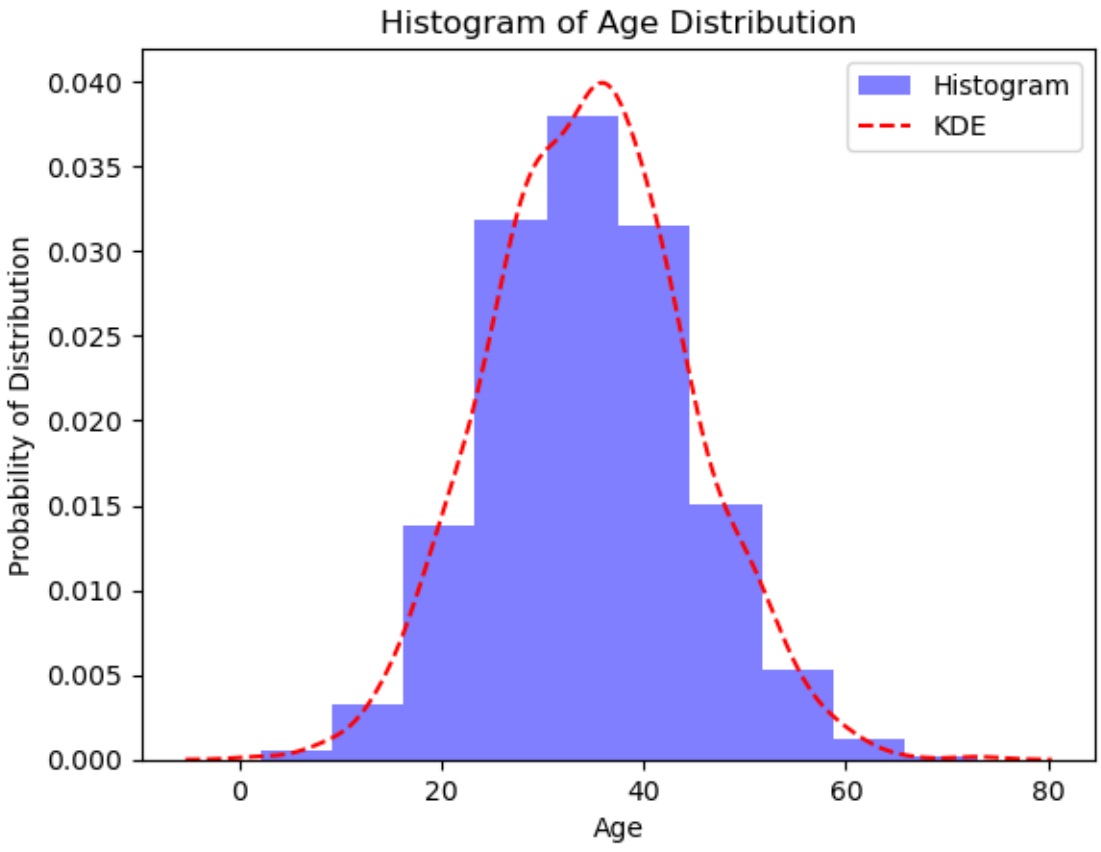
Final Project: Customer Segmentation and Behavior Prediction

Benjamin Nicholson
Seton Hill University



Introduction

With the exponential growth in computational power, cutting-edge techniques leveraging neural networks have emerged as the cornerstone for customer segmentation, revolutionizing the landscape of data-driven decision-making. This project is going to dig deep into clustering, neural networks, image processing and other tools that are utilised in data science.

