Journal of Quality Measurement and Analysis *JQMA* 21(4) 2025, 355-372
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*
https://doi.org/10.17576/jqma.2104.2025.18

UNVEILING SKILL–INDUSTRY ASSOCIATIONS IN MALAYSIA'S DATA PROFESSIONAL EMPLOYMENT MARKET

(*Analisis Perkaitan Kemahiran dan Industri Terhadap Pasaran Pekerjaan Profesional Data di Malaysia*)

JACQUELINE LOW YUN ZHI, ROHAYU MOHD SALLEH*, FONG LI XUAN & VINCENT LIM KANG CHIEN

ABSTRACT

Digital transformation and the shift to remote work accelerated Malaysia's entry into the Big Data Analytic (BDA) world. Rising demand for data-related jobs contrasts with limited workforce and unclear role-specific knowledge. This study analyzes the job growth of four data professional roles in Malaysia by using correspondence analysis to identify industry involvement and Apriori algorithm to determine the high-proficiency skills preferred by employers. The data were collected from Malaysian job advertisements on LinkedIn, JobStreet, and Indeed between 15 April and 15 October 2024, revealing sector-specific demand with uniquely preferred skills for each role. Data Analysts are highly sought in Services and Retail/F&B/Hospitality, requiring Python, Natural Language Preprocessing (NLP), Tableau, Excel, strong decision-making and organizational skills, alongside Warehouse Management Systems (WMS) and SAP Customer Data Platform expertise. Data Scientists are in demand in Science and Sales/Marketing sectors, emphasizing R Studio, Tableau, machine learning (ML) algorithms, and creativity. Data Engineers are preferred in Computer/IT and Healthcare, needing Python, NLP, ML algorithms, organizational skills, and cloud platforms Amazon Web Service (AWS) & Google Cloud Platform (GCP). Database Administrators are required in Admin/HR and Healthcare, with skills in NLP, Excel, Gorubi, decision-making, organizational skills, and systems like Cloud Formation, Snowflake, Databricks, WMS, and SAP Customer Data Platform. Generally, data professional job growth in Malaysia is developing rapidly but clear tasks boundaries are needed in bridging the gap between industry and institution for skill development in future career advancement.

Keywords: correspondence analysis; Apriori algorithm; data analyst; data scientist; data engineer; database administrator

ABSTRAK

Transformasi pesat ke arah dunia digital dan kesan selepas peralihan kepada bekerja dari rumah telah mempercepatkan kemasukan Malaysia ke alam Analisis Data Raya (BDA). Permintaan terhadap pekerjaan berkaitan data semakin meningkat, namun tenaga kerja terhad dan kekurangan pengetahuan mengenai peranan pekerjaan. Kajian ini menganalisis pertumbuhan pekerjaan bagi empat peranan profesional data di Malaysia dengan menggunakan analisis koresponden untuk mengenal pasti penglibatan industri manakala algoritma Apriori digunakan untuk menentukan kemahiran berkecekapan tinggi yang diutamakan oleh majikan. Data diarkibkan daripada iklan pekerjaan di Malaysia melalui platform *LinkedIn, JobStreet*, dan *Indeed* dari 15 April 2024 hingga 15 Oktober 2024, yang menunjukkan permintaan spesifik sektor dengan kemahiran unik bagi setiap peranan. Kemahiran profesional data menunjuk permintaan tinggi dalam sektor Perkhidmatan dan Runcit/*F&B*/Hospitaliti memerlukan kemahiran *Python, NLP*, *Tableau, Excel*, kemahiran membuat keputusan dan organisasi, serta kepakaran *WMS* dan *SAP Customer Data Platform*. Sektor Sains dan Jualan/Pemasaran memerlukan saintis data serta *R Studio*, *Tableau*, algoritma pembelajaran mesin, dan kreativiti. Jurutera data diperlukan dalam sektor *IT* dan Kesihatan, memerlukan kemahiran *Python*, *NLP*,

algoritma pembelajaran mesin, kemahiran organisasi, serta platform awan *AWS* dan *GCP*. Pentadbir pangkalan data diperlukan dalam sektor *Admin/HR* dan Kesihatan, dengan kemahiran dalam *NLP*, *Excel*, *Gurobi*, kemahiran membuat keputusan dan organisasi, serta sistem seperti *Cloud Formation*, *Snowflake*, *Databricks*, *WMS*, dan *SAP Customer Data Platform*. Secara keseluruhan, pertumbuhan pekerjaan profesional data di Malaysia berkembang pesat, namun sempadan tugas yang jelas diperlukan untuk merapatkan jurang antara industri dan institusi, serta menyokong pembangunan kemahiran untuk kemajuan kerjaya masa hadapan.

Kata kunci: analisis koresponden; algoritma Apriori; penganalisis data; saintis data; jurutera data; dan pentadbir pangkalan data

1. Introduction

In the rapid transformation into digitalization in Malaysia, the shifting development causes the demand for experts in data to skyrocket, yet the categories of data professionals and their duties remain unspecific. Data professionals are a group of people with the ability to work with data constantly where they investigate, arrange, manage, review, analyze, present, supervise, and protect data to make the greatest possible use in generating information (Kim 2016). The evolution to data-driven science has culminated in an urgent lack of data professionals and standard procedures to address the challenges associated with information development and data administration, collection, as well as analysis.

Despite the evolution of Big Data Analytic (BDA), the technology landscape has shifted dramatically during the previous decade. This transformational ability to collect, sift through, and comprehend massive and intricate datasets has enabled businesses to discover insights that were previously unattainable. Consequently, the centralized database system of National Big Data Analytics Centre would be generated and pushing Malaysia towards BDA market. The demand of data professionals steadily increases hence as transformation into BDA market is a must for Malaysia to be more statistical analyzed with data information. However, the employment scope for each data-related position is not indicated in the Malaysian job market, leading to situations where one is compensated for a data analyst title yet continues with data engineering work.

Job advertisements on platforms like Indeed, JobStreet, and LinkedIn show confusion in job roles, with some posts listing tasks for data scientists, data engineers, and machine learning analysts under the title of data analyst, making it hard to differentiate roles and leading companies to assume data analysts can handle all tasks. Thus, when an employee's standards are not met, the corporation may hold them accountable for their failure to finish tasks on schedule.

Past research showed that data analyst, data scientist and business data analyst were the common top 5 popular job titles of data professionals since fourth quarter in 2020 as published by Institute of Labour Market Information and Analysis in 2023. The job advertisements provided the information needed for a specific career, mainly job descriptions, skills and academic requirements which are representing the employers' perspective of the expertise needed by job seekers (Kim & Angnakoon 2016). Research on skill gap in graduates' employability analyse the cross-sectional data collected through online questionnaires (Yong & Ling 2023). The job requirements and course descriptions in more than 3000 job advertisements from LinkedIn job websites were analysed (Behpour *et al.* 2019).

Building professionalism in data-related roles requires the development of professional, soft, and hard skills. For simplicity, fresh graduates and newbies of data analysts who are not

equipped with proper skills before stepping into the BDA environment will result in difficulty when solving the assigned task. Hence, the market demand for data professionals in Malaysia is urgently needed to be captured. The main aim of this study is to identify the preferences of the main industries that require data professionals and essential skills needed by a data professional in Malaysia. This study is supported by data mining techniques.

2. Materials and Methods

This chapter presents the research methodologies. Graphical charts visualize the data, Correspondence Analysis examines industry preferences for data professionals, and the Apriori algorithm identifies the skills required by employers in Malaysia.

