# CUSTOMER EXPERIENCE & AI TOOLS

Applied Data Science – Data Analytics

18TH MARCH

RODNEY SIBANDA

# Contents

## 1. Abstract

This report details the stages of a data science pipeline applied to the Customer Satisfaction Response to AI, encompassing data acquisition, comprehension, exploration, cleaning, modelling, visualisation, and evaluation. The goal is to utilise multiple machine learning algorithms and assess their performance.

## 2. Data acquisition

The initial step in the pipeline involved obtaining the dataset. The Customer Satisfaction Response to AI data was gathered through an online survey conducted in 2021 and 2022. This dataset was sourced from the Kaggle website under the title "Customer Satisfaction Response to AI."

Through data exploration, it was established that the Product Clothing category is notably popular among AI Tool users. To delve deeper into this trend, a second dataset, titled "Shopping Trends," was sourced from the Kaggle website.

The two datasets were merged to provide deeper insights and uncover more meaningful patterns.

## 3. Data understanding

After acquiring the data, an effort was made to comprehend its structure and content. The dataset columns were described as below in the Kaggle.

The original dataset contains 656 rows and 23 columns, with all columns having data in string format. Below is a description of each column:

Country: The country of the consumer

Age: Age group of the consumer (e.g., Gen X, Gen Z).

Annual_Salary: The consumer's income level (e.g., Low, Medium High, High).

Gender: Gender of the consumer (e.g., Male, Female).

Education: Educational qualification (e.g., University Graduate, Masters' Degree).

Living_Region: Type of area the consumer resides in (e.g., Metropolitan, Rural Areas).

Online_Consumer: Indicates whether the consumer shops online (YES/NO).

Online_Service_Preference: Preference for online services (YES/NO).

AI_Endorsement: Indicates trust in AI technology (YES/NO).

AI_Privacy_No_Trust: Indicates concern about AI and privacy (YES/NO).

AI_Enhance_Experience: Whether AI enhances the shopping experience (YES/NO).

AI_Satisfaction: Satisfaction level with AI (YES/NO).

AI_Tools_Used_Chatbots: Whether the consumer uses chatbots (YES/NO).

AI_Tools_Used_Virtual_Assistant: Use of virtual assistants like Alexa (YES/NO).

AI_Tools_Used_Voice&Photo_Search: Use of AI tools for voice/photo search (YES/NO).

Payment_Method_COD: Use of cash on delivery as a payment method (YES/NO).

Payment_Method_Ewallet: Use of digital wallets for payment (YES/NO).

Payment_Method_Credit/Debit: Whether the consumer uses credit/debit cards for payments (YES/NO)

Product_Category_Appliances: Preference for purchasing appliances (YES/NO).

Product_Category_Electronics: Preference for purchasing electronics (YES/NO).

Product_Category_Groceries: Preference for purchasing groceries (YES/NO).

Product_Category_Personal_Care: Preference for purchasing personal care items (YES/NO).

Product_Category_Clothing: Preference for purchasing clothing (YES/NO).

The original dataset was augmented with data from 'Shopping Trends' using the 'Mapping' label as the primary key. The additional data is shown below.

Item Purchased - The item purchased (Integer)

Category - Category of the item purchased (String)

Purchase Amount - The amount of the purchase (Integer)

Size - Size of the purchased item (String)

Color - Color of the purchased item (String)

Season - Season during which the purchase was made (String)

Review Rating - Rating given by the customer for the purchased item (Float)

Shipping Type - Type of shipping chosen by the customer (String)

Previous Purchases - The total count of transactions concluded by the customer, excluding the ongoing transaction (Integer)

Frequency of Purchases - Frequency at which the customer makes purchases (e.g., Weekly, Fortnightly, Monthly)

# 4. Data cleaning

To explore the dataset, I had to import 'chardet' to determine how the csv file was encoded so that I could load and read the data using visual studio.

To discover more about the dataset using different machine learning algorithms I encoded the Yes/ No values for several features.

## 4.1. Feature Engineering

Three key features have been created:

- 'Number of AI Tools Used,' was introduced, counting the usage of tools like 'AI_Tools_Used_Chatbots,' 'AI_Tools_Used_Virtual_Assistant,' and 'AI_Tools_Used_Voice&Photo_Search' for each row.
- 'Number of Payment Methods Used,' was created to count payment methods such as 'Payment_Method_COD,' 'Payment_Method_Ewallet,' and 'Payment_Method_Credit/Debit' per row.
- 'Frequency of Purchases (Days)' was created to replace 'Frequency of Purchases'. 'Frequency of Purchases' uses string values; those values were converted to integers in 'Frequency of Purchases (Days)'. For example, 'Fortnightly' was converted to 14 days for 'Frequency of Purchases (Days)'.

Two target variables have been created:

- 'Segment' was generated by concatenating the values of 'Gender,' 'Age,' 'Living_Region,' and 'Annual_Salary,' for each row, resulting in 76 unique combinations/ target labels.
- 'Mapping,' was formed by combining the values of 'Gender' and 'Age,' for each row leading to 9 distinct patterns/ target labels. These labels are frequently used for visualisation.

## 4.2. Resampling

The model performance was poor because the dataset was not large enough.

Data resampling was utilised to generate a bootstrapped dataset, increasing its size to 4 times the original dataset.

This significantly improved the performance of all the models used.

