



Un modèle de prévision de la demande au Mexique

R- LADIES

MERCEDES SGOBBA 28.11.17

Hello
my name is

Mercedes SGOBBA

- ✓ Directrice Marketing, Digital & Data
- ✓ Ex - Danone, Coca-Cola, Shoptimise
- ✓ Certifiée 'Data Scientist' à l'Ensaï
- ✓ Certifiée 'Big Data pour entreprises numériques' à Centrale-Supelec
- ✓ Speaker: Media School Mastère 'Data Strategy', Edhec Lille MSc 'Data Analytics & Digital Business'



Plateforme web organisant des compétitions en data science

Création en 2010 par Anthony Goldbloom

Rachat par **Google** en mars 2017

Clients : facebook, AirBnB, Amazon, AXA, BNP, Criteo, Deloitte, GE, HP...

Competition 'Grupo Bimbo' en mai 2016

Kaggle Paris **meetup**

LE BUSINESS CASE



Grupo Bimbo, leader mondiale dans la production de produits « bakery »,

- leader aux USA, au Mexique, en Amérique Latine, en Espagne, Portugal et en Chine
- 100 marques



Au Mexique, Grupo Bimbo vend ses produits

- Dans > d'1 million de points de vente
- Prise de commande à la main, par les livreurs (qui récupèrent les retours)
- Beaucoup de produits ont une date de péremption 'courte' (de 5 jours à 3 mois)



AGENDA

- OBJECTIF ET METHODE
- PRESENTATION DES DONNÉES & TRAITEMENT ET MANIPULATION
- VISUALIZATION DES DONNÉES
- MODELES ET PREDICTIONS DE LA DEMANDE
- CONCLUSIONS

OBJECTIF & METHODE

- OBJECTIF & METHODE
- PRÉSENTATION DES DONNÉES &
TRAITEMENT ET MANIPULATION
- VISUALIZATION DES DONNÉES
- MODELES ET PREDICTIONS DE LA
DEMANDE
- CONCLUSIONS

OBJECTIF & METHODE

Développer un modèle de prévision de la demande quotidienne des produits Bimbo Mexique (demande de chaque produit pour chaque point de vente au net des retours) au fin de:



1. répondre correctement à la demande des consommateurs
2. réduire les rendus (invendu ou produits date courte), à date 1.7% en moyen



Langage de Programmation : R et ses outils de visualisation

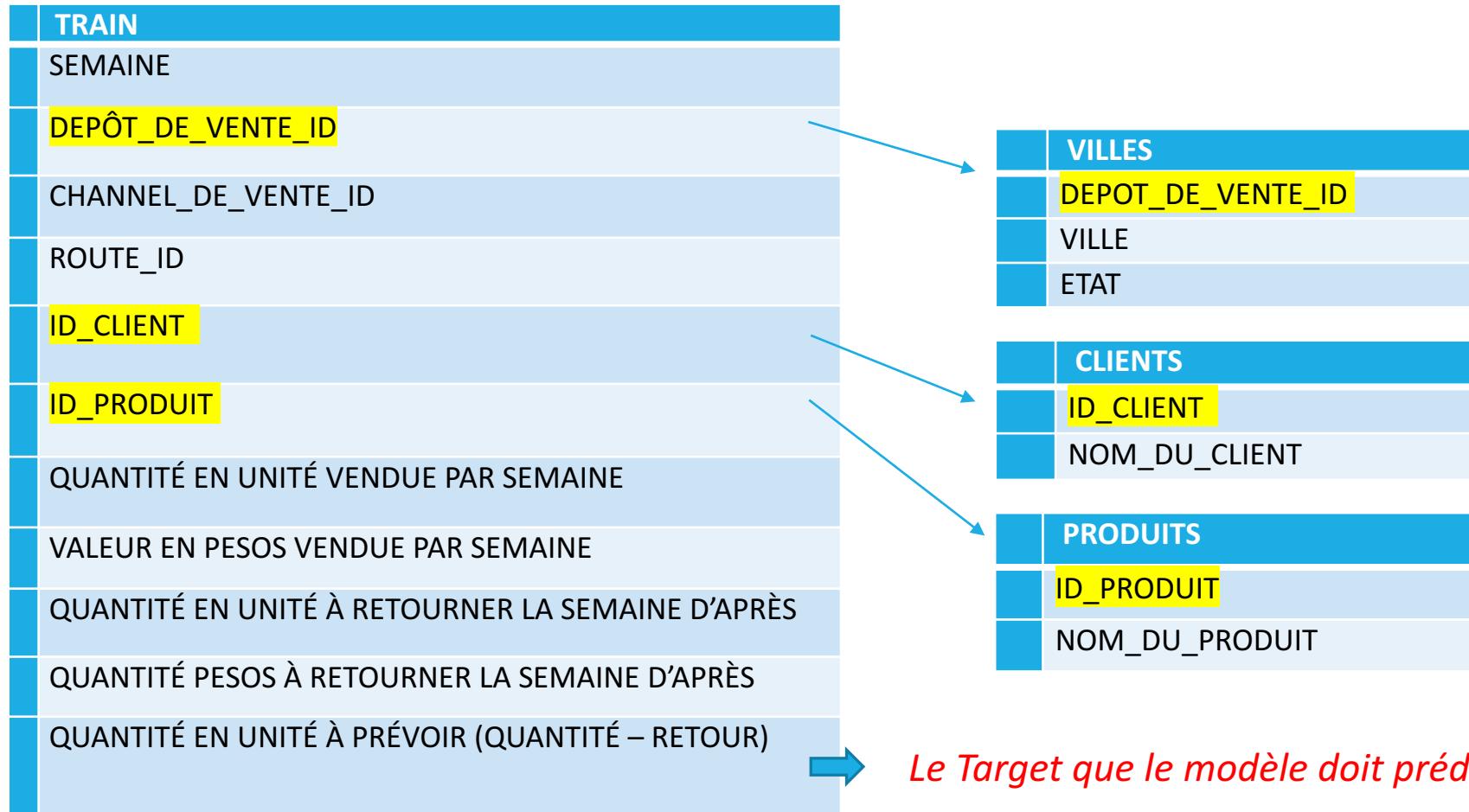


PRESENTATION DES DONNÉES, TRAITEMENT ET MANIPULATION

- ❑ OBJECTIF & METHODE
- ❑ PRESENTATION DES DONNÉES &
TRAITEMENT ET MANIPULATION
- ❑ VISUALIZATION DES DONNÉES
- ❑ MODELES ET PREDICTIONS DE LA
DEMANDE
- ❑ CONCLUSIONS



