Research Questions

Do products(**Consumer Packaged Goods**.) explicitly marketed to women cost more than close male analogs after controlling for pack size, brand, and features?

Sub-questions:

What is the average "pink tax" percentage across different product categories?

How do pricing patterns differ between Western markets (US/EU) vs. South Asian markets (Pakistan)?

Which product categories show the highest gender-based price discrimination?

What factors influence the acceptance and prevalence of gendered pricing?

We test: after making fair comparisons (same brand, similar size), are "for women" items **priced higher** than the "for men" version?

Are U.S. MFN tariffs systematically higher on women's apparel vs. men's, and how large is the aggregate burden?

We measure: how big that difference is by clothing type using the tariff schedule (HTS 61/62, HS6 codes).

HTS 61 & 62 = parts of the **Harmonized Tariff Schedule** the U.S. uses to set import taxes (**tariffs**) on clothing.

- **61** = knitted/crocheted apparel (e.g., T-shirts, socks).
- **62** = woven apparel (e.g., dress shirts, jeans).

HS6 = a **6-digit product code** used worldwide to identify what a thing is at the border (e.g., "6109.10" = cotton T-shirts). "HS6" is just the first 6 digits of that code.

If women's clothing pays a higher import tax than men's in a category, do store prices for the women's items end up higher by about that much?

This isn't a perfect causal test, but we can check if **categories with bigger tax gaps also have bigger store price gaps**.

Dataset

First, what we will collect:

- **Retail prices in Pakistan** for everyday products (deodorant, shampoo, body wash, lotion, razors) and a few clothing basics (T-shirts, socks, underwear).

- **Pakistan import tariffs** for clothing (women vs men) from the official tariff book.

Where online to look (all have public product pages you can read/scrape):

- **Daraz.pk** (huge marketplace). Daraz+3Daraz+3Daraz+3
- **Carrefour Pakistan** (grocery + personal care). Carrefour+2Carrefour+2
- **Al-Fatah** (national chain with an online store). Al-Fatah+2Al-Fatah+2
- **Metro Pakistan** (cash & carry; online catalog). Metro+2Metro Online+2
- **Imtiaz** (big local chain; online presence varies by city). Imtiaz+2Imtiaz+2

Official **Pakistan Customs Tariff** (the law book with HS codes & rates, often called **PCT** for "Pakistan Customs Tariff"): latest PDFs on the **FBR** site. You'll need **Chapter 61 (knit)** and **Chapter 62 (woven)** for clothing.

Tables you keep

- **products**: product_id, retailer, category, brand, title, gender_target (female/male/unisex), size_value, size_unit, url
- **prices**: product_id, date, price, sale_price, in_stock, unit_price_per_100g_ml (or per piece)
- **pairs**: pair_id, female_product_id, male_product_id, how_we_matched (strict/nearest)
- **hts_lines**: hs6, description, gender_label (men/women), mfn_rate, chapter (61/62)

Exactly how to collect the data (step by step)

A) Retail price data

1) Pick categories & pages

- **Everyday (CPG):** deodorant/antiperspirant, shampoo, body wash/soap, lotion, razors.
- **Clothing basics:** cotton T-shirts, socks, underwear (briefs).
- Find the category pages on each site and note their **URL patterns** (Daraz filters; Carrefour/Al-Fatah category pages). Al-Fatah+8Daraz+8Daraz+8

2) For each product page, save these fields

- retailer (daraz, carrefour, alfatah, metro, imtiaz)
- category (e.g., deodorant)
- brand (e.g., Dove, Fa)
- title (product name as shown on page)
- gender_target (female / male / unisex / unknown)
  - **How to label simply:** If the page text says **"women/ladies/her/for her/for women/خواتین/لیڈیز"** ➜ **female**. If it says **"men/gents/his/for men/مرد/جینٹس"** ➜ **male**. Otherwise **unisex**.
- pack_size_value and pack_size_unit (e.g., 200, ml)
- price (PKR) and, if shown, sale_price
- in_stock (true/false)
- product_url, image_url
- observed_at (date of scrape)

Tip: On marketplaces like **Daraz**, always capture the **variant** (e.g., 150 ml vs 200 ml); on grocery sites, capture the **per-piece** for socks/underwear multipacks.

3) Normalize the units (so prices are comparable)

- **Liquids/creams** (shampoo, body wash, lotion): compute **unit_price_per_100ml**.
- **Solids** (deodorant sticks/bars): **unit_price_per_100g**.
- **Razors**: **price per cartridge/blade**.
- **Clothing**: **price per piece** (one T-shirt, one pair of socks, one underwear).
- Save the computed unit price as a column (e.g., unit_price).

4) Make fair pairs (women vs men) for analysis

- **Strict match first**: same brand + same/similar size (±10%).
- If the brand doesn't have both genders, you can match via **nearest features** (same category, similar size, similar product type). Keep a match_method column: strict or nearest.

5) Keep it clean & bilingual

- Some titles are in **Urdu/English mix**. Keep the raw title, but store your **gender_label** using your rule above.
- Do a **quick manual check** of ~100 random items to see if your gender labels look right; adjust the keyword list if needed.

B) Pakistan tariff data (PCT / HS)

1) Get the official tariff book

- Download the **Pakistan Customs Tariff** (latest FY) from **FBR**. You'll use **Chapter 61 & 62** for clothing. Federal Board of Revenue+3FBR Download+3FBR Download+3

2) Build a tiny table of clothing lines you care about

For each basic garment type, find the **HS6** lines that split **men/boys** vs **women/girls** and note their **ad-valorem rate**:

- **T-shirts** (cotton knit — Chapter 61)
- **Socks** (knit — Chapter 61)
- **Underwear** (knit — Chapter 61)

If there's more than one HS6 for the same thing (e.g., different material blends), pick the most common (like 100% cotton) or keep a few and tag **material**.

