

# Club Power BI

La communauté Power BI francophone

**Nous démarrons à 18h05, vous pouvez vous présenter  
dans l'espace de conversation**



**/Club-Power-BI**



**@ClubPowerBI**



**/ClubPowerBI**



**/ClubPowerBI**



**/ClubPowerBI**



**@ClubPowerBI**



# Club Power BI

La communauté Power BI francophone

Téléchargez les fichiers de démonstration sur  
notre GitHub



/Club-Power-BI



@ClubPowerBI



/ClubPowerBI



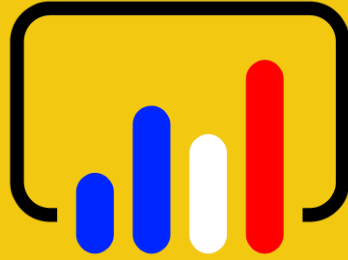
/ClubPowerBI



/ClubPowerBI



@ClubPowerBI



# Club Power BI

## Augmentez l'intelligence de vos rapports Power BI

**Avril 2020 – Nantes - Marseille**

**En partenariat avec Wild Code School**



@ClubPowerBI

# WILD TALK

Augmentez l'intelligence  
de vos rapports Power BI



avec Paul PETON et Joël CREST

27 avril 2020 - 18h



# **Avant de commencer ...**

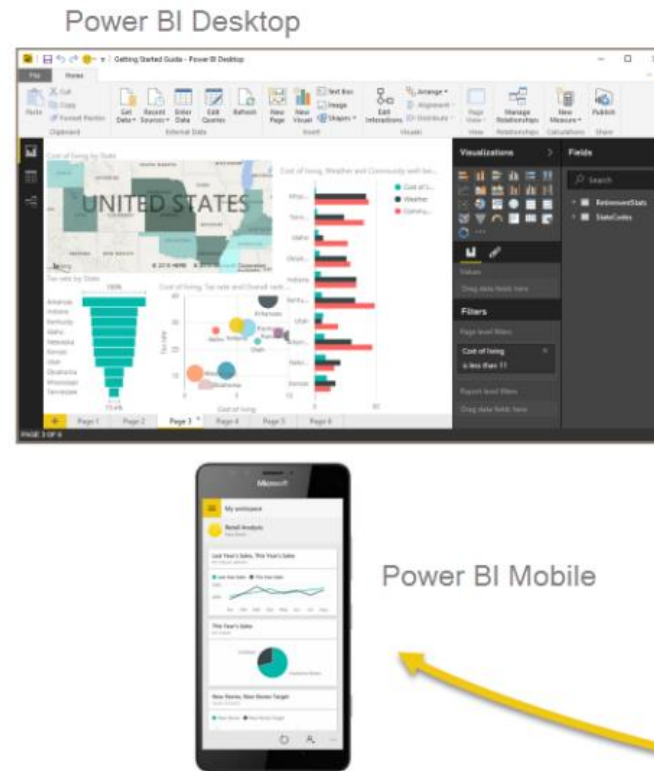
Power BI en 3 minutes !



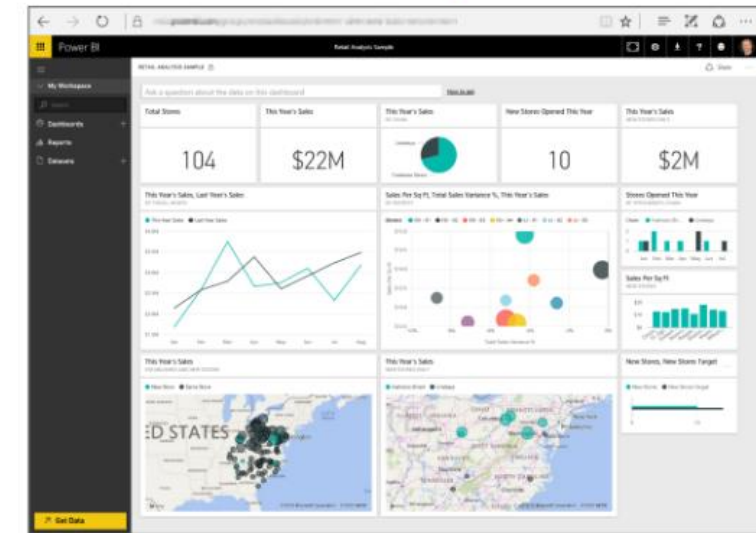
# Power BI en 3 minutes (mode easy)

## ❖ Les principales « briques » Power BI

Mise à jour du Power BI Desktop **tous les mois**

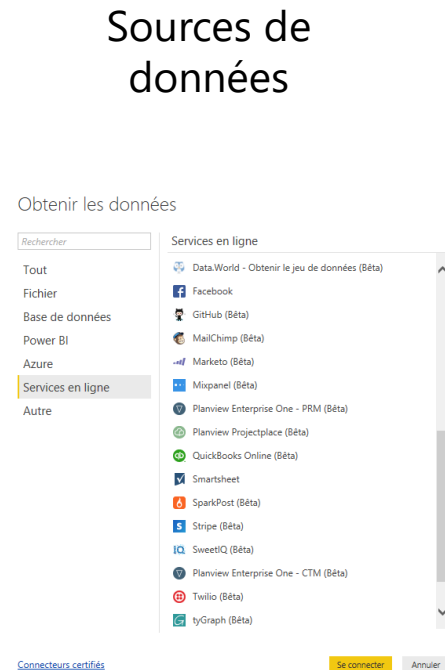


Power BI service



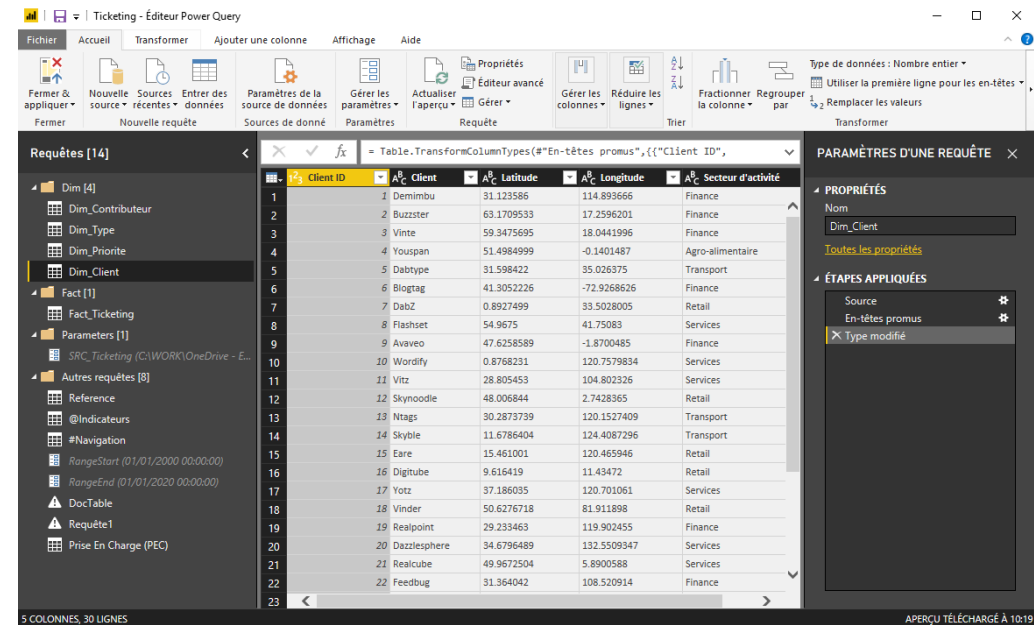
# Power BI en 3 minutes (mode easy)

## ❖ Préparation de données (langage M)



Connexion

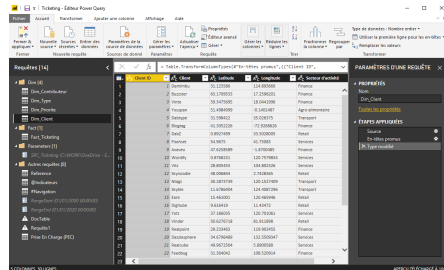
## Editeur Power Query



@ClubPowerBI

# Power BI en 3 minutes (mode easy)

## ❖ Modélisation de données (langage DAX)

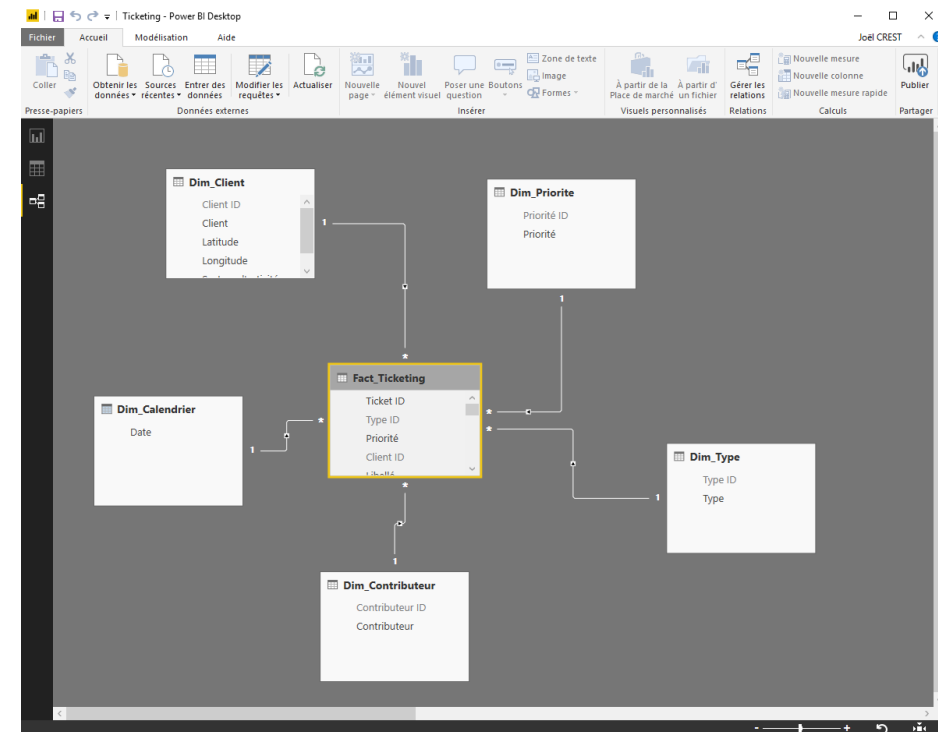


The screenshot shows the 'FactTicketing' table with the following columns: Client ID, Client, Latitude, Longitude, Ticket ID, Type ID, Priorité, and Contributeur ID. The table contains 10 rows of data.

