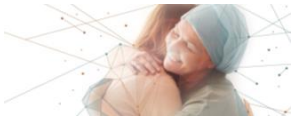
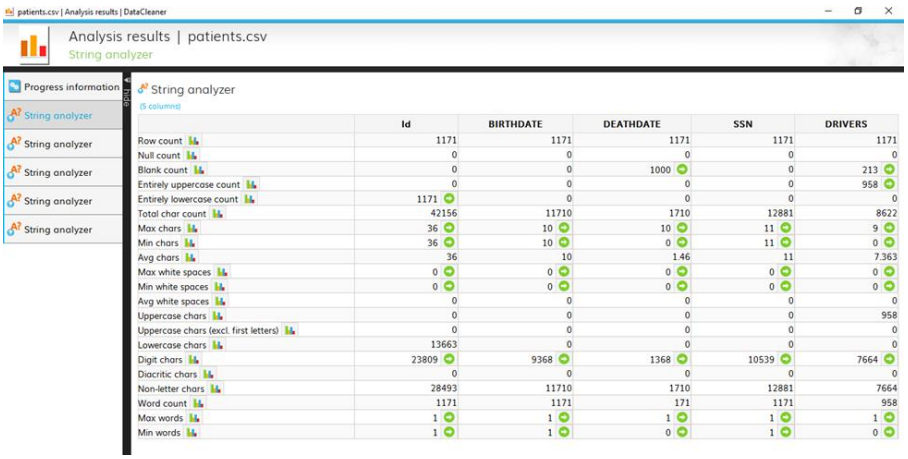


DATACLEANER

COMMUNITY EDITION - Limited

Data profiling, data cleaning, and data integration tool - offers integration with Pentaho



	Id	BIRTHDATE	DEATHDATE	SSN	DRIVERS
Row count	1171	1171	1171	1171	1171
Null count	0	0	0	0	0
Blank count	0	0	1000	0	213
Entirely uppercase count	0	0	0	0	958
Entirely lowercase count	1171	0	0	0	0
Total char count	42156	11710	1710	12881	8622
Max chars	36	10	10	11	9
Min chars	36	10	0	11	0
Avg chars	36	10	1.46	11	7.363
Max white spaces	0	0	0	0	0
Min white spaces	0	0	0	0	0
Avg white spaces	0	0	0	0	0
Uppercase chars	0	0	0	0	958
Uppercase chars (excl. first letters)	0	0	0	0	0
Lowercase chars	13663	0	0	0	0
Digit chars	23809	9368	1368	10539	7664
Diacritic chars	0	0	0	0	0
Non-letter chars	28493	11710	1710	12881	7664
Word count	1171	1171	171	1171	958
Max words	1	1	1	1	1
Min words	1	1	0	1	0

Links

- ✓ http://datacleaner.org/docs
- ✓ https://github.com/datacleaner/DataCleaner
- ✓ https://travis-ci.org/datacleaner/DataCleaner
- ✓ https://datacleaner.org/faq

License GNU Lesser General Public License v3.0
Version 5.7.0
Last Update 4/1/2019
OS Linux, macOS, Windows

System Requirements Computer (with a graphical display, except if run in command-line mode); Java Runtime Environment (JRE), version 7 or higher and a DataCleaner software license file for professional editions.

Description **Community Edition is free, otherwise you need a subscription**. DataCleaner is a Data Quality toolkit that allows you to profile, correct and enrich your data. People use it for ad-hoc analysis, recurring cleansing as well as a swiss-army knife in matching and Master Data Management solutions.
DataCleaner is a data profiling engine for discovering and analyzing the quality of user's data. It is built to handle data big and small from CSV files, Excel spreadsheets to RDBMs and NoSQL databases. User can build their own cleansing rules and compose them into several use scenarios or target databases whether it is simple search/replace rules, regular expressions, and pattern matching or completely custom transformations.
DataCleaner's Monitoring establishes the starting point and goals, and to ensure a process of following up on data quality issues. The monitoring server of DataCleaner enables users to make point-in-time profiles of users' data, and to schedule periodic data quality checks and receive notifications if quality KPIs get out of control.
DataCleaner's Data Quality Eco-System delivers out-of-the-box functionality, and application extensions, integrations and shared content.
DataCleaner's Duplicate detection feature builds on Machine Learning principles and inferential matching.

