Final Project

Group 3. Presentation for the board of directors.

Data for the next generation store

How to bring value to the Company through Data Science.

Table of contents:

Introduction

Data source and information available.
Objectives of the project.

Available information

Sales, calendar and prices for years 2011 to 2016.

Information from 10 stores across 3 regions: New York, Boston and Philadelphia.

Objectives

1. Go from global trends to understand the angles of the business.
2. Achieve a deeper understanding on how our sales behave.
3. Understand the products similarities, cycles and tendencies.
4. Build an accessible data repository with meaningful insights available for all our stack holders
5. Use our past data to predict future sales.
6. Propose a data-based Machine learning project implantation to optimize shop replenishment.

This is just the starting point of a journey to discover how data science can help us improving and optimizing our results.

NUCLIO

Sales analysis.

Discovering differences, similarities and key points across our network. Main conclusions of a first data analysis focusing in 3 cities: New York, Boston and Philadelphia.

Business insights – first level.

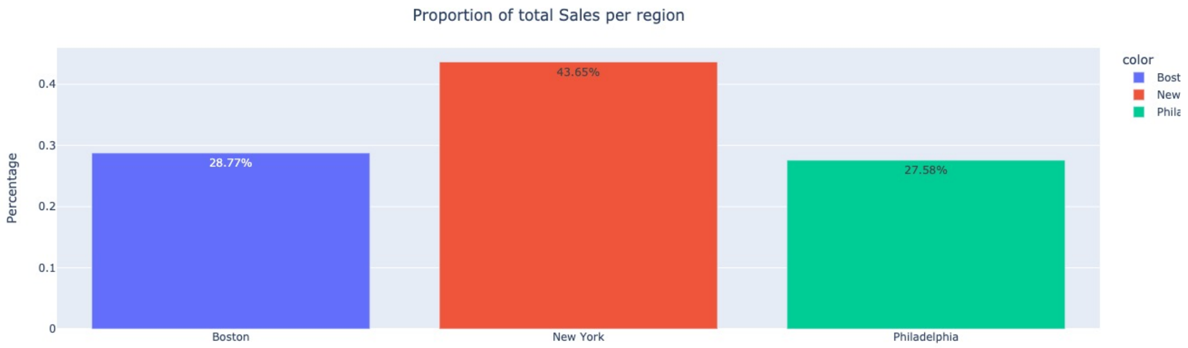
A separate presentation has been prepared for showing the results of the first data analysis in a very detailed way, and was already discussed in a previous meeting.

In any case, here we are just summarizing the key aspects of this first approach as a starting point of the rest of the project.

Regions and shops

Units sold and revenue per region (%)

Distribution of sales (units) per region



Distribution of revenue ($) per region

We checked that the proportion of total sales and items sold are aligned, meaning that the revenue of the business is stable across regions. We also saw a slightly difference in the New York and Philadelphia percentages, meaning that same number of sales in NY and PH have a slightly higher revenue in NY.
As we discussed, **New York accounts for the 43%** of the sales of the three areas.

Sales per store in each region (units)

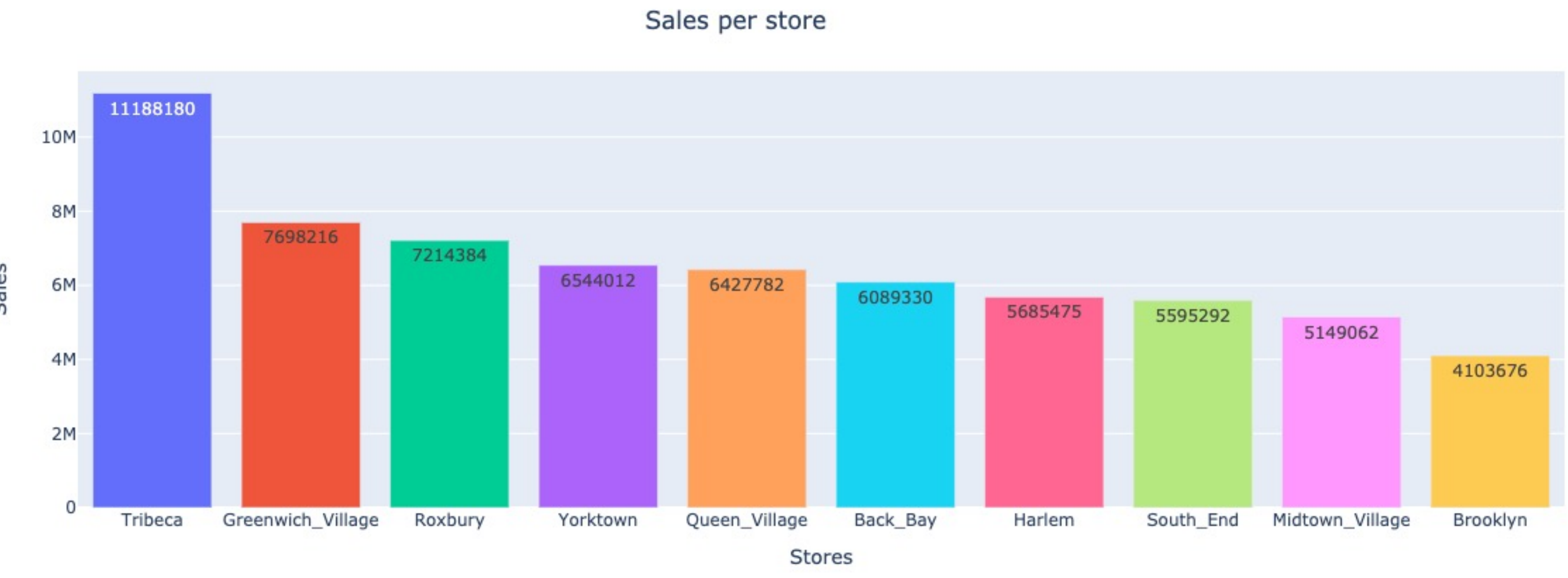
Considering NY as a key territory, we all agreed that it is clear how Tribeca is pushing up the sales of the NY area, being Greenwich_Village the best 2nd shop , Harlem an average store and Brooklyn the one with the lower sales number across the three regions. **Tribeca is a key store**.

NUCLIO

Sales per store ($)

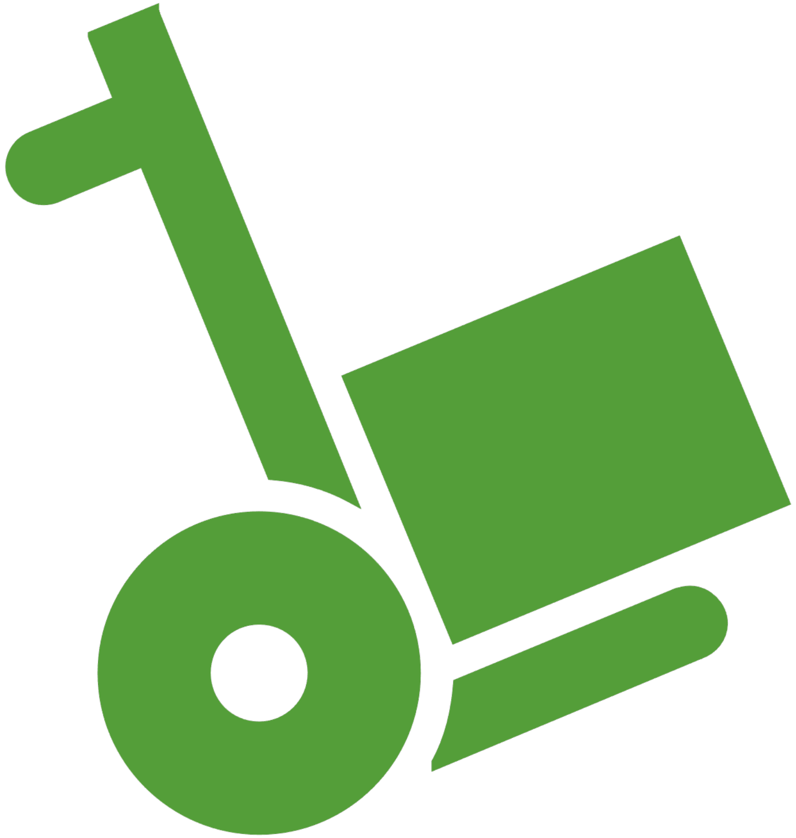
There is an interesting finding when comparing the sales per item and revenue across stores. As we have seen, some of our regions make a higher revenue with less sales (and others make a smaller revenue with more sales). We cannot affirm that the first ones are more profitable stores since we do not have enough information (purchase price, COL…)

