

---

## **Master de recherche en Informatique Module :**

### **Information Visualisation**

### **Travail Pratique**

---

## **Visualisation d'Information pour l'évaluation des maladies chroniques**

---

### Réalisé par :

- KETFI Raniya.
- Kessi Lamia.
- SI TAYEB Majda.

## Table des matières

<b>Introduction :</b>	4
<b>Application du processus de génération de visualisation :</b>	4
1- Analyse de données.....	4
Chargement des données :	5
Traitement des données manquantes :	5
Renommer des colonnes :	6
Normalisation :	7
2-Filtering :	8
3-Mapping:	9
4-Rendering :	12
Autres vues :	13
Conclusion :	15

## Table des figures

Figure 1: Chargement des données sous python.....	5
Figure 2: Pourcentage des valeurs null dans une colonne sous python .....	5
Figure 3: Suppression des colonnes sous Python.....	6
Figure 4: Suppression des colonnes avec les colonnes de scores n'ayant pas de sens .....	6
Figure 5: Renommage des colonnes sous python .....	7
Figure 6: Normalisation des données (Partie1) .....	7
Figure 7: Normalisation des données (Partie 2) .....	8
Figure 8: Projection des données par tranche d'âge .....	9
Figure 9: Projection par tranches d'âge et historique de maladie .....	9
Figure 10: Diagramme a barres par tranche d'âge.....	10
Figure 11: Diagramme a barres par tranches d'âge et maladie.....	11
Figure 12: Diagramme a barres avec deux interactions .....	12
Figure 13: Pie chart interactif (femmes) .....	13
Figure 14: Pie chart interactif (hommes) .....	13
Figure 15: Résultats du sondage nettoyés (Partie 1) .....	13
Figure 16: Résultats du sondage nettoyés (Partie 2) .....	14

# Résumé :

Aujourd'hui plus que jamais, les organisations utilisent la visualisation des données et des outils d'analyse des données pour poser de meilleures questions et prendre de meilleures décisions. Les technologies informatiques émergentes et les nouveaux logiciels intuitifs facilitent la compréhension de l'entreprise et permettent de prendre de meilleures décisions commerciales fondées sur les données.

L'accent mis sur les métriques de performance, les tableaux de bord de données et les indicateurs clés de performance (KPI) montre l'importance de la mesure et du suivi des données de l'entreprise.

Les visualisations de données sont devenues un outil essentiel pour les médias grand public d'aujourd'hui.

La visualisation de l'information est un ensemble de techniques permettant de représenter des données structurées. Le fait que l'on dispose d'informations sur les données distingue la visualisation de l'information de méthodes employées en statistique où l'on cherche justement à découvrir des relations entre les données. Le but de la visualisation de l'information est de représenter de façon cohérente et claire un nombre important de données afin qu'une personne puisse prendre conscience des informations structurelles présentes dans ces données. Pour cela, il faut tenir compte du fait que l'utilisateur peut être amené à manipuler la représentation qu'on lui offre (ce qui implique une visualisation et des interfaces adaptées).

Dans la majeure partie du temps, pour visualiser un ensemble de données il est important d'avoir recours à un processus appelé « Processus de génération de visualisation » qui contient un ensemble de phases nécessaires qui vont garantir l'obtention d'un résultat fiable, simple et rapide à interpréter. L'objectif de ce TP est d'appliquer ce processus sur une base de données d'évaluation des maladies chroniques en utilisant D3.JS.

## ***Introduction :***

De bonnes visualisations de données sont créées lorsque la communication, la science des données et le design s'assemblent. Les visualisations de données effectuées correctement offrent des aperçus clés sur des ensembles de données complexes d'une manière significative et intuitive. Le statisticien américain et professeur à Yale, Edward Tufte, pense que les excellentes visualisations de données consistent en des «idées complexes communiquées avec clarté, précision et efficacité».

Afin d'élaborer une bonne visualisation des données, il est nécessaire de commencer par des données claires, de sources sûres et complètes. Une fois que les données sont prêtes à être visualisées, il faut choisir le bon graphique. Cela peut être difficile, mais il existe de nombreuses ressources disponibles pour orienter le choix du type de graphique pour les données. Après avoir décidé quel type de graphique est le plus approprié, vient l'étape de la conception et la personnalisation de la visualisation. Le maître mot étant la simplicité car elle permet de rester clair et d'éviter les éléments qui rendent les données confuses.

Dans ce qui suit, nous présenterons plus en détail ce processus en l'appliquant sur des données d'évaluation des maladies chroniques.

