DATA-DRIVEN CURRICULUM DESIGN: LEVERAGING CLUSTERING ALGORITHMS FOR A BUSINESS AND AI MASTERS PROGRAM

1. INTRODUCTION:

To design a course curriculum for a unique and new program, we must identify the skills we need to develop in a student through the program. As a program that aims to couple both business and technical aspects of a field, the course must be dynamic and help the student achieve their technical and business-development goals. This isn't a research program, but a professional one, the students taking up the program are looking to improve their professional careers; thus, this must be aimed at helping the students to get better jobs and career prospects.

2. DATA COLLECTION AND CLEANING

To begin, I collected data by scraping job postings in one of the most used applications for searching for jobs, "Indeed", to find listings of jobs in potential career paths for the program-takers. From those postings, we collect information such as Company name, salary, job description, location, etc. I collected data from postings of the title "Data Scientist", "Data Analyst", "Manager of Analytics", and "Director of Analytics". These are some of the potential positions we want our program-takers to go into, thus understanding the requirements of these positions, we can develop a curriculum that targets these areas. The description column in the dataset was first cleaned, then tokenized, and lemmatized.

3. EXPLORATORY DATA ANALYSIS AND FEATURE ENGINEERING

It is important to know that the data set is quite imbalanced, thus the skills obtained are more biased toward data analytics and engineering.

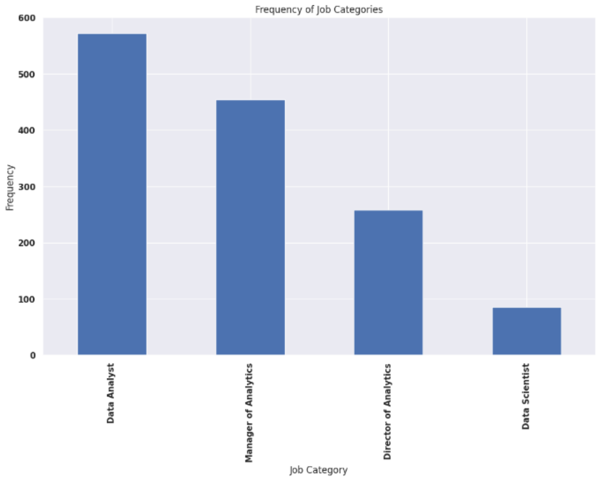


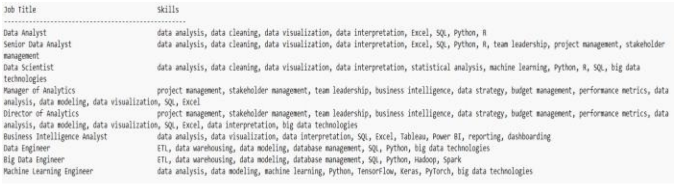
Fig. 1: Frequency distribution of each job category

From the plots in the previous exploration, I could identify stop words, that is, words to exclude. For example, let us have a look at the visuals below. I couldn't utilize Open Ai's ChatGPT 3.5 for identifying skills from the job descriptions as my free credits were over and my API limits were complete. So, I had to use ChatGPT to help me identify some of the skills for the related jobs, which was the closest I could do.

support'
team'
system'
product'
process'
project'
including'
program'
sexual' orientation'
andor'
need'
role'
department'
lead'
equal' opportunity'
client'
data' analysis'
solution'
organization'
work'
analysis'
service'
data'
customer'
technology'
national' origin'
year'
experience'
requirement'
company'
data' science'
gender' identity'
salary' range'
develop'
research'

Fig. 2: Word Cloud of stop words.

