Intale Report

Business Analytics and Personalization Techniques 2021-2022

Dataleak

Iliadis Viktoras 8180026

Ioannis Vagionakis 8180009

Antreas Sofos 8180119

Ioannis Vogas 8180013

Nikos Georgakopoulos 8180016

Dataset 1

200 Stores

Last 2 Years

Dataset 1

135,367 rows of sales data

Parameter	Info
Geography	Location , Type (Kiosk - MiniMarket)
Product	32 Categories
Measures	Quantity & Revenue
Time	Monthly Dec 2019 - Nov 2021

What we focused on

01 **Segmentation of the Stores**

02 **How the Segments Differ and Propositions**

03 **Lockdown Effect Analysis**

Dataset 1

Store Segmentation

Clustering Basis 200 Stores

01

Store Type and Location

02

Seasonality of Sales

03

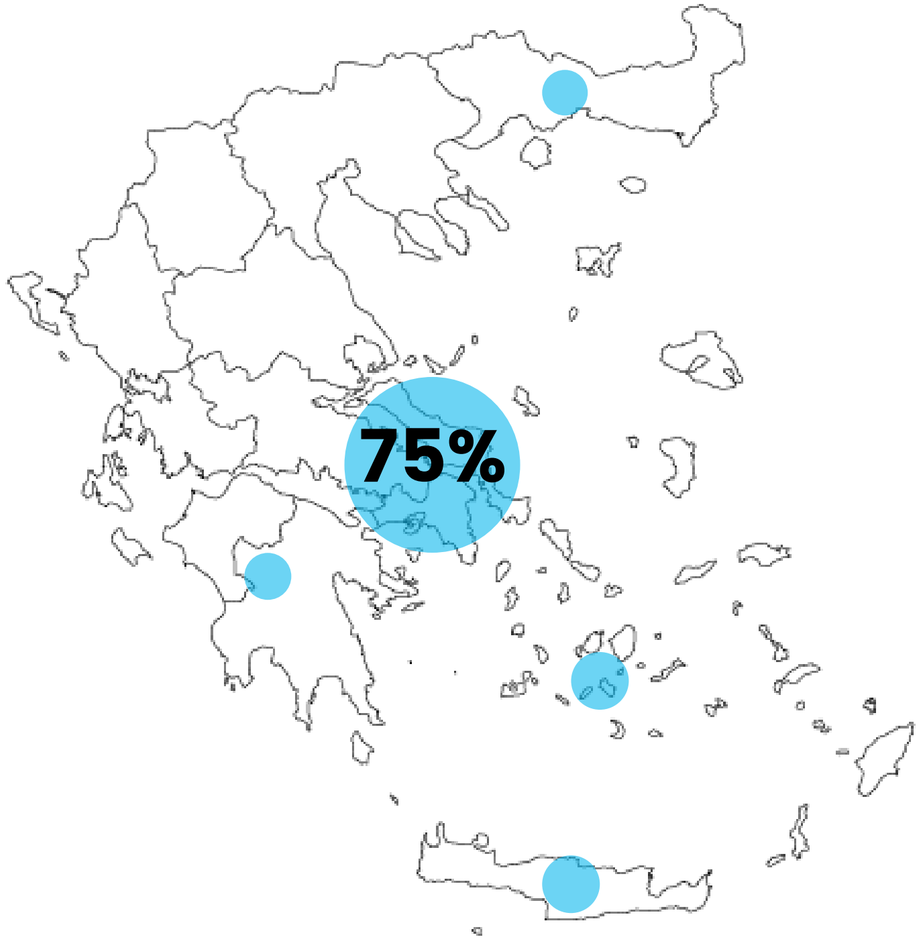
Sales - Quantity , Monthly Revenue, Retail Price

