**60.002 Project 2**

**Team 8**

1006878 *Gay Kai Feng Matthew*

1007182 *Koh Zi Kai Hubert*

1006988 *Destor Rose Evangeline Anne Dagman*

1006853 *Darrel Liew Jian Hong*

1006866 *Joel Lim*

Table of Contents

[Abstract 3](#_Toc164634256)

[Project Priority 3](#_Toc164634257)

[Project Process and Overview 4](#_Toc164634258)

[Data Sources 5](#_Toc164634259)

[Amazon Reviews and YouTube Comments 5](#_Toc164634260)

[YouTube Video Captions 5](#_Toc164634261)

[GSMArena 5](#_Toc164634262)

[Other explored data sources 5](#_Toc164634263)

[News outlets 5](#_Toc164634264)

[Other Social Media Platforms 6](#_Toc164634265)

[Official Customer Support & Patents 6](#_Toc164634266)

[Finalised Data Collection Method 6](#_Toc164634267)

[Selenium 7](#_Toc164634268)

[Google’s Youtube Data API 7](#_Toc164634269)

[Youtube Transcript API 7](#_Toc164634270)

[Explored Data Collection Methods 7](#_Toc164634271)

[Web Scraping Tools 7](#_Toc164634272)

[Google’s YouTube Data API for Captions 7](#_Toc164634273)

[Final Program 7](#_Toc164634274)

[EDA Explored 8](#_Toc164634275)

[EDA Model 8](#_Toc164634276)

[Graphing 8](#_Toc164634277)

[Benchmarking 10](#_Toc164634278)

[Constant reiteration of the data architecture 10](#_Toc164634279)

[Use of LLMs to guide the coding process 11](#_Toc164634280)

[Finding the appropriate help online 11](#_Toc164634281)

[Taking sufficient breaks 11](#_Toc164634282)

[Appendix 11](#_Toc164634283)

[Citations 12](#_Toc164634284)

# Abstract

The manual method of researching and making decisions processes in product design and development (PDS), is time consuming, labour intensive and prone to human error. This reduces the overall efficiency of recommending improvements for a product and analysing its competency against its competitors, alongside the continual shift in consumer preferences.

This report thus explores the opportunities created by automating the product design processes, specifically for tablets, from our data sources, finalised model, explored exploratory data analysis (EDA) method and its limitations. By leveraging automated algorithms, companies stand to gain significant time efficiency. With the obviation of manual labour, companies also enjoy the potential reduction of labour costs and overhead expenses. Human resources can also be directed to focus on the creative aspects of the design process.

However, design problems are inherently multifaceted and dynamic, leading to unforeseen variables and contingencies. In our case specifically, features such as the ecosystem surrounding the tablet and the user interface may be easily overlooked during the automated decision-making process due to the subjective nature of individual consumers. Consequently, human designers are still necessary to adapt and iterate in response to evolving the design requirements. Leveraging intuition and creativity, these designers can devise innovative solutions to address complex design challenges.

Overall, this program is not to replace human designers but rather to be used alongside human designers as a tool to speed up and enhance the research phase of the design process. This streamlines the product development process, reducing costs, fostering innovation, and ultimately enhancing competitiveness in the market.

# Project Priority

In line with the project's objectives in emphasizing versatility and accuracy, the developed program must possess the capability to analyse any given tablet model as input and give a precise and reliable design recommendations as output. Hence, it is important for our data sources to: (i) encompass a comprehensive information on all tablets available in the market, (ii) without compromising on their accuracy and reliability, and (iii) are up to date with the latest information available on the internet, ensuring the program's relevance and efficacy.

Moreover, the program should be able to facilitate comparative analyses between the target tablet and its competitors of similar objectives, enabling a comprehensive assessment of their features and functionalities. Visual representations of this comparison should be provided for ease of manual analysis, accompanied by succinct summaries outlining the distinguishing features of the competing tablets.

In summary, the program's key attributes must include versatility, accuracy, dynamism, and comprehensive comparative capabilities, to automate the tablet design process effectively and accurately.

# Project Process and Overview

The program is structured to automate the laborious and time-consuming tasks involved in researching and providing design recommendations, via the utilization of large language models (LLMs) that are capable of generating reasonable inferences and recommendations and pertinent data scraping techniques.