NUCLIO

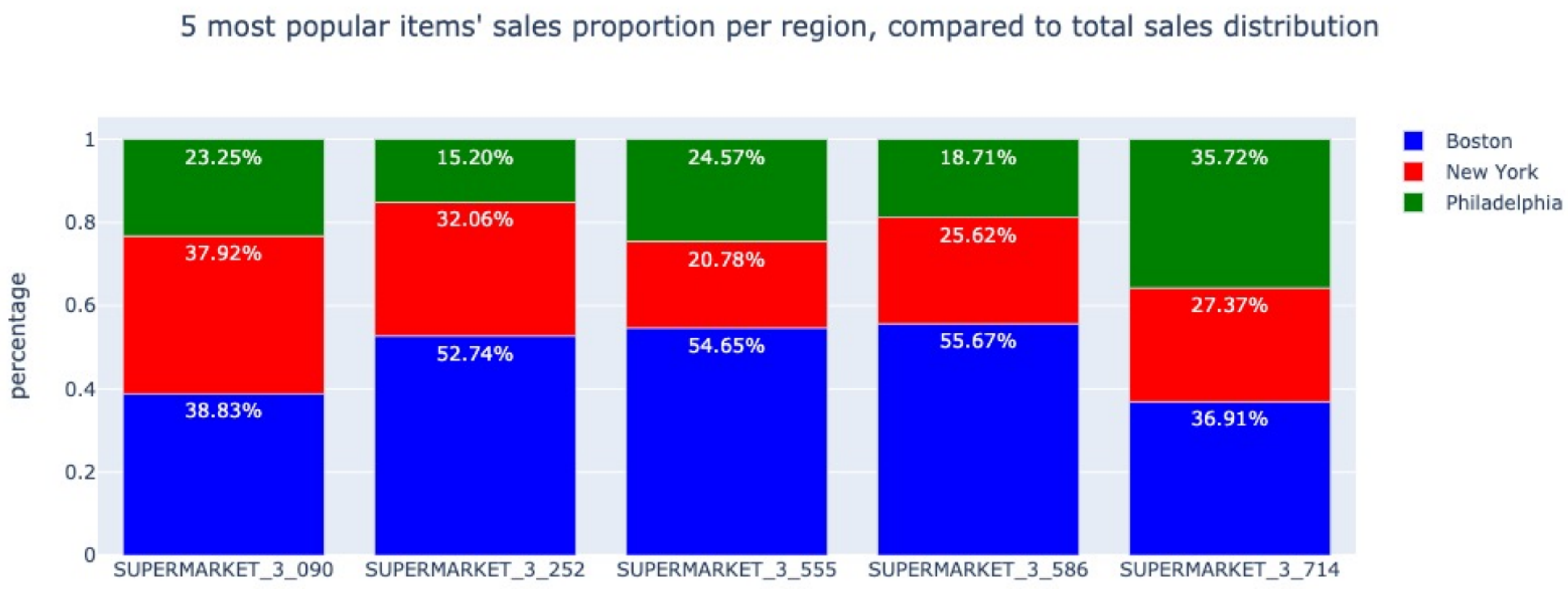


Yorktown, **Queen Village** and **Midtown Village** are our shops in PH. Being Yorktown and Queen Village the 4th and 5th stores by items sold, these are the 5th and 7th by revenue .

NUCLIO



Supermarket categories and products

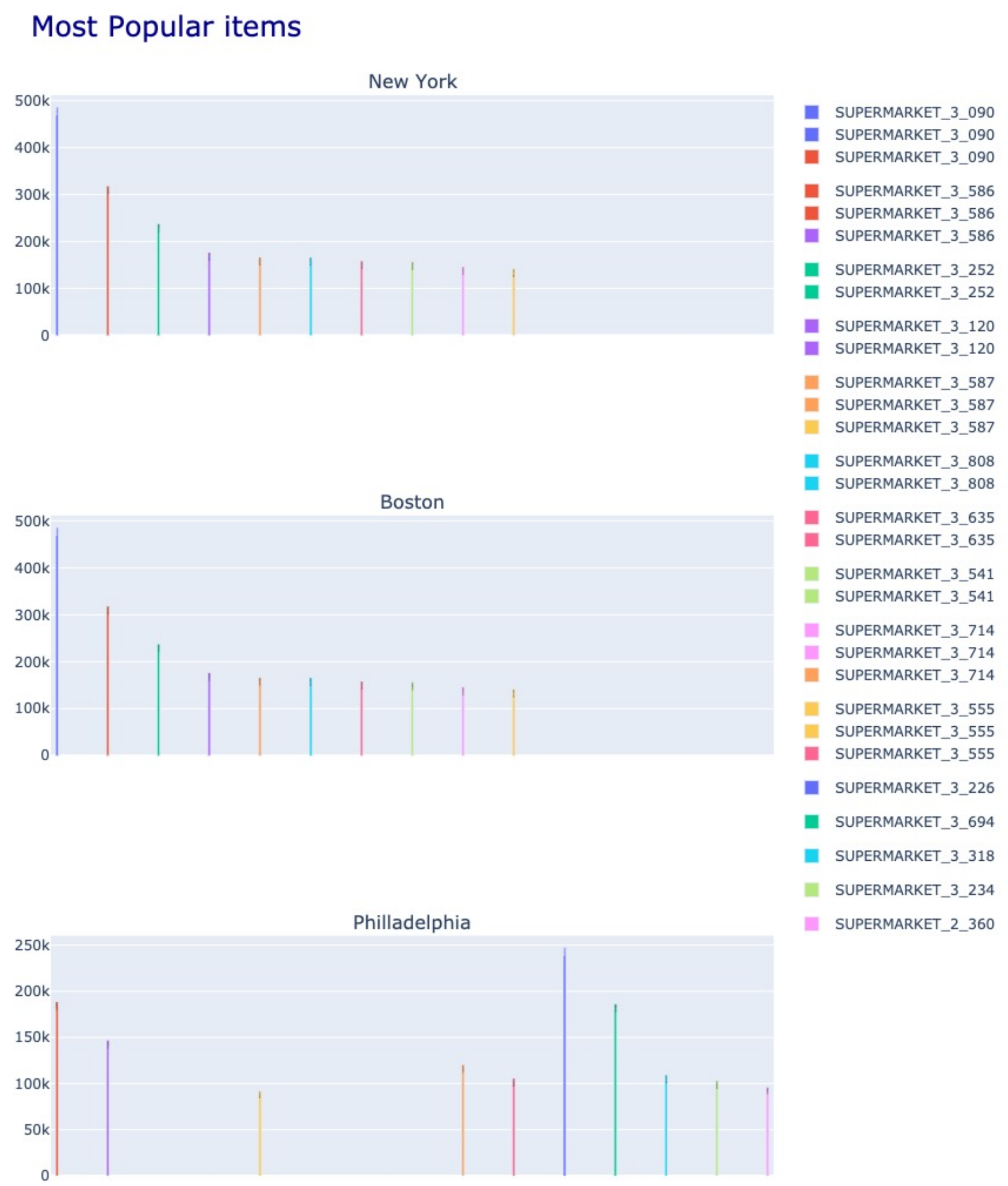


Item sales % per region

When considering the top 5 most popular products, we saw how they were not equally sold across the different regions.

Some insights: It is significant that the sales of 555 are percentualy higher in Philadelphia than in NY. 714 has almost the same weight in Philadelphia than in NY.

NUCLIO



Despite not being a very aesthetic graph, it allowed us to visually see how most popular items vary across regions.

We can also see how Boston and New York regions are more alike than Philadelphia, in terms of which of products are more sold.

This format of graph showed again in a very visual way the differences between the top 10 less sold products across regions.

Improvement areas

- We could only access to a **limited part of the data**. Accessing the company's database would help us evaluate possible optimizations in the process of data cleaning and preparation and identify additional potentially useful information.
- Having **additional information**, like the **commercial name** of the items, would had helped us in our analysis. Data is powerful, but **understandable human context** is necessary for a complete analysis. **Purchase price** would had helped us studying profitability.

NUCLIO

Self service Reporting and dashboards

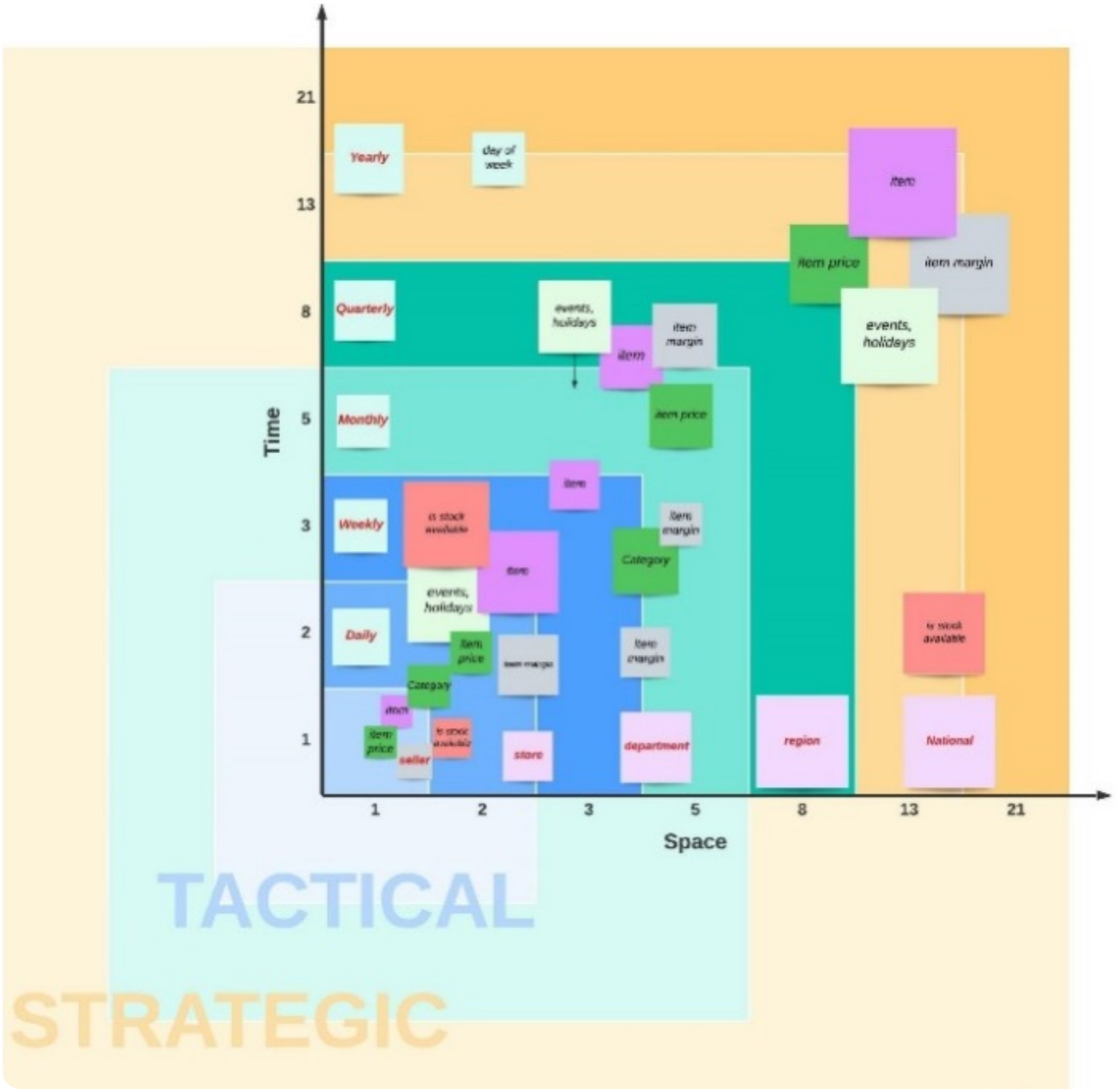
BI solution implementation

For data analysis to be useful within an organization, it needs to be accessible to all stakeholders. The steps toward a data driven decision company must include an easy and powerful way to access the available information.

