MIE 1624 Introduction to Data Science and Analytics – Winter 2023

Assignment 3

Name: Ayesha Patnaik, Student Number: 1008681696

The objective of this assignment is to design a course curriculum for a new "Master of Business and Management in Data Science and Artificial Intelligence" program at University of Toronto with focus not only on technical but also on business and soft skills, by extracting skills from Indeed job postings and using clustering algorithms to derive insights.

Part-1, Data Collection and Cleaning : In this data collection and cleaning part, I first performed web scraping of Indeed website for the job title "data scientist" for "Remote" locations that generated 150 job postings. Then I changed the location to "USA" which resulted in 525 job postings. Finally, I scrapped for the location "Vancouver" which resulted in 517 job postings. Hence, after combining the three datasets, my final collected data had 1192 job postings. The data was scrapped for job title, location, company name, job description, and salary. Please note that in order to avoid generating new results, I have stored the code block inside a no-call function for reference and proceeded with the first generated dataset with 1192 postings. Furthermore, I cleaned the job descriptions from the raw data in the following steps:

1. Converting all words to lower cases to maintain consistency for easier access.
2. Removing stopwords
3. Removing alpha-numeric characters.

Part-2, Exploratory Data Analysis and Feature Engineering: In this part, my objective was to extract skills required for data scientist job roles, encompassing both technical and business aspects. In order to do so, I first manually created two separate lists of skills for technical and business aspects respectively. This is created by leveraging my own knowledge as well as by analyzing raw word frequencies of the job descriptions. Thereafter, ChatGPT API was utilized to create another set of 20 skills. The model used for this task was *"gpt-3.5-turbo"* with a maximum token limit of 1024, a temperature of 0.6 and a top_p value as the default setting of 1. This new set of 20 skills was cross-referenced with the manually generated list to identify any unique skills. The additional unique skills were identified as technical skills. Subsequently, all the skills generated manually and those generated by ChatGPT were combined into two lists categorized as technical and non-technical. Please note that I have stored the code block for ChatGPT prompt in a no-call function to limit my free trial usage as well as to not generate any new results. The original generated results can be found in Figure-1. Next, N-gram models were used to extract skills from the job descriptions. This was done by defining a function that returns a list of word sequences. Two lists for bigrams and trigrams were generated from the job descriptions, which were then combined to form a single list of extracted words. The collected list was cross referenced with N-gram extracted list of skills to find out that all of my collected skills are in fact present in the job descriptions with a significantly high frequency of mentions. The results were visualized separately for technical and non-technical skills using word clouds and bar plots representing skill frequency in each category, as can be seen in Figure-2, Figure-3, Figure-4 and Figure-5. From the visualization, we can interpret that the technical skills like **"data science", "data analysis", "data engineering"** appear very frequently in job descriptions whereas non-technical skills like **"team work"**, **"business acumen"** and **"communication"** are deemed as important. We also infer that among technical skills, skills in R appear to be most required across all job descriptions close to 99% frequency and among non-technical skills, "communication" and "management" are in demand with almost 50% frequently mentioned. Before proceeding to the next sections, all the skills were combined into one list generating **55 skills in total.**

Part-3, Hierarchical clustering implementation: The objective here was to identify relationships between the skills. So I considered agglomerative clustering which combines the data points based on their proximities, instead of divisive clustering which was counterintuitive as it starts with all skills as one cluster and then divides them into smaller clusters.After experimenting with various distance calculation method such as euclidean, manhattan and cosine; as well as various linkage methods such as centroid,complete, single and ward, I chose the combination of cosine distance and ward linkage as this provided me with the best results. The clusters were visualized by plotting a dendrogram in Figure-6. Since the aim was to form at least 8 courses, a suitable cut-line was determined that resulted in the formation of 10 clusters. The interpretation of dendrogram and development of course curriculum are discussed in detail in Part-5 and Part-6.

Part-4, K-means or DBSCAN clustering implementation: For part-4 clustering, I decided to use K-means clustering. This is based on the fact that my data is relatively structured hence preferring DBSCAN may not be the right choice**.** First, I engineered 10 unique features that described key information about the skills. The following are my engineered features:

1. **Skill Frequency :** The number of times the skill has been mentioned in job postings.

2. **Average Salary :** Average salary across all jobs that demands the skill.
3. **Average Rating :** Average ratings of jobs that demand the skill.
4. **Average Work Experience :** Average work experience required for the jobs that demand the skill.
5. **Average Full-Time Demand :** Average number of full-time type jobs that demand the skill.
6. **Average Senior-Roles Demand :** Average number of senior-level jobs that demand the skill.
7. **Average Remote-Jobs Demand :** Average number of remote jobs that demand the skill ; an important feature especially after pandemic.
8. **Average Master's Degree Demand :** Average number of jobs with Master's Degree requirement demanding the skill.
9. **Average Demand in Finance :** Average number of jobs in finance that demand the skill.
10. **Average Demand in Healthcare:** Average number of jobs in healthcare that demand the skill.

A detailed description of how the features were engineered can be found in Table-1. Each feature was created separately into a list, which were subsequently combined to form a final dataframe containing 55 rows for each skill and 10 features for each. The data were then standardized due to varying order of magnitudes of the values. Furthermore, to avoid the effect of "curse of dimensionality", *Principal Component Analysis* was employed to reduce the dimensions from 10 to 2. This enabled me to determine the optimal number of clusters by the *Elbow Method* as seen in Figure-7. From the plot, we observed that roughly the rate of decrease in the inertias is significantly low after 10 and a flat line pertained thereafter. So **the optimal number of clusters was considered as 10**. K-Means clustering was performed on the dataset with optimal number of clusters and the results can be seen in the plot Figure-8. The results and interpretations are discussed in the next sections.

Part-5, Interpretation of results and visualizations:

- **Hierarchical clustering:** As per the dendrogram in Figure-6, we got 10 clusters out of which we designed 11 courses. The clusters formed were as follows:

1. Cluster-1 : data mining, statistical analysis, predictive modeling, java.
2. Cluster-2 : curiosity, data cleaning
3. Cluster-3 : deep learning,natural language processing,team work, pipelines, data engineering
4. Cluster-4 : data wrangling, hadoop, spark, problem solving, data analytics
5. Cluster-5 : tableau, data visualization,powerbi, consulting, leadership
6. Cluster-6 : business acumen, experimental design, business knowledge
7. Cluster-7 : database management, algorithm development, bayesian statistics,product management, a/b testing,data storytelling, data manipulation, data modeling, industry knowledge, microsoft azure,creativity, time series analysis, neural network, cloud computing, big data technologies, nosql
8. Cluster-8 : microsoft excel, risk modeling
9. Cluster-9 : research, sql, management, presentation,data analysis, mathematics, statistics
10. Cluster-10 : python, data science, communication, computer science, machine learning, r

Since Cluster 2 and Cluster-8 contain less than 3 skills, they were not considered while designing the courses as they didn't provide any meaningful insights. Hence, there are 8 clusters with at least 3 skills in each cluster. Moreover, since Cluster-7 includes a lot of skills, I decided to create sub-clusters. Given the relatively small distance between these skills and their placement within the same cluster, I decided to select skills that share similarities and that can complement each other. Hence the sub-clusters are as follows:

1. Sub-Cluster-1: cloud computing, neural network, microsoft azure
2. Sub-Cluster-2: time series analysis, algorithm development, bayesian statistics
3. Sub-Cluster-3: big data analytics, noSQL, A/B testing, data manipulation, data modeling, database management
4. Sub-Cluster-4 : industry knowledge, product management, data storytelling, creativity

Based on the cluster results, I decided to design a course curriculum containing 11 courses as shown in Figure 9. The courses are organized in a logical order based on difficulty level starting with introductory courses, then courses with exploring advanced topics, courses with implementation of the concepts in Big Data and finally by application based courses in various domains.