04

Categories Variety and Concentration of Sales in them

Attica - High Earners

80 Stores - 40% of Total Stores



01 75 % Attica - 25 % Highest Earnerns of other regions

02 63 Kiosks - 17 Mini-Markets

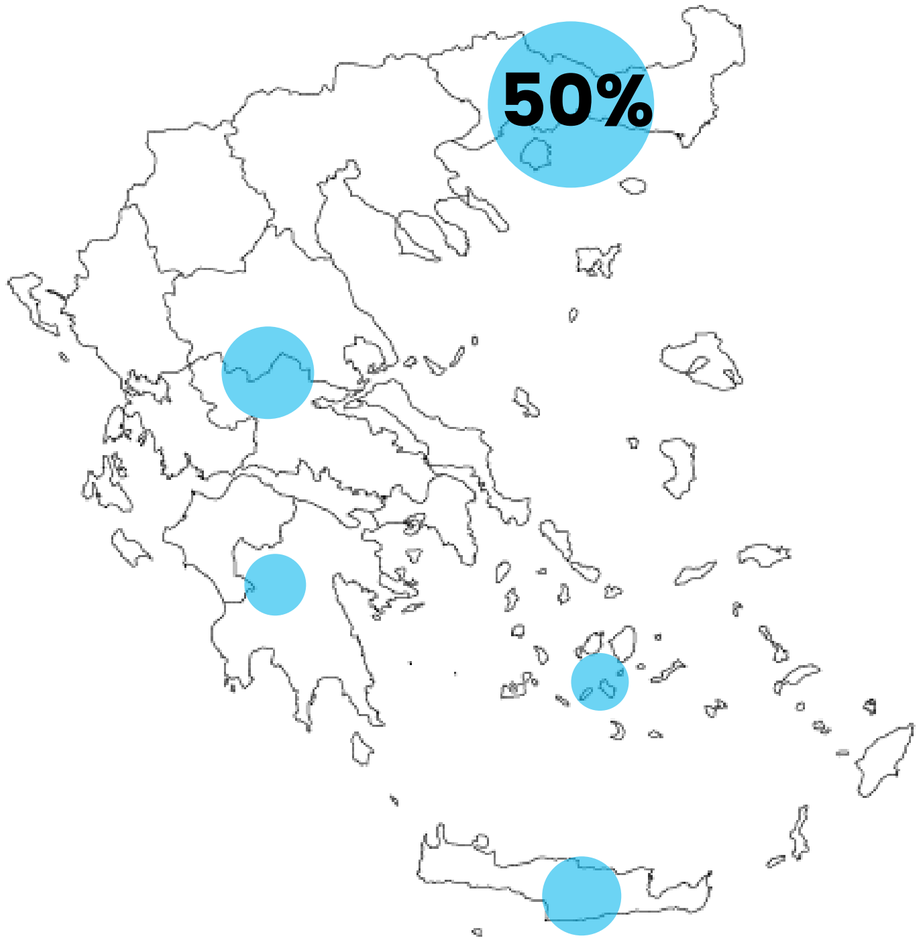
03 Lowest Summer to Year Performance

04 Offer AVG 28 Categories - 7% Produce 80 % of Rev

05 2.42 AVG Price Highest Observed , Double the sales of other Segments

Rural Mini Markets

60 Stores - 30% of Total Stores



01 50% Macedonia -50% Rest of Greece without Attica

02 Highest Distribution of Sales to different categories

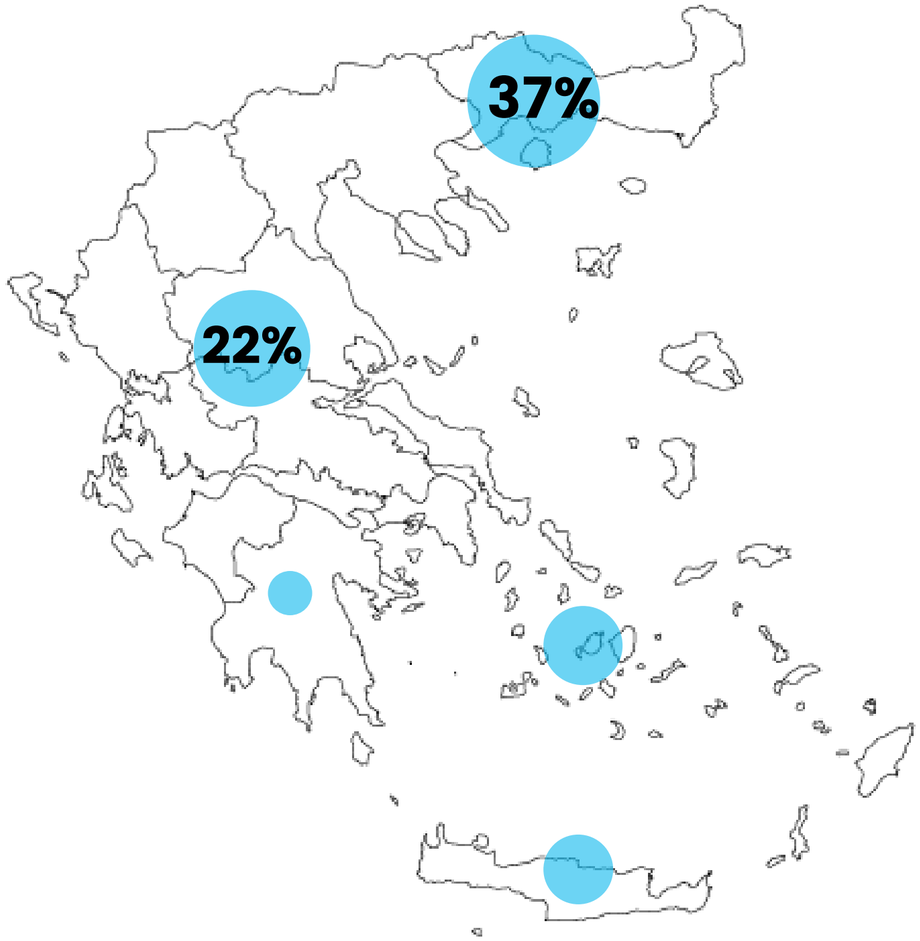
03 Largest Categories Variety

04 90% Mini Markets to 10% Kiosks with higher Rev to categories Distribution

05 2.26 Lowest AVG Price Observed , Lowest Revenue Second Lowest Quantities

Rural Kiosks

60 Stores - 30% of Total Stores



01 Relatively evened out Locations - Without Attica

02 80% Kiosks to 20% Mini - Markets

02 Slightly Higher Summer Performance

02 Only 5% Of categories amount for 80% of Revenue - Ciggaretes , Lowest amount of Categories Offered

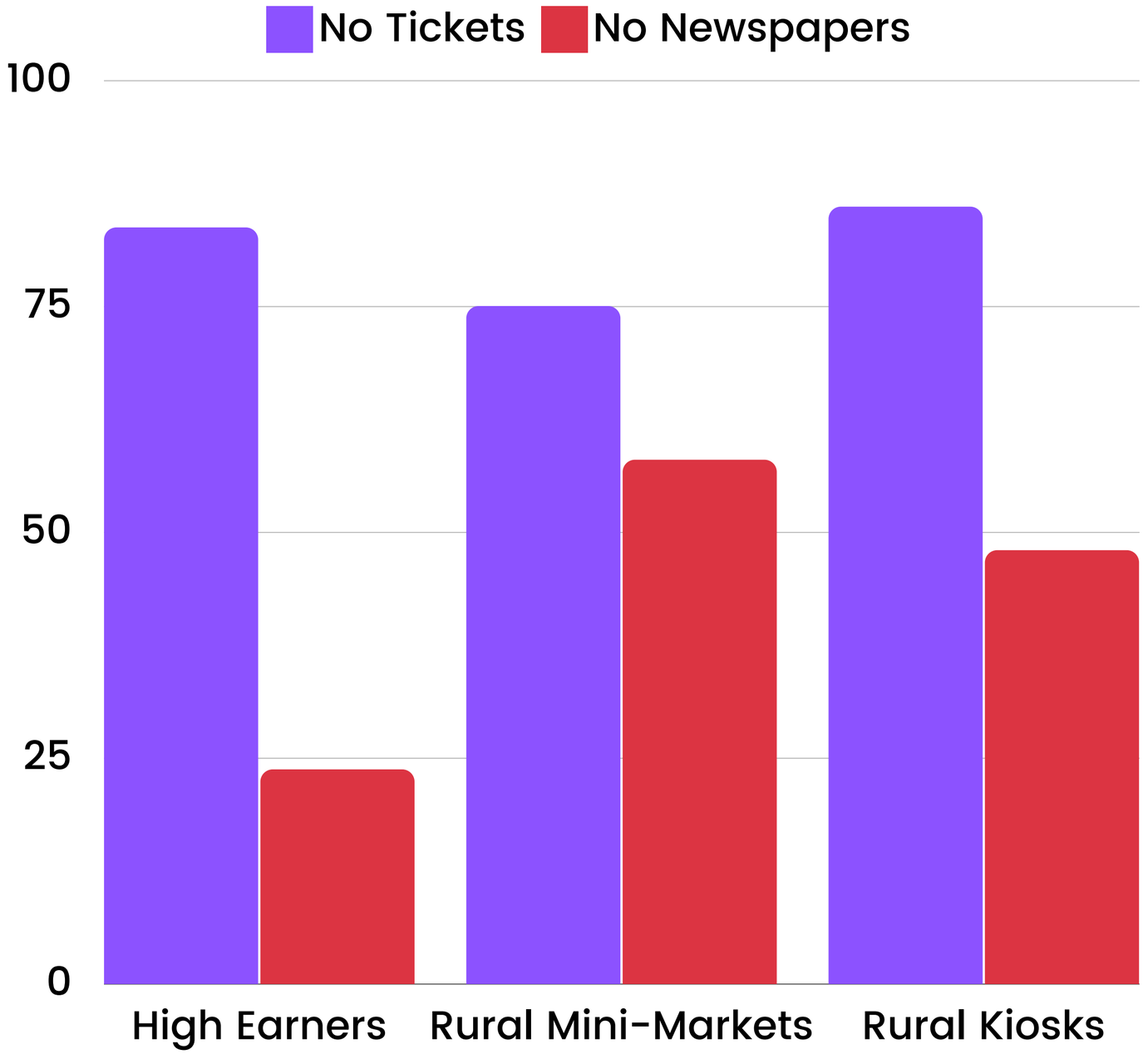
Segments Strategy

Cross - Selling

Newspapers and Tickets

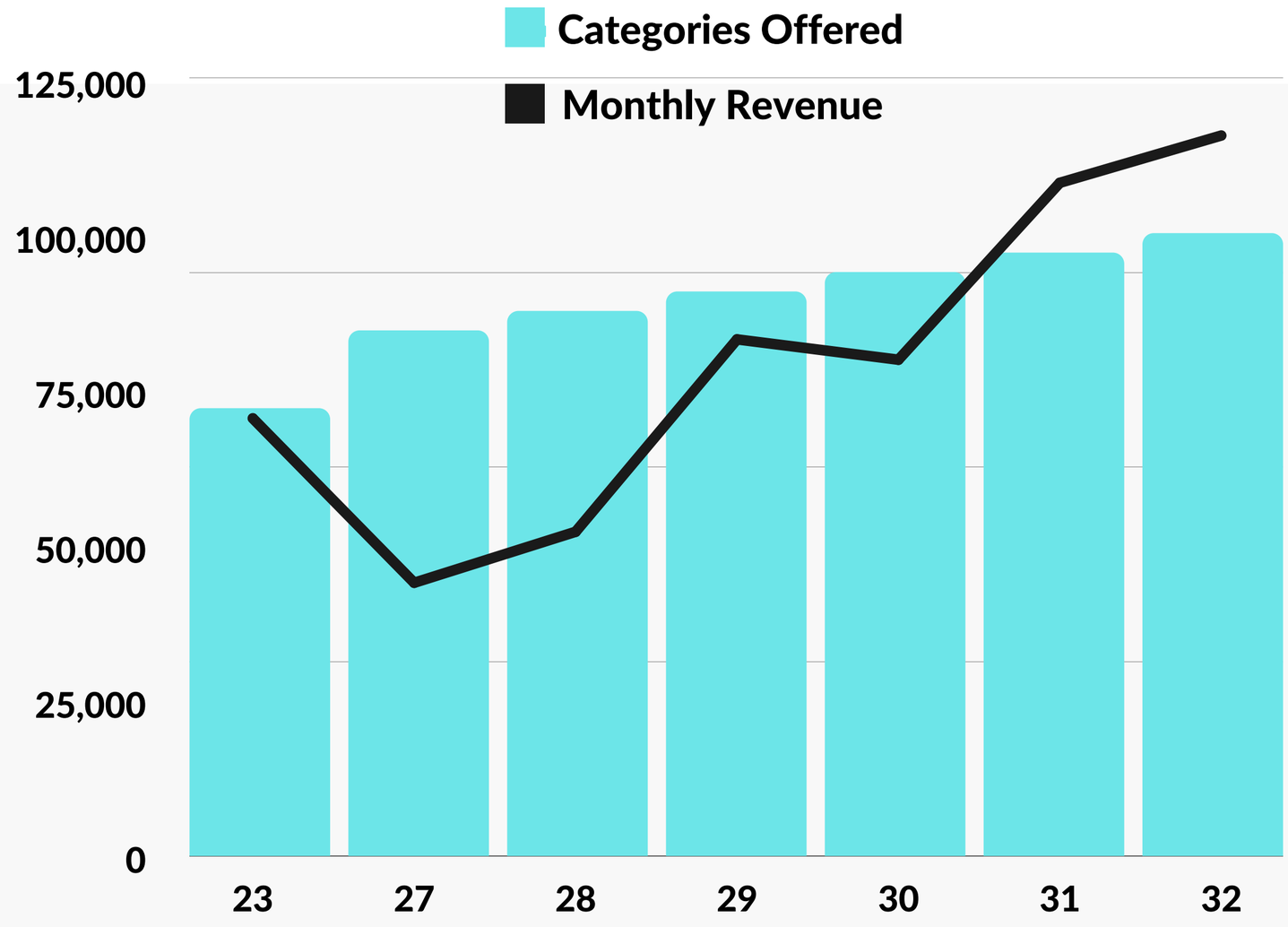
Large percentages of the stores especially the rural markets, do not sell newspapers and tickets perhaps because they have very small room for net profit.

But they are great opportunities for attracting customers into the store and cross selling



Rural Mini-Markets

Proposition



- Positive Correlation (0,5) Beetween Amount of Categories Sold and Sales (Quantity and Revenue)
- Negative Correlation beetween sales/categories concentration

Proposition for Rural Mini-Markets

Bring more product categories from suppliers - try to antagonize bigger channels, have broader baskets.

