Méthodes de classification pour les produits en Display

Réalisé par :

Badreddine SALEH

Hajar BOUCHANE

PLAN

- Problématique
- Analyse exploratoire des données
- Construction des modèles:
  - Approche 1: numérisation des variables catégorielles
  - Approche 2: discrétisation des variables continues
    - Comparaison de discrétisation avec MDLPC et discrétisation avec arbre de décision
- Conclusion

Problématique

- le mot "**display**" peut être traduit par "présentoir" ou "vitrine". Dans un supermarché, un display est une présentation visuelle de produits destinée à attirer l'attention des clients et à les inciter à acheter les produits.
- Dans ce sens s'inscrit notre projet pour aider à prendre la bonne décision: **quel produit mettre en Display?**

Analyse exploratoire des données

- Dans ce rapport, nous allons analyser un jeu de données pour comprendre les caractéristiques et la structure des données et les relations qui peuvent exister entre elles.
- Les résultats de cette analyse nous aideront à découvrir de nouvelles connaissances.

Analyse exploratoire des données

	Y	X1	X2	X3	X4	X5	X6	X7
1	Display	cor_sales_in_vol	cor_sales_in_val	CA_mag	value	ENSEIGNE	VenteConv	Feature
2	No_Displ	2	20.2	47400	36	CORA	72	No_Feat
3	No_Displ	2	11.9	62000	24	LECLERC	48	No_Feat
4	No_Displ	8	29.52	60661	60	AUCHAN	480	No_Feat
5	No_Displ	2	16.2	59677	19	CARREFOUR	38	No_Feat

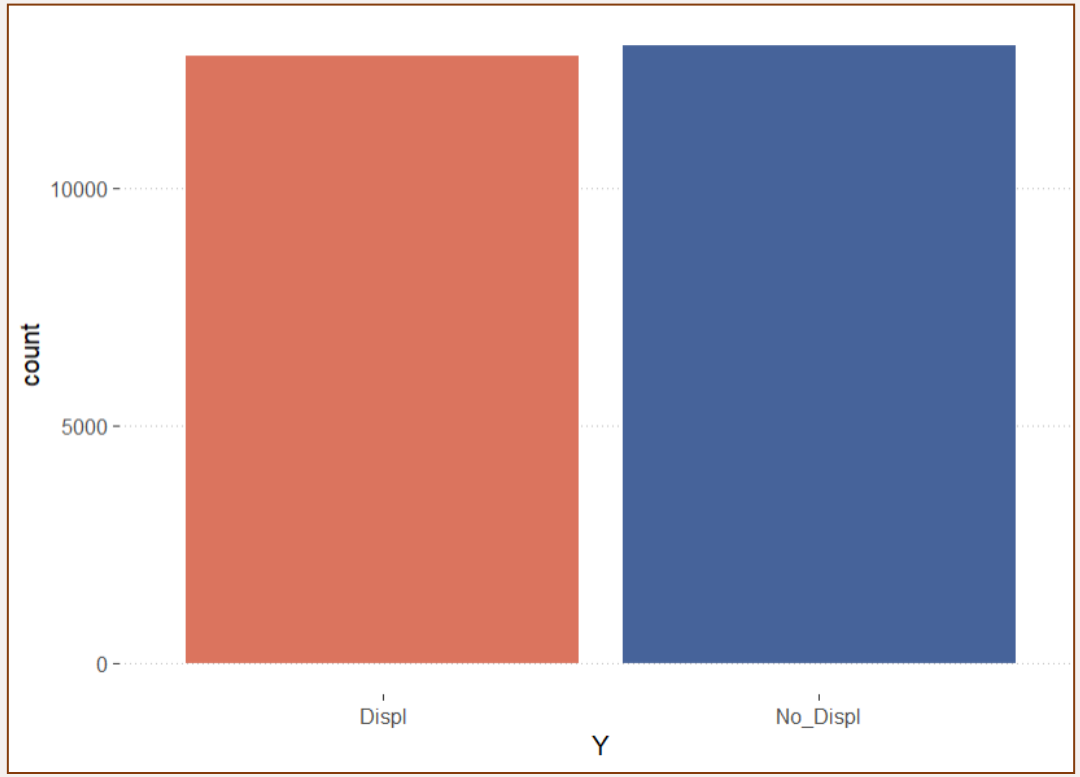
- Cette base de données contient 8 variables.
- La variable à expliquer : Y catégorielle
- Les variables explicatives : X1,X2,X3,X4 et X6 sont continues, X5 et X7 catégorielles

Statistiques descriptives univariées :

X1		X2		X3		X4		X6	
Min.	1.00	Min.	1.11	Min.	1693	Min.	1.00	Min.	1.0
1st Qu.	2.00	1st Qu.	13.36	1st Qu.	21394	1st Qu.	25.00	1st Qu.	50.0
Median	4.00	Median	31.05	Median	51522	Median	32.00	Median	120.0
Mean	13.77	Mean	130.89	Mean	64641	Mean	37.65	Mean	587.9
3rd Qu.	11.00	3rd Qu.	87.45	3rd Qu.	91000	3rd Qu.	40.00	3rd Qu.	380.0
Max.	1475.00	Max.	13589.22	Max.	284844	Max.	198.00	Max.	48816.0

- On a la moyenne est presque le triple de la médiane, pour les variables X1, X2 et X6, cela signifie que la distribution des données est décalée vers la droite (asymétrique), avec certaines valeurs très élevées.

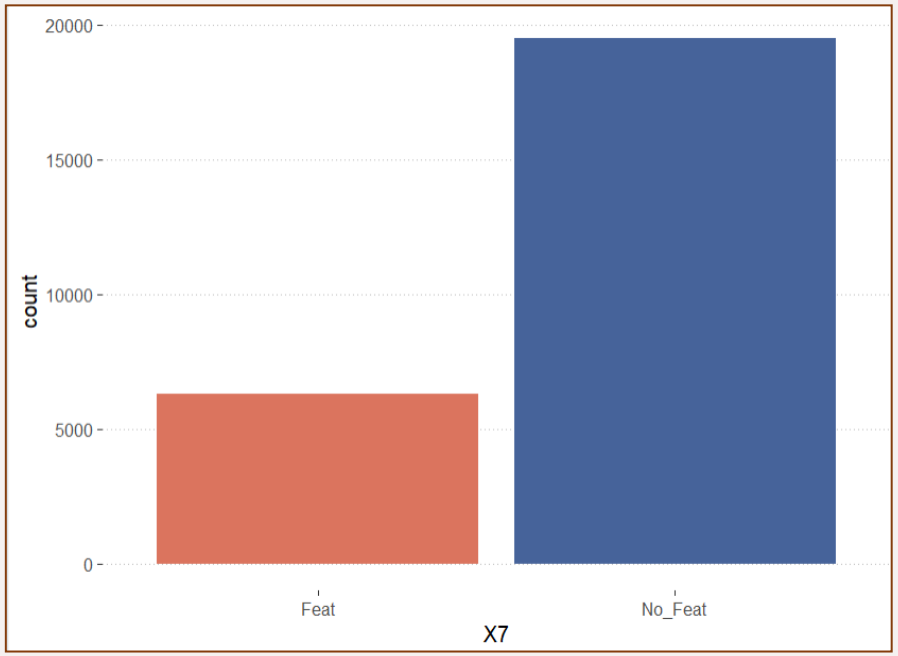
Analyse exploratoire des données

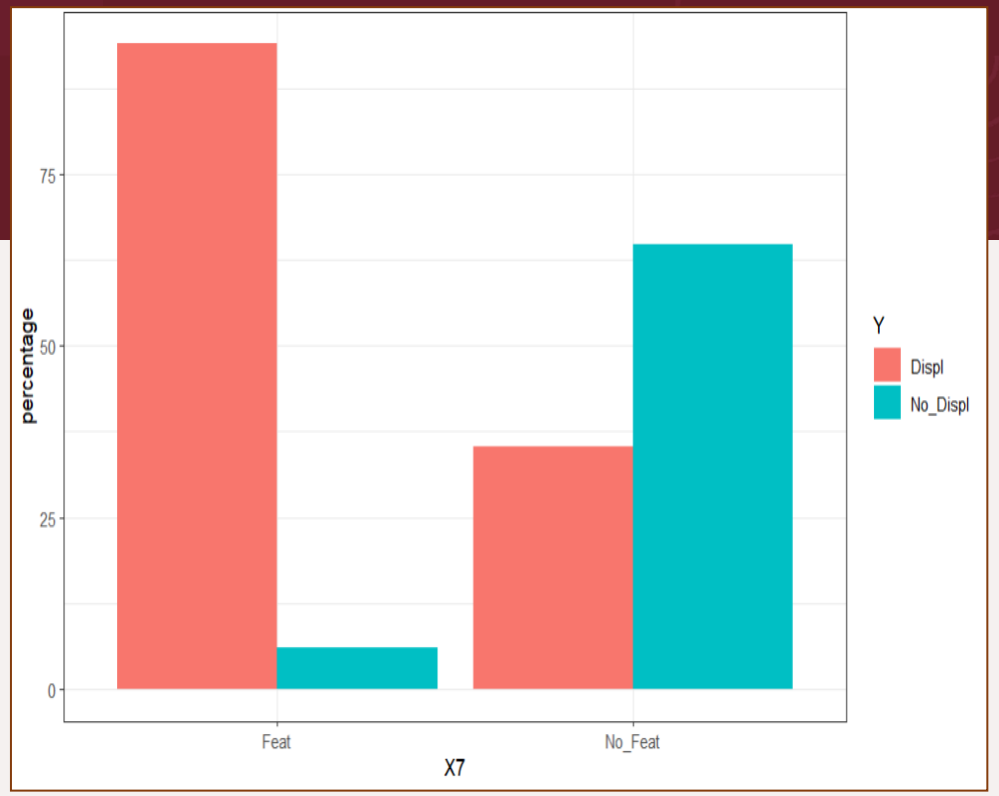


- Les deux classes de la variables Y sont bien équilibrées.

Analyse exploratoire des données

- Les deux classes de la variables X7 sont déséquilibrées, cela se justifie par le fait que les supermarchés ne peuvent pas promouvoir tous les produits.

