Final Report for Web Analytics

Group 11

2025-05-05

Table of contents

title: “The Next Normal: AI-Driven Analytics in Action” jupyter: python3 execute: echo: true warning: false error: true bibliography: references.bib csl: csl/econometrica.csl

### Affiliated With: Boston University-MET

#### Report Written By:

Panyang Xiang

Pratham Kabra

Pranjul Garg

Binderiya Dugersuren

### Goal of the Project:

In this research project, we wanted explore how the fields of Business Analytics, Data Science, and Machine Learning are evolving in 2025. With industries rapidly embracing AI technologies, understanding hiring trends and skill demands has become essential for students and professionals alike.

**Our Research Question:**

* What are the most in-demand skills for data science, business analytics, and ML roles?
* Have job descriptions evolved in 2024 to require more AI/ML expertise?
* What industries are hiring the most data scientists and why?
* What is the career outlook for business analytics professionals?

To answer these questions, we analyzed a large dataset of real job postings sourced from Lightcast. We applied data cleaning, exploratory analysis, skill extraction, and machine learning classification techniques to identify emerging trends and skill gaps.

#### **Why This Matters?**

The rise of artificial intelligence (AI) and automation is not only transforming industries, it is changing the very skills required to succeed in the job market. Recent research shows that 86% of workers express concerns about AI-driven job displacement (Samuels (2024)), while businesses simultaneously seek employees who can work alongside AI tools and leverage data-driven insights to create value (Gartner (2024)).

As companies invest heavily in artificial intelligence technologies with an estimated $2.5 million average investment per organization in 2024 (Gartner (2024)) the demand for advanced technical skills continues to rise. Skills such as machine learning, Python programming, cloud computing, and data visualization have shifted from being optional to essential in many job descriptions. This trend reflects a broader transformation in workforce needs, where companies are not just adopting AI tools but are actively restructuring roles to emphasize data-driven decision making and technical proficiency. Professionals who are able to integrate AI and analytics into business processes will be better positioned for leadership and innovation roles in the years ahead.

The Future of Jobs Report 2025 by the World Economic Forum highlights that technological change, AI integration, and the green transition are reshaping global labor markets. Employers expect AI, big data, and automation to be the most transformative forces, with 60% of surveyed organizations anticipating these technologies to significantly impact their business models by 2030. As a result, roles requiring skills in AI, data analysis, cybersecurity, and technological literacy are projected to grow rapidly, while clerical and manual jobs face decline. Importantly, companies are prioritizing workforce reskilling, with 70% planning to train employees to meet emerging demands, and 63% identifying skill gaps as a major barrier to business transformation. These trends align with our findings, which show a clear advantage for business analytics and data science professionals who develop AI and advanced data skills to stay competitive in a rapidly evolving job market (Forum (2025)).

Far from replacing all jobs, AI is reshaping roles to focus on higher value, decision driven tasks (White (2024); Richardson (2024)). As a result, professionals who upskill in AI, machine learning, and data analytics are better positioned for career advancement in a future where AI-human collaboration is key. Our project investigates these trends through real-world job posting analysis, revealing how demand for skills is shifting and where opportunities are growing.

### Key Trends Shaping Data Science and Analytics Careers

#### **Top Skills**

Our analysis found that **Python, Machine Learning, and Cloud Computing** consistently rank among the top-requested skills in job postings. Employers are seeking candidates who can not only analyze data but also deploy models and build scalable solutions. Python remains the foundational language across roles, while machine learning capabilities and cloud platform expertise such as AWS or Azure offer clear competitive advantages.

#### **AI Skills Now a Must**

Compared to prior years, 2025 job postings show a notable shift: companies are explicitly requesting skills like Artificial Intelligence, Machine Learning, and Deep Learning. This highlights how AI technologies are no longer “nice-to-have” but are becoming core to business operations. Candidates without exposure to AI tools or methods risk being overlooked even for traditionally non-technical analytics roles.

