Saving time and money in the kitchen

Andrea Cardinali, Adonis Granita Kingsley,
Matteo Simeoni

DATA ACQUISITION

GialloZafferano

Carrefour

	Recipe Name	Portions	Cooking Time (min)	Calories (Kcal)
0	Insalata di riso	2	35	660.0
1	Chili con carne	2	135	562.0
2	Gateau di patate	2	100	317.0
3	Paella de marisco	3	95	617.0
4	Piadina romagnola fatta in casa	2	34	488.0
...	...	...	...	...
500	Hosomaki vegetariano	4	50	274.0
501	Cornbread burger con porcino	3	100	322.0
502	Maki di kiwi e salmone	3	42	306.0
503	Demi baguette al salmone	2	35	728.0
504	Gateau di patate con zucca e taleggio	3	110	NaN

	product_name	brand	price	source_file
0	Carrefour Classic Burger di Prosciutto Cotto 2...	Carrefour	15,93	GASTRONOMIA
1	Carrefour Veg Medaglioni Bulgur, broccoli e po...	Carrefour Veg	14,39	GASTRONOMIA
2	Carrefour Veg Medaglioni Melanzane, farro e mi...	Carrefour Veg	14,39	GASTRONOMIA
3	Carrefour Extra Cannelloni Ricotta e Spinaci 3...	Carrefour	9,97	GASTRONOMIA
4	Carrefour Sensation Vegetal Cous Cous Aromatic...	Carrefour	15,95	GASTRONOMIA
...	...	...	...	...
7718	Gamberetti Boreali in salamoia	Polar Seafood	47,92	PESCE
7719	Bubble Tea Black Forest - Gusto ribes nero 450 ml	Flavour Drink	10,00	PESCE
7720	Bubble Tea Tropical Jungle - Gusto ibisco 450 ml	Flavour Drink	10,00	PESCE
7721	Bubble Tea Wild Straberry - Gusto fragola 450 ml	Flavour Drink	10,00	PESCE
7722	Bubble Tea California Peach - Gusto pesca 450 ml	Flavour Drink	10,00	PESCE

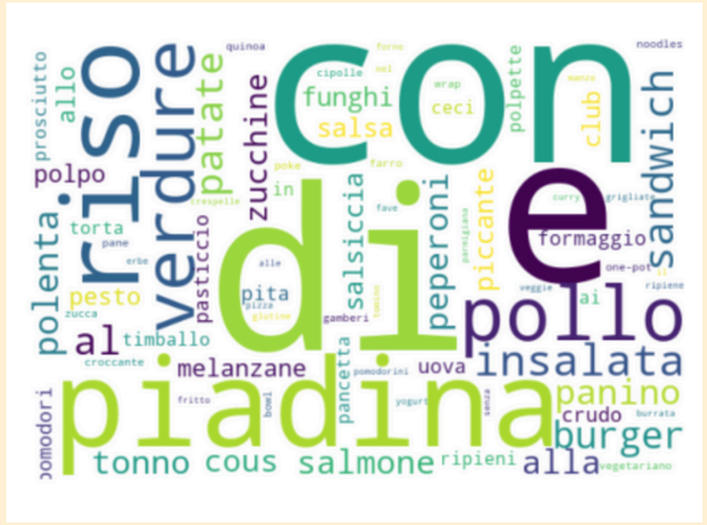
DATA CLEANING

Apostrophe Fix: Corrected "Olio extravergine d’oliva" to "Olio extravergine di oliva".
Symbol Encoding: Replaced symbol 1/2 with 0.5.
Unit Conversion: Standardized all quantities to grams (gr) for consistent pricing.

Used reference tables for converting units like "cucchiaio" and "cucchiaino". Assigned approximate values to vague quantities ("q.b." to 3g).

Brand Removal: Stripped brand names from product names.
Text Normalization: Removed information in parentheses, transformed text to lowercase, and eliminated non-relevant words.

