DATA301

Yaquub Ali

97166977

Project Report

Project Abstract/Summary: For my project, I intend to use the ‘RentTheRunway’ dataset which includes clothing fit feedback. The research question I have decided to investigate is how body image/type perception influences users' fit preferences and satisfaction with clothing items. Throughout my project I intend to use algorithms including clustering to make meaning of my data. I intend for this project to provide many different plots/visualisations and/or graphs to clearly and easily understandably portray the findings of my investigation. The significance of my findings will be of importance to those in the fashion/clothing industry to better understand what more they need to do to satisfy people of all different body types.

Introduction: The Rent the Runway dataset contains data about clothing item reviews, including 15 different fields of data. Throughout my project however, we will be focusing on the data fields ‘body type’, ‘fit’, ‘rating’, and ‘review text’. I have chosen these 4 data fields because I feel as if they are the most likely fields to draw meaningful outcomes from. Algorithms like clustering are used to group users based on their body types/image to understand how body image perceptions affect fit preferences and satisfaction.

As someone passionate about fashion and having struggled with sizing due to being slim, this research question resonates deeply with me. Exploring how body image influences fit preferences and satisfaction with clothing items directly connects to my personal experiences. Sharing the findings from this research could help others facing similar challenges by fostering understanding and appreciation for diverse needs in fashion.

The research question, "How does body image/type perception influence users' fit preferences and satisfaction with clothing items?" is directly relevant to the Rent the Runway dataset, which contains clothing item reviews and user feedback. Implementing clustering on the dataset will help segment users based on their body image perceptions and fit preferences and hopefully uncover trends/relationships which can help to answer the research question. This analysis will uncover patterns and correlations between these different groups, providing insights into how body image influences user satisfaction with clothing items.

Experimental Design and Methods: Within my project, I had a single colab notebook dedicated to data processing in which I handled and processed my code in a way which prevented me from having to import it every time I ran a new instance of the main notebook. This way, it became a lot more time-efficient for me to implement my code which translated to overall more time to complete the project. The workflow began with data pre-processing, where I downloaded and decompressed a large dataset of clothing reviews from the Rent the Runway platform. The data was filtered to retain only relevant entries containing body type, rating, review text, and fit information. This filtered data was then uploaded to Google Drive for persistent storage. Using Dask, the data was processed in parallel to handle the large size efficiently. Dask was also utilized to convert the data into a Dask DataFrame for further analysis.

Next, I did some basic exploratory data analysis (EDA) to get a feel for the dataset and potentially uncover some trends early on to further investigate when I introduce algorithms to the data. Dask and NLTK were employed to process text data, generating trigrams from review texts to identify common phrases associated with each body type. The mean ratings were calculated and visualized using matplotlib to identify trends in user satisfaction across body types. In the final step, clustering analysis was performed using scikit-learn. The data was pre-processed with one-hot encoding for categorical features and standardized before applying K-Means clustering. PCA was used for dimensionality reduction for visualization purposes. From here, I wanted to see the proportions of body types within each cluster. To do this, I melted the dataframe using pd.melt to a long format for plotting. Then, I filtered rows where 'presence' indicates the body type's presence, removed the 'presence' column, and adjusted 'body type' names for better readability. From here, I used matplotlib to visualize the distribution of body types within each cluster.

Libraries and Modules

1. **Dask**: Used for parallel processing and handling large datasets efficiently. Key functions include read_text, to_dataframe, and map.
2. **Google Colab & Google Drive Integration**: For storing and accessing large datasets easily across sessions using drive.mount.
3. **urllib & gzip**: Downloading and decompressing the dataset.
4. **Json**: Parsing JSON formatted data.
5. **pandas**: Data manipulation and conversion between Dask and Pandas DataFrames.
6. **scikit-learn**: Used for data preprocessing (StandardScaler, OneHotEncoder) and clustering (KMeans, PCA).
7. **NLTK**: For natural language processing tasks such as removing stop words and generating n-grams.
8. **matplotlib**: Visualization of mean ratings and clusters.

