PharmaFetch

BAX 422 - Data, Design and Representation

Winter 2023

UC Davis School of Management

Submitted By:

Arpita Mangal

Rahul Rajput

Rishika Chaudhary

Table of Contents:

Executive Summary

The healthcare industry in India is rapidly growing. To meet the needs of customers, online pharmacies and healthcare platforms like 1mg have emerged, providing a wide range of healthcare products and services at affordable costs, with the convenience of home delivery and online consultations with medical specialists. To gain insights into the products offered by 1mg, their prices, we scraped data from the website using Beautiful Soup in Python, and stored it in a MongoDB database. Compilation constraints and time constraints have limited us to 500 medicines for each alphabet. For storing the data, MongoDB was selected due to its flexibility and ability to handle unstructured data.

Our data can be used for various business scenarios like product pricing, market research, competitive analysis, and more. We have identified three top business requirements that can be addressed using this data: topic modeling, price prediction, and identifying substitutes. For inventory management, marketing strategies, and identifying research and development areas, topic modeling can be useful for clustering and labeling medicines based on the FDA's 40 categories. By determining which factors affect a medicine's cost, price prediction can help optimize pricing strategies, cost-benefit analysis, and revenue forecasting. Identifying substitutes can enhance the customer experience and increase customer loyalty.

Background

One of the largest and fastest-growing industries in the world is the healthcare sector. The healthcare sector in India is predicted to grow at a CAGR of 16–17% over the next few years, with a market value of about USD 280 billion. Some of the main factors driving this growth include the expanding prevalence of chronic diseases, rising disposable incomes, and the rising demand for healthcare services.Customers now have a popular and convenient option to get healthcare products and services through online pharmacies and healthcare platforms like 1mg. These platforms provide a wide selection of products at affordable costs, as well as the ease of home delivery and online consultations with medical specialists.

Data was taken from 1mg, a popular Indian online pharmacy and healthcare platform. The website offers a variety of healthcare products and services, such as prescription medications, over-the-counter medications, health supplements, and diagnostic tests. The platform also provides various health-related information, such as drug information, health articles, and wellness tips.

This data can be used for various business scenarios, like product pricing, market research, competitive analysis, etc. It can help provide insights about the types of products offered by 1mg, their prices and their customer reviews.

About the Data

Description of the Web-Scraping Routine

Web scraping is a method of extracting data from websites. Using Beautiful Soup in Python, the 1mg website was scraped for information about medicines, such as id as on website, manufacturer, non-proprietary name, dosage form, active ingredient, mechanism of action, name, price, description, product information, uses, benefits, workings, factbox, and substitutes. Beautiful Soup is a popular web scraping library that makes it easy to parse HTML and XML documents.

To obtain this information, we utilized Beautiful Soup in Python, to scrape URL's of medicines listed in alphabetical order on the website. After getting all the URL's we appended them in a list and then from the URL's we download the HTML file for each medicine and then extract relevant information from the HTML file.

Due to computation and time limitations, we have extracted top 500 medicines from each alphabet, so in total we have information of 500 * 26 = 13,000 medicines. This method of web scraping has advantages because it allows for a large amount of data to be efficiently extracted and stored without manual input. After getting the necessary information we store it in a MongoDB database in JSON format.

Database Design Choice

We chose MongoDB as the database for storing this data because of the unstructured nature of the data, and for the need of flexibility in the schema. In a traditional relational database like MYSQL, the schemas are rigid, and any deviation from the already defined schema will not be stored in the database. MongoDB is a NoSQL database that stores data in JSON format with dynamic schemas. As the schemas are dynamic, we can modify it as per our requirement. Another advantage of storing data MongoDB is when we are storing null values. In traditional relational databases, the presence of null values can result in wasted space, as the database reserves space for the null values even though they do not contain any data. This can lead to inefficient storage and retrieval of data. While MongoDB does not reserve space for null values and can lead to more efficient storage and retrieval of the data. Our data had null values, so this was another reason for using MongoDB as our database. The flexibility is ideal for storing data scraped from websites, where the attributes and structure of the data can vary widely.

