DataCare

BI platform based on ML & Analytics for pharmaceutical suppliers

Talia Rosenkranz
Diego Gonzalez
Gonzalo Vargas
Stefania Hau
Tania Cajal

Master in Big Data & AI solutions | Barcelona Technology School

Introduction

The DataCare project started with a question: how is data being used in the pharmaceutical industry? The answer depends on the country, the parties involved, and the technical capabilities of their teams. However, we noticed a pattern. Diego Gonzalez, our CEO from Chile, pointed out the huge price gap of medications sold between large pharmacies and small ones. We decided to look into it.

First, we acquired our key partner: Cenabast, a Chilean company dedicated to regulating medication prices. In talking to them, we found that a fundamental problem lies in the supply chain. Laboratories and other drug suppliers are relying on decades-old knowledge of the market and the traditional economic assumption; the big company will buy more for less, and the little one will buy less for more.

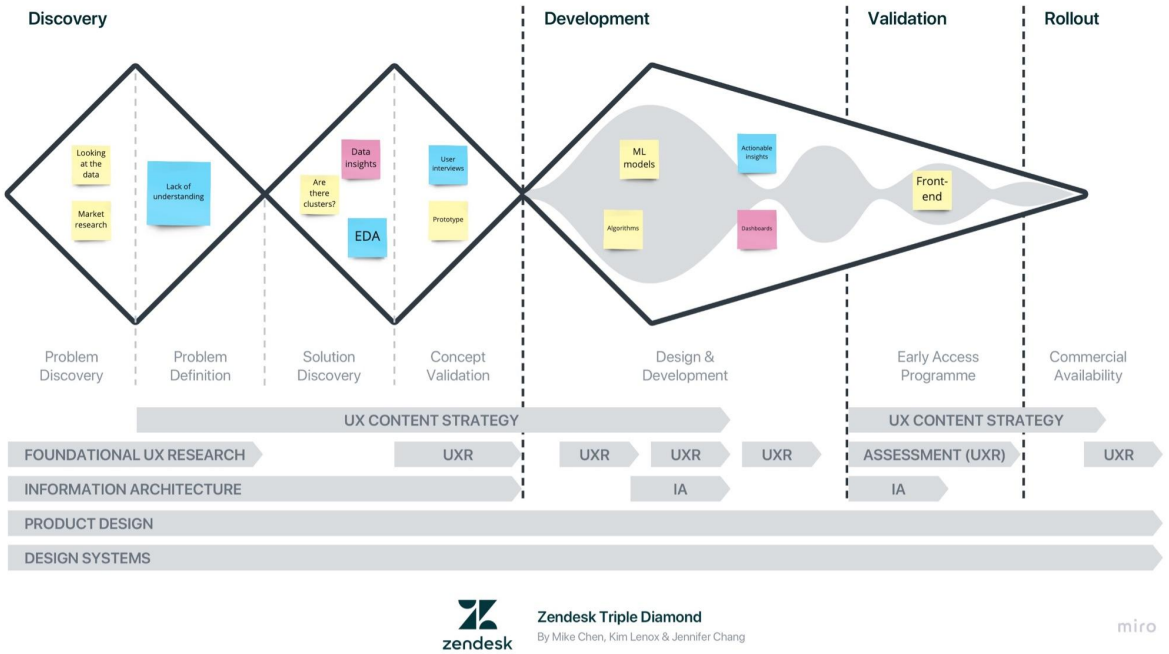
We wondered if there was something else to uncover. What if there were patterns beyond big versus small, specialized versus generic products? In the long run, efficient markets tend to result in welfare for the final consumer. Fortunately, big data and machine learning have a history of improving efficiency.

Continue reading to find how we are leveraging the tools available, targeting the key links in the supply chain, and improving the way they understand and interact with their customers (hence, pharmacies), resulting in a data-driven approach to drug distribution, consequently giving smaller-scale players a fighting chance.

Why us?
We are an international and multidisciplinary team:

- Diego Gonzalez, our Chief Executive Officer, is a data analyst from Chile with abundant experience in the pharmaceutical sector.
    - https://www.linkedin.com/in/diegonzalezava/
- Talia Rosenkranz, our Chief Information Officer, is a data scientist from Germany, with a sharp analytical eye and zero tolerance for mediocrity.
    - https://www.linkedin.com/in/talia-rosenkranz-b67766161/
- Tania Cajal, our Chief Artificial Intelligence Officer, is an artificial intelligence consultant ready to boost innovation.
    - https://www.linkedin.com/in/tgcajal/
- Stefania Hau, our Chief Financial Officer, is a business developer from Romania, with a magic touch with numbers and a passion for innovation.
    - https://www.linkedin.com/in/stefaniahau/
- Gonzalo Vargas, our Chief Administrative Officer, is a data analyst from Bolivia, has a talent to swoon clients, and oil the wheels of our team.
    - https://www.linkedin.com/in/gonzalo-vargas-toledo-809b3672/

Value Proposition



Discovery

Problem Delivery

We started off this project by analyzing the price discrepancy of the same medication sold within Chile, in different pharmacies. The data showed that there are two types of pharmacies. Those that are privately owned, and those that are corporalty operated. The issue lies within the fact that the same medication is sold at a significantly higher price in the corporate pharmacies compared to the smaller privately owned pharmacies (Appendix 1). Our research showed that the difference in price within the exact same product could be up to 637% (Cenabast, 2022). Based on these findings we created the concept of building an application where the user could find their product of interest at the lowest cost as well as the smallest proximity to their desired location (Appendix 2).

We began to validate the idea with potential users using an MVP but also contacted pharmacies and laboratories if they would have been interested to partner up with us. For the pharmacies, we would have been a good lead for new customers while for the laboratories, we would have sold reports with the collection of data on drugs sold countrywide. As we began closer conversations with the laboratories, we recognized the existence of the fundamental problem that they had no understanding of their customers' (the pharmacies) behavior in the first place. The

collection of data does exist within small, medium and large laboratories, however not even the third largest distributor of drugs in Chile has a tool or service to properly make use of their structured data. Upon this realization, we came to the conclusion to pivot our company goal. We became a soly B2B driven company in which we leverage the data of our clients, the laboratories, to support them in identifying the differences between their customers, depending on their purchase behavior as well as the type of products they are mainly buying. The understanding about the different kinds of products and pharmacies there are in Chile is now also helping us understand how to segment the customers (pharmacies) for our clients (laboratories).

Problem Definition

Pharmaceutical Laboratories, which are providers of drugs to pharmacies, have no tool or service to personalize the communication to their customers, understand the demand nor understand their customers' purchase behavior.

Solution Discovery

Due to the described problem understanding, we began to put together a set of algorithms that make best use of the data available, in order to create value from the raw data for laboratories. Our goal to create a business intelligence platform based on machine learning and analytics for pharmaceutical suppliers was defined. Throughout the development of the best suited solution, we continuously communicated and presented the product to our partners in order to gain feedback of which analysis tools were beneficial to them and to identify which ones were considered less important. Our vision, to help vendors understand their customers as you and I understand our best friend, took on shape.

Development

From a simple dashboard to a web app with multiple solutions, the design and development process of DataCare was full of ups, downs, and U-turns. We opted for an Agile approach using Jira to develop the main functions and build up from them. By talking to potential users, we settled on a front-end design and worked up the back-end architecture. Through continuous testing and experimentation, we turned possible users' feedback into results, with the right balance between visualizations and numbers.

Exclusive, demo-purposes GitHub access by clicking here.

Partners

We have three main partners at this time which have guided us through the process of identifying the problem, validating the idea and validating the prototype.

Cenabast is our most important partner as they have been providing us with the data we needed to train our models. They are the third largest pharmaceutical distributor in Chile. Initially they said that what they need is a product to understand their collected data of the country wide demand. Indopharma is a medium sized laboratory from which we understood that they are of desperate need for our product. They said that products sold most commonly depend on the population of people that live in specific areas. While they do have the data, they don't know how to leverage it. Similar with Geamed. they a very small pharmacy based in chile as well and are in need of basic tools such as what products are sold most and by whom. They have all also been validating our prototype throughout the process.

Problem Validation

Our user testing through interviews and many sparing conversions showed that real-life laboratories and drug suppliers have no understanding of the market on different levels. Here are some of their testimonies. The protocolled interview can be reviewed in the appendix 4.

Rollout

Based on what we have been told, interviews and our preliminary user testing, we have landed on an optimal user persona: a mid-sized drug supplier located in Chile or Spain, to begin with. While small and medium companies have the most potential to acquire our services. Bigger companies might be able to develop inhouse solutions with domestic talent, whereas small companies may not be liquid enough to cover our fees. We are targeting the medium laboratory and turning it into a top one.