<u>Data Compositions</u> and Self Service Strategy

- Key results measures:
  - **Revenue**
  - **Items sold**
- Available Dashboards revolve around **Purpose** and **Audience**
- **Relevance** of data can be influenced by the combination of **3 Main factors** :
  - **Time :** Date, Week, Month, Quarter, Year
  - **Space:** Store, Department, Region, Country
  - **Matter :** stock_available, id, item, category, price

NUCLIO



Reports matrix

In this matrix we can see how we need to go from the most **tactic shop daily** reports, to the ones used **strategically** each **year** to review margin and prices of the business.

We proposed Tableau as BI tool and shared with you its potential by creating some dashboards with the available information.

Top 12 performance

Top 12 Items performance (revenue)

Semana de dates

item	18 May 2015	25 May 2015	1 June 2015	8 June 2015	15 June 2015	22 June 2015	29 June 2015	6 July 2015	13 July 2015	20 July 2015	27 July 2015
ACCESORIES_1_158	1,963	9,603	8,348	9,763	9,422	8,995	10,799	9,752	9,058	10,142	9,599
ACCESORIES_1_354	2,528	14,053	12,974	12,417	12,943	10,630	12,574	12,669	12,361	11,838	12,111
HOME_&_GARDEN_1_110	579	3,894	4,449	4,004	3,780	3,990	4,240	3,964	4,350	3,798	4,185
SUPERMARKET_1_096	568	4,430	5,071	5,061	4,733	5,103	6,164	7,522	6,516	5,489	5,834
SUPERMARKET_3_090	5	10	16	1,548	7,260	6,521	7,331	6,793	7,412	7,163	6,204
SUPERMARKET_3_120	1,810	10,549	12,054	13,393	12,897	11,158	13,320	11,863	11,229	11,589	11,880
SUPERMARKET_3_202	1,178	5,902	7,323	15	66	46	66	8,767	9,101	8,049	8,322
SUPERMARKET_3_252	496	3,497	4,132	4,722	4,557	4,450	4,864	4,309	5,232	5,280	5,606
SUPERMARKET_3_282	463	3,136	3,611	4,459	4,841	3,593	4,053	4,784	3,359	2,820	3,037
SUPERMARKET_3_444	1,245	3,816	86	78	78	78	63				
SUPERMARKET_3_586	927	5,460	6,504	6,808	6,293	6,249	6,808	7,061	7,038	6,822	7,019
SUPERMARKET_3_587	808	5,075	5,971	6,792	5,938	4,737	4,867	5,800	5,896	5,515	5,560

Top 12 Item movement (units sold)

Semana de dates

item	18 May 2015	25 May 2015	1 June 2015	8 June 2015	15 June 2015	22 June 2015	29 June 2015	6 July 2015	13 July 2015	20 July 2015	27 July 2015
SUPERMARKET_3_090	3	6	10	947	4,436	3,984	4,478	4,154	4,531	4,377	3,789
SUPERMARKET_3_586	484	2,851	3,396	3,555	3,286	3,263	3,555	3,687	3,675	3,562	3,665
SUPERMARKET_3_252	272	1,917	2,265	2,588	2,498	2,439	2,666	2,362	2,868	2,894	3,073
SUPERMARKET_3_120	303	1,766	2,018	2,242	2,159	1,868	2,230	1,986	1,880	1,940	1,989
SUPERMARKET_3_587	269	1,687	1,987	2,259	1,978	1,576	1,619	1,927	1,964	1,830	1,844
SUPERMARKET_3_282	154	1,043	1,201	1,483	1,610	1,195	1,348	1,591	1,117	938	1,010
SUPERMARKET_3_202	231	1,159	1,437	3	13	9	13	1,717	1,784	1,579	1,633
HOME_&_GARDEN_1_110	127	854	976	878	829	875	930	869	954	833	918
SUPERMARKET_1_096	63	516	578	577	545	583	696	854	719	623	653
ACCESORIES_1_354	82	456	421	403	420	345	408	411	401	384	393
ACCESORIES_1_158	66	323	282	328	316	301	362	328	304	339	323
SUPERMARKET_3_444	157	481	11	10	10	10	8				

item

store
- (Todo)
- Back_Bay
- Brooklyn
- Greenwich_Village
- Harlem
- Midtown_Village
- Queen_Village
- Roxbury
- South_End
- Tribeca
- Yorktown

category
- (Todo)
- ACCESORIES
- HOME_&_GARDEN
- SUPERMARKET

Año de dates
- (Todo)
- 2014
- 2015
- 2016

dates

24/05/2015 24/04/2016

+ableau

This dashboard can lead us to conclusions like:

-Top 12 items vary over different time ranges.

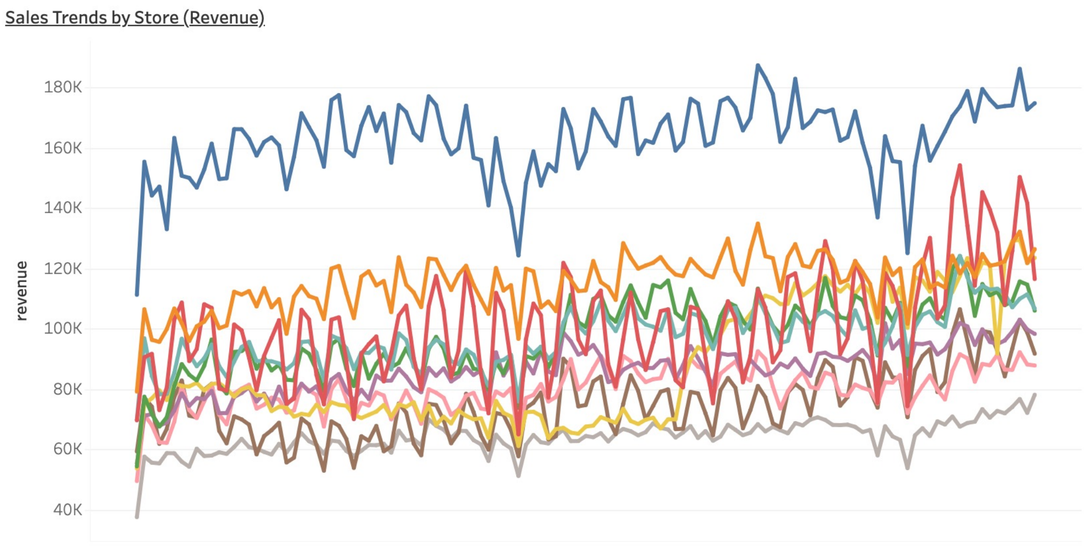
-Different stores have different top 12 items.

-Specific products may start and stop selling at a given time.

-There is high seasonality with some products: they may be top sellers at one point and be at the bottom at some other.

-There are some products that were popular at a given time but are not selling any more.

Sales by store (revenue)



Leader Board

store	revenue	unit_sold
Back_Bay	11,760,972	3,014,971
Brooklyn	7,748,062	2,013,616
Greenwich_Village	13,934,943	3,692,451
Harlem	10,614,240	2,705,376
Midtown_Village	10,399,570	2,862,559
Queen_Village	9,088,234	2,693,424
Roxbury	11,771,447	3,157,231
South_End	9,645,833	2,626,055
Tribeca	19,686,679	5,318,014
Yorktown	12,231,558	3,557,537

Semana de dates

store	30 Decemb..	6 January 2014	13 January ..	20 January ..	27 January ..	3 February..	10 Februa
Tribeca	111,233	155,286	144,084	147,066	132,933	163,189	150,62
Greenwich_Village	79,169	106,453	96,315	95,632	99,871	106,687	101,80
Yorktown	69,653	90,544	91,685	73,011	80,558	103,037	108,67
Roxbury	69,825	96,817	84,336	78,269	78,371	85,368	96,69
Back_Bay	54,431	77,495	71,792	67,584	70,661	81,010	88,27
Harlem	53,776	75,042	76,927	79,613	76,337	82,047	80,97
Midtown_Village	55,372	71,176	71,693	67,871	68,945	72,893	78,57
South_End	49,548	71,531	68,046	62,359	62,193	69,194	81,16
Queen_Village	59,349	77,189	74,201	61,861	66,847	77,358	82,87
Brooklyn	37,600	57,640	55,610	55,386	58,790	58,752	55,96

region
- (Todo)
- Boston
- New York
- Philadelphia

store
- (Todo)
- Back_Bay
- Brooklyn
- Greenwich_Village
- Harlem
- Midtown_Village
- Queen_Village
- Roxbury
- South_End

dates

01/01/2014 24/04/2016

store
- Tribeca
- Greenwich_Village
- Yorktown
- Roxbury
- Back_Bay
- Harlem
- Midtown_Village
- South_End
- Queen_Village

category
- (Todo)
- ACCESORIES
- HOME_&_GARDEN
- SUPERMARKET

revenue

37,600 187,222

This dashboard :

Shows Top 12 items and how they vary over different time ranges and stores.
It allows to interpret that specific products may start and stop selling at any given time.
Also that there is high seasonality with the different products, some products may be top sellers at one point and be at the bottom at some other .

