MIE 1624 Introduction to Data Science and Analytics – Winter 2023

Assignment 3

Anbumanivel Mohan Suganthi - 1008696653

April 9, 2023

The aim of this assignment is to design a course curriculum that meets the industry's needs for a new "Master of Business and Management in Data Science and Artificial Intelligence" program at the University of Toronto. To achieve this, I have obtained a dataset of 1800 job descriptions through Indeed's web scraping. These job postings are exclusively for Data Scientist and Data Analyst positions in the job market in the United States.

1. Data collection and cleaning

a). To begin with, I conducted web scraping of job postings separately for data scientists and data analysts in both remote and New York locations, resulting in a total of 1800 job postings. I extracted various details including the job title, company name, job location, rating, date posted, salary information, job description, and other relevant job details. Subsequently, I combined the extracted data from all four sources and saved it in a new CSV file named "webscraping_results_assignmnet3.csv".

For data cleaning, the salary column was the first to be addressed. The values in the salary column were in different formats such as salary per hour, salary per month, salary per week, and salary per year. Most of the values were in a range format, and the average was computed. For the missing values in the first 900 rows, the mode of the first 900 values was computed, and the same was done for the last 900 rows.

Next, duplicate columns and other non-relevant columns such as ‘Description‘, ‘Rating‘, ‘Date‘, and ‘Links‘ were dropped. Afterward, text preprocessing was performed on the 'Description' column, including converting all text to lowercase, removing emojis and special characters, removing rows with no non-whitespace characters, and removing stop words.

2. Exploratory data analysis and feature engineering

a). For this part, I began by manually defining a list of skills based on my own knowledge. Then, using the ChatGPT (gpt-3.5-turbo model), I extracted additional skills and stored them in a separate list. I then combined both lists, removing any duplicates. Each skill in the combined list was then compared to the pre-processed job descriptions column. If the skill was present in a job description, it was assigned a value of 1, otherwise, it was assigned a value of 0. Finally, a pandas dataframe was created with columns representing each skill and the values being either 0 or 1.

b). Several visualizations were created based on the skills identified in the previous step. Word clouds were generated for each skill. The percentage of job postings that mention each skill was plotted to identify the most in-demand skills. The number of job postings that require each skill was also plotted to identify the most requested skills. Finally, the average salary associated with each skill was plotted to identify the most lucrative skills in the market. All the plots are shown in the appendix.

3. Hierarchical clustering implementation

a). In this project, a distance matrix is computed by calculating the cosine distance between every pair of skills. The resulting values are then stored in a pandas dataframe where both the column and index represent each skill. Subsequently, a dendrogram is used to plot the clustering result with complete linkage. Each resulting cluster is color-coded and represented by a distinct color.

To create 8 to 12 courses, a threshold of 1.5 is applied to the distance matrix, which leads to the formation of 11 clusters. The resulting clusters are displayed in Figure 7.a in the appendix. Out of the 11 clusters, 9 have at least 3 skills, and the remaining 2 have 2 skills. For this assignment, the focus is exclusively on the 9 clusters that consist of at least 3 skills. The course curriculum is shown in Figure 7.b in the appendix.

4. K-means or DBSCAN clustering implementation

a). In this part, K-means is chosen. It works by randomly initializing k centroids, assigning each data point to the nearest centroid, and then updating the centroids based on the mean of the data points in each cluster. This process is repeated until the centroids no longer change significantly.

A set of 10 features were engineered to describe each skill for clustering. These features include Skill frequency, Average salary for skill, Binary indication of soft skill, Binary indication of hard skill, Distance matrix (taking the average between each skill), correlation (taking the average between each skill), Binary indication of if the skill is in demand (above average salary), Number of correlated skill, Standard deviation of the distance matrix, and Binary indication of if the skill is frequently requested. These features were carefully selected to ensure that they capture important characteristics of each skill, such as frequency, demand, correlation with other skills, and skill type.

The K-means clustering algorithm was utilized to group skills with similar feature values, using the aforementioned 10 engineered features and setting n_clusters to 10, in order to create a course curriculum consisting of 8 to 12 courses. The clusters obtained from this process are depicted in Figure 11.a in the appendix. Out of the 10 clusters formed, 9 clusters contain at least 3 skills while the remaining cluster consists of only 2 skills. With regards to this assignment, only the 9 clusters containing at least 3 skills were considered. The course curriculum can be found in Figure 11.b in the appendix.

b). The Elbow method was applied to identify the most suitable number of clusters (k) for the K-means algorithm. To do this, the k-means model was fitted with k values ranging from 1 to 15, and the corresponding inertia values were recorded. A plot of inertia versus the number of clusters was generated, and the elbow point was observed at k=5, indicating that five clusters are the optimal choice for this dataset. The plot is shown in Figure 13 in the appendix. The K-means algorithm was then run with k=5, and principal component analysis was performed to reduce the dimensions. Finally, a scatterplot was generated to visualize the resulting clusters. It is shown in Figure 14 in the appendix.