2.1. *Data description*

The data for this study were collected from three major Malaysian job portals: JobStreet, LinkedIn, and Indeed. Advertisements were gathered over six-months (15 April–15 October 2024), with five daily postings selected, resulting in an average of 500–1,000 postings per day. The sample focused on four data professional roles, including data analyst, data scientist, data engineer, and database administrator (Fergus 2023), following Wilkins (2021) research framework. Data samples were obtained by registering accounts on each platform without filters to avoid bias. The keyword 'data professionals' was used to identify relevant job advertisements, and details from the first five ads were manually recorded in Excel based on the variables needed. Manual collection helped minimize misinterpretation of job descriptions and skills. The information extracted from the job ads is shown in Table 1.

In this study, irrelevant or incomplete variables such as company name, job portals and location were removed, the key features were extracted in data preprocessing process to increase model accuracy and performances (Sangaiah *et al.* 2018). The company collected from job ads has been classified into 12 industries involved by company respectively to identify the preferences of main industries that needed data professionals specifically Accounting/Finance, Admin/Human Resources (HR), Science, Art/Media/Communications, Building/ Construction, Computer/Information Technology (IT), Education/Training, Healthcare, Sales/Marketing, Manufacturing/Engineering, Retail/Food & Beverages/Hospitality and Services. Each skill and tool from the advertisements was recorded as a separate binary variable (1 = present, 0 = absent) and later grouped into six broader categories based on Wilkins's framework. Skills were initially flexible to capture unique, role-relevant data, but were ultimately found to align with the reference categories.

2.2. *Methods*

The methods included visualizations to represent variable relationships. Correspondence Analysis (CA) involved Chi-Squared tests and asymmetric biplots, while association rule mining used the Apriori algorithm with Support, Confidence, and Lift to identify skill associations. All analyses were conducted in R software.

2.2.1. *Data visualization*

From the data collected mentioned in the previous section, a radar chart was used to visualize the composition of skill categories for each data professional, illustrating the key information from a set of data (International Business Machines Corporation (IBM) 2024), particularly in CA and association rule mining.

Table 1: The variables collected from job advertisements

Category	Variable	Data Type	Descriptions
Company Details	Company Name	Qualitative (Nominal)	Name of company posted in each portal.
	Company Location	Qualitative (Nominal)	States of company located.
Job Specialization	Job Specialization	Qualitative (Nominal)	Field of job advertisement.
Job Descriptions	Job Title	Qualitative (Nominal)	Essential types of data professionals.
	Job Position	Qualitative (Nominal)	Level of job positioning hired.
Job Requirements	Job Location	Qualitative (Nominal)	Job location during employment.
	Minimum Academic Qualifications	Qualitative (Ordinal)	Minimum academic qualification required.
	Language Requirements	Qualitative (Nominal)	Stated language required.
	Working Experience	Quantitative (Intervals)	Working experience requirement.
Salary	Salary	Quantitative (Intervals)	Salary listed in job advertisements.
Skill Requirements	Programming	Qualitative (Nominal)	Programming skills or software stated in job advertisement.
	Visualization	Qualitative (Nominal)	Visualization skills or software stated in job advertisement.
	Statistical	Qualitative (Nominal)	Statistical skills or software stated in job advertisement.
	Soft Skills	Qualitative (Nominal)	Soft skills requirement stated in job advertisement.
	Cloud	Qualitative (Nominal)	Cloud skills or software stated in job advertisement.
	Databases	Qualitative (Nominal)	Database skills or software stated in job advertisement.

2.2.2 *Chi-squared test of independence*

The Chi-Squared Test of Independence was employed to assess whether a statistically significant association exists between job titles and job specialization, based on frequency data in a two-way contingency table (McHugh 2013). The test statistics of Pearson's Chi-Square test are shown in Eq. (1),

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{1}$$

where O represents the observed frequency and E the expected frequency. A greater discrepancy between these values yields a larger chi-square statistic. The analysis compares the resulting p-value against a significance threshold of 0.05. A p-value below this threshold indicates a meaningful relationship between job titles and job specialization. Conversely, a p-value above 0.05 suggests no statistically significant association.

2.2.3 *Correspondence analysis*

CA was used to reveal relationships between categorical variables via a graphical plot. Rows (job titles) were plotted in standard coordinates, and columns (job specializations) as supplementary points in an asymmetric biplot, illustrating patterns of association and

behavioral similarities. The original data n_{ij} in terms of rows and columns in the bilinear CA model shown in Eq. (2),

$$p_{ij} = r_i c_j \left(1 + \sum_{k=1}^{K} \sqrt{\lambda_k} \phi_{ik} \gamma_{jk}\right) \tag{2}$$

where p_{ij} is the relative proportions of n_{ij}/n, where n is grand total of $\sum i \sum j n_{ij}$; $r_i = \sum_{j=1}^{J} p_{ij}$ is a row mass; $c_j = \sum_{i=1}^{I} p_{ij}$ is column mass, λ_k is the k[th] principal inertia, ϕ_{ik} is row standard and γ_{jk} is column standard coordinates.

This configuration preserves the true distances and relationships among job titles, while the positions of job specializations indicate their relative association with the job titles rather than with each other. The standard coordinates of both variables were extracted to complement the visual interpretation, and Euclidean distances between each job title and job specialization were calculated using the vector notation and shown in Eq (3).

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3}$$

where D is the distance between two objects in CA space by minimizing the sum of squared differences between the coordinates (Ward 1963). In this study, x denotes the coordinates of data professional roles and y represents job specialization. This numerical measure quantifies proximity, providing an objective basis for associations observed in the biplot.

2.2.4. *Association rule mining*

Association rule mining is finding the association between two itemsets, where one will affect another itemset in terms of frequency and counts (Yuan & Ding 2012), such as market basket analysis to identify frequent itemsets in the customer's shopping basket (Xie 2021). This method identified association rules between data professional roles and skill categories, evaluated using Lift, Support, and Confidence metrics with varying thresholds to determine significant itemsets.

Support (S), a fraction of any itemset that contains item A and item B union in a dataset (D). In this study, Support was used to identify the association of each data professional who needed the combination of skill A and skill B shown in Eq. (4)

$$S(A \Rightarrow B) = \frac{n(A \cap B)}{D} = P(A \cap B) \tag{4}$$

where $n(A \cap B)$ is the number of advertisements requiring A and B skills; D is the total number of advertisements collected; $P(A \cap B)$ is probability of advertisements requiring A and B skills. The minimum support threshold (*minsup*) value is designed in accepting the Support of the paired itemset in percentage (Cheng & Xiong 2010; Verma *et al.* 2014).

Confidence (C) computed the probability of a targeted itemset (consequent) being detected, provided that there is another itemset (antecedent) had already occurred in the circumstances. In this study, Confidence is useful to identify the confident occurrence of skill B when skill A is required in data professionals' career in Eq. (5),

$$C(A \Rightarrow B) = P(B|A) = \frac{P(A \cap B)}{P(B)} \tag{5}$$

where $P(B|A)$ is the conditional probability of skill B requiring based on skill A. The minimum confidence threshold (*minconf*) value is the threshold for accepting the Confidence of the paired itemset. The strong associations are interpreted when any association rule satisfied both *minsup* and *minconf* (Verma *et al.* 2014).

Lift determines the existence of association rules similarly to significance test, identify the existing association rule that follows a positive or negative relationship (Berry & Linoff 2004; Larose 2004; Zhang & Zhang 2002). In this study, Lift was used to check whether the proposed itemset is a significant association rule ranges positively, indicating a positive association using Eq. (6),

$$L(A \Rightarrow B) = \frac{P(A \cap B)}{P(A)P(B)} = \frac{S(A \Rightarrow B)}{P(A)P(B)} = \frac{C(A \Rightarrow B)}{P(A)} \quad (6)$$

where $P(A \cap B)$ is the probability of advertisements requiring A and B skills; $C(A \Rightarrow B)$is the confidence of the association of B when A occurred. When Lift equals 1, indicating skill A does not affect the occurrence of skill B. When Lift is less than one, A and B occur together less often than expected.

