B2W Pricing Challenge 2016

André Teixeira dos Santos
30 / 09 / 2016

OBJECTIVES

“**predict** the **quantity sold** for a each product given a **prescribed price**”

"we need metrics, relationships and descriptions of these data in order to understand the sales behavior. **What does the data tell us?** How are the different data sources related?”

“what were the **steps and your strategy** (approach to the problem)”

“ - Show a understanding of **SQL**;
- Use **Version Control** (Git for example);
- Show methods for **clustering;**”

CONTENTS

OVERVIEW

TECHNOLOGIES & APPROACH

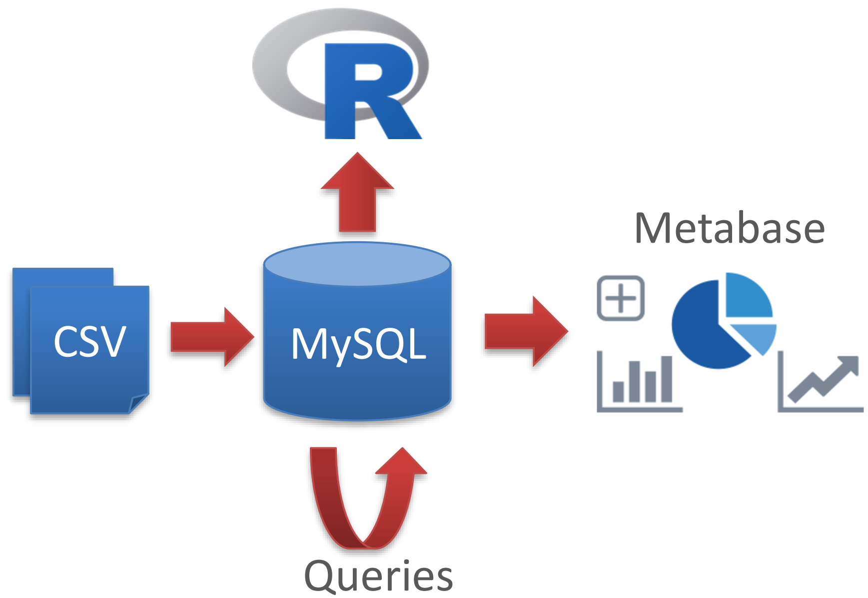
Load: CSVs checked and load at MySQL.

Data Preparation in MySQL: Indexing, renaming, horizontal enhancements (e.g., day of week), dimensional aggregations (e.g., summary of a day).

Basic Visualizations & Analysis: basic analysis using Metabase summarizations and time series views, averages of prices and volumes. R histograms, boxplots & scatterplots.

Analytics & Insights: development of indicators, creation of more aggregated table and summaries, compilation of insights throught MySQL and Metabase.

Modeling: development of models that tries do predict volume of sales for a given product, using avaiable information (not only price). Naive forecasting method used has benchmark.



PROCESS OVERVIEW

- basic data preparation
- basic analytics
- first modeling attempt: linear regression (cross-sectional predictors)
- second modeling attempt: vector autoregression (VAR & SVAR)
- advanced data preparation & analytics
- third modeling attempt: dynamic models (cross-sectional predictors + lagged and differentiated predictors)
- fourth modeling attempt: clustering + dynamic models
- consolidation of results & presentation

PREMISSES AND SIMPLIFICATIONS

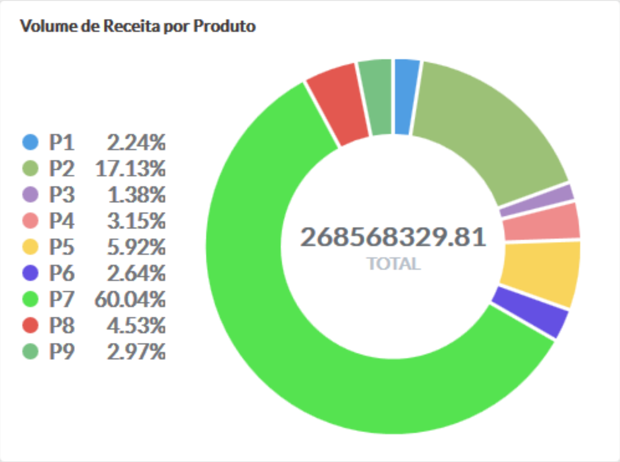
- the nominal/base price for a product, for a given day, was considered as the **maximum price** for that day
- **outliers** were not treated
- **inner joins** were used to combine sales and competitors prices to avoid dealing with **missing data**, resulting in a decreased dataset
- the models were developed **for 1 product (P2)** and inter-products effects were not considered (e.g., substitute goods), i.e., it was **assumed products sales were not negatively correlated**
- neural networks and other **non-linear models were not considered** due to increase in complexity and labor, and loss in interpretability.
- for all regression models the **mean absolute error** (MAE) was used to evaluated the model performance and for **benchmarking was used the MAE of naive and mean forecasting methods**. Other performance metrics like R^2 were not considered.

