Data Collection and Cleaning:

This report analyzes a dataset of 987 unique job postings (**after removing duplicate rows to ensure the data integration**) scraped from Indeed.com, focusing on data-related positions including **Python Analysts**, **Data Analysts**, **Data Scientists**, and **Business Analysts** across Canada. The dataset contains information including job titles, company names, locations, salary ranges, and detailed job descriptions. Initial analysis reveals that while most postings are complete in terms of basic job information, there are gaps in salary information, posting date and company ratings. The dataset structure includes 8 key columns, providing a solid foundation for deeper analysis to create courses.

Data processing, feature engineering, and visualization:

The feature engineering process for analyzing job posting skills began by using the OpenAI **GPT-3.5-turbo** API to generate skill lists. Through **five** distinct queries to ChatGPT, we gathered different categories of skills required for data-related positions, including general skills, programming-specific skills, technical/modeling skills, business intelligence capabilities, and soft skills. These generated lists were systematically organized into **five** main skill categories: programming/systems skills, visualization tools, technical skills, business skills, and personal skills. For programming skills, we tracked eight key technologies including Python, Matlab, R, and SQL. The visualization category focused on Tableau, PowerBI, and VBS, while technical skills encompassed areas like Data Management and Machine Learning. The feature extraction process implemented a **binary classification** process, creating indicators (0 or 1) for each skill's presence in job descriptions. Special attention was given to accurate detection of standalone terms, for example, programming language 'R', by implementing specific string-matching conditions to ensure I don't miss any terms. The extracted features were organized into separate Data Frames and ultimately combined with the original job posting data, resulting in a dataset with **46** columns representing different skills.

In term of Visualization, A **word cloud** was generated as an initial visualization, highlighting key terms frequently appearing in job descriptions, with prominent words like "**experience**," "**business**," "**team**," and "**management**" emerging as central themes(Appendix 1). The visualization tools distribution was illustrated through a **pie chart**, showing **Tableau** leading with 52.4% of mentions, followed closely by **PowerBI** at 47.1%, while **VBS** represented only 0.6% of the requirements(Appendix 2). In programming skills, **Excel, Python, and SQL** dominate with over 600 mentions each, while Matlab and C++ appear less frequently (Appendix 3). Technical capabilities show **Modeling and Data Management** leading with 400+ mentions, followed by **Machine Learning**, while specialized skills like Feature Engineering receive fewer mentions(Appendix 4). Business intelligence skills highlight **Reporting** as most crucial with nearly **400** mentions, followed by **Data Visualization** and **Consulting/Project Management** at around **150** mentions each(Appendix 5). In soft skills, **Communication** emerges as the top requirement with over 600 mentions, closely followed by Leadership, Collaboration, and Problem Solving(Appendix 6). After all I save skills as 2D array to prepare hierarchical clustering.

Hierarchical clustering implementation:

Using Ward's linkage method, the dendrogram visualization identified **13** distinct clusters at the chosen distance threshold at **0.4** (Appendix 7&8). However, based on the project required as least 3 distinct terms in one cluster. I manual created 8 cluster, based on my own knowledge and

thoughts(Appendix 9). After that I ask ChatGPT to create a course based on my manual cluster result (Appendix 10).

K-means clustering implementation:

The data engineering process for analyzing education and skill levels in job descriptions employed a sophisticated text analysis approach using regular expressions. For education levels, the code implemented a **5-tier hierarchical classification system**, ranging from Level 5 (PhD/doctorate) down to Level 1 (high school), with each level containing multiple variations of degree terminology. The system scans job descriptions using case-insensitive pattern matching, prioritizing higher education levels in its search sequence. For skill proficiency, a **3-level classification system** was implemented, categorizing requirements as Level 3 (advanced/expert), Level 2 (intermediate), or Level 1 (beginner). The code generates ten distinct features to comprehensively analyze skills in job postings. It begins with basic **Skill Frequency** counting, followed by calculating **Average Co-occurring Skills** to understand skill relationships by identifying how many other skills typically appear alongside each skill. The **Average Job Description Length** measures the complexity of job postings containing each skill. Two binary indicators are implemented: a **Soft Skills Indicator** to flag non-technical competencies and a **Language Skills Indicator** specifically for programming languages (Python, R, Java, etc.). The **Average Education Level** feature utilizes the previously extracted education levels to understand the educational requirements associated with each skill. For organizational context, a **Management Role Proportion** is calculated by identifying management-related keywords in job titles. The **Geographical Spread** measures each skill's demand across different locations by counting unique locations per skill. The **Average Skill Level** incorporates the previously computed skill proficiency levels. Finally, a **Visualization** indicator flags skills specifically related to data visualization tools.

