University of Glasgow | School of Computing Science

Honours Individual Project Dissertation

eCommerce Analytics: Real-Time Dashboard and LSH-based Product Recommendation

Sea Hao Feng, Sam

March 27, 2019

Abstract

Existing research and works on big data analytics are catered more towards established organizations with financial and technical resources while the findings are inadequate for newly start-up companies.

In the eCommerce industry, new platforms will face the data sparsity issue of not having enough records of customers purchase pattern and the dilemma of selecting a suitable cost-effective analytics platform. The system developed provides an inexpensive yet effective analytics engine to new entrants in the eCommerce landscape through easy analytics and a scalable clustering algorithm for recommending products while avoiding the data sparsity issue.

The proposed solution is the development of a real-time dashboard and an LSH-Based Product Recommender. The system is implemented in non-proprietary software so that it is cost-free and its community of users can collaborate modify the system according to their domains.

The final product underwent four evaluations – two of it being the usability assessment of the dashboard through a quantitative survey and qualitative feedbacks while the other half is the investigation of key parameters that could possibly fine-tune the recommender's performance.

Education Use Consent

I hereby grant my permission for this project to be stored, distributed and shown to other University of Glasgow students and staff for educational purposes. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Signature: Sea Hao Feng, Sam Date: 27 March 2019

Contents

1 Introduction

With the world increasingly moving towards digitalization, 2.5 quintillion bytes of data are created everyday[1]. The massive amount of data generated proves too much for traditional data mining tools and thus the term was coined Big Data by Mougalas[2].

The introduction of Big Data led to a wave of businesses incorporating data-driven decisions into their daily operations. Facebook[3] provides omni-channel [1] analytics services to help their clients grow, Spotify[5] uses a recommendation system to recommend songs to its customers, Shazam[6] implements a hash function for its audio fingerprint algorithm in querying music, Netflix[7] recommends movies to its users and Amazon[8] provides relevant products to its customers real-time.

In other studies regarding Big Data Analytics in eCommerce, Lee[9] reported that standard conversion rates were increased by 5.9% with recommenders and that eCommerce customers were guided by recommendation systems on a regular basis. Akter and Wamba[10] also shared that real-time decision making is the decisive factor of an eCommerce's success and simplifying the analytics process and result for front-line staff is the key challenge for the company's management.

For this project, the clients are from Trollee which is a new eCommerce start-up that has a vision of penetrating the eCommerce landscape by differentiating themselves from the already saturated market in Singapore.

1.1 Motivation

Analytics can be utilized to better comprehend the trends derived from these collected data which can help organizations in making forecasts based on data-derived insights, anticipating business opportunities and challenges to stay ahead of the curve.

In the context of eCommerce, data analytics can improve the understanding of customers behavior, optimize price models, enhance business intelligence, better target recommendations and promotions to different groups of users.

However, comprehensive analysis on a large data reservoir might dilute the focus of business objectives because the time and resources used to track non-significant metrics could have been allocated to gain more insights on key metrics that matters.

[1] Omni-channel is a "multi-channel approach to sales that provides customers with a seamless shopping experience" regardless of the device used[4]

1.2 Aims & Objectives

This project aims to devise and implement a novel analytics engine that provides easy insights and Business Intelligence (B.I.). via real-time dashboard visualizations and a product recommender that operates on Locality Sensitive Hashing (LSH). Having easy analytics helps companies to make business decisions more efficiently without filtering through comprehensive traditional analysis systems, assisting stakeholders to allocate resources for more pertinent issues.

1.3 Outline

This chapter introduced the motivations and aims behind the project. The remainder of this dissertation will cover the following chapters in sequence:

- Literature Review – This chapter reviews the terminologies in big data, types of visualization, analytics archetypes, existing dashboards in the market and recommendation systems by established eCommerce companies.

- Analysis/Requirements – This chapter includes the project's requirements by establishing eCommerce terms with the client along with functional and non-functional requirements.

- Design - This chapter contains the blueprint of the project that includes the system architecture, design decisions behind the wireframes and recommender methodology.

- Implementation – This chapter goes into details of the implementation process according to the requirements and design mentioned earlier.

- Evaluation – This chapter discusses the various ways the product was evaluated along with the evaluation metric used and variable tuning that can satisfy the requirements.

- Conclusion - This chapter summarizes the report, identifies future enhancements opportunities and presents a reflection on this project.

2 Literature Review

This chapter reviews terminologies and existing works on big data analytics such as types of data formats, data attributes, data preparation, visualization techniques, analytics archetypes, existing embedded analytics tools and recommendation models adopted by the eCommerce industry.

2.1 Types of data formats

Structured data in relational databases is no longer the only data format these days. Advancement in technology led to new media contents such as video-sharing websites (YouTube, Vimeo etc.) and social media (Facebook, Twitter, Instagram etc.) which produce unstructured data. This section will explain the two main data formats - structured and unstructured data.

2.1.1 Structured data

Structured data is normal data that have relational keys and can be stored in a pre-defined formatted data warehouse which is typically a relational database. Examples includes customer data (e.g. name, gender, region, ethnicity) and product attributes (e.g. category, size, colour, pricing information).

2.1.2 Unstructured data

Unlike structured data, relation databases are unsuitable for unstructured data because its structure does not adhere to pre-designed data schema. Instead, it must be stored in a non-relational database because its information could not fit properly in a relational database. Examples include customer emails, call centre support tickets, product reviews, social media comments, search log, page views etc.

2.2 Data Attributes

A data attribute describes the nature of a data object. The two main types of data attributes are Quantitative and Qualitative data.

Quantitative data is information that can be measured in numerical terms (e.g. age, shoe size, page visits) without the need for further interpretations and it is usually examined via statistical analysis. It can be further broken down into discrete and continuous data where discrete data can only accept certain values such as using whole numbers to record the number of household pets (half a pet is invalid) while continuous data can take any values within a range like a person's height, weight etc.

Qualitative data relates to non-numerical details recorded through open-ended questions, interviews and observations for theme analysis through subjective interpretations explaining how and why a particular pattern exists in the dataset. Its subset consists of nominal, binary and ordinal attributes. Nominal attributes present categorical values (e.g. type of colours), binary attributes contain only 2 states (e.g. Gender) and ordinal attributes describes a perceived ranking order of values.

2.3 Data Preparation

With over 2.5 billion gigabytes of data created every day and an upwards forecast of this trend[1] due to advancement in reducing information search-related costs and accelerated surge in data availability[11], it is expected that today's world will be overloaded with information. This section discusses the importance of data preparation through the investigation of information overload and data preparation methodologies.

2.3.1 Information Overload

Even though there is a popular proverb of "more data will result in better analysis", this might not be applicable every time; on the contrary, it could cloud an organization's business acumen. Given the accelerated availability of data, there will be a point where the human information processing capability hit its limit because decision-makers must deal with more data that they can handle. The relationship between decision making performance and information load can be portrayed as a negative quadratic curve (See Fig. 2.1) where the former declines after reaching the maximum amount of data it can handle[12][13][14].

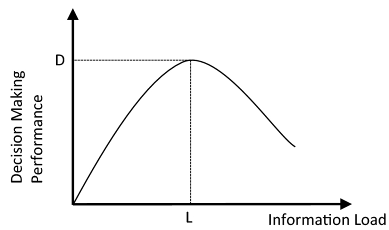


Figure 2.1: *Decision Making Performance VS Information Load*

Basex estimated that information overload costs the economy $650 billion in terms of in terms of "lost productivity and innovation"[15]. This was also reaffirmed by academic and industry reports, which revealed that 80% of marketers were unsure of turning data into insights and organizations only use 5% of the data they have collected [16][17][18]. In order to reduce information overload, we will look at a methodology based off the Cross-Industry Standard Process for Data Mining(CRISP-DM) [19] in the following section.

2.3.2 Data Preparation Methodology

The proposed methodology (Fig. 2.2) suggested by P´erez et. al[20] consists of two main preparation stages (General Data Preparation and Specific Data Preparation) and further broken down into two and three sub-phases respectively, with each sub-phase having its own tasks:

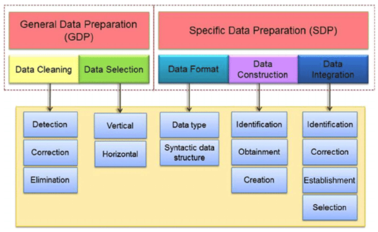


Figure 2.2: *Tasks of the data preparation methodology[20]*

2.3.2.1 General Data Preparation

This stage relates to a generic set of procedures which are applicable to other domains due to it being a separate cause from the specific analysis to be accomplished. Its sub-phases are as follow:

- Data Cleaning
  The aim of this process is to identify, calibrate and remove irregularities in the dataset such as spelling errors, duplicate records, incomplete information and erroneous parsing of data fields from different systems.

  In addition, methods shared by[21] to deal with outliers included:

  1. Discarding tuples with outliers
  2. Replacing missing entries with expected values enumerate environment.
  3. Estimating outliers' values through robust estimation techniques

- Data Selection
  This step is to select relevant records (rows) and data attributes (columns) that provide information of interest.

2.3.2.2 Specific Data Preparation

This stage stipulates a list of actions for making the dataset more relatable to the specific goal of the analysis. Its sub-phases are as follow:

- Data Format
  The purpose of this sub-phase is to modify the data type and the attributes' syntactic structure to a more suitable format for processing.

- Data Construction
  The next stage of Specific Data Preparation is to patch incomplete data and generate new attributes for the research domain if necessary.

- Data Integration
  The final preparation before deployment for visualization and modelling, the data will be integrated and checked for merge conflicts to fix any errors found in the integration process.

2.4 Data Visualization

Card, Mackinlay, and Shneidermann[22], all of whom stated that computer-generated, interactive visualization helps to increase the understanding of data. This section will discuss the benefits and various techniques of data visualization.

2.4.1 Benefits of Data Visualization

According to Heer[23], the main benefits of visualizing data includes:

1. Faster information processing
   Parkinson[24] shared that images are processed 60,000 times quicker than textual information. This claim is true as the human eye contains 70% of the sensory receptors[25] and MIT researchers discovered that only 13 milliseconds are required for the human brain to process pictorial information[26]. Therefore, visualizing data can help analysts to achieve a faster response to market developments and new opportunities.

2. Pattern discovery
   With the potential of an information overload, Ellstrøm[27] pointed out that visualization tools can facilitate the manipulation of data to discover underlying trends that the naked eye would have missed out. Thus, business stakeholders can identify new opportunities in the market.

3. Effective means of conveying data
   Along with the advantages from the previous points, data visualization can help analysts in story-telling for highlighting trends to the appropriate stakeholders in timely-fashion to ensure that everyone on the team is on the same page.

2.4.2 Visualization Techniques

Given the myriad of visualization tools today, it is imperative that analysts choose the appropriate type of graphical representation to simplify the analysis process as much as possible. Wang and Tao[28] stated that there is a need to streamline graphs for reducing the complexity in handling huge datasets.

Angra and Gardner[29] shared that amateur graphing users tend to select graphs based on its aesthetics features, construct visualizations according to instinct and utilize only raw data in the charts. In contrast, experienced users will choose graphs depending on the experiment's hypothesis and use re-modelled data for visualizations[29]. The section below section will review the purpose of utilizing specific types of graphs.

Line chart

According to Slutsky[30], line charts are suitable for showing a continuous dataset with the X-axis being the independent variable and the Y-axis as the dependent variable. A typical example will be temperature level over time (see Fig. 2.3)

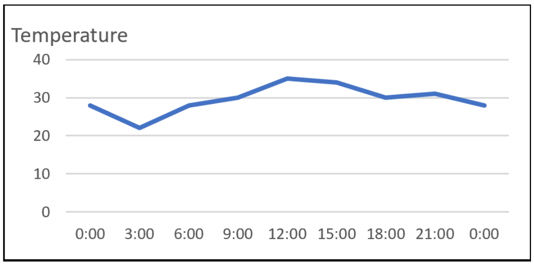


Figure 2.3: *Temperature level over Time*

Bar chart

There are two types of bar charts: Horizontal (see Fig. 2.4) and Vertical (See Fig. 2.5). The interpretation for both types of bar charts are the same – Longer bar corresponds to bigger values. Slutsky[30] stated that bar charts can be used for comparison of distinct objects or observation of objects over time.

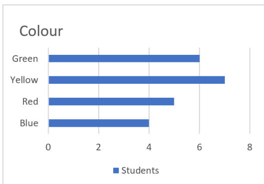


Figure 2.4: *Poll of students' favourite colour*

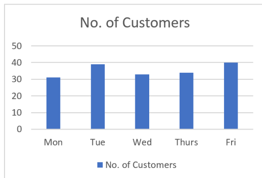


Figure 2.5: *Frequency of customers on weekdays*

Pie chart

A pie chart (see Fig. 2.6) shows the entire composition of a dataset with each segment representing a different category. It was recommended by Slutsky[30] that the number of segments in a pie chart should be within the range of three to ten categories.

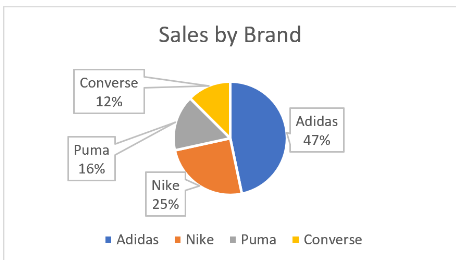


Figure 2.6: *Sales performance of brands*

Heatmap
Utilizing rows and columns to build a tiling cluster structure, a heat map (See Fig. 2.7) illustrates the correlation between two variables along with its significance which is indicated by its colour scale. Cerdas et. al[31] suggested a “colour scale of three hues” for a clearer distinction between the extreme ends of the data.

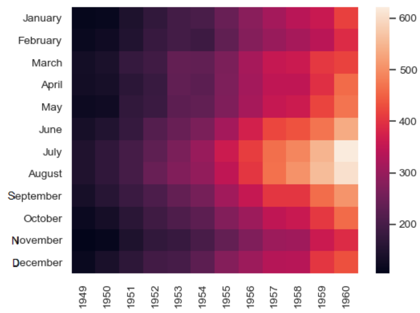


Figure 2.7: *Amount of Passengers from 1949-1960[32]*

2.5 Analytics Archetypes

As stated in[33], knowledge and insights are derived from big data analytics which comprises of several methods. In this section, we will look into the different methods that are applicable to the eCommerce landscape as mentioned in[34]:

2.5.1 Descriptive analytics

Descriptive analytics helps to summarize data in a meaningful way, providing a simpler interpretation of data, describing past and current events within the dataset and possibly establish relationships among entities. (e.g. Monetary Authority of Singapore Annual Report 2017/2018)

2.5.2 Exploratory analytics

This method aims to discover latent relationships among data points using correlation, graphical representations and summary statistics for pattern discovery and anomalies detection. (e.g. creation time of posts that received the most upvotes on reddit)

2.6 Existing Embedded Analytics Tools

With the increasing popularity of Internet of Things (IoT) and Big Data, there is a growing demand for embedded analytics to handle the massive quantity of data created daily by organisations[35]. This section will browse through existing embedded analytics tools available on the market.

2.6.1 IBM Watson Analytics

IBM Watson Analytics provides state-of-the-art analytics for story-telling without the complexity through its smart data discovery service, guided data exploration, automated predictive analytics and easy dashboard widgets creation[36].

Figure 2.8: Sample of IBM Watson Analytics[37]

2.6.2 Oracle Business Intelligence (B.I.)

Oracle Business Intelligence provides businesses with an extensive view of their progress by standardizing analytics policies (data model and metrics) on a centralized platform[38].

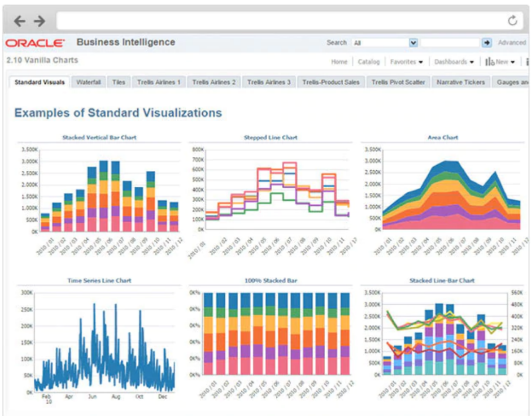


Figure 2.9: Sample of Oracle B.I. dashboard[38]

2.6.3 Tableau

Tableau is another business intelligence and visualization tool that provides quick insights with its flexibility to function within the users' enterprise architecture and data ecosystem[39].

Despite the attractive display of their state-of-the-art visualizations, the aforementioned embedded analytics tools might be overwhelming for users due to excessive features on the dashboard interface which purpose is to simplify analysis. Also, these analytics tools can be very expensive to scale in organizations because they are proprietary software.

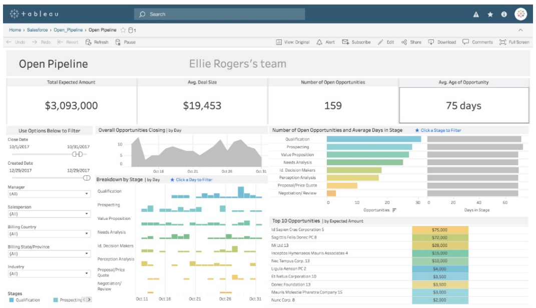


Figure 2.10: Sample of Tableau Dashboard[40]

2.7 Recommendation Systems

In today's eCommerce landscape, there exists various types of recommendation systems to help consumers navigate through the websites with lesser search costs and more product-fit[9]. In this section, we will examine the two main categories of recommendation systems (personalized and non- personalized) that are in practice within the technology industry.

2.7.1 Non-Personalized Recommendation System

As suggested by its name, a non-personalized recommendation system does not create customized recommendations for each individual. It takes the overall view of the eCommerce ecosystem and presents the information to everyone regardless of the customer's interactions with the system. The following are examples of non-personalized recommendation.

2.7.1.1 Top Purchases

Using the purchase statistics of all customers on the website, Top Purchases displays the best-selling and trending products across the board (Fig 2.11) or by product category (Fig 2.12) to all customers.

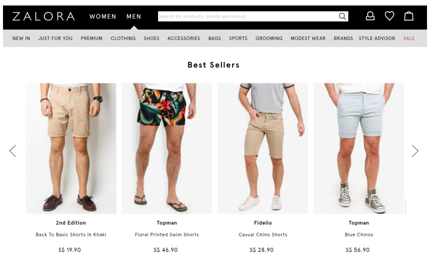


Figure 2.11: Zalora Best Sellers[41]

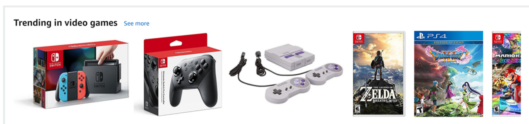


Figure 2.12: Amazon Category Trending (Video Games)[42]

2.7.1.2 Featured Items

In this version of non-personalized recommender, merchants (Fig 2.13) and products (Fig 2.14) can be featured regardless of their page views and/or product purchase frequency.