Client ID	Client	Latitude	Longitude	Ticket ID	Type ID	Priorité	Contributeur ID
1	Client 1	48.856614	2.351219	1	1	1	1
2	Client 2	48.856614	2.351219	2	1	1	1
3	Client 3	48.856614	2.351219	3	1	1	1
4	Client 4	48.856614	2.351219	4	1	1	1
5	Client 5	48.856614	2.351219	5	1	1	1
6	Client 6	48.856614	2.351219	6	1	1	1
7	Client 7	48.856614	2.351219	7	1	1	1
8	Client 8	48.856614	2.351219	8	1	1	1
9	Client 9	48.856614	2.351219	9	1	1	1
10	Client 10	48.856614	2.351219	10	1	1	1

➔  
Chargement  
données

Modèle de données

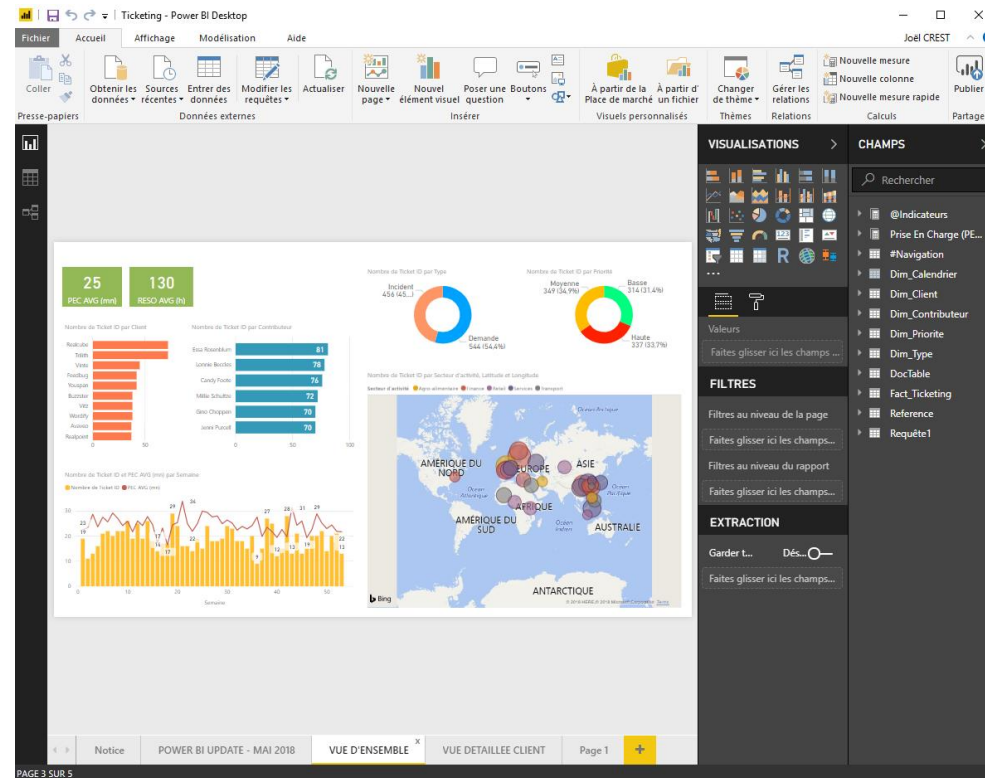




# Power BI en 3 minutes (mode easy)

## ❖ Visualisation de données

Possibilité de switcher  
entre 3 vues



# L'actu du Club Power BI

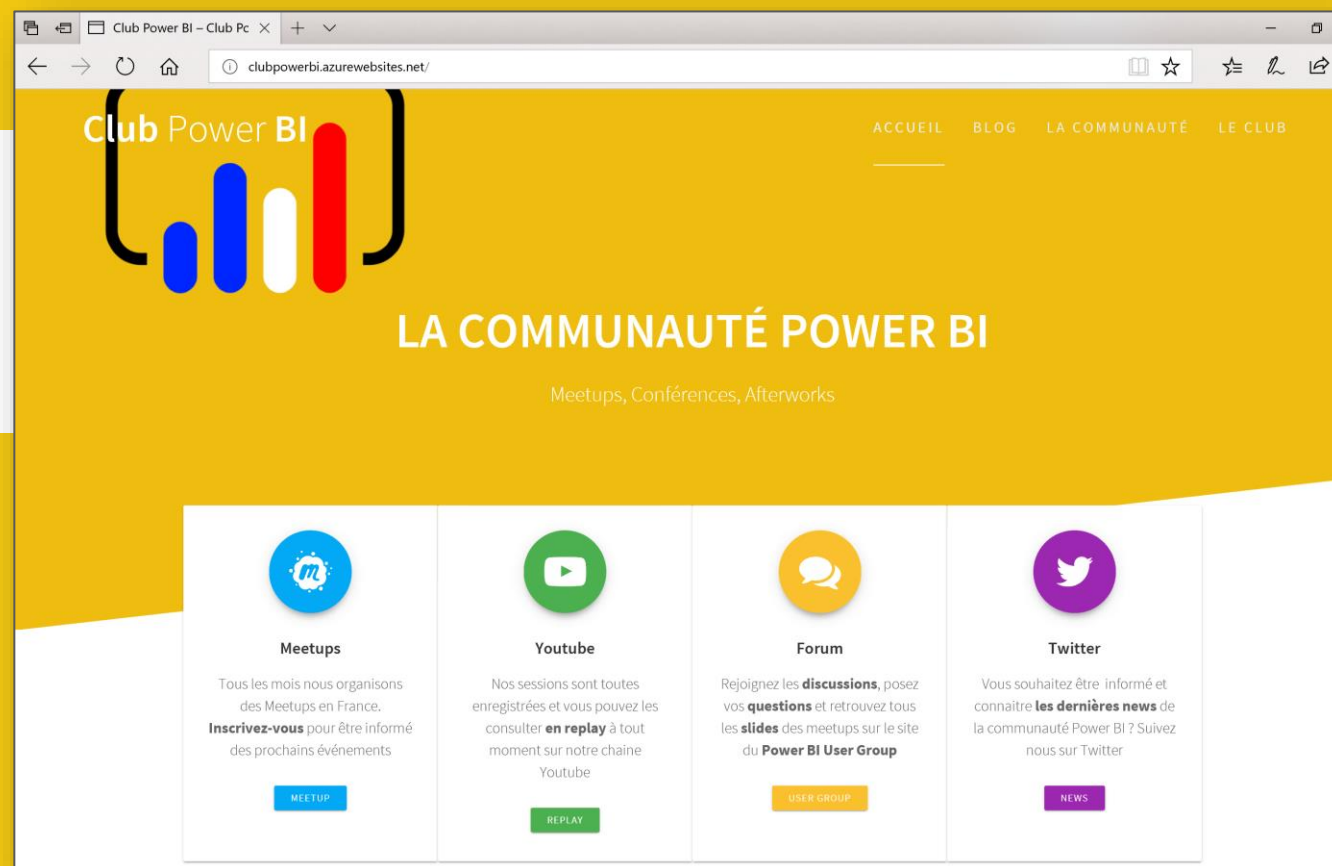
Joël CREST





# Club Power BI

<http://clubpowerbi.com>



## PARIS



Isabelle



Jean-Pierre



Tristan

## LILLE



Guillaume

## STRASBOURG



Philippe

## LYON



Marie



Thomas

## NANTES



Françoise



Paul



Mohamed

## AIX-MARSEILLE



Joël



Franck



# Club Power BI

Présentation de l'équipe



@ClubPowerBI

# Les prochains Meetups

**Online**

**4 mai**

*Dataflow vs Dataset  
avec Reza Rad*

**Online**

**7 mai**

*Modèles composites et  
agrégations*

**Online**

**6 juin**

**Power Saturday**



## PARIS, 5 ET 6 JUIN 2020

Power Saturday est la conférence qui vous donne les clés pour appréhender les **scénarios** et **tendances** qui nous guident, aussi bien les fonctions IT que les directions métiers : Self-Service BI, Collaboration, Citizen Development, Intelligence Artificielle, etc.

CELESTINS DE CONFÉRENCES L.L. - 253 RUE DU FAUBOURG SAINT-MARTIN, PARIS

RETROUVEZ CI-DESSOUS TOUS LES DÉTAILS DE L'ÉDITION 2019

# En Ligne - 6 juin

Data		Power BI / Power Platform		Office 365, SharePoint	
Amphi Microsoft	Bastille	Nation 4 AZEO	Nation 3 Umanis	Odéon 1 Exakis Neille	Odéon 4
Accueil et petit-déjeuner					
Big Intelligent Power Data Platform : tour d'horizon de la Data et + en 2019 Jean-Pierre Rogné & Co		Power BI Embedded Nicolas Sene, Core Giron	Power BI Cheat Sheet explained! David Ruellet, Marc Leblond	Office 365 Security & Compliance Khalid Hussein	Teams and Tricks for ITPros Francesco Sodano
Pause					
RED: Interactive query avec Azure SQL DWH Arnaud Volon	Running statefulset applications like SQL Server in K8s David Barbier	Les bonnes pratiques sur Power Query Thomas Deschamps	Capture Your Store Visit with PowerApps Eric Lecomte	DLP dans Office 365 et Sharepoint en ligne Sandra Aubert	La combinaison de Dynamics 365 avec la Power Platform Chloé Moreau
Pause					
Azure Data Factory Deep Dive Charles-Henri Sautet	SQL dans Azure Sarah Bessard	Power BI et la modélisation de données : Je vise les étoiles! Arnaud Deschamps	Construisons une solution de onboarding avec Graph et Flow en moins d'une heure Julie Rivelle, Gilles Ponsard	SharePoint unexplained Ivan Vagstad	Enabling External Sharing in Office 365, SharePoint and OneDrive Chirag Patel
Déjeuner					
Azure Data Factory Deep Dive	From relational to	CALCULATE - the swiss	Microsoft Flow advanced:	Sketchnoting &	The rise of the citizen



# Les dernières nouveautés sur Power BI

Mohamed CHELLY, tu nous manques !



# Augmentez l'intelligence de vos rapports

Joël CREST - Paul PETON, OM - FCNA





# Les questions fondamentales de la création de valeur



Quels sont mes clients qui ont le plus de valeur ?



Quels sont mes produits ou services les plus importants ?



Quelle sont mes campagnes (marketing, après-vente, etc.) les plus réussies ?

# Quatre questions face à la donnée



*Je doute de la qualité de mes données,  
comment puis-je la vérifier ?*



*Mon indicateur est à la baisse, j'aimerais  
comprendre pourquoi.*



*J'ai trop d'éléments détaillés, comment  
puis-je synthétiser ?*



*Je vois bien ce qui s'est passé mais  
j'aimerais bien me projeter dans le futur.*

# *Une méthode face à la donnée*



*Qualité : observer graphiquement les distributions, calculer des indicateurs de contrôle*



*Comprendre : trouver des phénomènes explicatifs (influent) par corrélation ou modélisation*



*Synthétiser : regrouper des éléments semblables à l'aide des classes ou du clustering*



*Prévoir : les prochaines valeurs d'une série temporelle, un nombre ou une catégorie à l'aide du Machine Learning supervisé*

**Toujours commencer par la qualité des données**

**BI**

**Infused AI**

**Data  
Science**

Comprendre

Résumer

Anticiper

# ***Statistics, back to basics***

*« Je vais dire des trucs simples (...) basiques » (Orelsan)*

# Moyenne *versus* médiane



- Si vous deviez choisir (en France, en 2015) entre le salaire moyen et le salaire médian, lequel choisiriez-vous ?
- Chiffres de 2015 :
  - Salaire moyen net annuel : 26 634 €
  - Salaire médian net annuel : 21 309 €
- La moyenne est « sensible aux valeurs extrêmes ».
- A observer : l'écart entre moyenne et médiane
  - Si moyenne < médiane, il existe des valeurs extrêmes basses
  - Si moyenne > médiane, il existe des valeurs extrêmes basses
  - Si moyenne ~ médiane... des valeurs extrêmes se compensent peut-être !