I asked ChatGPT for a list of skills for various positions. I used these skills in my model to extract the skills.



```
Job Title                         Skills
--------------------------------------------------
Data Analyst                      data analysis, data cleaning, data visualization, data interpretation, Excel, SQL, Python, R
Senior Data Analyst               data analysis, data cleaning, data visualization, data interpretation, Excel, SQL, Python, R, team leadership, project management, stakeholder
management
Data Scientist                    data analysis, data cleaning, data visualization, data interpretation, statistical analysis, machine learning, Python, R, SQL, big data
technologies
Manager of Analytics              project management, stakeholder management, team leadership, business intelligence, data strategy, budget management, performance metrics, data
analysis, data modeling, data visualization, SQL, Excel
Director of Analytics             project management, stakeholder management, team leadership, business intelligence, data strategy, budget management, performance metrics, data
analysis, data modeling, data visualization, SQL, Excel, data interpretation, big data technologies
Business Intelligence Analyst     data analysis, data visualization, data interpretation, SQL, Excel, Tableau, Power BI, reporting, dashboarding
Data Engineer                     ETL, data warehousing, data modeling, database management, SQL, Python, big data technologies
Big Data Engineer                 ETL, data warehousing, data modeling, database management, SQL, Python, Hadoop, Spark
Machine Learning Engineer         data analysis, data modeling, machine learning, Python, TensorFlow, Keras, PyTorch, big data technologies
```

Table 1. Skills for Job Title from Open Ai's ChatGPT

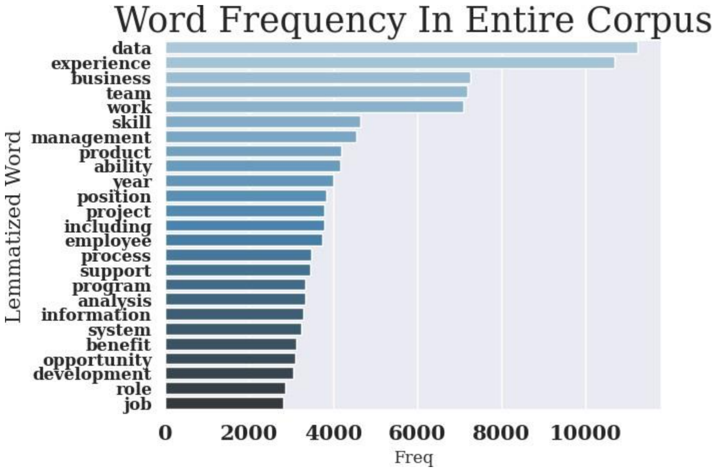


Fig. 3: Frequency Distribution of stop words.

Further, I utilized the Naive Bayes model to extract the skills from the job description with a bag of words and also new stop words on top of the ones from preprocessing. This model provided some beautiful visuals that paint the picture.

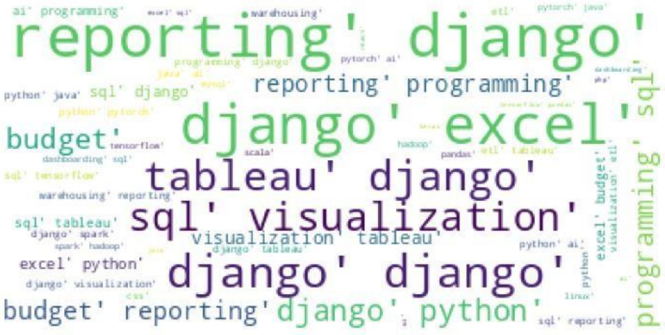


Fig 5. Word cloud of skills

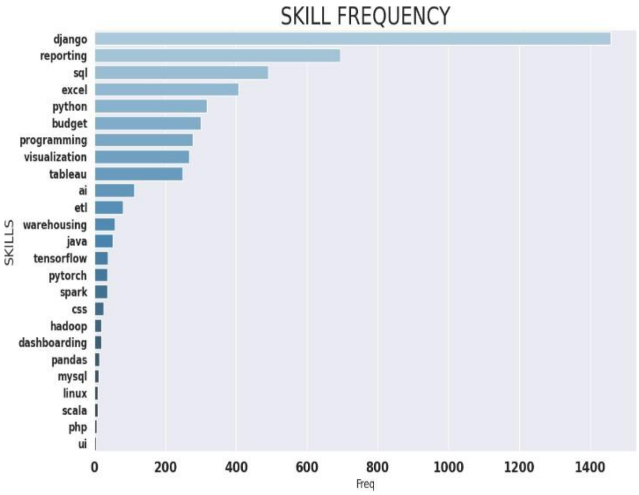


Fig 4. Skill frequency. From Naïve Bayes Model

To understand the skills required for a job even better, I have combined job titles and classified them into 4 main types: "Data Analyst", "Data Scientist", "Manager of Analytics", and "Director of Analytics". By doing this, a much clearer picture was painted.