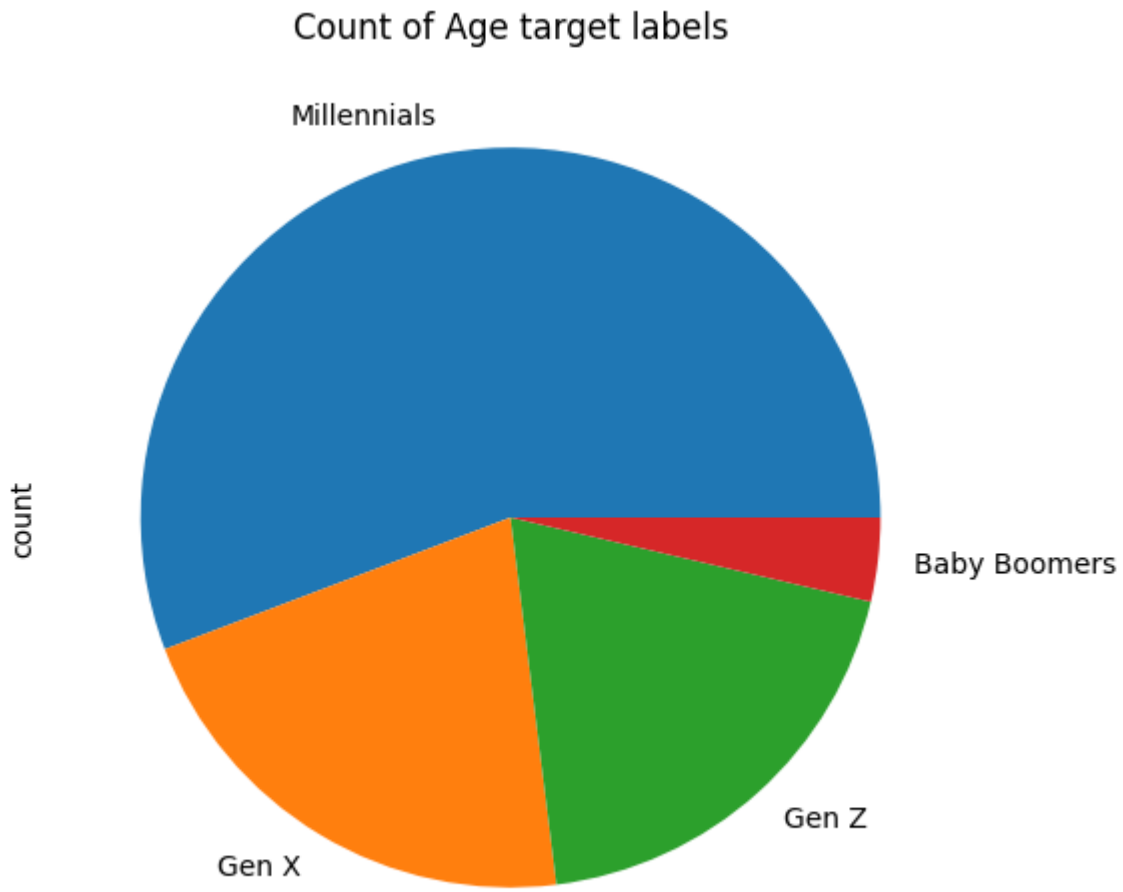
## 5. Data exploration

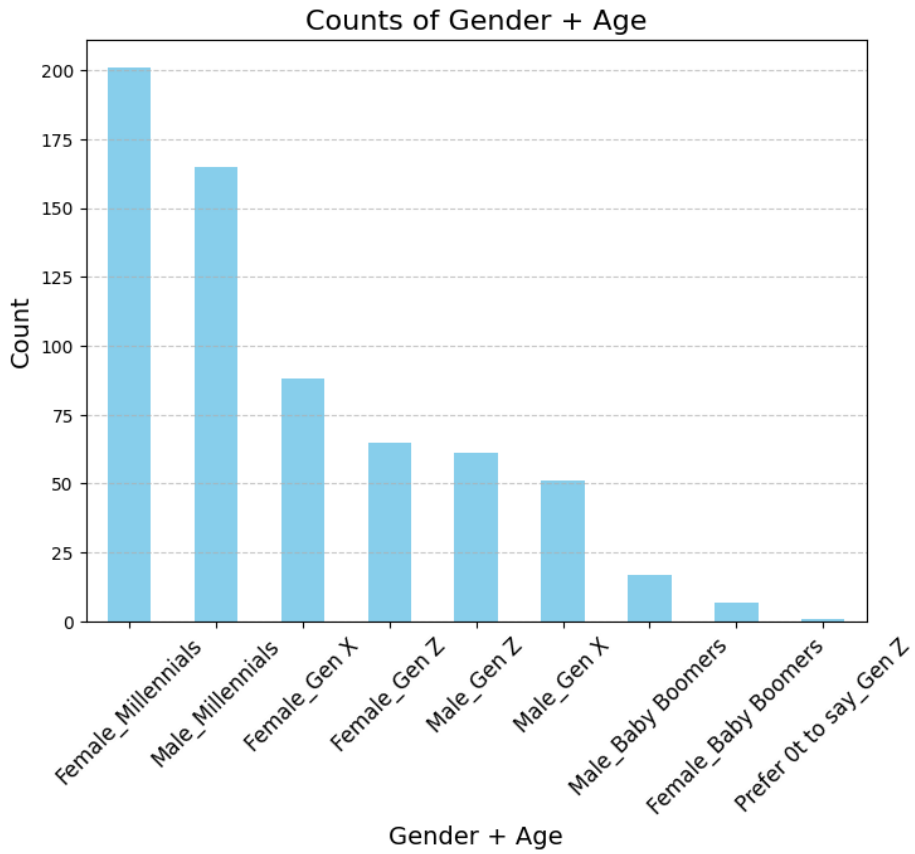
In the initial stages of data exploration, the dataset was segmented based on age and gender. This division facilitated a more detailed analysis of patterns and trends within these demographic groups, providing foundational insights for further exploration.

