Préparation et analyse des données d'une librairie

1

Données de 2021-03 à 2022-02

CLADIERE Nathan, Projet 4, Formation Data Analyst OC

Sommaire

- Nettoyage des données
- Analyse des datasets
  - Clients
  - Produits
  - Ventes
  - Chiffres d'affaires
- Relations caractéristiques clients/achats
- Eléments clés à retenir

Contexte et objectifs

- Premières analyses des données
- Compréhension globale des ventes de l'entreprise

- Objectifs
  - Identifier les problèmes dans les datasets
  - Caractéristiques principales (clients, produits et ventes)
  - Axes d'améliorations de l'algorithme de recommandation

Nettoyage des données

4

Nettoyage des données : Vérifications basiques

Doublons

```
produits.describe(include = "all")
```

	id_prod
count	3287
unique	3287

Valeurs Null

```
produits.isnull().sum()
```

```
id_prod     0
price       0
categ       0
dtype: int64
```

Valeurs NaN

```
produits.isna().sum()
```

```
id_prod     0
price       0
categ       0
dtype: int64
```

Nettoyage des données : valeurs aberrantes
Format des dates

Dates au mauvais format : «test »

```
ventesTest = ventes[ventes['date'].str.contains('.*test.*')]
```

	id_prod	date	session_id	client_id	is_id_prod	is_client_id
count	200	200	200	200	200	200
unique	1	39	1	2	1	1
top	T_0	test_2021-03-01 02:30:02.237413	s_0	ct_0	True	True
freq	200	13	200	106	200	200

Test pour un unique produit (T_0)

Nettoyage des données : valeurs aberrantes
Format des dates

Option 1 (choisie)

- Suppression des lignes avec « test »
  - Initialisation des caisses
  - Test des caisses

Option 2

- Reformater les dates

```
newVentes = ventes[ventes['date'].str.contains('.*test.*')== False]
```

Nettoyage des données : Valeurs manquantes Produit manquant

Comparaison des « id _prod » entre les datasets ventes et produits

```
#ajout d'une colone isin
ventes['is_id_prod']= ventes.id_prod.isin(produits["id_prod"])
#filtre colonne isn
ventesMauvaisProduit = ventes[ventes['is_id_prod']== False ]
```

id_prod	date	session_id	client_id	is_id_prod
0_2245	103	103	103	103

Nettoyage des données : valeurs manquantes
Produit manquant

Option 1 (choisie)

- Imputation du nouveau produit (moyenne des prix de sa catégorie)

Option 2

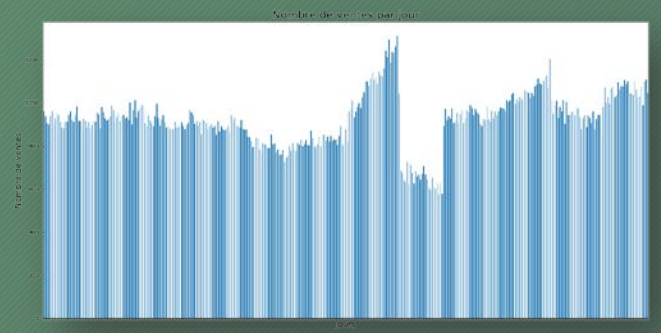
- Suppression des données

```
#Imputation du produit dans la bases de données produits (valeurs moyennes des produits de sa catégorie)
produits.loc[3288] = {'id_prod':"0_2245","price":produits[produits['categ']== 0].mean().iloc[0],"categ":0}
```

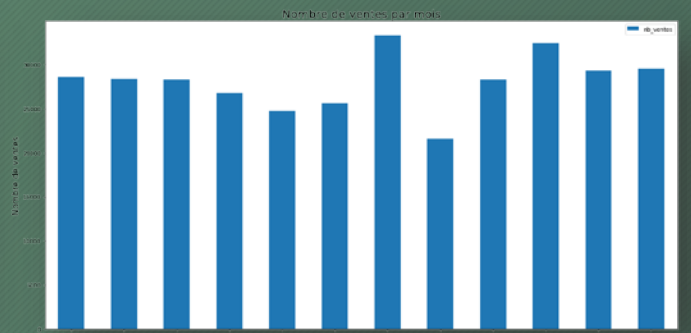
Nettoyage des données : valeurs manquantes
Continuité des ventes

