Laptop Recommender System

Smart Laptop Selection: A Data-Driven Recommendation System
From Web Scraping to Personalized Recommendations

Final project for DATA 607, CUNY SPS

Supervised by: Andrew Catlin

Presented by: Mehreen Ali Gillani
Date: 12-17-2025

The Problem

Can we build an intelligent system that understands user needs and recommends the perfect laptop?"

- Information overload: 1000+ laptop models online
- Technical jargon confusion: GHz, RAM, SSD, GPU, etc.
- Price vs. performance dilemma
- User needs vary greatly (student vs. gamer vs. professional)
- Time-consuming research process

Solution Overview

Web scraping: Real-time market data

Machine Learning: Feature engineering

Interactive Shiny App: User-friendly interface

Web scrapping https://serpapi.com

Data Points Collected:

==========

📈 AMAZON DATA COLLECTION SUMMARY

==========

📦 Unique products: 943

💰 Price range: $13.99 - $7678.99

💰 Average price: $701.18

⭐ Products with ratings: 870

🏷️ Unique brands: 19

🏷️ TOP 5 BRANDS:

HP: 288 products (30.5%)

Lenovo: 145 products (15.4%)

Dell: 128 products (13.6%)

Unknown: 94 products (10.0%)

ASUS: 90 products (9.5%)

==========

📊 WALMART DATA COLLECTION SUMMARY

==========

📦 Products collected: 615 (removed 285 duplicates)

💰 Price range: $47.99 - $7506.24

💰 Average price: $721.35

⭐ Average rating: 2.69/5.0

🏷️ Unique brands: 10

📊 DATA SUMMARY:

Total products: 1558

Products with ratings: 1485

Products with prices: 1558

Products with reviews: 1558

🛒 **SOURCE DISTRIBUTION:**

Amazon: 943 products (60.5%)

Walmart: 615 products (39.5%)