### 5.1 Understanding the demographics in the dataset

The Customer Satisfaction Response to AI uses four distinct age groups.

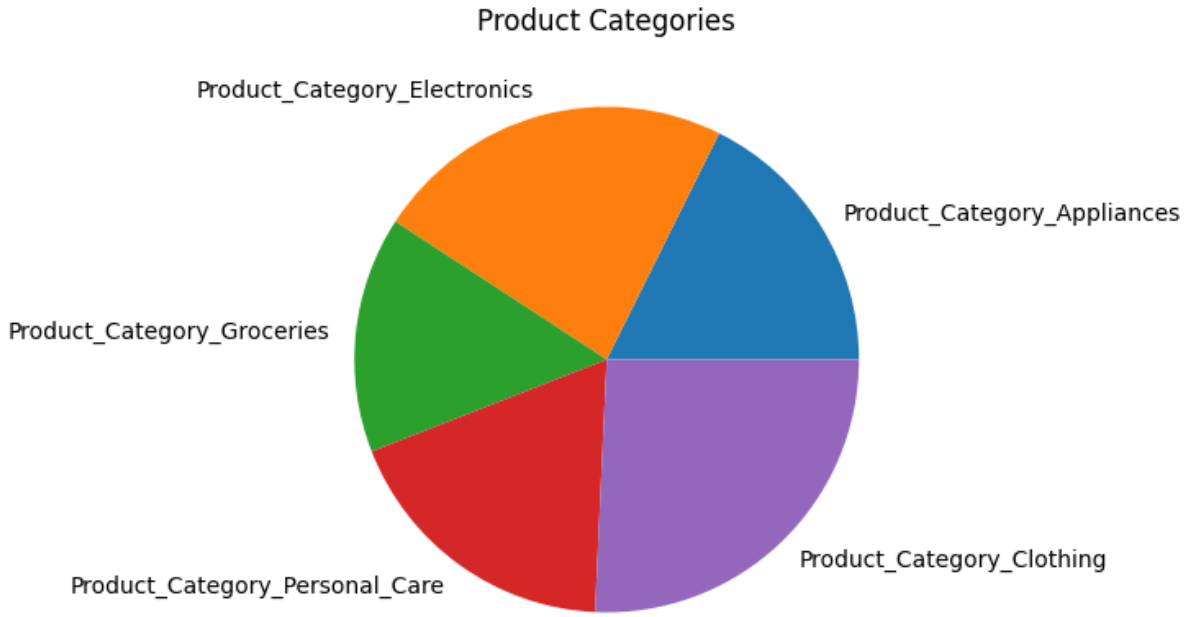
- **Baby Boomers** (born between 1946 and 1964)
- **Generation X** (born between 1965 and 1980)
- **Millennials** (born between 1981 and 1996)
- **Generation Z** (born between 1997 and 2012)






Female Millennials are dominant in this dataset.

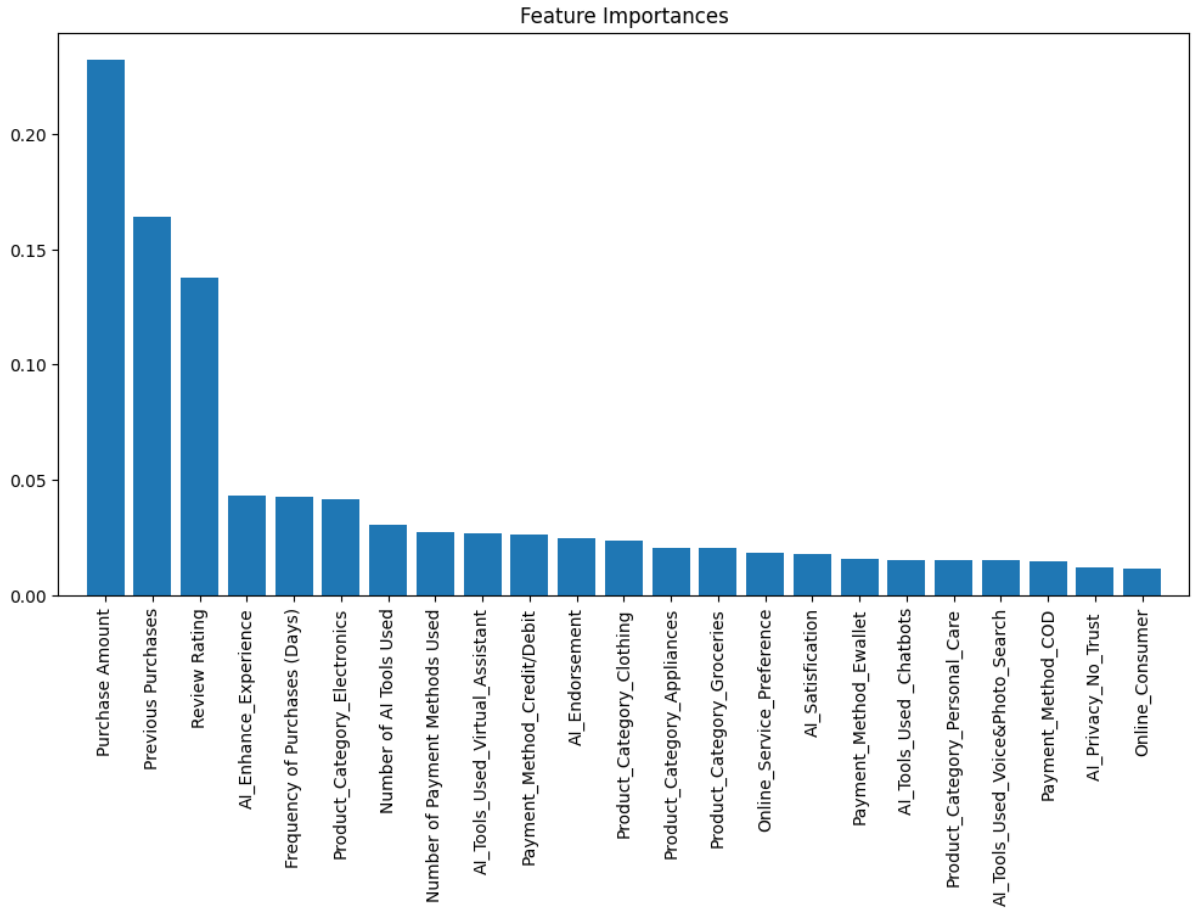
Clothing emerged as the most frequently utilised category for AI Tool Users in this sample.