Inspection en 4 étapes

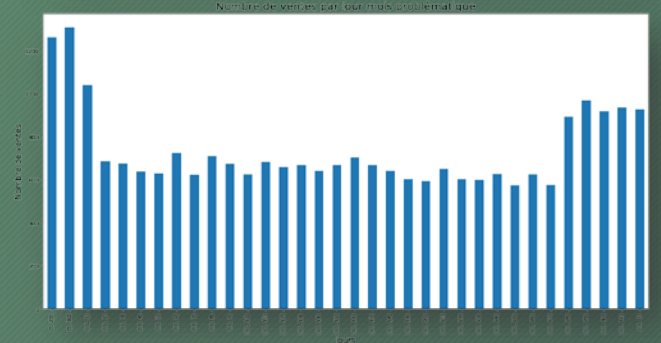
1



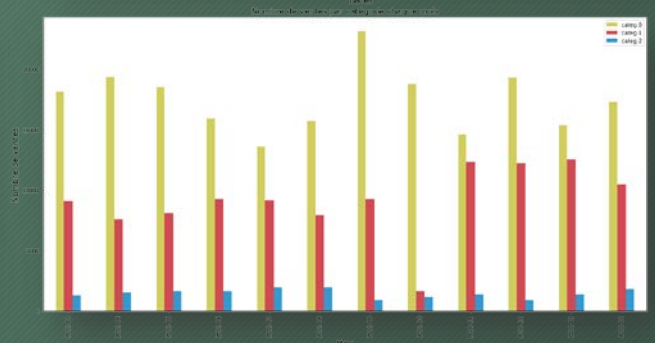
2



3



4



Nettoyage des données : valeurs manquantes
Continuité des ventes

Option 1 (choisie)

- Laisser les données telles quelles
  - Pondérer les analyses

Option 2

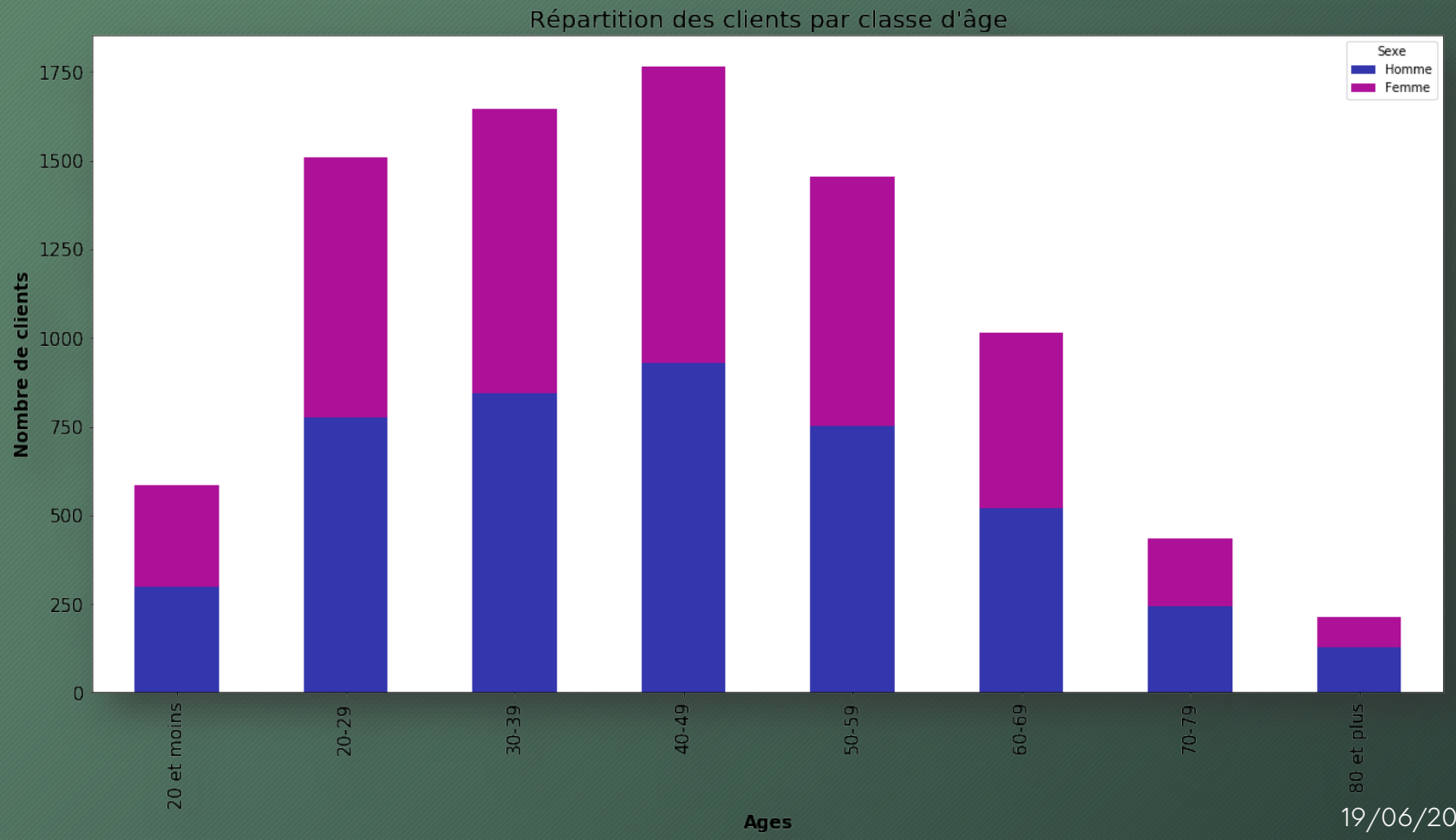
- Suppression des données

Analyse des données

16

Analyse des données : Qui sont nos clients ?

