GROUP PROJECT

(Python and Data Analysis)

glassdoor

We're **Group 2**

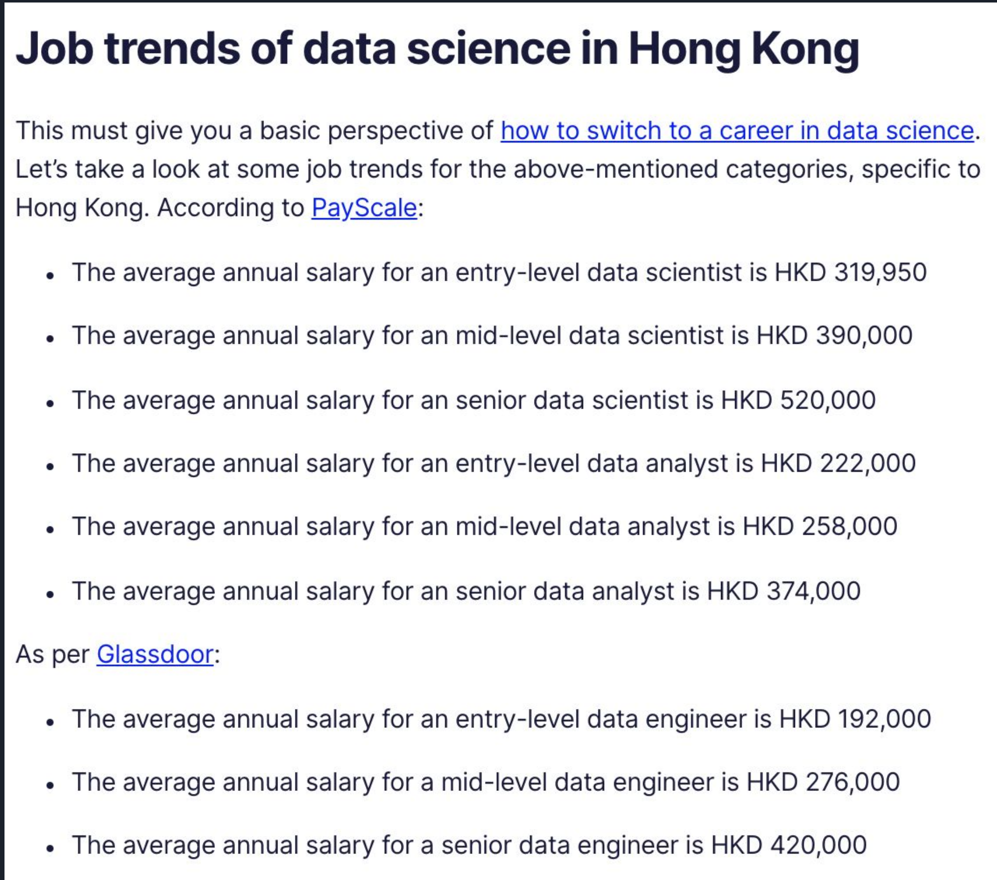
Why are we doing this?

Have you read this article?

Xccelerate

2021 Career Guide: Data Science & Machine Learning in Hong Kong

Thomas Ho
30 Jan 2021



Job trends of data science in Hong Kong

This must give you a basic perspective of how to switch to a career in data science. Let's take a look at some job trends for the above-mentioned categories, specific to Hong Kong. According to PayScale:

- The average annual salary for an entry-level data scientist is HKD 319,950
- The average annual salary for an mid-level data scientist is HKD 390,000
- The average annual salary for an senior data scientist is HKD 520,000
- The average annual salary for an entry-level data analyst is HKD 222,000
- The average annual salary for an mid-level data analyst is HKD 258,000
- The average annual salary for an senior data analyst is HKD 374,000

As per Glassdoor:

- The average annual salary for an entry-level data engineer is HKD 192,000
- The average annual salary for a mid-level data engineer is HKD 276,000
- The average annual salary for a senior data engineer is HKD 420,000

Our Goals

To increase transparency and clarify data-oriented job market in Hong Kong.

BUSINESS VALUE

Our Offers

- To offer a solution on how to **optimise job search** for prospective Data Analysts / Engineers / Data Scientists

- To provide corporate HRs / recruiting firms with an efficient way of conducting **competitive measurement** (quant vs qual)

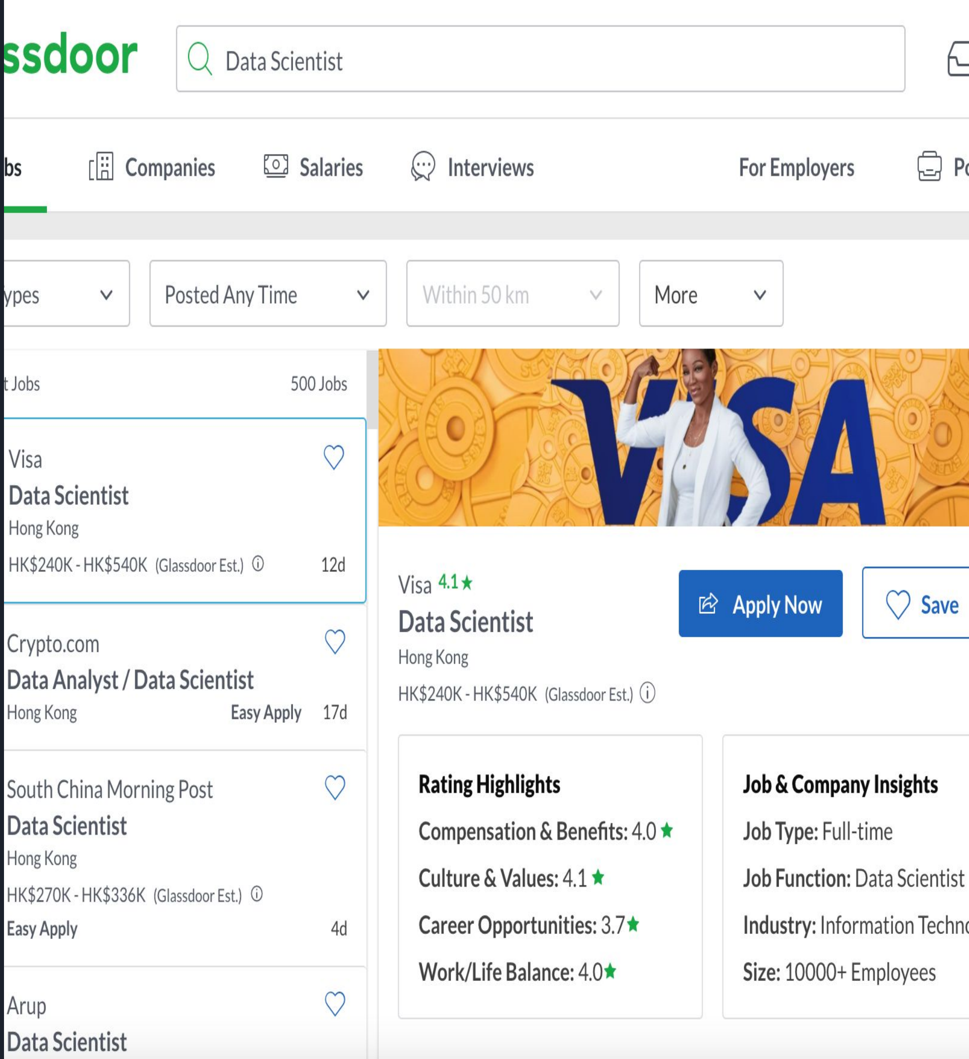
DATA COLLECTION

Glassdoor

Scraped multiple pages of job postings for the following roles:

- Data Analyst
- Data Engineer
- Data Scientist

glassdoor



ssdoor Data Scientist

bs Companies Salaries Interviews For Employers Po

ypes Posted Any Time Within 50 km More

t Jobs 500 Jobs

Visa
Data Scientist
Hong Kong
HK$240K - HK$540K (Glassdoor Est.) 12d

Crypto.com
Data Analyst / Data Scientist
Hong Kong Easy Apply 17d

South China Morning Post
Data Scientist
Hong Kong
HK$270K - HK$336K (Glassdoor Est.)
Easy Apply 4d

Arup
Data Scientist

Visa 4.1 ★
Data Scientist
Hong Kong
HK$240K - HK$540K (Glassdoor Est.)

Apply Now Save

Rating Highlights
Compensation & Benefits: 4.0 ★
Culture & Values: 4.1 ★
Career Opportunities: 3.7 ★
Work/Life Balance: 4.0 ★

Job & Company Insights
Job Type: Full-time
Job Function: Data Scientist
Industry: Information Techn
Size: 10000+ Employees

DATA PROCESSING

Our Method

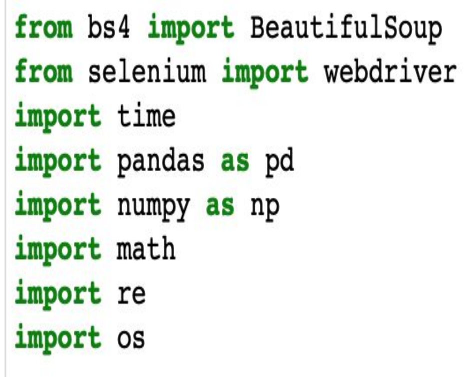
1. Scraping

Scraped Glassdoor for each role using:

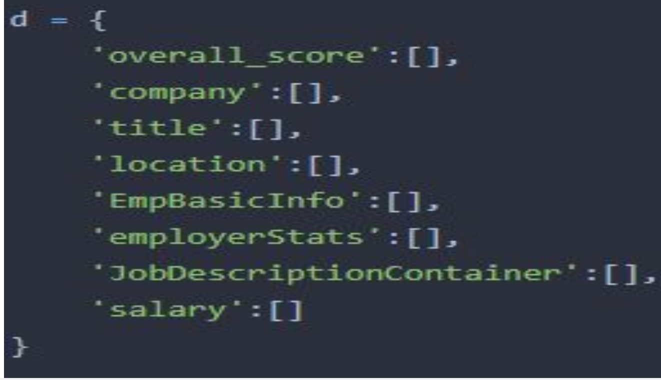
- Selenium

Data processing with:

