Présentation R

Arthur Lemoine, François Somville, Alexandre Antippas

Janvier 2019

Plan

1. Introduction
2. Exploration du Dataset
3. Statistiques descriptives
4. Analyse des composantes principales (ACP)
5. Clustering
6. Bonus

1. Introduction

Le dataset contient toutes les factures émise par un magasin en ligne basé en Angleterre. Le magasin vends des cadeaux uniques pour toutes les occasions.

Le dataset contient: nb de lignes

```
    InvoiceDate
1 1/12/10 08:26
2 1/12/10 08:26
3 1/12/10 08:26
4 1/12/10 08:26
5 1/12/10 08:26

          InvoiceDate
541905 9/12/11 12:50
541906 9/12/11 12:50
541907 9/12/11 12:50
541908 9/12/11 12:50
541909 9/12/11 12:50
```

Exploration des données dans le Dataset

Nombre d'enregistrements

```
[1] 541909
```

Les différentes variables

```
[1] "InvoiceNo"   "StockCode"   "Description" "Quantity"
[6] "UnitPrice"   "CustomerID"  "Country"
```

Exploration du Dataset

Nombre de pays différents

```
[1] 38
```

Nombre clients

```
[1] 4373
```

Nombre de factures

```
[1] 25900
```

Facture annulées

StockCode POST, D, ...

Nettoyage du dataset

Suppression des données inutiles

```
541909 lignes à la base

396337 lignes après nettoyage

Pourcentage retiré : 26.86281 %
```

Statistique descriptive

Visualisation des données

```
Nombre de factures (uniques): 25900
Nombre de produits (uniques): 4070
Nombre de clients (uniques): 4373
Nombre de pays: 38 Attention 37 + undefined
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-80995.00      1.00      3.00      9.55     10.00  80995.00
```