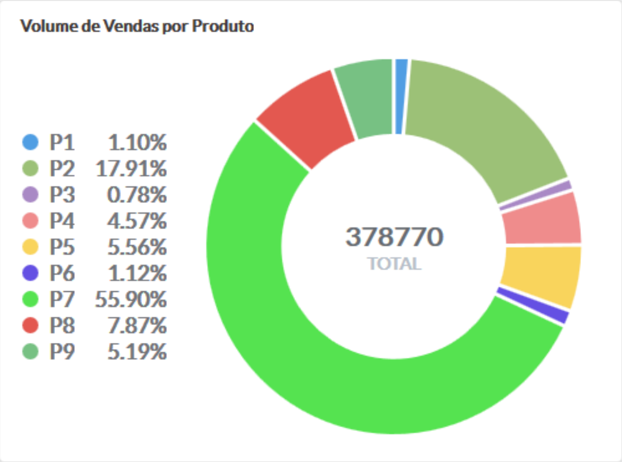
REMARKS ABOUT THE DATA

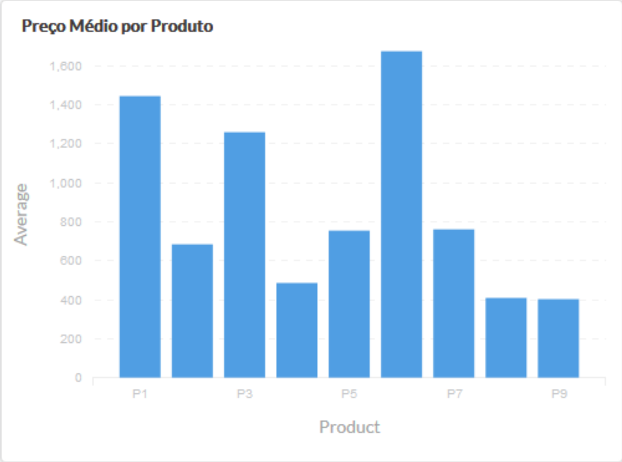
- some prices from the competitors monitoring csv looked like were multiplied by 10, so it was divided back in the staging area.
- each product had a different temporal window, being for sale or competitor monitoring.
- the claim that the price was captured twice a day was not precise, i.e., it could be more or less than that and also, it could be at the same time, which has no use.
- price is not a good predictor for volume, as the requirements led to believe.