Market Validation

To validate that a market exists for increased personalized communication through a better understanding of one's customer, we performed some research. According to Afshar, 2017 (Chief Digital Evangelist at Salesforce), 91% of customers don't complain when they are having issues with their vendor. They simply leave. This underlines the importance of the vendor recognizing their customers' unhappiness. DataCare is dedicated to help detect those red flags to decrease the churn rate.

Additionally, Afshar suggests that in 65% of the cases, clients will switch vendors due to the lack of personalized communication. 85% of vendors that use their customers' behavioral data outperform their competition.

According to the report by IT Digital Media Group, 2022, laboratories increase their revenue by 73 % by investing into digital products.

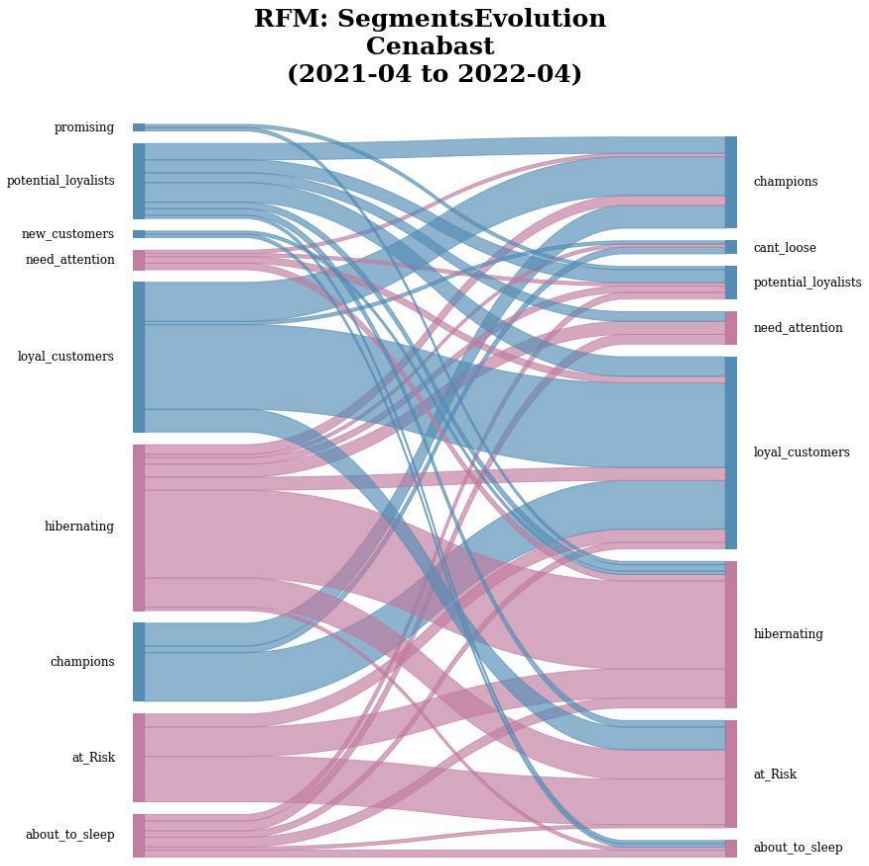
These statistics further underline the benefit laboratories will have from employing DataCares services.

Idea Validation

We began to validate the need of creating a platform that lets our clients understand their customers better. We began by calculating the churn rate as this would indicate that clients are lost and therefore sales are reduced. A decreased churn rate is a good KPI for us to understand whether our platform shows an effect

or not. With use of the provided data by Cenabast, we found that Cenabast has a historical churn of around 7%. This result is supported by data from the past two years. We calculated it by first gathering the average number of days between orders for each customer of cenabast. We then took the overall mean, calculated the standard deviation and multiplied it by a set confidence interval. All customers whose last purchase has been longer than the maximum number of days calculated are considered churn.

Furthermore, in order to understand the current and past status of the different clients that Cenabast has, we created a Recency - Frequency - Monetary analysis (RFM). This allowed us to find that close to 50% of Cenabasts' clients are placed within the negative segments such as "at risk" or "hibernating" (displayed in red below) which means that there is a high chance to lose these customers. We then compared each segment with the previous year in order to see how many customers had changed segments over the year.

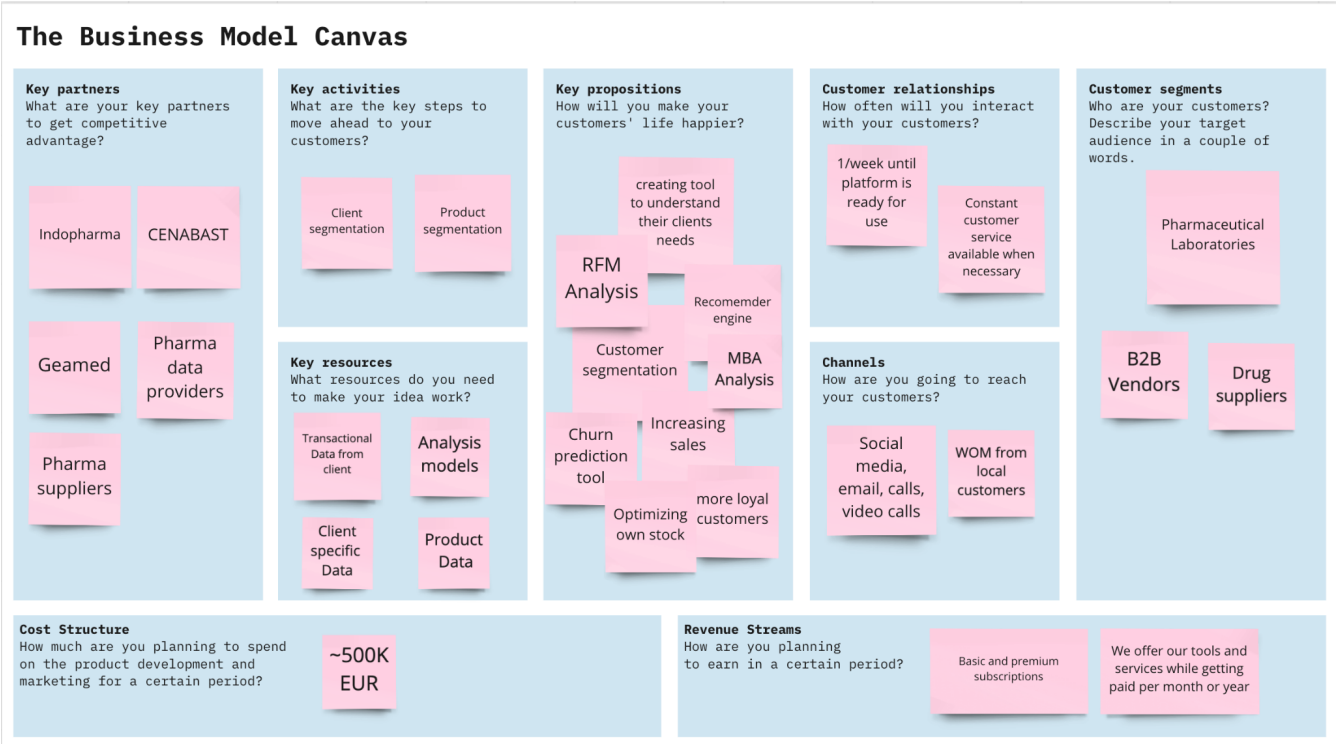


As we can see in the plot above, the segment "hibernating" is one of the largest represented segments with 23.7%. If we follow the lines of the hibernating segment, we can see that a large percentage of customers remained within this segment. This clearly calls for action on Cenabasts' behalf to wake up these clients through personalized communication and well suited product recommendations.

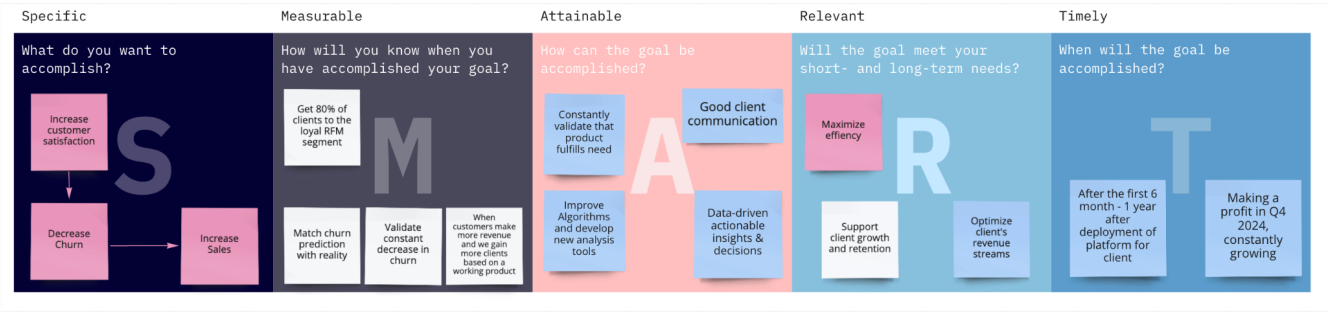
We can also recognize from this analysis that the “at risk” segment grew 3% over the year which further underlines the importance of reaching out to these clients in time in order to not lose these customers.

