Assignment 3

MIE1624 REPORT

Tongfei Li

1004759460

Part1
For the very first part of data collection and cleaning, I modified the given python codes, set the location to remote, and collected job postings of unique positions in Data Analyst (538 items) and Data Engineer (512 items), regardless of salary and company names.
Secondly, on my main ipynb file, I read the job postings data of both positions and combine them into one data frame.

Part2 (Two visual depictions graph in appendix)
On open AI website, I generated my API to access ChatGPT, and I used model 3.5 turbo to create conversation with ChatGPT and record the response in a list, ChatGPT response python list is in appendix. I combined gpt list with manually created list to extract skills as columns.
In Data Visualization, I created two bar plots, one is to visualize skills frequency of position data analyst and engineer respectfully. I extract salary of non-null observations, and calculated average salary of the position that requires specific skills and plot it in the second graph.
From the first plot, we can see that, the skills of communication, SQL, database, python, excel, report, AI, teamwork, and business are highly required in these positions, while the distribution is different. Analyst requires the skills of communication, excel and report much more frequently, and engineer requires more of machine learning, SQL, database, python, AI, cloud, and software. I would say that the role of data analyst is more like a storyteller of data, while data engineer is more programming based, it requires more technical skills in data manipulation.
Of all the job postings, the position that requires skills in statistical analysis, R programming, machine learning, data modelling, and data science seems to have higher salaries, while the positions that require skills related to spreadsheets, meeting, share point and power point seems to have lower average salary.

Part3 (clustering result and course curriculum in appendix)
To generate the distance matrix between skills, I calculate tuple wise distance between each pair of skills. Here, I used Euclidean as distance measurement to calculate pair-wise skills distances. For each two vectors of skills (each index is if the skill required in specific job posting), if the two skills are required in many postings at the same time, the subtraction difference would contain many zeros, therefore make the Euclidean result to be small, and vice versa.

Based on Hierachical clustering result, I chosed 5 clusters with more than 3 skills, and created course curriculum based on skills in each cluster. Some of them are independent and some of them requires prerequisite course, there are 10 courses that can be finished in full-time 2 semesters based on UofT academic settings.
Course list:
Based on skills from red cluster: excel, communication, report, teams, table, database, application, software, decision, leader
Course 1: Introduction to excel and database applications
Course 2: Effective Communication and Teamwork for Leaders
Course 3: Decision-Making and Leadership for data analysis (prerequisite course 2)

Basesd on skills from blue cluster: team, AI, business, SQL
Course 4: SQL for AI and business intelligence

Based on skills from green cluster: finance, meeting, power bi, presentation, data visualization, data management, machine learning
Course 5: Introduction in machine learning
Course 6: Financial data management and visualization with machine learning (prerequisite course 5)

Based on skills from yellow cluster: Jupyter notebook, data cleaning, R programming, data manipulation, spreadsheet, detail oriented, figure
Course 7: Introduction to R and Python
Course 8: data manipulation and visualization with R studio and Jupyter notebook (prerequisite course 7)

Based on skills from grey cluster: artificial intelligence, sharepoint, visualize, statistical analysis, database management
Course 9: Advanced Database Management and AI Techniques (prerequisite course 4)
Course 10: Sharepoint and Database management for statistical analysis

Part4 (clustering results and course curriculum are in appendix)
To implement K-means clustering, I engineer 11 features in addition of the distance matrix. For each skill (as rows), I calculated average salary, binary indication of if they belong to soft or technical skill, frequency of their occurrence in job position DA and DE, binary indication of if they belong to programming skills, visualization skills, teamwork skills, or data processing skills. And binary indication of if they are related to desktop applications and finance. Each feature is put into columns. I standardized the data and fit it with k-means of 26 clusters, it results in 7 clusters contain 3-8 skills, and 19 clusters with less than 3 skills. I develop a course curriculum with 9 courses. Some of them are independent and some of them requires prerequisite course, the courses can be finished in full-time 2 semesters based on UofT academic settings.

I iteratively calculated k-means sum of squared distance within cluster for k number of clusters from 2 to 45, and I plot distance vs k to visualize the change in distance towards cluster number. Using the Elbow method, I would choose 20 clusters as optimized k number, so my optimal model is k-means with 20 clusters. Since the sum of distances does not decrease much when we continue increase k. (Elbow method plot in appendix)

Part5 (a, b, c) visualization plots are in appendix.
From the labeled scatter plot of k-means clusters, with one color per cluster. There are many clusters of points with same color, so our k-means model successfully identify some similarity among skills and group the skills that are closed to each other. But I would say that the model may be a little bit overfit, since there are still some points that are closed to each other with different colors.

