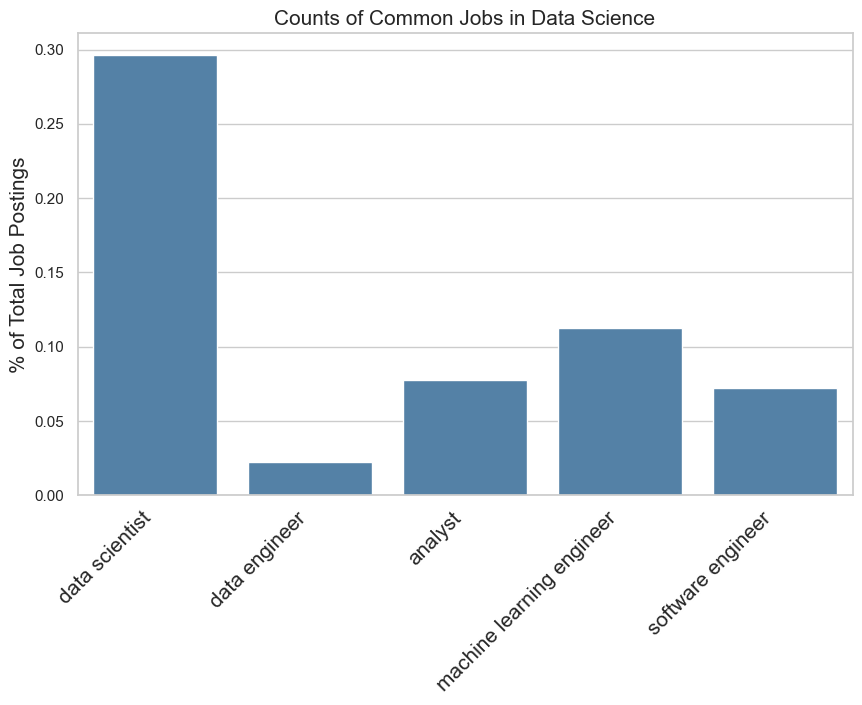
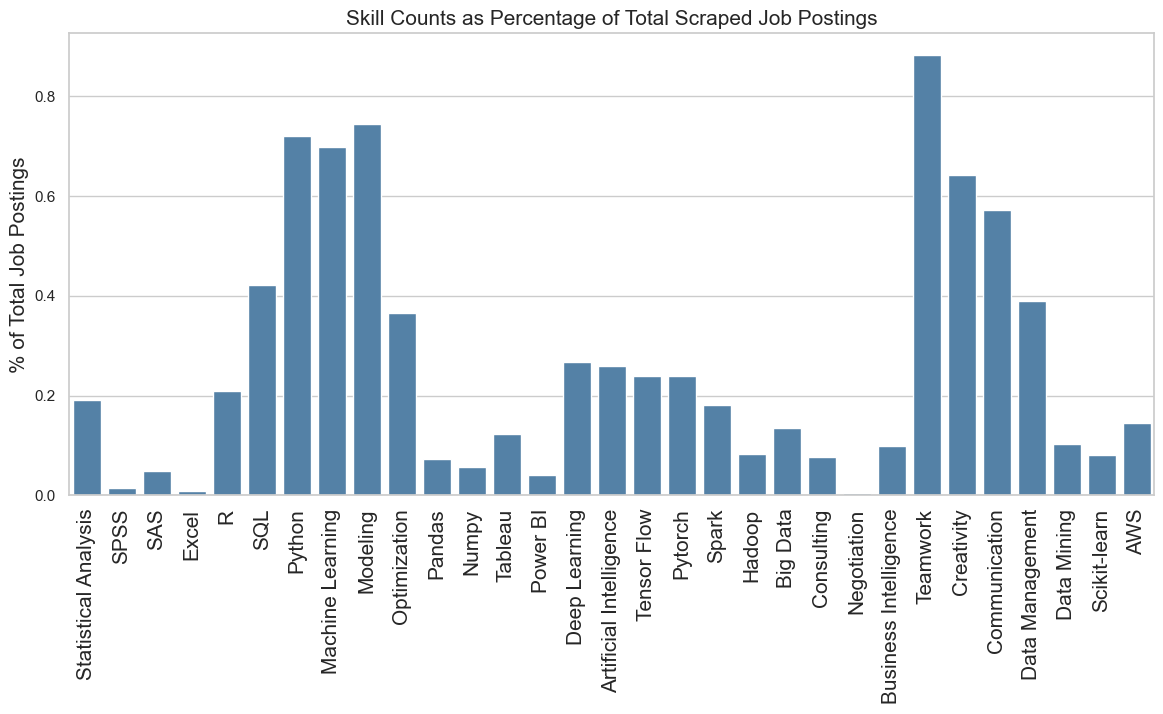
**Assignment 3**

**Data Collection and Cleaning**

Data was scraped from indeed using the web-scraping sample code provided. To manage the substantial computational load, the data collection process involved scraping in segments of 150 job postings, which were subsequently combined. The initial location was set to California, USA, and 1500 postings were scraped. After inspection and removal of duplicated postings, 800 unique postings remained. To supplement the additional 200 postings required, New York, USA, was used due to its similarity in the job market within the data science sector.

