📌 Domain: Pharmaceutical Product Data Analysis

The pharmaceutical domain involves the development, production, pricing, and distribution of medicines. Each product varies based on **form type** (tablet, syrup, suspension, cream, injection), **composition**, **strength**, **pack size**, and **manufacturer**.
Understanding these characteristics allows us to analyze **price variations**, **drug patterns**, and **market behavior**, which is essential for decision-making in healthcare and pharmaceutical retail.

📌 Use Case: Price Optimization and Drug Comparison

This project aims to analyze the pricing behavior of pharmaceutical products and identify:

- How medicine **price varies** across form types, pack sizes, manufacturers, and drug compositions.
- Which medicines and pack sizes are **most expensive** and **least expensive**.
- How **strength**, **composition**, and **form type** influence drug pricing.
- Patterns and outliers that can aid in **price optimization**, **drug comparison**, and **market segmentation**.

📌 1. Data Understanding

Dataset Overview

- **Source:** opendatabay.com
- **Total Rows:** 253,973
- **Total Columns:** 9

Columns

- **id** – Unique identifier
- **name** – Medicine name
- **price(₹)** – Price
- **Is_discontinued** – Whether the drug is discontinued
- **manufacturer_name** – Producing company
- **type** – Allopathy / Non-allopathy
- **pack_size_label** – Packaging details (e.g., strip of 10 tablets)
- **short_composition1** – Main active ingredient
- **short_composition2** – Secondary ingredient

Data Types

- **Numerical:** price(₹)
- **Categorical:** All other columns
- **Missing Values:** Mostly in short_composition2 (filled with "None")

Initial Statistics of Price

- **Min:** ₹0
- **Max:** ₹4,36,000
- **Mean:** ₹270
- **Median:** ₹79
- **Std Dev:** Very high due to extreme outliers

SAMPLE DATA OF THE indian_medicine_data.csv

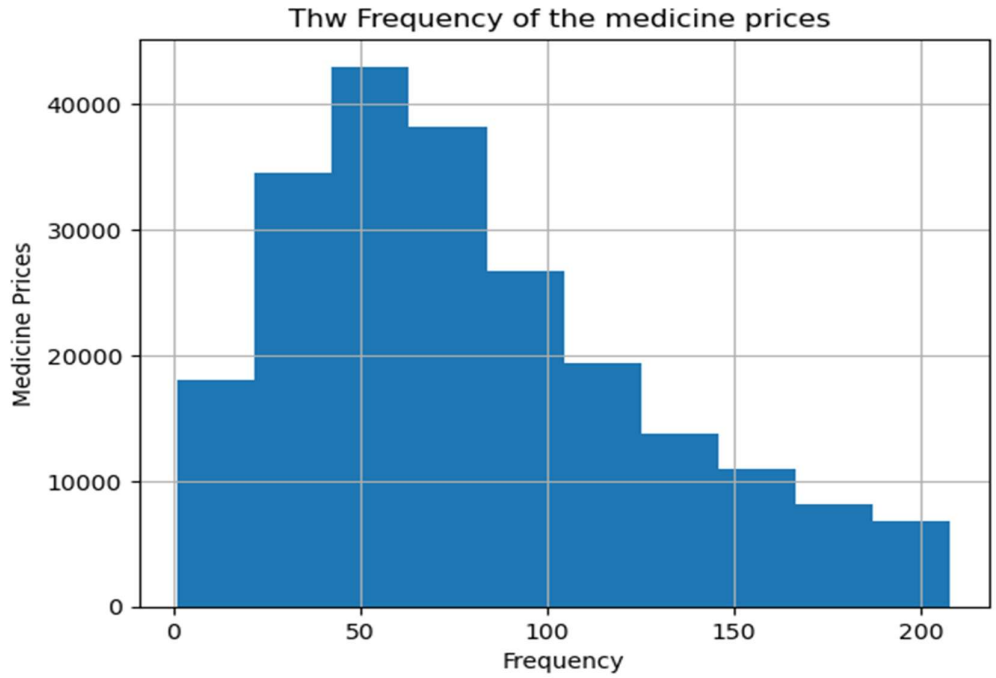
id	name	price (₹)	Is_discontinued	manufacturer_name	type	pack_size_label	short_composition1	short_composition2	
0	1	Augmentin 625 Duo Tablet	223.42	False	Glaxo SmithKline Pharmaceuticals Ltd	allopathy	strip of 10 tablets	Amoxycillin (500mg)	Clavulanic Acid (125mg)
1	2	Azithral 500 Tablet	132.36	False	Alembic Pharmaceuticals Ltd	allopathy	strip of 5 tablets	Azithromycin (500mg)	NaN
2	3	Ascoril LS Syrup	118.00	False	Glenmark Pharmaceuticals Ltd	allopathy	bottle of 100 ml Syrup	Ambroxol (30mg/5ml)	Levosalbutamol (1mg/5ml)
3	4	Allegra 120 tablet	218.81	False	Sanofi India Ltd	allopathy	strip of 10 tablets	Fexofenadine (120mg)	NaN