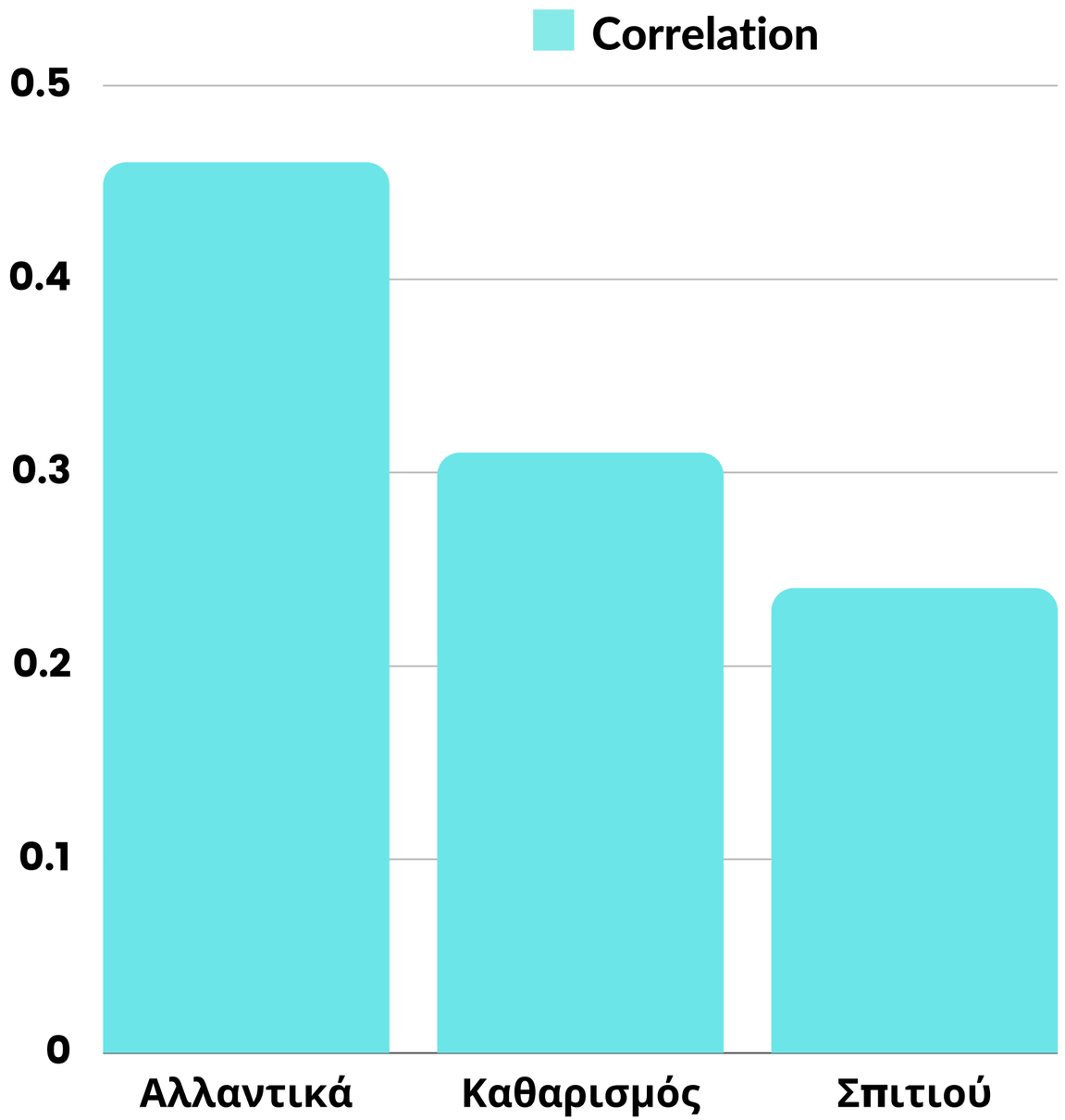
Rural Kiosks

Category Price to Produced Revenue

- Positive Correlation beetween certain products price , and monthly revenue

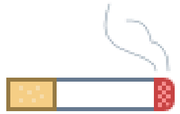
Proposition for Rural Kiosk Segment

increase prices on products people would buy in bigger chains , like cleaning and house merc.



Lockdown effect on Segments

Dec 19, Jan 20, Feb 20 vs
Dec 20, Jan 21, Feb 21

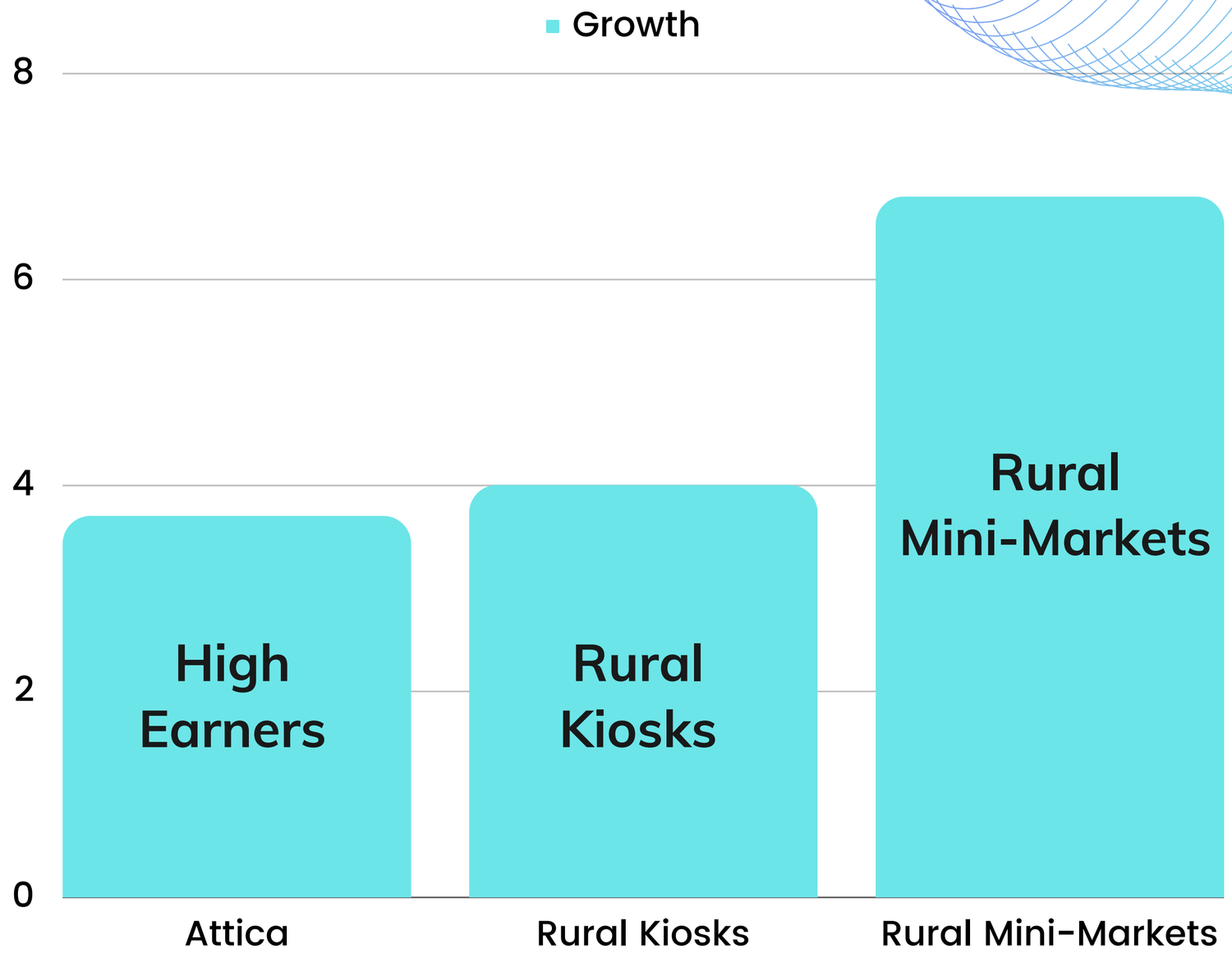
					
Attica - High Earners	61%	110%	75%	122%	89%
Rural Markets	67%	131%	55%	121%	131%
Rural Kiosks	53%	87%	73%	64%	77%

Segments Growth Last 2 Years

High-Earners 3.7%

Rural Kiosks 4%

Rural Mini-Markets 6.8%



The fear of coronavirus contagion in large and packed super-markets , created an opportunity for small retail to be a source for broader type of sales .

Dataset 2

10 Stores

October 2021

Dataset 2

128,129 rows of sales data

Parameter	Info
Geography	Location - Store Type
Product	32 Categories
Measures	Quantity & Revenue
Time	Day - Time /Octomber 2021