LES DONNÉES: 74 MILLIONS DE TRANSACTIONS SUR 9 SEMAINES



LE RESEAU DE DISTRIBUTION MEXICAIN

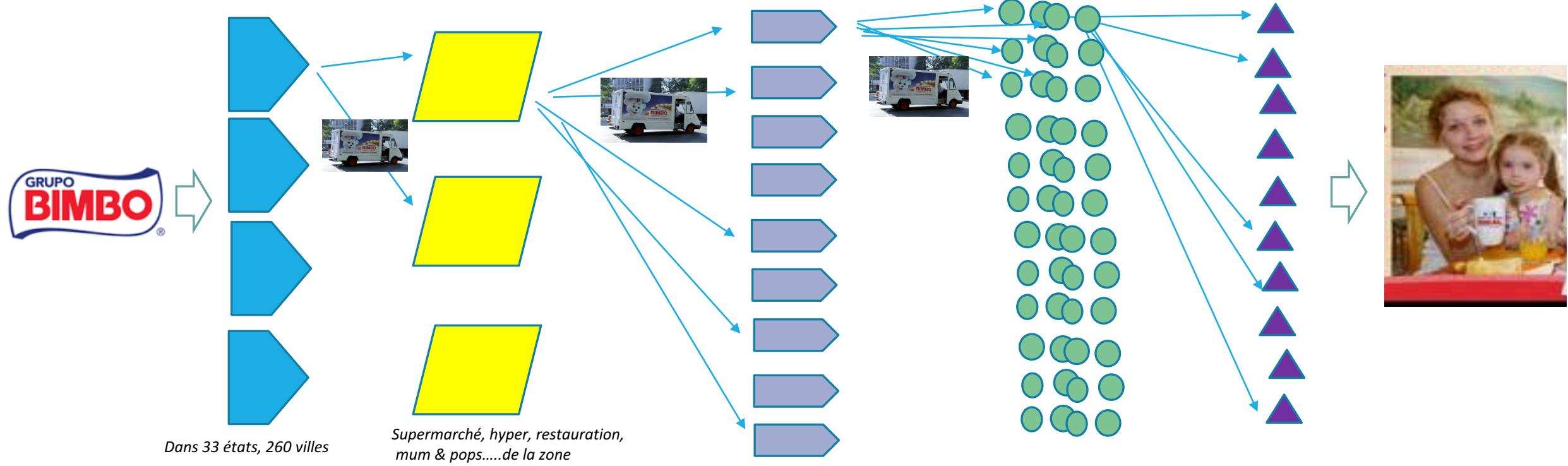
790
DEPOTS DE VENTE

9
CHANNELS

3 603
ROUTES

930 500
CLIENTS

2 592
PRODUITS



MANIPULATION DES DONNÉES

- Ouverture de ‘Train’ complexe avec ses 74 million d’observations => focus sur un échantillon de 15 millions
- Sur tous les fichiers:
 - transformation des vecteurs numériques (id client, N° route, ID produit...) en facteurs
 - Control que il n'y a pas de doublons (élimination des doublons dans le fichier ‘Clients’)
 - Fusion des fichiers entre eux
- Fichier ‘Produits’: décomposition du label de description ‘Bimboros Ext sAjonjoli 6p 480g BIM 41’.
- Calcule et ajoute des variables ‘prix moyen’ et ‘prix’ dans le fichier ‘Train’. Manquent les promotions ...

DES DONNÉES EXTERNES MANQUANTES

A ajouter « les prix » dans le fichier ‘train’. Ils manquent les promotions ...



Deux données externes qui peuvent fortement influencer la demande de produit alimentaires:

- ❖ la température
- ❖ le calendrier de la semaine/ fêtes (religieuses, nationales...)



=> aucune réponse du Grupo Bimbo et de Kaggle



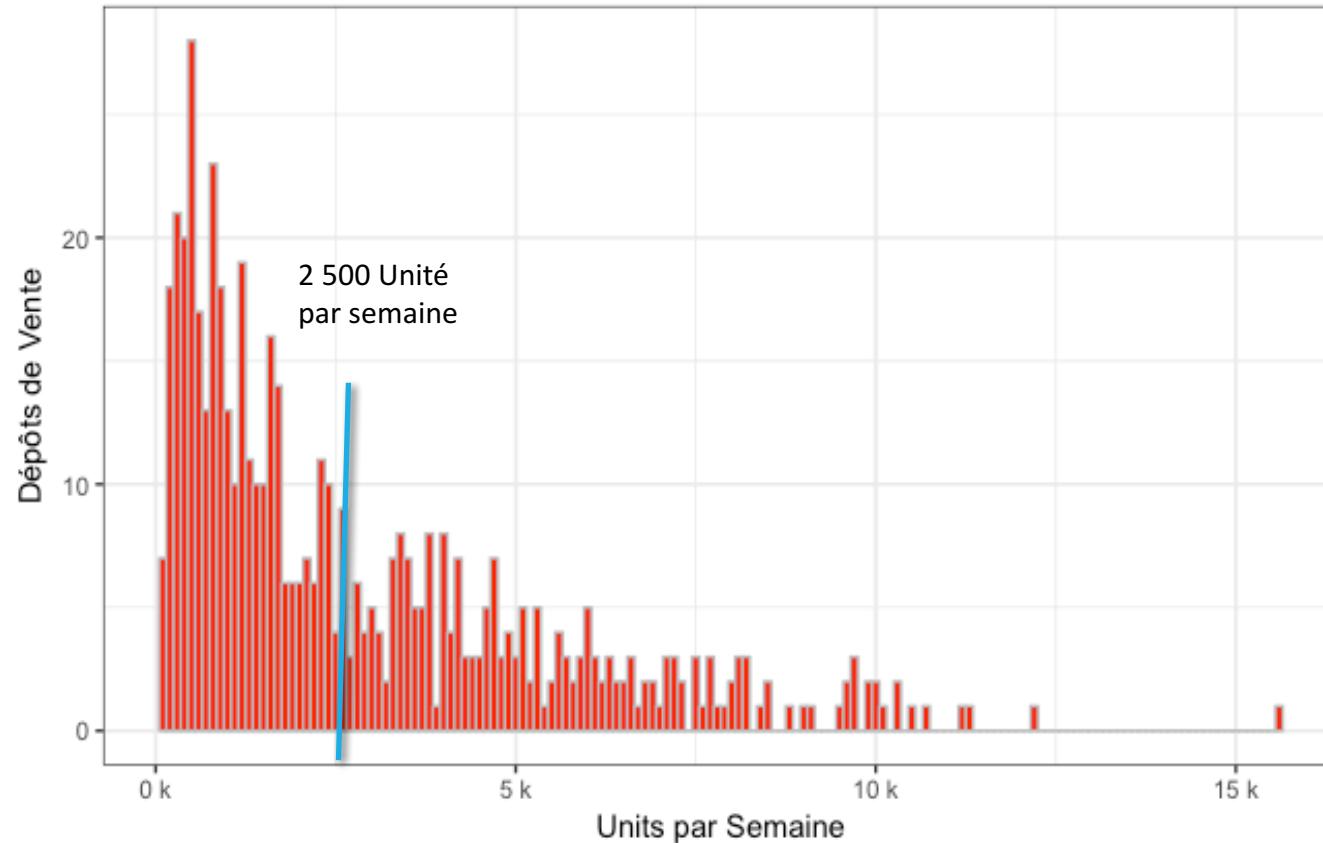
FICHIER « train » : les variables	CONTENU
1. SEMAINE	7 semaines (de la 3ème à 9ème)
2. DEPOT_DE_VENTE_ID	551 Dépôts (ils sont 552 dans le fichier Train complet)
3. CHANNEL_DE_VENTE_ID	9 Channels : de 1 à 11 (3 et 10 sont manquantes)
4. ROUTE_ID	2 327 routes (3 603 routes dans le fichier Train complet)
5. ID_CLIENT	541 097 clients (880 604 dans le fichier Train complet)
6. ID_PRODUIT	1 374 produits (1799 dans le fichier Train complet)
7. QUANT_UNIT_SEM	(valeur)
8. VALEUR_PESOS_SEM	(valeur)
9. RETOURS_UNIT_SEM_PRO	(valeur)
10. RETOURS_PESOS_SEM_PRO	(valeur)
11. QUANT_UNIT_PREVOIR	(valeur à calculer)
12. VILLE	257 villes (260 dans le fichier Train complet)
13. ETAT	33 états
14. PRIX	(valeur)
15. PRIX MOYEN	(valeur)
16. poids	(valeur)
17. volume	(valeur)
18. N° pièces	(valeur)
19. nom_du_produit	(valeur)
20. marque	(valeur)
21. num_marque	(valeur)

VISUALIZATION DES DONNÉES

- ❑ OBJECTIF & MÉTHODE
- ❑ PRÉSENTATION DES DONNÉES &
TRAITEMENT ET MANIPULATION
- ❑ **VISUALIZATION DES DONNÉES**
- ❑ MODÈLES ET PREDICTIONS DE LA
DEMANDE
- ❑ CONCLUSIONS

DEPÔT DE VENTE: QUANTITÉS VENDUES PAR SEMAINE

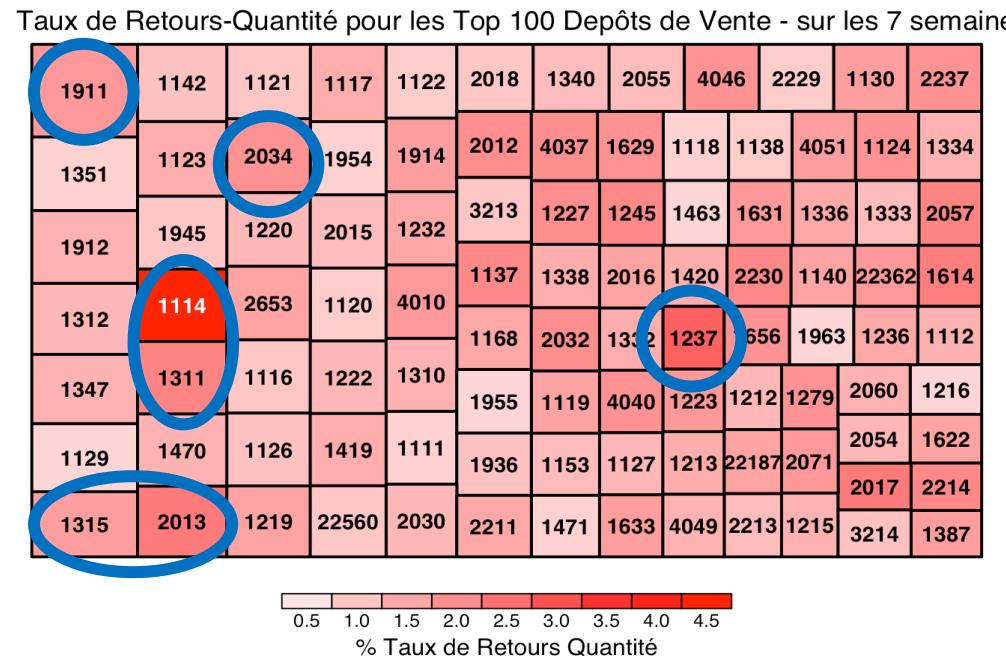
Quantités vendues par Semaine par Dépôts de Vente