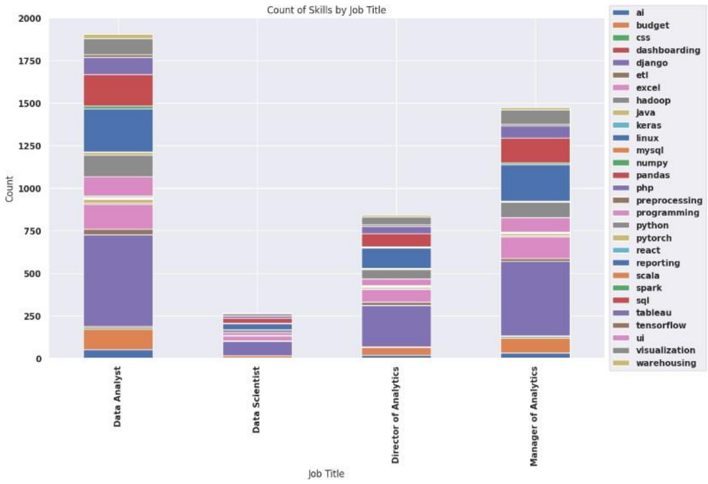


Fig 6. Count of Each Skill vs Job Title

4. COURSE CURRICULUM BASED ON INITIAL CLUSTERING RESULTS

I generated a distance matrix in the form of a dendrogram to develop a course curriculum of about 8-12 courses covering at least 3 topics/skills in each. Based on the results, I decided this to be the curriculum with skills covered in the brackets:

1. Database Systems and Data Warehousing (ETL, SQL, warehousing)
2. Data Engineering (Programming, Linux, Excel)
3. Introduction to Data Science (Python, NumPy, Pandas, SQL)
4. Data Visualization and Business Intelligence (Dashboarding, Tableau, Reporting, Visualization, Excel)
5. Big Data Technologies (Hadoop, Spark, Warehousing, SQL, ETL)
6. AI and Deep Learning (TensorFlow, PyTorch, Keras, AI)
7. AI in Finance and Management (AI, Budget, Reporting)
8. Capstone Project (All skills covered)

Further, I will engineer 10 features around the skills to better describe them and perform k-means clustering.

Based on the results of the 10 features, I performed k-means clustering with n=5, and developed a new course curriculum.

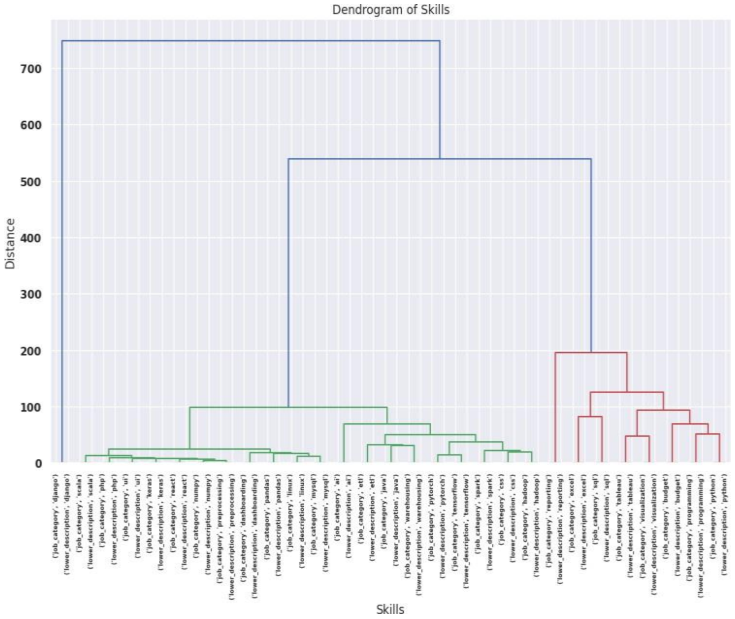


Fig. 7: Distance between each skill identified - initial clustering.