Upon this analysis we also understood the usefulness of the RFM analysis and therefore deployed it within our platform as well. This will be described in more detail below.

Business Model



SMART goals



Plan Elements

Background

The purpose of this work is to find out if there is a demand to develop a BI platform based on ML & Analytics for pharmaceutical suppliers in Spain.

The current situation in Spain shows there's an opportunity in the market due to the growth of the IT industry in recent years, in cities such as Barcelona, Madrid, Valencia and others.

The pharmaceutical industry is a key and strategic sector of the Spanish economy. In the last 25 years, the sector has gained enormous relevance, becoming an important driver of Spanish exports and of private investment in R&D. Despite this, its productive capacity still has room for improvement.

Characterization of the pharmaceutical industry

The pharmaceutical industry is not a sector of enormous dimensions in the Spanish economy, although it is not residual either. According to CaixaBank Research, 2022, the pharmaceutical industry directly generated 6,846 million euros of gross value added (GVA), 0.6% of the total Spanish economy. It is the eighth largest industrial sector in the country, which generates 5% of the GVA of the manufacturing industry.

Market analysis of DataCare

Competitor analysis

The competitors are companies offering similar services. With them we mainly compete over prices, services and customer satisfaction. We aim to be better and more complete in terms of analytical results. The main focus is offering services they do not get from elsewhere. We are not alone in the field, therefore, in order to research the market and find the customer base, our main target audience are companies working in the pharmaceutical sector. The growth of the platform has been based on the growth of similar platforms, such as Snowplow, Scalefast, Woopra.

Financial Projections Estimations

Below there is a summary of how financials estimations were done. Here is a more detailed explanation of how calculations were made.

Since our company is still in early stages, without having launched a product yet, the projections were based on some general assumptions, based on market research.

Since this application has not yet been launched and in order to find the right market for this application, we have made assumptions by looking at the potential market in Europe and South America, with a similar growth to that of competitive applications with similar services (i.e. Scalefast, Woopra, Snowplow,Kraz.ai) for the users.

In order to estimate the financial projections, we researched the pharmaceutical laboratories within the Spanish market. According to *Ranking Empresas Fabricación de especialidades farmacéuticas,* 2021, there are currently 188 laboratories divided into four categories: corporate, big, medium and small. Our intention for DataCare is to penetrate these markets and obtain the market share per each category as displayed in the table.

Etiquetas de fila	Cuenta de Posición Sector	Cuenta de Tipo	Marketshare
Big	73	38.83%	50%
Coorp	62	32.98%	15%
Medium	27	14.36%	60%
Small	26	13.83%	70%
Total general	**188**	**100.00%**	**Sum**

Based on the market data above, we have performed our analysis over the next 2 years, dividing the number of laboratories per category by 24 months. This represents our estimation per month per user per category of laboratories.

			Año 2						
Termino	Predictions	Variable	18	19	20	21	22	23	24
Total Laboratories		188	141	149	157	164	172	180	188
Total coorp laboratories	15%	62	46	49	52	54	57	59	62
Total big laboratories	50%	73	55	58	61	64	67	70	73
Total medium laboratories	60%	27	20	21	23	24	25	26	27
Total small laboratories	70%	26	20	21	22	23	24	25	26

Business Model

The subscription models which DataCare offers are “Basic” and “Pro”. The Basic is addressed to SME organizations and offers features such as uploading CSV files onto the platform, conducting analysis on data such as RFM, MBA, but with a limited computing time and power. In addition to the Basic model, the Pro version offers the user a custom setup, customer segmentation, market insights, churn analysis, company-based personalized services, API connections, cloud services and maintenance periodically. When considering the monthly price for our subscription models, we have looked at similar tools available on the market, the competitors prices and offerings.

Revenue

In order to estimate the revenue for each model per category of laboratories, we have estimated the percentages for each category for each model. For the Basic subscription, we will focus on targeting small laboratories (80%), coorp laboratories (40%), medium (30%) and big (30%). Similarly, for the Pro subscription, we will focus on targeting small laboratories (20%), coorp laboratories (60%), medium (70%) and big (70%). In order to calculate the estimated revenue, we have analyzed the number of users per month by the set price of the subscriptions for both models.

Costs

As most of the costs for DataCare are variable costs, and these costs vary a lot depending on how many users you have (i.e. AWS costs will change depending on how many users we have), we have had to make some assumptions to see what the costs would be like.

We then have added into consideration the costs such as: cloud services, marketing, salaries, HR and hardware for the next 24 months. For the purpose of this project, we have estimated that the costs for HR, hardware, marketing and the salaries will not change for the next two years.

Cloud Service	0.1
Total Operational cost	
Total Operational revenue	
HR	10,000
Hardware	10,000
Marketing	10,000
Salaries	60,000

Following, there is a summary of income, operational and fixed costs, as well as revenue before tax. Summary of financials, considering 25% corporate tax for the Spanish market.

Column1	Year 1	Year 2
Collaborators	93.996	187.992
Total Revenue Model 1	74,974.38	213,388.62
Total Revenue Model 2	216635.64	616578.36
Total Revenue	291,610.02	829,966.98
YoY Revenue Growth	0	185%
Variable Costs	21663.564	61657.836
Fixed Costs	480000	480000
Total Costs	501663.564	541657.836
EBTDA	-210053.544	288309.144
Corporate Tax (25%)	-52513.386	72077.286
Net Profit After Tax	(157,540.16)	216,231.86

Valuation DataCare

To perform a valuation for DataCare, we have used the 'Scorecard valuation method'. This method compares the target company to typical angel-funded startup ventures and adjusts the median valuation of recently funded companies in the region to establish a pre-money valuation of the target. The pre-money valuation for DataCare is €1.5M.

Comparison Factor	Range	Target Company	Factor
The strength of the Management Team	30%	100%	0.30
Size of the Opportunity	25%	125%	0.31
Product/Technology	15%	100%	0.15
Competitive Environment	10%	75%	0.08
Marketing/Sales Channels/Partnerships	10%	80%	0.08
Need for Additional Investment	5%	100%	0.05
Other	5%	100%	0.05
Sum			1.02
Average pre-money valuation	€1.5M		
Summed factor	1.02		
DataCare Pre-money valuation	€1.5M		

Big Data Prototype

The Data

While we initially collected data by parsing online catalogs of pharmacies, we noticed a handful of statistically significant problems. First, only medium and big pharmacies, especially large chains, had a near-full, updated catalog listed online. This reflected the price to the end user -clearly without the marginal revenue-, and was of no use to us as a B2B solution provider.

Luckily, upon pointing out the aforementioned gaps in knowledge in the industry, the Chilean pharmaceutical supplier Cenabast agreed to a mutually beneficial partnership with us.

Cenabast provided us with the following files:

- Transactional data
- Products data
- Clients data
- Market data

The historical transactional data from the past two years of sales which Cenabast made includes what products were purchased, how much was spent on each order, the average market price for the products ordered, the highest price at which the products are sold as well as the order date and many more insights. Furthermore, they provided us with a file on all drugs sold in Chile including detailed information about all products and what they are made of. This was particularly useful when segmenting the clients based on the type of products they were purchasing. Cenabast also provided us with a file of their customers they are providing products to. This file includes the name of the company, the different branches if available, the geolocation, each branch contact person, and many more details on each reseller. The final data we received gives us deep insights into the price distribution of the medications throughout the entire country. This is a file which is not client dependent and can be considered as a service/information we provide to our clients. It compares product prices to the average sales price in the market, the highest price at which the products are sold and further details. Having this data adds value to our company and makes us more reliable when clients choose a suitable SaaS.

In return for Cenabast giving us access to their data, we are providing them with our platform and parts of the code as well as tailoring some of the tools to their specific needs.

