



# Guide d'utilisation

## Visual Analytics Explorer

Apprendre à réaliser des fouilles ainsi que des explorations de données

Photowatt®



# **SOMMAIRE**

## Table des matières

Introduction SAS Visual Analytics.....	5
Accès à l'application Visual analytics .....	11
Ouverture d'une exploration SAS Visual Analytics.....	15
Choisir une nouvelle source de données .....	17
Liste des sources de données disponibles :.....	20
L'exploration de données.....	23
Le diagramme de Sankey.....	25
Boîte à moustache.....	31
Arbre de décision .....	35
Matrice de corrélation.....	41
Carte thermique .....	47
Nuage de mots .....	51
Distribution.....	55
Diagramme de réseau .....	59
Graphique à bulles.....	65



# Introduction SAS Visual Analytics

SAS Visual Analytics est un outil de visualisation, édité par la société SAS, qui permet l'exploration visuelle et la représentation graphique des données, quelles qu'en soient la volumétrie, la nature ou la provenance.

L'outil peut aider notamment à détecter des phénomènes ou des tendances invisibles de prime abord et permet une visualisation rapide d'un grand nombre de données contextualisées.

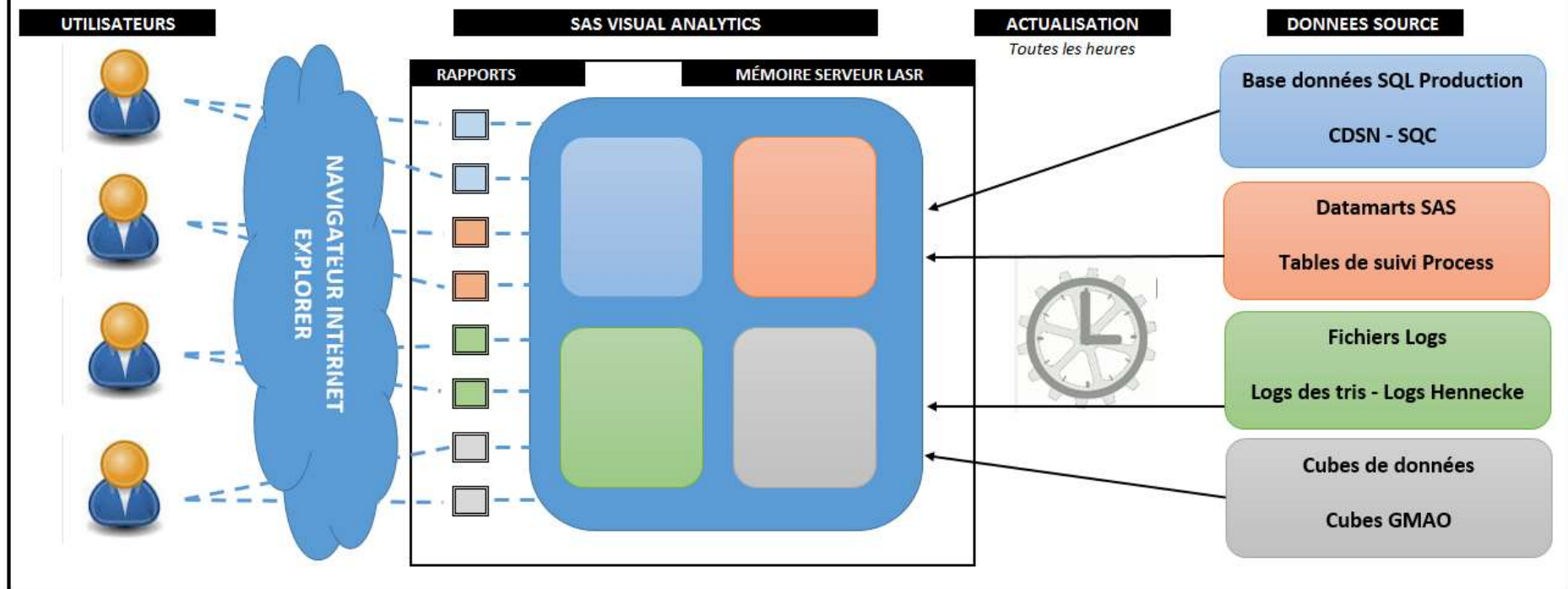
Les principaux atouts de SAS Visual Analytics sont les suivants :

- Accéder à de nombreuses sources de données en temps réel et sans temps d'actualisation,
- Disposer de gros volumes de données sur une grande profondeur et un grand historique,
- Illustrer les analyses par des explications visuelles synthétiques,
- Mieux évaluer certains facteurs en croisant les différentes sources de données,
- Intégrer plusieurs visualisations graphiques complémentaires sur le même rapport,
- Partager les différents rapports de données au sein d'un service,

C'est un outil accessible depuis un navigateur Web (Internet Explorer) qui ne nécessite pas l'installation d'un outil sur le poste de travail.

Dans le contexte PhotoWatt, il a vocation à remplacer l'add-in SAS Excel.

## ARCHITECTURE DU FONCTIONNEMENT DU REPORTING SAS VISUAL ANALYTICS



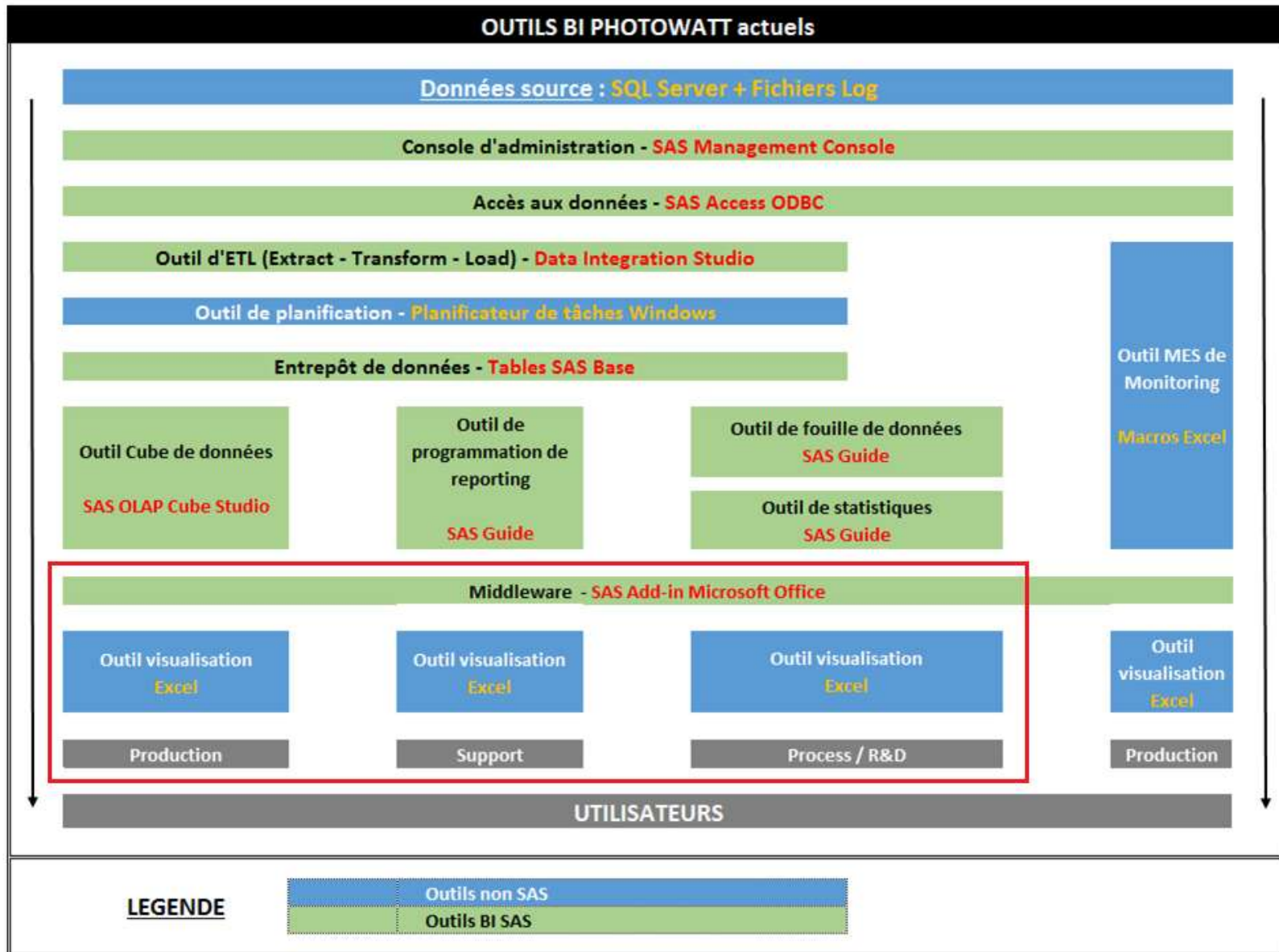
## COMPARATIF SAS VISUAL ANALYTICS - ADD-IN EXCEL

### SAS VISUAL ANALYTICS

- Accès rapide à de gros volumes de données stockées dans la mémoire du serveur
- Pas d'actualisation nécessaire car les données sont automatiquement rafraîchies
- La profondeur de données peut être très importante (de plusieurs mois à plusieurs années possibles)
- Aucune attente nécessaire à l'ouverture des rapports
- Possibilité d'effectuer des filtres élaborés sur les données en temps réel,
- Partage des rapports disponibles par le biais du navigateur Web
- Envoi d'email et export des rapports au format PDF
- Liaison entre les rapports créés
- Rendu au design Web moderne et ergonomique
- Intégration de paramètres dynamiques dans les rapports

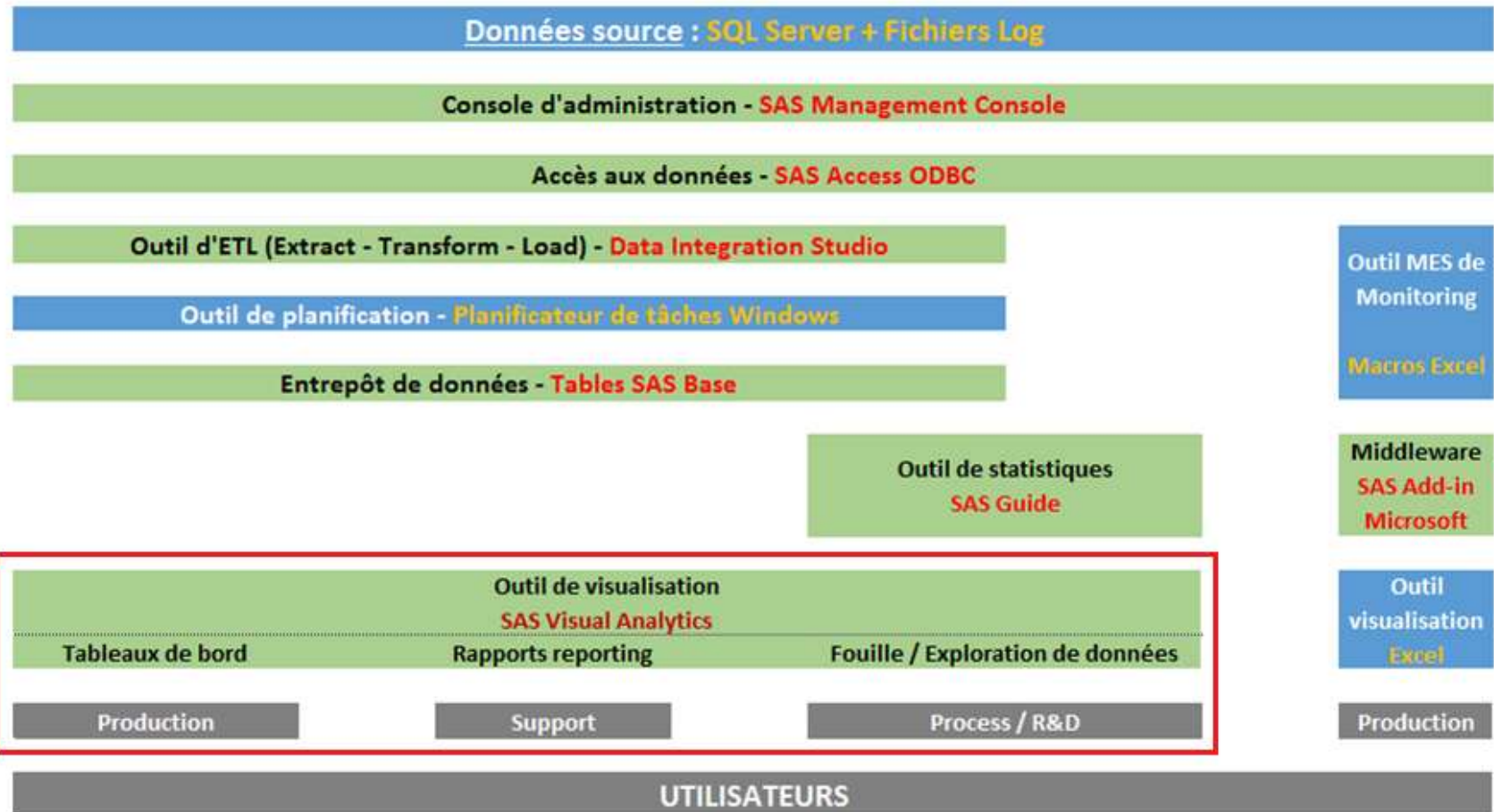
### SAS ADD-IN EXCEL

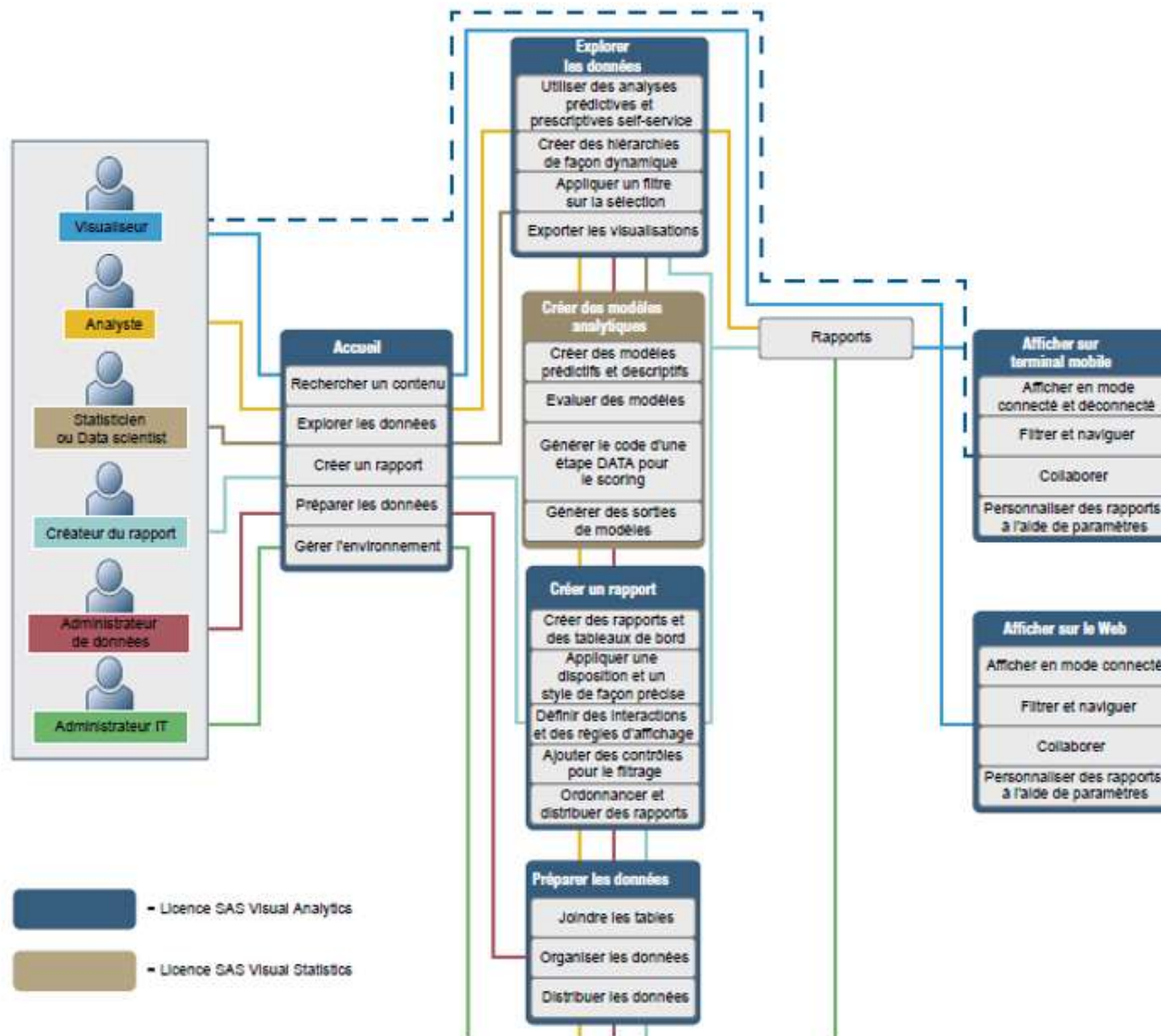
- Temps d'actualisation des données très long chaque matin
- Obligation de devoir limiter les données sur une courte période
- Pas de possibilité de travailler sur de très gros volumes de données
- Perte de connexion fréquente avec l'onglet des données
- Nombreux bugs et dysfonctionnements de l'add-in SAS Excel





## OUTILS BI PHOTOWATT avec intégration de Visual Analytics





## Accès à l'application Visual analytics

On accède au menu de SAS Visual Analytics par le biais de l'Intranet :



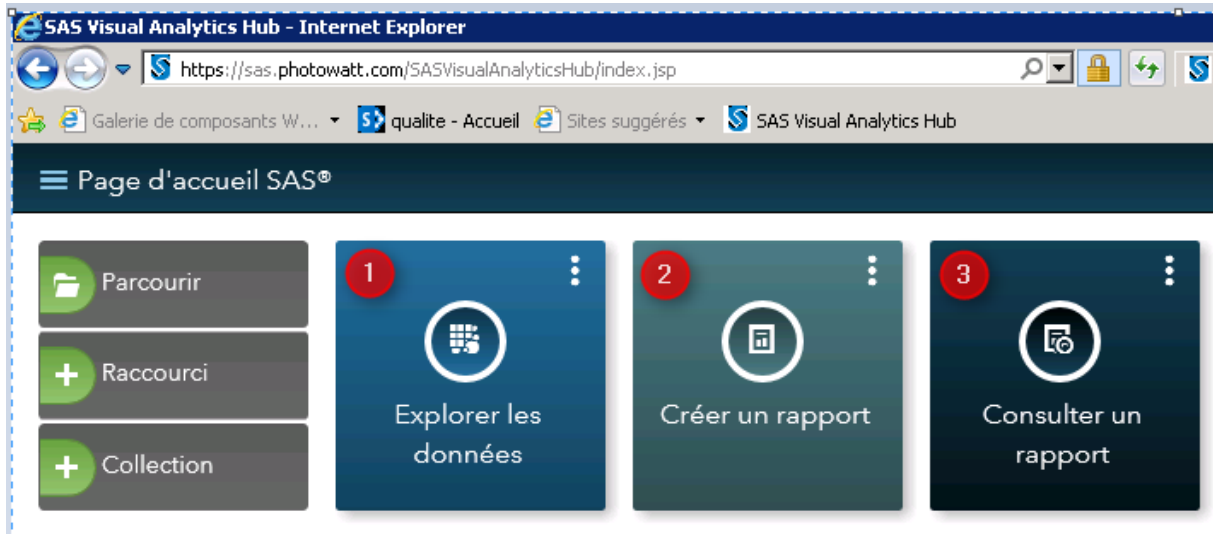
Puis en choisissant « Visual Analytics » dans le menu latéral gauche de la page d'accueil de la rubrique Intranet « Reporting » :



Il est, par ailleurs, tout à fait possible d'enregistrer le lien de d'application directement dans l'onglet des favoris du navigateur Internet Explorer :

<https://sas.photowatt.com/SASVisualAnalyticsHub/index.jsp>

L'accueil de SAS Visual Analytics se matérialise par un menu de 3 icônes :



- Explorer les données :

L'exploration des données permet travailler et fouiller les données de façon brute et improvisée afin de déceler des tendances, des dérives ou des corrélations entre facteurs dans les différentes source de données,

- Créer un rapport :

La création de rapports va permettre de construire des visualisations élaborées, que l'on va pouvoir consulter au fil du temps. La création d'un rapport nécessite de bien savoir ce que l'on veut suivre et observer. C'est pourquoi, cette étape intervient après l'exploration de données, une fois que les facteurs et la tendance à observer sont bien connus,

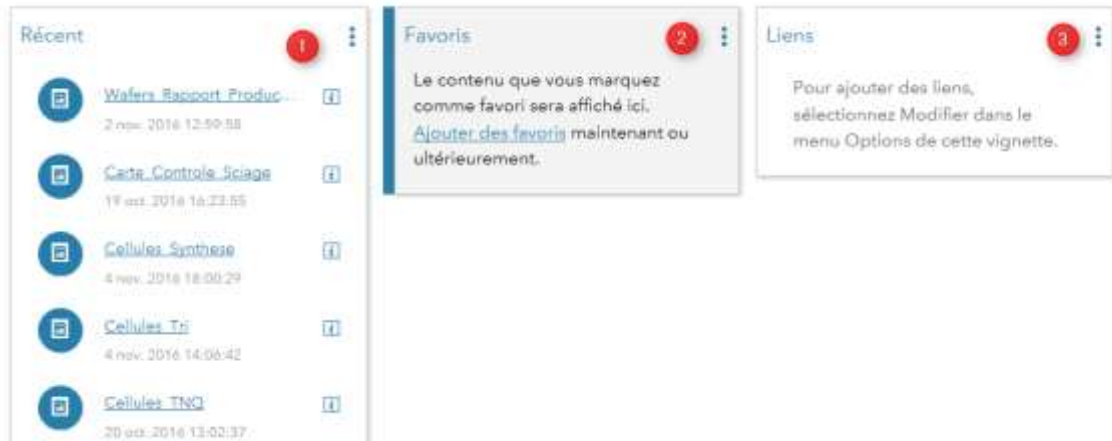
- Consulter un rapport :

La consultation de rapports est la dernière étape. Elle permet de consulter des tableaux de bord prêts à l'emploi, même si l'on n'est pas l'auteur de ces visualisations (à condition d'en avoir tout de même les droits de lecture).

