Generating New Course Curriculums for Data Science and Analytics

Section 1: Data Generation and Cleaning

The data was collected using the provided web scraping code to scrape for both American and Canadian data scientist and data analyst positions from www.indeed.com (ca.indeed.com for the Canadian postings). Data pertaining to job title, location, company name, job description, job rating, and salary were all collected. These data were then concatenated into a single dataframe and duplicate job postings were removed. The final dataset contained 2,429 unique jobs postings and 48 skills (Appendix A). All values relating to job title and company name were present, however, numerous salary and rating values were missing (>50%). Despite a large proportion of missing data in these columns, information in other columns, such as job description and job title, were useful for analysis. Ultimately, it was determined that the missing data would not be imputed (as such a large portion was missing it wouldn't add any additional insight into the data) but instead set to 0 and excluded from average value calculations for these columns.

Section 2: Feature Engineering and Extraction

Features (i.e., skills) were generated using three methods – first, the use of OpenAI's ChatGPT to suggest skills that were required for both data analyst and data scientist positions; second, use of personal domain knowledge about the industry; and third, through extracting keywords from the job descriptions using a pre-trained Natural Language Processing (NLP) algorithm.

To generate features from ChatGPT, the program was prompted to provide "at least 10 important skills…" needed for both data scientist and analyst roles (see Appendix B for more detail about the ChatGPT configuration). Using this output, domain knowledge was used to select important features from the text to create individual features (e.g., output: "Proficiency in SQL and other programming languages like Python or R" would create features "SQL", "Python" "R", and "Programming"). Keywords were also extracted from the job description field using the pre-trained unsupervised NLP algorithm, TextRank [1]. TextRank extracts keywords from the text data based on their frequency of occurrence in the data and ranks the outputs from most occurring to least. The TextRank output was assessed and any relevant keywords that were not selected from the previous Chat GPT outputs were added to the feature set. Finally, additional features were added to supplement this information using domain knowledge. Please see Appendix A for a full description of the feature set.

A new column was generated for each individual feature and populated with binary values to indicate if the feature term occurred in the job description, where '0' indicates the term is not present and '1' indicates that the term is present.

Additional feature engineering was performed to be used in the Kmeans clustering step, henceforth referred to as the Kmeans feature data. This was completed by creating an index of the skill features and populating the data frame by calculating various summary statistics relating to the feature set (e.g., average salary and rating associated with each skill Appendix C, Figure 1). The full list of the engineered features and the generation method can be seen in Appendix D.

Section 3: Hierarchical Dendrogram and Kmeans Clustering

Once the model features were determined (Appendix A) this information was used to generate a distance matrix for clustering . This distance matrix was then used to generate clusters using sch.linkage from SciPy [2] and the resulting hierarchical dendrogram was visualized (Appendix E, Figure 2) and course curriculums were determined by selecting closely related skill sets from the dendrogram. The resultant curriculum suggestions can be seen in Table 1.

The engineered Kmeans feature data (Appendix D) was used to generate clusters that would be used to determine alternative course curriculums to those generated using the dendrogram. The Kmeans clustering was performed using scikit-learn [3]. The Kmeans model was fit using the scaled feature values to nine clusters, of which eight clusters contained enough information to create the curriculum (one cluster contained only one skill and was excluded from the curriculum). The optimal k value was six (Appendix F, Figure 3).The resultant curriculum suggestions can be seen in Table 2 and a visualization of the clusters can be seen in Appendix G, Figure 4.

Table 1: Overview of the potential course curriculum based on the information in the hierarchical dendrogram.

Course (C) Number: Name	Skills Covered
C 1: Effective Communication in Research	Research, Communication, Data Visualization, Statistics
C 2: Fundamentals of ML	Coding, Science, Machine Learning
C 3: Database Management	SQL, Engineering, Critical Thinking, Problem Solving, C/C++, Python
C 4: Applications of Big Data for Data Analytics	Data Mining, Big Data, Data Analytics
C 5: Advanced Techniques in Excel for Database Design & Analysis	Excel, Data Exploration, Design
C 6: Applications of Cloud Computing	AI, Math, Cloud Computing
C 7: Data Visualization for Business	Tableau, Collaboration, Modeling
C 8: Developing Business Acumen	Business, Consulting, PowerBI

Table 2: Overview of the course curriculum developed using Kmeans clustering. Note, the suggested curriculum names were taken from Chat GPT suggestions (Section 4) based on the skills covered in each group.