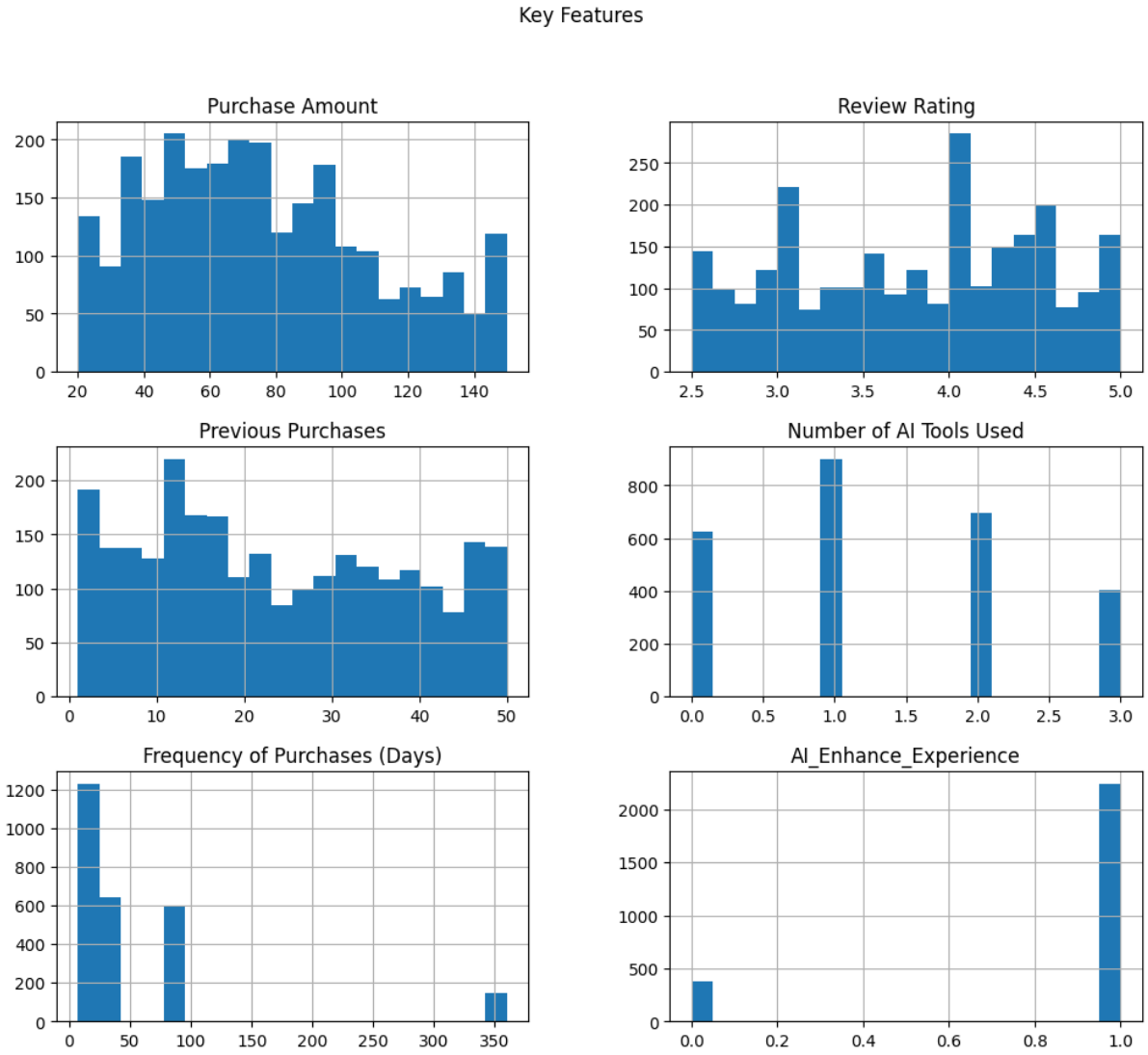
## 5.2 Understanding Feature Importance

After augmenting the original dataset with the Clothing Category Shopping Trends data there are 37 columns, 23 potential feature attributes are available (attributes using Boolean, floats, or integers).