Features

- Data Quality Analysis and Profiling
- Duplicate Detection
- Data Standardization and Cleansing
- Data Health Monitoring
- Data Quality Eco-System
-

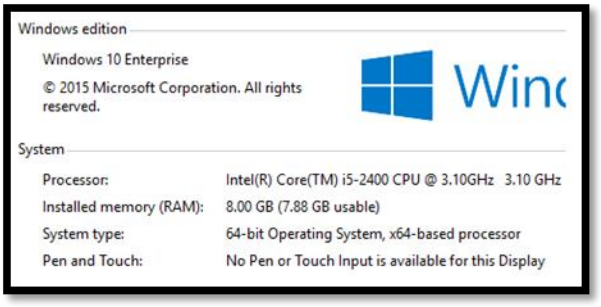
Connectivity / Supported Data Sources & Formats

- CSV files, Excel spreadsheets
- JDBC, MySQL, PostrgreSQL, SQL Server
- Salesforce, SugarCRM

Limitations

DataCleaner requires people with a technical background. The Open Source version of this tool is not able to process the 1.3M records.

Performance



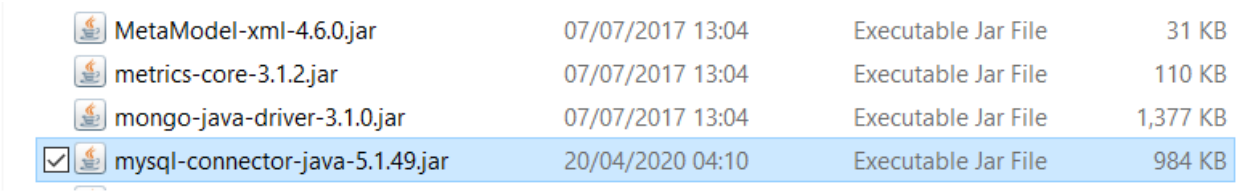
DataCleaner was not able to process 1.3 records

Feedback from Neonatal Team:

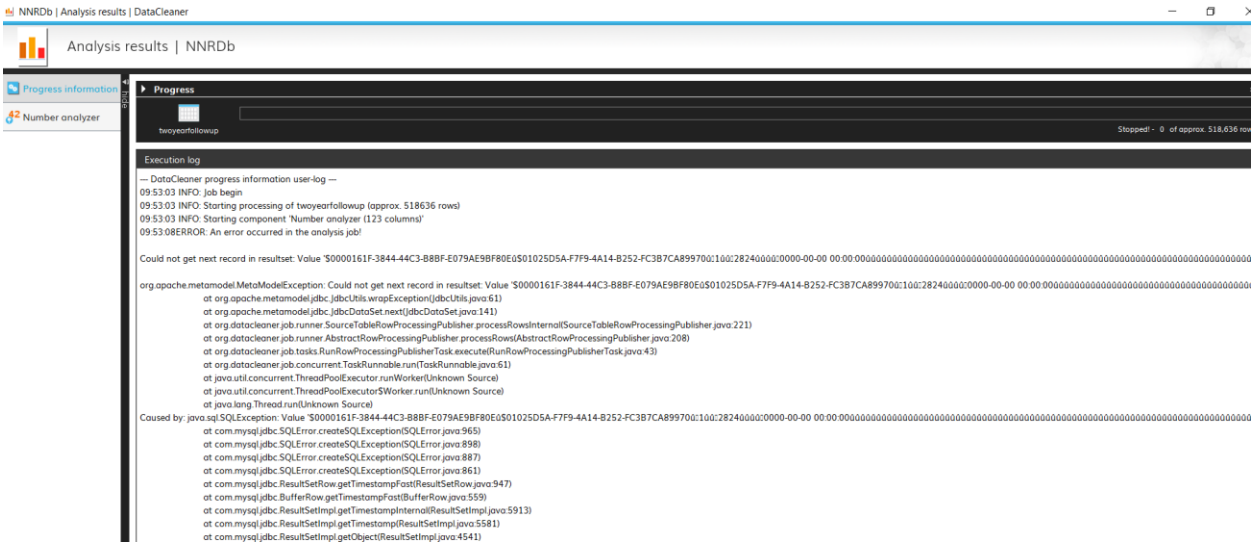
Victor L Banda, a Data Analyst from the Neonatal Data Analysis Unit, Neonatal Medicine Research Group, Imperial College London, Chelsea and Westminster Hospital provided the following feedback:

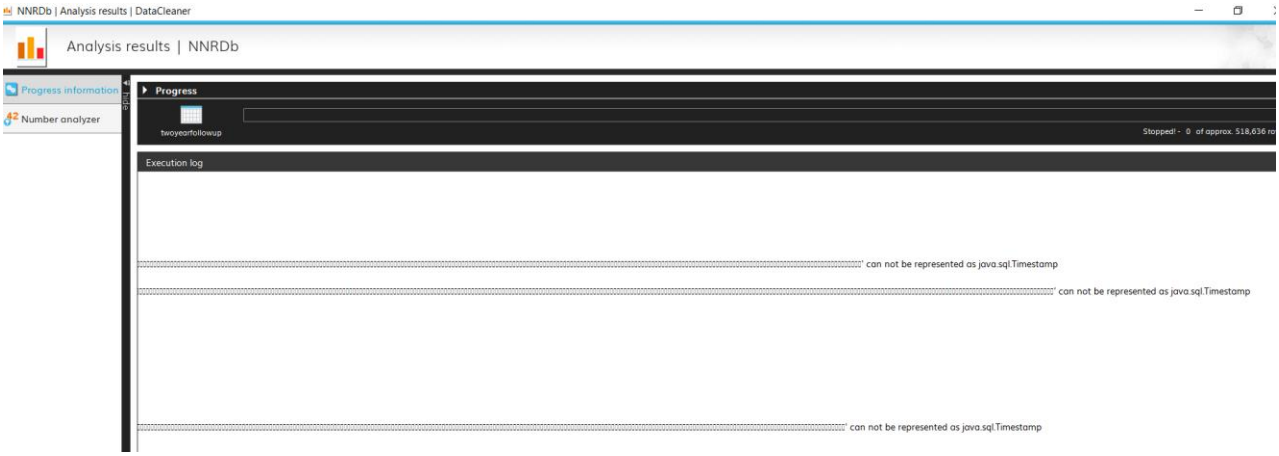
DataCleaner was relatively easy to setup, even on a windows machine. Was possible to link to a MySQL database. Was possible to run analysis on more than one table at once.

I had to change the MySQL connector to a more recent version to match support for a MySQL 5.7 database.



To access timeliness on criteria 3, a table with datetime fields was chosen, and due to a default value of 0000-00-00, the following error presented:





Which was resolved by adding a zeroDateTimeBehavior=convertToNull property to the connection string. i.e. jdbc:mysql://localhost:3306/NNRDb?defaultFetchSize=-2147483648&largeRowSizeThreshold=1024&zeroDateTimeBehavior=convertToNull

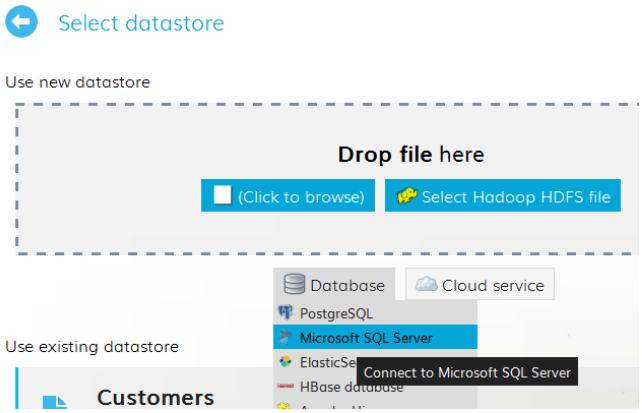
DataCleaner tends to crash when it reaches around 170,000 records.

The supporting document "**Data Quality Tool Assessment Request - Data Custodian Neonatal.docx**" contains all feedback provided by the Neonatal team.