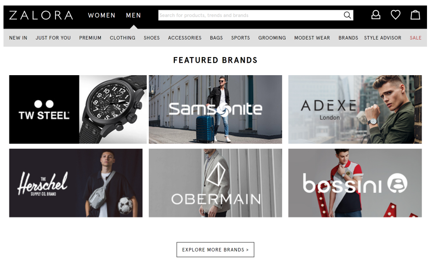


Figure 2.13: Zalora Featured Brands[41]

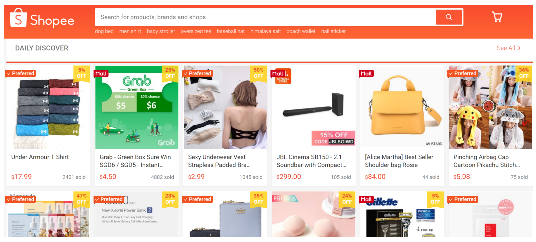


Figure 2.14: Shopee Daily Discover[42]

2.7.2 Personalized Recommendation System

Unlike the previous recommendation system, a personalized recommendation system takes the users' behaviour into account and generates recommendations based on their interaction with the website.

2.7.2.1 Content-Based Filtering

A content-based filtering recommender analyses a list of items that were previously rated by the user and create a profile based on the user's preference. It then uses the profile generated to recommend similar items more efficiently and accurately, effectively overcoming the 'cold start' problem of insufficient unique users that newly launched online stores encounter.

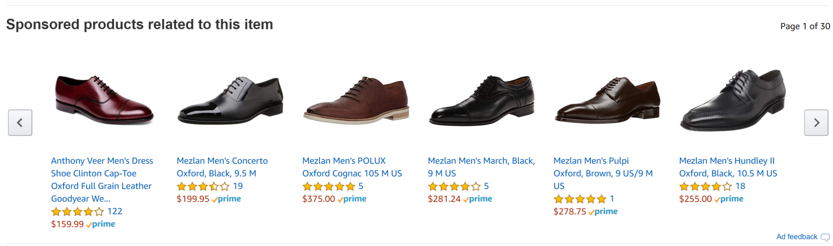


Figure 2.15: Amazon Related Products Carousel[42]

However, Alyari and Navimipour[43] reported that content-based filtering recommender systems face a drawback of limited diverse recommendations because the recommender only computes suggestions from the users' browsing history.

2.7.2.2 Collaborative Filtering

Contrary to the individual approach of content-based filtering, collaborative filtering uses the collective information within the community and make recommendations for a user derived from other individuals with similar purchase patterns and browsing history without the need for earlier insights about a product as mentioned by Shruthi and Gripsy[44]. van Meteren and van Someren[45] have reported that collaborative filtering can also provide a huge variety of recommendations to users because it is not dependent on only one source for operation. Below is an example of a recommendation using collaborative filtering:

Despite having the benefits of producing a diverse range of recommendations utilizing minimal prior knowledge of products, collaborative filtering also has its shortcoming. Alyari and

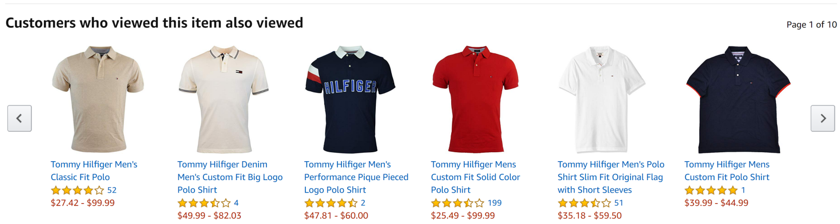


Figure 2.16: "Customer who viewed" Product Slider[42]

Navimipour[46] stated that this approach suffers from the 'cold start' problem because it requires a huge record of customers' interaction in order to capitalize on the behaviour similarity between users for product recommendation.

2.7.2.3 Hybrid Recommender

In other to overcome the limitations of Content-Based and Collaborative Filtering, Melville and Vikas Sindhwani[47] suggested that a hybrid approach of merging both methods can be adopted.

A famous literature for Hybrid Recommender is the Netflix Prize[48] where a team of researchers (BellKor's Pragmatic Chaos) were awarded USD $1 Million for improving its movie recommendation accuracy by 10% using a hybrid approach[49].

Other articles regarding hybrid recommenders hypothesize that it can provide satisfactory results even when faced with data sparsity[46][50] while Lagerstedt and Olsson[51] demonstrated that a hybrid approach can decrease the impact of 'cold start' issues by 30%.

2.8 Summary of Literature Review

From the findings of existing works, the project will focus on constructing a real-time dashboard with a minimalistic approach that provides descriptive and exploratory analysis through basic visualizations (e.g. heatmap, line, bar and pie chart).

Also, a non-personalized product recommender will be developed so that the client is still able to provide recommendations even with a small quantity of user records so that the client can overcome the 'cold-start' hurdle that new eCommerce companies will encounter.

3 Analysis/Requirements

According to Sommerville[52], requirements engineering allows the developer to lay out the requirements of the system by analysing the customer's specifications and evaluate its feasibility for development.

This chapter will outline the requirements gathering process from the project client (Trollee) by establishing the terminologies for eCommerce and identifying the actors, user stories, functional and non-functional requirements of the system for the system.

3.1 eCommerce terminologies

Before commencing on the requirements of the system itself, it is important to establish the terminology for this project's domain with the client. This section will explore the commonly used terms[53][54][55] in the eCommerce industry.

3.1.1 Customer

An individual that uses credit cards to make purchases from merchants on the eCommerce website.

3.1.2 Merchant

A retailer that offers products and/or services on the eCommerce website.

3.1.3 Traffic

The amount of visitors that a website or webpage receives.

3.1.4 Bounce Rate

The proportion of visitors that view only one page and exit the website without clicking on other items. The formula can be defined as:

$$\text{Bounce Rate} = \frac{\text{Single page visits}}{\text{Total number of visits to that page}}$$

3.1.5 Click

The event that occurs when a user access a hyperlink on the webpage.

3.1.6 Clickstream

The series of clicks that a user went through while visiting the eCommerce website.

3.1.7 Add-to-cart

The action taken by a user to add items on his virtual shopping list.

3.1.8 Cart Abandonment

The event that occurs when a user add items to cart but does not proceed to the payment page to complete the ordering process

3.1.9 Review

A feedback that a customer leaves on the product page after purchasing the item.

3.1.10 Conversion Rate

The percentage of users who have accomplished a task successfully over the total number of unique users that took part in the event. Examples of conversion activities include product sales, media subscription and new signups. The formula for conversion rate can be calculated as such:

$$\text{Conversion Rate} = \frac{\text{Total number of successful conversion events}}{\text{Unique visitors (for that event)}}$$

3.1.11 Impressions

An impression is an instance where an advertisement is being loaded on a webpage in the form of a traditional digital advertisement or recommended products[56].

3.2 Actors

Before going into functionalities of the system, developers must first determine the types of users (actors) that will use the system. The following actors were identified:

Actor	**Role**
Customer	Uses the system to view trending and recommended products
Merchant	Uses the system to view traffic and sales trends pertaining to his store
Administrator	Uses the system to view overall (eCommerce-wide) traffic and sales trends

Table 3.1: *System Actors*

3.3 User stories

As mentioned by Cohn[57], user stories are high-level artefacts that describes features required by stakeholders of the system. The suggested user story format of "As a <Role>, I want to <Objective>, So that <Reason>" in[57] will be adopted for this project development.

No.	As a	I want to	So that
1	Customer	View featured products	I can see the deals of the day
2	Customer	View top products by category	I can see what is trending in each category
3	Customer	View recommended items	I can find similar items with minimal effort
4	Merchant	View my store's incoming visitor traffic	I can see how many customers are viewing my store
5	Merchant	View my store's sales clickstream	I can find out my customers drop out at which stage of the transaction process
6	Merchant	View trending products across the eCommerce website	I know what the trending items are
7	Administrator	View eCommerce-wide incoming traffic	I can see how many users are active
8	Administrator	Analyse the eCommerce traffic by weeks	I can discover the highest traffic frequency period
9	Administrator	View trending products across the eCommerce website	I know what the trending items are
10	Administrator	View eCommerce-wide sales clickstream	I can analyse customers drop out at which phase of the transaction process
11	Administrator	Know the devices users used to visit the website	I can improve the quality of the respective devices' user interface
12	Administrator	View the product recommender's impressions and purchase	I can measure the performance of the recommender
13	Administrator	View sales of new and returning customers	I can understand if the website is attracting back customers
14	Administrator	Know the demographics of my returning customers	I can target promotions to boost sales from specific demographics
15	Administrator	View trending search queries	I can understand what customers are searching for

Table 3.2: *User Stories*

3.4 Prioritization technique

Sommerville[52] mentioned that as with any system development, requirements must be ranked to resolve any conflicts regarding the importance of a certain requirement due to the involvement of different stakeholders. This project will use the MoSCoW prioritization technique[58] with the following priority levels:

- **Must Have (M)** – These are the minimal features that must be implemented for the system to be useful or labelled as a Minimum Viable Product (MVP).
- **Should Have (S)** – These features are important to the users but it is not vital to the product launch.
- **Could Have (C)** – These features are 'good-to-have' if there is extra time to further develop the system
- **Won't Have (W)** – These are the features not considered within the current project timeline but can be included in future developments

This method of prioritization will be applied to the subsequent two sections (Functional and Non-Functional Requirements) to determine the importance of certain requirements.

3.5 Functional Requirements

Functional Requirements specifies the behaviour of the system, stating what it should do when given different types of inputs as shared by Sommerville[52].

No.	Requirement	Description	Priority
1	View product carousel	User must be able to view products on the product carousel	M
2	Browse product carousel	User must be able to iterate through the product carousel	M
3	Change product carousel	User must be able to select different categories on the product carousel	S
4	View live traffic	User must be able to view live incoming traffic	M
5	Change graph plots	User must be able to select specific timeframes and demographics for analysis	S
6	View tabulated data	User must be able to view trending items in tabulated form	S
7	View trending queries	User must be able to view trending search queries	C
8	Item-based recommender	The recommender must be able to recommend products despite the lack of users' purchase history pattern	M
9	Integrate recommender with Graphical User Interface (GUI)	The recommender should be deployed with a GUI for product demonstration	C
10	Integration with other modules	This system should integrate with the other parts of the eCommerce project that is developed by other students	C

Table 3.3: *Functional Requirements*

3.6 Non-Functional Requirements

Non-Functional Requirements are not features of a system but benchmarks to judge a system's overall operational quality.

No.	Requirement	Description	Priority
1	User-friendly interface	The dashboard should be intuitive to untrained users	S
2	Responsive screen size	The dashboard should fit any screen's resolution size	C
3	Attractive layout	The dashboard should not be too dull	C
4	Fast loading time	The dashboard should not take too long to load	S
5	Fast response time	The recommender should not take too long to respond to queries	S
6	Maintainable	Maintenance should be easily executed for this system	S
7	Scalable	This system should cater for future upgrades	S

Table 3.4: *Non-Functional Requirements*

3.7 Data Source

Two structured data sources were provided by the Project Supervisor:

- comScore 2004 Disaggregate Dataset[1]
- Brazilian E-Commerce Public Dataset by Olist[59]

This project will prioritize developing the recommender using the comScore 2004 Disaggregate Dataset.It has the necessary fields for developing an item-based recommender that recommends a product solely based on the item's attributes instead of the conventional item-based collaborative recommender which requires product purchase history patterns that will not be available in a newly launched eCommerce website.

