Analyse des méthodes de fusion

Outer join

Cette méthode fusionne les colonnes des deux tables même si elles ne correspondent pas. Dans notre cas cela nous permettra de pouvoir garder les colonnes présentes dans table2.csv mais qui ne sont pas présentes dans table1.csv comme la date de naissance. Si une valeur n'existe pas dans l'une des tables, les colonnes correspondantes contiendront des valeurs manquantes (NaN).

Avantages :

- **Conservation maximale des données** : Aucune donnée n'est perdue. On garde les informations à la fois des démissionnaires et des sociétaires actuels, qu'ils apparaissent dans une table ou dans les deux.
- Cela garantit de ne pas exclure des démissionnaires ou sociétaires potentiels qui sont présents dans l'une des tables, mais pas dans l'autre sachant qu'il n'y a pas de correspondance concernant leur id.

Inconvénients :

- **Données manquantes** : La table fusionnée peut être plus grande car elle contient potentiellement beaucoup de données manquantes (NaN).

Inner join

Description :

- Cette méthode fusionne uniquement les lignes qui ont une correspondance dans les deux tables selon le nom des colonnes. Les lignes qui ne correspondent pas dans les deux tables sont exclues.

Avantages :

- **Table plus concise** : La taille de la table résultante est réduite car seules les lignes communes aux deux tables sont conservées.
- **Pas de valeurs manquantes** : Aucune ligne avec des valeurs NaN, ce qui simplifie le nettoyage et l'analyse des données.

Inconvénients :

- **Perte d'informations** : Il y a forcément une perte des données sur des démissionnaires ou sociétaires actuels qui n'ont pas d'enregistrement correspondant dans les deux tables.
- **Non adaptée à notre besoin** : On souhaite garder des informations sur tous les démissionnaires et sociétaires actuels, cette méthode exclurait potentiellement des données importantes.

Left join

Description :

- Cette méthode garde toutes les lignes de la **première table** (ici `table1`, démissionnaires) et les complète avec les données correspondantes de la **deuxième table** (`table2`, sociétaires + démissionnaires). Si une ligne n'a pas de correspondance dans la deuxième table, elle aura des valeurs NaN dans les colonnes de cette table.

Avantages :

- **Conservation des démissionnaires** : Toutes les informations sur les démissionnaires sont préservées, même si elles n'ont pas de correspondance dans la table des sociétaires.
- **Pratique puisque table1 est plus importante** : `table1` contient les données principales (démissionnaires), cela permet de ne pas perdre ces données.

Inconvénients :

- **Potentielles valeurs manquantes** : Si beaucoup de lignes de `table1` ne trouvent pas de correspondance dans `table2`, il y aura des NaN dans certaines colonnes.
- **Perte des informations des sociétaires non démissionnaires** : Les sociétaires actuels qui ne sont pas dans `table1` (démissionnaires) seront perdus alors que nous en avons grandement besoin pour pouvoir prédire leur démission.

Right join

Description :

- L’inverse de la jointure gauche : elle conserve toutes les lignes de la **deuxième table** (`table2`, sociétaires + démissionnaires) et les complète avec les données correspondantes de la **première table** (`table1`, démissionnaires). Les lignes de `table2` sans correspondance dans `table1` auront des NaN.

Avantages :

- **Conservation des sociétaires** : Toutes les informations sur les sociétaires sont préservées, même si elles n'ont pas de correspondance dans `table1`.
- **Pratique si table2 est plus importante** : Si `table2` contient les données principales (sociétaires actuels), cela permet de ne pas perdre ces données.

Inconvénients :

- **Potentielles valeurs manquantes** : Comme pour la jointure gauche, il peut y avoir des NaN si les lignes ne trouvent pas de correspondance dans `table1`.
- **Perte des informations des démissionnaires non présents dans `table2`**.

Phase intermédiaire

A cet instant, nous hésitons entre deux méthodes de jointure qui nous semble pertinentes, entre autre :

- Outer join
- Left join
- Pas de fusion

Nous avons décidé de réaliser les implémentations de ces deux méthodes afin de pouvoir comparer les résultats et ainsi prendre la décision adéquate.

Méthode 1 (Outer join)