5. Interpretation of results and visualizations

a). In part 3, a dendrogram was generated to visualize the clustering results from the hierarchical clustering algorithm which is shown in Figure 9 in the appendix. Each resulting cluster is color-coded and represented by a distinct color. Out of the 11 clusters, 9 have at least 3 skills, and the remaining 2 have 2 skills. For this assignment, the focus is exclusively on the 9 clusters that consist of at least 3 skills.

b). In part 4, a scatterplot was plotted to visualize the clustering results obtained from the k-means algorithm. Due to the presence of 10 features, a principal component analysis (PCA) was performed to reduce the dimensions of the data and enable the visualization of the graph in 2 dimensions. The number of resulting clusters was 5, which was determined using the elbow method. Each of these 5 clusters was assigned a unique color, and they were found to be visually distinct from one another.

c). In part 4, an elbow plot was generated to determine the optimal number of clusters for K-Means clustering. The plot indicated inertia which is the sum of squared distances between each point in the cluster and the centroid of that cluster did not decrease significantly beyond 5 clusters. Hence, 5 was identified as the optimum number of clusters.

6. Discussion and final course curriculum

The course curriculum that has been finalized is based on the hierarchical clustering results from section 3. The rationale behind this decision is that hierarchical clustering is more appropriate when the number of clusters is not predetermined. Hierarchical clustering provides more flexibility than k-means clustering as it allows the creation of clusters at varying levels of granularity. Compared to k-means clustering, hierarchical clustering is more robust as it is not sensitive to the initial placement of centroids. Additionally, dendrograms, which are generated by hierarchical clustering, are beneficial in visualizing the clusters and can aid in the identification of significant clusters within the data. The Course curriculum for a new "Master of Business and Management in Data Science and Artificial Intelligence" program at the University of Toronto is mentioned in the figure below.

Course 1: Data Science Fundamentals: From Statistics to Visualization
Course material - Data Analysis, Communication, Data Visualization, R, Modeling, Statistics, Sql, Tableau
Course 2: Applied Deep Learning with Python Libraries
Course material - Keras, Pandas, Numpy, Tensorflow, Deep Learning, Pytorch
Course 3: Big Data and AI Infrastructure with Java and AWS
Course material - C, Spark, AWS, Hadoop, Big Data, Optimization, Nosql, Docker, Artificial Intelligence, Java, AI
Course 4: Data Analytics and Visualization for Business Professionals
Course material - Project Management, Powerpoint, Presentation, Power BI, Interpersonal, Excel
Course 5: Advanced Data Analysis: Statistical Modeling and Predictive Analytics
Course material - Statistical Analysis, Data Mining, Predictive Analytics
Course 6: Data Warehousing and ETL for Business Intelligence
Course material - Data Modeling, ETL, Data Warehousing
Course 7: Cloud Computing and Simulation with MATLAB on AWS
Course material - Cloud Computing, Amazon Web Services, Matlab, Simulation
Course 8: Professional Development for Effective Teamwork and Problem Solving
Course material - Business Acumen, Teamwork, Flexibility, Preparation, Critical Thinking, Problem Solving, Time Management
Course 9: Data Science Tools and Techniques: From Data Cleaning to Visualization
Course material - Data Cleaning, Html, Data Manipulation, Github, SPSS, Javascript

7. OpenAI to describe clustering results

Using OpenAI, I asked ChatGPT to describe the common characteristics shared by the clusters in the final course curriculum and to provide a course name for each cluster. The results can be found in Figure 15 in the appendix.

Appendix

	Title	Company	Location	Descriptions	Salary	Descriptions_pre-processed
0	Data Scientist	The Travelers Companies, Inc.	Remote in Hartford, CT	Who Are We?\nTaking care of our customers, our...	161250.0	taking care customers communities thats travel...
1	Data Scientist - Quality and Safety - REMOTE	MEDSTAR HEALTH	Remote in Washington, DC 20008	The Data Scientist supports clinical quality a...	150000.0	data scientist supports clinical quality safet...
2	Data Scientist - RWD	Norstella	Remote	Job Summary:\nWe are seeking an experienced Da...	162500.0	job summary seeking experienced data scientist...
3	Associate Data Scientist	Protective Life Corporation	Remote	The work we do has an impact on millions of li...	87500.0	work impact millions lives part help protect c...
4	Associate Data Scientist - Online Business Ana...	Home Depot / THD	Remote in Atlanta, GA 30301	Position Purpose:\nThe Associate Data Scientis...	150000.0	position purpose associate data scientist resp...
...	...	...	...	...	...	...
1546	Business Analyst II- HBA05-2	Horizon Blue Cross Blue Shield of New Jersey	Wall, NJ	Horizon BCBSNJ employees must live in New Jers...	78490.0	horizon bcbsnj employees must live new jersey ...
1547	Business Analyst	Granite Solutions Groupe	Manhattan, NY	Job Category: Business Analysis\nJob Location:...	80000.0	job category business analysis job location ma...
1548	Data Analyst	1199SEIU Training and Employment Funds	New York, NY	Responsibilities\n\nAssist Associate Director ...	80000.0	responsibilities assist associate director imp...
1549	Senior Business Analyst	Intuit	New York, NY 10012	Overview\nCome join our Business Operations (B...	80000.0	overview come join business operations bizops ...
1550	Data Analyst Trainee	Chubb INA Holdings Inc.	Whitehouse Station, NJ 08889	The position will be based in Whitehouse Stati...	80000.0	position based whitehouse station nj supports ...