- **K-Means Clustering:** Based on the K-Means clustering results with optimal elbow at 10, we got 10 clusters from which 10 courses were designed. The generated clusters are the following:

1. Cluster-1: data science, sql, research, statistical analysis, mathematics, spark, data cleaning, communication, management, presentation.
2. Cluster-2 : database management
3. Cluster-3 : product management

4. Cluster-4 : neural networks, industry knowledge, cloud computing, time series analysis, data mining, microsoft azure, a b testing, powerbi, big data technologies, nosql, data modeling, data manipulation, data visualization, tableau.
5. Cluster-5 : microsoft excel, risk modeling
6. Cluster-6 : predictive modeling, natural language processing, creativity, java
7. Cluster-7 : business knowledge, hadoop, data wrangling, experimental design, business acumen
8. Cluster-8 : team work, algorithm development, curiosity, pipelines, computer science, python, deep learning, r, machine learning, data engineering.
9. Cluster-9 : data storytelling
10. Cluster-10: bayesian statistics, data analysis, data analytics, problem solving, leadership, consulting

We observed that Cluster-2, Cluster-3, Cluster-5 and Cluster-9 contain less than 3 skills for which they were not considered due to lack of meaningful insights. In addition, we also observed that Cluster-1 and Cluster-4 contain a number of skills which were formed into sub-clusters, as was previously done for hierarchical clustering.
From Cluster-2, we have the following sub-clusters:

1. Sub-Cluster-1 : data science, data cleaning,spark, nosql
2. Sub-Cluster-2 : statistical analysis, mathematics, research
3. Sub-Cluster-3 : communication, management, presentation

From Cluster-4, we have the following sub-clusters:

4. Sub-CLuster-4 : cloud computing, neural network, microsoft azure
5. Sub-Cluster-5 : time series analysis, a/b testing, data mining, big data technologies, data modeling, data manipulation, nosql.
6. Sub-Cluster-6 : industry knowledge, powerbi,tableau, data visualization

Based on the cluster results, I follow the same logic and design a course curriculum as can be seen in Figure-10. The courses are divided into sections of Introduction containing introductory courses, Advanced Topic covering advanced topics, Advanced and Innovative that offers creative courses that employ advanced topics and require deep level understanding, and finally the Application section containing courses with real-time applications in various domains.

Part-6, Discussion and final course curriculum:

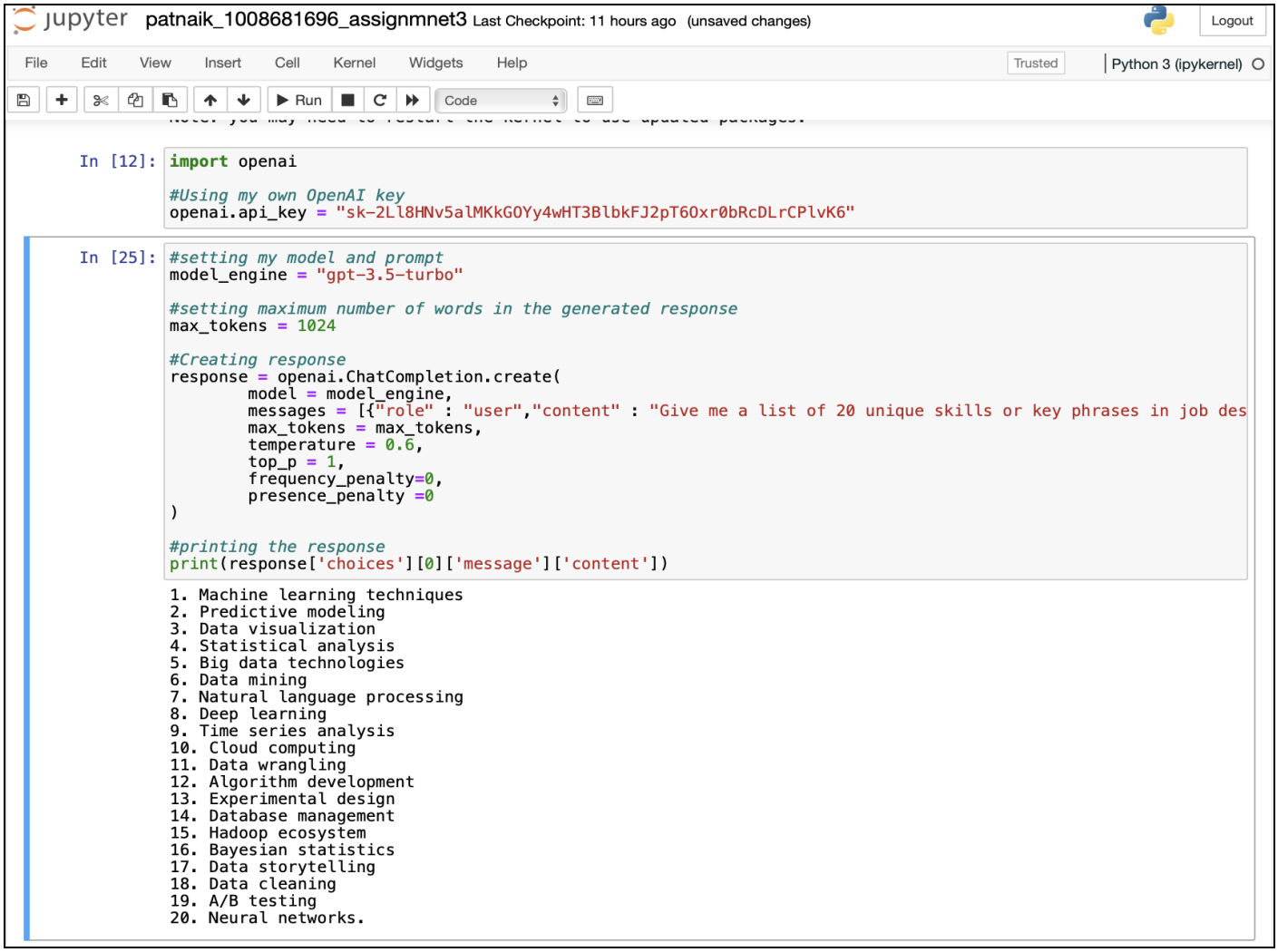
Comparing both the curriculum, we see many similar structured courses offered that can provide students with comprehensive knowledge on data science and make them proficient in the field. However, the course curriculum generated from Hierarchical Clustering covers a diverse range of topics required in the data science field at various difficulty levels. On the other hand, the course curriculum generated by K-Means offers fewer courses than the first curriculum and covers lesser topics, even though it has more in-depth knowledge based courses. This may be attributed to the fact that Hierarchical Clustering is robust in nature in which it does not require any prior knowledge of the data. Therefore it provided an accurate number of clusters. The clusters formed were also very much efficient since most of the clusters, except Cluster-7, were not too large and were reasonable. Conversely, K-Means requires a prior knowledge of data to determine the optimal number of clusters after which we could proceed to perform clustering. This is a disadvantage in a course curriculum design project because it defies the purpose of my problem statement. Hence, based on these reasonings, my final course curriculum is from Hierarchical Clustering, Figure-9.