## ***Application du processus de génération de visualisation :***

Comme le montre le schéma ci-dessus, le processus de génération de visualisation se divise en quatre étapes qui sont l'analyse de données (Data analysis), le filtrage (filtering), le plongement visuel (mapping) et le rendu (rendering). Nous allons expliquer comment nous avons appliqué chacune de ces étapes au dataset des malades chroniques.

### **1- Analyse de données**

Le fichier de données nécessite un prétraitement pour rendre le dataset utile, cette phase consiste en plusieurs opérations incluant le traitement des données manquantes, la correction des données erronées et la transformation des données.

Cette phase a été réalisée en utilisant Python et la bibliothèque pandas ( le code est rattaché au dossier du projet sous le nom `pre_processing_MNT.py`)

## Chargement des données :

```

Entrée [29]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

Entrée [33]: data=pd.read_excel('MNT.xlsx')
df=data.copy()

Entrée [34]: df.info(verbose=False)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 106 entries, 0 to 105
Columns: 75 entries, Horodateur to Unnamed: 74
dtypes: float64(24), int64(5), object(46)
memory usage: 62.2+ KB

```

Figure 1: Chargement des données sous python

Le jeu de données comprend 75 colonnes ( variables ou questions ) et 106 lignes qui représentent les réponses aux questions.

## Traitement des données manquantes :

Le dataset contient plusieurs colonnes ayant des valeurs NULL majoritairement ou un score NULL ,la fonction suivante permet de donner le pourcentage des valeurs NULL dans chaque colonne.

```

Entrée [36]: #prétraitement
missing_rate=df.isna().sum()/df.shape[0]

Entrée [37]: missing_rate*100

Out[37]: Horodateur                0.000000
Score total                0.000000
1.sexe                    0.000000
1.sexe [Score]             0.000000
1.sexe [Commentaires]     100.000000
...
24.avez vous d'autres maladies chroniques dans la famille? diabète, la tension, [Commentaires] 100.000000
25.si oui c'est quoi la maladies?                33.018868
25.si oui c'est quoi la maladies? [Score]         22.641509
25.si oui c'est quoi la maladies? [Commentaires] 100.000000
Unnamed: 74                92.452830
Length: 75, dtype: float64

```

Activer Wind

Figure 2: Pourcentage des valeurs null dans une colonne sous python

On remarque qu'il y a des colonnes inutiles ou toutes les données sont nulles ( 100 % ) ou manquantes( 92 % 33%..). C'est pourquoi, nous avons supprimé ces colonnes comme suit:

```
Entrée [38]: df.dropna(axis=1,inplace=True)
```

```
Entrée [39]: df
```

```
Out[39]:
```

	Horodateur	Score total	1.sexe	1.sexe [Score]	2.Quel âge avez-vous ? [Score]	2.Quel âge avez-vous ? [Score]	3.Quel est votre niveau d'étude [Score]	3.Quel est votre niveau d'étude [Score]	4.Quelles est votre activité professionnelle [Score]	4.Quelles est votre activité professionnelle [Score]	...	19.Est-ce que votre travail implique des activités physiques d'intensité modérée, comme une marche rapide ? [Score]	20.Est-ce que vous effectuez des trajets d'au moins 10 minutes à pied ou à vélo ? [Score]	20.Est-ce que vous effectuez des trajets d'au moins 10 minutes à pied ou à vélo ? [Score]	21.Habituellement, combien de fois par semaine effectuez-vous des trajets à pied ou à vélo ? [Score]
0	2020/10/12 10:19:31 PM UTC+1	0.00 / 0	homme	-- / 0	37	-- / 0	universitaire	-- / 0	Employé(e) de l'Etat	-- / 0	...	-- / 0	oui	-- / 0	-- / 0
1	2020/10/12 10:25:06 PM UTC+1	0.00 / 0	homme	-- / 0	37	-- / 0	universitaire	-- / 0	Employé(e) de l'Etat	-- / 0	...	-- / 0	oui	-- / 0	-- / 0
2	2020/10/12 10:28:55 PM UTC+1	0.00 / 0	femme	-- / 0	64	-- / 0	école secondaire	-- / 0	chomeur(se)	-- / 0	...	-- / 0	oui	-- / 0	-- / 0

