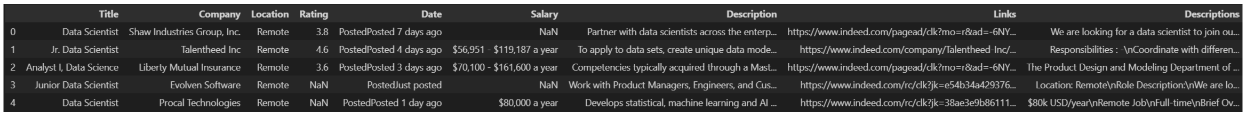
1 Data Collection and Cleaning

For first part, I web scrapped 1499 Data Scientiest job description from indeed.com, by checking the data by using isnull function, I found that the data set contains no empty features, therefore, no data cleaning is needed by far.

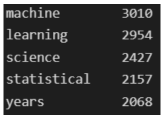


	Title	Company	Location	Rating	Date	Salary	Description	Links	Descriptions
0	Data Scientist	Shaw Industries Group, Inc.	Remote	3.8	PostedPosted 7 days ago	NaN	Partner with data scientists across the enterp...	https://www.indeed.com/pagead/clk?mo=r&ad=-6NY...	We are looking for a data scientist to join ou...
1	Jr. Data Scientist	Talentheed Inc	Remote	4.6	PostedPosted 4 days ago	$56,951 - $119,187 a year	To apply to data sets, create unique data mode...	https://www.indeed.com/company/Talentheed-Inc/...	Responsibilities : -\nCoordinate with differen...
2	Analyst I, Data Science	Liberty Mutual Insurance	Remote	3.6	PostedPosted 3 days ago	$70,100 - $161,600 a year	Competencies typically acquired through a Mast...	https://www.indeed.com/pagead/clk?mo=r&ad=-6NY...	The Product Design and Modeling Department of ...
3	Junior Data Scientist	Evolven Software	Remote	NaN	PostedJust posted	NaN	Work with Product Managers, Engineers, and Cus...	https://www.indeed.com/rc/clk?jk=e54b34a429376...	Location: Remote\nRole Description:\nWe are lo...
4	Data Scientist	Procal Technologies	Remote	NaN	PostedPosted 1 day ago	$80,000 a year	Develops statistical, machine learning and AI ...	https://www.indeed.com/rc/clk?jk=38ae3e9b86111...	$80k USD/year\nRemote Job\nFull-time\nBrief Ov...

Figure 1: web scrapped result

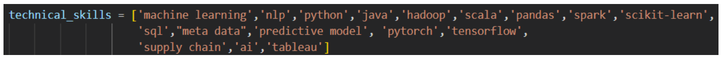
2 Exploratory data analysis and feature Engineering

For the skills extraction process, the idea is extract the common words in the job descriptions, since if a skill is important in this job field, then this skills must be mentioned most of the time in the description. However, in the frequency analysis, I found that the most common words are something like ”I”, ”we”, ”this” etc. Therefore, before I actually doing skills extraction, I have to remove those ”useless” words, and this can accomplished by importing stopwords from NLTK package. After removing common words and punctuation, I analysis the description by extract the most common words, the skills-like feature becomes appear. Then I can define some technical skills based on those common word and also our common sense. since the python can not understand the human language, so I have to express the sentence in the vector form, TfidVectorized did that for us. After convert job description and self-built library in to vector form, I can measure the similarity between them, if their similarity is higher than a specific number(Threshold), we can say that this job acquire this skills and we will append 1 to denote this condition. To maintain a higher accuracy, I instead of put all the extracted skills, I create an another library use to store the soft/business skills. To obtain the soft skills, I extract the adjective from the first 50 common words, since most of the soft skills are expressed as adjective, for example: be responsible, creative, those words are all adjective. Then apply TF-IDF to measure their importance,By observing the following figure, we can say that the basic technical skills every data scientist has to acquire are Python, Machine-Learning, SQL and the soft skills are communication, responsible and team work. Those extracted skills obeys the human logic.



