Project 3

Question 2b:

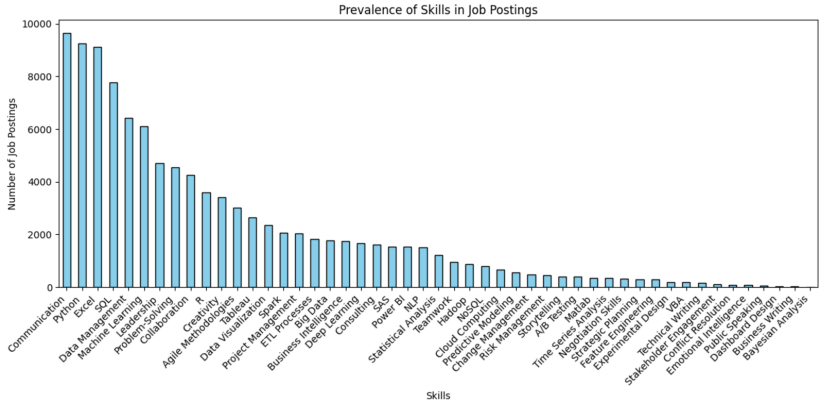

Figure 1: Prevalence of Skills in Job Postings

Figure 1 highlights the demand for both technical and soft skills in job postings. Communication is the most sought-after skill, followed by Python, Excel, and Agile Methodologies, emphasizing the importance of clear communication, programming, and organizational frameworks. Data management, machine learning, and data visualization are also highly valued, reflecting the growing demand for data expertise. Mid-level skills like project management and business intelligence underscore the need for strategic capabilities, while soft skills such as teamwork and leadership complement technical expertise. Niche skills like Bayesian analysis cater to specialized roles, offering differentiation opportunities. Professionals should prioritize communication, technical, and leadership skills to stay competitive.

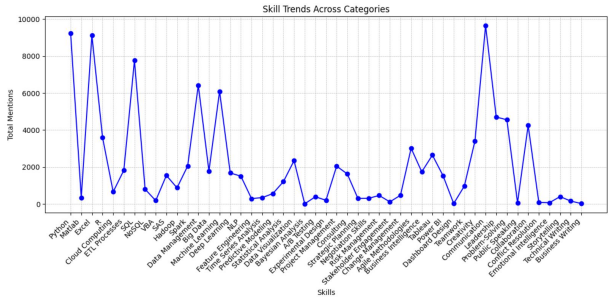

Figure 2: Skill Trends Across Categories

Figure 2 highlights skill trends across job categories, showing Python and Communication as the most mentioned, with nearly 10,000 mentions each. Skills like Excel, Cloud Computing, and ETL Processes also rank highly, reflecting the demand for technical expertise, while Teamwork emphasizes collaboration. Business-oriented skills such as Project Management and Business Intelligence maintain relevance, bridging technical and managerial roles. Niche skills like Bayesian Analysis and Experimental Design are less common, catering to specialized roles. The chart underscores the need for a balance of technical, analytical, and interpersonal skills,

encouraging professionals to combine expertise with collaboration and explore niche specializations.

Figure 3 displays the distribution of skills mentioned per job posting, revealing that most postings require 5 to 10 skills, with a peak around 6. This suggests employers value candidates with a balance of specialization and versatility. The sharp decline beyond 10 skills reflects the rarity of roles needing extensive expertise, while postings with fewer than 5 skills are uncommon, emphasizing the demand for multi-disciplinary capabilities. Job seekers should focus on developing 5–10 key skills to stay competitive in a market that values breadth and relevance.

Figure 4 shows the proportion of various skills in job postings, with the "Other" category making up 36.8%, reflecting diverse niche demands across industries. Communication is the most prevalent skill (9.3%), followed by Python (8.9%) and Excel (8.8%), highlighting the importance of programming and data manipulation. SQL (7.5%) and Data Management (6.2%) emphasize data-driven expertise, while soft skills like Leadership (5.9%), Collaboration (4.1%), and Problem-Solving (4.4%) are essential for teamwork and execution. Machine Learning (4.4%) and R (3.5%) highlight the rise of advanced analytics. The chart emphasizes balancing technical, analytical, and soft skills while exploring niche areas for specialization.

