

**MASTER MIAGE PARCOURS INFORMATIQUE ET INNOVATION.**

**UE: DATA MINING et BIG DATA**

**ENSEIGNANTS: Pr Antoine TABONNE.**

**Pr Olivier PERRIN**

***Rédigé par Alain NGUIDJOI BELL***

Rapport DU PROJET  
 DE BIG DATA

Sommaire

[Introduction 1](#_Toc26106360)

[1 Première partie: BIG DATA, K-MEANS 1](#_Toc26106361)

[1.1 Clustering par k-means avec k=3 1](#_Toc26106362)

[1.2 Normalisation des données. 1](#_Toc26106363)

[1.3 Détermination des clusters Par l’algorithme de K-Means 1](#_Toc26106364)

[2 Deuxième partie : Business Intelligence, TP POWERPLAY (Version 7) 4](#_Toc26106365)

[Conclusion 10](#_Toc26106366)

# Introduction

Ce document résume les étapes de mise en œuvre du projet de Big Data. Ce projet consiste à modéliser le taux de désabonnement ou churn grâce à deux algorithmes d’apprentissages que sont la régression logistique et le Random Forest, puis de comparer les résultats obtenus. Pour réaliser ce travail, nous avons tout d’abord effectué un bref aperçu des données grâce aux logiciel ANACONDA, ensuite grâce aux bibliothèques seaborn et panda nous avons effectués une analyse exploratoire des données, des transformations des certaines donnée ont ensuite été effectué, puis nous avons construit le pipeline de nos model grâce à Spark après entrainement de ces dernières nous avons terminé par une analyse comparative des performance de nos deux models.

# APERCUS DES DONNEES

En chargeant le fichier de donnée avec l’application Orange 3 du Logiciel ANACONDA, on obtient le tableau récapitulatif suivant :

|  |
| --- |
|  |
|  |
|  |

Ce tableau nous présente les types de donnée et les valeurs possibles de chaque features, en prenant en compte que notre label est représenté par la colonne **Churn** et que le champs **customerID** ne représente que les utilisateurs, On peut conclure que notre ensemble de donnée comporte :

19 Features dont trois numériques et 16 catégoriques, et un Label de type catégorique.

Si le récapitulatif nous indique bien qu’il n’y’a pas de valeur manquante ce qui est bien confortable pour les valeurs catégoriques ou les valeurs possibles sont déterminées, on ne connait pas les valeurs possibles des features de type numérique qui peuvent avoir des valeurs nulles. Nous allons donc vérifier l’existence de valeur nulle pour les trois features numériques.

|  |
| --- |
|  |

Nous allons dérouler l’algorithme des **K-Means** sur les données présentées dans le tableau précèdent et regrouper les valeurs de la conne **NUM** en trois groupes.

## Normalisation des données.

Nous allons préalablement normaliser les données des colonnes **AGE, SEXE, SALAIRE**, en appliquant la fonction :   sur chaque élément de la colonne **X** ꞓ **{ AGE, SEXE, SALAIRE }.** En appliquantcette normalisation nous obtenons le tableau normalisé suivant :

|  |  |  |  |
| --- | --- | --- | --- |
| **NUM** | **AGE** | **SEXE** | **SALAIRE** |
| **A** | 0,151515152 | 0 | 0,044444 |
| **B** | 0,878787879 | 1 | 0,566667 |
| **C** | 0,909090909 | 1 | 1 |
| **D** | 0,333333333 | 0 | 0,444444 |
| **E** | 0,696969697 | 1 | 0,333333 |
| **F** | 0,393939394 | 0 | 0,5 |
| **G** | 0 | 1 | 0 |
| **H** | 1 | 1 | 0,166667 |
| **I** | 0,606060606 | 0 | 0,322222 |

Nous allons donc prendre en input de l’algorithme de **K-Means** les données normalisées du tableau précèdent, en utilisant la distance d(X,Y)= n, étant le nombre de colonne

## Détermination des clusters Par l’algorithme de K-Means

Apres avoir normaliser les données nous allons dérouler l’algorithme des K-Means avec ici K=3,pour déterminer trois groupes ou clusters de l’ensemble des points ou vecteurs A, B, C ,D, E, F, G, H , issue de ces données normalisées.

1. **Première Itération**

* Choix des groupes de base :

Nous choisissons trois groupes.

**Groupe 1 = {A}, Groupe 2 = {D}, Groupe 3={F}**

* Déterminations des appartenances des point B, C, E, G, H, I au groupe choisi.

Nous allons calculer les distances de chaque point des groupes choisies par rapport au points B,C,E,G,H,I n’appartenant pas à ces groupes. On obtient le tableau suivant :

|  |  |  |  |
| --- | --- | --- | --- |
| **NUM** | **A** | **D** | **F** |
| **B** | 2,249495 | 1,667676768 | **1,551515** |
| **C** | 2,713131 | 2,131313131 | **2,015152** |
| **E** | 1,834343 | 1,474747475 | **1,469697** |
| **G** | **1,19596** | 1,777777778 | 1,893939 |
| **H** | 1,970707 | 1,944444444 | **1,939394** |
| **I** | 0,732323 | 0,394949495 | **0,389899** |

On en déduit les groupes suivants en prenant la distance minimale :

**Groupe 1 = {A, G}, Groupe 2 = {D}, Groupe 3 = {B, C, E, F, H, I}**

Ces groupes sont différents des groupes choisis, nous allons donc faire une nouvelle itération.