Part-7,Bonus Point,OpenAI to describe clustering results:

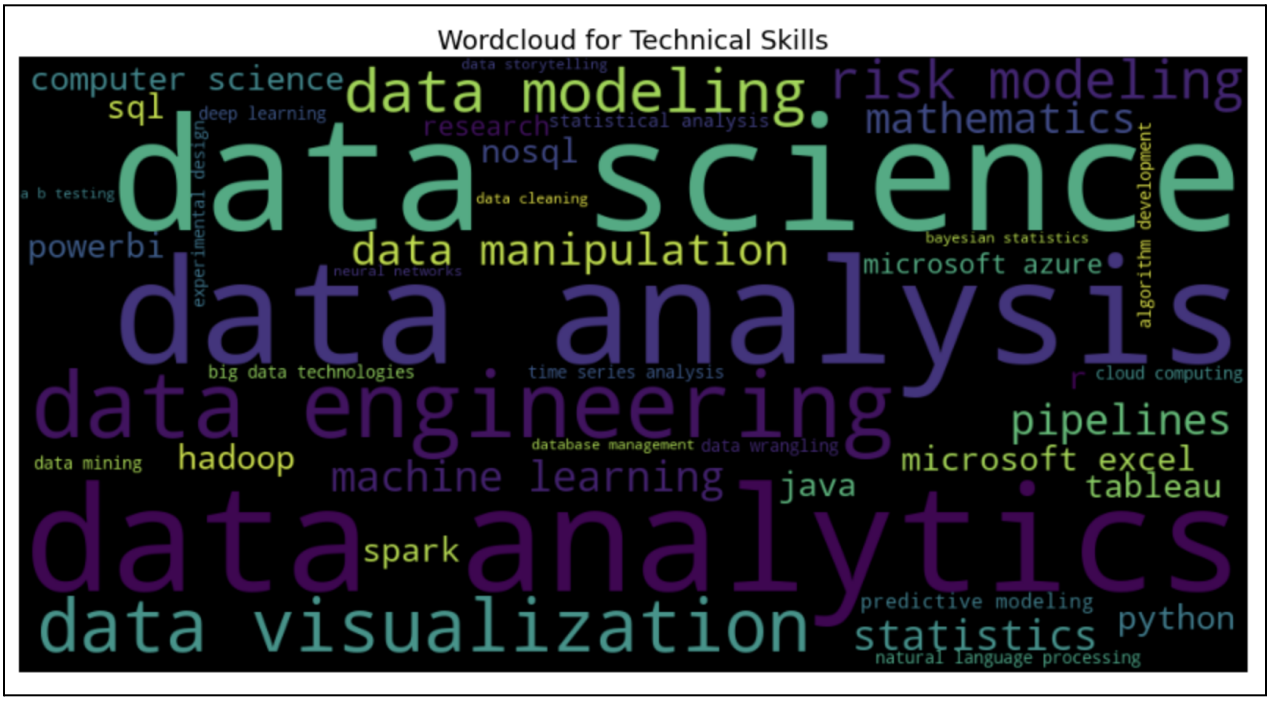
As a bonus point, I employed OpenAI to describe my clustering results for K-Means and the results can be seen in the notebook output. A snapshot of the code implementation can be seen in Figure-11. It describes various characteristics of each cluster in context of the features generated and provides a very good overview of the cluster. The generated description result is included in the Appendix under OpenAI Description Results in italics.

APPENDIX

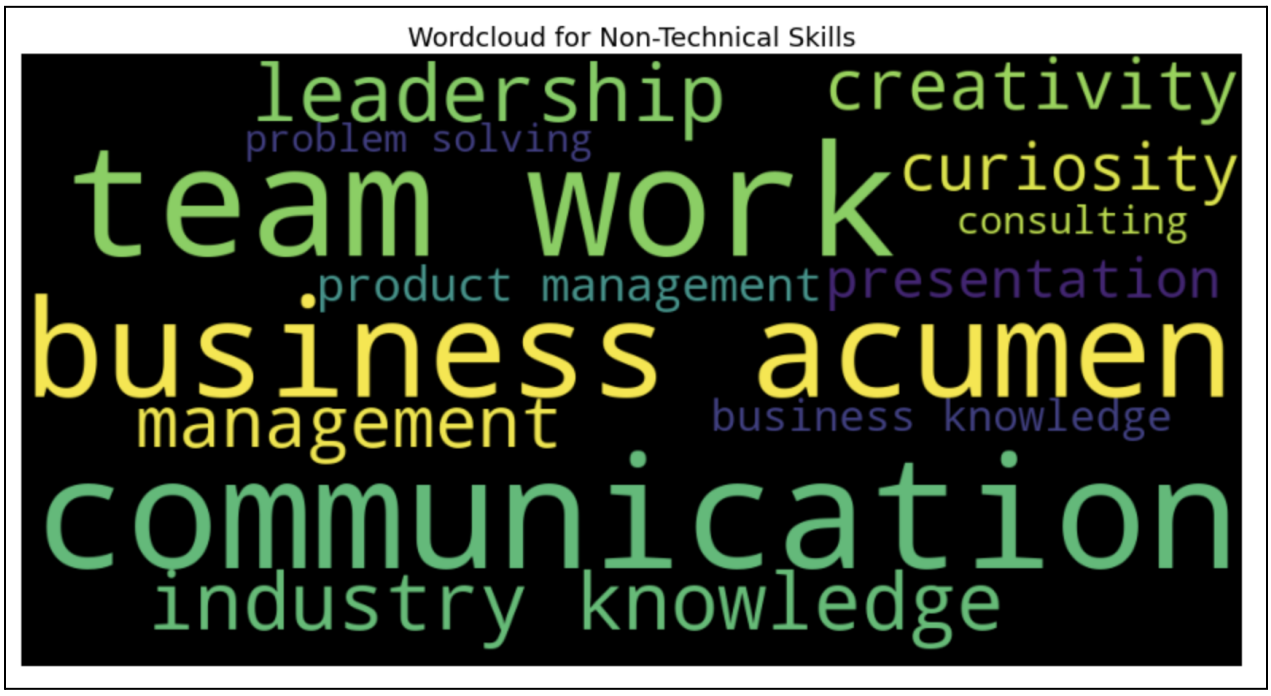
1. Figure-1: Results of skills generated by ChatGPT