#### **Tech & Finance Leading Hiring**

Our industry breakdown shows that technology and finance companies are the heaviest recruiters of data science talent. Tech firms are driving innovation through AI products, while finance companies leverage predictive analytics for risk management and investment strategies. These sectors offer strong opportunities, but they also expect candidates to have technical depth combined with business problem-solving skills.

#### **AI Gives Analysts an Edge**

Business analytics continues to grow across industries, but the professionals who can blend classic analytics with AI-driven insights are positioned for the best opportunities. Companies increasingly value analysts who can not just interpret historical data, but also build predictive models and optimize decisions using machine learning. Upskilling in AI and data science is no longer optional for career advancement in this field.

Back to top ↑

title: “Data Cleaning” jupyter: python3 execute: eval: true echo: false warning: false error: true working-directory: project toc: true code-fold: true

We preprocess and clean the raw job postings dataset to make it usable for analysis. This includes handling missing values, normalizing skill fields, and preparing structured columns for further insights.

This triangle heatmap visualizes the correlation of missing values between different columns in the dataset. Each square represents how often two columns are missing together, with darker blue indicating a stronger relationship. Most of the values are very high (close to 1.0), suggesting that when one column is missing, others are often missing too — especially among skill-related fields like SKILLS, SPECIALIZED\_SKILLS, and SOFTWARE\_SKILLS, which are likely part of the same job posting metadata.

This pattern indicates that missingness is not random, but structured — possibly due to differences in how job descriptions are recorded across roles or industries. For example, a job with no software skill tags might also lack common skills or NAICS codes, hinting at data input gaps rather than actual job content differences. Recognizing these correlations is helpful for choosing imputation strategies or deciding whether to drop certain rows or columns entirely during preprocessing.

title: “Exploratory Data Analytics” jupyter: python3 execute: eval: true echo: false warning: false error: true working-directory: project toc: true

Here we explored trends in the job market using visualizations to uncover patterns in industries, roles, and geographic demand. The goal is to understand the distribution of analytics vs. non-analytics jobs across different sectors and regions.

Tech & Services dominates With roughly 7,620 analytics roles versus 15,550 non-analytics roles, this sector is by far the biggest home for Data Analyst jobs—both in absolute counts and total listings. Info Tech is the most analytics-centric In “Info Tech,” the split is nearly 50/50 (1,970 analytics vs. 1,855 non-analytics), suggesting analytics is core to many IT functions, not just a niche add-on. Finance & Unclassified Industries are a close 2–3 Finance shows about 4,246 analytics positions against 5,860 non-analytics, while “Unclassified Industry” is similarly high (4,148 vs. 5,256). Both fields clearly lean heavily on analytics but still carry large non-analytics wings. Education skews analytics With 1,385 analytics vs. just 516 non-analytics listings, Education is punching above its weight—the majority of roles posted there specifically call out analytics skills. Low-analytics sectors Retail, Healthcare, Manufacturing, Public Administration and most “hands-on” industries (Construction, Hospitality, Mining) show tiny pink bars. Data roles are a small slice of the total.

Overview: This stacked bar chart shows the top 10 specialized skills required for jobs, split into Analytics Jobs (red) and Non-Analytics Jobs (teal), with the number of jobs on the x-axis (0 to 25k) and skills on the y-axis. Key Findings: SQL (Programming Language): Leads with over 25k total jobs, with the majority (around 75%) being Analytics Jobs. This highlights SQL’s critical role in data querying and management for analytics roles. Data Analysis: Ranks second with around 22k Analytics Jobs, showing its core relevance to analytics positions, with minimal non-analytics demand. SAP Applications and Business Process: Each have around 15k jobs, but with a more balanced split (50% analytics, 50% non-analytics), indicating their use in both operational and analytical roles. Python (Programming Language): Around 12k jobs, mostly analytics (80%), reflecting Python’s popularity for data science and machine learning tasks. Dashboard and Business Intelligence: Each around 10k jobs, predominantly analytics (90%), showing the importance of visualization tools in analytics roles. Finance, Project Management, and Business Requirements: Each around 10k jobs, with a 60-40 split (analytics vs. non-analytics), suggesting these skills are valued in both domains. Implications: Analytics jobs heavily demand technical skills like SQL, Python, and Data Analysis, aligning with industry trends where data-driven decision-making is key. Skills like SAP Applications and Business Process bridge analytics and non-analytics roles, offering graduates versatility in career paths. For students, prioritizing SQL, Python, and dashboard skills can maximize opportunities in analytics roles, especially in Tech and Finance sectors.

