MIT 805 Assignment 1 - Part 2

Armand de Wet u16005326

18 October 2024

1 Map-Reduce

1.1 Explanation of dataset

I made use of the Amazon reviews 2023 dataset [1], which is a large multi-modal dataset that can contribute to business insights and value. I chose to specifically focus on the electronics category and focused my analysis on the top performing attributes.

The data is structured in 2 distinct files, nl. user review data (numerical ratings, textual rating, helpfulness votes etc.) and item meta-data (descriptions, price, links to different raw images etc.). These 2 datasets can be jointly analysed by joining them on the parent-asin attribute that represents the parent ID of the product.

This data was downloaded in JSONL format, which allows for easier streaming and processing of individual objects [2] and I made use of the latest version of this dataset that was collected in 2023. It is up to date as of September 2023 and contains interactions from May 1996.

An extensive overview of the dataset is provided in [3].

1.2 Map-Reduce Algorithm Approach and Reasoning; Extra Analyses

I considered this task out of the point of view of an electronics retailer looking to optimise business value and product acquisition and sales. Thus the goal of my map reduce algorithms and other analyses was to provide information on what the most popular and best performing brands, categories and stores are as well as insights into customer behaviour through the analysis of the review data.

This mainly consisted of applying the map-reduce paradigm in python scripts and assigning the different attributes e.g. brand, main category, granular category, verified purchases etc. as key values and then aggregating the values either by summing them up or averaging over them. All the code used for the map

reduce jobs can be found in my GitHub in the mr-jobs folder. The outputs of the map-reduce procedures are stored in .csv files for visualisation.

Together with the map-reduce procedure I also performed some time-series analysis on parts of the data in order to visualise yearly and monthly performance of top brands and categories as well as sales of electronics as a whole. In order to perform these analyses I had to export the JSONL data to a postgreSQL database and join the data on the parent-asin attribute, thereafter this data needed to be extracted into a python environment for visualisation. The code utilised for the creation of the tables, the export of the data to postgreSQL, the joining of the tables and the extraction of the joined data to python can be found in my GitHub under the sql-scripts folder, data-exports folder and db-extracts-scripts folder respectively.

This resulted in outputs that allowed me to gain a clear picture of the electronics landscape (as sold on Amazon) as well as other behavioural and meta-data aggregations as mentioned in the next sub-section. Overall this provides information to optimise different operational processes such as which products to by, which categories to focus on and what constitutes a successful/valuable product review.

1.3 Results

The following summary statistics and trends were calculated with all visualisation provided in the next section:
Map-Reduce Summary Statistics outputs:

- Top 10 most reviewed brands (most popular).
- Top 10 most reviewed manufacturers (most popular).
- Top 10 most reviewed main categories (most popular).
- Top 10 most reviewed granular categories (most popular).
- Top 10 most reviewed stores (most popular).
- Average rating of most popular brands.
- Average rating of most popular manufacturers.
- Average rating of most popular main categories.
- Average rating of most popular granular categories.
- Average rating of most popular stores.
- Average number of review helpful votes by length of review.
- Average rating of verified and non-verified purchases.
- Average number of verified purchases by average review rating.
- Average number of helpful votes by image presence.

The outputs are saved in .csv format.

Time series outputs:

- Brand monthly and yearly performance (number of reviews).
- Granular category monthly and yearly performance (number of reviews).
- Total electronics monthly and yearly performance (number of reviews).

These outputs are saved in .xlsx format.

The data was thus summarised in a myriad of ways providing an overall view that can be used to make informed business decisions. Thus contributing value to an electronics retailer looking to start up or branch out.

2 Visualisation

2.1 Overview

To visualise the summarised data I made use of python and the graphical plotting library called matplotlib. All plots were generated in a Jupyter Notebook which can be found here.

Three different types of graphs were used nl. bar charts to show comparative performance, pie charts to show popularity share and line charts to visualise the time series behaviour.