- La distribution des 790 Dépôts de vente (classées par Q/semaine) ressemble à une distribution de Poisson
- La majorité vendent pour < 2500 unités, mais beaucoup vendent jusqu'à >15 000 unités

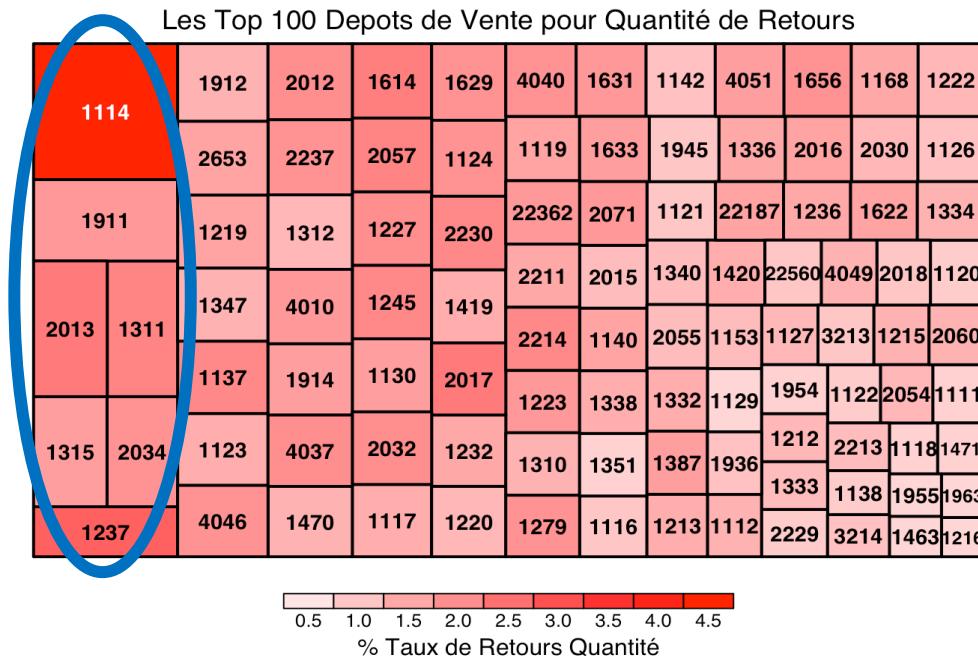
DISTRIBUTION EQUILIBRÉ DES TOP 100 DEPÔTS DE VENTE

100 DÉPÔTS (SUR 940) REPRÉSENTENT 47%
DES VOLUMES ET 52% DES RETOURS



Top 100 des Dépôts classés par Quantité Vendue

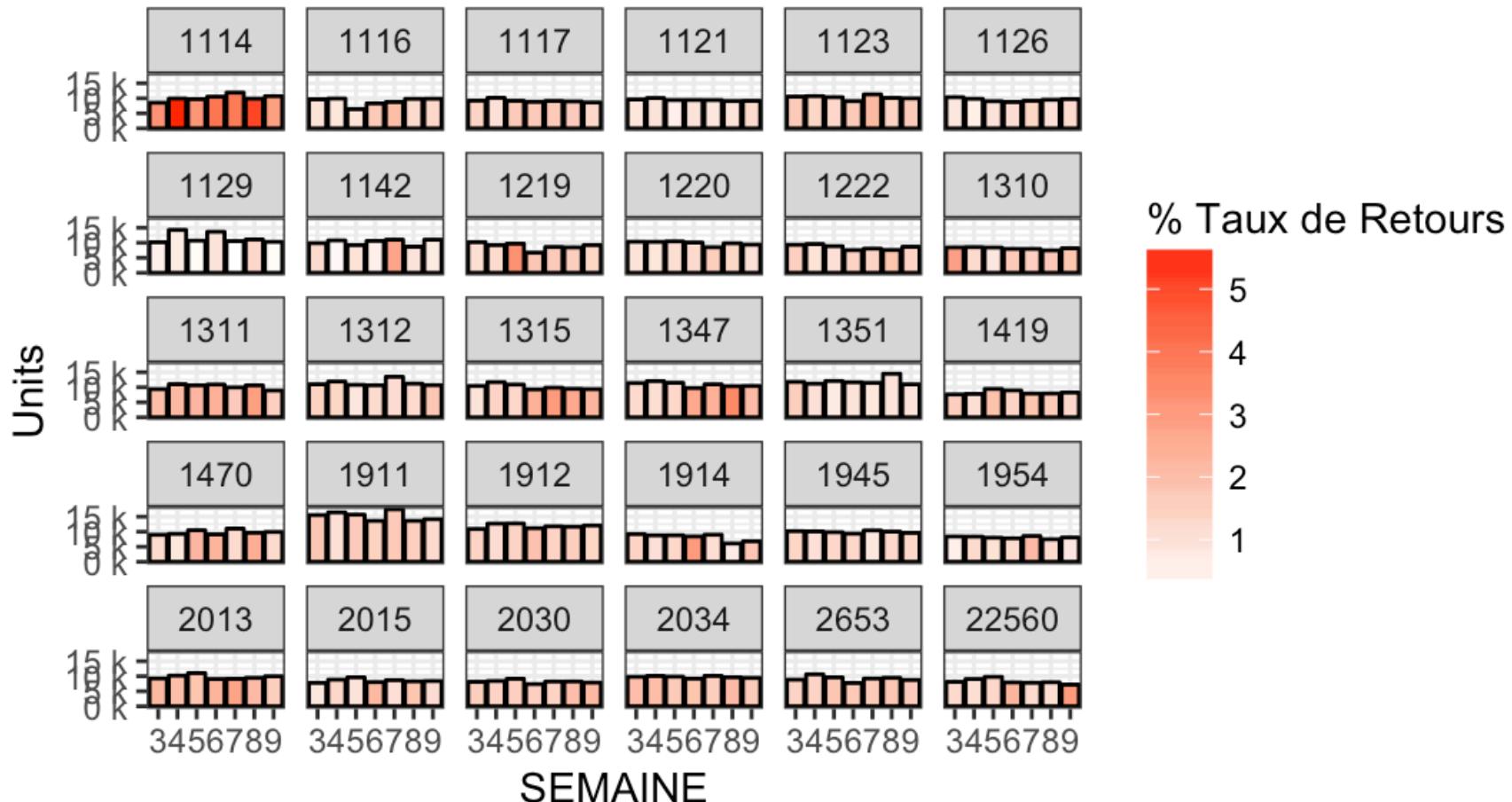
7 DÉPÔTS DE VENTE REPRÉSENTENT 9% DES
RETOURS



