MIE 1624: A3 Report

Date:	2023-04-06
Student:	William Hazen
Student ID:	1009231225
Institution:	University of Toronto
Faculty:	Department of Mechanical & Industrial Engineering

Objective

Design a course curriculum for a new "Master of Business and Management in Data Science and Artificial Intelligence" program at the University of Toronto with a focus not only on technical but also on business and soft skills.

1 Data collection and cleaning

For this assignment, data was collected from two versions of the job search website Indeed - *ca.indeed.com* and *www.indeed.com*. A modified web scraping code was used to extract job postings for the position of Data Scientist in the cities of Toronto, New York, and Boston. The use of *ca.indeed.com* was necessary to include Canadian-based jobs in the search. The rationale behind selecting these cities was to observe job opportunities in major cities on the east coast of North America. The number of unique jobs found in each location was the following,

- Toronto: 259 unique jobs
- New York: 469 unique jobs
- Boston: 406 unique jobs

A total of 1134 unique job postings were found and stored in the file *webscraping_ results_ assignmnet3.csv* - shown in Figure 1. The analysis was conducted using the following columns: job title, company, location, rating, date, salary, and job description. Data cleaning was performed on the job descriptions, including converting all text to lowercase and removing punctuation, to extract specific skills in a standardized manner.

2 Exploratory data analysis and feature engineering

For feature engineering, I started by manually creating a list of skills that are commonly associated with data scientists. This list included skills such as mathematics, programming, and business acumen. In addition to the manual list, I utilized the ChatGPT (GPT-3.5 Turbo) API to generate its own list of data science skills. These two lists were then combined, resulting in a set of 59 unique skills as shown in Figure 2. Initially, I attempted to use N-grams to identify if a job description contained skills from the unique list. However, since the unique list contained keywords as well as key phrases such as "AI" and "Machine Learning," using at least 2-grams was necessary. Unfortunately, this approach resulted in stopwords like "the," "their," and "and" being included in the N-gram data - shown in Figure 3. Consequently, I decided to try a different approach. I utilized the *str.contains* function in the Pandas library to determine if a skill - in the form of a string - was present in a job description. By doing so, I was able to vectorize the data and create a binary dummy column for each skill in our unique set - shown in Figure 4. With this vectorized data, the following plots can be shown in Figure 5 & 6. Figure 5 shows how frequently a skill shows in a job description, the results showed that the programming language *R* showed up 100% of the time in a job posting for data scientist, whereas *Python* showed up close to 70% of the time. Additionally, "AI" had around a 95% frequency rate and "machine learning" was over 60%. Moreover, looking at Figure 6, the average salary per skill is shown where we can see the most monetizable skills. The skill with the highest average salary is PyTorch with an average of $203199 and the second is TensorFlow with an average of $197762. Both are Python libraries that are widely used in the development of AI and deep learning applications, to which "deep learning" is the third most monetizable skill. It is clear that programming skills and knowledge in AI and deep learning applications are desirable to employers.

3 Hierarchical clustering implementation

Having a vectorized skill dataframe allows for the implementation of clustering in order to find similarities among the unique skills. I implemented a hierarchical clustering algorithm by computing the distance between each element to form a matrix. From here, I used SciPy to create the hierarchy to form the plot shown in Figure 7. In the figure, the orange vertical line acts as a cutoff for the clusters of around 8 where after some analysis, I group skills that were clustered to create the following courses.

1. **Introduction to Data Science** & AI (Skills: R, AI, Science)
2. **Introduction to Machine Learning** (Skills: Python, Machine Learning, Engineering, Excel, Communication)
3. **Math & Statistics for Data Science** (Skills: statistics, math, SQL)
4. **Cloud-based Data Analytics** (Skills: cloud, data analytics, mathematics)
5. **Introduction to Deep Learning** (Skills: Deep learning, PyTorch, TensorFlow)
6. **Introduction to Big Data** (Skills: Big data, Spark, Data Management)
7. **Data Science in Finance Analytics** (Skills: power bi, business intelligence, data mining, pandas, NumPy, probability)
8. **Data Fundamental** (Skills: Data governance, data architecture, data cleaning)

Notable insights from Figure 7 were some dense clusters appearing near the bottom of the plot. These skills had common themes like "data" appearing in the word or relative soft skills.

