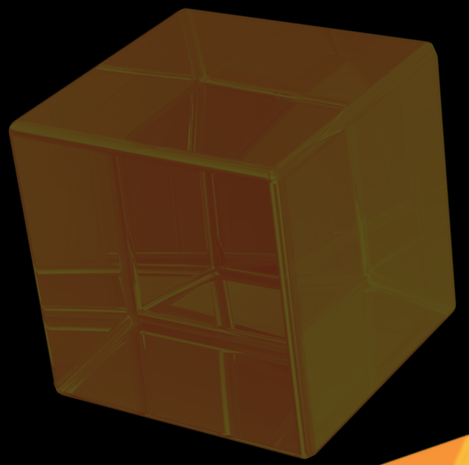
BIG DATA AND SMART DATA ANALYTICS

Alkemy Project

"The idea behind this business case is that there isn't a yes/no answer: there are many perspectives to visualise the problem." - Alkemy



Group O | Cosimo Poccianti | Olimpia Sannucci | Miro Confalone | Carlo Ardito

Goal

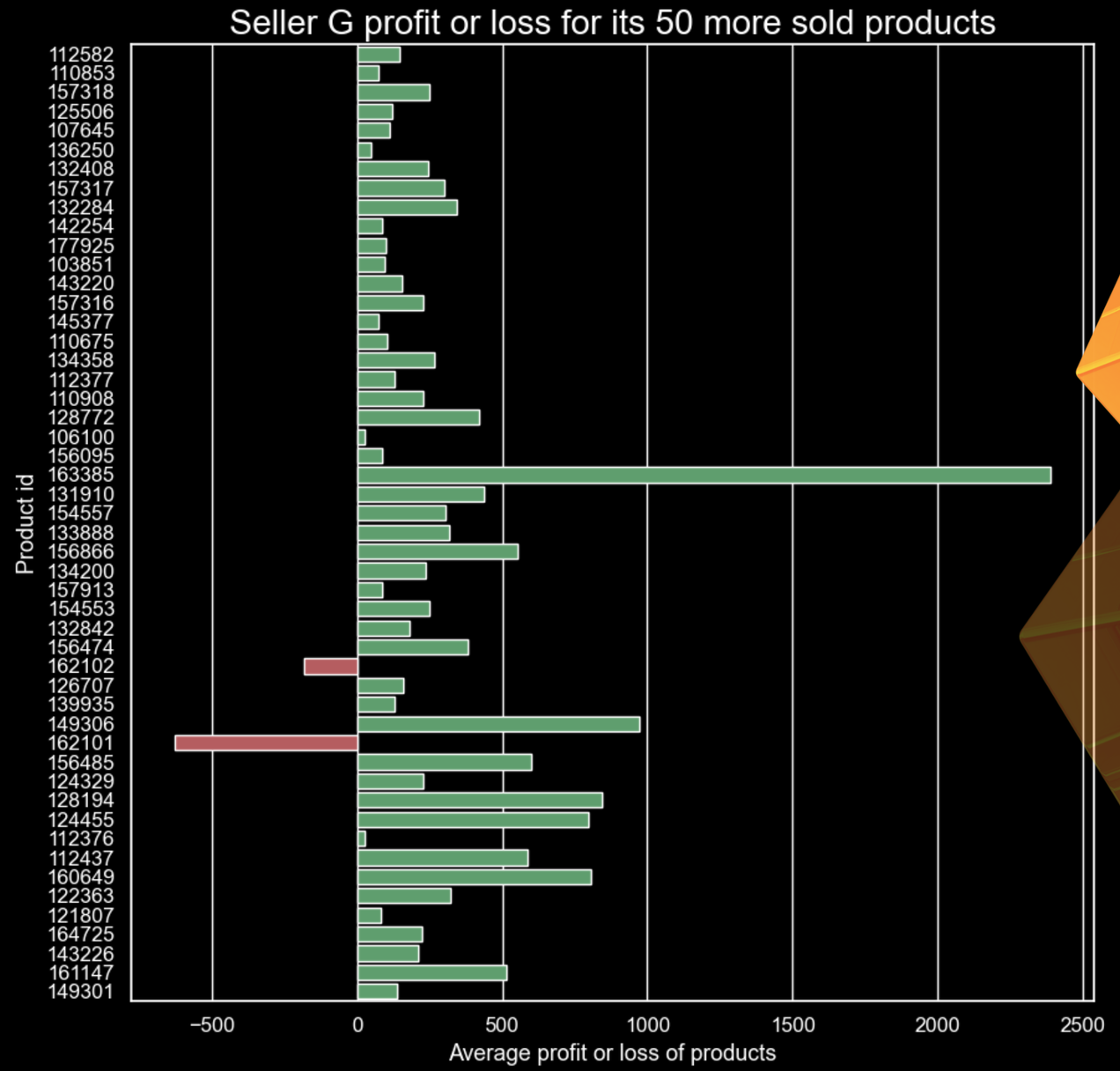
Empower managment in decision making processes
to optimise the company pricing strategy.

