**Group 1 – Business Running Case: Evaluating Personal Job Market Prospects**

**Business Analytics, Data Science, and Machine Learning Trends**

A blue light on a screen

AI-generated content may be incorrect.

Authors

Moiz Deshmukh

Prabu Jayabalan

Andrey Pafnutyev

Jason Stopas

Jitvan Vij

Published

April 29, 2025

Boston University Metropolitan College

**Table of Contents**

Rationale Statement ……………………………………………. Pg. 1

Literature Review ……………………………………………. Pgs. 2-5

Data Cleaning & Preprocessing ……………………………………………. Pgs. 6-7

Exploratory Data Analysis ……………………………………………. Pgs. 8-14

Skill Gap Analysis ……………………………………………. Pgs. 15-23

Career Strategy ……………………………………………. Pgs. 34-26

Supervised Machine Learning ……………………………………………. Pgs. 27-38

Unsupervised Machine Learning ……………………………………………. Pgs. 39-46

References ……………………………………………. Pg. 47

**Rationale Statement:**

Firstly, on behalf of all of us in Group 1 - welcome.

After careful consideration, we have collectively chosen to focus our research efforts on Business Analytics, Data Science, and Machine Learning Trends. The primary motivation for selecting this topic is that, in the long run, a deeper understanding of these areas will significantly strengthen our professional expertise and career prospects. Among the various research options available, we believe this theme holds the greatest potential for long-term impact as we pursue careers in the business analytics and data science disciplines.

Our goal is to conduct thorough and insightful research that provides a reliable understanding of the industry’s landscape at large. As is outlined in the assignment description, we aim to explore the most in-demand skills, examine how job descriptions have evolved in recent years, identify the leading employers of both business analysts and data scientists, and to assess the career outlook of these roles.

If approached strategically, the exploration will serve as a valuable resource, providing knowledge we can revisit and build upon throughout our careers - even as the industry continues to evolve.

Sincerely,

Group 1

**Literature Review: Data Science, Business Analytics, and Machine Learning Trends**

**Introduction**

The demand for professionals in data science, business analytics, and machine learning (ML) has surged as organizations increasingly rely on data-driven decision-making. Based on recent research, the most sought-after skills in these fields can be categorized into technical (hard skills) and non-technical (soft skills).

**In-Demand Skills for Data Science, Business Analytics, and ML Roles**

**Hard Skills**

* Hard skills remain a fundamental requirement for data science (Umamaheswaran et al., 2023), business analytics, and ML roles.

**Key technical competencies include:**

* Programming Languages: Python, R, SQL, and Java are among the most demanded languages, with Python being particularly crucial for ML and data analytics applications.
* Big Data and Cloud Technologies: Expertise in cloud platforms such as AWS, Azure, and Google Cloud, along with tools like Hadoop and Spark, is increasingly valuable.
* Machine Learning & AI: Knowledge of machine learning frameworks such as TensorFlow, PyTorch, and Keras, along with expertise in neural networks and deep learning, is becoming a core requirement for ML-based roles.
* Data Management and Visualization: Business Intelligence (BI) tools such as Tableau, Power BI, and data warehousing techniques are critical for business analytics roles.
* Mathematics & Statistics: A strong grasp of statistics, probability, and mathematical modeling is necessary for data-driven roles.

**Soft Skills**

While technical expertise is crucial (Musazade, N. (2022), employers also prioritize soft skills that enable professionals to apply their knowledge effectively:

* Communication & Stakeholder Management: Business analytics professionals require strong communication skills to translate data-driven insights into actionable business strategies.
* Problem-Solving & Critical Thinking: Employers value candidates who can apply analytical thinking to solve complex business problems and optimize decision-making processes.
* Teamwork & Leadership: Collaborating across teams and managing projects efficiently are critical skills, particularly for leadership roles in data science and analytics.

**Evolving Skill Trends**

The increasing integration of machine learning in business analytics suggests a convergence between data science and business intelligence roles. Furthermore, automation and AI-powered analytics are shaping the future of these professions, emphasizing the need for adaptability and continuous learning.

**The Essential Competencies of Data Scientists: A Framework for Hiring and Training**

The article, The Essential Competencies of Data Scientists: A Framework for Hiring and Training by Motahareh Zarefard and Nicola Marsden (Zarefard & Marsden, 2024), investigates the skills needed for data scientists in 2024 and ‘looking’ ahead into the future. As stated in the article, an increasing demand for data scientists and a reported difficulty in finding qualified professionals, the authors identify and categorize 130 distinct competencies across seven KSAEOs that are deemed to be essential for success within the role.

These include: 1) functional, 2) ethical, 3) cognitive, 4) awareness, 5) social, 6) organizational, and 7) behavioral skills.

* Functional includes technical skills like machine learning comprehension and programing.
* Ethical covers responsible data usage and compliance.
* Cognitive includes analytical thinking, problem-solving, and creativity.
* Awareness highlights organizational knowledge, social intelligence, and self-awareness.
* Social covers communication skills, teamwork, and leadership skills.
* Organizational relates to managerial and strategic proficiencies.
* Lastly, behavioral involves adaptableness, resilience, and even mentions the entrepreneurial spirit.

The study uses data analytics methods that combine literature review, statistical analysis, and text mining to classify and measure the defined skills. By combining information from job postings and from within industry, a model is created. The model provides a deeper insight into the competencies and can be used for optimized training and professional development (plus academic as well). Their conclusions ultimately demonstrate that those who are well-rounded [a balance of both technical and non-technical skills] are best positioned to succeed in data analytics roles.

**Evolution of Job Descriptions in 2024 with AI/ML Expertise**

According to the article Artificial Intelligence and Employment Transformation: A MultiSector Analysis of Workforce Disruption and Adaptation by Kanagarla Krishna Prasanth Brahmaji (2024), job descriptions have evolved significantly in 2024 to reflect the growing demand for AI/ML expertise (Kanagarla, 2024).

**The study highlights the following key changes:**

* Increased Demand for AI/ML Skills: A 75% increase in demand for data analytics skills and a 41% gap in AI/ML expertise across various sectors have emerged as critical concerns for employers.
* Sector-Specific Evolution: Manufacturing (43% adoption), Financial Services (39% implementation), Healthcare (41% growth) show significant AI/ML integration.
* Skill Profile Adjustments: 92% of job roles now require multi-domain digital competencies, with cross-functional expertise earning professionals a 45% salary premium.
* Training and Development Shifts: A 163% increase in digital skills training programs underscores the need for continuous learning.