# Interprétation de l'écart-type



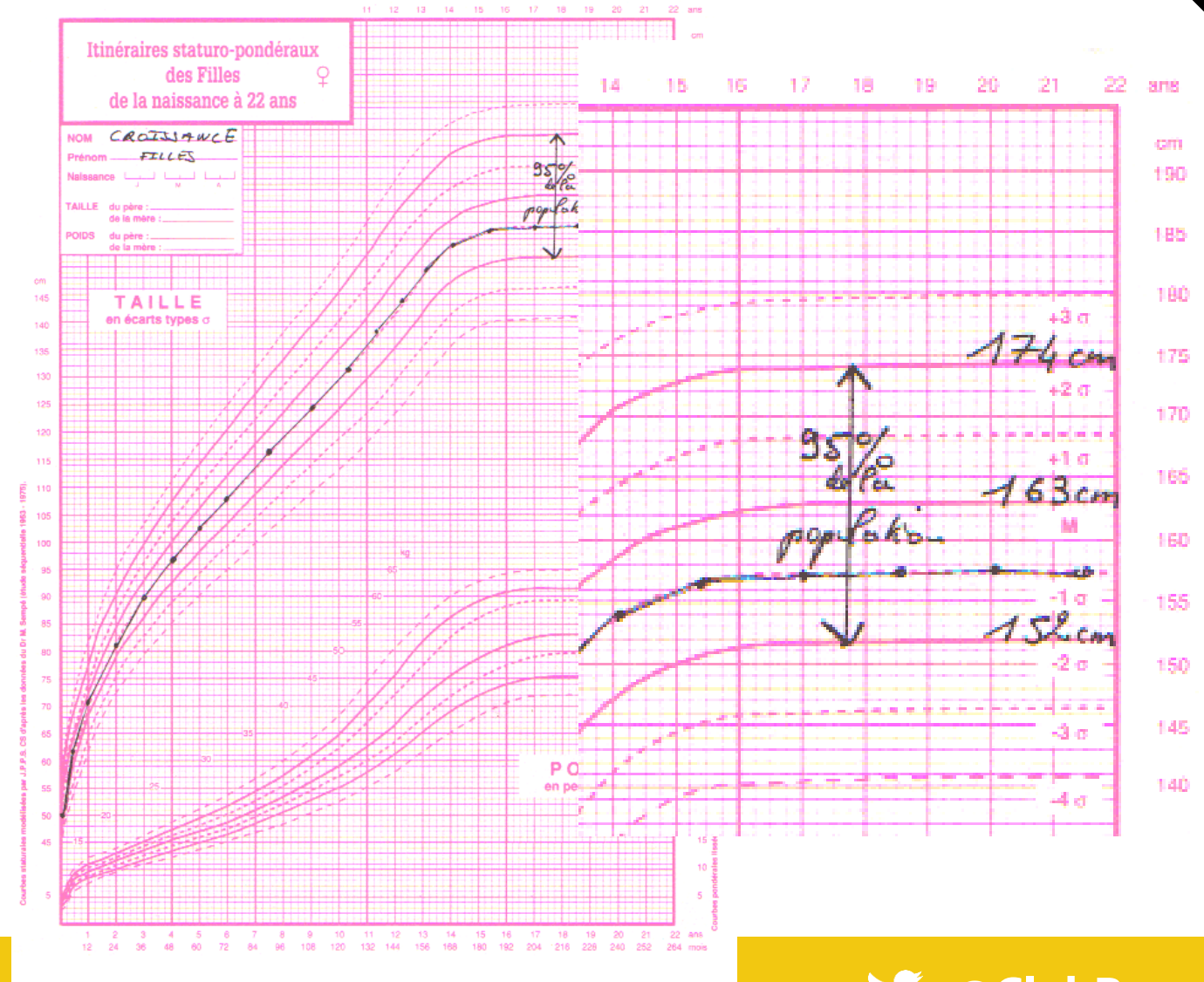
- Ecart-type : indicateur de dispersion d'une valeur numérique
  - Racine carré de la variance
  - Dans l'unité de la variable numérique
- Quelques repères :
  - Dispersion forte si l'écart-type dépasse une demi moyenne
  - Il est « normal » de trouver :
    - 30% des valeurs au-delà de  $\mp 1$  écart-type
    - 5% des valeurs au-delà de  $\mp 2$  écarts-types
    - 2/1000 des valeurs au-delà de  $\mp 3$  écarts-types
    - 1 /10000 des valeurs au-delà de  $\mp 4$  écarts-types
- Voir le quartet d'Anscombe



# Exemple de la courbe de croissance



- Trois lignes :
  - Moyenne +2 écarts-types
  - Moyenne
  - Moyenne - 2 écarts-types
- 95 % entre les lignes extérieures





# Marge d'erreur sur un échantillon



- 6 % des personnes sont atteintes de crises de fou rire
  - Sur la base d'un **échantillon** de 600 personnes
- Entre 4% et 8% des personnes sont atteintes de crises de fou rire
  - Avec un risque d'erreur de 5%
  - On se trompe 5 fois sur 100 en faisant cette conclusion
  - Sur 100 échantillons, 5 donneraient des valeurs  $<4\%$  ou  $>8\%$
- Attention aux strates :
  - 300 hommes : entre 3% et 9% atteints de crises de fou rire
  - 75 hommes de plus de 60 ans : entre 0,5% et 11,5%

# Différence (non) significative de moyenne

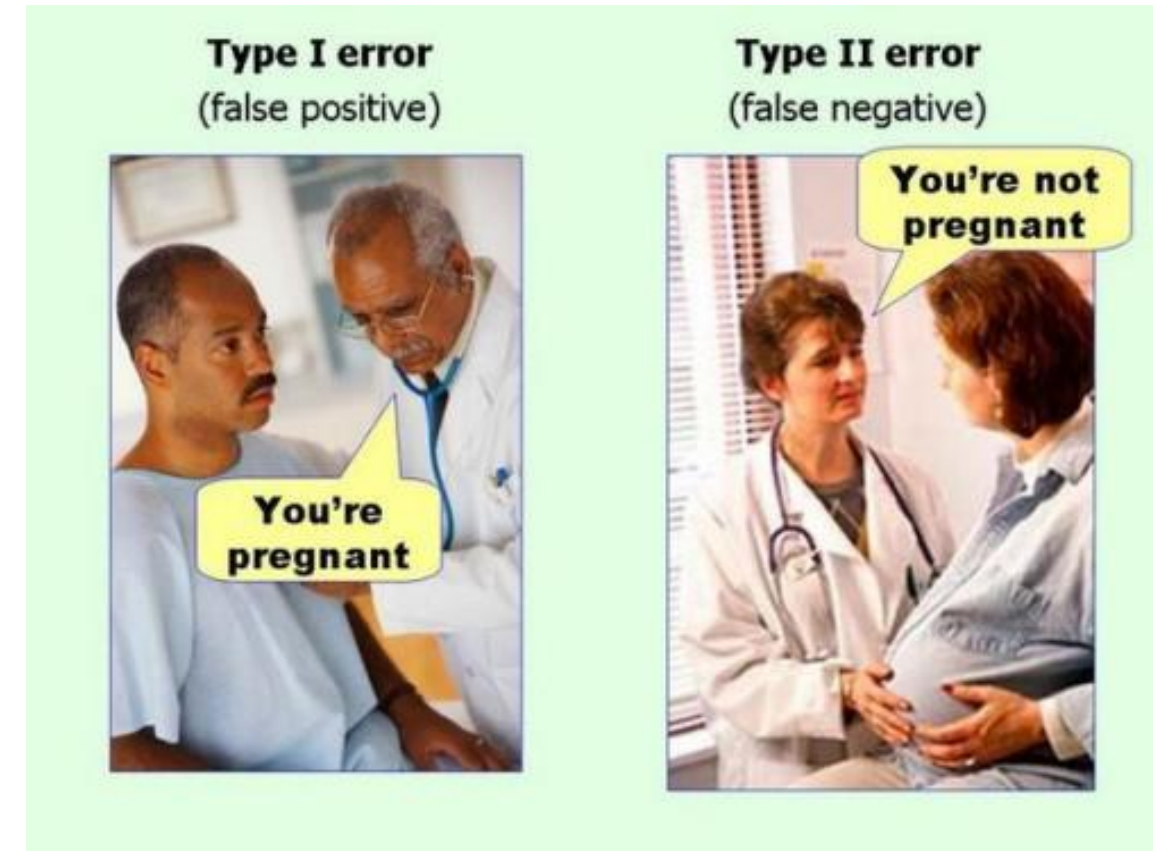


- Concentration moyenne d'endorphine : 10 pg/ml
- Concentration pour les personnes atteintes de fou rire :
  - Echantillon de 36 personnes
  - Moyenne : 13 pg/ml et Ecart-type : 10
- La concentration est comprise entre 9,6 et 16,4 pg/ml
  - Il existe donc un risque (faible) que la valeur ne soit pas significativement différente de celle de la population
  - Cela dépend de la taille de l'échantillon (et non de la population !) et de sa dispersion (la valeur de l'écart-type)

# Faux positifs et faux négatifs



- Les prévisions ne se font jamais avec une certitude totale
  - Ce serait un surajustement (*overfitting*)
- Lors d'une classification binaire, deux erreurs sont possibles :
  - Les faux positifs
  - Les faux négatifs
- La gravité de l'erreur se détermine en fonction du contexte des données



<https://effectsizefaq.com/>

# Biais de confirmation



- Privilégier les informations confirmant ses idées préconçues ou ses hypothèses
- Ex. : il y a plus d'accouchements lors des nouvelles lunes
- Voir aussi :
  - Rasoir d'Ockham : l'explication la plus simple semble la meilleure
  - Biais d'ancrage : se fier à la première information reçue
  - Et bien d'autres biais cognitifs
  - Paradoxe de Simpson

# Scénario données Hypermarché



- Expliquer le taux de retour
- Ventilation
  - Montrer la limite de l'approche manuelle
  - Expliquer la hausse / la baisse
  - Analyser la distribution
  - Influenceurs clés
- Regroupement
  - Données par commune => département (hiérarchisation naturelle)
  - Exemple de la tranche d'âge à partir de l'année de naissance => ajouter une colonne calculée
  - Clustering (auto Power BI) puis par code (trouver une illustration du multidim)
- Anticiper
  - Tendence linéaire
  - Forecast avec injection dans la saisonnalité (mais boîte noire)
  - Modélisation (automatedML) du taux de retour