Methodology

- Data Familiarization
- Features Extraction
- Market dynamics analysis
  - VAR Model
  - Granger Causality model
  - November analysis
- Popularity Index
  - Our Dashboard

Data Familiarization

As suggested, we have familiarized with the datasets by taking into consideration different approaches and points of view.



Feature Extraction

Task 1

Where possible, data were merged to get a better understanding of the information available

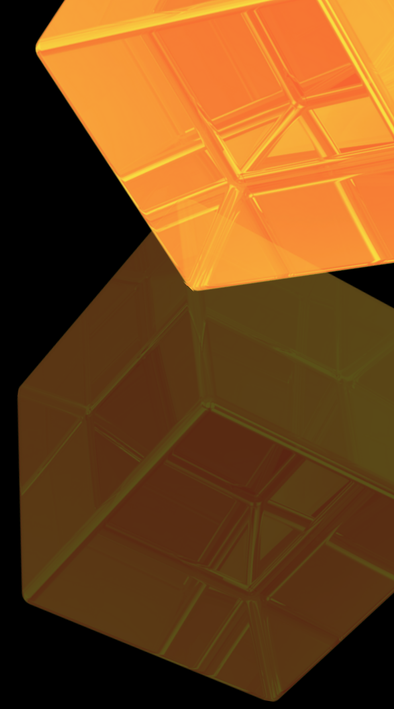
From price_competitor.csv

- Minimum and maximum price for each product
- Logarithmic price change of each product, for each seller
- The variance of the logarithmic price change, on each quarter

Using Apache Spark

Feature Extraction

Task 2



From sales_data.csv

Grouping on product id for each quarter + November

- The profit for each product
- The average mark-up for each product
- The sum of days in which the product has been sold

From clicks_regular.csv and clicks_bidding.csv

- The sum of the regular clicks and bidding per product
- The average position of both bidding and regular