We stored each medicine as a document in mongoDB database. We stored the manufacturer information as a dictionary as it includes the type as well as name of manufacturer, the could be utilised by business to identify which type of manufacturer have a larger market share. The information on non-proprietary name, dosage form, active ingredient, mechanism of action, name, price, description, product information, uses, workings and benefits were stored as key value pairs in the document with the key name as information it corresponds to. The data for side effects of a medicine is stored as a list as a medicine could have as many as 10 or more side effects to 1 or none. The facts regarding a medicine are stored as a dictionary with name factbox as there are four categories of facts for a medicine namely, Habit Forming, Therapeutic Class, Chemical Class, and Action Class. We used a dictionary as any of the facts is easily accessible using the fact name of interest. Lastly we also stored the substitute of medicine as a dictionary which contains information on substitute name and id as on the website. The substitute id could be used to get other information about the substitute using the id column in the database.

Discussion - Business Requirements

The dataset scraped from 1mg website has a lot of potential to answer business- relevant questions and provide insights that can help drive the business and create value. The top 3 business requirements that we would be looking into are:

1. Topic Modeling : Clustering and labeling medicines according to the 40 categories listed by the FDA can help provide valuable insights into the similarities and differences between different types of medicines. This information may be useful to optimize inventory management, implementing marketing strategies, and also identifying potential areas of research and development.
2. Price Prediction : Building a model to predict the price of a medicine based on its attributes could be useful for price optimization, cost benefit analysis and also revenue forecasting. By utilizing ML algorithms, it is possible to determine which factors affect the cost of the medicine and then help accurately predicting the price.
3. Identifying Substitutes : If we are able to correctly identify substitutes for medicines that may not be available in the market by utilizing their descriptions and attributes, it would lead to enhanced customer experience and also might increase customer loyalty. When there is demand for a medicine from a specific brand which might not be available or it is a highly priced medicine, then a substitute from a lesser known brand could be utilized.

Other business relevant questions may include, determining which manufacturers produce the most popular medicines in each category. This information can help identify market leaders and potential competitors, inform pricing strategies, and optimize supplier relationships. We can also look into how prices differ among the medicines and their substitutes to identify pricing opportunities, inform substitution strategies, and optimize revenue.

Conclusion

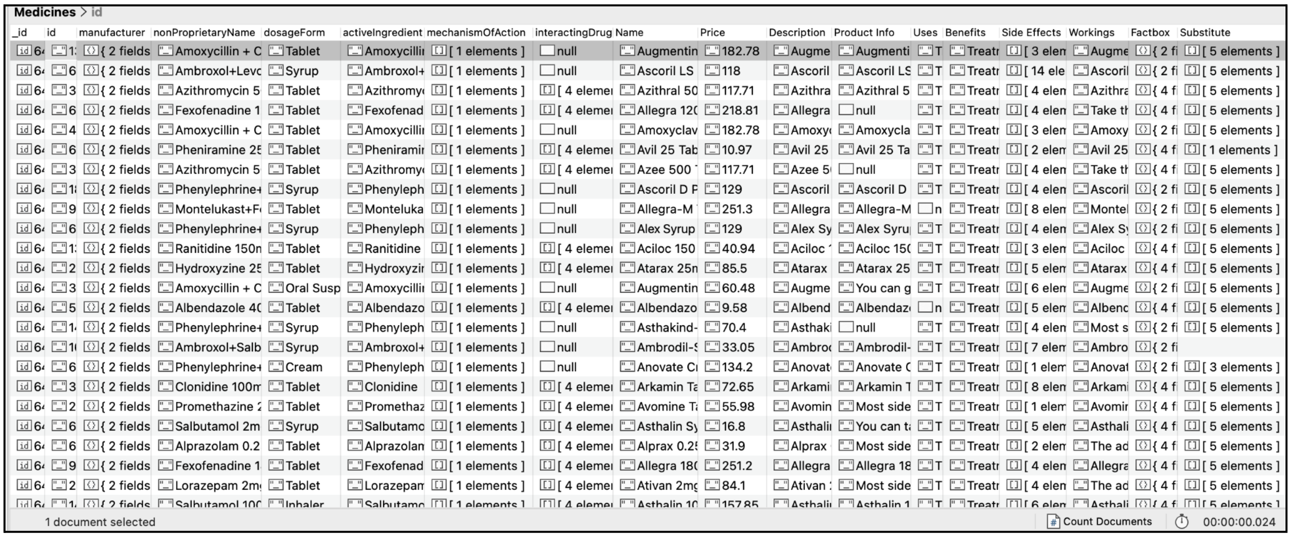
In conclusion, the dataset scraped from the 1mg website has great potential for providing valuable insights to drive business decisions and create value. Businesses can use a wealth of data about medications, including information on their attributes, costs, and feedback from customers, to improve supply chain management, pricing strategies, and inventory control.

