Skills commonly required for datas analyst and data scientists are manually defined based on my own knowledge and combined with the skills generated by ChatGPT. The combined list can be found in the Jupyter Notebook. I then cleaned the combined skill dataset and removed the common stop words using the NLTK package. With N-grams from NLTK, single-gram, bigrams and trigrams were generated. Comparing the results, bigrams is a more suitable option for the skills dataset.

For visualization, I first took the job posting dataset and plot the frequency of different skills as illustrated in Figure 1 (Appendix). As we can see, R has the highest frequency count, following by Python and SQL and communication. Therefore, for the course design, R, Python and communication should be prioritized for people want a career in data analysis and data science. On the contrary, project management, MATLAB and SPSS have very low frequency, therefore, they will not be priorized in course design later.

Next, I plotted the 'Location' column of the job posting using word cloud. WordCloud is a text visualization tool for quick identifying the frequent used words in the text. The bigger the word is, the more frequent is present in the text. As we can see in Figure 2, 'Remote' is the most frequent word which is expected, as the dataset is for remote jobs. Some states or city are frequently present in the job postings, such as San Francisco, New York, Boston, etc.

Next, I created a WordCloud of the 'Description' column in the job posting dataset. As shown in Figure 3, it demonstrates the frequency of different skills. For example, machine learning, visualization, model, research, are among the most frequently required skills.

To implement hierarchical clustering algorithm, I first created a squareform distance matrix of all the skills. The dendrogram was created based on the distance matrix with a color threshold of 100 (Figure 4). Based on the clustering result, a course curriculum is suggested as below:

1. Introduction to R (skills to focus: into to programming language, data cleaning, data manipulation)
    - R, Python and SQL and closely together on the Dentrogram, and based on Figure 1, R has the highest frequency in the job postings, therefore, I believe it is essential to build the fundation of R programming first.
2. Introduction to Python (skills to focus: Python programming language, data cleaning and manipulation, data visualization)
    - R, Python and SQL and closely together on the Dentrogram, and based on Figure 1, Python has the second highest frequency in the job postings.
3. Introduction to SQL and database (skills to focus: data query, data access control and data manipulation)
    - R, Python and SQL and closely together on the Dentrogram, and based on Figure 1, SQL is the third highest frequency in the job postings.
    - SQL focuses on database, which is very different from R and Python.
4. Data visualization (skills to focus: data analysis, design principles, effective story telling and communication)
    - Visualization is a form of communication that is essential for any data related career. It can help us effectively summarize and communication the insights and drive business related decision making. I believe it is such a critical skill that an entire course should be dedicated to this topic. In addition to the

programming skills and the data analysis skill, it is also very important to get a good understanding of the design principles, good visualization should be accessible, easy to understand, aesthetically pleasing, and it can also guide attention to important aspects of the data/insights.
    - Software in scope of this course: Python/R, Tableau
5. Data analysis for project manager (skills to focus: PowerBI, Excel, project management, presentation, communication)
    - This course is designed for people that wants to work in the tech/data industry and are enthusiastic about project management and communication.
    - This course focuses on the technical skills of a project management such as data analysis and visualization, as well as the soft skills
    - This course has no prerequisite of programming, and this course will not focus/require any programming skills
6. Indroduction to deep learning (skills: machine learning, deep learning, Pytorch, Spark)
    - This course is designed for people that are interested in deep learning, Spark, Pytorch or Tensorflow will be introduced during the course
7. Statistics and data science (skills: statistic focused data analysis and data science)
    - This course focused on the mathematically models and SAS application for research
8. Leadership in data science (skills: leadership, management, communication)

In the next part, 11 features were developed (including the distance matrix from the hierarchical clustering), the dataframe of the 10 features can be found in the jupyter notebook. Elbow method was used to determine the optimal number of clusters to perform kmean clustering. A total of 7 clusters were then generated for the skills dataset. To visualize the 7 clusters in a 2D scattered plot, PCA was applied to reduce the dimensionality to 2 before plotting. The scattered plot can be found in Figure 5. To understand what is in each cluster, I summed up the frequency of each skill mentioned in each cluster, and ranked them in descending orders. A course curriculum is designed based on the clustering results:

1. Fundamental of data science (R, Python, statistics)
2. Advanced data analysis (R, SQL, statistics)
3. Thesis: Imaging processing and machine learning (Python, deep learning)
4. Seminar: Consulting in data engineering (project based, consulting, leadership, communication, project management, visualization)
5. Data structure and big data (SQL, SAS)
6. Data mining (Java, Python)
7. Machine learning in health care (Python, Tableau, deep learning, visualization)
8. Machine learning in finance (C, PowerBI)

Both methods are great ways to cluster skills and use the clustering analysis for curriculum design. Hierarchical clustering is easier to understand visually, while k-mean clustering requires further data processing steps such as PCA for dimensionally reduction before we can visualize the results.