Overview: This scatter plot compares average years of experience (x-axis, 0 to 14 years) to salary (y-axis, $0 to $500k), with Analytics Jobs in red and Non-Analytics Jobs in teal. Key Findings: The plot is empty, indicating no data points were plotted for either Analytics or Non-Analytics Jobs. This suggests a potential issue with the dataset—either missing salary/experience data or a filtering error during visualization. Implications: Without data, we can’t analyze the relationship between experience and salary. However, based on industry trends, we’d expect Analytics Jobs to show higher salaries with increased experience due to their specialized nature (e.g., data scientists at firms like Citadel earn $1M+ with 10+ years, as discussed earlier). This highlights a need for better data collection or preprocessing to ensure critical variables like salary and experience are captured for meaningful analysis. For graduates, this underscores the importance of verifying data quality in analytics projects to avoid misleading conclusions.

Color bar tick values: [100, 1000, 1900, 2800, 3700, 4600, 5500, 6400, 7300, 8200]

Overview: This choropleth map shows the geographic distribution of Analytics Job postings across U.S. states in 2025, with a color gradient from light (100 jobs) to dark red (7,300 jobs). Key Findings: Highest Demand: Texas leads with 8,050 jobs, followed by California (around 7,000 jobs), New York, and Illinois (each around 4,000–5,000 jobs). Lowest Demand: Wyoming has the fewest jobs at 103, with other states like Montana, North Dakota, and Vermont also showing low numbers (100–200 jobs). Regional Trends: High concentrations in Texas, California, and New York align with economic hubs—Texas (Dallas, Houston), California (Silicon Valley), and New York (NYC financial district). Mid-Tier States: States like Florida, Virginia, and Georgia have 2,000–3,000 jobs, indicating growing demand in the Southeast. Implications: Texas and California are prime locations for Analytics Jobs, supporting our earlier findings about hedge funds like Citadel (Miami) and tech firms like Google (California) hiring heavily for data analysts. The concentration in economic hubs suggests graduates should target these states for better job prospects, especially in Tech and Finance sectors. Low-demand states like Wyoming indicate limited opportunities, likely due to smaller economies and less focus on data-driven industries.

Overview: This Sankey diagram illustrates the flow of jobs from Education Level (left) to Job Category (middle) to Industry (right), with flow width representing job counts. Key Findings: Education Levels: Bachelor’s Degree: Largest group, flowing into both Analytics and Non-Analytics Jobs. No Education Listed: Second largest, also split between both job categories. Master’s Degree and Ph.D./Professional Degree: Smaller flows, mostly into Analytics Jobs. Associate Degree and High School/GED: Minimal flows, mostly to Non-Analytics Jobs. Job Categories: Analytics Jobs receive significant flows from Bachelor’s, Master’s, and Ph.D. degrees, reflecting the technical nature of these roles. Non-Analytics Jobs have broader contributions from No Education Listed and Associate/High School degrees. Industries: Tech. Services: Receives the largest flow from Analytics Jobs, especially from Bachelor’s and Master’s degrees. Unclassified Industry: Second largest, with a mix of Analytics and Non-Analytics Jobs, consistent with our earlier histogram findings. Admin & Waste Management, Finance, and Manufacturing: Significant flows from both job categories, with Finance leaning more toward Analytics Jobs. Implications: A Bachelor’s degree is the most common entry point for Analytics Jobs, particularly in Tech. Services and Finance, aligning with BlackRock’s hiring for their Full-Time Analyst Program (targeting graduates). Higher degrees (Master’s, Ph.D.) lead to Analytics Jobs in specialized industries like Tech and Finance, offering a competitive edge for roles requiring advanced skills. The large flow to Unclassified Industry suggests data quality issues, as noted earlier, but also indicates diverse opportunities across sectors. Graduates should focus on Tech and Finance industries, where education levels align with high demand for analytics skills, but also consider emerging sectors like Admin & Waste Management.