📌 2. Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

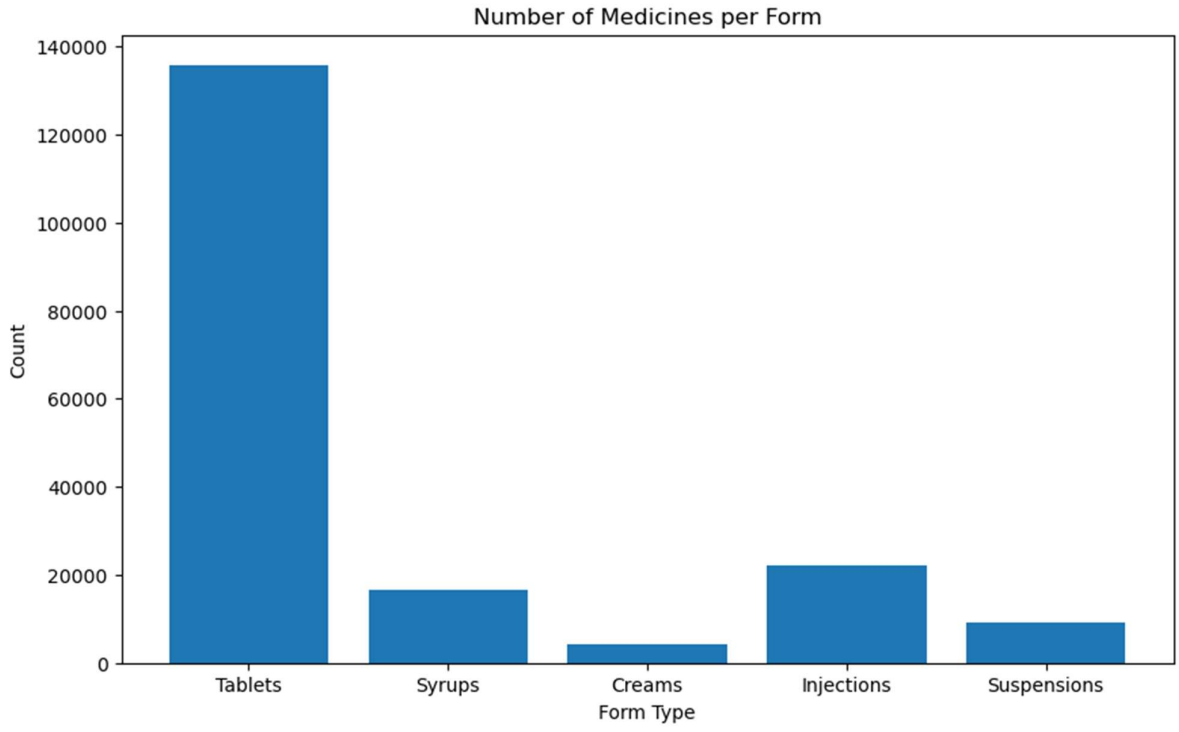
Histogram of Prices

- Highly **right-skewed** distribution.
- Most medicines cost **₹50–₹200**.



Bar Graphs of Categorical Columns

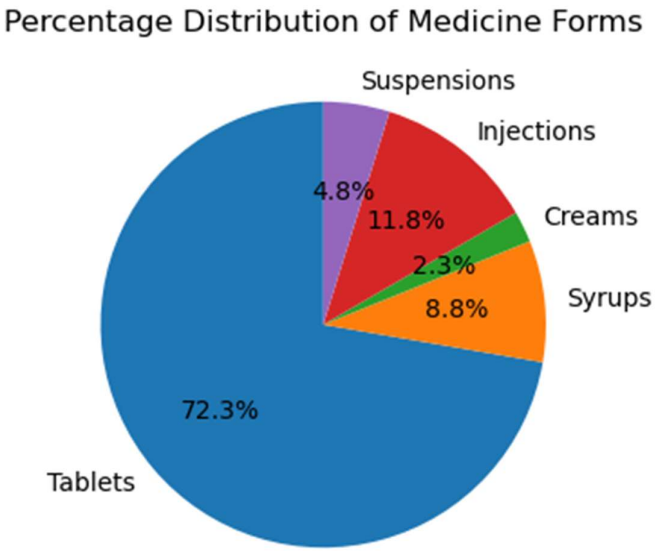
- **Tablets** are the most common form.