Question 3:

For figure 5, the hierarchical clustering dendrogram groups skills based on their co-occurrence in job postings, revealing distinct clusters of related competencies. Soft skills like communication, teamwork, and public speaking form one cluster, while technical skills such as Python, R, and machine learning form another, highlighting their demand in data science. Data visualization tools like Tableau and Power BI and business-oriented skills such as project management and strategic planning create separate clusters, reflecting their relevance in specific domains. Big data and cloud-related skills like Spark and NoSQL form specialized clusters, while versatile skills like SQL bridge multiple groups. The maximum dendrogram depth was set to max_d = 1.52, yielding over 9 meaningful clusters, each containing 3 or more related skills. This depth ensured practical groupings ideal for designing courses or training programs that focus on interconnected skills without being overly broad or narrow.

Table 1: Course Curriculum via hierarchical clustering algorithm

Courses	Topics
Course 1	Consulting, Cloud Computing, Data Visualization, Strategic Planning
Course 2	Deep Learning, Dashboard Design, Predictive Modeling, Power BI
Course 3	Creativity, Conflict Resolution, Emotional Intelligence, Storytelling, A/B Testing, Python, Problem-Solving
Course 4	Stakeholder Engagement, Collaboration, Big Data, SQL, Experimental Design
Course 5	Agile Methodologies, Change Management, Business Intelligence, SAS, Statistical Analysis, Excel, NLP
Course 6	ETL Processes, Project Management, Public Speaking
Course 7	VBA, Machine Learning, Risk Management
Course 8	Feature Engineering, Bayesian Analysis, Technical Writing, R, Hadoop
Course 9	Time Series Analysis, Teamwork, Negotiation Skills, Communication, Business Writing

This curriculum offers a well-rounded program to develop technical, analytical, and interpersonal skills for modern industries. Courses cover a range of topics: strategic and technical skills like

Cloud Computing and Data Visualization (Course 1), advanced analytics with Deep Learning and Predictive Modeling (Course 2), and creativity and emotional intelligence with Conflict Resolution and Storytelling (Course 3). Other courses focus on collaboration and data skills (Course 4), leadership and management (Course 5), data engineering and communication (Course 6), risk management and technical expertise (Course 7), advanced data engineering (Course 8), and interpersonal skills like Teamwork and Negotiation (Course 9). Together, these courses provide an industry-relevant foundation for diverse career paths.

Question 4:

For figure 6, the Elbow Method graph is used to determine the optimal number of clusters (k) for a dataset. In this case, the optimal value of k was chosen to be 11 because it provides a balance between compact clusters (low inertia) and meaningful grouping. The "elbow" point, where the rate of decrease in inertia slows significantly, occurs at k = 11, indicating that adding more clusters beyond this point results in diminishing returns in terms of reducing within-cluster variance.

By selecting k = 11, the analysis yielded 8 meaningful courses, each containing a minimum of 3 closely related skills. This approach ensures that the clusters are both interpretable and practical, as each course bundles a reasonable number of interconnected skills, making them relevant for training or educational purposes. This choice reflects a thoughtful trade-off between granularity and usability, aligning the clustering results with the objective of grouping skills into actionable course structures.

Figure 7 visualizes clusters from k-means clustering (k = 11) using PCA for dimensionality reduction. Each point represents a skill, colored by cluster, with axes capturing the most significant data variance. Most clusters are distinct, showing effective grouping of related skills, while slight overlaps suggest shared characteristics. Closely grouped skills are ideal for inclusion in the same course, while distant clusters highlight dissimilar skill sets requiring specialized focus. This visualization is a valuable tool for designing targeted courses by identifying complementary and unique skills for training programs.

Table 2: Course Curriculum via K-mean

Courses	Topics
Course 1	Matlab, Time Series Analysis, Bayesian Analysis
Course 2	Excel, SQL, Communication
Course 3	R, Agile Methodologies, Leadership
Course 4	Cloud Computing, Deep Learning, NLP
Course 5	ETL Processes, VBA, Spark
Course 6	NoSQL, SAS, Big Data
Course 7	Hadoop, Consulting, Power BI
Course 8	A/B Testing, Strategic Planning, Negotiation Skills

This curriculum equips students with a diverse range of skills for modern industries. Course 1 develops computational and statistical expertise with Matlab and Bayesian Analysis for data science roles. Course 2 builds foundational skills like Excel, SQL, and Communication for data management. Course 3 combines R, Agile Methodologies, and Leadership for project execution.