Course(C) Number: Name	Skills Covered
C 1: Professional Skills for Data Science	Communication, Problem Solving, Detail, Critical Thinking, Stakeholder Relations, Collaboration, Teamwork, Management, Writing
C 2: Machine Learning & Big Data Fundamentals	Machine Learning, Data Visualization, Data Cleaning, Big Data, Developer, Research, Coding, Java, C/C++, Cloud Computing, AI, Data Mining, Engineering
C 3: Programming for Data Science	Programming, MATLAB, R, Hadoop
C 4: Advanced Data Science Techniques	Probability, Spark, Deployment, Simulation
C 5: Data Analytics and Modeling	Data Analytics, Modeling, Data Exploration, Math, Python, Data Management
C6: Business Intelligence & Data Visualization	Business, Design, Tableau, Excel, SQL, Science
C7: Statistics and Data Science Research	Statistics, Data Visualization, Research, Coding

C 8: Leadership and Consulting for Data Science	Leadership, Consulting, Presentation Skills

Section 4: Interpretation of Suggested Curriculums

Ultimately, the offerings of the course curriculums are fairly similar and a student who had a comprehensive grasp on many of these skills would likely be successful in the workplace. However, the final course curriculum suggestion would be the one presented from the Kmeans clustering data (Table 2). The main reason for this is that the hierarchical dendrogram provided larger clusters of skills that had to be broken down into smaller skill sets for each curriculum by visually assessing the dendrogram figure (Appendix E) and incorporating domain knowledge of useful skills for this field to generate the curriculum suggestions. This was a more qualitative assessment than that done in the Kmeans clustering approach, where the number of clusters could be specified (nine, in this case) and the skills within each cluster could be considered a single course. Additionally, when plotting the clusters one can observe a more definite separation among clusters than that in the dendrogram (Appendix G, Figure 4) indicating that similar skill sets were grouped together. Validation of this assumption that similar skill sets were grouped together can be observed in Table 2, where technical programming skills tend to be paired together while more business-centric and communication/management skills tend to be paired together.

This analysis is supported by the analysis done using ChatGPT, where ChatGPT was presented with both curriculums and asked to compare the course offerings, output the main differences between the two curriculums, which curriculum would better prepare a student for working in these industries, and additional course suggestions to supplement these curriculums. ChatGPT outlined that the Kmeans curriculum provided more exposure to essential technical skills (such as Machine Learning & big data management) than the dendrogram curriculum. The Kmeans curriculum also contained a more comprehensive set of skills, such as the addition of specific professional skills courses (Table 2, C1 & C8), making the Kmeans curriculum the better choice. It also suggested that the curriculums be supplemented with additional project management, deep learning, and data storytelling skills to better prepare students. It is worth noting that these skills were included in the analysis (i.e., Power BI and deep learning) but excluded from the curriculum as the cluster only contained two skills.

There are some limitations of this work, primarily relating to the quality of the data and data processing steps. Notably, the application of TextRank was not overly effective as it couldn't filter out similar words (e.g., 'includes' and 'included' where both high ranking outputs) and the relevance of the terms varied drastically, as it only was assessing the frequency of the terms appearing in the data. Additionally, certain skills were missing any presence in the data (e.g., MATLAB, R, and the term "programming" did not appear once in the data). As all of these skills are utilized in data science and analytics, it is unusual that they were not captured in the job descriptions.

Future work should focus on implementing a more effective NLP method to gather skills from the job descriptions. Effective application of an NLP algorithm may also help mitigate the latter issue of certain terms not occurring in the dataset by utilizing the NLP algorithm to scan the job descriptions for keywords occurring, instead of only utilizing it to extract skills via frequently occurring terms.