4 K-means or DBSCAN clustering implementation

In addition to the hierarchical clustering algorithm, I implemented the K-Means cluster where I first engineered 12 new features for each skill being the title and skill frequency, average salary and rating, as well as binary indications to categorize certain skills such as soft, hard, programming, etc - shown in Figure 8. From here if first normalize the data as F2 (Average Salary per Skill) has very large values compared to the rest of the features, to which I get 8 unique clusters. The optimal value for k number of clusters was chosen by an elbow plot shown in Figure 9, where we can see the elbow point calculated by the *yellowbrick* python library stated k = 7. However, since the minimum number of courses in our curriculum is 8, the optimal k = 8. The cluster assignments obtained from the K-Means algorithm are also stored in the "Clusters" column shown in Figure 8, which was used to further analyze the data. Thus by printing out the unique skills categorized by their clusters, we can come up with course names for the following skills.

1. **Critical Thinking & Communication for Data Science** (Skills: critical thinking, decision making, excel, experimental design, presentation, teamwork, writing)
2. **Advanced Data Management and Analysis** (Skills: data architecture, data governance, data strategy, data warehousing, deep learning, java, natural language processing, nlp, PyTorch, TensorFlow)

3. **Statistics and Predictive Modeling** (Skills: data analysis, data modeling, math, mathematics, predictive analytics, predictive modeling, statistical analysis, statistical modeling, statistics)

4. **Artificial Intelligence and Machine Learning** (Skills: ai, engineering, machine learning, python, r, science)

5. **Big Data Technologies and Analytics** (Skills: big data, cloud, cloud computing, data mining, hadoop, power bi, spark, sql, tableau)

6. **Data Preprocessing and Analysis** (Skills: data preprocessing, data quality control, numpy, pandas, preprocessing, probability, sklearn)

7. **Business Intelligence for Data Science** (Skills: business acumen, 'business intelligence, communication, communication skills, eda, time management)

8. **Data Cleaning and Visualization** (Skills: data cleaning, data management, data science tools, 'data visualization)

5 Interpretation of results and visualizations

As for the interpretation of the results, looking at Figure 7, we can see that the hierarchical clustering was able to group skills based on the distance matrix, and we were able to find the optimal number of clusters from the elbow plot - Figure 9. However, some clusters have related and somewhat unrelated skills. For the K-Mean clustering, we can a visual of the clusters in the form of a scatterplot shown in Figure 10 where we can see 8 distinct clusters that correspond to the courses mentioned in part 4. To note, to ensure proper interpretations by the K-means plot, I used PCA to reduce the dimensions of the normalized data and projected it onto a two-dimensional space, which allowed for easier visualization of the clusters.

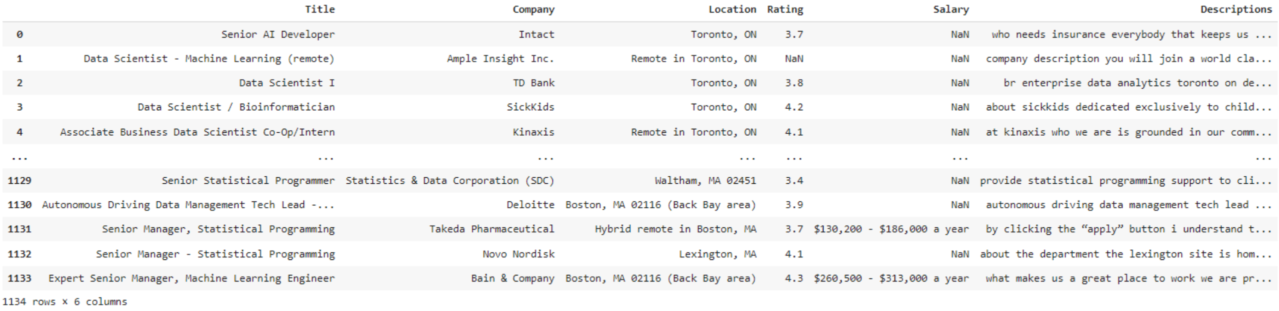
6 Discussion and final course curriculum