Course 4 focuses on advanced AI technologies like Cloud Computing and Deep Learning. Course 5 emphasizes data engineering with ETL Processes and Spark, while Course 6 covers Big Data tools like NoSQL and SAS. Course 7 integrates Hadoop, Consulting, and Power BI for business intelligence and client-facing roles. Finally, Course 8 enhances strategic abilities with A/B Testing and Negotiation Skills, fostering expertise in business strategy. Together, these courses offer a comprehensive skill set for diverse career paths.

Question 5:

Results are in the python file.

Question 6:

The course curriculum via k-means is a well-rounded program designed to equip students with essential technical, analytical, and interpersonal skills to excel in modern, data-driven industries. It features eight courses, each with a distinct focus. From foundational tools like Excel, SQL, and Communication to advanced topics such as Cloud Computing, Deep Learning, and NLP, the curriculum ensures a diverse skill set. Courses like Time Series Analysis and Bayesian Analysis cater to quantitative fields, while Agile Methodologies and Leadership prepare students for management roles. Data engineering concepts such as ETL Processes, Spark, and Hadoop are paired with business intelligence tools like Power BI and consulting skills, ensuring industry relevance. Soft skills, including Negotiation Skills, Strategic Planning, and Emotional Intelligence, are integrated to prepare students for client-facing and leadership positions. The curriculum achieves a balance between technical and soft skills, with a modular structure that allows students to customize their learning paths, making it highly adaptable to diverse career goals. This program addresses the needs of various industries, enabling students to stand out in a competitive job market.