2 030M €

Chiffre d'Affaires

13,9%

Taux de retour

76M €

Profits après retours

795

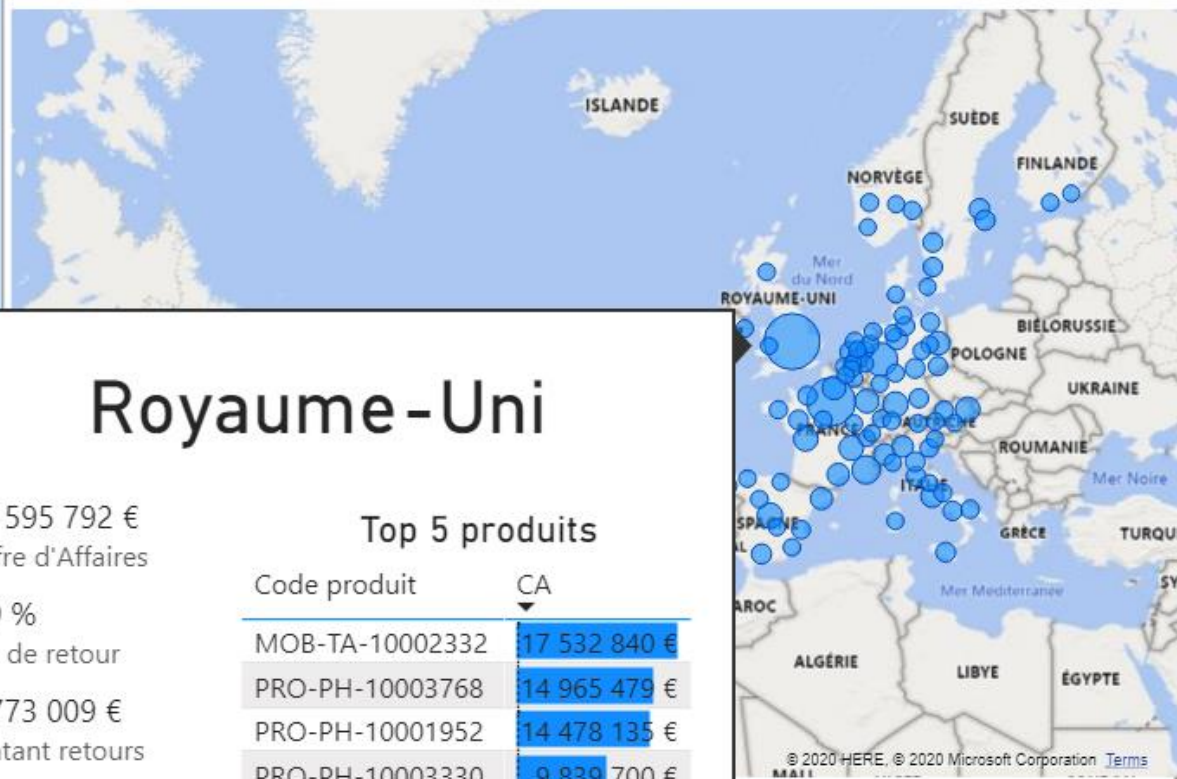
Nb clients actifs

#### Choisir un indicateur

- ☐ CA corrigé des retours
- ☒ Chiffre d'Affaires
- ☐ Délai moyen d'expédition
- ☐ Montant retours
- ☐ Profit corrigé des retours
- ☐ Profit total
- ☐ Quantité totale
- ☐ Taux de retour



Indicateur choisi



## Royaume-Uni

287 595 792 €  
Chiffre d'Affaires

12,9 %  
Taux de retour

10 773 009 €  
Montant retours

276 822 783 €  
CA après retours

16 977 513 €  
Profits après retours

### Top 5 produits

Code produit	CA
MOB-TA-10002332	17 532 840 €
PRO-PH-10003768	14 965 479 €
PRO-PH-10001952	14 478 135 €
PRO-PH-10003330	9 839 700 €
MOB-TA-10002231	9 151 155 €

Indicateur

2013

2014

2015

Année

2016

2017



#### Catalogue

Rechercher

- ☐ Sélectionner tout
- ☒ Fournitures de bureau
- ☒ Mobilier
- ☒ Produits technologiques
- ☒ Accessoires
- ☒ Machines
- ☒ Photocopieurs
- ☒ Téléphones

Poser une question sur vos données (en anglais)



Essayez l'une de ces suggestions pour commencer

what is the MdV min by



@ClubPowerBI

# Qualité des données

*« Garbage in, garbage out » (Jean-Pierre Riehl)*





# *Quelle confiance puis-je avoir dans les indicateurs ?*



- Faut-il remplacer les valeurs manquantes ?
  - Pas toujours car l'absence d'information peut être une information !
  - A faire lorsque l'absence de données est bloquante
- Par quoi remplacer les valeurs manquantes ?
  - Par la moyenne, la médiane, le mode, le « plus proche voisin »...
    - Délicat à faire dans Power BI
    - Sauf en R ou Python
    - Mais problèmes de déploiement sur le service
- Mon conseil : à faire avant d'importer les données dans Power BI





# *Quelle confiance puis-je avoir dans les indicateurs ?*



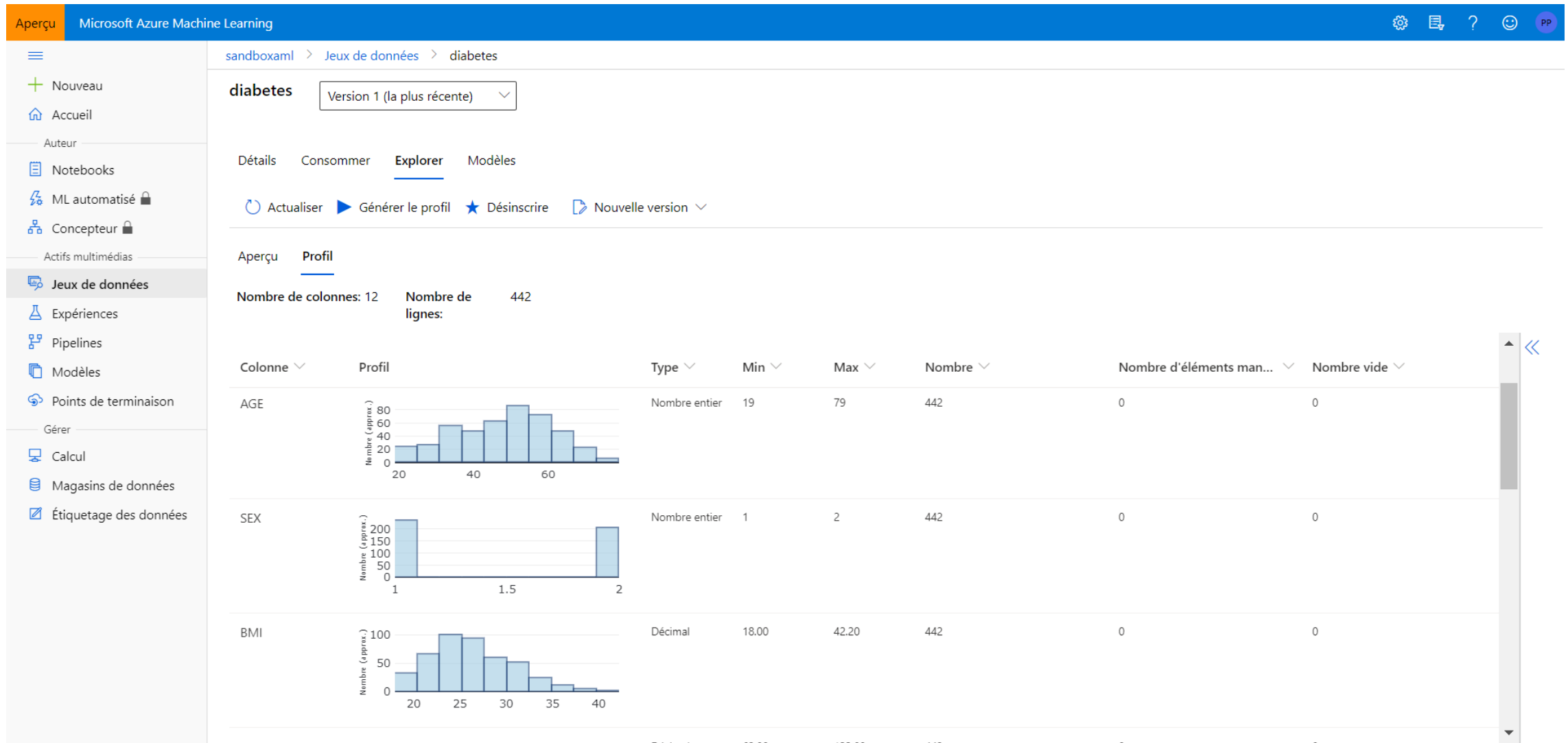
- Comment trouver les valeurs (numériques) aberrantes ?
  - Grâce aux graphiques de distribution
    - Diagramme de distribution (univariée)
    - Boxplot
    - Violin plot
  - En raisonnant « centrage-dispersion »
    - Comparer moyenne et médiane
    - Au-delà de 4 écarts-types, 1 valeur sur 10000 probable (en distribution normale)

# Distribution d'une donnée numérique



- Définition : tableau ou graphique qui associe des classes de valeurs à leurs fréquences d'apparition
  - Les classes sont des intervalles de valeurs
    - Qui ne sont pas forcément d'amplitude constante
    - Même si cela facilite l'interprétation
- La distribution peut s'observer à l'aide de différents graphiques :
  - Histogramme
  - Boîtes à moustaches (box plot)
  - Violin plot

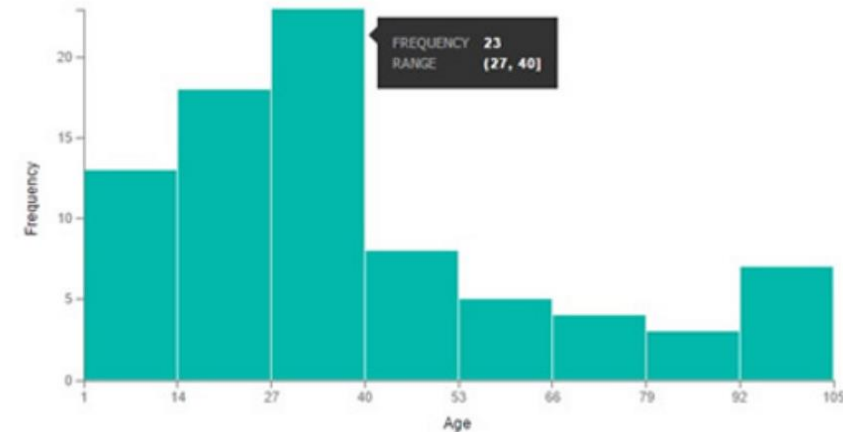
# Profil des données dans Azure ML



# (Véritable) Histogramme



- S'applique UNIQUEMENT à une variable quantitative
- Représente la **distribution** de la variable
- Il faut définir un nombre de classes
  - Automatiquement
  - Arbitrairement
- La distribution la plus fréquente est la distribution Normale
  - Courbe de Gauss ou « courbe en cloche »
  - S'interprète de manière probabiliste :
    - 95% des données dans l'intervalle moyenne  $\pm 2$  écarts-types



# Comprendre et interpréter

*« Si le seul outil que vous avez est un marteau, vous verrez tout problème comme un clou. » (Abraham Maslow)*



	BI	Infused AI	Data Science
Comprendre	Ventilation selon les dimensions  Arbre de décomposition	Expliquer la ↗, la ↘  Influenceurs clés  Cognitive Services	Corrélation, tests d'hypothèse et analyse de variance
Résumer			
Anticiper			

*(hors Power BI)*



*Donnez-moi toutes les mesures et toutes les dimensions, je veux tout croiser !*

- Mais par quel quoi dois-je commencer ?
- Je n'ai pas le temps de tout comparer.
- Je ne sais plus ce que j'ai filtré...
- Et quels sont les éléments vraiment pertinents ?



# Explosion du nombre de combinaisons

3 mesures

\* 4 dimensions

\* 5 champs

\* 2 valeurs

= 120 combinaisons !