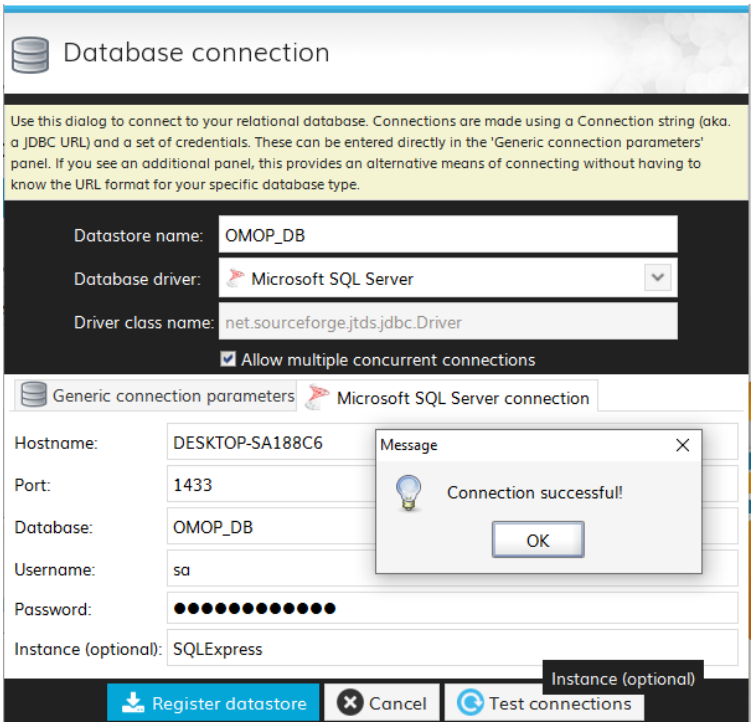
DATACLEANER DATA ANALYSIS

1. Download Data Cleaner from "https://datacleaner.org/get_datacleaner_ce"

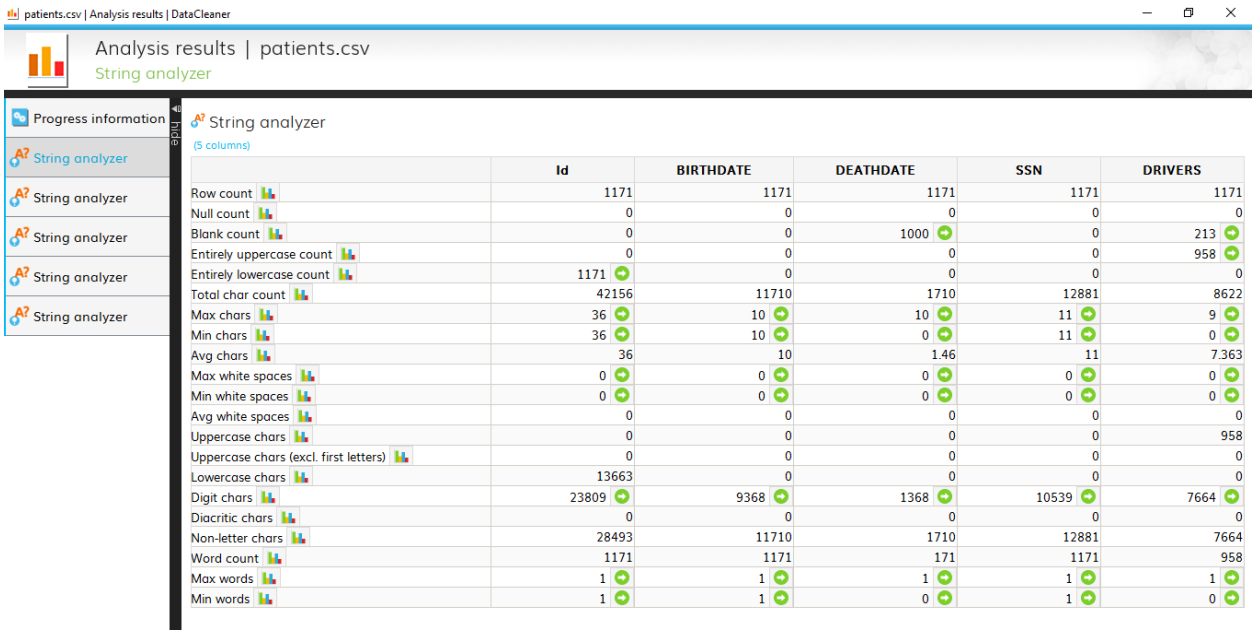
2. Open "DataCleaner.Exe" to start the application. Configure the DB as shown below,



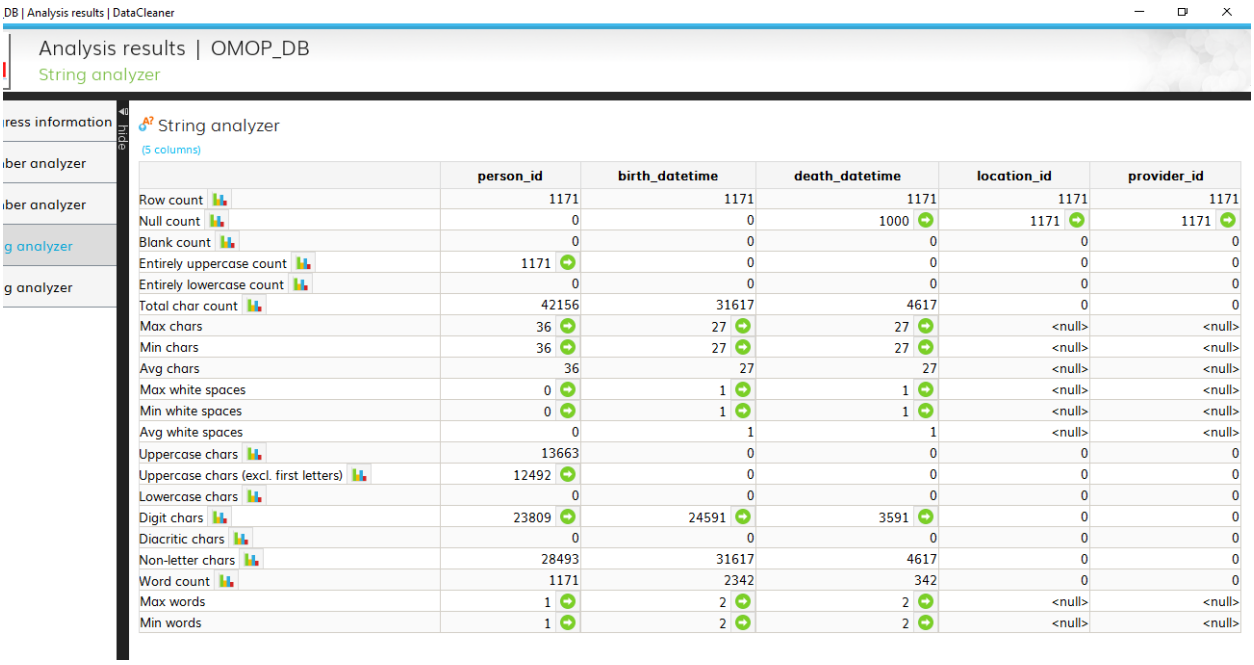
Step 2:



SAMPLE STATISTICS:

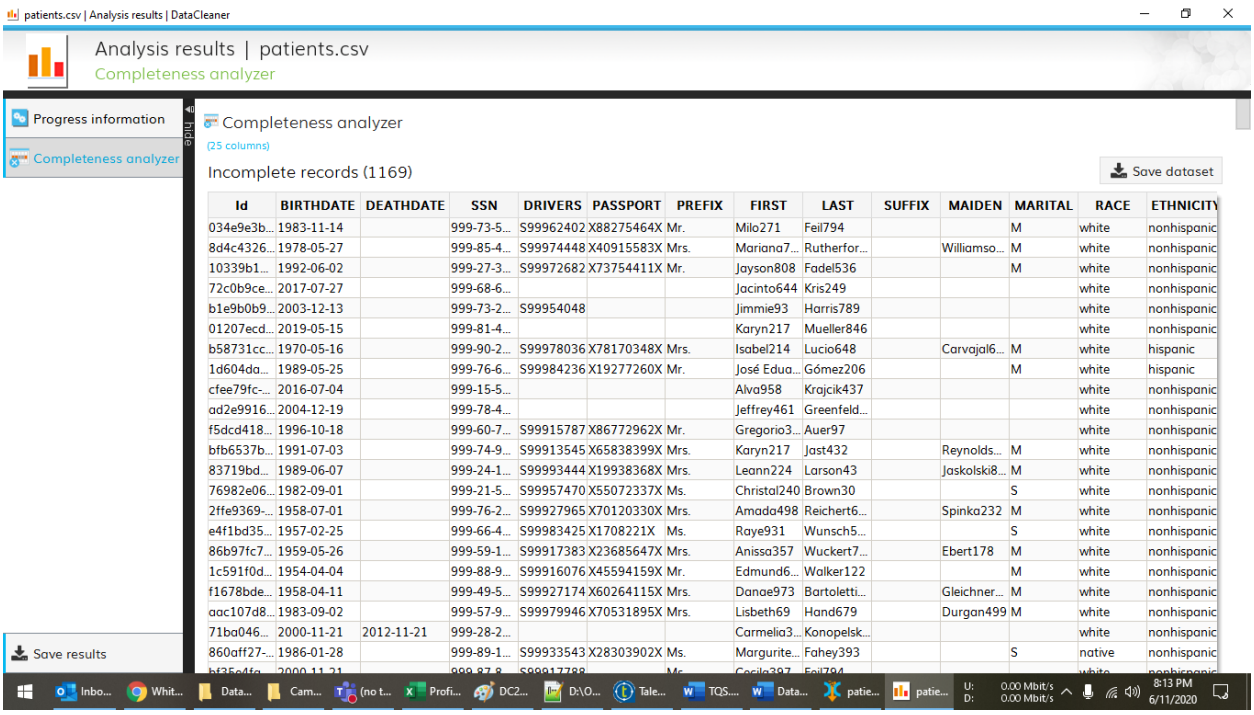


	Id	BIRTHDATE	DEATHDATE	SSN	DRIVERS
Row count	1171	1171	1171	1171	1171
Null count	0	0	0	0	0
Blank count	0	0	1000	0	213
Entirely uppercase count	0	0	0	0	958
Entirely lowercase count	1171	0	0	0	0
Total char count	42156	11710	1710	12881	8622
Max chars	36	10	10	11	9
Min chars	36	10	0	11	0
Avg chars	36	10	1.46	11	7.363
Max white spaces	0	0	0	0	0
Min white spaces	0	0	0	0	0
Avg white spaces	0	0	0	0	0
Uppercase chars	0	0	0	0	958
Uppercase chars (excl. first letters)	0	0	0	0	0
Lowercase chars	13663	0	0	0	0
Digit chars	23809	9368	1368	10539	7664
Diacritic chars	0	0	0	0	0
Non-letter chars	28493	11710	1710	12881	7664
Word count	1171	1171	171	1171	958
Max words	1	1	1	1	1
Min words	1	1	0	1	0



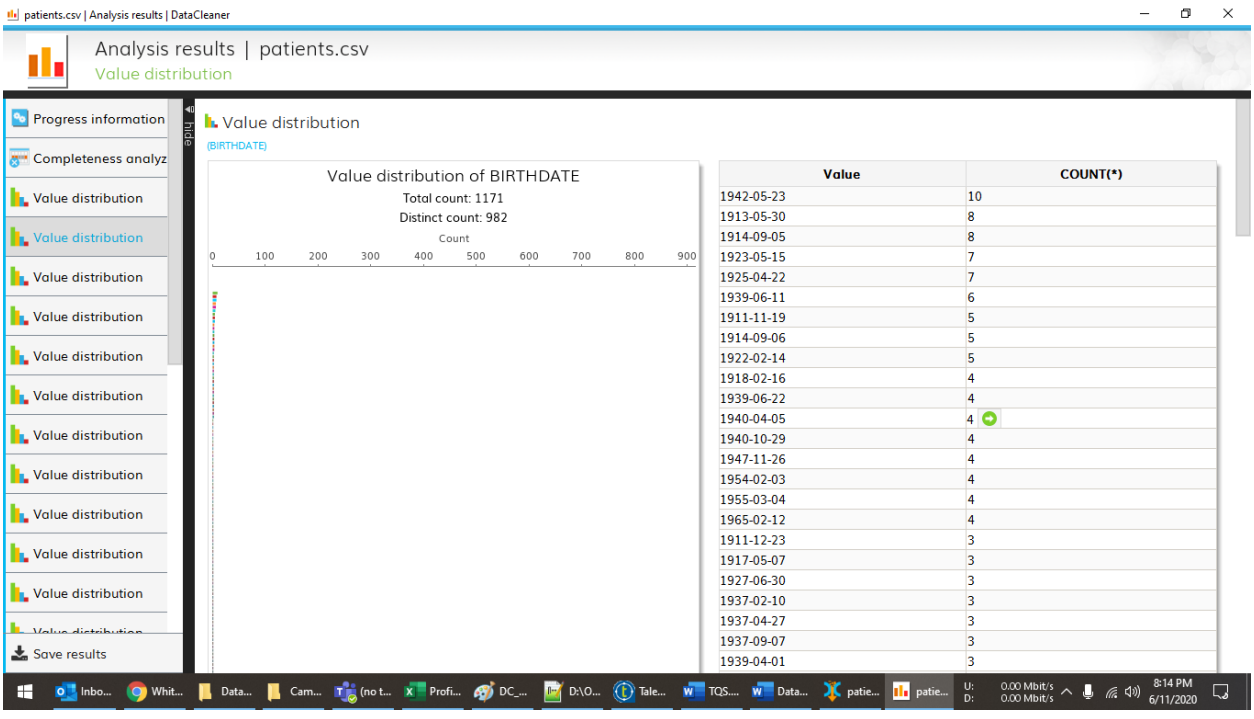
	person_id	birth_datetime	death_datetime	location_id	provider_id
Row count	1171	1171	1171	1171	1171
Null count	0	0	1000	1171	1171
Blank count	0	0	0	0	0
Entirely uppercase count	1171	0	0	0	0
Entirely lowercase count	0	0	0	0	0
Total char count	42156	31617	4617	0	0
Max chars	36	27	27	<null>	<null>
Min chars	36	27	27	<null>	<null>
Avg chars	36	27	27	<null>	<null>
Max white spaces	0	1	1	<null>	<null>
Min white spaces	0	1	1	<null>	<null>
Avg white spaces	0	1	1	<null>	<null>
Uppercase chars	13663	0	0	0	0
Uppercase chars (excl. first letters)	12492	0	0	0	0
Lowercase chars	0	0	0	0	0
Digit chars	23809	24591	3591	0	0
Diacritic chars	0	0	0	0	0
Non-letter chars	28493	31617	4617	0	0
Word count	1171	2342	342	0	0
Max words	1	2	2	<null>	<null>
Min words	1	2	2	<null>	<null>