1. **Deuxième Itération**

Groupe initiale : **Groupe 1 = {A, G}, Groupe 2 = {D}, Groupe 3={B,C,E,F,H,I}**

* Calcul des points Moyen de chaque groupe :

On obtient le tableau suivant comportant les point moyenne de chaque groupe

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Groupe 1 = {A, G}** | **Groupe 2 = {D}** | **Groupe 3 = {B, C, E, F, H, I}** |
|  | **Moyenne Groupe 1** | **Moyenne Groupe 2** | **Moyenne groupe 3** |
| **AGE** | 0,07575758 | 0,3333333 | 0,747475 |
| **SEXE** | 0,5 | 0 | 0,666667 |
| **SALAIRE** | 0,02222222 | 0,4444444 | 0,481481 |

* Calcul de la distance de chaque point par rapport au point Moyens et déterminations des groupes de chaque point.

Nous obtenons les distances contenues dans le tableau suivant :

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Groupe 1 = {A, G}** | **Groupe 2 = {D}** | **Groupe 3 = {B, C, E, F, H, I}** |
| **Moyenne Groupe 1** | **Moyenne Groupe 2** | **Moyenne Groupe 3** |
| A | 0,5979798 | **0,5818182** | 1,699663 |
| B | 1,84747475 | 1,6676768 | **0,549832** |
| C | 2,31111111 | 2,1313131 | **1,013468** |
| D | 1,17979798 | **0** | 1,117845 |
| E | 1,43232323 | 1,4747475 | **0,531987** |
| F | 1,2959596 | **0,1161616** | 1,038721 |
| G | **0,5979798** | 1,7777778 | 1,56229 |
| H | 1,56868687 | 1,9444444 | **0,900673** |
| I | 1,33030303 | **0,3949495** | 0,96734 |

En prenant la distance minimale de chaque point par rapport au trois moyenne de chacun des trois Groupes, on obtient à cette deuxième itérations les groupes suivants :

**Groupe 1= {G}, Groupe 2 = {A, D, F, I}, Groupe 3 = {B, C, E, H}.**

Ces groupes sont différents des groupes obtenus à l’issue de l’étape précédente, donc nous fairons une autre itération.

1. **Troisième itération**

Groupe initiale : **Groupe 1= {G}, Groupe 2 = {A, D, F, I}, Groupe 3 = {B, C, E, H}.**

* Calcul des points Moyen de chaque groupe :

On obtient le tableau suivant comportant les point moyenne de chaque groupe

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Groupe 1 = {G}** | **Groupe 2 = {A, D, F, I}** | **Groupe 3 = {B, C, E, H, I}** |
| **Moyenne Groupe 1** | **Moyenne Groupe 2** | **Moyenne groupe 3** |
| **AGE** | 0 | 0,469697 | 0,871212 |
| **SEXE** | 1 | 0 | 1 |
| **SALAIRE** | 0 | 0,3277778 | 0,516667 |

* Calcul de la distance de chaque point par rapport au point Moyens et déterminations des groupes de chaque point.

Nous obtenons les distances contenues dans le tableau suivant :

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Groupe 1 = {G}** | **Groupe 2 = {A, D, F, I}** | **Groupe 3 = {B, C, E, H, I}** |
| **Moyenne Groupe 1** | **Moyenne Groupe 2** | **Moyenne groupe 3** |
| **A** | 1,1959596 | **0,6015152** | 2,191919 |
| **B** | 1,44545455 | 1,6479798 | **0,057576** |
| **C** | 1,90909091 | 2,1116162 | **0,521212** |
| **D** | 1,77777778 | **0,2530303** | 1,610101 |
| **E** | 1,03030303 | 1,2328283 | **0,357576** |
| **F** | 1,89393939 | **0,2479798** | 1,493939 |
| **G** | **0** | 1,7974747 | 1,387879 |
| **H** | 1,16666667 | 1,6914141 | **0,478788** |
| **I** | 1,92828283 | **0,1419192** | 1,459596 |

En prenant la distance minimale de chaque point par rapport au trois moyenne de chacun des trois Groupes, on obtient à cette deuxième itérations les groupes suivants :

**Groupe 1= {G}, Groupe 2 = {A, D, F, I}, Groupe 3 = {B, C, E, H}.**

Ces groupes sont les mêmes que ceux obtenus à l’issue de l’étape précédente, donc nous arrêtons les itérations en retenant trois clusters ou groupes suivants :

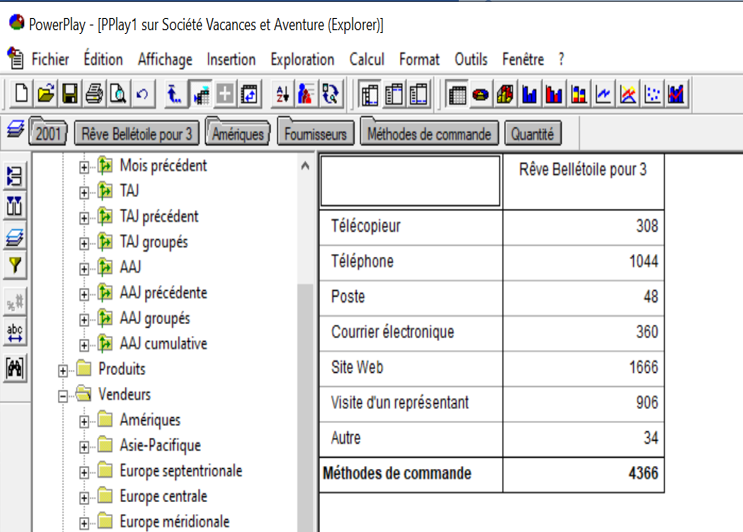
**RESULTAT : Groupe 1 = {G}, Groupe 2 = {A, D, F, I}, Groupe 3 = {B, C, E, H}.**

