

# Polytechnique Montréal Département Génie Informatique et Génie Logiciel INF8008 – Prétraitement de données .

## TP2 - Transformation, distribution et statistiques descriptives Hiver 2025 . Janvier 2025

### Introduction

Le TD2 porte sur la transformation, la distribution et les statistiques descriptives. **Nous survolons l'utilisation de fonctions de base de Pandas et de l'analyse de données numériques.** Les données du fichier *Alzheimer\_s\_Disease\_and\_Healthy\_Aging\_Data.csv* sont des données publiques provenant d'enquêtes sur le vieillissement et la santé, faites par le [Département de la Santé et des Services sociaux des États-Unis](#). Contrairement aux données du TP1 qui avaient été traitées au préalable, celles utilisées pour ce TP ne le sont pas. Vous devrez traiter les données brutes pour obtenir une version plus condensée, facilitant l'analyse des tendances et des sous-groupes de population.

Les champs principaux du fichier de données

**Alzheimer\_s\_Disease\_and\_Healthy\_Aging\_Data.csv** sont les suivants :

- **YearStart/YearEnd** : années de début et de fin des données
- **LocationAbbr** : abréviation du lieu
- **Class** : catégorie des données (ex. : Santé mentale)
- **Topic** : sujet spécifique (ex. : détresse mentale fréquente)
- **Question** : question étudiée
- **Data\_Value\_Unit** : unité de mesure des données (ex. : pourcentage)
- **Data\_Value** : valeur des données collectées
- **StratificationCategory1 / Stratification1** : catégorie et détail de la première stratification (ex. : âge, genre)
- **StratificationCategory2 / Stratification2** : catégorie et détail de la deuxième stratification (ex. : race, ethnie)

Ces données servent de base pour explorer les tendances, identifier des corrélations, et mieux comprendre les facteurs liés aux maladies neurodégénératives et à la santé mentale des populations vieillissantes. Votre objectif dans ce TP sera de préparer ces données pour qu'elles soient prêtes pour une analyse approfondie.

Voici les librairies python qui sera à utiliser pour ce TP :

- [pandas](#)
- [numpy](#)
- [matplotlib](#)

**À noter qu'au niveau de chaque question, il est recommandé de copier le DataFrame obtenu à la question précédente dans un nouveau DataFrame.**

*Veillez vous référer à l'énoncé PDF de ce TP pour voir la sortie attendue.*

```
import pandas as pd
import numpy as np
```

```
df = pd.read_csv('Alzheimer_s_Disease_and_Healthy_Aging_Data.csv')
df
```

		RowId	YearStart	YearEnd	\
0	BRFSS~2022~2022~42~Q03~TMC01~AGE~RACE		2022	2022	
1	BRFSS~2022~2022~46~Q03~TMC01~AGE~RACE		2022	2022	
2	BRFSS~2022~2022~16~Q03~TMC01~AGE~RACE		2022	2022	
3	BRFSS~2022~2022~24~Q03~TMC01~AGE~RACE		2022	2022	
4	BRFSS~2022~2022~55~Q03~TMC01~AGE~GENDER		2022	2022	
...		...	...	...	
284137	BRFSS~2016~2016~55~Q15~TSC02~AGE~RACE		2016	2016	
284138	BRFSS~2017~2017~56~Q45~TOC13~AGE~RACE		2017	2017	
284139	BRFSS~2015~2015~56~Q42~TCC04~AGE~RACE		2015	2015	
284140	BRFSS~2019~2019~54~Q46~TOC10~AGE~RACE		2019	2019	
284141	BRFSS~2015~2015~56~Q02~TNC02~AGE~RACE		2015	2015	

	LocationAbbr	LocationDesc	Datasource	\
0	PA	Pennsylvania	BRFSS	
1	SD	South Dakota	BRFSS	
2	ID	Idaho	BRFSS	
3	MD	Maryland	BRFSS	
4	WI	Wisconsin	BRFSS	
...	...	...	...	
284137	WI	Wisconsin	BRFSS	
284138	WY	Wyoming	BRFSS	
284139	WY	Wyoming	BRFSS	
284140	WV	West Virginia	BRFSS	
284141	WY	Wyoming	BRFSS	

