Price Optimization Analysis

> A good deal is a fair price, found fast.
>
> — Marketplace Motto

Workflow

01 Problem Statement

Identify how sellers can price products competitively while maintaining profitability.

02 Exploratory Data Analysis

Analyse product data to uncover pricing patterns, discount trends, and product competitiveness.

03 Modeling

Build and evaluate machine learning models to predict optimal product prices based on key features.

04 Deployment

Deploy the best-performing model as a user-friendly web app for real-time price prediction.

Project Collaboration

- Project Authors:
  - Hafsa M. Aden
  - Ryan Karimi
  - Rose Muthini
  - Lewis Karanja
  - Harrison Karime
  - Elizabeth Ogutu

Collaboration & Workflow

- **Notion** served as our central hub for organizing tasks. We used it to:
  - Share notes and useful links in one accessible space
  - Document meeting notes and decisions for easy reference. - Track deadlines and assigned responsibilities through task boards.
  - Enable real-time collaboration reducing miscommunication and delays.

Business Understanding

At what price should one sell a product to remain competitive while still making a profit?

Problem Statement

- Pricing is one of the most critical decisions for any seller. While costs, margins, and supplier agreements guide internal pricing, what’s often missing is visibility into how competitors price similar products. Without this context, sellers face two key risks:

  - Overpricing → losing customers to cheaper alternatives
  - Under pricing → cutting into profits unnecessarily.

- New sellers entering marketplaces often struggle with:

  - Price variability – Similar products can vary widely in price (e.g., KSh 499 vs KSh 8,900).
  - Discount strategy – Discounts strongly influence buyers, but the "sweet spot" is unclear.

Stakeholders

New Sellers

- Data-backed guidance on product selection & pricing.

Marketplace (e.g., Jumia)

- Gains from onboarding more sellers and improving customer experience

Business Development

- Use insights to attract and support vendors.

Data Science Team

- Build models
- Deliver insights

End Customers

- Fair prices
- Wider variety

Project Objectives

Support sellers in setting competitive yet profitable prices.

➢Generate data-driven insights from marketplace listings.

➢Benchmark products against competitors for market awareness.

RAID Summary

Risks

➢ Website Structure Changes

- If Jumia changes its webpage layout, the scraper could break, halting data collection.

Assumptions

➢ Internet Accessibility

- Users (especially small business owners) have access to reliable internet to use the online tool.

Issues

➢ Limited user feedback

➢

- few respondents to the questionnaire, reducing insight into real-world tool usability.

Dependencies

➢ This project depends on several external factors:

- Jumia
- Hosting Platform(Render)
- Tableau

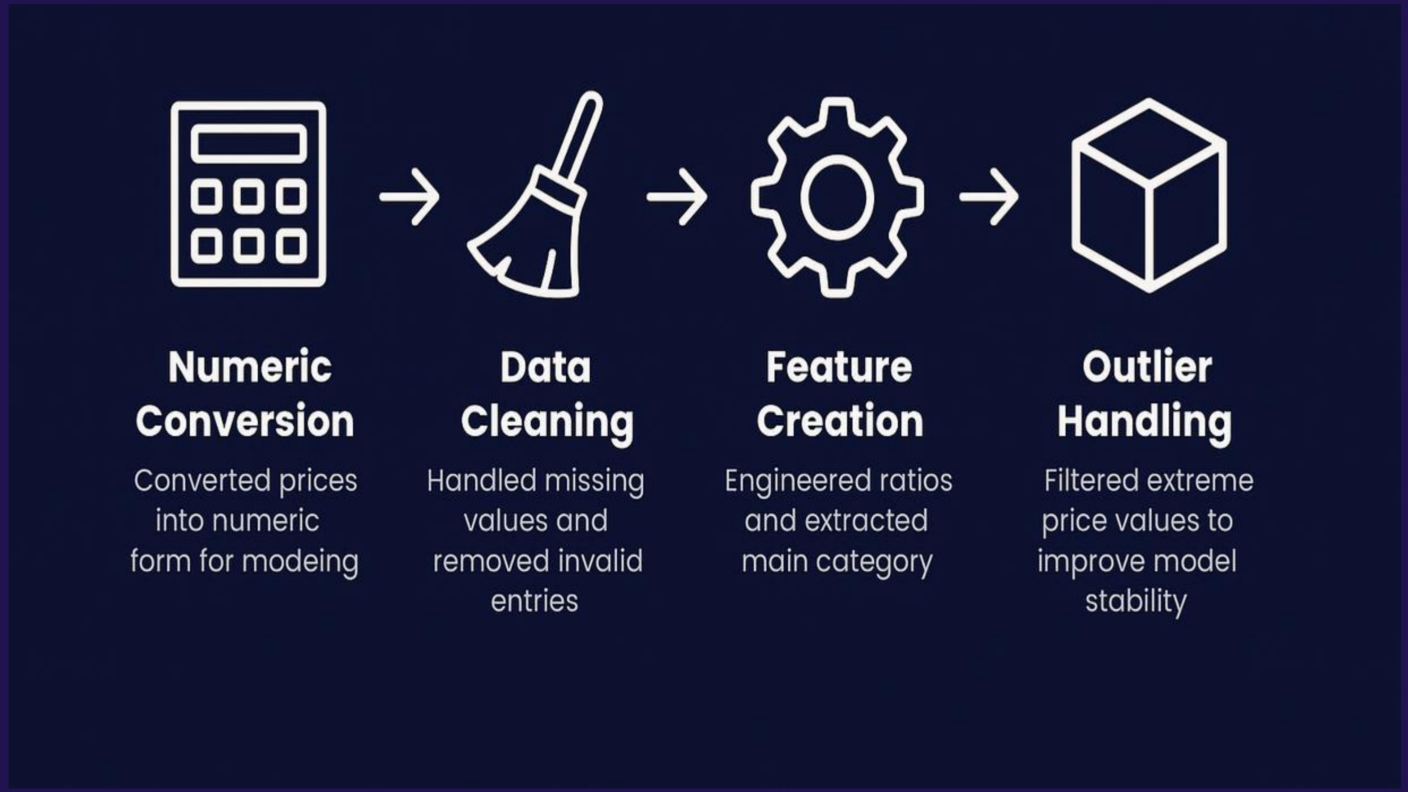
Data Understanding

Data Source

- The dataset we worked with was scraped from Jumia.
- It consist of 1,999 product listings with 13 features (columns) including:
  - current_price → Current product price.
  - original_price → Price before discount
  - discount → Discount percentage
  - main_category → Product category (electronics, fashion, etc.)
  - rating_number & verified_ratings → Customer satisfaction.
  - seller → Who is selling the item.
  - title → Product description.

