The link to the online version of this document can be found [here](https://docs.google.com/document/d/1Dn6-WnhDafuYkzLY5uYqMxsSqQqUmgbgQ7XcW87_7KY/edit?usp=sharing):

**CHAPTER THREE**

**Description Of The Existing System**

In recent times, a lot of systems have been built to address advertisement problems. These systems enable business owners to get a wide range of audience which in turn lead to higher revenue. These systems, after receiving payment from business owners, display their products and services either in a search engine result, social media platform news feed, other mobile and web applications etc.

Google search engine for an example displays an "Ads" label on the product/services it's advertising. It always displays in the form of a search result after a search is made on the engine.It could either be at the top of the search result page or at the bottom of the page.

Facebook and other social media platforms display advertisement products in their news feed which will enable the platform users to gain access to those products and services.

Youtube(now owned by Google), a video streaming platform intercepts the video a particular user is watching at a particular time with an advertisement mostly in the form of a short video describing the product or service.

There are also situations where a particular application is completely interrupted by an advertisement. This will lead to a complete hang up of the background application for the period in which the advertisement is being displayed.

In most cases, the user always has the option to skip a particular advertisement but it's usually after a short period of time, sometimes 5 seconds and sometimes more.

**Problem Of The Existing System**

According to the description of the above system(s), I identified the following problems:

1. Users are usually irritated and embarrassed by sudden pop ups of contents that disrupts their initial activities on a platform. It is a form of online disturbance to most users, especially those with zero intention of purchasing anything and are just looking for a smooth online experience either from, browsing, online research and watching their favourite series.

2. Users mostly close up advertisements because it sometimes contradicts their interest as the advertisement didn't take into consideration their interest, shopping details and demographics which is a loss for the business owner. Also, whenever there is a click on an advertisement without an actual purchase, It's a los for the business owner and sadly, this is always the case.

3. The cost of advertisement on a small product or services are mostly greater than the sales made from the advertisement which is a loss on the path of the business owner. This advertisement cost can also increase the price of products and services making them much more difficult for customers to afford.

4. It is easy for big industries to take control of the market by grossly increasing their advertisement which can also lead to better but small businesses going out of reach and out of market.

**System Design**

This system runs on a machine learning base model, The system returns personalised advertisement of available products to customers depending on their demographics and recent activities in the store by sending their details to the machine learning base model with determines the kinds of products that will be returned to customers as advertisement.

**Defining some terms**

**Machine Learning**: according to Wikipedia, Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks. it is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

**Python**: according to Wikipedia, Python is a high-level, interpreted, general-purpose programming language. Its design philosophy emphasises code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. The main project was built in python being the general language for data science and machine learning.

**Django**: according to Wikipedia, Django is a free and open-source, Python-based web framework that follows the model–template–views architectural pattern. The backend(Model deployment) aspect of this project was built in django.

**Machine Learning Algorithm**: is the method by which the AI(Artificial Intelligence) system conducts its task, generally predicting output values from given input data. The two main processes of machine learning algorithms are classification and regression. This project uses the Random Forest Classifier algorithm(supervised machine learning).

**Model**: according to Google, machine learning model is a file that has been trained to recognize certain types of patterns. A model is trained over a set of data, providing it an algorithm that it can use to reason over and learn from those data. The name of such a file usually ends with ".sav"and ".pkl" among others.

**HTML**: stands for HyperText Markup Language, it is used to structure a web page and its content.

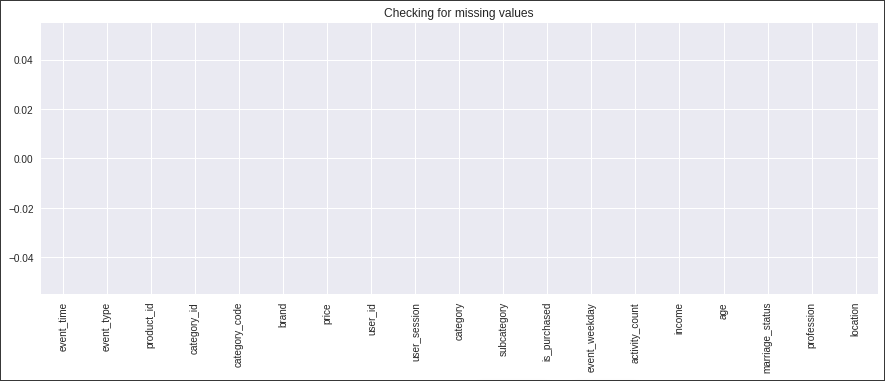
**CSS**: stands for Cascading Style Sheet, it is used to design the structure and content created using HTML.

**The architectural, algorithmic, mathematical and process analysis are as follows**:

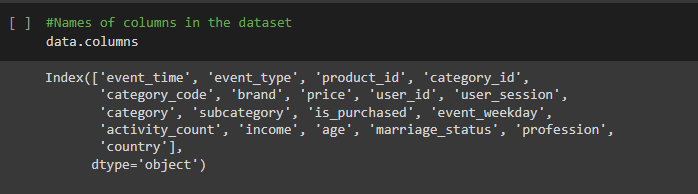
**1.** **Data Collection**: In every data science or machine learning project, The first and also an important stage is to gather data from appropriate source(s) which will both be used for analysis and also in the training of the machine learning algorithm. The data used for the project is a live observation of the behaviours of different users in an ecommerce online store which was obtained from kaggle at https://www.kaggle.com/datasets which is the section of the platform where real world datasets can be found. Kaggle is an online community platform for data scientists and machine learning enthusiasts. Kaggle allows users to collaborate with other users, find and publish datasets, use GPU integrated notebooks, and compete with other data scientists to solve data science challenges.

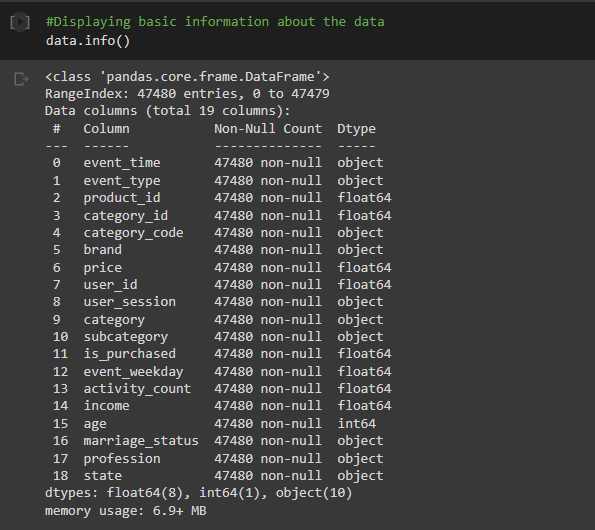
**2. Exploratory Data Analysis(EDA)**: This phase of the processes of understanding the data using visualisations and statistical tools. It is the process of digging into the data to get intuitions that will be needed in the data preprocessing phase. The following is the result of the EDA.

