PROJET N° 4 : Etude de marché

Analyse des ventes d'une entreprise

Ines Ebilitigué

Mission 1: Nettoyage des données

Nettoyage de la **table « customers »**

Présentation de la table « customers »

Colonnes	Nombre de lignes	Type
Client_id	8623	objet
sex	8623	objet
birth	8623	int64

Pas de valeurs nulles ni manquantes

Grâce à un tri croissant de la colonne client_id :

```
customers.sort_values(["client_id"], axis=0, ascending=False, inplace=True)
```

Constat de la présence de valeurs aberrantes: ct_0 et ct_1

Suppression de ces dernières :

```
customers.drop(customers.loc[customers['client_id']=='ct_1'].index, inplace=True)
```

Création d'une colonne âge à partir de la colonne birth et la librairie « datetime »

Tendances centrales

	birth
Moyenne	1978
Ecart type	16
min	1929
25%	1966
50%	1979
75%	1992
max	2004

Nettoyage de la table products

Présentation de la table « products »

Colonnes	Nombre de lignes	Type
id_prod	3287	objet
price	3287	float
categ	3287	int64

Recherche de valeurs nulles, manquantes aberrantes

Présence de prix inférieur à 0 (en lien avec le produit T_0 et la table transactions).
Retrait des valeurs <=0 et vérification Total des lignes 3786 pour chaque colonne
Pas de valeurs nulles, manquantes ni aberrantes

Tendances

	price
Nb de lignes	3287
Moyenne	21
Ecart type	29
min	0.62
25%	6.99
50%	13.06
75%	22.99
max	300

Nettoyage de la table « transactions »

Présentation de la table transactions

Colonnes	Nombre de lignes	Type
Id_prod	337016	objet
date	337016	objet
Session_id	337016	objet
Client_id	337016	objet

Vérification des valeurs nulles, manquantes, aberrantes et du format

Pas de valeurs nulle ni manquantes **Présence de 4 clients ("c_1609", "c_6714" , "c_3454", "c_4958") avec un CA élevé** **Présence d'id_prod aberrant: T_0 et 0_2245 (103 fois).**
Retrait des clients avec un CA trop élevé, car trop important dans le CA global, soit un peu plus de 7% de ce dernier
Suppression des client T_0 Remplacement de l'id_prod 0_2245 par une estimation de la moyenne des prix des produits de la catégorie associée (la catégorie est 0 et la moyenne des prix est de 12€)
Modification du type de la variable date

Mission 2: analyse des ventes

Mission 2: analyse des ventes

- **Création d'une table « ventes » via 2 jointure des tables transactions et products, via id_prod (en inner), puis entre ventes et customers via la clé client_id (en inner).**

