Université Sorbonne Paris Cité	UNIVERSITE PARIS 13
Institut Galilée	Année Universitaire
Laboratoire d'Informatique de Paris Nord	2019 / 2020
La Data e(s)t le monde de demain	**F. Boufarès**
La Data/La Donnée	boufares@lipn.univ-paris13.fr
Bases de Données Avancées- Entrepôts de Données	**iDQMS→ Smart Data**

DATA MANGEMENT & MACHINE LEARNING, the future!
(DATA BASE, DATA WAREHOUSE, DATA MANAGER, DATA LAKE, … BIG DATA!)

Think DIFFERENTLY, BIGGER and SMARTER!

Votre mission, si vous l'acceptez, est l'excellence

« EID : The Excellence in Data Use ! »

Si vous échouez, nous nierons avoir eu connaissance de vos agissements !

EID : L'Excellence dans l'Investigation des Données → EDI : Excellence in Data Investigation

Vous êtes le TOUBIB-TABIB des données : Un **Data-Logue** tels que par exemple les Cardio-logues, les Pneumo-logues, les Uro-logues,...) !

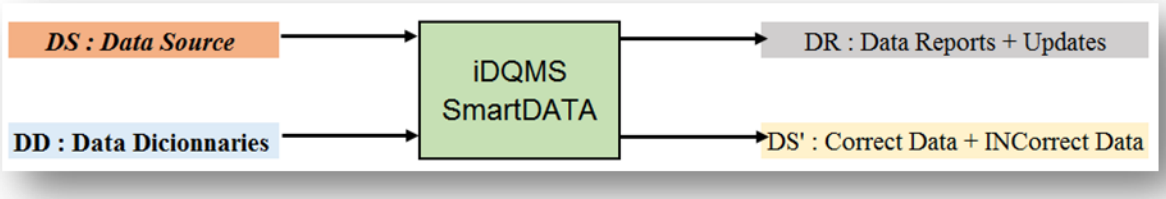
Vous disposez des outils (que vous avez à créer) qui vous permettent de DIAGNOSTIQUER les anomalies de votre patient la source de données DataSource !
Est-ce une mission possible ?! SiSi C PO CIBLE !!!

VOTRE MISSION SI VOUS L'ACCEPTEZ
EST DE DONNER UN SENS AUX DONNEES !
SI VOUS ECHOUEZ,
NOUS NIERONS AVOIR EU CONNAISSANCE DE VOS AGISSEMENTS !

Faites le diagnostic automatiquement qui vous permet :
- de détecter les erreurs (les anomalies), et ensuite,
- de corriger les anomalies

Vous êtes en train de développer un outil de qualité très intelligent : → >>>>> SmartDATA
Est-ce une mission impossible ?! SiSi C PO CIBLE !!!

iDQMS tool : An intelligent Data Quality Management System tool → >>>>> SmartDATA



DS : Data Source	Data with anomalies (Examples : **CSV** file, SQL **table**)
DD : Data Dictionnaries	Dictionnaries such as **DDRE** for Regular Expressions and **DDVS** for Valid Strings
DR : Data Reports	Reports that contain **diagnostics** (metrics & **mesures**), to help correcting INVALID DATA (**Cleaning**)
Updates	A set of actions to do to better correct Data (UPDATE DataSource SET ...=... WHERE ... ; Etc...)
Correct Data	VALID Records (with NO anomalies such as Heteregeneous Data, Null Value, Functional Dependency, Deduplication,...)
INCorrect Data	INVALID Records (with anomalies, at least one) DSWARNING

Etant donné une source de données au format CSV (**Comma-separated values**, connu sous le sigle **CSV**, est un format texte ouvert représentant des données tabulaires sous forme de valeurs séparées par des **virgules**), composée d'une seule « colonne » et de plusieurs lignes (enregistrements). Les valeurs, sur une ligne, sont séparées par le **séparateur « ; »** par exemple.

Le but est de l'éclater en plusieurs colonnes afin de la diagnostiquer (diagnostiquer chacune des colonnes), et donc de tenter d'homogénéiser les données, de détecter et ensuite de corriger d'éventuelles anomalies (**intra-colonne, inter-colonnes et inter-lignes**).

Plusieurs **rapports** sont à établir sur la source de données (DR : Les rapports sur la source de données ; The data source reports). Ils permettent d'établir, d'une part, les **diagnostics syntaxiques** des colonnes (The syntactic diagnosis of the columns), et d'autres part, les **diagnostics sémantiques** des colonnes (The semantic diagnosis of the columns).

Diagnostiquer chacune des colonnes pourrait aider à :

- **transformer et homogénéiser** le contenu des colonnes et

- **détecter** et ensuite de **corriger** d'éventuelles anomalies (intra-colonne, inter-colonnes et inter-lignes) : réaliser le **nettoyage** des données !

Les exemples ci-dessous permettent d'expliciter les **différentes étapes des diagnostics nécessaires** afin d'**aider à la détection et la correction des anomalies**. L'étape de l'analyse des données peut être appelée **profilage des données**.

Etant donné la source de la figure 1 ci-dessous.