Example of dashboard insights:

Unit Sales Trends by Day of Week | Tableau Public Date Range : 24/05/2015 – 04-24-2016

Tops **Revenue Sales** occur from Friday to Monday.

Around 650.000 USD was earned on all the Sundays versus 462.000 USD on all Wednesdays of the selected date range.

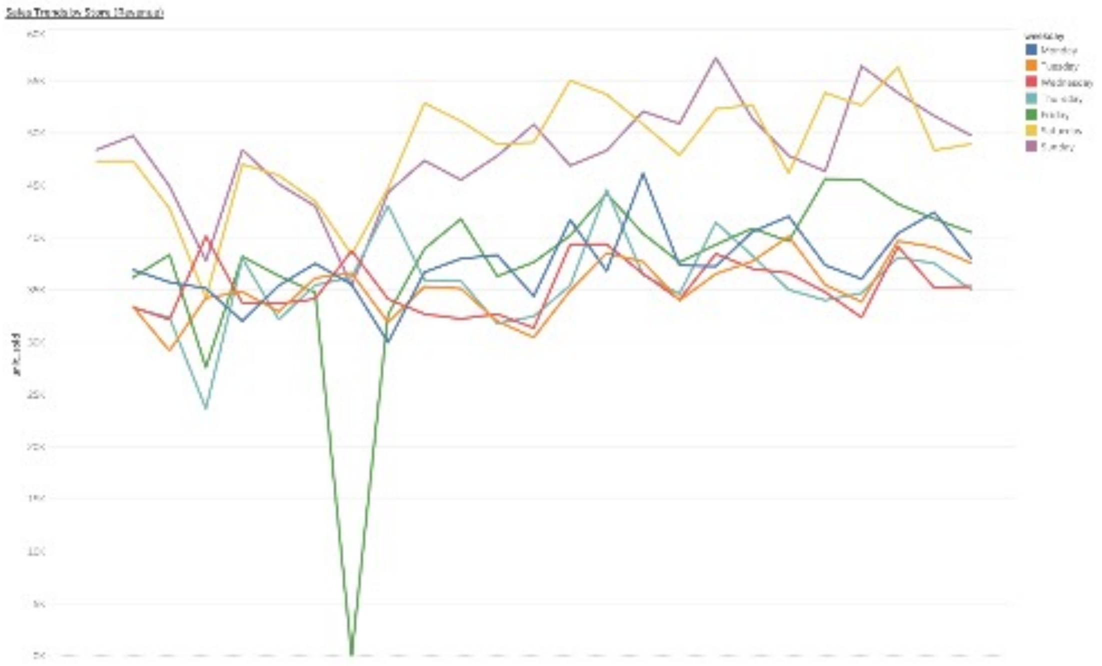
Average **revenue** across all stores on Sundays is 132 USD compared to lowest average of 97 USD on Wednesdays.

Around 147.000 **units** were sold on all the Sundays versus 100.000 **units** were sold on all Wednesdays of the selected date range.

An average of 30 units were sold on Sundays versus 21 units sold on Tuesdays and Wednesdays

Saturday is the peak day for Roxbury, Yorktown and Midtown Village

Top Days over time can spike up and down, this can significantly be seen during holidays such as Thanksgiving, Christmas and New Years



Link	Description	Possible applications
Global Sales by Day of the week | Tableau Public	Actual Average of Revenue sales by Day of the week compared against the forecasted average per day of week.	Weekly Tactical operational management benchmarking across stores and regions Target setting
Global Revenue and Unit Sales predictions | Tableau Public	Global Revenue predictions. Global Unit Sales predictions .	Strategic Global forecasting Top-Down goal setting Global budget and stock procurement allocation
Revenue trends by Store | Tableau Public	This shows store trends and comparison of store performance over selected periods of time.	Store sales performance management Trend analysis Sales step goals and kpi's
Top 12 Items Performance | Tableau Public	-Top 12 items vary over different time ranges -Different stores have different top 12 items -Specific products may start and stop at any given time -There is **high seasonality** with the different products, some products may be top sellers at one point and be at the bottom as some	Sales bundles for items that are likely to sell together Sales campaigns for items that have high affluence to seasons or events
Top and Bottom revenue earners by Store | Tableau Public	Shows total revenue and unit sold of Top 10 and Bottom 10 items by Store You can select different time periods to see change in top and bottom items and break this down by store	store product performance management operational sales / mktg management customer insights top / bottom by item
Top and Bottom Items (Revenue) by Region | Tableau Public	Shows total revenue and unit sold of Top 10 and Bottom 10 items by region You can select different time periods to see change in top and bottom items and break this down by store	regional product performance management operational sales / mktg management customer insights top / bottom by item
Unit Sales Trends by Day of Week | Tableau Public	Shows Trends over time of Unit Sales per day of the week	Evaluation. Model evaluation to see if day of the week campaigns are still relevant for both Stock replenishment and per store.
Global Units Sales Dashboard | Tableau Public	This Dashboard shows different reports representing actual unit sales values over different time breakdowns.	Evaluation of Inventory management against unit sales.

Task 2 - Clustering

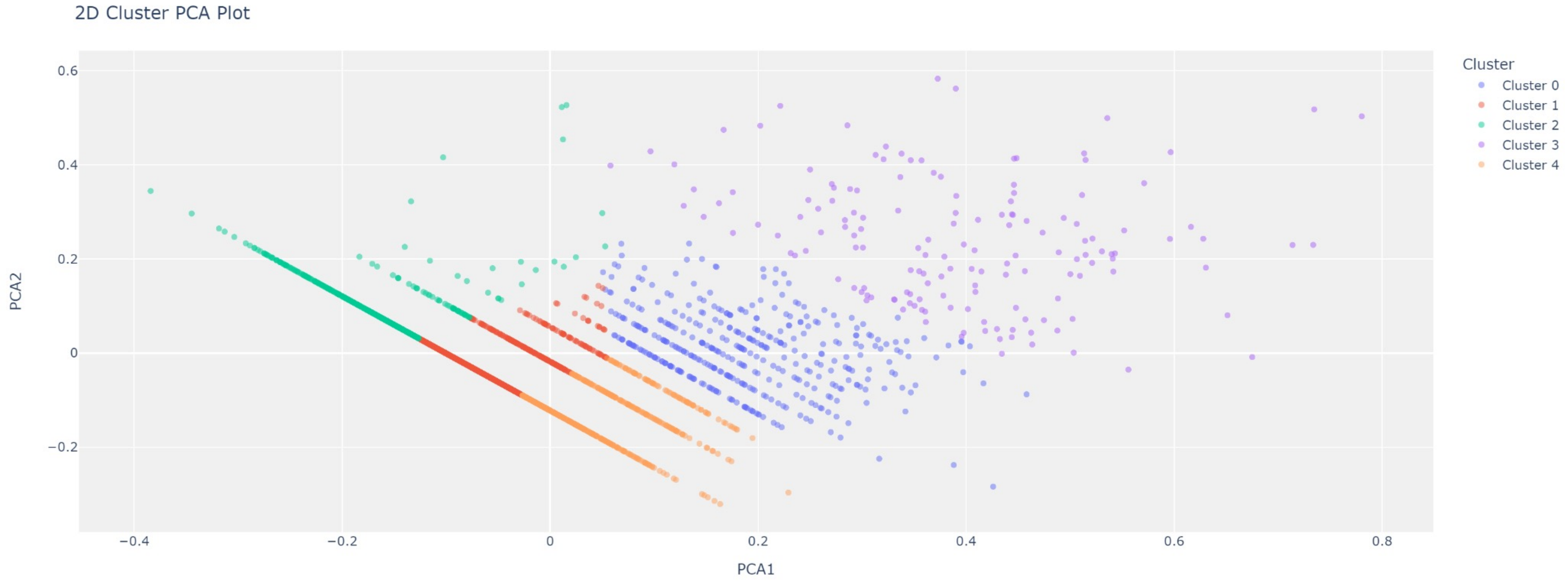
How identifying similarities and grouping products can help marketing and store management.

Clustering: grouping products by its behaviour.

- The clustering part is aiming at identifying **products that behave similarly**.
- The data from the **Tribeca** store (most selling store) were used to create the clusters.
- Clustering features created are **frequency** (total number of sold products per item) and **recency** (days since last purchase).
- The cluster model has been applied to the rest of the stores and their results are also available.