-

Pie Charts

- Tablets dominate the dataset.
- Syrups, creams, and suspensions follow.



4.2 Outlier Detection

- Boxplots reveal strong outliers in price.
- Outliers are mostly **high-cost injections**.
- Removed using IQR for cleaned analysis, but noted as **valid domain values**.

CODE:

```
Q1 = df['price(₹)'].quantile(0.25)
Q3 = df['price(₹)'].quantile(0.75)
IQR = Q3 - Q1

lower_limit = Q1 - 1.5 * IQR
upper_limit = Q3 + 1.5 * IQR

df = df[(df['price(₹)'] >= lower_limit) & (df['price(₹)'] <= upper_limit)]
```

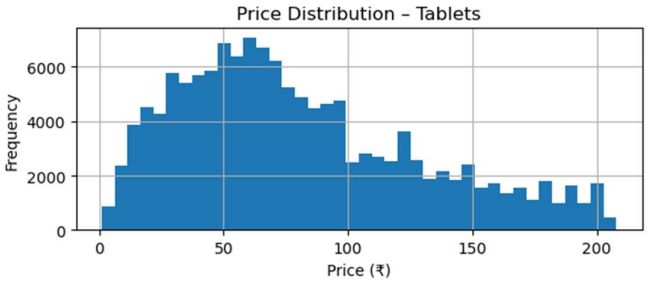
4.3 Handling Missing Values

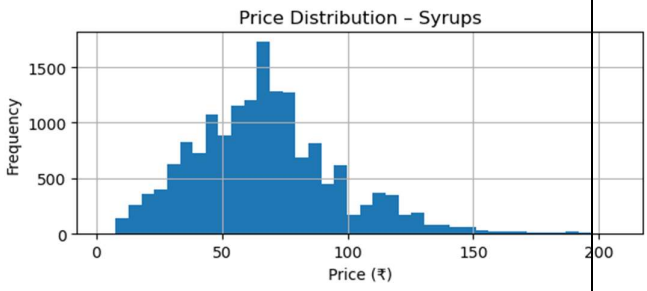
- **short_composition2** had many NaN values → replaced with "None".
- df['short_composition2'] = df['short_composition2'].fillna("None")

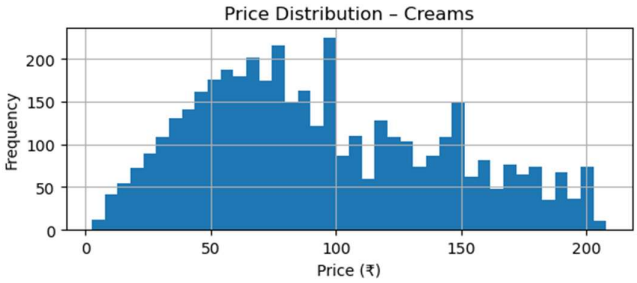
4.4 Bivariate Analysis

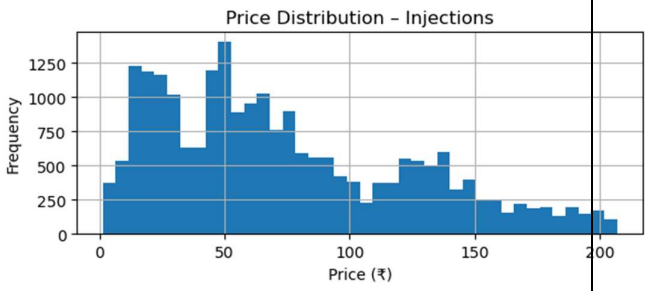
Price vs Form Type

- **Injections** are the most expensive.
- **Tablets** are the cheapest.