From stock.csv

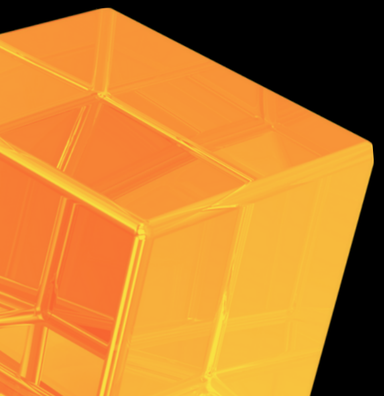
- Average of the stock for each product

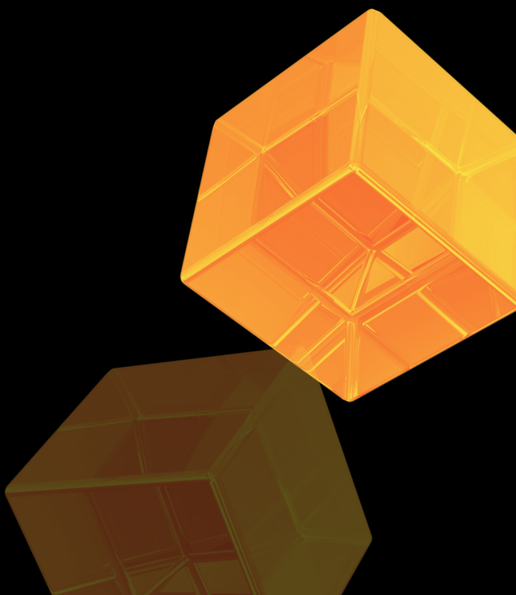
Using Apache Spark

Market dynamics analysis

Data Preparation

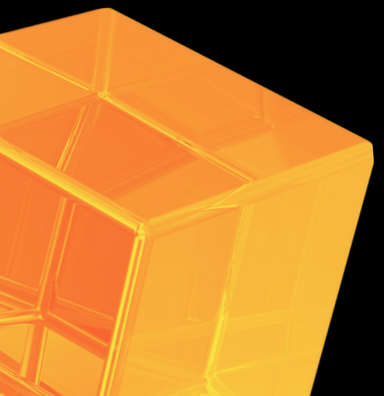
- To run Task 1 we had to aggregate first our products by category, based on the mean change log of the price
- Then we created pivot columns which had the date as index and each seller as column to create 9 different time series and as value there was the mean change price log
- Before of that some filtering was applied such as discarding those sellers which never change the price of their products.

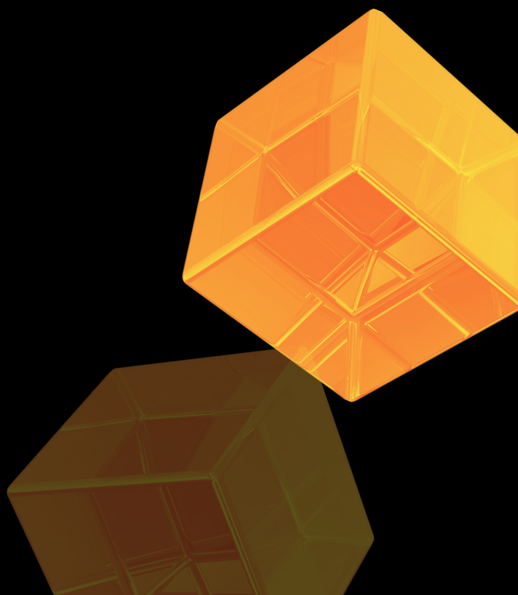




VAR Model

Vector autoregression (VAR) is a statistical model used to capture the relationship between multiple quantities as they change over time and to forecast a time series on itself and the other time series. We used this model to check if there was dynamic princing





VAR Model

Augmented Dickey-Fuller Test to check Stationarity of the series

- Non-Stationary → Differencing
  Obsrvation at **t** - Observation at t-1
- Stationarity

Johansen's Co-Integration Test → To identify whether more than two-time series variables have some similar deterministic trends that can be combined over a period of time.

Selection of the Lag according to of AIC, BIC, FPE and HQIC's results → Fit the model in each Pivot Table → Analysis of the results
Analyzing coefficients and relative p-values and correlation of residuals

Grangers Causation

The Granger causality test is a statistical hypothesis test for determining whether one time series is useful in forecasting another.