Appendix. Visualization

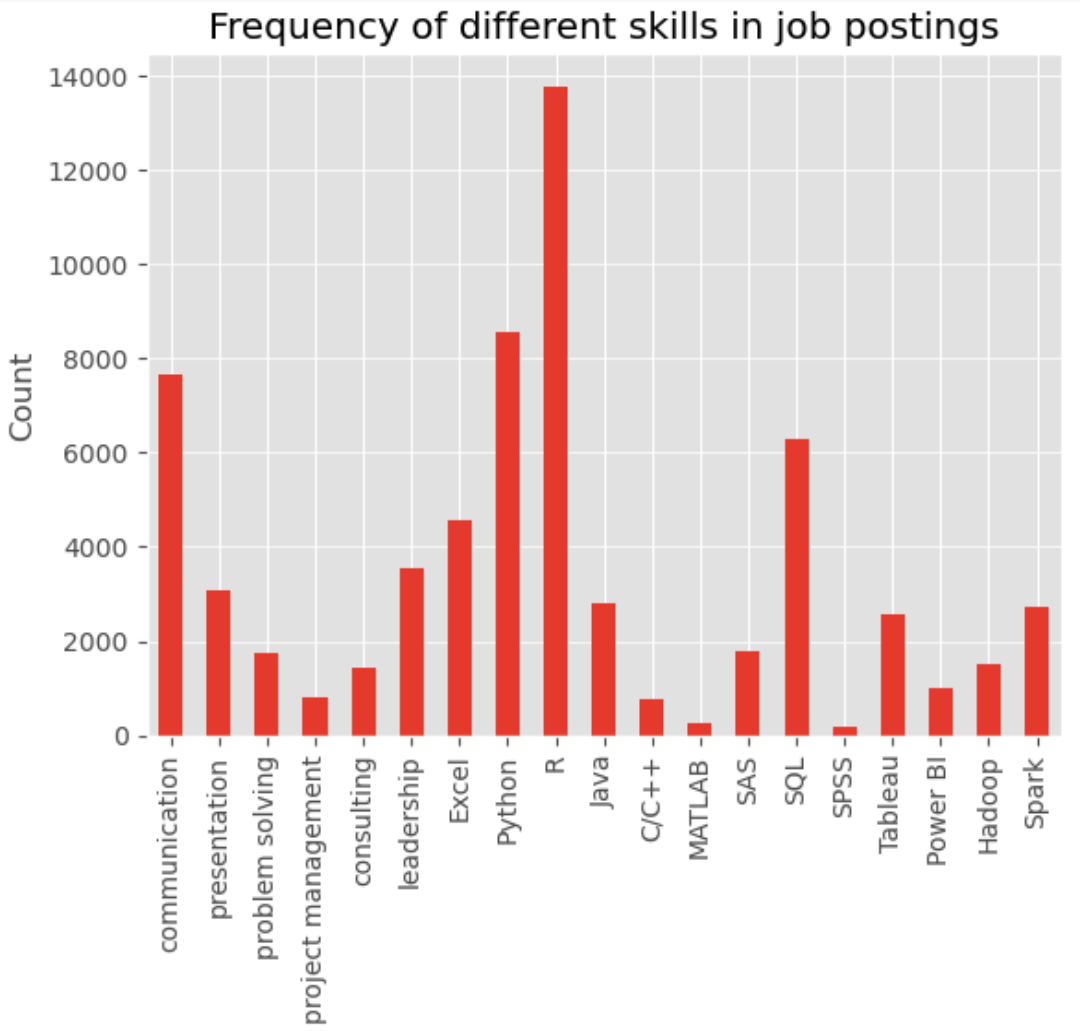

Figure 1. Frequency of different skillsin job postings.