1551 rows × 6 columns

Figure 1: Dataframe after data cleaning and pre-processing the descriptions column

Figure 2: Wordcloud for skills

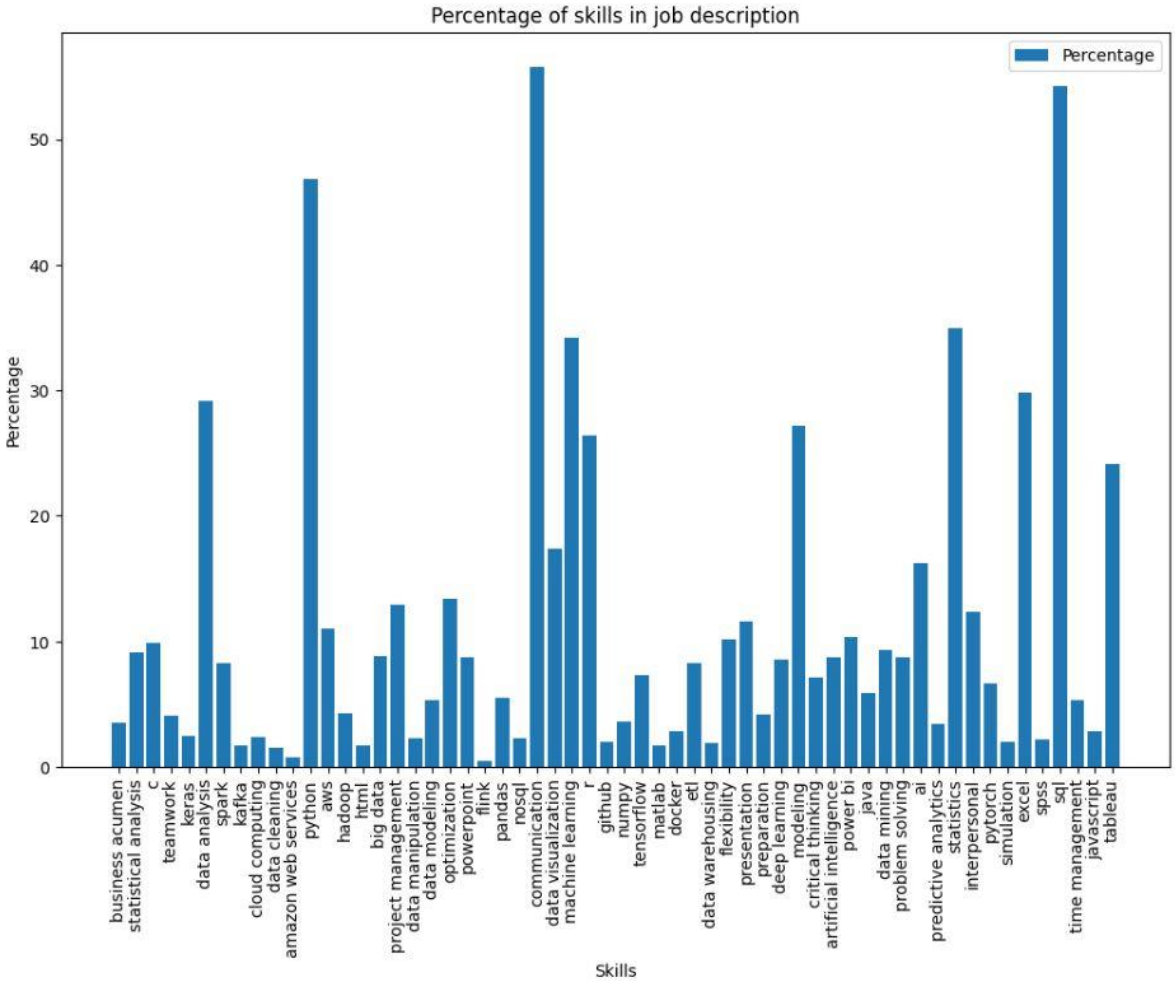

Figure 3: Percentage of skills in job description

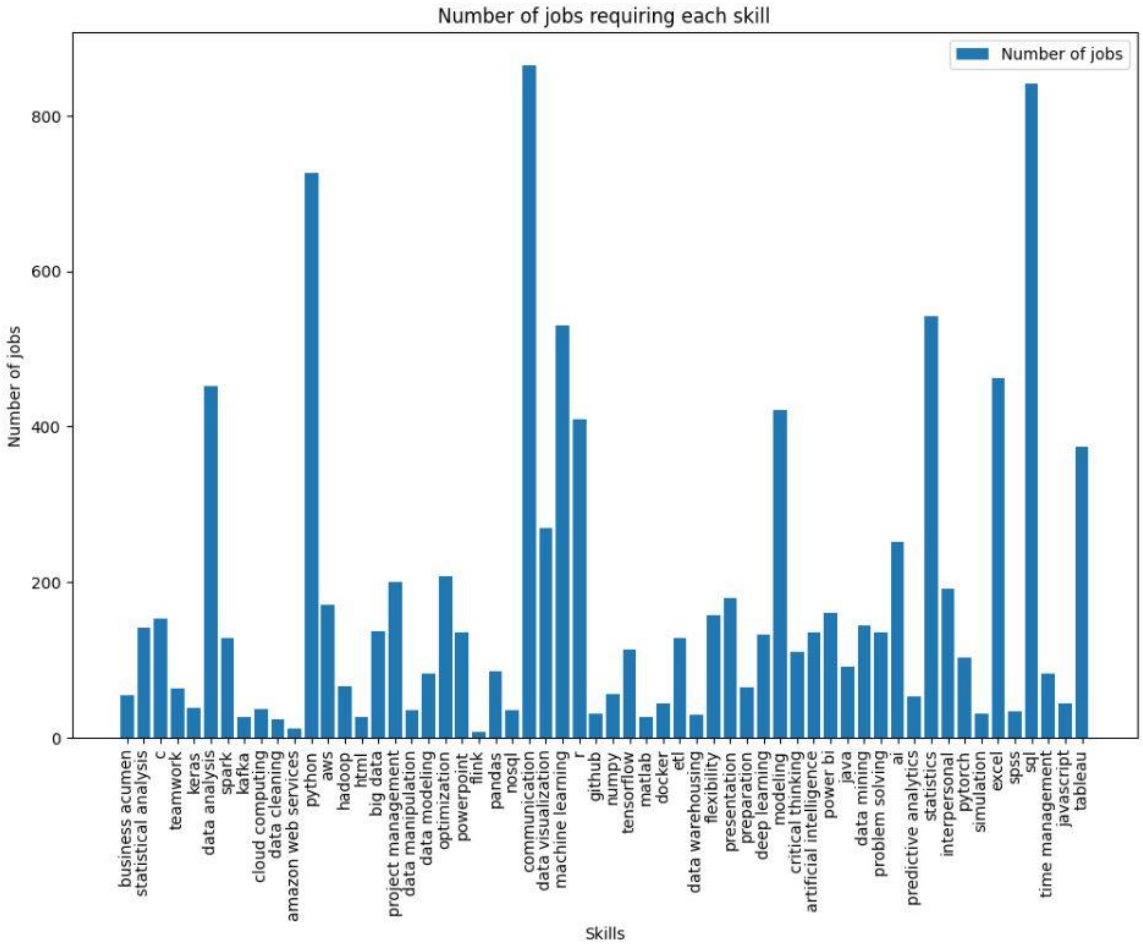

Figure 4: Number of jobs requiring each skill

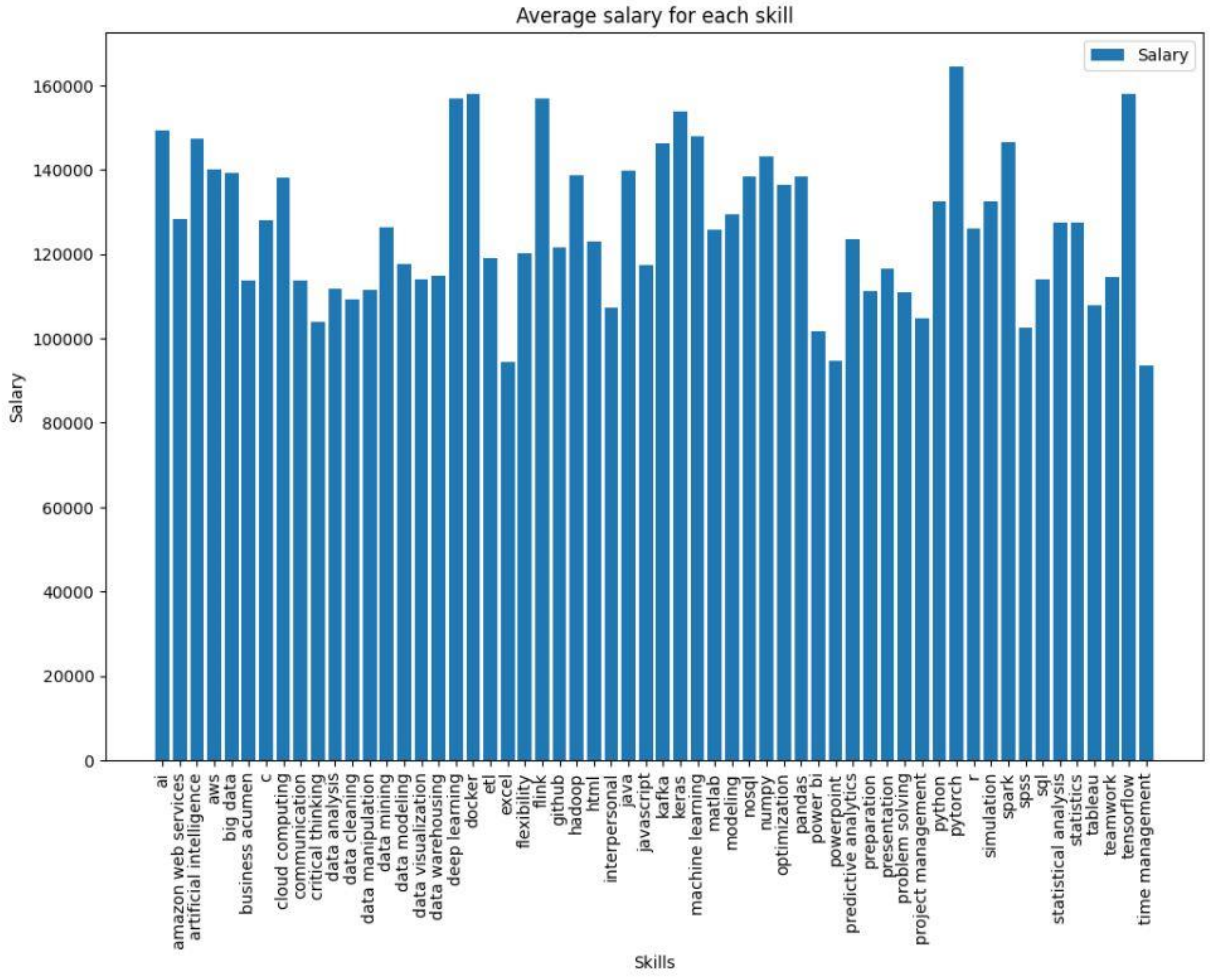

Figure 5: Average salary for each skill

	business acumen	statistical analysis	c	teamwork	keras	data analysis	spark	kafka	cloud computing	data cleaning	...	statistics	interpersonal	pytorch	simulation	excel	spss	sql	time management	javascript	tableau
business acumen	0.000000	0.919778	0.911987	0.931421	1.000000	0.916790	0.916130	0.973811	0.977628	1.000000	...	0.842178	0.941074	1.000000	0.951118	0.892371	1.000000	0.821685	0.969944	1.000000	0.852230
statistical analysis	0.919778	0.000000	0.931916	0.946949	0.972677	0.766292	0.918438	1.000000	0.944620	0.965619	...	0.746785	0.890601	0.950212	0.954624	0.855032	0.870015	0.732835	0.925600	0.936520	0.856296
c	0.911987	0.931916	0.000000	0.898145	0.934426	0.855500	0.864758	0.906648	0.906964	0.950493	...	0.753445	0.877475	0.848648	0.898358	0.849550	0.958405	0.779767	0.964289	0.865933	0.849505
teamwork	0.931421	0.946949	0.898145	0.000000	1.000000	0.881480	0.955629	0.975754	0.979288	0.974283	...	0.866355	0.845429	0.962758	0.977372	0.865186	0.978393	0.839256	0.846956	0.981007	0.882736
keras	1.000000	0.972677	0.934426	1.000000	0.000000	0.946588	0.828606	0.906341	0.839986	1.000000	...	0.839736	0.976585	0.568428	0.912593	0.992453	1.000000	0.904905	0.982086	0.975544	0.966447