Top 100 des Dépôts classés par Quantité Rendue

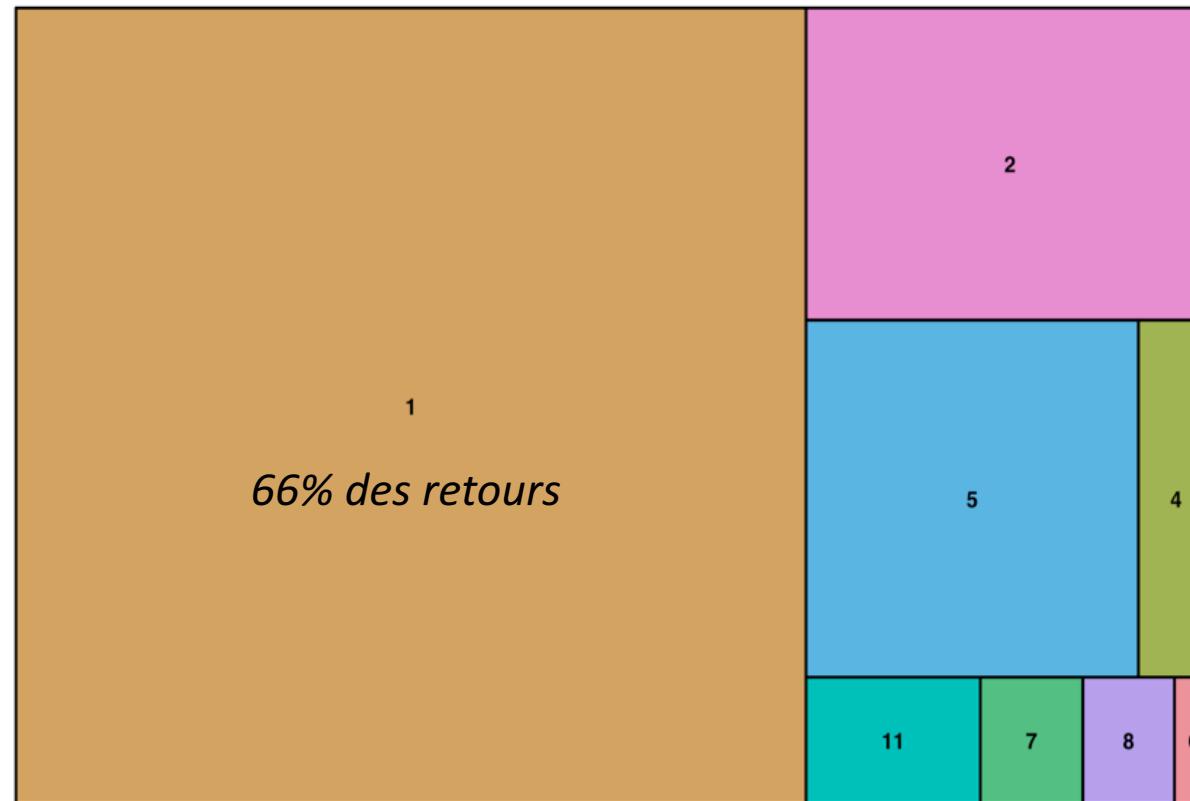
DISTRIBUTION ÉQUILIBRÉ DES RETOURS PAR SEMAINE, DANS LES TOP 30 DEPÔTS DE VENTE

% Taux de Retours Quantité par semaine
des Top 30 Dépôts de Vente



RÉPARTITION DES RETOURS SELON LES 9 CHANNELS

Le Channel 1
représente 72%
des volumes et
66% des retours

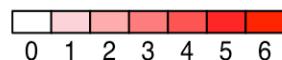


LES ROUTES

TOP 100 ROUTES selon les Ventes

Les TOP 100 routes selon les ventes (sur 3 603) représentent 29% des volumes totaux

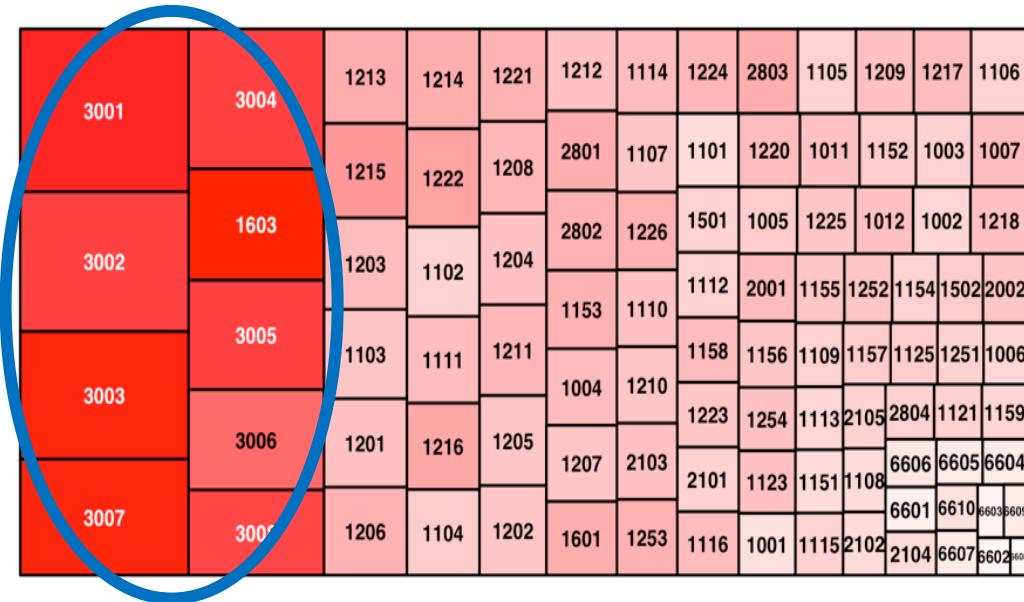
1101	6603	6604	1213	1002	1111	1113	1210	1004	1151	1209	1221	6607
6601	6602	1205	1001	3003	1005	1501	1114	3007	3006	1108	6608	900
3002	3001	1204	1214	1211	2103	2802	6609	1156	3005	1216	6610	1225
1201	6605	1212	1107		1152	1224	1155	1217	1011	1006	1115	1007
1103	1105	3004	1206	1112	1222	1154	1226	1602	1251	1157	2002	1218
1102	1203	1207	6606	2101	1110	1502	2801	1204	1121	2105	1116	1252
1104	1202	1106	1208	1003	2102	1220	1223	2001	1253	1254	1012	6611



% Taux de Retour Quantité

TOP 100 ROUTES selon les retours

Les TOP 100 routes selon les retours (sur 3 603) représentent 6% des volumes totaux et 37% des retours. Les premières 9 routes ont un taux de retours jusqu'à 7% (sur une moyenne de 1.7%)

