From Data to Insight: A Machine Learning Approach for Quantifying and Visualizing the Pink Tax Disparity.

Team_No: 11

Sl. No.	Reg. No.	Name of the student
1	BL.EN.U4CSE22202	Yasasree Lasya
2	BL.EN.U4CSE22203	Archana Reddy P
3	BL.EN.U4CSE22204	Asi Kuushalie
4	BL.EN.U4CSE22221	Divya Jyothi G

Project Advisor: Dr Sajitha Krishnan

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

Introduction

- Gender-based pricing discrimination persists across global retail markets.
- Pink Tax charges women more for similar products.
- Indian e-commerce remains largely unstudied for this issue.
- Dynamic pricing may unintentionally worsen gender price gaps.
- No large-scale Indian dataset exists to measure Pink Tax.
- Research gap limits consumer protection and policymaking.

Problem Statement

- Women pay higher prices for functionally identical items.
- Indian platforms lack transparency in gendered pricing.
- Algorithmic pricing may reinforce hidden discriminatory patterns.
- No standardized system tracks pricing disparities in India.
- Policymakers lack evidence for regulating price inequality.
- Consumers remain unaware of systematic gender-based overpricing.

Research Gaps

1. Limited Use of Advanced Machine Learning for Pink Tax Analysis
2. Heavy Reliance on Descriptive or Self-Reported Data
3. Lack of Dynamic Pricing Systems
4. Geographical and Dataset Limitations
5. BI and Visualization Papers Lack Predictive Depth
6. Lack of context based justification

Dataset

Dataset Source:

- Data collected through automated web scraping from major Indian e-commerce platforms: Amazon, Flipkart, Nykaa, Myntra.
- Scrapers handled dynamic layouts, JS-loaded data, product variations, and ensured multiple time-interval scrapes to reduce price fluctuations.
- Only publicly available product details were collected—no login, no policy violations.

Dataset Size & Structure

- Total scraped: 3,621 product entries.
- Final cleaned dataset: 2,000 entries → forming 1,000 male–female matched product pairs.
- Each record has 13 variables, including: productid, pairid, brand, category, subcategory, gendertarget, price, size, sizeunit, retailer, description, ingredients

Dataset

Category Distribution

- Personal Care – 45%
- Hygiene Products – 20%
- Shaving Products – 15%
- Health Products – 10%
- Clothing/Footwear – 10%

Target Variable

- Pink Tax (absolute): womenprice – menprice
- Pink Tax %: percentage difference relative to men's price
- Preliminary Insight:
  - Women's products cost more in 68.7% of pairs
  - Avg absolute markup: ₹110.79
  - Avg percentage markup: 33.6%

Data Preprocessing and Exploration

1. Data Cleaning
   - Normalized price formats from different sites into consistent numerical values.
   - Standardized brand names and corrected spelling variations.
   - Converted sizes into uniform units (ml, g, count, clothing sizes).
   - Removed duplicate listings and cross-platform repetitions.

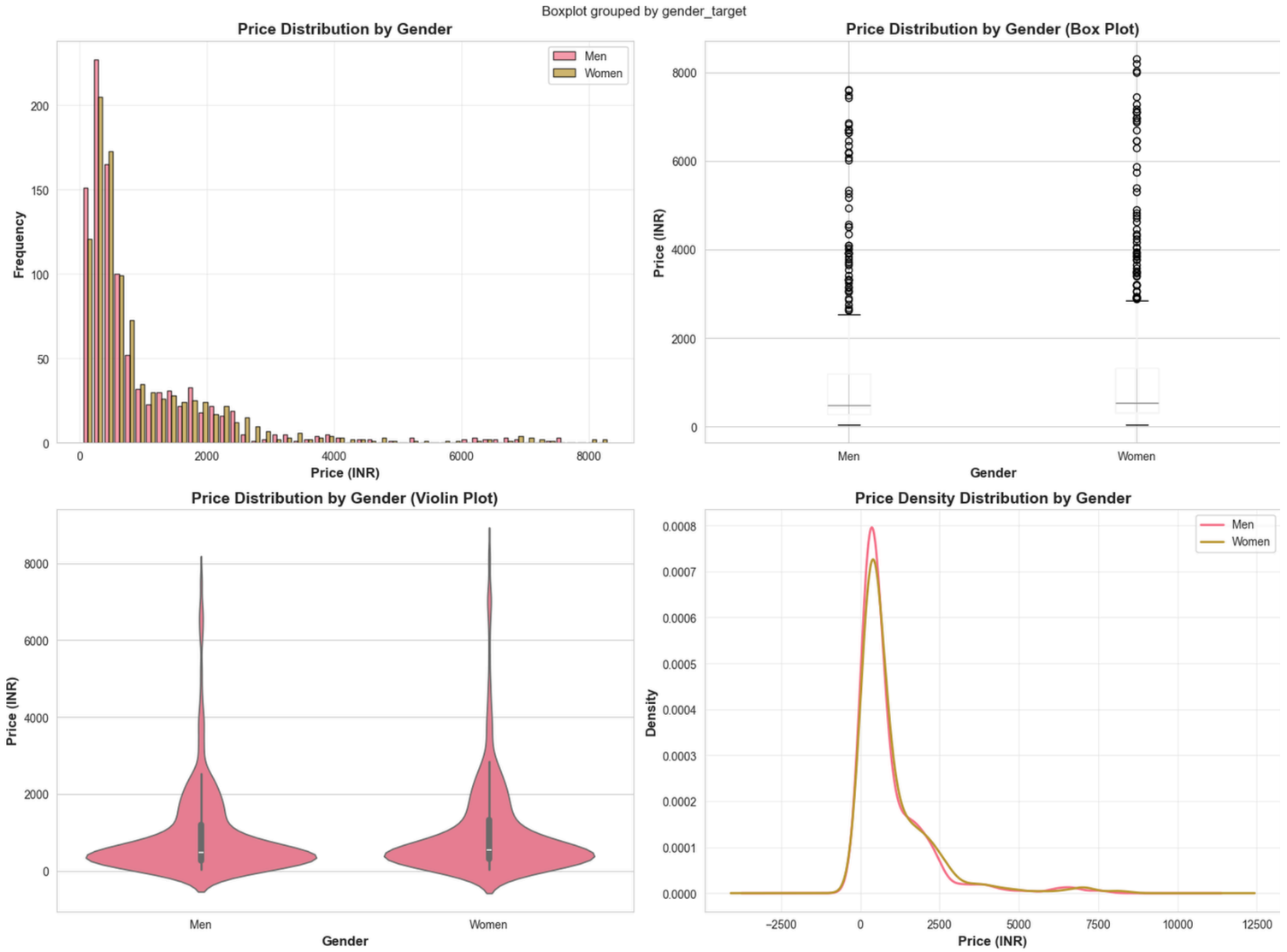
2. Gender Filtering & Pair Matching
   - Removed unisex items to keep only clearly gender-targeted products.
   - Implemented a matching algorithm based on: **brand, category & subcategory, size similarity, name similarity.**

