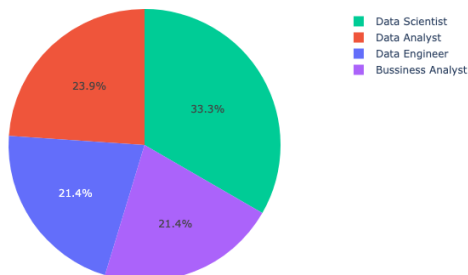


本專案想嘗試除了產業以外的公司特徵，由於近年來FAANG等公司的快速成長，對於求職者來說，這幾間公司所提供的職位也成了許多求職者心目中的Dream Job，因此針對這些公司所開出的職位進行分析，由以上的兩張長條圖可以發現，對於一般階層的職缺（左圖）來說，這幾家公司所開出的最多職位是資料科學家，而最少的是資料分析師，但以資深的職缺來說（右圖）反而是資料分析師和商業分析師的職缺相對多，或許是因為公司營運已經日漸成熟，在系統方面已經有穩定的基礎，所以在Senior的方面需要更多能夠第一線了解產品的職缺。

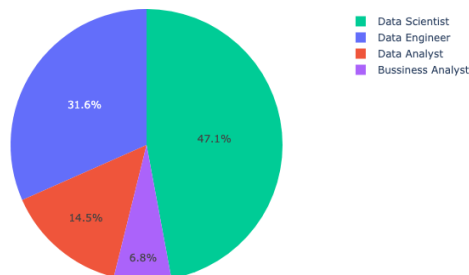
(6)工作類別與能力

[1] 語言要求

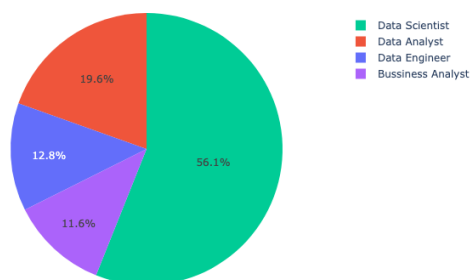
Pie chart of Job Type and SQL



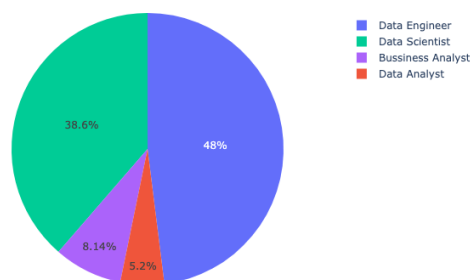
Pie chart of Job Type and Python



Pie chart of Job Type and R



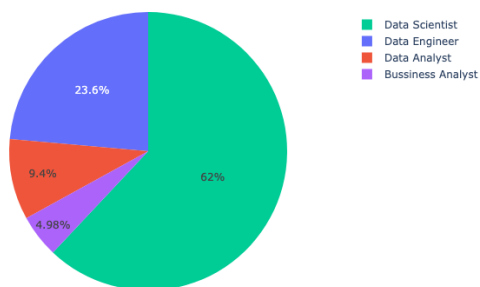
Pie chart of Job Type and Java



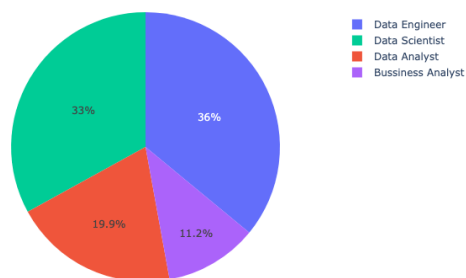
首先由使用的語言能力對應不同的工作類別製作圓餅圖，可以發現工作中要求會Python的人中，資料科學家為最多，其次為資料工程師、資料分析師、商業分析師。工作中要求會使用Java的人中，最多的是資料工程師其次為資料科學家、商業分析師、資料分析師。以上述兩點此可以推斷若求職者欲尋找資料科學家或資料工程師的工作，可能可以著重於Python或是Java的學習，而要求會使用SQL的工作中，每項工作類別的比例並沒有差太多，由此可以推斷四種工作都需要會使用SQL，可以將SQL判斷為較通用的技能，而除了資料庫語言SQL之外的三項語言，資料分析師和商業分析師兩職業對程式能力的要求都較低，這或許是造成薪資上下限差距的原因，在後續模型建立的過程中，也將程式能力作為特徵值納入模型之中。