After that the skill clustering analysis began by standardizing ten key features using StandardScaler and determining the optimal number of clusters through the **Elbow Method**. The elbow curve showed a clear inflection point at **k=7**, indicating the optimal number of clusters (Appendix 11). The K-means clustering algorithm was then applied to group the skills into seven distinct clusters, each representing different aspects of data science competencies. These clusters were then used to inform the development of a **10-course curriculum** that logically organizes related skills. The curriculum ranges from fundamental programming courses to specialized topics like data visualization and business consulting. Each course contains 3-4 complementary skills, ensuring comprehensive coverage while maintaining manageable scope. The courses were designed to progress from core technical skills (like Python and SQL) to more advanced topics (such as Machine Learning and Feature Engineering), and finally to business and management skills, creating a well-rounded educational pathway.

Interpretation of results using ChatGPT API:

The analysis used the ChatGPT API to interpret and organize the clustered skills into a coherent curriculum structure. The process began by formatting a prompt that presented the eight distinct skill clusters derived from the K-means analysis. Each cluster was formatted as a course with its associated skills: "Course 1: Modeling, SQL, Python, Leadership, Communication, Excel" and so on. When this prompt was sent to the GPT model through the OpenAI API, it generated a structured interpretation that organized the skills into three main categories: **Data Analysis and Programming**

(Courses 1-4), **Business and Management Skills** (Courses 5-6), and **Data Visualization and Business Intelligence** (Courses 7-8). The API response provided not only the categorization but also a valuable analysis of how these courses interconnect and their relevance to different career paths. This interpretation helped validate the clustering results and provided a practical framework for curriculum development.(Appendix 12).

Discussion and final course curriculum:

The proposed data science curriculum consists of eight carefully structured courses that provide a comprehensive education pathway. **Course 1: Foundations of Data Science and Programming** establishes the core technical foundation with Python, SQL, and Excel, ensuring students develop essential data manipulation skills. This leads into **Course 2: Advanced Data Techniques**, which deepens technical expertise through data cleaning, wrangling, mining, and ETL processes, preparing students to handle complex, real-world datasets. **Course 3: Data Visualization and Creativity** focuses on translating data into visual insights using tools like Power BI and Tableau, complemented by presentation skills and VBS knowledge. **Course 4: Machine Learning and Team Dynamics** combines technical depth in machine learning algorithms with R programming while emphasizing collaborative work environments.

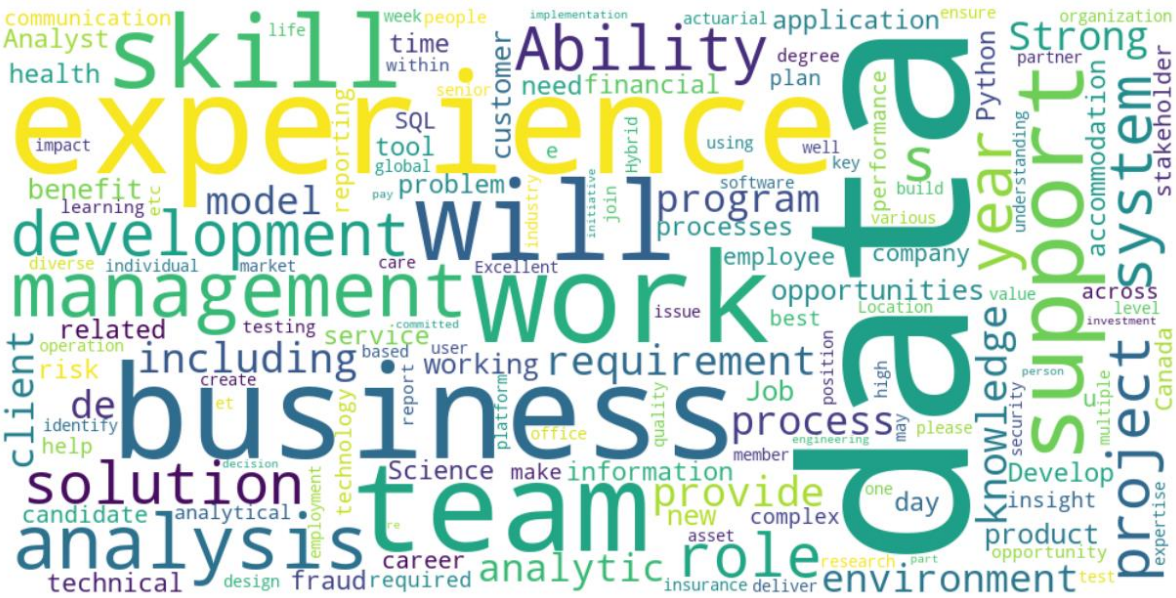
The curriculum then transitions to business-oriented courses, starting with **Course 5: Project Management and Consulting Skills**, which develops crucial soft skills in project execution, client relations, and budgeting. **Course 6: Leadership and Team Collaboration** builds on this foundation by focusing on interpersonal skills, communication, and adaptability in diverse team settings. For those seeking deeper technical expertise, **Course 7: Advanced Programming for Data Science** offers specialized training in C++, Java, and SAS for optimizing computational performance. The curriculum concludes with **Course 8: Business Intelligence and Decision Making**, which bridges technical and business domains, emphasizing critical thinking and risk management in business contexts. This carefully sequenced curriculum ensures students develop both technical proficiency and essential business acumen required for success in the data science field.