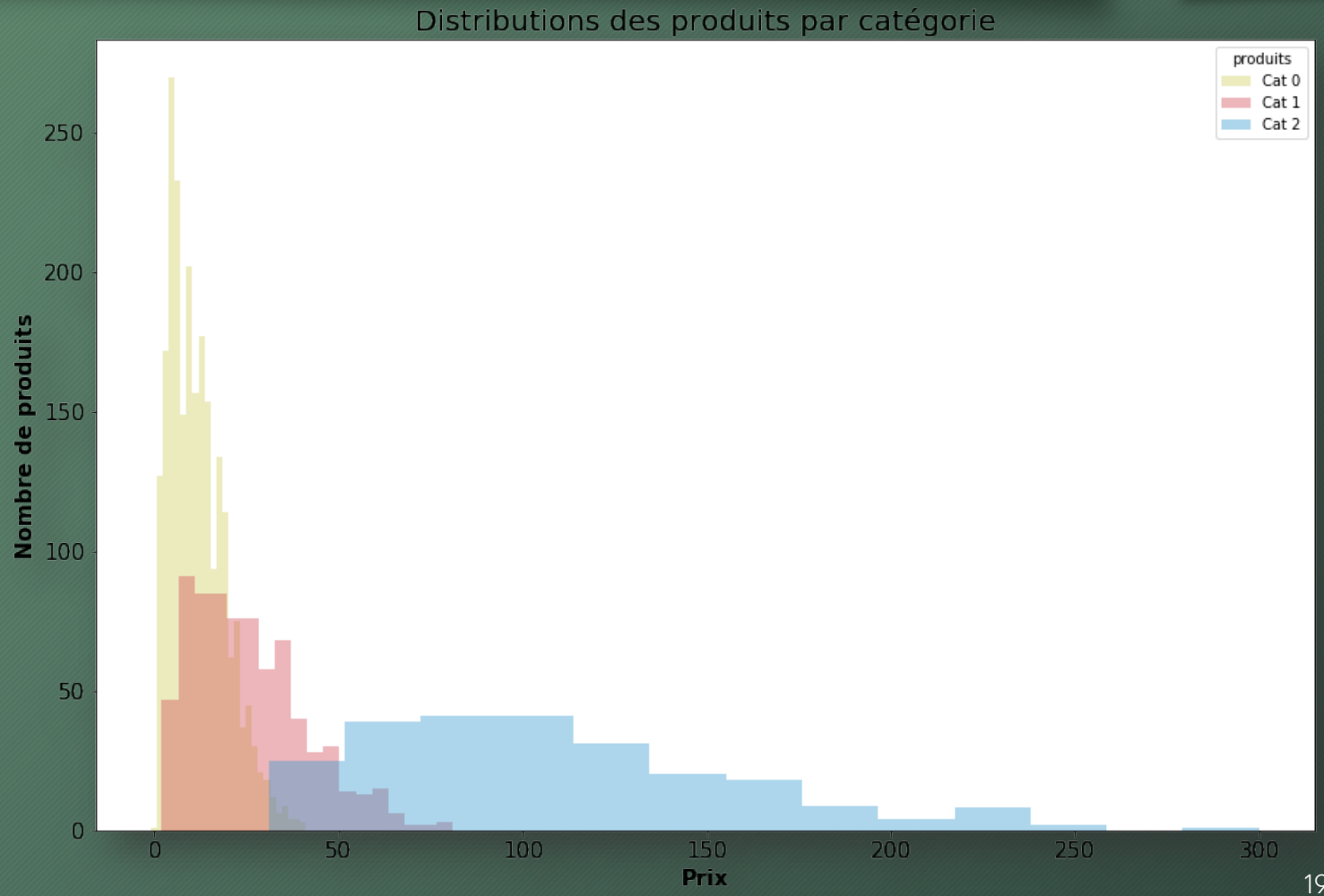
- Classe d'âge modale : 40-49 ans
- Age moyen : 43 ans
- Sexe :
  - Homme : 48 %
  - Femme : 52%