The following is a sample of the comScore dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 257182 entries, 0 to 257181
Data columns (total 20 columns):
domain_id              257182 non-null object
domain_name            257182 non-null object
machine_id             257182 non-null object
site_session_id        257182 non-null object
event_date             257182 non-null object
event_time             257182 non-null object
prod_name              254200 non-null object
prod_category_id       257182 non-null object
prod_qty               257182 non-null object
prod_totprice          257182 non-null object
basket_tot             257182 non-null object
hoh_most_education     257182 non-null object
hoh_oldest_age         257182 non-null object
census_region          257182 non-null object
household_size         257182 non-null object
household_income       257182 non-null object
children               257182 non-null object
racial_background      257182 non-null object
connection_speed       257182 non-null object
country_of_origin      257182 non-null object
dtypes: object(20)
```

Figure 3.1: *Fields of comScore 2004 Dataset*

[1] 2004 comScore Disaggregate Dataset is a record of 257,182 "web-wide visitation and transaction behaviour within 12 months based on a random sample of more than 1.5 million Internet users in the United States who have given comScore explicit permission to confidentially capture their Web-wide activity."

4 Design

This chapter discusses the system architecture along with the design considerations for the user dashboards and product recommender.

4.1 System Architecture

The Model-View-Controller (MVC) framework segregates the presentation and interaction of an application from the system's data source, effectively splitting them into three main components:

Component	Purpose
Model	Manages data and prepares it based on instructions coming in from the Controller
View	Renders the presentation of the data to the user based on his interactions with the system
Controller	Comprehends the user's action and send it as a command to the Model and updates the View

Sommerville[52] shared that besides the ability to achieve a loosely-coupled system, the MVC pattern is also used when there are various means to visualise and interact with the data.

In this system, the main components are user dashboard, business logic layer and the recommender algorithm. There might be a need to re-calibrate the business logic and recommender to suit the eCommerce because business requirements will change after the company has matured. Thus, this project will adopt the MVC framework as it caters for future developments through the concept of "separation of concerns". Below is the MVC diagram of the overall system architecture:

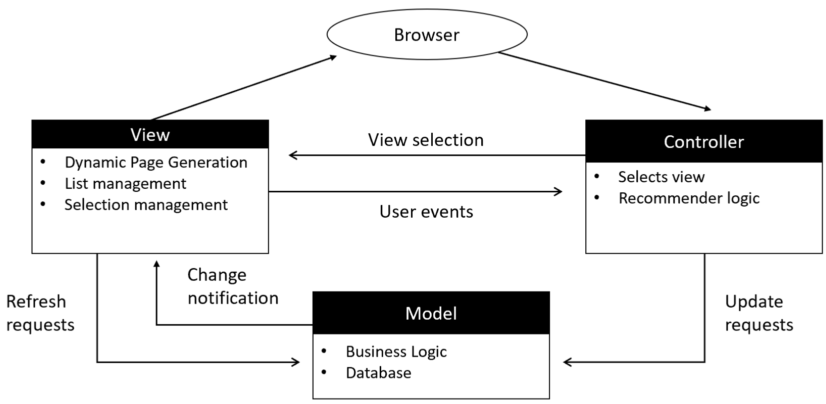


Figure 4.1: *MVC Diagram*

4.2 Dashboard Wireframes

A wireframe is a basic mock-up of screens without any functionalities, serving as a blueprint for a software development project. Babich[60] shared that wireframes should be produced before the start of in-depth user interface design so that any major changes to the visualization can be efficiently processed. This section will examine the entire wireframing process from its initial to final design.

4.2.1 Initial Design

The following table was the initial plan for the respective system actors[label]:

Actor	Page(s) to develop
Customer	Single page for trending deals
Merchant	Dashboard Activity
Administrator	Store/Product Inventory/Sales

4.2.1.1 Design Decisions

- **Customers View**
  The derived layout was derived from earlier functional requirements and the issues seen in existing eCommerce websites.

  For the daily deals section, Shopee's design was too overwhelming as the user would have to scroll through numerous products as seen below:

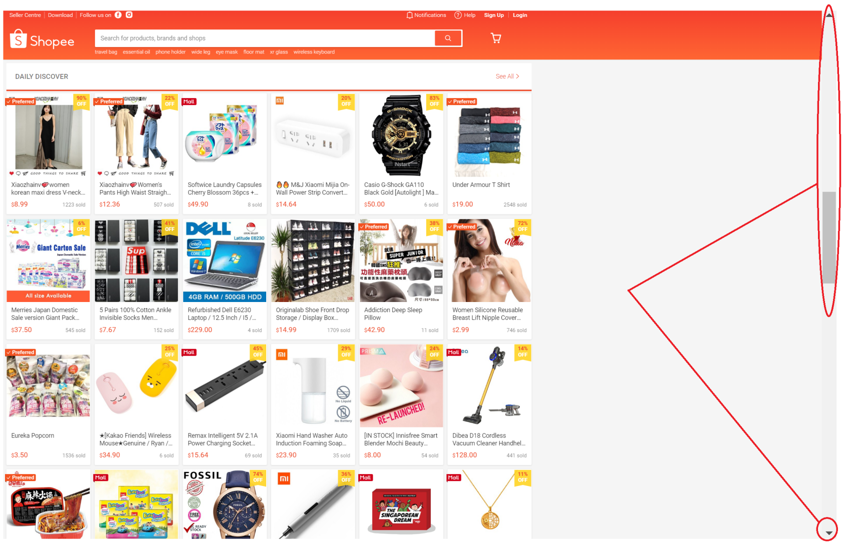


Figure 4.2: Shopee Daily Discover: Long browsing screen[34]

  Regarding the trending products section, users also have to scroll through the landing page of noon.com[61] and amazon.com[43] because they have different product carousels for each category individually.

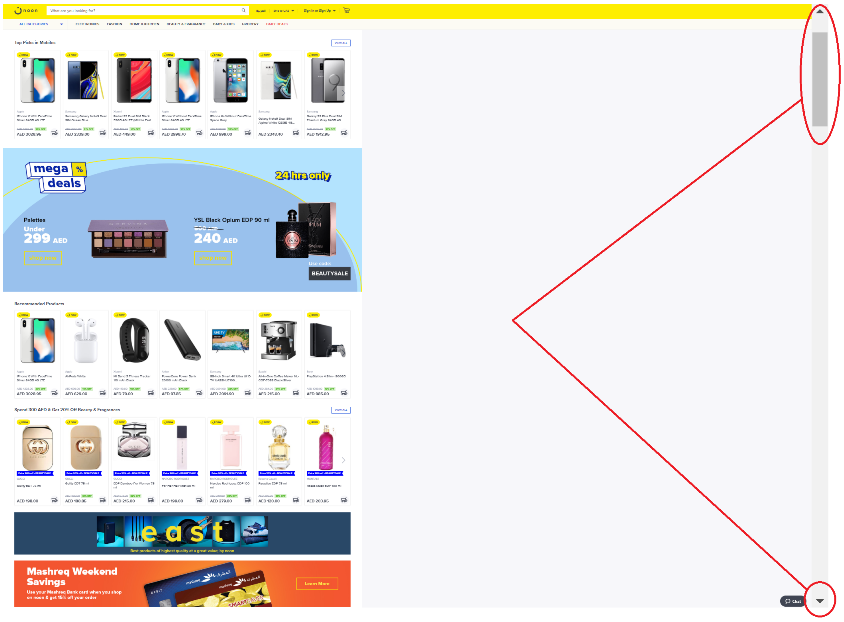


Figure 4.3: *Noon.com: Individual Category's Carousel[61]*

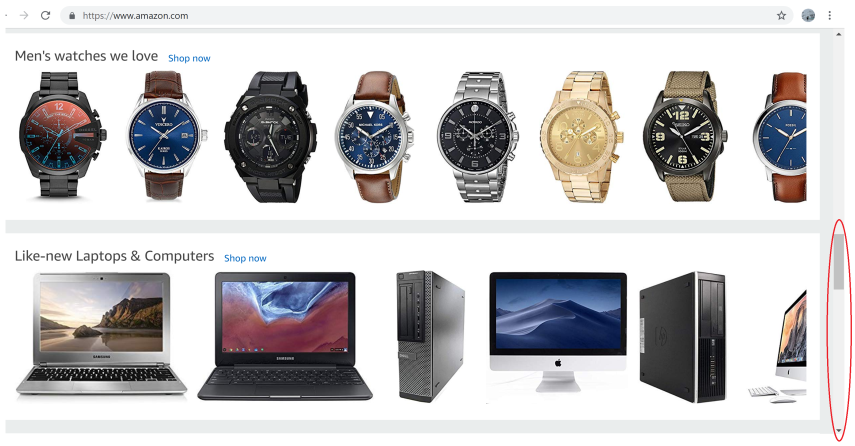


Figure 4.4: *Amazon: Individual Category's Carousel[43]*

Therefore, the consideration here was to keep scrolling to a minimum so that the page will not be overwhelming for the customers' eyes.

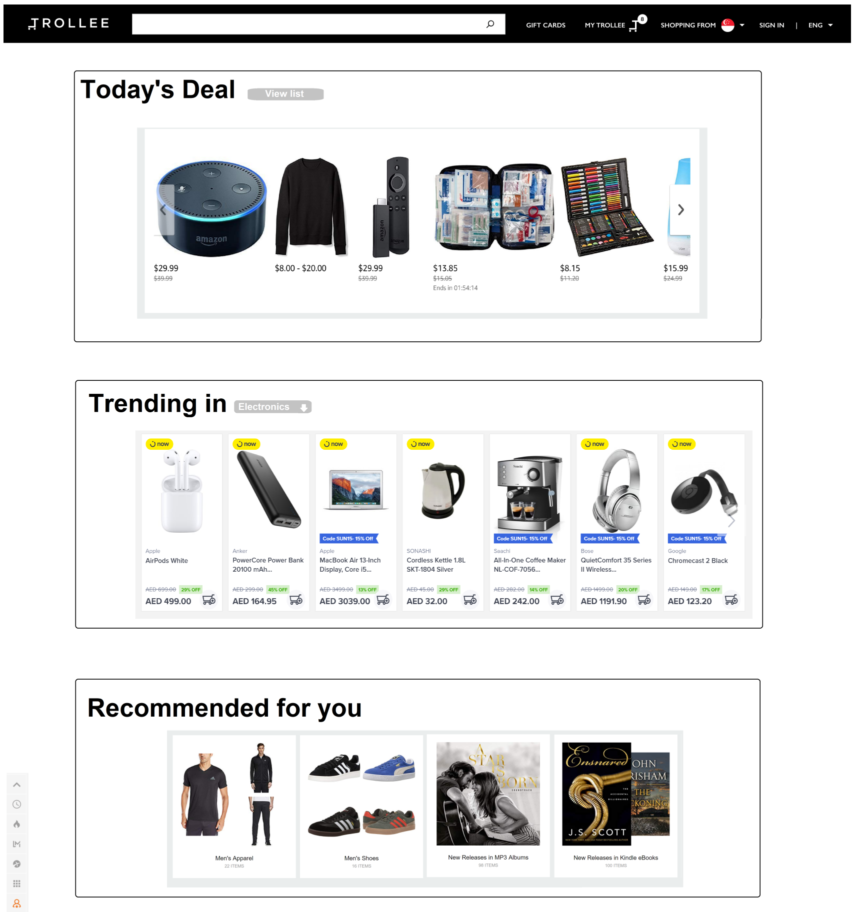


Figure 4.5: *Customer View: Initial Wireframe*

- **Merchants View**
  With reference to the functional requirements and the problems discovered in [existing dashboards], a minimalist approach will be adopted to reduce information overload. The Merchants View has a landing page (Dashboard) and three other pages (Activity, Store/Product and Inventory/Sales).

  The Dashboard page presents the overall performance of a merchant's store with summary statistics. The Activity page shows customers' interactions with the store while Store/Product page shows the store's products clickstream information and Inventory/Sales displays the store's sales performance, sales demographics, inventory management and campaigns tracking.

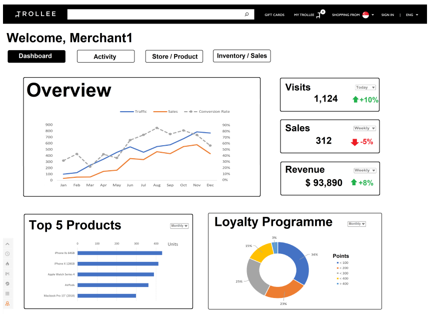


Figure 4.6: Initial Wireframe (Dashboard)

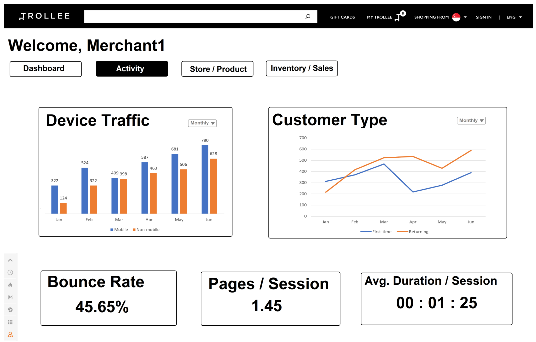


Figure 4.7: Initial Wireframe (Activity)

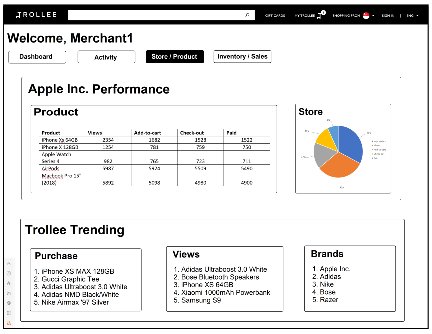


Figure 4.8: Initial Wireframe (Store/Product)

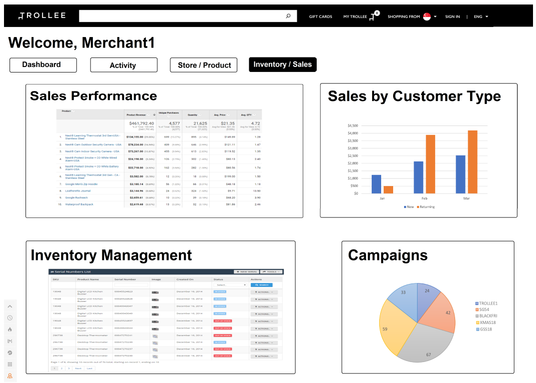


Figure 4.9: Initial Wireframe (Inventory/Sales)

- **Administrators View**
  Like the Merchants View, the Admin View have a similar structure however its contents will be on a broader scale since it displays overall information pertaining to the entire eCommerce. The following is a list of notable content differences in contrast to the Merchant View:

Page	**Difference**
Dashboard Page	Donut chart displays Payment Methods
Admin Page	Two additional line charts tracking active/inactive merchants and customers
Store/Product Page	Tracking of latest product addition to the eCommerce instead of tracking individual store's products
Inventory/Sales Page	Displays recommender's performance in place of inventory management

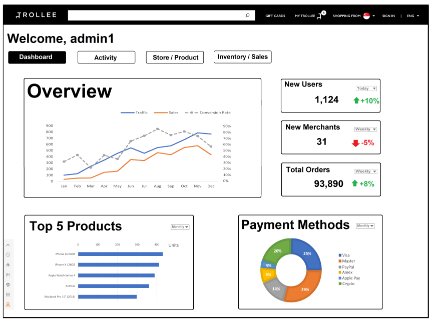


Figure 4.10: Initial Wireframe (Dashboard)

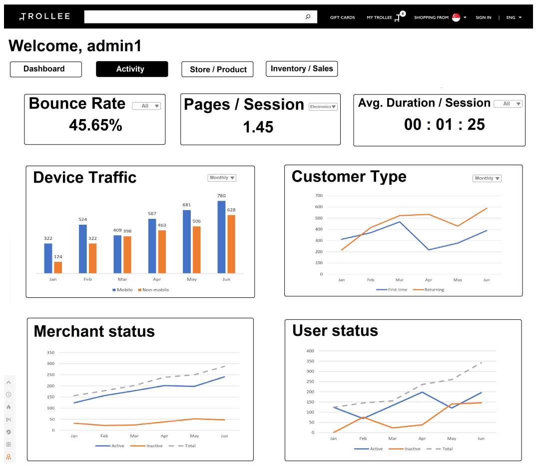


Figure 4.11: Initial Wireframe (Activity)

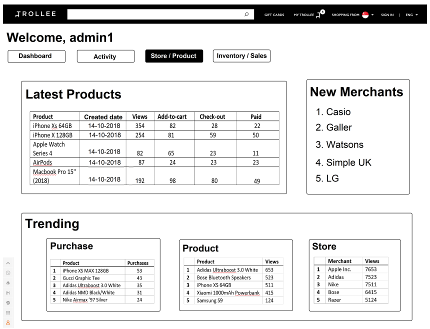


Figure 4.12: Initial Wireframe (Store/Product)

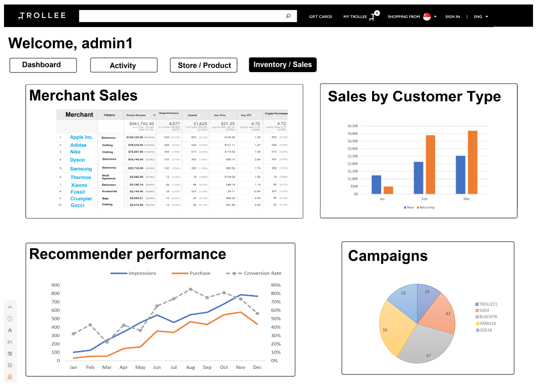


Figure 4.13: Initial Wireframe (Inventory/Sales)

4.2.2 Initial Wireframes Feedback

After presenting the designs to the clients and the project supervisor, the following feedbacks were gathered:

List of features to be removed

- Loyalty Program
- Top 5 Products
- Campaign Tracker
- Payment Methods

List of features to be modified

- "Customer Type" graph to be changed to "New Sign Up"
  - Visualization method to use Bar Graph

- "Existing Customers" graph to cater for the following trends:
  - Age
  - Gender
  - Frequency
  - Visits Vs Conversion

In summary, there was too much redundant information in the wireframes and this can potentially cause information overload that will defeat the purpose of this project. As suggested by the project supervisor, each view should be streamlined into a single page with the necessary visualizations to promote ease of navigation throughout the dashboard.

4.2.3 Final Design

Taking the feedbacks into considerations, the Merchants View and Admin View was streamlined into one page each. No changes were required for the Customers View. Below are the finalized wireframes prior to implementation:

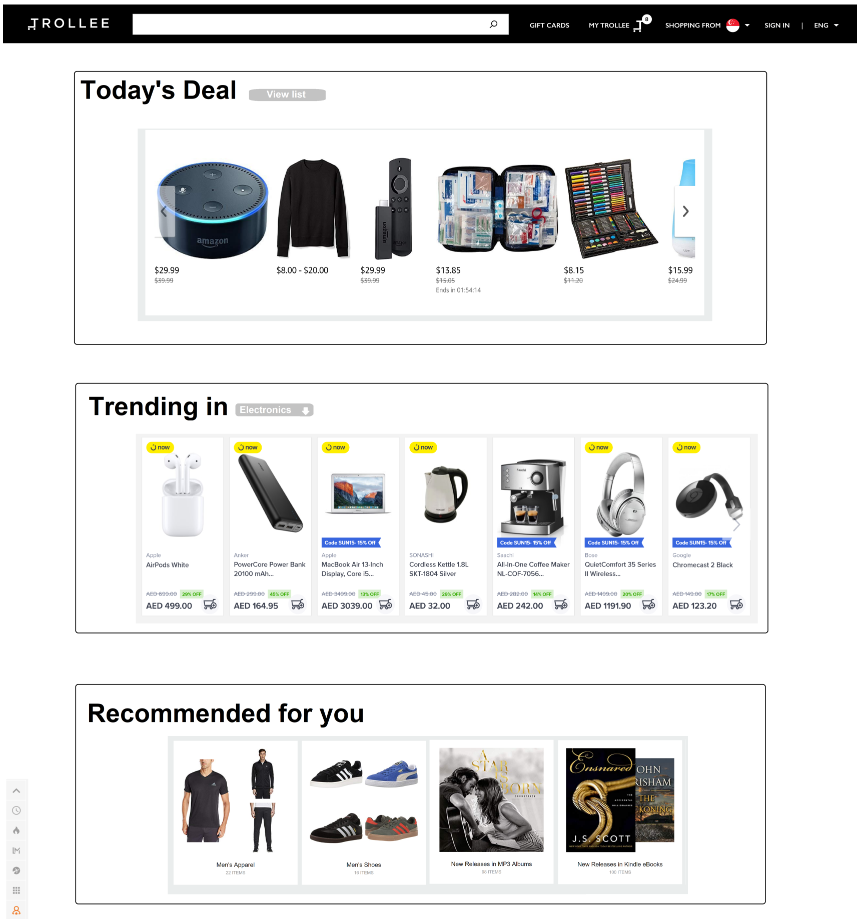


Figure 4.14: Customer View: Final Wireframe

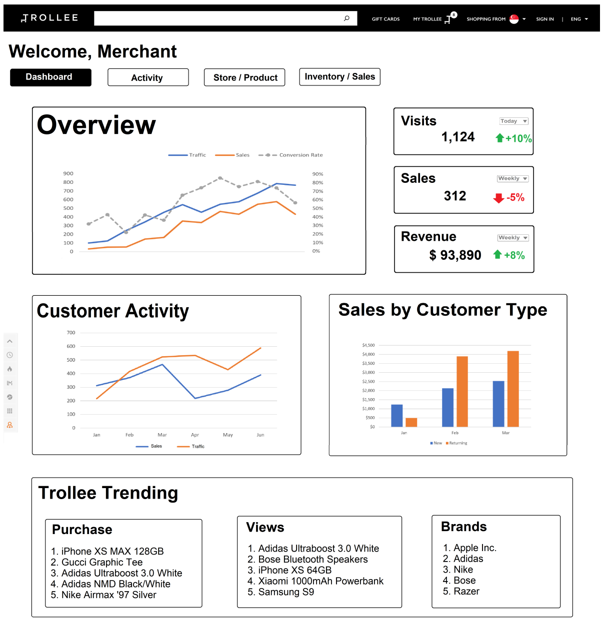


Figure 4.15: Merchant View: Final Wireframe

Figure 4.16: Administrator View: Final Wireframe

4.3 Recommender Design

As mentioned in the functional requirements, the recommender must be able to recommend products to the user regardless of the availability of users purchase history patterns. This requirement is of paramount importance due to the client being a newly set-up eCommerce – There will be insufficient user records to perform the conventional content-based and collaborative filtering.

4.3.1 Why Locality Sensitivity Hashing?

To overcome customer data sparsity faced by new eCommerce companies, Locality Sensitivity Hashing (LSH) can be introduced to minimize this 'cold-start' issue. LSH is a dimensionality reduction algorithm that uses Min Hash [1] to map nearby data points into similar buckets. LSH is also efficient as mentioned by Leskovec, Rajaraman and Ullman[63] that it can detect similarity without probing every pair in the dataset, achieving an O(N) time complexity (sub-linear performance) as reported by Datar et. al[64].

After interviewing the client about the eCommerce set-up process, it was established that merchants will be on-board before the website launches. Thus, we can the utilise the products' name from the on-boarded merchants and feed it to the recommender to make recommendations solely based on the name of an item that a user views (textual similarity) since LSH has the capability to cluster similar item efficiently.

Rajaraman[65] shared the following protocol (fig below) for constructing a LSH algorithm:

1. Splitting the text input into shingles (a set of characters with k length)
2. Use Min Hash to create similarity signature matrices
3. Execute the LSH phase by Hashing columns of the signature matrices and divide them into band partition

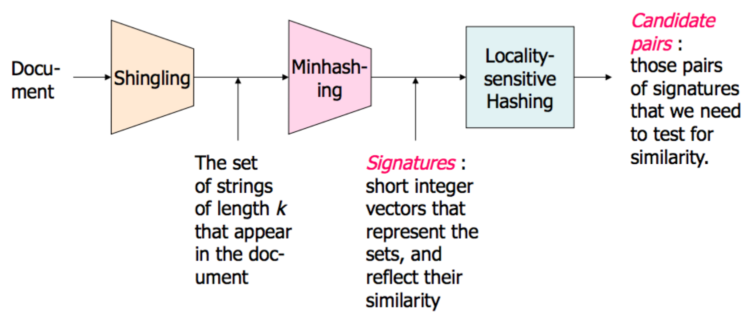


Figure 4.17: LSH Methodology[65]

[1] Min Hash is an algorithm that measures the similarity between two different sets quickly[62]

4.3.2 Design Methodology

With the information from the previous section, the recommender will be designed as an unpersonalized LSH-based product recommender so that the eCommerce will still be able to provide recommendations during its infancy stage while it collects purchase history patterns from its customers for collaborative filtering recommendation systems in the future.

The LSH-based recommendation methodology will be constructed using the Data Preparation Methodology in section.. and the procedure suggested by Rajaraman[65] in the earlier section:

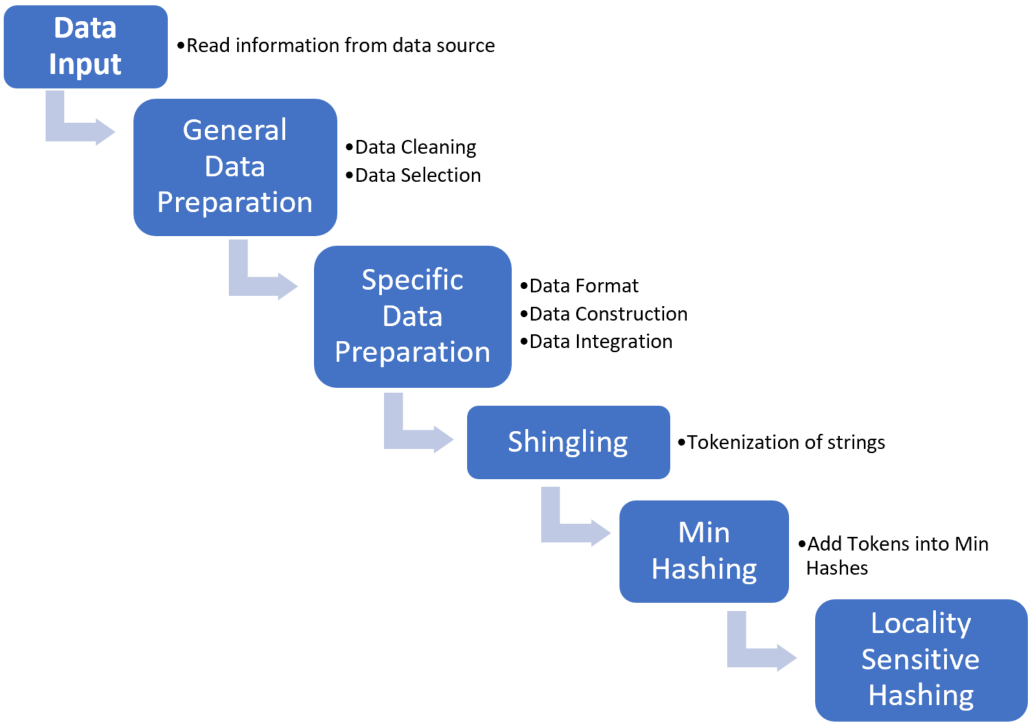


Figure 4.18: Recommender Methodology

5 Implementation

This chapter reviews the coding tools used to implement the design idea and outlines the coding process.
For a demo on how the product works, please visit: `https://youtu.be/lcvC6hqayhk`

5.1 Coding Tools

It is of utmost importance to choose the appropriate tools for development because it lays down the foundation of the system. This section examines the technologies used to develop the dashboard and recommender.

5.1.1 Dashboard Technologies

The dashboard Graphical User Interface (GUI) will adopt the HTML5, Bootstrap 3 and Javascript framework along with Bootstrap 3. The GUI will also incorporate jQuery for event handling purposes.

Visualization libraries were chosen based on the following criteria:

- Variety – Does the library have a huge variety of charts for free users?
- Code structure – Does the code have a neat structure?
- Intuitive Syntax – Is the code easy to understand?
- Documentation – Was the library well-documented?
- Community Support – Is there a large active community using the library?
- Watermarks – Is there any obvious watermarks on the charts?

A comparison was done among five graphing libraries with the aforementioned criteria and their line chart was used as a baseline to evaluate the code's structure:

Library	**Huge Variety**	**Neat Structure**	**Intuitive Syntax**	**Well-Documented**	**Community Support**	**Minimal Watermarks**
Google Charts[66]	✔	✔		✔		✔
CanvasJS[67]		✔	✔			
amCharts[68]			✔			
D3.js[69]	✔				✔	✔
Plotly[70]	✔	✔	✔	✔	✔	✔

Therefore, based on the comparison table above, Plotly was chosen as the main visualization library for the GUI and an exception was made for D3.js as it is required for the dynamic word cloud in the final product.

5.1.2 Recommender Technologies

Python was chosen as the programming language for constructing the product recommender because it is a dynamically-typed interpreted language which is efficient for data analysis. It also has an extensive ecosystem of free libraries for data science and it was the "Programming Language of the Year" for 2018 as reported in the TIOBE index[71].

The choice of text editor for Python was Jupyter as it is an Interactive Python Notebook (.ipynb) that lets users to run 'live' Python codes. This is very helpful when building a data model as users can run specific segments of their program instead of the entire script, allowing real-time debugging of the codes should the program encounter any error.

Alongside the programming language and text editor, several external libraries were selected for the construction of the product recommender. Numpy was chosen for scientific computation and array processing while Pandas was selected for data manipulation and analysis. The 'time' library was also included to measure the program's speed while Python's native 'Regular Expression' library(RE) and the Natural Language Toolkit(NLTK) was utilized to perform data cleaning, word tokenization and stop words[1] processing.

Datasketch was chosen as the library for building the LSH algorithm. Besides the basic MinHash LSH algorithm, Datasketch has the MinHash LSH Forest module which uses Top-K queries instead of threshold queries[73]. According to the findings by Bawa et al[74], LSH Forest surpassed the basic LSH algorithm's performance by 33%.

5.2 Dashboard Development

The first part of the GUI development was to structure the layout by specifying the HTML containers, rows and column widths. Next, visualization or tabulated data (shown in the wire-frames) were plotted using the respective external Javascript libraries mentioned in Dashboard Technologies. Javascript functions were also utilised to simulate 'Live' traffic (Merchant and Administrator view) and a dynamic word cloud (Administrator view). jQuery functions were also incorporated into the GUI to handle on-select events to display the chart of the selected option and hide the rest of the drop-down list.

The following sub-sections are visuals of the implemented product:

[1]Stop words are common words in a language that does not have any significant meaning[72]

5.2.1 Final Customer View

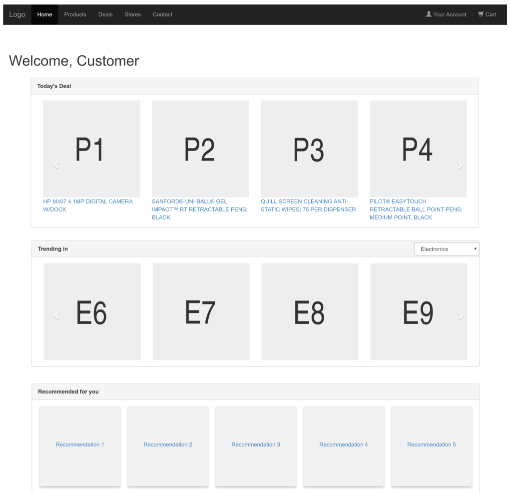
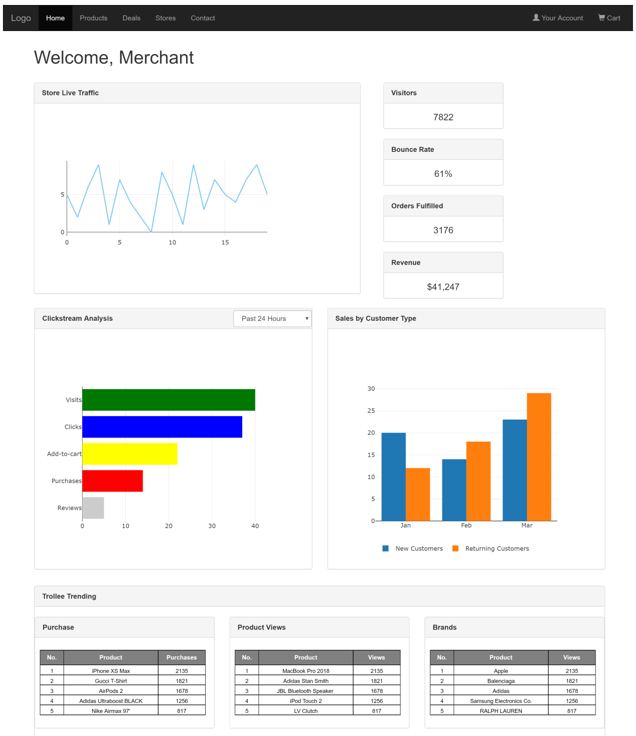


Figure 5.1: Implemented Customer View

5.2.2 Final Merchant View



No.	Product	Purchases
1	iPhone XS Max	2135
2	Gucci T-Shirt	1821
3	AirPods 2	1678
4	Adidas Ultraboost BLACK	1256
5	Nike Airmax 97'	817

No.	Product	Views
1	MacBook Pro 2018	2135
2	Adidas Stan Smith	1821
3	JBL Bluetooth Speaker	1678
4	iPod Touch 2	1256
5	LV Clutch	817

No.	Product	Views
1	Apple	2135
2	Balenciaga	1821
3	Adidas	1678
4	Samsung Electronics Co.	1256
5	RALPH LAUREN	817

Figure 5.2: Implemented Merchant View

5.2.3 Final Administrator View

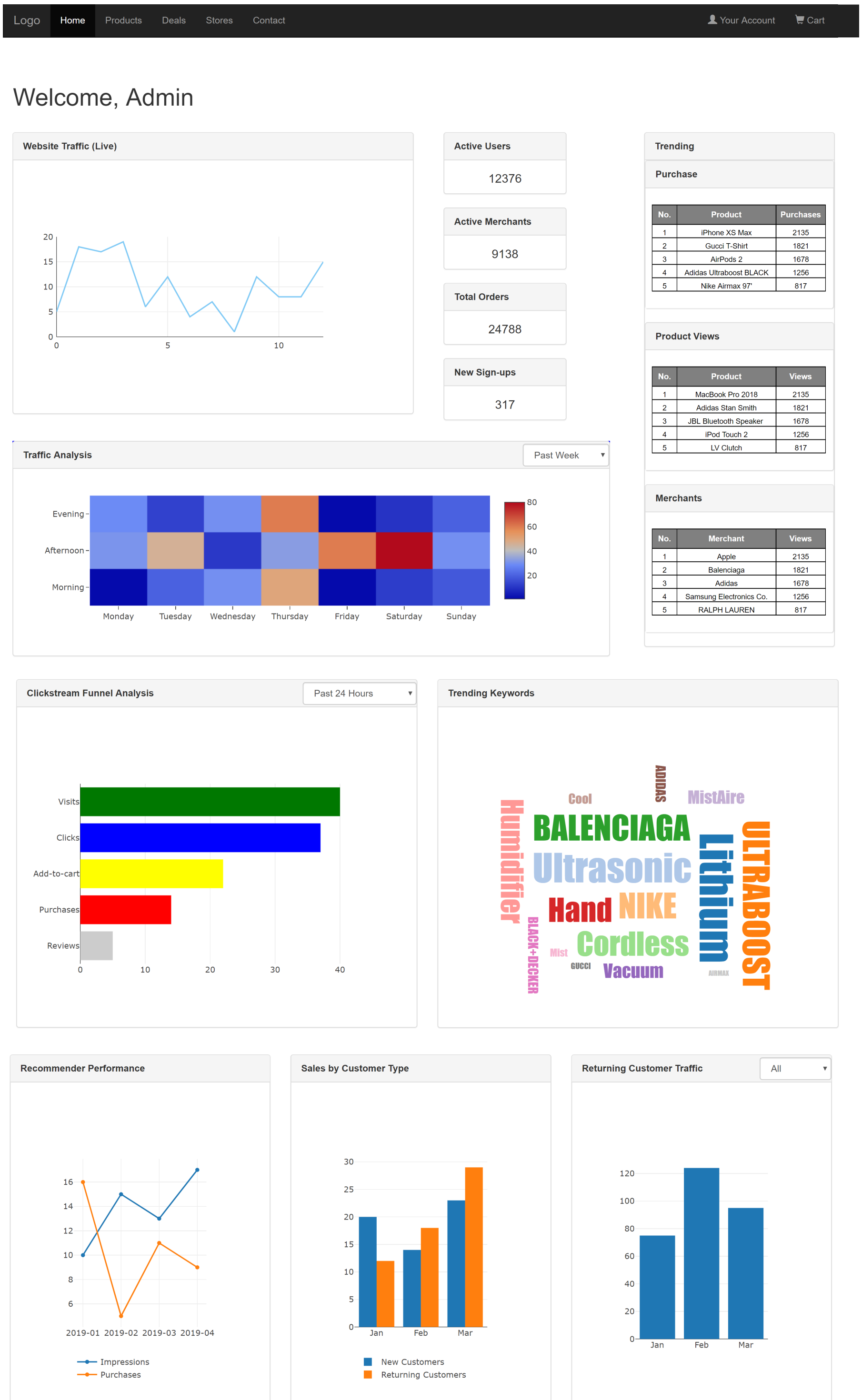


Figure 5.3: Implemented Administrator View

5.3 Recommender Development

Using the 4.18 methodology and technology tools, this section will describe the implementation process of the LSH-Based Product Recommender in detail.

5.3.1 Data Preparation

To start off the data preparation process, the dataset[data source] was surveyed to pick out the necessary columns for the recommendation script. It was also noted that several required fields such as product ID and product purchase count was unavailable however it can be derived from the identified columns. For clarity purposes, data frames used for data preparation were given a label of raw[number] while the final data frame for building the LSH Forest will be named 'data'.