BASIC ANALYTICS

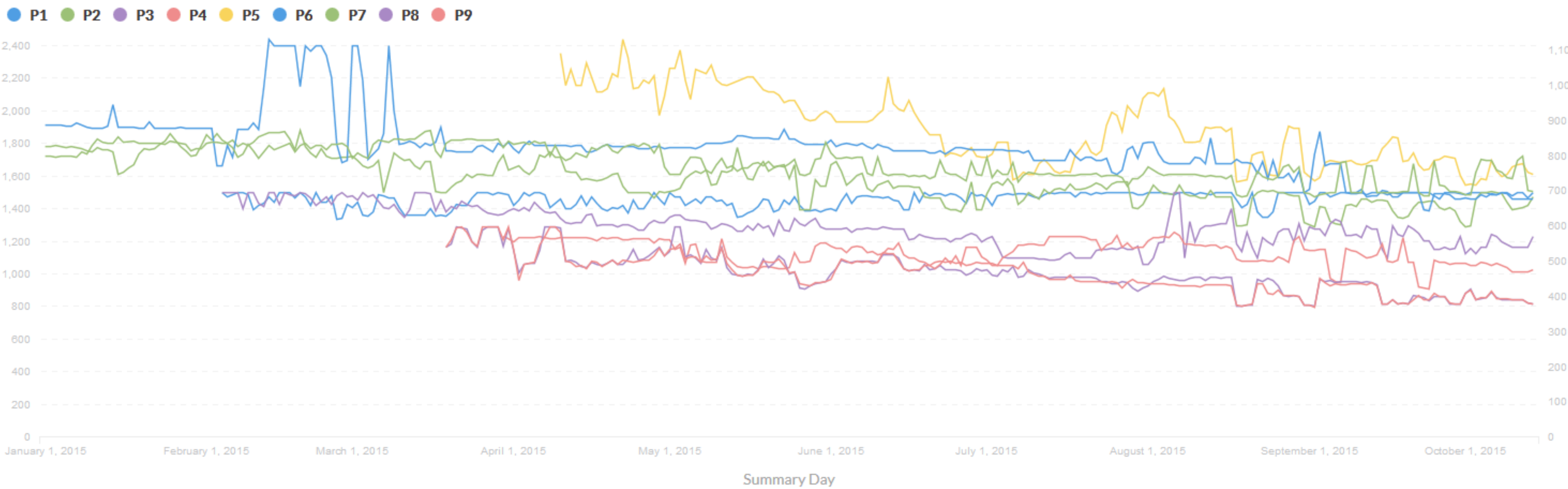
BASIC ANALYTICS – AGGREGATIONS





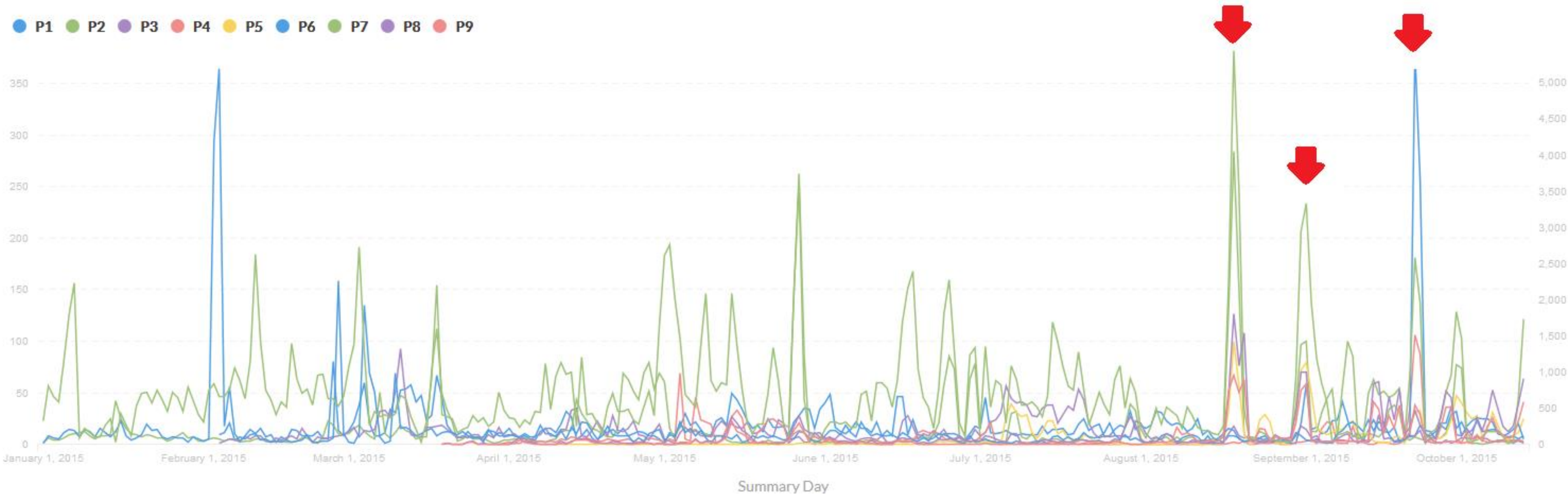


BASIC ANALYTICS – PRICE TIME SERIES



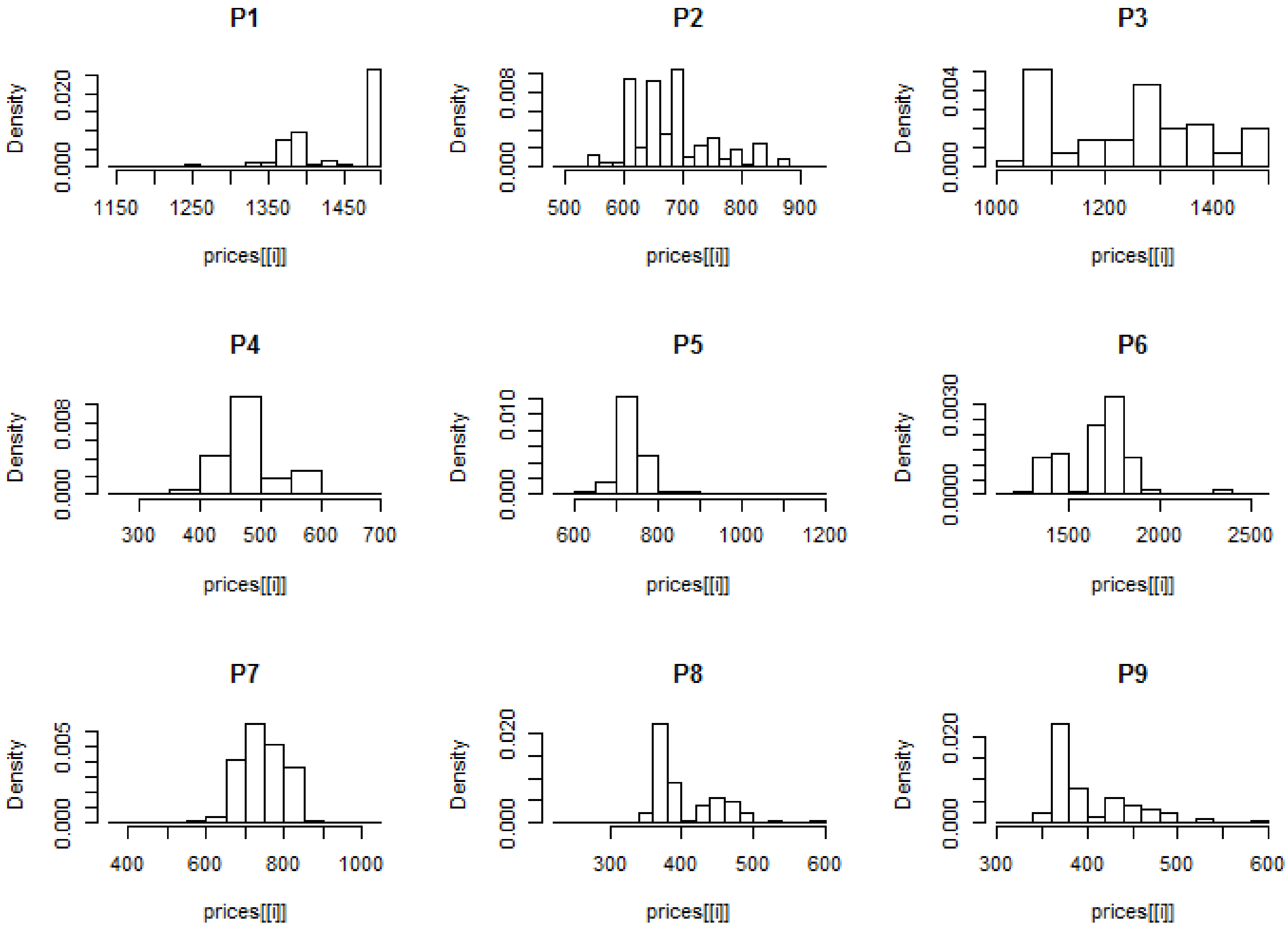
- Averege price in day
- Distinct windows of time
- Looks like white noise

