Watch Tower

A Forecast Analysis to look out over Teaté's Minimarkets Community

DS4A 2020 - TEAM 77

David Martin, Camilo Peña, Nicolás Rey, Darwin Fonseca, David Peña, Camilo Cano, Maria Quintero

Watch Tower

A Forecast Analysis to look out over Teaté's Minimarkets Community

Watch Tower will help Teaté to look out over some vulnerable minimarkets community to better drive their digital transformation and rebuild their social fabric

Executive summary

Small businesses in Colombia, located mostly at middle- and lower-class localities, are dedicated to real selling of massive consuming goods. The owners of these establishments are micro-merchants engaged in the retail sector, particularly in Fast Moving Consumer Goods through a traditional channel that contributes more than 50% of the sales of mass consumption in our country. In order to provide micro-merchants with digital channels that directly connect them with product manufacturers and mass consumption services, new companies like Teaté are connecting suppliers of goods and services with micro-merchants, thus allowing thousands of shopkeepers to supply their businesses through a mobile application. Although these initiatives could improve the efficacy of the entire retail distribution channel, these companies have some needs and gaps to fill. In particular, Teaté needs to gain knowledge about the market behavior and the demand of shopkeepers in order to avoid inefficiencies in the process and consequent losses of time and money. In this project we will use a Big Ben VEC Forecasting Algorithm to predict the demand in the marketplace using data sales from Teaté. We expect the results of this project will allow the company and its clients to allocate resources appropriately, plan their strategy for the future and take better decisions about the investments required to accomplish their goals. We think this would contribute to engine the social and economic growth in Colombia as we are supporting a strategy with the potential of rebuilding the social fabric along with the digital transformation of the micro-merchant retail sector.

INTRODUCTION

Currently, more than 10 million people work in the 3 million businesses and stores in Latin America. According to Fenalco, in Colombia there are about 266 thousand stores that engine the national economy, which has been in decline during the crisis and have forced the retail sector to reinvent itself and adapt to a growing need for modernization [1]. In response to this need, several digital platforms are offered for free to support small and medium-sized businesses, such as neighborhood stores, so that they can reach consumers through a digital channel to maintain, and even increase, their sales, and to reduce the impact of mobility restrictions due to health emergency during the current pandemic.

In a recent study, the Multiservice Technological Platform shows a total of 450,000 neighborhood stores or commercial establishments in the country that have been favored by the use of these platforms [2]. They not only support the shopkeepers to place their orders at any time, but also allows them to generate additional incomes of up to 800,000 per month. In addition, the initiative strengthens the relationship within shopkeepers and serves as the starting point for the emergence of a retail community where everyone benefits (table 1).

On the one hand, the retail business in Colombia, which in recent years has been involved in an increasingly competitive ecosystem with the entry of new players, can benefit from this initiative as it may compete by offering consumers a greater diversity of products while facilitating the selection of the best purchase option [2]. In this way, new strategies can emerge for the retail business to retain and conquer new customers. In addition, these platforms benefit shopkeepers by creating an efficient communication channel not only with clients but also with providers of goods and services that includes large brands.

On the other hand, suppliers of goods and services benefit from higher profits as they can directly access the micro-merchant community, thus optimizing incomes and outcomes, connecting their brands directly with sellers and customers, increasing the margin of their business within a collaborative model, exploring the possibility of expanding their market, and all the benefits that the digital economy can provide.

Lastly and most importantly, the entire society benefits from a digitalization process as small businesses get linked to social initiatives. These initiatives provide strategies for all the actors of the society to improve the civic culture and rebuild the social fabric. In particular, strategies such as the empowerment of shopkeepers have the potential to drive the people to learn and transform themselves and their families through activities that can generate important changes in their community. In this effort, companies, local governments and NGOs would join forces to support the strengthening of community initiatives that promote peacebuilding in the rural and urban areas of regions where the digital strategy is implemented. As these kinds of initiatives are usually implemented by leading shopkeepers who are part of the retail sector, the community around them and their families have been highly benefited [3].

Finally, being able to predict demand in the marketplace would optimize costs and losses, would reduce inventory costs, people allocation and hiring. Thus, in this preliminary work we want to predict the demand for small business sales to allow Teaté, a digital community with this kind of digital strategy, and its members, to allocate resources appropriately, plan their strategy for the future and take better decisions about the investments required to accomplish their goals. With this, we expect to indirectly contribute to engine the social and economic growth in Colombia by supporting strategies pointing at the digital transformation of the micro-merchant retail sector.

Table 1. Benefits for the main actors of the retail sector

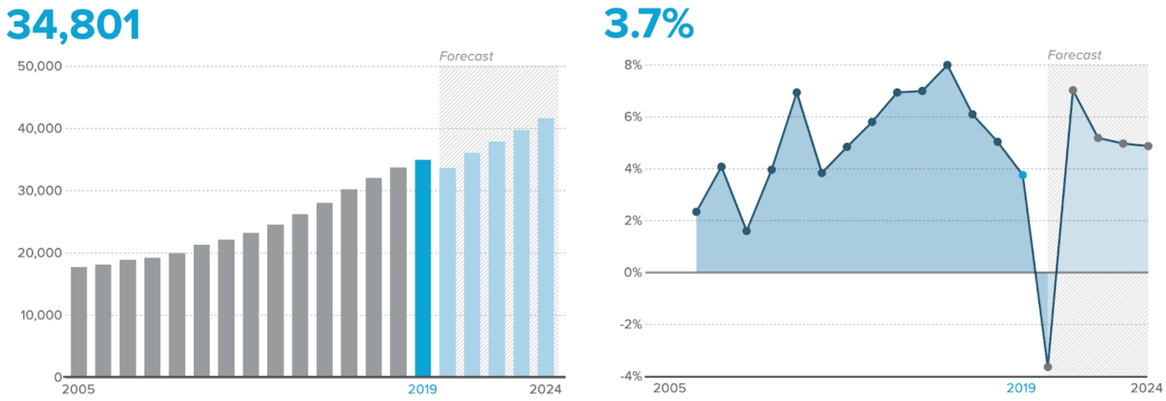
Small business	**Providers**	**Communities**
Digital transformation consolidated orders save time Place orders 7x24 Direct connection with suppliers direct discounts Access to microcredits Portfolio expansion Higher earnings	Direct connection with shopkeepers Connecting brands directly Increase the profit margin Expansion of the market Market data capture Market data analysis Entering the digital economy	Price reduction Strengthening communities Vulnerability reduction (loan sharks) Generation of social fabric Lower cost of living Improve supply chain profitability Community leadership Change management

Business understanding

About the retail market

Small retailers adjust their practices as a strategy to adapt to change, which makes them follow consumer demand and remain important in their local communities. In recent years their effort has been so important that it is reflected in a sales trend that seems to grow even when the distribution channel of this market is one of the most fragmented in Colombia, where the biggest player does not even account for a 1% of value share (figure 1). Despite the fragmentation of the channel, traditional grocery retailers' sales have increased by 4% to COP114.4 trillion in 2019. Indeed, traditional grocery retailers are expected to post a current value CAGR of 6% (3% 2019 constant value CAGR) over the forecast period, to reach COP151.2 trillion in 2024.

Fig. 1. Sales performance of traditional groceries retailers



*from Euromonitor International website

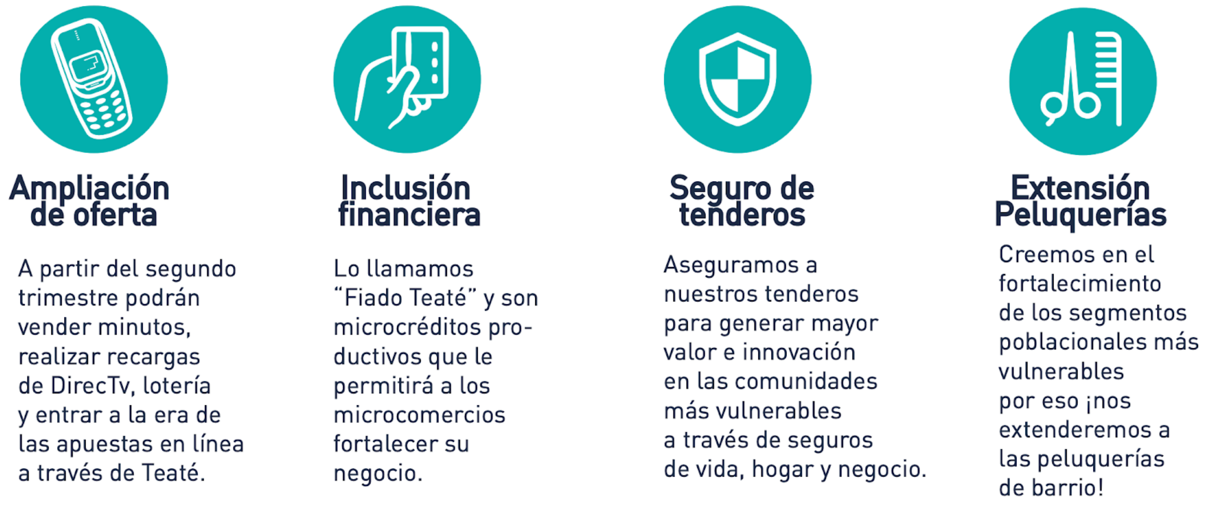
According to the most recent Market report from Passport [5], the Traditional retail channel will grow on average at a 6% rate year over year and it will be a 40 thousand Billion USD market in 4 years. However, given the uncertainty of these times, traditional grocery retailers are growing faster than predicted in the post pandemic cycle, along with a slight rise in outlets. These outlets remain crucial in the local retail environment and are particularly popular amongst lower-income consumers.

Digital transformation in the retail sector

Currently in Colombia, neighborhood stores still have a special role in the sale of products from the family basket. It is estimated that in large cities they capture 48% of the market, while in small cities they reach 62% [6]. Despite this, Colombians are migrating towards electronic commerce as a purchasing mechanism, a technology that is not usually found in conventional neighborhood stores. For these reasons, the use of technological platforms that allow consolidating the digital transformation in neighborhood stores has gained special importance in times of pandemic. Platforms such as *Tienda Registrada*, *Lapp Tienda*, *Mercadoni* and *Teaté*, are only few of the new companies that have emerged in the main cities of Colombia to provide technological tools that make the job of shopkeepers more efficient.

In addition, as a consequence of the increased competence between small retail business due to the pandemic, independent small grocers have been forced to innovate. To give just an example, they have learned to collect enough information about their clients and local people in order to select who they are eligible for owing products and even offer private credit to their customers. Moreover, as part of the adjustment strategy to face the growing competition from other channels, traditional grocery retailers are moving beyond selling typical groceries, by offering services such as paying utilities, purchasing vehicle insurance, recharging public transport cards and buying lottery tickets. In particular, as part of this trend *Teaté* have developed a specific product line to support micro merchant with different services (Fig 2).

Figure 2. New services offered by digital platforms like Teaté



Ampliación de oferta

A partir del segundo trimestre podrán vender minutos, realizar recargas de DirecTv, lotería y entrar a la era de las apuestas en línea a través de Teaté.

Inclusión financiera

Lo llamamos "Fiado Teaté" y son microcréditos productivos que le permitirá a los microcomercios fortalecer su negocio.

Seguro de tenderos

Aseguramos a nuestros tenderos para generar mayor valor e innovación en las comunidades más vulnerables a través de seguros de vida, hogar y negocio.

Extensión Peluquerías

Creemos en el fortalecimiento de los segmentos poblacionales más vulnerables por eso ¡nos extenderemos a las peluquerías de barrio!

*Taken from Teate's website

In line with this, to accept private branded cards from large stores such as Éxito, Alkosto and Colsubsidio have increased the possibility of purchase for clients, so small grocery stores have now the opportunity to become non-banking correspondents, which ensures them a greater customer flow. In addition to this strategic adjustment, bulk sales and packaging of complementary products such as a chocolate bar with a small milk carton promote household consumption.

Such strategies are expected to maintain strong growth for traditional grocery retailers from 2020 to 2024. With all these strategies, independent small grocers benefit as they centralize purchases and payable accounts with only one provider like Teaté. In addition, the provider can monitor product turnover through a proprietary information system to manage the operation, including all the new brand promotional material and signs, as well as advice on improving the store layout.

At the same time, all the actors of this market ecosystem can take advantage of the government via programs such as *Empresario Digital* or Mipyme Vive Digital, in which traditional retailers are being encouraged to use digital platforms to place orders with suppliers, record sales, manage inventories and accept means of payment besides cash, as well as to sell to customers. Engaging in e-commerce offers players the possibility to increase their competitiveness, taking advantage of internet access and the adoption of apps; many of them free of charge or financed by the government, empowering them to manage and modernize their businesses.