# Deuxième partie : Business Intelligence, TP POWERPLAY (Version 7)

Dans cette deuxième partie, il s’agit de répondre aux 15 questions de la partie 1, du TP **POWERPLAY et Transformer (Version 7),** cette partie ne concerne donc que **POWERPLAY**, nous reprendrons les questions et ensuite, les réponses à ces questions seront constituées d’impression écran des schémas, figures, tableaux, ou graphes réalisés.

1. Afficher le tableau donnant les quantités vendues de la tente **« Rêve Belleétoile pour 3** » pour l’année 2001 par les vendeurs d’Amérique ventilées par méthodes de commandes. Vous devez remarquer que le moyen le plus utilisé pour ces ventes a été le web.

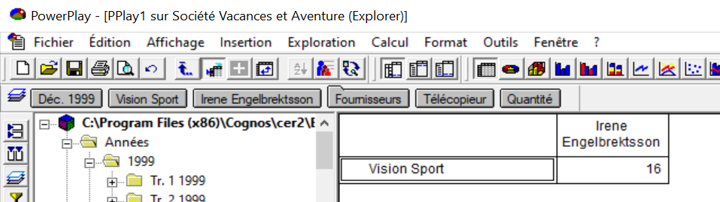
Le tableau ci-dessus, donnant les quantités vendues de la tente **« Rêve Belleétoile pour 3** » pour l’année 2001 par les vendeurs d’Amérique ventilées par méthodes de commandes, confirme bien que le moyen le plus utilisé pour ces ventes est le site web avec une quantité de **1666** vente.



*Figure*

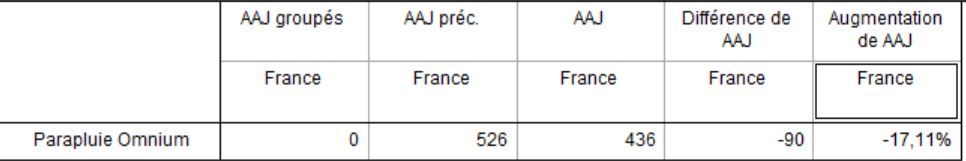
1. Retrouver la quantité de lunettes de type Vision Sport vendues par **Irène** **Engelbrektsson** la suédoise en décembre 1999.

Le tableau suivant nous permet de dire que cette quantité est de **16 lunettes.**

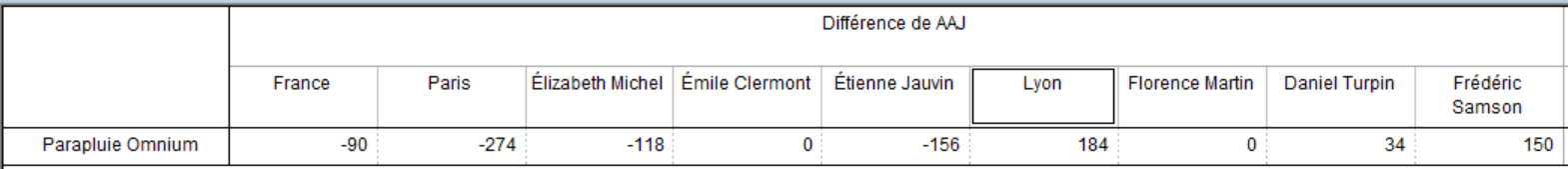
 : 

*Figure 1*

1. Combien de parapluies Omnium (matériel de Golf) en moins ont été vendu en France en 2001, on utilisera la dimension AAJ groupés ? Quel est le vendeur qui a fait le moins bon résultat ?



