

****

Bakery Store Data Set

Produce a Project Plan

**Authors**

Lorenzo Foà - [lorenzo.foa@gmail.com](mailto:lorenzo.foa@gmail.com)

José Manuel Pérez - [josemprb@gmail.com](mailto:josemprb@gmail.com)

# Index

[**Index**](#_v4nags4fl59j) **2**

[**Executive summary**](#_auvji2mqky5d) **3**

[**Background**](#_6cpfz8c805ns) **4**

[Project Context](#_7iicl5jjq9nc) 4

[Opportunity Statement](#_dhj2e3gp5f1k) 4

[Business Case](#_69qww7c7etj3) 5

[Service Change](#_vdr70t4dahuo) 5

[**Project Vision and Scope**](#_lylgh4elhk8c) **6**

[Project success measurements](#_som438sfwd5) 6

[Project scope](#_r1zhq1tz66qy) 6

[Project objectives and indicators](#_adp50ho46biq) 6

[Project architecture](#_uyvr9vbnrf1j) 7

[Project communication](#_e4e4yextp39o) 8

[**Project Plan**](#_ruq4hucmywmg) **9**

[Roles and responsibilities](#_z5nplrn1c3aq) 9

[Project Phases and Milestones](#_1j58b9qfejjg) 9

[Risks, assumptions, constraints and dependencies](#_73v9xhhphlgy) 9

[Tasks, schedule and budget](#_31jr5o2isgl) 9

# Executive summary

In this project, a bakery market will be analysed from a dataset containing transactions of different products. This document will serve as guideline to develop it furtherly using an organised structure for the whole project. The main goal of it is to increase the revenues of the bakery by offering different discounts in products that are not highly sold. These discounts should be memorized by the workers of the bakery and remembered to each customer whenever a transaction starts.

The scope of the project is made of a data set of 21.000 entries, which realizes a data analysis to lead into some conclusions. The staff in charge of analysing this data set and others possibly developed is made of two people: Lorenzo Foà and José Manuel Pérez. Both share the ownership of the project. The goals of the project are to fulfill the success measurements of the project, to increase the revenues for the bakery and to reduce the quantity of exceedings produced by the store. Those goals are aligned with indicators that measure their relative performance whenever the project is implemented.

The architecture for the modelling of the data will be done using software statistical tool R, and the algorithm to be applied is a priori. This learning association rules algorithm is a data mining solution that analyses the whole data set looking for associations in the values of the attributes. Once these rules are extracted, the same data scientists in charge for the project will have to make decisions whether or not such rules are important. If they are, they will be converted into offers or discounts for the least sold products. The communication of the project will be handled weekly, by terms of meetings between the communications managers of both parts: Ernestina Mensalvas, in representation of the bakery, and José Manuel Pérez, in representation of the project workers.

The roles and responsibilities for the project are almost equivalent for every employee. José will be in charge of communication management, while Lorenzo will administer the financial management. In parallel, both will develop the data analysis. The total duration of the project, including the different phases and milestones is to be performed in two days. The phases for the project are divided into:

1. Data understanding: Understand the data set for the project. Try to define the structure of the data and set objectives for the purpose of the project.
2. Data preparation: Clean the data so it can be analysed furtherly.
3. Data modelling: Apply different algorithms to the data, trying to find patterns in it.
4. Data evaluation and decision making: Conclude which offers should be recommended.

The tasks of the project are divided into two sections: development and implementation. These tasks are subdivided into more concrete tasks that are detailed in the Gantt chart provided. The total amount of money required to develop the project is broke down in the last section of the planning, the budget, and it is up to 1500€.

# Background

## Project Context

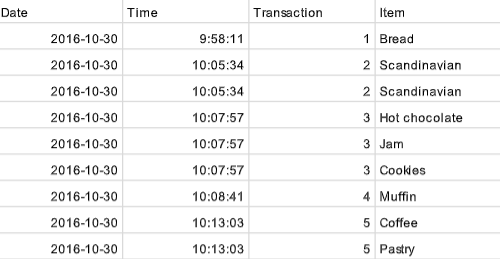
The business understanding of the problem always starts with an interesting question that should be solved. In this case, it is “How to improve the sales performance of a bakery?”. Therefore, it will be precise to develop a project in which some data mining will be applied. This task has to be completed within a dataset containing all the transactions in approximately six months, between the 30th October 2016 and the 9th April 2017. In this dataset, there are 21296 entries grouped in 9684 different transactions. Each entry of the dataset refers to a concrete product sold by the bakery in the time range mentioned aforehead. The dataset is made of the following attributes:

**Date (YYYY-MM-DD):** Date of the transaction expressed in the given format. It only includes values in the six months mentioned aforehead.