What we focused on

01 **Transactional Store Segmentation**

02 **Product Correlations**

Clustering Basis
10 Stores

01

AVG Revenue , Categories, Quantity per Transaction

02

AVG Daily Transactions

03

Off - to Peak Hours Transactions to Total

04

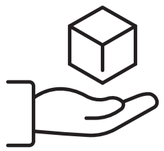
Weekend Transactions

Dataset 2
Standard Kiosk

Contains 7 stores

Tend to operate during night hours

Typically more profitable on weekends



Make more, but less lucritive, sales

Dataset 2
Provincial Mini Markets

Contains 2 stores

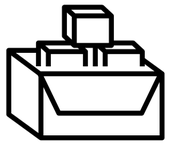
Operate during peak hours only

Typically more profitable on working days

Tend to sell a limited variety of products per transaction

Dataset 2
Large Kiosk in Central Location

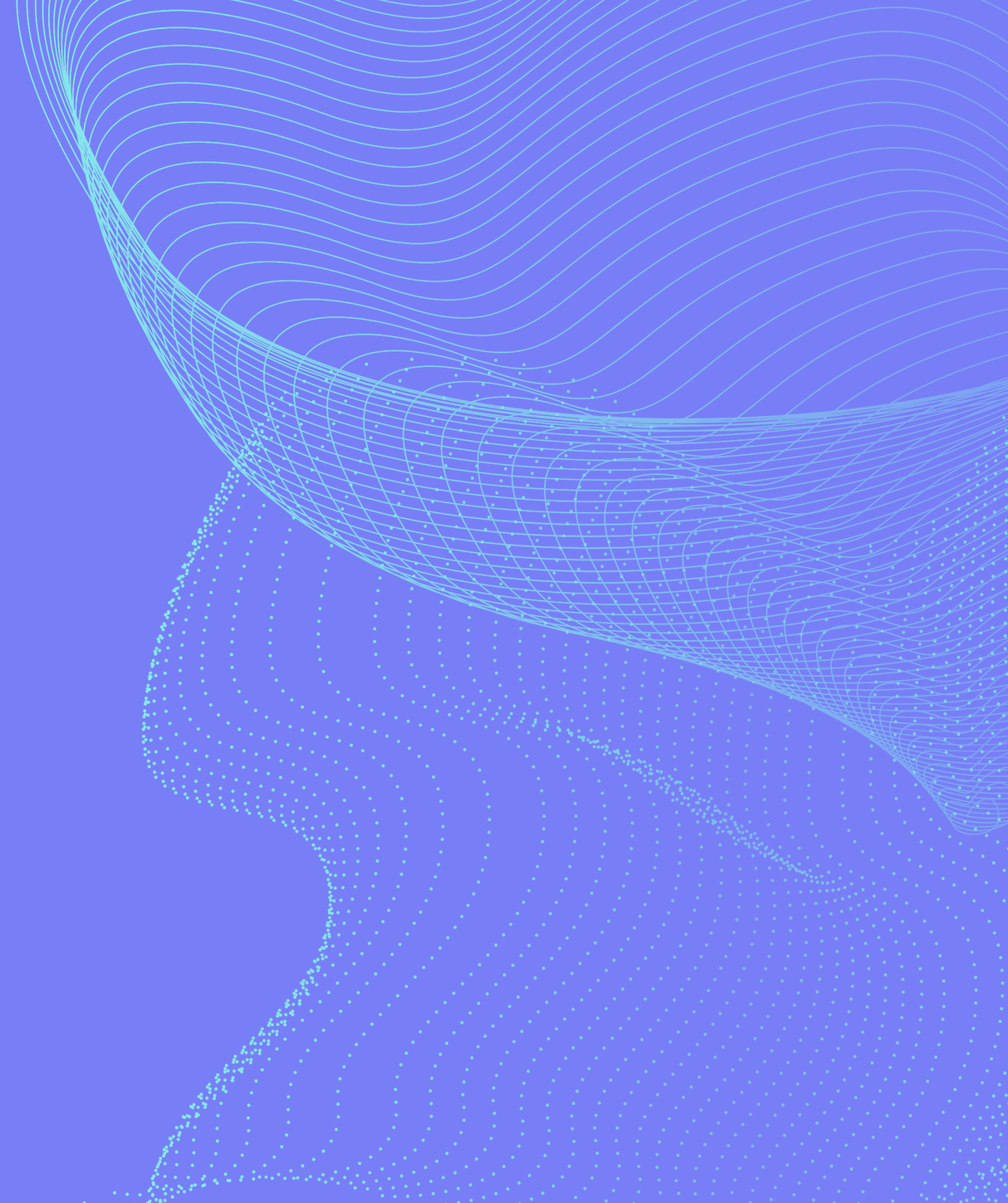
Contains 1 store



Its sales volume and variety are significantly bigger

Sales on working days and working hours make bigger part of its income than the average store

Cluster Analysis

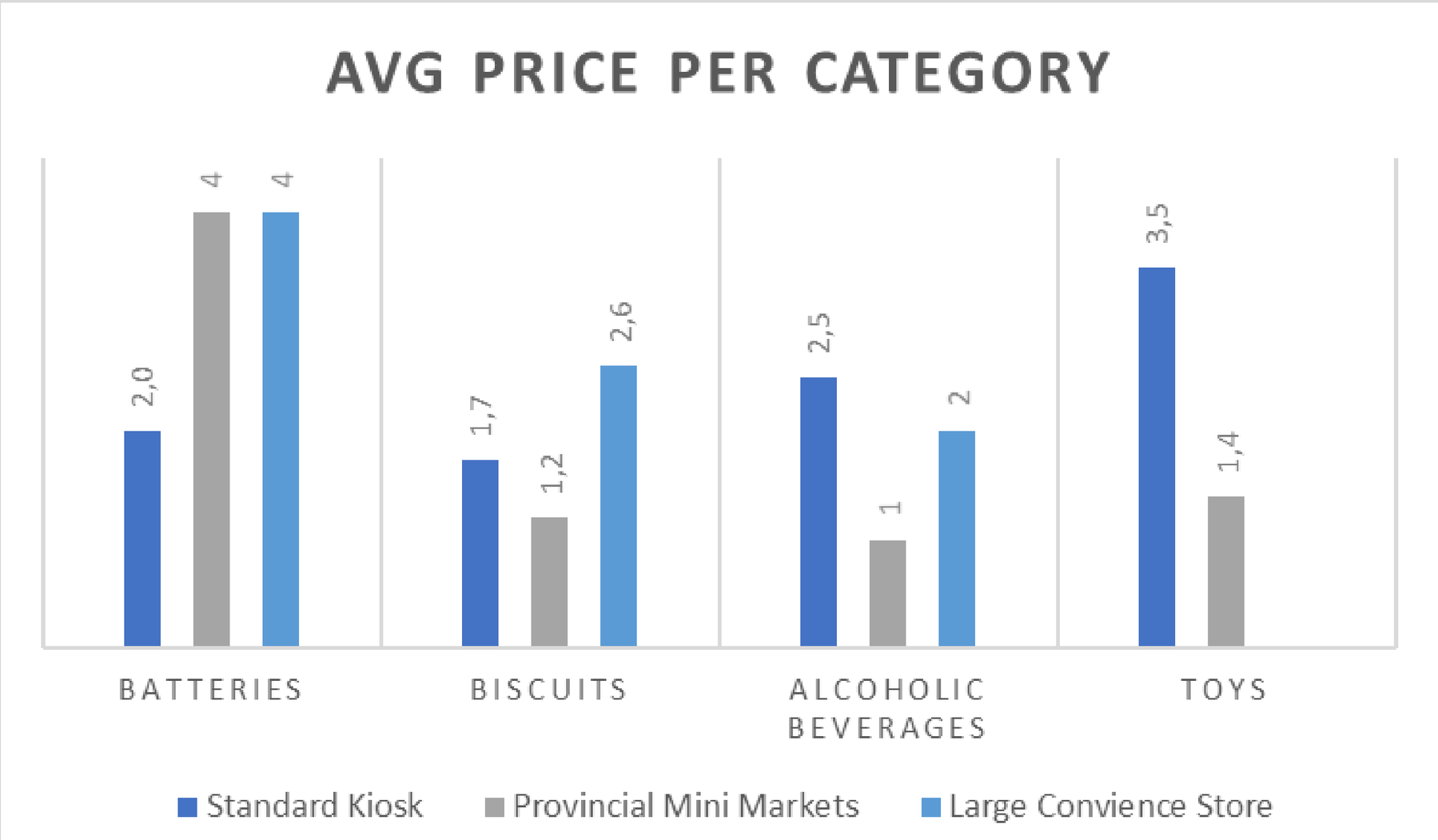


About baskets

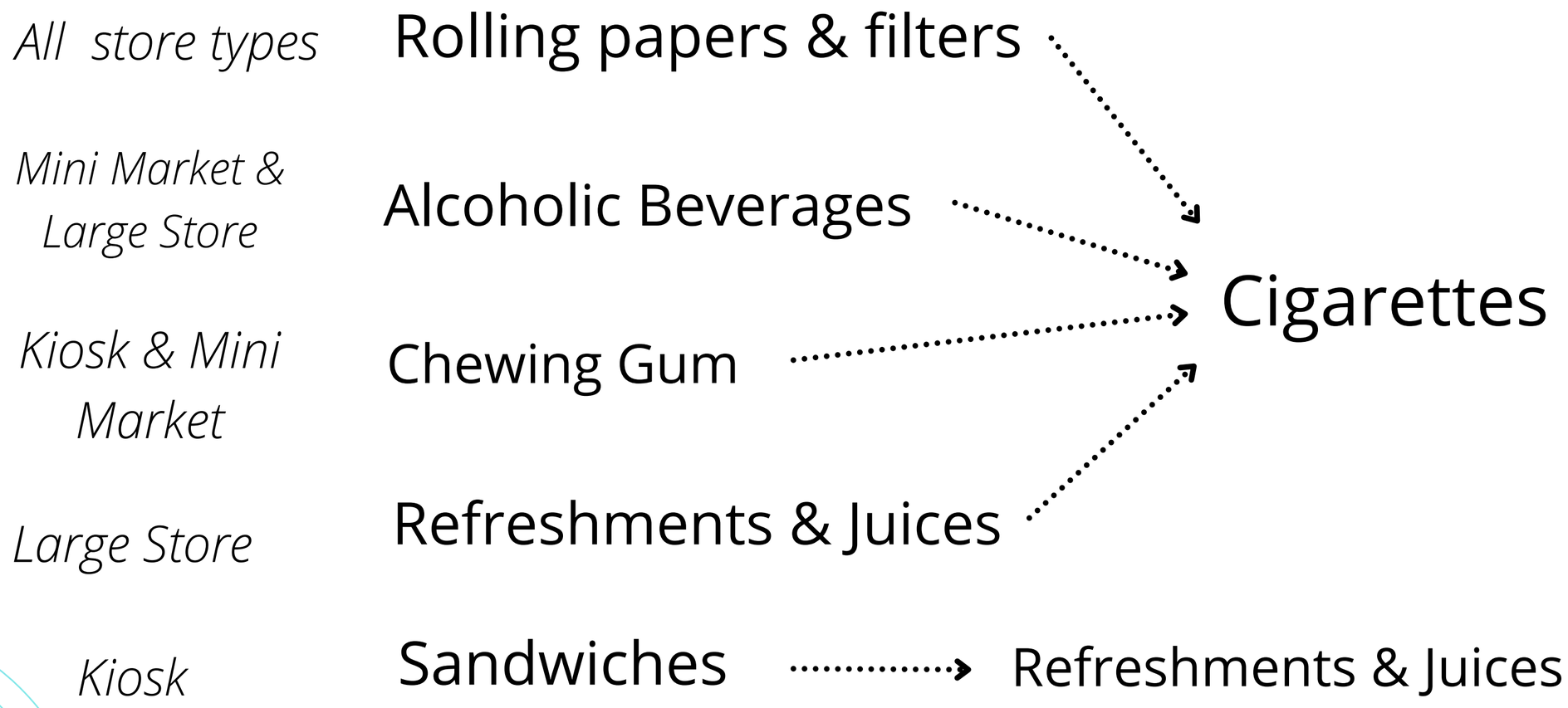
97.800 Baskets

	AVG basket size	AVG basket value	Percentage of total baskets
Kiosks	1,9	5,6	86%
Mini-Markets	2	7,1	7%
Large Stores	3,4	11,2	7%

AVG price product



Product correlations



Segments Strategy

Most "powerfull" product combination

for all store types

Snacks ⟷ Refreshments & Juices

Stores missing on this opportunity

Standard Kiosks

Store *4139* Store *7227* Store *5458*

Provincial Mini Markets

Store *3788*

Our proposal: "Bundle them!"