*Figure 2*

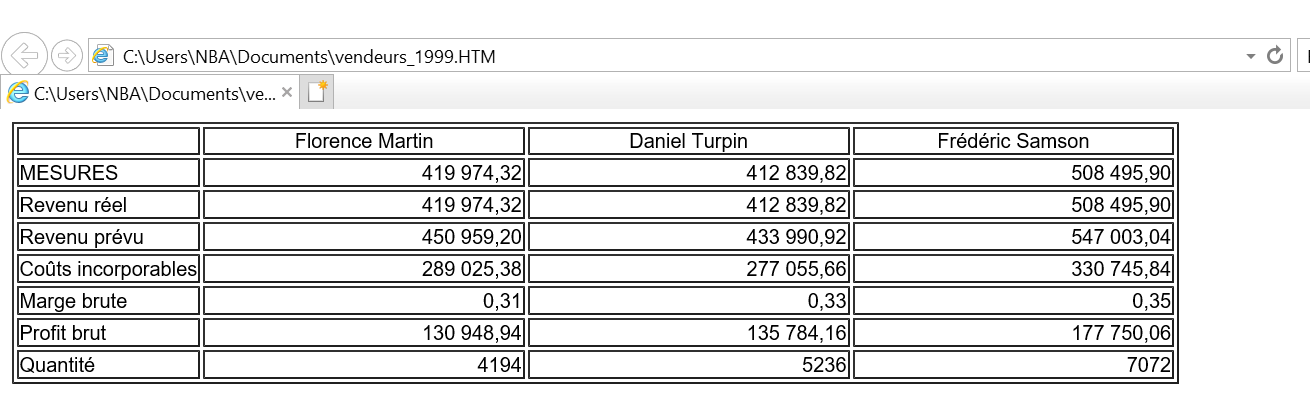


*Figure 3*

Les figures 2 et 3 ci-dessous, nous permettent d’affirmer que **90** parapluies Omnium ont été vendus en moins en 2001 que le vendeur ayant fait le moins bon résultat est : **Etienne Jauvin.**

1. Donner pour l’année 1999 et pour chacun des vendeurs de Lyon, les résultats obtenus selon toutes les mesures. Publier ce document sous format Html, visualisez-le avec votre « browser ».

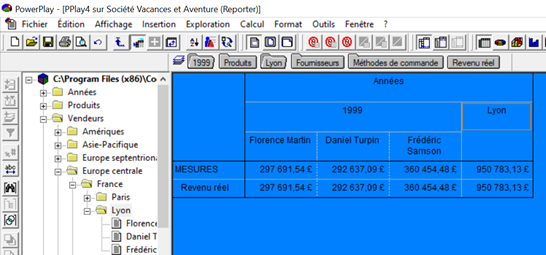
La figure suivante nous donne une capture d’écran du rapport des vendeurs de Lyon en 1999 selon toutes les mesures, visualisé avec le browser internet explorer :



*Figure 4*

1. Vous souhaitez envoyer les résultats en termes de revenus réels à vos collègues Britaniques, apportez la modification pour que les résultats s’affichent en Livre Sterling et modifiez le fond d’écran.

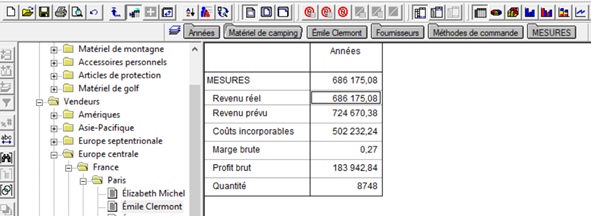
Le tableau suivant donne les résultats en livre sterling des revenu réels des vendeurs de Lyon pour l’année 1999 en fond d’écran de couleur bleu.



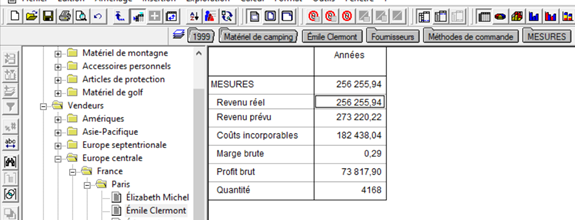
*Figure 5*

1. Pour Emile Clermont (Vendeur de Paris) donnez ses résultats selon toutes les mesures pour le matériel de camping. A-t-il réalisé son chiffre d’affaire ?

Les figures 6 pour toute les années et 7 pour l’année 1999 suivante donne les résultats pour toutes les mesures de Emile Clermont, on constate que son revenu réel est inferieur à son revenu prévu, ce qui nous pousse à conclure qu’il n’a pas réalisé son chiffre d’affaire.



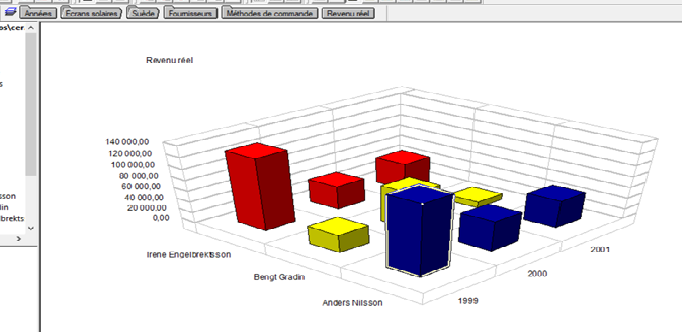
*Figure 6*



*Figure 7*

1. Faire un rapport (graphique) présentant l’évolution des revenus réels des vendeurs suédois pour les écrans solaires sous forme graphique.

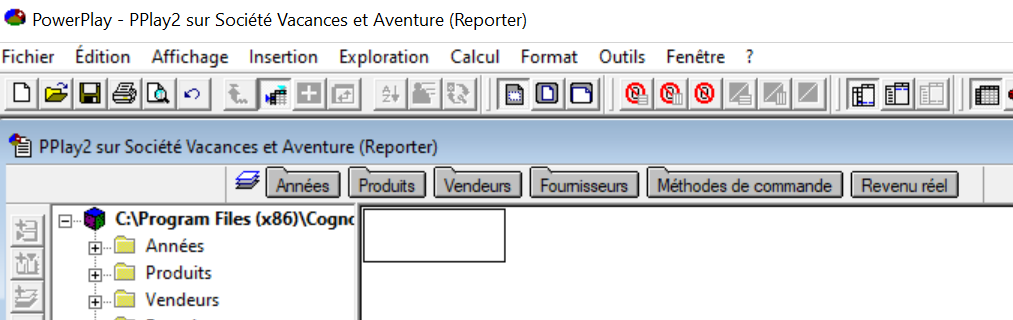
La figure 8, suivante représente le diagramme en barre 3D de l’évolution des revenus réels des vendeurs suédois.