Data Preparation

- To prepare the dataset for analysis, we applied the following steps:
  - Removed duplicates: Many sellers list the same product multiple times.
  - Handled missing values: Some products lacked ratings or discounts. Strategies included:
    - Filling missing numeric values with averages.
    - Dropping irrelevant text-only columns when not useful for modeling.
  - Converted datatypes: Prices were stored as strings (KSh 499 → 499). Converted to integers.
  - Created new features:
    - discount_amount = original_price - current_price
    - discount_ratio = % discount offered



Numeric Conversion
Converted prices into numeric form for modeing

Data Cleaning
Handled missing values and removed invalid entries

Feature Creation
Engineered ratios and extracted main category

Outlier Handling
Filtered extreme price values to improve model stability

Exploratory Data Analysis (EDA)

➢ We explored the dataset to understand patterns and relationships among products listed on Jumia Kenya, focusing on:

- pricing behaviour
- discount strategies
- customer ratings, and category-level
- competitiveness among products listed on Jumia Kenya.

➢ The process involved four main steps:

- data preparation
- visual exploration
- category-level analysis
- feature engineering.

Price distribution:

➢ To explore the relationship between original prices and discounts, a correlation analysis was conducted. The resulting correlation coefficient (0.08) indicates a very weak relationship between the two variables.

➢ This suggests that both low-priced and high-priced products are equally likely to receive discounts, meaning discount rates are not strongly influenced by a product's original price.

Exploratory Data Analysis (EDA)

Impact of discounts on product visibility.

➢ To understand how discounts differ across product price tiers, products were grouped into four ranges:

- Low (<1,000 KSh)
- Mid (1,000–5,000 KSh)
- High (5,000–10,000 KSh)
- Very High (>10,000 KSh)

➢ Discount by Price Range

- Discounts remain consistent across all price ranges.
- Whether products are low, mid, or high priced, they receive similar discount percentages, which is unexpected, as higher-priced items are often assumed to have larger discounts.

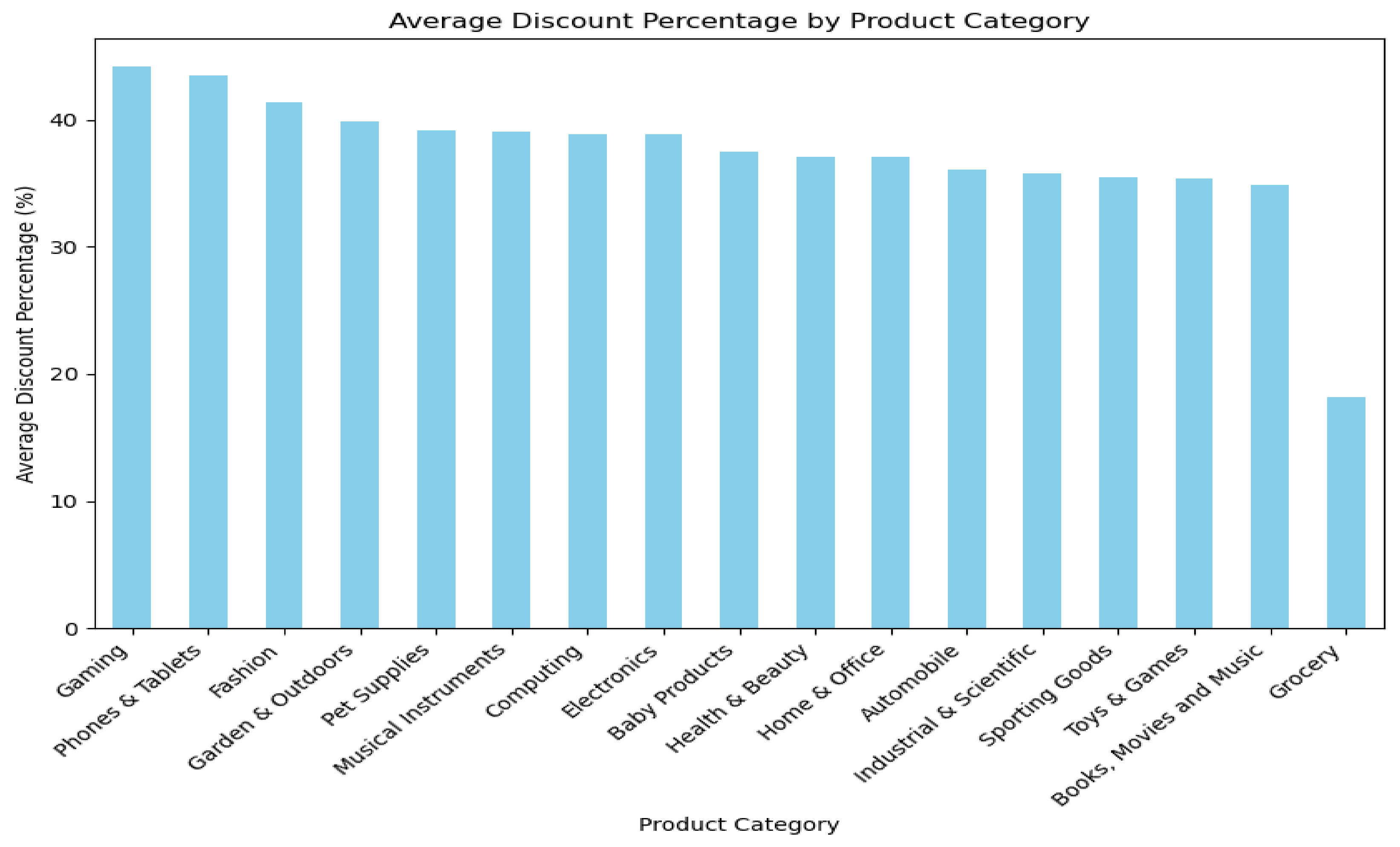
➢ Discount by Product Category:

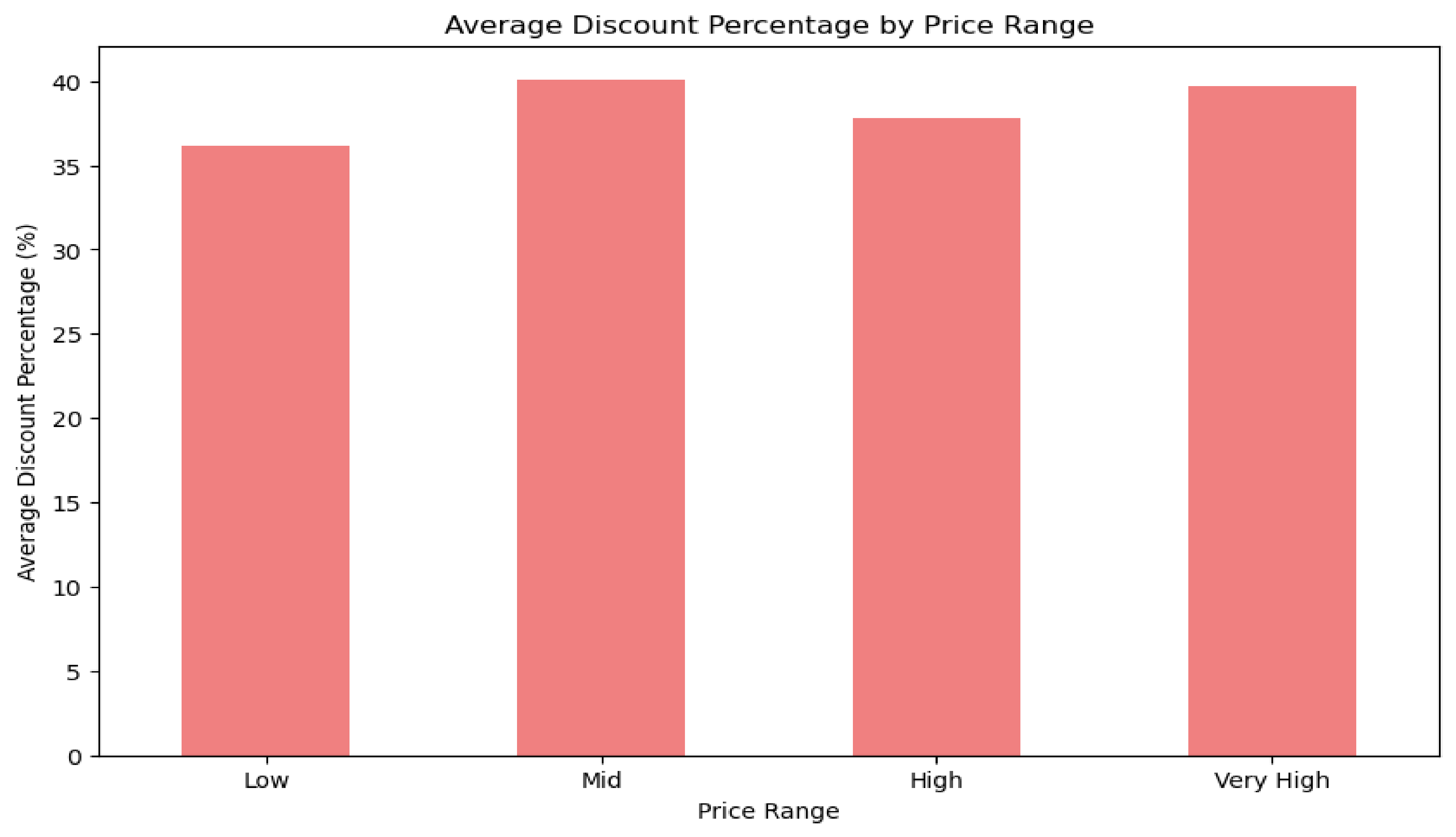
- *Gaming* and *Phones & Tablets* show the highest average discounts (40%+) likely due to strong competition and frequent model updates.
- *Groceries* have the lowest discounts, reflecting their smaller profit margins.
- *Fashion*, *Computing*, and *Health & Beauty* maintain moderate discounts (20–40%), suggesting similar pricing strategies across these segments.