Topic modeling, pricing prediction, and identifying alternatives are just a few examples of how this data might be used to support decision-making for the three business criteria mentioned in this paper. The dataset can also be further investigated to provide answers to other business-related queries, such as identifying market leaders and possible competitors, discovering price opportunities, and maximizing revenue. Using MongoDB as the database for storing this data allowed for flexibility in the schema, making it easier to modify and store data without rigid structures.

Overall, the insights offered by this dataset can assist companies in the healthcare sector in maintaining their competitiveness, enhancing the customer experience, and stimulating industry growth.

APPENDIX

Screenshot to show how the final data is stored in MongoDB –



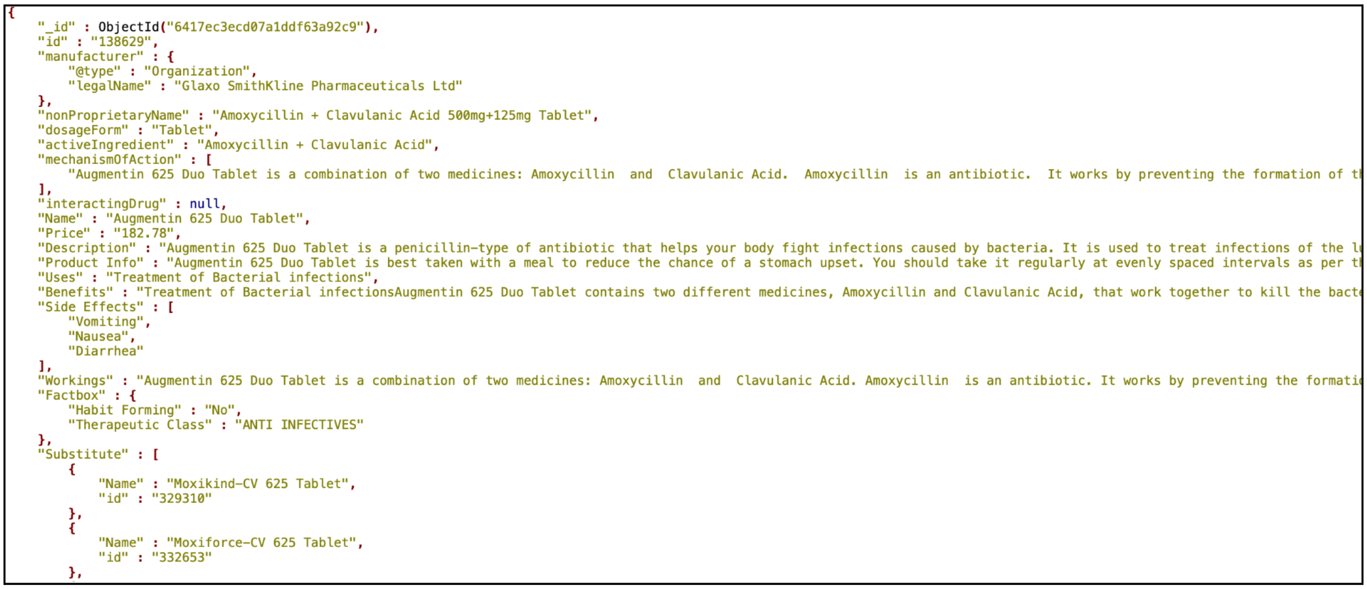
Medicines > id

_id	id	manufacturer	nonProprietaryName	dosageForm	activeIngredient	mechanismOfAction	interactingDrug	Name	Price	Description	Product Info	Uses	Benefits	Side Effects	Workings	Factbox	Substitute
64	13	{ 2 fields	Amoxycillin + C	Tablet	Amoxycilli	[1 elements]	null	Augmentin	182.78	Augme	Augmenti	T	Treatr	[3 elem	Augme	{ 2 fi	[5 elements]
64	6	{ 2 fields	Ambroxol+Levc	Syrup	Ambroxol+	[1 elements]	null	Ascoril LS	118	Ascoril	Ascoril LS	T	Treatr	[14 ele	Ascoril	{ 2 fi	[5 elements]
64	3	{ 2 fields	Azithromycin 5	Tablet	Azithromy	[1 elements]	[4 elemer	Azithral 50	117.71	Azithra	Azithral 5	T	Treatr	[4 elem	Azithra	{ 4 fi	[5 elements]
64	6	{ 2 fields	Fexofenadine 1	Tablet	Fexofenad	[1 elements]	[4 elemer	Allegra 12(	218.81	Allegra	null	T	Treatr	[4 elem	Take th	{ 4 fi	[5 elements]
64	4	{ 2 fields	Amoxycillin + C	Tablet	Amoxycilli	[1 elements]	null	Amoxyclav	182.78	Amoxy	Amoxycla	T	Treatr	[3 elem	Amoxy	{ 2 fi	[5 elements]
64	6	{ 2 fields	Pheniramine 25	Tablet	Pheniramir	[1 elements]	[4 elemer	Avil 25 Tab	10.97	Avil 25	Avil 25 Ta	T	Treatr	[2 elem	Avil 25	{ 4 fi	[1 elements]
64	3	{ 2 fields	Azithromycin 5	Tablet	Azithromy	[1 elements]	[4 elemer	Azee 500	117.71	Azee 5	null	T	Treatr	[4 elem	Take th	{ 4 fi	[5 elements]
64	18	{ 2 fields	Phenylephrine+	Syrup	Phenyleph	[1 elements]	null	Ascoril D P	129	Ascoril	Ascoril D	T	Treatr	[4 elem	Ascoril	{ 2 fi	[5 elements]