We built a matrix for each Pivot with the sellers on both rows and columns

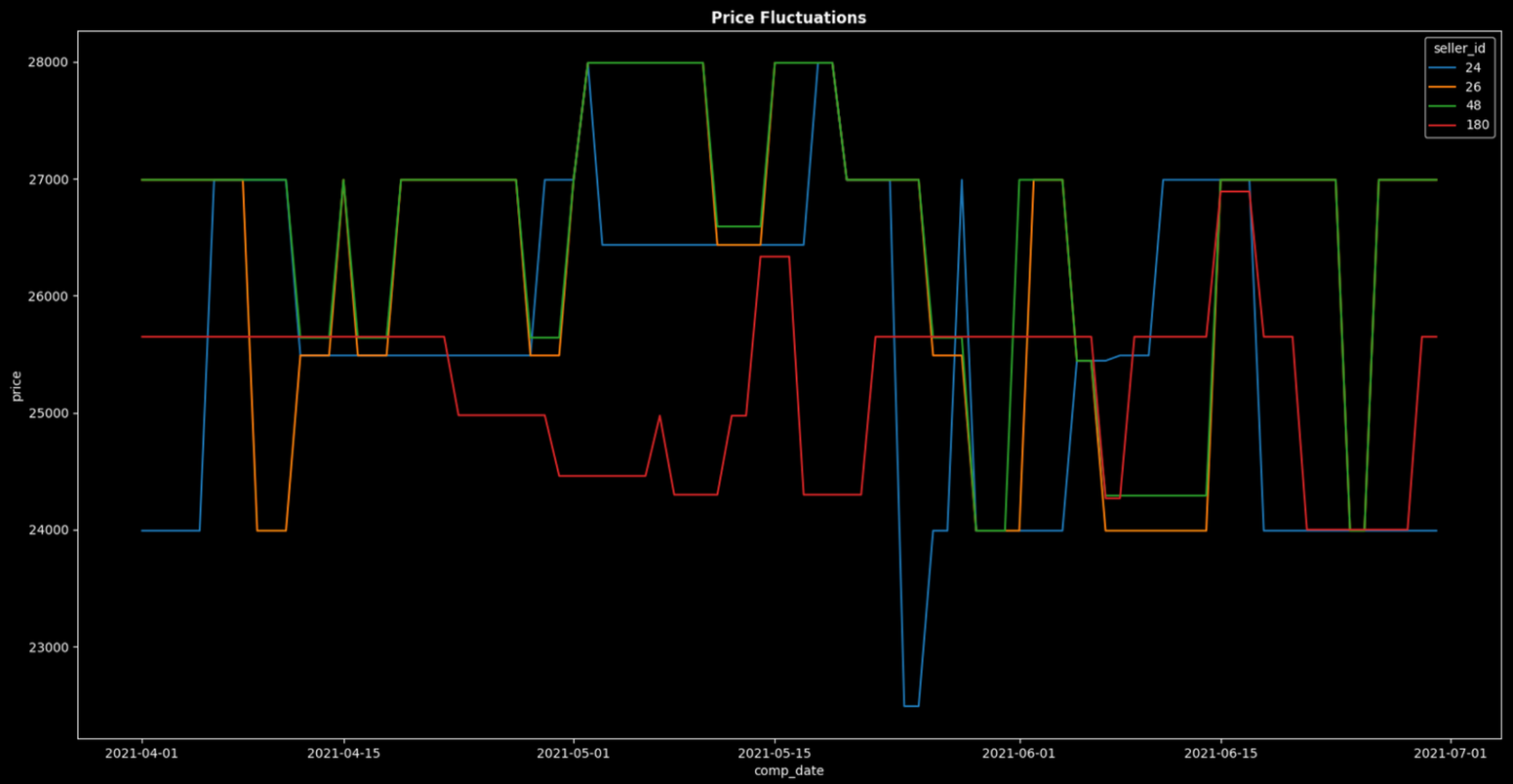
We performed the Chi Square Test, and took into consideration the number of lags obtain in the previous analysis and the minimum p-values was kept in the matrix

Column can be said to **Granger-cause** the row if p-value is < 0.05

According to the grangers causality test we determined, for each category , the seller leader and its followers.

Results

We then compared we value obtained from the VAR model and the granger causality test.



Here plotted, for the second quarter, the variation of the price for the sellers 24, 26, 48, e 180.
As can be seen, sellers 26 and 48 are highly correlated: in the VAR model correlation matrix of residuals, the value for 26 and 48 is 0.96, a sign of an automated price system for the two sellers. Furthermore, granger's causation test suggests that seller 24 is the leader for category 1676, of 26, 48 and 180. The results can be confirmed by the graph made for product 135135
of the 1676 category.

