Memoria

Jakub Bagiński, Natalia C., Borja M., María S., Daniel R.

1. Alcance del proyecto

Sales, product evolution and logistics in XXX: Integrating industrial and data science perspectives

In this project, we conducted an integrated analysis of XXX's business activity by combining sales data with demographic, economic, and geographic information. First, we analyzed multi-year sales patterns to detect trends and fluctuations relevant to the company's commercial strategy. We also explored the structure of the Spanish floriculture sector and assessed how regional market characteristics relate to XXX's demand.

To complement the market analysis, we studied XXX's logistics network, including the location and service areas of its distribution centers. Using geographical coordinates and transport estimates, we evaluated the suitability of current warehouses and examined the potential location for an additional one.

Finally, we produced interactive visualisations and maps to support interpretation and to create an accessible, decision-oriented view of the data. Taken together, these analyses offer a multifaceted understanding of XXX's market, logistics environment, and demand drivers.

1.1. Objetivos del proyecto

- Analyze XXX's sales patterns over time to identify trends that may influence its commercial strategy and distribution network.
- Explain demand in relation to the specific market conditions of the floriculture sector.
- Describe XXX's current logistics situation, including its distribution centers, delivery areas and transport network.
- Determine the optimal location for a new warehouse that would minimize overall delivery time and cost if XXX were to internalize its distribution.
- Understand the structure of the floriculture sector in Spain's, identifying the geographical distribution of florists to assess its potential impact on XXX's sales.

1.2. Utilidad del estudio

This study is valuable because it integrates several dimensions—commercial performance, demographic dynamics, geographic structures, and logistics—into a unified analytical framework specifically tailored to XXX. Unlike broad market analyses or purely academic case studies, this project is grounded in real operational data provided directly by the

company. This gives the study practical relevance: its findings can support decision-making related to market expansion, warehouse planning, and commercial strategy. Moreover, for us it is an opportunity to explore real-world business data by applying the analytical methods we have learned so far. Furthermore, the direct contact with the company - confirmed through a visit - enhanced the originality as well as contextual understanding and ensured that the analytical decisions aligned with actual business processes. This combination of academic methodology with real operational knowledge is not common in standard coursework projects.

2. Configuración del proyecto

2.1. Fuentes de datos

Our main data were provided by XXX, containing information about sales of the company from 2020 to 2024. The dataset includes the following variables: Year, Month, Euros, Quantity, Colour, Measure, Producer, Type, Category, Province, and Country. Together, these variables allow us to perform a detailed analysis of the company's operations but will require extra information to explain as a whole. We decided to focus only on 2022-2024 interval, because previous years were impacted by COVID-19 pandemic – there were less social or family events than usually. We decided to keep only records corresponding to Spain, Portugal, France, and Andorra, as these countries are geographically and logistically connected to XXX's operations and distribution network, and the representation of other countries was marginal and highly variable across the years.

Secondly, we enriched those data by adding travel distances between Spanish provinces and also their geographical coordinates. They were acquisited using public Open Route Service API.

Furthermore, we took data about the numer of registered Spanish florists and places where flowers are sold from Cámara de Comercio website. But here we faced a problem that the website shows only few entries while searching by category, so we had to automate the process and using webscrapping search by postal code and finally merge the data. They are all potential buyers.

Moreover, we acquired the data about colors, but instead of previously planned RGB codes, we decided to obtain only hue and value from HSV color model, as they are closer to how human eye work and can fully model all the deterministic flower colors existing in the data.

In addition, we obtained complementary economic and geographic indicators: population of Spanish provinces, GDP, numer of marriages and deaths. Those data were taken from Instituto Nacional de Estadística website. But in this case we have limited ourselves only to Spain, since the majority of the sales are within the borders of the country and other data, for example the number of flower shops, we have only for Spain. Comparing with previous assumptions we resigned from using regional festivities, as it could be hard to obtain and would not be of paramount importance in the analysis.

3. Integración y transformación de los datos

3.1. Data cleaning

We encountered few problems with XXX data that needed to be cleaned, such as:

- Mismatches in provinces and countries such as *"Sevilla – Portugal"* or *"Faro – Australia"*. Because each province belongs unambiguously to a single country, we corrected the country field based on the province's correct location, ensuring geographic consistency in the dataset.

- Quantity of sold flowers contained a small number of zero or negative values. According to the company, negative quantities correspond to bonuses granted to clients. Since these records were few and did not represent actual sales, we chose to remove them from the dataset.
- Measure variable (medida) contained a significant number of missing values. We used different techniques here to impute data.
- Colour variable contained duplicated spellings (e.g., Marron/Marrón), compound values (e.g., Blanco/Rosa), and ambiguous labels (Natural, Transparente, Mixto). Spelling inconsistencies were unified, compound colours were simplified by keeping the dominant (first) one, ambiguous values were set to NA, and "Mixto" were assigned random colours following the distribution observed for each product type.