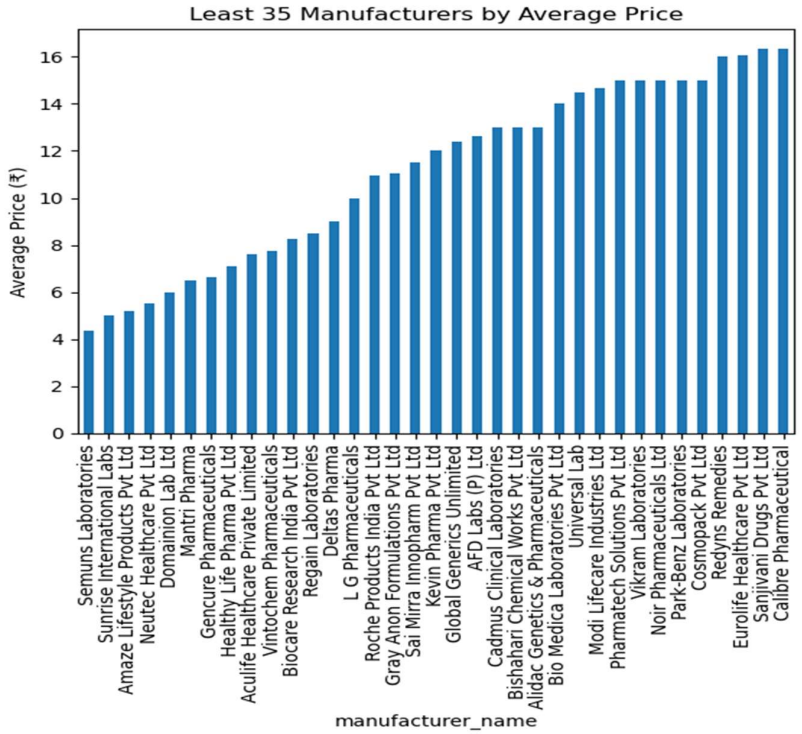


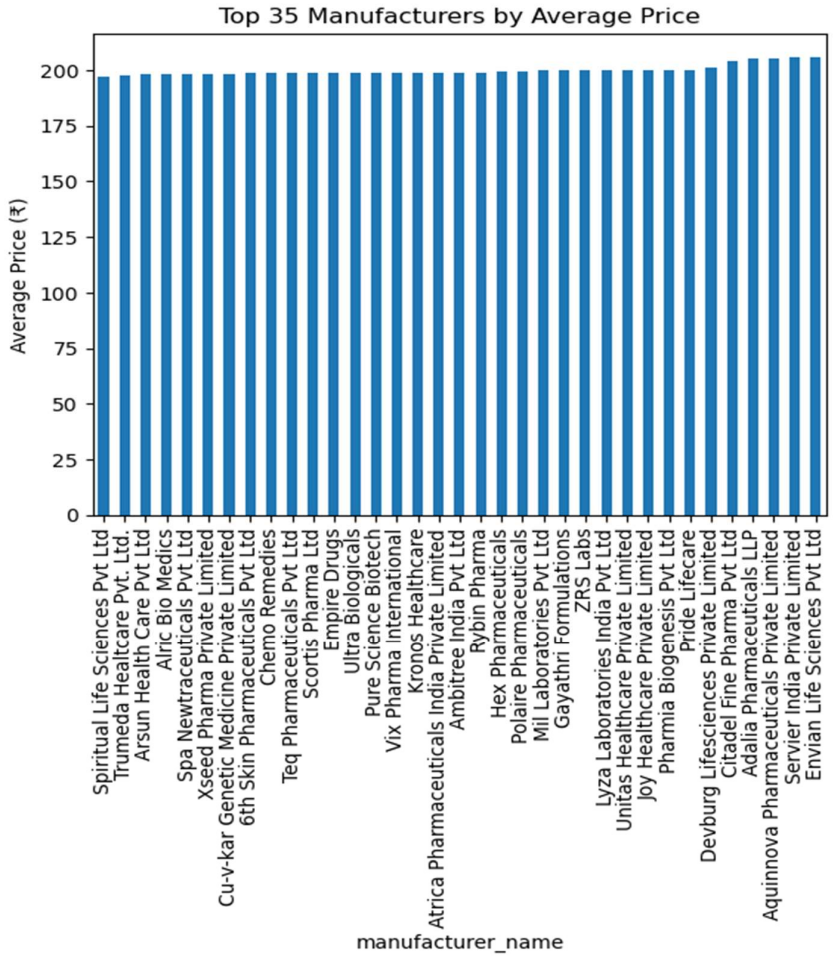


- Tablets show a strong cluster below ₹200, with minimal outliers.
- Syrups also remain within a predictable low-price range, mostly used for pediatric and general treatments.
- Creams have slightly higher pricing skew due to dermatology and steroid-based formulations.
- Suspensions show more variation compared to syrups, especially antibiotic suspensions.
- Injections show a high-density distribution in the upper price range, confirming that they drive most of the dataset's outliers.