Understanding the database

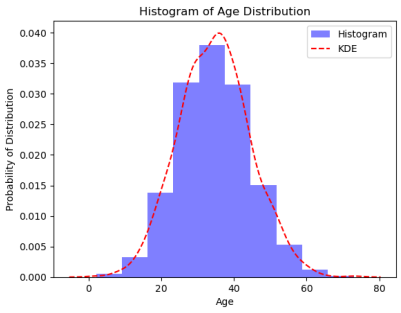
There are 1000 instances with 12 different parameters

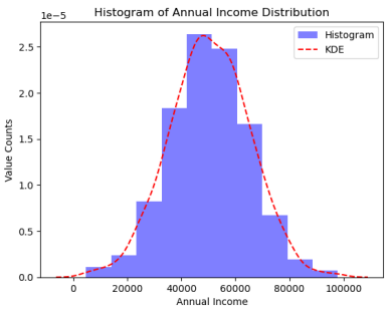
- Customer_ID: Unique identifier for the customer
- Age: Customer's age
- Gender: Customer's gender
- Annual_Income: Annual income of the customer
- Total_Purchases: Total number of purchases made by the customer
- Average_Purchase_Value: Average value of purchases
- Product_Category_Most_Purchased: Category of the most purchased products
- Website_Visits_Last_Month: Number of times the customer visited the website in the last month
- Marketing_Emails_Opened: Number of marketing emails opened by the customer
- Hours_Spent_on_Support_Calls: Total hours spent by the customer on support calls
- Churn: 1 if they are leaving as a customer, and 0 if they stay
- Future Purchases: They bought products in the future

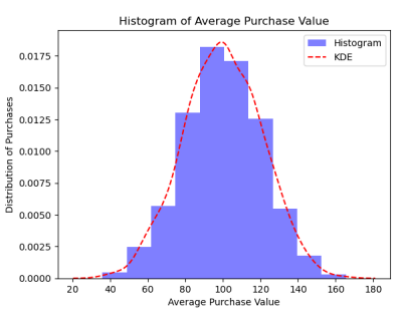
Preprocessing

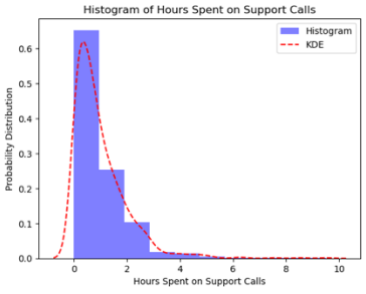
Data: original data that can be manipulated
data_original: original data that will not be touched
data_dummy: the original data that has dummy variables in place of the categories
data_dummy_integer_encoding: dummy variables and integer encoding
data_normalised: original data that has been normalised
data_label_categories: creates labels for groupings of data
percentages_df: percentage values for different spreads of distribution
value_counts = look at the value of distribution for segmented values