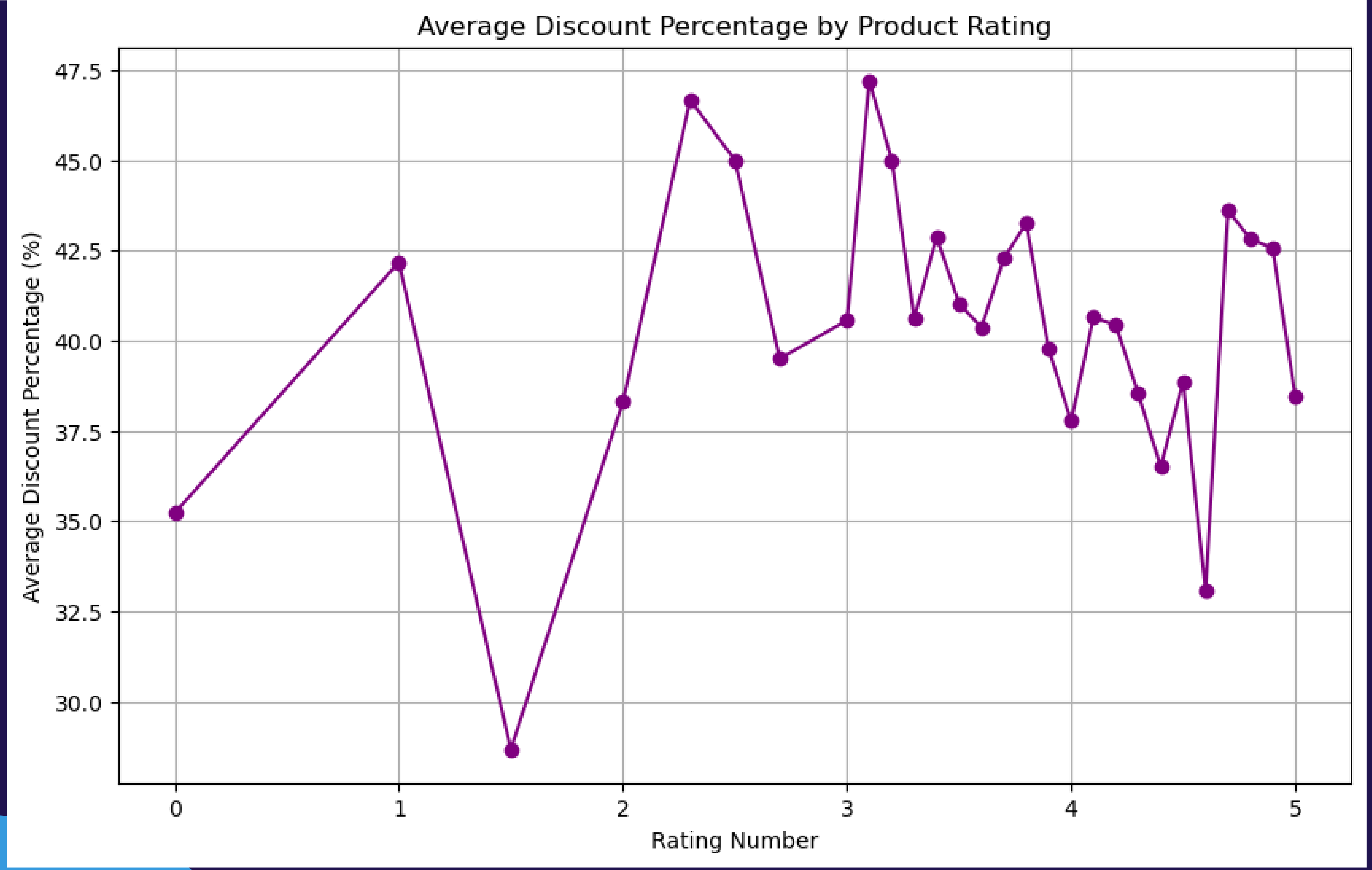
•Summary:

- Discount strategies are largely independent of product price, with consistent discount patterns across most price ranges and categories.
- *Gaming* and *Phones & Tablets* receive the highest discounts likely due to competition and frequent model updates, while *Groceries* show the lowest.
- Overall discounts do not appear strongly influenced by either product price or customer ratings.

Exploratory Data Analysis (EDA)