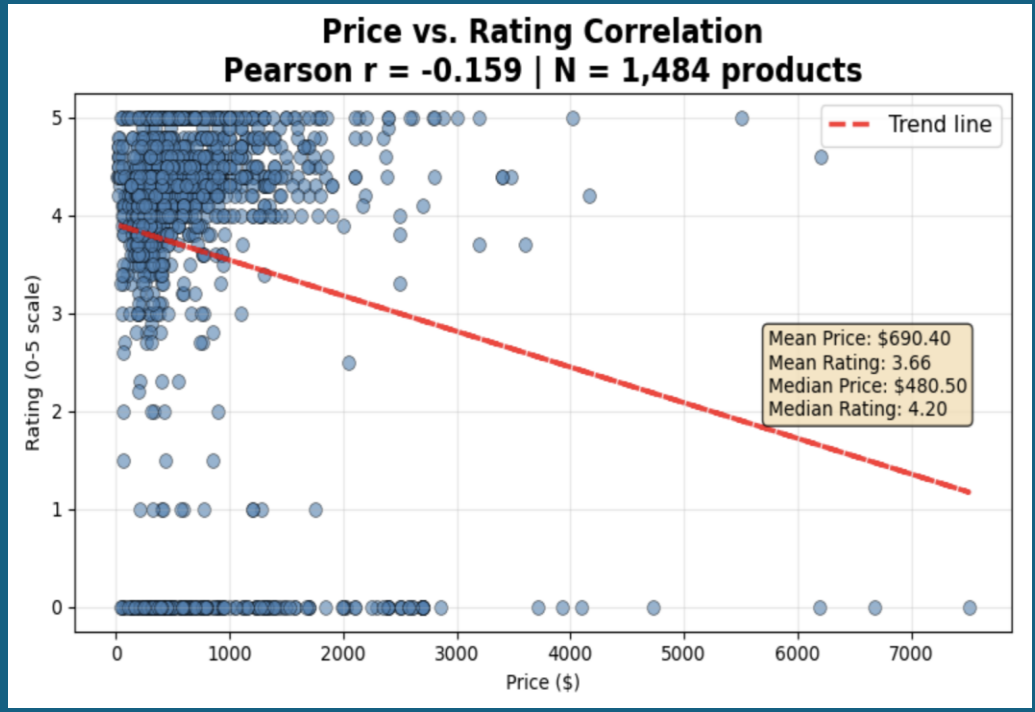
Average price: $708.67

Average rating: 3.66

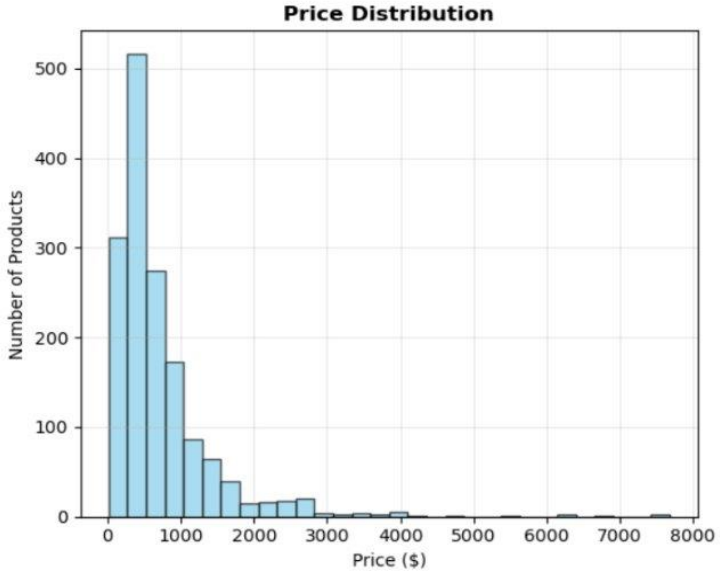
Price-Rating correlation: *-0.159*

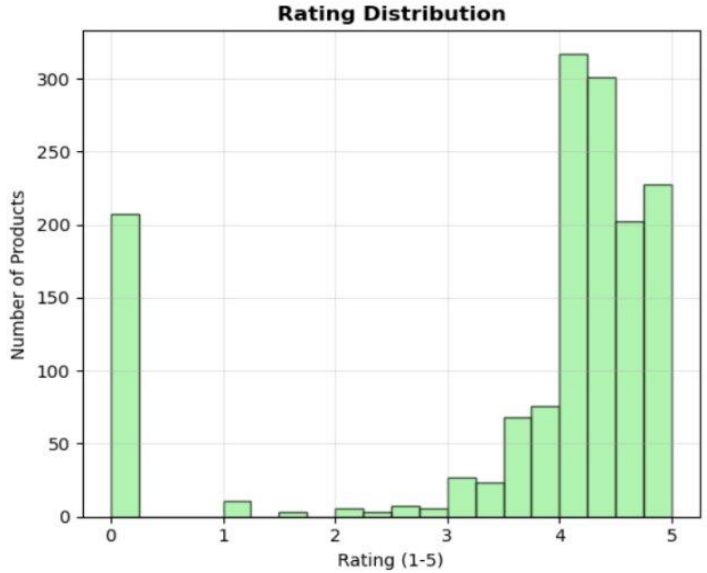
- INSIGHT: Price and rating show weak correlation

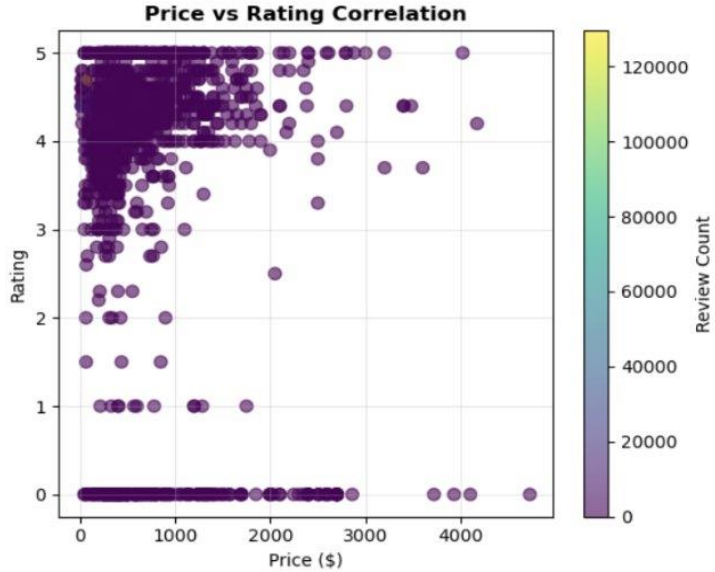
['product_id', 'source', 'position', 'sponsored',
'brand', 'title', 'rating', 'reviews', 'price',
'old_price', 'delivery', 'free_shipping','in_stock', 'seller', 'climate_pledge_friendly', 'product_category']