**Industries Hiring the Most Data Scientists and Why**

The demand for data scientists has surged across multiple industries as businesses increasingly rely on artificial intelligence (AI), machine learning, and big data analytics to drive decision-making. According to recent studies, the data science job market is projected to grow by 35% between 2022 and 2032.

**Key Industries Driving Data Science Hiring**

* Healthcare and Biomedicine: AI-driven medical research and predictive diagnostics are expanding (Bzdok et al., 2024), with the healthcare analytics market expected to reach $129.7 billion by 2028.
* Finance and Banking: The financial analytics market, growing at an 11.3% CAGR, utilizes data science for fraud detection, risk assessment, and algorithmic trading.
* Technology and AI: The technology sector drives demand for professionals in NLP, robotics, and cloud computing, with the AI market projected to hit $2 trillion by 2030.
* Retail and E-commerce: E-commerce companies use AI for supply chain optimization and customer personalization, with the market expected to reach $11.1 billion by 2028.
* Government and Public Policy: Data science aids urban planning, national security, and policy analysis.

**Career Outlook for Business Analytics Professionals**

The career outlook for business analytics professionals is exceptionally promising due to the growing reliance on data-driven decision-making across industries. Organizations are increasingly seeking individuals who can analyze complex datasets to optimize operations and support strategic decision-making (Noble Desktop, 2024).

**Factors Influencing Career Growth**

* High Demand Across Industries: Data analytics professionals are in high demand in finance, healthcare, technology, and consulting.
* Educational Pathways and Skill Development: Universities are introducing programs to address the skill gap.
* Career Advancement Opportunities: Roles such as senior business analyst, project manager, and CIO present potential growth paths.

**Conclusion**

The increasing reliance on data-driven insights is propelling the demand for professionals skilled in data science, business analytics, and machine learning. As industries continue to evolve with technological advancements, these fields promise robust career prospects and opportunities for innovation.

**Data Cleaning & Preprocessing**

During this project, we analyzed the “lightcast\_job\_postings.csv” dataset, which contains detailed job market information, including job titles, companies, locations, salaries, and various metadata. The dataset initially comprised 131 columns, offering a comprehensive view of job postings and associated attributes for evaluating personal job market prospects in 2024.

NOTE: *shape* of data frame at onset: (72498, 131)

**To prepare the dataset for analysis, we undertook a thorough data cleaning and preprocessing process, including:**

1. Dropping Irrelevant Columns:

* We removed columns that were either redundant or not relevant to our analysis. These included unique identifiers (ID), URLs of job postings (URL, ACTIVE\_URLS), and columns providing less granular versions of NAICS and SOC codes (NAICS2, SOC\_3, etc.).
* The rationale for dropping these columns was to enhance data efficiency and clarity, reduce the dataset size, and focus on the most granular and meaningful data.

A blue and white triangle with black lines

AI-generated content may be incorrect.

1. **Handling Missing Values:**

* As seen in the heatmap above, it reveals that the columns ACTIVE\_SOURCES\_INFO, MODELED\_DURATION, and MODELED\_EXPIRED contain significant missing values, suggesting potential data collection or extraction issues. Notably, DURATION and EXPIRED exhibit a moderate positive correlation (0.5), indicating that longer job durations might influence missingness in related columns.
* Additionally, ACTIVE\_SOURCES\_INFO shows a strong negative correlation with other variables, implying a pattern where missing data in this column might coincide with gaps in others.
* In contrast, columns like MSA, MSA\_NAME, and LIGHTCAST\_SECTORS have no missing values, providing a reliable foundation for further analysis. Addressing missing data in correlated columns through targeted imputation can prevent bias and enhance the accuracy of our analysis.

1. **Removing Duplicates:**

* To maintain unique job postings, duplicates were removed based on job title, company, location, and posting date.
* This ensured that each job opportunity was only represented once, preventing skewed analysis results.

**Impact of Data Cleaning**

The data cleaning process resulted in a more streamlined and manageable dataset, eliminating redundancy and potential inconsistencies. This step set the foundation for accurate and insightful analysis in the subsequent phases of the project.

**Exploratory Data Analysis (EDA)**

A group of blue and white graphs

AI-generated content may be incorrect.

**Top 15 Job Posting Industries**

A pie chart with numbers and text

AI-generated content may be incorrect.

In our analysis, we looked at the top industries with the most job postings to get a sense of where the demand is highest. The pie chart above shows that **22.6%** of the postings fall under the **Unclassified Industry** category, suggesting that many roles either span multiple sectors or lack clear classification. This made us consider the potential limitations in how industries are labeled in the data.

We also noticed a strong demand for skills in **technology and consulting**. For example, **Custom Computer Programming Services** made up **12.1%** of the postings, while **Management Consulting Services** accounted for **11.3%**. This highlights a significant need for both tech and management skills in today’s job market.

Interestingly, several tech-focused industries, such as **Software Publishers** and **Computer Systems Design Services**, showed up prominently in the chart. This aligns with the growing demand for IT and consulting professionals, which didn’t surprise us given the ongoing digital transformation across industries.

We also found that **finance and healthcare** sectors have a notable share of job postings, indicating steady demand in these fields. On the flip side, areas like **Accounting** and **Temporary Help Services** had fewer listings, suggesting these might be more niche markets.

**Monthly Trend of Data-Related Job Postings by Role**

A graph of data showing different colored lines

AI-generated content may be incorrect.

In the above Stacked Area Chart, we focused on understanding the monthly trends for various data-related job roles. To do this, we filtered the dataset to include only job titles containing keywords like “Data Analysts,” “Business Intelligence Analysts,” and “Data Engineers,” among others. After filtering, we converted the posting dates into a proper datetime format and created a new column to track monthly trends. We then grouped the data by month and job title to visualize how the demand for these roles evolved over time.

The stacked area chart we generated illustrates that the demand for data-related roles has shown a consistent upward trend, especially after mid-June 2024. One of the most noticeable patterns is the significant share held by roles such as **Big Data Analysts and Business Intelligence Analysts**, which suggests that organizations are increasingly seeking professionals who can handle large-scale data analysis and provide actionable insights.