Part 6

I would select my course curriculum in part 4 k-means model as my final course curriculum. The first reason is that k-means machine learning model is more flexible in building clusters in this case. Since we want to generate some specific number of clusters, with ideally 2-8 skills per cluster, so that we can design courses with the skills from one of the clusters. K-means model is sufficient in choosing number of clusters, to control the skills that are classified in the same group. When we want to see the result, we can easily print out the skills by choosing the cluster number. On the other hand, hierarchical clustering always results in one dendrogram, we have to manually choose the number of clusters from the graph and identify the skills that are sorted together. It is sometimes hard to identify the priority order that the skills are grouped, and it is possible that hierarchical clustering is putting too many skills together, which makes the topic of course vague and the content of each course messy.

Secondly, K-means clustering seems to do a better job in classifying related skills. For example, it puts programming related skills into cluster 21, machine learning modelling skills into cluster 15, data processing skills into cluster 23, presenting skills into cluster 18, visualization skills into cluster 22 and finance and business-related skills into cluster 19, it is easy to design courses based on skills with similarity, and there are outlined learning objectives for each course. On the other hand, hierarchical clustering does not have explicit characteristic among clusters, and similar skills seems to be separated into different group. For example, 'database', 'SQL', 'database management' are similar skills, but they are separated into different clusters. Additionally, visualization skills such as 'table', power bi', 'tableau' and 'figure' are far away from each other in dendrogram. Data related skills are also being put into different clusters. So hierarchical clustering has higher misclassification probability and it therefore makes curriculum design messy.

Part 7 (ChatGPT response in appendix)

Based on ChatGPT's answer from part 7, we have sufficient prove to say that k-means clustering on this dataset gives a better classification on the skills. Since ChatGPT's answer implies that clustering result of hierarchical clustering contains mixed of skills in different fields, while for k-means results, groups of skills are identified clearly. This is very similar to my discussion in part 6.

Appendix

ChatGPT response skill list:

```
: conversation = []
conversation.append({'role': 'system', 'content': 'Can you think of skills that are required for the position of Da
response = ChatGPT_conversation(conversation)
conversation.append({'role': response.choices[0].message.role, 'content': response.choices[0].message.content})
print('{0}: {1}\n'.format(conversation[-1]['role'].strip(), conversation[-1]['content'].strip()))
```