Columns to keep:

- hs6, description, gender_label (men/boys or women/girls), mfn_rate_percent, chapter (61/62), garment_type (tee/sock/underwear)

3) Compute TariffDiff per garment type

For each garment type, compute:

- TariffDiff_pp = (women_rate − men_rate) in **percentage points**

(Example: women 20% vs men 15% ➜ **+5 pp**)

Collection & cleaning pipeline

Scraping

- Playwright scripts per retailer + category search URLs.
- Parse pack size to numeric (regex + unit converters).
- **Gender targeting** detection:
    - Rules: keywords in title/breadcrumbs ("women", "her", "men", "his", "ladies", "boys/girls" for apparel).
    - Color words *don't* determine gender; keep only if explicitly stated.
    - Manual audit: sample 100 items for precision/recall, adjust rules.

Normalization

- Unit price: convert to per 100 g / mL; for razors, per blade or per cartridge; for deodorant, per g.
- Apparel sizes: restrict to one common size band (e.g., men M vs women M equivalent) or normalize by area proxy (but keep scope tight by focusing on socks/tees/underwear where size mapping is easy).

Matching

- **Strict brand+size matcher**: same brand & near-equal size within 5–10% (CPG).
- **Hedonic nearest-neighbor**: TF-IDF on title + key attributes; cosine similarity threshold for non-identical brands where features are comparable.
- Create pairs table; keep top-1 match per female SKU.

HTS processing

- Download HTS (ch. 61 & 62), parse rows with gendered descriptors; extract MFN ad valorem. hts.usitc.gov
- Aggregate to **garment type bins** (tees, underwear, socks) and compute women–men **tariff differential** per type. Cross-check context with USITC/PPI summaries. usitc.gov+1

Models

Model A — Matched Pairs (for RQ1)

- **Idea:** Compare **apples to apples**: pair a women's product with the **closest men's** version.
- **Example pair:** Dove deodorant 75g "for women" vs Dove 76g "for men".
- **Metric:** % difference =

  $$\text{\% gap} = 100 \times \big(\ln(\text{price per g}_\text{women}) - \ln(\text{price per g}_\text{men})\big)$$

- **Output:** an average "pink premium" by category with error bars.

Model B — Hedonic Regression (still RQ1, more careful)

· **Idea:** Price depends on features. We control for them to see the **"just being for women"** effect.

· **We predict:** unit price using features like brand, size, ingredients, scent, retailer, week.

· **Key variable:** FemaleTargeted (1 if the product literally says "women/for her/ladies").

· **Interpretation:** the FemaleTargeted coefficient tells you the **average premium** for women-targeted items **after** adjusting for features.

Model C — Tariff → Price Gap Link (for RQ3)

- **Idea:** For clothing, compute the **tariff difference** (women minus men) by garment type (e.g., tees).
  Then see if **pairs' store price gaps** are **bigger** in garment types with **bigger tariff differences**.
- **We run:**

$$\text{PriceGap(pair)} = \alpha + \beta \times \text{TariffDiff(garment type)} + \text{controls}$$

Read β: if $\beta > 0$, categories with bigger women>men tariff gaps also show bigger women>men price gaps—suggesting **pass-through** of the tax into prices.
(We'll call this **correlational**, not proof of causality.)

The Methods

Part A: Everyday products (CPG) — find the pink premium

1. **Collect prices** weekly for 5 categories (razors, shampoo, body wash, lotion, deodorant) from 2–3 big retailers.
   Save: title, brand, size, price, sale price, category breadcrumb, URL, timestamp.

2. **Clean & normalize:** convert all sizes to the **same unit** (e.g., price per 100 mL or per 100 g).

3. **Label gender:** if the page says "women/ladies/her" → **female**; "men/his" → **male**; else **unisex**.

4. **Make pairs:** for each women's item, find the **closest men's** item (same brand and similar size if possible).

5. **Compute gaps:** for each pair, compute the **% price gap**.

6. **Summaries:** show average gap by category and retailer; then run the **hedonic** model to adjust for features.

Part B: Clothing tariffs and prices — see if taxes line up with price gaps

1. **Tariff table:** download HTS chapters **61** (knit) and **62** (woven); find the **HS6** clothing lines that say men/boys vs women/girls; record the **tariff rates**.

2. **Tariff differences:** for each garment type (tees, socks, underwear), compute **women minus men tariff %**.

3. **Price pairs (small sample):** scrape prices for **basic items** where matching is easy (e.g., same brand cotton tees: women's vs men's). Normalize by piece (per T-shirt, per pair of socks).

4. **Link test:** run Model C to see if **bigger women>men tariff gaps** go with **bigger store price gaps**.

5. **Reality checks:** run a **placebo** on a no-tariff category (e.g., deodorant) where tariff diff ≈ 0 → expected β ≈ 0.

Visuals & deliverables

- **Dashboards** :
  - PPP by category/retailer; drill-down to brand.
  - Tariff differentials by garment type with interactive HS lines.
- **Key figures for paper/poster:**
  - Violin/box of PPP per category.
  - Hedonic coefficient plot (FemaleTargeted with 95% CI).
  - Bar chart: average MFN women vs men per garment.
  - Scatter: TariffDiff vs. mean within-pair price gaps (with regression line).