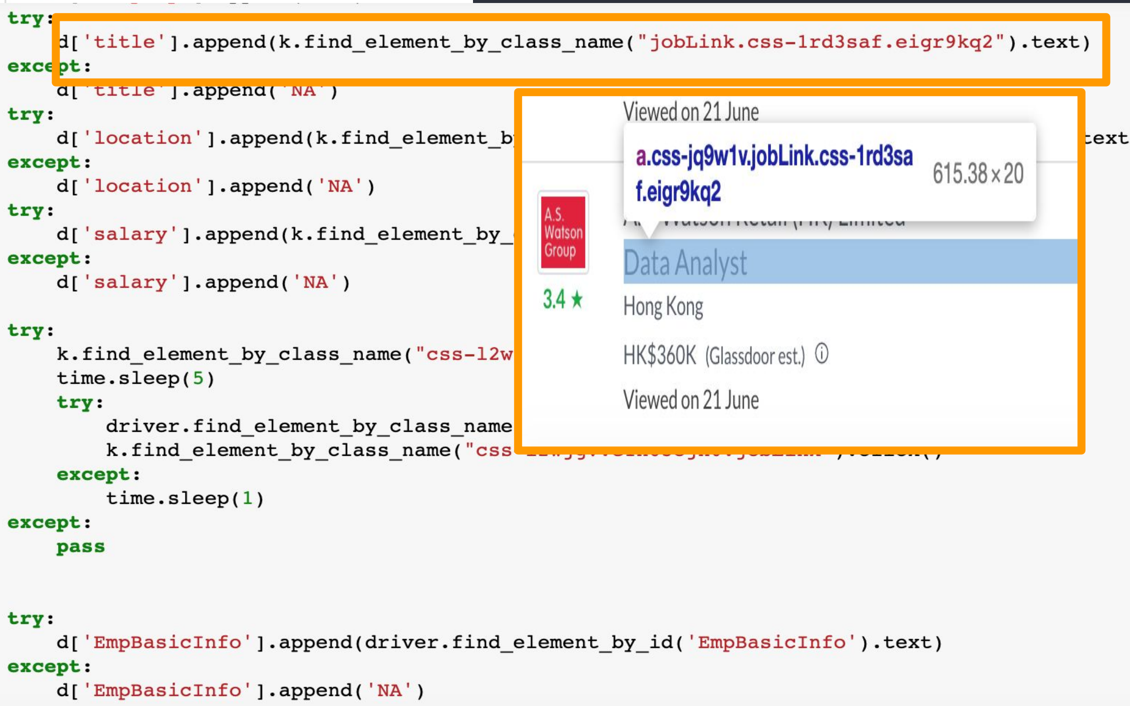
Pandas, Time, Numpy, OS, Math, Regex



```
from bs4 import BeautifulSoup
from selenium import webdriver
import time
import pandas as pd
import numpy as np
import math
import re
import os
```

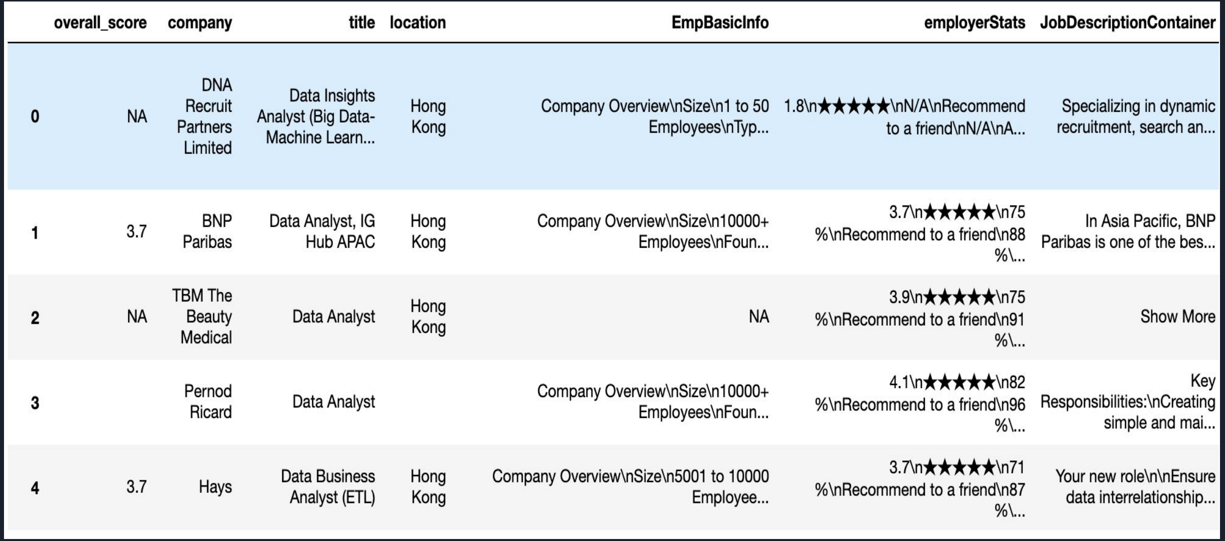


```
d = {
        'overall_score':[],
        'company':[],
        'title':[],
        'location':[],
        'EmpBasicInfo':[],
        'employerStats':[],
        'JobDescriptionContainer':[],
        'salary':[]
}
```



```
try:
    d['title'].append(k.find_element_by_class_name("jobLink.css-1rd3saf.eigr9kq2").text)
except:
    d['title'].append('NA')
try:
    d['location'].append(k.find_element_b
except:
    d['location'].append('NA')
try:
    d['salary'].append(k.find_element_by_
except:
    d['salary'].append('NA')

try:
    k.find_element_by_class_name("css-l2w
    time.sleep(5)
    try:
        driver.find_element_by_class_name
        k.find_element_by_class_name("css
    except:
        time.sleep(1)
except:
    pass


try:
    d['EmpBasicInfo'].append(driver.find_element_by_id('EmpBasicInfo').text)
except:
    d['EmpBasicInfo'].append('NA')
```

Our Method



	overall_score	company	title	location	EmpBasicInfo	employerStats	JobDescriptionContainer
0	NA	DNA Recruit Partners Limited	Data Insights Analyst (Big Data-Machine Learn...	Hong Kong	Company Overview\nSize\n1 to 50 Employees\nTyp...	1.8\n★★★★★\nN/A\nRecommend to a friend\nN/A\nA...	Specializing in dynamic recruitment, search an...
1	3.7	BNP Paribas	Data Analyst, IG Hub APAC	Hong Kong	Company Overview\nSize\n10000+ Employees\nFoun...	3.7\n★★★★★\n75 %\nRecommend to a friend\n88 %\...	In Asia Pacific, BNP Paribas is one of the bes...
2	NA	TBM The Beauty Medical	Data Analyst	Hong Kong	NA	3.9\n★★★★★\n75 %\nRecommend to a friend\n91 %\...	Show More
3		Pernod Ricard	Data Analyst		Company Overview\nSize\n10000+ Employees\nFoun...	4.1\n★★★★★\n82 %\nRecommend to a friend\n96 %\...	Key Responsibilities:\nCreating simple and mai...
4	3.7	Hays	Data Business Analyst (ETL)	Hong Kong	Company Overview\nSize\n5001 to 10000 Employee...	3.7\n★★★★★\n71 %\nRecommend to a friend\n87 %\...	Your new role\n\nEnsure data interrelationship...

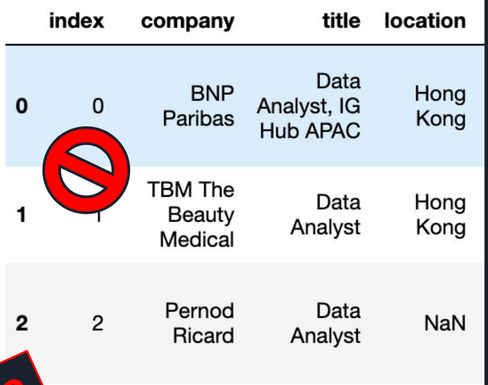
2. Merging

- Merge 3 data frames (DA, DE & DS) together
- Drop any duplicate rows
- Drop any entries without company names to ensure high-quality data

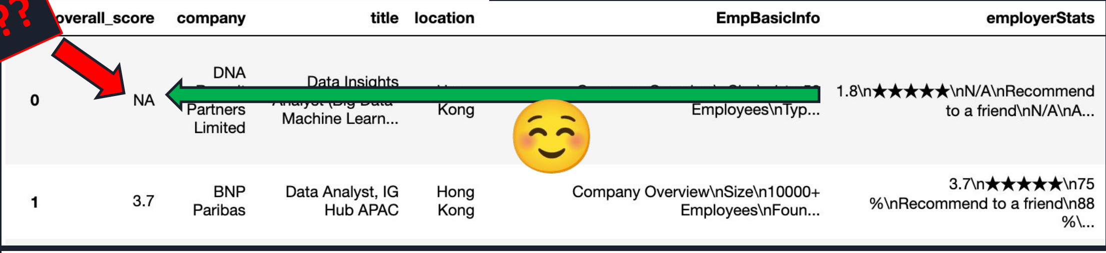
Our Method

3. Clean-up

- Drop unnecessary columns
- Ensure overall_score is extracted from the "*employerStats*" column
- New column to categorise each role to (1)*Data Analyst*, (2)*Data Engineer*, (3)*Data Scientist*
- Merge 3 data frames together

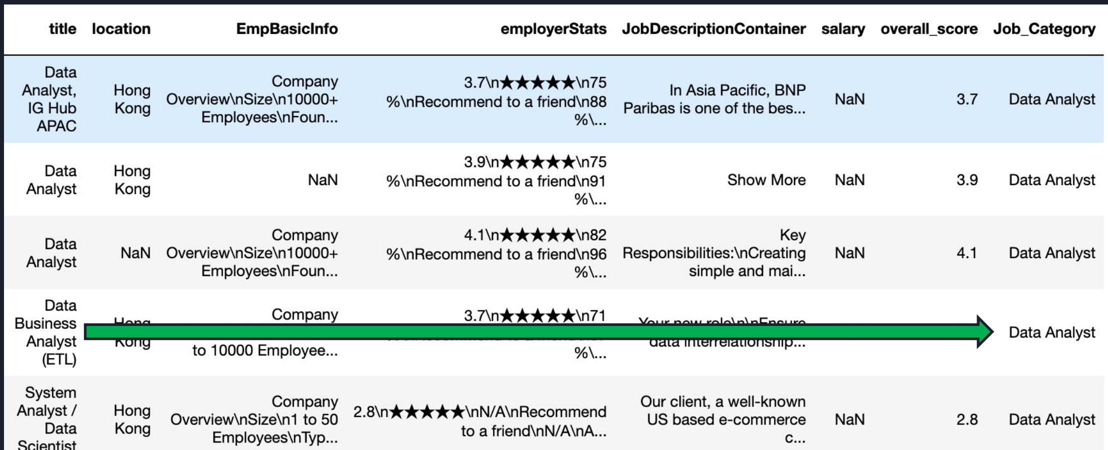


	index	company	title	location
0	0	BNP Paribas	Data Analyst, IG Hub APAC	Hong Kong
1	1	TBM The Beauty Medical	Data Analyst	Hong Kong
2	2	Pernod Ricard	Data Analyst	NaN



???

	overall_score	company	title	location	EmpBasicInfo	employerStats
0	NA	DNA [illegible] Partners Limited	Data Insights Analyst (Big Data Machine Learn...	Hong Kong	Company Overview\nSize\n1 to 50 Employees\nTyp...	1.8\n★★★★★\nN/A\nRecommend to a friend\nN/A\nA...
1	3.7	BNP Paribas	Data Analyst, IG Hub APAC	Hong Kong	Company Overview\nSize\n10000+ Employees\nFoun...	3.7\n★★★★★\n75 %\nRecommend to a friend\n88 %\...