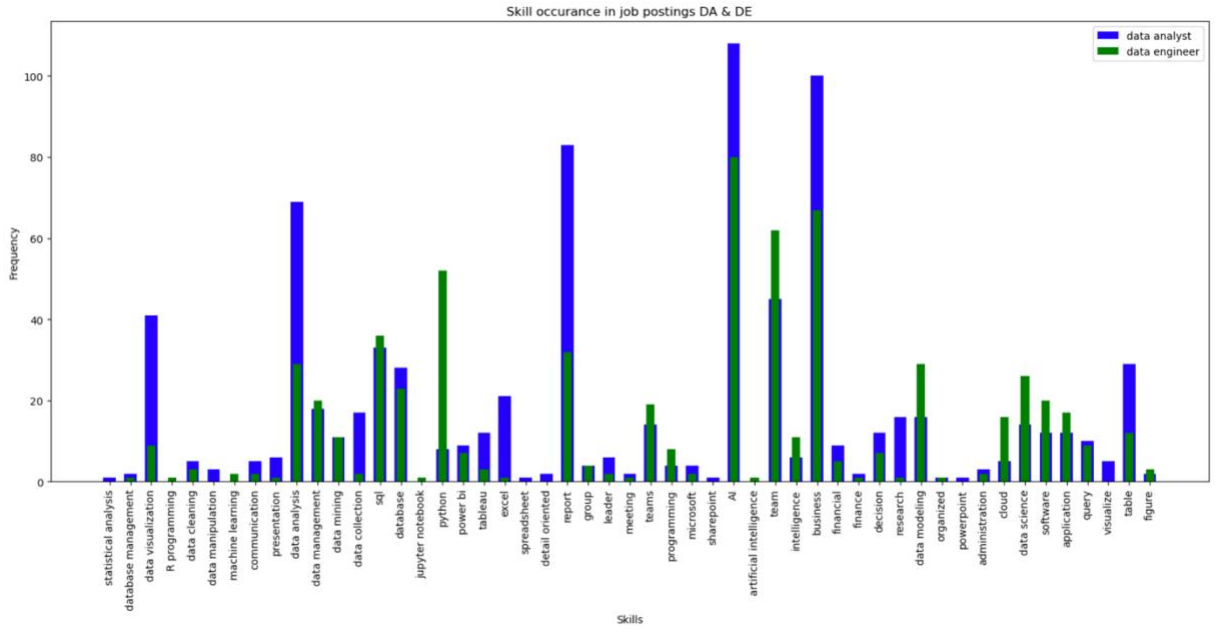
```
assistant: Some skills required for the position of Data Analyst may include:

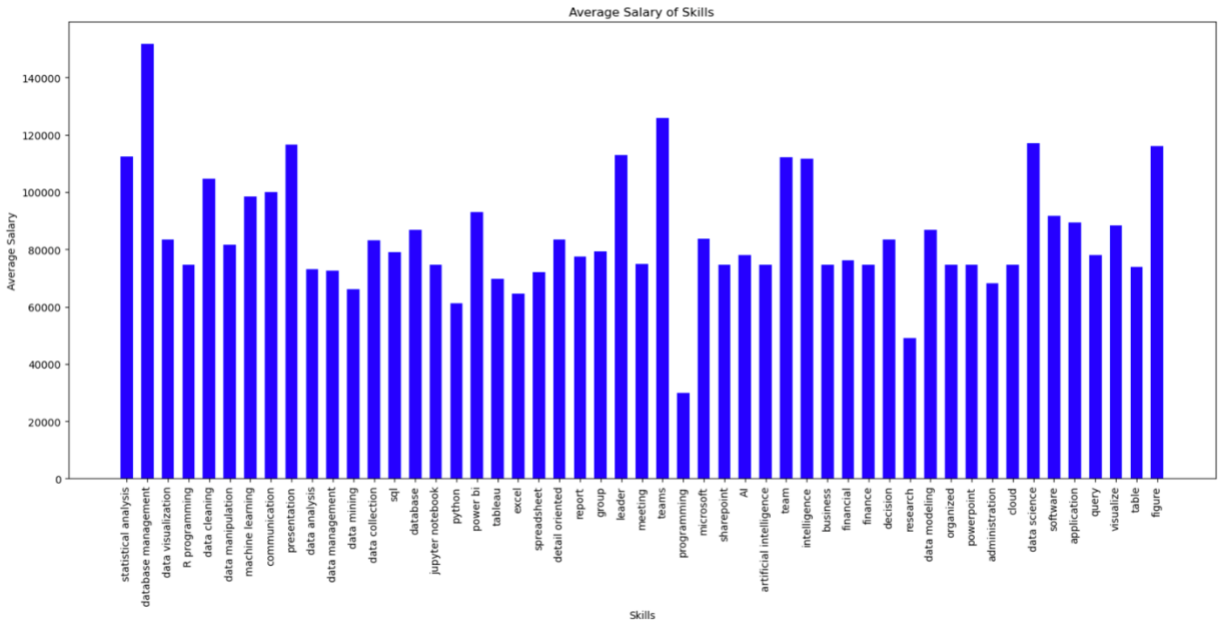
- Strong understanding of statistical analysis
- Knowledge of database management and querying
- Proficiency with data visualization tools
- Experience with programming languages such as Python, SQL, and R
- Ability to clean and manipulate data
- Understanding of machine learning concepts
- Effective communication and presentation skills