OpenAI to describe clustering results:

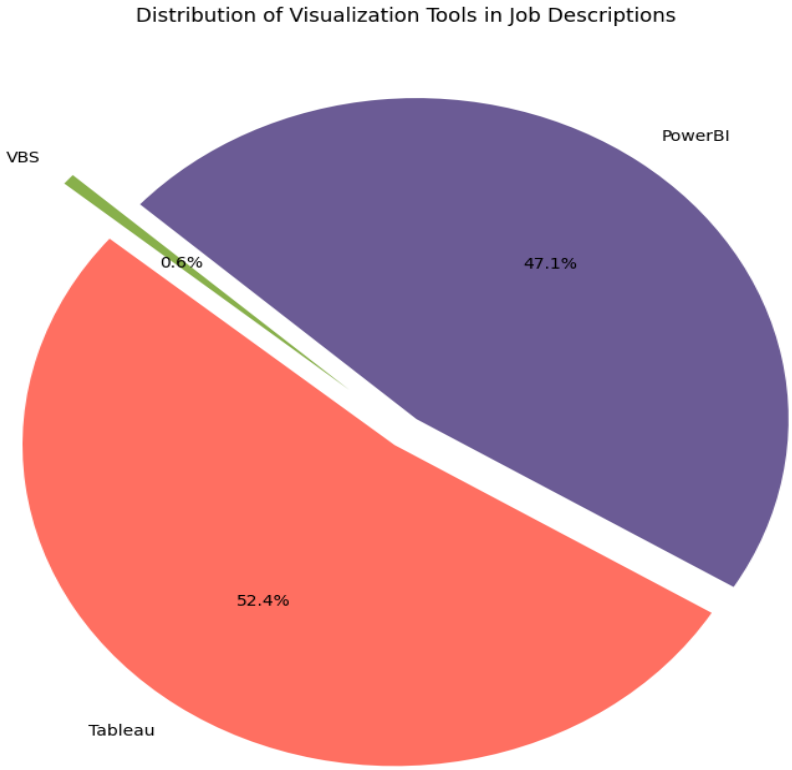
At last, I process utilized OpenAI's text embedding model (text-embedding-ada-002) and K-means clustering to analyze job descriptions and create a structured data science curriculum. The process began by generating embeddings for each job description, implementing error handling to manage potential API failures by substituting zero vectors when necessary. The resulting embeddings were then processed using K-means clustering with eight clusters to group similar job descriptions together. The final curriculum was organized into **twelve comprehensive courses**, ranging from foundational skills to specialized applications. The courses progressed from basic programming and data manipulation (Courses 1-2) through visualization and machine learning (Courses 3-4), to business skills (Courses 5-6), advanced programming (Course 7), and business intelligence (Course 8). The curriculum was further enhanced with advanced statistical methods (Course 9), cloud computing (Course 10), ethics (Course 11), and domain-specific applications (Course 12), creating a well-rounded educational program for aspiring data scientists.