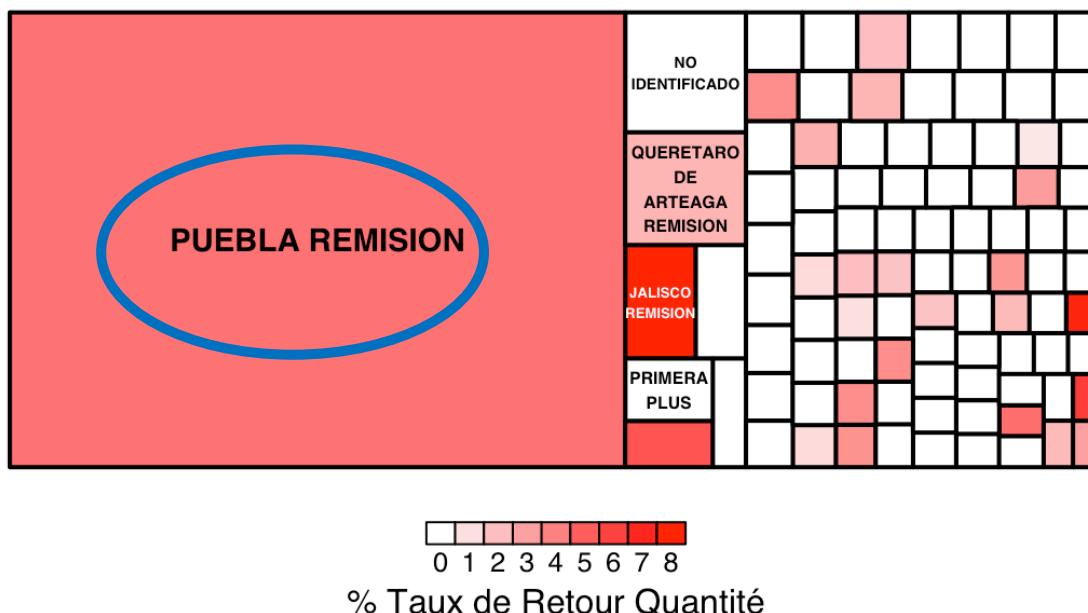


% Taux de Retour Quantité

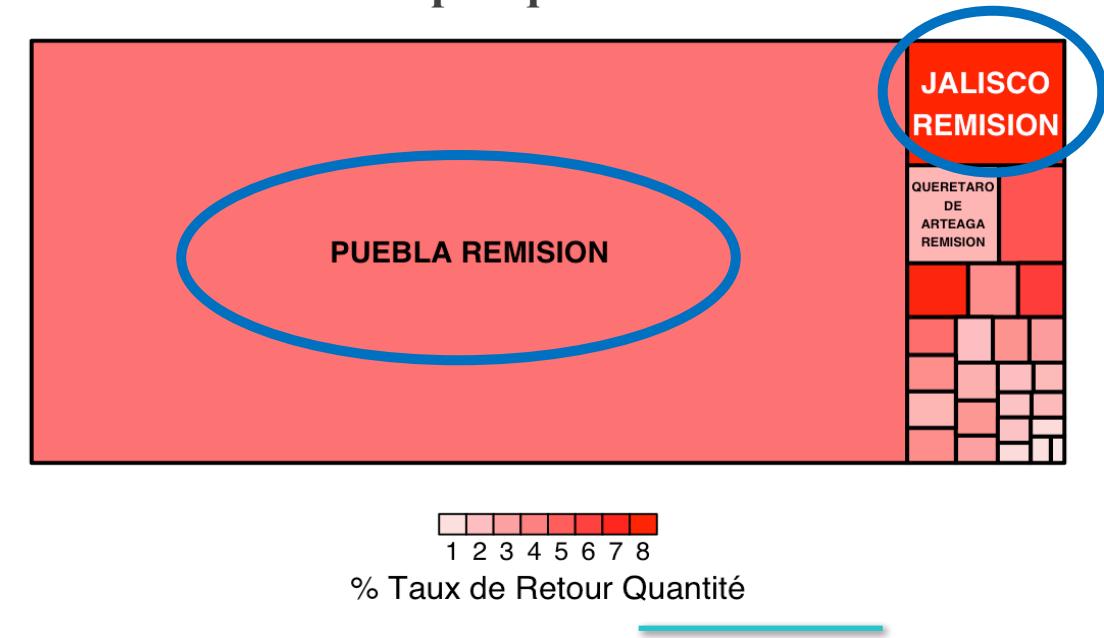
UN CLIENT TRÈS INFLUENT

- Enorme fragmentation des 930 500 clients . TOP 100 représente que 6% des quantités vendues et 23% des quantités rendues
- 'PUEBLA REMISION' représente la $\frac{1}{2}$ des quantités vendue par les TOP 100 et 2/3 des rendus des TOP 100
- Puebla, Quartegario et Jalisco sont les moins performants dans le TOP 100

Top 100 clients par Quantité Vendue



Top 100 clients par Quantité vendue,
classés par quantité rendue

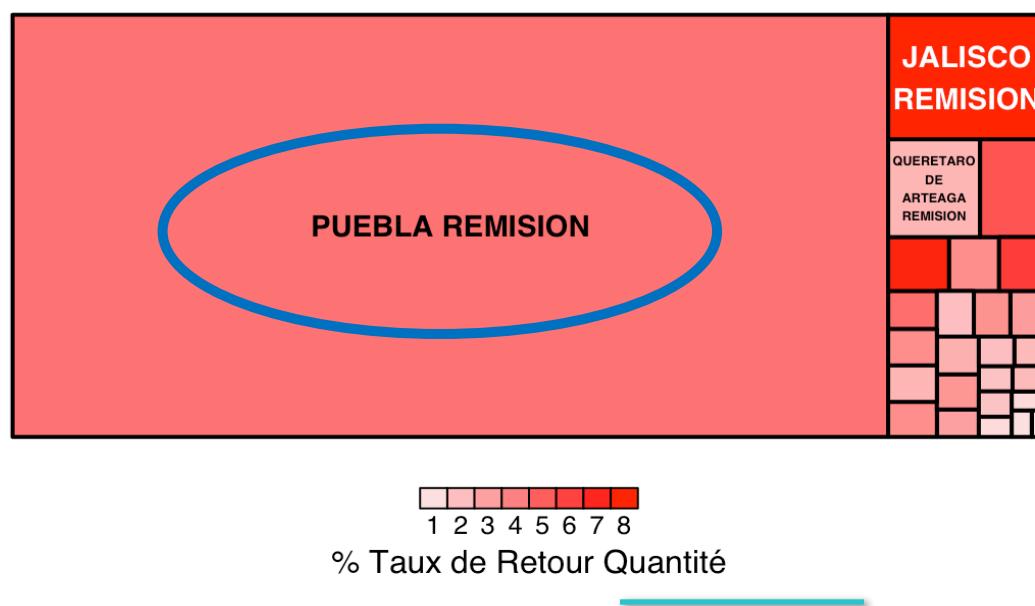


LES TOP 100 CLIENTS POUR RETOUR (surprise...)

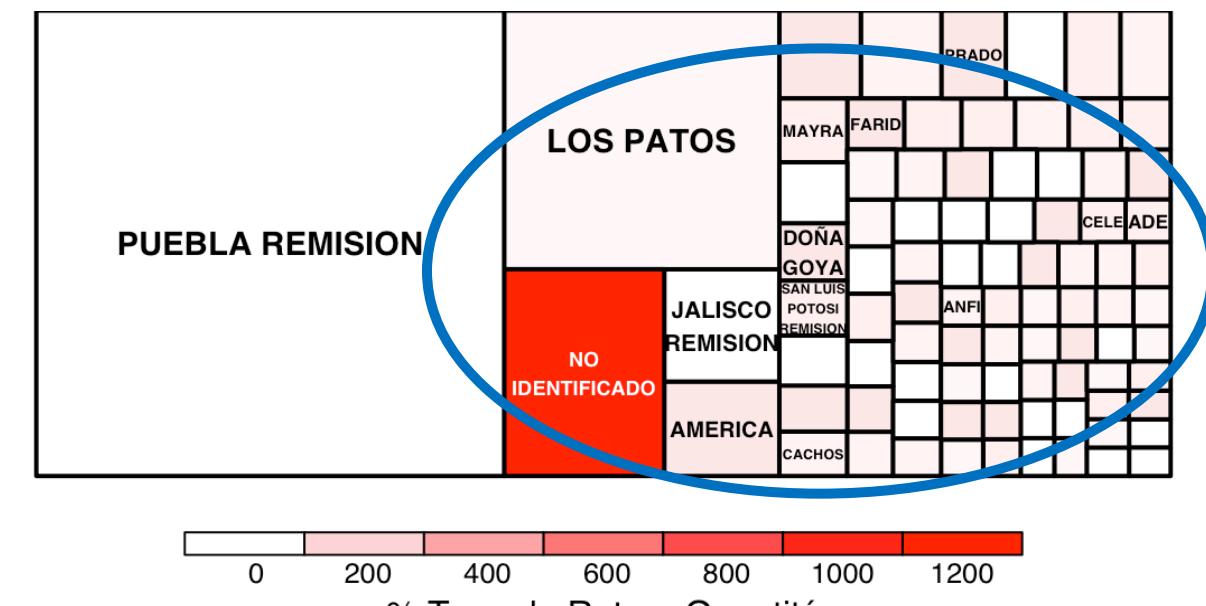


- Il y a plein de petits clients – hors du top 100 – avec taux de retour hors normes.
- Surtout il y a un client NO IDENTIFICADO...

TOP 100 CLIENT QUANTITE' VENDUE
classés par quantité rendue



TOP 100 CLIENTS PAR QUANTITÉ RENDUE



LISTE DES PREMIERES 12 CLIENTS POUR RETOUR

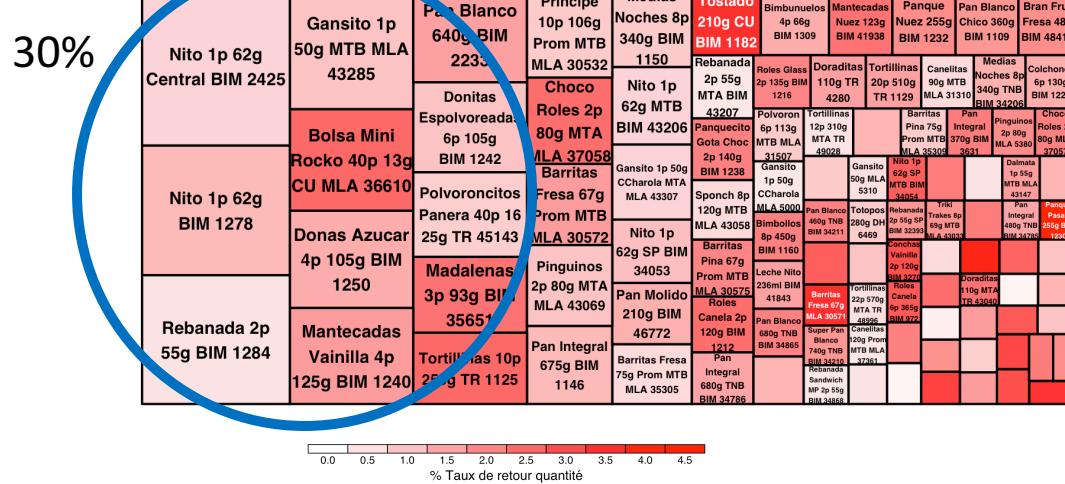
- Sur les premiers 12 clients par retour, 8 clients présentent des anomalies