	CSV Data Source
2	Id01;Alain;CLEMENT;21-02-1970;;B+;M;76,49 KG;France;Europe;73;32.14°C;11,51 m
3	Id02;Inès;MIGNONNE;;01-01-2020;A;F;40159,31 g;France;Europe;146,345;35;$1
4	Id03;Omar;BELLE;;?@;R;M;84,48 KG;France;Europe;39;26.93°C;14,35 m
5	Id04;Rayan;MIGNONNE;10-06-1961;rayan.mignonne@hotmail.com;A-;M;61,49 KG;Itale;;10;20.9°C;17,63 m
6	Id05;Eve;AIMANT;;eve.aimant@gmail.com;1;F;52606,37 g;Royaume-Uni;Europe;55,237;100.1°F;650,9 cm
7	Id06;Alain;MOUSTAFA;26-12-1969;alain.moustafa@yahoo.fr;++;M;76,49 KG;Argentine;Amérique;60;26.93°C;17,63 m
8	Id07;Sabrine;PRINTEMPS;14-novembre-1962;;sabrine.printemps@hotmail.com;AB+;F;54561,49g;Royaume-Uni;Europe;96,409;91.1°F;534,27 cm
9	Id08;Clément;DELOIN;14-11-1962;clément.deloin@gmail.com;AB+;A+;71,68 KG;Tunisie;Afrique;69;38.05°C;17,37 m
10	Id09;Emna;AIMANT;14-avril-1970;emna.aimant@hotmail.com;++;F;66360,8 g;Allemagne;Europe;109,344;87.53°F;424,22 cm
11	Id10;Claire;EXCELLE;02-mars-1970;claire.excelle@gmail.com;AB-;F;46082,53 g;Tunisie;Afrique;50,491;;744,55 cm
12	Id11;Mamadou;MOUSTAFA;02-03-1970;mamadou.moustafa@gmail.com;A+;M;58,83 KG;Algérie;Afrique;8;26.26°C;14,35 m
13	Id12;Jean;JOLIE;28-12-1978;01-01-2020;AB+;M;75,34 KG;Belgique;Europe;13;32.14°C;14,35 m
14	Id13;Faouzi;GRANDE;31-12-1982;faouzi.grande@yahoo.fr;++;M;71,68 KG;Tunisie;Afrique;27;26.26°C;9,58 m
15	Id14;Alexandre;FORT;31-01-1999;;B+;M;71,68 KG;Tunisie;Afrique;46;26.93°C;1
16	Id15;Clément;PARIS;;clément.paris@yahoo.fr;;0;61,49 KG;Itale;;60;20.9°C;0
17	Id16;Alexandre;JOLIE;;alexandre.jolie@gmail.com;B+;M;61,49 KG;Qatar;Asie;41;;1
18	Id17;Mamadou;JOLIE;31-01-1999;01-01-2020;++;M;76,49 KG;Chine;Asie;41;26.26°C;14,2 m
19	Id18;Alain;PARIS;21-02-1970;alain.paris@gmail.com;;M;0;Brésil;Amérique;51;26.93°C;17 m
20	Id19;Jean;FORT;14-04-1970;jean.fort@yahoo.fr;O+;M;76,49 KG;Algérie;Afrique;43;20.9°C;17 m
21	Id20;Médecin;FORT;;médecin.fort@hotmail.com;++;M;61,49 KG;Franc;;1;26.26°C;11,51 m
22	Id21;Emna;BON;31-janvier-1999;emna.bon@yahoo.fr;AB;F;66360,8 g;Brésil;Amérique;78,593;;530,6 cm
23	Id22;Ibrahim;MIGNONNE;07-04-1989;ibrahim.mignonne@gmail.com;;0;76,49 KG;France;Europe;23;32.14°C;19,85 m
24	Id23;Maria;MIGNONNE;31-janvier-1999;1;AB+;F;64873,28 g;Qatar;Asie;79,33;;530,6 cm
25	Id24;Adam;SPORTIF;26-12-1969;01-01-2020;A-;M;61,49 KG;Espagne;;56;32.14°C;11,51 m
26	Id25;Alexandre;SOLEIL;14-04-1970;alexandre.soleil@yahoo.fr;A+;M;84,48 KG;Japon;Asie;70;15.67°C;14,35 m
27	Id26;Omar;MOUSTAFA;31-12-1982;omar.moustafa@hotmail.com;++;M;1;Canada;Amérique;39;;0
28	Id27;Clémence;PRINTEMPS;03-avril-1970;clémence.printemps@gmail.com;;F;64873,28 g;Italie;Europe;45,291;87.53°F;0
29	Id28;Kenza;TROPFOR;10-juin-1961;kenza.tropfor@hotmail.com;A+;F;52606,37 g;Tunisie;;40,778;111.29°F;530,6 cm
30	Id29;Rayan;TROPFOR;14-11-1962;rayan.tropfor@hotmail.com;++;M;63,3 KG;Qatar;Asie;66;26.93°C;0
31	Id30;Alain;INFORME;;alain.informe@gmail.com;AB+;M;78,17 KG;Itale;;24;41.44 °c;17,37 m

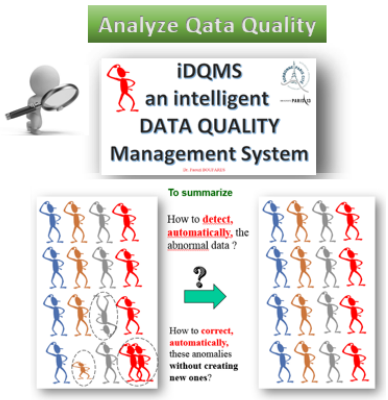


Figure 1 : La source de données **DS** (**CSVFile**$_j$, Une seule colonne)

L'éclatement de la source (un fichier CSV, une table d'une seule colonne) permet d'avoir la table suivante, de nom **CSV2TABCOLUMNS**, de la figure 2.