From the analysis, the dataset have 47482 observations(rows), 19 features(columns) with no missing value

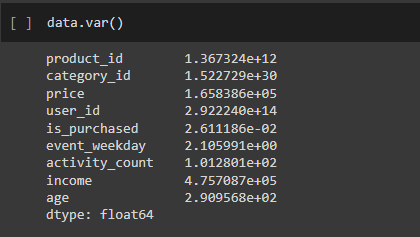


The total number of customers the store has is **32431** and the list of all the columns is shown below.

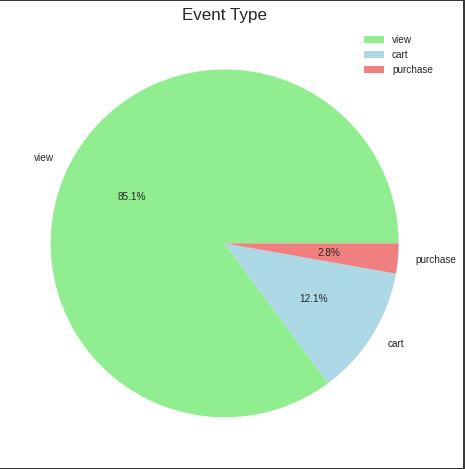




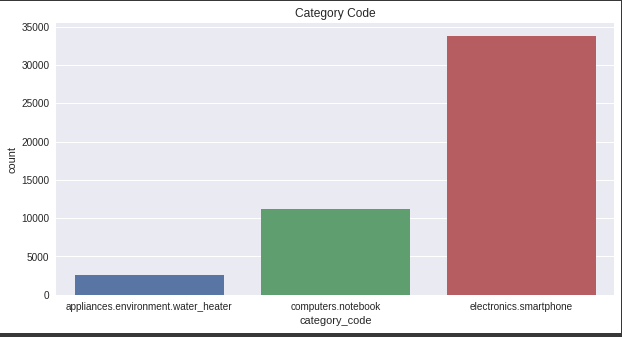
The variance of each numerical column is within the same range except for the income column that's slightly higher but not too high which is because the purchase power of an individual largely depends on their income.



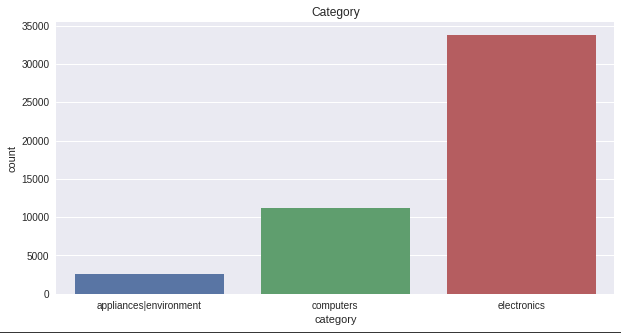
The chart below is a distribution of the proportion of each activity type, of which the highest was view, which is always true in an ecommerce store because anyone and everyone is permitted to view all the products in the store at any time. Cart follows view and the least activity is the actual purchase.

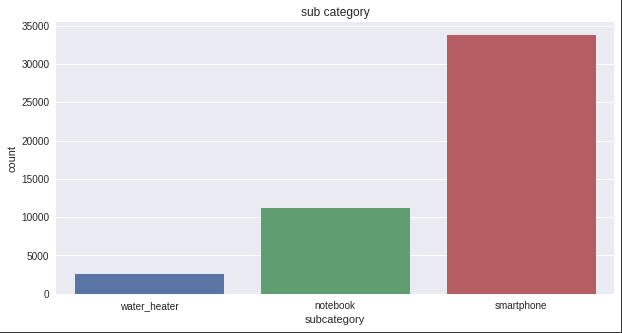


From the category code column, there are three values of which the one with the highest frequency is Electronic Smartphone. This means that there are more sales from smartphone



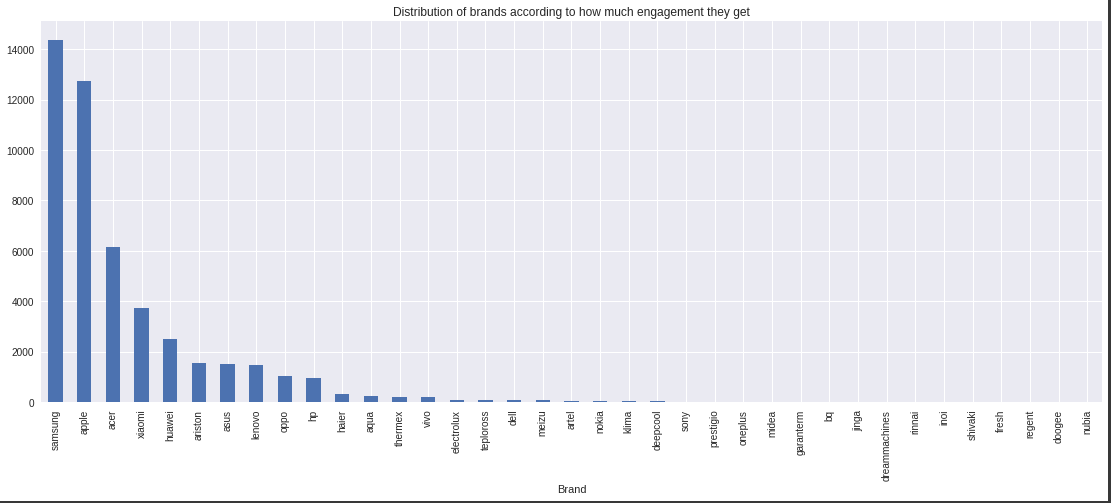
We also got similar charts from the category and sub category columns





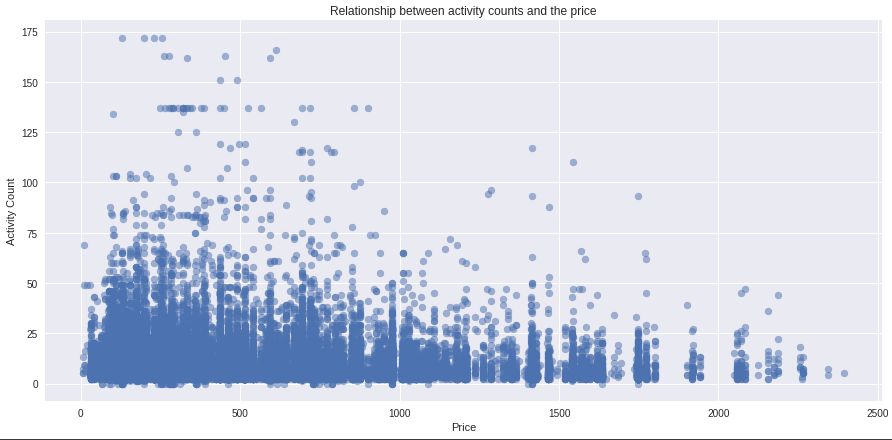
Due to the strong and obvious correlation between these three columns, only one of them will be considered for developing the model.