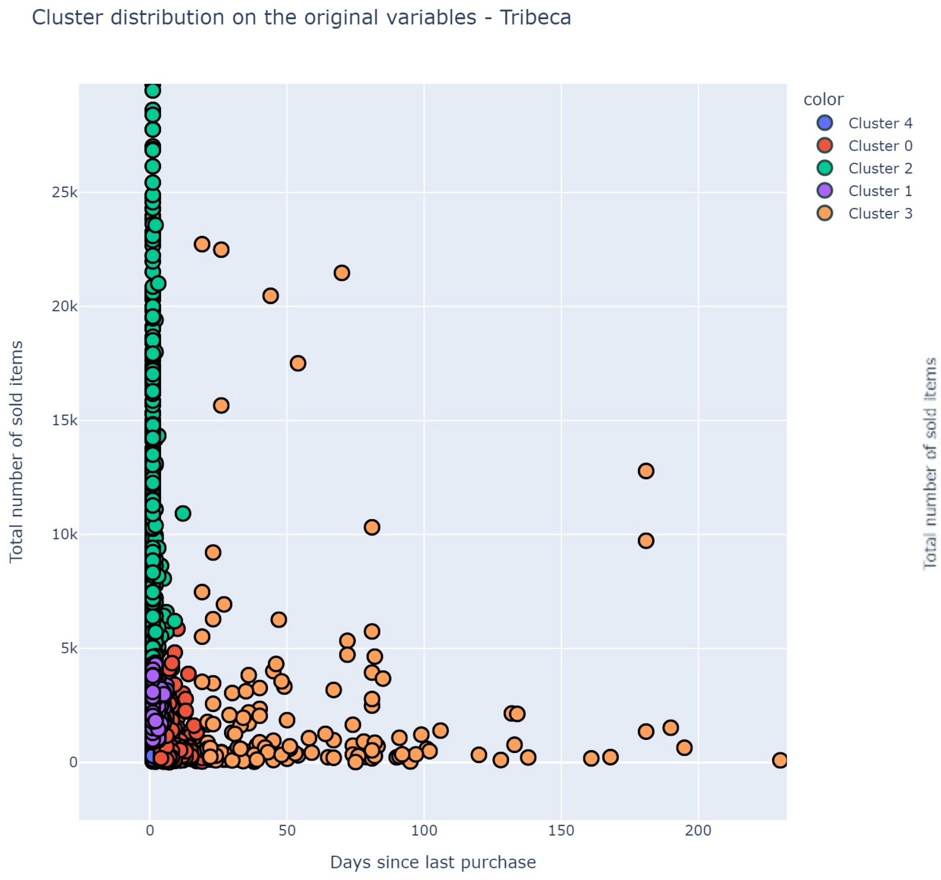
NUCLIO

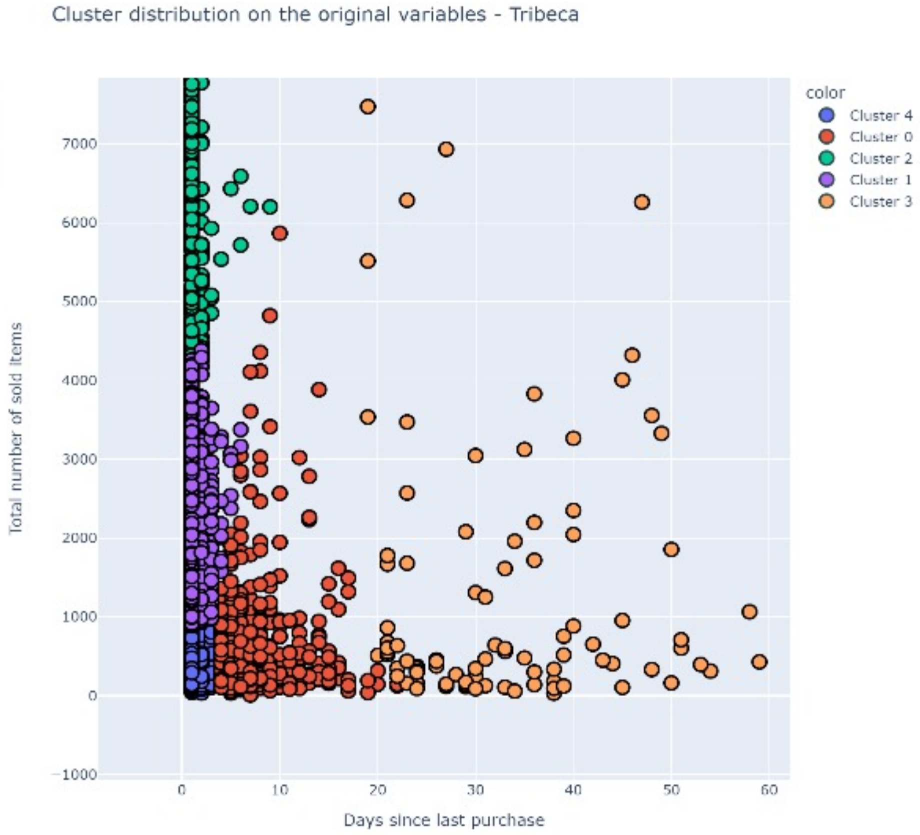
Cluster segmentation: Log transformed features.



NUCLIO

Cluster segmentation: Original variables

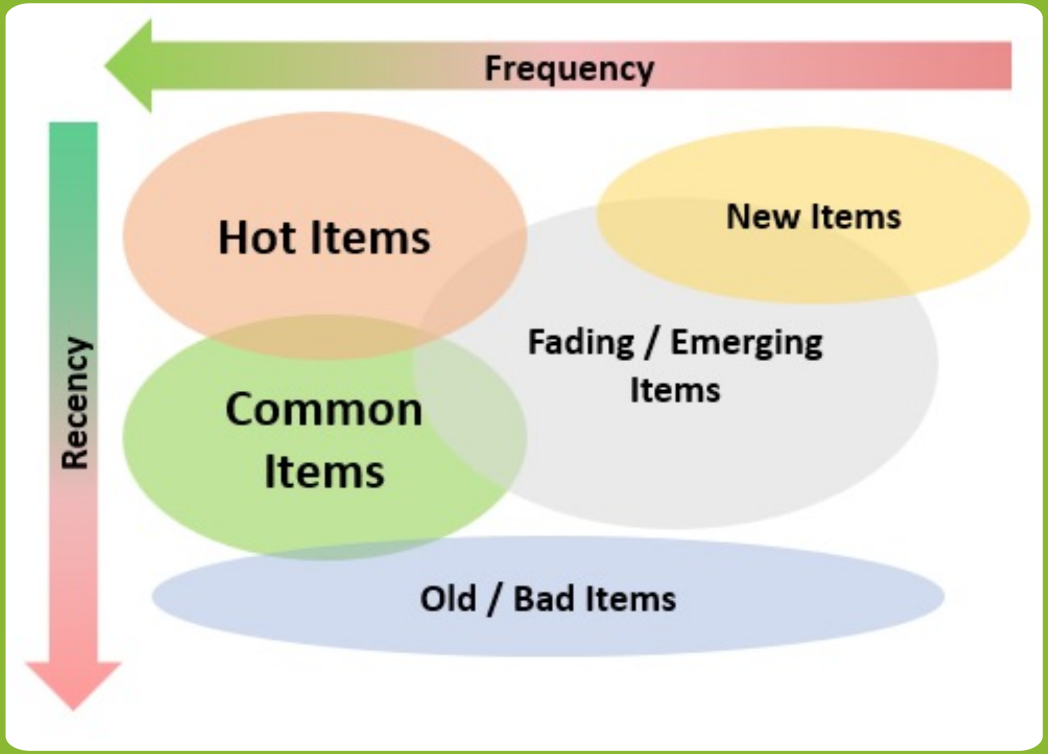




The plot on the right is a zoomed area of the left plot.

These results can be explored on the following dashboard: [Item Clusters | Tableau Public]

Cluster interpretation



Cluster interpretation:

Cluster 0 - **Transition Items** - Fading or emerging products: Medium recency and a wide range in frequency: they could be emerging products (lower-left part of the cluster) or fading products (upper-right) part of the cluster.

Cluster 1 - **Common Items**: Cluster with medium frequency and low recency: these are common products, should include most of the product and as a matter of fact it is the cluster with more items.

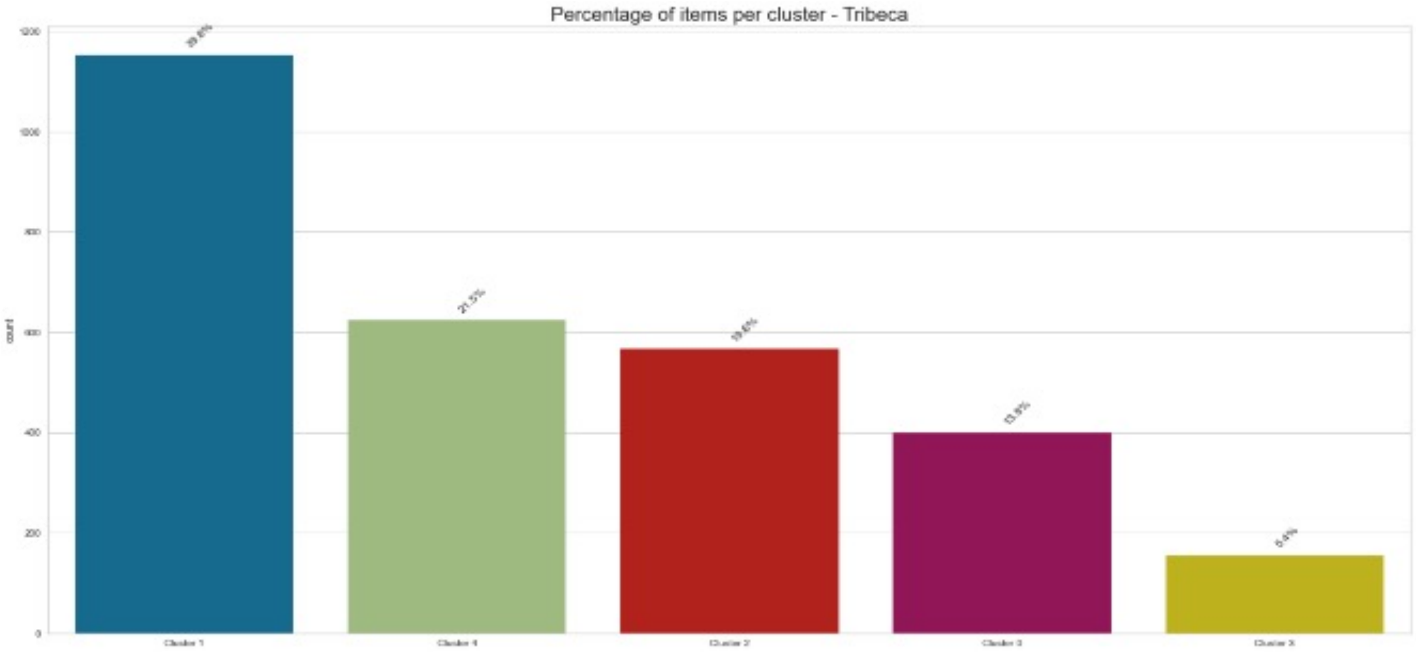
Cluster 2 - Hot Items: Cluster with high frequency and low recency: these are the hot/top selling products.