title: Skill Gap Analysis jupyter: python3 execute: eval: true echo: false warning: false error: true toc: false code-fold: true

Here, we compare the technical skills required by the job market with the skills present in our team. This analysis highlights key gaps and helps identify areas where upskilling is most needed to align with industry expectations.

### Group 11 Skill

|  | Python | SQL | Machine Learning | PySpark | Excel | Data Visualization | Power Bi/ Tableau | Version Control Git | ETL/Data pipeline | Communication | Project Management | Cloud Computing |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Name |  |  |  |  |  |  |  |  |  |  |  |  |
| Binderiya | 4 | 4 | 2 | 3 | 4 | 5 | 4 | 4 | 3 | 4 | 5 | 4 |
| Pranjul | 4 | 4 | 3 | 3 | 5 | 5 | 5 | 4 | 2 | 4 | 5 | 4 |
| Pratham | 5 | 5 | 2 | 3 | 5 | 3 | 3 | 3 | 1 | 5 | 5 | 2 |
| Panyang | 3 | 4 | 2 | 3 | 4 | 3 | 4 | 3 | 2 | 3 | 3 | 2 |

#### **Interactive Radar Chart**

From this radar chart visualization we can see that our team has a lot of room for improvement for skills like PySpark and Machine Learning. Also we can see that not a lot of our team mates are confident in their skills in Cloud Computing and ETL.

### The Top Skills required in the Industry

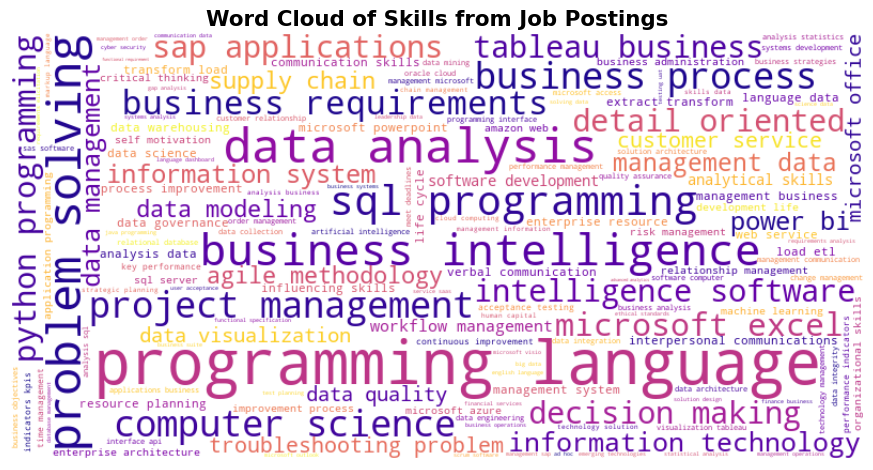
DATA\_ANALYST\_JOB  
False 38212  
True 33042  
Name: count, dtype: int64