Appendix

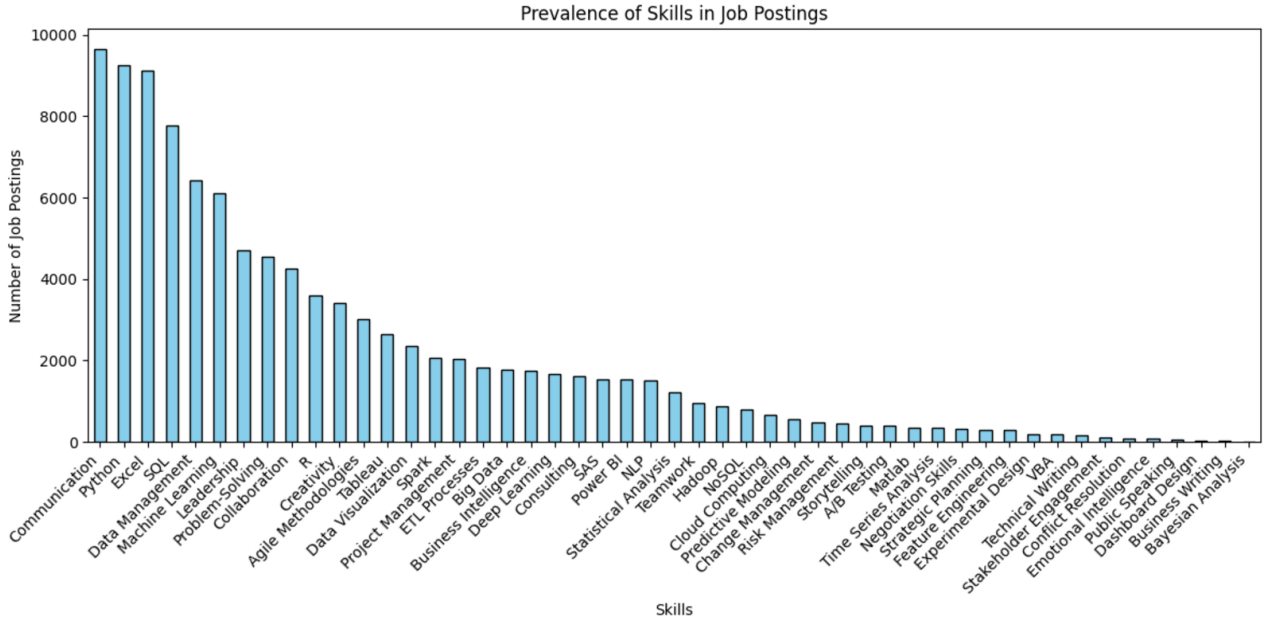

Figure 1: Prevalence of Skills in Job Postings

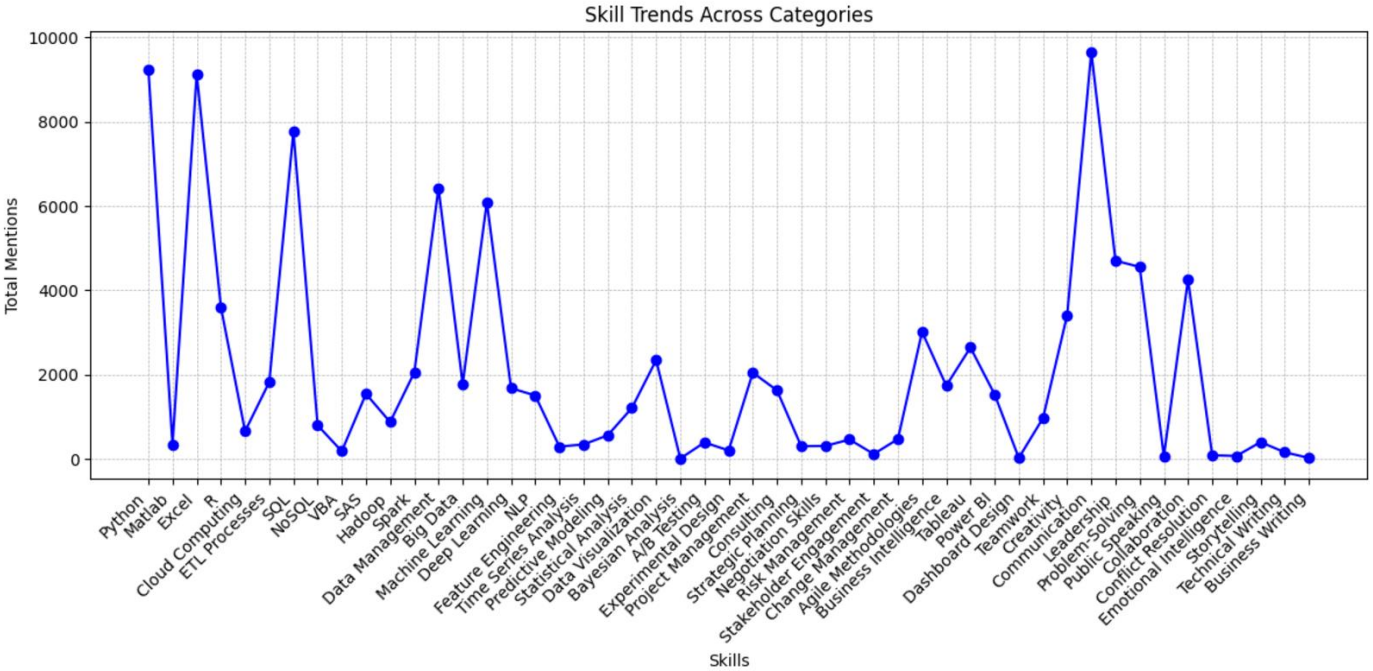

