LONDON METROPOLITAN UNIVERSITY

MSc Data Analytics Project

UK Data Job Insights 2025

An Analysis of In-Demand Skills, Job Trends, and Salary Expectations

May 2025

Kiran Correya

Student ID: 23027046

Email ID: KIC0211@my.londonmet.ac.uk

Supervised by

Dr. Subeksha Shrestha

Acknowledgments

I would like to express my sincere gratitude to Dr. Subeksha Shrestha, my tutor, for her exceptional guidance, continuous support, and invaluable feedback throughout this project. Her expertise, constructive criticism, and thoughtful suggestions have been instrumental in refining my work and pushing me to think critically at every stage of the research. I truly appreciate her unwavering encouragement, which has been crucial to the success of this dissertation.

On a personal note, I am happy I chose to solve this problem, as it has provided me with an opportunity to apply my skills to a real-world issue, furthering my growth as an aspiring data analyst. The challenges I encountered throughout this journey have been rewarding, and I am grateful for the knowledge and experience gained in the process.

Lastly, I would like to thank my family and friends for their unwavering support and belief in me during this endeavor.

Abstract

This study explores the demand for data professionals in the UK by analysing over 6,300 job postings from Reed.co.uk during March 2025. Using a custom pipeline integrating the Reed API and Selenium, roles such as Data Analyst, Data Scientist, Power BI Analyst, and Machine Learning Engineer were extracted and enriched with details including job titles, locations, salaries, and descriptions. A keyword matching technique, based on a curated list of technical skills, was applied to analyse job descriptions, offering interpretable and consistent results. Salaries were standardised to annual equivalents, and geographic data was normalised to city-level granularity. The findings, visualised through an interactive Power BI dashboard, reveal trends in skill demand, salary expectations, and regional hiring patterns. This real-time, data driven approach provides valuable insights for job seekers, educators, and recruiters navigating the evolving UK data job market.

Contents

List of Figures

List of Tables

Listings

Chapter 1

Introduction

1.1 Background

The proliferation of data in recent years has profoundly transformed the global economy. We now operate in an era where decisions—both strategic and operational—are increasingly driven by the availability and analysis of data. Across sectors such as healthcare, finance, retail, logistics, and the public sector, data has become central to how organisations innovate, compete, and deliver services. As a result, the importance of roles dedicated to extracting insights from data—such as Data Analysts, Data Scientists, Machine Learning Engineers, and Business Intelligence Developers—has grown exponentially.

In the United Kingdom, this demand is particularly pronounced. With the government prioritising digital transformation and businesses accelerating their adoption of technologies such as artificial intelligence, machine learning, and big data platforms, the need for qualified data professionals is reaching unprecedented levels. According to recent reports by the UK Department for Digital, Culture, Media and Sport (DCMS), there remains a persistent shortage of skilled data workers, a gap that is costing the UK economy billions in unrealised potential. Furthermore, this gap is not limited to niche technical roles—it spans across business functions, from strategic planning and marketing to customer analytics and operations.

Despite the visibility of this demand, one of the most pressing challenges in the UK's data labour market is a lack of clarity and consistency in how data roles are defined. Titles such as "Data Analyst," "Data Scientist," or "AI Specialist" are often used interchangeably, even though the actual responsibilities, required competencies, and salary expectations vary considerably across postings and organisations. This inconsistency presents significant difficulties for job seekers who must interpret what a given job title entails based solely on unstructured and often ambiguous job descriptions.

The issue is compounded for graduates and early-career professionals. Without a clear, data-informed understanding of what skills are in demand, which tools are considered

foundational, and how expectations shift across roles and regions, they risk misaligning their training and preparation. Likewise, academic institutions face growing pressure to align their curricula with evolving industry needs but often lack access to real-time, granular data on what employers are actually asking for in job listings.

This study aims to address these critical information gaps. By leveraging real-time job scraping and structured text analysis, the research examines over 6,000 job postings from Reed.co.uk to systematically extract, classify, and visualise skill requirements, job roles, salary expectations, and geographic trends within the UK data job market in 2025. The findings offer a data-driven foundation for improving decision-making among key stakeholders in the ecosystem: students, educators, employers, and policy advisors.

1.2 Research Motivation and Problem Identification

The motivation for this project stems from both personal experience and observed challenges within the data employment landscape. While exploring job opportunities in early 2025, I encountered several recurring issues that pointed to a broader systemic inconsistency in how data roles are communicated through job advertisements.

One of the most noticeable patterns was the mismatch between job titles and skill requirements. For instance, positions labelled as "Data Analyst" often included tasks and expectations typically associated with more advanced roles like "Data Scientist" or "Machine Learning Engineer"—including requirements for TensorFlow, PyTorch, or deep learning experience. Conversely, some roles designated as "Data Scientist" listed only basic tools such as Excel, SQL, or Python, with little mention of statistical modelling or algorithm development.

This inconsistency extended to salary ranges as well. Roles with similar titles and geographic locations offered vastly different salary bands, with no apparent explanation regarding the variation. One Data Analyst role might advertise £30,000 per year, while another—seemingly identical—offered up to £70,000. These disparities suggest a lack of standardisation not only in how roles are titled and described but also in how they are valued by employers.

In addition to the confusion experienced by job seekers, these inconsistencies pose challenges for other stakeholders. Educators and curriculum developers cannot confidently align teaching with job market expectations if those expectations are ambiguous or shifting. Employers, in turn, face difficulty attracting the right talent if job descriptions fail to communicate role expectations clearly. Recruiters may also struggle to accurately screen candidates when job postings are vague or overly broad.

These experiences highlighted a critical need for a systematic, empirical examination of the UK data job market—one that moves beyond anecdotal evidence and isolated job descriptions to provide a comprehensive, real-time overview of how roles are defined, what

skills are in demand, and how those elements connect to salary and geography.

Therefore, this study seeks to answer the following research question:

> *What are the most in demand technical and analytical skills for data analytics and data science roles in the UK job market, and how do these skill requirements correlate with salary expectations and job opportunities?*

By answering this question, the study aims to support evidence-based career planning, curriculum design, and recruitment practices within the rapidly evolving data economy of the UK.

1.3 Aims and Objectives

The principal aim of this research is to identify, categorise, and interpret the technical and analytical skills most frequently requested in UK-based job postings for data roles. In doing so, the project also seeks to explore how these skills align with role definitions, compensation structures, and regional demand. Through this comprehensive analysis, the research hopes to reduce ambiguity in the data labour market and provide actionable intelligence to those navigating or shaping it.

To achieve this aim, the study is guided by the following objectives:

Skill Extraction and Categorisation: Develop a curated keyword list representing key technical competencies, and use this to extract structured insights from unstructured job descriptions.

Job Role Mapping: Associate identified skills with specific job titles (e.g., Data Analyst, Data Scientist, Data Engineer) and evaluate how skill requirements vary by role.

Demand Quantification: Calculate the number of postings linked to each skill and role to determine which competencies are most in demand and how they are distributed across the labour market.

Salary Normalisation and Analysis: Convert all salary figures—whether hourly, daily, monthly, or annually—into standardised annual equivalents to enable accurate comparison across roles, regions, and industries.

Geographic Distribution Analysis: Normalise inconsistent or incomplete location data using postcode matching and city classification, then analyse spatial demand trends at the city level.

Data Preparation and Transformation: Apply a structured data cleaning pipeline to address missing values, inconsistent formats, and multi-source integration issues, ensuring a high-quality dataset for analysis.

Interactive Insight Generation: Create a Power BI dashboard that visualises findings across multiple dimensions (role, skill, location, salary) and enables stakeholders to explore the data through interactive filters and visual tools.

Each of these objectives contributes to the overarching goal of improving transparency in the UK data job market and equipping decision-makers with reliable, actionable intelligence.

1.4 Report Structure

This dissertation is structured into six chapters, each serving a unique purpose in the overall research process, building upon each previous section to create a comprehensive and coherent analytical journey.

Chapter 1: Introduction sets the stage by presenting the research context and identifying the knowledge gap that the study seeks to address. It outlines the motivations behind the research, clearly defining the key aims and objectives that guide the investigation.

Chapter 2: Literature Review explores existing academic and industry research related to data skills, job market analytics, and the methodologies used for labor market visualization. This chapter positions the current study within the broader academic discourse, providing insights into the existing literature and emphasizing the gaps that this dissertation aims to fill.

Chapter 3: Methodology dives into the details of the data sources, the scraping pipeline, and the preprocessing steps employed in the study. It also elaborates on the analytical models used, along with the techniques involved in constructing the final dashboard. The chapter reflects on the rationale behind each methodological choice, ensuring that readers understand the approach's strengths and limitations.

Chapter 4: Results and Discussion presents the key findings of the analysis, which include skill frequency tables, salary trends, role-skill co-occurrence networks, and geographic demand maps. This chapter not only reports the results but also provides a discussion of these findings in relation to the research question and the objectives outlined earlier.

Chapter 5: Conclusion, Recommendations, and Future Work concludes the dissertation by summarizing the project's contributions. It draws key conclusions from the analysis and offers targeted recommendations for various stakeholders, including students, educators, and hiring managers. This chapter also identifies potential directions for future research and suggests ways to improve the methodology, thereby extending the study's scope and impact.

Chapter 6: Limitations provides a reflective overview of the study's constraints, such as the selection of data sources, the assumptions made during salary standardization,

and the logic applied to role categorization. It acknowledges how these limitations might influence the interpretation of the results, offering transparency regarding the study's scope.

Each chapter has been designed to provide a structured, empirical, and stakeholder-oriented analysis of the UK data job market in 2025. The aim is to ensure that the dissertation is clear, relevant, and practically applicable, offering insights that can guide future research and industry practices.

Chapter 2

Literature Review

The rising demand for data professionals has prompted a wide range of empirical and methodological investigations into job market expectations, skill gaps, and hiring trends. This review focuses on both academic and industry-led studies that examine the landscape of data-related roles, particularly with respect to skill requirements, salary trends, and the effectiveness of analytical methods used in market assessment. The reviewed works span diverse geographies, methodological approaches, and platforms—providing a foundational context against which this project's contribution can be assessed.

Wilkins (2021) conducted a comprehensive analysis of 12,748 job postings for data-related roles, aiming to identify core competencies in demand and explore salary relationships. The study employed Excel-based processing techniques for keyword extraction and trend identification. While the large sample size provided valuable insights, the reliance on manual processing methods limited the scalability and adaptability of the analysis. This approach made it challenging to update the dataset in real-time, which is crucial for capturing the rapidly evolving dynamics of the data job market.Additionally, the static nature of the dataset reduced the study's relevance to current job market trends, as it lacked automated data collection processes. The absence of advanced text analytics and data visualization techniques further constrained the study's ability to provide interactive, real-time insights into labor market trends.[1]

Patacsil and Acosta (2021) investigated the skill requirements in the Philippine IT job market by applying association rule mining techniques. Specifically, they utilized the FP-Growth algorithm to identify frequent skill pairings and employed TF-IDF to assess word significance across job postings. This methodological approach effectively uncovered co-occurring skill sets and thematic patterns, offering valuable insights into skill relationships within job descriptions. However, the study faced several limitations that hindered its generalizability and scalability. Firstly, the research relied on manual data ingestion, which introduced potential biases and inefficiencies in data processing. As the job postings were manually sourced, the approach lacked the scalability and adaptability

required for continuously monitoring the dynamic job market. Additionally, the study's relatively narrow sample scope, which was focused on the Philippine IT sector, limited the generalizability of the findings to other regions or broader sectors. Furthermore, the study did not incorporate salary data, role-level categorization, or regional differentiation, which are crucial for understanding the economic value of skills and how they vary across different job roles or geographic locations. The absence of these elements means the study provided a more limited view of the market and lacked the depth required for drawing actionable, context-specific conclusions. These methodological constraints are important to consider when evaluating the applicability of Patacsil and Acosta's findings to other job markets or regions.[2]

The 365 Data Science (2025) report analysed 1,355 U.S.-based data analyst job postings sourced from Glassdoor with the objective of identifying prevalent technical skills and employer expectations. The methodology involved manual keyword extraction and frequency analysis to compile a list of high-demand competencies. While the study provided a useful snapshot of the skills sought by employers in the U.S. context, its analytical depth was limited by the lack of automation and reliance on a single job board. Furthermore, the exclusion of salary data and the absence of visualisation tools reduced its utility for comparative or strategic workforce planning. The report, while informative, ultimately offered a static and partial view of a dynamic labour market.[3]

Verma et al. (2019) conducted a structured content analysis of 1,235 job advertisements to investigate skill requirements across various data-related roles. The study employed a predefined classification framework to categorise technical and analytical skills and utilised pairwise comparisons to uncover role-based differences in employer expectations. This methodological structure added clarity to the skill taxonomy and highlighted distinctions across job titles. However, the study did not incorporate salary analysis, geographic segmentation, or visualisation techniques, limiting its capacity to assess labour market value or regional demand. While the research offered useful categorisation, its practical application for job seekers or educators was constrained by its static presentation and lack of contextual depth.[4]

Digital Skills and Careers UK Government (2024) commissioned a national policy report to examine the state of data and digital skills within the workforce, drawing on large-scale survey data and employer feedback. The study aimed to identify gaps between industry needs and current proficiency levels across sectors. While the breadth of the survey ensured extensive coverage, the report relied primarily on self-reported perceptions rather than direct analysis of job postings. This introduced potential response bias and reduced empirical precision. Furthermore, the report underutilised data visualisation and lacked role-specific or skill-level granularity, limiting its effectiveness for operational workforce planning or curriculum design. Despite its policy relevance, the study offered only a high-level overview rather than actionable, role-based insights.17

The Data Analyst Job Landscape (2024) examined 1,091 entry-level data analyst job postings in the United States using text mining and clustering algorithms. Its primary aim was to identify frequently requested technical skills and group them into thematic clusters for role segmentation. The methodology was computationally robust, and the clustering approach allowed for nuanced insights into employer expectations. However, the study lacked integration with salary data and did not explore geographic or industry-based variations, which limited its utility for economic or regional workforce planning. Additionally, the absence of interactive visualisation tools reduced accessibility and interpretability for non-technical stakeholders, such as job seekers or educational planners.[5]