When comparing the results from the hierarchical clustering algorithm and the k-means, we can see that both produce very different results. On one hand, the dendrogram shows the distances between skills where it grouped some common skills like 'data-oriented' skills and relative math/stats skills. When looking at k-means, we can see the cluster of skills is more sophisticated and more relatable such that coming up with the course titles was much easier. Additionally, the course diversity seems more structured such that students will be able to develop skills related to data analysis, cloud computing, business, management, and data science tools in a more comprehensive way. Thus I conclude that k-means was able to group like-skills better than the hierarchical clustering algorithm such that the final course curriculum will be based on the k-means clusters. However, it is important to note that the interpretation of the clusters is subjective and heavily depends on the domain knowledge of the researcher. As seen in the hierarchical clusters - Figure 7 - the clustering results may not always be completely accurate without the context of the problem being solved. Therefore, it is important to validate the clustering results through further analysis and domain-specific expertise.

7 OpenAI to describe clustering results

Lastly, by using OpenAI ChatGPT (GPT-3.5 Turbo), I was able to generate its interpretation of the K-Means clusters and give its overall opinion on whether the 8 courses would give a student the skills necessary to be a data scientist. It's output was "Overall, these eight courses cover a wide range of skills related to data science, including data analysis, data management, and data visualization. However, becoming a data scientist requires more than just completing these courses. It also requires practical experience and the ability to apply these skills to real-world problems".

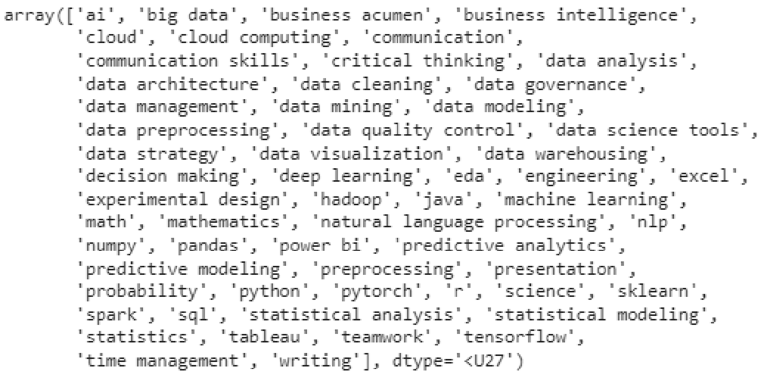
Appendix



	Title	Company	Location	Rating	Salary	Descriptions
0	Senior AI Developer	Intact	Toronto, ON	3.7	NaN	who needs insurance everybody that keeps us ...
1	Data Scientist - Machine Learning (remote)	Ample Insight Inc.	Remote in Toronto, ON	NaN	NaN	company description you will join a world cla...
2	Data Scientist I	TD Bank	Toronto, ON	3.8	NaN	br enterprise data analytics toronto on de...
3	Data Scientist / Bioinformatician	SickKids	Toronto, ON	4.2	NaN	about sickkids dedicated exclusively to child...
4	Associate Business Data Scientist Co-Op/Intern	Kinaxis	Remote in Toronto, ON	4.1	NaN	at kinaxis who we are is grounded in our comm...
...	...	...	...	...	...	...
1129	Senior Statistical Programmer	Statistics & Data Corporation (SDC)	Waltham, MA 02451	3.4	NaN	provide statistical programming support to cli...
1130	Autonomous Driving Data Management Tech Lead -...	Deloitte	Boston, MA 02116 (Back Bay area)	3.9	NaN	autonomous driving data management tech lead ...
1131	Senior Manager, Statistical Programming	Takeda Pharmaceutical	Hybrid remote in Boston, MA	3.7	$130,200 - $186,000 a year	by clicking the “apply” button i understand t...
1132	Senior Manager - Statistical Programming	Novo Nordisk	Lexington, MA	4.1	NaN	about the department the lexington site is hom...
1133	Expert Senior Manager, Machine Learning Engineer	Bain & Company	Boston, MA 02116 (Back Bay area)	4.3	$260,500 - $313,000 a year	what makes us a great place to work we are pr...