References

[1] P. Nathan, "pytextrank: Python implementation of TextRank as a spaCy pipeline extension, for graph-based natural language work plus related knowledge graph practices; used for phrase extraction and lightweight extractive summarization of text documents." Accessed: Apr. 08, 2023. [OS Independent]. Available: https://derwen.ai/docs/ptr/

[2] "SciPy." https://scipy.org/ (accessed Apr. 08, 2023).

[3] "scikit-learn: machine learning in Python — scikit-learn 1.2.2 documentation." https://scikit-learn.org/stable/ (accessed Apr. 08, 2023).

***Appendix A* – Feature Set Description**

The following list follows the format of **'Skill Name** – Description' where '1' indicated in the database means the skill was in the job description, '0' otherwise.

Programming – Any general reference to programming skills
Data_Analytics – Any general reference to required data analytics skills
Big_Data – The application of big data techniques are required for this posting
Business – General Business Acumen skills are required
Communication – General communication (oral, written, visual) is required
Problem_Solving – General problem solving skills are required
Detail – Shorthand for 'Attention to Detail' or 'Detail Oriented' skills that are required
Critical_Thinking – General critical thinking skills required
Data_Management – Skills relating specifically to data management are required
Developer – Used to capture software or program developer-specific skill sets
Research – General research skills required
Teamwork– General teamwork skills required
Coding – Similar to Programming, used to capture general programming skills that may have been referenced as 'coding'
BI – Shorthand used to capture skills relating to the use of Power BI
Design – Shorthand used to capture references to requiring reference to more general skills like experimental or database design
Deep_Learning – The application of deep learning techniques are required for this posting
Leadership – General leadership skills are required for this posting
Consulting – General consulting skills are required for this posting
Excel – Knowledge or use of Excel is required for this posting
Stakeholder_Relations – Knowledge or experience in stakeholder relations are required
Cloud_Computing – The application of cloud computing techniques are required
AI – Application or knowledge of Artificial Intelligence is needed for this role
Simulation – Application or knowledge of simulation models is needed for this role
Collaboration – General collaboration skills required
Management – General management skills required; used to capture skills relating to project management, team management, etc.
Probability – Knowledge or use of probability is required for this posting
Writing – Writing (e.g., reports, papers, articles) is needed for this role
Data_Exploration – Data exploration skills are required for this posting
Mining – Shorthand to refer to Data Mining skills required in the posting
Deployment – Used to capture skills relating to project/program deployment
Math – General math skills or applications are required
Science – General science skills or applications are required
Engineering – General engineering skills or applications are required
Presentation_Skills – Specific presentation communication skills are required

Feature names referring to a specific programming or database skill:

- Tableau
- Python
- SQL
- R
- MATLAB
- C/C++
- Hadoop
- Spark
- Java

Appendix B – ChatGPT Configuration

Chat GPT configuration was as follows:

```
##Code Start
message_log = [
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": """Please provide the following:
     - at least 10 important skills needed for data analysts
     - at least 10 important skills needed for data scientists"""}]

response = openai.ChatCompletion.create(
    model="gpt-3.5-turbo",
    messages= message_log,
    max_tokens = 1000,
    temperature = 0.5
)

print(response.choices[0].message.content)
## Code End
```

Bolded information corresponds to the parameters set in order to provide a specific question to the algorithm and alter the length and ‘creativity’ of the response.