From the visualisation of the brand column, we can see that the samsung brand has the highest interaction and engagement from the customers which means there's a strong preference of samsung brand/products to other brands according to the dataset. The brand column contains all the product brands available in the store.



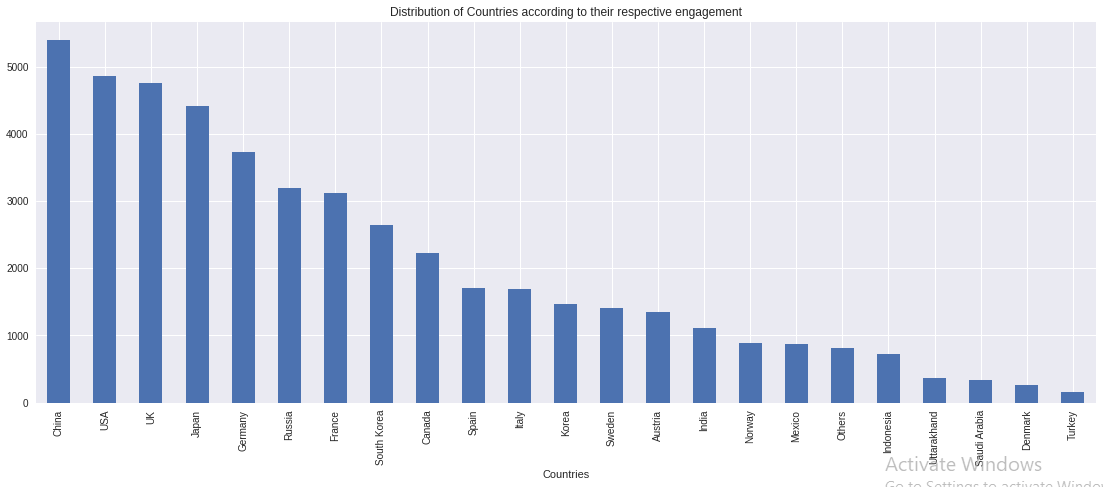
The second most popular product is Apple which can be justified by the love teenagers have for apple products and the third most popular product is acer, also a very common brand of laptops/computers.

The activity counts column records the number of activities performed by a user and the price column gives information about the price of the products. These two columns were plotted to check for relationships between them and a weak negative correlation was detected.



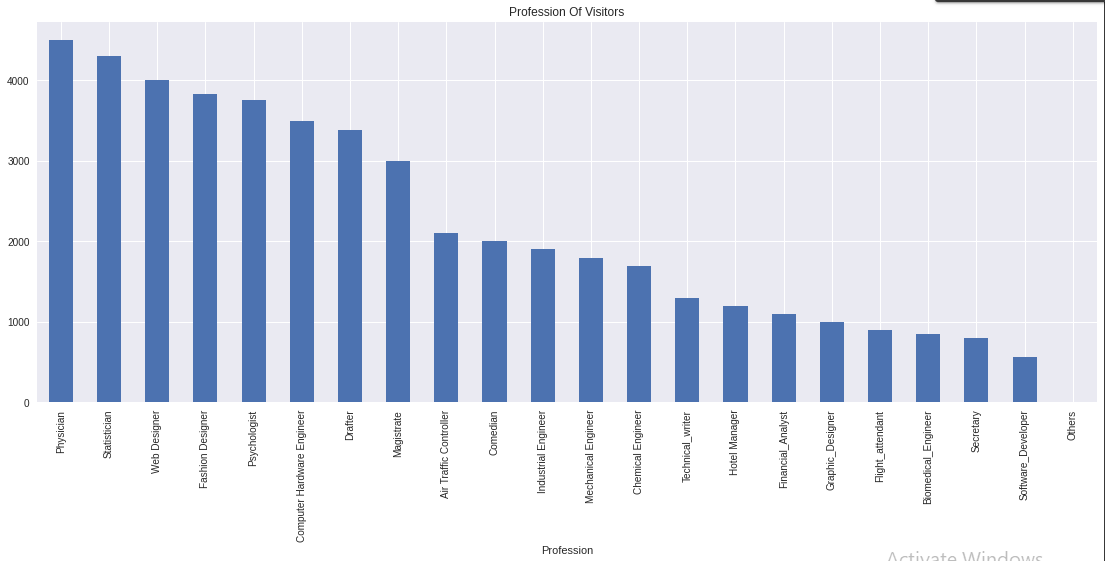
The negative correlation tells us that as the activity counts fall, the price of products tends to rise and vice versa.

The country column gives us information about the location of each customer

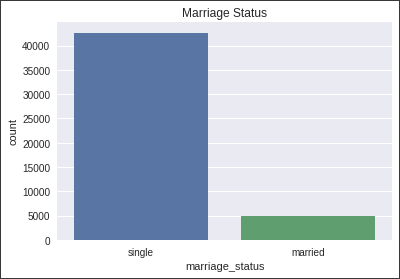


The country with the highest engagement is China followed by the USA. This is also justifiable because these two countries are part of the countries of the world that have large ecommerce markets.

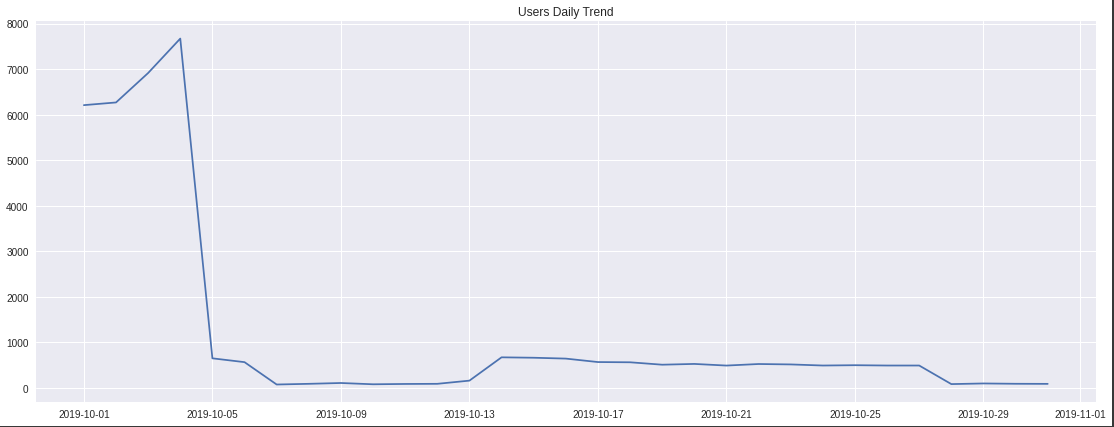
The profession column represent the job titles of customers and users of the ecommerce store with Physician being the highest profession of the platform users



Only a few customers are married and a large number are not married.