Put those products in close range inside the store

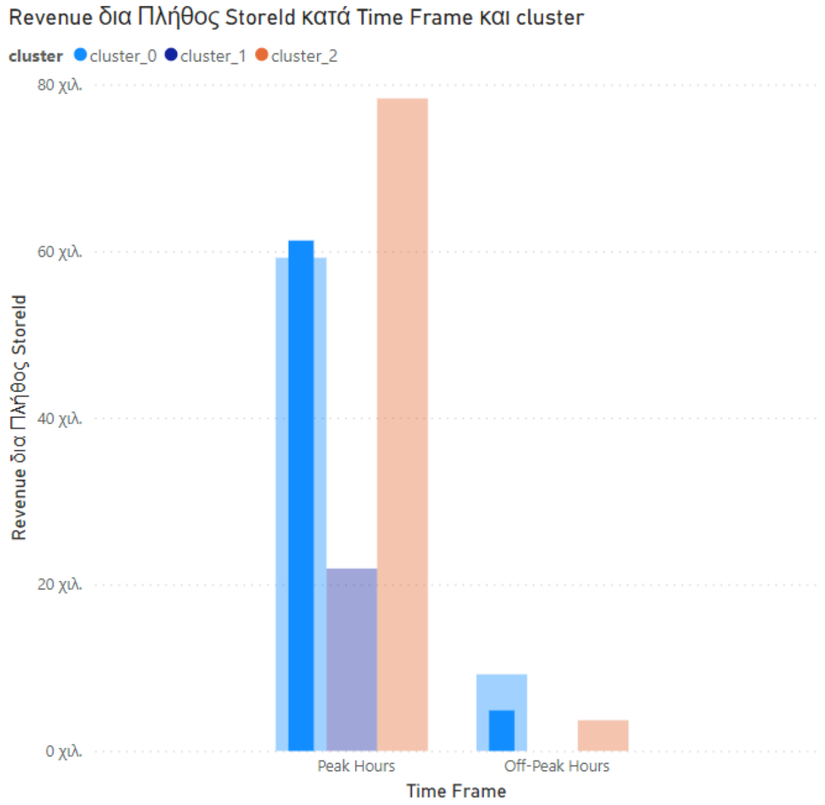
Display ads of these products on the kiosk screens

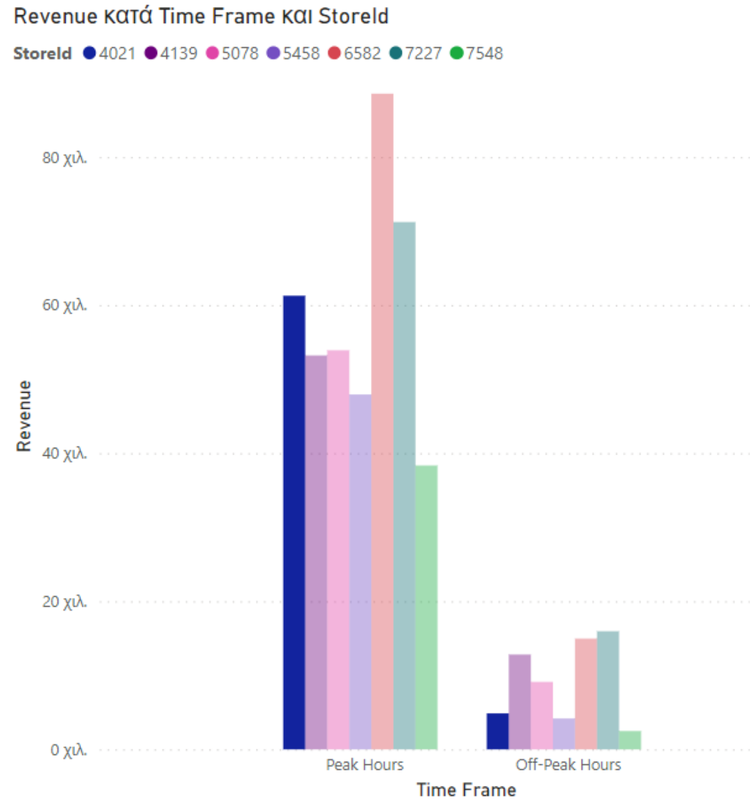
Introduce promotional offers that include these types of products as a bundle

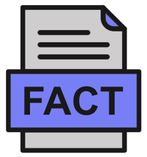
Some stores that perform better than others during peak hours don't do so on off-peak ones

Get a bigger share on off-peak hours

Expand their oppening hours and spread the word



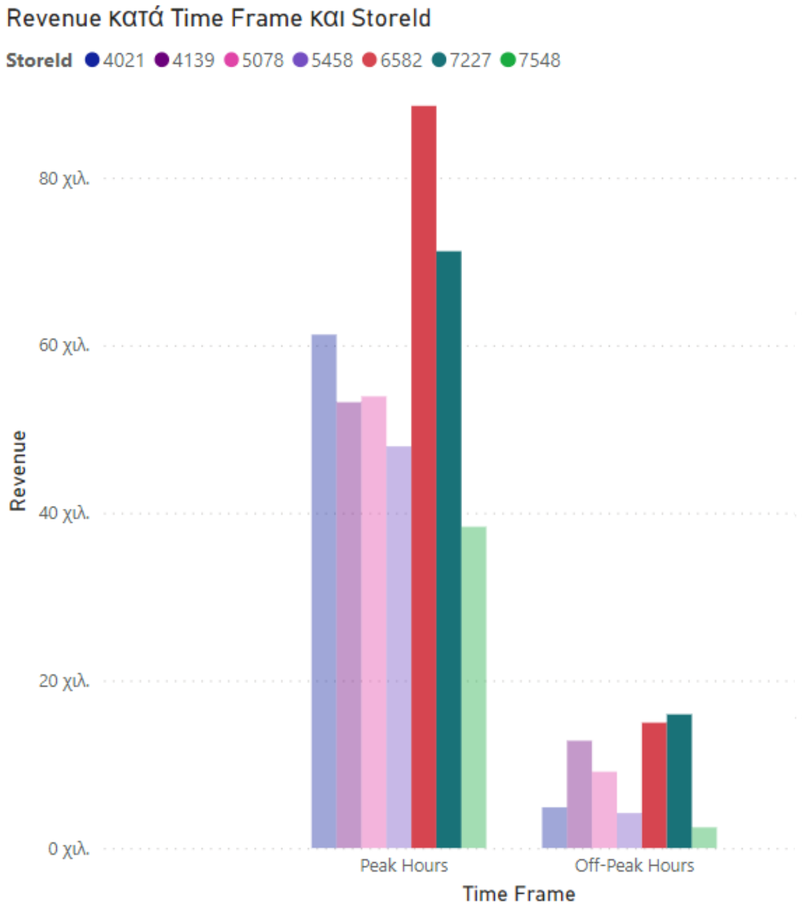


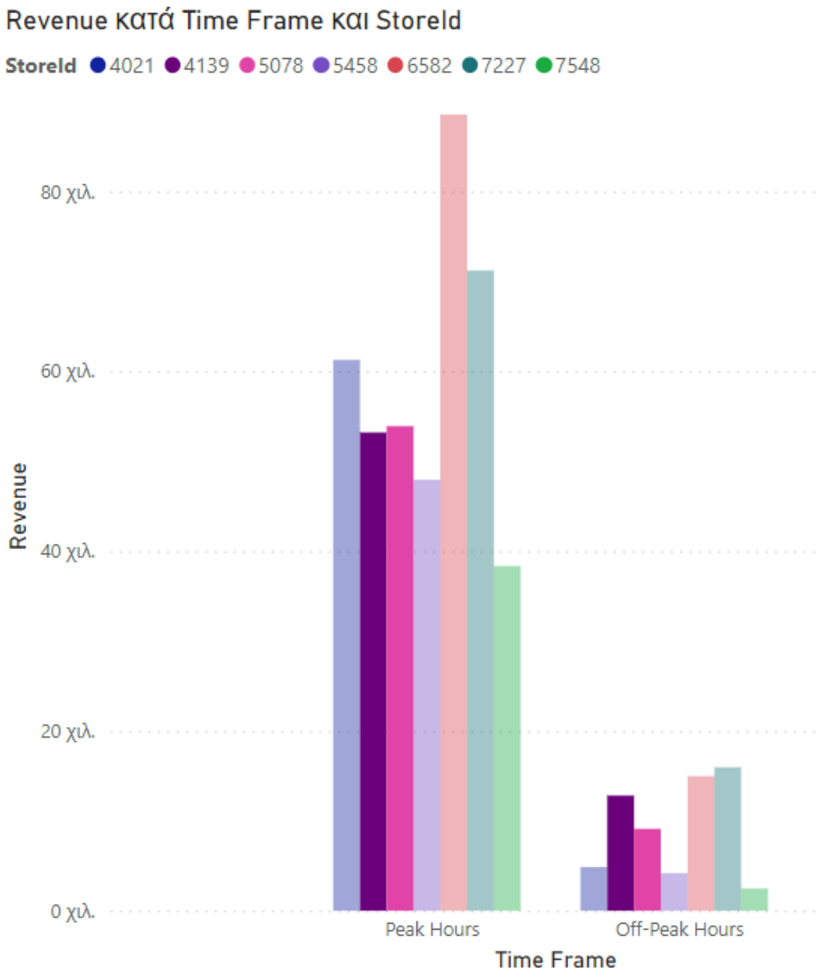


Some stores that perform better than others during peak hours don't do so on off-peak ones

Get a bigger share on off-peak hours/ make their opperational costs more effective in producing revenue

Shrink their oppening hours and/or their oppertional costs accordingly, or try to introduce promotional offers for that period