Exemple : les différents tableaux de bord de production Wafers et Cellules disponibles par le biais de la rubrique « Reporting » de l'Intranet,

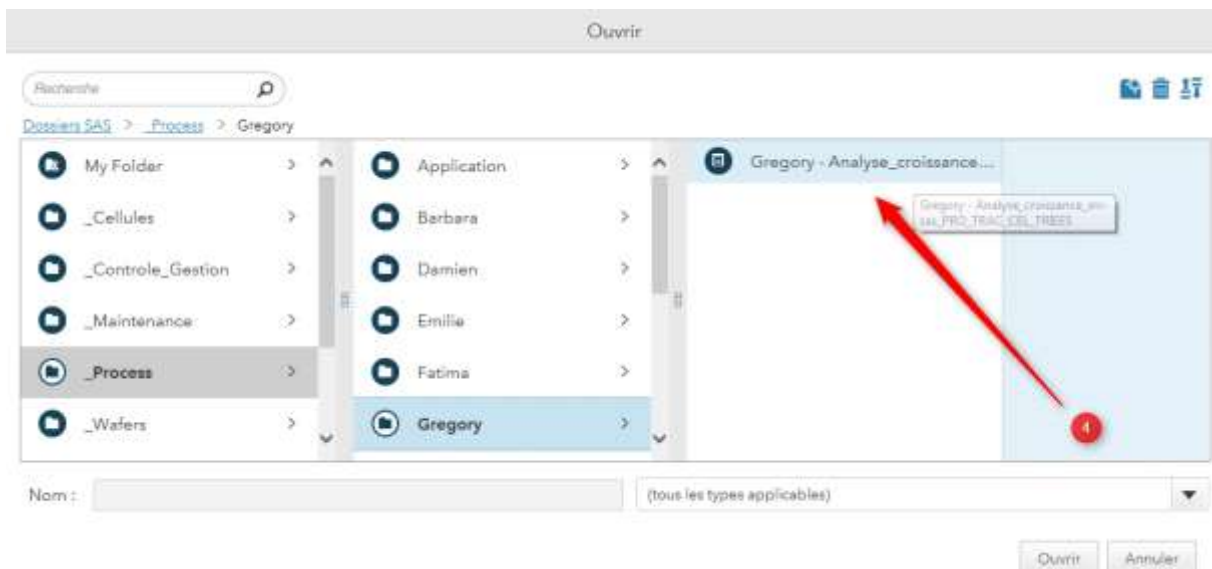
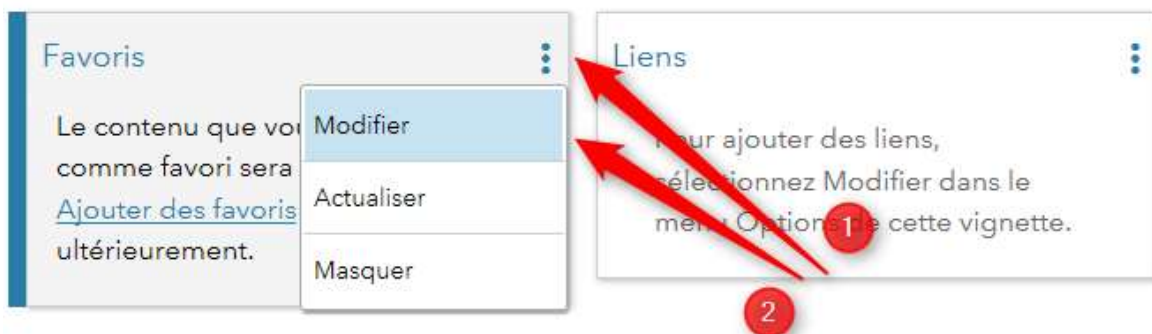
Pour accéder à l'une de ces sections, il suffit de cliquer sur l'icône correspondante.

Le restant de la page est composé de différents blocs contenant des URL d'accès aux rapports et visualisations de données.



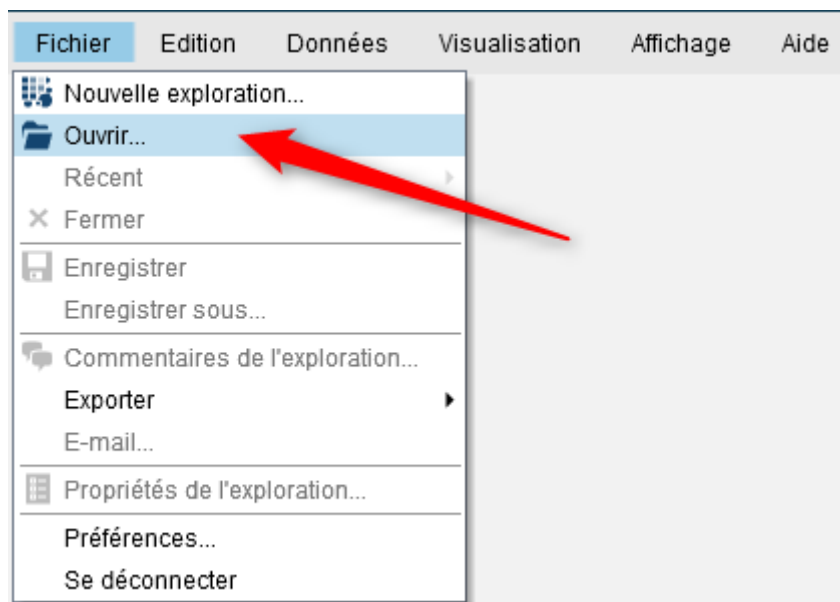
- 1. Le bloc de gauche contient tous les liens des derniers rapports que l'utilisateur a récemment ouverts,
- 2. Le bloc du milieu est configurable et permet d'enregistrer soi-même des liens vers les rapports souhaités,
- 3. Le bloc de droite permet de mettre à disposition des liens vers des rapports pré-définis à tous les utilisateurs de SAS Visual Analytics (cela nécessite de faire une demande auprès des administrateurs de l'application),

Pour configurer un lien, il suffit de cliquer sur les 3 points du bord droit supérieur du bloc :

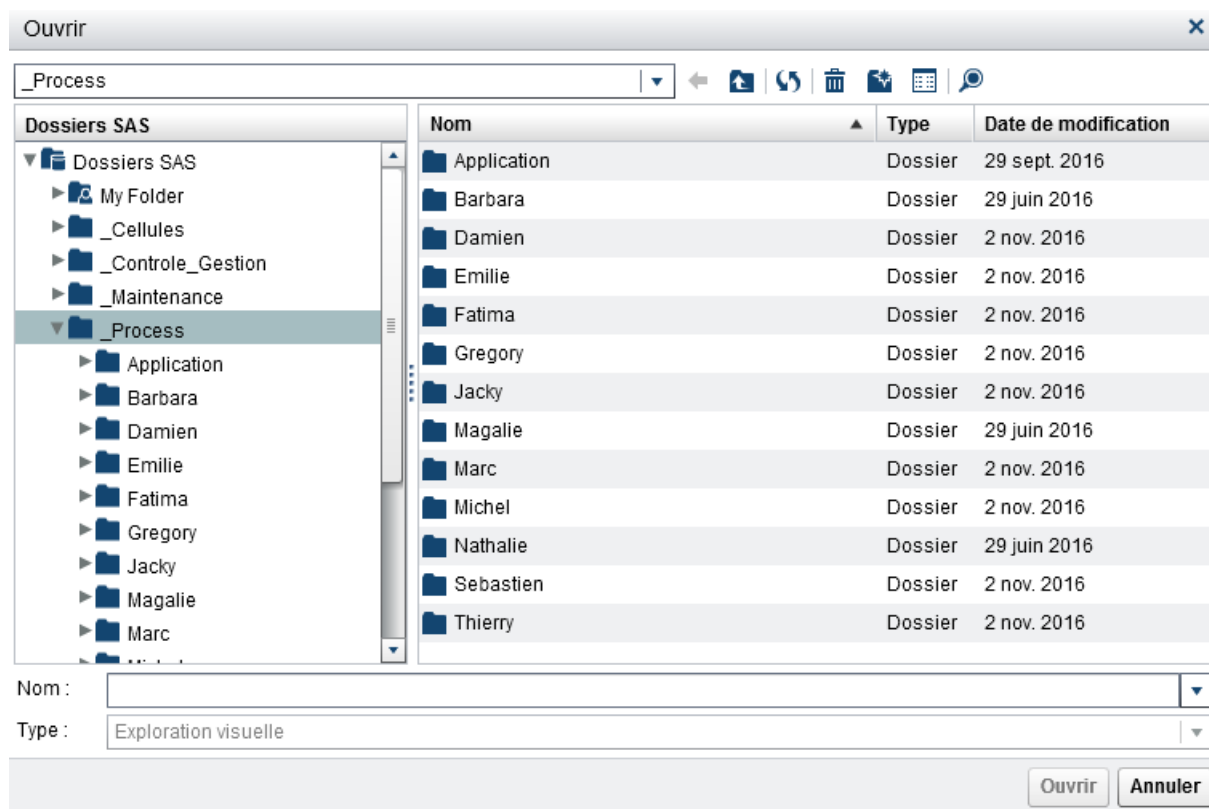


## Ouverture d'une exploration SAS Visual Analytics

Pour ouvrir une exploration Visual Analytics, il faut sélectionner « Ouvrir » dans le menu « Fichier » en haut de l'écran.



Il suffit ensuite de naviguer dans l'arborescence et de double cliquer sur le rapport souhaité :



L'arborescence des rapports Visual Analytics est construite de la sorte :

- Un répertoire par service :
  - UAP Cellules,
  - UAP Wafers,
  - Contrôle de gestion,
  - Maintenance,
  - Process,

Dans le répertoire du service Process, on a un dossier par membre de l'équipe (libellé selon le prénom):

- Barbara,
- Damien,
- Emilie,
- ....
- Thierry,

Tous les membres du service Process ont accès en lecture aux différents répertoires ce qui permet de diffuser des explorations et des rapports au sein du service.

En revanche, seule la personne propriétaire du répertoire a le droit d'effectuer des modifications en écriture sur le contenu du dossier.

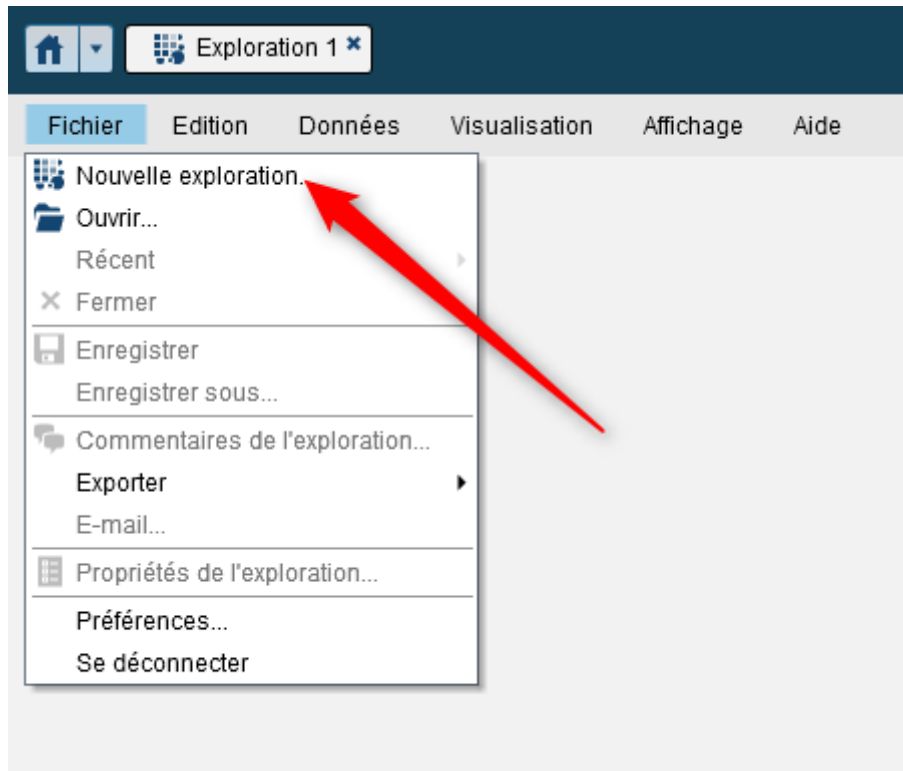
Par exemple, dans le répertoire « Nathalie », tout le service peut consulter les différents rapports contenus mais seule Nathalie pourra modifier ou créer de nouveaux éléments dans cet espace.

A noter, que le répertoire « Application » contient des rapports et des explorations qui sont mises à disposition de tout le service Process comme l'application Chrono par exemple.

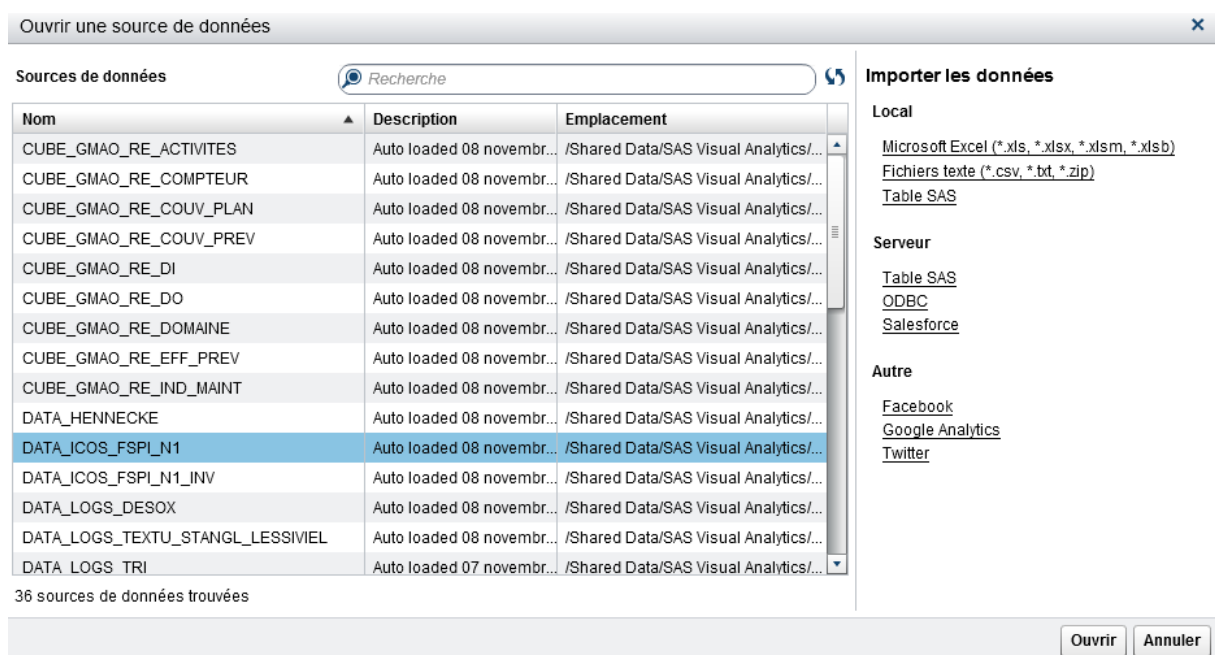


## Choisir une nouvelle source de données

Pour choisir une nouvelle source de données pour l'exploration, il faut cliquer sur « Nouvelle Exploration » dans le menu Fichier en haut de l'écran :



Il faut ensuite double cliquer sur la source de données que l'on souhaite ouvrir :



Les sources de données sont des tables stockées dans la mémoire du serveur Visual Analytics.

Ces tables sont immédiatement accessibles mais uniquement en lecture (aucune modification n'est possible).

Afin de classer l'information contenue dans ces différentes tables, une convention de nommage a été mise en place, ce qui permet de plus facilement trouver les données souhaitées.

Cette convention est la suivante :

- Préfixe AUDIT : Données pour le monitoring d'administration du serveur,
- Préfixe CUBE : Données maintenance issue de la GMAO,
- Préfixe DATA : Données de logs équipements,
- Préfixe DTM : Datamarts condensant l'ensemble des remontées de paramètres Process,
- Préfixe SPC : Données de maîtrise statistique des procédés pour les cartes de contrôle,
- Préfixe TDB : Données à destination des tableaux de bord de production,

35 tables de données sont actuellement disponibles dans la mémoire du serveur.

Ces tables de données sont actualisées selon une fréquence de mise à jour dépendant du type de données.

De même la profondeur de données (l'ancienneté de l'historique) est variable selon la table et le type de données.

A noter qu'il est possible de rajouter de nouvelle source de données, sur demande auprès de l'équipe informatique (dans la mesure des possibilités techniques).



## Liste des sources de données disponibles :

Type table	Destinataire	Objet	Actualisation	Profondeur de données	Tables
Système	Administration	Monitoring des accès SAS VA	Toutes les 15 minutes	2 mois	AUDIT_VISUAL_ANALYTICS
Cube de données	Maintenance	Suivi des activités des intervenants Maintenance	Quotidienne (à 8 h chaque matin)	3 ans	CUBE_GMAO_RE_ACTIVITES
		Suivi des relevés de compteur Maintenance			CUBE_GMAO_RE_COMPTEUR
		Couverture du plan de maintenance			CUBE_GMAO_RE_COUV_PLAN
		Couverture des préventifs / parc équipement			CUBE_GMAO_RE_COUV_PREV
		Suivi des demandes d'intervention GMAO			CUBE_GMAO_RE_DI
		Suivi des disponibilités opérationnelles			CUBE_GMAO_RE_DO
		Ventilation des demandes d'intervention / domaine			CUBE_GMAO_RE_DOMAINE
		Suivi des réalisations effectives des préventifs			CUBE_GMAO_RE_EFF_PREV
		Calcul des indicateurs maintenance de type MTTR			CUBE_GMAO_RE_IND_MAINT
Datamart Process	Service Process	Suivi des paramètres Process des lingots	Toutes les heures	2 ans	DTM_PRO_TRAC_LINGOTS
		Suivi des paramètres Process des Wafers		2 ans	DTM_PRO_TRAC_WAFERS
		Suivi des paramètres Process des cellules		2 ans	DTM_PRO_TRAC_CEL_TRIEES
		Suivi de production des cellules		5 ans	DTM_PRO_TBO_PVA
		Suivi des rendements électriques des fichiers de tri		6 mois	DTM_LOGS_TRI
		Statistiques des rendements des fichiers de tri		6 mois	DTM_LOGS_TRI_STATS
Tableaux de bord production	Cellules	Tableau de bord des TNQ Cellules	Toutes les heures	2 mois	TDB_DTM_PRO_PVA
		Tableau de bord des tris Cellules			TDB_V_LOTS
	Wafers	Tableau de bord Rapport production - Contrôle Wafers		2 mois	TDB_Wafers_Controlle
		Tableau de bord Rapport production - MEP			TDB_V_LOTS_MEP

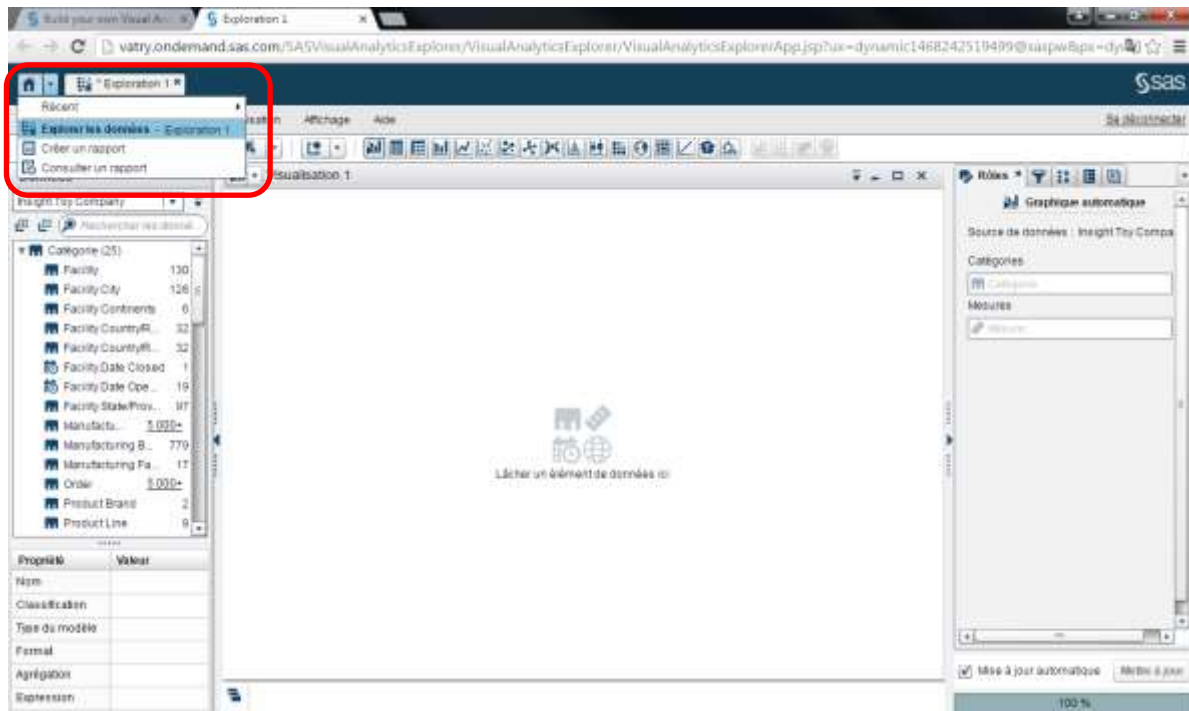
	Cellules / Fiabilité	Tableau de bord des consommations d'écran		2 ans	TDB_Conso_mation_Ecran
	Cellules / Gestion	Tableau de bord Synthèse - PW Diff à tri		2 mois	TDB_V_LOTS_Synthese_1
		Tableau de bord Synthèse - PVA MEP à Diff			TDB_V_LOTS_Synthese_2
		Tableau de bord Synthèse - En cours de stock			TDB_V_LOTS_Synthese_3
	Suleyman CETIN	Fichier Base Cellule pour responsable UAP		4 mois	TDB_BASE_CELLULES
Exploitation de logs équipement	Process / Qualité	Logs équipement de contrôle Hennecke	Quotidienne (à 7 h chaque matin)	1 an	DATA_LOGS_HENNECKE
	Méthodes	Logs équipement des palpeurs croissance		7 jours	DATA_PALPEUR
	Process	Logs des équipements de tri cellules		3 ans	DATA_LOGS_TRI
	Process / Qualité	Logs équipement sérigraphie ICOS FSPI N1		4 mois	DATA_ICOS_FSPI_N1
	Process	Logs équipement Textu - Stangl et Lessivielle		2 ans	DATA_LOGS_Textu_Stangl_Lessiviel
	Process	Logs équipement Désoxydation		2 ans	DATA_LOGS_DESOX
Cartes de contrôle	Process / Qualité	Cartes de contrôle de l'atelier sciage	Toutes les heures	3 mois	SPC_CC_Sciage
		Calcul des limites des cartes de contrôle sciage		4 mois	SPC_CC_SCIAGES_LIMITES
		Cartes de contrôle des mesures sérigraphie		2 ans	SPC_CC_MESURE_SERIGRAPHIE
		Cartes de contrôle des défauts ICOS FSPI N1		3 mois	SPC_CC_ICOS_N1
Etude de corrélation	Process / Qualité	Données des paramètres Slurry des scies	Ponctuel	1 an	SPC_DATA_SLURRY



# L'exploration de données

Quand on travaille sous SAS Visual Analytics, on a la possibilité d'effectuer de la fouille de données, c'est-à-dire d'explorer les différentes tendances et les corrélations sous-jacentes que décèlent les données, que l'on veut étudier.

