# Analyse des données importées

* 1. **Qui?**

Magasin au UK ...

* 1. **Quoi?**

On décrit les variables du dataset (length + fonction qui décrit les variables)

* 1. **Ou, Source?**

UK

* 1. **Quand?**

On utilise la fonction head et tail pour montrer la période couverte

* 1. **Combien?**

Nombre de lignes dans le data set (fonction dim)

# Nettoyage des données

* 1. **Données non pertinantes**
     1. Retirer les CustomerID vide
     2. C, C2, D, POST, Price < 0, Quantity <0, Dupliquées, M, BANK CHARGES, PADS DOT
  2. **Nombre de lignes Nettoyées**
     1. Nombre de lignes au total
     2. Problème avec les dates (2 formats différents qui ont été utilisés)
     3. % par rapport au dataset original

# Statistiques Descriptives

* 1. **Analyse Générale**
     1. Nb de pays
     2. Nb de factures
     3. Nb de clients
  2. **Analyse des ventes selon les pays**
     1. Piechart avec UK en comparaison avec les ventes en dehors
     2. Piechart des ventes en dehors de UK
     3. Piechart avec % ventes hors UK
     4. Pays avec le plus grand nombre de retour
  3. **Analyse des ventes**
     1. Summary (summary(invoiceData[2]), summary(salesPerProduct[2]))
     2. Boxplot des ventes (2 boxplots : quantity et sales per products)
     3. Month Graph
     4. Day Graph
     5. Hour Graph

# PCA

* 1. **DataSet par pays** (pairs())
     1. Pairs pays => non pertinent
     2. Pairs pays (sans UK) => pertinent
        1. Matrice des corrélations (cor(countyWithoutUKData))
        2. Sommaire des composants (summary(pcaCountyWithoutUK))
        3. Plot du pca (plot(pcaCountyWithoutUK))
        4. Poids des variables originales dans les composantes (Loadings(pcaCountyWithoutUK))
        5. Lien entre les composantes et les variables (score(pcaCountyWithoutUKAde4, xax=1) et score(pcaCountyWithoutUKAde4, xax=2))
        6. Cercle de corrélation (s.corcircle(pcaCountyWithoutUKAde4$co))
        7. scatter(pcaCountryWithoutUKAde4, posieig="none")
  2. **DataSet par produit** (pairs())
     1. Matrice des corrélations (cor(productData))
     2. Sommaire des composants (summary(pcaProduct))
     3. Plot du pca (plot(pcaProduct))
     4. Poids des variables originales dans les composantes (Loadings(pcaProduct))
     5. Lien entre les composantes et les variables (score(pcaProductAde4, xax=1) et score(pcaProductAde4, xax=2))
     6. Cercle de corrélation (s.corcircle(pcaProductAde4$co))
     7. Projection des données (scatter(pcaProductAde4, posieig="none", clab.row=0))

# Clustering