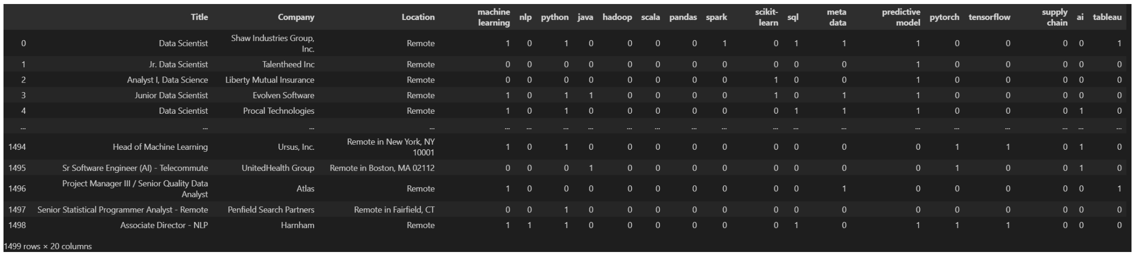
```
machine        3010
learning       2954
science        2427
statistical    2157
years          2068
```

Figure 2: Top 5 Common Words



```
technical_skills = ['machine learning','nlp','python','java','hadoop','scala','pandas','spark','scikit-learn',
                    'sql',"meta data",'predictive model', 'pytorch','tensorflow',
                    'supply chain','ai','tableau']
```

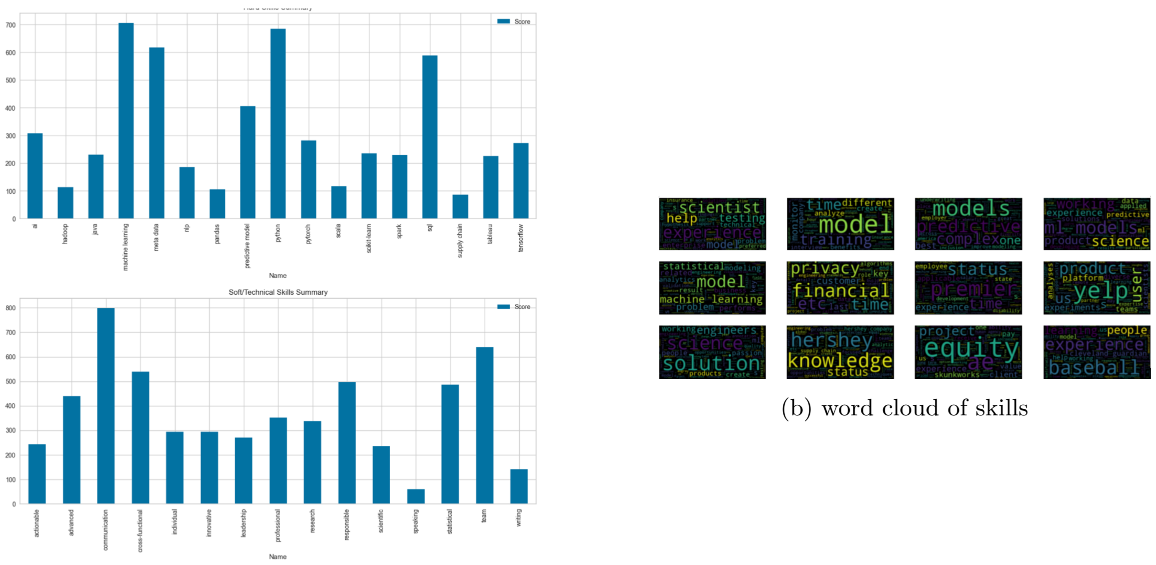
Figure 3: Self-built Technical Skills



	Title	Company	Location	machine learning	nlp	python	java	hadoop	scala	pandas	spark	scikit-learn	sql	meta data	predictive model	pytorch	tensorflow	supply chain	ai	tableau
0	Data Scientist	Shaw Industries Group, Inc.	Remote	1	0	1	0	0	0	0	1	0	1	1	1	0	0	0	0	1
1	Jr. Data Scientist	Talentheed Inc	Remote	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
2	Analyst I, Data Science	Liberty Mutual Insurance	Remote	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
3	Junior Data Scientist	Evolven Software	Remote	1	0	1	1	0	0	0	0	1	0	1	1	0	0	0	0	0
4	Data Scientist	Procal Technologies	Remote	1	0	1	0	0	0	0	0	0	1	1	1	0	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1494	Head of Machine Learning	Ursus, Inc.	Remote in New York, NY 10001	1	0	1	0	0	0	0	0	0	0	0	0	1	1	0	1	0
1495	Sr Software Engineer (AI) - Telecommute	UnitedHealth Group	Remote in Boston, MA 02112	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0
1496	Project Manager III / Senior Quality Data Analyst	Atlas	Remote	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
1497	Senior Statistical Programmer Analyst - Remote	Penfield Search Partners	Remote in Fairfield, CT	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1498	Associate Director - NLP	Harnham	Remote	1	1	1	0	0	0	0	0	0	1	0	1	1	1	0	0	0

1499 rows × 20 columns

Figure 4: logically formatted Technical Skills



(b) word cloud of skills

(a) hard/soft skills vs company

3 Hierarchical Clustering Implementation

To implement hierachy clustering, I use the cluster.hierarchy from scipy package, and then I used linkage method to build up the dendrogram, below is the dendrogram generated for all skills. Based on this plot, I decided to have 9 clusters, since we are designing 8-12 courses, so 10 clusters is reasonable, each cluster is a course, and the content of the course is the skills inside each cluster. For example, for the cluster, I would like to say it contain the skills like nlp,pytorch, tensorflow, predictive mode, research. Since all the skills contains is around data, building a model and predict, therefore, I like to say the designed course name is machine learning.