A chart of users' daily trend was also plotted with the event\_time column which contains information about the date of activities on the platform.

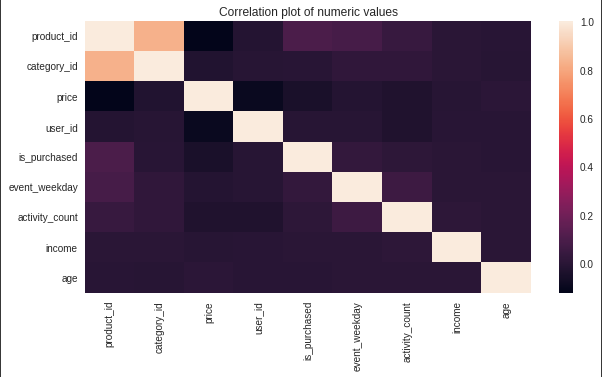


The highest visits occur on 2019-10-05(Saturday) which could be as a result of the weekend when there is little or no work done by most people.

Out of 32431 that visited the store, only 1123 bought a product and only 4524 carted a product. These values were obtained by first selecting the instances of the dataset where the value of event\_type is equal to purchase and then the unique method to get the unique values of user\_id. Similar process was carried out to determine the number of customers that carted a product.

Samsung products have the highest sales and also have the highest number of carts.

A correlation plot was made to detect values that correlated with one another. Correlated columns provide the same information which will not add any information to the machine learning model, it could in turn worsen the performance of the model.



The correlation between product id and category id is the highest, the two columns will be dropped as they won't give more information to the model(ideally, only one should be dropped). Also note that correlation can only happen between numerical values.

**3. Data Preparation**: This phase makes the data ready for modelling. It requires a numeric representation of raw data to be fitted to models. It is also a process of generating more features using domain knowledge to improve the performance of the mode

The following procedures were carried out to ensure we gave a great model.

The 'subcategory' and 'category\_code' columns were dropped as they provided the same information with the 'category' column as established from the analysis above in Exploratory Data Analysis.

The event\_type and is\_purchased column also provides similar information, hence only the event\_type column was used for modelling.

According to the subcategory and category\_code column, it is evident that Electronics represent smartphone and Computers represent computers. So, the 'appliances|environment' value in the category column will be replaced with the most frequent(smartphone) and other values will be renamed accordingly.

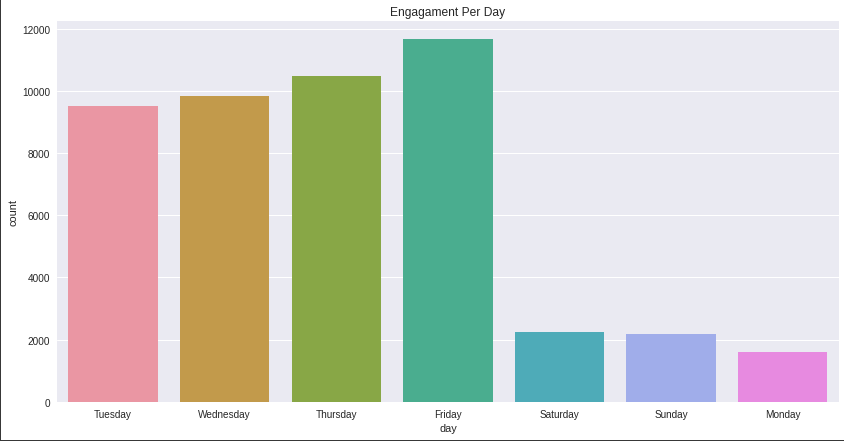
Three(3) new columns were added namely year, month and day. These columns were extracted from the event\_time column which was initially an object column but was converted to a datetime column since it contains the year, month, day and time altogether. Without the conversion to datetime, it will be impossible to extract those new features(column).

The new year feature has only a single value meaning all events happened in that single year(2019), hence, the column was dropped before there is nothing for the model to learn from it. same as the month column, all events happened in the month of October.

The week day column was also dropped because it was found to contains the same information as the new day(day of the week) column

The addition of the new column(the three column that was added now remain one because the year and the month column has been removed) gave rise to new analysis which follows:

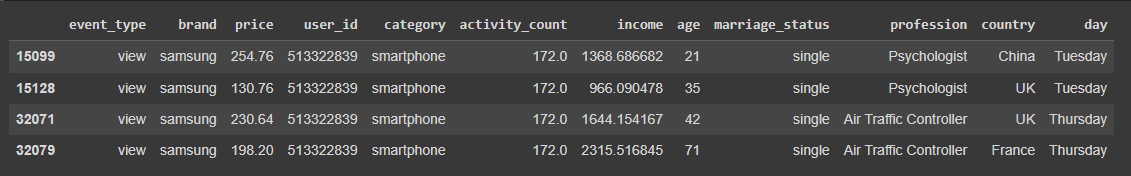
The visualisation of engagement per week day”



The day with the highest engagement is Friday followed by Thursday and the least on Monday. This could be due to the fact that there is always a lot of workload on Mondays being the first workday of the week and hence, there will be less time for shopping.

From the analysis, samsung products has the highest engagement on friday and was also the most bought product brand on fridays just as expected(due to the large engagement it gets)

The highest activity count is 172, the chart below shows the activity that happens at the highest activity count



High activity happens on Tuesday and Thursday,

Event\_type is view meaning the customers only view alot of samsung(brand is samsung) products on Tuesday and Thursday

Marriage Status is Single meaning the more unmarried people add more time for more activities on the ecommerce platform

Profession of those customers with the high activity are Psychologist and Air Traffic Controller

Countries: China, France and UK

Other Analysis:

The country with the highest purchases is China

The country with the highest carts is China

Age 72 is the age that buys more products according to the dataset

Age 68 is the age that Carts more products according to the dataset

The Profession with the highest purchases is Physician

The Profession with the highest carts is Magistrate

**Further Data Preparation**:

To get a better prediction, Only the top brands(samsung, apple, acer, xiaomi, huawei) with highest engagement will be used for the model development and hence, other values were replaced with others since they are small and negligible. This means the machine learning model that was built will be predicting six classes(samsung, apple, acer, xiaomi, huawei and Others).

The dataset was undersampled to avoid imbalance of the classes which can lead to biases of the model (Predicting only one/certain classes). Imbalance occurs when one or more classes have very low proportions in the training data as compared to the other classes because most machine learning algorithms assume that the classes are evenly distributed. After downsampling, the dataset will contain an equal number of each class of the dependent feature(brand). It also reduces the size of the machine learning model which makes it easier for further computation(deployment).