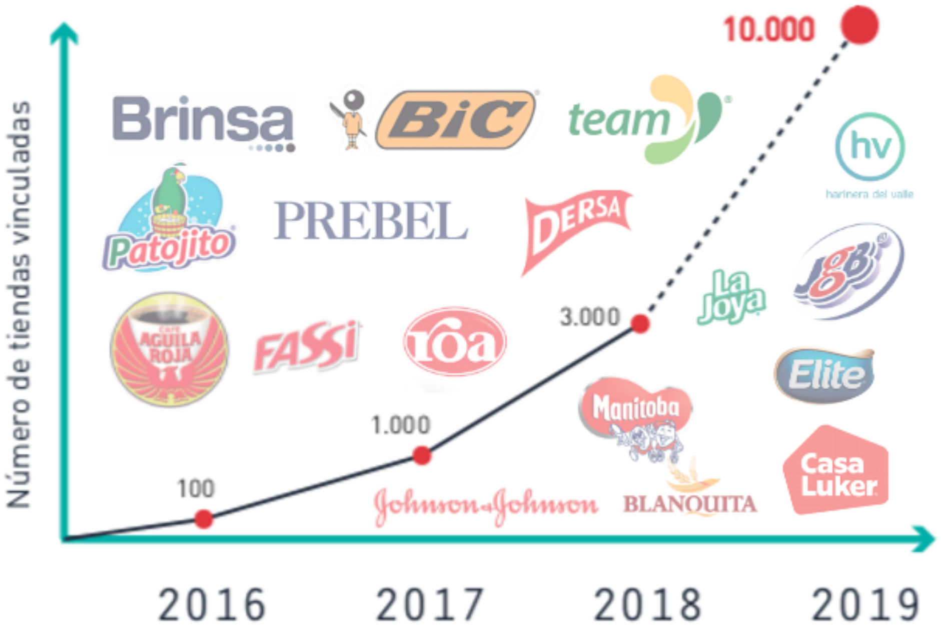
About the company

Teaté was born in 2016 as a startup in Cali, Colombia, and has consolidated itself as a digital community that directly connects micro-merchants with consumer goods and services companies. For this, Teaté seeks to transform the traditional channel through a platform that connects shopkeepers or small merchants with large companies, thus facilitating the supply process for stores [4].

Through their operation and the efficient use of technology Teaté strives to reduce the cost of living of people by making the retail chain of traditional supply a lot more inexpensive for micro merchants and for the families they supply. They do this by directly connecting suppliers of goods and services with micro-merchants to supply their businesses at the lowest costs. For that purpose, Teaté is creating deals with multiple convenience stores around the country to supply them on time, help them future demand and forecast inventories appropriately to supply local businesses.

With this strategy, Teaté has managed to maintain continuous growth since the beginning of its operation. In 2016 Teate had 3 suppliers for its market, and at the end of their second year, they managed to integrate the operation of 4,200 shopkeepers and 24 large manufacturers while integrating communities around their business. After 4 years of exponential growth they now do business with more than 30 suppliers that include main brands in Colombia (Fig 3), they continue growing in Colombia, start their expansion strategy in Latin America, and are exploring new markets in Asia. This has made the company to be identified as a driver of the economy and a key entrepreneurship opportunity for families, as well as an important part of the social network of local communities.

Figure 3. Growth in stores and related Colombian brands



Competition and digital transformation

The traditional grocery retailer market is a very fragmented channel, with nearly all outlets and value sales included under “others”. Most companies in this category consist of one to less than 10 outlets of what is called “minimercados”. This fragmentation makes the competence between mini-markets to become each day more aggressive. Some relevant competing players of this kind that are worth mentioning, are *Cooratiendas*, *Surtifruver* and other *Fruver* outlets.,

Another relevant competitor for *Teate* is *Alkosto*. Although being a hypermarket, it is recognized by traditional grocery retailers not only as a competitor, but also as an allied minimarket which shopkeepers visit for supplies. Alkosto supports traditional grocery retailers as well by creating a shopping club or association, so that outlet owners can buy in bulk, improve their offer and produce a marketing plan. This is done through what is called *Red Contigo*, whose participants are increasing and strengthening their portfolios while incorporating practices such as accepting debit and credit cards. Finally, other players like Bancolombia and TPaga are offering QR payments local payment and fintech applications with no associated costs to convince grocers to adopt this solution to receive electronic payments. Furthermore, digital marketplaces like Rappi and uber Eats are encouraging traditional grocery retailers to join the digital economy on their own store APPs/platform.

Strategic Partnerships

Big consumer brands keep traditional grocery retailers in their distribution landscape, as one of their best allies. In turn, by offering well-recognised brands, players in traditional grocery retailers have a good base for their product portfolios. Strengthening each other as strategic allies allows both to access final consumers, offering them products in a variety of sizes, in the way they require them and at an accessible price. This is possible due to the outlet owner’s knowledge, but at the same time is thanks to the knowledge of suppliers, which sell a wide product variety and quality so that products rotate quickly, thus benefiting both sides. However, some private label lines are now being introduced even in traditional grocery retailers.

Scope of the project

To support Teaté, this project's main target is to provide the company with a tool that allows them to forecast the demand of the products across their market. This tool is based on a Vector Error Correction model that fit the data collected by the company and forecasts time series for SKU categories on sub-zones for Cali and Medellin. As an additional result, this project will deliver a Dashboard that will provide a scheme from where the user can download the model forecast in a csv format so that they can keep track of relevant variables and understand the relationship between them.

Social impact

Not always the society benefits from a digitalization process unless it is accompanied with social initiatives that serve that purpose. In this regard, Teatés' initiative provides strategies for all the actors of the society to improve the civic culture and rebuild the social fabric. In particular, strategies such as the empowerment of shopkeepers have the potential to drive the people to learn and transform themselves and their families through activities that can generate important changes in their community. In this effort, companies, local governments and NGOs would join forces to support the strengthening of community initiatives that promote peacebuilding in the rural and urban areas of regions where the digital strategy is implemented. For that reason, we feel motivated to help Tetaé as we are indirectly contributing to engine the social and economic growth in Colombia by supporting a strategy that has the potential to rebuild the social fabric of vulnerable communities of the retail sector.

DESCRIPTION OF THE DATA

We obtained the entire data from Teaté, which came out in several installments:

- .csv files with the data of some variables related to the orders from the last 7 months of 2020.
- .cvs file with information on 17 months of sales (including 2019)
  Database of all clients
- New database that includes some months that were not in the previous installment (February, March, May and June 2020)
- 4-year annual database.

From those files we made some previous consolidations until we finally got two files that were structured enough to start working with. In total we obtained data from January 2019 to August 2020 and some data from September. With the last instalment we finally got a 4-year annual database.

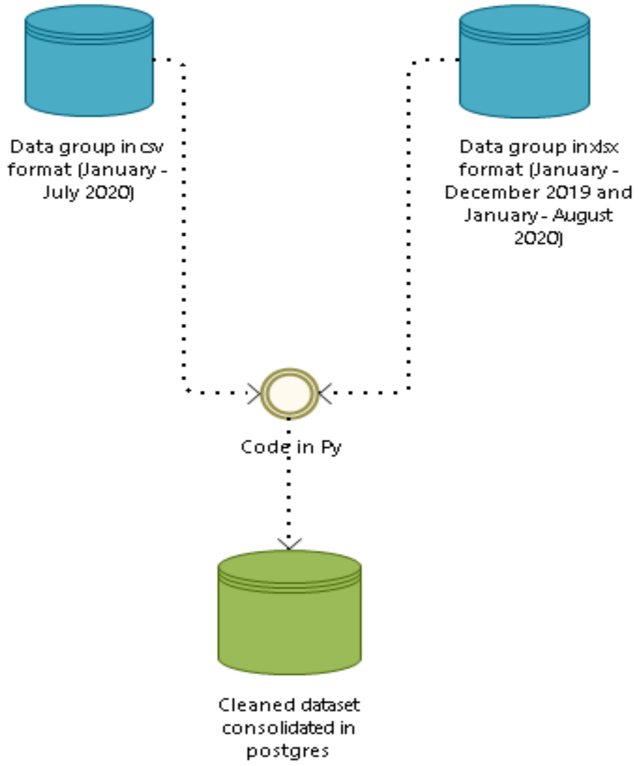
Data wrangling

Regarding the data structure we started with a plain big table with information of vendors, clients and products. The information contained in this table was duplicated for many objects; for example, in a specific month sugar appears 1000 times, so we have 1000 registries of the word "ounces" and for a same unitary price. Therefore, we leveraged the SQL infrastructure to separate this big table into 3 smaller tables with basic information for Order_to_clients, Vendor_information and products.

Moreover, we found numerical fields in different formats (with points and semicolons), different date formats, empty fields, different types of records and duplicate record numbers in the original databases, missing periods of city-level information, missing geographic information, as well as inflated values and important number of weeks of loosed information.

Given the amount and different formats of the incoming data, data wrangling was performed via scripts in Python and SQL (figure 4). We also organized the data so we can visualize relevant variables and possible behaviors or trends associated with sales, products, quantities, among others. It is also important to note that the date formats had differences between files, so this type of particularities were taken into account when cleaning the data.

Figure 4. General steps for data wrangling process


Data group in csv format (January - July 2020)
Data group in xlsx format (January - December 2019 and January - August 2020)
Code in Py
Cleaned dataset consolidated in postgres

Missing values

The consolidated information given by Teaté has 27 fields for which we search for missing values. We found that the complete records are related to the business information, so using queries in SQL we created tables grouped by functions. Besides grouping the complete data, we described the number of records with missing values using the function bar included in the missingo library. While carrying out the review of the consolidated data table, it was possible to determine that some variables had duplicated information and others did not contain complete information. The former variables were considered to not provide valuable additional information to what is already available.

There were also other variables with only null values, for which we decided to remove them from the consolidated data table, and other columns that needed to be analyzed (Figure 5 and Table 2). Based on previous analysis we removed unnamed fields and many other variables with names and values representing nothing. Finally, the consolidated data set contained the most relevant variables for the analysis: Order, Store, Manufacturer, Order date, Invoice date, Store creation date, Order quantity, Order unit value, Teaté Discount, Order entry value, Total value of the order. (See variable details in Appendix 1).

As a result of the data cleaning process, the rows that were found to be completely duplicated were removed from the consolidated database, which means that 1784 rows were removed. In the same way, we selected some variables that we found necessary for the projection and analysis of demand and that were candidates to be completed by filling the missing values using information from previous datasets and from external sources like Google API to capture some geographical information of the clients (Figure 6). Finally, the consolidated base was transformed into a 625,142 records dataset representing purchase orders information collected for more than 11,406 shopkeepers through Teatés' platform from January 2016 to August 2020.

Figure 5. Visualization of missing values for consolidated data set

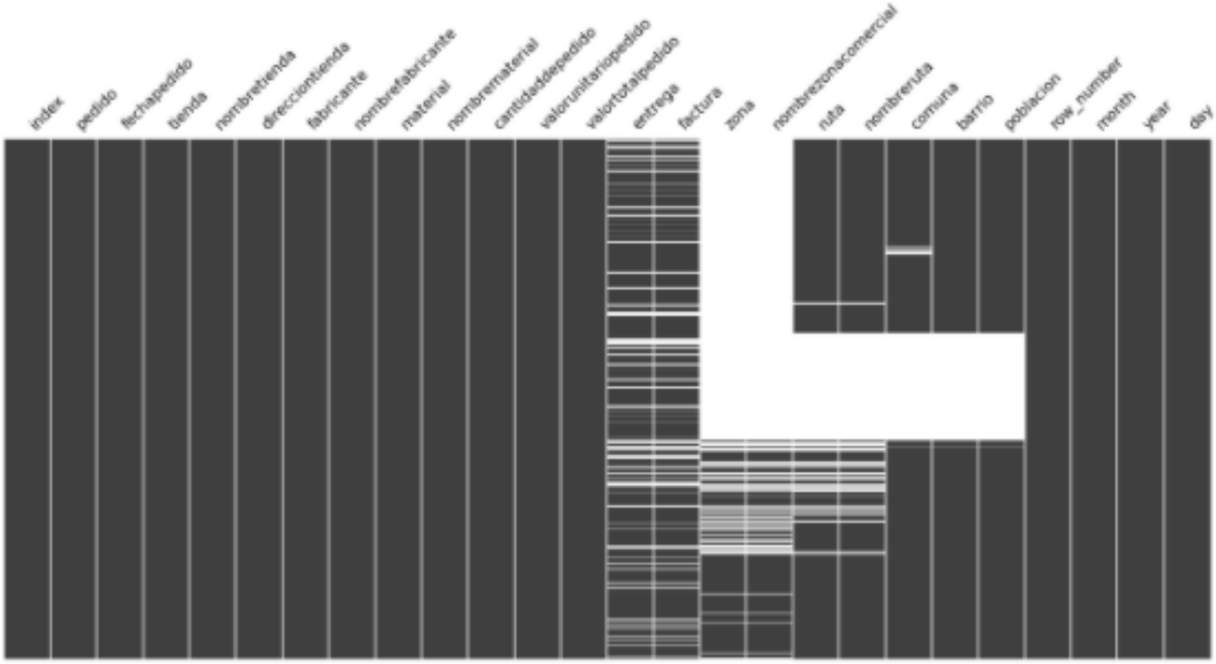
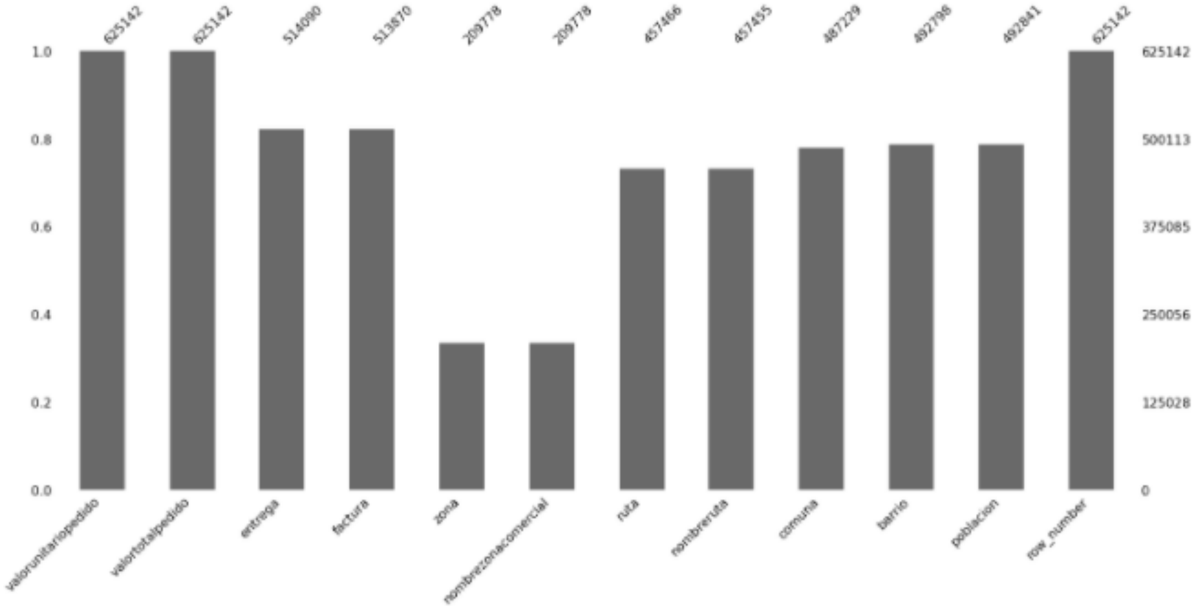