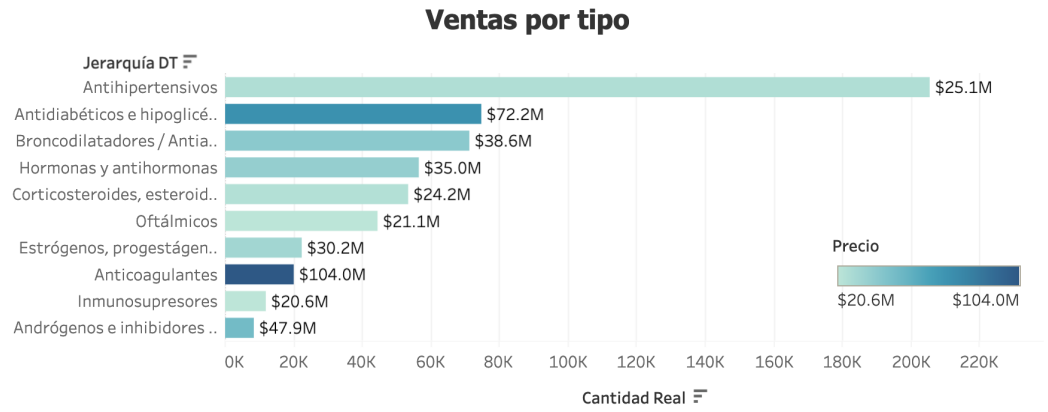
The Prototype

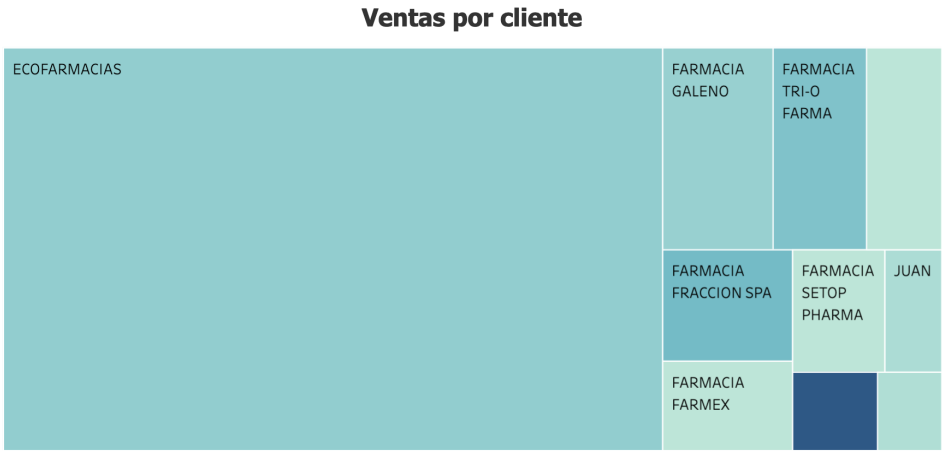
The main idea of the platform is for each of our clients to upload their data to the database and from there, the platform is automatically generated, creating all the analysis and visualizations. The data will have to be in a specific format, where the columns relevant to the analysis will have to be named correctly to match the code. This will either be done by the client themselves by an example provided to them, or the data pre-processing can be offered as a service by DataCare.

For ease of understanding, we will describe the platform now by the example of Cenabast, our first client and partner.

Page 1: Main Page

Once the necessary data from cenabast is uploaded to the database, what is currently a 5 page platform will appear. The first and main page shows some general insights about all the customers that Cenabast has. These customers are from different kinds of pharmacies and drug stores all over Chile. These insights include the sales by client and the most sold drug by type of medication. All of these visualizations are interactive. By placing the mouse over each of them, Cenabast has the possibility of getting some more details. These visualizations were created with Tableau are connected to streamlit through an HTML link.



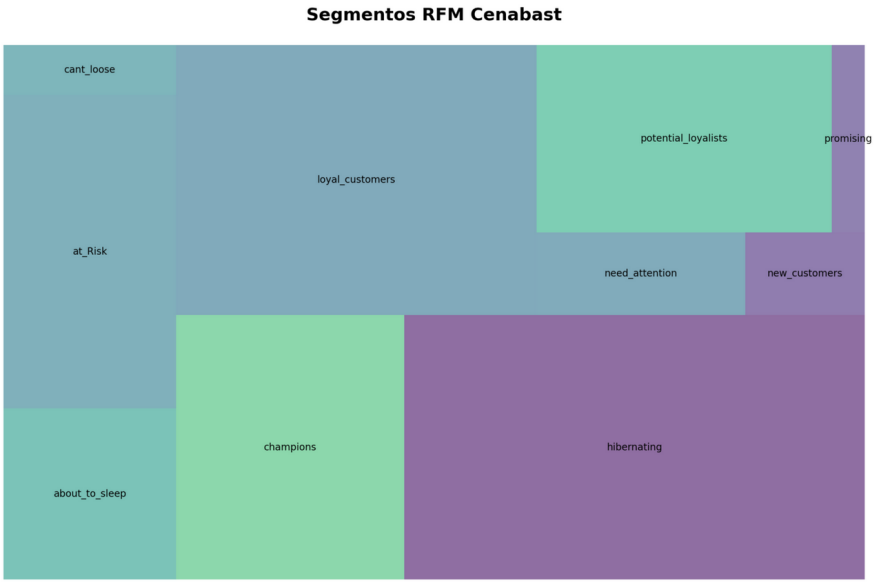


Page 2: Commercial analysis of clients

The second page displays the output of the RFM analysis. There we can see the different customer segments that exist. The size of each square indicates the number of customers Cenabast has within this segment. Here too, we can see that the "hibernating" segment is the most dominant segment of Cenabasts' clients. If Cenabast is then interested in their customers which have been segmented into a specific segment such as the "about to sleep" segment, they can select the segment of interest and receive a dataframe listing all the pharmacies that have been segmented accordingly. We are also displaying some key information about this customer segment such as the total number of clients it includes, the percentage of clients it includes, number of average days since the last purchase, the average frequency this segment places orders and the total revenue Cenabast generates with this segment. We then provide a definition of the segment as well as recommended steps to take in order to move the clients in the selected segment into a more positive and stable segment. Finally, we display a table including all the customers that fall within the selected segment, providing the user with some key information about each customer.

Segmentación RFM

Segmentación de clientes según su comportamiento respecto a las variables: dias desde la ultima compra (**Recencia = R**), Frecuencia (**Frequency = F**) y Monetario (**Monetary = M**)



Campeones

Total Clientes

49

Porcentaje del total de clientes

13.1%

Promedio días ultima compra

17.47

Promedio de frecuencia de compra

440.14

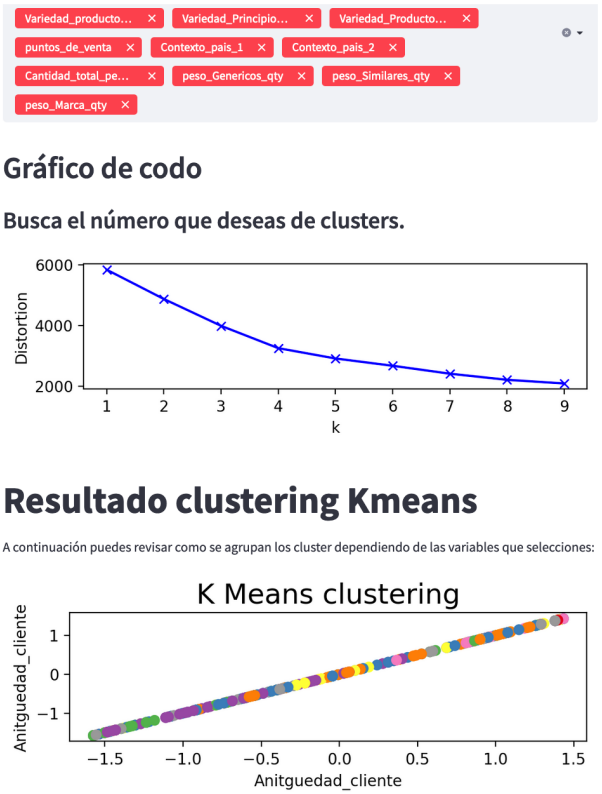
Monto total promedio (CLP)

$42,580,885.12

Page 3: Clustering Clients

Within the third page, we provide the possibility for Cenabast to interact with the customer segmentation based on the data model we created (also visualized on the platform). We are providing Cenabast with a list of features they can select, depending on what they would like to cluster their clients. These features include revenue of cenabast, total units sold, time as client and/or product type and many more. On the left side of the platform, the user can further interact with the clusters by choosing the size of the clusters as well as the total number of clusters the customers should be divided into. We are then deploying the K-means algorithm based on the criteria the user decided upon. The user can also decide upon what to

display in the plot. The revenue of cenabast, the average revenue of the entire market or their clients revenue, all according to time as client. Finally, the client can download a table as an excel file, including details about the client, the products they bought and details on their RFM score.