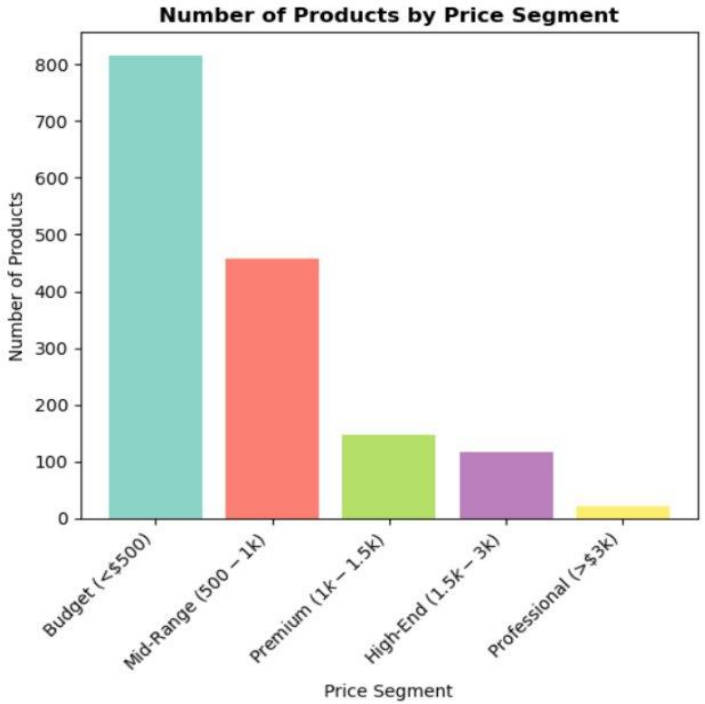


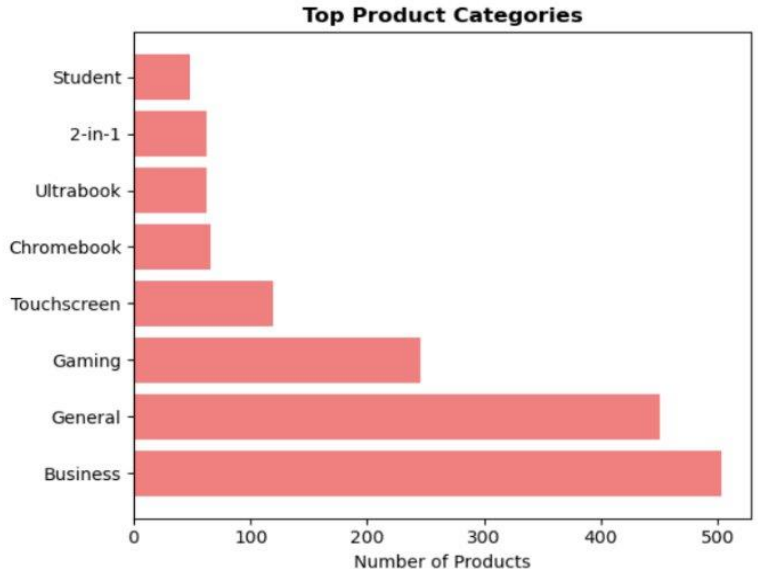
Amazon & Walmart Laptops - Descriptive Analytics Dashboard











Data Cleaning Challenges

1. Price Data Cleaning

- Converted $0 prices to NA (missing values)
- Removed invalid price entries (≤ 50)

2. Rating & Review Standardization

- Converted "N/A" and empty ratings to NA
- Parsed review counts (removed commas, handled text variations)

3. Brand Name Standardization

- Trimmed whitespace, replaced "N/A" brand entries with "Unknown"
- Applied case standardization:
- Examples: "ACER" → "Acer", "asus" → "ASUS"
- Consolidated brand variations for major manufacturers

4. Specification Extraction *(from product titles)*

- **RAM Detection**: Extracted GB values using regex patterns
- **Storage Capacity**: Priority-based extraction algorithm for SSD/HDD storage
- **Display Size**: Pattern matching for screen dimensions (8-24 inch range)
- **Processor Identification**: Classified Intel Core i3/i5/i7/i9, AMD Ryzen series, Apple M-series
- **OS Classification**: Windows 11/10, macOS, Chrome OS detection