Knowing fully well that machine learning algorithms work with numerical values, all categorical(non numbers) were then converted to integers using the map function to assign a new number to each categorical value of each class in each feature of the dataset.

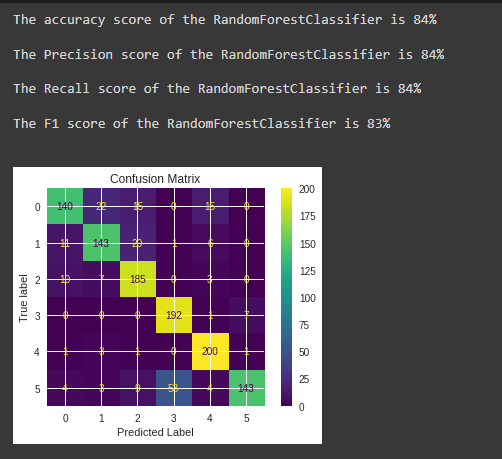
The new dataset with all features converted to integers is now ready for training after confirming it has no missing values.

But before the training, the dataset was separated into x and y variables with x containing the independent features(every other feature except brand) and y containing the feature to be predicted or target variable(brand). And then finally, a module called model selection from scikit-learn library(a major machine learning library in python) was used to split the data into training and test set in the ratio 90:10, 10% of the data for testing while 90% for training.

**4. Model Building**: The model building is usually a straightforward process of fitting the data splitted into the training set to a machine learning algorithm. In this project, the training data set was fitted to a Random Forest Classifier algorithm, hence we had a random forest model.

The Random Forest Algorithm is composed of different decision trees, each with the same nodes, but using different data that leads to different leaves. It merges the decisions of multiple decision trees in order to find an answer, which represents the average of all these decision trees. A decision tree predicts a class label for a record by starting from the root of the tree. it compares the values of the root attribute with the record’s attribute. On the basis of comparison, it follows the branch corresponding to that value and jumps to the next node. For more on random forest algorithm, you can check [here](https://towardsdatascience.com/understanding-random-forest-58381e0602d2)

The random forest model was evaluated with the metrics in the picture below



The Accuracy Score is the number(percentage) of correct predictions over all predictions. It is the base metric used for model evaluation often.

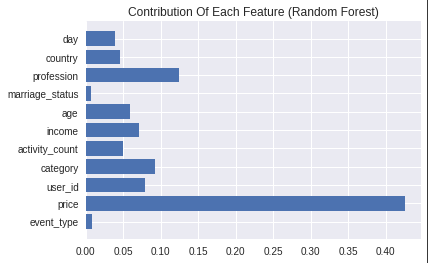
Precision Score is a measure / percentage of how many of the positive predictions made are correct.

Recall score is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made.

F1 score sums up the predictive performance of a model by combining precision and recall. It is a very important metric for the evaluation of machine learning models.

According to Google, a confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarised with count values and broken down by each class. This is the key to the confusion matrix.

Other algorithms like Logistics Regression, Naive\_Bayes, Decision tree were also trained, though the decision tree model performed slightly better than the random forest but the decision tree does not consider most of the features for it prediction while the random forest consider all of the features for it's prediction, hence it's more reliable to use the random forest model.



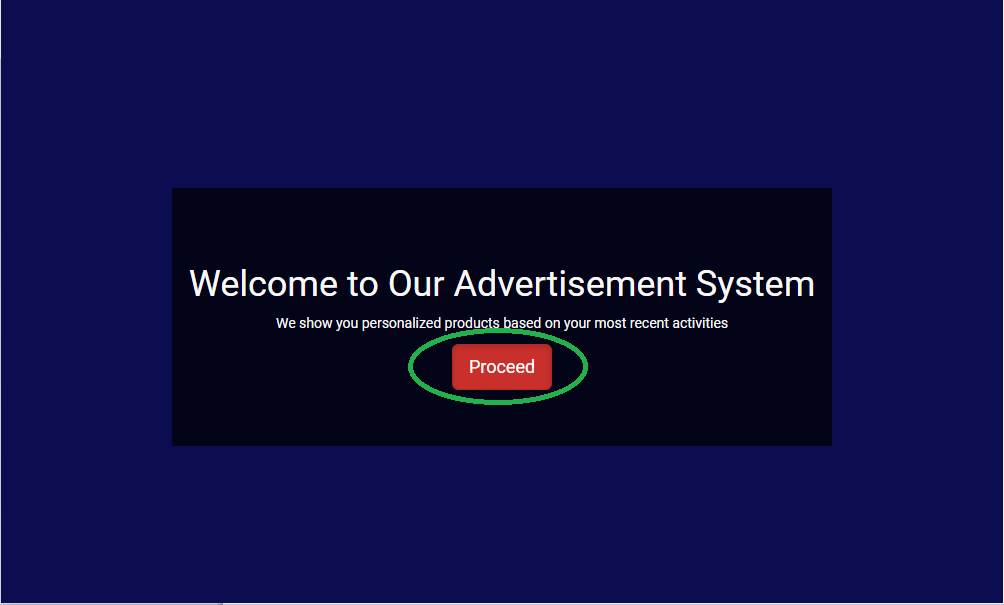
Generally, The price of a product is the most important feature anyone would consider before purchasing a product which is also obvious in the above chart.

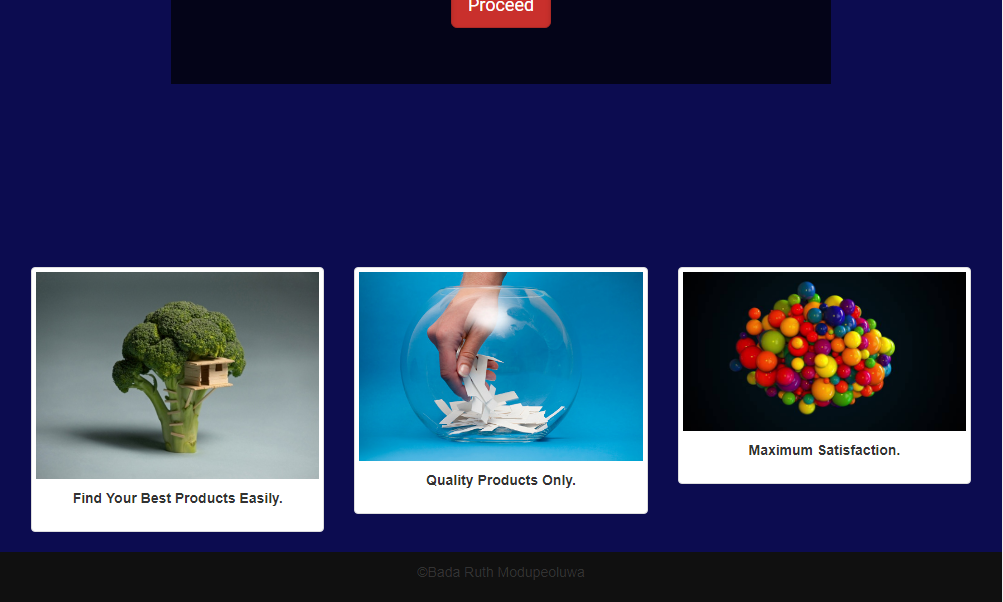
**5. Model Deployment and the main system** **development:**

The web application was built with html, css, django and bootstrap. Bootstrap is a free and open-source CSS framework directed at responsive, mobile-first front-end web development. It contains HTML, CSS and JavaScript-based design templates for typography, forms, buttons, navigation, and other interface components.