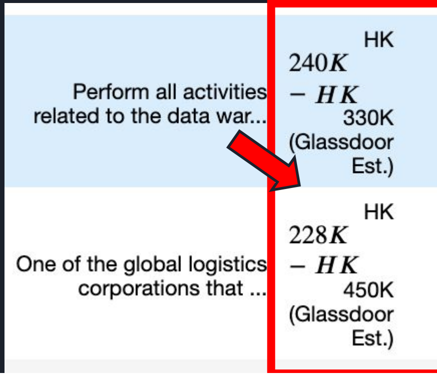
title	location	EmpBasicInfo	employerStats	JobDescriptionContainer	salary	overall_score	Job_Category
Data Analyst, IG Hub APAC	Hong Kong	Company Overview\nSize\n10000+ Employees\nFoun...	3.7\n★★★★★\n75 %\nRecommend to a friend\n88 %\...	In Asia Pacific, BNP Paribas is one of the bes...	NaN	3.7	Data Analyst
Data Analyst	Hong Kong	NaN	3.9\n★★★★★\n75 %\nRecommend to a friend\n91 %\...	Show More	NaN	3.9	Data Analyst
Data Analyst	NaN	Company Overview\nSize\n10000+ Employees\nFoun...	4.1\n★★★★★\n82 %\nRecommend to a friend\n96 %\...	Key Responsibilities:\nCreating simple and mai...	NaN	4.1	Data Analyst
Data Business Analyst (ETL)	Hong Kong	Company [illegible] to 10000 Employee...	3.7\n★★★★★\n71 [illegible] %\...	Your new role\n\nEnsure data interrelationship...			Data Analyst
System Analyst / Data Scientist	Hong Kong	Company Overview\nSize\n1 to 50 Employees\nTyp...	2.8\n★★★★★\nN/A\nRecommend to a friend\nN/A\nA...	Our client, a well-known US based e-commerce c...	NaN	2.8	Data Analyst

Our Method

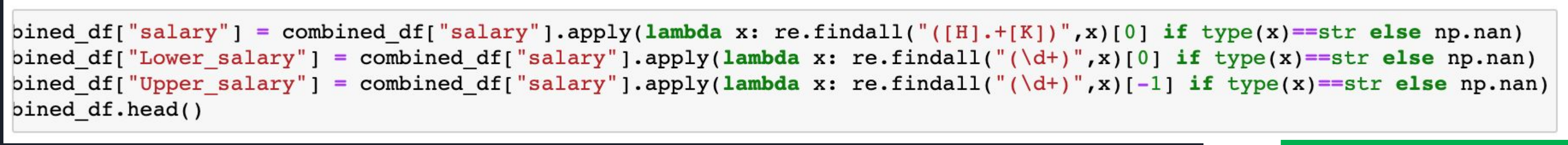
3. Clean-up

“salary”

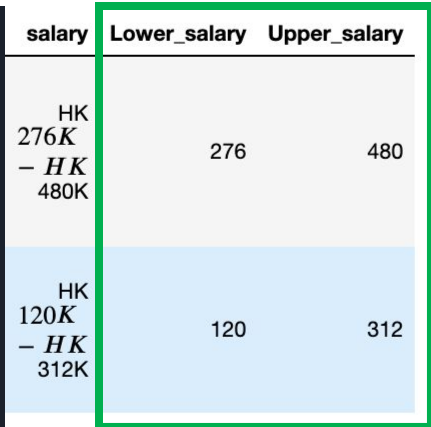
- Separate into Lower Range and Upper Range



Perform all activities related to the data war...	HK 240$K - HK$ 330K (Glassdoor Est.)
One of the global logistics corporations that ...	HK 228$K - HK$ 450K (Glassdoor Est.)



```
bined_df["salary"] = combined_df["salary"].apply(lambda x: re.findall("([H].+[K])",x)[0] if type(x)==str else np.nan)
bined_df["Lower_salary"] = combined_df["salary"].apply(lambda x: re.findall("(\d+)",x)[0] if type(x)==str else np.nan)
bined_df["Upper_salary"] = combined_df["salary"].apply(lambda x: re.findall("(\d+)",x)[-1] if type(x)==str else np.nan)
bined_df.head()
```

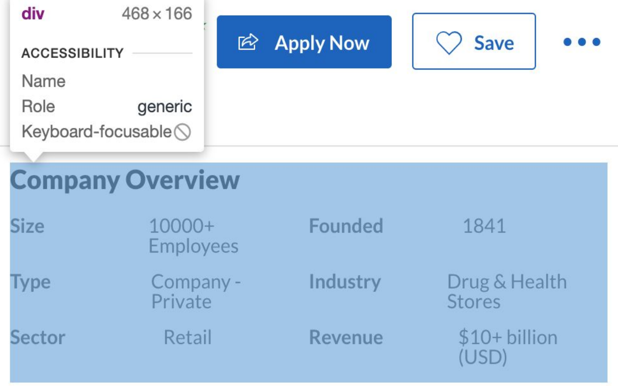


salary	Lower_salary	Upper_salary
HK 276$K - HK$ 480K	276	480
HK 120$K - HK$ 312K	120	312

Our Method

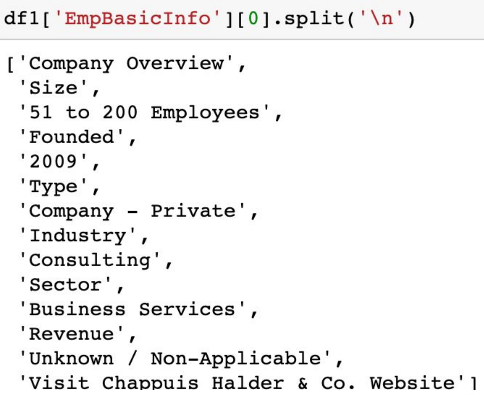
3. Clean-up *“EmpBasicInfo”*

- Split extracted strings
- Create dictionary for each job posting
- Build a temporal list for each intended independent variable
- Append each list with extracted value from all created dictionaries

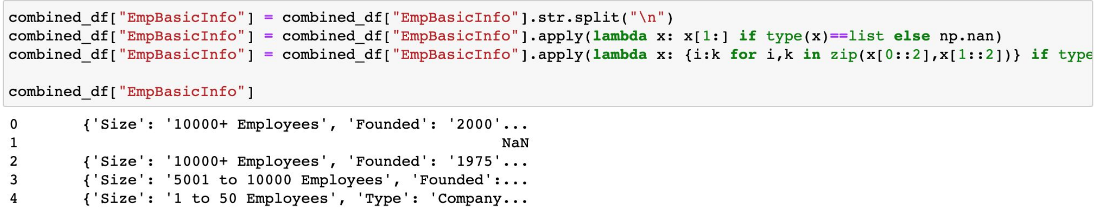


Company Overview

Size	10000+ Employees	Founded	1841
Type	Company - Private	Industry	Drug & Health Stores
Sector	Retail	Revenue	$10+ billion (USD)

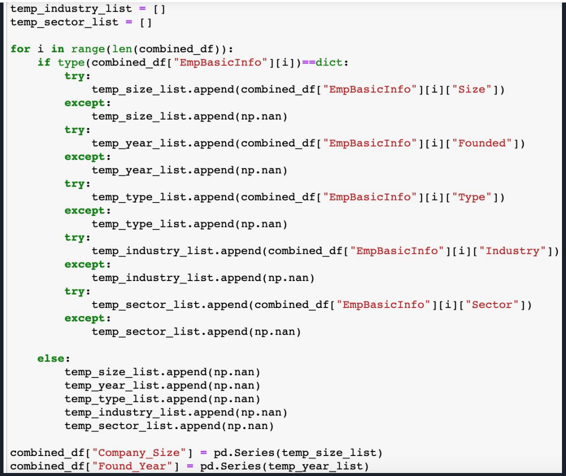


```
df1['EmpBasicInfo'][0].split('\n')

['Company Overview',
 'Size',
 '51 to 200 Employees',
 'Founded',
 '2009',
 'Type',
 'Company - Private',
 'Industry',
 'Consulting',
 'Sector',
 'Business Services',
 'Revenue',
 'Unknown / Non-Applicable',
 'Visit Chappuis Halder & Co. Website']
```