Table 2. List of variables that need to be analyzed

Variables	Proportion Missing Values
Entrega	17.76%
Factura	17.79%
Zona	66.44%
Nombre comercial zona	66.44%
Ruta	26.82%
Nombre Ruta	26.82%
Comuna	22.06%
Barrio	21.17%
Población	21.16%

Figure 6. Candidate fields to be filled using external sources



For these variables we needed more information regarding locations of main areas in the cities, so the next step in to request this information to build geofences (Polygons) for the main zones and use it to make aggregate forecasts by city level and give additional insights. This might be done in a posterior study by collecting geocoding information regarding the address of the deliveries (organized as columns for latitude and longitude) to generate heatmaps that identify geo-zones by tiers (High, Medium and low) based on the frequency of deliveries per week for each zone.

EXPLORATORY DATA ANALYSIS

In this section we present the analysis between variables and identification of trends and patterns needed to start extracting information that would help us determine the main factors that affect the demand of Teate's products. Here we present some graphs that we first considered decisive to visualize the possible relationships between variables to then study causality and select the most suitable model to predict demand. For that, the outliers, the distributions of the data grouped according to the variables and the behavior of the latter according to the main variables were visualized.

Outliers

One of the key variables that would be important to forecast is order values. In general, we found this data has a high variability between order values, ranging from 0 to 200,000 COP and a median around 70,000 COP (Figure 7). Out of this range we observe there is a high number of orders order with zero values, which might be due to missed or refunded deliveries, in which case they would serve as indicators of profit and loss, as well a as quality regarding delivery monitoring. Thus, such an indicator would be key in striving to stabilize deliveries as an initial strategy to absorb a greater market share.

Figure 7 also evidence an important number of out-of-range orders with values higher than COP200,000 per order. To investigate if atypical values can be explained as an effect of population density on delivery values, we groped the data by sectors (Figure 8). This plot suggests that such increased values in deliveries are evident in largest cities and popular sectors (zones 208, 214, 205 and 202).

Figure 7. Boxplots for total order values per month

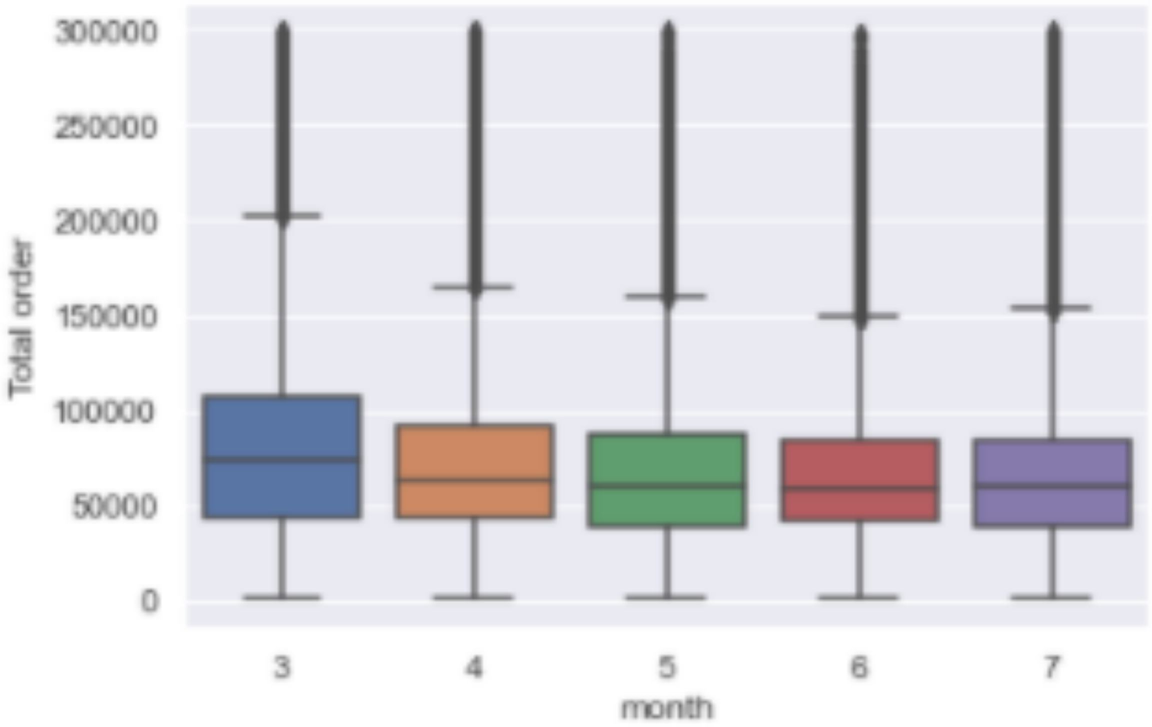
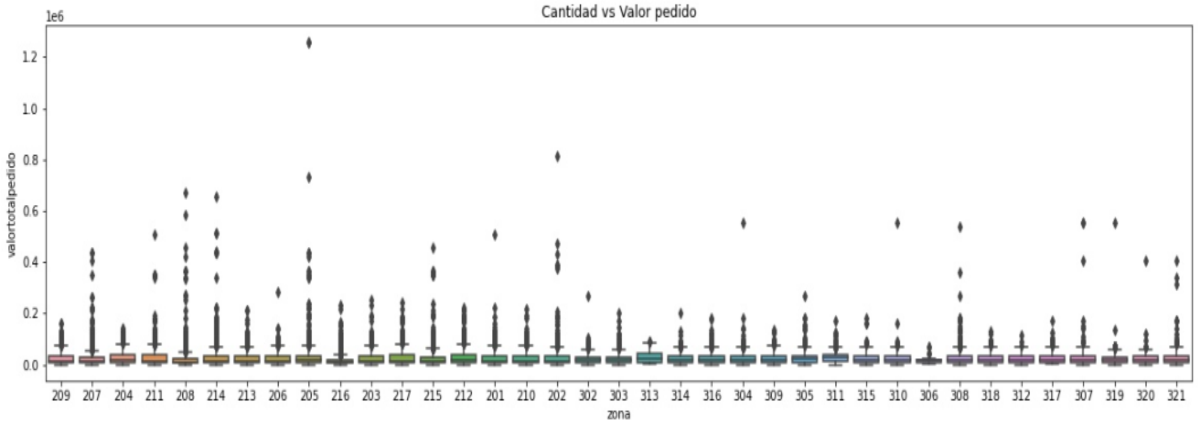
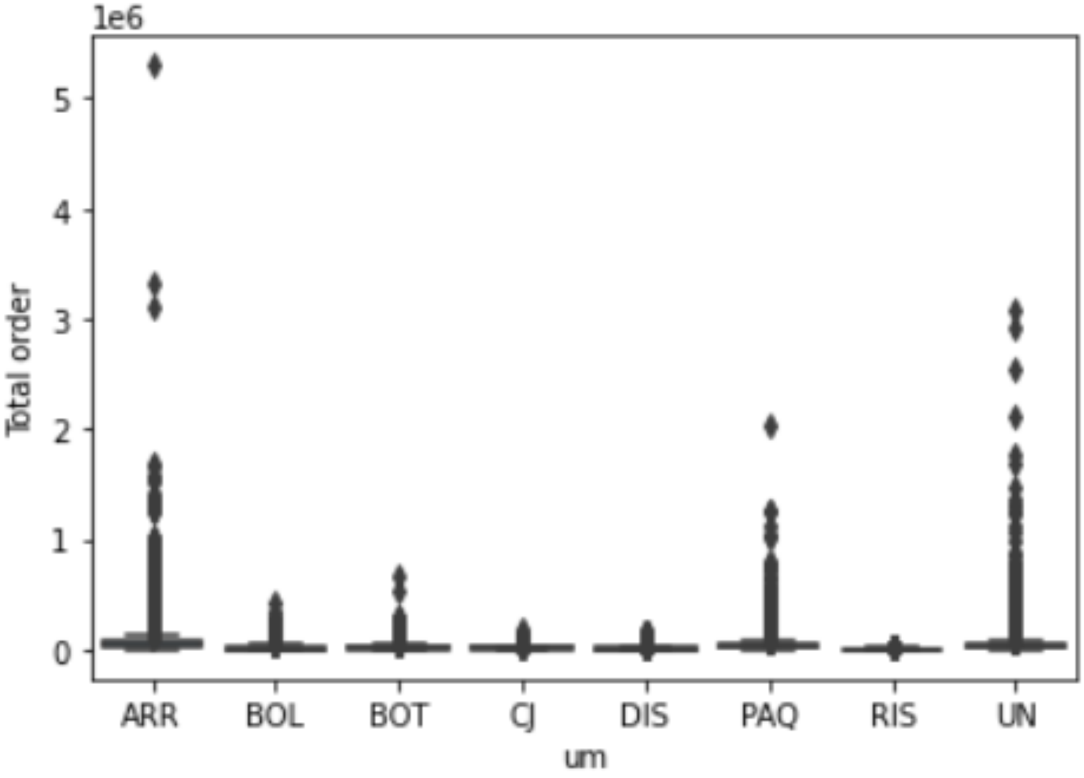


Figure 8. Boxplots for total order values per zone



Grouping by order to visualize the behavior of the units of measure of delivered products, several atypical data are evident, in particular for the ARR (arroba) unit measure (Figure 9). This can be due to the fact that Teatés' best-selling products are ordered in arrobas (See top 10 selling products section). However, it is evident that the data shows great dispersion and as with other variables, we cannot identify a clear trend in this one either.

Figure 9. Boxplots for total order values per unit of measure



Top 10 selling products, vendors and regions

The main products and players that we found to have a key role in the product marketing chain are shown in figure 10. It is evident that the most demanded products are mostly basic needs, that is, foods such as salt, coffee and rice. In particular, we observe that the products with the highest turnover are 25-unit Roa rice and 25-unit Florhuila rice, representing about 400,000 million pesos of sales. From this first observation we expect these products to have priority over others in the forecast. Moreover, to highlight their importance, we found that they have the highest turnover associated with rice brands, which supports what we found in our previous analysis for sales by product and by unit of measure (figs. 9 and 10). In line with this, the main suppliers are Roa S.A., Arroz Florhuila S.A, and Procon S.A., which account for just over 50% of purchases (Figure 11), distributed mainly in Santiago de Cali (Figure 12).

As a result of this first analysis, a strong dependence on Teaté for the sale of rice is evident, which translates into a strong commercial relationship with two of the largest rice companies operating in Colombia that are Organización ROA and Florhuila. In terms of the projection this can facilitate the analysis, since according to the data it seems that they follow the historical trend.