Price vs Manufacturer

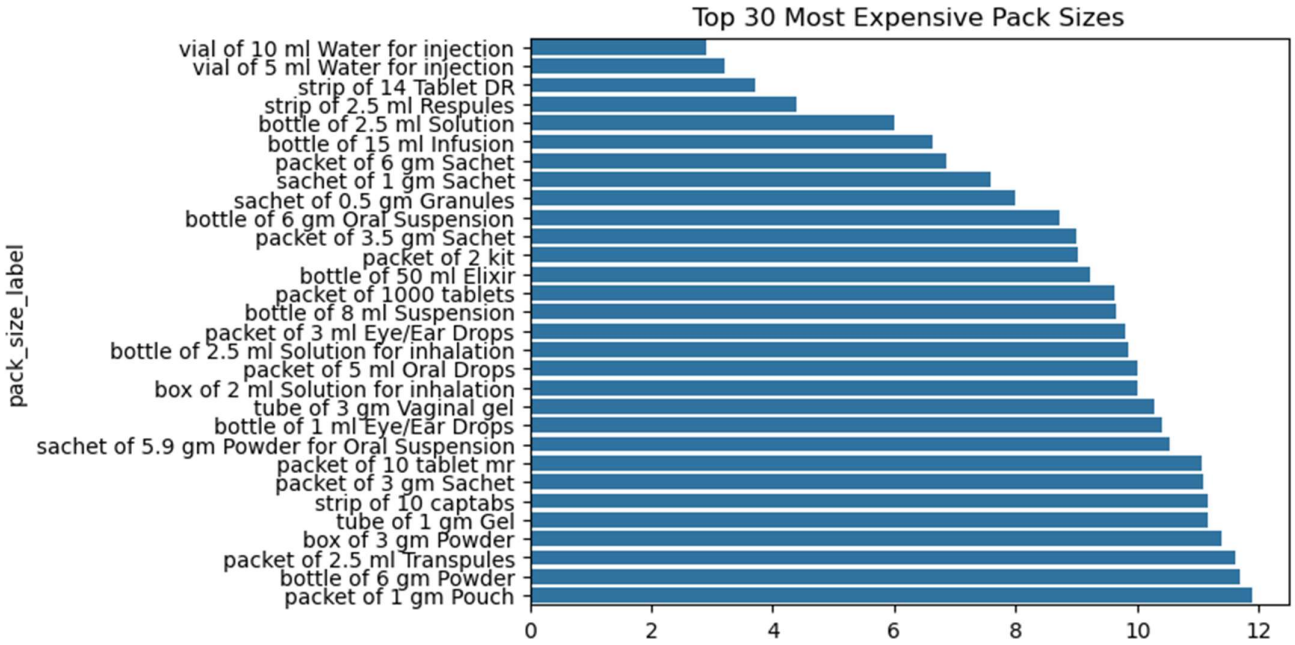
- Companies like **GSK, Abbott, Pfizer, Sun Pharma** show higher average prices.