```
raw = pd.read_csv("C:\\Users\\Sam\\Desktop\\TransactionsData04.csv" ,
    error_bad_lines=False, index_col=False, dtype='unicode')
```

Listing 5.1: Data Input

For the first phase of data preparation, a prod_id column was created by assigning a category code to each unique product name and the necessary columns were extracted out for processing.

```
raw2 = raw.assign(prod_id =(raw['prod_name']).astype('category').cat.codes)
raw3 = (raw2[['prod_id', 'prod_name', 'prod_category_id']]).apply(pd.to_numeric,
    errors='ignore')
```

Listing 5.2: Assignment of Category Codes

Next, invalid product ids were filtered out because they do not add any value to the analysis since their product_name fields were empty. They were assigned an id number of -1 or 0 by default.

```
raw4 = raw3.drop(raw3[raw3.prod_id <= 0].index)
```

Listing 5.3: Filtering Invalid Products

For raw5 to raw7, the purchase count of each unique product was calculated and merged back into raw4. After merging, duplicated entries of each product id was dropped from the new data frame (raw8). Meanwhile, another csv file containing the product category list was read into the program and mapped to raw8.

Lastly, the combined column of product_name and product_category (prod_name_cat) was assigned to the final data frame. A new column was duplicated from it and reassigned back to raw8 after removing all special characters for calculating the Jaccard Index.[2]

[2]Jaccard Index is an evaluation method that calculates the similarity of two sets by dividing their intersection over their union

5.3.2 Shingling

The next phase of the recommender development was to create shingles from the input data. Each entry will be checked for non-alphanumerical characters, switched into lower-case and tokenized.

An additional step here was performing stop words processing so that the LSH Forest will be able to cluster items more accurately without redundant information.

```
def preprocess(text):
   text = re.sub(r'[^A-Za-z0-9]+',' ',text)
   tokens = text.lower()
   tokens = tokens.split()
   filtered_words = [word for word in tokens if word not in stopENG]
   return filtered_words
```

Listing 5.4: *Shingling Process*

5.3.3 Min Hashing

After shingling has been completed, the tokens will be added to a MinHash object with a permutation setting of 128 and the MinHash object will be appended to an empty list.

```
minhash = []

   for text in data['text']:
      tokens = preprocess(text)
      m = MinHash(num_perm=perms)
      for s in tokens:
         m.update(s.encode('utf8'))
      minhash.append(m)
```

Listing 5.5: *Creation of MinHash Objects*

5.3.4 LSH Forest

```
forest = MinHashLSHForest(num_perm=perms)

   for i,m in enumerate(minhash):
      forest.add(i,m)