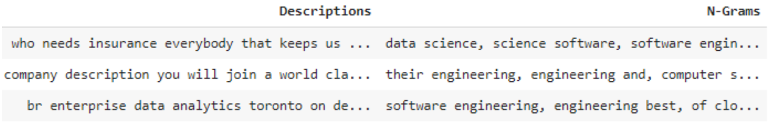
1134 rows × 6 columns

Figure 1: Q1: Web scraped raw dataframe



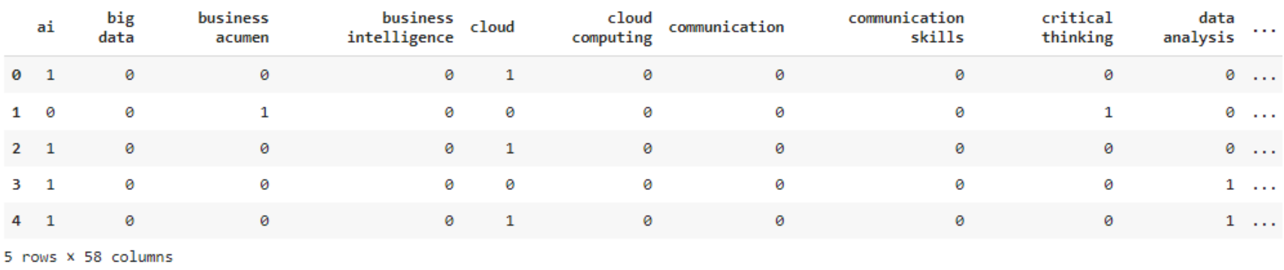
```
array(['ai', 'big data', 'business acumen', 'business intelligence',
       'cloud', 'cloud computing', 'communication',
       'communication skills', 'critical thinking', 'data analysis',
       'data architecture', 'data cleaning', 'data governance',
       'data management', 'data mining', 'data modeling',
       'data preprocessing', 'data quality control', 'data science tools',
       'data strategy', 'data visualization', 'data warehousing',
       'decision making', 'deep learning', 'eda', 'engineering', 'excel',
       'experimental design', 'hadoop', 'java', 'machine learning',
       'math', 'mathematics', 'natural language processing', 'nlp',
       'numpy', 'pandas', 'power bi', 'predictive analytics',
       'predictive modeling', 'preprocessing', 'presentation',
       'probability', 'python', 'pytorch', 'r', 'science', 'sklearn',
       'spark', 'sql', 'statistical analysis', 'statistical modeling',
       'statistics', 'tableau', 'teamwork', 'tensorflow',
       'time management', 'writing'], dtype='<U27')
```

Figure 2: Q2: Combined Skills



Descriptions	N-Grams
who needs insurance everybody that keeps us ...	data science, science software, software engin...
company description you will join a world cla...	their engineering, engineering and, computer s...
br enterprise data analytics toronto on de...	software engineering, engineering best, of clo...

Figure 3: Q2: N-Grams



	ai	big data	business acumen	business intelligence	cloud	cloud computing	communication	communication skills	critical thinking	data analysis	...
0	1	0	0	0	1	0	0	0	0	0	...
1	0	0	1	0	0	0	0	0	1	0	...
2	1	0	0	0	1	0	0	0	0	0	...
3	1	0	0	0	0	0	0	0	0	1	...
4	1	0	0	0	1	0	0	0	0	1	...