The final program takes any given tablet model (referenced as focal tablet) as input and outputs a design recommendation that considers the business direction and the tablet’s competitors as summarised below

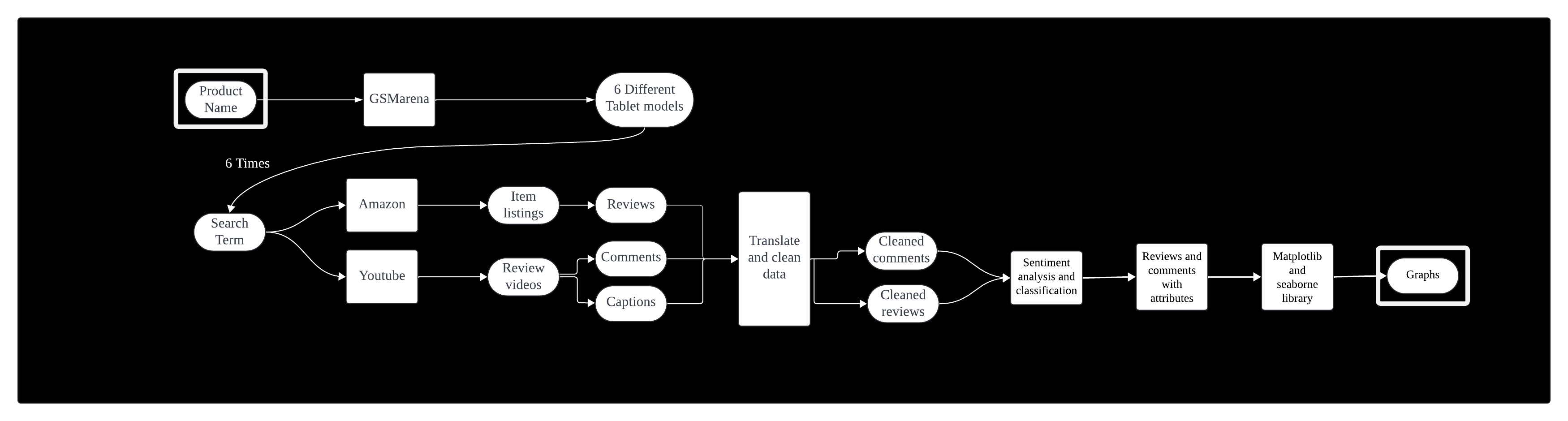
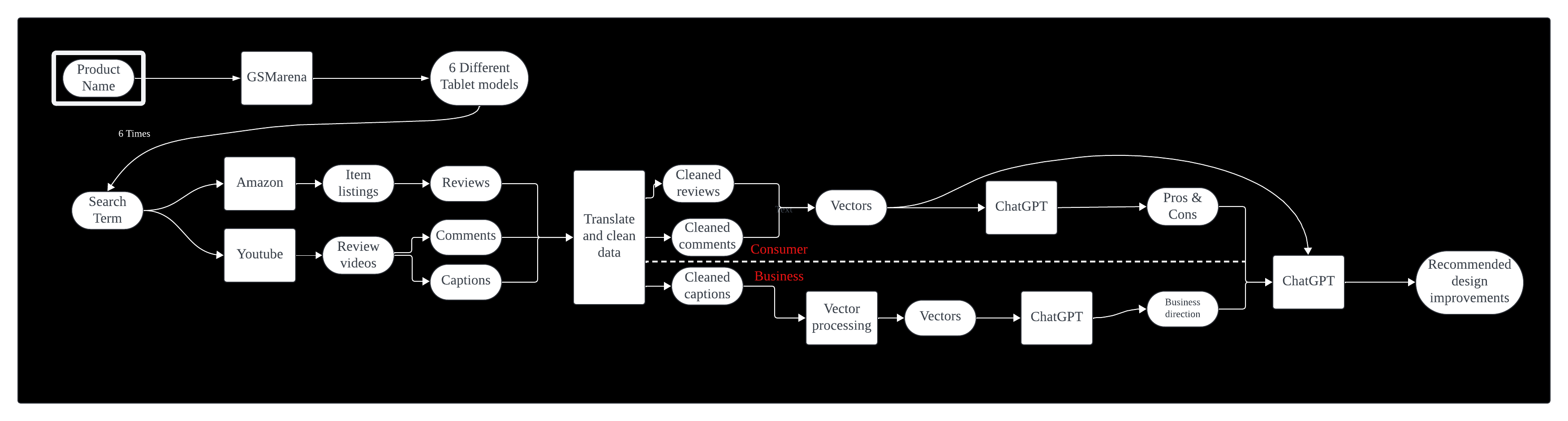


Figure 1: Data Architecture of the Final Model

The use of EDA for further analysis and benchmarking of the final model was attempted. However, its accuracy and performance were not satisfactory and needed further testing and fine-tuning. Nevertheless, it will be discussed in the report to showcase its potential use and application in data analysis with the data architecture as shown:

Figure 2: Data Architecture of EDA

# Data Sources

Many data sources were considered when making the final program. However, we decided to only include the following:

### Amazon Reviews and YouTube Comments

In the manual research process, initial steps involve examining consumer sentiment regarding the tablet, discerning preferences and issues encountered with the tablet. Of the various platforms possible, YouTube comments and Amazon official storefront reviews are chosen due to their consistent yield of the requisite data. Hence, our consumer consensus data comprises of YouTube comments and Amazon reviews. Both of which are treated identically for analytical purposes.

