

Designing a Master's Degree Course Curriculum using Clustering Algorithms:

1. Data Collection and Scraping

To design a 'Master of Business and Management in Data Science and Artificial Intelligence', curriculum was to understand the different positions that a graduate from such a degree would work in. Subsequently, job postings were scraped from the job portal **Indeed** to be used for analysis.

Identifying roles like **Data Scientist, Data Analyst, Machine Learning Engineer, Business Intelligence Analyst, Director of Analytics, Manager of Analytics, Chief Analytics Officer, and Machine Learning Consultant**, I chose California—Silicon Valley's hub—for its cutting-edge job market. The outputs from ChatGPT, offering insights on these roles, are detailed in the **appendix**, and were combined with what skills I deemed necessary based on my knowledge:

Technical Skills: Machine Learning, Data Engineering, Natural Language Processing, Mathematical Optimization, Big Data, Data Visualization, Statistical Analysis, Deep Learning

Software Skills: R, SQL, Python, TensorFlow, Pandas, Tableau, Power BI, Numpy, Hadoop, Spark, PyTorch, Excel

Soft Skills: Leadership, Problem Solving, Creativity, Teamwork, Communication, Project Management, Consulting, Presentation Skills

Job positions and skills were selected to match graduates' pursuit of technical and managerial roles in Data Science ensuring balanced analysis. 300 data points per role were extracted. This strategy supports the development of a nuanced and comprehensive curriculum beyond conventional frameworks

2. Data Cleaning, Skill Location, Encoding and Visualization of Key Information

After collecting data, I combined it, removed duplicates and irrelevant columns for reliable analysis, and tagged each job posting with its position searched for. Job descriptions were cleaned—lowercased, extra white spaces removed and **tokenized**—to efficiently locate skills with **n-grams**, facilitating the search for alternative phrases. **NLTK's everygram** library was utilized to detect skills in job descriptions accurately, which ensured **multi-word skills** like 'mathematical optimization' are accurately extracted. Also, by incorporating alternative expressions (e.g., 'statistical analysis' as 'statistics'), this method ensured comprehensive identification of skills, accommodating **variations in wording**.

In order to numerically encode whether a job description had a particular skill, A column was created for each skill to be searched for: If identified in a particular job description, a value of 1 was inputted, otherwise a value of 0 was assigned. Subsequently, 4 visuals were plotted (2 in **Appendix**) to capture key information

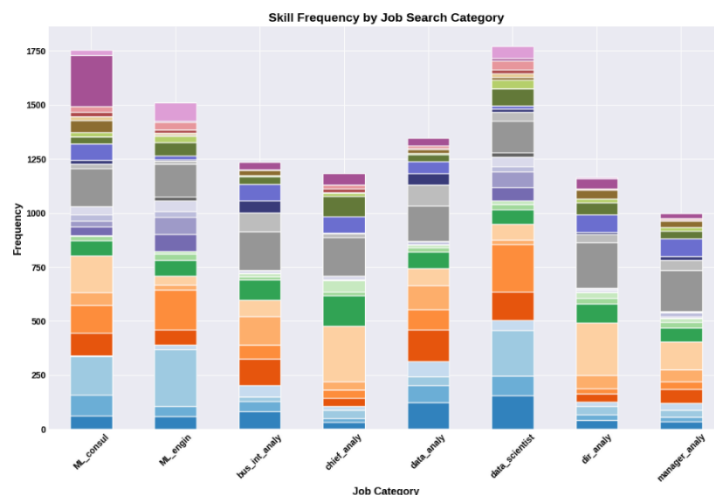


Figure 1: Skills most prevalent in respective job position that was searched for.