5 rows × 58 columns

Figure 4: Q4: Vectorized Skills

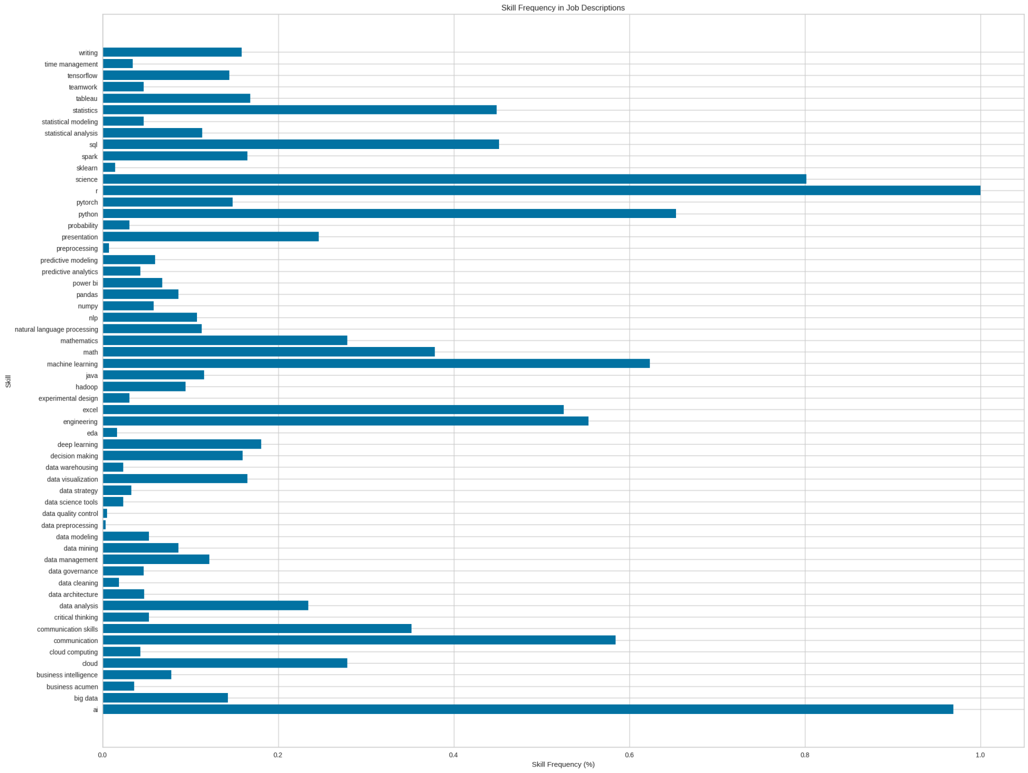


Figure 5: Q2: Skill Frequency