45025	4	3390	0	C	29/07/2002	2		31/12/1900			10		11613.0	03/04/1978	
45026	4	3441	1	U	30/07/2003	2		31/12/1900			10		12346.0	21/01/1960	
45027	4	3452	0	C	28/08/1989	2		31/12/1900			10		5189.0	12/04/1947	
45028	4	3452	2	U	13/08/2001	2		31/12/1900			10		10794.0	19/12/1952	
45029	4	3581	0	C	09/07/1971	2		31/12/1900			10		161.0	15/11/1948	
45030	4	3608	3	C	06/03/2006	2		31/12/1900			10		14140.0	15/05/1976	
45031	4	3658	0	C	13/08/1997	0	2.0	26/12/2005	DV		21	30.0			
45032	4	3735	0	C	29/09/1999	0	2.0	18/12/2003	DV		21	34.0			
45033	4	3767	0	C	05/10/2001	2		31/12/1900			10		10908.0	07/07/1963	
45034	4	3779	0	C	21/05/1979	2		31/12/1900			10		195.0	02/08/1949	
45035	4	3798	0	C	11/12/1978	2		31/12/1900			10		2897.0	29/08/1941	
45036	4	4087	0	U	22/06/1992	0		31/12/1900			10		6127.0	20/05/1956	
45037	4	4152	1	U	28/03/1997	2		31/12/1900			10		7922.0	12/10/1954	
45038	4	4300	0	C	18/04/1977	2		31/12/1900			10		9.0	26/09/1947	
45039	4	4360	1	C	15/11/1982	2		31/12/1900			10		3154.0	11/12/1962	
45040	4	4645	0	C	22/12/2003	2		31/12/1900			10		12649.0	31/03/1948	
45041	4	5263	3	D	23/11/2006	0		31/12/1900			25		14624.0	03/11/1953	
45042	4	5968	0	C	18/11/1999	2		31/12/1900			10		9420.0	23/07/1958	
45043	4	15244	0	C	01/10/1986	0	2.0	09/07/2004	DV		21	31.0			
45044	4	15244	0	C	31/08/1981	0	2.0	17/06/2005	DV		21	30.0			
45045	4	18300	1	C	29/03/2001	2		31/12/1900			10		10546.0	21/09/1968	
45046	4	28135	1	C	04/04/2001	2		31/12/1900			10		10557.0	09/12/1959	
45047															
45048															
45049															
45050															
45051															

Cette méthode est la plus pertinente car elle permet d'obtenir toutes les lignes des deux tables, certes avec plus de valeur manquantes. Ceci est normal car le schéma des deux tables est différent. Cependant nous n'avons pas de perte de données avec cette méthode, ce qui est crucial dans notre analyse.

Méthode 2 (Left join)

30310	2	0	0	M	02/09/1983	0	2	02/08/2004	DA	21	33			
30311	3	949	0	A	11/10/1983	0	2	09/09/2002	DA	21	40			
30312	2	0	0	M	19/10/1983	0	2	27/07/2004	DA	21	33			
30313	3	0	0	A	20/12/1983	0	2	07/03/2000	DA	21	39			
30314	2	0	0	A	21/12/1983	0	2	14/10/2004	DA	21	39			
30315	3	0	1	M	20/01/1984	0	2	02/01/2003	DA	21	34			
30316	2	3975	0	M	29/12/1983	0	2	21/07/2004	DA	21	43			
30317	2	0	0	M	10/01/1984	0	2	10/01/2006	DA	21	37			
30318	3	0	0	A	26/09/1983	0	2	07/07/2003	DV	21	42			
30319	2	0	0	A	17/08/1983	0	2	15/12/2003	DV	21	36	15017.0	03/10/1947	
30320	3	0	0	M	25/01/1980	0	2	28/01/2003	DA	21	36			
30321	3	0	0	M	23/01/1980	0	2	25/01/2006	DV	21	43			
30322	2	0	0	M	20/02/1980	0	2	21/12/2005	DV	21	46			
30323	3	0	0	M	24/06/1982	0	2	05/06/2000	DA	21	29			
30324	2	0	0	A	18/06/1980	0	2	11/10/2006	DV	21	34			
30325	3	0	0	A	06/06/1980	0	2	27/07/2004	DV	21	33			
30326	2	0	0	M	28/02/1980	0	2	13/12/2004	DV	21	50			
30327	3	0	0	A	27/03/1980	0	2	14/02/2000	DV	21	32			
30328	2	0	0	M	07/07/1982	0	2	10/05/2000	DV	21	30			
30329	2	0	0	M	16/06/1982	0	2	15/06/2005	DV	21	32			
30330	3	0	0	A	29/02/1980	0	2	27/06/2005	DV	21	36			
30331	3	0	0	A	17/09/1980	0	2	12/01/1999	DA	21	32			
30332	3	0	0	A	03/07/1980	0	2	03/06/1999	DV	21	31			
30333	2	0	0	A	29/07/1983	0	2	28/01/2005	RA	21	40			
30334														
30335														
30336														

Comme nous pouvons le voir, avec cette méthode nous avons gardé le même nombre de démissionnaires (uniquement ceux de la table 1) et avons fait une correspondance avec les colonnes présentes dans la table 2. Cette façon de procéder ne correspond pas à notre besoin car nous voulons à la fois fusionner les démissionnaires de la table 1 et de la table 2 sachant que leur id ne

correspondent pas et que ce sont des sociétaires distinct (pas de correspondance entre les attributs ID des deux tables).

Méthode 3 (Pas de fusion)

Cette méthode consiste en l’utilisation des données de la table1 pour entraîner le modèle avec les données des démissionnaires, avec donc 100% de démissionnaires pour maximiser la précision. La table2 sera utilisée pour effectuer des tests à partir du modèle que nous aurons entraîné au préalable.