Figure 3: Suppression des colonnes sous Python

Le nombre de colonnes se réduit à 36 colonnes , mais on constate aussi qu'il existe des colonnes de scores contenant la valeur ( -- / 0 ) qui n'a pas de sens , donc on va les supprimer avec les commandes suivantes :

```
Entrée [40]: df.drop(columns=df.columns[(df == '-- / 0').any()],inplace=True)
df.drop(columns=df.columns[(df == '0.00 / 0').any()],inplace=True)
df
```

```
Out[40]:
```

	Horodateur	1.sexe	2.Quel âge avez-vous ? [Score]	3.Quel est votre niveau d'étude [Score]	4.Quelles est votre activité professionnelle [Score]	5.Fumez-vous actuellement des produits à base de tabac tels que cigarettes, cigares ?	9.avez-vous consommé une boisson alcoolisée ? [Score]	14.Habituellement, combien de jours par semaine consommez-vous des fruits ?	15.Habituellement, combien de jours par semaine consommez-vous des légumes ?	16.Quelle sorte de matière grasse (huile, beurre...)utilisez-vous le plus souvent pour la préparation des repas à la maison ?	17.Habituellement, combien de fois par semaine effectuez-vous des activités physiques
0	2020/10/12 10:19:31 PM UTC+1	homme	37	universitaire	Employé(e) de l'Etat	Non	Non	0	3	Huile végétale	
1	2020/10/12 10:25:06 PM UTC+1	homme	37	universitaire	Employé(e) de l'Etat	oui	Non	3	5	Huile végétale	
2	2020/10/12 10:28:55 PM UTC+1	femme	64	école secondaire	chomeur(se)	Non	Non	5	5	Huile végétale	

Figure 4: Suppression des colonnes avec les colonnes de scores n'ayant pas de sens

Le nombre de colonnes ( variables ) se réduit à 17 colonnes seulement qui sont significatives.

## Renommer des colonnes :

On renomme les colonnes par des noms plus clairs et indicatifs

```
Entrée [47]: df.set_axis(['Date','Sexe','Age','Niveau etudes','activite professionnelle','Fumeur',
                        'alcoolique','consommation fruit par semaine','consommation de légumes par semaine',
                        'matiere grasse utilisée','activité physique','salle de sport','activité physique au travail',
                        'marche ou velo','marche ou velo par semaine','sport','maladie chronique'], axis=1, inplace=True)
```

```
Entrée [48]: df
```

```
Out[48]:
```

	Date	Sexe	Age	Niveau etudes	activite professionnelle	Fumeur	alcoolique	consommation fruit par semaine	consommation de légumes par semaine	matiere grasse utilisée	activité physique	salle de sport	activité physique au travail	marche ou velo
0	2020/10/12 10:19:31 PM UTC+1	homme	37	universitaire	Employé(e) de l'Etat	Non	Non	0	3	Huile végétale	5	Non	oui	oui
1	2020/10/12 10:25:06 PM UTC+1	homme	37	universitaire	Employé(e) de l'Etat	oui	Non	3	5	Huile végétale	3	oui	oui	oui
2	2020/10/12 10:28:55 PM UTC+1	femme	64	école secondaire	chomeur(se)	Non	Non	5	5	Huile végétale	5	Non	oui	oui
3	2020/10/12 10:29:54 PM UTC+1	homme	44	école secondaire	Employé(e) dans le privé	oui	Non	3	5	Huile végétale	1	Non	Non	oui
4	2020/10/12 10:36:22 PM UTC+1	homme	25	universitaire	étudiant	Non	Non	2	4	Zit ziton	2	Non	Non	Non

Figure 5: Renommage des colonnes sous python

## Normalisation :

On constate que la variable âge en tant qu'un nombre entier ne nous aide pas à visualiser correctement l'impact de l'âge sur le phénomène observé c'est pour cela que nous proposons de diviser la population sur des groupes par tranches d'âges : (18-28) -(28-38)-(38-48)-(48-58)-(58-68).

```
Entrée [93]: tranche_age=pd.cut(df.Age,bins=[0,18,28,38,48,58,68],labels=['0-18','18-28','28-38','38-48','48-58','58-68'])
df.insert(5,'tranche_age',tranche_age)
```

```
Entrée [94]: df
```

```
Out[94]:
```