[2] 其他能力要求

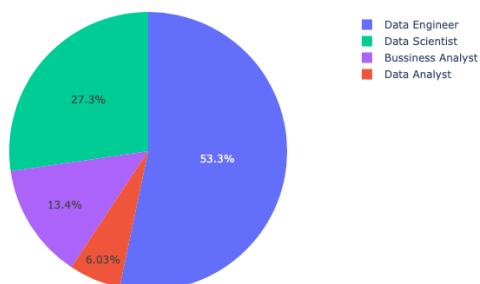
Pie chart of Job Type and ML



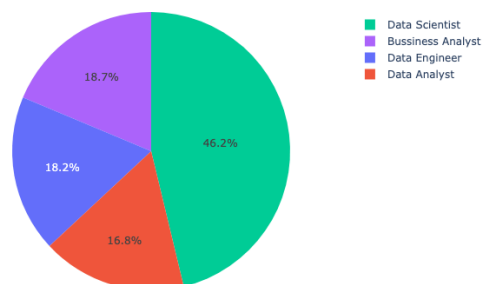
Pie chart of Job Type and ETL



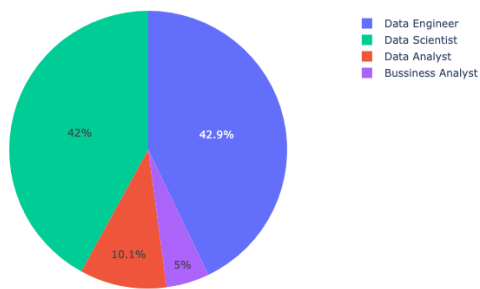
Pie chart of Job Type and OOP



Pie chart of Job Type and Modeling



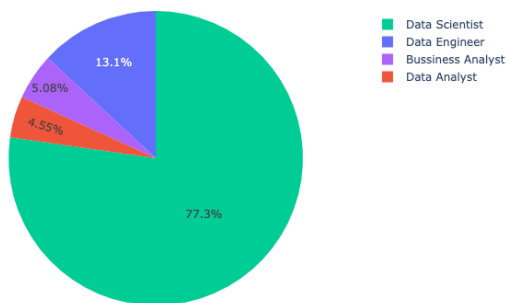
Pie chart of Job Type and Database



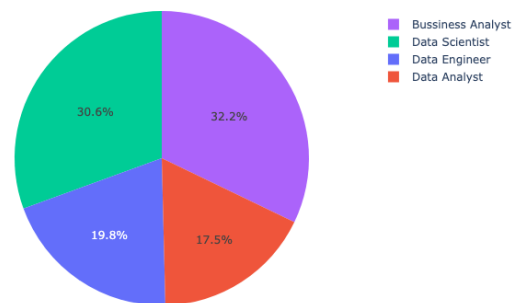
其次，對其他軟體能力要求和工作類別繪製圓餅圖，圖中可以發現，對於機器學習，模型建構，資料庫，ETL等能力的要求，資料科學家職位所佔的比例都為最高，而商業分析師都為最低。對於物件導向程式設計的要求，資料工程師為最高，而其次為資料科學家，由此可知因為資料工程師需要進行資料庫的設計和建立以及管理，因此要求OOP的工作也會較多，以上五張圖可以了解到，對於這些軟體能力來說，資料科學家及資料工程師職位的要求，相對於商業分析師和資料分析師來說高出許多。且對於機器學習和模型建構的能力要求，資料科學家的佔比最重，若求職者欲申請資料科學家的職位應更進一步的了解該方面的能力，以獲得較多機會。

[3] 學歷要求

Pie chart of Job Type and PHD



Pie chart of Job Type and MS



最後以學歷和工作類別進行圓餅圖的繪製，可以發現在碩士學歷的要求中，資料科學家和商業分析師佔最高的要求，而資料工程師和資料分析師的要求比例較低，推斷有可能是因為資料科學家所要求的工作技能較多及除了程式能力之外，需要對數學和統計學有更進一步的理解，因此在學歷的要求上會較高。而對商業分析師來說，需要即時在產品的第一線對業績和成長做出貢獻，為此則需要更多商業化或是營利的實戰經驗，因此對學歷的要求可能也會較高。在博士學歷的要求方面，資料科學家的職位佔了百分之70的比例，有可能是因為需要應用AI模型及不同的演算法，或是為公司進行新演算法開發或成為研究員，因此要求博士學歷的比例也最高。

4.模型建構

(程式碼為./model/model.py)

本專案透過XGBoost分類器，針對不同的薪資期望、學歷、欲尋找公司型態、技能等求職者本身的條件，分類預測欲尋找資料科學相關工作的求職者，目前自身情況所適合的工作類別，再由資料庫列出符合條件的工作職位進行推薦。目前所建構的模型，經過Tunning之後，其預測能力如下圖：

```
Test Result:

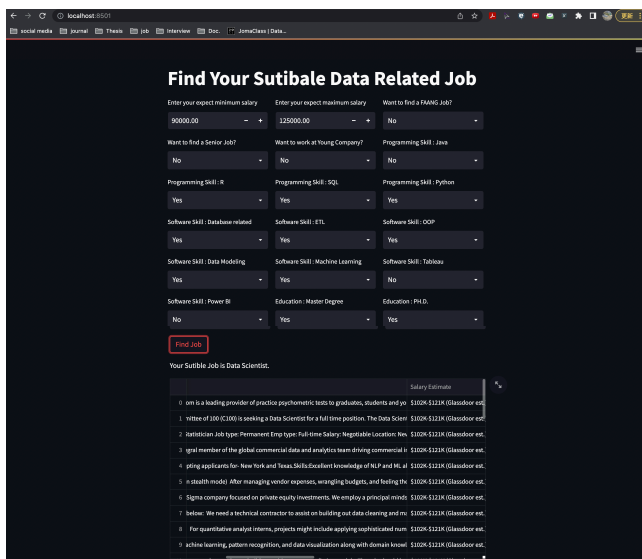
Accuracy Score: 95.90%
Precision Score: 95.90%
Recall Score: 95.90%
F1 score: 95.90%
Confusion Matrix:
[[754   2   8   3]
 [ 4 470   2  13]
 [ 7   2 411  24]
 [11   8  15 682]]
```

目前模型的準確度為95.90%，有可能在資料集中有尚未捕捉完全的特徵值。可以透過輸入自身的能力、學歷、和期望薪資，來知道自己目前在資料科學相關工作方面較適合的工作類別。

5. Result Visualization

(程式碼為./streamlit_app.py和./main.py)

本專案透過streamlit套件建立prototype在localhost中，透過簡易的UI選擇和輸入特徵值，並透過訓練完成的模型進行預測，將推薦的工作類別和適合的工作一併呈現在網站中,成果如下圖示意:



6. Future work

- (1) 更進一步進行更深入的Data mining 或加入其他資料提高模型的精準度
- (2) 擴大資料庫，加入Real time Data
- (3) 部署模型
- (4) 加入統計分析圖表幫助求職者判斷哪些能力可以優先加強