Interestingly, we observed that the demand for **Data Analysts and Data Analytics Engineers** also increased steadily, indicating a growing need for professionals who can not only interpret data but also manage the infrastructure required for analysis. This aligns with the broader industry trend of integrating advanced analytics into decision-making processes.

We also found that more specialized roles like **Clinical Data Analysts and Customer Data Analysts** are gaining traction, although they represent a smaller share of the total postings. This could imply that industries such as healthcare and customer analytics are gradually expanding their analytics teams.

**Remote vs. On-Site Jobs (Data Roles)**

A pie chart with numbers and a few different colored circles

AI-generated content may be incorrect.

We created a Pie Chart to explore the distribution of remote, on-site, and hybrid roles for data-related jobs. The chart shows that **73.8%** of postings do not specify a preference, suggesting either data gaps or employer flexibility. Among specified roles, **20.4% are remote**, indicating a strong demand for remote work. Hybrid roles account for **4%**, while fully on-site roles are only **1.79%**.

These findings suggest a clear shift towards remote and flexible work arrangements in the data field. Highlighting remote work skills could be advantageous for job seekers in this area.

**Data-Related Job Postings Across the USA**

A map of the united states

AI-generated content may be incorrect.

We created a **Choropleth Map** to understand the geographical distribution of data-related job postings across the United States. To do this, we aggregated the job counts by state and converted full state names to abbreviations to ensure consistency in the map. We then used a choropleth map to visualize how job postings are spread across different states.

From the map, we can see that **states like California, Texas, and Florida** have the highest concentrations of data-related job postings, indicated by the brighter yellow shades. This pattern suggests that these states are major hubs for data-related roles, likely due to their large tech industries and the presence of major corporations.

In contrast, **several states in the Midwest and some in the Mountain region** show darker shades, indicating fewer job opportunities in data-related fields. This could imply either a lower demand for such roles or a smaller tech industry presence in those areas.

**Industries Having Data-Related Roles**

A close-up of words

AI-generated content may be incorrect.

With this WordCloud, we aimed to identify the industries that most frequently offer data-related roles. We extracted industry names from the dataset, combined them into a single text string, and then generated a word cloud to visualize which industries appeared most often.

From the word cloud, we can see that **“Unclassified Industry,” “Programming Services,” “Consulting Services,” and “Insurance Carriers”** are some of the most prominent terms. The large size of these words suggests that these industries have a significant number of data-related job postings. This pattern highlights the growing importance of data roles in both tech-heavy sectors like programming and consulting as well as in more traditional fields such as insurance.

Other notable industries in the word cloud include **“Computer Systems Design,” “Employment Placement Agencies,” and “Direct Health Services.”** Their visibility suggests a broad demand for data professionals across diverse sectors, emphasizing that data analytics skills are valuable beyond just the tech industry.

This indicates that the demand for data roles is not limited to a few industries but is spread across both tech-focused and traditional sectors. This reinforces the idea that data skills are becoming essential across the board.

**Key Findings**

1. **Industry Demand:**

* Programming Services, Consulting Services, and Insurance emerged as leading industries for data-related roles, indicating a widespread need for data skills beyond traditional tech sectors.
* The prominence of Unclassified Industry suggests potential gaps in data classification or a diverse range of roles that do not fit into conventional categories.

1. **Geographical Distribution:**

* California, Texas, and Florida were identified as major hubs for data-related jobs, while several Midwestern and Mountain states showed fewer opportunities.
* This pattern suggests that professionals may find more job prospects by focusing on these high-demand states.

1. **Work Arrangement Preferences:**

* A significant share of job postings preferred remote and hybrid roles, with over 20% specifically offering remote options.
* The limited proportion of fully on-site roles reflects a broader shift towards flexible work models in the data industry.

1. **Role-Specific Trends:**

* There is a clear upward trend in demand for roles such as Big Data Analysts and Business Intelligence Analysts, highlighting the growing importance of both technical and analytical skills.
* Specialized roles like Clinical Data Analysts and Customer Data Analysts are also gaining traction, indicating expanding opportunities in niche areas.

**Conclusion**

Our analysis revealed that the demand for data-related skills is both substantial and diverse, spanning multiple industries and regions across the United States. Key sectors such as tech, consulting, and insurance show the most significant opportunities, while states like California, Texas, and Florida lead in job postings.

The strong preference for remote and hybrid roles highlights the importance of flexibility in the current job market. Meanwhile, the consistent rise in demand for specialized data roles suggests a promising outlook for professionals equipped with both analytical and domain-specific skills.

Overall, these findings suggest that focusing on high-demand industries, enhancing remote work capabilities, and acquiring specialized skills can significantly boost job prospects in the data field.

**Skill Gap Analysis**

**Team Skill Insights**

**A screenshot of a computer

AI-generated content may be incorrect.**

From the matrix above, we observe:

* **Jason** shows consistently strong proficiency across all listed tools, especially in SQL and Excel.
* **Jitvan** excels in Power BI and SQL, but may benefit from further development in Python.
* **Moiz** demonstrates advanced proficiency in Excel and Tableau, indicating strong visualization and reporting capabilities.
* **Andrey** is an expert in Python but may require further training in Power BI.
* **Prabu** maintains intermediate-to-advanced competency across all tools, making him a well-rounded contributor.

**Team Skill Levels – Heatmap Analysis**

A grid of red and blue squares

AI-generated content may be incorrect.

The heatmap above offers a visual overview of our team’s self-assessed proficiency across five core analytical tools: **Python, SQL, Excel, Power BI, and Tableau**. The color intensity reflects the skill level, with darker shades representing higher proficiency (scale: 1 = Beginner, 5 = Expert).

**Key Insights:**

* **Python**: Andrey stands out with expert-level skills (5), while Jitvan shows the lowest proficiency (2), indicating a potential area for development.
* **SQL**: Jason and Jitvan both score the highest (5), reflecting strong database querying capabilities, while others maintain intermediate competency.
* **Excel**: Moiz and Jason demonstrate the highest proficiency (5), useful for data wrangling and reporting.
* **Power BI**: Jitvan leads with a top score (5), suggesting strong data visualization capabilities; however, Andrey scores the lowest (2), indicating a potential gap.
* **Tableau**: All team members cluster around the intermediate level (3–4), showing consistent yet improvable skills across the board.