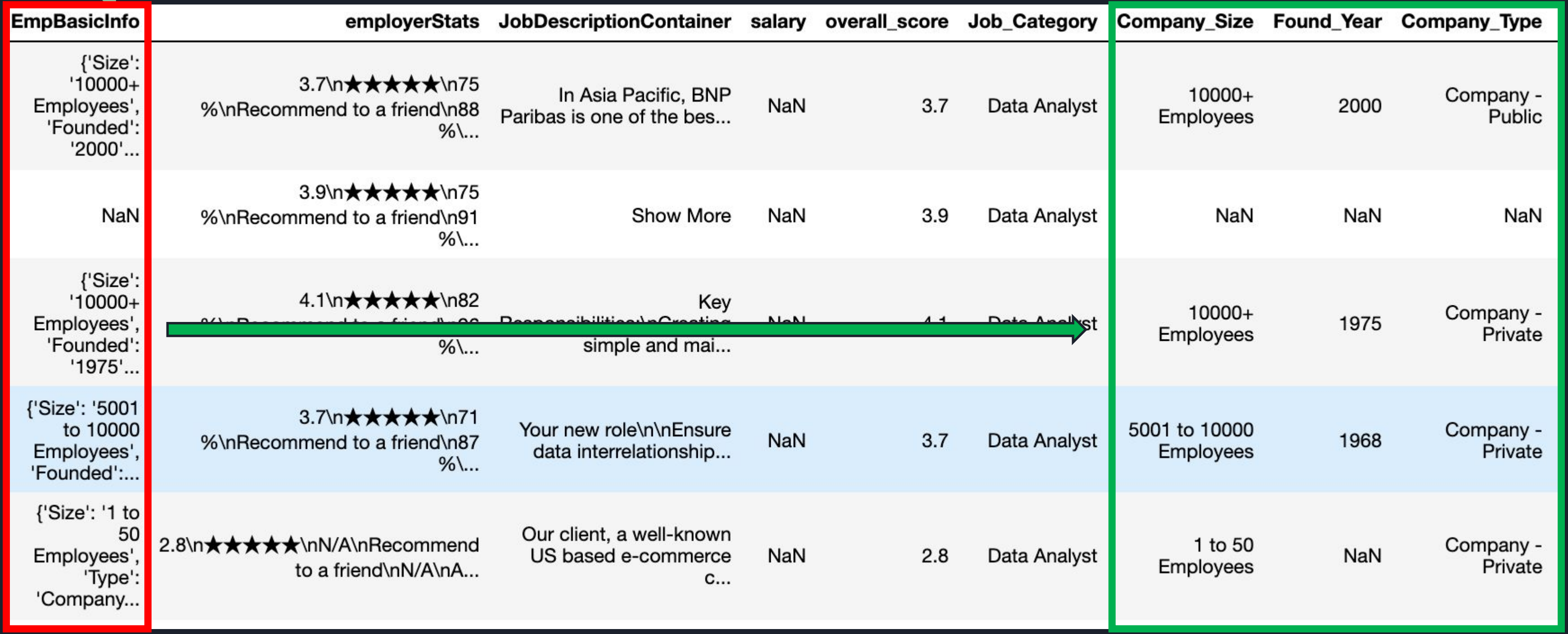


```
combined_df["EmpBasicInfo"] = combined_df["EmpBasicInfo"].str.split("\n")
combined_df["EmpBasicInfo"] = combined_df["EmpBasicInfo"].apply(lambda x: x[1:] if type(x)==list else np.nan)
combined_df["EmpBasicInfo"] = combined_df["EmpBasicInfo"].apply(lambda x: {i:k for i,k in zip(x[0::2],x[1::2])} if type

combined_df["EmpBasicInfo"]
```

```
0       {'Size': '10000+ Employees', 'Founded': '2000'...
1                                                     NaN
2       {'Size': '10000+ Employees', 'Founded': '1975'...
3       {'Size': '5001 to 10000 Employees', 'Founded':...
4       {'Size': '1 to 50 Employees', 'Type': 'Company...
```



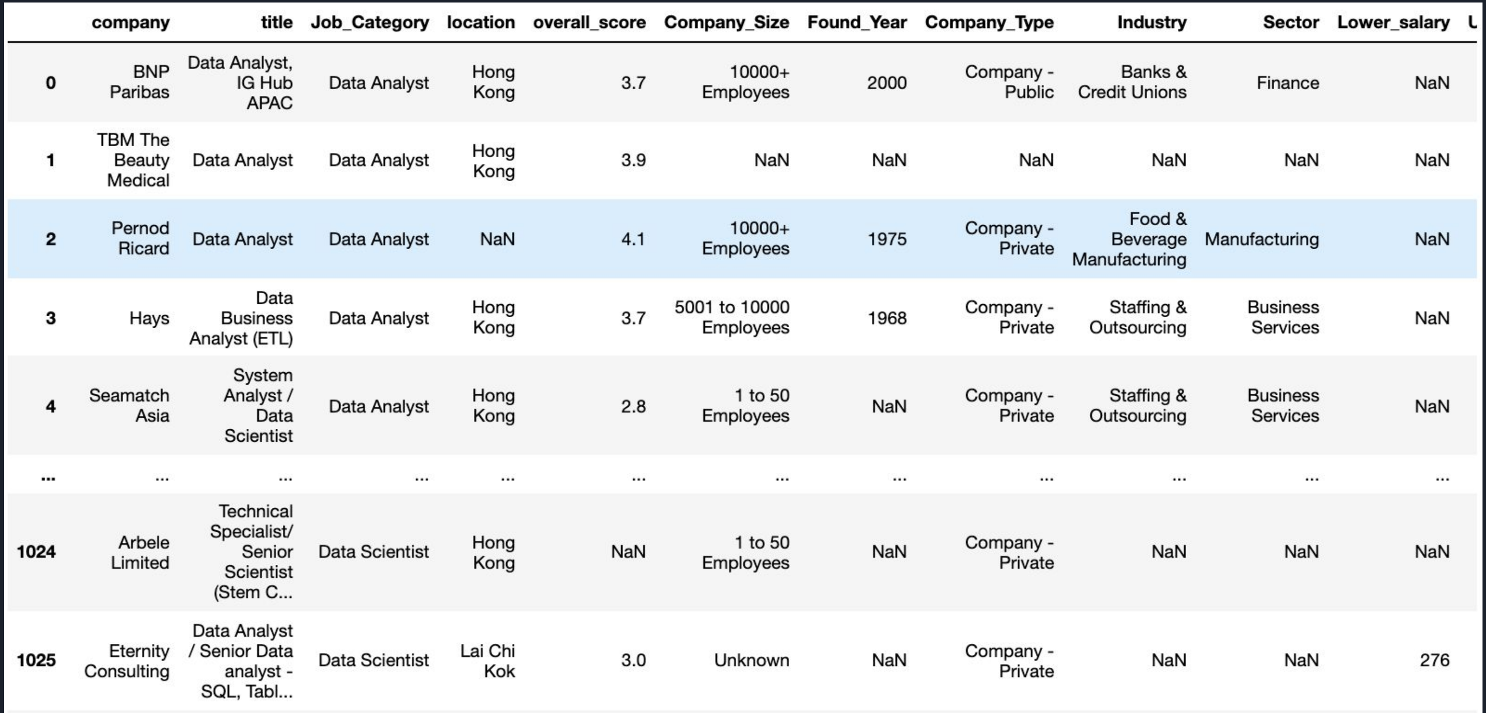
```
temp_industry_list = []
temp_sector_list = []

for i in range(len(combined_df)):
    if type(combined_df["EmpBasicInfo"][i])==dict:
        try:
            temp_size_list.append(combined_df["EmpBasicInfo"][i]["Size"])
        except:
            temp_size_list.append(np.nan)
        try:
            temp_year_list.append(combined_df["EmpBasicInfo"][i]["Founded"])
        except:
            temp_year_list.append(np.nan)
        try:
            temp_type_list.append(combined_df["EmpBasicInfo"][i]["Type"])
        except:
            temp_type_list.append(np.nan)
        try:
            temp_industry_list.append(combined_df["EmpBasicInfo"][i]["Industry"])
        except:
            temp_industry_list.append(np.nan)
        try:
            temp_sector_list.append(combined_df["EmpBasicInfo"][i]["Sector"])
        except:
            temp_sector_list.append(np.nan)

    else:
        temp_size_list.append(np.nan)
        temp_year_list.append(np.nan)
        temp_type_list.append(np.nan)
        temp_industry_list.append(np.nan)
        temp_sector_list.append(np.nan)

combined_df["Company_Size"] = pd.Series(temp_size_list)
combined_df["Found_Year"] = pd.Series(temp_year_list)
```



EmpBasicInfo	employerStats	JobDescriptionContainer	salary	overall_score	Job_Category	Company_Size	Found_Year	Company_Type
{'Size': '10000+ Employees', 'Founded': '2000'...	3.7\n★★★★★\n75 %\nRecommend to a friend\n88 %\...	In Asia Pacific, BNP Paribas is one of the bes...	NaN	3.7	Data Analyst	10000+ Employees	2000	Company - Public
NaN	3.9\n★★★★★\n75 %\nRecommend to a friend\n91 %\...	Show More	NaN	3.9	Data Analyst	NaN	NaN	NaN
{'Size': '10000+ Employees', 'Founded': '1975'...	4.1\n★★★★★\n82 %\nRecommend to a friend\n86 %\...	Key Responsibilities:\nCreating simple and mai...	NaN	4.1	Data Analyst	10000+ Employees	1975	Company - Private
{'Size': '5001 to 10000 Employees', 'Founded':...	3.7\n★★★★★\n71 %\nRecommend to a friend\n87 %\...	Your new role\n\nEnsure data interrelationship...	NaN	3.7	Data Analyst	5001 to 10000 Employees	1968	Company - Private
{'Size': '1 to 50 Employees', 'Type': 'Company...	2.8\n★★★★★\nN/A\nRecommend to a friend\nN/A\nA...	Our client, a well-known US based e-commerce c...	NaN	2.8	Data Analyst	1 to 50 Employees	NaN	Company - Private