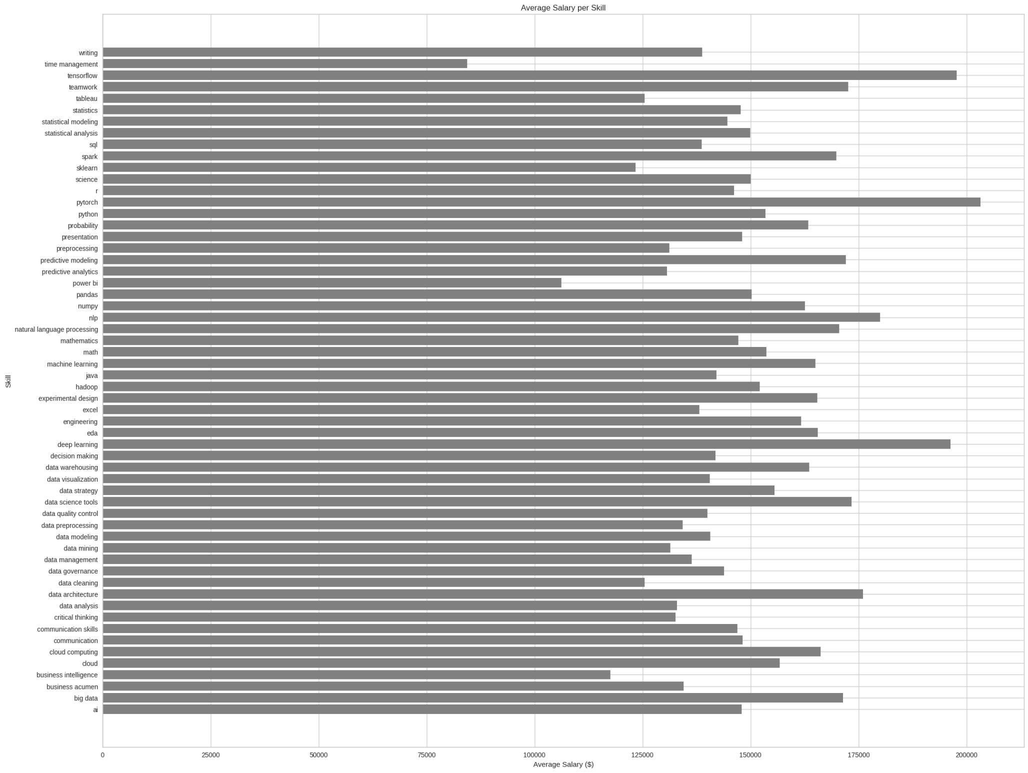


Figure 6: Q2: Average Salary per Skill

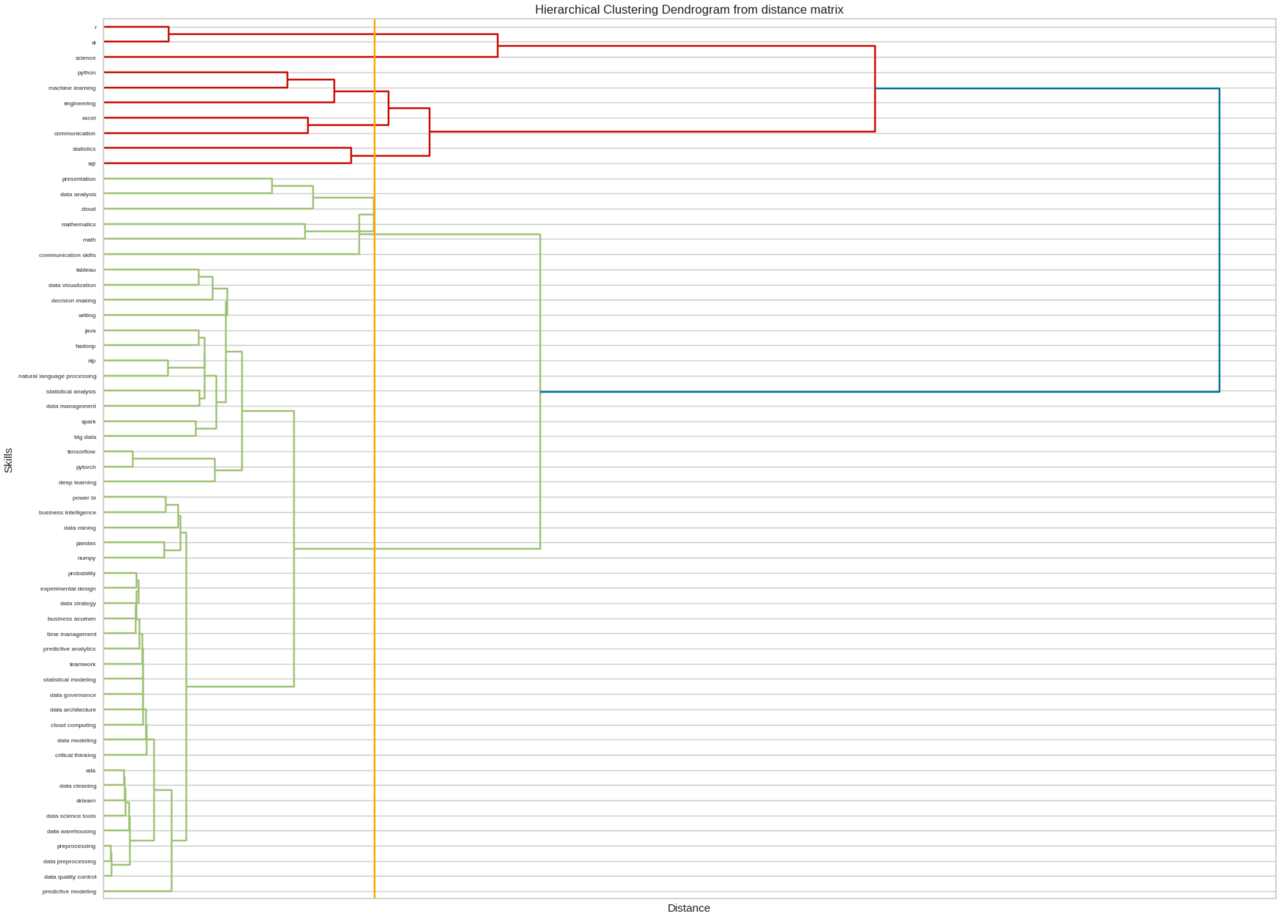
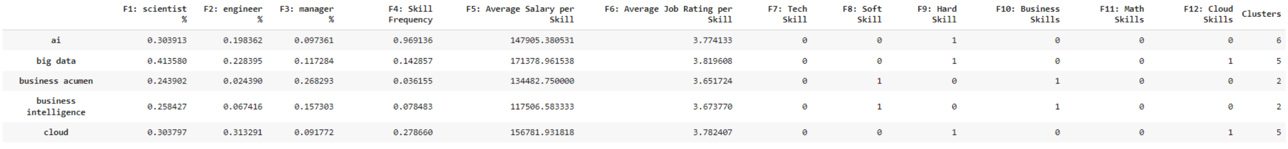