This heatmap complements the earlier skill matrix by allowing a quick comparative glance, which is particularly helpful in identifying areas of individual strength and skill gaps that may benefit from team-wide upskilling initiatives.

**Top 5 In-Demand Software Skills in IT (Industry-Level)**

A graph of a bar graph

AI-generated content may be incorrect.

The bar chart above highlights the top five most frequently requested software-related skills in IT job postings, based on the extracted data from the SOFTWARE\_SKILLS\_NAME column in the Lightcast dataset.

**Observations:**

* **SQL (Programming Language)** ranks as the most in-demand skill across job postings, signaling its foundational importance in data querying and backend systems.
* **Microsoft Excel** continues to be widely valued, especially in roles that demand reporting, data cleaning, and spreadsheet-based operations.
* **Python (Programming Language)**, a critical skill for automation, data analysis, and machine learning, ranks third — affirming its strong market relevance.
* **SAP Applications**, often used in enterprise resource planning (ERP) and supply chain systems, indicates a demand for professionals with domain-specific technical expertise.
* **Dashboard** skills, representing tools like Tableau, Power BI, and similar platforms, round out the top five — showcasing the need for data visualization and communication capabilities.

These findings serve as a benchmark for evaluating whether our team’s skill sets align with current industry expectations — and will help guide upskilling priorities in the subsequent gap analysis.

**Team Skills vs. Top 5 In-Demand IT Skills**

A diagram of a skill level

AI-generated content may be incorrect.

This heatmap visualizes the alignment between our team’s current skill levels and the top 5 most frequently requested IT skills from job postings: **SQL, Excel, Python, SAP, and Tableau**.

**Key Observations:**

* **SAP Applications** show a complete skills gap across the team — no member currently reports proficiency, indicating a strong potential area for upskilling based on industry demand.
* **SQL** and **Excel** are relatively well-covered, with multiple team members (e.g., Jason, Jitvan, Moiz) scoring between 4–5, reflecting strong database and spreadsheet competencies.
* **Python** remains an essential and in-demand skill. While most team members show intermediate-to-high proficiency, one member (Jitvan) has a lower score of 2, indicating a learning opportunity.
* **Tableau**, frequently requested under the category “Dashboard”, is evenly represented, with all members scoring between 3 and 4 — suggesting a consistent but improvable baseline.

Takeaway:

This mapped heatmap offers a focused lens on how well our collective capabilities align with current market expectations. Notably, **SAP** stands out as an urgent area for learning, while ongoing refinement in **Tableau** and **Python** can further improve the team’s employability and project readiness.

**TOP 5 IN DEMAND SPECIALIZED SKILLS IN IT**

**A graph of a graph with text

AI-generated content may be incorrect.**

The horizontal bar chart above identifies the five most sought-after specialized skills extracted from the SPECIALIZED\_SKILLS\_NAME column within the Lightcast job postings dataset. These reflect deeper domain-specific capabilities expected by employers across the IT sector.

**Observations:**

* **Data Analysis** emerged as the most frequently mentioned specialized skill, underlining its critical role in transforming raw information into actionable insights across industries.
* **SQL (Programming Language)** maintains its strong presence, reinforcing its dual role as both a software and specialized technical skill essential for data management and querying.
* **Computer Science** appears prominently, indicating employers’ preference for candidates with a foundational understanding of algorithms, system architecture, and programming principles.
* **Project Management** ranks high, suggesting that beyond technical proficiency, employers are seeking professionals who can also manage timelines, deliverables, and stakeholder expectations effectively.
* **Business Process** knowledge also stands out, pointing to demand for skills that bridge technical solutions with operational efficiency.

**Interpretation:**

These specialized skills reflect a blend of technical expertise and operational acumen. When compared to our team’s current skills, some gaps — such as **Project Management**, **Business Process**, and **Data Analysis** — may represent areas of opportunity for targeted learning and future curriculum alignment.

**Top In-Demand Common (Soft) Skills in IT**

A graph of a number of people

AI-generated content may be incorrect.

The chart above showcases the top five most frequently listed **common skills** — often referred to as **soft skills** — based on the COMMON\_SKILLS\_NAME column in the job postings dataset. These are non-technical but highly valued in IT roles across all organizational levels.

**Observations:**

* **Communication** ranks highest, emphasizing the critical need for professionals who can clearly articulate ideas, collaborate effectively across teams, and convey technical findings to non-technical stakeholders.
* **Management** and **Leadership** both appear in the top five, reflecting the industry’s growing need for individuals who can not only execute tasks but also lead initiatives, manage teams, and drive strategic outcomes.
* **Problem Solving** is another top-listed trait, aligning with the complex, analytical nature of most IT roles where troubleshooting and innovative thinking are routine.
* **Operations** rounds out the list, suggesting a demand for process-oriented thinking, efficiency, and understanding of business workflows.

**Interpretation:**

While often overlooked in technical training, these soft skills are crucial differentiators in hiring decisions. Regardless of specialization (data science, software development, or project management), **interpersonal and leadership abilities significantly enhance employability**. This insight reinforces the importance of combining technical expertise with strong communication and organizational skills in both curriculum design and individual development plans.

**Top 5 In-Demand IT Skills by Category – Treemap Visualization**

A screenshot of a computer screen

AI-generated content may be incorrect.

The treemap above presents a unified visual overview of the most in-demand IT skills, categorized into three primary groups: **Common**, **Specialized**, and **Software**. Each rectangle’s size represents the relative frequency of that skill across job postings.

**Category Breakdown:**

* **Common Skills (orange):**
  + This category dominates the overall space, with **Communication** and **Management** leading the way.
  + Other prominent soft skills include **Problem Solving**, **Leadership**, and **Operations**, reinforcing the industry’s continued emphasis on interpersonal and leadership competencies alongside technical abilities.
* **Specialized Skills (Mediumpurple)**:
  + **Data Analysis** and **SQL (Programming Language)** are top mentions, reflecting their value in strategic decision-making and backend infrastructure.
  + Skills such as **Computer Science**, **Project Management**, and **Business Process** suggest a need for foundational technical understanding coupled with domain-specific operational insights.
* **Software Skills (Skyblue)**:
  + These are tools and technologies actively used in IT workflows.
  + **SQL**, **Excel**, and **Python** appear again here, reaffirming their dual presence as both specialized knowledge and hands-on tools.
  + **SAP Applications** and **Dashboard** (referring to visualization platforms like Tableau and Power BI) reflect enterprise-level technical expectations.