Le but de l'exploration de données est de vérifier des hypothèses, faire des essais, creuser des intuitions, pour ensuite créer le rapport correspondant :



Pour l'exploration des données, on sélectionne cet icône :



Les objets pour l'analyse de données sont répertoriés sous forme d'icônes en haut de la fenêtre de visualisation :

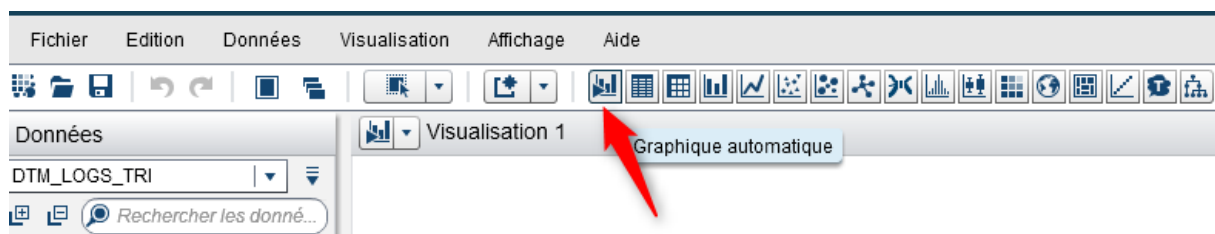


Les différentes explorations disponibles sont les suivantes :

- Diagramme de Sankey
- Boîte à moustache
- Arbre de décision
- Matrice de corrélation
  
- Carte thermique
- Nuage de mots
- Distribution
- Diagramme de réseau

Pour sélectionner une exploration, il suffit de cliquer sur l'icône correspondante.

A noter, qu'il est possible de laisser le système choisir l'exploration la plus appropriée aux données. Pour cela il suffit de cliquer sur l'icône suivante :

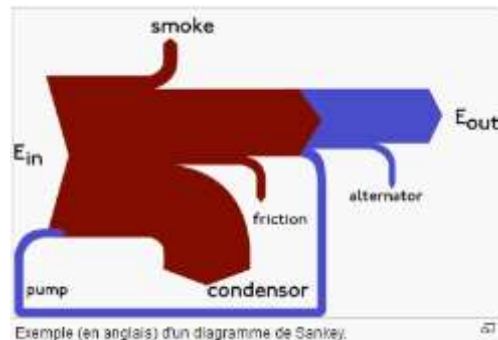




# Le diagramme de Sankey



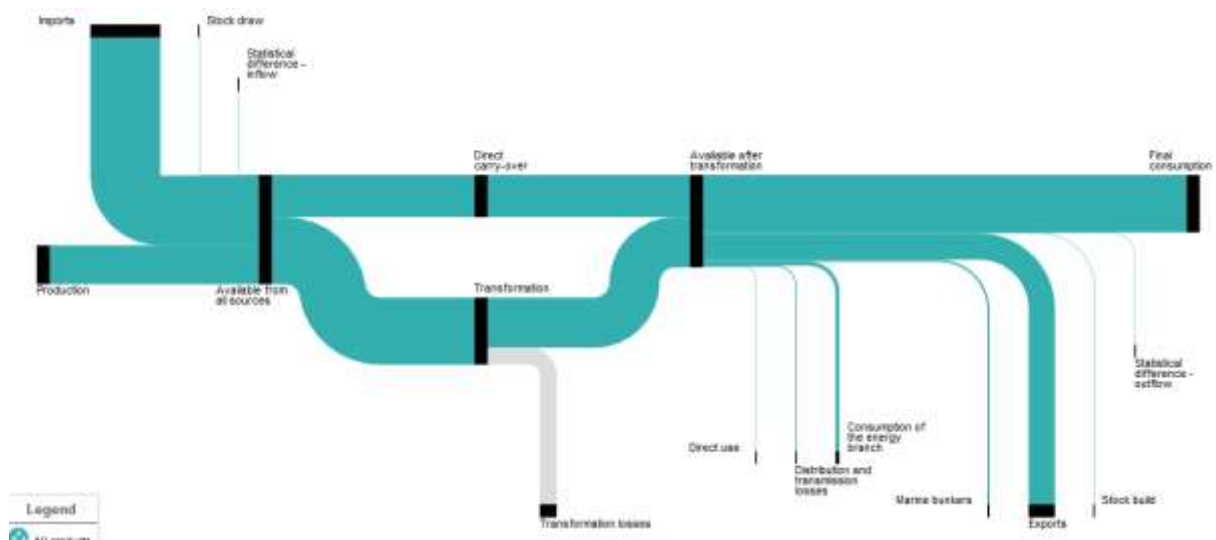
**Un diagramme de Sankey** ou diagramme Sankey est un type de diagramme de flux, dans lequel la largeur des flèches est proportionnelle au flux représenté. Le diagramme de Sankey est utilisé en particulier pour visualiser les transferts énergétiques, les coûts ou les pertes engendrées par un processus.



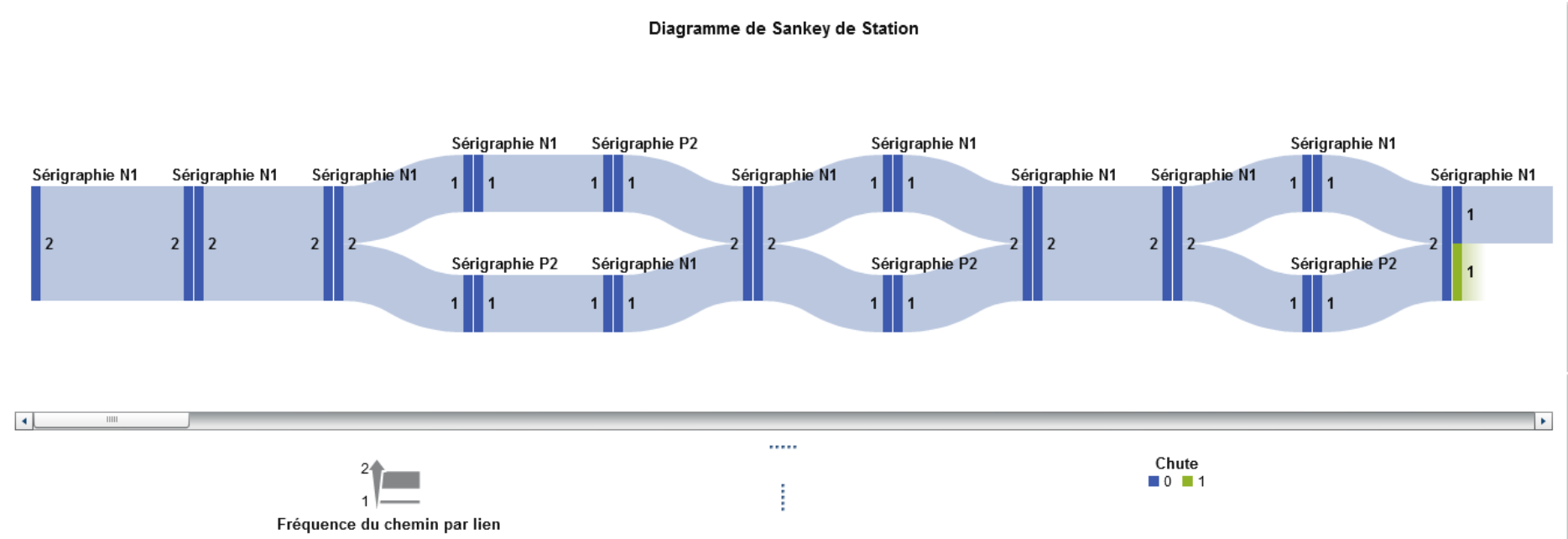
Les anglo-saxons nomment cette visualisation "diagramme de Sankey" en hommage au capitaine irlandais Matthew Henry Phineas Riall Sankey (1853 - 1926), qui a utilisé ce type de diagramme dès 1898 dans une publication sur l'efficacité énergétique d'une machine à vapeur.

Dans le cas de systèmes fonctionnant en cycle clos, les diagrammes de Sankey montrent les étapes de la conservation du système, comme la conservation de la masse ou de la conservation de l'énergie. À l'opposé, en système ouvert, on met en exergue les modifications de quantités dans le système, comme l'énergie.

Dans le cas où le transfert important observé est une perte, le diagramme de Sankey peut servir à identifier les points faibles d'un processus, et ainsi de définir les phases de ce processus sur lesquelles il est nécessaire de mettre l'accent en matière d'optimisation des performances, par exemple dans un procédé industriel ou mécanique.



Exemple de diagramme de Sankey :

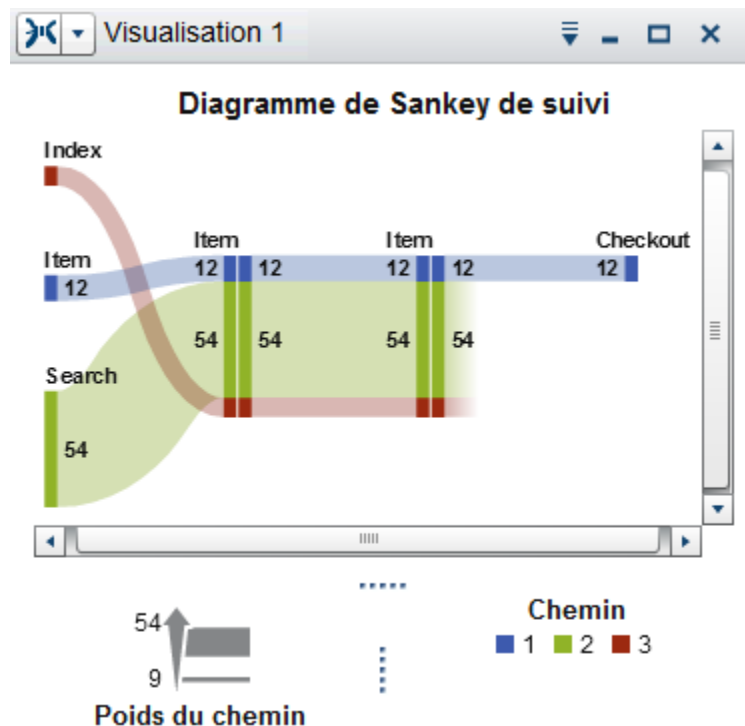


Changement des écrans sérigraphie en fonction de la station de l'équipement

**Cette exploration** affiche une série de noeuds liés, où la largeur de chaque lien indique la fréquence du lien ou la valeur d'une mesure. Un diagramme de Sankey permet d'effectuer une analyse des chemins. L'analyse des chemins affiche des flux de données d'un événement (valeur) à un autre, comme une série de chemins.

Exemple :

- flux d'énergie (Il permet de comprendre visuellement où et quand se passent les transferts les plus importants à l'aide de pourcentage de perte/récupération par rapport à un ou plusieurs flux initiaux.),
- analyse des parcours des clients sur les sites internet



On veut analyser le cheminement des utilisateurs d'un site internet, les pages visitées.

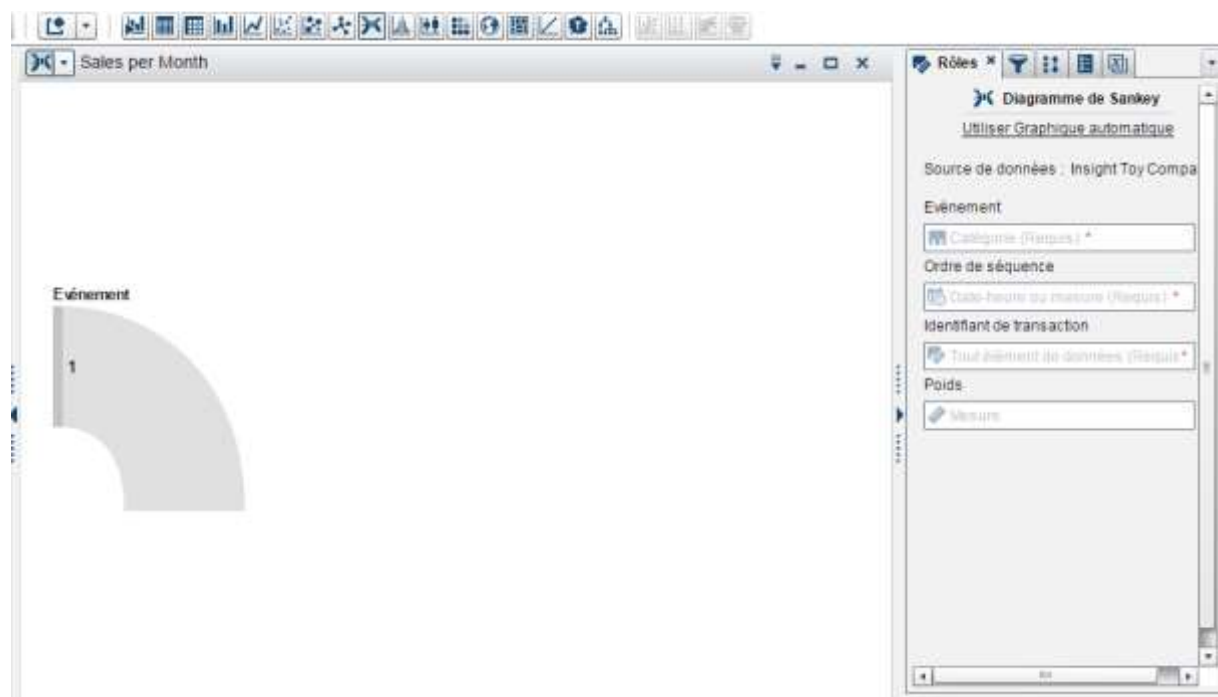
On dispose d'un jeu de données qui comprend le nom des utilisateurs (customer), un identifiant pour la connexion (transID), un identifiant pour les pages Web visitées (item) et l'ordre des pages visitées (sequence).

Le jeu de données est présenté ci-dessous :

	customer	transID	item	sequence
1	John	1	A	1
2	John	1	B	2
3	John	1	C	3
4	Jane	2	B	1
5	Jane	2	D	2
6	Jane	2	E	3
7	Jane	2	E	4
8	Jane	2	D	5
9	John	3	A	1
10	John	3	D	2
11	John	3	E	3
12	Bob	4	A	1
13	Bob	4	F	2
14	Bob	4	D	3

On double clic sur l'icône Diagramme de Sankey  dans la barre à outils « Objets ».

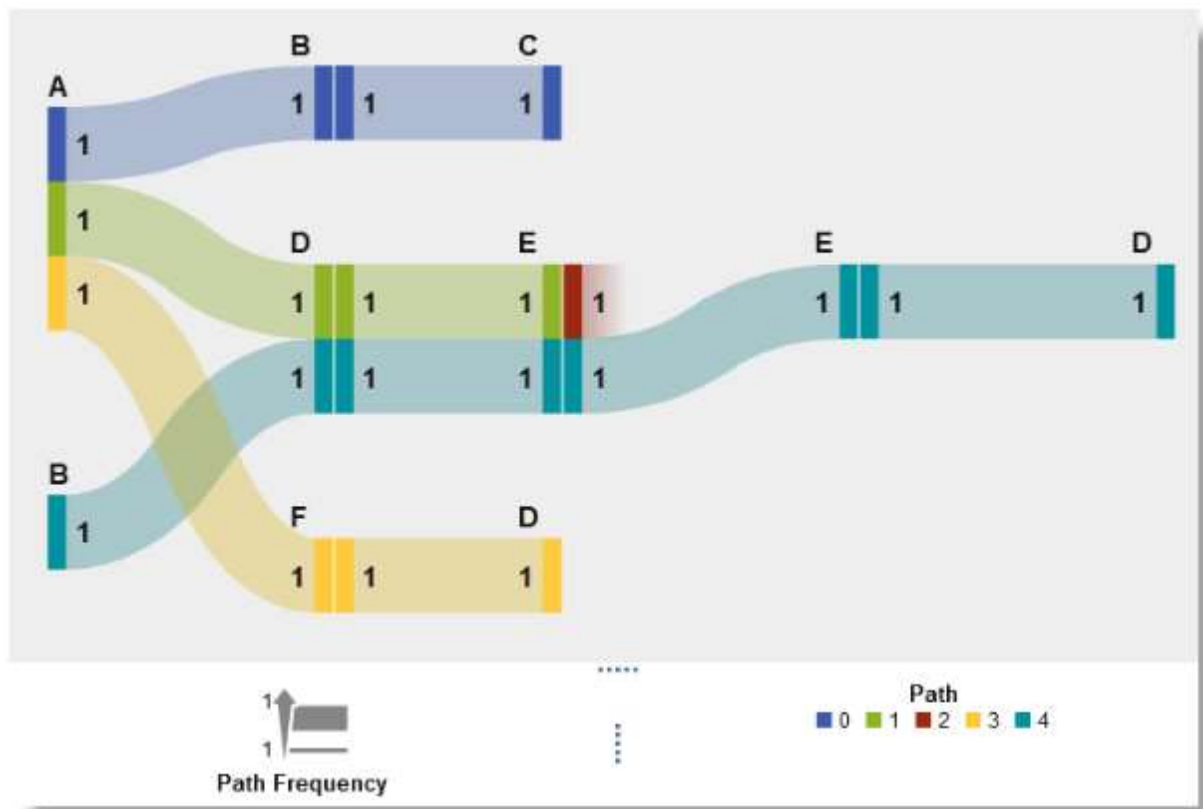
On obtient le graphique vierge :



Dans l'onglet « Rôles » à droite, on complète :

- Item pour l'Evénement
- Sequence pour l'Ordre de séquence
- TransID pour l'identifiant de la transaction

Le graphique se met à jour :



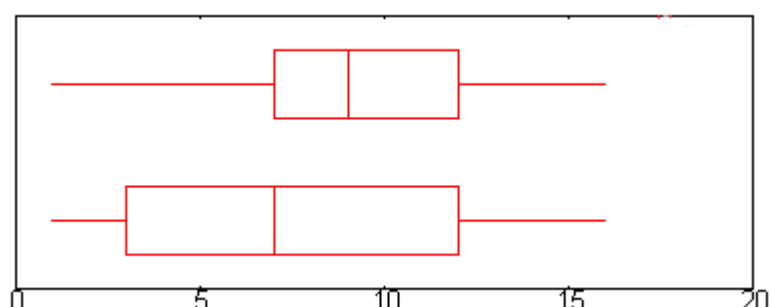
Le diagramme montre 5 chemins différents incluant une coupure (le chemin/path 2 en rouge). Il permet de mettre en avant que 2 chemins sont partiellement en commun (le vert/John et le turquoise/Jane). La coupure indique qu'un des 2 chemins, après une partie commune, s'est arrêté.



# Boîte à moustache



**Le Boxplot :** Dans les représentations graphiques de données statistiques, la boîte à moustaches (aussi appelée diagramme en boîte, boîte de Tukey ou box plot en anglais) est un moyen rapide de figurer le profil essentiel d'une série statistique quantitative. Elle a été inventée en 1977 par John Tukey, mais peut faire l'objet de certains aménagements selon les utilisateurs. Son nom est la traduction de Box and Whiskers Plot.



La boîte à moustaches résume seulement quelques caractéristiques de position du caractère étudié (médiane, quartiles, minimum, maximum ou déciles). Ce diagramme est utilisé principalement pour comparer un même caractère dans deux populations de tailles différentes.

Il s'agit de tracer un rectangle allant du premier quartile au troisième quartile et coupé par la médiane.

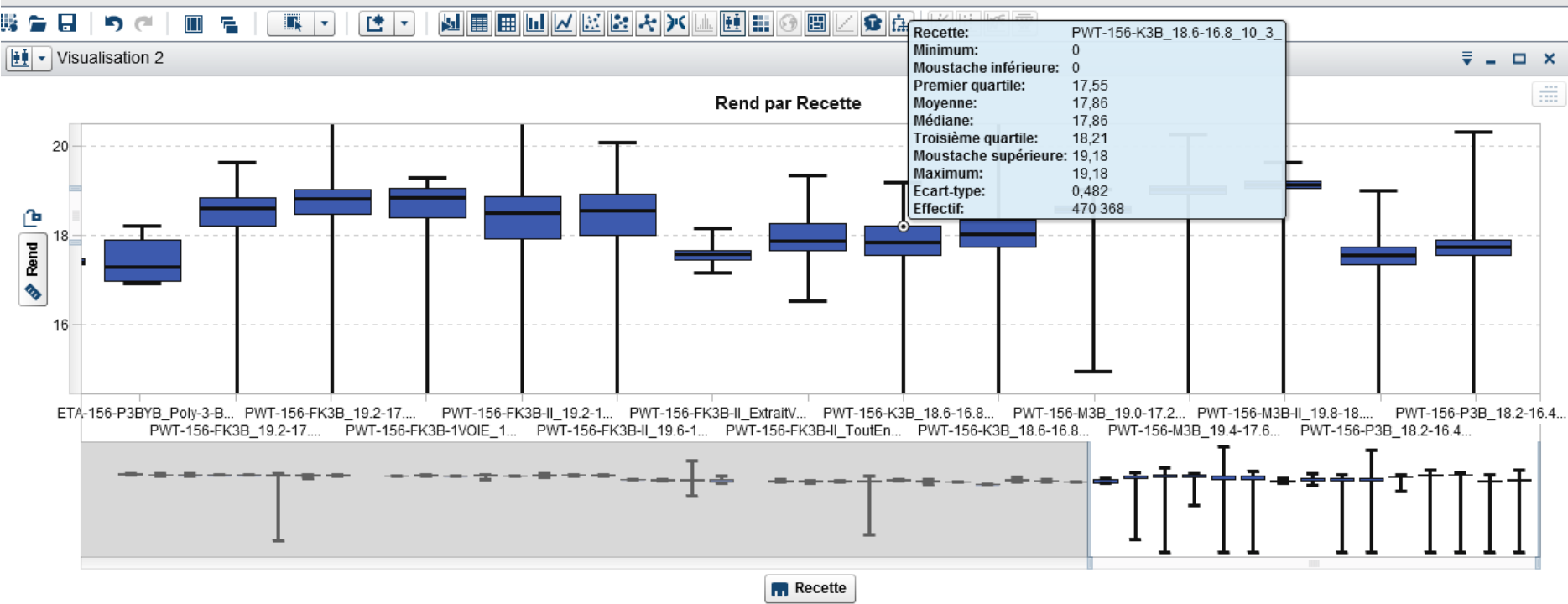
Une boîte à moustaches nous indique de façon visuelle les traits marquants de la série observée :

- La médiane nous renseigne sur le milieu de la série ;
- Les largeurs des deux parties de la boîte mesurent la dispersion des valeurs au centre de la série (la boîte contient 50% des observations : 25% à gauche et 25% à sa droite),
- La longueur des moustaches renseigne sur la dispersion des valeurs aux extrémités (les valeurs les plus petites correspondant à 25% des observations et vice versa),
- De façon générale, la boîte et les moustaches seront d'autant plus étendues que la dispersion de la série statistique est grande.