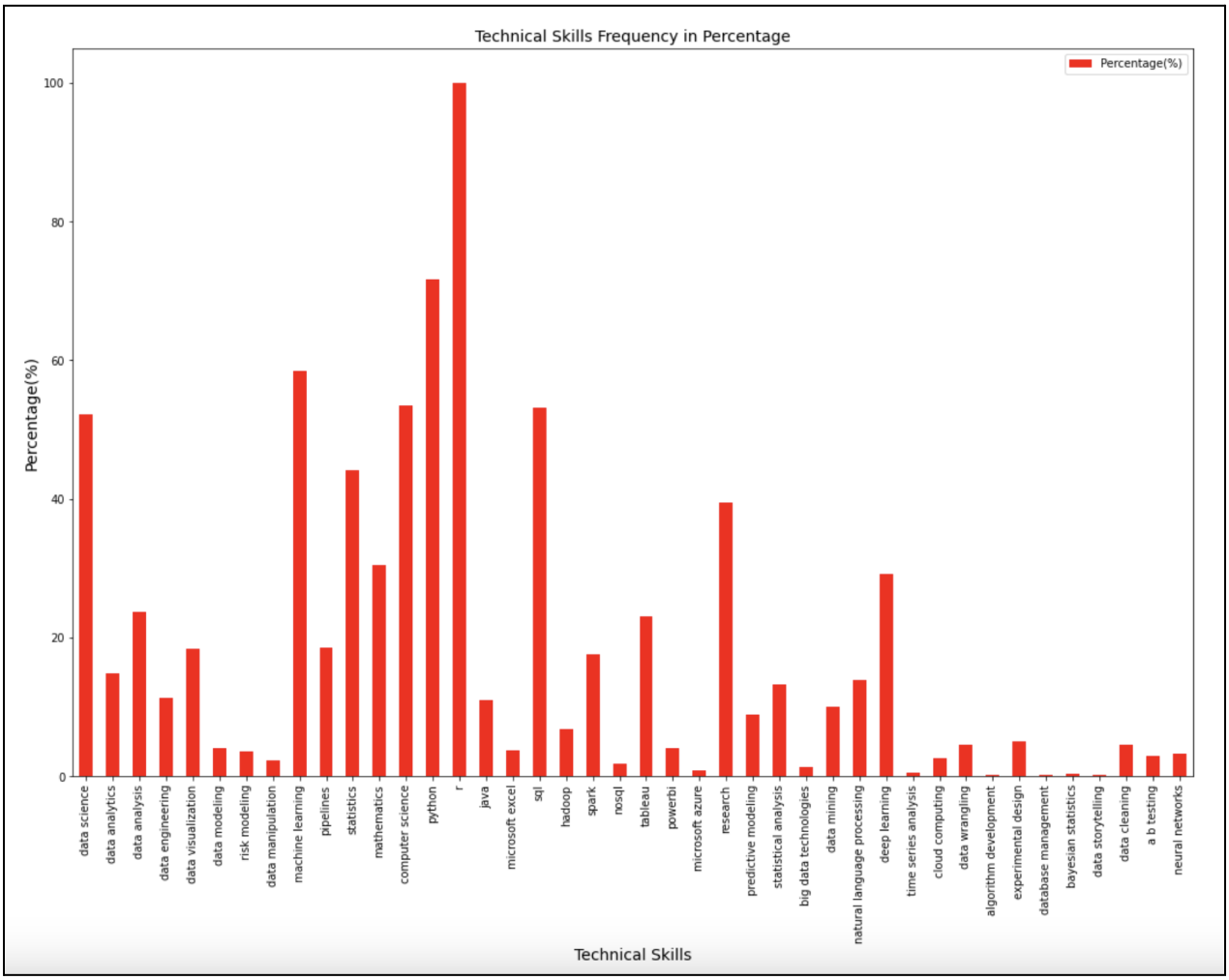
2. Figure-2: Word Cloud for Technical Skills



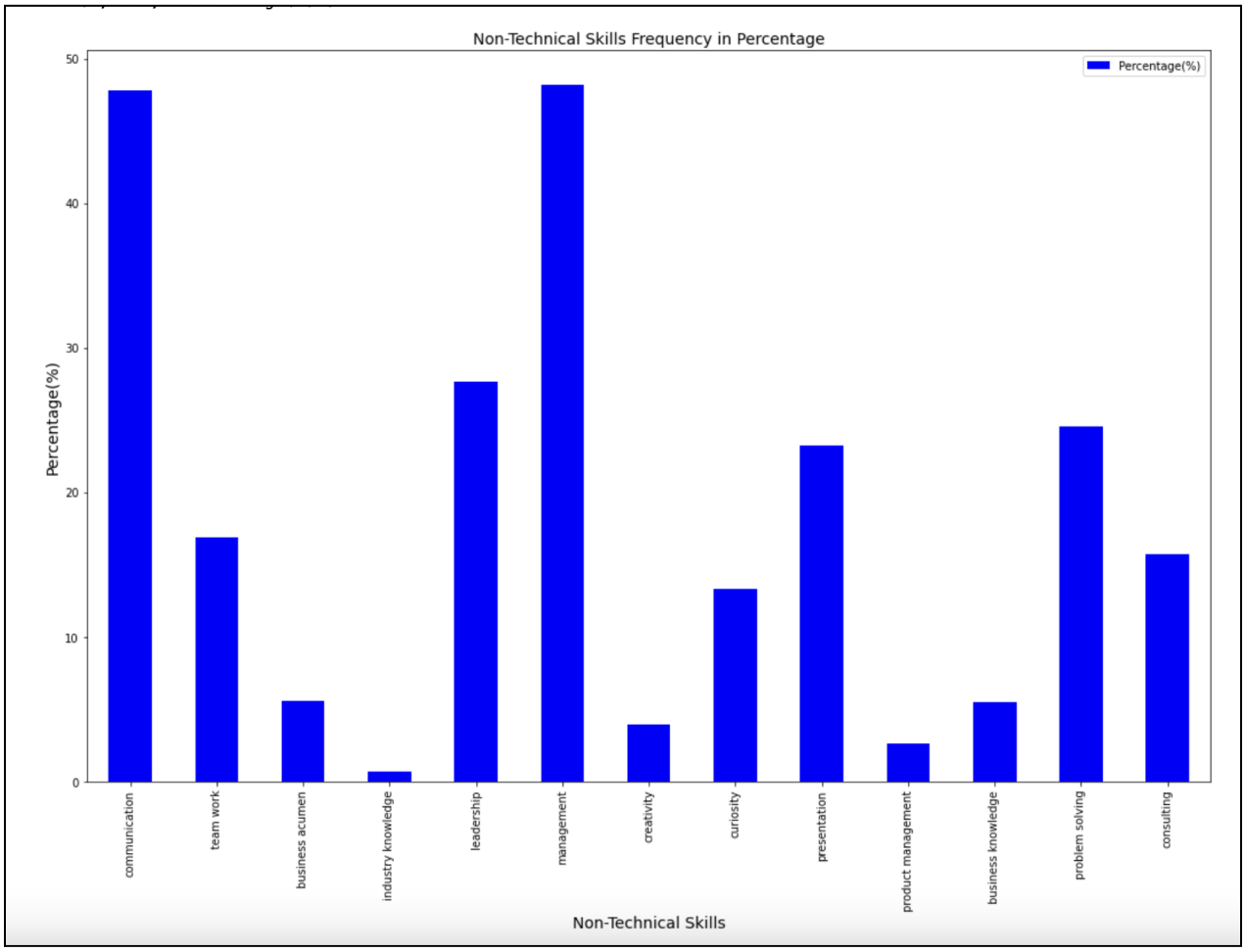
3. Figure-3: Word Cloud for Non-Technical Skills