### YouTube Video Captions

Our design recommendations must align with the product's business trajectory. For instance, in the context of the Samsung Galaxy Tablet S9 FE, designed to offer an affordable alternative within the high-end tablet market, recommendations should support its affordability while enhancing competitiveness against other high-end tablets. This strategic business direction was discerned manually through the analysis of YouTube reviews for the tablet. This process can be automated by aggregating the captions of tablet reviews on YouTube.

### GSMArena

In the manual research process, identification of competing tablets entails seeking devices with similar objectives to the focal tablet. In the case of the Samsung Galaxy Tablet S9 FE, this translates to identifying other high-end tablets as competitors. To automate this process, GSMArena serves as the primary data source for competitors. GSMArena is a regularly updated website housing specifications of all existing tablets. Utilizing both the information available on GSMArena and its search functionality allows for the collection of tablet specifications and identification of competing devices within the relevant market segment, all within the confines of the same website.

# Other explored data sources

Some data sources were explored but were not used in the final prototype program. This was due to either inconsistency or incomplete information.

### News outlets

One of the news outlets explored was The Verge, an American technology news website operated by Vox Media, to gather data concerning the business direction of the focal tablet. However, this source was ultimately excluded due to its inconsistency in providing relevant search results. For instance, searching for "Apple iPad Pro" on the website yields articles typically consisting of only two paragraphs, offering speculative insights from the author rather than concrete information about the tablet's direction.

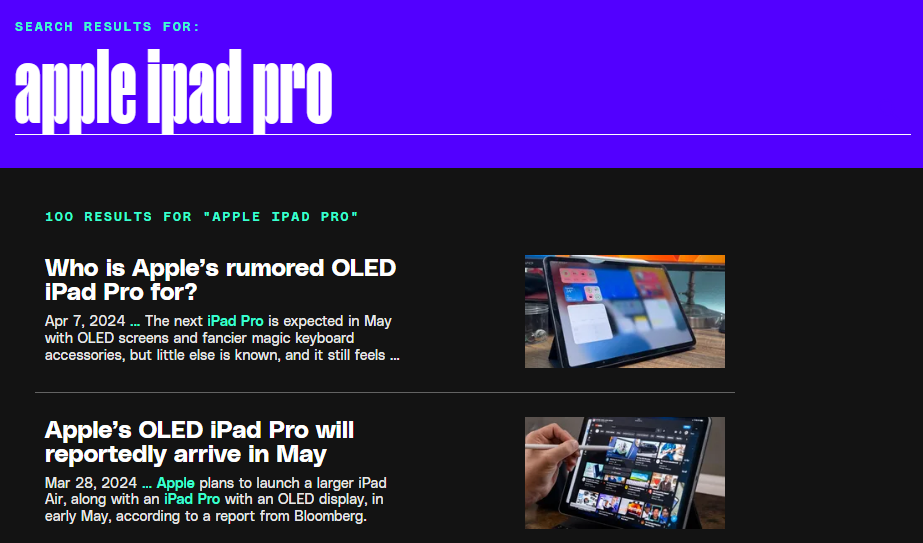


Figure 3: Verge search results

### Other Social Media Platforms

Reddit and Twitter were also explored to accurately represents consumers of diverse backgrounds and opinions. However, a significant portion of the Twitter search results comprises posts containing links to external blog pages, rendering the collected data either inconsistent or irrelevant. On Reddit, search results predominantly feature posts where users seek advice in choosing their tablet. Moreover, the demographic on Reddit tends to skew towards a younger, more tech-savvy audience. These factors collectively result in posts and comments appearing as remarks rather than inquiries, potentially leading to confusion for the model and its results. While Reddit serves as a valuable resource for understanding consumer preferences regarding tablets, it may not accurately reflect the consensus among users regarding the focal tablet.

### Official Customer Support & Patents

Patents and official customer support websites such as Apple Support and Samsung Support were investigated to uncover schematics and business direction. Ideally, processing information on the physical structure of tablet components would have yielded a detailed recommendation on the component structures. Unfortunately, the patents were not accessible on the internet, and the information retrieved from the support pages was largely irrelevant to the product's business direction.