Data Science Skills in the UK Workforce (2023) synthesised existing research and stakeholder perspectives on the state of data science skills within the UK workforce. Drawing on secondary sources, expert interviews, and employer consultations, the study provided a macro-level overview of skills demand, training gaps, and future workforce needs. While its breadth made it suitable for strategic policy discussions, the absence of primary data collection—particularly from job postings—limited the empirical strength of its findings. The lack of granular role-level analysis and minimal use of data visualisation further constrained its practical applicability for operational planning or curriculum design.[6]

Jaiswal et al. (2024) investigated the alignment between UK university AI curricula and the technical skills demanded in the job market. The study employed automated data scraping to collect job descriptions and applied a Naive Bayes classifier to perform textual analysis and classify skill categories. This use of machine learning added methodological sophistication and provided a scalable framework for analysing employer expectations. However, the study did not include salary data or explore regional hiring patterns, limiting its ability to assess the economic value of specific skills. Additionally, the absence of interactive visual outputs reduced the accessibility of the findings for wider educational and professional audiences. Despite these limitations, the study makes a valuable contribution to curriculum reform by highlighting critical gaps between academic training and industry requirements.[7]

Ya (2022) conducted a personal project focused on the data analytics job market in San Diego, using job postings scraped from Indeed. The study employed Python and SQL for data processing and implemented statistical techniques such as bootstrapping and principal component analysis (PCA) to explore job trends and skill patterns. While the project demonstrated strong technical capability and methodological experimentation, its practical impact was limited by a small sample size and narrow geographic focus. Additionally, the study did not incorporate salary analysis or soft skill identification, both of which are critical for a comprehensive understanding of job market expectations. The absence of structured role segmentation and visual storytelling further constrained its generalisability and stakeholder relevance.[8]

Dawson et al. (2019) presented an adaptive methodology for identifying skill shortages using a large-scale dataset of over 6.7 million Australian job advertisements. Their approach leveraged skill similarity networks and dynamic indicators—such as posting frequency, salary, and job ad predictability—to detect high-demand data science and analytics (DSA) roles. A key strength of the study is its scalability and methodological innovation; unlike traditional static analyses, it employed automated occupation selection grounded in granular skill groupings, enhancing its responsiveness to evolving labour market demands. The study successfully demonstrated that highly technical DSA roles consistently experienced demand–supply imbalances, supported by elevated salaries and persistent posting activity. However, several limitations constrain its generalisability. Firstly, the geographic scope is restricted to the Australian job market, potentially limiting the applicability of its findings to other regions such as the UK or US, where industry structures and recruitment norms differ. Secondly, the study lacked role-level interpretability—while skills were effectively grouped, it was less clear how specific job titles evolved or overlapped. Furthermore, although the model inferred demand strength, it did not directly assess employer expectations or candidate qualifications, reducing its relevance for individual job seekers or curriculum designers. Despite these limitations, the study's integration of adaptive modeling techniques represents a valuable progression in the field of labour market analytics and aligns with the present project's emphasis on real-time, skills-based job market intelligence.[9]

Loloshahvar (2023) undertook a large-scale analysis of 25,114 data-related job postings, using Power BI to visualise trends in skill demand and role distribution. The study's strength lay in its dataset size, which offered strong representativeness for general observations. Power BI was effectively used for dashboard creation, providing stakeholders with a visual overview of job market trends. However, the study did not include salary information, nor did it segment the data by region, industry, or seniority level—factors crucial for contextualising skill value and job accessibility. Additionally, the dashboard lacked advanced interactivity, limiting user-driven exploration. While useful for initial market scanning, the analysis did not fully capitalise on the potential of its large dataset for targeted or strategic insight generation.[10]

Zhang (2024) presents a comprehensive exploration into the application of Natural Language Processing (NLP) techniques for analyzing job market trends. The study introduces innovative methodologies, including weak supervision for skill extraction and taxonomy-aware pre-training, to enhance the accuracy of information retrieval from job postings. A notable contribution is the development of a retrieval-augmented model that leverages multiple skill extraction datasets, aiming to improve overall performance in identifying relevant skills demanded in the labor market. However, the study does present certain limitations. The reliance on annotated datasets and complex NLP models necessitates substantial computational resources and expertise, which may not be

feasible for all researchers or institutions. Additionally, the focus on developing sophisticated models may overlook practical challenges in data collection and preprocessing, such as inconsistencies in job titles and descriptions across different platforms. Furthermore, while the study emphasizes the technical advancements in NLP applications, it provides limited discussion on the interpretability of the results for stakeholders like educators and policymakers.[11]

Huang et al. (2023) investigated the data analytics job market across Australia and New Zealand using text mining techniques to identify key technical competencies associated with various professional roles. The study successfully uncovered role-specific tool usage and skill preferences, contributing to regional workforce understanding. However, it did not incorporate salary analysis or soft skill assessment, both of which are essential for evaluating the holistic expectations of employers. Additionally, the study lacked geographic visualisation and interactive features, limiting its practical application for job seekers, educators, or policymakers aiming to align training with regional needs. While methodologically sound, its insights were constrained by the absence of multidimensional analysis.[12]

2.1 Comparative Analysis

To evaluate the methodological strengths and limitations of prior research, a comparative summary of 12 key studies is provided in Table 2.1. The table outlines core features such as data sources, analytical techniques, geographic focus, and notable omissions in each study. This overview sets the foundation for identifying recurring gaps that this dissertation aims to address.

The existing body of literature on data job market analysis exhibits several methodological and practical limitations that constrain their applicability for dynamic labor market intelligence. A common shortcoming across multiple studies is the reliance on manual or semi-automated data processing techniques. For instance, Wilkins (2021) utilized Excel-based methods for keyword extraction, which, while transparent, lack scalability and hinder the ability to update datasets in real-time. Similarly, the 365 Data Science (2025) report focused exclusively on job postings from a single platform, Glassdoor, and did not incorporate salary data or advanced visualization tools, limiting its utility for comprehensive workforce planning.

Advanced computational methods have been employed in studies like Patacsil and Acosta (2021) and Jaiswal et al. (2024), which utilized association rule mining and supervised classification algorithms, respectively. However, these studies often suffer from limited data granularity, lacking integration of salary analysis and regional or role-specific filtering. Consequently, their findings, while methodologically sophisticated, are

not always actionable for practitioners seeking detailed labor market insights.

Policy-level assessments, such as the UK Government's Digital Skills and Careers report (2024) and the Data Science Skills in the UK Workforce (2023), provide broad overviews of digital skill gaps but are primarily based on self-reported survey data. This reliance introduces potential biases and lacks the empirical grounding provided by direct job posting analyses. Moreover, these reports often omit role-level skill mapping, reducing their effectiveness for targeted labor market interventions.

Studies focusing on skill taxonomy development, like Verma et al. (2019), offer valuable categorizations of technical and analytical skills across job titles. However, the absence of salary benchmarking, geographic context, and temporal trend analysis in these studies limits their applicability for career planning or curriculum design.

Some investigations, such as those by Loloshahvar (2023) and Ya (2022), have introduced interactive dashboards and visualizations to present job market trends. Nevertheless, these studies often lack scale or analytical depth, and the absence of standardized salary data and role normalization undermines the reliability of their findings for comparative analysis.

Zhang (2024) presents a comprehensive exploration into the application of Natural Language Processing (NLP) techniques for analyzing job market trends, introducing innovative methodologies like weak supervision for skill extraction and taxonomy-aware pre-training. While these approaches enhance the accuracy of information retrieval from job postings, they require substantial computational resources and expertise, potentially limiting their feasibility for broader application. Additionally, the focus on sophisticated models may overlook practical challenges in data collection and preprocessing, such as inconsistencies in job titles and descriptions across different platforms.

Unlike prior works, this study makes a distinctive contribution by combining several innovative approaches. First, it employs real-time and automated data collection at scale, which stands in contrast to the static or manual data sources typically used in previous studies. Additionally, the study introduces a standardized salary transformation logic, enabling valid economic comparisons across various job types, ensuring consistency in the salary data. Another key feature of this study is the use of network-based visualisation to explore role–skill relationships, providing interactive insights into the multi-role competencies that are essential for a deeper understanding of job market dynamics. Furthermore, a two-stage stopword refinement process is applied to the text cleaning phase, improving the semantic clarity of word cloud outputs and enhancing the interpretability of the findings. Lastly, the study incorporates role-level filtering and slicing of visuals, allowing for targeted exploration by job seekers, educators, or other stakeholders, ensuring that the data can be examined from different perspectives to extract the most relevant insights.

In doing so, it moves beyond static summaries and offers a dynamic, stakeholder-

oriented approach to labour market analysis. It bridges the gap between descriptive job analytics and applied career intelligence, contributing both methodologically and practically to the field.

Table 2.1: Comparative Overview of Reviewed Literature on Data Job Market Analysis

Author(s) & Year	**Methodology / Techniques**	**Key Limitation**
Wilkins (2021)[1]	Manual keyword extraction, regression analysis	Static dataset, lacks automation and real-time updates
Patacsil & Acosta (2021)[2]	FP-Growth algorithm, TF-IDF analysis	Manual data collection, no salary or regional analysis
365 Data Science (2025)[3]	Manual keyword extraction	Single platform focus, lacks salary data and visualization
Verma et al. (2019)[4]	Skill classification, pairwise comparison	No salary or geographic segmentation, static presentation
Digital Skills and Careers UK Government (2024)	Quantitative survey analysis, policy review	Self-reported data, lacks empirical scraping and granularity
Data Analyst Job Landscape (2024)[5]	Text mining, clustering algorithms	No salary or regional breakdown, limited accessibility
Data Science Skills in the UK Workforce (2023)[6]	Policy synthesis	No primary data collection, lacks role-level analysis
Jaiswal et al. (2024)[7]	Naive Bayes classifier, automated scraping	No salary data, limited regional analysis, weak visualization
Ya (2022)[8]	SQL, bootstrapping, PCA	Small sample size, no salary or soft skill analysis
Dawson et al. (2019)[9]	Skill similarity networks, dynamic indicators	Geographic limitation to Australia, lacks role-level interpretability
Loloshahvar (2023)[10]	Power BI dashboard visualization	No salary data, limited segmentation by region or seniority
Zhang (2024)[11]	NLP techniques, weak supervision, taxonomy-aware pre-training	High computational requirements, limited result interpretability
Huang et al. (2023)[12]	Text mining techniques	Omits salary and soft skill analysis, lacks interactive features

Chapter 3

Methodology

This research adopts the *Data Analytics Life Cycle (DALC)* as its core methodological framework. The DALC provides a structured yet adaptable approach to managing data-driven projects, particularly those involving real-time data collection, exploratory analysis, and interactive visualisation. Its cyclical and iterative nature makes it well-suited for navigating the complexities of unstructured data processing, skill extraction, network modelling, and dashboard development.[13]

The following diagram illustrates the Data Analytics Life Cycle used in this study:

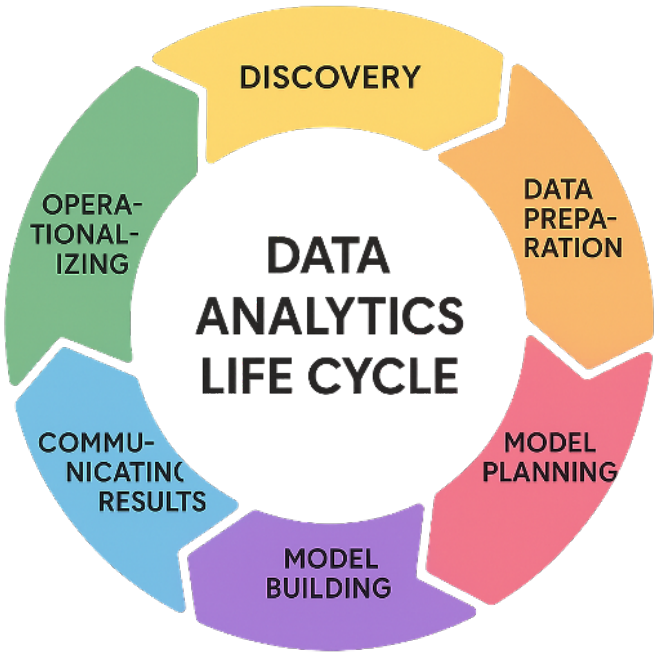


Figure 3.1: Data Analytics Life Cycle

The DALC consists of six interdependent phases: **Discovery**, **Data Preparation**, **Model Planning**, **Model Building**, **Communicating Results**, and **Operationaliz-**

ing. These stages collectively guide the project from initial problem formulation to the delivery of actionable insights and stakeholder-ready outputs.

3.1 Discovery

The central aim of this research is to identify the most in-demand technical and analytical skills for data analytics and data science roles within the UK job market. Specifically, the project seeks to examine how these skill requirements vary across job titles, geographic locations, and salary bands, using real-time job posting data. This problem is rooted in the persistent lack of transparency and standardisation in job advertisements, which often fail to clearly differentiate between roles such as "Data Analyst," "Data Scientist," and "Business Intelligence Analyst."

This ambiguity presents significant challenges for job seekers—particularly early-career professionals—navigating the job market without a clear understanding of employer expectations can result in skill mismatches and under-preparation. For educators and training providers, the absence of real-time industry feedback complicates efforts to align curricula with workforce needs. For employers and recruiters, unclear role definitions can hinder effective candidate assessment and job targeting.

In the context of 2025's rapidly evolving employment landscape, driven by continued digitisation and data integration across sectors, this research is both timely and essential. By providing a data-driven, real-time snapshot of market expectations, the study aims to support more informed career planning, curriculum development, and recruitment strategy formulation.

3.2 Data Preparation

This phase of the project focuses on obtaining, cleaning, and transforming data to ensure it is structured, complete, and suitable for analysis. Given the unstructured nature of job advertisements and the diversity of fields across postings, meticulous preparation is required to support reliable insights.

3.2.1 Data Sourcing

To analyse the UK data job market in a comprehensive and scalable manner, job posting data was collected from Reed.co.uk, a leading UK-based recruitment platform. The sourcing strategy integrated two distinct but complementary technologies: the Reed API and Selenium WebDriver. This dual approach was necessary to ensure both breadth and depth in data coverage.