Visualisations were employed to identify the key features for predicting the ‘Segment’ label for a new data record.



‘Purchase Amount’, ‘Previous Purchases, ‘Review Rating’, ‘AI Enhances Experience’ & ‘Frequency of Purchase (Days)’ are the key features for identifying ‘Segment’ & ‘Mapping’ class labels.



# 6. Data modelling and visualisation

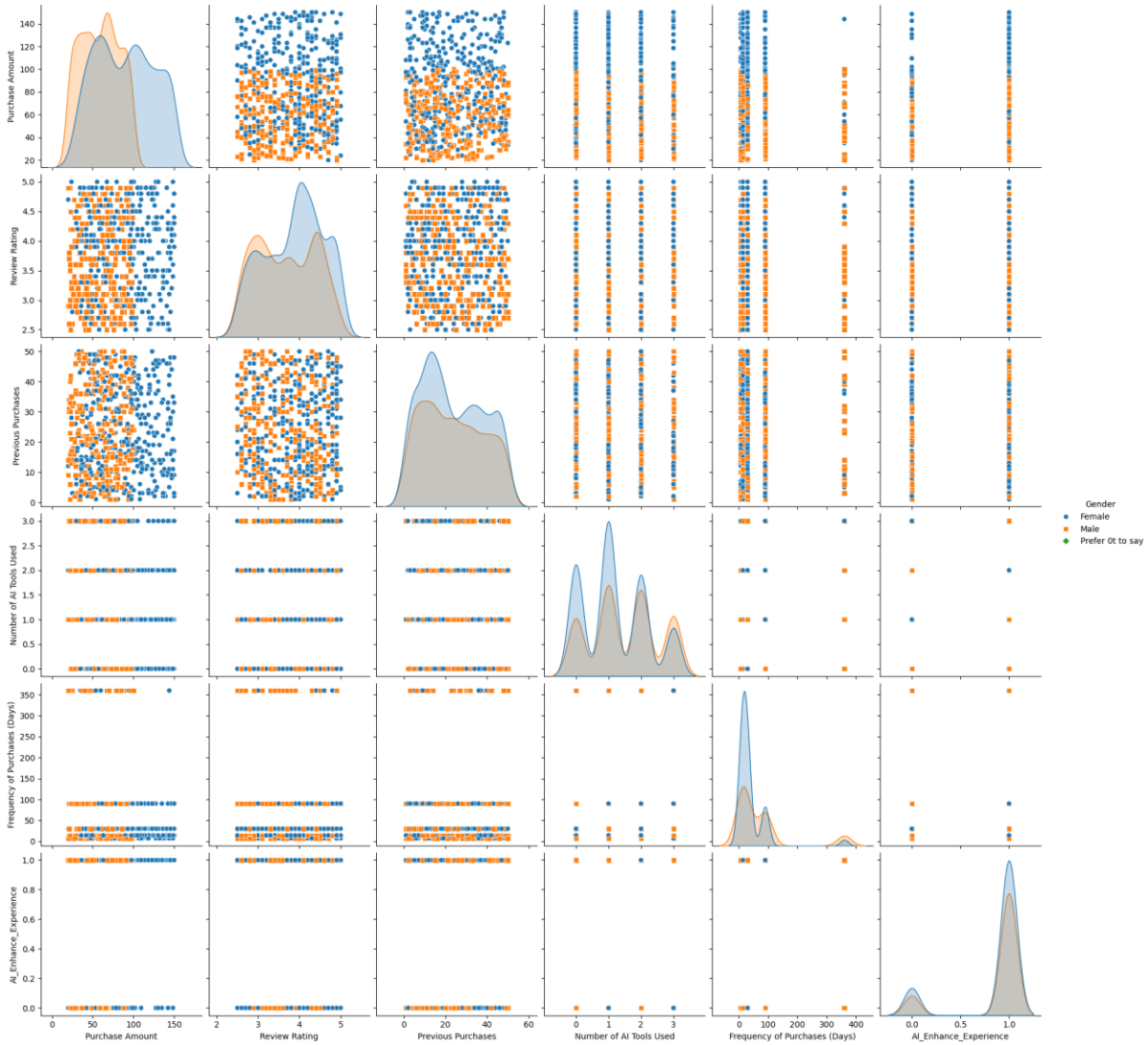
We start by looking at pair plots using ‘’Gender’ class labels and the most important features (‘Purchase Amount’, ’ Review Rating’, ‘Previous Purchases’, ‘Number of AI Tools’, ‘Frequency of Purchases (Days)’ & ‘AI Enhances Experiences’).

The next step is to drill into the correlations between variables/ features we see in the pair plots. We do this using ‘Mapping’ class labels or ‘Segment’ class labels.