# Finalised Data Collection Method

With the data sources finalized as YouTube comments and Amazon reviews for consumer consensus, and YouTube captions for business direction, the next step involves collating the information. This process is facilitated by utilizing Selenium for web scraping Amazon reviews, Google’s YouTube Data API for YouTube comments, and youtube-transcript-api for retrieving YouTube captions.

### Selenium

Selenium, a Python web scraping library that operates by launching a browser on the local machine. This characteristic makes it less prone to detection and more effective on dynamic websites with heavy JavaScript usage. Moreover, it is easy to implement in a program.

Selenium was used to navigate the Amazon website. Thereby, accessing reviews pertaining to the focal tablet and extracting pertinent data points including user ratings, comments, and review dates. This approach allows for the efficient collection of a substantial volume of real-world user feedback, serving as a valuable asset in informing our design recommendations and comprehending customer preferences.

### Google’s Youtube Data API

The Youtube data API is an API created by Google. The implementation of it is similar to what was covered in DAI AID class.

### Youtube Transcript API

Youtube Transcript API is a python library that sends the transcript with the Youtube video ID and preferred language as the input. The output is a string with the caption and the timestamp of the string.

# Explored Data Collection Methods

### Web Scraping Tools

Requests and BeautifulSoup4 libraries were initially utilized but are faced with numerous anti-scraping measures on GSMArena and Amazon. Such anti-scraping mechanisms included encryption of HTML code with JavaScript, Captcha requirements. Efforts were made to circumvent these obstacles by randomly generating valid user agents for requests. However, Amazon redirected requests to a sign-in page when attempting to access reviews beyond the first page. Subsequently, attempts were made to reverse engineer a Chrome extension capable of scraping reviews, but unfortunately, no success was achieved. Given the multitude of issues encountered, Selenium emerged as the preferred data collection method due to its ability to navigate through anti-scraping measures effectively.

### Google’s YouTube Data API for Captions

The YouTube Data API, initially intended for retrieving comments, was also considered for obtaining captions. However, it necessitated authorization from the video author to access the captions. Consequently, this approach was abandoned in favor of utilizing the youtube-transcript-api.

# Final Program

The final program created makes use of retrieval augmented generation (RAG) and ChatGPT to provide the design recommendation. Firstly, we vectorized the data sources. For each of the tablet, the vectorized consumer consensus data and the question “What are 3 pros and 3 cons of the tablet? Provide an overall summary of the tablet.” are passed into ChatGPT with RAG. ChatGPT then returns a contextualized response about the tablet. Meanwhile, the YouTube captions of the tablet of focus are subjected to the same process but with the question “Infer what the business direction of the company behind the tablet.”

These responses are then input into the final question “What are some general feedback about the tablet? Provide some recommended design changes to the product that aligns with the previously mentioned business direction.”

# EDA Explored

Exploratory Data Analysis (EDA) was attempted to be implemented into the final program (as shown in Figure 2). However, the accuracy and reliability of the EDA model was not up to standard.

### EDA Model

The EDA model was exclusively applied to the consumer consensus data. This decision stemmed from the potential usefulness of the results when applied to YouTube captions. The selected categories for tablets include ergonomics, build quality, portability, display, performance, connectivity, audio, and cost. Each comment undergoes analysis using the BART model from Hugging Face for each category. Consequently, each comment is processed eight times, with each result indicating a true or false threshold whether the comment pertains to the respective category. Sentiment analysis is also conducted on each comment using the RoBERTa base model from Hugging Face. Additional details can be accessed in the Jupyter notebook from the Appendix.

### Graphing

Following the categorization of comments and sentiment analysis, the subsequent step involves visualizing the gathered information through bar graphs, radar graphs, and pie charts.