- **Présentattion de la table ventes**

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 313651 entries, 0 to 313650
Data columns (total 12 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   id_prod     313651 non-null  object
 1   date        313651 non-null  datetime64[ns]
 2   session_id  313651 non-null  object
 3   client_id   313651 non-null  object
 4   date1       313651 non-null  object
 5   date_year   313651 non-null  float64
 6   date_month  313651 non-null  float64
 7   price       313651 non-null  float64
 8   categ       313651 non-null  int64
 9   sex         313651 non-null  object
 10  birth       313651 non-null  int64
 11  age         313651 non-null  int64
dtypes: datetime64[ns](1), float64(3), int64(3), object(5)
memory usage: 31.1+ MB
```

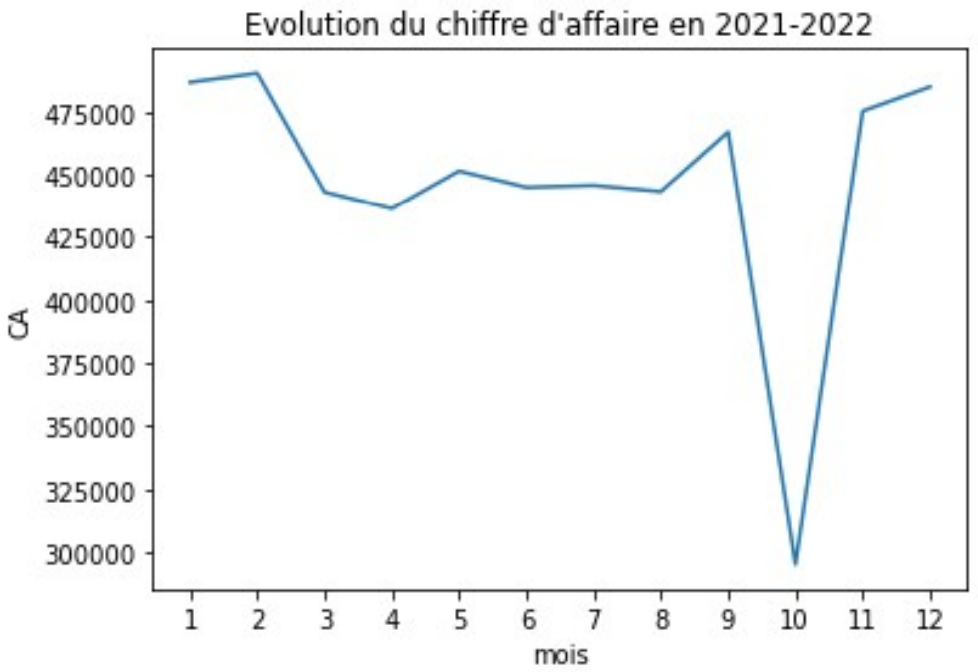