Page 4: Client statistics

The fourth page concentrates on each customer of Cenabast individually. The user can first select the pharmacy of interest and will then show details on that specific customer split into three tables. The first table includes details like the contact person at the selected pharmacy, the address of each of the branches in case there are multiple, as well as the district the pharmacy is located in as well as the region. The second table displayed, shows transactional information about the selected client such as the date of their first and most recent purchase, the number of years they have been with cenabast, and the total number of orders they have placed within this time frame. Additionally, it displays the revenue which Cenabast is generating with the selected pharmacy as well as the revenue that the pharmacy is generating with the products acquired from Cenabast. Finally, it also shows the revenue if the products were sold at the average market price. The final table is the results of our recommender engine (described in further detail below), displaying all the products that we believe would be useful to the pharmacy and should therefore

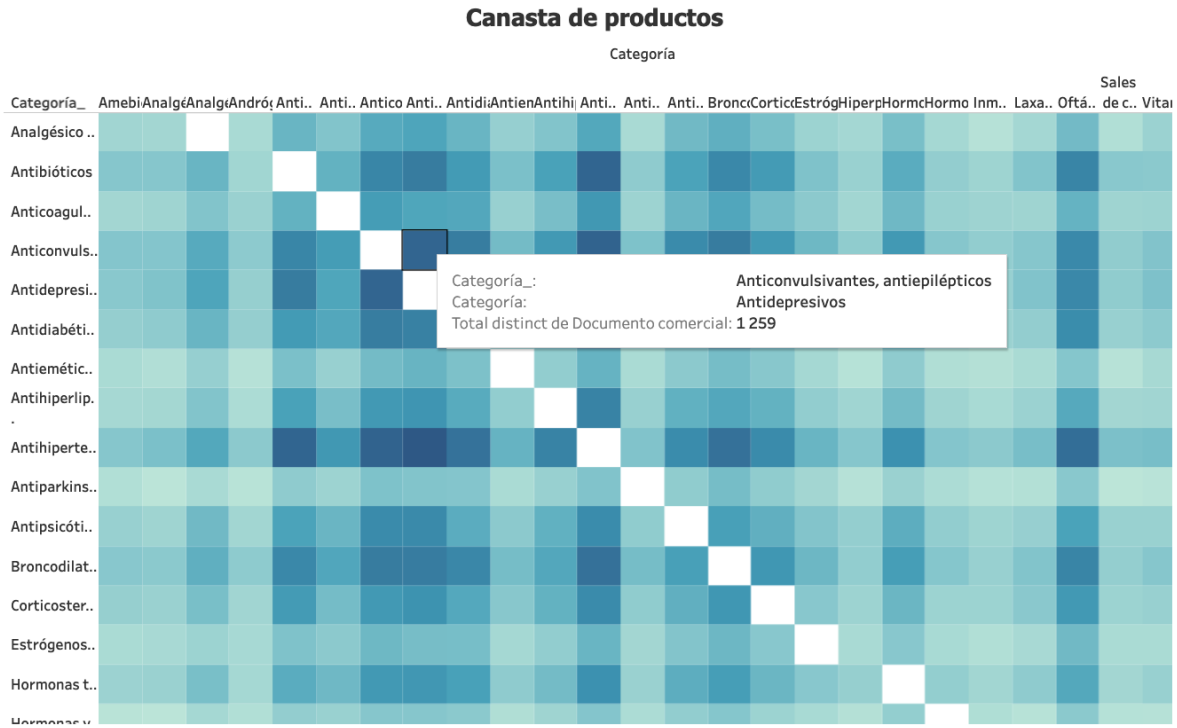
be an incitement for Cenabast to use in order to increase the personalized communication and support in increasing sales for the pharmacy.

ID Cliente	ZGEN	Nombre producto genérico	Jerarquía Cenabast
203913	100001289	SERTRALINA 50 MG CAJ 30 CM	Antidepresivos
203913	100000546	ENALAPRIL 20 MG CAJ 20 CM	Antihipertensivos
203913	100000378	CIPROFLOXACINO 500 MG CAJ 6 CM	Antibióticos
203913	100000694	FUROSEMIDA 40 MG CAJ 12 CM	Antihipertensivos
203913	100000415	CLOTRIMAZOL 1% 20 G CAJ 1 TU	Fungicidas o antimic
203913	100000806	IBUPROFENO 400 MG CM CAJ 20 CM	Analgésico Antipiréti
203913	100000806	IBUPROFENO 400 MG CM CAJ 20 CM REC	Analgésico Antipiréti
203913	100000168	ATORVASTATINA 20 MG CAJ 30 CM REC	Antihiperlipidémicos
203913	100003219	D-HISTAPLUS 5 MG CAJ 30 CM	Antihistamínicos
203913	100001232	QUETIAPINA 25 MG CAJ 30 CM	Antipsicóticos
203913	100003264	SULIX 0,4 MG LIB PROLONG CAJ 30 CP	Hiperplasia prostátic
203913	100003264	TAMSULOSINA 0,4 MG CAJ 30 CP LIB PROL	Hiperplasia prostátic
203912	100003210	ESCITALOPRAM 10 MG CAJ 30 CM	Antidepresivos
203912	100002057	NEVINEX 75 MG CAJ 30 CP	Anticonvulsivantes,
203912	100001289	SERTRALINA 50 MG CAJ 30 CM	Antidepresivos

Page 4: Market Basket Analysis

The final page displays the results of our MBA in the shape of a heatmap. Only the most relevant and correlated categories are displayed. Further detail is provided to the user when the mouse moves over the map.

The Models



The DataCare platform uses 4 main algorithms to perform analysis:

1. Recency, frequency, and monetary analysis (RFM)
2. Apriori algorithm for market basket analysis (MBA)
3. Recommendation engine
4. Customer/Product segmentation with unsupervised ML

1. RFM Analysis

By leveraging the transactional data, we are segmenting the customers of our client depending on how recent their last purchase was, how frequently they are placing orders with the laboratory and how much money they are spending on each order. By giving different weights to each category depending on what's most important to the laboratory, a score for each customer is calculated. This allows the laboratories to detect customers which might be particularly loyal customers or customers they are about to lose. It shows those customers that used to spend a lot of money but have not placed an order in a long time, and also new customers which might not be spending a lot of money yet. Depending on the segment the customer was categorized into, there are different recommendations of actions which we also provide. Within the platform, we are visualizing the size of each segment for each of our clients and depending on which customer the laboratory wants to focus on, we then provide the according suggestions of how to improve

this situation. As an example, for the ‘about to sleep’ segment it is important to reach out to the customer through personalized communication in order to keep them. For the ‘can’t lose’ segment, we would for example suggest them to make product recommendations to their customers in order for them to increase their sales.
The customer segment ‘champions’ are considered very loyal customers and could be leveraged to promote the laboratory to other pharmacies.

2. MBA

For our market basket analysis we used the Apriori algorithm. It’s an algorithm for association rule learning and frequent item set mining across relational databases. As long as such item sets exist in the database frequently enough, it moves forward by detecting the frequent individual items and extending them to larger and larger item sets. The association rules that highlight general trends in the database can be created using the frequent item sets discovered by Apriori.

For our big data prototype, we applied the algorithm to our data and decided to use Tableau to provide an interactive visualization for our clients in the form of a heatmap. This allows for quick identification of item sets.

3. Recommendation engine

The recommender engine we built is based on a unique calculation for each product, individual for each pharmacy. The rating includes the recency the product was bought, the quantity of the purchased product and the amount of money that was spent. Each of these categories are dependent on the client on that product. We then trained a matrix factorization model to predict these ratings for products based on each client. The top recommended products for the selected client are displayed on the platform.

4. Segmentation
From the transactional information, product information and market information, we receive from our client, we create a new data model with which we are segmenting the customers considering the commercial information (e.g. revenue) and the product details (e.g. what kind of product is it: generic or brand). The results are displayed on our platform as an interactive tool allowing our client to adjust the number of clusters and the variables they believe are relevant depending on the insight they want to gain from the segmentation.

The Infrastructure

The product we developed is dependent on the data provided by our clients. All of the raw data is loaded into a PostgreSQL database which is located in a docker container. It's a secure database protected by passwords.