Cluster interpretation:

Cluster 3 – Old/Bad Items: Cluster with high recency and a wide range in frequency: they could be bad products, that never sold (lower part of the cluster) or old products (upper part of the cluster) that sold in the past but now are not selling anymore.

Cluster 4 - New Items: Cluster with low frequency and low recency: these are new/appearing products.

NUCLIO



Item distribution histogram.

- The cluster with the most product is, as expected, the Common Items cluster.

- The cluster with less items is again, as expected, the Old/Bad Items

Model evaluation.

- Model checked against the 10 most sold and less sold products in the New York area.
  - All most sold products are in Cluster 2: Hot items.
  - Less sold products are in Cluster 0, 3, and 4 --> products that sold very little can be either bad/old products, new products, or transition products (emerging or fading products)
- This cluster model has been applied to data for all the other stores, with similar results. **The same item can correspond to a different cluster depending on the store**.

Considerations

- Products that were **hot items** but **haven't been sold recently** could be wrongly segmented as old/bad products. They would be better identified as *fading products* instead.
- The line between **Common/Hot** and **Common/New** products is a **subtle line**. It is important not to act on products that are on the border of clusters.
- **Seasonal** products could appear in **any cluster**, depending on the season when the model's data last date was, and if they are good or bad selling products.

Considerations

- **Fading/emerging** products will tend to move to other clusters as the time passes by.

Emerging products→ Common or Hot items cluster

Fading products →Old/Bad product cluster.

- **New items** will move to other clusters as the time passes by as well.

Good products → Common or Hot cluster

Bad products → Old/Bad product cluster.

- The model needs to be **re-trained regularly during different seasons** to see the evolving behaviour of the products and discover new Hot products or new Bad products to act upon. A good retraining time could be **quarterly,** to have a picture of the behaviour during all seasons.

Aplications:

Cluster segmentation can give an insight on products to act upon, e.g.:

- An insight of bad/old products could be used to evaluate if **removing** some of them from the store, or to **relocating** the stock where the product is still in a good selling cluster (to minimize stockage)

- An insight on hot products, could be used for example for developing a new **pricing strategy** for them intended increase the revenues.

Cluster segmentation can be used for **modelling predictions**.

Task 3 – Sales forecasting

Using past data to predict future needs.

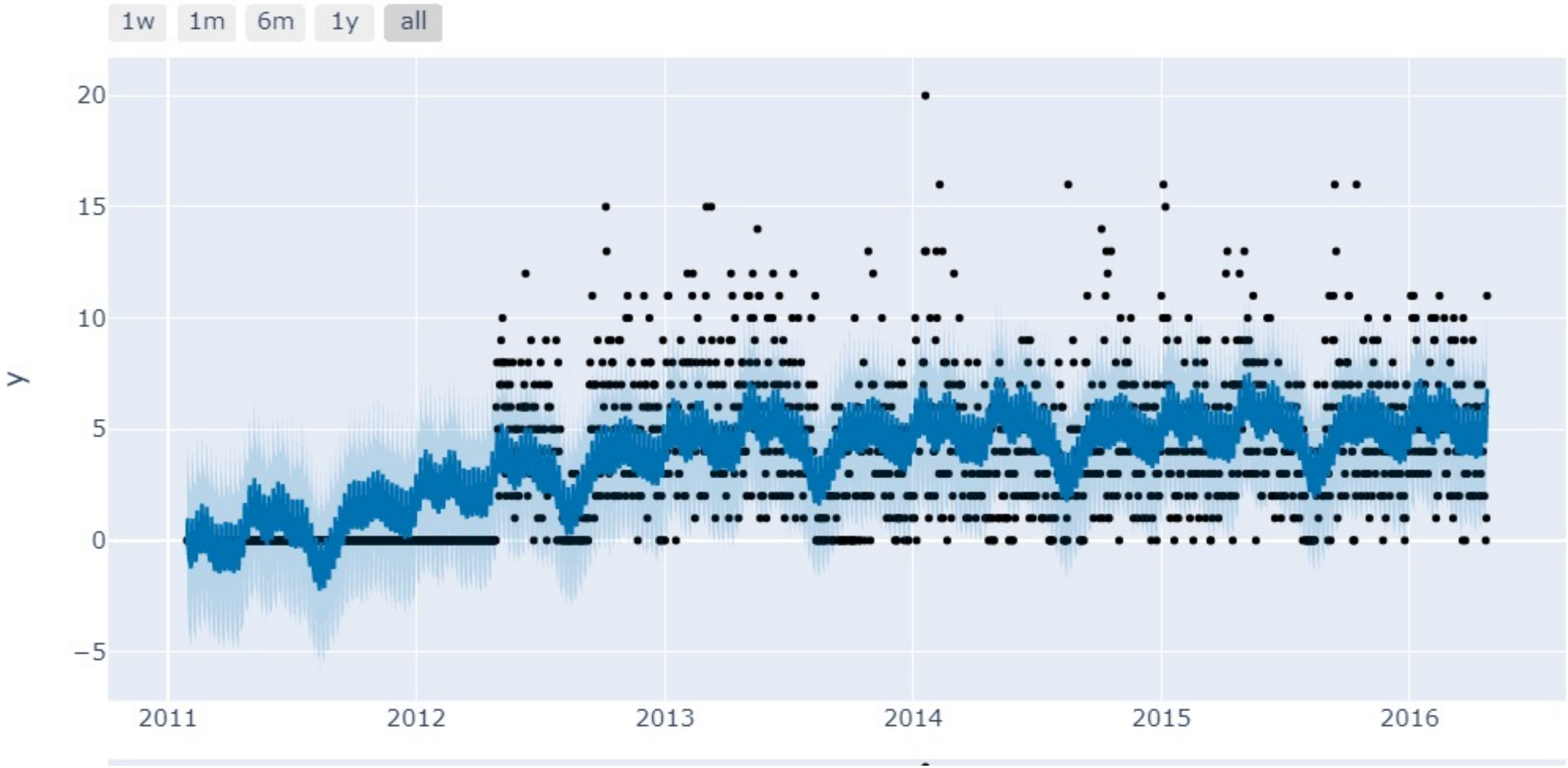
Initial considerations

First thought: Create a model for every product in each store. Constraint: we did not have the computer power required.

Thus, we used the **clusters priorly established** to **group the products** and **created a model for every cluster-**

As a test set, we decided to select **Tribeca**, and within this store we have selected the **50 items with the highest unit sold in the last 30 days**. The predictions for one of these items can be found in the next slide.