Results:

Scalability:

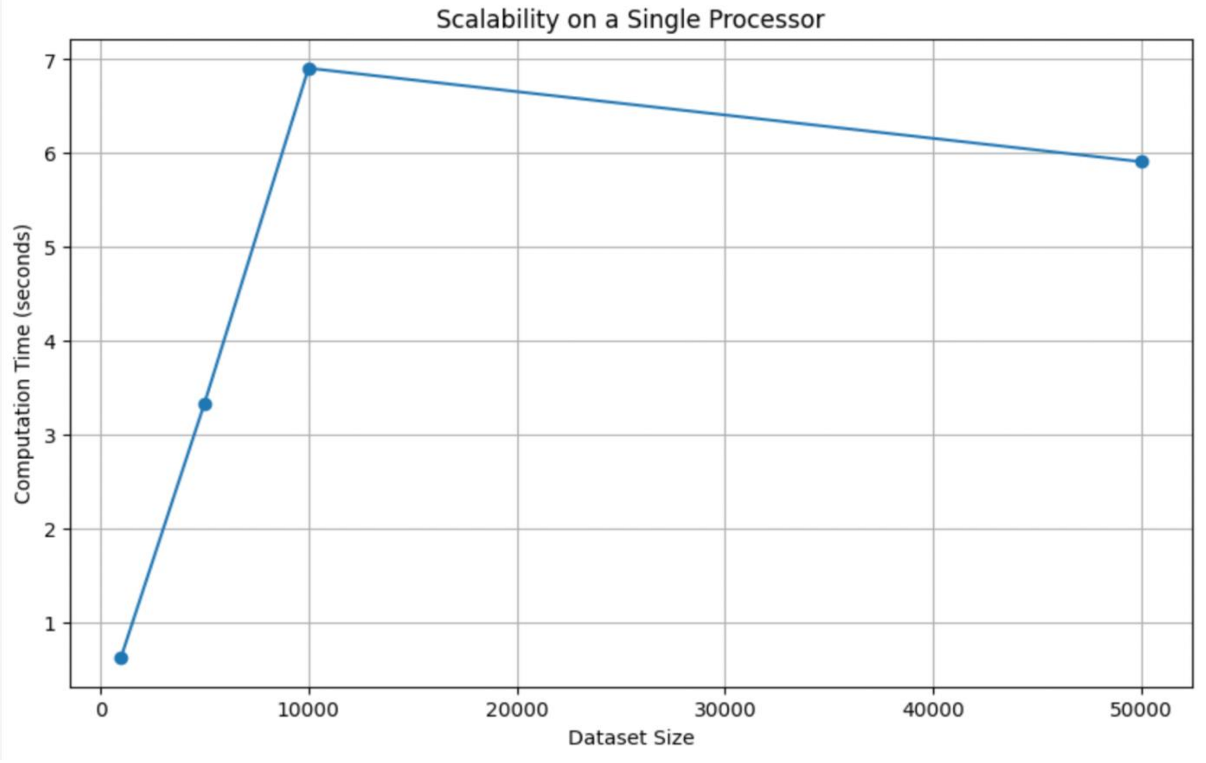


Figure 1: Scalability on a single processor