**Time (HH:MM:SS):** Time of the day in which the transaction occurs. It is usually registered as the moment in which products sold are saved into the system.

**Transaction:** Identification number for the transactions that occur in the bakery. It is a very useful attribute because it gives the opportunity of grouping different entries by transaction.

**Item:** A string value, which corresponds with the name of the product sold in each entry. More than one item could be sold in the same transaction.



From this dataset, it is possible to develop different data analyses. The most evident is to group the products sold (items) by transaction. From this grouping, it is possible to elaborate a production rule set of products sold in the same transaction. The timestamp of each item is given, so that is possible to know which products are being vended the most in different periods of the day, or the week.

## Opportunity Statement

Once the context is explained, we must set a business goal to focus during the project. In this case, it is to increase revenues of the bakery by creating customized offers for the least sold products of the bakery. This means that dataset will be analyzed to find least vended products, or products that are usually sold but not in every time window. Afterwards, depending on each situation, a different discount will be recommended. We will try to make at least five recommendations, one per each weekday. This strategy is usually followed in fast food restaurants, so that they increase the motivation of the customers when buying these low-trend products. As the bakery follows a similar model, in which clients are likely to buy at least two products, this approach could be very effective. However, as price is also an important factor to decide whether to spend money on a product or not, it has to be highly considered when creating the offers.

## Business Case

In every bakery, there are always products which are trendier, fancier, or even more famous. They are usually sold without problems and become the main incomes of the shop. However, others remain unsold because of different reasons. These ones cause a loss on the bakers, as products are left in the stock, reducing their benefits. If we do not include this offers, we are going to continue losing raw materials, working time and, therefore, money with this products. However, if we include these offers, the products will be sold and these resources will not be wasted. Offers must be customized again after a period of time, which can be controlled setting phases and milestones in the long term. If any promotion does not result as expected, it can be either changed or substituted by another that is likely to be bought.

Therefore, the proposition value for the project includes the following aspects:

* Remove the excess from the stock, by means of increasing the sales of the least fashionable products.
* Modernize the marketing of the store, including offers, discounts and promotion carried out by the workers.
* Increase the value of the bakery, complying with the success measurements.

## Service Change

In other terms, the service change in the bakery will be very low, which is an advantage of this implementation of the project. The staff must be aware of the discounts or promotions that are offered. Also, they should mention then whenever a client buys any product, in order to increase the advertising of the discounts. Apart from that, they just have to continue doing the same work they have been doing, without extra hours or any other type of solution that requires the active participation of the workers on the bakery.

Finally, if the bakery engages to continue with the development of the project, another remedial action will be considered for the staff. They will have to follow a list of questionnaires regarding the performance, the advices they have received by the customers or their own advices of the first discounts given. The feedback coming from the polls will be decisive to create the next lot of supplies.

# Project Vision and Scope

## Project success measurements

The vision of the project is a bakery that is able to fulfill all its success measures. The most important of all this measurements is the return on investment (ROI), that is expected to occur in six months from the delivery of the project. In that semester, not only the investment will be recovered, but it will also increase the revenues on a five percent. However, such increasement in the incomes cannot be capitalized the first semester, as the project cost has to be considered. With both of these requirements covered, the bakery will situate itself as one of the better examples of management of this type of small businesses, as they will have almost no wastings while increasing the revenues.

## Project scope

The project is being developed around a data set of approximately twenty one thousand entries, which can be extended to another data set to be analyzed after the implementation. In parallel, the indicators for the goals will require to analyze other simple data sets with information on the stocks and the incomes.

The staff required to do this job is made of two people, Lorenzo Foà and José Manuel Pérez, trying to develop a customized data mining solution for the store. Moreover, a filtering of the different data will be made before and after applying the learning association rule algorithm. The first filter will divide the data in different days of the week and in days of the month as well. However, the second filter is the most important one, as it decides in a rational way which offers are more interesting from the data perspective for the store. In this last part, the company can also help to make the decisions, giving the bakery’s perspective. As the concrete algorithm to be applied is a priori, machine learning is an area of interest for the solution. Also, the final document to be presented with the development is done in HTML, a language for web applications. This can be considered as an area of interest for the project.

The customer will receive different advertisement on the discounts offered by the bakery. On one side, the bakery will have different posters with the promotions. On the other, the employees must communicate these offers when buying a product. In a wider perspective, customer may start to buy only the discounted products and open a new clientele made of sporadic buyers who are just attracted by the discounts. This will only happen if the offers are really making an impact on a market segment and it will lead to more benefits for the shop.

