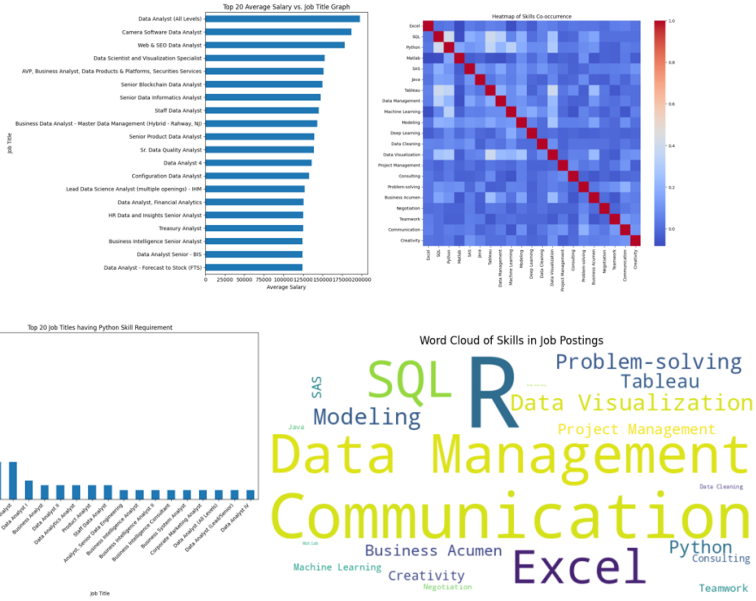
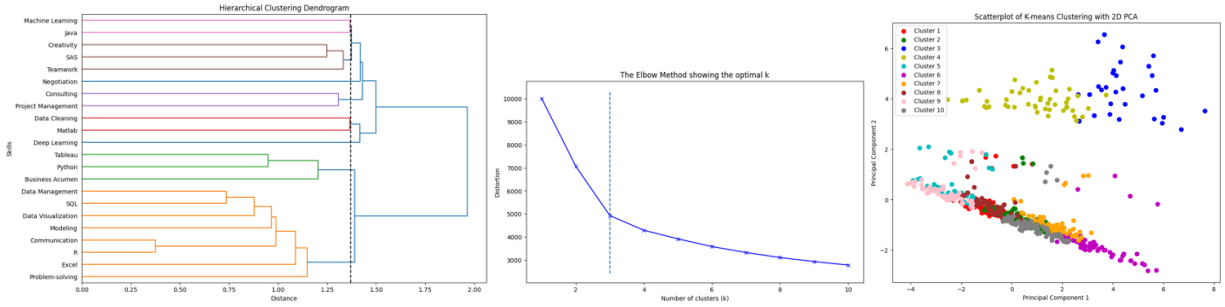
In this project, since I need to design a course curriculum for a new Master of Business and Management in Data Science and Artificial Intelligence program to help the students learn the crucial skills which can help them have a successful career in doing Data Science work, I decided to collect the data from the Indeed website to collect the job information in order to know which kind of skills are required for the job market regarding Data Science to design the courses which can help the students to achieve their goals. Therefore, I decided to search the job position which is data analyst in the USA. After collecting 1000 job posting information and removing all the duplicate values from the Indeed website, I used ChatGPT API to generate the skills which may highly required for the Data Science job which are technical skills include [“Excel”, “SQL”, “Python”, “R”, “Matlab”, “SAS”, “Java”, “Tableau”] and soft skills include [“Data Management”, “Machine Learning”, “Modelling”, “Deep Learning”, “Data Cleaning”, “Data Visualization”, “Project Management”, “Negotiation”, “Teamwork”, “Communication”, “Creatively”]. I used “0” to represent selected job does not require these skills and used “1” to represent selected job requires these skills.



After getting the modified dataset, the first thing I want to find out is which job title has the highest average salary. From the first graph above, we can see that the Data Analyst (All levels) has the highest annual salary which is nearly $200,000 which means that we should design our courses more related to this job title which can help our students to be more successful in the future. And then, the next thing I am going to find out is which kind of skills have more frequency appear together. From the heatmap above, the red squares indicate a higher frequency of co-occurrence and the blue squares indicate a lower frequency of co-occurrence. For example, the strong co-occurrence between “Python” and “Machine Learning” highlights the importance of Python programming skills for machine learning roles. This analysis helps in understanding the skill sets that are complementary and frequently sought together in the data

science field. Also, I am interested in knowing which kinds of jobs have the most demand for "Python" which currently is the most common coding technique. From the third graph above, we can see that Data Analyst has the highest Python demand which means that we should design the course which will introduce Python for the students who want to be a Data Analyst in the future. Lastly, in order to find which kinds of jobs have the most demand in job postings, I generated the word cloud. From the word cloud above, we can see that text like "Data Management", and "Communication" have the largest size which means that these kinds of jobs have the highest demand in the job market which shows that we should take more attention to design the course to teach these kinds of skills.