	Date	Sexe	Age	Niveau etudes	activite professionnelle	tranche_age	Fumeur	alcoolique	consommation fruit par semaine	consommation de légumes par semaine	matiere grasse utilisée	activité physique	salle de sport	activit physique au trava
0	2020/10/12 10:19:31 PM UTC+1	homme	37	universitaire	Employé(e) de l'Etat	28-38	Non	Non	0	3	Huile végétale	5	Non	oi
1	2020/10/12 10:25:06 PM UTC+1	homme	37	universitaire	Employé(e) de l'Etat	28-38	oui	Non	3	5	Huile végétale	3	oui	oi
2	2020/10/12 10:28:55 PM UTC+1	femme	64	école secondaire	chomeur(se)	58-68	Non	Non	5	5	Huile végétale	5	Non	oi
3	2020/10/12 10:29:54 PM UTC+1	homme	44	école secondaire	Employé(e) dans le privé	38-48	oui	Non	3	5	Huile végétale	1	Non	No
4	2020/10/12 10:36:22 PM UTC+1	homme	25	universitaire	étudiant	18-28	Non	Non	2	4	Zit ziton	2	Non	No

Figure 6: Normalisation des données (Partie1)

Pour les lignes ayant comme réponse Oui /Non , il vaut mieux les normaliser en les transformant à une variable booléenne pour accélérer le traitement des données

```
Entrée [117]: df =df.replace({'alcoolique': {'oui': True, 'Non': False}})
df =df.replace({'salle de sport': {'oui': True, 'Non': False}})
df =df.replace({'activité physique au travail': {'oui': True, 'Non': False}})
df =df.replace({'marche ou velo': {'oui': True, 'Non': False}})
df =df.replace({'sport': {'oui': True, 'Non': False}})
```

```
Entrée [118]: df
```

```
Out[118]:
```

	Date	Sexe	Age	Niveau etudes	activite professionnelle	tranche_age	Fumeur	alcoolique	consommation fruit par semaine	consommation de légumes par semaine	matiere grasse utilisée	activité physique	salle de sport	activit physiqu au trava
0	2020/10/12 10:19:31 PM UTC+1	homme	37	universitaire	Employé(e) de l'Etat	28-38	False	False	0	3	Huile végétale	5	False	Tru
1	2020/10/12 10:25:06 PM UTC+1	homme	37	universitaire	Employé(e) de l'Etat	28-38	True	False	3	5	Huile végétale	3	True	Tru
2	2020/10/12 10:28:55 PM UTC+1	femme	64	école secondaire	chomeur(se)	58-68	False	False	5	5	Huile végétale	5	False	Tru
3	2020/10/12 10:29:54 PM UTC+1	homme	44	école secondaire	Employé(e) dans le privé	38-48	True	False	3	5	Huile végétale	1	False	Fals

Figure 7: Normalisation des données (Partie 2)

## 2-Filtering :

Pour mieux comprendre le phénomène présent dans ce jeu de données, nous avons choisi nos variables en fonction des critères suivants:

1. La complétude des données de la variable choisi (la complétude concerne le taux de valeurs non null)
2. La pertinence des variables par rapport à la problématique traité. En effet, notre étude porte sur les maladies chroniques et les principaux facteurs qui influent sur l'apparition de cette maladie sont le sexe, la tranche d'âge, la consommation d'alcool...

Nous avons donc finalement opté pour les 5 variables suivantes :

1. sexe
2. tranche d'âge
3. Fumeur
4. alcoolique
5. maladie chronique

Pour alimenter notre code avec les données, on avait besoin des tableaux croisés dynamiques qui projettent nos données sur des groupes :

### 1. Par tranches d'âge :



```
Entrée [81]: table = pd.pivot_table(data=new,index=['tranche_age'],columns=['Fumeur','alcoolique','Sexe'],
aggfunc={'maladie_chronique':"count"})
table
```

Out[81]:

	maladie chronique							
	False				True			
	False		True		False		True	
	Sexe	femme	homme	femme	homme	femme	homme	femme
tranche_age								
0-18		7	3	0	0	0	0	0
18-28		31	33	0	2	0	1	1
28-38		6	10	0	3	0	2	1
38-48		1	1	0	0	0	1	0
48-58		1	1	0	0	0	0	0
58-68		1	0	0	0	0	0	0

Figure 8: Projection des données par tranche d'âge

## 2. Par tranches d'âge et historique de maladie chronique :

table

	femme non F non A	homme non F non A	femme non F A	homme non F A	femme F non A	homme F non A	femme F A	homme F A
0-18 pas de mal-chronique	1	1	0	0	0	0	0	0
0-18 mal-chronique	6	2	0	0	0	0	0	0
18-28 pas de mal-chronique	8	12	0	1	0	0	0	1
18-28 mal-chronique	23	21	0	1	0	1	0	0
28-38 pas de mal-chronique	0	5	0	0	0	1	0	0
28-38 mal-chronique	6	5	0	3	0	1	0	1
38-48 pas de mal-chronique	1	0	0	0	0	1	0	0
38-48 mal-chronique	0	1	0	0	0	0	0	0
48-58 pas de mal-chronique	0	0	0	0	0	0	0	0
48-58 mal-chronique	1	1	0	0	0	0	0	0
58-68 pas de mal-chronique	0	0	0	0	0	0	0	0
58-68 mal-chronique	1	0	0	0	0	0	0	0