data analysis	0.916790	0.766292	0.855500	0.881480	0.946588	0.000000	0.842631	0.972844	0.922673	0.903988	...	0.575722	0.769172	0.921212	0.864833	0.638928	0.870934	0.526396	0.880532	0.879454	0.625445
spark	0.916130	0.918438	0.864758	0.955629	0.828606	0.842631	0.000000	0.695003	0.782882	0.982028	...	0.712579	0.904688	0.722389	0.905120	0.938557	1.000000	0.714612	0.980554	0.893814	0.904393
kafka	0.973811	1.000000	0.906648	0.975754	0.906341	0.972844	0.695003	0.000000	0.968361	1.000000	...	0.942135	0.958333	0.905187	1.000000	0.964186	1.000000	0.880548	1.000000	0.970987	0.960195
cloud computing	0.977628	0.944620	0.906964	0.979288	0.839986	0.922673	0.782882	0.968361	0.000000	1.000000	...	0.872892	0.988136	0.870410	0.940946	0.961757	1.000000	0.886621	1.000000	0.950432	0.931993
data cleaning	1.000000	0.965619	0.950493	0.974283	1.000000	0.903988	0.982028	1.000000	1.000000	0.000000	...	0.886018	0.970537	1.000000	1.000000	0.943020	0.894979	0.873302	1.000000	0.969227	0.936670
amazon web services	1.000000	1.000000	0.953324	0.963630	1.000000	0.959266	0.898334	1.000000	0.952542	1.000000	...	0.950401	0.958333	0.914668	1.000000	0.973139	1.000000	0.930320	1.000000	1.000000	0.970146
python	0.853535	0.718703	0.684953	0.845697	0.789279	0.601987	0.624219	0.842865	0.810856	0.886364	...	0.394219	0.817866	0.674536	0.833356	0.753085	0.942716	0.324279	0.889341	0.832148	0.593152