**Interpretation:**

This treemap effectively integrates the earlier bar charts into one cohesive view, allowing for quick comparative analysis across skill types. It visually reinforces the **multi-dimensional expectations of IT professionals** — who must blend communication, strategic insight, and software proficiency to meet market demands.

As we move toward building an improvement plan, this treemap helps prioritize upskilling areas based on both category-level and individual skill-level significance.

**Carrer Strategy**

**Proposing an Improvement Strategy for our Team**

🔹 **Andrey**

* **Skills to Prioritize Learning**: Power BI, Tableau, SAP Applications
* **Courses or Resources**:
  + Microsoft Power BI Data Analyst
  + Introduction to Tableau – UC Davis
  + SAP Professional Fundamentals
* **Team Collaboration Suggestion**:
  + Can shadow Jason and Prabu on Excel reporting dashboards.
  + Organize peer-led walkthroughs with Moiz on Tableau once Moiz completes his course.

🔹 **Moiz**

* **Skills to Prioritize Learning**: Python, Power BI, SAP Applications
* **Courses or Resources**:
  + Python for Everybody – University of Michigan
  + Power BI for Beginners – Microsoft
  + Becoming an SAP Professional
* **Team Collaboration Suggestion**:
  + Pair with Jason to co-develop a dashboard and improve Power BI skills.
  + Conduct weekly learning swaps with Jitvan to practice Python.

🔹 **Jason**

* **Skills to Prioritize Learning**: Power BI, SAP Applications
* **Courses or Resources**:
  + Microsoft Power BI for Beginners
  + SAP Technology Consultant – SAP
* **Team Collaboration Suggestion**:
  + Lead weekly recap sessions on SQL with teammates.
  + Practice SAP module navigation with Moiz and Andrey after completing the course.

🔹 **Prabu**

* **Skills to Prioritize Learning**: SAP Applications
* **Courses or Resources**:
  + Implementing an SAP Solution
* **Team Collaboration Suggestion**:
  + Lead a team-wide “SAP Sunday” mini hackathon once a month for hands-on SAP tasks.
  + Assist others in brushing up Python and Excel via peer mentoring.

🔹 **Jitvan**

* **Skills to Prioritize Learning**: Python, Excel, Tableau
* **Courses or Resources**:
  + Crash Course on Python – Google
  + Excel Skills for Business – Macquarie University
  + Data Visualization with Tableau – UC Davis
* **Team Collaboration Suggestion**:
  + Co-present Tableau project findings with Moiz to practice and receive feedback.
  + Schedule weekly “code & coffee” Python debugging sessions with the team.

**Summary Collaboration Ideas**

* Assign a “Skill Champion” for each domain (e.g. SQL – Jason, Tableau – Moiz, Python - Prabu) to mentor others.
* Implement weekly 30-minute peer-learning check-ins.
* Maintain a shared Notion or Google Doc for learning progress, notes, and resources.
* Host monthly demo days to present completed mini-projects using new tools learned.

**Supervised Machine Learning**

To ensure data quality, we filtered out records with missing values in the SALARY column from the onset as SALARY is out ‘target’ variable. Note: This reduced our working dataset to 30808 rows.

**Annual Salary Distribution - Box Plot**

A graph of a graph

AI-generated content may be incorrect.

The box plot above provides a visual summary of the annual salary distribution within the dataset after filtering for non-null entries. The central box represents the interquartile range (IQR), which captures the middle 50% of salaries in the dataset. The line within the box denotes the **median salary**, which is approximately **$116,300**.

**Key observations:**

* The **minimum salary** is slightly above **$15,000**, while the **maximum non-outlier salary** is just below **$230,000**.
* Numerous data points are plotted as individual markers above the upper whisker, identifying them as **outliers**. These salaries extend up to **$500,000**, indicating the presence of extremely high-paying roles.
* The **interquartile range (IQR)** suggests that most salaries are concentrated between **$80,000 and $150,000**, which is a critical salary band for the job market analysis.

To ensure robust analysis and mitigate the impact of extreme values, salaries below **$50,000** and above **$230,000** were filtered out prior to regression modeling. This decision helps reduce skewness and prevents outliers from disproportionately influencing predictions.

**Handling Missing Experience Values**

A close-up of a sign

AI-generated content may be incorrect.

Next, we analyzed missing values in the MIN\_YEARS\_EXPERIENCE column. About 21.6% of the entries were missing. These were imputed using the **median value (5 years)**, which preserved the integrity of the dataset while addressing missingness in a statistically neutral way.

**Exploring the ONET Category**

We focused our analysis on a specific ONET category: **Business Intelligence Analysts**. The filtered dataset had approximately **28,971** entries under this job classification. Unique counts for ONET, ONET\_NAME, and ONET\_2019\_NAME confirmed the consistency of this subset.

A close-up of a card

AI-generated content may be incorrect.

**Bar Plot: Top 10 Business Intelligence Analyst Role Titles**

A graph of a business intelligence analyst

AI-generated content may be incorrect.

This bar chart illustrates the frequency distribution of the top 10 most common job titles under the ONET category “Business Intelligence Analysts.” The data was extracted from job postings where ONET classification was available and grouped by the TITLE\_CLEAN field.

The most frequently occurring role by a significant margin is “Data Analyst” with over 2,000 occurrences, followed by “Business Intelligence Analyst,” “Senior Data Analyst,” and “Enterprise Architect.” This suggests that the term “Data Analyst” is more widely used in job postings, even when referring to more specialized roles such as business intelligence positions.

The diversity of titles such as “Oracle HCM Cloud Implementation Lead,” “Lead Data Analyst,” and “Solution Architect” also reflects the multidisciplinary nature of business intelligence roles, spanning technical infrastructure, enterprise systems, and analytics.

This visualization supports the idea that while many job functions may fall under the business intelligence umbrella, employers label them in varied ways, highlighting the importance for job seekers to use a broad range of keywords when searching for roles in this domain.

**Data Cleaning and Refinement of Education-Level Information**

A screenshot of a computer

AI-generated content may be incorrect.