Figure 2. WordCloud of ‘Location’ column in job posting dataset

Figure 3. WordCloud of ‘Description’ column in job posting dataset

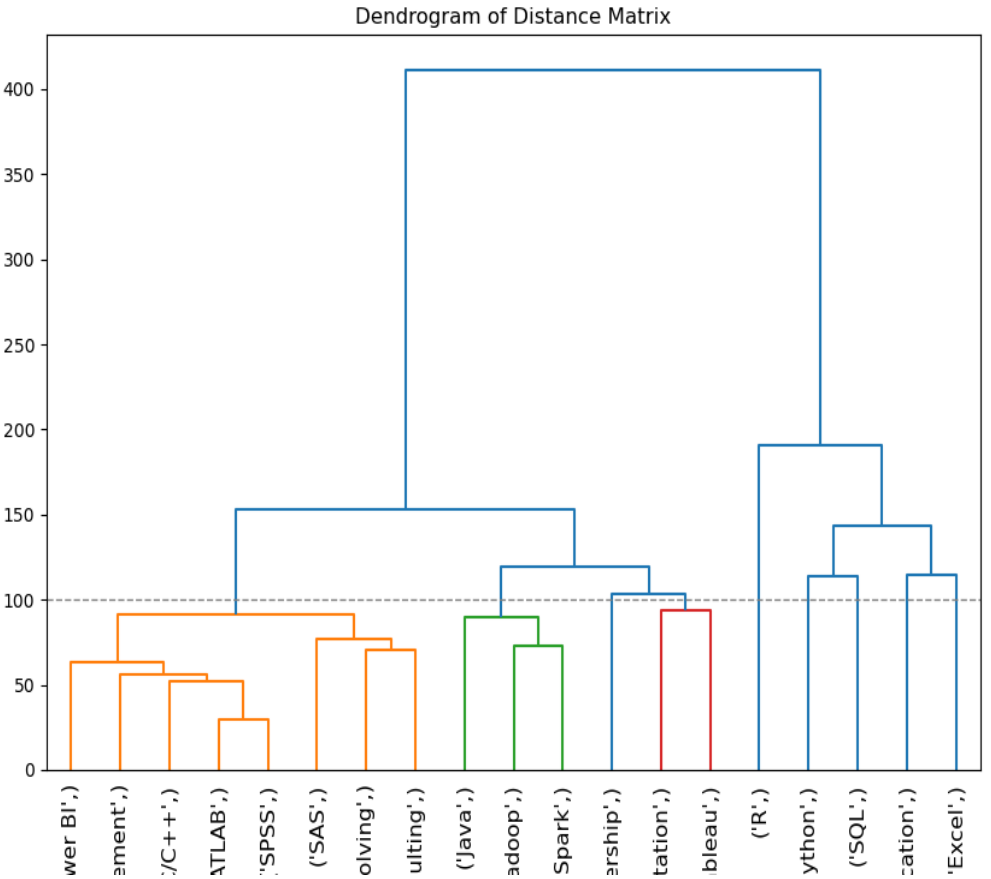

Figure 4. Dentrogram of different skills based on distance matrix.

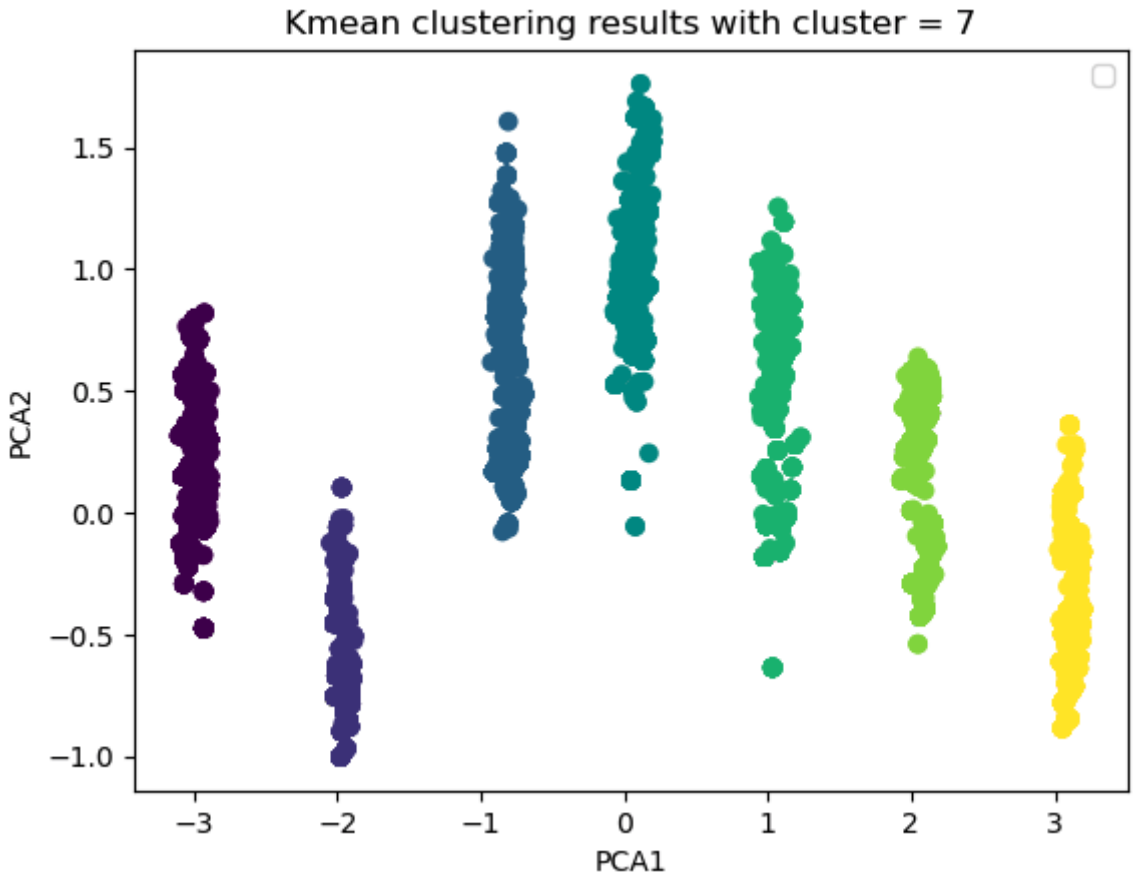

Figure 5. Kmeam clustering results with PCA = 2