The Apriori algorithm speeds up the flow of the original algorithm with all possible combinations of association rules of antecedent (A) and consequent (B) (Verma *et al.* 2014). It is based on a large dataset regarding the data professionals job market, with all needed skills extracted from job ads. The minimum threshold value of Support and Confidence based on the average utility value divided by the total existing transactions (Hikmawati *et al.* 2021). This method has been supported by two algorithms, Apriori and the Frequent Pattern (FP) Growth algorithm, eligible to determine the threshold value based on dataset characteristics.

In this study, the *minsup* and *minconf* were pre-set during each skill analysis. Any association rule that does not achieve the threshold value of support and confidence might not truly reflect the employer's perspective.

3. Results and Discussions

This section outlines the results obtained based on the job advertisements collected involving job specializations, job titles, and skills requirements. The results of the analysis are discussed with the supported interpretations and explanations, with summary to conclude this section.

Figure 1 depicts skill category preferences by data professional roles, with six consolidated categories. Min-max scaling was applied based on each professional's total skills in a category, using the total sum calculated for each professional within each category. In the employers' preferences, data analyst found to be expected to excel in programming, visualization, statistical, and soft skills, with cloud skills less critical as it is not the main skills required by data analyst. Data scientist requires databases and programming skills more than visualization skills as they carry out extraction and transformation of data into interpretable information, which do not necessarily need visualization skills.

Data engineer shows a high percentage at cloud skill followed by database as the big data framework undoubtedly works between database and cloud computing. Alternatively, visualization skill is not significantly affecting the role of data engineer in data pipelines for efficient data flow. Database administrators are not interpretable due to the output not significantly reflecting the database administrator role as they were less than 10% of the total job ads collected. The radar chart for the database administrator role appears noticeably smaller compared to other job titles, as only 128 job advertisements were collected out of the total 2760 advertisements analyzed.

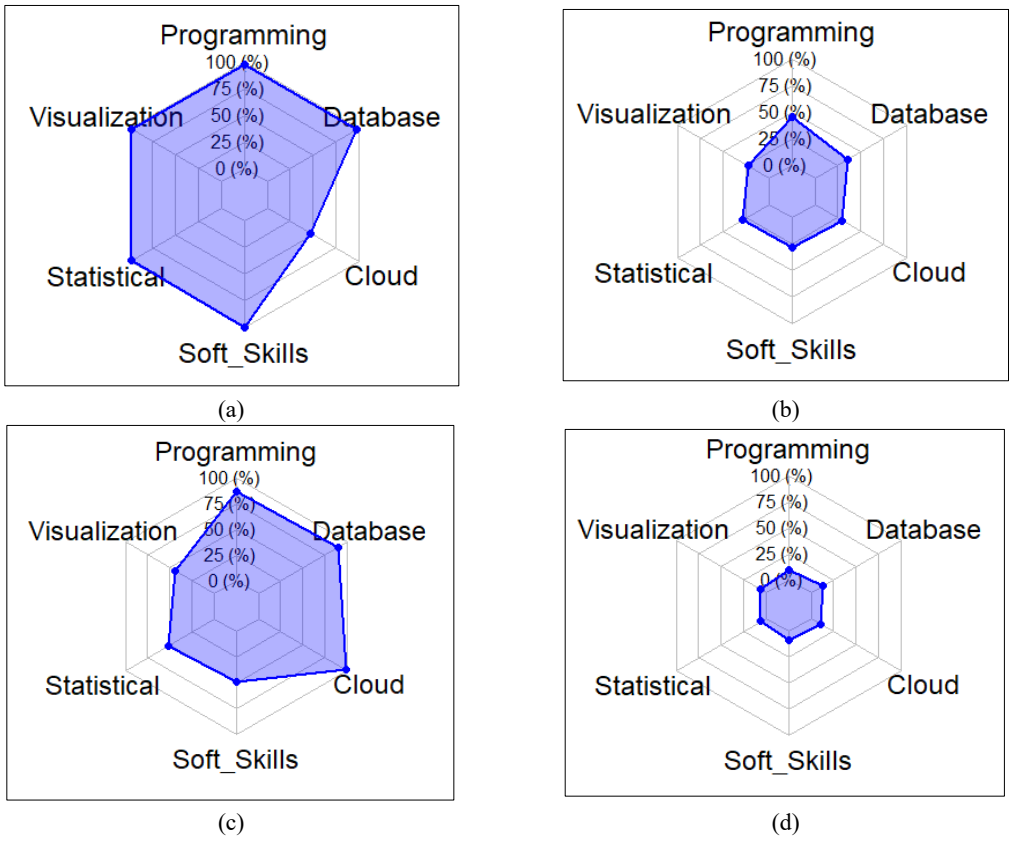

Figure 1: Skill categories preferences of (a) Data Analyst, (b) Data Scientist, (c) Data Engineer, (d) Database Administrator

3.1. *Association between job specialization and job titles*

The Chi-Square Test of Independence is constructed to analyze the relationship between job specialization and employees' job titles. The results are shown in Table 2.

Table 2: Chi-square test of independence results between job specialization and data professionals job titles

Pearson's Chi-Square Test	χ^2	df	$p-value$
Results	425.56	33	0.0000

The Pearson Chi-Square statistic of 425.56 is found to be statistically significant, as evidenced by a p-value below the conventional threshold of 0.05. The output concludes that there is an association between job specialization and data professionals' job titles collected from job advertisements, indicating that the correspondence analysis can be carried out. A scree plot is constructed to visualize the percentage explains that variance of each dimension in Figure 2.

Figure 2 shows an elbow point after the second dimension suggesting that the first two dimensions are sufficient to capture the main patterns in the data. Dimension 1 (Dim 1) shows 61.9% while dimension 2 (Dim 2) shows 28.7% of the variation in the data indicating that 90.6% of the biplots retains the first two dimensions (Dim 1 and Dim 2). The first two

dimensions (Dim 1 and Dim 2) show 90.6% of the variability in Figure 2, capturing most of the dataset's information. Therefore, a two-dimensional (2D) asymmetric biplot focusing on data professional job titles was generated in Figure 3.