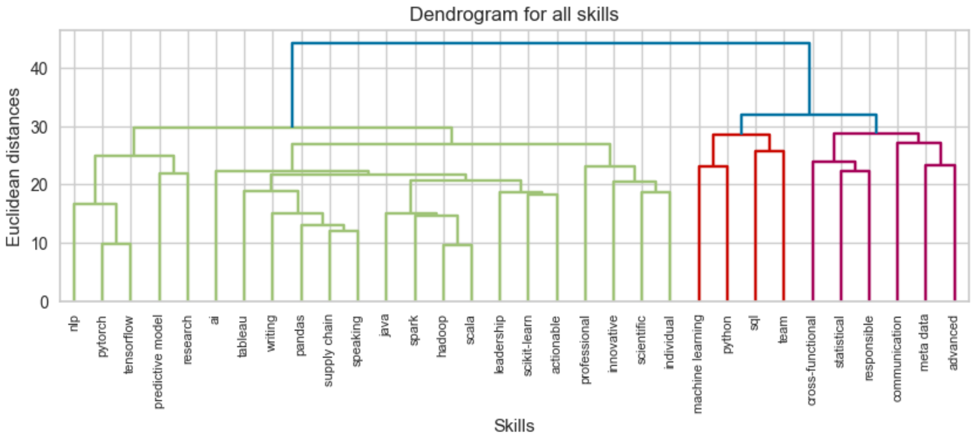
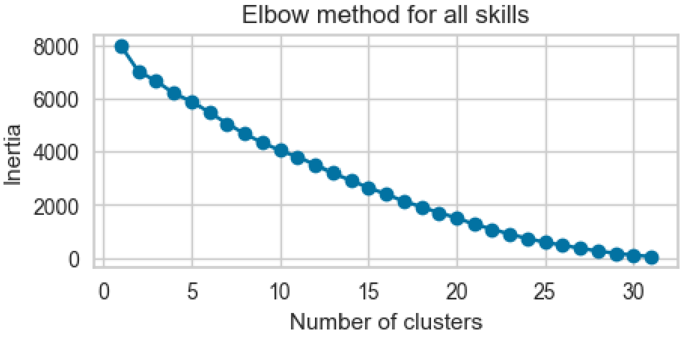


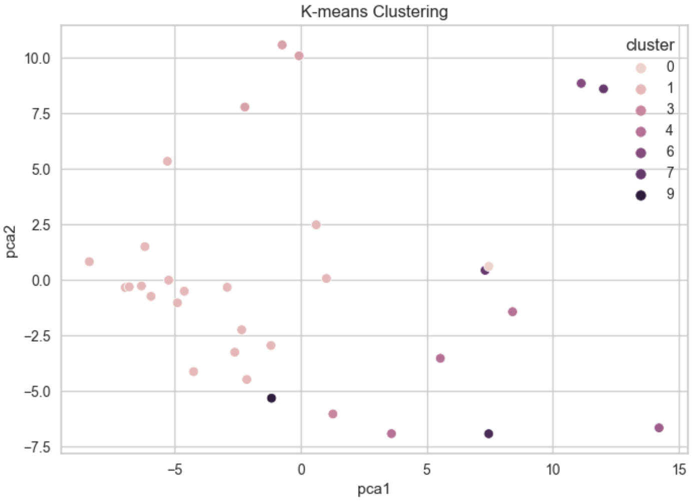
Figure 6: dendrogram for all skills

4 K-means or DBSCAN clustering implementation:

For this part, we can use KMeans function in sklearn.cluster, since we have to decide the best number of cluster, therefore, I implement a for loop to try every single cluster, then plot a diagram based on the interia. Then, we can decide how many cluster by using elbow method. From the generated plot, I would like to say the elbow points happens when cluster equals to 9, and the cluster number approximately obey the analysis from the previous part. Since the features in the dataset is multi-dimensional, therefore we need to first apply PCA to do the dimension reduction to plot scatter plot