Mission 2: analyse des ventes (suite 1)

- **Chiffre d'affaire annuel (CA)**

Année	CA
2021	4.386997e+06
2022	9.768639e+05

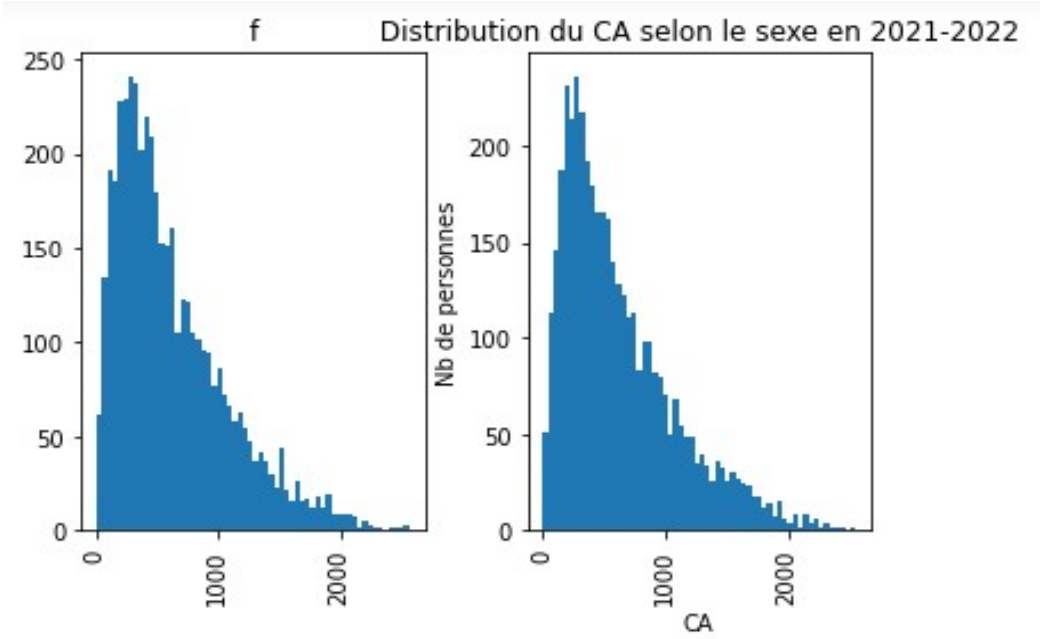
Chute du CA en octobre 2021 en raison de faible quantité de produits de catégorie 2 et de valeurs manquantes dans la catégorie 1

Evolution du CA



Mission 2: analyse des ventes (suite 2)

Distribution du CA selon le sexe en 2021

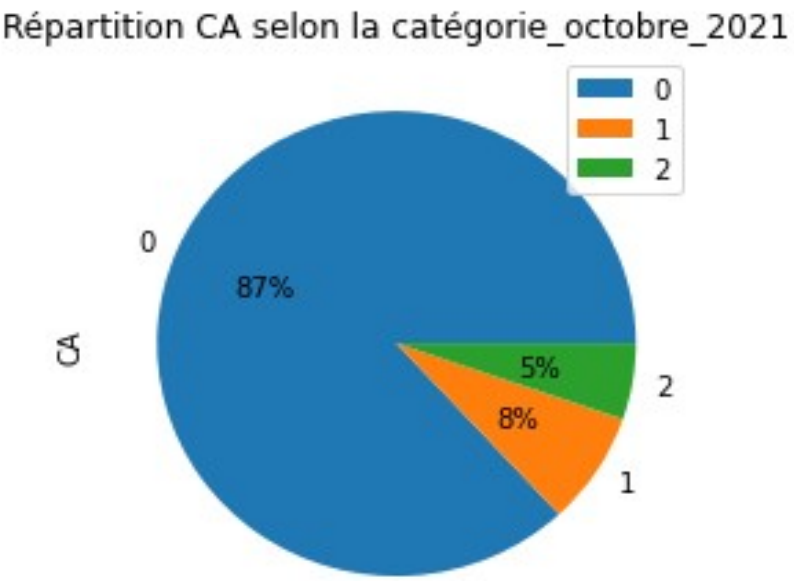


La répartition du CA par sexe révèle une faible différence de CA entre femmes et hommes.

Mission 2: analyse des ventes (suite 3)

Répartition du CA selon la catégories de produit

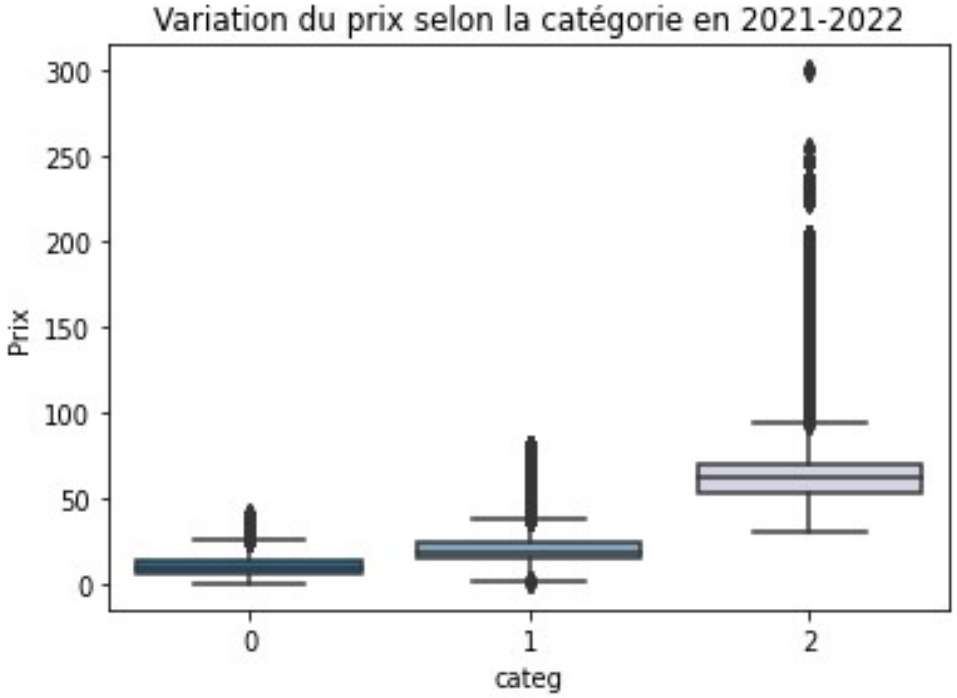
Il y a trois catégories de produits allant de 0, 1 et 2



Le graphique en secteur ici présenté révèle une inégale distribution du CA par catégorie. La catégorie 2 ne contribue au CA qu'à 5% alors que la catégorie 0 concentre plus de 80% du CA.