Here's a Python list that includes key phrases for these skills:

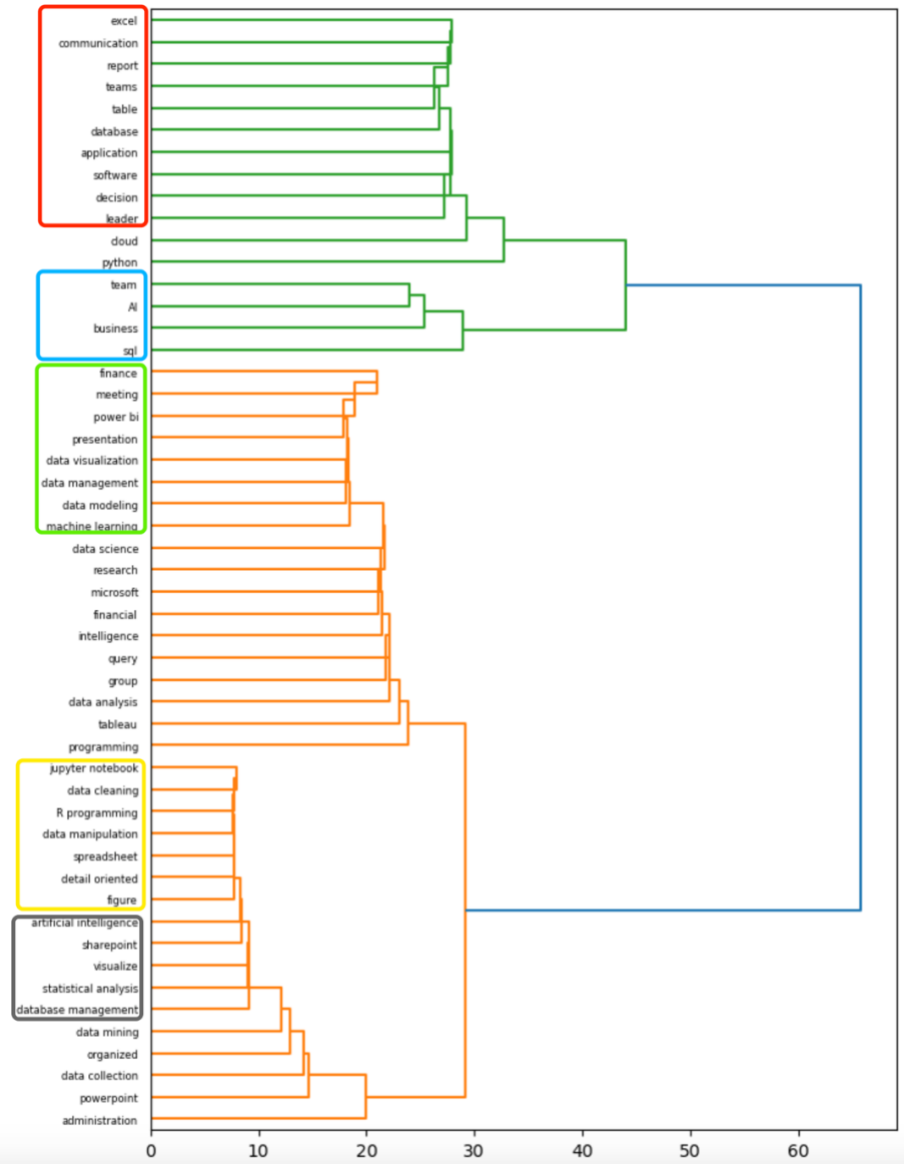
```python
skills = ['statistical analysis',
          'database management',
          'data querying',
          'data visualization',
          'Python programming',
          'SQL programming',
          'R programming',
          'data cleaning',
          'data manipulation',
          'machine learning',
          'communication',
          'presentation']
```
```