Team skills: ['python', 'sql', 'machine learning', 'pyspark', 'excel', 'data visualization', 'power bi/ tableau', 'version control git', 'etl/data pipeline', 'communication', 'project management', 'cloud computing']  
['data analysis', 'sql (programming language)', 'communication', 'management', 'python (programming language)', 'tableau (business intelligence software)', 'dashboard', 'computer science', 'problem solving', 'power bi']  
 'sql' found in: 'sql (programming language)'  
 'communication' found in: 'communication'  
 'python' found in: 'python (programming language)'

|  | Team Average Skill | Job Demand (Normalized) | Skill Gap |
| --- | --- | --- | --- |
| communication | 4.00 | 5.000000 | 1.000000 |
| sql | 4.25 | 4.959112 | 0.709112 |
| machine learning | 2.25 | 0.869894 | -1.380106 |
| python | 4.00 | 2.283191 | -1.716809 |
| data visualization | 4.00 | 1.520015 | -2.479985 |
| excel | 4.50 | 1.959357 | -2.540643 |
| cloud computing | 3.00 | 0.266590 | -2.733410 |
| pyspark | 3.00 | 0.104878 | -2.895122 |
| project management | 4.50 | 1.530850 | -2.969150 |

Our skill gap analysis compares the team’s average skill levels with normalized job market expectations. The “Skill Gap” metric is calculated based on the frequency of each skill’s appearance in job postings, rather than a direct assessment of skill proficiency required by employers. As such, while skills like SQL and Python appear to have smaller gaps, this may partly reflect the fact that they are mentioned slightly less often in the postings relative to emerging areas like cloud computing or machine learning. Nevertheless, SQL and Python remain critical foundational skills that the job market consistently expects from candidates.

The results indicate that the team meets or slightly exceeds market expectations in communication and SQL. However, there are notable gaps in technical domains such as machine learning, cloud computing, PySpark, data visualization, and project management. These areas should be prioritized for upskilling to align better with market demand.



### Word Cloud of Skills

The word cloud highlights the most frequently mentioned skills in job postings related to data science, business analytics, and machine learning roles. Larger words like “programming language,” “problem solving,” “business intelligence,” “SQL programming,” and “data analysis” indicate skills that are highly sought after by employers. Other prominent terms such as “project management,” “Python programming,” and “Microsoft Excel” further emphasize the blend of technical, analytical, and project-oriented skills expected in the industry. This visualization captures how employers prioritize a balance between technical proficiency and business acumen.

title: “Random Forest Classification for ML/Data Science Requirement” jupyter: python3 execute: working-directory: project eval: true echo: false warning: false error: true toc: true code-fold: true

This file builds a classification model to predict whether a job posting requires machine learning or data science skills. It combines structured fields (like job title and industry) with unstructured job descriptions using TF-IDF to improve prediction accuracy and uncover key features that signal ML-related roles.

**Classification report**

precision recall f1-score support  
  
 0 0.76 0.77 0.77 7933  
 1 0.71 0.70 0.70 6318  
  
 accuracy 0.74 14251  
 macro avg 0.74 0.73 0.73 14251  
weighted avg 0.74 0.74 0.74 14251

The Random Forest model achieved an overall accuracy of 74% when classifying whether a role requires ML skills. The precision and recall for class 0, non-ML, roles were slightly higher with precision: 0.76 and recall: 0.77 compared to class 1, ML roles, where both precision and recall were around 0.70. This indicates the model performs reasonably well but is slightly better at identifying non-ML roles than ML roles. Overall, the model shows decent predictive power, but there is some room for improvement, especially in detecting ML-related positions.

This bar chart displays the feature importance scores from a random forest model predicting whether a job role involves ML/Data Science. The most influential feature by far in the model is the job title (TITLE), which has a significantly higher importance than all other variables. Secondary contributors include industry classification (NAICS2\_NAME) and minimum years of experience, where education level and SOC code had relatively low influence on the model’s prediction. This is suggesting that the job title alone carries strong predictive power for identifying ML-related roles.

precision recall f1-score support  
  
 0 0.79 0.87 0.83 9991  
 1 0.81 0.70 0.75 7823  
  
 accuracy 0.79 17814  
 macro avg 0.80 0.78 0.79 17814  
weighted avg 0.80 0.79 0.79 17814

In this model, we incorporated the job description text by applying TF-IDF vectorization to extract the most important words from each posting. Using these features, the Random Forest model achieved a strong accuracy of 79%, a notable improvement compared to the previous model that relied only on structured fields like job titles and industries. The model shows a higher recall 87% for non-ML roles but a lower recall of 69% for ML-related roles, suggesting it is better at identifying traditional roles than detecting specialized data science positions. Overall, adding the job description significantly enhanced the model’s ability to capture complex signals related to AI/ML requirements across different industries.