Figure 10. Top 10 selling products

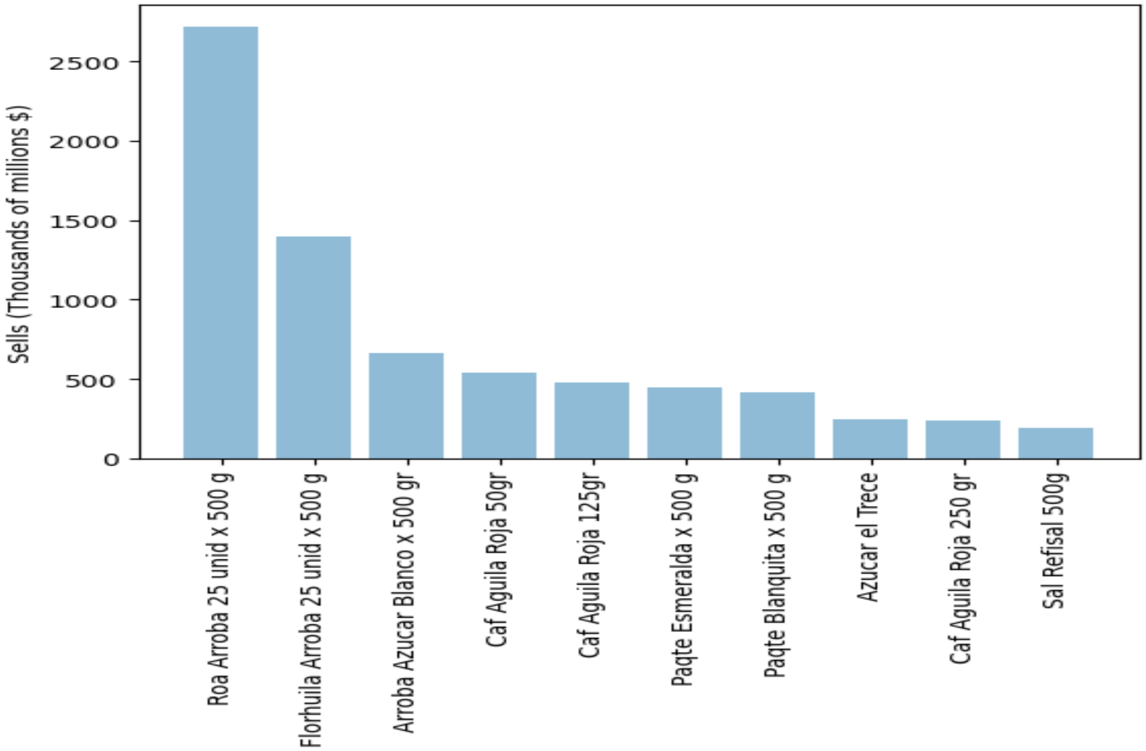


Figure 11. Top 10 selling manufacturers

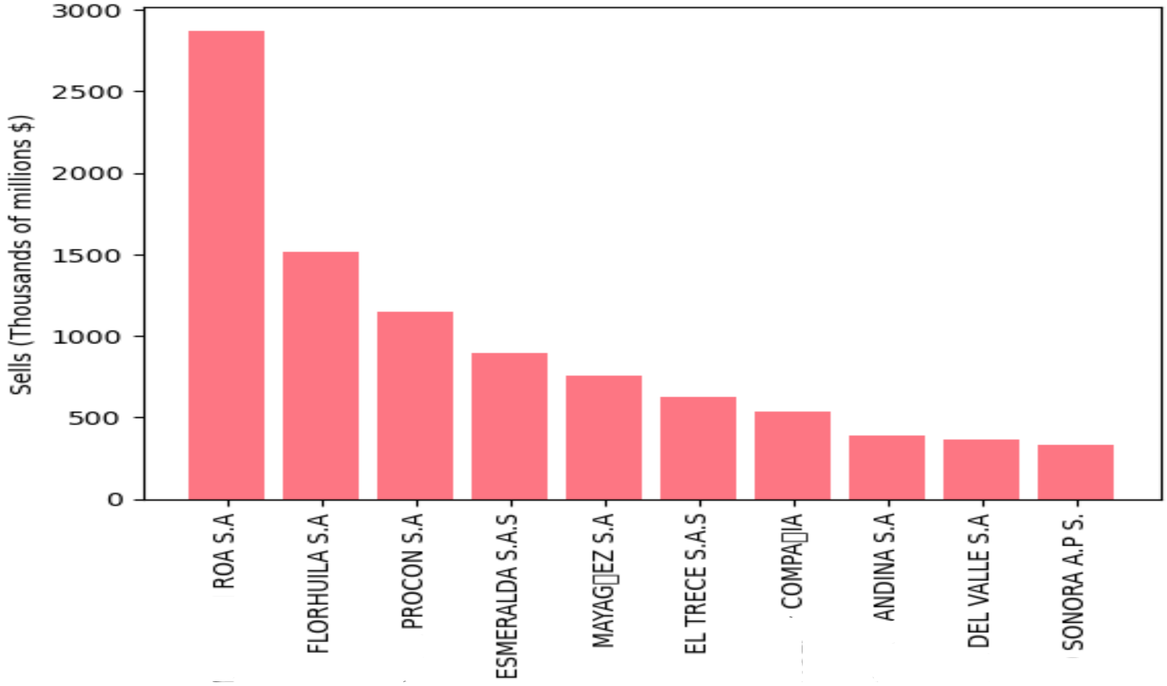
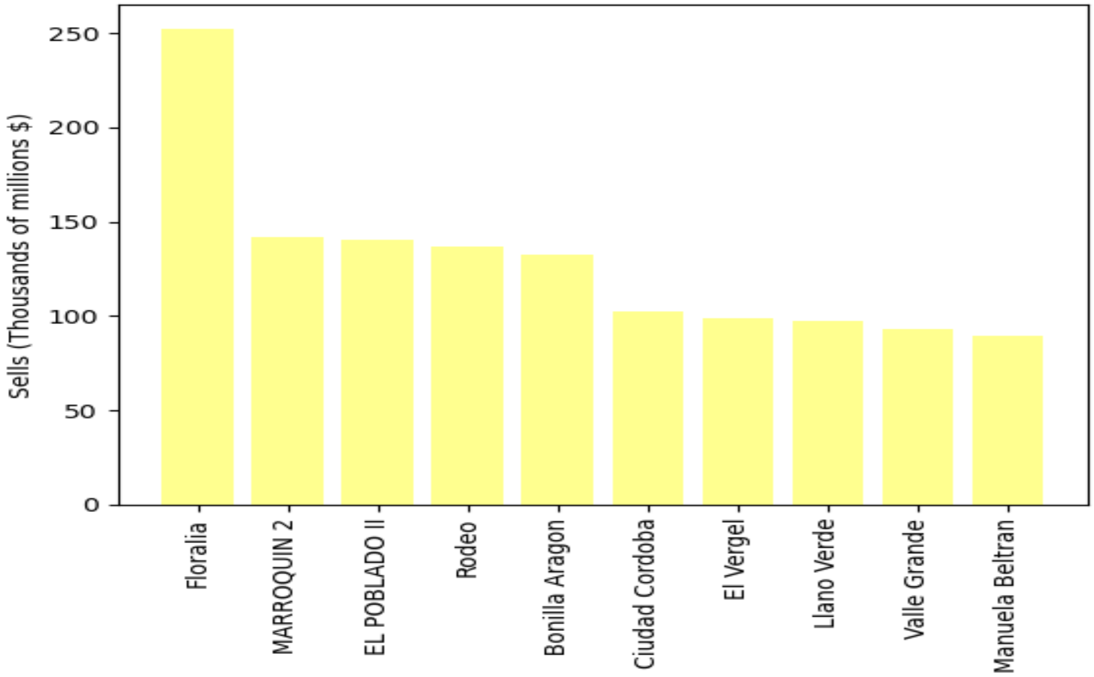


Figure 12. Top 10 selling neighborhoods



On the other hand, the concentration does not seem to be related only to the star products and their respective suppliers, there is also a strong relationship between sales and region where the company is located (i.e. most sales come from Santiago de Cali). However, given the structural conditions caused by COVID 19 in 2020, higher sales are evident in other places close to Cali and Medellín, though we also observe a growing factor of expansion of Teaté in other cities and municipalities of the country.

Searching for trends and patterns

One of the most important variables to understand the demand by product and by manufacturer is sales, however, since it is such an obvious relationship, it is necessary to carefully analyze the historical data, in which trends and patterns can be found as evidence in this section.

In figure 13, weekly sales between 200-300 million are evidenced and a strong peak is in the month of February and March. This peak evidence an important positive effect of the quarantine during the start of the pandemic; an expected response of the demand (as people in general reacted with increasing orders and purchases in response to the emergency) that tended to stabilize as months went by.

To explore in deep the demand behavior we plotted the time-series for purchase orders from Jan 19 to Sep 2020 (Figure 14). Here we can see that during 2019, the number of orders increased from 500 to 2000, growing in a random walk pattern with a growth rate around 150 per month. In 2020 however, the demand experiments a sudden increase during the months of March and April placing its average orders between 3500-4000. This result, that also confirmed the COVID effect that we observed for total sales values (Figure 15), is noteworthy since in times of pandemics we definitely detect an increase in sales that would propel the acceleration of the digital transformation of small businesses in the cities where Teaté operates.

Key drivers of sales and orders

Following what was presented above about the trend of historical sales, we continue with the analysis by disaggregating sales in greater detail. In figure 10 we present the main indicators for product sales' behavior to investigate how the different metrics would impact the demand. Here we can see that total orders by type of measure divided by year have a growing and stable behavior; however, in 2020, a peak of 3000 orders has been presented between week 10 and 15, which coincides with the start of the pandemic and the shortage of products due to the high demand for basic products. It is also noteworthy in 2020, product presentations in arrobas have the highest demand compared to other product lines, despite the fact that in 2019 they remained in a similar range between 900 and 1200 orders.

Figure 13. Time series for total sales values between 2019 and 2020

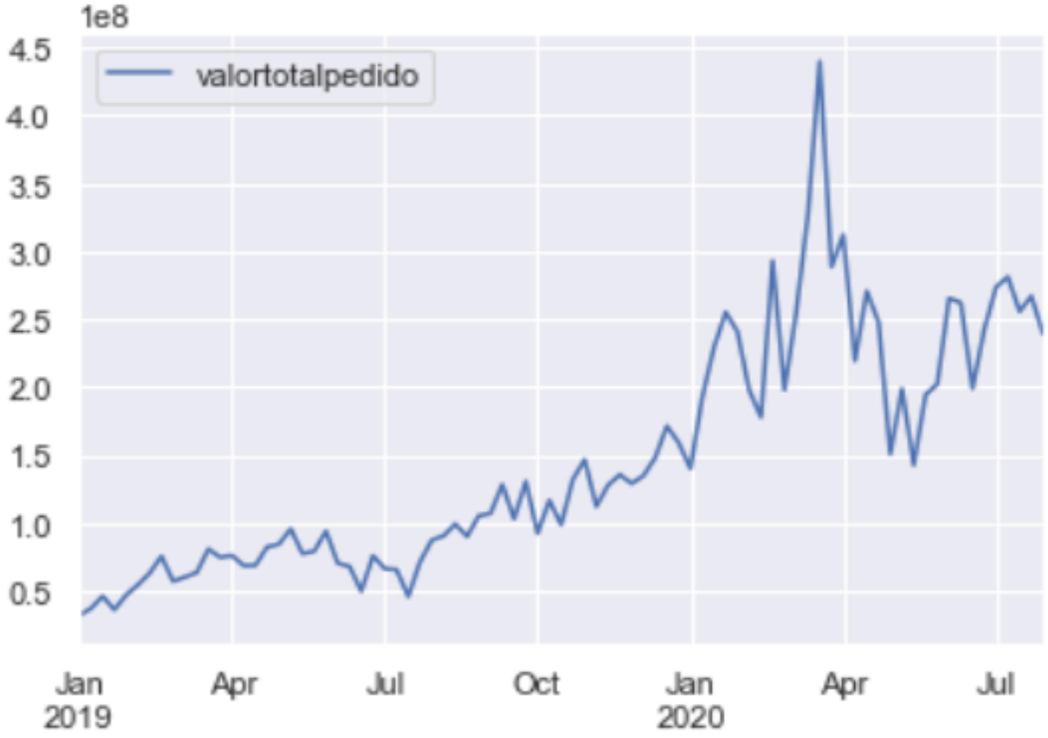
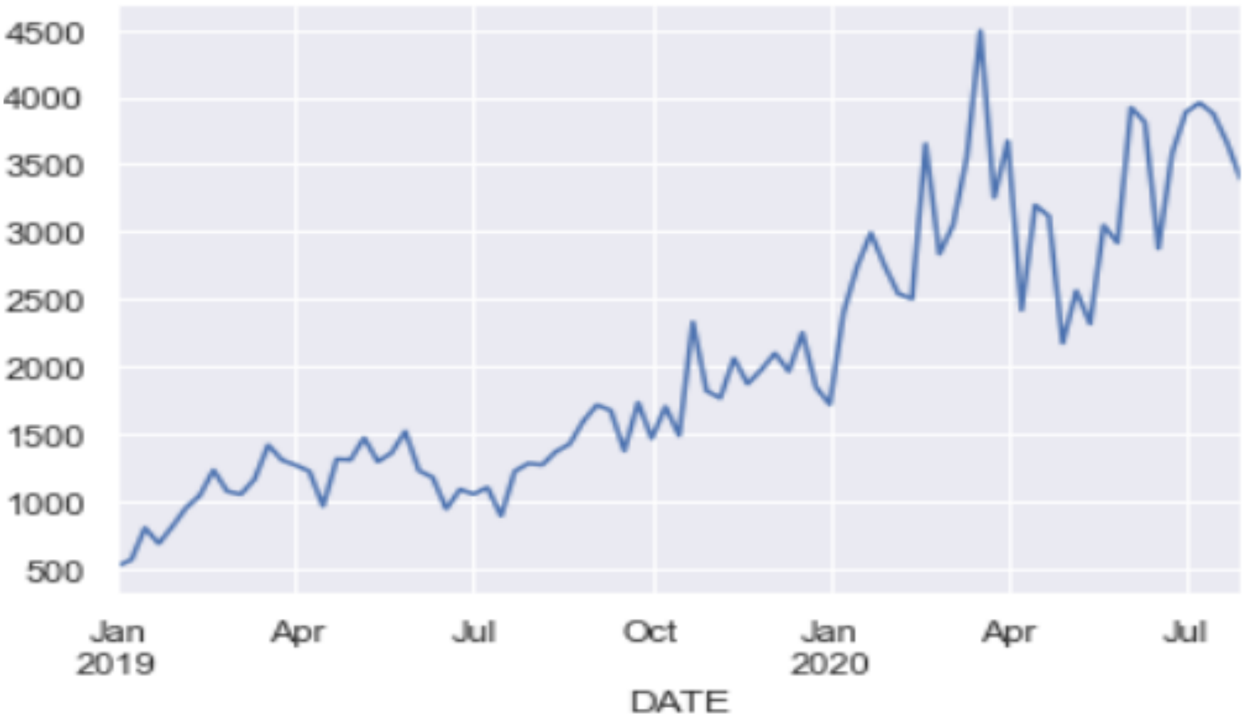


Figure 14. Time series for demand between 2019 and 2020



In addition, the total orders by unit measure by year shows a growing and stable behavior. However, in 2020, a peak of more than 3000 is evident between week 10 and 15, which coincides with the start of the pandemic and the shortage of products due to a high demand for basic products. It is also noteworthy in 2020 that product presentations in arrobas have the highest demand compared to other product lines, despite the fact that in 2019 they remained in a similar range between 900 and 1200 orders.