2.2 Results

2.2.1 Brands

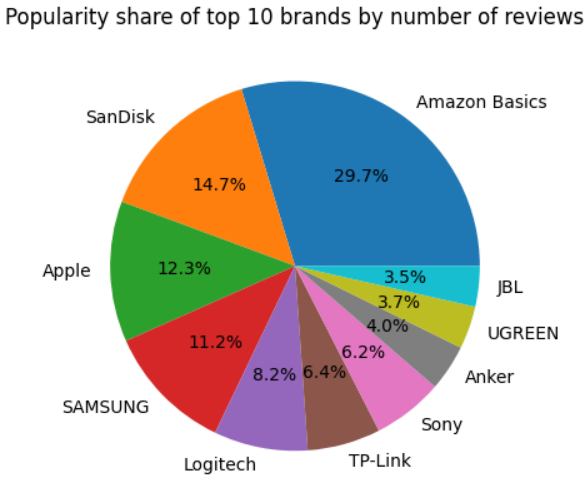

Figure 1: Popularity share of top 10 brands

Here we see the top 10 most popular brands reviewed expressed as a percentage of the number of reviews between them. Clearly Amazon branded products hold the vast majority followed by SanDisk, Apple and Samsung as the top brands. Therefore focusing on last three mentioned products as an electronics retailer would be most sensible.

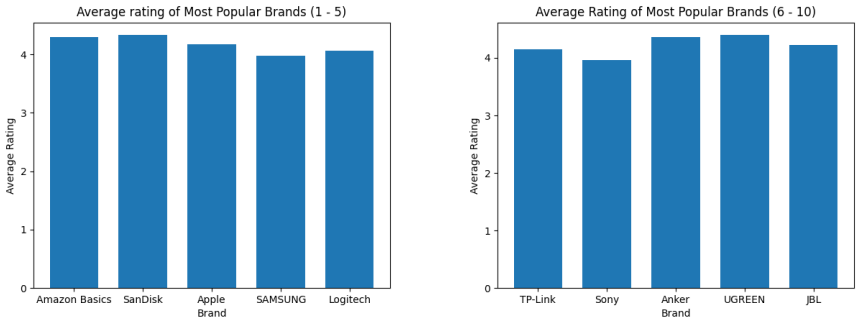

Figure 2: Average ratings of top 10 brands

Looking at these graphs we see that all of the top 10 brands are considered favourable by consumers, with only Sony and Samsung dipping slightly below an average of 4 out of 5.

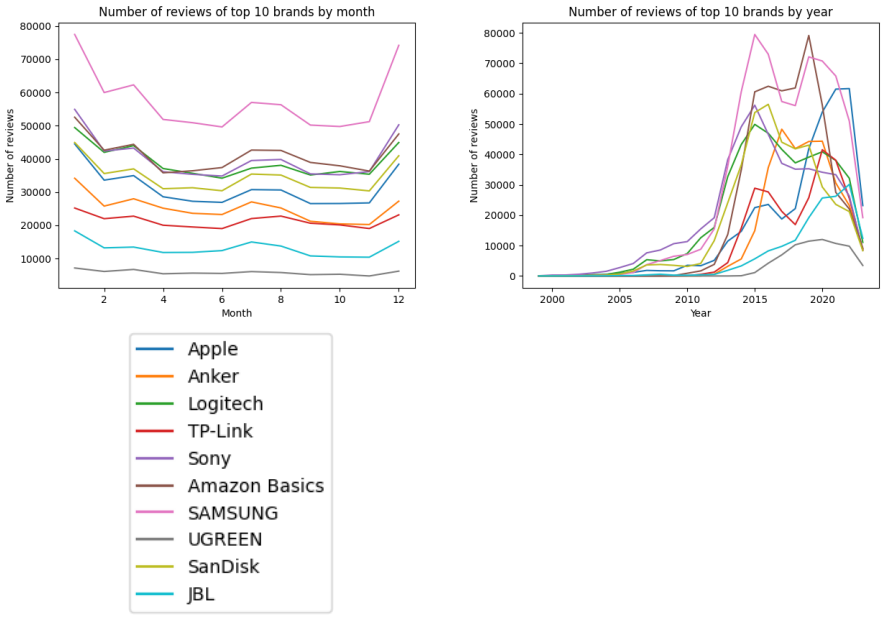

Figure 3: Time series performance of top 10 brands

For the monthly graph we see that all brands tend to have a slightly U-shaped curve, which corresponds with the Christmas holidays when sales tend to increase. Thus focusing marketing efforts during slower months that still exhibit good review numbers (perhaps between June and August) may increase overall revenue for an electronics retailer.

For the yearly graph we see a natural increase in number of reviews as technology improved and Amazon became more mature and popular. All the curves seem to dip towards 2024, which is the case because the values are only up to date to September 2023.