ID_CLIENT	Units	Return_Units	Return_Rate	NOM_DU_CLIENT
1653378	366591	17310	4.508975	PUEBLA REMISION
24102610	375	8064	95.556346	NO IDENTIFICADO
32215234	35	1508	97.731692	COBACH 15 COPERATIVA 2
44367822	46	1193	96.287328	NO IDENTIFICADO
5652850	10469	917	8.053750	JALISCO REMISION
64392197	16	902	98.257081	NO IDENTIFICADO
74344485	113	623	84.646739	NO IDENTIFICADO
81046764	82	530	86.601307	VAZQUEZ
9653039	17807	421	2.309634	QUERETARO DE ARTEAGA REMISION
10653058	154	345	69.138277	SAN LUIS POTOSI REMISION
1157450	26	338	92.857143	JUGOS Y LICUADOS
124241981	8	333	97.653959	NO IDENTIFICADO

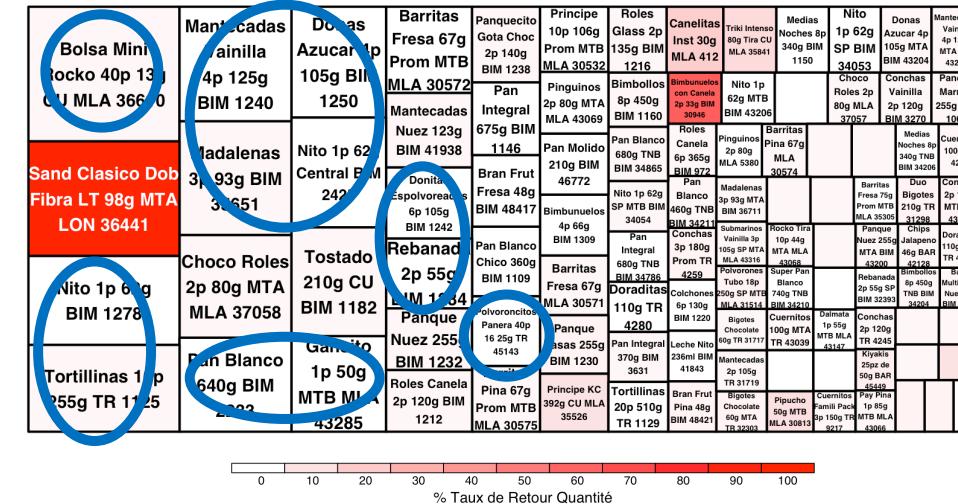
LES PRODUITS: VARIABLE INFLUENTE

- Forte concentration: sur 2 592 produits, les premières 12 produits représentent 30% des quantité vendues et 26% des Rendus
- Dans le ranking des TOP 100, bonne gestion du taux de retour en quantité (< moyenne 5%)

**TOP 100 PRODUITS PAR QUANTITÉ VENDUE
(74% DES TOT. QUANTITÉS VENDUES)**



**TOP 100 PRODUITS PAR QUANTITÉ RENDUE
(69% DES QUANTITÉS RENDUES)**



CONSIDERATIONS APRÈS VISUALIZATION

En conclusion, la quantité des retours est fortement influencée par :

1. Les clients (1 client : « Puebla Remision » et plein de petits clients). Quid fraude? **PRIORISÉ**
2. Les produits (surtout 3 produits : Nito leche, Bolsa Madalenas 3p 93g BIM, Mantecadas Vainilla 4p125 BIM) **PRIORISÉ**

A tenir d'œil:

1. le 1^{er} Channel
2. 9 routes

Les variables Etats, les Villes, les Dépôts de Ventes semblent impacter moins les retours

MODELES ET PREDICTIONS DE LA DEMANDE

- OBJECTIF & METHODE
- PRÉSENTATION DES DONNÉES &
TRAITEMENT ET MANIPULATION
- VISUALIZATION DES DONNÉES
- MODELES ET PREDICTIONS DE LA
DEMANDE
- CONCLUSIONS

1. TROIS MODELES LINAIRES SELON LA VARIABLE PRIX

ML: ~ prix + prix moyen

L'Adjusted R-squared est:

0.001611

ML: ~ prix moyen

L'Adjusted R-squared est:

0.001545

ML: ~ prix

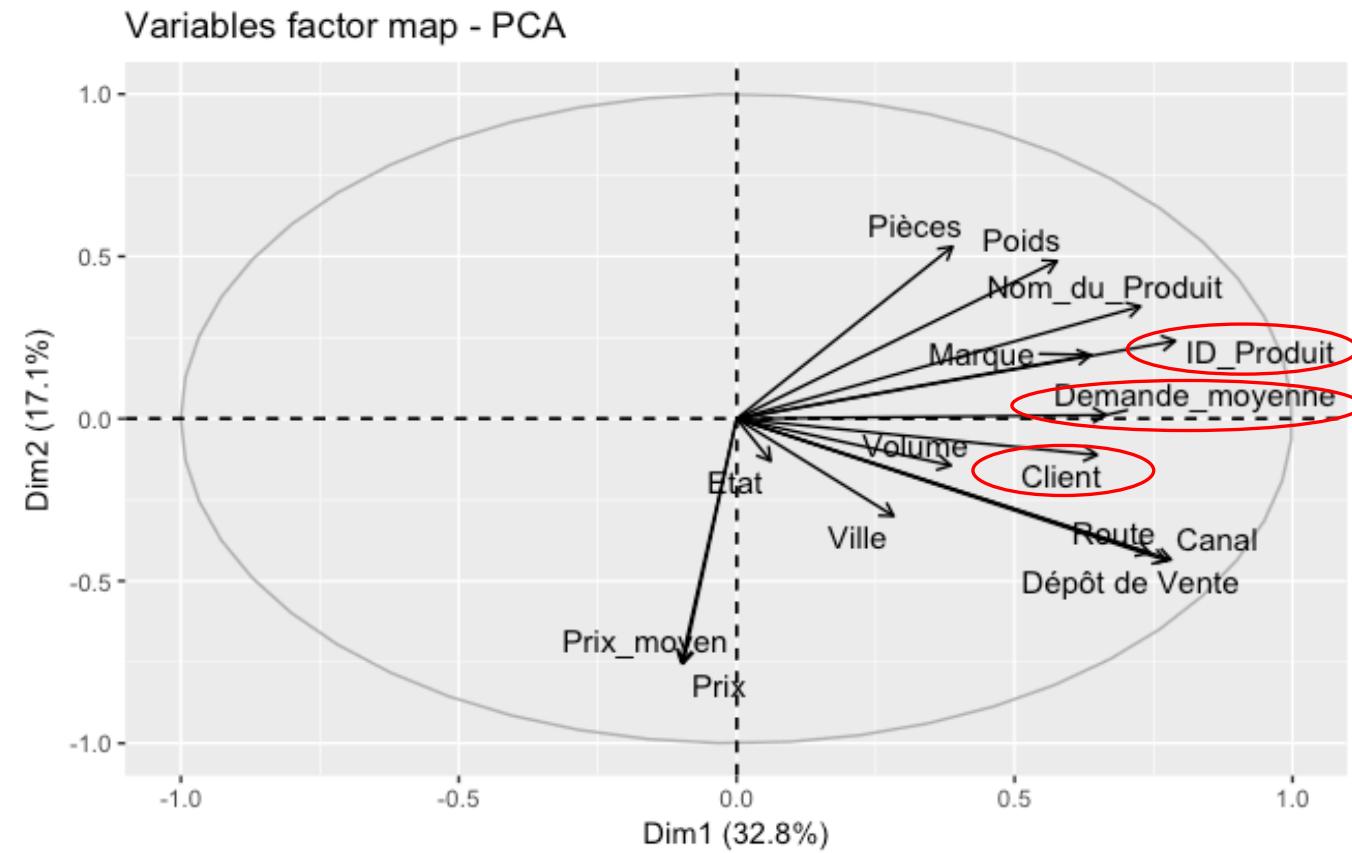
L'Adjusted R-squared est:

0.001512

NO GO....