	CSV Data Source, Only one column	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9	Col10	Col11	Col12	Col13
2	Id01;Alain;CLEMENT;21-02-1970;;B+;M;76,49 KG;France;Europe;73;32.14°	Id01	Alain	CLEMENT	21/02/1970		B+	M	76,49 KG	France	Europe	73	32.14°C	11,51 m
3	Id02;Inès;MIGNONNE;;01-01-2020;A;F;40159,31 g;France;Europe;146,345	Id02	Inès	MIGNONNE		01/01/2020	A	F	40159,31 g	France	Europe	146,345	35	$1
4	Id03;Omar;BELLE;;?@;R;M;84,48 KG;France;Europe;39;26.93°C;14,35 m	Id03	Omar	BELLE		?@	R	M	84,48 KG	France	Europe	39	26.93°C	14,35 m
5	Id04;Rayan;MIGNONNE;10-06-1961;rayan.mignonne@hotmail.com;A-;M	Id04	Rayan	MIGNONNE	10/06/1961	rayan.mignonne@hotmail.com	A-	M	61,49 KG	Itale		10	20.9°C	17,63 m
6	Id05;Eve;AIMANT;;eve.aimant@gmail.com;1;F;52606,37 g;Royaume-Uni;E	Id05	Eve	AIMANT		eve.aimant@gmail.com	1	F	52606,37 g	Royaume-Uı	Europe	55,237	100.1°F	650,9 cm
7	Id06;Alain;MOUSTAFA;26-12-1969;alain.moustafa@yahoo.fr;++;M;76,49	Id06	Alain	MOUSTAFA	26/12/1969	alain.moustafa@yahoo.fr	++	M	76,49 KG	Argentine	Amérique	60	26.93°C	17,63 m
8	Id07;Sabrine;PRINTEMPS;14-novembre-1962;;sabrine.printemps@hotmail.com;	Id07	Sabrine	PRINTEMPS	14-nov-62	sabrine.printemps@hotmail.com	AB+	F	54561,49g	Royaume-Uı	Europe	96,409	91.1°F	534,27 cm
9	Id08;Clément;DELOIN;14-11-1962;clément.deloin@gmail.com;AB+;A+;7	Id08	Clément	DELOIN	14/11/1962	clément.deloin@gmail.com	AB+	A+	71,68 KG	Tunisie	Afrique	69	38.05°C	17,37 m
10	Id09;Emna;AIMANT;14-avril-1970;emna.aimant@hotmail.com;++;F;663	Id09	Emna	AIMANT	14-avr-70	emna.aimant@hotmail.com	++	F	66360,8 g	Allemagne	Europe	109,344	87.53°F	424,22 cm
11	Id10;Claire;EXCELLE;02-mars-1970;claire.excelle@gmail.com;AB-;F;4608	Id10	Claire	EXCELLE	02-mars-70	claire.excelle@gmail.com	AB-	F	46082,53 g	Tunisie	Afrique	50,491		744,55 cm
12	Id11;Mamadou;MOUSTAFA;02-03-1970;mamadou.moustafa@gmail.co	Id11	Mamadou	MOUSTAFA	02/03/1970	mamadou.moustafa@gmail.com	A+	M	58,83 KG	Algérie	Afrique	8	26.26°C	14,35 m
13	Id12;Jean;JOLIE;28-12-1978;01-01-2020;AB+;M;75,34 KG;Belgique;Europe	Id12	Jean	JOLIE	28/12/1978	01/01/2020	AB+	M	75,34 KG	Belgique	Europe	13	32.14°C	14,35 m
14	Id13;Faouzi;GRANDE;31-12-1982;faouzi.grande@yahoo.fr;++;M;71,68 KG	Id13	Faouzi	GRANDE	31/12/1982	faouzi.grande@yahoo.fr	++	M	71,68 KG	Tunisie	Afrique	27	26.26°C	9,58 m
15	Id14;Alexandre;FORT;31-01-1999;;B+;M;71,68 KG;Tunisie;Afrique;46;26.9	Id14	Alexandre	FORT	31/01/1999		B+	M	71,68 KG	Tunisie	Afrique	46	26.93°C	1
16	Id15;Clément;PARIS;;clément.paris@yahoo.fr;;0;61,49 KG;Itale;;60;20.9°	Id15	Clément	PARIS		clément.paris@yahoo.fr		0	61,49 KG	Itale		60	20.9°C	0
17	Id16;Alexandre;JOLIE;;alexandre.jolie@gmail.com;B+;M;61,49 KG;Qata	Id16	Alexandre	JOLIE		alexandre.jolie@gmail.com	B+	M	61,49 KG	Qatar	Asie	41		1
18	Id17;Mamadou;JOLIE;31-01-1999;01-01-2020;++;M;76,49 KG;Chine;Asie;4	Id17	Mamadou	JOLIE	31/01/1999	01/01/2020	++	M	76,49 KG	Chine	Asie	41	26.26°C	14,2 m
19	Id18;Alain;PARIS;21-02-1970;alain.paris@gmail.com;;M;0;Brésil;Amérique	Id18	Alain	PARIS	21/02/1970	alain.paris@gmail.com		M	0	Brésil	Amérique	51	26.93°C	17 m
20	Id19;Jean;FORT;14-04-1970;jean.fort@yahoo.fr;O+;M;76,49 KG;Algérie;A	Id19	Jean	FORT	14/04/1970	jean.fort@yahoo.fr	O+	M	76,49 KG	Algérie	Afrique	43	20.9°C	17 m
21	Id20;Médecin;FORT;;médecin.fort@hotmail.com;++;M;61,49 KG;Franc;;1	Id20	Médecin	FORT		médecin.fort@hotmail.com	++	M	61,49 KG	Franc		16	26.26°C	11,51 m
22	Id21;Emna;BON;31-janvier-1999;emna.bon@yahoo.fr;AB;F;66360,8 g;Brés	Id21	Emna	BON	31-janv-99	emna.bon@yahoo.fr	AB	F	66360,8 g	Brésil	Amérique	78,593		530,6 cm
23	Id22;Ibrahim;MIGNONNE;07-04-1989;ibrahim.mignonne@gmail.com;;0	Id22	Ibrahim	MIGNONNE	07/04/1989	ibrahim.mignonne@gmail.com		0	76,49 KG	France	Europe	23	32.14°C	19,85 m
24	Id23;Maria;MIGNONNE;31-janvier-1999;1;AB+;F;64873,28 g;Qatar;Asie;7	Id23	Maria	MIGNONNE	31-janv-99	1	AB+	F	64873,28 g	Qatar	Asie	79,33		530,6 cm
25	Id24;Adam;SPORTIF;26-12-1969;01-01-2020;A-;M;61,49 KG;Espagne;;56;3	Id24	Adam	SPORTIF	26/12/1969	01/01/2020	A-	M	61,49 KG	Espagne		56	32.14°C	11,51 m
26	Id25;Alexandre;SOLEIL;14-04-1970;alexandre.soleil@yahoo.fr;A+;M;84,4	Id25	Alexandre	SOLEIL	14/04/1970	alexandre.soleil@yahoo.fr	A+	M	84,48 KG	Japon	Asie	70	15.67°C	14,35 m
27	Id26;Omar;MOUSTAFA;31-12-1982;omar.moustafa@hotmail.com;++;M;1	Id26	Omar	MOUSTAFA	31/12/1982	omar.moustafa@hotmail.com	++	M	1	Canada	Amérique	39		0
28	Id27;Clémence;PRINTEMPS;03-avril-1970;clémence.printemps@gmail.c	Id27	Clémence	PRINTEMPS	03-avr-70	clémence.printemps@gmail.com		F	64873,28 g	Italie	Europe	45,291	87.53°F	0
29	Id28;Kenza;TROPFOR;10-juin-1961;kenza.tropfor@hotmail.com;A+;F;52	Id28	Kenza	TROPFOR	10-juin-61	kenza.tropfor@hotmail.com	A+	F	52606,37 g	Tunisie		40,778	111.29°F	530,6 cm
30	Id29;Rayan;TROPFOR;14-11-1962;rayan.tropfor@hotmail.com;++;M;63,3	Id29	Rayan	TROPFOR	14/11/1962	rayan.tropfor@hotmail.com	++	M	63,3 KG	Qatar	Asie	66	26.93°C	0
31	Id30;Alain;INFORME;;alain.informe@gmail.com;AB+;M;78,17 KG;Itale;;24;41.44	Id30	Alain	INFORME		alain.informe@gmail.com	AB+	M	78,17 KG	Itale		24	41.44 °c	17,37 m

Figure 2 : La source de données éclatées en une table de plusieurs colonnes
Le résultat est stocké dans une table de nom : **CSV2TABCOLUMNS**

Soit N le nombre de lignes dans la source
Soit Pv_i = le nombre de « séparateur ; » pour la ligne i
Le nombre de colonnes = **Max(Pv_i) + 1** pour i de 1 à N

En faisant un « **ZOOM** » sur chaque colonne (Col_i, i de 1 à N) de la nouvelle source éclatée **CSV2TABCOLUMNS**, il est possible de détecter dans un premier temps des anomalies que l'on peut qualifier de **syntaxiques** pour ensuite analyser **sémantiquement** le contenu.