To implement this hybrid pipeline, a combination of Python libraries and web automation tools was used. Table 3.1 summarises the key technologies employed in the data collection process and their respective purposes:

Table 3.1: Key Tools and Libraries Used in the Data Collection Pipeline

Component	Tool/Library	Purpose
API Request and Handling	requests (Python)	Access Reed.co.uk's API and retrieve job posting metadata
Authentication	base64	Encode API key for secure HTTP header transmission
Data Storage and Processing	csv, pandas	Save raw data to CSV; structure and merge enriched job descriptions
Browser Automation	selenium, webdriver	Automate navigation to job URLs and extract full job descriptions
Dynamic Content Handling	WebDriverWait, By, EC	Wait for page elements to load before extraction
Execution Control	time	Introduce delays to manage API rate limits and page loading stability

While the API provided structured data fields (such as job title, employer, location, and salary), it frequently omitted detailed job descriptions. Therefore, browser automation was employed to enrich the dataset by extracting full job content directly from each listing.

The data extraction process began by defining a set of eleven job titles commonly associated with data-related roles. These included, but were not limited to, "Data Analyst", "Data Scientist", "BI Analyst", "Machine Learning Engineer", and "Power BI Developer". Each of these titles served as a keyword to query the Reed API in an iterative loop. The API request returned paginated job listings in batches of up to 100 entries per call, and pagination was controlled using the `resultsToSkip` parameter to ensure all available results were captured.

To prevent duplication, a Python set was used to track previously seen URLs. For every unique job posting retrieved, the script extracted metadata such as job title, employer, location, minimum and maximum salary, job type, contract type, number of applications, and job URL. These records were written to a CSV file in real-time. Each entry also retained the search keyword that triggered its inclusion, thus preserving traceability in role categorisation.

However, one of the major limitations of the Reed API is that it provides only truncated or summarised job descriptions. To address this, Selenium WebDriver was employed to access each job's URL in a headless Safari browser. Once the page had loaded, the script attempted to locate the full job description using a set of CSS selectors derived from the page's HTML structure. The use of multiple selectors was essential to account for layout variability across postings.

If the primary container was not found, fallback selectors were used to extract any residual content. This scraping step added a critical layer of detail, capturing narrative content that included required technical tools, behavioural competencies, and contextual language used by employers. The final descriptions were merged back into the dataset using the job URL as a primary key.

A flowchart illustrating this end-to-end data collection pipeline is shown below.

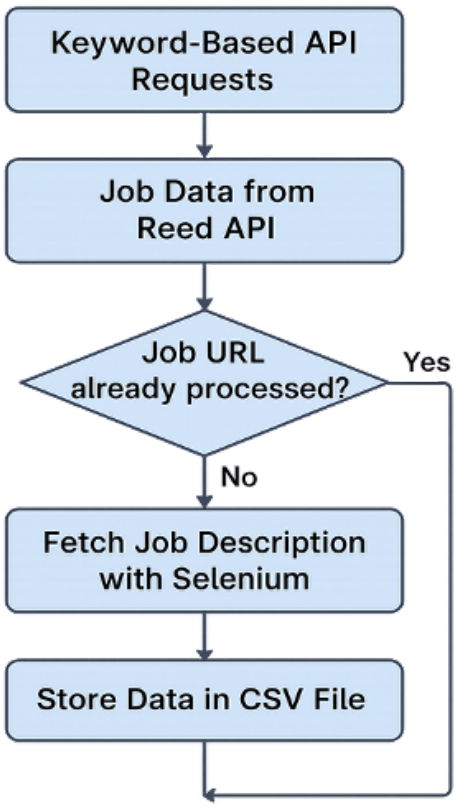


Figure 3.2: End-to-End Data Collection Pipeline

The use of both API-based and browser-based collection methods ensured that the dataset was not only broad in scope but also rich in content. By combining structured fields with detailed textual descriptions, the data sourcing process enabled subsequent phases of the analysis, including keyword-based skill extraction, salary standardisation, and geographic mapping.

This integrated and automated approach resulted in a final dataset comprising over 6,300 job records. Each entry in this dataset contained both structured fields and unstructured content, thus supporting a multi-layered analysis of the UK data job market landscape in 2025.

3.2.2 Data Cleaning and Transformation

Following the acquisition of raw job data, a series of systematic cleaning and transformation steps were performed to ensure the dataset's analytical integrity. These steps addressed missing values, inconsistent formats, and unstructured text content. The final dataset brought together structured fields obtained from the Reed API—such as job titles, employers, locations, and salary details—with the complete job descriptions extracted using Selenium. This integration of structured and unstructured data enabled a more comprehensive analysis of skill requirements, salary trends, and geographic patterns.

A. Initial Assessment and Data Structuring

The initial dataset consisted of 6,330 job postings retrieved through the combined use of the Reed API and Selenium WebDriver. Each entry included metadata such as job title, company, location, salary range, posting date, number of applications, and the full job description. A structured review of the raw data confirmed that the fundamental schema was consistent and analysis-ready at a structural level.

```
file_path = "/Users/kirancorreya/Downloads/Dessertation/fies/
   Dataset2025.xlsx"
df = pd.read_excel(file_path)
print("Dataset loaded. Shape:", df.shape)
print("Columns:", df.columns)
```

Listing 3.1: Python Code: Loading and Inspecting Excel Dataset

```
Dataset loaded. Shape: (6330, 10)
Columns: Index(['Job ID', 'Job Title', 'Company', 'Location', 'Minimum Salary',
       'Maximum Salary', 'Date Posted', 'Applications Count', 'Search Keyword',
       'Full Job Description'],
      dtype='object')
```

Figure 3.3: Output of Dataset Load and Column Inspection

An initial review of the dataset revealed a significant presence of irrelevant job postings. This occurred because the API searches returned results based on keyword presence in any part of the posting, often capturing irrelevant roles with vague or loosely related titles.

To address this, a keyword-matching filter was applied to job titles using regular expressions. A curated list of domain-relevant terms (e.g., "data analyst", "machine learning", "python", "business analyst") was used to scan each job title. A new column—`Matched Keywords`—was created to tag valid listings. Rows with "No Keyword" were excluded from further analysis to enhance dataset precision.

```
keywords = ['data', 'data analyst', 'power', 'AI', 'data engineer
    ', 'data scientist', 'data science',
             'Sales Analyst', 'Business Analyst', 'Finance Analyst
                ', 'Research Analyst', 'machine learning',
             'python', 'Intelligence']

def match_keywords_in_title(job_title):
    matched = [kw for kw in keywords if re.search(r'\b' + re.
        escape(kw.lower()) + r'\b', str(job_title).lower())]
    return ', '.join(matched) if matched else 'No Keyword'

if 'Matched Keywords' not in df.columns:
    df['Matched Keywords'] = df['Job Title'].apply(
        match_keywords_in_title)
```

Listing 3.2: Python Code: Keyword-Based Filtering of Job Titles

```
Matched Keywords column added.
Number of job titles with matched keyword: 1561
```

Figure 3.4: Status of Keyword-Based Filtering of Job Titles

This filtering step significantly improved the dataset's relevance by reducing noise introduced during the collection phase, narrowing the dataset from 6,330 to 1,561 relevant job postings.

Additionally, a `Job ID` column was introduced to facilitate unique identification and referencing of each record. This was particularly useful for downstream merging, filtering, and visualisation tasks. Job IDs were assigned sequentially based on the row index.

B. Salary Standardisation and Transformation

Job postings on Reed.co.uk often present salary information using a variety of time-based units, including hourly, daily, weekly, monthly, and annual formats. To facilitate consistent comparison and analysis across listings, all salaries were converted into standardised annual equivalents through a threshold-based classification system.

To define these thresholds, the 10^{th} percentile annual salary for data analysts in the UK (£25,313)—sourced from ITJobsWatch (2025)[14]—was reverse-engineered into corresponding hourly, daily, weekly, and monthly rates using standard UK working time assumptions (40 hours per week, 2080 hours per year). These calculations informed the logical banding system used to infer salary type based on value ranges:

A custom classification function, `detect_salary_type`, was used to infer the salary type (e.g., Hourly, Daily, Monthly) based on either the `Minimum Salary` or, if unavailable,

Table 3.2: Salary Type Classification Thresholds

Salary Type	Formula Used	Approximate Value	Threshold Range in Code
Hourly	£25,313 ÷ 2080 hrs	£12.18/hour	0–96
Daily	£12.18 × 8 hrs	£97.44/day	97–479
Weekly	£25,313 ÷ 52 weeks	£486.79/week	480–2099
Monthly	£25,313 ÷ 12 months	£2107.75/month	2100–6500
Yearly	N/A	N/A	¿6500

the `Maximum Salary`. The logic used clearly defined thresholds to categorise salary entries based on realistic UK pay bands.

```
def detect_salary_type(salary):
    if salary <= 96:
        return "Hourly"
    elif salary <= 479:
        return "Daily"
    elif salary <= 2099:
        return "Weekly"
    elif salary <= 6500:
        return "Monthly"
    else:
        return "Yearly"
```

Listing 3.3: Salary Type Classification and Annualisation

Assumption: Standard full-time UK employment = 40 hours/week → 2080 hours/year (52 weeks)

After the salary type was assigned using threshold logic, the second function computed the **Standardised Annual Salary** by:

- Calculating the average of the minimum and maximum salary, if both were available (or using whichever was present),
- Applying a multiplier based on the salary type:
  - Hourly × 40 × 52
  - Daily × 5 × 52
  - Weekly × 52
  - Monthly × 12
  - Yearly: unchanged

This two-step process ensured that all job listings were expressed in a comparable annual format, enabling accurate aggregation and comparison of compensation across roles, regions, and skills.

C. Skill Extraction from Job Descriptions

To identify the technical competencies required for data roles, a predefined list of relevant skills was compiled. This list included core tools and technologies commonly expected in data analytics and data science positions, such as:

Table 3.3: Curated Skill Keywords for Extraction

Skill
Python
SQL
R
Machine Learning
Tableau
Snowflake
AWS
Excel
Power BI

The extraction approach relied on direct keyword matching in order to maintain interpretability, consistency, and control over matching logic.

Table 3.3: Curated Skill Keywords for Extraction.

The extraction approach relied on direct keyword matching in order to maintain interpretability, consistency, and control over matching logic.

```
skills = ['python', 'sql', 'r', 'machine learning',
          'tableau', 'snowflake', 'aws', 'excel',
          'power bi']

short_skills = ['r']
long_skills = [skill for skill in skills if skill not in
   short_skills]

def extract_skills(description):
    description_cleaned = re.sub(r'[^a-zA-Z]', '', str(
       description).lower())
    matched_skills = []
```

Listing 3.4: Python Code: Skill Extraction from Job Descriptions

The extraction approach avoided natural language processing or machine learning models to maintain clarity and auditability.

Skill Matching Logic:

- For **multi-word and long skills** (e.g., *machine learning*, *power bi*):
  - Job description text was cleaned of non-alphabetic characters and converted to lowercase.
  - Each skill was normalised (spaces removed) and checked for presence in the cleaned string.
- For **short, ambiguous skills** like *r*:
  - A word-boundary-aware regular expression was used to prevent false positives (e.g., `r` in `report`).
  - Original uncleaned text was retained to preserve structure during matching.

Each job description was scanned using this logic, with results stored in a column titled `Matched Skills`. If no match was found, the record was marked as *"No skills found"*. This ensured a transparent and consistent mapping between job requirements and technical competencies.

D. Location Normalisation

Job location data extracted from Reed.co.uk exhibited inconsistent formats—some entries listed cities (e.g., *London, Manchester*), while others included postcode prefixes (e.g., *E14, M1*). This variability impaired regional analysis.

To resolve this, Power BI's Power Query was used to normalise location fields into standardised city names...

Step 1: Classification of Location Format

A new column, `Location Type`, was added to identify whether each entry represented a city or a postcode. This classification was based on detecting numeric characters in the location string. If numbers were present, the entry was labelled as `Postcode`; otherwise, it was marked as `City`.

```
Location Type = Table.AddColumn(#"Changed Type", "Location type",
    each if Text.Select([Location], {"0".."9"}) <> "" then "
   Postcode" else "City")
```

Listing 3.5: Power Query Logic: Location Type Classification

Step 2: Extraction of Postcode Prefix

For entries classified as postcodes, only the alphabetical prefix was extracted. This prefix (e.g., `E`, `M`, `LS`) is sufficient to identify the general geographic region.

```
Location Prefix= Table.AddColumn(#"Reordered Columns", "Location
   prefix", each if List.MatchesAny(Text.ToList([Location]), each
    _ >= "0" and _ <= "9") then
    Text.Upper(Text.Start([Location], List.PositionOfAny(Text.
       ToList([Location]), {"0".."9"})))
else
    [Location])
```

Listing 3.6: Power Query Logic: Extracting Location Prefix

Step 3: Mapping Postcodes to City Names

A reference table containing UK postcode prefixes and their corresponding city names was imported into Power BI from Wikipedia using the Web Table Import feature.

Source: List of postcode areas in the United Kingdom.[15]

This reference table was merged with the job dataset on the `Location Prefix` column to associate each postcode with a full city name.

```
Full City Name = Table.ExpandTableColumn(#"Merged Queries", "
   Postcode", {"Column2"}, {"Full City Name"})
```

Listing 3.7: Power Query Logic: Merging Location Prefix with City Table

Step 4: Final Standardisation

A new column, `Standardised Location`, was created using conditional logic. If a matching city name was retrieved from the reference table, it replaced the original location value. Otherwise, the original value was retained.

```
Standardised Location = Table.AddColumn(#"Renamed Columns", "
   Standardised Location ", each if [Location type] = "Postcode"
   and [Full City Name] <> null then [Full City Name] else [
   Location])
```

Listing 3.8: Power Query Logic: Final Location Transformation

This four-step pipeline ensured that all job listings were consistently mapped to city-level granularity, enabling reliable spatial analysis in Power BI.

3.3 Model Planning

Given the exploratory and descriptive objectives of this research, model planning focused on selecting analytical techniques that could extract interpretable patterns from unstructured text data and visualise complex relationships between variables. Specifically, text analytics and network visualisation methods were deemed most appropriate for uncovering trends in skill demand and mapping associations between job titles and technical

competencies. These models align with the study's goal of delivering both quantitative rigour and stakeholder-accessible insights through dynamic and intuitive visual formats.

Two primary models were planned:

a) Force-Directed Network Diagram (Skill–Role Mapping)

To visualise relationships between job roles and the technical skills extracted from descriptions, a force-directed network model was constructed in Power BI. This model depicted job roles and skills as nodes, with edges representing co-occurrence frequency.

Filters were incorporated to allow users to view:

- All roles and their linked skills,
- A specific skill and its associated roles,
- Or a particular role and its most common skill requirements.