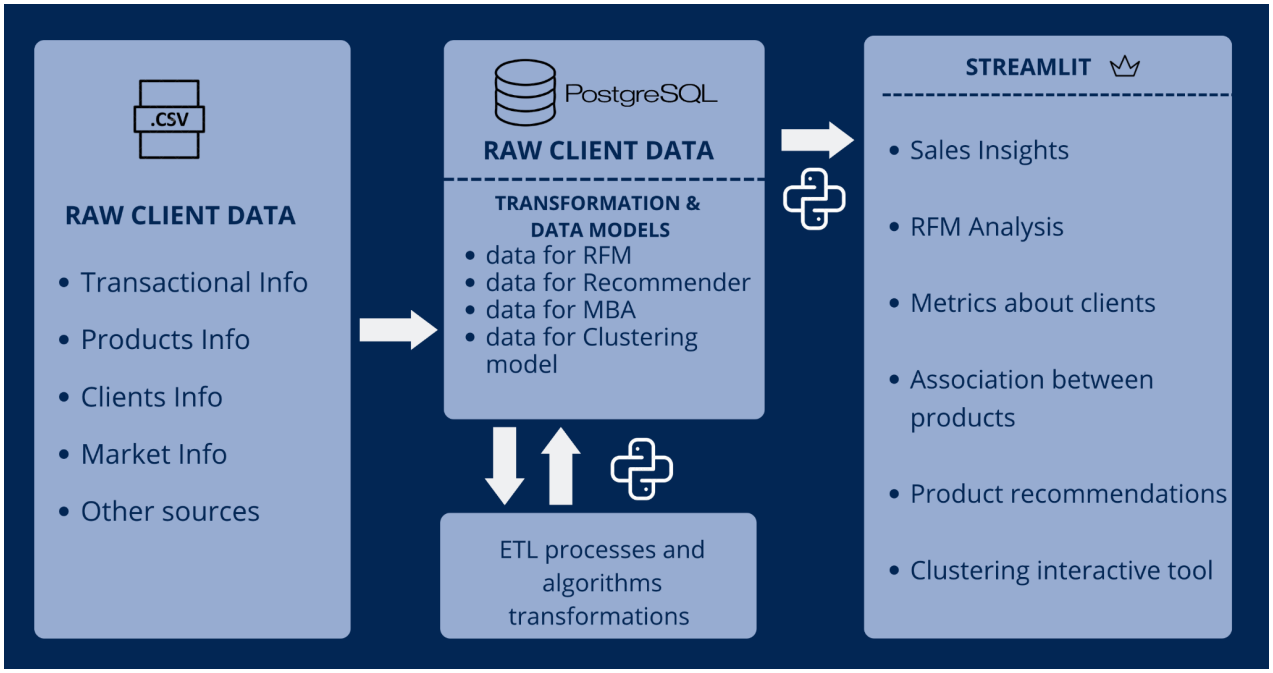
Next we are using an ETL process to then run the RFM analysis as here we are only transforming the data. We then run functions to run the data as a visualization in streamlit.

Another ETL process is created for the recommender system. First we are creating a data model that is based on the input of the user depending on which client they want to receive product recommendations for. This data model is then used to run the recommender engine within streamlit, which then displays the top recommendations for the selected customer in the user interface.

For the MBA we are running the apriori algorithm outside of streamlit, in the production environment, and then putting the results back into the database which is considered a new dataset. Finally, we are creating the heatmap visualization for the front end inside of streamlit.

We are creating the data model for the customer clustering algorithm, that includes some of the features chosen by our client. We are taking the data from the database, using an ETL process and then loading the resulting data back into the database. As we want to create an interactive application for our clients, we are giving them the option of interacting with the clustering tool. Selecting the features is like querying for the specific data. Once we have the data, the cluster is run in streamlit considering the parameters chosen by the client.

Once we run the functions that get the data from our database, we are saving processing time in Streamlit as we are saving the data into the cache memory. This action improves the speed of our application.



The Security Architecture

Since we're working with real data and sensitive information from real clients, cybersecurity is of utmost importance even before deployment. During the development and testing phase, we can't neglect our network architecture, as we learned in class. We need to incorporate it into the design and protect ourselves from any vulnerability we may foresee. These are technical and procedural controls we are working on.

Preventive controls

- User authentication to access the demo
- User authentication to access
- MySQL IP whitelist to access MySQL
- Streamlit secrets TOML file
- Passwords and tokens only shared with necessary members
- Privacy agreements — user specific access to their own data
- Streamlit app designed against SQL injections
- Regular password changes

Detective controls

- MySQL logs
- Streamlit logs
- MySQL IPS

- MySQL Firewall
- GitHub commits history reviewing

Corrective controls

- Encrypted data backups in each of the team member's local hard drive
- Change of Streamlit secrets file
- Change of MySQL IP whitelist
- Change of MySQL usernames and passwords
- New generation of tokens (i.e. Tableau)
- Generation of new links, new repositories, new access points
- Local anti-virus & anti-malware

Vulnerabilities

Upon deployment, our biggest vulnerability lies in the protocols to generate new passwords for users. As a recommended good practice, we will work on a way to let the user change their password upon generation of their credentials, and afterwards, the user will be prompted to change it every three months. Password forgetting, weak passports, access to user passwords by unauthorized team members are inherent risks that we will continue to address in the next stages of development.

The front-end

As mentioned before, we chose to use Streamlit for our front-end, mainly because it was built specifically for the creation of web applications with Python. Streamlit is an inexpensive solution, especially for a start-up. The development time was a lot shorter than with the traditional framework, and allowed us to focus on the functions we wanted to deliver to our client. It also allowed us to start testing our prototype in early stages as the visualizations of our analysis could quickly be implemented into the frontend.

The app loads the data fast thanks to Streamlit's features including the fact that there is no hidden state and there are no callbacks. There is a simple cache function that improves the speed, as well.

Another advantage of Streamlit is its seamless integration with dozens of useful tools. Aside from the essential Python libraries, Streamlit displays visualizations from JavaScript-based libraries and even Tableau, which we made use of to the extent that our clients asked us to.

Future plans

Within the near future, we aim to create more features to support our clients in reducing their churn. It is also important for us to let our clients understand how much they are truly benefitting from our service. Therefore, we want to add visualizations that track how the churn and sales change. A specific example would be to track how sales change due to the recommendations. How many products are purchased based on the recommendations and of these, are the products purchased repeatedly? By showing the client how much more they are profiting due to the service, they will be more likely to be loyal to us and also willing to pay for the service in case our monthly fee increases over time.

Furthermore, in order to grow, we need to acquire new clients constantly. By acquiring new clients, we also gain further insights of needs and will be able to develop further tools of analysis. Currently, our product has only been validated in Chile, however once reaching clients within other regions, our product is just as well suited to fulfill their needs. While the analysis tools have been built with a focus on supporting vendors within the pharmaceutical industry, the algorithms we have developed and employed can be of great advantage for many different sectors. Therefore, once we have signed the first paying clients and have received good feedback on the usefulness of our application, we will begin with our market expansion pains.

Final thoughts

To sum up, through this project, our team aimed to reflect all the concepts that we have learned in regards to Big Data and AI. We have aimed to solve a real problem in the market for a public institution in Chile. During this journey, we have applied many of the notions learned in class: working in an Agile way using Jira; using GIT as a version control, implemented security controls; used a range of ML algorithms and data processing techniques on large data sets.

References

Afshar, V. (2017, December 7). *50 Important Customer Experience Stats for Business Leaders*. HuffPost. https://www.huffpost.com/entry/50-important-customer-exp_b_8295772

CaixaBank, & Ibáñez De Aldecoa Fuster, J. (2022, June 30). *La industria farmacéutica española*. La industria farmacéutica española. https://www.caixabankresearch.com/es/analisis-sectorial/industria/industria-farmaceutica-espanola

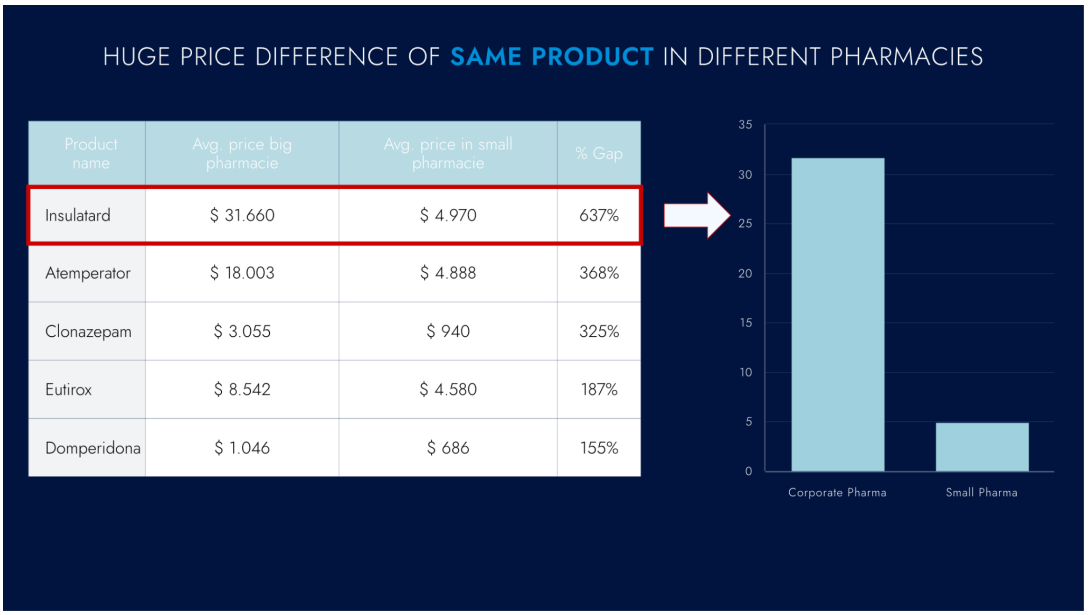
Cenabast. (2022). *Transactional data, Customer data, Market data, Products data* [Dataset].