BASIC ANALYTICS – VOLUME TIME SERIES

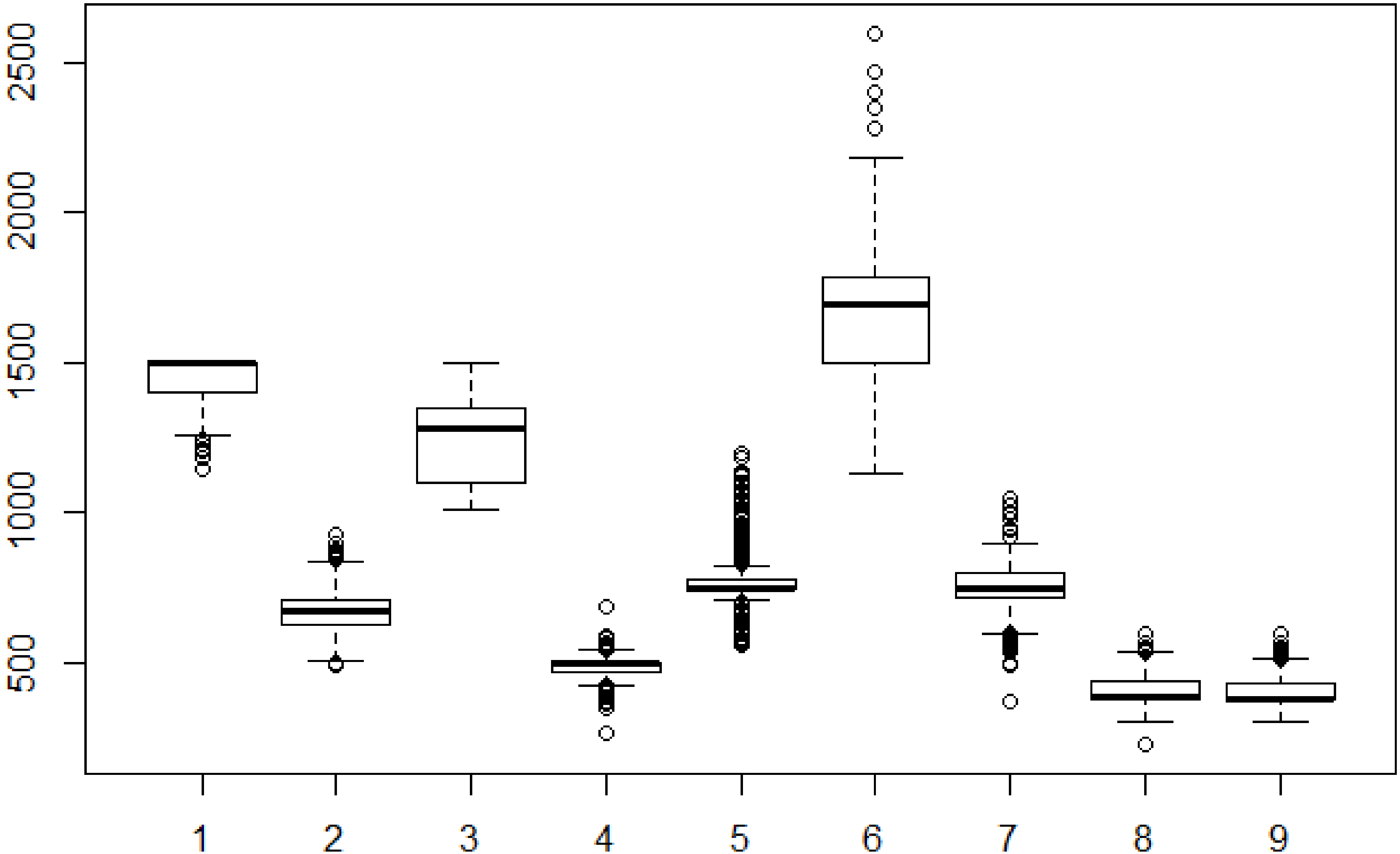


- Considerable oscillations
- Looks like that are market spikes for all products (global campaigns or intrinsic market seasonality)
- Average inter-product positive correlation

BASIC ANALYTICS – PRICES HISTOGRAM



BASIC ANALYTICS – PRICES BOXPLOTS



PRICING ANALYTICS

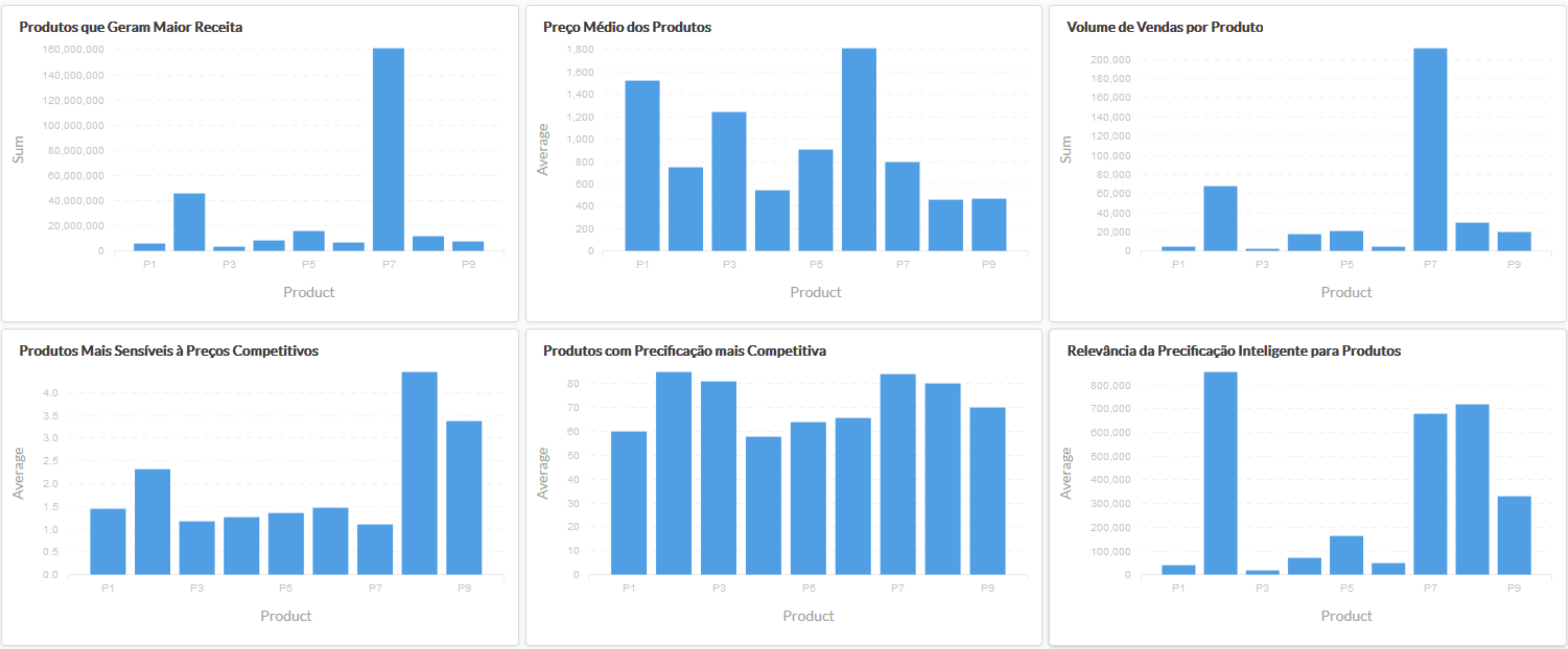
ANALYTICS – PRICING INDICATORS

pricing efficiency: how well is the competitor able to keep its prices bellow mine, for each product?