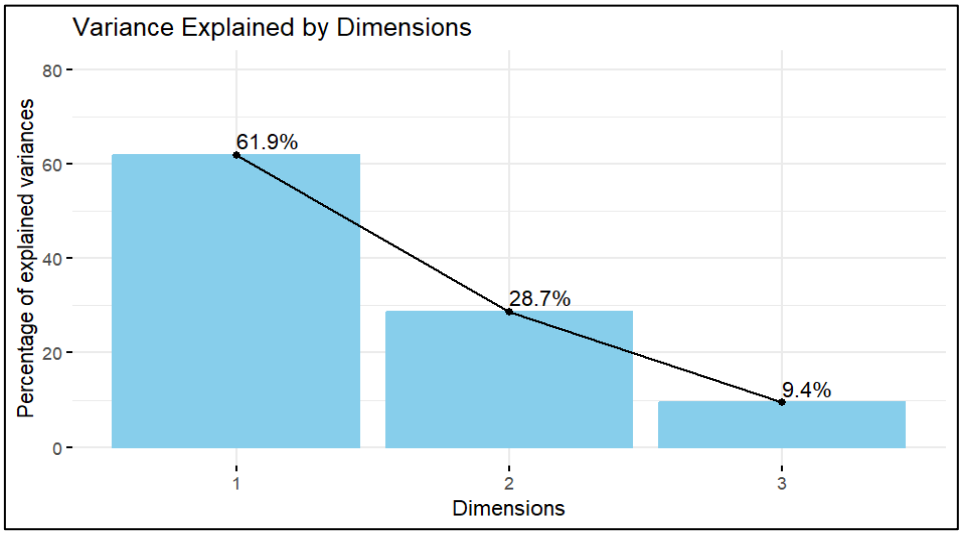

Figure 2: Percentage of explained variances for each dimension

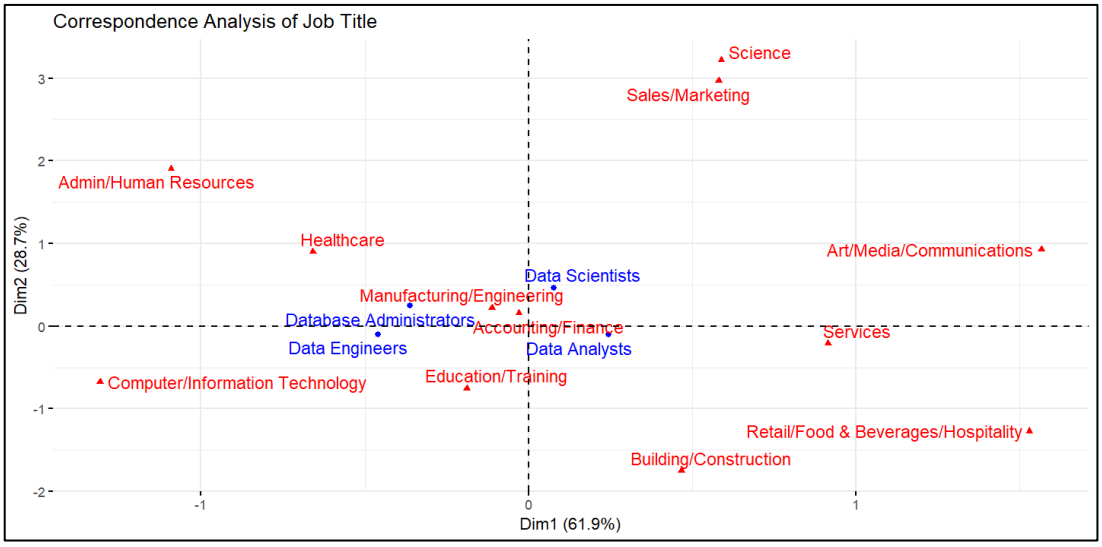

Figure 3: Asymmetric biplot of the correspondence analysis

Figure 3 shows the asymmetric biplot of the correspondence analysis. The relative positions of job specializations indicate their association with the job titles, but distances between specializations are not directly comparable. Visually, data scientist is positioned in the positive region while data analyst located on the negative region contributing to the Dim 1 in alternative magnitude, indicating contrasting association patterns. Science, Sales/Marketing, and Art/Media/Communications are positioned in the same positive region of Dim 1 as Data Scientists, suggesting potential associations. Conversely, Retail/F&B/Hospitality, Services, and Building/Construction appear closer in orientation to Data Analysts in the negative region of

Dim 1. In Dim 2, Database Administrators are positioned above Data Engineers, with apparent associations to Admin/HR, Healthcare, Manufacturing/Engineering and Accounting/Finance, while Data Engineers appear closer to Computer/IT and Education/Training. However, the visual inspection of the asymmetric biplot alone does not provide a sufficiently clear picture to make definitive conclusions about the strength of these associations.

Hence, Euclidean distances were calculated between each job title and job specialization based on their standard coordinates to verify and refine these visual interpretations. The calculated Euclidean Distance Matrix is presented in Table 3a, Table 3b and Table 3c.

Table 3a: The Euclidean distance matrix of 48 points

Job Title	Accounting/ Finance	Admin/ Human Resources	Art/Media/ Communications	Building/ Construction
Data Analyst	0.2882	0.7702	0.3838	0.2815
Data Scientist	0.4390	0.4189	0.4882	0.8346
Data Engineer	0.4686	0.5141	0.9875	0.6606
Database Administrator	0.4134	**0.1553**	0.8485	0.7976

Table 3b: The Euclidean distance matrix of 48 points

Job Title	Computer/ Information Technology	Education/ Training	Healthcare	Manufacturing/ Engineering
Data Analyst	0.6492	0.3067	0.5352	0.3170
Data Scientist	0.7736	0.6362	0.3922	0.4314
Data Engineer	**0.0696**	0.4050	**0.3850**	0.4486
Database Administrator	0.3918	0.5071	**0.1696**	0.3846

Table 3c: The Euclidean distance matrix of 48 points

Job Title	Retail/Food & Beverages/ Hospitality	Sales/Marketing	Science	Service
Data Analyst	**0.2805**	0.7313	0.7838	**0.0722**
Data Scientist	0.8308	**0.1901**	**0.2373**	0.5479
Data Engineer	0.9466	0.9641	1.0057	0.7435
Database Administrator	0.9811	0.6599	0.6932	0.7082

Data Scientists are positioned in the positive region alongside Science, Sales/Marketing, and Art/Media/Communications. Science (0.2373) and Sales/Marketing (0.1901) are the closest, while Art/Media/Communications (0.4882) is comparatively further apart. This indicates that Science and Sales/Marketing share stronger skillset overlapping with Data Scientists, particularly in statistical modelling, predictive analytics, and applying data-driven insights for strategic decision-making. Similarly, data analyst is most closely aligned with Services (0.0722) and Retail/F&B/Hospitality (0.2805), with Building/Construction (0.2815) also showing a moderate connection on Dim 1. Dim 1 can be understood to explain a shift which emphasizes a specific area of work that had an employee specializing in programming, statistics, and to some extent visualization techniques implying that industries which focused on providing business intelligence relied on the two roles of data scientists and data analysts.

In Dim 2, database administrator contributes positively as compared to data engineer. The job specialization that contributes the highest variability associated with database administrator are Admin/HR (0.1553) and Healthcare (0.1696). Data engineer contributes the most in

Computer/IT (0.0696) and Healthcare (0.3850). These two regions are suggesting that the job specialization in Dim 2 requires more specified technical skills specially in database and cloud computing skills in data infrastructure and management. The remaining specialization such as Accounting/Finance (0.2882), Building/Construction (0.2815), Education/Training (0.3067), and Manufacturing/Engineering (0.3170) are found to be moderately close to Data Analyst as compared to other professional roles. The result shows that the job specializations as aligned with the major role of each data professional.

3.2. *Association rule mining*

By applying the Apriori algorithm to identify associations between professional job titles and skill requirements derived from job advertisements, association rule mining successfully addressed the second research objective. The skills were grouped into six main categories, and analyzed by Support (S) and Confidence (C) measures, revealing common associations and patterns. In the network charts, larger circle nodes indicate higher confidence, while darker colors represent higher Lift values.

The rules were generated in R software with minimum thresholds of Support ≥ 0.05 and Confidence ≥ 0.60, selected to balance statistical significance and practical interpretability. Specifically, the support threshold ensures each rule represents at least 5% of job postings, reducing noise from rare skill combinations; the confidence threshold reflects a strong conditional probability that the consequent skill will occur given the antecedent. The Lift threshold highlights positive associations where co-occurrence is at least 1% higher than expected under independence due to the inability of the sample size to represent the entire population in Malaysia. For categories with more than ten rules, only the top ten (ranked by Lift value) are displayed to avoid redundancy and emphasize the most significant associations. The resulting network diagrams and detailed tables are used to present the rules associated with each job category across various data professional roles, as discussed in the subsequent subtopics.