64	9	{ 2 fields	Montelukast+F	Tablet	Monteluka	[1 elements]	null	Allegra-M	251.3	Allegra	Allegra-M	n	Treatr	[8 elem	Montel	{ 2 fi	[5 elements]
64	6	{ 2 fields	Phenylephrine+	Syrup	Phenyleph	[1 elements]	null	Alex Syrup	129	Alex Sy	Alex Syru	T	Treatr	[4 elem	Alex Sy	{ 2 fi	[5 elements]
64	13	{ 2 fields	Ranitidine 150n	Tablet	Ranitidine	[1 elements]	[4 elemer	Aciloc 150	40.94	Aciloc	Aciloc 15(	T	Treatr	[3 elem	Aciloc	{ 4 fi	[5 elements]
64	2	{ 2 fields	Hydroxyzine 25	Tablet	Hydroxyzir	[1 elements]	[4 elemer	Atarax 25n	85.5	Atarax	Atarax 25	T	Treatr	[5 elem	Atarax	{ 4 fi	[5 elements]
64	3	{ 2 fields	Amoxycillin + C	Oral Susp	Amoxycilli	[1 elements]	null	Augmentin	60.48	Augme	You can g	T	Treatr	[6 elem	Augme	{ 2 fi	[5 elements]
64	5	{ 2 fields	Albendazole 40	Tablet	Albendazo	[1 elements]	[4 elemer	Albendazo	9.58	Albend	Albendaz	n	Treatr	[5 elem	Albenc	{ 4 fi	[5 elements]
64	14	{ 2 fields	Phenylephrine+	Syrup	Phenyleph	[1 elements]	null	Asthakind-	70.4	Asthaki	null	T	Treatr	[4 elem	Most s	{ 2 fi	[5 elements]
64	10	{ 2 fields	Ambroxol+Salb	Syrup	Ambroxol+	[1 elements]	null	Ambrodil-S	33.05	Ambroc	Ambrodil-	T	Treatr	[7 elem	Ambro	{ 2 fi	
64	6	{ 2 fields	Phenylephrine+	Cream	Phenyleph	[1 elements]	null	Anovate Cr	134.2	Anovat	Anovate C	T	Treatr	[1 elem	Anovat	{ 2 fi	[3 elements]
64	3	{ 2 fields	Clonidine 100m	Tablet	Clonidine	[1 elements]	[4 elemer	Arkamin Ta	72.65	Arkami	Arkamin T	T	Treatr	[8 elem	Arkami	{ 4 fi	[5 elements]
64	2	{ 2 fields	Promethazine 2	Tablet	Promethaz	[1 elements]	[4 elemer	Avomine Ta	55.98	Avomin	Most side	T	Treatr	[1 elem	Avomir	{ 4 fi	[5 elements]
64	6	{ 2 fields	Salbutamol 2m	Syrup	Salbutamo	[1 elements]	[4 elemer	Asthalin Sy	16.8	Asthali	You can ta	T	Treatr	[5 elem	Asthali	{ 4 fi	[5 elements]
64	6	{ 2 fields	Alprazolam 0.2	Tablet	Alprazolan	[1 elements]	[4 elemer	Alprax 0.25	31.9	Alprax	Most side	T	Treatr	[2 elem	The ad	{ 4 fi	[5 elements]
64	9	{ 2 fields	Fexofenadine 1	Tablet	Fexofenad	[1 elements]	[4 elemer	Allegra 18(	251.2	Allegra	Allegra 18	T	Treatr	[4 elem	Allegra	{ 4 fi	[5 elements]
64	2	{ 2 fields	Lorazepam 2m	Tablet	Lorazepam	[1 elements]	[4 elemer	Ativan 2mg	84.1	Ativan	Most side	T	Treatr	[4 elem	The ad	{ 4 fi	[5 elements]
64	14	{ 2 fields	Salbutamol 100	Inhaler	Salbutamo	[1 elements]	[4 elemer	Asthalin 10	157.85	Asthali	Asthalin 1	T	Treatr	[6 elem	Asthali	{ 4 fi	[5 elements]