Appendix

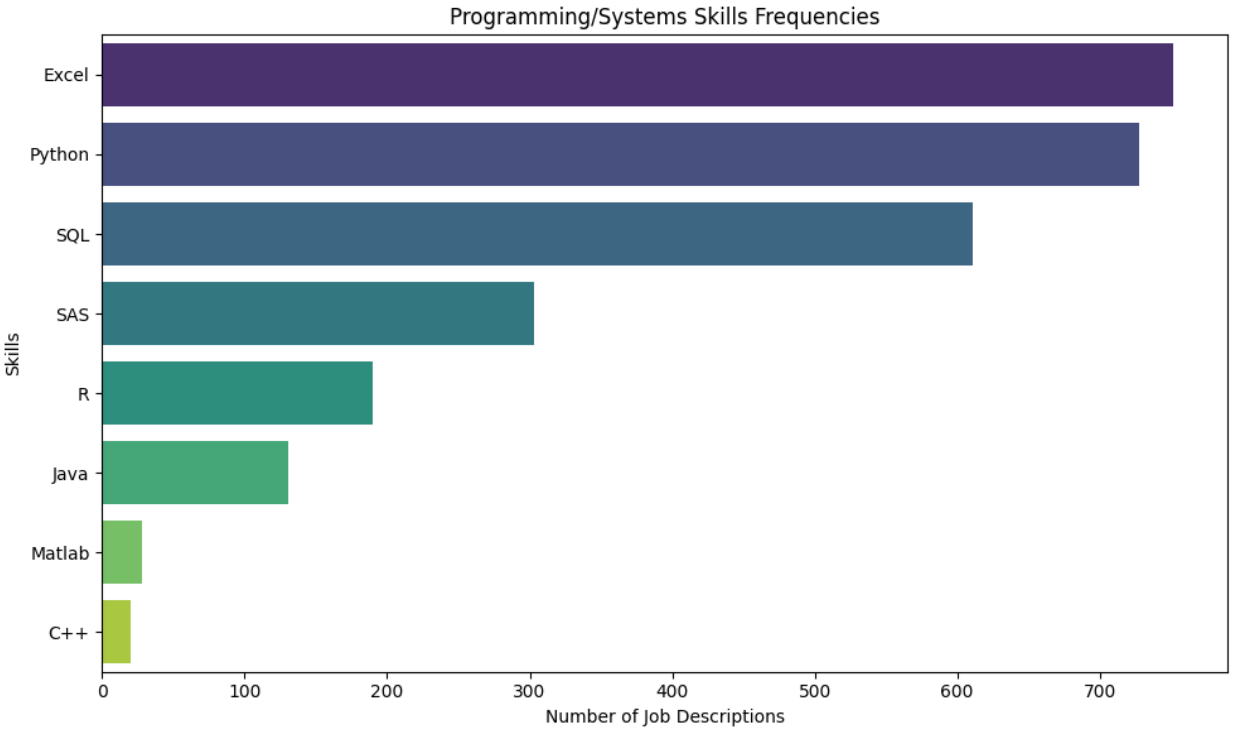
Appendix 1



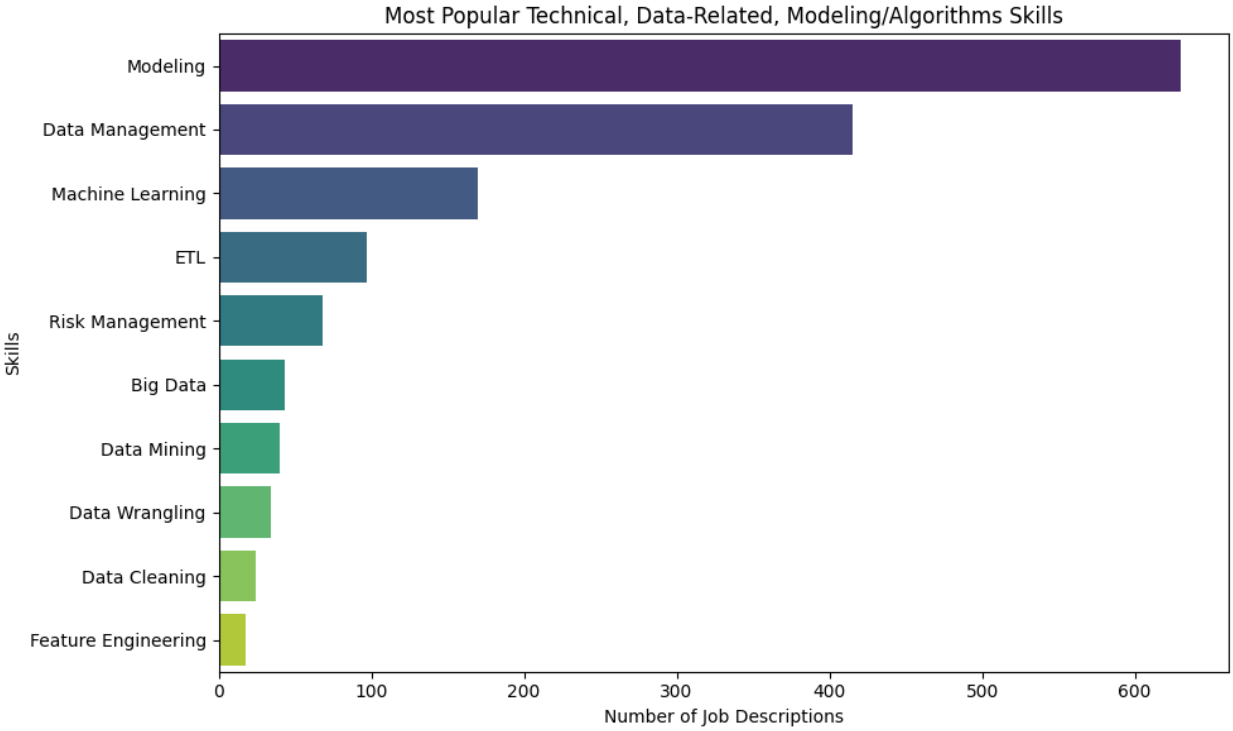
Appendix 2



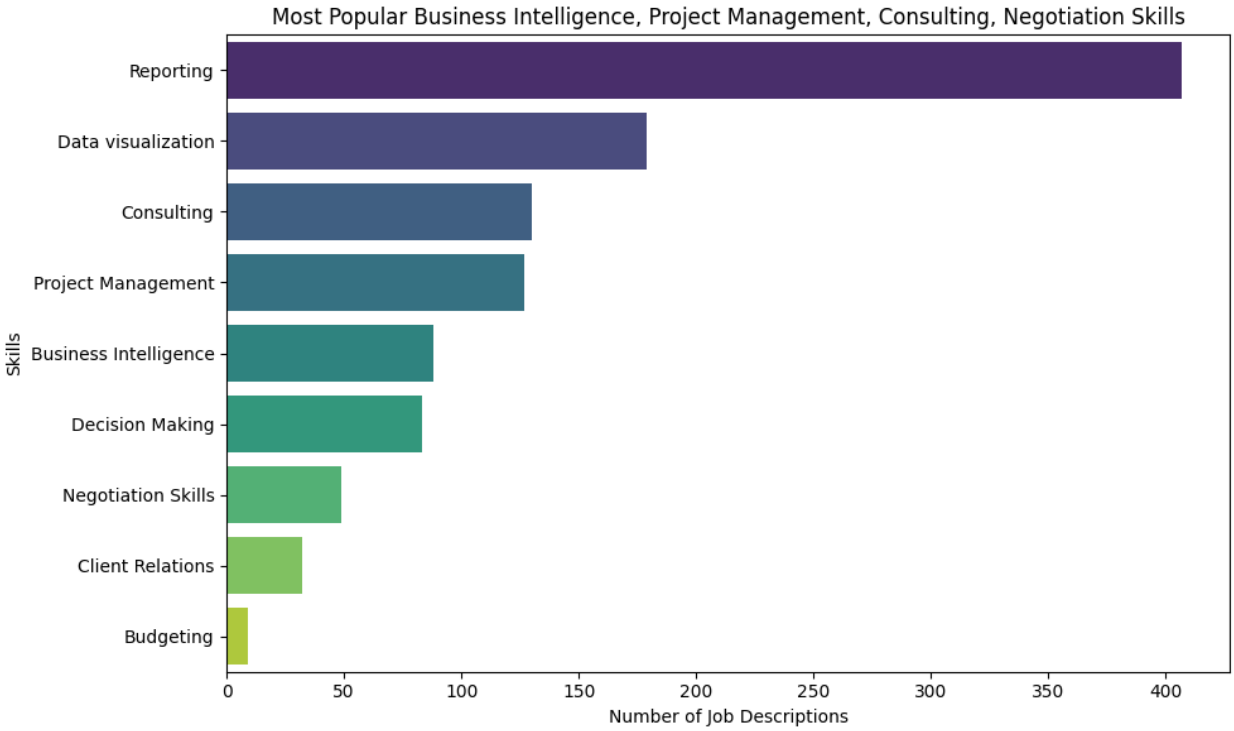
Appendix 3



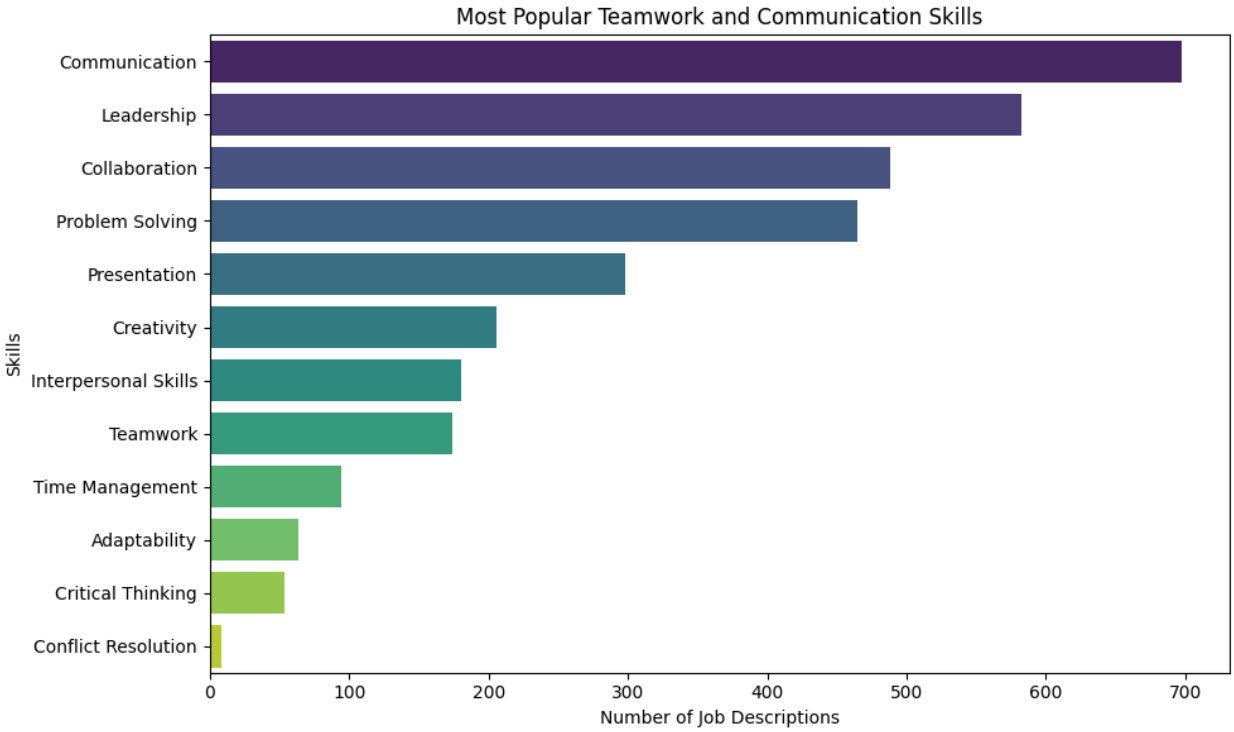
Appendix 4



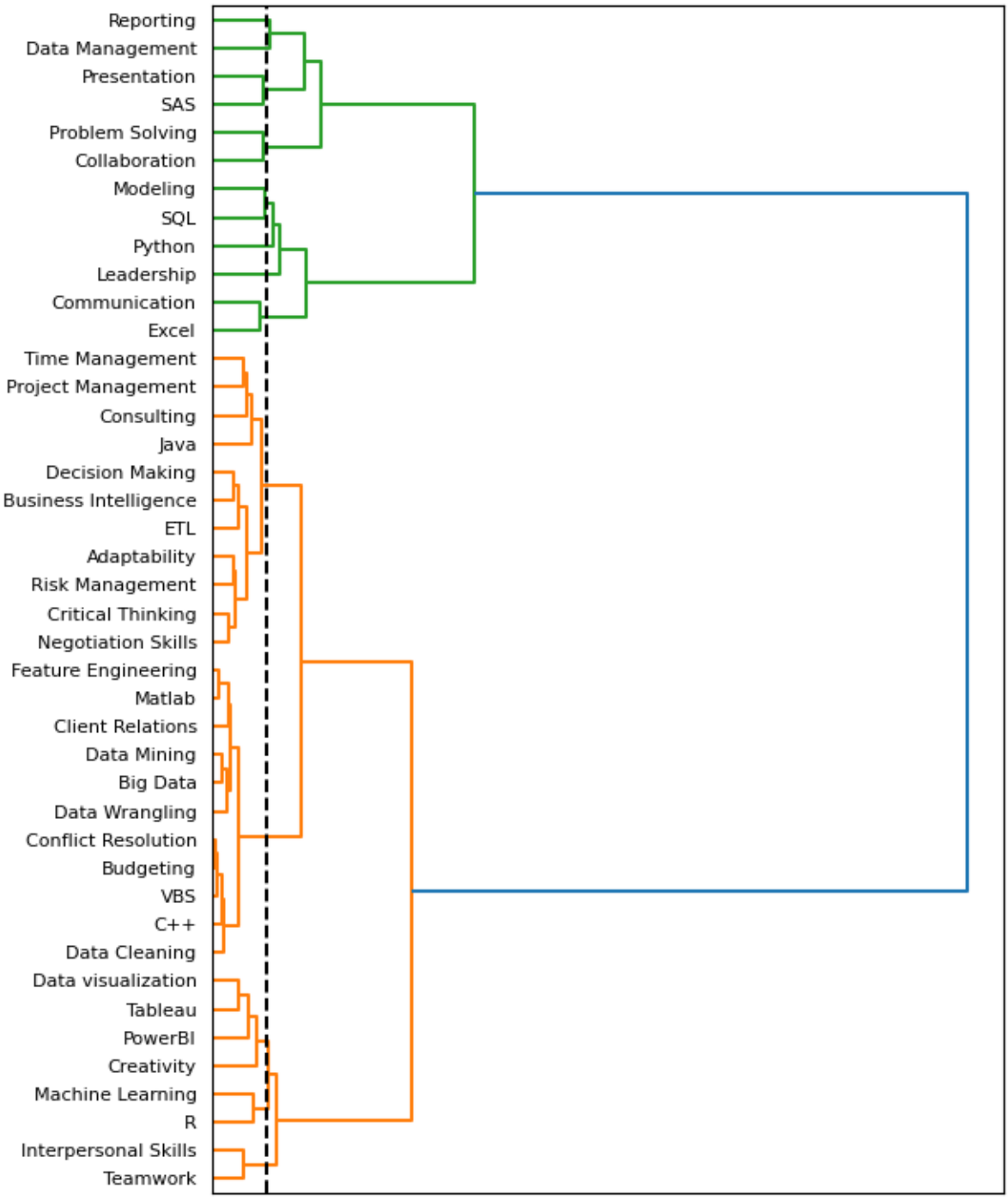
Appendix 5



Appendix 6



Appendix 7



Appendix 8

```
1 ['Teamwork', 'Interpersonal Skills']
2 ['R', 'Machine Learning']
3 ['Creativity', 'PowerBI', 'Tableau', 'Data visualization']
4 ['Data Cleaning', 'C++', 'VBS', 'Budgeting', 'Conflict Resolution', 'Data Wrangling', 'Big Data', 'Data Mining', 'Client Relations', 'Matlab', 'Feature Engineering']
5 ['Negotiation Skills', 'Critical Thinking', 'Risk Management', 'Adaptability', 'ETL', 'Business Intelligence', 'Decision Making', 'Java', 'Consulting', 'Project Management', 'Time Management']
6 ['Excel', 'Communication']
7 ['SQL', 'Modeling']
8 ['Python']
9 ['Leadership']
10 ['Collaboration', 'Problem Solving']
11 ['SAS', 'Presentation']
12 ['Data Management']
13 ['Reporting']
```