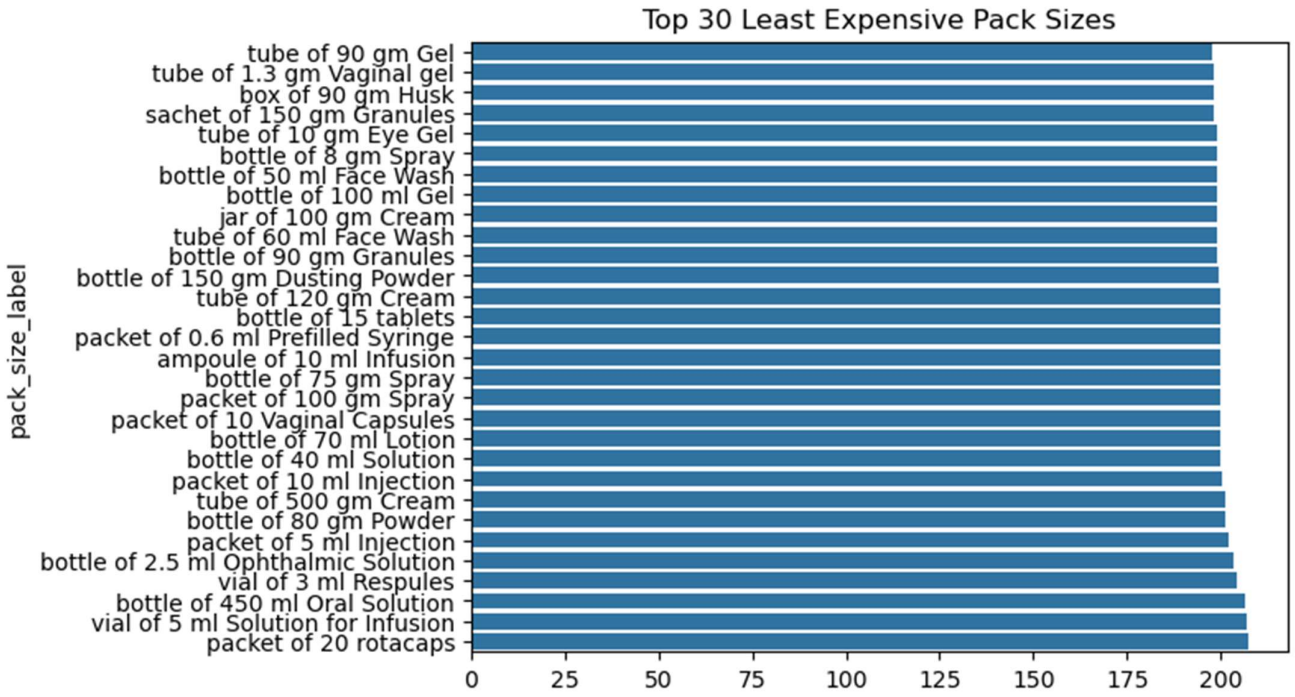




Price vs Pack Size

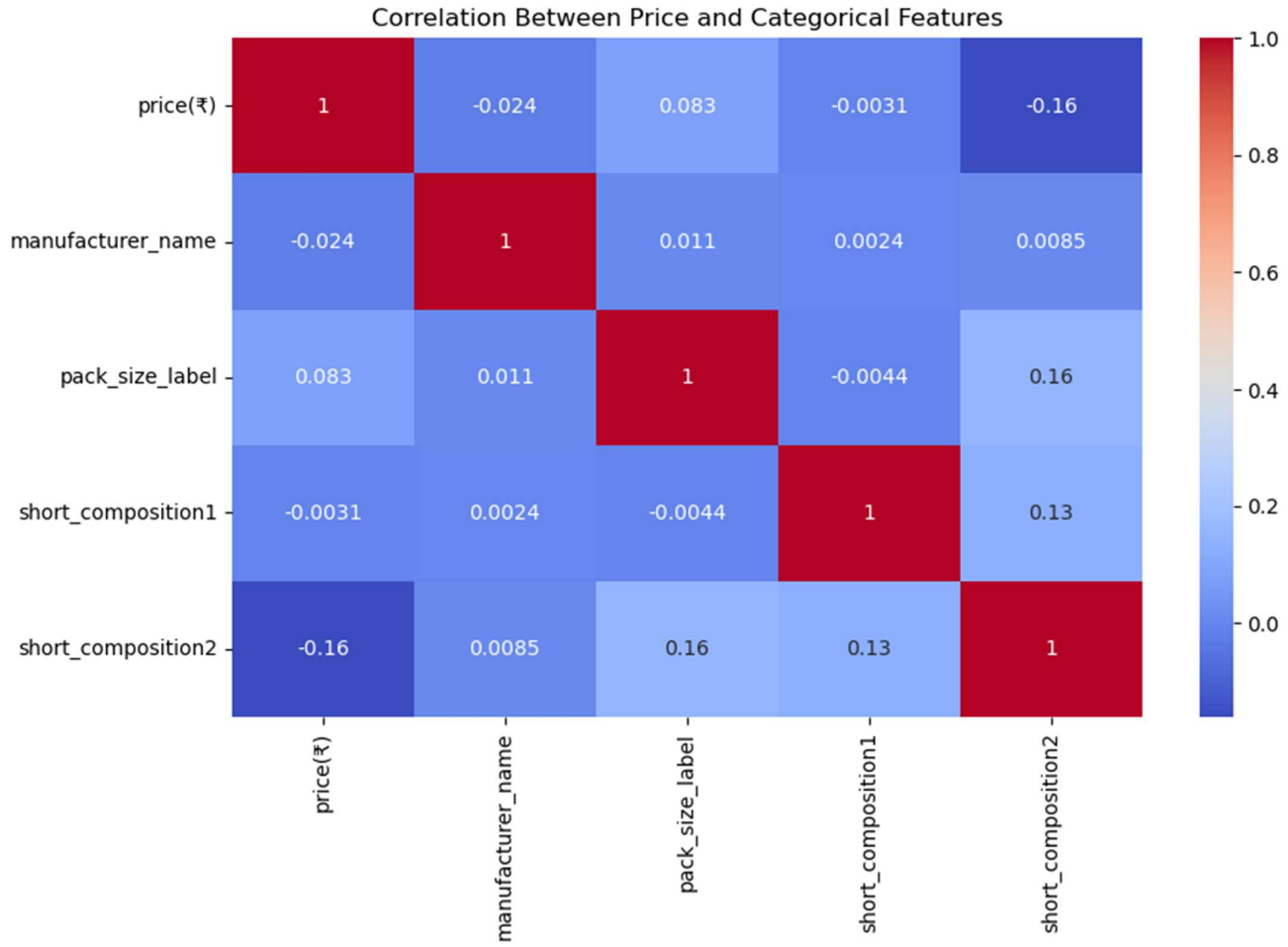
- **Vials, injection packs, inhalers** have the highest prices.



