Introduction:

This report will guide the course curriculum design for a new “Master of Business and Management in Data Science and Artificial Intelligence” program at the University of Toronto based on the skill clusters generated from unsupervised machine learning algorithms (Hierarchical clustering and K-means clustering).

Part 1: Data Collection and Cleaning

To gather the initial dataset, we scraped job postings for "data scientist" and "data analyst" roles from the U.S. section of Indeed.com. This was based on their popularity among data science and AI graduates. Our scraping resulted in a raw dataset of 1,500 job listings, each detailed across 9 columns. We detected some null values in our raw dataset: 791 in the **Salary** column and 1500 in the **Rating** column. Therefore, we dropped the **Rating** because it does not provide much information. For the **Salary**, we filled all null values with “Unknown” to indicate the salary entity was not successfully scraped. We also dropped **Links** and **Locations** because we believed that they would not be helpful for clustering.

To further prepare our dataset for later implementations, we cleaned the **Description** by only keeping the description itself. We also extracted numerical salary as the **Extracted Salary** column from the **Salary** by calculating the average in each salary entity. For rows with “Unknown” **Salary**, we filled their **Extracted Salary** with the average salary for data scientists in the US, which is $76835, according to Indeed.

Part 2: Exploratory Data Analysis (EDA) and Feature Engineering

Feature Engineering

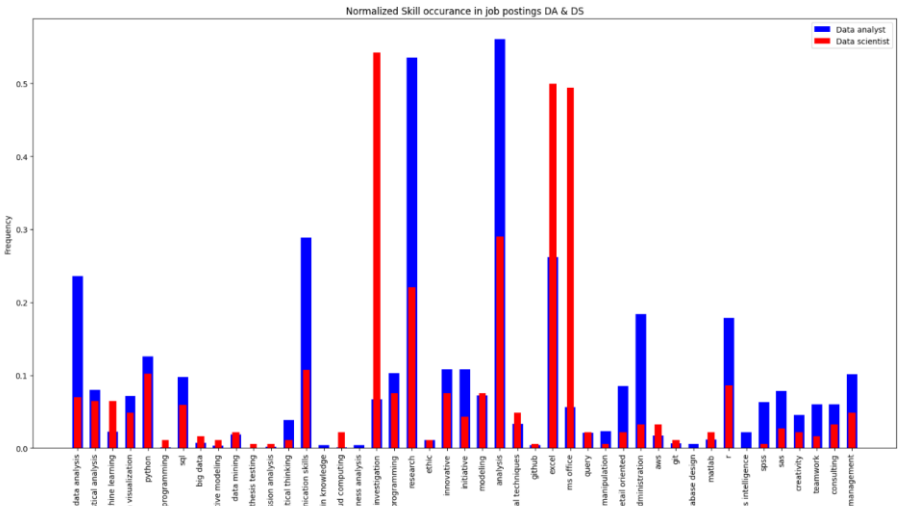
We first needed to generate a list of skills for skill extraction from the **Description**. In the creation of the list, we combined the manually generated skills and those that were generated by giving OpenAI API with the order "Please give me a list of relevant skills (Keywords) name for a data science related job.”

The algorithm to extract skills in our analysis was based on N-grams, representing a contiguous sequence of n items for a given text sample. We defined a function **extract_skills,** which first tokenizes the text into words. Then, the function would split a description into sequences with one, two and three of items (1-grams, 2-grams, and 3-grams). Lastly, the function compared every element in three sequences with the skill set and return all elements common to both sets. This method provides a way to compare potential skills in the **Description** and our skill set. However, N-grams can only capture limited context, since each description entity is restricted to the size of n.

We also generated text embeddings with **Bert** library on the **Description**. The text embeddings are the numerical representations of text. It can provide an additional aspect of understanding the distribution.

EDA (Comprehensive Analysis is included in the notebook)

Keywords in data analyst vs data scientist



Normalized skill occurrence in job postings data analyst vs data scientist

Part 3: Hierarchical Clustering Implementation

Since our skill columns were binary columns, the **Euclidean distance** was not suitable because it is more appropriate for continuous data. Instead, we used **Hamming Distance** to determine the difference between pairs of skills. It measures the proportion of binary variables that differ between two binary vectors. If two skills appeared simultaneously in many of our observations, we would conclude they are similar to each other, and their **Hamming Distance** would be closer to 0. Conversely, if they rarely appeared together, their **Hamming Distance** would be higher, approaching 1, indicating greater dissimilarity.

