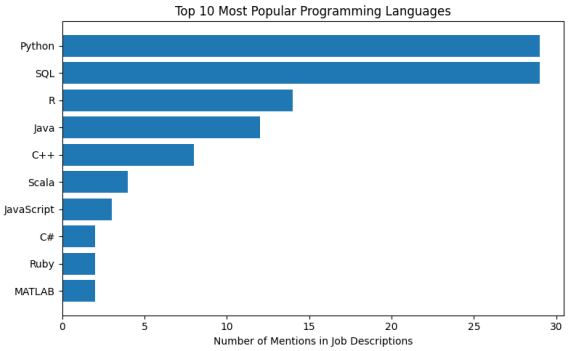
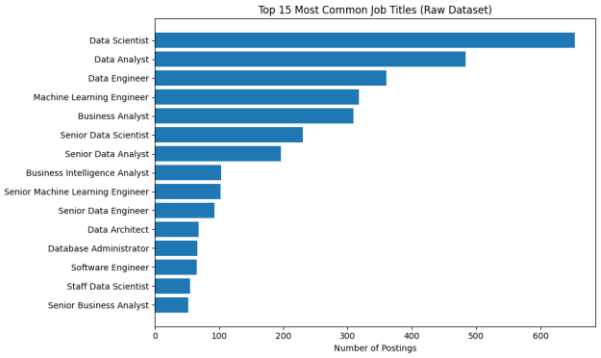
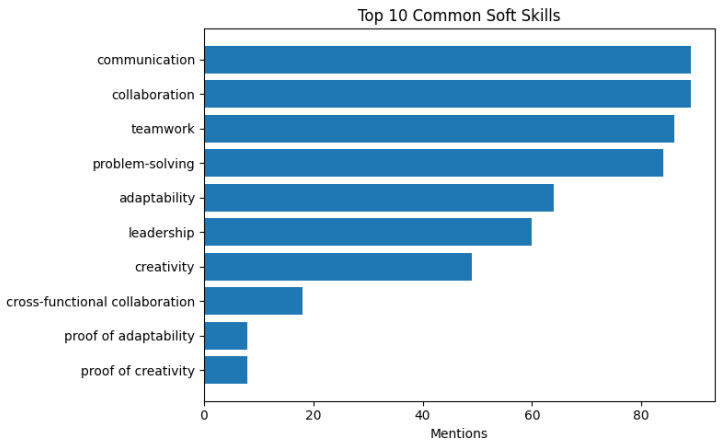
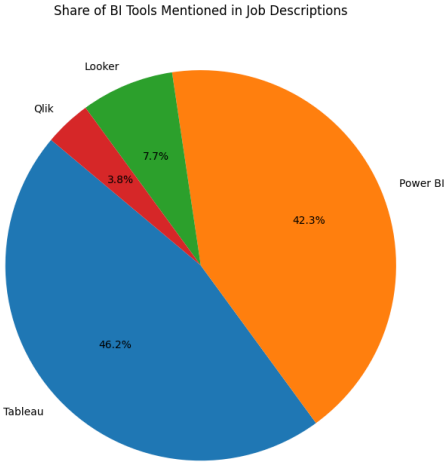
Q1 – Extracted Skills & Visualizations

Skills were extracted from job descriptions using the DeepSeek API, prompting the model to return both soft and hard skills in a comma-separated format. Each posting was truncated to 1000 characters and processed individually. Extracted skills were stored in a new column and used as the basis for downstream analysis. To explore the extracted skills, several visualizations were generated. Python, SQL, and R emerged as the most frequently mentioned programming languages, underscoring their centrality in data-focused roles. The most common job titles included Data Scientist and Data Analyst, reflecting the dataset's concentration on analytics-driven positions. Among soft skills, communication, teamwork, and problem-solving were most prominent, highlighting the consistent emphasis on interpersonal abilities in technical roles. In terms of business intelligence tools, Tableau and Power BI accounted for the vast majority of mentions, indicating their widespread adoption in the industry. These observations guided the downstream clustering and curriculum design by emphasizing the most relevant and commonly sought-after skills.