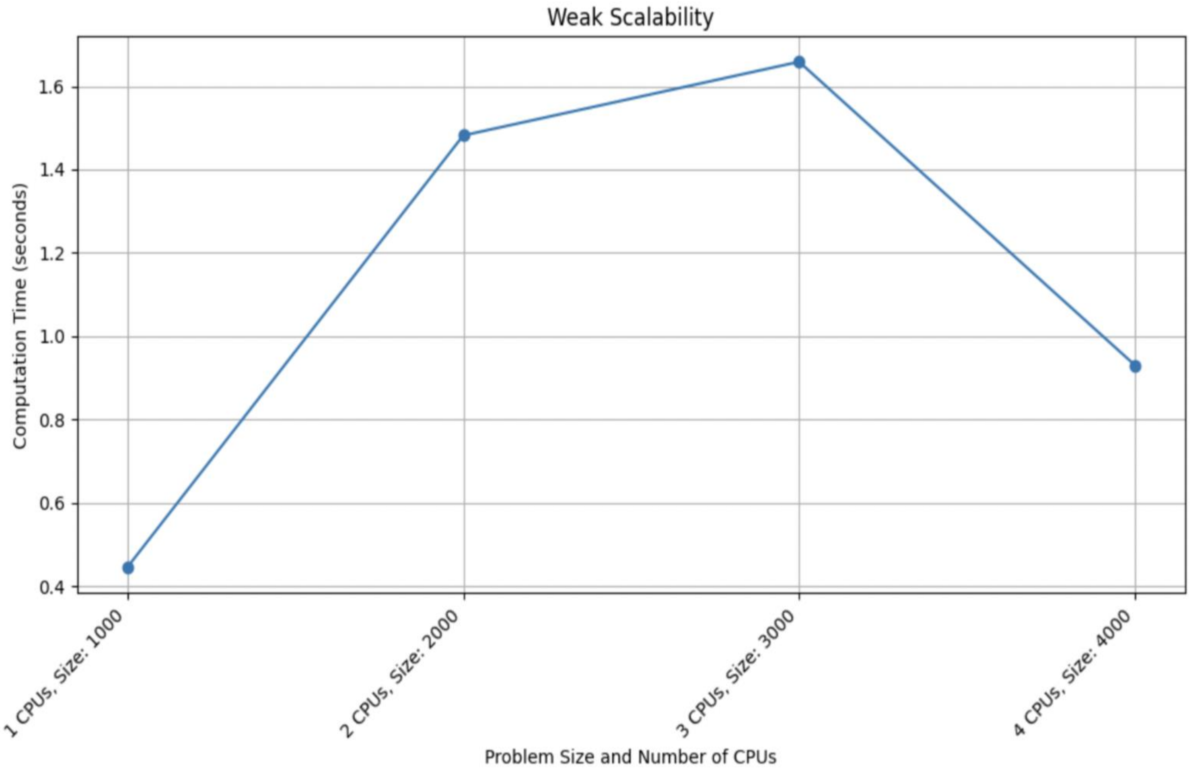


Figure 2: Weak Scalability

Data visualizations:

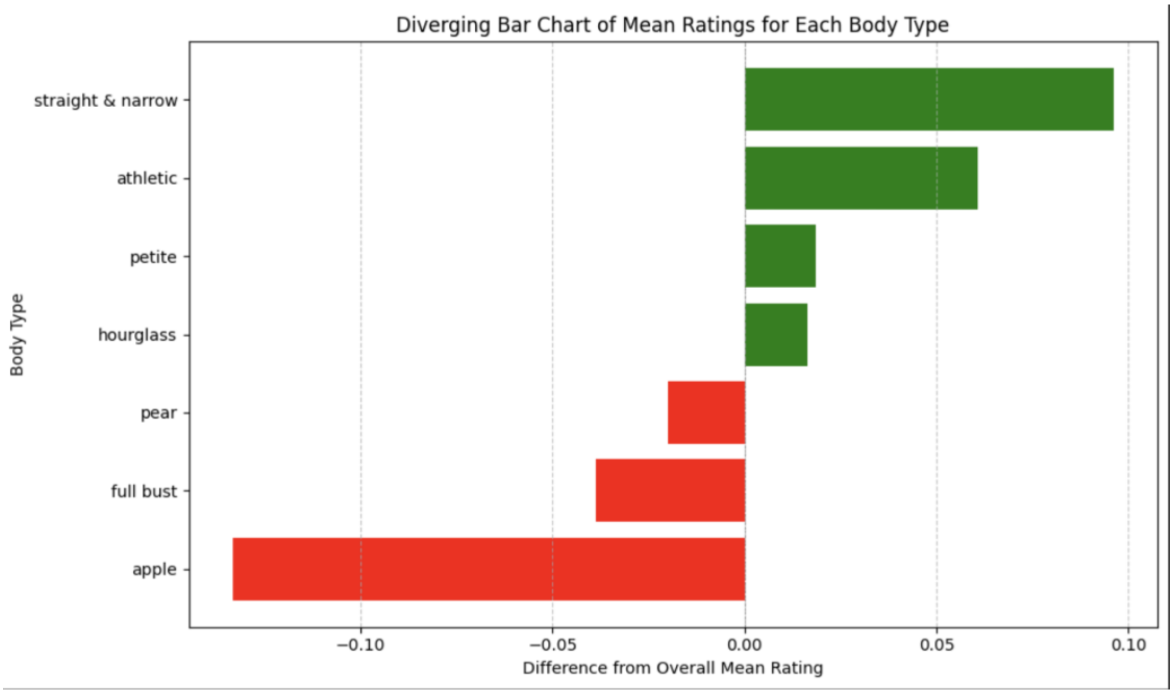


Figure 3: Diverging Bar Chart of Mean Ratings for Each Body Type