3.2.1. *Programming skill*

Figure 4 shows the top 10 strongest association rules between programming skills in the data analyst, data scientist, data engineer, and database administrator. Meanwhile, Table 4 summarizes the representative rule for each role, selected based on the highest Lift value. For data analyst skills shown in Figure 4(a), Scikit-learn emerges as the most consistent co-occurring skill (Lift = 2.0433), closely linked with Python, NumPy, Pandas, and NLP, while NumPy and Pandas also show strong associations (Support ≈ 0.49, Lift = 2.00), confirming their importance with Scikit-learn. Figure 4(b) reveals data scientist skill, NumPy, Pandas, and Scikit-learn form the core programming tools (Lift = 1.4807, Confidence = 1.0000), frequently listed together in job ads, whereas R Studio appears only in Rule 7-10 (Support ≈ 0.33), making it supplementary rather than essential. Skillset for data engineers in Figure 4(c) depicted NumPy and Pandas show a near-universal overlap (Lift = 1.3484, Confidence = 1.0000), while Scikit-learn consistently appears with Python and either library (Lift = 1.3533, Confidence = 1.0000), confirming the demand for both data manipulation and applied machine learning skills, with high support values (0.6533–0.7416) reinforcing their prevalence. For database administrators shown in Figure 4(d), Scikit-learn, Pandas, and NumPy co-occur with perfect Confidence (1.0000) and high Lift (2.51), marking them as essential, while the presence of NLP with Scikit-learn and Pandas highlights the value employers place on advanced text-processing capabilities. It can be concluded that NumPy, Pandas and Scikit-learn are the most essential programming skills for all roles.

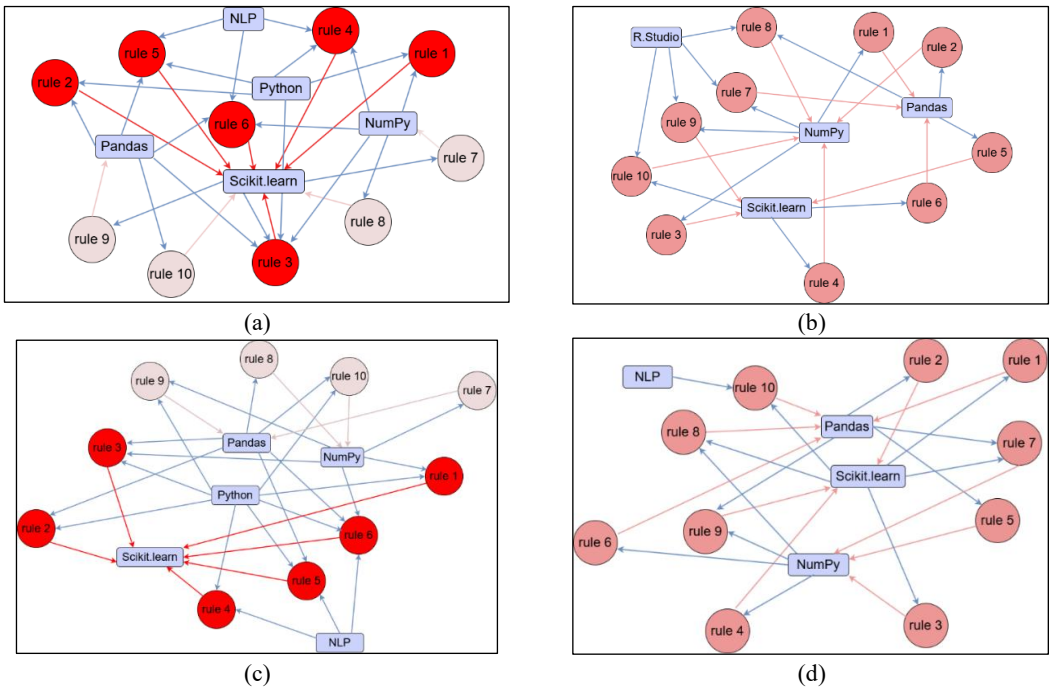

Figure 4**:** Network graph of association rules between programming skills in (a) Data Analyst; (b) Data Scientist; (c) Data Engineer; (d) Database Administrator

Table 4: Strongest association rules of programming skills in each role job advertisement

Job	Rule	LHS	RHS	Support	Confidence	Lift	Count
DA	1	{Python, NumPy}	{Scikit.learn}	0.3985	1.0000	2.0433	583
DS	1	{Scikit.learn}	{Pandas}	0.6754	1.0000	1.4807	285
DE	1	{Python, NumPy}	{Scikit.learn}	0.6693	1.0000	1.3533	500
DBA	1	{PowerBI}	{Data.Visualization}	0.3359	0.8600	1.3424	43

3.2.2. *Visualization skill*

Figure 5 illustrates the association rules related to visualization skills across the specified data professional roles. Table 5, meanwhile, presents the representative rule for each role, selected based on the highest Lift value. Data visualization is a key competency for data analysts displayed in Figure 5(a). Tableau and Power BI are strongly associated, co-occurring in 75.5% of postings with a Lift of 2.00, showing they appear together twice as often as expected.

General visualization skills strongly predict Power BI (Confidence = 0.93), though the low Lift (0.38) indicates the link is common but not exclusive. Power BI also implies visualization skills in 73.2% of cases, while visualization overall appears in 55.3% of postings, confirming its central role in analyst requirements. For data scientist shown in Figure 5(b), Tableau and Power BI also co-occur strongly (Confidence = 93.1%, Lift = 2.05), and Tableau paired with general visualization still predicts Power BI (Confidence = 89.7%), showing employers expect both principles and tool expertise. Visualization appears in 80.1% of data engineers' postings, shown in Figure 5(c), with Power BI closely tied to visualization skills (Confidence = 95.6%, Lift almost 1.2), underscoring growing expectations beyond system management to include effective data communication. Figure 5(d) outlines the required skills for database administrator

with Power BI is strongly associated with visualization (Confidence = 86%, Lift = 1.34), reflecting its role in making database outputs more accessible for business decision-making. However, data engineer and database administrator show less association rules with similar itemsets for visualization skills.

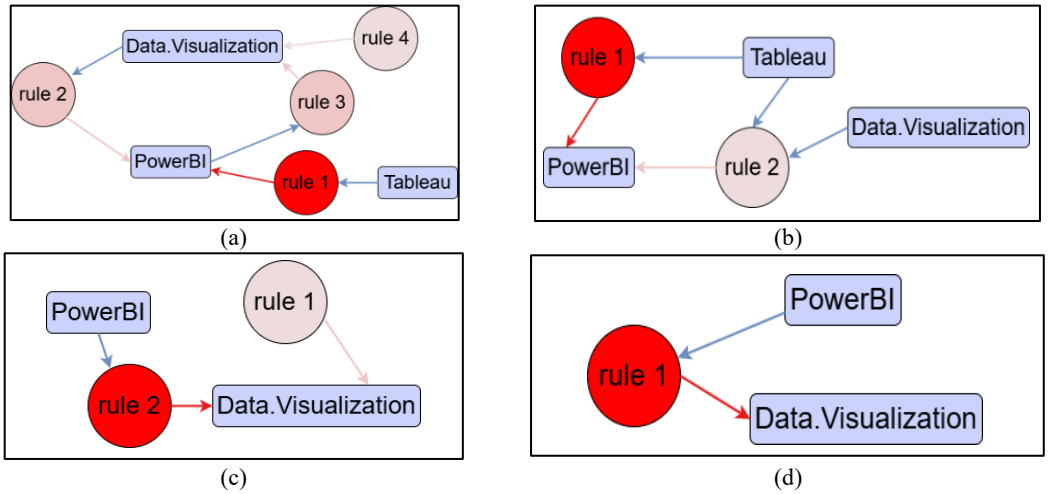

Figure 5: Network graph of association rules between visualization skills in (a) Data Analyst; (b) Data Scientist; (c) Data Engineer; (d) Database Administrator

Table 5: Strongest association rules of visualization skills in each role job advertisements

Job	Rule	LHS	RHS	Support	Confidence	Lift	Count
DA	1	{Tableau}	{PowerBI}	0.7553	0.7553	2	1105
DS	1	{Tableau}	{PowerBI}	0.2227	0.9307	2.0456	94
DE	2	{PowerBI}	{Data.Visualization}	0.3775	0.9559	1.1941	282
DBA	1	{PowerBI}	{Data.Visualization}	0.3359	0.8600	1.3424	43

3.2.3. *Statistical skill*

Figure 6 illustrates the association rules related to statistical skills across the four data professional roles, supported by the representative rules with the highest Lift values presented in Table 6. The data analyst skillset shown in Figure 6(a) characterizes optimization and data mining as co-occurring in 52.0% of postings, with high Confidence (0.82–0.89) and above-random Lift (1.36–1.45), indicating that they are often paired, especially when reinforced by Excel.