1. Profilage SYNTAXIQUE

Par exemple, pour la colonne 4 [A DATE], il est possible de dresser le bilan syntaxique ci-dessous présenté dans les figures 3 et 4. Un rapport sur les données, de nom **DR_CSVFile_Col_4**, est stocké selon le **modèle de la figure 4**.

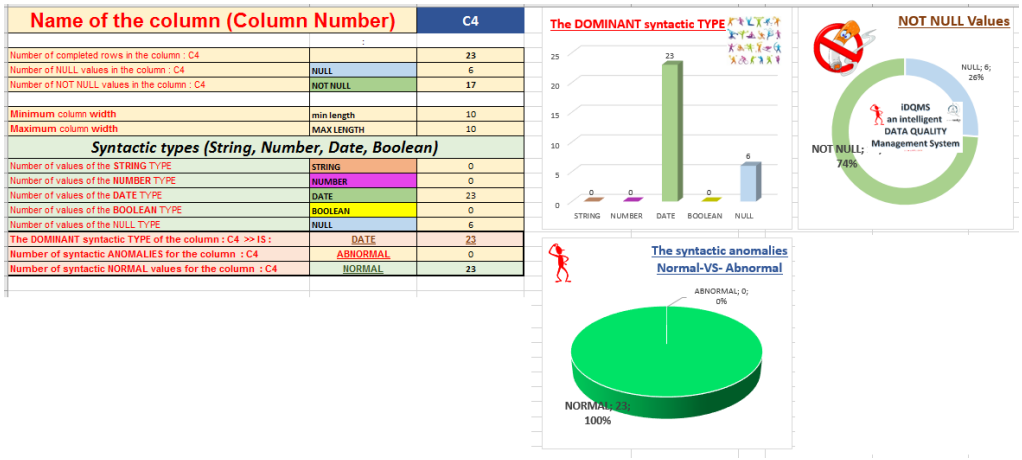


Figure 3 : Profilage/Analyse syntaxique de la colonne 4 (Colonne entière)

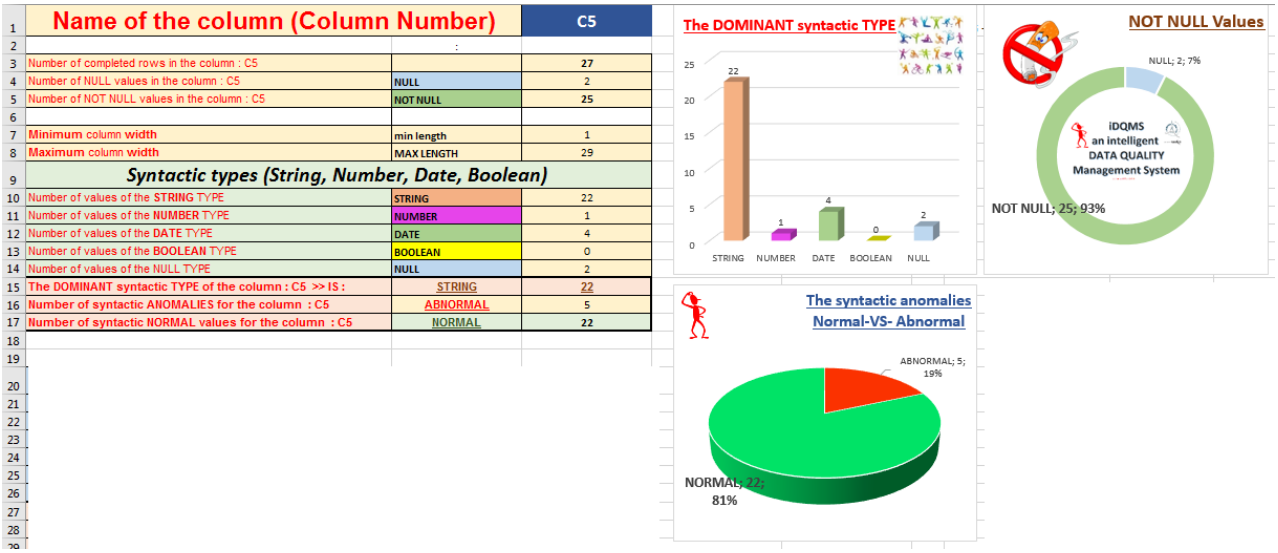
THE STARTING DATA	SYNTACTIC ANALYSIS			Observation	THE TRANSFORMED DATA
OLD VALUES	SYNTAXIC TYPE	COLUMN WIDTH	NUMBER OF WORDS	SYNTACTIC ANOMALY	NEW VALUES
21/02/1970	DATE	10	1		1970-02-21
	NULL	NULL	0	|<!>NULL	|<!>NULL
	NULL	NULL	0	|<!>NULL	|<!>NULL
10/06/1961	DATE	10	1		1961-06-10
	NULL	NULL	0	|<!>NULL	|<!>NULL
26/12/1969	DATE	10	1		1969-12-26
14-nov-62	DATE	10	1		1962-11-14
14/11/1962	DATE	10	1		1962-11-14
14-avr-70	DATE	10	1		1970-04-14
02-mars-70	DATE	10	1		1970-03-02
02/03/1970	DATE	10	1		1970-03-02
28/12/1978	DATE	10	1		1978-12-28
31/12/1982	DATE	10	1		1982-12-31
31/01/1999	DATE	10	1		1999-01-31
	NULL	NULL	0	|<!>NULL	|<!>NULL
	NULL	NULL	0	|<!>NULL	|<!>NULL
31/01/1999	DATE	10	1		1999-01-31
21/02/1970	DATE	10	1		1970-02-21
14/04/1970	DATE	10	1		1970-04-14
	NULL	NULL	0	|<!>NULL	|<!>NULL
31-janv-99	DATE	10	1		1999-01-31
07/04/1989	DATE	10	1		1989-04-07
31-janv-99	DATE	10	1		1999-01-31
26/12/1969	DATE	10	1		1969-12-26
14/04/1970	DATE	10	1		1970-04-14
31/12/1982	DATE	10	1		1982-12-31
03-avr-70	DATE	10	1		1970-04-03
10-juin-61	DATE	10	1		1961-06-10
14/11/1962	DATE	10	1		1962-11-14

Figure 4 : Profilage/Analyse syntaxique de la colonne 4 (Ligne par ligne)

Remarque : Homogénéisation du type DATE !

Les données de type date sont transformées en un **format par défaut** :
Année-Mois-Jour AAAA-MM-JJ / Year-Month-Day / **YYYY-MM-DD**

Par exemple, pour la colonne 5 [AN EMAIL], il est possible de dresser le bilan syntaxique ci-dessous présenté dans les figures 5 et 6. Un rapport sur les données, de nom **DR_CSVFile_Col_5**, est stocké selon le **modèle de la figure 6**.