Appendix C - Visualization of Average Salary and Rating for Each Skill

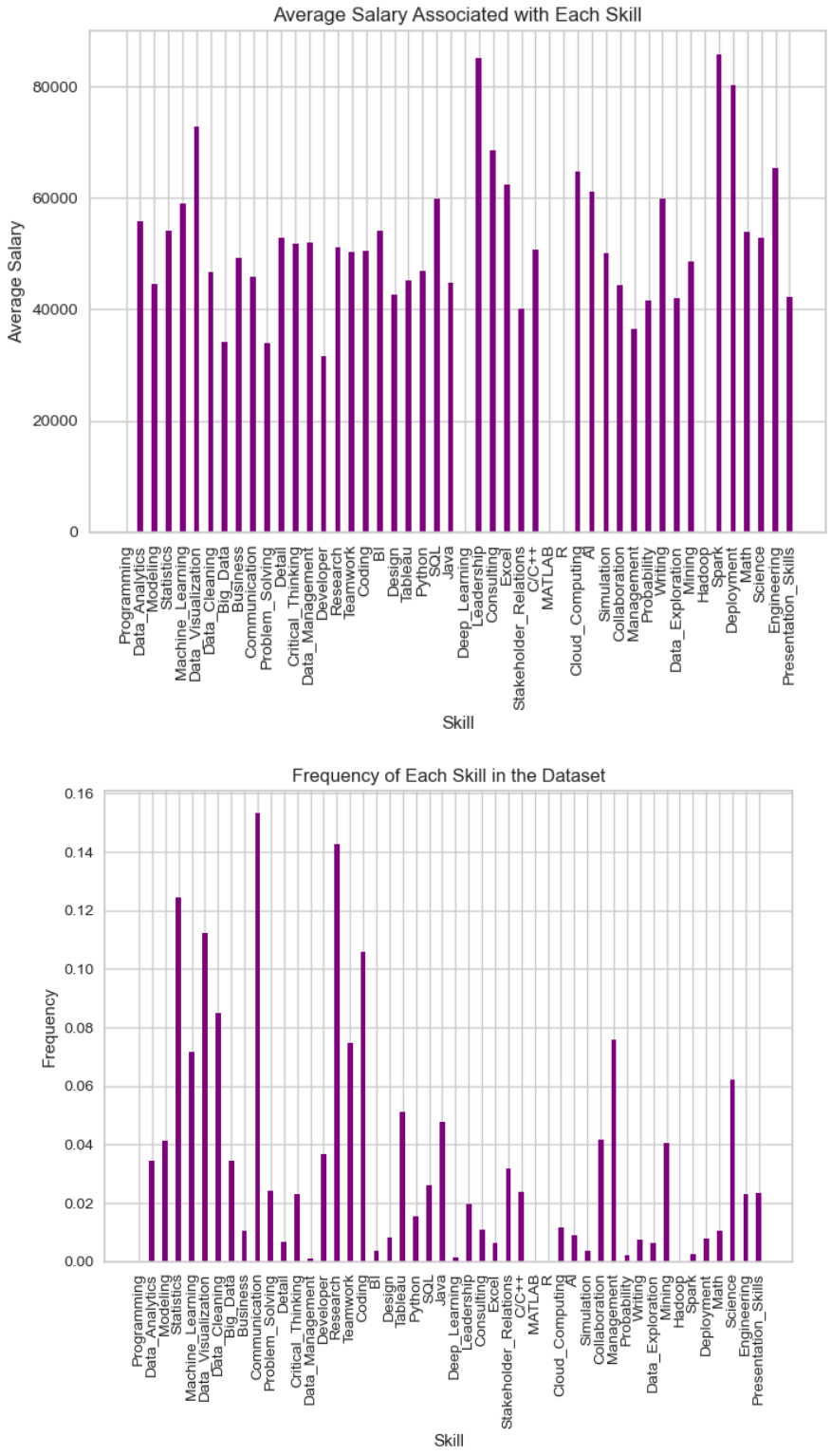

Figure 1: Barplot of the average salary (top) and average job rating (bottom) associated with each skill in the data set.

Appendix D – Summary of Engineered Features

Number	Engineered Feature Name	Description
1	*Skill_Frequency*	Frequency of the skill occurring for each skill
2	*Avg Salary*	Average salary for each skill
3	*hybrid_proportion*	The proportion of job roles that offered hybrid options for work
4	*remote_proportion*	The proportion of job roles that offered remote options for work
5	*is_hard_skill*	A binary value, where '0' represents not a hard skill, and '1' represents that it is considered a hard skill. List of hard skills was generated using domain knowledge.
6	*is_soft_skill*	A binary value, where '0' represents not a soft skill, and '1' represents that it is considered a soft skill. List of hard skills was generated using domain knowledge (see ipynb for list).
7	*jr_proportion*	Proportion of roles that are considered entry level positions. Keywords (such as 'junior' or 'entry') were searched for in each job title to determine if the role was an entry level position. For a full list of terms used see ipynb file.
8	*sr_proportion*	Proportion of roles that are considered mid-senior level positions. Keywords (such as 'senior' or 'manager') were searched for in each job title to determine if the role was a mid-senior level position. For a full list of terms used see ipynb file.
9	*is_business_skill*	A binary value, where '0' represents that it is not considered a business skill, and '1' represents that it is considered a business skill. List of business skills was generated using domain knowledge (see ipynb for list).
10	*is_developer_skill*	A binary value, where '0' represents not a developer/programmer skill, and '1' represents that it is considered a developer/programmer skill. List of skills to identify these job postings can be seen in the ipynb file.