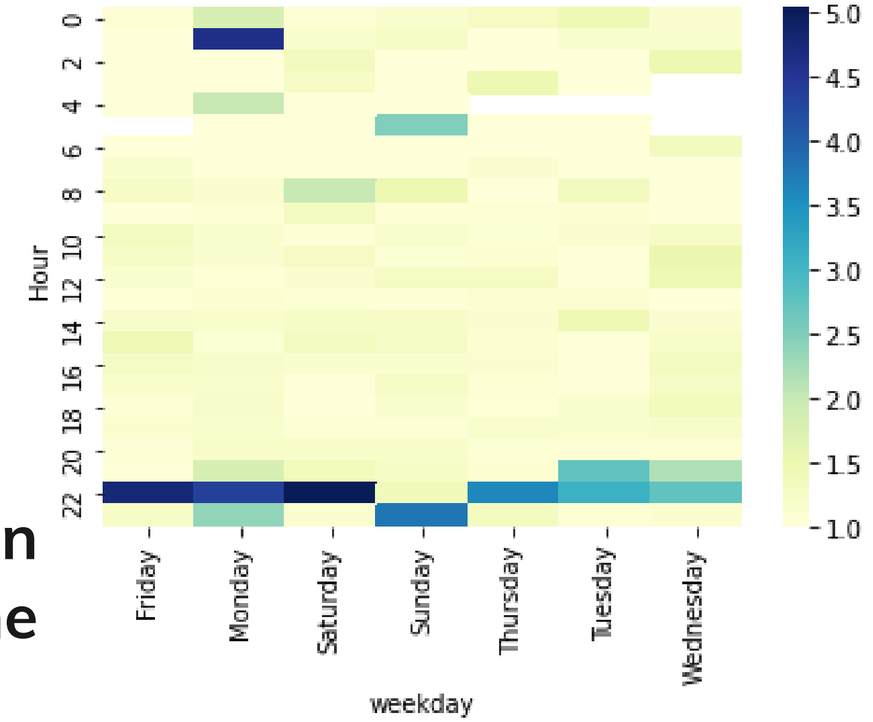
Dataset 2
Standard Kiosk

Most phone cards and cigarettes sales are concentrated around 22:00

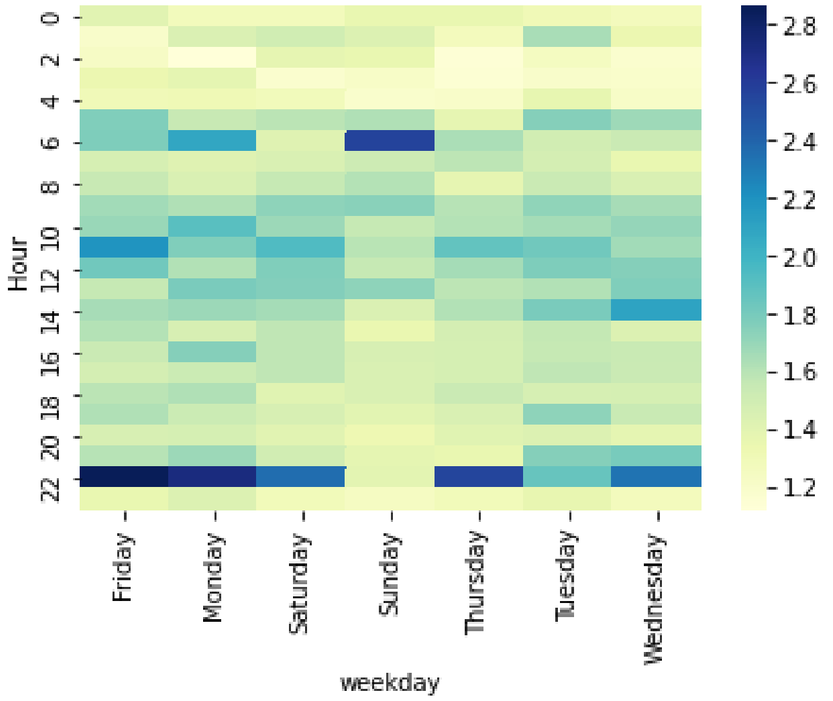
Suggestions:

- **Retailers should preload stock to target this time period**
- **Manufacturers should run media exactly at this time**

Καρτες Τηλεφωνίας

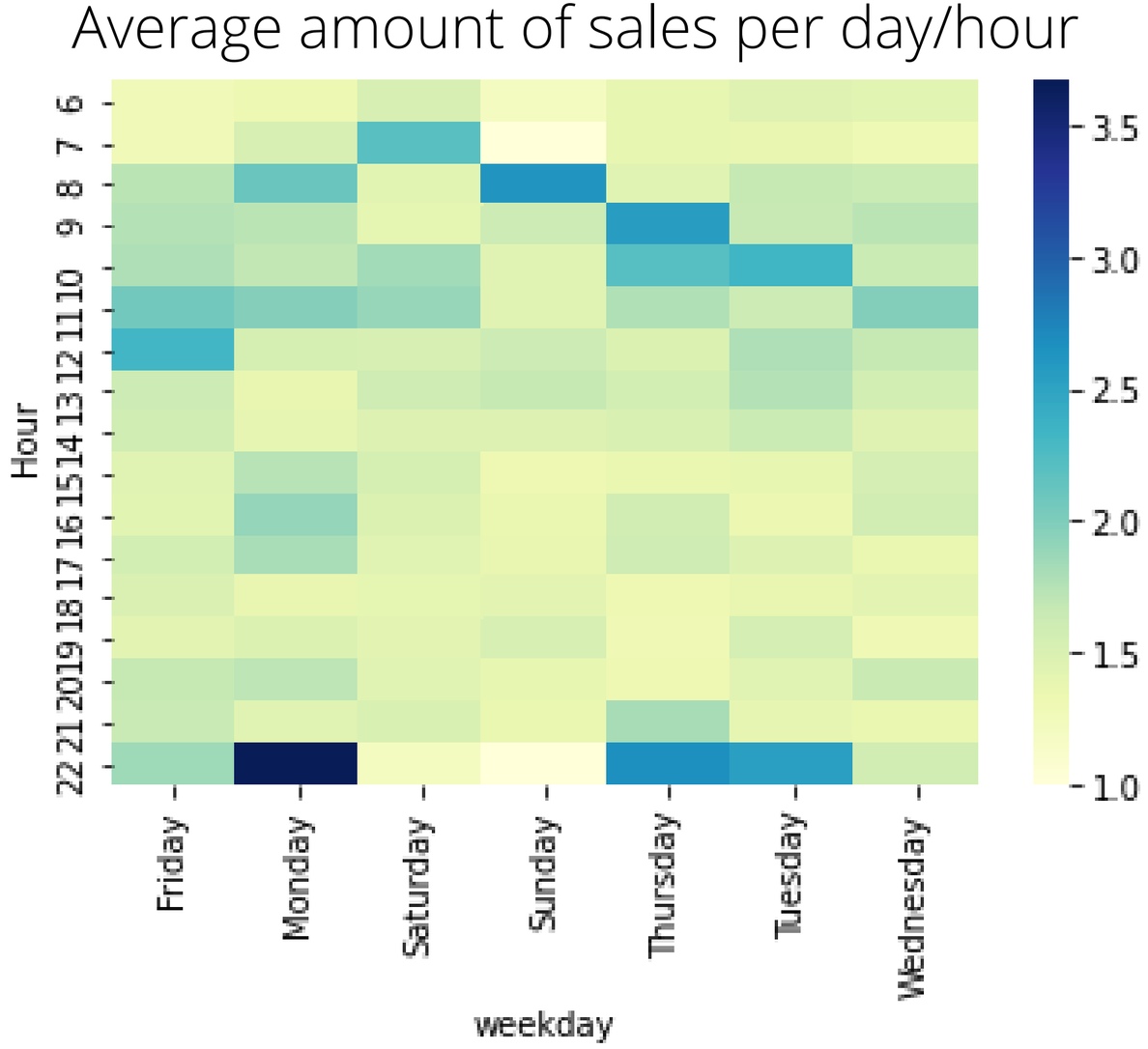


Τσιγαρα



Dataset 2
Provincial Mini Markets

The average amount of sales per hour is negligible. We would recommend considering alternative channels of sales such as delivery, in order to increase the amount of sales and revenue



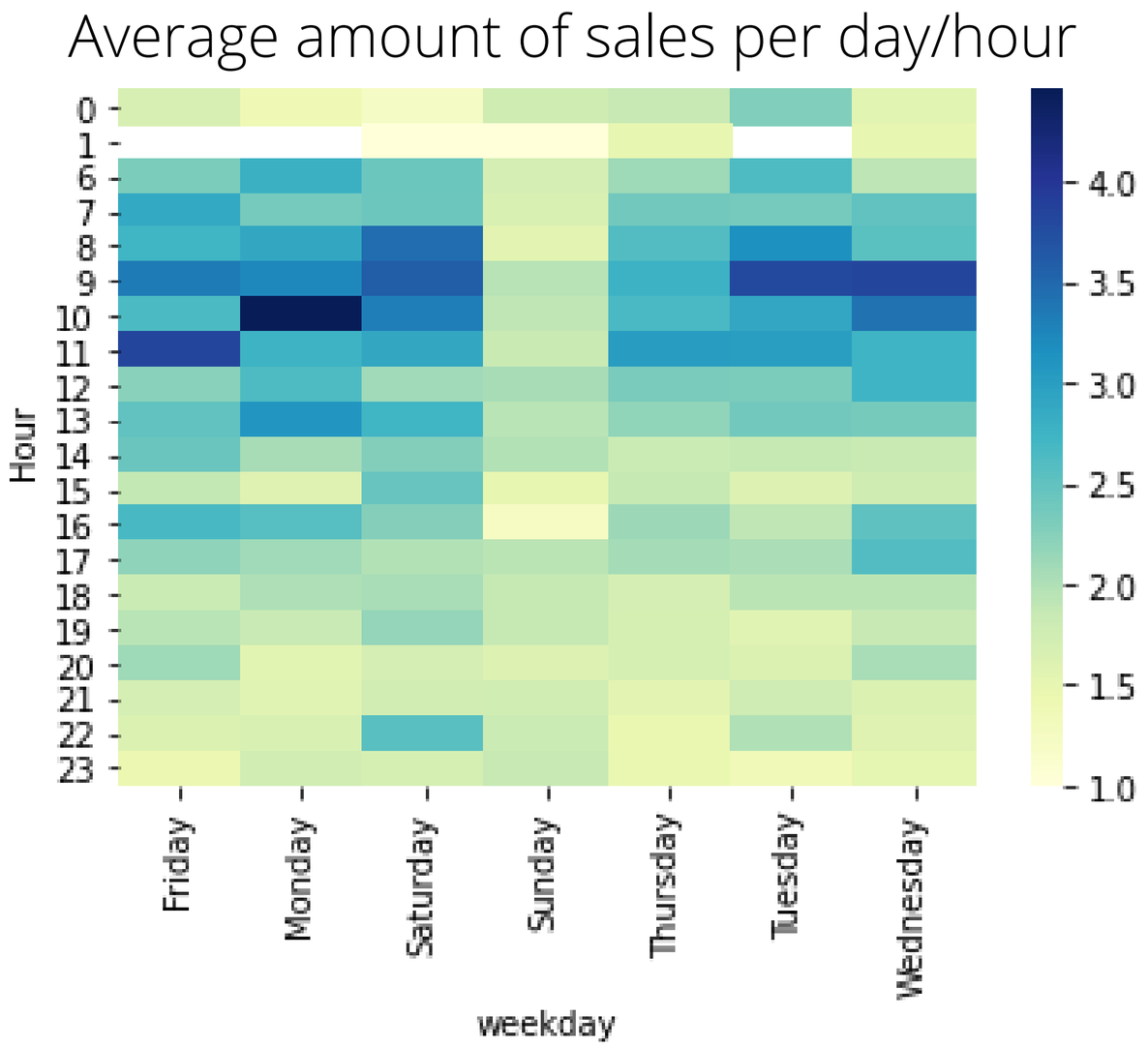
Dataset 2
Large Kiosk in Central Location