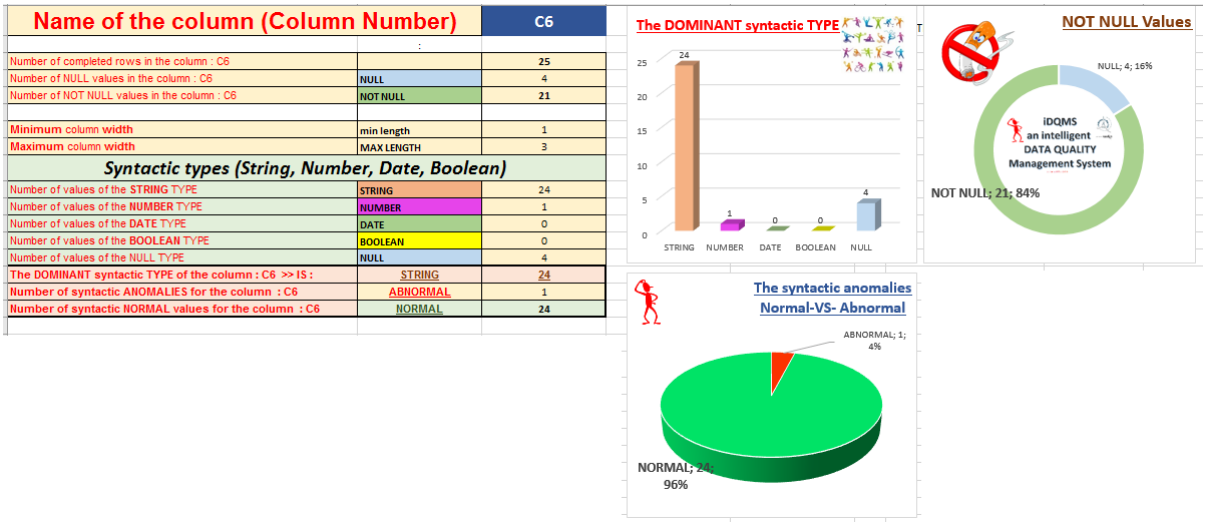
Name of the column (Column Number)		C5
Number of completed rows in the column : C5		27
Number of NULL values in the column : C5	NULL	2
Number of NOT NULL values in the column : C5	NOT NULL	25
Minimum column width	min length	1
Maximum column width	MAX LENGTH	29
Syntactic types (String, Number, Date, Boolean)		
Number of values of the STRING TYPE	STRING	22
Number of values of the NUMBER TYPE	NUMBER	1
Number of values of the DATE TYPE	DATE	4
Number of values of the BOOLEAN TYPE	BOOLEAN	0
Number of values of the NULL TYPE	NULL	2
The DOMINANT syntactic TYPE of the column : C5 >> IS :	STRING	22
Number of syntactic ANOMALIES for the column : C5	ABNORMAL	5
Number of syntactic NORMAL values for the column : C5	NORMAL	22

Figure 5 : Profilage/Analyse syntaxique de la colonne 5 (Colonne entière)

THE STARTING DATA	SYNTACTIC ANALYSIS			Observation	THE TRANSFORMED DATA
OLD VALUES	SYNTAXIC TYPE	COLUMN WIDTH	NUMBER OF WORDS	SYNTACTIC ANOMALY	NEW VALUES
	NULL	NULL	0	|<!>NULL	|<!>NULL
01/01/2020	DATE	10	1	01/01/2020|<!?!>ANOMALY	01/01/2020|<!?!>ANOMALY
?@	STRING	2	1		?@
rayan.mignonne@hotmail.com	STRING	26	1		rayan.mignonne@hotmail.com
eve.aimant@gmail.com	STRING	20	1		eve.aimant@gmail.com
alain.moustafa@yahoo.fr	STRING	23	1		alain.moustafa@yahoo.fr
sabrine.printemps@hotmail.com	STRING	29	1		sabrine.printemps@hotmail.com
clément.deloin@gmail.com	STRING	24	1		clément.deloin@gmail.com
emna.aimant@hotmail.com	STRING	23	1		emna.aimant@hotmail.com
claire.excelle@gmail.com	STRING	24	1		claire.excelle@gmail.com
mamadou.moustafa@gmail.com	STRING	26	1		mamadou.moustafa@gmail.com
01/01/2020	DATE	10	1	01/01/2020|<!?!>ANOMALY	01/01/2020|<!?!>ANOMALY
faouzi.grande@yahoo.fr	STRING	22	1		faouzi.grande@yahoo.fr
	NULL	NULL	0	|<!>NULL	|<!>NULL
clément.paris@yahoo.fr	STRING	22	1		clément.paris@yahoo.fr
alexandre.jolie@gmail.com	STRING	25	1		alexandre.jolie@gmail.com
01/01/2020	DATE	10	1	01/01/2020|<!?!>ANOMALY	01/01/2020|<!?!>ANOMALY
alain.paris@gmail.com	STRING	21	1		alain.paris@gmail.com
jean.fort@yahoo.fr	STRING	18	1		jean.fort@yahoo.fr
médecin.fort@hotmail.com	STRING	24	1		médecin.fort@hotmail.com
emna.bon@yahoo.fr	STRING	17	1		emna.bon@yahoo.fr
ibrahim.mignonne@gmail.com	STRING	26	1		ibrahim.mignonne@gmail.com
1	NUMBER	1	1	1|<!?!>ANOMALY	1|<!?!>ANOMALY
01/01/2020	DATE	10	1	01/01/2020|<!?!>ANOMALY	01/01/2020|<!?!>ANOMALY
alexandre.soleil@yahoo.fr	STRING	25	1		alexandre.soleil@yahoo.fr
omar.moustafa@hotmail.com	STRING	25	1		omar.moustafa@hotmail.com
clémence.printemps@gmail.com	STRING	28	1		clémence.printemps@gmail.com
kenza.tropfor@hotmail.com	STRING	25	1		kenza.tropfor@hotmail.com
rayan.tropfor@hotmail.com	STRING	25	1		rayan.tropfor@hotmail.com

Figure 6 : Profilage/Analyse syntaxique de la colonne 5 (Ligne par ligne)

Par exemple, pour la colonne 6 [A BLOOD GROUP], il est possible de dresser le bilan syntaxique ci-dessous présenté dans les figures 7 et 8. Un rapport sur les données, de nom **DR_CSVFile_Col_6**, est stocké selon le **modèle de la figure 8**.