		Class	\
0		Mental Health	
1		Mental Health	
2		Mental Health	
3		Mental Health	
4		Mental Health	
...		...	
284137	Screenings and Vaccines		
284138	Overall Health		
284139	Cognitive Decline		
284140	Overall Health		
284141	Nutrition/Physical Activity/Obesity		

		Topic	\
0	Frequent mental distress		
1	Frequent mental distress		
2	Frequent mental distress		
3	Frequent mental distress		

4	Frequent mental distress
...	...
284137	Colorectal cancer screening
284138	Fair or poor health among older adults with ar...
284139	Talked with health care professional about sub...
284140	Disability status, including sensory or mobili...
284141	Eating 3 or more vegetables daily

Data_Value_Unit \	Question
0	Percentage of older adults who are experiencin...
%	
1	Percentage of older adults who are experiencin...
%	
2	Percentage of older adults who are experiencin...
%	
3	Percentage of older adults who are experiencin...
%	
4	Percentage of older adults who are experiencin...
%	
...	...
...	
284137	Percentage of older adults who had either a ho...
%	
284138	Fair or poor health among older adults with do...
%	
284139	Percentage of older adults with subjective cog...
%	
284140	Percentage of older adults who report having a...
%	
284141	Percentage of older adults who are eating 3 or...
%	

...	Stratification2
Geolocation \	
0	Native Am/Alaskan Native POINT (-77.86070029 40.79373015)
1	Asian/Pacific Islander POINT (-100.3735306 44.35313005)
2	Black, non-Hispanic POINT (-114.36373 43.68263001)
3	Black, non-Hispanic POINT (-76.60926011 39.29058096)
4	Male POINT (-89.81637074 44.39319117)
...	...
...	...
284137	Black, non-Hispanic POINT (-89.81637074 44.39319117)

284138 ... Hispanic POINT (-108.1098304  
43.23554134)  
284139 ... Asian/Pacific Islander POINT (-108.1098304  
43.23554134)  
284140 ... Hispanic POINT (-80.71264013  
38.6655102)  
284141 ... Native Am/Alaskan Native POINT (-108.1098304  
43.23554134)

	ClassID	TopicID	QuestionID	LocationID
StratificationCategoryID1 \				
0	C05	TMC01	Q03	42
AGE				
1	C05	TMC01	Q03	46
AGE				
2	C05	TMC01	Q03	16
AGE				
3	C05	TMC01	Q03	24
AGE				
4	C05	TMC01	Q03	55
AGE				
...	...	...	...	...
...				
284137	C03	TSC02	Q15	55
AGE				
284138	C01	T0C13	Q45	56
AGE				
284139	C06	TCC04	Q42	56
AGE				
284140	C01	T0C10	Q46	54
AGE				
284141	C02	TNC02	Q02	56
AGE				

	StratificationID1	StratificationCategoryID2	StratificationID2
0	5064	RACE	NAA
1	65PLUS	RACE	ASN
2	65PLUS	RACE	BLK
3	65PLUS	RACE	BLK
4	65PLUS	GENDER	MALE
...	...	...	...
284137	AGE_OVERALL	RACE	BLK

284138	5064	RACE	HIS
284139	AGE_OVERALL	RACE	ASN
284140	65PLUS	RACE	HIS
284141	5064	RACE	NAA
[284142 rows x 31 columns]			

A)

Vous remarquerez que ce jeu de données est assez large, avec 284142 lignes et 31 colonnes.

Avec des ensembles de données de cette taille, on peut souvent trouver des défauts, comme des doublons de lignes. Vérifiez donc s'il existe des valeurs en double dans le DataFrame. (2 points)

```
#TODO:
has_duplicates = "oui" if df.duplicated().any() else "non"
print("Le DataFrame contient-il des doublons ? :", has_duplicates)
```

Le DataFrame contient-il des doublons ? : non

B)

Il est possible d'extraire la durée du sondage en soustrayant l'année de début de l'année de fin. Utilisez lambda, ainsi que cette soustraction, pour garder les lignes avec une durée de sondage de moins d'1 an. (3 points)

```
#TODO:
df = df[(df['YearEnd'] - df['YearStart']).apply(lambda x : x<1)]
```

C)