aws	0.947967	0.896958	0.839258	0.951827	0.838730	0.812959	0.676817	0.749811	0.798850	0.953171	...	0.747074	0.917217	0.751345	0.917591	0.939517	0.986885	0.702023	0.949331	0.919300	0.849738
hadoop	0.916874	0.917692	0.861725	0.938432	0.940544	0.896565	0.494448	0.788396	0.899577	0.975062	...	0.784847	0.947099	0.903698	0.912231	0.943162	1.000000	0.755661	1.000000	0.926329	0.898924
html	1.000000	0.950452	1.000000	0.925875	1.000000	0.953877	0.982733	0.962257	0.967759	1.000000	...	0.924185	0.971693	0.980676	1.000000	0.927007	1.000000	0.878273	0.978343	0.733909	0.908732

Figure 6: Distance matrix of the first 15 skills out of 58 skills

```
Cluster 1: ['python', 'machine learning']
Cluster 2: ['data analysis', 'communication', 'data visualization', 'r', 'modeling', 'statistics', 'sql', 'tableau']
Cluster 3: ['keras', 'pandas', 'numpy', 'tensorflow', 'deep learning', 'pytorch']
Cluster 4: ['c', 'spark', 'aws', 'hadoop', 'big data', 'optimization', 'nosql', 'docker', 'artificial intelligence', 'java', 'ai']
Cluster 5: ['project management', 'powerpoint', 'presentation', 'power bi', 'interpersonal', 'excel']
Cluster 6: ['statistical analysis', 'data mining', 'predictive analytics']
Cluster 7: ['kafka', 'flink']
Cluster 8: ['data modeling', 'etl', 'data warehousing']
Cluster 9: ['cloud computing', 'amazon web services', 'matlab', 'simulation']
Cluster 10: ['business acumen', 'teamwork', 'flexibility', 'preparation', 'critical thinking', 'problem solving', 'time management']
Cluster 11: ['data cleaning', 'html', 'data manipulation', 'github', 'spss', 'javascript']
```