How to check for dynamic pricing?

Check VAR Coefficients and their relative p-value
Correlation in the residuals to check a close movement

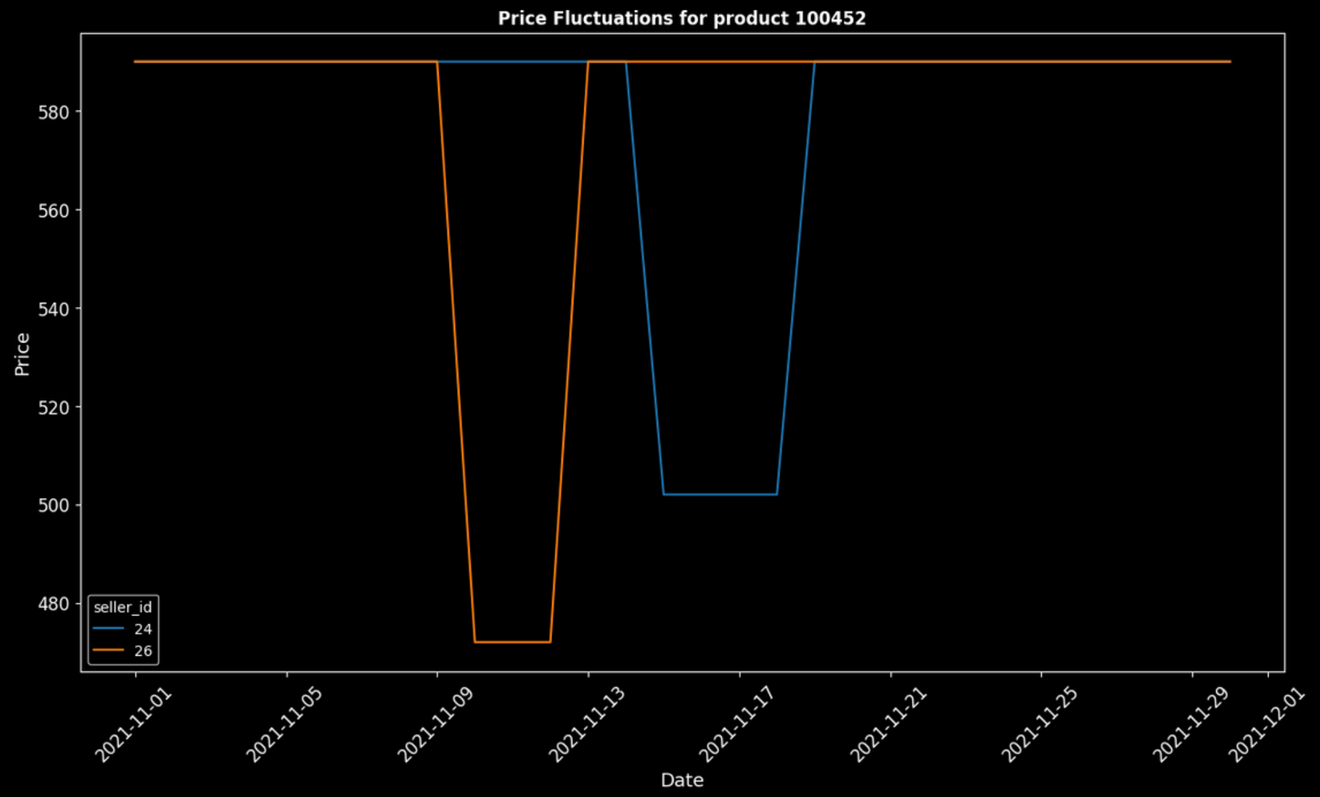
How to individuate Leaders or Followers?