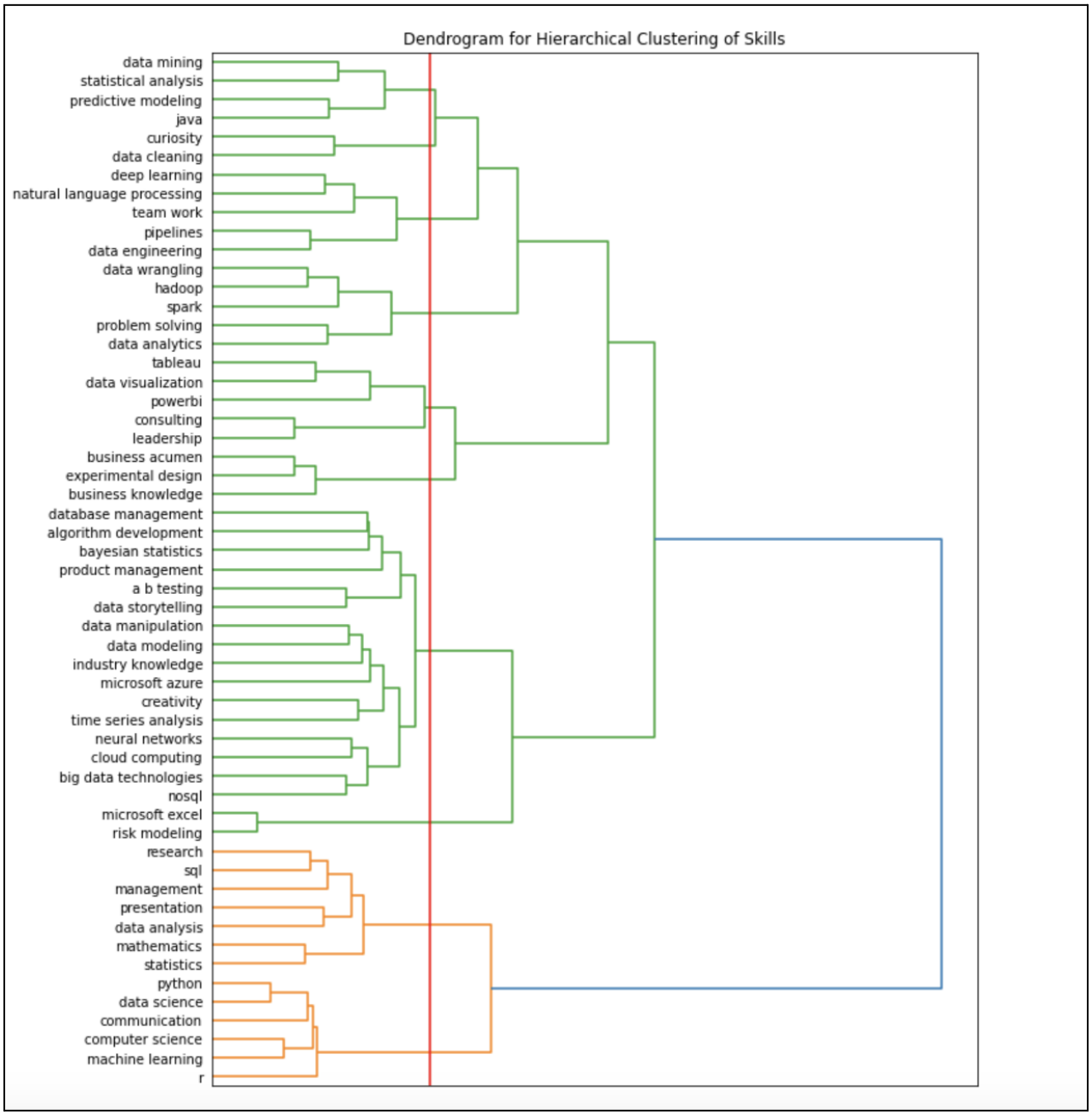
4. Figure-4: Technical Skills Frequency in Percentage



5. Figure-5 : Non-Technical Skills Frequency in Percentage



6. Figure-6: Dendrogram for Hierarchical Clustering of Skills



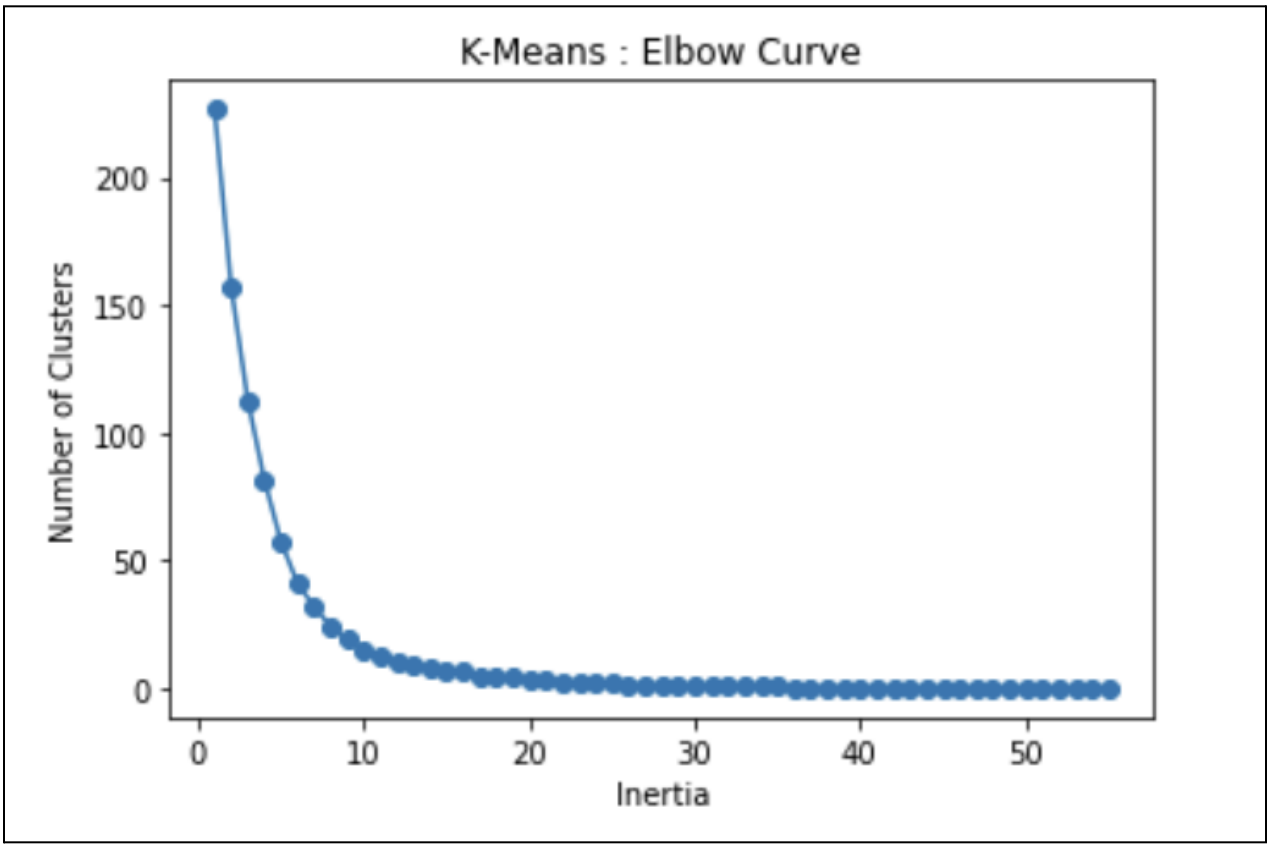
7. Table-1 : Methodology for Feature Engineering

Sl.No.	Features	Methodologies
1	Skill Frequency	Calculated the count of each skill across the job descriptions.
2	Average Salary	1. Extracted the lower and upper bound of salary range from the "Salary" column which is in string format. 2. Converted the string value to numerical value for lower and upper bound.