   forest.index()
```

Listing 5.6: *LSH Forest Construction*

5.4 Final Recommender View

The following is the final implementation of the LSH-Based Product Recommender.

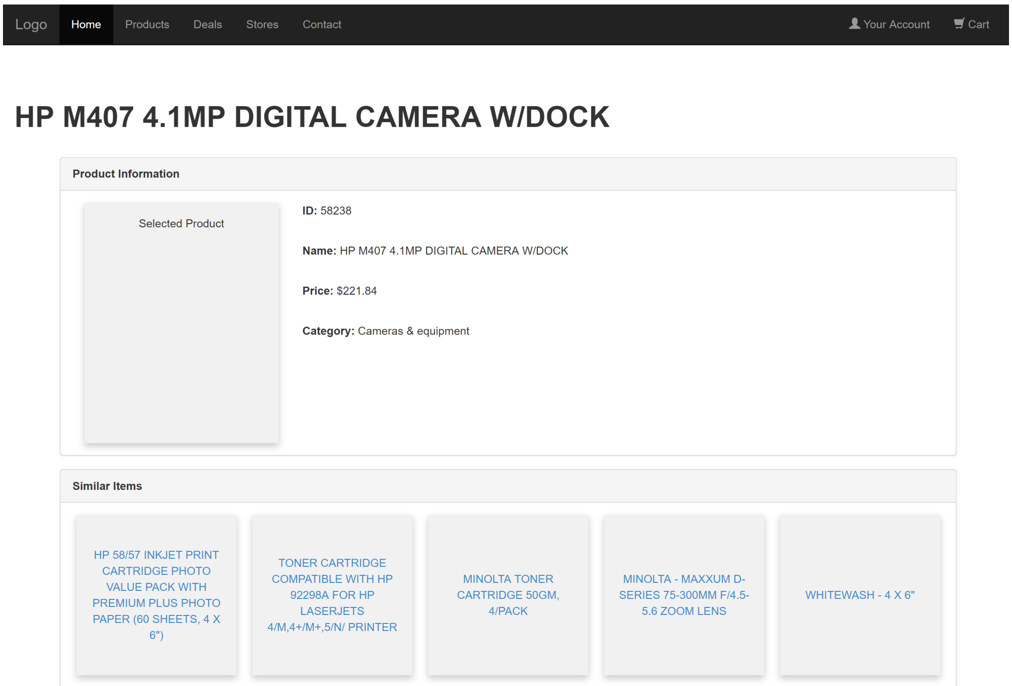


Figure 5.4: Implemented Recommender View

6 Evaluation

Towards the end of a project, it is always important to conduct evaluations on the final product. Patton[75] shared that evaluations helps designers to understand the product's effectiveness in serving its purpose. Two main evaluations were conducted for the Dashboard and LSH-Based Product Recommender.

6.1 Dashboard Evaluations

The dashboard evaluation was conducted with 18 participants, 4 of which were the clients of this project. They underwent a System Usability Scale survey followed by a Think-Aloud Evaluation.

6.1.1 System Usability Scale

The System Usability Scale(SUS) is a fast non-proprietary evaluation method that was created by John Brooke in 1996 for assessing the usability of any electronic interface[76]. It consists of ten questions with a rating scale of 1 to 5 (Strongly Disagree to Strongly Agree) that provides a high-level overview of the system.

The SUS survey was done on Google Forms for easy post-survey analytics. For a sample of the SUS survey and visualization of its scores, please refer Appendix.

6.1.1.1 SUS Results

The dashboard achieved a total score of 86.88% which was an excellent and acceptable rating according to the Adjective Grading Scale version of the SUS proposed by Bangor et. al[76].

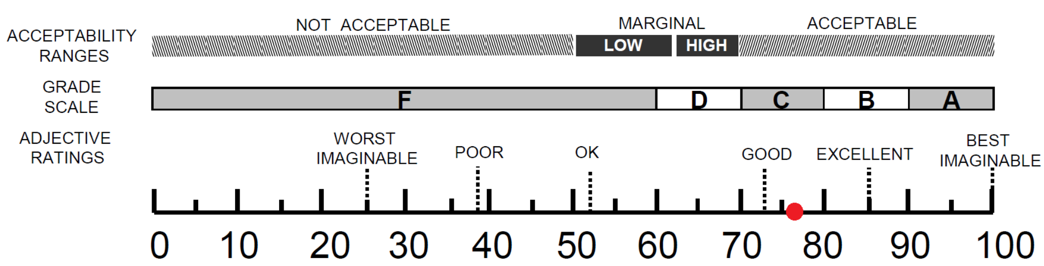


Figure 6.1: SUS Evalutation Result

The following is the breakdown of scores for the SUS survey:

No.	Question	Average Rating	Contribution
1	I think that I would like to use this website frequently	4.0625	3.0625
2	I found this website unnecessarily complex	1.6875	3.3125
3	I thought this website was easy to use	4.8125	3.8125
4	I think that I would need assistance to be able to use this website	1.8125	3.1875
5	I found the various functions in this website were well integrated	3.3125	2.3125
6	I thought there was too much inconsistency in this website	2.25	2.75
7	I would imagine that most people would learn to use this website very quickly	4.8125	3.8125
8	I found this website very cumbersome/awkward to use	2	3
9	I felt very confident using this website	4.4375	3.4375
10	I needed to learn a lot of things before I could get going with this website	2.9375	2.0625
Sub-Total		30.75	
Final Score		30.75 * 2.5 = 76.875%	

6.1.2 Think-Aloud Evaluation

Besides the SUS survey which gathered the overall impression of the system, this project also seeks to understand the users' thought while using the dashboard. For that, the Think-Aloud evaluation was also included in the dashboard evaluation as it allows users to speak aloud while performing a task and it is convincing because developers get to hear the first-hand honest opinion about their product from the users[77].

Below was the list of tasks for the Think-Aloud evaluation for each view:

View	Tasks
Customer	1) Navigate through application (inform evaluator what each component represents) 2) Browse products 3) Change product categories
Merchant	1) Navigate through Application 2) Filter clickstream by time
Administrator	1) Navigate through Application 2) Change heatmap 3) Filter clickstream by time 4) Filter returning customers by demographic

Feedback on the dashboard views was largely satisfactory. Majority of participants stated that the components within each view are intuitive and the word cloud is an interesting concept to be placed in a dashboard. However, one participant reported that the title of the clickstream analysis chart was too technical and mentioned that there should be a section for investigating bounce rate in detail.

Alongside the list of tasks, below are two additional qualitive questions for participants to elaborate their ideas on improving the dashboard:

- What would you like to add to this dashboard?
  - Page is easy to use, no need to further complicate
  - Try to incorporate real data when integrating with other modules of this project

- Is there any difficulty in using this dashboard?
  - Less technical terms for graph labels
  - More granular details to investigate statistics
  - No, it's easy to use

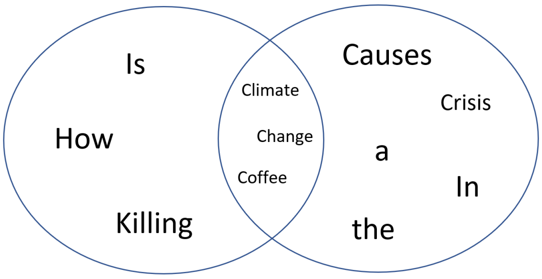
6.2 Recommender Evaluation

Two features of the LSH Forest were investigated for their impact on building and querying the LSH Forest for performance tuning purposes. The Jaccard Index was the metric chosen for evaluating the accuracy of the recommender's recommendations as it computes similarity between two sets in a straightforward manner[78]. The formula for the Jaccard Index[79] is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

The following Venn diagram illustrates how this evaluation method works, given two different sets of text (S1 and S2):

S1: "How Climate Change Is Killing Coffee"
S2: "Climate Change Causes a Crisis in the Coffee Market"



Words in common: 3
Total number of unique words: 11

Therefore, the Jaccard Index for the above example is (3/11) × 100% = 27% similarity

6.2.1 Evaluation Findings

In testing the permutations effect on build and query time, the number of permutations were incremented by 10 from the default value of 128. It was observed that an increase in number of permutations will result in a longer build and query time. (See Figure 6.2 and 6.3)

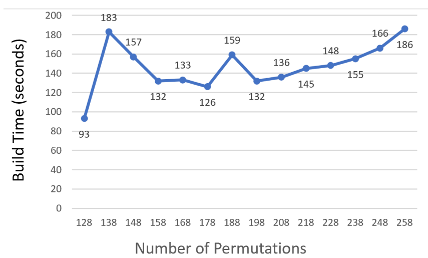


Figure 6.2: *Permutations VS Build Time*

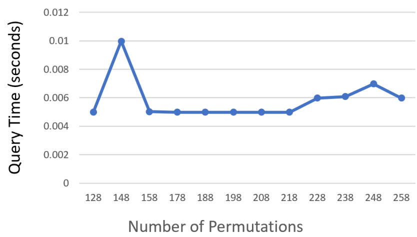


Figure 6.3: *Permutations VS Query Time*

For evaluation of the Jaccard Index, 128 was set as the number of permutations while the product "HP M407 4.1MP DIGITAL CAMERA W/DOCK" was chosen to the queried item and the amount of recommendations to be return was set to five because it is the minimum amount of recommendations that existing eCommerce provides.

Despite the notion of increased accuracy with increased permutations, it was discovered that increased permutations do not necessarily guarantee an increase in accuracy as shown in the figure below:

With regards to the correlation between the number of prefix trees and Jaccard Index, it was noticed that only one point (Prefix Tree = 20) beyond the default prefix tree quantity achieved a better score.

However, the build and query time of an LSH Forest with 20 prefix trees were significantly larger than that of the default amount of prefix tree as recorded in the following table:

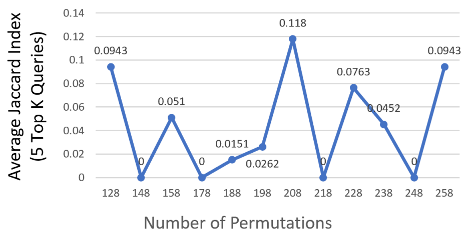


Figure 6.4: *Permutations VS Jaccard Index*

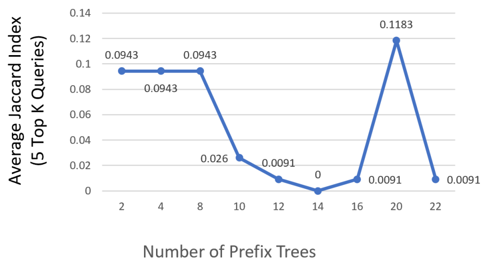


Figure 6.5: *Prefix VS Jaccard Index*

Prefix Trees	Build Time (seconds)	Query Time (seconds)
8	92.692	0.00499
20	109.125	0.00698

Therefore, the final specifications of the LSH-Based Product Recommender were set to 128 permutations and 8 prefix trees. Below is an output of the queried product:

```
Top Recommendation(s) for [HP M407 4.1MP DIGITAL CAMERA W/DOCK] is(are)
```

	prod_id	prod_name	prod_cat	jaccard_sim
0	124648	TONER CARTRIDGE COMPATIBLE WITH HP 92298A FOR HP LASERJETS 4/M,4+/M+,5/N/ PRINTER	Other computer supplies	0.100000
1	76908	MINOLTA TONER CARTRIDGE 50GM, 4/PACK	Other computer supplies	0.058824
2	131980	WHITEWASH - 4 X 6"	Photo printing services	0.062500
3	58104	HP 25-PACK 8X DVD+R (4.7GB) IN SPINDLE	Printers, monitors & peripherals	0.100000
4	76894	MINOLTA - MAXXUM D-SERIES 75-300MM F/4.5-5.6 ZOOM LENS	Cameras & equipment	0.150000

Figure 6.6: *Recommender Results*

7 Conclusion

7.1 Summary

This project strived to bring a different concept to the data analytics landscape by introducing unconventional methods to the dashboard and product recommender development. Through the evaluations of the Real-Time Dashboard and LSH-Based Product Recommender, I am contended with the implementation of the final product as it completed the high priority Functional and Non-Functional Requirements ("Must Have" and "Should Have") and achieved an excellent user rating. The key stakeholders (clients) were also very excited to deploy the final product on their fully integrated platform which was concurrently developed by other students in the eCommerce FYP Project.

7.2 Future Works

Due to time constraints, two "Could Have" functional requirements were not implemented:

- Integrate recommender with Graphical User Interface (GUI)
- Integration with other modules

The following functionalities were not within this project's scope however these could be explored if I were to further develop this project:

- Dataset adaptability so that the LSH-based product recommender can be applicable to other datasets
- Merging of the LSH-based product recommender with the conventional content-based and collaborative based filtering recommenders to create a hybrid recommender that can operate in any maturity state of the eCommerce
- Incorporation of deep learning models when the eCommerce matures as deep learning excels in processing large amount of data and natural language processing

7.3 Reflection

All the knowledge that I have attained in UGS for the past two years culminated in this Level 4 Individual Project as it allowed me to apply what I have learnt in the university's curriculum (e.g. Professional Software Development, Team Project 3, Interactive Systems, Human Computer Interaction and Web Science) as the baseline for the project. Developing a dashboard and a recommender might seemed easy as there were already numerous examples available however it can be an uphill task to convince others why they should use your product over other vendors. This can be done by showing value through the differentiating factor of your product. Therefore, the key learning points from this project were the theme of simplicity and novelty which I will carry forward into my career in the technology industry.

8 Appendices

This chapter contains the following:

- Product Video URL
- SUS Survey URL
- Data Source Headers
- Initial Wireframes
- Visualization of SUS Scores
- Clients' SUS forms

.1 Product Video

https://youtu.be/lcvC6hqayhk

.2 SUS Survey

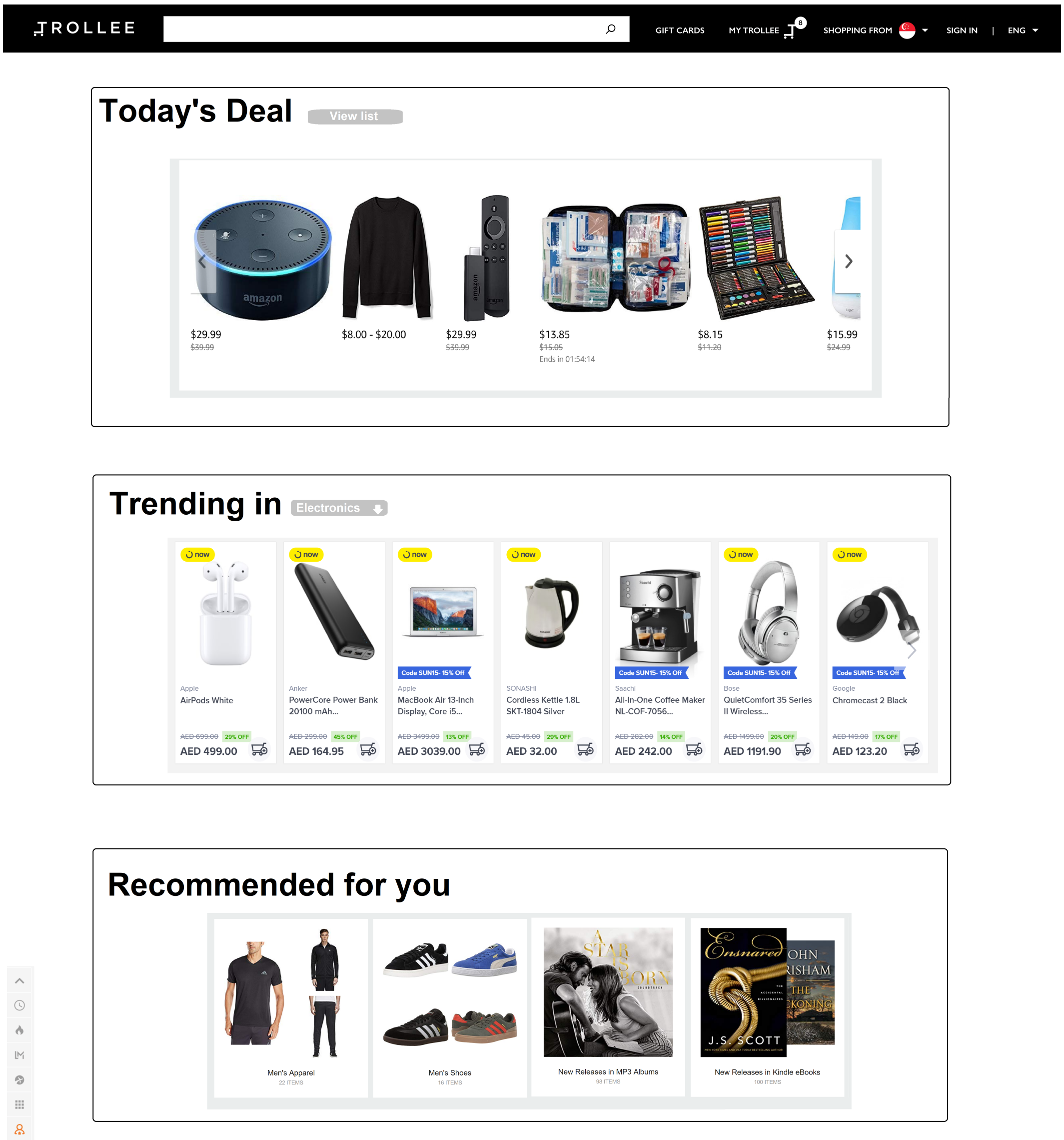
https://forms.gle/b5eBmq7sQ3P3642J8

.3 Data Source Headers

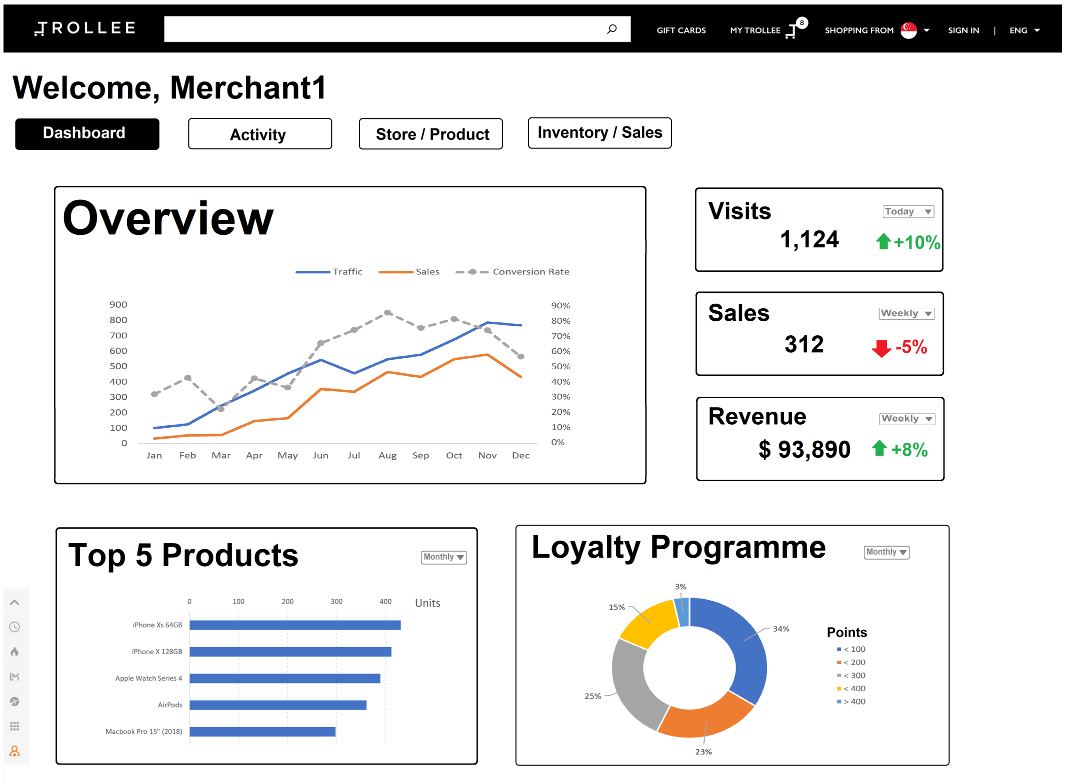
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 257182 entries, 0 to 257181
Data columns (total 20 columns):
domain_id              257182 non-null object
domain_name            257182 non-null object
machine_id             257182 non-null object
site_session_id        257182 non-null object
event_date             257182 non-null object
event_time             257182 non-null object
prod_name              254200 non-null object
prod_category_id       257182 non-null object
prod_qty               257182 non-null object
prod_totprice          257182 non-null object
basket_tot             257182 non-null object
hoh_most_education     257182 non-null object
hoh_oldest_age         257182 non-null object
census_region          257182 non-null object
household_size         257182 non-null object
household_income       257182 non-null object
children               257182 non-null object
racial_background      257182 non-null object
connection_speed       257182 non-null object
country_of_origin      257182 non-null object
dtypes: object(20)
```