Check graphically with the coefficients of VAR model if small amount of products or use other statistical tests such as the Granger-Causality test

Month of November

Detection of dynamic pricing

- Closeness of the change in price between sellers
- Similar frequency indicates synchronicity in setting prices



Leaders & Followers

- Inclination in following the same trend
- Same number of price changes
- Earliest date = Leader

Seller id	Leader %
26	29,7 %
24	26 %
48	17,6 %
407	7,8 %
188	6,3 %
41	4,7 %

Popularity Index

Business Context

Our team has concentrated its attention towards the company's decision-makers, and we have posed ourselves with multiple questions:

- What is the goal of the company?
- What is the idea of popularity for the company?
- How does the decision-making process works?
- What are the missing data that could support our decisions?
- etc.

We have focused on multiple definitions of popularity indexes to empower Management to access information freely and customise it to make conscious decisions.

Indexes

Products could be indexed based on the following:

- Profits - the profit of a product over the total profit of all products
- Mark Up - average profit per product per quarter
- Sell Frequency - how many times the product was sold
- Quantity - the units of a product over the total units of all products sold
- Stocks - the average amount of product in the warehouse
- Clicks Regular - number of clicks regularly received per product*
- Clicks Bidding - number of clicks bidding received per product*

+ ABC Analysis - Classification of products into three categories: A, B and C. Classified according to value, given by their price multiplied by the demand for that product (quantity sold)

*= yes, we know data about clicks is inconsistent!

Using Apache Spark

Our Dashboard

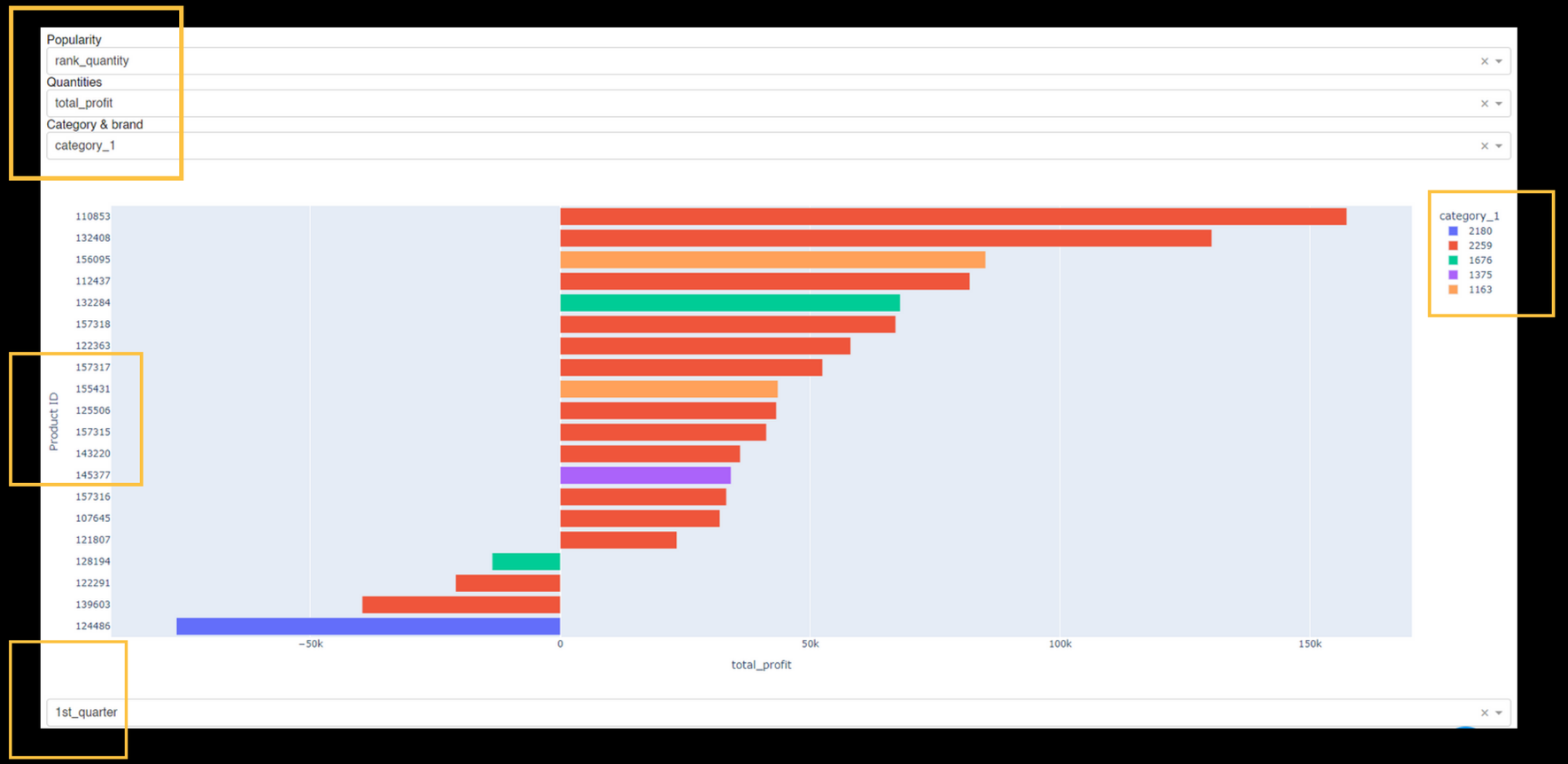
We have decided to rank the product based on sold quantity, with the bars representing the total profit made by those specific products.