Appendix E – Hierarchical Dendrogram Visualization

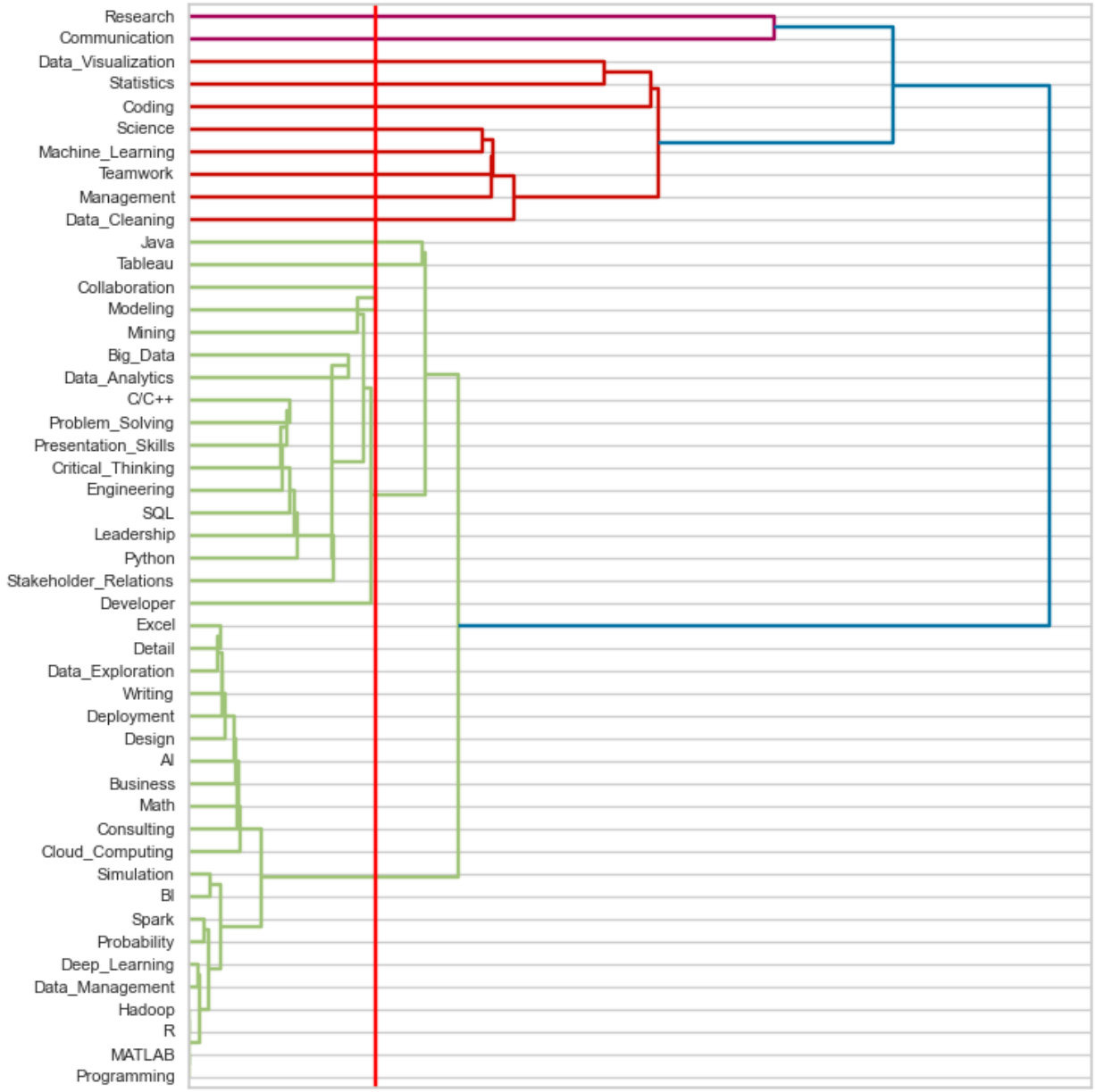

Figure 2: Hierarchical dendrogram visualizing the relationships between the different skills in the feature set.

