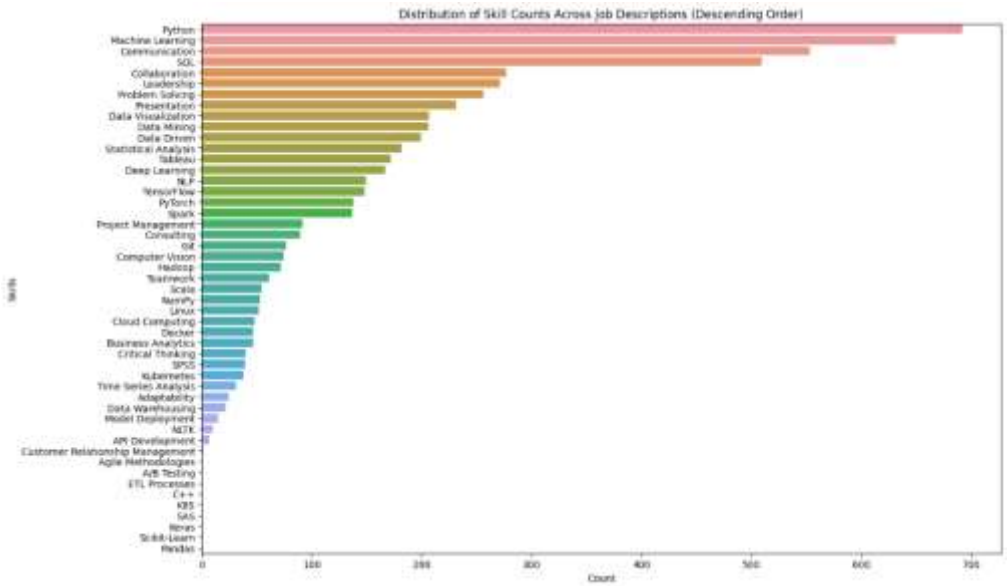
Assignment 3 Report

Data collection and Cleaning

For this part, I used Undetected Chromedriver for web scraping because it is optimized for bot detection bypassing so that it won't be identified by Cloudlfare.
For location and title, I chose "united states" and "data scientist" because the size and diversity of the search result for this combination would be better.
The data collected includes "Title", "Company", "Location", "Rating", "Date", "Salary", "Description", "Links". Among those field, "Rating" is not being fetched from the website and all the results are NaN, thus dropped. "Location" and "Salary" are parsed and converted to city name and average salary. Other data cleaning methods for "Description" include lemmatization, stop words removal, tabulation, and punctuation removal, etc. This prepares the job description ready for key words extraction using N-gram.

Exploratory data analysis and feature engineering

For feature engineering, OpenAI API is utilized to generate a list of skills that are required for data scientists. The result is manually adjusted before getting tokenized as bi-grams and tri-grams. And then for each key word, it will be matched against the set of n-grams for every job description. If the keyword is present in the set, label the skill as 1, otherwise label it as 0. In this way, we get a list of features which contains binary integer results.
After visualizing the distribution of skill counts, we get the following graph as an insight:

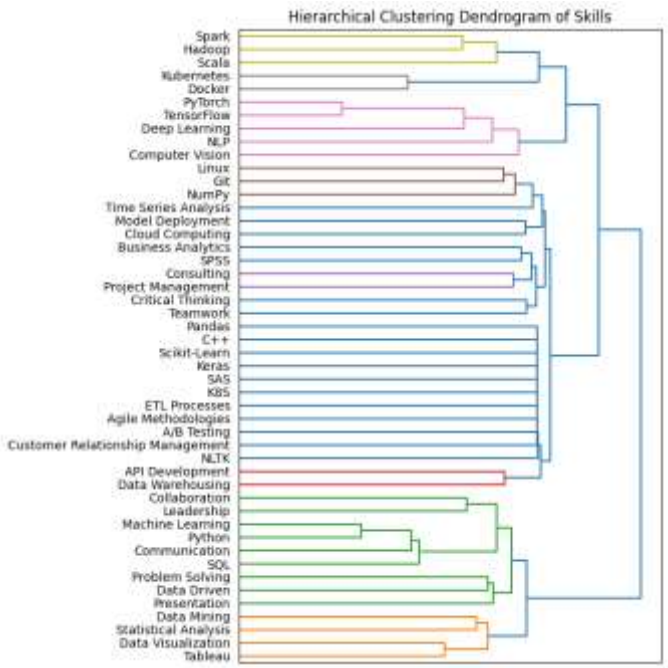


As we can see from the graph, Python is the most popular skill demanded, as it is now the "default" programming language for data science and machine learning. Surprisingly,

“Communication” is ranked top 3, which tells us that data scientists should also develop their soft skills.