Dans les diagrammes en boîte de Tukey, la longueur des « moustaches » vaut 1,5 fois l'écart interquartile.

*L'usage des boîtes à moustaches permet de visualiser les concepts de centralité et de dispersion (de même que de symétrie ou bien d'asymétrie). Elles sont donc alors particulièrement recommandées lorsqu'on veut ainsi comparer des séries statistiques ou bien des distributions entre elles.*

Exemple de BoxPlot sur les données des logs de tri de l'année 2016 :

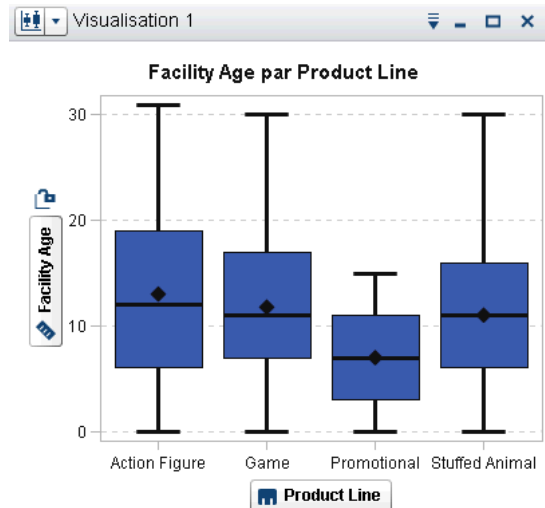


Distribution des rendements selon les différents types de recette



**Cette exploration** affiche les données sous forme de boîte à moustaches. La distribution des valeurs est affichée pour une seule mesure à l'aide d'une boîte et de moustaches. La taille et l'emplacement de la boîte indique la plage de valeurs entre le 25ème et le 75ème centile. Des informations statistiques supplémentaires sont représentées par d'autres fonctionnalités visuelles.

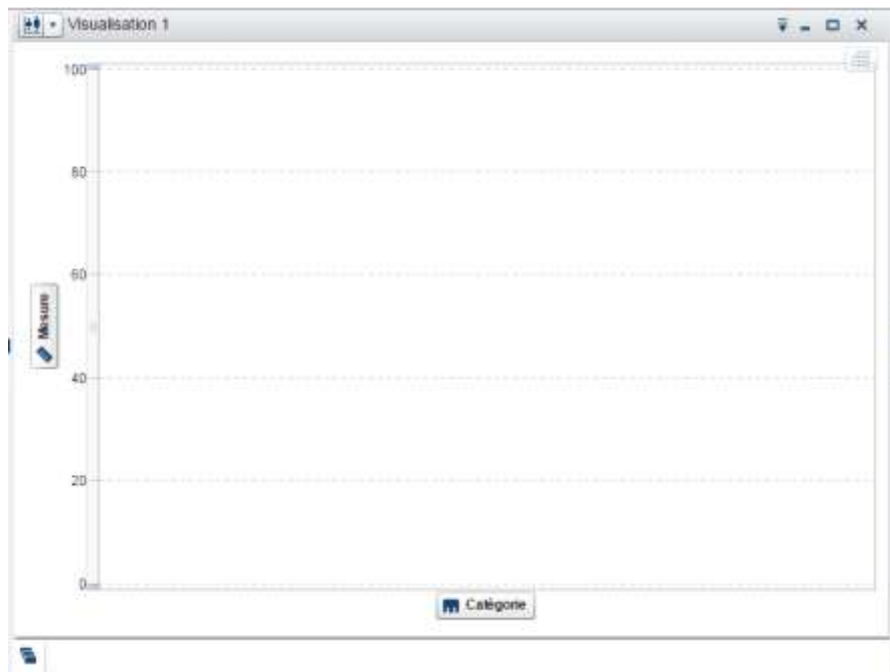
Vous pouvez créer des treillis et indiquer si la valeur moyenne (mean) et les valeurs hors norme sont affichées pour chaque boîte.



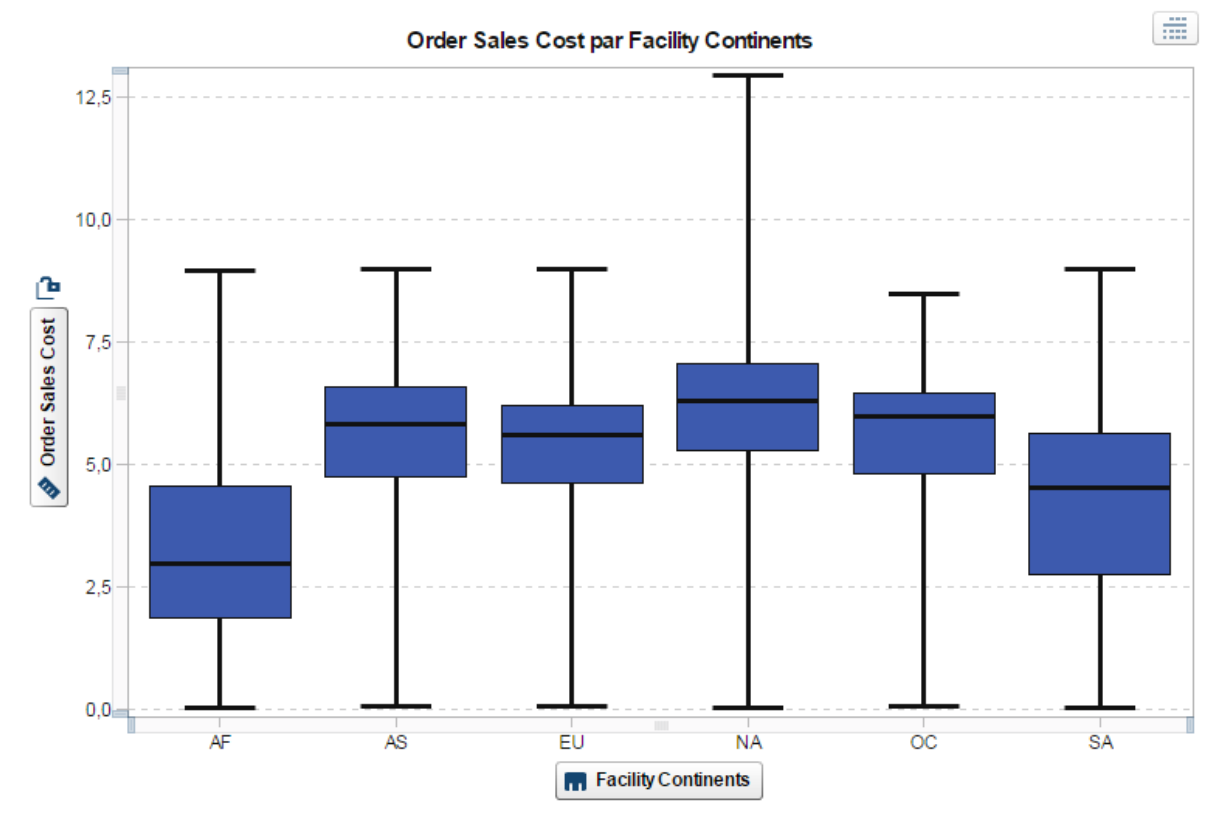
On veut exprimer le coût des bons de commande (Order Sales Cost) par continent.

On double clic sur l'icône Boîte à moustache  dans la barre à outils « Objets ».

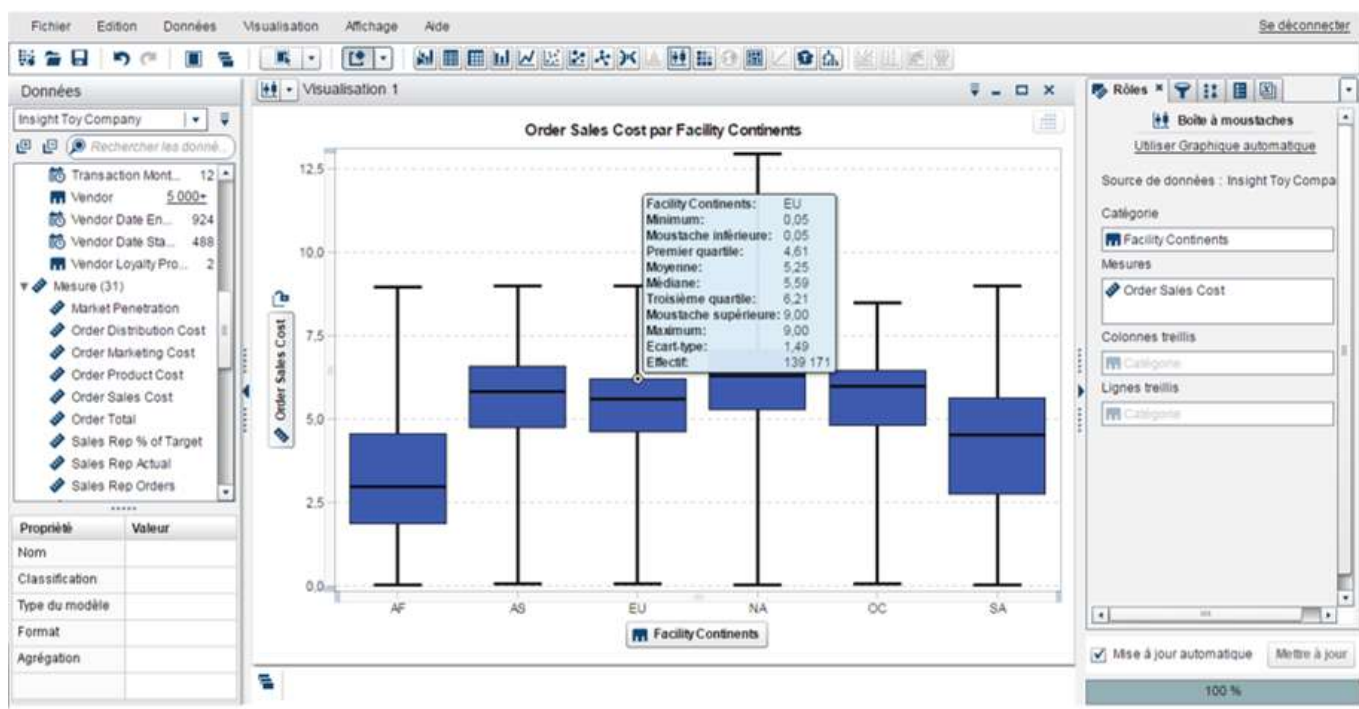
On obtient le graphique vierge :



Déplacer la variable Catégorie « Facility Continents » vers l'axe des abscisses et la variable Mesure « Order Sales Cost » vers l'axe des ordonnées, et le graphique est mis à jour :



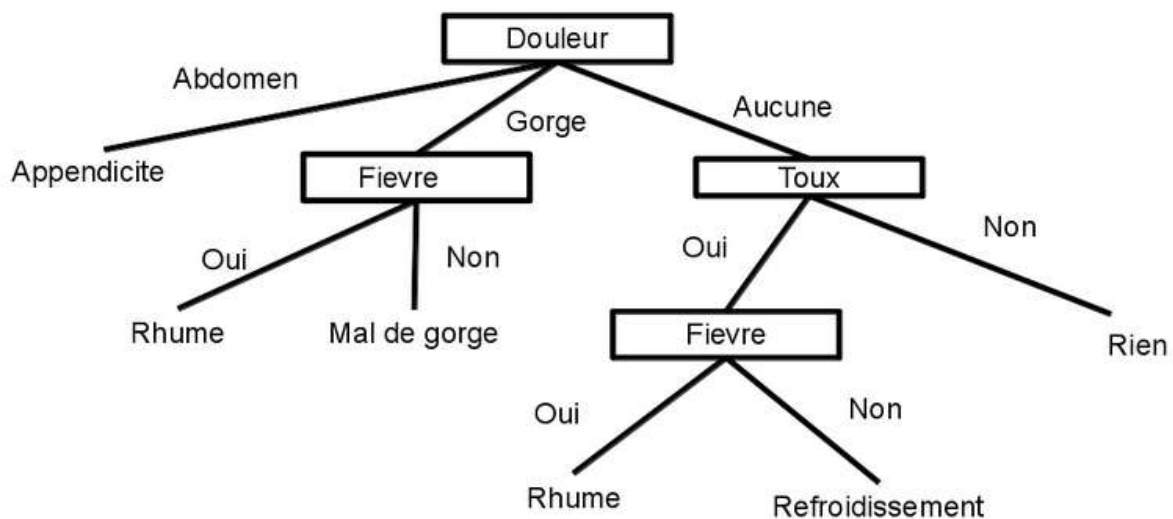
En plaçant le curseur sur une des boîtes à moustaches, on obtient le détail des statistiques :



## Arbre de décision



**L'arbre de décision :** Un arbre de décision est un outil d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont situées aux extrémités des branches (les « feuilles » de l'arbre), et sont atteints en fonction de décisions prises à chaque étape. L'arbre de décision est un outil utilisé dans des domaines variés tels que la sécurité, la fouille de données, l'industrie, etc. Il a l'avantage d'être lisible et rapide à exécuter. Il s'agit de plus d'une représentation calculable automatiquement par des algorithmes d'apprentissage supervisé.



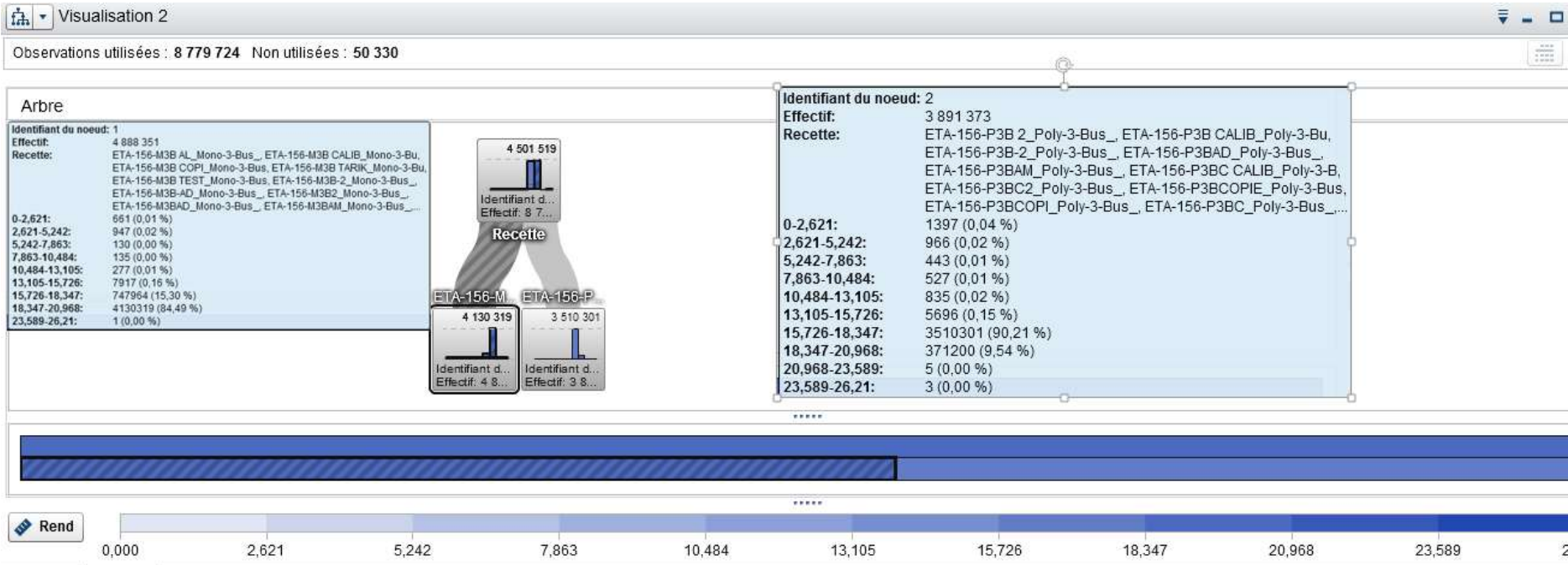
Les arbres de décision sont utilisés dans des domaines d'aide à la décision (par exemple l'informatique décisionnelle) ou l'exploration de données.

Ils décrivent comment répartir une population d'individus (clients d'une entreprise, produits d'une industrie, ...) en groupes homogènes selon un ensemble de variables discriminantes (âge, matériau, nombre de défauts, catégorie socio-professionnelle, ...) et en fonction d'un objectif fixé (aussi appelé « variable d'intérêt » ou « variable de sortie » ; par exemple : rendement, chiffre d'affaires, probabilité de cliquer sur une publicité, ...).

Un avantage majeur des arbres de décision est qu'ils peuvent être calculés automatiquement à partir de bases de données par des algorithmes d'apprentissage supervisé. Ces algorithmes sélectionnent automatiquement les variables discriminantes et peuvent permettre d'extraire des règles logiques de cause à effet (des déterminismes) qui n'apparaissaient pas initialement dans les données brutes.

*Globalement, l'arbre de décision permet de créer des groupes homogènes par les données, ce qui permet de mettre en évidence certaines tendances explicatives (exemple : mettre en valeur un facteur influant sur le rendement électrique).*

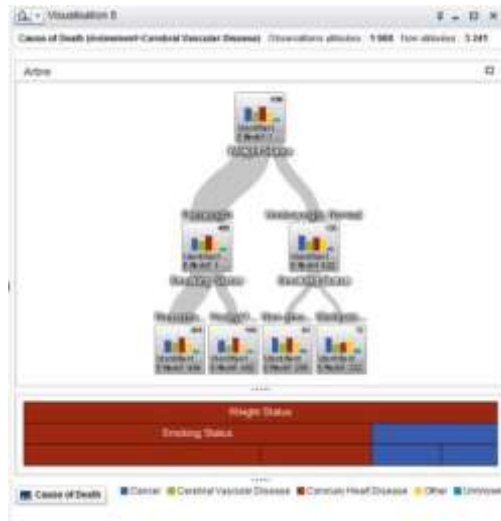
**Exemple d'arbre de décision sur les données logs de tri :**



Classement des recettes en 2 groupes selon la valeur de leur rendement

**Cette exploration** affiche les données sous forme d'arbre de décision. Un arbre de décision affiche une série de nœuds sous forme d'arborescence, où le nœud supérieur est l'élément de données de réponse, et chaque branche de l'arbre représente une division dans les valeurs d'un élément de données prédicteur.

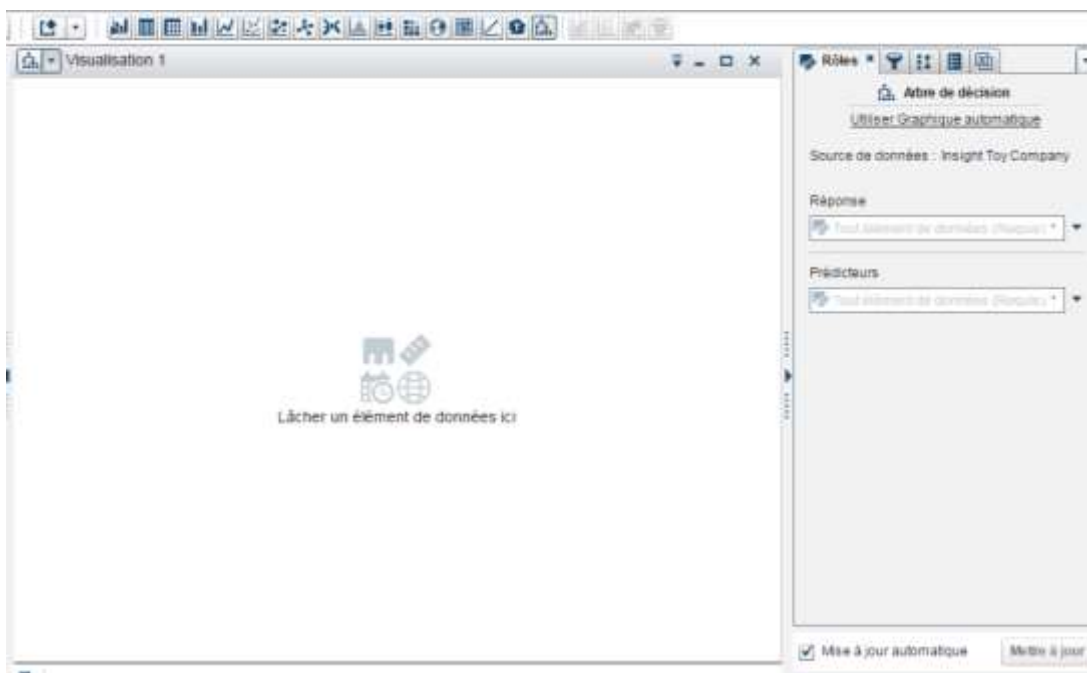
Les divisions permettent de voir quelles valeurs de l'élément de données prédicteur correspondent aux différentes distributions des valeurs dans l'élément de données de réponse.



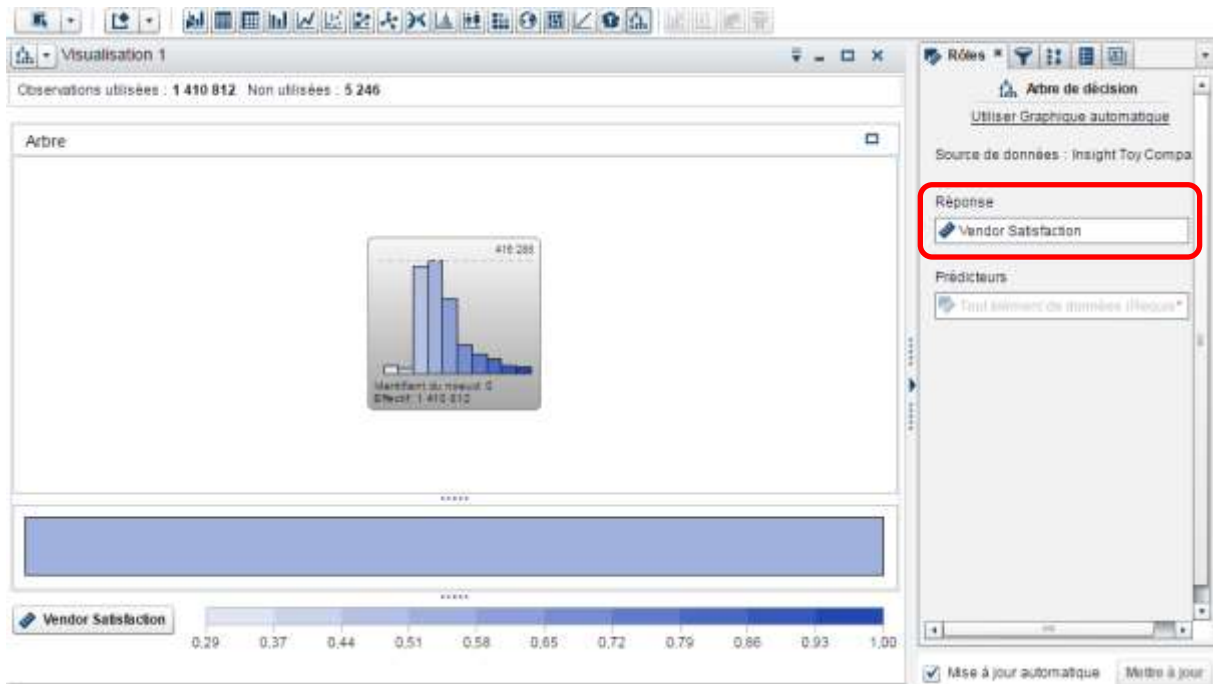
On voudrait prédire la satisfaction des vendeurs (Vendor Satisfaction) par rapport à leur note (Vendor Rating) et leur distance (Vendor Distance).