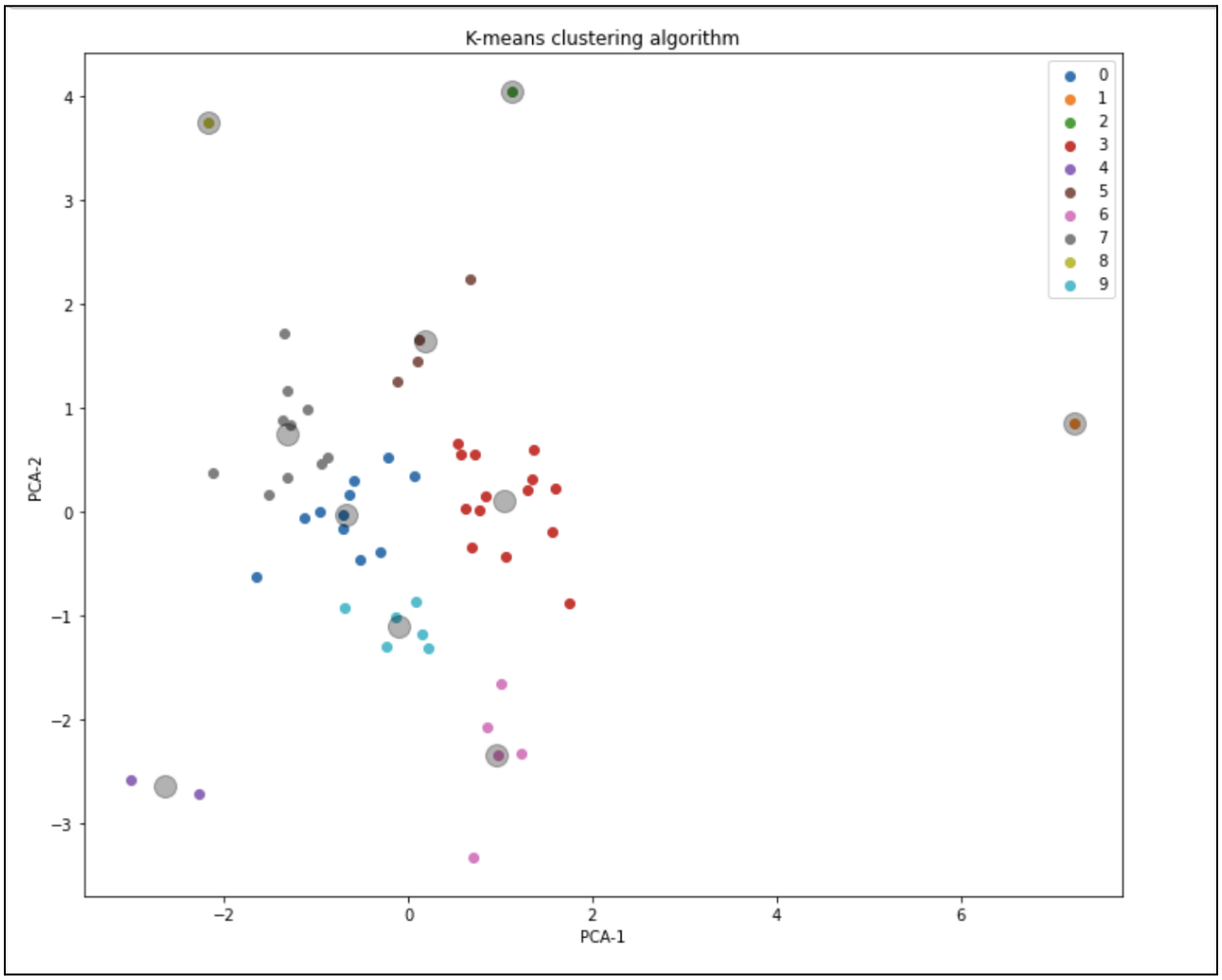
		3. Calculated the average salary for each skill.
3	Average Rating	1. Extracted the job ratings from the "Rating" column. 2. Calculated the average rating for each skill.
4	Average Work Experience	1. Extracted the number of years of work experience required from the job description of each posting and store it in a list. 2. Filtered out the list elements by removing any element value greater than 10 and replacing it with 0. This is done because upon observing the job descriptions, I found that the descriptions sometimes mention the years of establishment of that company which is often greater than 10. Overall, this is ofcourse not an accurate method but will definitely give us an approximate result. 3. Added the list as a new column of work experience to the full version of the dataframe. 4. For each skill, average work experience is then calculated by using the dataframe.
5	Average Full-Time Demand	1. Extracted the full time type of employment mentioned in the job descriptions for each job posting into a list, meaning if the job description mentions full time type, we will add 1 to the list. 2. Added the list as a new column to the full versioned dataframe. 3. Calculated the average frequency of full-time employment type jobs required for each skill using the dataframe.
6	Average Senior-Roles Demand	1. Collected job postings for senior level roles in a list, meaning that if the description is for senior level role, I'll add 1 to my list, identifying it as a senior-level-job-posting. 2. Added the list as a new feature to the full versioned dataframe. 3. Calculated the average frequency of the senior-level-job-postings for each skill.
7	Average Remote-Jobs Demand	1. Collected job postings for remote jobs in a list, meaning that if the location for the job is remote, I'll add 1 to my list, identifying it as a senior-level-job-posting. 2. Added the list as a new feature to the full versioned dataframe. 3. Calculated the average frequency of the remote jobs

		requirement for each skill.
8	Average Master's Degree Demand	1. Collected job postings for master's degree in a list, meaning that if the location for the job is remote, I'll add 1 to my list, identifying it as a senior-level-job-posting. 2. Added the list as a new feature to the full versioned dataframe. 3. Calculated the average frequency of the master's degree requirement for each skill.
9	Average Demand in Finance	1. Collected job postings for finance domain in a list, meaning that if the job is in finance, I'll add 1 to my list. 2. Added the list as a new feature to the full versioned dataframe. 3. Calculated the average frequency of skill in the finance domain.
10	Average Demand in Healthcare	1. Collected job postings for healthcare domain in a list, meaning that if the job is in healthcare, I'll add 1 to my list. 2. Added the list as a new feature to the full versioned dataframe. 3. Calculated the average frequency of skill in the healthcare domain.

8. Figure-7: Elbow Curve



9. Figure-8: K-Means Clustering



10. Figure-9 : Course Curriculum based on Hierarchical Clustering.

----INTRODUCTION--------

1. **Course-1 (Pre-requisite Course for all)** (from Cluster-10) : **Introduction to Data Science and Machine Learning** (Course will provide an introduction to fundamental concepts in data science and machine learning concepts using Python and R. This will cover general areas of computer science required in analytics and will also include data communication.)
2. **Course-2** (from Cluster-1) : **Data Mining and Predictive Modelling in Java** (Course will teach the fundamnetal concepts of data mining, statistical analysis and predictive modelling using Java language)
3. **Course-3** (from Cluster-5): **Data Visualization for Management** (Course will be designed for data visualization and consulting skills. It will teach techniques of data visulaization using Tableau and PowerBI, and will require students to derive insights for decision making. The course will also cover key areas required in consulting and leadership. The skills can be then put into test for implementation by including a Industry Capstone project.

------ADVANCED TOPICS-------

4. **Course-4**(from Cluster-3) : **Applications of Deep Learning and Natural Language Processing in Data Engineering** (Course will teach core techniques in deep learning and natural language processing, and demonstrate their applications in data engineering. The course will test team wok by including a group project component where students will work with text data to to build pipelines for data engineering using the deep learning model and NLP techniques.
5. **Course-5**(from Sub-Cluster-1): **Cloud based Data Science** (Provides an introduction to cloud based neural network modelling using Microsoft Azure
6. **Course-6** (from Sub-Cluster-2): **Advanced Data Science for Time Series Analysis** (Course provides an in-depth approach on various aspects of time series analysis including algorithm development, model development and bayesian statistics)

------BIG DATA---------

7. **Course-7 (Pre-requisite for Course-8)** (from Cluster-4): **Data Wrangling in Big Data Analytics** (Course will introduce data wrangling techniques for big data problems using platforms of Hadoop and Spark)
8. **Course-8** (from Sub-Cluster-3) : **Big Data Management with NoSQL** (The course will cover the fundamental principles and practices of big data management using NoSQL and will also demonstrate the importance of A/B testing to derive decisions).