Result: 1,000 reliable male–female matched product pairs.

3. Exploratory Analysis
   - Price distribution found to be right-skewed.
   - Identified clear pricing differences across: categories, brands, retailers
   - Verified consistency of product sizes and prices before statistical modelling.

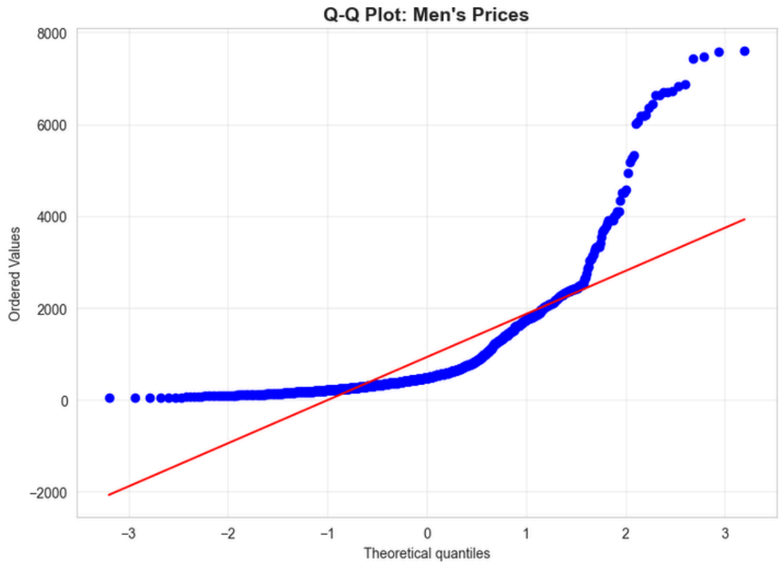
AMRITA VISHWA VIDYAPEETHAM

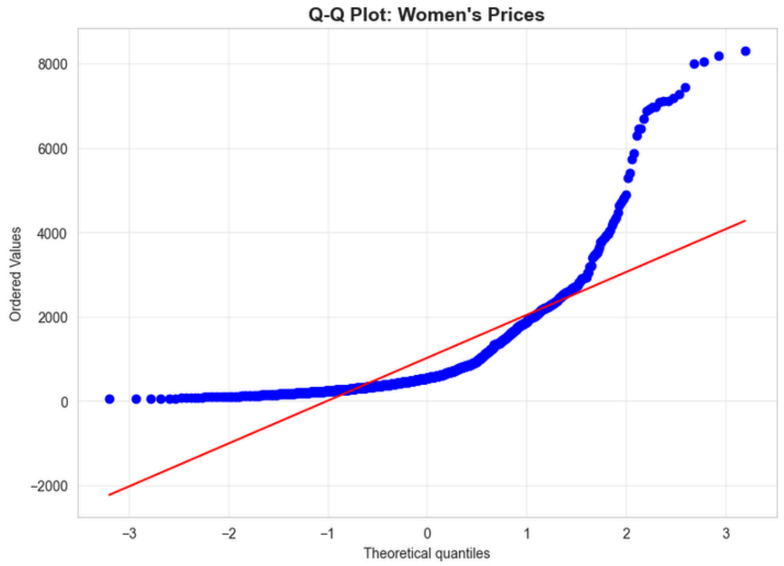
Data Exploration: Price Distrbution by Gender