Figure 9: Projection par tranches d'âge et historique de maladie

## 3-Mapping:

Avec D3.JS, nous avons construit des diagrammes à barres pour représenter l'évaluation des nombres de malades par rapport aux variables citées ci- dessus. Ces diagrammes servent à montrer la comparaison entre les catégories de ces dernières.

Choix de couleur : vu le nombre de classes obtenues ( 8 classes ) , on s'est trouvé dans l'obligation de choisir les plus distinctes possibles ( opposées) pour pouvoir différencier les classes

. Les couleurs ont été sélectionnés en fonction du sexe et du rapport alcool/tabac. Le bleu pour les hommes et l'orange pour les femmes, de plus nous avons utilisé une dégradation de gris et bleu pour montrer le rapport alcool/tabac.

### 3.1 Diagramme à barre par tranche d'âge:

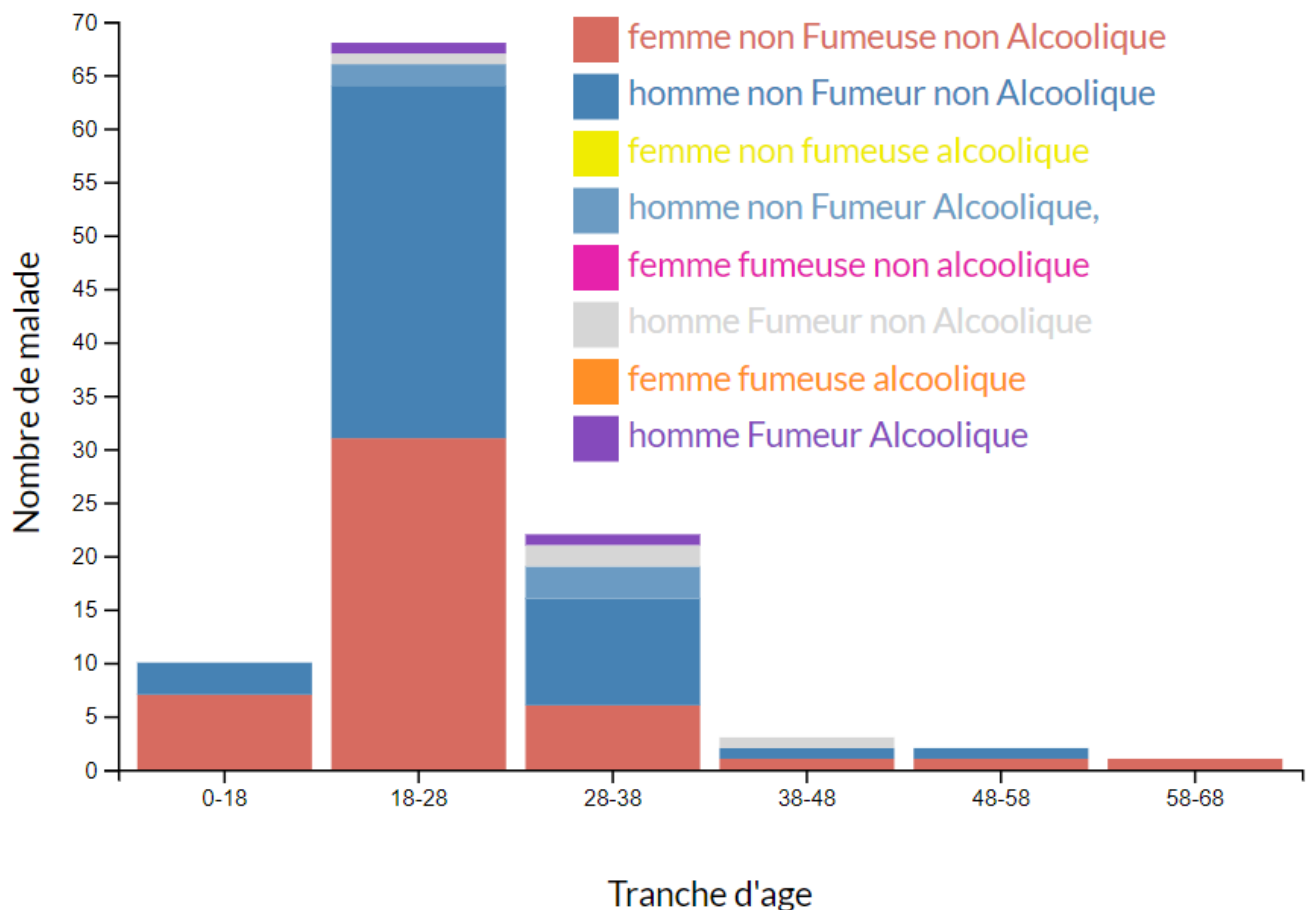


Figure 10: Diagramme a barres par tranche d'âge

### Discussion :

On remarque qu'un grand nombre de malades est compris entre 18 ans et 28 ans, ce qui semble anormal en considérant que les jeunes sont généralement en bonne santé. En analysant les résultats du formulaire nous constatons que la majorité des réponses appartiennent à des personnes dont l'âge est entre 18 et 28 ans.

De plus, nous avons remarqué que le tabac et l'alcool ne sont pas des facteurs clés dans l'augmentation du nombre de malades (un grand pourcentage des personnes observées ne sont pas alcoolique et fumeurs), cela est peut-être dû à des raisons sociales mais aussi aux réponses au formulaire dans lequel nous constatons qu'il n'y a ni des femmes fumeuses ni des femmes alcooliques.

Pour cela nous avons effectué une autre analyse et catégorisation en ajoutant les maladies chroniques à la tranche d'âge pour mieux comprendre notre phénomène.