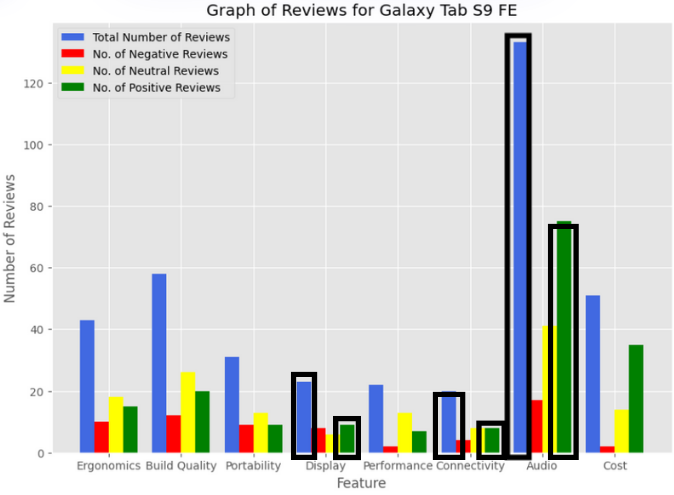


Figure 4: Bar graph for a single tablet model

The bar graph provides an overview of consumer consensus regarding the tablet. For instance, if a comment or review pertains to display, connectivity, and audio, and expresses a positive sentiment, it contributes to the count displayed in the corresponding category of the bar graph. This visualization method aids in comprehending the overall perception of the tablet among consumers. This graphing procedure is replicated for each tablet under examination.

Calculating the ratio of positive comments/reviews to the total number of positive and negative comments/reviews yields a radar graph depicting the consumer rating of each component. Ratings range from 1 to 5, with a score of 5 indicating the most positive sentiment regarding that component of the tablet.

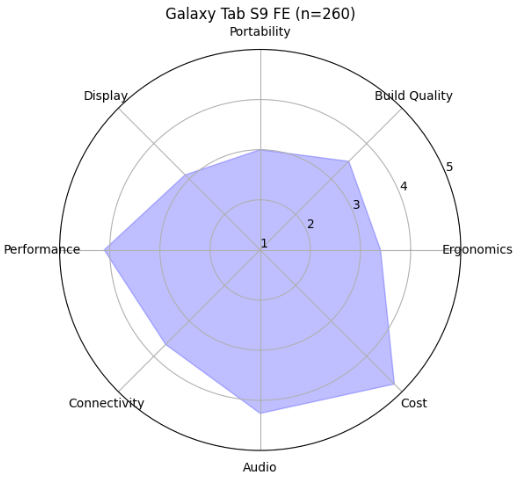


Figure 5a: Radar graph of a single tablet model

This radar graph serves to visualize the strengths and weaknesses of the tablet model, thus facilitating the decision-making process for the tablet redesign. This graphing procedure is replicated for each tablet under consideration.