Analyse des données : Détails sur nos catégorie de produits

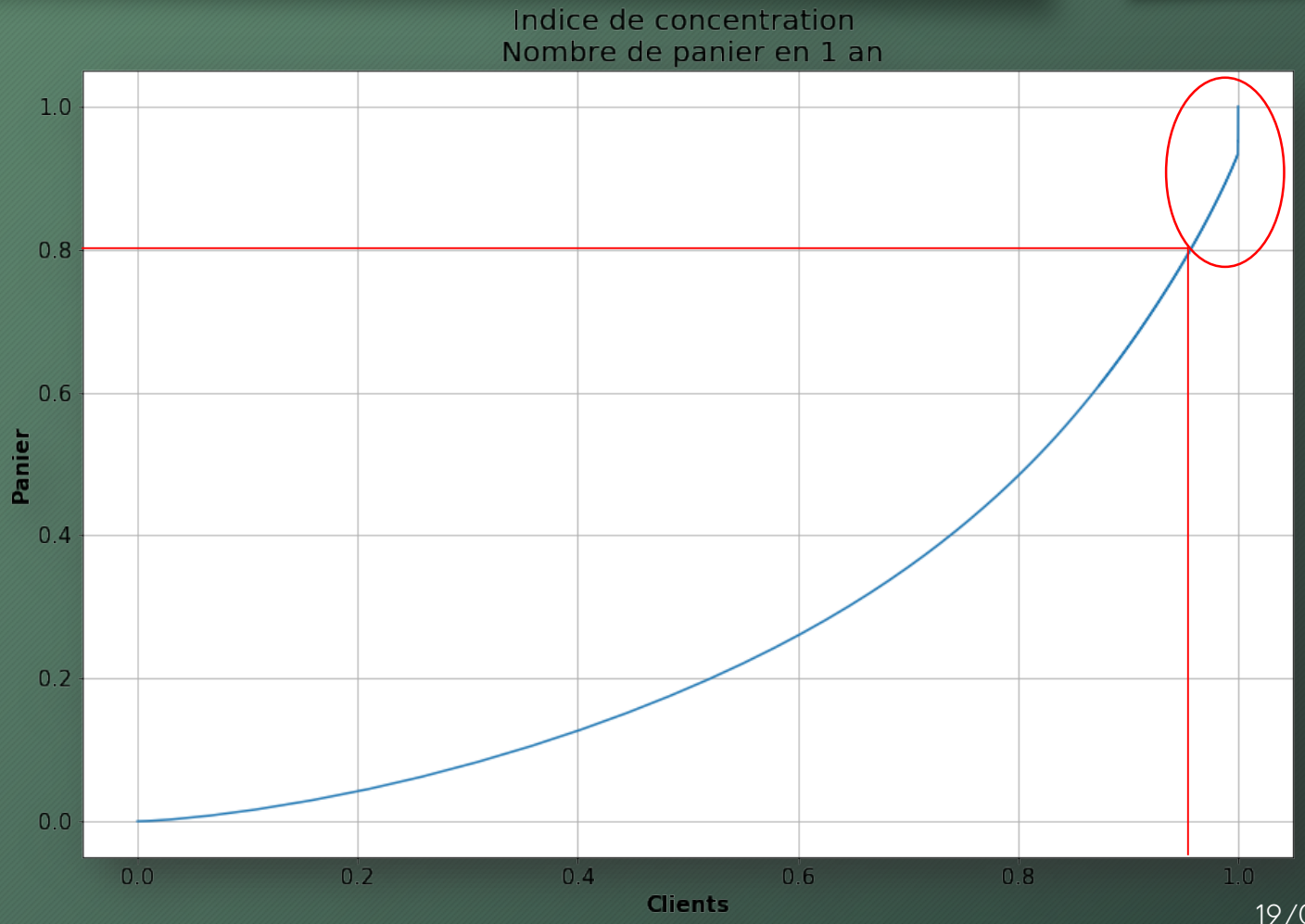
- 3 catégories distinctes

- Catégorie 0 :
  - Moy :11 €
  - Skw: 0,83
  - Fournitures et papeteries
- Catégorie 1 :
  - Moy :25 €
  - Skw: 0,81
  - Livres et magazines
- Catégorie 2 :
  - Moy :108 €
  - Skw: 0,9
  - Beaux livres et high-tech



Analyse des données : Répartition inégale des ventes

- Indice de gini : 0,47
  - Certains clients achètent beaucoup plus que d'autres
- Interprétation Lorenz:
  - 20% des paniers sont dues à 5 % des clients



Analyse des données : Répartition inégale des ventes

Qui sont-ils ?