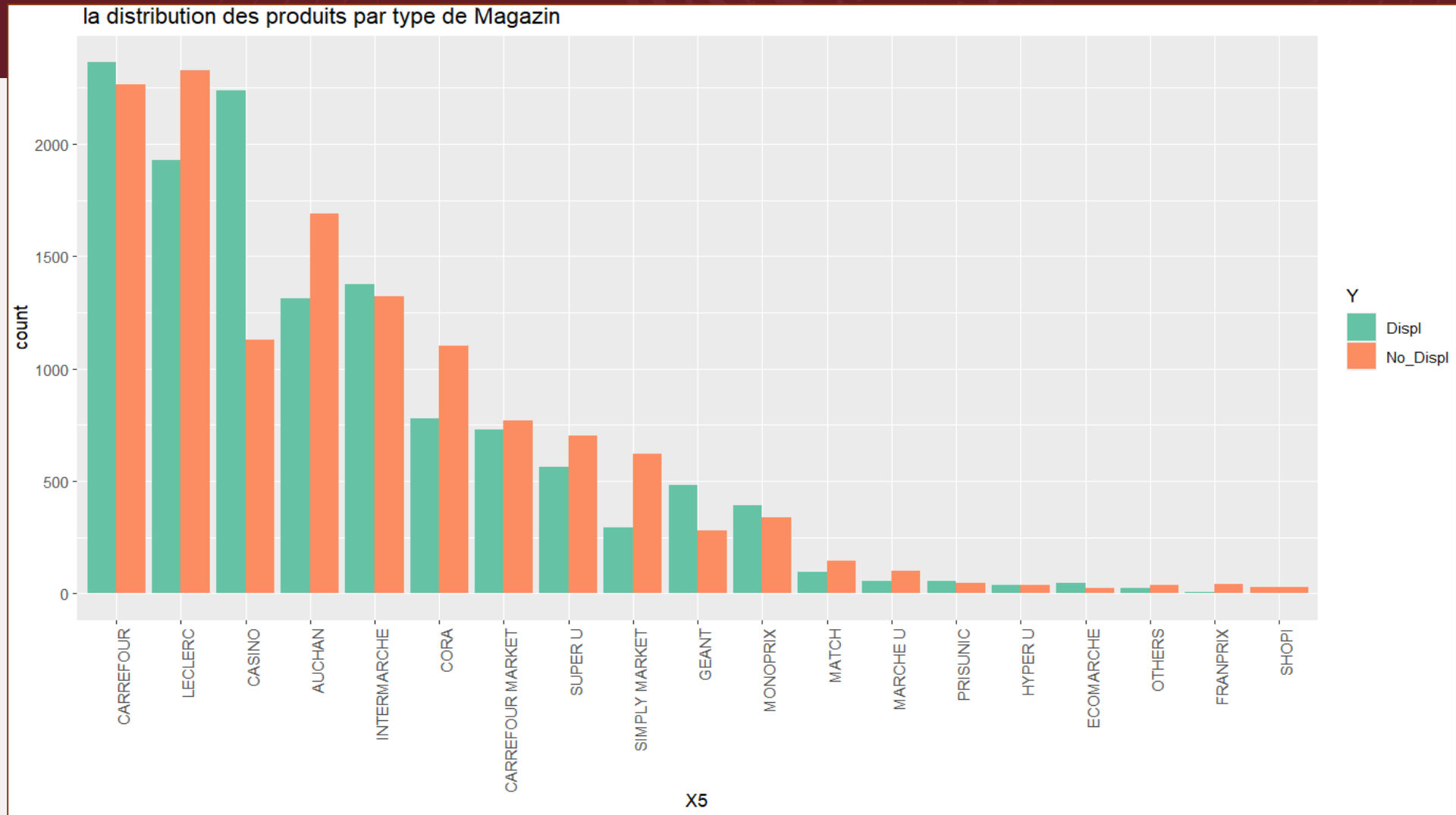




- 94% des produits promotionnels sont en Display. Ce qui montre à quel point la mise en Display du produit fait partie importante de la stratégie de promotion visant à attirer l'attention des consommateurs.

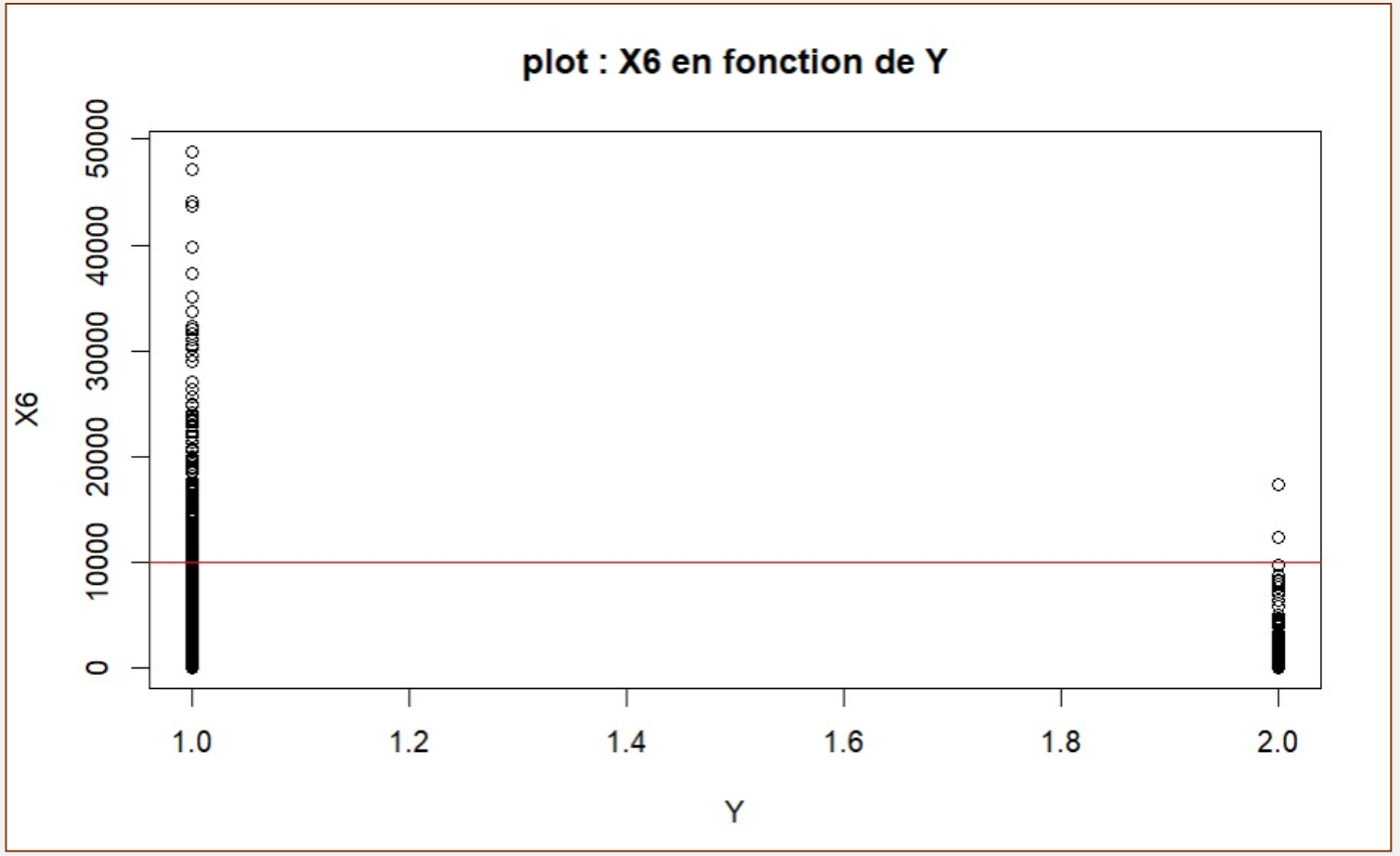
Analyse exploratoire des données

- La distribution des produits par type de Magazin.



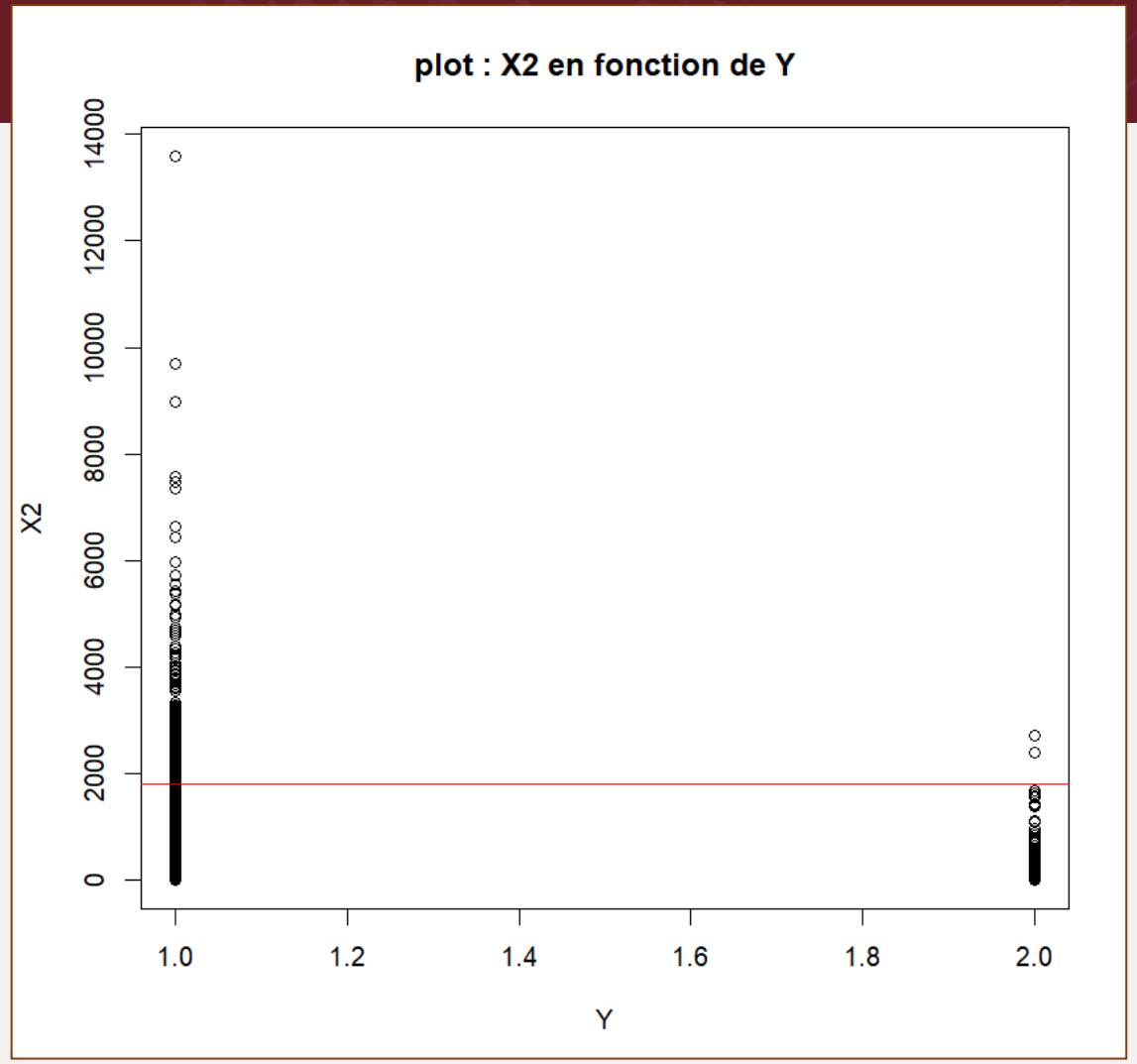
Analyse exploratoire des données

- Les produits qui ont un profit total des ventes plus que 10.000 sont toujours en Display.