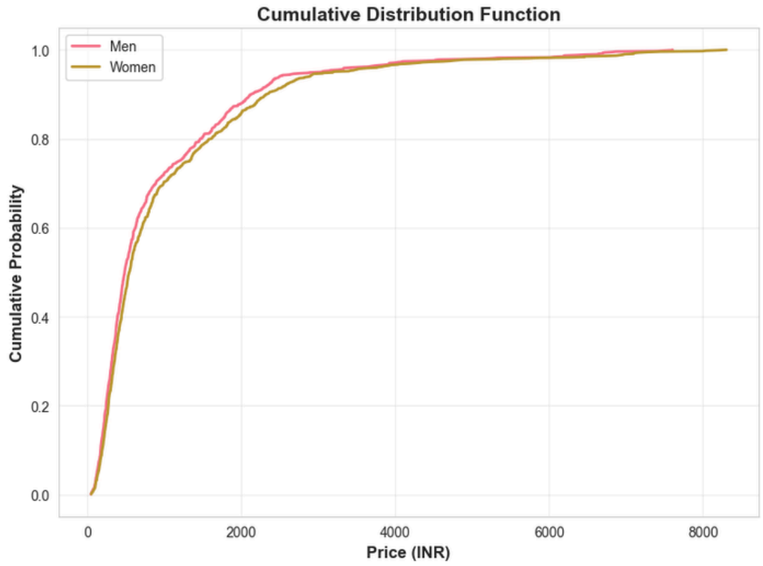


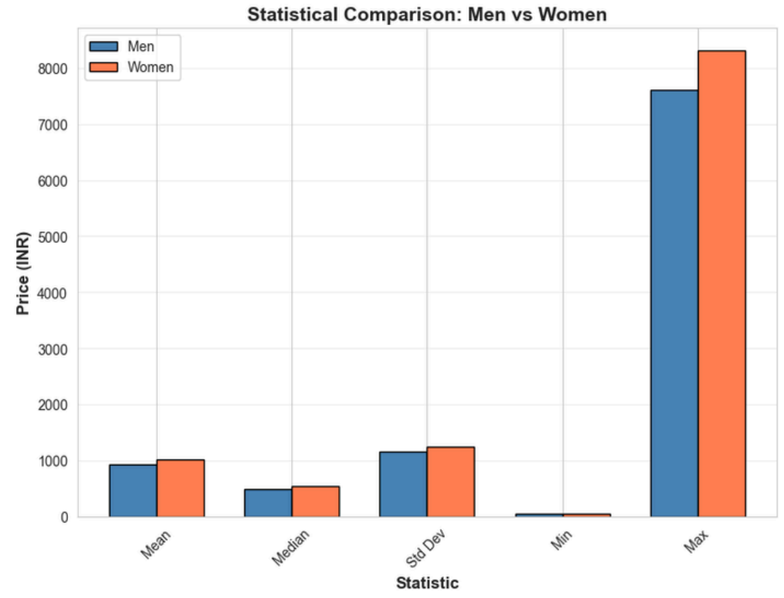
AMRITA VISHWA VIDYAPEETHAM

Data Exploration: Statistical Comparison



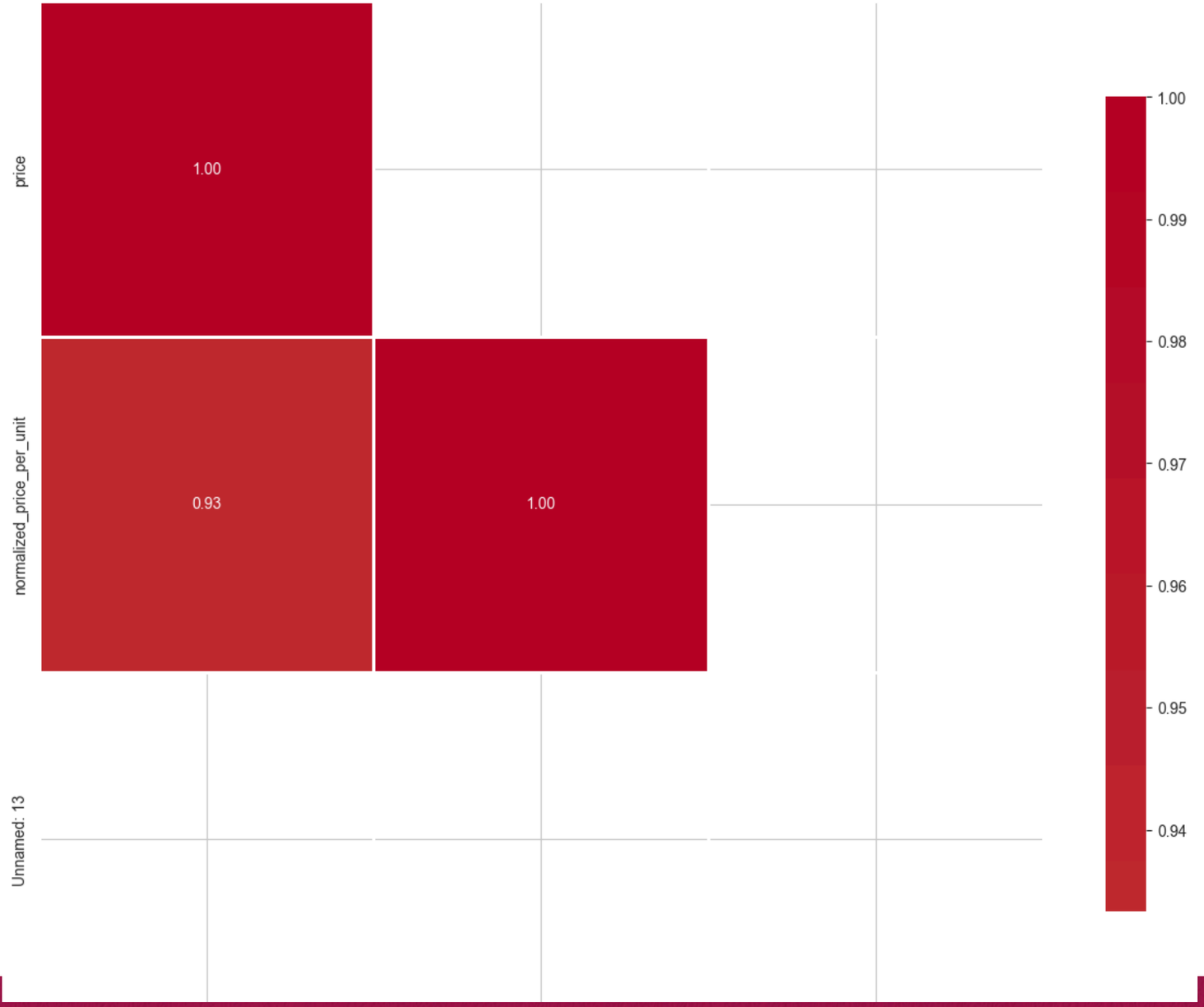






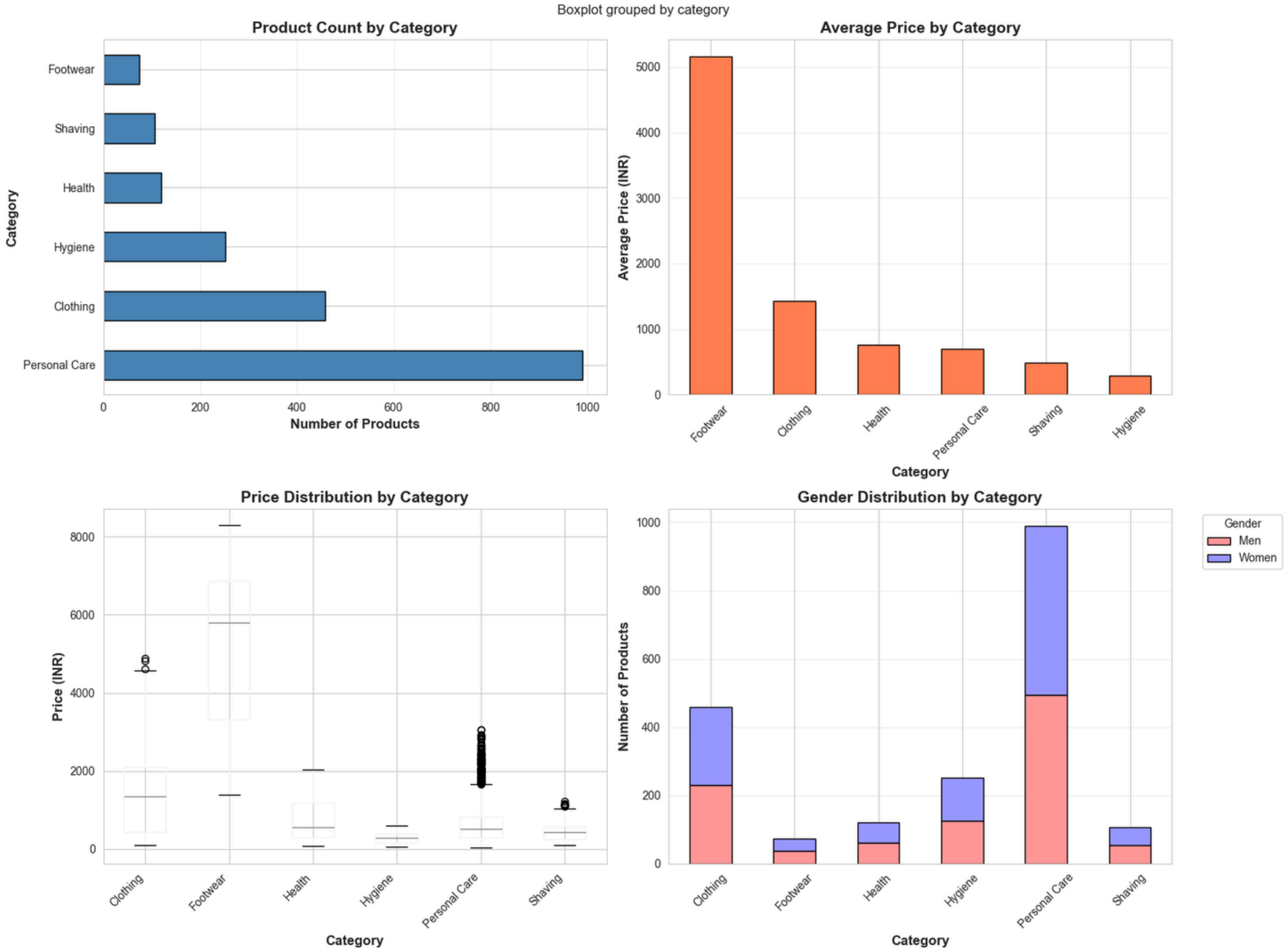
AMRITA VISHWA VIDYAPEETHAM

Data Exploration: Correlation Heatmap



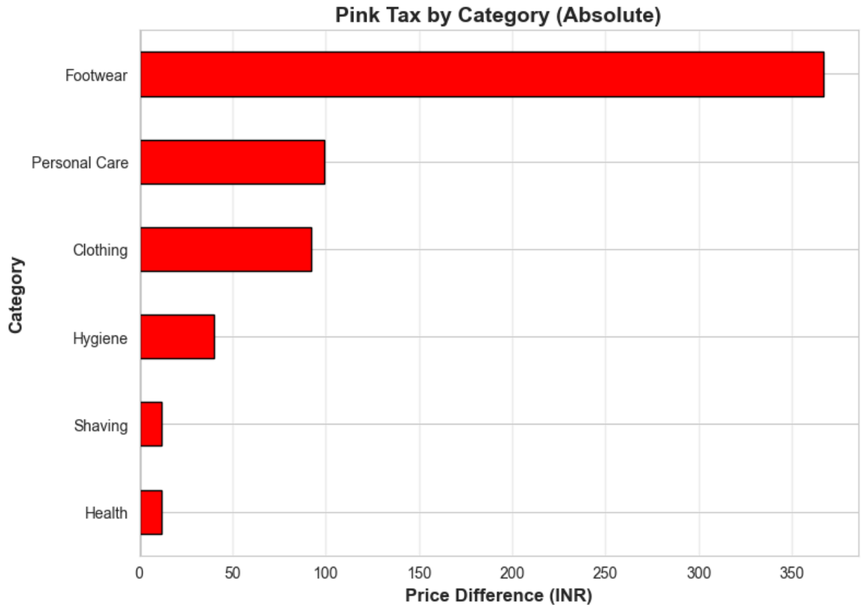
AMRITA VISHWA VIDYAPEETHAM

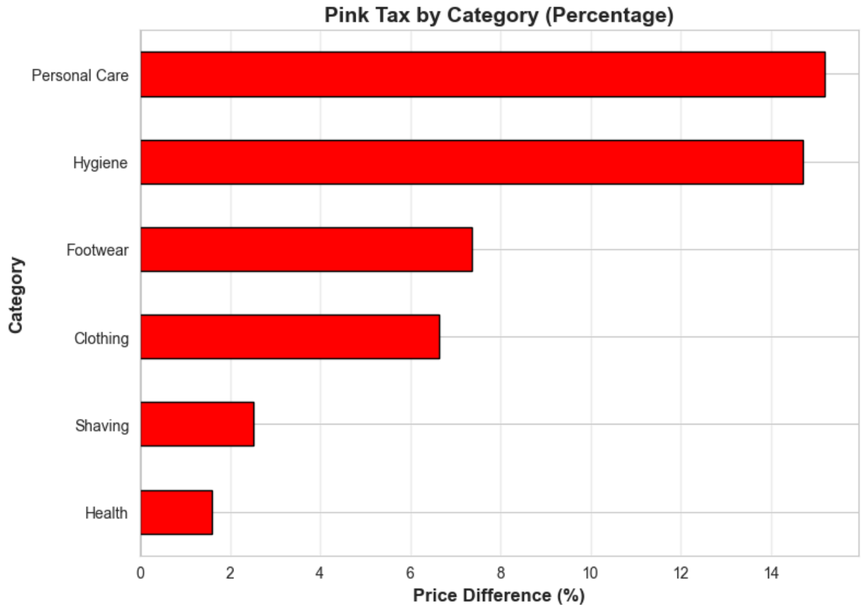
Data Exploration: Category

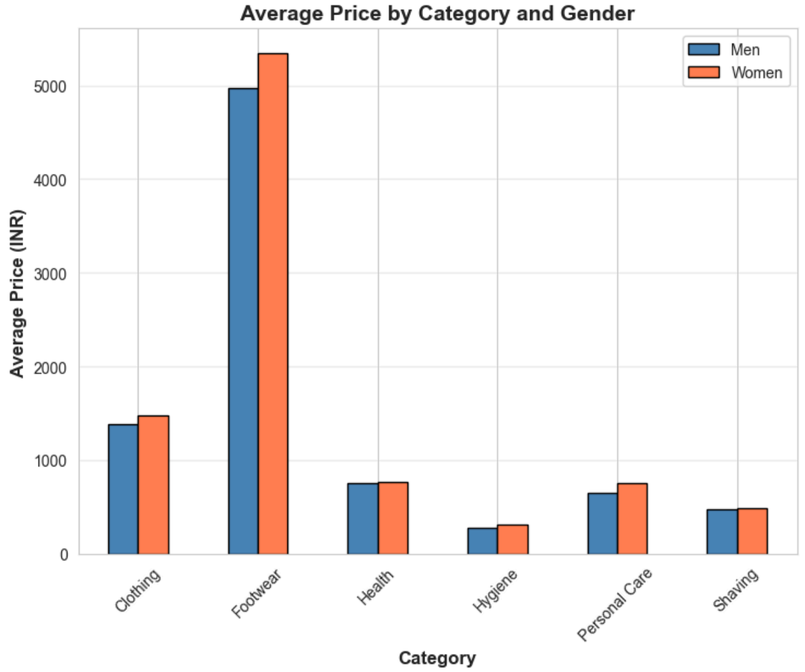


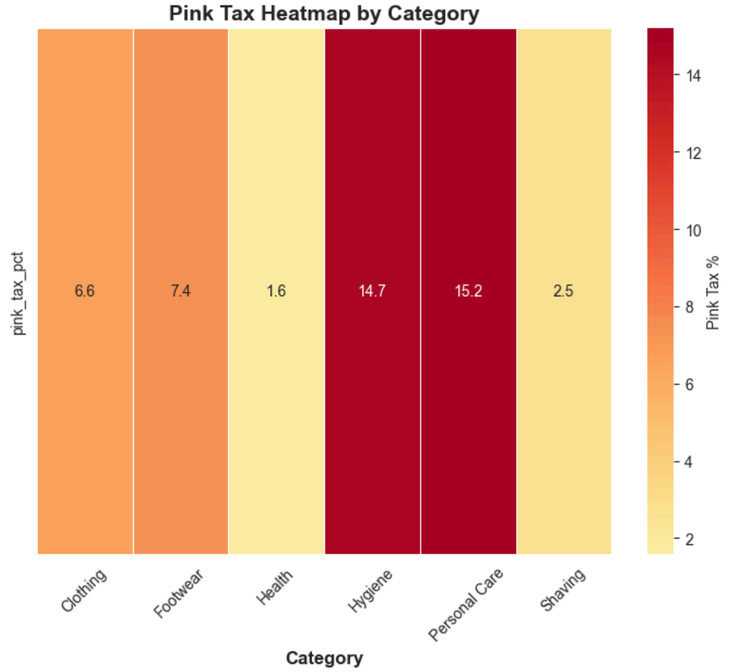
AMRITA VISHWA VIDYAPEETHAM

Data Exploration: Subcategory



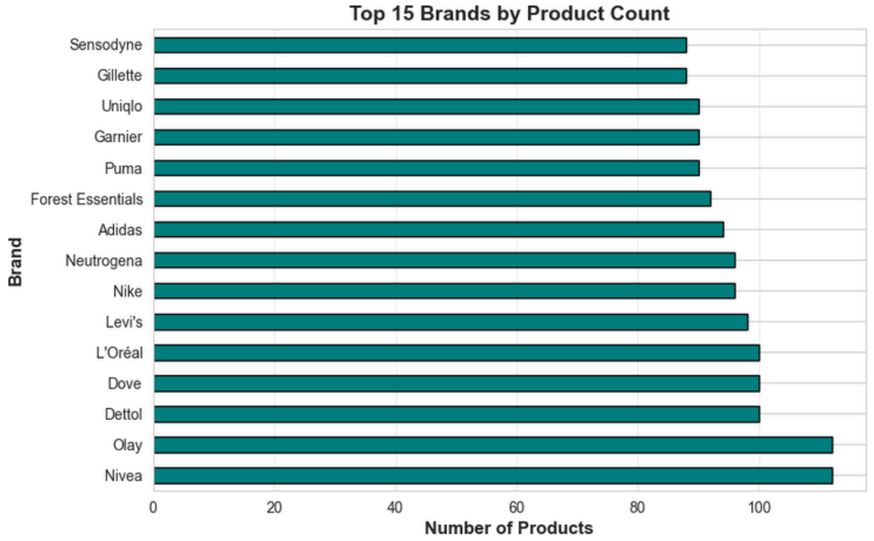


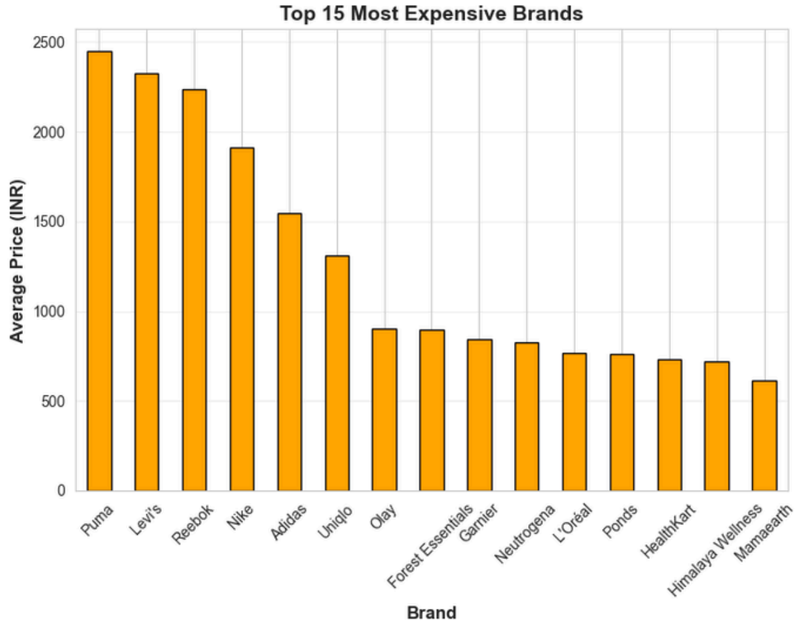


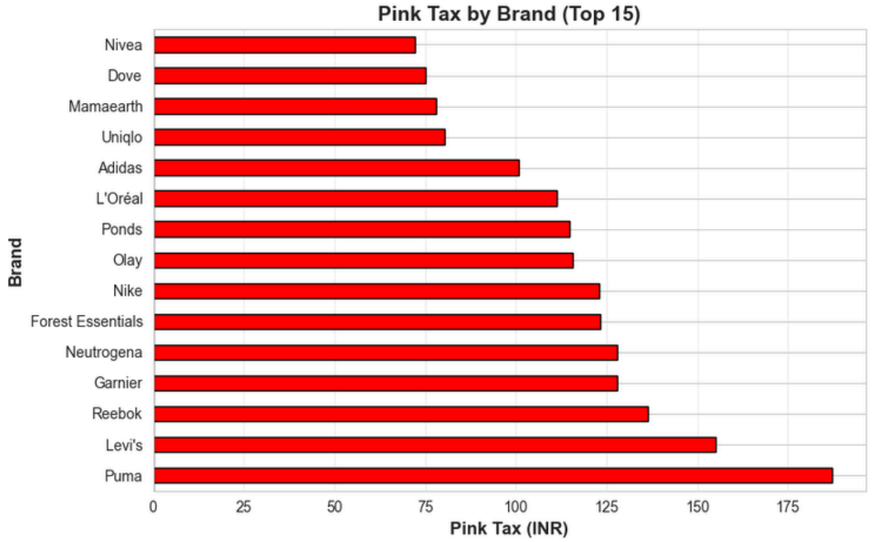


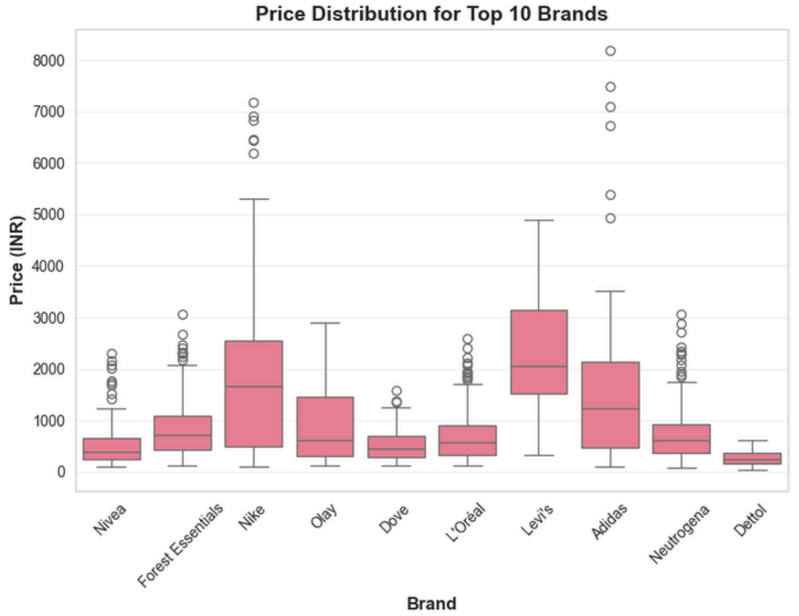
AMRITA VISHWA VIDYAPEETHAM