- Entreprises
- Comptes internes

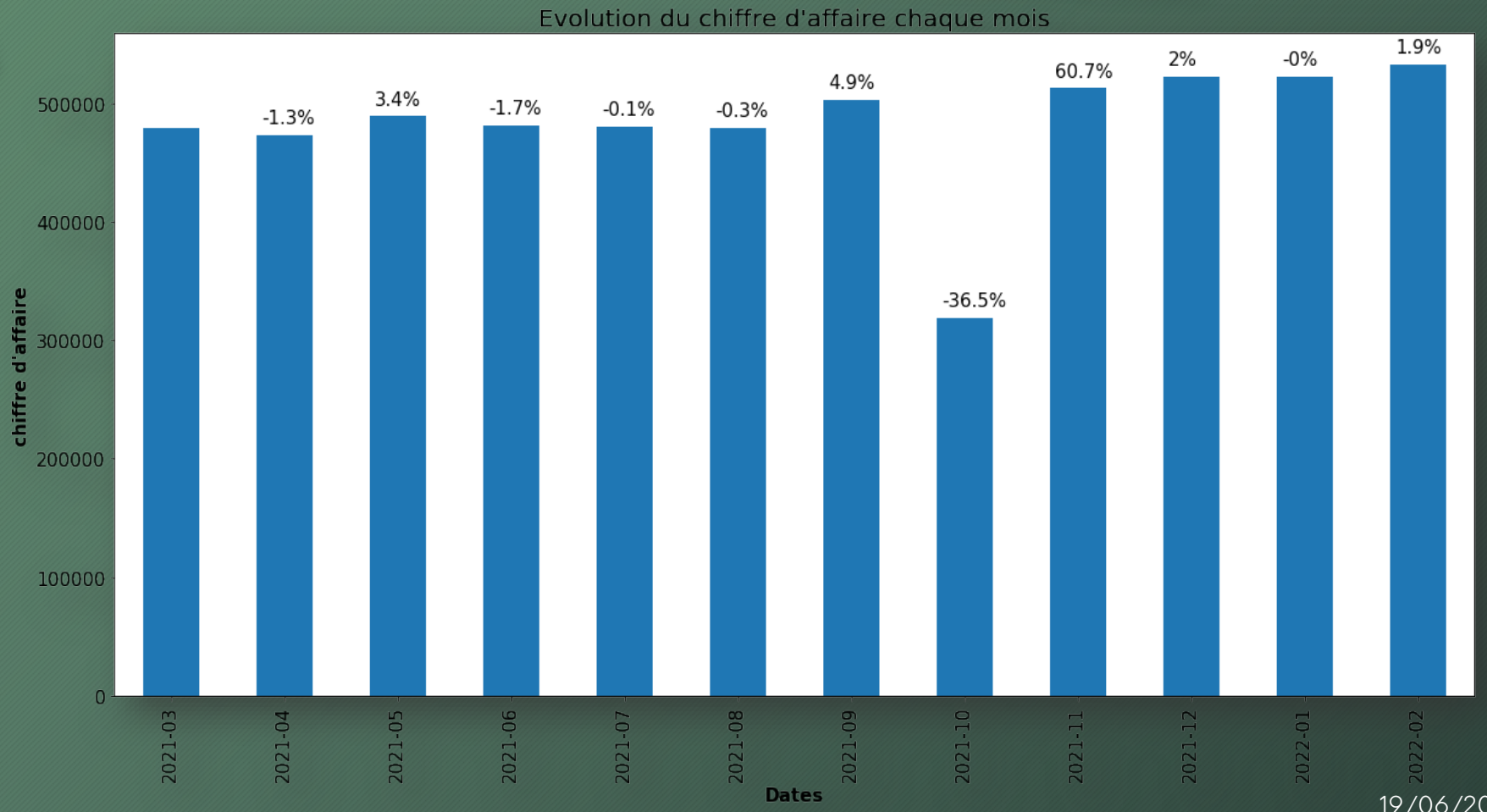
Clients	Nombre de paniers
c_1609	5501
c_3454	2711
c_4958	1888
c_6714	1286
c_682	84
c_8392	79

Chiffres avec et sans ces clients

- Sans :
  - Moyenne : 18
  - Variance : 230
  - Ecart-type : 15
- Avec
  - Moyenne : 20
  - Variance : 5160
  - Ecart-type : 71

Analyse des données : Chiffre d'affaire, croissance régulière

- Problème vu dans le nettoyage des données
- Hausse régulière malgré quelques stagnations