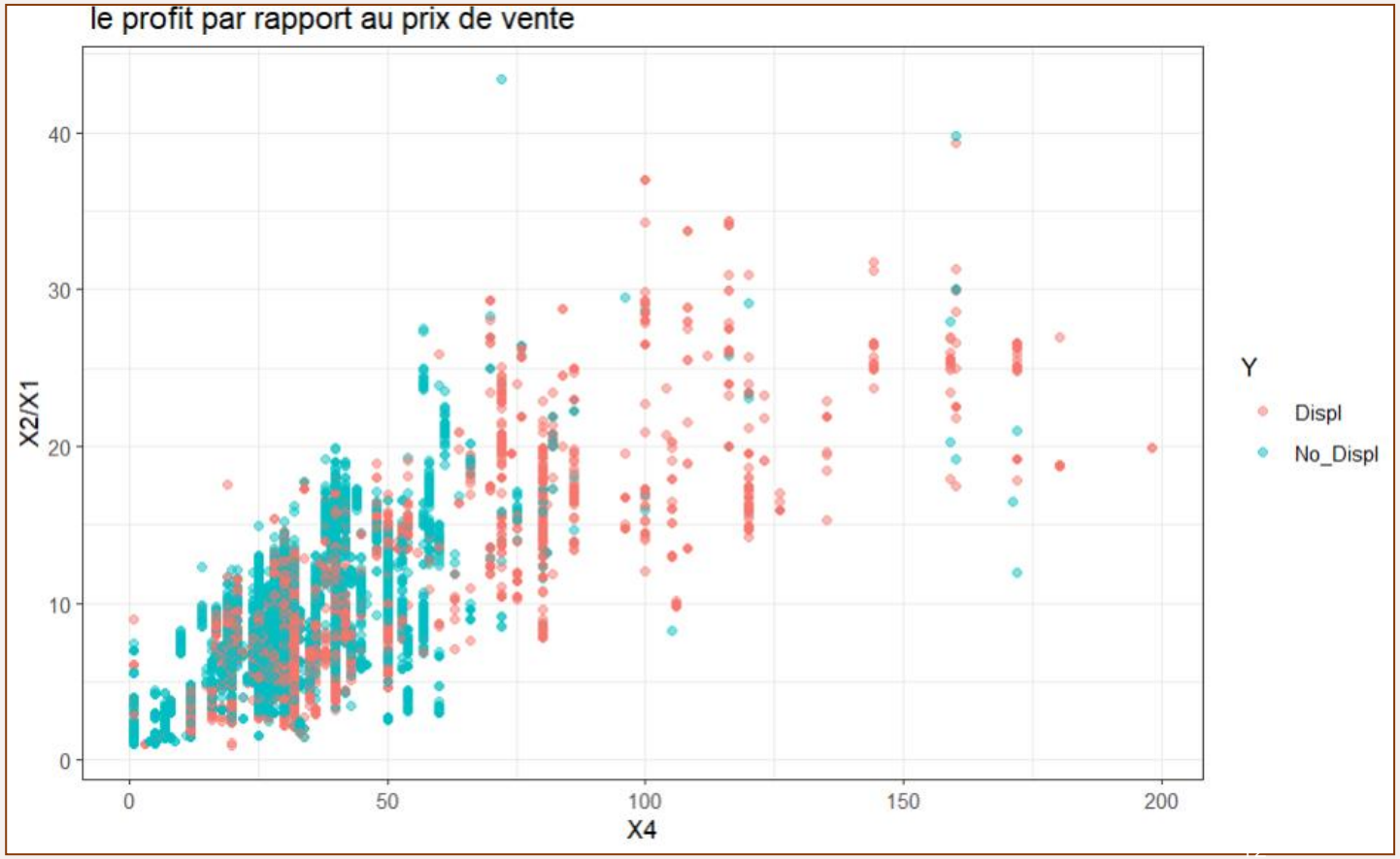
Analyse exploratoire des données

- Les produits qui ont un profit totale supérieur à 1900 sont toujours en Display.



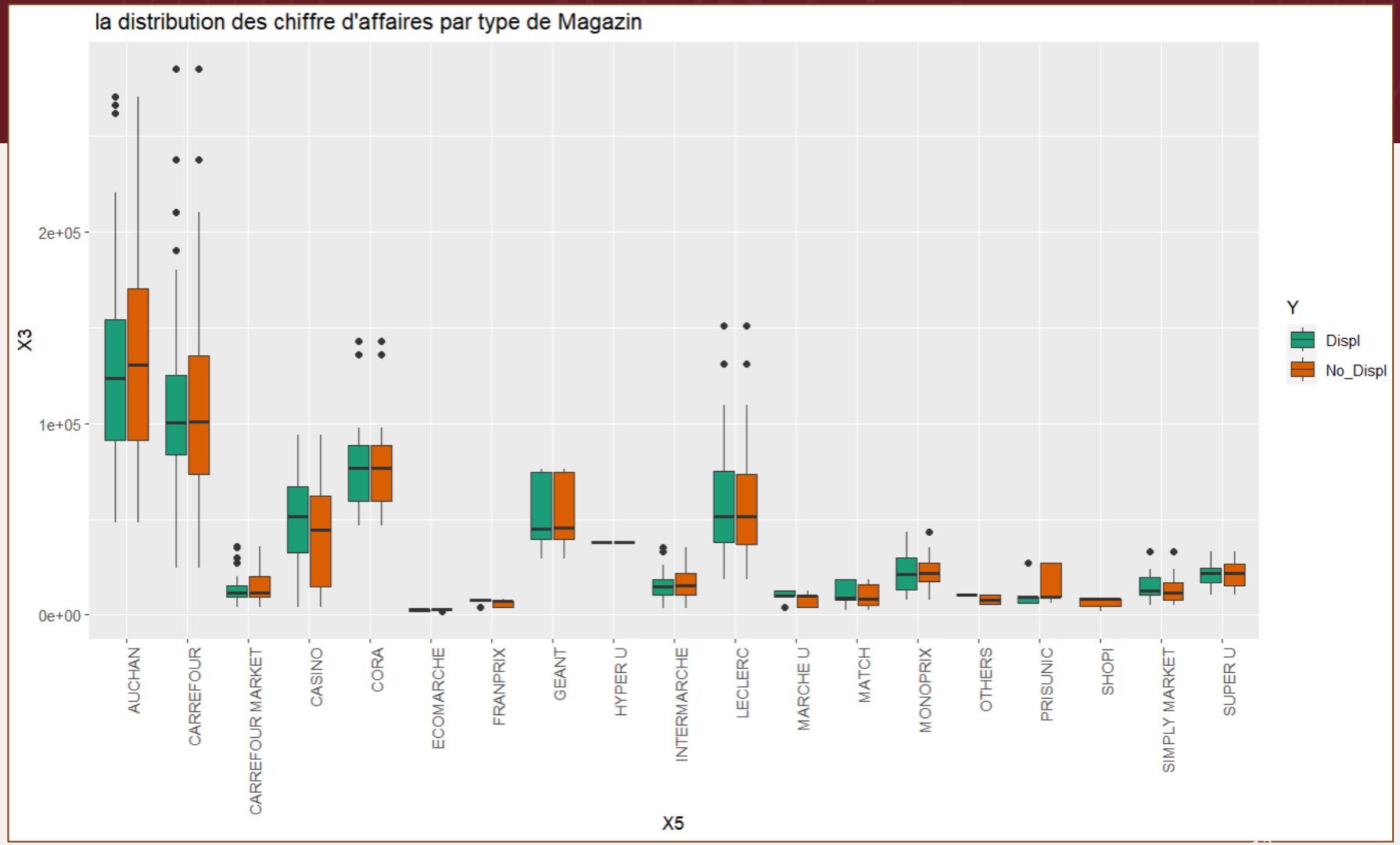
Analyse exploratoire des données

- Le profit par unité par rapport au prix de vente, qui sont positivement corrélé.
- Les produits ayant un prix de vente plus élevé ont une forte probabilité d'être en Display.



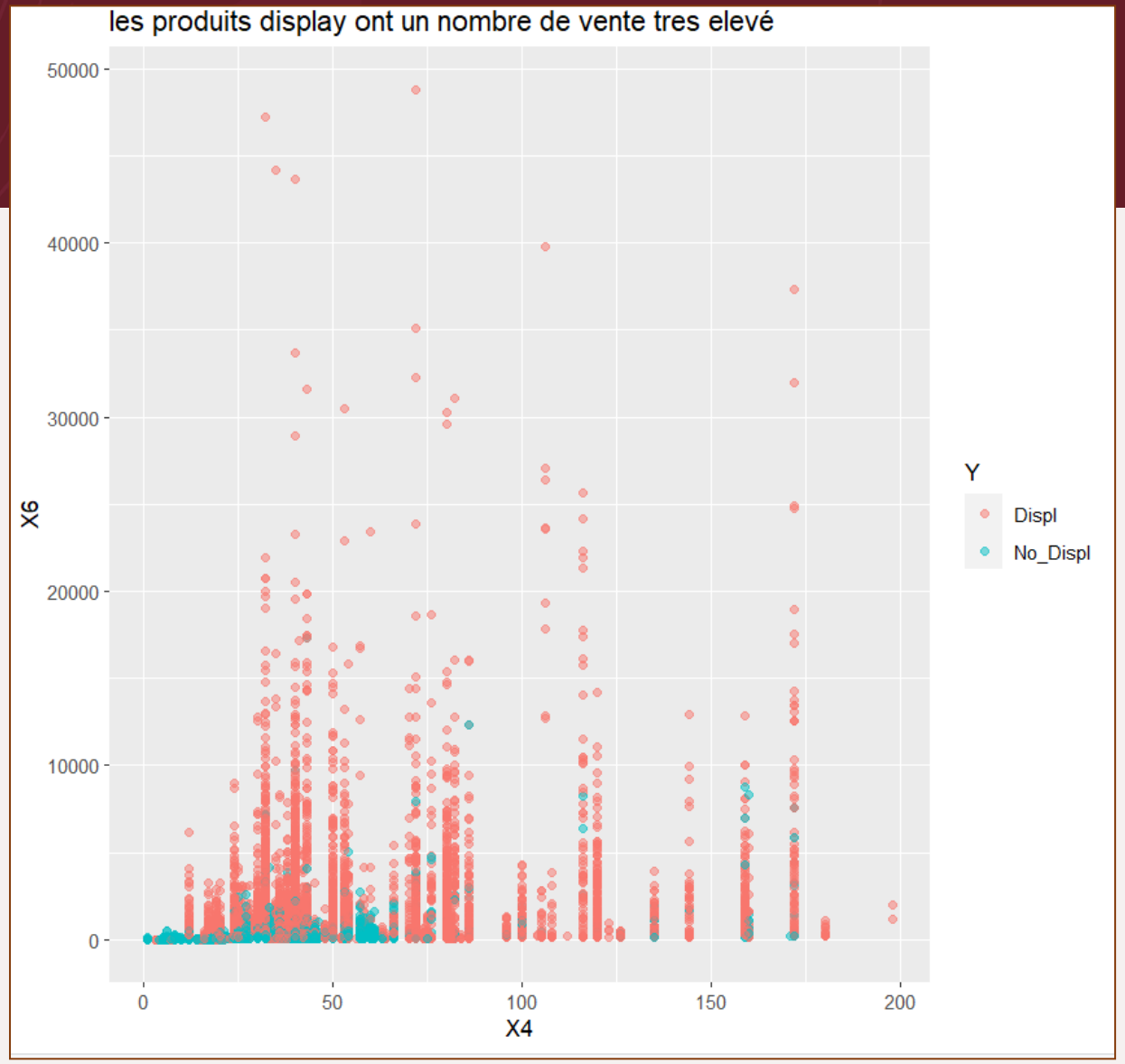
Analyse exploratoire des données

- Chiffre d'affaire par type de magazin.