## 6.1 Pair plot with ‘Gender’ class labels.

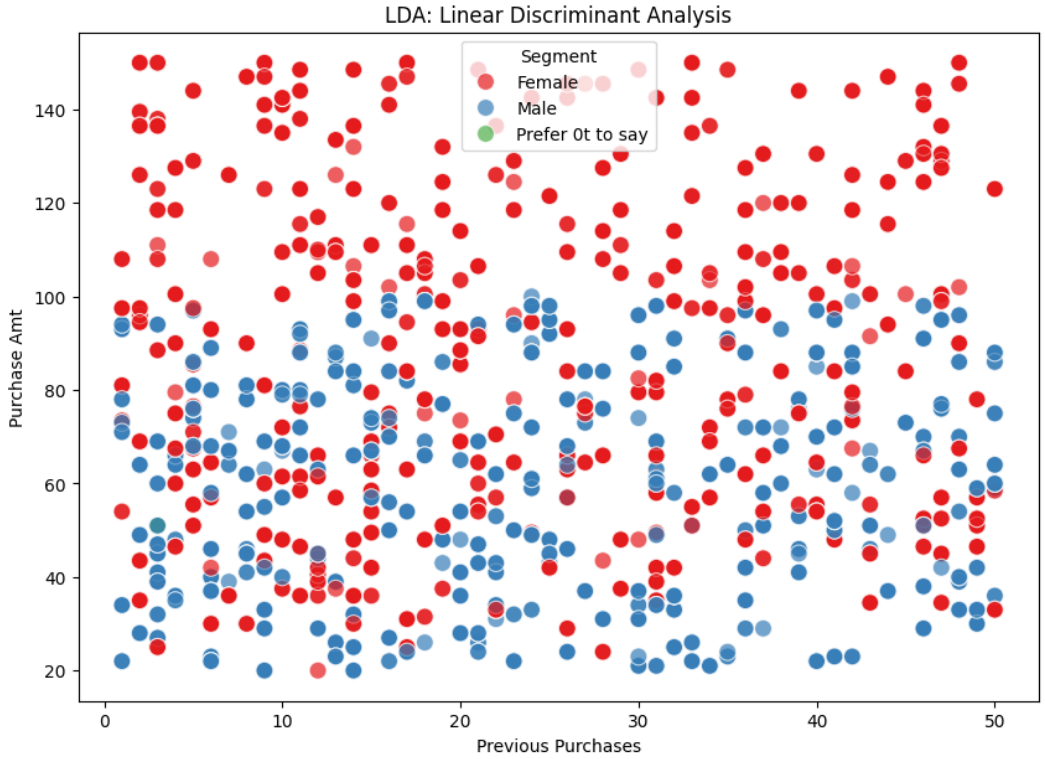
‘Gender’ class labels are used for this pair plot. I have only used the key features for this visualisation. Both men & women had an overwhelmingly positive view about AI Tools enhancing the shopping experience.

Men typically have longer intervals between purchases than women.



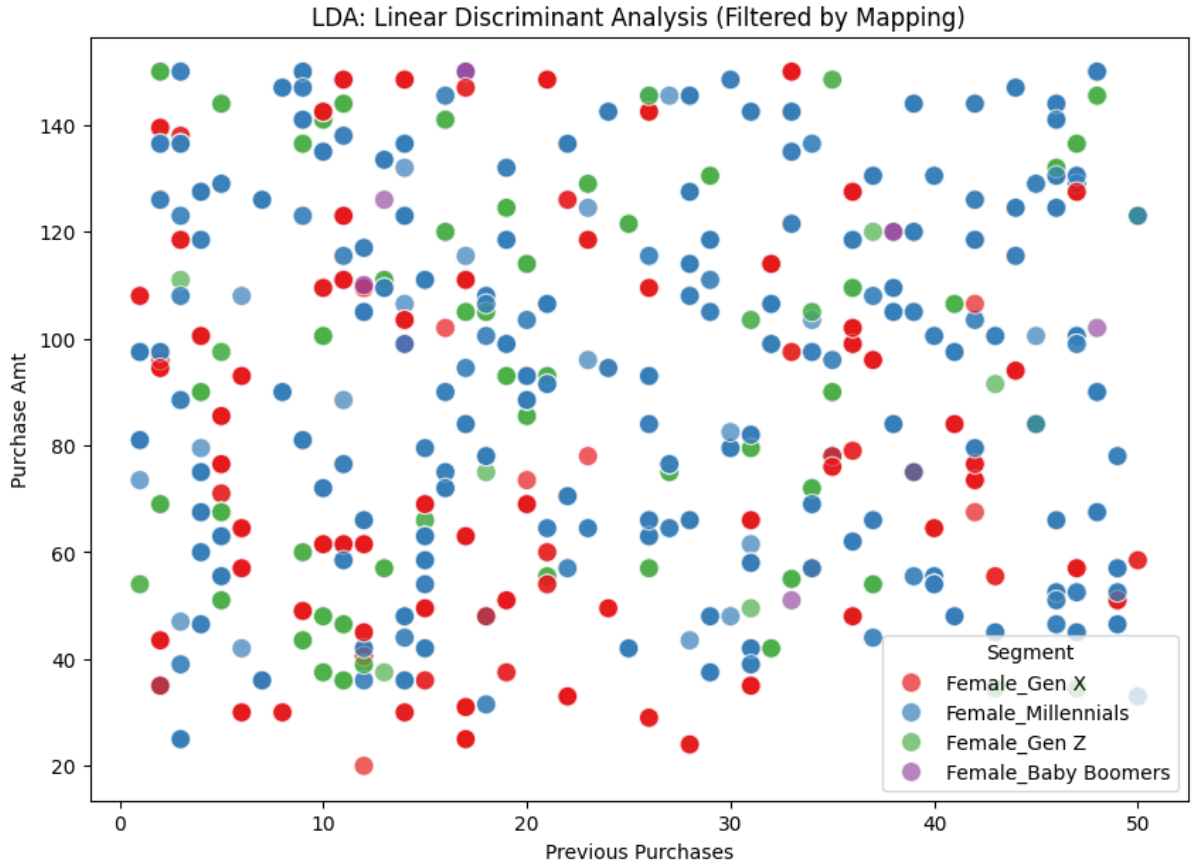
## 6.2 Scatter plot with ‘Gender’ class labels.