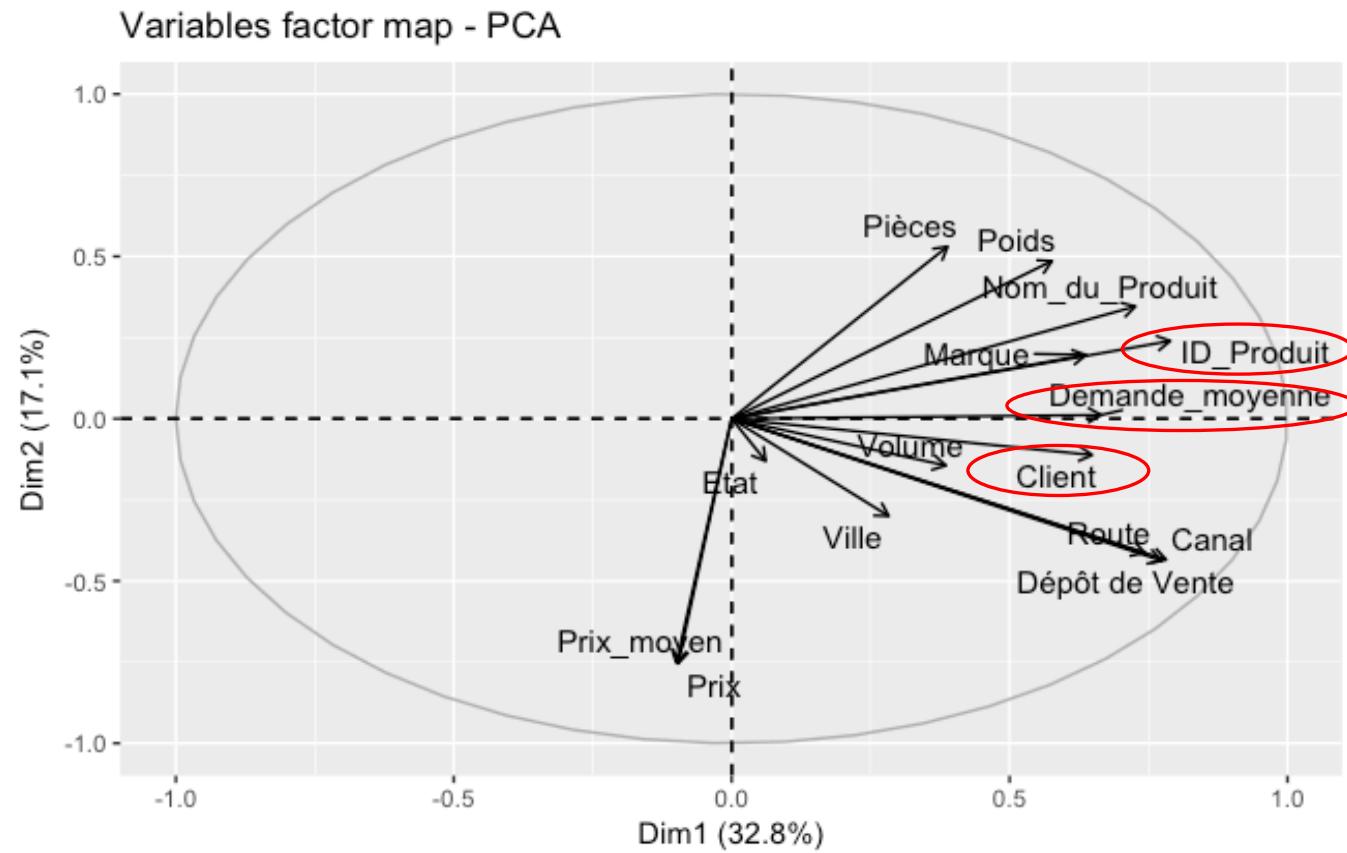
2. ACP POUR COMPRENDRE LA RELATION ENTRE LES VARIABLES

- ID Produit et Client très liés à la Demande moyenne



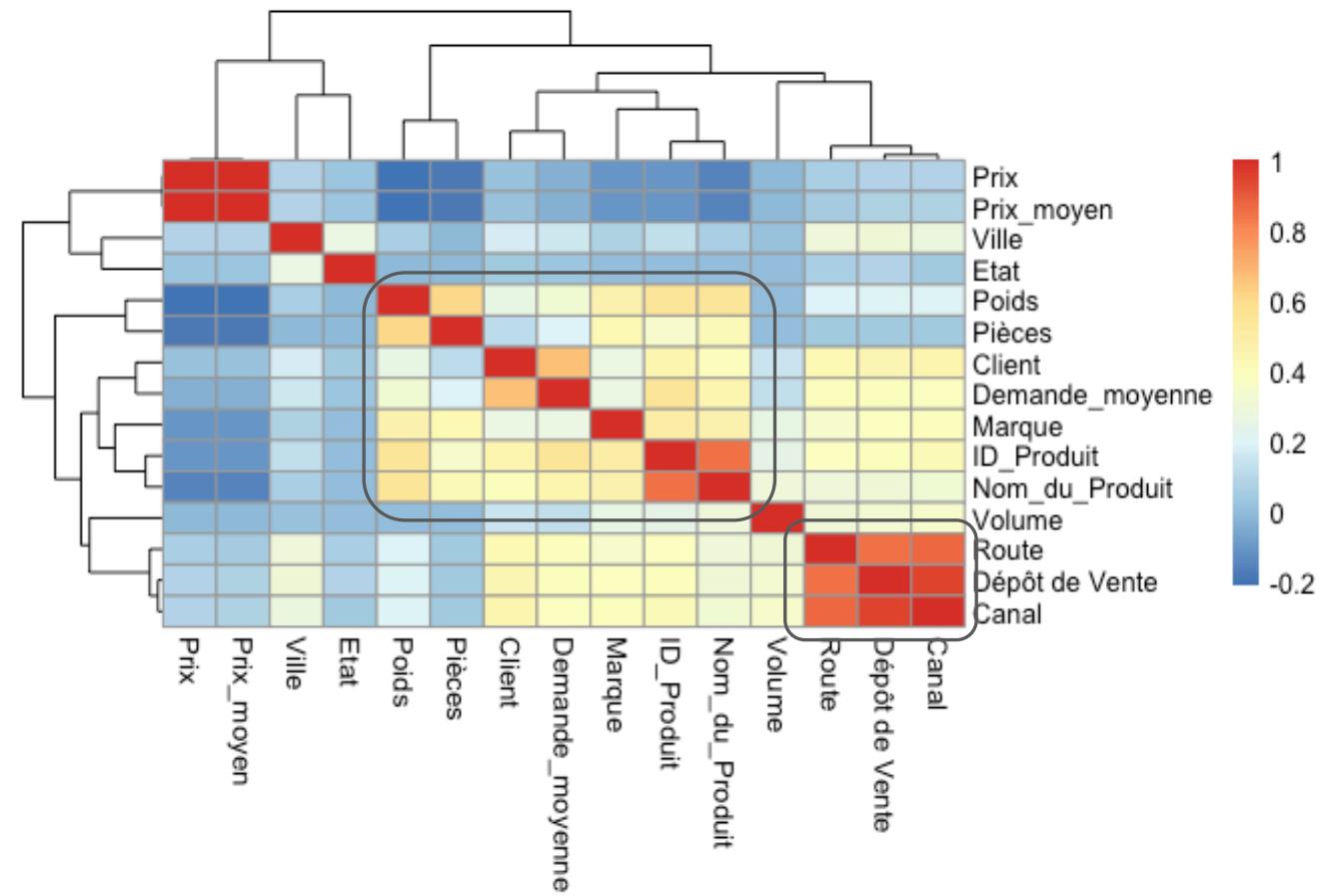
ACP

- ID Produit et Client très liés à la Demande moyenne



MATRICE DE CORRELATION

- Indice de corrélation > 0.5 prédictif d'une relation linéaire entre les variables et Quantité à prévoir



2. DEUX AUTRES MODÈLES LINAIRES ET EVALUATION

- Dans le fichier ‘Train’ , substitution de la valeur des variables factor avec la médiane de la variable (plus performante pour distribution style ‘Poisson’)
- Évaluation sur jeux de données: 75% Train et 25% Test

- ML ~client et produit

- Client et Produit: variables significatives

- R2 adjusted: **0,56**

- ML ~toutes les variables

- Toutes les variables sont significatives

- R2 adjusted: **0,58**

- Pas concluant RandomForest sur toutes les données

Evaluation du 1^{er} et du 2^{ème} MODÈLE LINAIRe

2. En fonction du client et produit

```
Console ~/Documents/Certificat Big Data Centrale Supelec/Memoire Big Data/Bimbo
lm(formula = Demande_moyenne ~ ., data = new_train)

Residuals:
    Min      1Q   Median      3Q      Max 
-2130.4   -2.3   -1.1     0.2  3591.7 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.1785994 0.0255292 6.996 2.64e-12 ***
Client       0.8879613 0.0009597 925.239 < 2e-16 *** 
ID_Produit   0.5928358 0.0013003 455.917 < 2e-16 *** 
Prix        -0.3900873 0.0456700 -8.541 < 2e-16 *** 
Prix_moyen   0.3234599 0.0457135  7.076 1.49e-12 *** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.31 on 1499995 degrees of freedom
Multiple R-squared:  0.551,    Adjusted R-squared:  0.5509 
F-statistic: 4.601e+05 on 4 and 1499995 DF,  p-value: < 2.2e-16
```