5. Product Feature Flagging

- **Refurbished Products**: Flagged renewed/restored items
- **Gaming Laptops**: Identified gaming-specific keywords (ROG, Alienware, etc.)
- **2-in-1 Convertibles**: Detected convertible/touchscreen models

6. Derived Feature Creation

- **Value Metrics**:
  - RAM in GB
  - Storage in **GB**
- **Popularity Tiers**: Based on review counts

•**Product Categories**: Gaming, Workstation, Standard, Apple Premium classification

7. Quality Control & Validation

- Extraction success rate tracking
- Range validation for numeric features
- Handling missing data appropriately
- Cross-source consistency checks

8. Extraction success rates:

- RAM: 92.8%
- Display: 87.6%
- Processor: 92.1%
- Gaming laptops: 246
- 2-in-1 laptops: 337
- refurbished laptops: 279

Feature Engineering

Objective: Create actionable business insights from cleaned data

1. Customer-Centric Rating Classification

- **Rating Categories** for intuitive interpretation:
  - **Poor** (< 3.5 stars)
  - **Average** (3.5 - 4.0 stars)
  - **Good** (4.0 - 4.5 stars)
  - **Excellent** (4.5 - 5.0 stars)
- **Popularity Index**: Categorized based on review counts:
  - **No Reviews**
  - **Few Reviews** (< 10 reviews)
  - **Some Reviews** (10-99 reviews)
  - **Popular** (100-999 reviews)
  - **Very Popular** (≥ 1000 reviews)

2. Product Type Classification

- **Workstation**: High RAM (≥ 32GB) systems
- **Premium Laptop**: Price > $1,500
- **Budget Laptop**: Price < $500
- **Standard Laptop**: Middle-ground category

3. Price Category:

Created price segmentation categories:

- Budget (< $500)
- Mid-Range ($500 - $1k)
- Premium ($1k - $1.5k)
- High-End ($1.5k - $3k)
- Luxury (> $3k)

4. Brand Categorization

- **Brand Tier System**: Converted brand categories into simplified tiers:
  - **Premium** → HP, Lenovo, Dell, ASUS, Apple, Microsoft, Samsung, LG
  - **Mid-Tier** → Acer, MSI, Razer, Gateway, Alienware
  - **Budget** → Chinese brands (NIMO, CHUWI, ZOLWAYTAC, etc.)
  - **Generic** → Unknown/no-name brands

5. Price-Performance Value Metrics

Storage Extraction: 91.8%

- **Price per GB Storage**: Calculated `price_per_gb_storage` = Price ÷ Storage (GB)
- **Price per GB RAM**: Calculated `price_per_gb_ram` = Price ÷ RAM (GB)
- **Robust Error Handling**: Includes validation for:
  - Zero/negative prices
  - Missing storage/RAM values
  - Invalid calculations
- **Business Interpretation**: Lower values indicate better price-performance ratio

6. Retailer Comparison Features

- **Amazon vs. Walmart** differential analysis:
  - Average price comparison
  - Product category distribution
  - Specification differences
  - Rating distribution

Exploratory Data Analysis (EDA)

Objective: Uncover patterns, differences, and opportunities in Amazon vs. Walmart laptop markets

source	products	avg_price	avg_rating	avg_ram	avg_storage	premium_brands_pct	gaming_pct	convertible_pct
Amazon	924	715.01	4.34	20.1	949	80.2	15.6	24.6
Walmart	609	726.87	2.69	18.7	712	70.0	16.7	18.1