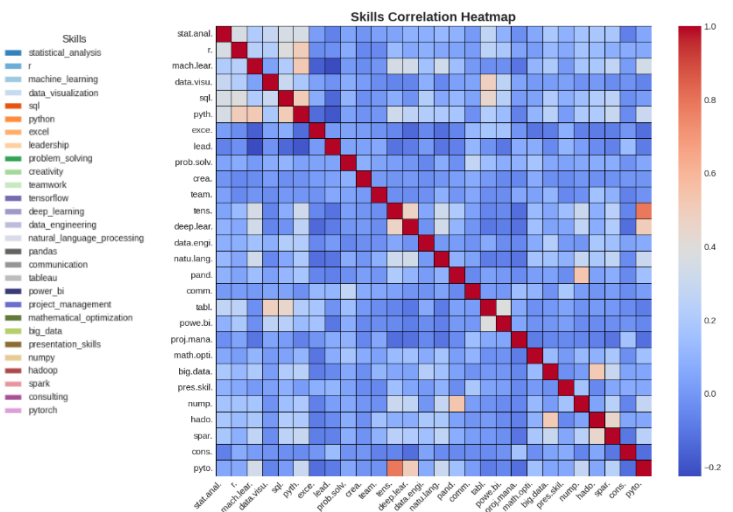


Figure 2: Correlation heatmap between skills per their mutual presence or absence.

3. Hierarchical Clustering

To perform hierarchical clustering, a distance method with our 1, 0 encoding was used, comparing skills via absolute difference. This technique identifies skill co-occurrence in the same job postings (If they both appear, together, distance is 0. If one appears but another does not a value of 1 is added to the distance). Distances are summed for each pair over all job descriptions. Smaller distances between a pair of skills indicate higher affinity and vice versa. The 'average' linkage was chosen because it gave the most empirically correct clusters and were visualized through a dendrogram in **Figure 5**.

- A Max D of 375 was chosen visually and balanced clustering by avoiding distant skill grouping while preventing excessive granularity. This identified 21 ideal clusters, but curriculum size of 12 courses and minimum of 3 skills per prompted use of manual adjustment.

- Manual cluster adjustments utilized dendrogram distances to group related skills, such as merging teamwork, creativity, presentation; consulting, project management, Excel; and R, statistical analysis, SQL, based on their proximity in the dendrogram.

- Intuitive adjustments were made for specific skills, clustering python with NumPy and Pandas, and Hadoop with Big Data, Data Engineering, and Spark, to align with logical associations.

- Only 1 course comprised 2 skills - machine learning & optimization, for their interconnection, Intensive and foundational nature

- While manual adjustments were made, the dendrogram's distance relationships were largely preserved to maintain objectivity and ensure data-driven curriculum development.

As per above, the following course curriculum was designed:

1. Efficient coding in Python: (Python, Numpy, Pandas)
2. Mathematics & Implementation - Machine Learning Algorithms: (Machine Learning, Mathematical Optimization)
3. Leveraging Statistical Methods for powerful Insights: (Statistical Analysis, SQL, R)
4. Visual Analytics for Impactful and Persuasive Data Presentations: (Tableau, Power BI, Data Visualization)
5. Skillful handling of Big Data: (Big Data, Spark, Data Engineering, Hadoop)
6. Advanced Machine Learning Algorithms: (PyTorch, Natural Language Processing, TensorFlow, Deep Learning)
7. Delivering Effective Presentations and Results: (Teamwork, Presentation Skills, Creativity)
8. Effective Consulting using Project Management Techniques: (Consulting, Project Management, Excel)
9. Effective Spearheading of Analytical Projects – Data-Thons!: (Leadership, Problem Solving, Communication)

4. Feature Engineering - K-Means Clustering:

For next algorithm of clustering - K-Means, I engineered features that describe skills. All Features extracted from job descriptions except (2), (3) which were based on domain knowledge. (4), (5), (6), (9), (10) used everygrams with alternative wording. (7) used Regex with alternate phrases to locate most instances. Proportion used for (6), (9), (10), to prevent bias due to skewed skill counts. (2), (5), (7) is nominal, (3) is ordinal. (1) sums 1,0 encoding done in sect. 2