Our Method

4. Final Data Frame

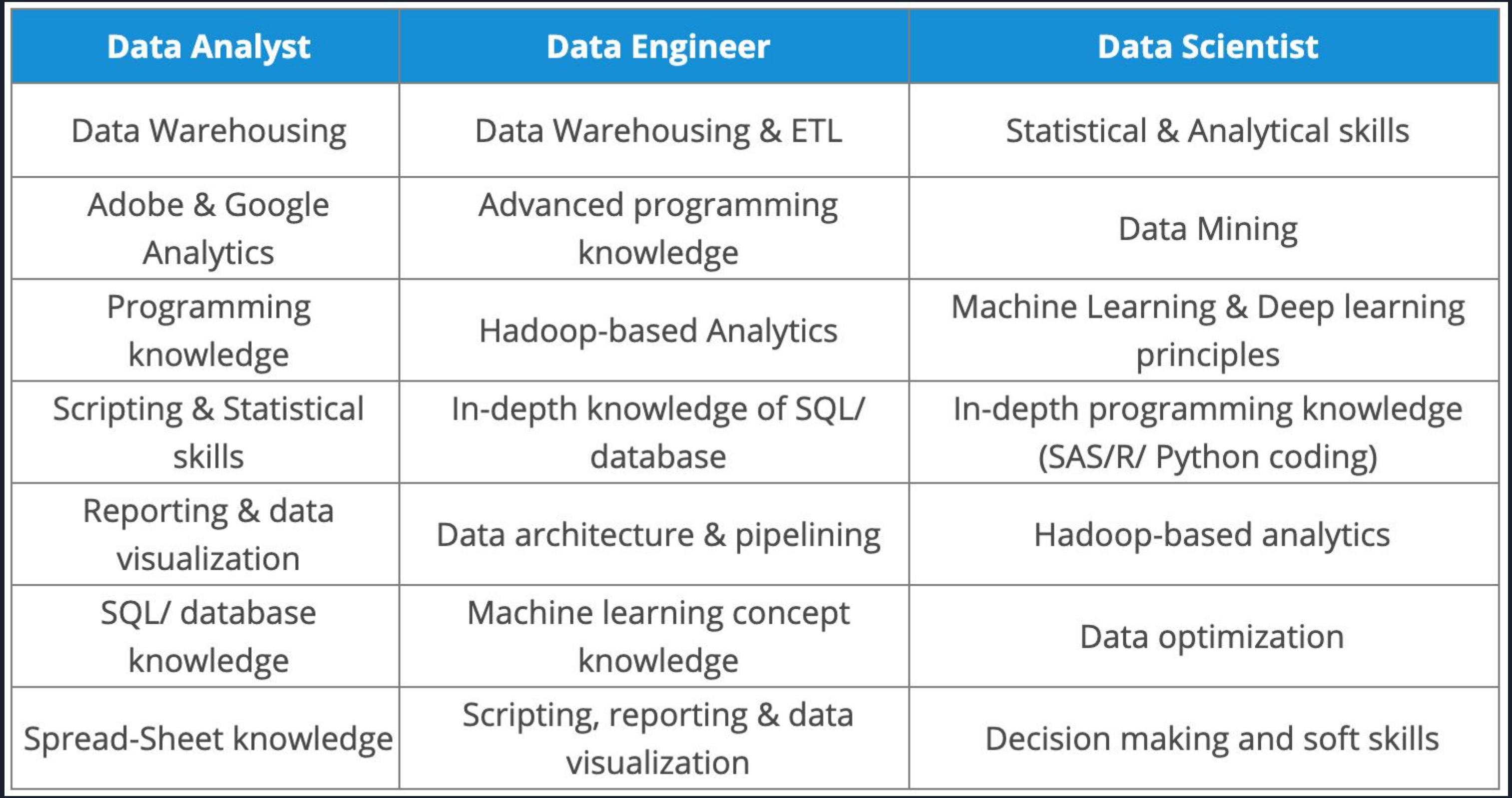


	company	title	Job_Category	location	overall_score	Company_Size	Found_Year	Company_Type	Industry	Sector	Lower_salary	U
0	BNP Paribas	Data Analyst, IG Hub APAC	Data Analyst	Hong Kong	3.7	10000+ Employees	2000	Company - Public	Banks & Credit Unions	Finance	NaN	
1	TBM The Beauty Medical	Data Analyst	Data Analyst	Hong Kong	3.9	NaN	NaN	NaN	NaN	NaN	NaN	
2	Pernod Ricard	Data Analyst	Data Analyst	NaN	4.1	10000+ Employees	1975	Company - Private	Food & Beverage Manufacturing	Manufacturing	NaN	
3	Hays	Data Business Analyst (ETL)	Data Analyst	Hong Kong	3.7	5001 to 10000 Employees	1968	Company - Private	Staffing & Outsourcing	Business Services	NaN	
4	Seamatch Asia	System Analyst / Data Scientist	Data Analyst	Hong Kong	2.8	1 to 50 Employees	NaN	Company - Private	Staffing & Outsourcing	Business Services	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	
1024	Arbele Limited	Technical Specialist/ Senior Scientist (Stem C...	Data Scientist	Hong Kong	NaN	1 to 50 Employees	NaN	Company - Private	NaN	NaN	NaN	
1025	Eternity Consulting	Data Analyst / Senior Data analyst - SQL, Tabl...	Data Scientist	Lai Chi Kok	3.0	Unknown	NaN	Company - Private	NaN	NaN	276	

ANALYSIS

The Job Market

Differentiating between roles

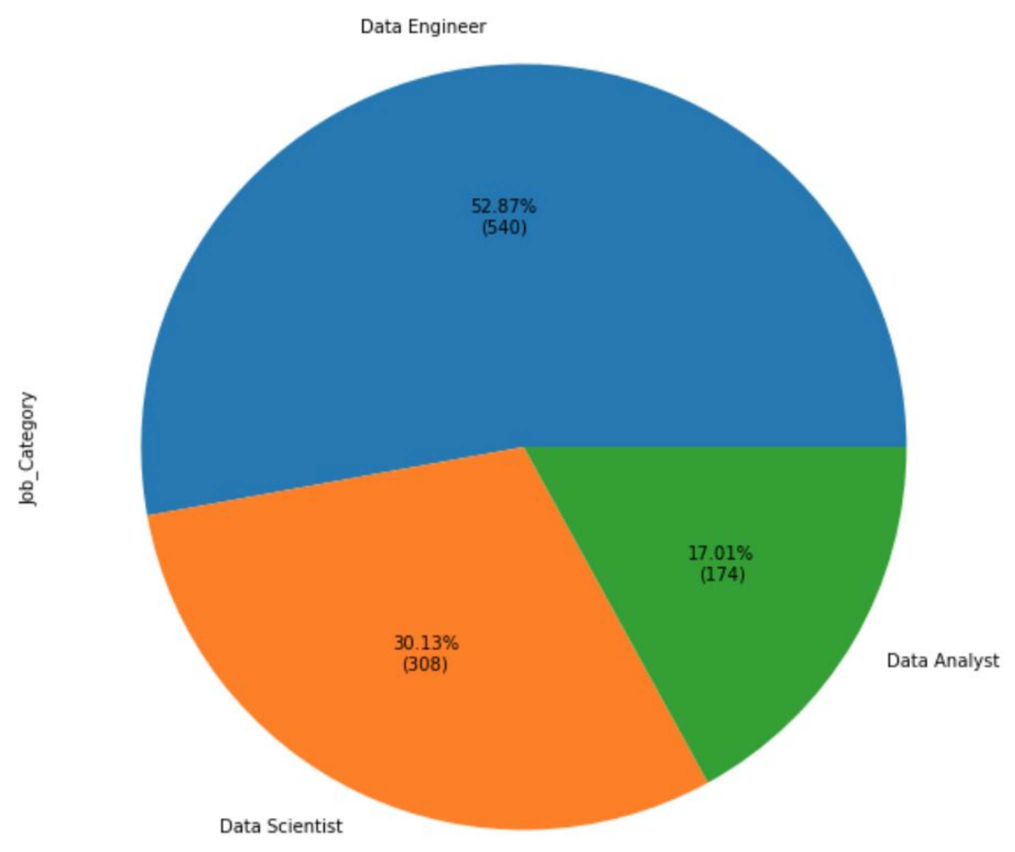


Data Analyst	Data Engineer	Data Scientist
Data Warehousing	Data Warehousing & ETL	Statistical & Analytical skills
Adobe & Google Analytics	Advanced programming knowledge	Data Mining
Programming knowledge	Hadoop-based Analytics	Machine Learning & Deep learning principles
Scripting & Statistical skills	In-depth knowledge of SQL/ database	In-depth programming knowledge (SAS/R/ Python coding)
Reporting & data visualization	Data architecture & pipelining	Hadoop-based analytics
SQL/ database knowledge	Machine learning concept knowledge	Data optimization
Spread-Sheet knowledge	Scripting, reporting & data visualization	Decision making and soft skills

Overall Market

As of **20 June 2021**, job postings were dominated by *Data Engineer* by an **exceptionally** large margin:

- **540** Data Engineers
- **308** Data Scientists
- **174** Data Analysts



Sectoral Demand

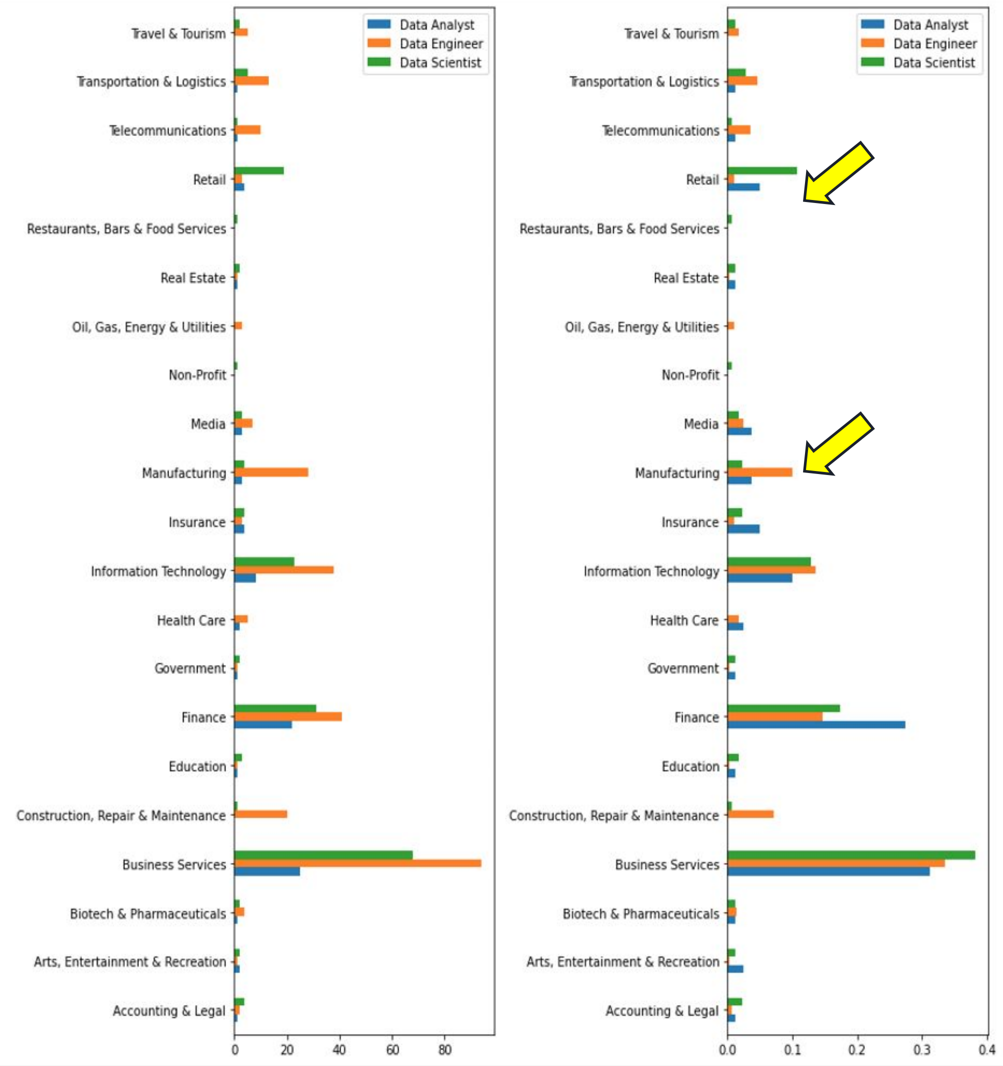
The following sectors lead the demand for data-oriented roles in terms of :

1. Business Services
2. Finance
3. Information Technology

Aside from the TOP 3 (>60%) above:

Applying for *Data Engineer* roles?
Be on the lookout for ones in *Manufacturing*.