- Il faut laisser travailler l'ordinateur
  - grâce aux méthodes statistiques non supervisées
  - puis interpréter
  - et critiquer les résultats obtenus



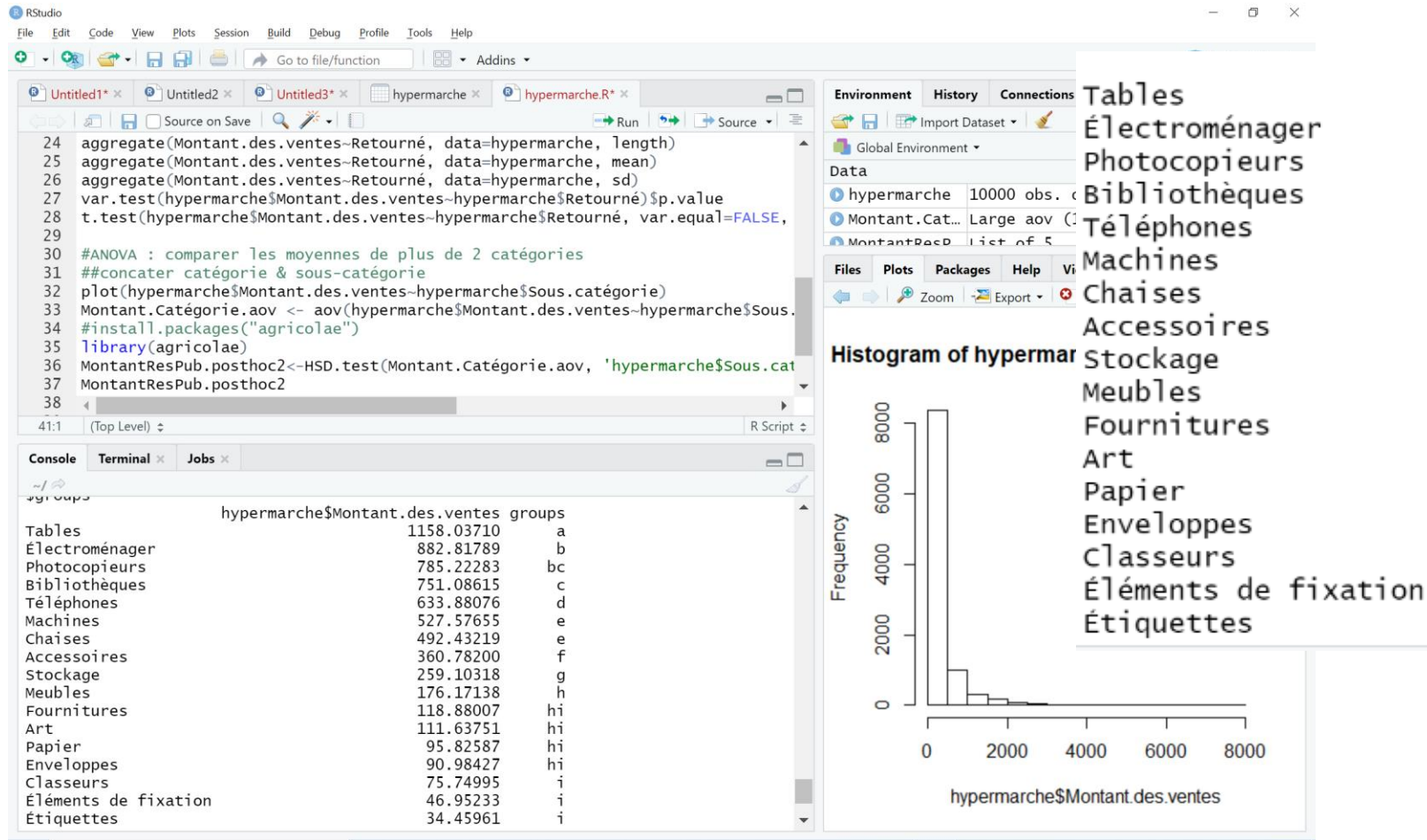
# Comparer des valeurs entre catégories



- Quand la différence entre deux catégories est-elle significative ?
  - Faire un test statistique et interpréter une p-value
  - *Les fonctions de tests n'existent pas en DAX (mais présentes dans Excel)*
- Et comment automatiser entre plusieurs catégories ?
  - Réaliser une ANOVA et des tests post-hoc
- Lire les pourcentages lignes et colonnes
  - Analyser le test du Khi Deux
- *Mais attention à l'effet du volume de données !*
  - Il faut *a minima* 30 valeurs (par catégorie).
  - A l'inverse, les tests seront toujours significatifs avec un très grand volume de données.



# Réaliser les calculs statistiques dans R Studio



hypermarche\$Montant.des.ventes	groups
1158.03710	a
882.81789	b
785.22283	bc
751.08615	c
633.88076	d
527.57655	e
492.43219	e
360.78200	f
259.10318	g
176.17138	h
118.88007	hi
111.63751	hi
95.82587	hi
90.98427	hi
75.74995	i
46.95233	i
34.45961	i

*Regroupement de catégories  
par test post-hoc*

# Analyser la hausse ou la baisse



- Fonctionnalité contextuelle et basée sur le point de données immédiatement précédent
- L'algorithme prend toutes les autres colonnes du modèle (non masquées) et calcule le décompte en fonction de cette colonne pour les périodes *antérieure* et *postérieure*, déterminant l'importance de l'évolution observée dans ce décompte et retournant ensuite les colonnes qui présentent la plus forte évolution.
- Pour identifier la colonne qui présente les plus grandes différences en termes de contributions relatives, voici les éléments qui sont pris en considération :
  - La cardinalité est prise en compte, car une différence est statistiquement moins significative et moins intéressante quand une colonne présente une cardinalité importante.
  - Les différences concernant les catégories dont les valeurs d'origine étaient très élevées ou proches de zéro ont plus de poids que les autres. Par exemple, une catégorie qui comptait pour seulement 1 % des ventes et qui est passée à 6 % est statistiquement plus significative et donc plus intéressante qu'une catégorie dont la contribution est passée de 50 % à 55 %.
  - Plusieurs méthodes heuristiques sont utilisées pour sélectionner les résultats les plus significatifs, par exemple en prenant en considération d'autres relations entre les données.
- Après examen des différentes colonnes, celles qui sont choisies et affichées sont celles dont la contribution relative a le plus fortement évolué. Pour chacune, les valeurs qui correspondent à la plus forte évolution sur le plan de la contribution sont représentées dans la description. Il en va de même des valeurs qui ont le plus augmenté ou baissé.



# Fonctionnalités non prises en charge



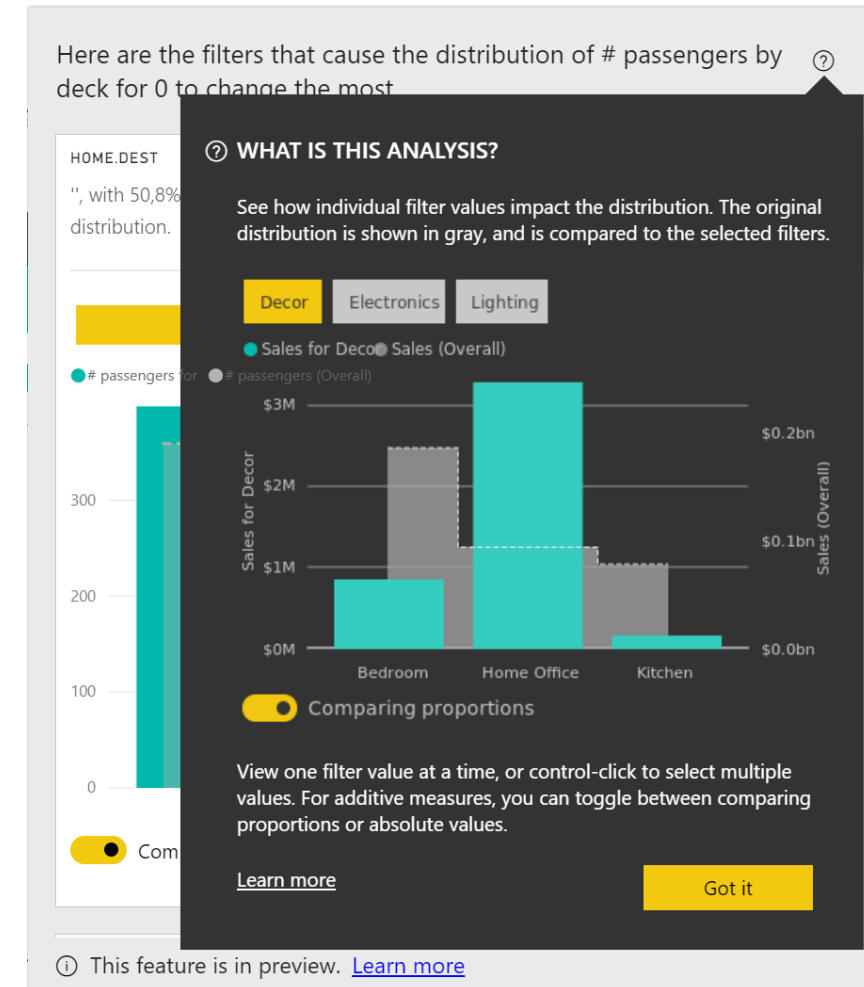
- La liste suivante répertorie les scénarios actuellement non pris en charge pour **expliquer les hausses/baisses** :
  - Filtres TopN
  - Filtres Inclure/Exclure
  - Filtres de mesures
  - Mesures non numériques
  - Utilisation de « Afficher la valeur comme »
  - Mesures filtrées (les mesures filtrées sont des calculs effectués au niveau du visuel auxquels est appliqué un filtre spécifique (par exemple, *Total des ventes pour la France*) ; elles sont utilisées dans certains visuels créés par la fonctionnalité d'insights.
  - Colonnes de catégorie sur l'axe X, sauf si celui-ci définit un tri par colonne de type scalaire. Si vous utilisez une hiérarchie, chaque colonne dans la hiérarchie active doit remplir cette condition.
- De plus, les sources de données et les types de modèles suivants ne sont actuellement pas pris en charge pour la fonctionnalité d'affichage d'informations :
  - Direct Query
  - Live Connection
  - Reporting Services en local
  - Incorporation



# Trouver où la distribution est différente



- Objectif :
  - Savoir si cette répartition serait identique pour des sous-populations différentes
  - Mettre en évidence les différences trouvées entre la répartition globale (telle que présentée dans le visuel d'origine) et la valeur après application du filtre spécifique.
- Pour les mesures additives simples, la comparaison s'effectue à partir de valeurs relatives, et non absolues.
- Pour les mesures non additives, l'algorithme recherche des différences dans la valeur absolue.