COMPLETENESS:



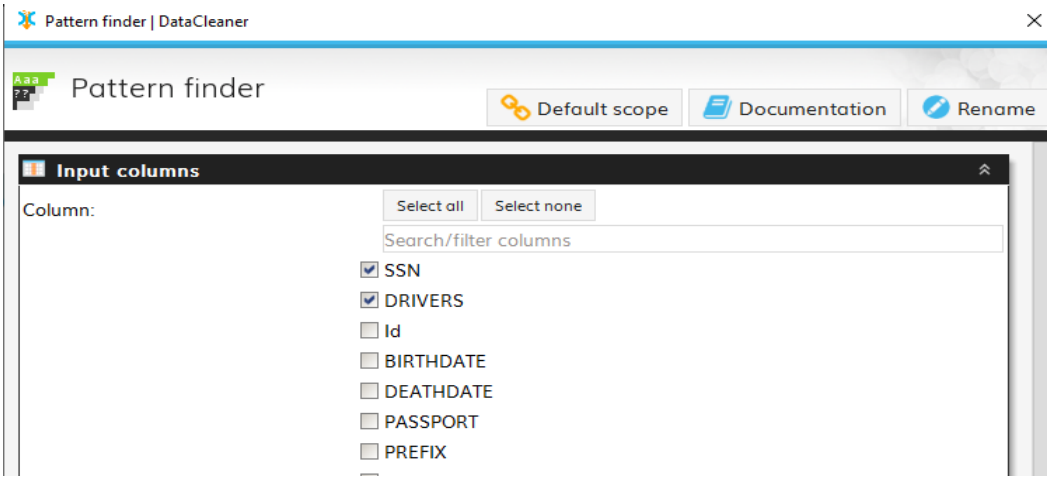
Id	BIRTHDATE	DEATHDATE	SSN	DRIVERS	PASSPORT	PREFIX	FIRST	LAST	SUFFIX	MAIDEN	MARITAL	RACE	ETHNICITY
034e9e3b...	1983-11-14		999-73-5...	S99962402	X88275464X	Mr.	Milo271	Feil794			M	white	nonhispanic
8d4c4326...	1978-05-27		999-85-4...	S99974448	X40915583X	Mrs.	Mariana7...	Rutherfor...		Williamso...	M	white	nonhispanic
10339b1...	1992-06-02		999-27-3...	S99972682	X73754411X	Mr.	Jayson808	Fadel536			M	white	nonhispanic
72c0b9ce...	2017-07-27		999-68-6...				Jacinto644	Kris249				white	nonhispanic
b1e9b0b9...	2003-12-13		999-73-2...	S99954048			Jimmie93	Harris789				white	nonhispanic
01207ecd...	2019-05-15		999-81-4...				Karyn217	Mueller846				white	nonhispanic
b58731cc...	1970-05-16		999-90-2...	S99978036	X78170348X	Mrs.	Isabel214	Lucio648		Carvajal6...	M	white	hispanic
1d604da...	1989-05-25		999-76-6...	S99984236	X19277260X	Mr.	José Edua...	Gómez206			M	white	hispanic
cfee79fc-...	2016-07-04		999-15-5...				Alva958	Krajcik437				white	nonhispanic
ad2e9916...	2004-12-19		999-78-4...				Jeffrey461	Greenfeld...				white	nonhispanic
f5dcd418...	1996-10-18		999-60-7...	S99915787	X86772962X	Mr.	Gregorio3...	Auer97				white	nonhispanic
bfb6537b...	1991-07-03		999-74-9...	S99913545	X65838399X	Mrs.	Karyn217	Jast432		Reynolds...	M	white	nonhispanic
83719bd...	1989-06-07		999-24-1...	S99993444	X19938368X	Mrs.	Leann224	Larson43		Jaskolski8...	M	white	nonhispanic
76982e06...	1982-09-01		999-21-5...	S99957470	X55072337X	Ms.	Christal240	Brown30			S	white	nonhispanic
2ffe9369-...	1958-07-01		999-76-2...	S99927965	X70120330X	Mrs.	Amada498	Reichert6...		Spinka232	M	white	nonhispanic
e4f1bd35...	1957-02-25		999-66-4...	S99983425	X1708221X	Ms.	Raye931	Wunsch5...			S	white	nonhispanic
86b97fc7...	1959-05-26		999-59-1...	S99917383	X23685647X	Mrs.	Anissa357	Wuckert7...		Ebert178	M	white	nonhispanic
1c591f0d...	1954-04-04		999-88-9...	S99916076	X45594159X	Mr.	Edmund6...	Walker122			M	white	nonhispanic
f1678bde...	1958-04-11		999-49-5...	S99927174	X60264115X	Mrs.	Danae973	Bartoletti...		Gleichner...	M	white	nonhispanic
aac107d8...	1983-09-02		999-57-9...	S99979946	X70531895X	Mrs.	Lisbeth69	Hand679		Durgan499	M	white	nonhispanic
71ba046...	2000-11-21	2012-11-21	999-28-2...				Carmelia3...	Konopelsk...				white	nonhispanic
860aff27-...	1986-01-28		999-89-1...	S99933543	X28303902X	Ms.	Margurite...	Fahey393			S	native	nonhispanic