* To assess the role of education in job salary trends, we began by exploring the MIN\_EDULEVELS and MIN\_EDULEVELS\_NAME columns. These columns represent the minimum educational qualifications required for each job posting. The dataset includes a range of educational levels, from “High school or GED” to “Ph.D. or professional degree”, along with a placeholder value 99 indicating “No Education Listed”.
* A quick check confirmed that there were no missing (NaN) values in either column, which eliminated the need for imputation or data repair.
* We then examined the frequency distribution of different education levels. It was observed that “Bachelor’s degree” was the most commonly required qualification, followed by “No Education Listed” and “Associate degree”.
* To ensure analytical clarity and avoid skewing insights with ambiguous values, we removed rows where the education level was listed as “No Education Listed” (coded as 99). This helped refine the dataset to include only those records where education expectations were clearly stated, ensuring better model training and more interpretable results in subsequent steps.

**Correlation Analysis Between Salary and Job-Related Features**

To explore the relationship between salary and other variables such as job titles, experience, and education, we conducted a correlation analysis using the top 10 most frequent job titles.

**Data Preparation**:

* We identified the top 10 job titles using frequency counts and filtered the dataset accordingly.
* The relevant features for analysis included: SALARY, MIN\_YEARS\_EXPERIENCE, MIN\_EDULEVELS, and TITLE\_CLEAN.
* One-hot encoding was applied to the TITLE\_CLEAN column to make it suitable for correlation analysis.

**Correlation Matrix**:

A white screen with black text

AI-generated content may be incorrect.

* A correlation matrix was computed to quantify the strength and direction of relationships between SALARY and other features.
* The .corr() function was used to generate Pearson correlation coefficients.

**Findings**:

* MIN\_YEARS\_EXPERIENCE showed a moderately strong positive correlation with salary (0.47), indicating that more experience is generally associated with higher salaries.
* Certain job titles such as enterprise architect (0.42), oracle hcm cloud manager (0.39), and oracle hcm core hr module (0.29) also exhibited notable positive correlations with salary.
* Conversely, roles such as data analyst (0.45) and data and reporting professional (−0.17) demonstrated negative correlations, suggesting these positions typically receive lower salaries.

A colorful chart with red and blue squares

AI-generated content may be incorrect.

**Heatmap Interpretation**:

* The heatmap visually confirms these trends, with warmer tones (e.g., orange) representing stronger positive correlations and cooler tones (e.g., blue) showing negative correlations.
* Diagonal values are all 1.00, reflecting perfect self-correlation.
* Cross-comparisons (e.g., between titles and experience/education) indicate subtle multicollinearity that might be worth considering when selecting features for regression modeling.

This analysis was crucial in identifying the most informative features for predicting salary and helped guide our feature selection for the next steps in supervised modeling.

**Random Forest Model**

**A blue and black text

AI-generated content may be incorrect.**

**Feature Selection and Model Training**

To predict salary outcomes, we selected a set of relevant features including MIN\_YEARS\_EXPERIENCE, MIN\_EDULEVELS, and the top 10 job titles obtained from the cleaned TITLE\_CLEAN column through one-hot encoding. The dataset was split into training and testing sets with a 70/30 ratio using a fixed random state to ensure reproducibility. We implemented a Random Forest Regressor with 100 estimators.

A graph with text and numbers

AI-generated content may be incorrect.

**Feature Importance Visualization**

A bar chart was generated to visualize the importance of each feature in predicting salary. MIN\_YEARS\_EXPERIENCE emerged as the most influential feature, followed by titles such as enterprise architect, data analyst, and oracle hcm cloud implementation lead core hr module. This suggests that experience and specific job roles have a strong influence on salary prediction, while education level (MIN\_EDULEVELS) had moderately less impact in the model.

A computer error message

AI-generated content may be incorrect.

**Model Performance Evaluation**

The performance of the Random Forest Regression model was evaluated using the following metrics:

* R² Score: 0.5959
* Mean Absolute Error (MAE): 15,409.01
* Mean Squared Error (MSE): 513,392,249.80
* Root Mean Squared Error (RMSE): 22,658.16

These results indicate that while the model captures general salary trends based on experience and job role, there is room for improvement, potentially through feature engineering, additional variables, or advanced tuning techniques.

**Interpretation**

* A moderate R² score indicates that about 60% of the variance in salaries can be explained by the model.
* The residual error values suggest variability in salary predictions, potentially due to unobserved variables such as company reputation, geographic location, or job-specific skills.

**Residual Plot for the Random Forest Model**

**A graph with blue dots

AI-generated content may be incorrect.**

The residual plot above visualizes the difference between the actual salaries and the predicted salaries generated by the Random Forest Regressor. On the Y-axis, we have the residuals, which are calculated as:

**Residual = Actual Salary - Predicted Salary**

On the X-axis, we have the actual salary values. The horizontal red dashed line at Y=0 represents the ideal line where residuals would lie if predictions were perfect.

Observations:

* Most residuals are centered around the zero line, which indicates that the model’s predictions are generally close to the actual values.
* However, there is a noticeable **spread** of residuals at both the lower and higher ends of the salary range. This suggests **heteroscedasticity**, meaning the variance of the residuals increases with the actual salary.
* Some **outliers** are visible, where the predicted salary significantly underestimates or overestimates the actual salary by large margins.
* The plot maintains a relatively linear pattern without any distinct curvature, which is desirable as it supports the assumption that the model captures the relationship without strong systematic bias.

Conclusion:

While the model performs reasonably well for most observations, the presence of outliers and increasing residual spread at higher salary ranges indicate that the model may benefit from additional feature engineering or alternative modeling techniques to improve accuracy across the entire range.

**Salary Categorization and Confusion Matrix Interpretation**

To enhance interpretability of salary prediction outputs, we categorized the continuous salary values into three distinct bins: - **Low**: Salaries less than $60,000 - **Medium**: Salaries between $60,000 and $120,000 - **High**: Salaries greater than $120,000

These bins were applied to both the actual test set and the model’s predicted salaries. A confusion matrix was then constructed to assess the model’s classification performance across these categories.

A chart with numbers and a yellow square

AI-generated content may be incorrect.

Confusion Matrix Summary:

* **Low-End Salaries**: None were correctly classified. All 40 instances were incorrectly predicted as belonging to the “Medium” category.
* **Average (Medium) Salaries**: Out of 612 actual medium salary records, 587 were correctly classified, and 25 were misclassified as “High”.
* **High Earners**: Among 230 actual high-salary records, 155 were correctly classified, while 75 were misclassified as “Medium”.