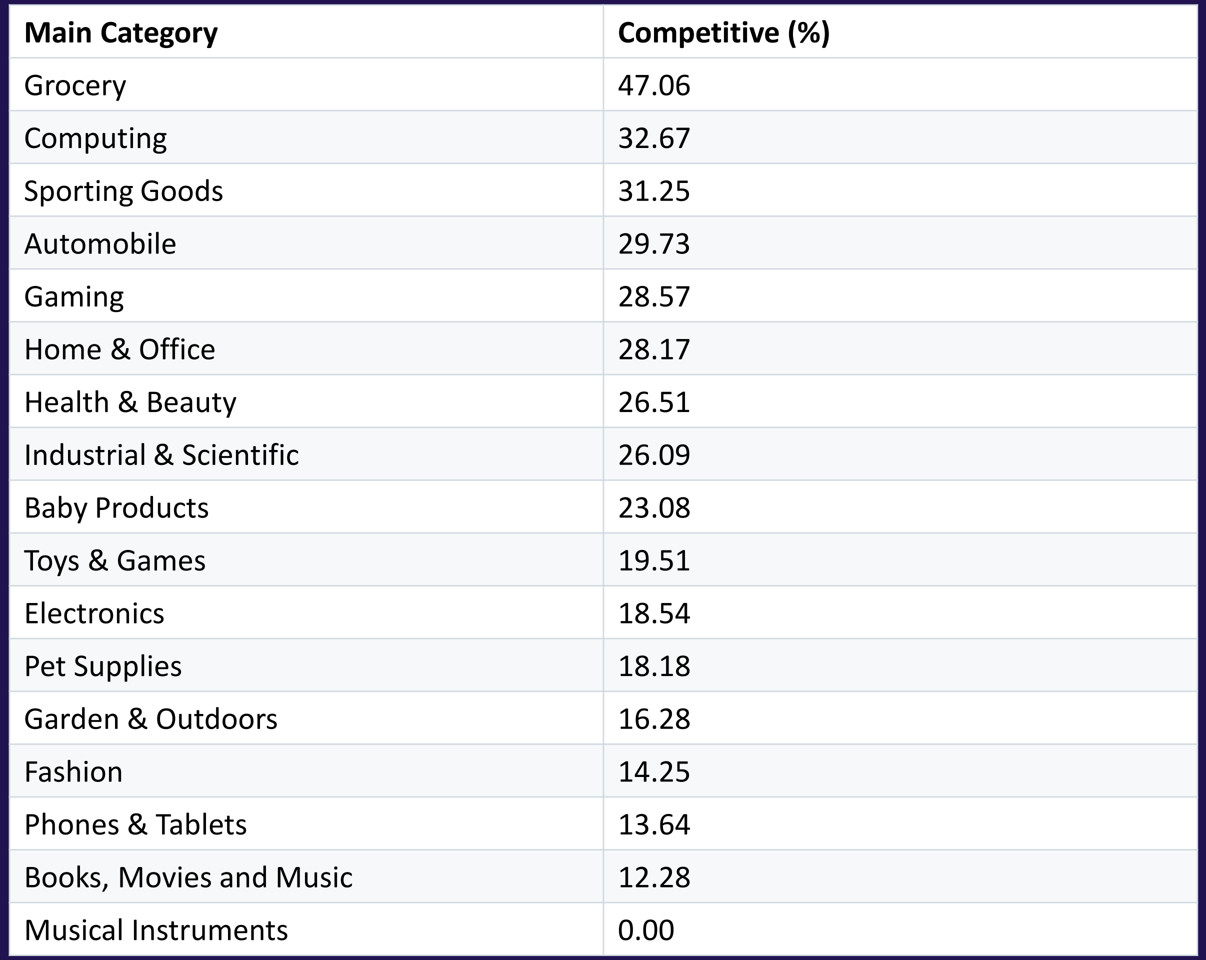


Exploratory Data Analysis (EDA)

Product Competitiveness.

- To identify the best-value or “deal” products, we assessed competitiveness within each category. A product was considered competitive if:
  - Its price was at or below the category’s median price, and
  - It had a rating of 4.0 or higher.

•This method highlights products that are both affordable and well-rated, revealing which categories offer the best balance between price and customer satisfaction.



Main Category	Competitive (%)
Grocery	47.06
Computing	32.67
Sporting Goods	31.25
Automobile	29.73
Gaming	28.57
Home & Office	28.17
Health & Beauty	26.51
Industrial & Scientific	26.09
Baby Products	23.08
Toys & Games	19.51
Electronics	18.54
Pet Supplies	18.18
Garden & Outdoors	16.28
Fashion	14.25
Phones & Tablets	13.64
Books, Movies and Music	12.28
Musical Instruments	0.00

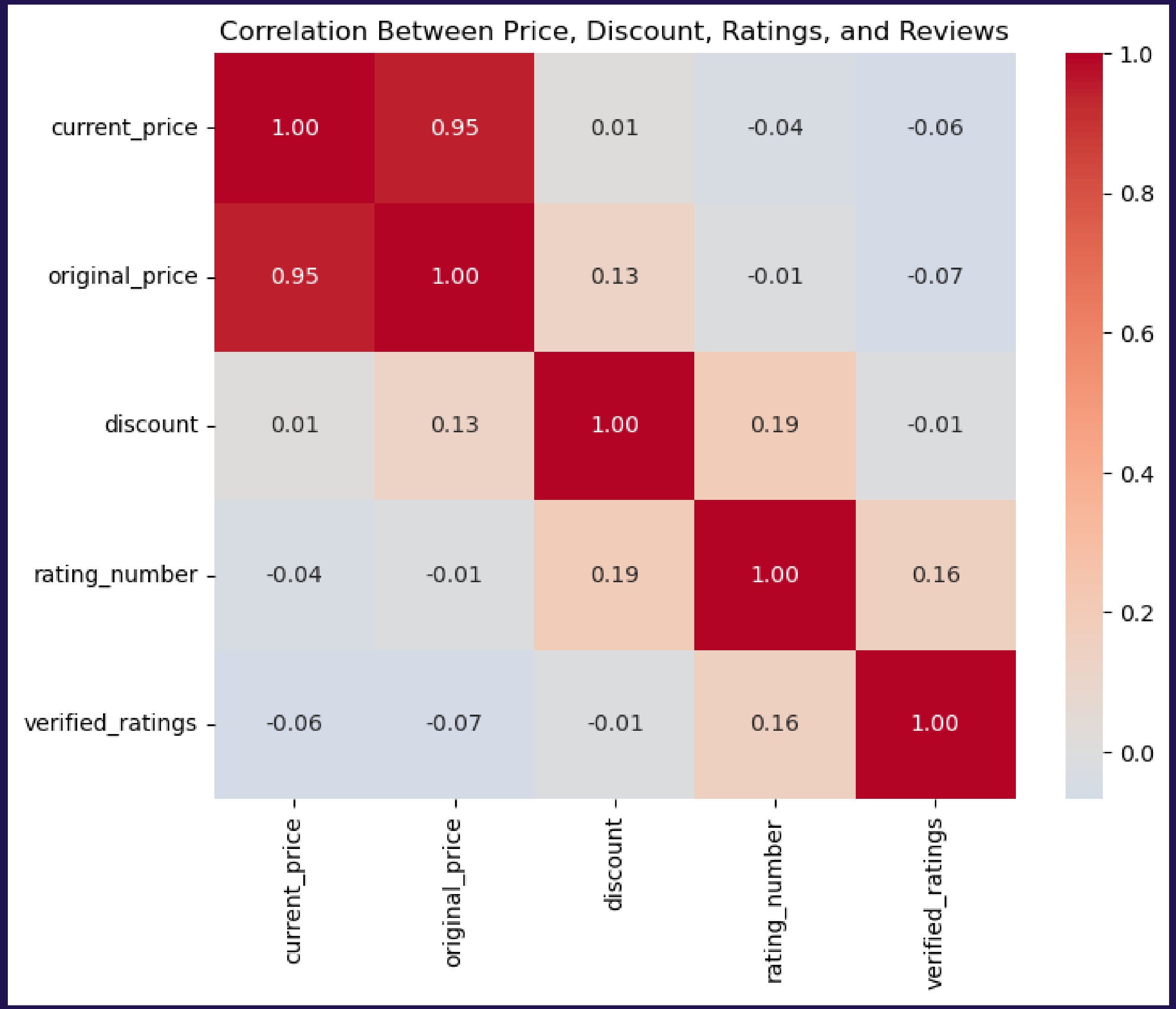
- key Insights:
  - *Grocery* items are the most competitive (47%), showing strong affordability and consistent ratings.
  - *Computing*, *Sporting Goods*, and *Automobile* categories follow closely (~30%).
  - *Fashion*, *Phones & Tablets*, and *Books* & *Music* have the lowest competitiveness, likely due to higher prices or lower ratings.

