INTRO

* N Brown has provided a dataset containing valuable information, including discount values, units sold per day, product numbers, and product department descriptions. The time range of this information is between January 2021 – February 2023
* The objective of this project is to pre-process the data, conduct exploratory data analysis (EDA), perform statistical analysis, and apply data modelling techniques to address the selected research questions.
* Through this process, we aim to uncover insights into the relationship between discount levels and units sold.

RESEARCH QUESTIONS

What **factors affect elasticity**? Does price elasticity change periodically?

**Optimal discount** percentage to **maximize revenue** ?

What will happen if we **remove the discounts all together**? Will sales drop?

Do number of orders differ due to **seasonality factors**?

What departments have the **highest/lowest** **elasticity**?

**Suggestions** for N Brown, based on our analysis?

APPROACH

We will break down the problem in four steps:

1. Data preprocessing (General)
2. EDA
3. Statistical analysis
4. Modelling.

PREPROCESSING

We have both datasets, we will explore the upper\_data set over the lower\_dataset for now. If we have more time we will explore the lower\_dataset.

**General approach & Repetition in prices**

* Handle missing data, Impute missing values for 'discount' and 'orders' with their respective median.
* In the dataset there were cases where the same product was sold at a different price on the same day.
* Separated the data set into two data sets. One with the upper dataset and one with the lower dataset. We did not take the mean of the values as there were significant differences in price for the same product.
* For now, we will only consider the data set with the higher prices. If there is time we will consider the data set with the lower prices.

**Identifying Imbalance:**

* Significant number of products displayed extreme discount behaviors.
* Products with no discounts(extreme right), products discounted every day (extreme left) lead to data bias
* 368 products displaying such extreme behavior were removed
* "To ensure a robust analysis, we first identified and addressed the imbalance in our dataset caused by extreme discount behaviors. Our initial dataset comprised 4662 products, many of which exhibited extreme patterns. We systematically cleaned the data in three steps. First, we removed products that were never discounted, reducing our dataset to 4416 items. Next, we eliminated those with only one discount slab, which brought us down to 4294. Finally, we focused on products that had discounts applied at least 40-60%

**Outliers Removal**

* Custom outlier removal using percentiles for both the upper and lower datasets.
* Removed the top 1st and 99th percentile of data for each dataset.
* This removed less than 1% of the data. We do not want to remove too much data unnecessarily as we may lose valuable information
* This was done more as a convention (as outlier removal is commonly used in data pre-processing)
* We will filter the data, more meaningfully, later. Did not want to remove too much data as we may lose valuable information about certain products etc.

EDA SLIDE ONE

**Total Orders: Discounted vs Non-Discounted**

*Non- Discounted products show less variability over months, unlike discounted orders which display high variability.*

**Distribution of Discounts (All-Time)**

*Most discount ranges between 20%-40%*

**Departments by Total Orders**

*Top two departments with total orders, were men's branded clothing, and men's own brand tops*

EDA SLIDE TWO

Variance in Seasonality

* The number of items sold for products varies across months suggesting seasonality has an influence on the number of orders sold. We need to consider this as part of the modelling stage. Such as using summer, holiday, weekend, winter.
* For example: MENS OWN BRAND FOOTWEAR sees an increase in orders sold for the between 2022-11 and 2023-01 period which could be attributed to Christmas/new year.

LITERATURE REVIEW

1. Price reductions can lead to increased sales volumes, but their long-term effects depend on how the price reduction aligns with the perceived value of the product. Consumers are more likely to increase their purchase volume if they perceive the discount as significant enough to justify a change in their behavior *(cite 1,4)*
   * *This can be seen in the Distribution of Discounts (all time), the best range was between 20-30%*
2. To understand customer behavior and purchasing decisions, retailers use data, business analytics, and machine learning to analyze customer psychology and shopping patterns. These patterns are influenced by factors such as demographics, purchase history, reactions to past promotions, social influences, and psychological factors (cite 1)
   * Methods we used to do our analysis
3. The success of a retailer's efforts to influence consumer choices depends on a strategic approach to reach shoppers at the optimal time to maximize sales potential *(cite 1, 3)*
   * *Looks like NBrown is able to approach shoppers during optimal peak seasons or seasons where shopping is increased overall, which is displayed on various graphs that show monthly orders*
4. While high discounts may initially attract consumers, they may also raise concerns about the reasons for such drastic reductions. Consumers might wonder if the product is defective, outdated, or simply of inferior quality*(cite 3)*