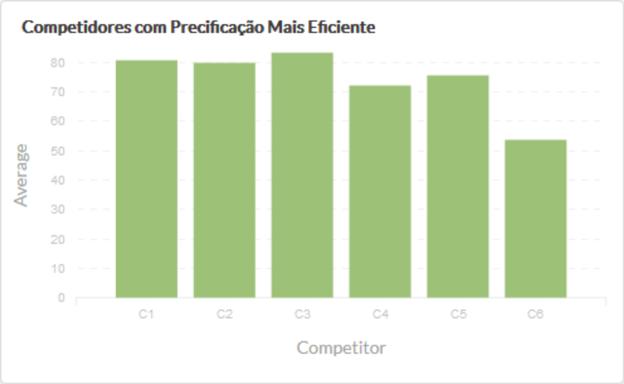
pricing influence: how does the competitors prices impacts my sales, for each product, i.e., when he has lower prices, does my sales volume decreases?

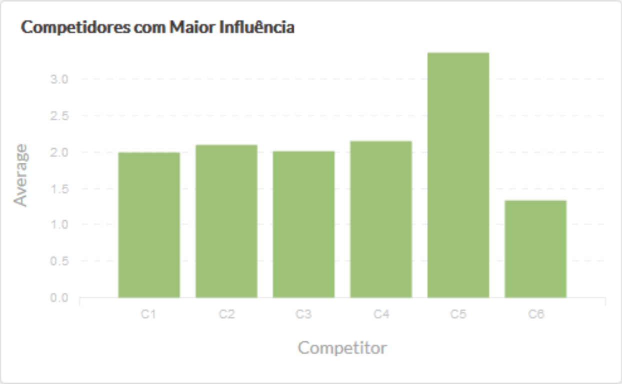
pricing relevancy: what products should I focus on, to give a more intelligent pricing strategy, i.e., what products represents a greater revenue for me and are more sensible to competitive prices?