(a) elbow method

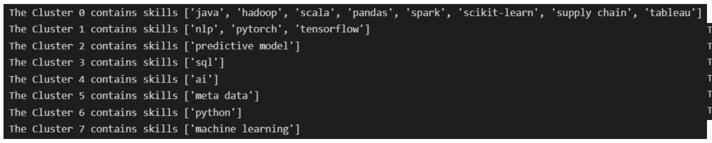


(b) kmeans visualization

5 Interpretation of results, discussion and final course curriculum:

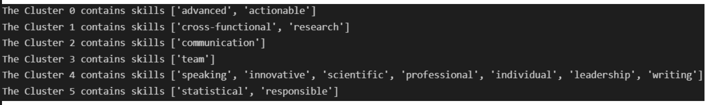
In this part, I attach the plot which shows what skills in each cluster, the result greatly follows the logic, but need a little bit modification. Based on three plots shows, I would to design 9 courses as follows: Introduction to AI, Introduction to Machine-Learning, Data analysis, Introduction to Business, Deep Learning, Programming with different language, Supply Chain, Big Data Analysis and Innovative Thinking.

Setting the course of Introduction to AI is based on hard skill cluster 5. Setting the course Introduction to Machine Learning is based on hard skill cluster number 0, Data Analysis is based on all skill cluster 0 ,5 and cluster 8 which contains the skills like statistical and cross-functional and SQL, and the course Introduction to Business is based on soft skills cluster 2 which need student to do research and writing essay about some successful company , presentation, leadership. The next course Deep Learning is based on hard skills cluster 2, which need students to know how to use TensorFlow. The course Programming with different language is based on hard skills cluster 1, student who enroll this course should know and use programming language such as Python, Java, etc and solving problems by using some pre-installed packages such as pandas, scikit-learn etc. In addition, the course Supply chain is chosen because it also shows up in all skills cluster 1, and the course Big Data analysis is based on the hard skill cluster 3,4,5,6,7, the content should include how to use and process meta data and using some tools like predictive model or scikit-learn to do some advanced analysis. The last course prepared for the student is Innovative Thinking, student should learn the skills in soft skills cluster 1, such as independent thinking, organize the idea and state them in a scientific way, and be responsible.



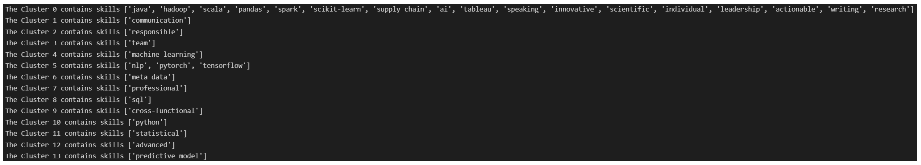
```
The Cluster 0 contains skills ['java', 'hadoop', 'scala', 'pandas', 'spark', 'scikit-learn', 'supply chain', 'tableau']
The Cluster 1 contains skills ['nlp', 'pytorch', 'tensorflow']
The Cluster 2 contains skills ['predictive model']
The Cluster 3 contains skills ['sql']
The Cluster 4 contains skills ['ai']
The Cluster 5 contains skills ['meta data']
The Cluster 6 contains skills ['python']
The Cluster 7 contains skills ['machine learning']
```

(a) soft skills cluster



```
The Cluster 0 contains skills ['advanced', 'actionable']
The Cluster 1 contains skills ['cross-functional', 'research']
The Cluster 2 contains skills ['communication']
The Cluster 3 contains skills ['team']
The Cluster 4 contains skills ['speaking', 'innovative', 'scientific', 'professional', 'individual', 'leadership', 'writing']
The Cluster 5 contains skills ['statistical', 'responsible']
```

(b) hard skills cluster



```
The Cluster 0 contains skills ['java', 'hadoop', 'scala', 'pandas', 'spark', 'scikit-learn', 'supply chain', 'ai', 'tableau', 'speaking', 'innovative', 'scientific', 'individual', 'leadership', 'actionable', 'writing', 'research']
The Cluster 1 contains skills ['communication']
The Cluster 2 contains skills ['responsible']
The Cluster 3 contains skills ['team']
The Cluster 4 contains skills ['machine learning']
The Cluster 5 contains skills ['nlp', 'pytorch', 'tensorflow']
The Cluster 6 contains skills ['meta data']
The Cluster 7 contains skills ['professional']
The Cluster 8 contains skills ['sql']
The Cluster 9 contains skills ['cross-functional']
The Cluster 10 contains skills ['python']
The Cluster 11 contains skills ['statistical']
The Cluster 12 contains skills ['advanced']
The Cluster 13 contains skills ['predictive model']
```

(c) all skills cluster