Data Exploration: Brand Based Classsification









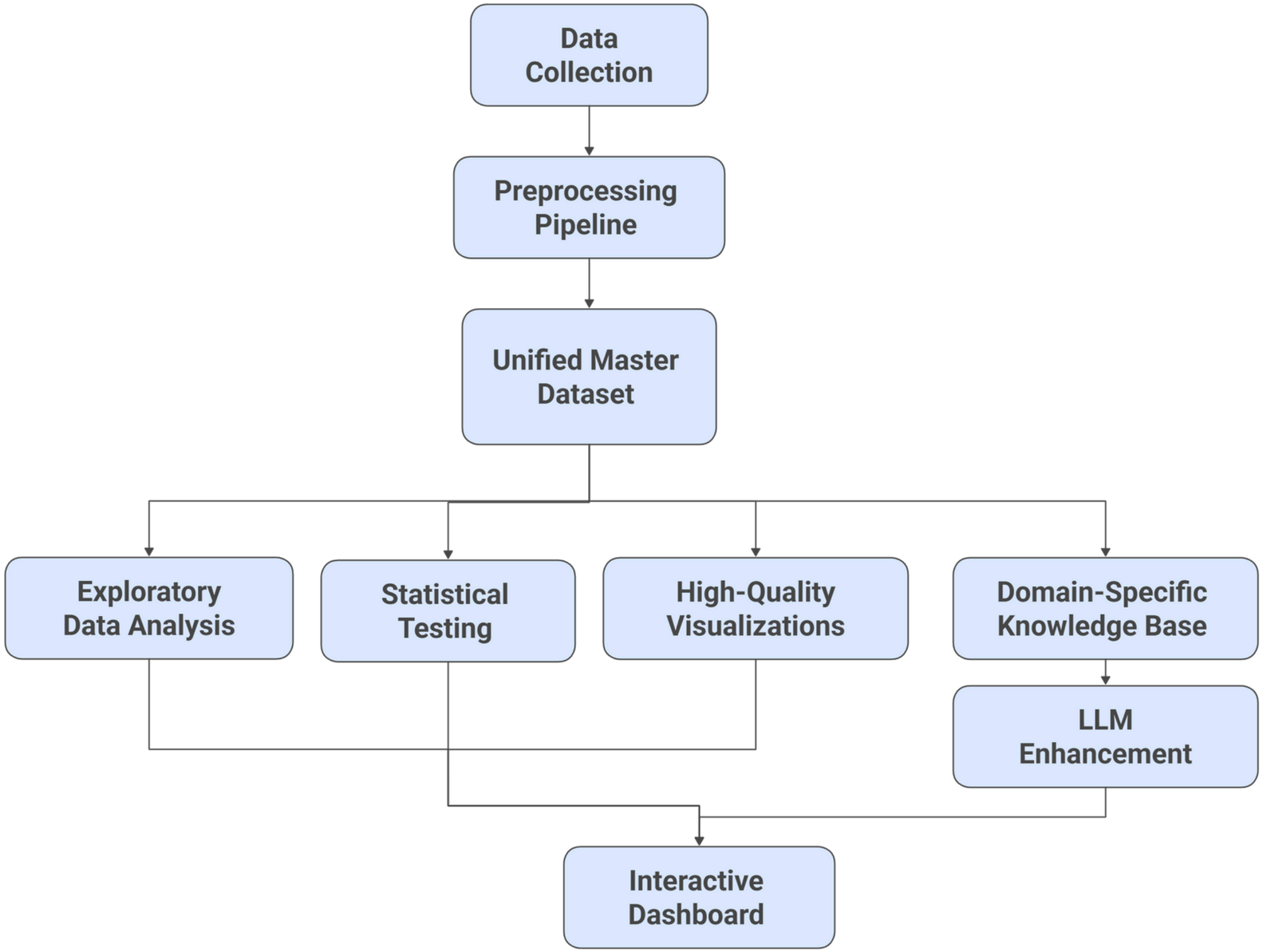
AMRITA VISHWA VIDYAPEETHAM

Tools Used

- Python for data processing and automation
- Pandas & NumPy for cleaning and analysis
- Matplotlib & Seaborn for visualizations
- BeautifulSoup / Requests for data collection
- Flask for chatbot backend API
- Ollama / LLM for AI-driven explanations
- HTML, CSS, JavaScript, Chart.js for the dashboard
- JSON & CSV for data storage

AMRITA VISHWA VIDYAPEETHAM

Architecture Diagram



AMRITA VISHWA VIDYAPEETHAM

Methodology

1. Data Acquisition

- Collect raw product data from major e-commerce websites.
- Use website scraping to extract product names, prices, categories, and gender labels.
- Ensure enough volume and variation for accurate analysis.