Name of the column (Column Number)		C6
	:	
Number of completed rows in the column : C6		25
Number of NULL values in the column : C6	NULL	4
Number of NOT NULL values in the column : C6	NOT NULL	21
Minimum column width	min length	1
Maximum column width	MAX LENGTH	3
Syntactic types (String, Number, Date, Boolean)		
Number of values of the STRING TYPE	STRING	24
Number of values of the NUMBER TYPE	NUMBER	1
Number of values of the DATE TYPE	DATE	0
Number of values of the BOOLEAN TYPE	BOOLEAN	0
Number of values of the NULL TYPE	NULL	4
The DOMINANT syntactic TYPE of the column : C6 >> IS :	STRING	24
Number of syntactic ANOMALIES for the column : C6	ABNORMAL	1
Number of syntactic NORMAL values for the column : C6	NORMAL	24

Figure 7 : Profilage/Analyse syntaxique de la colonne 6 (Colonne entière)

THE STARTING DATA	SYNTACTIC ANALYSIS			Observation	THE TRANSFORMED DATA
OLD VALUES	SYNTAXIC TYPE	COLUMN WIDTH	Number Of Words	SYNTACTIC ANOMALY	NEW VALUES
B+	STRING	2	1		B+
A	STRING	1	1		A
R	STRING	1	1		R
A-	STRING	2	1		A-
1	NUMBER	1	1	1 |<!?!>ANOMALY	1 |<!?!>ANOMALY
++	STRING	2	1		++
AB+	STRING	3	1		AB+
AB+	STRING	3	1		AB+
++	STRING	2	1		++
AB-	STRING	3	1		AB-
A+	STRING	2	1		A+
AB+	STRING	3	1		AB+
++	STRING	2	1		++ SEE NEXT STEP ! >>> !!!
B+	STRING	2	1		B+
	NULL	NULL	0	|<!>NULL	|<!>NULL
B+	STRING	2	1		B+
++	STRING	2	1		++
	NULL	NULL	0	|<!>NULL	|<!>NULL
O+	STRING	2	1		O+
++	STRING	2	1		++
AB	STRING	2	1		AB
	NULL	NULL	0	|<!>NULL	|<!>NULL
AB+	STRING	3	1		AB+
A-	STRING	2	1		A-
A+	STRING	2	1		A+
++	STRING	2	1		++
	NULL	NULL	0	|<!>NULL	|<!>NULL
A+	STRING	2	1		A+
++	STRING	2	1		++

Figure 8 : Profilage/Analyse syntaxique de la colonne 6 (Ligne par ligne)

Le résultat des différentes étapes de profilage, réalisé pour chacune des colonnes de la source de données, est stocké dans une méta-table, un rapport sur les données, de nom **DR_CSVFile_TabCol**, est stocké selon le **modèle de la figure 9**.

1	OLD Name or Column Number	NEW Column Name	Number of rows	Number of NULL values	Number of NOT NULL values	min length (characters)	MAX LENGTH(characters)	Number of Words	Number of values of the STRING TYPE	Number of values of the NUMBER TYPE	Number of values of the DATE TYPE	Number of values of the BOOLEAN TYPE	Number of values of the NULL TYPE	Number of DIFFERENT values	The DOMINANT syntactic TYPE	Number of syntactic ANOMALIES	Number of syntactic NORMAL values		The DOMINANT semantic CATEGORY	M115…M200…Etc Mesures
2	OLDName	NEWName	M000	M100	M101	M102	M103	M104	M105	M106	M107	M108	M109	M110	M111	M112	M113		M114	
3	Col1	Col1_STRING__UNKNOWN	30	0	30	4	4								STRING					
4	Col2	Col2_STRING__FIRSTNAME	30	0	30	3	9								STRING					
5	Col3	Col3_STRING__UNKNOWN	30	0	30										STRING					
6	Col4	Col4_DATE__DATE_AAAAMMDD	30	6	24	10	10								DATE					
7	Col5	Col5_STRING__EMAIL	29	2	28	1	29	1	22	1	4	0	2		STRING	5	22			
8	Col6	Col6_STRING__BLOODGROUP	30	4	26										STRING					
9	Col7	Col7_STRING__GENDER	30												STRING					
10	Col8	Col8_STRING__WEIGHT_KG	30												NUMBER					
11	Col9	Col9_STRING__COUNTRY	30												STRING					
12	Col10	Col10_STRING__CONTINENT	30												STRING					
13	Col11	Col11_NUMBER__UNKNOWN	30												NUMBER					
14	Col12	Col12_STRING__TEMPERATURE	30												NUMBER					
15	Col13	Col13_STRING__MESURE_M	30												NUMBER					

Figure 9 : Résultat des différentes étapes de Profilage/Analyse syntaxique de toutes les colonnes

Profilage des données : Exemples des différentes mesures à effectuer

Data profiling : Examples of the different measurements to be made

OLD Name or Column Number	OLDName
NEW Column Name	NEWName
Number of rows	**M000**
Number of **NULL** values	**M100**
Number of **NOT NULL** values	**M101**
min length (characters)	**M102**
MAX LENGTH (characters)	**M103**
Number of **Words**	**M104**
Number of values of the **STRING TYPE**	**M105**
Number of values of the **NUMBER TYPE**	**M106**
Number of values of the **DATE TYPE**	**M107**
Number of values of the **BOOLEAN TYPE**	**M108**
Number of values of the **NULL TYPE**	**M109**
Number of **DIFFERENT values**	**M110**
The **DOMINANT syntactic TYPE**	**M111**
Number of **syntactic ANOMALIES**	**M112**
Number of **syntactic NORMAL values**	**M113**
The **DOMINANT semantic CATEGORY**	**M114**
Number of **semantic ANOMALIES**	**M115**
Number of **semantic NORMAL values**	**M116**
Etc…	**M200…**

SQL :

DROP TABLE **DR_CSVFile_Col_i** ;

CREATE TABLE **DR_CSVFile_Col_i**
(

REFERENCES	VARCHAR2(100),	-- NomDuFichier CSV_ DateSystème_Col$_i$
OLDVALUES	VARCHAR2(1000),	
SYNTACTICTYPE	VARCHAR2(6),	
COLUMNWIDHT	NUMBER(5),	
NUMBEROFWORDS	NUMBER(2),	
OBSERVATION	VARCHAR2(100),	
NEWVALUES	VARCHAR2(1000),	
SEMANTICCATEGORY	VARCHAR2(1000),	
SEMANTICSUBCATEGORY	VARCHAR2(1000)	

);

DROP TABLE **DR_CSVFile_TabCol ;**
CREATE TABLE **DR_CSVFile_TabCol**
(

REFERENCES	VARCHAR2(100),	-- NomDuFichier CSV_ DateSystème
OLDNAME	VARCHAR2(100),	
NEWNAME	VARCHAR2(100),	
M000	NUMBER(5),	
M100	NUMBER(5),	
M101	NUMBER(5),	
M102	NUMBER(5),	
Etc…		

);

En fin de profilage, les anomalies dans la source de données sont détectées, la table devrait ressembler à celle données ci-dessous en figure 10.