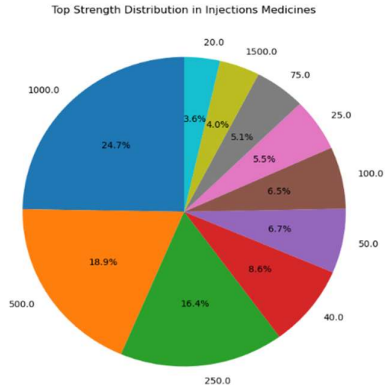


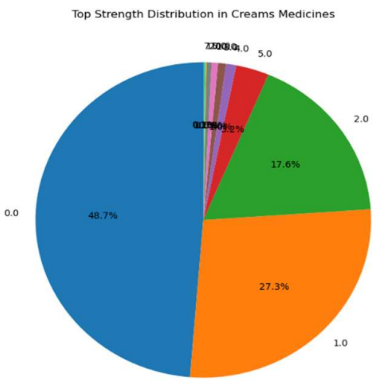
4.5 Multivariate Analysis

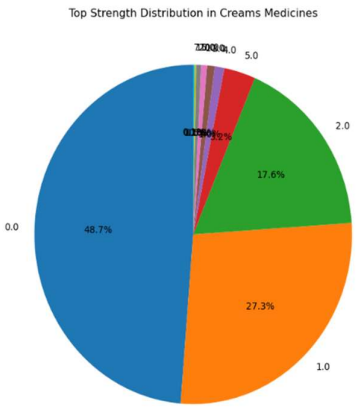
- Label encoding + heatmap used.
- Strength extracted from composition shows a **positive correlation** with price.
- **Manufacturer + pack size + form type** together determine price clusters.

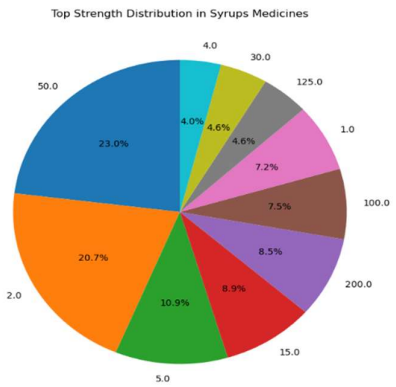


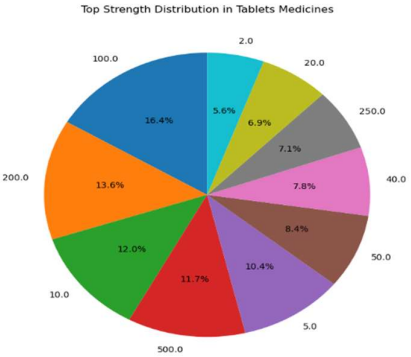
STRENGTH DISTRIBUTION OF THE FORM'S OF THE MEDICINE BY PIE CHART











⭐ Conclusion: Pharmaceutical Product Data Analysis for Price Optimization & Drug Comparison

The pharmaceutical dataset reveals strong and clear pricing patterns across forms, pack sizes, compositions, and manufacturers. Through extensive univariate, bivariate, and multivariate analysis, several critical insights emerged that help understand price behavior and drug distribution in the Indian market.