DISTRIBUTIONS:



Value	COUNT(*)
1942-05-23	10
1913-05-30	8
1914-09-05	8
1923-05-15	7
1925-04-22	7
1939-06-11	6
1911-11-19	5
1914-09-06	5
1922-02-14	5
1918-02-16	4
1939-06-22	4
1940-04-05	4
1940-10-29	4
1947-11-26	4
1954-02-03	4
1955-03-04	4
1965-02-12	4
1911-12-23	3
1917-05-07	3
1927-06-30	3
1937-02-10	3
1937-04-27	3
1937-09-07	3
1939-04-01	3

PATTERN MATCHING/FINDER: EXPECTED VALUES VS UNEXPECTED

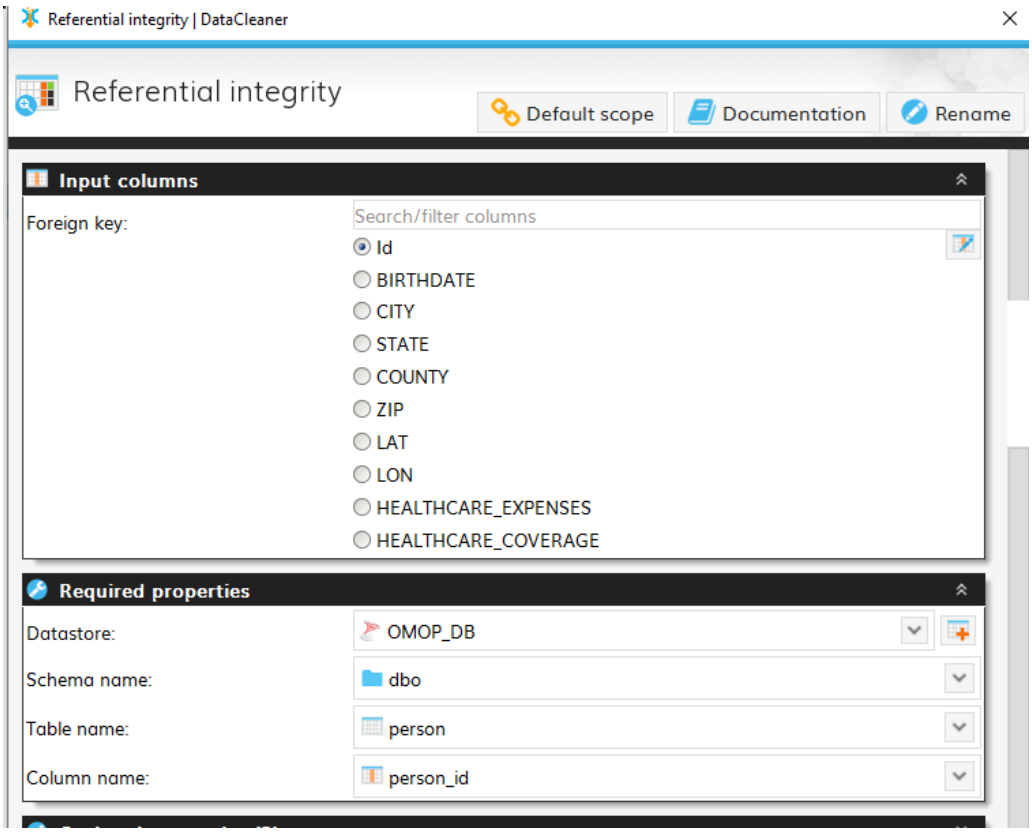


Pattern finder

(DRIVERS)

	Match count	Sample
?????????	958	S99984236
<blank>	213	<blank>

REFERNTIAL INTEGRITY:



LIMITATIONS:

- The tool is not able to process the 1.3M records.