	?AnomalyNull	C1_STRING_ _UNKNOWN	C2_STRING_ _FIRSTNAME	C3_STRING_ _UNKNOWN	C4_DATE__DATE_ AAAAMMDD	C5 - STRING -- EMAIL	C6 - STRING -- BLOODGROUP	C7 - STRING -- GENDER	C8 - STRING -- WEIGHT _KG	C9 - STRING -- COUNTRY	C10 - STRING -- CONTINENT	C11 - NUMBER - - UNKNOWN	C12 - STRING -- TEMPERATURE	C13 - STRING -- MESURE_M
2		Id01	Alain	CLEMENT	1970-02-21	|<!>NULL	B+	M	76,49 KG	France	Europe	73	32.14°C	11,51 m
3		Id02	Inès	MIGNONNE	|<!>NULL	01/01/2020 |<!?!>ANOMALY	A	F	40159,31 g	France	Europe	146,35 €	35 |<!?!>ANOMALY	1 €|SEM#<!?!>ANOMALY
4		Id03	Omar	BELLE	|<!>NULL	?@ |SEM#<!?!>ANOMALY	R |SEM#<!?!>ANOMALY	M	84,48 KG	France	Europe	39	26.93°C	14,35 m
5		Id04	Rayan	MIGNONNE	1961-06-10	rayan.mignonne@hotmail.com	A-	M	61,49 KG	Italie	|<!>NULL	10	20.9°C	17,63 m
6		Id05	Eve	AIMANT	|<!>NULL	eve.aimant@gmail.com	1 |<!?!>ANOMALY	F	52606,37 g	Royaume-Uni	Europe	55,237	100.1°F	650,9 cm
7		Id06	Alain	MOUSTAFA	1969-12-26	alain.moustafa@yahoo.fr	++ |SEM#<!?!>ANOMALY	M	76,49 KG	Argentine	Amérique	60	26.93°C	17,63 m
8		Id07	Sabrine	PRINTEMPS	1962-11-14	sabrine.printemps@hotmail.com	AB+	F	54561,49g	Royaume-Uni	Europe	96,41 €	91.1°F	534,27 cm
9		Id08	Clément	DELOIN	1962-11-14	clément.deloin@gmail.com	AB+	A+|SEM#<!?!>ANOMALY	71,68 KG	Tunisie	Afrique	69	38.05°C	17,37 m
10		Id09	Emna	AIMANT	1970-04-14	emna.aimant@hotmail.com	++ |SEM#<!?!>ANOMALY	F	66360,8 g	Allemagne	Europe	109,344	87.53°F	424,22 cm
11		Id10	Claire	EXCELLE	1970-03-02	claire.excelle@gmail.com	AB-	F	46082,53 g	Tunisie	Afrique	50,491	|<!>NULL	744,55 cm
12		Id11	Mamadou	MOUSTAFA	1970-03-02	mamadou.moustafa@gmail.com	A+	M	58,83 KG	Algérie	Afrique	8	26.26°C	14,35 m
13		Id12	Jean	JOLIE	1978-12-28	01/01/2020 |<!?!>ANOMALY	AB+	M	75,34 KG	Belgique	Europe	13	32.14°C	14,35 m
14		Id13	Faouzi	GRANDE	1982-12-31	faouzi.grande@yahoo.fr	++ |SEM#<!?!>ANOMALY	M	71,68 KG	Tunisie	Afrique	27	26.26°C	9,58 m
15		Id14	Alexandre	FORT	1999-01-31	|<!>NULL	B+	M	71,68 KG	Tunisie	Afrique	46	26.93°C	1 |<!?!>ANOMALY
16		Id15	Clément	PARIS	|<!>NULL	clément.paris@yahoo.fr	|<!>NULL	0 |<!?!>ANOMALY	61,49 KG	Italie	|<!>NULL	60	20.9°C	0 |<!?!>ANOMALY
17		Id16	Alexandre	JOLIE	|<!>NULL	alexandre.jolie@gmail.com	B+	M	61,49 KG	Qatar	Asie	41	|<!>NULL	1 |<!?!>ANOMALY
18		Id17	Mamadou	JOLIE	1999-01-31	01/01/2020 |<!?!>ANOMALY	++ |SEM#<!?!>ANOMALY	M	76,49 KG	Chine	Asie	41	26.26°C	14,2 m
19		Id18	Alain	PARIS	1970-02-21	alain.paris@gmail.com	|<!>NULL	M	0 |<!?!>ANOMALY	Brésil	Amérique	51	26.93°C	17 m
20		Id19	Jean	FORT	1970-04-14	jean.fort@yahoo.fr	O+	M	76,49 KG	Algérie	Afrique	43	20.9°C	17 m
21		Id20	Médecin	FORT	|<!>NULL	médecin.fort@hotmail.com	++ |SEM#<!?!>ANOMALY	M	61,49 KG	Franc	|<!>NULL	16	26.26°C	11,51 m
22		Id21	Emna	BON	1999-01-31	emna.bon@yahoo.fr	AB	F	66360,8 g	Brésil	Amérique	78,593	|<!>NULL	530,6 cm
23		Id22	Ibrahim	MIGNONNE	1989-04-07	ibrahim.mignonne@gmail.com	|<!>NULL	0 |<!?!>ANOMALY	76,49 KG	France	Europe	23	32.14°C	19,85 m
24		Id23	Maria	MIGNONNE	1999-01-31	1 |<!?!>ANOMALY	AB+	F	64873,28 g	Qatar	Asie	79,33	|<!>NULL	530,6 cm
25		Id24	Adam	SPORTIF	1969-12-26	01/01/2020 |<!?!>ANOMALY	A-	M	61,49 KG	Espagne	|<!>NULL	56	32.14°C	11,51 m
26		Id25	Alexandre	SOLEIL	1970-04-14	alexandre.soleil@yahoo.fr	A+	M	84,48 KG	Japon	Asie	70	15.67°C	14,35 m
27		Id26	Omar	MOUSTAFA	1982-12-31	omar.moustafa@hotmail.com	++ |SEM#<!?!>ANOMALY	M	1 |<!?!>ANOMALY	Canada	Amérique	08/02/1900	|<!>NULL	0 |<!?!>ANOMALY
28		Id27	Clémence	PRINTEMPS	1970-04-03	clémence.printemps@gmail.com	|<!>NULL	F	64873,28 g	Italie	Europe	45,291	87.53°F	0 |<!?!>ANOMALY
29		Id28	Kenza	TROPFOR	1961-06-10	kenza.tropfor@hotmail.com	A+	F	52606,37 g	Tunisie	|<!>NULL	09/02/1900	111.29°F	530,6 cm
30		Id29	Rayan	TROPFOR	1962-11-14	rayan.tropfor@hotmail.com	++ |SEM#<!?!>ANOMALY	M	63,3 KG	Qatar	Asie	66	26.93°C	0 |<!?!>ANOMALY