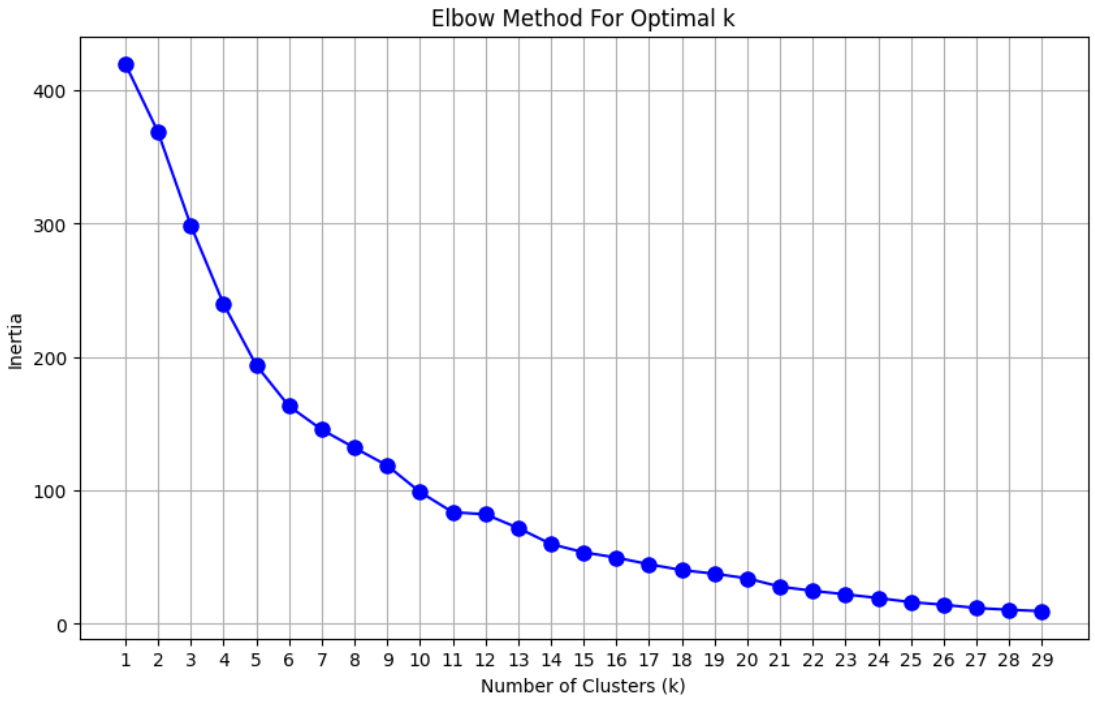
Appendix 9

```python
# Crate a dictionary with the clusters
manual_clusters = {}
manual_clusters[0] = ['Modeling', 'SQL','Python', 'Leadership', 'Communcation', 'Excel']
manual_clusters[1] = ['Data Cleaning','Data Wrangling', 'Data Mining', 'Big Data', 'Feature Engineering', 'Matlab', 'C++', 'VBS']
manual_clusters[2] = ['Reporting', 'Presentation', 'SAS', 'Data Management']
manual_clusters[3] = ['Machine Learning', 'R', 'Team Work']
manual_clusters[4] = ['Conflict Resolution', 'Budgeting', 'Client Relations']
manual_clusters[5] = ['Consulting', 'Project Management', 'Time Management']
manual_clusters[6] = ['PowerBI', 'Tableau', 'Creativity', 'Data Visualization']
manual_clusters[7] = ['Negotiation Skills', 'Critical Thinking', 'Risk Management', 'Adaptability', 'ETL', 'Business Intelligence', 'Decision Making']
```

Appendix 10

```
Course 1: Modeling, SQL, Python, Leadership, Communcation, Excel; Course 2: Data Cleaning, Data Wrangling, Data Mining, Big Data, Feature Engineering, Matlab, C++, VBS; Cours
e 3: Reporting, Presentation, SAS, Data Management; Course 4: Machine Learning, R, Team Work; Course 5: Conflict Resolution, Budgeting, Client Relations; Course 6: Consultin
g, Project Management, Time Management; Course 7: PowerBI, Tableau, Creativity, Data Visualization; Course 8: Negotiation Skills, Critical Thinking, Risk Management, Adaptabi
lity, ETL, Business Intelligence, Decision Making
```

Appendix 11



Appendix 12

It looks like you have a list of courses covering a wide range of skills related to data, business, and personal development. Here's a structured overview of these courses, categorized by themes, along with the skills and tools associated with each course:

Course Overview

Data Analysis and Programming
- **Course 1**:
  - Skills/Tools: Modeling, SQL, Python, Excel
- **Course 2**:
  - Skills/Tools: Data Cleaning, Data Wrangling, Data Mining, Big Data, Feature Engineering, Matlab, C++, VBS
- **Course 3**:
  - Skills/Tools: Reporting, Presentation, SAS, Data Management
- **Course 4**:
  - Skills/Tools: Machine Learning, R, Team Work

Business and Management Skills
- **Course 5**:
  - Skills/Tools: Conflict Resolution, Budgeting, Client Relations
- **Course 6**:
  - Skills/Tools: Consulting, Project Management, Time Management

Data Visualization and Business Intelligence
- **Course 7**:
  - Skills/Tools: PowerBI, Tableau, Creativity, Data Visualization
- **Course 8**:
  - Skills/Tools: Negotiation Skills, Critical Thinking, Risk Management, Adaptability, ETL, Business Intelligence, Decision Making

Summary
This collection of courses offers a solid foundation in both technical data skills and essential business competencies. You can choose courses based on your career goals, whether you want to focus on data analysis, visualization, programming, or enhancing your business acumen and leadership skills.

If you're looking for advice on how to prioritize these courses or how they might interconnect in a given field, please let me know!