Maintenant que cette étape est faite, les colonnes YearStart et YearEnd contiennent la même information. Renommez une des deux colonnes à Year, et supprimez l'autre. (2 points)

```
#TODO:
df = df.rename(columns={'YearStart': 'Year'}).drop(columns=['YearEnd'])
```

D)

Certaines colonnes contiennent des données redondantes ou inutiles pour notre analyse. Éliminez toutes les colonnes inutiles en ne conservant que celles mentionnées dans l'introduction. Combien de colonnes reste-t-il ? (2 points)

*#TODO:*

```
columns_to_keep = ['Year',  
'LocationAbbr', 'Class', 'Topic', 'Question', 'Data_Value_Unit', 'Data_Value',  
'StratificationCategory1', 'Stratification1', 'StratificationCategory2', 'Stratification2']  
df = df[columns_to_keep]  
num_columns = df.shape[1]  
print(f"Nombre de colonnes restantes : {num_columns}")
```

Nombre de colonnes restantes : 11

E)

Comme vu dans le module 1, le prétraitement des données consiste à gérer les défauts des données collectées, comme les valeurs nulles. La colonne `Data_Value` est importante pour notre analyse.

Vérifiez donc s'il existe des données manquantes dans la colonne `Data_Value`. Quel est le pourcentage de valeurs manquantes ? (3 points)

*#TODO:*

```
Nan_values = 'oui' if df['Data_Value'].isna().any() else 'non'  
Nan_values_per = df['Data_Value'].isna().sum() / len(df['Data_Value'])  
  
print(f'Existe-il des valeurs manquantes ? {Nan_values} \n '  
      f'Pourcentages de valeurs manquantes {Nan_values_per:.2f}%')
```

Existe-il des valeurs manquantes ? oui  
Pourcentages de valeurs manquantes 0.32%

F)

Deux façons de traiter les données manquantes: les remplacer par la valeur médiane ou les éliminer complètement.

Il n'existe pas de solution **unique ou meilleure**. Tout dépend de l'analyse effectuée. Il est essentiel d'examiner les effets de chacun de ces choix sur l'analyse ultérieure. C'est pourquoi, dans ce TP, nous essayerons les deux méthodes.

Vous devez donc:

1. Créez deux copies de l'ensemble de données.
2. Supprimez les valeurs manquantes d'une des copies.
3. Remplacez les valeurs manquantes d'une autre copie par la médiane.

Affichez les nouveaux dataframes. Vous devriez avoir autour de 186595 lignes pour l'un et 274881 lignes pour l'autre. (4 points)

#TODO:

```
df_filled = df.copy()
df_dropped = df.copy()
```

```
df_dropped = df_dropped[df_dropped['Data_Value'].isna() == False]
df_filled['Data_Value'].fillna(df_filled['Data_Value'].median(),
inplace=True)
```

```
rows_dropped = df_dropped.shape[0]
rows_filled = df_filled.shape[0]
```

```
print(f"Nombre de lignes après suppression des valeurs manquantes :
{rows_dropped}")
print(f"Nombre de lignes après remplacement par la médiane :
{rows_filled}")
```

Nombre de lignes après suppression des valeurs manquantes : 186595  
Nombre de lignes après remplacement par la médiane : 274881

df\_filled

	Year	LocationAbbr	Class \
0	2022	PA	Mental Health
1	2022	SD	Mental Health
2	2022	ID	Mental Health
3	2022	MD	Mental Health
4	2022	WI	Mental Health
...	...	...	...
284137	2016	WI	Screenings and Vaccines
284138	2017	WY	Overall Health
284139	2015	WY	Cognitive Decline
284140	2019	WV	Overall Health
284141	2015	WY	Nutrition/Physical Activity/Obesity

	Topic \
0	Frequent mental distress
1	Frequent mental distress
2	Frequent mental distress
3	Frequent mental distress
4	Frequent mental distress
...	...
284137	Colorectal cancer screening
284138	Fair or poor health among older adults with ar...
284139	Talked with health care professional about sub...
284140	Disability status, including sensory or mobili...
284141	Eating 3 or more vegetables daily