IT Digital Media Group. (2022, April 6). *El 73% de las empresas farmacéuticas incrementaron sus ingresos digitales en 2021*. Estrategias digitales | IT User. https://www.ituser.es/estrategias-digitales/2022/04/el-73-de-las-empresas-farmaceuticas-incrementaron-sus-ingresos-digitales-en-2021

Ranking Empresas Fabricación de especialidades farmacéuticas | Ranking Empresas. (2021). Directorio Ranking Empresas - Ranking de las principales empresas españolas. https://ranking-empresas.eleconomista.es/sector-2120.html

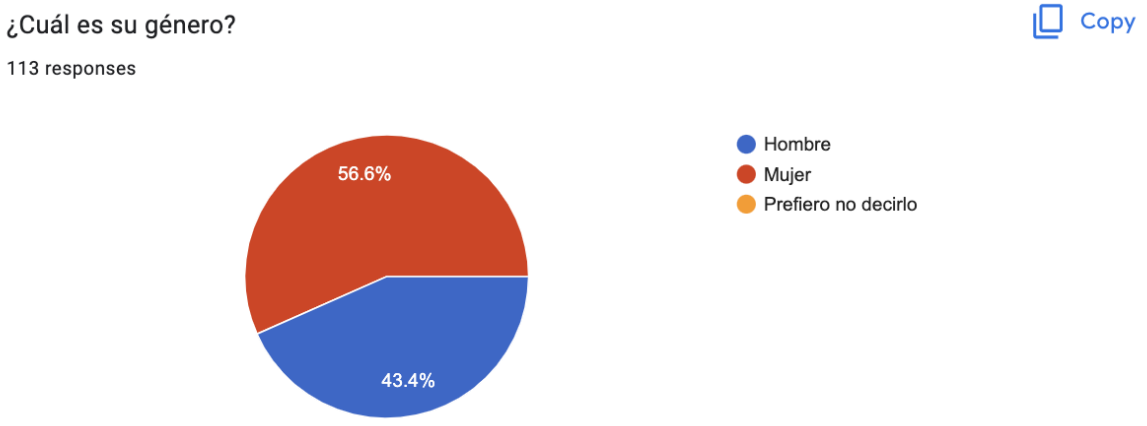
Appendix

Appendix 1: Initial project idea validation



Appendix 2: Interview for the initial solution

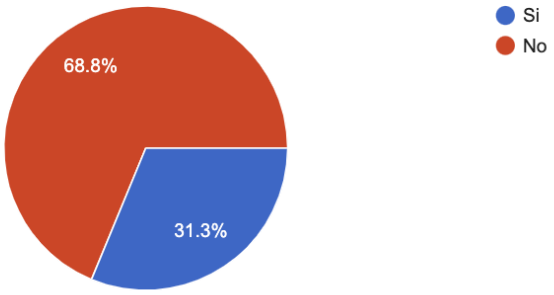
For validating the initial idea, we have conducted a survey in order to understand and study how Chileans are buying and using medicaments.



¿Tienes una enfermedad crónica?

Copy

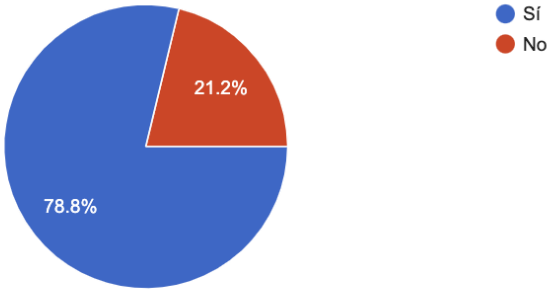
112 responses



¿Sabes qué son los productos Genéricos?

Copy

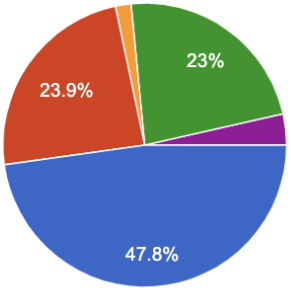
113 responses



¿Cual es tu ocupacion?

Copy

113 responses



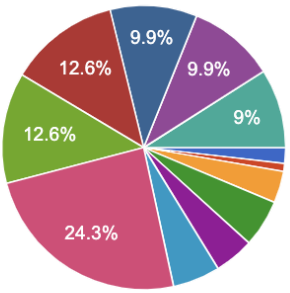
- Trabajador/a dependiente
- Trabajador/a independiente
- Desocupado/a o desempleado/a
- Tercera edad
- Estudiante

¿Cuál es aproximadamente su ingreso mensual?

Copy

Copy chart

111 responses



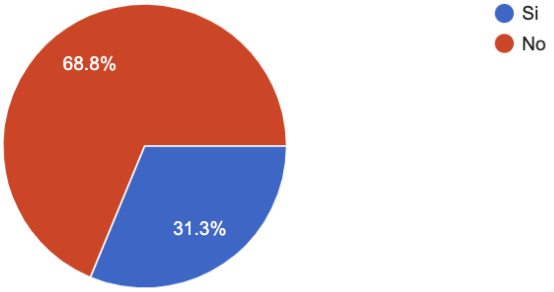
- Menos de $ 500.000
- $ 500.000 - $ 750.000
- $ 750.000 - $ 1.000.000
- $1.000.000 - $ 1.500.000
- $1.500.000 - $ 2.000.000
- Sobre $ 2.000.000
- Sobre 2000.000$
- 1000.000$ - 1500.000$

1/2

¿Tienes una enfermedad crónica?

Copy

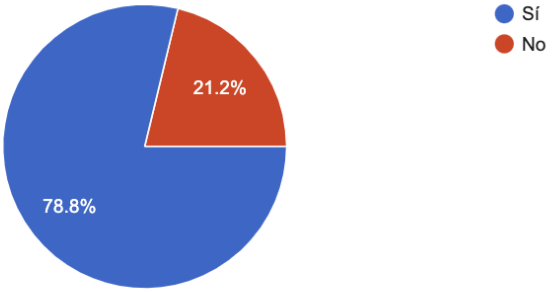
112 responses



¿Sabes qué son los productos Genéricos?

Copy

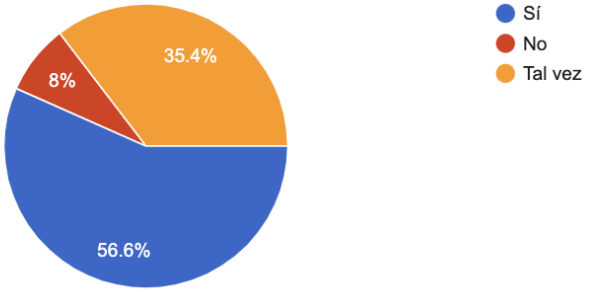
113 responses



¿Prefieres productos genéricos?

Copy

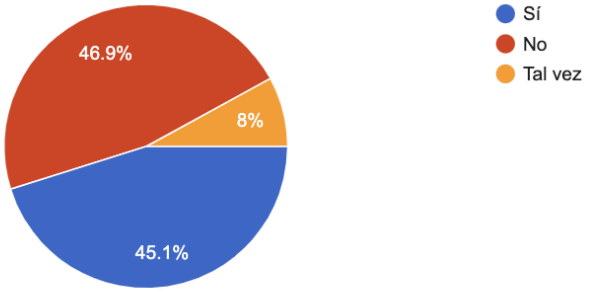
113 responses



¿Requiere constantemente medicamentos en su vida habitual?