Figure 10 : Les anomalies dans la source de données sont détectées

2. Profilage SEMANTIQUE

CSV2TABLE (From a CSV file TO a TABLE with a set of columns)

A CSV File	A TABLE with COLUMNS (DBMS)					
Comma-separated values, a file with only one column Each value is seen as a string of characters	Col1	Col2	Col3	Col4	Col5	Col6
	STRING	STRING	STRING	NUMBER	DATE	NUMBER
	FIRSTNAME	CITY	GENDER			TEMPERATURE
	7	7	1		10	
	INITCAP	UPPERCASE	UPPERCASE (F/M)		Year-Month-Day	°C
Adam;Paris;M;19;19-06-2001;38°C	Adam	PARIS	M	19	2001-06-19	38°C
Eve;Paris;F;23;16-10-1996;37°C	Eve	PARIS	F	23	1996-10-16	37°C
Gabriel;Paris;m;18;17-09-2002;36,5°C	Gabriel	PARIS	M	18	2002-09-17	36,5°C
Mariam;Paris;F;41;13-08-1978;38Celcius	Mariam	PARIS	F	41	1978-08-13	38°C
Nadia;Londres;f;55;10-10-1965;95°F	Nadia	LONDRES	F	55	1965-10-10	35°C
Inès;Madrid;F;50;22-11-1969;99,5°F	Inès	MADRID	F	50	1969-11-22	37,5°C
Inconnu;77;12-12-2012	Inconnu			77	2012-12-12	
Abnomly;Rome;1;88;02-10-2019;38°C	Abnomly	ROME	1	88	2019-10-02	38°C
Anomalies;Tunis;f;99;25-30-2020;x	Anomalies	TUNIS	F	99	25-30-2020	x

A vous d'entamer la réflexion ! Vous les expert.e.s !

1. Découpage du fichier CSF (composé d'une seule colonne) en une table composée de plusieurs colonnes
2. Profilage de chacune des colonnes (Type de données, contenu)
3. Homogénéisation des données
4. Détection des anomalies syntaxiques
5. Détection des anomalies sémantiques
6. Corrections des anomalies
7. Etc…
8.

Ce travail est à tester sur plusieurs SGBD différents tels que : Oracle, MySQL, MongoDB, Access, SQLServer…

FROM a CSV file TO a Table with Columns! CSV-2-TABLE

```
Adam;Paris;M;19;19-06-2001;38°C
Eve;Paris;F;23;16-10-1996;37°C
Gabriel;Paris;m;18;17-09-2002;36,5°C
Mariam;Paris;F;41;13-08-1978;38Celcius
Nadia;Londres;f;55;10-10-1965;95°F
Inès;Madrid;F;50;22-11-1969;99,5°F
Inconnu;77;12-12-2012
Abnomly;Rome;1;88;02-10-2019;38°C
Anomalies;Tunis;f;99;25-30-2020;x
Adam;Paris;M;19;19-06-2001;38°C
Eve;Paris;F;23;16-10-1996;37°C
Marie;Pari;F;41;17-09-1979;38Celcius
```

```
CREATE TABLE CSVfile (Col VARCHAR2(1000));

INSERT INTO CSVfile VALUES ('Adam;Paris;M;19;19-06-2001;38°C');
INSERT INTO CSVfile VALUES ('Eve;Paris;F;23;16-10-1996;37°C');
INSERT INTO CSVfile VALUES ('Gabriel;Paris;m;18;17-09-2002;36,5°C');
INSERT INTO CSVfile VALUES ('Mariam;Paris;F;41;13-08-1978;38Celcius');
INSERT INTO CSVfile VALUES ('Nadia;Londres;f;55;10-10-1965;95°F');
INSERT INTO CSVfile VALUES ('Inès;Madrid;F;50;22-11-1969;99,5°F');
INSERT INTO CSVfile VALUES ('Inconnu;77;12-12-2012');
INSERT INTO CSVfile VALUES ('Abnomly;Rome;1;88;02-10-2019;38°C');
INSERT INTO CSVfile VALUES ('Anomalies;Tunis;f;99;25-30-2020;x');
INSERT INTO CSVfile VALUES ('Adam;Paris;M;19;19-06-2001;38°C');
INSERT INTO CSVfile VALUES ('Eve;Paris;F;23;16-10-1996;37°C');
INSERT INTO CSVfile VALUES ('Marie;Pari;F;41;17-09-1979;38Celcius');

COMMIT;
```

3. Correction de quelques anomalies

NB : **Quelques idées pour corriger : Autocorrection basée sur la fréquence afin de choisir la valeur correcte.**

Autocorrection basée sur fréquence afin de choisir la valeur correcte (intra-colonne).

Parcours de chaque colonne de la table, Calculer pour chaque valeur sa fréquence puis rapprocher les valeurs similaires grâce à une combinaison des algorithmes de Levenshtein (Levenshtein, 1966) et du Soundex (Cohen, 1998).
Pour ce faire, on utilise des vues intermédiaires afin d'aider dans le traitement.
En premier lieu on va créer une vue V1 contenant les valeurs, leurs occurrences et par exemple leurs soundex respectifs.
Ensuite, une nouvelle vue V2 sera créée pour chaque valeur.
Pour finir, on récupère la valeur maximale qui sera considérée comme la valeur de correction.
On peut générer un fichier contenant les mises à jour à réaliser.
Une fois la valeur correcte identifiée, les valeurs similaires seront remplacées par celle-ci.

Val	Num	soundex
	24	
Algérie	802	A426
Allemangne	381	A4525
Almania	4	A450
Argentine	323	A62535
Belgique	282	B420
Brésil	289	B624

val	num	soundex	levenshtein
Algeria	1	A426	1
algeri	1	A426	0
Algérie	802	A426	1

col	col2
Alger	Algérie
Algeri	Algérie
Algeria	Algérie
Algérie	Algérie

Autocorrection basée sur la fréquence afin de choisir la valeur correcte (inter-colonnes).

PaysContinent	num	sound1	sound2
-Afrique	10		A162
-Asie	6		A200
-Europe	8		E610
Algérie-	2	A426	
Algérie-Afrique	800	A426	A162
Allemangne-Europe	381	A4525	E610
Almania-Ourouppa	4	A450	O610
Argentine-Amérique	323	A62535	A562
Belgique-Europe	282	B420	E610
Brésil-	1	B624	
Brésil-Amérique	288	B624	A562
Britania-Ourouppa	4	B635	O610
Canada-	1	C530	
Canada-Amérique	323	C530	A562

Val	Num
Tunisie-	1
Tunisie-Afrique	1449