Focusing on brands that have a less steep drop-off may be wise for an electronics retailer, as this shows a less drastic fall in popularity.

2.2.2 Manufacturers

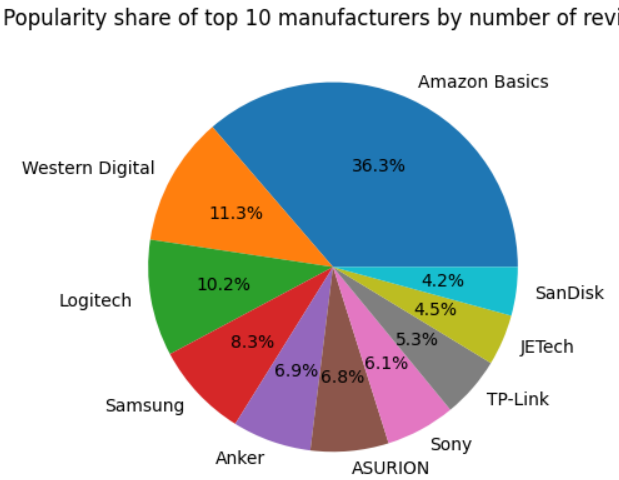

Figure 4: Popularity share of top 10 manufacturers

Here we again see Amazon as the main manufacturer, followed Western Digital, Logitech and Samsung as top performers. Thus building good relationships with these manufacturers will be good for business.

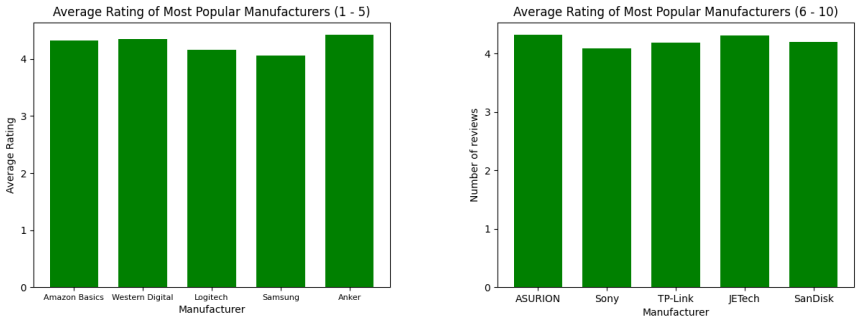

Figure 5: Average ratings of top 10 manufacturers

All manufacturers are highly rated, thus indicating that working with them and selling there products should be a safe and wise option.

I did not include the time series performance as it showcases the same behaviour

and information as the brand time series graphs and thus would have been redundant to include.

2.2.3 Main Categories

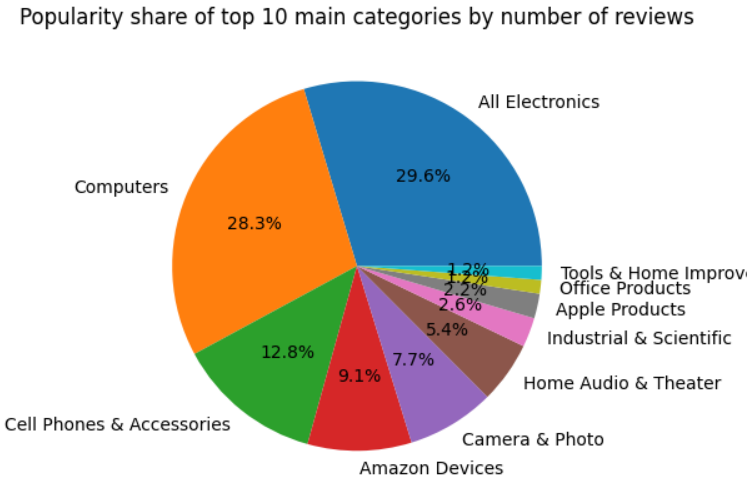

Figure 6: Popularity share of top 10 main categories

Here we see the top 10 most popular main categories. The graph indicates that the category of 'All Electronics' is most popular which is a general category. Thus focusing and providing product in the categories of computers, cell phones and accessories and devices may be most profitable. But in general focusing on any of these categories may be good business practice.