The force-directed layout enabled a dynamic, spatial understanding of skill centrality and cross-role applicability, offering deeper insights into skill clusters and multi-role skill relevance.

These models were selected for their suitability in non-parametric, text-heavy datasets, and their ability to support interactive exploration through Power BI.

b) Word Cloud Analysis (Role-Based)

To examine the recurring terminology and responsibilities within job descriptions, word cloud visualisation was employed. A cleaned version of each job description was used to generate visual frequency maps, where word size represented relative occurrence.

Importantly, a job role filter was introduced to allow viewers to isolate word clouds for specific roles—such as Data Analyst, Machine Learning Engineer, or Business Intelligence Analyst. This role-based segmentation enabled comparison of the language and skill expectations across job categories.

This method was chosen for its interpretability, ease of communication, and ability to highlight high-frequency terms in an intuitive format.

3.4 Model Building

Following the model planning phase, two primary visual analytics components were developed to support exploration of skill demand and job-role relationships: a word cloud visualisation and a force-directed network diagram, both integrated into an interactive Power BI dashboard.

3.4.1 Word Cloud Construction and Refinement

To extract qualitative insights from job descriptions and highlight non-technical attributes such as soft skills, work environment descriptors, and behavioural expectations, a word

cloud model was developed using Power BI. The process involved custom text cleaning, frequency analysis, and domain-specific filtering to ensure that the final output was both informative and targeted.

a) Initial Cleaning Using Generic Stop Words

The first step involved removing common English stop words using a Python-based cleaning script. A list of 851 generic stop words was sourced from and loaded from an Excel file. These included standard fillers such as `and`, `with`, `the`, and `at`.

b) Word Cloud Output and Emergence of Domain-Generic Noise

After importing the cleaned dataset into Power BI and rendering the word cloud, the initial results were not as interpretable as expected. Although generic fillers were removed, many of the top words in the visual were domain-generic but semantically weak, such as:

- team, career, opportunity, support, responsible, requirements

These did not help differentiate roles or reflect specific skill requirements — instead, they represented recruitment boilerplate language common across job postings.

c) Frequency Refinement and Domain Stop Word Expansion [16]

```
# Combine all cleaned job descriptions into one string
all_words = ' '.join(df['Cleaned Job Description'].dropna()).
    split()

# Count frequency of each word
word_freq = Counter(all_words)

# Convert to DataFrame
freq_df = pd.DataFrame(word_freq.items(), columns=['Word', '
    Frequency'])
```

Listing 3.9: Python Code: Word Frequency and Stop Word Cleaning

The frequency count output from the above code was used to review generic keywords. The top 200 most frequent words were manually reviewed. From this list, 181 domain-specific stop words were identified and added to the original stop word file. These included words such as *management*, *company*, *projects*, *working*, and *services*—terms that are common but not analytically valuable.

The dataset was re-cleaned using the updated stop word list to remove this second layer of noise.

d) Final Visual Configuration and Role Filtering

The final cleaned dataset was imported back into Power BI, where the word cloud was configured as follows:

Table 3.4: Word Cloud Visual Configuration

Parameter	Configuration
Category Field	Cleaned Job Description
Stop Words	Fully applied via preprocessing (no in-visual exclusions used)
Slicer	Job Role (Search Keyword)
Word Display Limit	Top 200

This cleaned and filtered version produced a significantly more interpretable cloud, surfacing phrases such as *problem solving*, *data driven*, *forward thinking*, *cross functional*, and *cutting edge*. The visual was especially effective when filtered by role, offering targeted insights into how employers describe expectations for different data job titles.

3.4.2 Network Analysis of Role–Skill Relationships

To analyse the co-occurrence between job roles and associated technical skills, a force-directed network diagram was developed in Power BI. The purpose of this visual was to uncover patterns in how specific skills cluster around different job roles and to surface both core and niche competencies across the UK data job market.

a) Dataset Construction and Skill Explosion

The original dataset contained a column `Matched Skills` where multiple skills were stored as a single, comma-separated string (e.g., `python,sql,excel`). This structure was not suitable for network analysis, as each skill needed to be represented independently for accurate node-edge mapping.

To transform this into an analysable format, a Python-based data preprocessing step was implemented: firstly, all job postings where `Matched Keywords` were `"No Keyword"` were excluded to retain only relevant data roles. Secondly, the `Matched Skills` field was cleaned and split into individual skills using string operations and list comprehension.

The dataset was then exploded so that each job–skill pair appeared on a separate row. This transformation produced a simplified structure with the following fields:

Table 3.5: Fields Used in Network Analysis Dataset

Field	Description
Job ID	Used for tracking individual listings
Search Keyword	Used as the Source node (Job Role)
Skill	Used as the Target node

```
df_filtered['Matched Skills'] = df_filtered['Matched Skills'].
   fillna('')
df_filtered['Matched Skills'] = df_filtered['Matched Skills'].
   apply(
    lambda x: [skill.strip() for skill in x.split(',') if skill.
       strip()]
)
```

Listing 3.10: Python Code 3.4.2: Exploding Skill List for Network Graph

This processed dataset was saved as an Excel file and imported into Power BI. The final file used for this model was named: `Skills Filtered`

b) Interactive Filtering and Exploration

In Power BI, a custom force-directed graph visual was configured using:

Table 3.6: Force-Directed Graph Configuration Fields

Field	Description
Source	Job Role
Target	Skill
Weight	Frequency of occurrence (computed via aggregation in Power BI)

Interactive slicers were added to enable filtering by:

- Job Title
- Skill

This interactive filtering capability enabled users to engage with the network diagram in a flexible and targeted manner. By selecting specific job roles or skills, users could examine the full spectrum of technical competencies associated with a given position, as well as observe which tools were most frequently shared across multiple job categories. For example, skills such as `SQL` and `Python` appeared prominently across various roles, indicating their cross-functional relevance. In contrast, tools like `Power BI` and `Snowflake` emerged as more role-specific, reflecting their specialised use within distinct job families such as Business Intelligence or Data Engineering.

3.5 Communicating Results

To enable effective exploration and communication of insights, an interactive Power BI dashboard was developed. It presents both high-level indicators and granular visualisa-

tions to help users understand job trends, skill distributions, salary patterns, and role–skill relationships in the UK data job market.

Slicers and Interactivity A tile-style slicer was implemented for Job Role, supporting multiple selections. Most visuals on the dashboard respond dynamically to this input, allowing users to filter all metrics and charts based on specific job categories (e.g., Data Analyst, Data Engineer).

KPI Cards

Three DAX-driven KPI cards were used to provide immediate context:

Table 3.7: KPI Cards Used in the Dashboard

KPI Metric	Description
Total Jobs Scraped	Represents the total number of job postings collected from the initial scraping process
Total Relevant Jobs	Indicates the number of job postings retained after keyword-based relevance filtering
Average Annual Salary	Calculated from standardised salary values to provide compensation benchmarking

These KPIs are interactive and automatically update based on slicer input.

Skills Summary Table A dynamic table visual lists all skills extracted across the dataset along with the total job count in which each appears. This table updates when slicers are used, allowing role-specific skill filtering.

Clustered Column Chart – Average Annual Salary by Role This chart displays the average standardised annual salary across different job roles. It supports drill-down analysis using the slicers.

Clustered Bar Chart – Total Jobs by Role This bar chart shows the number of job postings for each role, derived from the search keywords used during API scraping.

Treemap – Top 10 Recruiters This treemap visualises the top 10 most active employers or recruitment agencies by job posting volume, based on the `Employer Name` field.

Donut Chart – Jobs by City A donut chart presents job volume by standardised location, highlighting the top five UK cities with the highest job concentrations.

Word Cloud and Network Analysis These two visuals are hosted on separate dashboard pages:

- **Word Cloud:** Based on cleaned job description text and role-filterable, this visual helps identify the most common terminology and soft skills across roles.
- **Force-Directed Network Diagram:** This model maps relationships between job roles and technical skills using node–edge visualisation. It supports filtering by both skill and role, dynamically adjusting node size and edge weight based on frequency.

Table 3.8: Dashboard Visuals Summary

Visual	Type	Purpose
KPI Cards	Summary Tiles	Display total jobs scraped, total relevant jobs, and average salary
Skills Table	Table	Lists each skill and total jobs it appears in
Average Salary by Role	Clustered Column Chart	Compare average standardised salary across job roles
Total Jobs by Role	Clustered Bar Chart	Show number of job postings per role
Top 10 Recruiters	Treemap	Visualise the most active companies or agencies
Jobs by City	Donut Chart	Show job volume in top 5 cleaned locations
Word Cloud	Text Frequency Visual	Display common keywords in job descriptions by role
Force-Directed Network Diagram	Network Graph	Map role–skill relationships and co-occurrence strength

3.3 Operationalising

The final stage of the Data Analytics Life Cycle focused on converting analytical outputs into a usable and shareable form. In this project, the primary output was an interactive Power BI dashboard, designed for practical use by stakeholders such as job seekers, university curriculum designers, and hiring professionals.

To support real-world usability, all key insights were embedded into interactive visual elements, including charts, tables, and networks. Additionally, slicers for job role, skill, and location were incorporated, enabling tailored exploration of the data. The visuals were strategically grouped into thematic pages, such as Salary, Skills, and Role Trends, to ensure intuitive navigation for users.

Although the dashboard was not deployed in a public web environment, it was structured to allow for easy sharing. With minimal adjustments, it could be further operationalized in institutional settings, such as academic counseling or HR analytics platforms.

Chapter 4

Results and Discussion

4.1 Key Performance Indicators (KPI Summary)

The dashboard features three KPI cards that provide a high-level overview of the dataset and establish the foundation for deeper analysis:

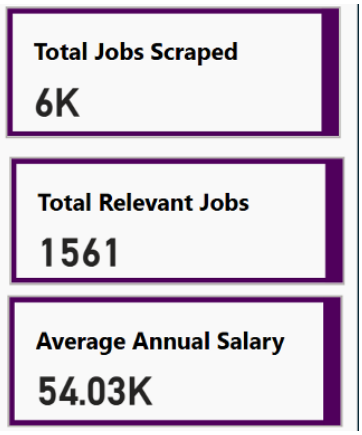


Figure 4.1: KPI Summary Cards – Total Jobs, Relevant Jobs, Average Salary

These figures highlight two key insights:

1. **Relevance Filtering Impact:** Out of 6,000 job postings initially scraped from Reed.co.uk, only 1,561 were deemed relevant after keyword-based filtering. This indicates that approximately 74% of the scraped data did not align closely with core data roles, underscoring the importance of rigorous post-processing to eliminate noise and improve data quality.

2. **Salary Benchmarking:** The average standardised salary across relevant roles was calculated as £54,030, based on transformed salary fields and consistent annualisation logic. This figure serves as a benchmark for comparing salary distributions across different job titles and regions in subsequent visuals.

Together, these KPIs provide essential context about the size, quality, and value profile of the UK data job market as captured through this study.

4.2 Job Distribution by Role

The clustered bar chart illustrates the distribution of job postings across various roles within the UK data job market. The chart highlights the number of relevant postings identified for each role as follows:

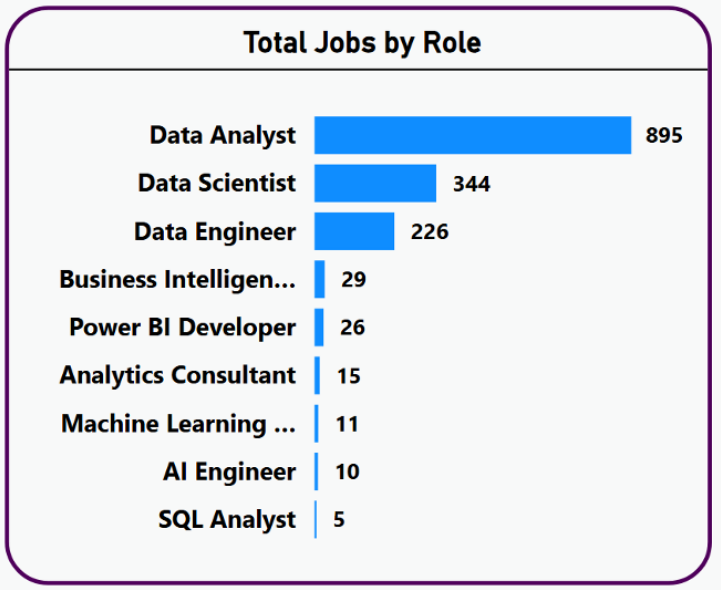


Figure 4.2: Total Jobs by Role – Clustered Bar Chart

Data Analyst roles are by far the most frequently advertised, accounting for over 57% of all relevant postings (895 out of 1,561). This suggests high demand for professionals with generalist analytical skills, often in business and operations-focused environments.

Data Scientist positions are the second most common, with 344 roles — reflecting a steady need for advanced analytical, statistical, and machine learning capabilities. **Data Engineer** postings (226) indicate solid demand for backend data infrastructure skills, though lower in volume compared to analysts and scientists. **BI Analyst** and **Power BI Developer** roles appear much less frequently. This may reflect narrower specialization, integration into broader analyst roles, or job title variation across companies.

This distribution establishes a foundation for deeper comparisons across salary, skill requirements, and geographic demand, particularly focusing on the top three roles: Data Analyst, Data Scientist, and Data Engineer.

4.3 Skill Frequency and Demand

This section examines the frequency of technical skills mentioned across job descriptions to identify which competencies are most in demand within the UK data job market. A dedicated table visual in Power BI was used to display the number of job postings in which each skill appeared, both in aggregate and segmented by job role.