Orders and sales in different regions

Figure 16 shows the number of orders per week in each city or municipality in 2019 and 2020. From these time series we can see that Santiago de Cali is the main sales place for Teaté, followed by Medellín. Sales and orders in other cities can be considered as marginal and small-growth movements in comparison with Cali and the tendency makes remarkable in 2020 as Cali continues to be the largest market, though a considerable growth is observed for Medellín. Another element that is relevant in this analysis is the lack of information that is notorious in 2020 from week 1 to 10. This data is essential to understand the behavior of demand and to carry out the projection. In terms of total sales, the behavior is similar (Figure 17). There is a noticeable growth in the sale levels of orders, reaching values of 100,000,000 at the end of the year. In 2020, with the information available, the total value of the orders went from 150.000.000 to 200.000.000.

Figure 15. Total orders per week in 2019 and 2020

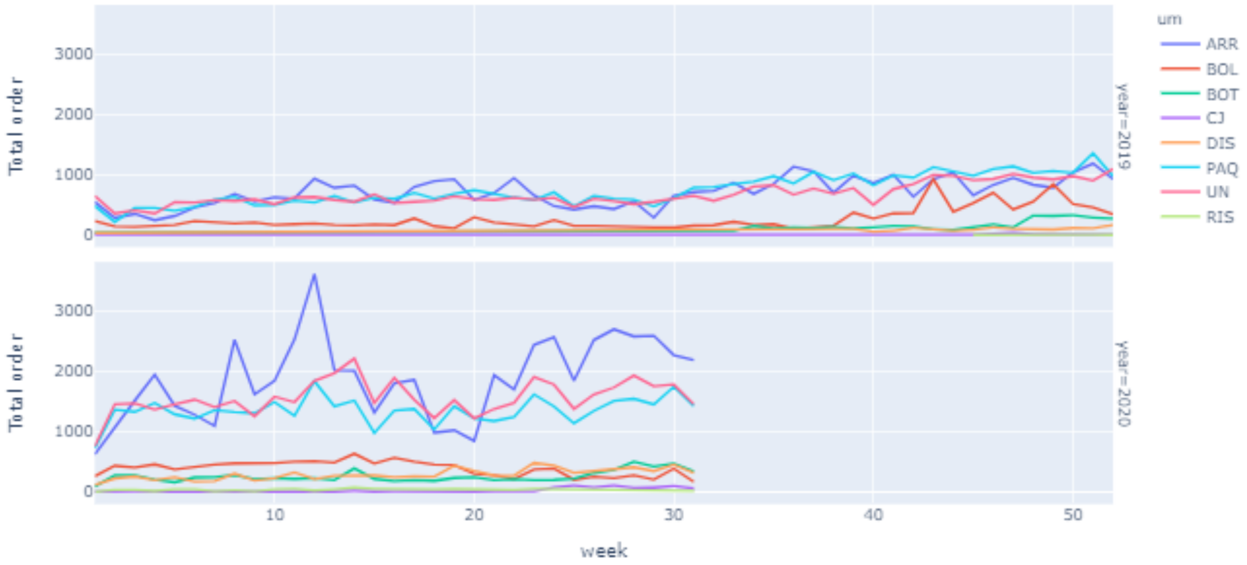


Figure 16. Number of total orders per city in 2019 and 2020

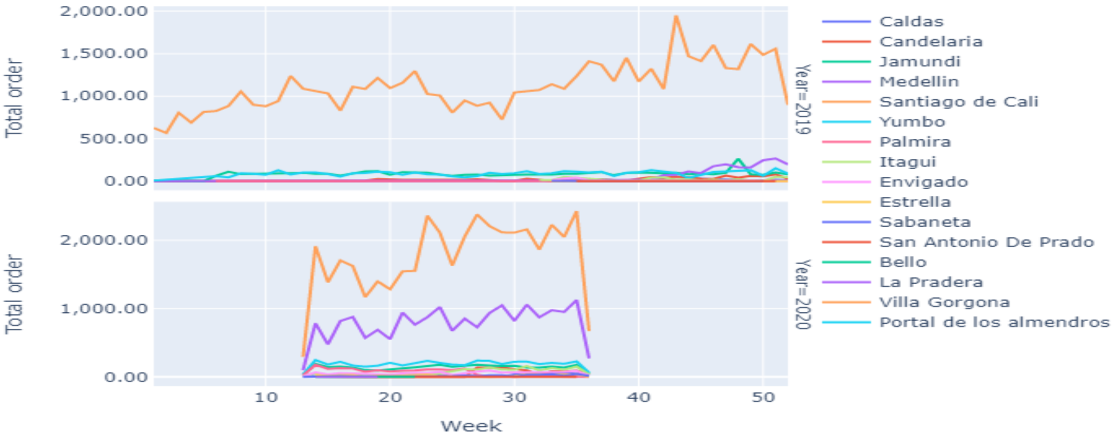
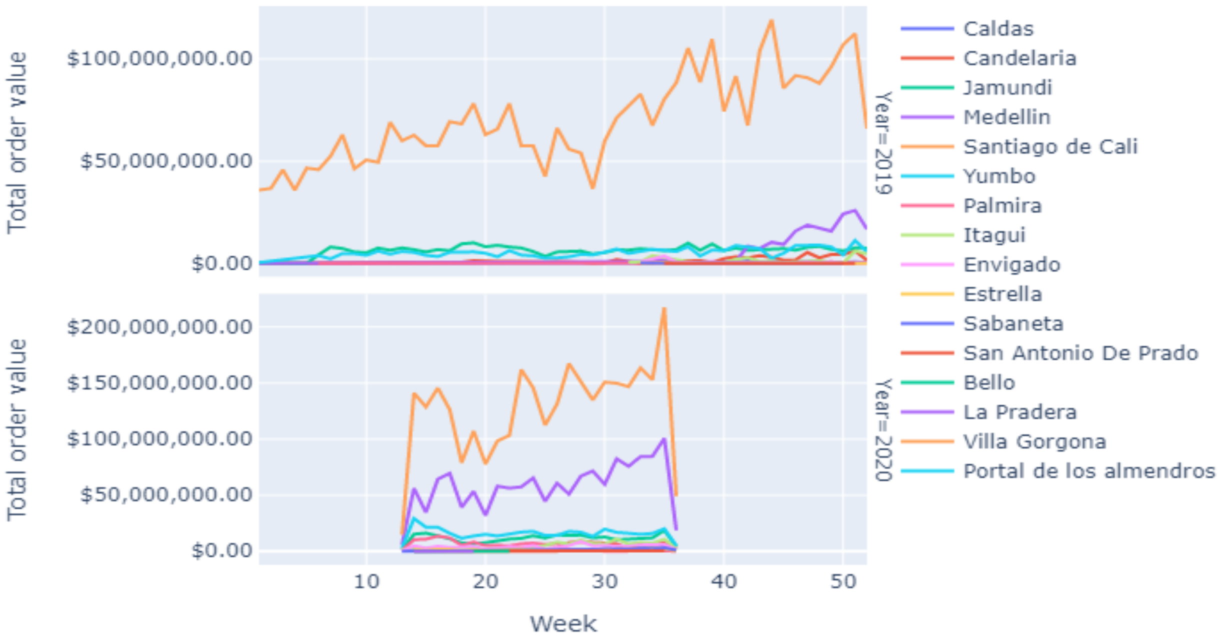


Figure 17. Total sales per city in 2019 and 2020



METHODS

In this section we present the methodological approach we used in this project to start developing a tool to help Teaté to watch over the demand based on the time series data we analyzed above. To do that, we first investigated if the past of the series could explain the number of orders in the present to determine whether a time series model would result in an appropriate methodological approach for this project. After meeting the conditions and assumptions for a time series analysis, we used a combination of three time series models in order to predict the demand from the data sales. A description of the complete process including the projection and selection of the model that showed the best performance is also presented.

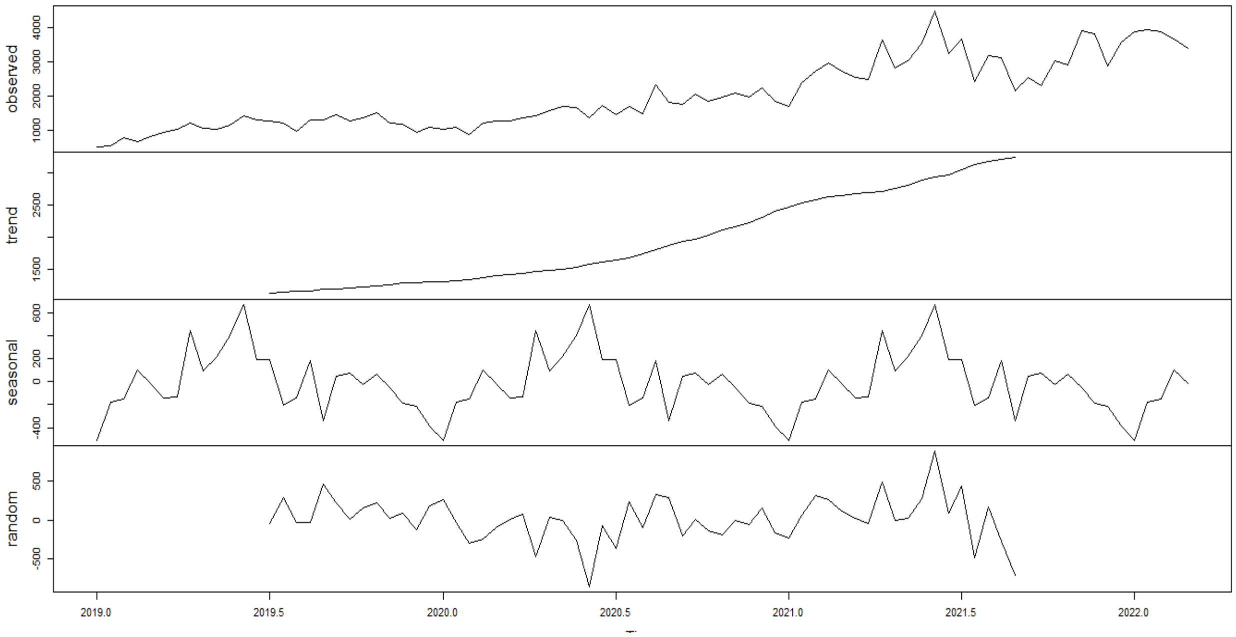
Structure of the time series

To investigate the structure of the time series, we present its decomposition in figure 18. Here we observe a semestral trend for the number of orders that would be associated to the well-known six-month cycles that the economy experiments in Colombia due to semestral wages bonifications. This effect might have a positive impact on general orders that would benefit micro-merchants and would contribute to the increased effect of sales observed in figure 10 from the end of 2019 to the beginning of 2020. On the other hand, the second component of increased trend we observe in this figure can be attributed to the increasing number of clients that Teaté is gaining and the market potential of the company, which have been strengthening in Medellin and Cali.

Time series analysis

First, it is important to note that the same as when building machine learning models, we should avoid multicollinear features to time series models as well. For that purpose, we first searched for optimum features or order of auto-regressive (AR) and moving averages (MA) series using both complete and partial autocorrelation functions (ACF and PACF, respectively).

Figure 18. Decomposition of additive time series



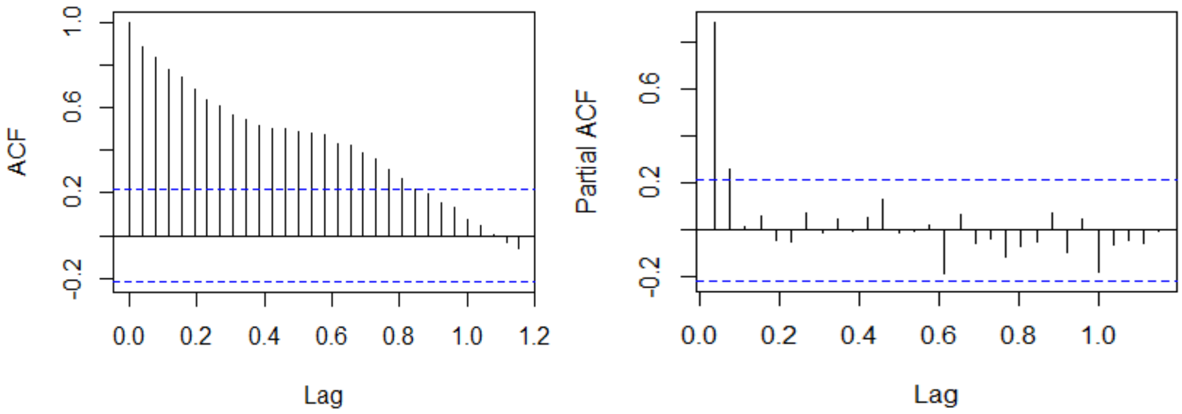
As the time series described above presented trend and seasonality (Figure 18), we used an ACF to consider these components while finding correlations, and a PACF to remove variations explained by earlier lags (Figure 19). Here, looking at the values of auto-correlation of the series with its lagged values along with the confidence band in the ACF plot, we can conclude that the present value of the series is related with its past values.