- Use simple Python tools for scraping and creating sample data.

2. Data Cleaning And Normalization

- Clean messy raw data and convert it into a standard format.
- Fix inconsistent price formats and convert them into numeric values.
- Standardize size units (e.g., L → mL, g → kg).
- Remove unwanted characters and correct brand/product names.
- Ensure all data columns follow the same structure.

Methodology

3. Product Matching & Pink Tax Calculation

- Match similar men's and women's products using a common product key.
- Remove gender-specific words to find equivalent items.
- Pair comparable products based on brand, size, and features.
- Calculate the price difference between men's and women's versions.
- Compute the pink tax percentage and absolute difference.

4. Product Matching & Pink Tax Calculation

- Study overall price trends and distributions in men's vs women's products.
- Use statistical tests to check if price differences are significant.
- Identify which categories show the highest pink tax.
- Detect unusual price values or outliers.
- Confirm that results are real and not due to random chance.

Methodology

5. Visualization & Reporting

- Convert findings into simple, clear charts and graphs.
- Show comparisons using bar charts, pie charts, box plots, and histograms.
- Build an interactive dashboard for easy exploration.
- Allow users to filter products and see real-time comparisons.
- Present insights visually for better understanding.

6. AI-Driven Analysis & Interaction

- Add a smart chatbot that explains pink tax findings.