Interpretation:

* The model demonstrates strong predictive performance for the **medium salary category**, which could be attributed to its higher representation in the training data.
* However, it **struggles to identify low-end salary jobs**, consistently misclassifying them as medium. This could be due to a class imbalance or model bias toward the center of the salary range.
* **High salaries are partially misclassified**, likely due to overlap in feature values with medium-salary roles.

This analysis highlights the need for further feature engineering or possibly applying class balancing techniques to improve classification performance across all salary categories.

**Salary Predictions Using Random Forest Model**

The Random Forest Regressor model was utilized to predict salaries for four hypothetical candidates based on their experience, education level, and job title. Below is a summary and interpretation of the predicted results:

Predicted Salary 01: $74,467.49

**Scenario 1:**

* Entry-level Data Analyst (Bachelor’s Degree, 1 Year of Experience)
* Predicted Salary: **$74,467.49**

This prediction reflects an early-career role, showing a relatively lower salary aligned with market expectations for individuals with minimal experience and a bachelor’s degree.

Predicted Salary 02: $123,480.98

**Scenario 2:**

* Entry-level Data Analyst (Master’s Degree, 1 Year of Experience)
* Predicted Salary: **$123,480.98**

With the same experience level as Scenario 1 but a higher education qualification, the model predicts a substantially higher salary. This illustrates how education level, particularly a master’s degree, significantly boosts compensation potential for data roles.

Predicted Salary 03: $149,786.45

**Scenario 3:**

* Senior Data Analyst (7 Years of Experience, Master’s Degree)
* Predicted Salary: **$149,786.45**

The increase in experience leads to a proportionate increase in predicted salary. This aligns with industry trends where senior analysts with advanced degrees command higher pay due to their skill maturity and potential leadership responsibilities.

Predicted Salary 04: $162,616.71

**Scenario 4:**

* Enterprise Architect (7 Years of Experience, Master’s Degree)
* Predicted Salary: **$162,616.71**

Among all scenarios, this role has the highest salary estimate. The title “Enterprise Architect” along with senior-level experience and advanced education demonstrates the highest value according to the model, consistent with real-world compensation structures for such strategic positions.

**Conclusion:**

These predictions showcase how variables such as years of experience, minimum education level, and job title contribute to salary estimations. The model effectively differentiates salary bands across role types and qualifications, providing valuable insights for career planning and market benchmarking.

**Unsupervised Learning Model**

**Text Preprocessing: Combine Job Title and Skills into a Single Field for TF-IDF**

* We combined the job title and skills into a single text field (combined\_text) to create a richer, unified input for the TF-IDF vectorizer. This improves the quality of feature extraction by capturing more context about each job, enabling better clustering and analysis.

**Unique Value Counts in Job Titles and Skill Fields**

**A close-up of a white background

AI-generated content may be incorrect.**

*Note: It helps to assess the diversity and granularity of job titles and skill mentions before clustering or vectorization, which is important for understanding feature richness and potential noise in the dataset.*

**Text Vectorization and Feature Scaling for Clustering**

* We performed text vectorization and feature scaling, which are essential preprocessing steps before clustering
* Tfidf Vectorizer converts the cleaned job and skills text (combined\_text) into a numeric matrix based on word importance (TF-IDF), enabling text-based clustering.
* StandardScaler scales the TF-IDF features to have zero mean and unit variance, which is important because KMeans is sensitive to feature magnitudes.

**KMeans Clustering and Evaluation with NAICS 6-Digit Labels**



**Evaluate Clustering Using Multiple Reference Labels (NAICS, SOC, ONET)**

**A white rectangular object with a black background

AI-generated content may be incorrect.**

* *NAICS\_2022\_6\_NAME has the highest agreement with your clusters (though still very low), suggesting a slight alignment with industry-based classification.*
* *SOC and ONET labels have zero alignment — meaning the clusters derived from TF-IDF features of job titles + skills do not correspond to occupation-based taxonomies.*

**Visualize TF-IDF-Based Clusters with PCA and Plotly**

A yellow and pink dots

AI-generated content may be incorrect.

**Interpretation:**

* *The three clusters (color-coded) are distinct in the PCA space, suggesting that the clustering algorithm was able to differentiate based on text patterns.*
* *This model is capturing textual similarity (e.g., shared tools, terms, or phrasing in job descriptions), not necessarily formal job classifications.*
* *Cluster boundaries are data-driven, not taxonomy-aligned.*

**Top Terms Representing Each Cluster (TF-IDF Feature Importance)**

Cluster 0: pmi, apple, institute, ios, android, vmware, desktop, methodology, expectation, zachman, windows, infrastructure, capability, operating, subcontracting

Cluster 1: data, language, programming, sql, intelligence, python, tableau, analysis, dashboard, bi, power, statistics, visualization, analyst, analytics

Cluster 2: sap, enterprise, consultant, applications, oracle, functional, management, planning, cloud, architect, architecture, solution, design, erp, resource

**Nomenclature of our Clusters**

* *Cluster 0 = “IT Infrastructure & Support”*
* *Cluster 1 = “Data Analytics & BI”*
* *Cluster 2 = “Enterprise Applications & Consulting”*

**Visualizing Representative Job Titles Across Clusters**

A screenshot of a graph

AI-generated content may be incorrect.

**Cluster 0 (“IT Infrastructure & Support”)**

→ Jobs like enterprise support analyst, senior IT analyst, data integration analyst, IT enterprise architect.

→ These titles are support, IT system maintenance, integration, and architecture focused.

**Cluster 1 (“Data Analytics & BI”)**

→ Jobs like sr BI analyst, data analyst, data scientist, data research analyst.

→ Heavy analytics, business intelligence (BI), data science skills — matches perfectly.

**Cluster 2 (“Enterprise Applications & Consulting”)**

→ Jobs like SAP BTP consultant, ERP integrations analyst, applications consultant, product architect.

→ Clearly related to enterprise software (SAP, ERP) and consulting roles.

**Preprocessing Software Skills for Analysis**

**Clustered Software Skill Visualization**

A screenshot of a graph

AI-generated content may be incorrect.

**Interpretation**

**Cluster 0 (“IT Infrastructure & Support”)**

**Common skills:** Microsoft Excel, Microsoft SharePoint, Docusign, SAP Applications, TOGAF, automated cost tools.

→ These tools are typical for IT operations, documentation, system architecture support.