Finally, to remove variations explained by earlier lags so we get only the relevant features, we used a PACF that helped us find correlation of the remaining residuals with the next lag value, thus removing previous variations before we find the next correlation.

The models

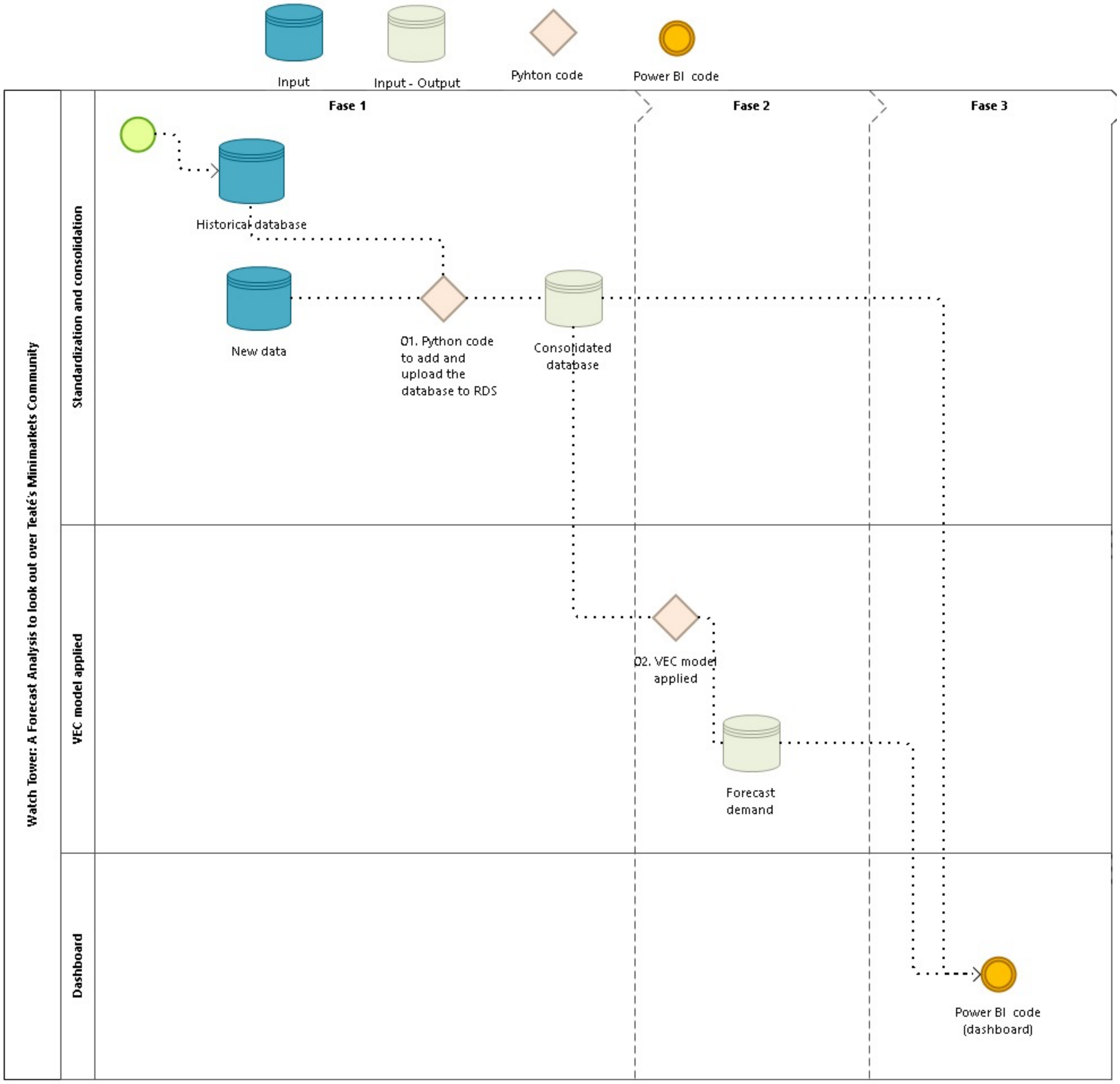
Given the nature of the problem and the structure of the data (exhibiting both a trend and a seasonal variation), we used a combination of three time series models capable of modelling this kind of time series in order to predict the demand from the data sales. For that we used the VEC, the ARIMA and the Holt-Winters exponential smoothing model (see Apendix II for description of the model). As in previous analysis, the information in both ACF and PACF graphs evidenced that by choosing a time series approach, after passing the unit root tests, it is appropriate to use of a time series approach to the problem of what information of the past can influence a forecast. To do that, each model was independently trained in order to forcast the transformed time series data.

Figure 19. ACF and PACF plots



In the first place, each one of the series of the input dataframe (value demanded by product or subgroup of products) was segmented into the training and test sets, using a partition defined by the number of periods to be projected. Each of the 3 models were trained with the series to make the projection of the value for the number of periods selected. This projection was contrasted with the values of the test set by estimating the mean square error of the projection of each model for each particular series.

Figure 20. Flujogram of the process


Input
Input - Output
Pyhton code
Power BI code
Watch Tower: A Forecast Analysis to look out over Teaté's Minimarkets Community
Standardization and consolidation
VEC model applied
Dashboard
Fase 1
Fase 2
Fase 3
Historical database
New data
01. Python code to add and upload the database to RDS
Consolidated database
02. VEC model applied
Forecast demand
Power BI code (dashboard)

Second, once the precision of the projection generated by each model was estimated, we chose the one that showed the best performance for the test set. The same process was carried out with each one of the series of the input dataframe, in such a way that we obtained the best model to carry out the future projection of the required periods for each series.

Finally, the algorithm uses the best model to generate an output dataframe with the same number of columns as the input dataframe and with a number of rows corresponding to the number of projected periods. In this way, we ensure that each time a projection is made, the algorithm selects the model that best captures the intrinsic behavior of the input series. A diagram of the entire process is described in figure 20.

RESULTS

We used the vector autoregressive (VAR) model as a general framework to describe the dynamic interrelationship among stationary variables. To do this we first determined whether the levels of the data were stationary and take the first differences of the series. We found the log-levels of the time series to be stationary, so we could use a family of VAR/VEC models to predict the demand.

While looking for a VEC model to emulate a vector autoregressive technique, the challenge was the generation of categories homogeneous enough among themselves, to maintain a constant over time. Before the assumptions were tested, the description of the time series between the different presentations were found similar and some kind of interdependence between products was expected (Figure 21). This interdependence can be explained with the fact that the products we were analyzing are basic necessities that can be either complementary or substitute, with a demand that remains almost constant. To give just an example, products such as milk, toilet paper, grains and basic necessities are always part of the purchases of an average family.

It is also important to note that even when a VEC model is a powerful tool to make predictions, it has some limitations. For example, the model it is not effective when there are is no clear correlation between different time series, as in the cases where the correlation is low. However, alternative models such as SARIMA and Holt-Winter resulted useful to describe some time series in this project. To choose the best model for each time series, we used the RSME as an indicator (Table 2).

Figure 21. Correlation matrix between most important substitute products

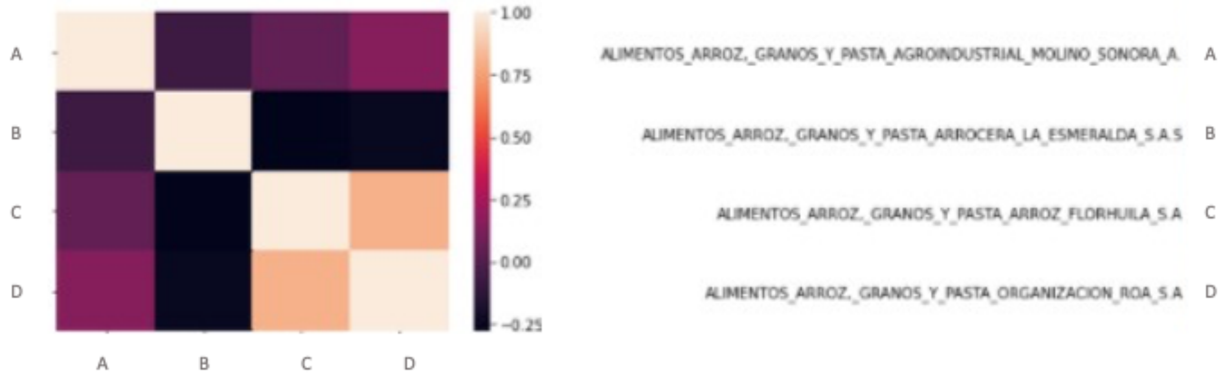
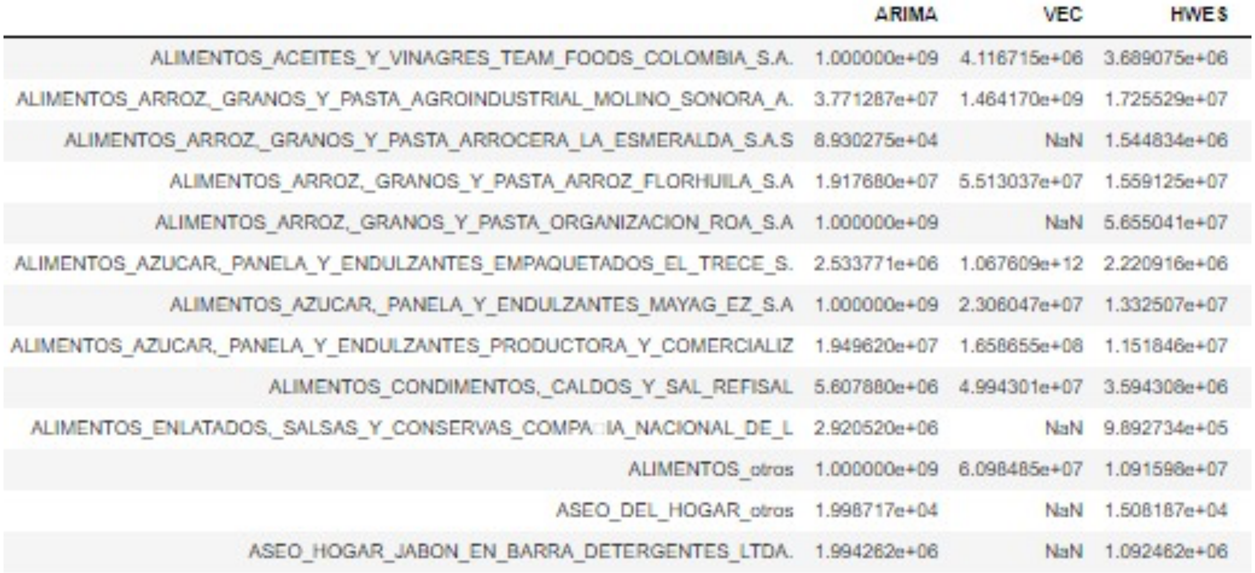