- **Most sales on working days and early hours probably before work.**
- **Almost 50% of sales are tobacco related products**

Suggestions:

- **Starbucks coffees**
- **Puff pastry/ pies such as Stergiou**

Try to catch some coffee and breakfast sales-products



Thank you!

Questions?

Technical Appendix

Dataset 1

Analyzed with Pandas

Clustering Basis – Page 5

Clustering Algorithm K-Prototypes , an algorithm for mixed variables numerical and categorical .

1. Store Type – String (Mini-Market , Kiosk)
2. Location - String 6 Geographical Locations
3. 4 Variables for 4 Seasons – Percentage of yearly sales that occurred on that season .
4. AVG Retail Price – Total revenue divided by total items sold
5. AVG Monthly Revenue – Grouped Sales Revenue by Month and calculated the average.
6. AVG Items Sold - Grouped Items sold by Month and calculated the average.
7. Categories Variety – Amount of different categories sold .
8. Categories Concentration – How many categories amount for 80% of revenue.

Cluster Analysis – Pages 5, 6, 7

The Metrics used in the presentation are calculated by labeling the stores (0,1,2) for their respective segments , and then calculating for each features metrics , grouping by those labels .

Ex.

```
Dataset2Cluster.groupby('Cluster')['Winter'].mean()
```

```
Cluster
0    0.235473
1    0.246038
2    0.222527
Name: Winter, dtype: float64
```

Cross-Selling on Tickets and Newspapers Proposition – Page 10

Created a binary table for each store with 2 columns valued 0 to 1 for the Tickets and Newspapers , if we had a record of sales for either with scored it with 1 if not we scored it with zero , we summed the 1’s grouping by segment .

Rural Mini Markets Proposition – Page 11

We isolated the 3 segments , and used the .corr command , the highest Pearson correlation came from the amount of categories a store sold and its monthly revenue for the rural mini markets , so we plotted both the categories and revenue sorted and came up with the diagram you see in page 11 .

Rural Kiosks Price Sensitivity – Page 12

Following the same method we created a df with index the stores id and features the avg price the corresponding store sold the products of each category , and we chose with the same metho corr , the most highly correlated ones .

Lockdown effect on segments – Page 13

For the top 5 selling categories (Cigarettes , Soft Drinks , Alcohol , Phone Cards , and Tobacco Products) , we summed the revenue for a 3 month period where covid restrictions were active (Dec 2020 to Feb 2021) , and we compared it to the sales of the year before for exactly the same months (Dec 2019 to Feb 2020) .

Code snippet

```
def findLockdownChanges(categories, storeids, cluster):
    data = readExcel()
    lockdown_revenues = {str(categories[0]): 0, str(categories[1]): 0, str( # initialize category revenue sums per year
        categories[2]): 0, str(categories[3]): 0, str(categories[4]): 0}
    free_revenues = {str(categories[0]): 0, str(categories[1]): 0, str(
        categories[2]): 0, str(categories[3]): 0, str(categories[4]): 0}
    for iD in storeids:
        for row in data.iterrows():
            for category in categories:
                rowObj = row[1].to_dict() # get relevant fields
                year = rowObj['Date_Year']
                month = rowObj['Date_MonthName']
                id = rowObj['StoreId']
                cat = rowObj['CategoryId']
                if(str(id) == str(iD)):
                    if(str(category) == str(cat)): #sum revenue to the appropriate free or lockdown revenue per category
                        if(str(year) == '2019'):
                            revenue = rowObj['Revenue']
                            free_revenues[str(category)] += revenue*(1/13)
                        if(str(year) == '2020'):
                            if(str(month) == "December"):
                                revenue = rowObj['Revenue']
                                lockdown_revenues[str(category)] += revenue
                            elif(str(month) == "January" or str(month) == "February"):
                                revenue = rowObj['Revenue']
                                free_revenues[str(category)] += revenue
                        elif(str(year) == '2021'):
                            if(str(month) == "January" or str(month) == "February"):
                                revenue = rowObj['Revenue']
                                lockdown_revenues[str(category)] += revenue

    print("Cluster" + str(cluster))
```

Segments Growth – Page 14

Grouped by with the help of pandas by the year (2020, 2021) and by segment

(0, 1, 2)

So we got 6 results, the revenue for the 3 segments for the 2 year's and we divided them to get the change in the revenue.

Dataset 2

Analyzed with RM, PowerBI & Pandas

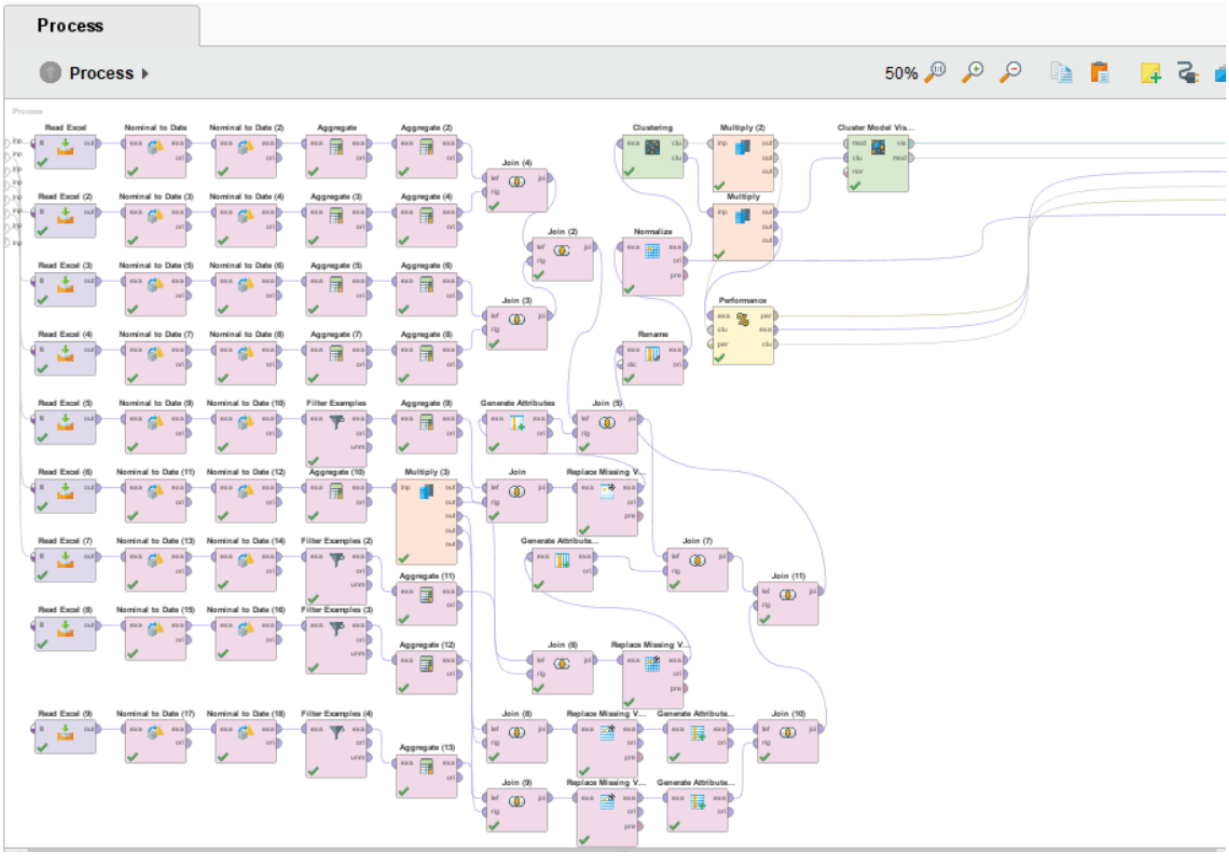
Clustering Basis for Store Segmentation – Page 18

K-Means Clustering, using RapidMiner, taking as variables (normalized):

1. Average Daily Transactions
2. Average Revenue/Transactions
3. Average No of Categories per Transaction
4. Average Items per Transaction
5. Weekend Revenue (as percentage of Total Revenue) – 2 variables, one for weekdays and one for weekends
6. Off-Peak (7:00-23:00) Revenue (as Percentage of Total Revenue) – 2 variables, one for peak and one for off-peak hours

Row No.	StoreId	AVG daily tra...	AVG revenue/...	AVG cat...	AVG products...	Off Peak...	Peak Hours...	Weekend...	Weekdays ...
1	3788	46.452	4.283	1.268	1.692	0	1	0.016	0.984
2	4021	371.097	5.748	1.366	2.071	0.073	0.927	0.033	0.967
3	4097	151.065	8.036	1.254	2.093	0	1	0.021	0.979
4	4139	426.387	4.994	1.249	1.776	0.194	0.806	0.038	0.962
5	5078	364.129	5.582	1.306	1.811	0.145	0.855	0.029	0.971
6	5458	325.645	5.162	1.236	1.813	0.080	0.920	0.042	0.958
7	6463	235.548	11.232	1.710	3.474	0.045	0.955	0.029	0.971
8	6582	554	6.028	1.359	2.105	0.145	0.855	0.028	0.972
9	7227	448.129	6.275	1.083	2.245	0.183	0.817	0.035	0.965
10	7548	232.387	5.666	1.420	2.111	0.061	0.939	0.036	0.964

Fact table (normalized) for the k-Means algorithm



Each “line”/”stream” of operators creates each one of the variables used in the clustering process

Store Segmentation Clusters – Pages 19, 20, 21

We chose the gravest characteristics of its cluster after the RM process bellow was run.