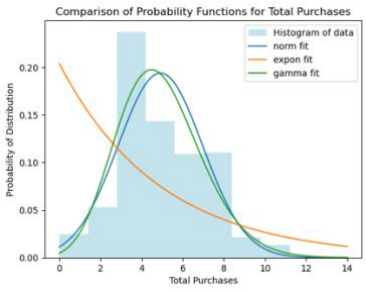
Exploratory Data Analysis





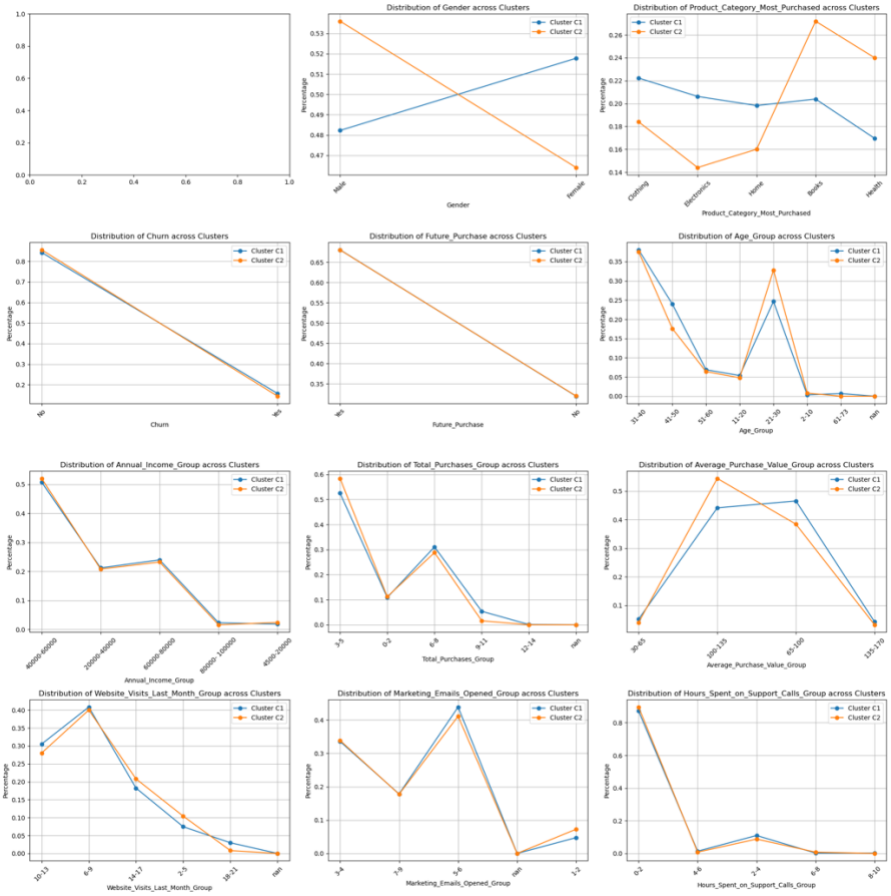




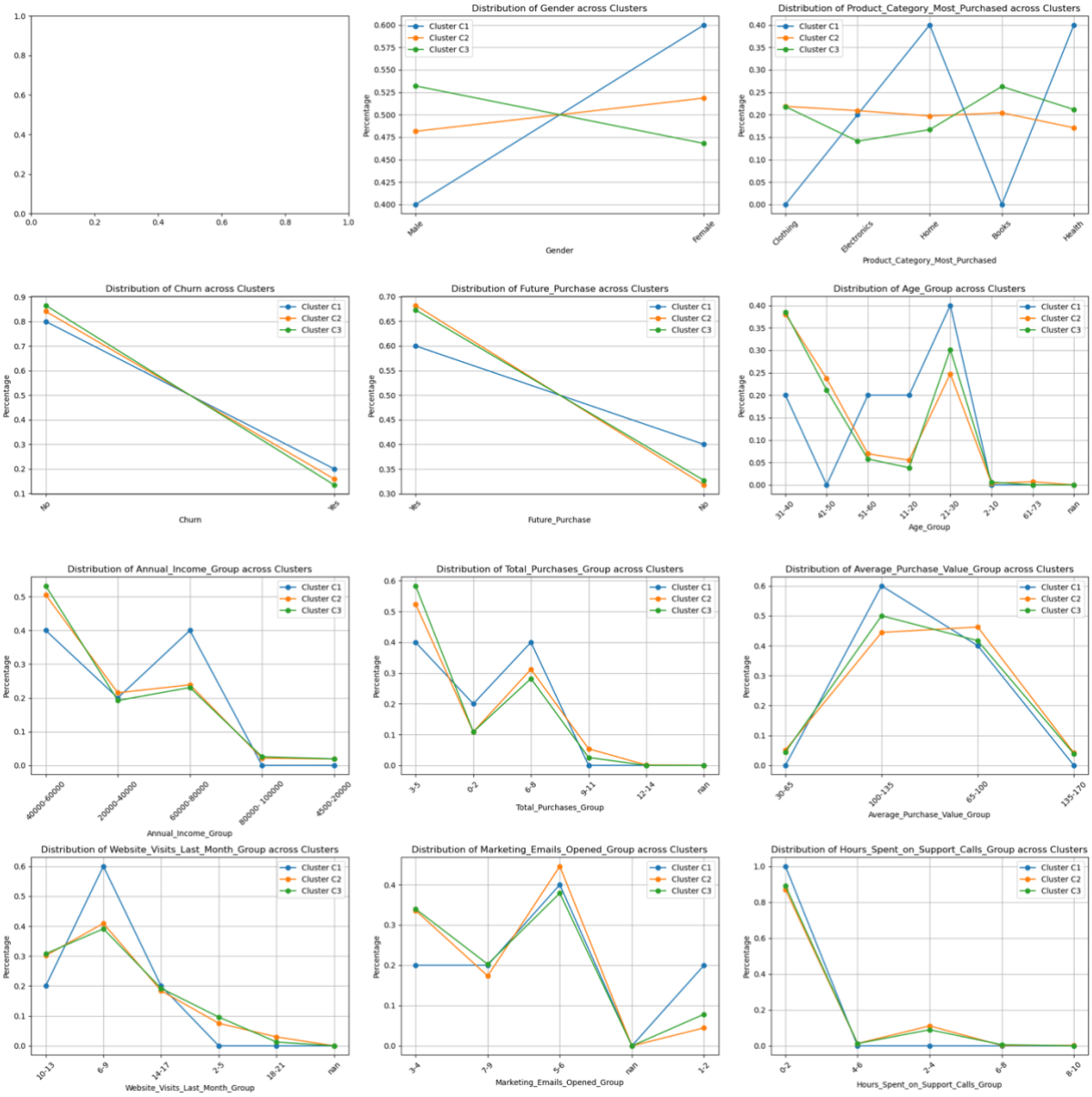


Looking at the spreads of distribution is an important component of understanding clusters. Wherever there are differences in the distribution it