Figure 7.a: Clustering results for Hierarchical Clustering

Cluster 1: Data Analysis, Communication, Data Visualization, R, Modeling, Statistics, Sql, Tableau
Cluster 2: Keras, Pandas, Numpy, Tensorflow, Deep Learning, Pytorch
Cluster 3: C, Spark, AWS, Hadoop, Big Data, Optimization, Nosql, Docker, Artificial Intelligence, Java, AI
Cluster 4: Project Management, Powerpoint, Presentation, Power BI, Interpersonal, Excel
Cluster 5: Statistical Analysis, Data Mining, Predictive Analytics
Cluster 6: Data Modeling, ETL, Data Warehousing
Cluster 7: Cloud Computing, Amazon Web Services, Matlab, Simulation
Cluster 8: Business Acumen, Teamwork, Flexibility, Preparation, Critical Thinking, Problem Solving, Time Management
Cluster 9: Data Cleaning, Html, Data Manipulation, Github, SPSS, Javascript

Figure 7.b: Clusters selected for course curriculum - Hierarchical Clustering

Course 1: Data Science Fundamentals: From Statistics to Visualization
Course material - Data Analysis, Communication, Data Visualization, R, Modeling, Statistics, Sql, Tableau
Course 2: Applied Deep Learning with Python Libraries
Course material - Keras, Pandas, Numpy, Tensorflow, Deep Learning, Pytorch
Course 3: Big Data and AI Infrastructure with Java and AWS
Course material - C, Spark, AWS, Hadoop, Big Data, Optimization, Nosql, Docker, Artificial Intelligence, Java, AI
Course 4: Data Analytics and Visualization for Business Professionals
Course material - Project Management, Powerpoint, Presentation, Power BI, Interpersonal, Excel
Course 5: Advanced Data Analysis: Statistical Modeling and Predictive Analytics
Course material - Statistical Analysis, Data Mining, Predictive Analytics
Course 6: Data Warehousing and ETL for Business Intelligence
Course material - Data Modeling, ETL, Data Warehousing
Course 7: Cloud Computing and Simulation with MATLAB on AWS
Course material - Cloud Computing, Amazon Web Services, Matlab, Simulation
Course 8: Professional Development for Effective Teamwork and Problem Solving
Course material - Business Acumen, Teamwork, Flexibility, Preparation, Critical Thinking, Problem Solving, Time Management
Course 9: Data Science Tools and Techniques: From Data Cleaning to Visualization
Course material - Data Cleaning, Html, Data Manipulation, Github, SPSS, Javascript