# Fonctionnalités non prises en charge

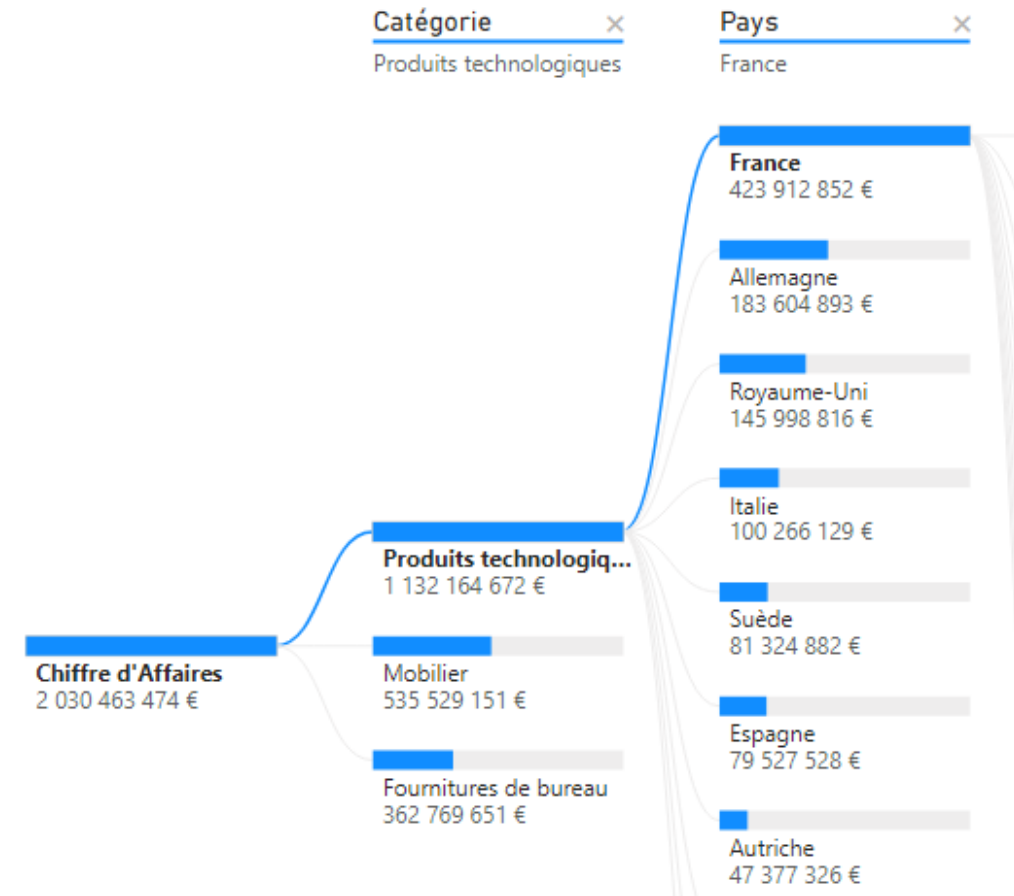


- La liste suivante répertorie les scénarios actuellement non pris en charge :
  - Filtres TopN
  - Filtres de mesures
  - Mesures non numériques
  - Utilisation de « Afficher la valeur comme »
  - Mesures filtrées
    - les mesures filtrées sont des calculs effectués au niveau du visuel auxquels est appliqué un filtre spécifique (par exemple, *Total des ventes pour la France*)
- Les sources de données et les types de modèles suivants ne sont actuellement pas pris en charge :
  - DirectQuery
  - Live Connection
  - Report Server
  - Embedded



# Arbre de décomposition

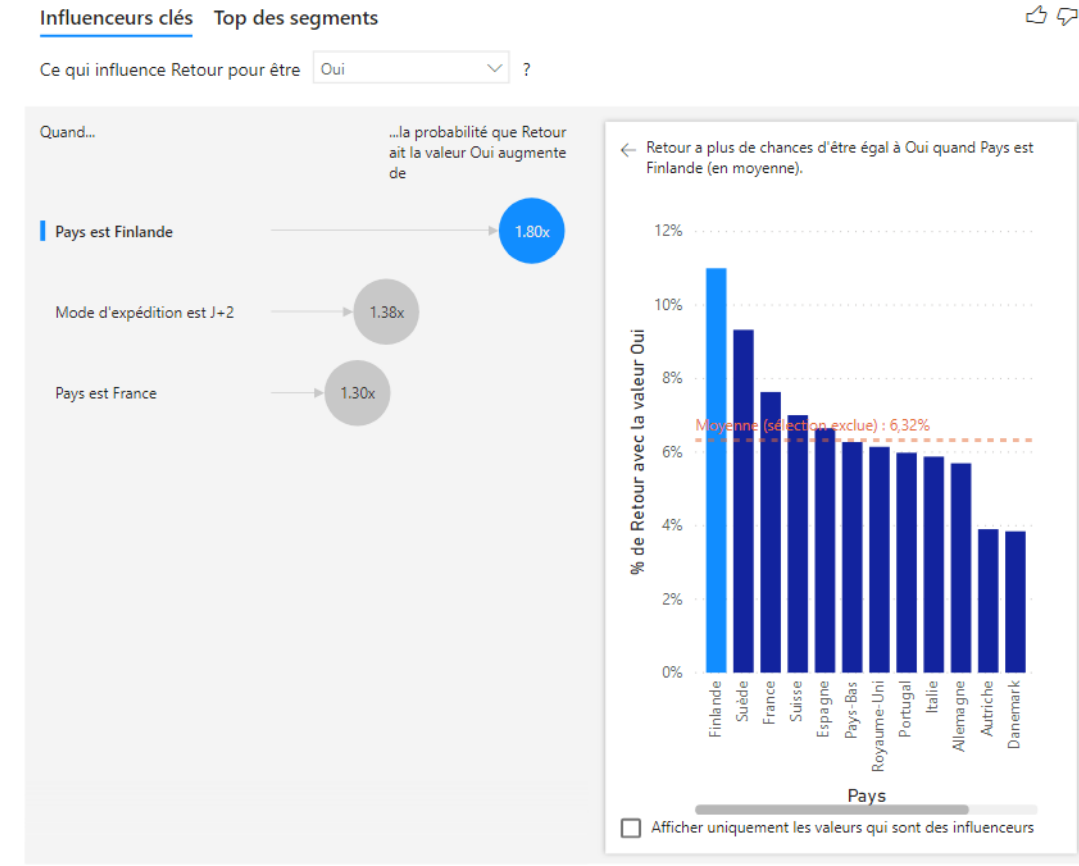
- Visuel permettant de répartir une mesure sur différentes valeurs d'une dimension.
- Si plusieurs champs de dimensions sont proposés, le visuel peut déterminer celui qui présente l'élément le plus représenté.
- Ce n'est pas pour autant une méthode d'IA.



# Visualisation des influenceurs clés



- Deux types d'analyse
  - Key influencers
  - (Top) Segments
- Analyse une variable catégorielle
  - Avec un nombre de catégories assez faible
  - Par des variables catégorielles ou numériques
- Analyse une variable numérique
  - A la hausse ou à la baisse
- Visuel récent
  - Voir les fonctionnalités non prises en charge





# Principes de fonctionnement



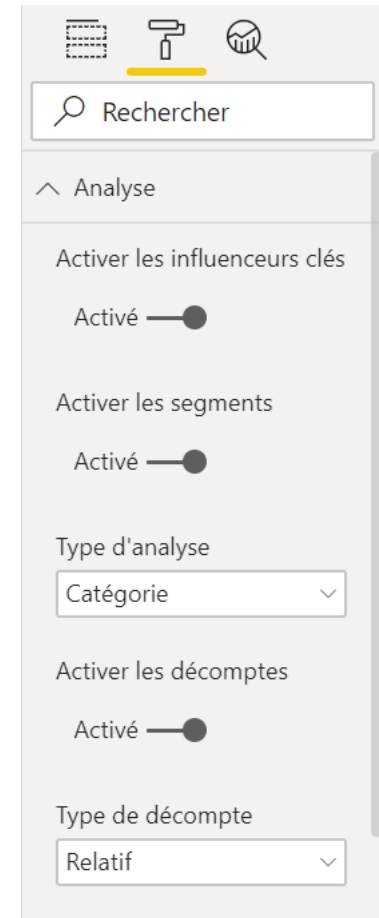
- Appel à la plateforme ML.NET
  - open source machine learning framework, created by Microsoft, for the .NET developer platform
- Key influencer : algorithme de régression logistique qui compare différents groupes de données entre eux
  - Classe les mesures qui comptent et se concentre sur les facteurs qui influencent la mesure clé
- Top segments : algorithme d'arbre de décision qui crée des sous-groupes de points de données ayant une forte influence sur la mesure analysée
  - Combine des mesures qui, ensemble, influencent la mesure clé



# Options de l'influenceur clé



- Possibilité de désactiver les influenceurs clés ou les segments si l'un des deux n'a pas de signification
- Type d'analyse : catégorie ou continu
- Type de décompte (absolu ou relatif) : ajoute une portion de cercle autour du coefficient de l'influenceur
  - Permet de visualiser la proportion concernée dans la population
- Champ « développé par »
  - Pour un champ ou une mesure synthétisée (« summarized »)



# Fonctionnalités non prises en charge



Le visual ne permet pas d'utiliser les fonctionnalités suivantes :

- Analyzing metrics that are aggregates/measures
- Consuming the visual in embedded
- Consuming the visual on Power BI mobile
- RLS support
- DirectQuery support
- Live Query support



# Limites de l'approche Key Influencers



- La qualité de la réponse dépend... de la qualité de la donnée fournie !
  - Il faut un minimum de données
  - Les événements rares (ex.: spam à 2%) sont plus difficiles à discriminer
- Il faut être critique par rapport aux résultats présentés
  - En particulier la taille des segments
  - Ne pas « forcer » les interprétations
  - Eviter le biais du rasoir d'Ockham : l'explication qui nous paraît la plus simple nous semble toujours être la meilleure !



# Est-ce vraiment l'IA dans Power BI ?



- Non, simplement une régression logistique

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

- Interprétation des coefficients du modèle :
  - $\text{Exp}(-1,37821) = 0,25$
  - $\text{Exp}(1,37771) = 3,97$

=> ~4 fois plus de chances de survivre en étant une femme plutôt qu'un homme

Feature	Weight
Bias	1,45599
sex_male_1	-1,37821
sex_female_0	1,37771
age	-0,944348
cabin_C22 C26_73	-0,511961
cabin_G6_171	-0,504166
embarked_C_0	0,286755
cabin_A34_11	0,191287
cabin_E24_142	0,167725
sibsp	0,154279
cabin_E25_143	0,133692
cabin_B96 B98_55	0,129552
cabin_C55 C57_88	-0,122333
embarked_S_2	-0,0606757
cabin_E34_146	0,0188288

# Résumer

*"Rien de beau ne peut se résumer." (Alan Turing)*



	BI	Infused AI	Data Science
Comprendre			
Résumer	Regroupement métier (« <i>a priori</i> »)	Regroupement naturel ( <i>clustering</i> )	Clustering à plus de 3 dimensions
Anticiper			

(hors Power BI)



@ClubPowerBI

# Créer des groupes

- Deux méthodes disponibles
  - Compartiments
    - Définis par leur taille ou nombre
  - Liste
    - Groupes manuels à partir des valeurs
- Manque une méthode pour des groupes équilibrés
  - Utiliser les percentiles
  - Ex. : Q1, médiane et Q3 pour créer 4 groupes

## Groupes

Nom  Champ

Type de groupe  Valeur minimale

Type de compartiment  Valeur maximale

Le compartimentage répartit les données numériques ou de date/heure en groupes de même taille. La taille du compartiment par défaut est calculée en fonction de vos données.