From the result of our **hierarchical clustering (Appendix I)**, we developed a course curriculum including 8 courses in total as follows:
Course 1: Communicate with Data (Analysis, Research, Excel, Communication Skills)
Course 2: Introduction to Data Analytics (R, Python, Initiative, Innovative, Investigation)
Course 3: Introduction to Data Science & Machine Learning (Data Visualization, Modeling, Machine Learning)
Course 4: Data-Driven Reporting (Critical Thinking, SQL, detail Oriented, Programming)
Course 5: Additional Quantitative Fundamental (MATLAB, Git, AWS, Ethic)
Course 6: Cloud-Based Data Analysis (Database Design, Big-data, Predictive Modeling)
Course 7: Data Engineering (Data Preprocessing, Data Engineering, Data Cleansing)
Course 8: Advanced Data Algorithm (Deep Learning, Neural Networks, Time Series Analysis)

Part 4: K-means Implementation

Another clustering algorithm was the K-means algorithm. We started by applying a K-means algorithm using the text-embedded description column to cluster job posts into 5 categories by their descriptions. Then, we generated **5 binary feature columns** to indicate which cluster the skills appear in the most frequently (e.g. 1 if the skill appears in the cluster the most frequently; otherwise, 0). Aligned with these 5 columns, we manually engineered a few more features including **skill frequency**, **average salary**, and **binary indicators** (data science, communication, data engineering) **of skills** (e.g. 1 if the skill is needed for data science; otherwise, 0). Finally, we generated a skill data set with 45 rows of

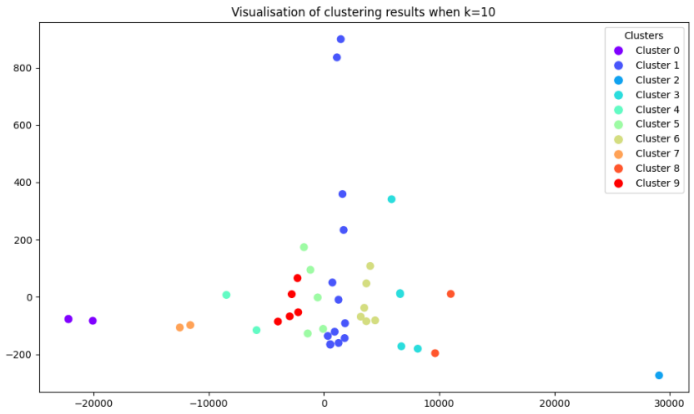
skills and 10 columns of features corresponding to each row of skills. With this dataset, we were able to implement a k-means algorithm with k=10 (check the result at Appendix II).

The elbow method was also used to determine the optimal number of clusters for our k-means clustering. As increasing the number of clusters will naturally improve the fit of the model, one increase of k will give increasingly less information after certain k. Therefore, elbow method can find a point where adding more clusters does not give much better clustering results. The interpretation of K-means results will be discussed in part 5 and part 6.

Part 5: Interpretation of Results and Visualizations

a) Although we were able to generate a course curriculum with eight courses according to the result of our hierarchical clustering, the limitation seemed to be obvious: the distribution of skills across clusters was uneven (see Appendix I). In other words, some clusters had a large number of skills, while others had very few. The potential issues of using this hierarchical clustering include some courses being overloaded with content while some skills being overlooked. Therefore, we should refer to the second clustering algorithm, k-means, to design our final course curriculum.

b) To visualize the clusters, we applied PCA to reduce the dimension of the features columns into 2.



- We observed that skills were separated into 10 clusters successfully.
- There is a point on the far right of the graph (Cluster 2). It is the point representing the skill “hypothesis test.”

c) The visualization of the elbow method shows that the optimal **k** is 4 where the rate of decrease sharply changes (Appendix III). However, the optimal **k** does not meet our requirement to generate a course curriculum with at least eight courses.

Part 6: Discussion and Final Course Curriculum

Our final course curriculum was based on k-means algorithm with **k=10**. The reasons why we chose this method over the other two were discussed in part 5 a) and c). Furthermore, according to our observations in part 5 b), the data point that represents the skill “hypothesis test” seems to be an outlier in our skill data set and a cluster with only one skill is not informative enough to form a course. Therefore, we should ignore it in the final course curriculum.