Prospective *Data Scientists*?
Pay attention to the *Retail* sector.



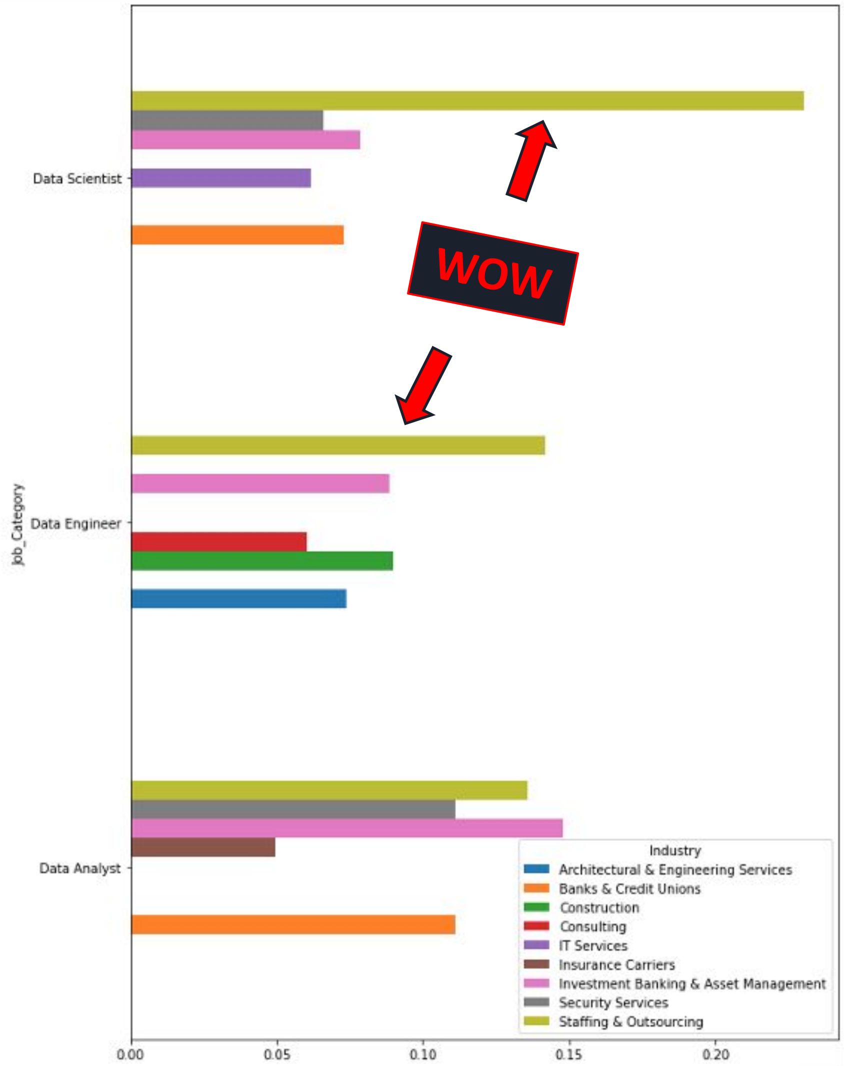
Recruiter Galore!

Data Analyst

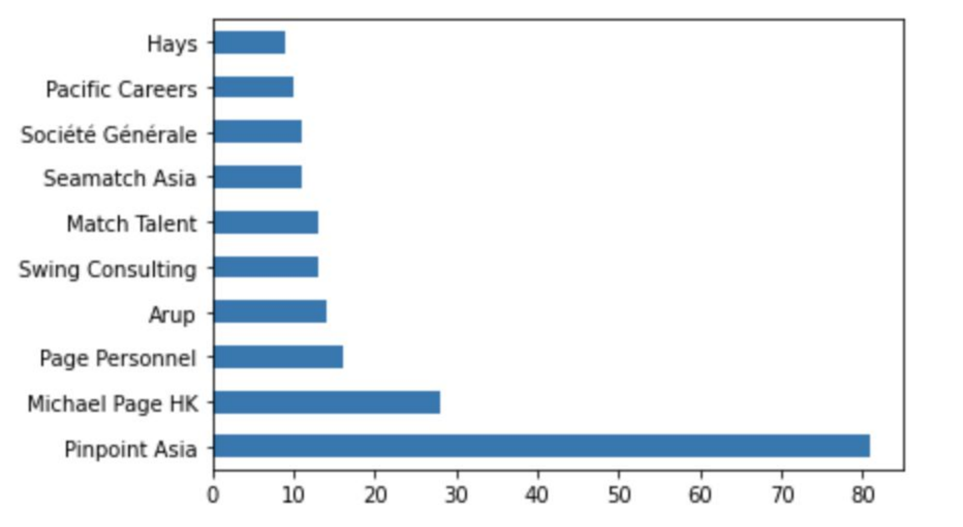
Multiple industries already account for 10%+ of job postings, indicating that corporates are familiar with the scope of *Data Analyst* roles.

Data Engineer / Data Scientist

Largely posted by *Staffing & Outsourcing* firms, representing ~15% and ~24% of such job posts.



Recruiter Galore!



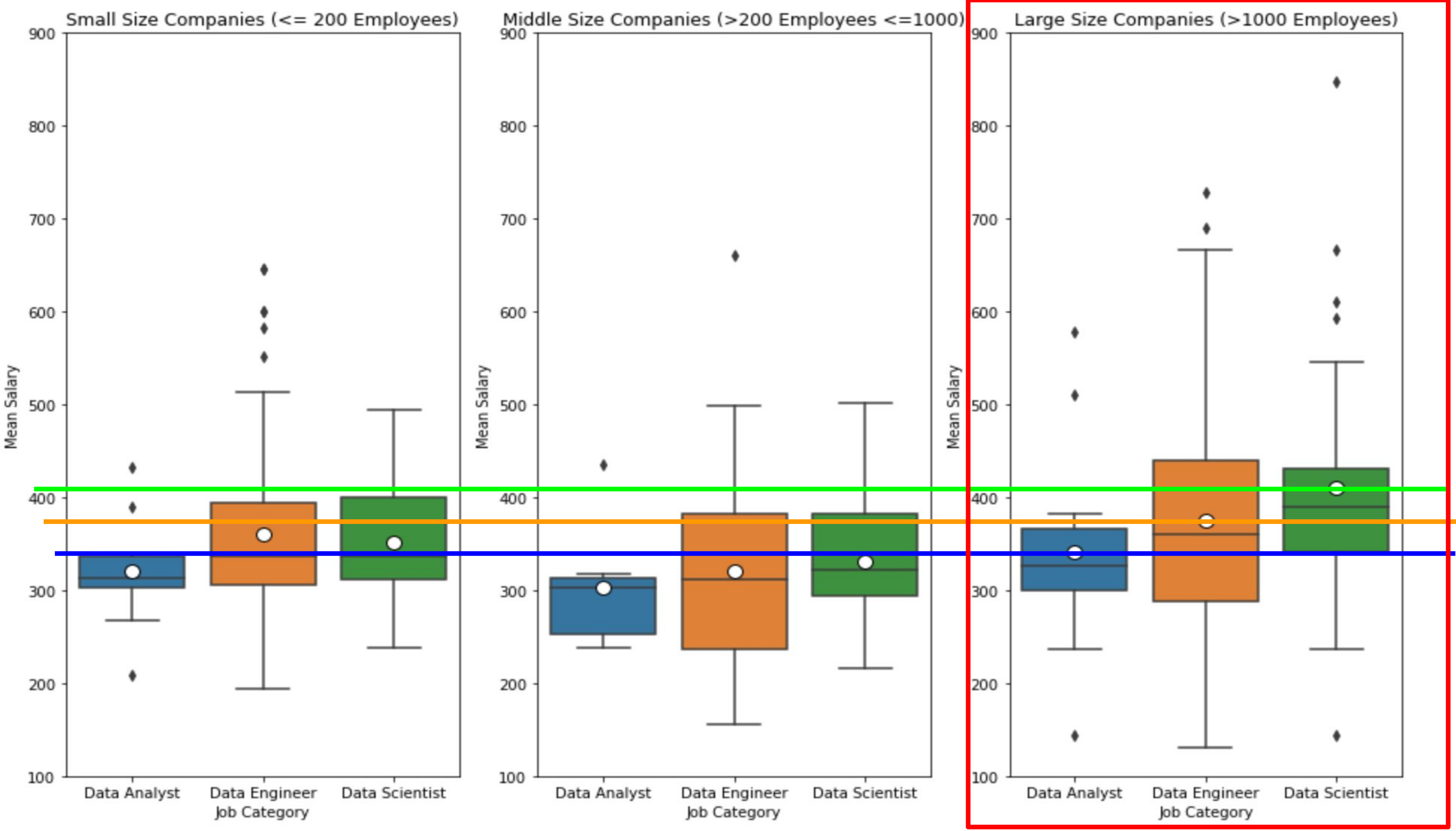
9 out of 10 of the companies are professional recruiting specialists

Applying for *Data Engineer / Data Scientist* roles?

Be prepared to face the recruiters first!