Hierarchical clustering implementation

When it comes to hierarchical clustering, the choice of methods for describing similarity is important. Since the data for features are binary (0 for not present, 1 for present), I used Jaccard distance for building the distance matrix. It measures the dissimilarity between sets by comparing the size of the intersection with the size of the union of the sample sets. In the context of skills data, where each skill can either be present (1) or absent (0) in a job description, Jaccard distance effectively captures the similarity between different skills based on their co-occurrence across job descriptions.



Three methods are chose to generate the dendrogram: complete, average and centroid. Among these three results, the complete method does its job, because the clustering is more diverse and the relationship among the cluster members are more obvious. This graph is later used for distance level selection. After experimenting with the max_d value, max_d = 1.42 cut the dendrogram just fine and create eight clusters which are used in our course curriculum.

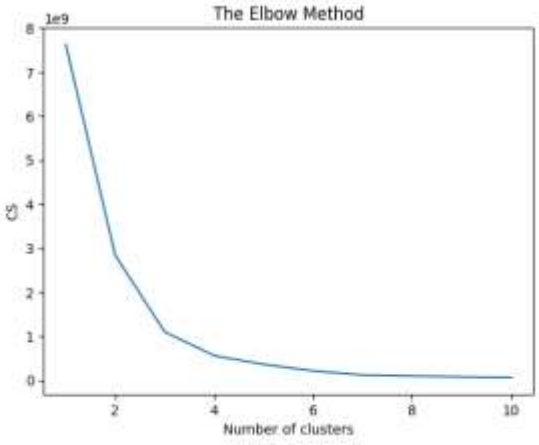
K-means clustering implementation.

The most challenging part for this is feature engineering. I came up with 11 unique features (including the text embedding from part 1). The features created are:

Skill Frequency	Count the number of times each skill appears across all job postings.
Average Salary for Skill	Calculate the average salary for job postings that mention each skill.
Skill Category (Soft/Hard)	Classify each skill as a soft skill or a hard skill.
Geographic Prevalence	Determine the frequency of each skill in different geographic locations.
Job Title Seniority Level	Categorize job postings based on the seniority level indicated in the job title.

Skill Relation to Remote Work	Create a feature indicating whether a job is remote or not.
Length of Job Description	Calculate the length of each job description.
Text Embedding Aggregated	Aggregate the text embeddings for each skill across all job postings using weighted sum

- The seniority level expands to four categories: Junior, Senior, Management and Mid Level
- Text Embedding vectors are added and averaged for each skill category.



Then by using the elbow method, we find the kink at k=3, it means from that value on, the marginal benefit is very small for reducing errors, so we choose k=3 for k-means clustering.

Course Curriculum

Here is the course curriculum in order:

Course 1: Foundations of Data Science and Business Leadership:
Course 2: Big Data Technologies:
Course 3: Data Warehousing and Integration:
Course 4: Data Mining and Visualization:
Course 5: Business Analytics and Consulting Skills:
Course 6: Cloud Computing and Model Deployment:
Course 7: Data Science Tools and Time Series Analysis:
Course 8: Advanced Deep Learning and AI:
Course 9: Advanced Data Science Programming and Tools:

Students will learn advanced programming skills in C++ and Python, with a focus on libraries such as Pandas, Scikit-Learn, NLTK, and Keras for data manipulation, machine learning, natural language processing, and deep learning. The course also introduces SAS for analytics, Agile methodologies for project management, A/B testing for hypothesis evaluation, and CRM (Customer Relationship Management) systems.

The course sequence is based on 1. Difficulty 2. Prerequisites for each course. For example, course 1 comes with no prerequisites and is beginner friendly. Whereas course 9 requires students to have a deep understanding of data science and technical tools.