Two visual depictions

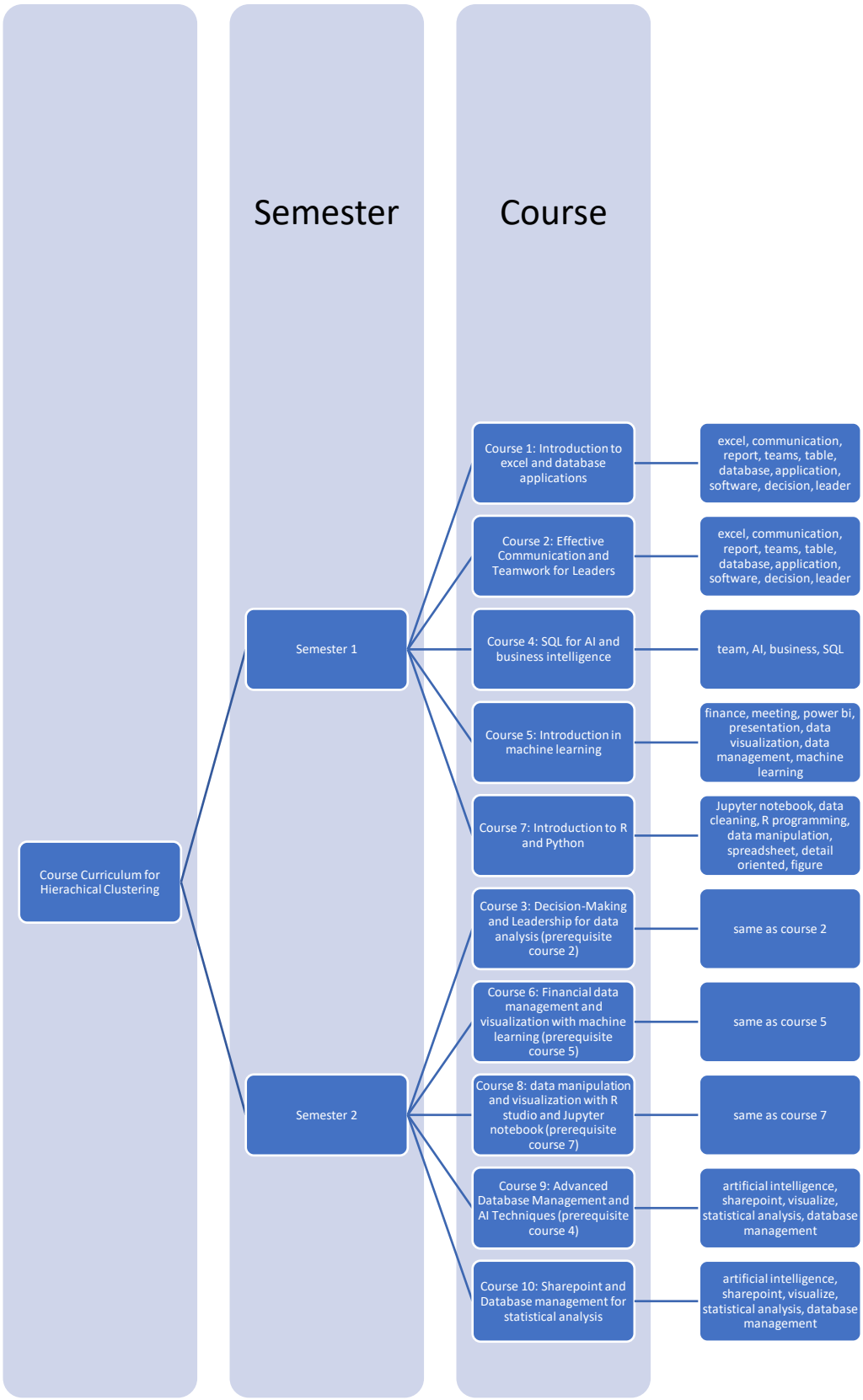




Clustering result for **hierarchical clustering**



Course curriculum for hierarchical clustering:


Semester
Course
Course Curriculum for Hierachical Clustering
Semester 1
Semester 2
Course 1: Introduction to excel and database applications
Course 2: Effective Communication and Teamwork for Leaders
Course 4: SQL for AI and business intelligence
Course 5: Introduction in machine learning
Course 7: Introduction to R and Python
Course 3: Decision-Making and Leadership for data analysis (prerequisite course 2)
Course 6: Financial data management and visualization with machine learning (prerequisite course 5)
Course 8: data manipulation and visualization with R studio and Jupyter notebook (prerequisite course 7)
Course 9: Advanced Database Management and AI Techniques (prerequisite course 4)
Course 10: Sharepoint and Database management for statistical analysis
excel, communication, report, teams, table, database, application, software, decision, leader
excel, communication, report, teams, table, database, application, software, decision, leader
team, AI, business, SQL
finance, meeting, power bi, presentation, data visualization, data management, machine learning
Jupyter notebook, data cleaning, R programming, data manipulation, spreadsheet, detail oriented, figure
same as course 2
same as course 5
same as course 7
artificial intelligence, sharepoint, visualize, statistical analysis, database management
artificial intelligence, sharepoint, visualize, statistical analysis, database management

K-means Clusters for courses

Course 1: Programming foundations in R and Python

Course 2: AI techniques using R studio and Jupyter Notebook (prerequisite: course 1)

```
kmean_label[kmean_label.label == 21]
```

	label
R programming	21
jupyter notebook	21
artificial intelligence	21

Course 3: Introduction to machine learning models (prerequisite: course 1)

```
kmean_label[kmean_label.label == 15]
```

	label
machine learning	15
data management	15
data modeling	15
data science	15

Course 4: Introduction to data collection and data manipulation in database

Course 5: Visualization of database management (prerequisite: course 4)

```
kmean_label[kmean_label.label == 23]
```

	label
statistical analysis	23
database management	23
data cleaning	23
data manipulation	23
data collection	23
visualize	23

Course 6: Teamwork for Data Analyst

```
kmean_label[kmean_label.label == 18]
```

	label
presentation	18
group	18
meeting	18

Course 7: Presentation of data mining results (prerequisite: course 6)

```
kmean_label[kmean_label.label == 1]
```

	label
data mining	1
spreadsheet	1
detail oriented	1
sharepoint	1
organized	1
powerpoint	1
figure	1

Course 8: Data visualization using power bi and tableau

```
kmean_label[kmean_label.label == 22]
```