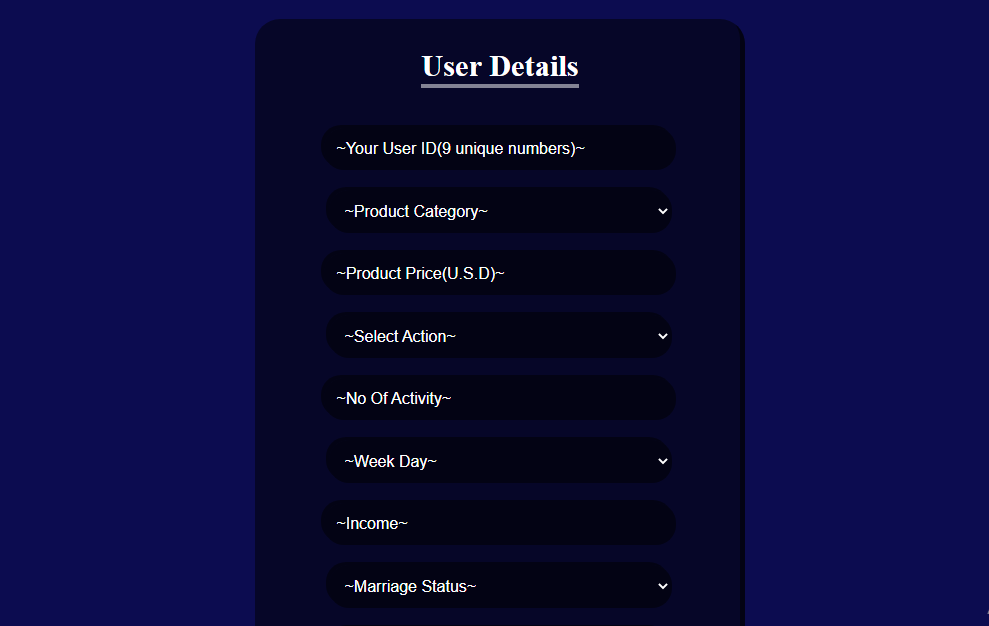
The user interface(frontend) was built with html, css and bootstrap while the server side(backend) was built with django.

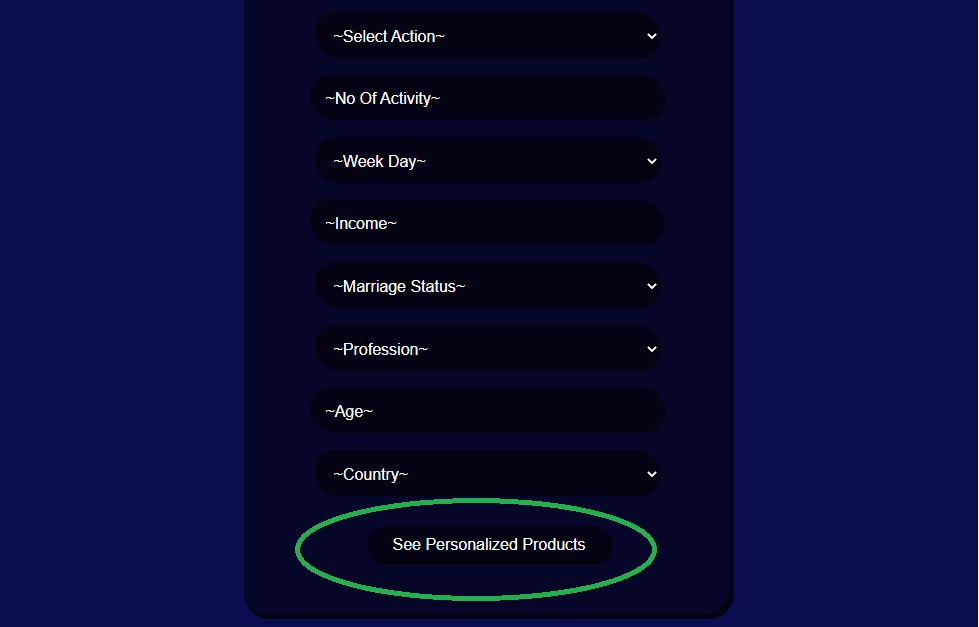
When the application runs, it first displays a welcome page to the user as shown below



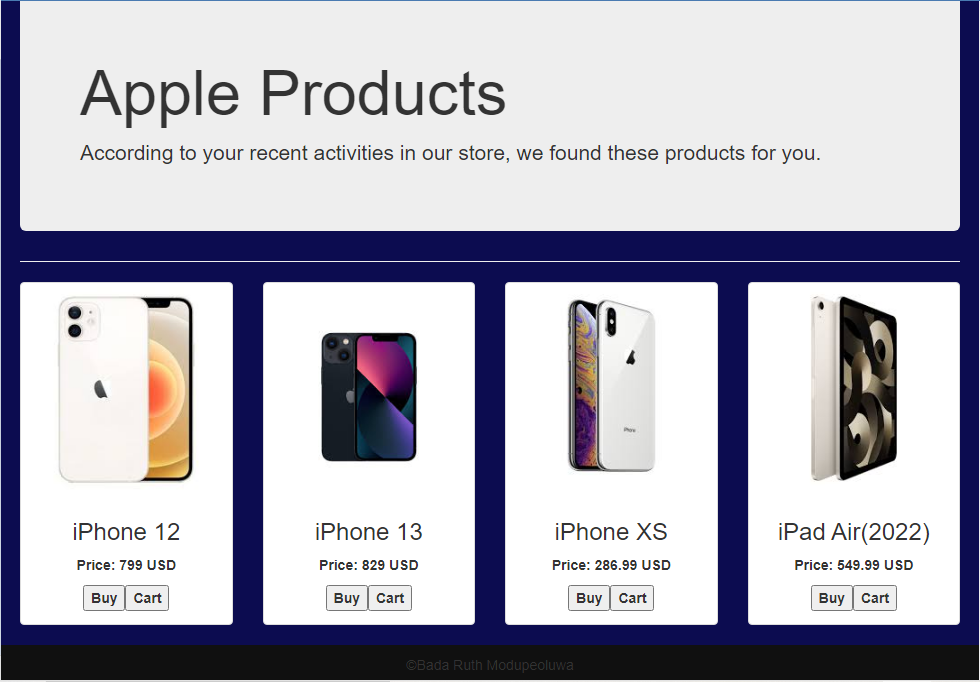


and the user can then click on the proceed button to continue with inputting his details(last shopping details and demographic) in the next page shown below





After the details have been filled by the user, these details are sent to the machine learning base model which predicts the product brand. Based on the predicted brand, a set of products are then returned back to the user as Advertisement which is also shown below.