Figure 7: Q3: Dendrograms



	F1: scientist %	F2: engineer %	F3: manager %	F4: Skill Frequency	F5: Average Salary per Skill	F6: Average Job Rating per Skill	F7: Tech Skill	F8: Soft Skill	F9: Hard Skill	F10: Business Skills	F11: Math Skills	F12: Cloud Skills	Clusters
ai	0.303913	0.198362	0.097361	0.969136	147905.380531	3.774133	0	0	1	0	0	0	6
big data	0.413580	0.228395	0.117284	0.142857	171378.961538	3.819608	0	0	1	0	0	1	5
business acumen	0.243902	0.024390	0.268293	0.036155	134482.750000	3.651724	0	1	0	1	0	0	2
business intelligence	0.258427	0.067416	0.157303	0.078483	117506.583333	3.673770	0	1	0	1	0	0	2
cloud	0.303797	0.313291	0.091772	0.278660	156781.931818	3.782407	0	0	1	0	0	1	5

Figure 8: Q4: New Features

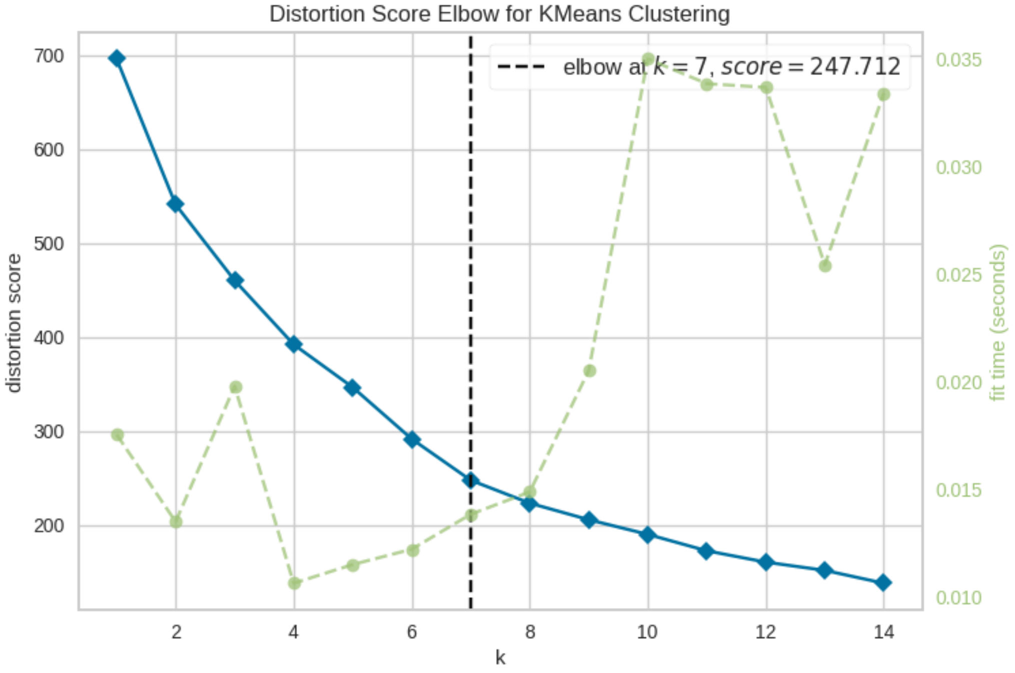


Figure 9: Q4: K-Means Elbow Plot

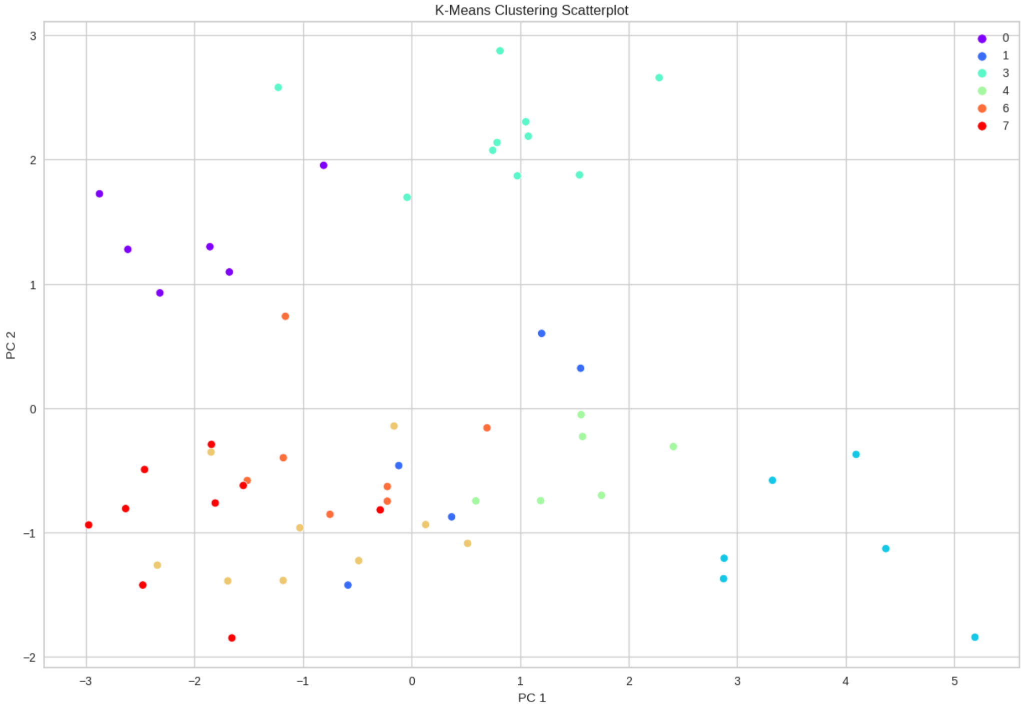
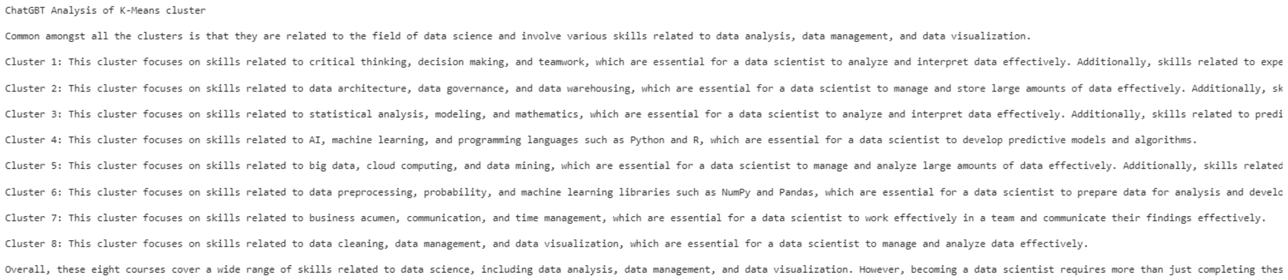


Figure 10: Q4: K-Means Elbow Plot



```
ChatGBT Analysis of K-Means cluster

Common amongst all the clusters is that they are related to the field of data science and involve various skills related to data analysis, data management, and data visualization.

Cluster 1: This cluster focuses on skills related to critical thinking, decision making, and teamwork, which are essential for a data scientist to analyze and interpret data effectively. Additionally, skills related to expe

Cluster 2: This cluster focuses on skills related to data architecture, data governance, and data warehousing, which are essential for a data scientist to manage and store large amounts of data effectively. Additionally, sk

Cluster 3: This cluster focuses on skills related to statistical analysis, modeling, and mathematics, which are essential for a data scientist to analyze and interpret data effectively. Additionally, skills related to predi

Cluster 4: This cluster focuses on skills related to AI, machine learning, and programming languages such as Python and R, which are essential for a data scientist to develop predictive models and algorithms.

Cluster 5: This cluster focuses on skills related to big data, cloud computing, and data mining, which are essential for a data scientist to manage and analyze large amounts of data effectively. Additionally, skills related

Cluster 6: This cluster focuses on skills related to data preprocessing, probability, and machine learning libraries such as NumPy and Pandas, which are essential for a data scientist to prepare data for analysis and develo

Cluster 7: This cluster focuses on skills related to business acumen, communication, and time management, which are essential for a data scientist to work effectively in a team and communicate their findings effectively.

Cluster 8: This cluster focuses on skills related to data cleaning, data management, and data visualization, which are essential for a data scientist to manage and analyze data effectively.

Overall, these eight courses cover a wide range of skills related to data science, including data analysis, data management, and data visualization. However, becoming a data scientist requires more than just completing thes
```

Figure 11: Q7: ChatGBT interpretation of K-Means cluster results