5. FEATURE ENGINEERING AND K-MEANS CLUSTERING

I engineered 10 features, they are:

1. Skill frequency: the number of times the skill appears in the dataset.
2. The average salary for skill: the average salary for jobs that require the skill.
3. Binary indication of soft or hard skill: a binary value indicating whether the skill is considered a hard skill or soft skill (based on external sources)
4. Skill level: a measure of the skill level required (e.g., beginner, intermediate, advanced)
5. Level of Education: the level of education typically required the skill.
6. Job titles: a count of the number of job titles that require the skill.
7. Distance Matrix: Using cosine similarity of skill pairs.
8. Certification Requirement: Create a binary feature indicating whether the skill requires a specific certification or not. This can be based on external sources or industry knowledge.
9. Skill Category: Create a new feature that categorizes skills into broader categories. You can create these categories based on external sources or industry knowledge.
10. Industry Demand: A measure of the current demand for the skill in the job market. High-1, Low-0.

Based on the features and clustering, I stuck with the same course curriculum.

1. **Introduction to Information Technology** (UI, Linux, java, php, python)
2. **Software Engineering** (Python, Java, Django, SQL, ETL)
3. **Database Systems and Data Warehousing** (ETL, SQL, Warehousing)
4. **Business Analytics and Intelligence** (Visualization, Dashboarding, Tableau, Excel, Reporting)
5. **Machine Learning** (Python, TensorFlow, PyTorch, Keras)
6. **Introduction to Artificial Intelligence** (AI, NLTK, TensorFlow, PyTorch, Python)
7. **Financial Risk Management and Analysis** (AI, Python, TensorFlow, PyTorch, Budget, Reporting)
8. **Capstone Project** (All Skills covered)

Further, Elbow Method was utilized to find the ideal number of clusters.

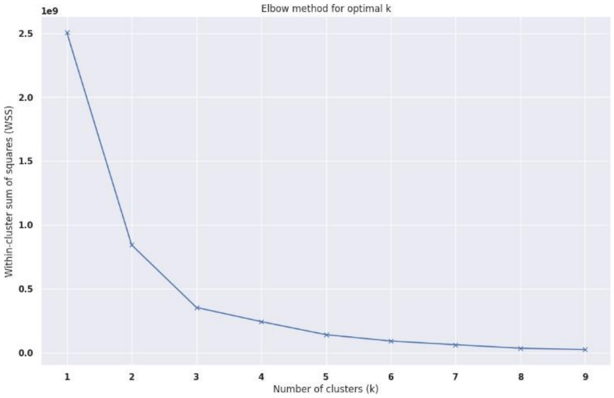


Fig 8: Elbow Method Results

A scatter plot based on the clustering results with the ideal number of clusters "2" is shown below. The elbow method's visualization can be referred from Fig. 7.

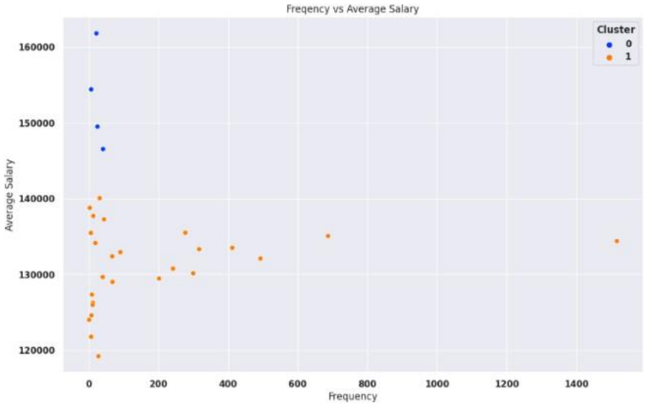


Fig 10: Frequency vs Average Salary

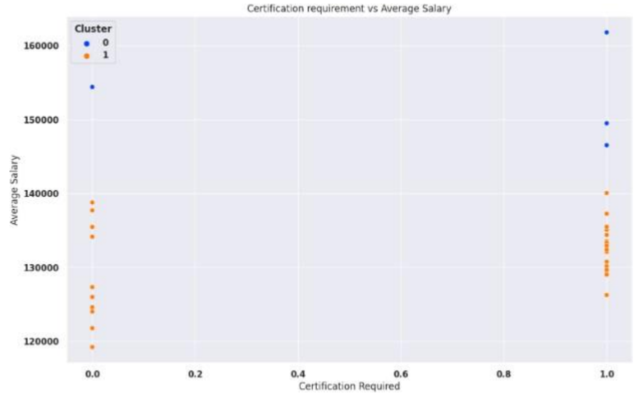


Fig. 11: Certification Requirement vs Average Salary

Based on this, the ideal number of clusters is 2, and the resulting dendrogram is:

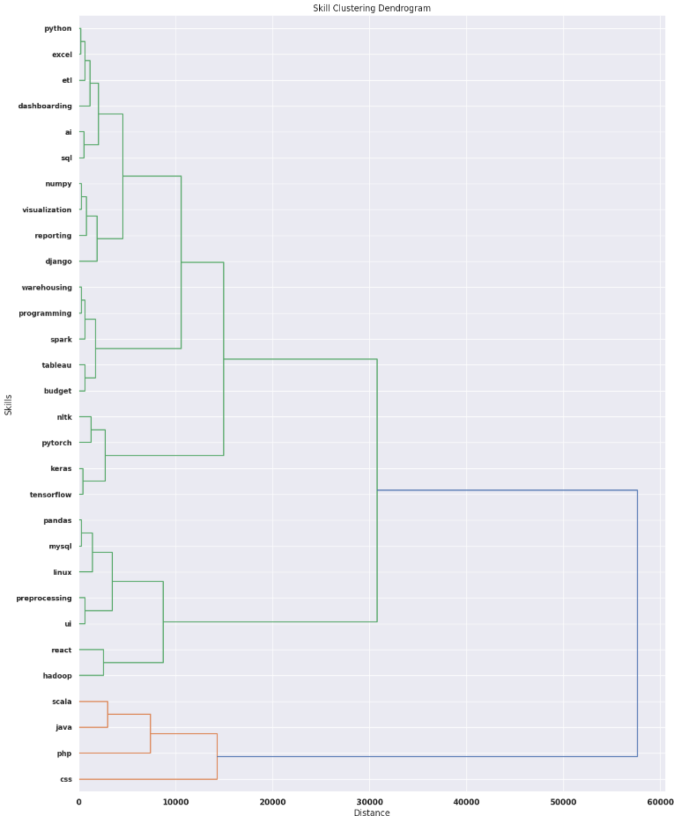


Fig 9: Dendrogram of Skills based on cluster results.

6. CONCLUSION AND FINAL COURSE CURRICULUM

Both of the course curricula look quite comprehensive and cover various skills and topics pertinent to the realm of data science and technology. However, my recommendation would be to finalize Curriculum 1 based on the dendrograms, which indicates that it provides a more structured approach to progressing through the skills and topics.

Dendrogram 1 showcases a clear clustering of skills that are associated with data engineering, data science, data visualization, big data technologies, and AI. Curriculum 1 complements this clustering by starting with courses related to database systems and data warehousing, moving on to data engineering, followed by an introduction to data science, data visualization and business intelligence, big data technologies, and AI and deep learning. This sequence of courses appears to follow a natural progression of building foundational skills before advancing to more intricate topics.

On the other hand, Dendrogram 2 shows a broader spread of skills with less structuring of clustering. Curriculum 2 starts with an introduction to information technology, followed by software engineering, and jumps back to database systems and data warehousing before covering business analytics and intelligence, machine learning, and introduction to artificial intelligence. Even though all of these topics are essential, the flow may not be as structured as Curriculum 1.

In conclusion, Curriculum 1 aligns more closely with the dendrogram and appears to provide a more organized approach to advancing through skills and topics, making it the preferable option for a course curriculum.

Thus, the final curriculum can be re-ordered to highlight the logical order in which the courses can be taken, such as ones with prerequisites.

6.1 FINAL COURSE CURRICULUM

The final course curriculum based on the analysis for the program "Master of Business and Management in Data Science and Artificial Intelligence" is:

1. **Database Systems and Data Warehousing** (ETL, SQL, warehousing)
2. **Data Engineering** (Programming, Linux, Excel)
3. **Introduction to Data Science** (Python, NumPy, Pandas, SQL)
4. **Data Visualization and Business Intelligence** (Dashboarding, Tableau, Reporting, Visualization, Excel)
5. **Big Data Technologies** (Hadoop, Spark, Warehousing, SQL, ETL)
6. **AI and Deep Learning** (TensorFlow, PyTorch, Keras, AI)
7. **AI in Finance and Management** (AI, Budget, Reporting)
8. **Capstone Project** (All skills covered).