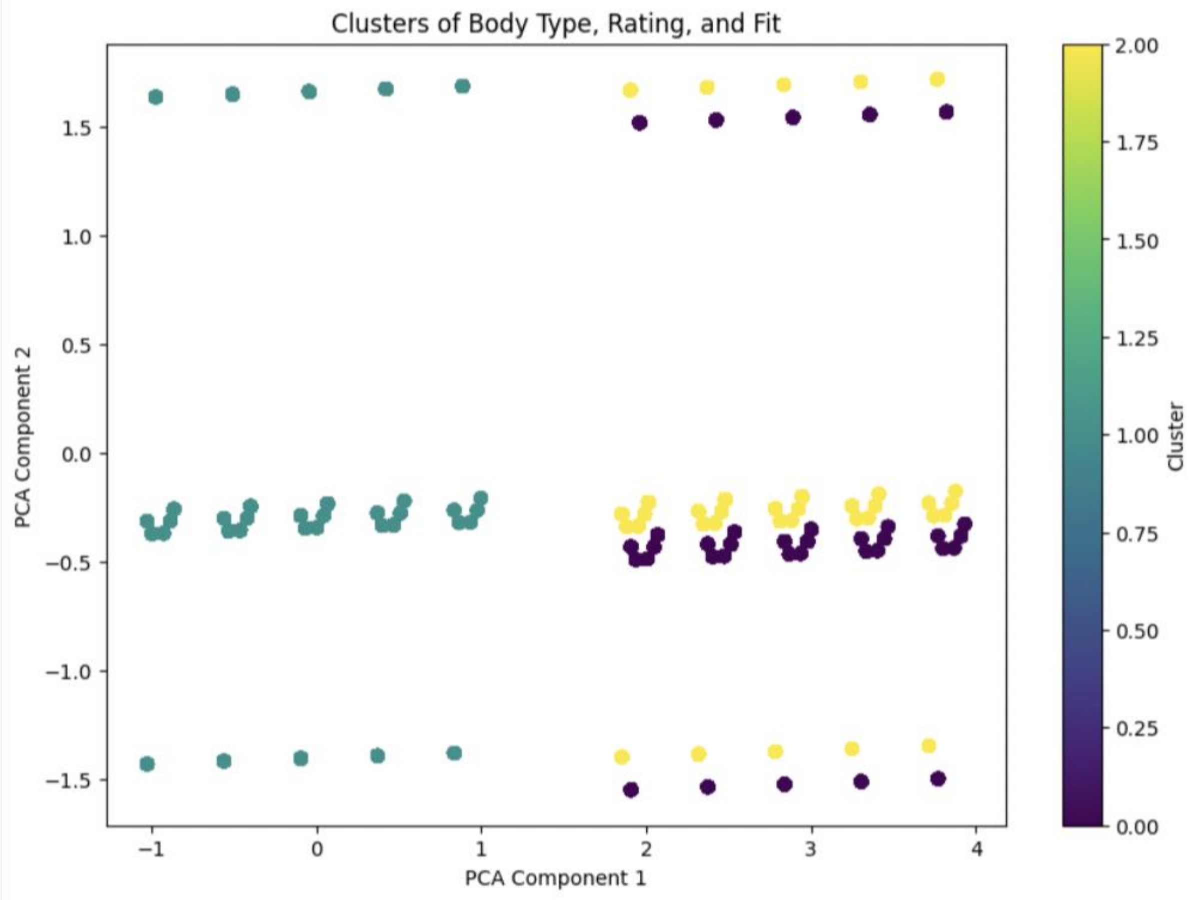
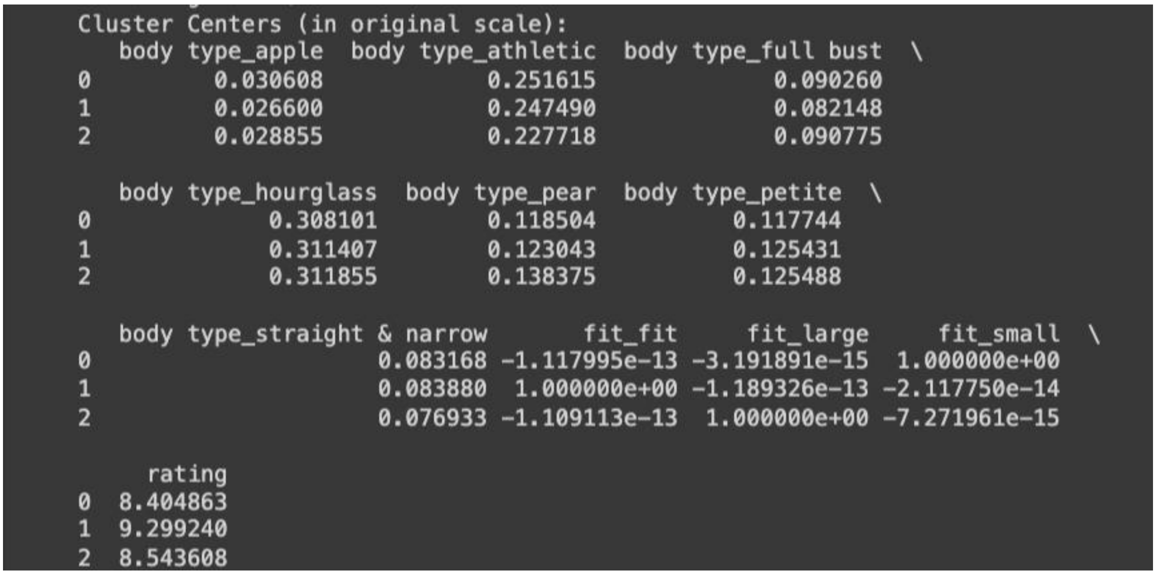