2.2.4 Granular Categories

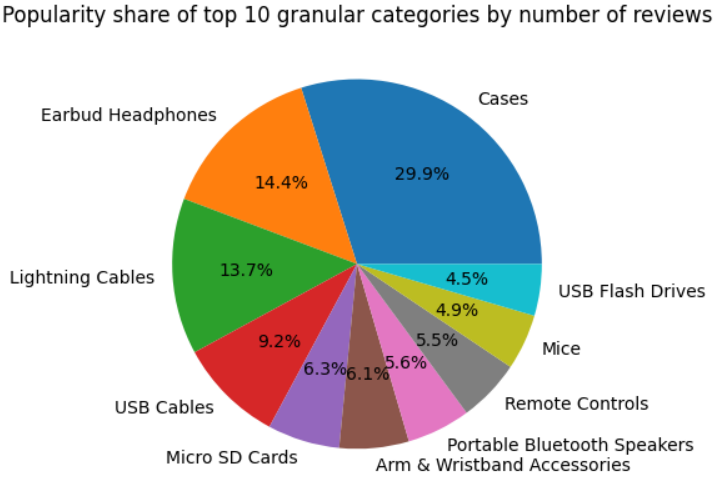

Figure 7: Popularity share of top 10 granular categories

This pie chart shows review share of the 10 most popular granular categories (more specific than the main category). The best performing categories are cases, earbud headphones, lightning cables and USB cables. These indicate the most popular specific items reviewed and thus provide a good idea of specific items to include in stock purchases.

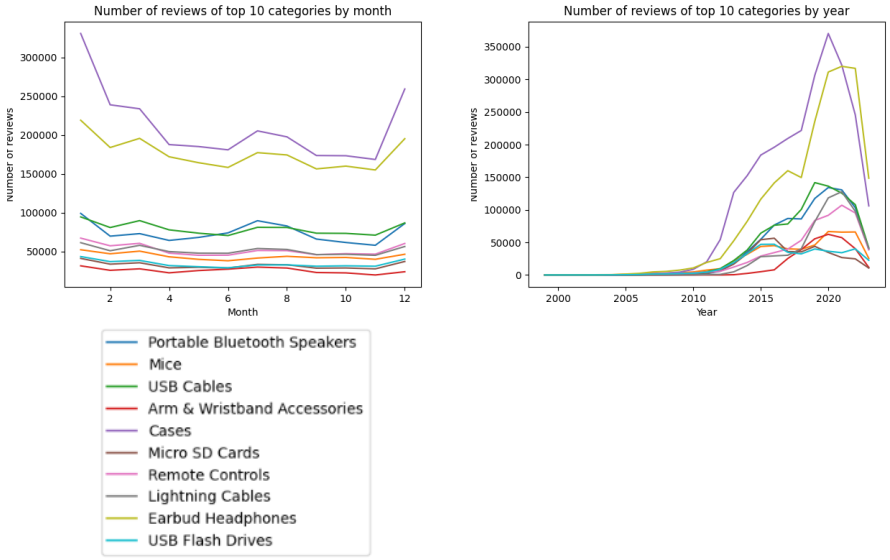

Figure 8: Time series performance of top 10 granular categories

These time series graphs exhibit the same behaviour as seen in the brands graphs and the same logic and reasoning holds for explaining the behaviour.

2.2.5 Stores

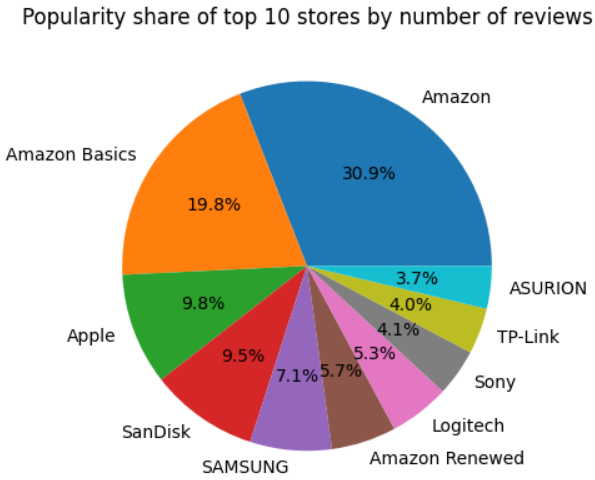

Figure 9: Popularity share of top 10 stores

This gives us an indication of the highest selling stores and can thus be used to gauge possible competition for an electronics retailer. The main seller is Amazon itself followed by Apple, SanDisk and Samsung which corresponds with some of the previous results.