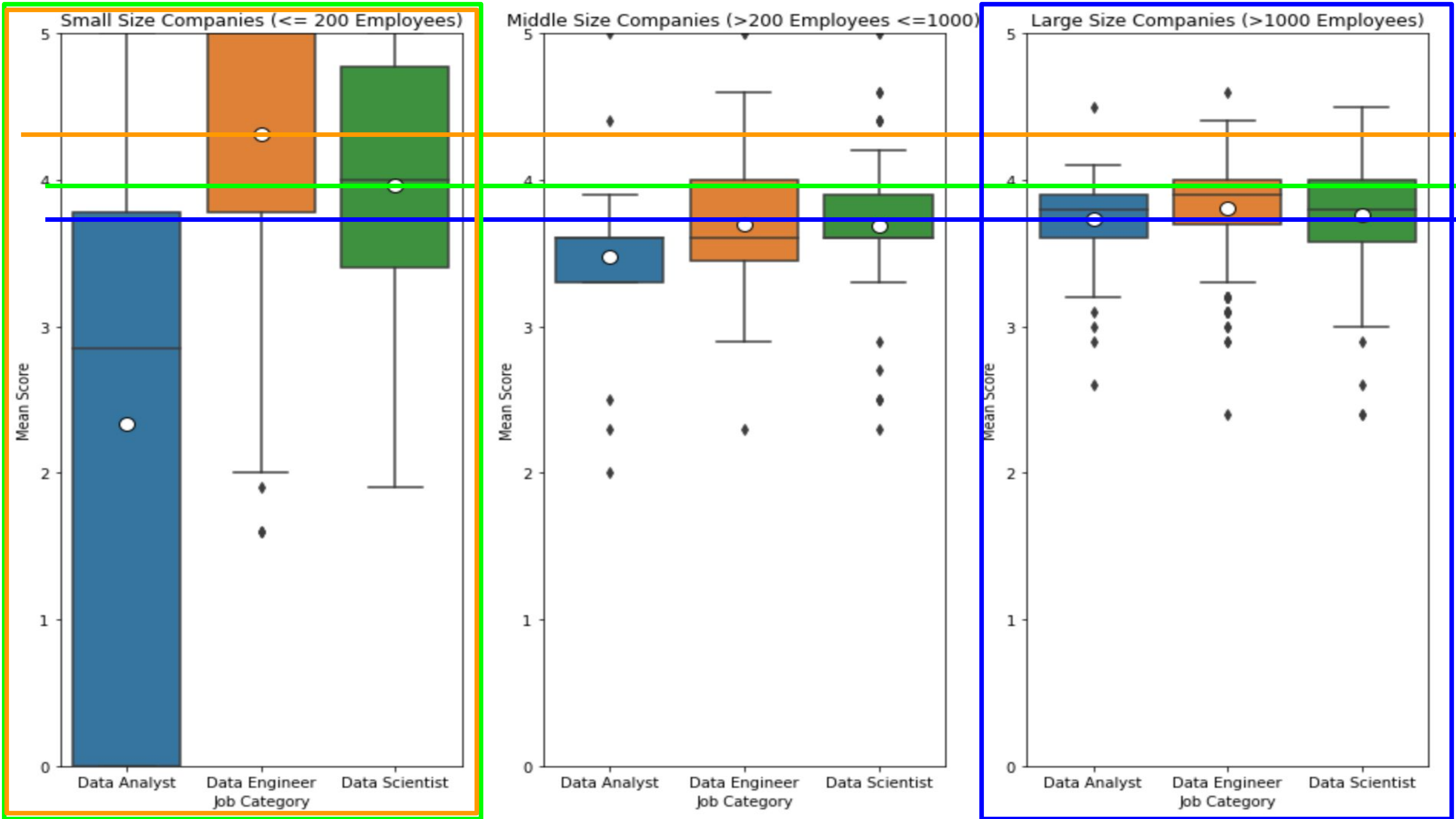
ANALYSIS

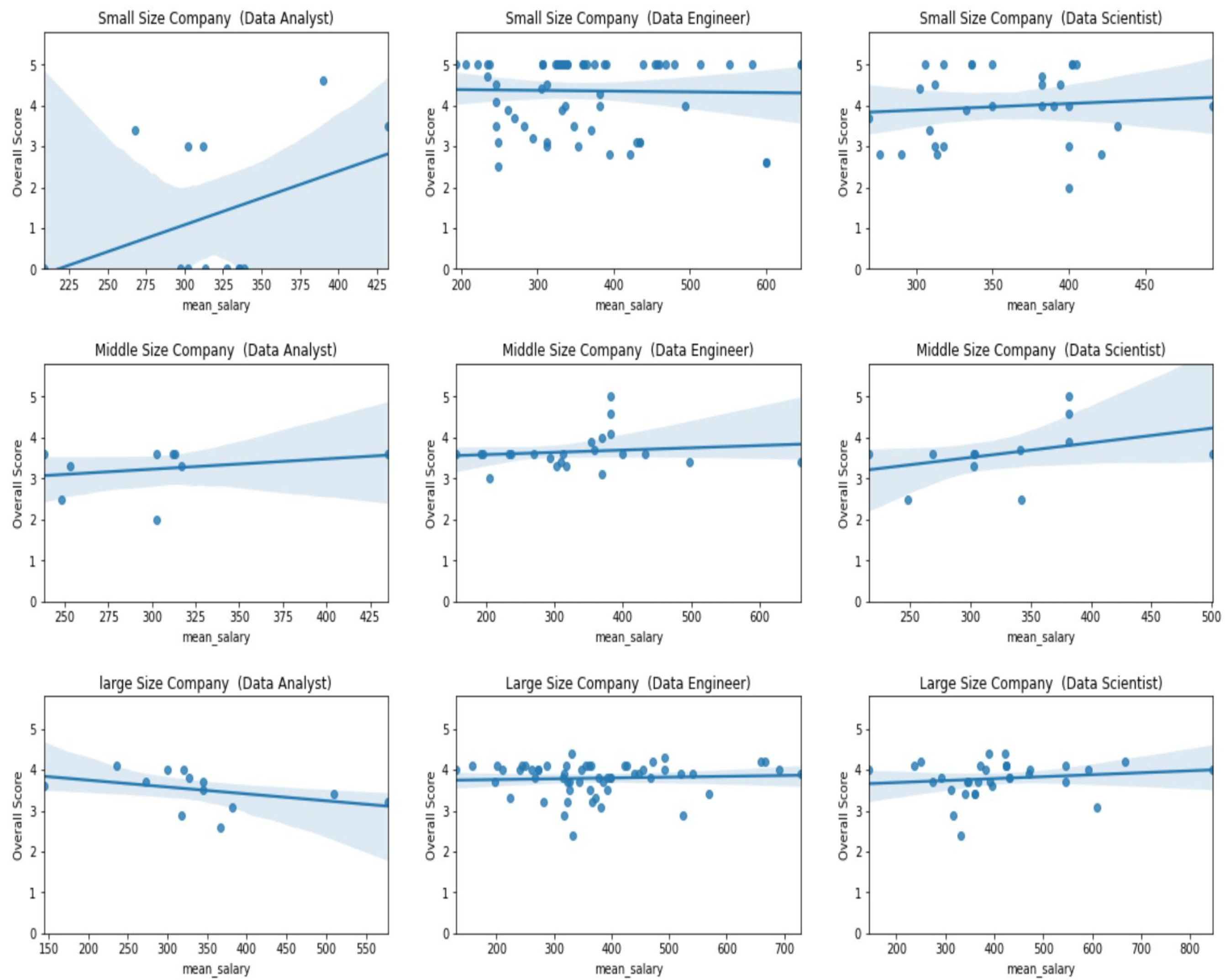
Salary per size of companies

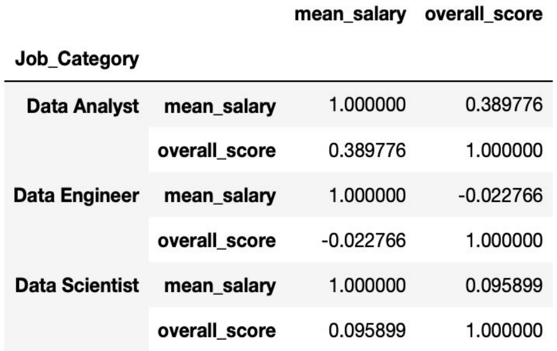

Small Size Companies (<= 200 Employees)
Middle Size Companies (>200 Employees <=1000)
Large Size Companies (>1000 Employees)
Mean Salary
900
800
700
600
500
400
300
200
100
Data Analyst
Data Engineer
Data Scientist
Job Category

ANALYSIS

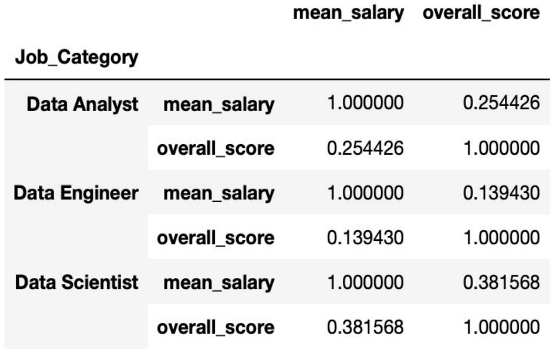
Overall score per size of companies


Small Size Companies (<= 200 Employees)
Middle Size Companies (>200 Employees <=1000)
Large Size Companies (>1000 Employees)
Mean Score
0
1
2
3
4
5
Data Analyst
Data Engineer
Data Scientist
Job Category

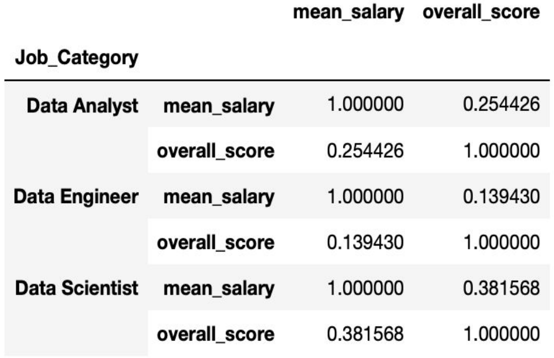




Job_Category		mean_salary	overall_score
Data Analyst	mean_salary	1.000000	0.389776
	overall_score	0.389776	1.000000
Data Engineer	mean_salary	1.000000	-0.022766
	overall_score	-0.022766	1.000000
Data Scientist	mean_salary	1.000000	0.095899
	overall_score	0.095899	1.000000



Job_Category		mean_salary	overall_score
Data Analyst	mean_salary	1.000000	0.254426
	overall_score	0.254426	1.000000
Data Engineer	mean_salary	1.000000	0.139430
	overall_score	0.139430	1.000000
Data Scientist	mean_salary	1.000000	0.381568
	overall_score	0.381568	1.000000



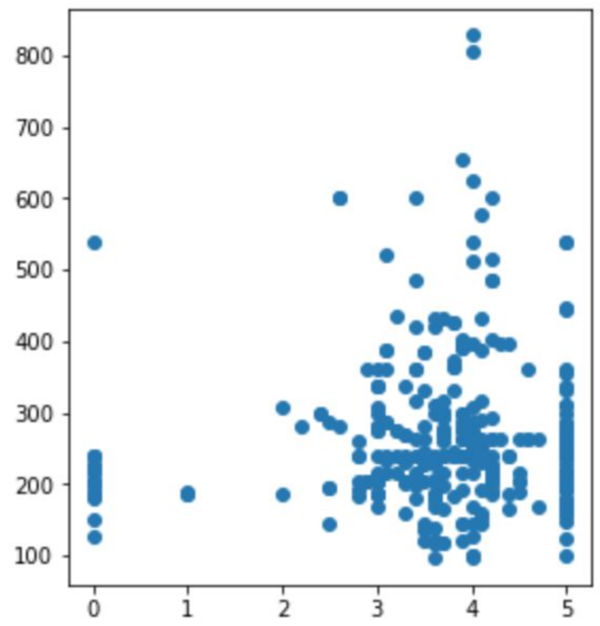
Job_Category		mean_salary	overall_score
Data Analyst	mean_salary	1.000000	0.254426
	overall_score	0.254426	1.000000
Data Engineer	mean_salary	1.000000	0.139430
	overall_score	0.139430	1.000000
Data Scientist	mean_salary	1.000000	0.381568
	overall_score	0.381568	1.000000