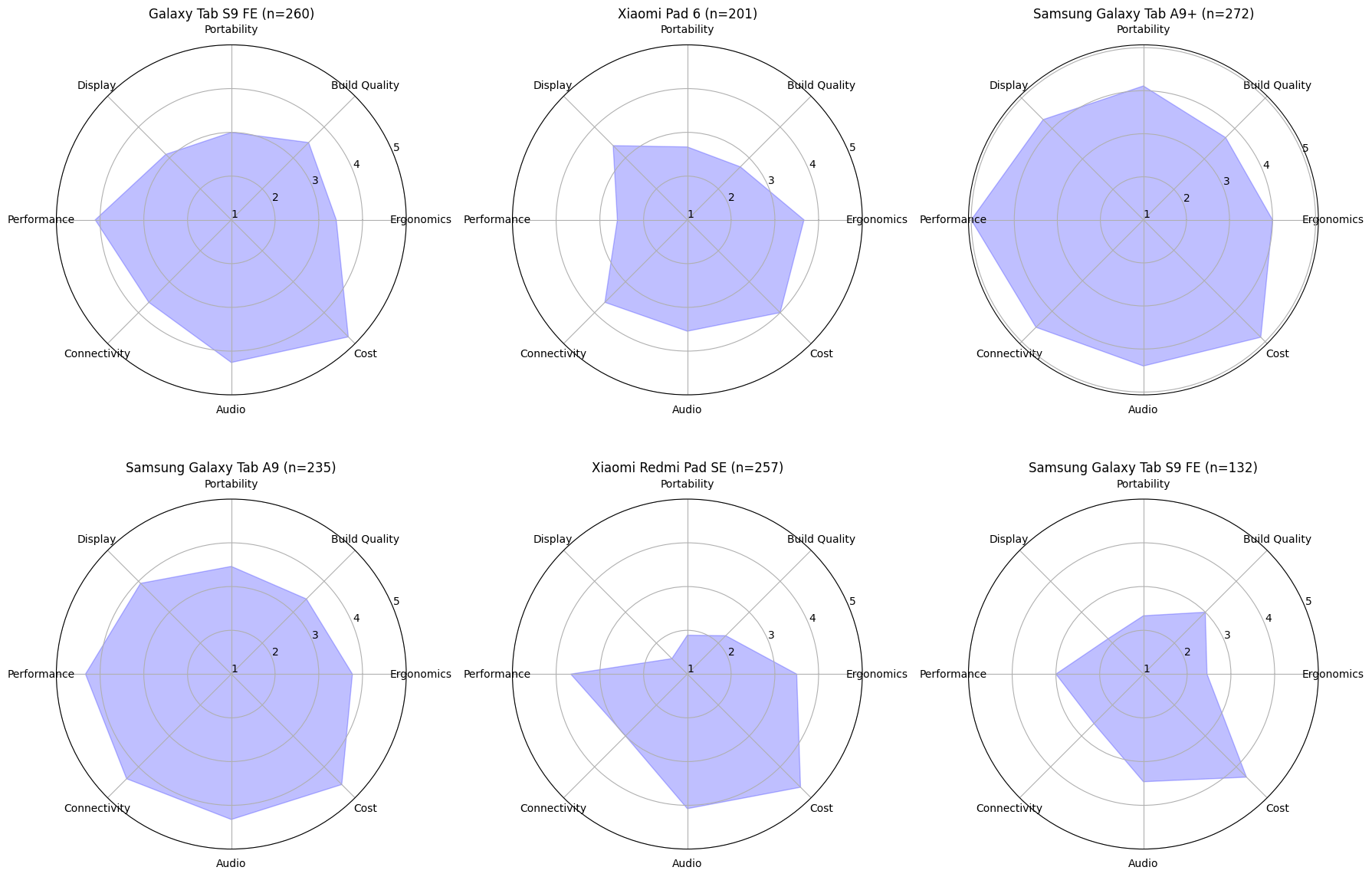


Figure 5b: All radar graphs for focal tablet and its competitors

The collection of radar graphs serves to showcase the consumer consensus of each tablet, thereby facilitating the decision-making process when determining the next business direction. For instance, if the client aims to enhance the performance of the Galaxy Tab S9 FE, insights gleaned from studying the components of the Galaxy Tab A9+ can inspire design recommendations for the Galaxy Tab S9 FE.

To determine the consumer priority for each component, we can generate a pie chart aggregating all mentions of a component. This involves consolidating the blue bars from the bar graph into a single pie chart. The resulting visualization illustrates consumer priorities in a tablet, operating under the assumption that mentions of a component reflect consumers' indications of their priority for that aspect in a tablet.

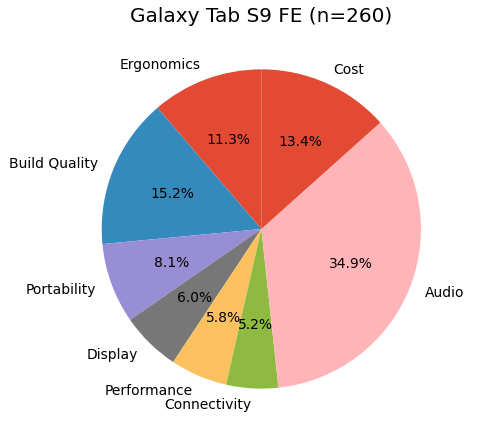
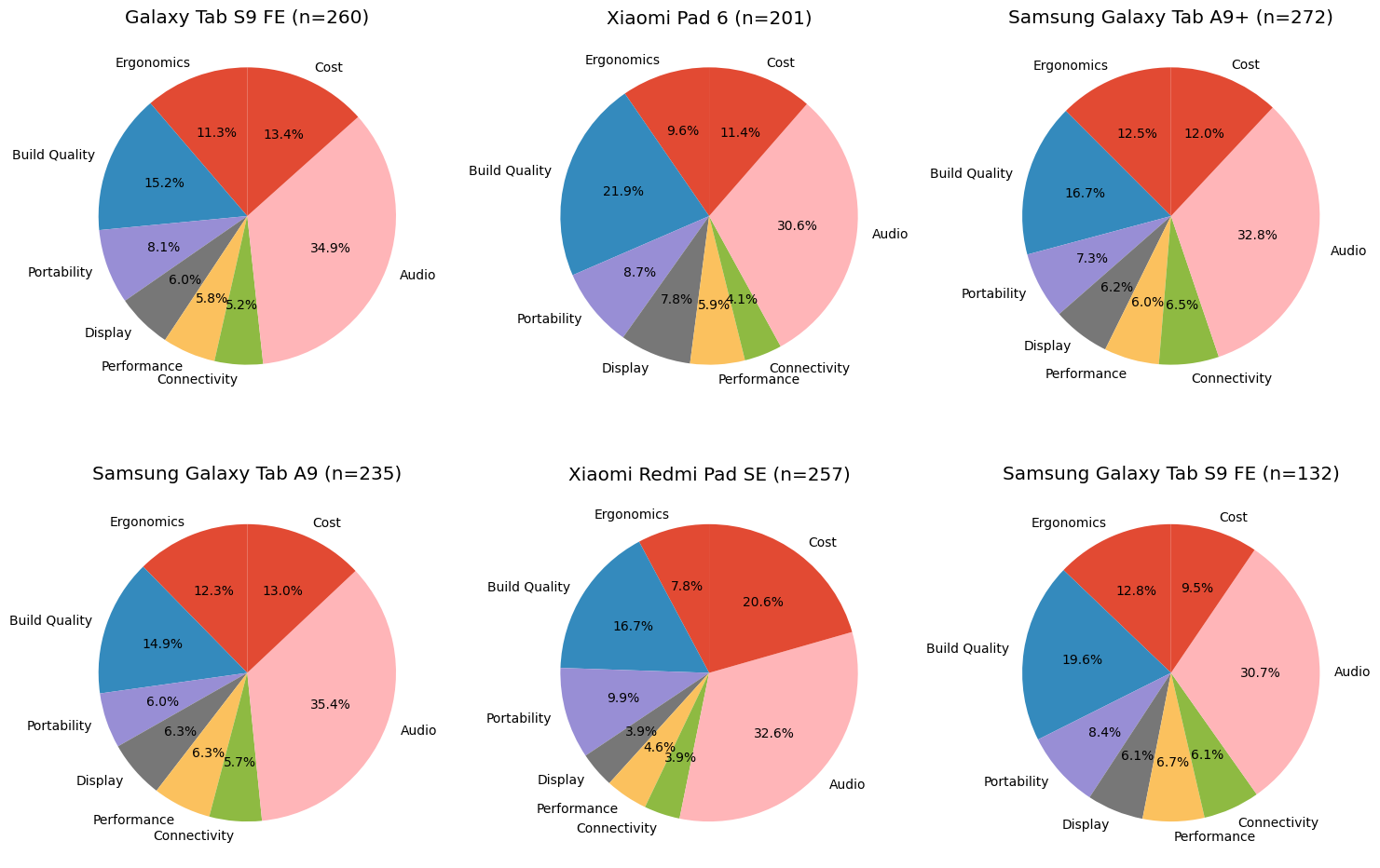