Figure 8: Course curriculum based on Hierarchical Clustering

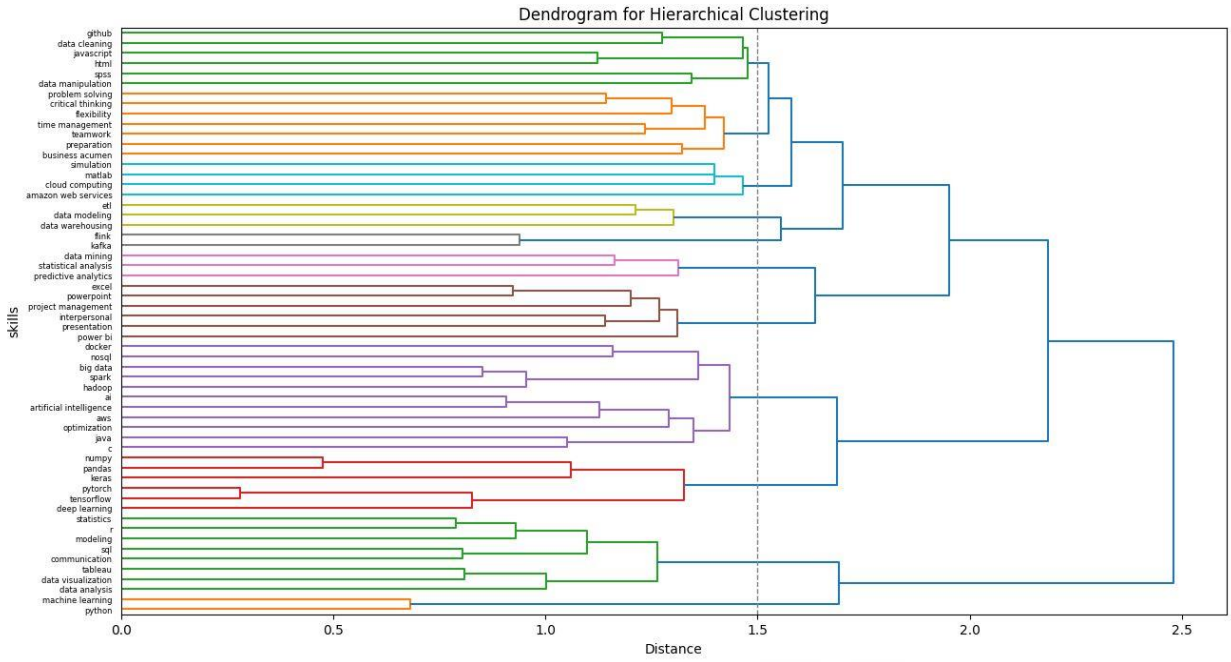

Figure 9: Dendrogram for Hierarchical Clustering

	Skill Frequency	Average Salary	Soft Skill	Hard Skill	Average_distance_matrix	Average_correlation	In Demand	Correlated Skills	Standard Deviation Distance Matrix	Frequently Requested
business acumen	54	149329.425794	1	0	0.920050	0.026994	1	0	0.135111	0
statistical analysis	141	128216.833333	0	1	0.878107	0.037554	1	0	0.139727	0
c	153	147225.270588	0	1	0.859558	0.052781	1	0	0.133938	0
teamwork	63	140014.551462	1	0	0.909602	0.032975	1	0	0.131171	0
keras	38	139300.170073	0	1	0.893077	0.063770	1	0	0.162099	0
data analysis	452	113620.879630	0	1	0.811279	0.039309	0	0	0.153664	1
spark	129	127880.422222	0	1	0.817565	0.107439	1	0	0.157742	0
kafka	27	138149.583784	0	1	0.900361	0.063010	1	0	0.150278	0
cloud computing	37	113702.990509	0	1	0.899180	0.058329	0	0	0.133640	0
data cleaning	24	104025.744545	0	1	0.940158	0.023985	0	0	0.131754	0

Figure 10: Dataframe of Skills of 10 Unique Features out of 58 Skills

```
Cluster 1: ['data analysis', 'cloud computing', 'html', 'big data', 'project management', 'github', 'power bi', 'simulation', 'time management']
Cluster 2: ['business acumen', 'c', 'numpy', 'matlab', 'flexibility', 'interpersonal']
Cluster 3: ['statistical analysis', 'spark', 'hadoop', 'docker', 'etl', 'predictive analytics', 'excel', 'spss']
Cluster 4: ['powerpoint', 'modeling', 'tableau']
Cluster 5: ['data manipulation', 'data modeling', 'pandas', 'tensorflow', 'ai', 'javascript']
Cluster 6: ['teamwork', 'keras', 'kafka', 'communication', 'r', 'data warehousing', 'presentation', 'preparation']
Cluster 7: ['data cleaning', 'deep learning', 'data mining', 'pytorch']
Cluster 8: ['amazon web services', 'python', 'aws', 'machine learning', 'artificial intelligence', 'java', 'sql']
Cluster 9: ['optimization', 'flink', 'nosql', 'data visualization', 'critical thinking']
Cluster 10: ['problem solving', 'statistics']
```

Figure 11.a: Clustering results for K-means Clustering

Cluster 1: Data Analysis, Cloud Computing, Html, Big Data, Project Management, Github, Power BI, Simulation, Time Management
Cluster 2: Business Acumen, C, Numpy, Matlab, Flexibility, Interpersonal
Cluster 3: Statistical Analysis, Spark, Hadoop, Docker, ETL, Predictive Analytics, Excel, SPSS
Cluster 4: Powerpoint, Modeling, Tableau
Cluster 5: Data Manipulation, Data Modeling, Pandas, Tensorflow, AI, Javascript
Cluster 6: Teamwork, Keras, Kafka, Communication, R, Data Warehousing, Presentation, Preparation
Cluster 7: Data Cleaning, Deep Learning, Data Mining, Pytorch
Cluster 8: Amazon Web Services, Python, AWS, Machine Learning, Artificial Intelligence, Java, Sql
Cluster 9: Optimization, Flink, Nosql, Data Visualization, Critical Thinking

Figure 11.b: Clusters selected for course curriculum - K-means Clustering

Course 1: Data Science and Cloud Computing: Strategies for Successful Project Management
Course material - Data Analysis, Cloud Computing, Html, Big Data, Project Management, Github, Power BI, Simulation, Time Management
Course 2: Applied Data Science with Business Acumen and Interpersonal Skills
Course material - Business Acumen, C, Numpy, Matlab, Flexibility, Interpersonal
Course 3: Big Data Analytics with Spark, Hadoop, and Predictive Analytics
Course material - Statistical Analysis, Spark, Hadoop, Docker, ETL, Predictive Analytics, Excel, SPSS
Course 4: Data Visualization and Presentation
Course material - Powerpoint, Modeling, Tableau
Course 5: Data Engineering with Python and Tensorflow
Course material - Data Manipulation, Data Modeling, Pandas, Tensorflow, AI, Javascript
Course 6: Data Modeling and Database Design for Analytics
Course material - Teamwork, Keras, Kafka, Communication, R, Data Warehousing, Presentation, Preparation
Course 7: Fundamentals of Deep Learning
Course material - Data Cleaning, Deep Learning, Data Mining, Pytorch
Course 8: Developing AI Applications
Course material - Amazon Web Services, Python, AWS, Machine Learning, Artificial Intelligence, Java, Sql
Course 9: The Art of Data Analysis
Course material - Optimization, Flink, Nosql, Data Visualization, Critical Thinking