	Question
Data_Value_Unit \	

0	Percentage of older adults who are experiencin...
%	
1	Percentage of older adults who are experiencin...
%	
2	Percentage of older adults who are experiencin...
%	
3	Percentage of older adults who are experiencin...
%	
4	Percentage of older adults who are experiencin...
%	
...	...
...	
284137	Percentage of older adults who had either a ho...
%	
284138	Fair or poor health among older adults with do...
%	
284139	Percentage of older adults with subjective cog...
%	
284140	Percentage of older adults who report having a...
%	
284141	Percentage of older adults who are eating 3 or...
%	

	Data_Value	StratificationCategory1	Stratification1	\
0	33.0	Age Group	50-64 years	
1	33.0	Age Group	65 years or older	
2	33.0	Age Group	65 years or older	
3	9.0	Age Group	65 years or older	
4	5.6	Age Group	65 years or older	
...	...			
284137	70.7	Age Group	Overall	
284138	33.0	Age Group	50-64 years	
284139	33.0	Age Group	Overall	
284140	33.0	Age Group	65 years or older	
284141	33.0	Age Group	50-64 years	

	StratificationCategory2	Stratification2
0	Race/Ethnicity	Native Am/Alaskan Native
1	Race/Ethnicity	Asian/Pacific Islander
2	Race/Ethnicity	Black, non-Hispanic
3	Race/Ethnicity	Black, non-Hispanic
4	Gender	Male
...	...	
284137	Race/Ethnicity	Black, non-Hispanic
284138	Race/Ethnicity	Hispanic
284139	Race/Ethnicity	Asian/Pacific Islander
284140	Race/Ethnicity	Hispanic
284141	Race/Ethnicity	Native Am/Alaskan Native

[274881 rows x 11 columns]



df\_dropped

	Year	LocationAbbr	Class \
3	2022	MD	Mental Health
4	2022	WI	Mental Health
6	2022	OK	Mental Health
7	2022	PA	Mental Health
8	2022	AZ	Overall Health
...	...	...	...
284131	2015	WI	Cognitive Decline
284132	2020	US	Caregiving
284134	2016	WV	Nutrition/Physical Activity/Obesity
284135	2017	WV	Overall Health
284137	2016	WI	Screenings and Vaccines

	Topic \
3	Frequent mental distress
4	Frequent mental distress
6	Frequent mental distress
7	Frequent mental distress
8	Recent activity limitations in past month
...	...
284131	Need assistance with day-to-day activities bec...
284132	Provide care for someone with cognitive impair...
284134	Obesity
284135	Arthritis among older adults
284137	Colorectal cancer screening

	Question
Data_Value_Unit \	
3	Percentage of older adults who are experiencin...
%	
4	Percentage of older adults who are experiencin...
%	
6	Percentage of older adults who are experiencin...
%	
7	Percentage of older adults who are experiencin...
%	
8	Mean number of days with activity limitations ...
Number	
...	...
...	
284131	Percentage of older adults who reported that a...
%	
284132	Percentage of older adults who provided care f...
%	
284134	Percentage of older adults who are currently o...
%	
284135	Percentage of older adults ever told they have...
%	

284137 Percentage of older adults who had either a ho...  
%

	Data_Value	StratificationCategory1	Stratification1 \
3	9.0	Age Group	65 years or older
4	5.6	Age Group	65 years or older
6	21.5	Age Group	Overall
7	10.0	Age Group	Overall
8	6.1	Age Group	65 years or older
...	...	...	...
284131	29.2	Age Group	50-64 years
284132	28.5	Age Group	65 years or older
284134	44.5	Age Group	50-64 years
284135	57.7	Age Group	Overall
284137	70.7	Age Group	Overall

	StratificationCategory2	Stratification2
3	Race/Ethnicity	Black, non-Hispanic
4	Gender	Male
6	Race/Ethnicity	Native Am/Alaskan Native
7	Race/Ethnicity	White, non-Hispanic
8	Race/Ethnicity	White, non-Hispanic
...	...	...
284131	Gender	Female
284132	Race/Ethnicity	Hispanic
284134	Gender	Female
284135	NaN	NaN
284137	Race/Ethnicity	Black, non-Hispanic

[186595 rows x 11 columns]

**Note :** Pour la suite du travail, chaque étape devra être réalisée sur les deux copies de l'ensemble de données.

G)