Ultimately, our final course curriculum with 9 courses, including skills to learn (STL) for each course, is as follows:

- **Course 1** - Integrate Tech-Business Skills (STL: Domain Knowledge in Business, Business Analysis)
- **Course 2** - Data Analytics and Data Management (STL: Data Analysis, R programming, SQL, Predictive Modeling)

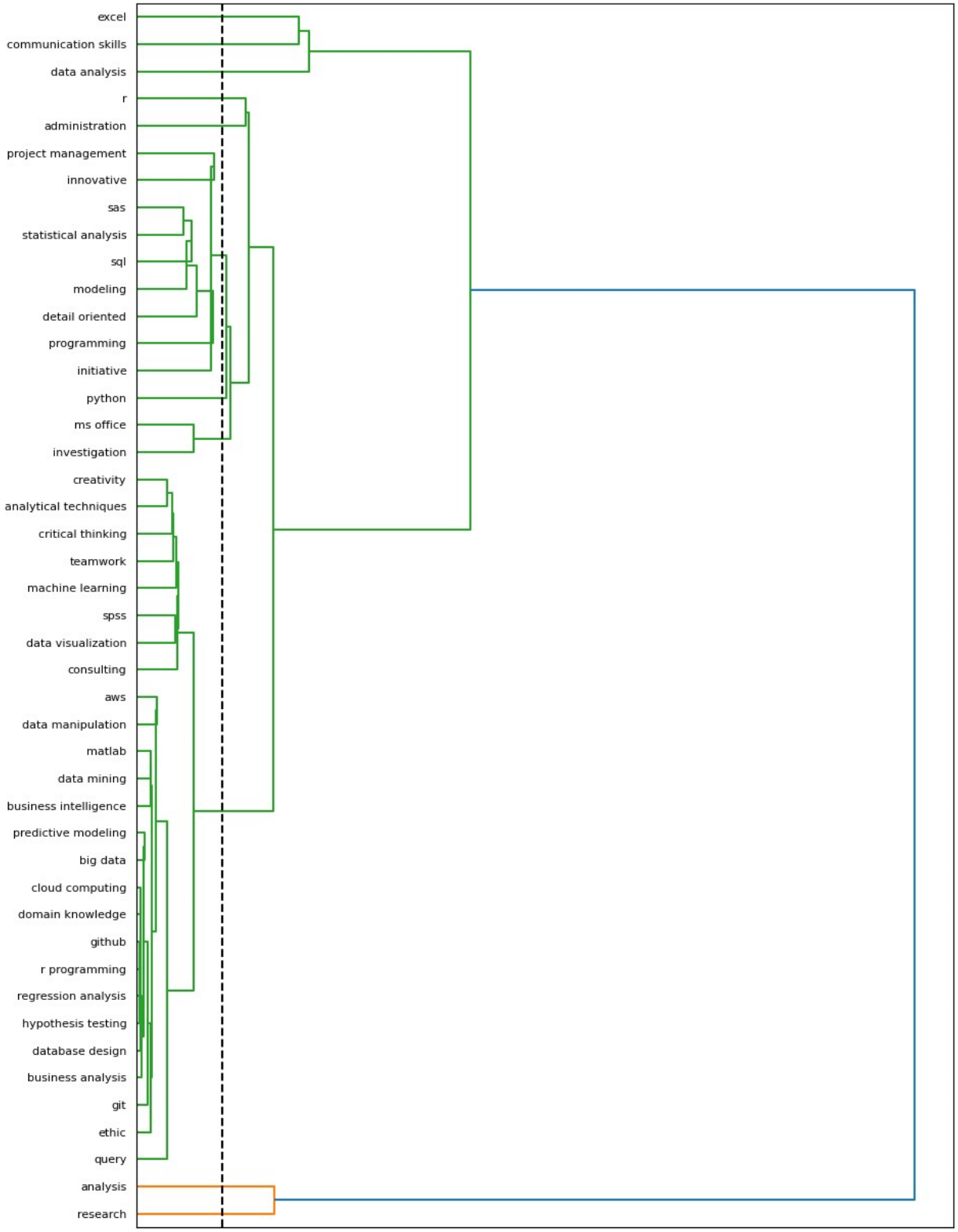
- **Course 3** - Investigation and Office Skills (STL: Investigation, Ethic, Initiative, Excel)
- **Course 4** - Data Mining and Modeling (STL: Data Mining, Modeling, Data Factory, Cloud Computing, IoT)
- **Course 5** - Technical Foundation of Data Science (STL: Python, SAS, AWS)
- **Course 6** - Business Analytics (STL: Data Visualization, Innovative, Detail-oriented, Consulting)
- **Course 7** - Big Data and Quantitative Analytics (STL: Big Data, MATLAB, Azure, Streaming Analysis)
- **Course 8** - Data Administration (STL: MS Office, Database Design, Data Warehousing, Data Cleaning and Preparation)
- **Course 9** - Advanced Data Science (STL: Statistical Analysis, Machine Learning, SPSS, Creativity, Programming)

Part 7: OpenAI to Describe Clustering Results for bonus.