2.2.6 Total Performance

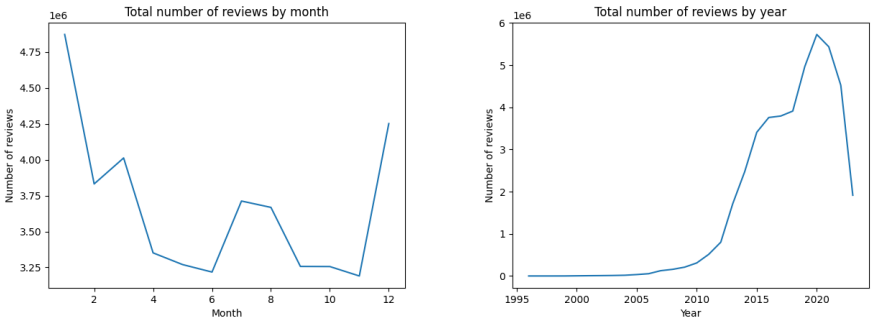

Figure 10: Time series performance of all products

Again, these graphs exhibit the same behaviour as for the brands and granular categories. Showing increased monthly sales over the Christmas holidays and increasing yearly then declining towards 2024.

This data can provide retailers with sales benchmarks to gauge there own performance and to gauge performance of initiatives such as advertisements and special offers etc.

2.2.7 Review Meta Data Evaluation

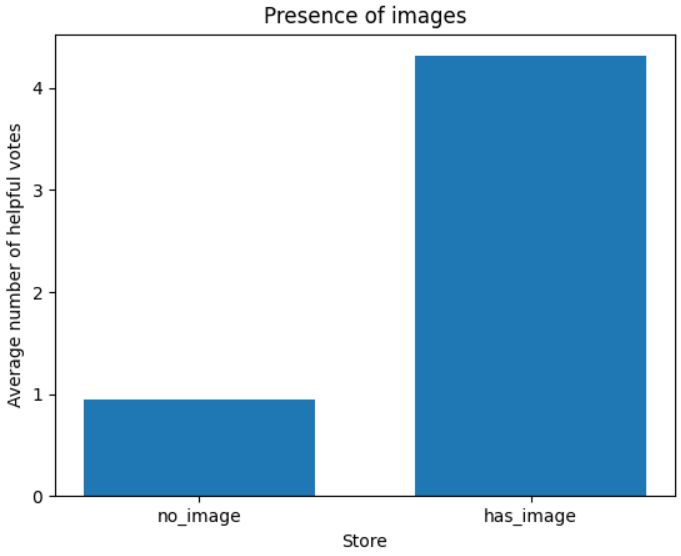

Figure 11: Average number of review helpful votes by image presence

Here we can clearly see that reviews that include actual images of the product are seen to be more than 4 times more helpful than reviews without images.

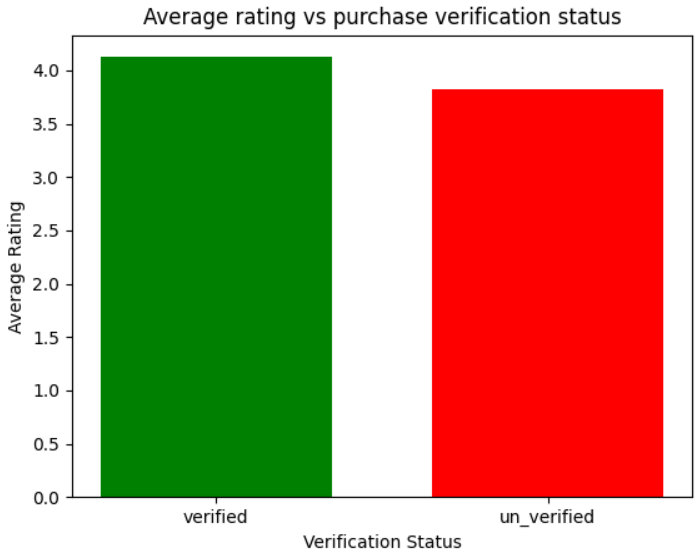

Figure 12: Average product rating by verified purchase status

Products with reviews with a verified purchase have higher average ratings than products with reviews without a verified purchase. This makes sense since people tend to by products that have better ratings.