Q2 – Hierarchical Clustering Implementation

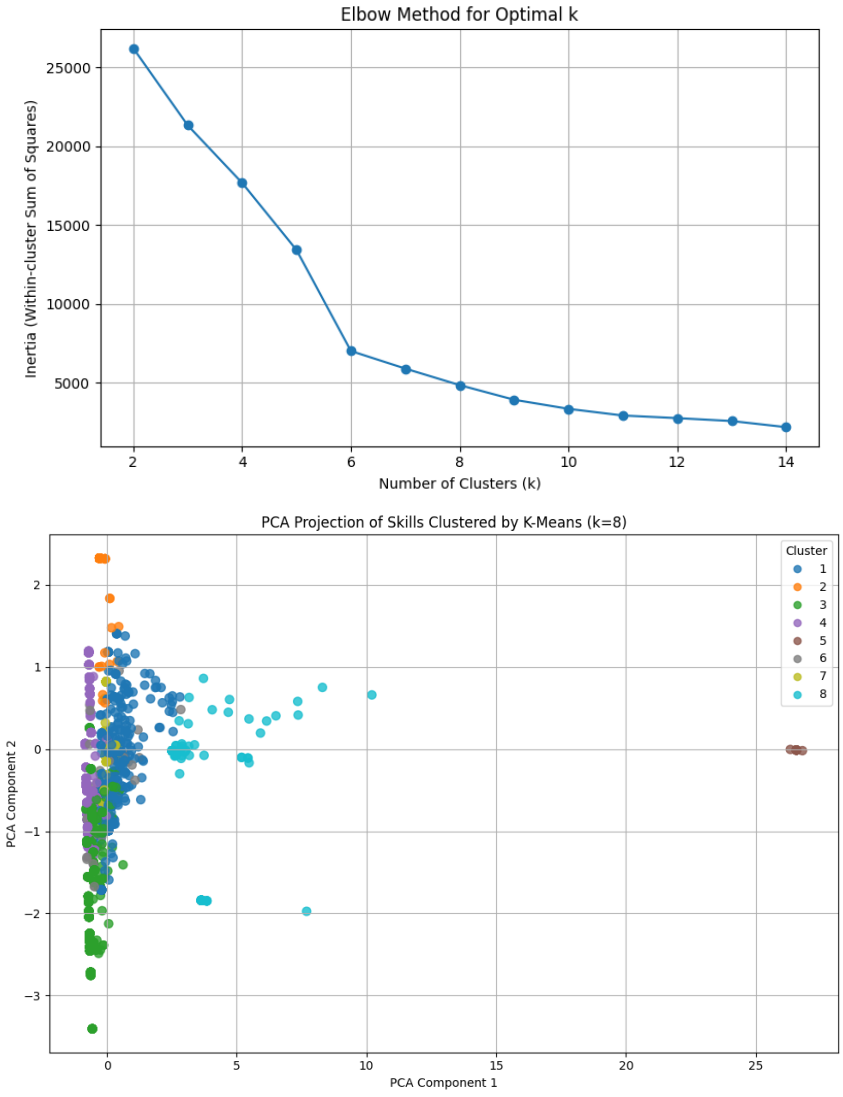
We constructed a skill co-occurrence matrix by first filtering out infrequent skills (less than 10 appearances) to reduce sparsity. A Jaccard distance matrix was then computed to represent similarity between skills based on their shared appearance across job descriptions. This matrix served as input to agglomerative hierarchical clustering. To visualize the clustering structure, we generated a dendrogram and initially applied a flat threshold to create clusters. However, some large clusters were overly broad (e.g., containing more than 60 skills), while others had only one or two. As a result, we manually split the largest cluster into four thematic courses: Data Engineering, Data Visualization, Machine Learning, and Business Strategy. These groups were curated to ensure skill coherence and meet the assignment's requirement of 8–12 courses, each covering at least 3 skills. Smaller, semantically consistent clusters from the dendrogram were retained as additional courses (e.g., "Data Management", "Scalability"), while a separate soft-skills course was assembled from recurring interpersonal and communication-related terms. In total, eight courses were defined, combining manual refinement and algorithmic grouping, and avoiding duplication of skills across them; Silhouette analysis was considered but ultimately omitted due to the unsuitability of the Jaccard distance matrix and binary encoding of skills for this metric. Instead, interpretability was prioritized: dendrogram structure and manual validation ensured clusters grouped related skills while maintaining instructional coherence.