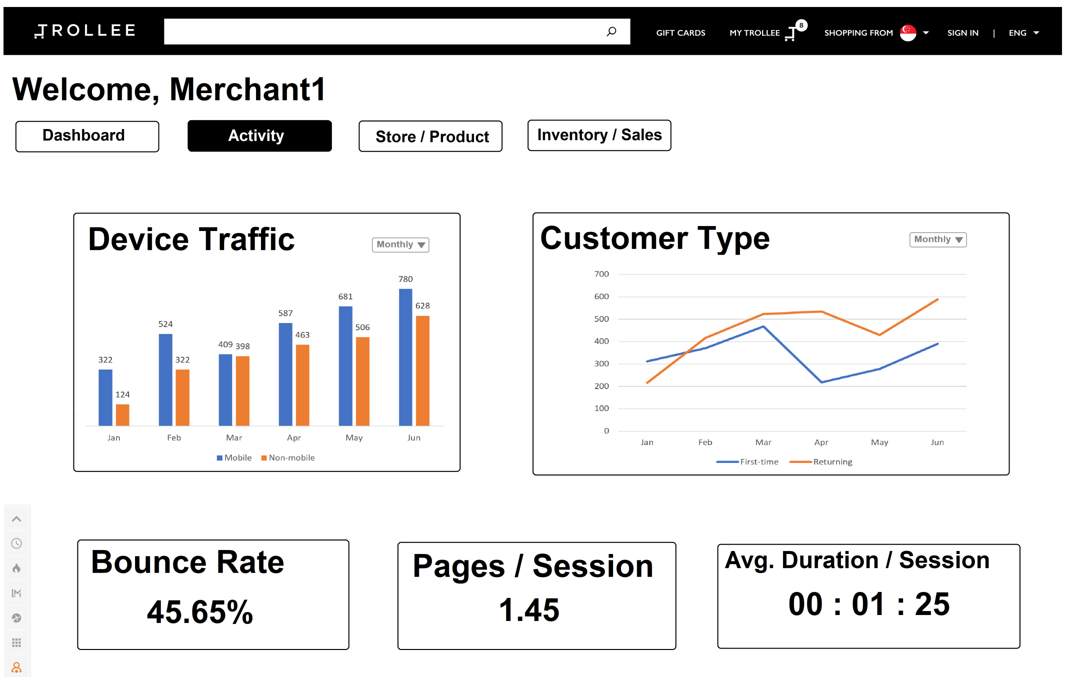
.4 Initial Wireframes

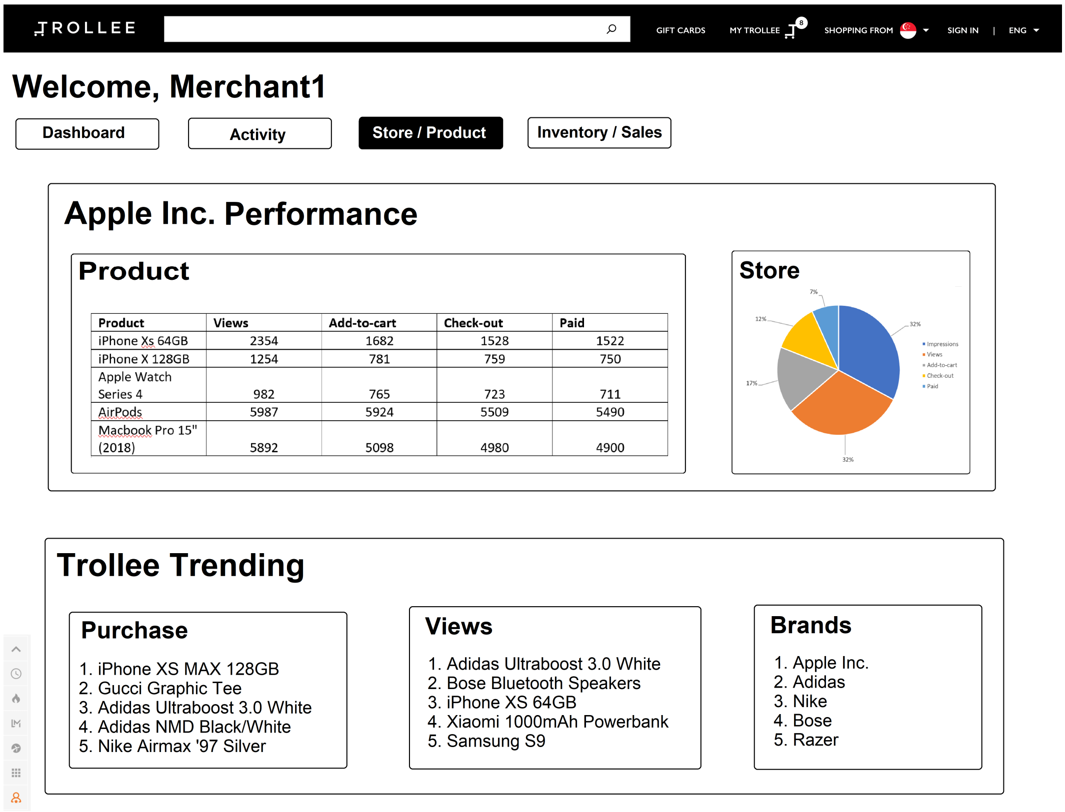
.4.1 Customer

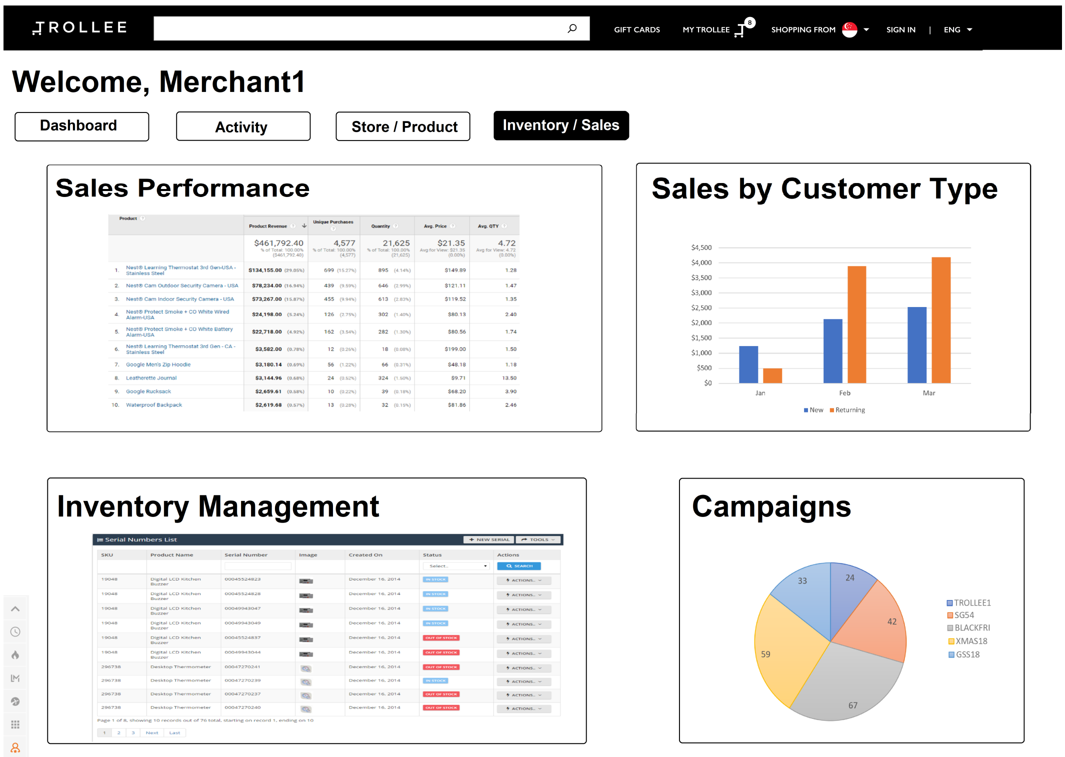


.4.2 Merchant

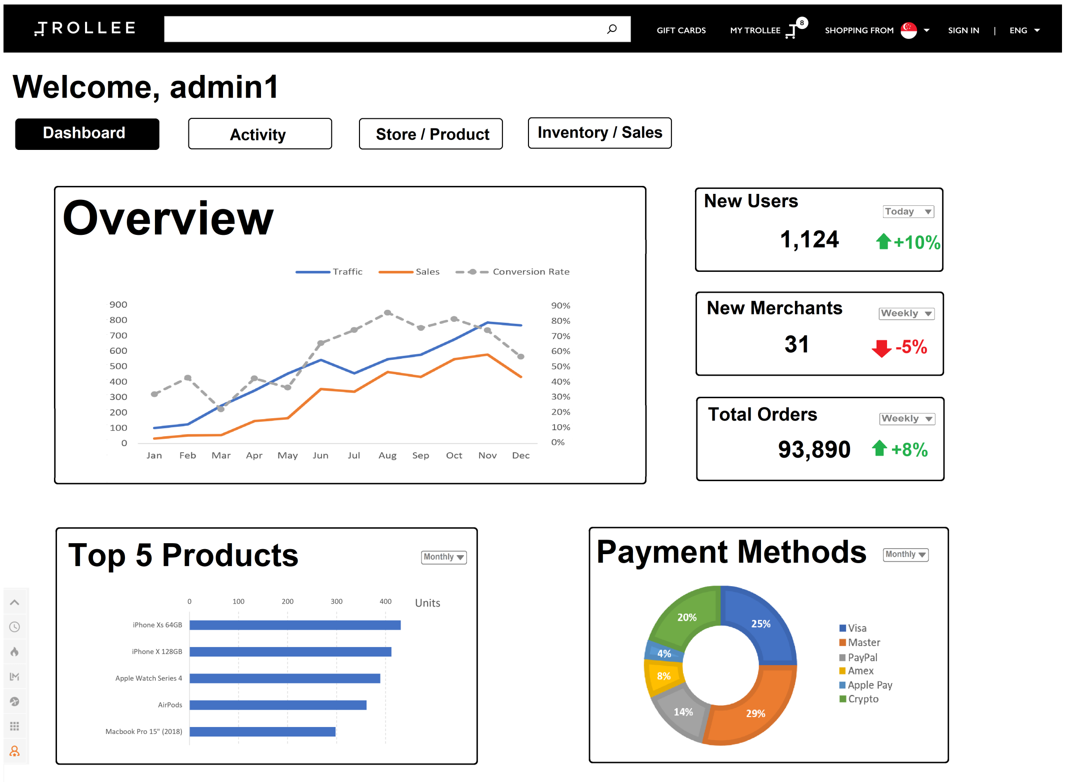

TROLLEE
GIFT CARDS
MY TROLLEE
SHOPPING FROM
SIGN IN
ENG
Welcome, Merchant1
Dashboard
Activity
Store / Product
Inventory / Sales
Overview
Traffic
Sales
Conversion Rate
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Visits
Today
1,124
+10%
Sales
Weekly
312
-5%
Revenue
Weekly
$ 93,890
+8%
Top 5 Products
Monthly
Units
iPhone Xs 64GB
iPhone X 128GB
Apple Watch Series 4
AirPods
Macbook Pro 15" (2018)
Loyalty Programme
Monthly
Points
< 100
< 200
< 300
< 400
> 400
3%
15%
34%
25%
23%

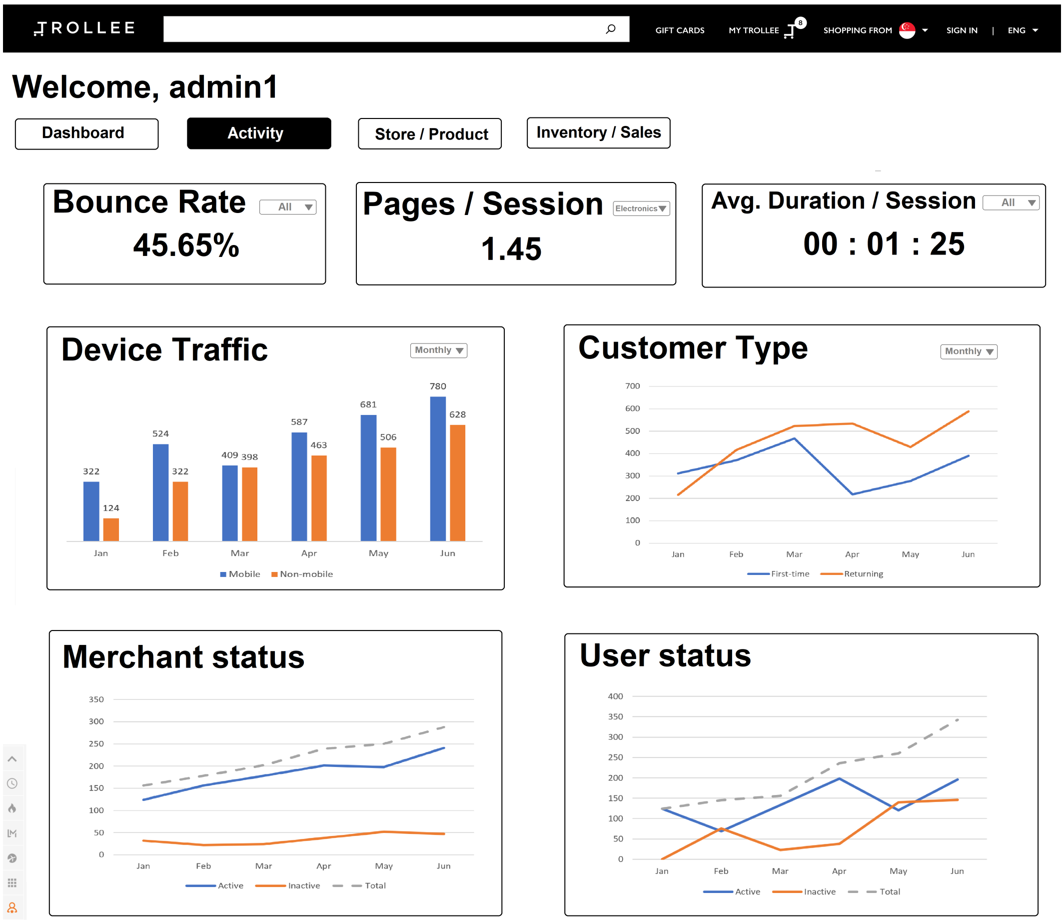

TROLLEE
GIFT CARDS
MY TROLLEE
SHOPPING FROM
SIGN IN
ENG
Welcome, Merchant1
Dashboard
Activity
Store / Product
Inventory / Sales
Device Traffic
Monthly
322 124
524 322
409 398
587 463
681 506
780 628
Jan Feb Mar Apr May Jun
Mobile
Non-mobile
Customer Type
Monthly
First-time
Returning
Bounce Rate
45.65%
Pages / Session
1.45
Avg. Duration / Session
00 : 01 : 25

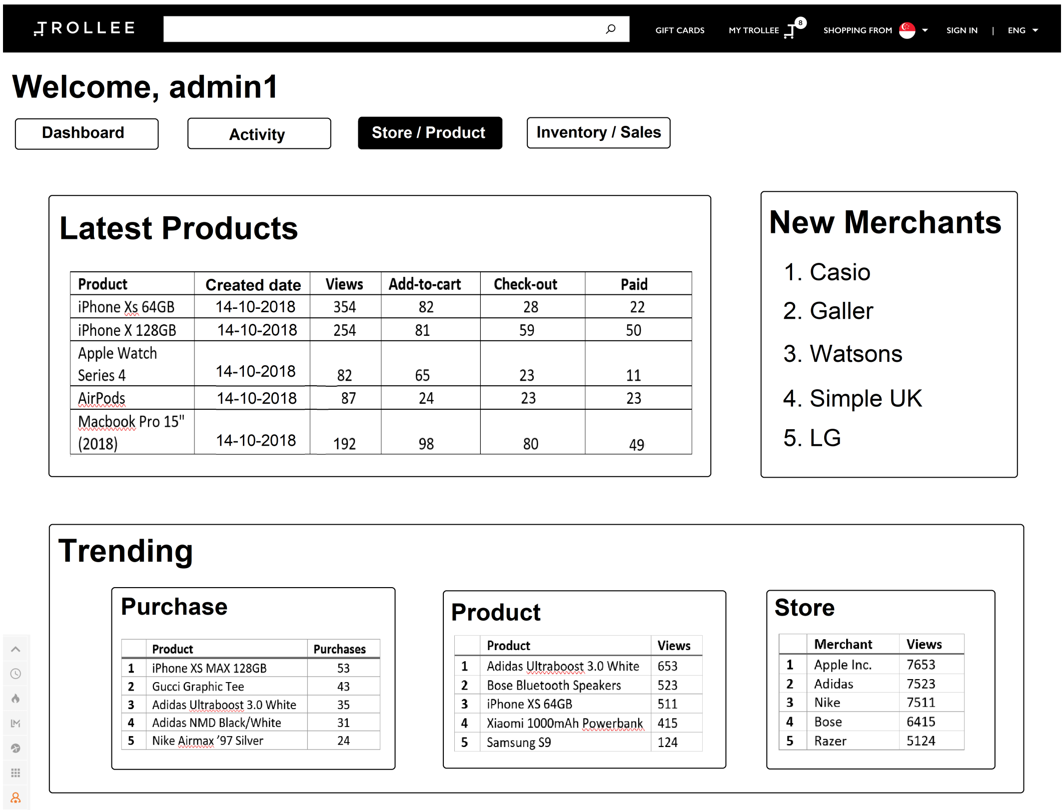

TROLLEE
GIFT CARDS
MY TROLLEE
SHOPPING FROM
SIGN IN
ENG
Welcome, Merchant1
Dashboard
Activity
Store / Product
Inventory / Sales
Apple Inc. Performance
Product
Product | Views | Add-to-cart | Check-out | Paid
iPhone Xs 64GB | 2354 | 1682 | 1528 | 1522
iPhone X 128GB | 1254 | 781 | 759 | 750
Apple Watch Series 4 | 982 | 765 | 723 | 711
AirPods | 5987 | 5924 | 5509 | 5490
Macbook Pro 15" (2018) | 5892 | 5098 | 4980 | 4900
Store
Trollee Trending
Purchase
1. iPhone XS MAX 128GB
2. Gucci Graphic Tee
3. Adidas Ultraboost 3.0 White
4. Adidas NMD Black/White
5. Nike Airmax '97 Silver
Views
1. Adidas Ultraboost 3.0 White
2. Bose Bluetooth Speakers
3. iPhone XS 64GB
4. Xiaomi 1000mAh Powerbank
5. Samsung S9
Brands
1. Apple Inc.
2. Adidas
3. Nike
4. Bose
5. Razer