Women spend more than men. ‘Previous Purchases’ trends are the similar for both men and women.



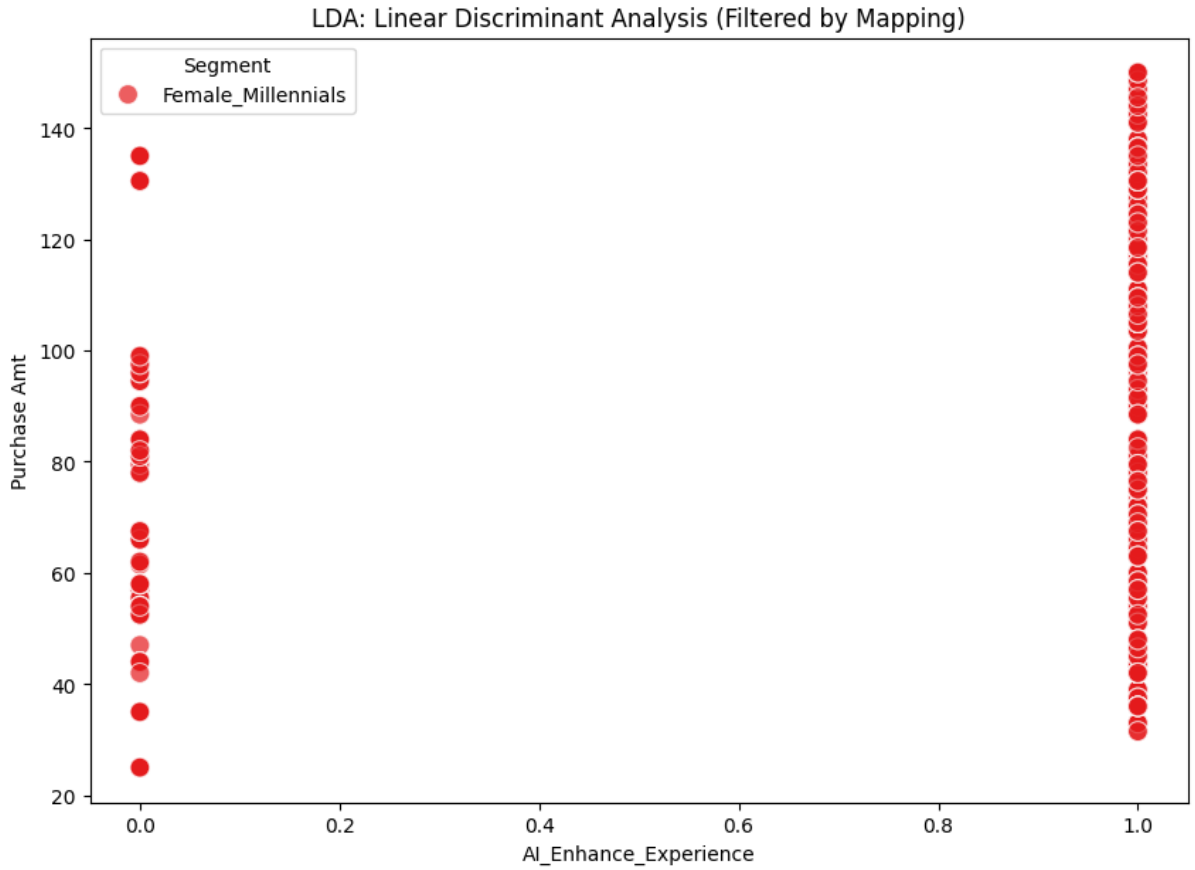
## 6.3 Scatter plot with ‘Mapping’ class labels.

There is no material difference in how much women spend based on age.