Instead of manually giving course names according to the skills in clusters, we can use OpenAI to describe the clustering result for us by using OpenAI API. Please refer to Appendix IV for the result.

Appendix

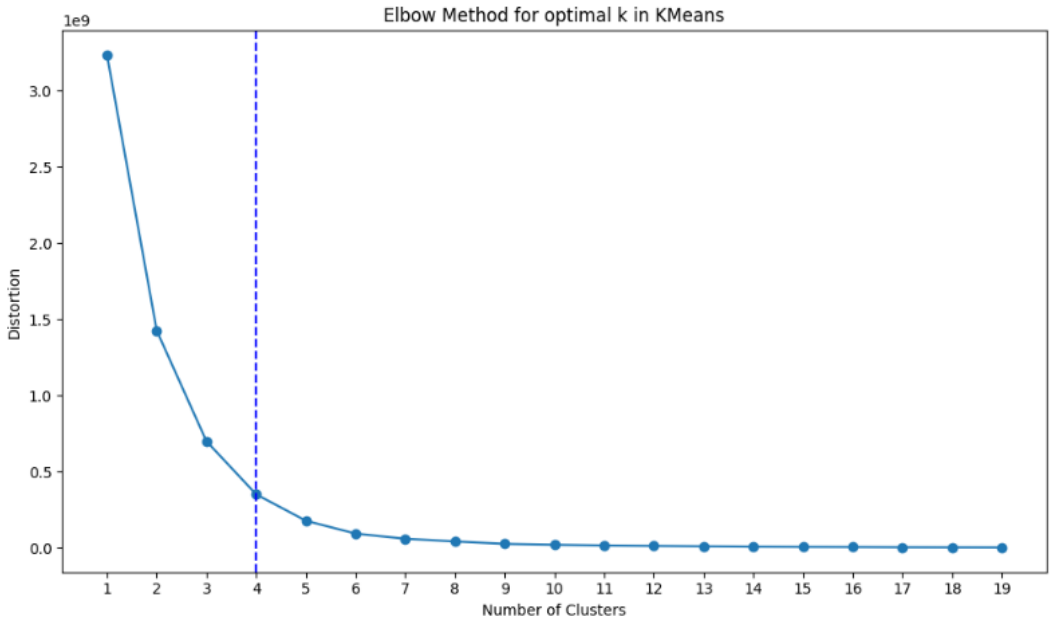
Appendix I:



Appendix II:

```
Skills in cluster 0:
['domain knowledge', 'cloud computing', 'business analysis']
Skills in cluster 1:
['data analysis', 'r programming', 'sql', 'predictive modeling', 'regression analysis', 'critical thinking', 'communication skills', 'research', 'analysis', 'github', 'query', 'data manipulation', 'business intelligence', 'project management']
Skills in cluster 2:
['hypothesis testing']
Skills in cluster 3:
['investigation', 'ethic', 'initiative', 'excel', 'git']
Skills in cluster 4:
['data mining', 'modeling']
Skills in cluster 5:
['python', 'analytical techniques', 'aws', 'r', 'sas']
Skills in cluster 6:
['data visualization', 'innovative', 'detail oriented', 'administration', 'teamwork', 'consulting']
Skills in cluster 7:
['big data', 'matlab']
Skills in cluster 8:
['ms office', 'database design']
Skills in cluster 9:
['statistical analysis', 'machine learning', 'programming', 'spss', 'creativity']
```

Appendix III:



Appendix IV:

Based on the listed skills, common themes can be observed among the clusters:

Cluster 0: Skills related to technical expertise such as domain knowledge, cloud computing, and business analysis.

Cluster 1: Skills related to data analysis and statistical modeling, along with strong communication and critical thinking skills. Experience with tools like R programming, SQL, and GitHub are also present.

Cluster 2: Skills related to hypothesis testing, indicating a focus on statistical analysis and research methodologies.

Cluster 3: Skills related to investigation and ethical practices, along with proficiency in using tools like Excel and Git for data management and version control.

Cluster 4: Skills related to data mining and modeling, indicating a focus on extracting insights and developing predictive models.

Cluster 5: Skills related to programming, data analysis, and cloud computing, with proficiency in Python, AWS, R, and SAS.

Cluster 6: Skills related to data visualization, administration, and teamwork, along with an innovative and detail-oriented mindset. Consulting abilities are also mentioned.

Cluster 7: Skills related to handling big data and proficiency in using MATLAB for data analysis and modeling.

Cluster 8: Skills related to database design and proficiency in MS Office suite.

Cluster 9: Skills related to statistical analysis, machine learning, and programming, along with proficiency in SPSS and a emphasis on creativity in data analysis and modeling.