*Figure 8*

1. L'outil qui vient d'être utilisé s'appelle **Explorer*.*** Il permet de trouver les secteurs d'activité que l'on désire analyser plus finement. Pour en faire un rapport pertinent il faut ensuite basculer vers l'outil **Reporter** en utilisant l’icône « Modifier le mode de document » : ou en utilisant le menu Exploration puis « Basculer vers Reporter ».

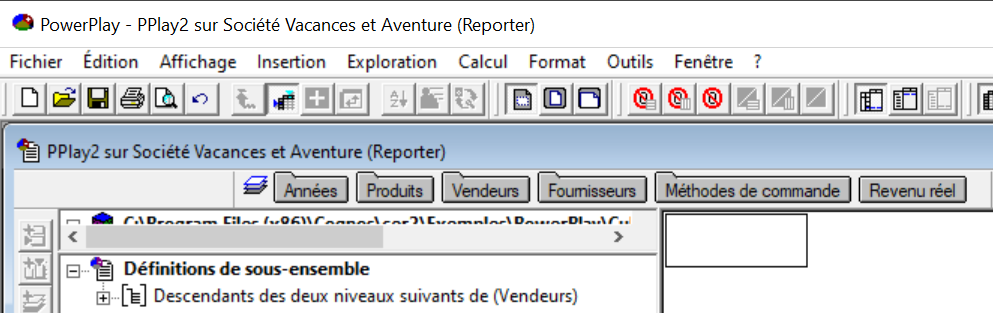
La figure 9, suivante confirme le basculement au mode Reporter fais en utilisant le menu Exploration puis « Basculer vers Reporter ».



*Figure 9*

1. Créer une sous-catégorie composée des descendants des deux niveaux suivants de la dimension Vendeurs en utilisant l’cône.

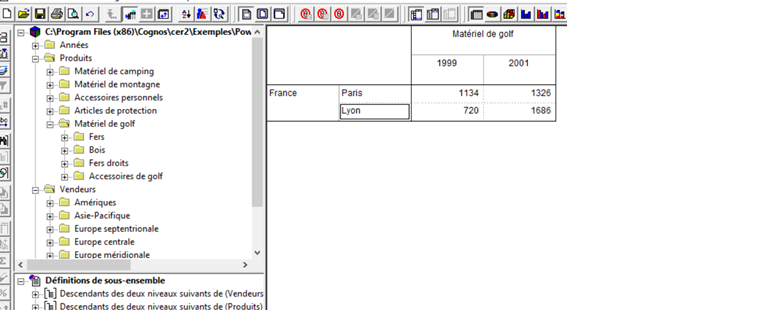
La figure 10 suivante confirme la création des descendants des deux niveaux suivants de la dimension Vendeurs.



*Figure 10*

1. Créer un rapport avec les sous ensemble précédent pour les années 99 et 01 en prenant comme mesure la quantité.

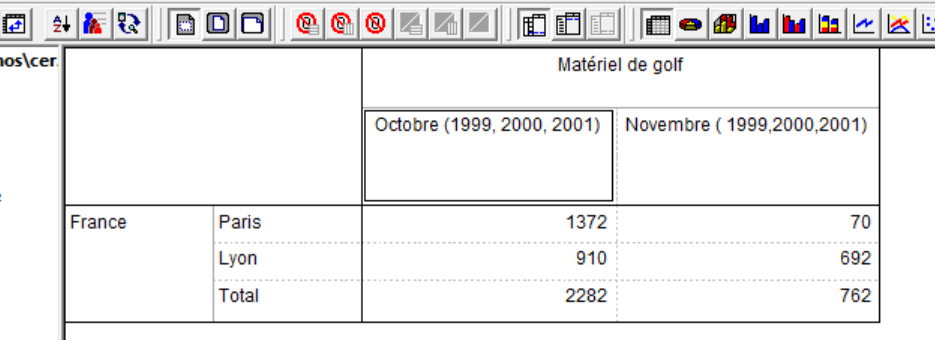
La figure suivante présente le rapport de vente en France du Matériel de Golf de 1999 et de 2001, elle présente les deux derniers niveaux de la dimensions Vendeurs.



*Figure 11*

1. Créer un sous-ensemble évolué en utilisant l’icône. Il doit contenir les matériels de Golf vendu en France en octobre et novembre. Chacune de ces conditions doit faire l’objet d’une définition à part. Il n’est pas possible d’utiliser les opérateurs logiques.