is likely that there is a cluster. If they are all normally distributed then clusters are not going to form. Where as if you look at total purchases, hours spent on support calls and the normally distributed functions it is clear that some clusters will be created.

K-Means Clustering – Using integer encoded with dummy variables



Hierachial Clustering – Using normalised dataframe with dummy variables



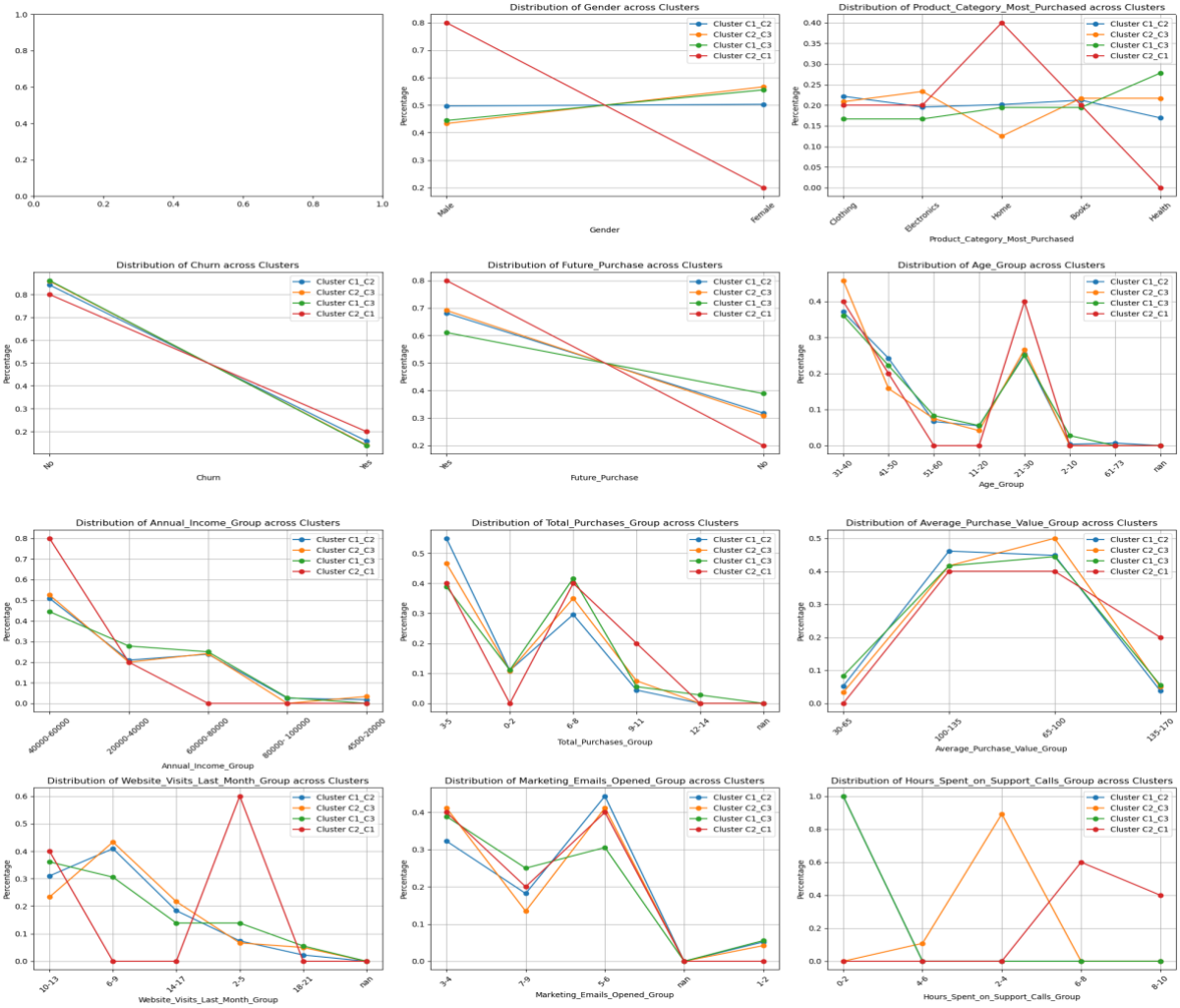
K-Means clustering and Hierarchical clustering have created two separate clusters based on their own different parameters and understanding of creating a cluster.