Salary won’t buy your employees’ happiness

Overall score vs Salary

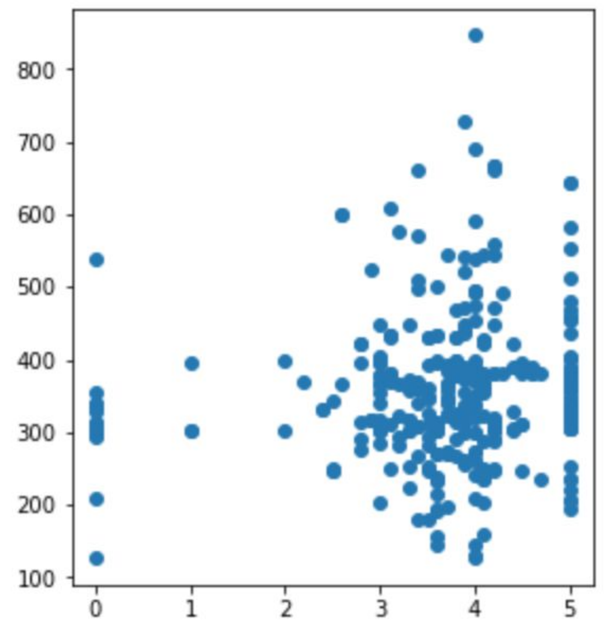
Lower Salary

correlation = 6.5%



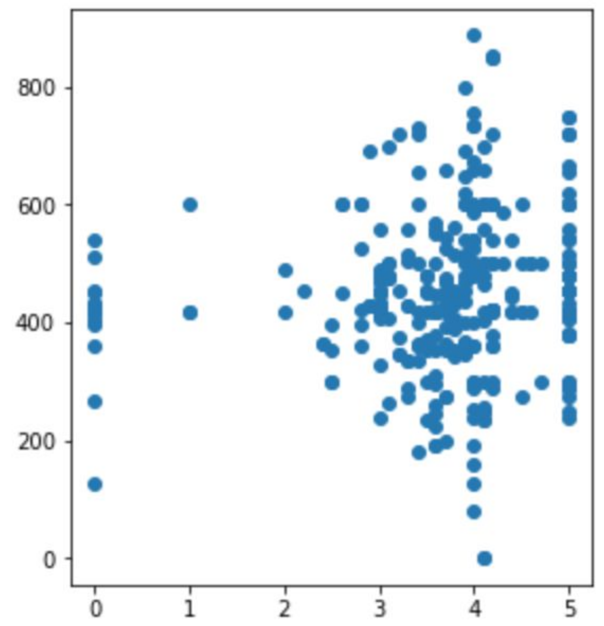
Mean Salary

correlation = 10.3%



Upper Salary

correlation = 10.7%



Findings and suggestion in terms of company's score & salary

Overall Score is the average of the scores below:

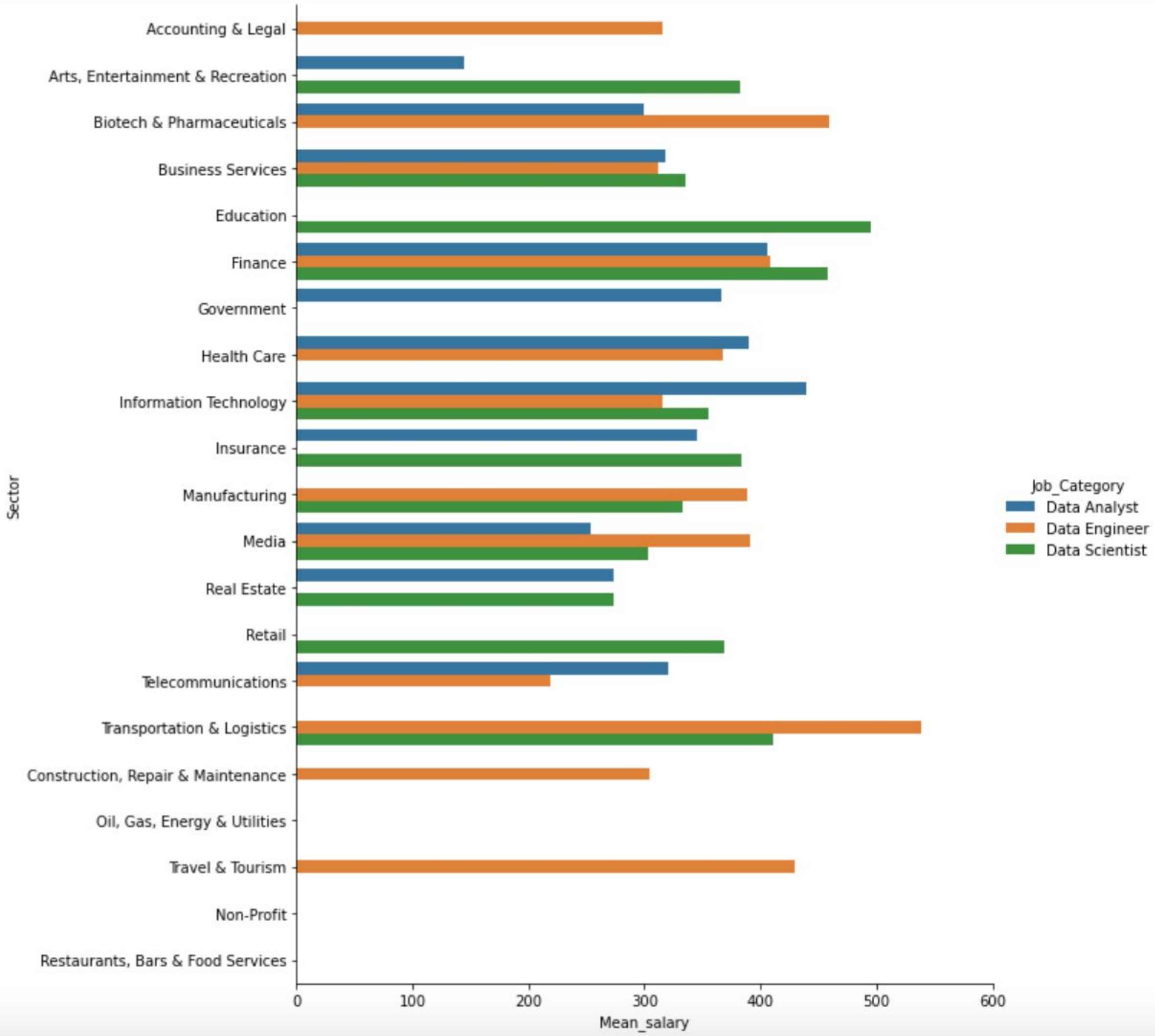
1) Comp & Benefits
2) Culture & Values
3) Career Opportunities
4) Work/Life Balance
5) Senior Management

For **middle-size companies**:

- They pay the lowest salary and score the lowest overall score from employees
- Given no obvious correlation between salary and overall score, our suggestions for them in order to improve the above-mentioned factors to gain higher score from employees are:
  → Retain talents in the company
  → Reduce the turnover cost

ANALYSIS

Mean Salary Analysis


Accounting & Legal
Arts, Entertainment & Recreation
Biotech & Pharmaceuticals
Business Services
Education
Finance
Government
Health Care
Information Technology
Insurance
Manufacturing
Media
Real Estate
Retail
Telecommunications
Transportation & Logistics
Construction, Repair & Maintenance
Oil, Gas, Energy & Utilities
Travel & Tourism
Non-Profit
Restaurants, Bars & Food Services
Sector
0
100
200
300
400
500
600
Mean_salary
Job_Category
Data Analyst
Data Engineer
Data Scientist

Where can you get the **highest salary**?

	Data Analyst	Data Engineer	Data Scientist
Sector 1	**Information Technology** HKD 440,000	**Transportation & Logistics** HKD 538,100	**Education** HKD 495,500
Sector 2	**Finance** HKD 406,667	**Biotech & Pharmaceuticals** HKD 458,833	**Finance** HKD 458,458
Sector 3	**Health Care** HKD 390,000	**Travel & Tourism** HKD 430,000	**Transportation & Logistics** HKD 411,167

Where to consider avoiding?

*But we **don't** condone you to entirely avoid these sectors

	Data Analyst	Data Engineer	Data Scientist
Sector 1	**Arts, Entertainment, & Recreation** HKD 144,000 ???	**Telecommunications** HKD 219,000	**Real Estate** HKD 273,000
Sector 2	**Media** HKD 253,000	**Construction** HKD 304,300	**Media** HKD 303,000
Sector 3	**Biotech & Pharmaceuticals** HKD 273,000	**Business Services** HKD 311,938	**Manufacturing** HKD 332,500

CHALLENGES & NEXT STEPS

Moving Forward

	Data Availability	Static Analysis	Incomplete Data
Challenges	Out of **1,022** scraped job postings, only **35.2%** of them have salary information	The analysis is conducted based on particular data that was extracted in one single timeframe, that is **20 June 2021 afternoon**.	Not enough information regarding required experience level for each job, hard to analyse correlation between **experience level** & **mean salary**
Next Steps	Use a more reliable resource with particular emphasis on salary data, e.g. *Payscale*	Apply automation / machine- centred operation to enable real-time tracking, perhaps **machine learning**?	Likely apply machine learning to more accurately classify job category and seniority of each posting

CONCLUSION

Our Takeaways

Project

- High-quality analysis is not possible without high-quality data
  - Scrape data with correct extraction methodologies
  - Data has to be cleaned up properly… and it takes **a lot of time**

- Write your codes clearly
  - After hundreds of lines, codes become hard to follow
  - Ensure code writing is easy to trace and understand
  - Create variables with relevant names

Our Takeaways

Findings

- Overall score of a company doesn't correlate to salary
  - Corporates should stop thinking they can make their employees happy by offering higher salary
  - Focus more on culture/values, senior management, work/life balance, advancement opportunities, etc.

- Data-oriented roles are becoming more important in Hong Kong
  - Jan '21 (Xccelerate): **HKD 192,000** (entry) → **HKD 420,000** (senior)
  - Jun '21 (our analysis): **HKD 265,000** (entry) → **HKD 444,000** (senior)

Q&A