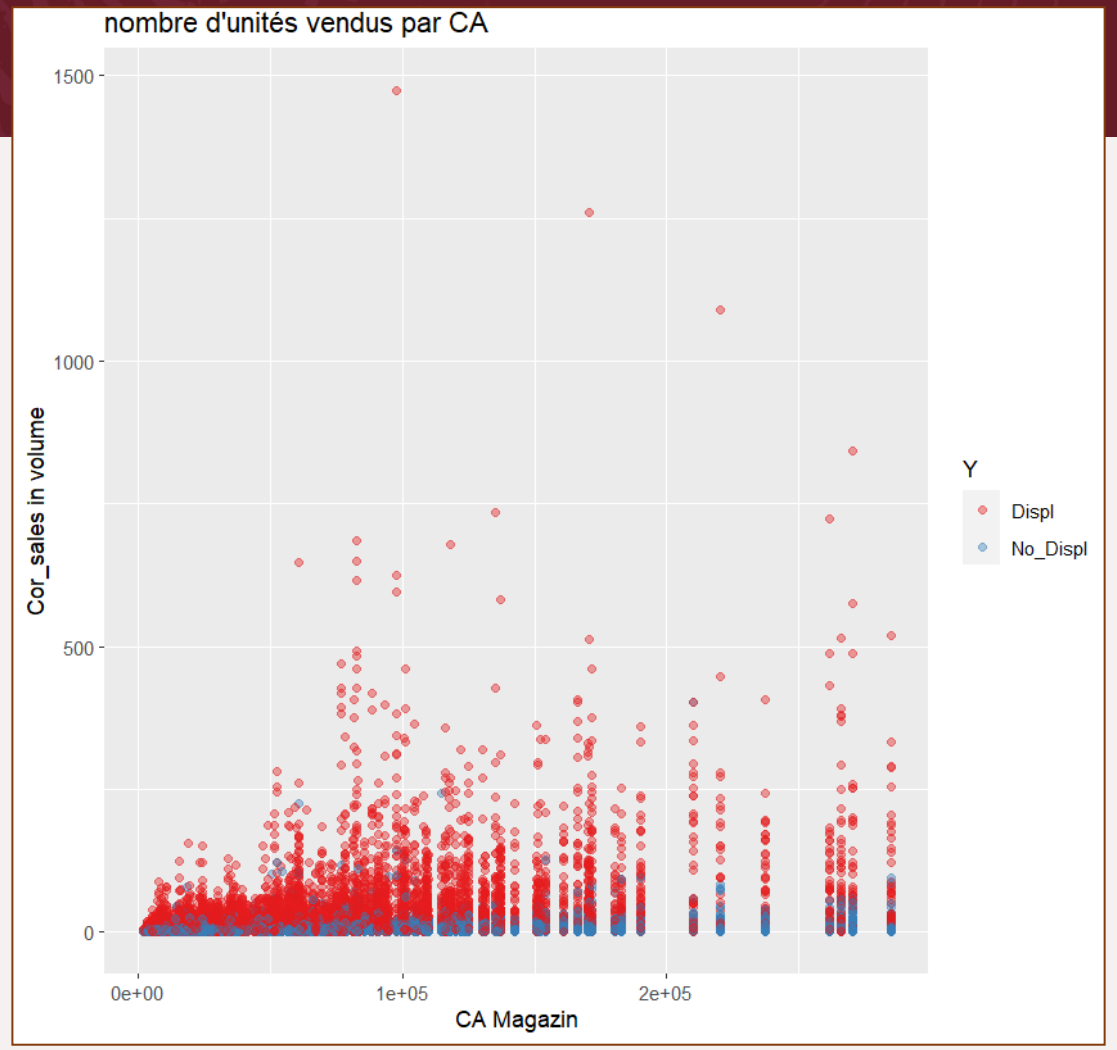


Analyse exploratoire des données

- Les produits Display génèrent un profit plus élevé.



Analyse exploratoire des données



- Les produits les plus vendus sont les produits Display.

Mesure de symétrie de la distribution statistique des variables (skewness)

- Mesure de l'asymétrie de Galton (mesure par des quantiles):
- Elle est comprise entre -1 et 1.
- 0 est pour une distribution parfaitement symétrique.

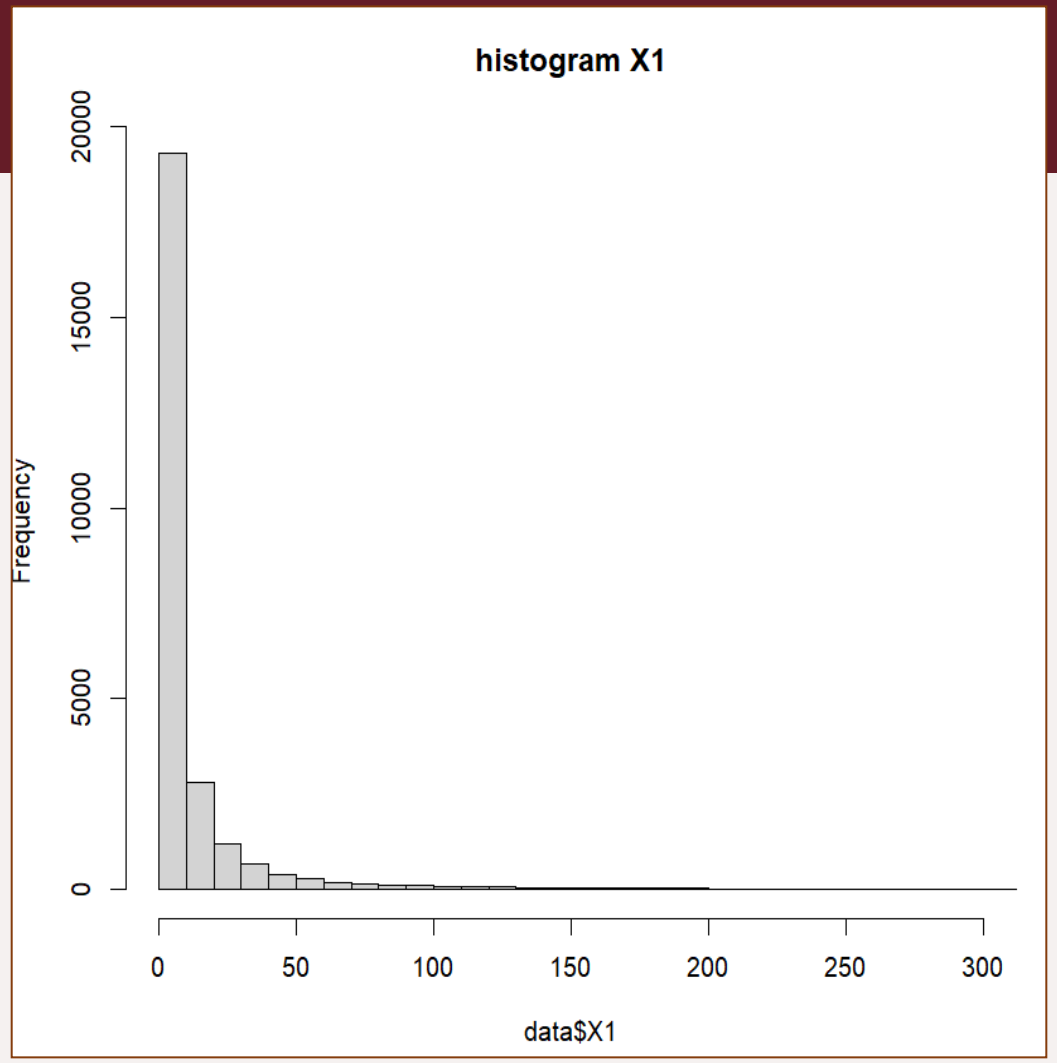
$$\frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

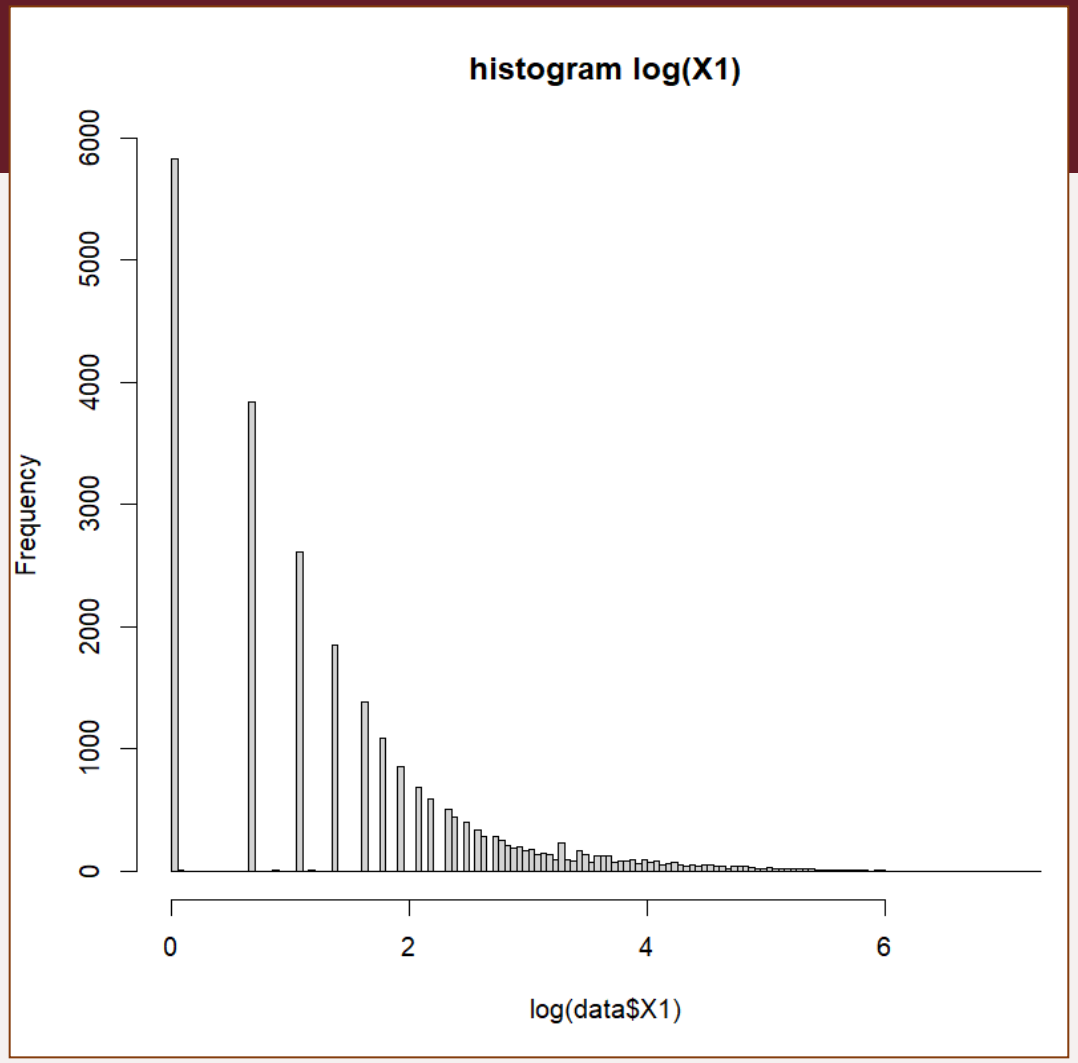
```
> galtonskew.proc(data$X1)
[1] 0.5555556
> galtonskew.proc(data$X2)
[1] 0.5224727
> galtonskew.proc(data$X3)
[1] 0.1343275
> galtonskew.proc(data$X4)
[1] 0.06666667
> galtonskew.proc(data$X6)
[1] 0.5757576
>
```

- Les variables X1,X2 et X6 ont une distribution très asymétriques.