*● Something for NBrown to take into consideration, as at least 4-5 different departments which are on the top orders and the average top discounts graphs show that the discount does not look like it is working effectively as those departments have very low total orders*

1. The perceived value of a discount is essential, and a discount can drive purchase volume only if it's seen as significant enough to change consumer behavior. Consumers may base their perception of the discount's significance on various factors, including the magnitude of the discount, the original price of the product, and their personal perception of value *(cite 3)*

*● This is displayed in Departments by Total Orders, and Average Discounts for top 10 departments, although mens branded clothing had the most orders/sales; mens own brand outwear had the largest discount 7-8% more*

ANALYSIS SLIDE ONE

Hierarchical Understanding**[1]**:

Department Level: Gives broad strategic insights about which product categories are most price-sensitive

Product Level: Provides granular understanding of individual item performance

This multi-level approach matches how retail businesses operate - both at category management and individual SKU levels

Time-based Analysis

Monthly: Captures longer-term price sensitivity patterns and seasonal trends

Weekly: Shows immediate price response and short-term consumer behaviour

Different time windows help identify both immediate price reactions and sustained effects

Used standard economic price elasticity formula:

***Price Elasticity of Demand = % Change in Quantity Demanded / % Change in Price [2].***

Applied to monthly sales data by department Results identified MENS OWN BRAND BOTTOMS (22.88) as most elastic department"

ANALYSIS SLIDE TWO

Variance in Elssticity

* The weekly elasticity of products was calculated.
* There also seems to be variance in elasticity across the days for products.
* Currently, our model just considers, week, month, year.
* So, going forward we will implement enhanced feature engineering.

Statistical Analysis

T-test for department: MENS OWN BRAND BOTTOMS

H0: In the MENS OWN BRAND BOTTOMS department, the mean number of orders for products with low discounts is less than or equal to the mean for products with high discounts

H1: In the MENS OWN BRAND BOTTOMS department, the mean number of orders for products with low discounts is greater than the mean for products with high discounts

t-statistic: 25.6775

p-value: 0.00008

Conclusion: Reject H0

Detailed conclusion: There is statistically significant evidence that in the MENS OWN BRAND BOTTOMS department, the mean number of orders for products with lower discounts is greater than for products with higher discounts.

* This observation runs counter to the initial assumption that higher discounts would lead to higher overall orders across all departments. It suggests there may be more nuanced relationships between discount levels and customer demand that need to be investigated further.
* To better understand this dynamic, the analysis should explore how the discount-order relationship varies across the different product departments. This will be an important consideration as part of the modelling process, as the impact of discounts on orders may not be uniform across the business.

MODELLING

1. DATA

Selected the three departments with the highest elasticity

Peak sales periods appear around July-September 2022.

Lower sales volumes are seen around January-March periods in both years.

All departments seem to follow somewhat similar seasonal patterns

1. Feature Enginerering

Created temporal features from date

Generated lag features – to take into consideration variation in over weeks

Rolling statistics e.g. rolling mean discount

1. Data Pre-Processing

Created temporal features from date

Generated lag features – to take into consideration variation in over weeks

Scaled features using StandardScaler

MODEL AND RESULTS

Trained each model on scaled training data.

Special handling for Bayesian Ridge (uncertainty estimation)

CONCLUSIONS

Modelling Results Conclusion

* Highest R² score of 0.817 (81.7% of variance explained) Lowest RMSE (5.85) and MAE (2.99)
* This indicates Gradient Boosting captures both linear and non-linear relationships in the data most effectively.
* The lower error metrics suggest more reliable predictions compared to other models

What’s Next

**Time Series Specific Handling**

* Replace random train-test split with time series cross-validation
* Implement proper temporal validation strategy
* Consider time series specific models
* Add stationarity tests and appropriate transformations

**Enhanced Feature Engineering**

* Add seasonal decomposition components (trend, seasonality, residual)
* Include cyclic encoding for temporal features
* Create price elasticity features
* Add interaction terms between discount and seasonality
* Include business-specific features (promotions, holidays)
* Product specific features.