Mission 2: analyse des ventes (suite 4)

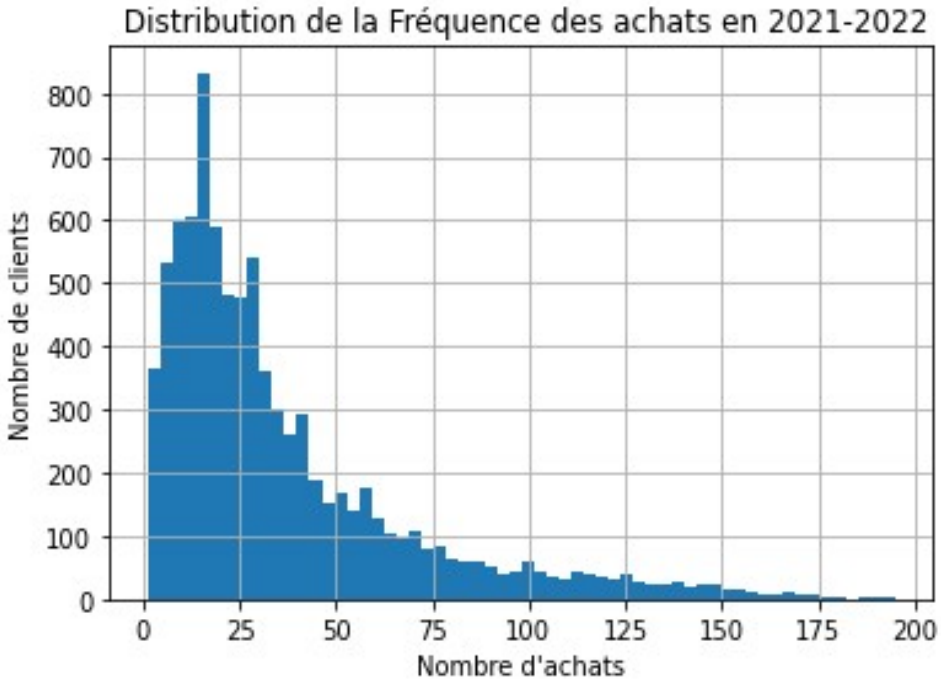
- **Variation du prix par catégorie de produit**



Le lien entre la catégorie de produit et le prix est manifeste. Plus la catégorie est élevée, plus le prix l'est .

Mission 2: analyse des ventes (suite 5)

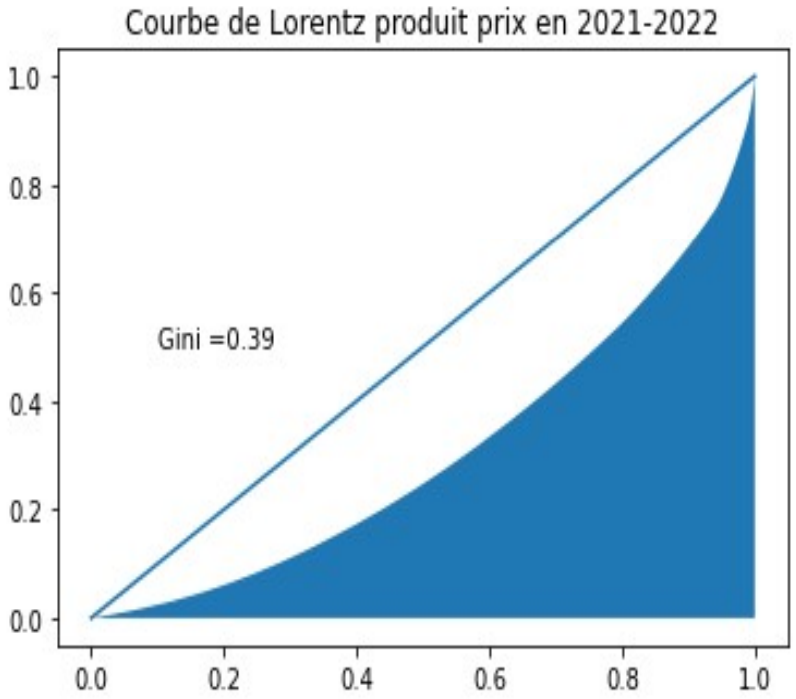
➢ Distribution de la fréquence des achats



La répartition de la fréquence des achats est asymétrique. Une majorité de clients consomment peu. Le nombre de clients décroit lorsque celui des achats augmente.

Mission 2: analyse des ventes (suite 6)

- **Relation entre l'âge et le prix du produit**



La courbe de Lorentz révèle une inégale répartition des prix entre les produits. L'indice de Gini, une mesure statistique permettant de rendre compte