Figure 6a: Pie Chart of a single tablet model

For example, in the case of the Galaxy Tab S9 FE, the pie chart may indicate that consumers prioritize audio and build quality. This process is replicated for each tablet.

  
Figure 6b: All Pie Charts for focal tablet and its competitors

The collection of pie charts serves to showcase the consumer priority of each tablet, thereby facilitating the decision-making process when determining the next business direction. For instance, if the client aims to target customers of the Xiaomi Redmi Pad SE, the Galaxy Tab S9 FE can focus on lowering the price of the Tab S9 FE, potentially sacrificing connectivity if necessary.

# Benchmarking

The EDA (Exploratory Data Analysis) model did not make it to the final model due to its unsatisfactory accuracy, as revealed during benchmarking. Collating 15 comments and 15 reviews showed that sentiment analysis achieved an accuracy of 86.7%. However, categorization accuracy was only 25%.

Upon closer examination of the comments and reviews and their categorization, it was observed that the BART Hugging Face model tended to misclassify short and irrelevant comments or reviews that were not related to tablets. This rendered the decision to classify each comment eight times irrelevant.

While potential solutions such as adjusting category names, utilizing a different Hugging Face model, or fine-tuning the current model could mitigate this issue, these options were deemed excessively time-consuming and were not fully explored.

Limitations

The poor performance of the EDA posed challenges in automating the process of verifying the accuracy of our ChatGPT responses. This issue was further compounded by the fact that the responses detailing the pros and cons of each tablet were fed back into ChatGPT, potentially leading to the generation of nonsensical responses on top of existing ones. To mitigate this, we can keep automating the verification of data sources for the response and requesting another response until an appropriate one is received could be implemented.

Additionally, the sample size of Amazon reviews is significantly smaller compared to the number of YouTube comments. This discrepancy arises from the absence of duplicate postings for tablets on Amazon. Moreover, due to the page limit of reviews being set to 10, a maximum of 100 Amazon reviews can be collated. In contrast, YouTube allows for multiple videos to be posted about the same topic, with each video having no limit on the number of comments. With the max result variable in the program set to 10, the ratio of reviews to comments reached approximately 100:250. This disparity skews the representation of consumer feedback towards the YouTube audience.