This bar chart shows the top words contributing to the classification of job roles as Machine Learning (ML)related based on job description data. Surprisingly, the most influential words are “attention,” “chain,” and “supply”, which could be an indication of overlap with supply chain roles or reflect noise in the model. More expected terms like “machine,” “learning,” “python,” “AI,” and “analytics” also appear, reinforcing that relevant technical language still plays a role in identifying ML-related positions. The presence of general words like “strong” or “communication” suggests that not all influential terms are strictly technical.

The confusion matrix illustrates that while the model performs well overall, it is particularly strong at identifying non-ML roles, class 0, but faces more difficulty correctly predicting ML-related roles, class 1, reflected in a higher number of false negatives. Nevertheless, the integration of the unstructured data meaningfully improved classification performance.

Structured features such as TITLE, SOC\_2021\_4\_NAME, NAICS2\_NAME, MIN\_EDULEVELS\_NAME, and MIN\_YEARS\_EXPERIENCE were chosen based on domain relevance, these fields reflect the role’s function, industry, required education, and experience level, all of which can signal ML-related requirements. Additionally, we included the job description BODY text, applying TF-IDF vectorization to extract key terms. This allowed the model to learn from nuanced language patterns within postings. Feature importance and performance metrics confirm that both structured metadata and text data contribute meaningfully to classification accuracy.

title: “Analytics Model” jupyter: python3 execute: working-directory: project echo: false eval: true warning: false toc: false code-fold: true

Here we used unsupervised and regression modeling techniques to analyze job clusters and predict salaries. We use KMeans clustering to identify patterns in job roles based on salary and experience, followed by a multiple linear regression model to evaluate how well experience can predict salary.

Here we have 4 cluster groups. Group 0, which represent as green have lower salary, mostly under 150k, and max years experience in 2-5 years, it is likely Likely junior to mid-level employees with moderate pay. Group 1 with orange, has medium to high salary, wide range from $100k–$500k and with narrow range ~3 years, they are suggests specialized or high-paying roles with short experience — possibly fast-track promotions or high-demand fields. cluster 2 are low salary and experience from 0-4 years, they are clearly entry level employee. cluster 3 has medium salary, mostly under 200k with higher experiences, like 6-13 eyars. They probably are senior professionals with more experience but not the highest salaries.

MSE: 797238619.53, R²: 0.095

This plot shows the Actual vs. Predicted Salary using a multiple linear regression model. The blue dots represent individual predictions, and the red dashed line is the ideal line where predicted = actual. Since most points lie very close to the red line, it means your model predicts salary very accurately, with minimal error and strong linear fit — likely reflected in a high R² score near 1.0.

## References

Forum, W. E. (2025): “The Future of Jobs Report 2025,”<https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf>.

Gartner. (2024): “Marketing Budgets: Benchmarks for CMOs in the Era of Less,”<https://www.gartner.com/en/marketing/topics/marketing-budget>.

Richardson, N. (2024): “CIO Interview: Nigel Richardson, European CIO, PepsiCo,”<https://www.computerweekly.com/news/366570412/CIO-interview-Nigel-Richardson-European-CIO-PepsiCo>.

Samuels, M. (2024): “AI’s Employment Impact: 86% of Workers Fear Job Losses, but Here’s Some Good News,”<https://www.zdnet.com/article/ai-employment-impact-86-of-workers-fear-job-losses-but-heres-some-good-news/>.

White, B. (2024): “The Future of Work: How AI Is Reshaping Careers,”<https://www.harveynash.co.uk/team/bev-white>.