KEY INSIGHTS

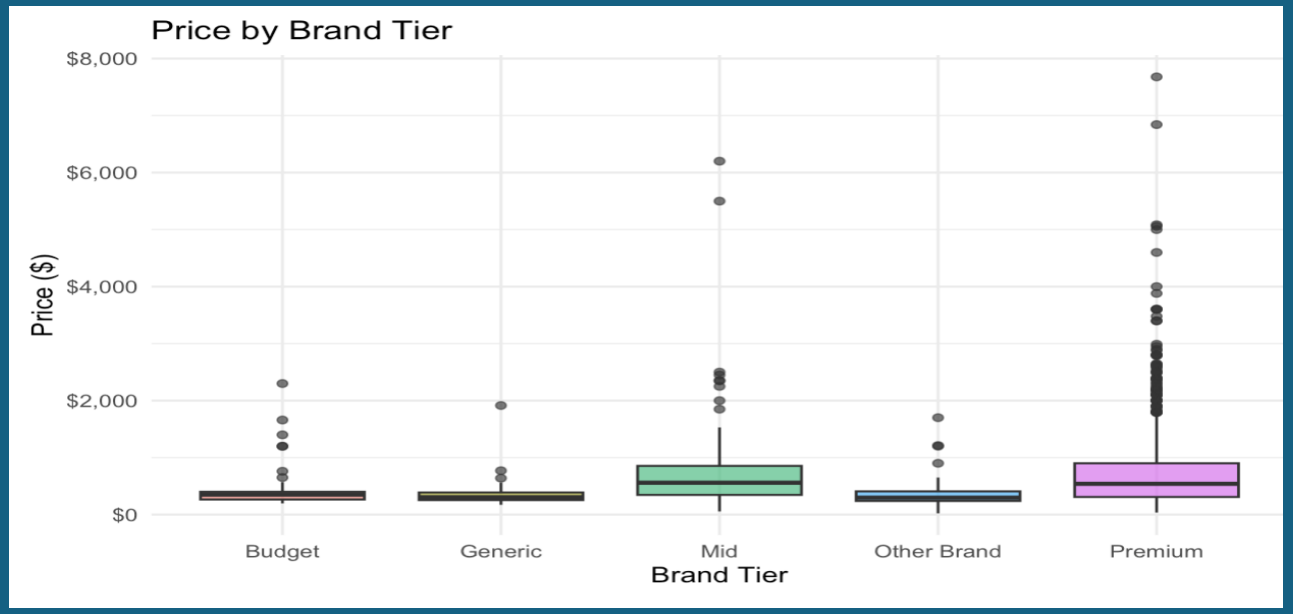
Amazon:

Much higher average rating (4.34 vs. 2.69),

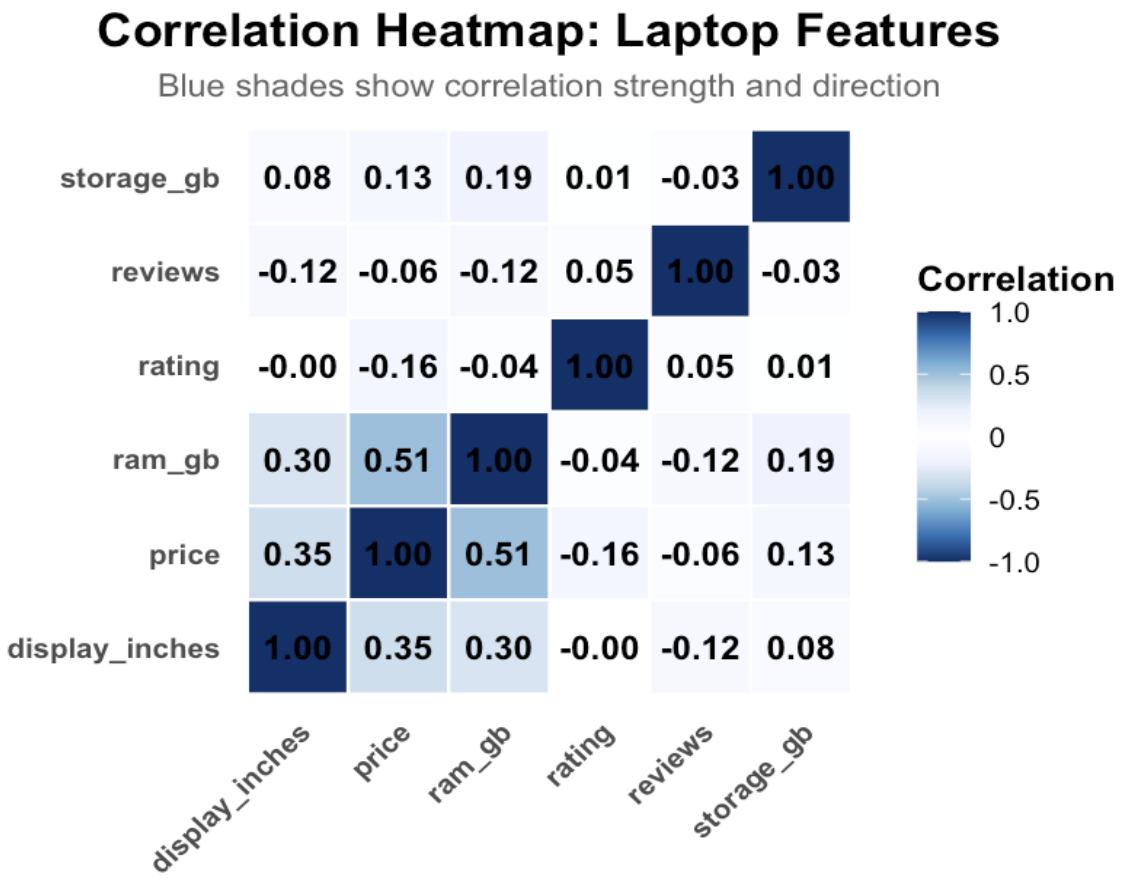
Greater share of premium brands (80.2% vs. 70.0%)