Tribeca top 50 products last 30 days



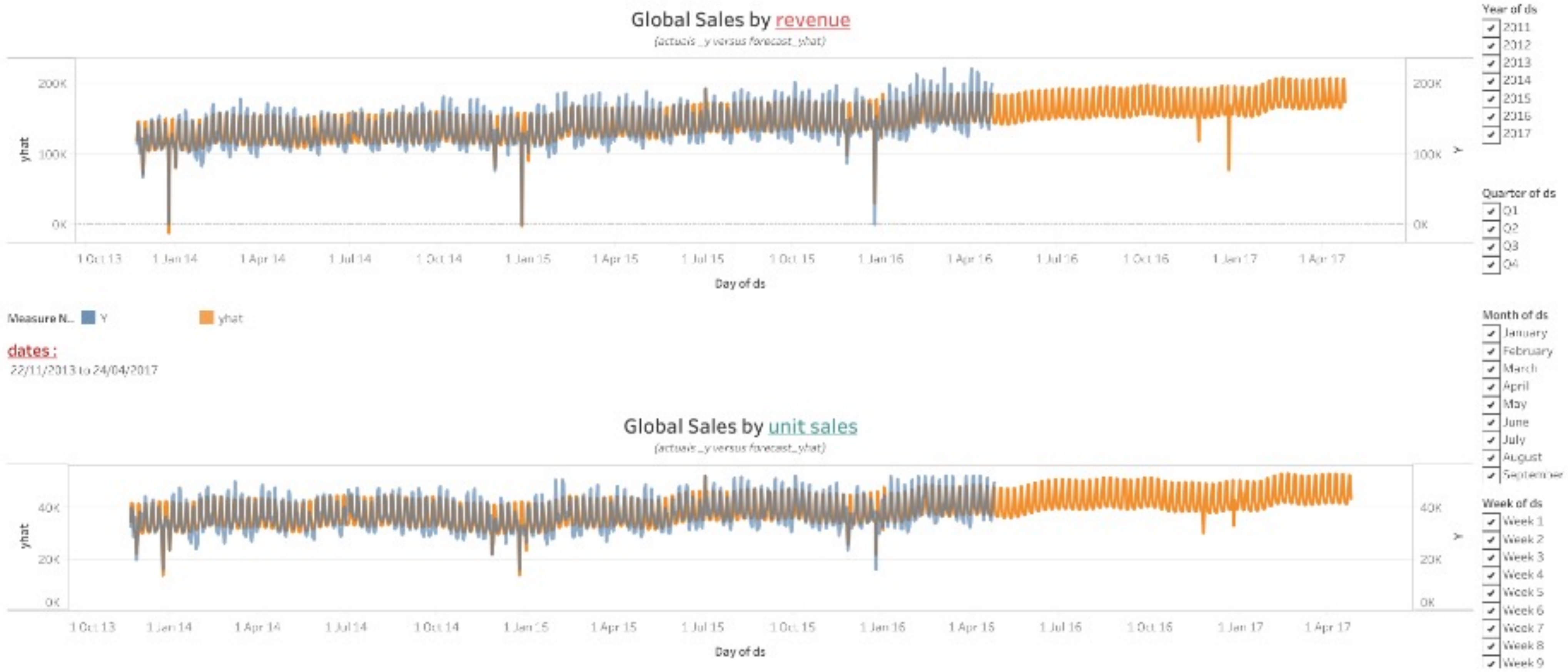
NUCLIO

Validation considerations

The model is unable to catch all the seasonality in the data, this is because it is a generalised model on a cluster of products. Therefore, the inventory strategy that would match our model is an **average strategy**.

Once our model is validated, we can go on to **create the models for every cluster in each store and predict the units sold for each of these products**. For consolidated predictions (i.e. sales per category, department) we would simply sum the individual predictions.

Model user case value (national)



Model user case value:

Potential **1,400,000M USD** Value of Model quarterly with even with an accuracy of 70 percent .

Assuming model is run **across** all stores of the United States

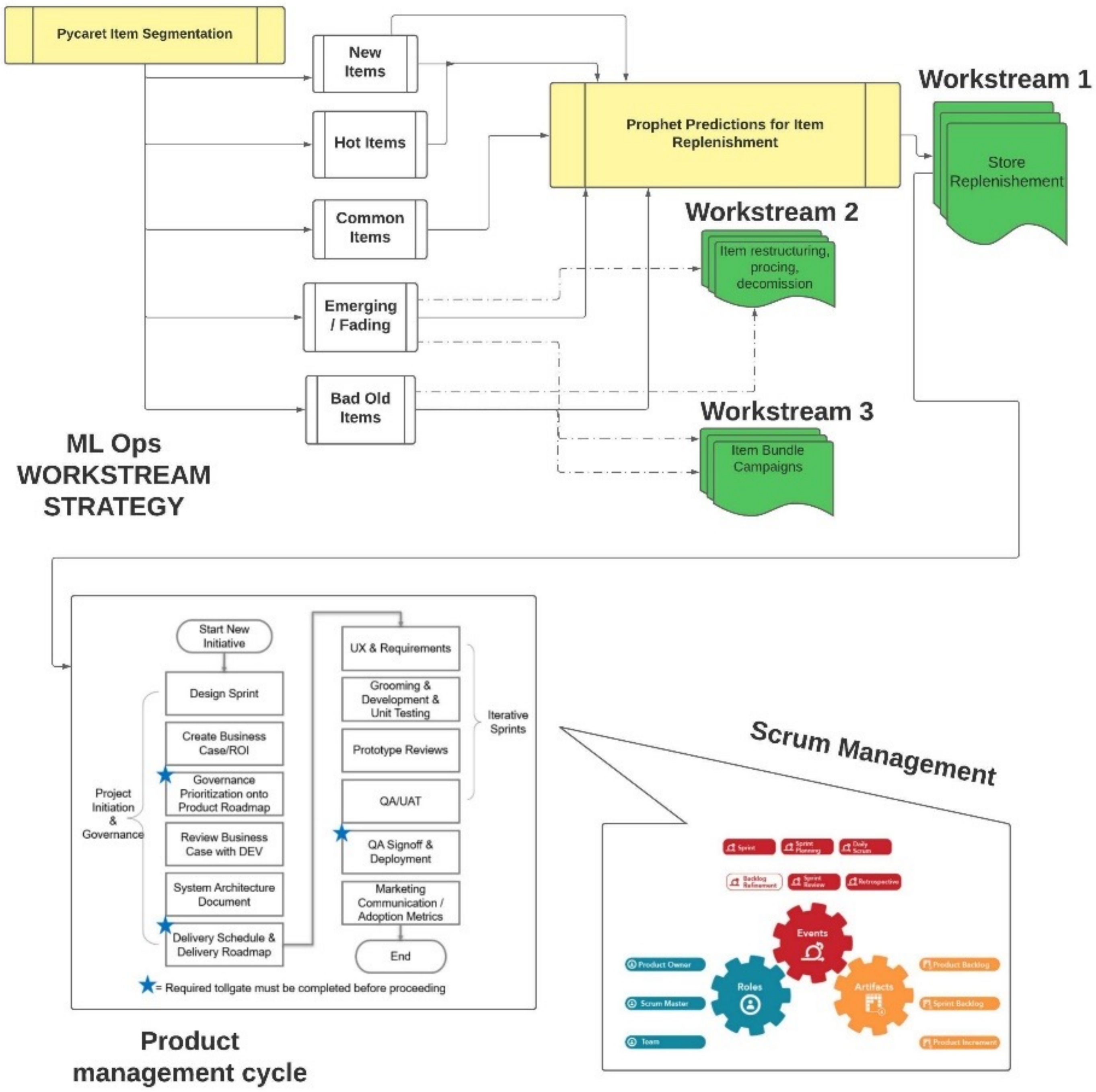
Comparision to previous year's actual performance for the quarter to the same quarter predictions for Quarter4-2016 and Quarter 1-2017.

Model to be deployment assumption is computed for Quarter4-2016 and Quarter 1-2017

	Sum of yhat	Sum of y		
2014	**$ 46 857 636,60**	**$ 46 685 289,36**		deviation
Qtr1	$ 11 376 245,96	$ 11 151 306,90	$ -224 939,05	-1,98%
Qtr2	$ 11 607 184,82	$ 11 642 899,54	$ 35 714,72	0,31%
Qtr3	$ 12 080 579,50	$ 11 962 669,30	$ -117 910,20	-0,98%
Qtr4	$ 11 793 626,32	$ 11 928 413,61	$ 134 787,29	1,14%
2015	**$ 52 117 898,86**	**$ 51 950 761,22**	$ -167 137,64	-0,32%
Qtr1	$ 12 276 813,18	$ 12 251 301,88	$ -25 511,30	-0,21%
Qtr2	$ 12 787 365,55	$ 12 802 010,39	$ 14 644,84	0,11%
Qtr3	$ 13 604 060,14	$ 13 469 445,75	$ -134 614,39	-0,99%
Qtr4	$ 13 449 659,99	$ 13 428 003,19	$ -21 656,80	-0,16%
2016	**$ 59 642 637,00**	**$ 18 245 489,01**	$ -41 397 147,99	
Qtr1	$ 14 162 388,11	$ 14 195 256,65	$ 32 868,54	0,23%
Qtr2	$ 14 655 602,24	$ 4 050 232,37	$ -10 605 369,87	
Qtr3	$ 15 440 524,44		$ -15 440 524,44	
Qtr4	$ 15 384 122,22		$ -15 384 122,22	
2017	**$ 20 225 488,57**		$ -20 225 488,57	
Qtr1	$ 15 844 147,59		$ -15 844 147,59	
Qtr2	$ 4 381 340,98		$ -4 381 340,98	
Grand Total	**293660853,7**	**231855179,6**		

Task 4 – Store supply use case (with ML Ops)

Let’s put our data to work.



MLOps Project Management: Artefacts, Infrastructure, Cadence, Adoption, Evaluation

MLOps Project Management

- Models are to branch out into multiple workstreams, the main workstream is the **Store replenishment use case**.
- **Standard Product management cycles** will be observed through to deployment and monitoring until closure of project.
- Standard **Scrum management process** will be deployed within the management of sprints.
- **Kanban boards** will be used to manage tasks using tools such as Trello or java to manage backlog.