In this case, the Apple product brand was selected and other related products were also displayed. The user can then choose to buy or cart the products.

**CHAPTER FOUR**

**Output Specification and Design**

The section contains the details of the project functionalities. In other words, It contains the details of what the end users can do with the product/project.

This project is designed to show advertisements that only correspond with the user's demographics and shopping experiences over time. It is also design to reduce the cost of advertisement by removing the additional advertisement expenses after the complete integration of the system into any ecommerce store or platform.

It will also make every business unique without unwanted interference and spy on other platforms. It gives privacy to users as their data will not be transferred or monitored elsewhere.

Model Deployment and the main system development above also contains more information about the output specification and design.

**Input Specification and Design**

Input specialisation is the exact procedures that must be put in place while building the system. For this project, The input specialisation can be deduce from initial system design in chapter 3.

**The data collection process**: The dataset that will be used for the project must be a live observation of the behaviours of different users in an ecommerce online store.

**Exploratory Data Analysis(EDA):** All columns in the dataset must be carefully explained both in words and graphs where necessary. Relationship graphs between two or more columns must be recorded and explained. Correlation plot and every other needed information that aids in the understanding of the dataset must be duly recorded.

**Data Preparation**: All irrelevant and highly correlated columns must be removed as they will not add any information to the model that is to be built, missing values must be replaced with any suitable method, new indispensable columns must be created from the existing one(s). Like, extracting the data, month and year from a datetime column and all processes that will result in the emergence of a very good model must be observed.

**Modelling**: The prepared data must be tested on different different algorithm to see the only that works best after evaluation with the appropriate classification metrics like f1 score, accuracy score, recall score, precision score etc of which none of the already mentioned scores should be less than 0.8(80%). The model might also be fine tuned(boost/upgraded) if a lower metric is recorded.

**Model Deployment and the main system development**: the model deployment might be done with any suitable framework like django in this case. An html form must be created to collect the user's information(shopping details and demographic) which will then be sent to the model and then the model returns a response which will then be used to determine the response(advertisement) that will be sent back to the user.

**File Design**

The files and folders used for the development of the project is as follows:

**The mlads folder**: every django project must have a project name, hence mlads is the name given to the django project, It contains the following file:

->settings.py: is a core file in django projects which holds all the configuration values that the project needs to work

->urls.py: the urls.py file in the mlads folder was used to connect the django project(mlads) with the django application(mldeploy).

All other files were left in the default state as no change was required of them.

**The mldeploy folder**: mldeploy is the name of the django application inside the django project. it contains the following files:

->urls.py: the urls.py in the django application receives a web request and sends it to a matching function in the views.py file.

->views.py: contains the main functions on which the project is built. it take the web(http) request from urls.py and returns http response(html pages in this case) according to different tasks and mission

All other files in the mldeploy folder were left in their default state as no change was required of them.

**db.sqlite3 file:** is a database file that stores all the data generated by the system in the local host.

**manage.py file**: it was used to perform basic operations like creating a django project, django application, running the localhost server etc.

**Justification of Programming Language Used**

The python programming language is the major programming language used for the project. It was used for the main data science procedures because:

->Due to the complexity of data science, a lot of help is needed to simplify its process and python offers that help in form of rich in built libraries, tools, packages and frameworks

->it is very popular and widely accepted as the general language for data science. Hence data science project written in python is highly scalable

->Python has simple syntax which makes it easy for anyone to pick up. These are owned to the fact that python is free and open-source language, high level programming, python is interpreted(an interpreter executes the statements of code "one-by-one")

->Python is a unique language that has powerful growth and opens multiple career opportunities for Data Scientists.

Django(Python web framework) was also used for the deployment because it's a framework that enables rapid development of secure and maintainable products. it brings with itself the simplicity of utilising the syntax structure of Python and enables developers to build meaningful and viable web applications effortlessly. Products built in django are also highly scalable.

Other tools used are CSS, HTML and bootstrap which has all been explained the chapter 3

**System Requirement**

This is both hardware and software configuration needed by a system to run a particular program or application effectively and smoothly. To run this project on the localhost server, the following system hardware requirement is needed:

->The system must have a minimum Processor of 1.8 GHz but for better performance faster processors is highly recommended.

->The system must have a minimum of 800MB available hard disk space.

->Works on both 32-bit and 64-bit operating systems but 64-bit operating systems are highly recommended.

->Minimum Installed RAM of 2GB, 8GB recommended

->Video card that supports a minimum display resolution of 720p (1280 by 720)

The software requirement include:

->A minimum of windows 7 but windows 10 is recommended. It also works on mac and other system versions.

->The following applications must be installed/available on the system: Visual Studio Code, Command Prompt, Git, Anaconda to run Jupyter Notebook.(The entire system requirement varies slightly according to the version of these applications installed).

->The following tools and libraries must be installed on the system: Python 3.0 upward, django, sklearn, numpy, pandas, colorama, shap, joblib and other basic python libraries which mostly comes together with the installation of python and anaconda.

When the advertisement system is completely integrated to any ecommerce platform, the requirements will be an internet connection, a browser and a link that gives one access to the platform.

**CHAPTER FIVE**

**Summary**

This project is a demonstration of how any company(e-commerce stores) can get the most with advertisements by showing them only the products that best suit them as a result of making predictions from their past details and demographics. with this type of advertisement, users can never receive advertisements not related to their interest which is an effective method of advertisement.

When the system is completely integrated, the system will be a feature that will be available to end users. Then the buy and cart button will now do their real job of directing the user to a payment platform of adding the store to their profile as carts.

Some of the advantages of this system are as follows:

->Users gets directed to the exact page where they will get the product they like best without having to go through the catalogue of large number of products available in the store

->It doesn't interrupt user session or causes any disturbance

->The returned advertisement will always be relevant to the user because it's a reflection of their personal details.

->This product will also save time and also yield to the growth of the store since users are able to derive the satisfaction and convenience they want.

->It is free of cost once it is completely integrated into any ecommerce store. Then the buy and cart button on the advertised product can now direct to a payment gateway where payment can be made and the cart button also performs its function as it should.

->It is very easy to use.

**System Maintenance**

System maintenance is the process of modifying and updating a system to keep up with its purpose of creation and customer needs.

The system maintenance process will be done by updating the dataset by collecting more data from the user as they use the platform(s) more. This will enable the machine learning model to make more accurate predictions as it is able to gain much more experience from the platform users.