**Exploratory Data Analysis**

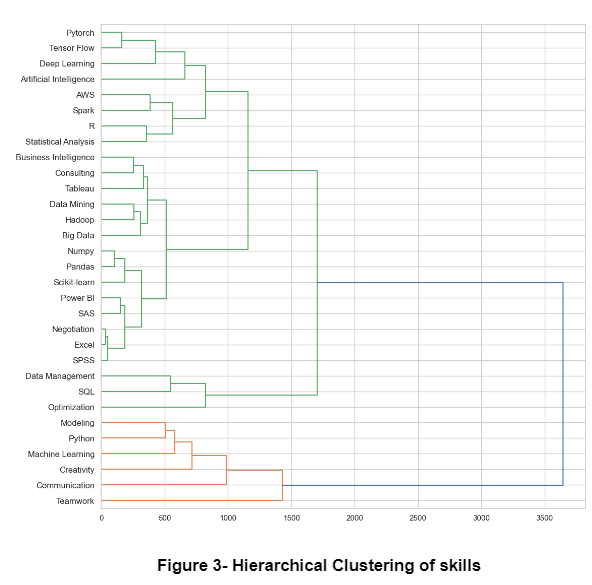
By employing both the chatGPT API and manual research, I identified crucial skills for data science roles. These skills were then used to assess their occurrence in job descriptions, with binary indicators (1 or 0) assigned to columns indicating skills in a new dataframe. Exploratory Data Analysis unveiled notable trends. Figure 1 illustrates soft skills like teamwork and creativity, accounting for over 80% of postings. Figure 2 highlights the distribution of common data science roles, with 'data scientist' comprising the largest segment at 30% of all scraped postings.



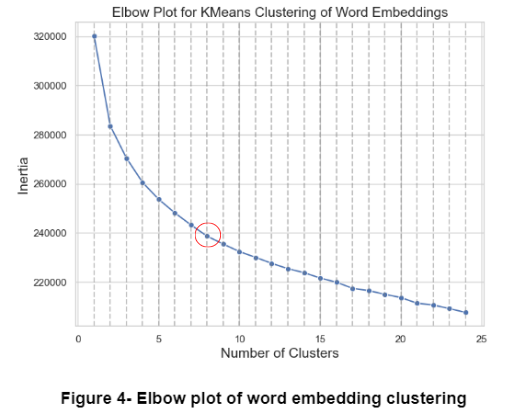
**Figure 1- Skill counts (%) of total scraped job postings Figure 2- Counts of common jobs in**