## 3.2 Diagramme à barre par tranche d'âge et maladie :

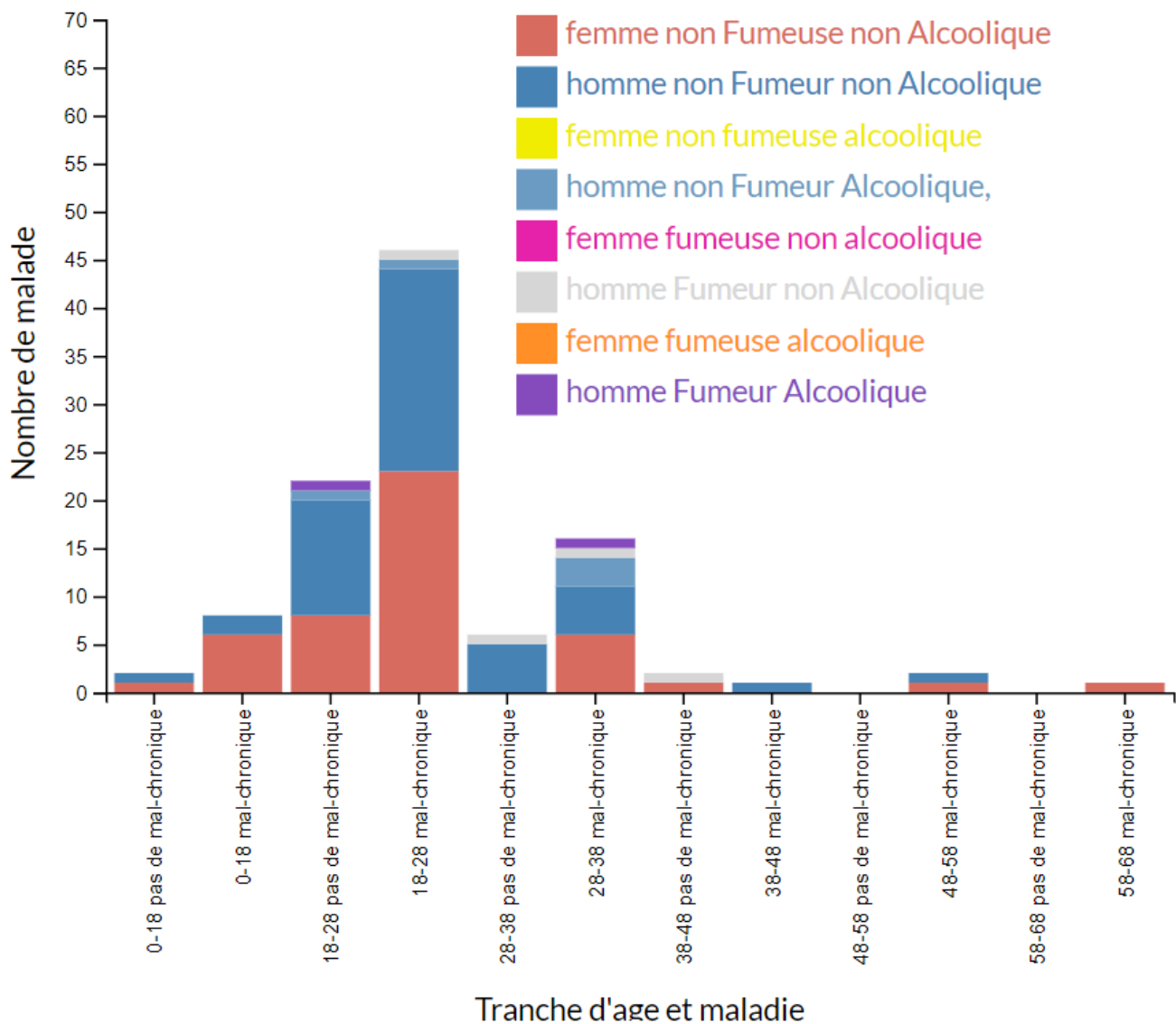


Figure 11: Diagramme a barres par tranches d'âge et maladie

Discussion :

En ajoutant la variable “maladie chronique” nous avons pu obtenir une image plus claire. En fait, les personnes dont l'âge est compris entre 18 et 28 ans qui ne fument pas et ne boivent pas ont des antécédents de maladies chroniques dans la famille ce qui justifie le grand nombre de malades dans cette catégorie.

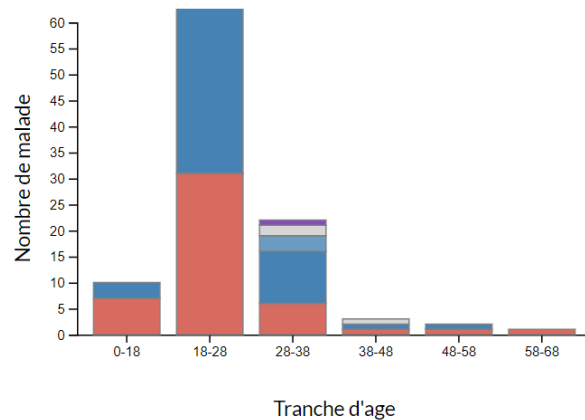
De plus, nous avons constaté que la plupart des malades ont des antécédents de maladies chroniques dans la famille.

## 4-Rendering :

Pour améliorer la vue visuelle des diagrammes à barres, nous avons ajouté deux tâches interactives qui seront mises en valeur quand l'utilisateur clique sur la surface concernée (la barre ou une partie de la barre), qui sont:

1. Affichage du groupe,
2. Affichage du nombre de personnes dans le groupe.

Pour réaliser ces tâches nous avons utilisé une interaction de type "tooltip":



Classe: femme non F non A  
nombre de malade: 31

*Figure 12: Diagramme à barres avec deux interactions*

### Remarque :

- Le dataset étant chargé dans le GitHub , il faut avoir une connexion interne pour visualiser les fichiers .html envoyés.
- Le lien de d3.js est un lien en ligne, il faut avoir une connexion interne pour visualiser les graphes.

## Autres vues :

Les graphes suivants rendent compte de l'analyse de notre dataset sur les maladies chroniques en utilisant un autre type de visualisation à savoir le pie chart. Les variables qui nous ont semblé les plus pertinentes pour notre modélisation sont la tranche d'âge et le sexe. Notre graphe exprime donc la répartition des malades par tranche d'âge et il est possible grâce à l'interaction que nous avons mis en œuvre, de visualiser cette répartition que ce soit pour les hommes ou pour les femmes en cliquant sur un bouton.