Taille du compartiment

Rétablir les valeurs par défaut

## Groupes

Nom  Champ

Type de groupe

### Valeurs non groupées

27  
27.794117647058822  
29.823529411764707  
31.5  
32  
34.125  
39.75  
44.57142857142857  
54.142857142857146  
54.52173913043478

### Groupes et membres

▲ - de 25

- 6
- 9
- 10.928571428571429
- 12
- 23.510204081632654
- 24

▲ Autre

- Contient toutes les valeurs non groupées

Groupe

Dissocier

☒ Inclure un autre groupe ⓘ

OK

Annuler



@ClubPowerBI



# Regrouper automatiquement : clustering



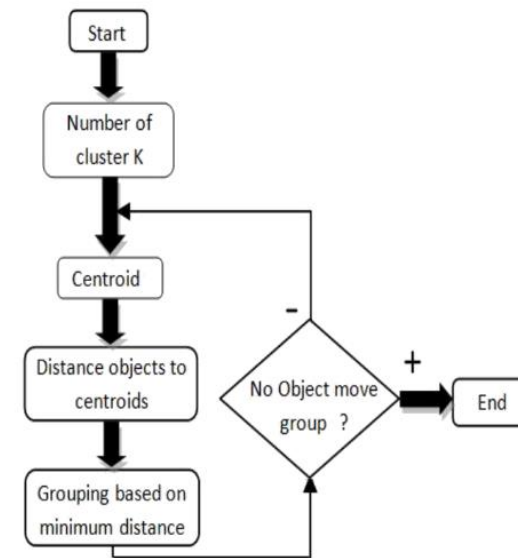
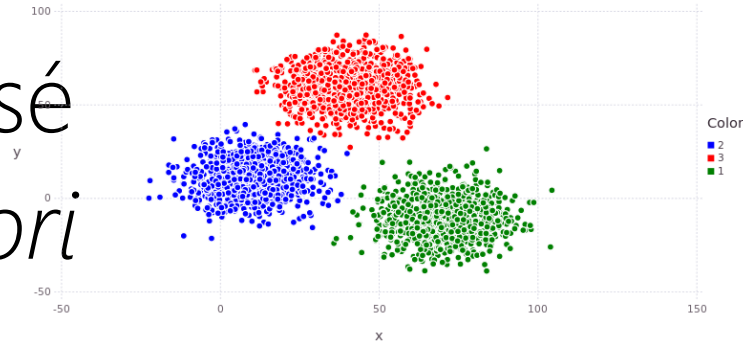
- Choix du nombre de classe
  - A priori
  - Automatique : quelle méthode ? (Silhouette ?)
- Limitations de Power BI :
  - Seules les deux mesures en X et en Y pour le calcul des distances
  - Ne pas utiliser le champ segment : il sera remplacé par le cluster
- Attention, en français, on traduit clustering par classification
  - A ne pas confondre avec la classification de l'apprentissage supervisé
  - Pour le métier, cela est souvent synonyme de segmentation ou de typologie



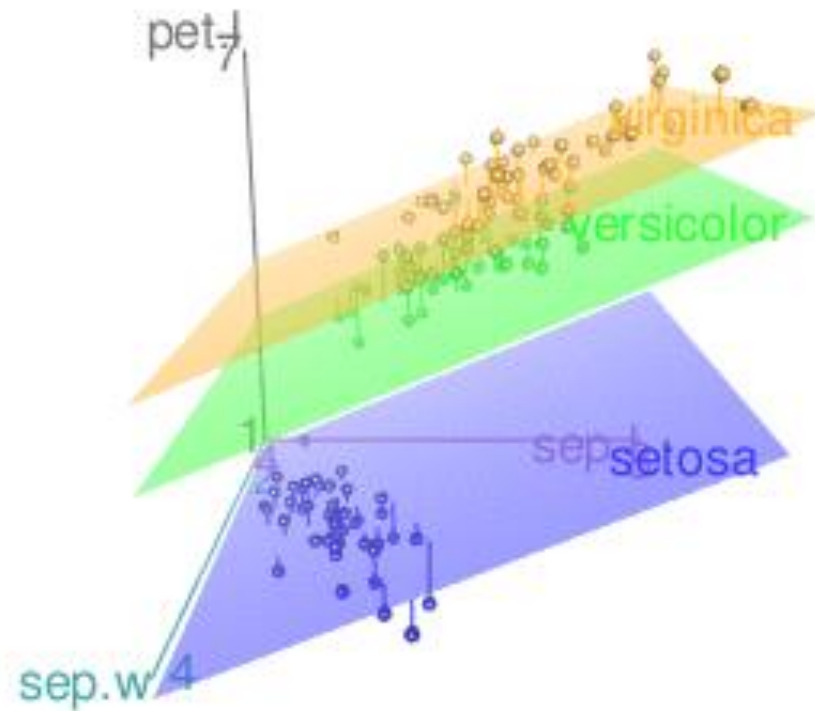
# Le clustering



- Méthode d'apprentissage non supervisé
- On recherche une structure sans *a priori*
  - Une forte similarité intraclasse
  - Une faible similarité interclasse
- Algorithme K-Means :
  - Choisir K éléments initiaux "centres" des K groupes
  - Placer les objets dans le groupe de centre le plus proche
  - Recalculer le centre de gravité de chaque groupe
  - Itérer l'algorithme jusqu'à ce que les objets ne changent plus de groupe



# Clustering en trois dimensions (iris dataset)



# Rechercher automatiquement les clusters



- Fonctionnement :
  - Dans un nuage de points
  - Ajouter 2 (voire 3) données numériques (X, Y et éventuellement taille)
  - Utiliser le champ détails, ne pas utiliser le champ légende
  - Choisir le nombre de clusters ou laisser l'algorithme trouver le plus pertinent
- Limites :
  - Peu de données numériques (3 maximum) ou de points (30000 maximum)
  - Impossible de projeter de nouveaux points sur les coordonnées
  - Sensibilité aux valeurs extrêmes
- Il sera ensuite nécessaire d'étudier les groupes constitués pour les interpréter.



# Anticiper

*« J'ai vu plus loin que les autres (...) juché sur les épaules de géants. »  
(Isaac Newton)*



	BI	Infused AI	Data Science
Comprendre			
Résumer			
Anticiper	Tendance linéaire (« <i>tirer un trait</i> »)	Prévision ( <i>forecast</i> )	Régression Classification Automated ML

(hors Power BI)



# Prévision (*forecast*) par décomposition



- Une série temporelle peut être décomposée en différents éléments :
  - Une saisonnalité (hebdomadaire, mensuelle...)
  - Une tendance (à la hausse, à la baisse)
  - Un bruit
  - Eventuellement, un cycle économique
- Ces différents éléments s'additionnent ou se multiplient.
- Sur l'hypothèse d'une saisonnalité constante, la tendance (à court terme) peut être prolongée pour constituer la prévision.



# Paramétrer le composant Prévvision



- Disponible dans le menu Analytique
- Pour un visuel de type Courbe uniquement
- Indisponible pour les « mesures filtrées »
- Paramètres :
  - Unité : point, élément de date ou d'heure
  - Ignorer les premiers points d'un cycle s'il n'est pas complet
  - Il faudrait vérifier le caractère saisonnier
  - L'intervalle de confiance est un intervalle de prévision

Longueur de la prévision

12

Mois

Ignorer le dernier

0 Mois

Intervalle de confiance

95%

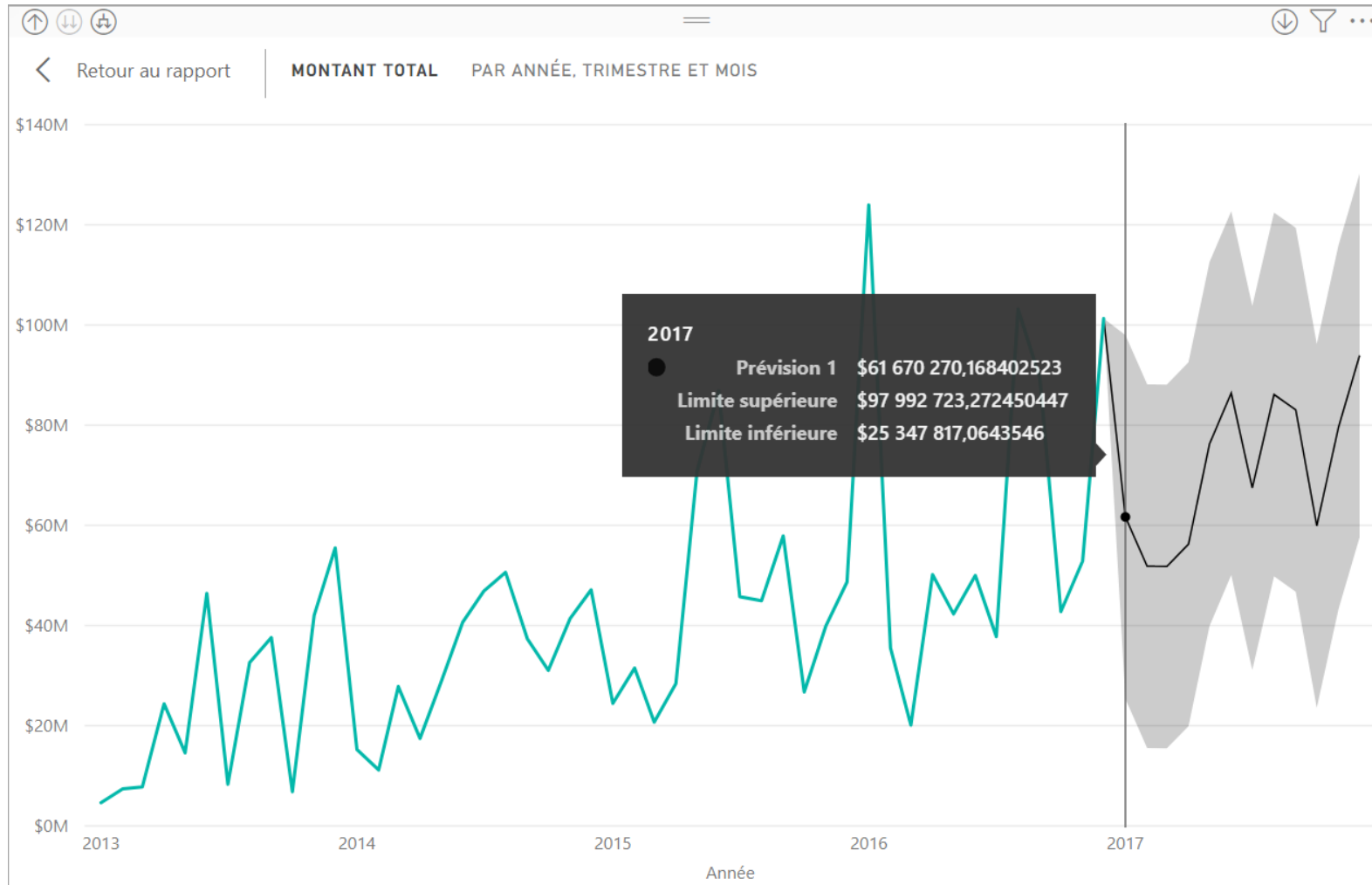
Caractère saisonnier

12 Point(s)





# Afficher un composant Prédiction



- Les limites inférieure et supérieure apparaissent seulement dans l'info-bulle
- Possible de le faire dans Excel (2016+)