On double clic sur l'icône Arbre de décision  dans la barre à outils « Objets ».

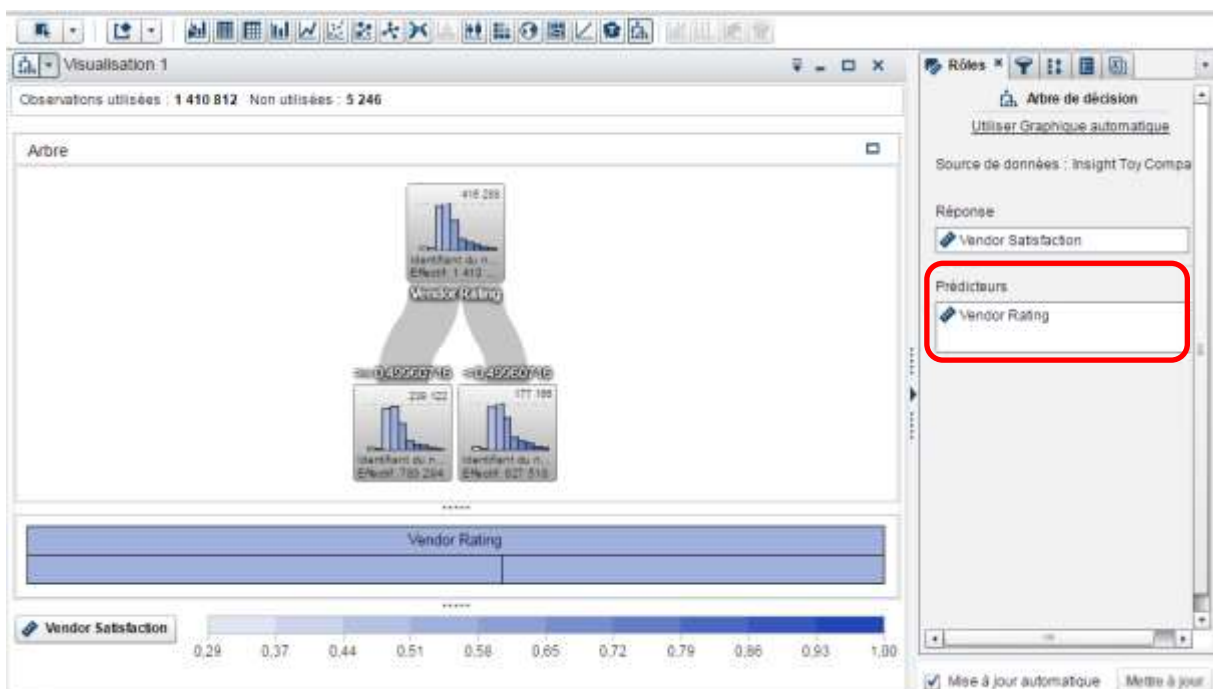
Il reste à définir les données qui correspondent à « Réponse » (Target) et à « Prédicteurs » (Predictor) dans l'onglet « Rôles » en haut à droite.



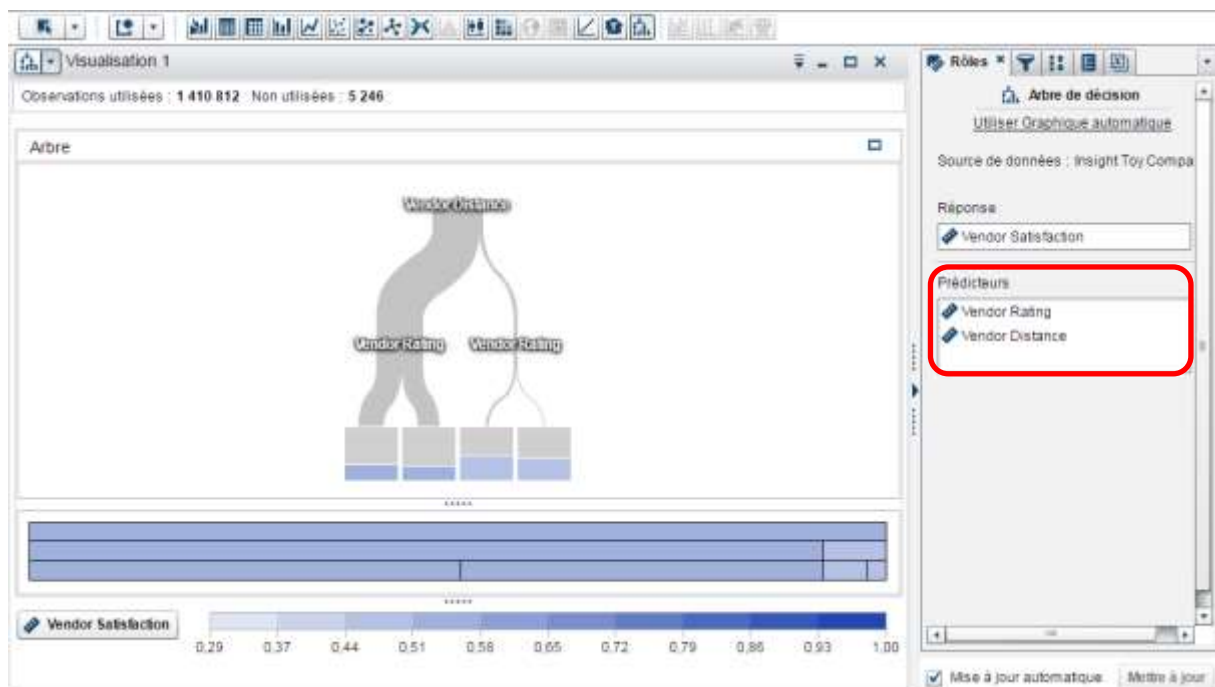
On sélectionne la donnée « Vendor Satisfaction » et on la glisse sur la droite à l'emplacement « Réponse ». La répartition de cette variable est alors affichée.



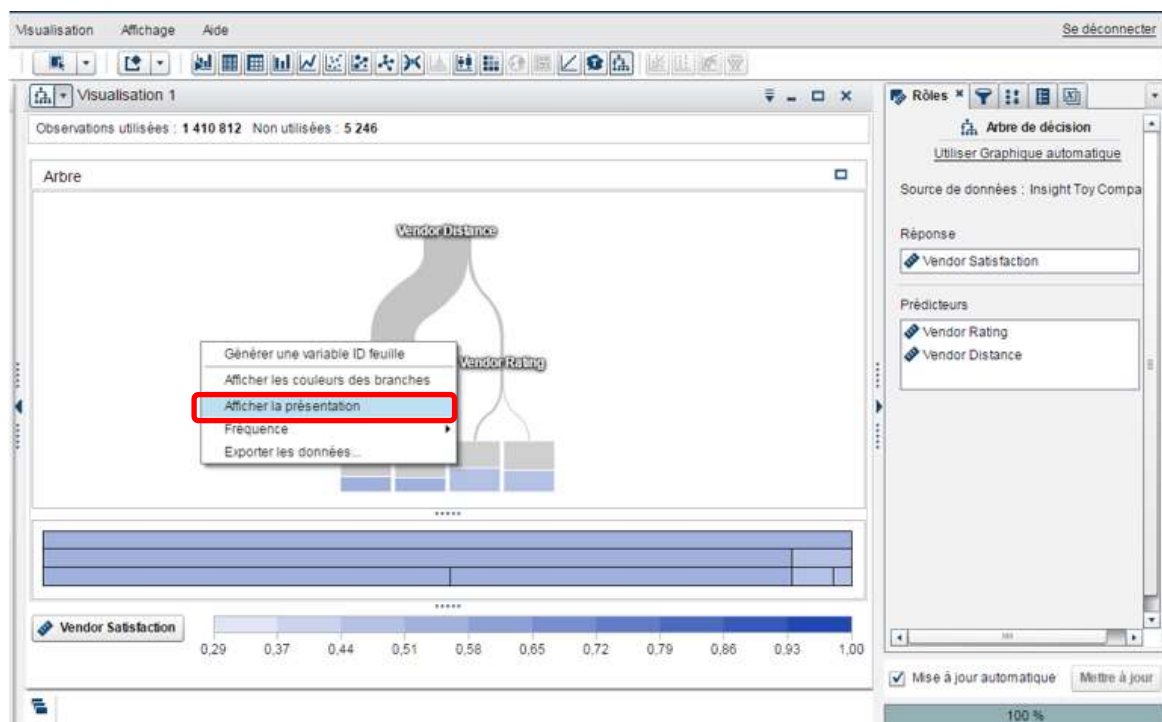
On sélectionne alors la donnée « Vendor Rating » et on la glisse sur la droite à l'emplacement « Prédicteurs ». L'arbre de décision est alors créé.



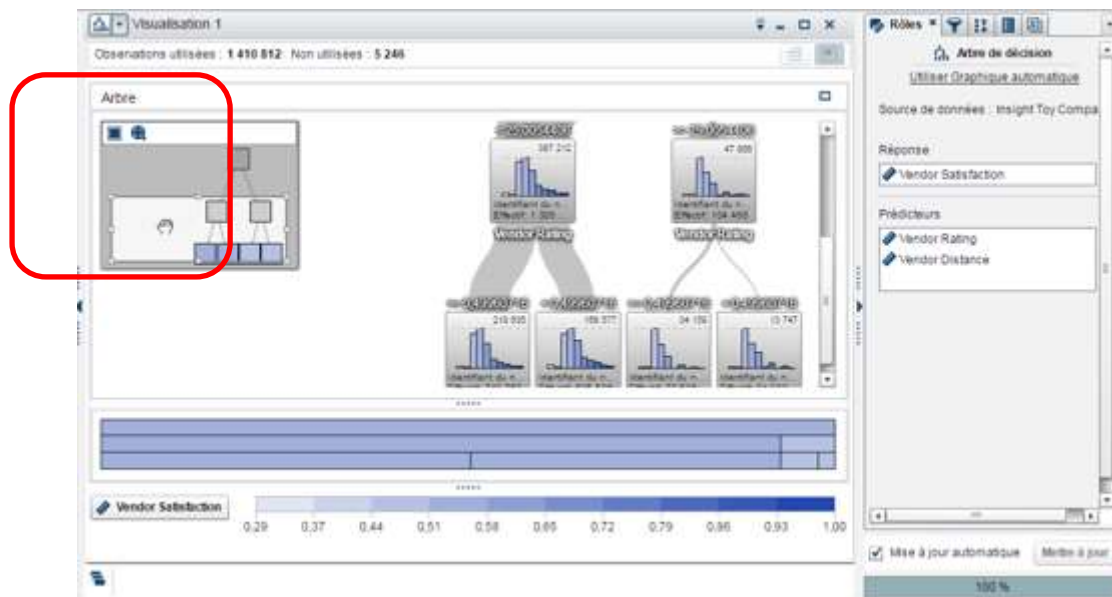
On sélectionne alors une 2e donnée prédictive « Vendor Distance». L'arbre de décision est alors mis à jour.



L'arbre de décision est alors trop grand, on peut alors faire un clic-droit sur le graphique et choisir « afficher la présentation ».



Une fenêtre apparaît en haut à gauche, elle permet de zoomer sur une partie de l'arbre :

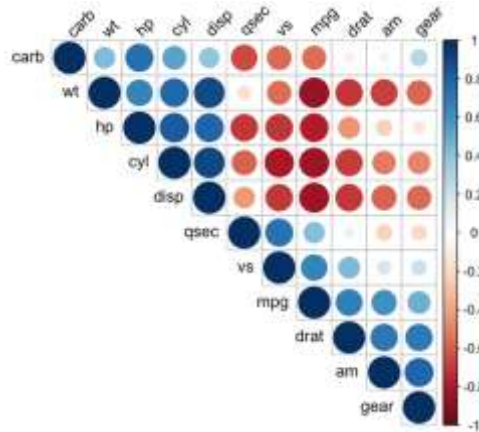




# Matrice de corrélation



**La matrice de corrélation :** En probabilités et en statistiques, étudier la corrélation entre deux ou plusieurs variables, c'est étudier l'intensité de la liaison qui peut exister entre ces variables.



Le fait que deux variables soient « fortement corrélées » ne démontre pas qu'il y ait une relation de causalité entre l'une et l'autre. Le contre-exemple le plus typique est celui où elles sont en fait liées par une causalité commune. Cette confusion est connue sous l'expression « Cum hoc ergo propter hoc ».

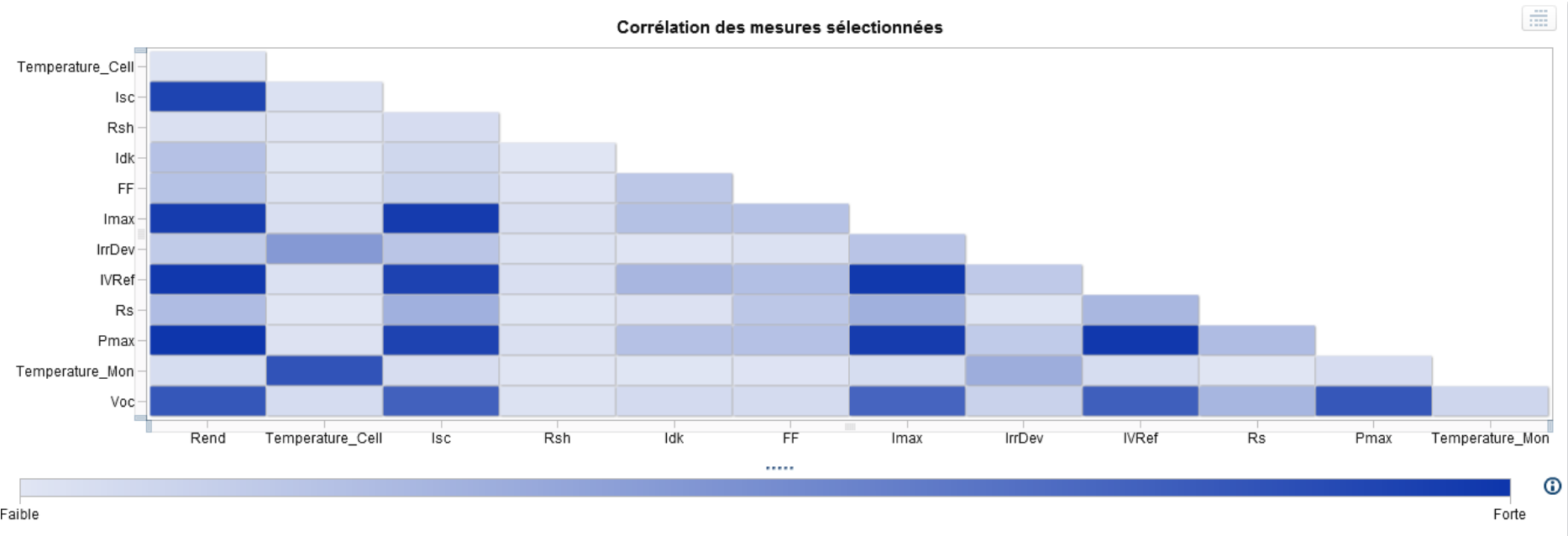
Une matrice de corrélation est utilisée pour évaluer la dépendance entre plusieurs variables en même temps. Le résultat est une table contenant les coefficients de corrélation entre chaque variable et les autres.

Une corrélation entre variables peut varier entre -1 et 1. Les valeurs intermédiaires renseignent sur le degré de dépendance linéaire entre les deux variables. Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation linéaire entre les variables est forte ; on emploie simplement l'expression « fortement corrélées » pour qualifier les deux variables. Une corrélation égale à 0 signifie que les variables ne sont pas corrélées linéairement.

Le coefficient de corrélation n'est pas sensible aux unités de chacune des variables. Ainsi, par exemple, le coefficient de corrélation linéaire entre l'âge et le poids d'un individu sera identique que l'âge soit mesuré en semaines, en mois ou en années.

En revanche, ce coefficient de corrélation est extrêmement sensible à la présence de valeurs aberrantes ou extrêmes (valeurs très éloignées de la majorité des autres, pouvant être considérées comme des exceptions).

**Exemple de matrice de corrélation sur les données de tri :**




**Corrélation entre le rendement et les différents types de courant et autres mesures de tri**

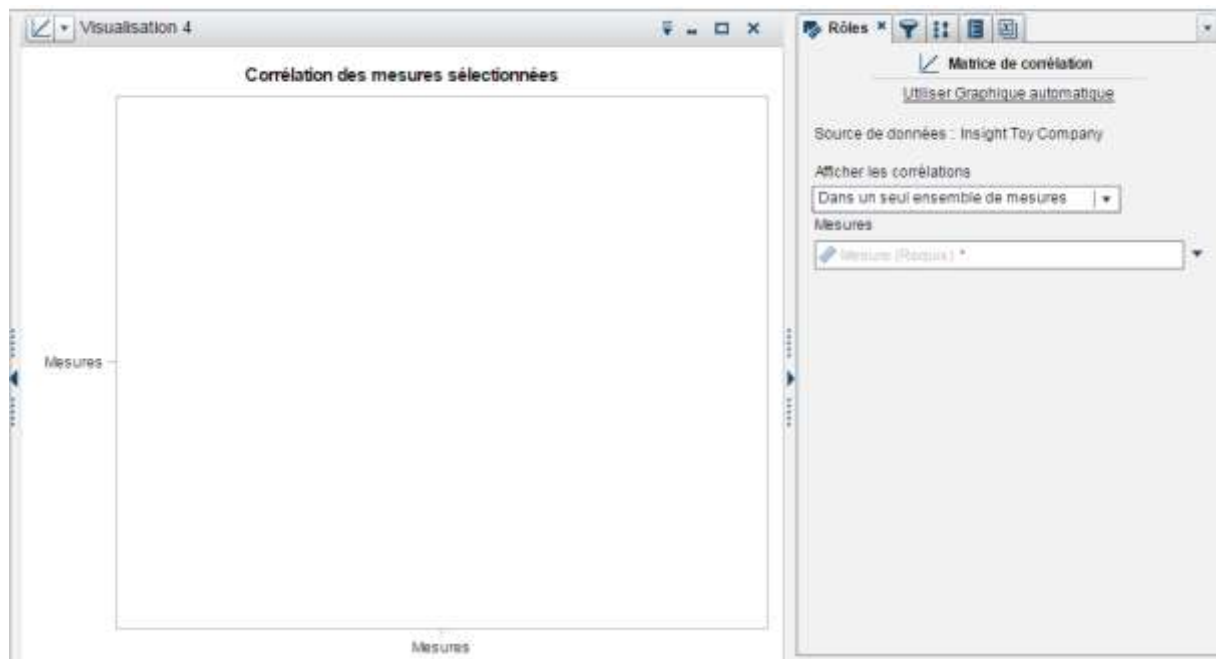
**Cette exploration** affiche les données sous forme de matrice de corrélation. Une matrice de corrélation affiche le degré de corrélation entre les mesures sous forme d'une série de rectangles de couleur. La couleur de chaque rectangle indique la force de la corrélation.



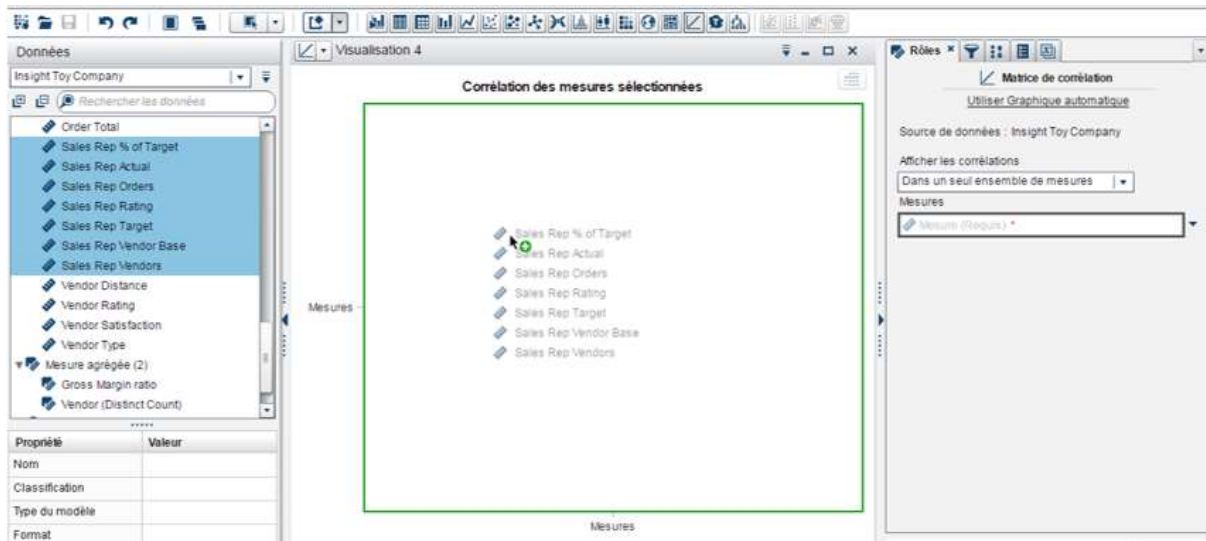
On souhaite connaître les corrélations entre toutes les mesures liées aux représentants commerciaux.

On double clic sur l'icône Matrice de corrélation  dans la barre à outils « Objets ».