DATA EXPLORAITON



Frequent recipes

Frequent ingredients

Frequent brand

DATA ENRICHMENT

Objective: Enhance the recipe dataset to provide better insights into:

- Cost per Portion
- Preparation Time
- Caloric Content

Enrichment Data: Added information on calories, portions, and preparation time for each recipe.

- **Final Output**: Created a comprehensive JSON file that includes:
  - Recipe details
  - Ingredients and their prices
  - Number of portions
  - Caloric content per portion
  - Total preparation time

MATCHING

Final datasets:

- Price of each ingredients
- Total price per recipe

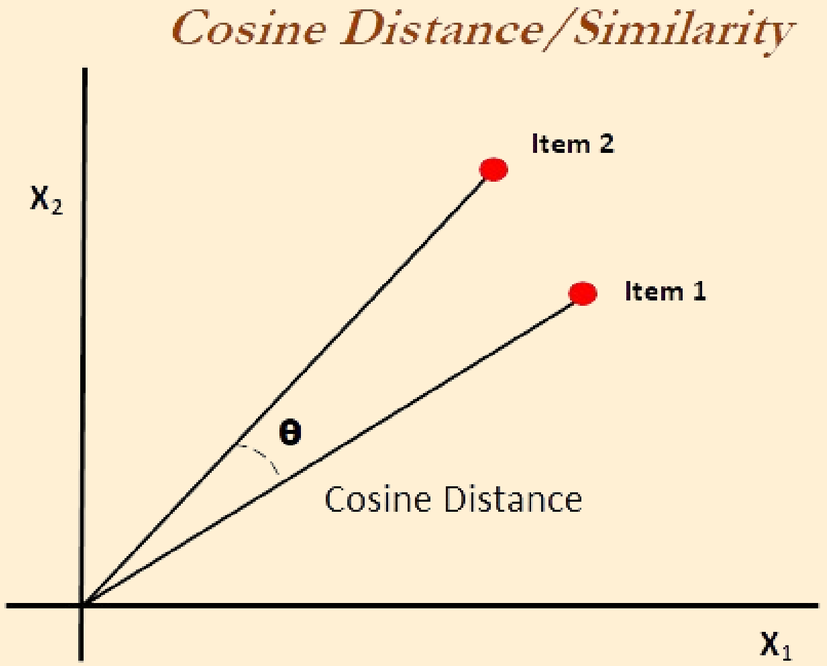
	ingrediente	price_max	price_min
0	"Acciughe sottolio	49.83	49.08
1	"Acciughe sottolio filetti	47.60	38.40
2	"Burrata (4 da 125 g luna)	20.72	13.52
3	"Cosciotto dagnello	3.99	3.99
4	"Filetto di salmone da 200 g luno	74.50	17.45
...	...	...	...
858	macinato grossolanamente)	NaN	NaN
859	o di pesce volante)	69.75	6.64
860	pulito e rifilato	19.90	19.90
861	pulito e rifilato)	19.90	19.90
862	sgocciolato	NaN	NaN

Matching

The Similarity Index was computed through a two-step process.

- **First**, the text was transformed into numerical vectors using the Term Frequency-Inverse Document Frequency(TF-IDF) method.

- **Next,** we calculated the cosine similarity between these vectors.



DATA QUALITY

prioritize data completeness over accuracy.

- Achieving very high accuracy proved to be extremely challenging.
- we took several measures to achieve the best possible accuracy while maintaining a good level of completeness.

To do so we performed a **2 step matching process.**

DATA QUALITY

FIRST MATCHING

- similarity threshold of > 0.7
- Removed unecessary words that could alter the similarity index.
- multiple matches of producst for each ingredient

SECOND MATCHING

Performed only for the remaining **ingredients without a match**

- reassigned words present in the remaining ingredients that we had previously eliminated.
- similarity threshold lowered to >0.4
- highest similarity index for each ingredients

DATA STORAGE

Non-relational model:

- Best for handling unstructured data from different sources
- Optimized for storing JSON files
- schema flexibility

mongoDB

FINAL RESULTS

The 5 cheapest per portion recipes with low cost products

The 5 recipes with cooking time < 20 min and with the cheapest price per portion.