# Comprendre l'intervalle de prévision



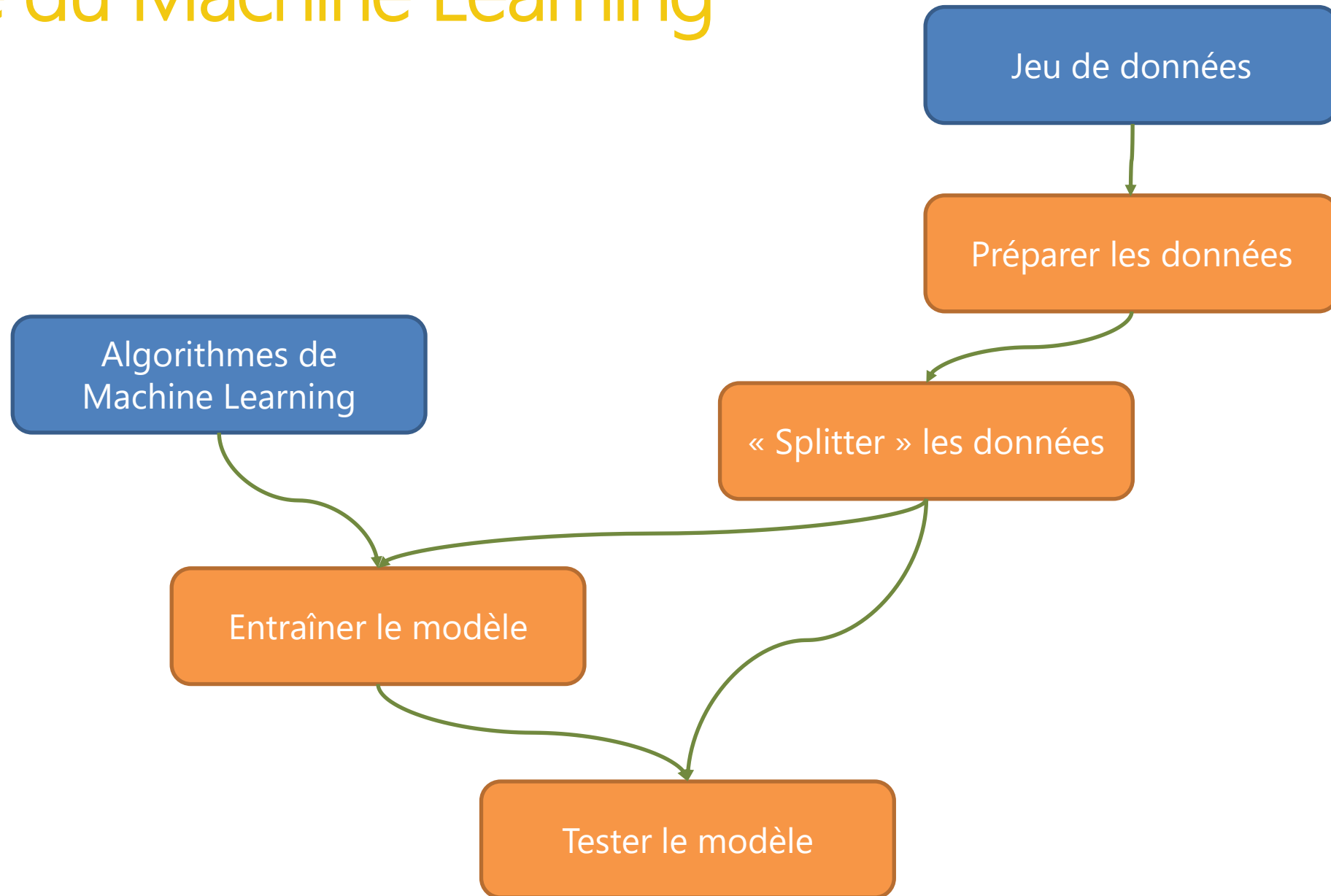
- Niveau de confiance : 95% (ou risque d'erreur : 5%)
- La valeur est donnée par la formule suivante :

$$\text{Var}(\varepsilon_{n+1}) = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

[http://www.jybaudot.fr/Correl\\_regress/intprev.html](http://www.jybaudot.fr/Correl_regress/intprev.html)

- A ne pas confondre avec l'intervalle de confiance :
  - L'intervalle de prédiction est toujours plus large que l'intervalle de confiance et s'amplifie plus la prévision s'éloigne

# Principe du Machine Learning



# Rapport de performance

## ❖ Model Performance

- Matrice de confusion
- Analyse coût / bénéfice

## ❖ Accuracy Report

- Graphe de gains cumulatifs
- Courbe ROC

## ❖ Training Details



# Matrice de confusion

## ❖ VP = Vrai Positif

- Prédiction = survivant
- Réalité = survivant

## ❖ FP = Faux Positif

- Prédiction = survivant
- Réalité = disparu

## ❖ FN = Faux Négatif

- Prédiction = disparu
- Réalité = survivant

## ❖ VN = Vrai Négatif

- Prédiction = disparu
- Réalité = disparu

	Predicted Survivant	Predicted Disparu
Actual Survivant	<b>VP</b>	<b>FN</b>
Actual Disparu	<b>FP</b>	<b>VN</b>



# Notion de précision

## ❖ Précision

- Sur l'ensemble des identifications positives
- Quel est le nombre d'identifications effectivement correctes ?

$$\text{Précision} = \text{VP} / (\text{VP} + \text{FP})$$

67%  
Precision

	Predicted Survivant	Predicted Disparu
Actual Survivant	64,00	9,00
Actual Disparu	30,00	75,00



# Notion de rappel (recall)

## ❖ Rappel

- Sur l'ensemble des identifications devant être positives
- Quel est le nombre d'identifications effectivement correctes ?

$$\text{Rappel} = \text{VP} / (\text{VP} + \text{FN})$$

88%  
Recall

	Predicted Survivant	Predicted Disparu
Actual Survivant	64,00	9,00
Actual Disparu	30,00	75,00



# Générer un point de terminaison prédictif

- ❖ Automated ML est aussi disponible
- ❖ On déploie une image Docker comme une API REST
- ❖ On l'interroge en envoyant de nouvelles données au format JSON
- ❖ Cette URL peut être utilisée dans Power BI (Premium)

The screenshot displays the Microsoft Azure Machine Learning interface. The left sidebar contains a navigation menu with options like 'Nouveau', 'Accueil', 'Notebooks', 'ML automatisé', 'Concepteur', 'Jeux de données', 'Expériences', 'Pipelines', 'Modèles', and 'Points de terminaison'. The main panel shows the 'hypermarcheaci' endpoint under the 'Consommer' tab. It provides the REST endpoint URL: `http://89c38734-9cd0-414c-8ced-37db9c2c630b.westus2.azurecontainer.io/score`.





# Fonctions DAX

*"CALCULATETABLE (*  
*SUMMARIZE ( Sales, Sales[SalesOrderNumber] )..." (Marco Russo)*



# Choisir la bonne fonction statistique



- STDEV.S Function
- STDEV.P Function
- L'écart-type doit être artificiellement augmenté lorsque l'on ne dispose pas de l'exhaustivité des données
  - Il y a toujours des données manquantes
  - Prendre la fonction en .S comme « sample »
  - Et non en .P comme « population »
- PERCENTILE.EXC Function
- PERCENTILE.INC Function
- A choisir selon que l'on considère les bornes de l'intervalle  $[0;1]$  incluses ou exclues.
  - On travaille généralement avec la fonction PERCENTILE.INC
  - Le percentile 0 est le minimum, le percentile 1 le maximum



# Calcul de l'intervalle de confiance



- CONFIDENCE.NORM Function
- CONFIDENCE.T Function
- La fonction renvoie la « demi-amplitude » de l'intervalle de confiance.
- Il faut donc la retrancher et l'ajouter à la valeur encadrée pour obtenir le véritable intervalle de confiance.
- Utiliser la fonction .T (version de Student) si les données comportent moins de 30 lignes, la fonction .NORM sinon (version « normale »).



# Fonctions DAX du Key Influencers



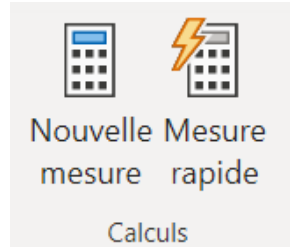
Ces fonctions ne sont pas directement disponibles mais apparaissent lors de l'exécution du visuel Key Influencers :

- AI.SAMPLESTRATIFIED()
- AI.TRAIN()
- AI.KEYDRIVERS()
- AI.EXTRACTPROFILEFILTERS()
- SampleAxisWithLocalMinMax

```
1 DEFINE
2   VAR __DATA__ =
3     SELECTCOLUMNS(
4       'Acc',
5       "[MODEL_TARGET]", 'Acc'[Accident_Severity] IN {(1)},
6       "[Sandbox.Acc.v1st_Road_Class]", 'Acc'[1st_Road_Class],
7       "[Sandbox.Acc.Light_Conditions]", 'Acc'[Light_Conditions]
8     )
9
10  VAR __LABEL_COUNT__ =
11    GROUPBY(
12      __DATA__,
13      "[MODEL_TARGET]",
14      "COUNT", SUMX(
15        CURRENTGROUP(),
16        1
17      )
18    )
19
20  VAR __SAMPLED_DATA__ =
21    AI.SAMPLESTRATIFIED(
22      __DATA__,
23      "[MODEL_TARGET]",
24      __LABEL_COUNT__,
25      10000
26    )
27
28  VAR __FEATURE_ROLE_MAPPING__ =
29    SELECTCOLUMNS(
30      {
31        (0, "[Sandbox.Acc.v1st_Road_Class]", 0),
32        (1, "[Sandbox.Acc.Light_Conditions]", 0)
33      },
34      "FeatureId", [value1],
35      "FeatureColumn", [value2],
36      "FeatureRole", [value3]
37    )
38
39  VAR __MODEL__ =
40    AI.TRAIN(
41      __SAMPLED_DATA__,
42      "[MODEL_TARGET]",
43      __LABEL_COUNT__,
44      __FEATURE_ROLE_MAPPING__
45    )
46
47  EVALUATE
48    AI.KEYDRIVERS(__MODEL__)
49
50  EVALUATE
51    AI.EXTRACTPROFILEFILTERS(__MODEL__)
52
```



# Mesure rapide : corrélation (linéaire)



- Doit être faite pour tout couple de deux mesures
- Croisées avec une dimension
- Ingérable !
- A faire dans un autre outil
  - Voir les matrices de corrélations
  - L'approche linéaire n'est pas la seule

## Mesures rapides

### Calcul

Coefficient de corrélation ▼

Calcule le coefficient de corrélation entre deux valeurs sur une catégorie. Initialement suggéré par Daniil Maslyuk dans la galerie de mesures rapides.

[En savoir plus](#)

### Catégorie ⓘ

Pays ×

### Mesure X ⓘ

Délai moyen d'expédition ×

### Mesure Y ⓘ

Nb retours ×

### Champs

Rechercher

#### Indicateurs

- CA A-1
- CA après retours
- Chiffre d'Affaires
- Délai moyen d'expédition
- Montant moyen
- Montant retours
- Nb clients actifs
- Nb commandes
- Nb produits commandés
- Nb retours
- Profits après retours



	BI	Infused AI	Data Science
Comprendre	Ventilation selon les dimensions  Arbre de décomposition	Expliquer la ↗, la ↘  Influenceurs clés  Cognitive Services	Corrélation, tests d'hypothèse et analyse de variance
Résumer	Regroupement métier (« <i>a priori</i> »)	Regroupement naturel ( <i>clustering</i> )	Clustering à plus de 3 dimensions
Anticiper	Tendance linéaire (« <i>tirer un trait</i> »)	Prévision ( <i>forecast</i> )	Régression Classification Automated ML

(hors Power BI)



@ClubPowerBI

# Parfois, l'absence de résultat est aussi un résultat...



Aucun insight

# Q&A





Rendez-vous lundi 4 mai 20h30  
avec Tristan, Jean-Pierre et Reza

**Merci !**