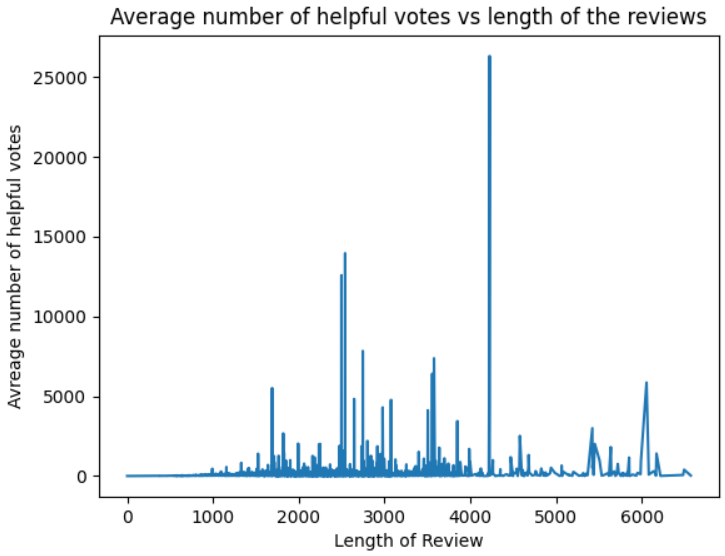

Figure 13: Average number of helpful votes by review length (number of characters)

We see here than longer review tend to be seen as more helpful with the caveat being that after a certain length this behaviour degrades. the 'sweet spot' seems to be around 2500 to 4500 characters. Any longer or any shorter results in less average helpful votes.

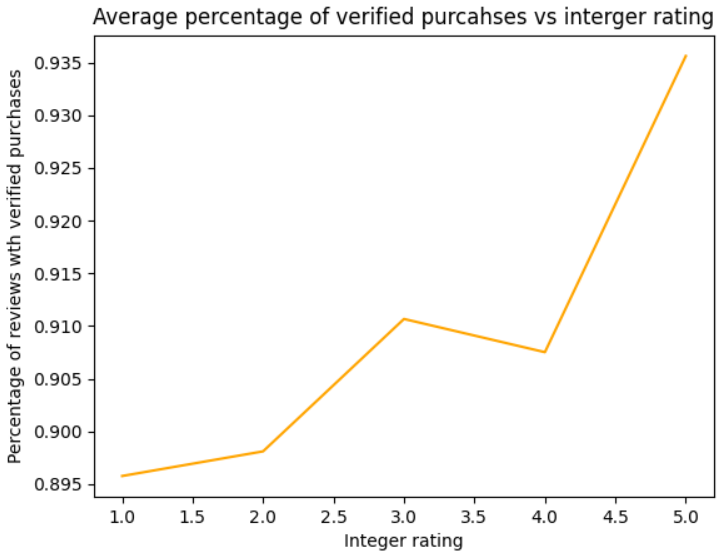

Figure 14: Percentage of reviews with verified purchases by product integer rating

We see the higher the rating the higher the percentage of verified purchases. Thus 'better' products have a higher successful purchase rate which aligns with common sense.

3 Conclusion

The Amazon Reviews 2023 Electronics dataset proved to be a valuable source of big data that has the potential to guide electronics retailers' business decisions and help optimise operational processes.

The map-reduce paradigm as well as other analytical techniques and technologies (such as database construction, JSON data manipulation, python scripting etc.) were implemented to produce visualisation that give a meaningful overview of the dataset and can be used in decision making and business optimisation processes of applicable electronics retailers.

4 Appendix

- Link to GitHub repository
- Link to graph images
- Link to datasets used for visualisation.

- Link to raw JSONL Amazon dataset.

References

[1] Y. Hou, J. Li, Z. He, A. Yan, X. Chen, and J. McAuley, "Bridging language and items for retrieval and recommendation," *arXiv preprint arXiv:2403.03952*, 2024.

[2] S. Mudadla. "Difference between ordinary json and json lines?" (2023), [Online]. Available: `https://medium.com/@sujathamudadla1213/difference-between-ordinary-json-and-json-lines-fc746f93d75e#:~:text=In%20summary%2C%20the%20key%20difference,and%20processing%20of%20individual%20objects.` (visited on 2024).

[3] A. de Wet, "Mit 805 assignment 1 - part 1," *MIT 805*, 2024.