## Project objectives and indicators

With the scope presented, the next issue to be solved is to adjust the different goals of the project with one indicator each. One of the most important objectives is to increase the revenues, which is already detailed in the success measures for the project. However, there are others aims that we should emphasize into them. The detachment of those objectives is shown is the following list:

* Increase the items sold.
* Reduce the surplus.

The first objective, which was increase the revenues should be evolved into a measurement that could be measured. To do so, it is needed to follow some steps. Firstly, found a metric to measure. In this case, it would be increase the revenues received monthly. Once it is found, try to set an indicator for such metric, and compare it with the current value. The indicator for the goal is, as said before, increase the average revenues in a five percent for next semester. The current incomes that the bakery receives are approximately twelve thousand euros monthly, as there are sold over four thousand products a month with an average price of three euros per item. This can be verified just by asking to the bakery how the revenues they calculate have evolved in the last months, and compare the results with the previous one. Also, starting from the axiom of the average price of an item, we can compute the revenues increasement from the number of items sold.

The quantification of the increasement of items sold can be measured in two ways, being one of them more optimal for the task. The first is to use as metric the number of items sold per transactions. The other is to use the number of items sold per month, or even each day. This second metric is better, as adjusts really well with the objective. The indicator in this part is to increase it a ten percent. Currently, the average of products sold in a month is approximately four thousand items. This is only possible if we do not consider the months of October and April, which are not shown completely in the data set. This measurement would be verified when analysing a second dataset that should be sent by the bakery with the items sold since the project was implemented.

To reduce the excess, the metric is more difficult to be very precise. It has been decided that it could be the actual stock that is thrown every month by the bakers. We will have to ask for it to monitor correctly the information received. Actually, this value is in a 8.7%. The objective would be to reduce it to less than a five percent of wasted resources.

## Project architecture

As the source of data is only one schema, the complexity of this architecture will not be very high. This source is given in a csv file, that should be read in some way to be processed. After this read operation, the data will follow an analysis leading into some information. This information that outcome the process should be evaluated by experts rather than being directly communicated to the store. In such evaluation, decisions should be made in order to choose between different offers. Once this task is performed, the results should be communicated to the company in the most suitable way.

The main model that will be applied to develop this project is a priori algorithm, which generates a set of association rules which depend on two parameters: support and confidence. The support of an event is the probability of occuring a concrete condition of the event. The confidence of one event on another is the probability of occuring a multiple constraint (a condition in both events) which is conditioned by another constraint (a condition in the first event). It is mathematically expressed as the support of happening both divided by the support of the first. These parameters should be treated in some way to deliver a production set of rules. Afterwards, the set of rules is analysed by the experts to reach a conclusion on which are the best proposals for the project.

As we mentioned before, the complexity of the project is low. Then, the resources are in consonance with that assumption. The tool that will be explored and exploited is RStudio. This software tool is very useful when handling simple schemas, as it contains hundreds of thousands of functions and thousands of libraries and packages for the purpose. Almost everything can be solved using this tool, as it also leaves the possibility of programming new customizable functions.

## Project communication

One of the most important parts that should be considered is the feedback implemented for the project. The person in charge of reporting this feedback from the bakery is Ernestina Mensalvas. As it is a very short project which can be improved also in a long term, both situations must be explained separately.

On one way, the project itself is completed in no more than two days, so the interaction with the bakery has to be very specific. At first, a meeting will be scheduled before starting the project just to explain how the project will work. Marketing questions will be asked to the bakers in order to know if they already have feedback from the customers of different products that are overpriced, or that they are likely to buy two products at the same time. From that meeting on, every day in the morning both attendants have to talk for fifteen minutes approximately just to be aware of every update that should be taken into account.

On the other way, the improvements that can be done for the project have to be managed in a different manner. Further analysis on the datasets resulting when applying the offers can compare the efficiency of the project before the offers and after them, or it can also compare which promotions are better. If the bakery agrees to settle a long term contract for the exploration and exploitation of their data, the communication must be handled by experts. Therefore, it will be needed to hire a specialist for this purpose. Both daily telephone calls and weekly meetings can be arranged, but it is not compulsory to do so. Efficiency is improved, as we give freedom to a specialist to decide whether the meeting or the telephone calls are really needed or not.

