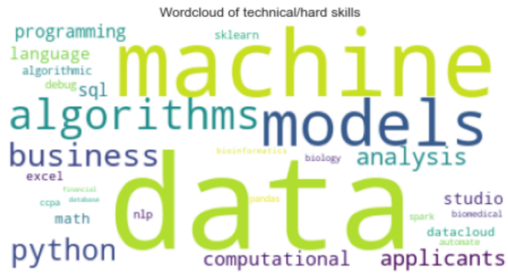
In part 1, I choose the position of Data Analyst and location in Vancouver, BC. After website scraping, there are 1289 positions in the data frame.

In part 2, I build a list of hard skills and soft skills to explore data analysis and do feature engineering. Then I create a bag of words from the columns "Descriptions" in the data frame of jobs. I select all the columns in my list of skills to create a new data frame called "dataframe_new", then combine "dataframe_new" and term frequency data frame as "data".

```
data.head()
```

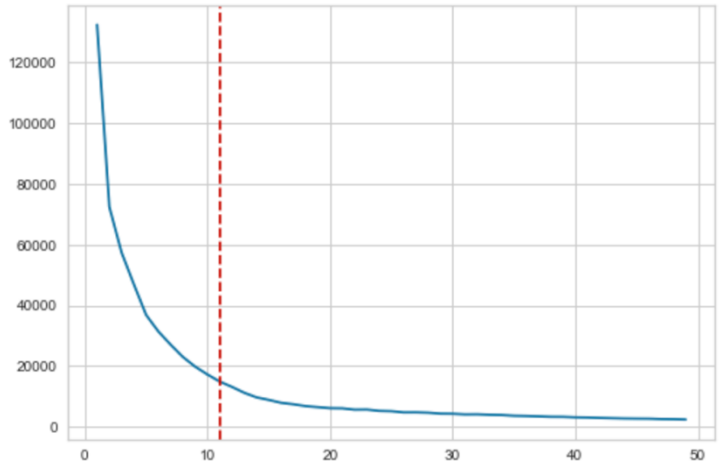
	Title	Company	Location	Rating	Salary	active	accountability	decision	planning	writing	...	analysis	algorithms	algorithmic	applicants	auto
0	Data Science Graduate (Canada)	Novo Nordisk	Canada	4.1	NaN	0	0	0	0	0	...	0	0	0	0	
1	Junior Data Analyst	Fasken SLP	Vancouver, BC	3.4	NaN	0	0	1	0	1	...	1	0	0	1	
2	Senior Performance and Data Analyst	The City of Vancouver	Remote in Vancouver, BC	3.7	102, 256–127,830 a year	0	2	3	2	0	...	6	0	0	2	
3	Senior Data Scientist	Mastercard	Vancouver, BC	4.0	NaN	0	0	0	0	0	...	0	0	0	0	
4	Data Scientist	Minpraxis Solutions Ltd.	Hybrid remote in Vancouver, BC	NaN	6, 600 – 7,500 a month	0	0	0	0	0	...	4	2	0	0	

The data frame with hard skills is from the 47th column in "data", the data frame with soft skills is from the 5th to the 46th column, and the data frame with all skills is from the 5th column in "data". Then I use the method wordcloud to show the importance of skills, which is explained by the frequency of a word in the data frame "tf_new". Plots of word could of hard skills, soft skills and all skills are shown below:



In part 3, I try to implement a hierarchical clustering algorithm. Based on the codes provided in tutorials, I show the plot of dendrograms and the vertical line = 3000. After locating the vertical line, I count how many times horizontal lines are crossed by it :10 times. So, 10 seems a good indication of the number of clusters that have the most distance between them, which is between 8 to 12 courses.

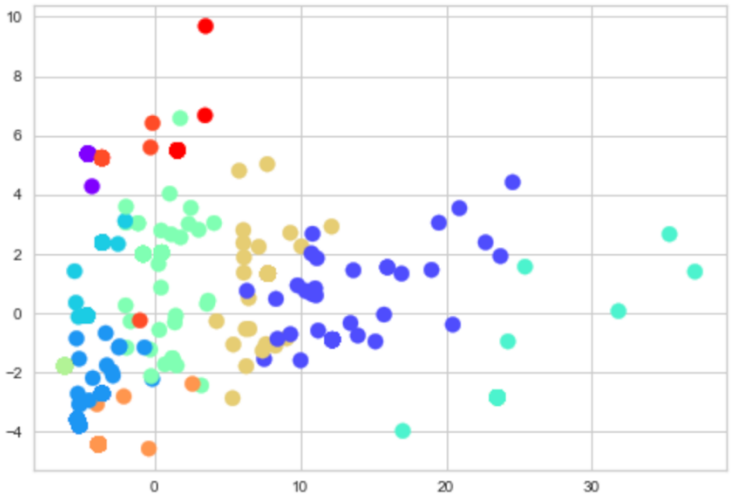
In part 4, I implement k-means clustering algorithm. I use elbow method to determine the optimal number of clusters. X axis stands for how many clusters we choose. Y axis stands for sum of squared distances. Based on the plot, as the number of clusters is 11, the sum squared distances is around 19000. This is small enough. Plot is shown below:



Thus, I decide to choose 11 clusters in part 4. Next, I add a new column of cluster number to the end of the data frame "data_all". Every row corresponding to each job is distributed to different clusters.

coordination	innovation	experimentation	...	algorithms	algorithmic	applicants	automate	bioinformatics	biology	biomedical	business	ccpa	cluster
0	0	0	...	0	0	0	0	1	1	0	0	0	1
0	0	0	...	0	0	1	0	0	0	0	5	0	1
0	0	0	...	0	0	2	0	0	0	0	18	0	1
0	1	1	...	0	0	0	0	0	0	0	3	0	1
0	0	0	...	2	0	0	0	0	0	0	0	0	7

And the clustering results are visualized as below:



In part 5, I design courses' names and topics for methods in part 3 and part 4. There are 10 courses in part 3 and 11 courses in part 4.

```
curricula_p3 = {'Course_a1': ['machine','design','models','time'],
                'Course_a2': ['computational','innovation','advanced','leadership'],
                'Course_a3': ['applicants','python','analysis'],
                'Course_a4': ['business','management','algorithms'],
                'Course_a5': ['programming','datacloud','intelligence'],
                'Course_a6': ['communication','project','writing','speaking'],
                'Course_a7': ['hadoop','studio','math'],
                'Course_a8': ['financial','presentation','coordination'],
                'Course_a9': ['communication','project','data'],
                'Course_a10': ['sql','excel','matlab','spark']}
```

```
Course1's name and topics:
Machine Learning and Algorithms
['machine', 'algorithms', 'model designing']
Course2's name and topics:
Data Analysis and Intelligence in Business field
['data analysis', 'intelligence', 'business']
Course3's name and topics:
Data Design by R studio in Math field
['R studio application', 'design model', 'train data', 'math']
Course4's name and topics:
Data and models in Machine Learning
['machine learning', 'models designing', 'data testing']
Course5's name and topics:
Data Management and Analysis
['data management', 'data analysis', 'assistance in group project']
Course6's name and topics:
Data and Models in the fields of Machine Learning and Business by python
['creating models', 'process data', 'Machine Learning', 'Business', 'python']
Course7's name and topics:
Advanced Computational Design and Innovation in field of Machine
['advanced computational design', 'innovation', 'machine']
Course8's name and topics:
Design Models to Solve Data Problems and Time Management
['models designing', 'time management', 'project']
Course9's name and topics:
Design Writing and Management Projects in field of Business
['writing', 'business project', 'management project']
Course10's name and topics:
Models in Machine Learning and Business
['machine learning', 'business', 'leadership in group', 'data process']
Course11's name and topics:
Data Analysis and Datacloud in Machine Learning
['data analysis', 'datacloud', 'time management', 'machine learning']
```

I use word cloud to visualize course with its name and topics in part 3 and part 4. The word cloud plots of course 1 are shown below:

```
# This course's name is:
Course_a1 = 'Models designing and Time Management in Machine Learning'
# This course focus on these topics:
Course_a1_topics = ['machine','design','models','time']
```

```
(1289, 4)
```


```
Course1 = 'Machine Learning and Algorithms'
```

Course 1 focus on machine, algorithms, and model designing

```
Course1_topics = ['machine', 'algorithms', 'model designing']
```

According to the dendrograms in part 3, I found many courses contain one topic only. And some courses contain lots of unrelated topics. I redesign the curriculum by myself, distribute 3 to 4 related topics to each course. Based on KNN, there are 11 clusters, which are designed by model directly. This method is more accurate and smarter than that in part 3. I found most topics are related with each other in each course in part 4, but there are some similar topics between courses. Overall, I choose outputs from part 4 as my final curriculum design.