Relations caractéristiques clients/achats

22

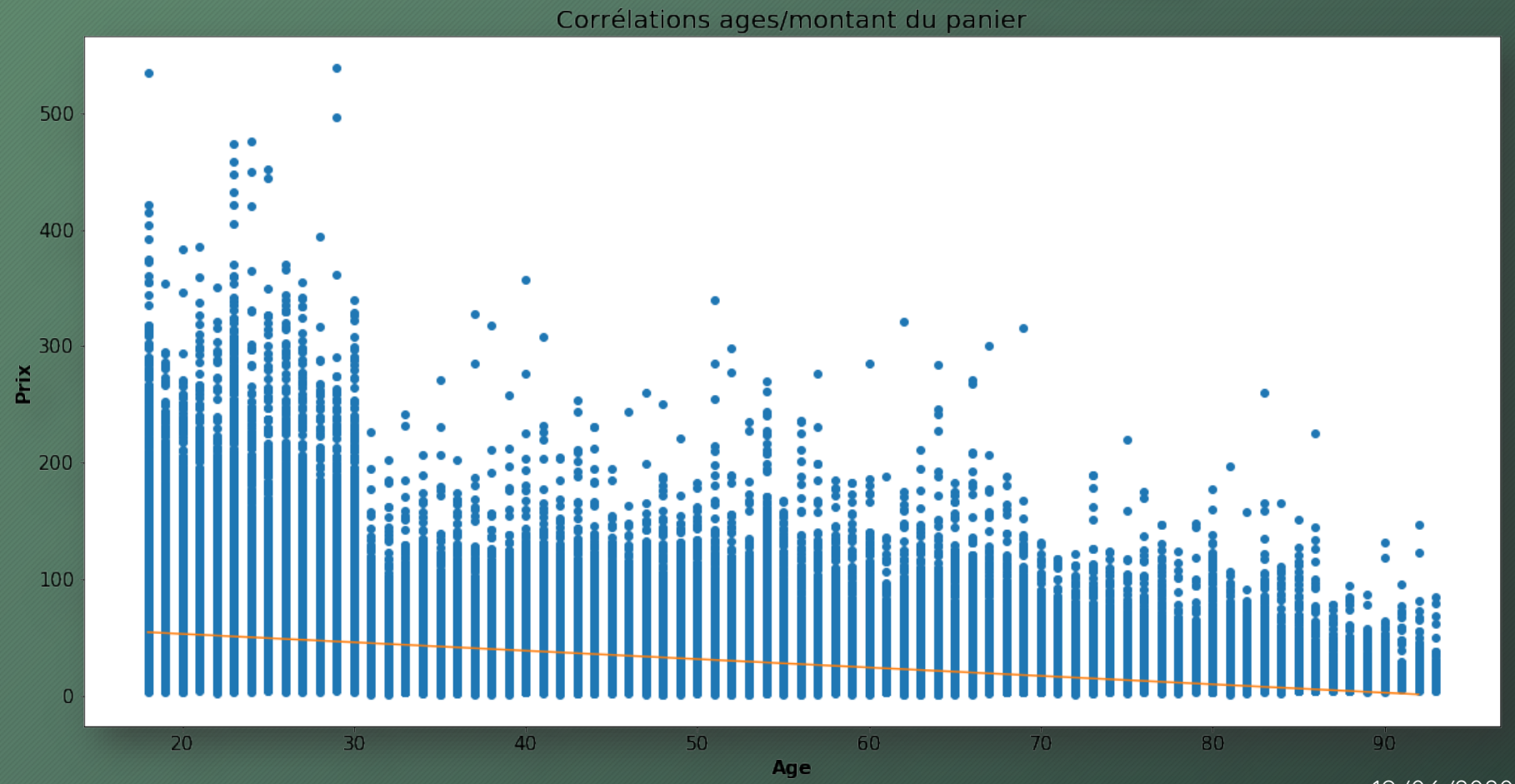
Corrélations

Le sexe n’influe par sur la catégorie de produit achetée

- Pas de différence entre les sexes
- Catégorie 0 significativement plus achetée dans les deux cas

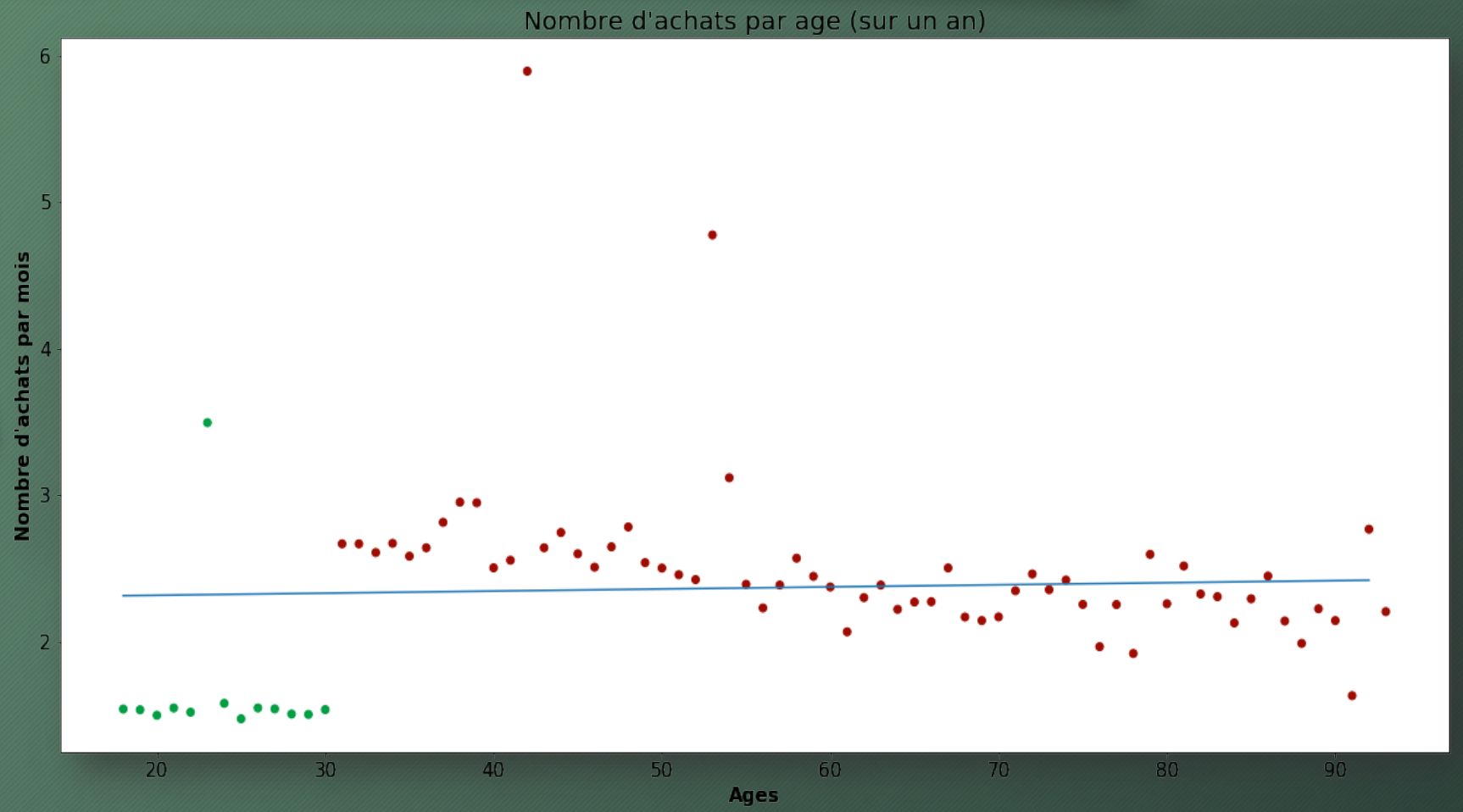
Les plus jeunes ont des panier plus élevés