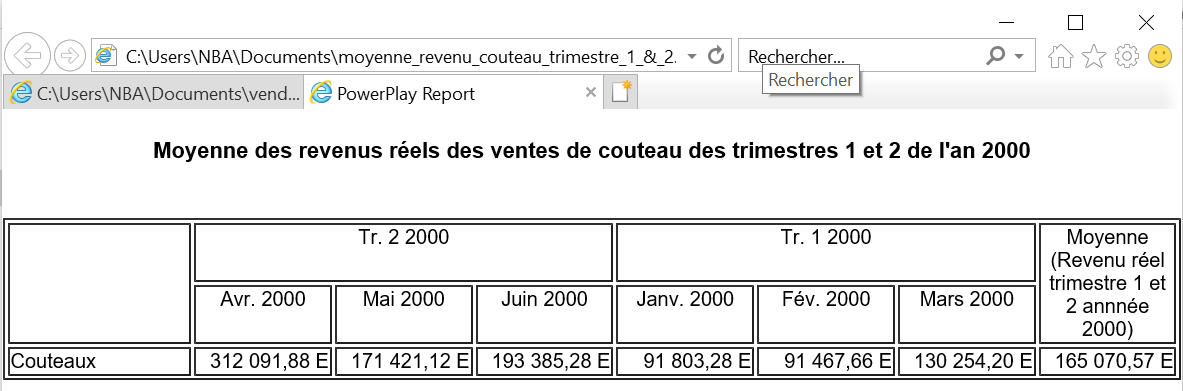
Nous avons crée deux sous-ensembles évolués l’un comportant le matériel de Golf, l’autre comportant les ventes en France, et nous avons crée également une sous-catégorie comportant les descendants de niveau de plus fin de la dimension année. Ainsi nous avons obtenu le rapport suivant comportant les Matériel de Golfs vendu en France en Octobre et Novembre de chaque année puis nous avons créé une colonne calculant la somme de chaque mois, ensuite en masquant les colonnes et en gardant la colonne des sommes nous avons obtenu le rapport ci-dessus :



*Figure 12*

1. Calculer la moyenne du revenu réel des deux premiers trimestres de l’an 2000 des couteaux. Publier le sous format html.

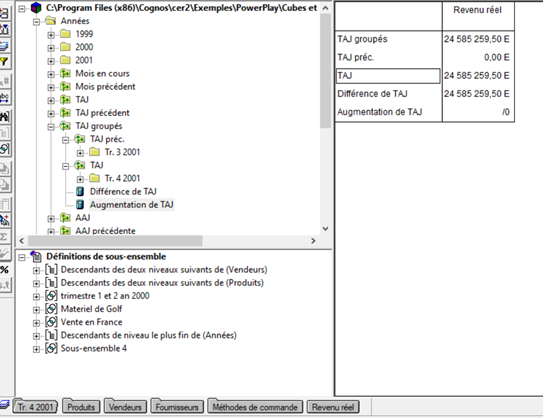
La figure suivante illustre le rapport contenant la moyenne des deux premiers trimestres de l’an 2000 des couteaux au format html et afficher avec le browser internet explorer.



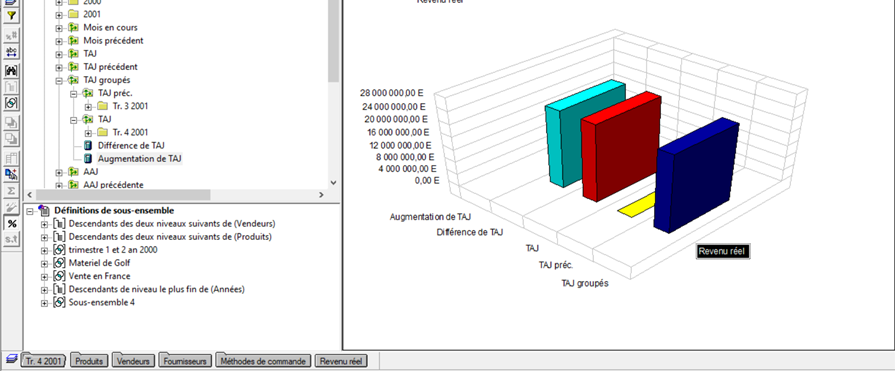
*Figure 13*

1. On veut visualiser la croissance du revenu du 4ème trimestre de 2001. Pour ce faire utiliser la sous la sous-catégorie TAJ groupés (trimestre à ce jour).

Le tableau suivant figure 14 et le diagramme en barre 3D figure 15, représente la croissance du 4 ème trimestre 2001.



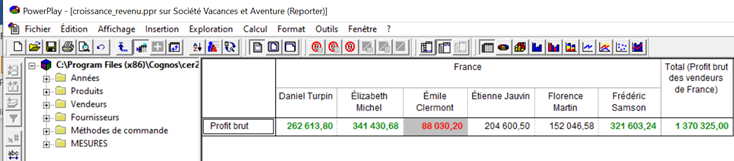
*Figure 14*



*Figure 15*

1. Créer un nouveau rapport avec en colonnes les vendeurs de France et en ligne la mesure **Profit Brut** en se limitant à l’année 2001. Choisir dans le menu **Exploration** l'option **Exceptions personnalisées.** Nommer votre exception : Profit et définissez des intervalles de valeur pour des styles que vous devez créer. Par exemple un profit brut en dessous de 100000 sera mauvais (Afficher en rouge) et au-dessus de 250 000 il sera bon (Afficher en vert).