Machine learning also predicts data mining (Confidence = 0.85, Support = 30.2%), though it appears less frequently overall. Skillset illustrated in Figure 6(b) describes that data mining and optimization co-occur in 50.9% of data scientist's postings (Confidence = 0.82–0.84, Lift = 1.35), indicating both are core requirements. When combined with machine learning, their mutual prediction strengthens (Confidence up to 87.9%, Lift = 1.44), highlighting stronger-than-chance associations. Among data engineers shown in Figure 6(c), machine learning and optimization almost always imply data mining. Data mining dominates about 80.1%, occurrence acting as a baseline skill closely tied to optimization and machine learning occurred more than 50% postings (Confidence = 0.90–0.97, Lift = 1.12–1.21). Figure 6(d) shows that

optimization emerges as the key statistical skill for database administrators, appearing with Gurobi at 100% confidence (Lift = 2.17). The strong optimization linked with data mining individually or paired with Microsoft Excel shows confidence at 88% and 96% respectively. The Lift values (1.90–2.17) indicate that optimization is a distinguishing requirement compared to general postings.

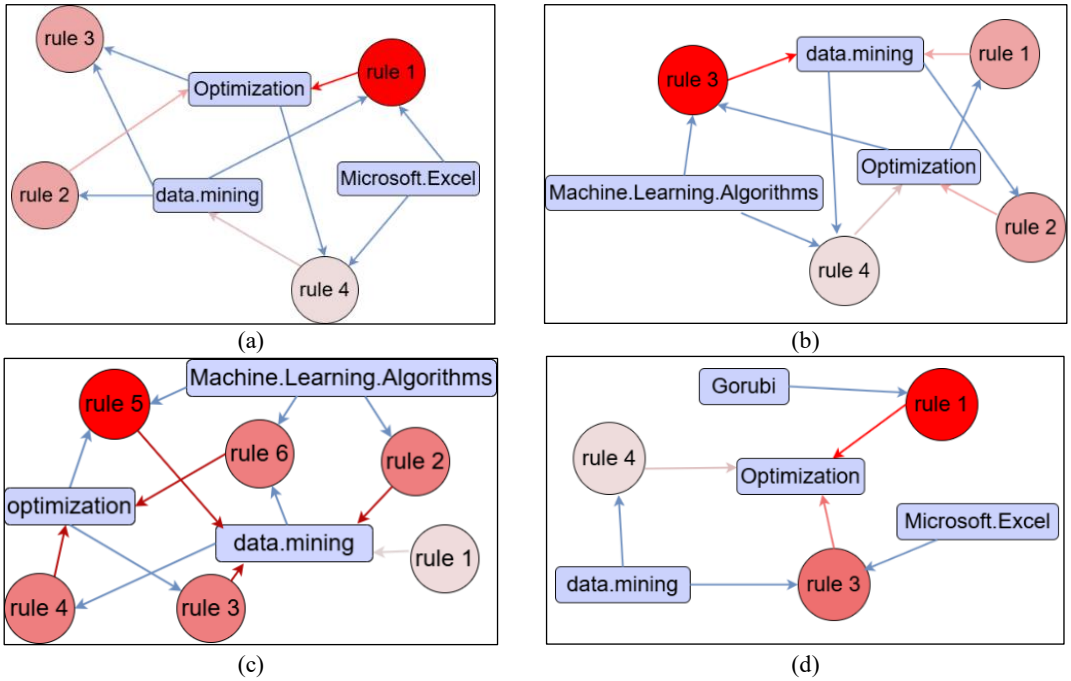

Figure 6: Network graph of association rules between statistical skills in (a) Data Analyst; (b) Data Scientist; (c) Data Engineer; (d) Database Administrator

Table 6: Strongest association rules of statistical skills in each role job advertisements

Job	Rule	LHS	RHS	Support	Confidence	Lift	Count
DA	1	{data.mining, Microsoft.Excel}	{optimization}	0.4033	0.8538	1.4542	590
DS	3	{Machine.Learning.Algorithms,optimization}	{data.mining}	0.3460	0.8795	1.4442	146
DE	5	{Machine.Learning.Algorithms,optimization}	{data.mining}	0.5114	0.9720	1.2142	382
DBA	1	{Gorubi}	{optimization}	0.1094	1.0000	2.1695	14

3.2.4. *Soft skill*

Figure 7 shows the top 10 strongest association rules between soft skills of all roles, supported by the representative rules with the highest Lift values presented in Table 7. The association rules in Figure 7(a) display that organizational skills emerging as a key requirement with critical thinking and decision making supported by 60% of job postings and more than 95% confidence highlighted their frequent co-occurrence with positive association under independence. Problem-solving, communication, and analytical skills also frequently appear alongside organizational skills. Data analysts are expected to combine cognitive skills with

practical task management to perform efficiently in their roles. For data scientist, critical thinking appearing in over 97% of cases when creative is listed. Problem-solving, analytical skills, communication, creativity and organizational abilities shown 95% confidence in the co-occurrence when critical thinking prioritized, although these associations show 4% higher than expected under independence. Critical thinking, creative solutions provider and communication skills are expected in data scientist. Figure 7(c) revealed data engineer's skillset that critical thinking is a core soft appearing in 96% job postings where problem-solving and communication skills are strongly associated with 98% confidence and slightly frequent occurrence than chance. Organizational skills, analytical skills, and decision-making appear alongside critical thinking, reflecting the importance of structured reasoning, collaboration and problem-solving abilities in data engineer. The association rules in Figure 7(d) show that critical thinking is expected whenever decision-making and organizational skills are required confidently around 95% and supported by above 60% job postings. Problem-solving, communication, and analytical skills appear consistently across the top rules, with 100% confidence in some cases. The positive association suggests all associations co-occur more often than expected by chance.

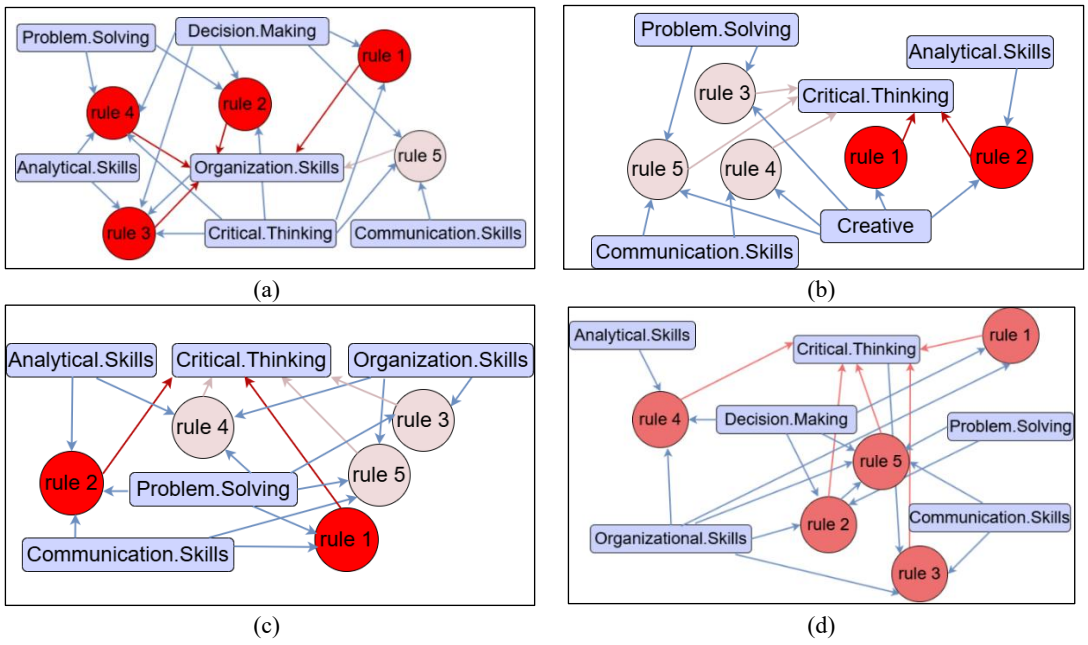

Figure 7: Network graph of association rules between soft skills in (a) Data Analyst; (b) Data Scientist; (c) Data Engineer; (d) Database Administrator