# Project Plan

## Roles and responsibilities

Since the project is not so big, two persons working on it will be enough. Both Lorenzo and José will work together on the development of the product. The whole development will take us two full days of work, which are lately explained in the Gantt chart. Since José has better social capabilities and spanish knowledge, he will also be in charge of the communication between us and the bakery, explaining them how to effectively implement the new marketing strategy. In order to have fully covered the whole project, Lorenzo will be in charge of the finances. He is considered an expert in his field, as he has been in charge of the financial management in so many projects. Therefore, we can graphically see the roles of each employee in the following table:

|  |  |  |
| --- | --- | --- |
| Workers | Roles | |
| Lorenzo Foà | Data scientist | Financial manager |
| José Manuel Pérez | Data scientist | Communication manager |

## Project Phases and Milestones

The phases in order to have the project done are the following:

1. Data understanding: we have a first look at the dataset. We try to figure out what kind of analysis we can do depending on the informations contained in the dataset. we also check the dataset has a low number of errors or missing values so that we can make an exhaustive analysis.
2. We prepare the data before running Apriori algorithm to get some sets of association rules modelling the dataset. Then we will pick the best set according to some ‘fitness’ function we will later identify.
3. Product testing. Once the best set of rules is decided we are able to build up a set of offers and discount to offer to the customer.
4. Final session in the bakery with José for deciding the prices together with the bakers. The session will end with José training bakery’s employees.

The four phases can not be parallelized. Each step has as input the output of the previous one. So we have to the jobs sequentially, and no job can be started until the previous one is finished. This can be dangerous as a retard in doing one of them could make the whole project to be late.

Between the second and the third job will be generated a document containing different sets of rules: one document for the best set of rule day by day of the week, another containing the best rules for each month and a last one containing the product that rarely get buyed.

## Risks, assumptions, constraints and dependencies

On the economical side the risks of the project are almost nonexistent. The only riskful thing would be that the new price after applying special offers causes loss to the bakery, but it can easily be avoided by correctly computing the proper cost internally in the bakery.

In order to exploit our data, and to make our analysis as effective as possible, we should have clean data from the transaction. And then we need to be in the position to carry out our project in the shop.

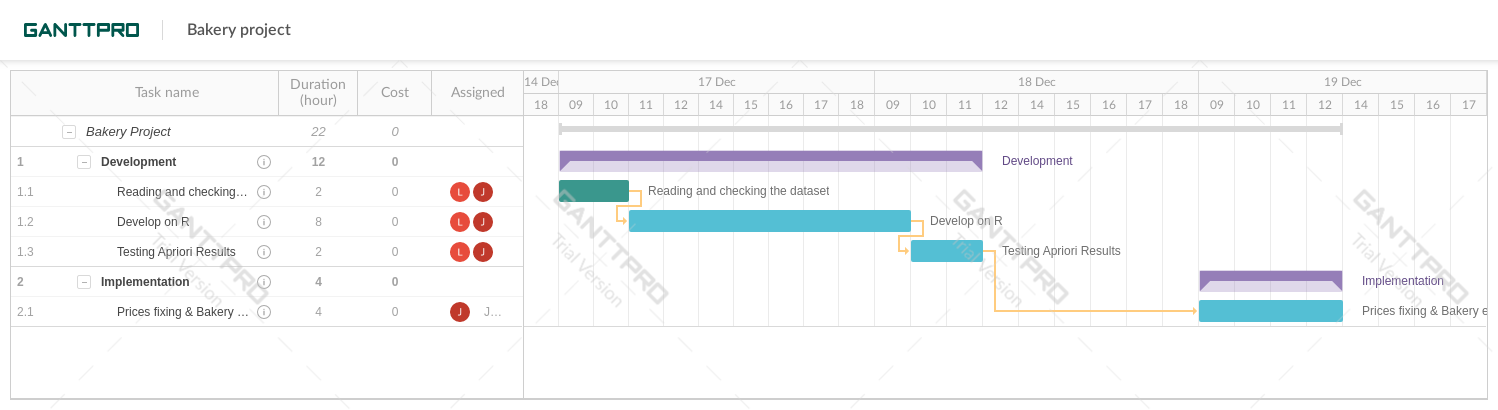
The assumptions we make, in order to be able to have the job completed, are:

* The data set has been correctly received, with the minor possible number of mistakes.
* The bakery will apply the offers as suggested by us.
* Also, the bakery will follow any recommendation given in terms of service change or business change.

The main constraint, on the side of the analysis we can do, are data themselves. Our analysis indeed can only use the attributes given in the dataset, without taking into account many other factors that could be very interesting for solving his problem.

We are depending on the reliability of the computers we will use and on the bakery’s collaboration.

## Tasks and Schedule (Gantt Chart)



In the above diagram, it is shown the activities and the time they will require to be done. They are divided into two sections: development and implementation.

## Budget

The customer get charged 50€ for each developing hour of a junior engineer. The special training session for the bakery employees will have a cost of 300€. The total development hours are 24, divided in the tasks as listed above. So the total cost of the project will be 1500€.