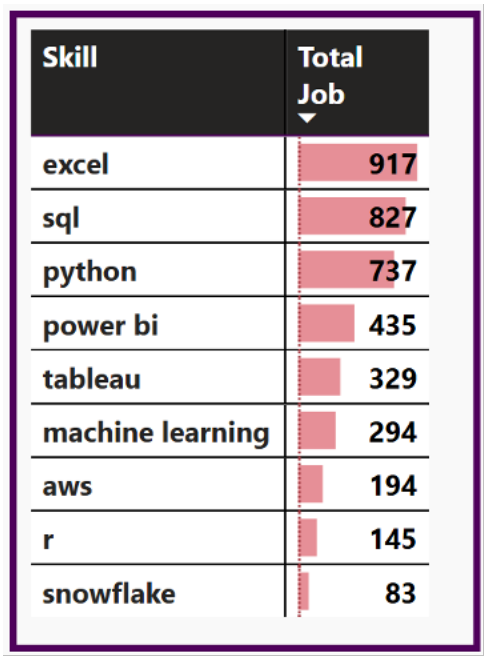


Figure 4.3: Skill Frequency Table – Overall

Skill Distribution by Job Role

A breakdown of the skills by top 3 job roles reveals how demand varies across Data Analyst, Data Engineer, and Data Scientist roles:

Skill	Data Analyst	Data Engineer	Data Scientist
Excel	658	80	124
SQL	483	136	160
Python	356	126	232
Power BI	241	50	86
Tableau	293	–	–
AWS	–	53	–
Machine Learning	–	–	157

Table 4.1: Skill Distribution by Top 3 Job Roles

The table above illustrates clear differentiation in tool preferences based on job role:

Power BI is prominently associated with Data Analyst and Data Scientist roles, reinforcing its value in dashboarding and visual reporting tasks.

Tableau appears almost exclusively in Data Analyst roles, suggesting either employer preference for Power BI or a concentration of Tableau usage in specific sectors.

AWS is primarily mentioned in Data Engineer roles, aligning with cloud infrastructure and platform development needs.

Machine Learning is specific to Data Scientist roles, indicating the modelling and experimentation focus of that domain.

Noteworthy Observation: Python Surpassing Visualisation Tools

One of the most significant findings is that Python is mentioned more frequently than either Power BI or Tableau, even in roles traditionally associated with reporting and visualisation. This indicates a growing expectation for candidates to have programmatic data capabilities, such as automation and custom analytics, across all job roles.

4.4 Salary Analysis by Job Role

To examine how compensation varies across different job titles, a clustered column chart was used to visualise the average standardised annual salary (in £000s) for each role. This metric was derived through salary transformation logic applied uniformly to ensure comparability across hourly, daily, monthly, and yearly postings.

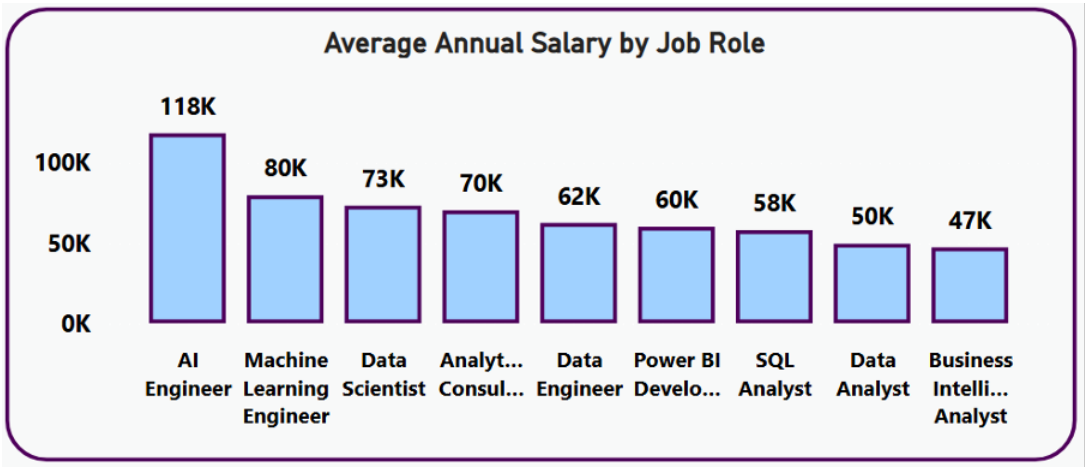


Figure 4.4: Average Annual Salary by Job Role – Clustered Column Chart

Insights:

Top-Tier Roles: AI & ML Engineering

AI Engineer roles command the highest average salary at £118,000, followed by Machine Learning Engineers at £80,000. These roles typically involve deep technical expertise in

model deployment, advanced algorithms, and production-grade AI solutions, justifying the higher compensation.

Strategic and Technical Roles Pay More

Data Scientists and Analytics Consultants earn average salaries of £73,000 and £70,000, respectively, showing strong market value for roles that combine statistical depth with business-facing insights.

Engineering and Visualisation Roles

Data Engineers are well-compensated at £62,000, reflecting the importance of data pipeline and infrastructure work. Power BI Developers and SQL Analysts fall into a mid-tier range of £58,000–£60,000, blending technical implementation with reporting tasks.

Foundational Analyst Roles Offer Lower Salaries

Data Analysts and BI Analysts, often entry-level or mid-tier roles, sit at the lower end of the spectrum with £50,000 and £47,000 respectively. This may reflect broader candidate supply and fewer technical entry barriers compared to specialist positions.

Observation

This salary analysis confirms a clear correlation between technical complexity and compensation. Roles that involve engineering, machine learning, or AI systems tend to command significantly higher salaries than general analyst positions. These findings reinforce the market value of advanced skillsets—such as programming (Python), cloud infrastructure (AWS), and machine learning frameworks—for professionals aiming to increase their earning potential in the UK data job market.

However, while roles such as AI Engineer and Machine Learning Engineer offer the highest average salaries, they represent a relatively small portion of the market, as noted in Section 4.2. This suggests that such roles, although financially attractive, are also more competitive and less frequently advertised compared to broader roles like Data Analyst or Data Scientist.

4.5 Geographic Distribution of Data Roles

A donut chart was used to visualise the top five UK cities with the highest concentration of relevant data job postings. The chart was based on cleaned and standardised location data derived from job postings, using postcode parsing and city-matching techniques (as detailed in Section 3.2.2).

Insights

1. **London's Dominance:**
   London accounts for over 36% of all relevant job postings (566 out of 1,561), making

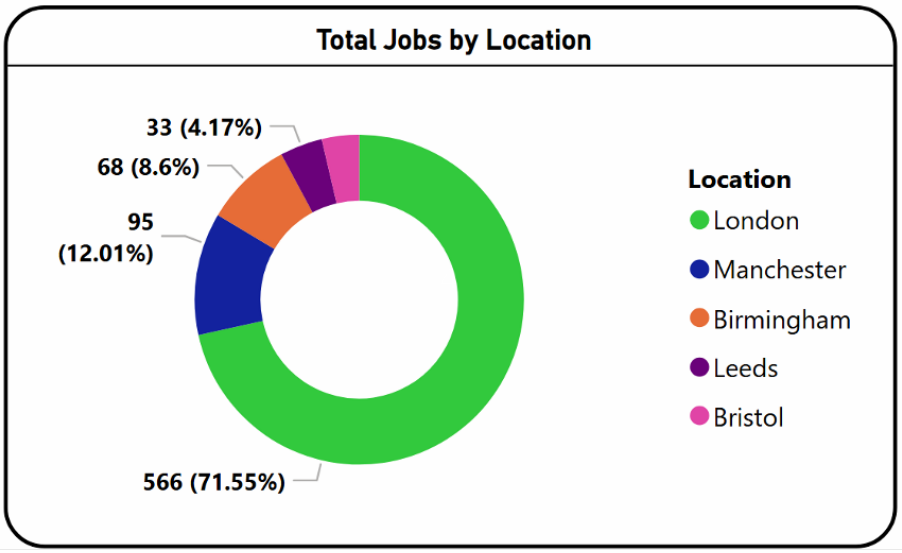


Figure 4.5: Jobs by City – Donut Chart

Table 4.2: Job Postings by Top 5 Cities and Role

City	Total Jobs	Data Analyst	Data Scientist	Data Engineer
London	566	283	145	95
Manchester	95	38	28	22
Birmingham	68	40	18	8
Leeds	33	13	9	9
Bristol	29	12	9	6

it the clear hub for data-related hiring in the UK. This concentration spans all roles—Data Analyst, Scientist, and Engineer—reflecting London's position as the country's primary economic and technological centre.

2. **Regional Opportunities in Key Cities:**
   Beyond London, cities like Manchester, Birmingham, and Leeds present secondary hubs of opportunity, especially for analyst roles. Manchester alone accounts for 95 postings, showing strong northern demand in line with its growing tech sector.

3. **Data Engineering Roles Follow Urban Concentration:**
   Data Engineer roles show a consistent pattern across cities, often tracking closely with Data Scientist demand. These roles tend to cluster in locations with more mature or infrastructure-heavy organisations, such as financial firms or digital consultancies.

4. **Smaller Cities Have Limited but Diverse Demand:**
   Leeds and Bristol, while contributing fewer jobs overall, still show representation across all three roles. This suggests a distributed—though less dense—demand landscape outside of Tier 1 cities.

Conclusion

The data highlights a strong geographic skew, with London clearly leading but other cities offering credible opportunities, especially for candidates open to relocation. These patterns are valuable for both job seekers planning regional applications and for educators or policymakers assessing the spatial dynamics of digital employment in the UK.

4.6 Recruiter Landscape

A treemap visualisation was employed to depict the distribution of relevant job postings among recruitment agencies. This analysis highlights the key players in the UK's data job market.

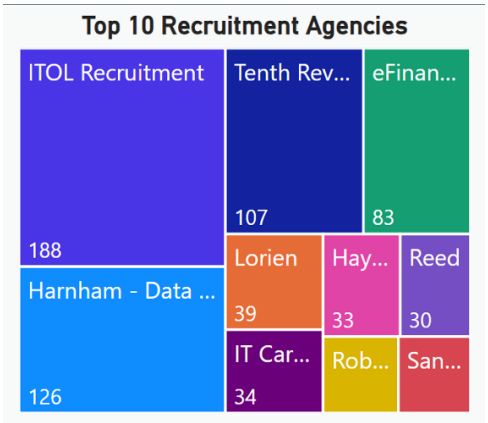


Figure 4.6: Top 10 Recruiters – Treemap Visualisation

Table 4.3: Top 5 Recruiters by Job Postings

Recruiter	Job Openings
ITOL Recruitment	188
Harnham – Data Analytics Recruitment	126
Tenth Revolution Group	107
eFinancialCareers	83
Lorien	39

Recruiter Profiles

- **ITOL Recruitment:** Focuses on developing newly qualified professionals and connecting them with employers in project management, business analysis, IT, and cybersecurity. They address UK-wide skills shortages with staff who contribute from the outset.

- **Harnham – Data Analytics Recruitment:** A specialist agency in the Data & Analytics sector with over 15 years of experience. They offer services across the UK, US, and EU, covering data science, AI, and NLP.
- **Tenth Revolution Group:** A global leader in technology talent solutions. They focus on closing the tech skills gap by finding, training, and deploying professionals, including through reskilling and consulting services.
- **eFinancialCareers:** A global platform connecting finance and tech professionals with opportunities. They operate in 23 territories, serving as a niche recruitment tool in financial services and tech.
- **Lorien:** A specialist in technology and transformation talent solutions. They provide services to diverse clients, combining tech sector knowledge with scalable talent offerings for digital competitiveness.

Insights

1. **Dominance of Specialized Recruiters:** The top recruiters are niche firms focusing on tech and data, indicating a market preference for domain-specific hiring expertise.
2. **Geographical Concentration:** These agencies predominantly operate within the UK, reflecting an intent to address domestic demand.
3. **Role-Specific Focus:** Agencies like Harnham and Tenth Revolution Group specialise in data roles, aligning well with observed labour market trends.

4.7 Network Analysis: Role–Skill Relationships

To explore the relationship between job roles and their associated technical skills, a force-directed network diagram was created using Power BI. In this network:

- **Job roles** serve as source nodes (e.g., Data Analyst, Data Scientist).
- **Skills** are target nodes (e.g., Python, Power BI).
- **Edge thickness** reflects the number of job postings linking the two.
- **Slicers** allow interactive filtering by job role or skill.

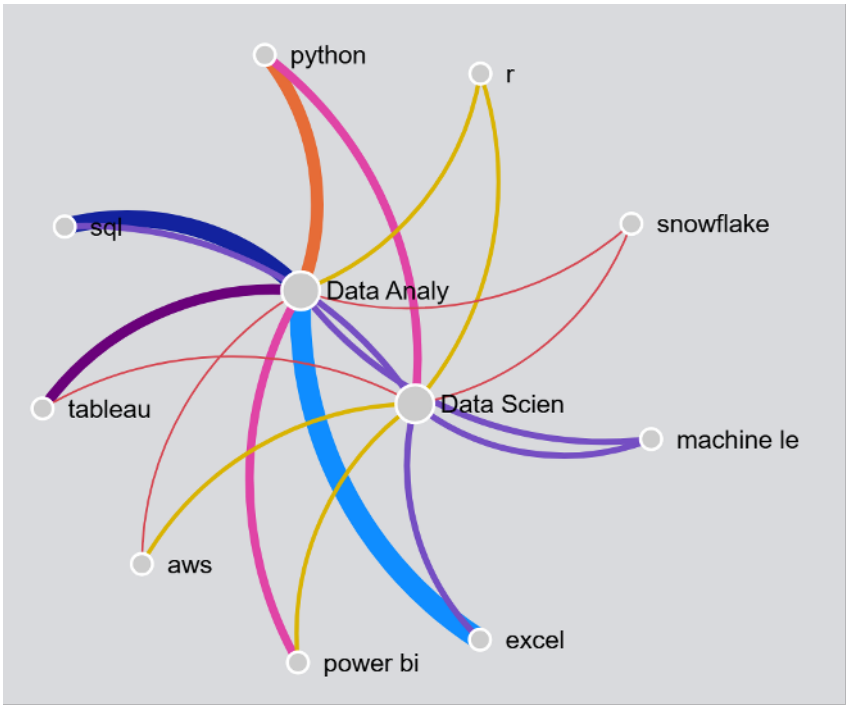


Figure 4.7: Force-Directed Network Graph – Role–Skill Relationships

Observations

Observation 1: High Interconnectivity Across Roles and Skills

The network exhibits a high degree of overlap, with most skills connected to multiple job roles. This reflects the cross-functional nature of the UK data job market, where core skills such as Python, SQL, and Excel are essential across a variety of positions. This universality, while reducing role differentiation in the unfiltered graph, confirms that foundational tools remain central across job types.

Observation 2: Data Analyst Volume Skews Visual Prominence

Due to the large number of Data Analyst job postings (as discussed in Section 4.1), the edges linking this role to skills appear visually dominant. This highlights a methodological nuance: edge thickness is a function of posting volume, not necessarily skill criticality. As such, slicer-based filtering is recommended to uncover role-specific patterns.