- Overall, competitiveness varies widely across categories, showing that strong performance (good price + good ratings) is not uniform across product types.

Exploratory Data Analysis (EDA)

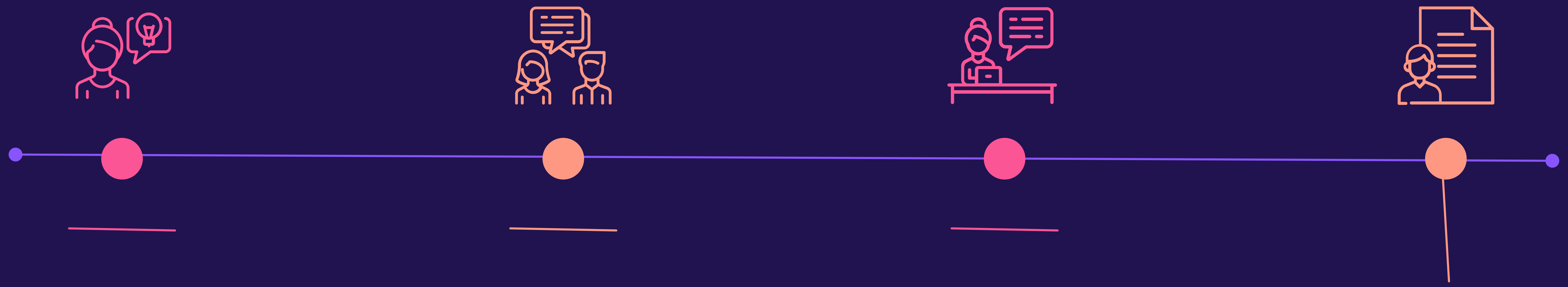
Correlation

- There is a strong positive correlation between current and original prices as current prices are derived from discounts.
- Discounts shows weak influence on price, suggesting they serve more as marketing tools than true price drivers.
- There’s no clear link between price and ratings, indicating that expensive products don’t necessarily receive better reviews.
- While discounted products may attract slightly more ratings, the effect is minimal. As expected, products with more reviews tend to have slightly higher average ratings.



Modeling

- Objective :
  - Build a reliable model that predicts fair, competitive prices while adapting to category trends.



Data Preparation

- Clean prices and categories
- Handling missing values and outliers

Feature Engineering

- Compute median price per category

Test Multiple Models

- Logistic Regression K-Nearest Neighbours (KNN) Random Forest, and Neural Network

Model Optimization & Evaluation

- Neural Networks/KNN
- Compare accuracy, precision, recall
- Select best-performing model for deployment

Modelling

Data Source

➢ Before building predictive models, the data was cleaned and pre-processed. Missing values were imputed, categorical variables were encoded, and numerical features were scaled to ensure consistency and improve model performance. The dataset was then split into training (80%) and testing (20%) subsets to evaluate how well each model generalizes to unseen data.

Models Tested

➢ Several regression models were developed to predict the target variable based on product and pricing features:

- Linear Regression served as a baseline for performance comparison.
- K-Nearest Neighbours (KNN) Regressor captured local relationships in the data.
- Decision Tree Regressor modelled non-linear patterns and feature interactions.
- Random Forest Regressor leveraged multiple decision trees for better generalization.
- Neural Network (MLP Regressor) tested for its ability to model complex, non-linear relationships.

• Pipeline & Training

➢ All models were trained using a consistent preprocessing pipeline built with Column Transformer and Pipeline, ensuring reproducibility and efficient feature handling.

Model Performance

➢ Initial model results showed the following trends:

- Linear Regression performed strongly, achieving an R^2 of 0.96, indicating a solid linear fit.
- KNN and Decision Tree achieved R^2 scores of 0.91 and 0.88 respectively, suggesting some overfitting or sensitivity to data structure.
- Neural Network showed moderate performance (R^2 of 0.87) but required longer training and did not converge fully.
- Random Forest delivered a high R^2 of 0.94, significantly reducing error and improving predictive stability.

Modelling

Model Optimization

- To further improve accuracy, GridSearchCV was applied for hyperparameter tuning across three advanced models:
  - Ridge Regression
  - Lasso Regression
  - Random Forest Regressor
- Each model was fine-tuned using 3-fold cross-validation to identify best performing parameters.

Model Selection

- Model performance was compared using the following metrics:
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - Root Mean Squared Error (RMSE)
  - R-squared (R^2).

Model	**Mean Absolute Error (MAE)**	**Mean Squared Error (MSE)**	**Root Mean Squared Error (RMSE)**	**R-squared (R^2)**
Random Forest	285.90	721,254	849.27	0.9923
Ridge Regression	1001.57	3,691,000	1921.20	0.9606
Lasso Regression	815.11	3,692,944	1921.70	0.9606

Interpretation:

- The Random Forest Regressor emerged as the best-performing model achieving:
  - MAE: 285.90
  - MSE: 721,254
  - R^2: 0.99

- MAE measures the average difference between predicted and actual values (model's predictions deviate by about 286 units on average).

- The Mean Squared Error (MSE) penalizes larger errors more heavily, indicating overall low prediction errors.

- R^2 = 0.99 means the model explains 99% of the variance in the target variable, making it the most accurate and robust choice for deployment.

Deployment

- The project was deployed as a Flask web app and hosted on Render.

- For the Frontend, we created a simple site where people can explore the project. The site has different pages:
    - Home → welcoming page with project overview.
    - Data → a page describing the dataset used.
    - Dashboard → Interactive Tableau Dashboard.
    - Team → information about the contributors.