Appendix F - Visualization of Elbow Plot for Optimal K value in Kmeans Clustering

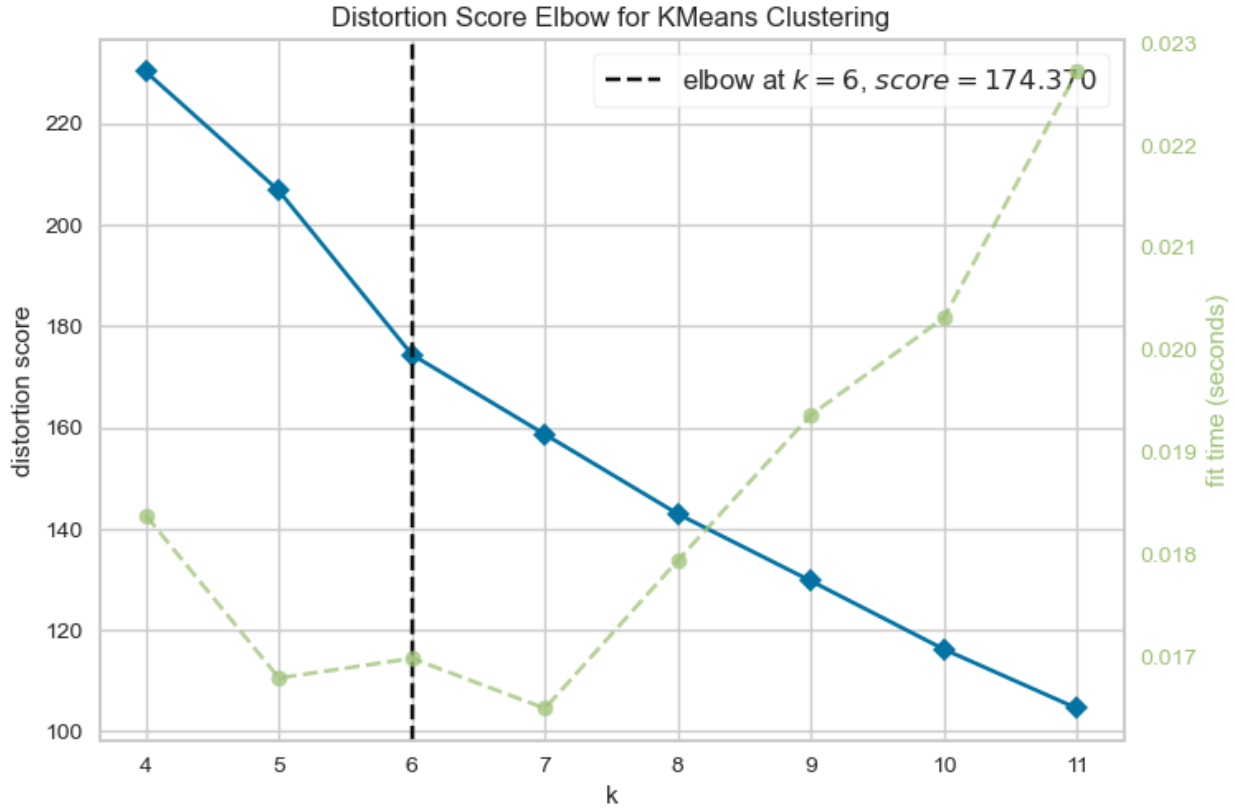

Figure 3: Elbow plot indicating the optimal *k* value for the Kmeans clustering analysis and the correlation of the time, in seconds, of how long it took the model to fit for each *k* value. The optimal *k* value is 6 for this model.