de la répartition d'une variable au sein d'une population (ici la répartition des prix sur les produis) indique une faible inégalité des prix entre produits, car il varie entre 0 (égalité parfaite) et 1 (inégalité extrême). Entre 0 et 1, l'inégalité est d'autant plus forte que l'indice de Gini est élevé

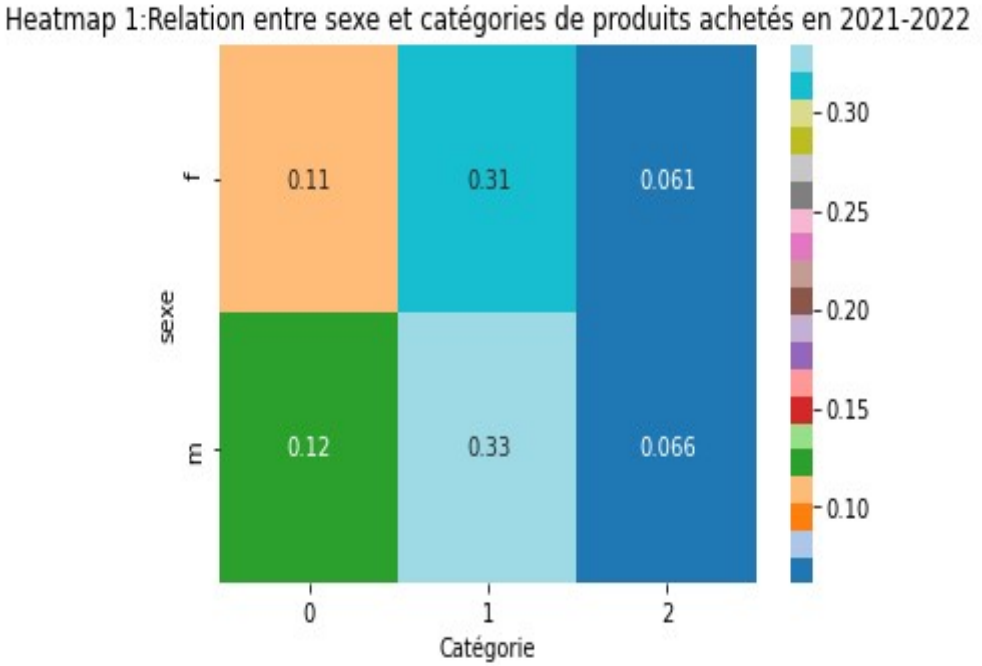
Mission 3: analyse des ventes

Mission 3: analyse des ventes

➢ Corrélation entre le sexe des clients et les catégories de produits achetés

Création d'une table de contingence
Création de la matrice "valeurs attendues«
Produit matriciel. On utilise pd.T pour pivoter la table.
Représentation graphique d'une carte des chaleur, pratique pour observer le comportement de clients

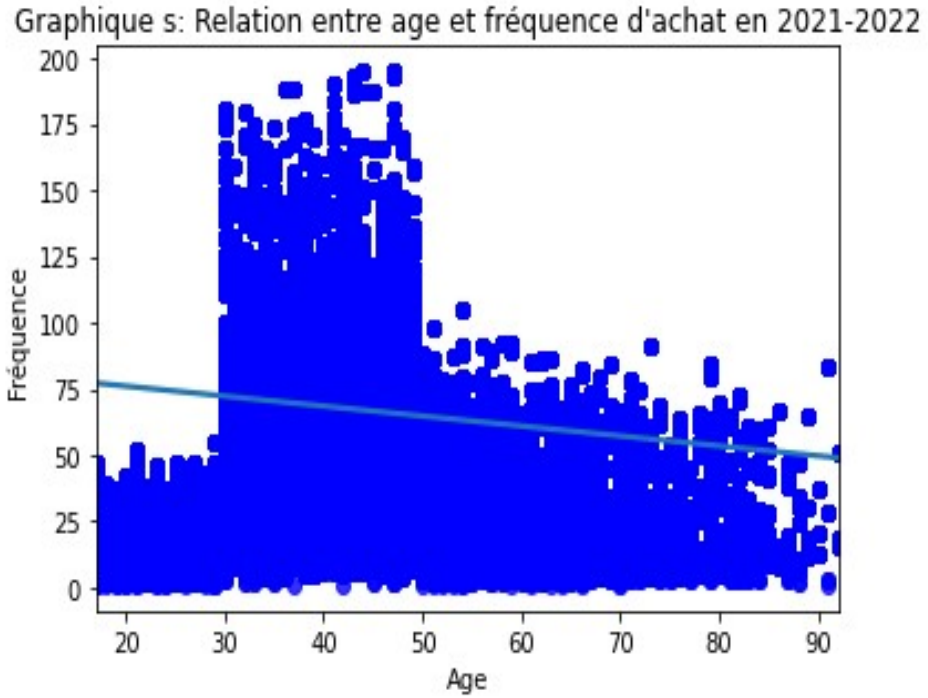
Le khi2 est de 10.110865513726571.