Heat map underlying the most important differences between clusters

Segments Strategy – Page 26

Based on a process created on RM that identifies association rules for each of the three clusters, we used the *lift* metric to spot itemsets whose correlation is indifferent to the popularity of the individual products.

Show rules matching

all of these conclusions:

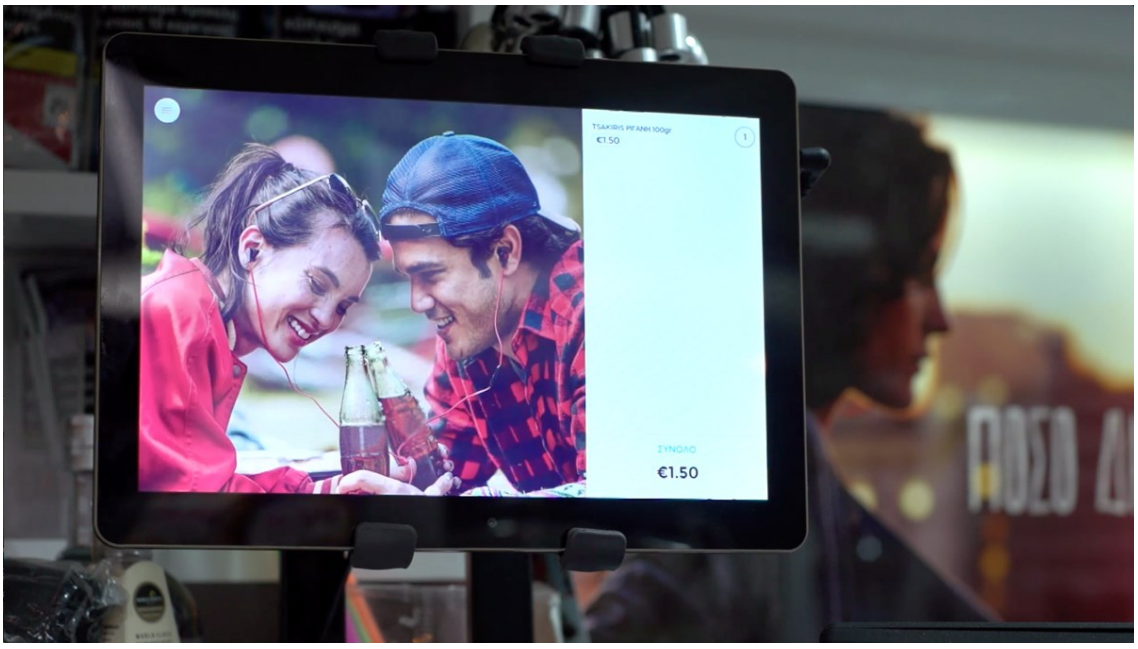
count(Quantity)_28
count(Quantity)_27
count(Quantity)_115
count(Quantity)_36
count(Quantity)_47
count(Quantity)_34
count(Quantity)_17
count(Quantity)_25
count(Quantity)_106

No.	Premises	Conclusion	Lift ↓	Support	Confidence	LaPlace	Gain
20	count(Quantity)_115	count(Quantity)_36	2.062	0.041	0.219	0.876	-0.337
29	count(Quantity)_36	count(Quantity)_115	2.062	0.041	0.390	0.941	-0.171
14	count(Quantity)_115	count(Quantity)_25	1.889	0.023	0.120	0.860	-0.355
27	count(Quantity)_25	count(Quantity)_115	1.889	0.023	0.357	0.962	-0.105
17	count(Quantity)_115	count(Quantity)_34	1.878	0.028	0.146	0.864	-0.350
26	count(Quantity)_34	count(Quantity)_115	1.878	0.028	0.355	0.953	-0.128
13	count(Quantity)_115	count(Quantity)_47	1.339	0.023	0.120	0.860	-0.355
23	count(Quantity)_47	count(Quantity)_115	1.339	0.023	0.253	0.939	-0.156
28	count(Quantity)_28, count(Quantity)_115	count(Quantity)_27	1.296	0.029	0.364	0.953	-0.132

One of the most powerful associations, was that between products from Category 36 and 115 (Snacks and Refreshments). *Lift* values greater than one show a positive relationship between the items.

Promotion of a specific product bundle – Pages 27, 28

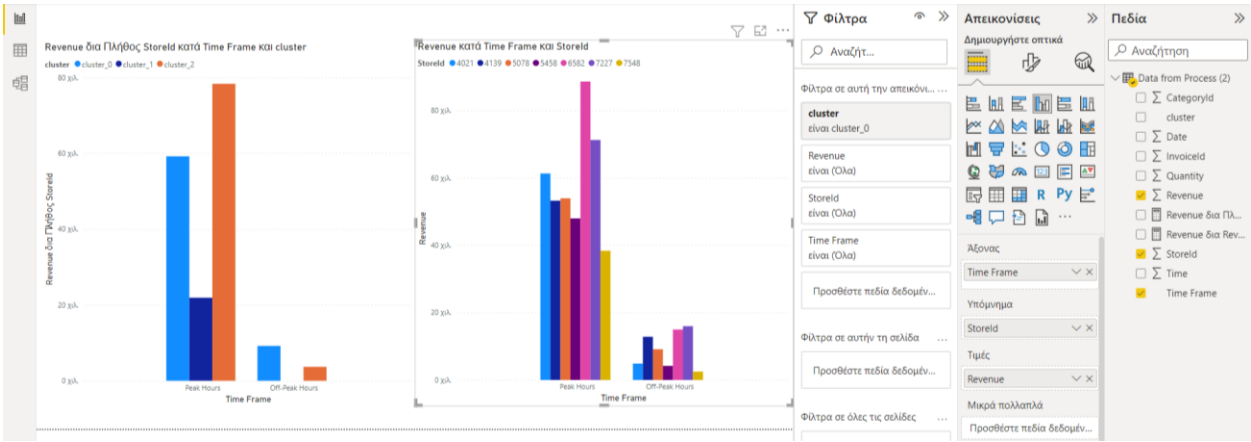
The above process was repeated for each StoreID and indicated that some stores miss on the opportunity to exploit the strong correlation between the purchase of snacks and beverages. Thus we proposed some promotional actions for the bundle of the aforementioned products.



Advertisements displayed on screens next to the cash desk could be put to good use in promoting the product(s) believed to be able to make bigger sales.

Data visualization using PowerBI – Pages 29, 30

PowerBI was used to visualize data concerning the clustered stores. The purpose of the analysis was to see how each store's performance (in terms of sales made) during off-peak hours differentiates. What the results of the process suggested, was that some stores have more room for improvement than others of the during that time frame (23:00 – 7:00) and should consider expanding their hours, while others should cut off unnecessary operational costs or even shrink their opening hours.



Sales during peak and off-peak hours for all stores – right chart – as well as the median of each cluster – left chart.

Strategy Propositions on Segments – Pages 31, 32, 33

For each segment we grouped by time and day , and created a pivot table that showed for each day and hour the avg sales . Ex.

weekday	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
Hour							
6	1.280000	1.333333	1.526316	1.214286	1.387097	1.466667	1.434783
7	1.275862	1.523810	2.200000	1.000000	1.388889	1.380952	1.291667
8	1.719298	2.122449	1.428571	2.628571	1.437500	1.676471	1.634921
9	1.750000	1.720000	1.410959	1.617021	2.553571	1.659574	1.728814
10	1.786667	1.697368	1.838095	1.456790	2.218182	2.333333	1.625000
11	2.072464	1.971429	1.894737	1.439394	1.771429	1.620690	1.986667
12	2.324074	1.551724	1.530435	1.619048	1.494845	1.782609	1.670886
13	1.623762	1.379310	1.607143	1.675676	1.569231	1.758065	1.566265
14	1.588235	1.400000	1.484211	1.486111	1.517241	1.632653	1.461538
15	1.452830	1.741379	1.541667	1.315789	1.367347	1.387755	1.559322
16	1.434783	1.900000	1.489796	1.362069	1.574468	1.352941	1.576923
17	1.568627	1.809524	1.451220	1.378049	1.609375	1.476190	1.370968
18	1.508475	1.375000	1.415663	1.418803	1.316456	1.376344	1.419355
19	1.423841	1.494118	1.417647	1.525424	1.308411	1.561798	1.308411
20	1.671642	1.718310	1.455882	1.390476	1.312500	1.457831	1.636364
21	1.650602	1.457143	1.516484	1.360656	1.813953	1.396552	1.372549
22	1.857143	3.666667	1.200000	1.000000	2.666667	2.533333	1.583333

Finally we used the above table and the ones we created for the other segments to create heatmaps so we could find where the sales were most concentrated .