1. Skill Frequency: Counts job postings per skill, indicating demand and niche status.
2. Skill Category: Categorizes skills as technical, software, or soft, highlighting domain relevance.
3. Skill Transferability: Ranks from low, medium and high, showing versatility across industries.
4. Average Skill Mentions: Averages no. of mentions in description, if skill mentioned, reflects skill importance within role.
5. Common Job Type: Most common work setup, by highest count of remote, hybrid, in-person mention for skills.
6. Education Level Proportion: Reveals education distribution, indicating required education levels for a skill.
7. Average Experience Needed: Calculates average years of experience required for each skill, gauging entry barriers.
8. Frequent Job Position: Finds most frequent job title (Data Scientist, ML Engin.), clarifying roles of importance for skill
9. Certificate Proportion: Prop. of jobs that require skill that also mention certification, indicate skill validation requirement
10. Job Role Proportion: Distribution of roles (entry, moderate, senior), showing which level of careers different skills required for.

Features (Ordinal and Continuous) were standardized, and K-Means was performed

- Elbow method revealed no sharp WCSS decrease but showed an almost linear decline after **k=10**, establishing **k=10** as the optimal cluster number.

- Adjustments were made to clusters with fewer than 3 skills or illogical groupings, using PCA plots and domain knowledge, with sparing use of PCA because of dimensional reduction inaccuracies.

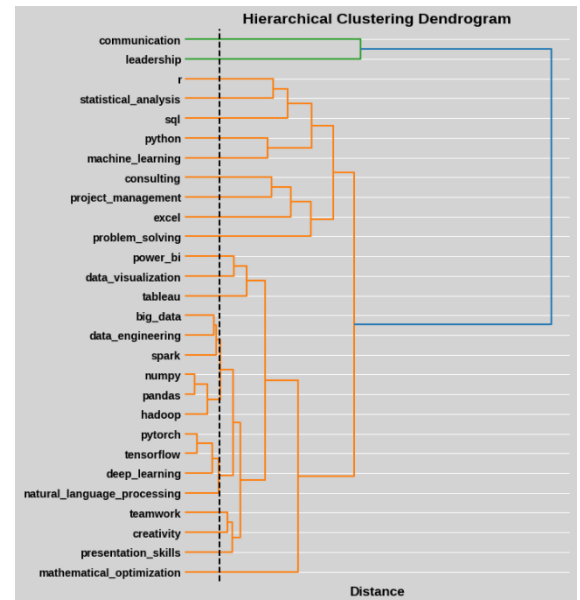


Figure 5: Hierarchical Clustering Dendrogram with Max D Line

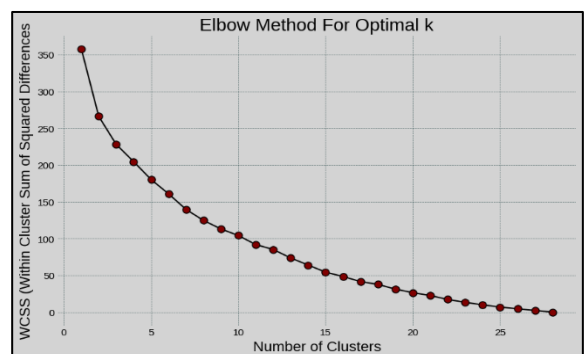


Figure 6: Elbow Plot for K-Means

- PyTorch (Cluster 5) was moved to TensorFlow, Natural Language Processing, Deep Learning (Cluster 8). Due to their proximities in the PCA plot and the logic of it being in the same cluster. Similarly, Spark (Cluster 5), was moved to Big Data, Hadoop, Pandas (Cluster 10), for the same reasons.

- Most clusters were manually adjusted and needed to be broken with just domain knowledge such as statistical analysis and machine learning (Cluster 9), and was moved respectively to R, Mathematical Optimization, Python (Cluster 2) and NumPy (Cluster 5), respectively. Similarly, python was moved from Cluster 2 to Cluster 5.

- Teamwork and Consulting form a unique two-skill cluster, chosen for their synergy in fostering collaboration and real-world project skills, which takes significant amount of time to properly learn.