Also economic, demographic and geographic data such as population, GDP or the number of flower shops included different naming conventions for provinces (e.g., "València" vs. "Valencia"), so we standardized names to match those used in the main dataset and enable merging. The GDP data also contained additional economic indicators such as *"Agriculture and fishing"* or *"Retail and transport"*; we filtered these out, keeping only the total GDP variable as it was the most relevant for studying the relationship between economic activity and sales. Minor cleaning adjustments were also needed, such as converting "2022 (P)" to "2022" and removing unnecessary columns. This dataset contained no missing values or outliers.

3.2. Data transformation

After cleaning the data, we started with some transformations to meet the requirements of the objectives:

- Color: The Color variable was originally in text format and needed to be transformed into a numerical representation to analyze similarities, group comparable tones, and study their distribution. After assigning a single colour to each product, we converted it to Hue and Value components from the HSV colour space. Although some colors such as white, black, and grey have no resonable hue, all colours retain a value component, allowing for a consistent numerical comparison.
- Unit price: price per one flower in a single sale.
- Trend coefficient: we created a variable that reflects structural growth or decline — not seasonal peaks. It captures long-term shifts in product popularity, market expansion, or changing customer volumes.
- Weighted group ages: during a research we found a report from Ministry of Agriculture saying that different age groups have different relevance in market and demand for flowers. So here we transformed population in age groups with proper coefficients taken from the report.

3.3. New variables

We also decided to create some new variables that would faciliate the analysis or make it more interpretable and explainable.

- Region: Autonomous Community of a province. May not only help to decrease the number of provinces, but also provinces within tchem similar cultural habits
- Quantity category: discretising of cantidad variable as it presents skewed distrbution
- Seasonality: this variable describes changes in sales of given type (Tipo variable) and using clustering technique assigns a name for seasonality trend

4. Objetivos

Objetivo 5: Understand the structure of the floriculture sector in Spain's, identifying the geographical distribution of florists to assess its potential impact on XXX's sales.

In this objective we merged the economic and demographic data (population, GDP, marriages, deaths, population in age groups) from INE website with geographic data about provinces (distance, latitude, longitude) from Open Route Service API with the data about the number of flower shops from Cámara de Comercio website.

With those data we were able to analyse what are the most important factors deciding the number of flower shops in provinces. Also what is the origin of those factors – whether they have demographic, economic or geographic background.

Additionally, we created weighted age groups and total populations with regard to the Ministry of Agriculture report that shows different importance of them to the flower's market. Also the majority of variables were calculated per capita to avoid scaling issues with bigger provinces, where almost all of the variables (except geographic ones) had higher values.

This objective complements the others by providing the spatial context needed to interpret sales patterns, evaluate socioeconomic influences, and design an efficient logistics strategy. Understanding where florists are located helps explain why sales vary across regions, supports predictive models of demand, and is helpful for optimizing the location and impact of a new distribution center and the resulting logistics network.

In this objective we took a look how variables correspond with each other, also calculated per capita and whether they have direct impact to the number of flower shops. Then we discovered that almost all of the variables per capita are changing in a linear way in time, through different years. It allowed us to model them using linear approximation (and other variables that did not match the linearity were encoded with proper modelling). Given those modelled variables we normalised them and created the linear regression model to predict the number of flower shops from economic, geographic and demographic indicators. Finally, using the information from the model about the most important variables, we visualised those factors on interactive maps, that give us meaningful insight into their spatial distribution, changes in time and distinctive provinces.

Obtained results could help the company while extending their sales to other countries, how they may differ by province. Moreover, it can help understand how change in time of different variables may affect the market of flower shops. Furthermore, results of this analysis may help potential business investors to decided where in Spain is the best place for a new florist or which places are characterised by the biggest threat.

We consider this objective to be of average difficulty, giving it a score of 3. It required to combine data from a few data sources and unify them. Also few transformations and insights into data to understand their meaning and behaviour but this part was not very difficult. Also it contained a regression model, so it needed careful handling of data. Together it required a bit of work, but probably it was not the hardest objective of this project.

The majority of the knowledge and skills needed to this work has been acquired during this subject, Project II, but also some of the knowledge comes from my home university, Warsaw

University of Technology, where I had a course about an introduction to artificial intelligence and in the part about machine learning, among other topics, used here linear regression was discussed.

This objective focused only on open, public data, so there are no issues with the data being private or sensitive.