- Use NLP to understand user questions about pricing differences.
- Build a knowledge base from the cleaned dataset.
- Use an LLM (Llama3 via Ollama) to generate clear insights and suggestions.
- Provide personalized explanations through a Flask-powered API.

AMRITA VISHWA VIDYAPEETHAM

Results

Accuracy Metrics	Value
Accuracy	98.65%
Precision	95.90%
Recall	95.90%
F1 Score	92.81%

Confusion Matrix Metrics	Value
False Positive Rate	5.30%
False Negative Rate	5.30%
True Positive Rate	95.90%
True Negative Rate	89.23%

Results

Justification Quality Metrics	Value
Justified Price MAE	₹10
Justified Price Median AE	₹10
Upgrade Completeness	85.60%
Upgrade Precision	77.80%
Upgrade F1 Score	87.20%

Consistency Metrics	Value
Verdict Consistency	98%
Verdict Flip Rate	4.50%
Justified Price Std Dev	₹10
Brand Consistency	100%
Overall Consistency	98.50%

Results

Response Quality Metrics	Value
Average Explanation Length	24 words
Clarity Score	4.6
Evidence Citation Rate	73%
High Confidence Accuracy	85%
Confidence Calibration	92.80%

Conclusion and Future Scope

- The project collected, cleaned, and analyzed product data to identify price gaps between men's and women's products.
- Results showed clear and significant Pink Tax across several categories, supported by visual and statistical evidence.
- An interactive dashboard with AI insights made the findings easy to explore and understand.
- Future work includes expanding the dataset, improving matching with ML methods, adding real-time pricing APIs, and deploying the dashboard online with a stronger LLM.

THANK YOU