ANALYTICS – PRODUCTS INDICATORS

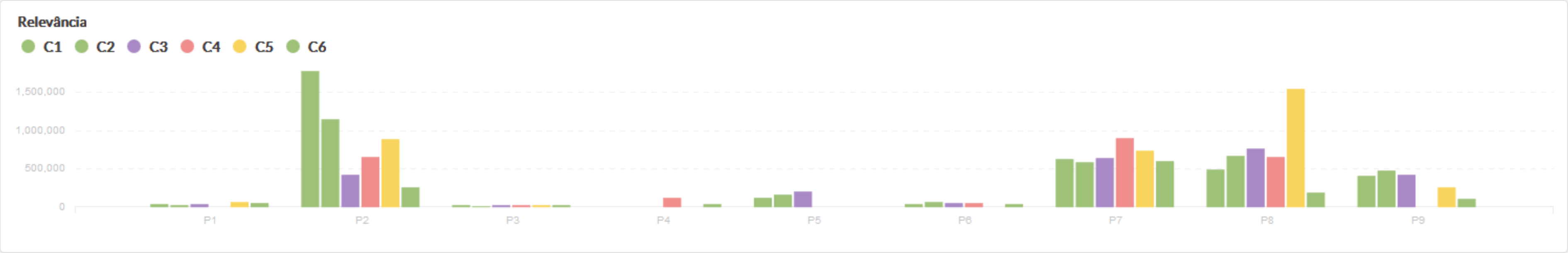


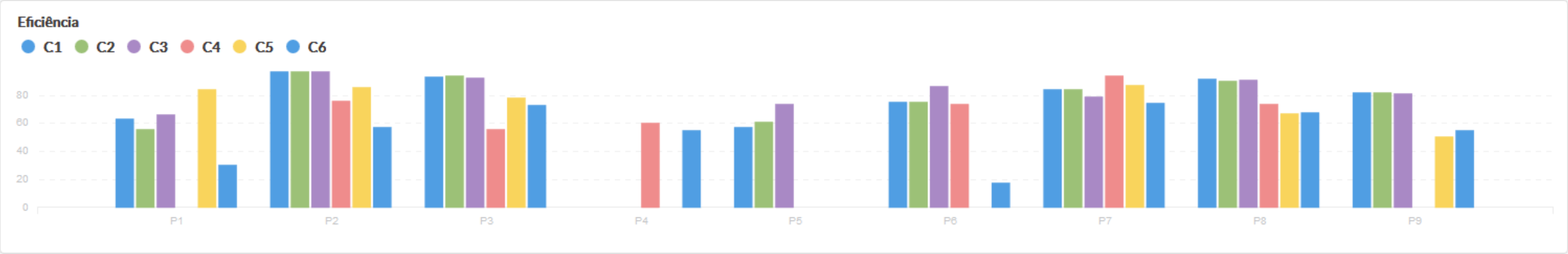
ANALYTICS – COMPETITORS INDICATORS

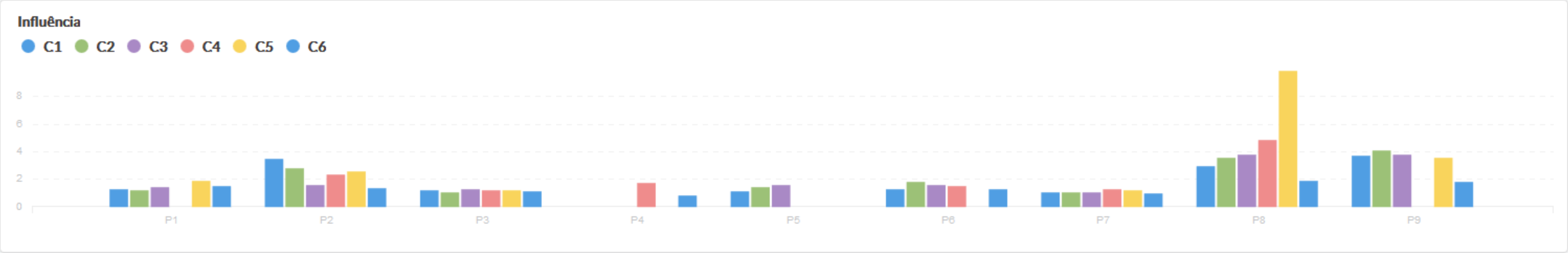




ANALYTICS – GRAIN INDICATORS





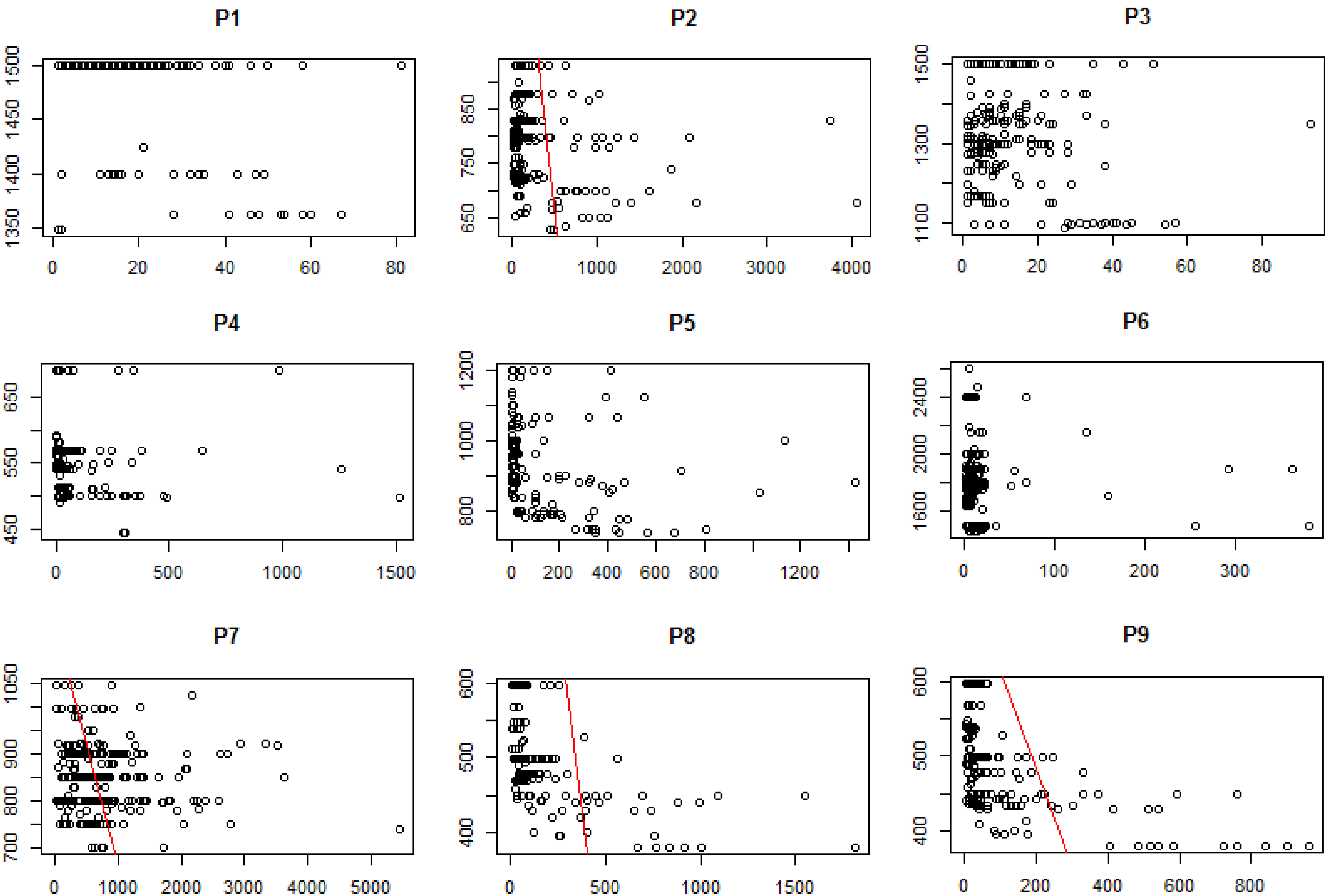


ANALYTICS – SOME INSIGHTS

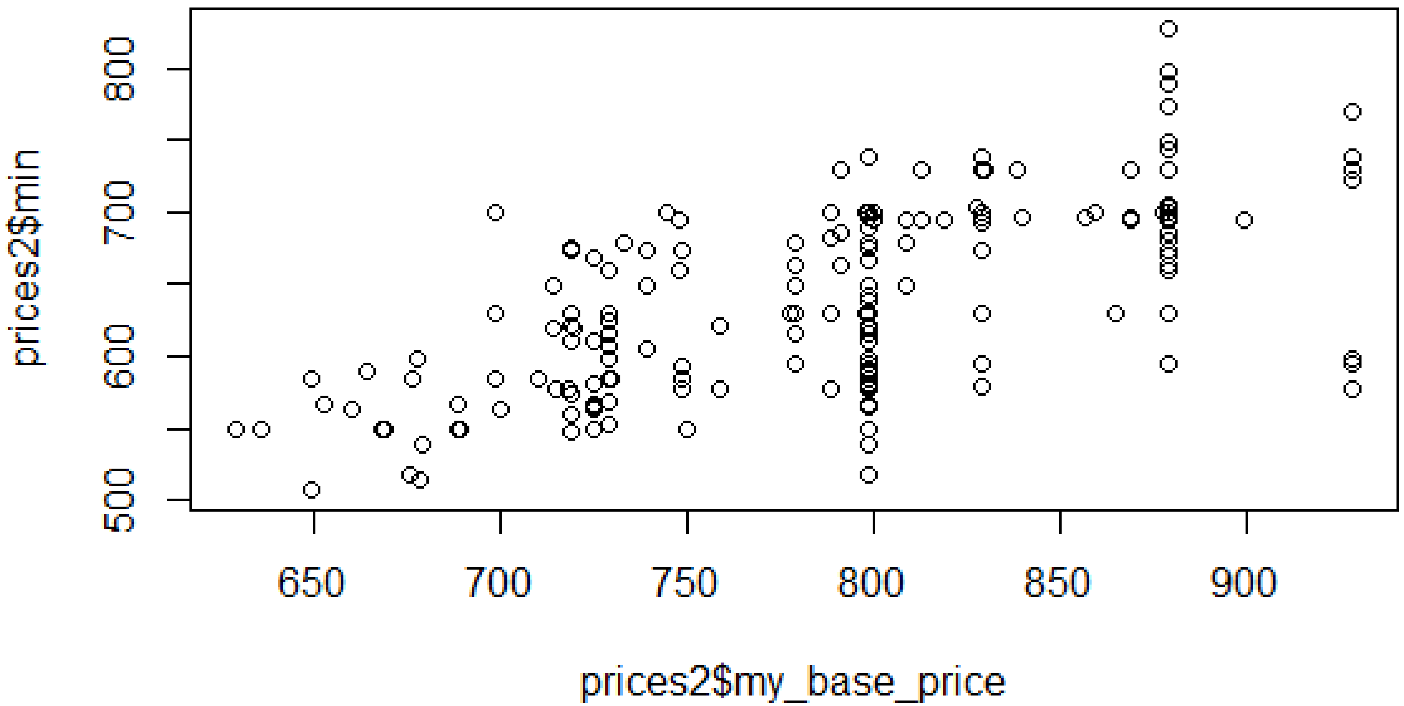
- C5 is the main competitor
- P7 is the main product
- P2, P8 and P7 are the products who mostly require smart pricing
- Despite its prices, P2 is preferred by consumers to be bought at competitors
- It’s probable that consumers looks at C5 for P2 prices first.
- C6 is niche, i.e., consumers buy there no matter the price.