Table 7: Strongest association rules of soft skills in each role job advertisements

Job	Rule	LHS	RHS	Support	Confidence	Lift	Count
DA	1	{Critical.Thinking, Decision.Making}	{Organizational. Skills}	0.6220	0.9609	1.0523	910
DS	1	{Creative}	{Critical. Thinking}	0.3934	0.9708	1.0397	166
DE	1	{Problem.Solving, Communication.Skills}	{Critical. Thinking}	0.9692	0.9904	1.0219	724
DBA	1	{Decision.Making, Organizational.Skills}	{Critical. Thinking}	0.6328	0.9529	1.1294	81

3.2.5. *Cloud skill*

Figure 8 illustrates the cloud skills required for each data professional role, while Table 8 presents the representative rules with the highest Lift values. Azure and Cloud skills confidently appear together in only 12.3% of data analyst job postings but high lift value at 6.04 stronger than expected under independence. This relatively low proportion suggests that cloud expertise is not a core requirement for all data analyst positions. Similarly, data scientists found Cloud implying Azure in 97.1% of cases and support reaching 31.5%, though with a lower Lift (2.13) due to Azure's broader baseline demand. Cloud skills are clearly important for Data Engineers with only the top 10 strongest rules shown in Figure 8(c) with AWS, Azure, and GCP frequently required together. AWS emerges as the most common complementary skill, appearing in up to 95% of postings when paired with Azure and GCP. The highest Support (52.9%) is seen when Azure and AWS imply Cloud, reflecting that data engineer roles typically demand proficiency across multiple cloud platforms. Cloud-related skills frequently appear in combination for database administrator, particularly Snowflake, Databricks, and Cloud Formation, along with Azure and general Cloud skills. Although these skill pairs occur in less than 20% of job postings (Support 0.10–0.20), they show high confidence values above 80%, indicating that when one skill is required, the others are very likely to be needed as well as shown in Figure 8(d). The strong co-occurrence of these skills is further highlighted by Lift value above 3. Snowflake, Databricks, and Cloud Formation display tighter internal connections, while Azure and Cloud appear more independent. Database administrators are expected to be proficient in specialized cloud platforms, with certain tools strongly associated together, whereas more general cloud knowledge may be required separately.

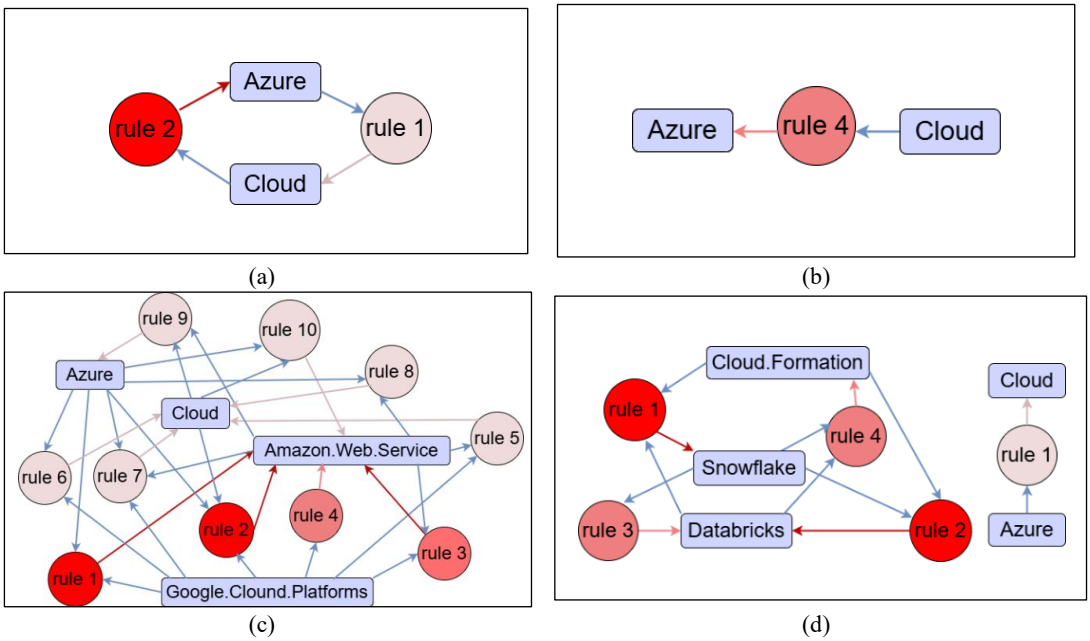

Figure 8: Network graph of association rules between cloud skills in (a) Data Analyst; (b) Data Scientist; (c) Data Engineer; (d) Database Administrator

Table 8: Strongest association rules of cloud skills in each role job advertisements

Job	Rule	LHS	RHS	Support	Confidence	Lift	Count
DA	2	{Cloud}	{Azure}	0.1230	0.7965	6.0374	180
DS	4	{Cloud}	{Azure}	0.3152	0.9708	2.1337	133
DE	1	{Azure,Google.Cloud. Platforms}	{Amazon. Web.Service}	0.3066	0.9502	1.6023	229
DBA	1	{Databricks, Cloud.Formation}	{Snowflake}	0.1563	0.9524	5.3002	20

3.2.6. *Database skill*

Data Model & Pipelines emerges as the core database skill depicted in Figure 9(a) for data analyst, strongly linked by occurrence of SAP Customer Data Platform, Warehouse Management System, and SQL up to 90% confidence with at least 10% increase over independence, highlighting its role in integrated data tasks. Data scientist is found to have SQL and Data Model & Pipelines each appear in over 70% of postings shown in Figure 9(b), but their co-occurrence is weaker (Support 50.5%, Confidence 70.8%, Lift <1), suggesting employers typically prioritize one over the other depending on context. Data engineer shown in Figure 9(c) has the strongest reliance on Data Model & Pipelines (Support = 85.8%, Confidence = 85.8%) when SQL as the consequent skill with 92.5% confidence, but only 71.0% in the reverse, positioning pipelines as the baseline skill. The most associated skills for database administrators are Data Modeling and Pipelines, Warehouse Management Systems, SAP Customer Data Platforms, and SQL showing the interrelated demand in Figure 9(d) with Lift <1 signifying positive association and above 70% confidence of tendency to co-occur when either skill as antecedent, despite support values being slightly above 0.3. Together, these findings show pipelines as central for analysts and engineers, SQL as flexible for scientists, and broader multi-tool integration for administrators.