70.0%

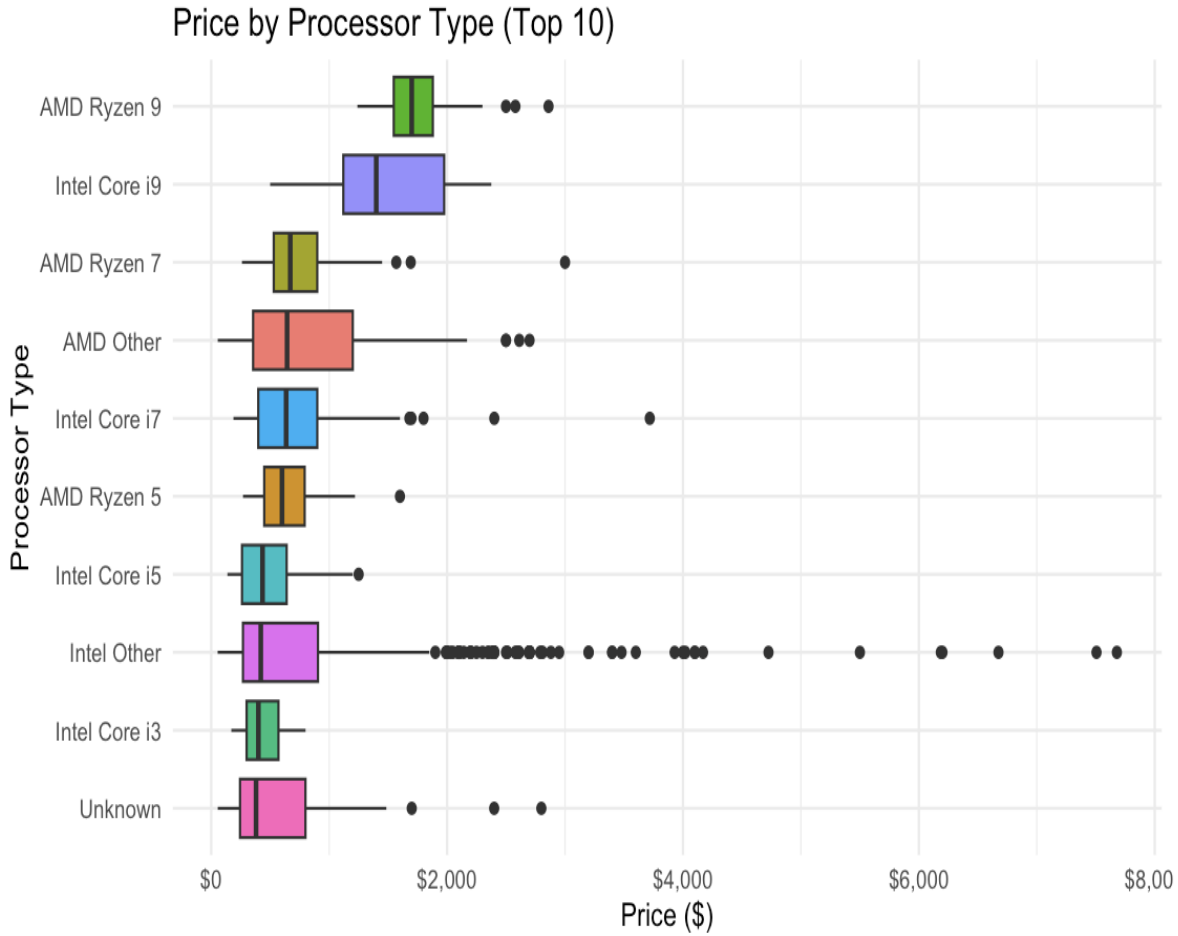
Best value: $1.24 per GB RAM



Correlation Heatmap:



Price by Processor Type



Personalized Laptop Recommendation System

Step 1: Define Case Weight

```
CASE_WEIGHT = {
  "Gaming": 1.5,
  "Student": 1.2,
  "2-in-1": 1.3,
  "General": 1.0
}
```

Step 2: Dynamic Budget Range

BUDGET_MIN = User_Budget × 0.8

BUDGET_MAX = User_Budget × 1.2

Step 3: Core Filters

FILTERED_POOL = All_Laptops WHERE:

- Price ∈ [BUDGET_MIN, BUDGET_MAX]
- Rating ≥ User_Min_Rating
- RAM ≥ User_Min_RAM
- Brand = User_Preference (if specified)

Phase 2: Intelligent Fallback (If Results < 5)

EXPANDED_POOL = All_Laptops WHERE:

- Price ∈ [Budget×0.7, Budget×1.5]
- Rating ≥ MAX(3.0, User_Rating×0.9)
- RAM ≥ MAX(4GB, User_RAM×0.7)
- Apply Case-Specific Filters
- Remove Duplicates

Phase 3: Multi-Dimensional Scoring

Component Scores (0-1 Scale)

1. PRICE_SCORE = 1 - (|Actual_Price - Budget| ÷ Budget)
   → +10% bonus if price ≤ budget

2. RAM_SCORE = (RAM_GB - Min_RAM_in_Set) ÷ (Max_RAM_in_Set - Min_RAM_in_Set)

3. STORAGE_SCORE = {
   if missing: 0.3
   if <128GB: 0.4
   if ≥2000GB: 1.0
   else: Storage_GB ÷ 2000 }

4. RATING_SCORE = Rating ÷ 5

5. PROCESSOR_SCORE = {
   "High-End": 1.0,
   "Performance": 0.8,
   "Mid-Range": 0.6,
   "Entry-Level": 0.4,
   default: 0.3 }

Phase 4: Weighted Final Score

```
FINAL_SCORE =
  (PRICE_SCORE × Weight_Price) +
  (RAM_SCORE × Weight_Performance × 0.6) +
  (PROCESSOR_SCORE × Weight_Performance × 0.4) +
  (STORAGE_SCORE × Weight_Features) +
  (RATING_SCORE × 0.3)

FINAL_SCORE_NORMALIZED = 100 × (FINAL_SCORE ÷
Max_FINAL_SCORE)

VALUE_METRIC = ((RAM_GB × $8) + (Storage_GB ×
$0.05)) ÷ Price
```

USER INPUT
- Budget: $1,200
- Use Case: Gaming
- Priority: Performance (70%)

▼

SMART FILTERING
- Gaming flag = TRUE
- RAM ≥ 16GB
- Budget range: $960-$1,440
- Case weight = 1.5×

▼

SCORING ENGINE

Price: 0.92 (under budget)
RAM: 0.85 (32GB vs 16-64GB)
Processor: 1.0 (High-End)
Storage: 0.9 (1.8TB SSD)
Rating: 0.94 (4.7/5)