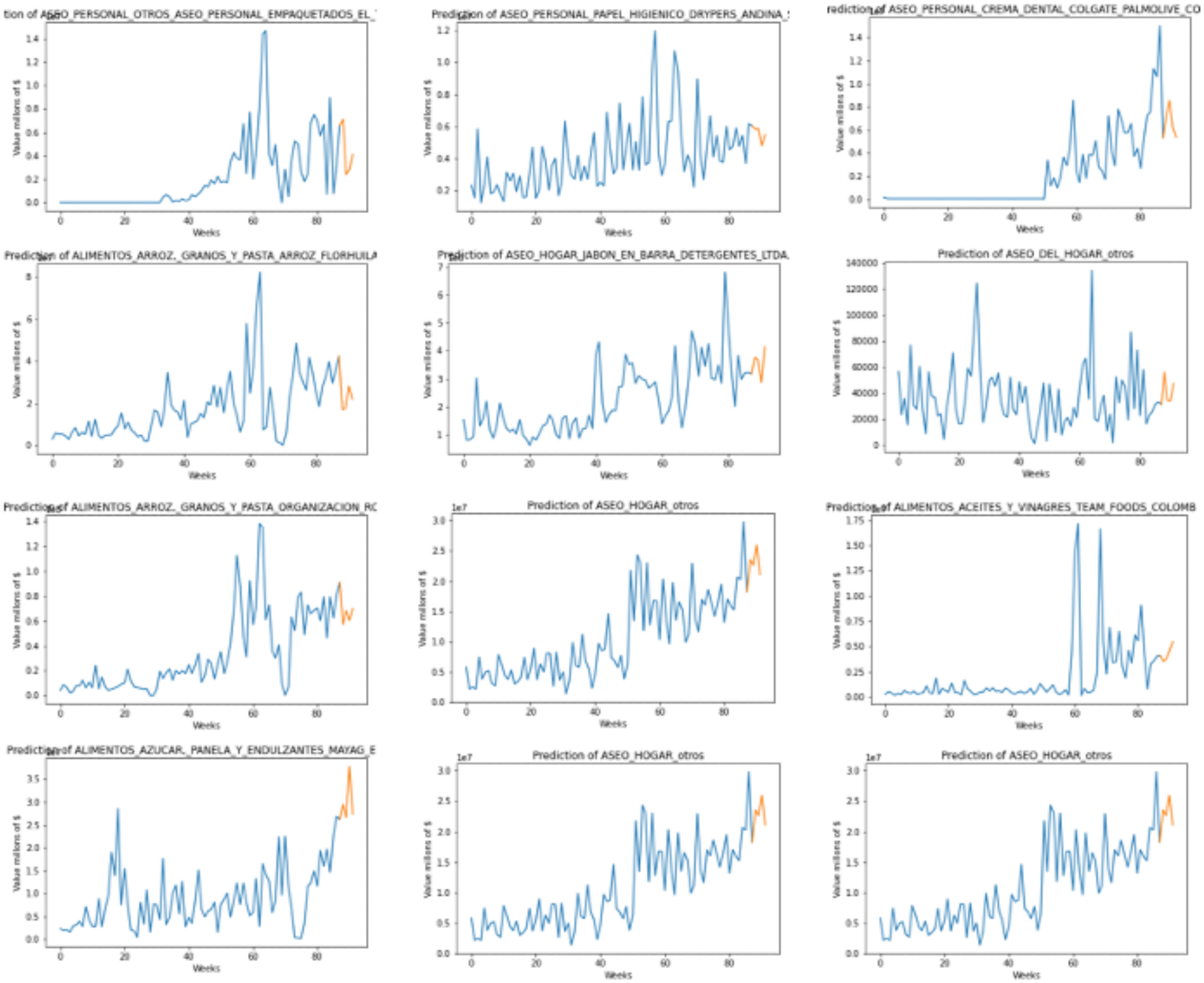
Table 2. Calculated error for each method using the RSME



	ARIMA	VEC	HWES
ALIMENTOS_ACEITES_Y_VINAGRES_TEAM_FOODS_COLOMBIA_S.A.	1.000000e+09	4.116715e+06	3.689075e+06
ALIMENTOS_ARROZ,_GRANOS_Y_PASTA_AGROINDUSTRIAL_MOLINO_SONORA_A.	3.771287e+07	1.464170e+09	1.725529e+07
ALIMENTOS_ARROZ,_GRANOS_Y_PASTA_ARROCERA_LA_ESMERALDA_S.A.S	8.930275e+04	NaN	1.544834e+06
ALIMENTOS_ARROZ,_GRANOS_Y_PASTA_ARROZ_FLORHUILA_S.A	1.917680e+07	5.513037e+07	1.559125e+07
ALIMENTOS_ARROZ,_GRANOS_Y_PASTA_ORGANIZACION_ROA_S.A	1.000000e+09	NaN	5.655041e+07
ALIMENTOS_AZUCAR,_PANELA_Y_ENDULZANTES_EMPAQUETADOS_EL_TRECE_S.	2.533771e+06	1.067609e+12	2.220916e+06
ALIMENTOS_AZUCAR,_PANELA_Y_ENDULZANTES_MAYAG_EZ_S.A	1.000000e+09	2.306047e+07	1.332507e+07
ALIMENTOS_AZUCAR,_PANELA_Y_ENDULZANTES_PRODUCTORA_Y_COMERCIALIZ	1.949620e+07	1.658655e+08	1.151846e+07
ALIMENTOS_CONDIMENTOS,_CALDOS_Y_SAL_REFISAL	5.607880e+06	4.994301e+07	3.594308e+06
ALIMENTOS_ENLATADOS,_SALSAS_Y_CONSERVAS_COMPA IA_NACIONAL_DE_L	2.920520e+06	NaN	9.892734e+05
ALIMENTOS_otros	1.000000e+09	6.098485e+07	1.091598e+07
ASEO_DEL_HOGAR_otros	1.998717e+04	NaN	1.508187e+04
ASEO_HOGAR_JABON_EN_BARRA_DETERGENTES_LTDA.	1.994262e+06	NaN	1.092462e+06

The projections obtained from this combination of methods resulted in accordance to what was expected for the most important products. In figure 22 it can be seen that the historical series (blue lines) and the forecasts obtained with the models (orange lines) are in agreement and the projections are relatively stable despite the variations. As we expected, the projections follow a trend and do not present exaggerated peaks.

Figure 22. Projections for some of the major products and distributors



Frontend MockUP

To simplify the transference process with the final user of the platform we designed the frontend using Power BI, which is the business analytics service that Teaté has. We created the MockUP using the app.moqups.com tool which is very useful for making drafts for an application. With the free tier, we were able to make a formal design of what the application would be in the future that allows the Teaté company to estimate demand per product per week (See Watch Tower).

The dashboard is composed of 4 sections; index, reporting, forecast, and validation. All the sections consist on informative panels that presents the project and its members and the results of the forecast for business intelligence as well. Here it collects the information on reporting issues that provides specific historical information. A forecasts tab shows the projections (at 3 months?) at different dimensions, and the last back-testing tab is to visualize the difference between real and expected sellings during a given time interval.

Index

The Index is a brief introduction to our project. In this section we will analyze in depth the company we are helping and its business context, the definition of the problem we are facing, the description of what the App does and its functionalities and finally a brief description of the members of the group in the section called developers.

Reporting

This section is for business intelligence (i.e. to collect information on the reporting issues). Here Teaté will find information about some sales and product turnover indicators. In this part we have collected specific historical information that is useful for the company, such as the sales of a specific product in a specific area, or how many units were delivered in a particular time. This reporting space is specially designed to provide a much more detailed visualization of the data behavior. In this section the idea is to show the funnel of the orders, how the total of the orders are made, and follow the process until having a successful delivery.

On the other hand, is a visualization of the orders closed week by week for each of the products. It is important to clarify that in this report and in the following ones, the user can perform filters to observe specific behavior of, for example, products, skus, cities, manufacturers, minimarkets and metrics. Another important visualization that we can find in the report is a heat map that specifies where the largest number of orders is concentrated. Finally, we will have a vision of the best-selling products, top SKUs and top manufacturers.

Forecast

Forecasts tab allows Teaté to see the projections (at 6 months?) for different dimensions (example: by product, brand or area).This third point is the main part of the algorithm (forecast), and in this visualization the idea is to show the results of the demand prediction for each of the products in the different cities for each week in the next 6 months. This part of the application will receive as input what the model calculates for the demand and will display it in a more acceptable way. Again we include the filters so that the user can select the parameters they want to see at their convenience and in addition to this, we include a button to download the information in table form.

Validation

The final section is a back-testing page that allows to visualize the difference between expected sales and what was actually sold in a given time interval. With this information the company will be able to determine the error of the expected demand and supply. This segment of the table is also for the validation of the forecast results. The idea of this visualization is to have feedback and keep track of how the model is behaving with respect to reality and what error the model has. The first thing to highlight is that we have a foot graph which shows the percentage of success in the model taking into account whether the demand was close enough to the prediction taking into account a level of error. After that, we have a graph that will show us with respect to time the error of the model when implementing it. Finally, a table that allows us to observe in which products we have had a deficit with respect to real demand and in which we have overestimated demand.

Finally, the platform generates a report that contains a header with the identification of the company, the team, and the area of the company to which it belongs. In addition, the report contains all the sales information filtered by the most relevant variables and synthesized by means of some informative graphics that contain the most important aspects of sales and their trends. The dashboard and information project are available in googlesites portal https://sites.google.com/view/team77-watchtower/main?authuser=0. The portal contains embedded the powerBI dashboard across an iframe.

DISCUSSION

Since this project was focused on estimating the demand for each product and by manufacturer, the methodological approach of a range of models related the time series resulted appropriate due to the nature of the problem and the structure of the data as we noted in the exploratory data analysis section.

However, for a more complete analysis we need more information about the main areas in the cities to continue exploring the effect of different locations on the demand of products. For that we need to request or build geofences (Polygons) for the main zones so that we can aggregate a forecast under city level that could generate additional insights breakdowns. We can also add geocoding information regarding the address of the deliveries so that we can generate heatmaps by means of including columns for latitude and longitude.

In addition, we would like to design and implement a hierarchy of the product lines that are being analyzed. For example, it would be interesting to request the barcode of each SKU so that we can build the hierarchy internally. In this way we would be able to forecast beers instead of just “Corona” or “Club Colombia” brands.

Next steps

The large number of variables provided a broader picture to propose demand forecasting tools and address the main problem. However, there are many ideas that still can be implemented un a second part of this project. In particular, additional variables that identifies clients by tiers (1 ,2,3) based on the frequency by which each client makes orders would be included so we can analyze the demand behavior in relation on clients-classification basis (i.e. tier 1 are the 33% most frequent clients while tier 3 are the 30% least frequent).

We can also add a column that identifies geo-zones by tiers (High, Medium and low) based on the frequency of deliveries per week for each zone (i.e. "high" represents a geo-zone with a high amount of deliveries per week). Along the same lines, heat maps are proposed to geographically illustrate the demand by zones according to the classification shared in the data set. Subsequently, it is proposed to cluster the most requested products, which can be seen by means of a scatter plot: In principle these graphs would serve to make a deeper approach to the data shared by Teaté, however, as the project progresses, they can emerge other equally valid alternatives to visualize the data in a concrete and forceful way.

REFERENCES

1. Teaté. (March 5, 2020) Dinero.com Newsletter
2. Tiendas de barrio (October 2, 2020) Dinero.com Newsletter
3. Tiendas de barrio y cambio social (October 19 2020) Teaté Digital
4. La plataforma caleña Teaté llegará a las ciudades de Bogotá y Medellín (Mayo 18 2019) La República from https://www.larepublica.co/internet-economy/la-plataforma-calena-teate-llegara-a-bogota-y-medellin-2863461
5. Euromonitor International, APR 2020, Traditional Grocery Retailers in Colombia COUNTRY REPORT from www-portal-euromonitor-com reviewed on 2020- oct 03 at 10:30 pm
6. ¿Transformación digital para que las tiendas de barrio no se marchiten? (Septiembre 23, 2019) El Espectador
7. Creating the traditional channel , 2019, from https://teate.co/wp-content/uploads/2019/02/Nuestro-Balance-2018.pdf on 2020 - oct -03

APPENDIX I

The VAR/VEC model

The vector autoregressive model is a framework used to describe the dynamic interrelationship among stationary variables. The model involves multiple independent variables and has more than one equation. Each equation uses the lags of all the variables as explanatory variables to predict the most likely trend.

Engle and Granger combined cointegration and error correction models to establish the trace error correction model. As long as there is a cointegration relationship between variables, the error correction model can be derived from the autoregressive distributed lag model. Each equation in the VAR model is thus an autoregressive distributed lag model so the VEC model can be considered as a VAR model with cointegration constraints. Because there is a cointegration relationship in the VEC model, when there is a large range of short-term dynamic fluctuations, VEC expressions can restrict long-term behavior of the endogenous variables so they converge to their cointegration relationship.

Calculation steps

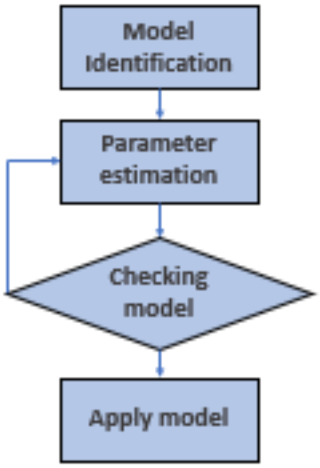
The method needs at least 50 observations as a requirement and it consists on 4 steps of identification, fitting, and checking for autoregressive and moving average time series models (Figure 3.1). The identification step is a filtering or whitening process that transforms the data in order to help identify which lags of x predict y, so that patterns for intervention effects can be identified and interpreted. For that, the best estimate of the likelihood of the models is selected using a set of information criterium such as the Hannan Quinn (HQ), the Bayesian (BIC), and the Akaike information criterion (AIC):

$$HQ(m) = ln|\sum_u(m)| + \frac{2}{T}mk^2$$

$$BIC(m) = ln|\sum_u(m)| + \frac{lnT}{T}mk^2$$

$$AIC(m) = ln|\sum_u(m)| + \frac{2mk^2}{T}$$

Figure 3.1. Steps of the VAR/VEC model



Stage	Techniques
Stationary Seasonality	ACF PACF Unit root test
SARIMA VAR	Least Squares Maximum likelihood Bayesian methods
Residuals	Rao's test Doonik and Hansen test
Granger causality Impulse response	Cholesky decomposition

A third step consists on estimating a set of parameters that minimize the error while maximizing the probability of the results using a combination of methods such as the maximum likelihood (ML) that maximizes a likelihood function, so that under assumed statistical model the observed data is most probable. Another method is the least squares (LS) that minimizes the squared discrepancies between observed data and their expected values. In addition, Bayesian methods are used as decision rules that minimizes the posterior expected value of a loss function (equivalently, it maximizes the posterior expectation of a utility function).