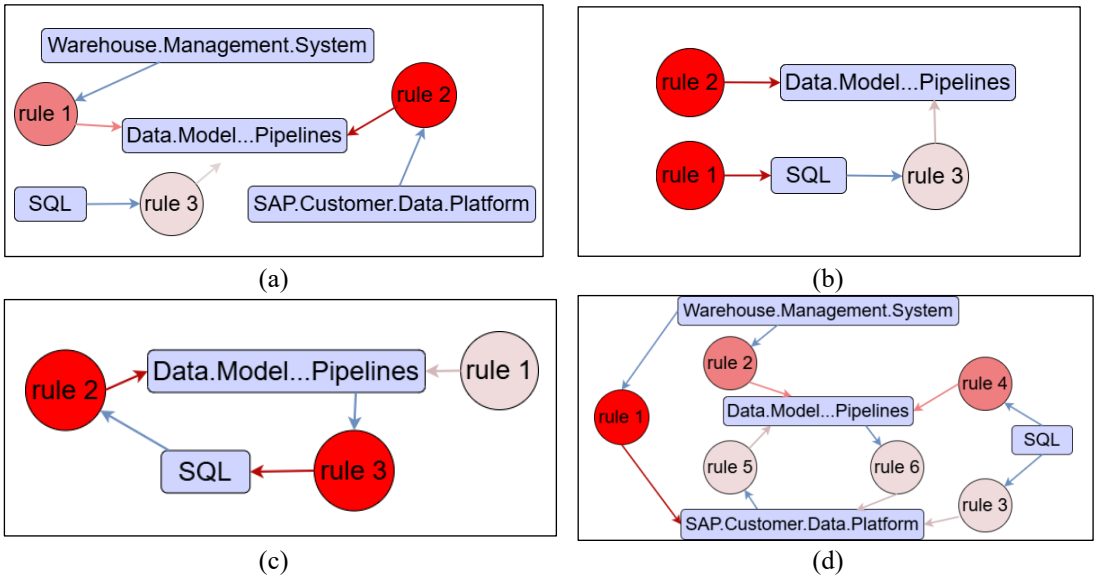

Figure 9: Network graph of association rules between database skills in (a) Data Analyst; (b) Data Scientist; (c) Data Engineer; (d) Database Administrator

Table 9: Strongest association rules of database skills in each role job advertisements

Job	Rule	LHS	RHS	Support	Confidence	Lift	Count
DA	2	{SAP.Customer. Data.Platform}	{Data.Model... Pipelines}	0.1408	0.8957	1.3117	206
DS	1	{}	{SQL}	0.7133	0.7133	1.0000	301
DE	2	{SQL}	{Data.Model... Pipelines }	0.6091	0.9248	1.0777	455
DBA	1	{Warehouse. Management.System }	{ SAP.Customer. Data.Platform }	0.3516	0.8491	1.6221	45

4. Conclusion

There are four essential types of data professionals in Malaysia: data analyst, data scientist, data engineer and database administrator. This study demonstrates that the demand for data professionals in Malaysia varies by sector, with data analysts sought in Services and Retail/F&B/Hospitality, data scientists in Science and Sales/Marketing, data engineers in Computer/IT and Healthcare, and database administrators in Admin/HR and Healthcare. The data professionals in Malaysia share several core skills such as NumPy, Pandas, Scikit-learn for programming, data mining and optimization for statistics, Power BI for visualization, and fundamental soft skills. Cloud platforms, particularly Azure, are essential across all roles, with Data Engineers and Database Administrators requiring more complex cloud skills. Data Model & Pipeline is a critical database skill common to all roles. Although there is overlap in skill requirements, each role is distinguished by specialized knowledge demanded by employers where specific roles require additional specialized competencies. Data analysts, data scientists, data engineers, and database administrators exhibit sector-specific preferences, emphasizing the importance of Python, NLP, machine learning algorithms, visualization tools, cloud platforms, and organizational skills. The findings underscore critical gaps in workforce readiness and role-specific knowledge, emphasizing for clear role definitions. Strengthening collaboration between industry and educational institutions is essential to develop relevant skills, ensuring that future professionals are better aligning workforce capabilities for career advancement in the evolving Big Data Analytics landscape.

Acknowledgments

The authors would like to thank the anonymous reviewers for their thorough assessment and constructive comments. Their expertise and valuable insights have greatly enhanced the quality and clarity of the content.

References

Behpour S., Hawamdeh S. & Gourarzi A. 2019. Employer's Perspective on data science; Analysis of job requirement & course description. *Proceedings of the Association for Library and Information Science Education Annual Conference: ALISE 2019*, pp. 177–182.

Berry M.J.A. & Linoff G. 2004. *Data Mining Techniques: For Marketing, Sales, and Customer Support*. 2nd Ed. New York, NY: John Wiley & Sons.

Cheng Y. & Xiong Y. 2010. Research and improvement of apriori algorithm for association rules. *Proceedings of the 2010 2nd International Workshop on Intelligent Systems and Applications*, pp. 1–4.

Institute of Labour Market Information and Analysis. 2023. Data Professionals Job Market Insights 2020 to 2023. Institute of Labour Market Information and Analysis. https://www.ilmia.gov.my/index.php/en/bda-jmi-2020-q4 (1 May 2025).

Fergus S. 2023. 6 Essential Types of Data Professionals. Shipyardapp Blog. https://www.shipyardapp.com/blog/data-professionals/ (15 April 2025).

Hikmawati E., Maulidevi N.U. & Surendro K. 2021. Minimum threshold determination method based on dataset characteristics in association rule mining. *Journal of Big Data* **8**(1): 146.
International Business Machines Corporation. 2024. What is data visualization? IBM. https://www.ibm.com/think/topics/data-visualization#:~:text=IBM-,What%20is%20data%20visualization%3F,that%20is%20easy%20to%20understand.
Kim J. 2016. Who is teaching data: Meeting the demand for data professionals. *Journal of Education for Library and Information Science Online* **57**(2): 161–173.
Kim J. & Angnakoon P. 2016. Research using job advertisements: A methodological assessment. *Library & Information Science Research* **38**(4): 327–335.
Larose D.T. 2004. *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, NJ: John Wiley & Sons.
McHugh M.L. 2013. The chi-square test of independence. *Biochemia Medica* **23**(2): 143–149.
Sangaiah A.K., Zhang Z. & Sheng M. Eds. 2018. *Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications*. London: Academic Press.
Verma A., Khan S.D., Maiti J. & Krishna O.B. 2014. Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. *Safety Science* **70**: 89–98.
Ward J.H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**(301): 236–244.
Wilkins D.E. 2021. An in-depth analysis of the data analytics job market. Senior Honors Thesis. University of New Hampshire.
Xie H. 2021. Research and case analysis of apriori algorithm based on mining frequent item-sets. *Open Journal of Social Sciences* **9**(4): 458–468.
Yong B.P.P. & Ling Y.-L. 2023. Skills Gap: The importance of soft skills in graduate employability as perceived by employers and graduates. *Online Journal for TVET Practitioners* **8**(1): 25–42.
Yuan J. & Ding S. 2012. Research and improvement on association rule algorithm based on FP-growth. *Proceedings of the International Conference on Web Information Systems and Mining*, pp. 306–313.
Zhang C. & Zhang S. Eds. 2002. *Association Rule Mining: Models and Algorithms*. *Lecture Notes in Computer Science, Vol. 2307*. Berlin, Heidelberg: Springer.

Department of Mathematics and Statistics
Faculty of Applied Sciences and Technology
Universiti Tun Hussein Onn Malaysia
Kampus Cawangan Pagoh
Hab Pendidikan Tinggi Pagoh
KM 1, Jalan Panchor
86400 Pagoh, Muar
Johor, MALAYSIA
E-mail: jacquelineyunzhi@yahoo.com, msrohayu@uthm.edu.my[*]*, lixuan053@gmail.com, vincentlim0221@gmail.com*

Received: 9 June 2025
Accepted: 24 August 2025

[*]Corresponding author