Observation 3: Network Filtering Enhances Interpretability

Applying slicers to isolate individual roles transforms the graph from dense to informative:

- Filtering for *Data Scientist* surfaces high-value skills such as Python, SQL, and Machine Learning.
- Filtering for *Power BI Developer* restricts the network to Power BI, Excel, and SQL.

This feature allows users to explore customised skill maps that better reflect employer expectations for specific roles.

Observation 4: Limited Visibility of Niche or Domain-Specific Tools

Although tools such as Snowflake and AWS were included in the skill extraction list, they did not form prominent or visually distinct clusters within the network. This is primarily due to their low frequency of mention across job postings.

Notably, these domain-specific tools occasionally appeared connected to broadly defined roles like Data Analyst—an unexpected pairing, given that such tools are typically associated with more specialised roles like Data Engineer. This may point to inconsistencies in job title usage, where some postings assign general titles to technically demanding roles. Alternatively, it may reflect the use of generic recruitment language, where job descriptions include a wide range of skills, not all of which are central to the actual role.

Observation 5: Role Centrality and Skill Reach

Among all roles, Data Analyst stands out as the most connected node, linked to nearly every skill in the dataset. This supports the view that the role is often broadly defined, requiring proficiency in tools ranging from Excel and Power BI to Python and SQL.

In contrast, Data Engineer roles, while strongly connected to back-end skills like SQL, Python, and AWS, show no significant association with Machine Learning. This confirms that engineering responsibilities are generally infrastructure-focused rather than analytical.

On the other hand, Machine Learning appeared almost exclusively linked to Data Scientist roles and only marginally to Data Analyst, reinforcing its domain-specific nature in model development, experimentation, and algorithmic work.

Conclusion

While the full network is dense, it effectively visualises the multi-role demand for core competencies like Python and SQL. When filtered by role, the diagram reveals meaningful structural relationships and helps identify the skill ecosystems tied to specific job titles. The method also exposes potential inconsistencies in job titling, offering insights not just into skill requirements but also into the fluidity and ambiguity of role definitions in the current UK data job market.

4.8 Word Cloud Analysis: Employer Language and Role Expectations

To explore the qualitative framing of data-related roles, a word cloud analysis was conducted using cleaned job descriptions sourced from Reed.co.uk postings. This visualisation highlighted the most frequently occurring non-technical terms across the dataset,

offering insight into how employers frame roles not just by skill requirements, but also through behavioural expectations, organisational culture, and strategic focus.

Figure 4.8: Word Cloud of Job Descriptions Filtered by Role

The word cloud was generated using the Cleaned Job Description column, preprocessed using Python to remove 1,032 stopwords (including both generic English terms and 181 domain-specific filler words). This ensured that only high-context, meaningful terms remained in the analysis. A slicer was added for job role, allowing dynamic filtering of the visual by position.

General Observations Across All Roles

Across all job roles, the word cloud consistently surfaced a blend of technical tools and behavioural descriptors. The most frequent terms included:

Table 4.4: Key Technical and Behavioural Terms from Word Cloud

Category	**Terms**
Technical	SQL, Python, AI, developer
Behavioural/Cultural	problem solving, collaboration, leadership, continuous learning, dynamic, implementation, innovative

This mixture underscores that employers are increasingly looking for well-rounded candidates who not only possess hard skills but also demonstrate adaptive thinking, communication, and a growth mindset.

Notably, the equal prominence of soft skills such as *problem solving* and *collaboration* alongside core technical terms like *SQL* and *Python* suggests that these attributes are perceived as equally essential in modern data roles.

Role-Specific Variations

Filtering the word cloud by individual roles revealed distinct emphasis patterns:

Data Analyst: Keywords such as *analyse*, *SQL*, *Excel*, *strategic*, *implementation*, and *transformation* dominated, reflecting the role's hybrid position between business and data processing functions.

Data Scientist: The cloud featured *machine learning*, *Python*, *cloud*, *innovation*, *leadership*, and *collaborative* — indicating a technically advanced, research-driven role.

Data Engineer: Terms like *mechanical*, *equipment*, *safety*, and *construction* appeared, reflecting engineering-oriented language, possibly due to industry-specific needs.

Surprising and Unexpected Findings

The presence of *problem solving*, *collaboration*, and *continuous learning* at the same visual prominence as technical tools was notable. This indicates that soft skills are not merely "nice to have" but are core expectations.

Interpretation

This analysis offers several key insights:

How Employers Communicate Roles: Employers articulate job roles using a balanced combination of technical requirements and soft skill expectations. The language reflects a growing need for candidates who are both tool-proficient and capable of thriving in collaborative, innovative environments.

Value for Job Seekers and Educators: The findings reinforce the importance of developing soft skills alongside technical capabilities. For curriculum designers, it underscores the need to incorporate behavioural competencies into academic programs.

Chapter 5

Conclusion and Recommendation

This project was driven by the central research question: *"What are the most in-demand technical and analytical skills for data analytics and data science roles in the UK job market, and how do these skill requirements correlate with salary expectations and job opportunities?"* To address this, a multi-method approach combining structured rule-based models and exploratory visual analysis techniques was employed. Over 6,000 job postings were scraped from `Reed.co.uk` using a custom-built pipeline integrating the Reed API and Selenium automation. This enabled the extraction of structured data (such as job title, salary, location) and full-text job descriptions, which were enriched through skill matching, salary standardization, and job role categorization.

The structured analysis revealed that **SQL**, **Python**, and **Excel** are consistently the most requested technical skills, with role-specific skills such as **Power BI**, **AWS**, and **Machine Learning** emerging in more specialized positions. Salary analysis demonstrated a strong correlation between technical complexity and compensation, with roles such as **AI Engineer** and **Data Scientist** offering salaries significantly above the industry average.

Complementing these findings, exploratory text models—such as word clouds and network analysis—provided qualitative insights into employer expectations. The word cloud highlighted that soft skills like **problem solving**, **collaboration**, and **continuous learning** are just as prominently featured as technical skills. The force-directed network diagram offered a visual map of skill–role relationships, confirming both widespread tool overlap and distinct skill clusters when filtered by job title.

A key contribution of this project was the two-phase stop word refinement strategy, which improved clarity in the word cloud output. Additionally, a rule-based salary band detection and transformation model allowed for the normalization of diverse salary formats into comparable annual values, providing reliable compensation insights.

Together, these methods fulfilled the project's objectives, offering a comprehensive, data-driven overview of the UK's 2025 data job market. The insights derived from this

research are valuable for job seekers, educators, and recruiters, helping them navigate the complexities of role expectations, skill demand, and compensation in the evolving data job landscape.

5.1 Recommendations

5.1 Recommendations

a. For Job Seekers: For job seekers, it is essential to not only focus on core technical skills such as **SQL** and **Python** but also to actively develop and highlight soft skills like **problem-solving**, **adaptability**, and **teamwork**. These skills are increasingly sought after by employers and can significantly improve one's employability. Additionally, job seekers aiming for higher salary bands should consider pursuing specialized learning paths in advanced fields like **artificial intelligence**, **cloud platforms**, or **machine learning**, as these areas are in high demand and often command premium salaries.

b. For Educators and Curriculum Designers: For educators and curriculum designers, it is crucial to maintain an emphasis on foundational analytics tools, while also integrating real-world skill combinations that are highly valued in the industry. For example, pairing **SQL with storytelling** or **Python with communication** can create a more holistic skill set for students, equipping them to handle both technical and business-related aspects of data analysis. Furthermore, introducing dedicated modules on applied behavioural skills, such as **leadership**, **collaboration**, and other traits emphasized in employer language, would better align curricula with the evolving expectations of employers.

c. Future Work

While the current project provides robust insights into the UK data job market, there are several areas for further enhancement that could expand both its scope and analytical depth.

One potential improvement lies in the **dynamic classification of salary bands**. The existing model, which uses predefined value thresholds to classify salaries by unit (e.g., hourly, daily, monthly, yearly), assumes that pay structures remain constant across the market. A more advanced approach could involve developing a data-driven salary banding model that adjusts thresholds based on real-time benchmarking data from sources such as ITJobsWatch, the Office for National Statistics (ONS), or Glassdoor APIs. This would allow for more accurate salary classifications, particularly for atypical listings, and enable better salary comparisons across job levels and regions, adapting to market shifts over time.

Another avenue for future research involves **expanding data sources** to capture a

broader spectrum of job roles. The current study relies solely on data from Reed.co.uk, which, while extensive, may not fully represent the diversity of the market. Job postings on platforms like LinkedIn, Indeed, and sector-specific sites such as CWJobs or Technojobs often feature niche or senior positions that are underrepresented in Reed's dataset. Future studies could enhance the representativeness of the sample by integrating data from multiple job boards, reducing platform bias, and providing a more comprehensive view of job role definitions and salary trends across various platforms.

Moreover, the current analysis presents a **snapshot of the job market** at a specific point in time. Given the highly dynamic nature of the data job market, it would be beneficial to incorporate temporal and trend analysis into future work. This could involve tracking the evolution of skill demand over time, through methods such as time-series analysis, to observe how tools like Power BI, Snowflake, or Python fluctuate in popularity. Additionally, identifying seasonal hiring patterns or shifts in demand due to policy changes or technological advancements—such as the rise of AI-related roles—would provide valuable insights. Integrating time filters into the dashboard would allow for real-time trend visualisation, making it a more interactive tool for users.

Future versions of this study could also benefit from the adoption of **more advanced text analysis techniques**. While the current study employs refined keyword matching and frequency-based filtering, methods such as topic modelling (e.g., Latent Dirichlet Allocation, or LDA) could uncover hidden themes within job descriptions. Similarly, Named Entity Recognition (NER) could help identify domain-specific terms or certifications, while models like Word2Vec or BERT could capture semantic relationships between skills and job titles, offering a deeper understanding of the language used in job postings.

Additionally, the current analysis primarily focuses on **technical skills**, yet findings suggest that **domain-specific skills and soft skills** are equally important in data roles. Future work should expand the analysis to include soft skills like communication, problem-solving, and teamwork, which are often emphasized by employers alongside technical expertise. Furthermore, incorporating domain-specific skills related to particular industries or sectors could provide more nuanced insights into the specific competencies required for roles across different data domains.

Finally, improving the **role ontology and title normalisation process** would enhance the consistency of job role classifications. Many job titles in the dataset overlap or vary in ways that could complicate role-to-skill mapping (e.g., "BI Developer" vs. "Business Intelligence Analyst"). Future work could involve building a standardized taxonomy for job titles, allowing for better comparison and consistency across roles. This would improve dashboard filtering and make cross-role comparisons more intuitive and effective.

In conclusion, while this project has met its objectives by offering a detailed overview of the UK data job market, these enhancements would significantly improve the depth,

accuracy, and applicability of the findings. The insights gained from this research provide valuable guidance for job seekers and educators, while also paving the way for future studies that can further refine the understanding of the data job market and its evolving trends.

Chapter 6

Limitations

While this project achieved its objectives and provided valuable insights into the UK data job market, several limitations should be acknowledged:

Data Source Restriction Initially, the study aimed to collect job postings from multiple platforms, including LinkedIn and Indeed, to improve representativeness and reduce platform bias. However, both platforms have strict API limitations and scraping restrictions, which hindered direct data access. As a result, the dataset was exclusively sourced from Reed.co.uk, which offered an accessible API and consistent data structure.

While Reed is a prominent UK job board with substantial listing volume, it may not fully reflect the breadth of the national job market. Platforms like LinkedIn and Indeed often feature senior, niche, or contract roles that may be underrepresented on Reed.

Salary Band Assumptions The salary standardisation model relies on static thresholds to classify salaries by type (e.g., daily, hourly, annual). While the thresholds were derived from real-world benchmarks, they may not account for industry-specific or location-based pay structures, and misclassification is possible for atypical salary figures or contract roles.

Role Title Inconsistency Job roles were inferred based on the search keyword used in data scraping, not the full job title in every case. This could introduce categorisation bias, especially for multi-role listings (e.g., "Data Scientist/Data Engineer") or creatively titled postings (e.g., "Data Rockstar"). Some inconsistencies in role–skill pairing may arise from this ambiguity.

Irrelevant Job Captures During Scraping Despite using targeted search keywords, a notable number of irrelevant or misclassified job postings were captured during the initial scraping phase. This was primarily due to broad keyword matching logic applied via the Reed API, where job titles or descriptions included generic terms (e.g., "analyst," "engineer") without being directly relevant to data-focused roles.

To mitigate this, a post-scraping filtering step was implemented using keyword matching on job titles. However, this filtering was rule-based and not context-aware, which

means some irrelevant listings may still have remained, or borderline-relevant postings could have been excluded. This limitation may have introduced minor noise or bias in skill frequency and role-based distributions, particularly affecting precision in skill–role pairings.

Temporal Static Snapshot The data reflects a single point in time rather than a continuous collection period. This limits the ability to detect trends or seasonality in hiring demand and skill requirements. Employer priorities can shift rapidly, particularly in response to technological advances, economic changes, or regulatory events.

Generalisation of Word Cloud Insights Although refined through a two-stage stopword filtering process, the word cloud model remains a frequency-based tool. It does not capture the context in which terms are used (e.g., "leadership required" vs. "leadership support provided") and therefore may oversimplify semantic meanings.

References

[1] Dagny Elaine Wilkins. "An In-Depth Analysis of the Data Analytics Job Market". Honors Thesis. University of New Hampshire, 2021. URL: https://scholars.unh.edu/honors/597/.

[2] Frederick Patacsil and Michael Acosta. "Analyzing the Relationship Between Information Technology Jobs Advertised Online and Skills Requirements Using Association Rules". In: *Bulletin of Electrical Engineering and Informatics* 10.5 (2021), pp. 2771–2779. DOI: 10.11591/eei.v10i5.2590. URL: https://www.researchgate.net/publication/354370658.

[3] Sophie Magnet. *Data Analyst Job Outlook 2025: Trends, Salaries, and Skills*. 2025. URL: https://365datascience.com/career-advice/data-analyst-job-outlook-2025/.

[4] Yulia V Yurova Amit Verma Peggy L Lane. "An investigation of skill requirements for business and data analytics positions". In: *International Journal of Computer Applications* 182.43 (2019), pp. 1–5. URL: https://www.researchgate.net/publication/330156506.