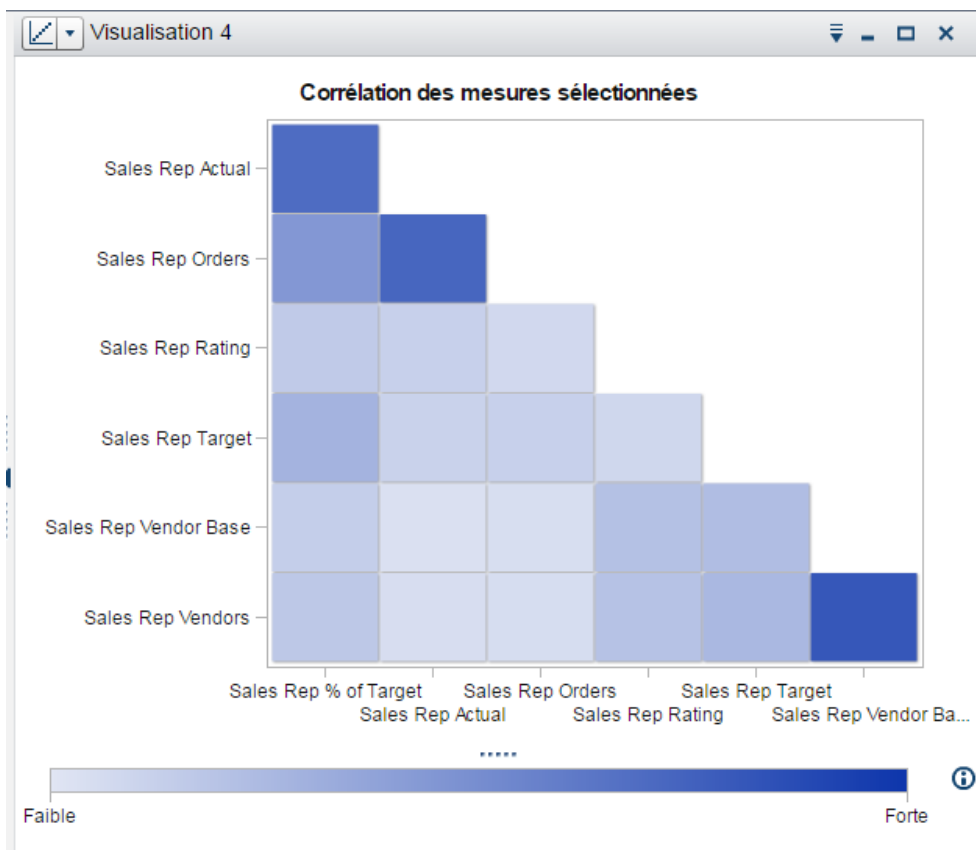
On obtient le graphique vierge suivant :



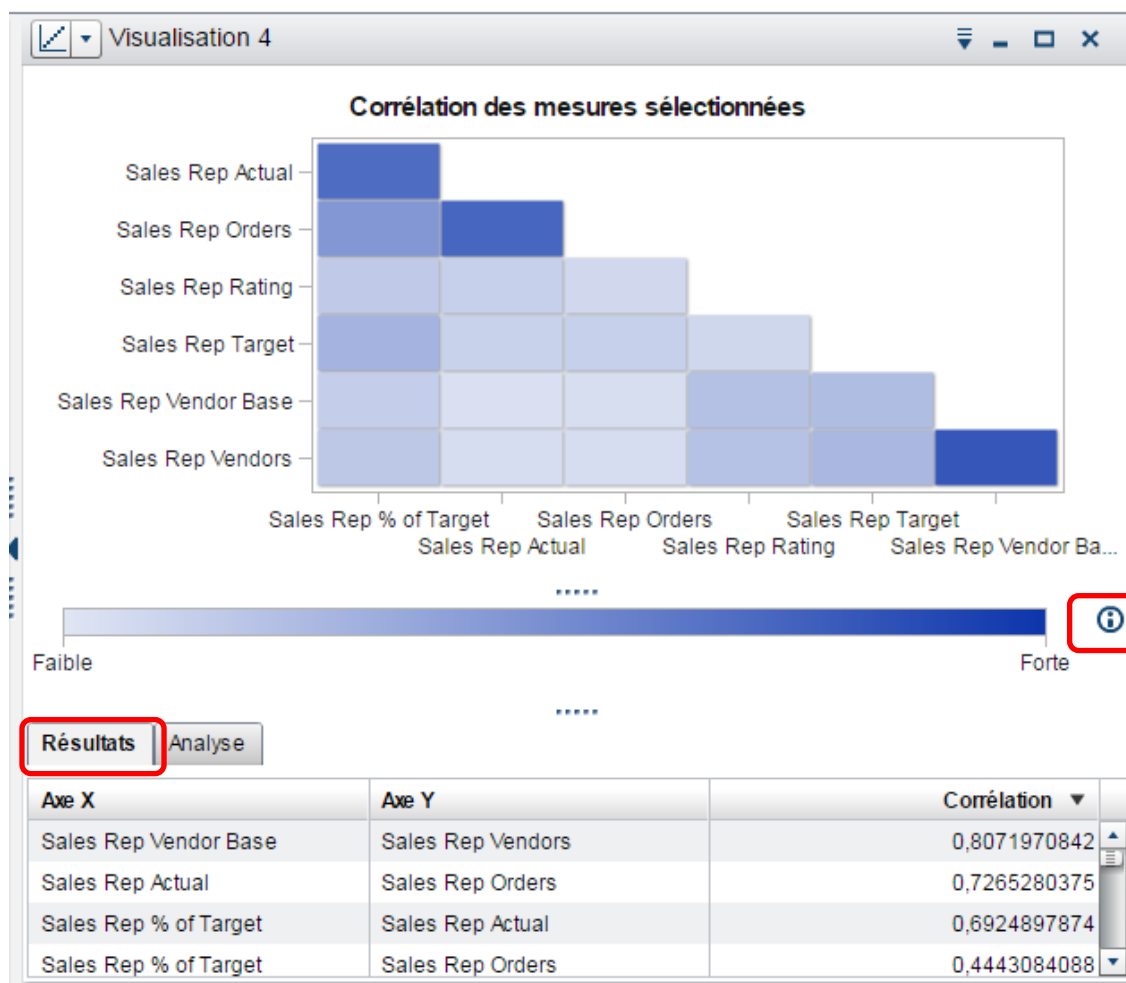
On va sélectionner toutes les mesures liées aux représentants commerciaux, elles commencent par « Sales Rep » et les glisser au milieu du graphique.



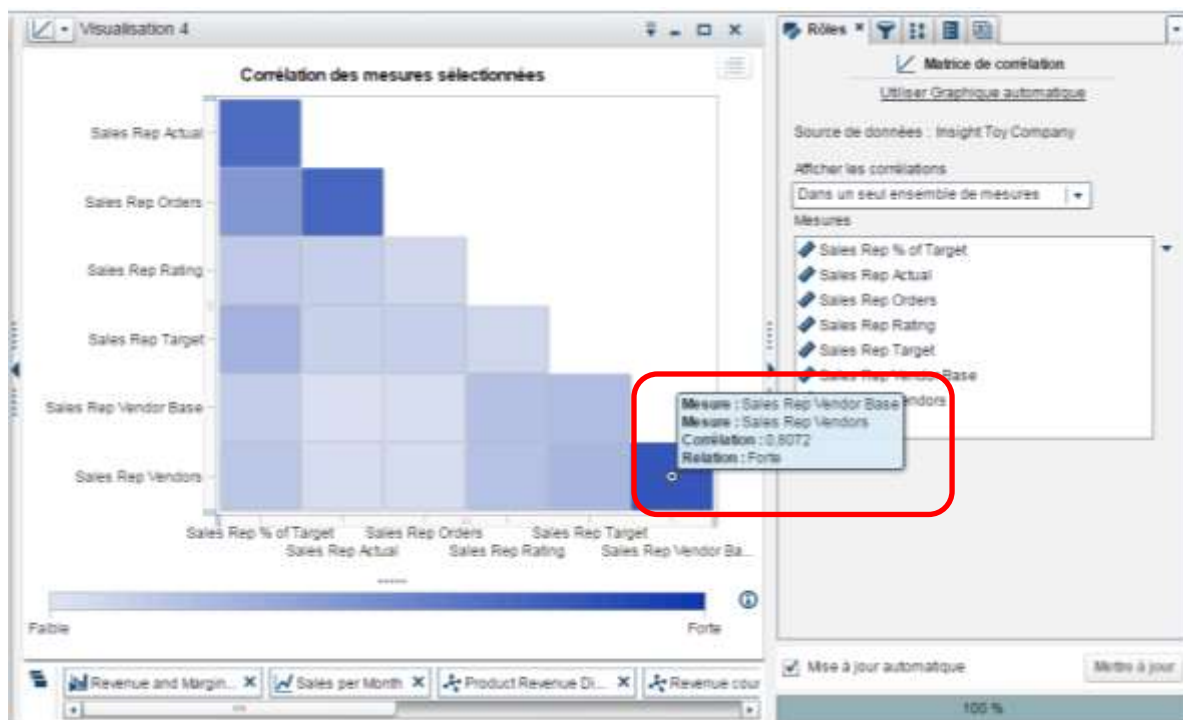
Le graphique se met à jour et met en valeur les corrélations plus ou moins fortes entre les mesures associées.



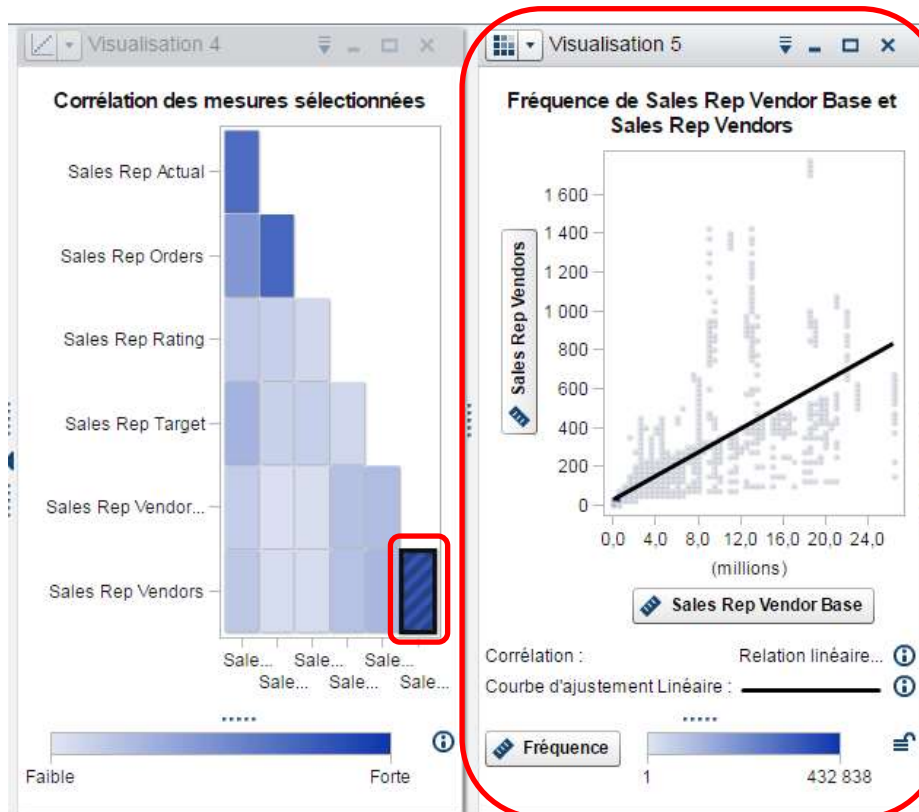
En cliquant sur l'icône « i » en bas à droite du graphique, puis sur l'onglet « Résultats », on obtient le résumé de toutes les valeurs de corrélation.



En déplaçant le curseur sur un des carrés de la matrice de corrélation, on obtient le résumé pour la corrélation entre les 2 mesures choisies.



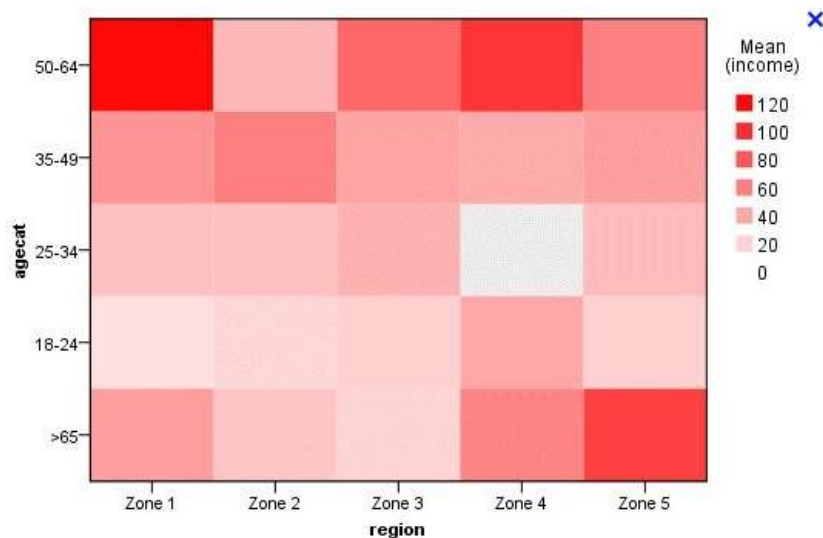
En double cliquant sur un des carrés de la matrice de corrélation, le graphique se met à jour et propose des informations plus fines sur cette corrélation dont la courbe d'ajustement linéaire.



## Carte thermique



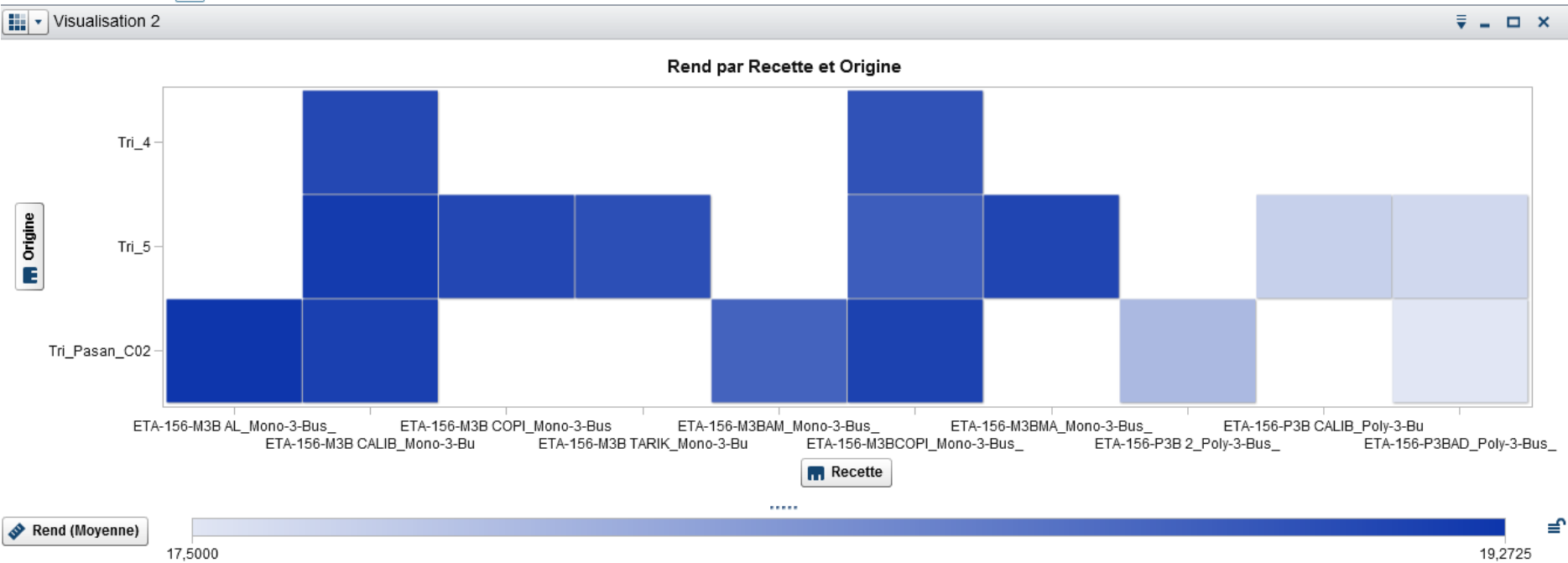
**La carte thermique :** Une heat map (carte thermique ou carte de chaleur) est une représentation graphique de données statistiques qui fait correspondre à l'intensité d'une grandeur variable une gamme de tons ou un nuancier de couleurs sur une matrice à deux dimensions (qui peut elle-même représenter une zone géographique). Ce procédé permet de donner à des données un aspect visuel plus facile à saisir qu'un tableau de chiffres.



Les heat maps servent notamment à montrer la fréquence du passage d'évènements (par exemple, des pannes ou des anomalies) ou d'événements (par exemple, des défauts ou des incidents de production) dans une zone donnée.

Une heat map peut par exemple servir à mettre en exergue les parties d'une ligne de production le plus souvent frappée par des arrêts ou des anomalies de fonctionnement.

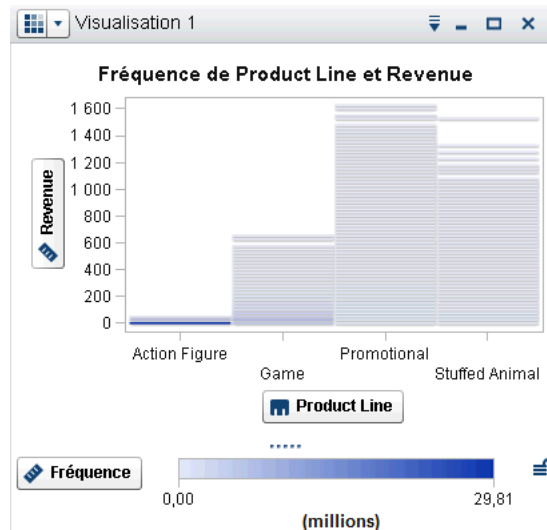
Exemple de carte thermique sur les données de tri :




Valeur du rendement en fonction de la recette et du poste de tri d'origine



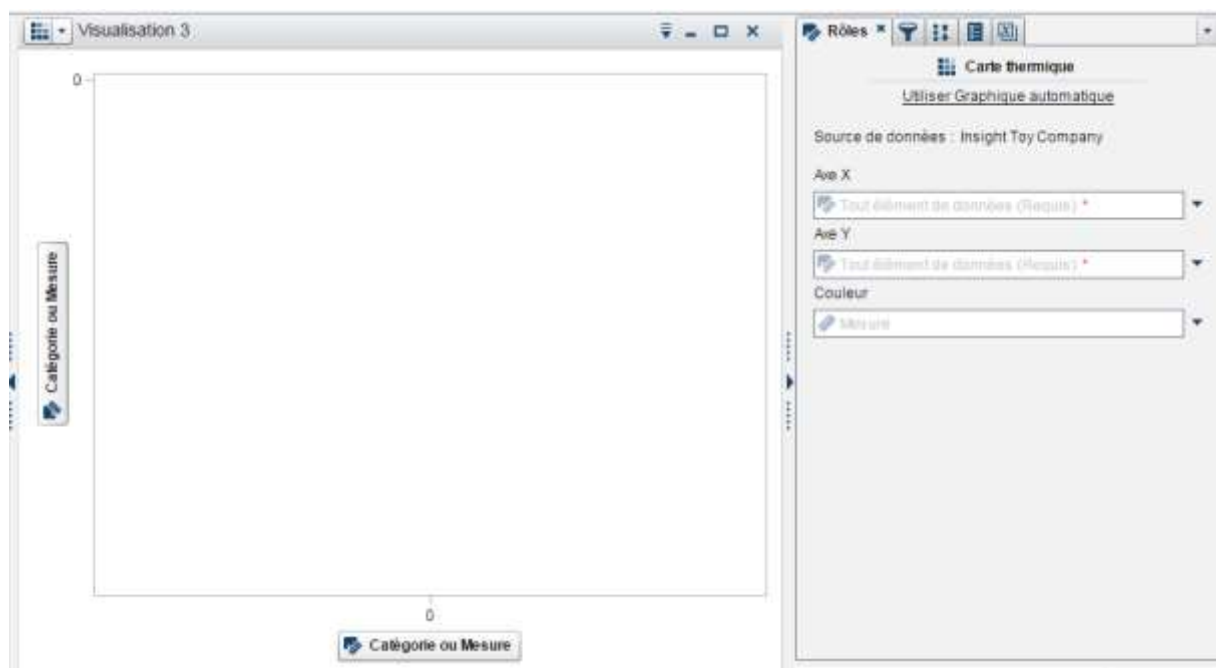
**Cette exploration** affiche les données sous forme de carte thermique. Une carte thermique affiche la distribution des valeurs pour deux éléments de données à l'aide d'un tableau avec des cellules de couleur. Si vous n'affectez pas de mesure au rôle de données Couleur, une couleur de cellule représente la fréquence de chaque intersection de valeurs. Si vous affectez une mesure au rôle de données Couleur, une couleur de cellule représente la valeur de mesure de chaque intersection de valeurs.



On souhaite connaître la distribution de la mesure de coût de distribution de commande (Order Distribution Cost) en fonction des continents.

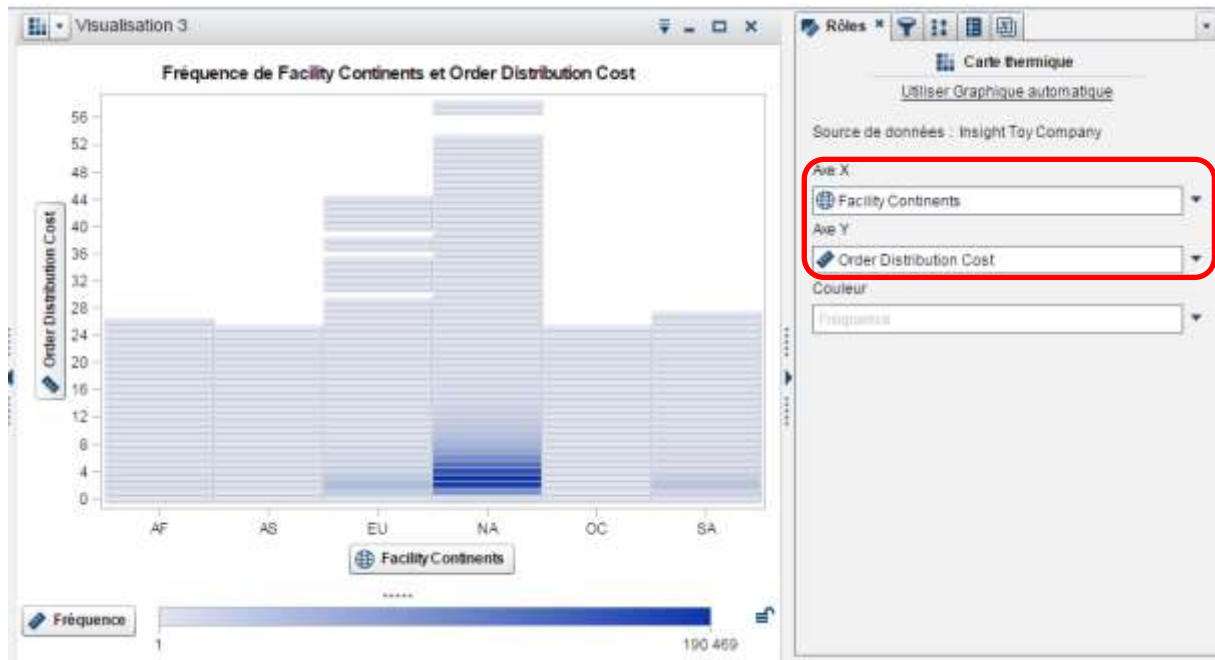
On double clic sur l'icône Carte thermique  dans la barre à outils « Objets ».

On obtient le graphique vierge suivant.



Dans l'onglet « Rôles », on choisit « Facility Continents » pour l'axe X et « Order Distribution Cost » pour l'axe Y, le graphique se met à jour.