Figure 4: PCA of Clusters by Body type, Fit, and Satisfaction Ratings



```
Cluster Centers (in original scale):
   body type_apple  body type_athletic  body type_full bust  \
0         0.030608            0.251615             0.090260   
1         0.026600            0.247490             0.082148   
2         0.028855            0.227718             0.090775   

   body type_hourglass  body type_pear  body type_petite  \
0             0.308101        0.118504          0.117744   
1             0.311407        0.123043          0.125431   
2             0.311855        0.138375          0.125488   

   body type_straight & narrow       fit_fit     fit_large     fit_small  \
0                     0.083168 -1.117995e-13 -3.191891e-15  1.000000e+00   
1                     0.083880  1.000000e+00 -1.189326e-13 -2.117750e-14   
2                     0.076933 -1.109113e-13  1.000000e+00 -7.271961e-15   

     rating  
0  8.404863  
1  9.299240  
2  8.543608  
```

Figure 5: Cluster Centers

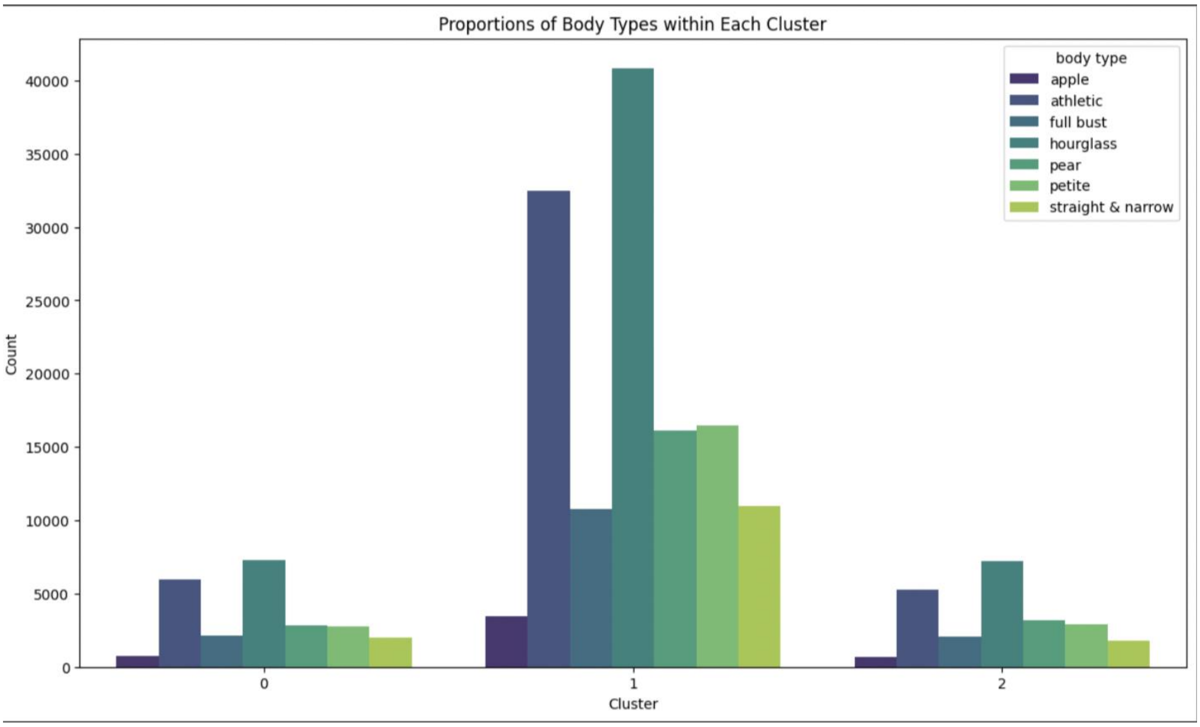


Figure 6: Proportion of Body Types within Each Cluster

Answer and discuss the hypothesis or research question as best as you can with data:

- From the 4 figures provided above, it's clear to see that users body type/perception has an effect on their satisfaction ratings. The first graph in figure 1 above shows a Diverging Bar Chart of Mean Ratings for Each Body Type. This bar chart shows that users with straight & narrow and athletic body types are most satisfied with clothing fit, as indicated by their positive ratings above the overall mean. Petite and hourglass types also show slight positive satisfaction, while pear types are neutral. In contrast, full bust and apple body types are less satisfied, with apple showing the largest negative rating.

- From the clustering visualizations above in Figure 1, 2, 3, and 4, we started to see which groups of users were similar to each other in terms of body types, fit preferences and satisfaction ratings. We can see that Cluster 0 is dominated by users with hourglass body type (30.8%), which is shown clearly through figure 4. A significant portion of users also find the fit to be small. The average rating for this cluster is 8.40. Cluster 1 is similarly dominated by users with hourglass body type (31.1%), but these users generally find the fit to be accurate. This cluster has the highest average rating of 9.29. This higher rating is shown on figure 2 by how separated cluster 1 is from the others. Cluster 2 also contains a notable portion of hourglass body types (31.2%) followed closely by athletic body types (22.7%) and pear body types (13.8%). Users in this cluster predominantly find the fit to be large and the average rating is 8.54. Cluster 1's higher ratings suggest that accurate fit correlates strongly with higher satisfaction ratings. The hourglass body type is also highly prevalent across clusters, suggesting that it could potentially be a significant factor in determining fit and satisfaction.