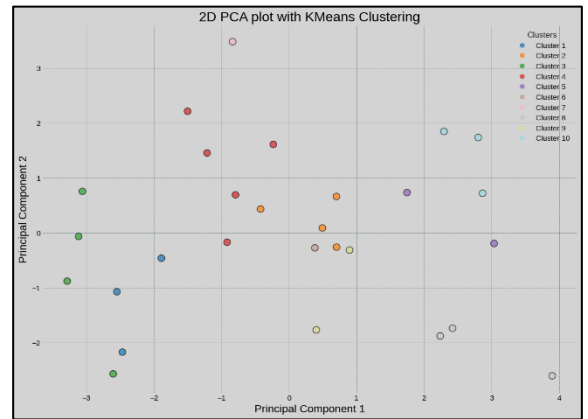


Figure 7: Scatter Plot using PCA for Clustering Visual

The following course curriculum was designed based on K-Means Clustering:

1. Introduction to Python and Machine Learning Algorithms: (Python, NumPy, Pandas, Machine Learning)
2. Navigating Databases & using data for visualization: (Power Bi, Tableau, Data Visualization, Creativity, SQL)
3. Data Analysis from different perspectives: (R, Statistical Analysis, Mathematical Optimization)
4. Skillful handling of Big Data: (Big Data, Spark, Data Engineering, Hadoop)
5. Advanced Machine Learning Algorithms: (PyTorch, TensorFlow, Deep Learning, Natural Language Processing)
6. Managing Projects under Strict Deadlines: (Project Management, Excel, Presentation Skills)
7. Collaborative Consulting - Real Life Data Science Projects: (Consulting, Teamwork)
8. Effective spearheading of Analytical Projects: (Leadership, Problem Solving, Communication)

5. Discussion and Final Course Curriculum

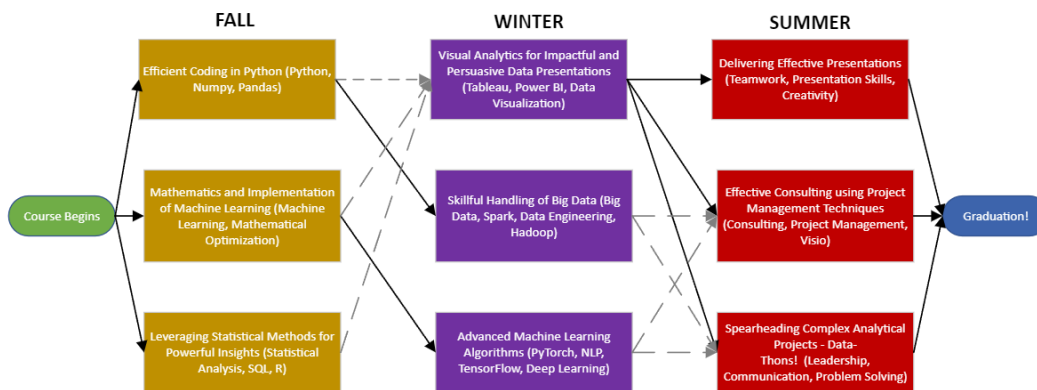


Figure 8: Flowchart of our Designed Course Curriculum

- The course curriculum, derived from hierarchical clustering, is outlined in a flowchart I created where solid lines represent essential prerequisites, and dashed lines state beneficial but non-essential skill connections, illustrating the structured sequence and prerequisite relationships among courses. Each semester contains courses that are independent yet complementary, offering insight into the curriculum's organization.
- Hierarchical clustering, guided by a dendrogram, provided a structured approach to grouping courses, allowing us to efficiently combine skills based on data and domain knowledge. This method produced more logical clusters, like combining Spark, Data Engineering, and Big Data, unlike K-Means, which initially suggested less cohesive groupings.
- K-Means clustering, grouped Pandas with Big Data and Hadoop, and statistical analysis with machine learning—combinations less intuitive for course design. Hierarchical clustering's clear visualization led to more effective grouping formations compared to PCA's visualizations for interpretation due to dimensionality reduction.
- Curriculum emphasizes following technical knowledge with soft skills to equip students for management in analytical fields, stressing collaboration and creativity for corporate success. A summer dedicated to real-life projects further bolsters job readiness and provides networking opportunities, blending technical skills with practical experience.