1 document selected

Count Documents 00:00:00.024

Fig 1 - Studio3T Screenshot



```
{
    "_id" : ObjectId("6417ec3ecd07a1ddf63a92c9"),
    "id" : "138629",
    "manufacturer" : {
        "@type" : "Organization",
        "legalName" : "Glaxo SmithKline Pharmaceuticals Ltd"
    },
    "nonProprietaryName" : "Amoxycillin + Clavulanic Acid 500mg+125mg Tablet",
    "dosageForm" : "Tablet",
    "activeIngredient" : "Amoxycillin + Clavulanic Acid",
    "mechanismOfAction" : [
        "Augmentin 625 Duo Tablet is a combination of two medicines: Amoxycillin  and  Clavulanic Acid.  Amoxycillin  is an antibiotic.  It works by preventing the formation of th
    ],
    "interactingDrug" : null,
    "Name" : "Augmentin 625 Duo Tablet",
    "Price" : "182.78",
    "Description" : "Augmentin 625 Duo Tablet is a penicillin-type of antibiotic that helps your body fight infections caused by bacteria. It is used to treat infections of the lu
    "Product Info" : "Augmentin 625 Duo Tablet is best taken with a meal to reduce the chance of a stomach upset. You should take it regularly at evenly spaced intervals as per th
    "Uses" : "Treatment of Bacterial infections",
    "Benefits" : "Treatment of Bacterial infectionsAugmentin 625 Duo Tablet contains two different medicines, Amoxycillin and Clavulanic Acid, that work together to kill the bacte
    "Side Effects" : [
        "Vomiting",
        "Nausea",
        "Diarrhea"
    ],
    "Workings" : "Augmentin 625 Duo Tablet is a combination of two medicines: Amoxycillin  and  Clavulanic Acid. Amoxycillin  is an antibiotic. It works by preventing the formatio
    "Factbox" : {
        "Habit Forming" : "No",
        "Therapeutic Class" : "ANTI INFECTIVES"
    },
    "Substitute" : [
        {
            "Name" : "Moxikind-CV 625 Tablet",
            "id" : "329310"
        },
        {
            "Name" : "Moxiforce-CV 625 Tablet",
            "id" : "332653"
        },
```

Fig 2 - Studio3T Screenshot