-----APPLICATIONS---------

9. **Course-9** (from Cluster-9) : **Data Science for Research and Management** (Course will cover a comprehensive introduction to data science fundamentals in research and management settings that will include SQL, mathematics and statistics. The course will also include a presentation component to showcase the understanding of the concepts.
10. **Course-10** (from Cluster-6): **Data Science for Business Strategies** (Course will teach the applictaion of data science in business problems that will include experimental design of strategies, data analysis and business acumen. The skills can be put into test for implementation by inclusing a business oriented Final Project.)
11. **Course-11** (from Sub-Cluster-4) : **Data Science for Industry and Product Management** (Course will apply data science techniques in industry and product management settings and will enhance presentation & data storytelling approach).

11. Figure-10 : Course Curriculum based on K-Means Clustering.

-------INTRODUCTION-----------

1. **Course-1 (pre-requisite for all)** (from Sub-Cluster-2): **Mathematical Foundations for Data Science.** (Covers the mathematical and statistical foundations behind data science and doing a literature review.)
2. **Course-2** (from Sub-Cluster-6): **Introduction to Data Visualization** (Provides an introduction to data visulaization and general visualization practices across various industries.)

------ADVANCED TOPICS------------

3. **Course-3** (from Sub-Cluster-1): **Advanced Data Science with Spark and SQL.** (Covers advanced data science concepts with data cleaning and large scale machine learning models using SQL and Spark databases.)
4. **Course-4** (from Sub-Cluster-4): **Cloud based Data Science** (Provides an introduction to cloud based neural network modelling using Microsoft Azure.)
5. **Course-5 (pre-requisite for Course-6)** (from Cluster-8): **Advanced Data Science Pipeline and Engineering** (Provides an in-depth understanding of pipelines and their applications in engineering. It covers data engineering and data cleaning on deep learning models using python and R. It will also cover algorithm development and fundamnetaks of computer science.The course will include a team project to foster team work and curiousity.)

------ADVANCED & INNOVATIVE--------

6. **Course-6** (from Sub-Cluster-5): **Time Series Analysis in Big Data** (Provides an advanced introduction to time series analysis using big data that will include data mining, data manipulation and data modelling followed by demonstrating the importance of alpha-beta testing)
7. **Course-7** (from Cluster-6): **Creative Predictive Modelling using Java and NLP** (Aim to teach how to creatively think and develop predictive modelling using Java and NLP techniques.)

------APPLICATIONS-----------

8. **Course-8** (from Sub-Cluster-3): **Data Science in Management and Communication.** (Designed to manage data science projects and effectively communicating complex data science results to non-technical stakeholders.)
9. **Course-9** (from Cluster-7): **Data Science for Business Analytics.** (Aims to master the data science skills required in business analytics that covers techniques of data wrangling, experimental design and uses large dataset using Hadoop.)
10. **Course-10** (from Cluster-10): **Data Science in Leadership and Consulting** (Aim to teach solving complex problems using bayesian statistics and deriving data-driven insights for decision making in leadership and consulting roles.)

12. Figure-11 : OpenAI description of clustering results

In this part, I will be suing ChatGPT to describe results from K-Means clustering.

In [66]:
```
#Using my own OpenAI key
openai.api_key = "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"

descriptions = []

#setting my model and prompt
model_engine = "gpt-3.5-turbo"

#setting maximum number of words in the generated response
max_tokens = 500

for i in range(0,10):
    cluster = df_kmeans[df_kmeans["kmeans_cluster"] == i]
    cluster = cluster.drop("kmeans_cluster", axis=1)

    description = openai.ChatCompletion.create(
        model = model_engine,
        messages = [{"role" : "user","content" : f"Describe the common characteristics of the skills in cluste
        max_tokens= max_tokens,
        temperature = 0.6,
        top_p = 1,
    )
    descriptions.append(description['choices'][0]['message']['content'])

# Print the descriptions
for i, desc in enumerate(descriptions):
    print(f"Cluster {i+1} description: {desc}")
    print(" ")
```

```
Cluster 1 description: Cluster 1 skills have an average frequency of 470 and an average salary of $72,496. Th
ey have an average rating of 2.5 and an average experience of 1.8 years. The demand for these skills is high
with an average full-time demand of 0.3 and an average senior-roles demand of 0.3. The demand for remote jobs
is moderate with an average demand of 0.1. These skills have a moderate demand for master's degree holders wi
th an average demand of 0.3. The demand for these skills is moderate in finance and healthcare industries wit
h an average demand of 0.06 and 0.23, respectively. Overall, these skills are in high demand and require mode
rate experience and education. They also have moderate to high demand in various industries.

Cluster 2 description: Based on the given data, the common characteristics of the skills in cluster 2 are:

1. Frequency: The skills in cluster 2 appear with a frequency of 2, which means that there are only two skill
s that belong to this cluster.

2. Average Salary: The average salary for the skills in cluster 2 is 43750.0, which is relatively low compare
d to other clusters.

3. Average Rating: The skills in cluster 2 have an average rating of 4.2, which indicates that they are highl
y rated by employers and job seekers.

4. Average Experience: The skills in cluster 2 require an average of 1.0 years of experience, which means tha
t they are entry-level skills.
```

13. OpenAI Descriptions Results

Cluster 1 description: Cluster 1 skills have an average frequency of 470 and an average salary of $72,496. They have an average rating of 2.5 and an average experience of 1.8 years. The demand for these skills is high with an average full-time demand of 0.3 and an average senior-roles demand of 0.3. The demand for remote jobs is moderate with an average demand of 0.1. These skills have a moderate demand for master's degree holders with an average demand of 0.3. The demand for these skills is moderate in finance and healthcare industries with an average demand of 0.06 and 0.23, respectively. Overall, these skills are in high demand and require moderate experience and education. They also have moderate to high demand in various industries.

Cluster 2 description: Based on the given data, the common characteristics of the skills in cluster 2 are:

1. Frequency: The skills in cluster 2 appear with a frequency of 2, which means that there are only two skills that belong to this cluster.

2. Average Salary: The average salary for the skills in cluster 2 is 43750.0, which is relatively low compared to other clusters.

3. Average Rating: The skills in cluster 2 have an average rating of 4.2, which indicates that they are highly rated by employers and job seekers.

4. Average Experience: The skills in cluster 2 require an average of 1.0 years of experience, which means that they are entry-level skills.

5. Average Full-Time Demand: The skills in cluster 2 have an average full-time demand of 1.0, which indicates that they are in high demand for full-time positions.

6. Average Senior-Roles Demand: The skills in cluster 2 have an average senior-roles demand of 0.0, which means that they are not typically required for senior-level positions.

7. Average Remote-Jobs Demand: The skills in cluster 2 have an average remote-jobs demand of 0.0, which indicates that they are not commonly available as remote work opportunities.

8. Average Master's Degree Demand: The skills in cluster 2 have an average master's degree demand of 1.0, which means that they are typically required for positions that require a master's degree.