7. Appendix

7.1. ChatGPT Prompts outputs for Skill Extraction

Prompt 1: "What are the top skills required for a Data Scientist and Data Analyst job. Give me 12 most relevant skills in total for both that jobs look for in applicants, each skill being in 1 word, give me specific software and softskills"

Output 1:

Data Scientist:

1. Python
2. SQL
3. Machine Learning
4. Data Visualization
5. Statistical Analysis
6. Hadoop

Data Analyst:

1. Excel
2. Tableau
3. PowerBI
4. R
5. Critical Thinking
6. Communication

Prompt 2: "What are the top skills required for a Machine Learning Engineer and Business Intelligence Analyst job. Give me 12 most relevant skills in total for both that jobs look for in applicants, each skill being in 1 word, give me specific software and softskills"

Output 2:

Machine Learning Engineer:

1. Python
2. Algorithms
3. TensorFlow
4. Deep learning
5. Data visualization
6. Statistical modeling

Business Intelligence Analyst:

1. SQL
2. Tableau
3. PowerBI
4. Data analysis
5. Problem-solving
6. Critical thinking

Prompt 3: "What are the top skills required for a Manager of Analytics and Chief Analytics Officer job. Give me 12 most relevant skills jobs in total for both that jobs look for in applicants, each skill being in 1 word, give me specific software and soft skills"

Output 3:

1. Data analysis
2. Statistical modeling
3. Machine learning
4. Data visualization

5. SQL
6. Python
7. R
8. Tableau
9. Power BI
10. Leadership
11. Problem-solving
12. Communication

Prompt 4: "What are the top skills required for a Director of Analytics and Machine Learning Consultant job. Give me 12 most relevant skills in total for both that jobs look for in applicants, each skill being in 1 word, give me specific software and softskills

Output 4:

1. Python
2. R
3. SQL
4. Machine learning algorithms
5. Data visualization
6. Statistical analysis
7. Deep learning frameworks (e.g., TensorFlow, PyTorch)
8. Natural language processing (NLP)
9. Problem-solving
10. Effective communication
11. Project management
12. Team leadership

7.2 - 2 Other visualizations of our data to highlight key information (Question 2)

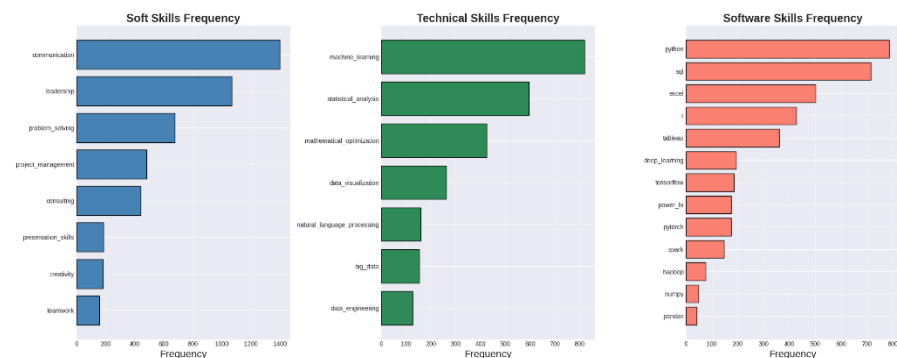


Figure 3: Distribution of skills within their respective category (soft, technical or software).

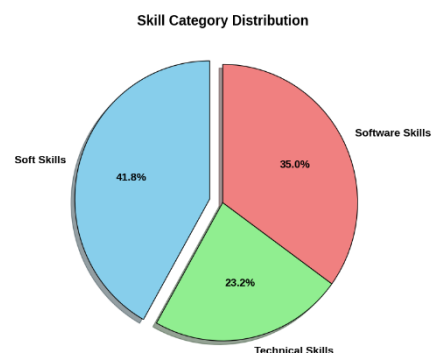


Figure 4: Overall proportion of the frequency of each type of skill found