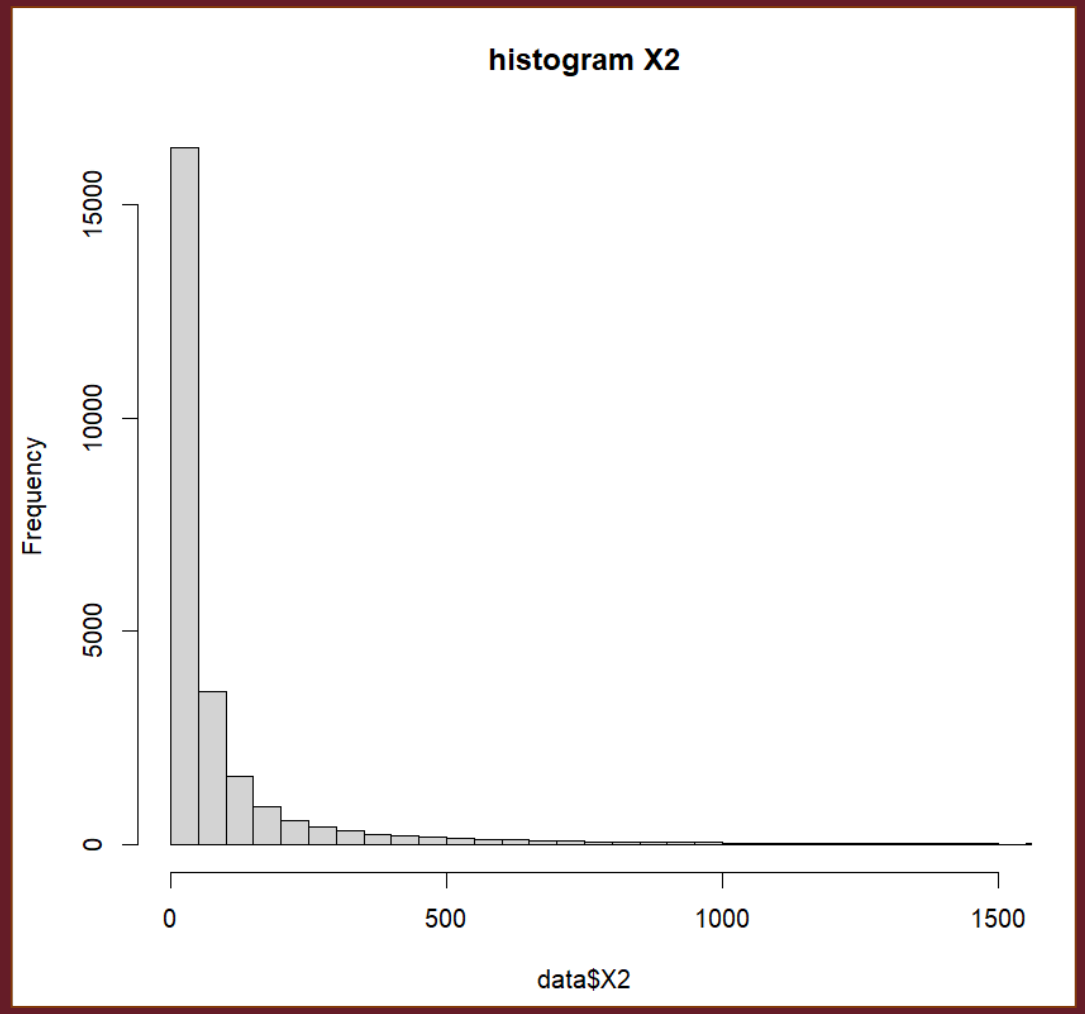
- Donc on a appliqué le Log pour rendre la distribution plus symétrique.

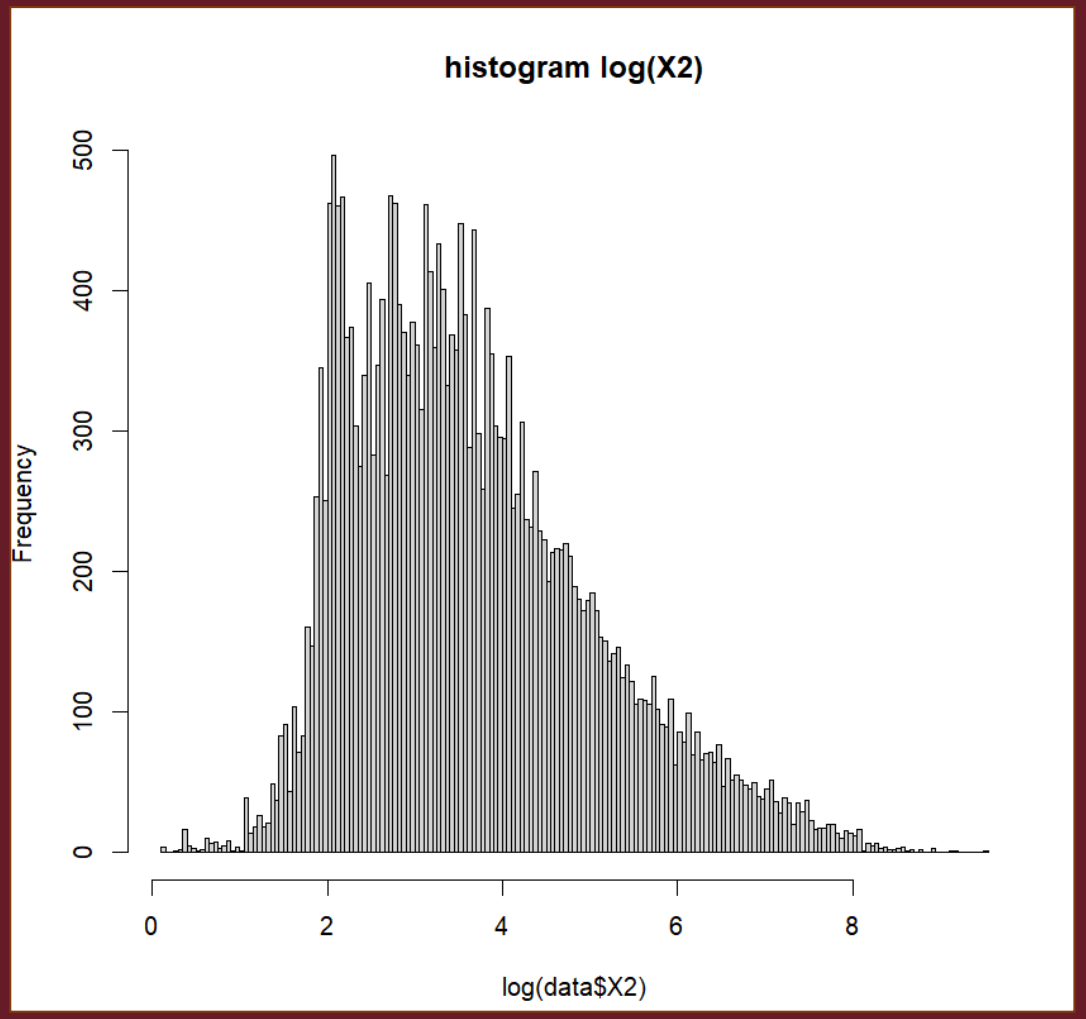
- La distribution de X1, avant et après la transformation par Log.



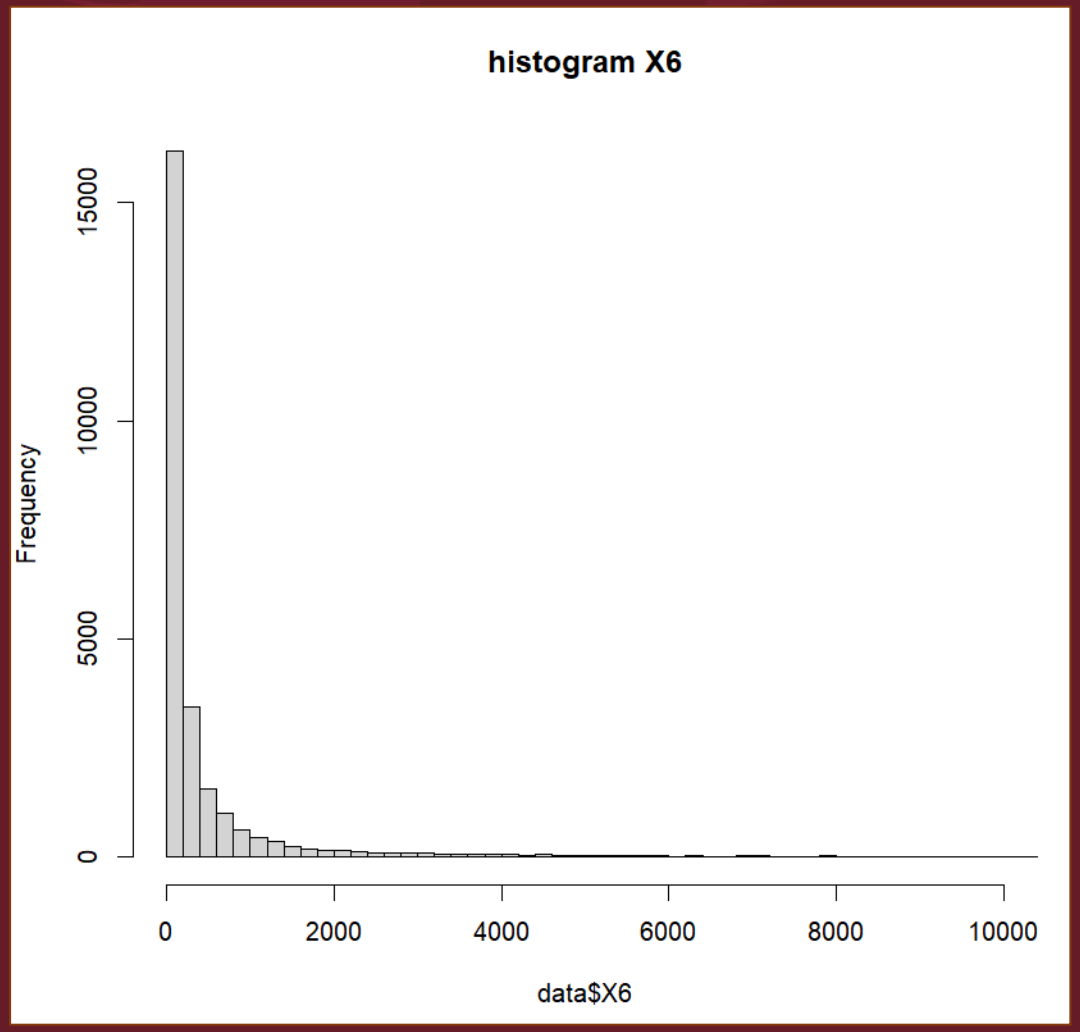


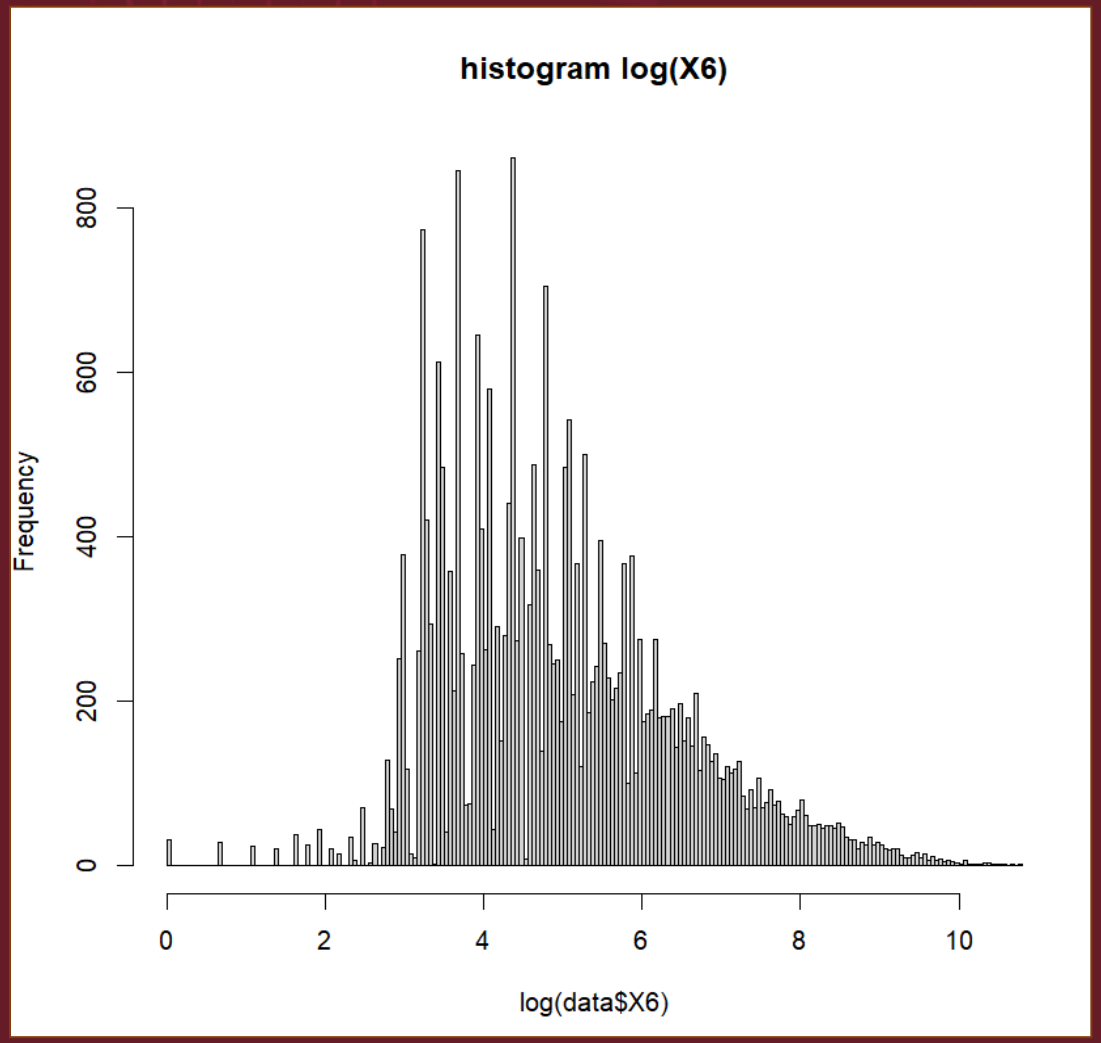
- La distribution de X2, avant et après la transformation par Log.