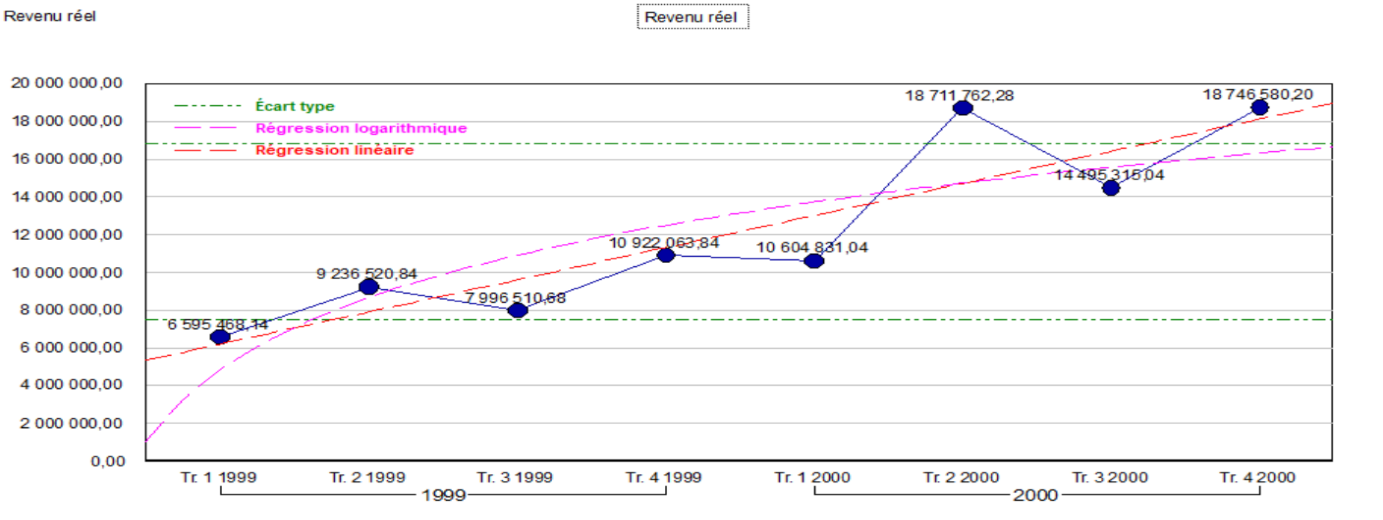
Le tableau suivant présente le rapport contenant les profits des vendeurs de France en 2001 avec la réalisation de l’exception Profit, permettant de colorier en vert les valeurs supérieures à 250 000 et en rouge les valeurs inferieures à 100 000, les valeurs entre 100 000 et 250 000 gardent la couleur de police par défaut (le noir).



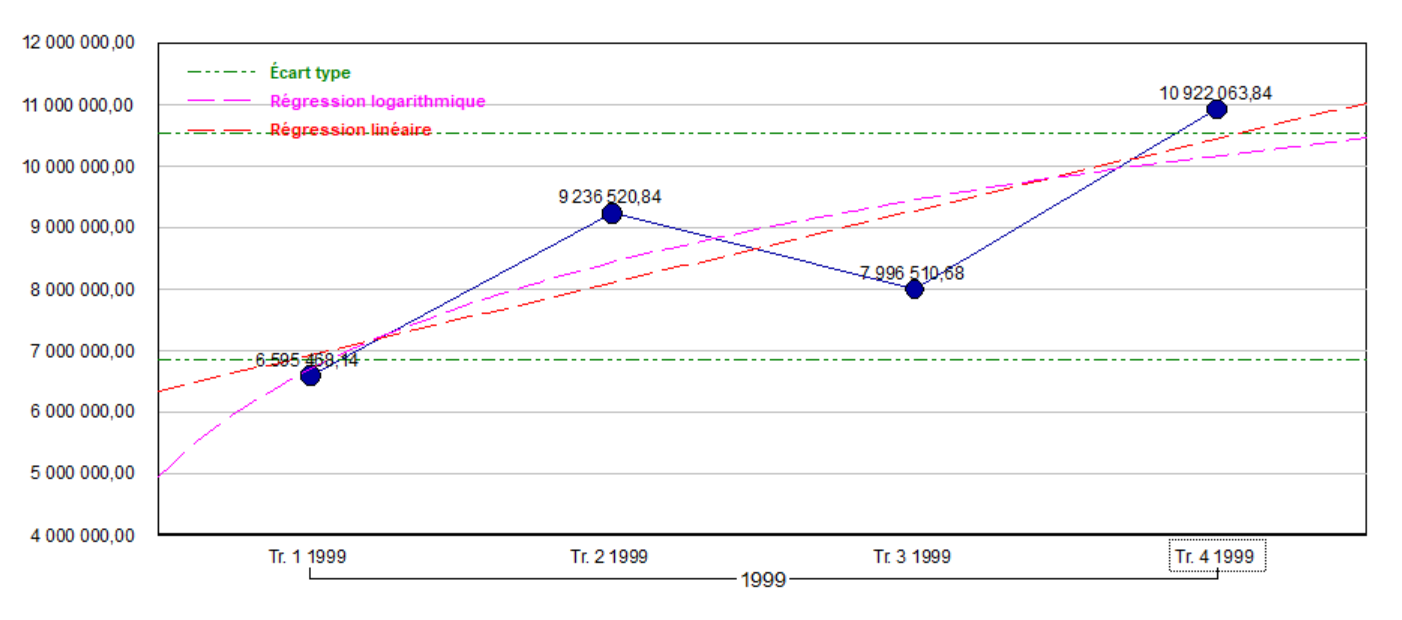
*Figure 16*

1. Tracer la courbe représentant l’évolution du revenu réel pour 1999 et 2000. Afficher également l’écart type, la régression linéaire et logarithmique pour chaque année.