	Course_ID	Skills
0	Course 1: Data Engineering & Big Data Tools	[Hadoop, Spark, SQL, Python, R, AWS, database ...
1	Course 2: Data Analysis & Visualization	[data analysis, data visualization, Tableau, P...
2	Course 3: Machine Learning & AI	[AI, machine learning, deep learning, predicti...
3	Course 4: Business Strategy & Project Management	[project management, business intelligence, bu...
4	Course 5: Introduction to Data Science	[data science, analytics, analytical skills]
5	Course 6: Applied Data Modeling	[data modeling, data interpretation, trend ana...
6	Course 7: Data Engineering Basics	[data pipelines, data cleaning, automation]
7	Course 8: Professional Skills for Data Roles	[critical thinking, communication, conflict re...

Q3 – KMeans Clustering

We applied K-means clustering on a set of engineered features derived from each extracted skill. These features included frequency, co-occurrence metrics, document coverage, presence in job titles, average salary, and indicators for remote and senior roles. After scaling the feature matrix, we ran K-means with varying k values. The elbow method indicated diminishing returns beyond k=6; however, we selected k=8 to better meet the requirement of producing 8–12 courses. The clustering output grouped skills into semantically coherent clusters, such as analytics, machine learning, programming, and soft skills. The resulting clusters were treated as course modules, with each containing 3 or more related skills. To visualize the clustering results, we used PCA to reduce dimensionality. The PCA scatterplot showed reasonable separation between clusters, with some overlap in softer or general-purpose skills. The silhouette score for k=8 was 0.6390,

indicating strong intra-cluster similarity and inter-cluster separation. This metric supported our chosen k and confirmed that the engineered features effectively differentiated between distinct skill categories. Combined with human interpretation of cluster coherence, this validated the quality of the K-means clustering.



Q4 – Ensemble Method

We implemented an ensemble method by intersecting the results from hierarchical and K-means clustering. Each skill was assigned a composite label based on its (Hierarchical Cluster ID, K-Means Cluster ID) pair. This allowed us to capture more nuanced relationships between skills, reducing the reliance on a single clustering structure. Skills with identical ensemble labels were grouped together. This approach helped surface granular yet coherent clusters, for example, “predictive modeling” was isolated from broader machine learning terms, and skills like “data

interpretation" and "data modeling" were grouped independently. In total, the ensemble method yielded 16 distinct skill clusters, improving curriculum modularity while preserving consistency.

Q5 – Please find the corresponding section in Jupyter notebook

Q6 – Final Course Curriculum & Discussion

I selected the final curriculum from the hierarchical clustering results (Section 2), refined through manual adjustments and thematic interpretation. This structure balances technical depth with managerial and professional skill development and was ultimately validated through an ensemble of clustering insights and ChatGPT-powered narrative analysis.

The final curriculum consists of 8 well-defined courses, categorized into three key focus areas: technical (Courses 1, 2, 3, 6, 7), managerial (Course 4), and professional (Courses 5, 8). Technical courses emphasize hands-on skills including programming, machine learning, data visualization, and data engineering. Course 4 bridges technical and business contexts with themes in project management and strategic planning. Meanwhile, soft skills are grouped into a dedicated module covering communication, problem-solving, and collaboration, skills that support career progression but are rarely taught directly.

This structure follows a logical learning progression: from foundational concepts (Course 5), through technical specialization (Courses 1–3, 6–7), to leadership and workplace readiness (Courses 4 and 8). Overlaps between clusters (e.g., debugging in both engineering and ML) reinforce the real-world interconnectedness of skills.