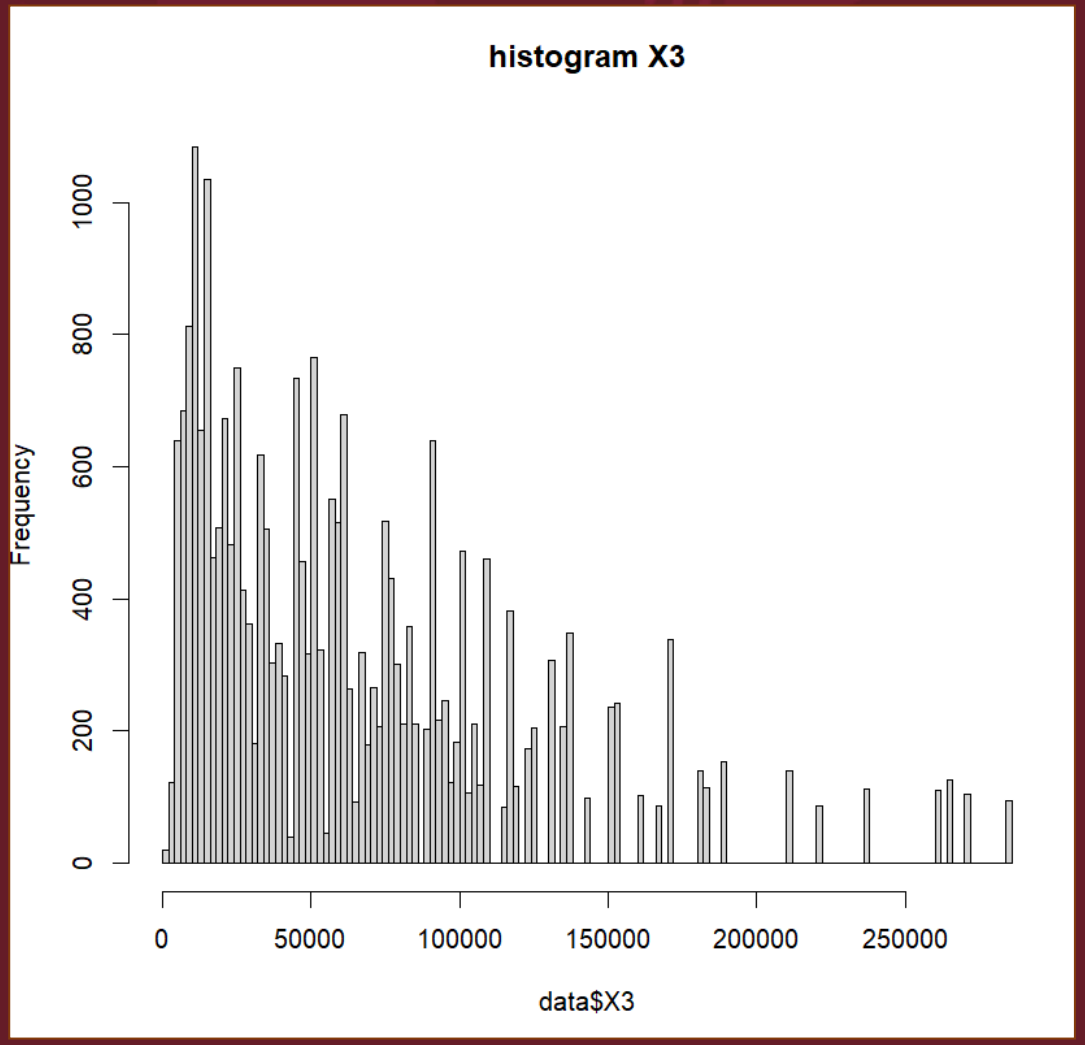


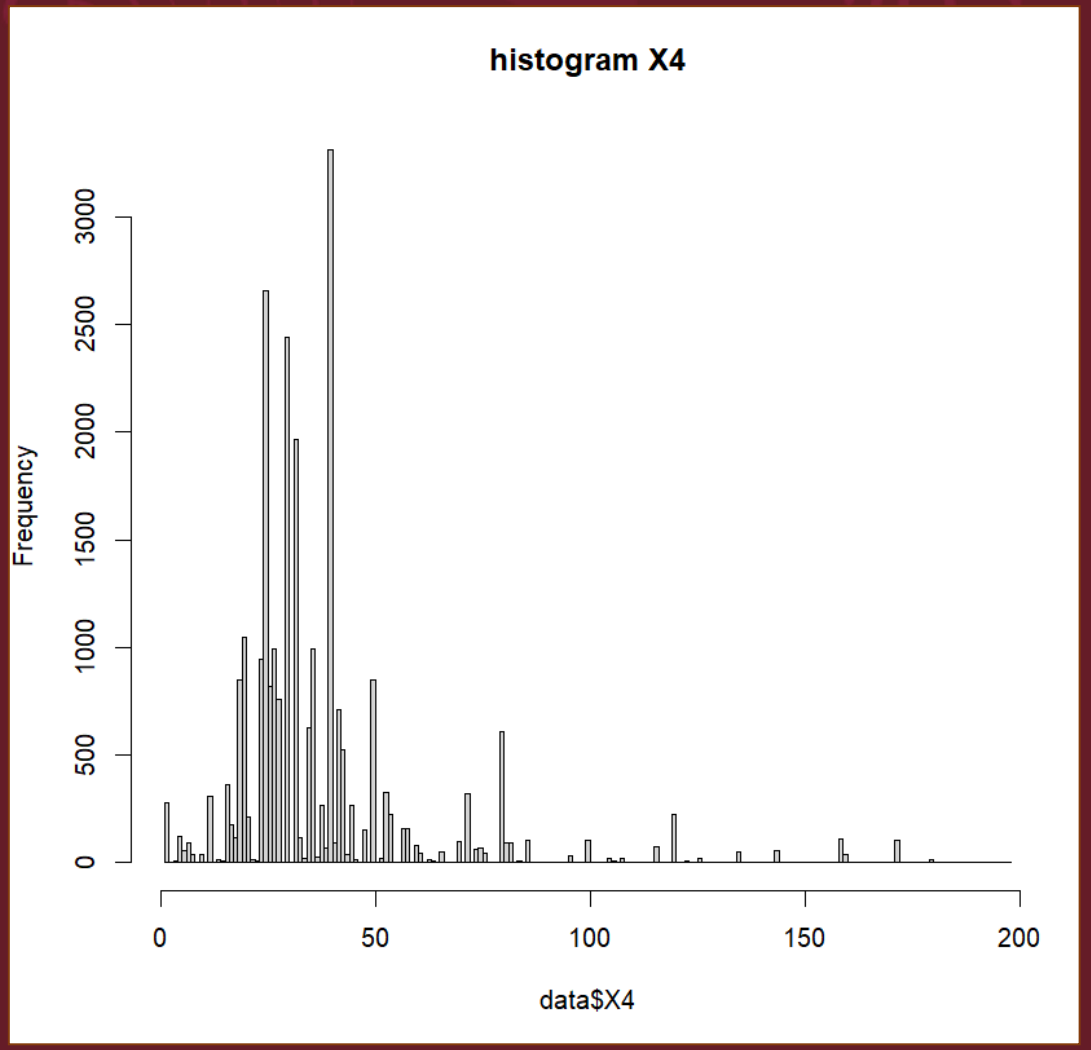
- La distribution de X6, avant et après la transformation par Log.





- La distribution de X3 et X4 sans transformation.





Construction des modèles:

Approche 1: numérisation des variables catégorielles

- Il existe plusieurs techniques de numérisation des variables catégorielles, dans notre cas on a utilisé le one hot encoding.

Construction des modèles:

- One hot encoding est une technique de codage de données qui consiste à convertir une variable catégorique en une série de variables binaires, où chaque variable représente une seule catégorie.
- Et on supprime une colonne, car elle est déduite à partir des autres.

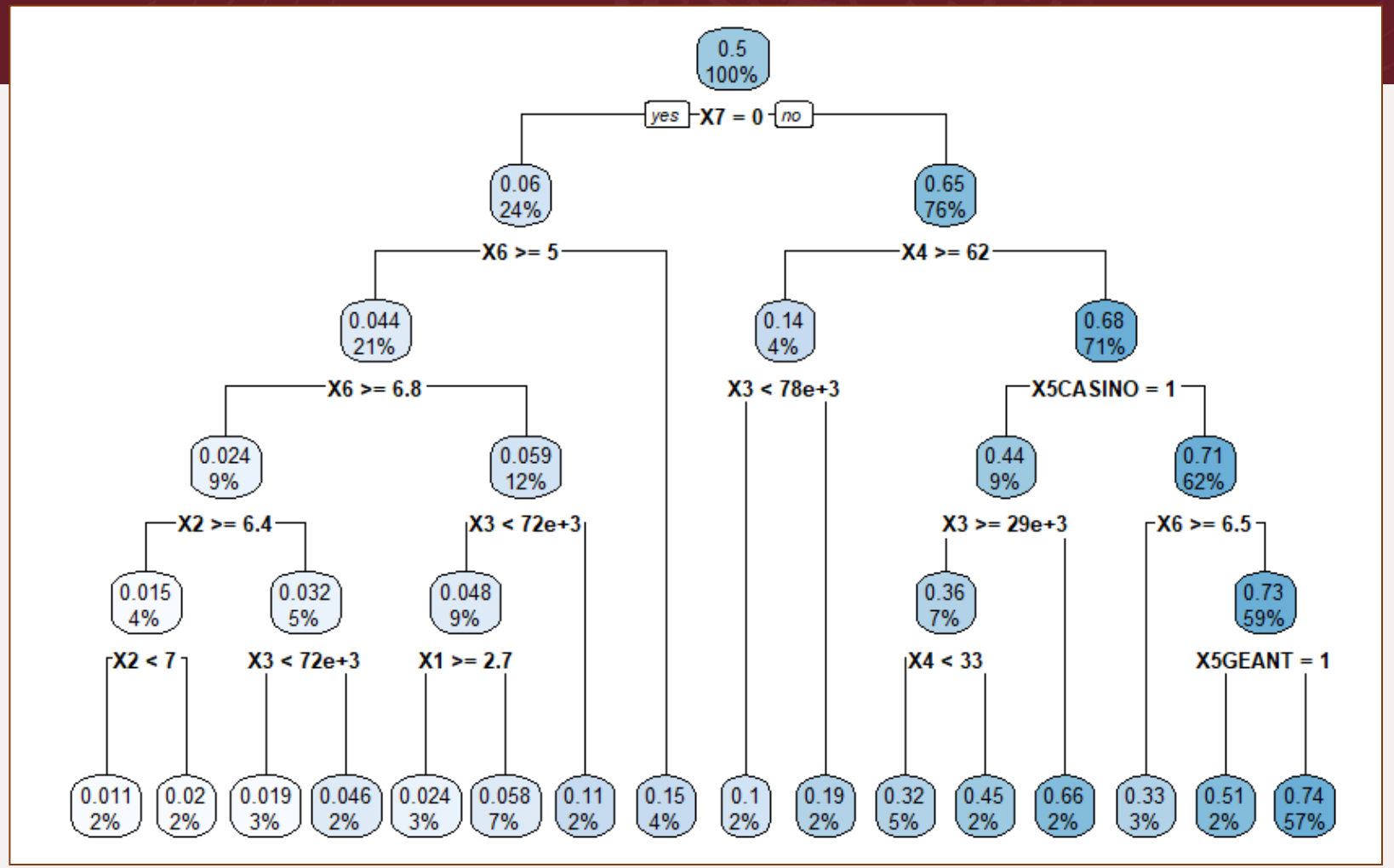
X5CARREFOUR	X5CARREFOUR.MARKET	X5CASINO	X5CORA	X5ECOMARCHE	X5FRANPRIX	X5GEANT
0	0	0	1	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
1	0	0	0	0	0	0
0	0	0	1	0	0	0
0	0	1	0	0	0	0
0	0	0	0	0	0	0
0	0	1	0	0	0	0
1	0	0	0	0	0	0
1	0	0	0	0	0	0
0	0	0	1	0	0	0
0	0	1	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Construction des modèles:

- Etape 1: définition des groupes homogènes par approche supervisée en utilisant une arbre de décision.
- Etape 2: utilisation des modèles de classification sur ces groupes supervisés.

Construction des modèles:

- **Etape 1**: construction de l' arbre de décision.



Construction des modèles:

- On ajoute une colonne « Node », chaque feuille terminal est une catégorie d'observations.

X5OTHERS	X5PRISUNIC	X5SHOPI	X5SIMPLY.MARKET	X5SUPER.U	X6	X7	Node
0	0	0	0	0	72.00	1	10
0	0	0	0	0	48.00	1	11
0	0	0	0	0	480.00	1	14
0	0	0	0	0	38.00	1	10
0	0	0	0	0	250.00	1	13
0	0	0	0	0	19.00	1	11
0	0	0	0	0	80.00	1	10
0	0	0	0	0	6.00	1	14
0	0	0	0	0	120.00	1	13
0	0	0	0	0	810.00	1	6
0	0	0	0	0	32.00	1	11
0	0	0	0	0	28.00	1	11
0	0	0	0	0	57.00	1	10
0	0	0	0	1	40.00	1	15
0	0	0	0	0	105.00	1	14

Construction des modèles:

- **Etape 2:**
- on applique le modèle sur chaque classe d'observations, et on fait une moyenne pondérée par l'effectif de chaque classe pour calculer les mesures de performances globales.

- Et on trouve les résultats suivants:

XGBoost

- accuracy is : 84.07509 %
- F1 score is : 83.0485 %
- recall is 83.11286 %
- Precision is 83.212 %

Construction des modèles:

Etape 2:

Random Forest

- accuracy is : 84.86848 %
- F1 score is : 88.89216 %
- recall is 91.34441 %
- Precision is 86.82864 %

Construction des modèles:

Etape 2:

Adaboost

- accuracy is : 83.33496 %
- F1 score is : 84.99065 %
- recall is 87.09625 %
- Precision is 86.25086 %

Comparaison des modèles

	xgboost	Random forest	adaboost
Accuracy	84.07509 %	84.86848 %	83.33496 %
F1 score	83.0485 %	88.89216 %	84.99065 %
Recall	83.11286 %	91.34441 %	87.09625 %
precision	83.212 %	86.82864 %	86.25086 %

- Les résultats obtenus par ces modèles sont très proche.
- Ces modèles ont atteint une accuracy de 84% sur le jeu de données de test.

Construction des modèles:

Approche 2: discrétisation des variables continues

- Il existe plusieurs techniques de catégorisation des variables continues, dans notre cas, on a comparé deux méthodes de discrétisation : la méthode MDLPC et la discrétisation avec arbre de décision.

Discrétisation avec MDLPC

- Cette fonction discrétise les variables continus à l'aide du critère d'entropie.
- Ce tableau donne le nombre de classes supervisées obtenues pour chaque colonne.

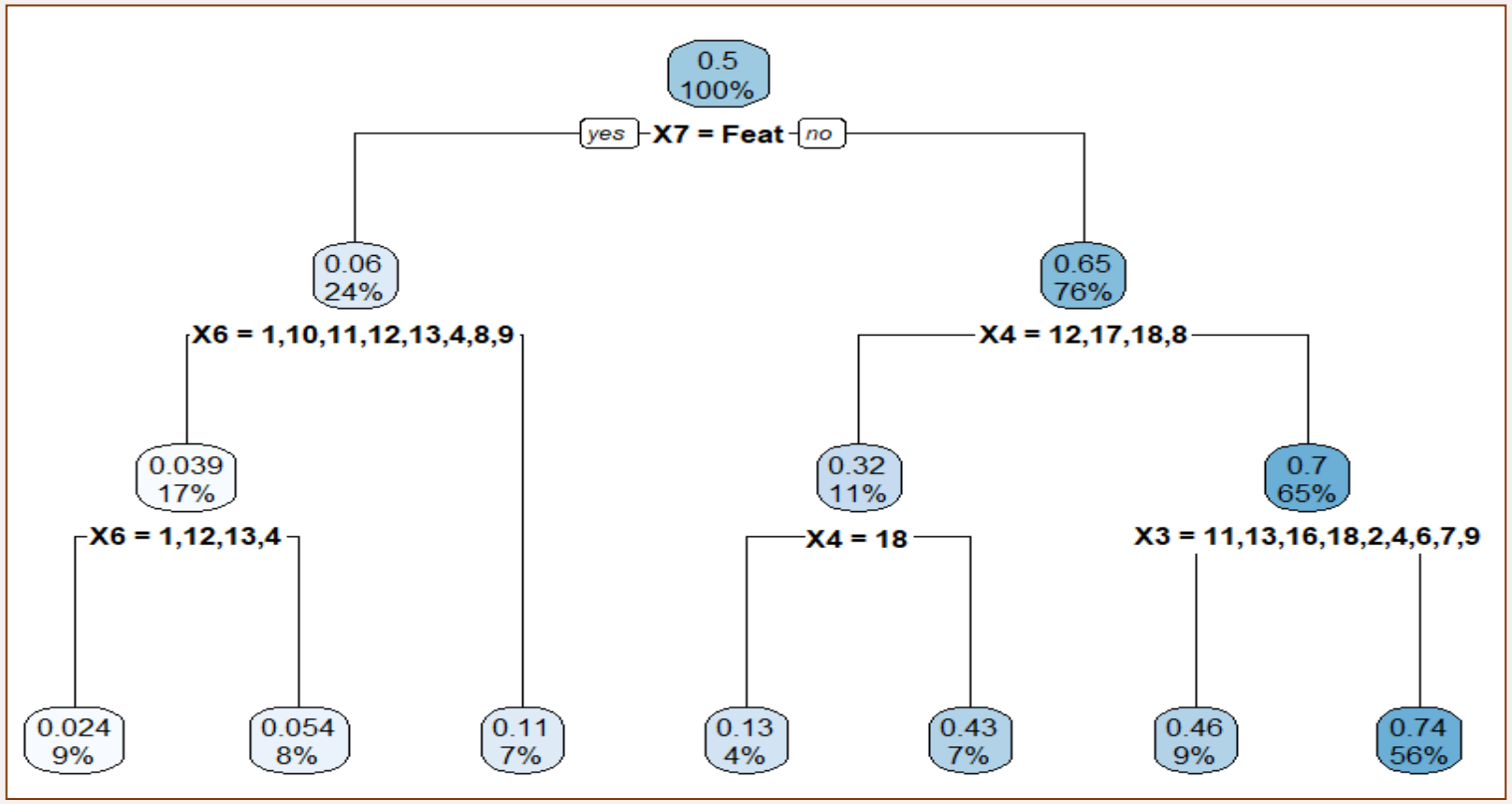
```
> length(table(data$X1))
[1] 7
> length(table(data$X2))
[1] 8
> length(table(data$X3))
[1] 19
> length(table(data$X4))
[1] 18
> length(table(data$X6))
[1] 13
```

Discrétisation avec MDLPC

- On va adopté la méthode suivante:
  - **Etape 1**: définition des groupes homogènes par approche supervisée en utilisant une arbre de décision sur la data qui contient les variables discrétisées.
  - **Etape 2**: utilisation des modèles de classification sur ces groupes supervisés.

Discrétisation avec MDLPC

- **Etape 1**



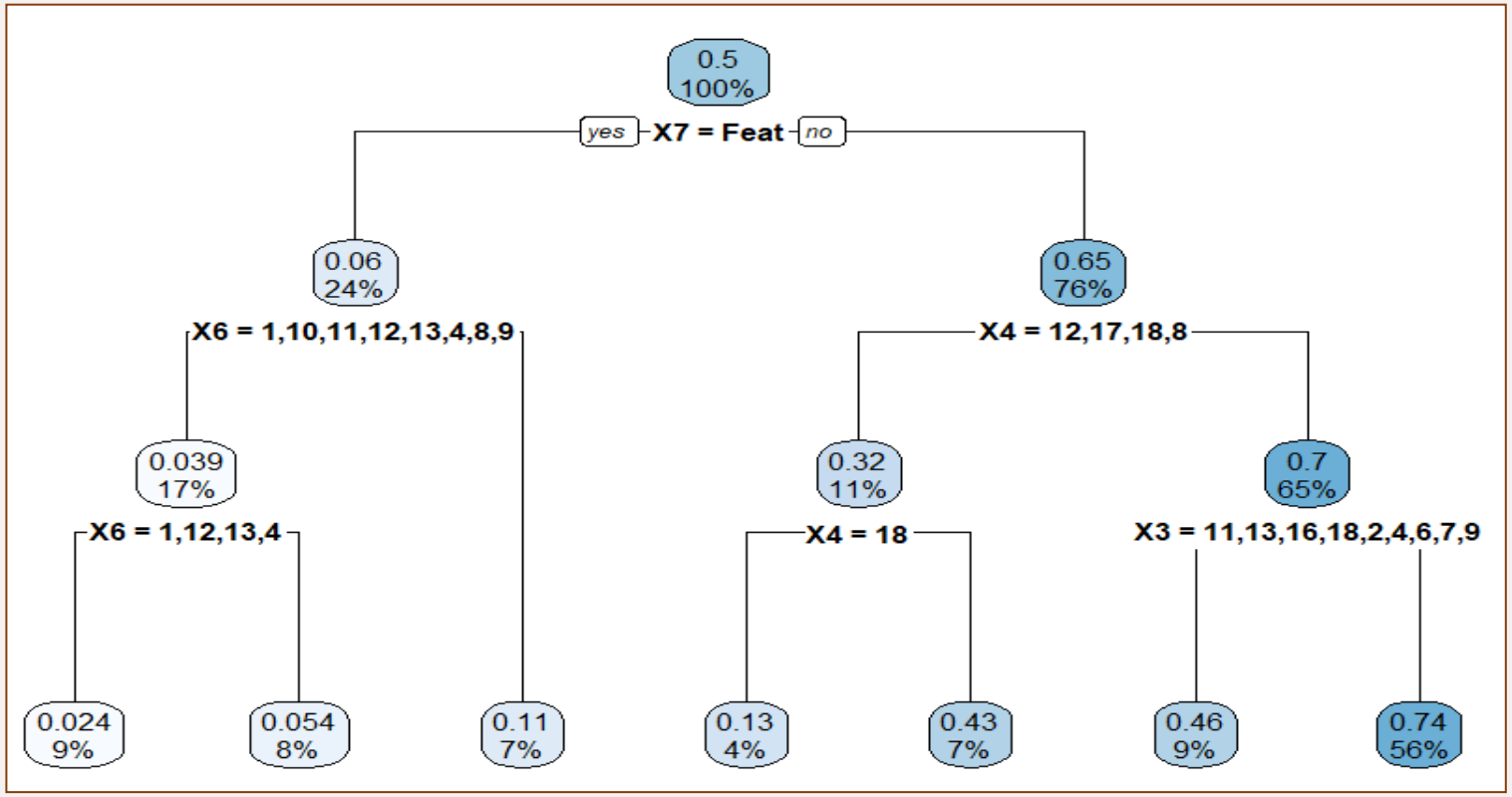
- **Etape 2**

NaiveBayes

- accuracy is : 82.42839 %
- F1 score is : 78.05997 %
- recall is 78.2898 %
- Precision is 82.28681 %