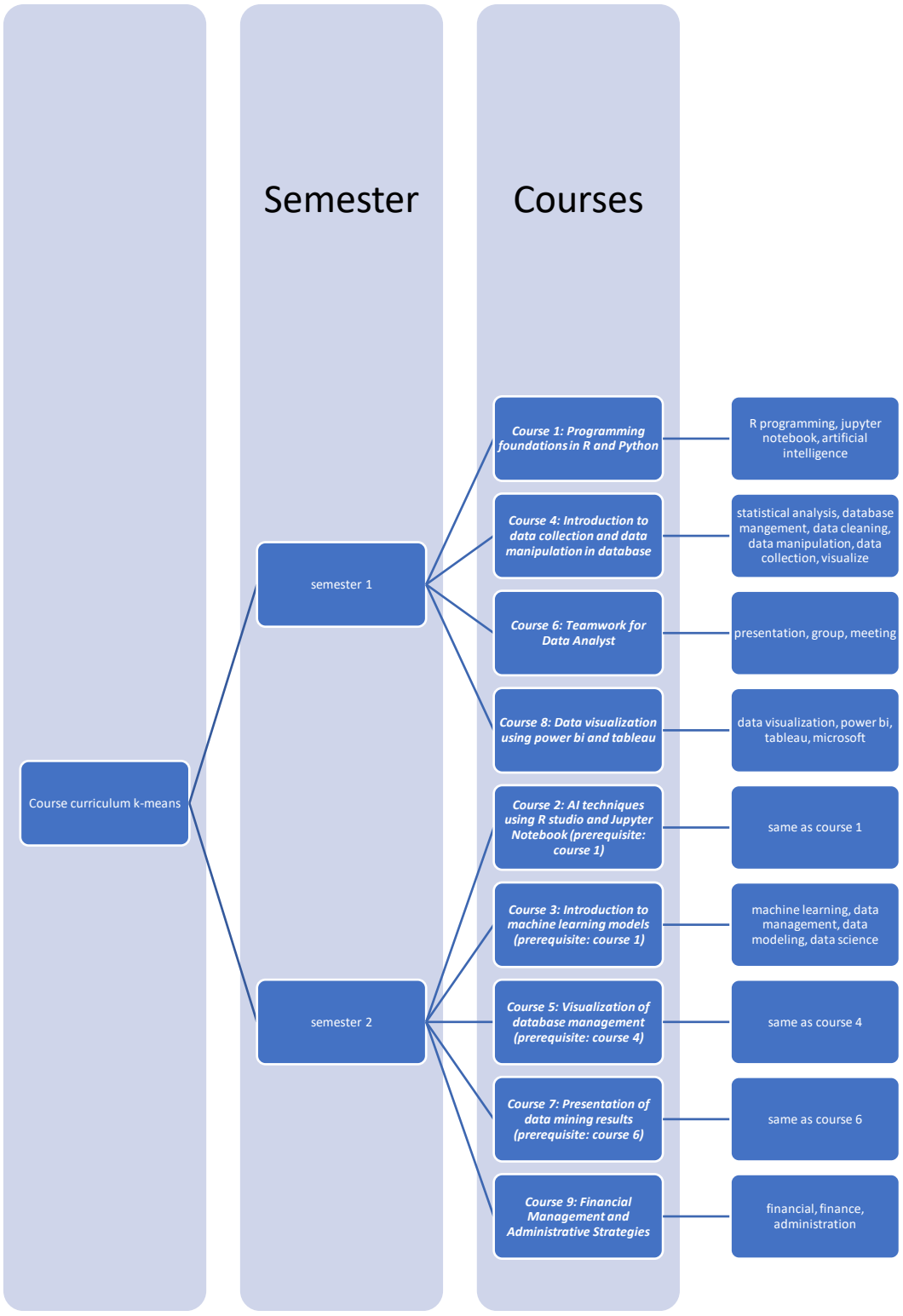
	label
data visualization	22
power bi	22
tableau	22
microsoft	22

Course 9: Financial Management and Administrative Strategies

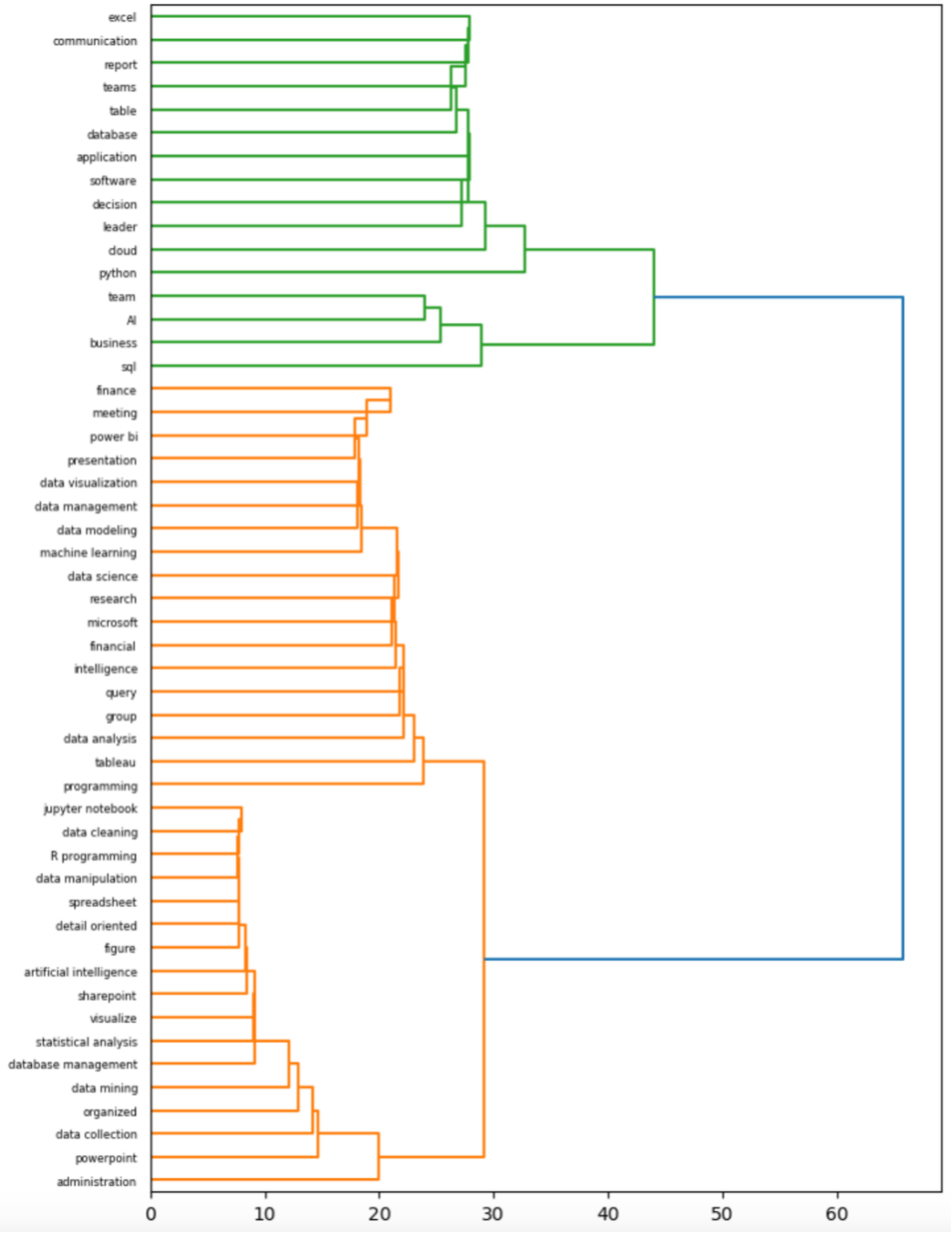
```
kmean_label[kmean_label.label == 19]
```