**Data Science**

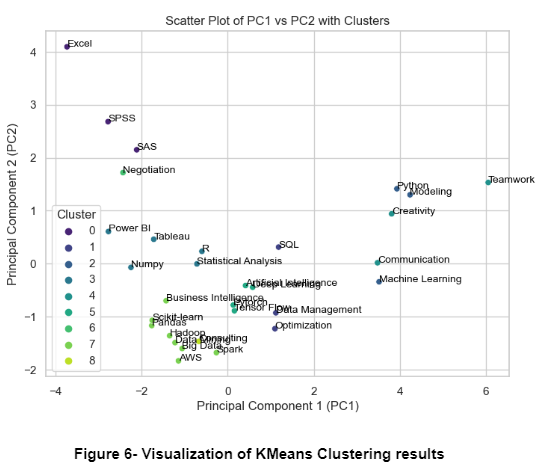
**Hierarchical Clustering**

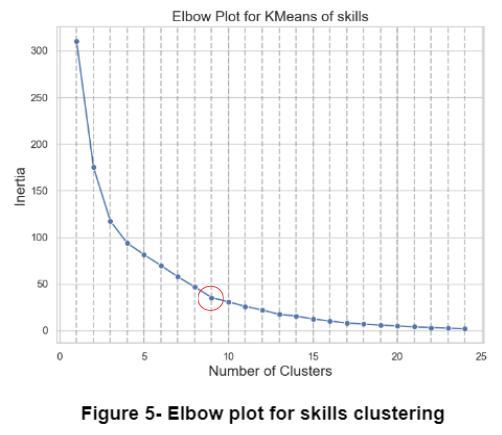
The dataframe compiled from the binary indicators of data science-specific skill occurrences was then used to construct a hierarchical clustering of skill sets. The hierarchical clustering results reveal distinct clusters of skills with similar characteristics. In Cluster 1, which encompasses PyTorch, TensorFlow, Deep Learning, and Artificial Intelligence, the course will focus on advanced machine learning and AI techniques, exploring the latest developments in neural network frameworks. Cluster 2, featuring AWS, Spark, R, and Statistical Analysis, suggests a concentration on cloud computing, big data processing, and advanced statistical modeling. Cluster 3, comprising Business Intelligence, Consulting, and Tableau, indicates a focus on data visualization, strategic decision-making, and communication skills essential for consultancy roles.Moving to Cluster 4, including Data Mining, Hadoop, and Big Data, the course will focus on large-scale data processing, data mining techniques, and the management of big data infrastructures. Cluster 5, with Numpy, Pandas, and Scikit Learn, underscores the importance of foundational data analysis tools, making it an ideal module for mastering Python-based data manipulation and machine learning libraries. In Cluster 6, which includes Power BI, SAS, SPSS, Negotiation, and Excel, the curriculum will cover a diverse set of skills, from data visualization and statistical analysis to negotiation and spreadsheet proficiency, preparing students for roles that demand a versatile skill set. The remaining clusters emphasize skills related to database management and optimization (Cluster 7), advanced modeling and machine learning with Python (Cluster 8), and soft skills such as creativity, communication, and teamwork (Cluster 9). This course design aligns with industry trends and equips students with a comprehensive skill set tailored to various roles within the data science and artificial intelligence domain.

**Feature Engineering for KMeans Clustering**

A new dataset was compiled with rows indicating skills, and a total of 10 columns as 10 features that explain the skills within the context of the web-scraped job postings. The first feature used the clustering of word embeddings from the job description via KMeans clustering. Word2vec was used to extract word embedding vectors. The optimal number of clusters, based on the elbow point in Figure 4, was determined as 8. Subsequently, the number of data points in each cluster for each skill was calculated, and the cluster with the maximum number of data points for each skill was extracted. Additional features comprising of skill frequency, mean salary, count of bachelor’s degree, count of Master’s degree, count of PHD, average years of experience required, the average length of job title, and count of managerial jobs for each skill were calculated, and compiled in the columns as features for the identified skills.

**KMeans Clustering on skills**

Similar to the KMeans Clustering of word embeddings, the dataframe of features was fed into the KMeans algorithm, and the optimal number of clusters, as seen in Figure 5, was determined as 9. In order to visualize the clusters, pca was used to reduce the number of features in the dataset to 2. In Cluster 1, which currently includes SPSS, SAS, and Excel, the course will emphasize proficiency in statistical software and spreadsheet tools, while accommodating negotiation skills (which was initially placed alone in a different cluster), which align more closely with the analytical focus of this cluster. Cluster 2, encompassing SQL, optimization, and data management, will form the foundation for a module concentrating on comprehensive database management, optimization techniques, and effective data handling. Cluster 3, housing Python, machine learning, and modeling skills, emphasizes programming expertise, machine learning algorithms, and advanced modeling techniques. Statistical analysis, R, Numpy, Tableau, and Power BI in Cluster 4 will be explored extensively to cover a broad spectrum of statistical tools, data visualization, and business intelligence. Soft skills, including Teamwork, Creativity, and Communication, combined in Cluster 5, will be an integral part of the curriculum to enhance collaborative and communicative abilities. Cluster 6 focuses on cutting-edge technologies like Deep Learning, Artificial Intelligence, Tensor Flow, and Pytorch. Negotiation skills, initially placed in Cluster 8, will be seamlessly integrated into Cluster 1, aligning more cohesively with the analytical and statistical focus. Similarly, Consulting skills, currently the only skill in Cluster 10, will be relocated to Cluster 9 to harmonize with the comprehensive toolkit of Pandas, Spark, Hadoop, Big Data, Business Intelligence, Data Mining, Scikit-learn, and AWS.



**Final Course Curriculum**

The KMeans clustering amalgamated diverse skills like Pandas, Spark, Hadoop, Big Data, Business Intelligence, Data Mining, Scikit-learn, and AWS into a single cluster, neglecting the distinct domains within data science that these skills represent. Furthermore, negotiation and consulting skills were isolated into separate clusters, overlooking their potential integration into core courses. In contrast, the Hierarchical clustering approach produced more cohesive results, effectively segmenting the skills essential for success in data science roles. While both clustering methods yielded similar outcomes in certain areas, the Hierarchical clustering approach demonstrates a superior ability to delineate and organize the multifaceted skills required for data science job applicants. Therefore, I advocate for the course designs derived from the Hierarchical clustering section as the preferred and more comprehensive proposal.