[5] Andreas Sorling Jeric Bryan Lim. *The Data Analyst Job Landscape*. 2024. URL: https://du.diva-portal.org/smash/get/diva2%3A1889854/FULLTEXT01.pdf.

[6] Clare Lally Josh Fearns Lydia Harriss. *Data Science Skills in the UK Workforce*. 2023. URL: https://post.parliament.uk/research-briefings/post-pn-0697/.

[7] Khushi Jaiswal, Ievgeniia Kuzminykh, and Sanjay Modgil. "Understanding the Skills Gap Between Higher Education and Industry in the UK in Artificial Intelligence Sector". In: *SAGE Open* (2024). DOI: 10.1177/09504222241280441. URL: https://journals.sagepub.com/doi/abs/10.1177/09504222241280441.

[8] Ya. *Analyzing the Data Analytics Job Market — A SQL and Python Project*. 2022. URL: https://medium.com/@yali0514/analyzing-the-data-analytics-job-market-a-sql-and-python-project-introduction-de8f816e23df.

[9] Nik Dawson et al. “Adaptively Selecting Occupations to Detect Skill Shortages from Online Job Ads”. In: *arXiv preprint* (2019). URL: https://arxiv.org/abs/1911.02302.

[10] Negar Loloshahvar. *Power BI Job Market Analysis*. 2023. URL: https://github.com/negarloloshahvar/Case-Study-Analyzing-Job-Market-Data-in-Power-BI.

[11] Mike Zhang. “Computational Job Market Analysis with Natural Language Processing”. In: *arXiv preprint* (2024). URL: https://arxiv.org/abs/2404.18977.

[12] Farhad Mehdipour George Huang Emre Erturk. “Data Analytics Job Market in Australia and New Zealand”. In: *International Journal of Data Science* 5.2 (2024), pp. 45–52. URL: https://www.researchgate.net/publication/384370999.

[13] RudderStack Editorial Team. *The Data Analytics Lifecycle: What It Is and How It Works*. Accessed: 2025-04-13. 2024. URL: https://www.rudderstack.com/learn/data-analytics/data-analytics-lifecycle/.

[14] ITJobsWatch. *Data Analyst Salary Trends in the UK*. Accessed: 2025-05-13. 2025. URL: https://www.itjobswatch.co.uk/jobs/uk/data%20analyst.do.

[15] Wikipedia Contributors. *List of Postcode Areas in the United Kingdom*. Accessed: 2025-05-13. 2025. URL: https://en.wikipedia.org/wiki/List_of_postcode_areas_in_the_United_Kingdom.

[16] Abdallah Ashraf. *How to Handle Stop Words in NLP*. Accessed: 2025-05-10. 2023. URL: https://medium.com/@abdallahashraf90x/how-to-handle-stop-words-in-nlp-7fa5c7eb40f6.

[17] Joe Lewis, Andrew Powell, Nerys Robert, Matthew Ward, *Digital Skills And Careers*, 2024, available at: https://researchbriefings.files.parliament.uk/documents/CDP-2024-0073/CDP-2024-0073.pdf.

Appendices

Appendix A: Power BI Dashboard

.1 Power BI Link

[Access the Power BI Dashboard]

.2 Power BI Overview Page

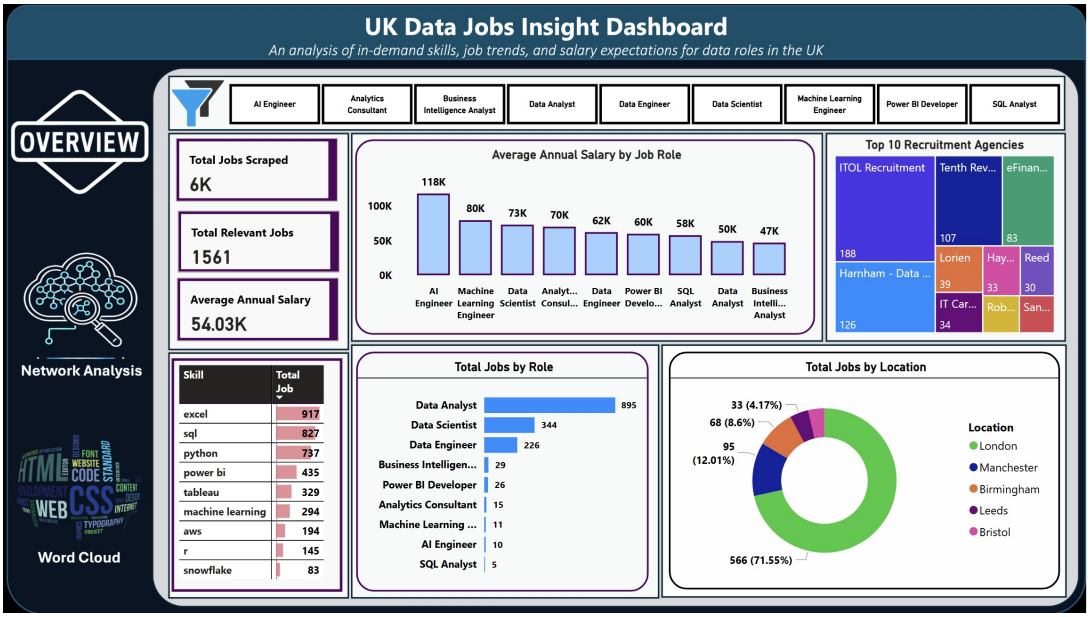


Figure 1: Power BI Overview Page

.3 Power BI Network Analysis Page

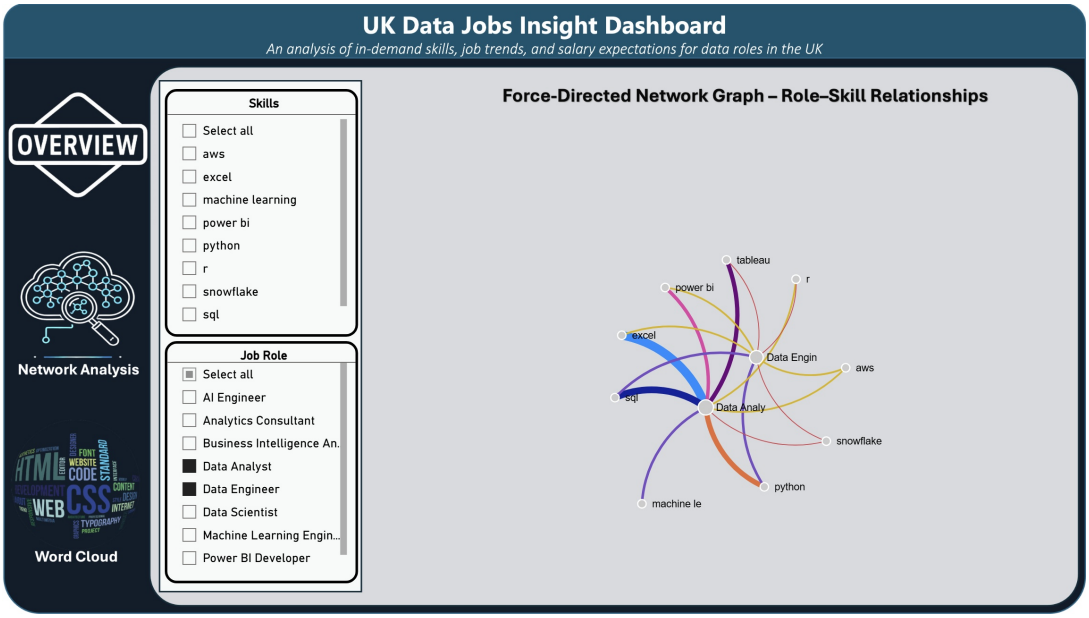


Figure 2: Power BI Network Analysis Page

.4 Power BI Word Cloud Page

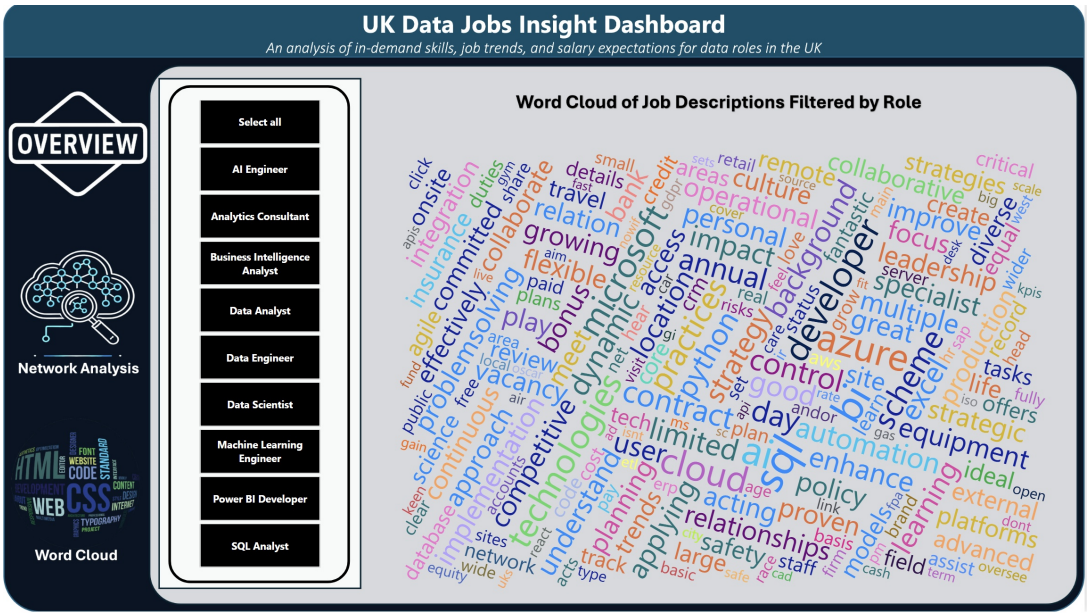


Figure 3: Power BI Word Cloud Page

Appendix 2: Python Codebase and Data Processing

.1 Data Scraping Code Link

Access Data Scraping Code

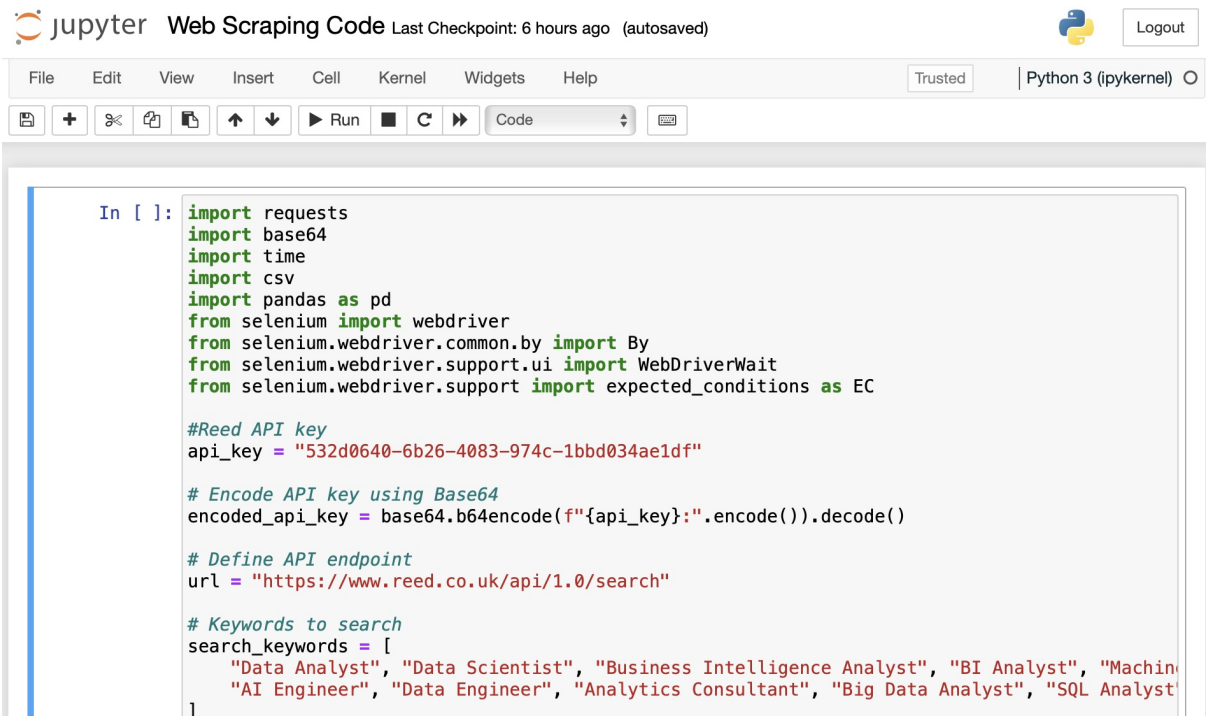
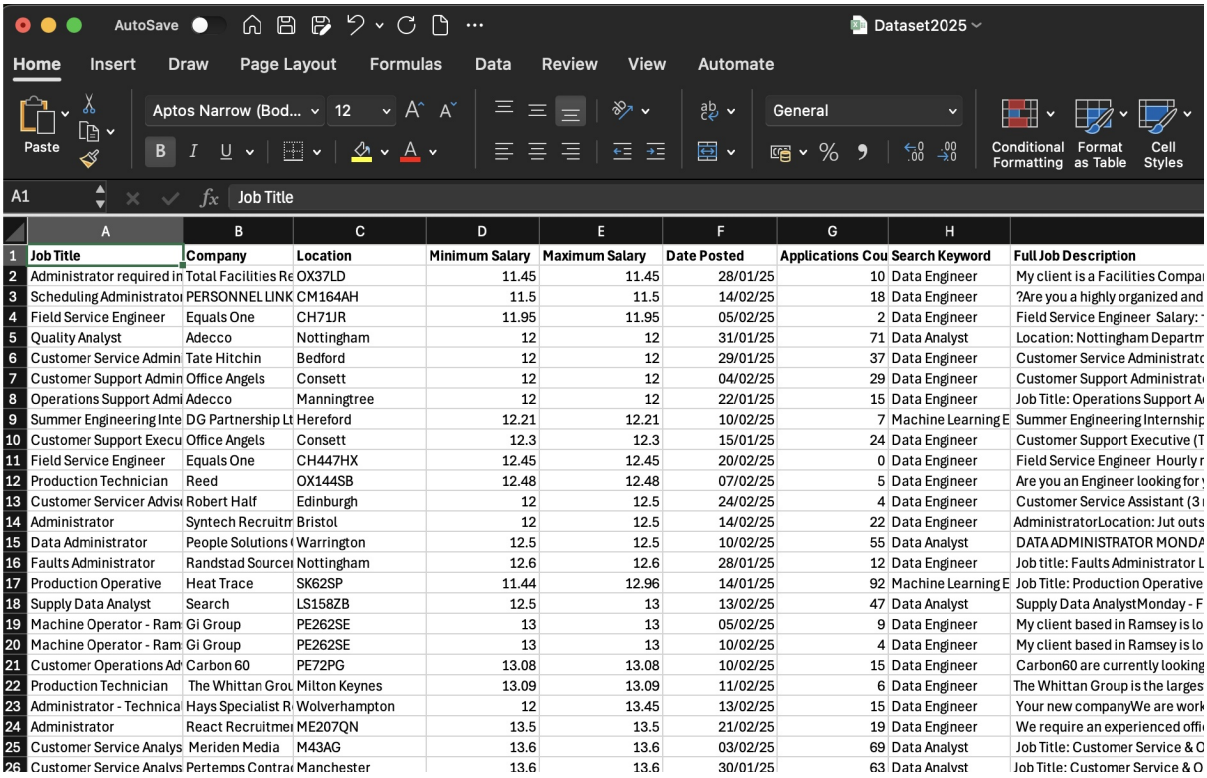