K-Means:
Gender: More Males in C2
Age: More 21-30 year olds in C2 but less 41-50 years old
Most Purchased: Books & Health for C2
Market to 21-30 year old males, Books and Health items

Hierarchal
Gender: C1 less males than females
Age: C1 had no 41-50 year olds but far greater 31-40,51-60, 11-20 and 21-30 year olds (younger)
Annual Income: No High income earners but had more 60000-80000 (lower income earners)
C1 is lower income, younger women who opened a lot less marketing emails.

Future Purchases: Lower
Hours spent on support call: All C1 clusters was between 0-2 hours
Website Visits: Between 6-9 was the highest category
Products: Home and Health
Target home and health to young men who had lower incomes with home and health products

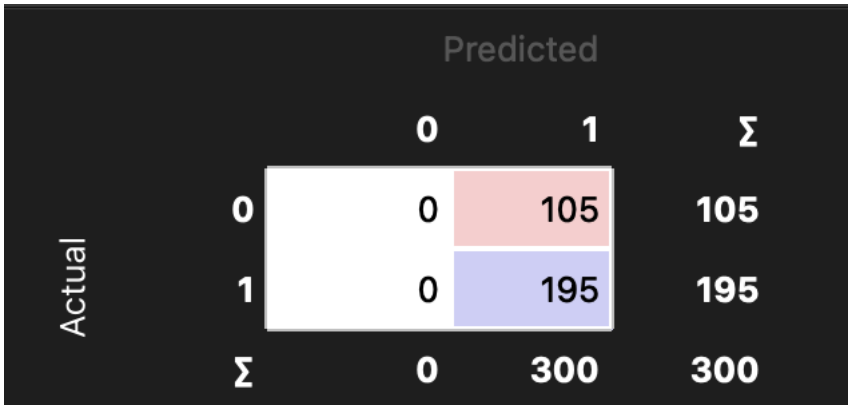
Combined clusters



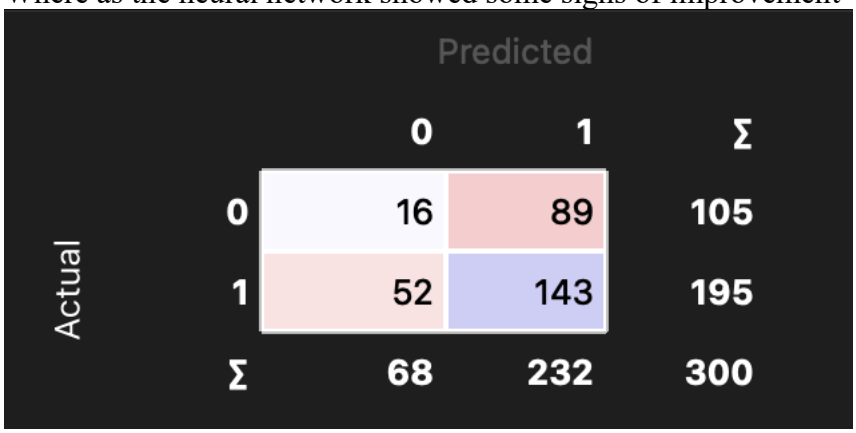
Gender: C2_C1 mainly male
Annual Income: All was between 20,000-60000 (lower income)
Future Purchases: 0.8
Hours Spent on Support Calls: 6-10
Products: Home

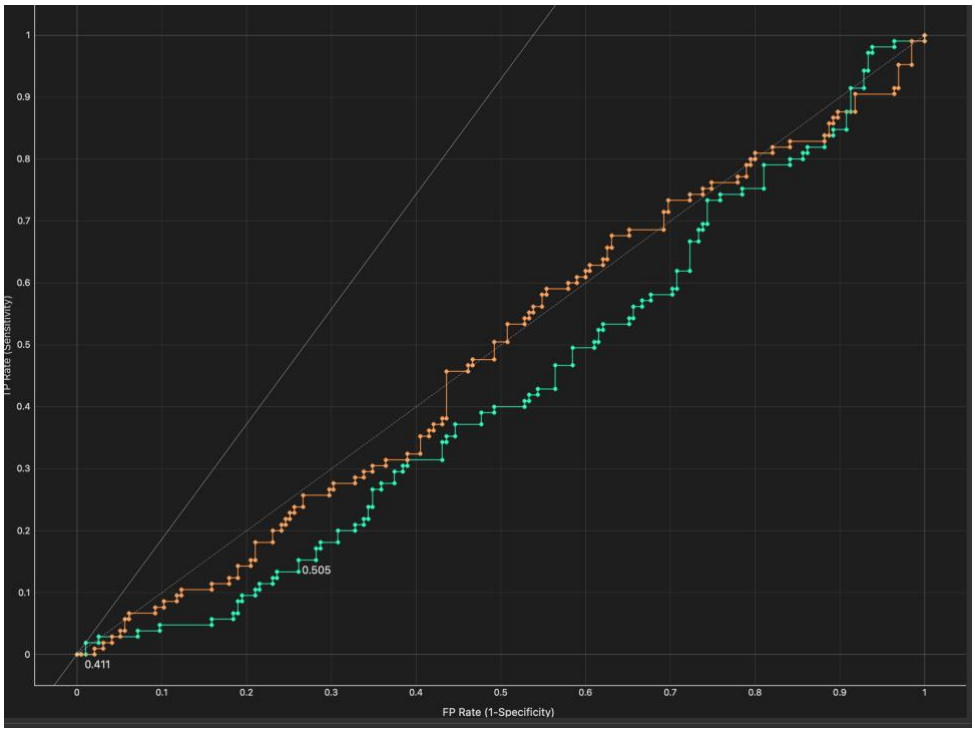
Understanding how the different distribution is spread across these different clusters can help to understand the type of customers and how to market certain products.

The logistic regression for this data due to the lack of patterns essentially was random. There was no way to differentiate the data so the AUC and ROC displayed this. So when trying to predict whether the customer would make a purchase, the logistic regression was unable to make any definitve answers.