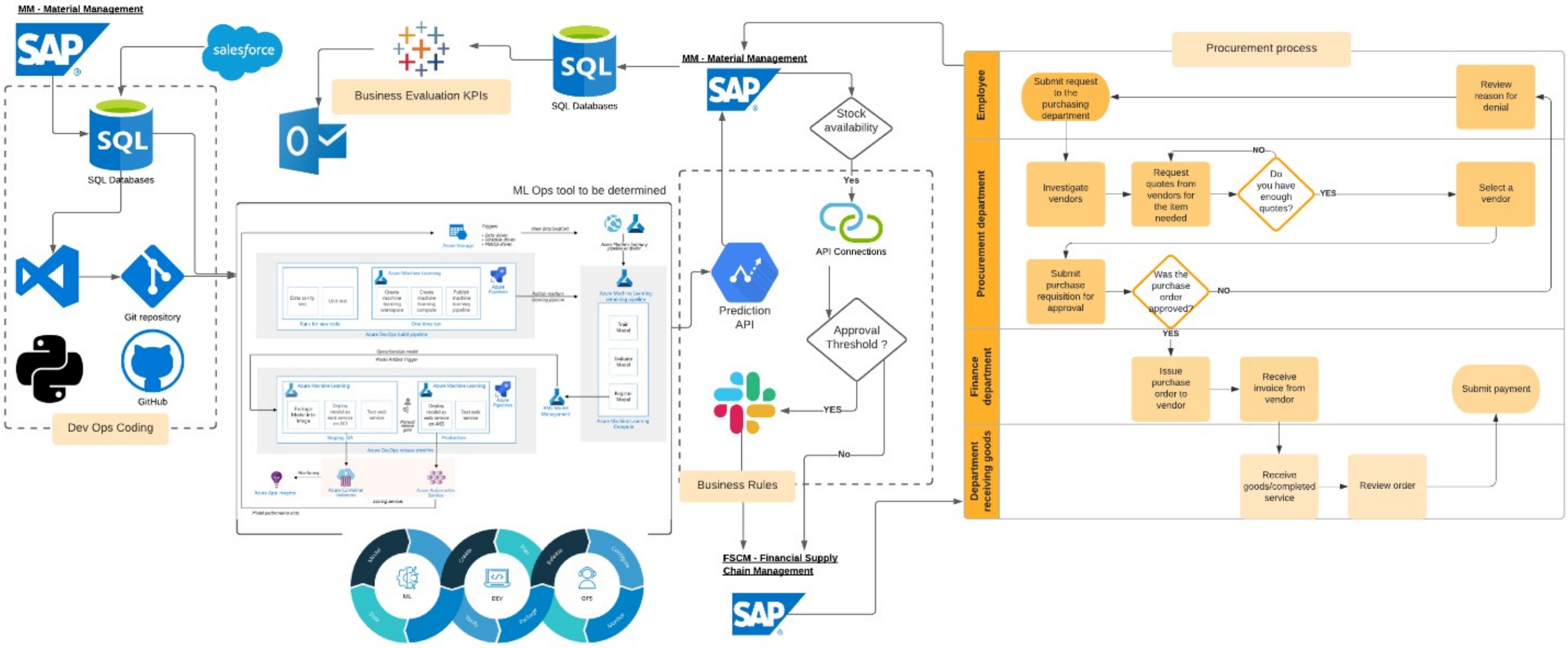
Business considerations

- **New York** as **region** comprises 43 percent of National Sales
- **Tribeca** is the top **store** nationally and in its **region**
- Sales occur the most on Saturdays and Sundays and the **least** on **Tuesdays** and **Wednesdays**

Top 12 Items Performance | Tableau Public_Top 10 of the products only comprise of around 3 percent of items sold 950,000 items sold over 3years

NUCLIO

Infrastructure: Workstream. Store Supply Use Case



Infrastructure: Workstream. Store Supply Use Case

List of Top MLOps Tools 2022 (trustradius.com)

Data sources input from **SAP** and **Salesforce** relating to sales, events and prices.

Data engineering and coding will be occurring through **VS Studio**

Git being the repository and collaboration venue

ML Ops is to be determined- in diagram; we are looking at **azure**.

Api to connect prediction models to SAP Material Management system to trigger workflows and business rules that would initiate automation.

Infrastructure: Workstream. Store Supply Use Case

Slack will be used as an approval workflow tool and a communication chanel

SAP FSCM Module will once again be consulted to trigger automated Purchase orders and payment approvals

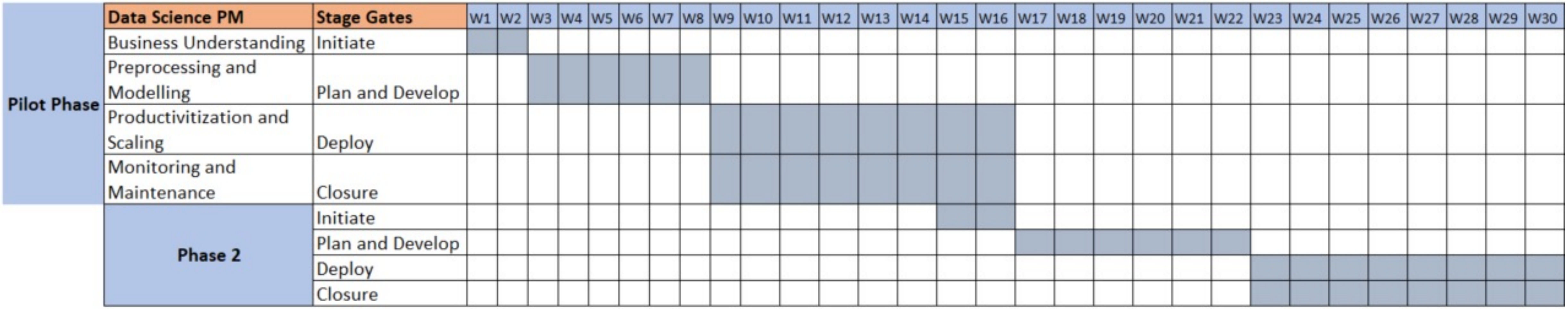
This would lead back to **SAP material Management** to update stock

Tableau will be used as dashboards to measure business evaluation KPIs

Automated reports will be made available through **Outlook.** Governance triggers will be also sent through Outlook and Slack.

Tabpy will be used to link VS Studio and Python if MLOps tool has limitations on Business reporting data

Project timeline proposal



Evaluation

Metrics to keep our model accurate and relevant.

Model evaluation

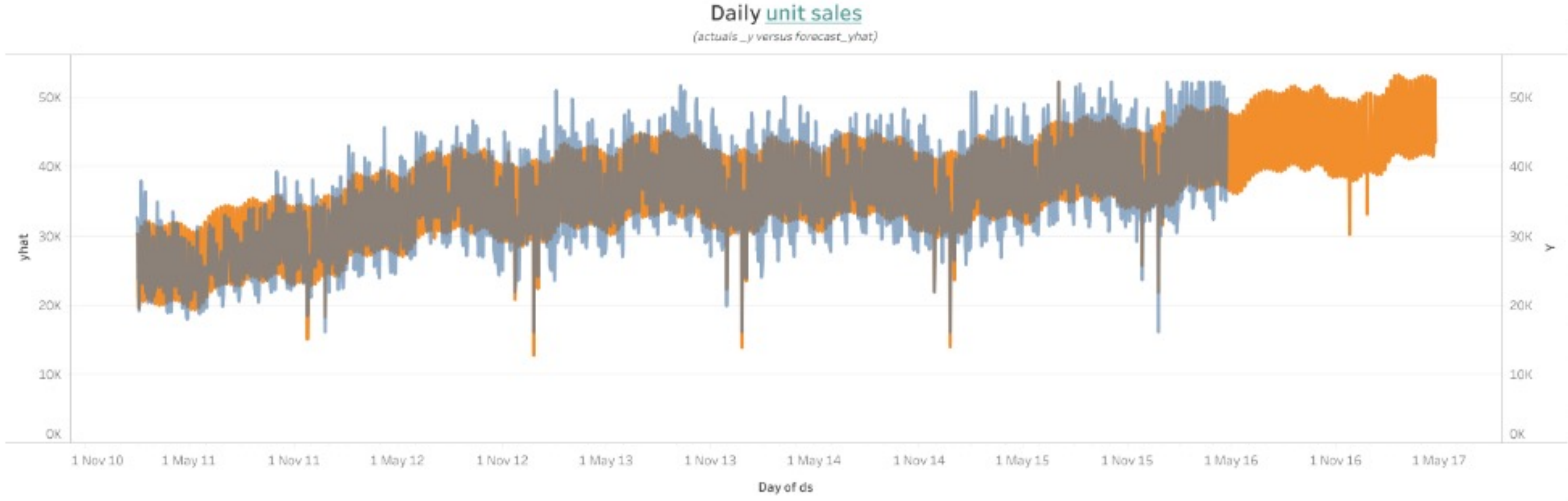
- **Accuracy metrics** will be added to the dashboard comparing projected with actuals.
- **Relevancy metrics** will be used to check if assumptions are still accurate and current.

Global Units Sales Dashboard | Tableau Public

Global Revenue and Unit Sales predictions | Tableau Public

Global Sales by Day of the week | Tableau Public

Dashboards for model evaluation



Week of ds	Y	yhat	additive_te..
22 February 2016	276,431	290,295	31,204
29 February 2016	297,516	287,525	19,395
7 March 2016	298,133	286,500	11,402
14 March 2016	287,415	287,659	9,560
21 March 2016	285,830	285,807	-5,327
28 March 2016	286,956	290,723	11,887
4 April 2016	305,045	291,689	15,217
11 April 2016	296,004	290,497	12,444
18 April 2016	285,271	288,050	4,026
25 April 2016		285,002	-8,739
2 May 2016		282,677	-19,917
9 May 2016		282,678	-22,379
16 May 2016		285,798	-14,151
23 May 2016		291,210	-974
30 May 2016		298,206	15,914
6 June 2016		301,472	11,237
13 June 2016		304,399	10,900
20 June 2016		306,293	13,394
27 June 2016		307,403	18,863
4 July 2016		306,187	15,145

Business Continuity - Governance

Accuracy measures

- Swapping models and running models in parallel to swap with better accuracy

- Retraining notifications when decrease in accuracy by 5percent.

NUCLIO

Business rules… keeping us humble

Remaining stock for an item below sales predicted for the item in next X days == trigger reorder.

Sales of an item go +20% above predictions for X days == trigger email with suggested new reorder quantity for the product and get approval through slack for redefinition and reorder. ==trigger email to Purchases, Ops and Marketing for business and Marketing actions.

Sales of an item go -20% below predictions for X days == trigger email (to Purchases, Ops and Marketing) and hold reorders until approval process through slack.

Business rules… keeping us humble

- Out of stock == trigger reorder
    - == trigger email to Purchases and Ops.
- Total sales go +20% above predictions for X days == trigger email to Purchases with items with risk of out of stock and suggested reorder.
- Total sales go -20% under predictions for X days == trigger email with next order quantities proposal for over-stocked products.

Questions?

Thank you