Figure 2: Skill Trends Across Categories

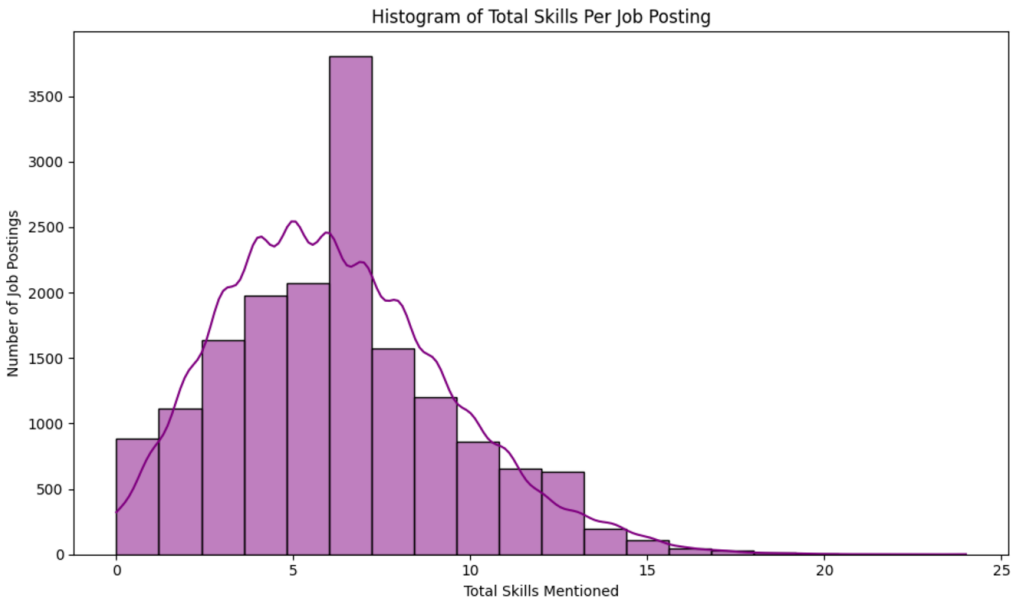

Figure 3: Histogram of Total Skills Per Job Posting

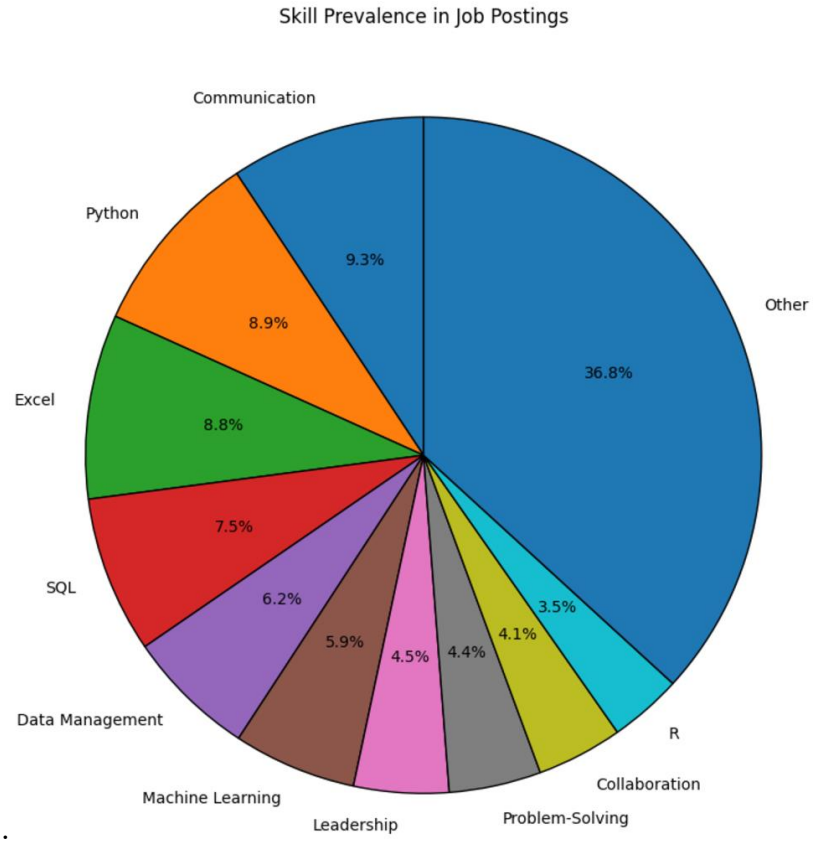