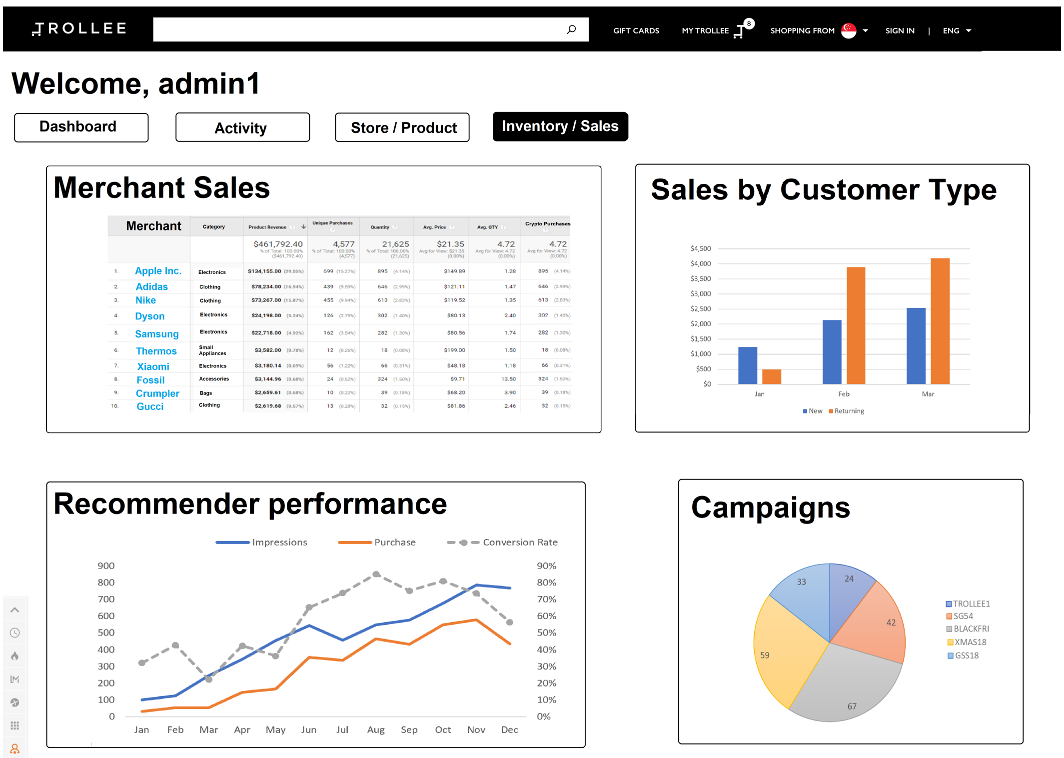

TROLLEE
GIFT CARDS
MY TROLLEE
SHOPPING FROM
SIGN IN
ENG
Welcome, Merchant1
Dashboard
Activity
Store / Product
Inventory / Sales
Sales Performance
Sales by Customer Type
Jan
Feb
Mar
New
Returning
Inventory Management
Campaigns
TROLLEE1
SGS4
BLACKFRI
XMAS18
GSS18

.4.3 Administrator




TROLLEE
GIFT CARDS
MY TROLLEE
SHOPPING FROM
SIGN IN
ENG
Welcome, admin1
Dashboard
Activity
Store / Product
Inventory / Sales
Bounce Rate
All
45.65%
Pages / Session
Electronics
1.45
Avg. Duration / Session
All
00 : 01 : 25
Device Traffic
Monthly
322 124 524 322 409 398 587 463 681 506 780 628
Jan Feb Mar Apr May Jun
Mobile
Non-mobile
Customer Type
Monthly
700 600 500 400 300 200 100 0
Jan Feb Mar Apr May Jun
First-time
Returning
Merchant status
350 300 250 200 150 100 50 0
Jan Feb Mar Apr May Jun
Active
Inactive
Total
User status
400 350 300 250 200 150 100 50 0
Jan Feb Mar Apr May Jun
Active
Inactive
Total


TROLLEE
GIFT CARDS
MY TROLLEE
SHOPPING FROM
SIGN IN
ENG
Welcome, admin1
Dashboard
Activity
Store / Product
Inventory / Sales
Latest Products
Product | Created date | Views | Add-to-cart | Check-out | Paid
iPhone Xs 64GB | 14-10-2018 | 354 | 82 | 28 | 22
iPhone X 128GB | 14-10-2018 | 254 | 81 | 59 | 50
Apple Watch Series 4 | 14-10-2018 | 82 | 65 | 23 | 11
AirPods | 14-10-2018 | 87 | 24 | 23 | 23
Macbook Pro 15" (2018) | 14-10-2018 | 192 | 98 | 80 | 49
New Merchants
1. Casio
2. Galler
3. Watsons
4. Simple UK
5. LG
Trending
Purchase
Product | Purchases
1 | iPhone XS MAX 128GB | 53
2 | Gucci Graphic Tee | 43
3 | Adidas Ultraboost 3.0 White | 35
4 | Adidas NMD Black/White | 31
5 | Nike Airmax '97 Silver | 24
Product
Product | Views
1 | Adidas Ultraboost 3.0 White | 653
2 | Bose Bluetooth Speakers | 523
3 | iPhone XS 64GB | 511
4 | Xiaomi 1000mAh Powerbank | 415
5 | Samsung S9 | 124
Store
Merchant | Views
1 | Apple Inc. | 7653
2 | Adidas | 7523
3 | Nike | 7511
4 | Bose | 6415
5 | Razer | 5124


TROLLEE
GIFT CARDS
MY TROLLEE
SHOPPING FROM
SIGN IN
ENG
Welcome, admin1
Dashboard
Activity
Store / Product
Inventory / Sales
Merchant Sales
Merchant
$461,792.40
4,577
21,625
$21.35
4.72
4.72
Apple Inc.
Adidas
Nike
Dyson
Samsung
Thermos
Xiaomi
Fossil
Crumpler
Gucci
Sales by Customer Type
$4,500
$4,000
$3,500
$3,000
$2,500
$2,000
$1,500
$1,000
$500
$0
Jan
Feb
Mar
New
Returning
Recommender performance
Impressions
Purchase
Conversion Rate
900
800
700
600
500
400
300
200
100
0
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Campaigns
24
42
67
59
33
TROLLEE1
SG54
BLACKFRI
XMAS18
GSS18

.5 Visualization of SUS Score per question

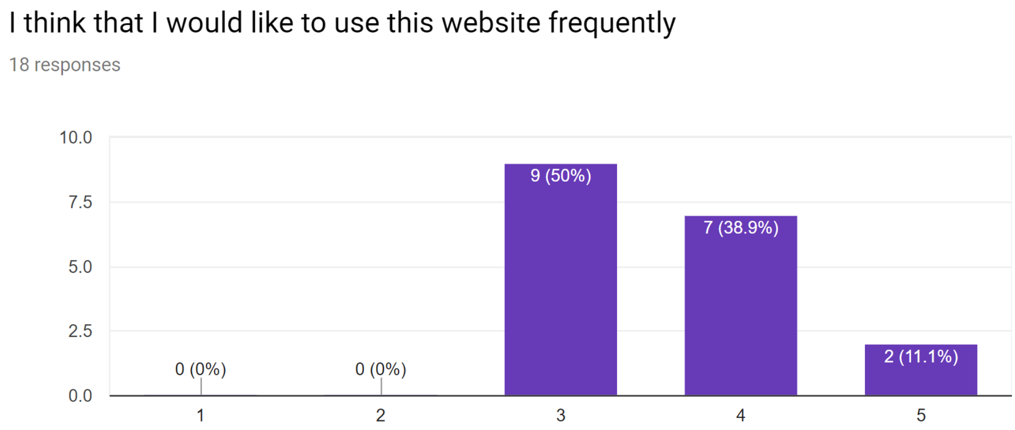


Figure 1: Q1 Results

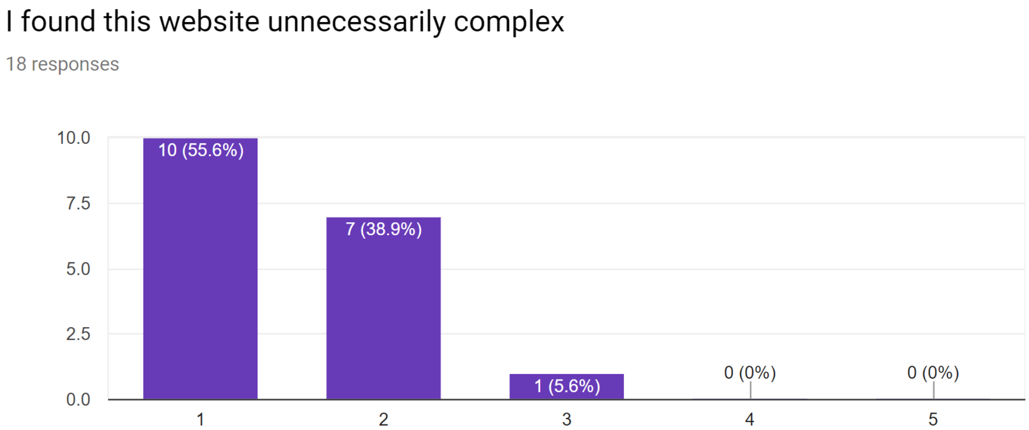


Figure 2: Q2 Results

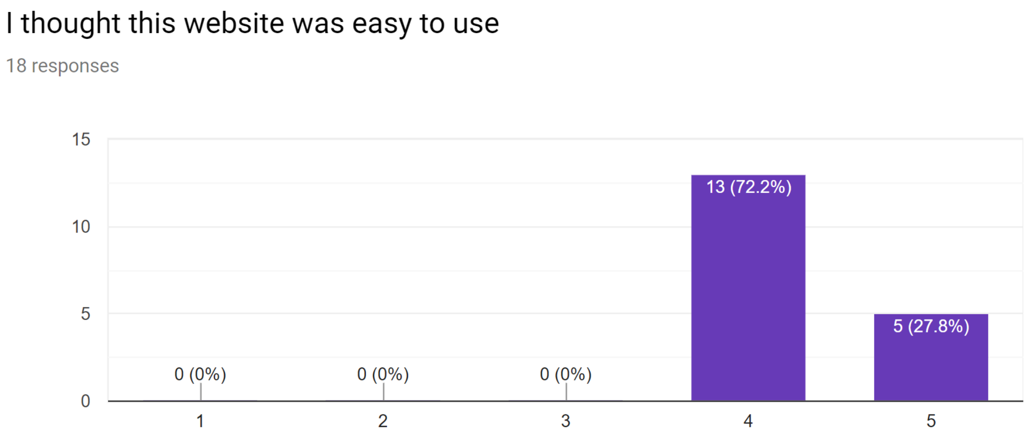


Figure 3: Q3 Results

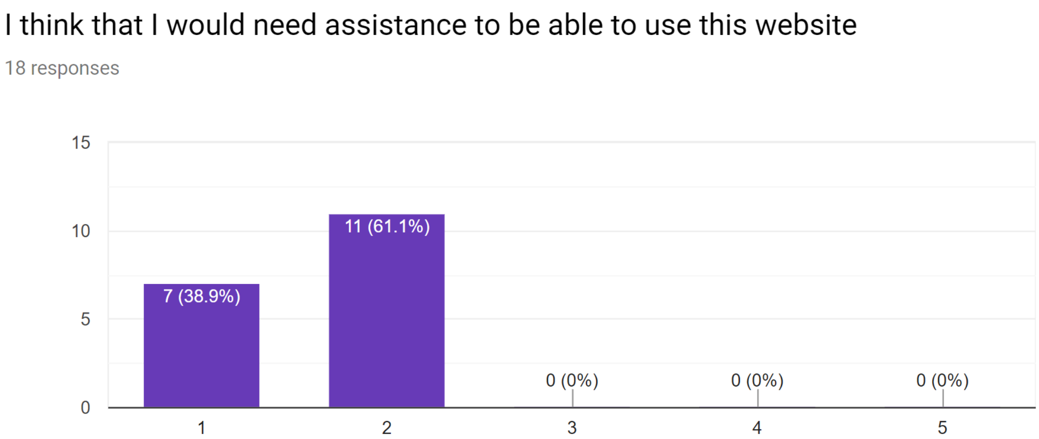


Figure 4: Q4 Results

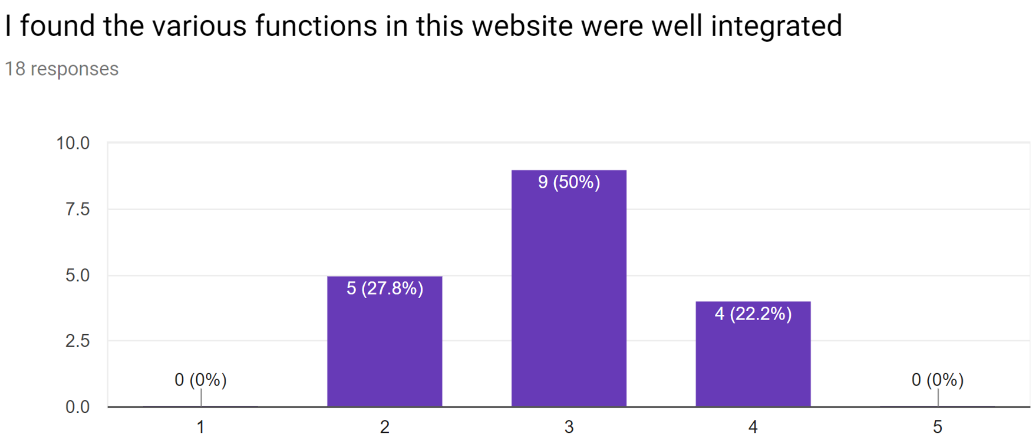


Figure 5: Q5 Results

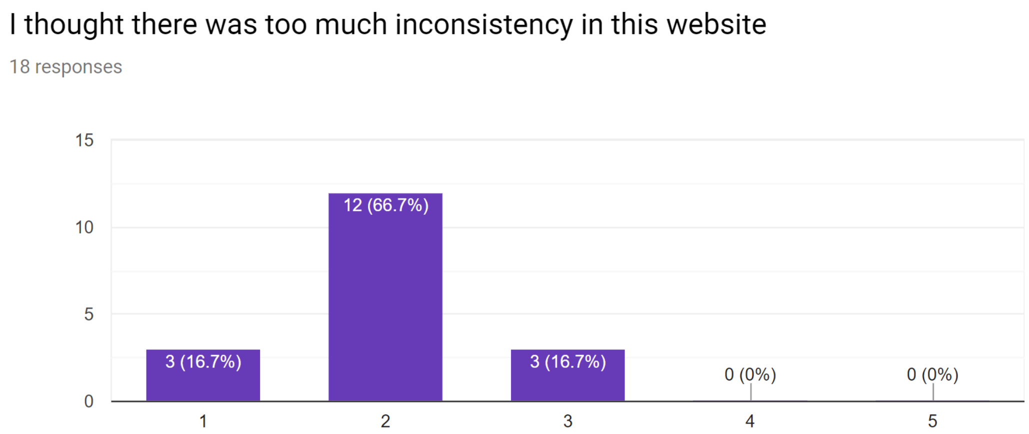


Figure 6: Q6 Results

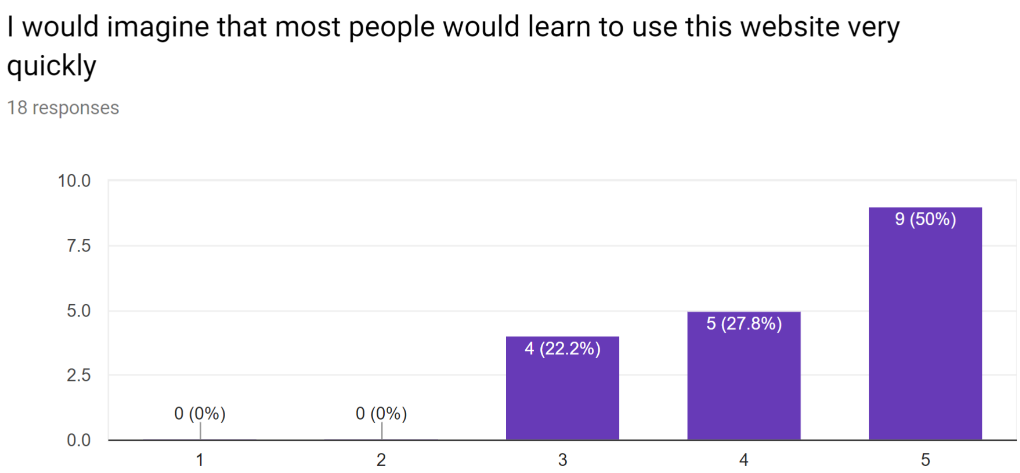


Figure 7: Q7 Results

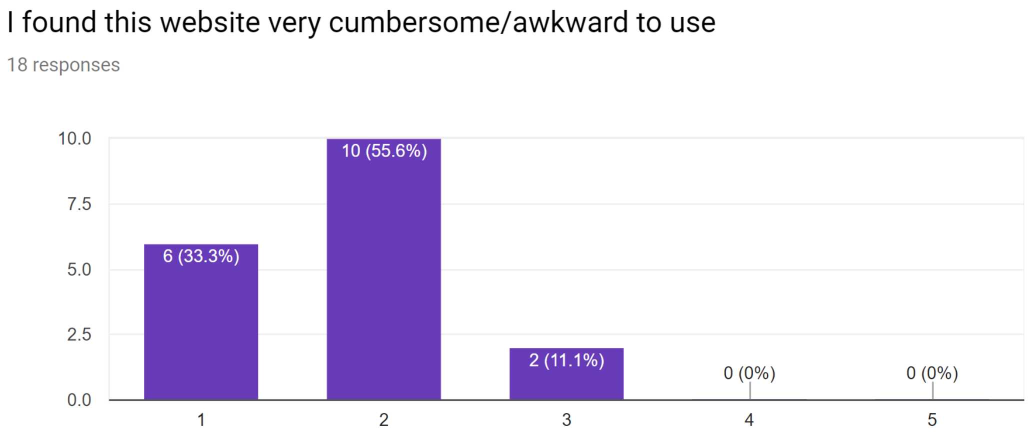


Figure 8: Q8 Results

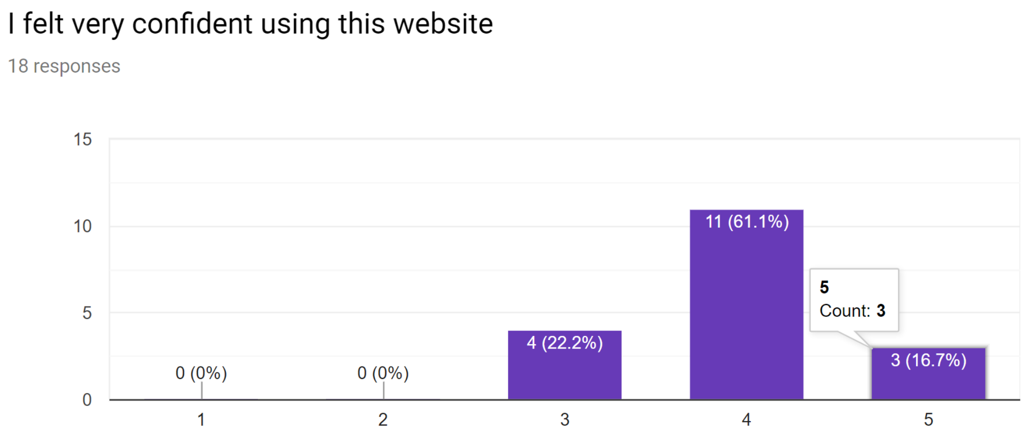


Figure 9: Q9 Results

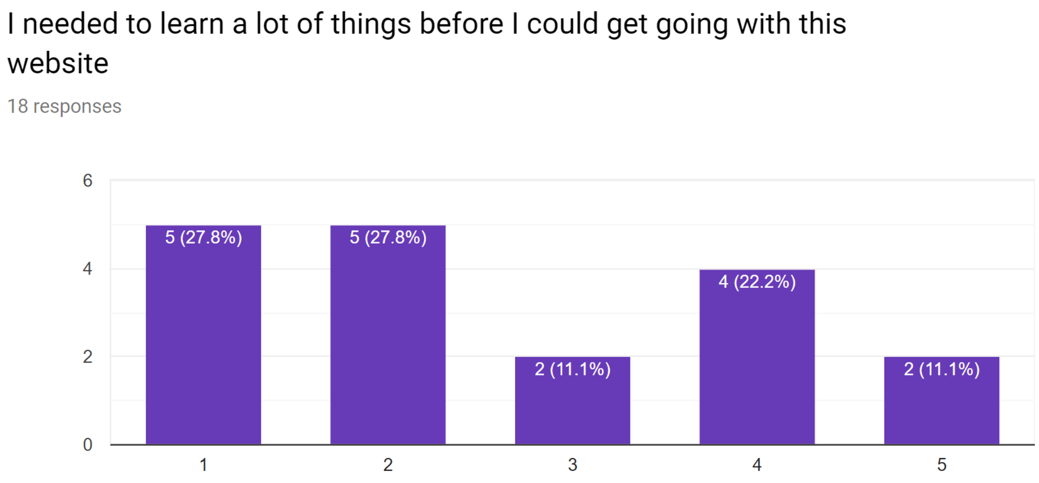


Figure 10: *Q10 Results*

.6 Clients' SUS Form

3/27/2019 System Usability Scale (Trollee Analytics)

