Smartphone Pricing Analysis on Amazon (2005–2025)

1.Project Overview

This is a self-directed data analysis project aiming to explore what factors influence smartphone pricing on Amazon over the past two decades. Starting from a broad curiosity — why some phones are cheap but highly rated, while others are expensive with mixed reviews — I decided to investigate whether configuration or brand has more explanatory power on smartphone prices.

The study involved collecting real-world data, cleaning it, conducting exploratory analysis, running regression models, and designing a value-based scoring system to evaluate smartphone cost-effectiveness. This project also helped me practice data interpretation, model evaluation, and presentation.

2.Data Source & Features

brand	price_USD	rating_out_of_5_stars	memory_storage_capaci	ram_GB	standing_screen_displ	ram_size	ram_size	ram_category	memory_storage_size	screen_ _size(inches)
CUBOT Store	299.99	4.6	256	32	6.58		32	>16GB	128-256GB	6.5-6.7inch
DOOGEE Store	239.99	4.3	256	24	6.6		24	>16GB	128-256GB	6.5-6.7inch
AGM Store	699.99	3.4	128	8	6.53		8	8-16GB	64-128GB	6.5-6.7inch
NUU Store	139.99	4.5	128	6	6.7		6	4-6GB	64-128GB	6.5-6.7inch
Plum	39	3.1	32	4	2.4		4	4-6GB	32-64GB	<5inch
OSCAL Store	189.99	4.4	256	24	6.67		24	>16GB	128-256GB	6.5-6.7inch
CUBOT Store	279.99	3.5	512	24	6.58		24	>16GB	>256GB	6.5-6.7inch
artfone Store	53	3.8	0.13	0.046875	1.8		0.046875	<1GB	<32GB	<5inch
Straight Talk	144.83	4.5	64	64	4.7		64	>16GB	32-64GB	<5inch
OUKITEL Store	169.99	3.8	256	12	6.6		12	8-16GB	128-256GB	6.5-6.7inch
Blackview Store	97.31	3.8	32	4	6.5		4	4-6GB	32-64GB	6-6.5inch
Gabb Store	199.99	4.2	128	4	6.5		4	4-6GB	64-128GB	6-6.5inch
OUKITEL Store	139.98	4.2	128	12	6		12	8-16GB	64-128GB	5-6inch

Amazon Data head(1)

The dataset includes 1,000+ smartphone entries scraped from Amazon (2005–2025). Key columns:

Product Info: Title, Brand, Model, Price

Specifications: RAM (GB), Storage (GB), Screen Size (in)

User Feedback: Rating (0–5), Number of Reviews

Others: Some product metadata like image URL, description text (not used in this version)

I cleaned and standardized the columns to make them machine-readable. RAM, storage, and screen were transformed into numeric values; prices were cleaned and converted to USD.

3.Methodology

The project followed three stages of analysis:

3.1. Descriptive Statistics

1) Explored price distribution, rating patterns, and review counts.
2) Identified dominant brands and popular configurations.
3) Goal: Understand general trends and anomalies.
4) Benefit: Offers an intuitive entry point.
5) Limitation: Cannot quantify relationships.

3.2. Multivariate Regression (with & without brand)

1) Objective: Quantify how much each variable (RAM, storage, screen, rating, brand) affects price.
2) Step 1: Run regression without brand → R^2 = 0.019
3) Step 2: Add brand as a categorical variable → R^2 = 0.34
4) Interpretation: Configuration alone explains little; brand strongly influences pricing.
5) Benefit: Simple, interpretable model.
6) Limitation: Assumes linear relationships; ignores interaction effects and unobserved factors (e.g., launch date).

3.3. Value Score Ranking

- Question: Are high-rating low-cost phones really a good deal?
- Formula: **Value Score = (Rating / Price) × log(Review Count + 1)**
- Logic: Reward high rating, discount high price, boost high review count (but log-scaled).
- Result: Top-scoring phones were often lesser-known brands — suggesting that "value" phones may be underrated.
- Benefit: Translates multiple features into one interpretable metric.
- Limitation: Still correlational, sensitive to fake reviews.

4.Key Findings

1. **Brand matters more than specs:** Brand alone accounts for over 30% of price variance; RAM, storage, screen size add little explanatory power.
2. **Low-price phones can have high ratings**: Some underdog phones with decent specs and excellent reviews rank high in value score.

3. **Popular phones are still big brands:** When limiting to high-rating (4.5+) and high-review (100+) phones, they are mostly from Samsung, Xiaomi, and Huawei.

5.Limitations & Future Directions

1. **Sample Bias:** Data is from Amazon, which differs from global/mobile-dominant markets like China. Apple is underrepresented, so conclusions don't generalize.

2. **Time Span:** Spanning 20 years introduces noise from tech evolution. E.g., 1GB RAM in 2005 and in 2025 has vastly different meaning.

3. **Omitted Variables:** Launch year, OS, market tier (flagship vs. budget), and locked/unlocked status are missing.

4. **Model Assumptions:** Linear regression assumes constant marginal effects (e.g., 1GB RAM = +\$10). Likely oversimplified.

5. **Correlation vs Causation:** This project shows co-movements but not causality. Brand reputation and marketing may drive both price and rating.

6. **Review Manipulation:** Rating and review counts can be faked, introducing noise into regression and value score.

Suggested Improvements

- Narrow scope to 2020–2025 data for more relevant findings.
- Add controls for release year, model tier, and OS.
- Try polynomial regression or random forest for non-linear relationships.
- Apply NLP to analyze review content or product descriptions.