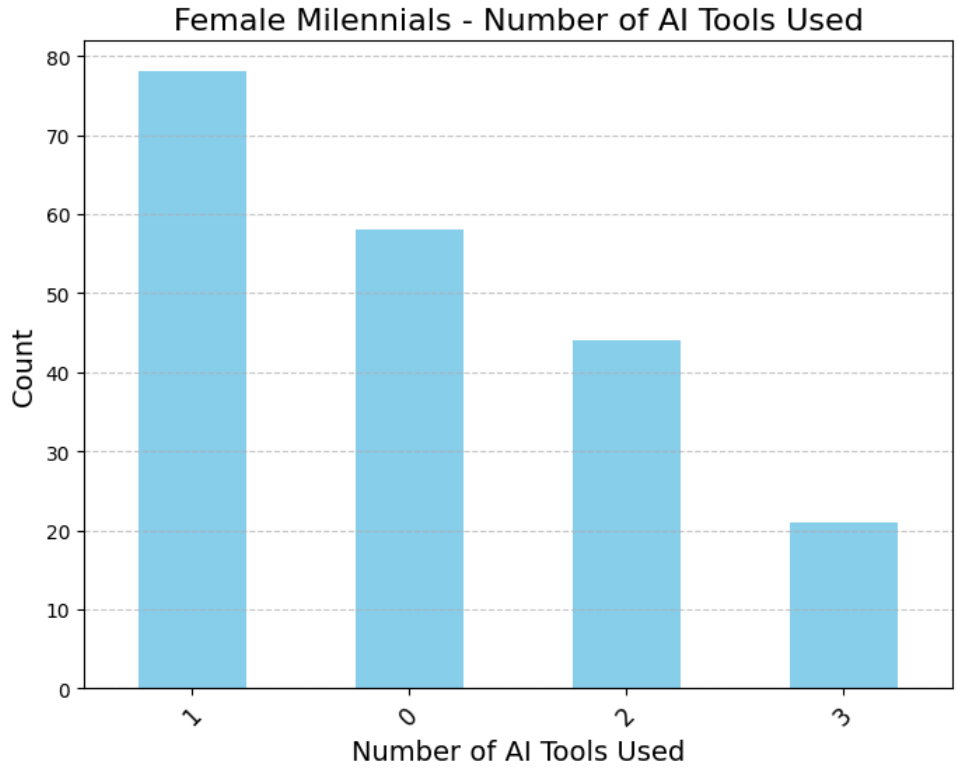


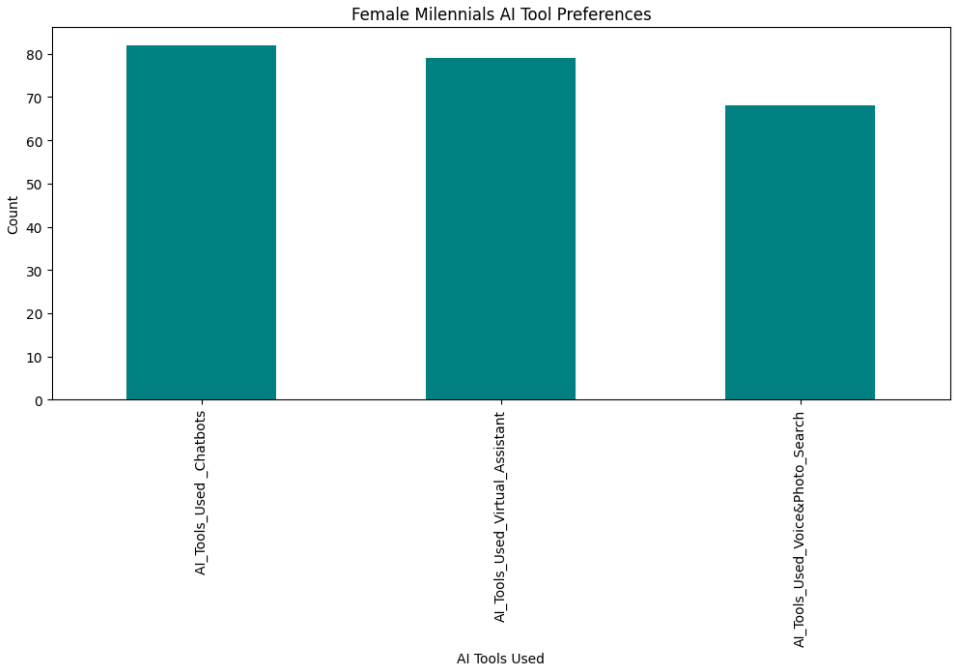
## 6.4 Deep dive into ‘Female Millennials’

Significantly more Female Millennials think AI enhances their shopping experience. Female Millennials, with a positive view of AI Tools, spend significantly more than Female Millennials that do not.



Female Millennials prefer to use a single AI Tool, but there is no clear preference for any single AI Tool.



# 7. Model Evaluation

Six models were evaluated to determine which one would most accurately label a new data record, the label was ‘Segment’.

## 7.1 Cross-validation of different models

The models used are:

- Decision Tree Classifier
- KNeighbors Classifier
- Linear Discrimination Analysis
- XGBoost
- GaussianNB
- Logistic Regression

I used cross-validation to evaluate each of these models.

The least populated class in ‘Segment’ has only 2 members, so the ‘n_split’ for each model evaluations is == 2.

XGBoost & the Decision Tree Classifier were the best performing models with mean test scores of +0.89.

KNeighbors Classifier had a mean test score of +0.68.

Linear Discrimination Analysis, GaussianNB, and Logistic Regression all performed poorly with mean test scores below 0.14.

Decision Tree Classifier is the preference. Decision Tree Classifier does not require the target variable to be encoded, XGBoost does.

# 8. Conclusion

## 8.1. Overview

There is a correlation between Purchase Amounts, customers perception of AI and specific 'Mapping' classes.

Those correlations enable specific models to accurately predict class labels for new data records.

Identifying those labels enables retailers to tailor their marketing strategies to individual preferences.

## 8.2. Improving model performance.

It is necessary to enhance the models' performance by focusing on class labels containing fewer than two members. The viable options include:

- Remove class labels with <2 members.
- Increase the size of the synthetic/ bootstrapped data.
- Use synthetic/ bootstrapped data to increase class labels with <2 members to >5 members.

Resolving this issue will allow for more 'n_splits', which should improve the performance of the models.

## 8.3. Is the dataset/ sample representative?

Are any of the demographic groups under or over-represented? The decision to use Female Millennials as a ball-point is dependent on whether this dataset reflects wider societal dynamics.