Figure 12: Course curriculum based on K-means Clustering

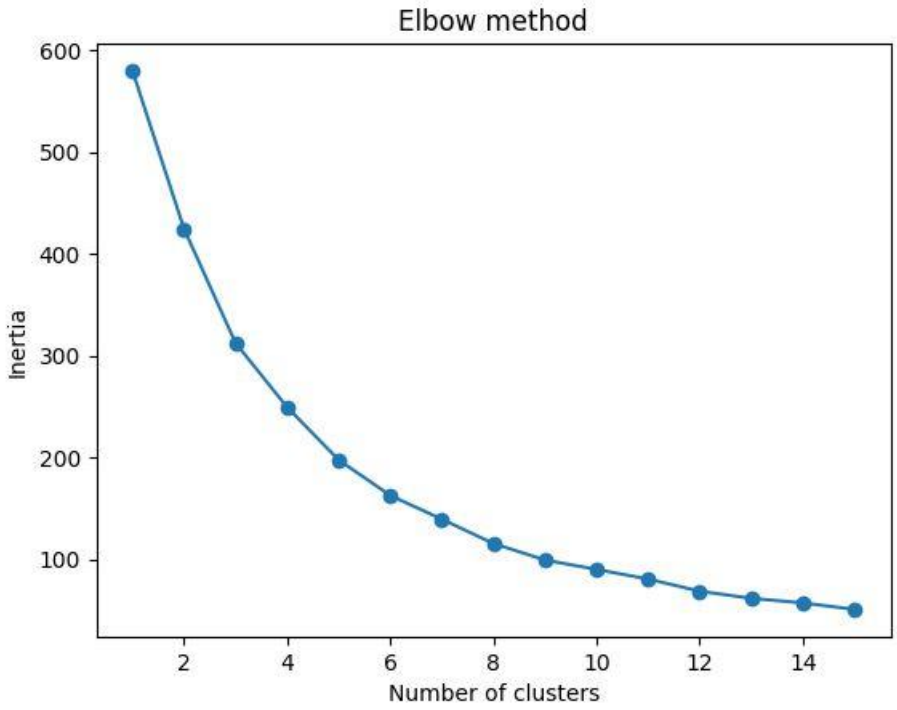

Figure 13: Plot of Elbow method to find optimal k value

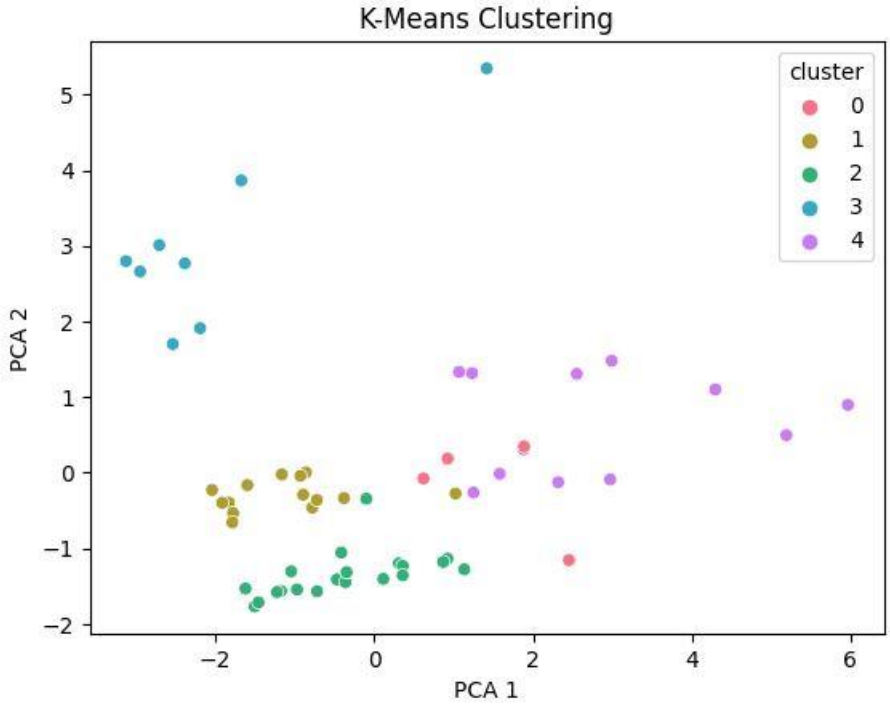

Figure 14: Scatterplot of K-means Clustering

What is common among all the clusters is that they all involve skills and knowledge related to data and its analysis, along with various tools and technologies used for this purpose.

Course 1: "Data Analysis and Visualization" - covering topics such as data analysis, communication, data visualization, R, modeling, statistics, SQL, and Tableau.

Course 2: "Deep Learning with Python" - covering topics such as Keras, Pandas, Numpy, Tensorflow, Deep Learning, and PyTorch.

Course 3: "Big Data and AI" - covering topics such as C, Spark, AWS, Hadoop, NoSQL, Optimization, Docker, Artificial Intelligence, Java, and AI.

Course 4: "Project Management and Presentation Skills" - covering topics such as project management, PowerPoint, presentation, Power BI, interpersonal skills, and Excel.

Course 5: "Predictive Analytics and Data Mining" - covering topics such as statistical analysis, data mining, and predictive analytics.

Course 6: "Data Modeling and Warehousing" - covering topics such as data modeling, ETL, and data warehousing.

Course 7: "Cloud Computing and Simulation" - covering topics such as cloud computing, Amazon Web Services, Matlab, and simulation.

Course 8: "Business Skills for Data Professionals" - covering topics such as business acumen, teamwork, flexibility, preparation, critical thinking, problem-solving, and time management.

Course 9: "Data Cleaning, Manipulation and Programming" - covering topics such as data cleaning, HTML, data manipulation, Github, SPSS, and Javascript.

Figure 15: ChatGPT description about the final course curriculum and course name