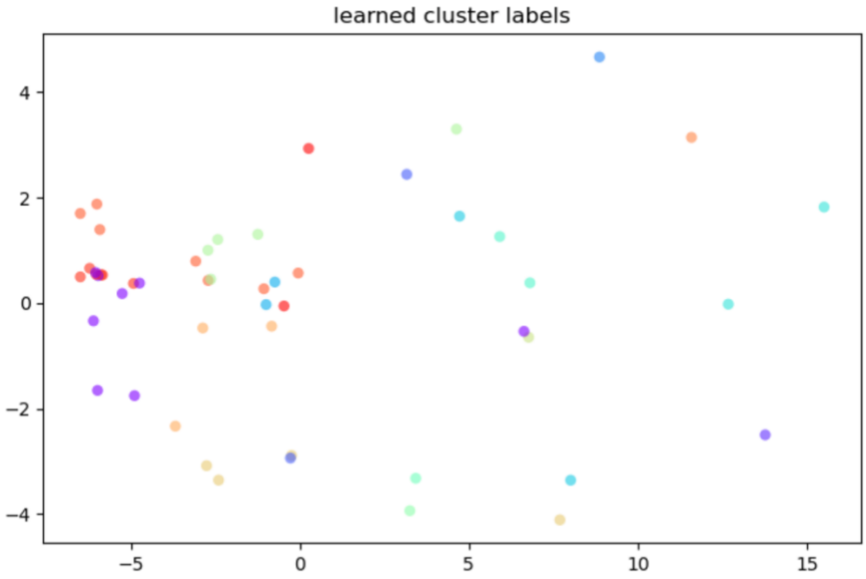
	label
financial	19
finance	19
administration	19