Degré de liberté à préciser le degré de liberté est égal à (3-1)*(2-1), soit 2
La marge d'erreur est de 0,05 qui renvoie à 5,991. Il y a une corrélation entre les variables sexe et catégorie.



Mission 3: analyse des ventes (suite 1)

➢ Relation entre l'âge des clients et la fréquence d'achat

- création table âge
- création table frequence des achats: freq
- Jointure des deux tables
- Représentation graphique de la relation âge et fréquence des achats

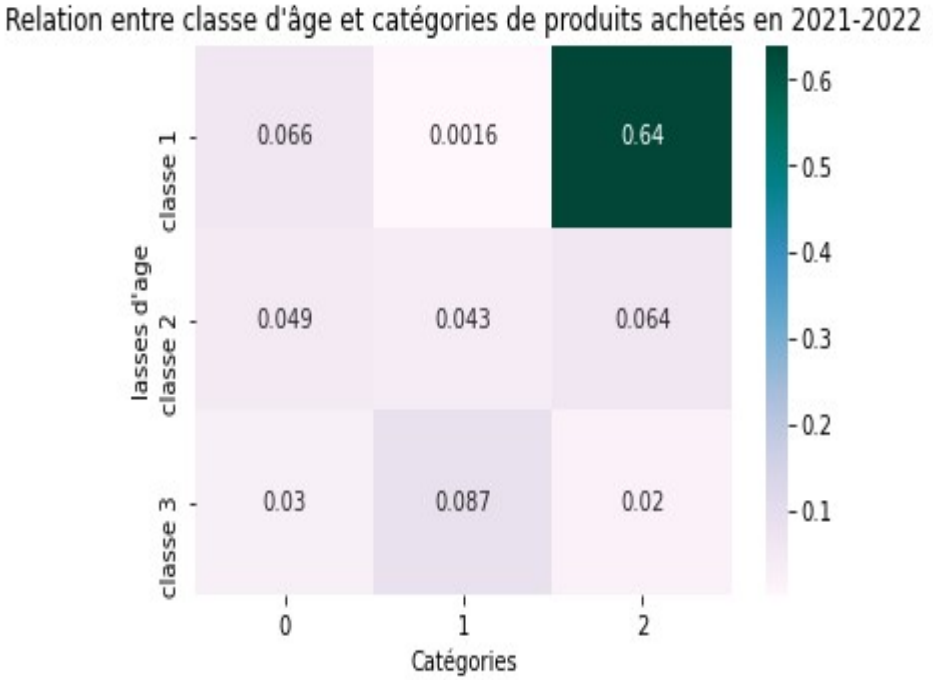


le coefficient de Pearson est négatif, soit -0.11608681317122682 .

il indique une faible relation linéaire entre les variables âge et fréquence. Cette interprétation est confortée par la P-value < 0, 05.

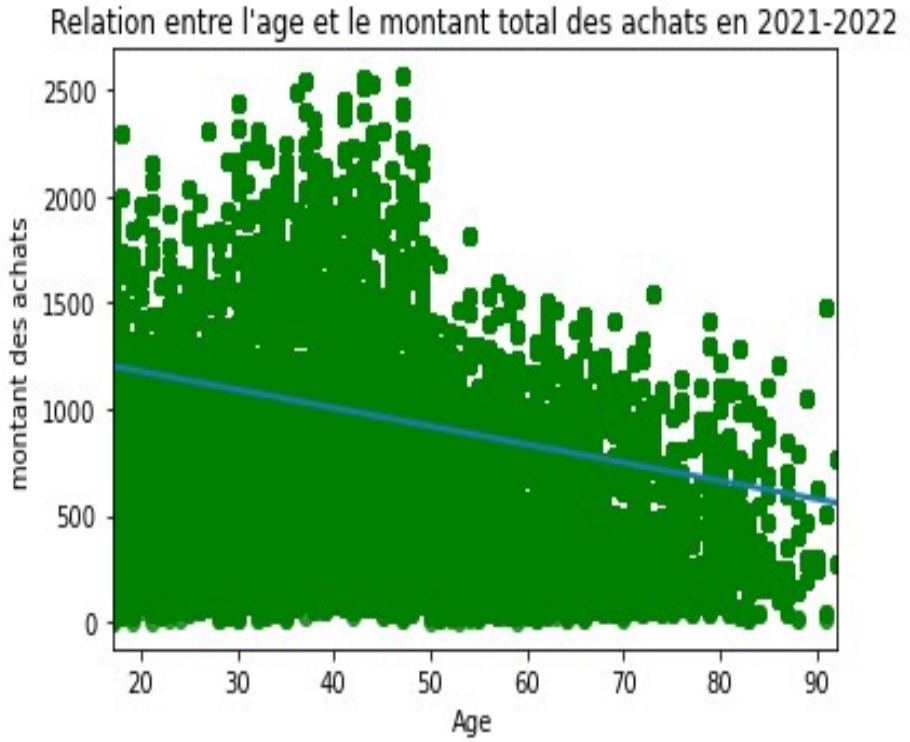
Mission 3: analyse des ventes (suite 2)

- **Relation entre âge et la taille du panier moyen (en nombre d'articles)**