Appendix G - Principal Components Analysis (PCA) Scatter Plot of the Kmeans Clusters

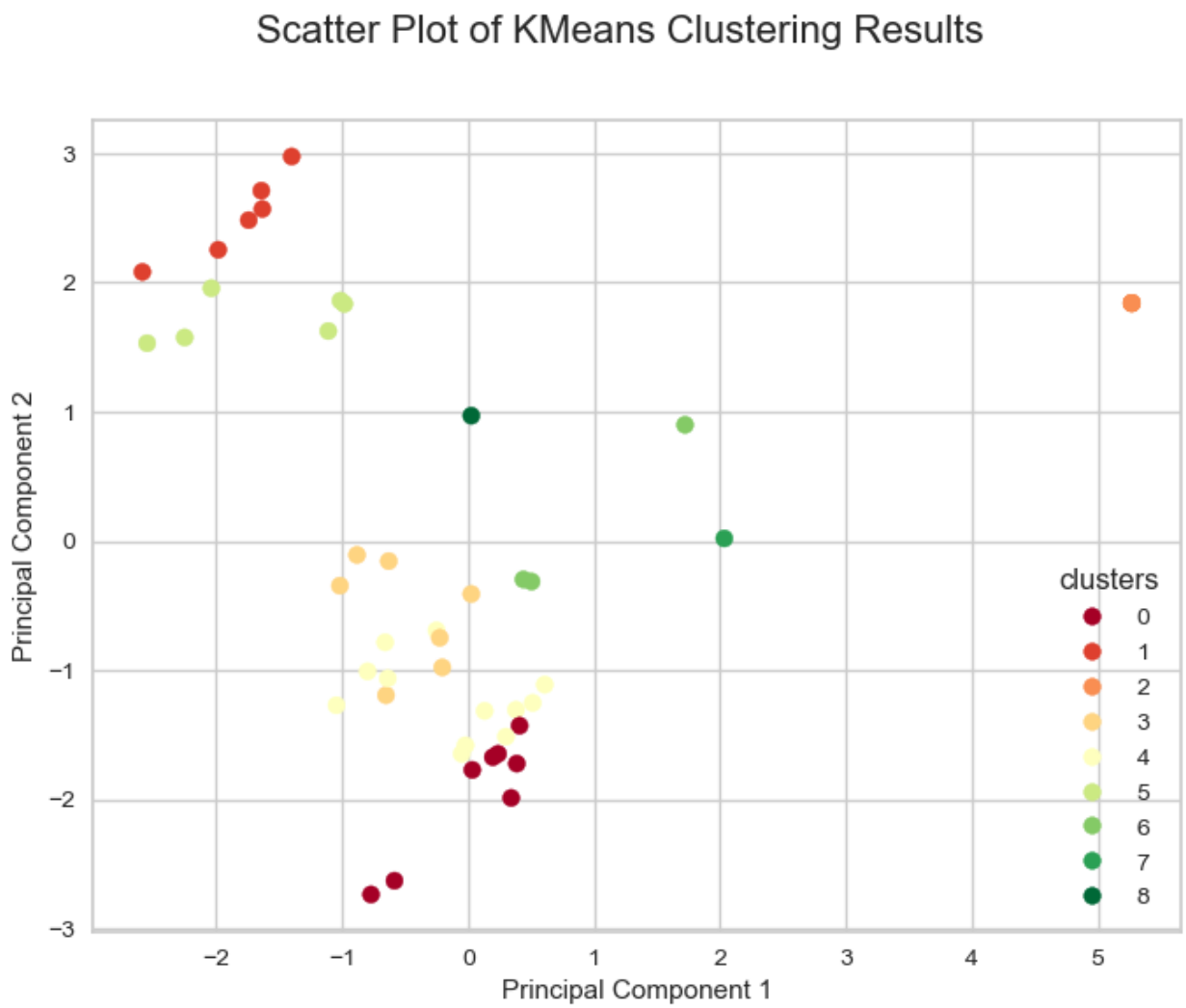

Figure 4: Output of the PCA scatter plot of the kmeans clusters. One can observe some distinct separations among certain clusters, such as clusters 0, 1, and 5, and some distinction among clusters 3,4, and 6.