On voit que la fréquence est la plus élevée pour le continent North America (NA) et pour les valeurs basses de coût de distribution de commande.



## Nuage de mots



**Le nuage de mots** est une représentation visuelle des mots-clefs (tags) les plus utilisés sur un site web. Généralement, les mots s'affichent dans des polices de caractères d'autant plus grandes qu'ils sont utilisés ou populaires.



Le « nuage de mots-clefs » est une sorte de condensé sémantique d'un document dans lequel les concepts clefs évoqués sont dotés d'une unité de taille (dans le sens du poids de la typographie utilisée) permettant de faire ressortir leur importance dans le site Web en cours ou dans les annuaires de sites utilisant ce même principe de fonctionnement. Il est possible de hiérarchiser ce système selon un ordre alphabétique, de popularité ou encore de représentation dans le site en cours.

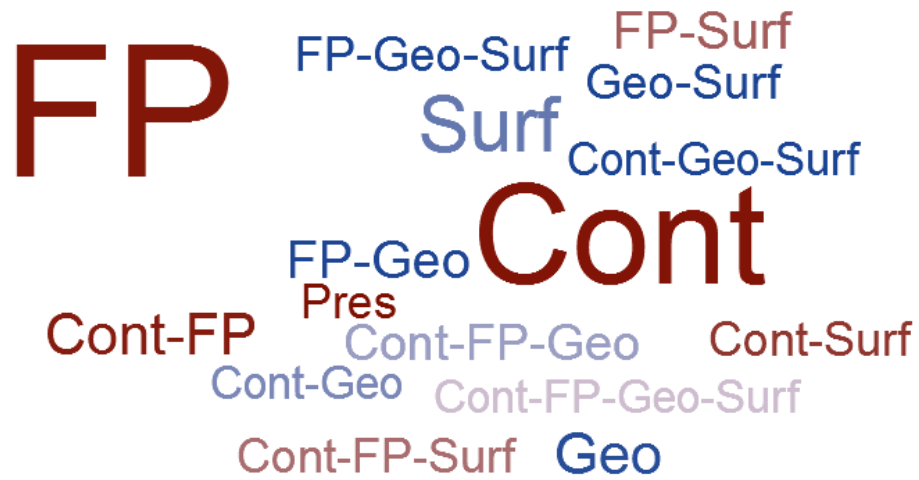
Il existe deux grandes familles de nuages de mots-clefs. C'est plus par leur valeur sémantique que par leur apparence que l'on distingue ces catégories.

La première famille de nuage de mots-clefs classe les concepts selon le critère de la répétition d'un mot dans un article. Il s'agit donc d'une méta-donnée permettant de symboliser par ordre d'importance les concepts que recouvre l'article en cours.

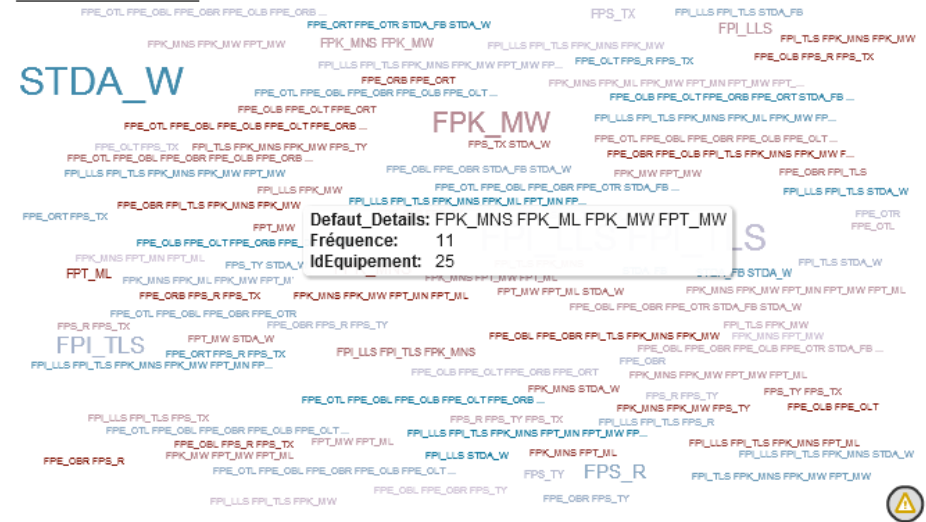
La seconde, plus transversale, regroupe en nuage de mots-clefs les mots-clefs revenant le plus souvent dans un site ou dans un annuaire de sites. Il s'agit donc là de mettre en avant la popularité d'un concept, qui a fédéré plusieurs rattachements dans un site ou un ensemble de sites. Cela est particulièrement utile à une navigation transversale, permettant de balayer l'intégralité du contenu d'un site à travers le fil conducteur du mot-clef auquel on s'intéresse.

### Exemple de nuage de mots sur les données ICOS N1 :

### Famille de défauts principaux



### Détails des défauts

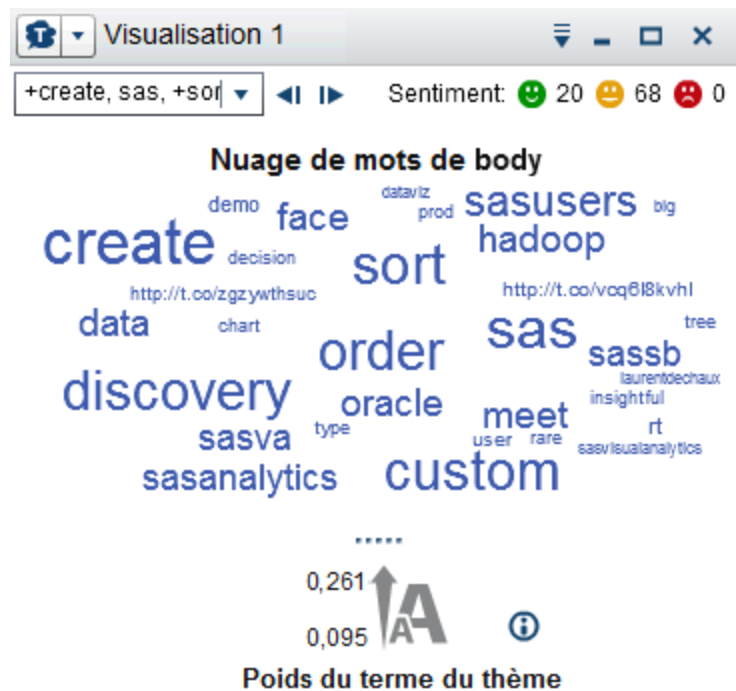


**Type de défaut remonté sur la supervision de la station sérigraphie N1**

**Cette exploration** affiche un ensemble de mots à partir d'un élément de données alphanumérique. Selon le type de mot-clé et vos rôles de données, la taille de chaque mot dans le nuage peut indiquer la pertinence du mot par rapport à une rubrique, la fréquence du mot dans une catégorie ou la valeur d'une mesure.

Vous pouvez utiliser l'analyse de texte dans un nuage de mots pour identifier les thèmes et les termes qui apparaissent ensemble dans vos données et analyser le sentiment qui ressort des documents dans un thème donné.

**NB** : la fonctionnalité « Nuage des mots » est également disponible dans la partie « Rapport ».



On veut savoir quelles villes ressortent par rapport à la mesure de coût de distribution de commande (Order Distribution Cost).

On double clic sur l'icône Nuage de mots  dans la barre à outils « Objets ».

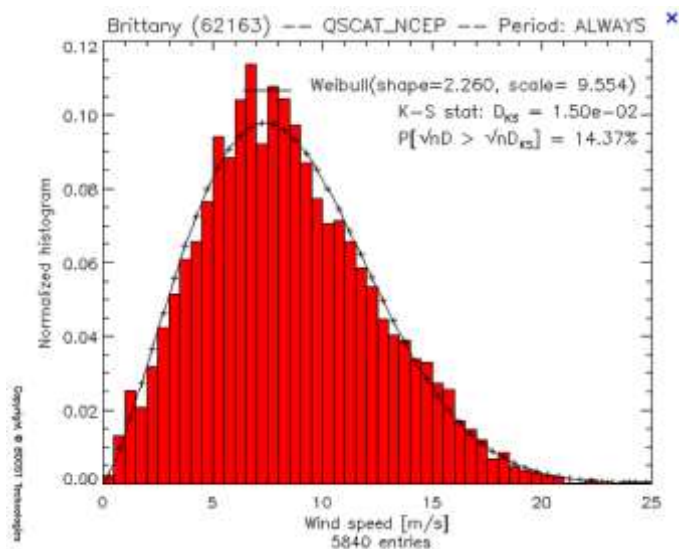
Plus la taille des caractères des mots est grande, plus le coût de distribution de commande est important.



# Distribution



**La distribution** : En statistique, la distribution, distribution empirique ou distribution des fréquences, est un tableau qui associe des classes de valeurs obtenues lors d'une expérience à leurs fréquences d'apparition. Ce tableau de valeurs est modélisé en théorie des probabilités par une loi de probabilité.



Dans le cas général, les classes sont des intervalles de valeurs. Dans le cas de valeurs discrètes, une classe peut ne regrouper qu'une seule valeur. Pour que les calculs statistiques aient un sens, il faut que l'effectif de chaque classe soit suffisant.

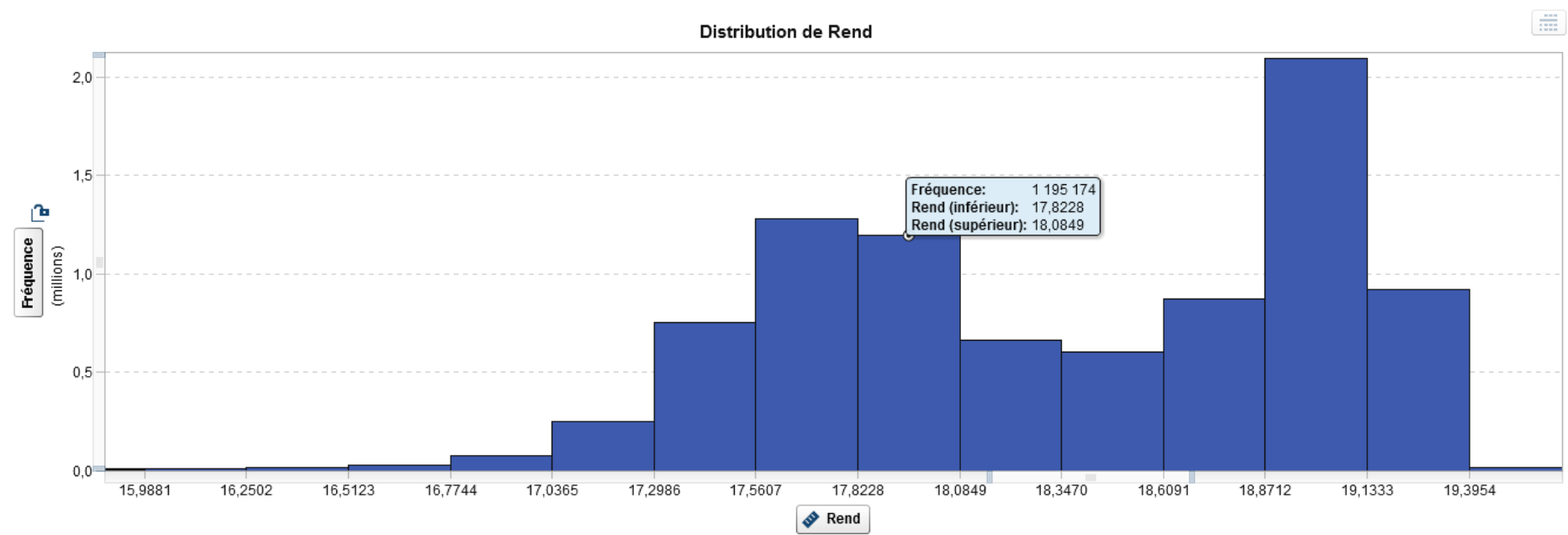
Une modélisation probabiliste est associée : chaque résultat est modélisé par une variable aléatoire. La distribution statistique est alors modélisée par une loi de probabilité. Plus précisément, la fréquence observée ou empirique d'appartenance à la classe A peut être modélisée par une valeur théorique qui est la probabilité de l'évènement.

La modélisation se justifie par le fait que plus l'échantillon est grand plus la distribution statistique est proche (au sens des lois de probabilités) de la loi de probabilité.

La loi des grands nombres assure que :

La fréquence observée converge vers une valeur  $P(A)$  qui est la probabilité qu'un individu pris au hasard appartienne à la classe A.

**Exemple de distribution sur les données de tri :**

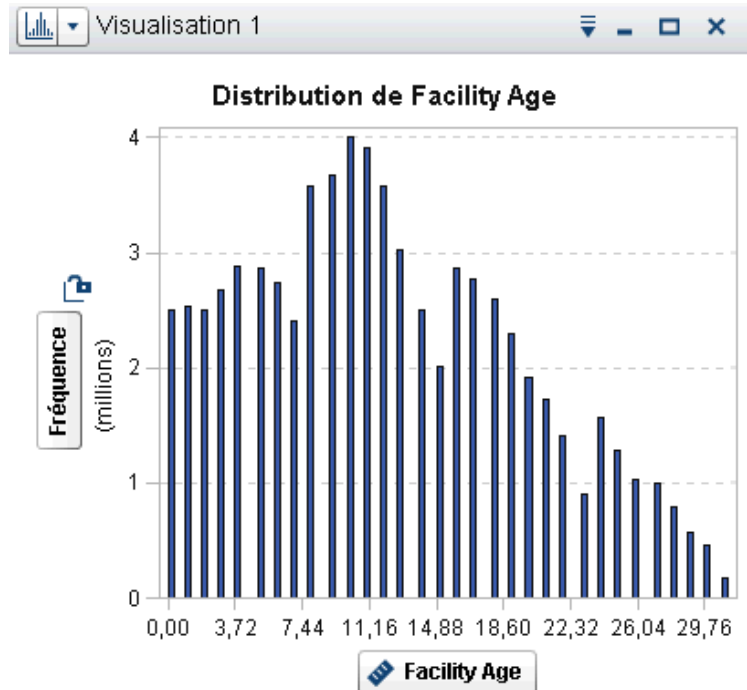


Distribution en classes de rendement des cellules sur une année de mesure aux différents poste de tri



**Cette exploration** affiche les données sous forme de diagramme en bâtons. Un diagramme en bâtons affiche la distribution des valeurs pour une seule mesure.

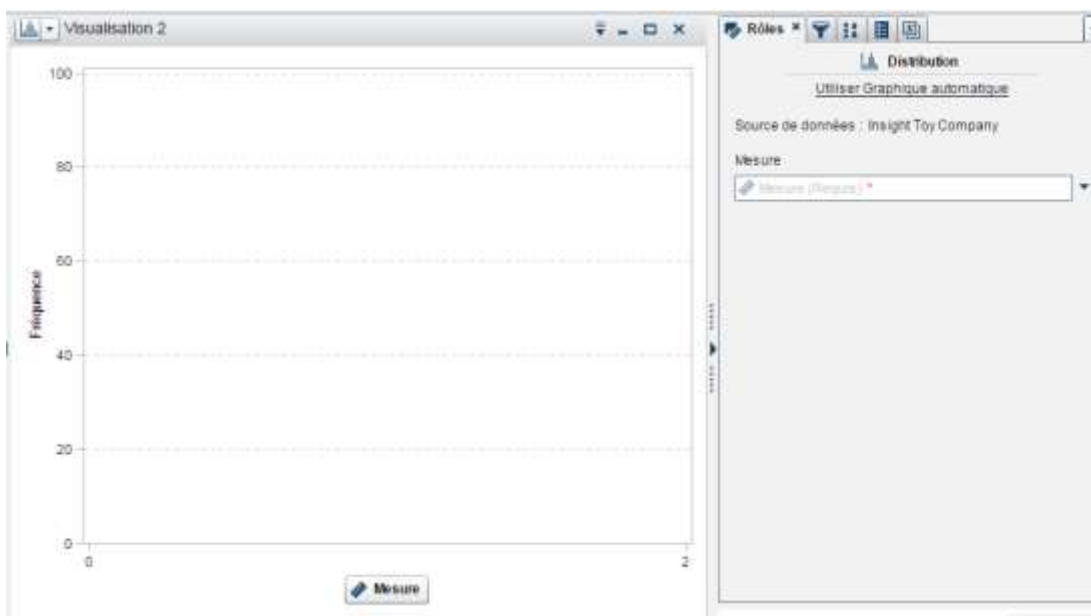
Vous pouvez sélectionner l'orientation des barres et indiquer si les valeurs sont affichées sous forme de pourcentage ou d'effectif



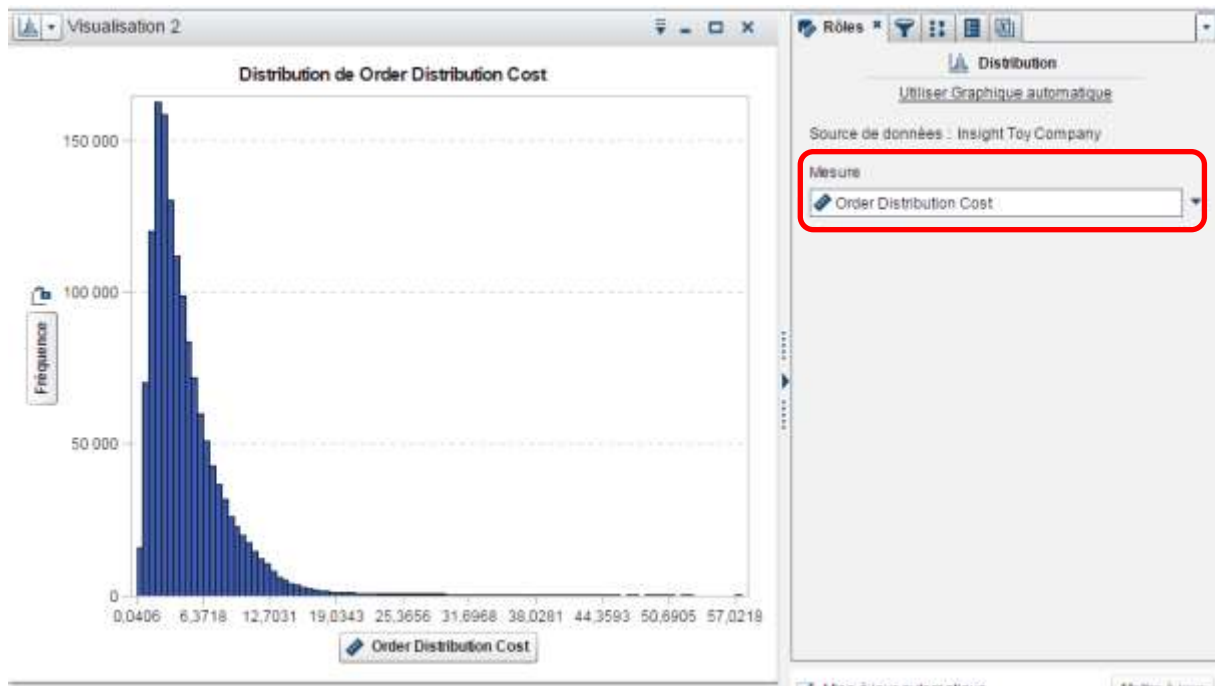
On souhaite connaître la distribution du coût de distribution de commande (Order Distribution Cost).

On double clic sur l'icône Distribution  dans la barre à outils « Objets ».

On obtient le graphique vierge suivant.

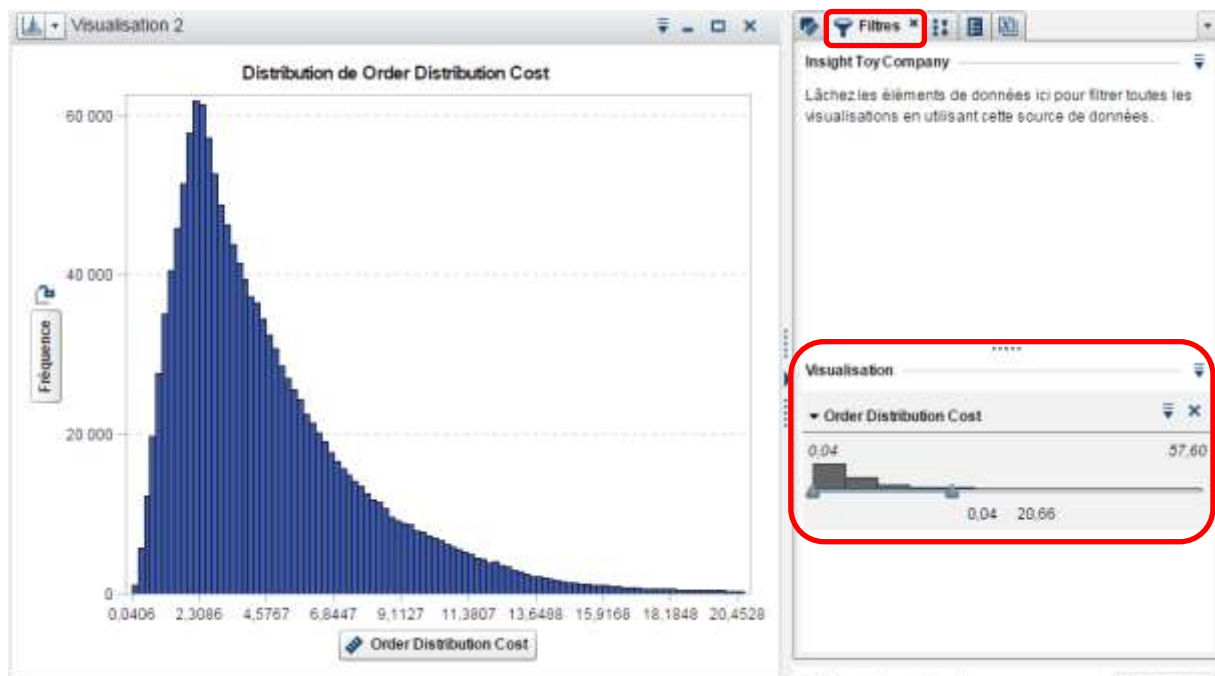


Dans l'onglet « Rôles », on choisit la mesure « Order Distribution Cost » pour la section « Mesure », le graphique se met à jour :



On souhaite filtrer la distribution pour ne pas visualiser les valeurs extrêmes.

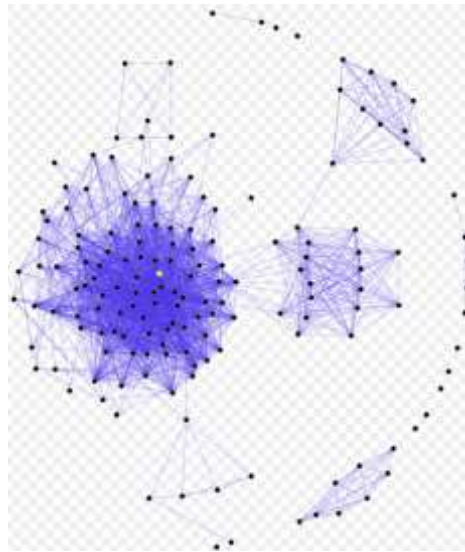
On clique sur l'onglet « Filtre », puis on fait glisser la mesure « Order Distribution Cost » vers la section « Visualisation ». On peut alors déplacer les curseurs pour choisir la fourchette des mesures.