System Usability Scale (Trollee Analytics)

* Required

1. **What is your name?** *

   Nigel

2. **What is your title?** *

   CEO & Founder

3. **I think that I would like to use this website frequently** *

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree					✓	Strongly Agree

4. **I found this website unnecessarily complex** *

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree		✓				Strongly Agree

5. **I thought this website was easy to use** *

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree				✓		Strongly Agree

6. **I think that I would need assistance to be able to use this website** *

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	✓					Strongly Agree

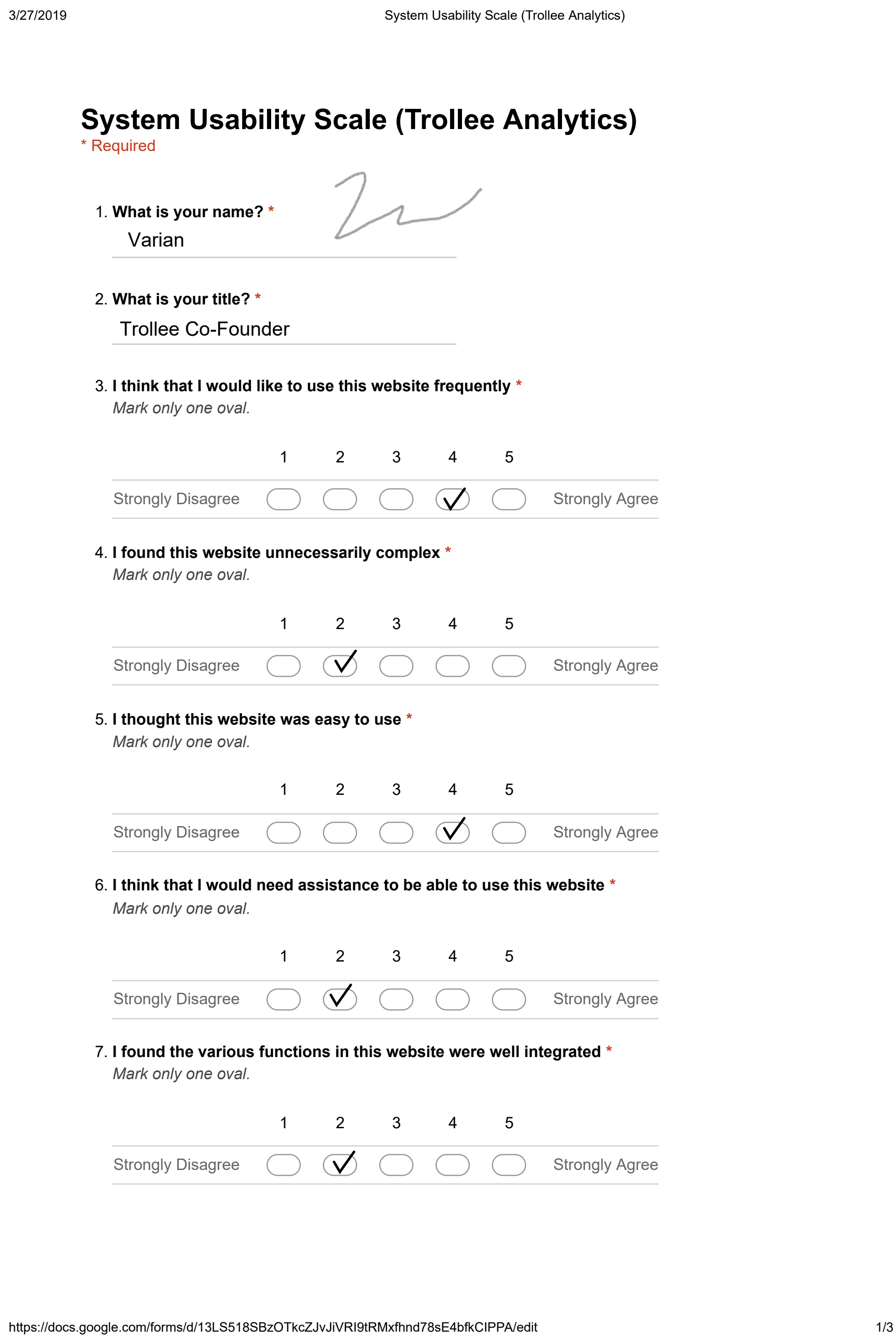
7. **I found the various functions in this website were well integrated** *

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree				✓		Strongly Agree

https://docs.google.com/forms/d/13LS518SBzOTkcZJvJiVRI9tRMxfhnd78sE4bfkCIPPA/edit 1/3

Figure 11: Trollee Founder & CEO's SUS


3/27/2019 System Usability Scale (Trollee Analytics)

System Usability Scale (Trollee Analytics)

* Required

1. **What is your name?** *

   Varian

2. **What is your title?** *

   Trollee Co-Founder

3. **I think that I would like to use this website frequently** *

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree				✓		Strongly Agree

4. **I found this website unnecessarily complex** *

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree		✓				Strongly Agree

5. **I thought this website was easy to use** *

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree				✓		Strongly Agree

6. **I think that I would need assistance to be able to use this website** *

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree		✓				Strongly Agree

7. **I found the various functions in this website were well integrated** *

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree		✓				Strongly Agree

https://docs.google.com/forms/d/13LS518SBzOTkcZJvJiVRI9tRMxfhnd78sE4bfkCIPPA/edit 1/3

Figure 12: Trollee Co-Founder's SUS

System Usability Scale (Trollee Analytics)

* Required

1. **What is your name? ***

   Dr. Danny Poo

2. **What is your title? ***

   Consultant

3. **I think that I would like to use this website frequently ***

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree				✓		Strongly Agree

4. **I found this website unnecessarily complex ***

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree		✓				Strongly Agree

5. **I thought this website was easy to use ***

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree				✓		Strongly Agree

6. **I think that I would need assistance to be able to use this website ***

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree		✓				Strongly Agree

7. **I found the various functions in this website were well integrated ***

   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree				✓		Strongly Agree

Figure 13: Trollee Consultant's SUS

System Usability Scale (Trollee Analytics)

* Required

1. **What is your name?** *
   Sapumal

2. **What is your title?** *
   External

3. **I think that I would like to use this website frequently** *
   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree					✓	Strongly Agree

4. **I found this website unnecessarily complex** *
   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree		✓				Strongly Agree

5. **I thought this website was easy to use** *
   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree					✓	Strongly Agree

6. **I think that I would need assistance to be able to use this website** *
   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree		✓				Strongly Agree

7. **I found the various functions in this website were well integrated** *
   Mark only one oval.

	1	2	3	4	5	
Strongly Disagree			✓			Strongly Agree

Figure 14: Trollee External Consultant's SUS

8 Bibliography

[1] Domo, "Becoming a data-driven ceo," 2018.

[2] M. van Rijmenam, "A short history of big data," Available at https://datafloq.com/read/big-data-history/239.

[3] Facebook, "People-first analytics for an omni-channel world," Available at https://analytics.facebook.com.

[4] M. Rouse, "What is an omnichannel?" Available at https://searchcio.techtarget.com/definition/omnichannel.

[5] Spotify, "Get recommendations based on seeds," Available at https://developer.spotify.com/documentation/web-api/reference/browse/get-recommendations.

[6] A. L.-C. Wang, "An industrial-strength audio search algorithm," 2003.

[7] Netflix, "How netflix's recommendations system works," Available at https://help.netflix.com/en/node/100639.

[8] Amazon, "Real-time product recommendations," Available at https://aws.amazon.com/mp/scenarios/bi/recommendation/.

[9] D. Lee, "Three essays on big data consumer analytics in ecommerce," 2015.

[10] S. F. W. Shahriar Akter, "Big data analytics in e-commerce: A systematic review and agenda for future research," 2016.

[11] P. G. Roetzel, "Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development," 2018.

[12] M. Driver and S. Streufert, "Integrative complexity: An approach to individuals and groups as information processing systems," *Administrative Science Quarterly*, vol. 14, no. 2, pp. 272–285, 1969.

[13] Domo, "The dynamic decision maker, 1st ed. lincoln: iuniverse," 1990.

[14] M. D. Schroder, H. and S. Streufert, "Human information processing. new york: Holt, rinehart and winston," 1967.

[15] S. Lohr, "Is information overload a $650 billion drag on the economy?" 2007.

[16] N. K. G. F. 7) D. Kiron, R. Shockley and M. Haydock, "Analytics: The widening divide," *MIT Sloan Management Review*, vol. 53, no. 2, pp. 1–22, 2011.

[17] D. Rogers and D. Sexton, "Marketing roi in the era of big data: The 2012 brite and nyama marketing in transition study," Tech. Rep., 2012.

[18] Monetate, "From big data to big personalization," Tech. Rep., 2014.

[19] I. K. Center, "Crisp-dm help overview."

[20] V. O. M. H. A. M. N. A. Joaqu´ın P´erez, Emmanuel Iturbide, "A data preparation methodology in data mining applied to mortality population databases," 2015.

[21] J. H. K. Sang Kyu Kwak, "Statistical data preparation: management of missing values and outliers," 2015.

[22] M. Card and Shneidermann, "Readings in information visualization using vision to think. morgan kaufmann," 1999.

[23] J. Heer, "Value of visualization," Lecture Notes Available at https://courses.cs.washington.edu/courses/cse442/17au/lectures/CSE442-ValueOfVisualization.pdf (28/9/2017).

[24] M. Parkinson, *Mike Parkinson's Do-it-Yourself Billion Dollar Business Graphics: 3 Fast - the power of visual communication*, 3rd ed., 2007.

[25] K. Merieb, E. N. & Hoehn, *Human Anatomy & Physiology*, 7th ed. Pearson International Edition, 2007.

[26] A. Trafton, "In the blink of an eye," 2014.

[27] M. Ellstrøm, "Benefits of presenting information visually and guidelines on how to do it - an introduction to static information visualizations for novices," 2013.

[28] C. Wang and J. Tao, "Graphs in scientific visualization: A survey," 2017.

[29] A. Angra and S. M. Gardner, "Reflecting on graphs: Attributes of graph choice and construction practices in biology," 2018.

[30] S. D. J., "The effective use of graphs," *Journal of wrist surgery*, vol. 3, no. 2, pp. 67–68, 2014.

[31] S. E.-A. S. B. C. H. Felipe Cerdasa, Alexander Kaluzaa, "Improved visualization in lca through the application of cluster heat maps," 2017.

[32] Seaborn, "seaborn.heatmap," Available at 23)https://seaborn.pydata.org/generated/seaborn.heatmap.html.

[33] A. Gandomi and . M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, pp. 137–144, 2015.

[34] J. Phillips, *Ecommerce Analytics: Analyze and Improve the Impact of Your Digital Strategy*. PH Professional Business, 2016.

[35] Z. M. Research, "Global embedded analytics market will reach usd 77.3 billion by 2025," Available at https://globenewswire.com/news-release/2019/02/20/1738031/0/en/Global-Embedded-Analytics-Market-Will-Reach-USD-77-3-Billion-By-2025-Zion-Market-Research.html.

[36] IBM, "Ibm watson analytics," Available at https://developer.ibm.com/watson-analytics/.

[37] ——, "Ibm watson customer experience analytics," Available at https://www.ibm.com/us-en/marketplace/customer-experience-analytics.

[38] Oracle, "Oracle business intelligence 12c—features," Available at https://www.oracle.com/solutions/business-analytics/business-intelligence/features.html.

[39] Tableau, "Products," Available at https://www.tableau.com/products.

[40] ——, "Dashboard starters," Available at https://www.tableau.com/products/dashboard-starters.

[41] Zalora, "Zalora," Available at https://www.zalora.sg/men/?catalogtype=Main.

[42] Shopee, "Shopee," Available at https://shopee.sg/.

[43] Amazon, "Amazon," Available at https://www.amazon.com/.

[44] D. J. V. G. S. Shruthi, "An effective product recommendation system for e-commerce website using hybrid recommendation systems," 2017.

[45] R. van Meteren and M. van Someren, "Using content-based filtering for recommendation," 2000.

[46] N. J. N. Fatemeh Alyari, "Recommender systems: A systematic review of the state of the art literature and suggestions for future research," *Kybernetes*, vol. 47, no. 5, pp. 985–1017, 2018.

[47] P. Melville and V. Sindhwani, "Recommender systems, encyclopedia of machine learning," 2010.

[48] Netflix, "Netflix prize," Available at https://www.netflixprize.com/.

[49] Y. Koren, "The bellkor solution to the netflix grand prize," 2009.

[50] Irvine, "Knowledge-based recommender systems," 2000.

[51] M. L. Marcus Olsson, "A hybrid recommender system for usage within e-commerce," 2017.

[52] I. Sommerville, *Software Engineering 9th Edition Chapter 4*, 2010.

[53] MITECOMMERCE, "E-commerce glossary of terms," Available at http://web.mit.edu/ecommerce/www/glossary.html.

[54] ecommerceIQ, "Ecommerce glossary," Available at https://ecommerceiq.asia/ecommerce-glossary/.

[55] IBM, "20 ecommerce terms in simple language," Available at https://www.ibm.com/blogs/watson-customer-engagement/2015/04/24/20-ecommerce-terms-in-simple-language/.

[56] Investopedia, "Impression," Available at https://www.investopedia.com/terms/i/impression.asp.

[57] M. Cohn, "User stories applied: For agile software development," 2004.

[58] K. Waters, "Prioritization using moscow," 2009.

[59] Olist, "Brazilian e-commerce public dataset by olist," 2018.

[60] N. Babich, "Everything you need to know about wireframes and prototypes," 2017.

[61] Noon.com, "Noon.com," Available at https://www.noon.com/uae-en/.

[62] J. M. Phillips, "Min hashing," 2013.

[63] A. R. J. D. U. Jure Leskovec, *Mining of Massive Datasets. Chapter 3 - Finding Similar Items*, 2014.

[64] P. I. V. S. M. Mayur Datar, Nicole Immorlica, "Locality-sensitive hashing scheme based on p-stable distributions," 2003.

[65] A. Rajaraman, "Near neighbor search in high dimensional data (1)."

[66] Google, “Google charts line chart,” Available at https://developers.google.com/chart/interactive/docs/gallery/linechart.

[67] CanvasJS, “Canvasjs line chart,” Available at https://canvasjs.com/html5-javascript-line-chart/.

[68] amCharts, “amcharts xy chart with value-based line graphs,” Available at https://www.amcharts.com/demos/smoothed-line-chart/.

[69] D3js, “D3 v5 line chart,” Available at https://bl.ocks.org/gordlea/27370d1eea8464b04538e6d8ced39e89.

[70] Plotly, “Line charts in plotly.js,” Available at https://plot.ly/javascript/line-charts/.

[71] TIOBE, “Tiobe index,” Available at https://www.tiobe.com/tiobe-index/.

[72] P. R. Christopher D. Manning and H. Schütze, “Introduction to information retrieval, chapter 2 the term vocabulary and postings lists, determining the vocabulary of terms, dropping common terms: stop words,” 2008.

[73] E. Zhu, “Minhash lsh forest,” Available at https://ekzhu.github.io/datasketch/lshforest.html#minhash-lsh-forest.

[74] P. G. Mayank Bawa, Tyson Condie, “Lsh forest: Self-tuning indexes for similarity search,” 2005.

[75] M. Patton, “Qualitative evaluation and research methods (pp. 169-186),” 1990.

[76] A. Bangor, P. T. Kortum, and J. T. Miller, “An empirical evaluation of the system usability scale,” *International Journal of Human-Computer Interaction*, vol. 24, no. 6, pp. 574–594, 2008.

[77] J. Nielsen, “Thinking aloud: The 1 usability tool,” 2012.

[78] Fletcher and Islam, “Comparing sets of patterns with the jaccard index,” 2018.

[79] P. Jaccard, ““etude comparative de la distribution florale dans une portion des alpes et du jura”, bulletin de la soci´et´e vaudoise des sciences naturelles 37(1): 547–579.” 1901.