3. En fonction de toutes les variables

```
Console ~/Documents/Certificat Big Data Centrale Supelec/Memoire Big Data/Bimbo data
```

Residuals:

Min	1Q	Median	3Q	Max
-2048.7	-2.4	-1.0	0.5	3570.2

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.1285685	0.0984889	1.305	0.191753
Semaine	NA	NA	NA	NA
`Dépôt de Vente`	0.3435380	0.0055781	61.587	< 2e-16 ***
Canal	-0.8662183	0.0063516	-136.377	< 2e-16 ***
Route	0.7745040	0.0026016	297.706	< 2e-16 ***
Client	0.8535551	0.0009789	871.981	< 2e-16 ***
ID_Produit	0.7059849	0.0021947	321.671	< 2e-16 ***
Ville	0.0229373	0.0063634	3.605	0.000313 ***
Etat	0.1611751	0.0283411	5.687	1.29e-08 ***
Poids	-0.2164987	0.0026362	-82.125	< 2e-16 ***
Volume	-0.5408492	0.0087570	-61.762	< 2e-16 ***
Pièces	0.5110669	0.0043739	116.844	< 2e-16 ***
Nom_du_Produit	-0.1089734	0.0025680	-42.435	< 2e-16 ***
Marque	-0.1607688	0.0032457	-49.533	< 2e-16 ***
Prix	-0.6625268	0.0449389	-14.743	< 2e-16 ***
Prix_moyen	0.5863959	0.0449419	13.048	< 2e-16 ***

Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

Residual standard error: 14.72 on 1499985 degrees of freedom

Multiple R-squared: 0.5847, Adjusted R-squared: 0.5847

F-statistic: 1.509e+05 on 14 and 1499985 DF, p-value: < 2.2e-16

474ème sur 1 600 sur Kaggle

470	↑158	liaoyuanZhu	0.48818	1	Wed, 03 Aug 2016 09:18:39
471	↑158	Farouk TANGAO	0.48818	40	Wed, 03 Aug 2016 23:12:33 (-0.6h)
472	↑158	mikepb	0.48818	9	Thu, 25 Aug 2016 02:31:59 (-21.2d)
473	↑17	ShHs	0.48818	16	Tue, 30 Aug 2016 23:00:59 (-3.7d)
474	↑18	Eduardo Franco	0.48818	24	Tue, 30 Aug 2016 21:32:30 (-0.6h)
-		Mercedes	0.48818	-	Mon, 16 Jan 2017 16:01:11 Post-Deadline
Post-Deadline Entry					
If you would have submitted this entry during the competition, you would have been around here on the leaderboard.					
475	↑18	Reza	0.48818	7	Fri, 26 Aug 2016 18:38:41 (-2.4h)
476	↑18	RajatKumar	0.48818	9	Sat, 27 Aug 2016 12:14:24 (-0h)
477	↑18	Polarbearthu	0.48818	3	Sat, 27 Aug 2016 21:22:25
478	↑18	NavdeepGill	0.48818	24	Mon, 29 Aug 2016 00:30:56 (-16.6h)
479	↑18	vanilla-ic	0.48818	5	Sun, 28 Aug 2016 08:46:35

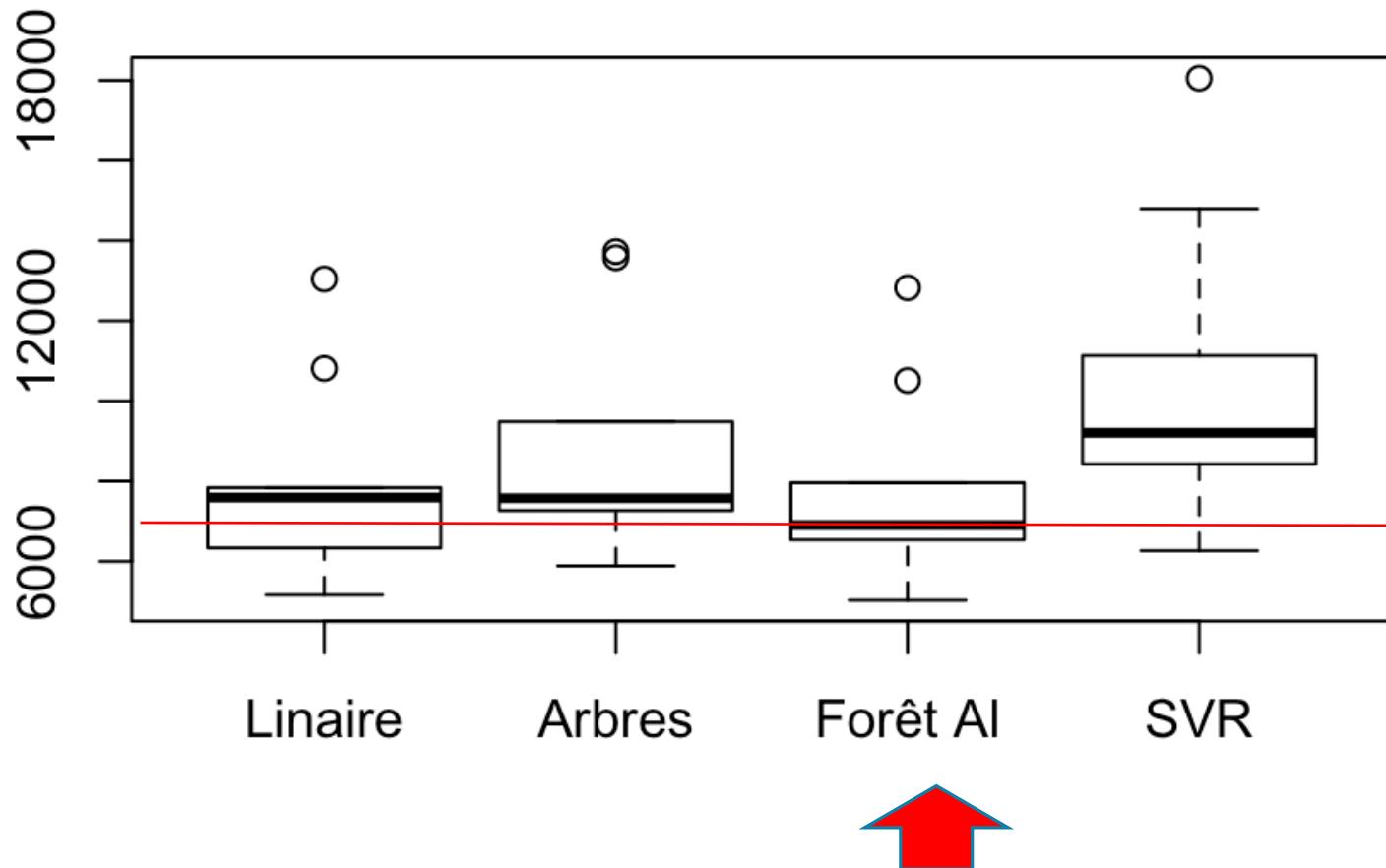
Pour aller plus loin

Focus que sur Puebla Remision (3% quantité vendue, 15% quantité rendue)

Reformulation sans les « possibles fraudeurs »

Comparaison de 4 modèles sur 2 cas (1/Puebla Rémission, 2/sans Fraudeurs)

- Le même boxplot dans les deux cas
- Forêts Aléatoires à choisir



CONCLUSIONS

- ❑ OBJECTIF & METHODE
- ❑ PRÉSENTATION DES DONNÉES &
TRAITEMENT ET MANIPULATION
- ❑ VISUALIZATION DES DONNÉES
- ❑ MODELES ET PREDICTIONS DE LA
DEMANDE
- ❑ CONCLUSIONS

CONCLUSIONS

-
- Essentielle la collaboration entre Data Scientist – Metier
 - On recommande de choisir un model RandomForest (en enlevant les « possibles fraudeurs »). Tester un model xgboost pour aller plus loin
 - Dans un contexte réel, il aurait fallu inclure - dans les variables - la température, des éléments du calendrier et les promotions réalisées
 - Bénéfice économique estimé, à date :
 - GAIN sur 2 mois: - 1 million de Pesos dans les ristournes / - 5 millions d'unités retournées
 - GAIN sur 1 ans: - 6 millions de Pesos de ristournes/ - 30 millions d'unités

MERCI

mercedes.sgobba@gmail.com

06 86 26 00 37

@MercedesSgobba