# Diagramme de réseau



**Le diagramme de réseau** ou diagramme de liens. Un lien, ou entrelacs, en topologie est un ensemble de courbes de l'espace, fermées et d'intersection vide deux à deux. La théorie des nœuds étudie les configurations possibles de ces objets (un nœud est un lien ne possédant qu'une seule courbe). Un diagramme de lien est une représentation bidimensionnelle d'un lien.

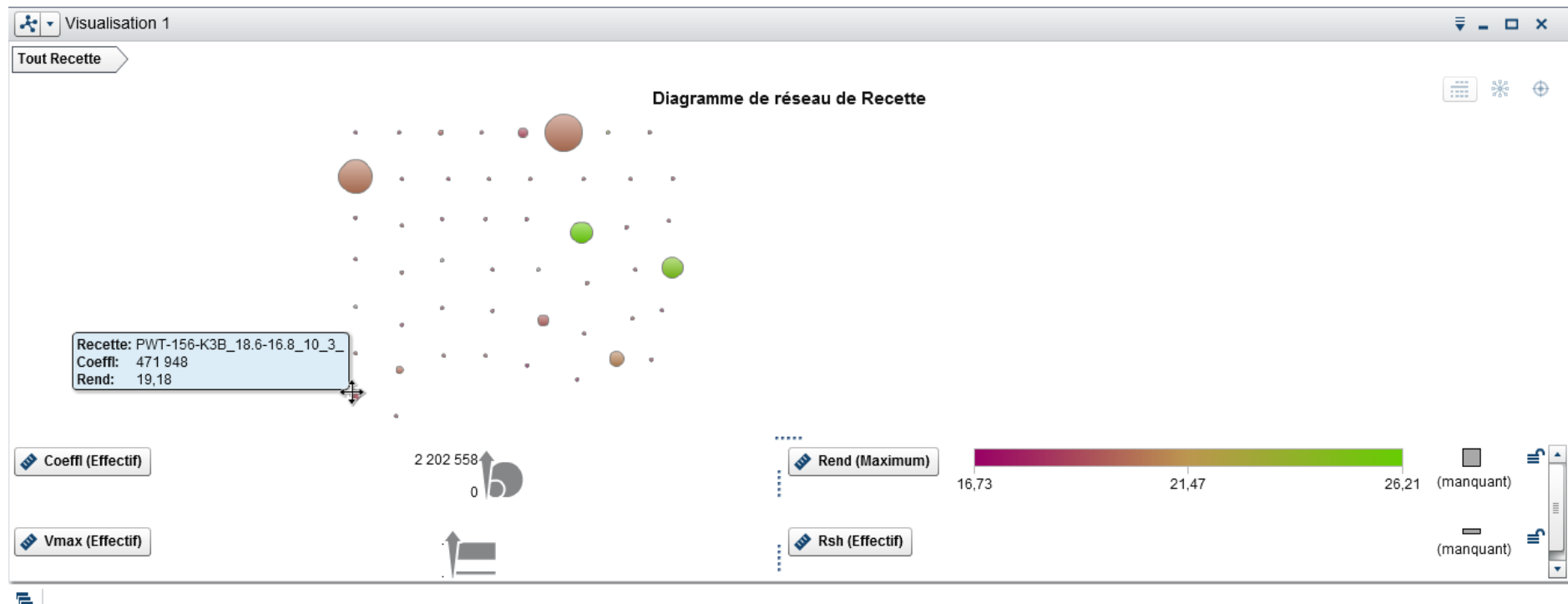


On peut avoir différentes projections du même objet, par exemple en changeant de direction d'observation. Cependant il existe un moyen de relier des diagrammes entre eux.

Si on veut relier deux diagrammes qui sont la projection d'un même lien ou de ses déformations sans couper aucune courbe, on utilise les relations de Reidemeister. Ces relations sont une conséquence des règles de projection. Elles relient entre elles les différentes situations que l'on obtient en le résolvant (en changeant légèrement de direction de projection).

Plus globalement, un diagramme de réseau permet de représenter l'intensité de la relation entre différents groupes ou ensembles de données.

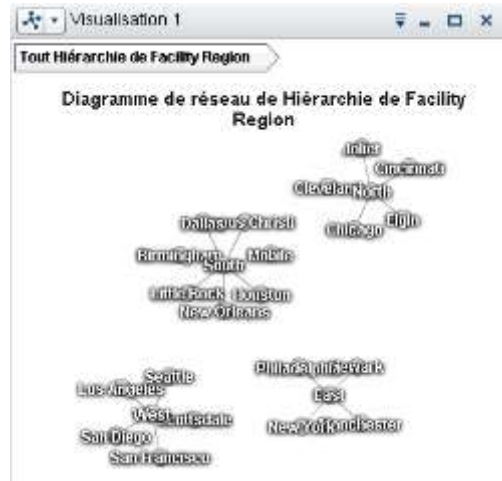
## Exemple de diagramme de réseau sur les données de tri :




Exemple de diagramme de réseau sur les différentes recettes mesurées sur les équipements de tri

**Cette exploration** affiche une série de noeuds liés. Un diagramme de réseau affiche les relations entre les valeurs de catégories ou les niveaux de hiérarchie.

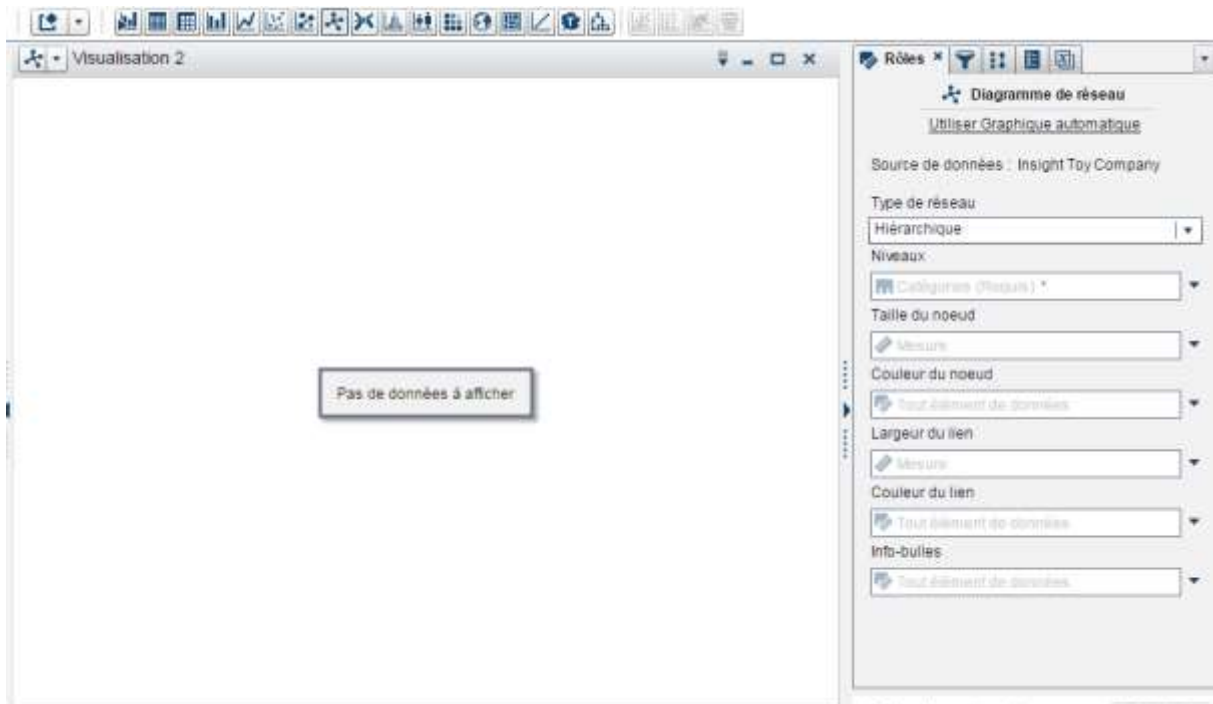
Vous pouvez indiquer les valeurs des mesures via les dimensions et les couleurs des noeuds et les liens des noeuds.



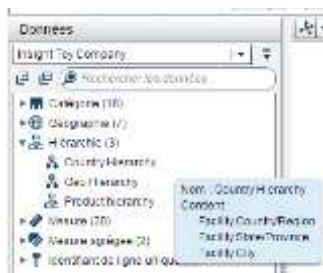
On voudrait exprimer la répartition entre les zones géographiques en terme de coût de distribution de commande (Order Distribution Cost). On veut ensuite projeter ce diagramme sur une carte.

On double clic sur l'icône Diagramme de réseau  dans la barre à outils « Objets ».

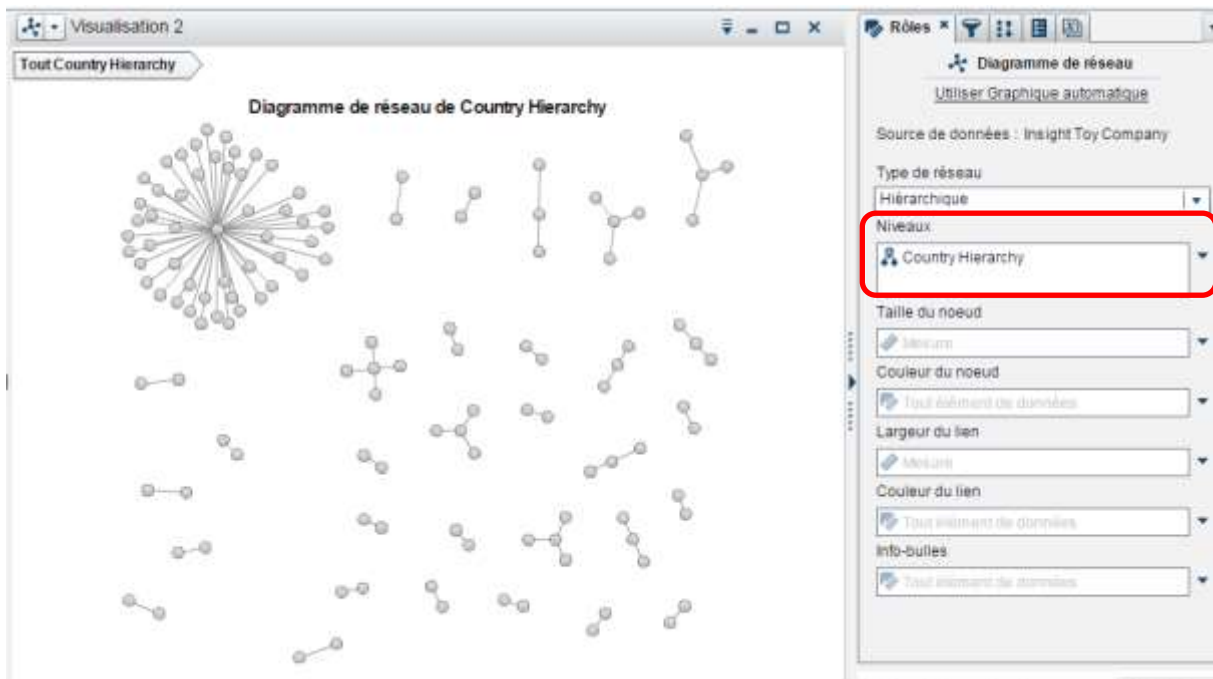
On obtient le graphique vierge :



Pour le rôle « Niveaux », on dispose d'une hiérarchie « Country Hierarchy ».

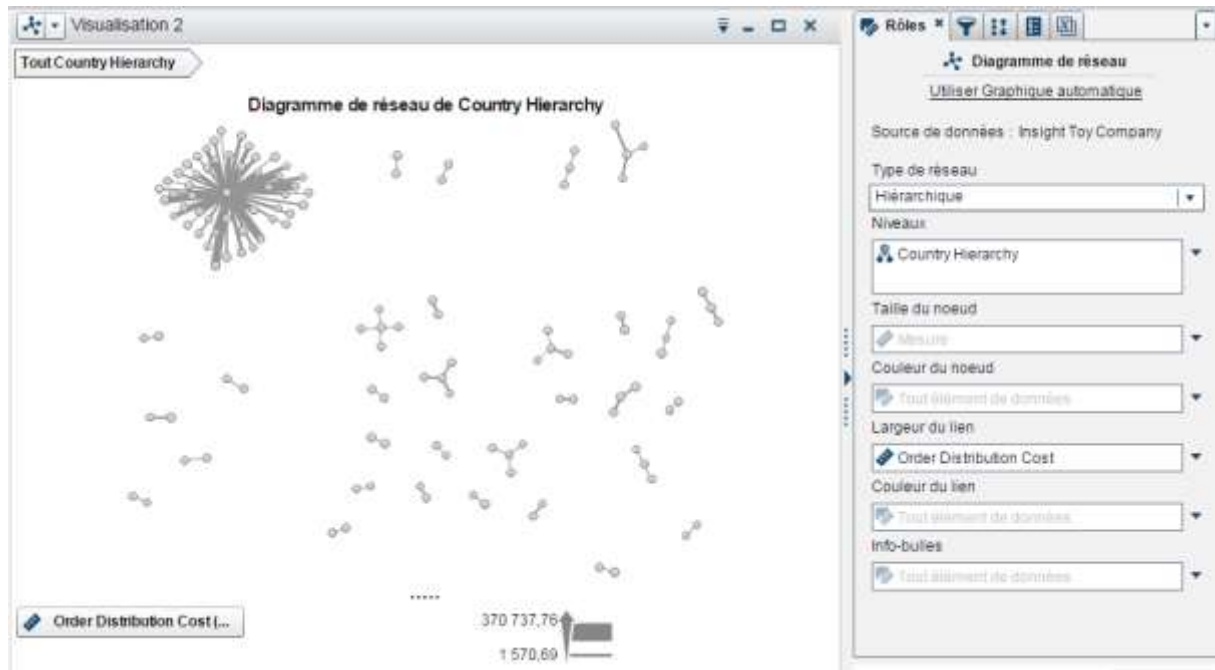


On déplace « Country Hierarchy » vers le rôle « Niveaux », le graphique se met à jour.

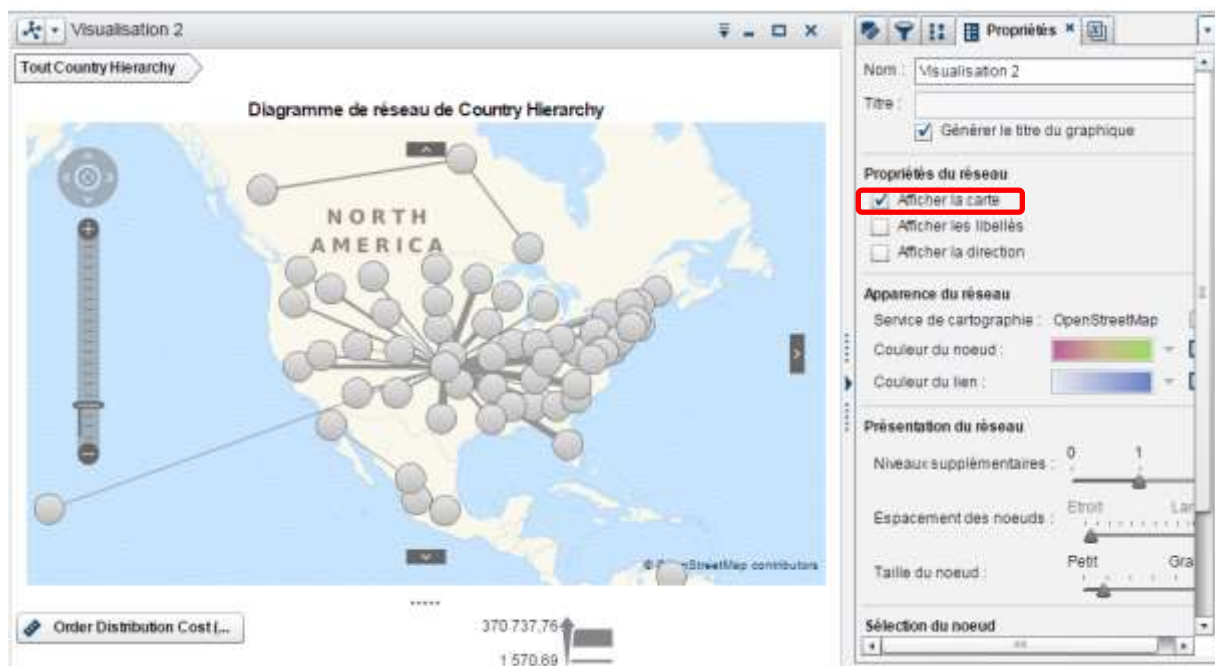


Il est possible de paramétrer la taille et la couleur du nœud, la largeur et la couleur du lien ainsi que les infos-bulles.

On choisit alors comme largeur du lien « Order Distribution Cost », on voit alors que certains liens sont plus épais.



En allant sur l'onglet « Propriétés », on peut cocher l'option « Afficher la carte » dans la catégorie « Propriétés du réseau », le diagramme de réseau est alors projeté sur une carte géographique.





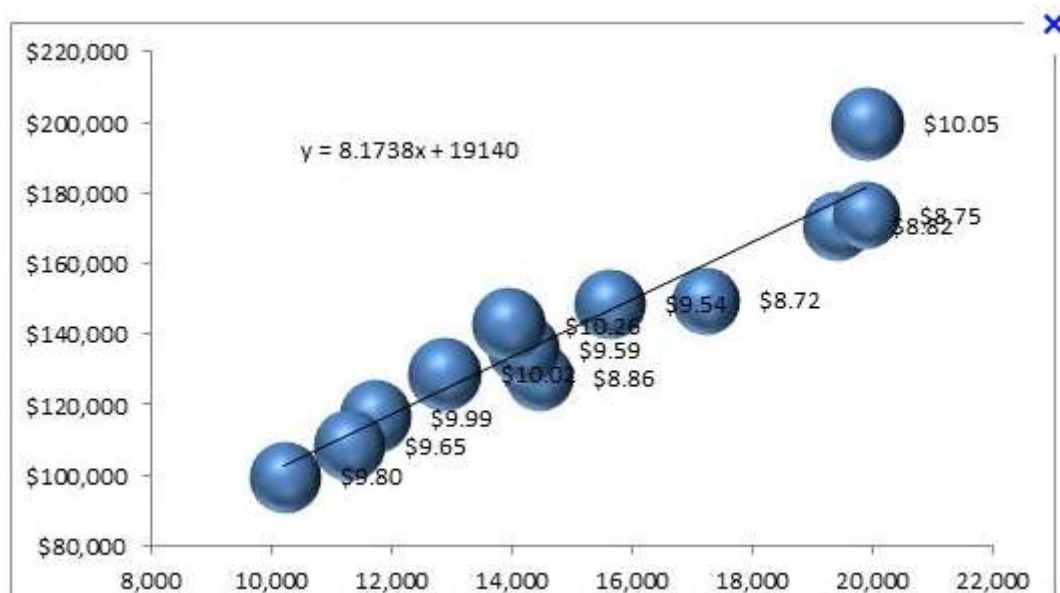


## Graphique à bulles



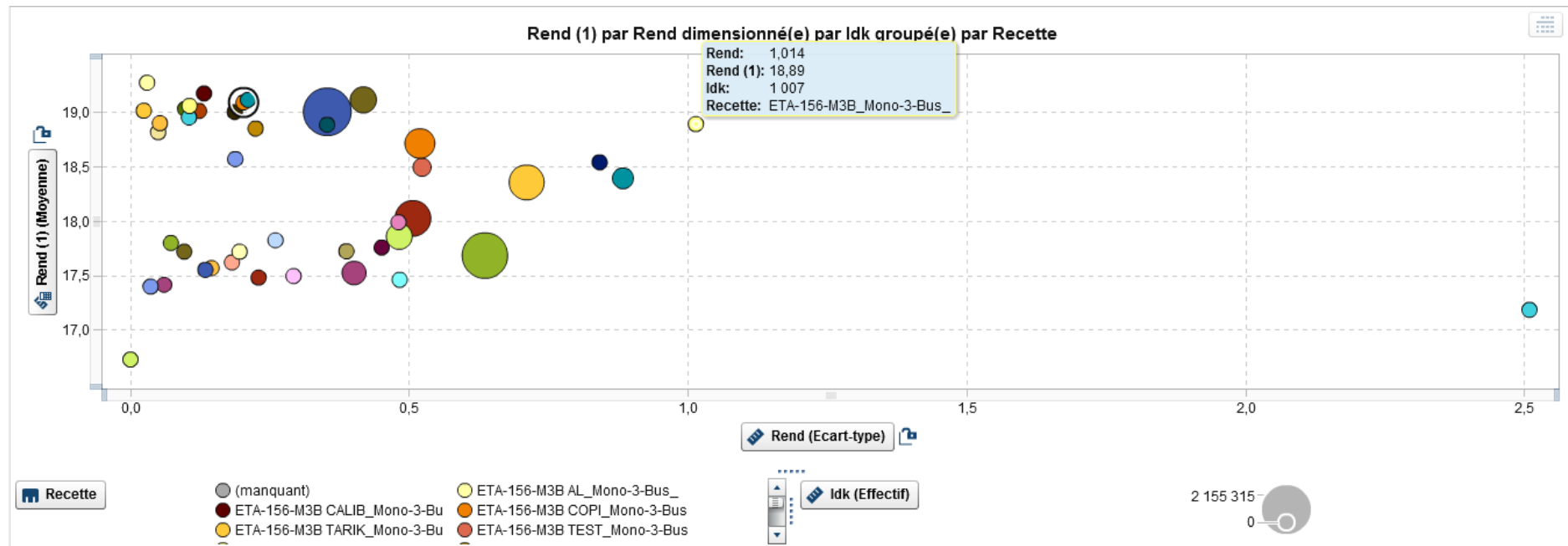
Le graphique à bulles permet de représenter différentes catégories en positionnant des objets de type « bulle » sur un plan cartésien en fonction de mesures numériques en abscisse et en ordonnée, et en dimensionnant la taille de la bulle en fonction d'un paramètre déterminé.

Les graphiques en nuage de points et les graphiques à bulles sont fortement similaires, car ils utilisent tous deux un tracé à coordonnées (X,Y) pour visualiser le contenu des données. Cependant, ils se différencient par le style de marqueurs en bulles utilisés pour les points de données individuels.



Cette visualisation permet notamment de mettre en évidence certains liens et certaines corrélations entre différents facteurs en observant la répartition géographique et la taille des différentes bulles sur l'espace représenté.

### Exemple de graphique à bulles sur les données de tri :



Chaque bulle représente une recette positionnée en fonction :

- de l'écart type de son rendement en abscisse
- de la moyenne de son rendement en ordonnée,
- du nombre de cellules mesurées au tri pour la taille de la bulle,