Lessons Learned

The journey of creating the program involved a significant amount of trial and error, which is typical for any project. However, much of this can be mitigated through proper planning. Additionally, mastering coding and basic HTML knowledge typically requires self-teaching and hands-on experience to resolve the issues encountered.

### Constant reiteration of the data architecture

The final program differed significantly from the original plan and expectations due to the discovery of additional information during the development process. Consequently, the key lesson learned is to exhibit flexibility and openness to change direction based on new discoveries. Furthermore, it is crucial to invest serious thought into each iteration of the data architecture to avoid unnecessary work while maintaining efficiency and focus throughout the project.

### Use of LLMs to guide the coding process

ChatGPT and its variations, such as Phind and Gemini, are excellent tools that expedite and guide the coding process. They assist in debugging issues, suggesting solutions, and even writing code for the user when used correctly. However, it is crucial to avoid overreliance on Large Language Models (LLMs). It is easy to fall into the trap of blindly sending questions to the LLM, resulting in not only useless but also potentially misleading outcomes. Therefore, it is essential to use these tools judiciously and complement their capabilities with critical thinking and domain expertise.

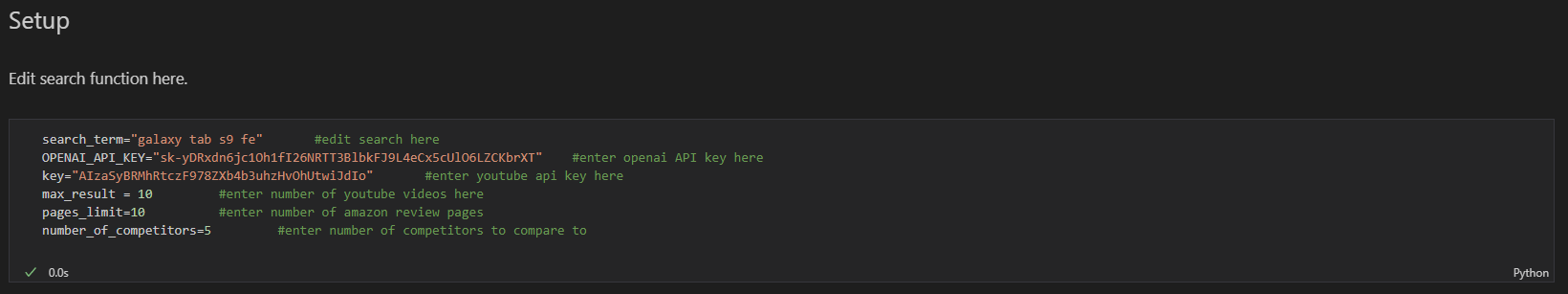
### Finding the appropriate help online

Streamlining the search for coding solutions is aided by adept Google searching for relevant terms. When encountering unfamiliar libraries, watching short guides or tutorials on YouTube can provide valuable insights. Exploring alternatives via Google and ChatGPT before finalizing library choices ensures optimal selection and understanding of the libraries’ functionality before implementation.

### Taking sufficient breaks

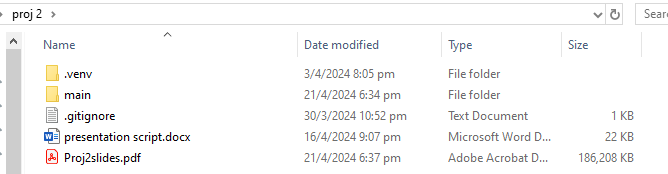
Taking breaks during program creation is essential. We have observed that adequate rest not only enhances our mental well-being but also improves efficiency and proficiency when focusing on development. Breaks help maintain focus and clarity of thought, which are crucial when tackling complex programming tasks.

# Appendix

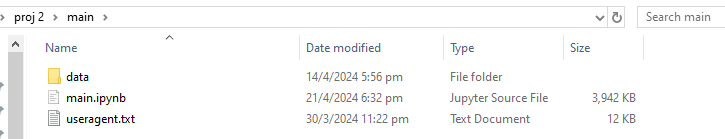
Example of Input for the program

File directory of Submission folder:

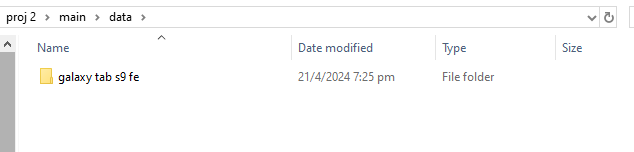
Contains presentation slides and script



Contains the main program in a Jupyter File



Contains the data for each previously completed search term



Contains the information for the focal tablet and relevant competitors. 