Plusieurs classes existent. On va évaluer la santé mentale "Mental Health". Filtrez les données de la colonne "class" pour la valeur "Mental Health", puis déterminez la moyenne de Data\_Value par Year et Topic. (2 points)

```
#TODO:
## for df_filled
mean_filled = df_filled[df_filled['Class'] == 'Mental
Health'].groupby(['Year' , 'Topic'])['Data_Value'].mean()
mean_filled
```

Year	Topic	
2015	Frequent mental distress	18.601085
	Lifetime diagnosis of depression	23.518149

2016	Frequent mental distress	18.775391
	Lifetime diagnosis of depression	22.530156
2017	Frequent mental distress	18.710397
	Lifetime diagnosis of depression	23.513863
2018	Frequent mental distress	18.829718
	Lifetime diagnosis of depression	22.814389
2019	Frequent mental distress	18.950699
	Lifetime diagnosis of depression	22.567108
2020	Frequent mental distress	19.122905
	Lifetime diagnosis of depression	22.561561
2021	Frequent mental distress	19.217354
	Lifetime diagnosis of depression	22.662039
2022	Frequent mental distress	19.408730
	Lifetime diagnosis of depression	23.028531

Name: Data\_Value, dtype: float64

```
# for df_dropped
mean_dropped = df_dropped[df_dropped['Class'] == 'Mental
Health'].groupby(['Year', 'Topic'])['Data_Value'].mean()
mean_dropped
```

Year	Topic	
2015	Frequent mental distress	10.370795
	Lifetime diagnosis of depression	18.838661
2016	Frequent mental distress	10.678795
	Lifetime diagnosis of depression	17.439112
2017	Frequent mental distress	10.960913
	Lifetime diagnosis of depression	19.199265
2018	Frequent mental distress	10.930743
	Lifetime diagnosis of depression	17.982196
2019	Frequent mental distress	10.901620
	Lifetime diagnosis of depression	17.487637
2020	Frequent mental distress	10.770949
	Lifetime diagnosis of depression	17.193873
2021	Frequent mental distress	11.314676
	Lifetime diagnosis of depression	17.323026
2022	Frequent mental distress	11.792802
	Lifetime diagnosis of depression	17.957388

Name: Data\_Value, dtype: float64

H)

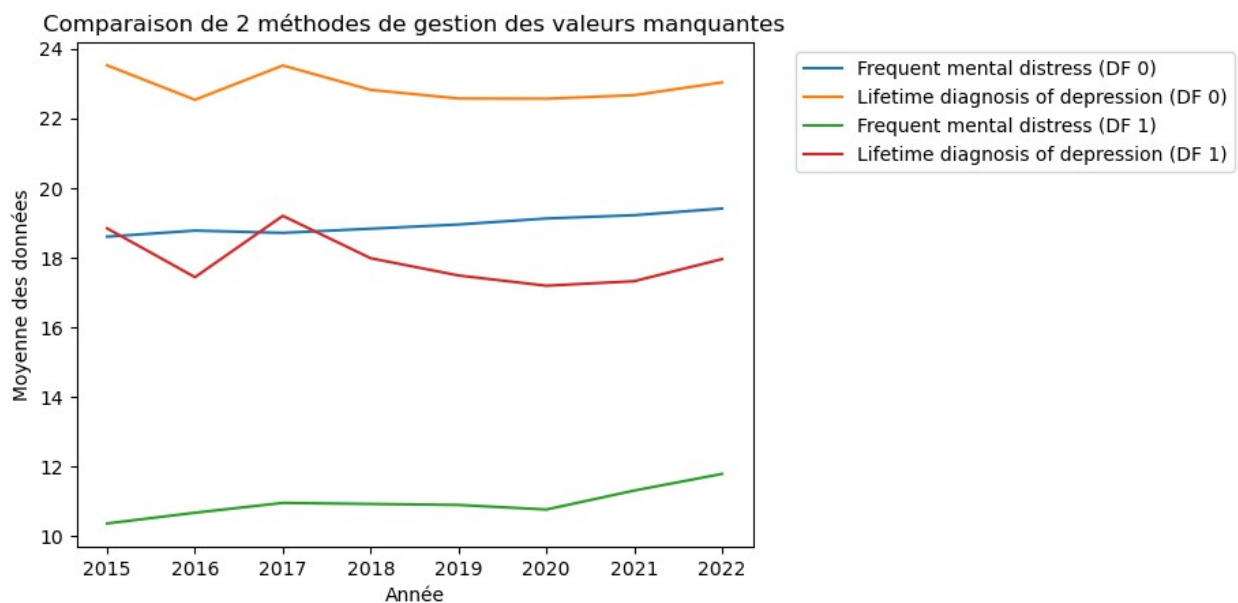
Il est temps de comparer la suppression des données manquantes vs leur remplacement par la médiane. Pour cela, affichez les valeurs moyennes de `Data_Value` par année, pour chaque groupe et chaque topic. (3 points)

```
#TODO:
import matplotlib.pyplot as plt
plt.figure( facecolor='white')
```

```

for i,mean in enumerate([mean_filled , mean_dropped]) :
    mean = mean.reset_index()
    for topic in mean['Topic'].unique():
        subset = mean[mean['Topic'] == topic]
        plt.plot(subset['Year'], subset['Data_Value'], linestyle='--',
label=topic+f" (DF {i})")
plt.xlabel('Année')
plt.ylabel('Moyenne des données')
plt.title('Comparaison de 2 méthodes de gestion des valeurs
manquantes')
plt.legend( bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()