Figure 4: Screenshot of Web Scraping Code

.1.1 Dataset Sample Screenshot After Web Scraping

Access Dataset



	A	B	C	D	E	F	G	H	
1	Job Title	Company	Location	Minimum Salary	Maximum Salary	Date Posted	Applications Cou	Search Keyword	Full Job Description
2	Administrator required in	Total Facilities Re	OX37LD	11.45	11.45	28/01/25	10	Data Engineer	My client is a Facilities Compa
3	Scheduling Administrator	PERSONNEL LINK	CM164AH	11.5	11.5	14/02/25	18	Data Engineer	?Are you a highly organized and
4	Field Service Engineer	Equals One	CH71JR	11.95	11.95	05/02/25	2	Data Engineer	Field Service Engineer Salary:
5	Quality Analyst	Adecco	Nottingham	12	12	31/01/25	71	Data Analyst	Location: Nottingham Departm
6	Customer Service Admin	Tate Hitchin	Bedford	12	12	29/01/25	37	Data Engineer	Customer Service Administrato
7	Customer Support Admin	Office Angels	Consett	12	12	04/02/25	29	Data Engineer	Customer Support Administrat
8	Operations Support Admi	Adecco	Manningtree	12	12	22/01/25	15	Data Engineer	Job Title: Operations Support A
9	Summer Engineering Inte	DG Partnership Lt	Hereford	12.21	12.21	10/02/25	7	Machine Learning E	Summer Engineering Internship
10	Customer Support Execu	Office Angels	Consett	12.3	12.3	15/01/25	24	Data Engineer	Customer Support Executive (T
11	Field Service Engineer	Equals One	CH447HX	12.45	12.45	20/02/25	0	Data Engineer	Field Service Engineer Hourly r
12	Production Technician	Reed	OX144SB	12.48	12.48	07/02/25	5	Data Engineer	Are you an Engineer looking for
13	Customer Servicer Advis	Robert Half	Edinburgh	12	12.5	24/02/25	4	Data Engineer	Customer Service Assistant (3
14	Administrator	Syntech Recruitm	Bristol	12	12.5	14/02/25	22	Data Engineer	AdministratorLocation: Jut outs
15	Data Administrator	People Solutions (	Warrington	12.5	12.5	10/02/25	55	Data Analyst	DATA ADMINISTRATOR MONDA
16	Faults Administrator	Randstad Sourcer	Nottingham	12.6	12.6	28/01/25	12	Data Engineer	Job title: Faults Administrator L
17	Production Operative	Heat Trace	SK62SP	11.44	12.96	14/01/25	92	Machine Learning E	Job Title: Production Operative
18	Supply Data Analyst	Search	LS158ZB	12.5	13	13/02/25	47	Data Analyst	Supply Data AnalystMonday - F
19	Machine Operator - Ram	Gi Group	PE262SE	13	13	05/02/25	9	Data Engineer	My client based in Ramsey is lo
20	Machine Operator - Ram	Gi Group	PE262SE	13	13	10/02/25	4	Data Engineer	My client based in Ramsey is lo
21	Customer Operations Ad	Carbon 60	PE72PG	13.08	13.08	10/02/25	15	Data Engineer	Carbon60 are currently looking
22	Production Technician	The Whittan Grou	Milton Keynes	13.09	13.09	11/02/25	6	Data Engineer	The Whittan Group is the larges
23	Administrator - Technica	Hays Specialist R	Wolverhampton	12	13.45	13/02/25	15	Data Engineer	Your new companyWe are work
24	Administrator	React Recruitmer	ME207QN	13.5	13.5	21/02/25	19	Data Engineer	We require an experienced offi
25	Customer Service Analys	Meriden Media	M43AG	13.6	13.6	03/02/25	69	Data Analyst	Job Title: Customer Service & C
26	Customer Service Analys	Pertemps Contrac	Manchester	13.6	13.6	30/01/25	63	Data Analyst	Job Title: Customer Service & C

Figure 5: Screenshot of Dataset Sample After Web Scraping

.2 Data Cleaning Code Link

Access Data Cleaning and Processing Code

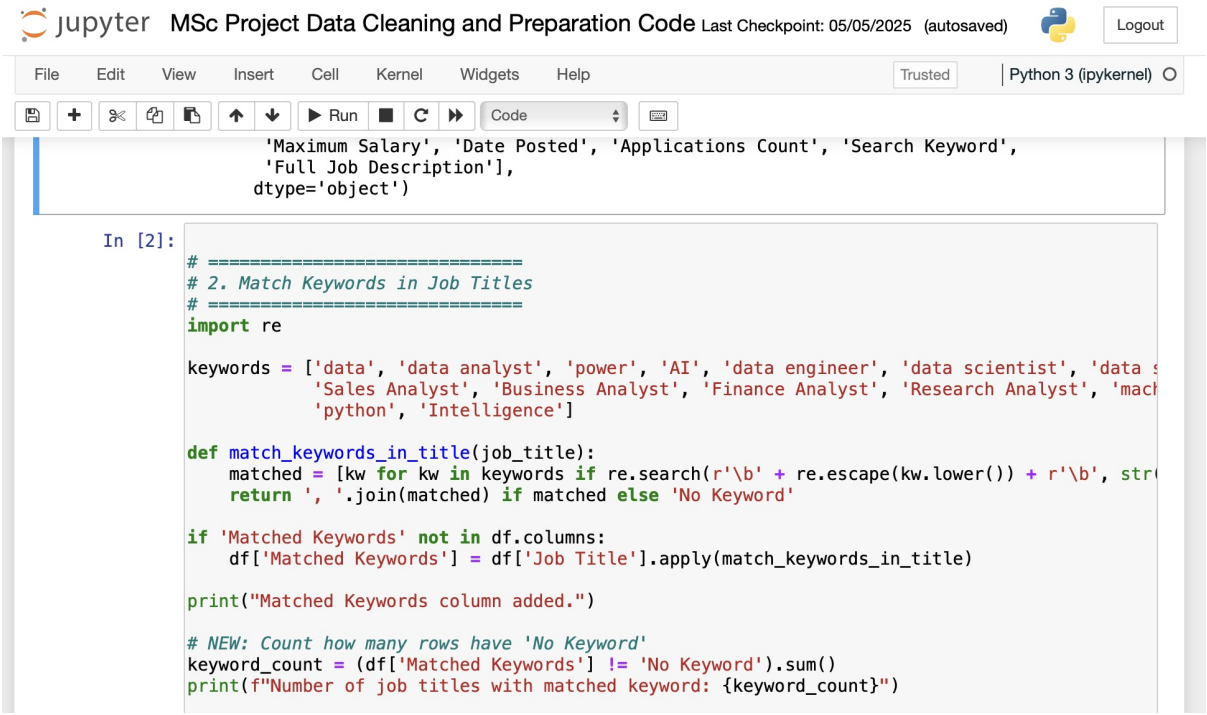


Figure 6: Screenshot of Data Cleaning Code

.3 Skill Extraction Code (Network Analysis)

Access Skill Extraction Code

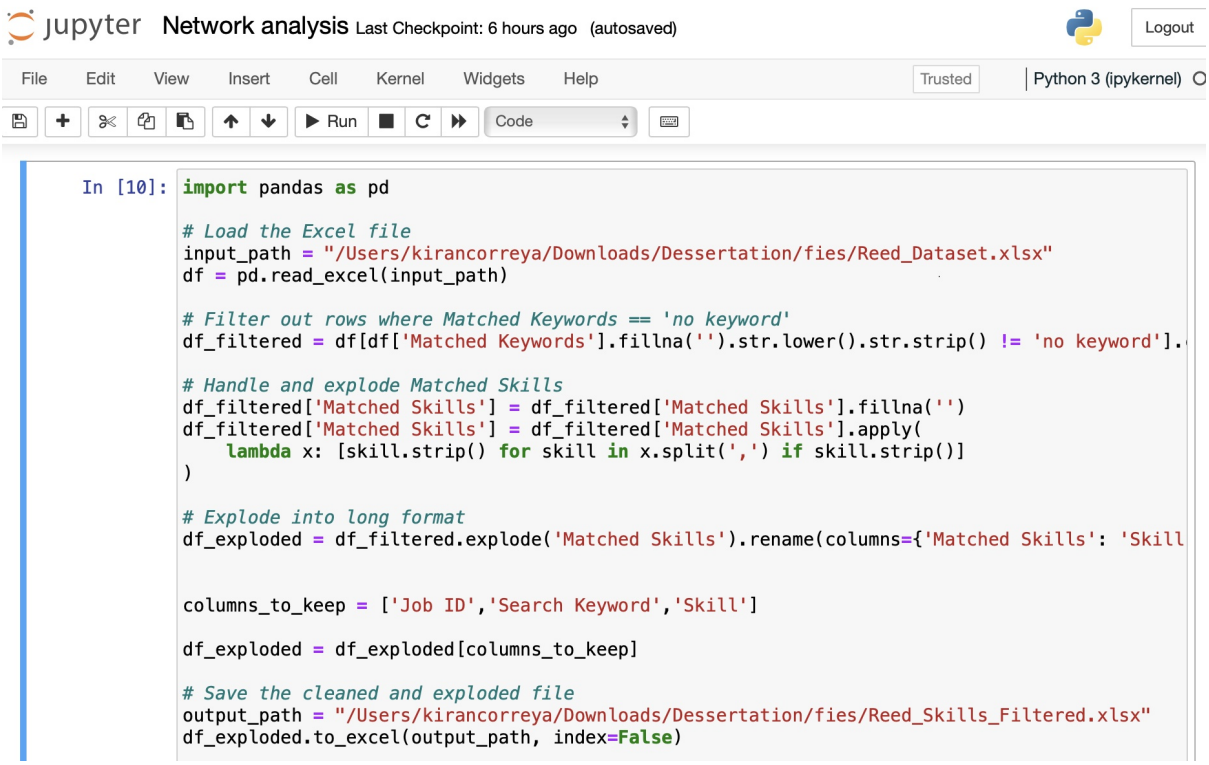


Figure 7: Screenshot of Skill Extraction Code

.4 Word Cloud Job Description Cleaning Code Link

Access Word Cloud Generation Code

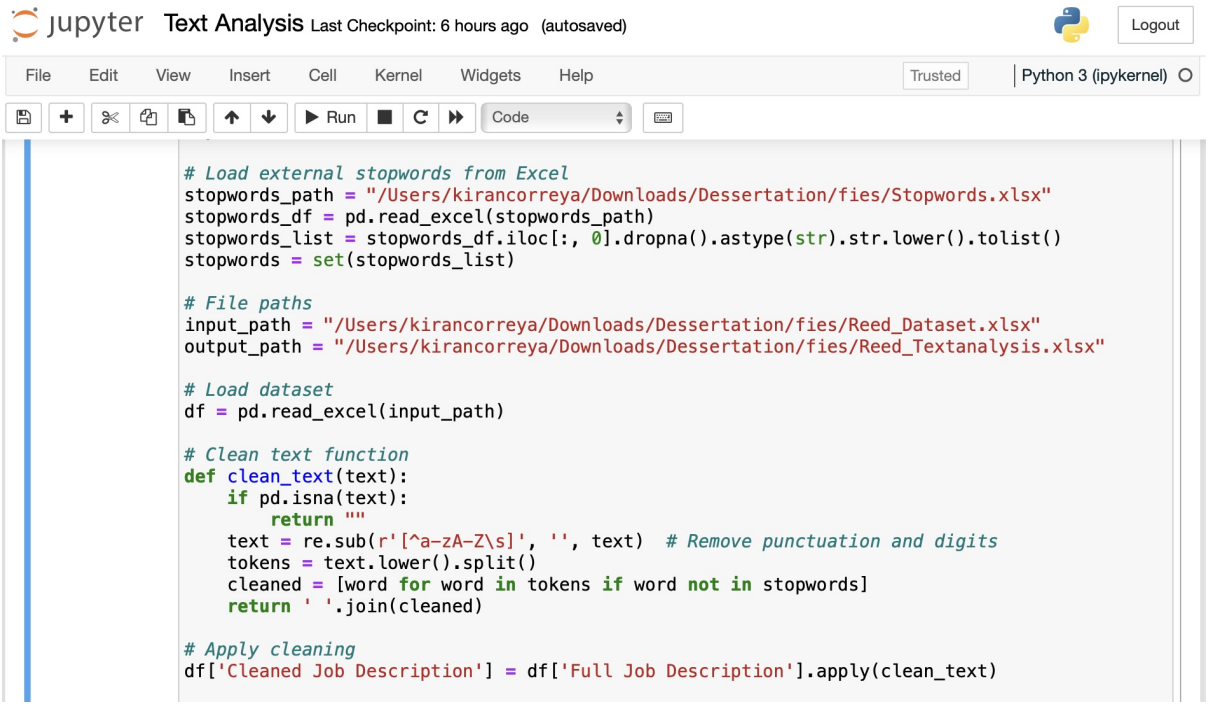


Figure 8: Screenshot of Job Description Cleaning Code