Customise:

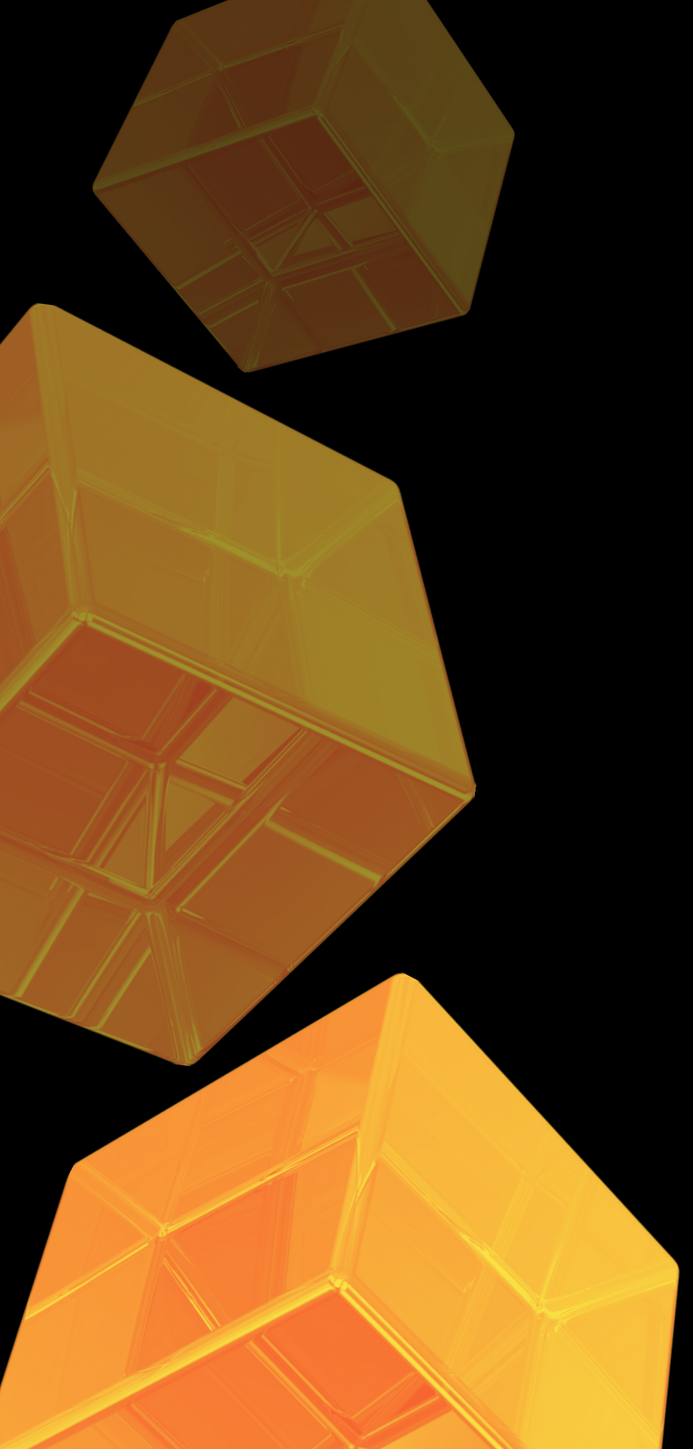
- Ranks
- Values of products
- Highlight categories or brands with colours