-créer des classe d'âges pour faciliter la lecture des données
-Création d'une table de contingence
-Remplacement des valeurs nulles par 0
-Création de la heatmap.
st_chi2: 137137.97556981342
-Forte corrélation entre les deux variables

Mission 3: analyse des ventes (suite 3)

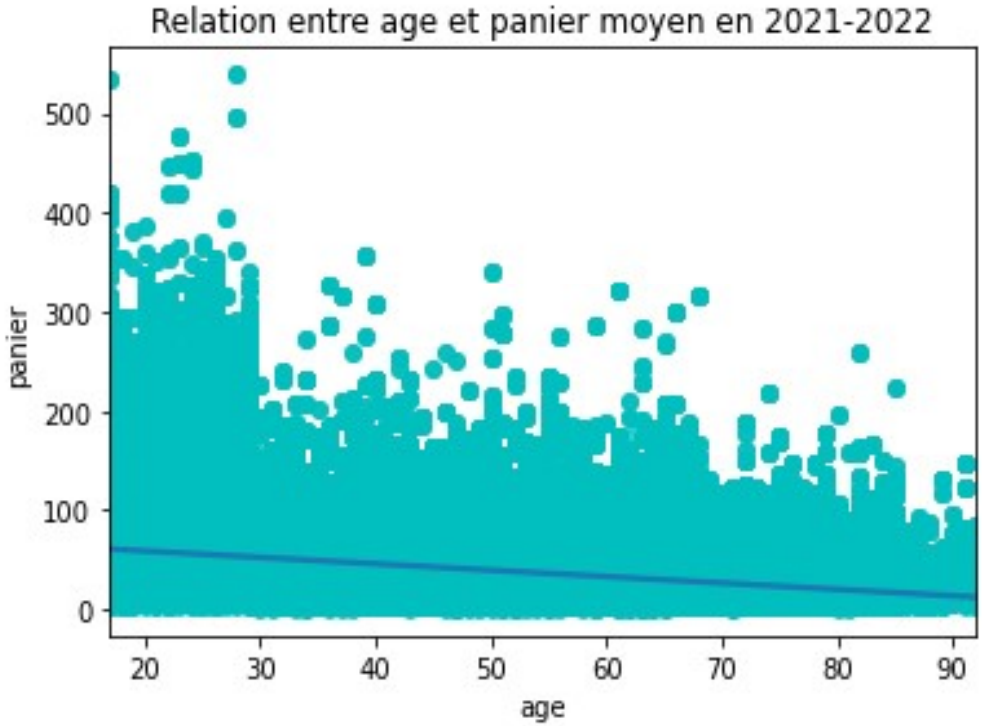


Relation entre âge et le montant total des achats

Le coefficient de Pearson est négatif, ce qui indique une faible relation entre les variables âge et montant total du panier. La p value conforte cette interprétation.

Mission 3: analyse des ventes (suite et fin)

- **Corrélation entre âge et panier moyen**



Le test de Pearson est de -0.2537483623525303
Il y a une faible corrélation ente les variables âge et le panier moyen en nombre de marchandises
Cette interprétation est soutenue par la p_value, inférieure à 0,05.