Where as the neural network showed some signs of improvement





There was a similar pattern in trying to discover the churn:

- Using binning to categories values into ranges and using the following columns
- Gender
- Age group
- Future purchases
- Website visits last month
- Hours spent on support calls
- Products category most purchased

Using ADAM with 10,000 layers

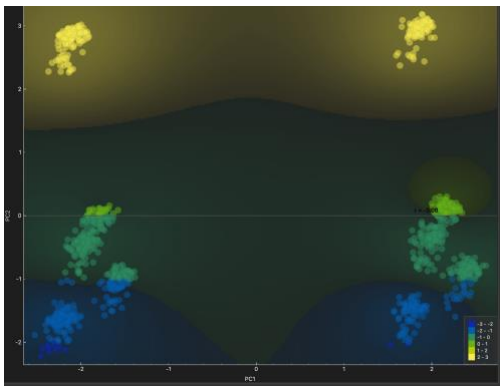
		Predicted		
		No	**Yes**	**Σ**
Actual	**No**	234	23	**257**
	Yes	36	7	**43**
	Σ	**270**	**30**	**300**

Neural Network (3)	0.499	0.803	0.788	0.776	0.803	0.086
Logistic Regression (3)	0.464	0.857	0.791	0.734	0.857	0.000

Neural networks seem to outperform logistic regression likely due to the larger size of the dataset. The number of hidden layers was an important factor in maximising the accuracy of the neural network.

PCA Analysis:

In clustering and customer segmentation, PCA helps by reducing the dimensionality of the data, making it easier to identify patterns and group similar customers together. It does this by transforming the original features into a smaller set of uncorrelated variables called principal components, which retain most of the variance in the data. This simplifies the clustering process and improves interpretability while preserving the essence of the original data.