- Coefficient de pearson :-0,3
- Très marqué jusqu'à 30 ans



Fréquence des achats:
2 catégories de personnes

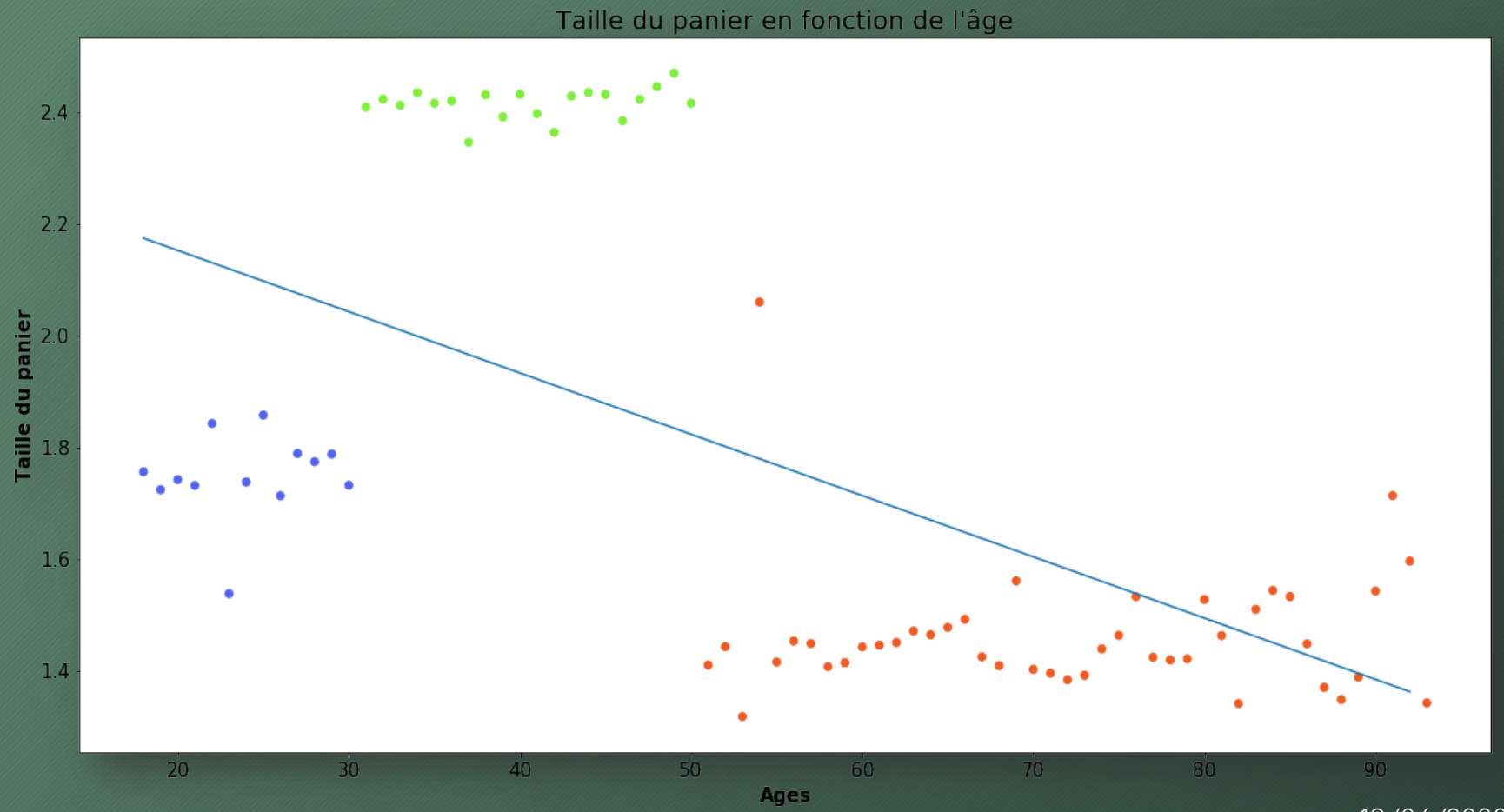
- Moins de 30 ans,
  Fq achat moyen : 1,6
- Plus de 30 ans,
  Fq achat moyen : 2,4



