Data cleaning – Historique

1. Donnees Excel: FR\_Laundry
   * Split
     + SKU Variante (Format, Conc, HDLS, Nom)
     + Pack 2nd
     + Promo/Std
     + Refill/Std
   * MAXIFs/SUMIFs
2. Importation des données Excel dans R
3. Suppression des lignes SKU "" (total, vide)
4. Retirer les lignes où is.na( Taille SKU) == TRUE qui ne sont pas des lignes TOTAL (SM Mir)
   * *Non fait pour les autres fichiers normaux et \_s*
   * *Particulier aux données NIELSEN*
5. Split de la colonne SKU *Déjà fait dans Excel, copier-coller les colonnes traitées par Exel HM et SM*
6. Suppression des doublons
   * Critères d'unicité
     + Retail
     + Variante : Nom, Conc, Format, HDLS
     + Pack secondaire
   * Critères écartés
     + Promo/STD
     + STD/Refill
7. Ajout d'une colonne Total?
8. Passage en long format
   * Pour les tables distribution, ordre Retail puis SKU
   * Pour les autres tables, même ordre
   * Copier la colonne de données (promo value, promo volume…) dans la table longue globale hm\_g ou sm\_g
9. Tables globales: HM ou SM sans les retailers : hm\_gg et sm\_gg
10. Format liquide : hm\_ggl et sm\_ggl
11. Elimination des SKUs en dessous d'un certain seuil:
    * Seuil HM et SM séparés: 0.9
      + Sumifs des ventes PNP.Value par SKU dans hm\_ggls
      + Etude de la somme cumulée des ventes: graphique somme cumulée / #SKU
        1. Détermination du seuil
        2. Seuil choisi: 90% ventes, reste 250 SKU HM
      + Filtration de hm\_ggl selon seuil
    * Seuils HM et SM commun: 0.95 pour chacun des deux, reste 206 SKUs
      + HM: 71% PNP Value et Volume captés par ces 206 SKUs
      + SM: 75% PNP Value et Volume captés par ces 206 SKUs
12. Définition des variables globales liquides
    * Nom des variables
      + hm\_ggl.period: global sur tous les produits liquides
    * Variables calculées par période
      + Volume NP / P / PNP: total et moyenne
      + Value NP / P / PNP: total et moyenne
      + Prix moyen NP / P / PNP
      + Value.WD moyenne PNP
      + Proportion Volume P / (Volume PNP)
13. Autre seuil: seuil des périodes? NON pour le moment
    * Mettre les période en dessous d'une WD de 2 à 0 🡪 évite les parasites car sinon l'arondi du WD fausse les mesures (0.5 arrondi à 1 par exemple)
14. Variables renormalisées
    * Renormalisation non centrée NonPromo.Volume.Rel
      + P/NP/PNP Volume : divisé par total période
      + P/NP/PNP Value: divisé par total période
    * Renormalisation centrée en écart relatif à la moyenne
      + P/NP/PNP Volume
      + P/NP/PNP Value
      + Prix P/NP/PNP
      + Value/WD
15. Variables de packs
    * Catégorie pack : 1, 2, 3, 4
16. Variables d'environnement concurrentiel
    * # SKU dans la catégorie Format x Catégorie Pack x Concentration
    * # SKU dans Format x Catégorie Pack x Concentration x Marque
17. Fichier \_num: on ne garde que les variables numériques – matrice de corrélation
18. Fichiers \_num.red : variables numériques réduites (kick NonPromo/Promo Volume…)
19. Fonctions de seuillage automatique dfSeuil39
    * Pour les fichiers du type hm\_ggl : toutes les périodes
    * Seuil sur les ventes
20. Fonctions de renormalisation centrée renormCenter
    * pour les fichiers du type hm\_ggls9, donne hm\_ggls9rcs9 ou hm\_ggls9rc selon le 2nd argument de la fonction
    * renorm pour les fichiers seuillés vente
21. Fonction seuillage WD dfSeuilWD39
    * Pour les fichiers du type hm\_ggl : toutes les périodes
    * On élimine les lignes où le WD est en dessous d'un certain seuil
22. Fonctions de renormalisation centrée renormCenterWD
    * renorm pour les fichiers seuillés WD (possible seuil ventes également avant)
23. Update de la fonction insert.CountCompetitors et insert.CountBrandCategory, nouveau nom insert.CountCompetitorWD
    * Prend en compte le seuillage WD 🡪 Si un SKU a été éliminé car son WD est en dessous du seuil, alors il n'est pas compté dans la concurrence
    * La fonction WD appliqué à un data frame non seuillé par WD fonctionne aussi, elle donne le même résultat que l'ancienne fonction. Son résultat ne dépend cependant pas de chaque période