Copy

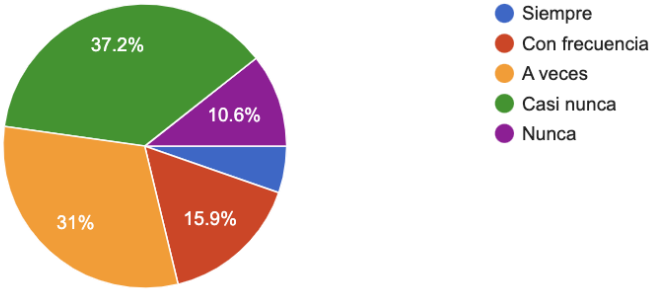
113 responses



¿Con que frecuencia cambia la marca de los medicamentos que compra?

Copy

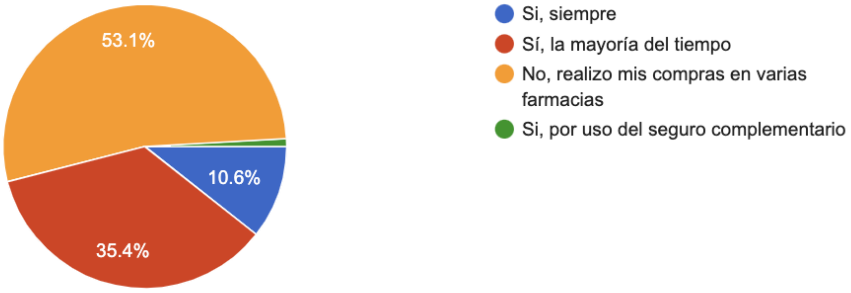
113 responses



¿Sueles hacer compras en una sola farmacia?

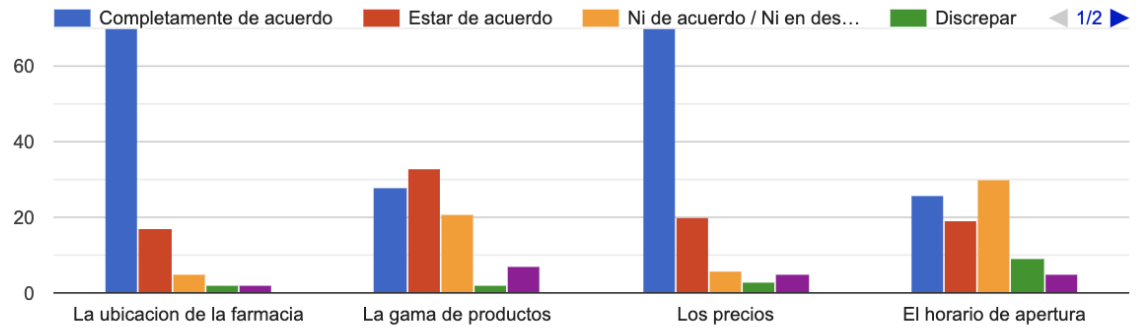
Copy

113 responses



Mi selección de farmacias está influenciada por...

Copy



Cuando compra medicamentos de venta libre (sin receta medica), por lo general:

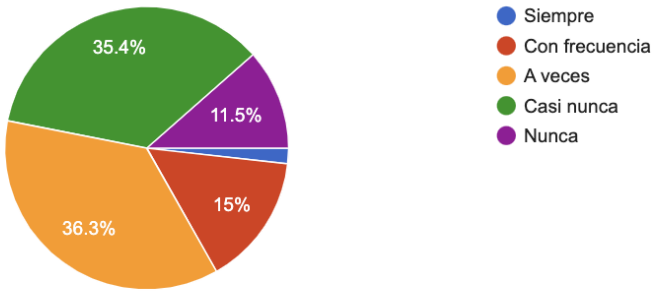
Copy

113 responses

¿Con qué frecuencia recibe recomendaciones del farmacéutico?

Copy

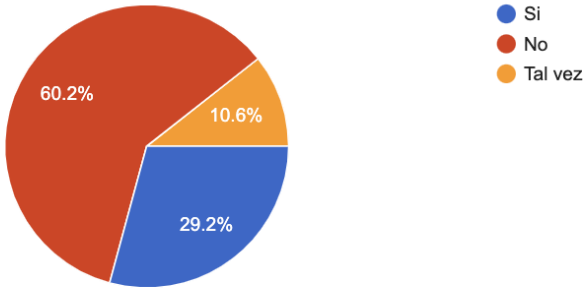
113 responses



¿Usas habitualmente tu dispositivo (Smartphone, Tablet o PC) para buscar una farmacia?

Copy

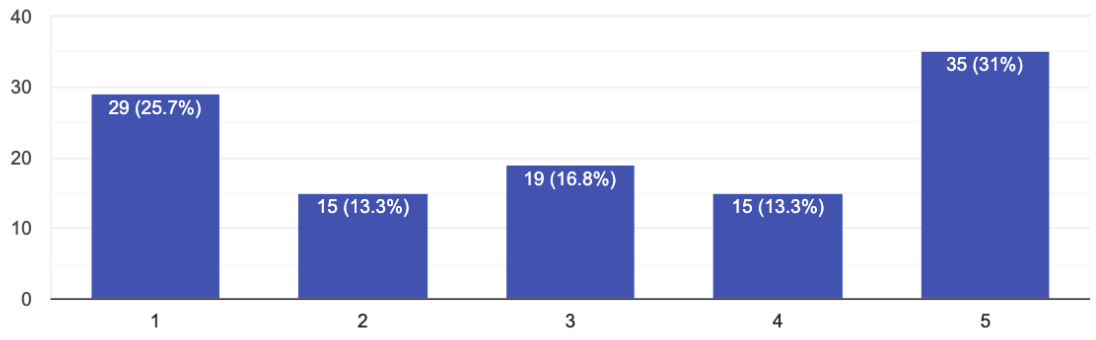
113 responses



¿Qué tan probable es que utilice una aplicación web que localice la farmacia más cercana con el precio más barato para sus productos?

Copy

113 responses



Appendix 3: For our initial project idea, a main component was to gather product information from pharmacies online shops. We had used web scraping in order to gather more data related to the description of medications, difference in prices, location,etc. For the web scraping we used the Selenium library. The purpose was to use this information in order to find out which pharmacy sold the products of interest of the user at the lowest price. We faced many challenges in doing this process, as every website has a different structure, so code adjustments had to be made for every pharmacy. Also, we were not able to find out if the website has been updated regularly, so we were at risk of promoting products that were not available in the stock. We then understood that this process was time consuming and not scalable. Here the data we extracted for Fraccion pharmacy in Chile can be found as an example

.

Once we pivoted the project goal, all of this gathered data was not necessary anymore. We could have still used some insights about the price distribution for our current solution, however, it was more convenient to use the data which Cenabast provided us with (they have access since they are a governmental institution).

Appendix 4: Interview with Indopharma

The supplier is an indian company, a special stock is needed in the warehouse: ask product -> sell it further ; the indian lab needs at least 5 months to send to Chile; The stock is chosen on the contracts , currently there are 30 products registrated with the Chilean law and there is a regular demand every month; Excel forecast how the demand for the next year will be; there is a problem with forecasting exactly the needed stock; In case of an urgency the products are sent via plane : 3 months to create the products with the standard of Chile and 2 months to ship to Chile; Usually the number of units per products is between 200,000 ; the life of one product is 36 months

24 to 48 hours to have the products delivered to pharmacies; They have a delivery company that sends the products with appropriate delivery methods ; In case of an urgency they send the medicine in a couple of hours ~2-3 hours or the hospital goes straight to the warehouse

He feels there is no problem with the current way of ordering products, but there have been situations when there was an overstock of products;

The pharmacies order the same product but not always the same quantity; Hospitals, private hospitals, drug stores, pharmacies ; there is a minimum amount to sell products to the pharmacies other wise there is no partnership; there is difficult to understand the behavior of little pharmacies - WE HAVE DATA BY SEGMENT; there is no publicly data available to find data on private vs public system;

Little pharmacies place orders once a month , drug stores three times a month , hospitals between 1-2 times a month, cenabast once pm ; actively growing clients ;

There is no way of targeting new pharmacies right now ; there could be channels of marketing; they do not have a list of pharmacies to reach out to ; There are products ODC that can be bought with no prescription and you can advertise those on TV;

Trending products: by contract they need to provide future demand (next 12 months) per month; little pharmacies do not have this restrictions ; Levels of priority: hospitals have always the same no of products, pharmacies are very;

There is no way of understanding which products are most sold; everything depends on the population of people that live in the specific area ; there are also hospitals with specializations and they would have different products

We know which products are sold the most ; the ones that bring more revenue and which not

The market is very aggressive; it's expensive to buy different data; exploit the data; lack of abilities to work with big data ;

Report to understand the demand

Reaching out to customers through phone calls; there are 2 segments of clients : the ones that have their system automatized (SKU); the little Pharma / drug stores do not have SKU so they call ; sometimes they forget to ask for the product through a request ; file that tracks all the calls with the pharmacies.