MODELING

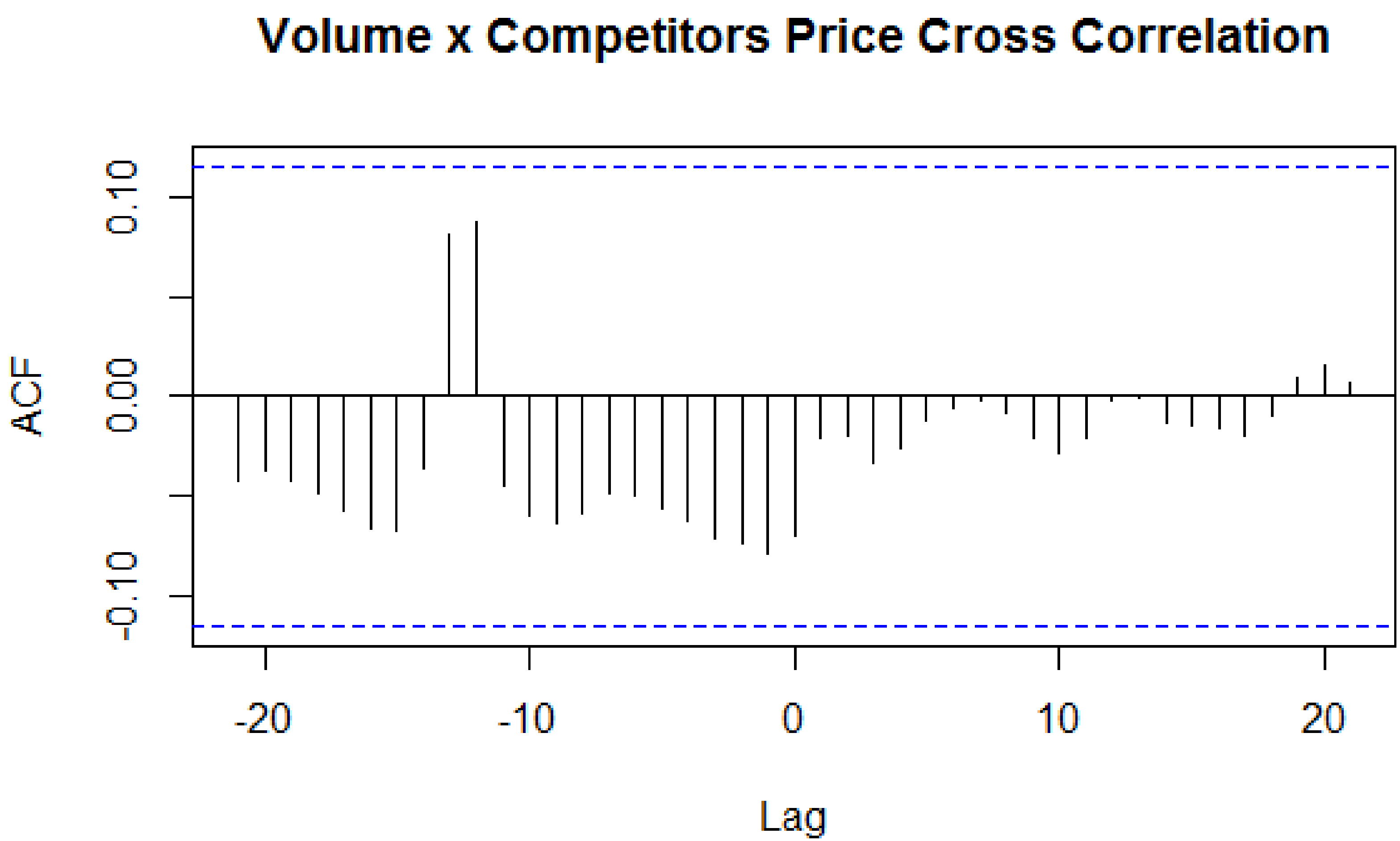
MODELING – VOLUME x PRICE SCATTERPLOTS



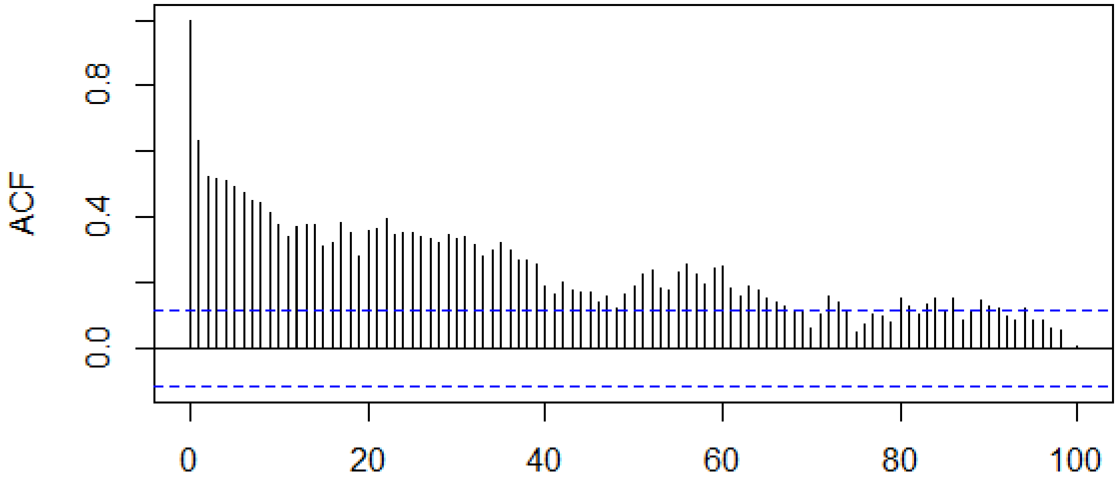
MODELING – PRICE x COMPETITORS PRICE



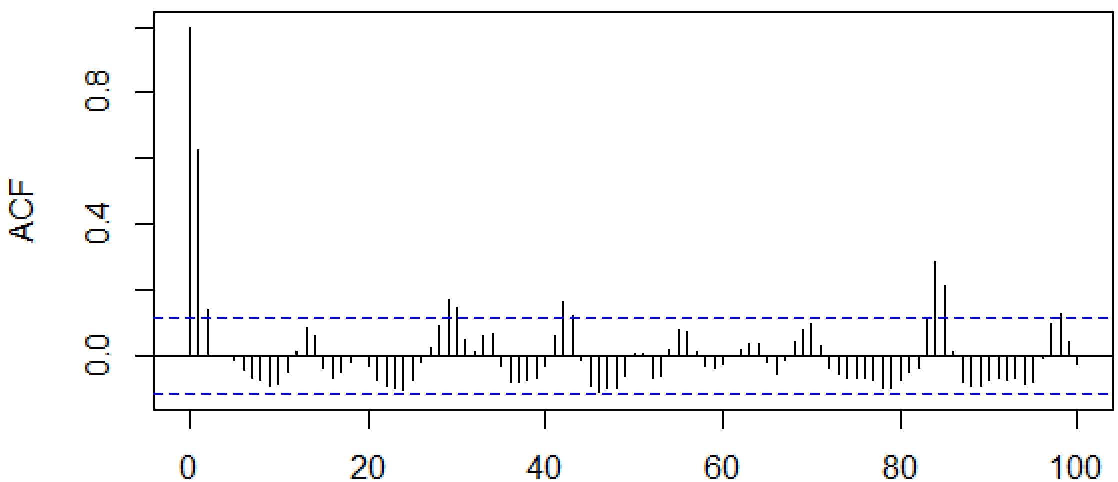
MODELING – VOLUME x COMPETITORS PRICE



MODELING – AUTOCORRELATION



PRICE



VOLUME

MODELING – STRATEGIES & APPROACHES

1. Linear Regression
   - Price alone isn’t good predictors
   - Use other cross-sectional predictors, e.g., day of week, competitors price, day of month
   - Mean absolute error (MAE) didn’t improve against benchmark (naive forecasting method)

2. Vector Autoregression (VAR & SVAR)
   - Use lagged values of volume, price and competitors prices time series to predict volume
   - VAR didn’t attend test requirements and SVAR would be too complex.

3. Dynamic Models
   - Manually combine cross-sectional data with lagged and differentiated value from multiple time series into a linear regression
   - MAE didn’t improve against benchmark

4. **Clustering + Dynamic Models**
   - Predictor hidden somewhere
   - “Show mtehods for clustering”
   - Clustering the days of the year (days that have a particular effect on volume behaviour)
   - Kmeans with 6 centers chosen (based on whitin cluster sum of squares evolution)
   - Improvement of 75% against benchmark!
   - Predictors used: price, cluster, competitors minimum price, volume of the day before (lag 1)

MODELING – RESULTS FOR **P2**

Average volume: **291 items per day**

Mean absolute error for benchmark (naive): **200 items**

70% of average volume

Mean absolute **error** for model with clustering: **46 items**

16% of average volume

75% decreased error against benchmark!

END