After that, we are going to use the hierarchical clustering method to find which skills are more similar and closer linked to each other in order to design the course for our students to take. However, I found that some clusters' distance is very far away from the others, so I think I can make little adjustments to my course curriculum based on the output which lets the skills which have the closer distance to become into one course (more similar to each other) after implementing hierarchical clustering algorithm. Also, since I only implemented 22 skills, it means that there should only be two skills covered in two of the designed courses. Under these conditions, we can finally get 8 courses after the little adjustment based on the output dendrogram above which are [Course 1: Machine Learning, Java], [Course 2: Creativity, SAS, Teamwork], [Course 3: Negotiation, Consulting, Project Management], [Course 4: Data Cleaning, Matlab, Deep Learning], [Course 5: Tableau, Python, Business Acumen], [Course 6: Data Management, SQL, Data Visualization], [Course 7: Communication, R], and [Course 8: Modeling, Excel, Problem-solving].

And then, we use the 10 features regarding job postings in which we are interested which are ["Skill Frequency", "Average Salary", "Job Posting Descriptions Length", "Technical Skills Count", "Soft Skills Count", "Presence in High Salary Jobs", "Presence in Job Title", "Skill Diversity", "Skill Correlation with Salary", "Company Demand for Skill"] to do the k-means clustering to help us to design the course. After using the elbow method to determine the optimal k number of clusters, we can get the optimal number of k is 3 from the output graph above. However, after doing the k-means clustering, we only can get two courses which is not enough for designing the course curriculum. Therefore, I applied adjustment to my cluster which changed the value

for k to 10 instead of 3. By applying this adjustment, I can design 10 courses which are [Course 1: R, Communication, Data Management], [Course 2: R, Communication, SQL], [Course 3: R, SQL, Data Visualization], [Course 4: R, Communication, Modeling], [Course 5: R, Communication, Data Management], [Course 6: R, Communication, SQL], [Course 7: R, Communication, Data Management], [Course 8: R, Communication, Problem-solving], [Course 9: R, Data Management, SQL], and [Course 10: R, Communication, Data Management]. However, since the dimension for our output is quite large, as the scatterplot below, we use the PCA method to reduce the dimensions into two dimensions with each point representing a job posting and its placement being determined by the principal components that capture the most variance in the dataset which are shown in the scatterplot above.

Based on my clustering results, I wrote two prompts to ChatGPT API. My first prompt let the ChatGPT successfully underscores my program's unique approach, blending rigorous technical training in R, SQL, and data management with crucial soft skills like communication and problem-solving. And then, my second prompt let the ChatGPT analysis further delineate my program's courses into clusters, revealing a balanced emphasis on foundational technical skills and the integration of soft skills. Overall, I think that ChatGPT's interpretation successfully enriches my curriculum's description and makes my curriculum to be attractive option for prospective students.

I decided to use the course curriculum which I got by using the hierarchical clustering method as my final course curriculum to address the multifaceted nature of the data science filed. I chose this approach because this curriculum meticulously balances a spectrum of analytical skills, with a focus on programming languages such as R, SQL, and Python, essential for data manipulation and analysis. Also, this technical foundation is complemented by a strong emphasis on soft skills, notably communication and negotiation, equipping students to effectively articulate insights and collaborate across diverse teams. The inclusion of advanced topics like machine learning and deep learning alongside fundamental data management practices ensures graduates are not just prepared for current industry standards but are also forward-looking, capable of adapting to and innovating in a rapidly evolving technological landscape. Therefore, I think that this curriculum stands as a comprehensive blueprint that prepares students for a successful career in data science, bridging the gap between theoretical knowledge and its practical application in solving real-world problems.

Lastly, after creating text embeddings of the job descriptions by using the ChatGPT API, I used embeddings to perform clustering by using the k-means clustering method with the 10 for k value. The new course curriculum got by this new method is shown in my code part and which looks similar with the course curriculum generated by the original k-means clustering method.