05/08/15 : suppression des doublons

* Paramètres retenus pour l'unicité:
  + Retail
  + Variante
  + Pack secondaire
  + Marque
  + HDLS
  + Format
  + Standard vs Concentré
  + Nb lavages
  + Refill kické : maxif cohérent car on reste dans les mêmes magasins et sumif
* Ventes: sumifs
* Distribution: maxif
  + Vérifié sur 10 SKUs: deux cas de figure compatibles avec le maxif
    - Doublon nul
    - Les ventes et la distribution sont transféré presque en une seule période d'un item à l'autre
    - Refill: cohérent de prendre le maxif car réapprovisionnement dans les mêmes magasins

**Index des tables:**

**Données brutes**

hm\_g: tous les SKUs avec toutes les variables brutes et tous les retailers

hm\_gg: tous les SKUs avec toutes les variables brutes avec retailers agrégés

hm\_ggl: tous les SKUs liquides avec toutes les variables brutes

**Données seuillées**

hm\_ggls: tous les SKUs liquides au-dessus du seuil des 0.95 HM&SM avec toutes les variables brutes

hm\_ggls9: tous les SKUs liquides au-dessus du seuil des 0.90 HM avec toutes les variables brutes

**Données renormalisées**

hm\_gglr: tous les SKUs liquides avec toutes les variables renormalisée par division par la valeur des ventes totales sur chaque période 🡪 saisonnalité corrigée

hm\_gglrg: tous les SKUs liquides avec toutes les variables renormalisée par division par la valeur des ventes totales sur toutes les périodes 🡪 saisonnalité non corrigée

hm\_gglsr: tous les SKUs liquides au dessus du seuil des 0.95 HM&SM avec les variables renormalisées par division par la valeur totale de la période pour l’ensemble des SKU non centrées

hm\_gglsrc: tous les SKUs liquides au dessus du seuil des 0.95 HM&SM avec les variables renormalisées centrées : écart à la moyenne de la période – totaux globaux

hm\_gglsrs : tous les SKUs liquides au dessus du seuil des 0.95 HM&SM avec les variables renormalisées – totaux seuillés

hm\_gglsrcs: tous les SKUs liquides au dessus du seuil des 0.95 HM&SM avec les variables renormalisées centrées – totaux seuillés

sm\_ggls9: tous les SKUs liquides au dessus du seuil des 0.90 SM avec toutes les variables brutes

\*) Tables de sommes par SKU sur toutes les périodes

hm\_ggl.sku: tous les SKUs avec tous les variables brutes sur l'ensemble des périodes

hm\_gglr.sku: tous les SKUs avec tous les variables renormalisée – divisée par total de tous les SKUs - sur l'ensemble des périodes

hm\_gglrc.sku: tous les SKUs avec tous les variables renormalisée – écart à la moyenne de tous les SKUs - sur l'ensemble des périodes

hm\_ggls.sku: tous les SKUs liquides au-dessus du seuil des 0.95 HM&SM avec tous les variables brutes sur toutes les périodes

hm\_gglsr.sku: tous les SKUs au-dessus du seuil des 0.95 HM&SM avec tous les variables renormalisée – divisée par total de tous les SKUs - sur l'ensemble des périodes

hm\_gglsrc.sku: tous les SKUs au dessus du seuil des 0.95 HM&SM avec tous les variables renormalisée – écart à la moyenne de tous les SKUs - sur l'ensemble des périodes

hm\_gglsrs.sku: tous les SKUs au dessus du seuil des 0.95 HM&SM avec tous les variables renormalisée – divisée par total de tous les SKUs au dessus du seuil des 0.95 HM&SM - sur l'ensemble des périodes

hm\_gglsrcs.sku: tous les SKUs au dessus du seuil des 0.95 HM&SM avec tous les variables renormalisée – écart à la moyenne de tous les SKUs au dessus du seuil des 0.95 HM&SM - sur l'ensemble des périodes

\*) Tables de sommes par période de tous les SKUs

hm\_ggl.period: totaux et moyennes pour tous les SKUs liquides par période

hm\_ggls.period: totaux et moyennes pour tous les SKUs liquides au dessus du seuil 0.95 HM&SM par période

**Historique**

**17/08**

* Catégorie Taille
* # SKU catégorie
* Plots durée de vie SKU : Value.WD.Diff / Période : RAS.
  + Seul effet: court effet de lancement mais non significatif statistiquement ?
  + Davantage lié aux effets de bords dûs à la faible WD
  + Possible correction avec un palier de WD

**18/08**

* Reg lin simple sur hm\_ggls9rcs9
* Reg lin polynomiale deg 2 sur hm\_ggls9rcs9
* rpart sur hm\_ggls9rcs9
* randomForest sur hm\_ggls9rcs9