The model will also undergo frequent and consistent validation to make sure the performance doesn't deteriorate. Backdoors will also be checked for. Backdoor attacks are usually hidden patterns in a dataset which will make the system perform in a way different from the purpose of its creation whenever its corresponding action is triggered. This is usually used by attackers to corrupt the system.

All data generated will also be kept secured and backed up. This backups will be done at regular intervals.

The user interface will also be improved to be more user friendly and attract more users.

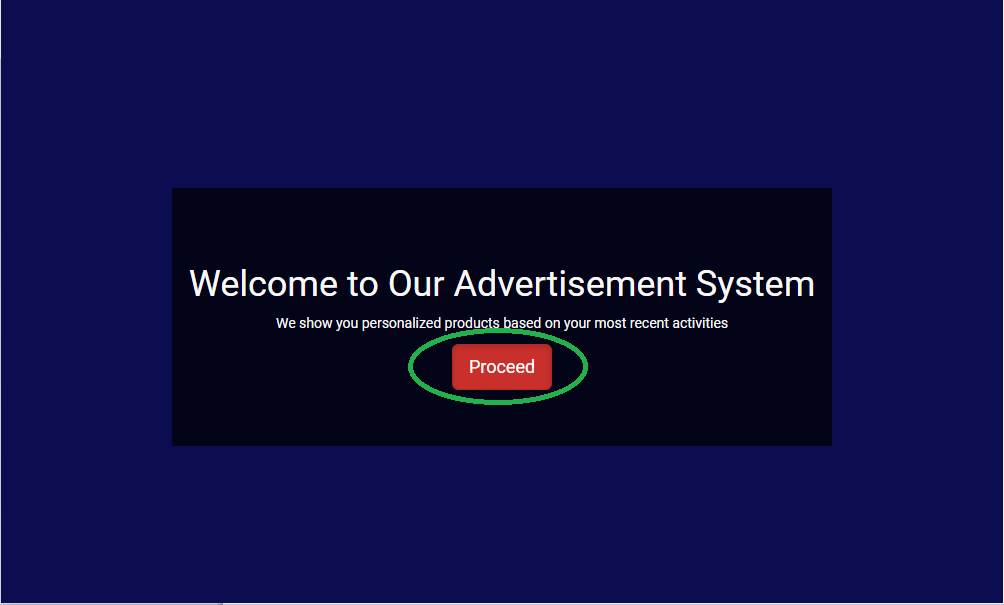
Users will also be able to provide their feedback which will be integrated into the later versions.

**Suggested Area for Further Studies**

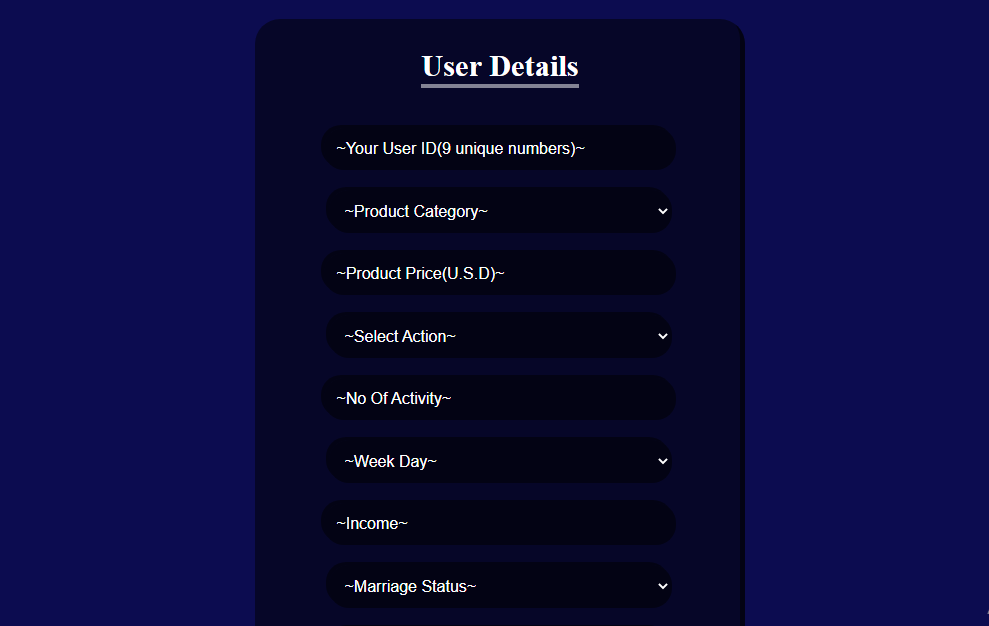
**User Manual**

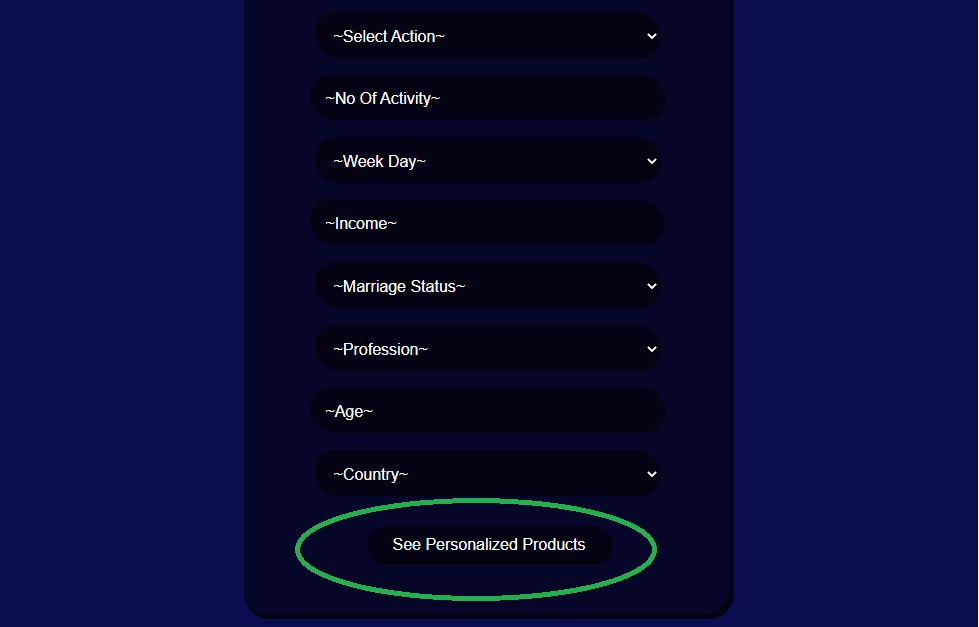
The section explains how users can navigate the system

The image below is the landing page of the system



The green circled area when clicked by the user will be directed to a form page where the requested information must be filled accordingly. image below





After filling the form, the user must click on the green circled area to see personalised products LIke the image below.