**Cluster 1 (“Data Analytics & BI”)**

**Common skills:** Python, SQL (PL/SQL), Looker, Tableau, Power BI, Google Analytics, Qlik Sense.

→ Clear focus on analytics, data visualization, and programming languages.

**Cluster 2 (“Enterprise Applications & Consulting”)**

**Common skills:** SAP Sales and Distribution, Google Cloud Platform (GCP), Microsoft OneNote, IBM Maximo.

→ These are enterprise-level software systems for consulting, ERP, and large infrastructure projects.

**Conclusion:**

*The software skills distribution perfectly matches the previously assigned cluster themes based on job titles and top terms.*

**Average salary per cluster**

A group of black text

AI-generated content may be incorrect.

**Salary Distribution Across Clusters**

**A graph with different colored squares

AI-generated content may be incorrect.**

**Interpretation for the Boxplot**

*Boxplot shows the salary distribution across three clusters derived from unsupervised KMeans clustering on job titles and skills:*

**Cluster 0:**

* Has a relatively high median salary (~$125K) and moderate spread, suggesting roles with consistent mid-to-high pay (e.g., enterprise or management roles).

**Cluster 1:**

* Has the lowest median salary (~$95K) with many outliers, indicating entry-to-mid-level roles with high variance (e.g., data or analyst roles).

**Cluster 2:**

* Shows the widest salary range with the highest outliers (up to $500K), implying this cluster contains senior or highly specialized roles (e.g., consultants or architects).
* Overall, the plot reflects meaningful salary differences between the clusters, supporting the relevance of clustering for job role segmentation.

**Conclusion and Key Takeaways**

*we applied KMeans clustering to job postings using text data from job titles and associated skills (software + specialized). Despite relatively low alignment with external classification labels like SOC, and ONET (as shown by ARI and NMI scores), our analysis still uncovered distinct, interpretable clusters with practical insights.*

**Clustering Pipeline Summary**

* Text Preprocessing: Combined job title, software, and specialized skills into a combined\_text field.
* Vectorization: Used TfidfVectorizer to convert text to numerical features.
* Scaling: Applied StandardScaler to normalize TF-IDF vectors.
* Clustering: Ran KMeans with k=3 clusters.

**Evaluation:**

* ARI (Adjusted Rand Index): Max ~0.009 with NAICS\_2022\_6\_NAME
* NMI (Normalized Mutual Info): Max ~0.033

*Interpretation: Clusters do not align well with predefined industry/occupation codes, which is expected in unsupervised learning.*

**Key Visual Insights**

**1. PCA Projection**

The PCA plot revealed clear separation between clusters, indicating the clustering algorithm did find structural patterns in job descriptions.

**2. Top Terms per Cluster**

* Cluster 0: Keywords like apple, ios, vmware, infrastructure suggest tech roles focused on devices, systems, and IT frameworks.
* Cluster 1: Terms like sql, tableau, python, analysis indicate data-related roles (analysts, BI, data scientists).
* Cluster 2: Words like sap, oracle, consultant, planning suggest enterprise solutions, consultants, or ERP specialists.

**3. Sample Job Titles by Cluster**

Confirms term-based interpretations:

* Cluster 0: IT infrastructure & support
* Cluster 1: Data analysts and BI roles
* Cluster 2: SAP/ERP consultants and architects

**4. Software Skills by Cluster**

* Cluster 0: Excel, SharePoint, PowerPoint – general office + support tools
* Cluster 1: Python, Tableau, Power BI – analytics & data tools
* Cluster 2: SAP, Oracle, GCP – enterprise software and cloud tools

**5. Salary Distribution by Cluster**

* Cluster 0: Mid-range salaries, low outliers – stable IT roles
* Cluster 1: Lower median salaries, wide spread – junior data roles
* Cluster 2: High median and extreme outliers – senior consultants & architects

**Final Takeaways**

Even with low overlap to government taxonomies (NAICS/SOC/ONET), clustering successfully revealed latent role patterns based on real-world skills and job titles.

Unsupervised clustering can meaningfully group job postings by functional role, skill stack, and salary range, offering powerful segmentation for:

* Career recommendation systems
* Skill gap analysis
* Compensation benchmarking
* Targeted recruitment strategies

**References**

* Umamaheswaran, S., Fernandes, S., Venkatesh, V. G., Avula, N., & Shi, Y. (2023). What Do Employers Look for in Business Analytics Roles? – A Skill Mining Analysis. Information Systems Frontiers. https://doi.org/10.1007/s10796-023-10437-y
* Musazade, N. (2022). Understanding the relevant skills for data analytics-related positions: An empirical study of job advertisements [Master’s thesis]. https://www.doria.fi/bitstream/handle/10024/184705/musazade\_nurlan.pdf
* Zarefard, M., & Marsden, N. (2024). The essential competencies of data scientists: A framework for hiring and training. In H. Mori & Y. Asahi (Eds.), HCII 2024, LNCS 14691 (pp. 397–418). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-60125-5\_27
* Noble Desktop. (2024). Business Analyst Job Outlook. https://www.nobledesktop.com/careers/business-analyst/job-outlook
* Tandfonline. (2024). Trends in Business Analytics Education. Journal of Business Analytics, 35(3), 325-376. https://doi.org/10.1080/26939169.2024.2393427
* Sacred Heart University. (2024). Everything You Should Know About the Business Analytics Career Path. https://shuconnect.sacredheart.edu/blog/2024/02/09/everythingyou-should-know-about-the-business-analytics-career-path/
* Kanagarla, K. P. B. (2024). Artificial intelligence and employment transformation: A multisector analysis of workforce disruption and adaptation. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5015970
* Bzdok, D., Thieme, A., Levkovskyy, O., Wren, P., Ray, T., & Reddy, S. (2024). Data science opportunities of large language models for neuroscience and biomedicine. Neuron, 112(1), 1-16. https://doi.org/10.1016/j.neuron.2024.01.016
* Yosifova, A. (2024, February 29). Data scientist job market 2024: Analysis, trends, opportunities. 365 Data Science. https://365datascience.com/career-advice/data-scientistjob-market/
* U.S. Bureau of Labor Statistics (BLS). (2024). Occupational Outlook for Data Scientists. https://www.bls.gov/ooh/math/data-scientists.htm