Figure 4: Skill Prevalence in Job Postings

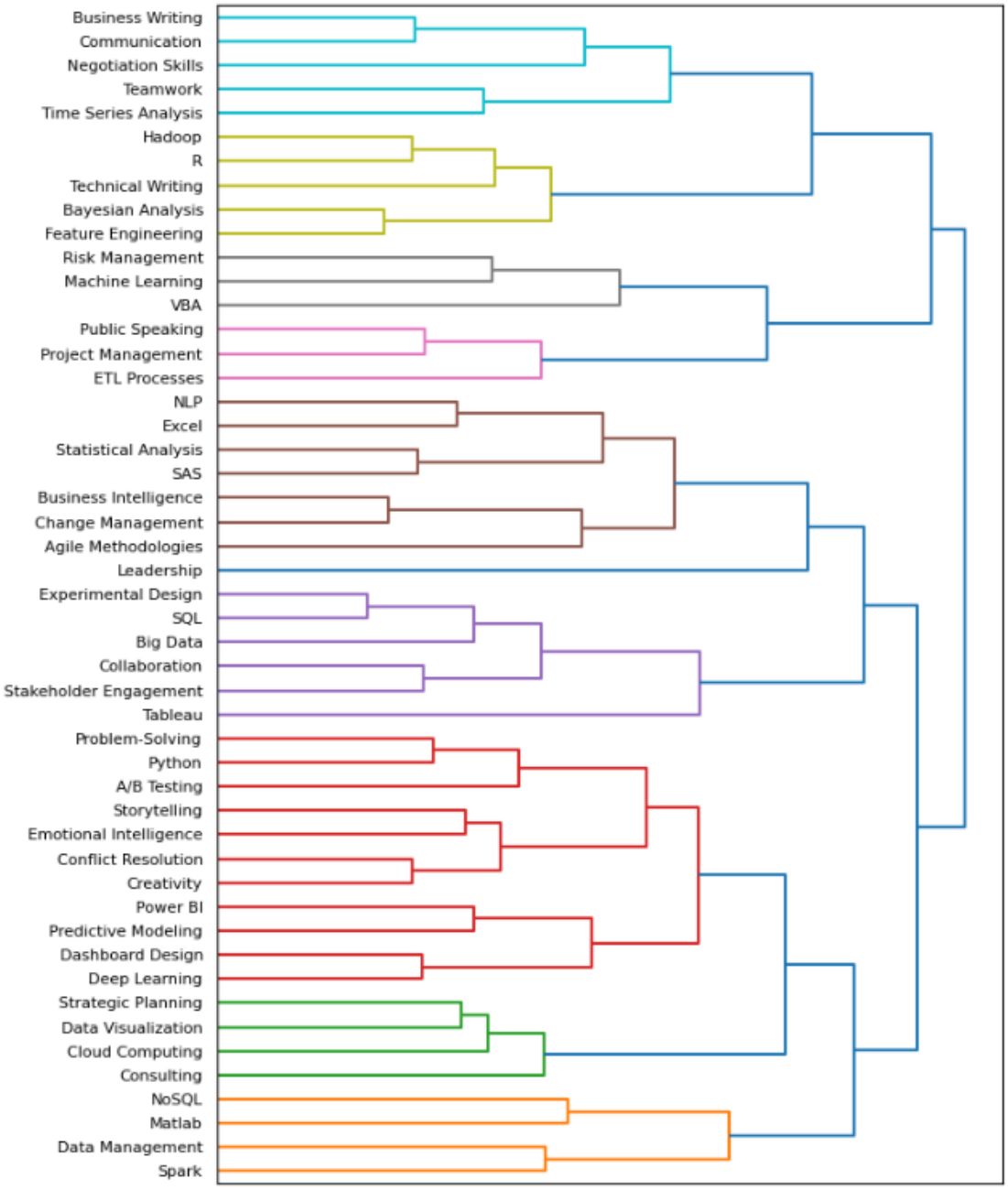

Figure 5: Hierarchical Clustering Dendrogram