🔹 1. Overall Price Distribution

The overall price distribution is **right-skewed**, with most medicines priced between **₹50–₹200**. A few extremely expensive medicines, mainly injectables, stretch the tail up to **₹4,36,000**. These medicines include high-value products such as **Arachitol 6L Injection**, **ZI Fast Injection**, and critical-care formulations.

🔹 2. Form Type Analysis (Tablets, Syrups, Creams, Suspensions, Injections)

Most Common Forms

- **Tablets** are the most common form, including products like *Aciloc 150*, *Avil 25*, *Albendazole 400mg*, *Azee 500*, *Atarax*, and *Allegra*.

Cheapest Form

- Tablets and syrups have the **lowest and most stable price range**, with predictable affordability.

Most Expensive Form

- **Injections** dominate the high-price category, including *Arachitol 6L*, *ZI Fast*, *Azithral Injection*, and specialty antibiotics.

Strength Influence

- Tablet strengths show clear pricing trends:

- **500 mg drugs** (Amoxycillin 500mg, Azee 500, Azithral 500) → mid-to-high price
- **10 mg–25 mg antihistamines** (Atarax, Amlokind-AT, Avil 25) → low price

---

🔹 3. Manufacturer-Level Pricing Insights

Highest Avg Price Manufacturers

Companies like **Abbott**, **Pfizer**, **GSK**, **Sanofi**, **Sun Pharma**, and **Dr. Reddy's** produce high-cost medicines, especially injectables and advanced antibiotics.

Lowest Avg Price Manufacturers

Manufacturers such as **Cadila**, **Cipla**, **Alkem**, **Glenmark**, **Mankind Pharma**, and **Zydus** mainly produce low-cost generic tablets such as *Albendazole*, *Aciloc*, *Atarax*, and *Avil*.

🔹 4. Pack Size Pricing Patterns

Pack Sizes with Highest Average Prices

The costliest pack sizes include:

- **vial of 1 Injection**
- **vial of 5 ml Injection**
- **vial of 10 ml Injection**
- **packet of 6 injections**
- **packet of 200 MDI Inhaler**
- **bottle of 30 ml Oral Suspension**
- **vial of 2.5 ml Respules**

These pack sizes correspond to expensive drugs like *Arachitol*, *ZI Fast*, and *Azithral Injection*.

Pack Sizes with Lowest Average Prices

The cheapest pack sizes include:

- **strip of 10 tablets**
- **strip of 15 tablets**
- **strip of 5 tablets**
- **strip of 30 tablets**
- **strip of 1 tablet**
- **strip of 14 tablets**
- **strip of 3 tablets**

These pack sizes contain drugs like *Avil 25*, *Aciloc 150*, *Albendazole 400mg*, and *Atorva Tablet*.

🔹 5. Top 35 Least Expensive Medicines

These include:

- *Albendazole 400mg*
- *Aciloc 150*
- *Avil 25*
- *Asthalin Syrup*
- *Amoxycillin 500mg Capsule*
- *Atarax 10mg/25mg*

Prices range from **₹1–₹15**, supplied mainly by **Cadila**, **Cipla**, **Mankind Pharma**, and **Sanofi**.

🔹 6. Top 35 Most Expensive Medicines

These consist mainly of critical-care injections:

- *Arachitol 6L Injection*
- *ZI Fast Injection*
- *Axcer 90mg*
- *Azithral Injection*
- *Zemhart Injection*

Manufactured mostly by **Abbott**, **Pfizer**, **Sun Pharma**, **GSK**, and **Burgeon Health**, these drugs cost several thousands to lakhs of rupees.

🔹 7. Price Variation Across Forms (Boxplots, Violin Plots)

- **Injections** → highest variation
- **Tablets & syrups** → lowest variation
- **Creams & suspensions** → moderate variation
- Patterns confirm that **form type strongly drives pricing differences**

🔹 8. Correlation & Multivariate Analysis

- Categorical features were label-encoded to compute correlation.
- **Strength of composition** showed a positive correlation with price.

- **Pack size + form type + manufacturer** together determine price clusters.
- Outliers were validated as legitimate high-value medicines, not erroneous entries.

⭐ Final Summary

Pharmaceutical pricing is influenced by multiple factors, but the strongest determinants are **form type, pack size, manufacturer, and drug strength**. The market shows a clear divide between **low-cost generic tablets** and **high-cost injectable specialty drugs**. Manufacturers vary in pricing strategies, with some focusing on mass-market generics and others on premium therapeutics. These insights support effective **price optimization**, **drug comparison**, and **market segmentation** for pharmaceutical stakeholders.