Discrétisation avec MDLPC

- **Etape 1**



- **Etape 2**

Random forrest

- accuracy is : 81.69142 %
- F1 score is : 81.07977 %
- recall is 81.50738 %
- Precision is 81.48032 %

Comparaison des modèles

	NaiveBayes	Random forest
Accuracy	82.42839 %	81.69142 %
F1 score	78.05997 %	81.07977 %
Recall	78.2898 %	81.50738 %
precision	82.28681 %	81.48032 %

- Les résultats obtenus par ces modèles sont très proche.
- Ces modèles ont atteint une accuracy de 82% sur le jeu de données de test.

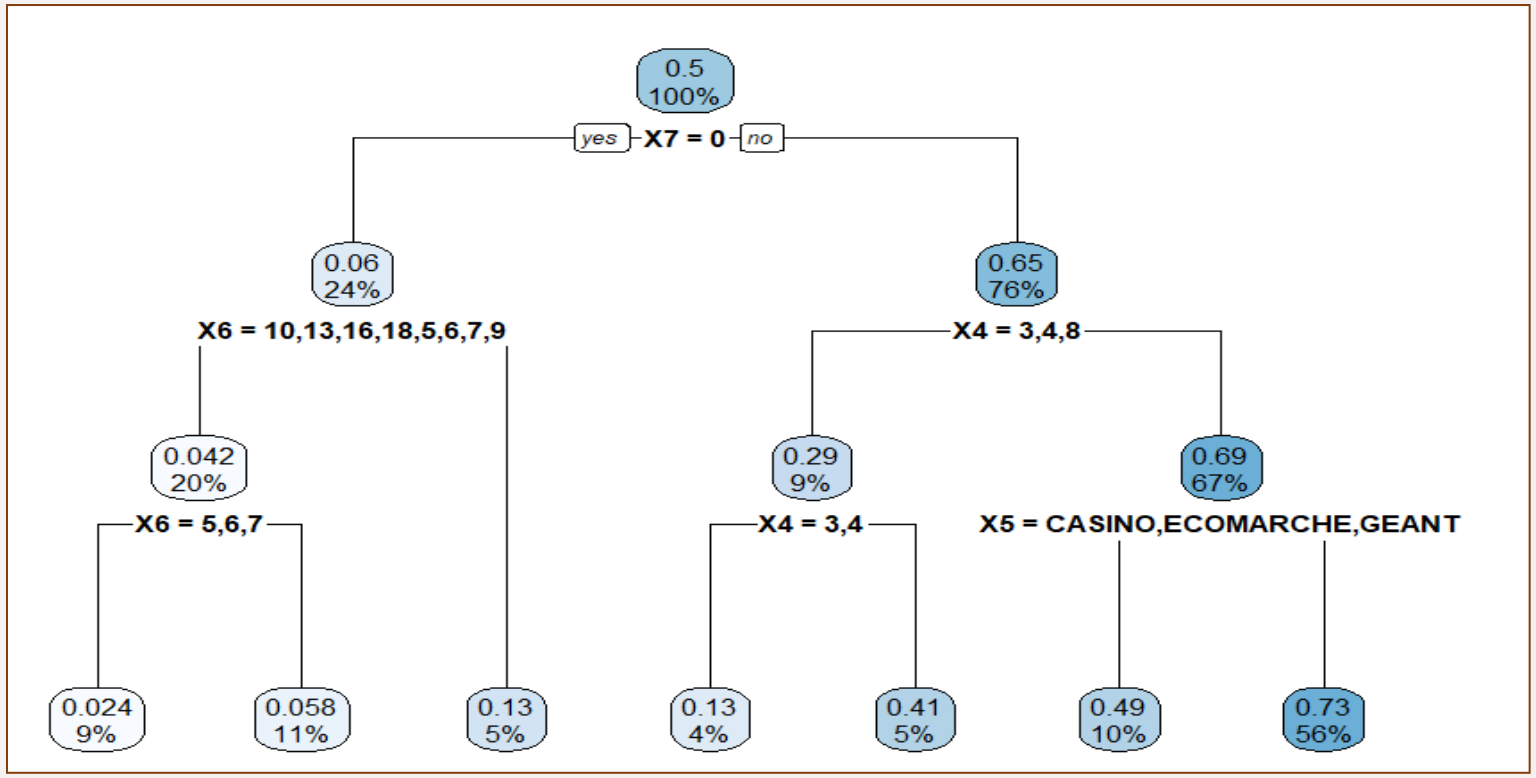
Discrétisation avec arbre de décision

- La discrétisation avec des arbres de décision consiste à utiliser un arbre de décision pour identifier les points de découpage optimaux qui détermineraient les intervalles contigus :
- **Étape 1** : Tout d'abord, on construit un arbre de décision de profondeur limitée (2, 3 ou 4) en utilisant seulement la variable que nous voulons discrétiser pour prédire la variable cible.
- **Étape 2** : Les valeurs des variables d'origine sont ensuite remplacées par l'indice de nœud terminal renvoyée par l'arbre. La classe est la même pour toutes les observations à l'intérieur d'un même Noeud terminal
- Ce tableau donne le nombre de classes supervisées obtenues pour chaque colonne.

```
> length(table(data$X1))
[1] 11
> length(table(data$X2))
[1] 14
> length(table(data$X3))
[1] 15
> length(table(data$X4))
[1] 12
> length(table(data$X6))
[1] 15
```

Discrétisation avec arbre de décision

- **Etape 1**



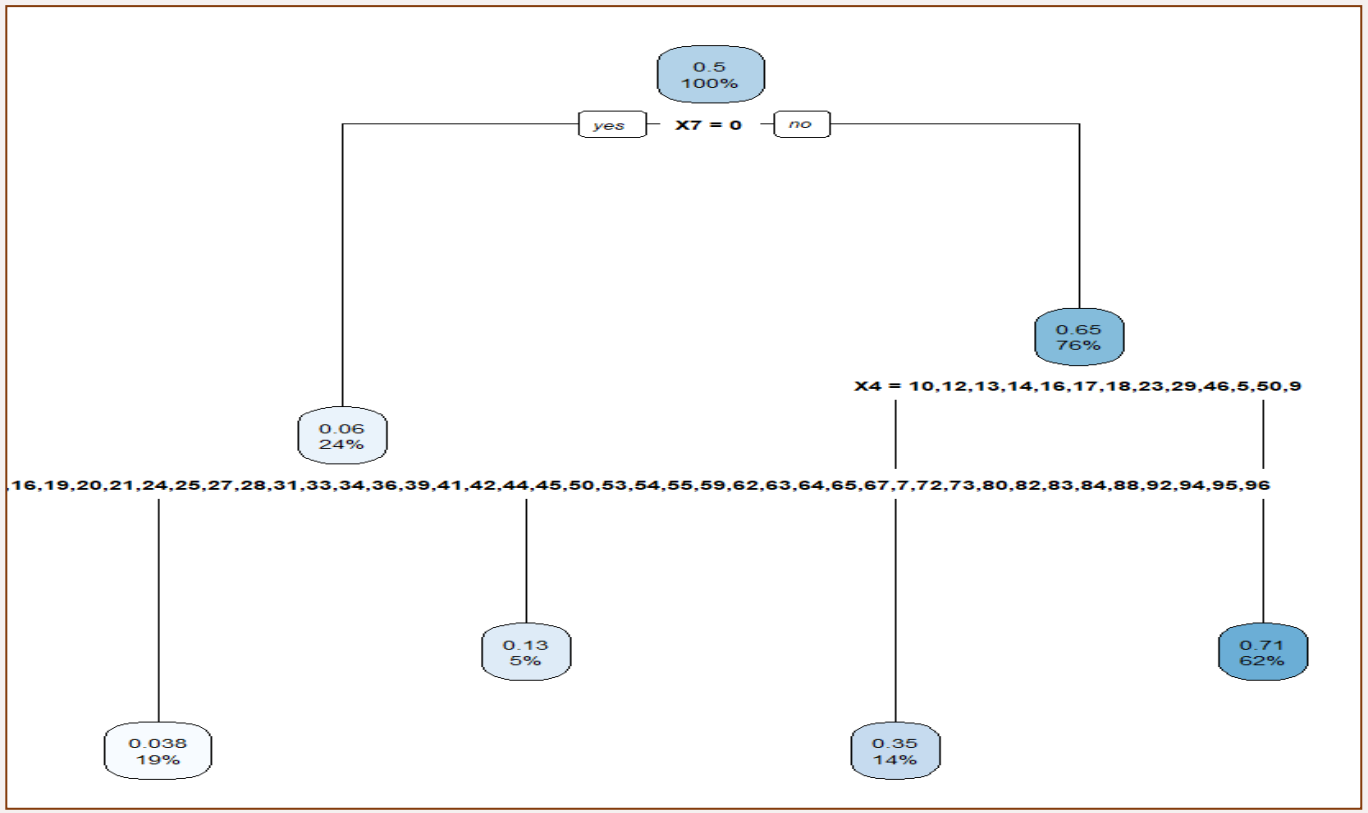
- **Etape 2**

NaiveBayes

- accuracy is : 82.55675 %
- F1 score is : 79.12049 %
- recall is 79.25106 %
- Precision is 81.8591 %

Discrétisation avec arbre de décision

- **Etape 1**



- **Etape 2**

Random forrest

- accuracy is : 83.90617 %
- F1 score is : 83.23388 %
- recall is 84.77523 %
- Precision is 82.23531 %

Comparaison des modèles

	NaiveBayes	Random forest
Accuracy	82.55675 %	83.90617 %
F1 score	79.12049 %	83.23388 %
Recall	79.25106 %	84.77523 %
precision	81.8591 %	82.23531 %

- Ces modèles ont atteint une accuracy de 83% sur le jeu de données de test.

Comparaison de discrétisation avec MDLPC et discrétisation avec arbre de décision.

- Pour tester la qualité des deux discrétisation on a appliqué un test univarié sur chaque variable discrétisée et Y.
- Le **Cramer's V** est une mesure d'association qui va de 0 à 1.
- Avec une valeur de 1 indiquant une association parfaite et une valeur de 0 indiquant aucune association. Il est calculé comme la racine carrée de la statistique du chi deux divisée par le nombre total d'observations.

	MDLPC	decision tree
X1	0.3808	0.3826
X2	0.4229	0.4291
X3	0.1642	0.1508
X4	0.3893	0.3689
X5	0.4261	0.429341
X6	0.1743	0.1743
X7	0.5049	0.5049

Conclusion

- En conclusion il en ressort que la méthode de numérisation des variables catégorielles possède les meilleur score dans les tests de performance adoptés.

Merci pour votre attention