Ranks of products
Listed from the 1st ranked to the 20th

Choose the period



Try : https://colab.research.google.com/drive/1BRQDONL646k7R2B5EWH7CinlExgqa281?usp=sharing



Thank You

Appendix

Dataset Preparation

Joined

seller_list.csv
product_catalog.csv
prices_competitor.csv

Data cleaning, manipulation and feature extraction before splitting the dataset into four quarter + the month of november (PySpark):

- We computed the minimum and maximum prices for each product over the entire year 2021.
- We checked when the price had been uploaded twice on the same day, by the same seller, for the same product and cleaned out the second product price change, removing 545,359 rows. We also removed 5 price values that were 0, judging this as an error during data collection.

Data cleaning, manipulation and feature extraction after splitting the dataset into four quarter + the month of november. We then use the power of Python since the size of the dataset was significantly reduced:

- Calculated the logarithm of the price change for each seller's product, and calculated its variance. For a product we eliminated the sellers that present variance equal to zero, this kept only the sellers that applied a price change.
- We then removed all products not sold by seller 24 and those that were generally sold by only one seller

At the end, we create a pivot tables for each of categories of coded_cat_1 in which for each seller, in each date, there are the mean of log prices change between products of each category. Pivot tables were used to perform the following statistical analysis.

Features Extraction 1

From sales_data.csv:

- We extracted profits by sales_price - purcased_price
- Mark-up by profit/purchased_price
- Number of days in which there is a transaction for a specific product

We then grouped on the product_id, computing:

- The sum of profit for each product
- The average mark-up for each product
- The sum of days in which the product has been sold

From product_catalog:

- We took all the information about the categories and the brand of each product

Using Apache Spark

Features Extraction 2

From clicks_regular.csv and clicks_bidding.csv:

- We extracted the number of clicks per product per day, both regular and bidding
- The average position, both regular and bidding, of each product per day

Then we grouped on product_id, computing:

- The sum of the regular clicks and bidding per product on the different quarters,
- The average position of both bidding and regular

From stock.csv:

We compute the average of the stock for each product during the quarter

Using Apache Spark

ABC Analysis 1

ABC analysis is an inventory management technique that determines the value of inventory items based on their importance to the business. It is performed on purchase cost and demand data (based on the quantity sold).

The aim of the function is:

- Classify as A the first 10% of the demanded product that accounts for around 60% of the value of the inventory.
- Classify as B the following 20% of items that account for around 30% of the value.
- Classify as C the remaining 70% of products that account for the remaining 10% of the inventory.

	category	percentage_of_items	cumulative_percentage_of_items	percentage_of_use_by_value	cumulative_percentage_of_use_by_value
0	A	10	10	59.519544	59.519544
1	B	20	30	27.694776	87.214321
2	C	70	100	12.785679	100.000000

ABC Analysis 2

The graph shows how much of the quarter use increases for every 10% increase in the number of items. We see that the curve is steeper for the first products and then flattens out, meaning that the last products have the lowest values of quarter use.