```



### 3. LIVRABLES

Vous devez remettre sur Moodle un fichier compressé .zip contenant :

1) Le code : Un Jupyter notebook en Python qui contient le code tel implanté avec les librairies minimales demandées pour ce TP (Python, Pandas, Matplotlib). Le code doit être exécutable sans erreur et accompagné des commentaires appropriés dans le notebook de manière. Tous vos résultats doivent être reproductibles avec le code dans le notebook. *Attention, en aucun cas votre code ne doit avoir été copié de d'ailleurs.*

2) Un fichier pdf représentant votre notebook complètement exécuté sous format pdf (obtenu via latex ou imprimé en pdf avec le navigateur). Assurez-vous que le PDF est entièrement lisible. [Tutoriel youtube](#)

ATTENTION: assurez-vous que votre fichier compressé .zip ne dépasse pas la taille limite acceptée sur Moodle.

**ÉVALUATION** Votre TP sera évalué sur les points suivants :

### Critères :

1. Implantation correcte et efficace
2. Qualité du code (noms significatifs, structure, performance, gestion d'exception, etc.) (1 point)
3. Réponses correctes/sensées aux questions de réflexion ou d'analyse

### CODE D'HONNEUR

- **Règle 1:** Le plagiat de code est bien évidemment interdit. Toute utilisation de code doit être référencée adéquatement. Vous **ne pouvez pas** soumettre un code, écrit par quelqu'un d'autre. Dans le cas contraire, cela sera considéré comme du plagiat.
- **Règle 2:** Vous êtes libres de discuter avec d'autres équipes. Cependant, vous ne pouvez en aucun cas incorporer leur code dans votre TP.
- **Règle 3:** Vous ne pouvez pas partager votre code publiquement (par exemple, dans un dépôt GitHub public) tant que le cours n'est pas fini.

### Conversion en PDF sur Google Colab

```
%%capture
!sudo apt-get install texlive-xetex texlive-fonts-recommended texlive-plain-generic
```

Assurez vous d'avoir téléchargé le TP complété en notebook sur votre ordinateur, puis importé ce fichier dans le répertoire "content" avant de rouler la ligne suivante.

```
!jupyter nbconvert --to pdf /content/TP1.ipynb

This application is used to convert notebook files (*.ipynb)
to various other formats.

WARNING: THE COMMANDLINE INTERFACE MAY CHANGE IN FUTURE
RELEASES.

Options
=====
The options below are convenience aliases to configurable class-
options,
as listed in the "Equivalent to" description-line of the aliases.
To see all configurable class-options for some <cmd>, use:
    <cmd> --help-all

--debug
    set log level to logging.DEBUG (maximize logging output)
    Equivalent to: [--Application.log_level=10]
--show-config
    Show the application's configuration (human-readable format)
    Equivalent to: [--Application.show_config=True]
--show-config-json
    Show the application's configuration (json format)
```