Taille du panier qui diminue fortement pour les plus de 50 ans

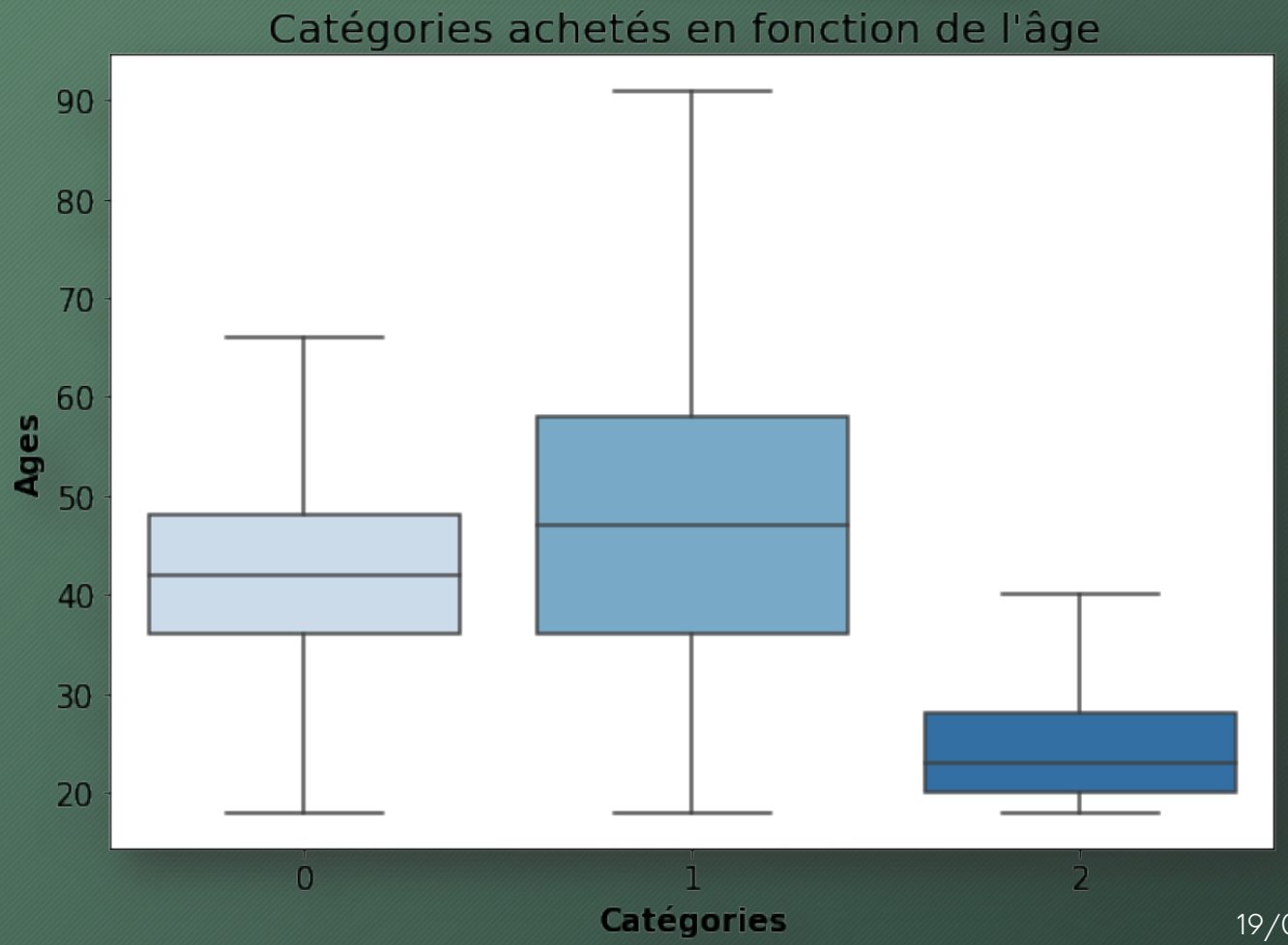
3 catégories de personnes :

- Moins de 30ans, Taille du panier moyen : 1,7
- 30 à 50 ans, Taille du panier moyen : 2,4
- 50 ans et plus, Taille du panier moyen :1,4



La catégorie de produit acheté varie en fonction de l'Age

- $\eta^2 = 0{,}28$
- Tendance
  - Catégorie 2 acheté par les plus jeunes
  - Catégorie 1 et 0 pour tout le monde



Conclusion

- Identifier le problème dans la continuité des ventes
- Identifier nos clients importants
- Classes modales clients 40-49 ans (et 30-39 ans)
  - Taille du panier élevés
  - Produits de catégorie 1
  - Fréquence achats : 2,4 par mois
- Les moins de 30 ans
  - Panier élevés
  - Produits de catégories 2