9. Average Demand in Finance: The skills in cluster 2 have an average demand of 0.5 in finance, which means that they are not typically required for finance-related positions.

10. Average Demand in Healthcare: The skills in cluster 2 have an average demand of 0.0 in healthcare, which means that they are not typically required for healthcare-related positions.

Cluster 3 description: The skills in cluster 3 have the following common characteristics:

1. High average salary: The skills in this cluster have an average salary of $133,474. This suggests that these skills are in demand and are likely to be high-paying.

2. Moderate average rating: The skills in this cluster have an average rating of 3.6, which suggests that they are rated positively by employers and employees.

3. Moderate average experience: The skills in this cluster have an average experience of 0.4 years, suggesting that they are relatively new skills.

4. High full-time demand: The skills in this cluster have a high demand for full-time roles, with an average demand of 0.9.

5. Low senior-roles demand: The skills in this cluster have a low demand for senior roles, with an average demand of 0.

6. Low remote-jobs demand: The skills in this cluster have a low demand for remote jobs, with an average demand of 0.1.

7. Low master's degree demand: The skills in this cluster have a low demand for master's degree holders, with an average demand of 0.06.

8. Low demand in finance: The skills in this cluster have a low demand in finance, with an average demand of 0.1.

9. Low demand in healthcare: The skills in this cluster have a low demand in healthcare, with an average demand of 0.03.

Cluster 4 description: The skills in cluster 4 have the following common characteristics:

1. Average salary: The average salary in this cluster is around $50,000, which is relatively lower compared to other clusters.

2. Average rating: The average rating of the skills in this cluster is around 2.5, which suggests that they are moderately rated by employers.

3. Average experience: The average experience required for these skills is around 1-2 years, which suggests that they are entry-level or mid-level skills.

4. Full-time demand: The demand for these skills in full-time positions is moderate, with an average demand of around 30%.

5. Senior-roles demand: The demand for these skills in senior-level positions is relatively low, with an average demand of around 20%.

6. Remote-jobs demand: The demand for these skills in remote jobs is relatively low, with an average demand of around 10%.

7. Master's degree demand: The demand for these skills for candidates with a master's degree is moderate, with an average demand of around 30%.

8. Finance demand: The demand for these skills in finance is relatively low, with an average demand of around 5%.

9. Healthcare demand: The demand for these skills in healthcare is relatively low, with an average demand of around 20%.

Overall, the skills in cluster 4 are entry-level or mid-level skills with moderate demand for full-time positions, moderate demand for candidates with a master's degree, and relatively low demand for senior-level positions, remote jobs, finance, and healthcare.

Cluster 5 description: Cluster 5 skills are characterized by a moderate frequency, average salary, and experience level. These skills have an average rating and are in moderate demand for full-time and senior roles. They have a relatively high demand for remote jobs and a low demand for a master's degree. In terms of industry demand, these skills have a moderate demand in finance and healthcare.

Cluster 6 description: The common characteristics of the skills in cluster 6 are:

1. Average salary is relatively high compared to other clusters.
2. Average rating is moderate.
3. Average experience required is relatively low compared to other clusters.

4. Average full-time demand is moderate.
5. Average senior-roles demand is moderate.
6. Average remote-jobs demand is low.
7. Average master's degree demand is moderate.
8. Average demand in finance is low.
9. Average demand in healthcare is moderate.

Overall, the skills in this cluster are high-paying, require moderate experience, and have moderate demand for senior roles and full-time positions. However, they have low demand for remote jobs and are not strongly associated with finance or healthcare industries.

Cluster 7 description: Based on the given data, the common characteristics of the skills in cluster 7 are:

1. Average experience: The skills in this cluster require an average of 2-4 years of experience.

2. High demand for full-time roles: The average demand for full-time positions is higher than other types of employment.

3. High demand for senior roles: The average demand for senior-level positions is relatively high.

4. Low demand for remote jobs: The average demand for remote jobs is relatively low.

5. High demand for master's degree: The average demand for candidates with a master's degree is relatively high.

6. High demand in finance: The average demand for these skills is higher in the finance industry.

7. High demand in healthcare: The average demand for these skills is also relatively high in the healthcare industry.

8. Average salary: The average salary for these skills is relatively low, ranging from $13,000 to $51,000 per year.

9. Average rating: The average rating for these skills is around 1.5-3.5, indicating average to above-average performance.

Overall, the skills in cluster 7 require some level of experience and education, and are in relatively high demand in the finance and healthcare industries. However, the average salary for these skills is not very high, and the demand for remote jobs is relatively low.

Cluster 8 description: The common characteristics of the skills in cluster 8 are:

1. High average salary: The skills in this cluster have an average salary ranging from $70,767 to $108,547, which is higher than the overall average salary.

2. Moderate to high demand: The skills in this cluster have an average full-time demand ranging from 0.26 to 0.35, indicating moderate to high demand in the job market.

3. Moderate experience required: The skills in this cluster have an average experience ranging from 1.31 to 2.91 years, indicating that they require moderate experience.

4. Low to moderate senior-roles demand: The skills in this cluster have an average senior-roles demand ranging from 0.25 to 0.53, indicating low to moderate demand for senior positions.

5. Low to moderate remote-jobs demand: The skills in this cluster have an average remote-jobs demand ranging from 0.03 to 0.23, indicating low to moderate demand for remote jobs.

6. Moderate to high demand in finance: The skills in this cluster have an average demand in finance ranging from 0.01 to 0.06, indicating moderate to high demand in the finance industry.

7. Low to moderate demand in healthcare: The skills in this cluster have an average demand in healthcare ranging from 0.05 to 0.06, indicating low to moderate demand in the healthcare industry.

8. Low to moderate average rating: The skills in this cluster have an average rating ranging from 1.68 to 3.70, indicating low to moderate satisfaction among employees in these roles.

Cluster 9 description: Unfortunately, as an AI language model, I cannot provide the common characteristics of the skills in cluster 9 without additional information. Please provide more context or data.

Cluster 10 description: Based on the given data, the common characteristics of the skills in cluster 10 are:

1. Average Salary: The average salary for skills in cluster 10 is between $35,058 and $54,454, which is relatively high compared to other clusters.

2. Average Rating: The average rating for skills in cluster 10 is between 2.23 and 3.43, indicating that these skills are generally well-regarded by employers.

3. Average Experience: The average experience for skills in cluster 10 is between 1.74 and 2.60 years, suggesting that these skills are more suitable for mid-level professionals.

4. Average Full-Time Demand: The average full-time demand for skills in cluster 10 is between 0.09 and 0.22, indicating that these skills are not in high demand for full-time positions.

5. Average Senior-Roles Demand: The average demand for senior roles for skills in cluster 10 is between 0.16 and 0.35, suggesting that these skills are more suitable for mid-level positions.

6. Average Remote-Jobs Demand: The average demand for remote jobs for skills in cluster 10 is between 0.04 and 0.14, indicating that these skills are not in high demand for remote positions.

7. Average Master's Degree Demand: The average demand for a master's degree for skills in cluster 10 is between 0.25 and 0.47, suggesting that having a master's degree may be beneficial for these skills.

8. Average Demand in Finance: The average demand for skills in cluster 10 in the finance industry is between 0.04 and 0.08, indicating that these skills are not in high demand in finance.

9. Average Demand in Healthcare: The average demand for skills in cluster 10 in the healthcare industry is between 0.14 and 0.47, suggesting that these skills may be more in demand in healthcare than in other industries.