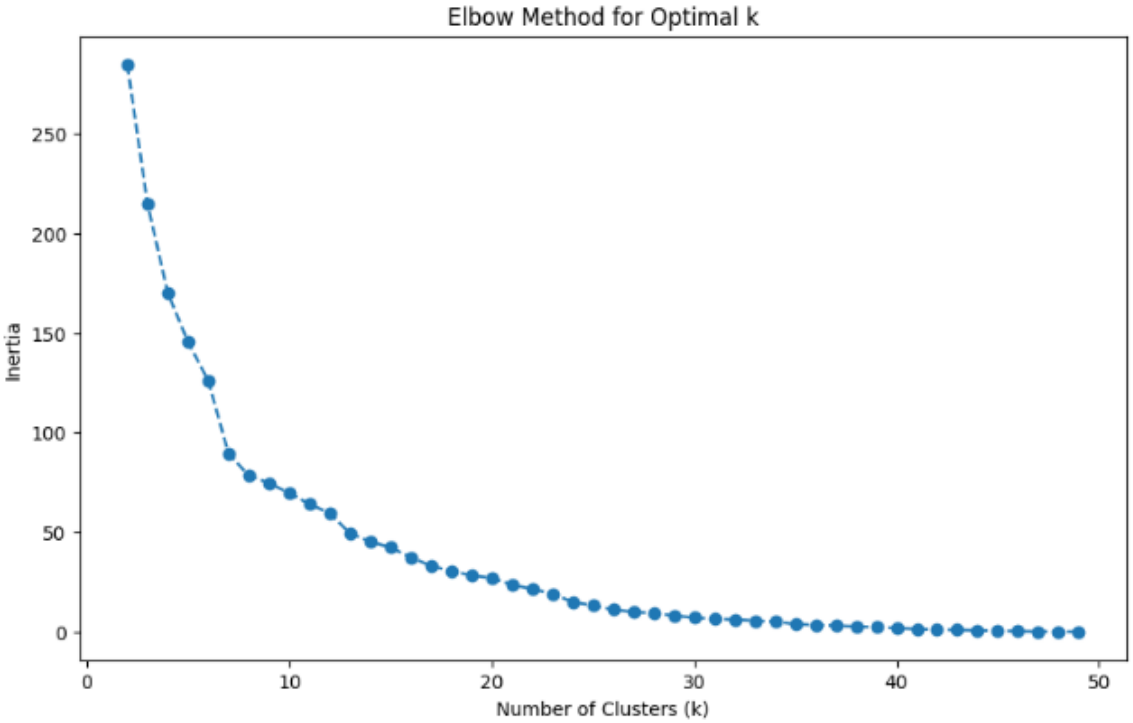

Figure 6: Elbow Method for Optimal k

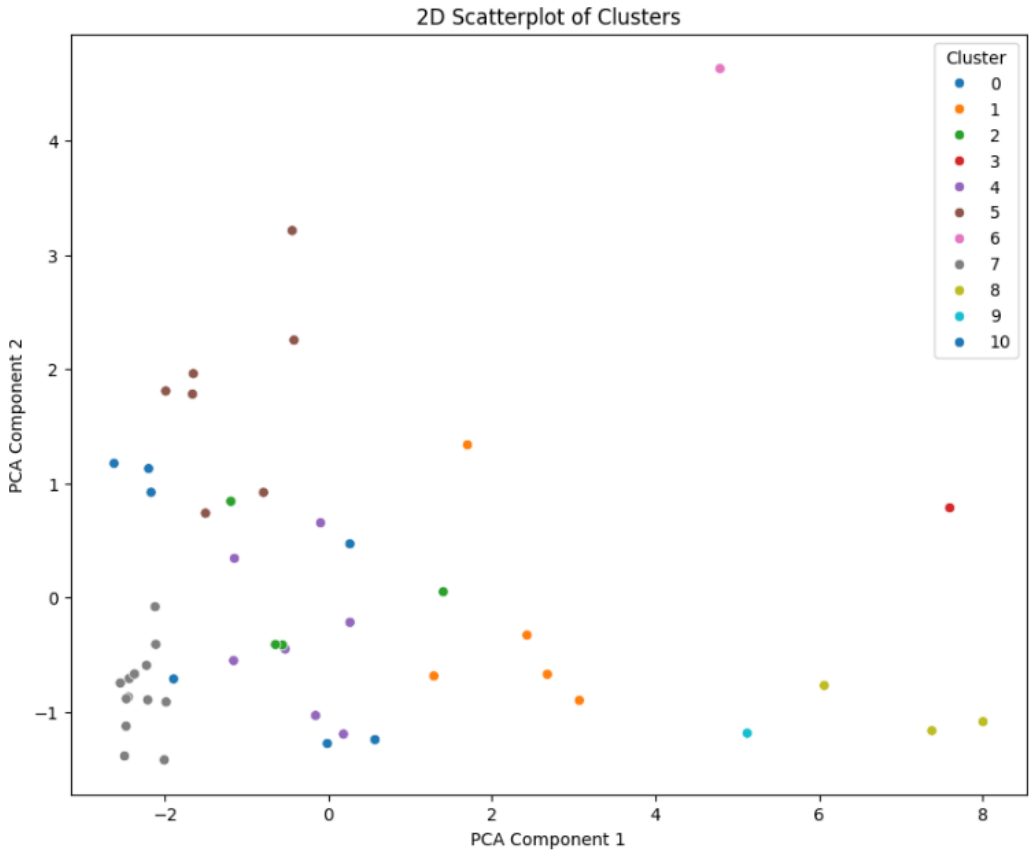

Figure 7: 2D Scatterplot of Clusters