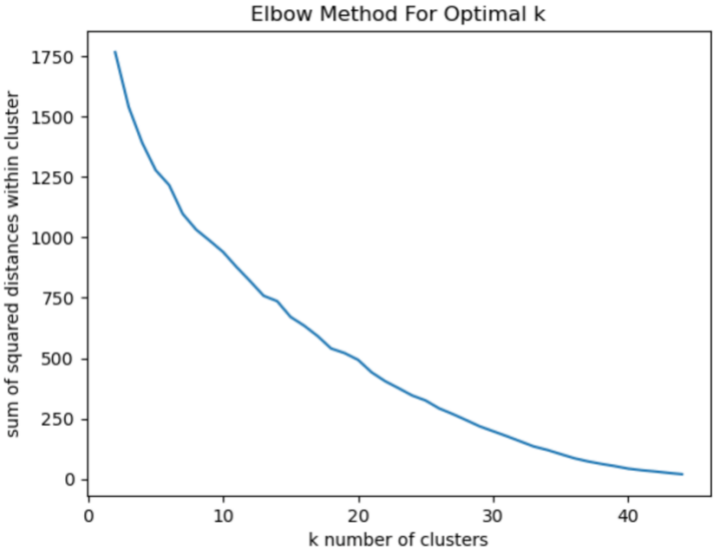
Course Curriculum for K-means


Semester
Courses
Course curriculum k-means
semester 1
semester 2
Course 1: Programming foundations in R and Python
R programming, jupyter notebook, artificial intelligence
Course 4: Introduction to data collection and data manipulation in database
statistical analysis, database mangement, data cleaning, data manipulation, data collection, visualize
Course 6: Teamwork for Data Analyst
presentation, group, meeting
Course 8: Data visualization using power bi and tableau
data visualization, power bi, tableau, microsoft
Course 2: AI techniques using R studio and Jupyter Notebook (prerequisite: course 1)
same as course 1
Course 3: Introduction to machine learning models (prerequisite: course 1)
machine learning, data management, data modeling, data science
Course 5: Visualization of database management (prerequisite: course 4)
same as course 4
Course 7: Presentation of data mining results (prerequisite: course 6)
same as course 6
Course 9: Financial Management and Administrative Strategies
financial, finance, administration

Part5 plots


excel
communication
report
teams
table
database
application
software
decision
leader
cloud
python
team
AI
business
sql
finance
meeting
power bi
presentation
data visualization
data management
data modeling
machine learning
data science
research
microsoft
financial
intelligence
query
group
data analysis
tableau
programming
jupyter notebook
data cleaning
R programming
data manipulation
spreadsheet
detail oriented
figure
artificial intelligence
sharepoint
visualize
statistical analysis
database management
data mining
organized
data collection
powerpoint
administration
0
10
20
30
40
50
60





Question to ask (hierarchical clustering):

```
This is my clustering result based on hierachical clustering, with some human soft skills and technical skills, cou
ld you please descripe my result and discuss what is common amongst the clusters.
 Cluster 1: excel, communication, report, teams, table, database, application, software, decision, leader
 Cluster 2: team, AI, business, SQL
 Cluster 3: finance, meeting, power bi, presentation, data visualization, data management, machine learning
 cluster 4: Jupyter notebook, data cleaning, R programming, data manipulation, spreadsheet, detail oriented, figure
 cluster 5: artificial intelligence, sharepoint, visualize, statistical analysis, database management
```

Answer by ChatGPT:

assistant: Based on your clustering result, it seems that there are some common themes among the clusters. Cluster 1 seems to be focused on general soft skills and technical skills that are useful in a variety of settings, such as communication, decision making, and database management. Cluster 2 appears to be focused on technical skills related to AI and business, such as SQL. Cluster 3 seems to be focused on technical skills related to finance and data management, such as data visualization and machine learning. Cluster 4 appears to be focused on technical skills related to data cleaning and manipulation, as well as being detail-oriented. Finally, cluster 5 appears to be focused on technical skills related to artificial intelligence and database management.

In general, it seems like there is a mix of both general soft skills and technical skills with a focus on data and technology. Additionally, there is a focus on skills related to AI and business, as well as data management and visualization. This suggests that a well-rounded set of skills is important for success in the data and technology field.

Question to ask (k-means):

This is my clustering result based on k-means, with some human soft skills and technical skills, could you please describe my result and discuss what is common amongst the clusters?
cluster 1: R programming, jupyter notebook, artificial intelligence.
Cluster 2: statistical analysis, database mangement, data cleaning, data manipulation, data collection, visualize.
Cluster 3: presentation, group, meeting.
Cluster 4: data visualization, power bi, tableau, microsoft.
Cluster 5: machine learning, data management, data modeling, data science.
Cluster 6: financial, finance, administration.

Answer by ChatGPT:

assistant: Based on the clustering result, it appears that there are six distinct groups of skills that have been identified. Within Cluster 1, there is a clear emphasis on technical programming skills related to artificial intelligence, specifically involving R programming and jupyter notebook. Cluster 2 involves a variety of data processing and management skills, ranging from statistical analysis to database management, data cleaning, data manipulation, data collection, and data visualization. Cluster 3 focuses on soft skills relating to presentations, group work, and meetings. Cluster 4 specifically relates to data visualization and data reporting tools, such as power bi, tableau, and microsoft. Cluster 5, similar to Cluster 1, emphasizes technical skills relating to machine learning, data management, data modeling, and data science. Finally, Cluster 6 is comprised of skills related to finance and administration.

There are some common themes that can be identified across the clusters. For example, there is a clear emphasis on both technical programming and data processing, specifically relating to artificial intelligence and machine learning in Clusters 1 and 5. Additionally, many of the skills in Clusters 2 and 4 specifically focus on data visualization and reporting. Finally, there is a clear distinction between soft skills and technical skills, with Cluster 3 specifically relating to the former, and Clusters 1, 2, 4, and 5 relating primarily to the latter.