▼

WEIGHTED CALCULATION

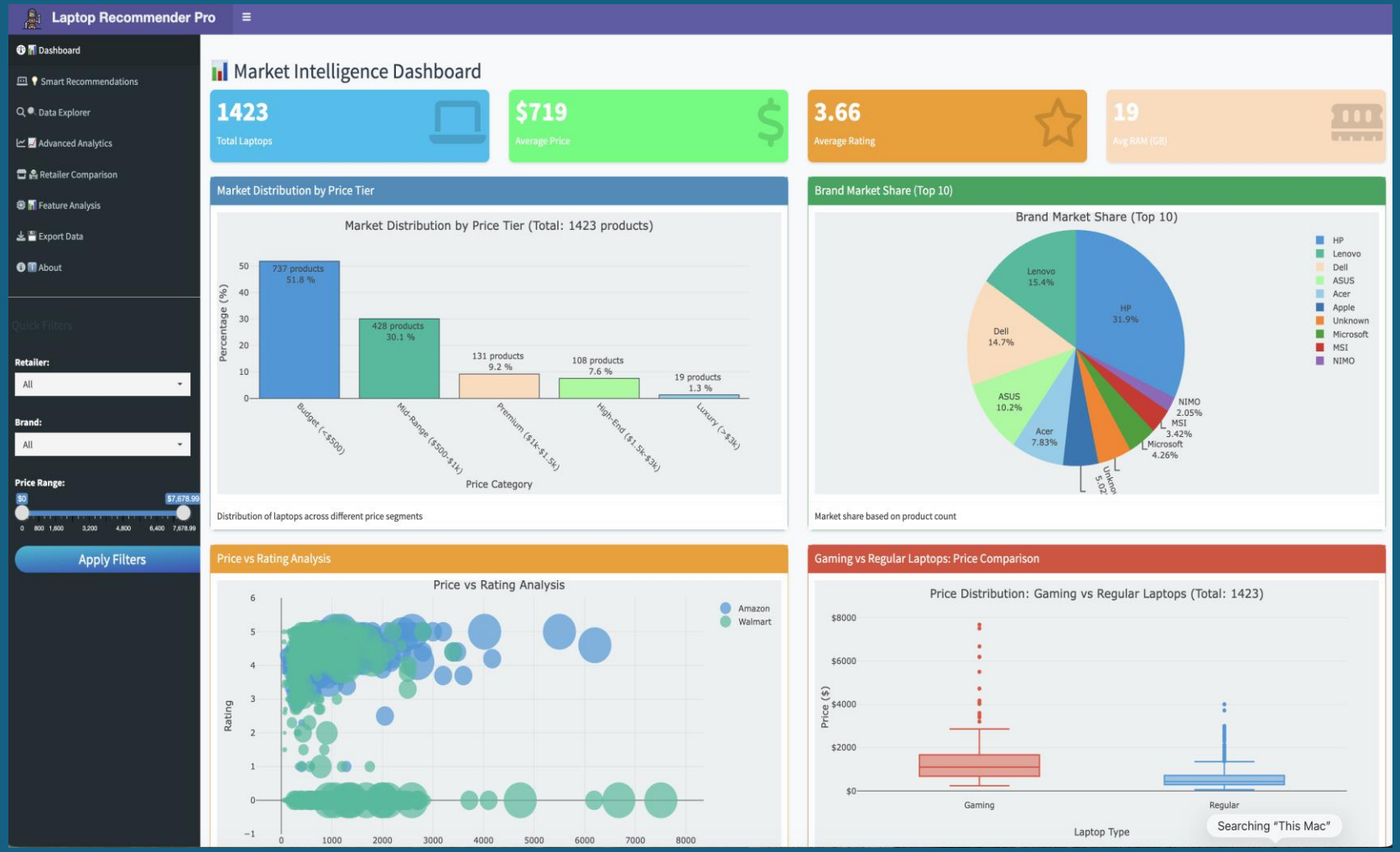
0.92×0.15 + 0.85×0.42 + 1.0×0.28 + 0.9×0.15 + 0.94×0.3 = 0.89

×1.5 (gaming weight) → FINAL: 1.34

▼

NORMALIZATION & RANKING
- Final score: 96.2/100
- Value score: $304 ÷ $1,199 = 0.254
- Recommendation: TOP MATCH ✓

Shiny App Demo



FInal Project DATA-607
Laptop Recommendation system

Conclusion

Challenges Faced During Project:

1. **Data Cleaning from Web Scraping:** Raw scraped data required significant cleaning to handle inconsistencies, missing values, and varied formatting across different product pages.
2. **Mastering R Shiny:** Developing the interactive front-end involved a steep learning curve to understand Shiny's reactive programming model and UI/server architecture.
3. **API Limitations (SERPAPI):** Access to real-time search data was constrained by SERPAPI's usage limits and costs, requiring careful quota management and efficient data caching strategies.

Future Enhancement:

- Broader Price & Specs Comparison
- Driving Better Value
- DATA from other retailers