Checking the model

$$u_t = D_t u_{t-1} + ... + D_h u_{t-h} + v_t$$

$$H_0 : D_1 = D_2 = ... + D_h = 0$$

$$H_1 : D_i \neq 0$$

No significance of cross-correlations

Normality in residuals: Doornik and Hansen normality test, it is based on the skewness and kurtosis of the VAR residuals.

$$H_0 : (u_{1t}, ..., u_{kt}) \sim N(0, \sum_u)$$

The mean of residuals is zero: the expected value of each residual serie should be zero

$$E(u_t) = 0$$

$$H_0 : \mu_i = 0$$

$$H_0 : \mu_i \neq 0$$

Apply the model

Impulse Response Function (IRF)

IRFs trace the effects of an innovation shock to one variable on the response of all variables in the system. In contrast, the forecast error variance decomposition (FEVD) provides information about the relative importance of each innovation in affecting all variables in the system.

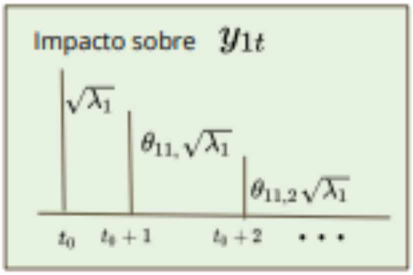
$$\mathbf{y_t} = \sum_{i=0}^{\infty} \Phi_j \omega\omega^{-1} \mathbf{u_{t-j}}$$

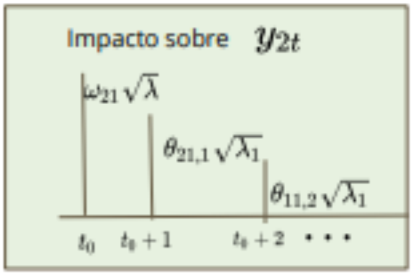
$$\mathbf{y_t} = \sum_{i=0}^{\infty} \theta_j \mathbf{w_{t-j}} \;\; con \; \Sigma_w = I_k$$

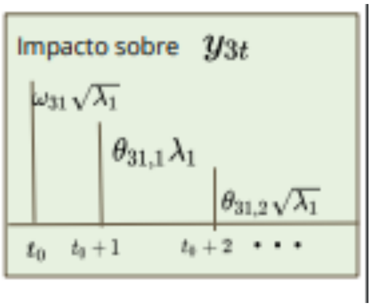
$$\begin{bmatrix} y_{1t} \\ y_{1t} \\ y_{3t} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \omega_1 & 1 & 0 \\ \omega_2 & \omega_3 & 1 \end{bmatrix} \begin{bmatrix} \omega_{1t} \\ \omega_{2t} \\ \omega_{3t} \end{bmatrix} + \begin{bmatrix} \theta_{11,1} & \theta_{12,1} & \theta_{13,1} \\ \theta_{21,1} & \theta_{22,1} & \theta_{23,1} \\ \theta_{31,1} & \theta_{32,1} & \theta_{33,1} \end{bmatrix} \begin{bmatrix} \omega_{1,t-1} \\ \omega_{2,t-1} \\ \omega_{3,t-1} \end{bmatrix} + \begin{bmatrix} \theta_{11,2} & \theta_{12,2} & \theta_{13,2} \\ \theta_{31,2} & \theta_{22,2} & \theta_{23,2} \\ \theta_{31,2} & \theta_{32,2} & \theta_{33,2} \end{bmatrix} \begin{bmatrix} \omega_{1,t-2} \\ \omega_{2,t-2} \\ \omega_{3,t-2} \end{bmatrix} + \cdots$$

$$\begin{bmatrix} y_{1t} \\ y_{1t} \\ y_{3t} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \omega_{21} & 1 & 0 \\ \omega_{31} & \omega_{32} & 1 \end{bmatrix} \begin{bmatrix} \omega_{1t} \\ \omega_{2t} \\ \omega_{3t} \end{bmatrix} + \begin{bmatrix} \theta_{11,1} & \theta_{12,1} & \theta_{13,1} \\ \theta_{21,1} & \theta_{22,1} & \theta_{23,1} \\ \theta_{31,1} & \theta_{32,1} & \theta_{33,1} \end{bmatrix} \begin{bmatrix} \omega_{1,t-1} \\ \omega_{2,t-1} \\ \omega_{3,t-1} \end{bmatrix} + \begin{bmatrix} \theta_{11,2} & \theta_{12,2} & \theta_{13,2} \\ \theta_{21,2} & \theta_{22,2} & \theta_{23,2} \\ \theta_{31,2} & \theta_{32,2} & \theta_{33,2} \end{bmatrix} \begin{bmatrix} \omega_{1,t-2} \\ \omega_{2,t-2} \\ \omega_{3,t-2} \end{bmatrix} + \cdots$$

$$\omega_{1t} = \begin{cases} \sqrt{\lambda_1}, & \text{si } t = t_0, \\ 0, & \text{si } t \neq t_0 \end{cases} \qquad \omega_{2t} = 0 \;\; y \;\; \omega_{3t} = 0$$

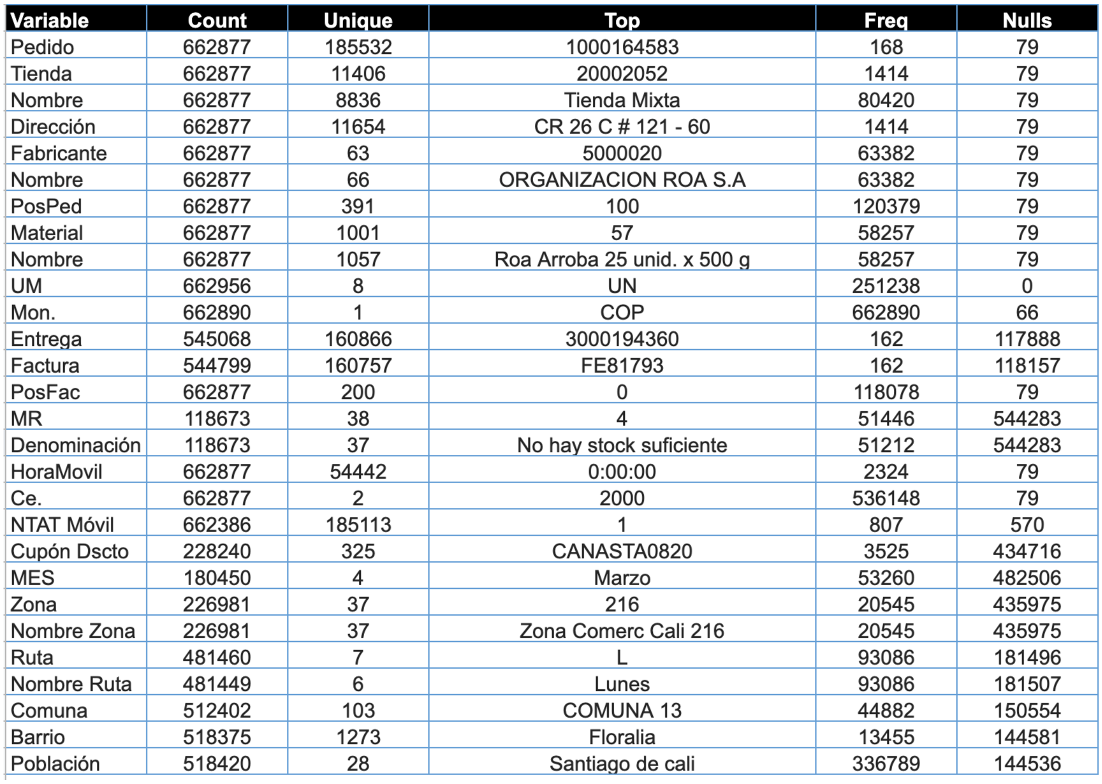



Impacto sobre y_{2t}
$\omega_{21}\sqrt{\lambda}$
$\theta_{21,1}\sqrt{\lambda_1}$
$\theta_{11,2}\sqrt{\lambda_1}$
t_0
t_0+1
t_0+2
• • •



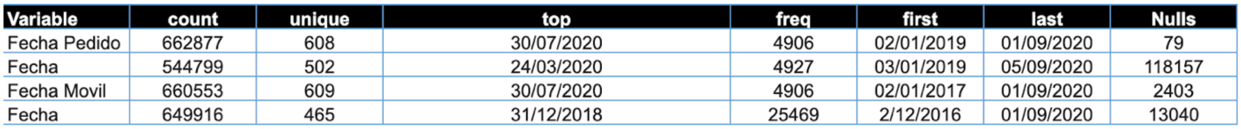
APENDIX II

Variables Details



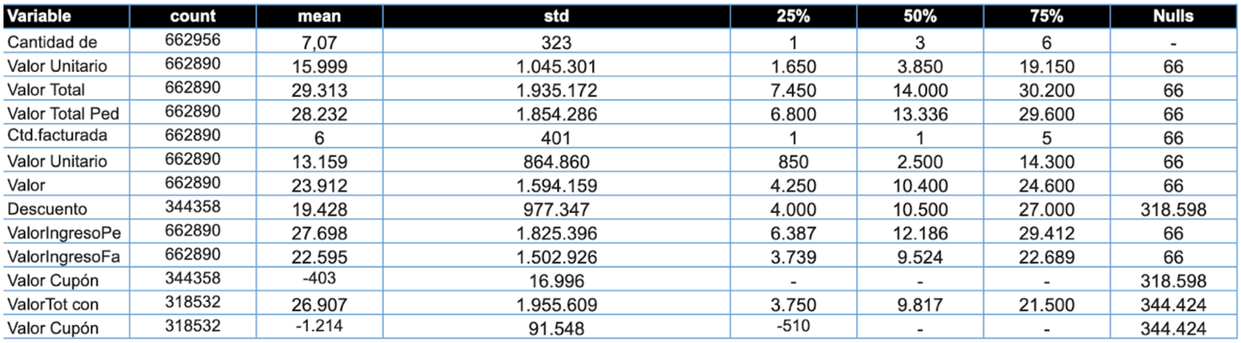
Variable	Count	Unique	Top	Freq	Nulls
Pedido	662877	185532	1000164583	168	79
Tienda	662877	11406	20002052	1414	79
Nombre	662877	8836	Tienda Mixta	80420	79
Dirección	662877	11654	CR 26 C # 121 - 60	1414	79
Fabricante	662877	63	5000020	63382	79
Nombre	662877	66	ORGANIZACION ROA S.A	63382	79
PosPed	662877	391	100	120379	79
Material	662877	1001	57	58257	79
Nombre	662877	1057	Roa Arroba 25 unid. x 500 g	58257	79
UM	662956	8	UN	251238	0
Mon.	662890	1	COP	662890	66
Entrega	545068	160866	3000194360	162	117888
Factura	544799	160757	FE81793	162	118157
PosFac	662877	200	0	118078	79
MR	118673	38	4	51446	544283
Denominación	118673	37	No hay stock suficiente	51212	544283
HoraMovil	662877	54442	0:00:00	2324	79
Ce.	662877	2	2000	536148	79
NTAT Móvil	662386	185113	1	807	570
Cupón Dscto	228240	325	CANASTA0820	3525	434716
MES	180450	4	Marzo	53260	482506
Zona	226981	37	216	20545	435975
Nombre Zona	226981	37	Zona Comerc Cali 216	20545	435975
Ruta	481460	7	L	93086	181496
Nombre Ruta	481449	6	Lunes	93086	181507
Comuna	512402	103	COMUNA 13	44882	150554
Barrio	518375	1273	Floralia	13455	144581
Población	518420	28	Santiago de cali	336789	144536

Datetime:



Variable	count	unique	top	freq	first	last	Nulls
Fecha Pedido	662877	608	30/07/2020	4906	02/01/2019	01/09/2020	79
Fecha	544799	502	24/03/2020	4927	03/01/2019	05/09/2020	118157
Fecha Movil	660553	609	30/07/2020	4906	02/01/2017	01/09/2020	2403
Fecha	649916	465	31/12/2018	25469	2/12/2016	01/09/2020	13040

NUMERIC:



Variable	count	mean	std	25%	50%	75%	Nulls
Cantidad de	662956	7,07	323	1	3	6	-
Valor Unitario	662890	15.999	1.045.301	1.650	3.850	19.150	66
Valor Total	662890	29.313	1.935.172	7.450	14.000	30.200	66
Valor Total Ped	662890	28.232	1.854.286	6.800	13.336	29.600	66
Ctd.facturada	662890	6	401	1	1	5	66
Valor Unitario	662890	13.159	864.860	850	2.500	14.300	66
Valor	662890	23.912	1.594.159	4.250	10.400	24.600	66
Descuento	344358	19.428	977.347	4.000	10.500	27.000	318.598
ValorIngresoPe	662890	27.698	1.825.396	6.387	12.186	29.412	66
ValorIngresoFa	662890	22.595	1.502.926	3.739	9.524	22.689	66
Valor Cupón	344358	-403	16.996	-	-	-	318.598
ValorTot con	318532	26.907	1.955.609	3.750	9.817	21.500	344.424
Valor Cupón	318532	-1.214	91.548	-510	-	-	344.424

APENDIX III

CONTRIBUTIONS

Interaction with the company
Data acquisition
Data cleaning - wrangling
Data analysis
Coding
Modeling
Results generation
Results analysis
Ingesta de datos
Manuscript
Video
Presentation