- There’s also a prediction form where a user can enter product details and instantly see the predicted price.

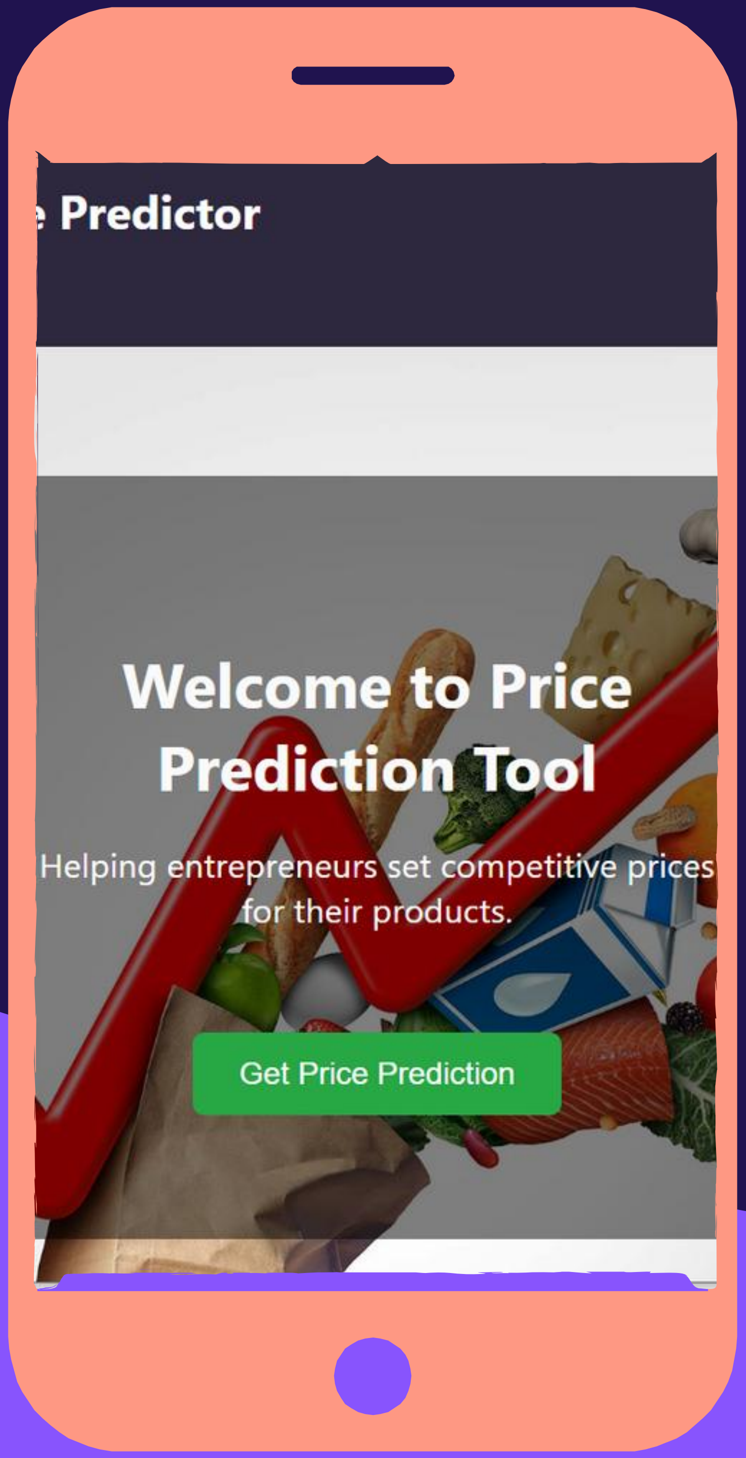
- The backend used Random Forest model that was trained on our dataset. The outcome is a live, user-friendly web app that makes predictions in real time.

- 🔗 The tool is live and accessed here. You can explore the tool, input a product name and view competitor based pricing insights.

User Feedback

- To gather user insights, we shared a short questionnaire with small business owners after they explored the deployed tool.
- The goal was to understand:
    - Whether the tool is helpful for their pricing decisions
    - If it solves their challenge of knowing competitor prices
    - What improvements or extra features they would like
- You can view or take the Feedback questionnaire here:

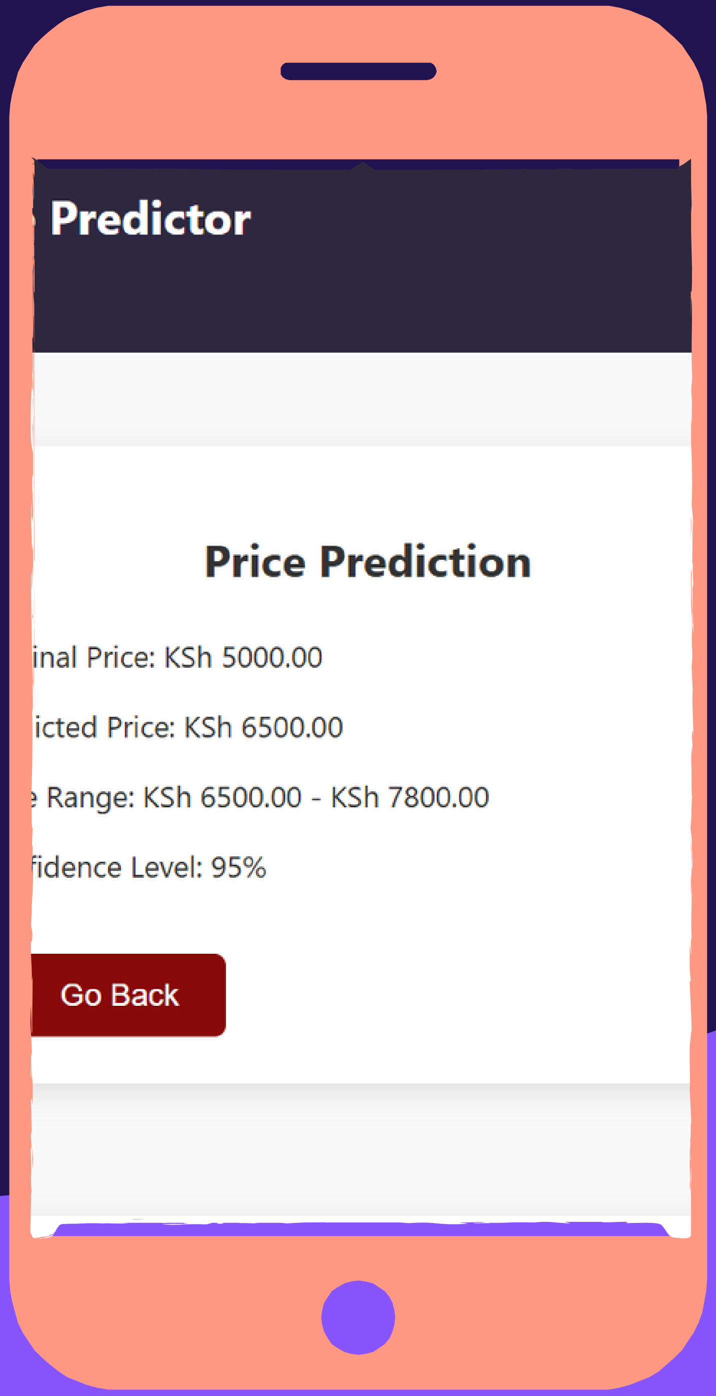
Deployment



e Predictor

Welcome to Price Prediction Tool

Helping entrepreneurs set competitive prices for their products.



Predictor

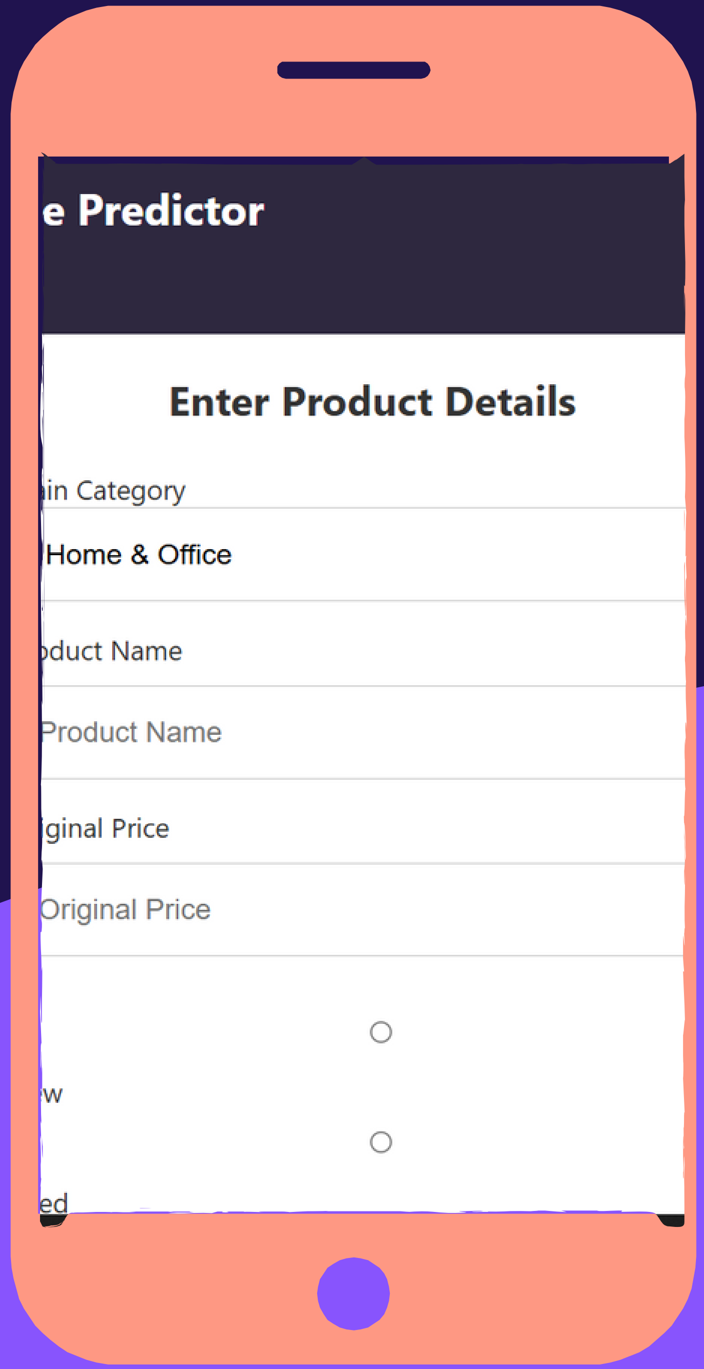
Price Prediction

inal Price: KSh 5000.00

icted Price: KSh 6500.00

e Range: KSh 6500.00 - KSh 7800.00

fidence Level: 95%



e Predictor

Enter Product Details

in Category

Home & Office

duct Name

Product Name

ginal Price

Original Price

○
w
○
ed

Conclusions

- This project demonstrates the end-to-end process from exploring the dataset and extracting insights, to training a robust predictive model, and finally transforming it into a live, user-friendly web app.
- The result is a practical tool that allows anyone to input product details and instantly receive a predicted price, effectively bridging the gap between data and real-world application. It also shows how machine learning can support better decision-making in e-commerce, such as pricing strategies, deal evaluation, and customer awareness.

Thanks You

Presented by:
Harrison Kuria, Ryan Karimi, Elizabeth Ogutu, Hafsa Mohamed Aden, Rose Muthini, Lewis Mbugua

➢ Other resourceful Links

- Github
- Tableau Dashboard