Conclusion:

- **Were you able to answer your hypothesis / research questions? Explain how and why (or why not):** I believe I was able to answer my research question. I was able to provide enough evidence to show that certain body types are significantly underrepresented within the 'rentherunway' dataset as well as the fact that more 'sought after' body types such as 'athletic' and 'hourglass' body types seemed to have a very heavy effect on ratings which was shown through our clustering analysis. Even though in my opinion I believe I was able to answer my research question, I don't believe any of my findings were in any way out of the ordinary and almost could've been predicted beforehand. With this being said, I still think my findings will be of interests to businesses/companies within the clothing industry.

- **What implications do your results have?** I believe my results would be of interest to businesses and companies within the fashion industry as it provides them an insight into the areas that they excel in, while also highlighting the areas where they could potentially improve and increase customer satisfaction. For example, from my clustering analysis it was clear to see that the clusters were dominated by 'hourglass' and 'athletic' body types while others were significantly misrepresented. Aside from this, body type 'apple' has the significantly lowest average mean. From this, businesses within the clothing industry could potentially request user feedback from customers who fall within these groups and gain knowledge as to how their experiences could be improved.

- **What future questions or directions would you take with your project?** For future directions, I would be interested to incorporate factors such as age, gender, and cultural background to potentially investigate how they influence satisfaction with clothing items. It would also be interesting to incorporate longitudinal studies to track changes in preferences and satisfaction over time which could potentially reveal trends and shifts in clothing/fashion culture.

Critique of Design and Project:

- One part of the design that could have worked better with a different approach is the data filtering and preprocessing step, specifically the method used for handling and updating the JSON data. The current method involves reading each line of the JSON file, filtering entries, and then rewriting the filtered entries back to the file. This approach, while straightforward, can be inefficient and slow for large datasets. It involves reading and writing the entire dataset line by line, which is not optimal for performance and can result in high memory usage and long processing times which can be seen in the scalability graphs above.

- A more efficient approach would be to use Dask for the initial filtering and preprocessing of the JSON data. As Dask is designed to handle large datasets by parallelizing operations across multiple cores, the data filtering and transformation steps can be significantly sped up. Instead of manually reading and writing the JSON file, the entire dataset could be loaded into a Dask bag, filtered in parallel, and then written out in a single operation. This approach not only improves performance but also enhances reliability, as Dask provides better handling of large-scale data processing and can recover from interruptions. Aside from this, using Dask from the beginning maintains consistency in the data flow, reducing the complexity of switching between different data handling methods. However, due to unfamiliarity with Dask, and also time challenges, I wasn't able to do this.

Reflection:

- **List course concepts and tools you found useful for completing the project:** Throughout my project various course concepts aided my software implementation as well as report. The lectures on Dask Dataframes, debugging, clustering, and also matplotlib were all extremely useful to me

- **What did you learn from the project?** The project was my opportunity to put into practice all that we have learnt throughout the course and the labs. For example, in the weekly labs, most of the data handling and preprocessing with dask was all done for us, and we then wrote functions to implement on the ready processed data. For the project, however, it was upon us to download our data, clean it and process it using dask for further analysis/function implementation. Personally, I learn things by applying them, so this project gave me a strong insight into the advantages of using dask in the real world.

References:

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge University Press.

Matplotlib tutorial: https://lectures.scientific-python.org/intro/matplotlib/index.html

Parallel programming introduction: https://carpentries-incubator.github.io/lesson-parallel-python/

Dask programming guide: https://docs.dask.org/en/stable/

Sample project code: Helped to generate trigrams for my code. https://colab.research.google.com/drive/1P7fYsSzJj_ZaUh_CH6nBhPfVtdrBWmas

Clustering guide using scikit-learn: https://scikit-learn.org/stable/modules/clustering.html