# PAN PARAGON × ZABKA BASKET ANALYSIS — PROJECT OVERVIEW

This project analyzes Zabka retail invoice data provided by Pan Paragon, aiming to uncover sales patterns and cross-selling opportunities – especially around Food Corner (FC) products such as hot meals, salads, sandwiches, deserts and coffee.  
It’s part of a collaboration between Matias Glusman (gastrologic adviser) and the data-science team exploring how Pan Paragon’s invoice-keeping platform can be monetized through business analytics and recommendation systems.

The end goal is still open to interpretation and exploration since the project is in its first steps. We need to understand:

1. Whether there is a product in it.
2. How it should be designed (app, dashboard, LLM, traditional ML etc.)
3. Who it can attract and how it can be monetized.

## Optional Business Case

Use real consumer purchase data to understand when and what customers buy, identify hidden cross-selling opportunities, and ultimately prototype an app-driven recommendation engine for retail chains like Zabka.  
The envisioned product would:

* analyze a shopper’s digital receipts in real time when uploaded to the app
* detect behavioral patterns (“morning coffee → sandwich + fruit”) by modelling and forecasting data
* and deliver personalized sale suggestions or category-mix insights to boost basket value

This analytical phase validates that Pan Paragon’s data can reliably produce such insights.

## Data scope

The initial data is based on a collaboration with Pan Paragon, but the data ingestion model may change in the future. This data will help us construct a baseline and experiment with different possibilities for the final product.

The data comprises of ≈ 100,000 invoices, delivered in 13 CSV files (1–10000.csv → 120001–123743.csv).  
Each file contains line-level receipt data with columns such as ID Paragonu, Data zakupu, Godzina zakupu, EAN, Nazwa produktu, Ilość, Cena jednostkowa brutto, Kasjer, Metoda płatności, etc.  
Files are merged into one Parquet dataset (data/\_invoices.parquet) for efficient processing.

## Analytical objectives

1. Normalize & merge all invoice shards into a single dataset.
2. Fuzzy-match product names to a canonical Food Corner menu.
3. Create a view of the data which combines products into simpler categories (all FC pizzas are pizza, all FC burgers are burger) for easier decision making.
4. Compute KPIs describing shopper behavior across day-parts.
5. Visualize Food Corner performance (heatmaps, rankings, basket metrics).
6. Detect co-purchases to power recommendation logic (“what sells with what”).

## Time-slot definitions

|  |  |  |
| --- | --- | --- |
| Slot | Hours | Label |
| 1 | 05:30 – 09:00 | Going to work |
| 2 | 09:00 – 12:00 | Morning groceries |
| 3 | 12:00 – 17:00 | Lunch time |
| 4 | 17:00 – 23:30 | After work |
| 5 | 23:30 – 05:30 | Probably outliers |

## Deliverables

All reports (CSVs + PNGs) should be stored under outputs\_large/.  
They include:

* Basket KPIs (average value, items, receipts by slot)
* Food Corner metrics by slot and product
* Mix analysis (FC vs non-FC within receipts)
* Rankings (top products overall and per slot)
* Heatmap (product × time-slot, red/blue)
* Receipt-level FC share histogram
* Co-purchase one-pagers (top-5 FC anchors and aggregated ANY-top-5)
* Optional lift metrics for co-purchase strength

Deliverables are expected to be presentable and easily interpretable by stakeholders who are not data savvy. We should decide whether a Grafana dashboard makes sense at the early stage to allow for navigation in the data.

## Toolchain

Python 3.10 +, Pandas, RapidFuzz (for fuzzy matching), Matplotlib, DuckDB (optional for fast IO).  
At this stage, we work in a notebook environment before we lay out a more complex architecture.  
Data folder: data/PANPARAGON-1114/  
Outputs folder: outputs\_large/

## Suggested Plots

### Core KPIs by time slot (ALL items)

* Avg basket value overall by slot – bar, x=slot, y=avg basket value.
* Receipts overall by slot – bar, x=slot, y=unique receipts.

**Why:** baseline shopping rhythm independent of FC.

### Food-Corner (FC) activity by time slot

* Food corner sales value by slot – bar, x=slot, y=FC sales value.
* KPIs fc only by slot — avg FC value/items per receipt-slot where FC appears.

**Why:** “hot hours” for FC and FC intensity per receipt.

### FC vs Non-FC mix inside FC receipts

* Mix within fc receipts (columns: fc\_label, value, units, value\_share, units\_share)
* Mix value fc vs nonfc – stacked (or two bars) value comparison.

**Why:** cross-sell headroom; how dominant FC is when it appears.

### FC product rankings (global)

* Rank fc all products (product, sales\_value, units)
* Fast movers top 20 value – horizontal bars for top-20 by value.

**Why:** identify heroes and tail.

### Per-slot product rankings

* Top products by value per slot (top-30 per slot)
* Top products by units per slot (top-30 per slot)

**Why:** day-part merchandising/cooking cues.

### FC heatmap (product × slot) — red/blue

* FC heatmap matrix share (rows=product, cols=slot, values=share)
* FC heatmap share bwr

**Why:** which FC items are slot-anchored vs. all-day.

### FC share per receipt (diagnostic)

* FC share per receipt (receipt\_id, basket\_value, fc\_value, fc\_share)
* FC share per receipt – only receipts that contain FC items

**Why:** upsell runway (how much of a basket FC captures).

### Co-purchase insights (recommendation seeds)

* Top-5 FC items → their top co-purchases
* Copurchases top5 FC (fc\_item, co\_item, co\_count)
* Aggregated: ANY of Top-5 FC → top co-purchases

**Optional enhancement:** add %lift columns:

## Visual style & parameters

* **Time slots (fixed order):** by analysis time slots disregarding outliers (night sales that are either rare or an artifact)
* **Axes:** x-tick rotation 30°, ha='right'.
* **Top-N defaults:** heatmap rows N=30, fast movers N=20, per-slot plots N=15, co-purchases per anchor K=8, aggregated K=15.
* **Heatmap colormap:** logically, I don’t see an intuitive idea to have negative relations, so we should probably stick to 0-1 with intensity of relation, unless you see a reason to have a -1 to 1. Let me know.
* **Bar labels:** annotate counts for co-purchase plots; leave KPI bars uncluttered.

**Data contract (inputs we rely on)**

* Canonical menu list: zabka\_food\_corner\_menu\_canonical.csv (we already built).
* Auto-mapping output: auto\_fc\_mapping\_from\_menu\_v2.csv with columns:  
  product, product\_norm, lines\_count, match\_item, match\_category, score, is\_food\_corner\_auto  
  (threshold default 0.55; adjust if needed.)

## Pipeline steps (high level)

1. **Ingest** 13 CSVs → normalize headers → write **Parquet**.
2. **Fuzzy map** invoice products → canonical FC (token blocking + rapidfuzz if available). For the future we need to figure out a way to build this on a larger scale for any store, not Zabka specific. This should be able to distinguish between FC and non-FC items, e.g. frozen pizza for home and a takeaway pizza.
3. **Derive** ts, hour, **slot** logically, by opening hours.
4. **Dashboard pipeline** – future development?