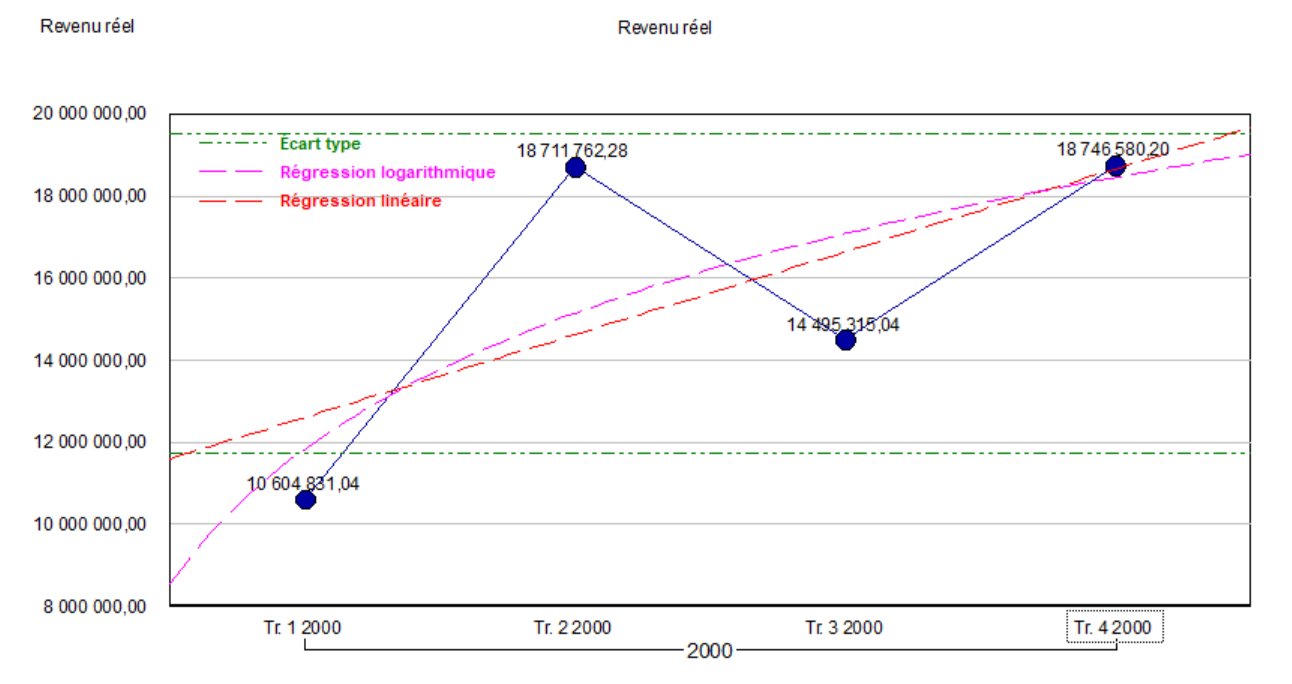
La figure 17 et 18 et 19 ci-dessus représente l’évolution du revenu réel pour 1999 et 2000 avec l’affichage sur cette courbe de l’écart type et la régression linéaire et Logarithmique pour les années 1999 et 2000 puis 1999 , suivi de 2000.Pour des besoins de lisibilité nous avons choisie comme unité de temps le trimestre, mais on peut facilement basculer ces courbe en mois en créant tout simplement une définition de sous ensemble évoluée et des restrictions sur les mois des années 1999 et 2000 .



*Figure 17 :* l’évolution du revenu réel pour 1999 et 2000



*Figure 18 :* l’évolution du revenu réel pour l’année 1999



*Figure 19 :* l’évolution du revenu réel pour l’année 2000.

# Conclusion

Au terme de ces deux travaux pratiques, force est de constater d’une part, la richesse des fonctionnalités de **POWERPLAY** bien qu’il ne soit pas possible de modifier les courbes à afficher à l’aide de fonction mathématiques personnelles, et que les définitions de sous catégories soient restrictives sur le nombre de contraintes, sa richesse dans la création de rapport a permis la réalisation du travail demandé.

D’autre part l’exécution manuelle de l’algorithme de K-Means utilisé dans le cadre de Clustering en Big Data, nous a permis de comprendre réellement son utilité ainsi que son fonctionnement.

Il aurait été souhaitable qu’il en fut ainsi pour l’ensemble des algorithmes utilisés en Big Data, et plus particulièrement les réseaux de neurones; de manière à maitriser les méandres de l’apprentissage en profondeur, dont les applications sont très nombreuses de nos jours, néanmoins force est de constater que le déroulement de cette algorithme ainsi que son apprentissage nécessite la maitrise d’un ensemble d’outils mathématiques sans lesquels ils n’est pas possible de véritablement les comprendre, ce qui confirme la pensée de **Severinius BOECE (480-524, Rome)** , selon laquelle : **"Tout a été créé par les nombres qui étaient le modèle exemplaire dans l'esprit du créateur"**.