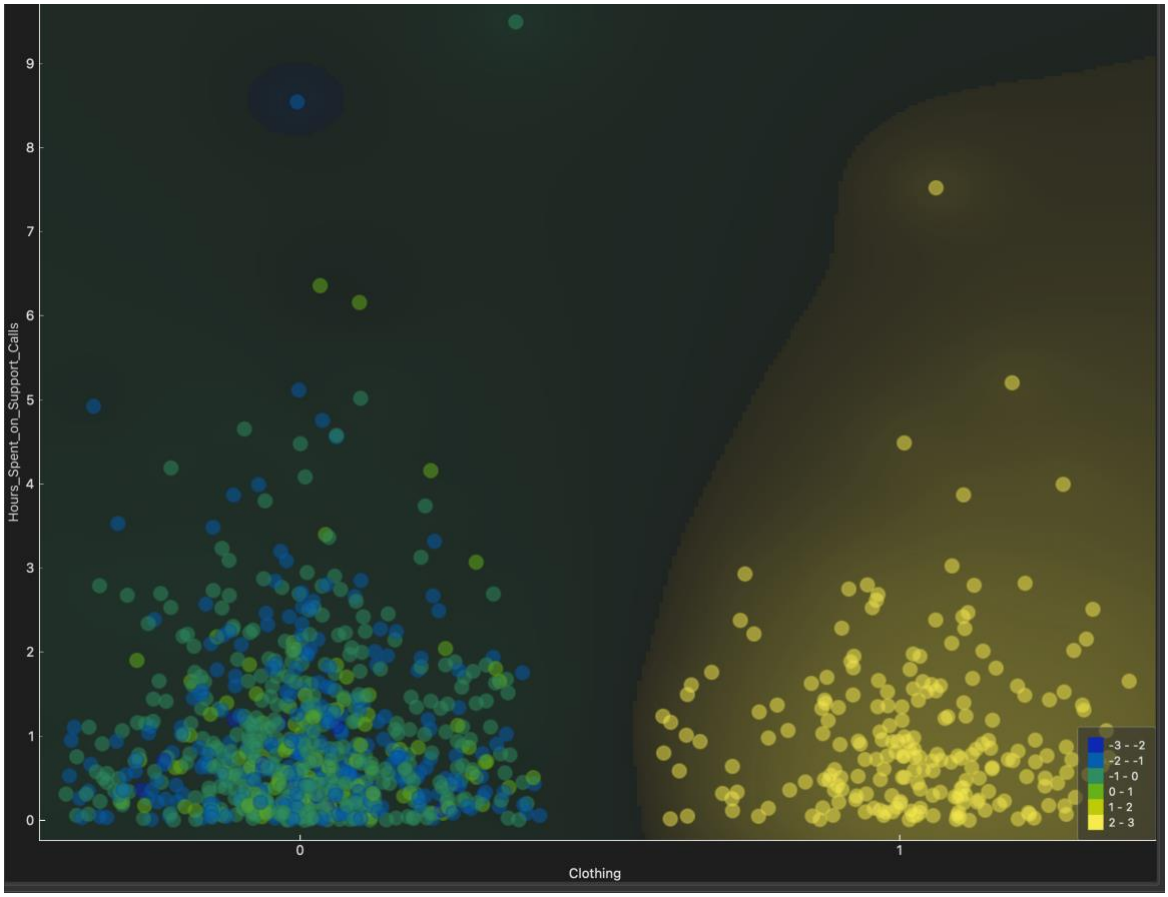
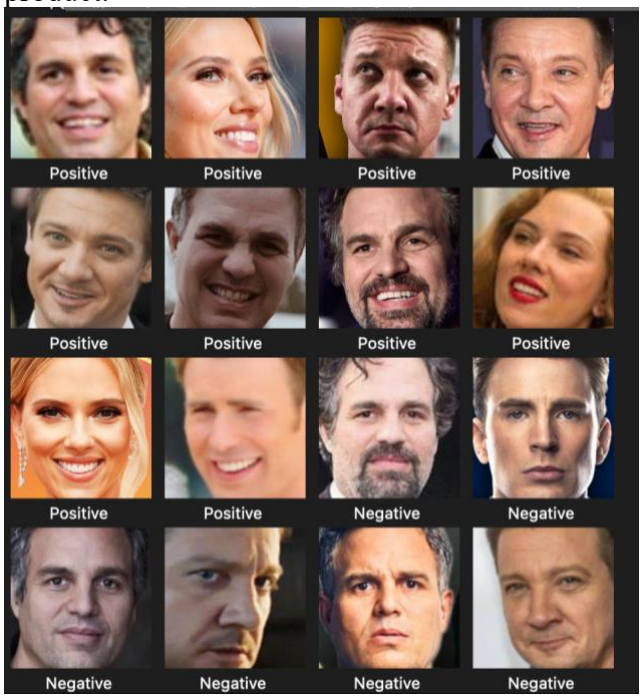
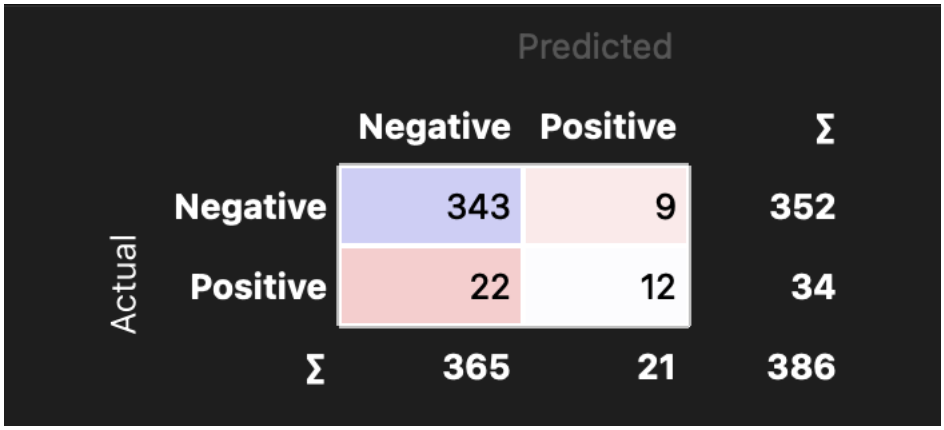


Image Analysis

When looking at customers you want to see happy smiley faces, so creating a neural network that can recognise these faces is important to understand whether a customer likes the product.





This was the result of a sample of around 400 images of people purchasing products.