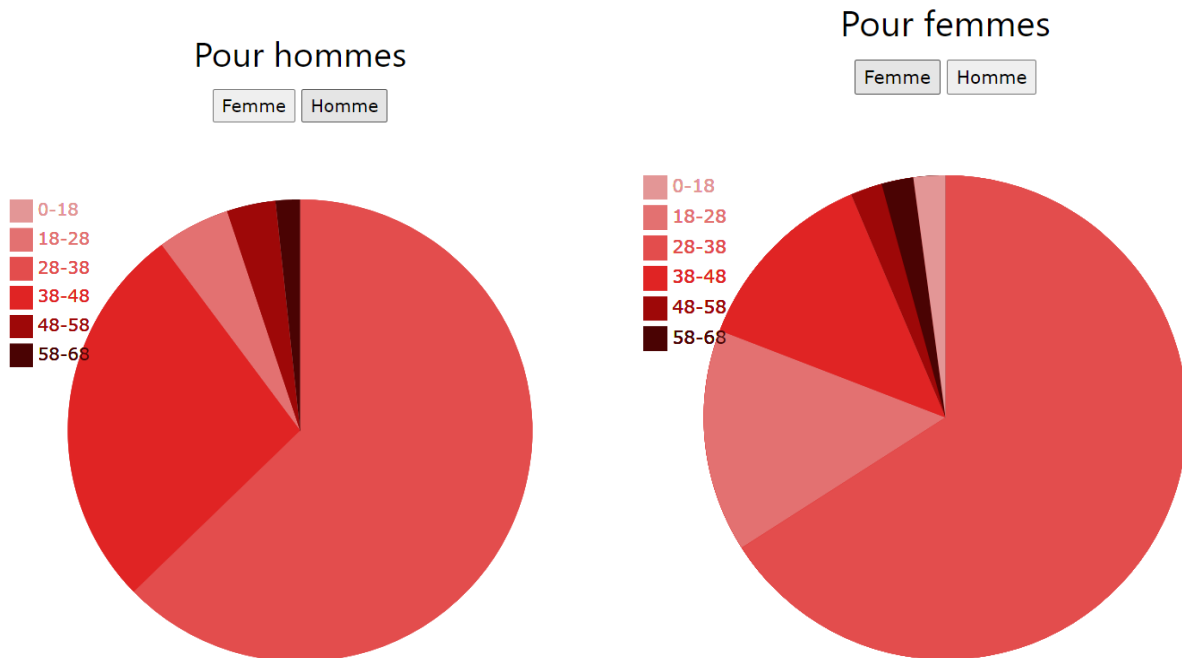


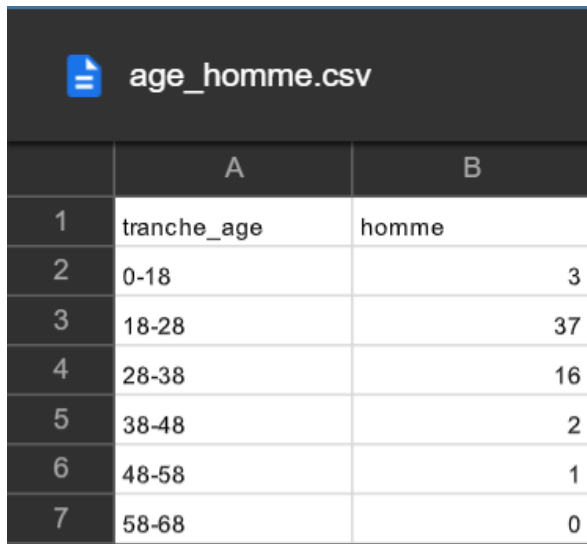
Figure 13: Pie chart interactif (femmes)

Figure 14: Pie chart interactif (hommes)

Nous avons nettoyé les résultats du sondage mené pour en extraire les 2 variables qui nous intéressent comme le montre la figure suivante :

age_femme.csv		
	A	B
1	tranche_age	femme
2	0-18	7
3	18-28	31
4	28-38	6
5	38-48	1
6	48-58	1
7	58-68	1

Figure 15: Résultats du sondage nettoyés (Partie 1)



	A	B
1	tranche_age	homme
2	0-18	3
3	18-28	37
4	28-38	16
5	38-48	2
6	48-58	1
7	58-68	0

Figure 16: Résultats du sondage nettoyés (Partie 2)

## Conclusion :

A travers ce TP, nous avons pris conscience de l'importance de la visualisation pour comprendre et mieux analyser des données brutes et non structurées.

Nous avons pu mettre en pratique les notions étudiées en cours en particulier le processus de génération de visualisation qui s'est avéré indispensable pour faciliter l'analyse des données, leur compréhension et enfin leur visualisation.

En effet, nous avons appliqué ce processus sur des données d'évaluation des maladies chroniques qui étaient au départ incomplètes et inexploitable à première vue. Notre travail a dans ce sens, consisté à produire des visualisations pour des futures interprétations en utilisant D3.JS et ce en ayant recours à des diagrammes de barres et des diagrammes en secteurs.

Les perspectives pour ce travail seraient d'envisager l'utilisation d'un dataset plus volumineux avec des données très difficilement interprétables en vue de mettre en valeur de manière encore plus frappante l'impact de la visualisation dans la compréhension d'une masse d'information.